NAND RESEARCH

RESEARCH NOTE

UALINK ALLIANCE FOR ACCELERATOR INTERCONNECTS

STEVE MCDOWELL, CHIEF ANALYST
MAY 30, 2024

CONTEXT

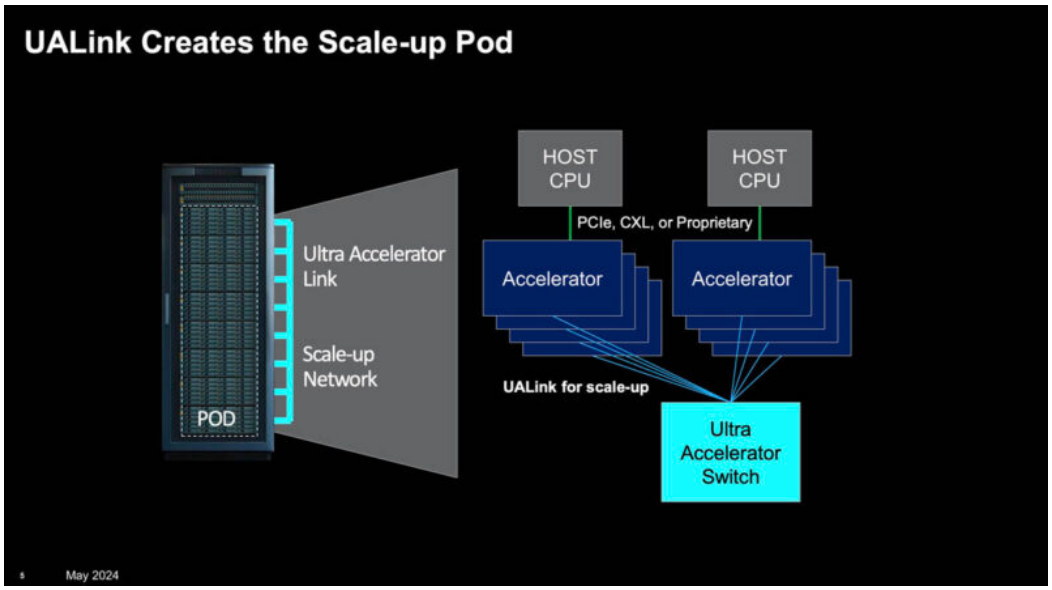
UALink is a new open standard designed to rival NVIDIA's proprietary NVLink technology. It facilitates high-speed, direct GPU-to-GPU communication crucial for scaling out complex computational tasks across multiple graphics processing units (GPUs) or accelerators within servers or computing pods.

The effort emerged from a new consortium spearheaded by AMD, including industry heavyweights such as Intel, Broadcom, Cisco, Google, HPE, Meta, and Microsoft. Notably absent are Arm and NVIDIA.

WHAT IS UALINK?

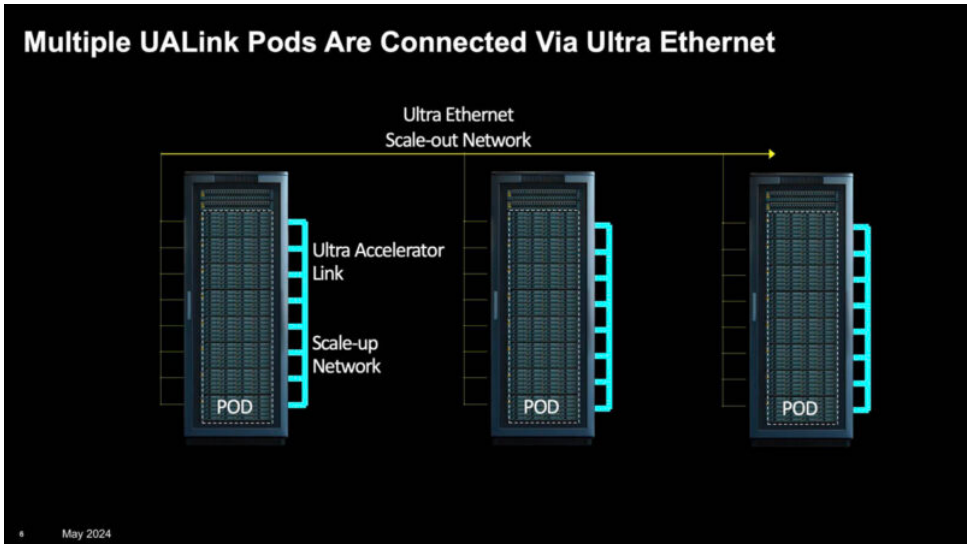
UALink, or Ultra Accelerator Link, is an open-standard technology developed as an open alternative to NVIDIA's proprietary NVLink. It focuses on enhancing GPU interconnectivity for high-performance computing and AI applications, supporting up to 1,024 accelerators in a shared memory pod by Q3 2024, with further enhancements planned.

