CABOT PARTNERS
OPTIMIZING BUSINESS VALUE

Open Compute Project Global Summit 2023: Innovations Show How AI Will Change the Entire Technology Stack

Jean S. Bozman, President,
Cloud Architects Advisors LLC
M. R. Pamidi, Ph.D., Principal Analyst

October 2023

Cabot Partners Group, Inc.
100 Woodcrest Lane, Danbury, CT 06810
www.cabotpartners.com
info@cabotpartners.com

Open Compute Project Global Summit 2023

Innovations Show How AI Will Change the Entire Technology Stack

AI technology is upsetting the applecart for the Cloud and On-Premises infrastructure that supports them because running AI efficiently – and scaling up resources smoothly – will require changing nearly every aspect of the infrastructure "stack" from hardware to software. As analysts who have studied technology trends over decades, we expect that the vendors, Cloud Service Providers (CSPs), and corporate customers worldwide will see changes up and down the technology stack, affecting each level along the way.

The OCP Global Summit 2023 ("Summit") made the dimensions of AI's technical challenge to traditional infrastructure very clear to the 4,400+ professionals attending the Summit in San José, California, October 17-19, 2022. Clearly, there will be near-term changes to their IT landscape of servers, storage, networking switches, and power/cooling management, all of which are being addressed by the Open Computing Project's (OCP's) working groups across dozens of industry-wide technology standards.

These AI-enabling technology changes are expected to affect CSPs (e.g., AWS, Microsoft Azure, Google Cloud, IBM Cloud, Oracle Cloud, Baidu, ByteDance, and others), who often build their own custom-designed systems, leveraging industry-standard components – and assembling these systems to run AI workloads and cloud services on behalf of their customers. However, the move to AI-enablement will affect large IT organizations, their data centers, and their service providers (e.g., CSPs, MSPs, and NSPs).

The extent of the reach of these AI-enabling changes will soon be evident – including new hardware specifications for data-center racks and other equipment; support for high-speed data transfers; fast network interconnects, and CXL-based systems; changes to widely accepted Ethernet standards; the addition of liquid-cooling, immersion cooling and cold-plate technology to manage heat generated by AI systems; and many types of software, including new software tools that promote and support interoperability throughout the infrastructure supporting AI, HPC, and corporate business computing.

History and Context

The OCP was founded in 2011 to find the emerging CSP computing requirements that would lead to new types of innovative system design for computer racks, servers, storage, and switches to support Cloud and AI requirements. One of the OCP's principal founders, Meta (formerly Facebook), inspired its mission as it prepared to build a state-of-the-art large-scale data center for cloud services in Prineville, Oregon.

As Wikipedia put it, "The Open Compute Project Foundation (OCP) was initiated in 2011 with a mission to apply the benefits of open source and open collaboration to hardware and [to] rapidly increase the pace of innovation in, near, and around the data center's networking equipment, general purpose and GPU servers, storage devices, and appliances."

The OCP has clearly stated the objectives of the Summit's activities, including work within its projects now and throughout the year: to create standards, to build ecosystems, and to encourage the adoption of those standards across IT infrastructure for a wide range of workloads (e.g., Artificial intelligence (AI), the Cloud, the Edge, and HPC workloads).

Since 2011-2012, the organization has grown substantially, as has its worldwide ecosystem of vendors and industry partners that contribute to approved OCP Contributions (e.g., specifications, white papers, use cases). The OCP working groups' contributions are adopted by data center facilities and companies that manufacture and provide hardware, silicon, and software to customers worldwide.

Many CSPs – including the largest ones – have joined the OCP ecosystem, along with a wide range of system vendors, semiconductor manufacturers, software firms, data center equipment providers, and some of the world's largest commercial companies and government agencies. Many of these customers have their own expansive IT data centers and dense, powerful IT resources – and the large CSPs build their own customized infrastructure by leveraging industry-wide OCP specifications.

AI Rocked the World This Year

Top changes for 2024/2025 include: adjusting the power/cooling envelope to remove heat surrounding highly dense CPU and GPU processors; designing a new generation of semiconductors, processors, and chiplets built for AI systems; accelerating the switching speeds for network equipment; adding optical switches and photonic links; updating the Ethernet networks already installed in data centers and CSPs sites worldwide; and writing new software to navigate new and emerging data "fabrics" tying all the equipment together.

