

Data Center Networking for AI and Cloud

Accelerating today's and tomorrow's workloads

RESEARCH BRIEF



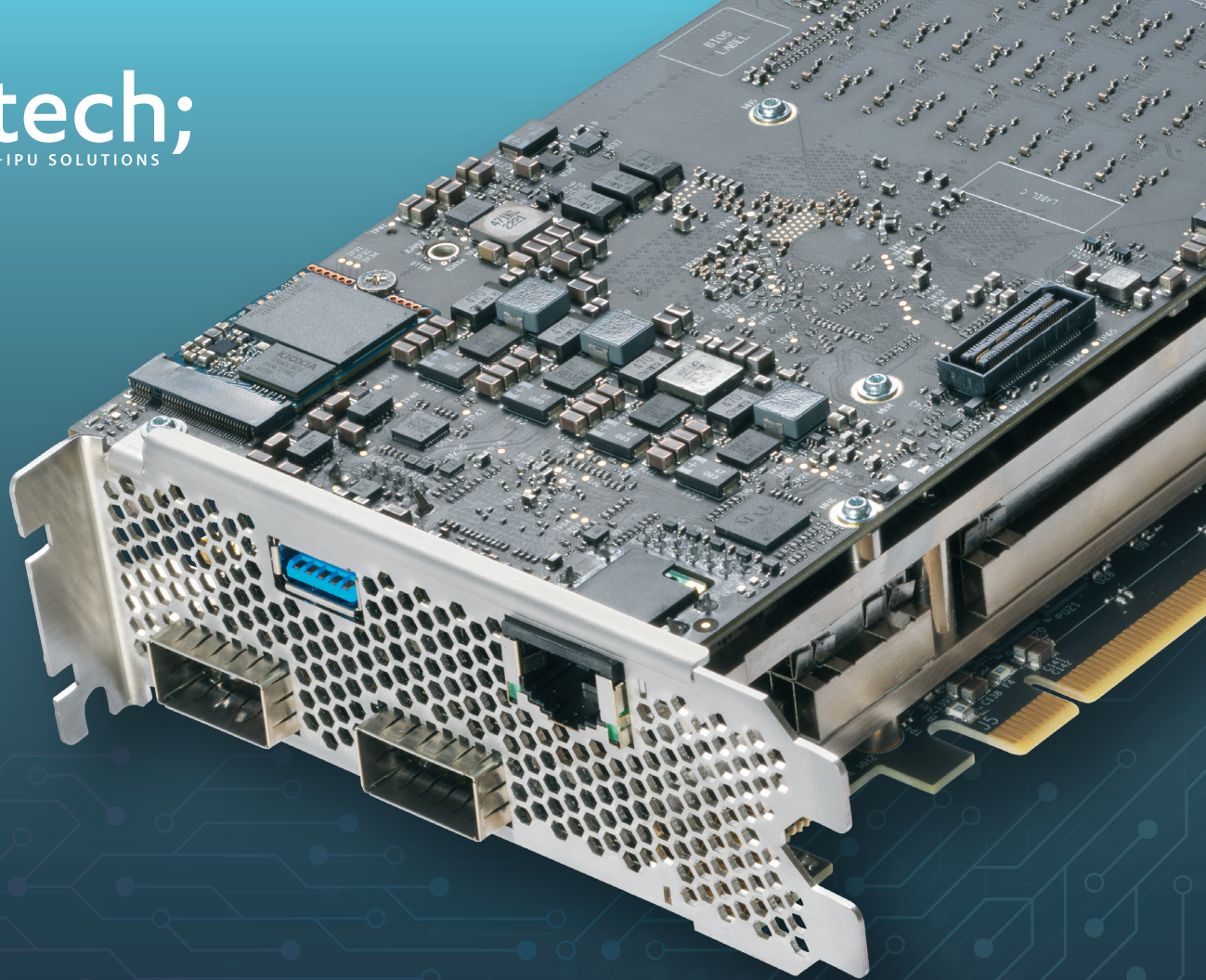
Table of Contents

| | |
|--|-----------|
| Introduction — What's Changed in Data Center Networking | 1 |
| GenAI Arrives with a Bang | 1 |
| Enterprise and Consumer Cloud Workloads Persist | 1 |
| The Workloads Driving Data Center Architectures | 2 |
| AI/Machine Learning and GenAI | 3 |
| Next-Gen Communications: 5G, Fiber, and Network Innovations..... | 4 |
| Interactive and Performance-Sensitive Edge Workloads | 4 |
| Cybersecurity: A Growing Imperative | 4 |
| Application Architecture Evolution | 5 |
| Microservices: The New Norm in App Development | 5 |
| Cloud Application Traffic Patterns: North-South to East-West | 5 |
| You Get a GenAI Cluster, Everyone Gets a GenAI Cluster | 5 |
| GenAI: Revisiting HPC..... | 5 |
| The Shifting Network Architectures of Data Centers | 8 |
| Bifurcation of the Data Center | 8 |
| The AI Section of the Data Center | 9 |
| The Cloud Section of the Data Center | 11 |
| Evolution of Hardware Network Acceleration | 12 |
| SmartNICs/SuperNICs/IPUs/DPUs - What's in a Name?..... | 13 |
| OCP Open Domain-Specific Architecture (ODSA) and Chiplets..... | 13 |
| Network Acceleration Updates - Hyperscaler and Vendors | 13 |
| Amazon Web Services | 13 |
| Google Cloud Platform | 14 |
| Microsoft Azure | 14 |
| Network Acceleration Vendors | 15 |
| Summary and Recommendations | 16 |

Research Briefs are independent content created by analysts working for AvidThink LLC. These reports are made possible through the sponsorship of our commercial supporters. Sponsors do not have any editorial control over the report content, and the views represented herein are solely those of AvidThink LLC. For more information about report sponsorship, please reach out to us at research@avidthink.com.

About AvidThink

AvidThink is a research and analysis firm focused on providing cutting-edge insights into the latest in infrastructure technologies. Formerly SDxCentral's research group, AvidThink launched as an independent company in October 2018. AvidThink's coverage includes 5G infrastructure, private wireless, edge computing, SD-WAN, SASE, SSE, ZTNA, cloud and containers, SDN, NFV, and infrastructure security. Our clients range from Fortune 500 enterprises and hyperscalers to tier-1 communications service providers, fast-growing unicorns, and innovative startups. AvidThink's research has been quoted by the Wall Street Journal, Light Reading, Fierce Telecom, Fierce Wireless, Mobile World Live, and other major publications. Visit AvidThink at [avidthink.com](https://www.avidthink.com).



Bring the benefits of IPU into your network

The Napatech F2070X is the perfect solution for network, storage and security offload and acceleration, enabling virtualized, containerized or bare-metal deployments with tenant isolation.

The 2x100 gigabit Ethernet PCIe card is powered by an Intel Agilex® AGFC023 FPGA and an Intel® Xeon® D processor. The unique combination of FPGA and general-purpose Xeon CPU on a PCI card allows for unique offload capabilities.

Explore the benefits at napatech.com



Unleash the power of distributed AI

Seamless, scalable, unified
connectivity with ACE-AI



Network Different™

arrcus.com

Data Center Networking for AI and Cloud

Accelerating Today's and Tomorrow's Workloads

Introduction – What's Changed in Data Center Networking

In the 2022 edition of our “SmartNICs and Infrastructure Acceleration” report, we examined the changing workloads in data centers and trends in silicon and systems architecture that were driving networking infrastructure. We discussed software networking libraries and hardware accelerators' role in speeding up workloads. And we took stock of the ecosystem's leading network acceleration (SmartNIC) players.

GenAI Arrives with a Bang

2023 has seen a shift in data center workloads driven by generative AI (GenAI). ChatGPT from OpenAI swept the world, hitting 100M monthly active users in January 2023, barely two months after launch¹, and has continued to gain traction.

