Optimizing Data Center Networks with NetOpt.Design

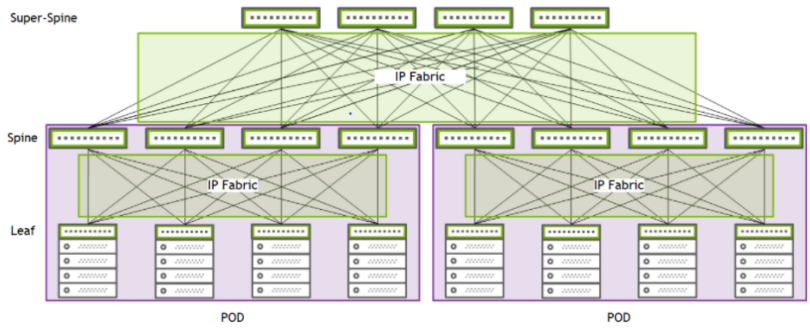
Introduction

Modern data center networks are typically built in a **leaf–spine (Clos)** topology, which provides a scalable, high-bandwidth fabric for server racks. However, the explosive growth of data center traffic – especially driven by AI workloads – is straining traditional designs. In fact, cloud data center traffic has been doubling roughly every year, far outpacing the historical pace of switch capacity improvements cacm.acm.org. This means simply scaling up with bigger switches is no longer sufficient. Network architects must explore new design optimizations to meet demand without exorbitant cost.

NetOpt.Design (NetOpt) is a multi-layer network planning tool purpose-built to tackle these challenges. It can intelligently optimize network topology across IP and optical layers simultaneously, while accounting for specific application requirements netopt.design. In other words, NetOpt provides a unified optimization framework that considers routers, switches, fiber links, and even WDM optical paths in one model netopt.design. By breaking the traditional siloed planning approach, it enables **holistic design** of data center networks that meet performance and reliability targets at minimal cost. The following sections discuss how NetOpt can bring additional value to leaf–spine data center fabrics – including scaling to extra tiers, leveraging optical connectivity, and accommodating stringent AI workload needs – with references to industry cases and best practices.

Scaling Leaf–Spine Fabrics with Additional Layers

Leaf–spine architectures typically have two tiers: **leaf switches** at the rack level and **spine switches** that interconnect all leaves in a full mesh (see figure below). This two-tier Clos fabric ensures any two servers communicate with at most one intermediate spine hop docs.nvidia.com docs.nvidia.com. As data centers grow, however, a single spine layer can hit scaling limits (e.g. port counts or hop count constraints). To support more racks and servers, operators often introduce a **third tier** – sometimes called a **super-spine** or core layer – that sits above the spines docs.nvidia.com. Each spine layer can then be segmented into pods, with the super-spine providing connectivity between pods, forming a three-tier Clos topology docs.nvidia.com.



Example of a three-tier Clos network (leaf–spine–super-spine). Leaf/spine pods are aggregated via an additional super-spine layer to scale out the data center fabric docs.nvidia.com.