It is a tall order – a demanding task – for system designers, semiconductor companies, "fabs", and software firms, among other technical contributors. And yet, it must be done, or AI technology will not be able to meet the fast-growing expectations of large customers and the largest CSPs that provide cloud services to customers. That's why the OCP Community focuses on meeting the emerging hardware and software requirements that will enable AI/ML workloads, regardless of where they are run.

In the wake of ChatGPT's release in November 2022, the past year has seen some remarkable events advancing the cause of AI-enabled systems for business and HPC. Among them:

- ChatGPT attracted 100 million users within two months of its launch.
- NVIDIA sales doubled in one quarter from Q1FY24 to Q2FY24.
- Google announced that its BARD AI is taking on ChatGPT and OpenAI for search.
- Microsoft launched ChatGPT with Microsoft Bing and Microsoft 365 Copilot.
- Meta launched Meta AI and the open-source LLaMA (Large Language Model Meta AI) for AI/ML.
- AWS announced its Bedrock AI-as-a-Service, offered through the AWS public cloud.

OCP Innovations for AI/ML, Cloud, and HPC

Given the drive toward AI enablement in the Data Center and the Cloud, the OCP Community is working to drive more innovations for computing infrastructure. This means that the OCP supports the vendor community. Its work enables a collaborative process that is driving industry standardizations – and the vendors in the OCP Community are delivering their own implementations of hardware products and software that meet those specifications.

Here are some of the most recent areas of OCP innovation, as announced and discussed at the Summit:

- **Security:** OCP announced hardware and firmware, delivered by an ecosystem of project participants, to improve security for cloud-based workloads. The S.A.F.E. program (Security, Appraisal, Framework, and Enablement) results from this security effort. As AI becomes more widely used in the enterprise, having safeguards against firmware-level security breaches is more important than ever because any breach could jeopardize an organization's security environment and mission-critical data.
- **Data Formats:** The OCP ecosystem is focusing on data-center hardware and firmware security as an important initiative to improve the overall efficiency of AI processing. An extensive list of tech vendors has joined this OCP initiative, including AMD, ARM, Intel, Meta, Microsoft, NVIDIA, and Qualcomm.
- **Firmware:** OCP committees are working to advance firmware technology to provide advanced features for systems supporting AI/ML, HPC, and other demanding workloads.
- **Falcon for faster networking:** At the Summit, Google announced that it has decided to open its Falcon technology, a reliable low-latency hardware support for improved

networking, to the ecosystem through the OCP for further innovations. This announcement reinforced the customers' view that the OCP Community provides open standards that systems vendors and software companies will implement. Going forward, OCP projects can use an open-systems development process for future innovation, supporting joint development of those standards.

AI Challenges, Demands, and Growth

Increased densities: Datacenter densities are growing, demanding high-speed optics and a massive increase in power consumption (Figure 1). At the same time, increased density is creating new IT challenges for power and cooling. Reducing power/cooling requirements will avoid shutdowns due to excessive heat and/or insufficient power during data-center operations for CSPs and the large, commercial, HPC, and scientific/research customers building and operating large-scale data centers.

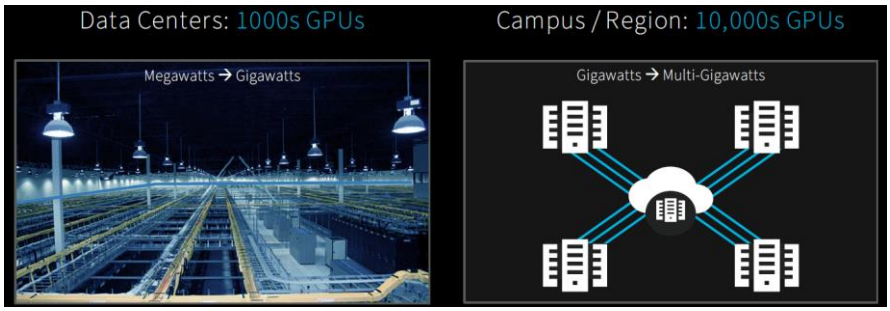

Figure 1: Denser and Power-Hungry Datacenters (Source: Marvell)

Long-distance connections: Large AI clusters can span hundreds or thousands of meters with hundreds of servers and switches and thousands of GPUs and optics. Therefore, data transfers and latency in AI task completion must be avoided if on-site data centers or campus hubs are to remain useful infrastructure for end-to-end AI-enabled systems.

