Accelerating high-performance computing with hybrid platforms

October 2010

Contents

1.0 Introduction

Despite steady advancements in processing speed, many complex computational problems require even greater computational power to be tractable. Also, in light of recent technological advances, researchers are reexamining problems previously limited in either scale or scope by technologies available at the time. Given these realities, it is not surprising that high-performance computing (HPC) is constantly looking to advance computing capabilities. As they do so, the HPC community struggles with many forces, including (1) the increasing programming complexity of problems that require Peta-scale and in the future Exa-scale computing resources, (2) the infrastructure cost and operational complexity of these HPC deployments, and (3) the need for simpler and more agile forms of accessing available compute power.

The advent of *compute accelerators*, special purpose coprocessors that significantly improve the performance of traditional host-based central processing unit (CPU) computations, represents a major transition in the tools available to achieve supercomputing power. One such class of accelerators is graphics processing units (GPUs) used for general purpose computing, or GPU Compute solutions. The very desirable price/performance of CPU/GPU-accelerated computing and the ease of new programming models make GPUs an attractive compute accelerator for the increasingly powerful CPU offerings in the market.

This paper assumes the reader is familiar with the use of GPUs in graphics processing and has an elementary understanding of general purpose GPU acceleration. The structure of this paper is as follows: first, it discusses HPC acceleration with GPUs as well as the computational characteristics of applications and algorithms that benefit from this type of compute acceleration; next, it presents the current state of the industry in terms of GPU offerings and GPU-CPU integration techniques; finally, it presents Dell solution offerings and shows how they push the envelope of available solutions that bring immediate added value to HPC users in terms of integration, flexibility, and efficiency of hybrid compute solutions.

2.0 The promise of general purpose computing with GPUs

Compute accelerators have been available in one form or another for a long time. Early floating point coprocessors accelerated and improved the precision of floating point computations, while GPUs have been used for years as dedicated processors that accelerate graphics stream computations.

Modern-day accelerators exploit parallelism and, in some cases, special-purpose instruction set architectures (ISAs) to improve computational performance. They may implement different models of parallelism, ranging from multiple-instruction multiple-data stream (MIMD) models, to single-instruction multiple-data stream (SIMD) models, or even to multiple single-instruction multiple stream (MSIMD) models, as is the case for present-day GPUs.

A GPU is a massively parallel and fully programmable many-core processor where each core has very high-speed local memory and shares a common global memory with other on-chip cores and with host CPUs (see Figure 1). The instruction set architecture (ISA) of a GPU is optimized for carrying out parallel floating-point computations, essential for 3D graphics rendering. The potential benefit of GPUs to HPC has become evident with the realization that small changes to GPU architecture make these processors suitable not only for graphics stream calculations, but also for general purpose computation. Recent advances in GPU technology confirm this to be the case.

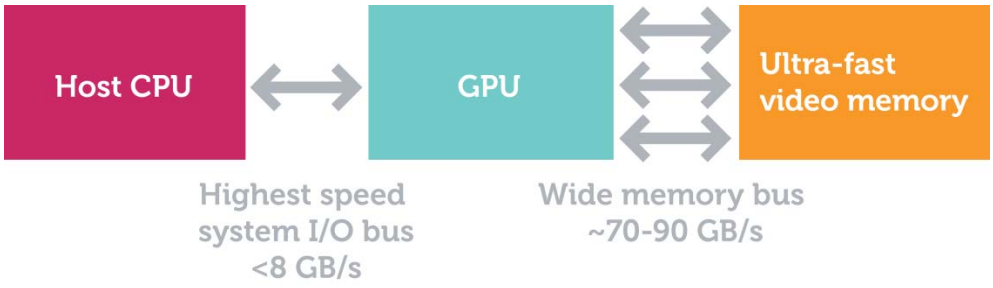

Figure 1. A typical graphics processing unit (GPU)

GPUs are still in their infancy as general purpose processors but are rapidly gaining momentum. GPU processing resources and their associated memory systems are heavily optimized to execute large numbers of operations in parallel. They maintain high efficiency through the use of multi-core designs that employ both hardware multithreading and thread execution interleaving to deal with longer memory latencies. Although GPUs currently use SIMD or MSIMD computational models, near-future offerings might also exploit multi-block, many-thread MIMD models that do not have the inefficiencies of SIMD architectures associated with processing plain sequential operation streams.

