SCTE Technical Journal

Volume 5, Issue 1 | August 2025

Table of Contents

Editorial Correspondence: If there are errors or omissions to the information provided in this journal, corrections may be sent to our editorial department at journals@scte.org.

Submissions: Have an idea or topic for a future journal article? Send your submission to journals@scte.org. All submissions are subject to peer review and publication is at the discretion of SCTE.

Table of Contents

Editorial Correspondence: If there are errors or omissions to the information provided in this journal, corrections may be sent to our editorial department at journals@scte.org.

Submissions: Have an idea or topic for a future journal article? Send your submission to journals@scte.org. All submissions are subject to peer review and publication is at the discretion of SCTE.

Foreword

A Note from SCTE

SCTE TechExpo is where the NETWORK (R)EVOLUTION comes to life. This year's theme, *"Delivering the Seamless Experience,"* reflects the industry's bold shift toward networks that unify wireline, wireless, and converged infrastructures into a foundation for limitless possibilities. Broadband operators and their partners are embracing AI, automation, advanced planning, and new approaches to security, operations, and construction, all with a singular focus: delivering the seamless experiences customers demand. This issue of the SCTE Journal captures that momentum, presenting eight papers that illustrate how automation, intelligence, collaboration, and sustainability are reshaping telecommunications from the ground up.

Enhancing Network Operations with Automation explores how AI-driven diagnostics and DevNetOps practices reduce manual intervention, predict failures, and enable scalable, efficient operations. Extending beyond the individual operator, Novel Network Sharing Architectures examines new multi-operator and non-3GPP models that promote collaboration, improve spectrum use, and accelerate service deployment.

Intelligence-driven optimization is a recurring theme. AI Optimization in HFC Networks demonstrates how adaptive analytics transform cable systems into predictive infrastructures, while Preventing Disruptions with a Digital Twin highlights how digital replicas can validate planned cutovers in optical networks and minimize service risk. At the physical layer, Drop Cable Analysis with OFDM Channel Estimates shows how proactive monitoring techniques provide precise fault detection in coaxial cables.

This issue also considers operational and environmental sustainability. LLM-Enhanced Part Similarity applies large language models to streamline inventory management and reduce waste, while Passive Thermal Management for Telecom Cabinets proposes climate-specific cooling solutions to lower energy use. Finally, Energy Projections for Grid Demand and Backup Needs examines future grid challenges in high-density areas, underscoring the importance of energy resilience.

Together, these contributions demonstrate a clear trajectory: networks are becoming more automated, collaborative, intelligent, and sustainable. These qualities will define the next generation of communication infrastructure and ensure its ability to meet the demands of a connected world.

As we look forward to gathering at TechExpo25, these papers provide both the technical depth and the forward-looking vision that will drive the conversations, collaborations, and solutions on display at the event. We hope this issue inspires your thinking and fuels meaningful dialogue in the months ahead, and we look forward to seeing you at TechExpo, where these ideas will come to life.

Enhancing Network Operations

Leveraging Automation for Efficiency, Reliability, and Scalability

An Operational Practice prepared for SCTE by

Allen Maharaj, Sr Manager – HFC Network Operations, Rogers Communications
8200 Dixie Rd
Brampton, ON, L6T 0C1
Allen.maharaj@rci.rogers.com
416-450-4820

Christian Corbin, Sr DevOps Tech Lead, Rogers Communications
861 Cloverdale Ave
Victoria, BC, V8X 4S7
Christian.corbin@rci.rogers.com
403-648-5770

Table of Contents

1. Introduction

1.1. Overview of Workforce Shortages in Network Operations

The telecommunications industry is at a critical juncture where demand for high-capacity, low-latency, and always-on connectivity continues to intensify. From streaming and online gaming to remote work, smart homes, and IoT deployments, modern networks are expected to deliver near-perfect performance under increasingly complex conditions. As networks evolve in capability and scale, the workforce responsible for building, operating, and maintaining them is shrinking in both size and availability.

A significant contributor to this challenge is a systemic shortage of skilled network professionals. Many operators report difficulty filling technical roles, particularly those requiring hybrid expertise in IP networking, RF engineering, software automation, and cloud infrastructure[1]. Contributing factors include an aging workforce, a limited pipeline of new technicians entering the field, and an education-to-execution gap where traditional certifications and degree programs lag real-world network demands. At the same time, experienced engineers are tasked with supporting both legacy infrastructure (e.g., coaxial and TDM systems) and emerging technologies like DOCSIS 4.0, PON, and 5G, creating an unsustainable burden.

This labor shortage not only leads to increased workload and burnout but also impacts operational KPIs. Routine maintenance, fault resolution, and provisioning require manual intervention and on-site visits, resulting in elevated operational costs, longer time-to-resolve (TTR), and customer dissatisfaction. Without intervention, the gap between operational demand and available technical resources will continue to widen, threatening the scalability and reliability of critical communications infrastructure.

1.2. Role of Automation in Network Operations

To bridge this widening gap, automation has become an essential pillar of modern network operations. By leveraging tools and technologies that can perform routine, time-consuming, or repetitive tasks, network operators are increasingly able to decouple operational scalability from workforce growth. Automation is no longer limited to basic configuration scripting; it now encompasses end-to-end capabilities such as diagnostics, self-healing mechanisms, provisioning, network assurance analytics, and zero-touch deployments.

For example, predictive maintenance systems use telemetry data, machine learning models, and historical trends to identify network elements with declining performance before this degradation causes service disruption. Predictive analysis enables proactive resolution before faults trigger customer complaints, without relying on expensive periodic field inspections. Similarly, remote troubleshooting platforms integrated with CMTS/OLT systems, cable modems, or ONTs can assess signal impairments, initiate resets, push firmware updates, or re-optimize service configurations—all without dispatching a field technician.

These advancements drastically reduce truck rolls, improve mean-time-to-repair (MTTR), and increase first-time resolution rates. More importantly, they allow skilled engineers to focus their efforts on higher-value functions—such as root cause analysis, infrastructure design, or performance tuning, rather than investigating routine escalations.

[1] Tech talent gap | Deloitte Insights

Contrary to concerns that automation will render technical roles obsolete, it can instead be seen as a force multiplier that enhances productivity, consistency, and accuracy.

1.3. The Emergence of DevNetOps

As networks grow more complex and software-driven, the traditional divide between development, network engineering, and operations is becoming a bottleneck. **DevNetOps** is a modern approach that unifies these disciplines by applying **DevOps principles**—such as automation, continuous integration/deployment (CI/CD), version control, and rapid feedback loops—to network infrastructure.

This methodology treats network configurations as code, enabling repeatable, automated deployments with improved reliability and speed. By adopting **infrastructure-as-code**, leveraging **APIs**, and encouraging cross-functional collaboration, DevNetOps allows network teams to manage large, dynamic environments with the same agility and efficiency seen in cloud and software engineering.

At its core, DevNetOps is not just a set of tools—it's a cultural shift. It breaks down organizational silos, accelerates delivery, and empowers engineers to build, test, and operate networks in a unified, software-centric way.

2. Current Workforce Challenges in Network Operations

2.1. Skills Gap in Network Operations

The shortage of skilled network professionals has become one of the most critical threats to the sustainability of network operations. As legacy copper and coaxial infrastructures converge with modern optical, virtualized, and cloud-based platforms, the role of the network technician has evolved dramatically. Today's field and NOC personnel are expected not only to understand RF modulation and Layer 2/3 protocols but also to be familiar with technologies like spine-leaf CIN, SD-WAN, DOCSIS 4.0, Passive Optical Networks (PON), Distributed Access Architecture (DAA), MPLS, IPv6, SDN, NFV, Containerization, Linux, Kubernetes, and an ever-increasing list of buzzword modernization. This shift demands a generalist hybrid skill set that is increasingly difficult to find.

Contributing to this problem is a rapidly aging workforce. In North America and Europe, many of the most experienced RF and plant engineers are approaching retirement, and the replacement pipeline is inadequate. [2]Technical schools and certification programs often lag behind industry needs, failing to incorporate software-centric skills such as scripting, API integration, and automation frameworks. The result is a growing experience and competency gap that places strain on both operations teams and service delivery timelines.

Compounding this is the increased reliance on contractors and third-party integrators, which can introduce inconsistency, reduce institutional knowledge, and hinder long-term process standardization. However, contractors can also provide important short-term advantages, like offer scalability and agility, enabling operators to meet temporary spikes in demand without long-term staffing commitments. Contractors will often bring specialized expertise from varied projects and vendors. In an industry where uptime, precision, and accountability are paramount, the labor shortfall is not merely an HR problem—it is a risk to service quality and business continuity.

[2] Urgent Need to Recruit and Train Nearly 180,000 Workers to Complete Federal- and State-Funded Broadband Networks - Fiber Broadband Association

2.2. Operational Inefficiencies

Networks are no longer static systems designed solely to deliver internet access. They are now dynamic service fabrics supporting a wide range of applications including cloud gaming, telemedicine, smart city infrastructure, IoT, 5G backhaul, and latency-sensitive business services. This evolution has significantly increased both architectural and operational complexity.

Operators now deploy hybrid networks composed of HFC, fiber, fixed wireless access (FWA), and even low-earth orbit satellite backhaul in certain markets. These technologies are supported by a variety of systems including virtual cable modem termination systems (vCMTS), software-defined access nodes, and cloud-native orchestration layers. Service assurance, which was once based largely on SNMP polling and alarm thresholds, now requires real-time telemetry, data correlation across multiple layers, and automated root cause analysis.

The advent of distributed access architectures (DAA) like Remote PHY and Remote MACPHY pushes functionality deeper into the network, dispersing control and operational visibility. A fault in a single RPD (Remote PHY Device) can manifest as customer issues across dozens of homes, requiring operators to trace problems across transport, power, optical links, and RF domains.

The number of "moving parts" in these networks—ranging from microservice-based provisioning systems to traffic management platforms, makes it increasingly difficult for engineers to troubleshoot using manual methods alone. Even identifying the boundary of a fault domain often requires sophisticated tools and cross-domain expertise.

In this environment, human operators are overwhelmed by data volume and system complexity. Without automation and intelligent decision support, the mean time to identify and resolve (MTTI and MTTR) increases, ultimately degrading customer experience and impacting key performance indicators like uptime and churn.

2.3. Impact of Workforce Shortages

Despite the digital transformation of network infrastructure, a large portion of fault detection, diagnosis, and resolution still relies on manual processes. This often results in reactive operations characterized by long mean time to repair (MTTR), inefficient use of technical staff, and costly truck rolls.

Truck rolls—dispatching technicians to troubleshoot issues on-site, are not only financially burdensome but also resource-inefficient. Each visit incurs direct costs in fuel, vehicle maintenance, and technician time, along with opportunity costs from diverting staff away from preventive or strategic tasks. According to industry benchmarks, the average cost per truck roll ranges from $150 to over $300 depending on geography, labor market, and operational scope[3]. When multiplied across hundreds of thousands of service calls annually, this represents a significant drain on OPEX.

More troubling is that many of these site visits are avoidable. Studies across cable and fiber operators reveal that 30–50% of truck rolls are dispatched for issues that could have been resolved remotely with the right automation and diagnostic tools[4].

[3] What is a truck roll and how to reduce them? | CareAR
[4] AI is cables secret weapon for 10g reliability

Examples include:

- Devices operating with outdated firmware that could be pushed remotely.
- Subscriber modems or ONTs misconfigured due to incorrect provisioning.
- Ingress or noise in the plant detectable via proactive network monitoring tools.
- Power issues at remote cabinets that could be diagnosed with telemetry from smart power supplies.

Manual troubleshooting also introduces variability and risk. Human diagnosis can be inconsistent, and repetitive tasks—like reading and interpreting upstream SNR values or re-provisioning a modem—are prone to error or omission. Field work conducted without a clear understanding of root cause can lead to "no-fault found" outcomes, repeat visits, or temporary fixes that don't address the underlying issue.

Reactive troubleshooting delays resolution, erodes customer trust, and undermines SLA compliance, resulting in financial penalties and reputational damage —especially in business-grade or high-availability service tiers. This reactive model is fundamentally incompatible with the scale and expectations of modern broadband networks. The modern, and customer focused approach, is software assisting with processing and understanding telemetry, and human developers reacting with code that can either mitigate the problem remotely or dispatch a technician with clear instructions.

2.4. Synthesis: A Converging Problem Space

The shortage of skilled labor, combined with increasing network complexity and unsustainable reliance on manual operations, presents a clear imperative: service providers must evolve beyond traditional operating models. Incremental hiring alone will not solve the scalability gap, especially as competition for skilled workers grows and networks continue to outpace staffing capacity.

Instead, operators must embrace operational transformation through automation, data analytics, and process reengineering. By adopting principles of DevNetOps and investing in intelligent platforms that reduce human bottlenecks, network teams can maintain high service quality while making the most of limited workforce capacity. In doing so, they can shift from reactive maintenance to proactive assurance, elevate the role of human expertise, and future-proof operations for the next generation of connectivity, for building a sustainable, adaptable, and resilient network infrastructure capable of meeting the evolving demands of the digital age

3. Automation and AI-Driven Solutions in Network Operations

As network operators face growing pressure to scale operations with constrained human resources, automation and AI-driven technologies are emerging as critical enablers of operational efficiency, agility, and service quality. These tools not only reduce the need for repetitive manual tasks but also enhance situational awareness, support faster decision-making, and enable proactive maintenance. When integrated into a DevNetOps framework—where development, network operations, and automation practices converge—these technologies form the backbone of a modern, software-defined operational model.

3.1. Automation-Based Diagnostics

Traditional diagnostics in network operations rely heavily on rule-based monitoring and reactive ticketing systems. While functional, these systems are limited by static thresholds, fragmented data sources, and their inability to detect subtle degradations before they escalate into customer-affecting incidents.

Automation based diagnostics, particularly those using machine learning (ML) models trained on historical telemetry, offer a significant upgrade. These systems can continuously analyze large volumes of network data (e.g., RF metrics, optical power levels, error logs, device telemetry), correlate data points and identify patterns that precede faults. By recognizing early indicators—such as gradual SNR degradation or intermittent packet loss, automation can trigger alerts well before service degradation is noticeable to the end user.

For example, predictive models can:

- Forecast modem or ONT failures based on behavioral deviations.
- Identify deteriorating drop cables by correlating temperature, attenuation, and return path noise.
- Detect power supply anomalies in nodes and cabinets prior to full failure.

These insights support **predictive maintenance**, allowing operators to address issues during planned windows or bundle repairs with other maintenance tasks—minimizing disruption and cost. AI-powered platforms such as CableLabs' Proactive Network Maintenance (PNM) suite and vendor solutions like CommScope's ServAssure or NetOps AI modules from Cisco are already demonstrating the practical viability of such systems in the field.

3.2. Remote Troubleshooting and Self-Healing Networks

Remote troubleshooting, enabled by telemetry-rich infrastructure and programmable interfaces, allows network operators to resolve many issues without dispatching technicians. Combined with automation logic and AI decision trees, this functionality enables **self-healing** capabilities in modern networks.

Key examples include:

- **Dynamic Configuration Remediation**: Devices that fail provisioning due to incorrect configuration files or MAC mismatches can be automatically identified and corrected by centralized systems.
- **Remote Reboot and Factory Reset**: Automated systems can perform controlled reboots, resets, or firmware updates in response to anomalies—either autonomously or under operator review.
- **RF and Optical Path Diagnostics**: When impairments are detected, systems can trigger diagnostics across the return path or fiber segment to isolate the fault domain and recommend specific field actions, reducing field time.

Self-healing logic, enabled by intent-based networking or rule-based automation platforms, can also dynamically reroute traffic, disable noisy channels, or throttle problematic devices. These features dramatically reduce the frequency and duration of service disruptions, particularly in access networks where last-mile failures are common.

3.3. Automated Provisioning and Fault Resolution

Provisioning and fault resolution are core operational activities that benefit significantly from automation. In traditional workflows, service activation and troubleshooting involve multiple steps: validating inventory, generating config files, applying service flows, and confirming connectivity. Errors in any step—often human-induced—can cause delays, repeat calls, or service fallout.

Automation platforms can streamline these processes through:

- **Zero-Touch Provisioning (ZTP)**: Devices request and receive configuration automatically upon installation, using pre-registration or at time of installation mobile triggered backend workflows. This reduces errors and allows equipment installers to self-service and complete work orders without engaging operations staff.

- **Automated Fault Resolution**: When a known fault condition is detected—e.g., a modem reporting "partial service"—automation can initiate a sequence of corrective actions: configuration refresh, firmware update, RF re-tune, or customer communication. These playbooks are triggered via APIs and can resolve many issues without intervention.

These systems not only reduce time-to-resolution but also enable **closed-loop assurance**—where the network verifies its own performance post-action, adjusting further if required. Integration with orchestration tools like Ansible, Netconf/YANG-based controllers, and DOCSIS provisioning backends makes this possible even in heterogeneous environments.

4. The Role of DevNetOps in Enhancing Network Operations

4.1. Defining DevNetOps and its Relevance

As networks grow increasingly dynamic, programmable, and software-defined, there is a critical need to integrate automation into network workflows with the same agility and rigor as modern software engineering. DevNetOps emerges in response to this need—a cultural and technical movement that blends Development (Dev), Network Engineering (Net), and Operations (Ops) into a unified operational model.

Rooted in the principles of DevOps, DevNetOps applies practices such as Agile development, CI/CD pipelines, version control, automated testing, and metrics-driven feedback loops to network infrastructure. It treats network configurations, policies, and provisioning scripts as code—commonly referred to as Infrastructure as Code (IaC)—allowing for scalable, modular, and testable automation that can be deployed rapidly and reliably.

In traditional IT workflows, the network has long been a bottleneck—a separate domain requiring manual provisioning and delayed handoffs. DevNetOps eliminates this by exposing network-as-code interfaces, enabling networks to be orchestrated alongside compute and storage in a single deployment pipeline. This shift accelerates change cycles, reduces human error, and transforms networking from an operational laggard to a driver of innovation.

Technically, DevNetOps encompasses:

- Use of tools like Ansible, Terraform, Python, Git, and CI/CD platforms for provisioning and version-controlled changes.

- Adoption of GitOps, Agile, and Scrum to enable collaborative, iterative workflows.
- Integration of model-driven APIs (e.g., NETCONF/YANG, gNMI, RESTCONF, gRPC) and telemetry pipelines for observability and intent-based networking.
- Utilization of finite state machines, regex-based parsers, or screen-scraping techniques to automate legacy systems.
- Deployment of analytics engines and proactive monitoring via automated telemetry.

Yet the greatest challenge isn't technical—it's cultural. For DevNetOps to succeed, organizations must break down long-standing silos between developers, network engineers, operations teams, and planners. DevNetOps demands a shift to shared ownership, fast feedback loops, and cross-functional collaboration, where automation and operational resilience are everyone's responsibility.

This cultural transformation is especially timely for telecoms and large network operators, who are facing workforce shortages and increasingly complex hybrid environments. DevNetOps empowers traditional network engineers to evolve into roles like Network Automation Engineer, Site Reliability Engineer (SRE), or Network DevOps Specialist, combining domain expertise with software proficiency. These hybrid roles not only drive innovation and agility but also ensure long-term career relevance in a landscape where networks are managed through code and API's.

By embracing DevNetOps, organizations can modernize their operational practices, reduce toil, and seamlessly integrate networks into the same automation pipelines that have transformed software delivery. In doing so, they not only accelerate change but also empower their teams to build more reliable, secure, and adaptable networks—networks that evolve as quickly as the demands placed on them, creating a future-proof, competitive advantage.

4.2. Key Pillars of DevNetOps

The first, most important, and hardest part of embracing automation and DevNetOps culture is the cultural change. Leaders must embrace, encourage, reward, and challenge employees to make the essential cultural changes to embracing DevOps culture: tearing down siloes, blameless incident response and postmortem, short circuited and fast feedback loops, metrics-based improvement, automated configuration, infrastructure as code, shared responsibility & ownership, and continuous learning & experimentation.

Break down siloes by encouraging cross-functional collaboration between teams of engineers, developers, operations, security, product, and business teams. Slicing work groups for workstreams, initiatives, and projects across teams will provide multiple benefits. Cross functional teams will short circuit and speed up feedback, improving delivery times and project cohesion. It will build relationships between teams, corrode old animosities and competitions, remove the fog of war, and encourage management to also build relationships and collaborate. Team managers will have to support their team members, as well as build relationships with other managers and help support those managers direct reports.

Increased communication without delays at team borders will short-circuit old communication pathways, speeding project delivery by identifying and resolving problems much sooner. Instead of "tossing the football" between teams at defined boundaries, this *shared responsibility & ownership* of cross-functional teams can own the entire lifecycle of applications, systems, and networks providing fast feedback and

solutions as resources are available and engaged at every project phase from concept to delivery, and maintenance. Maintaining the cross-functional team after delivery and into operations and maintenance maintains group and tribal knowledge and speeds the mean time to resolution. Expertise can be brought in to improve reliability, an efficient system can free the cross-functional teams to extend to new products, systems, and improvements.

Shifting from blame and post-mortem to learning-focused retrospectives allows employees to quickly self-organize and identify improvements in how they work, and how products and systems work. Cross-functional collaboration enhances and brings systems-based thinking, instead of component or role-based thinking. The reward for these new behavior patterns is an improvement in network performance and network functionality delivery, as a whole.

Automated configuration management, and infrastructure as code (IaC) ensure consistency, scalability, and repeatability in deployments. With higher deployment and change accuracy, higher velocity can be found, with less manual toil.

An important key pillar to upskilling and role evolution is *continuous learning & experimentation*. Employees across cross-functional teams must be encouraged and rewarded for continuous learning, and time must be provided for learning and experimentation. Without putting new practices or technology through their paces, employees cannot be assured to be gaining required knowledge through both experience and occasionally even struggle. By experimenting, learning and overcoming challenges, employees gain hard fought knowledge and experience.

Below is the summary of the key pillars of DevNetOps

1. Cross-Functional Collaboration & Shared Ownership

- Encourage cross-team collaboration and full-lifecycle responsibility across Development, Operations, Security, QA, and Product.
- Break down silos to promote unified delivery.
- Foster shared accountability from build to run.

2. Blameless Culture & Psychological Safety

- Build trust through transparency and learning, not blame.
- Conduct learning-focused postmortems.
- Normalize open incident response and knowledge sharing.

3. Fast Feedback Loops & Iterative Delivery

- Accelerate delivery and learning through rapid iteration.
- Shorten cycle times with CI/CD pipelines and automation.
- Embed continuous feedback across code, users, and systems.

4. Resilience & Data-Driven Improvement

- Design systems to withstand failure and evolve through insight.
- Build for reliability with observability and self-healing patterns.
- Use metrics to guide continuous improvement.

5. Automation & Infrastructure as Code

- Ensure repeatability, consistency, and speed at scale.
- Automate configuration, testing, and deployment workflows.
- Use IaC to reduce manual effort and drift.

6. Continuous Learning & Experimentation

- Empower teams to grow and adapt through knowledge and autonomy.
- Provide time and tools for upskilling and experimentation.
- Encourage self-organization and evolving roles.

4.3. DevNetOps Practices in Network Automation

Automation of network device configuration is one of the most obvious use cases in DevNetOps for using code and automation practices. The goal of network automation is clear: faster, more reliable deployments, fewer human errors, and rapid adaptation to evolving conditions.

Defining networks through infrastructure as code allows for development and solutions teams to self-serve and deploy or modify networks as their work and solutions require. Additionally, it allows for high change accuracy and response, as solutions can be pre-tested and put on rails.

Defining and deploying configurations through code allows for rapid testing, avoidance of errors, building for failure tolerance, automated pre-and-post change validation, and design and as-built auto-documentation.

Automation allows personnel formerly occupied with network changes to focus on strategic initiatives, enhancing innovation and efficiency in areas previously limited by personnel bandwidth.

With time affluence created by network automation, additional functions can be added to the network through network function virtualization, software defined networks, or network analyzing and adapting applications.

Metrics allow DevNetOps teams to proactively detect and respond to performance degradation—automating fixes through code while maintaining a standardized, homogenous network configuration. Rather than relying on bespoke configurations, teams can define consistent baselines and apply business logic to adapt where needed—enabling real-time reporting and dynamic responses to changing conditions.

When changes are deployed as code, the system can also monitor itself. If metrics or health checks detect an issue, automated rollbacks or corrective adaptations can trigger in an instant, preserving or even enhancing network stability without manual intervention.

Additionally, with a network adapting application deployed, solutions can be built to be failure proof. Configuration management encounters a fault during deployment? Automatically restart and resume. Network is built to be redundant? Have app-powered circuit breakers interrupt and test the network by introducing faults or chaos, and the network is resilient.

For configuration management, this failure tolerance can automatically resume configuration work as the system restores itself, ensuring change work completes, or backs it out and pages an operator if required.

For network resiliency, constant random testing of failure scenarios brings the confidence that the network is ready to survive when a failure occurs. This resiliency is proven because as automated agents are introducing failures, the network adapts, and engineers can demonstrate the network is provably fault tolerant. Instead of worrying about the "big outage" and the mean time to repair, introducing constant chaos could ensure peace of mind for operators, customers, and on-call employees.

4.4. Case Study – We're Full of BEANS Now

Since 2007, configuration complexity for Cable Modem Termination Systems (CMTS) and Converged Cable Access Platforms (CCAP) has skyrocketed. That year saw the debut of DOCSIS 3.0 and its channel-bonding feature, which immediately multiplied per-serving-group configuration lines by roughly 4×—and, over time, by as much as 32–48×. Eight years later, DOCSIS 3.1 introduced OFDM and OFDMA channels, each bringing dozens of additional tunable parameters. At the same time, the proliferation of services—VoIP, video streaming, business services over DOCSIS (BSoD), and more—layered on further Quality of Service, routing, and MPLS configurations. Altogether, these drivers have ballooned CMTS/CCAP configurations into tens or hundreds of thousands of CLI configuration lines—each one requiring absolute accuracy to ensure reliable customer service and underscore an urgent need for automation.

Over a decade of converging services into CCAP streamlined headend hardware but dramatically increased per-node configuration complexity. The subsequent shift to a disaggregated broadband edge via Distributed Access Architecture (DAA) multiplied provisioning tasks—pushing setup down to countless remote PHY devices (RPDs) and introducing a new Converged Interconnect Network (CIN).

To tackle this explosion of manual work and meet growing demands for real-time analytics, the Access Network Automation team built "BEANS" (Broadband Equipment Automation Networking System): an in-house, Python-driven platform powered by Jupyter Notebooks and a suite of microservices.

BEANS automates:

- CLI parsing
- Configuration generation and application
- SSH session management and multiplexing
- Configuration auditing

- Network data-science workflows

These capabilities transform once arduous tasks into one-click operations, autonomous change execution, and zero-touch provisioning. BEANS can configure or analyze hundreds of devices in parallel and multiplex connections from multiple automation sources—microservices, runbooks, and APIs—delivering speed, scale, and safety.

In 2016, BEANS began as a Python project, an off-the-desk initiative by a network engineer. It leveraged Jupyter notebooks and live IPython (Interactive Python) kernels to rapidly parse device CLI commands for automating information retrieval and configuration auditing. IPython's interactive interpreter allowed out-of-order script execution using cells, enabling fast testing and iterative parsing development without rerunning entire scripts or repeatedly connecting to test devices. As the utility of the tool became evident, fellow employees began requesting audits and data collection through BEANS as well. To streamline usage and reduce the author's time commitment, an email-based trigger system was introduced. Rather than designing a new user interface, workers could simply use the familiar medium of email: by sending a preconfigured subject line and required input in the message body, they would receive a reply with either the processed information or a generated configuration script.

BEANS, originally a Windows-based prototype running on a laptop, was expanded by two employees into a robust platform for rapid and accessible automation. Rather than centralize all automation efforts through a dedicated team, they championed a cultural shift—advocating for a democratized approach where automation was prioritized by everyone across the organization. This bottom-up strategy encouraged broader participation and integration of automation into daily workflows. Building on its success, BEANS was migrated to a dedicated Windows server to enhance reliability and enable seamless integration with Microsoft Outlook. Running continuously on Jupyter Notebooks and powered by ZEO DB (a Python object database), the platform could monitor email traffic 24/7, process audits, generate configuration scripts, and send automated replies—all without human intervention.

Faced with hiring restrictions and the challenge of deploying Distributed Access Architecture (DAA) in late 2016, management sought innovative solutions to improve network configuration speed, reliability, and Remote PHY provisioning. Recognizing the limits of their manual change management process—where workers executed procedures overnight—two employees were chosen to pioneer a new automation approach. With early automation successes and agile training led by experienced developers, the company embraced a strategy of upskilling network professionals to become programmers, rather than teaching programmers complex DOCSIS protocols and device CLI.

The importance of this shift was underscored in 2017 when a voluntary departure program (VDP) reduced staff by approximately 25% [5], while DAA adoption accelerated. Remaining employees were encouraged to think digital-first, embrace agile principles, and grow their technical skillsets. The creators of BEANS embraced this transition, building a network automation platform using Python-based automation runbooks and Jupyter Notebooks. The system evolved onto Linux, adopted containerization for rapid deployment and iteration, and eventually included a CMTS configuration API for orchestrating Remote PHY services—highlighting the agility of Python-driven automation.

More than an automation team, they became force multipliers—providing APIs, services, and a platform empowering others to automate. Adopting DevOps practices of tested code, continuous integration (CI),

[5] Approx 3300 Shaw employees accepted buyouts

continuous delivery (CD), metrics-based feedback loops, blameless retrospectives, Agile, and continuous learning through weekly code development and mentoring meetings with platform users.

Network teams self-organized and volunteered for runbook automation development training, writing automated runbooks and building services capable of validating change events, executing configuration changes, and triggering alerts when needed.

By 2024, this automation-first culture transformed operations: the team achieved a 99.9% success rate across 2,815 configuration changes, with over 73% completed via full automation, 19% with semi-automation, 7% manual implementation with scripting, leaving only 1% fully manual tasks—an evolution from 0% just seven years prior.

5. Addressing the Fear of Job Displacement

While automation promises significant operational gains, it often triggers a parallel concern within organizations: job security. Technicians and engineers, particularly those whose roles involve repetitive or manual tasks, may fear that automation and AI systems are intended to replace them. A 2024 survey by FlexJobs found that 34% of workers fear job displacement due to AI and automation, while nearly half view these technologies as potentially threatening to their current roles[6]. The report emphasized the importance of clear employer communication and proactive upskilling to help workers adapt to ongoing technological change. This fear, if unaddressed, can lead to resistance, morale issues, and delays in adoption. History shows that technological advancement, when managed thoughtfully, enhances human work rather than eliminating it. This section explores how to reframe the automation narrative, supported by workforce strategies and real-world case studies.

5.1. Historical Perspective: Technology as a Workforce Multiplier

Historically, every major wave of automation — from the industrial revolution to the computer age — has introduced initial fears of job displacement. In each case, some job functions were inevitably made obsolete, but new roles and industries emerged to replace them. The net effect over time was increased productivity, higher-value work, and a more specialized labor force.

For example, the introduction of automated telephone switching systems in the early 20th century eliminated the need for manual switchboard operators. However, it also created demand for telecom technicians, electronics engineers, and eventually, software developers. Similarly, network automation does not eliminate the need for network professionals — it changes the nature of their work.

The shift underway today is comparable: rather than focusing on manual CLI commands, technicians are being empowered to manage networks through APIs, scripting, telemetry analysis, and policy-driven orchestration. Automation replaces repetitive effort — not expertise.

5.2. Automation as a Complement to Human Expertise

Automation excels at performing tasks that are repetitive, time-sensitive, or data-heavy — such as polling devices for telemetry, comparing performance trends, or reconfiguring large groups of devices. These tasks disproportionately consume skilled engineers' time. By offloading this burden, automation enables personnel to focus on higher-order problems, including:

[6] Workers fear job displacement due to AI

- **Root cause analysis** that spans physical and virtual infrastructure.
- **Service design** for emerging applications with strict latency or reliability requirements.
- **Infrastructure optimization** and capacity planning based on traffic trends.
- **Building automation scripts or tools** for internal use, making their teams more agile.

In this way, automation becomes a **force multiplier** for technical talent. Rather than replacing engineers, it equips them with tools to be more effective and innovative. As a result, network teams can scale operations without scaling headcount at the same rate.

5.3. The Human Element of Automation Success

Ultimately, the success of any automation initiative is not defined solely by technical KPIs — it depends on whether the workforce embraces the tools. An empowered team that sees automation as a lever for professional growth, rather than obsolescence, will drive faster adoption, more innovation, and stronger long-term outcomes. Failing to address workforce concerns can undermine even the most technically sound initiatives.

Automation, AI, and DevNetOps are not replacements for people — they are **enablers of people**. With the right strategy, these technologies transform fear into fuel for the next generation of network operations.

6. Optimizing Technical Talent Through Automation

As the tech and telecom sectors navigate an evolving workforce landscape, a broader shift is underway—not merely driven by cost-cutting, but by a transformation in how work gets done. Organizations are transitioning from “people-powered” repetition to intelligent automation, leveraging code, AI copilots, and machine learning to reshape roles and workflows.

Automation isn’t about eliminating jobs; it's about elevating the value of human talent. By relieving teams of routine tasks and procedural toil, automation unlocks new capacity for innovation, problem solving, and customer experience improvement. What once lingered on the wish list of “someday projects” now becomes feasible, thanks to the time affluence created by machine-accelerated workflows.

6.1. Shifting Focus to High-Value Tasks

The true opportunity of automation lies in redirecting human effort toward what matters most. Instead of spending valuable time manually following playbooks and device procedures, technicians and engineers can now devote their cognitive capacity to optimization, innovation, and reliability improvements. By integrating automation into the heart of operations, organizations free up bandwidth to tackle initiatives that were previously deferred or deprioritized. This means stronger systems, faster delivery, and ultimately, greater business value—without burning out teams on low-leverage tasks.

6.2. Upskilling and Role Evolution

With repetitive labor reduced, employees can engage more meaningfully adding greater value to the organization and finding renewed satisfaction in their roles. Upskilling becomes a natural outcome of this shift, as workers adopt new tools, embrace new challenges, and evolve into new responsibilities.

Forward-thinking organizations enable this transformation by:

Offering self-paced learning and experimentation time

- Supporting cultural evolution toward automation and DevNetOps practices
- Actively challenging workers to grow, rather than passively inviting them

Rather than forcing change, the most effective strategy begins with early adopters—those excited by innovation. These champions create a ripple effect through peer influence, showcasing automation's benefits organically. A "pull, not push" approach builds intrinsic motivation, sparking a cultural shift grounded in curiosity, then demand, not coercion.

Hands-on experimentation, even in non-production environments, builds confidence. As teams upskill and collaborate across siloed domains, traditional roles begin to blend, self-organize, and realign around outcomes.

Embracing this organic reorganization is essential for lasting change. Conway's Law reminds us: "Organizations which design systems are constrained to produce designs which are copies of the communication structures of these organizations." [7] To build more dynamic, resilient systems, we must reshape how people collaborate—starting with how they learn, experiment, and grow.

Self-organization is a powerful catalyst for role evolution. By giving employees the autonomy to reshape internal communication patterns and team structures, organizations unlock the insights of those closest to the work. Employees on the ground know where the friction points lie, which processes slow them down, and who among them are natural problem-solvers and emerging leaders. When empowered to realign teams and roles organically, these employees can pave the way towards more efficient, resilient, and adaptive operations.

6.3. Automation Success Case Studies

6.3.1. Case Study: "C&R" — From Side-of-Desk Scripts to Platform Powerhouse

R began his role on a network implementation team, executing Method of Procedures (MOPs) created by designers and planners. A self-driven contributor with a growth mindset, R quickly grew restless with repetitive manual tasks. Driven by his own initiative, he began learning Python and developing small automations to streamline his daily workflow. Due to their positive impact, his scripts gained recognition, and soon, colleagues were lining up to have R run his tools to validate and accelerate their work as well.

[7] Conway's law - Wikipedia

Meanwhile, C worked as a network planner, using code to build internal tools that enhanced collaboration and automated routine planning tasks. One of C's innovations—a system that reported hardware deployment activity to vendors—earned the company millions in competitive replacement credits, eliminating the need for manual reporting across multiple team members and fully eliminating hundreds of staff hours. As network complexity increased and the volume of planning and change work grew, leadership took note. Recognizing the complementary strengths of C and R, they formally brought the two together to launch a new Access Network Automation Team.

In their first two years, C and R:

- Evolved R's desk-side tooling into a prototype runbook automation system.
- Re-platformed it to Linux.
- Containerized the solution into a scalable microservices architecture (Docker, Kubernetes).
- Developed a fully functional API for automated provisioningDOCSIS Remote PHY Devices (RPDs) on Cable Modem Termination System (CMTS/CCAP) devices.

Today, C and R lead a DevOps team of eight developers delivering BEANS: a modular runbook automation and microservices platform.

BEANS features include:

- A browser-based IDE for writing and deploying automations.
- Private-cloud microservices operator that multiplexes SSH connections and device sessions.
- Template-based configuration deployment operator with automatic rollback.
- Data parsing mining operator for bulk programmatic device configuration ingestion, CLI to modeled objects.
- Built-in resiliency: containerized services automatically recover and resume execution after system or hardware failure.

What began as "side-of-desk" tinkering has become a force multiplier—accelerating operational efficiency, reducing manual toil, and transforming how network changes are executed at scale.

6.3.2. Case Study: "L" — From CLI Fatigue to Automation Developer

L began his career on a network implementation team, executing Method of Procedure (MOP) documents written by network designers and planners. Working in a DOCSIS wireline access environment, he spent nearly three years manually configuring network devices via CLI—a process that, over time, became repetitive and unfulfilling. With little left to learn and diminishing job satisfaction, L greeted early discussions of network automation with apprehension: *"Are we automating ourselves out of a job?"*

That fear turned to fascination when L was introduced to a new internal automation platform called BEANS, built on Python, Linux, and Jupyter notebooks. The potential to reduce manual toil, eliminate errors, and increase accuracy ignited his interest. He began to see how automation could accelerate

change processes while improving work-life balance—replacing night shifts and urgent change windows with more stable, family-friendly hours.

While still fulfilling his original responsibilities, L took the initiative to build a new automation service on the BEANS platform. His tool could:

- Retrieve and identify upcoming network change events,
- Perform pre-change validations,
- Execute changes automatically,
- Conduct post-change checks,
- Resolve or roll back automatically if errors were detected.

By integrating with ServiceNow and PagerDuty, the system could alert operators if human intervention was needed—or revert changes automatically to avert critical faults.

L's role evolved from repetitive change execution to building the very tools that would transform operations. He reports significantly higher job satisfaction and continued engagement through ongoing learning and innovation. Today, L is a full-time member of the Access Automation & Tools team, actively developing his automation service into a core, sustainable element of the broader platform. Due to the success of network automation, L and five other contractors were converted to full time employees increasing headcount in defiance of expectations.

6.3.3. Case Study: "V" — From Change Executor to Network Innovator

Prior to her team's shift toward automation, V worked as a full-time change implementer in a wireline access environment, primarily surrounded by contractor peers. Her responsibilities included reviewing Method of Procedure (MOP) documents for technical accuracy and executing configuration changes to network devices via CLI.

When Python-based runbooks were introduced, V immediately recognized their potential to streamline operations, reduce human error, and standardize deployment workflows. Motivated by a desire to increase change success rates, she enthusiastically adopted the new tools—contributing runbooks, refining validation steps, and driving faster, more reliable implementations.

As her familiarity with the platform grew, so did her ambitions. Rather than continuing exclusively in scripting and implementation, V organically transitioned into a cross-functional DevNetOps role. She began working closely with engineers and architects, offering prompt feedback on new systems and providing actionable technical insight into how emerging designs could be implemented more effectively.

This collaborative exposure led to a natural evolution of her role. Today, V focuses on introducing new technologies, refining architectural designs, and improving the performance and reliability of existing wireline access networks. Her experience across implementation, automation, and design gives her a uniquely generalist skill set. She contributes meaningfully to cross-functional teams by translating operational needs into architectural requirements and helping guide others through adoption.

V's journey highlights the benefits of self-organization, upskilling, and task elevation in modern network operations. By delegating repetitive change implementation to automation, she unlocked new opportunities to deliver higher-value contributions—and her progression illustrates how network automation can foster individual growth and accelerate team-wide innovation.

7. Roadmap to Adoption: Enabling Organizational and Cultural Transformation

7.1. Define the Vision and Secure Executive Sponsorship

An important part of embracing network automation and DevNetOps principles is cultural change. This transformation must be led and empowered from the top. Mandates can face resistance, and issuing directives is not enough to change the course of an organization. A challenge is better, and the first people to embrace the challenge and overcome resistance have to be the executives.

Executive leadership needs to be prepared for the organization to shake things up, to try new ideas, to fail and fail fast, moving on to new experiments. The organization will likely need to reorganize, and self-organize, and executives will need to be prepared as well as find the energy to encourage and foster this growth of new ideas and culture.

Employees buy into mission, not mandates, therefore executives need to buy in and issue a challenge, set up support structures, and find the key players who can start to initiate change. Laslo Bock, formerly Google's Chief People Officer, noted that it was more effective to nudge behavior, not enforce it. Executives should sponsor pilot programs, provide resources, and remove barriers—but let teams self-organize and innovate.

To provide the nudge realise that employees want to do great work, and they want to solve problems. Many already want to make systems and organizations better. A big initial step is to help find and identify a problem that requires a solution. As Simon Sinek says: vision often comes *after* identifying a problem worth solving. Leaders should focus on the pain points—manual or error-prone change processes, slow deployments, lack of collaboration—provide a nudge and let the vision grow from there. The most impactful message a leader can send is "How can I help?", rather than "Here's what you should do."

7.2. Start Small with Champions and Pilot Projects

Instead of mandating change universally, start small and intentionally. Experts suggest identifying employees with an automation mindset—those who naturally seek innovation and thrive on evolving ideas—and forming an exclusive pilot group. Provide these early adopters with dedicated time, training, and the permission to experiment freely. Let them self-organize, prove that automation works, and return to their teams as advocates and examples.

This strategy aligns The Law of Diffusion of Innovations[8], which shows that cultural transformation begins when about 15–18% of a population adopts a new idea. Once this threshold is crossed, momentum builds, ideas spread organically, and resistance decreases. These pioneers become trusted voices—demonstrating results with code, runbooks, and improved workflows—and soon other teams will ask where their training and tools are.

[8] *Rogers, E. M. (2003). Diffusion of Innovations (5th ed.). Free Press*

Change must be authentic, and value driven. Mandates may spark friction but demonstrated effectiveness and improved work life inspires demand. When employees witness automation eliminating toil and empowering decision-making, they begin seeking it out themselves. This is when leaders should transition into servant leadership—stepping aside, asking how they can support, and enabling their teams to shape how work gets done. Often, the frontline employees know exactly where inefficiencies lie. Instead of presenting a rigid plan, empower thought leaders to initiate grassroots transformation. Use exclusivity, word of mouth, and visible results to make the change not only inevitable—but desirable.

Automation is an opportunity to empower teams, to improve their work-life, and bring what workers and businesses both want: efficiency, cost reduction, and improved service quality. The keys to achieve this are therefore: servant leadership, employee support and empowerment, embracing and empowering cultural change, and providing the psychological and job safety for employees to change the work culture, and how they get work done, without fear of being replaced or eliminated.

7.3. Strategies to Mitigate Fear and Drive Adoption

To ensure that automation is viewed as an opportunity — not a threat — operators must invest in thoughtful change management strategies:

A. Transparent Communication

Leaders must clearly articulate the purpose of automation: not to eliminate jobs, but to enhance performance, reduce burnout, and prepare for future demands. Framing automation as a partnership between humans and machines helps dispel zero-sum narratives.

B. Reskilling and Upskilling Programs

Offer training paths that enable employees to evolve with the technology. This includes scripting languages (Python, Bash), automation tools (Ansible, Terraform), and telemetry frameworks (gNMI, SNMP, Kafka). Partnerships with vendors, universities, or internal centers of excellence can accelerate this.

C. Include Employees in the Transformation

Invite field and NOC specialists to contribute to automation projects. Their frontline insights often lead to better solutions, and their involvement reinforces a sense of ownership and relevance. Teams that build their own tools are far more likely to adopt them.

7.4. Reskill the Workforce and Redesign Roles

Once you've tackled vision and identified and empowered the thought leaders who will initiate your climb to the 18% of workforce buy-in typically required initiate an organic, viral, change in culture, the next step is to set up the support structures required for change to succeed. Keep in mind that to achieve rapid and high-performance culture, leaders will need to trust, empower, and unleash their employees.

Laslo Bock wrote in his book *Work Rules!*: "Give your people slightly more trust, freedom, and authority than you are comfortable giving them. If you're not nervous, you haven't given them enough." [9]

With vision, and freedom granted, next set up support systems and structures so employees can begin to learn, upskill, and experiment. This could be providing self-learning websites and tools or granting approvals for requested training.

Inspire employees to put automation first and then get out of the way and encourage them to build. Training and adopting DevNetOps tools, like CI/CD, Agile planning, observability, is a solid idea, but be careful to offer, encourage, and not mandate adoption of any particular tools, at least until the balls is rolling and inertia and a clearer path kick in.

Offering a visible career path in DevNetOps and providing roles or titles changes as employees self-organize and tackle problems is a great way to show up and help. This will not only inspire employees to adopt this cultural shift but reassure them that there is a future role in the organization if they hop on board and ride the wave of change.

8. Business and Operational Benefits

The business case for automation in network operations is no longer theoretical, it is well established, driven by tangible results across cost, scale, performance, and customer experience. As networks grow in complexity and customer expectations rise, traditional, manual operations models struggle to keep up. Automated diagnostics, and DevNetOps practices together unlock a set of transformative benefits that directly impact both the bottom line and strategic competitiveness. This section explores these benefits in detail.

8.1. Reduced Operational Costs

Truck rolls remain one of the most expensive and logistically challenging aspects of network operations. As stated in section 2.3., it is estimated that each truck roll can cost between $150-$300 depending on geography, complexity, and whether the issue is resolved on the first visit. Some of these visits, however, are unnecessary, often due to problems that could be diagnosed or resolved remotely.

Automation directly addresses this issue:

- Automated diagnostics can determine whether an issue lies within the home, on the drop, or in the access network, eliminating "blind" dispatches.
- Remote remediation workflows can resolve provisioning issues, reconfigure modems, or restart services without technician intervention.
- Predictive analytics can identify likely points of failure and schedule proactive maintenance before an outage occurs.

[9] *Bock, L. (2015). Work Rules!: Insights from Inside Google That Will Transform How You Live and Lead. Twelve.*

By reducing reliance on truck rolls, operators can significantly lower OPEX while improving technician productivity. Some early adopters have reported reductions of 20–30% in dispatch volume after implementing remote automation platforms according to a McKinsey and Company article from 2019 [10].

8.2. Faster Issue Resolution and Improved Service Reliability

Customer satisfaction in broadband and wireless services is tightly correlated with perceived reliability and the speed of issue resolution. Traditional operations models often involve long mean time to identify (MTTI) and mean time to repair (MTTR) due to fragmented tools, manual triage, and tiered escalation.

Automation changes this dynamic:

- **Faults are detected earlier**, sometimes before customers are even aware.
- **Triage and root cause analysis** are accelerated through AI pattern recognition and integrated telemetry.
- **Remediation actions** are executed automatically or through guided playbooks, reducing dependency on human availability.

This results in:

- Lower call volumes to support centers.
- Shorter resolution windows when issues do occur.
- Higher Net Promoter Scores (NPS) and customer retention.

Operators who've adopted self-healing and remote diagnostics systems frequently report improvements in first-contact resolution rates, leading to better customer experience outcomes with fewer internal resources.

8.3. Increased Operational Scale Without Linearly Increasing Headcount

Traditionally, growing the network — adding customers, nodes, or capacity — required a corresponding growth in human resources. However, automation breaks this linear relationship.

Through centralized control, infrastructure-as-code, and telemetry-driven feedback loops:

- One engineer can manage more devices with consistent quality.
- Upgrades and deployments are executed through scripts and pipelines rather than manual logins.
- Monitoring and alerting systems can scale across millions of endpoints without overwhelming human teams.

[10] *The case for digital reinvention*

This scalability is especially critical in the era of fiber densification, 5G rollouts, and DOCSIS node splits — all of which exponentially increase the number of elements under management. Automation allows operators to meet these scaling demands without proportionally increasing cost or complexity.

8.4. Enhanced SLA Compliance and Proactive Service Assurance

For commercial and enterprise customers, Service Level Agreements (SLAs) are binding commitments — and failing to meet them results in financial penalties and reputational damage. Manual monitoring often falls short in identifying service degradation until after thresholds are breached.

Automation supports real-time SLA tracking and proactive assurance:

- Telemetry feeds can be mapped to SLA metrics, such as latency, jitter, and packet loss.
- Thresholds can trigger automated escalation, including pre-emptive rerouting or bandwidth reallocation.
- Compliance reports can be generated automatically and distributed to clients or internal teams.

This level of visibility and control strengthens relationships with high-value customers and provides differentiation in competitive B2B markets.

8.5. Improved Agility and Time-to-Market for New Services

With customer demands evolving rapidly — from low-latency cloud gaming to smart home orchestration — the ability to deploy new services quickly is essential. Automation and DevNetOps approaches enable this agility.

Instead of months of manual design, testing, and rollout, operators can:

- Use CI/CD pipelines to develop and test configurations in sandbox environments.
- Deploy services via templated workflows, ensuring consistency and reducing human error.
- Roll back or update services with minimal downtime and risk.

As a result, service innovation cycles shrink from quarters to weeks or even days. enabling operators to respond faster to market shifts and technology trends.

8.6. Future-Proofing Operations

Finally, automation helps future-proof network operations by preparing the organization for continued digital transformation:

- The move toward cloud-native networks, microservices, and disaggregated architectures makes manual management unsustainable.
- Vendor-neutral orchestration via APIs ensures flexibility and reduces lock-in.

- Training the workforce on automation tools and frameworks creates a culture of adaptability and continuous improvement.

Organizations that invest in automation today position themselves to adopt future technologies — such as intent-based networking, AI-driven orchestration, and even closed-loop autonomous networks — with greater ease.

9. Conclusion

9.1. Summary of Key Findings

This paper has explored the critical role of automation, AI-driven diagnostics, and DevNetOps in addressing operational and workforce challenges across modern network environments. As service providers confront increasing complexity, labor shortages, and rising performance expectations, traditional models of network management — heavily reliant on manual intervention and reactive troubleshooting — are proving inadequate.

The research highlights several key findings:

- Automation reduces operational strain by eliminating repetitive tasks, enabling remote resolution of common issues, and significantly decreasing the volume of truck rolls.
- Automated diagnostics and predictive maintenance improve service reliability and fault resolution by identifying issues before they impact customers, enhancing mean time to repair (MTTR) and first-contact resolution.
- DevNetOps practices bring a structured, agile, and programmable approach to network operations, allowing for scalable configuration management, rapid service deployment, and continuous optimization.
- The fear of automation-driven job loss is a legitimate concern, but with proactive communication and investment in reskilling, automation can be a catalyst for workforce empowerment and technical upskilling.
- Business benefits are tangible and measurable: reduced OPEX, improved SLA compliance, faster service innovation, and enhanced customer experience.

Together, these findings make a compelling case for network operators to transition toward intelligent, automated, and software-driven operations models.

9.2. Future Outlook

Looking ahead, the convergence of AI, automation, and software-defined networking will become even more integral to network operations. Emerging technologies such as intent-based networking, autonomous service assurance, and real-time telemetry analytics are poised to redefine how networks are monitored, managed, and optimized. Operators will increasingly rely on AI not only to detect and remediate issues, but to anticipate network needs, optimize resource allocation, and enable self-configuring infrastructure.

The DevNetOps movement will continue to mature, with more network engineers adopting infrastructure-as-code principles and participating in collaborative, version-controlled workflows. This evolution will reshape the skill profiles of operational teams — making fluency in scripting, APIs, and data interpretation as essential as traditional networking knowledge.

Importantly, the shift to automation must be paired with a shift in mindset. Operators will need to build cultures that embrace continuous improvement, data-driven decision-making, and human-machine collaboration. Workforce transformation — through training, certification, and new role creation — will be as vital to success as any technology deployment.

In the long term, service providers that embrace this transformation holistically will not only reduce operational risk and cost — they will differentiate themselves through service quality, agility, and innovation. Automation is no longer optional; it is foundational to building a network operation that is resilient, intelligent, and prepared for the next decade of connectivity demands.

10. Abbreviations and Definitions

10.1. Abbreviations

AI	Artificial Intelligence – Simulation of human intelligence in machines capable of learning and problem-solving.
BEANS	An internal automation platform built using Python and Jupyter Notebooks. Originally Beans. Acronym: Broadband Equipment Automated Networking System.
CLI	Command-Line Interface – A text-based interface used to interact with network devices.
CMTS	Cable Modem Termination System – A device that connects cable modem subscribers to a broadband network.
CCAP	Converged Cable Access Platform – Integrates CMTS and video edge functions.
DAA	Distributed Access Architecture – Moves key network functions from the headend to the edge.
DevNetOps	A cultural and technical practice that blends development, networking, and operations to automate and streamline network operations.
IaC	Infrastructure as Code – The practice of managing and provisioning infrastructure through code instead of manual processes.
MTTR	Mean Time to Repair – Average time required to repair a failed component or system.
MOP	Method of Procedure – A detailed set of steps for implementing network changes.
NOC	Network Operations Center – A centralized location where technicians monitor and manage network performance.

OPEX	Operational Expenditure – The ongoing cost of running a business, including labor, maintenance, and utilities.
PNM	Proactive Network Maintenance – A set of tools and processes for identifying and resolving impairments before they impact service.
RPD	Remote PHY Device – A key component in DAA that moves PHY layer functionality to the access edge.
SDN	Software-Defined Networking – An approach to networking that uses software-based controllers or APIs to manage hardware.
SLA	Service Level Agreement – A contractual commitment to a certain level of service availability and performance.
ZTP	Zero-Touch Provisioning. Automated setup of network devices without manual configuration.

10.2. Definitions

Tool/Platform	**Purpose**	**Notes**
BEANS	Internal network automation platform	Developed in-house to automate CMTS/CCAP config, CLI parsing, and zero-touch provisioning
Ansible	Configuration management	Agentless automation tool widely used in network scripting
Terraform	Infrastructure provisioning	Declarative, version-controlled infrastructure deployment
Python	General-purpose scripting language	Common for automation, data parsing, and orchestration
Git / GitLab / GitHub	Version control	Used to track, test, and deploy infrastructure-as-code
Jupyter Notebooks	Interactive coding environment	Used by BEANS for rapid script development and execution
ServiceNow	ITSM and workflow automation	Integrated for ticketing, change management, and automation triggers
PagerDuty	Incident response platform	Enables alerting and on-call escalation tied to automation systems

NETCONF/YANG, gNMI, RESTCONF	Network telemetry and API interfaces	Used for model-driven programmatic device management
CI/CD Pipelines (e.g., Jenkins, GitLab CI)	Continuous Integration / Continuous Deployment	Enables rapid, tested configuration delivery

11. Bibliography and References

1. Tech talent gap | Deloitte Insights - https://www.deloitte.com/us/en/insights/industry/technology/tech-talent-gap-and-skills-shortage-make-recruitment-difficult.html

2. Urgent Need to Recruit and Train Nearly 180,000 Workers to Complete Federal- and State-Funded Broadband Networks - Fiber Broadband Association - https://fiberbroadband.org/2024/07/29/urgent-need-to-recruit-and-train-nearly-180000-workers-to-complete-federal-and-state-funded-broadband-networks/

3. What is a truck roll and how to reduce them? | CareAR - https://carear.com/blog/what-is-a-truck-roll-how-to-reduce-them/

4. AI is cables secret weapon for 10g reliability - https://broadbandlibrary.com/ai-cables-secret-weapon-for-10g-reliability/

5. Approx 3300 Shaw employees accepted buyouts - https://mobilesyrup.com/2018/02/15/approximately-3300-shaw-employees-accepted-buyouts/

6. Workers fear job displacement due to AI - https://www.flexjobs.com/blog/post/flexjobs-report-workers-fear-job-displacement-due-to-ai

7. Conway's law – Wikipedia - https://en.wikipedia.org/wiki/Conway%27s_law

8. Rogers, E. M. (2003). Diffusion of Innovations (5th ed.). Free Press

9. Bock, L. (2015). Work Rules!: Insights from Inside Google That Will Transform How You Live and Lead. Twelve.

10. The case for digital reinvention - https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-case-for-digital-reinvention

Click here to navigate back to Table of Contents

Novel Non-3GPP Access Network Sharing Architectures

Innovative Approaches of Network Sharing for Seamless Network Convergence

A Technical Paper prepared for SCTE by

Omkar Dharmadhikari, Manager & Principal Architect, CableLabs
o.dharmadhikari@cablelabs.com

Rahil Gandotra, Lead Architect, CableLabs
r.gandotra@cablelabs.com

Table of Contents

List of Figures

List of Tables

1. Introduction

In today's highly interconnected, competitive and rapidly evolving landscape of wireless communications, concepts such as network sharing, particularly Multi-Operator Core Network (MOCN) and Indirect Network Sharing (INS) are increasingly vital for multiple system operators (MSOs) that are evolving into hybrid or converged service providers. These operators deliver a broad suite of integrated services including video, internet, voice, and IoT over a diverse array of infrastructures such as cable, fiber, satellite, and wireless networks. Network sharing enables these providers to maximize infrastructure utilization, reduce capital and operational expenditures, and accelerate service deployment, all while maintaining service quality and customer satisfaction. Spectrum sharing further amplifies these benefits by enabling operators to dynamically allocate spectrum resources across shared and dedicated bands, improving spectrum efficiency, capacity, and flexibility needed to support diverse applications. Moreover, innovative network sharing architectures tailored for hybrid and multi-infrastructure environments can facilitate more mobile virtual network operators (MVNOs) like relationships between operators, fostering a more modular, flexible, and collaborative ecosystem. Such relationships enable operators to offer tailored services, extend coverage swiftly, and respond more effectively to evolving market demands, ultimately leading to a more agile, cost-effective, and user-centric communication infrastructure.

This paper examines the evolution of network sharing frameworks, explores their application beyond traditional models, and discusses how these novel architectures can unlock new opportunities across the converged, multi-technology landscape. This paper explores "Novel Non-3GPP Access Network Sharing Architectures," aiming to extend traditional sharing paradigms beyond the confines of 3GPP-defined frameworks. Section 2 highlights the significance of network sharing in enabling operators to optimize infrastructure utilization, reduce deployment costs, and accelerate service rollout, especially in the era of 5G and beyond. Section 3 traces the evolution of network sharing, starting from passive and active sharing models, through advanced concepts like MORAN, MOCN, GWCN, and INS, illustrating how these arrangements have matured to address various operational and economic needs. Section 4 delves into the 3GPP-specified architectures for 5G network sharing, with a focus on MOCN and INS, and provides details about their mechanisms and capabilities. Building upon this foundation, Section 5 introduces novel architectures that extend beyond traditional 3GPP models, incorporating supplementary access technologies such as Wi-Fi, and explores how network sharing can be innovatively re-envisioned considering these access networks. An assessment of the potential benefits, including improved coverage, spectrum efficiency, and service agility alongside the associated technical and regulatory challenges, is also described. Section 6 concludes the paper with key takeaways, emphasizing how extending the 3GPP network sharing architectures to supplementary access networks can reshape future multi-technology networks, fostering greater flexibility, inclusivity, and innovation in connectivity solutions.

2. Network Sharing and Its Significance

Network sharing is a collaborative arrangement where two or more network operators agree to share network resources or infrastructure in order to reduce the costs of building and maintaining separate networks [1].

The key benefits of network sharing for operators include:

- Reduction of CAPEX due to sharing of network resources between operators
- Reduction of OPEX needed to manage and support separate network infrastructures
- Efficient spectrum utilization by spectrum sharing

- Ability to provide optimized QoS using differentiating core services for their own subscribers via the shared network
- Helping smaller operators compete more effectively with larger incumbents

The key benefits for the end users include:

- Always being connected to the network
- Getting a seamless connectivity
- Guaranteed last mile service with increased coverage
- Lower service costs as the cost savings operators achieve through network sharing can translate into more competitive pricing for consumers

Other benefits include providing environmental remuneration by reducing the number of sites and improving the landscape.

Beyond the cost dimension, sharing of assets across multiple operators can offer economic and strategic advantages realizing complementary capabilities and synergies between the sharing entities. For example, in a situation where one entity owns spectrum but has limited infrastructure, and a second entity has limited spectrum but established infrastructure, network sharing can yield benefits to both. Also, sharing can help alleviate the difficulty in acquiring sites to install base stations, resulting from the need to densify networks to address coverage demands in indoor environments [2].

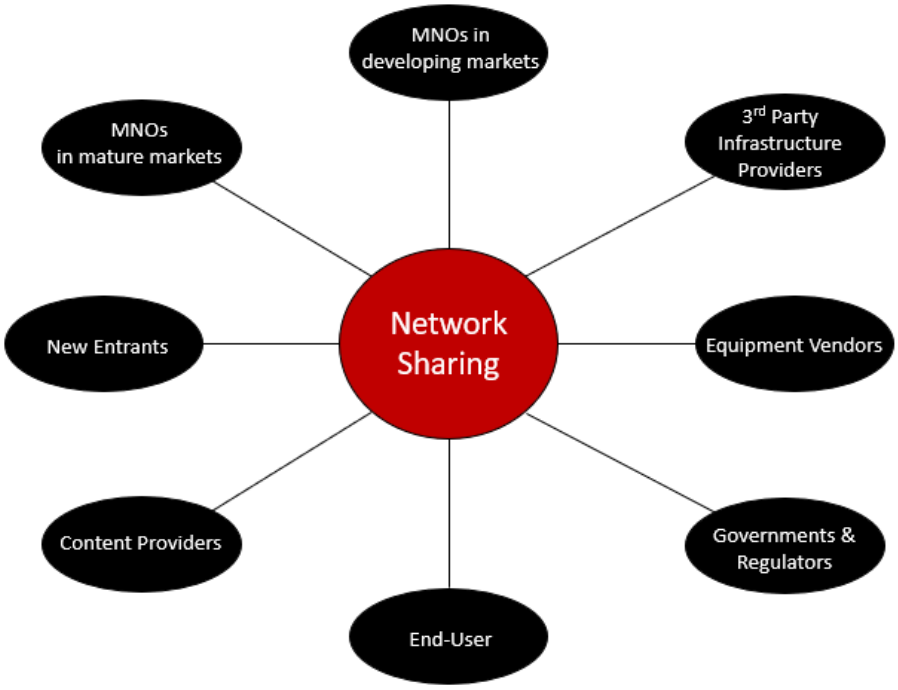


Figure 1 - Stakeholders That Could Benefit from Network Sharing

As shown in Figure 1, network sharing can benefit:

- New entrants or smaller operators (e.g.: Mobile Virtual Network Operators (MVNOs)) by providing faster entry to the markets without the need to roll out new infrastructure and without needing to purchase additional spectrum.
- Mobile Network Operators (MNOs) in mature markets to share their networks with other operators as an increased source of revenue.
- Mobile Network Operators (MNOs) in developing markets to reduce capital and operational expenditure by sharing infrastructure from the start of the build-out.
- Third-party infrastructure providers by leasing site locations, cabinets and towers to multiple operators.
- Network equipment manufacturers may have a revenue reduction, since now with multiple operators sharing the network, less equipment is needed. However, this can also be considered as an opportunity for equipment manufacturers to assist in the network planning process and offer managed network services differentiating their offerings.
- Government/regulators from efficient spectrum utilization, lowering energy consumption and reduced barrier to entry promoting competitive market.
- Consumers with increased network availability and reduced service costs.
- Content providers by reaching a wider audience through more reliable and high-performing networks, ultimately enhancing delivery, increasing consumption, and driving greater revenue.

3. Network Sharing Evolution

The major driver of network sharing continues to be the potential for cost savings [3]. However, the amount an operator can save depends upon the depth of the sharing arrangement. The potential cost savings and benefits increase as the depth of the sharing increases but so is the complexity and considerations. This brings us to the three broad categories of network sharing both sharing different network elements as shown in Table 1:

- Passive network sharing involves sharing of the passive elements of the network such as towers, rooftops, cabinets, etc., reducing costs by sharing physical infrastructure, suitable for rural or low-density deployments.
- Active network sharing involves sharing active elements of the network like the access network, core network, spectrum, etc. enabling more efficient capacity use in dense or high-demand areas but with increased complexity.
- Virtual network sharing involves sharing underlying physical resources through virtualization technologies to create isolated, customizable network slices for different operators or services leveraging NFV and SDN to create multiple logical networks over the same physical infrastructure

Active network sharing vs passive network sharing vs virtualized network sharing			
Metric	**Passive network sharing**	**Active network sharing**	**Virtualized network sharing**
Definition	Sharing of physical infrastructure like towers, sites, power supplies, and cabins without sharing active radio or core equipment	Sharing of active network elements such as base stations, radio hardware, and sometimes spectrum resources in addition to sharing of physical infrastructure	Sharing of underlying physical resources through virtualization, creating multiple logical networks or slices over the same infrastructure
Elements Shared	Infrastructure only (towers, shelters, power, antennas, site access)	Both physical (radio equipment, base stations) and possibly spectrum, core network components	Physical infrastructure, managed via software (VNFs, SDN) to create multiple logical networks
Complexity	Lower complexity; mainly site management and infrastructure maintenance	Higher complexity due to sharing active RF equipment, interference management, and coordination	High; involves software orchestration, resource slicing, and dynamic management
Ownership & Control	Separate network control, each operator maintains and operates their own core and spectrum	Usually shared or jointly managed radio resources; operators often retain core separation	Infrastructure owner controls physical layers, operators control their own virtual networks/slices
Cost Savings	Significant reduction in site acquisition, construction, and maintenance costs	Greater cost savings through sharing of active radio assets and spectrum	Potentially highest; maximizes resource utilization through virtualization & slicing
Interference Management	Minimal; interference mainly addressed at physical layer (antenna tilts, site planning)	More complex; requires advanced interference management, coordination, and potentially strict synchronization	Managed via software; flexible interference mitigation and resource allocation
Deployment Flexibility	Simplest; limited to site sharing, no change to network operation	More flexible; can support multi-operator or multi-band configurations, but requires detailed planning	Very flexible; supports rapid deployment of diverse services, slices, and tenants
Use Cases	Rural or cost-sensitive deployments, site sharing agreements	Urban deployments, high traffic areas, multi-operator RAN sharing scenarios	Multi-tenant networks, network slicing for different services or industries
Impact on Network Secrecy & Data	Operator control over core network, no shared active elements	Potential for shared spectrum and radio, requiring careful interference and security management	Strong logical separation, individual operator security policies delivered via slicing

Table 1 - Passive Network Sharing vs Active Network Sharing vs Virtualized Network Sharing

3.1. Passive Network Sharing

Passive network sharing involves sharing of the physical infrastructure, which is of great value to new entrants with site acquisition being one of the biggest challenges.