Post-ChatGPT's launch, analyst firm Gartner forecast that more than 80% of enterprises will use generative APIs or deploy GenAI-enabled applications by 2026². Marc Ganzi, CEO of DigitalBridge (a digital infrastructure investment company with Switch, DataBank, and Vantage data centers in its portfolio), has projected that the GenAI opportunity could reach 38 gigawatts (GW) in data center power consumption soon, dwarfing today's 13 GW public cloud capacity³. Consultancy McKinsey corroborates Ganzi's estimates, predicting that total data center power consumption in the US will rise to 35 GW by 2030, up from 17 GW in 2022⁴ (the US currently accounts for about 40% of the global data center market)

Whether GenAI lives up to its hype and becomes the operating system for all business and consumer software, it will likely be a primary consumer of data center resources over the next five to ten years. GenAI is a CEO and board-level concern, and we, therefore, pursue two tracks in this report. We will examine the impact on data center networking by (1) AI/ML/GenAI workloads and (2) other cloud workloads.

Enterprise and Consumer Cloud Workloads Persist

Even as GenAI dominates the headlines, businesses, governments, and consumers continue progressing on their digitization journeys. Non-GenAI workloads and cloud adoption drive substantial computing, networking, and storage use. Mobile and internet-of-things (IoT) devices are growing, and broadband speeds are increasing. Global 5G subscriptions will reach 1.6B by the end of 2023, and mobile data traffic consumption per smartphone will increase from 21GB in 2023 to 56 GB per month at the end of 2029⁵.

¹ “ChatGPT sets record for fastest-growing user base – analyst note,” Reuters, February 2, 2023. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>

² “Gartner says more than 80% of enterprises will have used generative AI APIs or deployed generative AI-enabled applications by 2026,” Gartner, 2023. <https://www.gartner.com/en/newsroom/press-releases/2023-10-11-gartner-says-more-than-80-percent-of-enterprises-will-have-used-generative-ai-apis-or-deployed-generative-ai-enabled-applications-by-2026>

³ “Generative AI is a 38GW data center opportunity, says DigitalBridge CEO Marc Ganzi,” Data Center Dynamics, August 9, 2023. <https://www.datacenterdynamics.com/en/news/ai-a-38gw-data-center-opportunity-digitalbridge-ceo-marc-ganzi-believes/>

⁴ “Investing in the rising data center economy,” McKinsey, January 17, 2023. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/investing-in-the-rising-data-center-economy>

⁵ “Ericsson Mobility Report,” Ericsson.com, 2023. <https://www.ericsson.com/en/reports-and-papers/mobility-report>

Video consumption is likewise growing, with short-form videos taking up a good part of the bandwidth mix. At the end of 2023, video traffic is estimated to account for 73 percent of all mobile data traffic. Virtual reality (VR) and augmented reality (AR) have been slower on the uptake and will likely drive edge data center workloads shortly. In the meantime, independent of AR/VR, increasingly interactive real-time retail, sports, and entertainment applications require fast responses from edge and cloud data centers today.

These workloads (especially AI) are driving changes in data center and computing architecture. Data center networking is evolving to accommodate these new applications – supporting real-time needs, enabling multi-cloud flexibility, improving security, establishing reliability, scaling throughput, and lowering latencies.

This year’s AvidThink research brief updates our 2022 report. We’ll continue to examine the evolution of network accelerators, whether termed SmartNICs, IPUs, DPUs, or SuperNICs⁶! We will touch on other key data center networking topics, including automation, visibility, troubleshooting, and security.


For executives and technologists at communication service providers (CSPs) and enterprises, we hope this helps you understand the changing landscape in data center networking and supports you in making informed decisions. As always, we welcome reader feedback at research@avidthink.com.

The Workloads Driving Data Center Architectures


Let’s start with examining leading use cases and workloads that drive data center systems and networking architectures. Since our last report in early 2022, many of the drivers remain, but there’s an exceptional new arrival: GenAI.

⁶ SmartNICs - smart network interface cards, IPU - infrastructure processing unit, DPU - data processing unit, SuperNIC - self-explanatory marketing term


WORKLOADS DRIVING DATA CENTER ARCHITECTURES




AI/ML/Generative AI




Digital Transformation




Next-Gen Communications



Interactive/RT Edge Workloads



Cybersecurity



avidthink.com

AI/Machine Learning and GenAI

Industry pundits, top management consultants, and business and political leaders proclaim the varying impact of GenAI on our global economy. With limited data from premature studies, they predict that we'll see a global GDP uplift ranging from multiple tens of billions to trillions over the next decade. Putting aside concerns about artificial general intelligence (AGI) and super-intelligence, or any doomsday theories, GenAI as a technology capability has been rapidly adopted by all classes of enterprise software. From CRM and sales-force automation to office suites and software development, integrated GenAI capabilities are unlocking new levels of productivity and efficiency. New GenAI technology for generating video, images, and other forms of content is fueling the development of new creative applications.

Unfortunately, GenAI consumes significant computing capacity and power, potentially detrimental to sustainability efforts. Cloud providers and co-location companies are entering into power leases for huge volumes of electricity. The increasing power demand has made data center leases for up to 100 Megawatts (MW) prevalent. Racks within data centers that used to house 10-12 KW are now pushing 30-50 KW for AI workloads, with potential peaks of 100 KW⁷. This trend underlines a scramble for power as operators and hyperscalers pursue availability in regions offering ample land for construction near cheap power sources with the needed network bandwidth.

Hyperscalers and other computing service providers are scrounging supply chains for GPUs. Demand currently outstrips supply⁸. Whether training a new GenAI foundation model (FM), fine-tuning pre-trained FMs, or employing related techniques like retrieval augmented generation (RAG), GPUs⁹ are needed. Techniques like pruning and quantization can enable CPUs to perform fast enough inferencing, but GPUs will dominate training and inferencing for the foreseeable future. This means that GenAI clusters for training and inferencing will be a common sight across hyperscaler, enterprise, CSP, and edge data centers (for low-latency inferencing workloads).

Data center networks will need to accommodate these new high-performance computing-type (HPC) workloads, providing the necessary capacity, latency, reliability, and performance to support training and inferencing data transfer patterns. We'll come back to GenAI shortly. Let's quickly run through other cloud workloads, many of which we've discussed in past reports, but we'll reaffirm their continued importance.

Enterprise Digital and Cloud Transformation

GenAI aside, digital and cloud transformation continues at enterprises. Many organizations want to mimic what digital pioneers like Uber, Lyft, Amazon, and CapitalOne have achieved in transforming traditional business models into digital powerhouses. Today, SaaS, PaaS, and other IaaS hosting traditional VM-based enterprise applications form a large part of data center workloads.

In parallel with cloud adoption, we're noticing more enterprises taking a multi-cloud/hybrid cloud approach to their cloud transformation. Whether deliberate or accidental (organic), many large enterprises are multi-cloud – with different applications running on multiple public and private clouds. Increasingly, enterprises will need to establish data transfer across clouds for analytics or to allow collaboration between applications and cross-API calls. This drives the need for secure, flexible cross-cloud, cross-cluster networking and communication. Public clouds' direct or express connection solutions, private backbones, transit gateways, and virtual secure overlay networks are designed to help with this need.

Meanwhile, the increasing digitization of business processes and accelerated information gathering from physical, real-world activity (manufacturing, transportation, medical care, building management) drives increased data volume. The use of digital

⁷"The Silicon Valley Co-Los Know What's Really Going On With AI - The Next Platform," The Next Platform, May 10, 2023. <https://www.nextplatform.com/2023/05/10/the-silicon-valley-co-los-know-whats-really-going-on-with-ai/>

⁸B. Fung, "The big bottleneck for AI: a shortage of powerful chips," CNN, August 6, 2023. <https://www.cnn.com/2023/08/06/tech/ai-chips-supply-chain/index.html>

⁹We'll use the term GPU as inclusive of other AI acceleration silicon like Google's Tensor Processing Units (TPU), AWS Inferentia2 and Trainium2 chips, Microsoft Maia 100, Cerebras Wafer-Scale Engine (WSE), and other AI acceleration silicon.

twins in IoT and IIoT and advanced data analytics using AI/ML will result in escalating network traffic and growth in processing demands, memory needs, and storage requirements.