A hybrid model of computation naturally ensues (see Figure 2). CPUs and GPUs each have their own computational "sweet spots." The challenge at the center of heterogeneous HPC computing is the judicious decomposition of algorithms and computations into chunks to be computed by CPUs and chunks to be computed by GPUs, in ways that exploit the advantages and avoid the problems of each technology. Appropriately and artfully decomposing the computational work can result in a hybrid

computational model providing orders of magnitude performance improvements over traditional approaches using only one processor technology. GPUs are a good match for algorithms with high computational complexity that need great numbers of floating point operations per second (FLOPS). They are also well suited to some algorithms that spend most of their time in small, inner loops of code or that have high computation versus communication needs.

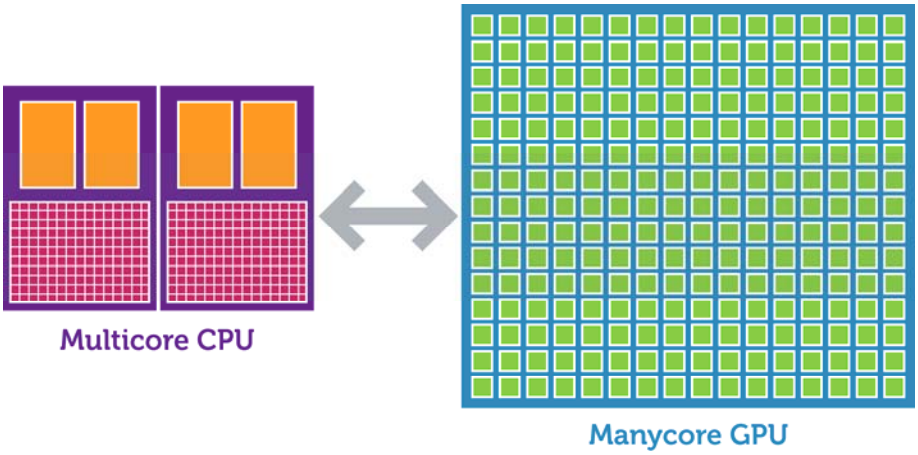

Figure 2. Hybrid CPU-GPU computing model

GPUs differ from CPUs in that they assume a highly parallel workload. While a GPU maximizes throughput of all threads, limited only by resource availability (e.g., bandwidth, memory, etc.) with control logic distributed and shared over many threads, a CPU minimizes the latency experienced by each thread using large on-chip caches and very sophisticated control logic. A GPU inherits its computational characteristics from the graphics pipeline: namely, instruction throughput is very important, multithreading can hide computational latency and threads are ephemeral commodities that can be created, executed and deleted by the thousands of millions per second.

Although GPUs promise massive performance increases in compute speed (hundreds of percentage points) whenever an algorithm is amenable to GPUs and the code can be ported to take advantage of the processing power of GPUs, no "magic" allows an existing application to be rebuilt for GPUs. This process requires carefully reexamining both applications and algorithms, which in many cases must be modified, and of course, ported to run on a GPU.

2.1 GPU-friendly algorithms

Not all algorithms work well on GPUs; in fact, some work much better on CPUs. However, when an application is ported to GPUs and its algorithmic structure is well matched to the strengths of GPUs, the performance gains can be stunning.

So what, besides graphics processing, are GPUs good for? GPU-based hybrid computing has done very well in areas such as weather simulation, bioinformatics, molecular dynamics, oil and gas, astrophysics, Monte Carlo analysis, and many others.

Several considerations are important when analyzing how best to execute a computation on a hybrid CPU/GPU platform.

A typical computation offloaded to one or more GPUs may have the following structure for each GPU:

1. CPU loads GPU with code to execute and data chunks to use as input.
2. GPU computes result from input data using host-provided code, and makes result available to CPU.
3. CPU retrieves result from GPU, combines it with other results.
4. CPU repeats steps 1 to 3 until CPU/GPU obtains a final result.

Aside from how parallelizable an algorithm is, the efficiency of executing a computation that uses it in a GPU depends on other characteristics of the CPU/GPU combination related to 1) data movement in the CPU/GPU channel; 2) balance between computation and communication/synchronization time requirements and 3) the GPU to CPU ratio, that is, the ratio of GPUs to CPUs involved in the computation.

