ciena CORNING Enablence Infinera VIAVI

LIGHTWAVE® + BTR

State of the Market:

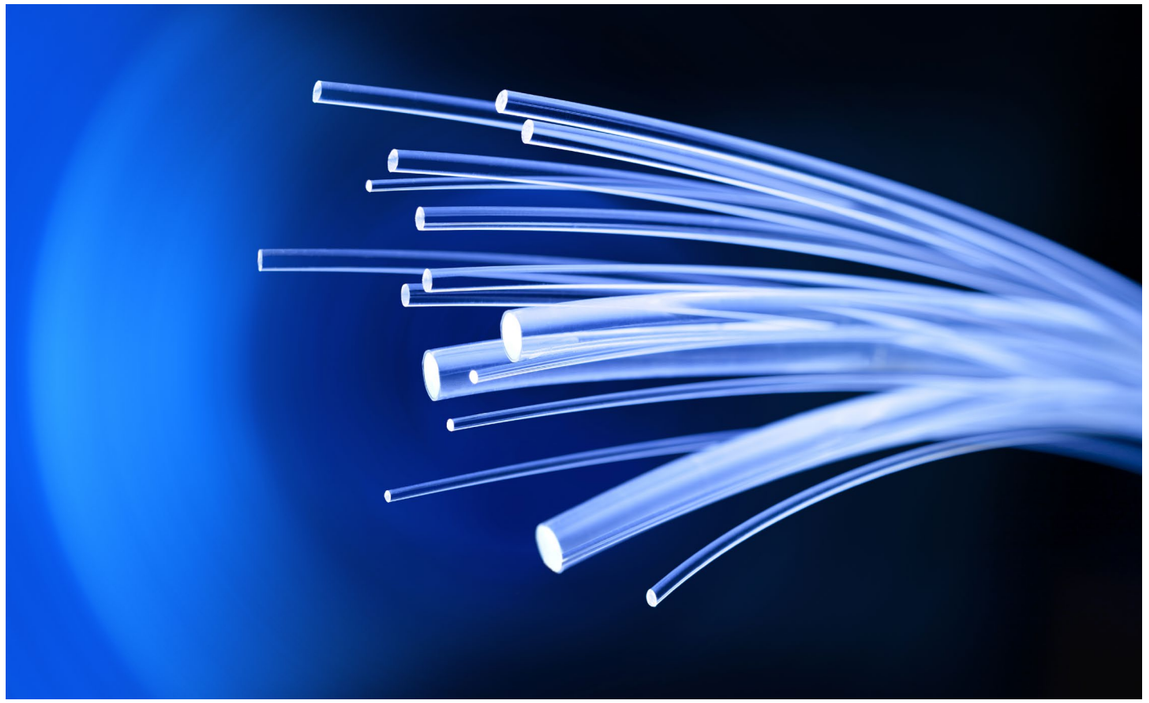
AI is Driving New Thinking in the Optical Industry

AI is Driving New Thinking in the Optical Industry

TABLE OF CONTENTS

EXECUTIVE SUMMARY

The year 2024 marked an inflection point for AI. In August, OpenAI's ChatGPT reached 200 million weekly active users.[1] Meanwhile, McKinsey reported that 72% of respondents in a business survey said that their organizations had adopted AI, compared to 55% just 10 months before. More tellingly, 55% of participants said their companies were using AI heavily, nearly double the 33% listed in the previous survey.[2] Although these numbers are just snapshots, they capture trends across the industry and the general population that have profound implications for AI-driven network traffic — and for the optical communications industry that serves it.



AI/ML applications depend on the near-constant exchange of massive data. In the early days of AI, training might have required a few hundred GPUs operating in parallel, but then the trained AI model could be stored and run on a single server. With the large multimodal models (LMMs) used for modern applications like generative AI, training now requires tens of thousands of GPUs. Afterward, the trained models need to be shared across multiple GPUs. Pure bandwidth is no longer enough — modern AI/ML demands highly synchronized, lossless communications to interlink massive numbers of GPUs with one another, CPUs, storage servers, and end users in the outside world.

In early 2024, *Lightwave* surveyed to explore how the optical communications ecosystem viewed the challenge of AI/ML traffic. Given the surge in AI adoption as the year has worn on, we decided to revisit the topic more granularly to understand how demands have changed and how companies are responding.

AI 2.0, our current survey, drew on a broad pool of participants from the *Lightwave+BTR* and *Data Center Frontiers* audience. Organization types included carriers, data center operators, internet content providers (ICPs), and more (see "About This Survey" at the end of this report). With annual revenues ranging from less than $1 million to over $5 billion, the companies represented various slices of the optical communications ecosystem, from startups to hyperscalers, from enterprise networks to international carriers. We had strong representation from management to provide macroeconomic insights and from all levels of engineering and R&D to shed light on the issues presented by AI/ML and technologies considered to solve them.

Read on to discover trends in AI/ML in the optical communications industry.

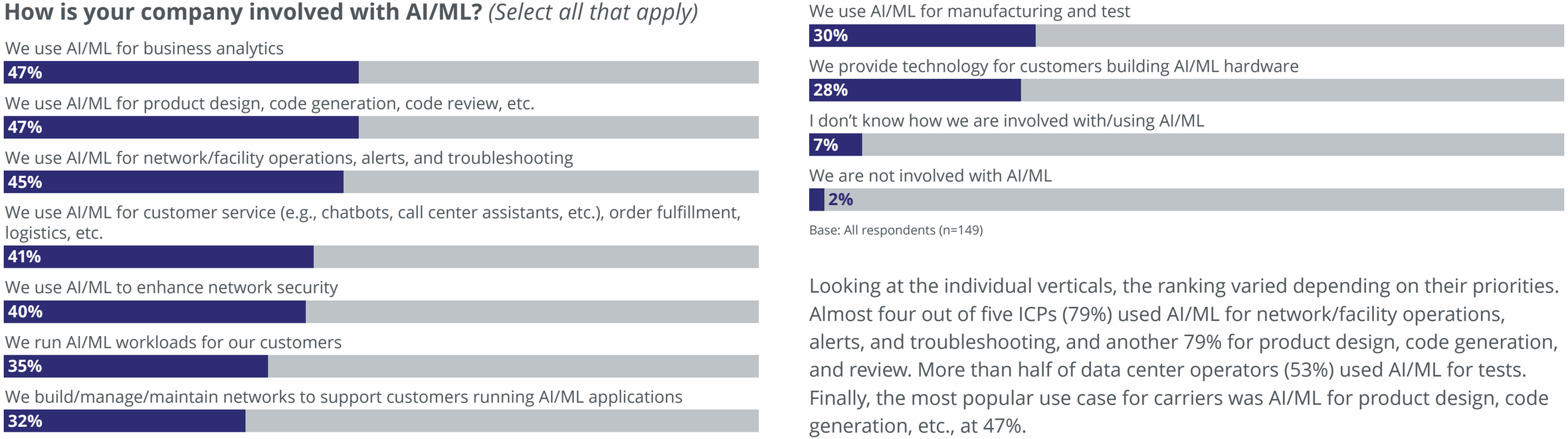
For the optical networking community, AI/ML is both a service and a tool

