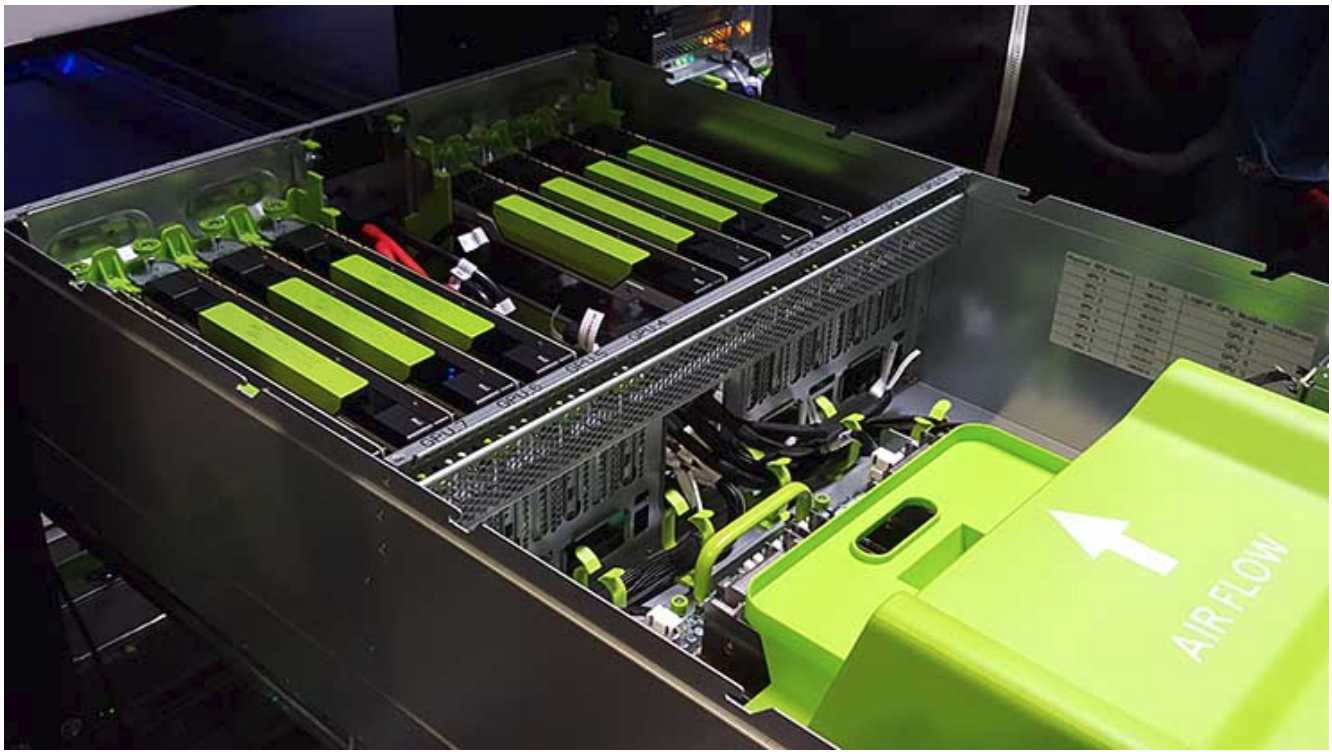
DATA CENTER Frontier

AI Boom Boosts GPU Adoption, High-Density Cooling

BY **RICH MILLER** - APRIL 12, 2017 — LEAVE A COMMENT



A row of eight NVIDIA graphics processing units (GPUs) packed into a Big Sur machine learning server at Facebook's data center in Prineville, Oregon. (Photo: Rich Miller)

The data center team at eBay is plenty familiar with high density data centers. The e-commerce giant has been running racks with more than 30 kilowatts (kW) of power density at the SUPERNAP in Las Vegas, seeking to fill every available slot in racks whenever possible.

But as eBay has begun applying artificial intelligence (AI) to its IT operations, the company has deployed more servers using graphics processing units (GPUs) instead of traditional CPUs.

"From a data center power and cooling perspective, they're a real challenge," said Serena DeVito, an Advanced Data Center Engineer at eBay. "Most data centers are not ready for them. These are really power hungry little boxes."

The rise of artificial intelligence, and the GPU computing hardware that often supports it, is reshaping the data center industry's relationship with power density. New hardware for AI workloads is packing more computing power into each piece of equipment, boosting the power density – the amount of electricity used by servers and storage in a rack or cabinet – and the accompanying heat. The trend is challenging traditional practices in data center cooling, and prompting data center operators to adapt new strategies and designs.

All signs suggest that we are in the early phase of the adoption of AI hardware by data center users. For the moment, the trend is focused on hyperscale players, who are pursuing AI and machine learning at Internet scale. But soon there will be a larger group of companies and industries hoping to integrate AI into their products, and in many cases, their data centers.

The Rise of Accelerated Computing

Amazon Web Services, Microsoft Azure, Google Cloud Platform and IBM all offer GPU cloud servers. Facebook and Microsoft have each developed GPU-accelerated servers for their in-house machine learning operations, while Google went a step further, designing and building its own custom silicon for AI.

"AI is the fastest-growing segment of the data center, but it is still nascent," said Diane Bryant, the Executive VP and General Manager of Intel's Data Center Group. Bryant says that 7 percent of servers sold in 2016 were dedicated for AI workloads. While that is still a small percentage of its business, Intel has invested more than $32 billion in acquisitions of Altera, Nervana and MobilEye to prepare for a world in which specialized computing for AI workloads will become more important.