Concerning data movement, a very important parameter is CPU/GPU channel bandwidth, which determines the speed at which data and code can be loaded from host to GPU and at which results can be retrieved from the GPU. Its importance depends on the ratio between a computation's expected data transfer time and expected compute time. If compute time at the GPU is very large relative to data load/retrieval time, channel bandwidth constraints can be relaxed. Otherwise, increases in channel bandwidth can result in improved performance.

Another aspect to consider is algorithm computation time vs. communication time. Algorithms that need to communicate intermediate results to neighbors very frequently compared to time spent generating the result are generally poor targets for GPU acceleration, especially if communication needs to leave the server. Increasing the GPU/CPU ratio may alleviate this problem by creating localized pools of GPU resources that in turn localize communication to the server. This can dramatically increase the hybrid compute performance of some algorithms. Interestingly, with some HPC algorithms, an increase of the GPU/CPU ratio may eventually degrade overall performance so careful understanding and testing of potential algorithms is critical to successful GPU deployments.

The industry is currently finding new and innovative ways to exploit the functionality and performance of GPUs for general purpose computing. The downside of this constant change is the need for flexibility in the hardware and software platforms being deployed for development and production to allow for the fast uptake of these new advances.

3.0 Hybrid accelerated HPC solutions

An optimized HPC compute infrastructure for a specific problem is highly application dependent. GPU-accelerated applications can have a wide set of infrastructure needs, depending on the nature of the problem being tackled and the actual algorithms being used. The optimal architecture for a problem and application may change dramatically as new ideas and code are developed. The pace of change is currently very rapid, creating a need for great infrastructure flexibility. This section explores the various aspects of an accelerated HPC solution as well as current approaches to hybrid compute integration.

3.1 GPU compute accelerators

Accelerators are key components of a hybrid HPC solution. GPUs currently are at the forefront of the race of compute accelerators, not only because they are commodities—making their price/performance ratio very advantageous—but also because they are much simpler to program than any other accelerator solution (e.g., FPGAs, ASICs, etc.).

Standard GPU offerings have very large numbers of cores (from 240 to 800 per chip), very high memory bandwidth (over 100 GB/s), and very high single precision (SP) throughput (over 1,000 gigaFLOPS, or GFLOPS). Double precision (DP) GPU throughput has been much lower in the past because most graphics computations can be done in single precision. Because both SP and DP performance are very important to most HPC applications, GPU manufacturers have concentrated on increasing DP throughput so that GPUs can address standard general computation needs. The latest generation, released to the market in mid-2010, has greatly increased the delivered DP performance. The compute power of GPUs is increasing very rapidly. Next-generation GPUs with 1,600 cores, and delivering over 2.7 teraFLOPS (TFLOPS) SP and 512 GFLOPS DP compute power are just around the corner. GPUs are typically available as full-length, double-width cards at prices comparable to those of server CPUs ($500-$2,300).

Two representative commodity GPU components are the AMD FireStream™ 9270 and the NVIDIA® Tesla™ C2050/C2070, originally codenamed Fermi. The AMD FireStream 9270 GPU has 800 cores and 2GB of GDDR5 memory with a memory bandwidth of 108 GB/s. It delivers up to 1.2 TFLOPS in SP and up to 240 GFLOPS in DP computations. The NVIDIA Tesla C2050/C2070 GPU has 14 multiprocessors of 32 cores each, for a total of 448 cores. It can be configured with up to 6 GB of memory depending on the model, with a memory bandwidth of 144 GB/s with true ECC memory throughout the whole processor. It delivers peak performance of 1.03 TFLOPS in single-precision and up to 515 GFLOPS in double-precision computations per GPU. The AMD FireStream 9270 has power ratings of 160W typical and 200W peak. The NVIDIA Tesla C2050/C2070 has a peak power rating of 238W.

3.2 Hybrid GPU/CPU integration

The prevalent industry model to provide GPU/CPU integration is to package a multi-socket server with one or more GPU cards in the same rack enclosure, communicating via an internal high-speed PCIe bus. NVIDIA is producing GPU cards for this kind of on-server solution. These cards are not the

standard GPU cards found in desktops. They come in a fanless, passive cooling packaging designed to fit inside standard server CPU enclosures and to be cooled by the server. Recently, a few high-end offerings under this model have started appearing that increase the number of internal GPU cards per server enclosure, in fixed 4U rack mount systems.