The next frontier for enterprises includes Meta's metaverse and Nvidia's omniverse. Integrating AR/VR into the unified 'X-verse' concept will drive more network traffic and demand dramatically, markedly lowering latencies and increasing reliability.

Next-Gen Communications: 5G, Fiber, and Network Innovations

CSPs are using the transition to 5G to drive their digitization. From back office and IT estates to customer-facing portals and applications, CSPs have virtualized and cloudified their workloads, embracing virtual machines and cloud-native applications across hybrid clouds (private data centers and public clouds).

The modernization of CSP networks has lagged, but we see that the processing of communications traffic is increasingly handled less by proprietary hardware and more by data center-hosted general-purpose computing as driven by key initiatives:

- **Software-defined networking (SDN) and Network Virtualization (NV):** Abstracting networks from physical connections offers agility and efficiency to businesses and telecom operators. Using software to centralize control and create programmable networks allows for flexibility, resiliency, and innovation.
- **Network Functions Virtualization (NFV) and Cloud-Native Network Functions (CNF):** Transitioning from physical hardware to software solutions on standard servers, enhancing flexibility and standardization. 5G's services-based architecture (SBA) promotes a cloud-native framework, enabling core telecom functions to run in standard cloud data centers.
- **Network Disaggregation:** Initiatives like open RAN (and open optical) promote disaggregation in 5G and high-speed networks. Network functions like routing on open and disaggregated servers bring performance-sensitive workloads to carrier on-premises locations, edge data centers, and regional data centers.

The goal is to manage increasing network traffic efficiently and cost-effectively by leveraging standard data center servers to run telco workloads. Concurrently, telecom operators are looking to provide enterprises with improved end-to-end SLAs and private slices of public networks for enterprise workloads – together with network slicing options across their 5G networks.

The goal is to manage increasing network traffic efficiently and cost-effectively by leveraging standard data center servers to run telco workloads.

Interactive and Performance-Sensitive Edge Workloads

Edge computing, crucial in the 5G era, supports applications requiring low latency and high-speed processing like the latency-sensitive 5G user plane function (UPF). The edge is also essential for multi-user gaming, content distribution, IoT, near-real-time analytics for industrial applications, and AI/ML applications like computer vision and video surveillance. The edge may be a metro-area or local data center on-premises to accommodate performance or data privacy requirements. Challenges at the edge continue to be power, space, cooling management, and physical security, especially in remote edge locations.

Cybersecurity: A Growing Imperative

As digital and physical infrastructures intertwine, cybersecurity becomes more critical. Cybersecurity will represent an increasing workload in metro and edge data centers, as security services use data center servers to process and inspect traffic, often decrypting and re-encrypting as needed to perform deeper content inspection.

Simultaneously, cybersecurity is a concern in data center networking. Ensuring data safety across networks involves encryption, monitoring, and defense against attacks. The demand for security-related processing is increasing, adding to the computational overhead for data transfers. Over the last year, an increasing number of data center and wide-area networking vendors have touted their adherence to zero-trust principles and increased use of telemetry and AI/ML to detect anomalous access and network traffic.

Application Architecture Evolution

Data center operators, including the hyperscalers, have recognized the need to support different classes of applications, hence the large families of machine instance types. AWS, Microsoft Azure, and Google Cloud Platform (GCP) provide variations that support CPU-intensive, memory bandwidth, and capacity-intensive, or I/O-intensive workloads, as well as different storage speed and capacities tiers. On top of these instance types run today's cloud-based applications, ranging from traditional VM-hosted lift-and-shift enterprise applications to modern cloud-native web-facing applications.

Microservices: The New Norm in App Development

Examining non-GenAI/AI/ML apps, we see software architecture shifting from bulky, monolithic structures to modular, distributed multi-tier systems. These applications are hosted in centralized data centers, often in containers on Kubernetes clusters, and interface with users through sleek web-based UIs or mobile apps via RESTful APIs. This caters to the demand for rapid software development, enhancing team productivity, scaling user capacity, better resilience, and cost efficiency.

At the same time, portions of the applications — mostly static data and video content are hosted on edge data centers within content-delivery networks (CDNs). Increasingly, to improve interactivity and user experience, even dynamic content is being rendered in these edge data centers, driving more computation versus storage.

Simultaneously, a shift towards serverless computing (somewhat a misnomer as servers haven't gone away) is gaining momentum. This approach, focusing on function-based programming, offers unparalleled portability, scalability, and efficient distribution of application logic. The rise in serverless was clear at AWS re:Invent 2023, where the hyperscaler announced a series of new capabilities focused on serverless across caching, database, and data warehouse services.

Cloud Application Traffic Patterns: North-South to East-West

The evolution of multi-tiered cloud-hosted applications has altered traffic patterns. Historically, the North-South (user-to-server) traffic dominated, but now, East-West (inter-component) communication prevails in data centers, driven by microservices architecture. While estimates vary, E-W traffic outweighs N-S, with ratios like 85/15 or 80/20 being common.

This change has major implications: where N-S traffic was managed by a series of network appliances for various functions, E-W traffic, amplified by microservices, demands advanced acceleration, scaling, and security. Modern solutions include service meshes or similar proxies that manage service discovery, load distribution, and security protocols like encryption and mutual authentication.

You Get a GenAI Cluster, Everyone Gets a GenAI Cluster

In parallel with cloud workloads, most data centers will experience substantial Gen AI demand. HPC clusters from hyperscalers and other cloud providers will rapidly expand and evolve to accommodate unique GenAI training, fine-tuning, and inferencing needs. HPC has been the realm of a limited number of specialized research-focused organizations, but with GenAI, enterprises, and even telecommunications providers could desire access to high-performance clusters.

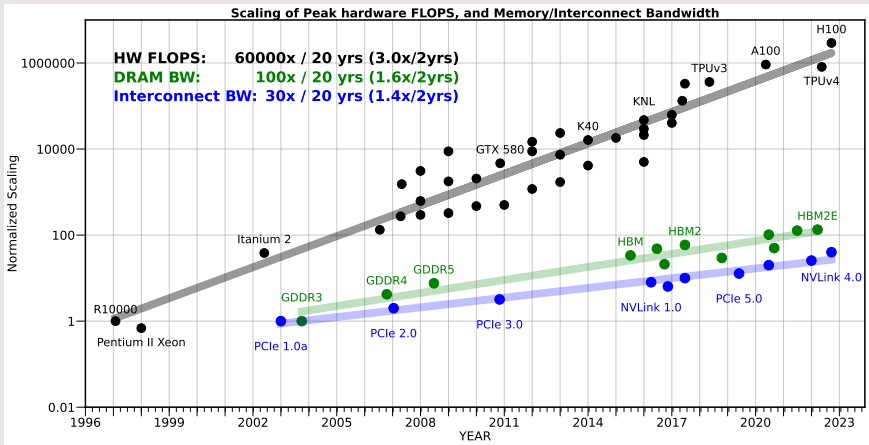
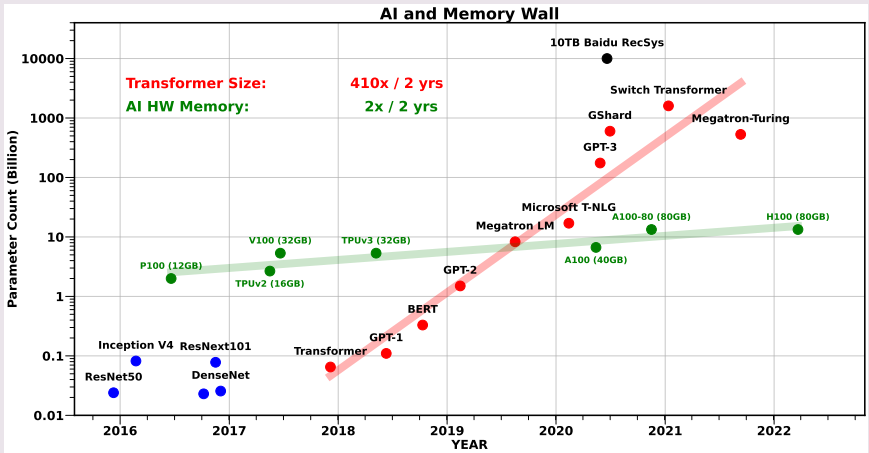
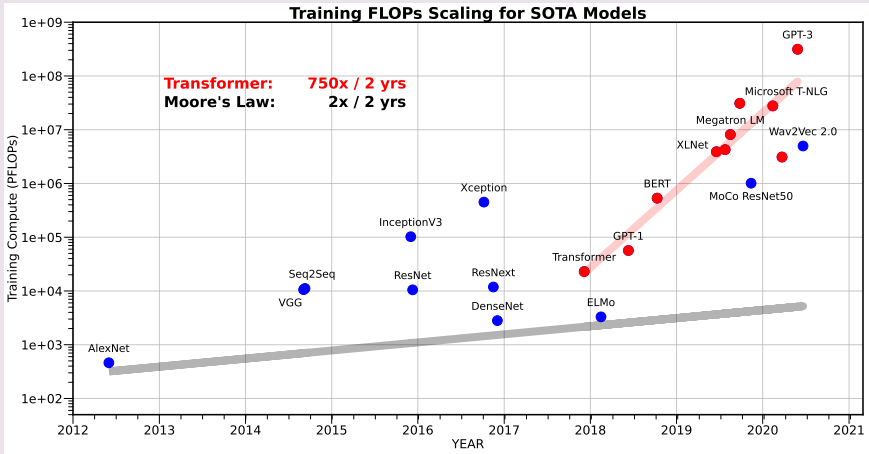
GenAI: Revisiting HPC