Demand for Bandwidth. AI accelerates bandwidth demand (Figure 2). At the Summit, many speakers cited the phrase "The Network is the Computer." Originally coined by Sun Microsystems in 1984, the phrase points to the critical role of faster network switches and network fabrics as essential parts of the overall IT landscape for AI and HPC. Many Summit talks centered on improving network efficiency, network speed, and network interconnects – including using optical switches to speed data transmission while reducing power/cooling characteristics for network devices and network switches.

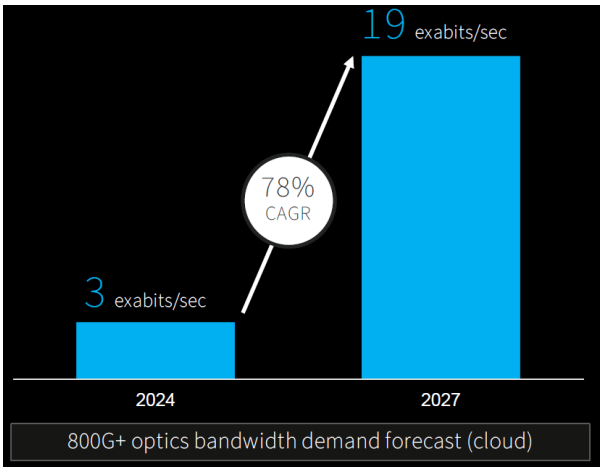

Figure 2: Bandwidth Growth Forecast (Source: Marvell)

Summit Breakout Discussions

Because AI is, quite literally, impacting every level of IT infrastructure, cooperation and interoperability of systems are key to ensuring that end-to-end infrastructure works well, and can be extended by adding new technologies that conform with the standards, over time.

All equipment vendors are attempting to make data centers more sustainable with advanced modularity vision (as seen in Figure 3), more efficient computing (shown in Figure 4), and hardware-aware software (shown in Figure 5). Intel, for instance, said at the OCP keynote session that it aims to improve processor efficiency by 50%, reduce platform-level carbon by up to 30%, and datacenter-level efficiency by up to 30%.

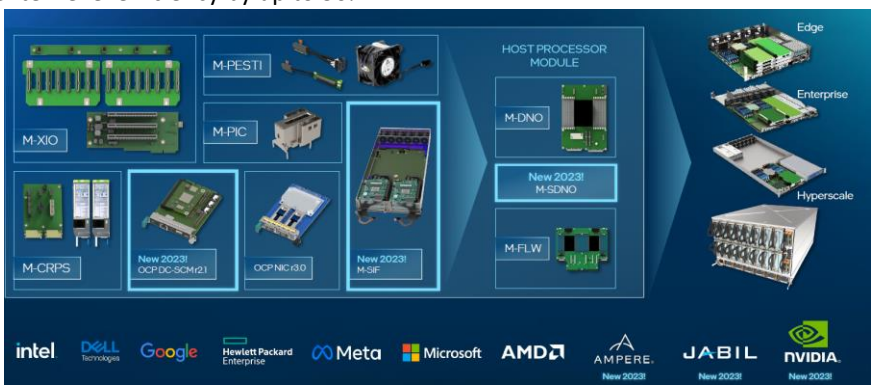

Figure 3: Hardware Modularity (Source: Intel)

Figure 4: Improved Efficiency (Source: Intel)

