Optimizing AI Traffic Over WAN: How NetOpt Meets Next-Generation Demands

Introduction

Artificial intelligence (AI) workloads – especially **training** (model learning) and **inference** (model querying) – are introducing new types of traffic on wide-area networks. Unlike traditional internet services (e.g. web browsing or streaming video), AI training and inference generate **highly data-intensive and performance-sensitive** traffic patterns. For example, streaming video today dominates global consumer traffic (Netflix and YouTube alone account for over 26% of volume), but AI traffic behaves very differently. Training an AI model involves distributed computations across many GPUs in different sites, which must exchange large volumes of parameters and gradients in real time. Inference involves serving AI model outputs (e.g. answering a user's query or analyzing a sensor feed), often under tight latency requirements and sometimes from **edge locations** rather than central data centers. These emerging AI-driven flows are growing rapidly – industry forecasts predict AI network traffic could increase at **35%+ CAGR** in coming years lightreading.com – and they challenge networks that were originally optimized for more predictable web and video usage.

Critically, **AI training traffic** and **AI inference traffic** have distinct characteristics that set them apart from traditional traffic. Training typically consists of long-running, high-bandwidth data exchanges between clusters of machines (often in different data centers) that are synchronized at microsecond scale. Inference traffic, on the other hand, involves many distributed requests/responses (sometimes from edge "inference hubs") and can spike unpredictably with user demand. These differences mean that both **technical requirements** (throughput, latency, reliability) and **network design strategies** for AI workloads diverge from past norms. Below, we examine the characteristics and demands of training vs. inference traffic, how telecom carriers are adapting their networks for AI, and how new tools like **NetOpt** can help plan wide-area networks to accommodate the AI era.

AI Traffic Characteristics and Technical Demands

Training Traffic: Synchronized, High-Volume, Low-Latency Requirements

AI *training* traffic occurs when large machine learning models (like deep neural networks) are being trained on distributed compute clusters. This traffic is **latency-sensitive, extremely high-volume, and tightly synchronized** across nodes:

- **Ultra-Low Latency & High Bandwidth:** To keep many GPUs in lockstep, training networks demand *very low latency* (often microseconds to a few milliseconds) and massive throughput. A recent analysis notes that *"ultra-high bandwidth, ultra-low latency, and ultra-high reliability are the requirements for network connectivity in large-scale AI training."*naddod.com In distributed training, nodes frequently exchange weight updates and gradients; any network delay directly slows the training iteration. Thus, the interconnect must provide maximal speed and minimal jitter.
- **Synchronized Burst Patterns:** Training traffic often follows a pattern of brief *bursts* of extremely high data exchange, synchronized across all nodes at each training step. For example, after computing on local data, each GPU might perform an all-reduce operation with others, flooding the network with parameter data simultaneously. This bursty, synchronized behavior is unlike the smoother, asynchronous flows of streaming video. It requires networks engineered for *peak load* rather than just average throughput, and with sufficient buffering or capacity to handle concurrent surges without loss.
- **Specialized Protocols (Credit-Based Flow Control):** Because of the need for consistent low latency and zero packet loss, AI training often runs on specialized, lossless transport protocols that are *incompatible with standard TCP/IP congestion control*. High-performance computing fabrics like InfiniBand or RoCE (RDMA over Converged Ethernet) use **credit-based flow control** to avoid drops – the sender only transmits when the receiver signals buffer availability naddod.medium.com. This ensures no packets are lost (unlike TCP which intentionally drops packets to signal congestion), preserving training throughput and stability. However, it means training clusters cannot tolerate random packet loss or throughput throttling in the way typical internet traffic might. To achieve this on wide-area links, operators increasingly use dedicated infrastructure.
- **Dedicated Fiber and Wavelength Circuits:** Given the unique demands, AI training traffic is often carried over **dedicated optical circuits** rather than mixed with general-purpose traffic. Credit-based RDMA protocols expect a virtually lossless environment, which is hard to guarantee on congested public IP networks. Instead, hyperscalers are leasing dark fiber or custom wavelength services between their data centers for training backbones. These private circuits provide guaranteed bandwidth and direct control over optical latency. In effect, an AI training cluster spanning multiple sites might have its **own fiber wavelengths** interconnecting the GPUs, bypassing the public internet entirely. Major carriers report exploding demand for such high-capacity circuits: industry data shows AI applications are now a "major purchase driver" for 100+ Gbps wavelength links (with hyperscalers often buying dedicated 400 Gbps waves) fierce-network.com. The traffic on these training waves is orders of magnitude heavier and more synchronized than typical enterprise flows.