High-performance computing (HPC) has been advancing in parallel across custom clusters in research centers worldwide. Public cloud providers have invested in "rentable" HPC clusters providing reservable GPU/CPU clusters with high-bandwidth interconnects (both Ethernet and Infiniband-based).

Cloud providers are launching services focused on GenAI training, fine-tuning, and inference workloads, and AI-specialist clouds are sprouting up. Nvidia is building its own AI cloud while partnering with hyperscalers. It's also funding and enabling new providers like CoreWeave by allocating GPUs in short supply to them. Regardless, all these cloud providers will face significant challenges in building computing architecture for training (and other GenAI operations). The following series of trend plots from researchers at UC Berkeley's RiseLab¹⁰ helps make the point.

¹⁰Gholami A, Yao Z, Kim S, Mahoney MW, Keutzer K. AI and Memory Wall. RiseLab Medium Blog Post, University of California Berkeley, 2021, March 29, 2021 https://github.com/amirgholami/ai_and_memory_wall

GENAI MODELS GROWTH RATES EXCEED COMPUTE, MEMORY, INTERCONNECT GROWTH



Source: UC Berkeley RiseLab



avidthink.com

The scaling of model sizes is growing faster than Moore's law (or CPU scaling), likewise with GPU memory capacity and interconnect speeds. Most FMs are bigger than will fit on the memory for a single GPU node.

Even as silicon and systems manufacturers work on **scaling up** each GPU node using proprietary high-speed interconnect like Nvidia's NVLink, the reality is that any useful training or inferencing GPU clusters on the order of thousands or tens of thousands of GPUs will need to **scale out** using a high-speed network fabric to connect multi-GPU nodes.

As an example of the scale of traffic and sensitivity to network performance, Broadcom and Chinese hyperscaler Tencent shared learnings from training FMs with more than 100B parameters at the Open Compute Project (OCP) 2023 Summit¹¹. They saw 10TB+ data transfers per model training iteration and saw that waiting for network communication could idle the GPUs for up to 50% of job completion times in the worst case. And 0.1% of packet loss in their RDMA network could cost up to 50% effective computation power loss. Given that the largest models are many 100Bs of parameters and GPT-4 is estimated at past a trillion parameters (who knows how large Google's multi-modal Gemini is), it's clear that the network will have to scale up, out, and improve in performance (reliability, latency) to support ongoing FM training.

GenAI training as an application is a more extreme case of what we saw in the past with big data analytics (remember Hadoop and MapReduce operations?). Big data analytics spread data and computation over multiple nodes, as is much of GenAI/ML training today. Parallelism techniques allow large models to be spread over multiple nodes of limited hardware and speed up training time. Whether the training uses data, tensor, pipeline parallelism, or more complex sharding¹², synchronization results in bursts of E-W traffic across multiple nodes. Common training operations like Broadcast, AllReduce, AllGather, and ReduceScatter (as part of multi-GPU and multi-node communication libraries like Nvidia's NCCL that power distributed training) could result in traffic bursts.

While fine-tuning and inferencing demands on the network will be less, data center networks will need to successfully support the training cycle's unique and extreme traffic patterns.

Silicon Evolution: Navigating the New Era of Computing

In our next section, we'll dig into the implications of AI and cloud workloads on data center networking. Before we do that, we'll take a slight digression to touch on a topic we've covered in detail in **last year's report** but we will do a quick run-through to catch new readers up.

SmartNICs/DPUs/IPUs are increasingly popular in cloud environments – every hyperscaler has its variation on the theme. Domain-specific architectures are gaining momentum, with GPUs being one example and DPUs/IPUs another. Specialization is not just playing out on silicon substrates but extends to system and data center architectures, driving multi-core and more parallel systems, domain-specific architecture (DSA), and large-scale distributed computing and storage systems.

Understanding key principles and guiding observations in silicon evolution will help explain why DSAs are gaining momentum:

- **Moore's Law (circa 1965):** Gordon Moore's observation of the doubling of transistor counts in integrated circuits approximately every two years has been a cornerstone in projecting semiconductor density and performance advancements.
- **Amdahl's Principle/Law (circa 1967):** Gene Amdahl's principle states that the performance improvement of optimizing a single system component is limited by the fraction of time the improved component is utilized. It highlights the theoretical limits of performance gains based on workload parallelization.
- **Dennard Scaling Principle (circa 1974):** Articulated by Robert H. Dennard and his team, this principle posits that as transistors get smaller, their power density remains constant, allowing higher switching speeds without increased power consumption. This scaling initially played a crucial role in enabling higher clock frequencies in chip designs.

¹¹ Broadcom and Tencent, "Telemetry based load balancing of AI/ML workloads," OCP Summit 2023. <https://www.opencompute.org/events/past-events/2023-ocp-global-summit>

¹² "Model Parallelism," Huggingface.co, 2020. <https://huggingface.co/transformers/v4.9.2/parallelism.html>

The breakdown of Dennard Scaling around the mid-2000s due to issues like power leakage and thermal effects has been a pivotal point in chip design. This limitation led to a stagnation in increasing CPU clock frequencies, signaling the end of consistent performance gains from traditional scaling methods.

This shift led to a strategic pivot towards multi-core processors. However, these systems' performance improvements are inherently limited by Amdahl's Law, which governs the extent to which parallelization can enhance processing speed. The industry has, therefore, responded by taking a multi-pronged approach with:

- **Heterogeneous Multi-Core Designs:** Moving beyond traditional CPU-centric designs to incorporate a mix of processor types, each optimized for specific tasks (e.g., performance versus efficiency cores in Apple Silicon).
- **Hardware Customization and Specialization:** Developing specialized hardware solutions tailored for specific applications like high-performance computing (HPC), AI/ML, and graphics.
- **Innovations in Energy Efficiency:** Advancing techniques in power management to maximize performance within thermal and power constraints. This is coupled with cooling techniques like liquid cooling in data centers.
- **Optimized Data Movement and Interconnect Technologies:** Enhancing data transfer efficiency within and between chips crucial for high-performance computing systems. Increasing use of optical interconnects side-by-side with electrical.
- **Co-Evolution of Software and Hardware:** Strengthening the interplay between software and hardware to maximize the efficiency and effectiveness.