The two common ways to deploy multiple GPUs into a single system are 1U and 4U rack-mounted server form factors.

The 1U server option generally has a dual-socket Intel or AMD motherboard with two x16 PCIe ports for connection to the GPUs and an additional port for high-speed I/O. These systems usually provide two GPUs integrated into the chassis and deliver a base-level building block for previous GPU cluster designs.

The 4U server option generally has been used by end users looking to build a system with more than two GPUs. The most common configuration is a dual-socket Intel or AMD server with up to four x16 PCIe ports connected to four or more GPUs. Additional options provide for four CPU sockets on the motherboard or additional GPU connectivity up to a maximum of eight in a single system.

Both of these two options have required the end user to decide between system density and GPU density. To increase GPU density, the end user must either choose the 4U system upfront to enable the increase or upgrade the entire platform from the 1U server to the 4U server. This type of design issue complicates the move from development to production by requiring the moving of platforms or by greatly increasing the space necessary to deploy large clusters.

Although both these offerings support larger numbers of GPUs, they are still fixed in function, one of their major limitations.

3.3 Limitations and benefits of current industry solutions

Although their tight packaging has real estate and operational advantages, these standard industry GPU acceleration offerings are limited in their flexibility and growth. GPU appliances provide only a small number of GPUs constraining GPU/CPU ratios, and their configurability is limited by design. In situations where HPC algorithm performance dramatically increases with every additional GPU added per CPU, an HPC shop will not be able to reap the potential benefits of adding infrastructure should a traditional approach be used. Equally important, it is very difficult to reassign GPU resources to CPU resources with the available fixed geometries. Changing GPU/CPU mixes in an HPC installation that uses these technologies as workloads change in time becomes very difficult if not impossible.

4.0 Dell HPC accelerated solution approach

There are many ways of integrating GPUs and CPUs. The latest accelerated HPC offerings from Dell™ address the issues of lack of flexibility and ease of growth present in current industry GPU acceleration approaches.

4.1 External GPU extension chassis: The Dell PowerEdge C410x

The Dell PowerEdge™ C410x is an external expansion chassis, with sixteen x16 PCIe slots for GPGPUs in a 3U enclosure. It provides for very flexible configuration, supporting from one to eight servers accessing PCIe devices. Host servers are attached via eight external PCIe x16 connectors. This flexibility is critical when algorithm development or application workloads benefit from dynamically changing the GPU compute ratios; doing so is not possible in designs with internal GPUs but is very easy with a Dell PowerEdge C410x expansion chassis.

Dell intended from the outset for the PowerEdge C410x to be commercial grade, "production ready" for deploying complex HPC applications in research and corporate environments directly from the development environment where the GPU ratio flexibility is critical to algorithm development. To this end, Dell designed the PowerEdge C410x to increase reliability and compute density by allowing more GPUs to be configured per rack U, and to allow for varying GPU/host ratios (1:1, 2:1, 3:1, 4:1 … up to a possible 16:1). Current technology limitations restrict the GPU/host ratio to a maximum of 4:1 per HIC. Achieving a 16:1 ratio requires server functionality that will soon be available from Dell.

The Dell PowerEdge C410x provides redundant power supplies, which are essential to resilient compute acceleration. All PCIe modules, as well as their associated power supplies and fans, are individually serviceable. When configured with NVIDIA Tesla M2050 GPUs, the Dell PowerEdge C410x can deliver up to 16.5 SP TFLOPS. Although the Dell PowerEdge C410x allows for a mix of different PCIe cards, including GPUs as well as other PCIe devices such as solid state disk storage devices or multi-port InfiniBand® (IB) adaptors, it does not let you mix different NVIDIA GPU cards on the same host. Also, while the system technically supports any PCIe device, no general purpose sleds are available yet that accommodate all possible kinds of such devices.

The Dell PowerEdge C410x is specifically qualified for use with the Dell PowerEdge C6100 server. The system will be enabled in the future for connection to virtually any server on the market with the appropriate host interface card. The extension chassis features N+1 Power (3+1) and N+1 Cooling (7+1).

4.2 Benefits of the Dell PowerEdge C410x GPU extension chassis

Dell is taking a novel approach that is simple yet very forward-looking. The idea is to design for GPUs as a shareable compute resource and to separate them from server CPUs using an external enclosure that houses only the GPU resources.

