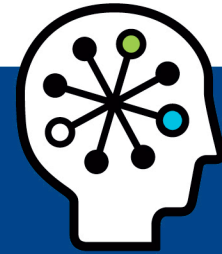




# Catalysts for HPC Innovation Program

Bringing innovative solutions to HP customers



## Contents

Introduction .....	2
Grid .....	4
Visualization .....	6
Accelerators .....	9
Converged Fabrics .....	11
Multi-core Optimization .....	13
Power & Cooling .....	15
Dense Computing .....	16
For more information .....	19

## Introduction

### HP's focus on High Performance Computing

The high performance computing market is traditionally an area where you will find customers with the most demanding computing requirements. Their computing solutions require the largest systems and the latest technology to achieve their outcomes. By focusing on high performance computing HP offers an environment where we create innovative solutions for our customers and in doing so bring those innovative technologies back into HP. This enables us to build on our industry standard systems and allows us to deliver solutions with greater economies and greater simplification and do it in such a way that builds on our years of experience in building large-scale systems for the high performance computing market.

### HP's leadership in the HPC Market

HP has the benefit of being one of the most innovative technology companies in the world. By leveraging the technology investments that HP is making in virtually every area, we are able to build on those new, innovative technologies and deliver HPC solutions that meet or exceed our customers' requirements.

For instance, we build on investments offered by HP Labs in the areas of nanotechnology and smart cooling. We build on capabilities offered by Enterprise Computing's focus on adaptive infrastructure by providing robust mission critical enterprise solutions for our enterprise customers with scale-up and scale-out solutions that they depend on for meeting the needs of their engineering environment for analytics or scientific research.

And finally, we build on the great wealth of industry standard products that HP provides to the industry. On top of this, we have focused HPC investments in areas such as advanced development programs for new technologies. Our Catalysts for HPC Innovation Program enables us to focus on collaborative projects with our customers to develop specific technologies that enhance the state of the art in high performance computing.

### Research at HP

HP has a long tradition of investing in innovation and successfully transferring technology breakthroughs into commercial success. HP Labs, our worldwide core research institution has a history of fundamental innovations to its credit. HP University relationships have scores of active collaborations with institutions in the United States, Europe and Asia. And the HP product divisions routinely adapt this emerging technology to commercial products.

We know that collaboration and innovation in High Performance Computing is best driven through research projects with customers, partners and research institutions. In our HPC disciplines, we value working closely with our customers to understand their business challenges and to achieve the results they demand. Through our customer-driven collaboration programs and our research we help our customers optimize and improve their computing environment, on a local and global scale.

The HPC organizations are part of the Scalable Computing and Infrastructure Organization (SCI), which is responsible for all aspects of the value chain for scale-out computing—including business management, sales, supply chain, marketing, and R&D.

### Catalyst for HPC Innovation Program

The SCI organization launched the Catalysts for Innovation Program to drive adoption for the next wave of emerging high performance technologies through a series of advanced development projects and customer collaborations. The basic idea is to incubate new technologies in real customer situations in order to solve some demanding problems. Each project in the program lasts about a year and will result in technology demonstrations and potential new product introductions. The goal for each project is to include at least one customer collaboration in each of our geographical regions.

HP formed the Catalysts for Innovation Program because customers demand innovation and collaboration to solve new problems. Investments in this space return dividends both in the broad, fast growing HPC market and eventually in the commercial space through trickle down effects.

The need for 'processing power' to solve critical problems is key to maintaining a competitive advantage; and this requirement is at the forefront for most HPC customers. Whether it is an oil company racing to cut decision time by an order of magnitude for billion dollar drilling investments; an aerospace company evaluating the multi-physics effects of wing shape on fuel consumption and safety; or a financial services company dashing to evaluate literally thousands of risk position alternatives in fractions of seconds; our customers value performance above all else.

Consequently, the mission and behavior of HPC customers frequently places our market at the forefront of key market trends - demanding and rewarding innovation. Many disruptive technologies incubate in HPC and evolve into broader commercial market developments. Several of HP's key technologies were incubated in HPC, evolving into full-fledged initiatives and products, thus demonstrating the potential of providing innovative technologies to the community at the earliest stages of the development process. For instance, scale-out computing and Linux were evident more than five years ago in HPC. HPC pioneered the use of AMD processors as a viable alternative to Intel products. Our market is leading the adoption of bladed infrastructures. HPC's early focus on Grid computing has led to a leadership position on the emerging Cloud computing front. HPC academic researchers are heavily engaged on work to deal with looming many-core architectures, which will require unprecedented parallelism to effectively utilize future processor designs from Intel and AMD.

