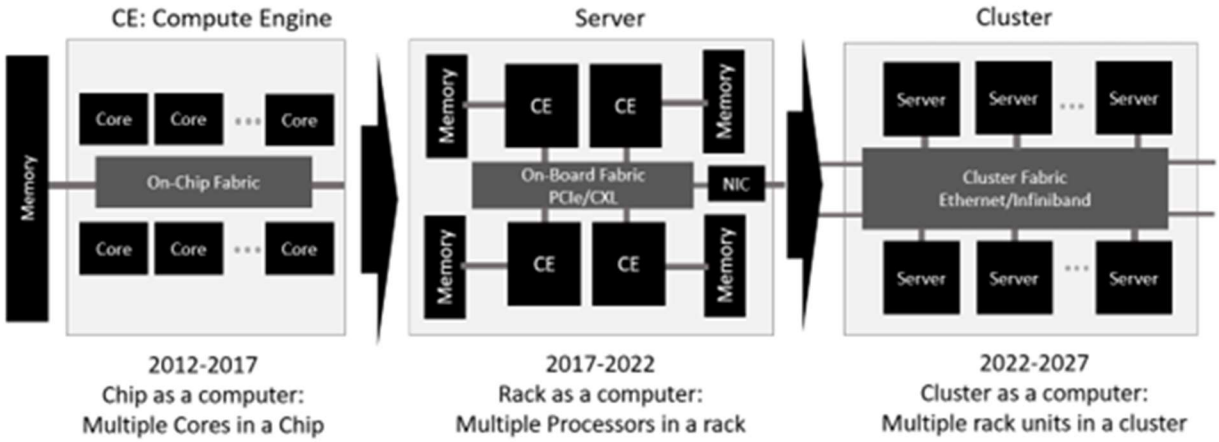
Enabling Next-Generation AI Applications with 800G/1.6T Ethernet

While recent Artificial Intelligence (AI) breakthroughs such as ChatGPT have showcased the power of AI to users, AI applications in the form of predictive search queries, recommendation engines, and image recognition have been evolving over the past 15 years.

These AI advancements are the result of machine learning, where machines “learn” autonomously. Machine learning, crucial for enabling AI applications, relies on vast datasets for training. Essentially, the better the model and the larger the dataset, the more impressive the AI results. This underscores the need for continual growth in computing power to fuel the next generation of AI applications.

Fundamental principles of computer architecture suggest computing throughput can be enhanced by aggregating multiple computing cores and sharing associated memories. In the pre-2012 era, the first generation of acceleration in computing increased throughput by boosting the performance of individual compute cores and enhancing processor memory interface speed.

The subsequent phase, spanning from 2012 to 2017, witnessed an increase in compute throughput through the integration of multiple processing cores into a single chip or Compute Engine.



The second phase of enhancing compute throughput, observed between 2017 and 2022, involved treating a rack as a compute unit by connecting multiple rack units in a cache-coherent manner. During this period, machine learning applications began to take shape in compute clusters organized as data center networks, with server racks serving

as the building blocks. In this phase, one rack unit of a server rack can be considered a compute unit. Technologies such as CXL facilitated memory pooling, while Ethernet-enabled rack unit-to-rack unit connectivity became a reality.

Today's AI applications are driven by Large Language Models (LLMs), which are trained on vast unstructured data. The effectiveness of LLMs is directly proportional to the number of parameters used in training. For instance, GPT-3 boasted 175 billion parameters, and GPT-4 is anticipated to exceed 1 trillion parameters. To keep pace with the anticipated improvements in AI performance, it is projected that LLM parameters will double every four months. Cluster-scale computing, where multiple racks are interconnected to function as a single compute unit, is essential to meet the exponential compute demands of next-generation AI in 2022-2027 timeframe and the significance of a faster interconnect rate in enabling more efficient connection of compute racks in this era becomes apparent and necessitates next-generation low-latency Ethernet.

The IEEE P802.3df task force has recently released version 1.0 of the 800GbE specification, while the IEEE802.3dj task force is actively working on formulating the 1.6TbE standard. These efforts are paving the way for open and interoperable 1.6 Tbps Ethernet links for chip-to-chip and electro-optical interfaces, essential for facilitating the compute aggregation required by next-generation AI applications.

Source: https://ethernetalliance.org/blog/2024/03/13/enabling-next-generation-ai-applications-with-800g-1-6t-ethernet/