In summary, training traffic pushes networks to their limits on **throughput and consistency**. It tolerates virtually no packet loss or variability. These requirements explain why training networks are being built as *parallel infrastructures* – dedicated fibers, advanced optical technologies, and custom transport protocols – separate from the traditional IP internet. The next section will discuss how telecom providers are addressing this need.

Inference Traffic: Distributed, Burst-Prone, and Variable SLAs

AI *inference* traffic occurs when deployed AI models are used to generate outputs (predictions, classifications, etc.) in real time. This type of traffic has a different profile, characterized by **widely distributed sources, bursty demand, and shifting latency/availability needs**:

- **Distributed Inference Hubs:** Unlike training (which might happen in a few big data centers), inference is often served from many distributed *edge or regional data centers*. To reduce user-perceived latency, companies deploy inference engines closer to end-users – for example, in regional cloud zones or telecom edge sites. These **inference hubs** inject traffic into the network in a more distributed fashion. Requests from users (or IoT devices) are routed to a nearby inference server, which processes the query using the AI model and sends back a result. The network must connect all these hubs and users with low delay. In practice, telecom carriers have begun hosting inference infrastructure at the network edge. A notable example is Verizon's 5G Edge partnership with AWS Wavelength, which *"embeds AWS compute and storage services at service access points on the edge of the 5G network,"* allowing inference traffic to reach local servers **without traversing the broader internet** rcrwireless.com. This dramatically cuts latency for applications like AR/VR, industrial automation, and real-time analytics. Such setups illustrate the emerging model of inference traffic: many localized pockets of AI processing, each requiring high-speed connectivity to users and back to core data centers for occasional model updates.
- **Burstiness and Unpredictable Load:** Inference demand can be highly **spiky**. While video streaming has diurnal patterns that smooth out over millions of users, AI inference might see sudden surges when a new application or feature goes viral, or when an unusual event triggers many AI queries at once. For instance, a popular AI chatbot or image generator can experience a flood of requests within minutes of a trending post. This burstiness means network capacity between inference servers and users (and between inference sites and core clusters) must scale rapidly. It's difficult to cache or buffer AI responses the way video streams are buffered, because each query is unique and processed on the fly. Therefore, the network must handle **instantaneous spikes** in traffic, often with elastic routing or on-demand bandwidth. The traffic profile is closer to transaction bursts or CDN cache misses than to steady streaming. Operators must plan for peak concurrent inference loads that may be an order of magnitude higher than averages.
- **Large Model Updates and Data Backhaul:** Another aspect of inference traffic is the need to distribute *AI model updates* and gather data backhaul. When an AI model is retrained or improved (which might happen weekly or even daily for rapid iteration), the updated model – potentially hundreds of gigabytes in size for large neural networks – must be pushed out to all the inference servers globally. This creates a **massive transfer** (essentially a multi-gigabyte content distribution) across the network from central training sites to edge inference nodes. Similarly, logs or user data from edge sites might be sent back to core for further training, adding to backhaul traffic. These episodic large transfers resemble big software distribution or backup traffic, but with greater frequency

as AI models evolve quickly. The network must accommodate these very large files in addition to the real-time query traffic. Dedicated circuits or scheduled transfer windows may be used to handle model update propagation.