As a starting point, we asked participants how their organizations were involved with AI/ML. More than a third of respondents (35%) run AI/ML workloads for customers. A more granular look at the input revealed that more than half of data center operators (53%) and ICPs (57%) included serving AI/ML customers as one of their top use cases. This category also tilted toward larger companies; in particular, three out of five (62%) of organizations with annual revenues of $1 billion per year provided AI/ML as a service. It was not, however — and this is important — their primary AI/ML use case. Instead, AI/ML is fast becoming essential to how optical communications companies run their businesses (see Figure 1).

Like companies across the broader economy, optical communications organizations use AI/ML for standard applications like business analytics (47%) and customer service (41%). That's not surprising. What is surprising is that AI/ML networking tools are also in use to simplify highly technical tasks (more on that later). The two most common were AI/ML for product design, code generation, code review, etc. (also 47%), and network/facility management and monitoring (45%). The areas with the least AI/ML involvement across the survey population used AI/ML for manufacturing and testing (30%) and providing AI/ML hardware (28%).

FIGURE 1:

How is your company involved with AI/ML? *(Select all that apply)*



Base: All respondents (n=149)

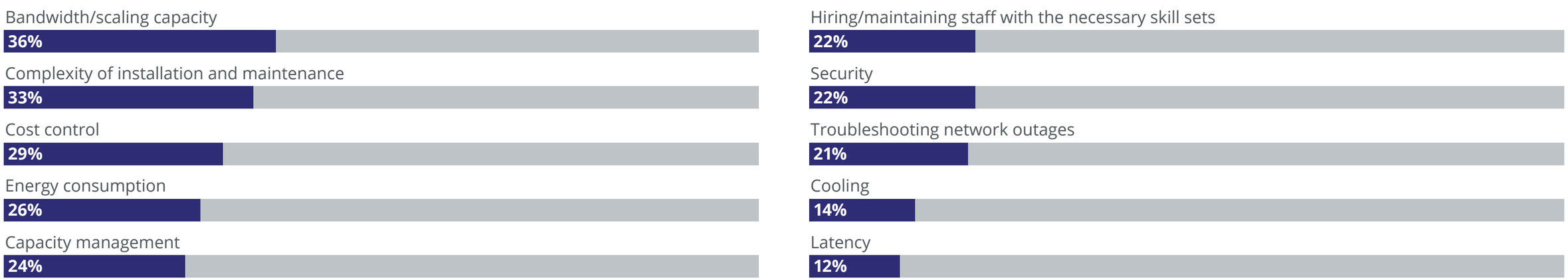
Looking at the individual verticals, the ranking varied depending on their priorities. Almost four out of five ICPs (79%) used AI/ML for network/facility operations, alerts, and troubleshooting, and another 79% for product design, code generation, and review. More than half of data center operators (53%) used AI/ML for tests. Finally, the most popular use case for carriers was AI/ML for product design, code generation, etc., at 47%.

The pain points in running AI/ML workloads

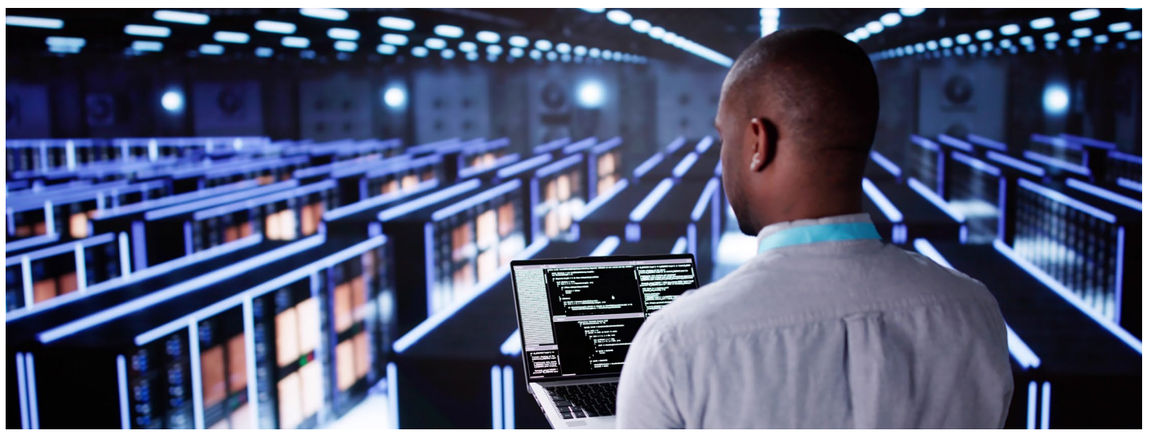
To understand the problems that AI/ML applications create, we turned to the engineers in our survey who work with it every day: design engineers, network and systems engineers, and IT/IS. Their top pain points in running AI/ML workloads were bandwidth/scaling capacity (36%), complexity of installation and maintenance (33%), cost control (29%), and energy consumption (26%; see Figure 2).

FIGURE 2:

What are your pain points in running AI/ML workloads, either for internal purposes or for customers? *(Select all that apply)*



Base: Engineering titles (n=58)



Engineers at service providers ranked capacity scaling (45%) and installation and maintenance complexity (30%) as top concerns, followed by capacity management, cost control, energy consumption, and the skills gap, all at 27%.

