insideHPC

*insideHPC Special Research Report*

# Practical Hardware Design Strategies for Modern HPC Workloads

*by Douglas Eadline*

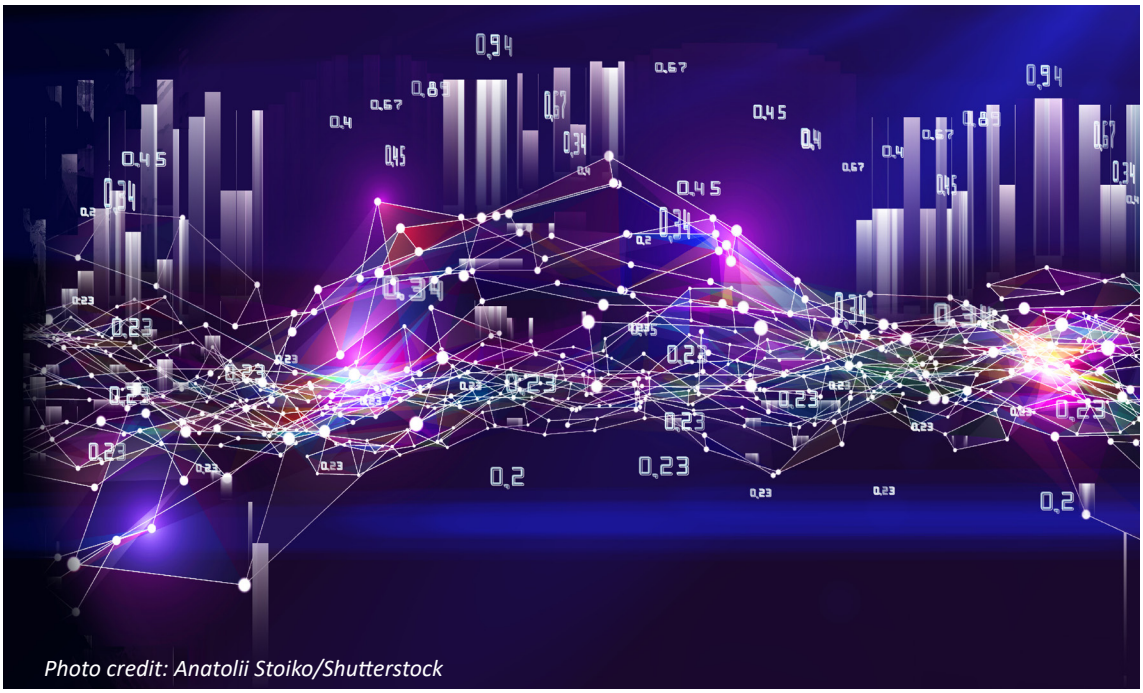

*Photo credit: Anatolii Stoiko/Shutterstock*

BROUGHT TO YOU BY TYAN

## Contents



# Executive Summary

Many new technologies used in High Performance Computing (HPC) have allowed new application areas to become possible. Advances like multi-core, GPU, NVMe, and others have created application verticals that include accelerator assisted HPC, GPU based Deep Learning, Fast storage and parallel file systems, and Big Data Analytics systems.

The verticals can be broken into three general design types:

1. **Accelerated HPC Computation –** includes both traditional HPC and Deep Learning systems
2. **IO-Heavy HPC Computing –** includes systems that provide fast NVMe implementations for local IO or as part of a parallel file system
3. **Big Data (Database) Computing –** includes system designed for high density bulk storage of large amounts of data

Depending on application needs, the Thunder HX FT83-B7119 can handle both HPC and Deep Leaning applications and is available in four versions based on PCIe bus routing topology and storage options.

Various design goals and a discussion of balanced vs. centralized PCIe topology are explained.

IO-Heavy applications should consider solid state U.2 connected NVMe devices that provide up to 4GB/s of throughput. An excellent starting point for an IO-Heavy computing systems is the TYAN Thunder SX GT62H-B7106 platform.

Big Data (and database) computing requires both high performance and bulk storage using 3.5 inch spinning disk drives. The TYAN Thunder SX GT93-B7106 chassis provides a solid platform to create or grow a Big Data computing systems.

In terms of accelerated HPC computing, the TYAN Thunder HX FT83-B7119 is a 10-GPU supercomputing system in a compact 4U rack-mount chassis. Depending on application needs, the Thunder HX FT83-B7119 can handle both HPC and Deep Leaning applications and is available in four versions based on PCIe bus routing topology and storage options.

