How AI Capital Expenditures Impact Semiconductor Markets

Max Xing

March 2026

1 Introduction

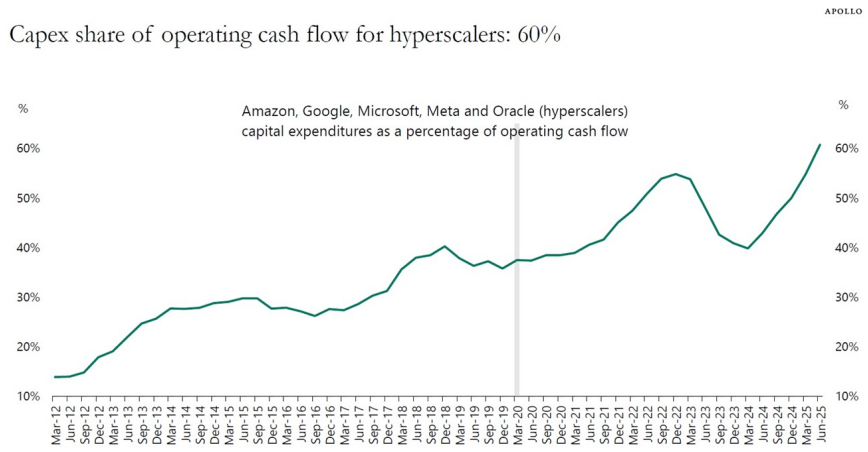
Artificial intelligence has quickly become one of, if not the most important topics in the financial markets. Massive tech companies, many of whom were the largest winners of the last decade of software expansion, have begun pouring billions of dollars into AI investments. However, although news sources and the broader public love focusing on Large Language Models (LLMs) such as OpenAI's ChatGPT or Anthropic's Claude, what I believe is far more important is the underlying hardware that supports this AI spending boom. Specifically, although the market knows names such as NVIDIA and AMD as drivers for the AI boom, there is significantly less focus from the general public on companies such as Astera Labs or Cerebras Systems. In this report, I argue that the true winners from the massive AI capital expenditure spend will reside in two subgroups: companies that develop bespoke inference chips, and companies that design optical chips for data center networking.

2 Investment Thesis

AI spending is driving a multi-year semiconductor super cycle, but will likely also create many losers and a few winners. Due to the structural constraints of AI scaling, I believe that the two sectors that will accrue the most value in the near future are inference computing and optical AI networking.

Most importantly, AI capital expenditures are reaching an all-time high. The chart below from Apollo highlights how capital expenditure spend is quickly rising as a percentage of hyperscalers' operating cash flows.

Figure 1: Capital Expenditures as a share of Operating Cash Flow

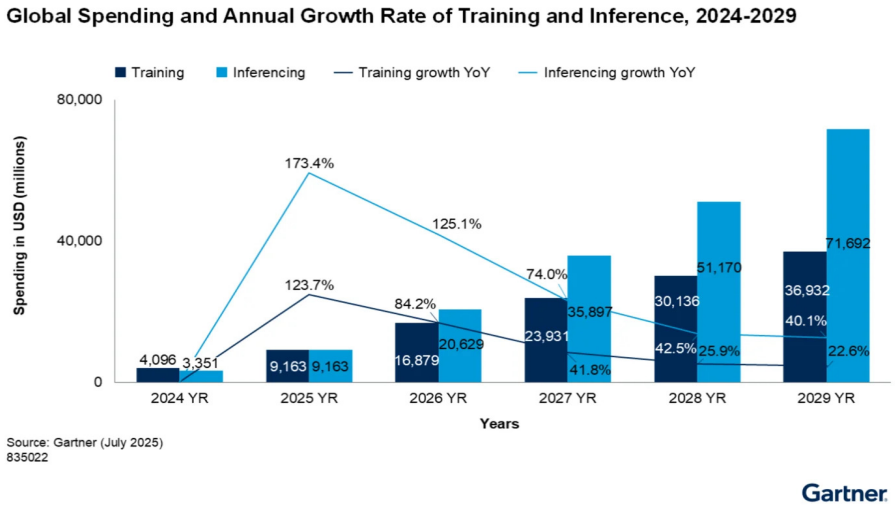


Furthermore, worldwide AI capital expenditure spend is projected to hit $1.7 trillion by 2030, funded primarily by US companies (which account for roughly 60-70% of global spend).