- **Dynamic Latency and Availability Requirements:** The sensitivity of inference traffic to latency and uptime varies by application – and can even shift over time or based on user context. For example, an AI-powered medical diagnostic service might require **ultra-reliable, low-latency** connectivity (e.g. <10 ms and redundant paths) to ensure instant, always-on analyses. In contrast, an AI batch analytics job may tolerate higher latency or occasional delay. Even for the same service, there may be periods where low latency is critical (e.g. during live usage) and other times when it's less urgent. This means that **quality-of-service demands for inference traffic are variable and context-dependent**. Networks carrying inference must therefore be flexible – able to provide high-priority, low-latency paths for certain flows on demand, and to ensure high availability (e.g. redundant routes) when required. Compared to streaming video – which generally tolerates a few seconds of buffering delay and uses best-effort internet paths – many AI inferences cannot be buffered and may represent mission-critical transactions. Thus, service providers are considering new traffic engineering approaches (like segment routing with strict latency SLAs, or edge computing placements) to meet these dynamic needs.

In short, inference traffic is **more distributed and erratic** than traditional content distribution. It calls for pushing compute closer to users (to meet latency goals) and for networks that can rapidly scale and reconfigure to handle surges. The combination of *burstiness* and *strict performance targets* (varying by use-case) makes inference one of the most challenging new traffic types to accommodate on a wide-area network.

How Telecom Carriers Are Responding

Recognizing these challenges, telecom carriers and network operators are rapidly evolving their backbone and metro networks to support AI training and inference traffic. In the U.S., major carriers such as **Lumen**, **Verizon**, **AT&T**, and **Zayo** have announced initiatives ranging from new fiber builds to specialized services for AI connectivity. Below we highlight examples of how each is adapting:

Lumen (Level 3): Dedicated Fiber Builds for AI and Edge Inference

Lumen Technologies has positioned itself as a key provider of **private fiber infrastructure** for AI workloads. In 2024, Lumen revealed it had signed over **$8 billion in private custom fiber (PCF) agreements** with hyperscalers Microsoft, Amazon, Google, and Meta specifically to carry their AI training data lightreading.com. These deals essentially task Lumen with building out dedicated fiber routes and wavelength circuits linking the cloud giants' data centers, creating high-capacity "AI superhighways" between training sites. Lumen's CEO Kate Johnson noted, *"Big Tech is choosing Lumen because our conduit-based fiber network was built for this moment… No other*

telco that owns a fiber network is doing this," underscoring Lumen's unique advantage of extensive long-haul fiber inherited from its Level 3 network lightreading.com. Analysts concur that Lumen's U.S. fiber footprint (with many spare conduits from past overbuilds) makes it *"uniquely positioned to transport AI traffic across the country."*lightreading.com

To meet the surge in demand, Lumen is dramatically expanding its fiber capacity. Notably, it struck an agreement to reserve **10% of Corning's entire fiber-optic cable production** for the next two years – Lumen's largest cable purchase ever – to ensure it can *more than double* its intercity fiber route miles in support of generative AI growth datacenterfrontier.com datacenterfrontier.com. This includes deploying Corning's latest "Gen AI" ultra-dense fiber cables, which allow 2–4× more fiber strands to be packed into existing conduits datacenterfrontier.com. The rationale is clear: *"Generative AI…creates significant demand for passive optical connectivity. All data centers consist of a front-end network connecting CPUs. To meet AI's demands, customers are building a new fiber-rich second network to connect GPUs,"* explained Corning's CEO datacenterfrontier.com. In other words, beyond the traditional data paths, AI training needs parallel fiber networks – and Lumen is literally laying those foundations.