As computing demands evolve, driven by new applications and workloads, the industry's approach to silicon design and system architecture continues to adapt. These new diverse and heterogeneous components will make up the new building blocks in today's and tomorrow's data centers.

The Shifting Network Architectures of Data Centers

Now that we've covered the evolution of workloads, application architectures, and silicon architecture, let's look at the data center and the underlying network infrastructure.

The unprecedented uptake of GenAI has shifted both data center rack design and power budgets. Even cooling strategies are driven by GenAI workloads. For instance, liquid cooling, which has been on the cusp of the mainstream, could finally see wider adoption. Industry legend Andy Bechtolsheim chose to talk about heat management and cooling strategies in his keynote at the recent OCP Global Summit 2023. And Silicon Valley-based co-location facility Colovore touts their AI-ready data center with liquid cooling¹³. GenAI is likewise pushing network architecture changes for inter-node communication in GPU clusters.

Different workload requirements for power, cooling, networking performance, and computing and memory intensity across AI and cloud applications will drive the separation of networked computing and storage resources.

Bifurcation of the Data Center

In these early days of GenAI, we are seeing a bifurcation of data center architectures with parts of (and perhaps eventually, entire) data centers dedicated to GenAI and HPC workloads and the remainder dedicated to cloud workloads.

Different workload requirements for power, cooling, networking performance, and computing and memory intensity across AI and cloud applications will drive the separation of networked computing and storage resources.

¹³R. Miller, "Colovore Building New Data Center for Extreme Density AI Workloads," Datacenterfrontier.com, November 3, 2022. <https://www.datacenterfrontier.com/cooling/article/21437353/colovore-building-new-data-center-for-extreme-density-ai-workloads>

The requirements for non-AI/GenAI workloads are different from enterprise applications:

| Enterprise Cloud Workloads | AI/ML/GenAI Workloads |
|--|---|
| Loose coupling, unsynchronized | Tight coupling, synchronized |
| Unpredictable, could be bursty | Quasi-periodic bursts, somewhat predictable (training) |
| Varying flow types based on app | Mix of large flows (gradient, weight exchange) with small orchestration flows |
| Small to medium volumes of data exchanged | Large volumes of data exchanged |
| Mostly latency and loss insensitive | Highly-latency and loss sensitive |
| Jitter tolerant | Jitter intolerant |
| Small-scale clusters of tens or hundreds servers or loosely collaborating services | Large-scale clusters - up to 10Ks of GPU nodes |
| <100Gbps of E-W BW | > 400Gbps of E-W BW |

The above is a coarse generalization of cloud workloads. There are non-GenAI workloads like media streaming that will require high bandwidth, mobile 5G network functions that require low latency, and gaming applications that prefer low jitter.

Nevertheless, what's clear is that the AI/GenAI section of the data center will require a different network – one focused on driving high throughput with low latency.

The AI Section of the Data Center

To envision on a large scale what a GenAI-specialized cluster looks like, we only have to look at Nvidia's DGX GH200 design or Microsoft's Open AI cluster used for training Open AI's GPT models. These clusters look similar to past HPC clusters, except they are larger, denser, and more power-hungry.

At AWS re:Invent 2023, the CEOs of AWS (Adam Selipsky) and Nvidia (Jensen Huang) announced Project Ceiba¹⁴, a 16,384 GPU 65-Exaflop cluster powered by Nvidia's Grace Hopper GH200¹⁵ super chip (combining a CPU and GPU) connected via Nvidia's NVLink (GH200 NVL32 connects 32 GH200s via a high-speed switch) inside each node, with AWS EFA (Elastic Fabric Adapter) enabled by AWS Nitro connecting across nodes.

For AI clusters, aside from the intra-node/intra-server connectivity between GPUs that is likely going to be proprietary for now (possibly migrating to standards-based CXL in the future), the question of how each GPU node talks to other GPU nodes (E-W) and how the cluster interacts with the outside world (or rest of data center) (more N-S) needs answers. Some of the reference architectures we saw at OCP Summit 2023 and what Nvidia details in its white paper "**Next-Generation Networking for the Next Wave of AI**" recommends high-speed 400Gbps (RoCE) using their BlueField-3 SuperNICs for backend E-W links and 200Gbps BlueField-3 DPU powered frontend N-S links.

Whether frontend/backend networks are necessary for AI training will be sorted out in the near future. Other architectures use a single high-speed fabric to interconnect all the GPUs without delineating two distinct networks. Google takes a dramatically different approach, using an optical circuit switch (OCS) to interconnect their TPUs as part of their training cluster¹⁶, of which each rack is 64 4x4x4 cubes, connected with 48 OCS¹⁷. That shows the potential diversity of approaches as vendors and hyperscalers innovate to train larger and larger models faster.

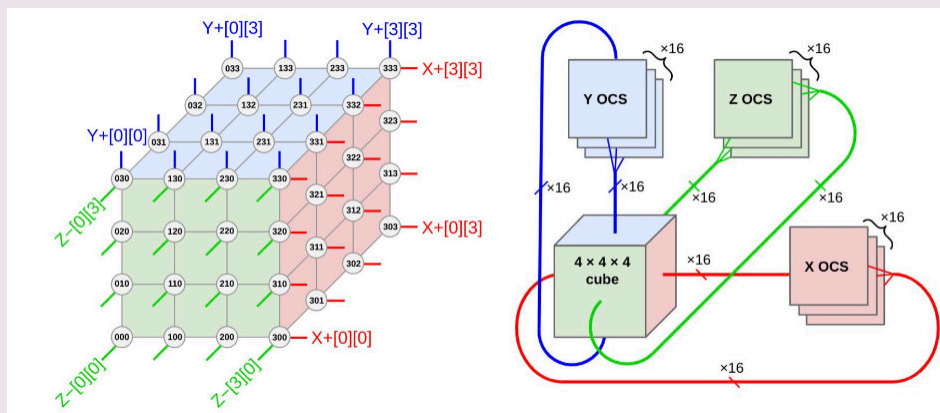
¹⁴ "AWS and NVIDIA Announce Strategic Collaboration to Offer New Supercomputing Infrastructure, Software and Services for Generative AI," NVIDIA, 2023. <https://nvidianews.nvidia.com/news/aws-nvidia-strategic-collaboration-for-generative-ai>

¹⁵ Details on the DGX GH200 GPU are available from Nvidia's site. <https://www.nvidia.com/en-us/data-center/dgx-gh200/>

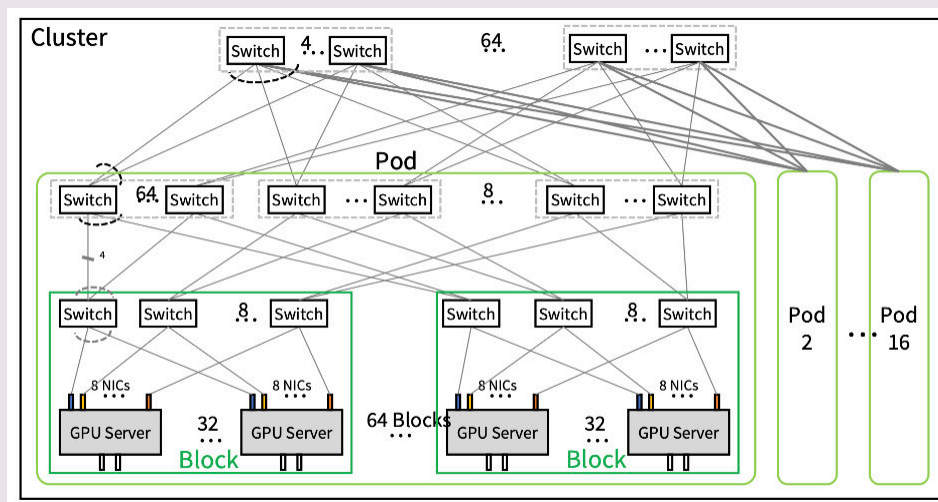
¹⁶ L. Poutievski et al., "Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Defined Networking," Google Research, 2022. <https://research.google/pubs/pub51587/>

¹⁷ Norman Jouppi & Andy Swing, Google, "A Machine Learning Supercomputer With An Optically Reconfigurable Interconnect and Embeddings Support", HotChips 2023. https://hc2023.hotchips.org/assets/program/conference/day2/ML%20training/HC2023.Session5.ML_Training.Google.Norm_Jouppi.Andy_Swing.Final_2023-08-25.pdf

EXAMPLE AI NETWORK TOPOLOGIES



Source: Google (HotChips 2023 Presentation)



Source: Broadcom, Tencent "Telemetry Based Load Balancing of AI Workloads" (OCP 2023)



avidthink.com

Esoteric architectures aside, many high-speed networks for AI training are typically RDMA over 400G/800G Infiniband or lossless Ethernet (RoCE or RoCEv2) from the leading networking vendors, built using standard merchant silicon.