The Catalyst Program includes projects in the areas of computation, data management, and visualization. In the computation area we are investigating multi-core optimization, advanced parallel languages, tools and computing environments, attached hardware accelerators, dense computing technologies, and grid. The HP Multi-core Optimization Program is a broad initiative to optimize HPC applications on multi-core systems and enable customers to realize the full performance benefits of multi-core technology. Our Accelerator project is an on-going effort to research various accelerator technologies such as General Purpose Graphical Processing Units (GPGPUs), Field Programmable Gate Arrays (FPGAs), and custom ASICs. We build and optimize typical HPC applications and benchmark them to determine which accelerators are most beneficial to HPC customers. The Dense Computing project is researching new server designs to achieve higher price/performance with lower power in less space than the commodity server roadmap. The Grid Catalyst Project is working with customers and ISVs to develop best practices for grid projects and is grid-enabling HP management software.

Data management projects include converged fabrics and remote caching. In the visualization area, we are exploring parallel compositing that simplifies the development and use of parallel applications on graphics clusters. And, of course, we are also working to extend our leadership technology in areas that address environmental concerns, such as power and cooling.

# Grid

## Introduction

Grid is a software environment that makes it possible to share disparate, loosely coupled IT resources across organizations and geographies. IT resources are freed from their physical boundaries and offered as services. They can potentially include almost any IT component – computer cycles, storage spaces, databases, applications, files, sensors or scientific instruments.

In grid computing, resources can be dynamically provisioned to users or applications that need them. Resources can be shared with a workgroup, department or enterprise; among different organizations and geographies; and even with groups outside the enterprise in collaborative projects. Grids can be designed to support various business processes. Grid technologies use Web services standards such as XML, SOAP and WSDL.

Grid technologies have long been used for scientific and technical work, where dispersed computers are linked to create virtual supercomputers that rapidly process vast amounts of information. Now, with the success of e-commerce and the Internet, the commercial enterprise is moving to an IT model based on Web services, in which software can be offered and consumed as services – a service-oriented architecture.

HP is devoting considerable resources to bringing the benefits of grid computing to the enterprise. Grid is not a ready-made solution, but rather a set of components and protocols pulled together to create a solution. HP views grid computing as a powerful way to virtualize resources and create a service-oriented architecture, where IT provides resources to business on demand, like a utility.

## Grid technologies and customer usage

Grid technologies are valuable because they help customers break down silos and virtualize the IT environment. This results in better utilization of resources across organizations. Customers typically deploy grid technologies because they are looking for measurable benefits such as cost savings and improved time-to-market for new products and services. They often find that there are additional benefits that are more difficult to measure, but are very real and valuable. For example, customers have found that the use of Grids fosters collaboration across organizations that share IT resources.

The early adopters of Grid technologies are in industries that have applications that run well in a distributed environment. For example, (1) Financial analysis and modeling in banking, securities, insurance (2) pharmaceutical and biotechnology research, and (3) engineering analysis and automation. The applications that are amenable to be run on the Grid are loosely asynchronous and do not require much data movement.

## Challenges with Grid technology

First, grids are heterogeneous environments both from a hardware and software perspective. Secondly, they are large (~5-10,000+cores and getting larger) and third, they can be geographically distributed. As you might imagine, there are many challenges with grids. One issue that keeps surfacing with customers is the lack of adequate management tools for grids. You need to be able to provision, monitor, discover new resources and control the grid from one place. Other issues that good management tools solve, are to make the environment fault-tolerant thereby recover from server outages in a graceful manner. This leads to the notion of self-healing or predictive healing and this is getting more and more important as grids get bigger. We also see a synergy between grid and adaptive infrastructure and our customers now request a “24 x 7 light-out automated computing” environment based on standard building blocks and delivered through services.

Access to data is also a major problem. The applications running on the grid need to access data. In some cases, the data that is dispersed across the grid is not public. The data resides in geographically distributed areas, within separate ownership domains, and separate management domains. So not only can data access be a problem, there are also security issues such as authentication and identity management and managing access to the data.

Another problem is to have the environment manage a workload of both batch and service jobs across a geographically distributed grid and make sure that the jobs meet their SLAs. Service jobs are

always running and grow or shrink in response to external triggers such as users or scheduled activities. The charge back and reporting can be a challenge.