In addition to raw connectivity, Lumen is also investing in **edge infrastructure for AI inference**. Its nationwide edge cloud nodes (reaching 95% of U.S. businesses within <5 ms latency) are being leveraged to host inference services in proximity to users crn.com. Lumen's recent partnership with IBM Watsonx exemplifies this strategy: by pairing IBM's AI platform with Lumen's edge computing and fiber network, they aim to bring *"real-time AI inferencing closer to where data is generated"* for enterprises crn.com. This reduces latency and offloads backbone traffic, since local data can be processed on local AI nodes instead of traveling back to a core cloud. Lumen has created a dedicated operations team for AI networking services lightreading.com and is marketing itself as *"the backbone for the AI economy."* It is essentially offering end-to-end solutions – from **wavelength circuits for training** to **edge hosting for inference** – to both hyperscalers and enterprise customers requiring AI connectivity.

Verizon: Upgrading Optical Backbone and Launching "AI Connect" Services

Verizon Communications is likewise adapting its network for the AI era, focusing on **upgrading capacity** and integrating network + compute services under its new **AI Connect** portfolio. On the infrastructure side, Verizon has been an early mover in deploying next-generation optical transport to handle AI's massive bandwidth needs. In late 2024, Verizon announced a successful field trial of a single-wave **1.6 Tbps optical channel** over its live fiber network – doubling the per-wavelength throughput of its previous 800 Gbps trials rcrwireless.com. This trial, using Ciena's cutting-edge WaveLogic 6e coherent optics, ran over a 118 km metro route with multiple ROADMs and confirmed the ability to carry 1.6 Tbps traffic on a single lambda in a real network rcrwireless.com. Verizon explicitly framed this achievement in context of AI, stating it is *"prepping its network for the higher speed and capacity demand expected to be generated by artificial intelligence workloads."*rcrwireless.com Such ultra-high-bandwidth channels will likely be deployed between Verizon's core data centers and interconnection points to carry aggregated AI flows (for both its cloud partners and large enterprise customers).

Verizon is also expanding fiber capacity in the metro and long-haul domains to assure low latency paths for AI. Adam Koeppe, Verizon's SVP for Technology Strategy, noted that Verizon's multi-year network transformation – including a **cloud-native core, massive fiber capacity upgrades, and intelligent edge deployments** – enables it to *"power the processes and movements of AI-generated activity"* across its infrastructure rcrwireless.com. In essence, Verizon is beefing up its backbone so it can be the "provider of choice" to transport AI data for hyperscalers and businesses. This includes leveraging its extensive metro fiber (e.g. the One Fiber initiative covering 70+ metro areas) and even its last-mile fiber (FiOS footprint) where applicable to link edge compute locations lightreading.com.

On the services front, Verizon in 2024 rolled out **AI Connect**, a suite of offerings combining network connectivity with data center and edge resources to support AI at scale lightreading.com lightreading.com. Verizon AI Connect essentially bundles high-performance connectivity (fiber, 5G, private networks) with colocation space, power, and cloud partnerships to create end-to-end solutions for AI workload hosting. *"Verizon AI Connect is a strategy and suite of products to manage AI resource-intensive workloads at scale… with our network's low latency, high bandwidth and robust intelligent edge capabilities,"* Verizon states verizon.com. In practical terms, this means Verizon will provide everything from **lit and dark fiber links** (for connecting AI training clusters) to **edge compute nodes** with GPUs (for running inference close to users). Verizon Business has even integrated with partners like NVIDIA and GPU-as-a-service provider Vultr to deploy GPU hardware in Verizon's facilities, extending AI computing into the network lightreading.com lightreading.com. The AI Connect portfolio underscores that *connectivity alone isn't enough* – Verizon is packaging network, compute, and storage to help customers run AI workloads "everywhere" (core, cloud, edge) with seamless connectivity in between.