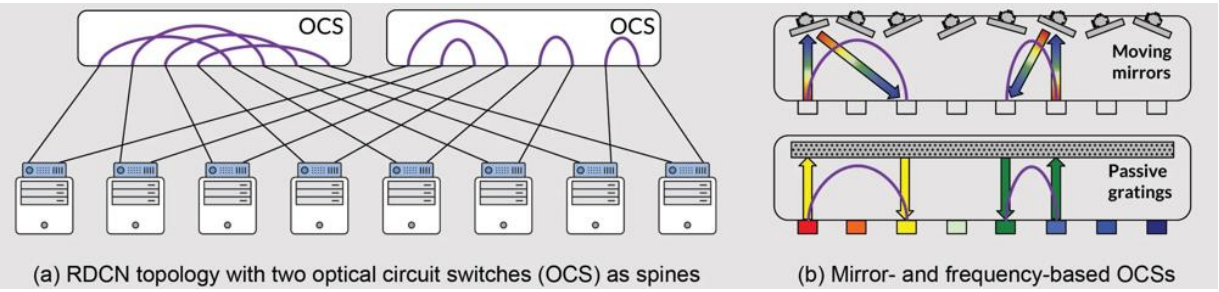
Industry experience shows this evolution is necessary for massive-scale data centers. As early as 2019, hyperscalers deploying tens of thousands of servers found that two-tier leaf–spine designs had to "scale out" into three-tier Clos fabrics to meet growing capacity requirements cisco.com. In such designs, **server pods** are formed – each pod is a two-tier leaf–spine fabric on its own, and multiple pods are then interconnected through the third-tier core cisco.com cisco.com. The super-spine layer increases the network's bisection bandwidth and node count by aggregating multiple spine pods docs.nvidia.com. Key parameters like the number of spine switches per pod, number of leaf uplinks, and links to the super-spine must be determined carefully, since they dictate the **oversubscription ratio** and overall capacity of the network. For example, if each leaf has 8 uplinks, it can connect to up to 8 spine switches; if each spine has 64 downlink ports, it can connect up to 64 leaf switches (minus any ports reserved for super-spine connections) cisco.com cisco.com. These factors determine whether the fabric is non-blocking or oversubscribed, and by how much. High-performance environments (e.g. AI supercomputing clusters) often require a non-oversubscribed (1:1) fabric for certain portions of the network, whereas enterprise data centers might allow some oversubscription to save cost.

NetOpt brings value here by **automating the optimal Clos design** given the data center's requirements. Rather than manually guessing the number of spines or links needed, architects can let NetOpt evaluate many topology variants to find the best fit. The tool can determine how many spine and super-spine switches are required and how to interconnect them to meet a target oversubscription or latency budget, while minimizing capital expense. NetOpt's optimizer considers the trade-offs between adding more switches/links vs. the performance gained. It ensures the fabric has enough path diversity and bandwidth to handle peak loads without bottlenecks, but also avoids over-provisioning idle capacity. For instance, NetOpt can find the minimal number of super-spines needed to support a given number of pods and traffic matrix, and even decide the optimal connectivity (which pod spines should link to which super-spines) to balance utilization. By **right-sizing** the Clos fabric, NetOpt prevents both under-design (which would cause congestion) and over-design (which wastes cost and power). This is especially crucial as networks scale to very large sizes. One recent AI cluster design ("AIPod") illustrates the scale at which careful Clos planning is needed: it used a 3-tier non-blocking Clos with ~400 switches to support over 16,000 GPUs, with each server having 8×400 GbE links into the fabric fibermall.com fibermall.com. In that design, an extra Super-Spine layer connects multiple spine

"channels," enabling full bandwidth across the entire cluster with no oversubscription fibermall.com. NetOpt can facilitate such designs by evaluating different multi-tier topologies and ensuring they meet the enormous bandwidth demands of distributed AI workloads, all while keeping the switch count as low as possible.

Incorporating Optical Connectivity in the Spine/Core

Another optimization avenue is to leverage **optical technologies** in the spine or core layers of the data center network. Traditionally, leaf and spine switches are electrical packet switches. But emerging approaches use optical circuit switching or optical cross-connects as a high-speed backbone to complement or even replace part of the electrical network. The motivation is that optical circuit switches (OCS) can provide *reconfigurable, direct connections* between top-of-rack switches with very high bandwidth and low per-bit cost, essentially acting as an optical spine layer cacm.acm.org cacm.acm.org. Instead of always routing traffic through multiple fixed spine hops, an optical layer can dynamically establish a **shortcut** between two racks (or two spine pods) that have heavy traffic between them cacm.acm.org cacm.acm.org. By doing so, it reduces the "bandwidth tax" of multi-hop paths in a static topology – for example, a static fat-tree might have a network diameter of 4 hops, whereas an optical circuit fabric can sometimes cut the path down to 1 hop for those specific communications cacm.acm.org. This yields lower latency and frees up capacity in the rest of the network. Optical switches also operate transparently with respect to data rate and protocol, simply forwarding light; they require no packet buffering or header processing, and avoid optical-electrical-optical conversions in the core cacm.acm.org. As a result, an optical spine can significantly reduce power consumption and cost per Gbps compared to adding more electronic switches cacm.acm.org. It can also improve reliability by reducing the number of active components data flows must traverse – essentially giving a more direct, express route that bypasses intermediate failure points.