Looking beyond RoCEv2, many of the companies involved in AI are signing onto the Ultra Ethernet Consortium (a growing list that includes AMD, Arista, Broadcom, Cisco, HPE, Intel, Meta, Microsoft, and Oracle) to advance new capabilities for Ethernet, improving throughput and lowering latency. As **compute express link (CXL)** matures and shows value as a generic switching interconnect that links composable components like CPUs, GPUs, memory, and peripherals, it could be used as an alternative to proprietary technologies like NVLink switches.

At the same time, interconnects are moving from copper to fiber and from electrical to optical for AI workloads. Proposals for using optical to connect computing components were presented at OCP Global Summit 2023, and in-server optical interconnects could be used to meet increasing performance requirements. Ongoing advancements in optical and silicon photonics to drive higher speeds like 400 ZR, 800 ZR, and 1.6T, or co-packaged optics (CPO) to address BW and power challenges will be needed to support increasing AI workloads.

As this transition happens, system vendors and data center operators will be balancing interconnect properties like max distance (managing bit error rates), throughput, power consumption (pJ/bit), latency, total cost, and even cable weight.

Techniques are being developed and refined at a system and software level. Advanced buffer and queue management, congestion avoidance and handling techniques like end-to-end data center quantized congestion notification (DCQCN) in RoCEv2, head-of-line blocking avoidance, improved hashing for ECMP to mitigate low entropy headers, end-to-end scheduling, virtual output queues, and adaptive load balancing are part of the laundry list of efforts being advanced across networking vendors. In-band telemetry and the use of AI/ML to manage data flows better are also part of this equation.

All the major networking vendors, Cisco, Juniper, Nvidia (yes, they are also a networking vendor with the prior acquisition of Mellanox and Cumulus Networks), Arista, and merchant silicon providers like Broadcom and Marvell are touting capabilities in these areas. Independent network operating system vendors like Arrcus (sponsor of the report) are likewise advertising their AI capabilities in the form of **ACE-AI**.

Unlike cloud workloads, AI workloads will be accelerated mostly through higher throughput and lower latency links, higher radix switches with more ports for flatter networks, and better and smarter traffic engineering and management, all to facilitate getting the model weights in the data packets into memory for the GPUs to get to work on as quickly as possible.

The Cloud Section of the Data Center

On the cloud side, depending on the workload, the networking infrastructure can potentially add more value in co-processing and pre-processing data packets flowing to and from the CPUs. Further, what used to be the networking infrastructure is turning into more of an orchestration and supervisory system, especially within the hyperscale data centers.

As we've shared before, a Google study from 2015 of 20K machines over three years showed that 30% of processing cycles are taken up by non-application tasks "datacenter tax", all ripe for acceleration.

Software Acceleration – Still a Thing

Not all acceleration needs to be in hardware. The software acceleration techniques we touched on in previous reports continue to be used and deployed. Extended Berkeley Packet Filter (eBPF) that enables applications to compile and run packet handling code safely and efficiently within the kernel has seen significant interest, as evidenced by the many companies supporting it at KubeCon 2023¹⁸ for acceleration, security, and visibility. Data Plane Development Kit (DPDK) and Vector Packet Processing (VPP) are likewise being deployed across many network-centric workloads and are a key part of many telco workloads for wireline and 5G deployments.

Multi-Cloud – The Need Exists

Most enterprise application components reside in a single cloud, even though the organization might use multiple public or private clouds for different use cases. However, for centralized analytics or when applications need to invoke APIs and services in other clouds, multi-cloud networking capabilities can be helpful. There are different strategies to managing this, ranging from using multiple direct private connections from co-location facilities with virtual routers to allow cross-cloud connections, software-powered middle-mile network vendors like Graphiant, to enabling multicloud fabrics using vendors like Arrcus with their FlexMCN, or products from Alkira and others (also covered in our **Enterprise Edge and Cloud Networking report**).

Enabling End-to-End Quality of Service

Highly interactive near real-time and real-time applications are gaining traction. As application developers try to improve customer experiences, as enterprise latency-sensitive workloads come online, enabling end-to-end SLAs becomes more important.

¹⁸"Industry Voices: Kubernetes reigns — Observations from KubeCon," Silverlinings, Nov. 14, 2023, <https://www.silverliningsinfo.com/apps-services/industry-voices-kubernetes-reigns-observations-kubecon>

Even for AI/ML, providing end-to-end networks with QoS for real-time inferencing (video surveillance, IoT, and industrial workloads) can unlock digital transformation across businesses. Here, SRv6 and other means for improving traffic management and routing will play a role — enabling the distribution of trained machine models to the edge or providing network paths with strict QoS from customer premises to metro edge data centers across both 5G (with network slicing enabled) and wireline last mile and middle mile networks.

Vendors like Arrcus, who have been working on providing a scalable and distributed network fabric using SRv6, are demonstrating the value of their ACE-AI solution for both distribution of trained ML models (including GenAI models) to the edge.

SONiC — Gaining Traction

At the recent OCP Global Summit 2023, multiple switch vendors touted their **SONiC** compatibility. There were presentations on the use of SONiC in AI fabrics, supporting RoCE with advanced telemetry. While not technically part of the data center network acceleration, SONiC is an important trend in the evolution of open data center switching. As a network operating system that supports multiple switches (and switch ASICs), it's been adopted by organizations like Alibaba, Tencent, Microsoft Azure, Comcast, eBay, and Target to lower the cost of operations and increase automation. There are at least two startups, Aviz Networks, and Hedgehog, that are stepping in to provide enterprise support with SONiC. SONiC continues to look promising and remains on our watch-closely list.

Evolution of Hardware Network Acceleration

The field of SmartNICs, DPUs, IPUs, and now SuperNICs (Nvidia's new term) can sometimes be messy. In general, SmartNICs offload processing from the main CPU, performing tasks like encryption and TCP/IP processing. They may incorporate various technologies like DPUs, FPGAs, ASICs, and GPUs and are adaptable for use by hypervisors, OS, virtual machines, and containers.

Typically, the key components of a SmartNIC include:

- High-speed NIC for network connection.
- Packet acceleration logic with technologies like DPUs and FPGAs.
- Additional logic for functions like crypto and compression.
- Embedded CPUs for control and management tasks.
- Memory controllers and high-bandwidth memory.
- Comprehensive software stacks.
- Secure subsystems, including secure boot and root-of-trust.

These technologies are evolving rapidly, with specialized form factors emerging, such as Dell's Open RAN Accelerator Card with Marvell and HPE's Qualcomm-powered card for 5G workloads.