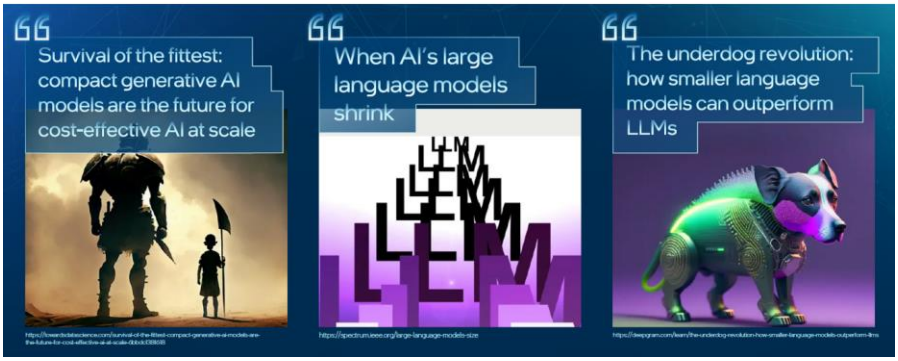

Figure 5: Hardware-aware Software (Source: Intel)

Similarly, Google speakers said that is aiming for sustainable, secure, and scalable systems. Partha Ranganathan, VP and Engineering Fellow, at Google, included these points in his keynote talk.

- For sustainability, Google plans to run on carbon-free energy 24x7 + net-zero carbon emissions across all operations and value chains by 2030. Partnering with Microsoft and Schneider Electric, it announced on September 28 the Net Zero Innovation Hub for Data

Centers with an ambitious agenda across all scopes, including Waste Heat Reuse and Grid Availability. It also includes using "Greener Concrete" since concrete is 11% of total carbon emissions.

- For security, Google is promoting Trusted Computing comprising secure, confidential, and reliable computing. Partnering with AMD, Microsoft, and NVIDIA, it has developed Caliptra for SoC integration into CPUs, GPUs, DPUS, and TPUs. It jointly developed S.A.F.E. (discussed earlier), a standardized approach for provenance, code quality, and software supply chain for firmware releases developed in partnership with Microsoft and OCP Security Project. Finally, it has open-sourced Silent Data Corruption (SDC) frameworks for CPUs and GPUs, building towards a common framework for both.
- For scalability, Google said it is working on the Accelerator Ecosystem, a foundational infrastructure to enable GenAI from chip to compilers, and enabling AI to transform computer architecture; Falcon, a reliable low-latency hardware support for improved networking. As a hardware-assisted transport layer, Falcon is designed to be reliable, high performance, and low latency, and leverages production-proven technologies including Carousel, Snap, Swift, PLB, and CSIG. Google contributed to the OCP Datacenter NVMe 2.5 spec for storage, focusing on flexible data placement, security, and telemetry. Google, Microsoft, Samsung, Kioxia, and Solidigm also develop an open-source NVMe Key Management block.

Covering the Waterfront of Infrastructure Change

Key elements of widely deployed standardized OCP-compliant building blocks include the following – all of which were topics at the Summit:

- Emerging standards for system design, including the Computer Express Link (CXL) standard for a wide range of interconnects.
- Security protection to reduce data loss via cyberattacks and ransomware.
- Interoperability of compute devices, supporting mixed-chip environments. This capability is vital to cloud service providers, who must support a variety of processor types to run many types of customer workloads.
- Next-generation CPUs and GPUs designed for scale-up AI and scale-out AI.
- Support for denser server systems and storage systems, bringing components and devices inside the systems closer together.
- Faster network switches, including a new generation of optical (light-based) switches, to accelerate inter-chip and inter-board connections.
- CXL consortia interconnects incorporated into systems by the OCP Community to speed performance and reduce latency in large-scale infrastructure for the AI/ML, Cloud, and the Edge; these CXL interconnects, designed according to new industry standards, also support multi-generation, multi-vendor processor deployments for high-speed computing, and faster data transport.
- Power/cooling systems and heat-dissipation devices to improve operational performance and to meet, and exceed, sustainability requirements.
- Liquid cooling systems, including full immersion of circuit boards in transparent hydro-carbon liquids for improved cooling performance in closely packed data centers.

AI Acceleration in a Highly Dense Compute Environment

Ingredients for building next-generation AI systems as a foundation for performance, reliability, and security must address these factors:

- **Datacenter densities are growing, demanding high-speed optics for low latency high bandwidth networking, and driving a huge increase in power consumption.** At the same time, increased density is creating new IT challenges for power and cooling, avoiding shutdowns due to excessive heat and/or insufficient power during data-center operations for cloud providers (CSPs) and large commercial customers building and operating large-scale data centers.
- **Data transfer rates must be accelerated.** Large AI clusters span hundreds or thousands of meters to meet these demands with hundreds of servers and switches and thousands of GPUs and optics. Therefore, data transfers and latency in AI task completion must be

avoided, especially if on-site data centers or campus hubs are to remain useful infrastructure supporting end-to-end AI-enabled systems.

