PRODUCT BRIEF

# OPTIONS FOR ACCESSING PCIE GPUS IN A HIGH PERFORMANCE SERVER ARCHITECTURE

*Understanding Configuration Options for Supermicro GPU Servers Delivers Maximum Performance for Workloads*

TABLE OF CONTENTS



## Executive Summary


GPU servers offer a tremendous benefit in terms of performance for AI and HPC applications compared to a traditional CPU only server. A wide range of applications can be executed on these systems, and the performance increase for applications that take advantage of the GPUs has been widely documented. While GPU focused servers contain single or dual CPUs and up to 10 PCIe GPUs, how the system is architected can impact the application speed and flexibility of the server. There are three ways to design a GPU server, resulting in a more optimized system for various workloads. The data flow between the CPU and GPUs is crucial when choosing a GPU server.


## PCIE GPU Access Choices

Supermicro GPU servers are designed for applications requiring several GPUs within the server. Although many servers can handle a 1:1 ratio of CPUs to GPUs through a PCIe slot, servers designed for high acceleration require a ratio with significantly more GPUs than CPUs. GPU servers are available in two general architectures:

- PCIe based GPUs where up to 10 GPUs are installed in PCIe slots

- SXM/OAM based GPUs servers where the GPUs are mounted on their own board and have only 1 PCIe connection to the CPUs.

Most GPU servers have two CPU sockets, with DRAM memory attached to each socket. The CPUs communicate via high speed communication paths (UPI for Intel based systems and xGMI for AMD based systems).

Digging further into the PCIe-based servers, three distinct system architectures are designed for various workloads.

- Single Root
- Dual Root
- Direct Attach

## Single Root Option Explained

The single root architecture is ideal for applications that reside on a single CPU but require access to multiple GPUs. A single root system dedicates one of the CPUs (out of two) to manage all communications with the GPUs. As shown in Figure 1, the CPU that communicates with the GPUs does so through a PCI switch (PLX). Each PLX switch is connected to the CPU via 2 PCIe x16 lanes and then can communicate up to five double-width GPUs. This results in using a maximum of 10 GPUs in a single server. A single root system is tailored for deep learning applications where most of the computation takes place on the GPU.

Advantages of a Single Root configuration:

- A single CPU has access to up to 10 GPUs. Applications that need direct access to all of the GPUs will benefit from this configuration.

Applications – When peer-to-peer communication (GPU to GPU) performance is not critical.

The general configuration of a Single Root system is shown in Figure 1.

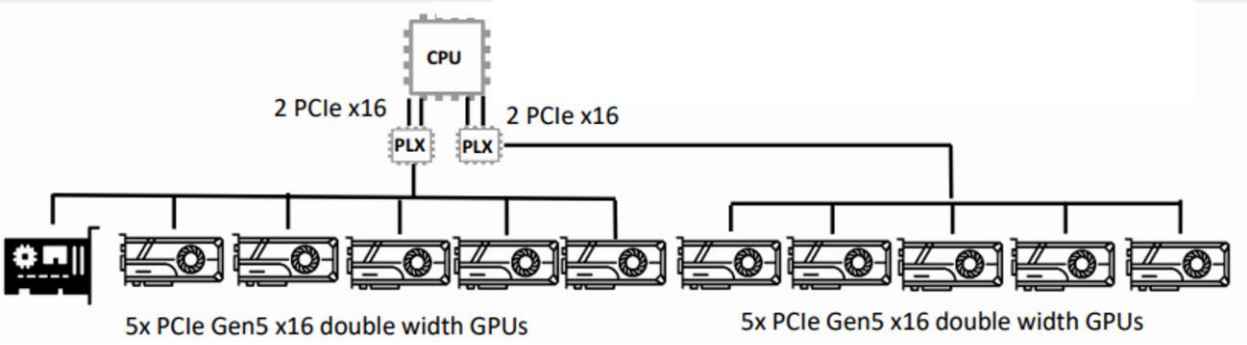


*Figure 1 - Single Root Schematic*

## Dual Root Option Explained

A dual root setup connects each CPU to several GPUs through a PLX switch. Currently, the maximum number of GPUs that can be addressed in total is 10. The distribution of GPUs attached through the PLX switch does not have to be equal per CPU, as a workload(s) assigned to a system may not be easily distributed among the CPUs. Each CPU can easily communicate with each

other, and the combinations of PCIe devices attached to each PLX switch can be very flexible. In the figure below, each CPU (and PLX switch) has 4 GPUs, 2 AOC cards, and 4 NVMe storage devices. This type of system is the most common configuration for Omniverse environments. This configuration will benefit applications that are balanced between the CPU and GPU.

Advantages of a Dual Root configuration:

- Workloads can be assigned to a CPU, each with up to 10 total GPUs or other devices accessible through a PLX switch.

Applications – Where data needs to be shared and communicated between the two CPUs.This balanced system enables the CPUs to communicate efficiently with GPUs, networking cards, or storage devices. Examples include AI Compute/Model Training/Deep Learning and High-performance Computing (HPC).

