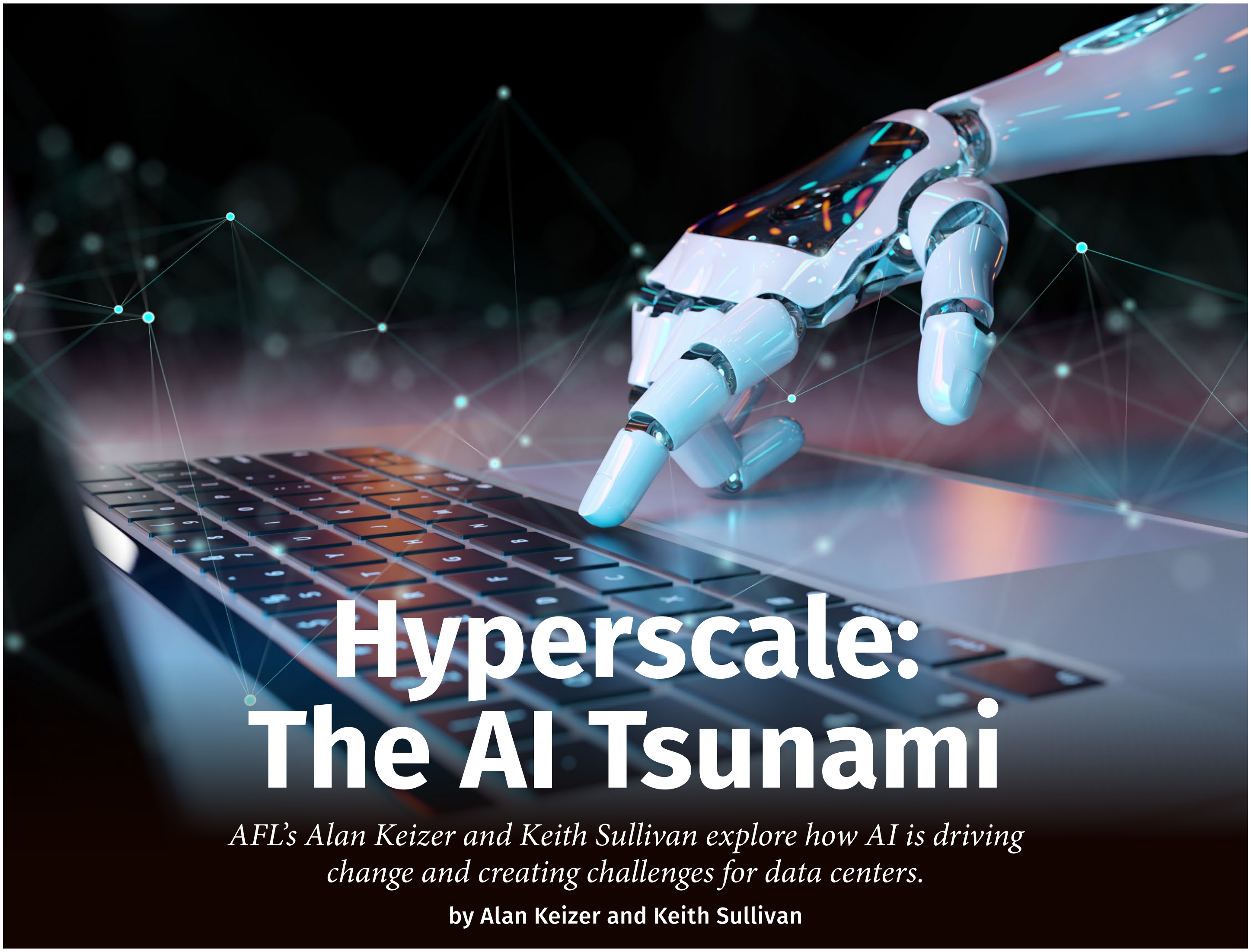
VOICES OF THE INDUSTRY Published on the DATA CENTER Frontier website



Hyperscale: The AI Tsunami

AFL's Alan Keizer and Keith Sullivan explore how AI is driving change and creating challenges for data centers.

by Alan Keizer and Keith Sullivan

©2019 sdecoret/Shutterstock.

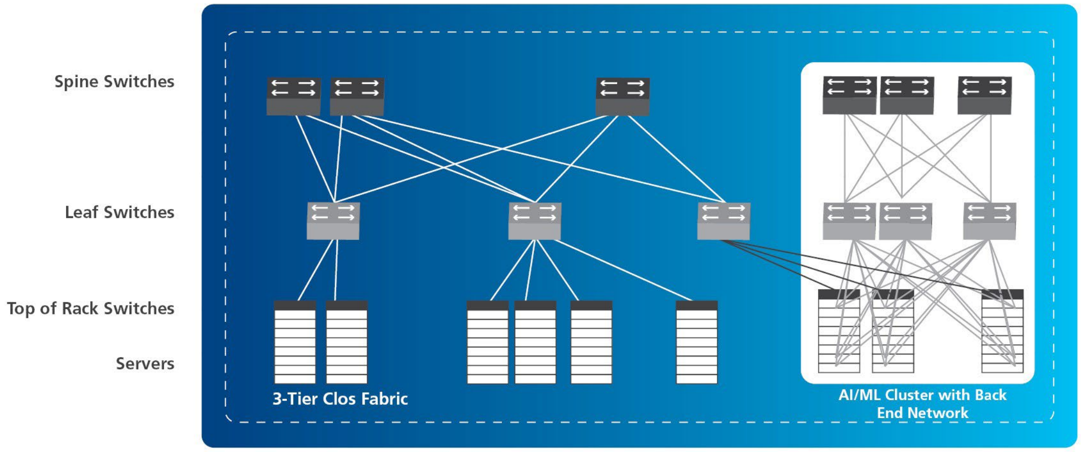
Artificial Intelligence (AI) is increasingly becoming a part of our everyday lives, integrating itself into different sectors like advertising, healthcare, education, transportation, and finance. This incorporation not only promises greater efficiency with tasks, personalized services, and innovations, but also signifies a transformative period that will continue to unfold in the coming decade.