## What is HP doing for Grid?

HP ensures that HP systems, servers, and clusters are “grid-enabled” meaning that these systems can be used in grids and that grid middleware runs well on them. We work with the grid middleware ISVs to ensure that their software is tested on HP systems.

HP Management Software for provisioning, monitoring, and control can be used to simplify Grid deployment, executive and management. For example, many of the same HP components that are valuable for managing clusters are also valuable for managing grids.

We often say that “Grids are built not bought.” HP Consulting & Integration Services offers services to help our customers design, deploy, and manage grids.

HP Flexible Computing Services (FCS) is HP’s utility offerings. Customers can use systems in HP data centers on a “pay-per-use” basis. HP FCS interoperates with customer grids that run many of the popular grid middleware packages such as Platform Computing LSF and Symphony, DataSynapse GridServer, and Altair Engineering PBS Professional.

HP also provides guidance on best practices. The Catalysts Grid team works with customers who are deploying grids, and documents best practices for deploying and managing grids.

## HP Grid guidance and best practices

The Grid Catalyst Program is developing a knowledge base with regards to leading grid technologies in various verticals. The idea is to collaborate with our customers and partners to create joint solutions that would solve problems encountered in the Grid. For example, the financial market is a growth area for grid. We have focused on the financial services industry first and then on Life Science and Pharma. We have spent time looking at grid middleware solutions have authored white papers which will help our customers install and run middleware applications on HP grids.

We have also investigated Cluster Resources Grid Workload Manager Moab and authored a Quickstart that will help our customers use Moab on HP grid.

## Grid and Financial Services Industry customers

In the financial industry we see compute grids but also compute grids combined with data grids. The three most popular grid middleware for compute grids are Platform Symphony, DataSynapse GridServer and DigiPede. All three of these products are workload and resource managers for service jobs. The first two schedulers are available for both Linux and Windows whereas the latter is only available on Windows.

This new type of service scheduler allows our customers to integrate their application with the scheduler via a specific API and increases parallelism in their applications. They also allow jobs to be dispatched onto any system within the grid. In the end, it is all about breaking down the silos to increase the utilization of their existing systems.

Another type of application needs to access data. Today some of our FSI customers are looking into distributed data caching to solve their problem. By Distributed Data Caching we mean that we deal with the issue of how to distribute large number of small files or portion of larger files across the nodes. In this type of DataGrid the data is directly coupled to the application and the amount of data being dealt with is overall likely to be small. To take advantage of data caching the application usually will do few writes and many reads. Today we see three main products in this area: Gemstone, Oracle Coherence (formerly known as Tangasol) and GigaSpaces.

While grids are getting bigger and bigger we see one major problem appearing—that is how do you manage these large grids? This was the reason behind the team looking into to making HP management tools more suitable for large grid as well as working on creating use cases for the Grid Dynamic Workload Utility in FSI.

## Learning from the Grid Catalyst Program

The Grid Catalyst team has been working with major customers to help them apply the learning’s from this project. They are also working on generalizing the learning’s from these custom projects and defining solutions and configurations based on the teams’ experiences.

# Visualization

## Introduction

High-performance computing often involves modeling and simulation, such as the flow of air around a vehicle, the effects of a head-on automobile collision, and weather modeling. These simulations produce vast amounts of data. Instruments and sensors also produce large amounts of data, such as seismic data for oil and gas exploration and data from medical instruments such as CT scanners. Consequently, large data sets need to be analyzed and visualized to gain insights that will lead to innovation, better design, advanced research and science outcomes and improved time-to-market.

## Visualization Challenges

HPC visualization uses desktops varying in size and price, largely depending on the graphics card and amount of system memory. Usually, these desktops visualize data sets copied from HPC systems to the office systems. Tremendous advances in graphics technology have occurred in recent years, yet visualization requirements for very large HPC data sets are well beyond those of a typical desktop. For example, seismic data sets can range from 10's to 100's of Gigabytes. Putting an expensive desktop system in an office and copying such large data files from an HPC system to the desktop isn't necessarily the best approach.

The size of particular data sets and the collaboration requirements usually drive many of the requirements unique to HPC visualization. Cost is also an important factor, as customers want an approach that is based on industry standards and COTS technology, which results in economies of scale and performance.

Large data sets produce images that require of substantial screen real estate. Visualization users work with extremely large models with fine details that must be seen in context. Displays with 4 to 8 times the resolution of a typical desktop display or more may be needed, such as quad-HD projectors and LCD with resolutions of 3840x2160 (over 8 MPixel compared to typical desktop display of less than 2 MPixel). These displays are driven by up to 4 video inputs and display large data sets that have been rendered in great detail.