Figure 2 shows Passive network sharing that involves sharing of passive components.

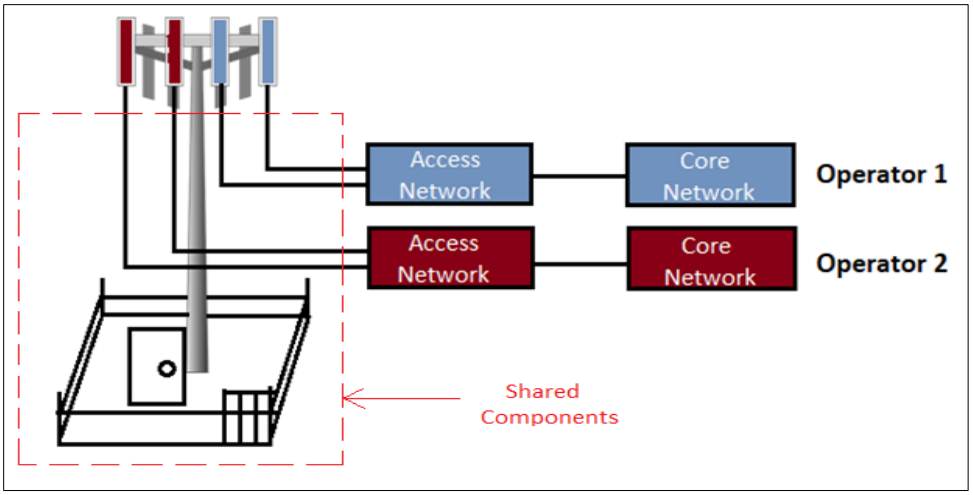


Figure 2 - Passive Network Sharing

Passive network sharing is preferred by operators, as sharing the site locations or passive components does not have a direct impact on operators' network performance [4].

The benefits of passive network sharing include:

- Avoiding duplication of passive network components
- Reducing site acquisition costs and eliminating difficulties in obtaining permits to build new sites
- Reducing upfront investments by sharing the towers, rooftops, etc.
- Minimizing the impact on environment by reducing the number of sites

Passive network sharing, although being simpler to implement than active network sharing, does have some considerations including:

- Load bearing capacity of the tower or a rooftop - which considers how many radio heads and antennas are mounted on tower, of what magnitude and how much weight.
- Height, Tilt and Azimuth considerations – to determine the coverage needed to achieve without interfering with signals coming from radio heads and antennas of other operators sharing the same tower. Operators might also consider the amount of space they might need to house additional equipment in future.

Passive network sharing is the most common form of network sharing and is supported and encouraged in most countries from a regulatory standpoint. One of the operators could own the physical infrastructure that is shared by other operators, or this type of sharing can also allow third-party infrastructure providers,

such as Crown Castle in the USA and China Tower in China, to enter the market specifically offering sites where multiple operators can install their own equipment [5].

3.2. Active network sharing

Active network sharing is when operators share network resources that have a direct impact on network performance.

Active network sharing deployments are generally not favored by operators, especially in the US market, as they have a lot of considerations and strict service level agreements (SLAs) that need to be enforced between the sharing entities.

Active network sharing is beneficial to host operators to:

- Leverage its own infrastructure and
- Share certain active network elements as a source of earning more revenue

For tenants or operators participating in network sharing, active network sharing provides:

- An opportunity to roll out their services faster with reduced time to market
- Cost savings for initial rollout, deployment and management of network
- Extended coverage by serving previously unserved or under-served areas
- Additional capacity in congested urban areas where new site acquisition is difficult and space for sites and towers is limited.

Active network sharing includes basic sharing of antennas, feeder cable and backhaul network. Or active network sharing can use one of the three deployment models:

- Multi-Operator Radio Access Network (MORAN)
- Multi-Operator core network (MOCN)
- Gateway Core Network (GWCN)

3.2.1. Multi-Operator Radio Access Network (MORAN)

MORAN allows multiple mobile network operators (MNOs) to share the same radio access network infrastructure, such as base stations, antennas, and towers, while maintaining their own core networks and spectrum allocations [6] as shown in Figure 3. In MORAN, operators share the access network but use their own dedicated spectrum, broadcasting their own individual network identifiers, or PLMN IDs. It retains logical separation of radio elements across multiple operators while enabling cost savings through physical integration and shared access network equipment. Common site-level parameters like antenna down tilt may need to be agreed upon collaboratively, but operators are free to independently control cell-level parameters.

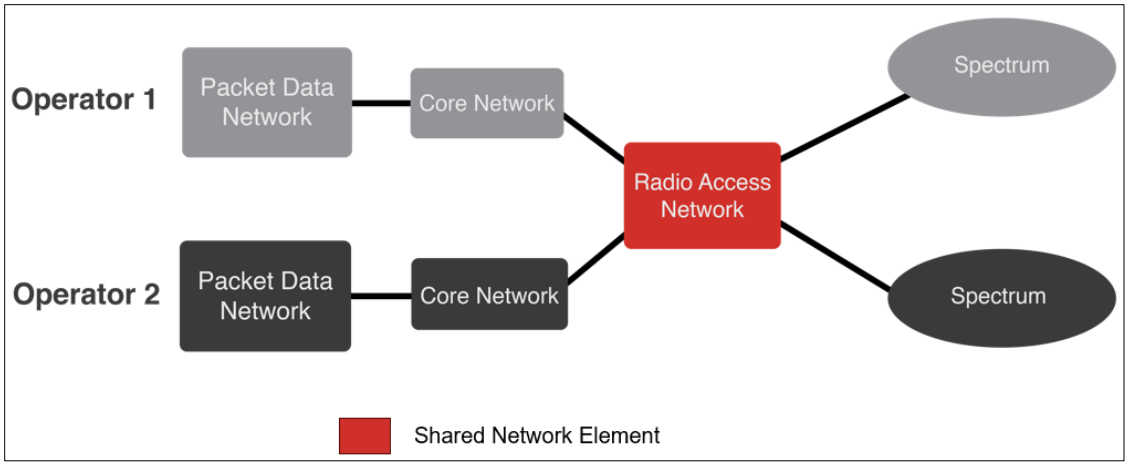


Figure 3 - Multi-Operator Radio Access Network (MORAN)

With physically independent radios and power amplifiers, each operator can operate exclusively in its own spectrum, ensuring service differentiation. Ultimately, MNOs have independent cell coverages to serve their own subscribers. This setup significantly reduces capital and operational costs by consolidating infrastructure deployment and site management, making it especially useful for expanding coverage in underserved or remote areas. It enhances operational efficiency and accelerates service rollout, but also involves challenges such as navigating regulatory compliance, establishing clear inter-operator agreements, and managing the complexities of shared maintenance and performance monitoring. Overall, MORAN offers a balanced approach to network sharing, achieving cost savings and coverage improvements while allowing operators to preserve their competitive identity.

3.2.2. Multi-Operator Core Network (MOCN)

MOCN is a network sharing arrangement where multiple operators share the same radio access network infrastructure while maintaining separate core networks [7]. Unlike MORAN, which allows operators to use their own spectrum on a shared access network, MOCN involves the sharing of both the radio access network and the spectrum, with each operator broadcasting their own individual network identifiers, or PLMN IDs, as shown in Figure 4. This setup allows for more efficient use of spectrum and radio resources, leading to cost savings and faster network deployment, especially in areas where multiple operators seek to extend coverage without duplicating infrastructure [8]. Additionally, MOCN supports coexistence of different services and technologies within the shared RAN, ensuring each operator's subscriber’s data remains separate while benefiting from the shared physical platform. The shared RAN is managed through centralized network control, which helps optimize overall performance, reduce interference, and improve capacity. Ultimately, MNOs would be under the same cell coverage.

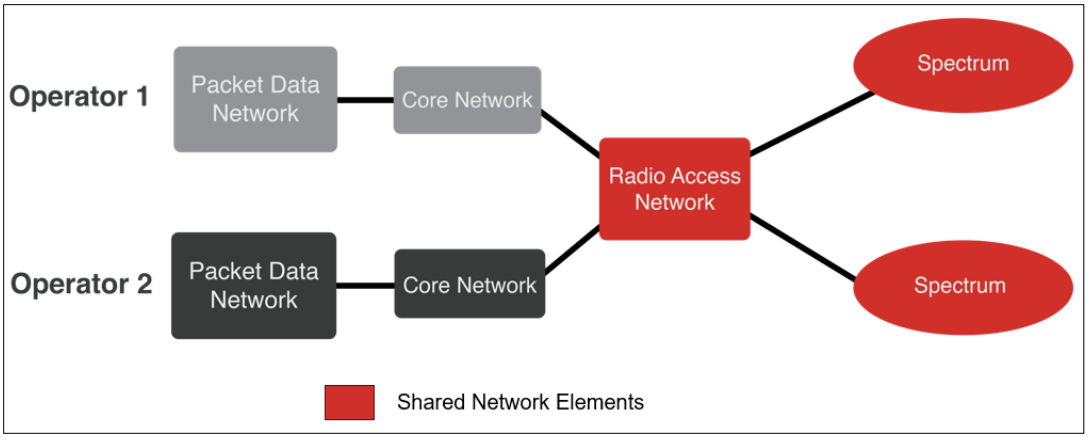


Figure 4 - Multi-Operator Core Network (MOCN)

MOCN provides a practical means to expand coverage, improve capacity, and accelerate rollout, especially in rural or cost-sensitive environments, while preserving operator independence in core network control and customer management. Its implementation demands careful coordination at the site level for parameters like antenna tilt, power, and interference management, but it offers a balanced approach to achieving operational efficiency, cost savings, and service quality in multi-operator deployments.

3.2.2.1. Shared Spectrum with MOCN

Shared spectrum can complement and enhance network sharing by introducing a dynamic and efficient model for utilizing spectrum. For example, CBRS (Citizens Broadband Radio Service) enables spectrum sharing in the 3.5 GHz band in the United States. CBRS operating under a three-tiered access framework promotes flexible spectrum use by allowing multiple entities, such as mobile network operators, enterprises, and other users, to share frequency bands dynamically. This aligns with network sharing principles where infrastructure is shared among multiple operators to optimize resource utilization. Spectrum sharing in CBRS maximizes efficiency by enabling opportunistic access to available frequencies, similar to shared RAN or core resources in network sharing which aim to optimize physical and logical network elements. The regulatory framework of CBRS provides a model for spectrum sharing that parallels the clear agreements and standards needed in network sharing, fostering innovative business models and service delivery akin to how network sharing reduces costs and promotes infrastructure efficiency. By allowing operators and new entrants to enhance coverage and capacity without significant licensing costs, The Neutral Host Network (NHN) option enabled by CBRS promotes a converged approach where shared spectrum and infrastructure drives improved service reach and quality. Thus, shared spectrum with MOCN represents a shift towards maximizing shared radio resources, embodying the broader goals of both spectrum and network sharing to foster a more innovative, cost-effective, and comprehensive connectivity ecosystem.

3.2.3. Gateway Core Network (GWCN)

GWCN is a network sharing model where multiple operators share a common core network entity along with sharing the radio infrastructure and optionally the spectrum (similar to MORAN), as shown in Figure 5. Unlike models where operators maintain entirely independent core networks, GWCN allows

each operator to retain its own dedicated core functions—such as subscriber data management, policy control, and billing—while leveraging a shared mobility management entity, session management entity, and some user plane functions. By centralizing these functions, GWCN reduces operational redundancy and infrastructure costs, offering streamlined connectivity and interoperability between networks. This model promotes efficient use of resources and infrastructure, facilitating the rapid rollout of services across geographical areas while ensuring that operators can independently manage their customer experience and regulatory requirements. Additionally, GWCN can enhance network reliability and capacity, providing a scalable solution that adjusts to various traffic demands across shared and unique spectrum allocations. It promotes operational efficiency, cost savings, and quicker deployment, while maintaining a clear separation of core functions and subscriber data for each operator, ensuring privacy and regulatory compliance.

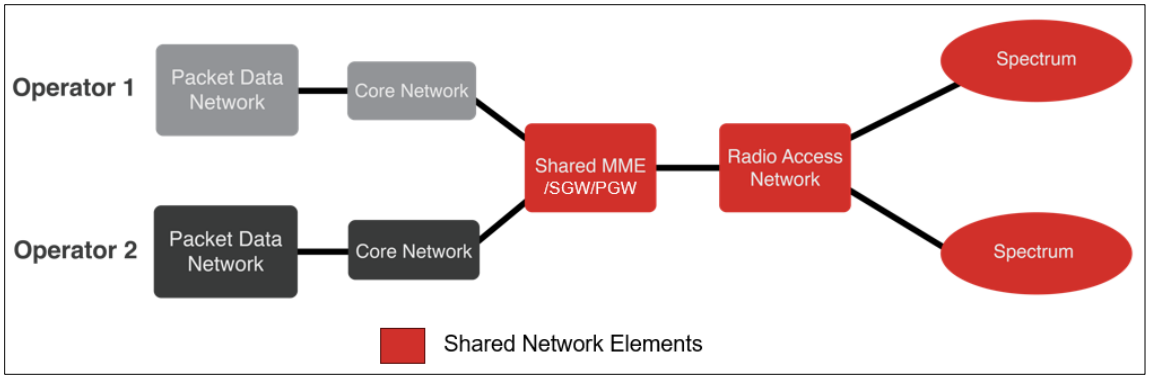


Figure 5 - Gateway Core Network (GWCN)

GWCN is particularly advantageous for smaller operators, as it allows them to implement a lightweight core network instead of investing in the development and maintenance of a fully functional standalone core. By sharing key components of the core network, smaller operators can focus resources on subscriber management, policy control, and billing functions without the financial and operational burden of maintaining an extensive core network infrastructure. This shared approach enables smaller operators to enter the market more feasibly, offering competitive services alongside larger operators by leveraging the robust, centralized capabilities of the GWCN framework. Additionally, it allows them to scale their operations gradually, expanding subscriber services and network capabilities in alignment with growing demand, without the immediate need for significant capital expenditure. This adaptability makes GWCN an attractive solution for smaller operators looking to innovate and expand their service offerings, particularly in regions or markets where comprehensive, independent core networks might be economically unviable.

3.3. Virtual Network Sharing

Virtual network sharing is an advanced approach that leverages network virtualization technologies to enable multiple operators to share the same physical infrastructure while maintaining isolated, customizable, and flexible virtual networks or slices [9]. Instead of physically sharing base stations or core network elements, virtualization abstracts underlying hardware resources—such as compute, storage, and radio assets—into multiple logical, independent network segments. Each operator can deploy its own tailored services, Quality of Service (QoS) parameters, security policies, and subscriber management within these isolated slices, all maintained over a common physical platform. This method offers

significant advantages in terms of agility, scalability, and cost-efficiency, allowing operators to rapidly deploy diverse services suited for different verticals like IoT, enterprise, or consumer applications without the need for duplicating hardware. Moreover, virtual network sharing supports dynamic resource allocation, enabling operators to optimize capacity based on real-time demand, facilitate rapid service innovation, and easily adapt to evolving market needs. As a result, virtualization-driven sharing solutions are foundational to future network architectures, promoting a more flexible, efficient, and multi-tenant ecosystem across 5G and beyond.

3.3.1. Network Slicing

Network slicing, as shown in Figure 6, enabled by virtualization and software-defined networking (SDN) technologies, allows a single physical network infrastructure to be partitioned into multiple, logically isolated virtual networks, or "slices," each tailored to specific service requirements or application profiles. This approach leverages foundational concepts such as network function virtualization (NFV), containerization, and flexible transport architectures to create distinct, customizable slices over shared hardware, ensuring each slice has dedicated logical resources like bandwidth, processing power, and security features [10]. By supporting diverse use cases such as enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC), network slicing provides the agility necessary for multi-industry and multi-service deployments. Operator strategies typically involve collocating all relevant Virtual Network Functions (VNFs) within shared or dedicated bare-metal servers, with resource sharing managed through function clustering, service level agreements (SLAs), and logical partitioning, enabling secure and efficient coexistence of multiple tenants and applications. This architecture is particularly vital for converged, multi-operator environments, as it facilitates rapid, flexible deployment of tailored network slices on a common infrastructure, thereby supporting diverse service demands and accelerating time-to-market for new applications and services [11].

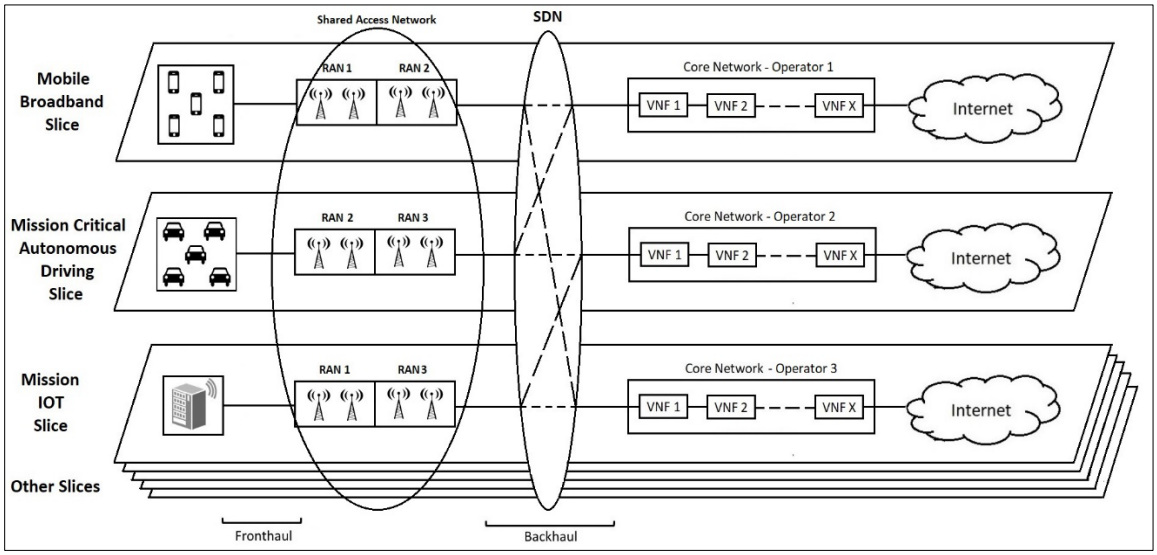


Figure 6 - Virtual Network Sharing Enabled by Network Slicing

3.3.2. Multi-access Edge Computing (MEC)

Multi-access Edge Computing (MEC) plays a pivotal role in the realm of virtualized network sharing, enabling multiple operators to deliver low-latency, high-bandwidth services by leveraging shared physical infrastructure at the network edge [12]. MEC introduces cloud-native, virtualized compute, storage, and network resources collocated close to end-users, allowing operators to deploy and manage edge applications and services dynamically across a unified platform. Figure 7 shows how the MEC system can be deployed in an integrated manner in 5G network. In a shared environment, MEC enables multiple tenants to utilize the same edge infrastructure securely and efficiently, facilitated by virtualization and containerization technologies that isolate resources and workloads. This shared MEC framework supports a wide range of use cases, including industrial automation, extended reality, and IoT, by providing dedicated, low-latency, and QoS-guaranteed environments tailored to each tenant's needs. Operators benefit from reduced operational costs, faster service deployment, and enhanced agility—enabled by cloud-native management practices like continuous integration/continuous delivery (CI/CD) and DevOps. When integrated into a converged or multi-operator network, MEC further enhances resource utilization, fosters ecosystem collaboration, and accelerates innovation by allowing different operators to deliver differentiated services without the need for separate edge deployments, thereby strengthening the overall network ecosystem's adaptability and responsiveness [13].

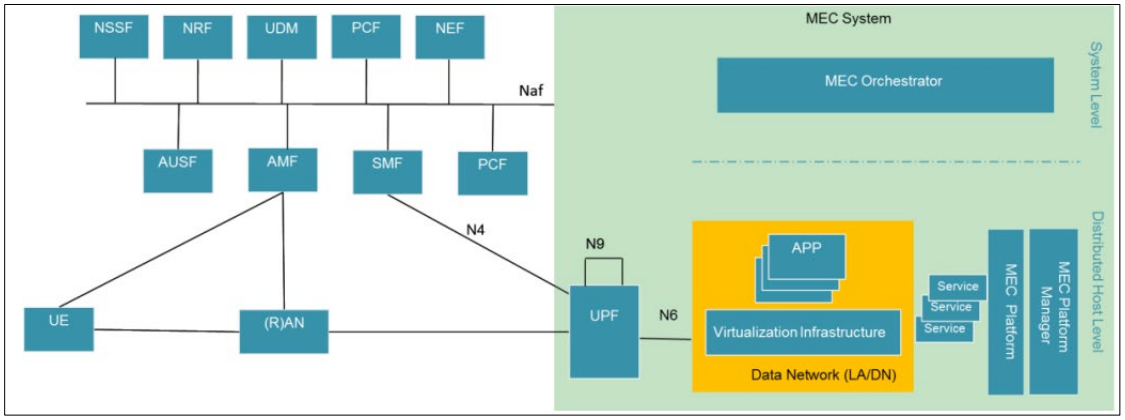


Figure 7 - Integrated MEC Deployment in 5G Network

3.3.3. Dedicated Core (DECOR) Networks (DCNs)

The 3GPP concept of Dedicated Core Networks (DCNs), introduced in Release 13, is a precursor to and a foundational element of virtual network sharing. DCNs allow operators to deploy separate core network instances tailored to specific services, user groups, or device types, operating over a shared RAN infrastructure. While DCNs provided logical separation of core network functions, they often relied on dedicated hardware resources for each core instance. In contrast, virtual network sharing, enabled by Network Function Virtualization (NFV) and Software-Defined Networking (SDN), abstracts network functions into virtualized components that can be dynamically deployed and scaled on a shared, common infrastructure. Virtualization takes the DCN concept a step further, enhancing resource efficiency, agility, and service customization. It enables operators to create highly isolated, end-to-end network slices that meet the specific requirements of diverse applications, offering greater flexibility and scalability than static DCN deployments [14]. Thus, DCNs laid the groundwork for virtual network sharing by

demonstrating the benefits of logical core separation, while subsequent advancements in NFV and SDN have transformed this concept into a dynamic, virtualized, and highly efficient reality.

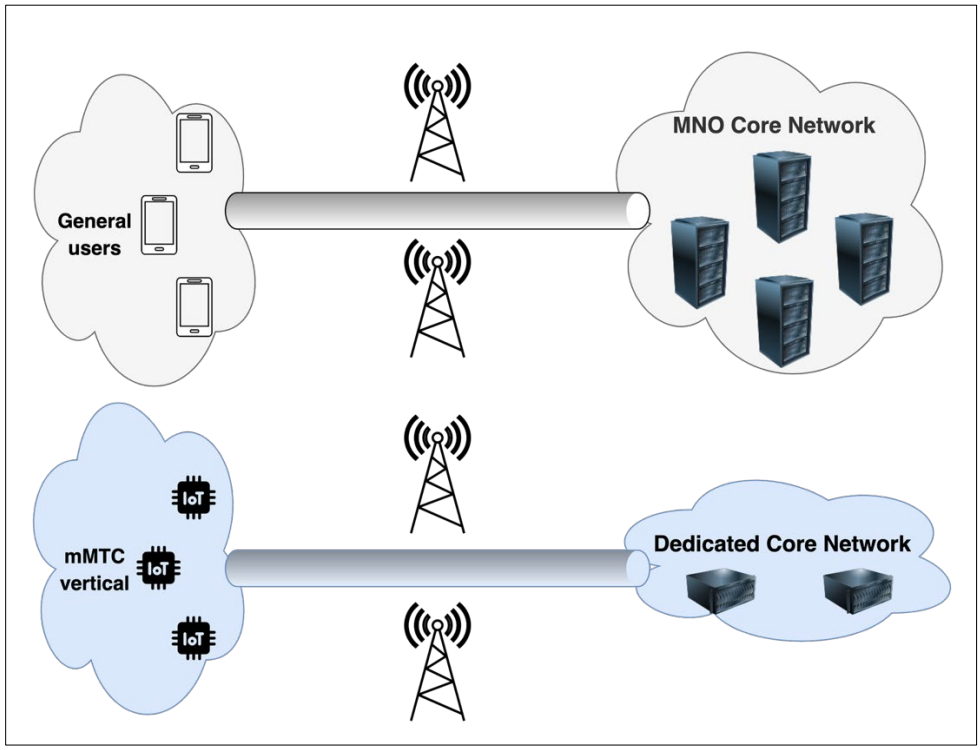


Figure 8 - Dedicated Core (DECOR) Network Representation

3.3.4. Control and User Plane Separation (CUPS) and Service Chaining

CUPS and Service Chaining are key concepts that underpin the effectiveness and flexibility of virtual network sharing [4]. CUPS, a 3GPP-defined architectural principle as shown in Figure 9, decouples the user plane (handling data traffic) from the control plane (managing signaling and control functions), allowing each to scale and evolve independently. This decoupling is crucial for network slicing, enabling operators to customize and optimize each slice according to specific service requirements, for instance, prioritizing low latency for URLLC services or high bandwidth for eMBB. Service Chaining builds upon this by enabling the dynamic composition of network functions to create customized service delivery paths. Operators can orchestrate sequences of virtualized network functions (VNFs), such as firewalls, traffic shapers, or load balancers, to meet the precise needs of each network slice, without requiring physical rewiring or reconfiguration. By leveraging CUPS and service chaining, virtual network sharing allows for highly agile, adaptable, and resource-efficient networks that support a diverse range of applications and service models across a shared infrastructure [15].

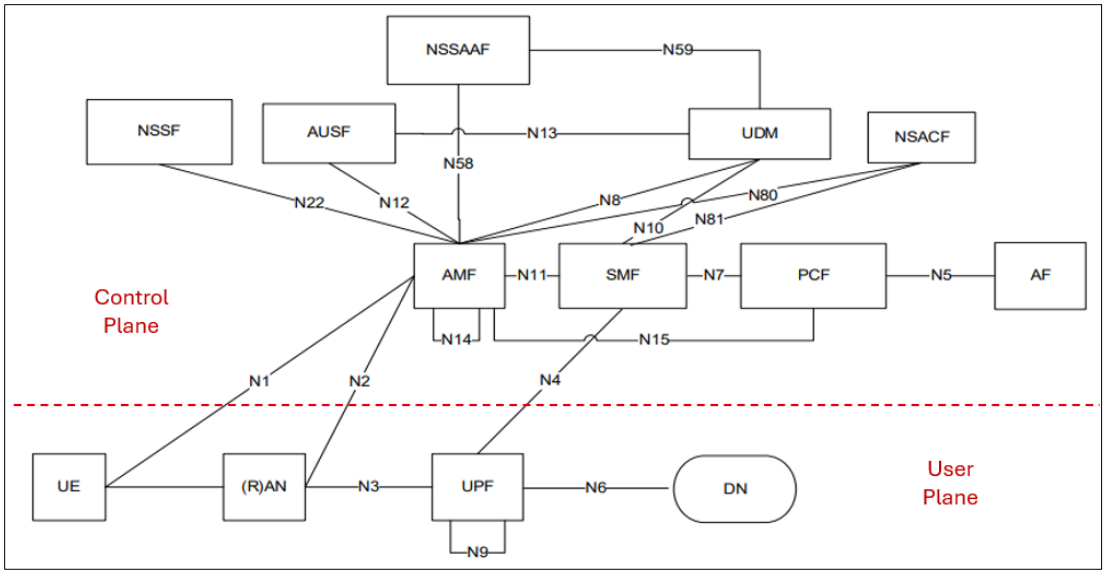


Figure 9- 5G CUPS System Architecture in Reference Point Representation

4. 3GPP Specified 5G Network Sharing Architectures

4.1. Multi-Operator Core Network (MOCN)

For smaller mobile network operators that lack licensing, spectrum and network resources, network sharing using MOCN deployment model can help them offer services with shared infrastructure, as shown in Figure 10 [16]. MOCN offers reductions in both capital expenses and operating expenses for each operator while meeting market needs in terms of extended coverage and additional capacity. MOCN also improves spectral efficiency. Both these benefits are valuable in scenarios where achieving sustainable returns on investments is difficult (e.g. regions with low population density, dense areas where a large number of small cells must be deployed, or where it might be difficult to build sites due to limited resource availability). MOCN was standardized and introduced by 3GPP in Release 8 for LTE with improvements in further releases, realizing the cost effectiveness for commercial network deployments and operations. MOCN enables core networks of multiple operators to share radio access network and spectrum. Each mobile network operator has its own core network managed independently, in a similar manner to the deployments without active sharing. The shared access network is designed to support propagation and recognition of PLMN IDs across multiple operators. The shared access network has a 1:1 mapping of PLMN to a S1 MME (4G)/N2 AMF (5G) IP address of a core network of a specific operator. This configuration allows UEs with services enabled via multiple operators to co-exist with radio services provided by a single shared access network.

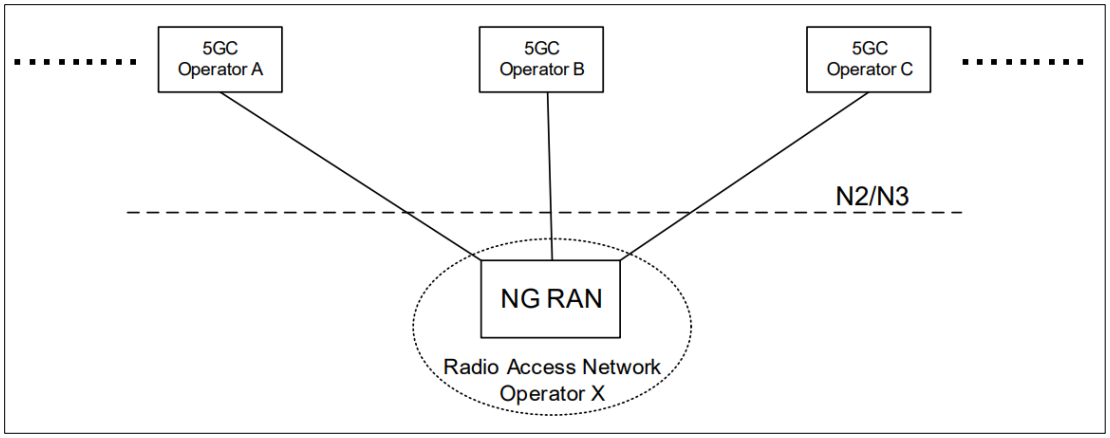


Figure 10 - 3GPP Specified 5G Multiple Operator Core Network (MOCN)

In the case of MOCN, since the access network is shared, the eNodeB/gNB scheduler plays a key role in deciding how the multiple operator subscribers are served. In traditional networks, scheduler algorithms are proprietary, however with MOCN, the scheduling method should be agreed upon by all the sharing entities. Whether the scheduler uses a 'proportional-fair' method, where no user is allowed to be starved providing a certain degree of fairness, or whether it uses an 'opportunistic approach', where users are assigned resources based on dynamic radio conditions using spectrum efficiently, depends on SLAs agreed upon by participating operators sharing the access network.

Current standards support the shared radio access network to connect to up to six core networks across a single radio carrier.

4.1.1. MOCN Testing at CLs

CableLabs tested the MOCN deployment in 2018 leveraging their in-house lab [17]. For the demo, four different virtualized EPC solutions simultaneously connected to a single eNB. Both Ethernet as backhaul and DOCSIS as backhaul, with cable modems and cable modem termination system between the LTE radio access network and core network, were tested. A variety of end user devices, including a plug-in adapter, an indoor CPE, an android phone and a USB dongle were used to connect to the shared eNB and one of the four core networks simultaneously.

Figure 11 shows a diagrammatic representation of the setup tested. The eNB is configured with multiple PLMN IDs, one for each of the core networks. Each PLMN has a 1:1 mapping to the S1 MME IP address of a specific core network. The eNB detects the PLMN in the RRC Connection Request from the end device and forwards the initial attach request to the correct S1 MME IP address of a core network to which the PLMN belongs.

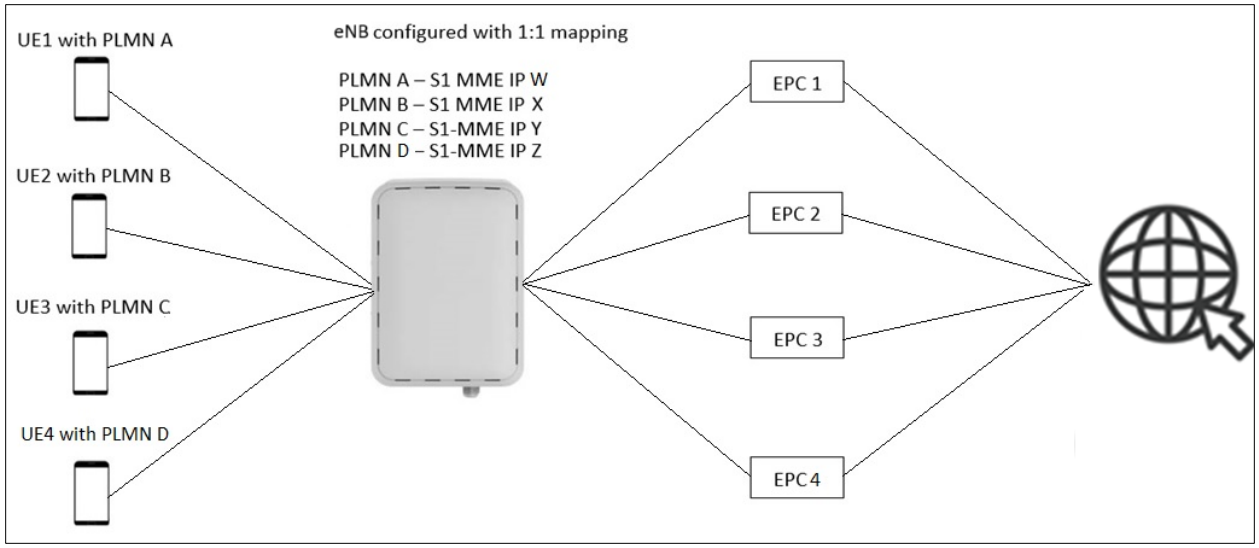


Figure 11 - Logical Setup of CLs MOCN Testing

Figure 12 shows the demo setup of a successful working of an end-to-end system with MOCN implementation, showcasing internet browsing by playing a YouTube video on laptops connected to each of the end devices, and on the android phone. Each of the end devices has a SIM card configured to connect to a different core network via the shared eNB simultaneously. This demo showcased the viability of the MOCN feature in 4G, highlighting the capability for multiple operators to utilize the shared eNB allowing their subscribers to transparently connect to their core network. It demonstrates how operators can leverage a standardized solution to maximize resource utilization, reduce costs and extend coverage while retaining control over their core networks and services.

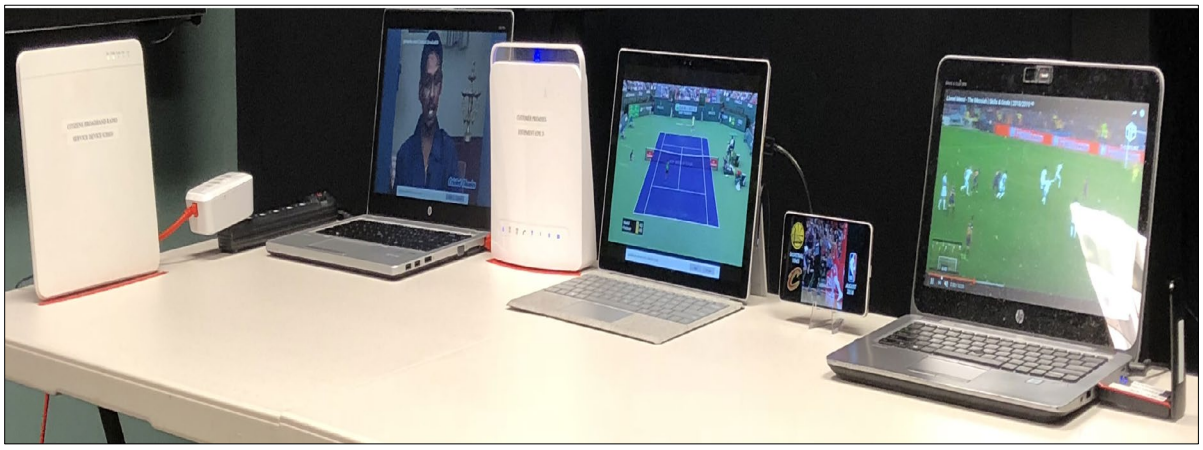


Figure 12 - Demo Setup of CLs MOCN Testing

4.2. Indirect Network Sharing (INS)

INS, similar to MOCN, enables users to access another operator's 5G network when they are beyond their own network's coverage, ensuring uninterrupted 5G services [18]. This approach enhances 5G resource

utilization, extends coverage to remote areas, improves user experience, and fosters the advanced development of communication infrastructure. While MOCN allows operators to share a single NG-RAN while maintaining independently operated 5GCs, it presents challenges associated with maintaining numerous N2 and N3 interfaces across the shared RANs and multiple core networks [19]. By collaborating to build a shared network, operators can share infrastructure without mandatory direct links between the shared NG-RAN of the host operator and the core networks of participating operators. For users, network services remain seamless, allowing UEs to access their subscribed PLMN and hosted services through participating operators when they connect to the shared NG-RAN of the host operator.

As shown in Figure 13, with indirect network sharing, a shared RAN broadcasts multiple PLMN IDs—one for the hosting operator and others for participating operators. The serving AMF of the host operator handles these PLMN IDs, enabling UEs from participating operators to select and connect to their respective PLMNs based on existing procedures. The core network functions (such as AMF, SMF, UDM, AUSF) serve UEs based on the PLMN ID of the participating operator or the hosting operator, guided by principles such as home routing. During procedures like PDU Session Establishment, the serving PLMN ID is chosen according to whether functions are interacting across the shared or the hosting network, ensuring correct routing and policy application in a multi-operator, shared infrastructure environment [21].

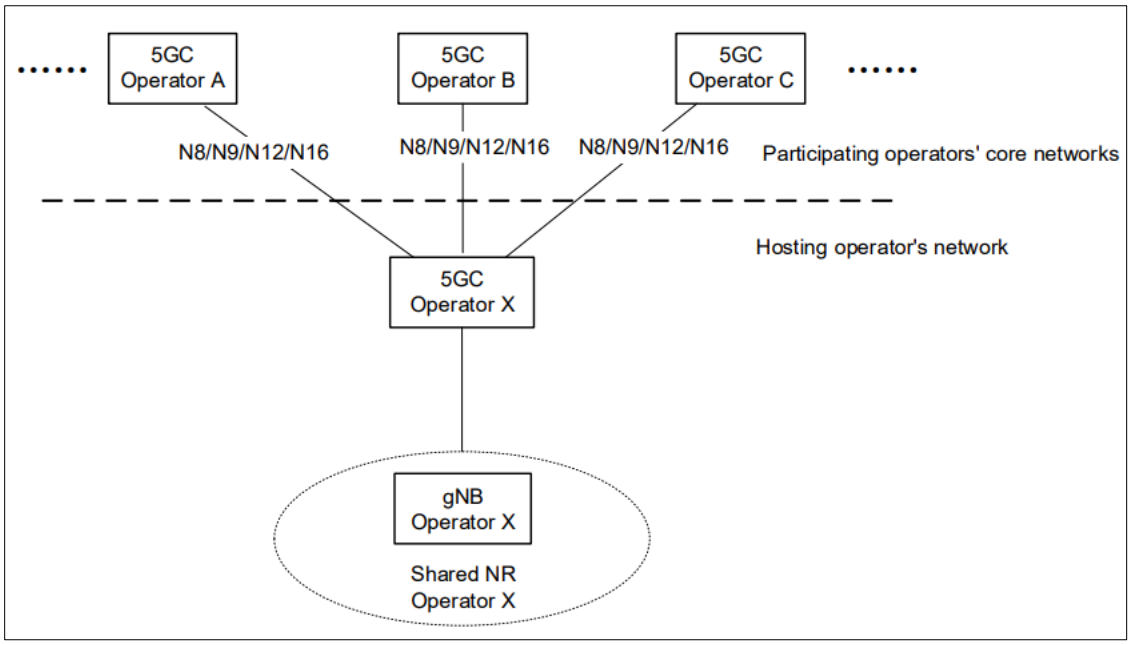


Figure 13 - 3GPP Specified 5G Indirect Network Sharing (INS)

Sharing RAN infrastructure and roaming to a visited network are two distinct concepts in mobile networking. RAN sharing involves multiple operators jointly utilizing the same physical radio access network infrastructure—such as towers, antennas, and sometimes even base station equipment—within a specific geographic area. This allows operators to reduce costs, improve network coverage and capacity, and optimize resource utilization, while each operator retains its own core network and subscriber management. On the other hand, roaming allows a subscriber from one operator to access network services through another operator's network when outside their home network's coverage area. In this scenario, the user's home network manages billing and subscriber services, and the connection typically

incurs higher charges due to inter-operator agreements. Unlike RAN sharing, roaming is characterized by temporary access to a foreign network rather than a permanent infrastructure sharing arrangement, and it often involves different regulatory and operational frameworks. Additionally, network sharing is invisible to end users - they connect to their home operator's service running on shared infrastructure without knowing about the arrangement. Roaming is, instead, explicit, with users receiving notifications when switching to a partner network.

5. Novel non-3GPP Access Network Sharing Architectures

Existing 3GPP specifications focus primarily on network sharing for 3GPP radio access technology (RAT). Extending existing network sharing architecture to include non-3GPP access allows operators to create a more comprehensive and versatile network footprint. Integrating these disparate access types enables seamless mobility and service continuity, opening new revenue streams through bundled service offerings, while simplifying inter-operator network sharing across RATs. By embracing non-3GPP access in network sharing, operators can create a more efficient, flexible, and customer-centric telecommunications landscape.

5.1. Multi-Operator Core Network (MOCN) for Non-3GPP Access

As shown in Figure 14, implementing Multi-Operator Core Network (MOCN) for non-3GPP access involves extending the core-sharing principles established within 3GPP NR systems to heterogeneous access technologies such as Wi-Fi (untrusted or trusted). Currently, in 3GPP NR, MOCN enables multiple operators to share the same radio access network while maintaining separate subscriber data and core networks, with shared control and radio resources managed by a centralized Radio Resource Management (RRM) system.

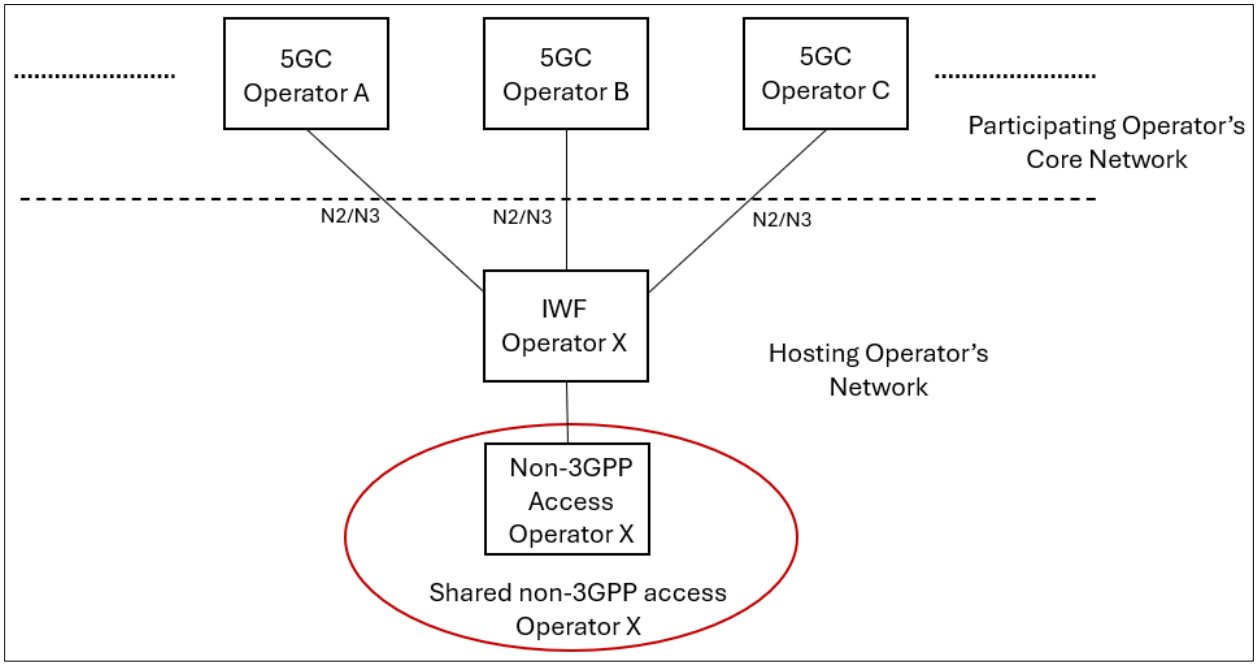


Figure 14- Proposed Multi-Operator Core Network Architecture for Non-3GPP Access Sharing

Although, an interworking function is similar to an access network node that supports N2/N3 interfacing with core networks, they are defined in TS 23.501 under the "Network Functions" section while the NG-RAN functions and entities are described in TS 38.300 and TS 38.401. In majority of the commercial deployments, where the 3GPP and non-3GPP access are owned and operated by different operators, the non-3GPP access is treated as untrusted and the N3IWF is owned by the participating operator rather than the host operator. The primary reason is the reluctance of the operators to open up N2/N3 interfacing to their AMF. The 3GPP standards define multiple IWFs connecting to a serving network (e.g.: different IWFs supporting different slices), but the 3GPP standards do not explicitly define procedures for connecting an IWF (N3IWF or TNGF) directly to multiple operator core networks simultaneously.

Table 2 below highlights key aspects of 3GPP MOCN for NR access sharing versus new proposed MOCN architectures for non-3GPP access sharing.

Table 2 - MOCN for 3GPP Access Sharing vs MOCN for Non-3GPP Access Sharing

Comparison metric	3GPP MOCN for NR access sharing	Proposed MOCN for non-3GPP access sharing
Shared Access Technology	5G New Radio (NR)	Wi-Fi (untrusted or trusted)
Other Network Components	Core networks remain separate; shared RAN infrastructure broadcasts multiple PLMN IDs, with user selection based on subscription	Core networks remain separate; shared TNAN advertises information about supported PLMNs (hosting PLMN and participating PLMNs)
Authentication	Authentication handled via 3GPP authentication procedures (e.g., 5G-AKA) at the core network	Utilizes authentication methods compatible with non-3GPP access and the operator cores, like EAP/802.1X for Wi-Fi, incorporating 3GPP credentials for enhanced security
Mobility Management	Handover procedures defined within the 5G NR standards; seamless mobility across the shared RAN	Requires interworking solutions to support mobility between shared non-3GPP access points and potentially to 3GPP networks, often involving tunnel management and session continuity

Spectrum Handling	Each operator controls and uses its dedicated licensed spectrum, but shares radio resources within the MOCN	Spectrum management involves unlicensed spectrum (e.g., Wi-Fi), requiring coordination and interference mitigation
Interworking	Not applicable, as RAN natively connects to the 5GC	IWFs act as a bridge connecting non-3GPP access to the different operator cores, managing protocol translation, security, and session management
Security	5G NR security framework is inherently enforced	Security relies on a combination of native non-3GPP security mechanisms (e.g., WPA3- Enterprise for Wi-Fi) and secure tunneling to the core network via the IWF
Scalability and Flexibility	Supports a wide range of 5G NR deployment scenarios and network configurations	Scalability depends on the IWF's capabilities to manage concurrent connections and throughput; flexibility depends on the configuration options provided by IWF and its integration with the cores
Key Benefits	Reduced infrastructure costs, extended coverage, enhanced network utilization, operator control over their subscriber management	Cost-effective expansion of services for mobile and non-mobile operators, seamless integration of non-3GPP access into 5G ecosystems, extended reach, higher user capacity, service innovations
Key Challenges	Requires tight coordination and standardization across operators, especially for resource scheduling and interference management	Involves complex interworking function design, protocol translation, and management of heterogeneous security frameworks, ensuring mobility and QoS across multiple cores and access methods

5.2. Indirect Network Sharing (INS) for Non-3GPP Access

In essence, indirect network sharing (INS) relates to supporting RAN sharing without a direct connection between the shared NG-RAN of the hosting operator and the participating operator's core network. The 3GPP Rel-19 specified INS architecture enables communication between the shared RAN and the participating operator's core network being routed through the hosting RAN operator's core network. INS addresses a key challenge for the operators with regards to the maintenance of the interconnections (e.g., number of network interfaces) between the shared RAN and participating operators' core networks, especially in the case of a large number of the shared base stations. Although this issue does not exist with the deployments of non-3GPP access, there are other potential challenges.

In 5G, for inter-operator scenario, the untrusted non-3GPP access network connection to a 3GPP core is via the non-3GPP interworking function (N3IWF). Typically, regardless of whether the non-3GPP access is owned and operated by the hosting network (managed Wi-Fi) or the non-3GPP access is a public Wi-Fi (unmanaged Wi-Fi), it's treated as untrusted non-3GPP access by the participating operator. In case of untrusted non-3GPP access, the 3GPP and non-3GPP access registrations are independent of one another. Once the UE registers with untrusted non-3GPP access using non-3GPP access credentials, and has internet access, it connects to the N3IWF by formulating an FQDN that resolves to a public IP address, where the traffic flows from the non-3GPP access to the N3IWF over the internet. This is an unmanaged connection where the operator has no control over how the traffic traverses from the untrusted non-3GPP access to N3IWF which may result in potentially high latency.

One way to resolve this issue could be with using the INS architecture for non-3GPP access sharing similar to the specified INS architecture for 3GPP access. Figure 15 shows the proposed INS architecture for non-3GPP access sharing. In this scenario if the non-3GPP access is owned and operated by the hosting operator, the host operator can treat the non-3GPP access as a trusted non-3GPP access network. Trusted non-3GPP access can benefit with direct authenticated access to the Wi-Fi network using the UE's SIM credentials, thereby creating a more streamlined authentication process and reducing latency, unlike untrusted non-3GPP access where the UE has to authenticate twice, once to gain access to the Wi-Fi network and then for the untrusted access to the 5GC via N3IWF. Furthermore, with non-3GPP access network owned and operated by the host operator, the operator can have complete control over how the traffic traverses to participating operator's core network via the host operator network. This enables end-to-end QoS policies and traffic prioritization that are typically difficult to implement across untrusted networks.

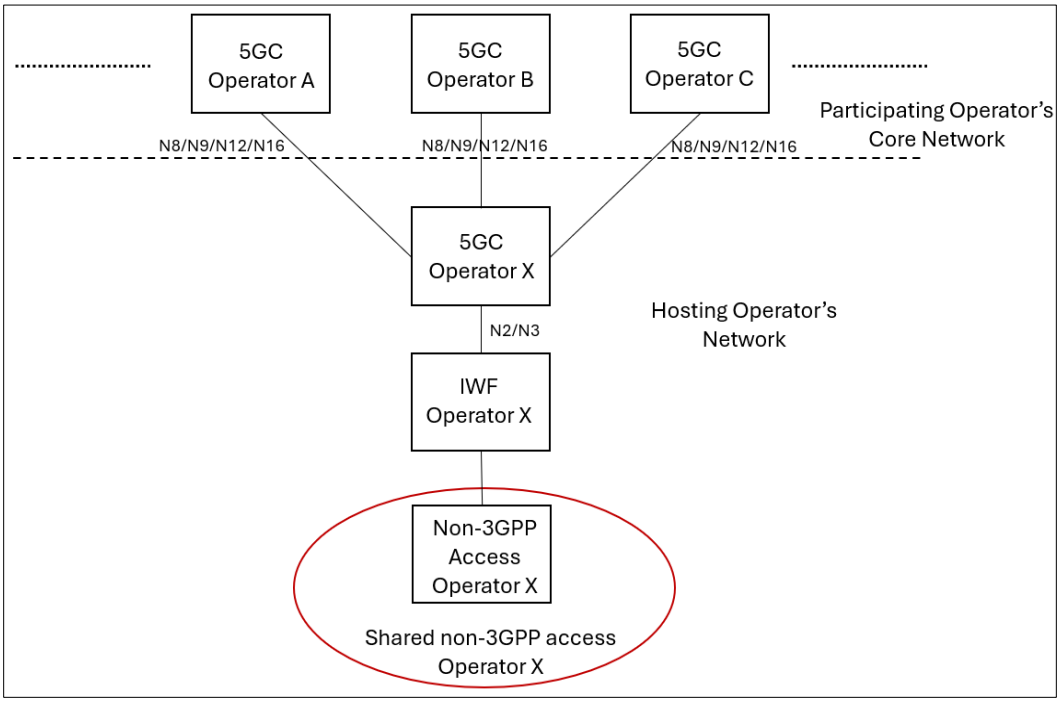


Figure 15 - Proposed Indirect Network Sharing Architecture for Non-3GPP Access Sharing

Additional capabilities enabled through sharing of the hosting operator's non-3GPP access as trusted to the participating operator include unified policy management becomes feasible allowing operator policies for data usage, content filtering, and service prioritization to be applied consistently regardless of the access technology, better network optimization because the hosting operator has direct visibility into trusted access networks, and reducing processing overhead and complexity by allowing the traffic of the participating operator's UE to be carried without additional encryption layers.

Table 3 below highlights key aspects of INS for 3GPP NR access versus proposed INS architectures for non-3GPP access sharing.

Table 3 - INS for 3GPP Access Sharing vs INS for Non-3GPP Access Sharing

Comparison metric	3GPP INS for NR access sharing	Proposed INS for non-3GPP access sharing
Access Technology	5G New Radio (NR)	Wi-Fi (untrusted or trusted)
Shared Network Components	RAN is shared and broadcasts multiple PLMN IDs; the serving AMF (Access and Mobility Management Function) selects core network functions in the PLMN of the participating operator based on routing architecture principle	Hosting Wi-Fi operator owns the WLAN and the serving core network. Guest or partner mobile operators connect through secure core-to-core interconnections. The shared TNAN advertises information about hosting PLMN and participating PLMNs; authentication routes the user to the appropriate core
Authentication	Authentication occurs in the PLMN of the participating operator (based on existing 3GPP procedures)	Authentication occurs in the hosting Wi-Fi operator's core (could use the existing Wi-Fi AAA servers for underlay Wi-Fi network in case of untrusted non-3GPP access), with user credentials mapped to the appropriate partner mobile operator's core, relying on protocols like EAP
Mobility Management	Seamless handover procedures between shared RAN resources, within the 5G system	Requires interworking solutions to handle mobility between the Wi-Fi network and the partner mobile operator's network, often involving tunneling protocols and session continuity techniques.

Spectrum Handling	Each operator utilizes licensed spectrum; shared RAN coordinates resource allocation and ensures non-interference	Spectrum management involves unlicensed spectrum (e.g., Wi-Fi), where spectrum management involves contention-based mechanisms requiring coordination and interference mitigation
Interworking	Serving AMF acts as the primary interworking function within the shared RAN	IWFs act as a bridge connecting non-3GPP access to the different operator cores, managing protocol translation, security, and session management
Security	5G NR security is end-to-end, within the 3GPP framework	Security relies on tunneling protocols, authentication methods (EAP), and policy enforcement within the Wi-Fi operator's core. Core-to-core connection has enhanced encryption.
Scalability and Flexibility	Highly scalable and flexible for multi-operator deployments within the 5G ecosystem	Depends on the performance of the tunnels, and the architecture of the primary Wi-Fi operator's core.
Key Benefits	Enables rapid and efficient deployment of 5G services, optimized resource utilization, and reduced infrastructure costs.	Leverages existing Wi-Fi infrastructure for expanded service offerings, allows mobile operators to offload traffic seamlessly, provides cost-effective solutions, and enables new business models.
Key Challenges	Requires careful planning for inter-operator coordination, policy enforcement, and resource allocation to guarantee performance guarantees for different subscribers.	Relies on established core-to-core agreements that can be complex to negotiate, may introduce latency overhead due to tunneling, and can be limited by the capacity and performance of peering agreements.

5.2.1. New Deployments Envisioned with Indirect Network Sharing (INS) for Non-3GPP Access

Sharing non-3GPP access through Indirect Network Sharing (INS) has the potential to transform the industry landscape for MSOs and MVNOs, especially for those whose unique selling proposition (USP) is their extensive cable infrastructure [20]. By enabling these operators to leverage shared access

technologies such as LTE, 5G, and Wi-Fi without heavy investments in physical infrastructure or spectrum, INS allows them to rapidly deploy next-generation connectivity services. This approach significantly reduces deployment costs and time-to-market, empowering MSOs and MVNOs to expand their service portfolio into areas like private LTE/5G, IoT, enterprise solutions, and 5G-based fixed wireless access—subscription models that align with their existing cable-centric business. Moreover, INS fosters a more flexible, scalable, and converged network architecture, enabling these operators to unify their services across multiple access technologies while maintaining control over user experience. This strategic shift can disrupt traditional industry boundaries, positioning MSOs and MVNOs as converged service providers capable of competing more aggressively in the future digital ecosystem, extending their core cable advantage into the wireless and IoT domains and unlocking new revenue streams without compromising their USP of extensive cable footprint.

5.2.1.1. Indirect Network Sharing of Both 3GPP and Non-3GPP Access of Hosting Operator

Figure 16 shows the host operator X, sharing both 3GPP and non-3GPP accesses leveraging the specified INS architecture. The same inter-core interfaces between the host operator's core network and participating operator's core network specified as part of the INS architecture for 3GPP access sharing can be utilized for sharing the 3GPP and non-3GPP access of the host operator. Both the 3GPP and non-3GPP access networks owned and operated by Operator X and shared by the participating operators integrate with the host operators core network which interfaces with the participating operators' core networks. Thus, the participating operators can share the host operators access networks without directly interfacing with them. This is similar to a home routed model being applied to both 3GPP and non-3GPP access sharing, by routing all the traffic from the serving network (host operator's network) to the home network (participating operator's network). As discussed earlier, the only difference being the shared (R)AN of Operator X broadcasting the network identifiers for Operator X as well as for all the participating operators (Operators A/B/C).

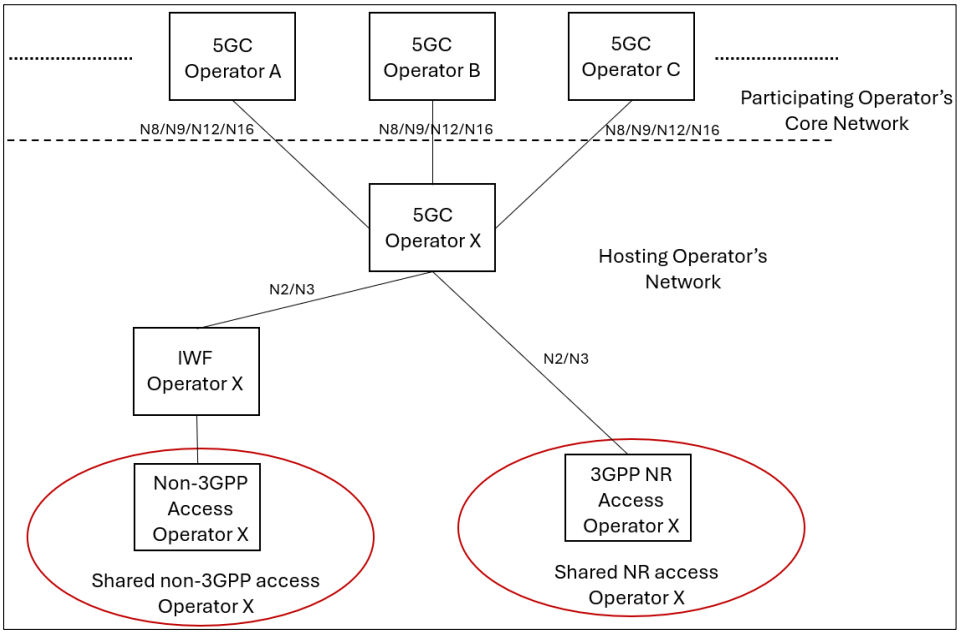


Figure 16 - Indirect Network Sharing of Both 3GPP and Non-3GPP Access of Hosting Operator

Since both 3GPP and non-3GPP access networks belong to and are managed by Operator X, the non-3GPP access network could be treated as trusted access, using a trusted non-3GPP access network (TNAN) and Trusted Network Gateway Function (TNGF) as the interworking function (IWF). In the scenario of trusted access, the TNAN may advertise the PLMNs or SNPNs for which the access network supports trusted connectivity. Thus, the non-3GPP access network selection could be performed using the UE's 5GS credentials, by first selecting a PLMN/SNPN and then selecting a non-3GPP access network (a TNAN) that supports trusted connectivity to the selected PLMN/SNPN. In this case, the non-3GPP access network selection is affected by the PLMN/SNPN selection. Another alternative is for the UE to perform non-3GPP access network selection using its 5GS credentials without having to register with the 5GS, by implementing a Non-Seamless WLAN Offload Function (NSWOF) instead of the IWF which would interface with the Wireless Local Access Network (WLAN) of Operator X and the Authentication Server Functions (AUSFs) within the 5GC of participating operator's network.

5.2.1.2. Indirect Network Sharing of Different RATs Across Operators

Another architecture that could be envisioned is when two operators share each other's RAT. Figure 17 shows Operator X who owns and operates a non-3GPP access network that is shared with Operator Y, and Operator Y who owns and operates a 3GPP access network shared with Operator X. So, for non-3GPP access sharing, Operator X is the hosting operator and Operator Y is the participating operator, while for 3GPP access sharing, Operator Y is the hosting operator and Operator X is the participating operator. And the same inter-core interfaces shared between the Operator X and Operator Y networks are leveraged for indirect network sharing of different RATs each operator owns.

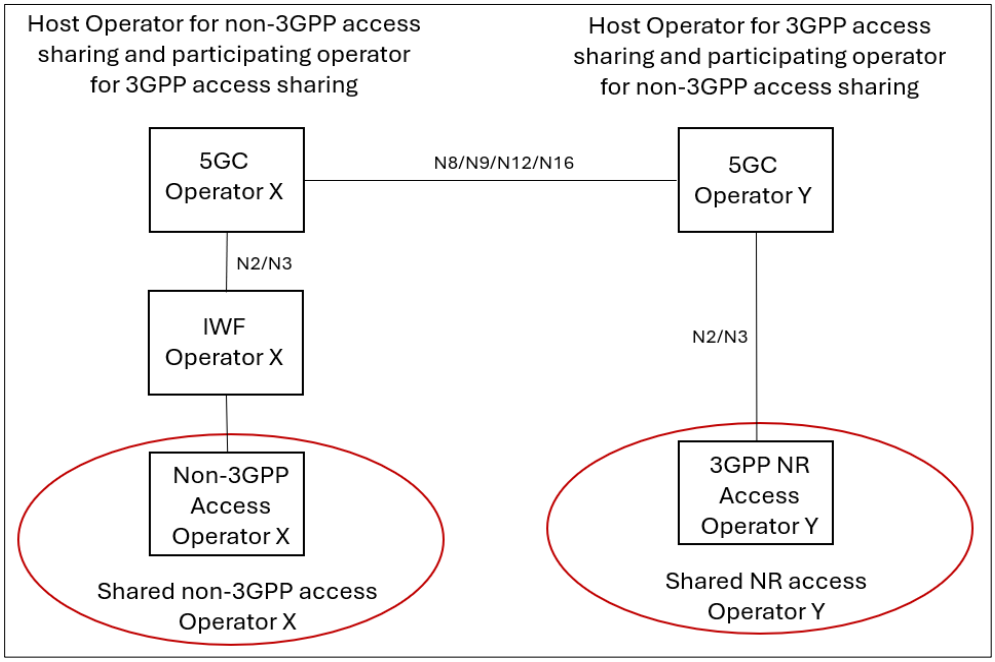


Figure 17 - Indirect Network Sharing of Different RATs Across Operators

5.3. Advantages of Non-3GPP Access Network Sharing Architectures: MOCN and INS

Extending network sharing paradigms such as Multi Operator Core Network (MOCN) and Indirect Network Sharing (INS) to non-3GPP access networks (e.g., Wi-Fi) offers a suite of synergistic benefits that dramatically enhance the operational capabilities, cost-effectiveness, and innovation potential for Wi-Fi operators, mobile operators or hybrid operators.

- **Extended Coverage Footprint:** Leverages diverse access technologies to provide ubiquitous coverage, especially in remote areas, improving service availability for users regardless of location.
- **Seamless Multi-Access Connectivity:** Creates a truly converged network environment where users can move between cellular and non-3GPP access points without service interruption, delivering a seamless experience.
- **Increased Network Resilience:** The diversity of access technologies in a shared network enhances robustness and reliability, ensuring backup connectivity options during network outages or congestion events.
- **Faster Time-to-Market:** By leveraging existing infrastructure and standardized interfaces, operators can quickly roll out new services and applications, such as fixed wireless access, IoT solutions, or location-based services, without lengthy deployment cycles.
- **Service Customization and Differentiation:** Each operator can retain control over its subscriber management, billing, and core network functions, enabling them to offer differentiated service packages, customized user experiences, and unique value-added services while sharing the underlying infrastructure.
- **Flexibility for Emerging Applications:** Supports cutting-edge applications like augmented reality (AR), virtual reality (VR), and low-latency gaming by providing the necessary bandwidth, QoS guarantees, and mobility management across RATs.
- **New Revenue Streams:** Enables new revenue opportunities by enabling private network offerings, and service bundling combining cellular and non-3GPP connectivity.
- **Ecosystem Collaboration:** Fosters collaboration between operators owning diverse infrastructures, creating an innovative ecosystem around converged network solutions.

In essence, non-3GPP access network sharing architectures unlock a new era of convergence, driving unprecedented efficiency, innovation, and profitability for all participants while providing users with a seamless, high-quality, and value-driven connectivity experience.

6. Conclusions

This paper has presented innovative architectural frameworks for extending the well-established concepts of MOCN and Indirect Network Sharing (INS), traditionally defined within the 3GPP ecosystem, to encompass a diverse range of non-3GPP access technologies. By addressing interoperability challenges and proposing new deployment models, we demonstrate how these architectures can enable a more

flexible, scalable, and unified wireless ecosystem that integrates multiple access methods such as Wi-Fi, with cellular. While Wi-Fi network sharing already occurs in practice, where operators share Wi-Fi infrastructure in venues like airports, shopping centers, or public spaces while maintaining separate authentication and billing relationships with their customers, the adoption of such novel architectures promises significant benefits, including accelerated deployment of converged services, optimized resource utilization across heterogeneous networks, and streamlined operational management for operators.