For context, the flow of capital expenditure spend can be roughly described as follows: hyperscalers spend billions of dollars on data centers, which are built out of physical servers and racks of semiconductors. For example, AWS might spend billions to construct data centers with servers assembled by an OEM/ODM such as Super Micro Computer, from chips designed by NVIDIA, SK Hynix, and Marvell before being manufactured by TSMC. This trickle-down structure means that the performance of these fabless semiconductor design companies depends almost entirely on end-market demand, which currently looks promising.

The first group of winners will be Application-Specific Integrated Circuits, or ASICs, for two main reasons. Firstly, as the industry pivots away from training models and towards applications, inference, and thus companies such as Baseten and Groq, become the primary focus. The following chart from Gartner shows the projected overtake of training spend by inference.

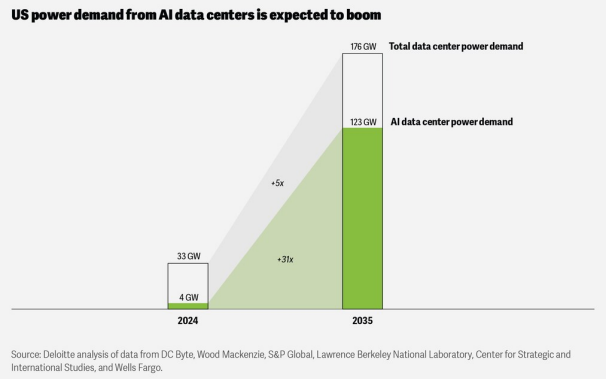
Figure 2: Inference vs. Training Spend



As inference workloads scale, general-purpose GPUs become less economically efficient than specialized silicon. This shift creates an opportunity for ASIC-based architectures optimized specifically for inference workloads.

Secondly, current energy constraints are rapidly forming the limits of what computer systems can do. This Deloitte chart shows how energy demand is projected to explode in the next couple of years.

Figure 3: AI Compute Energy Demand

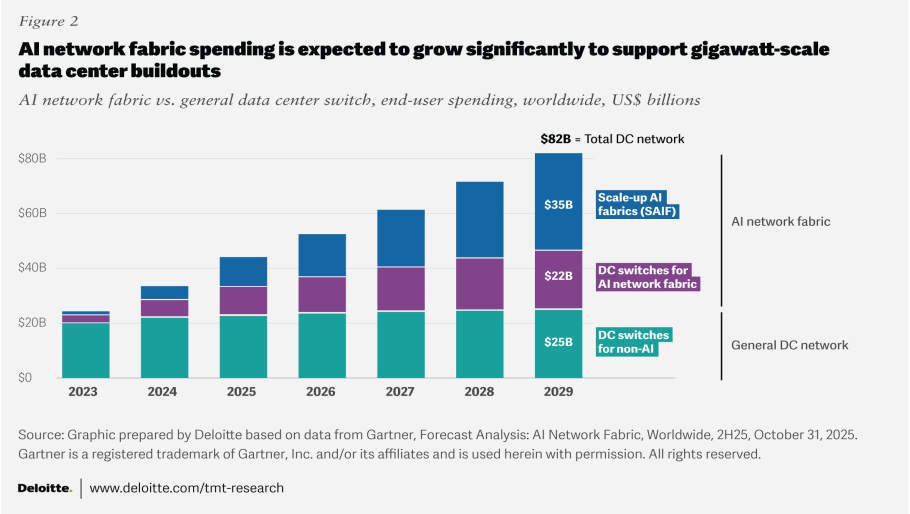


ASICs, due to their bespoke nature, are able to deliver higher performance-per-Watt. For example, Cerebras Systems is a company that pioneered wafer-scale

