OCP Future Technologies Symposium Paper: Optical Interconnect – Pathways to an Open AI Infrastructure

Presented at the 2023 Symposium by Matthew Williams, Chief Technology Officer (CTO), Cerio. matt@cerio.io

***Abstract*— Artificial intelligence (AI) is transforming every industry, driving demand for greater processing capacity and specialized hardware. Open infrastructure optimizes capacity and agility without the costly overhead and inefficiencies that make the current data center system model unsustainable. Distributed architecture with passive direct interconnect technology are key capabilities for delivering dynamic scale using standard optics at a significantly lower cost and footprint.**

I. INTRODUCTION

Artificial intelligence (AI) and Machine Learning (ML) applications have a wide range of point-to-point latency and bandwidth requirements with multi-directional traffic that is highly sensitive to congestion. Closed server systems designed for homogeneous compute are hitting a wall, increasing overhead, limiting choice and creating an unsustainable environmental impact.

Cerio designs open, highly scalable AI infrastructure systems to optimize capacity and heterogeneity without inflating costs or power consumption. Data centers can apply supercomputing topologies using off-the-shelf pluggable optical transceivers to build advanced networks that aren't limited by the constraints of spine-and-leaf architectures. By decoupling the optics from the underlying switching, Cerio delivers scalable performance at a significantly lower cost and footprint using cabling that is entirely passive and requires zero power or cooling.

Cerio's fully distributed architecture eliminates centralized switching by embedding the control plane, data plane and data forwarding function into every node, connected by a completely passive direct interconnect (called a SHFL). Prewired for the target topology, the SHFL can be used with off-the-shelf optics and scale to hundreds of nodes in servers or PCIe expansion chassis for a high-radix infrastructure with 15-150 Tbps of fabric bandwidth. Topology-and-transport independence also ensures full heterogeneity for best-fit optics and hardware components.

Prewired supercomputing topologies are now being applied to composable PCIe systems to optimize AI/ML traffic flow and increase the utilization of resources at data center scale. Implementing PCIe composability at scale for GPU capacity use cases delivers a 35% reduction on average in per-rack power consumption for AI infrastructure.

The approach is also advancing scalable CXL memory pooling years in advance of CXL 3.x and 4.x availability. Cerio SHFLs are designed to ensure that servers accessing remote memory pools have one or more direct connections to the memory pool, enabling a wide range of use cases.

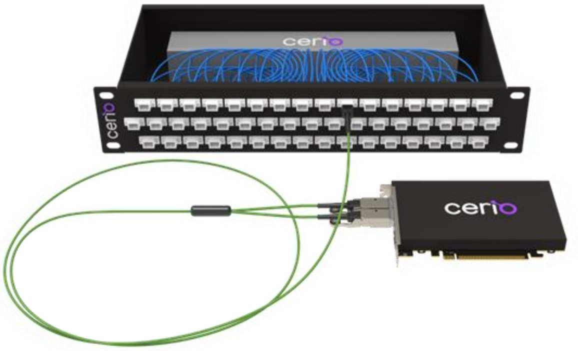

Figure 1: The Cerio platform's passive direct interconnect requires zero power or cooling, uses off-the-shelf optics and scales to 100s of nodes in a fully distributed architecture.

II. BREAKING THE COST-PERFORMANCE CONNECTION

Traditional network fabrics are composed of point-to-point links between endpoints and first-level switches, and a set of point-to-point links between each switch layer. Commonly, links are composed of four, eight or 16 parallel lanes, with each lane supporting usable bandwidth of 25, 50, 100 or 200 Gbps. In centralized networks, higher per-lane bandwidth drives significantly higher transceiver cost and power consumption along with corresponding increases in SERDES power consumption and complexity. In data centers, the cost of optics is rapidly growing as a percentage of total system cost, with estimates as high as 25%. The additional cost and power consumption of high per-lane bandwidth transceivers often exceeds the performance gain, creating an overall increase in both dollar/performance and power/performance.

Cerio breaks the performance-cost dependency by eliminating layers of centralized switching. Each node in the network contains both a host interface (PCIe, CXL, Ethernet, etc.) and a distributed hardware-based switch that treats each lane as an individual link used to create direct connections to a large community of direct-neighbor nodes. Inherent "multipathing" capability is used to optimize AI/ML traffic flows and create very high path diversity.

Distributed architecture also removes the scaling breakpoint in centralized switching where large number of transceivers in each path create a non-linear transceiver cost and power at scale. By decoupling the optics from the switching, any scale of switch can be used with any scale of optics without the corresponding step function in cost and energy consumption.

III. PREWIRING COMPLEX TOPOLOGIES

One of the challenges with traditional interconnects has been the complexity of manually wiring the target topology. To address this wiring challenge in the distributed model, Cerio uses a commodity high-density passive optical cable (e.g., MTP/MPO 24/32) to connect 12 or 16 links from each node to a port on a passive SHFL. This cable contains multiple fiber pairs that carry the independent single-lane links.