This open standard seeks to foster innovation and provide more affordable and diverse options for AI server nodes and rackscale systems, challenging Nvidia’s proprietary solutions.



Key Features of UALink include:

1. **Scalability**: UALink is designed to scale efficiently, aiming to support up to 1,024 accelerators in a single shared memory pod by its initial specification release. Future updates are expected to push this scalability even further.

2. 

3. **Open Standard**: Unlike proprietary solutions that lock users into specific technologies or vendors, UALink is an open standard. This approach allows for broader industry collaboration and integration, giving various hardware and software vendors the flexibility to adopt and adapt the technology without restrictive licensing or compatibility issues.

4. **Ethernet Integration**: UALink plans to leverage Ethernet's Layer 1 transport level, combined with Infinity Fabric on top, to create a scalable and high-performance interconnect. This approach aims to combine the advantages of

Ethernet's broad and flat scaling capabilities with the advanced memory semantics of Infinity Fabric.

5. **Memory Semantic Fabric**: UALink will utilize AMD's Infinity Fabric as the primary shared memory protocol, which enables efficient memory sharing across accelerators. AMD's more limited, GPU-specific xGMI protocol will also be supported.

6. **PCI-Express Future-proof**: While UALink will support Infinity Fabric and xGMI protocols initially, it is also expected to be compatible with future PCI-Express switches, like Broadcom's upcoming "Atlas 4" switch adhering to PCI-Express 7.0, ready for market in 2025, ensuring that UALink can evolve with advancements in interconnect technology.

7. **Collaborative Development**: Spearheaded by AMD, with participation from other tech giants like Intel, Broadcom, Cisco, Google, HPE, Meta, and Microsoft, UALink benefits from the collective expertise and resources. This allows multiple vendors to contribute and benefit from a unified interconnect protocol, avoiding the monopolization seen with proprietary solutions.

ANALYSIS

The initiative to create UALink reflects a growing recognition within the industry of the need for scalable, open architecture in high-performance computing. As AI and data-intensive applications evolve, the demand for such technologies will likely increase, positioning UALink as a critical component in the next generation of computational tools.

By offering a credible and open alternative to NVIDIA's NVLink and NVSwitch, UALink aims to break NVIDIA's dominance in the GPU interconnect market. This is particularly important as AI workloads and models continue to grow, requiring more scalable and efficient interconnect solutions.

The introduction of UALink puts pressure on NVIDIA to continue innovating and improving its own technologies. To maintain its competitive edge, NVIDIA will need to enhance the performance, scalability, and cost-effectiveness of NVLink and NVLink Switch, potentially accelerating the pace of innovation in the industry.

The open-source nature of UALink encourages a collaborative approach to development. Moving away from proprietary systems fosters innovation and allows for rapid adoption and improvement of the technology.

UALink's potential goes beyond just providing an alternative to NVLink. It's about creating an ecosystem where companies can contribute to and benefit from advanced interconnect technologies regardless of their size. For hyper-scalers and large data centers, adopting UALink could mean significant cost savings, greater flexibility in hardware deployments, and enhanced performance.

UALink is a significant move by a consortium of technology companies to limit proprietary interconnects for AI accelerators, promoting openness, competition, and innovation. Its impact on the market is likely to be substantial, fostering a more dynamic and cost-effective ecosystem for AI and accelerator-based systems. It's hard to find a negative in that mission.

For NVIDIA, UALink poses a competitive challenge that will necessitate strategic and technological responses to maintain its leadership in the high-performance interconnect space.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.