Engineers at data centers considered energy consumption and cost control as their primary challenges at 57%. These were followed by capacity scaling and complexity, each at 43%.

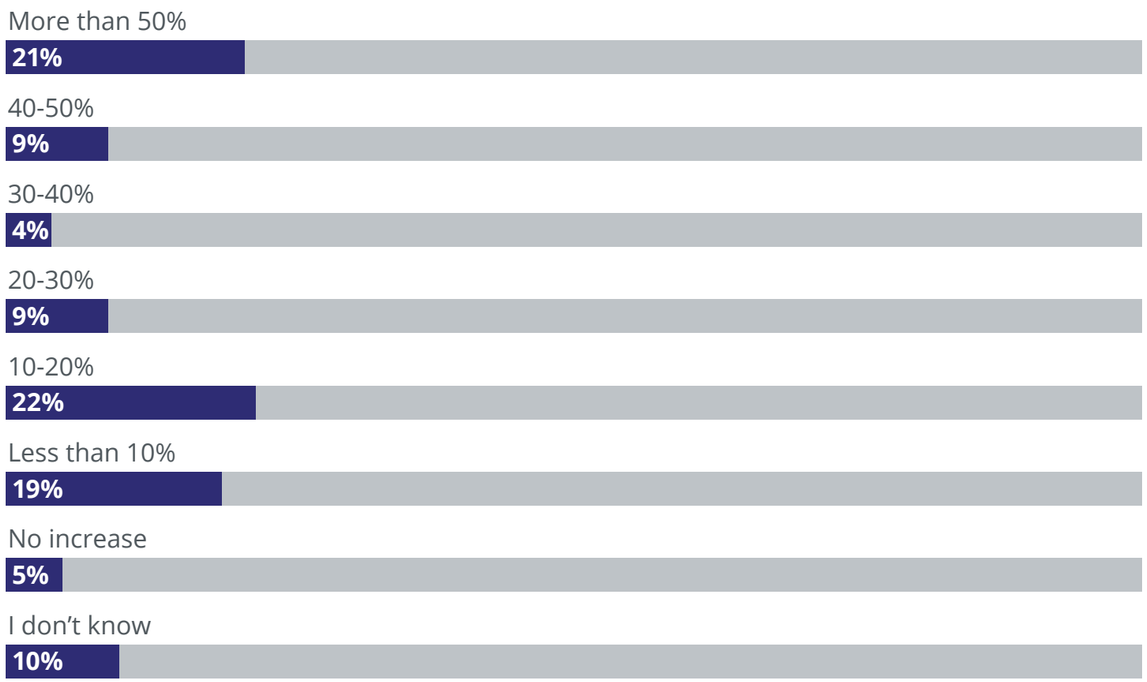
Let's dive a little more deeply into some of the key points.

AI/ML workloads are driving up demand for optical capacity

In the data center: Two in five respondents (41%) said that AI/ML workloads have increased optical capacity demand inside the data center by 20% or less, while one in five (21%) said it has grown by more than 50% (see Figure 3).

FIGURE 3:

To what extent are AI/ML workloads (e.g., generative AI, operational AI/ML, etc.) increasing optical capacity demand at your organization? – Inside the data center



Base: All respondents (n=149).

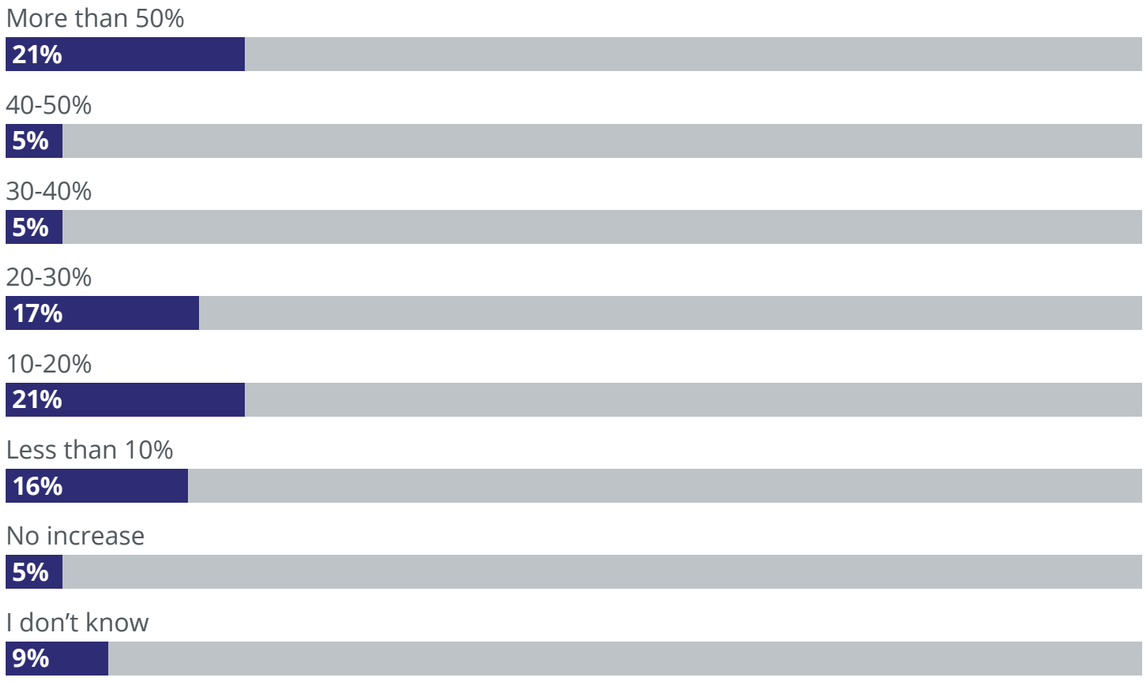
The split in the data, with a larger percentage of participants reporting more modest growth, merits discussion. For starters, it's worth noting that although AI/ML has seen a significant jump in usage this year, we're still in the early stages of the ramp.