# Introduction and Background

While High Performance Computing (HPC) hardware and software have become much more turn-key than in the past, the choice of hardware is still important for optimal performance. Traditional clustered HPC systems have been built from off-the-shelf servers using x86 processors and high speed networks. End users often contributed to designs based on their application needs. For instance, systems designed on a fixed budget had to strike a balance between numbers of processors (cores), memory, and quality of the interconnect. A high performance interconnect, such as InfiniBand was more expensive than traditional Ethernet (usually 10 GbE) and thus inclusion reduced the number of servers in the cluster (or the cores and/or amounts of memory per server). Often times, a balanced system was designed that would provide reasonable performance across the spectrum of user applications. This approach generally worked and many applications were successfully deployed on these systems.

As hardware continued to develop, technologies like multi-core, GPU, NVMe, and others have allowed new application areas to become possible. These application areas include accelerator assisted HPC, GPU based Deep learning, and Big Data Analytics systems. Unfortunately, implementing a general purpose balanced system solution is not possible for these applications. To achieve the best price-to-performance in each of these application verticals, attention to hardware features and design is most important.

# Differentiation in Modern HPC Workloads

Most HPC traditional workloads consist of "number crunching" whereby large amounts of floating point calculations are run to simulate or model complicated processes. These can include, materials and molecular systems, weather forecast and astronomy, fluid dynamics, financial markets, oil and gas, physics, bioscience, and many others. All of these share the need for large amounts of calculations in a reasonable amount of time (i.e. there is no use in trying to predict tomorrow's weather if it takes two days for the model to run).

> Many traditional HPC practitioners are turning to Big Data Analytics and Deep Learning to further their understanding of many natural phenomena.

Many HPC applications can be considered "compute bound," where the limiting step is how much compute performance they can provide over time. There are other applications that are IO bound, where the amount of disk IO can be a limiting factor.

There are other types of applications that have overlap in the HPC market. These include Big Data Analytics and Deep Learning. It can be argued that these two areas are not strictly "HPC," but since a clear definition of what constitutes an "HPC" problem is rather vague and both of these application areas seek to increase performance by adding more hardware, it seems reasonable to include them as part of the high performance ecosystem. Indeed, many traditional HPC practitioners are turning to Big Data Analytics and Deep Learning to further their understanding of many natural phenomena.

## Accelerated HPC Computation

Historically, many of HPC simulations and models were distributed across clustered servers (also called "nodes") that worked in concert to produce a solution. These types of applications are often considered "compute bound," because the amount of computation is the limiting factor in application progress. In many instances, adding more servers (CPU/memory resources) allowed the problem to be scaled as more computation was required. The ability to scale a problem size often comes with some limitations (due to the need to move data and the nature of the problem at hand) and at some point begins to level off (i.e. applications do not get any faster when adding more servers).

The use of GPU based accelerators has become a popular way of increasing performance on computational nodes.

One way around this limitation is to increase the computation rate on the server. While 2nd Gen Intel® Xeon® Scalable Processors have shown a steading increase in performance and core counts, specialized accelerator processors have gained favor in recent years. Most notably the use of GPU based accelerators has become a popular way of increasing performance on computational nodes.

The GPU's main advantage is the ability to perform a single instruction across large amounts of data at the same time. This type of operation is common in graphics applications and occurs in many HPC applications as well — particularly array operations (linear algebra). Typically, the host Intel Xeon Scalable Processor provides the basic computing platform (large memory, computational cores, operating system, IO, networking, etc.) and uses one or more GPUs as accelerators for certain parallel operations.

## IO-Heavy HPC Computing

This level of computing usually requires large amounts of data to be read and written to disk as part of the computation. These types of applications are often considered "IO bound" problems because the speed at which data can be read/written to disk defines the performance.

The speed of storage IO on the servers used to create the filesystem have a large influence on how well the filesystem can perform.

For example, an application may need to write temporary intermediate files during the course of the program because the amount of data in these files is too large to keep in memory. Note that even "compute bound" applications can become IO bound at times if they are writing checkpoint/restart files during the course of the application.

These types of applications can benefit from things like local NVMe storage where storage is directly accessible to the processor and does not have to traverse a network.

Using nodes fast IO also comes into play when building out fast parallel file systems. In this case, the speed of storage IO on the servers used to create the filesystem have a large influence on how well the filesystem can perform. Examples of these files systems may include Lustre, Gluster, and Ceph.

## Big Data Computing

Big Data computing is similar to High-IO computing; however, the goal is both speed and bulk storage. With Big Data computing, the need is more focused on distributed bulk storage than single node performance.