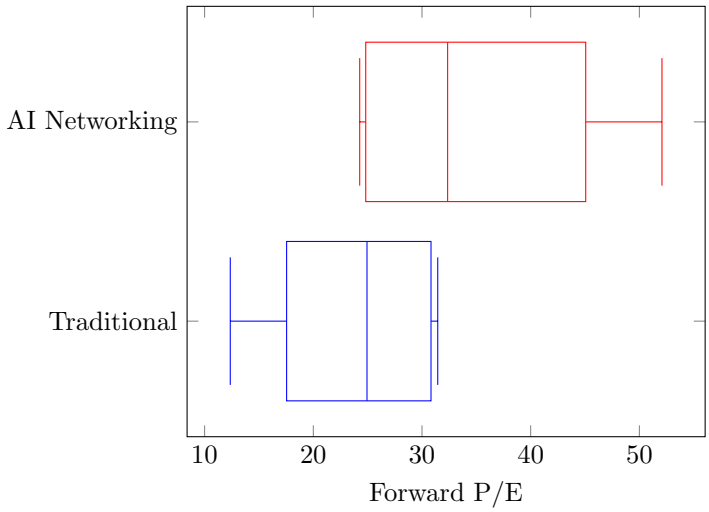
engineering (WSE), a successful approach primarily because it delivers over a 2x improvement in performance-per-Watt over general-purpose GPUs. In January, they received a $10 billion investment from OpenAI to build out 750 megawatts of computing power, a partnership that validated the startup as an important player in the AI inference space. So even within inference ASICs, I believe that the specific winners will be those that can create meaningful (e.g., orders of magnitude) power/energy improvements from existing systems.

The second sub-vertical of semiconductors in which I believe there will be significant winners is in networking. The figure below from Deloitte shows the increase in AI networking fabric spend as data center build-out increases.

Figure 4: AI Networking Fabric Spend



When comparing public market performances, fabless semiconductor companies focused on AI networking specifically tend to outperform traditional design companies. The following box graph compares the forward P/E multiples (as of mid-March) between the two groups, with the traditional class consisting of NVIDIA, AMD, TSMC, Qualcomm, and Applied Materials, whereas the AI networking category contains Credo, Astera Labs, Marvell, Broadcom, and Juniper Networks.

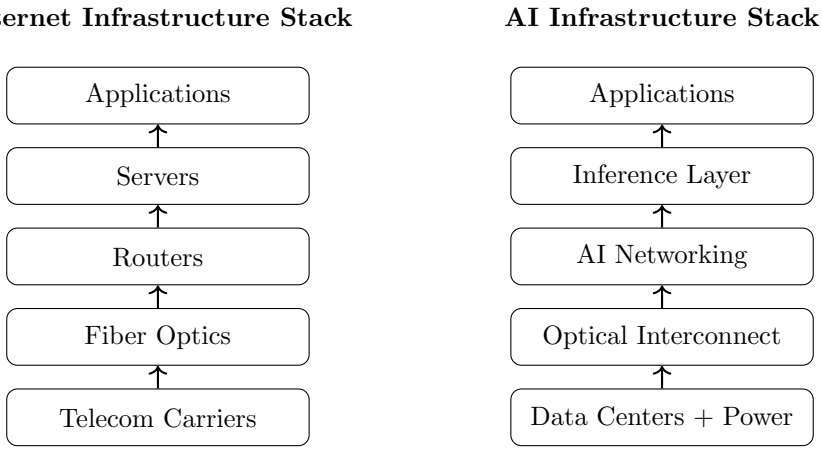


We can see that companies that focus on AI-specific networking tend to perform better and are thus command higher valuations. However, I believe that the market is still underpricing the imminent transition from copper to optics. We are currently at an inflection point where traditional copper is becoming obsolete and unable to handle rapidly increasing data rates. The industry solution is optics, and specifically, co-package optics (CPO), where optical transceivers are built directly in the same package. This avoids the need for electrical signals to travel from the switch ASIC across the circuit board to a separate optical module, which is an extremely power-intensive process. I believe that the true winners will be CPO and optics design companies that can gracefully navigate the manufacturing difficulties (i.e., incorporating CPO into Chip-on-Wafer-on-Substrate, or CoWoS) that will inevitably arise.