A more granular look at the data also provides insights. If we correlate changes in capacity demand with annual revenues, the majority of organizations associated with 50% plus growth are those reporting annual revenues in excess of $1 billion per year — hyperscalers and large ICPs. At the other end of the spectrum, lower growth is correlated with organizations posting annual revenues of $25 million or less, such as private network operators and small cable providers. Half of ICPs (50%) cited capacity growth at the high end, whereas for data centers, it appeared to be more evenly split by the scale of the operation.

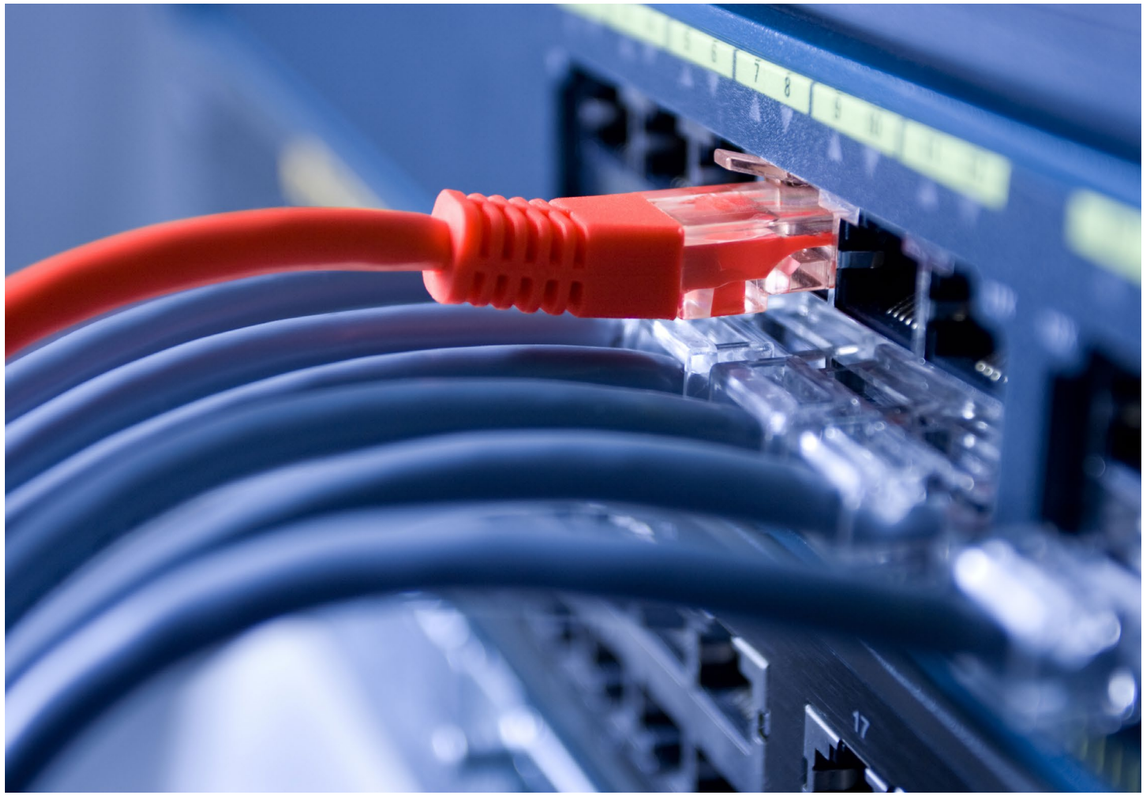
Over DCIs: In the DCIs segment, results were similarly divided. Two in five respondents (38%) said that AI/ML workloads have increased their optical capacity demand over DCIs by between 10% and 30%, while one in five (21%) saw increases of more than 50% (see Figure 4).

FIGURE 4:

To what extent are AI/ML workloads (e.g., generative AI, operational AI/ML, etc.) increasing optical capacity demand at your organization? – Over data center interconnects



Base: All respondents (n=149).



A closer look at the data revealed the same correlation between annual revenues and growth as for data center results, with large carriers and hyperscalers more likely to report top-line growth. Drilling down into verticals, half of ICP respondents cited 50% plus growth. More than a quarter of data center operators (26%) also fell into this group; notably, another 31% cited growth of 20% or less.

Power is the prevailing problem

Excess power consumption in the data center has been a concern for over a decade, but it has been intensified by AI's vastly increased computational requirements. For example, xAI's Colossus supercomputer incorporates 100,000 GPUs, with plans to double it in the future. Meta and Microsoft are likewise in the hunt. Running a cluster of 100,000 Nvidia H100 GPUs requires an estimated 150 MW of power.

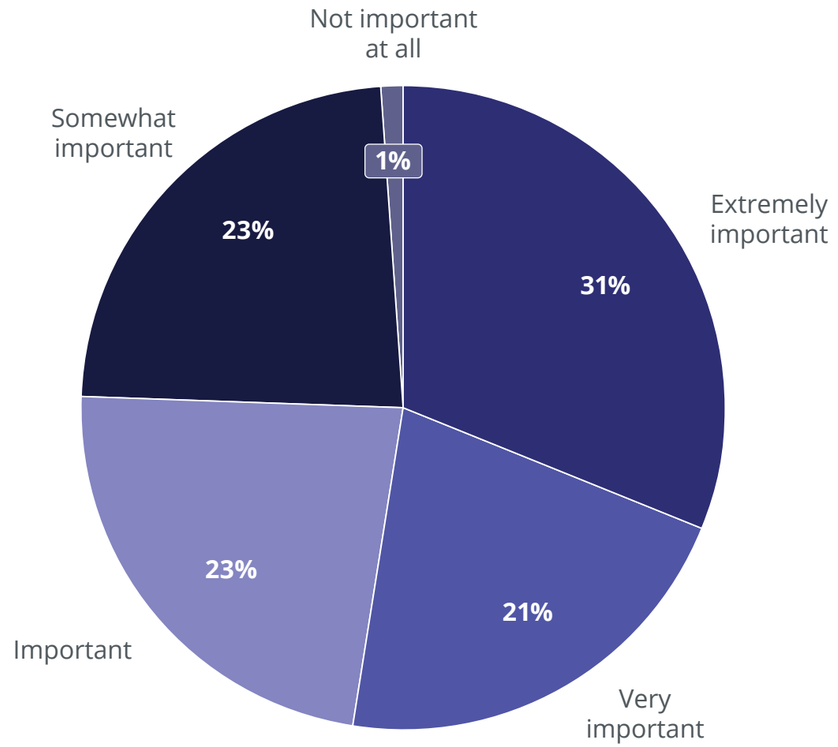
As a result, the location of these types of data centers is critical to success. Bringing 100,000 GPUs online is meaningless if there's insufficient power to run them. That might explain why the Colossus facility is located in Memphis, Tennessee, where it recently received approval for 150 MW of power from the region's Tennessee Valley Authority utility.