Crucially, these architectures open new avenues for MSOs, MVNOs, and other industry players to leverage their existing infrastructure—such as cable, Wi-Fi, and private LTE/5G, while maintaining independence, control, and service differentiation. This approach supports rapid service innovation, cost-effective expansion into new verticals like IoT and enterprise solutions and enhances user experience through seamless connectivity. Embracing these innovative sharing models that incorporate non-3GPP accesses will be pivotal in shaping the future of network convergence, fostering a more integrated, efficient, and industry-disruptive wireless landscape that meets the evolving demands of next-generation connectivity. Ultimately, such architectures will enable stakeholders to build resilient, agile, and commercially viable networks that drive competitive advantage and industry growth in a rapidly changing digital world.

7. Abbreviations and Definitions

7.1. Abbreviations

3GPP	3rd Generation Partnership Project
5G	fifth generation
5GC	fifth generation core
AUSF	authentication server function
CAPEX	capital expenditure
CBRS	Citizens Broadband Radio Service
CD	continuous development
CI	continuous integration
CLs	CableLabs
CPE	customer premise equipment
CUPS	control and user plane separation
DECOR	dedicated core
DCN	dedicated core network
DOCSIS	Data Over Cable Service Interface Specification
eMBB	enhanced mobile broadband
eNB	evolved node base station
EPC	evolved packet core
FQDN	gully qualified domain name
GWCN	gateway core network
INS	indirect network sharing
IoT	internet of things

IP	internet protocol
IWF	interworking function
LTE	long term evolution
MEC	multi-access edge computing
MME	mobility management entity
MNO	mobile network operator
MOCN	multiple operator core network
MORAN	multiple operator radio access network
MSO	multiple systems operator
MVNO	mobile virtual network operator
N3IWF	non-3GPP interworking function
NFV	network function virtualization
NG-RAN	next generation radio access network
NHN	neutral host network
NR	new radio
NSWO	non-seamless WLAN offload
NSWOF	non-seamless WLAN offload function
OPEX	operational expenditure
PLMN	Public Land Mobile Network
QoS	quality of service
RAN	radio access network
RAT	radio access technology
RRC	radio resource control
RRM	radio resource management
SDN	software defined networks
SIM	subscriber identity module
SLA	service level agreement
SNPN	standalone non-public network
TNAN	trusted non-3GPP access network
TNGF	trusted network gateway function
TS	technical specification
UE	user equipment
URLCC	ultra-reliable low-latency communication
USP	unique selling proposition
VNF	virtualized network function
WLAN	wireless local access network

7.2. Definitions

5G Access Network	An access network comprising a NG-RAN and/or non-3GPP AN connecting to a 5G Core Network.
5G NSWO	The 5G NSWO is the capability provided by 5G system and by UE to enable the connection to a WLAN access network using 5GS credentials without registration to 5GS.
Hosting NG-RAN Operator	The operator that has operational control of a Shared NG-RAN.
Hosting RAN	The Shared RAN that is owned or controlled by the Hosting RAN Operator.
hybrid access	Access consisting of multiple different access types combined, such as fixed wireless access and wireline access.
Indirect Network Sharing	A type of NG-RAN Sharing in which the communication between the Shared NG-RAN and the Participating Operator's core network is routed through the Hosting NG-RAN Operator's core network.
NG-RAN	A radio access network connecting to the 5G core network which uses NR, E-UTRA, or both.
NG-RAN Sharing	The sharing of NG-RAN among a number of operators.
non-public network	A network that is intended for non-public use.
Non-Seamless Non-3GPP offload	The offload of user plane traffic via non-3GPP access without traversing either N3IWF/TNGF or UPF.
Non-Seamless WLAN offload	Non-Seamless Non-3GPP offload when the non-3GPP access network is WLAN.
NR	The new 5G radio access technology.
Participating NG-RAN Operator	Authorized operator that is using Shared NG-RAN resources provided by a Hosting NG-RAN Operator.
Stand-alone Non-Public Network	A non-public network not relying on network functions provided by a PLMN
User Equipment	An equipment that allows a user access to network services via 3GPP and/or non-3GPP accesses.

8. Bibliography and References

[1] 3GPP, TS 22.261, "Service requirements for the 5G System (5GS)," v20.3.0, Jun. 2025.

[2] S. Marek, "UScellular CEO says network sharing will be a necessity," Fierce Network, Aug. 07, 2022. https://www.fierce-network.com/wireless/uscellular-ceo-says-network-sharing-will-be-necessity

[3] GSMA, "5G Network Co-Construction and Sharing Guide," Feb. 2023. https://www.gsma.com/get-involved/gsma-foundry/wp-content/uploads/2023/02/5G-NCCS_GSMA-Guide_27.02.2023.pdf

[4] A. Simmons, "Who Owns the 5G Cellular Towers in the United States?," Dgtl Infra, Dec. 18, 2020. https://dgtlinfra.com/who-owns-the-5g-cellular-towers/

[5] P. Koutroumpis, P. Castells, and K. Bahia, "To share or not to share? The impact of mobile network sharing for consumers and operators," in Information Economics and Policy, Dec. 2023.

[6] B. Fletcher, "KDDI, SoftBank tap Ericsson for multi-operator RAN," Fierce Network, Jun. 23, 2021. https://www.fierce-network.com/operators/kddi-softbank-tap-ericsson-for-multi-operator-ran

[7] 3GPP, TR 28.835, "Study on management aspects of 5G Multiple Operator Core Network (MOCN) network sharing phase 2," v18.0.1, Jul. 2023.

[8] R. Sharma, "VodafoneThree Integration Delivers Faster 4G & 5G Nationwide with Multi-Operator Core Network," Thefastmode.com, Aug. 15, 2025. https://www.thefastmode.com/technology-solutions/44004-vodafonethree-integration-delivers-faster-4g-5g-nationwide-with-multi-operator-core-network

[9] K. Samdanis, X. Costa-Perez, V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," in IEEE Communications Magazine. vol. 54, no. 7, pp. 32-39, Jul. 2016.

[10] J. O'Halloran, "BT unveils 5G Standalone network slicing deployment first in Belfast," ComputerWeekly.com, 2025. https://www.computerweekly.com/news/366617971/BT-unveils-5G-Standalone-network-slicing-deployment-first-in-Belfast

[11] X. Limani, M. Camelo, J.M. Marquez-Barja, N. Slamnik-Krieštorac, "Unveiling network sharing capabilities enabled by 5G network slicing," in IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-3, Mar. 2025.

[12] ETSI, "MEC in 5G networks," Jun. 2018. https://www.etsi.org/images/files/ETSIWhitePapers/etsi_wp28_mec_in_5G_FINAL.pdf

[13] C. S. Nin, "A closer look at Verizon's 5G MEC strategy," RCR Wireless News, Jan. 08, 2021. https://www.rcrwireless.com/20210108/5g/a-closer-look-at-verizons-5g-mec-strategy

[14] K. Ziser, "FirstNet's dedication starts with a dedicated mobile core," Lightreading.com, 2025. https://www.lightreading.com/mobile-core/firstnet-s-dedication-starts-with-a-dedicated-mobile-core

[15] Samsung U.S. Newsroom, "Samsung and KT Complete Korea's First 5G SA and NSA Common Core Network Deployment," Samsung.com, Nov. 04, 2020. https://news.samsung.com/us/samsung-kt-complete-koreas-first-5g-sa-nsa-common-core-network-deployment/

[16] 3GPP, TS 23.501, "System architecture for the 5G System (5GS)," v19.4.0, Jun. 2025.

[17] CableLabs. Webinar: CBRS Neutral Host Network (NHN) using Multi-Operator Core Network (MOCN). (Oct. 30, 2018). Accessed: Aug. 5, 2025. [Online Video]. Available: https://youtu.be/6ZsXsH2BUjk

[18] 3GPP, TR 22.851, "Feasibility Study on Network Sharing Aspect," v19.1.0, Sep. 2023.

[19] Q. Wei, T. Xing, Z. Wang, and C. Li, "Network sharing evolution," 3gpp.org, Mar. 2025. https://www.3gpp.org/technologies/net-share

[20] 3GPP, TS 23.502, "Procedures for the 5G System (5GS)," v19.4.0, Jun. 2025.

[21] O. Dharmadhikari, O. Choksi, and J. Kim, "Evolved MVNO Architectures for Converged Wireless Deployments," in Proceedings of the 2021 Fall Technical Forum, SCTE TechExpo, Oct. 2021. https://www.nctatechnicalpapers.com/Paper/2021/2021-evolved-mvno-architectures-for-converged-wireless-deployments

Click here to navigate back to Table of Contents

AI-Enhanced Optimization in HFC Networks: Real-Time Data Analytics and Adaptive Control

Modernizing HFC Maintenance Through Autonomous Network Intelligence

A Technical Paper prepared for SCTE by

Sahil Yadav, Sr. Director, Product Management, AOI
13139 Jess Pirtle Blvd.
Sugar Land, TX 77478
Sahil_Yadav@ao-inc.com
949-396-9611

Diana Linton, Principal Engineer I, Charter Communications
14810 Grasslands Dr,
Englewood, CO 80112
Diana.linton@charter.com
720-536-1027

Esteban Sandino, Distinguished Engineer, Charter Communications
14810 Grasslands Dr,
Englewood, CO 80112
Esteban.Sandino@charter.com
720-518-2318

Table of Contents

List of Figures

List of Tables

1. Introduction

Hybrid Fiber-Coaxial (HFC) networks form the backbone of broadband delivery for millions worldwide. Initially built for unidirectional analog broadcasting, these networks have evolved into complex, bidirectional systems capable of providing gigabit services via Data Over Cable Service Interface Specification (DOCSIS) standards [3]. As service expectations continue to grow, driven by video streaming, remote work, and latency-sensitive applications, ensuring network reliability and resilience has become a top priority for operators.

However, as infrastructure scales spanning over 1.5 million miles [4] in the U.S. alone, traditional maintenance practices relying on manual inspections and static threshold alarms are proving insufficient in today's scale-intensive and dynamic network environments. This reactive approach does not allow prediction of emerging issues. It often fails to identify root causes accurately and contributes to unnecessary technician dispatches and/or prolonged outages.

This paper introduces the concept of autonomous networks that use artificial intelligence (AI) to proactively improve network reliability, operational efficiency, and customer satisfaction. By predicting issues like signal drift or hardware aging, AI enables proactive resolution before customer service is impacted.

At its core, this framework shifts the operation model from static and reactive to dynamic, predictive, and adaptive. By leveraging real-time telemetry, adaptive algorithms, and closed-loop automation, the system supports continuous tuning of performance parameters even under growing demands from smart homes, remote work, and high-bandwidth streaming.

In the sections that follow, this paper outlines the system architecture, implementation methodology and challenges, and practical high impact use cases that validate the AI-powered autonomous network framework.

2. The Shift to Intelligence in HFC Networks: Enabling Reliable, Adaptive Operations

HFC networks have evolved significantly since their 1990s origins as analog, one-way broadcast systems. Driven by successive DOCSIS standards, they now support symmetrical multi-gigabit broadband and advanced video services [3]. This progression has added substantial complexity, blending optical and radio frequency (RF) technologies across fiber nodes, amplifiers, taps, and passive components.

Although advanced cable modem termination systems (CMTS) features and distributed access architectures have embedded deeper intelligence into networks, impairments such as ingress noise, temperature-induced drift, and power fluctuations can still pose challenges to operators. Traditional Simple Network Management Protocol (SNMP)-based tools and rule-based automation lack the adaptability needed to resolve these dynamic conditions, especially in complex HFC infrastructures.

Insights from a case study [1] revealed that static amplifier design based on worst-case design assumptions contributes to excessive signal spread and RF imbalance across customer legs. This study untapped the value of mass telemetry data. The analysis of cables modem received (Rx) and transmit (Tx) RF levels revealed that one-size-fits-all approaches lead to both underperformance and overcompensation. This underscores the need for dynamic, topology-aware intelligence within the network. By aggregating

metrics such as Rx/T RF levels, modulation error ratio (MER), and topological metadata, operators can expose systemic RF imbalances, identify poorly compensated legs, and detect early degradation patterns.

While machine learning has found applications in congestion prediction and RF pattern detection, many implementations remain siloed, reactive, and disconnected from real-time orchestration. These systems often rely on static thresholds or alarms, offering only minimal context about where, why, or how faults occur. To achieve autonomy, a system must overcome legacy integration challenges, limited data quality, and a lack of predictive, closed-loop automation.

Looking forward, the convergence of AI, machine learning (ML) models and software-defined networking (SDN) unlocks the potential for real-time control of network elements such as active settings, adaptive gain strategies, and spectral load balancing, which are key enablers of scalable, resilient HFC networks. By combining historical analysis with forward-looking control, autonomous architecture can deliver measurable improvements in mean time to repair (MTTR), signal consistency, and network reliability.

Ultimately, the mass availability of high-resolution telemetry reshapes how AI systems and field technicians interact with HFC networks. By providing granular, real-time insight into RF performance, signal integrity, and device health across the topology, telemetry empowers AI engines to detect, diagnose, and resolve faults autonomously, often before customer impact occurs. Meanwhile, it improves field intervention precision while reducing diagnostic time for technicians. Having a dual benefit for detection, root cause analysis, and remediation forms the foundation for autonomous, resilient access network.

3. Intelligence System Design for Autonomous Networks

The proposed AI-powered architecture follows a closed-loop model designed to continuously sense, interpret, and respond to dynamic network conditions in real time. This system encompasses four core functional layers as shown in Figure 1.

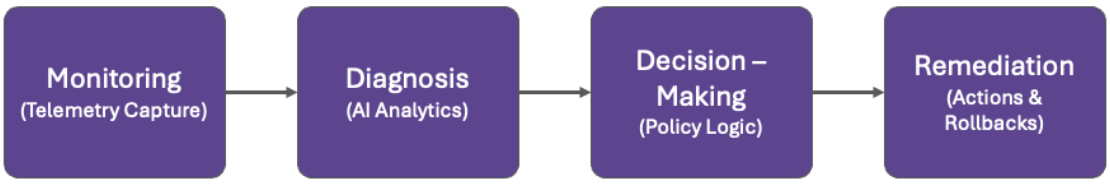


Figure 1 - Intelligence System Design for Autonomous Networks

3.1. Monitoring

The monitoring layer serves as the foundation of an autonomous system, providing continuous, high-frequency data capture across the HFC access network. This includes telemetry from active devices such as amplifiers, optical nodes, power supplies, and customer premises equipment (CPEs) as well as passive insights inferred from network behavior, historical trends, and topological metadata.

Key metrics collected include but are not limited to:

- **RF Performance:** Received and transmitted RF levels, downstream MER, and upstream signal-to-noise ratio (SNR), which are essential for identifying signal degradation, alignment issues, and noise ingress.

- **Device Health:** Temperature measurements, voltage fluctuations, and current draw, especially from power supplies and active network elements, which can indicate early signs of hardware failure, thermal drift, or localized power instability.

- **Error Metrics:** Correctable and uncorrectable codeword counts and modem reset statistics, which help identify chronic impairments, transient faults, or downstream impairments not visible through RF-only diagnostics.

- **Topological Context:** Cascade position, amplifier leg identifier (ID), and node relationships are captured or derived to support correlation and cascade-aware diagnostics.

This telemetry is ingested into a centralized data platform in near real time, enabling streaming analytics, anomaly detection, and historical trend comparisons. In the context of autonomous systems, the monitoring layer not only flags deviations from expected behavior but provides raw input for the next layers. It is this pervasive, high-resolution visibility that enables the network to evolve from a static, manually managed system to a dynamic, self-aware infrastructure capable of continuous adaptation.

3.2. Diagnosis

The diagnosis layer interprets the vast telemetry collected by the monitoring system and converts it into actionable intelligence. At this stage, ML models are applied to detect abnormal behavior, classify fault types, and locate root causes across the HFC topology.

Key capabilities include but are not limited to:

- **Signal Pattern Recognition:** ML models analyze real-time RF performance data such as MER dips, increases on uncorrectable errors, or Tx/Rx power fluctuations to detect patterns indicative of specific faults (e.g., amplifier misalignment, ingress noise, or physical damage).

- **Historical Trend Analysis:** Time-series models, including long short-term memory (LSTM) and regression frameworks, compare current telemetry against historical baselines. By detecting early degradation trends, like a gradual decline in MERs or increasing power consumption, the system can detect performance degradation before it severely impacts customers.

- **Alarm Correlation and Causal Inference:** When multiple alarms or performance anomalies occur concurrently (e.g., SNR drops, modem resets, power draw spikes), the system uses correlation matrices, Bayesian networks, and topology-aware logic to group symptoms and trace them back to a common root cause such as a failing power supply or a plant ingress point.

- **Customer and Field Insight Integration:** The diagnosis layer can ingest structured input from external sources such as customer trouble tickets, technician notes, or customer relationship management (CRM)-integrated complaint patterns. Natural language processing (NLP) models

help categorize qualitative data and correlate it with telemetry-based indicators to validate or refine a diagnosis.

- **Topological Awareness:** Each diagnostic decision is enriched with knowledge of the device's position in the cascade (e.g., upstream vs. end-of-line amplifier), allowing the system to differentiate between local faults and those propagated across portions of the network.

It ensures the system doesn't just react to thresholds, but understands the why, where, and severity of anomalies as well. The result is a highly targeted and context-aware fault classification engine that feeds directly into automated or technician-assisted remediation workflows.

3.3. Decision-Making

The decision-making layer serves as an intelligent hub for an autonomous network architecture. It takes the classified faults from the diagnosis engine and determines the optimal remediation strategy. This is based on a combination of real-time context, operational constraints, and learned optimization strategies. This layer shifts the system from reactive fault response to proactive, autonomous decision execution.

Key functionalities include but are not limited to:

- **Remediation Strategy Evaluation:** For each diagnosed issue, the system evaluates a range of possible corrective actions such as amplifier gain/slope adjustments, profile resets, spectrum migration, or load balancing. These options are scored based on predicted impact, implementation complexity, and time-to-effectiveness.

- **Contextual Decision Logic:** Decisions are made based on real-time variables including customer density, service level agreement (SLA) tier (e.g., residential vs. enterprise), affected service groups, and current load or maintenance windows. For example, a corrective action that temporarily disrupts service may be deferred during peak usage hours or blocked entirely for critical accounts.

- **Policy and Constraint Enforcement:** The system adheres to operator-defined policies such as maximum allowed gain adjustments, locked frequency plans, DOCSIS configuration limits, and failover conditions. It ensures that automated actions never violate plant rules or compromise regulatory or service requirements.

- **Prioritization via Reinforcement Learning:** Reinforcement learning (RL) algorithms are employed to continuously learn from past outcomes and improve future decisions. These agents adjust the prioritization of alarms and actions based on effectiveness, false positive rates, and post-remediation verification. Over time, the system learns which actions resolve specific fault classes most reliably in different topologies and service environments.

- **Topology-Aware Decisioning:** Recognizing the role of cascade depth, amplifier leg performance, and network segment criticality, the decision engine adapts its response to optimize signal continuity, MER uniformity, and energy efficiency across the physical plant.

- **Multi-Tiered Decision Models:** The system operates at three decision tiers: emergency (e.g., immediate outage correction), operational (e.g., RF balancing and optimization), and strategic

(e.g., long-term gain plan revisions or predictive upgrades) each with its own logic and execution scope.

This layer ensures that every action taken by the autonomous system is not only technically valid but operationally intelligent, balancing speed, risk, and network impact. It enables the network to respond autonomously to decisions that traditionally require coordinated human analysis and intervention.

3.4. Remediation

The remediation layer executes targeted, policy-compliant actions that restore service performance based on fault classification and decision logic provided by previous layers. It acts as an automation engine that applies corrective actions directly to network elements, with real-time verification to ensure success and prevent unintended side effects.

Key functions include but are not limited to:

- **Action Execution:** The system supports a wide range of actions, from low-impact adjustments such as amplifier gain/slope tuning, power level corrections, and modulation profile resets to substantial interventions like upstream channel migration, service group reallocation, or node segmentation**.** The selected response matches the severity and scope of the diagnosed issue.
- **Policy-Driven Enforcement**: Every remediation action is governed by operator-defined policies that define allowable parameters, escalation thresholds, rollback triggers, and priority rules. For example, auto-adjustment may be limited to ±2 decibel (dB) under certain conditions or disabled entirely during outages.
- **Closed-Loop Feedback Validation**: After remediation is applied, the system immediately monitors telemetry from affected devices to verify improvement in key metrics such as MER, SNR, power levels, and error counts. If the action results in no improvement or causes deterioration, the system initiates a roll-back and flags the event for further review.
- **Risk-Aware Execution**: For high-impact actions (e.g., service group shifts), the system performs impact forecasting before execution, estimating the likelihood of success, customer impact, and operational trade-offs. Only actions meeting confidence and policy thresholds are executed autonomously.
- **Rollback and Human Escalation**: When remediation fails or telemetry responses are ambiguous, the system reverts to the previous state and escalates the issue to an operator with detailed information about the action tried, the fault diagnosed, and the metrics that did not normalize.
- **Multi-Device Coordination**: To maintain signal continuity across the topology, the remediation engine coordinates actions across cascaded devices (e.g., power supplies to multiple amplifiers). This ensures that any changes made to one device do not disrupt the overall system. The engine monitors the impact of coordinated actions in real-time, adjusting as needed to maintain optimal performance and minimize service interruptions.

By combining speed, precision, and accountability, the remediation layer completes the remediation loop. Automated fault response replaces reactive technician dispatches with controlled, intelligent automation, which continuously strengthens network reliability while maintaining operational control and oversight.

With technologies such as LSTM neural networks and clustering models, fault prediction and classification are made possible, while reinforcement learning helps tailor strategies to regions and customer density. The success of the system is not solely determined by fault resolution rates but also by traditional and evolving key performance indicators (KPIs), such as detection accuracy, false positive rates, root cause correlation strength, mean time to detect (MTTD), and remediation success [9].

4. Implementation Framework for Scalable, Modular & Autonomous HFC Networks

The system architecture, as shown in Figure 2, is inspired by the Monitor–Analyze–Plan–Execute–Knowledge (MAPE-K) model [8], which facilitates continuous feedback, adaptive learning, and policy-aligned action across a distributed HFC infrastructure. This framework supports both autonomy and operator visibility, which is critical for integrating AI into high-availability networks.

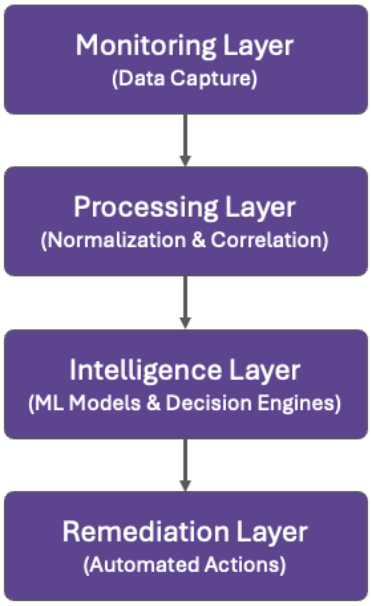


Figure 2 - Framework Implementation: Modular, Scalable, and Autonomous.

4.1. Data collection

A cornerstone of autonomous HFC networks is the ability to acquire continuous, high-fidelity telemetry from all critical network layers, from actives in the field to decision engines in the cloud. This section

outlines the architecture of the data collection pipeline that enables real-time monitoring, context-aware diagnostics, and autonomous remediation.

Architecture and Flow

1. **Smart Amplifiers**:
   Positioned throughout HFC networks, smart amplifiers serve as critical monitoring stations for real-time network visibility and control. These devices track RF and operational information, including input/output signal levels, gain settings, tilt, SNR, MER, power consumption, and thermal status. Amplifier health and RF conditions are monitored via embedded sensors and processed locally by integrated microcontrollers or digital signal processors (DSPs).

   As well as monitoring these parameters, smart amplifiers can also make closed-loop adjustments such as automatic level and slope control (ALSC). When integrated with AI-driven telemetry systems, they become active participants in autonomous network workflows.

2. **Transponders**:
   Each amplifier is paired with a transponder that acts as a telemetry gateway. Transponders transmit structured performance data using DOCSIS, Bluetooth low-energy (BLE), LoRaWAN, HSM or Ethernet/IP, depending on the site design. These devices package amplifier and environmental data and pass it upstream in near real-time. This enables amplifiers to transform from passive devices to smart, connected points, providing critical visibility into signal quality and impairment conditions across the HFC network.

3. **Node Aggregation**:
   Transponder outputs are ingested at the optical node level, where telemetry from multiple amplifiers and power supplies is consolidated. Here, initial filtering, timestamping, and metadata tagging occur, assigning each record a location context (e.g., node ID, leg ID, cascade level).

4. **Remote PHY Devices (RPDs)**:
   RPDs capture physical (PHY)-layer and DOCSIS telemetry from modems and CPEs, aggregating statistics like received and transmit levels, uncorrectable codeword rates, burst noise impact, upstream SNR, and active modulation profiles. They serve as high-resolution sensors for the physical layer of the access network.

5. **CMTS / CCAP Collection**:
   At the core, CMTS or converged cable access platform (CCAP) platforms aggregate telemetry from RPDs and directly from modems. These systems normalize disparate data sources and ensure synchronization of measurements across RF domains and time intervals.

Telemetry Aggregation and AI Engine:
All upstream telemetry is funneled into a centralized analytics engine, enriched with topological metadata (e.g., amplifier location, subscriber cluster, service group alignment). The pipeline supports both event-driven and scheduled ingestion models to ensure responsiveness and historical coverage.

Functional Benefits

- **Full-path visibility**: Enables fault localization from modem to amplifier.

- **Topology-aware correlation**: Contextual data allows for cascade-level fault tracing.
- **Real-time responsiveness**: Streaming architecture enables anomaly detection within seconds.
- **Integration flexibility**: Adapters support legacy devices lacking native telemetry.

Data normalization and enrichment: Ensures analytics operate with spatial and temporal awareness.

This end-to-end pipeline forms the digital nervous system of the autonomous network framework, allowing decisions to be made with high granularity, automation to be executed safely, and operators to have full situational awareness across even the most complex HFC footprints.

4.2. Analytics pipeline

Real-time and batch data are processed via hybrid polling/event-driven pipelines, enriched with topological metadata. These hybrid processing models blend real-time stream analytics with batch trend analysis.

- Streaming analytics enable low-latency detection of anomalies or performance drift as they emerge (e.g., sudden MER drop).
- Batch processing supports historical trend detection, periodic model retraining, and evaluation of network-wide patterns over long-time intervals.

Data is further enriched with topological metadata, historical baselines, and contextual overlays (e.g., service tier, customer impact). This fusion of live and historical data ensures temporal context and granularity for ML models and decision engines.

4.3. Machine Learning

The diagnostic and decision layers are powered by a diverse model ensemble tailored to specific operational goals:

- **Isolation Forests:** Detect anomalous behaviors and deviations from established baselines (e.g., unexpected MER fluctuations).
- **Ensemble Models:** Combine classification and regression approaches to predict fault types and failure likelihoods with high confidence.
- **Bayesian Networks & Gradient Boosting Trees:** Used for fault diagnosis, root-cause inference, and optimization of remediation strategies.
- **LSTM and Time-Series Regression:** Model performance degradation over time and anticipate service-impacting failures days in advance.
- **Reinforcement Learning Agents:** Continuously refine alarm prioritization, remediation sequence decisions, and policy-based escalation, learning from historical success rates and failure patterns.

- Training and validation are performed using real-world telemetry (e.g., modem data) to ensure operational validity and robustness.

4.4. Remediation logic

Once the AI engine has successfully identified and diagnosed a network fault, the system transitions from insight to action. This transition is powered by a remediation architecture, with a remediation flow sequence shown in Figure 3, that translates fault classification into targeted, policy-compliant corrective responses across the HFC infrastructure.

Core Components of the remediation Layer

1. **Remediation Policy Engine**
   This subsystem governs which corrective actions are allowed, under what conditions, and with what constraints. It leverages:

   - **Predefined action libraries** tailored to fault classes (e.g., amplifier imbalance, ingress noise).
   - **Operational policies** (e.g., max gain delta, peak-hour restrictions).
   - **Service-aware rules**, distinguishing between residential, business, and SLAs.

2. **Remediation Decision Selector**
   Based on the fault's characteristics, this selector chooses the optimal action path by evaluating:

   - Location (e.g., cascade level, amplifier leg)
   - Severity (e.g., service impact, performance degradation rate)
   - Historical effectiveness of remediation strategies
   - Real-time network conditions (e.g., upstream congestion)

   Techniques such as multi-objective optimization and reinforcement learning guide selection maximize success while minimizing disruption.

3. **Device Command Dispatcher**
   This interface translates remediation intent into actionable instructions, e.g.:

   - For amplifiers: gain/slope adjustments via SNMP, Message Queuing Telemetry Transport (MQTT), or proprietary application programming interfaces (APIs)
   - For modems: profile resets, transmit power tuning via Technical Report – 069 (TR-069) or CM configuration
   - For nodes/RPDs: spectrum reallocation or DOCSIS profile modification
   - For CMTS: upstream channel switching or scheduling adjustments

   All commands are validated for protocol compliance and transactional safety.

4. **Execution Validator (Closed-Loop Feedback)**
   Immediately after remediation is triggered, telemetry from affected devices is re-evaluated to:

   - Confirm normalization of KPIs (MER, SNR, error rates)

- Detect unintended side effects (e.g., impact on neighboring legs)
- Monitor rollback thresholds

If success is validated, the event is marked resolved. If not, the rollback mechanism is triggered, and the incident is flagged for operator escalation.

5. **Rollback Controller and Escalation Module**
   Should the remediation action fail, or metrics worsen:

   - Prior configuration states are reinstated
   - A human operator is notified through the dashboard with diagnostic summaries
   - Escalation metadata includes fault history, attempted fix, failure reason, and confidence scores

6. **Telemetry Logging and Learning Feedback Loop**
   All remediation attempts and outcomes feed into the ML training corpus:

   - Successful actions reinforce their weighting in future scenarios
   - Failed attempts are used to refine strategy selection and model sensitivity

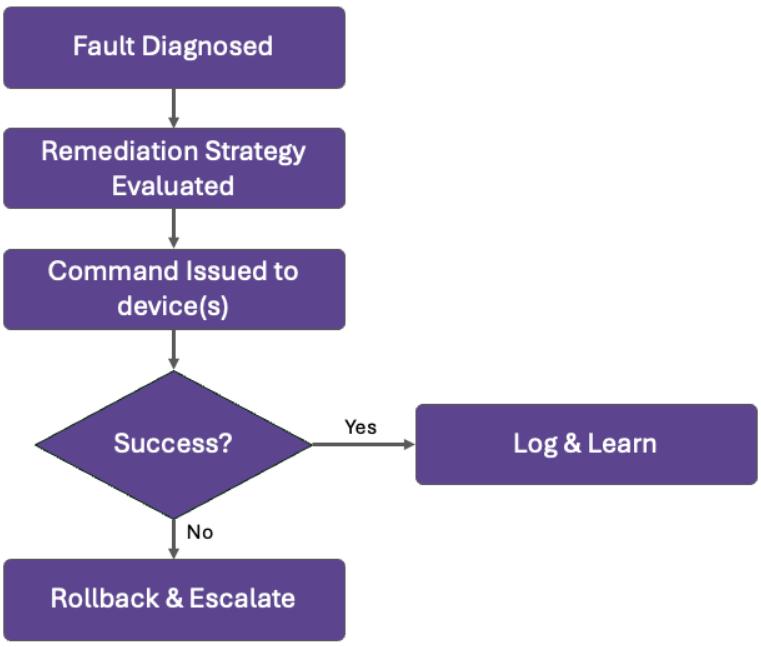


Figure 3 - Remediation Flow Sequence: Closed-Loop Automation with Validation and Escalation

5. Production Networks Operations: Modeling Autonomous Behavior in HFC Systems

The operational framework for AI-driven autonomous systems is built to withstand the unpredictability and velocity of real-world HFC environment. By aligning automation workflows with maintenance priorities and service-level constraints, this framework enables the practical execution of remediation logic across heterogeneous infrastructures, including legacy elements. It serves as the convergence point where AI insight becomes actionable remediation, with measurable impact on uptime, customer experience, and field operations.

5.1. Real-Time Operations

Real-time operation is a cornerstone of effective remediation in HFC networks, particularly as impairments can emerge and escalate within minutes. To support rapid detection and resolution, lambda architecture is used to combine stream (immediate anomaly detection) and batch (trend analysis) processing. Dimensionality reduction, correlation matrices, and frequency transformations enable root-cause tracing at scale.

The decision-making framework operates on three tiers, each powered by appropriate AI techniques and risk assessment models.

- **Emergency (rapid fault resolution):** supports rapid fault resolution for critical outages or significant RF anomalies. Triggers include but are not limited to sharp MER drops, group modem resets, or sudden power loss. This tier prioritizes autonomous RF tuning or failover actions.

- **Operational (balancing signal loads):** manages continuous optimization, such as balancing signal levels across amplifier cascades or adjusting network device settings based on time-of-day or load conditions. Models training based on field data enhances operational routines by correlating network performance KPIs to maintenance routines or unexpected changes that can be mitigated with active or modem settings adjustments.

- **Strategic (capacity planning):** focuses on capacity planning, hardware prioritization, and predictive resource allocation. Trend forecasting from LSTM models informs proactive amplifier reconfiguration or scheduling field interventions before quality falls below SLA thresholds.

Remediation paths include SNMP command execution, dynamic CMTS policy updates, SDN-based traffic routing, and technician dispatch with precise pre-diagnostics.

5.2. Integration with Legacy HFC Equipment

Integration with legacy equipment is achieved through a combination of protocol adapters, proxy agents, and retrofit monitoring devices. For manageable devices lacking modern telemetry capabilities, we implement enhanced polling that extracts additional diagnostic information through command sequences. Adapter modules convert AI-driven intents into vendor-specific commands for older amplifiers or CMTS gear. Security wrappers ensure encrypted, auditable transactions even on legacy protocols like SNMP.

For unmanageable passive components, we deploy inference models that estimate status based on upstream and downstream signal characteristics. Where direct integration is impossible, strategically

placed sensors monitor key parameters including power levels, temperature, and signal quality metrics. This hybrid approach enables comprehensive visibility across heterogeneous equipment deployments regardless of age or management capabilities.

5.3. Human Oversight and Intervention Interfaces

While the system is highly automated, human oversight remains essential. Human operators can intervene in critical or ambiguous scenarios through explainable AI interfaces and escalation workflows. A centralized operator dashboard allows for real-time monitoring, manual overrides, and policy tuning. Critical or ambiguous decisions can be escalated for human review, with all AI actions logged for transparency and auditability.

Intervention workflows support human-in-the-loop control during high-impact scenarios, while alert routing ensures unresolved or low-confidence cases are flagged for technician input. This balance of automation and human control ensures safe, accountable operations across dynamic HFC environments.

A hallmark of mature AI remediation logic is its **graduated autonomy,** as shown in Figure 4.

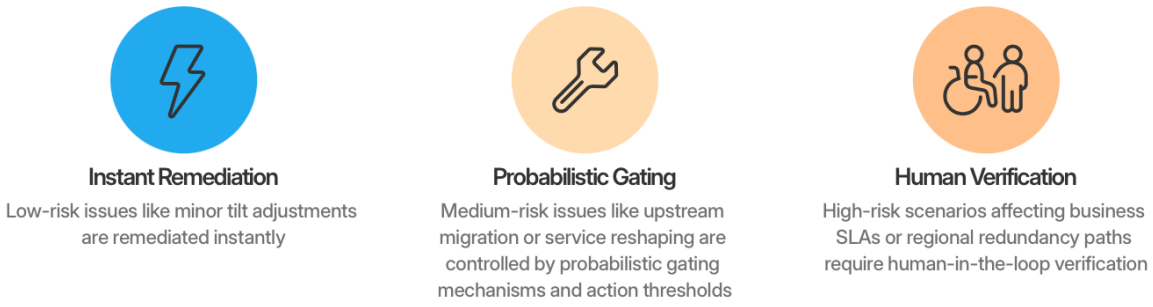


Figure 4 - Graduated Autonomy Model for Remediation Control

This layered governance model enhances operational trust, allowing organizations to incrementally expand AI authority as accuracy and operator confidence grow.

5.4. Detect–Mitigate–Prioritize–Resolve (DMPR) Model in Autonomous Systems

In HFC networks, autonomous systems must strike a balance between immediate continuity and long-term structural stability. A Detect–Mitigate–Prioritize–Resolve (DMPR) model, shown in Figure 5, separates real-time stabilization from permanent corrective actions, so networks remain resilient despite dynamic or degraded conditions.

Upon detection of anomalies the system initiates targeted mitigation actions to preserve service performance, even under degraded conditions. These actions are non-invasive and reversible, designed to stabilize the network while deep diagnostics are performed. Meanwhile, the model prioritizes faults based on service impact, topology risk, and probability of recurrence, enabling operators to allocate resources effectively.

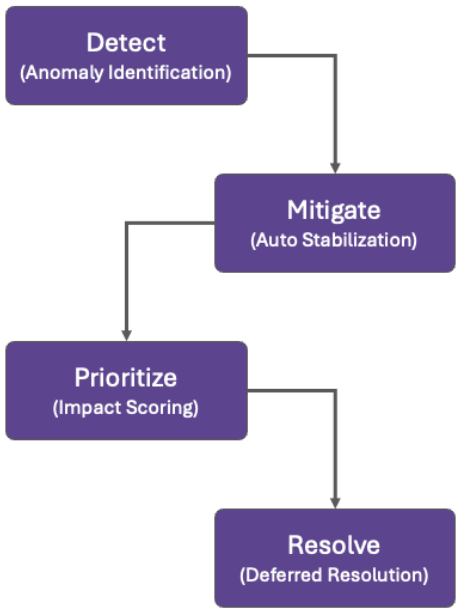


Figure 5 - DMPF Model: Detect–Mitigate–Prioritize–Resolve Workflow

5.4.1. Detect

The system continuously monitors high-resolution telemetry streams to detect deviations from learned baselines. Using anomaly detection, time-series forecasting, and cross-layer correlation, faults are identified before they result in noticeable customer impact. Rather than relying on static thresholds or reactive alarms, it uses a combination of anomaly detection algorithms, time-series forecasting, and cross-layer correlation to identify emerging faults.

Example: A sudden drop in upstream SNR, repeated modem resets, or rising uncorrectable codewords may indicate amplifier drift or ingress noise.

5.4.2. Mitigate

Once detected, the remediation layer attempts an immediate, policy-compliant mitigation, such as temporary gain adjustments or modulation profile resets. The goal is to contain service impact and stabilize key performance indicators, even if the root cause remains partially unresolved. Mitigation is not a permanent fix; it is a temporary containment strategy design to reduce service impact. For instance, if reserve gain is available, output power may be temporarily increased to normalize MER without triggering excessive distortion.

5.4.3. Prioritize

Post-mitigation, the system scores each incident for post-resolution based on:

- Severity of the original fault
- Number of subscribers impacted
- Duration of mitigation
- Proximity to other network risks
- Business importance of the affected service group

These scores feed into Proactive Maintenance & Assurance systems, allowing operators to triage deferred fixes based on impact and urgency.

5.4.4. Resolve

The final step involves dispatching field resources or initiating a structured repair, whether that means replacing degraded coax, replacing a waterlogged tap, or performing a firmware upgrade. Fixes are logged and correlated with the mitigation actions taken earlier to refine future policy.

Why DMPR Matters

- Prevents service-impacting escalations while buying time for engineering teams.
- Reduces operational noise by avoiding repeated alarms for the same root issue.
- Enables hybrid human-machine workflows, where AI handles mitigation and humans manage strategic repair.
- Supports opportunistic and consolidated repairs, allowing operators to bundle multiple deferred issues into a single field visit based on predictive maintenance insights, reducing truck rolls and increasing operational efficiency.

This model ensures that remediation or the fix does not become short-term patching at the expense of long-term network health. Instead, it orchestrates intelligent triage, automating what can be stabilized now, and prioritizing what must be permanently resolved later

5.5. Outcome Alignment with Network KPIs

Well-implemented remediation logic delivers:

- **Reduced MTTR** through autonomous triage and correction
- **Fewer technician dispatches**, as up to 31% faults are resolved autonomously [2]
- **Improved network stability**, due to closed-loop action verification
- **Operator control**, through override interfaces, explainability dashboards, and policy enforcement layers

By embedding the remediation engine directly into the AI architecture and tightly integrating it with real-time telemetry and decision logic, the framework ensures that each corrective action is traceable, testable, and safe, forming the final stage in a truly autonomous remediation loop. Notably, intermittent issues might be resolved significantly faster, potentially reducing repair times by up to 95% [7], particularly in legacy infrastructure, indicating possible value in maintenance-intensive environments.

6. Applying Intelligent Automation in the Field: Real-World HFC Use Cases

The deployment of AI-powered autonomous systems in HFC networks is justified not only by theoretical design but also by the tangible use cases it enables. These use cases captured in Table 1, define the operational, diagnostic, and performance-enhancing capabilities that such systems are expected to deliver. Below are six critical use cases, each aligned with key network management challenges and describing how the AI architecture fulfills them.

Table 1 - Use Cases

Use Case	Objective	System Functions	Expected Outcome
Baseline Metrics Establishment	Define normative performance benchmarks and detect deviations.	Use clustering models to define good operating ranges and context-aware baselines.	Improved fault detection sensitivity with fewer false positives.
Correlation – based root cause analysis	Trace anomalies to their root cause using multi-layer correlation.	Apply correlation matrices and causal models to group symptoms and pinpoint causes.	Fewer redundant alerts and faster root-cause resolution.
Alarm Prioritization	Rank alarms based on impact, urgency, and scope.	Score alarms using multi-factor analysis and historical resolution patterns.	Efficient resource allocation and higher resolution rates.
Performance Optimization	Identify latent issues and optimize network performance.	Leverage residual analysis and AI to identify inefficiencies or hidden degradations.	Improved long-term reliability and early problem detection.

Use Case	Objective	System Functions	Expected Outcome
AI-Based Remediation	Autonomously resolve diagnosed issues with minimal human intervention.	Trigger policy-driven remediation actions, verify results, and roll back if needed.	Reduced MTTR and fewer technician dispatches.
Data-Driven Design	Utilize telemetry data to improve HFC network design accuracy with dynamic updates.	Implement a data-driven approach, where real-time field data, instead of fixed specs, guides design	Accurate designs that result in measurable improvements in signal consistency, energy efficiency, and operational reliability.

6.1. Establishing Baseline Metrics and Optimal Operating Ranges

Challenge: Static thresholds fail to account for topological variation, environmental factors, or aging infrastructure. This leads to missed early warnings.

Objective: To define normative performance benchmarks for HFC network elements, enabling detection of failures, deviation or drift.

System Functionality: The monitoring layer continuously collects high-frequency telemetry, such as SNR, MER, uncorrectable codewords, temperature, and noise from nodes, amplifiers, and CPE. AI models, particularly unsupervised clustering algorithms, establish data distributions under "healthy" network conditions. These models adapt to geographical differences, topological variation and seasonal variations, recognizing what constitutes normal performance within different environmental and topological contexts.

Expected Outcome: This establishes dynamic baselines that improve fault detection sensitivity. For example, rather than relying on static thresholds (e.g., fixed dB levels), the system flags outliers relative to contextual baselines. This minimizes false alarms and enables early detection of performance deterioration that may precede outright failures.

6.2. Root Cause Analysis via Correlation of Multi-Layer Issues

Challenge: Symptoms commonly appear across layers (PHY, media access control (MAC), Internet Protocol (IP)), and/or devices, making it hard to isolate the root cause, which can lead to ineffective and incorrect troubleshooting.

Objective: To move beyond surface-level fault indicators and trace the causal chain of performance issues.

System Functionality: The analysis layer applies correlation matrices and temporal sequencing to understand interdependence among network events. For example, a sudden drop in upstream SNR across a node may correlate with a power fluctuation in a mid-span amplifier. Causal inference models, using techniques like Granger causality and Bayesian networks, trace symptoms back to their source. Cross-layer integration (PHY, MAC, IP) helps differentiate between local vs. propagated impairments.

Expected Outcome: The system reduces diagnostic ambiguity. Instead of generating isolated alerts for each affected device, it groups related anomalies into a single correlated event. This allows the remediation engine to target the true fault, e.g., a damaged tap rather than multiple falsely implicated modems, thereby avoiding redundant interventions.

6.3. Prioritization of Alarms Based on Service Impact

Challenge: There is a significant volume of alarms and similar fault signatures across devices or layers, leading to ambiguous or competing priorities. This can lead to overworked network operations center (NOC) teams and delays in identifying the most critical problems.

Objective: To intelligently rank concurrent alarms by their severity, impact scope, and urgency.

System Functionality: The decision engine incorporates multi-factor impact models, shown in Figure 6, that consider:

- The number of affected subscribers or service groups
- The class of service (e.g., business SLA vs. residential tier)
- Traffic load and utilization patterns
- Historical incident recurrence
- Risk of cascading failures

Each alarm is assigned a weighted priority score. Reinforcement learning fine-tunes this scoring based on historical outcomes, penalizing false positives and missed escalations.

Expected Outcome: This triage system ensures that field resources and automated remediations are allocated efficiently. For instance, a node-wide upstream degradation affecting 500 users takes precedence over a marginal downstream tilt on a single subscriber drop. In high-load scenarios (e.g., during storms), the system can escalate only the highest-impact issues for immediate resolution, preserving service continuity.

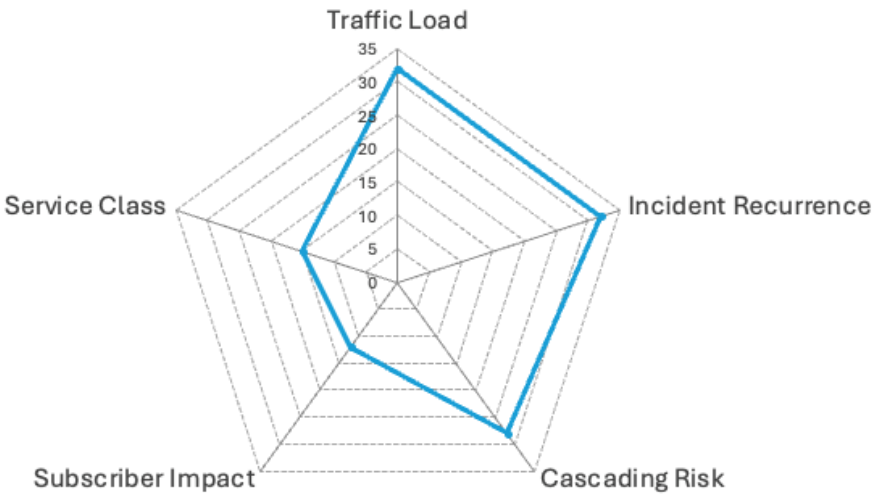


Figure 6 - Alarm Prioritization Scoring Factors. A radar chart displaying how alarms are ranked based on multiple operational impact criteria.

6.4. Performance Optimization and Latent Issue Detection

Challenge: When sub-threshold degradations occur, such as slow RF level drift, they tend to evade detection until there is a noticeable impact on the customer.

Objective: To uncover sub-threshold degradations or inefficiencies that do not trigger alarms but can impair long-term performance.

System Functionality: Pattern recognition and residual analysis models identify subtle shifts from baseline. For example, recurring but minor MER dips during temperature swings may indicate impending equipment failure or damaged components that are degrading. The analytics engine flags these as low urgency "latent anomalies," classifying them as candidates for proactive maintenance. A gradual pattern of signal degradation, transient optical level instability, or increasing uncorrectable errors may appear in the affected node area several weeks or days before a complete fiber break occurs, especially in cases where a fiber sheath fault is developing or an impending cut is imminent. These pre-failure signatures can alert operators to dispatch field teams for inspection or fiber reinforcement, potentially avoiding a broad outage.

Additionally, AI models detect chronic underutilization or spectral imbalance across service groups. Recommendations for profile tuning, load redistribution, or amplifier equalization are generated to optimize network efficiency.

Expected Outcome: Stretch goal capabilities allow operators to transition from reactive to truly predictive maintenance. These insights reduce customer-impacting incidents by addressing issues before they manifest as alarms, while also guiding long-term network health improvements.

6.5. AI-Driven Automomous Remediation Through Real-Time Fault Identification

Challenge: Reactive workflows are sometimes used in HFC networks: faults are identified by static threshold alarms, customer complaints, and SNMP polling. This can lead to long MTTR, duplicated truck rolls, and unnecessary escalation paths.

Objective: To autonomously correct faults once they are detected and diagnosed by the AI system.

System Functionality: Upon identifying the root cause, the remediation engine selects corrective actions from a library of domain-constrained, policy-aware interventions. These actions are based on fault type, device location, fault severity, and operational constraints.

With AI-driven autonomous systems, faults can be detected, diagnosed, and resolved autonomously in near real-time thanks to several key advantages:

- Nuanced performance changes can be captured by telemetry monitoring, before service impact is reported.
- Machine learning models such as clustering, anomaly detection, and time-series analysis can help identify emerging faults with higher precision than rule-based logic.
- The root cause inference engine correlates symptoms across layers and localizes the issue to specific devices, mitigating guesswork during diagnostics.
- Automated policy-driven remediation engines can apply corrective actions, verify effectiveness through feedback loops, and roll back, if necessary, without requiring human intervention.
- To drastically reduce field time, an AI system provides pre-diagnosed insights, and fault location estimates to guide technicians.

Common corrective actions include but are not limited to:

- Remote gain/slope adjustments on amplifiers
- Upstream channel migration
- Modem profile resets
- Transmit power tuning
- Service group realignment
- Dynamic spectrum reallocation
- Device reboot or remote reset
- Preemptive maintenance flagging

All actions are subject to policy controls and risk evaluation. The system verifies post-action metrics to confirm remediation effectiveness, and in the case of failure, initiates rollback or flags for human intervention.

Expected Outcome: Faults are resolved in minutes instead of hours, often without technician involvement. This enables high autonomous remediation rates and significantly reduces mean time to repair (MTTR), particularly for intermittent or transient faults that traditionally require repeated field visits.

6.6. Data-Driven Design Approach

Challenge: As highlighted in [1], traditional HFC network design relies on worst-case assumptions, aiming to ensure signal delivery to the farthest or most lossy customer locations. However, this method often leads to over-compensation, underperformance, and operational inefficiency.

Objective: enhance HFC network design using telemetry data to design adaptive profiles that reduce RF imbalance and improve network inefficiencies.

System Functionality: For instance, unsupervised clustering techniques can group cable modems into segments and proposed adaptive gain profiles for the amplifiers feeding them matched to the statistical characteristics of each group. This case supports the shift toward data-driven amplifier configuration, where real-time field data, instead of fixed specs, guides design. The approach lays the foundation for closed-loop, SDN-enabled amplifier control where cascade depth, topology, and performance targets dynamically inform active settings in real time.

Expected Outcome: The adoption of telemetry-informed, adaptive design can result in measurable improvements in signal consistency, energy efficiency, and operational reliability. By aligning amplifier configurations with actual network conditions rather than worst-case assumptions, operators observed:

- Improvement in average MER across the network segments.
- Reduction in amplifier energy use due to optimized gain distribution.
- Lower rate of post-installation technician visits caused by RF imbalance.
- The capability to provide operators with virtual twin models that simulate real-world plant behavior, allowing them to simulate network upgrades, test gain strategies, and evaluate performance impacts in advance.
- Enhanced cross-functional visibility, enabling Wi-Fi, mobile, and broadband engineering teams to evaluate how their services interact over shared HFC infrastructure. It facilitates better design, troubleshooting, and technology coexistence decisions.

7. Future work

Looking ahead, the promising results observed in initial implementations set the stage for more comprehensive deployments and enhanced capabilities:

- **Advanced Model Development:** Leverage the observed telemetry data patterns to train advanced machine learning models such as LSTM networks and Isolation Forests. These models will further refine anomaly detection, improve predictive accuracy, and enhance the system's ability to proactively address potential faults.

- **Operational Validation in Live Networks:** Conduct integrated case studies within live network environments to rigorously test and validate the framework’s real-world performance and reliability. This will include extensive scenario testing to evaluate the system’s capability to autonomously remediate faults under diverse and dynamic network conditions.

- **Expanded Legacy Integration:** Address integration complexities and extend compatibility to encompass additional legacy HFC components lacking direct telemetry capabilities. Efforts will focus on developing robust inference models and deploying strategic sensor arrays to ensure comprehensive monitoring coverage.

- **Standards Alignment:** Actively engage with CableLabs standards and working groups to further enhance and accelerate the development and deployment of the AI-driven autonomous framework, ensuring ongoing compatibility and adoption across industry practices.

These steps will solidify the practical applicability of the AI-driven autonomous framework and further its evolution towards fully autonomous HFC network management.

8. Challenges and Limitations:

8.1. Regulatory Considerations in Service Remediation

AI-driven autonomous framework introduces regulatory complexities requiring strict adherence to service and privacy laws. Emergency Alert System (EAS) requirements restrict automated actions like channel switching during alerts [6]. Signal leakage regulations enforce FCC limits [5], requiring special handling during amplifier and passive device adjustments.

Privacy laws limit telemetry collection from subscriber devices, impacting diagnostics. These were addressed using anonymization, compliance checks prior to execution, and regulation-aware decision-making. All orchestration decisions are constrained by FCC guidelines on signal levels and customer communication policies, necessitating policy-aware frameworks that blend autonomy with legal compliance.

8.2. Legacy Equipment Integration Challenges

Legacy HFC equipment lacks telemetry and modern APIs, creating monitoring blind spots and limited automation capabilities. Older CMTS platforms needed proxy agents and model tuning to work with noisy signal data. Heterogeneous infrastructure increased diagnostic complexity, requiring larger training datasets and more advanced algorithms.

Organizations might find success by retrofitting key nodes, employing inference to estimate the states of otherwise unmanageable devices, and strategically placing additional sensors to fill in observational gaps.

8.3. Data Quality and Availability Constraints

Sensor drift and intermittent connectivity create false baselines and data gaps, disrupting analytics and triggering incorrect remediations. Unstructured data from various vendors requires normalization, while poor historical labeling hampers supervised learning.

To counter this, organizations can deploy data validation pipelines, uncertainty-aware analytics, and semi-supervised learning to extract value from limited labeled data. Initial rollouts could be focused on areas with strong data quality, building the case for improving monitoring elsewhere.

8.4. Security Considerations

AI-driven RF control must be secure. Digital policy signing and anomaly detection on control flows are used to ensure only authenticated and validated actions are executed within the network.

8.5. Service Prioritization and Ethical Considerations

Automated prioritization introduces ethical and operational complexities. Prioritization decisions may be influenced by specific service-level agreements (SLAs) or perceived impact, which can inadvertently delay resolution for other customer segments. To maintain trust, operators should ensure transparent communication about prioritization criteria and strive for equitable service across all customer groups.

Frameworks should incorporate multifactor prioritization, balancing severity, customer SLA, historical behavior, and geographic equity. Human-in-the-loop controls and explainability dashboards are critical to maintain accountability and trust in automation.

Table 2 summarizes key challenges associated with AI-driven autonomous network implementation, along with potential mitigation strategies and the relative complexity of their implementation.

Table 2 - AI-driven Autonomous Network Implementation Challenges and Mitigation Strategies

Challenge Category	Specific Issues	Mitigation Strategies	Implementation Complexity
Regulatory Considerations	Emergency Alert System compliance	Pre-approval of remediation strategies with regulatory affairs	Medium
	Signal leakage management	Automated compliance verification before execution	High
	Subscriber privacy in telemetry	Data anonymization and aggregation techniques	Medium
Legacy Equipment	Limited telemetry capabilities	Strategic sensor deployment and proxy monitoring	High
	Proprietary management interfaces	Protocol adapter development and reverse engineering	High
	Heterogeneous equipment base	Equipment-specific diagnostic models	Medium

Challenge Category	Specific Issues	Mitigation Strategies	Implementation Complexity
Data Quality	Sensor drift and calibration	Automated drift detection and compensation	Medium
	Intermittent monitoring connectivity	Edge analytics with store-and-forward capabilities	Medium
	Unlabeled historical incidents	Semi-supervised learning approaches	High
Service Prioritization	Balancing business vs. residential	Multi-factor prioritization frameworks	Medium
	Geographic service equity	Fairness auditing in remediation patterns	Medium
	Transparency in prioritization	Clear communication of service expectations	Low

9. Conclusion

As HFC networks evolve and support increasingly complex and converged services, real-time intelligence and adaptive control have never been more crucial. The proposed framework serves as a strategic roadmap for AI-powered automation adoption in HFC networks. By integrating real-time telemetry, intelligent fault detection, and closed-loop remediation, it provides operators with a scalable reference architecture to enhance network resilience and service assurance.

This framework enables proactive network optimization by predicting faults, continually tuning performance based on live conditions, and guiding decisions with data. It reduces operational risk and downtime while enhancing agility, efficiency, and confidence in multi-service convergence.

10. Abbreviations

AI	artificial intelligence
ALSC	automatic level and slope control
API	application programming interface
BLE	Bluetooth Low Energy
CCAP	converged cable access platform
CM	cable modem
CMTS	cable modem termination system
CRM	customer relationship management
dB	decibel
DMPR	detect, mitigate, prioritize, repair
DOCSIS	Data Over Cable Service Interface Specification
DSP	digital signal processing
EAS	emergency alert system
FCC	Federal Communications Commission
HFC	hybrid fiber-coaxial
ID	identifier
IP	Internet Protocol
KPI	key performance indicator
LSTM	Long short-term memory
MAC	media access control
MAPE-K	monitor, analyze, plan, execute, knowledge
MTTD	mean time to detect
MTTR	mean time to repair
MER	modulation error ratio
MIB	management information base
ML	machine learning
NLP	natural language processing
NOC	network operations center
PHY	physical
PNM	proactive network maintenance
RF	radio frequency
RPD	remote PHY device
Rx	received
SCTE	Society of Cable Telecommunications Engineers
SLA	service level agreement
SNMP	Simple Network Management Protocol
TDR	time domain reflectometry
TFTP	trivial file transfer protocol
TR-069	Technical Report 069
Tx	transmit

11. Bibliography and References

[1] Linton, D., Sandino, E., Foroughi, N. Exploring the Benefits of Network Intelligence Applications to Optimize HFC Networks Using a Data Driven Design Approach. SCTE Expo 2023.xpo 2023.Available: https://wagtail-prod-storage.s3.amazonaws.com/documents/3585_Linton_5314_paper.pdf

[2] AIT365, "How AI-Driven Predictive Network Analytics is Redefining Tech Resilience," AIT365, [Online]. Available: https://aitech365.com/automation-in-ai/predictive-analytics/how-ai-driven-predictive-network-analytics-is-redefining-tech-resilience/. [Accessed: May 2024].

[3] CableLabs, "Decoding DOCSIS 3.1: What it Means for Cable Networks," Network World, [Online]. Available: https://www.networkworld.com/article/964534/decoding-docsis-3-1.html. [Accessed: May 2024].

[4] National Cable & Telecommunications Association (NCTA), "Making Rational HFC Upstream Migration Decisions in the Midst of Chaos," NCTA Technical Papers, 2013. [Online]. Available: https://www.nctatechnicalpapers.com/Paper/2013/2013-making-rational-hfc-upstream-migration-decisions-in-the-midst-of-chaos. [Accessed: May 2024].

[5] Federal Communications Commission (FCC), "Signal Leakage Standards for Cable Television," FCC Regulations, [Online]. Available: https://www.fcc.gov/document/cable-signal-leakage-standards. [Accessed: May 2024].

[6] Electronic Code of Federal Regulations (eCFR), "Emergency Alert System (EAS) Requirements," eCFR, [Online]. Available: https://www.ecfr.gov/current/title-47/chapter-I/subchapter-A/part-11. [Accessed: May 2024].

[7] Nanites AI, "The High Cost of Network Downtime: How Agentic AI Reduces MTTR," Nanites AI, [Online]. Available: https://www.nanites.ai/post/the-high-cost-of-network-downtime-how-agentic-ai-reduces-mttr. [Accessed: May 2024].

[8] S. R. White, J. E. Hanson, et al., "An Architectural Approach to Autonomic Computing," IEEE, 01 Jun. 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1301340. [Accessed: May 2024].

[9] M. B. da Silva, C. Bezerra, et al., "A Catalog of Performance Measures for Self-Adaptive Systems," ACM Digital Library, 14 Dec. 2021. [Online]. Available: https://dl.acm.org/doi/10.1145/3493244.3493259. [Accessed: May 2024].

Click here to navigate back to Table of Contents

Preventing Service Disruptions with a Network Digital Twin

Cutover Validation in Large Optical Transport Networks

A Technical Paper prepared for SCTE by

Nehuen Gonzalez-Montoro, Professor, Universidad Nacional de Córdoba
Av. Velez Sarsfield 1611
Córdoba, Argentina
nehuen.gonzalez@unc.edu.ar
+54 351 5353800 int. 64

Renato Cherini, Professor, Universidad Nacional de Córdoba
Av. Velez Sarsfield 1611
Córdoba, Argentina
renato.cherini@unc.edu.ar
+54 351 5353800 int. 64

Luis Compagnucci, Head of Network Solutions, Ceragon Networks
3801 E Plano Parkway
Plano, Texas 75074, USA
luisc@ceragon.com
+1 407 496 0315

Jorge M. Finochietto, Professor, Universidad Nacional de Córdoba
Av. Velez Sarsfield 1611
Córdoba, Argentina
jorge.finochietto@unc.edu.ar
+54 351 5353800 int. 64

Table of Contents

List of Figures

1. Introduction

Large-scale network operation has become a more demanding task over the recent years. The increasing reach and heterogeneity of these networks, commonly involving the integration of multi-vendor systems, and the wide variety of services they support, boost the complexity of planning, operations, and maintenance tasks [1]. Over the past decade, substantial efforts have been made to move away from traditional approaches that rely on operator expertise, specific models, and vendor-specific tools [2], toward advanced, flexible, standardized, and closed-loop management and control strategies [3], and toward software-defined networks (SDN) [4], recently framed within the network digital twin (NDT) paradigm [5].

An NDT dynamically maintains a digital replica of the physical network, enabling real-time monitoring and decision-making. As a result, operators can leverage digital twins to address the challenges of network design, operation, and maintenance. This approach supports more comprehensive, integrated, automatable, and resilient practices across all stages of the network lifecycle. It also facilitates optimization, what-if analysis, troubleshooting, and impact assessment. Companies such as Ceragon [6], Nokia [7], Viavi [8], Huawei [9], and Forward Networks [10] offer commercial products aligned with the NDT paradigm.

In our previous works [11], [12], we presented a comprehensive overview of building operator-centric digital twins for optical transport networks (DTON), and analyzed in detail a soft failure detection case study. In this article, we further investigate how a DTON can address the challenges posed by cutover validation. We explore how a NDT can be leveraged to validate and approve planned network cutovers, ensuring that maintenance actions neither affect paths within the same protection group nor impact the same circuit multiple times within a maintenance window. In particular, a digital replica enables the integration of multiple network layers, from fiber ducts to service paths, which can be used to determine potential impacts.

This problem has previously been studied as the task of finding a schedule that maximizes total flow over time [13], and as the challenge of minimizing outages while scheduling all maintenance tasks [14]. In [15], the authors studied the task scheduling problem in optical networks focusing on finding algorithms to solve it. In this work, we present a real-world case study where this capability was developed and deployed across a large-scale optical network, using an operator-centric digital twin model enriched with geospatial, relational, and historical data. We apply integer linear programming (ILP) to solve the maintenance task scheduling problem and discuss the associated implementation challenges. Although our approach relied solely on ILP formulations, which performed very well at the scale of the deployed network, we plan to explore algorithmic, heuristic, and even artificial intelligence–based alternatives in future work.

We adopt the definition of a network digital twin consistent with ITU-T recommendations [16] [17] and prevailing industry literature: a persistent, digital replica of a network that is continuously synchronized with its physical counterpart, capable of simulating, analyzing, and predicting network behavior under various scenarios. While an NDT can support all three of these capabilities, the presented use case focuses primarily on the analysis. More specifically, evaluating the impact of planned maintenance cutovers on network services and protection schemes. Simulation and prediction remain part of the broader NDT framework and could be integrated into this workflow in future developments.

By integrating these kinds of scheduling capabilities, a NDT empowers operators to evaluate planned changes, proactively flag potential risks, and make informed decisions. This results in enhanced network stability and a significant reduction in service outages caused by misaligned maintenance.

The remainder of this paper is structured as follows. Section 2 introduces the principal elements required to model DTON and its implementation. Section 3 analyzes a real-world application focused on planned cutover validation and approval. Section 4 discusses future perspectives, and section 5 summarizes the key findings of our work.

2. Digital Twin for Optical Transport Network

One of the main challenges in designing a NDT is developing a data model that accurately represents the network with sufficient detail while abstracting irrelevant particulars. Such a model should provide enough detail to adequately document the network yet offer sufficient abstraction to facilitate effective problem solving. Moreover, the data model significantly influences the types of problems that can be addressed and solved using the NDT. An overly abstract model may result in poor network representation and solutions that fail to meet operators' actual needs. Conversely, excessive detail can limit the scope of solvable problems. For example, highly detailed models can render the acquisition and maintenance of necessary information impractical, while computational costs may also become prohibitive in some scenarios. In general terms, the data model can be conceptualized as a large temporal multigraph. Addressing most of the common problems requires sophisticated algorithms capable of leveraging multiple, often recursive relationships and extensive historical data. Consequently, adopting computationally efficient data representations is essential to ensure these problems remain tractable.

A second issue arises from the fact that the assumption of complete data availability is rarely satisfied in the practice. Incompleteness can arise from various factors, such as the lack of integration with equipment from specific vendors, the absence of critical information that must be gathered manually, or the inherent complexity of modeling network segments not fully owned or managed by the operator. As a result, some parts of the network are richly documented, while others remain sparsely described. Even though this situation could be understood as a real-world limitation that could make the implementation of NDT at optical networks unfeasible, designing a flexible and adaptable data model can help to overcome this limitation. Moreover, the presence of incomplete data does not reduce the value of the digital twin. On the contrary, it underscores the need for a flexible and adaptive data model, one that can accommodate varying levels of detail and still extract meaningful insights from the available information. This adaptability is key to ensuring the digital twin remains a useful tool for network simulation, analysis, and management, even in the face of persistent data gaps. Using such a data model can cope with ideal scenarios where the full network data is available or collectable, but also with real-world scenarios where some data might be missing or not collectable at some point. For example, in a multi-layer representation of a network, a rigid data model must rely on the underlying layers to compute interdependence of two network links at a top layer. Therefore a data gap in a bottom layer could greatly limit the capabilities of the NDT. Instead, a flexible and adaptable data model can accept abstract relationships between top-level entities that temporarily replace dependence computation until the bottom layers data is retrieved, enabling the application of NDT in this kind of situation.