This use of external GPUs with the Dell PowerEdge C410x gives maximum flexibility to an HPC installation. Compute nodes can be designed for peak CPU performance and optimal power and

cooling requirements. GPUs can be housed in chassis specifically designed for their high power and cooling requirements. GPUs can be upgraded independently from compute nodes and vice-versa. You can also mix and match GPUs from different manufacturers. Available GPUs can be allocated to various compute nodes as needed: if a specific node requires more GPUs, they can be allocated without changing the chassis, which would be impossible by design if the node only had internal GPUs. This external arrangement has even further advantages: redundant power in the chassis increases reliability, and hot-swap GPU cards make it easy to repair them as needed, without disrupting execution of a project or processing of research results.

4.3 Host computing: The Dell PowerEdge server line

The Dell PowerEdge C Server line is designed for and meets the performance, reliability and price sensitivity demands of the HPC market. The Dell PowerEdge C6100 server is the preferred server to provide CPU compute power in hybrid solutions that use the Dell PowerEdge C410x extension chassis, although any server can be connected to the PowerEdge C410x if it has the appropriate hardware interconnection card. The Dell PowerEdge C6100 is a 2U rack mountable unit with four two-socket Westmere EP nodes. Each system board has 12 DIMMS and supports two GigE (Intel) interconnects. The chassis has one PCIe x8 on-board daughter card slot that can be used for 10GigE or QDR IB interconnects, and an optional SAS controller for use in place of IB. It also has one PCIe x16 slot for half-length, half-height cards. The Dell PowerEdge C6100 supports hot plug, individually serviceable system boards and nodes, and can be configured with up to 12 x 3.5-inch drives (three per node) or up to 24 x 2.5-inch drives (six per node). The Dell PowerEdge C6100 is both NVIDIA HIC certified and DDR/QDR IB PCIe card certified.

4.4 Benefits of the Dell PowerEdge C6100 server

The Dell PowerEdge C6100 is a perfect server candidate for GPU-accelerated HPC. It is a hyper-scale-inspired building block for high-performance computing (HPC) and research computing (RC), NVIDIA certified and well matched to the Dell PowerEdge C410x GPU extension chassis.

4.5 Host computing: The Dell PowerEdge M610x blade server

Some situations require achieving very high compute performance and density, while minimizing overall infrastructure and operational costs with solutions that are designed for ease of adoption and reduction in overall power consumption and real estate. The Dell PowerEdge M610x blade server provides such a compute alternative, providing GPU computing with blade form factor and benefits as well as full enterprise-class reliability and manageability.

The Dell PowerEdge M610x is a full-height, 2S Intel blade (up to eight per chassis), that can house up to two GPUs and a QDR IB card. It can be configured with up to 12 DIMMs, two hot-swap HDs, hardware RAID, and a 4 x 1 GigE LOM. The blade has slots for two I/O mezzanine cards and has two x16 Gen2 expansion slots for full-height, full-length PCIe cards. It has 2 x 250W and 1 x 300W platinum level (+94%) supplemental power. One of its main advantages is that it can be cooled within existing Dell blade chassis–a strong feature of Dell blade chassis, which can handle the power and thermal

requirements of GPUs. The combination of Dell PowerEdge M610x blades and the Dell PowerEdge C410x provides yet another alternative for added flexibility in real estate, power consumption and accelerated HPC compute power planning.

4.6 Benefits of the Dell PowerEdge M610x blade server

Of course, GPUs can be housed in the same enclosure as the CPU, which is convenient when a blade form factor is desired. In this case, one sacrifices configuration flexibility for simplicity and convenience of deployment and for enterprise-class hardware, a tradeoff that may make sense for small HPC configurations and for enterprise organizations with an HPC component.

5.0 Host-to-GPU interconnection patterns

Several characteristics of GPU acceleration can influence the overall throughput and performance of a GPU-accelerated configuration. Among these are (1) the GPU-to-host ratio, i.e., the number of GPUs connected to a single host; (2) the host-to-GPU interconnect bandwidth; (3) the host-to-GPU data path width; and (4) the nature of the host-to-GPU interconnect, whether internal (on-board GPUs) or external (GPU banks or farms).