Acceleration hardware building blocks

- **System-on-Chips (SoCs)** — SoCs, usually based on Arm ISA, integrate multiple CPU cores with hardware accelerators for various functions like encryption and compression.
- **FPGAs** — FPGAs, used for custom logic in fields where speed is critical, offer reconfigurable hardware acceleration but at a higher cost and power consumption compared to ASICs.
- **GPUs** — Originally for computer graphics, GPUs' multiple core architectures and ability to do fast matrix and vector operations make them useful for parallelizable tasks like AI/ML and some networking tasks.
- **ASICs** — ASICs offer a cost-effective, power-efficient solution for high-volume, stable designs, including analog circuits like transceivers alongside CPU cores.
- **DPUs** — DPUs, or Data Processing Units, represent a next-generation approach, processing data more efficiently than CPUs. The architecture varies, sometimes combining CPU cores with custom logic.

SmartNICs, as a market category, have seen mixed success to date. While AMD acquired high-profile Pensando Systems last year for \$1.9B¹⁹, activity in the market for AMD SmartNICs has been quiet thus far. And it's unclear whether Pensando's collaboration with HPE Aruba networking has yielded significant revenue. Early this year, Microsoft announced they had acquired Fungible²⁰, founded by Juniper Network's former co-founder and CEO Pradeep Sindhu. While numbers were not shared, the deal size was rumored to be \$190M, a big step down from the \$300M they raised.

Nonetheless, SmartNICs and DPUs, as part of hyperscaler cloud networking, have proven pivotal and foundational. As key elements in mobile 5G core and open RAN initiatives, as well as private 5G networks, SmartNICs/DPUs/IPUs hold potential. Fellow analyst firm Dell'Oro Group estimates that the market will grow at a 42% CAGR and exceed \$5 Billion in 2027²¹.

SmartNICs/SuperNICs/IPUs/DPUs - What's in a Name?

Last year, we had to contend with vendors attempting to differentiate their network acceleration products by attempting to define or redefine terms like DPU. Some viewed DPU as a component of a SmartNIC. Others saw the DPU as an evolved SmartNIC. Yet others invented new terms like IPU. And then this year, Nvidia introduced the SuperNIC, another variant of its BlueField-3-based cards. The Nvidia SuperNIC is smaller and less power-hungry than the BlueField-3 DPU and is optimized for high-bandwidth, low-latency data flows, ideal for AI fabrics.

The variations across all these accelerator cards tend to be along the lines of how programmable these cards are, how many GP-CPU cores are available for additional functions (i.e., running both control and data planes or storage protocol stacks), whether a scaled-down OS or hypervisor can run on the card, and the security capabilities – whether there's a hardware root-of-trust or Secure Enclave. By drawing boundary lines above or below a list of functional blocks, vendors hope to demonstrate differentiation. One effort to unify the definition, architecture, and programming model of a SmartNIC/DPU/IPU is the **Open Programmable Infrastructure** (OPI) project hosted by the Linux Foundation. While still early, if successful, it could simplify the lives of developers looking to take advantage of network accelerators.

OCP Open Domain-Specific Architecture (ODSA) and Chipllets

We first brought up chipllets in last year's report. The OCP Open Domain-Specific Architecture (ODSA) Sub-Project under the OCP Server Project will be increasingly important. Many of the vendors we're speaking with aim to use chipllets to assemble appropriate components for future SoCs. ODSA chipllets, and CXL, are likewise on our watch list.

Network Acceleration Updates - Hyperscaler and Vendors

Regardless of the name, we'll quickly run the latest from the leading hyperscalers and vendors in the network acceleration space. Many of these capabilities are important for cloud workloads, where these accelerators can offload important layer 4-7 functions or specialized functions like open RAN distributed unit (DU) L1 processing. Further, as we observe, an increasing number of vendors are buying into the Nitro-type approach of using the SmartNIC/DPU as a hardware "hypervisor" or resource management, scheduler, and security layer to protect and manage computing, memory, and storage resources in a server.

Amazon Web Services

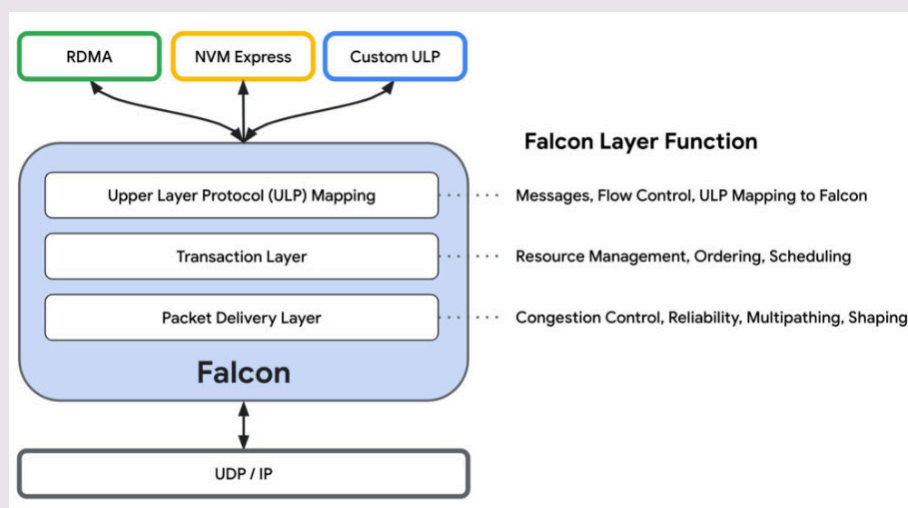
Amazon's Nitro system is the underlying platform for their Elastic Compute Cloud (EC2) web service that offloads networking, storage, and management services from the host servers to dedicated hardware. It's on its fifth generation now, with AWS Nitro v5. Nitro continues to be front and center in AWS's security and performance story (including at AWS re:Invent 2023), even as AWS touts its AI and GenAI muscle. Nitro powers both their Elastic Network Adapter (ENA) and their Elastic Fabric Adapter (EFA), which are used to tie together their large AI clusters (including the new collaboration with Nvidia).

¹⁹"AMD Expands Data Center Solutions Capabilities with Acquisition of Pensando," AMD, May 26, 2022. <https://www.amd.com/en/newsroom/press-releases/2022-5-26-amd-expands-data-center-solutions-capabilities-wit.html>

²⁰K. Wiggers, "Microsoft acquires Fungible, a maker of data processing units, to bolster Azure," TechCrunch, January 9, 2023. <https://techcrunch.com/2023/01/09/microsoft-acquires-fungible-a-maker-of-data-processing-units-to-bolster-azure/>

²¹"Smart NIC Market to Grow at a 42 Percent CAGR Through 2027," Dell'Oro Group, July 28, 2023. <https://www.delloro.com/news/smart-nic-market-to-grow-at-a-42-percent-cagr-through-2027/>

GOOGLE FALCON LOW-LATENCY HARDWARE TRANSPORT



Source: Google (OCP Global Summit 2023)



avidthink.com

Google Cloud Platform

Since our last report, GCP has not shared additional updates about its networking efforts. However, a new big reveal is GCP's release of their hardware transport, Falcon, to the OCP community (announced at OCP Global Summit 2023). We have not heard much otherwise on their Andromeda virtual network stack utilizing Nvidia GPUs and Intel QuickData DMA Engines.

Falcon²² is a reliable and low-latency hardware transport that Google has been using and is now providing to the world. It's a hardware-assisted layer sitting on top of UDP/IP that provides packet delivery, transactions, and maps to upper-layer protocols like RDMA, NVMe, and other custom protocols. It handles congestion control, reliable delivery, multi-pathing, shaping, resource scheduling, etc. Falcon will help with AI/ML workloads and other cloud workloads.

Microsoft Azure

Microsoft's Azure Accelerated Networking and the Microsoft Azure Network Adapter (MANA) are related to the company's efforts to enhance the performance of its cloud services, similar to AWS's Nitro System. MANA is their next-generation network interface for Windows and Linux operating systems and will be available via its Azure Boost service²³ (private preview currently). It's unclear what assets of the recently-acquired Fungible were used for MANA.

The MANA 200Gbps SmartNIC offloads storage data plane operations, accelerating storage performance and improved disk caching. MANA also offers an SR-IOV NIC as a virtual function guest OSes on Hyper-V. MANA is the successor to their previous Azure Accelerated Network (AccelNet) stack, which has been available since 2016.