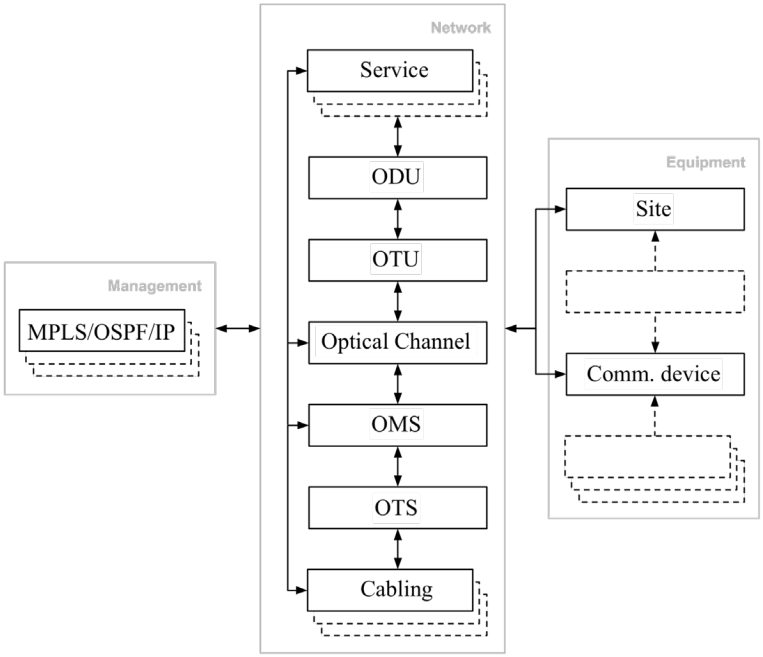


Figure 1 - Schematic Structure of the DTON Data Model

Regarding the first challenge, determining the optimal level of abstraction in optical networks is particularly challenging due to the numerous physical, logical, and management entities involved. A common starting point is the ITU recommendations for Optical Transport Networks (OTN) [18], [19], which offer a hierarchical layer model ranging from optical sections at the bottom optical domain layer to fully digital domain entities such as Optical Data Units (ODUs). Each layer acts as a transport layer for the one above it, with relationships typically involving parallel and/or series multiplexing schemes. The use of this standard presents an advantage because operators' networks are primarily structured around entities that can be mapped to it. Nevertheless, additional requirements remain unaddressed, necessitating extensions to the model to include elements like cabling, communication devices, and their physical layouts, as well as services delivered to end users and their corresponding management. This also includes interactions with management system components from multiple vendors, end-to-end circuit configurations, protection schemes, and related aspects.

Figure 1 presents a simplified schematic of the data model employed. Due to confidentiality constraints, further details cannot be disclosed. The diagram centers on the OTN model layers and the aggregation relationships among them, primarily capturing multiplexing structures. Below these layers are additional ones illustrating the geolocated physical implementation of Optical Transmission Section (OTS) entities, such as fiber optic cables and ducts, along with co-location and geographic proximity. Above, we incorporate layers mapping customers' end-to-end circuits, and redundancy relationships between them. Given that operators often lack certain information, additional relationships bypassing intermediate layers have been implemented. On the figure's left side, administrative and organizational management entities are shown, typically including Muliprotocol Label Switching (MPLS) domains, Open Shortest Path First (OSPF) areas, and subnet divisions, which define routing scopes, fault isolation boundaries, and policy enforcement zones.

Every entity representing a communication link, at any layer, is delimited by sites. This structure aligns perfectly with a graph representation, facilitating the integration of connectivity and routing algorithms. However, communication links depend on the availability of physical communication equipment. Therefore, the data model includes entities for actual devices, cards, ports, and so forth. Additionally, we also model their organization within sites and the occupancy relationships with communication links. This illustrates the coexistence of multiple redundant relationships—ranging from abstract to concrete—for different purposes.

As we will see in the following section, modeling the interrelationships between network entities across different layers is essential to accurately assess the impact of planned maintenance activities, which often affect entities that are not only indirectly, but also distantly related through long chains of intermediate dependencies.

While not directly tied to the approach proposed in this work, the entities comprising the NDT exhibit additional dimensions—beyond their structural organization—that merit consideration. The temporal (historical) dimension is especially relevant for operational and performance data of hardware components. Capturing and storing time-series metrics enables both near-real-time monitoring of network state and the addressing of problems relying on the temporal evolution of specific parameters. For instance, tracking historical optical performance indicators, such as attenuation, is instrumental in detecting soft failures along optical spans. Another dimension concerns representing the current network state and its temporal variations. Maintaining records of active, decommissioned, and planned entities facilitates analysis of hypothetical scenarios and testing of operational strategies. Incorporating these dimensions into the data model enables the forward-looking perspectives outlined in Section 4.

3. Case Study: Cutover Validation

Cutover validation refers to the problem of verifying that a planned network change—such as the activation of a new optical path, the migration of services, or a topology reconfiguration—has been executed successfully and without unintended impact. The challenge arises from the complexity of modern networks and the potential for indirect dependencies between entities. Effective cutover validation is critical to ensure service continuity, minimize downtime, and maintain the integrity of the network after a transition. In this context, the design of a working plan becomes critical to ensure that maintenance activities neither disrupt paths within the same protection group nor result in multiple impacts on the same circuit during a single maintenance window.

This key point can be illustrated better with an example. Consider the simple network shown in Figure 2, which consists of five ROADM nodes (*s1*, *s2*, *s3*, *s4*, and *s5*) connected by fiber optic spans made up of multiple Optical Transmission Section (OTS) segments at the bottom layer. Above these Optical Multiplex Sections (OMS) segments lies the lightpath topology, comprising three optical channels connecting *s1* to *s4*, *s1* to *s5*, and *s4* to *s5*. The network supports only two services (*service 1* and *service 2*), each with one working path and one protection path. These electrical paths are carried over Optical Channel Transport Units (OTUs) shown in the third layer (for simplicity, the ODU layer is omitted). *Service 1*'s working path uses *OTU 1* exclusively, while its protection path is formed by *OTU 2* and *OTU 3*. *Service 2* follows a similar configuration.

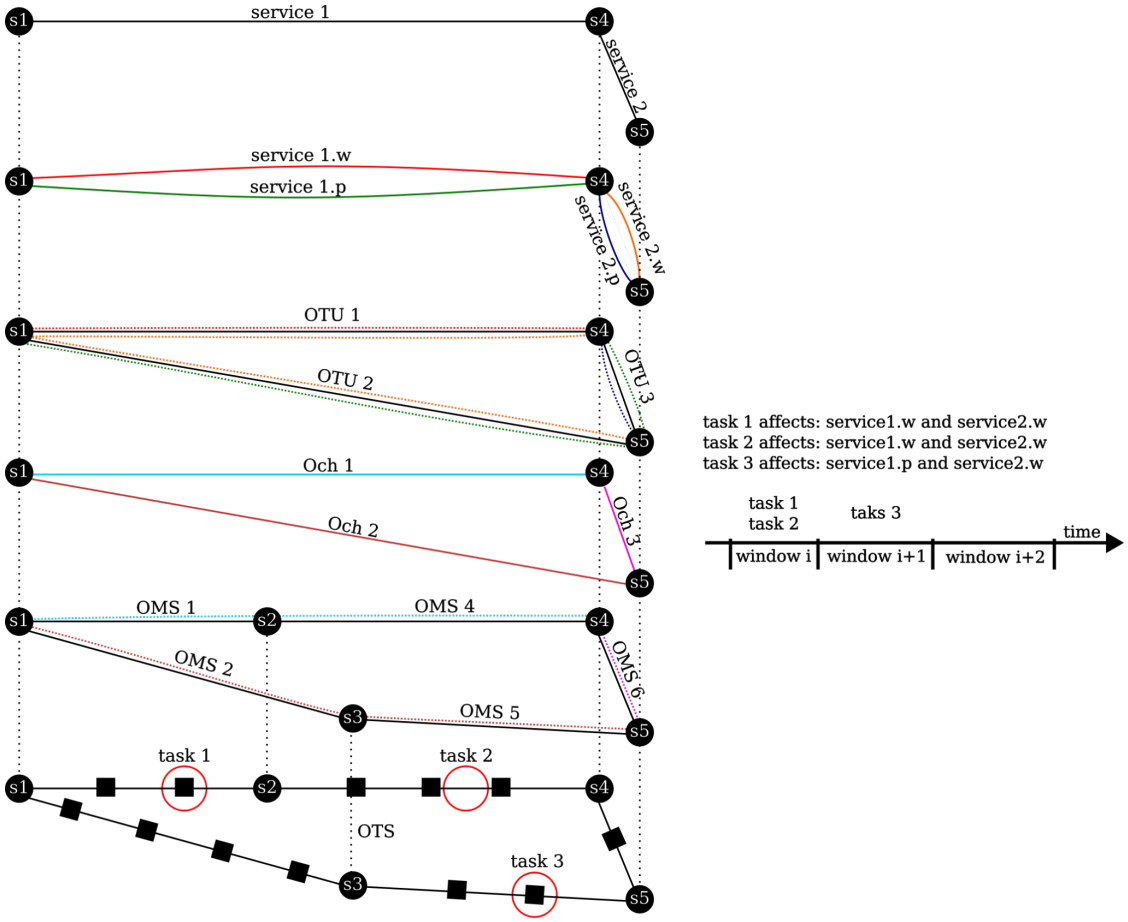


Figure 2 - Simple Task Scheduling Example

Up to this point, we have only shown a multilayer description of a simple network. So, what happens when maintenance tasks need to be performed? In this example, we assume there are two optical amplifiers and one optical section requiring maintenance: *task 1* targets an amplifier located at *OMS 1*, *task 2* targets a fiber section at *OMS 4*, and *task 3* targets an amplifier at *OMS 5*. If all three tasks are executed simultaneously, or even partially overlap, *service 1* will experience downtime. This raises the question: is it possible to schedule these three tasks in a way that avoids any service disruption?

In order to find such a schedule, maintenance windows are defined, and each task must be assigned to one of them. In this case, a valid solution is straightforward: *task 1* and *task 2* can be scheduled in the same window, as they affect exactly the same working and protection paths. *Task 3*, however, must be placed in a separate window to prevent service disruption. A possible solution, as shown on the right side of the figure, is to schedule *task 1* and *task 2* in the first available window, and *task 3* in the next.

It is worth noting that, although this example considers protection at the service level, the same approach applies when protection is implemented at the optical layer (i.e., working and protection lightpaths). The specific query used to identify dependencies in the NDT data model may differ, but the core problem remains unchanged: finding a schedule that avoids simultaneous impact on both the working and protection path or lightpath of a given service or OTU. While this example focuses on maintenance tasks involving amplifiers or fibers, the problem formulation accommodates any type of maintenance activity. This

flexibility is only achievable with a well-designed and adaptable data model, as presented in the previous section.

3.1. Integer Linear Programming

Integer Linear Programming (ILP) serves both as a formalism for describing optimization problems and as a method for solving them. It involves optimizing a linear objective function subject to a set of linear equality and inequality constraints, where all variables are restricted to integer values. These constraints define a feasible solution space, and the goal is to find the solution that maximizes or minimizes the objective function.

ILP is a powerful tool for tackling optimization problems. However, since solving ILP instances is itself NP-hard, scalability can become an issue. Over the years, a wide range of algorithms and heuristics have been developed to address this, and there are several open-source and commercial solvers available. Some commonly used ones include SCIP [20], HiGHS [21], CBC [22], IBM ILOG CPLEX [23], and GUROBI [24]. In addition, many libraries simplify the definition and instantiation of ILP models in various programming languages. Thanks to the optimizations implemented in these solvers, ILP models remain applicable in many scenarios. Performance generally depends on the number of variables involved, as well as the size of the feasible solution space.

In the following subsections, we present two ILP models for solving the task scheduling problem. These models not only provide a formal definition of the problem but also offer a practical way to compute solutions. We begin by defining the variables and coefficients, followed by the two model formulations.

3.2. Problem Definition

Leveraging the NDT and its efficiently implemented data model, the impact of planned maintenance tasks on paths can be determined through straightforward queries. For example, it is possible to directly identify all working and protection paths affected when a specific set of underlying resources is scheduled for maintenance.

For instance, if there are N required tasks and M services, each with a defined working and protection path, then we can determine whether the j-th working or protection path will be affected by the i-th task. For modeling purposes, this information can be encoded as a set of coefficients:

$$W_{ij} = \begin{cases} 1 & \text{iff task i affects at least one resource serving the working path of service j} \\ 0 & \text{otherwise} \end{cases}$$

$$P_{ij} = \begin{cases} 1 & \text{iff task i affects at least one resource serving the protection path of service j} \\ 0 & \text{otherwise} \end{cases}$$

$$with \quad \imath \in \{1, 2, ..., N\} \quad and \quad j \in \{1, 2, ..., M\}$$

The task scheduling problem can be addressed in multiple ways, depending on how its dimensions are represented, for example, task duration, availability of technical personnel, or material stock. In our case, we adopted two simple approaches based on slotting time in maintenance windows, under the assumption that both technicians and materials are always available. This representation fits well with our application

scenario, as the availability of stock and maintenance personnel can be verified beforehand. Consequently, the scheduling process only needs to account for task interdependencies.

We solved the task scheduling problem with the following two approaches: in one hand scheduling all tasks across multiple windows and, on the other hand, maximizing usage of the next maintenance window. In the first case, the aim is to minimize the total number of windows needed to complete all planned tasks. In the second, the objective is to fit as many tasks as possible into the next available window. When the full set of maintenance tasks is known in advance, the multi-window scheduling approach is generally preferred, as it can yield shorter overall completion times. However, in more dynamic scenarios—where new tasks may arise at each maintenance window—it may be more practical to focus on one window at a time, prioritizing maximum task execution within it.

Given *N* required tasks, *M* services and *K* maintenance windows (with $K \leq N$), we can define the following variables:

$$x_{ik} = \begin{cases} 1 & \text{iff task i is scheduled for the window k} \\ 0 & \text{otherwise} \end{cases}$$

$$x_k = \begin{cases} 1 & \text{iff window k has at least one task scheduled} \\ 0 & \text{otherwise} \end{cases}$$

$$w_{jk} = \begin{cases} 1 & \text{iff the working path of service j is affected in window k} \\ 0 & \text{otherwise} \end{cases}$$

$$p_{jk} = \begin{cases} 1 & \text{iff the protection path of service j is affected in window k} \\ 0 & \text{otherwise} \end{cases}$$

3.2.1. Task Scheduling Problem

The first approach can be formulated as follows: given a set of required maintenance tasks, schedule all tasks while minimizing the number of maintenance windows needed, ensuring that no service experiences disruption. In terms of ILP, it can be stated as:

Minimize:

$$z = \sum_{k=1}^{N} k x_k \qquad (1)$$

Subject to:

$$\sum_{k=1}^{N} x_{ik} = 1 \qquad \forall i \in \{1, 2, ..., N\} \qquad (2)$$

$$x_k \geq x_{ik} \qquad \forall k \in \{1, 2, ..., N\} \quad \forall i \in \{1, 2, ..., N\} \qquad (3)$$

$$W_{ij}x_{ik} \leq w_{jk} \qquad \forall i \in \{1,2,..,N\}, \forall j \in \{1,2,...,M\}, \forall k \in \{1,2,...,N\} \qquad (4)$$

$$P_{ij}x_{ik} \leq p_{jk} \qquad \forall i \in \{1,2,...,N\}, \forall j \in \{1,2,...,M\}, \forall k \in \{1,2,...,N\} \qquad (5)$$

$$w_{jk} + p_{jk} \leq 1 \qquad \forall k \in \{1,2,...,N\}, \forall j \in \{1,2,...,M\} \qquad (6)$$

Equation 1 defines the objective: minimizing the number of maintenance windows required. Since x_k indicates whether window *k* is used, the goal is to minimize the sum of all these variables. The sum is weighted by the position *k* to prioritize earlier windows over later ones. Equation 2 ensures that each task is assigned to exactly one maintenance window. Equation 3 sets a window's status to "used" whenever at least one task is scheduled within it. Equations 4 and 5 indicate whether the working or protection path of service *j* is affected by any task scheduled in maintenance window *k*. Finally, Equation 6 prevents both the working and protection paths of a service from being affected simultaneously in the same maintenance window.

3.2.2. Next Window Maximum Usage Problem

The second approach can be described as follows: given a set of required maintenance tasks, schedule the maximum possible number within the next maintenance window, ensuring no service disruption. Since only one window is considered, *K* equals 1, and the following ILP formulation is sufficient:

Maximize:

$$\sum_{i=1}^{N} C_i x_{i1} \qquad (7)$$

Subject to:

$$W_{ij}x_{i1} \leq w_{j1} \qquad \forall i \in \{1,2,...,N\}, \forall j \in \{1,2,...,M\} \qquad (8)$$

$$P_{ij}x_{i1} \leq p_{j1} \qquad \forall i \in \{1,2,...,N\}, \forall j \in \{1,2,...,M\} \qquad (9)$$

$$w_{j1} + p_{j1} \leq 1 \qquad \forall j \in \{1,2,...,M\} \qquad (10)$$

The objective now is to maximize the number of tasks scheduled within the next maintenance window. The coefficients C_i in Equation 7 represent task priorities; if all coefficients are equal, all tasks are treated with the same priority. Equations 8 and 9 indicate whether the working and/or protection paths of service *j* are affected by any selected task. Finally, Equation 10 restricts the solution to prevent simultaneous impact on both the working and protection paths of a service.

It is worth noting that the window duration can be adjusted according to the operator's needs. Although in our application scenario a fixed window duration is sufficient, the models can be easily extended to allow multiple slot allocations per task, providing a more flexible and efficient representation. However, we discarded this option because task duration is highly unpredictable; consequently, maintenance planners within the operators prefer to allocate an entire maintenance window to each task.

The first model ensures that all tasks are scheduled, but in the second case a task can remain unscheduled for multiple windows and potentially for ever. This is an unwanted behaviour and we cope with it by computing the C_i priority coefficients after each assignment in order to prioritize tasks that have been unscheduled for more consecutive windows.

3.3. Application Scenario and Results

Although the previous example might suggest that this problem is straightforward, the scale of a real-world, nation-wide network makes finding solutions by intuition or simple tools very challenging. We applied the described models to a network with 1,100 optical nodes, 849 optical sections, and over 2,500 protected services. Prior to implementing the DTON equipped with the ILP models, the complexity had grown so much that maintenance management limited itself to scheduling only one task per maintenance window, which caused significant delays and prioritization problems. Although in some cases the scale of the problem can be reduced by solving it for specific sections of the network, obtaining an optimal solution without proper modeling becomes highly challenging. As a result, operators often prefer conservative, suboptimal solutions in order to minimize the risk of unexpected behavior.

Our initial approach relied on the task scheduling model, which produced very good results. However, as new tasks continued to emerge before previously scheduled ones were completed, the solution quickly became outdated. Currently, both models are available, and their use depends on the task flow. For instance, the second model is preferred during periods of high task volume, whereas the first is applied when the task flow is lower and a global optimal solution can be achieved.

Integrating these models into the DTON has provided maintenance managers with a reliable tool, enabling better, data-driven decisions that reduce risks, minimize service disruptions, and shorten the time required to complete all maintenance tasks.

4. Future Perspective

Following the definitions provided in the ITU-T recommendations [16][17] we cover the representation level and partially the optimization level of NDTs. Looking forward to covering all five levels, the use case presented in this article paves the way toward a fully automated maintenance planning ecosystem that encompasses failure and anomaly detection, required actions definition, task scheduling, and post-execution validation.

The data collection capabilities enabled by the DTON allow for the continuous monitoring of network elements, supporting the early detection and prediction of soft failures, degradations, and anomalies in both optical equipment and fiber infrastructure. In previous work [8, 9], we demonstrated a specific use case that validates this potential. Nonetheless, there remain numerous open challenges, both in research and in implementation, particularly around the integration of more sophisticated artificial intelligence techniques for reliable and scalable forecasting.

Transitioning from anomaly and failure prediction to automated task planning becomes feasible when the DTON also integrates contextual information about available resources, such as personnel, spares, and operational constraints. Once the required actions are defined, scheduling models, such as the ones presented in this work, can be used to optimize execution while minimizing service impact. The cycle is completed with a post-maintenance validation process, where the digital twin can be queried to assess whether the intervention effectively addressed the issue and improved network health.

Artificial intelligence can play a central role throughout this process. In terms of failure prediction, anomaly detection and degradation prediction can be achieved over fiber spans attenuation time series. A comprehensive survey of deep learning applied to time series anomaly detection can be found in [25]. Also, machine learning can enhance task planning and scheduling strategies. For example, ILP formulations provide a valuable ground truth that can be used to generate real or synthetic datasets representing scheduling scenarios in complex networks. These datasets could be leveraged to train machine learning models capable of approximating optimal scheduling decisions with significantly reduced computational cost. In large-scale networks where solving ILP models becomes intractable in real time, such models could offer high-quality heuristic alternatives.

Moreover, reinforcement learning could be explored for adaptive scheduling in dynamic environments, where tasks and network states evolve over time. In this case, one or more heuristic can be learned in order to schedule tasks in a per window fashion. The use of reinforcement learning in generic task scheduling problems has been widely studied. A comprehensive review of related works can be found in [26]. Given the high availability requirements of communication networks, learning-based approaches can and should be complemented with feasibility checks and optimization models in order to guarantee no service affectation. Such integration would enable hybrid systems that balance precision and scalability, ultimately leading to more resilient and autonomous network operations.

In summary, this work lays the baseline for future developments that can couple the rigorous modeling capabilities of digital twins with the predictive and adaptive strengths of artificial intelligence, aiming to build a proactive, efficient, and fully closed-loop maintenance framework.

5. Conclusion

In this work, we presented a practical application of the DTON in the context of maintenance task scheduling and validation. By integrating a flexible multilayer data model with ILP optimization techniques, we demonstrated how a digital twin can be used to systematically plan maintenance windows while avoiding service disruptions.

Although the scheduling problem could be approached using a static representation of network routing, our implementation operates on a continuously updated, multi-layer digital model of the actual network, integrating physical topology, logical paths, protection relationships, and service configurations. This dynamic aspect, together with its ability to ingest live operational data and historical trends, aligns with key characteristics of a NDT as defined in ITU-T and industry practice. The scheduling models are not standalone algorithms but are embedded within this dynamic environment, allowing them to reflect the current state of the network and assess maintenance risks in near-real time. This distinction is critical: it transforms the solution from a one-time network map into an ongoing operational tool that evolves alongside the physical network.

This case study showed that leveraging a digital twin not only enables impact analysis of maintenance activities but also supports optimal scheduling decisions. Decisions that would be nearly impossible to reach by intuition alone due to network complexity. The combined use of the two scheduling models, minimizing the total number of windows versus maximizing task execution within the next window, proved to be an effective and adaptable strategy in real-world, dynamic operational environments.

Moreover, the implementation of these methods significantly improved the efficiency of maintenance planning, reducing risk and delays, and enabling operators to carry out more tasks in less time without compromising service availability.

Looking ahead, this work lays the groundwork for a broader and more integrated maintenance ecosystem. Future developments may include predictive analytics, AI-assisted planning, and fully automated validation workflows, reinforcing the central role of digital twins in modern network operations.

6. Abbreviations

NDT	network digital twin
DTON	digital twin for optical networks
OTN	optical transport networks
ODU	optical data unit
OTS	optical transmission section
OMS	optical multiplex section
OTU	optical channel transport unit
Och	optical channel
ILP	integer linear programming

7. Bibliography and References

1. A. Martinez, M. Yannuzzi, V. López, D. López, W. Ramírez, R. Serral-Gracià, X. Masip-Bruin, M. Maciejewski, and J. Altmann, “Network management challenges and trends in multi-layer and multi-vendor settings for carrier-grade networks”, IEEE Commun. Surv. & Tutorials 16, 2207–2230 (2014).

2. H. Kim and N. Feamster, “Improving network management with software defined networking”, IEEE Commun. Mag. 51, 114–119 (2013).

3. Andriolli, N., Giorgetti, A., Castoldi, P., Cecchetti, G., Cerutti, I., Sambo, N., ... & Paolucci, F. (2022). Optical networks management and control: A review and recent challenges. *Optical Switching and Networking*, *44*, 100652.

4. Nsafoa-Yeboah, K., Tchao, E. T., Yeboah-Akowuah, B., Kommey, B., Agbemenu, A. S., Keelson, E., & Monirujjaman Khan, M. (2022). Software-defined networks for optical networks using flexible orchestration: Advances, challenges, and opportunities. *Journal of Computer Networks and Communications*, *2022*(1), 5037702.

5. Wang, D., Song, Y., Zhang, Y., Jiang, X., Dong, J., Khan, F. N., ... & Zhang, M. (2024). Digital twin of optical networks: a review of recent advances and future trends. *Journal of Lightwave Technology*, *42*(12), 4233-4259.

6. Ceragon, “Ceragon network digital twin”, (2024). Accessed: 2024-12-13.

7. Nokia, “Network digital twin for manufacturing”, (2024). Accessed: 2024-12-13.

8. VIAVI Solutions, “Network digital twin”, (2024). Accessed: 2024-12-13.

9. C. Janz, Y. You, M. Hemmati, A. Javadtalab, Z. Jiang, H. Li, and F. Haoyu, “Network digital twins for optical networks”, Opt. Fiber Technol. 89, 104068 (2025).

10. Forward Networks, “Forward enterprise: Network digital twin solution”, (2024). Accessed: 2024-12-13.

11. R. Cherini, N. Gonzalez-Montoro, L. Compagnucci, and J. M. Finochietto, “Bottom-up design of digital twins for optical transport networks: Challenges and lessons learned [invited]”, in 2024 International Conference on Optical Network Design and Modeling (ONDM), pp. 1–6, (2024)

12. R. Cherini, N. Gonzalez-Montoro, A. Gonzalez-Montoro, L. Compagnucci, and J. M. Finochietto, “Building operator-centric digital twins for optical transport networks: from data models to real applications”, Journal of Optical Communications and Networking, 17(6), B1-B14, (2025).

13. N. Boland, T. Kalinowski, H. Waterer, and L. Zheng, “Scheduling arc maintenance jobs in a network to maximize total flow over time”, Discrete Applied Mathematics, 163, 34-52, (2014)

14. F. Abed, L. Chen, Y. Disser, M. Groß, N. Megow, J. Meißner, and R. Rischke, “Scheduling maintenance jobs in networks”, Theoretical Computer Science, 754, 107-121, (2019)

15. X. Qin, T. An, X. Gao, Y. Gong, Q. Hu, X. Huo, and B. Zhang, “Intelligent task scheduling for cutover in optical networks”, Optics Express, 33(13), 27334-27347, (2025)

16. International Telecommunication Union, “Y.3090: Digital twin network – Requirements and architecture,” Tech. Rep. Y.3090, ITU-T (2022).

17. International Telecommunication Union, “Y.3091: Digital twin network – Capability levels and evaluation methods,” Tech. Rep. Y.3091, ITU-T (2023).

18. International Telecommunication Union, “G.807: Generic Functional Architecture of the Optical Transport Network,” Tech. Rep. G.807, ITU-T (2020).

19. International Telecommunication Union, “G.872: Architecture of the Optical Transport Network,” Tech. Rep. G.872, ITU-T (2019).

20. S. Bolusani, M. Besançon, K. Bestuzheva, A. Chmiela, J. Dionísio, T. Donkiewicz, J. van Doornmalen, L. Eifler, M. Ghannam, A. Gleixner, C. Graczyk, K. Halbig, I. Hedtke, A. Hoen, C. Hojny, R. van der Hulst, D. Kamp, T. Koch, K. Kofler, J. Lentz, J. Manns, G. Mexi, E. Mühmer, M. E. Pfetsch, F. Schlösser, F. Serrano, Y. Shinano, M. Turner, S. Vigerske, D. Weninger, and L. Xu, “The SCIP Optimization Suite 9.0,” Available at https://www.scipopt.org/ February 2024

21. High performance software for linear optimization, https://github.com/ERGO-Code/HiGHS

22. COIN-OR Branch and Cut Solver, https://github.com/coin-or/Cbc

23. IBM ILOG CPLEX optimization solver https://www.ibm.com/es-es/products/ilog-cplex-optimization-studio

24. Gurobi optimization software https://www.gurobi.com/

25. Zamanzadeh Darban, Z., Webb, G. I., Pan, S., Aggarwal, C., & Salehi, M. (2024). Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, *57*(1), 1-42.

26. Shyalika, C., Silva, T., & Karunananda, A. (2020). Reinforcement learning in dynamic task scheduling: A review. *SN Computer Science*, *1*(6), 306.

Click here to navigate back to Table of Contents

PNM OFDM Channel Estimates for Drop Cable Analysis

Lab Experiments Using PNM, Project: “Echo Delta”

A Technical Paper prepared for SCTE by

Alexander (Shony) Podarevsky, CTO, Promptlink
4005 Avenida De La Plata
Oceanside, CA 92056
shony@promptlink.com
760-994-3153

Larry Wolcott, Engineering Fellow, Comcast
183 Inverness Drive West
Englewood, CO 80112
Larry_Wolcott@cable.comcast.com
303-726-1596

Denys Podarevsky, Lab Technician, Promptlink
Angelo Wolcott, Lab Technician, Independent
Nico Wolcott, Lab Technician, Independent

Editors:
Roger Fish, Ron Hranac, Tom Kolze, Jason Rupe, Brady Volpe

Table of Contents

List of Figures

List of Tables

1. Introduction

This experiment utilizes the proactive network maintenance (PNM) high-resolution orthogonal frequency-division multiplexing (OFDM) channel estimation capabilities of Data-Over-Cable Service Interface Specifications (DOCSIS®) 3.1 OFDM cable modems to detect, localize, and characterize impedance mismatches present in coaxial cable networks. When a signal travels along a coaxial cable and encounters a normal impedance discontinuity, such as a cable modem input, a connector, or an open cable, a portion of the signal energy is reflected back toward the source (Figure 1). These reflected RF signals are captured as discrete echoes in the time-domain impulse response, which is derived by applying an inverse fast Fourier transform (IFFT) to the complex frequency-domain coefficients reported by the cable modem. An example of this processing is described in [1] Tom Williams et al, "OFDMA Predistortion Coefficient and OFDM Channel Estimation Decoding and Analysis" and in [2] Larry Wolcott, et al, "OFDM and OFDMA Bring New, Laser-Guided Precision for Plant Fault Distancing." It has been demonstrated to provide highly accurate distance measurements to within a few meters or better, depending on the number of occupied subcarriers. This enables precise mapping of the cable network and rapid identification of changes, such as those caused by a cable cut or a disconnected modem.

These experiments will demonstrate some of the capabilities offered by PNM's OFDM channel estimates and expose the reader to important principles of operation for interpreting them in useful ways. These concepts will include:

- Demonstrate that drop cable lengths and cable modem connections can be accurately measured
- Identify when a drop cable has been cut and accurately report the distance
- Show that splices can be detected within the drop cables
- Provide length or distance information about tap spans
- Give insight about tap values, return loss and the importance port-to-port isolation

2. Reflection Dynamics and Triple-Transit Effects

In real-world cable networks, multiple impedance mismatches can exist, leading to complex reflection patterns. One important phenomenon is the triple-transit reflection. This occurs when a signal reflects off one discontinuity, travels back to another (such as the tap), and then reflects again, towards the receiver. The result is a secondary echo with a longer path delay and lower amplitude.

The time delay of each echo corresponds to the round-trip travel time of the impedance cavity (also called echo cavity or reflection cavity) created by two reflectors (impedance mismatches). The physical length of the echo cavity is calculated using a formula, described in [2] section 4.4. The formula requires the number of active subcarriers and velocity of propagation of the cable. For series-6 drop cables, a typical velocity factor (VF) is 0.85, which corresponds to velocity of propagation (VoP) 85% of the speed of light in vacuum.

For example, a signal might:

1. Reflect from an impedance mismatch at, say, a cable modem inside the subscriber premises,
2. Travel back through the drop cable to the tap and reflect again,
3. Then travel back through the drop cable to the first modem, appearing as an echo with a time delay equal to the round-trip propagation delay related to the drop cable's length and velocity factor. See Figure 1.

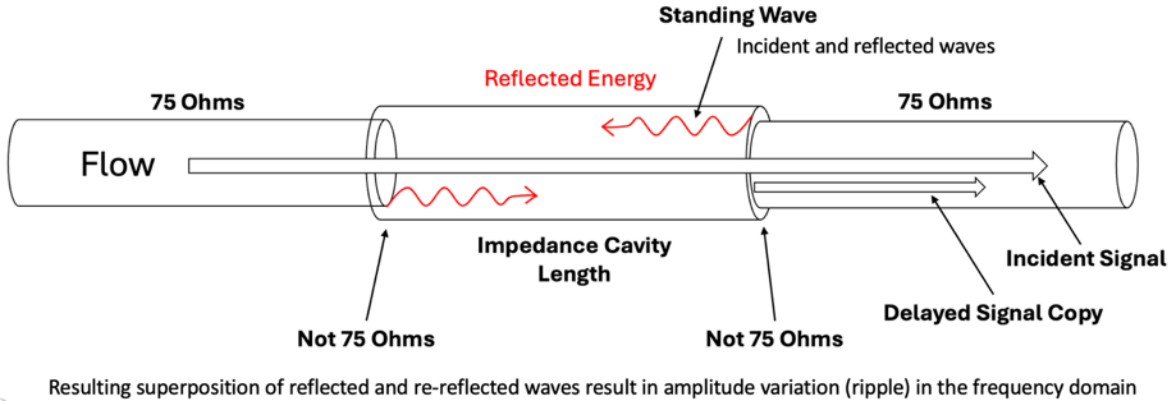


Figure 1 - Example Triple-Transit Reflections

Triple-transit reflections are distinguishable in the time-domain impulse response due to their longer delays and reduced amplitude, often significantly lower than incident signal, due to additional cable loss and passive losses, such as splitters and couplers. The high frequency resolution of DOCSIS 3.1 OFDM subcarriers (25 kHz or 50 kHz subcarrier spacing) allows these subtle echoes to be resolved and analyzed, providing insight into the network's topology, even in the presence of many different echoes. See Figure 2.

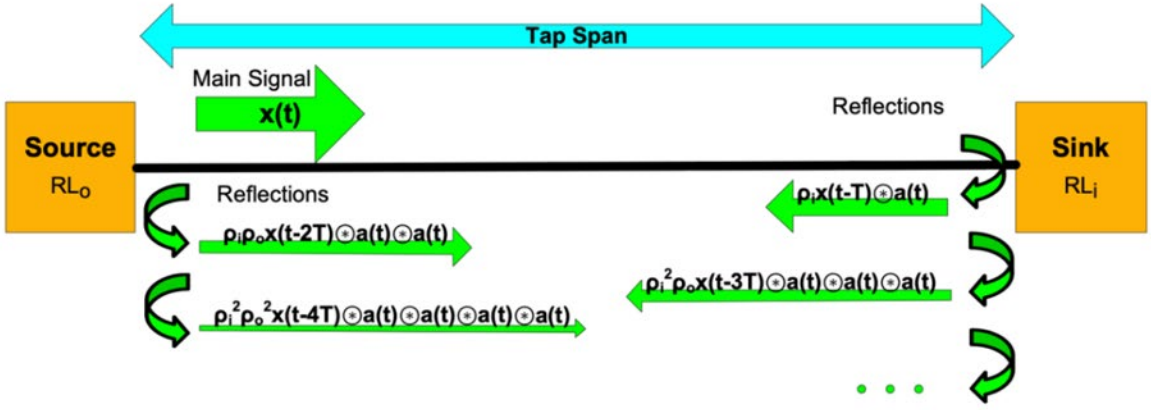


Figure 2 – Multiple Reflections and Their Coefficients (Courtesy: Dr. Richard Prodan)

3. Foundational Work

The methods used in this experiment build upon extensive research and standardization by the CableLabs Proactive Network Maintenance Working Group and the SCTE Network Operations Subcommittee (NOS) Working Group 7 (WG7) for PNM. These groups have:

1. Defined standardized methods for extracting and interpreting complex channel estimation coefficients from DOCSIS devices ([1] Section 4).
2. Developed best practices for phase correction and digital signal processing, enabling accurate conversion from frequency domain to time domain and reliable distance estimation ([1] Section 4).
3. Established classification schemes for identifying and distinguishing between open-circuit, short-circuit, and other fault signatures in the plant ([1] Section 5.3).

Notably, the work of Williams et al [1] and subsequent contributions by the CableLabs PNM and SCTE NOS WG7 communities have demonstrated that phase-corrected OFDM coefficients can accurately resolve echo cavity lengths within a few feet. This has enabled operators to move from broad, manual troubleshooting, to highly targeted, data-driven maintenance.

4. Drop Cable Lengths and Fault Detection

When a drop cable is cut, the modem at that location is no longer on the network, and a part of the drop cable remains connected at the tap and open where cut. The open end of a cable ideally presents a nearly infinite impedance (~∞ ohms) compared to the significantly larger (~75 ohms nominal) impedance of a connected modem. Practically speaking, the open end of the coax has less than infinite impedance (but still quite high) because of fringing capacitance and related factors. The very high impedance at the coaxial cable's open end results in a strong reflection with a reflection coefficient much larger than a properly terminated cable. Due to limited isolation between tap ports (for example, ~20 dB), this reflection can often be observed in the channel estimates of other modems connected to the same tap. The system detects this event by:

1. Noting the disappearance of the terminated modem
2. Observing the appearance of a new, higher-amplitude echo at a shorter distance (corresponding to the newly open cable end)
3. And correlating these changes across all active modems to localize the fault.

This approach enables near real-time, remote assessment of the drop network, allowing for rapid detection and localization of cut drops or other failures, especially valuable in environments prone to physical damage, such as during storms or construction activity. The result is a more resilient and maintainable cable network, as envisioned by the ongoing work of the CableLabs PNM and SCTE NOS working groups.

4.1. Laboratory Setup

These PNM experiments were conducted on-site at the Promptlink laboratory in Oceanside, California (Figure 3). The lab space was facilitated by our three volunteer lab technicians (Figure 4) for the benefit of the PNM community. The technicians participated in test setup, cable preparation, and cable testing with a time domain reflectometer (TDR) (Figure 5).

Figure 3 – The Authors, Inspecting the Test Facilities

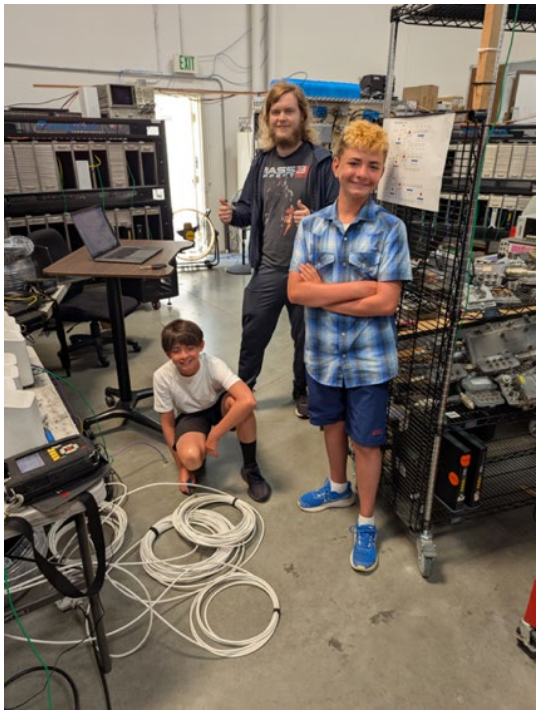


Figure 4 – The Lab Technicians, Inspecting the Cables

Figure 5 – The Lab Technicians, TDR Testing

4.2. Network Topology

A node was configured with a single downstream OFDM channel. The first four-port, 26 dB tap was connected to the node through 300 ft of series 11 coaxial cable, which was then validated with the TDR and factory length markings. After the first tap, 234 ft of 0.500 diameter feeder cable was used to connect a second tap (four-port, 23 dB). The second tap was terminated. See Figure 6 for a diagram of the test network topology.

For simplicity of the experiments in this technical paper, there are no modems connected to the second tap and all the tap ports are properly terminated with 75-ohm F-type terminators (CM5 - CM8) (Figure 7).

All the cable modems were of the same brand, model and firmware revision to avoid any inconsistency in reporting values. All the cable modems in the experiments were consistently connected to the same tap ports throughout each of the experiments.

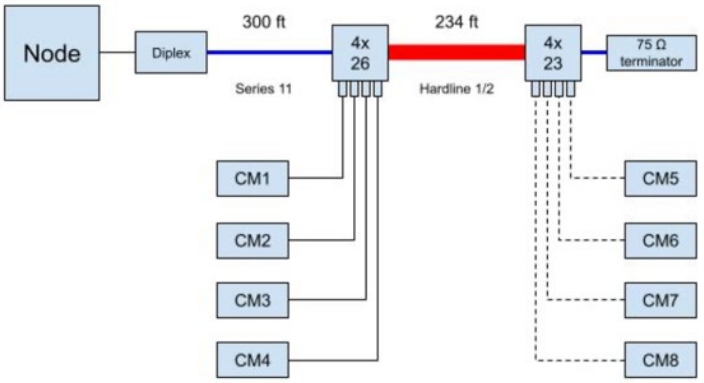


Figure 6 – Network Topology of Test

Figure 7 – Second Tap Terminated

4.3. Experiments

4.3.1. Experiment 1 : High-Value Tap, All Cables Intact

The first experiment was designed to validate the drop cable connections and lengths, using the OFDM channel estimates to measure the secondary reflections. The first tap is a high-value tap (four-port, 26 dB) and all the cables connected are intact as outlined in Table 1.

There are four cable modems connected to tap 1 (left side) according to the network topology diagram in Figure 6.

Table 1 – Experiment 1 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	Intact, Connected

The last tap at the end of line (right side, Figure 6) has all the ports terminated and output of the tap terminated (Figure 7).

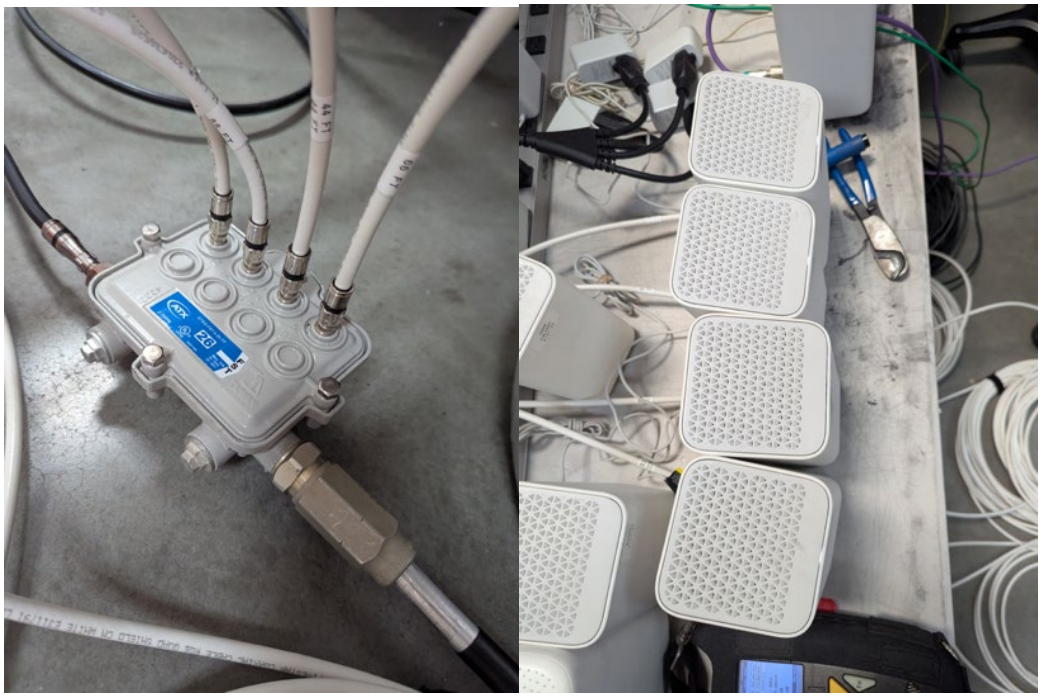


Figure 8 – 4 Test Modems and Their High-Value Tap Connections

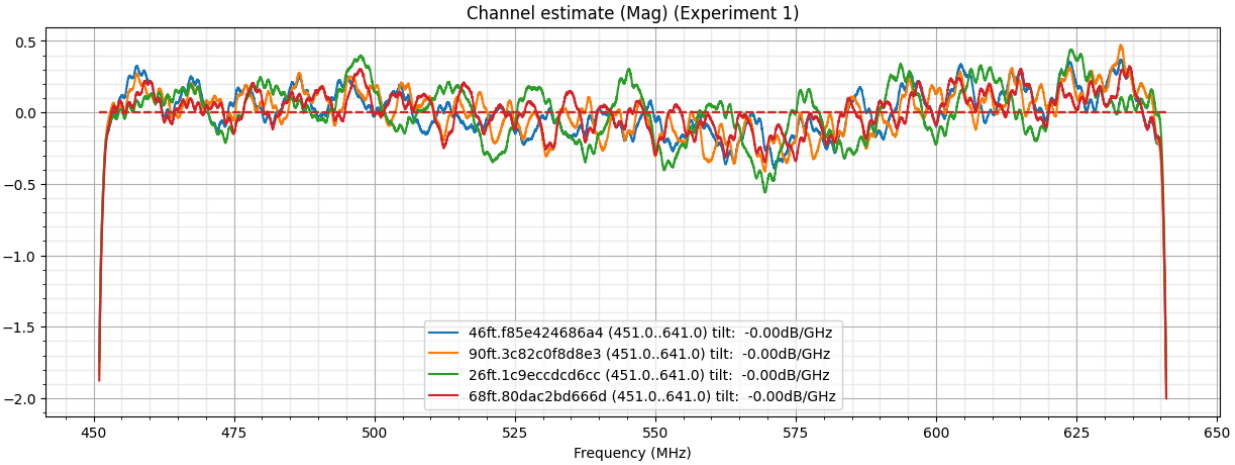


Figure 9 – Experiment 1, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

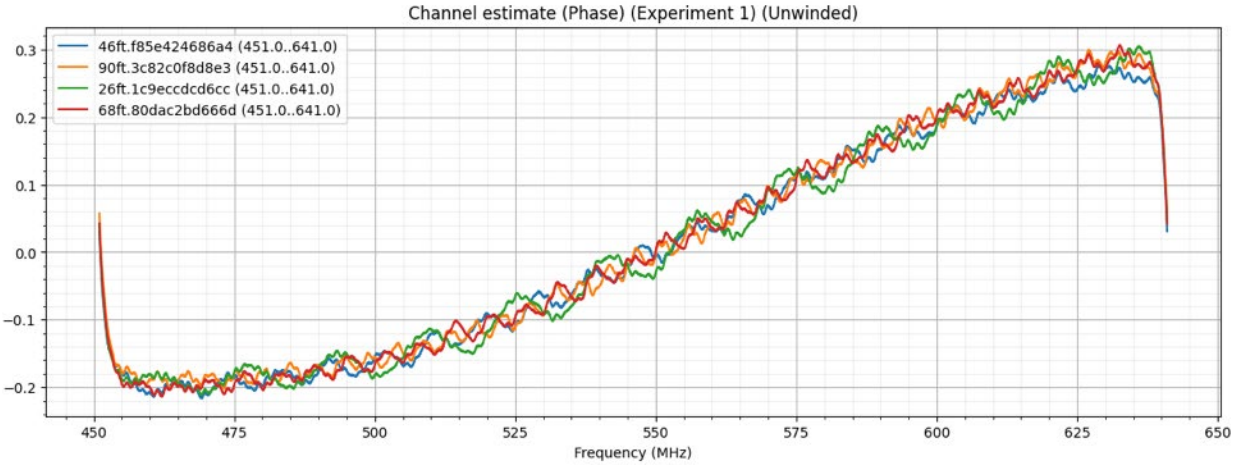


Figure 10 – Experiment 1, Channel Estimate, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 9) and frequency versus phase response (Figure 10) are collected, parsed, processed and graphed, using the PNM OFDM channel estimates.

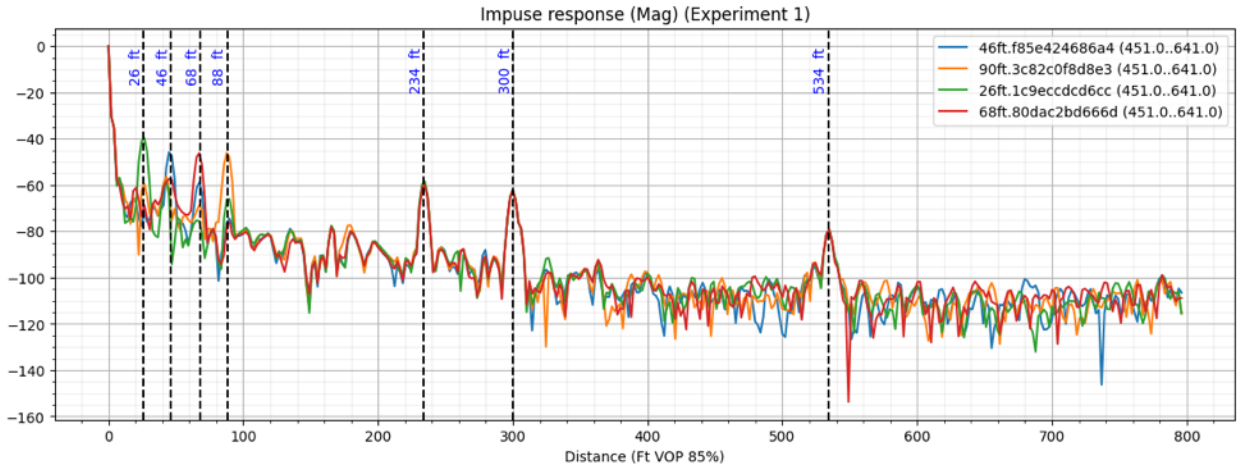


Figure 11 – Experiment 1, Time Domain Impulse Response, 800 Feet

Figure 12 – Experiment 1, Time-Domain Impulse Response, 400 Feet

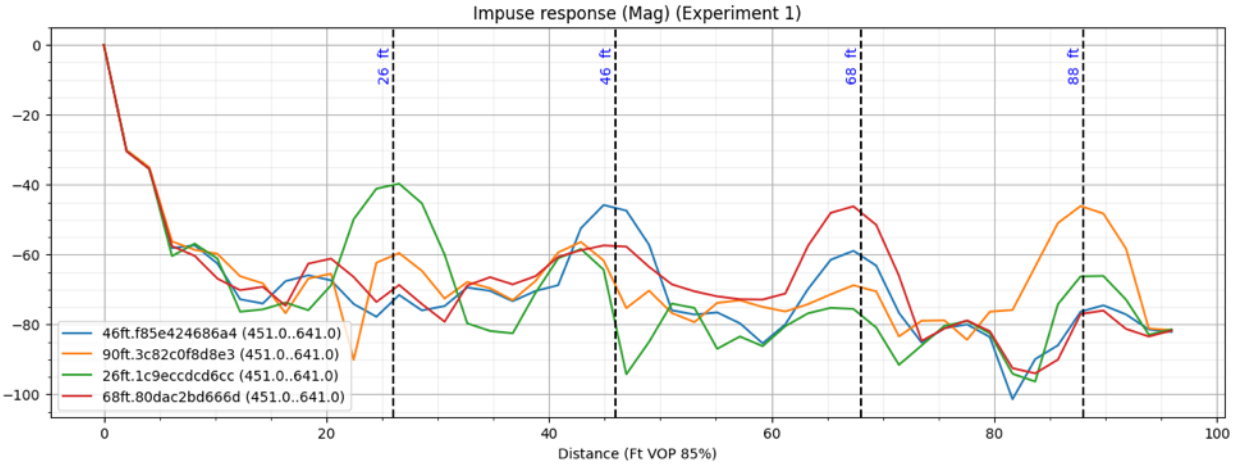


Figure 13 – Experiment 1, Time-Domain Impulse Response, 100 Feet

4.3.1.1. Analysis

The results in Figure 11, Figure 12, and Figure 13 are all time-domain impulse responses of the same channel estimate from each modem, truncated to 800, 400 and 100 feet to facilitate interpretation (note that cable modems in the plots listed in same order and color, and is consistent throughout all the plots). These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. The reflection magnitudes are largest at their directly connected modems because they are not subject to the port-to-port isolation. The 234 ft and 300 ft tap reflections are clearly seen, notably the highly similar magnitude (and phase, not shown). The magnitude and phase similarities are cause by the "common transit" of those echoes, taking the same signal path and electrical delay. Lastly, note the 534 ft reflection, which is caused by the impedance cavity between the diplexer output and the terminated end-of-line tap. That is, the 534 ft reflection is the convolution of the two smaller segments (234 ft and 300 ft). All time domain echo distances are within the maximum tolerance of 2.2 feet, given the operating parameters of this OFDM channel. The OFDM channel configured to 192 MHz (190 MHz of useable spectrum) and 25 kHz subcarrier spacing.

4.3.2. Experiment 2 : High-Value Tap, Cable 68 ft Cut at Random Distance

The second experiment was designed to demonstrate the effects of changing the impedance cavity lengths and return loss expected when a drop cable is cut or unterminated. The first tap is a high-value tap (four-port, 26 dB) and all the cables are connected and intact, except the 68 ft cable. The cable is cut between cable modem and tap port 4, leaving a 32 ft unterminated drop connected to the tap (Figure 14). All the cables are connected as summarized in Table 2.

Table 2 – Experiment 2 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	Cut 32 ft from the tap, Connected

The last tap at the end of line (right side, Figure 6) has all the ports terminated and output of the tap terminated (Figure 7).

Figure 14 – Cut Cable at 32 ft (Randomly Selected Length)

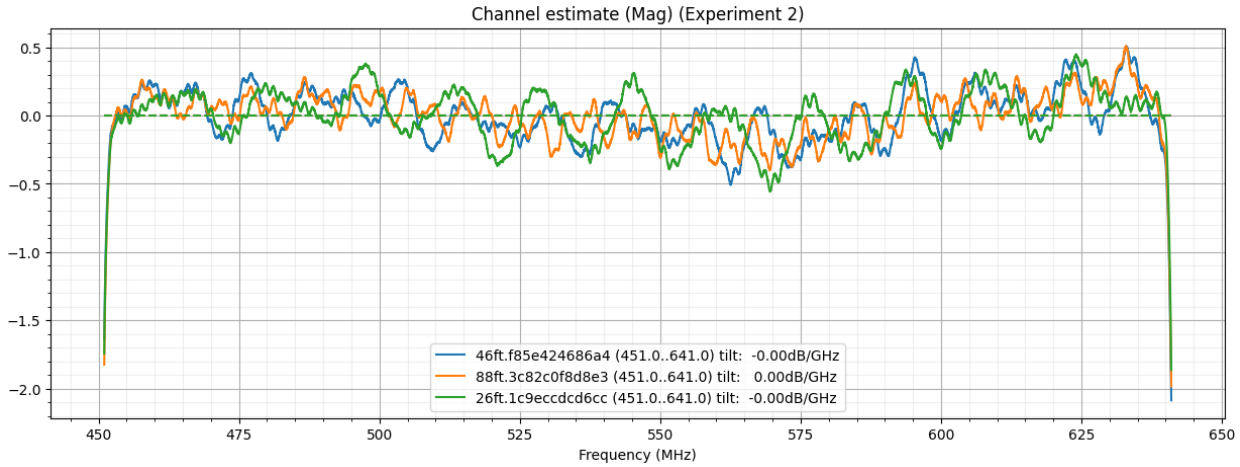


Figure 15 – Experiment 2, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

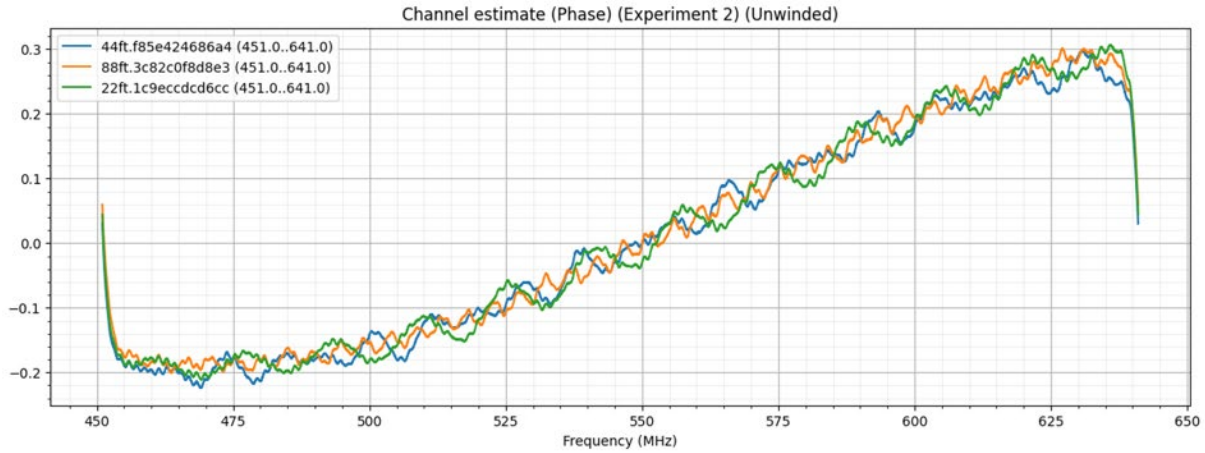


Figure 16 – Experiment 2, Channel Estimate, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 15) and frequency versus phase response (Figure 16) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates.

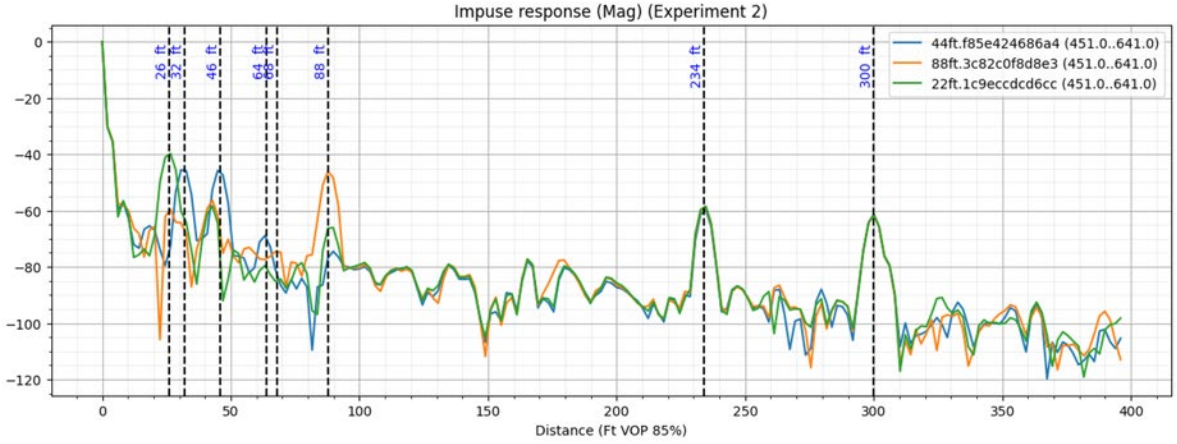


Figure 17 – Experiment 2, Time-Domain Impulse Response, 400 Feet

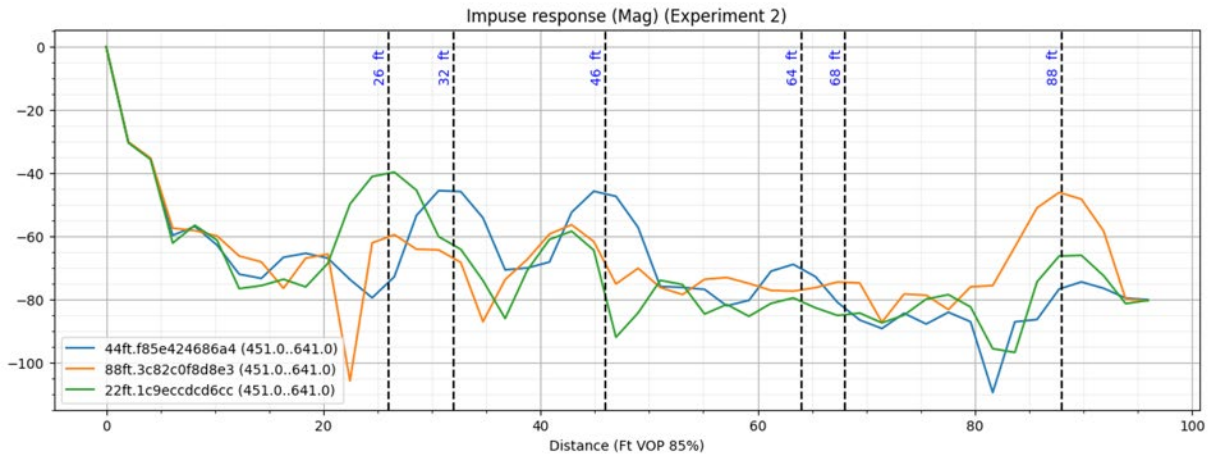


Figure 18 – Experiment 2, Time-Domain Impulse Response, 100 Feet, Cut at 32 and 64 Feet,

4.3.2.1. Analysis

The results in Figure 17 and Figure 18 are time-domain impulse responses of the same channel estimate from each modem, truncated to 400 and 100 ft to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. Note the red trace (measured from modem 80da.c2bd.666d) having the 68 ft drop cable length is now missing, because the cable modem is offline. Also note the elevated levels appearing at 32 ft and 64 ft which were not present in experiment 1. This is the new echo cavity that was created between the tap port 4 and the end of the cut cable, correctly distanced at 32 ft and 64 ft.

4.3.3. Experiment 3 : High-Value Tap, Cable 68 ft Cut Disconnected from the Tap and Tap Port Terminated with 75 ohm Terminator

The third experiment was designed to demonstrate the effects of properly terminated tap ports on the port-top-port isolation of the tap. As before, the first tap is high-value tap (four-port, 26 dB) and all the cables connected are intact, except the 68 ft. cable was disconnected from tap port 4 and properly terminated with a 75 ohm terminator. All the cables are connected as summarized in Table 3.

Table 3 – Experiment 3 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	Not Connected, Terminated

The last tap at the end of line (right side, Figure 6) has all the output ports terminated in addition to the OUT port of the tap, which is also terminated (Figure 7).

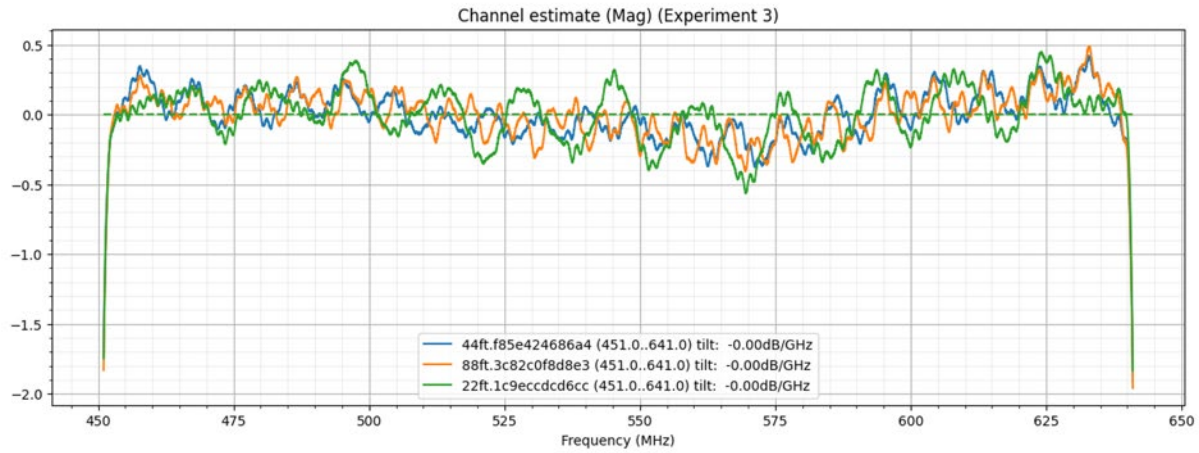


Figure 19 – Experiment 3, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

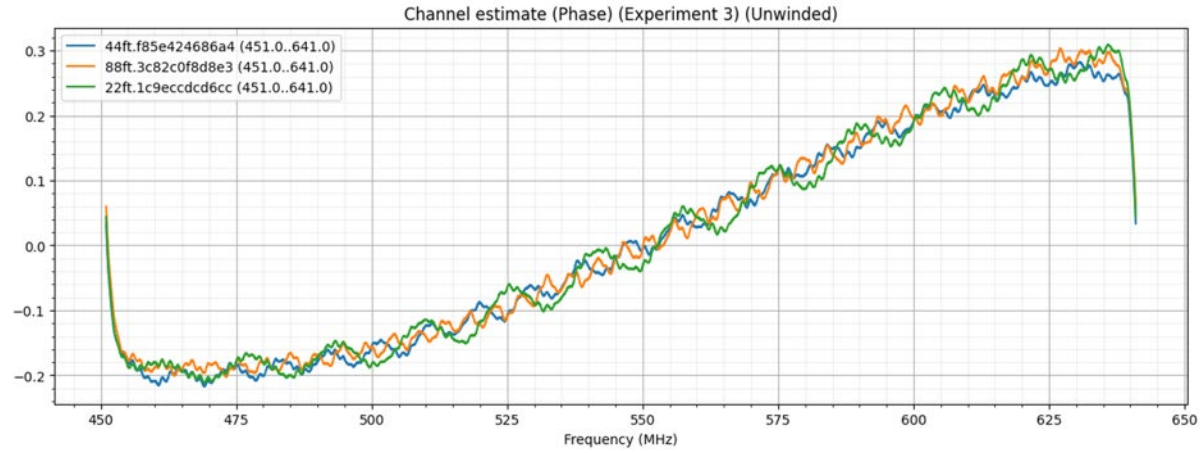


Figure 20 – Experiment 3, Channel Estimate, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 19) and frequency versus phase response (Figure 20) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates data.