Concept of an optical circuit-switched spine layer. Here, top-of-rack switches connect to optical circuit switches (OCS) that dynamically create direct optical circuits between racks (purple arcs), rather than always sending traffic up through multiple electrical spine switchescacm.acm.org. In (b), examples of two OCS technologies are shown: one uses MEMS mirrors and another uses wavelength-based switching cacm.acm.org cacm.acm.org.

NetOpt.Design can help network planners evaluate the **value of a hybrid or all-optical core** in a quantitative way. It can incorporate optical switching nodes and fiber links into the optimization

model alongside traditional switches. Because NetOpt considers IP layer and WDM/fiber layer jointly, it can decide when investing in an optical circuit pays off. For example, if two pods frequently exchange large volumes of traffic, NetOpt might find it optimal to deploy an OCS or a dedicated fiber link between them rather than routing everything through a hierarchical spine. The tool "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" to meet performance goals netopt.design. All layers are optimized under a single objective function, trading off IP switch costs vs. optical transport costs and capacities. This holistic approach might reveal, for instance, that adding a few optical cross-connects at the core could allow using smaller (cheaper) spine switches or fewer spine layers, netting a cost savings while still meeting latency and throughput targets. In essence, NetOpt can quantify the benefit of optical bypass: it will deploy optical circuits in the design *if and where* they give a net benefit in cost, latency, or resilience.

It's worth noting that industry leaders are already exploring optical fabrics in their data centers. Google's latest-generation network (Jupiter) famously integrated an optical circuit switching layer that **eliminated the traditional spine tier** – instead of a rigid fat-tree, Google connected clusters of racks in a flexible optical mesh using MEMS-based OCS devices cacm.acm.org. This demand-aware, slowly reconfigurable topology is adapted on the scale of days based on traffic patterns cacm.acm.org. It allows Google to treat the network fabric almost like a programmable resource, matching the physical topology to prevailing communication patterns for improved efficiency cacm.acm.org cacm.acm.org. NetOpt enables such forward-thinking designs for a broader range of operators by providing the tools to **plan and evaluate reconfigurable topologies**. Planners can model scenarios with a purely optical super-spine or hybrid electrical/optical layers and let NetOpt determine the optimal configuration. Furthermore, because NetOpt natively handles multi-layer resiliency, it can ensure that introducing optical circuits does not compromise availability – for instance, it can design protection routes (or keep parallel electronic paths) to reroute traffic if an optical link is reconfiguring or if a fault occurs. In summary, by using NetOpt to assess optical spine options, data center designers can unlock potential improvements in **cost-per-bit, latency, and scalability** that static electrical networks might miss, all while rigorously meeting reliability requirements.

Meeting AI Workload Requirements for Latency and Availability

AI and machine learning workloads place **unique demands** on data center networks. In AI training clusters, two types of parallelism are common: **data parallelism** (a.k.a. batch parallelism) and **model parallelism**. In data-parallel training, multiple nodes independently train on different data batches (each node has a replica of the model); they periodically synchronize gradients, which requires high throughput but can tolerate slight delays between iterations. In model-parallel training, on the other hand, a single neural network's layers or parameters are split across multiple nodes, which means those nodes must communicate *in lockstep for every forward or backward pass*. Model parallelism (and its variant, pipeline parallelism) demands **extremely low latency and jitter** – if one node is slow to deliver results to the next, it stalls the entire training step, leaving expensive GPUs idle wevolver.com wevolver.com. Similarly, AI