The speed of storage IO on the servers used to create the filesystem have a large influence on how well the filesystem can perform.

For instance, applications like Hadoop-Hive, Spark, and the Hadoop Distributed File Systems (HDFS) are a popular platform on which to build Big Data solutions. Typically HDFS is used to manage the large amounts of data used for these analytics operations. Due to the sheer size of the data, backing up the data is nearly impossible. In this case, the HDFS filesystem has built-in redundancy so that if one or two servers fail (at the same time), the file system continues to operate. In addition, HDFS is designed to easily scale-up (adding more storage to a storage server) and scale-out (adding more storage servers).

In addition to Hadoop, NoSQL databases rely on dense storage nodes. Similar to HDFS, these databases are designed to use multiple servers, offer redundant performance, and provide a more flexible column-based storage mechanism versus a more traditional row-based SQL transactional SQL database.

In both of these cases, storage nodes need large amount of bulk storage and typically employ 3.5 inch (spinning) disks due to the greater capacity and lower cost per byte than solid state drives.

### Deep Learning

The final type of HPC workload is very similar to HPC compute bound applications, however, the type of computation, linear algebra, is focused on one type of learning problem. These types of problems can require massive amounts of computation to provide usable results (i.e. the learning model can successfully predict a certain percentage of outcomes from new data).

The amount of computation for Deep Learning is quite large and multiple GPUs are often dedicated to a single problem. When running Deep Learning applications, all data are normally loaded onto the GPU and less dependence on CPU memory transfer is needed, however, for large models, Deep Learning can benefit from local caching of model epochs (learning steps).

> Quite often the limiting step in GPU accelerated HPC codes is data movement to and from the GPU (over the PCI bus).

Servers designed for Deep Learning have a definite overlap with accelerated HPC computing mentioned above. Quite often the limiting step in GPU accelerated HPC codes is data movement to and from the GPU (over the PCI bus). For this reason, there can be a point where adding GPUs does not increase the performance of HPC applications.

## Workload Design Strategies

Given the variety of today's HPC workloads, it is important to understand how this maps out to actual hardware platforms. Before considering specific hardware, there is an important design aspect that system designers and users should consider. The PCIe bus is designed for Intel Xeon Scalable systems can be implemented in two ways. Both have advantages and disadvantages depending on your workload.

### Balanced vs. Centralized PCIe Topology

Traditional servers have two processors (sockets) with multiple cores and additional memory channels. Memory attached to each processor is shared with the other processor across high speed links. On Intel platforms, these links are called Ultra Path Interconnect (UPI) or Intel QuickPath Interconnect (QPI). In a similar fashion, each processor has a certain number of PCIe bus connections (lanes) that may be shared in one of two ways:

- The first method is the balanced PCIe topology, where a device on the PCI bus may need to traverse the inter-processor links when it is accessed by the processor that is not providing the actual PCIe lanes. This design can create two levels of PCIe access, direct and over the inter-processor link.
- The second method is the centralized PCIe topology, which connects both PCIe buses using a PCIe switch to one of the processors. The unified PCIe bus does not require the inter-processor links and all PCIe access (and device speeds) are consistent to a single processor.

Depending on how the server is used, both methods may have advantages and disadvantages:

- First, consider the centralized PCIe topology, this approach is used in many Deep Learning systems because it allows fast memory movement/access from one GPU to another without the need to traverse the inter-processor links.
- In addition to GPU performance advantages, this architecture can use a lower-end Intel CPU SKU processor with slower UPI/QPI speed at lower cost.
- On the other hand, the balanced PCIe topology for multiple GPU Deep Learning applications creates an uneven GPU-to-GPU memory movement environment that may adversely affect performance.

While the "balanced-root" design may seem less advantageous than centralized-root systems, there are times when separating the PCIe devices may have an advantage. To be clear, all PCIe devices are still visible by both processors. Access to the device may have to travel over the UPI/QPI links. Splitting the "bus" can have an advantage when using GPUs or high speed (NVMe) storage devices to accelerate HPC applications. Since HPC application often relies more heavily on CPU to GPU memory movement or CPU to a storage device, separate PCIe connections can be used to improve performance. There may be cases where assigning processor exclusive access to a subset of GPUs or storage devices can lead to better performance (i.e. By its nature the PCIe bus is a shared bus and segmenting the bus can help with traffic mitigation when multiple applications are using the same server).

