BOXX AI Acceleration Platforms

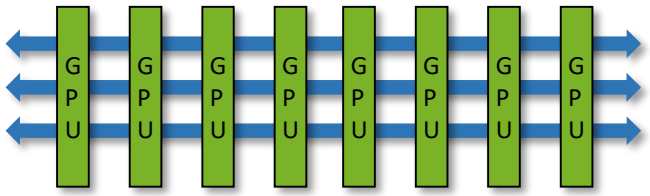
GX8 Multi-GPU AI Acceleration Server

Ultra-dense platform for AI Model Training

AI is expected to deliver superlative value in many data-driven industries, but connecting with that value can be elusive because of the substantial experience required to assemble a high-performing AI development platform. The defining step that shifts an AI program from experimentation to creating real business value is the move to a **multi-GPU platform**. The BOXX GX8 leverages the scalability and enterprise-class bandwidth of cutting edge data center technologies to radically scale the number of GPUs running in tandem in a single server. Minimized bus latencies, large system memory, and well matched CPUs combine with **up to eight GPU cards** to provide a powerful, very dense and highly cost effective AI development platform. Connect with BOXX AI infrastructure experts to learn more about how the BOXX GX8 can **shorten your time-to-model**.

Purpose-built Server for AI Development

High-bandwidth GPU-to-GPU Interconnect Fabric



Up to 8x NVIDIA V100 or Titan V GPU Cards working as one

Maximum AI performance in an ultra-dense format

In AI development, higher processing performance translates directly into faster time-to-market for organizations focused on AI. It results in higher quality models faster for Natural Language Processing, self-driving vehicles, and medical imagery analysis. Whether you are running Tensorflow, MXNET, Caffe2 or PyTorch, or using pretrained DL models like AlexNet or Inception, the BOXX GX8 delivers maximum power to accelerate time-to-model.

- Accelerates AI development with up to 8 NVIDIA Tesla or Quadro GPUs running in parallel
- Maximizes GPU throughput by minimizing internal bottlenecks with data center-class system design
- Minimizes data input delays with top-of-the-line AMD EPYC CPUs and dual Intel Xeon configurations
- Specifically designed to maximize performance with popular deep learning containers
- Supported by a team that understands AI research at any scale: from exploratory projects to industry-leading model training programs.

PYTORCH mxnet Inception-v3 Pretrained Model

TensorFlow Caffe2 AlexNet PRETRAINED MODEL

"Single Root" Architecture: The Key to Performance

System topology matters when maximizing the efficiency of AI development. Higher bandwidth between GPUs reduces latencies for increased training effectiveness per hour of iteration. In that respect, the BOXX GX8 features the most advanced version of the PCIe interconnect fabric with 96 lanes of high-speed GPU-to-GPU connectivity. General purpose computing platforms may claim a substantial number of lanes electrically but these connections do not have the GPU-to-GPU transfer bandwidth required to take advantage of the full throughput of each GPU and get the most out of a cluster of up to 8 GPUs:

- The GX8 leverages multi-lane PCIe Gen3 performance designed to enable next-gen parallel computing architectures for AI and HPC.
- High performance switching architecture supports multiple hosts in a single root configuration
- Scalable, high bandwidth fabric enables multiple clustered GPUs.
- Best price/performance ratio vs. other advanced interconnection fabrics like InfiniBand or Ethernet.

Powered by NVIDIA Tesla & Quadro GPUs

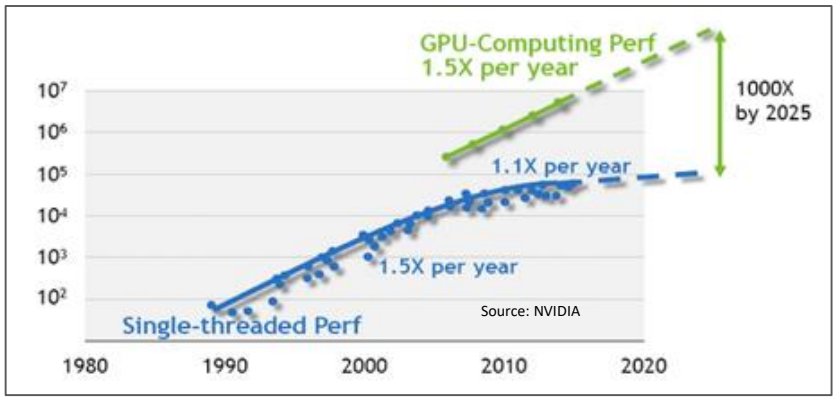
NVIDIA

Talk to a BOXX AI infrastructure expert to define the right GX8 for your specific type of AI development.

Multi-GPU AI Acceleration Server

GX8: the right mix of processing technologies for training AI models

Achieving the right system configuration is key to maximizing the throughput with your chosen machine learning container. While all machine learning projects benefit from more GPU chips and more cores, Natural Language Processing and Image Classification, for example, have different performance dependencies. **A BOXX/Cirrascale Cloud Services AI infrastructure expert** can help you achieve the best configuration for your data science program.

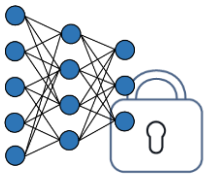


Data Input Phase:
When CPU Performance matters

Although AI model training occurs mostly in the GPU, loading the data and preparing it for processing is a key preliminary step which is highly dependent on the performance of the CPU. The GX8 features top-of-the-line AMD EPYC or Dual Xeon processors with sufficient CPU cycles to minimize the bottlenecks that are created by a complex data input pipeline.

AI Researchers and Data Scientists can Rely on Our Unique Experience

BOXX & Cirrascale Cloud Services have extensive experience in engineering high-performance distributed computing platforms based on CPUs *and* GPUs, from computing clusters for mass rendering of high-resolution digital images, to advanced parametric modelling for the AEC space, to cloud-based AI development platforms.



Data and Infrastructure Security

For many AI focused organizations, maintaining the confidentiality of the data and of the proprietary development processes is of paramount importance. The BOXX GX8 is the ideal building block of a both optimized and secure AI program. Compact, dense and scalable, it can be secured through best-of-breed data center security protocols incl.: Multifactor authentication, Data encryption,…

BOXX and Cirrascale Cloud Services Combine Advanced Solution Design for AI/ML/DL

BOXX and its sister company Cirrascale Cloud Services combine their deep expertise in high performance AI development platforms: on-premise - deskside and server-based- and also private cloud-based AI development infrastructure on an as-a-service basis.

Our commitment to a Sustainable Future

AI development will mobilize a large portion of worldwide computing resources putting pressure on carbon footprints, so power Consumption is an important factor to consider in AI Research. BOXX is committed to achieving the best possible training and inferencing performance per Watt consumed by its AI platforms. BOXX dedicates a substantial portion of its research effort to benchmarking the performance and power consumption of its distributed computing platforms for both AI and HPC.

To learn more about BOXX/Cirrascale Cloud Services private cloud infrastructure for AI development, call: 888 942-3800