inference clusters (serving real-time predictions) often require predictably low latency between inference stages. As a Cisco AI networking blueprint notes, when cluster communication suffers high latency or packet loss, "the learning job can take much longer to complete, or in some cases fail," so AI workloads have *stringent infrastructure requirements*, expecting *ultra-low latency* and *lossless* packet delivery cisco.com. In short, sections of the network fabric that carry model-parallel or other latency-sensitive traffic must be engineered for maximal performance and reliability, even if that means additional cost. At the same time, less sensitive portions of the network (such as the storage backup network, or training data ingest for batch jobs) can be designed with slightly higher latency or slight oversubscription to save cost. The challenge for designers is to **strike the right balance** – deliver **guaranteed low latency** and near-zero packet loss where it matters, but avoid over-engineering the entire network to those same extreme levels, which would be prohibitively expensive.

NetOpt is exceptionally well-suited to help planners navigate these trade-offs. It allows one to specify **differentiated service requirements** for different traffic flows or network segments, and it will optimize the design to honor those constraints netopt.design netopt.design. For example, using NetOpt, an architect can declare that all connections between a set of AI compute pods must have a end-to-end latency below, say, 100 microseconds (or a very low hop-count), and that this portion of the network must survive at least one link or device failure without disruption (99.99% availability). At the same time, the planner can allow the network connecting general-purpose IT servers or non-critical workloads to have a higher latency budget or an oversubscription ratio (trading some performance for cost efficiency). NetOpt will then **build a network design that meets these mixed SLAs** – perhaps by assigning more direct paths, higher link speeds, or extra redundancy to the AI pod interconnects, while using longer, cost-efficient routes for lower-priority traffic netopt.design. Crucially, because NetOpt's engine works across all layers, it can ensure that, say, two redundant IP paths for a critical AI flow do not accidentally share the same physical fiber or device, which would undermine true high availability netopt.design netopt.design. The tool can enforce such *strict disjointness* by design. If meeting a latency cap is challenging (e.g. for geographically dispersed datacenters running a distributed model), NetOpt might even suggest deploying an intermediate regeneration or edge computing site to shorten the longest links netopt.design – options a conventional single-layer planner would not consider.

In practice, applying these techniques means the network can be optimized to support AI training/inference without simply overprovisioning everything. A traditional approach to ensure low latency might be to build a completely non-blocking network with the highest-speed switches everywhere and two fully separate networks for redundancy – an extremely costly solution. NetOpt instead finds clever solutions to meet the **explicit latency and uptime targets** with minimal overbuild. For example, it might design the fabric so that certain latency-critical GPU-to-GPU communication flows go through only one spine hop (or an express optical circuit) and travel over short-range fiber paths, guaranteeing, for instance, sub-50µs latency netopt.design. For those links, it could also allocate fast failover paths on diverse hardware to achieve the 99.99% availability requirement netopt.design. Meanwhile, less critical traffic might be routed through a slightly longer path or a second-tier spine that introduces a few