## Servers for High-IO HPC Computing

Choosing a server for IO bound requires the fastest available storage devices. Currently, solid state U.2 connected NVMe devices can use 4 PCIe lanes and provide a speed of 4GB/s of throughput. Of course the NVMe drive specifications will determine the final performance, but the use of NVMe ensures the fastest connection.

The Thunder SX GT62H-B7106 design allows for processes running on each CPU to have direct access to high speed networking and local NVMe storage, resulting in the fastest possible access times and the lowest possible latency.

As shown in Figure 1, an excellent starting point for a High-IO computing system is the TYAN Thunder SX GT62H-B7106 platform. As part of TYAN's leading Intel Xeon Scalable Processor-based storage product line, this 1U server provides many important features for high-IO computing including ten NVMe U.2 drive bays, dual Intel Xeon Scable

*Figure 1: Thunder SX GT62H-B7106 - 1U All-Flash Dual Socket Storage Server*

Processor sockets, large memory capacity, 2 PCIe x16 slots for high performance NICs, IPMI with Redfish support, and (1+1) 800W redundant power supplies. The SX GT62H-B7106 is a good building block for both single node and multi-node storage systems, including parallel file systems and software defined storage.

In addition to the NVMe support, the Thunder SX GT62H-B7106 presents a balanced system where each CPU socket gets its own PCIe x16 slot as well as a handful of NVMe drives (One x16 and 4 NVMe bays for CPU0, and one x16 slot and 6 NVMe bays for CPU1). This design allows for processes running on each CPU to have direct access to high speed networking and local NVMe storage, resulting in the fastest possible access times and the lowest possible latency.



 

## Servers for Big Data Computing

As mentioned Big Data (and database) computing requires both high performance and bulk storage. The best cost-per-byte of storage is still with the 3.5-inch spinning disk drives. The TYAN Thunder SX GT93-B7106 chassis provides a solid platform to create or grow a Big Data computing system. Featuring dual socket Intel Xeon Scalable Processor support with up to 2TB of DDR4-2933 memory, twelve (12) internal easy-swap 3.5" SATA 6G drive bays, one internal 2.5" SATA 6G drive bay, one PCIe x16 OCP (Open Compute Project) v2.0 dual-port LAN Mezzanine slot, one Low Profile PCIe x16 slot, and (1+1) 650W redundant power supplies.

*Figure 2: Thunder SX GT93-B7106 - 1U Density Storage Server with 12 Internal 3.5" SATA 6G Drive Bays and dual socket Intel Xeon Scalable Processor*

The SX GT93-B7106 has the prefect "simple" design that allows Hadoop, Spark, and NoSQL database systems to deliver solutions using large amounts of data and scale-out or scale-up as needed. Using today's large 16 TByte drives, a single SX GT93-B7106 has the ability to deliver 192 TBytes of raw storage in a compact 1U platform. By providing options for large amounts of memory and dual Intel Xeon Cascade Lake Refresh CPU, the SX GT93-B7106 can be configured to fit users' needs. The optional OCP dual-port LAN Mezzanine slot allows for high speed networking to be added if needed to this versatile storage server.

## Servers for HPC and Deep Learning Computation

The TYAN Thunder HX FT83-B7119 is a 10-GPU supercomputing system in a compact 4U rack-mount chassis. When configured it can support GPU assisted HPC jobs and/or Deep Learning applications.

The base motherboard provides dual sockets for 2nd Gen Intel Xeon Scalable Processors, up to 3TB DDR4-2933 memory, either twelve 3.5" SATA 6G bays or eight SATA plus four NVMe U.2 bays, ten double-width PCIe x16 slots for GPUs, a PCIe x16 slot for a high performance NIC, a BMC with Redfish support, and (3+1) 4800W redundant power supplies.

Depending on application needs, the HX FT83-B7119 is available in four versions based on PCIe bus routing topology and storage options. As outlined in Table 1, each system comes with twelve 3.5 inch SATA bays and two models provide NVMe U.2 support for fast IO in four of the bays (All systems have twelve bays total). Each of these storage options is variable in two different CPU-GPU link topologies.

For multi-GPU Deep Learning, all GPUs connected to a single CPU should provide the best performance. In addition, the four bays of NVMe storage are also a good choice because large models may need fast storage for intermediate results.