Collaboration is an equally important reason for large multi-screen displays. Because multi-projector or multi-LCD displays scale-up to even higher resolutions, they are used when people need to collaborate on a visualization project. Examples include researchers conducting joint analysis, engineers deciding where to drill for oil, or designers working together to create new aircraft models.

Other challenges also exist in visualizing increasingly large data-sets. Rendering large data sets can exceed what is feasible to do at interactive frame rates on a single graphics card, even when the image will only fill a single display. To achieve this goal, one must distribute the problem over multiple systems, typically a cluster of systems connected by a high-speed interconnect, such as InfiniBand. This approach is similar to distributing a computation across multiple nodes, only in the visualization scenario, the systems are equipped with graphics cards and the results are now images.

Another issue with large data sets is that they are inconvenient to copy from the HPC systems (where they were computed) to where they need to be visualized over a campus or miles apart. Sending these large data sets from their source slows the workflow and can stress backup systems. Users want to be able to visualize large data sets from their offices, without the time and cost of copying those data sets to their offices.

Companies would also prefer to avoid putting expensive high-end graphics systems in every office. Engineers and scientists may only need the high-end capabilities some of the time, but they want to be able to access them when needed.

A solution to both of these challenges is to pool together and share high-end graphics systems integrated with HPC systems that have high-speed access to large data sets. The graphics systems can then be accessed by remote visualization. Remote visualization allows users to run a visualization application on a high-end system in the lab but display the results locally on an office desktop. Users can interact with the application as if the keyboard, mouse, and display were plugged into the system in the lab.

To summarize the HPC challenges:

- HPC applications produce very large data sets that need to be visualized.
- Visualizing these data sets can require multi-screen displays and multiple graphics systems working together to drive these displays or distribute the work of producing the images.
- The HPC computing and storage systems are often located at some distance from the users' offices, but users want to visualize the results from their offices. Copying very large files and dedicating high-end desktops to every individual is a costly approach.

## HP Visualization Solutions

To address these HPC challenges, HP provides a range of hardware and software solutions.

HP offers rack-mount servers with graphics cards that provide graphics systems in as little as 1U of rack-space. These systems include the management and monitoring features of a server and the reliability and price-performance that customers have come to expect from HP servers. For example, the HP DL160G5 Xeon-based server is particularly well suited for interfacing to graphics systems, since it supports PCI-Express 16x Gen2 cards.

HP also offers a 2U option, the Opteron-based DL385G5 server. Top-of-the-line graphics cards from NVIDIA, such as the FX 5600, are available. These graphics servers keep the data storage, high-end visualization, and computing systems consolidated as shared resources, while allowing engineers to visualize the data remotely from inexpensive desktops and even laptops.

HP also offers clusters of graphics systems that includes many systems managed together that share a high-speed interconnect. These systems can also share access to a large-scale parallel file system over this interconnect. The compute-systems can also be a part of the same interconnect, allowing the compute, visualization, and storage resources to all be integrated. These clusters can be used to run applications that drive the multi-screen displays (or CAVE environments) and scale-up to handle very large data sets by distributing image rendering across multiple systems. Or the cluster can be used to efficiently manage a shared pool of high-end graphics systems that are used remotely.

## HP Scalable Visualization Array (SVA)

To tie all the pieces together, HP provides the Scalable Visualization Array (SVA) for managing and using graphics clusters. SVA is part of the Linux-based HP cluster system software and Unified Cluster Portfolio (UCP). SVA supports a range of hardware platforms and graphics cards and is available factory integrated and tested with HP worldwide support and service.

SVA enables scaling of interactive visualization to large data sets, using the Parallel Compositing Library (described later). SVA provides mechanisms that describe multi-screen displays, allocate cluster nodes to users, and launch visualization applications on the cluster.

Related to this, SVA provides features for running remote graphics software such as HP RGS sessions on a cluster. Consequently, the graphics cluster in the HPC environment becomes, in essence, a managed, shared pool of high-end graphics systems.

HP provides remote visualization via Remote Graphics Software (RGS). Customers use RGS to deliver excellent image quality to the desktop using standard Ethernet at interactive frame rates. SVA also provides similar support for running the open source remote visualization tools VirtualGL and TurboVNC. Besides giving remote access to high-end shared graphics systems, these tools enable remote collaboration. Other users can join a session, view, and interact with the application.