We asked our survey participants whether energy availability from their local grid affected their choice of data center locations. Almost two-thirds of data center operators (63%) said yes.

Meanwhile, efforts are afoot to reduce energy consumption wherever possible, including in the network. Conversations on ways to minimize picojoules per bit and cents per gigabit have dominated conferences and meetings with product engineers and customers. Indeed, half of the respondents in our survey (52%) said that the energy efficiency of their networking gear and components was extremely or very important to their organizations (see Figure 5). A quarter (24%) said that it wasn't important (Somewhat important + Not important at all).

FIGURE 5:

How important to your organization is the energy efficiency of your networking gear and components?



Base: All respondents (n=149).

Not surprisingly, there were variations, depending on the verticals. Sixty-four percent of ICPs considered power efficiency to be extremely important or very important, while 58% of data center operators gave that response.

TECHNOLOGIES TO SERVE AI/ML

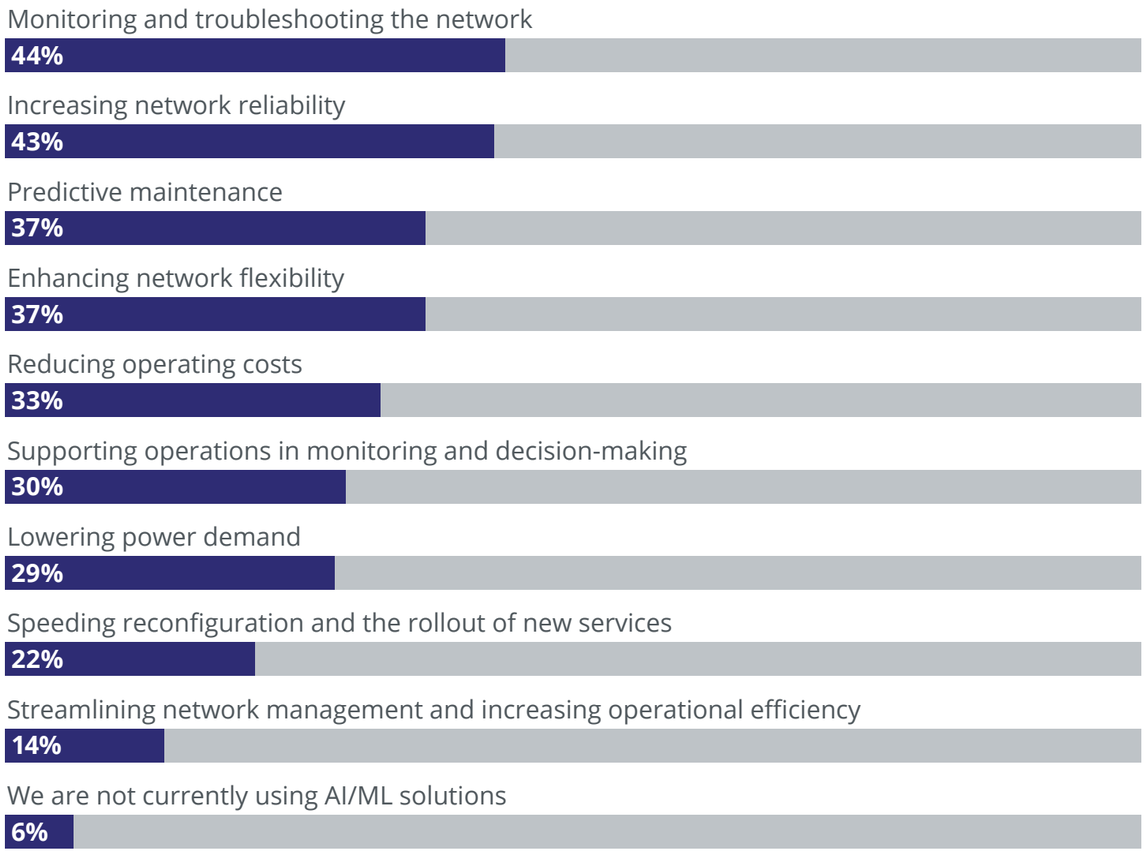
To understand how organizations are building and maintaining networks to support AI/ML traffic and address the pain points identified above, we returned to our group of design engineers, network/systems engineers, and IT/IS.

AI/ML networking solutions help tackle the AI/ML challenge

Organizations are using AI/ML tools to help address many of the pain points identified above. The three most popular categories concerned solutions that combine solutions network availability: monitoring and troubleshooting the network (44%), increasing network reliability (43%), and predictive maintenance (37%; see Figure 6). Organizations use AI/ML tools to reduce operating costs (33%). They're targeting the skills gap by using AI/ML networking solutions to assist staff in monitoring and decision-making (30%). Network management techniques can be used to reduce power demands and more efficiently leverage bandwidth. Almost three in 10 organizations are applying the technology in this way.