Figure 21 – Experiment 3, Time Domain Impulse Response, 400 Feet

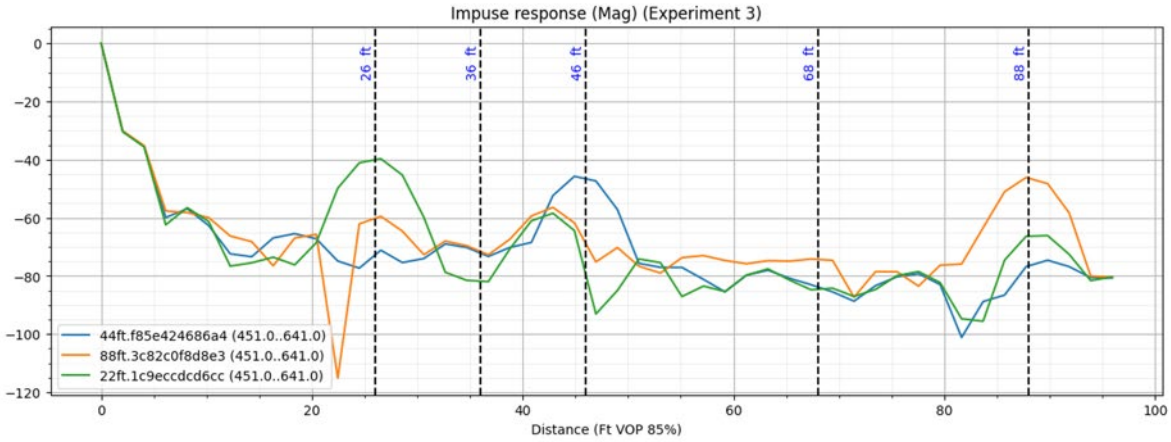


Figure 22 – Experiment 3, Time Domain Impulse Response, 100 Feet

4.3.3.1. Analysis

The results in Figure 21 and Figure 22 are time-domain impulse responses of the channel estimate from each modem, truncated to 400 and 100 ft to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. In this experiment, the 32 ft reflection from the cut drop cable is now missing, because the port 4 was properly terminated.

4.3.4. Experiment 4 : High-Value Tap, Cable 68 ft Cut Repaired with F81 Splice

The fourth experiment was intended to determine if the splice in a repaired cable could be detected in the channel estimate. As before, the first tap is high-value tap (4-port, 26 value) with all the cables connected and intact except the 68 ft drop cable. The cable is cut between the cable modem and tap port 4 (32 ft away from the tap and 36 ft away from the cable modem). The cut was repaired (Figure 23) using an F-81 (barrel) connector. All the cables are connected as summarized in Table 4.

Figure 23 – Lab Technicial Repairing Cut Cable With Splice

Table 4 – Experiment 4 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	F81 splice 36 ft from CM, Connected

The last tap at the end of line (right side, Figure 6) has all the output ports terminated and OUT of the tap also terminated (Figure 7).

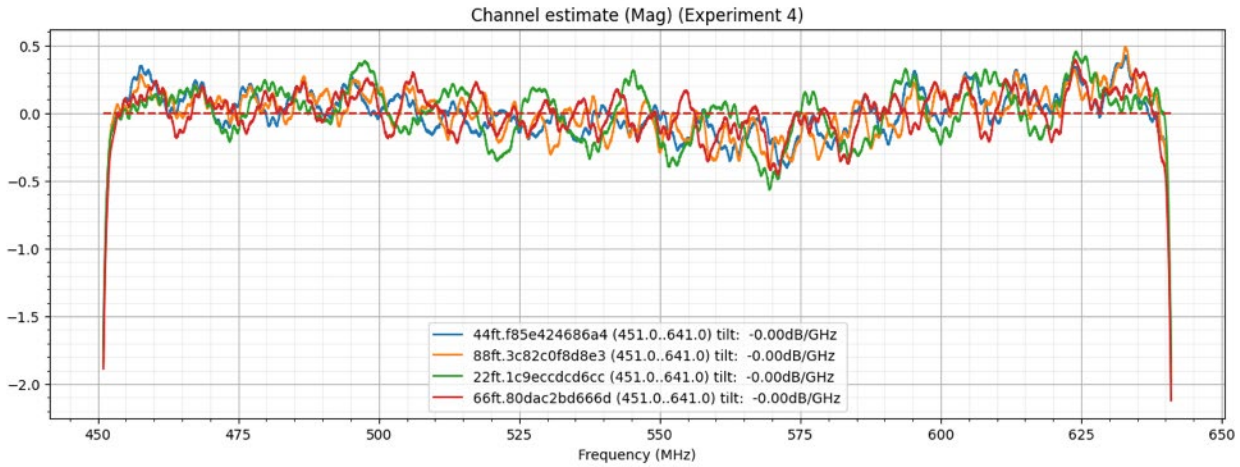


Figure 24 – Experiment 4, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

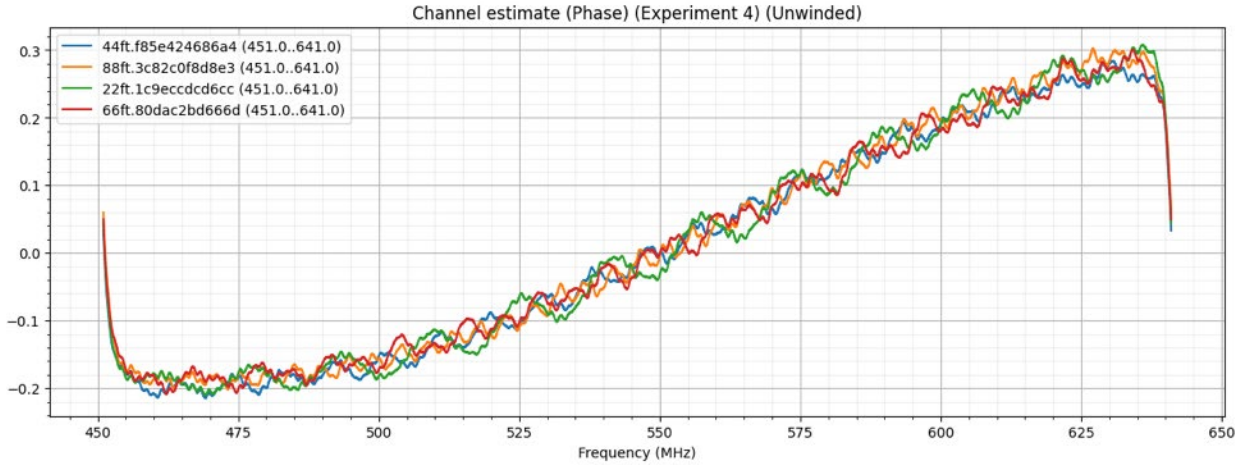


Figure 25 – Experiment 4, Channel Estimate, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 24) and frequency versus phase response (Figure 25) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates data.

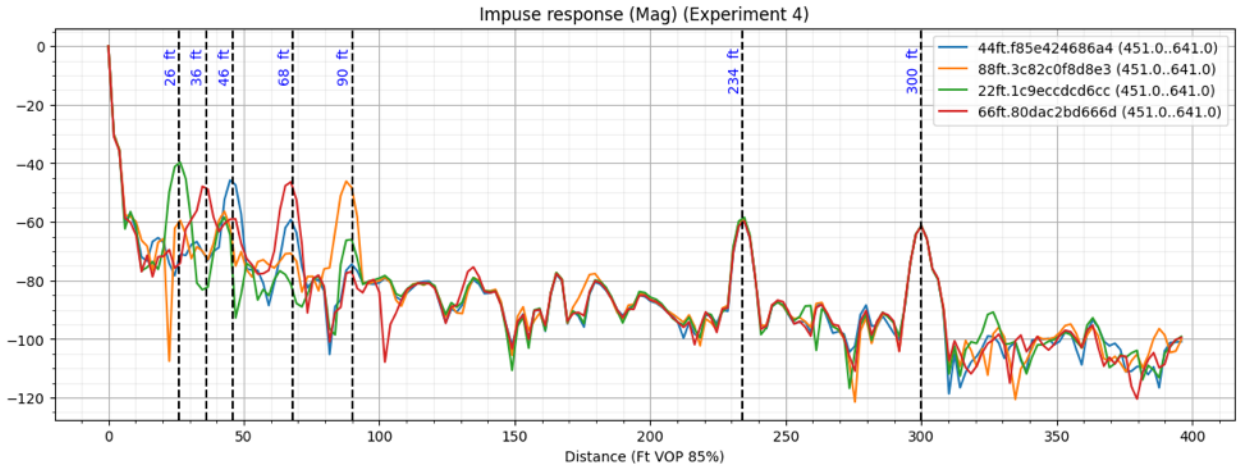


Figure 26 – Experiment 4, Time Domain Impulse Response, 400 Feet

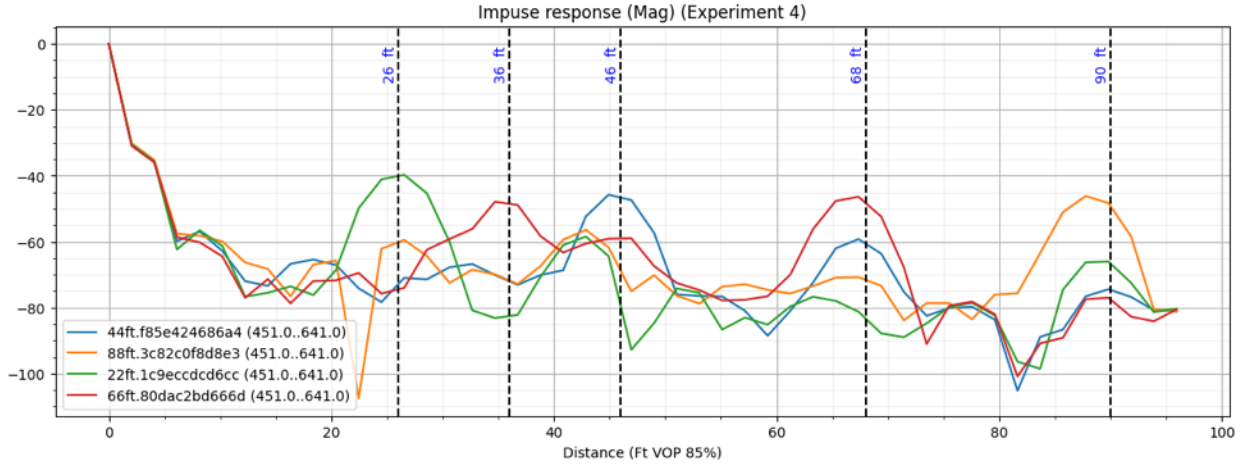


Figure 27 – Experiment 4, Time Domain Impulse Response, 100 Feet

4.3.4.1. Analysis

The results in Figure 26 and Figure 27 are time-domain impulse responses of the channel estimate from each modem, truncated to 400 and 100 ft to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. In this experiment, the red trace (measured from modem 80da.c2bd.666d) at 68 ft has returned since the cable modem is now back online. At the same time, the reflection at 36 ft is visible but significantly reduced, because the return loss of the F-81 splice is much better than the open end of a cut drop cable.

4.3.5. Experiment 5 : Low-Value Tap, All Cables Intact

The fifth experiment was designed to repeat the first experiment but using a low-value instead of a high-value tap. The first tap is low-value tap (four-port, 11 dB) and all the cables are connected and intact as outlined in Table 5

There are four cable modems connected to tap 1 (left side) according to the network topology diagram in Figure 6.

Table 5 – Experiment 5 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	Intact, Connected

The last tap at the end of line (right side, Figure 6) has all the output ports terminated and the OUT port of the tap also terminated (Figure 7).

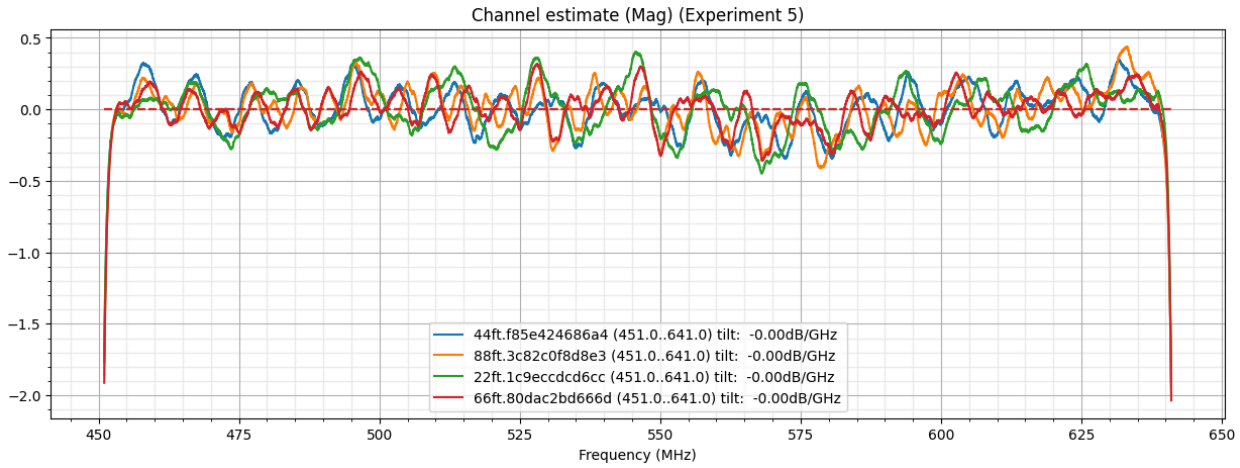


Figure 28 – Experiment 5, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

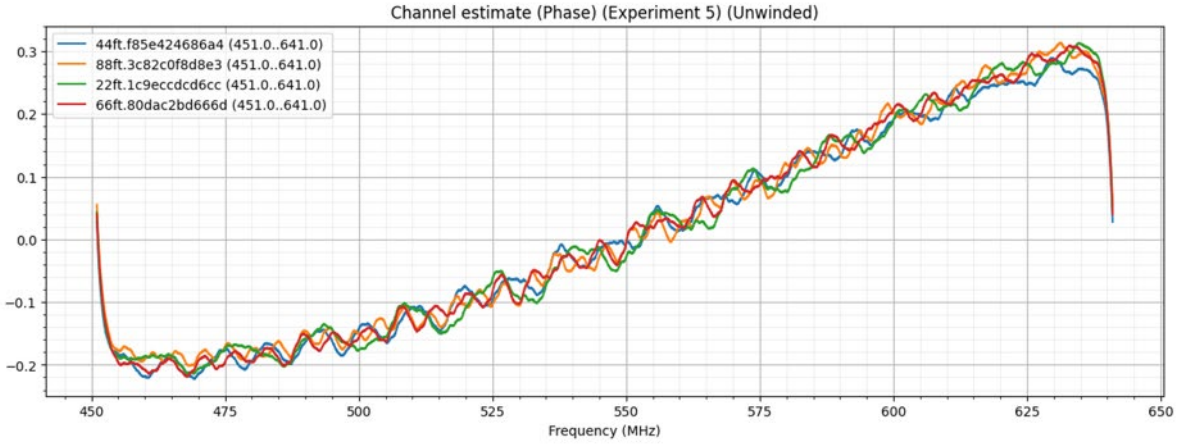


Figure 29 – Experiment 5, Channel Estimate, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 28) and frequency versus phase response (Figure 29) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates data.

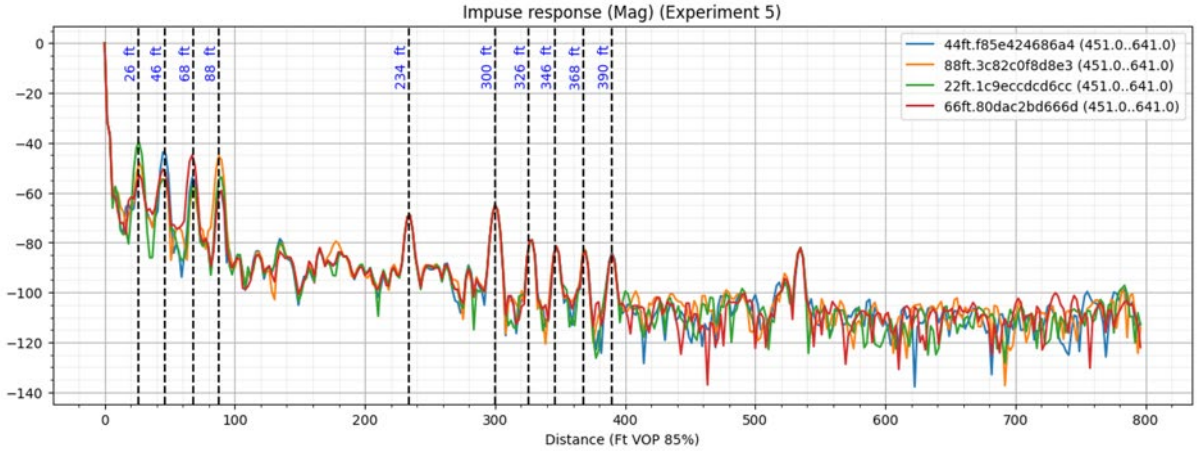


Figure 30 – Experiment 5, Time Domain Impulse Response, 800 Feet

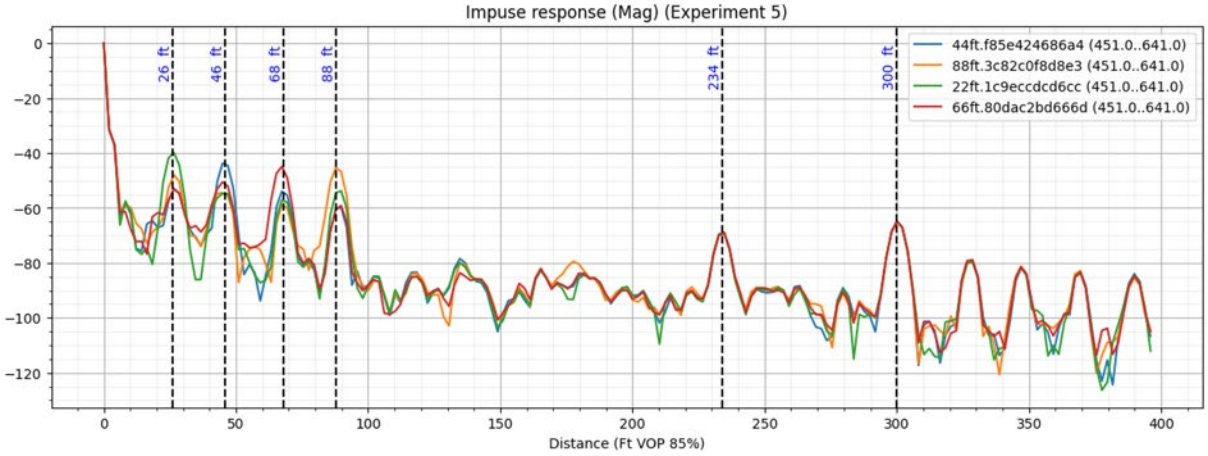


Figure 31 – Experiment 5, Time Domain Impulse Response, 400 Feet

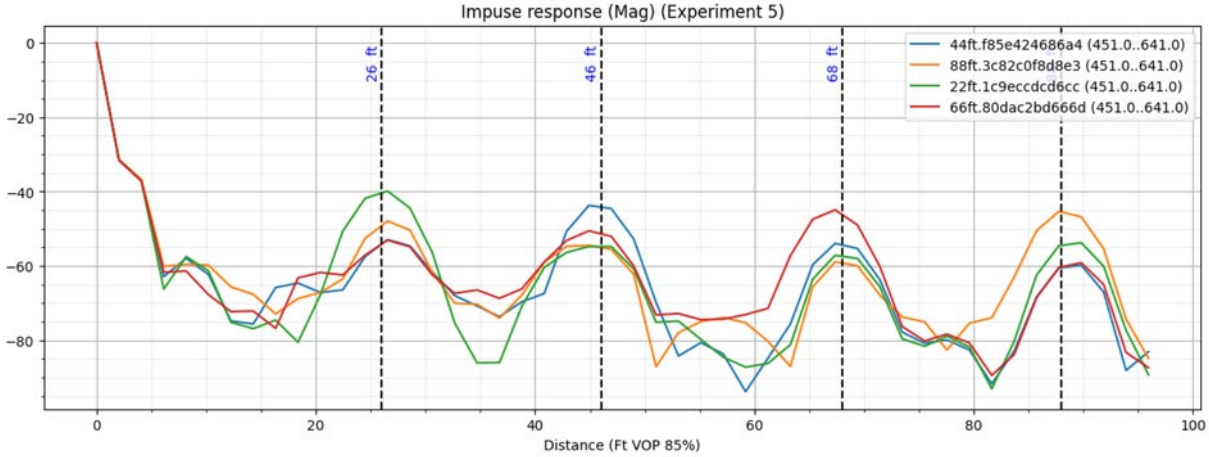


Figure 32 – Experiment 5, Time Domain Impulse Response, 100 Feet

4.3.5.1. Analysis

The results in Figure 31 and Figure 32 are time-domain impulse responses of the channel estimate from each modem, truncated to 800 and 400 ft, respectively, to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. The drop cable lengths are all present and consistent with experiment 1, however, due to the reduced tap attenuation, another set of echo transits can be seen. The cable lengths at markers 26 ft, 46 ft, 68 ft and 88 ft are now repeated after the 300 ft tap echo (see Figure 30). These can be seen at markers 326 ft, 346 ft, 368 ft, and 390 ft. The phenomena has been described as a common transit by the working group.

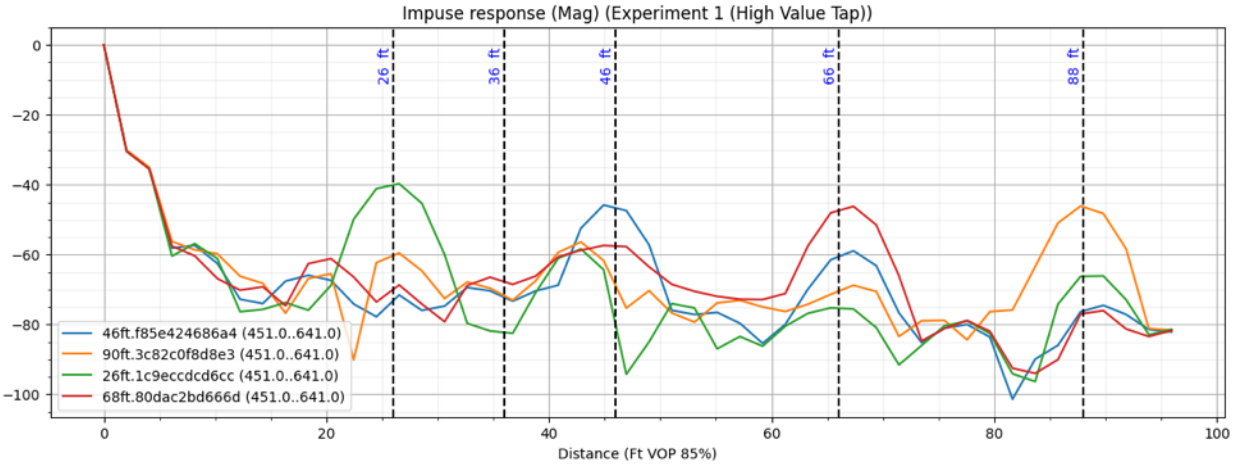


Figure 33 – Experiment 1 High-Value Tap, Time Domain Impulse Response, 100 Feet

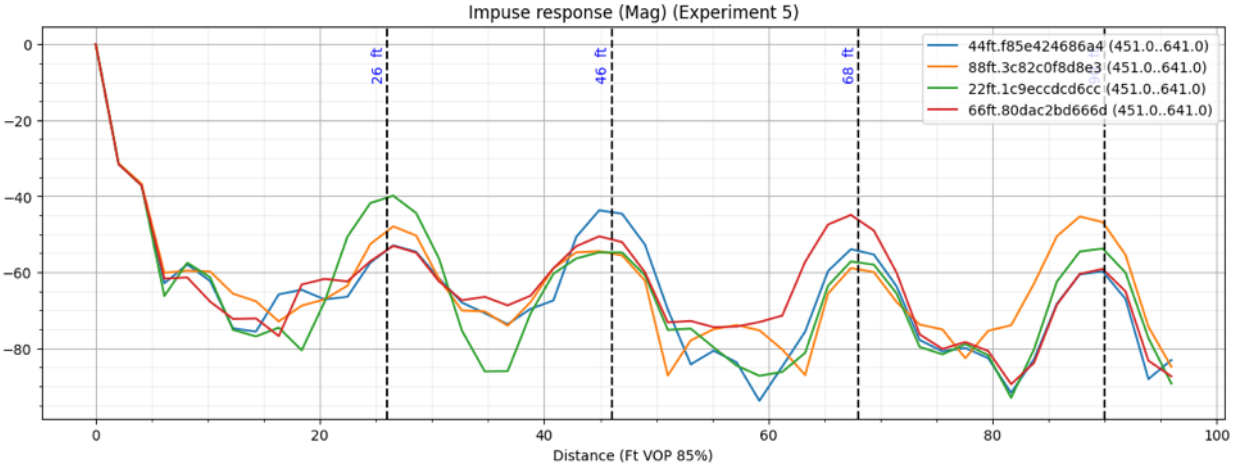


Figure 34 – Experiment 5, Low-Value Tap, Time Domain Impulse Response, 100 Feet

Figure 33 and Figure 34 are time-domain impulse responses of the channel estimate from each modem, truncated to 100 ft to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. However, the traces in Figure 33 demonstrate the effects of higher tap attenuation and port-to-port isolation. Figure 34, having lower tap attenuation and port-to-port isolation shows higher consistency with the reflection magnitudes seen by each of the modems.

4.3.6. Experiment 6 : Low-Value Tap, Cable 68 ft Cut at Random Distance

The next experiment is identical to experiment 2 with first tap replaced with low-value tap. The first tap is a low-value tap (four-port, 11 dB) and all the cables connected are intact except 68 ft. It is cut between the cable modem and tap port 4, leaving the now unterminated drop (at 32 ft) connected to the tap (Figure 14). All the cables are connected as summarized in Table 6.

Table 6 – Experiment 6 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	Cut 32 ft from the tap, Connected

The last tap at the end of line (right side, Figure 6) has all the output ports terminated and OUT port of the tap also terminated (Figure 7).

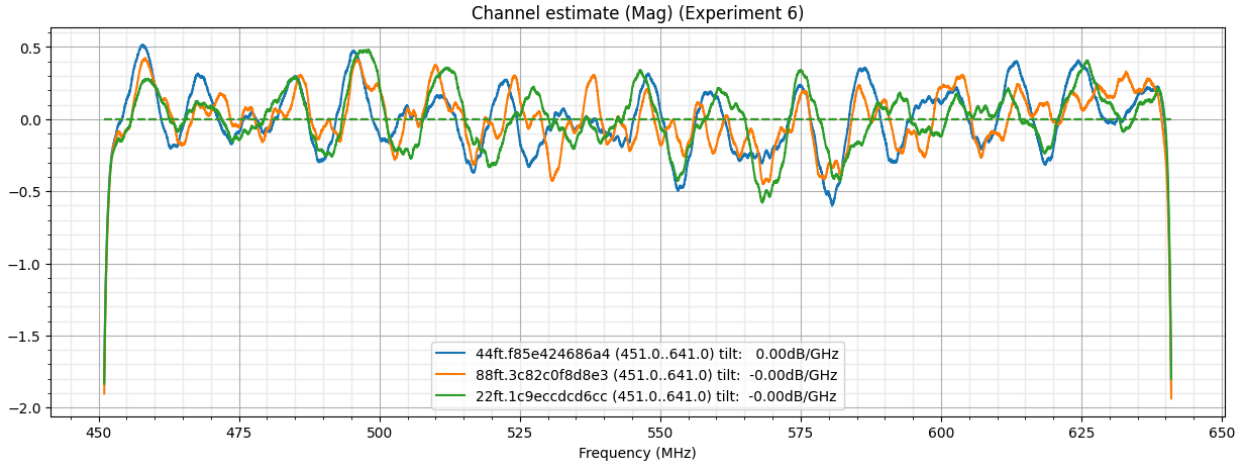


Figure 35 – Experiment 6, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

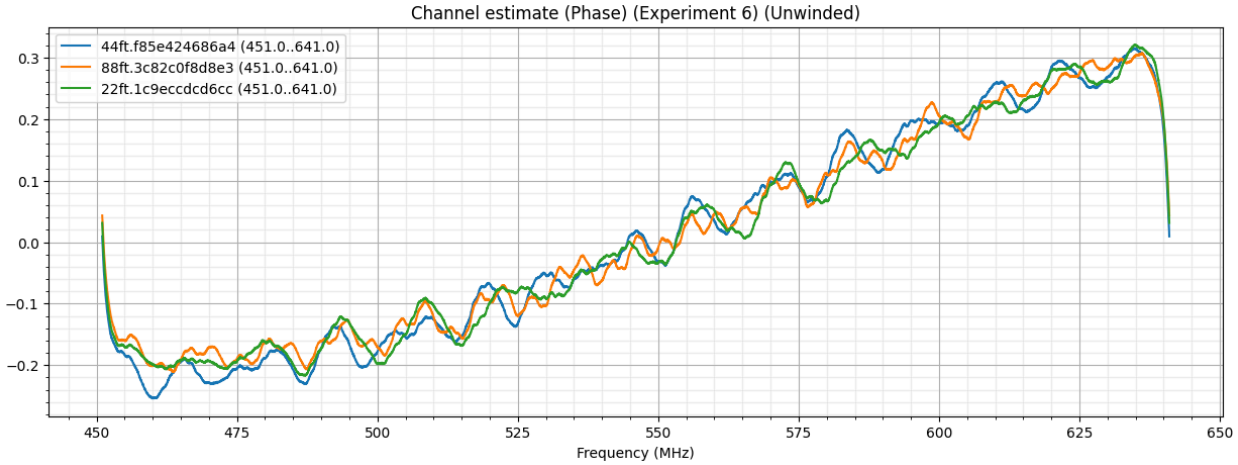


Figure 36 – Experiment 6, Channel Estimate, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 35) and frequency versus phase response (Figure 36) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates data.

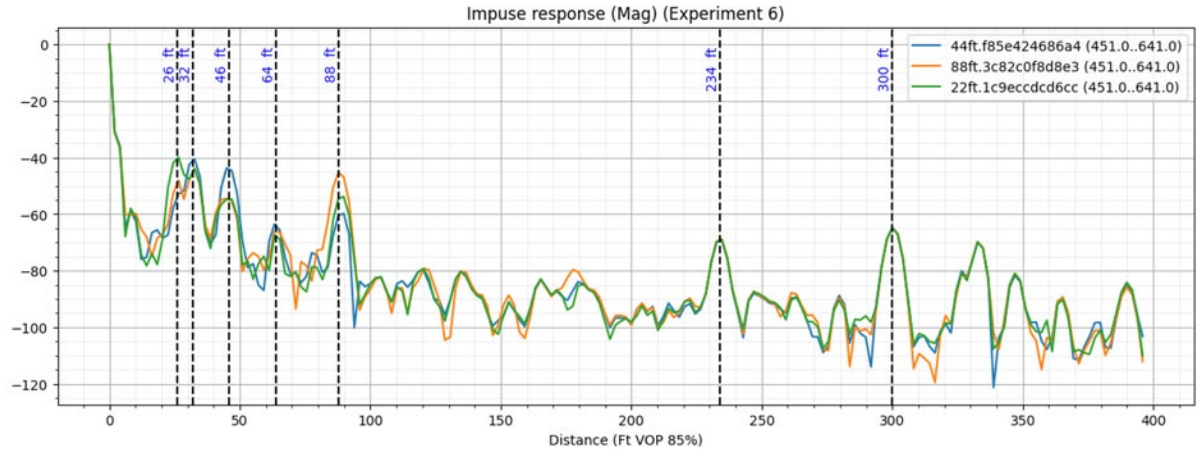


Figure 37 – Experiment 6, Time Domain Impulse Response, 400 Feet

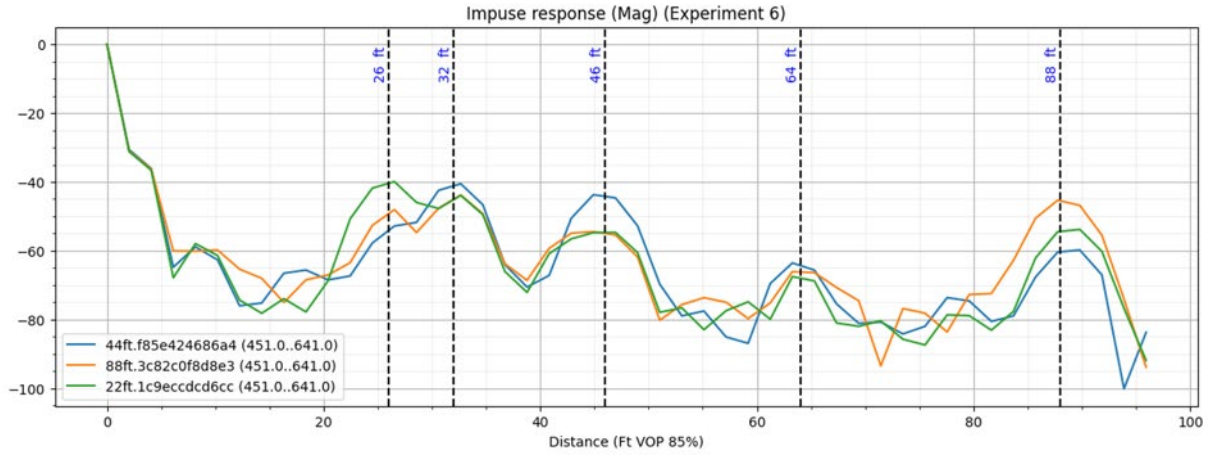


Figure 38 – Experiment 6, Time Domain Impulse Response, 100 Feet

4.3.6.1. Analysis

The results in Figure 37 and Figure 38 are time-domain impulse responses of the channel estimate from each modem, truncated to 400 and 100 ft, respectively, to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. Note the red trace (measured from modem 80da.c2bd.666d) having the 68 ft drop cable length is now missing, because the cable modem is offline. Also note the new trace appearing at 32 ft which was not present in experiment 1 and experiment 5. This is the new echo cavity that was created between the tap output port 4 and the open end of the cut cable, correctly distanced at 32 ft. Having a lower-value tap attenuation and port-to-port isolation results in increased level of reflections on modems attached to different ports of the same tap. The same can be observed in the longer-distance common transit reflections.

4.3.7. Experiment 7 : Low-Value Tap, Cable 68 ft Cut Disconnected from the Tap and Tap port Terminated With 75 ohm Terminator

Experiment 7 is identical to Experiment 3 with first tap replaced with low-value tap (four-port, 11 dB). This experiment was designed to demonstrate the effects of changing the impedance cavity lengths and return loss expected with a cut drop cable. All the cables are connected and intact except the 68 ft. cable disconnected from the tap port 4 and properly terminated with 75 ohm terminator. All the cables are connected as summarized in Table 3.

Table 7 – Experiment 7 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	Not Connected, Terminated

The last tap at the end of line (right side, Figure 6) has all the output ports terminated and the OUT port of the tap also terminated (Figure 7).

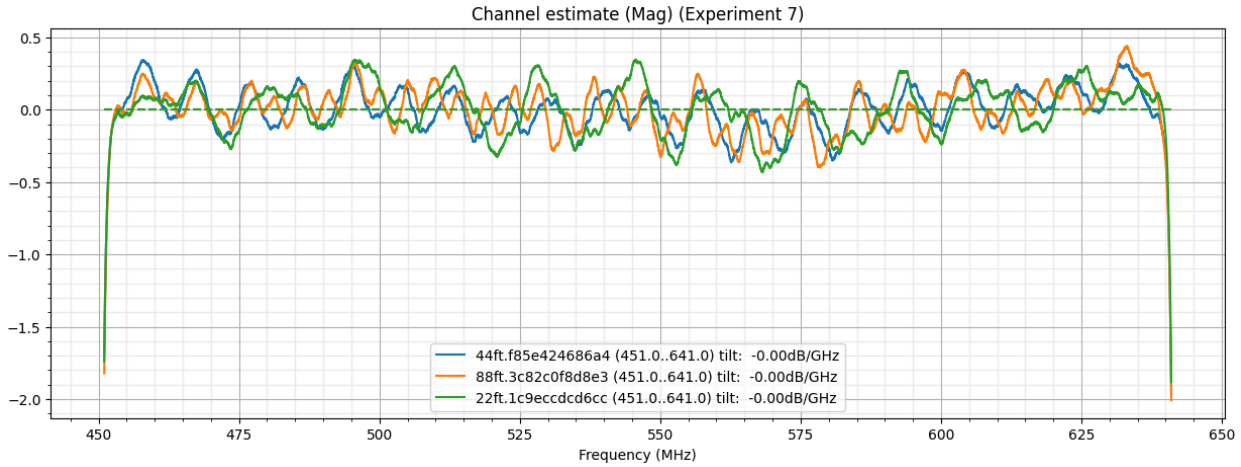


Figure 39 – Experiment 7, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

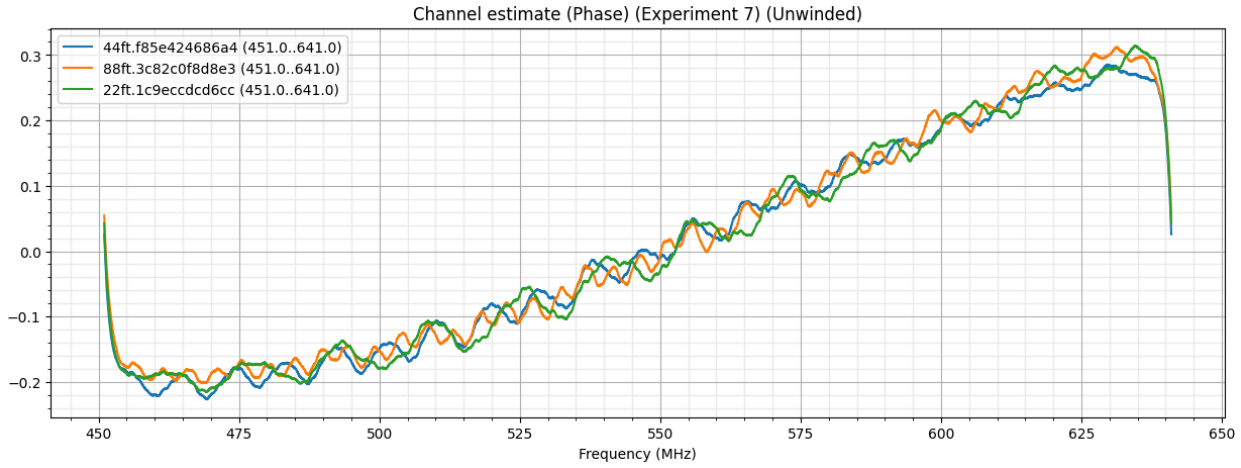


Figure 40 – Experiment 7 Channel Estimates, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 39) and frequency versus phase response (Figure 40) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates data.

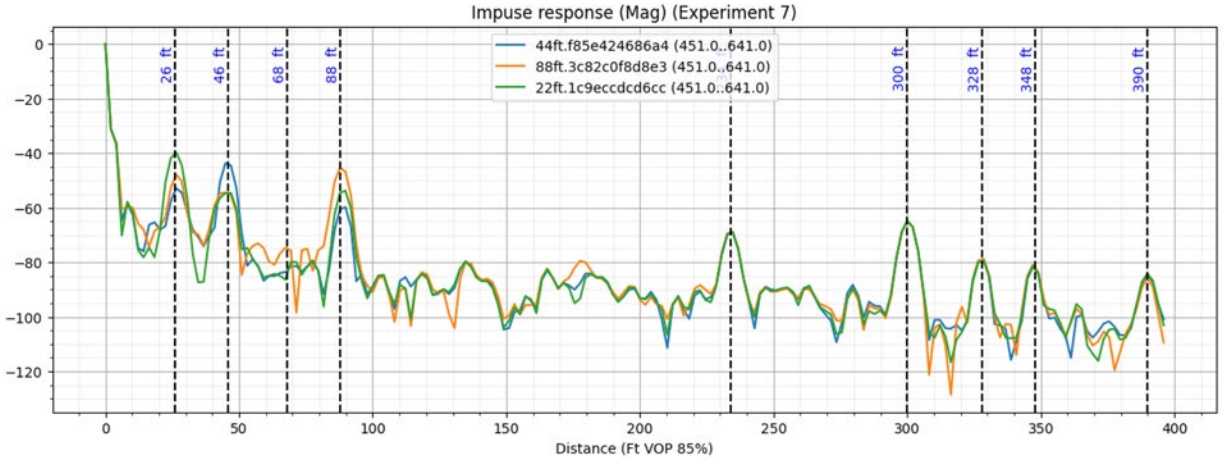


Figure 41 – Experiment 7, Time Domain Impulse Response, 400 Feet

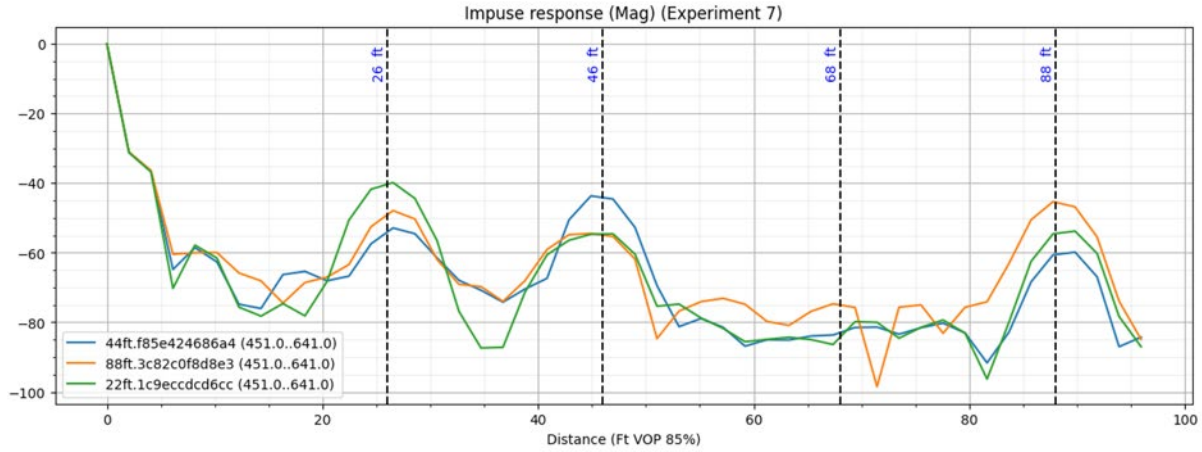


Figure 42 – Experiment 6, Time Domain Impulse Response, 100 Feet

4.3.7.1. Analysis

The results in Figure 41 and Figure 42 are time-domain impulse responses of the same channel estimate from each modem, truncated to 400 and 100 ft, respectively, to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. In this experiment, the 32 ft reflection from the cut drop cable is now missing, because tap port 4 was now properly terminated. The same common transit reflection pattern can be seen beyond the 300 ft tap echo, lacking a 332 ft echo from terminating the port where the 32 ft cut drop was previously connected.

4.3.8. Experiment 8 : Low-Value Tap, Cable 68 ft cut repaired with F81 splice

The final experiment is identical to experiment 4 and intended to determine if the splice in a repaired cable could be detected. However, the high-value tap was replaced with a low-value tap. The first tap is high-value tap (4-port, 11 dB) and all the cables connected and intact except the 68 ft drop cable on tap port 4. It is cut between the cable modem and tap port 4 (32 ft away from the tap and 36 ft away from the cable modem). The cut was repaired (Figure 23) using an F-81 (barrel) connector. All the cables are connected as summarized in Table 8.

Table 8 – Experiment 8 Tap Wiring

Cable Modem	Cable Distance	Tap Port	Cable status
f85e.4246.86a4	46 ft	5	OK
3c82.c0f8.d8e3	88 ft	3	OK
1c9e.ccdc.d6cc	26 ft	2	OK
80da.c2bd.666d	68 ft	4	F81 splice 36 ft from CM, Connected

The last tap at the end of line (right side, Figure 6) has all the output ports terminated and the OUT port of the tap also terminated (Figure 7).

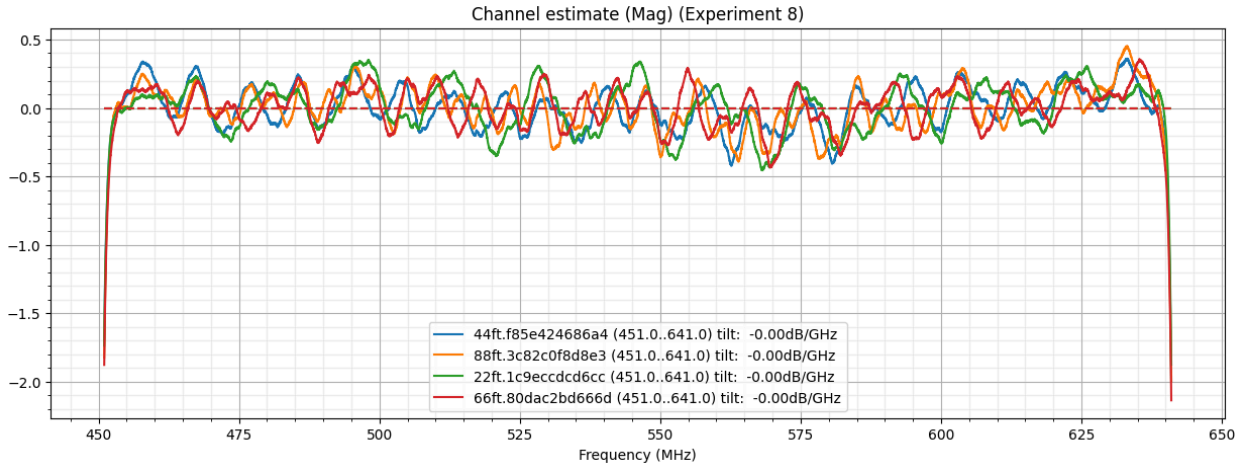


Figure 43 – Experiment 8, Channel Estimates, Magnitude Response (dB) versus Frequency (MHz)

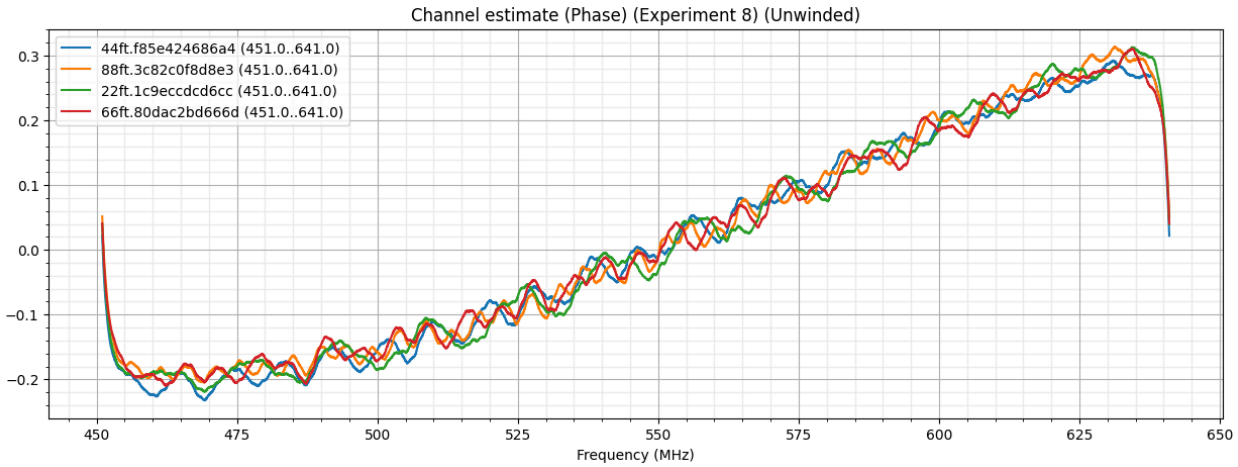


Figure 44 – Experiment 7 Channel Estimates, Phase Response (cycles) versus Frequency (MHz)

The frequency versus magnitude response (Figure 43) and frequency versus phase response (Figure 44) are collected from the cable modems, parsed, processed and graphed, using the PNM OFDM channel estimates.

Figure 45 – Experiment 8, Time-Domain Impulse Response, 400 Feet

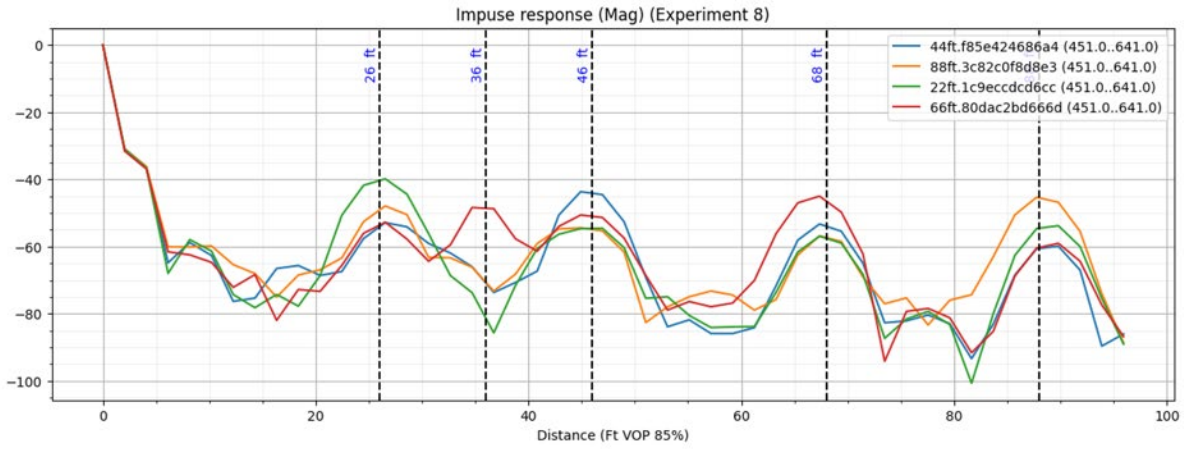


Figure 46 – Experiment 8, Time-Domain Impulse Response, 100 Feet

4.3.8.1. Analysis

The results in Figure 45 and Figure 46 are time-domain impulse responses of the channel estimate from each modem, truncated to 400 and 100 ft to facilitate interpretation. These results show the correct drop lengths, with the peak magnitudes associated to their respective cable modems. In this experiment, the red trace (measured from modem 80da.c2bd.666d) at 68 ft has returned since the cable modem is now back online. At the same time, the reflection at 36 ft is visible but significantly reduced, because the return loss of the F-81 splice is much better than the open end of a cut drop cable. It is noteworthy that the splice's echo magnitude is the approximately the same when measured on low-value and high-value taps. This is because the impedance cavity between the splice and cable modem remain constant, regardless of the tap's impedance properties.

5. Comparative Analysis

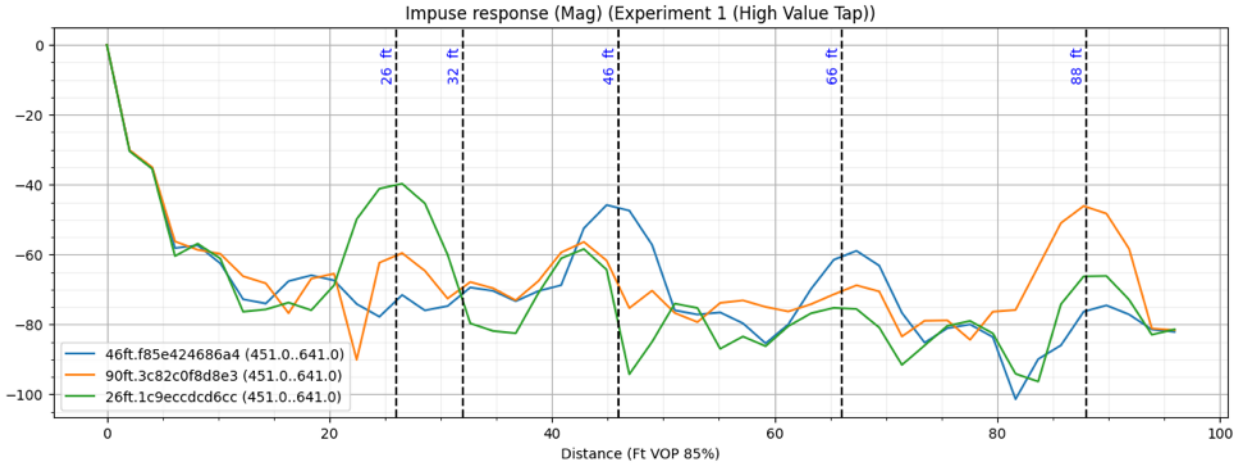


Figure 47 – Experiment 1, High-Value Tap, Time-Domain Impulse Response, 100 Feet

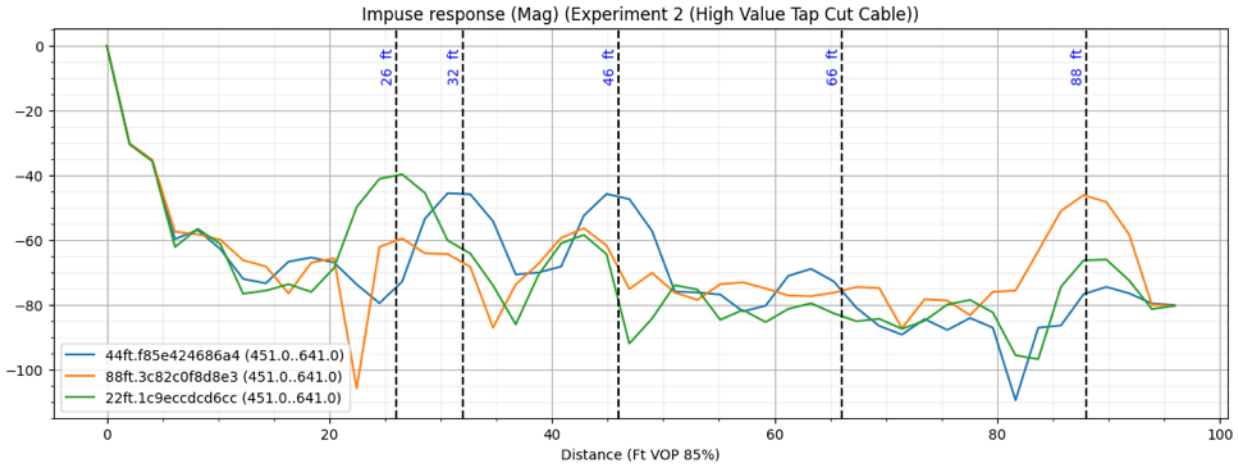


Figure 48 – Experiment 2, High-Value Tap, Cut Drop, Time-Domain Impulse Response, 100 Feet

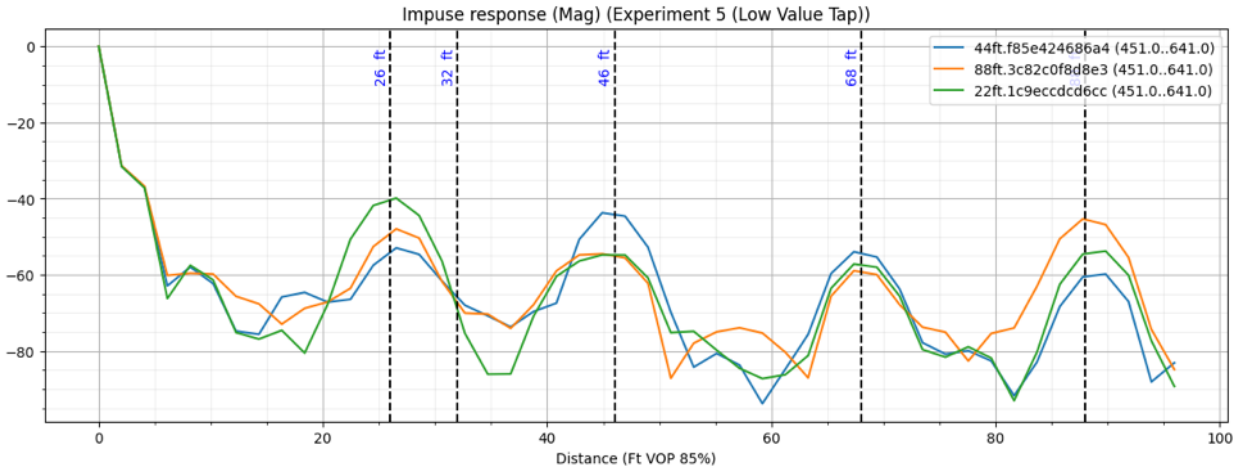


Figure 49 – Experiment 5, Low-Value Tap, Time-Domain Impulse Response, 100 Feet

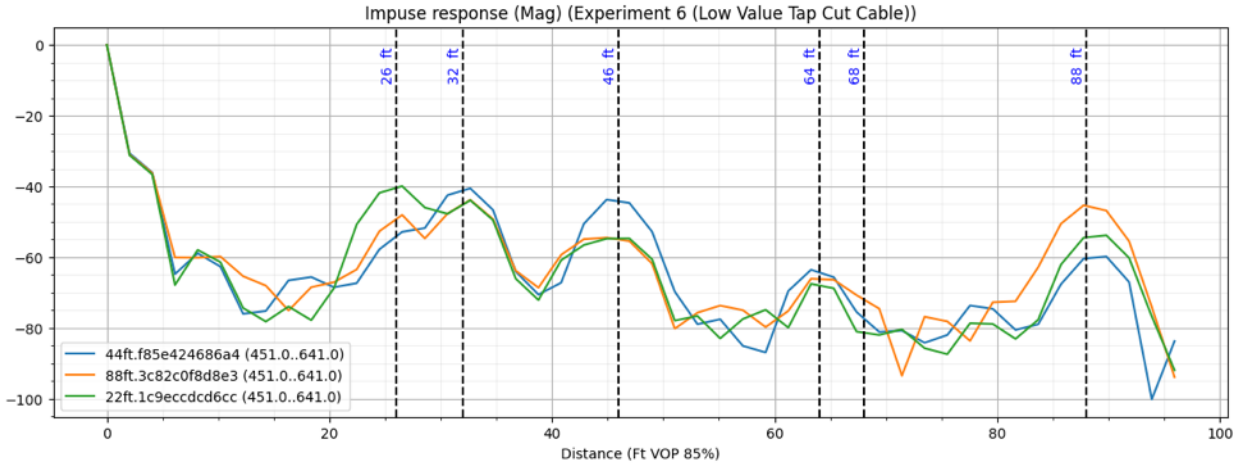


Figure 50 – Experiment 6, Low-Value Tap, Cut Drop, Time-Domain Impulse Response, 100 Feet

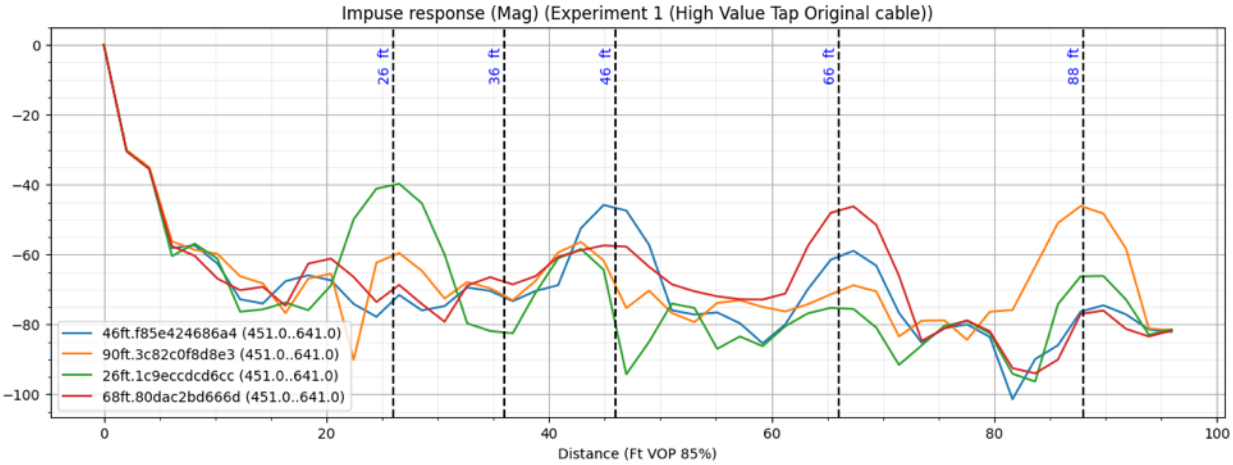


Figure 51 – Experiment 1, High-Value Tap, Time-Domain Impulse Response, 100 Bins

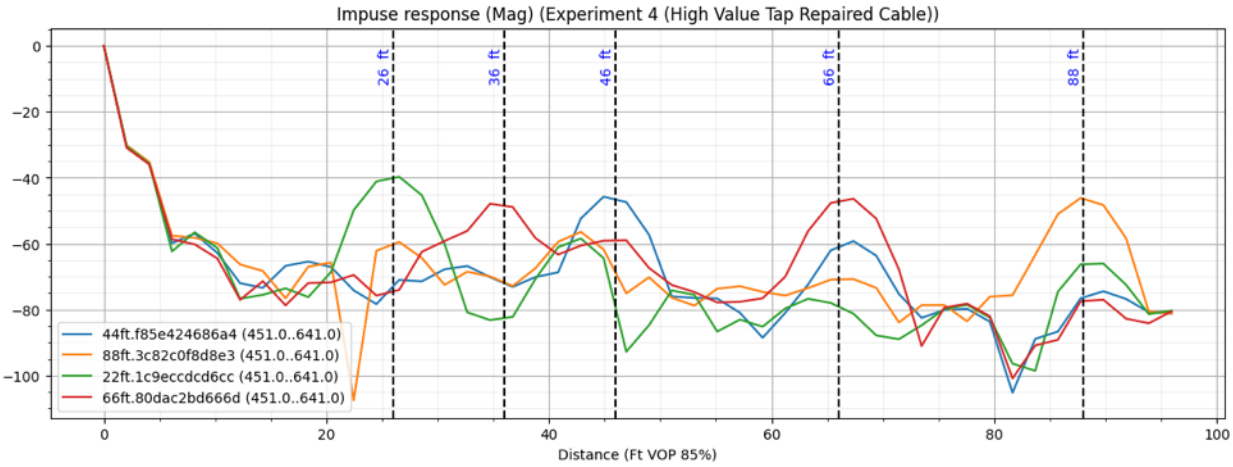


Figure 52 – Experiment 4, High-Value Tap, Repaired, Time-Domain Impulse Response, 100 Feet

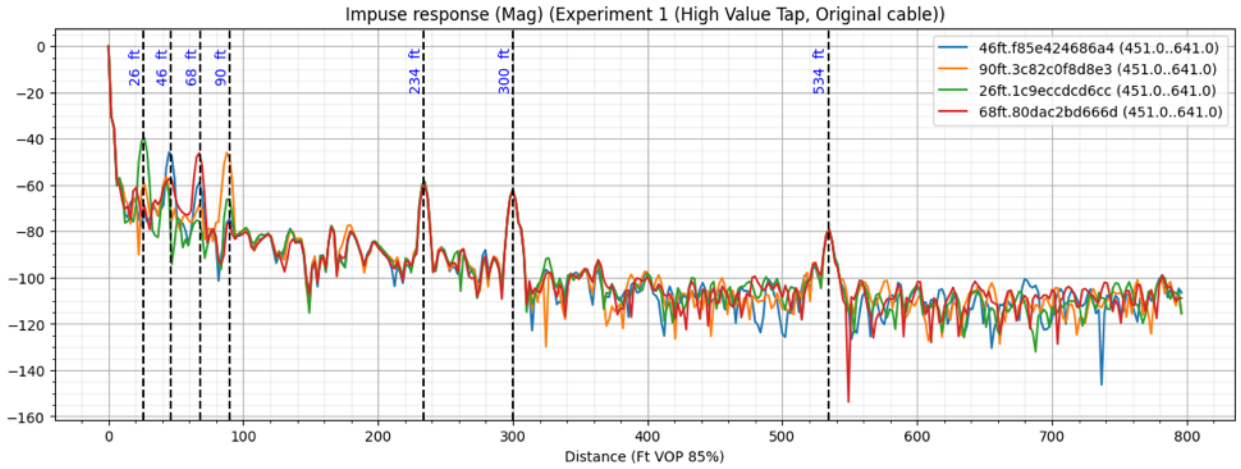


Figure 53 – Experiment 1, High-Value Tap, Time-Domain Impulse Response, 800 Feet

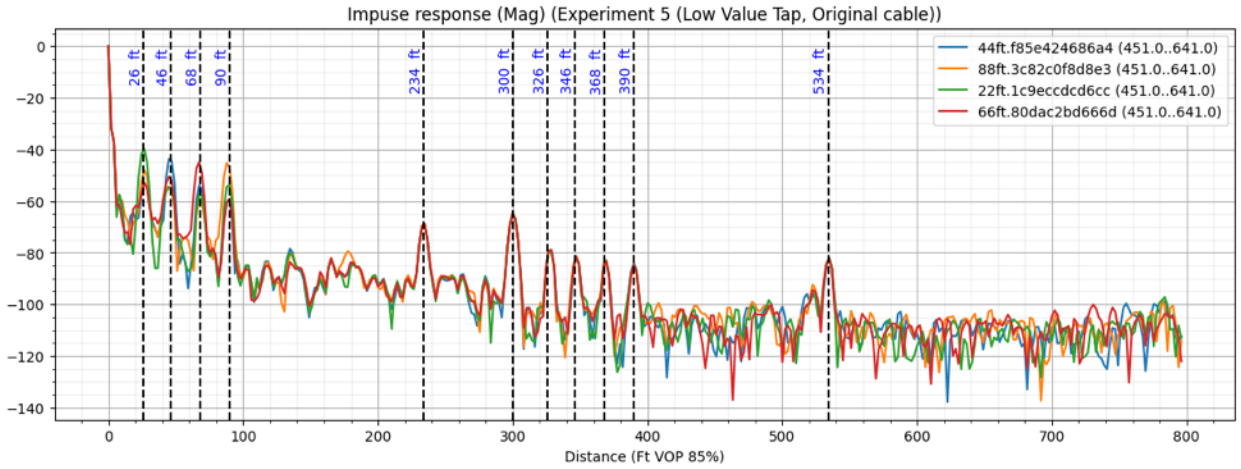


Figure 54 – Experiment 5, Low-Value Tap, Time-Domain Impulse Response, 800 Feet

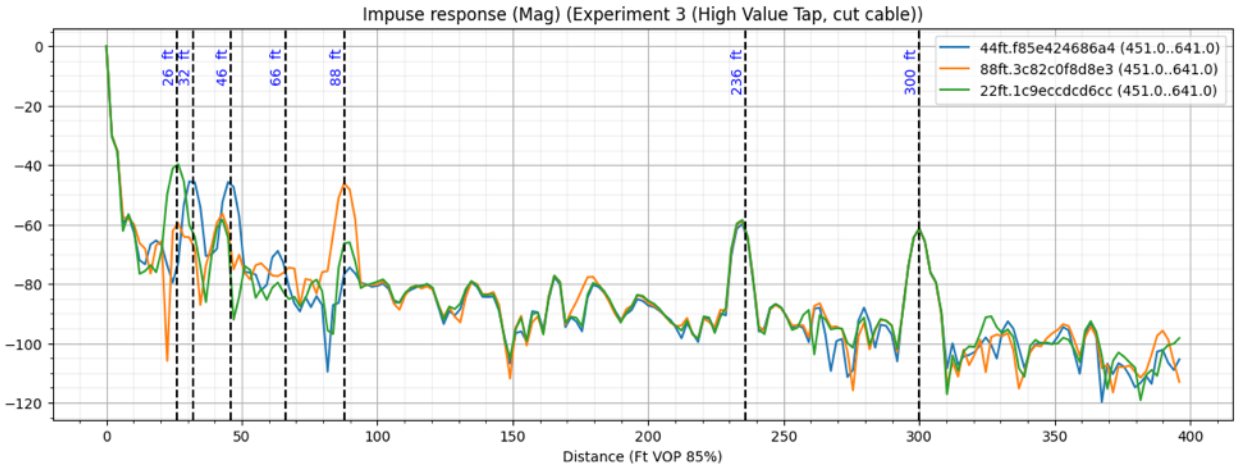


Figure 55 – Experiment 3, High-Value Tap, Cut Cable, Time-Domain Impulse Response, 400 Feet