The appetite for accelerated computing shows up most clearly at NVIDIA, the market leader in GPU computing, which has seen its revenue from data center customers leap 205 percent over the past year. NVIDIA's prowess in parallel processing was seen first in supercomputing and high-performance computing (HPC), and supported by facilities with specialized cooling using water or refrigerants. The arrival of HPC-style density in data centers is driven by the broad application of machine learning technologies.

"Deep learning on Nvidia GPUs, a breakthrough approach to AI, is helping to tackle challenges such as self-driving cars, early cancer detection and weather prediction," said Nvidia cofounder and CEO Jen-Hsun Huang. "We can now see that GPU-based deep learning will revolutionize major industries, from consumer internet and transportation to health care and manufacturing. The era of AI is upon us."

And with the dawn of the AI era comes a rise in rack density, first at the hyperscale players and soon at multi-tenant colocation centers.

Pushing the Power Envelope

How much density are we talking about? "A kilowatt per rack unit is common with these GPUs," said Peter Harrison, the co-founder and Chief Technical Officer at Colovore, a Silicon Valley colocation business that specializes in high-density hosting. "These are real deployments. These customers are pushing to the point where 30kW or 40kW loads (per cabinet) are easily possible today."

A good example is CirraScale, a service provider that specializes in GPU-powered cloud services for AI and machine learning. CirraScale hosts some of its infrastructure in custom high-density cabinets at the ScaleMatrix data center in San Diego.

"These technologies are pushing the envelope," said Chris Orlando, the Chief Sales and Marketing Officer and a co-founder of ScaleMatrix. "We have people from around the country seeking us out because they have dense platforms that are pushing the limits of what their data centers can handle. With densities and workloads changing rapidly, it's hard to see the future."

Cirrascale, the successor to the Verari HPC business, operates several rows of cabinets at ScaleMatrix, which house between 11 and 14 GPU servers per cabinet, including some connecting eight NVIDIA GPUs using PCIe – a configuration also seen in Facebook's Big Sur AI appliance and the NVIDIA DGX-1 "supercomputer in a box."

Strategies for Managing Extreme Density

Over the past decade, there have been numerous predictions of the imminent arrival of higher rack power densities. Yet extreme densities remain limited, primarily seen in HPC. The

consensus view is that most data centers average 3kW to 6kW a rack, with hyperscale facilities running at about 10kW per rack.

Yet the interest in AI extends beyond the HPC environments at universities and research labs, bringing these workloads into cloud data centers. Service providers specializing in high-density computing have also seen growing business from machine learning and AI workloads. These companies use different strategies and designs to cool high-density cabinets.

A TSCIF aisle containment system inside the SUPERNAP campus in Las Vegas. (Photo: Switch)

The primary strategy is containment, which creates a physical separation between cold air and hot air in the data hall. One of the pioneers in containment has been Switch, whose SUPERNAP data centers use a hot-aisle containment system to handle workloads of 30kW a rack and beyond. This capability has won the business of many large customers, allowing them to pack more computing power into a smaller footprint. Prominent customers include eBay, with its historic focus on density, which hosts its GPU-powered AI hardware at the SUPERNAPs in Las Vegas.

For hyperscale operators, data center economics dictates a middle path on the density spectrum. Facebook, Google and Microsoft operate their data centers at higher temperatures, often above 80 degrees in the cold aisle. This saves money on power and cooling, but those

higher temperatures make it difficult to manage HPC-style density. Facebook, for example, seeks to keep racks around 10 kW, so it runs just four of its Big Sur and Big Basin AI servers in each rack. The units are each 3U in depth.

AI Challenges Extend Beyond GPUs

Facebook's machine learning servers feature eight NVIDIA GPUs, which the custom chassis design places directly in front of the cool air being drawn into the system, removing preheat from other components and improving the overall thermal efficiency. Microsoft's HGX-1 machine learning server, developed with NVIDIA and Ingrasys/Foxconn, also features eight GPUs.

While much of the AI density discussion has focused on NVIDIA gear, GPUs aren't the only hardware being adopted for artificial intelligence computing, and just about all of these chips result in higher power densities.

A custom rack in a Google data center packed with Tensor Processing Unit hardware for machine learning. (Photo: Google)

Google decided to design and build its own AI hardware centered on the Tensor Processing Unit (TPU), a custom ASIC tailored for Google's TensorFlow open source software library for machine learning. An ASIC (Application Specific Integrated Circuits) is a chip that can be customized to perform a specific task, squeezing more operations per second into the silicon. A board with a TPU fits into a hard disk drive slot in a data center rack.

"Those TPUs are more energy dense than a traditional x86 server," said Joe Kava, the Vice

President of Data Centers at Google. "If you have a full rack of TPUs, it will draw more power than a traditional rack. It hasn't really changed anything for us. We have the ability in our data center design to adapt for higher density. As a percentage of the total fleet, it's not a majority of our (hardware)."

About Rich Miller

I write about the places where the Internet lives, telling the story of data centers and the people who build them. I founded Data Center Knowledge, the data center industry's leading news site. Now I'm exploring the future of cloud computing at Data Center Frontier.

ABOUT US

Charting the future of data centers and cloud computing. We write about what's next for the Internet, and the innovations that will take us there. We tell the story of the digital economy through the data center facilities that power cloud computing and the people who build them. **Read more ...**

ABOUT OUR FOUNDER

Data Center Frontier is edited by Rich Miller, the data center industry's most experienced journalist. For more than 15 years, Rich has profiled the key role played by data centers in the Internet revolution. **Meet the DCF team**.

TOPICS

Cloud

Colo

Cooling

Design

Energy

Executive Roundtable

Featured

Internet of Things

Machine Learning

Network