Initial designs for GPU compute systems sought to maximize the bandwidth connection from the host to the GPU adaptor for fear of limiting the performance by starving the GPU of data. Experimental data on state-of-the-art applications shows that host-GPU data path width and the use of dedicated x16 links may not be as important to overall performance as once thought. They also show that external connections minimally affect performance but provide substantive added flexibility over internal connections. Host-to-GPU interconnect bandwidth matters in so far as host-GPU and GPU-host data transfers account for a substantial fraction of the overall algorithm execution time.

Of all these factors, the GPU-to-host ratio is critically important for overall performance efficiency. Its impact is highly application and algorithm dependent. There are a number of applications that scale well beyond one or two GPUs, and show dramatic performance improvements as the GPU-to-host ratio increases. It is important to know the characteristics of the applications intended for GPU acceleration and to provide a path for upgradability in the architecture to allow for new applications and algorithms to be deployed on existing infrastructure as they are rolled out to production.

Flexibility is very important to address such unknown or unpredictable situations. The Dell PowerEdge C410x introduces the density and flexibility needed to cope with such requirements, and separates the host from the GPUs. The combination of Dell PowerEdge servers and Dell PowerEdge C410x extension chassis allows for very flexible interconnection patterns that can satisfy widely dissimilar HPC application and algorithm architectural and compute requirements.

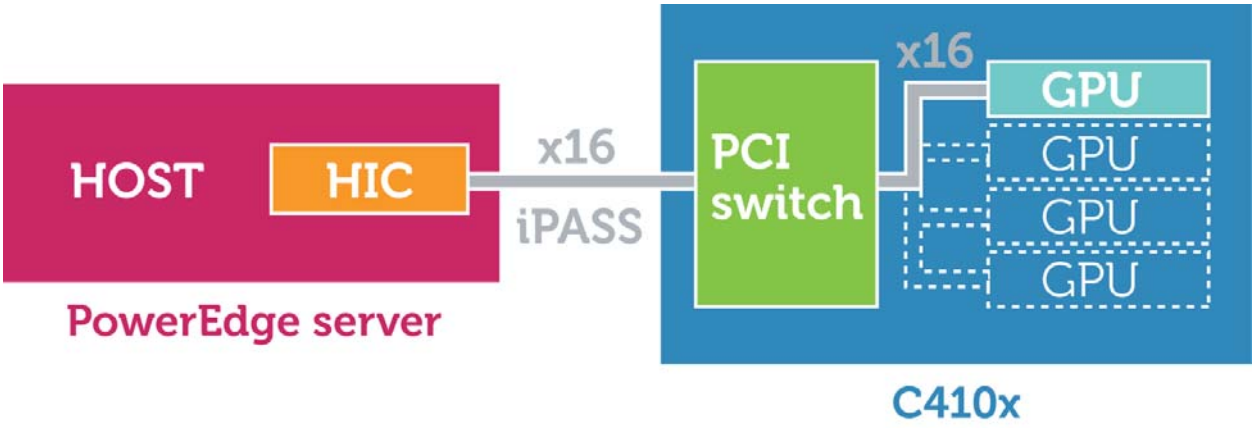

Figure 3. Dell PowerEdge C410x flexibility: Single host with single x16 interconnect

Figure 3 shows a single Dell PowerEdge host server connected via an x16 iPASS cable to a Dell PowerEdge C410x. This specific connection pattern allows for 1:1, 2:1, 3:1, and 4:1 host-to-GPU ratios.

Adding GPUs requires only plugging a GPU module into the Dell PowerEdge C410x chassis and providing the necessary power and data connections.

The overall throughput can be increased by using a dual-ported x16 interconnect. Figure 4 shows an interconnect pattern that can expand up to eight GPUs, using two host interface cards and two x16 iPASS connections between the host and the GPU chassis.