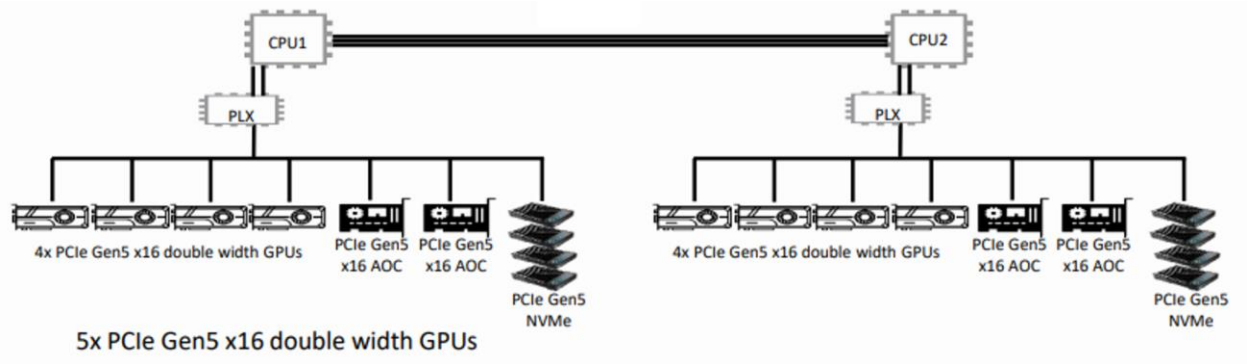


*Figure 2 - Dual Root Schematic*

## Direct Attached Option Explained

In a Direct Attached setup, each of the CPUs has direct PCIe access up to four full size GPUs, for a total of eight per system. The benefit of this configuration is that no PLX switches are required, and each CPU has a direct connection to four GPUs. A direct attached setup is most common for HPC applications. In this case, the PLX chip, while allowing for more PCIe devices, may increase latencies between the CPUs and the devices.

Advantages:

- Each application running on a CPU has access to four GPUs.

- Each CPU has equal access to GPUs and I/O capabilities.

Applications – Excellent for computing environments where more than one application may run concurrently, or a single application can be divided or separated and assigned to different CPUs.

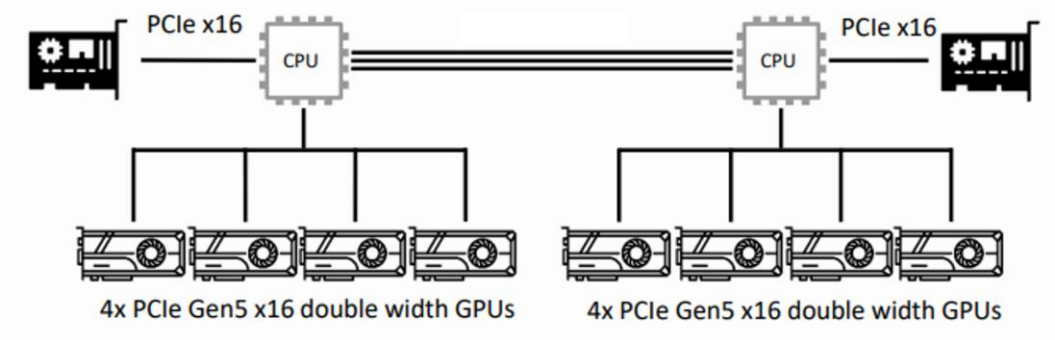


*Figure 3 - Direct Attached Schematic*

## Supermicro GPU 4U/5U Systems with PCIE GPU Options

| Configuration | Supermicro X12/X13 Servers intel | Supermicro H13 Servers AMD |
|---|---|---|
| Single-Root | SYS-421GE-TNRT2 (Coming Soon) | AS -4125GS-TNRT1 |
| Dual-Root | SYS-521GE-TNRT<br>SYS-421GE-TNRT (OVX - Omniverse Optimized)<br>SYS-420GP-TNR | AS -4125GS-TNRT2 |
| Direct Attached | SYS-421GE-TNRT3 | AS -4125GS-TNRT |

## PCIE GPU Support Options

The Supermicro X13 GPU Servers support the following GPUs in the above configurations:

- NVIDIA® H100 Tensor Core GPU PCIe form factor

The Supermicro H13 GPU Servers support the following GPUs in the above configurations:

- NVIDIA® H100 Tensor Core GPU PCIe form factor
- AMD Instinct™ MI200 Series

## Summary

Depending on the application workload, different configurations of GPU servers can be obtained from Supermicro. Whether Direct Attached, Single Root, or Dual Root, applications will perform well if the proper combinations of CPUs and GPUs are selected. Users must understand the differences and match their workloads to the servers.

For more information about Supermicro's GPU servers, visit: https://www.supermicro.com/en/products/gpu

For more information about Supermicro's NVIDIA solution portfolio, visit: https://www.supermicro.com/en/accelerators/nvidia