FIGURE 6:

In what areas, if any, is your organization using/investigating NETWORKING SOLUTIONS powered by AI/ML? *(Select all that apply)*



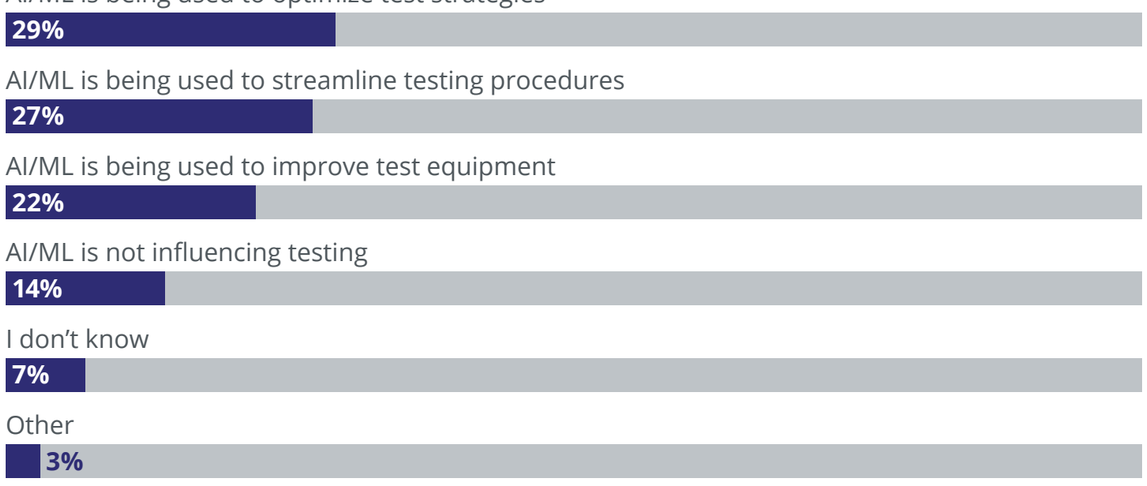
Base: Engineering titles (n=58)

AI/ML is changing the face of testing

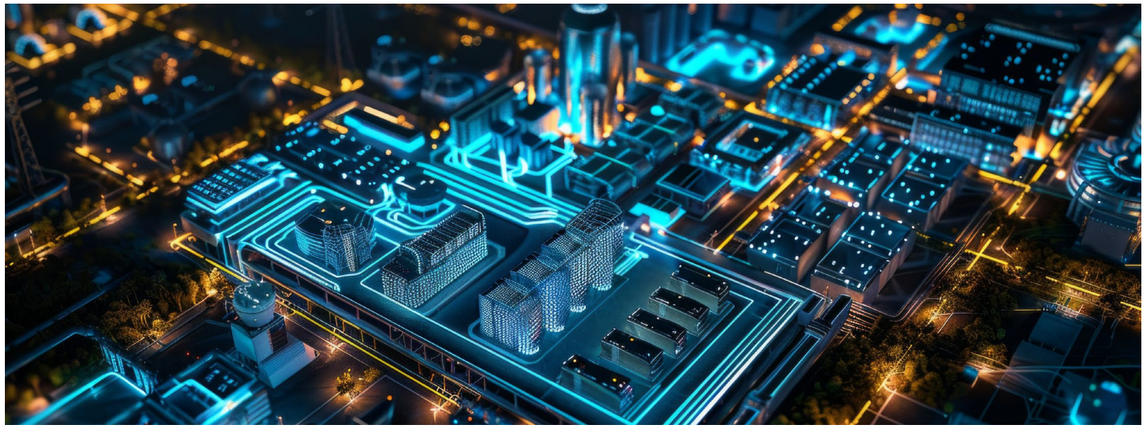
Testing is fundamental to network design and operation. To understand the effect of AI technology on testing, we focused on a subset consisting of design engineers, network and systems engineers, IT/IS, and R&D staff. Nearly three in 10 in this subset (29%) said AI/ML is used to optimize test strategies (see Figure 7). Another 27% said that it's being used to streamline testing procedures. Less than one in six (14%) said they didn't think AI/ML influenced testing.

FIGURE 7:

How do you see AI/ML influencing testing approaches in the optical networking industry?



Base: All technical titles (n=74)

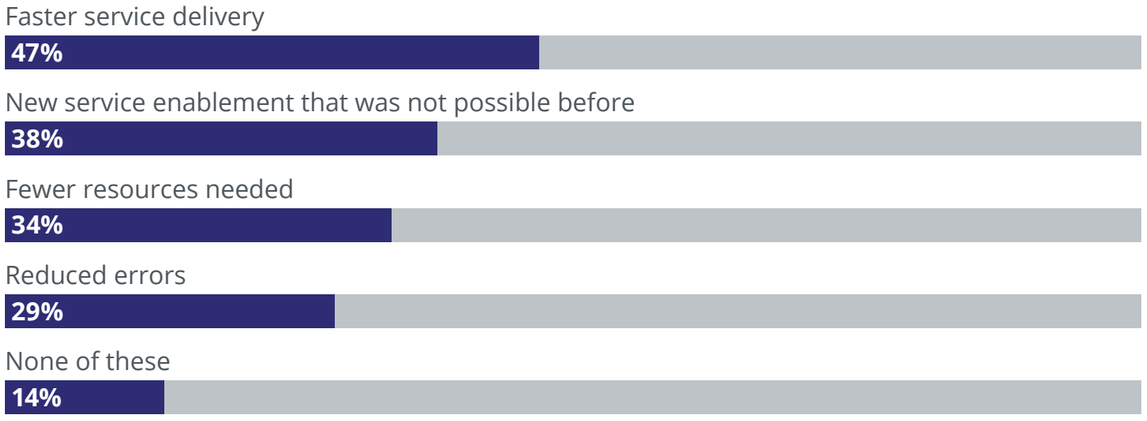


What were the benefits provided by AI/ML technology?

Engineering respondents saw the top benefits when AI/ML technology was applied to existing projects: faster service delivery (47%), new service enablement not previously possible (38%), and fewer resources needed (34%; see Figure 8).

FIGURE 8:

For existing projects where you have applied AI/ML technology, what benefits have you seen? *(Select all that apply)*



Base: Engineering titles (n=58)

Transmission rates throughout the network

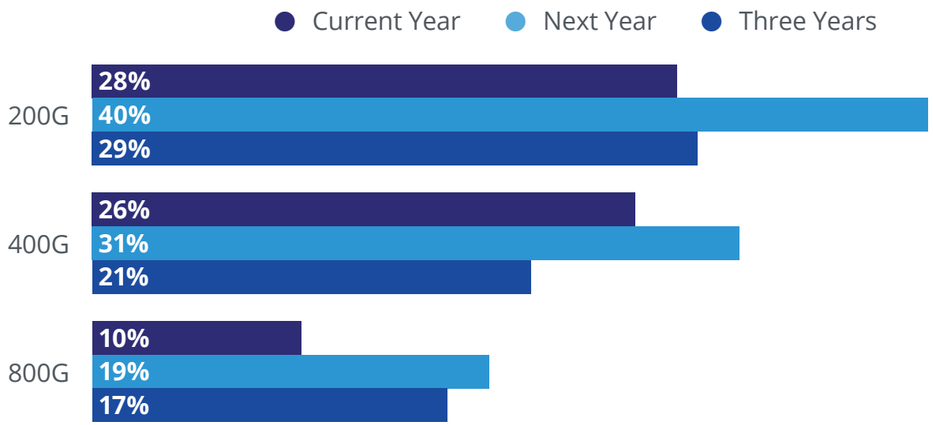
To gauge the maturity of the AI/ML evolution, we asked respondents about their current and planned transceiver usage. Because of the many variations of the transceiver types and network architectures, we grouped data by aggregate transmission rates. Engineering respondents could choose from:

- 200G: 8 x 25G, 4 x 50G, 1 x 200G
- 400G: 8 x 50G, 4 x 100G, 2 x 200G
- 800G: 8 x 100G, 4 x 200G

For communications between GPU clusters, almost three in 10 (28%) reported using 200G (see Figure 9). That number rises to 40% for next year, likely as a result of slower (e.g., 10G) links being upgraded. 400G had a strong foothold, which is supported by anecdotal descriptions of Meta and xAI both using 400G in their new large-scale clusters. 800G makes an appearance, growing over time.

FIGURE 9:

What aggregate transmission rates are your organization using to connect GPU clusters this year, next year, and in three years?
(Select all that apply)



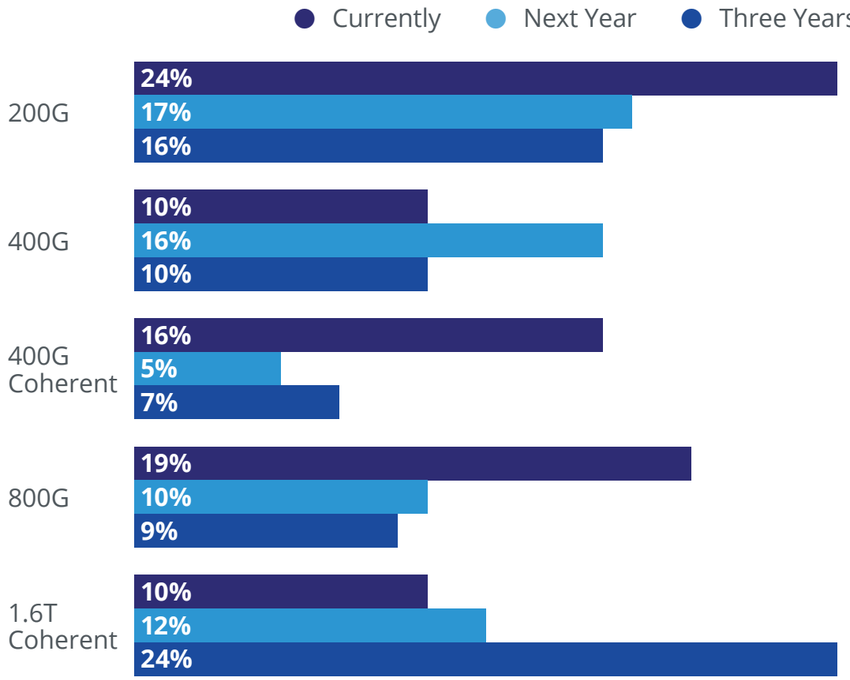
Base: Engineering titles (n=58)

We also examined trends in aggregate data rates for DCIs. We added 400G coherent and 1.6T coherent for this data set. While 1.6T coherent solutions based on 4 x 400G coherent transceivers are broadly available, 2 x 800G coherent transceivers are only reaching the market and will not be broadly available until next year.

The current and planned aggregate transmission rates for DCIs show some anticipated behaviors (see Figure 10). 200G is currently used by 24% of respondents but tails off over the coming years. 800G (19%) and 400G coherent (16%) are the workhorse technologies. Notably, 1.6T coherent solutions have a light foothold but are poised to explode in the coming years.

FIGURE 10:

What aggregate transmission rates are your organization using for DCIs this year, next year, and in three years? *(Select all that apply)*



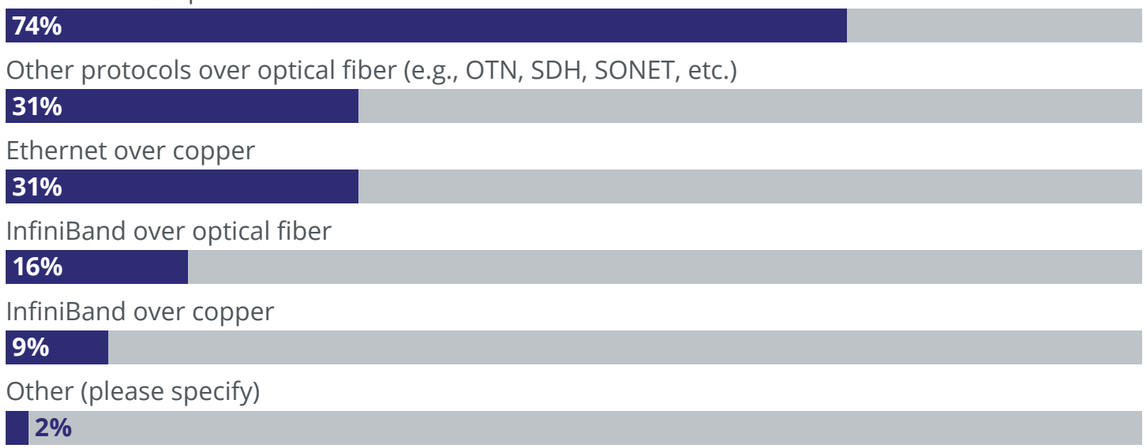
Base: Engineering titles (n=58)

Optics is winning the transmission protocol sweepstakes

There was a time when copper dominated data center communication. Although it still plays a role today, even in large-scale AI/ML GPU clusters, optical communications is ascendant. Almost three-quarters of respondents (74%) said that their organizations were using Ethernet over optical fiber to accommodate AI growth; this included 100% of engineers working at data centers (see Figure 11). The transmission protocol that was least used was InfiniBand over copper (9%).