Building Scalable AI Systems

Ingredients for building next-generation AI systems include these technology pillars as a foundation for performance, reliability, and security. Here is a listing of some of the newest developments for next-generation infrastructure, across key focus areas:

- **Sustainability:** Partnering with Microsoft and Schneider Electric, Google Cloud announced in September 2023 the Net Zero Innovation Hub for Data Centers with an ambitious agenda across all scopes, including Waste Heat Reuse and Grid Availability. It also includes using "Greener Concrete" since concrete is 11% of total carbon emissions.
- **Security:** Partnering with AMD, Microsoft, and NVIDIA, the OCP Community has developed Caliptra for hardware root of trust security and the OCP S.A.F.E. Program for firmware security, both discussed earlier.
- **Scalability:** The OCP Community is working on the Accelerator Ecosystem, a foundational infrastructure to enable GenAI from chip to compilers and enabling AI to transform computer architecture.
- **Composable architecture:** Modern software applications rely on Kubernetes orchestration to move software containers (e.g., Red Hat OpenShift and Docker containers) to run on alternative hardware resources, as needed. Composable architecture allows software-defined hardware systems to be built "on the fly" by using fast interconnects and widely implemented software-defined standards. This approach allows customers to flexibly meet changing demands for computing resources and system capacity.

Thermal Cooling Emerges as An Important Key for AI Systems

As the number of transistors per semiconductor chip continues to grow, Moore's Law is falling short of Intel's original expectations for Moore's Law in the late 1960s (doubling performance every two years). For history buffs, the Intel 4044, a 4-bit microprocessor with 2,300 transistors was originally shipped in 1971 – and Moore's Law predicted the performance curve for Xeon processors well into the 2000s.

Today, chip density for modern microprocessors is much higher than it was in Intel processors of the 1980s and 1990s. Another example of the trend: the Apple M2 Ultra chip, for instance, has 130 billion transistors and the AMD M300X has 153 billion transistors. But these are general-purpose processors (CPUs) – and the trend now is to add special-purpose processors (e.g., GPUs, DPUs, IPUs, NPUs, and TPUs), as well as smaller chiplets for specific functions that will be included inside the larger CPUs and specialized ASICs

Processors will continue to get denser to meet the more demanding needs of AI – that much is clear from recent years of ever-larger, ever-more powerful microprocessors. However, the increased density implies increased power consumption and heat generation, resulting in design challenges and operational techniques to alleviate these problems. Traditional air-cooling methods are inadequate, prompting some to use liquid-based cooling inside the server racks.

Liquid Cooling is the new buzzphrase. Yet, people tend to forget that this approach to reducing overall power/cooling requirements was used in the Cray X-MP supercomputers in the 1980s! The impressive Cray supercomputer design of that era was circular, allowing IT personnel to sit "around" the Cray system and watch the refrigerant in its cooling tubes descend as the water moved through the system.

Today, however, major oil companies, such as ExxonMobil, Shell, and BP (owner of Castrol), and specialty companies, such as Lubrizol, are providing synthetic hydrocarbon liquids that do not interact with the electronics on the system circuit boards. These companies use a combination of hydrocarbon-based chemicals developed by them to "immerse" electronic hardware, surrounding

it with heat-dissipating fluids (Figure 6). The fluids are then circulated to heat exchangers that are usually located at the "back" of the racks and cabinets in the data center.

At the Summit, many liquid cooling systems, including models from Intel, Wiwynn, and others, were shown on the exhibit floor. Several methods are used, including "cold plate" systems that draw heat away from the processor boards: Single-Phase liquid cooling using circulating hydrocarbon-based transparent fluids. By contrast, Two-Phase liquid cooling uses a "bubbling" process to draw heat away from circuit-board components and processors within a system.