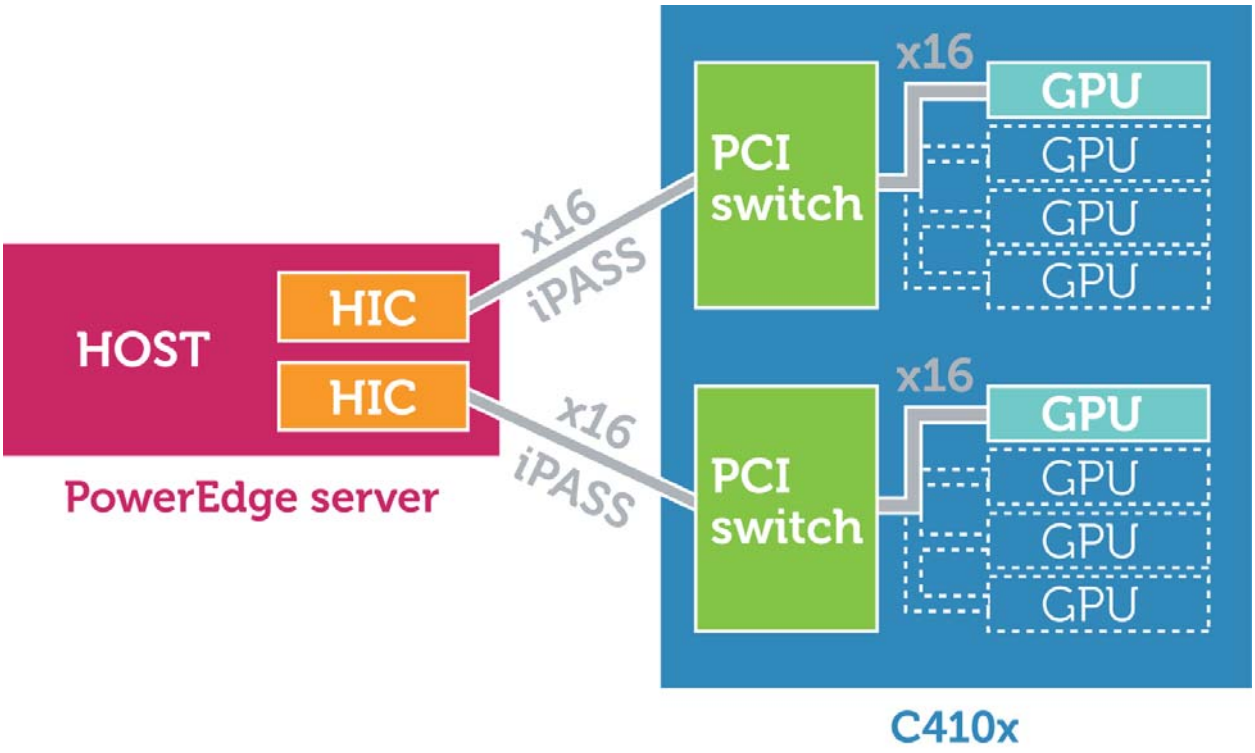

Figure 4. Dell PowerEdge C410x flexibility: Single host with dual x16 interconnect

6.0 Dell: Accelerating time to the dream

Dell offers its HPC customers maximum flexibility in developing configurations and solutions that can easily change over time, and in particular, where compute accelerators such as GPU hardware can easily be replaced as new technologies become available, or as other accelerator technologies become more convenient to use. This allows CPU hardware and GPU hardware to be deployed independently of each other. Additionally, Dell solutions as described in this document can easily grow with changing needs. Dell customers can purchase a system today and be able to add GPUs at a later date without compromising the system's performance or cost.

The net result of such flexibility and resilience becomes visible during the development and production stages of HPC applications. During the development stage, it allows the same hardware infrastructure to provide a wide set of configurations and capabilities for ease of new algorithm development and testing. During the production stage, the Dell hardware infrastructure's flexibility allows for optimization of the system configuration for different application needs based on overall or high priority workload demand.

Dell brings to market systems optimized to distinct environments, leveraging its open model to provide flexible platforms for internal and external GPU adoption. Dell is driving the next wave of HPC technology integration – delivering powerful GPU Compute solutions that are quickly becoming the most powerful computational hardware for the dollar.

Dell is committed to increasing the efficiency of HPC systems through innovation and investment in forward-looking solutions that integrate new technologies, such as GPGPUs, in state-of-the-art platforms for new computing models that produce results faster in a cost-effective manner. Dell delivers efficiency, not only through standardization, simplification, and automation, but directly through innovation, accelerating time to the dream.

7.0 References

Presentation by Mark Fernandez, Dell Advanced Systems Group. NVIDIA GPU Conference. San Jose, California. September 2010.

Presentation by Jeffrey Layton, "Accelerators," Dell HPC Group internal document.