Inside the SHFL, each of these links is broken out of the high density-cable and connected to the port associated with the correct target node. A single passive optical cable connects each node to a SHFL, and additional passive optical cables connect SHFLs together to create large-scale inter-rack connectivity. The power- and cost-savings, simplicity, and flexibility of passive cabling deliver significant advantages over multi-layer centralized switching.

The pattern of connectivity of the target topology is prewired into the rack-mountable Cerio SHFL. The SHFL breaks out each of the links and physically routes them to the target neighbor node, forming a direct optical path. While multi-lane pluggables (QSFP, OSFP, etc.) are generally used for “A” to “B” connections, the SHFL connects “A” to multiple neighbors to create much more advanced topologies with any number of racks and any number of nodes in every rack. The model supports efficient inter-chassis and intra-chassis communication for CPU-x, GPU-GPU or any resource type. Any number of nodes can be added (or removed) at any time for simplified in-place scaling.

IV. ACHIEVING PCIe COMPOSABILITY AT SCALE

Composable Infrastructure enables cloud-like provisioning by connecting GPUs or any remote PCIe device into nodes at job-execution time. The Cerio SHFL brings the simplicity of prewired topologies to composable systems by making predefined connections between Fabric Nodes in the host and chassis. Network topology and route information automatically update when Fabric Nodes are added, removed, become unavailable, or if an individual link path is down. The topology can be changed at any time by adding a new SHFL configuration.

Cerio overcomes the limitations of PCIe system scaling by providing transport optimization services that decouple the native PCIe transport from the underlay fabric. From a host point of view, it looks like the native protocol, but the emulated PCIe switch can now leverage Cerio’s transport optimization services at scale while enabling reliable, 24x7 operation with built-in failover and hot-swapping capability. A remote PCIe switch complex enables local communication between devices in a physical chassis. Placing an end-to-end reliable network between the host PCIe root complex and a remote PCIe device addresses the reliability issue that causes system failures inherent to native PCIe expansion. The PCIe service translates between host and chassis PCIe domains, alleviating memory space exhaustion by isolating remote devices to their own local memory spaces.

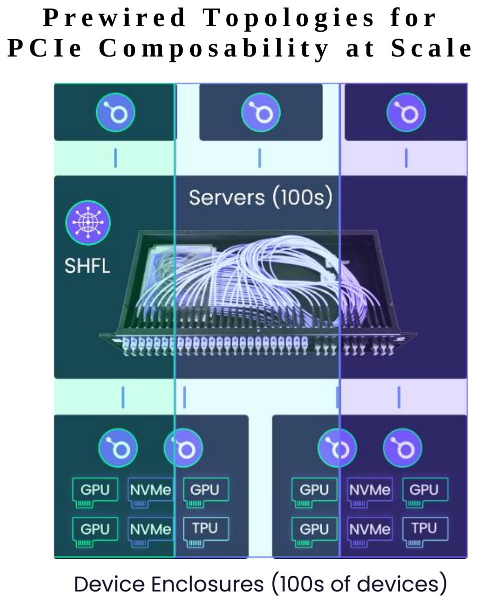

Figure 2: The Cerio SHFL makes predefined connections between Fabric Nodes in the host and chassis to simplify composed systems.

V. ADVANCING CXL MEMORY POOLING

By decoupling the host protocol from the underlay fabric, Cerio can bring CXL Type 3 memory pooling early with advanced topologies and transport capabilities at scale. Cerio SHFLs are designed to ensure that servers accessing remote memory pools have one or more direct connections to the memory pool, enabling a wide range of memory pooling use cases.

Since Cerio is protocol agnostic, it’s possible to take CXL 1.x and 2.x systems and scale them to hundreds or even thousands of endpoints years before CXL 3.x with full forward-compatibility.

Cerio’s decoupling approach also simplifies upgrades to future generations of CXL. Only the firmware in the Cerio Fabric Nodes within each endpoint needs to be upgraded. The existing node-to-SHFL cabling, SHFLs, and SHFL-to-SHFL cabling can remain in place, making upgrades time- and cost-efficient. Updating the topology of an existing deployment only requires the replacement of the SHFL, with no changes to the Fabric Nodes or cabling. The Fabric Nodes will simply discover the new topology and build sets of source routes to the new destinations.

VI. CONCLUSION

Optimizing composable infrastructure at data center scale provides cloud-like agility with greater control over the choice, utilization and life cycle of resources. Using prewired topologies and off-the-shelf optics, data centers can compose highly dynamic and efficient systems at the lowest possible cost. Fully distributed and protocol agnostic, Cerio simplifies the evolution of AI infrastructure, from scalable PCIe composability to CXL memory pooling and beyond. The open approach to AI infrastructure gives data centers a more sustainable way to support unpredictable demand and heterogeneous workloads.