Altogether, the semiconductor industry always trends towards a few huge winners in each "bucket". For example, in the GPU sub-vertical, NVIDIA and AMD are the clear favorites, whereas in memory, SK Hynix and Samsung dominate. Although inference and networking are witnessing a huge uptick of startups touting new innovations, eventually, given the capital intensive nature of research and development in the space, few players emerge victorious. It remains to be seen who the winners actually are.

3 Historical Comparison: The Internet Boom

I believe the current AI build-out closely resembles that of the Internet in the late 1990s. Generally, the respective stacks for the Internet and AI can be described as follows.



There are parallels in the development phases of the two as well. They both begin with an infrastructure build-out, with fiber cables/routers in the Internet and data centers for AI. Then, there's a rapid investment and resulting expansion of this infrastructure, a phase that we are currently in for AI. Only after the hardware is laid out do applications and the top usage layer become increasingly sophisticated.

Importantly, a lesson to be learned is that many of the real winners in the Internet boom were not necessarily the companies that built the most infrastructure, but rather those that solved crucial scaling bottlenecks and developed the application layer. As data centers and infrastructure building scale, the question shifts to how to move and use the data. That's why I believe that the true winners will be the inference specific semiconductor companies that abstract away from infrastructure, as well as the networking companies that solve current constraints by efficiently moving data.

4 Risks

The semiconductor supply chain is famously long, complicated, and therefore fragile. Simply, most chips are designed in Silicon Valley before being manufactured in foundries in Taiwan (likely TSMC). However, the raw materials (e.g., silicon) for the chips come from places such as Japan/China, and the machines that perform EUV lithography originate from ASML, a Dutch company. Even different types of chips originate from different geographies: although the US dominates in microprocessor design (CPUs, GPUs, etc.), South Korea is well known for its memory chips with companies such as SK Hynix and Samsung. Thus, any complications at any point in the chain can cause ripple effects throughout the entire semiconductor market.

As a result, the first huge potential pitfall for the semis market moving forward

is geopolitical uncertainty. Taiwan has long been a point of contention between the United States and China, and any conflict would prove to be catastrophic for the semis market. Although TSMC has floated the idea of and even tried implementing new foundries in the US (e.g., in Arizona), the bulk of manufacturing and fabs remains back on the island. However, I believe that this scenario is extremely unlikely. China imports more semiconductors than oil, and many of their domestic companies rely heavily on chips manufactured by TSMC. China's historical strategy regarding Taiwan has also always been economic and political pressure/coercion rather than full-on destruction.

From a manufacturing standpoint, both inference ASICs and CPO are early-stage technologies with relatively unproven track records. Given that ASICs are definitionally application specific, they are less general than traditional microprocessors and thus require specific specializations when built. They also tend to be more complex by design, forcing higher precision lower yields. On the other hand, CPO are a huge step up in complexity due to their goal of incorporating optical transceivers directly inside the same package. Furthermore, there are significant thermal constraints: AI switches run extremely hot, whereas optics thrive in lower temperatures. Developing a production strategy that produces CPO in an effective and efficient manner with a satisfactory yield will likely prove to be a difficult engineering exercise.

5 Recommendation

In order to take advantage of this cyclical transition from infrastructure/training towards networking/inference, I recommend a pairs investing strategy that involves longing ASIC inference and CPO networking companies while simultaneously shorting general-purpose microprocessor design and traditional enterprise hardware companies. While it is unlikely that GPUs will lose all value, I believe that their growth rate has peaked relative to the networking side. As hyperscalers realize their massive GPU clusters are bottlenecked by energy constraints and data transfer speeds, the marginal next dollar of capital expenditures will shift away from microprocessors and toward buying more inference ASICs and optical transceivers. Furthermore, by combining the two strategies, the trade is relatively isolated from broader market risk/beta and instead capitalizes on the shift in growth momentum.

I'm overall bullish on these two specific semiconductor trades. Unlike the dot-com bubble, the AI boom necessitates physical data centers and compute, a commodity that will always be in demand. So even if the end-market demand for AI compute doesn't measure up to current projections, the construction of physical infrastructure should always generate returns.