Notably, Verizon cites strong demand driving this move. The company's leadership pointed to estimates of **$1 trillion** in AI infrastructure investment over the next decade and highlighted that all this will *"need to be underpinned by secure network connectivity"* bridging distributed AI compute lightreading.com lightreading.com. Verizon already sees over $1 billion in its sales pipeline tied to AI projects that leverage its existing network assets lightreading.com. In summary, Verizon's response to AI traffic is twofold: **scale up the network** (optical innovations, fiber expansions) to haul unprecedented data loads, and **offer integrated solutions** (AI Connect's mix of fiber + edge compute) so that enterprises and hyperscalers can deploy AI services with the carrier's network as the fabric interconnecting everything.

AT&T: 1.6 Tbps Trials and Core Network Preparedness

AT&T is similarly bracing its core network for the onslaught of AI data. In early 2025 AT&T announced it had conducted a **1.6 Tbps single-wavelength trial** using Ciena coherent optics, demonstrating the ability to carry trans-Pacific-scale traffic on one channel across its fiber backbone fierce-network.com fierce-network.com. The successful trial, spanning 296 km, was framed by AT&T as a step towards readying its network for *"AI and other bandwidth-intensive applications in the future."*fierce-network.com Like Verizon, AT&T recognizes that services like

generative AI and advanced cloud applications are fueling a surge in demand for high-capacity transport. Vertical Systems Group noted that the **wavelength services market is "running hot" due to AI** – many customers are now ordering 100 Gbps+ links specifically to support AI workloads fierce-network.com. AT&T is one of the carriers meeting this demand by upgrading its optical layer to 800G and 1.6T speeds. Jimmy Yu, a Dell'Oro Group analyst, highlighted the importance of such trials, noting it gets *"harder and harder to increase span length as wavelength speed goes up,"* so achieving 1.6T over nearly 300 km was significant fierce-network.com. This implies AT&T's network will be able to deploy 1.6T waves on long-haul routes – critical for connecting far-flung AI data centers with minimal intermediate equipment.

AT&T has indicated these high-speed optics will help accommodate **AI, cloud computing, streaming and more** on the same backbone fierce-network.com. However, AI stands out as a major driver: *"As expected, AI applications are a major purchase driver for 100+ Gbps wavelength orders,"* reports Vertical Systems Group fierce-network.com. Hyperscalers, financial firms, and media companies are all buying new high-bandwidth circuits for AI connectivity, and AT&T is one provider for such services. In addition to optical capacity, AT&T is leveraging its extensive fiber deployments (including ongoing expansions of metro and regional fiber) to ensure it can meet latency requirements. For example, AT&T's fiber routes, combined with intelligent routing via its IP/MPLS network, are being optimized to provide alternate low-latency paths and diversity for critical AI links (much as they have done for financial trading networks in the past). AT&T is also involved in edge computing initiatives – it has previously partnered with cloud providers (like Microsoft Azure) to host edge compute nodes on AT&T premises, which could be used for AI inference scenarios requiring ultra-low latency at the network edge business.att.com.

While AT&T's public announcements have been a bit less AI-specific than Lumen's or Verizon's, the company's messaging makes clear it sees **high-capacity transport and low-latency networking as key to future services**. AT&T's CTO Andre Fuetsch (in previous discussions) emphasized that technologies like **routing automation, fiber densification, and edge computing** all interplay to support next-gen applications like AI. Indeed, AT&T is integrating technologies such as SRv6 (segment routing) and adaptive core networks, which could allow it to steer traffic on optimal paths meeting certain SLAs – useful when an AI application requires, say, sub-50 ms latency between two data centers. In summary, AT&T's response focuses on **future-proofing its backbone** with leading-edge optical rates and software-defined networking, to ensure that as AI traffic ramps up, the network can handle it without bottlenecks.