Large Language Models

AI's recent advances in Large Language Models (LLMs) and generative AI systems underscore this transformative phase. As digital leaders continue to train larger neural networks with more parameters on massive data sets, the resulting generative AI models consistently demonstrate impressive capabilities. However, the training and deployment of these models, coupled with the delivery

Dr. Alan Keizer is Senior Technical Advisor for AFL Hyperscale. Alan earned a Ph.D. in Physics from Cornell University and has worked at the intersection of computers and networking for the past four decades. He has been an individual technologist, a business development leader and senior executive. Joining FibreFab in 2006, he was a key contributor to the business which was acquired by AFL Telecommunications in 2013 and became a key component of today's AFL Hyperscale. Since 2010, his main focus has been development of optical fiber connectivity solutions for hyperscale computing.

Keith Sullivan is Director, Strategic Innovation for AFL. With over 27 years of experience, Keith is an accomplished business leader in the fiber-optics industry with a specialty in data centers. He started his career in 1995 in a factory in England producing specialty fiber assemblies for high-speed devices in the telecom industry. Since then, he has worked for industry-leading organizations in sales, marketing, and product line management.



Integrated hyperscale data center network with main data and storage fabric and embedded AI cluster.
Dr. Alan Keizer, Senior Technical Advisor for AFL

of large-scale inference at scale, requires extremely large, specialized computing systems. Consequently, hyperscalers embed data centers with high-performance, AI-optimized dense computing and storage, further solidifying AI's pivotal role in shaping the future of connectivity.

Artificial Neural Networks

AI involves the simulation of human intelligence by machines. AI systems can learn from data and experience, becoming more capable with continued use. This ability to learn on their own is called machine learning (ML). A subset of ML is deep learning where algorithms train themselves without human guidance. At the core of AI systems sit Artificial Neural Networks (ANNs), modelled on biological neural networks. ANNs comprise layers of weighted parameters, enabling richer, multi-layer data analysis. Parameter training and optimization involves massive amounts of data analysis.

Large AI models, including some LLMs exceeding one trillion parameters, offer enhanced analysis. Training LLMs is very compute intensive, demanding large and distributed data storage and fast, low-latency networking to connect the worker accelerators in large clusters. These accelerators, such as Graphic Processor Units (GPUs) and Tensor Processor Units (TPUs), form the backbone of AI training and inference. Organizing accelerators into nodes and interconnected clusters delivers the computational muscle required for AI processes.

Transformers and sequence processing

Transformers provide the necessary neural network architecture for sequence processing. What makes Transformers unique is their use of attention mechanisms. These mechanisms allow the model to focus on important parts of a long input sequence. The result is more efficient audio classification, speech recognition, language translation, text summarization, image and video analysis, and text/code generation.

AI cluster networking

When dealing with large scale systems and AI clusters, networking emerges as a pivotal consideration in efficient AI computing. AI cluster networking's characteristics include high bandwidth, low latency, and low-processor load. Standard network technologies like Ethernet and InfiniBand (IB) are utilized, as well as proprietary technologies like Nvidia's NVLink. Remote Direct Memory Access (RDMA), embedded in IB and achievable with Ethernet using RDMA Over Converged Ethernet (ROCE), enhances efficiency. AI cluster networking is also capable of very high connection densities—achieving a bandwidth per node of up to 6.4 Tbps.

Hyperscale brings together the requisite power capacity, cooling capabilities, and data storage for large AI computing. Cloud operators are rapidly deploying AI clusters, causing the biggest disruption of data center design in the history of hyperscale computing. Compared to legacy architectures, hyperscale's increased power consumption typically requires smarter use of white space, enhanced cooling techniques, and sparser installments.

Outlook

The pace is accelerating. Much has happened in the AI field since we wrote "The AI Tsunami" in November 2023. Here are just some of the events we have noted:

- Introduction of new large language models; Claude 3 (Anthropic), Mistral 8x7B (Mistral), Sora (Open AI), and Gemini 1.5 (Google)
- Announcement of investments and partnerships; Mistral/Microsoft and Antrhopic/AWS
- AI hardware announcements, Blackwell accelerators and GB superchips (Nvidia), CS3 wafer scale processor (Cerebrus AI), LPU Inference Engine (Groq), Maia 100 (Microsoft) and MTIA (Meta) accelerators
- Regulatory and government actions; EU AI Act and U.S. government restrictions on semiconductors and semiconductor technology

As we look toward the rest of 2024 and beyond, we will see even higher levels of AI cluster construction, along with further rapid evolution of AI hardware, models, applications, and data center design. This will be a time of change, of learning, and of challenge for us all.

For more information about the AI tsunami, and to accelerate your journey into AI and hyperscale thought leadership, view and download our e-book, Hyperscale: The AI Tsunami.