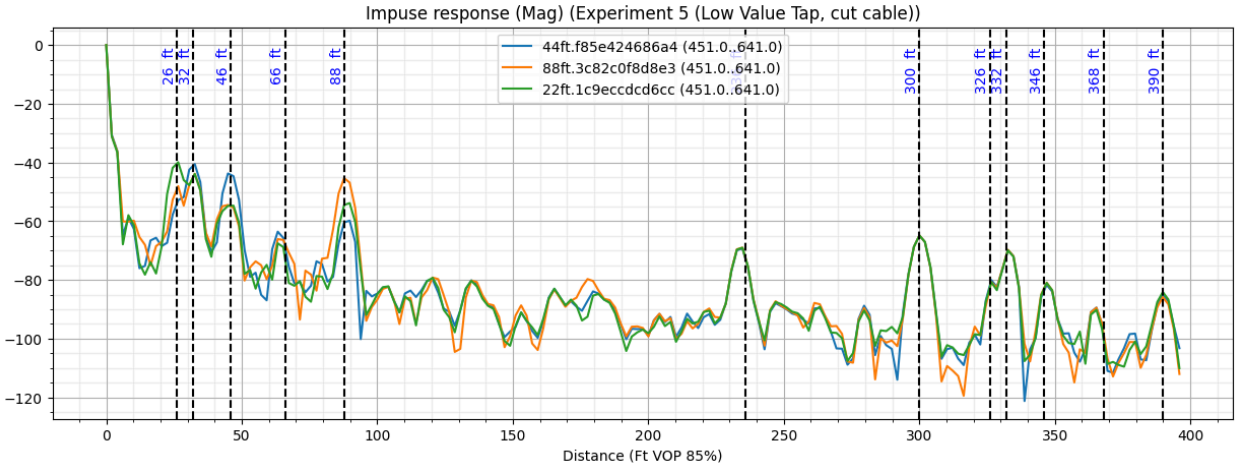


Figure 56 – Experiment 5, Low-Value Tap, Cut Cable, Time-Domain Impulse Response, 400 Feet

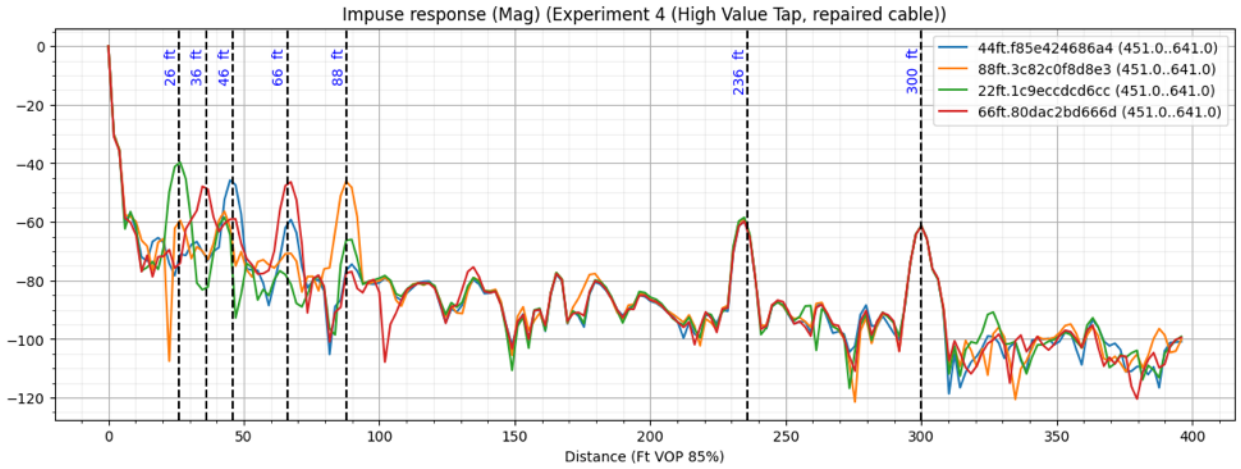


Figure 57 – Experiment 4, High-Value Tap, Repaired, Time-Domain Impulse Response, 400 Feet

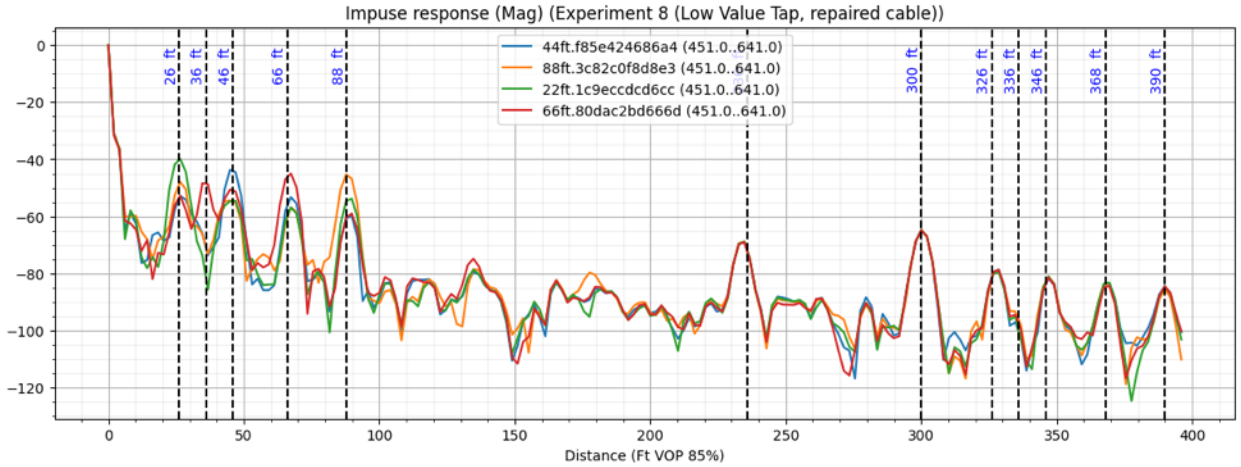


Figure 58 – Experiment 4, Low-Value Tap, Repaired, Time-Domain Impulse Response, 400 Feet

Figure 59 – Experiment 3, High-Value Tap, Terminated, Time-Domain Impulse Response, 400 Feet

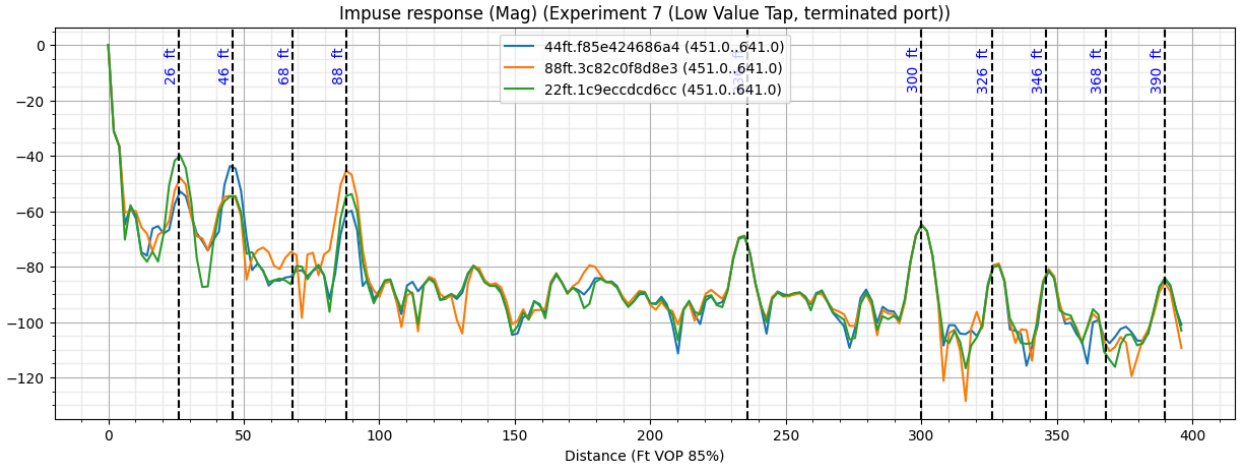


Figure 60 – Experiment 3, Low-Value Tap, Terminated, Time-Domain Impulse Response, 400 Feet

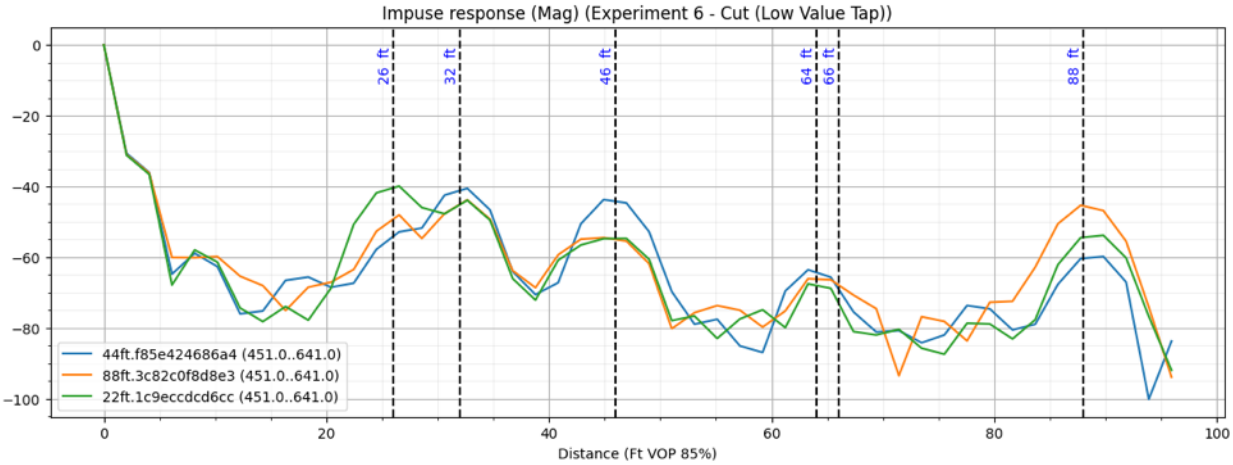


Figure 61 – Experiment 6, Low-Value Tap, Cut Cable, Time-Domain Impulse Response, 100 Feet

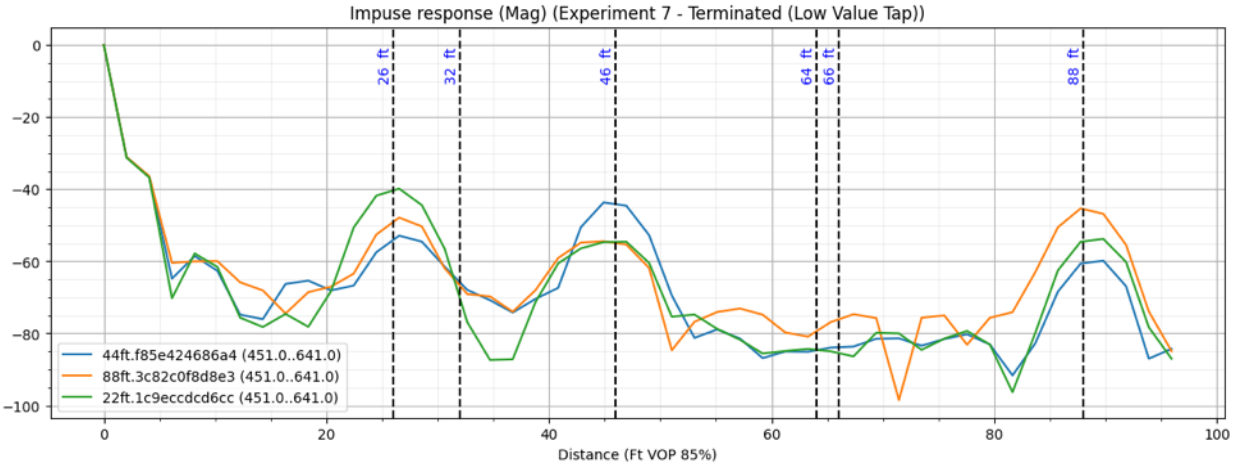


Figure 62 – Experiment 7, Low-Value Tap, Terminated, Time-Domain Impulse Response, 100 Feet

6. Conclusions

These experiments successfully demonstrated the capabilities of DOCSIS 3.1 PNM using OFDM channel estimation to detect, localize, and characterize impedance mismatches in coaxial cable networks. The key findings support the goals outlined in the introduction:

1. Accurate Measurement of Drop Cable Lengths and Connections: The impulse response data enabled precise identification of drop cable lengths and modem connection points. For example, ports 2 and 3 (26 and 90 ft) and ports 4 and 5 (46 and 68 ft) on the 4-way taps exhibited characteristics consistent with 2-way splitter behavior, confirming accurate length estimation.
2. Detection and Localization of Cable Cuts: The presence of a cut or unterminated drop cable was clearly visible in the impulse responses of nearby ports, particularly on low-value taps with limited port-to-port isolation. This confirms the system's ability to detect and localize faults.
3. Identification of Splices in Drop Cables: A repaired drop cable containing an F-81 splice was detectable by the connected modem. The reflection from the modem to the splice was clearly observed, validating the ability to identify and locate splices within the drop.
4. Characterization of Tap Spans and Return Loss: The experiments revealed significant second-order reflections (common transit) corresponding to the 300+ ft drop length, observable across all modems on the low-value tap. This provided insight into tap span behavior and return loss characteristics.
5. Evaluation of Tap Port Isolation: The high-value tap demonstrated superior port-to-port isolation compared to the low-value tap. This was evident in the reduced visibility of reflected energy across ports, reinforcing the importance of isolation in network diagnostics.

These results illustrate how OFDM channel estimation can be leveraged for PNM, enabling operators to interpret reflections meaningfully and maintain network integrity with high sensitivity and distance resolution.

7. Abbreviations and Definitions

CM	cable modem
CTO	chief technical officer
dB	decibel
DOCSIS	Data-Over-Cable-Service-Interface-Specifications
IFFT	inverse fast Fourier transform
kHz	kilohertz
NOS	Network Operations Subcommittee
OFDM	orthogonal frequency-division multiplexing
PNM	proactive network maintenance
SCTE	Society of Cable Telecommunications Engineers
TDR	time domain reflectometer
VF	velocity factor
VoP	velocity of propagation
WG	working group

8. Bibliography and References

[1] OFDMA Predistortion Coefficient and OFDM Channel Estimation Decoding and Analysis Remove the linear delay, examine the group delay, Tom Williams et al, SCTE CableTec Expo 2021

[2] The Proactive Network Maintenance (PNM) Comeback, OFDM and OFDMA Bring New, Laser-Guided Precision for Plant Fault Distancing, Wolcott et al, SCTE CableTec Expo 2024

Click here to navigate back to Table of Contents

Revolutionizing Part-Part Similarity with LLM-Enhanced Embeddings

A Technical Paper prepared for SCTE by

Wei Cai, Senior Lead Data Scientist, Cox Communications, SCTE Member
wei.cai@cox.com

Le Vo, Sr Manager of Architecture, Cox Communications
Le.Vo@cox.com

Uki Molina, Lead Architect, Cox Communications
Uki.Molina@cox.com

Alex Vang, Data Analyst II, Cox Communications
Alexander.Vang@cox.com

Table of Contents

List of Figures

List of Tables

1. Introduction

Since the late 1990s, the number of part numbers in telecommunications supply chains has expanded significantly. This growth has been driven by evolving network architectures, transitions from legacy systems, and shifting business requirements. Today, many operators manage thousands of unique part numbers across Inside Plant and Outside Plant categories to support diverse regional markets. These parts include critical components such as pedestals, conduits, fiber jumpers, power supplies, and other distribution materials essential to network deployment and maintenance.

A major challenge lies in the proliferation of functionally similar parts sourced from multiple vendors, often without a clear one-to-many relationship or substitution logic. For example, a pedestal may be sourced from different manufacturers — each performing the same function but treated as distinct Stock Keeping Units (SKUs), which serve as a standardized references for inventory tracking and logistics. The absence of a unified cross-reference or interchangeability system leads to duplication of part numbers, excess inventory, missed substitution opportunities, and inconsistent forecasting. Additionally, as network projects evolve in real time, architecture changes frequently occur mid-execution, leaving the supply chain with overages, shortages, or stranded inventory. This fragmented part number system contributes to procurement inefficiencies, deployment delays, and significant annual write-offs.

Artificial intelligence (AI) refers to the capability of machines to perform tasks that typically require human intelligence, such as pattern recognition, decision-making, and language understanding. Machine learning (ML), a subset of AI, enables systems to learn from structured and unstructured data to improve performance over time without explicit programming. To address these challenges, our use case proposes leveraging AI and ML to analyze part descriptions, technical specifications, and usage data to identify and consolidate functionally equivalent parts across the enterprise. This approach aims to standardize components across markets, enable intelligent substitution logic for field execution, and support the creation of pre-defined kits tailored to specific job types. By improving lifecycle visibility and enabling strategic sourcing through vendor rationalization, organizations can enhance forecasting accuracy and reduce operational complexity.

Ultimately, an AI-driven approach to part simplification offers a significant opportunity to reduce excess inventory, minimize write-offs, and streamline execution across platforms. By transforming the supply chain to operate more like a factory floor, telecom providers can improve operational efficiency, reduce costs, and better support the rapid pace of network deployment across markets.

The paper aims to illustrate how the integration of Sentence Transformers and Large Language Models (LLMs) can transform the management of network inventory data (e.g., part descriptions) in the cable industry. The analysis will underscore the significant advantages of this advanced semantic method over traditional Fuzzy String and pattern matching techniques.

Our paper makes two key contributions: Firstly, we add to the expanding body of literature on AI and LLMs by introducing a two-tier AI agent-based workflow to the cable industry. This contribution is particularly significant for the cable industry to stay ahead in AI adoption. Secondly, while AI and LLMs have been increasingly adopted in other areas of the cable industry, their application in network supply chain and inventory management has been limited. Thus, our work not only enriches the broader literature on employing AI to solve business challenges more efficiently but also pioneers new ways to leverage AI for improved inventory management in the context of the cable industry.

The structure of this paper is as follows: Section 2 traces the progression from traditional fuzzy string matching to Sentence Transformers and LLMs focusing on how each step has improved semantic understanding. Section 3 introduces the Two-Tier Mapping workflow and explains how it is implemented in practice. Section 4 offers a detailed comparative analysis, examining the advantages of combining sentence embeddings with a language model when working with real-world data. Finally, Section 5 reflects on the broader implications of the paper and provides a succinct conclusion.

2. Related Work

In this section, we delve into contemporary research on fuzzy string matching, with a particular focus on text classification using sentence transformer embeddings and LLMs. In light of the research gap made from text mapping, we highlight the key issues, thoughts, and methods that affect how we understand this research.

The rapid evolution of network technologies and the growing number of parts and devices pose a substantial challenge for traditional techniques such as fuzzy string matching to continue working effectively in inventory management systems, as textual descriptions are becoming more noisy or inconsistent. While commonly used, those popular techniques are limited in capturing the subtleties of language. In this approach, algorithms like *Levenshtein* distance, *Jaccard* similarity, and *N-gram* models are used to evaluate text by comparing edit distances, token overlaps, or character sequences and character patterns. Although useful for linking records and organizing information, they only focus on how words are spelled or structured rather than what they conceptually represent. This can be problematic in areas like telecommunications, where terms may have identical meanings but different wording (e.g., “EQ” vs “EQUALIZER”, “AMP” vs “AMPLIFIER”, “ATT” vs “ATTENUATOR”, “Attenuator” vs “PAD”, “Gain” vs “Amplifier”) or same technical function with different labeling (e.g., “F-CONNECTOR TERMINATOR, 75 OHM” vs “CABLE TERMINATION DEVICE”, “END OF LINE CAP, 75OHM” vs “F-TYPE TERMINATION PLUG”).

New methods have emerged to address the limitation of character-based techniques used in Fuzzy String matching, in text classification such as Sentence Transformer models, derived from Bidirectional Encoder Representations from Transformers (BERT), a deep learning model developed by Google for natural language understanding. In contrast to look-based embeddings like Word2Vec, these change forms are ideal for applications in semantic textual analysis [3]. With tuning techniques such as learning and training with the datasets of Natural Language Inference (NLI) languages, they can learn the differences between the sets of functionally different, yet linguistically similar descriptions of sets.

LLMs further enhance this capability by extracting information, classifying it accurately while also facilitating scalable analysis. Models such as GPT and LLaMA have been proven very powerful for their ability to adapt well and conduct contextual reasoning of text data through their pre-training on extensive datasets [1],[2]. Those models can be successfully used to enhance the classification of part descriptions in network inventory management, gather useful information from unstructured text, and result in the classification tasks that extend beyond conventional methods.

Using sentence embeddings along with LLMs can create a foundation for handling intricate text comparison assignments in network inventories effectively. This approach involves incorporating matching and contextual reasoning alongside sense similarity assessments to surpass the limitations of matching methods, in terms of precision and practicality – in scenarios where part descriptions from different vendors show significant distinctions in terminology, formatting, and specificity.

3. Methodology

In this section, we'll delve into great details about the Two-Tier AI Mapping workflow (Figure 1). This cutting-edge analytical framework combines rule-based pattern matching, transformer-based semantic modeling and LLM enhancements. We design this workflow to strike a balance, between efficiency and analytical depth in inventory data to overcome the limitation of individual methods through careful integration of different approaches.

In Tier 1, we combine transformer embeddings with rule-based pattern recognition to identify attributes like colors and technical details in part descriptions efficiently and reliably. While pattern recognition yields consistent output, it struggles to adapt to patterns without rule adjustments or contextual understanding. For instance, descriptions might contain internal identifiers or shorthand, which poses challenges for those outside the organization (e.g., "EQ MAXNET 3.5DB GRAY" vs. "Equalizer 3.5 dB, plug-in, Maxnet, gray cover") to easily interpret. Pattern matching can also fail when descriptions use different token order or contain special punctuations ("3DB PAD PLUG-IN" ≠ "PLUG-IN PAD 3DB"). To overcome these challenges, we propose combining pattern matching with transformer-based embeddings, such as MiniLM, to effectively understand semantic relationships in part descriptions. This parallel processing approach offers the benefits of improving efficiency and precision in matching unstructured text tasks.

To further optimize distinct pattern recognition and process efficiency, we also apply an initial screening process in Tier 1 to filter out unlikely matches before using LLMs to conduct detailed contextual matching reasoning with part descriptions. This screening takes into account both pattern recognition scores and embedding similarity measures to establish confidence levels to guide the LLM deep analysis of part description. Tier 1 phase deals, with around 80% to 90% of matching situations; meanwhile the LLM conducts in-depth analysis on the logic for the remaining 20% to 10% thereby reducing the load on subsequent analytical phases significantly.

Tier 2 employs advanced language models to enhance context comprehension and assess part technical functionality effectively. Because these models have domain-specific reasoning capabilities, they can provide detailed analysis and explanations of text classification. However, they are resource-intensive and may produce misleading results if used beyond their training knowledge base. Tier 2 is specifically used when contextual understanding or domain-specific knowledge is required. Part descriptions are first evaluated through a gateway,-a filtering mechanism that determines whether a case warrants escalation to the next Tier. If a case is straightforward and Tier 1 gives a high-confidence match, it moves on without further review. But if the part description is more complicated or ambiguous, it goes through a deeper review that includes compatibility checks and context-based analysis using LLMs.

The optional integration of Retrieval-Augmented Generation (RAG) provides users with access to detailed documents and specific manufacturer information for unique situations requiring external data sources. The strategic use of these resource-intensive tools enriches the analysis by incorporating contextual insights beyond the initial data inputs. However, implementing RAG integration requires a strong and scalable infrastructure, as well as regular maintenance to ensure accuracy and relevance. Therefore, RAG is used solely as a supplementary module.

Throughout each stage, verification and review by an agent are essential to ensure the accuracy and reliability of the data presented throughout the process.

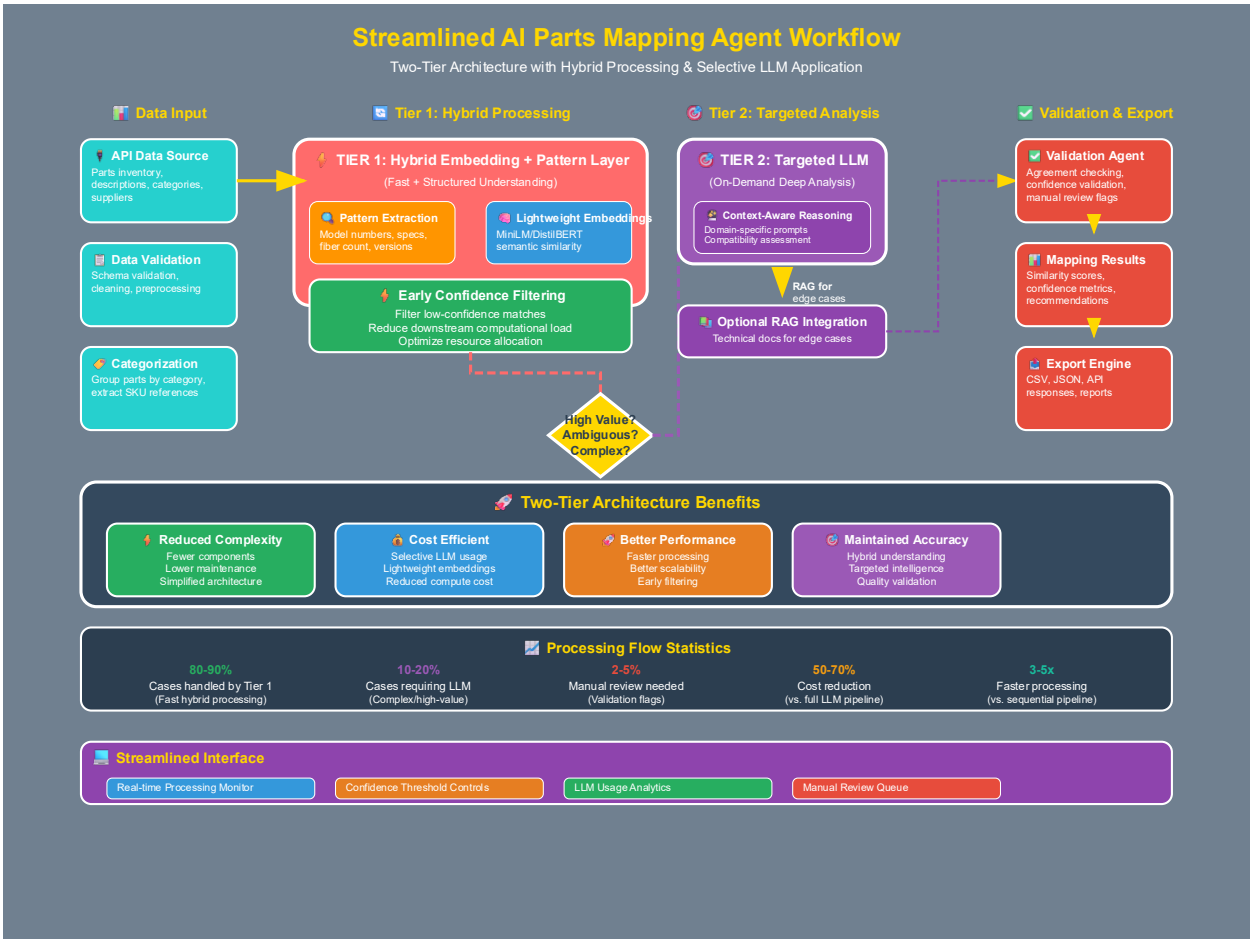


Figure 1 - Streamlined AI Parts Mapping Agent Workflow

4. Discussion

In this study, we compared three different methods to measure how similar network part descriptions are and to understand which approach works best in different scenarios. The goal was not just to find the most accurate method overall, but to explore the benefits of combining sentence embedding with LLMs when applied to real-world data.

- Fuzzy string matching
- Sentence embedding using a lightweight transformer model (MiniLM)
- A combination of sentence embedding and LLM (MiniLM + LLaMa3)

Each method views the data from a different angle—one focuses on surface-level word similarity, another captures the general meaning of the full description, and the third adds deeper contextual understanding. In the sections that follow, we break down how each method performed, where they excelled, and where they fell short. Figure 2 provides illustration on how similarity scores are derived for each method.

Figure 2 – Similarity Score by Method

4.1. Fuzzy String Matching

Fuzzy String method breaks down part names and descriptions into their textual tokens and applies various matching algorithms to uncover similarities. It assigns different weights to aspects of textual similarity. In this study, we calculate fuzzy string similarity score using the following weight allocation: basic character matching accounts for 30%, partial matches for 20%, token sorting for 25%, and token set analysis for the remaining 25%.

This approach works best when part descriptions show similar patterns and have token overlaps. For example, when comparing “Cisco RG6150FT” and “Cisco RG6100 FT”, even though they are written differently in format, a fuzzy string method can easily give a similarity score of 0.923 to suggest their close resemblance. However, it fails when it comes to match part descriptions with distinct word order, punctuation or with extra or missing stop words. It is often very sensitive to distinct words but similar functions or when abbreviations of part components are used in the descriptions. For instance, “EQUALIZER, PLUG-IN, 870MHZ, 12DB, YELLOW COVER – SA” and “EQUALIZER FORWARD 870 MHZ 12DB YELLOW COVER SA” can be lower than expected despite those two parts are effectively the same or very similar.

4.2. Sentence Transformer Embedding

Unlike Fuzzy String matching, sentence embedding encodes contextual and semantic information, and generate embeddings of a fixed size specifically designed for tasks such as text classification or matching

strings. A Sentence Transformer differs from an LLM in that it has no need for task prompts or extensive inference to compare semantic meanings. They utilize models like triplet networks to learn from sets of sentences and excel at gauging the similarity between two phrases or descriptions [3]. This approach is notably useful when dealing with terminologies, technical equivalents or explanations that convey alike concepts using different words such as "PLUG-IN" and "FORWARD" in the context of network part and devices. However, this approach might fail to realize different jargons carry the same meaning if the underlying pre-trained model is not adjusted.

4.3. Sentence Transformer Embedding + LLM

In a real-life scenario, there exists examples like the following Parts A and B, in which only an extensive LLM can identify a resemblance. Methods like string matching and sentence embeddings may not yield as optimal results, in comparison.

Table 1 - Sample Descriptions I

Part A	Part B
HOUSING, AMP500, LE 1.2GHz, w/ EQ and ATT, HARDENED	NODE SHELL, 1.2GHz, HARDENED, INCLUDES PAD/EQ, LEGACY AMP MODEL

A model AMP500 contains equalization and attenuation elements. However, they are presented in different writing styles. Description A employs abbreviations such as "AMP500" and "with EQ and ATT, whereas Description B chooses terms and subtle hints, like "NODE SHELL" and INCLUDES PAD/EQUALIZER". Between those two descriptions, "Housing" (Part A) and "Shell" (Part B) are synonymous, and "Att and Eq" (Part A) represent "Pad/Eq" (Part B). Meanwhile when we mention the "Amp500", as the Legacy Amp Model, we are delving into models that capture domain nuances and underlying connotations often overlooked by embeddings or character level comparisons.

Matching based on string logic faces challenges when distinct tokens, such as "NODE SHELL" and "HOUSING", have similar meanings but no overlapping character sequences. This lack of textual similarity makes it difficult for traditional string-matching algorithms to recognize them as related. Moreover, descriptions often vary in word order and phrasing, such as "PAD/EW" versus "EW and ATT," further complicating similarity scoring. As a result, the anticipated similarity score from string-based matching methods may fall below 50%, even when the parts could be functionally equivalent.

This low score could also happen when the sentence transformer isn't adjusted for cable or network hardware jargon. It might not realize that "AMP500" and "LEGACY AMP MODEL" carry the same meaning. This discrepancy in terminology could be the issue causing an anticipated SBERT similarity score range of 0.70– 0.80 or possibly even lower based on the model being used.

Unlike Sentence transformers concentrating on using embeddings for specific tasks such as semantic searches or comparison, LLMs are trained to understand language broadly by learning patterns, grammar, and meaning from massive text datasets. They excel at understanding contexts and specialized language nuances which helps in distinguishing components like "NODE SHELL" and "HOUSING". LLM can also excel in identifying key differences in part descriptions even when they appear alike. For instance, there exists a pair of part descriptions as follows:

Table 2 - Sample Descriptions II

Part A	Part B
EQUALIZER,FORWARD,862MHZ,2DB,SEQ-862-02 SERIES,DIGICOMM - C-COR	EQUALIZER,FORWARD,862MH916DB,SEQ-862-16 SERIES,DIGICOMM - C-COR

On the face of it these appear alike. Both are forward equalizers, at 862MHz from the manufacturer and product line which may be seen as highly similar, by Sentence Transformers. However, LLMs can spot the disparities such, as 2dB versus 16dB and 02 versus 16 series as they can identify these as notable technical differences rather than interchangeable elements.

Hence, we propose utilizing a mix of these two approaches and examining part details. With complex part descriptions, LLMs provide language intelligence while Sentence Transformers convert it into quick and practical similarity ratings to enhance the accuracy and scalability of string matching in world inventory and network system applications.

The results in Figure 3 clearly show that traditional string-matching methods fail to capture the complexity of technical part descriptions, returning near-zero similarity scores across the board. Semantic methods, by contrast, proved far more effective. While MiniLM embeddings alone delivered solid performance, integrating an LLM boosted accuracy by an additional 6.6%, achieving the highest overall mean similarity score (0.626). Importantly, nearly 40% of semantic comparisons reached high-confidence thresholds (>0.7), making them strong candidates for automation.

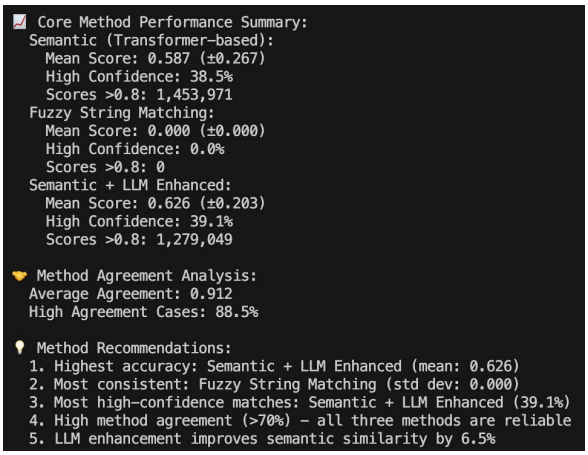


Figure 3 – Performance by Model

The data also revealed strong consistency between the semantic models, with high correlation and low variance across different part categories. Based on these results, the recommendation is to fully adopt the Semantic + LLM approach for production use, discontinue fuzzy string matching, and reserve manual review for lower-confidence cases.

5. Results and Conclusion

In this paper we exploit the hybrid of pattern matching, sentence transformer embedding and LLMs to learn word-level and sentence-level information of part description. Our study demonstrates the superiority of this method and concludes that sentence embeddings with an LLM output conventional fuzzy string matching and greatly increased the accuracy of mapping part description effectively.

The two-tier architecture offers a balanced method of managing part matching workflows, achieving significant performance improvements while maintaining analytical depth. Using LLMs only in the most ambiguous or high-value situations, which are typically impacted by 10% to 20% of the data, reduces computational demands by 50% to 70%. Compared to sequential designs, early filtering and parallel processing increase throughput often results in a 5x to 10x speed increase compared to fully sequential designs.

Efficiency and strategic complexity are key components of this architecture's design. The first tier combines rule-based pattern extraction with lightweight transformer embedding models. Simultaneously, the design allows for high precision with an LLM where it is genuinely necessary.

The LLM acts most importantly for situations requiring deeper understanding—such as those with inconsistent terminology, missing context, or non-obvious functional relationships. Due to its ability to effectively reason across domain-specific language, identify compatibility, and make nuanced decisions, deterministic methods cannot independently produce this functionality. It bridges gap between human intuition and automated processing. In short, the LLM is not just an extra layer. It plays a high-impact role in identifying key differentiators in part descriptions. The architecture calls on LLMs when deeper understanding is needed.

Scalability is another feature we want to highlight with this layered architecture. Since most of the workload is maintained by Tier 1, projected resources turn simple. High-compute capacity may be located and extended based on the number of LLM-qualified cases rather than for the overall input volume.

Finally, to ensure accuracy, a validation layer evaluates the alignment between outputs from both tiers and assigns a confidence score. New matching patterns identified during processing are continuously added to the knowledge base. This not only helps improve future matching performance but also enables the system to flag low-confidence results for further review. As a result, the system becomes more reliable and adaptive over time, even as the complexity and variability of the input data increase.

6. Abbreviations and Definitions

AI	artificial intelligence
AMP	amplifier
ATT	attenuator
EQ	equalizer
LLM	large language model
RAG	retrieval augmented generation
SKU	stock keeping units

7. Bibliography and References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, 33, 1877–1901. https://arxiv.org/abs/2005.14165

[2] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint. https://arxiv.org/abs/2307.09288

[3] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint. https://arxiv.org/abs/1908.10084

Click here to navigate back to Table of Contents

SCTE STANDARDS JOURNAL
SCTE a subsidiary of CableLabs

Matching Passive Thermal Management Solutions for Outdoor Telecommunication Cabinets to the Variable U.S. Climate Zones

A Technical Paper prepared for SCTE by

Isaac Bua
https://www.linkedin.com/in/bua-isaac/

Zach Taylor
https://www.linkedin.com/in/zachary-taylor-58016946/

Kamil Aliyev
https://www.linkedin.com/in/kamil-aliyev-b98156222/

Queen Okon
https://www.linkedin.com/in/okonqueen/

Abolfazl Baloochiyan
https://www.linkedin.com/in/abolfazl-baloochiyan-a69a17208/

Bolape Alade
https://www.linkedin.com/in/bolape-alade-3a214a10a/

Villanova University
800 East Lancaster Ave.
Villanova, PA 19085
610-519-4500

Acknowledgement

This project was completed by a team of six Villanova students, four in the Sustainable Engineering (SUSE) program and two in the Mechanical Engineering program. This project was conducted as part of the SUSE program's Resilient Innovation through Sustainable Engineering (RISE) initiative. The completion of this project was supported by Villanova faculty advisor, Karl Schmidt, and by the Society of Cable Telecommunications Engineers (SCTE). Thank you to SCTE and its partners for all their guidance and support throughout this project, with special thanks to Derek DiGiacomo - SCTE Senior Director, Energy Management Program & Business Continuity SCTE, Scott Kenny Technical Product and Program Management at Cable labs and Margeret Bernroth Program Manager at SCTE.

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Telecommunication Outdoor Cabinets

Outside plant telecommunication cabinets are crucial elements of the telecommunication infrastructure, acting as key nodes that connect and protect network components. These enclosures house active electronic equipment and are designed to safeguard from elements such as moisture, extreme temperature, dust, and corrosive agents. Because they are often installed in remote areas, these cabinets must be designed for durability and performance to avoid constant monitoring and maintenance. While relatively small, telecommunication cabinets make up the largest portion of the cable broadband industry's power footprint, consuming between 44 – 54% of total energy used by the industry (figure 2) [1]. This means that cabinets are subsequently responsible for a large portion of the industry's scope 2 greenhouse gas emissions, caused indirectly from the purchase of electricity [2].

One of the key technical challenges related to the energy required to power these cabinets is maintaining the required temperature range for operation within the cabinets. The outside air temperature, the solar load, heat generated by both the electronic components and cooling equipment, all contribute to elevating the internal temperature. Left unchecked, the range can quickly exceed the safe operating threshold for the IT equipment. Different government agencies, including the US Labor Department's OSHA (Occupational Safety & Health Administration) and the Consumer Product Safety Commission (CPSC), set standards for safe working or operating temperatures. More notably, general equipment manufacturer temperature limits operation to be low 46 °C [3] Exceeding this limit can lead to performance degradation, equipment failure, costly service disruptions, and compromise network reliability.

Figure 1- Sample of Telecommunication Outside Plant Cabinet [1][4]

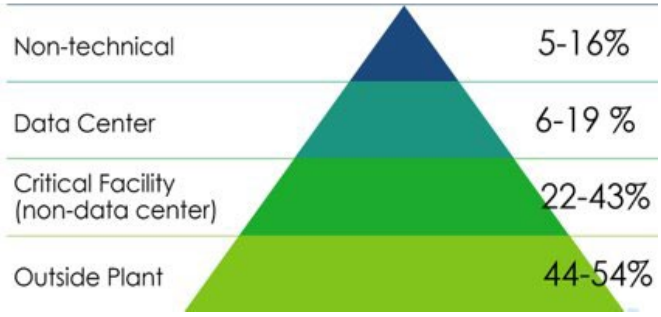


Figure 2 - Cable Broadband Industry Power Footprint [1]

1.2. Overview of Thermal Management in Telecommunication Outdoor Cabinets

The thermal regulation of outside plant cabinets has typically relied on two main technologies: cooling fans and air conditioning units. Fans function by facilitating the exchange of internal cabinet air with the outside ambient air, pushing out heat in the process. However, this method is inherently limited by environmental conditions—fans cannot lower cabinet temperatures below that of the surrounding air [5]. Consequently, these technologies are most effective in cooler climates or when internal heat exceeds external temperatures. Furthermore, fans offer minimal protection against environmental debris such as dust, moisture, or corrosive particles. These limitations highlight the growing need for more adaptable, sustainable, and efficient thermal management technologies capable of meeting evolving performance and environmental demands.

Air conditioners, by contrast, use a closed-loop refrigeration cycle involving refrigerants, compressors, evaporators, and condensers sometimes supplemented by fans to achieve more precise control. This allows for cabinet cooling even when ambient outside temperatures are high, making air conditioners the preferred choice in hotter environments to maintain the required temperature thresholds. Despite their effectiveness, air conditioners feature some drawbacks. They have high energy consumption, contributing to increased operational costs and, depending on the power source, higher greenhouse gas emissions. Their size and configuration also limit installation flexibility. Typically mounted externally on cabinet walls, these units require careful attention to monitor refrigerant leakage, which can lead to environmental harm such as damage to the ozone layer. Additionally, air conditioners require regular maintenance and monitoring to ensure optimized performance [6].

Both fans and air conditioners are referred to as active cooling solutions, meaning that they consume energy to cool their intended targets. In addition, it should be noted that these technologies produce heat as a byproduct of their operation, driving additional energy consumption. In contrast to active cooling solutions, passive cooling solutions are those that require no energy to achieve their desired cooling effect. Passive cooling solutions today are often used in support of active cooling solutions, especially within the realm of electronics [6].

1.3. Villanova University & SCTE RISE Partnership

The Society of Cable and Telecommunication Engineers (SCTE) is a nonprofit organization facilitating the advancement and deployment of technology, standards, as well as professional development education within the industry. SCTE's maintains a partnership with Villanova University's graduate Sustainable Engineering program by way of the Resilient Innovation through Sustainable Engineering (RISE)

initiative. The RISE program allows engineering students to contribute and work towards various company and industry sustainability goals by conducting research on specific topics [7]. This partnership has long been a focal point for the RISE program, with projects being highlighted biannually at Villanova's academic RISE forum.

2. Objectives

2.1. Goal and Scope

This project was conducted as a continuation of a previous RISE project with some members overlapping. The previous project featured a literature review of both active and passive solutions and was also featured in a previous SCTE Technical Journal [6]. The primary goal of this project was to identify, evaluate, and recommend effective passive thermal management solutions for outdoor telecommunication cabinets that could be deployed across diverse U.S. climate zones. As telecom infrastructure increasingly supports critical digital connectivity, ensuring cabinet thermal stability without excessive reliance on active cooling is essential for both system reliability and energy usage. This initiative involved a literature review of technical papers on the topic of passive cooling technologies, focusing on methods that minimize external heat gain and enhance internal heat dissipation.

Key components of the project scope included.

- Comprehensive literature review, synthesizing existing research and case studies on passive thermal control methods applicable to telecom enclosures.
- Quantitative benefit analysis by estimating potential temperature reductions and associated energy savings when passive strategies are applied, leveraging data modeling and climate-based simulations.
- Screening criteria development based on multi-criteria evaluation framework to assess each technology's performance, feasibility, and adaptability.
- Climate zone mapping aimed at aligning the most suitable passive cooling technologies with specific U.S. climate zones, plus recommendations based on regional environmental conditions such as temperature extremes, humidity, and solar radiation intensity.

2.2. Approach

This study began by investigating different passive cooling methods (not previously covered) for controlling thermal loads that could be utilized by outdoor telecommunications enclosures. In addition to evaluating thermal performance, each solution (found by our group and the previous) was assessed and paired with ASHRAE climate zones for recommended testing and implementation. The pairing of technology to climate zone represents the areas of the country that have the potential to produce the highest energy savings upon implementation of the cited technology.

2.3. Methodology

The passive thermal management solutions previously covered were thermosyphons, radiation shielding, double wall enclosures, and thermochromic coating. Passive solutions do not require energy to cool, however can often be paired with and used to support active solutions (for example, thermosyphons will often be paired with fans). In our research, passive solutions were not omitted from our scope if they were paired with active solutions, so long as they were able to net energy savings.

As a continuation, papers were identified and compiled from various academic journals and ranked on the applicability and potential to implement in a telecommunication cabinet (rankings being not likely applicable, could be applicable, and applicable).

Compiling nearly fifty academic papers and resources, our group identified developing and trending technologies from the readings and chose those that appeared to have high applicability and likelihood to net energy savings for telecommunication cabinets. After refining the number of sources, the group explored three different passive cooling techniques: designing for heat transfer, heat pipes, and radiative daytime cooling. From there, approximate ranges of energy savings from our readings were compiled to summarize conservative estimates.

3. Passive Cooling Methods

3.1. Designing for Heat Transfer

The first passive cooling technique the group came across was like that of the previously identified double wall enclosures. Meaning not a specific technology but an approach and methodology on how to design the layout and structure of a telecommunication cabinet to optimize the flow of heat. Recent experimental findings by Jobby et al. provide valuable insights into how internal configuration particularly spacing and orientation affects convective heat transfer [8].

3.1.1. Vertical (90°) Inclination

Jobby et al. experimented with modifying the orientation of heat generating components within a small cabinet. These fabricated heat generating components served as stand-ins for the typical electronic equipment used in a telecommunication cabinet and produced heat at a rate over a surface area comparable to the typical functioning equipment [8].

Among all inclination angles tested, the 90° vertical orientation emerged as the most viable for industrial adoption in our opinion. This configuration, which aligns with standard vertical rack-mounting practices, yielded a 9.99% increase in heat transfer under natural convection compared to horizontal (0°) mounting. This improvement is attributed to the alignment of buoyancy-driven airflow within the vertical gaps between components, minimizing stagnant zones and enhancing upward heat dissipation. Importantly, the 90° orientation also improved airflow over the larger heated surfaces of the stand-in components, thus more effectively allowing that heat to dissipate and not accumulate between components [8].

Unlike intermediate inclinations between 30°– 60° (which produced the highest heat transfer improvement), which may necessitate custom mechanical solutions, vertical mounting is presumably compatible with existing cabinet infrastructure and more typical arrangements. It offers a practical and low-risk path to improving passive heat transfer, with no need for additional power [7]. These findings could be implemented to change the most common arrangement during initial fabrication. That is to say that components are placed vertically (lengthwise) in the direction of the flow of heat (bottom to top). How easily internal components of an already in use cabinet can be changed and result in energy savings remains to be seen.

3.1.2. Stream-Wise Spacing

A recurring finding in the same study is the strong positive correlation between increased stream-wise spacing and convective heat dissipation. Jobby et al. confirmed that increasing vertical spacing between

components from 29 mm to 108 mm improved the heat transfer (Nu) by 25.9% under natural convection and 17.8% under forced convection. These results are consistent with prior numerical and experimental work [9][10] which attribute the improvement to reduced flow recirculation and enhanced air throughput between heated surfaces. Notably, the benefit diminishes beyond a certain threshold, suggesting that excessive spacing yields reduced marginal gains [8]. Nonetheless, moderate increases in stream-wise spacing represent a cost-effective passive cooling enhancement readily adaptable to standard cabinet designs.

Figure 3 - Control Arrangement, 90° Rotation, and 60° Rotation [8]

3.2. Heat Pipes

Heat pipes are passive thermal devices designed to transfer heat with extremely high efficiency by leveraging the principles of phase change and capillary action. Structurally, a heat pipe consists of a sealed tube commonly made from copper or aluminum containing a small amount of working fluid such as water, ammonia, or acetone. The inside of the pipe typically includes a wick structure, often composed of sintered metal, screen mesh, or grooved surfaces, which enables the capillary return of liquid to the heat source.

When one end of the heat pipe, known as the evaporator section, is exposed to a heat source such as equipment inside a telecommunication cabinet, the working fluid inside absorbs thermal energy and evaporates. The vapor then travels along the hollow core of the pipe to the cooler end, called the condenser, where it releases its latent heat to the surroundings and condenses back into liquid form. The wick structure then draws the liquid back to the evaporator through capillary action, thereby completing a continuous thermodynamic cycle without the need for any mechanical component [11][12].

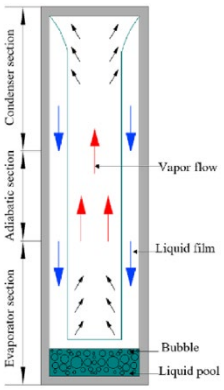


Figure 4 - Heat Pipe Schematic [12]

In the context of outdoor telecommunications cabinets, this technology is typically implemented in the form of air-to-air heat pipe exchangers. These systems operate in a closed-loop configuration, meaning that ambient outside air never enters the cabinet. Instead, hot internal air is circulated across the evaporator fins of the heat pipe exchanger. The heat is absorbed and transferred through the phase change process inside the pipe and then expelled to the external environment via condenser and an external airflow. The only moving parts in this configuration are small axial fans, typically rated under 100 watts, which circulate air across the condenser sections of the group of heat pipes. This closed-loop design is especially important in telecom applications because it isolates sensitive electronics from dust, humidity, and other outdoor contaminants that would otherwise compromise performance or lifespan [11][12].

The primary advantage of using heat pipes is their high thermal efficiency and passive operation. Because no compressor or refrigerant cycle is required, energy consumption is dramatically lower than in conventional air conditioning systems. In favorable ambient conditions—specifically when outdoor air is cooler than the internal—the heat pipe system alone can maintain cabinet temperatures within acceptable operating limits. This translates to significantly lower operating costs and less wear on mechanical components, which also improves system reliability and reduces maintenance needs [11][12].

However, heat pipes have an inherent limitation: they cannot cool a cabinet to a temperature below that of the outdoor environment. Therefore, in regions or seasons where outside temperatures exceed safe thresholds for telecommunications equipment, a heat pipe system on its own may not be sufficient. To overcome this, cooling configurations use a hybrid approach, combining a heat pipe exchanger as the primary cooling method with a supplemental compressor-based air conditioner that activates only during periods of excessive external heat. In such hybrid systems, the heat pipe carries the cooling load for most of the year, while the air conditioner provides backup during peak heat events. This integrated strategy optimizes energy efficiency while ensuring the required thermal range [12]. Materials used in heat pipe systems are selected for both thermal performance and environmental durability. Copper and aluminum are standard choices for the pipe body and fins, while the wick structures are designed with sintered metals of a matching material. Because the system is sealed and contains no external connections, it is particularly well-suited for remote, outdoor applications where maintenance access is limited, and reliability is critical [11], [12].

To address the rising energy demands of outdoor communication cabinets, particularly in the context of 5G infrastructure, Cui et al. constructed a hybrid cooling system that integrated heat pipes to an already in-use cabinet air conditioning system. The heat pipe functions as the primary cooling mechanism, with

the air conditioner serving as a secondary backup. Field tests conducted over one year in Zhengzhou, China demonstrated that this hybrid system significantly lowered internal cabinet temperatures reducing high temperature alarms and achieved a substantial annual cooling energy savings of 60% when compared to traditional systems [11]. The findings indicate the system's capability to maintain compliance with national thermal management standards while providing a scalable and energy-efficient solution for telecommunications operators.

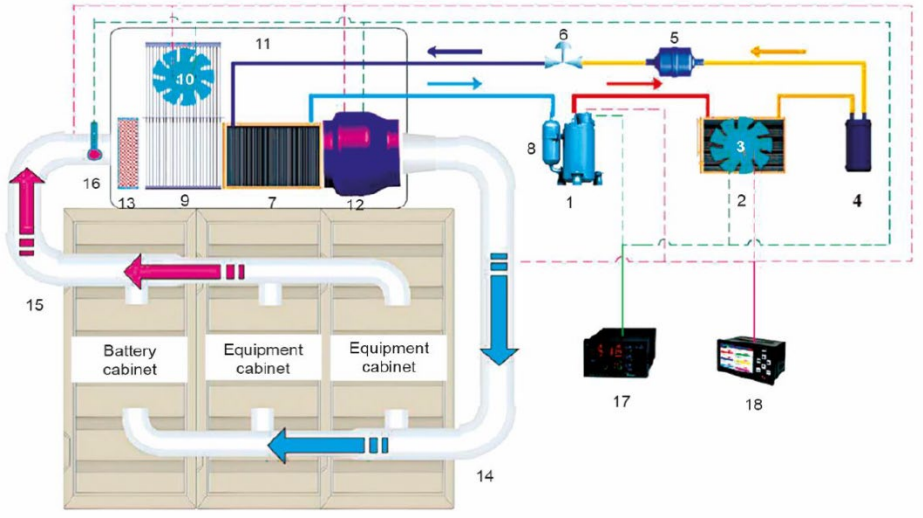


Figure 5 - Working Principle of Temperature Control Equipment for Outdoor Cabinets [11]

1 – air conditioning compressor, 2 – condenser, 3 – condensing fan, 4 – reservoir, 5-filter drier, 6 – expansion valve, 7 – evaporator, 8 – gas-liquid separator, 9 – heat pipe, 10 – heat pipe cold end fan, 11 – equipment integration box, 12 – tube guide fan, 13 – filter dust screen, 14 – air supply pipe, 15 – return air duct, 16 – temperature sensor, 17 – controller, 18 – data collector

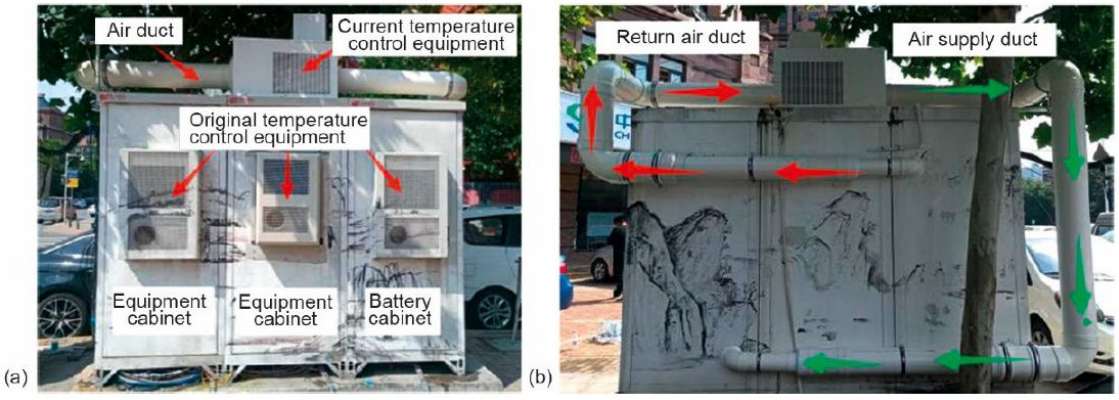


Figure 6 - Physical Site Diagram of Temperature Control Equipment (a) Front View (b) Back View [11]

In a separate study, Cui et al. further extended the investigation into hybrid cooling by developing and testing with an integrated heat pipe and air conditioning unit for another 5G communication cabinet. This system employed a heat pipe configuration for primary cooling and a vapor-compression air conditioner for supplemental cooling as well. The system demonstrated reliable thermal regulation across all seasons, maintaining temperatures within the limits defined by Chinese, ETSI, and U.S. standards. It delivered significant energy savings up to 89.9% in winter and 68.5% annually while reducing the annual Power Usage Effectiveness (PUE) from 1.67 to 1.21. With a payback period ranging from 0.97 to 1.41 years across diverse climate zones in China, the integrated heat pipe technology offers a highly efficient and economically viable alternative for the thermal management of communication cabinets [12].

Overall, heat pipe technology offers a compact, low-power, and highly effective solution for thermal management in outdoor communication cabinets. When integrated with smart controls and auxiliary systems, it can deliver substantial energy savings while maintaining the thermal stability of essential network infrastructure.

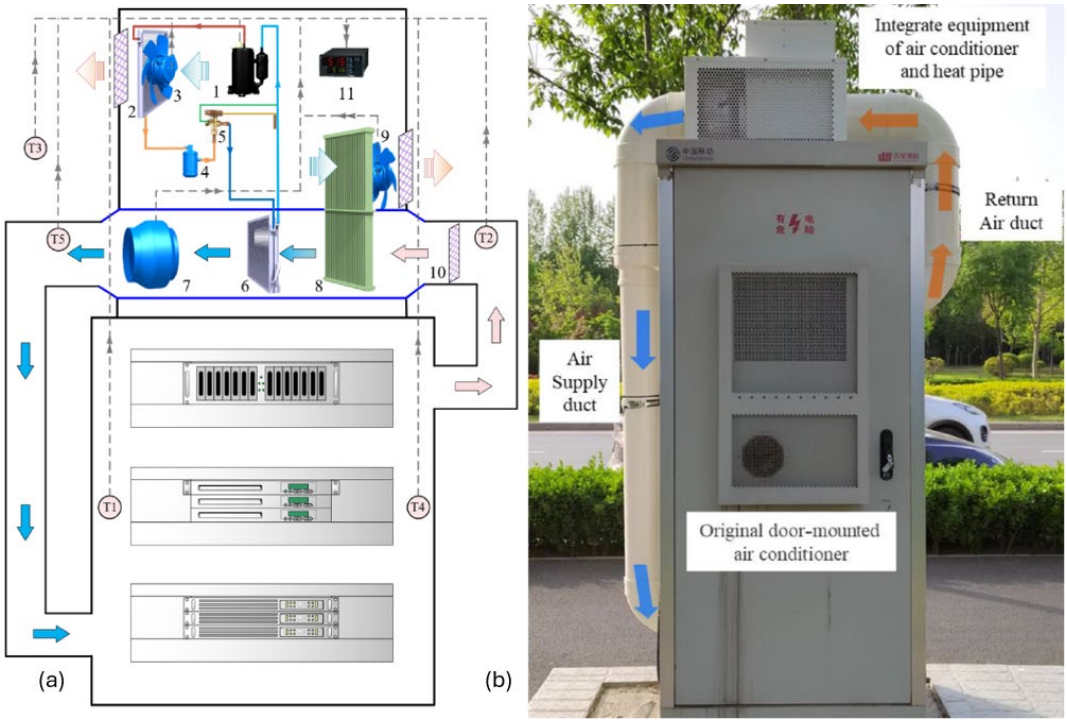


Figure 7 - (a) The Working Principle of Integrated Heat Pipes and Layout of Measuring Points. (b) Field Diagram of 5G Base Station [12].

3.3. Radiative Daytime Cooling

Radiative daytime cooling (RDC) is a passive thermal management strategy that enables sub-ambient cooling by emitting heat in the form of infrared radiation through the atmospheric transparency window (8–13 μm). Recent advancements in photonic materials and nanotechnology have overcome prior limitations that confined radiative cooling to nocturnal conditions. Radiative cooling exploits the Earth's energy balance by enabling objects to dissipate thermal energy to the seemingly endless heat sink that is outer space. While historically limited to nighttime applications, the emergence of materials with tailored optical properties has enabled effective daytime radiative cooling [13].

Radiative daytime cooling operates by leveraging two concurrent phenomena: high solar reflectivity to minimize solar heat gain and strong thermal emissivity in the mid-infrared range to emit thermal radiation into space. Under appropriate atmospheric conditions and with materials designed for spectral selectivity, surfaces can achieve steady-state temperatures below ambient air temperature even under direct sunlight. The effectiveness of radiative daytime cooling lies in the ability to reflect most of the solar spectrum while simultaneously emitting thermal energy via the atmospheric window with minimal interference [13].

The phenomenon of radiative cooling is observable in nature, particularly in arid regions where clear skies and low humidity permit significant heat loss at night. The Saharan Silver Ant can traverse the desert during the day, when it is typically too hot for other life forms, due to the small silver hairs on its body. These hairs also feature the properties of being highly reflective and emissive [13][14].

Historically, radiative cooling was harnessed in ancient Iran, where water left in shallow, insulated clay containers would freeze overnight through passive cooling—a method that prefigured modern refrigeration [15].

Within industry radiative cooling panels, composed of photonic multilayers and/or metamaterials, can achieve thermal emissivity in the relevant infrared band while maintaining high solar reflectance. These panels, which require no electricity, are suitable for use on rooftops, in industrial settings, and as components of integrated building systems such as HVAC and refrigeration [14].

In parallel, scalable emissive coatings and paints have emerged as an economical and flexible solution. Radiative cooling technology has only very recently been developed and studied in the form of coatings.

Yang et al. conducted an extensive field study on a superamphiphobic self-cleaning passive sub-ambient daytime radiative cooling coating applied to telecommunication base stations across multiple regions in China. The coating consistently reduced the interior cabinet temperatures by up to 10.6 °C and maintained sub-ambient conditions under full sun. As a result, base stations achieved energy savings ranging from 31.7% to 64.9% in cooling electricity usage compared to uncoated or conventionally coated stations. These gains were consistent across the tested locations of Chengdu, Panzhihua, Kangding, Yinchuan, and Ziyang within China, with the testing in Ziyang being conducted over the course of one year. Combined with minor door ventilation retrofits, the system enabled air-conditioner-free operation in many cases, reinforcing its viability as a high-impact, scalable cooling strategy for telecom infrastructure [16].

Figure 8 - Photos of the 4G Base Stations (a) with and (b) Without a Superamphiphobic Self-Cleaning PSDRC Coating Located in Ziyang, Sichuan Province, China. [16]

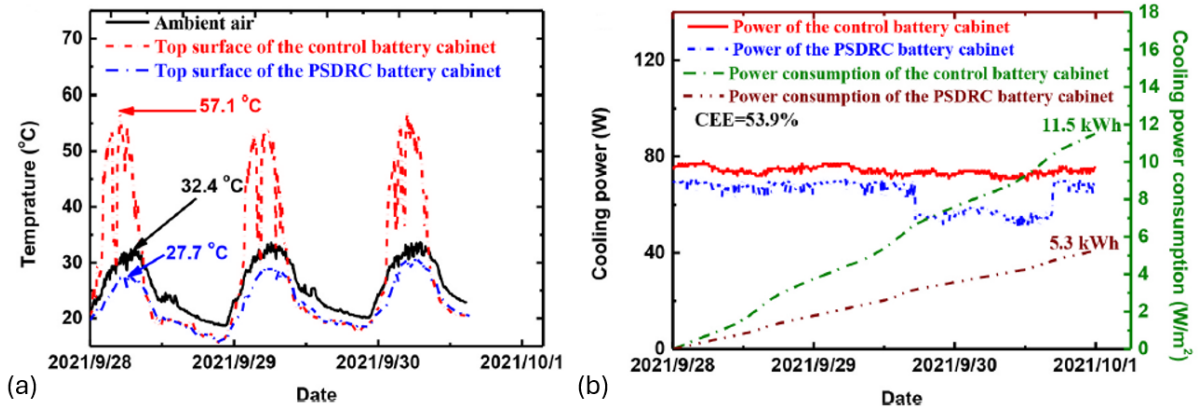


Figure 9 - (a) Top Surface Temperatures of the Coated and Uncoated Cabinets in Ziyang (b) Cooling Power and Power Consumption of Coated and Uncoated Battery Cabinets. [16]

Separately, Das et al. developed a low-cost, scalable MgO–PVDF nanocomposite paint optimized for passive daytime radiative cooling, also achieving exceptional performance. Field tests in Bangalore, India, showed surface temperature reductions of up to 13 °C under direct sunlight, with average cooling of 7 °C below ambient conditions. Compared to commercial white paints, the nanocomposite achieved 3 °C greater cooling and demonstrated water resistance, substrate compatibility, and strong adhesion. The study highlights the nanocomposite's utility for applications such as building surfaces, pavers, and rooftops, particularly in hot climates [17].