microseconds more latency – which is acceptable for that traffic – thereby allowing the use of cheaper switching or a modest oversubscription in that part of the network. This **fine-grained allocation of network performance** is a game changer. It lets data center operators maximize GPU utilization and job throughput for their most sensitive AI workloads, while still optimizing cost and power by not over-engineering every link for the worst case. Supporting this approach, Cisco reports that AI clusters today often use techniques like RDMA over Converged Ethernet (RoCE) and priority flow control to ensure no packet loss, and dual-connected GPUs for fast failover cisco.com cisco.com – all measures aimed at keeping GPU-to-GPU communication speedy and reliable. NetOpt can incorporate all these design aspects (topology, link-level features, redundancy schemes) into a unified optimization. The end result is a network blueprint where **low-latency, high-bandwidth paths** are guaranteed for model-parallel or distributed training traffic, and **cost-effective scalability** is achieved elsewhere. Indeed, one analysis showed that to reach near-linear speedups in large-scale training, each GPU needed an All-Reduce bandwidth of ~20 GB/s (which equates to a 200–400 Gbps network link fully utilized), whereas at 5 GB/s per GPU the scaling efficiency dropped to ~70% fibermall.com. This underlines how critical network capacity is to AI performance. By using NetOpt, planners can identify exactly where such capacity is needed and provision it accordingly, rather than oversizing the entire network. Likewise, because large AI jobs can run for weeks, high availability is paramount – even a 99.9% reliable component will likely fail over a month in a cluster of 1000+ GPUs (60% chance of interruption) fibermall.com. NetOpt can design the network with sufficient path redundancy to meet far higher effective uptime (for example, dual fabrics or fast re-route that guarantee no single point of failure for critical traffic). In summary, NetOpt enables **AI-centric network engineering**: it enforces latency and resiliency targets per traffic class, ensuring that distributed AI workloads run at full speed without idle GPU time, while simultaneously containing costs by not overbuilding parts of the network that don't need it.

Additional Benefits and Use Cases of NetOpt.Design

Beyond the specific optimizations above, NetOpt.Design offers a number of general benefits for data center network planning:

- **Demand-Aware Topology Design:** Traditional data center networks are typically designed for an assumed worst-case "all-to-all" traffic pattern, which often leads to underutilized capacity (since real traffic is usually skewed or localized) cacm.acm.org. NetOpt allows **demand-aware** planning – the topology and link capacities can be optimized for the actual expected traffic matrix or patterns. Planners can input projected traffic between each pair of racks (or applications) and NetOpt will tailor the network to those flows. This can include asymmetric designs (more bandwidth between certain clusters) or even *slowly evolving topologies* that change over time. For instance, Google implemented a demand-aware network that reconfigures on a daily basis based on traffic predictions cacm.acm.org. With NetOpt, operators could achieve similar outcomes by planning multiple scenarios. The result is a more efficient fabric that provides bandwidth *where it's needed most*, rather than uniformly over-provisioning everywhere "just in case." This

demand-driven approach can significantly reduce cost while still delivering required performance for known workloads.

- **Statistical Traffic Engineering:** NetOpt can model traffic **variability** and provision the network to handle bursts probabilistically, instead of using a single peak number. In practice, data center traffic (especially AI and big data workloads) can be highly bursty – designing for the absolute peak leads to a lot of idle capacity. NetOpt solves this by allowing traffic demands to be input as distributions or time-series (e.g. day vs. night, or normal load plus occasional surge) netopt.design netopt.design. Planners can specify, for example, that an inference service usually needs 10 Gbps but might spike to 100 Gbps with 5% probability. NetOpt will then ensure the network meets a chosen **confidence level** – say, 90% of the time the bursts can be served without congestion netopt.design netopt.design. By optimizing to a percentile (the 90th percentile load in this example), the tool avoids the extremes of either overbuilding for a rare worst-case or under-provisioning for common surges. It right-sizes capacity such that there is an acceptable (and quantifiable) risk of slight clipping only in the most extreme 5–10% of scenarios netopt.design. This approach is far more efficient than the old method of adding huge safety margins at each layer "just in case" netopt.design. Networks designed with NetOpt's statistical traffic engineering will meet AI and cloud demand spikes with high probability yet have much less unused headroom during normal periods – directly translating to cost savings.
- **Avoiding Over-Redundancy (Cross-Layer Optimization):** Large-scale networks often suffer from **"margin stacking"** – each layer (IP, transport, etc.) is planned with its own safety margin and redundancy, resulting in duplicated buffers that waste capital netopt.design. NetOpt's integrated planning helps avoid this. Because it sees the full end-to-end picture, it can coordinate redundancy across layers. For example, if the optical layer provides 1+1 protected fiber paths between two sites, NetOpt might decide you don't also need two completely separate IP routes for the same connection, and instead can rely on the optical protection with a simpler IP routing configuration netopt.design netopt.design. Conversely, if fast IP reroute can cover certain failures, it may allow using cost-effective unprotected links at the physical layer. In a data center context, this might translate to using a single redundant super-spine layer instead of duplicating the entire spine fabric, because the system knows that any single device failure can be handled at one layer or the other. Industry experts have emphasized moving away from "worst-case at every layer" planning toward such a unified strategy netopt.design. Using NetOpt, one European carrier found they could carry the same traffic with significantly **fewer devices and less idle capacity** once the IP and optical plans were co-optimized netopt.design netopt.design. For data center operators, this means a leaner network that still meets reliability targets – eliminating hidden inefficiencies and lowering power, space, and management overhead.
- **Rapid What-If Analysis and Tech Innovation:** NetOpt's algorithmic approach enables quickly exploring **what-if scenarios**. Data center planners can ask questions like "What if we upgrade to 200 Gbps server NICs?" or "What if we double the number of AI servers next year?" and see how the network design changes. The tool can automatically re-optimize the topology and capacity for the new inputs, providing a clear roadmap for