According to ExxonMobil, the typical temperature of the 'pond' is 50 deg. C, and the flash point of most of the coolants is between 150 deg. C and 200 deg. C, so there is no danger of fires. This is still a fairly new technology in data centers and the issues of cleaning and maintenance of the hardware and the long-term effects of hardware exposed to chemicals are still to be addressed.

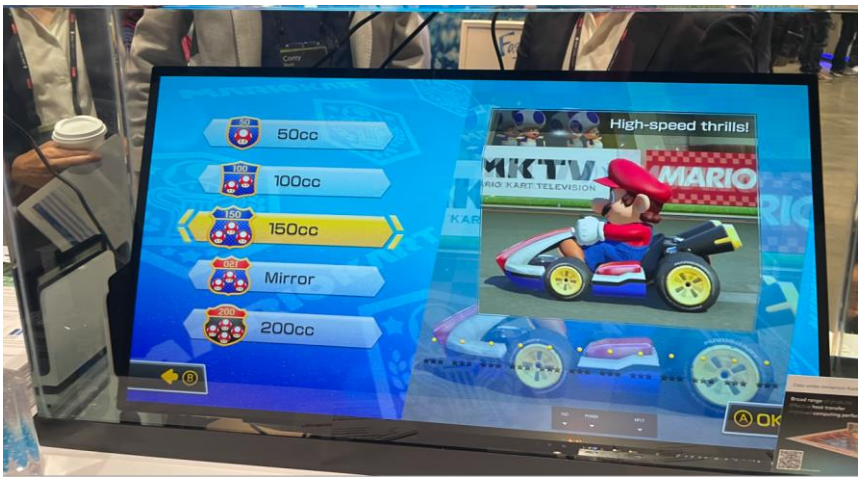

Figure 6: Gaming laptop immersed in coolant (Source: ExxonMobil)

We note here that there is a new group of small vendors in the liquid-cooling space.

Taiwan-based Delta Electronics offers hybrid air-assisted liquid cooling (AALC) systems (See Figure 7). The company's solutions are designed to fulfill the highly demanding power, heat dissipation, and high-speed networking demands of AI and HPC data centers. such as, Delta boasts a leading peak efficiency up to 97.5% with cooling density up to 2.5 times compared to traditional air-cooling systems.

Figure 7: Hybrid Cooling (Delta Electronics)

Another example of a liquid-cooling solution for AI and HPC systems is ZutaCore, a small firm with R&D in Israel and offices in San Jose, CA. ZutaCore offers an elegant (ZutaCore's HyperCool) solution that uses a highly efficient, two-phase boiling and condensation process, moving large amounts of heat off the processors and away from servers. It is a complete, closed-loop solution for cooling a server's heat-emitting components such as the CPU, GPU, and FPGA. It uses a waterless, direct-on-chip method, one of the most effective forms of cooling, to apply coolants directly to the chips to extract and disperse heat. No water is used in the system, so the equipment is protected from corrosion and other water-related threats.

Figure 8: Waterless On-Chip-Coolong (Source: ZutaCore)

Summary: Today's AI Challenges Will Spark Performance in 2024/2025

We believe that customers' need for flexible infrastructure in the age of AI and the Cloud will drive the design of a new generation of devices and software based on inventive and new technologies, AI will affect every aspect of modern computing, causing changes in hardware specifications and software standards to support AI/ML workloads efficiently and cost-effectively.

Overall, innovations will be required to address a constellation of challenges brought about by rapid growth in AI adoption. Among these are: increased density of processors and components inside systems; power/reduction to cool things down and save energy costs; and the capacity to "scale up" and "scale out" depending on the type of infrastructure deployed by cloud service providers and large companies, on behalf of their end-customers.

The pace of innovation for AI-enabled systems is accelerating. The long list of challenges must be addressed, so that the rapid pace of AI adoption for cloud and commercial sites can continue on its steep trajectory. To achieve these technology innovations, the OCP projects are defining and publishing new approved OCP Contributions (e.g., specifications, white papers, use cases) so that customers of its vendor community will be able to transform their current IT architectures for the Age of AI.

CABOT PARTNERS

OPTIMIZING BUSINESS VALUE

Cabot Partners Group, Inc.

100 Woodcrest Lane, Danbury, CT 06810

www.cabotpartners.com| info@cabotpartners.com