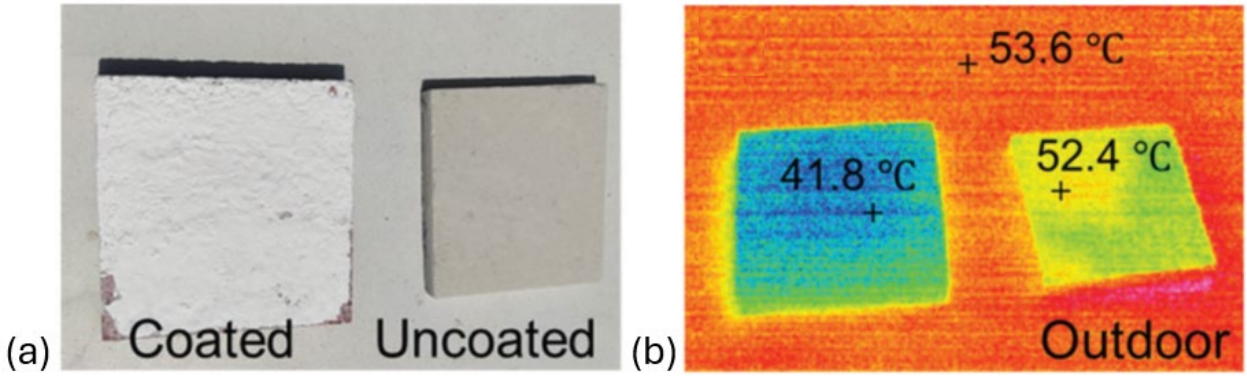


Figure 10 - (a) Photograph of a Coated and Uncoated Ceramic Paver (b) Outdoor Thermal Image of Coated and Uncoated Pavers [17]

These coatings do have their limitations. Because these coatings emit through the surface, out into space, their performance requires that the coated surfaces remain clean. As dirt and debris accumulates, the passive cooling performance of the coated surfaces diminishes. The two studies highlighted above were tested in conjunction with self-cleaning agents and as such, would be required to replicate similar results [16][17][19]. Furthermore, cloud coverage impacts the cooling impact of these coatings [13][14][16]. So, the typical climate these coatings are potentially implemented in must be considered and evaluated prior to use.

It is also worth mentioning that studies have looked at applying these coatings to buildings to lessen the urban heat island effect. Due to the canopy effect within urban areas, as well as the the surface area to volume ratio of large-scale buildings, the theoretical results have yet to warrant scaled physical testing [14]. Telecommunication cabinets seem uniquely positioned to utilize this lesser-known passive cooling technology. In summary, for these coatings to net a cooling effect they require a clear path to a clear sky.

Daytime radiative cooling represents a pivotal advancement in passive thermal management. While other technologies require more involved modifications, coatings can be easily implemented. In addition, these coatings have been presented as low-cost, supporting widespread deployment in energy-conscious applications (cost of self-cleaning agents not included) [16][17][18][19].

3.4. Mapping of Cooling Methods Over Different Climate Regions Within the U. S.

Numerous passive thermal management technologies have been discussed in the literature; some were covered in previous issues of the SCTE Technical Journal, others are highlighted in the current issue, and many remain unreviewed. To optimize the application of each technology for outdoor telecommunication cabinets, it is essential to select the most appropriate passive cooling method based on the specific characteristics of each location. Environmental conditions alone can provide a useful starting point for mapping suitable passive thermal management solutions to different climate regions across the United States. This mapping can serve as a preliminary guide to help cable telecommunication providers estimate which passive cooling approaches are most appropriate for outdoor cabinets in various U.S. climates, based on available research. However, it is important to emphasize that this serves only as an initial approximation. Several other critical factors, such as internal heat load generated by power electronic equipment, seasonal temperature variations, and other site-specific considerations, must also be considered when selecting an appropriate passive cooling strategy.

The United States Climate Zone Map, shown in Figure 10 and developed by ASHRAE in *ASHRAE Standard 169-2013*, is used to correlate various passive thermal management technologies with different climate regions across the U.S. This map categorizes regions based on heating and cooling requirements, reflecting the significant variations in environmental conditions across zones. Each thermal zone is assigned a number from 1 to 8, representing specific climate types as follows:

- Zone 1 - Very Hot
- Zone 2 - Hot
- Zone 3 - Warm
- Zone 4 - Mixed
- Zone 5 - Cool
- Zone 6 - Cold
- Zone 7 – Very Cold
- Zone 8 - Subarctic/Arctic

Additionally, these zone numbers are subdivided into three moisture categories:

- A – Moist/Humid
- B – Dry
- C - Marine/Coastal

A typical zone designation combines the zone number with the moisture category (e.g., 5A, 6B), as indicated on the map.

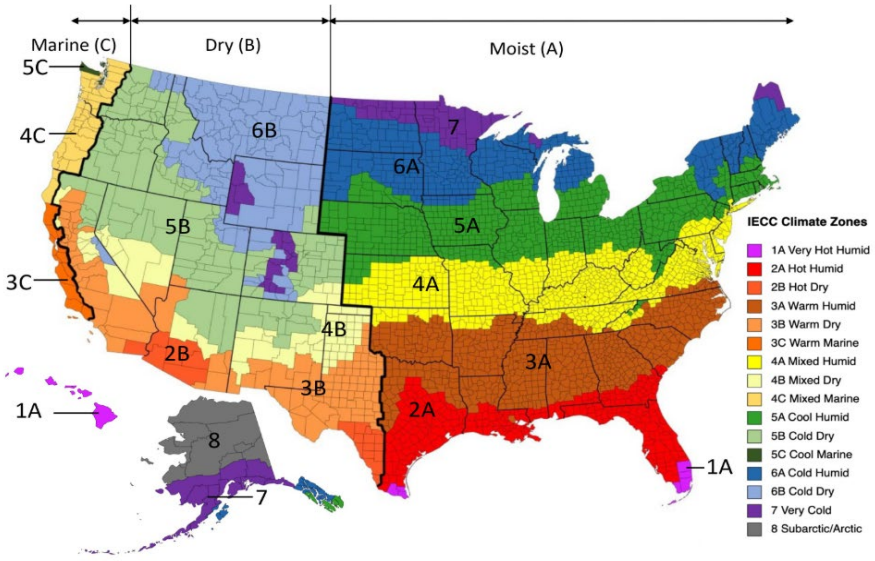


Figure 11 - United States Climate Zone Map [20], [21]

A detailed breakdown of climate zones by U.S. County and state can be found in Climatic Data for Building Design Standard, as outlined in ASHRAE Standard 169-2013 [20]

Based on a review of relevant research literature, Table 1 provides a preliminary comparison of five passive cooling technologies for outdoor telecommunication cabinets (not including designing for heat transfer as this is a methodology applicable everywhere), focusing on their suitability for various climate zones and their potential for energy savings. This table serves as an initial guide for cable

telecommunication providers seeking to enhance energy efficiency by selecting climate-appropriate passive cooling solutions. It is important to note that the passive cooling methods' estimated energy savings are conservative estimates reflective of our readings. The studies highlighted in this paper test the technologies across a wide range of temperature scenarios. As a result, the best-suited climate zones would likely result in savings beyond the ranges approximated below.

Table 1 - Energy-Efficient Passive Cooling Methods for Telecom Cabinets by Climate Zone

Cooling Method	Best-Suited Climate Zones	Energy Savings for Telecom Cabinet
Radiation Shielding	2A, 2B, 3A, 3B, 4A (Hot and sunny climates)	10–15% savings by reducing solar heat gain
Thermosyphons	5A, 5B, 6A, 6B, 7, 8 (Cool to very cold)	15–25% annual HVAC energy savings
Heat Pipes	4A, 4B, 5A, 5B, 6A (Mixed and cool climates)	10–18% cooling energy reduction
Thermochromic Coating	3A, 3B, 3C, 4A, 4B, 4C (Warm/Mixed zones)	5–10% cooling load reduction, depending on solar exposure
Radiative Cooling Coatings	2B, 3B, 4A, 4B, 5B (Dry with clear night skies)	10–20% reduction in cooling energy, especially at night

4. Conclusion

An array of passive cooling technologies can be utilized by telecommunication cabinets to minimize internal temperatures and ensure component operation and reliability, while minimizing cost and emissions. Coatings can be applied to the outside of the cabinets either during initial fabrication or after installation. The same can be said for radiation shielding, thermosyphons, and heat pipes. However, in these later instances the work involved in updating existing cabinets may be more involved.

More testing would give additional insight into how these technologies could be implemented and where. With greater insight, the deployment of these technologies can be done at scale with a greater degree of confidence that energy savings will ensure.

These passive cooling technologies mitigate risks, reduce costs, can drive growth, and enhance the overall brand of the companies that chose to address their energy consumption, resulting emissions, and their impact on our environment [22].

4.1. Recommendations

The various passive cooling technologies show potential to save significant energy in their implementation of telecommunication cabinets. While recommendations for their use have been given in a broad sense, site-specific studies like the ones highlighted in this literature review are recommended. Data related to telecommunication cabinet location, power consumption, internal components, layout, maintenance, and cost would help further understanding and could be used for testing either in computer programs or on-site. Testing can be done on implementing just one of the highlighted technologies or multiple to better understand the potential total sum of energy savings.

5. Abbreviations and Definitions

5.1. Abbreviations

CPSC	Consumer Product Safety Commission
OSHA	Occupational Safety & Health Administration
PUE	Power Usage Efficiency
RDC	Radiative daytime cooling
RISE	Resilient Innovation through Sustainable Engineering
SCTE	Society of Cable Telecommunication Engineers
SUSE	Sustainable Engineering

6. Bibliography and References

[1] Energy 20/20, “SCTE,” https://account.scte.org/standards/library/energy-2020-powering-cables-success/.

[2] nationalgrid.com, “What are scope 1, 2 and 3 carbon emissions?,” https://www.nationalgrid.com/stories/energy-explained/what-are-scope-1-2-3-carbon-emissions.

[3] Telcordia Technologies Generic Requirements, GR-487-CORE, Issue 3, April 2009.

[4] ESTEL, “Which is Better for Outdoor Telecom Cabinets Air Conditioner or Fan Cooling,” https://blog.outdoortelecomcabinet.com/air-conditioner-vs-fan-cooling-telecom-cabinets/.

[5] Thermal edge Inc, “5 LIMITATIONS OF AN ELECTRICAL ENCLOSURE COOLING FAN,” https://thermaledge.com/5-limitations-of-an-electrical-enclosure-cooling-fan/.

[6] Q. Okon, A. Murawski, K. Stansbury, H. Lange, A. Danial, and K. Schmidt, "Thermal Management of Outdoor Telecommunication Cabinets," *SCTE Technical Journal*, vol. 4, no. 2, pp. 5–25, Aug. 2024. [Online]. Available: https://www.scte.org/

[7] Villanova University, “RESILIENT INNOVATION THROUGH SUSTAINABLE ENGINEERING (RISE) FORUM,” https://www1.villanova.edu/university/engineering/academic-programs/departments/sustainable/rise.html.

[8] A. Jobby, M. Khatamifar, and W. Lin, “Alternative Internal Configurations for Enhancing Heat Transfer in Telecommunication Cabinets,” *Energies (Basel)*, vol. 16, no. 8, p. 3505, Apr. 2023, doi: 10.3390/en16083505.

[9] A. Hamouche and R. Bessaïh, "Mixed convection air cooling of protruding heat sources mounted in a horizontal channel," *International Communications in Heat and Mass Transfer*, vol. 36, no. 8, pp. 841–849, 2009. [Online]. Available: https://doi.org/10.1016/j.icheatmasstransfer.2009.04.009

[10] K. Bilen and S. Yapici, "Heat transfer from a surface fitted with rectangular blocks at different orientation angle," Heat and Mass Transfer, vol. 38, no. 7, pp. 649–655, 2002. [Online]. Available: https://doi.org/10.1007/s002310100275ResearchGate

[11] S. Cui, Y. Zhang, J. Bai, H. Zheng, H. Li, and C. Li, "Application of the integrated technology of heat pipe and air conditioner in communication cabinet cooling," *Case Studies in Thermal Engineering*, vol. 61, p. 104923, Sep. 2024, doi: 10.1016/j.csite.2024.104923.

[12] S. Cui *et al.*, "Energy-saving measures and temperature control for outdoor communication cabinets," *Thermal Science*, vol. 28, no. 3 Part A, pp. 2015–2022, 2024, doi: 10.2298/TSCI2403015C.

[13] D. Zhao *et al.*, "Radiative sky cooling: Fundamental principles, materials, and applications," *Appl Phys Rev*, vol. 6, no. 2, Jun. 2019, doi: 10.1063/1.5087281.

[14] S. Suhendri, M. Hu, Y. Su, J. Darkwa, and S. Riffat, "Implementation of passive radiative cooling technology in buildings: A review," *Buildings*, vol. 10, no. 12, p. 215, 2020. doi: 10.3390/buildings10120215. Accessed: May 22, 2025.

[15] M. N. Bahadori, "Passive cooling systems in Iranian architecture," *Scientific American*, vol. 238, no. 2, pp. 144–154, Feb. 1978. Accessed: May 22, 2025. [Online]. Available: *Scientific American Archives*

[16] Z. Yang *et al.*, "Sub-ambient daytime cooling effects and cooling energy efficiency of a passive sub-ambient daytime radiative cooling coating applied to telecommunication base stations— Part 1: Distributed base stations and long-lasting self-cleaning properties," *Energy*, vol. 283, p. 128750, Nov. 2023, doi: 10.1016/j.energy.2023.128750.

[17] P. Das, S. Rudra, K. C. Maurya, and B. Saha, "Ultra-Emissive MgO-PVDF Polymer Nanocomposite Paint for Passive Daytime Radiative Cooling," *Adv Mater Technol*, vol. 8, no. 24, Dec. 2023, doi: 10.1002/admt.202301174.

[18] SRI, "Self-Cooling Paint ™ – A Passive Radiative Cooling Solution," https://www.sri.com/fcd_technology/self-cooling-paint-a-passive-radiative-cooling-solution/.

[19] H. Zhang *et al*., "Superamphiphobic coatings with subambient daytime radiative cooling—Part 1: Optical and self-cleaning features," *Solar Energy Materials and Solar Cells*, vol. 245, p. 111859, 2022. [Online]. Available: https://doi.org/10.1016/j.solmat.2022.111859

[20] ANSI/ASHRAE Addendum a to ANSI/ASHRAE Standard 169-2013," 2020. [Online]. Available: www.ashrae.org

[21] U. States Department of Energy, "Preliminary Energy Savings Analysis: ANSI/ASHRAE/IES Standard 90.1-2019," 2021.

[22] K. Schmidt, Topic" Business Case for Sustainability", SUSE 8112, College of Engineering, Villanova University, Villanova, PA, Jan 2025.

Click here to navigate back to Table of Contents

SCTE STANDARDS JOURNAL
SCTE a subsidiary of CableLabs

Energy Projections for Grid Demand in High Density Areas and Backup Power Needs/Options

Pennsylvania, New York and Arizona

A Technical Paper prepared for SCTE by

Isaac Bua
https://www.linkedin.com/in/bua-isaac/

Zach Taylor
https://www.linkedin.com/in/zachary-taylor-58016946/

Kamil Aliyev
https://www.linkedin.com/in/kamil-aliyev-b98156222/

Queen Okon
https://www.linkedin.com/in/okonqueen/

Abolfazl Baloochiyan
https://www.linkedin.com/in/abolfazl-baloochiyan-a69a17208/

Bolape Alade
https://www.linkedin.com/in/bolape-alade-3a214a10a/

Villanova University
800 East Lancaster Ave.
Villanova, PA 19085
610-519-4500

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Energy Sources: Primary and Secondary

The U.S produces and consumes various types of energy, categorized as primary and secondary energy. Primary energy sources include fossil fuel options such as petroleum, natural gas, and coal, nuclear energy, biomass, renewables (solar, wind, hydro, geothermal) and can be consumed directly or converted to secondary energy form. Electricity is the flow of electrical power or charge and is often referred to as secondary energy source, because it is generated from primary sources [1]. Just over a century ago, there were no electric motors, lightbulbs, refrigerators, or air conditioners. Even now, over 750 million people globally lack access to electricity according to the International Energy Agency 2023, with majority in Sub-Saharan Africa. However, the electric power business has grown throughout time to become one of the world's largest machines, it is also one of the most polluted. The sector was responsible for 34% of 2019 global greenhouse gas emissions majorly from the burning of coal, natural gas, and oil for electricity, and a major share of particulate matter and hazardous heavy metals [2]. Excessive amounts of greenhouse gasses in the atmosphere can lead to global warming. As a result, the sector is currently seeing a big shift away from fossil fuel-based energy sources with significant reduction in the amount of coal fired plants to renewables options. This is in line with the IPCC recommendation to limit global warming at 1.5 ℃ hence the need to decarbonize the grid and become NetZero. The electricity we use goes through different stages: generation, transmission and distribution to different customers with different regulations at each stage to ensure reliability.

In the U.S, the North American Electric Reliability Corporation (NERC) is responsible for ensuring the reliability of North America's bulk power system. Authorized by the Federal Energy Regulatory Commission (FERC), NERC enforces reliability standards and ensures compliance across the industry. North America's power grid includes approximately 522,000 miles of high-voltage transmission lines (100,000 volts and above) and represents over $1 trillion in infrastructure assets. The North American Electric Reliability Corporation (NERC) oversees the reliability of the bulk power system and operates through six regional reliability entities. For more granular reliability assessments, NERC divides these regions into 20 assessment areas to better evaluate resource adequacy and address deliverability constraints.

1.2. Regional Breakdown

- Pennsylvania is part of the ReliabilityFirst (RF) region and is primarily managed by PJM Interconnection, a regional transmission organization (RTO) responsible for coordinating the movement of wholesale electricity across 13 states and the District of Columbia [3].
- New York falls under the Northeast Power Coordinating Council (NPCC) and is managed by the New York Independent System Operator (NYISO), which oversees grid operations and electricity markets within the state [4].
- Arizona is within the Western Electricity Coordinating Council (WECC) and is part of the Southwest Reserve Sharing Group (SRSG), which helps ensure power system reliability across the western U.S. [5]

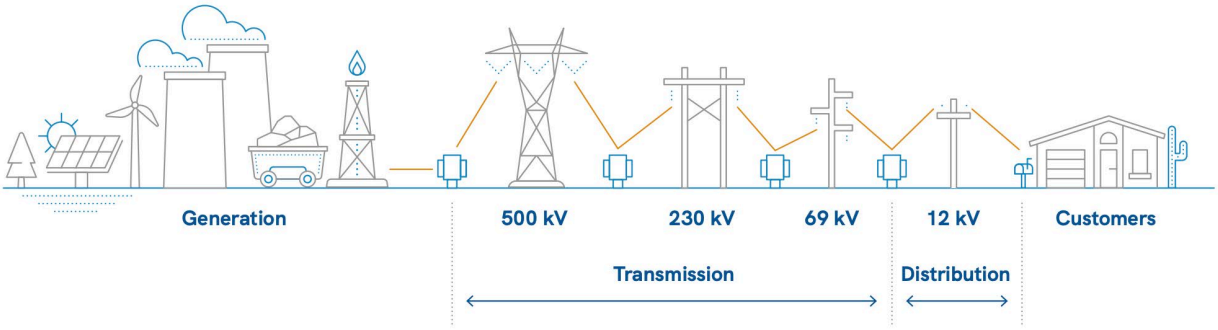


Figure 1 - Schematic of Electricity Production to Final Consumers.

2. Objectives

3. Goal and Scope

The primary goal of this project is to evaluate future grid demand projections in New York, Pennsylvania, Arizona over the next ten (+10) years and to identify optimal non-fossil fuel backup power options for critical sites with limited grid supply. This study will focus on,

- Reviewing energy projections for grid demand in the specified regions.
- Identifying suitable non-fossil fuel backup power options (e.g., nuclear, fuel cells, microgrids, battery storage etc.) for areas with constrained energy supply.
- Assessing these backup options using a comprehensive screening framework based on social, technological, environmental, economic, and political (STEEP) criteria.

3.1. Methodology

The group begun by conducting an in-depth literature search comparing published information from utility companies, the U.S. Department of Energy (DOE), Energy Information Administration, published research, news articles and other national research laboratories publications and whitepapers on the topic. Additionally, analyzing publicly available data from authoritative sources such as the National Renewable Energy Laboratory (NREL) and the Energy Information Administration (EIA) to support projections and validate findings on the different energy sources for each state. This was followed by developing and applying STEPP (social, technological, environmental, economic, and political) as a screening criterion to assess the benefits and challenges of various non-fossil fuel backup power options and finally integrating the findings from the literature review and data analysis to generate actionable conclusions and recommendations for enhancing energy supply reliability and low carbon backup power solutions for critical telecom sites.

3.2. U.S Energy Mix Outlook.

Different energy sources are measured in specific units' liquid fuels in barrels or gallons, natural gas in cubic feet, coal in short tons, and electricity in kilowatts and kilowatt-hours. To compare different energy types, the U.S. commonly uses the British thermal unit (Btu), which measures heat energy. In 2023, the country's total primary energy consumption was approximately 94 quadrillion Btu. Despite significant advancements in energy efficiency, the United States continues to lose approximately 61% of its total

energy, with more than 18% of that loss occurring in the electric sector. This emphasizes the need for ongoing improvements, such as upgrading power plants to more efficient technologies, modernizing the system to reduce transmission losses, and expanding demand-response programs to better match electricity supply and demand. By investing in these areas, the electric sector may help reduce overall energy waste while also cutting costs and environmental implications. [6]

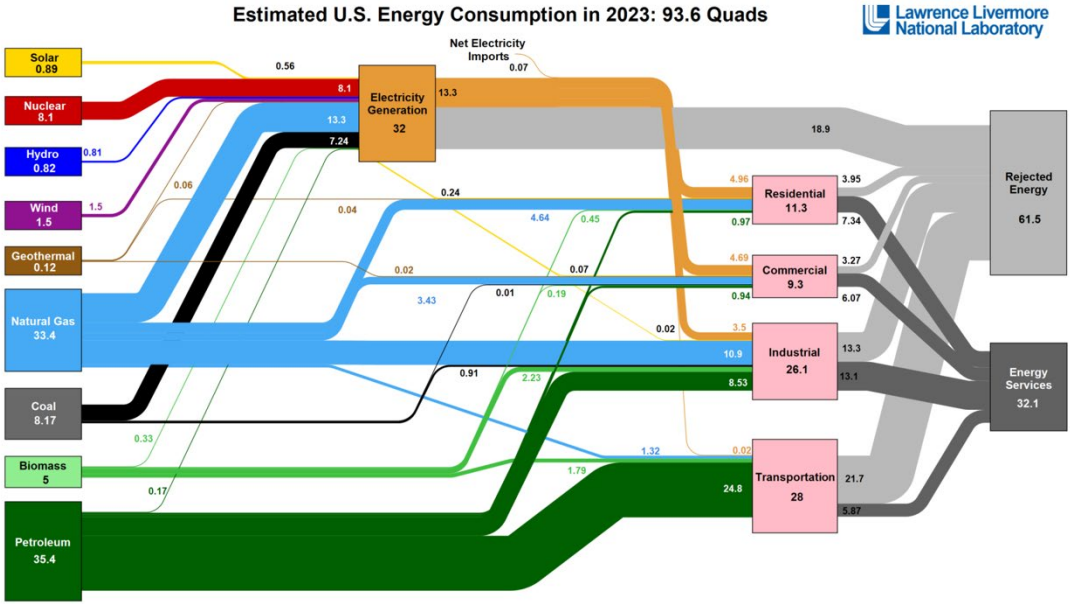


Figure 2 - Estimate U.S Energy Consumption in 2023.

3.3. Overall U.S Electricity Demand Projections.

After nearly two decades of stability, electricity usage increased by 2% in 2024 and is anticipated to rise by another 2% in 2025 and 2026. This expansion is mostly driven by increased demand from a rebound in US manufacturing, semiconductor and battery production facilities, as well as data centers supporting Artificial Intelligence (AI) [7]. Natural gas and solar power are expected to be the major sources to meet the projected demand, with the electric power industry planned installation of 26 gigawatts (GW) of solar capacity in 2025 and another 22 GW in 2026. These expansions are estimated to increase US solar power by 34% in 2025 and 17% in 2026. Energy storage is required to provide critical reliability services due to the intermittent nature of solar energy; interconnection queues are growing across the country for grid scale battery storage [8].

3.4. The U.S. Energy-Related Carbon Emissions, 2019–2023

U.S. energy-related carbon dioxide emissions declined slightly in 2023 compared to 2022, with a total reduction of about 3% (134 million metric tons) [9]. While emissions fell across several sectors, more than 80% of the overall reduction came from the electric power sector, which saw a 7% drop in emissions compared to the previous year [10]. This significant decline in the electric power sector was primarily due to a 19% decrease in coal-fired electricity generation, as natural gas and solar power took on a larger share of the generation mix. Natural gas generation increased by 7%, and solar generation rose by 14% in 2023. Since coal is the most carbon-intensive fossil fuel, this shift in the energy mix led to a substantial reduction in emissions from electricity production.

Table 1 - Showing Carbon Emission in Metric Tons 2019-2023

Sector	2023	2022	2021	2020	2019
Residential	311	339	325	319	347
Commercial	250	261	245	233	255
Industrial	963	960	977	952	1,007
Transportation	1,856	1,840	1,807	1,630	1,921
Electric power	1,427	1,542	1,551	1,450	1,618
Total	**4,807**	**4,941**	**4,905**	**4,584**	**5,147**

Data source: U.S. Energy Information Administration, *Monthly Energy Review*, March 2024, Tables 11.1–11.6
Note: Totals may not equal sum of components due to independent rounding.

Over the past decade, the United States has retired a significant amount of coal-fired power capacity. A record 14.9 gigawatts (GW) of coal capacity was shut down in 2015. From 2015 to 2020, annual retirements averaged 11.0 GW, before dropping to 5.6 GW in 2021. Retirements then rose again to 11.5 GW in 2022. In 2023, power plant owners and operators planned to retire 8.9 GW of coal-fired capacity, which represented about 4.5% of the total U.S. coal capacity at the beginning of that year. Most of the remaining coal plants in operation were built in the 1970s and 1980s, and many are being phased out due to competition from more efficient natural gas plants and low-cost renewable energy sources like wind and solar. [11]

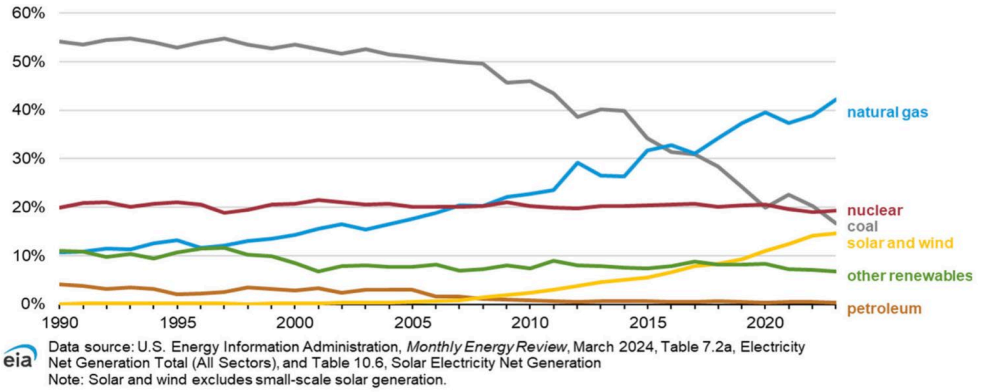


Figure 3 - Share of U.S. Electric Power Sector Generation by Fuel Source, 1990–2023

4. State Level Electricity Projection: Pennsylvania, New York and Arizona

4.1. Pennsylvania Electricity Projections

To start with, the Pennsylvania Public Utility Commission's 2023 Electric Power Outlook concludes that the Commonwealth has sufficient generation, transmission, and distribution capacity to meet customer demand for the foreseeable future, with regional reliability expected to be sustained through at least 2033 assuming timely completion of planned infrastructure projects. According to the North American Electric Reliability Corporation (NERC), PJM has consistently exceeded its reserve margin requirements. In 2024, PJM's required reserve margin was 14.7%, while the anticipated available reserve was 36.8%, signaling a

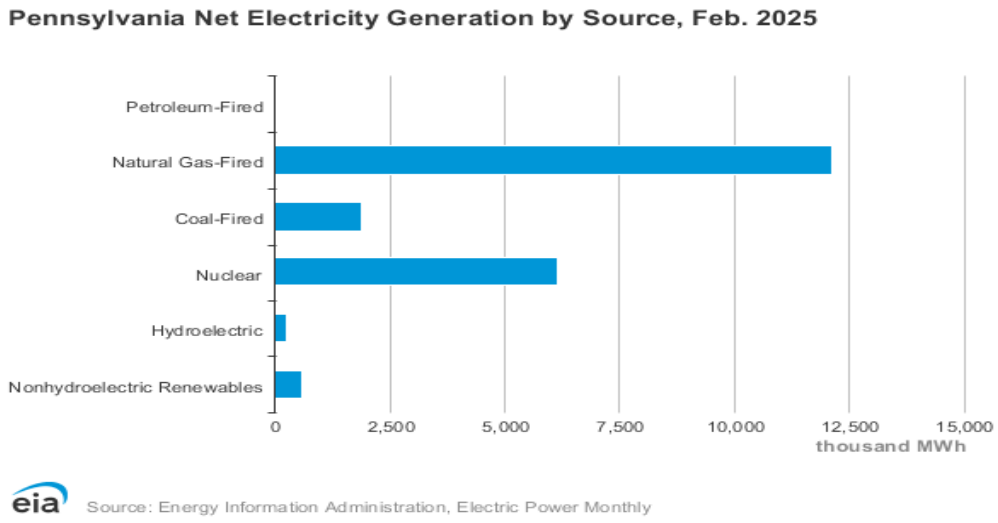
healthy buffer. However, risks loom beyond 2033, primarily due to a mismatch between power plant retirements, increased load growth, and the slower pace of new generation development. The ongoing retirement of coal and gas-fired generation, combined with a shift toward renewable and variable energy resources (VERs), may introduce future resource adequacy and grid reliability concerns—especially during extreme weather events like severe cold or heatwaves. Electricity consumption in Pennsylvania declined in 2023, with total net usage at 138,744 GWh, marking a 3.44% year-over-year (YOY) decrease from 2022. Notably:

- Residential consumption fell by 6.74%
- Commercial usage declined by 2.55%
- Industrial demand decreased marginally by 0.40%

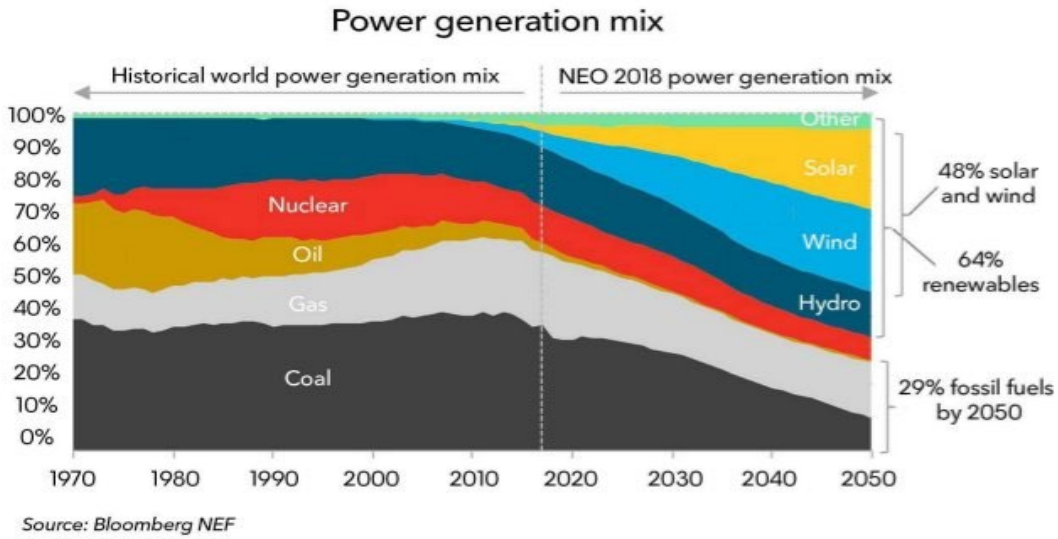
Despite this recent dip, long-term projections show modest but steady growth, with total energy demand expected to increase by 0.76% annually over the next five years.

- Residential demand is projected to grow at 1.39% annually
- Commercial growth is projected at 0.6% annually
- Industrial usage is expected to remain relatively flat

As Pennsylvania continues to transition towards a cleaner energy mix, careful planning, coordinated policy support, and resilient grid investments is still essential. Pennsylvania remains the second largest natural gas producing state. [12]



4.1.1. Power Generation Mix in PA Since 1970 - 2050.



The above figure illustrates Pennsylvania's electricity generation capacity by energy source, based on 2016 data from the U.S. Energy Information Administration. It showcases the state's diverse energy mix, with significant contributions from nuclear, natural gas, and coal. Renewable sources, including hydroelectric, wind, solar, biomass, and others, constitute a smaller portion of the total capacity. The projections indicate fading coal sources and emerging new sources like solar and wind as important sources, while some sources like natural gas remain almost the same. [13]

4.2. New York Electricity Projections

New York ranks among the top ten U.S. states in both total net electricity generation and carbon dioxide emissions [14]. The state is currently pursuing aggressive decarbonization policies and yet experiences massive economic growth. Electricity demand in New York is projected to rise significantly somewhere between 50% to 90% over the next two decades [15]. This surge is largely being driven by electrification of heating systems, transportation (especially electric vehicles), and the energy needs to power data centers and other emerging technologies.

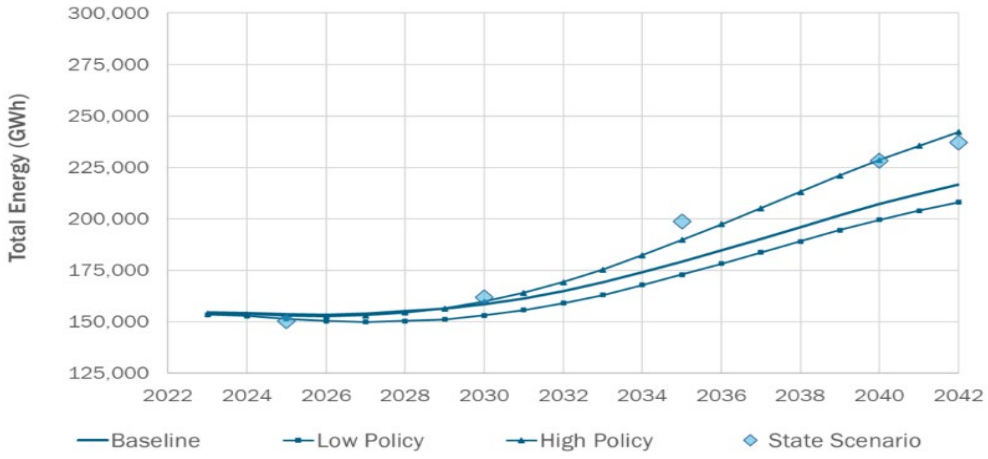


Figure 4 - Electricity Demand Projection for New York

To meet this demand sustainably, New York is planning a major shift away from fossil fuels. One ambitious scenario envisions the state becoming fully independent of fossil fuels and nuclear power by 2050, relying instead on wind, solar, hydropower, and hydrogen fuel cells [16]. This transition would require the deployment of thousands of renewable energy generation plants, including wind turbines, solar farms, and geothermal systems [17]. However, the path forward is not without challenges. The plan faces economic, political, and technical hurdles, including the need for grid modernization.

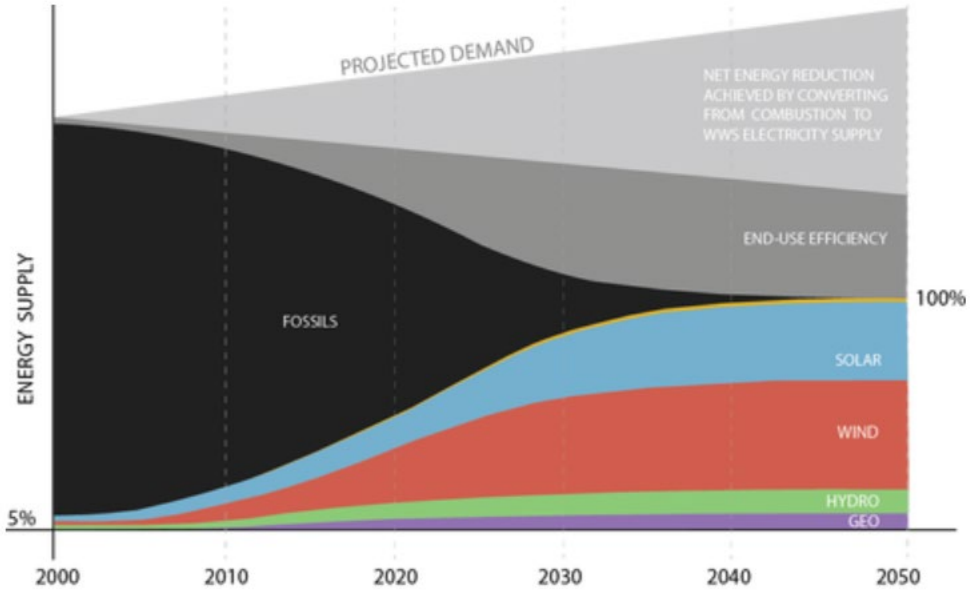


Figure 5 - Showing a Modeled Scenario of New York Complete Independence from Fossil Fuels

4.3. Arizona Energy Outlook

Over the next three decades, Arizona's electricity demand is projected to increase by 60%, driven primarily by population growth and economic expansion, even without the full-scale adoption of

electrification such as the use of electric vehicles or electric heating [18]. To meet this projected demand, the state is expected to add 57 gigawatts (GW) of new generating capacity, a 206% increase over current levels. Approximately 80% of this new capacity is planned to come from intermittent renewable sources, predominantly wind and solar. While these technologies have seen significant cost declines such as a 90% drop in utility-scale solar installation costs over the past decade reliance on them introduces challenges related to reliability and the need for substantial backup or dispatchable capacity, especially during periods of low renewable output [19].

Arizona's strategy has so far differed from other states like Texas, by maintaining a higher proportion of generation, from natural gas and nuclear. This has helped the state avoid the kinds of rolling blackouts that have plagued other parts of the U.S. during recent extreme weather events. For example, during the record-setting heatwave in the summer of 2023, Arizona's more balanced generation mix ensured grid stability, while other states struggled with insufficient supply.

By 2050, under the baseline scenario, Arizona's electricity grid is projected to be 67% renewable and 87% carbon-free, a dramatic shift from today's figures of 8% and 43%, respectively. However, the increased use of intermittent sources will require Arizona to "overbuild" its grid, meaning the state must develop more capacity than peak demand requires, to compensate for times when renewable output is low. Without this, the risk of outages and grid instability rises significantly. [20]

5. Backup Power Needs and Options Available

5.1. Solar

The sun emits solar radiation in the form of light. Solar energy technologies capture this radiation and turn it into useful forms of energy. There are two main types of solar energy technologies, photovoltaics (PV) and concentrating solar-thermal power (CSP). To effectively discuss the potential differences in solar energy sources in the three states, we will analyze various aspects such as solar irradiation, levelized cost of electricity, capacity factor, total capacity, and recent developments in solar energy technologies in these states.

5.2. Average Solar Radiation for the 3 states.

The study and measurement of solar irradiance have several important applications, including the prediction of energy generation from solar power plants, and the heating and cooling loads of objects. Solar irradiance is measured in watts per square meter (W/m2) in SI units.

Global horizontal irradiance (GHI) is the type of the solar irradiance which the total irradiance from the Sun on a horizontal surface on Earth. It is the sum of direct irradiance (after accounting for the solar zenith angle of the Sun z) and Diffuse horizontal irradiance [21]. By referencing data from NREL, we found the total irradiation for the three states as follows:

Table 2 - Monthly Global Horizontal Irradiance for 3 Cities

Months	January	February	March	April	May	June	July	August	September	October	Novomber	December	Average
New York State GHI(2023)kW/m2	44	76.5	117.5	141.7	211.2	173.6	191	159	119	98.5	70.6	46	120.7kW/m2
Pennsylvania GHI(2023)kW/m2	44.5	82.8	116.5	150.6	198.4	177.3	186.5	162.9	129.3	106.3	75	43.8	122.8kW/m2
Arizona GHI(2023)kW/m2	92.1	123.7	161	223.1	255.6	261.7	245.4	206.3	186.5	164.6	116.3	96.7	177.7kW/m2

kW/m² is the measure of solar energy per square meter. If a solar panel efficiency is 20%, then it can possibly convert 200W from that 1 kW/m² into actual electricity per square meter. From the values, we can see how much variation there is in irradiation in the three states. The irradiation levels in Pennsylvania and New York are almost the same, whereas Arizona is significantly different from New York and Pennsylvania with much higher irradiation level. This indicates that in Arizona, SCTE would generate more electricity taking up lesser space.

5.3. Pennsylvania

Pennsylvania is set for a significant solar energy expansion, aiming to generate 10% of its electricity from in-state solar sources by 2030 up from less than 1% today. This ambitious target, outlined in the Pennsylvania Solar Future Plan, would require the development of over 10 – 12 gigawatts (GW) of solar capacity. Two modeled scenarios suggest this would necessitate using between 56,800 and 79,200 acres of land, primarily farmland, representing just 0.8% to 1.1% of the state's total operated farmland. While the land requirement is relatively modest in proportion, the expansion faces notable challenges. Farmland is attractive for solar development due to its flat, sunny terrain and proximity to infrastructure. However, rural communities express concern over the long-term loss of productive farmland, the social and ecological impacts, and the fairness of lease agreements. Agrivoltaics that is integrating agriculture with solar panels, remains underutilized due to vague lease incentives and implementation barriers [22]. At least over 74% of new request in the generation queue is solar projects.

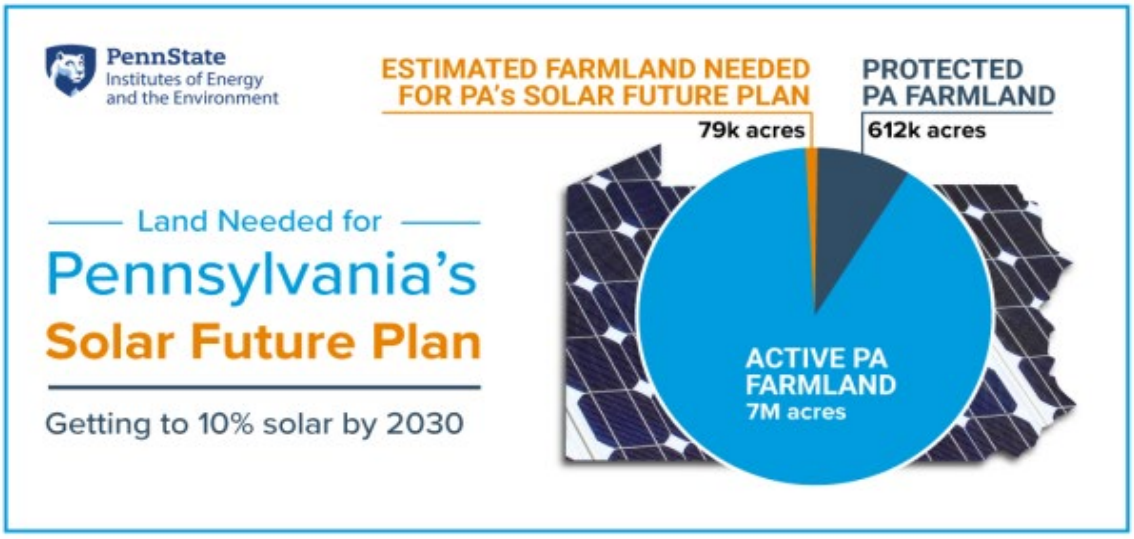


Figure 6 - Showing Land Requirement to Meet Solar Expansion

5.4. Economic and Technological Factors Important for Solar Panel Installation.

Levelized cost of electricity (LCOE) is a measure of the average net present cost of electricity generation for a generator over its lifetime. It is used for investment planning and to compare different methods of electricity generation on a consistent basis [23]. LCOE is an estimate of electricity cost of production. There is also The Site Levelized Cost of Electricity (LCOE), the cost of electricity produced by solar panels in this site, considering solar irradiance, land cost, and efficiency. The Site LCOE is lower than the Total LCOE, as the total includes transmission and others beyond local generation. The Levelized Cost of

Transmission (LCOT) is a measure of the cost of transporting electricity over a transmission network. It represents the cost per unit of electricity ($/MWh) associated with building, operating, and maintaining transmission infrastructure over its lifetime.

Table 3 - Levelized Cost of Energy ($/MWh) for PV Systems

States	LCOE Residential Rooftop PV($/MWh)	LCOE Commercial Rooftop PV ($/MWh)
New York	122	78
Pennsylvania	120	77
Arizona	83	53

As seen from the chart, the LCOE is marginally more in Pennsylvania and New York because it's directly proportional to lesser available area and lesser solar irradiance. Both these factors lead to less energy output, thereby resulting in marginally more LCOE there.

Key Factors Affecting LCOT

- Distance & Infrastructure Costs – Longer transmission lines increase costs.
- Technology Used – AC vs. HVDC transmission impacts efficiency and losses.
- Energy Losses – Transmission efficiency affects the actual energy delivered.
- Lifespan & Utilization Rate – Higher usage lowers LCOT per MWh.
- Geographical Challenges – Terrain, permitting, and regulatory costs impact cost variations [24].

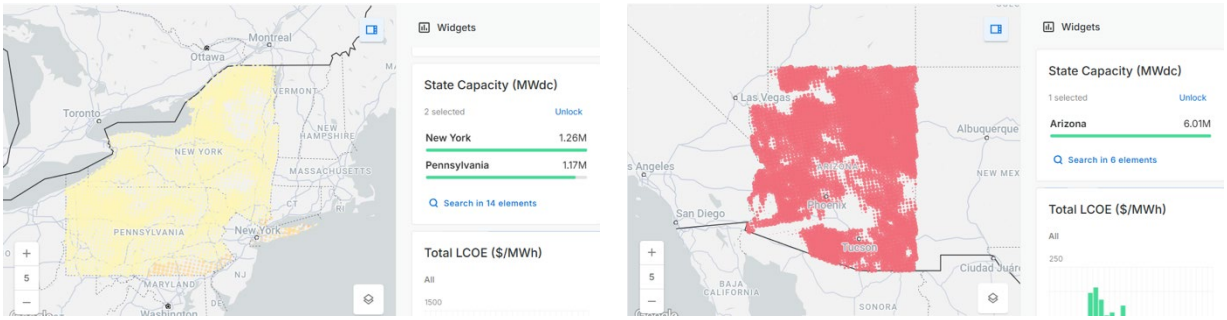
Capacity factor is a measure of how efficiently a power plant is used. It's the ratio of the actual energy produced by a plant over a specific period to the maximum possible energy it could have produced if it ran at full capacity during that same period. Since the amount of sunlight varies both with the time of the day and the seasons of the year, the capacity factor is typically computed on an annual basis.

Table 4 - Average Capacity Factor for PV Systems Based on AC and DC Ratings

States	Avergae Capacity Factor AC(%)	Avergae Capacity Factor DC(%)
New York	0.223	0.166
Pennsylvania	0.224	0.167
Arizona	0.338	0.252

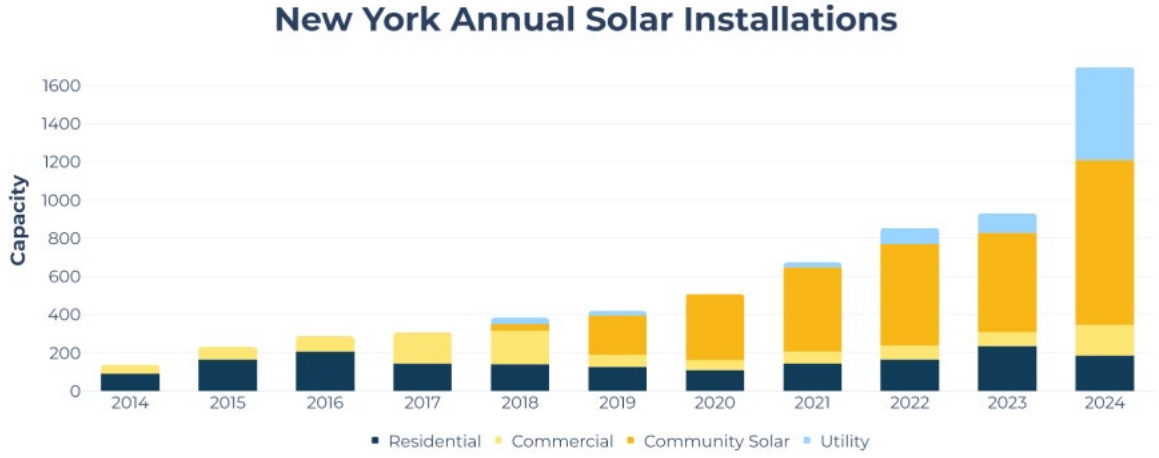
Arizona has a high-capacity factor since it receives higher solar irradiance levels, indicating more potential to always produce enough energy to satisfy anticipated demand.

5.5. Solar Energy Capacity in Three U.S. States

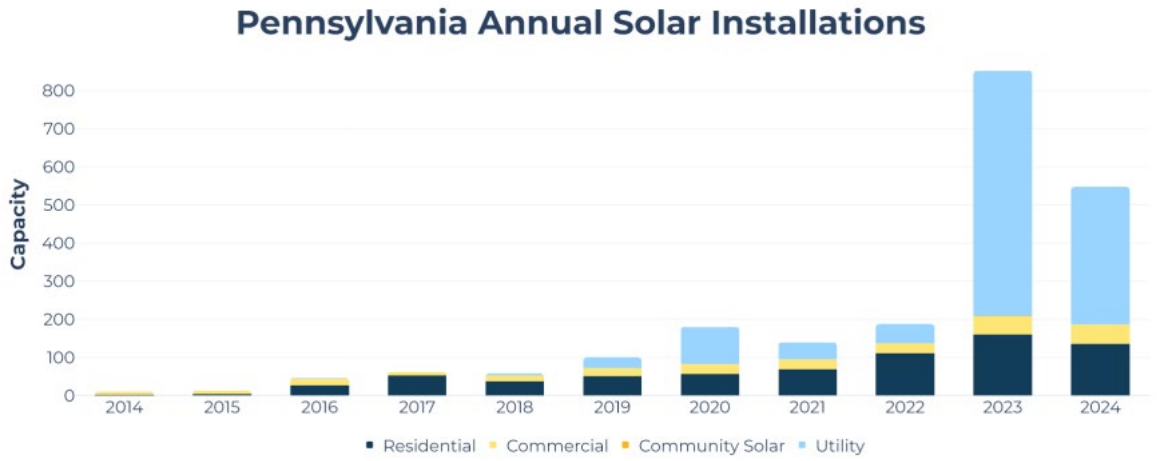


According to the Data, Arizona is the leading among the three states in solar energy capacity with 6.01 MWdc, far ahead of New York and Pennsylvania, which are at capacities of 1.26 MWdc and 1.17 MWdc, respectively. This can be assumed that the geographical and climatic situation of Arizona offers greater possibilities to produce solar power than the northeastern states.

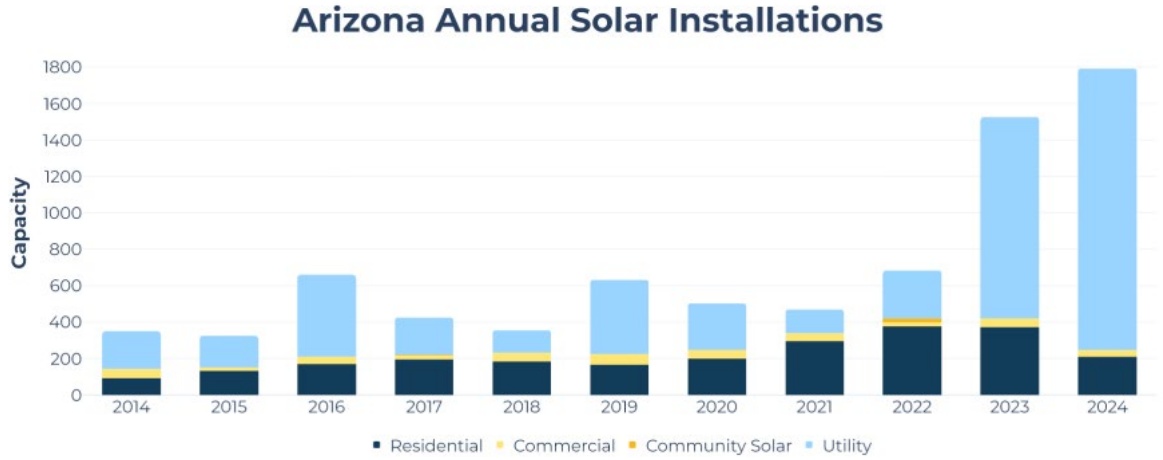
Solar PV accounted for 5.4% of total global electricity generation, and it remains the third largest renewable electricity technology behind hydropower and wind [25]. There are approximately 699 solar companies in New York City, including 82 solar manufacturers, 219 solar developers, and 398 other solar companies. New York showed the most diversified growth, with a huge surge in 2024 installations in residential, commercial, community, and utility market segments, supported by nearly 700 solar companies [26].



There are 440 solar companies in Pennsylvania, including 100 solar manufacturers, 212 solar developers, and 128 other solar companies [27]. Pennsylvania saw an unexpected rise in 2023 utility solar installations, which could be attributed to the recent policy or investment boosts, although it has fewer firms (440) than New York.



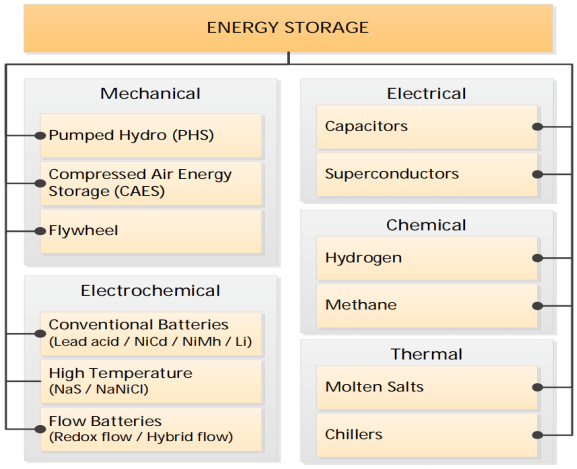
There are approximately 342 Solar Companies in Arizona State which 67 Solar Manufacturers, 156 Solar Developers and 119 Other Solar Companies [26]. Arizona, with 342 solar companies, had the most consistent long-term growth, especially utility-scale, with steady installation levels before its massive ramp-up in 2023 and 2024.



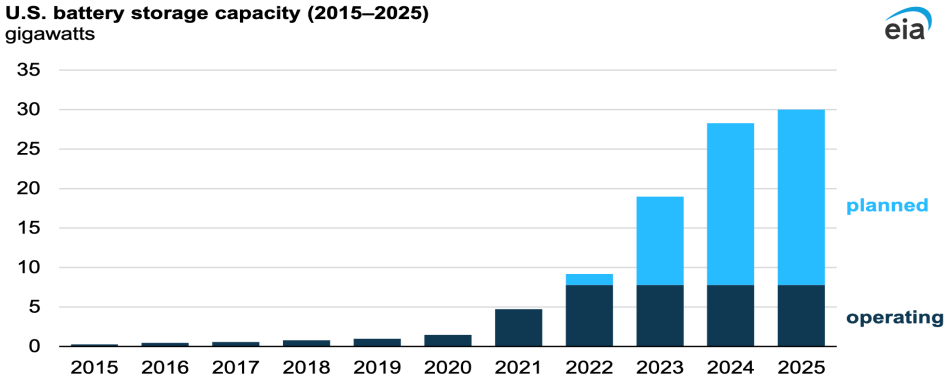
6. Energy Storage - Battery Technology

The term “energy storage” applies to technologies for absorbing energy in some form at one time and releasing later. Battery storage can offer transmission-related services like curtailment reduction, congestion relief, and contingency management. There are two distinct regulatory frameworks that govern storage as transmission: Storage as a Transmission Only Asset (SATOA) and Storage as a Transmission Asset (SATA). A storage asset that is solely capable of offering transmission-related services and lacks market-focused activities is referred to as SATOA. SATA, on the other hand, refers to a storage asset that is utilized for transmission-related services, but it can also be employed for market activities including capacity auction bidding, ancillary service provision, and wholesale energy buying and selling. Storage technologies can be grouped by the similarities of the storage medium. The figure below shows such a classification scheme [28].

6.1. Classification of Battery Energy Storage Technologies



The U.S. is rapidly expanding its utility-scale battery storage capacity, primarily to support the integration of intermittent renewable energy sources like solar and wind. According to the U.S. Energy Information Administration (EIA), battery storage capacity is projected to reach 30.0 gigawatts (GW) by the end of 2025, a dramatic increase from just 1.5 GW in 2020. This growth is driven by the need for grid stability and flexibility as renewable energy becomes a larger share of the generation mix. [29]



Batteries in general work by using chemical reactions in multiple electrochemical cells to move electrons. Essentially, they store extra energy when it's abundant and release it when needed. This flexibility makes them suitable for both short-term bursts and long-term energy needs. They are not only efficient and scalable but also versatile enough to be integrated at various points within the energy grid, from localized distributed setups to larger centralized systems that serve both mobile and stationary applications. However, despite their many advantages, challenges such as limited energy density, variability in power performance, finite lifespans, charging limitations, as well as environmental, safety, and cost concerns, still restrict their broader implementation.

6.1.1. Different Battery Types

- Lead-acid batteries are the oldest and most widely used rechargeable batteries, known for their moderate efficiency and they have a lifespan of 5–15 years, commonly used in vehicles, off-grid systems, and uninterruptible power supplies.
- Lithium-ion batteries, have a higher energy density, lower self-discharge, and better charging efficiency. Lithium-ion batteries have increasingly become a preferred replacement for lead-acid batteries, finding applications in portable electronics, electric vehicles, and power-grid stabilization.
- Sodium Sulphur batteries operate at high temperatures and deliver high power and energy density, although they require external heat sources and pose certain safety risks, making them more suitable for large-scale grid support.
- Flow batteries, on the other hand, store energy in two separate electrolyte solutions and offer moderate efficiency along with a long cycle life and operational flexibility, which, despite their increased complexity, makes them ideal for scalable, large-scale applications.

There are two important performance considerations, and these are the amount of energy the technology can hold – its round-trip efficiency and how much of that energy can be released in a short time that is the discharge time and duration. Today, the primary form of storage is pumped hydro, comprising 92% of total capacity across the country and the remaining 8% is comprised of thermal storage, compressed air energy storage, flywheels and batteries.

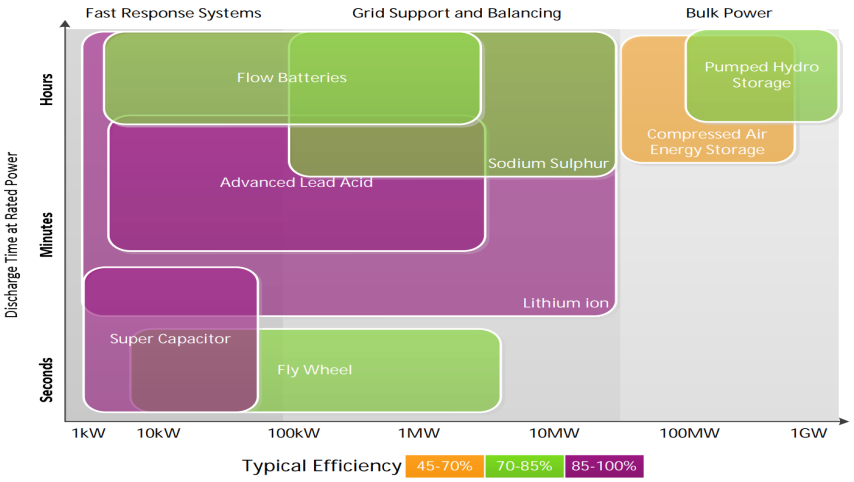


Figure 7 - Electricity Storage Technologies Comparison – Discharge Time vs. Power Capacity (MW)

6.2. Battery Storage Capacity in Pennsylvania

The Pennsylvania Public Utilities Commission (PUC) finalized a policy statement in April 2024 that allows utilities to use energy storage, including batteries, to enhance grid reliability and resilience, with a case-by-case approach to asset ownership. The policy statement notes that electricity-storage technology, namely batteries, can be used by EDCs (electricity distribution companies) instead of or in addition to more traditional "wired" solutions, to maintain or to increase the reliability or the resilience of the electric

distribution system, along with potential benefits. As of 2024, there is over 5,300 MW of Energy Storage in PJM and 96% being pumped hydro and just 4% include other energy storage technologies. At least 300MW being batteries, of which 50% batteries connected to the distribution system and 50% connected to the bulk electric system. [30]

6.3. Battery Storage Capacity in New York

The New York State Public Service Commission (PSC) adopted a bold storage goal in June 2024: 6,000 MW, 6 gigawatts (GW) statewide by 2030, which would account for 20% of the state's peak load. As outlined in the 2019 Climate Leadership and Community Protection Act (CLCPA), the energy storage aim is intended to assist the state in meeting its goal of having a zero-emission electrical grid by 2040 [31].

The NYISO is investigating various use cases for storage as transmission to potentially define future operational guidelines. In September 2023, they evaluated several scenarios, including an N-1 contingency strategy that employs a 200 MW/200 MWh battery at the Shore Road 345 kV substation to provide operators with enough time to respond to a contingency, ensuring the line remains within its rating and reducing congestion. Another scenario involves a 50 MW/50 MWh battery at the Oswego complex near the Edic 345 kV Substation, intended to supply voltage support and maintain a steady Central East interface transfer capability when generators in the Oswego area are offline. Additionally, they explored a reduced local capacity requirement case by proposing a 200 MW/200 MWh battery in Zone J at the Mott Haven 345 kV substation, which would enhance transmission security limits in that zone, thus improving reliability and reducing the need for additional installed capacity. Lastly, the curtailment reduction scenario in Northern New York envisions assets that would charge and discharge on a daily schedule—charging during periods of high renewable generation and discharging when renewable output is low, with timing adjustments made seasonally. [32]

6.4. Battery Storage Capacity in Arizona

In 2023, Arizona had 923 MW of installed battery storage capacity, ranking third in the US behind California and Texas. The state plans to add 2,000 MW by 2025. Key projects include the Sonoran Solar Energy Center (260 MW solar with 1 GWh battery) and the Storey Energy Center (88 MW solar and battery). Batteries deployed in BESS currently use lithium and Arizona has some lithium deposits. An open-pit lithium mining project, Big Sandy, is in development in northern Arizona. These projects support Arizona's renewable energy goals and help meet electricity needs, including for data centers like Google's. Arizona is set for further growth in battery storage capacity with additional projects and investments in clean energy infrastructure. [33]

Figure 8 - Battery Storage for Grid in Arizona

6.5. Electric Vehicles

Electric vehicle (EV) is not the central topic of this report, yet their growing presence has significant implications for both the power grid and the energy storage market. Current projections suggest that passenger EVs could be responsible for roughly 80–90% of worldwide demand for lithium-ion batteries by 2030. EVs also have the potential to serve as storage resources within the grid by strategically charging and discharging at optimal times. This concept known as vehicle-grid integration (VGI) can include a bidirectional flow of electricity, where EV batteries send power back to the grid as the number of EVs on the road continues to expand. Combined U.S. sales of hybrid vehicles, plug-in hybrid electric vehicles, and battery electric vehicles (BEVs) increased from 17.8% of total new light-duty vehicle (LDV) sales in first quarter of 2024 to 18.7% in second quarter 2024, according to estimates from Wards Intelligence.[34]

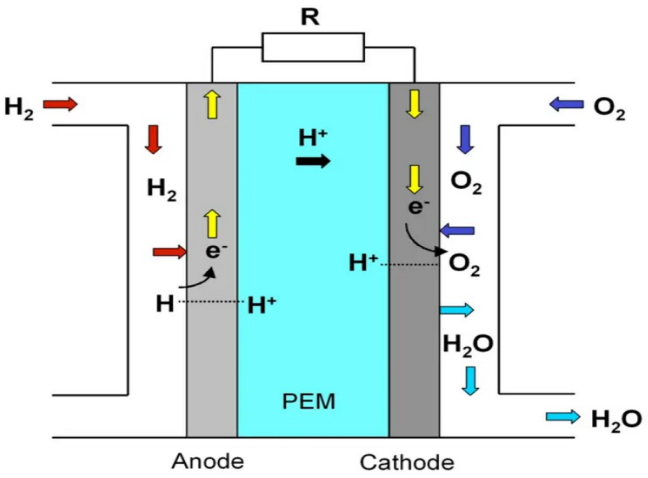
6.6. Fuel Cell

A fuel cell converts the chemical energy of hydrogen or other fuels into clean and efficient power. If hydrogen is the fuel, the only products are electricity, water, and heat. Fuel cells are unusual in terms of the breadth of their potential uses; they may employ a wide range of fuels and feedstocks [35]. Operators of several natural gas-fired power plants are exploring hydrogen as a supplement or replacement for natural gas. Fuel cells can be used to power a variety of applications, including transportation, industrial/commercial/residential structures, and long-term grid energy storage in reversible systems. Nasa used fuel cells to power the Apollo and Space shuttle mission [36].

Fuel cells provide significant advantages over traditional combustion-based technology utilized in many power plants and cars and can function at higher efficiencies than combustion engines, converting the chemical energy in fuel straight to electrical energy with efficiencies of more than 60%. Compared to combustion engines, fuel cells emit less or no emissions [37]. Hydrogen fuel cells emit just water, addressing key climate concerns with no carbon dioxide emissions. There are also no air contaminants that produce smog or health issues at the point of operation. Fuel cells are quiet because they have few moving parts. [38]

6.6.1. How Fuel Cells Work

Fuel cells function similarly to batteries, except they do not run out of power or require recharging. They generate electricity and heat so long as fuel is available. A fuel cell is made up of two electrodes, one negative (or anode) and one positive (or cathode), sandwiched by an electrolyte. The anode receives a fuel, such as hydrogen, while the cathode receives air. In a hydrogen fuel cell, a catalyst at the anode splits hydrogen molecules into protons and electrons, which travel separately to the cathode. The electrons pass through an external circuit, resulting in an electrical current. The protons flow through the electrolyte to the cathode, where they combine with oxygen and electrons to form water and heat.