scaling. This is invaluable for capacity planning and capital budgeting – it provides quantifiable outputs (like number of new switches or fibers needed) for future growth scenarios, taking the guesswork out of expansion planning. Additionally, NetOpt can help evaluate **new technologies** in the network. For example, as 800 Gbps and 1.6 Tbps switch ports emerge, or as new optical switching techniques become available, NetOpt can incorporate their cost and performance parameters to see at what point they become beneficial in the design. In this way, the tool acts as a kind of **digital twin** for the network, where architects and business stakeholders can test ideas (e.g. adding an optical bypass, using an alternative topology like a flattened butterfly, etc.) in simulation before making real investments. This accelerates innovation in data center networking by providing confidence backed by optimization and data.

Conclusion

As data centers evolve – particularly with the rise of AI superclusters and cloud-scale applications – network design is becoming a critical differentiator. Leaf–spine topologies must scale to unprecedented levels, support diverse workloads with differing SLAs, and possibly integrate new optical technologies to keep up with demand. NetOpt.Design offers a powerful solution for navigating this complexity. By optimizing across layers and across multiple objectives, it produces designs that meet high-bandwidth, low-latency requirements **at minimal cost** netopt.design netopt.design. It enables network architects to **unlock design flexibilities** (like multi-tier Clos, optical spines, and demand-specific provisioning) that were difficult to fully exploit with manual planning. The end result is a data center network that is *tailored* to the operator's unique needs: capable of supporting AI training traffic and mission-critical services with assured performance, while remaining efficient and agile as those needs evolve. In short, NetOpt helps build data center networks that are **AI-ready, cost-effective, and future-proof** – delivering the low-latency, high-capacity connectivity required by today's distributed applications, without breaking the budget netopt.design netopt.design.

Sources: The insights and data points above are drawn from a combination of NetOpt.Design's technical whitepapers and industry references on modern data center networking. Key references include NetOpt's AI networking whitepaper (2025) for multi-layer optimization features netopt.design netopt.design, Cisco's validated design guides for AI/ML clusters emphasizing low-latency, lossless fabrics cisco.com, recent research on reconfigurable optical datacenter topologies from Communications of the ACM cacm.acm.org cacm.acm.org, and case studies of large-scale AI network deployments (e.g. FiberMall's AIPod cluster and Google's Jupiter network) that illustrate the design strategies discussed fibermall.com cacm.acm.org. These references highlight the state-of-the-art approaches that NetOpt.Design encapsulates into an intelligent planning tool for network architects and planners.