²² D. Lenoski and Nandita Dukkupati, "Introducing Falcon: a reliable low-latency hardware transport," Google Cloud Blog, October 17, 2023. <https://cloud.google.com/blog/topics/systems/introducing-falcon-a-reliable-low-latency-hardware-transport>

²³ "Introducing Microsoft Azure Boost Preview," Microsoft Tech Community, November 21, 2023. <https://techcommunity.microsoft.com/t5/azure-infrastructure-blog/introducing-microsoft-azure-boost-preview/ba-p/3876742>

Network Acceleration Vendors

Select vendors are discussed in this section, and this is not meant to be an exhaustive list. These vendors were mentioned by network operators in both the cloud and telecommunications space during AvidThink's recent research. This space continues to include several vendors who focus on developing unique intellectual property, sometimes for niche applications. In the interest of brevity, we will not cover all those players. If you believe there's a vendor that we've not listed below but that we should have feel free to reach out to us.

AMD (Pensando, Solarflare, and Xilinx)

AMD continues to sell and market Pensando as part of its DPU offerings. Their collaboration with HPE Aruba continues with the Aruba CX 10000 switch using Pensando for 800Gbps distributed firewall for E-W traffic, zero trust segmentation, and telemetry. Meanwhile, the AMD Pensando DSC2-200 is used by cloud providers (i.e., Microsoft Azure and Oracle Cloud) as a SmartNIC offering for workload acceleration.

Likewise, under the XtremeScale brand, the Solarflare products continued to be sold. The X2 SmartNIC Ethernet adapters with Onload kernel bypass technology with DPDK support provide real-time packet and flow information across thousands of virtual NICs.

On the Xilinx front, it continues to offer its T1 and T2 accelerator cards for 5G deployments, accelerating front-haul and L1 workloads. And under the Alveo brand, they offer a family of SmartNICs from U25N to the U55C series, and their U200 to U280 series, all of which include FPGAs, RAM, large lookup tables, and Arm cores as needed for added programmability.

Xilinx's solutions address various solutions, from custom logic in HFT and other financial applications to AI/ML, video analytics, data analytics, networking, and security use cases.

Intel

Intel uses IPU to refer to their newer acceleration SoC and cards, reserving the SmartNICs moniker for their older lines of products.

SmartNICs currently include the N6000-PL, built with Agilex FPGA and targeted at telco workloads, and Silicom FPGA SmartNIC N5010 (developed with partner Silicom).

Intel's IPU family includes the E2000 (Mount Evans), a 200Gbps SoC resulting from co-design efforts with hyperscaler Google. Mount Evans includes 16 Neoverse N1 high-frequency cores from Arm and crypto and compression engines, supports PCIe Gen4 and can be equipped with up to 48GB DRAM.

The F2000X-PL (Oak Springs Canyon) is built using Agilex FPGA and an Xeon-D SoC that provides 2x100GbE (with a 200Gbps crypto block) and supports PCIe Gen4. It has virtual switch acceleration, NVMe over Fabric, and RoCE support for storage acceleration.

Finally, the C5000X-PL IPU is built on their Stratix 10 DX FPGA, with an Xeon-D processor, providing 2x25GbE and networking and storage acceleration platform. Production units are only available from their OEM partners, Silicon and Inventec.

Marvell MARVELL

Marvell first announced their OCTEON 10 DPU platform in 2021. They recently announced two new OCTEON 10 DPUs, CN102 and CN103²⁴. OCTEON 10 CN102 is a lower-priced version designed with 10G SerDes to better suit entry-level requirements, while CN103 contains 56G SerDes for higher throughput.

Containing up to 8 Arm Neoverse N2 cores, OCTEON 10 CN102 and CN103 improve the performance over last year's version while reducing power consumption by up to 50%.

²⁴ "Two New Marvell OCTEON 10 Processors Bring Server-Class Performance to Networking Devices," Marvell.com, December 6, 2023. <https://www.marvell.com/company/newsroom/two-new-marvell-octeon-10-processors-for-networking-devices.htm>

The two new DPUs complement the existing OCTEON 10 CN106 designed for cloud, enterprise, and baseband for 5G wireless networks. As we described in last year's report, the CN106 also has AI/ML cores, VPP hardware acceleration, and up to 24 Arm Neoverse N2 server processor cores. The other existing OCTEON 10 DPU is the OCTEON 10 Fusion CN105, which has fewer server cores but has signal processing capabilities added to support massive MIMO and other 5G RAN deployments. The OCTEON 10 Fusion is meant for inclusion into mobile equipment and L1 acceleration hardware.

Nvidia Networking NVIDIA

Nvidia's ConnectX NIC (the "Smart" moniker has been dropped) leverages built-in acceleration engines for RDMA over Converged Ethernet (RoCE), TLS/IPsec crypto offloads, accelerated switch, and packet processing (ASAP²) for virtual switching/routing, and NVMe over Fabrics storage, in addition to traditional networking offloads. ConnectX-7 is the most recent release in the ConnectX family and supports 400Gbps of throughput.

Their BlueField-3 DPU SmartNIC features 16 64-bit Arm v8 A78 Hercules cores in a single SOC. The DPU offers either 400GbE or NDR 400Gbps Infiniband connectivity. The DPU also features 32 GB onboard DDR5 memory and supports 32 lanes of PCIe Gen 5. BlueField accelerates control and data plane performance and functionality for cloud and edge, such as NVMe SNAP storage virtualization, bare metal networking, or cloud-native and security applications, including analytics, micro-segmentation, and firewalls.

In addition, Nvidia also has the BlueField-3 SuperNIC, providing up to 400Gbps of RoCE connectivity between GPU servers and focusing on AI training workloads.

Napatech napatech

Napatech (report sponsor) delivers integrated hardware/software solutions based on SmartNICs and DPUs/IPUs. Since 2003, their packet capture and security offload solutions have been deployed in applications like cybersecurity, financial systems, telecom infrastructure, data centers, and monitoring.

Napatech's system-level solutions comprise Intel's IPU hardware and Napatech's optimized, standards-compatible software stacks that offload and accelerate specific infrastructure workloads. As an example, Napatech's F2070X IPU, based on Intel's F2000X platform, is packaged with software for storage offload (NVMe-over-TCP), security offload (TLS and TCP) and network offload (Open vSwitch plus other networking protocols). Napatech's integrated hardware-software IPU solution can deliver high-performance networking and storage while improving system security and freeing up the general-purpose host CPU to run applications.

Summary and Recommendations

For cloud workloads, we expect SmartNICs/DPUs/IPUs to continue to be critical in supporting 5G, media, interactive, AR/VR, and security applications. We expect improved content awareness and content-specific acceleration for media, web, and security workloads.

In addition, all the hyperscalers are banking on network acceleration hardware to drive performance and free up CPU cycles for application tasks. Many are using the same hardware as management and security oversight for server resources. We anticipate the same will happen in other cloud and enterprise data centers.

On the AI workload front, we will continue to see improvements in raw throughput, especially in optical networking, higher radix switches, and smarter network congestion management to improve utilization. Likewise, we expect network topology to evolve across regular multi-tier CLOS architectures, separation of frontend N-S and backend E-W fabrics, and innovative new topologies like Google's network with OCS. As models get larger and we learn more about distributed training, the network will inevitably be pushed hard to support increasing demands for faster training speeds. When we revisit this topic in 2024, we're certain that the industry will have innovated significantly.

We hope this research brief has armed you with the necessary information to engage with the vendors and cloud providers. As always, the research team at AvidThink welcomes feedback and further conversation. You can reach us at research@avidthink.com.



AvidThink, LLC
1900 Camden Ave
San Jose, California 95124 USA
avidthink.com

©2023 AvidThink LLC. All Rights Reserved.
This material may not be copied, reproduced, or modified in whole or in part for any purpose except with express written permission from an authorized representative of AvidThink LLC. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgment of the authors of the document and all applicable portions of the copyright notice must be clearly referenced.