The anode oxidizes hydrogen, whereas the cathode reduces oxygen. Fuel cells come in many sizes, types, and applications. Fuel cell power plants provide electricity for individual establishments and have potential applications in microgrids and isolated areas that are not connected to electric power grids. Several automotive manufacturers have developed fuel cells to power automobiles. Currently, three major types of fuel cells are under research or in use. Each type has distinct advantages, limitations, and possible applications. The following table outlines the major properties of the various types of fuel cells:

Table 5 - Different Types of Fuel Cells

Fuel cell type	Operating temperature	Cathode gas	Electrolyte	Anode gas
Proton exchange membrane	50-100 °C	Air ($O_2 + N_2$)	Polymer (H+)	H_2
Alkaline	50-100 °C	Air ($O_2 + N_2$)	KOH	H_2
Solid oxide	500-1000 °C	Air ($O_2 + N_2$)	ZrO_2 (O^{2-})	CH_4, CO

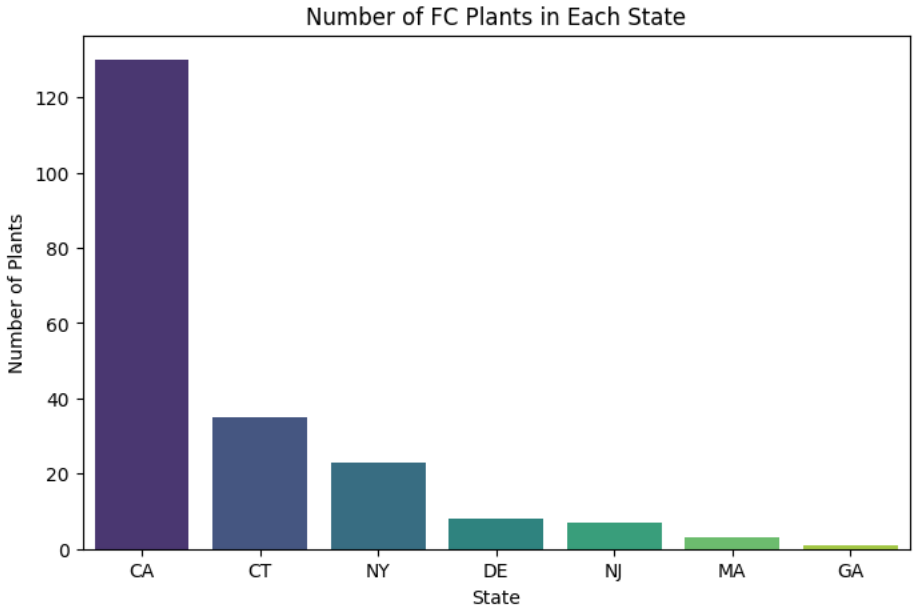
6.6.2. Proton Exchange Membrane (PEM) Fuel Cells

These fuel cells are the most common and use a solid polymer as an electrolyte and have porous carbon electrodes integrated with a platinum catalyst. They run on only hydrogen, oxygen from the air, and water, avoiding the requirement for corrosive fluids that other fuel cells may require. Typically, these fuel cells are fueled by pure hydrogen obtained from storage tanks or onboard reformers. PEM fuel cells operate at comparatively low temperatures of 50-100°C. However, the platinum catalyst is extremely vulnerable to CO poisoning. As a result, an extra reactor is required to remove CO in the fuel gas if the hydrogen is derived from alcohol or hydrocarbon fuels. To overcome this issue, developers are looking on platinum/ruthenium catalysts that are more resistant to CO.

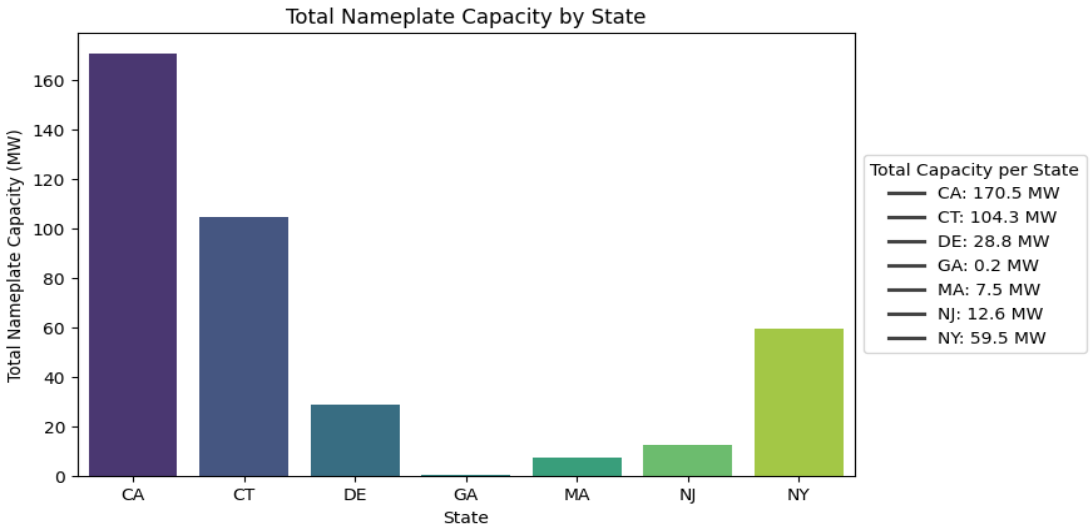
6.7. Total Capacity

As of the end of March 2024, the United States had around 210 active fuel cell electric power generators at 151 locations, totaling 384 megawatts (MW) of nameplate electric generation capacity. The nameplate capacities range from the largest single-fuel cell, with around 17 MW nameplate capacity, at the Bridgeport Fuel Cell, LLC plant in Connecticut, to ten fuel cells with 0.1 MW capacity each at the California Institute of Technology. Most functioning fuel cells use pipeline natural gas to generate hydrogen, although five use biogases from wastewater treatment and one uses landfill gas. In 2022, fuel cells accounted for less than 1% of total annual US electricity generation.

6.8. State Level Plants



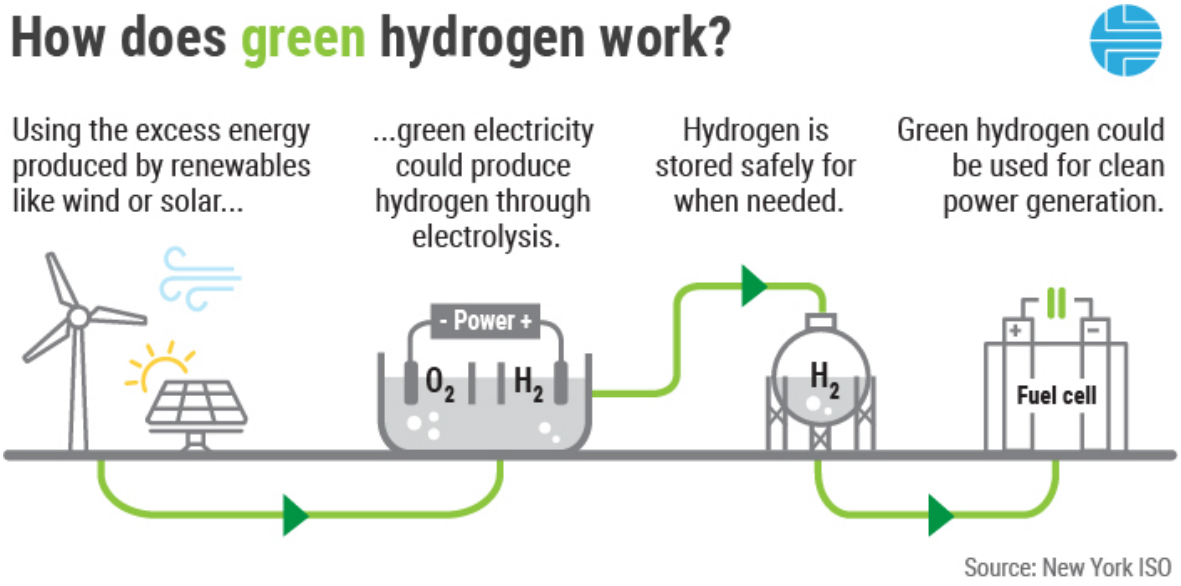
6.9. Capacity from Each State



6.10. The Future of Green Hydrogen for Fuel Cells

As New York strives to achieve a zero-emission energy grid over the next two decades, hydrogen could be a key player, and it might just be a green one. Green hydrogen is produced by splitting water using renewable sources like solar and wind power. It also offers a smart way to store energy from renewable sources. Unlike batteries that are ideal for short-term energy storage, hydrogen can be kept for months or even years. Imagine strong winds in northern New York, where many wind farms are located [39]. Sometimes, the wind is so powerful that the turbines generate more electricity than the grid can handle.

Traditionally, the excess power would be curtailed meaning the turbines would be slowed down or halted to keep the system balanced. Now, by channeling that surplus energy into hydrogen production, the excess electricity is used to create hydrogen, which can be stored similarly to how natural gas is kept underground. This stored hydrogen can then be tapped when energy demand peaks, helping to stabilize the grid [40].



6.11. Challenges Faced by Fuel Cells

Cost and durability remain the primary challenges for commercializing fuel cells, although the specific obstacles differ depending on their use.

- Cost. To compete with conventional technologies, fuel cell systems need significant cost reductions. Today, automotive internal-combustion engines cost about $25–$35 per kW, and fuel cells for transportation must reach around $30 per kW. Stationary applications face even higher price points, typically $400–$750 per kW for widespread use and up to $1000 per kW initially.
- Durability and Reliability. Fuel cells have yet to demonstrate the durability required for commercial acceptance. Transportation systems, for example, must match the longevity of current engines (approximately 5,000 hours or 150,000 miles) and operate reliably across a broad temperature range (40°C to 80°C). Stationary systems need to perform reliably for more than 40,000 hours, even in temperatures ranging from –35°C to 40°C.
- System Size. The current size and weight of fuel cell systems, including all supporting components like fuel processors, compressors, and sensors, must be reduced to meet the compact packaging needs of automobiles.
- Air, Thermal, and Water Management. Effective air management is difficult because existing compressors are not well suited for automotive fuel cell applications. Additionally, managing heat and water is challenging due to the small temperature difference between operating and ambient conditions, which necessitates the use of large heat exchangers.
- Improved Heat Recovery Systems. The relatively low operating temperatures of PEM fuel cells limit the recoverable heat for combined heat and power applications. Developing technologies that enable higher operating temperatures or more efficient heat recovery—and innovative system

designs—could improve CHP efficiencies beyond 80%. Evaluating cooling solutions that leverage the low waste heat, such as through regenerating desiccants, is also essential.

7. Micro Grid

Microgrids are versatile systems composed of distributed energy resources and loads that can seamlessly disconnect from and reconnect to the main utility grid. This capability ensures that connected loads receive power during outages. These systems are common in remote areas without access to a larger grid. While some backup generators can operate independently during outages, they are only classified as part of a microgrid if they serve additional functions, such as meeting daily power needs or participating in utility markets [41].

Microgrids function either continuously or conditionally under normal conditions. Continuous-operation microgrids regularly supply power to their host sites in both grid-connected and island modes, ensuring reliable power during prolonged outages. Conversely, conditional operation microgrids do not provide constant power while connected to the grid. Instead, their power supply depends on the availability of renewable resources or market signals like demand response or real-time pricing. This dual operational capability makes microgrids a critical component in enhancing energy resilience and supporting the integration of renewable energy sources [42].

Microgrids account for only a small portion of electricity generation in the United States. At the start of 2023, the United States had 692 microgrids installed, totaling about 4.4 gigawatts. More than 212 of them, totaling more than 419 MW, have come online in the last four years. The majority of microgrid projects are in Alaska, California, Georgia, Maryland, New York, Oklahoma, and Texas.

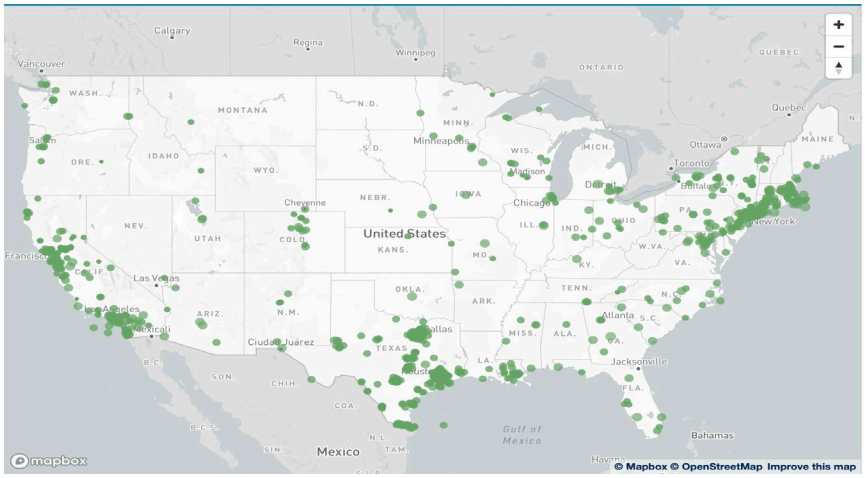


Figure 9 - The Concentration and Location of Different Microgrids Across U.S

Historically, microgrids produced electricity through fossil fuel-fired combined heat and power (CHP) and reciprocating engine generators. Today, however, projects are increasingly reliant on more sustainable resources such as solar power and energy storage. Microgrids can also serve as the primary electrical source for a hospital, university, or community. Microgrids are appealing to many significant U.S. corporations committed to working alone and in conjunction with governments to transition to a sustainable low-carbon economy. [43]

Technologies and applications of microgrid

Applications	
Agriculture	Farms that grow and process foods or other products
Airport	Commerical airports
City/Community	Multiple customers in a community or municipal setting
College/University	College and university campuses
Commercial	Commercial buildings not otherwise classified
Hospital/Healthcare	Buildings and campuses providing health care or assisted living services
Military	U.S. military bases and facilities
Multi-Family	Apartment buildings, public housing facilities, and housing cooperatives
Public Institution	Federal/state/municipal buildings that serve a public function
Research Facility	Research and development facilities and laboratories
Schools	K-12 schools
Water Utility/Treatment	Facilities that provide direct treatement of water
Other	Sites that do not fall into listed categories

Technologies		
CHP	CHP	Combined heat and power installation
BG	Biogas	Non-CHP biogas-fueled generator(s)
DSL	Diesel	Non-CHP diesel-fueled generator(s)
FC	Fuel Cell	Non-CHP fuel cell system
H	Hydro	Hydroelectric power facility
NG	Natural Gas	Non-CHP natural gas-fueled generator(s)
PV	Solar	Photovoltaic (PV) solar panels
STG	Storage	Energy storage technologies
W	Wind	Wind turbines
MCTRL	Advanced Controls	Advanced microgrid controls system

7.1.1. Benefits

Microgrids have several benefits to the environment, to utility operators, and to customers.

- Microgrids offer the opportunity to deploy more zero-emission electricity sources, thereby reducing greenhouse gas emissions.
- Microgrids can make use of on-site energy that would otherwise be lost through transmission lines and heat that would otherwise be lost up the smokestack.
- Microgrids can improve local management of power supply and demand, which can help defer costly investments by utilities in new power generation.
- Microgrids can enhance grid resilience to more extreme weather or cyber-attacks.

7.1.2. Challenges Facing Microgrids

- Complexity in Technology Integration. Microgrid projects typically combine multiple energy technologies such as solar, fuel cells, and others with varying scales of electric generation. Each project is unique in terms of the mix of technologies, load demands, geographic conditions, weather variability, and local market dynamics. This complexity makes it difficult for investors to apply a one-size-fits-all financial model.
- Financial and Tax Credit Variability. Although there are tax credits and preferential tax treatments for certain technologies (e.g., federal investment tax credit for solar and fuel cells), these benefits are not universally accessible. For instance, a municipality without a federal tax obligation might not be able to take advantage of the credit, necessitating highly customized financial solutions for each project.
- Legal and Regulatory Uncertainty. Most states do not have a standardized legal definition for a microgrid. This absence leads to:
    a) Activity Restrictions: Some laws outright prohibit or limit specific activities critical to microgrid operation.
    b) Increased Costs: Certain regulations may raise the cost of doing business.
    c) There is always the risk that new regulations could be introduced, imposing unforeseen restrictions or costs that were not anticipated during the project's planning and development stages.[44]

7.2. State Summary

State Summary for Arizona

Technology	Sites*	Capacity (KW)
Total	5	134,605
CHP	1	16,200
Diesel	5	99,480
Solar	3	17,675
Storage	2	1,250

* Most microgrids contain multiple technologies, so site totals by technology may be greater than the total number of microgrid sites in the state.

State Summary for New York

Technology	Sites*	Capacity (KW)
Total	105	580,427
CHP	60	448,478
Diesel	6	5,500
Fuel Cell	39	11,600
Hydro	2	1,700
Natural Gas	1	30
Solar	10	4,426
Storage	13	7,593
Wind	2	101,100

* Most microgrids contain multiple technologies, so site totals by technology may be greater than the total number of microgrid sites in the state.

8. Wind Energy

Wind turbines work on a simple principle that is instead of using electricity to make wind like a fan wind turbine use wind to make electricity [45]. The power available to a wind turbine is based on the density of the air (usually about 1.2 kg/m^3), the swept area of the turbine blades (picture a big circle being made by the spinning blades), and the velocity of the wind. Of these, clearly, the most variable input is wind speed, and the power output is defined by the equation below.

$$\text{Power(W)} = \frac{1}{2}\,\rho \times A \times V^3$$

- Power (SI Unit: Watts)
- ρ (rho, a Greek letter) = density of the air in kg/m^3
- A = cross-sectional area of the wind in m^2
- v = velocity of the wind in m/s [46]

Wind speed is a critical factor for generating electricity and initiating the rotation of wind turbine blades. According to the formula above and this assumption, we will first identify the wind speeds in the three states to evaluate the feasibility of wind turbine applications.

8.1. Average Wind Speed for 3 Cities

In general, wind turbines begin to produce power at wind speeds of about 6.7 mph (3 m/s). A turbine will achieve its nominal, or rated, power at approximately 26 mph to 30 mph (12 m/s to 13 m/s); this value is often used to describe the turbine's generating capacity (or nameplate capacity). The turbine will reach its cut-out speed at approximately 55 mph (25 m/s).

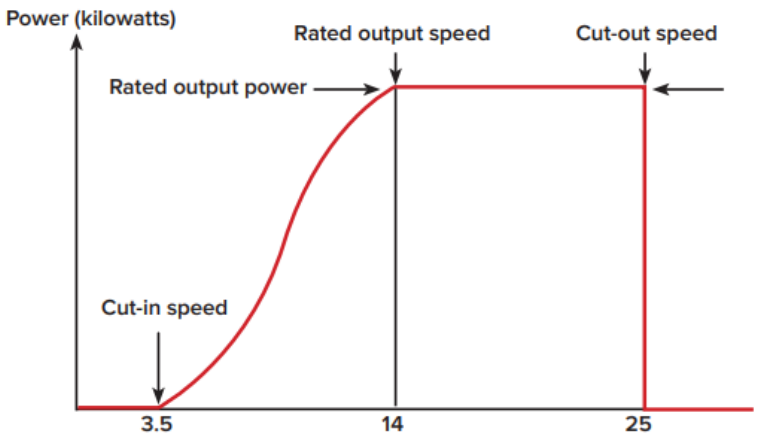


Figure 10 - Relationship of Wind Speed to Power Production (meter/second)

Based on NREL data to the wind speeds in three states are as follows.

States	Average Wind Speed(Onshore)	Average Wind Speed(Offshore)
New York	16.43 MPH	20 MPH
Pennsylvania	16.96 MPH	N/A
Arizona	12.81 MPH	N/A

Figure 11 - Average Wind Speed in Three States: Offshore and Onshore

We can observe that the average wind speed in New York and Pennsylvania is close to the nominal speed required to initiate turbine rotation and operate in a near-optimal regime. However, the average wind speed in Arizona is lower, which may pose a challenge during certain parts of the day in generating electricity.

8.2. Economic and Technological Factors Important for Wind Turbine Installation

The Levelized Cost of Electricity (LCOE) is typically higher for offshore wind turbines compared to onshore ones due to several factors, including higher initial construction and maintenance costs. Offshore wind farms require specialized infrastructure, more complex installation and maintenance operations, and longer distances for transmission of electricity [47].

Table 6 - Average LCOE for Wind Turbine Electricity Generation in Three States

States	Average Total LCOE (Land)($/MWh)	Average LCOE (Offshore) ($/MWh)
New York	77.38	124.83
Pennsylvania	58.3	N/A
Arizona	72.7	N/A

Pennsylvania shows a low levelized cost of electricity, indicating a strong investment potential for onshore wind turbines in the state compared to New York and Arizona.

8.3. Capacity Factors for Land and Offshore Base Turbines

The capacity factor of a wind turbine is its average power output divided by its maximum power capability. Capacity factor of onshore wind turbines in the U.S. ranges from 9% to 53% and averages 37% [48].

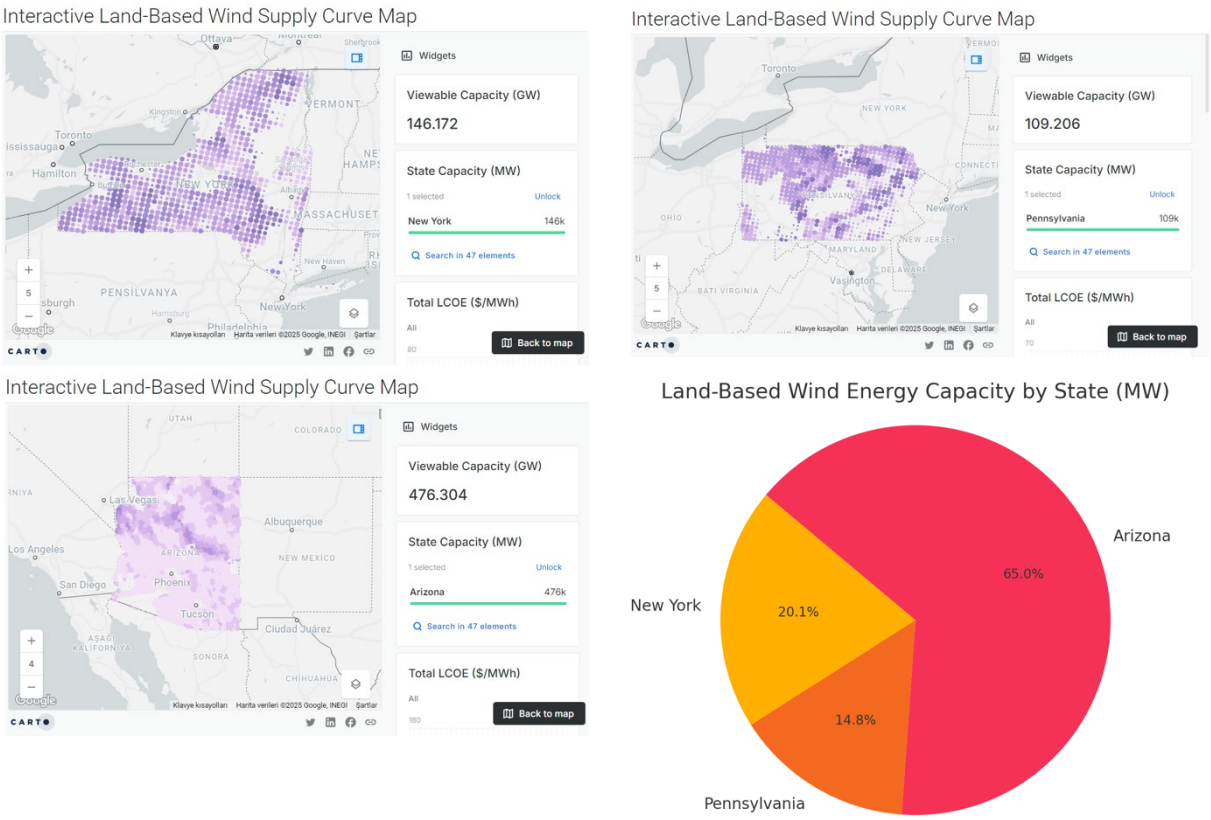
Table 7 - The Average Capacity Factor for Land and Offshore Wind Turbines

States	Average Capacity Factor(Land) (%)	Average Capacity Factor(Offshore) (%)
New York	0.444	0.5
Pennsylvania	0.463	N/A
Ariozna	0.262	N/A

The higher capacity factor of offshore wind turbines compared to onshore wind turbines is primarily due to the stronger and more consistent winds that occur over the sea. Offshore locations benefit from fewer obstructions like buildings or trees that can disrupt wind flow, leading to increased energy generation potential [49].

8.4. Total Capacity

Interactive map of the United States showing land-based wind supply curves. Reference Siting ATB Moderate for 2030 indicates a capacity range of 0.02 MW–398 MW, with an average of 137 MW. Capacity, mean wind speed, and mean capacity factor are all taken into consideration when determining the quantity and quality of land-based wind resources [50]. Viewable Capacity (GW) on the Interactive Land-Based Wind Supply Curve Map refers to the realistic amount of wind energy generation capacity, measured in gigawatts (GW), that can potentially be developed within the United States. Viewable capacity considers several critical factors like Siting Constraints, Technical Feasibility, and Economic Viability [51].



Viewable Capacity for Land-Based Wind Supply in 3 states

Based on the latest data from NREL's Wind Supply Curves, the viewable (developable) wind capacity varies significantly across states due to geographic, technical, and siting constraints. Arizona possesses the highest viewable capacity among the three states, with 476.304 gigawatts (GW) of potentially developable land-based wind energy, demonstrating its vast land availability.

8.4.1. Developable Area

This represents the total land area available and suitable for developing a wind energy project.

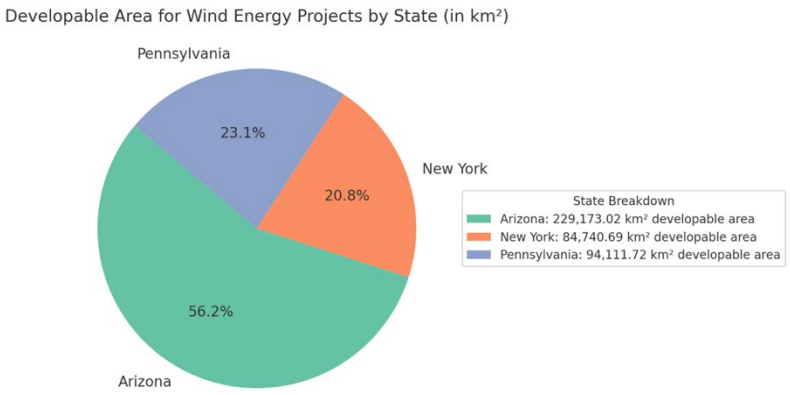
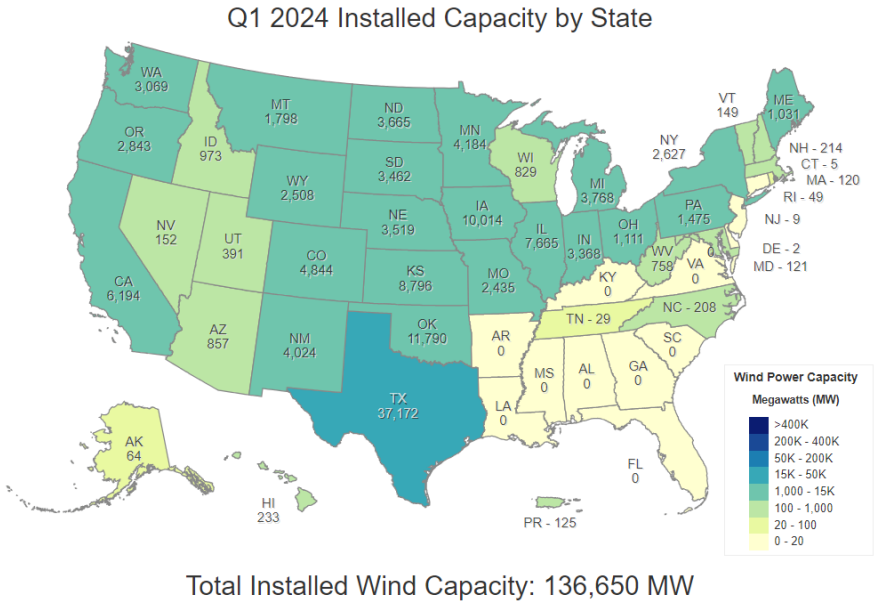


Figure 12 - Developable Land Area for Wind Energy in the Three States

This pie chart depicts the amount of developable area available in the three states. From the graph, we can see that Arizona has significantly more available area compared to Pennsylvania and New York for the application of wind turbines.



9. Nuclear Energy

The definition of clean energy does not always include nuclear energy, but it is one of the world's largest sources of low-carbon electricity, second only to hydropower. When considering greenhouse gas emissions, nuclear has the potential to be a key part of clean energy solutions [52]. The thermal efficiency, η_{th}, of any heat engine is defined as the ratio of the work it does, W to the heat input at the high temperature, Q_H

$$\eta_{th} = \frac{W}{Q_H}$$

The thermal efficiency, ηth, is the ratio of heat, QH, developed as work. It is a measure of the performance of a thermal energy utilizing engine such as a steam turbine, an internal combustion engine, or a refrigerator. The overall thermal efficiency in modern nuclear power plants is about one-third (33%), and therefore 3000 MW of thermal power produced during fission is needed to generate 1000 MWe of electrical power.

Nominal thermal power of an average reactor is approximately 3400MW. On April 30, 2024, there were 54 operating commercial nuclear power stations with 94 nuclear power reactors across 28 states. There are 19 Plants with a single reactor, 31 with two reactors, 4 with three reactors, and 1 with four reactors. Georgia's Alvin W. Vogtle Electric Generating Plant is the largest U.S. nuclear power plant with 4 reactors and an aggregate nameplate electricity generating capacity of around 4,658 MW and a total net summer electricity generating capacity of 4,530 MW. New York State's Ginna Nuclear Power Plant is the smallest single-reactor plant that is of 614 MW nameplate generation [53].

Table 8 - Nuclear Reactor, State, and Net Capacity [54].

State	Plant Name	Generator ID	2022 Summer Capacity Net MW(e)1
NY	R E Ginna Nuclear Power Plant	6122	580.3
NY	James A Fitzpatrick	6110	831.3
NY	Nine Mile Point Nuclear Station	2589	620.9
NY	Nine Mile Point Nuclear Station	2589	1272.1
PA	Beaver Valley	6040	907
PA	Beaver Valley	6040	901
PA	Limerick	6105	1119.7
PA	Limerick	6105	1122.1
PA	Peach Bottom	3166	1264.7
PA	Peach Bottom	3166	1284.7
PA	TalenEnergy Susquehanna	6103	1247
AZ	Palo Verde	6008	1312

9.1. Levelized Cost of Electricity (LCOE)

States	Average Total LCOE ($/MWh)
New York	84
Pennsylvania	84
Arizona	81

Neighbourhood	Average Total LCOE ($/MWh)
New Jersey(NY & PA)	85
Connecticut(NY)	87
Massachusetts(NY)	86
Vermont(NY)	82
Delaware(PA)	87
Maryland(PA)	87
West Virginia(PA)	83
Ohio(PA)	82
California(AZ)	86
Nevada(AZ)	84
Utah(AZ)	81
Colorado(AZ)	81
New Mexico(AZ)	81

LCOE for Nuclear Energy in New York, Pennsylvania, Arizona and Neighborhoods

Generally, Generation costs/kWh for new nuclear (including fuel & O&M but not distribution to customers) are likely to be from 25 – 30 cents/kWh. This high cost may destroy the very demand the plant was built to serve. High electric rates may seriously impact utility customers and make nuclear utilities' service areas noncompetitive with other regions of the U.S. which are developing lower-cost electricity [55].

10. Hydropower

Hydropower, or hydroelectric power, is a renewable source of energy that generates power by using a dam or diversion structure that alter the natural flow of a river or other body of water. Hydropower relies on the endless, constantly recharging system of the water cycle to produce electricity, using a fuel water that is not reduced or eliminated in the process [56].

World-wide, about 20% of all electricity is generated by hydropower. Hydropower provides about 10% of the electricity in the United States. The United States is the second largest producer of hydropower in the world after Canada.

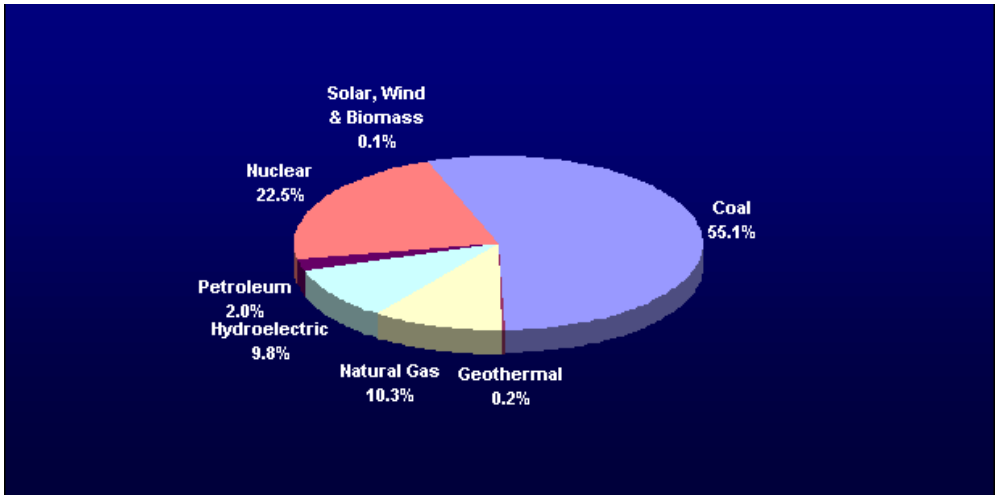


Figure 13 - U.S. Electric Utility Net Generation of Electricity

Recent data shows that in Wisconsin hydropower is produced for less than one cent per kwh. This is about one-half the cost of nuclear and one-third the cost of fossil fuel. Hydropower is the most efficient way to generate electricity. Modern hydro turbines can convert as much as 90% of the available energy into electricity. The best fossil fuel plants are only about 50% efficient.

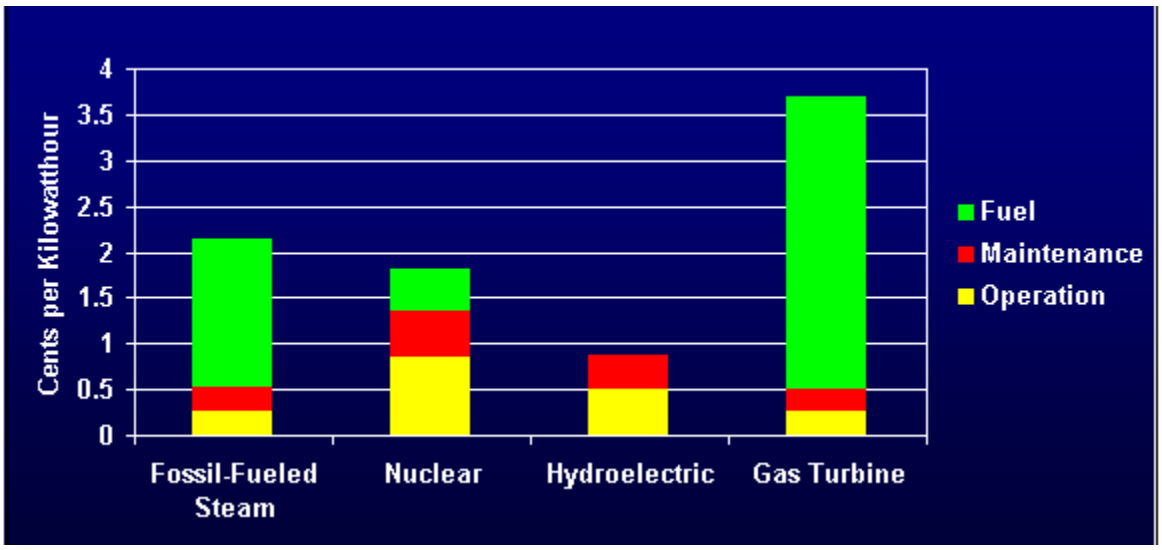


Figure 14 - Average Power Production Expense per KWh

Hydropower is the leading source of renewable energy. It provides more than 97% of all electricity generated by renewable sources. Other sources including solar, geothermal, wind, and biomass account for less than 3% of renewable electricity production [57].

10.1. Power Calculations

$$Power = \frac{(Height of Dam) \times (River Flow) \times (Efficiency)}{11.8}$$

Variables	Description
Power	The electric power in kilowatts (one kilowatt equals 1,000 watts)
Height of Dam	The distance the water falls measured in feet.
River Flow	The amount of water flowing in the river measured in cubic feet per second.
Efficiency	How well the turbine and generator convert the power of falling water into electric power. For older, poorly maintained hydro-plants this might be 60% (0.60) while for newer, well operated plants this might be as high as 90% (0.90).
11.8	Converts units of feet and seconds into kilowatts

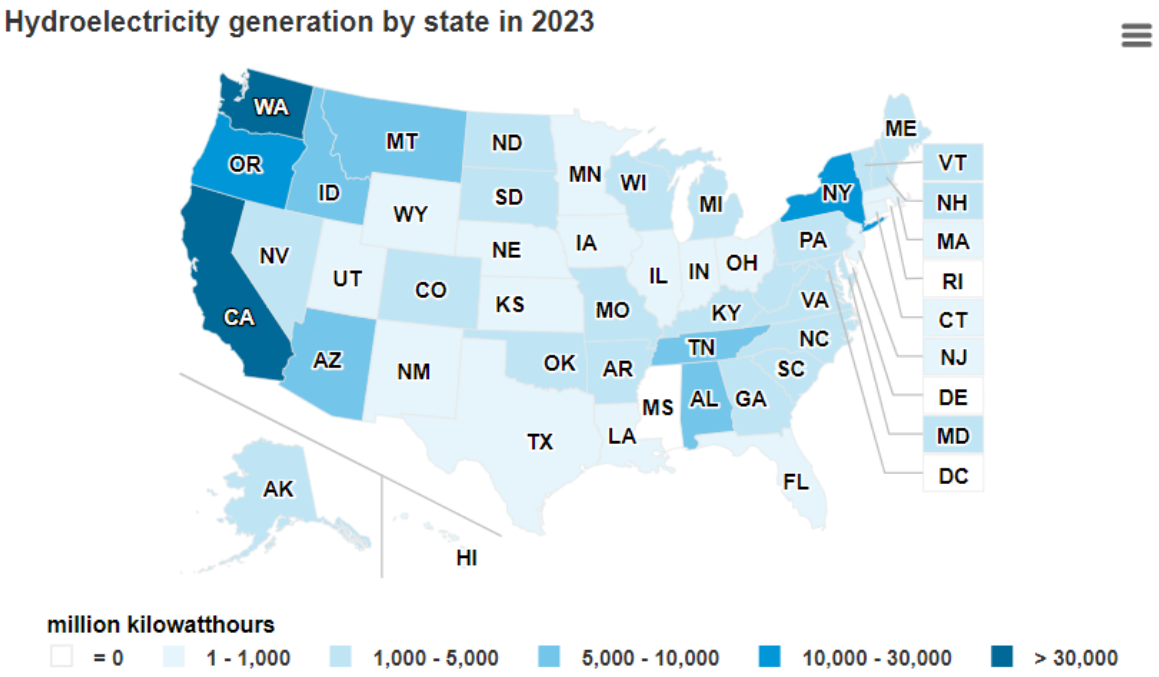


Figure 15 - U.S. Energy Information Administration, Electric Power Monthly, February 2024

The top five states and their percentage shares of U.S. total conventional hydroelectricity net summer generation capacity in 2023 were:[2]

27%	13%	10%	6%	4%
Washington	California	Oregon	New York	Alabama

Capacity Factor

The capacity factor for conventional hydropower resources across the US has remained relatively consistent through the years, between 35 and 45%.

Total Capacity

States	Average Capacity Per State (GWh)
New York	27,700
Pennsylvania	2,731
Arizona	5,805

Neighborhood	Average Capacity Per State (GWh)
California (AZ)	31,887
Nevada (AZ)	1,303

Hydropower energy is more available in New York compared to Pennsylvania and Arizona. Even though Pennsylvania has minor sources of hydropower, they are insufficient to meet the state's high energy requirement. With this consideration, the most appropriate option is New York. Moreover, in consideration of the possibilities of getting energy transmissions from neighboring states, the other neighboring states relevant to Arizona are also discussed in this paper.

Levelized Cost of Electricity (LCOE)

States	Average Total LCOE (Hydropower) ($/MWh)
New York	14
Pennsylvania	13
Arizona	11
Neighborhood	Average Total LCOE (Hydropower) ($/MWh)
California (AZ)	14
Nevada (AZ)	23

Overall, Hydropower produces very low-cost electricity over its long lifetime, despite having relatively high up-front construction costs. The weighted global average cost of electricity from hydropower plants in 2022 was US$0.061 per kWh, the lowest-cost source of electricity in most markets (IRENA). [58].

11. Geothermal Energy

Geothermal energy is heat within the earth. The word geothermal comes from the Greek words geo (earth) and therme (heat). Geothermal energy is a renewable energy source because heat is continuously produced inside the earth. People use geothermal heat for bathing, for heating buildings, and for

generating electricity [59]. In 2023, the United States had geothermal power plants in seven states, which produced about 0.4% (17 billion kilowatt-hours) of total U.S. utility-scale electricity generation. Utility-scale power plants have at least 1,000 kilowatts (1 megawatt) of electricity generation capacity [60].

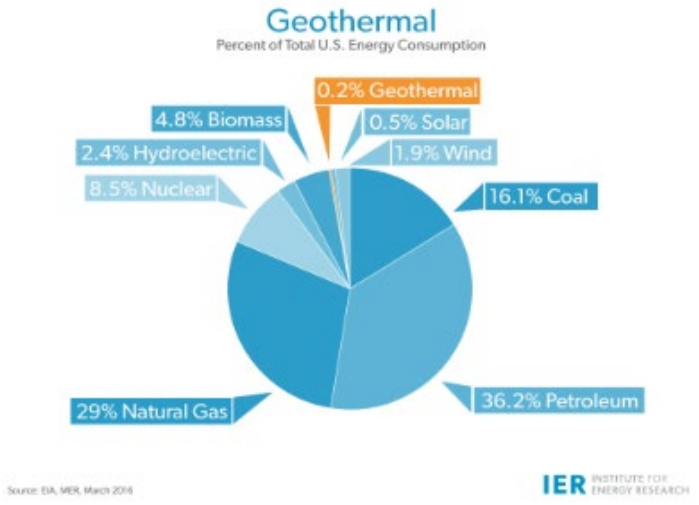
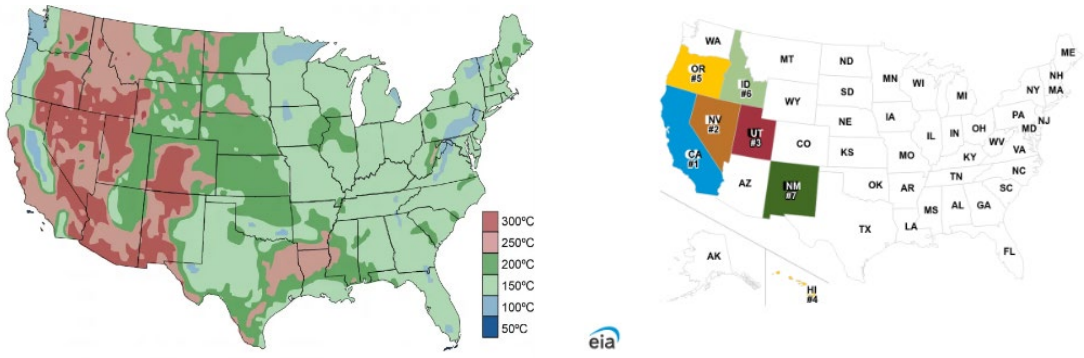


Figure 16 - Percent of Total U.S. Energy Consumption

Even though geothermal energy sources are limited for energy production, they are available 24 hours a day, 365 days a year, regardless of weather conditions, making geothermal one of the most reliable energy sources.



U.S. Geothermal Resources at 10 km depth and state ranking

The left map displays the geothermal resource temperature across the U.S., with the highest potential (≥250°C) occurring in the western states, and the right map displays the top 7 geothermal electricity-generating states—California, Nevada, and Utah leading the pack—indicating a close correlation between subsurface high temperatures and actual geothermal power generation.

Geothermal power plants have a high-capacity factor typically 90% or higher meaning that they can operate at maximum capacity nearly all the time. These factors mean that geothermal can balance

intermittent sources of energy like wind and solar, making it a critical part of the national renewable energy mix.

	State share of total U.S. geothermal electricity generation	Geothermal share of total state electricity generation
California	66.6%	5.1%
Nevada	26.1%	10.1%
Utah	3.2%	1.5%
Hawaii	2.1%	3.7%
Oregon	1.3%	0.4%
Idaho	0.5%	0.6%
New Mexico	0.2%	0.1%

Figure 17 - States with Geothermal Power Plants in 2023

State	Production (TWh)	Share of U.S total
California	10,962	66.6%
Nevada	4,296	26.1%
Utah	521	3.2%
Hawaii	348	2.1%
Oregon	212	1.3%
Idaho	86	0.5%
New Mexico[23]	36	0.2%
Total	16,462	100%

Figure 18 - Geothermal Production in Terrawatt Hours (TWh) by State as of December 2023

Although geothermal energy is not readily available within the three states, Arizona has better accessibility to geothermal resources due to its proximity to geothermally active neighboring states.

11.1. Capacity Factor

The capacity factor of geothermal power plants stood at an average of 69.4 percent in 2023, slightly increasing from the previous year. Geothermal energy has the highest capacity factor among renewable energy sources [61].

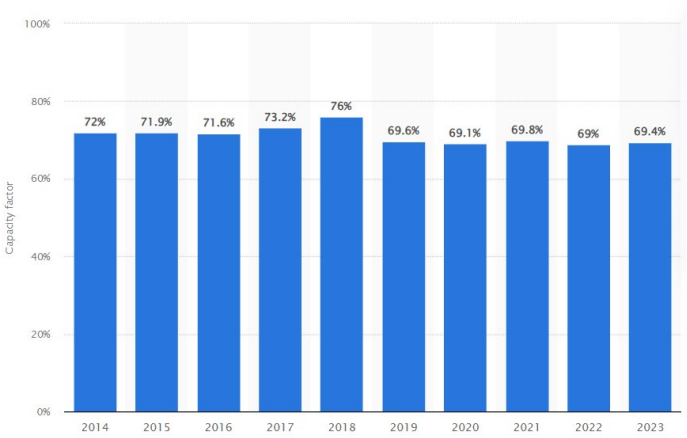


Figure 19 - Capacity Factor of Geothermal Power Plants in the United States from 2014 to 2023

Levelized Cost of Electricity (LCOE)

States	Average Total LCOE (Geothermal) ($/MWh)
New York	N/A
Pennsylvania	N/A
Arizona	326

Neighborhood	Average Total LCOE (Hydropower) ($/MWh)
California (AZ)	61
Neveda (AZ)	61
Utah (AZ)	61
New Mexico (AZ)	241

Generally, for geothermal power projects, between 2022 and 2023 the global weighted average LCOE of the seven projects commissioned increased by 23%, to USD 0.071/kWh [27].

12. Conclusion and STEEP Analysis of Backup Power Options

12.1. Pennsylvania

Pennsylvania	Factors	Solar Energy	Wind Turbine	Nuclear Energy	Hydropower	Geothermal	Fuel Cells	Battery Storage(Li-ion)
Social	Health	2	0	-1	0	0	0	0
	Community Acceptance	2	1	-2	0	1	1	1
Technological	Reliability	-1	-1	2	2	1	0	1
	Power Density	1.25	1	2	-1	0.5	1.5	1
Environmental	Waste Generation	-1	0	-2	2	1.5	1	-1
	Land Use & Wildlife	-1	-2	0	-1	2	1	-1
Economical	LCOE	1	1.5	1	1.75	(-)	-1.75	-1
	REC	2	1.5	(-)	1	0.5	1	(-)
Political	Incentives	0.75	0.75	0.5	1	0.5	1	0.5
	Regulations/Permits	1	1	1	1	0.75	0.75	0.75
Availability	Total Capacity	0.75	1	(-)	0.5	(-)	(-)	0.5
	Resource Availability	1.5	1.25	2	1	(-)	(-)	0.5

Nuclear and solar excel in power density in Pennsylvania. Hydropower is economically superior with lowest Levelized Cost of Energy (LCOE) and highest Renewable Energy Credit (REC) value. Solar has the highest REC values and medium incentives. Fuel cells are faced with negative LCOE, which makes them costly. Politically, all technologies are moderately favored but solar and hydropower are slightly favored with higher permit and incentive case. Nuclear is leading in resource availability and geothermal and fuel cells are highly constrained in this aspect. Overall, from the STEEP analysis of Pennsylvania, the results indicate that the best technology is nuclear power, solar power, and battery technology. However, the choice should be based on the average power demand of the system.

12.2. New York

New York	Factors	Solar Energy	Wind Turbine	Nuclear Energy	Hydropower	Geothermal	Fuel Cells	Battery Storage(Li-ion)
Social	Health	2	0	-1	0	0	0	0
	Community Acceptance	2	1	-2	0	1	1	1
Technological	Reliability	-1	-1	2	2	1	0	1
	Power Density	1	1.25	2	-1	0.5	1.5	1
Environmental	Waste Generation	-1	0	-2	2	1.5	1	-1
	Land Use & Wildlife	-1	-2	0	-1	2	1	-1
Economical	LCOE	1	1	1	1.75	(-)	-1.5	1
	REC	(-)	1.5	(-)	1	0.5	1.5	(-)
Political	Incentives	1.5	1.5	1	2	0.5	1.5	1.5
	Regulations/Permits	1.75	1.5	1.25	1.5	1.25	1.25	1.75
Availability	Total Capacity	1	1.5	(-)	2	(-)	1.25	1
	Resource Availability	1.5	1.75	1.25	2	(-)	1	1

In New York, hydropower stands out with the highest scores in both LCOE and total capacity, making it a highly economical and available option. Offshore wind also performs very well, especially in resource availability and permitting ease, making it a leading choice among wind technologies. Solar and battery storage benefit from strong political support through top incentives and regulations, which enhances their deployment potential. Fuel cells face challenges due to negative LCOE, while geothermal suffers from poor resource availability. Overall, hydropower, offshore wind, solar, and battery storage show strategic advantages in New York's clean energy mix. In conclusion, the STEEP analysis highlights hydropower, offshore wind, nuclear, solar energy, and battery storage as the most favorable options due to their strong economic, political, and technological indicators. However, the final energy mix should be tailored to New York's average power demand and grid requirements to ensure optimal performance and reliability.

12.3. Arizona

Arizona	Factors	Solar Energy	Wind Turbine	Nuclear Energy	Hydropower	Geothermal	Fuel Cells	Battery Storage(Li-ion)
Social	Health	2	0	-1	0	0	0	0
	Community Acceptance	2	1	-2	0	1	1	1
Technological	Reliability	-1	-1	2	2	1	0	1
	Power Density	2	0.75	2	-1	0.5	1.5	1
Environmental	Waste Generation	-1	0	-2	2	1.5	1	-1
	Land Use & Wildlife	-1	-2	0	-1	2	1	-1
Economical	LCOE	1.75	1.25	1.25	2	-2	-1.75	2
	REC	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Political	Incentives	1	1	0	1.5	1	0.5	1
	Regulations/Permits	1.5	0.5	0.5	0.25	0.25	0.5	1
Availability	Total Capacity	2	1.75	(-)	0.75	0.25	(-)	2
	Resource Availability	2	1.5	1	1.5	0.25	(-)	2

Solar power is the leader in Arizona with the highest rating for power density, LCOE, and overall capacity, reflecting its strong economic viability and high resources available. Its storage is also affordable and accessible, supplementing the balance of intermittent renewables. The performance of wind power is moderate, and that of nuclear and hydropower is also reliable but less in terms of overall capacity and resource availability. Geothermal is both economically and resource-restricted, and fuel cells are low both economically and on the politics. Political favor is for solar and hydropower with excellent incentives and permitting ease, and the others receive stricter regulatory requirements. Solar power and battery storage are most strategically advantageous options in Arizona's energy mix overall. In conclusion, based on the STEEP analysis for Arizona, solar, hydropower, and geothermal offer the best outcomes. However, the final choice will have to be dependent on the system's mean power need in a bid to achieve the highest reliability and operation.

Weighting for Each Color Code

Scale Level	Description
2	Best Case
1	Generally Acceptable
0	Neutral
-1	Needs improvement
-2	Worst Case

13. Abbreviations

AI	artificial intelligence
BEV	battery electric vehicle
BTU	british thermal unit
CHP	combined heat and power
CLCPA	Climate Leadership and Community Protection Act
CSP	concentrating solar-thermal power
DOE	U.S. Department of Energy

EDC	electricity distribution company
EIA	Energy Information Administration
EV	electric vehicle
FERC	Federal Energy Regulatory Commission
GHI	global horizontal irradiance
GW	gigawatt
LCOE	levelized cost of electricity
LCOT	levelized cost of transmission (LCOT)
LDV	light-duty vehicle
MW	megawatt
NERC	North American Electric Reliability Corporation
NPCC	Northeast Power Coordinating Council
NREL	National Renewable Energy Laboratory
NYISO	New York Independent System Operator
PEM	proton exchange membrane
PSC	Public Service Commission
PUC	Pennsylvania Public Utilities Commission
PV	photovoltaics
REC	renewable energy credit
RF	ReliabilityFirst
RTO	regional transmission organization
SATA	storage as a transmission asset
SATOA	storage as a transmission only asset
SCTE	Society of Cable Telecommunications Engineers
SRSG	Southwest Reserve Sharing Group
STEEP	social, technological, environmental, economic, and political
TWh	Terrawatt Hours
VER	variable energy resource
WECC	Western Electricity Coordinating Council
YOY	year-over-year
$/MWh	cost per unit of electricity

14. Bibliography and References

[1] EIA, “Electricity explained,” https://www.eia.gov/energyexplained/electricity/.

[2] EPA, “Global Greenhouse Gas Overview,” https://www.epa.gov/ghgemissions/global-greenhouse-gas-overview#:~:text=(Note:%20Emissions%20from%20industrial%20electricity,instead%20of%20the%20Energy%20sector.

[3] “PJM - Regional Transmission Expansion Plan (RTEP).” Accessed: May 19, 2025. [Online]. Available: https://www.pjm.com/library/reports-notices/rtep-documents.aspx

[4] “Planning - NYISO.” Accessed: May 19, 2025. [Online]. Available: https://www.nyiso.com/planning

[5] NERC, “Reliability Assessments,” https://www.nerc.com/pa/RAPA/ra/Pages/default.aspx.

[6] Lawrence Livermore National Laboratory, “Energy Flow Charts: Charting the Complex Relationships among Energy, Water, and Carbon,” https://flowcharts.llnl.gov/sites/flowcharts/files/2024-12/energy-2023-united-states.png.

[7] “After more than a decade of little change, U.S. electricity consumption is rising again - U.S. Energy Information Administration (EIA).” Accessed: May 19, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=65264

[8] “EIA extends five key energy forecasts through December 2026 - U.S. Energy Information Administration (EIA).” Accessed: May 19, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=64264

[9] “U.S. Energy Information Administration - EIA - Independent Statistics and Analysis.” Accessed: May 19, 2025. [Online]. Available: https://www.eia.gov/environment/emissions/carbon/

[10] “U.S. energy-related CO2 emissions decreased by 3% in 2023 - U.S. Energy Information Administration (EIA).” Accessed: May 19, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=61928

[11] “Coal and natural gas plants will account for 98% of U.S. capacity retirements in 2023 - U.S. Energy Information Administration (EIA).” Accessed: May 19, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=55439

[12] “Electric Power Outlook Report | PA PUC.” Accessed: May 20, 2025. [Online]. Available: https://www.puc.pa.gov/filing-resources/reports/electric-power-outlook-report/

[13] “Report Card on Pennsylvania’s Electricity Generation Infrastructure | Alpha 3 Consulting, LLC.” Accessed: May 20, 2025. [Online]. Available: https://www.alphathree.com/news/report-card-pennsylvanias-electricity-generation-infrastructure

[14] “New York Electricity Profile 2023 - U.S. Energy Information Administration (EIA).” Accessed: May 20, 2025. [Online]. Available: https://www.eia.gov/electricity/state/NewYork/

[15] “Grid Ready Mapping Tool.” Accessed: May 20, 2025. [Online]. Available: https://maps.urbangreencouncil.org/

[16] “Can Wind, Water and Sunlight Power New York by 2050? - The New York Times.” Accessed: May 20, 2025. [Online]. Available: https://archive.nytimes.com/dotearth.blogs.nytimes.com/2013/03/12/can-wind-water-and-sunlight-power-new-york-by-2050/

[17] “Grid Ready: Powering NYC’s All-Electric Buildings - Urban Green Council.” Accessed: May 20, 2025. [Online]. Available: https://www.urbangreencouncil.org/grid-ready-powering-nycs-all-electric-buildings/

[18] “State Energy Profile Data.” Accessed: May 20, 2025. [Online]. Available: https://www.eia.gov/state/data.php?sid=AZ

[19] “Electrification Futures Study | Energy Analysis | NREL.” Accessed: May 20, 2025. [Online]. Available: https://www.nrel.gov/analysis/electrification-futures

[20] C. Murphy *et al.*, "High electrification futures: Impacts to the U.S. bulk power system," *Electricity Journal*, vol. 33, no. 10, Dec. 2020, doi: 10.1016/J.TEJ.2020.106878.

[21] "Solar Irradiance and Solar Irradiation - An Overview." Accessed: May 20, 2025. [Online]. Available: https://www.alternative-energy-tutorials.com/solar-power/solar-irradiance.html#google_vignette

[22] "Understanding transitions to solar energy on Pennsylvania farmland | Institute of Energy and the Environment." Accessed: May 20, 2025. [Online]. Available: https://iee.psu.edu/news/blog/understanding-transitions-solar-energy-pennsylvania-farmland

[23] A. Martinez and G. Iglesias, "Mapping of the levelised cost of energy for floating offshore wind in the European Atlantic," *Renewable and Sustainable Energy Reviews*, vol. 154, p. 111889, Feb. 2022, doi: 10.1016/j.rser.2021.111889.

[24] K. Branker, M. J. M. Pathak, and J. M. Pearce, "A Review of Solar Photovoltaic Levelized Cost of Electricity," *Renewable and Sustainable Energy Reviews*, vol. 15, no. 9, pp. 4470–4482, Dec. 2011, doi: 10.1016/j.rser.2011.07.104.

[25] "Solar - IEA." Accessed: May 20, 2025. [Online]. Available: https://www.iea.org/energy-system/renewables/solar-pv

[26] "Solar State By State – SEIA." Accessed: May 20, 2025. [Online]. Available: https://seia.org/solar-state-by-state/

[27] IRENA, "Renewable power generation costs in 2023: Executive summary," https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2024/Sep/IRENA_Renewable_power_generation_costs_in_2023_executive_summary.pdf.

[28] "Utility-Scale Battery Storage | Electricity | 2024 | ATB | NREL." Accessed: May 20, 2025. [Online]. Available: https://atb.nrel.gov/electricity/2024/utility-scale_battery_storage

[29] "U.S. battery storage capacity will increase significantly by 2025 - U.S. Energy Information Administration (EIA)." Accessed: May 20, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=54939

[30] "PJM - Regional Transmission Expansion Plan (RTEP)." Accessed: May 20, 2025. [Online]. Available: https://www.pjm.com/library/reports-notices/rtep-documents.aspx

[31] "Approval of New York's Nation-Leading Six Gigawatt Energy Storage Roadmap Announced - NYSERDA." Accessed: May 20, 2025. [Online]. Available: https://www.nyserda.ny.gov/About/Newsroom/2024-Announcements/2024_06_20-Governor-Hochul-Announces-Approval-Of-New-Yorks-Nation-Leading

[32] "Storage as a Grid Asset – NYSERDA." Accessed: May 20, 2025. [Online]. Available: https://gridconnect.nyserda.ny.gov/nygc-home/priorities/storage-as-a-grid-asset/

[33] "Solar, battery storage to lead new U.S. generating capacity additions in 2025 - U.S. Energy Information Administration (EIA)." Accessed: May 21, 2025. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=64586

[34] NRELS, "Projecting Electric Vehicle Electricity Demands and Charging Loads," https://docs.nrel.gov/docs/fy24osti/89775.pdf.

[35] "Hydrogen and Fuel Cell Technologies Office | Department of Energy." Accessed: May 21, 2025. [Online]. Available: https://www.energy.gov/eere/fuelcells/hydrogen-and-fuel-cell-technologies-office

[36] "NASA exploring space applications of hydrogen and fuel cells – Climate Change: Vital Signs of the Planet." Accessed: May 21, 2025. [Online]. Available: https://climate.nasa.gov/news/788/nasa-exploring-space-applications-of-hydrogen-and-fuel-cells/

[37] "Hydrogen and Fuel Cells | NREL." Accessed: May 21, 2025. [Online]. Available: https://www.nrel.gov/hydrogen/

[38] NASA Glenn Research Center, "NASA Activities in Hydrogen and Fuel Cell Technologies," https://ntrs.nasa.gov/api/citations/20220004667/downloads/NASA%20Activities%20in%20Hydrogen%20and%20Fuel%20Cell%20Technologies%20-%2020220320.pdf.

[39] "Innovation at NYSERDA Power Generation & Storage Program - NYSERDA." Accessed: May 21, 2025. [Online]. Available: https://www.nyserda.ny.gov/All-Programs/Innovation-at-NYSERDA/Focus-Areas/Power-Generation-and-Storage

[40] "Wind-to-Hydrogen Project | Hydrogen and Fuel Cells | NREL." Accessed: May 21, 2025. [Online]. Available: https://www.nrel.gov/hydrogen/wind-to-hydrogen

[41] "Department of Energy Releases New Tool Tracking Microgrid Installations in the United States | Department of Energy." Accessed: May 21, 2025. [Online]. Available: https://www.energy.gov/eere/iedo/articles/department-energy-releases-new-tool-tracking-microgrid-installations-united

[42] "Microgrid and CHP Installation Databases | Better Buildings Initiative." Accessed: May 21, 2025. [Online]. Available: https://betterbuildingssolutioncenter.energy.gov/onsite-energy/microgrid-and-chp-installation-databases

[43] "U.S. Department of Energy Combined Heat & Power and Microgrid Installation Databases." Accessed: May 21, 2025. [Online]. Available: https://doe.icfwebservices.com/index

[44] "Microgrids Could Enhance Grid Resilience | NREL." Accessed: May 21, 2025. [Online]. Available: https://www.nrel.gov/news/detail/program/2025/microgrids-could-enhance-grid-resilience

[45] "How Do Wind Turbines Work? | Department of Energy." Accessed: May 21, 2025. [Online]. Available: https://www.energy.gov/eere/wind/how-do-wind-turbines-work

[46] "Wind Energy and Power Calculations | EM SC 470: Applied Sustainability in Contemporary Culture." Accessed: May 21, 2025. [Online]. Available: https://www.e-education.psu.edu/emsc297/node/649

[47] S. Tumse, M. Bilgili, A. Yildirim, and B. Sahin, "Comparative Analysis of Global Onshore and Offshore Wind Energy Characteristics and Potentials," *Sustainability*, vol. 16, no. 15, p. 6614, Aug. 2024, doi: 10.3390/su16156614.

[48] C. A. S. Hall, J. G. Lambert, and S. B. Balogh, "EROI of different fuels and the implications for society," *Energy Policy*, vol. 64, pp. 141–152, Jan. 2014, doi: 10.1016/j.enpol.2013.05.049.

[49] B. Desalegn, D. Gebeyehu, B. Tamrat, T. Tadiwose, and A. Lata, "Onshore versus offshore wind power trends and recent study practices in modeling of wind turbines' life-cycle impact assessments," *Clean Eng Technol*, vol. 17, p. 100691, Dec. 2023, doi: 10.1016/j.clet.2023.100691.

[50] A. Lopez, T. Mai, E. Lantz, D. Harrison-Atlas, T. Williams, and G. Maclaurin, "Land use and turbine technology influences on wind potential in the United States," *Energy*, vol. 223, May 2021, doi: 10.1016/j.energy.2021.120044.

[51] T. Mai, A. Lopez, M. Mowers, and E. Lantz, "Interactions of wind energy project siting, wind resource potential, and the evolution of the U.S. power system," *Energy*, vol. 223, May 2021, doi: 10.1016/j.energy.2021.119998.

[52] "Nuclear-Renewable Synergies for Clean Energy Solutions | NREL." Accessed: May 21, 2025. [Online]. Available: https://www.nrel.gov/news/detail/program/2020/nuclear-renewable-synergies-for-clean-energy-solutions

[53] "Frequently Asked Questions (FAQs) - U.S. Energy Information Administration (EIA)." Accessed: May 21, 2025. [Online]. Available: https://www.eia.gov/tools/faqs/faq.php?id=207&t=21

[54] "Nuclear reactor Ownership." Accessed: May 21, 2025. [Online]. Available: https://www.eia.gov/nuclear/reactors/reactorcapacity.php

[55] "The Staggering Cost of New Nuclear Power - Center for American Progress." Accessed: May 21, 2025. [Online]. Available: https://www.americanprogress.org/article/the-staggering-cost-of-new-nuclear-power/

[56] "Hydropower Basics | Department of Energy." Accessed: May 21, 2025. [Online]. Available: https://www.energy.gov/eere/water/hydropower-basics

[57] "Facts About Hydropower | Wisconsin Valley Improvement Company." Accessed: May 21, 2025. [Online]. Available: https://www.wvic.com/Content/Facts_About_Hydropower.cfm

[58] "Facts about Hydropower." Accessed: May 21, 2025. [Online]. Available: https://www.hydropower.org/iha/discover-facts-about-hydropower

[59] "Geothermal explained - U.S. Energy Information Administration (EIA)." Accessed: May 21, 2025. [Online]. Available: https://www.eia.gov/energyexplained/geothermal/

[60] "Use of geothermal energy - U.S. Energy Information Administration (EIA)." Accessed: May 21, 2025. [Online]. Available: https://www.eia.gov/energyexplained/geothermal/use-of-geothermal-energy.php

[61] "U.S. geothermal capacity factor 2023| Statista." Accessed: May 21, 2025. [Online]. Available: https://www.statista.com/statistics/1559055/average-capacity-factors-geothermal-electricity-united-states/

Click here to navigate back to Table of Contents

SCTE
a subsidiary of CableLabs®

140 Philips Road, Exton, PA 19341-1318 | 800.542.5040 | **scte.org**