Zayo: Dark Fiber and Wavelength Scalability for Hyperscalers

Zayo Group, a major fiber operator heavily used by cloud and content providers, has been directly responding to **hyperscaler demand for AI connectivity**. Zayo is not a traditional telecom with consumer services; its business is providing fiber and bandwidth infrastructure to other companies. In 2024, Zayo revealed plans to build **5,000 miles of new long-haul fiber routes** over the next five years specifically to *"support the artificial intelligence (AI) boom"* and growing data center clusters fierce-network.com. These new routes are targeted between key **AI data center hubs** – locations where hyperscalers are concentrating GPU-heavy infrastructure. The proactive

build reflects Zayo's view (shared with others) that AI is driving a *step-change* in backbone traffic needs. Zayo's Chief Product Officer noted that previously, a typical long-haul fiber sale was 8–12 fiber strands, but starting in 2023, customers began requesting **144, 288, even 432 fiber pairs in a single order** fierce-network.com. This is a massive increase in capacity purchasing, indicating that cloud providers are essentially reserving entire fiber cables for their AI networks. Likewise, **Windstream** (another wholesale fiber provider) observed some customers asking for up to 864 fibers at once to connect GPU farms fierce-network.com. These figures underscore how AI traffic is prompting *huge scale-ups in raw fiber capacity*. The traditional approach of a few 10 Gbps or 100 Gbps leased lines is giving way to hyperscalers buying huge bundles of dark fiber to light with whatever bandwidth their equipment can push, ensuring dedicated, isolated capacity for AI.

In addition to new fiber builds, Zayo has been upgrading its lit network services to cater to AI. It boasts one of the largest 400G-enabled networks in North America, and is already looking ahead to offering 800G waves as optical vendor technology allows zayo.comfierce-network.com. Zayo markets its **wavelength services** as ideal for "moving massive datasets, supporting AI training, and powering data-intensive applications"zayo.com. In effect, Zayo is selling **guaranteed high-bandwidth pipes** that AI developers can rely on between their sites. For example, a research lab training large models might lease multiple 400 Gbps wavelengths from Zayo to a cloud GPU cluster, rather than sending that traffic over commodity internet routes. Zayo is also working on automation ("Waves on Demand") to let customers turn up new circuits quickly as their AI needs spike zayo.com, reflecting the more dynamic provisioning that AI projects may require.

Another adaptation by Zayo is ensuring **route diversity and low latency** on key corridors. They are expanding fiber between major cloud regions (e.g. between Silicon Valley, Los Angeles, Phoenix – all growing AI hubs – and between Eastern hubs like Ashburn, Atlanta, Dallas) to provide multiple routes. This helps AI customers achieve the high availability they need (e.g. dual diverse paths in case one fiber route is cut) and to meet latency targets by selecting the shortest path. Zayo even notes the correlation of fiber with power: anticipating that *"more data centers will crop up where they have enough power"* for AI, and thus building fiber to connect those places fierce-network.com. In summary, Zayo's response is to **massively scale capacity** (both dark fiber and lit wavelengths) and to do so in a way that hyperscalers can flexibly consume, since AI demands can ramp up quickly. The company saw the AI capacity spike coming and moved early to cater to it, which is paying off as cloud players race to interconnect their AI infrastructure.

How NetOpt Solves AI Traffic Planning Challenges

The shift to AI-centric traffic puts enormous strain on network planning – especially for carriers and hyperscalers designing their wide-area networks to meet these new demands. This is where **NetOpt (NetOpt)**, a multi-layer network planning tool, offers a powerful solution. NetOpt is purpose-built to handle the complexity of modern network design by optimizing across IP, MPLS, and optical layers simultaneously, with the intelligence to incorporate specific application requirements (like those of training vs. inference traffic). Leveraging the capabilities described in the NetOpt whitepaper, we highlight three ways NetOpt addresses AI traffic planning challenges:

- Joint IP/Optical Optimization for Training and Inference: NetOpt breaks down the traditional silos between IP/MPLS planning and optical transport planning. In legacy processes, engineers plan the IP layer (routing, traffic engineering) and the optical layer (fiber routes, wavelengths) separately, often leading to suboptimal results. NetOpt instead *"simultaneously considers the IP/MPLS layer, the optical/WDM layer, and even physical fiber constraints, under a single optimization framework."* This means it can design an end-to-end solution that accounts for inference traffic at the IP layer and training traffic at the optical layer together. For example, inference traffic might need diverse IP paths between edge sites and core clouds, while training traffic might need dedicated optical circuits between data centers – NetOpt can trade off decisions across these layers. It doesn't assume a fixed optical topology; it can decide to route an IP flow over a new fiber route or add a wavelength if needed. All layers are optimized with one objective function, considering both IP router costs and optical transport costs in tandem. This multi-layer optimization enables truly holistic planning – NetOpt can simultaneously optimize, say, an MPLS overlay for inference distribution *and* the fiber paths for bulk training transfers, finding a global optimum that meets both needs. The result is a network design that efficiently supports *both* types of AI traffic without the usual over-provisioning (which siloed planning would resort to). In short, NetOpt allows carriers to plan IP and optical jointly – crucial for AI scenarios where inference (packet traffic) and training (circuit traffic) intersect.
- Latency- and Availability-Aware Path Design: AI applications often come with strict latency budgets and high availability requirements, as discussed earlier. NetOpt directly incorporates such Service Level Agreements (SLAs) into the planning model. Planners using NetOpt can specify demand-specific latency and resiliency requirements, and the tool will build network paths to honor those. As the whitepaper describes, *"Planners can specify, for instance, that between site A and B, at least one path < 50 ms latency is required (for low-latency apps), or that connectivity must survive any single fiber cut with 99.99% availability. NetOpt will then incorporate these in the design – perhaps choosing shorter fiber routes for latency-critical sites, or providing diverse routing to meet availability."* Because NetOpt works across layers, it can enforce truly disjoint routing (e.g. ensuring redundant IP paths do not share the same physical fiber) to achieve a required availability. For latency-critical demands, it might choose a more direct fiber route or even suggest placing an intermediate regen site to shorten a path. This is especially relevant for AI inference traffic, where certain hub-to-hub links might need guaranteed low delay, and for training traffic that might need protected circuits to avoid job interruption. Traditional tools would struggle to account for these needs across layers – e.g. an IP tool might not "see" fiber distance, and an optical tool wouldn't normally consider alternate IP reroutes. NetOpt unifies this, allowing it to meet strict SLAs without overbuilding. For example, rather than provisioning two completely separate networks for reliability, NetOpt might find a clever way to share capacity while still mathematically guaranteeing the 99.99% uptime by using diverse fibers and fast reroute in the IP layerfile-8c66pywhnrjc25csmzauqa. This capability to plan with explicit latency and availability targets is a game-changer for AI traffic engineering, where different flows have very different SLA profiles.

- **Statistical Traffic Modeling for Bursty Demand:** Planning for peak AI traffic can be challenging when demand is highly variable. NetOpt helps solve this by allowing **traffic demands to be modeled as distributions or multiple scenarios**, rather than a single fixed number. In the NetOpt approach, traffic can be input as a time series (e.g. busy hour vs. off-peak, or various load levels) and even as probabilistic distributions for bursts. One of NetOpt's innovations is *modeling traffic not as a static matrix, but over multiple time periods and scenarios*, then optimizing the network across thosefile-8c66pywhnrjc25csmzauqa. Practically, this means planners can incorporate **bursty traffic patterns** – for instance, specifying an inference demand as an average load plus a potential burst to X Gbps with Y probability. NetOpt can then ensure the design meets a certain **confidence level** (say 90% of the time the bursts can be served without congestion). By optimizing to a statistical percentile of load, NetOpt avoids the extremes of either overbuilding for a worst-case that rarely occurs or under-provisioning for peaks. Instead, it can right-size capacity such that, for example, there is a 90% probability that inference traffic spikes will be handled, with known risk of clipping only in the most extreme 10% cases (which might be acceptable or handled via traffic management). The tool's support for multi-period optimization and even **time-zone dependent demands** aligns well with AI traffic patterns – e.g. it can exploit the fact that not all regions peak in AI usage at the same time, sharing capacity across time zones. By modeling randomness (using techniques like exponential distributions for inter-arrival times of bursts, etc.), NetOpt enables planning for **bursty AI workloads in a rigorous, quantitative way**. Planners can choose a percentile (confidence level) for optimization, ensuring the network is statistically robust to surges. This is far more efficient than the old approach of adding huge safety margins at each layer "just in case." NetOpt's statistical traffic engineering results in networks that meet AI demand spikes with high probability, yet still minimize excess idle capacity.