In terms of HPC, all GPUs evenly connected to two CPUs may provide a better solution if multiple GPU based jobs are run on the system. This configuration

| Model | Storage Options | CPU-GPU Link Topology | Pre-installed NVMe Adapter |
|---|---|---|---|
| B7119F83V12HR-2T-NS | (12) 3.5" SATA 6G | All GPUs connected to a single CPU | - |
| B7119F83V12HR-2T-N | (12) 3.5" SATA 6G | All GPUs evenly connected to 2 CPUs | - |
| B7119F83V8E4HR-2T-NS | (12) 3.5" SATA 6G w/ (4) NVMe U.2 support | All GPUs connected to a single CPU | - |
| B7119F83V8E4HR-2T-N | (12) 3.5" SATA 6G w/ (4) NVMe U.2 support | All GPUs evenly connected to 2 CPUs | (1) M7106-4E |

*Table 1: Thunder HX FT83-B7119 Models*



 

will allow multiple conversations to occur over the split PCIe bus at the same time. Recall that memory movement is most important when running HPC applications on CPU-GPU systems.

As interest in Deep Learning continues to increase in traditional HPC, the HX FT83-B7119 is a good choice for both types of applications. With support for Intel Xeon Scalable Processors, a large memory footprint, and plenty of storage, the HX FT83-B7119 provides a great base for many HPC and Deep Learning applications. There is no reason why HPC and Deep Learning applications cannot operate on the same data within the same hardware platform. Figure 3 provides an image of the Tyan Thunder HX FT83-B7119.

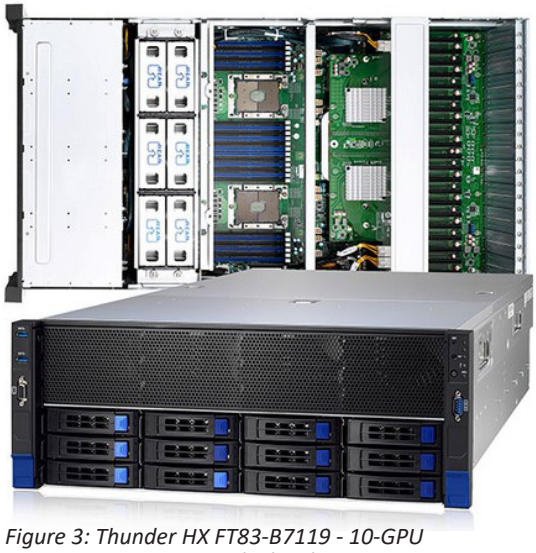

*Figure 3: Thunder HX FT83-B7119 - 10-GPU Server Platform for HPC/AI/ML/DNN workloads*

## Conclusion

In the past, HPC system design was focused on core counts, memory size, and networking. Modern high performance systems can be broken to three basic categories.

1. Parallel file systems like Lustre, Gluster, and Ceph require fast balanced storage. In addition, local IO-Heavy computing can take advantage of fast NVMe.
2. The need for high density bulk storage of data continues in the Data Analytics market. These systems include Hadoop/Spark and scalable noSQL systems.
3. Applications that require accelerated HPC computation include areas such as materials and molecular science, weather forecasting and astronomy, fluid dynamics, financial engineering, oil and gas exploration, pharmacology, and many others. This category also includes Deep Learning systems that are currently seeing high usage in many fields.

Designing for these types of systems requires a thorough understanding of your application space. The following recommendations will provide optimum performance within a given application vertical.

- IO-Heavy applications such as parallel file systems of local IO nodes should consider systems that provide balanced IO from solid state NVMe devices. Consider leading edge systems like the TYAN Thunder SX GT62H-B7106 platform that provides ten balanced NVMe U.2 drive bays, dual socket 2nd Gen Intel Xeon Scalable Processors, and large memory capacity in compact 1U rack-mount system.
- Big Data (and database) computing requires both high performance and bulk storage with spinning disk drives. The TYAN Thunder SX GT93-B7106 chassis provides a solid platform to create or grow a Big Data computing systems with dual socket 2nd Gen Intel Xeon Scalable Processor, up to 2TB of DDR4-2933 memory, and twelve (12) internal easy-swap 3.5” SATA 6G drive bays in a compact 1U rack-mount system.
- In terms of accelerated HPC computing, the TYAN Thunder HX FT83-B7119 is a good choice for a 10-GPU supercomputing system for both HPC and Deep Learning applications. The base system provides dual-socket 2nd Gen Intel Xeon Scalable Processors, up to 3TB of memory, either twelve 3.5” SATA 6G bays or eight SATA plus four NVMe U.2 bays, and the ability to support ten GPUs. The Thunder HX FT83-B7119 provides options to select a balanced or centralized PCIe topology.

In addition to the above mentioned models, Tyan also offers a complete line of leading edge server chassis to build systems optimized for your workload.



 