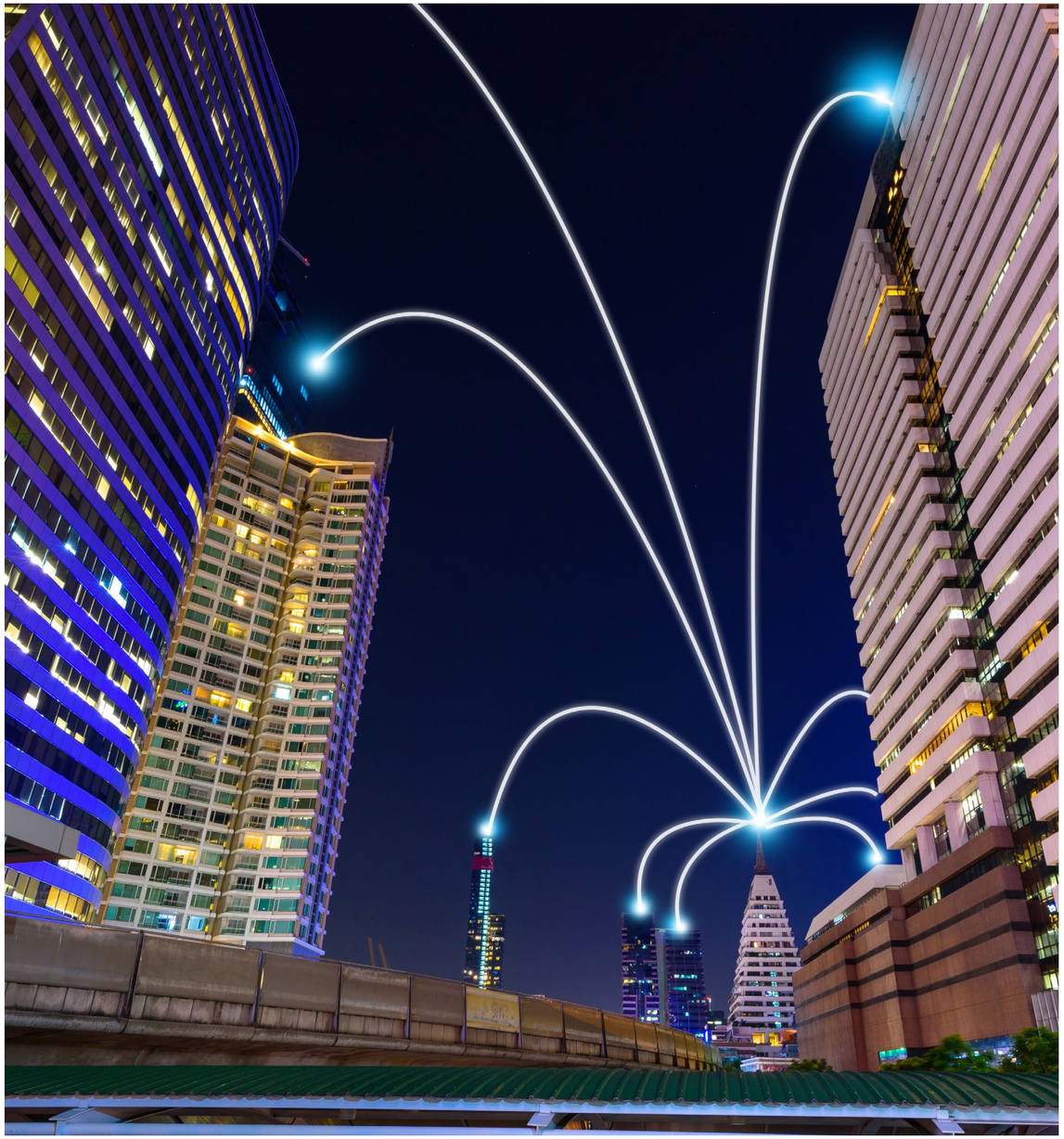
FIGURE 11:

What transmission protocols are your organization using to accommodate AI growth? *(Select all that apply)*



Base: Engineering titles (n=58)

The usage rates of InfiniBand relative to optical Ethernet may appear low, given InfiniBand's history of having been tailored for high-performance computing. Recent modifications to the Ethernet standard have addressed issues of latency and packet loss, however, making Ethernet's performance more competitive with InfiniBand. As a result, the choice of protocol arrays from project to project. xAI's Colossus, for example, is using Ethernet for communications, as are Meta and AWS.



CONCLUSION

The rapid adoption of AI/ML presents the optical communications sector with both challenges and opportunities. The top challenges include searching capacity demand, the complexity of optical networking technologies, rising costs, and increased power consumption. Organizations are addressing some concerns using AI/ML networking solutions. They're also looking to the vendor community to develop components and systems to address the pain points above. Co-packaged optics (CPOs), for example, promise to reduce power consumption, cost, complexity, and latency while increasing bandwidth. Survey respondents are investigating these and other near-term solutions, along with emerging high-speed coherent pluggable transceivers. The market for AI/ML is poised to explode, but when it does, the optical networking ecosystem will be ready with solutions.

ABOUT THIS SURVEY

The pool of participants consisted of 149 qualified industry insiders. More than two in five respondents (43%) identified as upper management, followed by network/systems engineering (16%) and design engineer/engineering (15%). Combine the engineering titles with R&D and IT/IS management; nearly half of the respondents (49%) had technical backgrounds. Three-quarters of respondents (81%) were in North America, followed by Europe, Africa, Asia, and the rest of the world.

The business types were tilted toward users of optical communications technology rather than vendors, with 55% identifying as national/international or regional carriers, 13% identifying as data center operators, and 12% identifying as other network operators, such as enterprise, utility, small government, education, etc. The remaining cohorts include internet content providers, cable operators, and wholesale fiber providers.

The survey cast a wide net, with more than a third (37%) of respondents reporting annual revenues of more than $500 million, with another third (32%) reporting less than $25 million annually. Cross-correlation with primary business focus indicates that respondents from hyperscalers and large national/international carriers weighed in.

REFERENCES

1. *OpenAI says ChatGPT's weekly users have grown to 200 million*
2. *The state of AI in early 2024*