In summary, NetOpt provides a **unified, intelligent planning framework** that directly tackles the dual challenges introduced by AI traffic. It can simultaneously design optical circuits for huge training flows and IP paths for distributed inference, all while respecting the nuanced SLAs and bursty nature of these demands. The outcome is an optimized network blueprint that meets AI requirements at minimal cost – often discovering that far less over-provisioning is needed than one would assume with siloed planning. By using tools like NetOpt, telecom operators and hyperscalers can efficiently build networks that are *AI-ready*: capable of low-latency, high-volume data distribution with guaranteed performance where needed.

Conclusion

AI is reshaping network traffic on a fundamental level – shifting the dominant paradigm from relatively predictable, cache-friendly streams (like video) to dynamic, high-volume data flows that strain every layer of the network. The rise of distributed training jobs and latency-critical inference services has created new **demands for bandwidth, latency, and reliability** that far exceed business-as-usual. We are witnessing a major industry response: telecom carriers are

reinventing their wide-area networks with more fiber, faster optics, and edge computing deployments to accommodate AI. From Lumen's continent-spanning fiber builds for hyperscaler training clusters, to Verizon's integration of fiber, 5G, and edge GPUs under AI Connect, to AT&T and others pushing the envelope on optical speeds, it's clear that supporting AI traffic has become a top strategic priority.

The network designs of the coming decade will be heavily influenced by these AI-driven requirements. Traditional planning approaches – with siloed IP and optical decisions and static worst-case provisioning – are ill-suited to the task. Instead, carriers and cloud operators will need holistic, agile planning tools to efficiently harness their infrastructure for AI. NetOpt exemplifies this new generation of planning solutions. By optimizing across all network layers and explicitly accounting for latency SLAs and statistical traffic variations, NetOpt enables engineers to create future-proof network plans that can carry both today's internet traffic and tomorrow's immense AI workloads. It helps avoid the 10× over-provisioning trap by smartly sharing capacity and aligning network build-outs with actual demand growthfile-8c66pywhnrjc25csmzauqafile-8c66pywhnrjc25csmzauqa. In a world where AI development is moving at breakneck speed, NetOpt provides a way to stay ahead – ensuring that carrier and hyperscaler WANs can meet AI's needs without breaking the bank.

In conclusion, the shift from web/video-dominated traffic to AI training and inference traffic represents a generational change for wide-area networks. Those networks must become ultra-fast, ultra-scalable, and application-aware. Telecom operators that embrace modern planning tools and invest in the right infrastructure (dense fiber, advanced optics, edge compute) will be well positioned to serve as the digital highways for the AI era. By leveraging solutions like NetOpt to intelligently plan multi-layer upgrades, operators can future-proof their WANs for AI, delivering the low-latency, high-capacity connectivity that the next wave of innovation will demand – all while maintaining efficiency and resiliency in their networks.

Sources: References supporting the insights and data points in this whitepaper include carrier press releases and briefings (e.g. Lumen, Verizon, AT&T, Zayo announcements), technical analyses from industry media (Light Reading, Fierce Telecom, RCR Wireless) on AI networking, and the NetOpt.Design whitepaper for multi-layer planning capabilities. These are cited throughout the document for further reading.

 Contact: info@NetOpt.design