Applications use distributed rendering to scale-up to handle very large data sets. Basically, they divide the data across multiple cluster nodes, and each node produces an image for part of the data. Next, these partial images are combined to produce the complete image. The images are typically combined, pixel-by-pixel, using either depth or transparency information associated with each pixel. A number of such applications are available commercially and as open source.

## The HP Parallel Compositing Library

To help develop parallel cluster-aware applications, HP developed the Parallel Compositing Library. Using information about how the partial images overlap, the library routes pixels among nodes, combines the pixels, and delivers the resulting image to a display, which may be a multi-screen

display. This library grew out of advanced development work that had its roots in an HP project to develop compositing hardware (SEPIA) and the work that HP did with others to define an API for parallel compositing. Within HP-CCN (HP Collaboration and Competency Network), HP continues to work with customers, research partners, and ISV's to enhance the library and incorporate it into applications and visualization libraries.

The Parallel Compositing Library simplifies the task of distributing an application by implementing optimal techniques for moving pixels within and between nodes. The Library has been optimized for reading and combining pixels and then transmitting those pixels over GigE and InfiniBand interconnects. This process frees the visualization application developers from learning and implementing interconnect-specific procedures, which may very well be outside their area of interest and expertise.

In a sense, the Parallel Compositing Library does for distributed visualization applications what MPI does for computational applications—it makes clusters a more approachable platform for scalable visualization.

Recognizing that application developers would be hesitant to have a dependency on a proprietary library, HP made this library open source for use and further enhancement by the open community. The Parallel Compositing Library is available on SourceForge.net. The project web page includes pointers to some of the applications that are using this library, including the source code.

In summary, HP is working with its customers and partners to understand the challenges unique to HPC visualization and to deliver cost-effective high-performance solutions:

- HP is providing graphics-enabled systems, based on industry-standard components. Visualization applications can scale-up to the data set and display sizes needed to handle the large data sets produced by HPC compute applications.
- HP also provides a remote visualization capability to help consolidate high-end systems and integrated them with other HPC storage and compute resources.
- With SVA and XC, HP is providing a tested and supported cluster software system that makes it easy to manage and use the shared visualization resources of a cluster.
- The HP-developed Parallel Compositing Library is available as open source. The Library makes it easier for developers to create distributed scale-up visualization applications to handle very large data sets.

# Accelerators

## Introduction

Accelerators are co-processing components containing massive numbers of functional units, together with memory and control systems that can be added to computers to speed up applications. Accelerators are similar to turbochargers in an automobile with the purpose of increasing the speed of an application with a low incremental use in power, for faster time to business, engineering, or scientific outcomes.

## Challenges with accelerator technologies

Now that clock speeds are improving only slowly for process technology reasons, microprocessor vendors propose to offer improved performance by increasing the number of cores per chip.

This approach does not automatically enable microprocessors to increase application performance at the rates we've come to expect. In fact, some codes actually run slower on the new multi-core chips because of contention for system resources. Therefore, many vendors and users are proposing alternative technologies such as General Purpose Graphical Processing Units (GPGPUs), Field Programmable Gate Arrays (FPGAs), and custom ASICs that will deliver substantial increases in application performance on industry standard platforms.

## What is HP doing for Accelerators?

The HP Accelerator Program is an example of innovation that is coming from HP as part of the Catalyst for HPC Innovation Program. In this program, HP is tailoring the ProLiant and Blade server products to accept a wide range of third party hardware accelerators for HPC and making them available to our customers. This helps mitigate risk for customers as HP qualifies different 3<sup>rd</sup> party accelerators and benchmarks results against the servers. This also means HP can offer advice to our customers on which type of hardware platform, with which type of software will work best for which type of application that they wish to accelerate.

## Applications best suited for accelerators

Customers are using accelerators in a number of industries. First and foremost, Financial Services organizations are using accelerators in option pricing, risk modeling, and other computational intensive functions where time is critically important. Biosciences are using accelerators for genetic sequencing and chemistry, molecular dynamics and drug discovery. The government has been using accelerators for a long time to do searching and encryption. Additionally, interest from the Oil and Gas industry has been seen as they seek to locate new energy reserves. And, a number of medical equipment companies are using accelerators for CT scan, ultrasound, X-ray and MRI. The Electronic Design Automation industry has also started to use accelerators and technical tools, such as MathLab and LabView are very amenable to accelerators.

Because the technology is changing so rapidly, customers really need to make a careful choice on when to bring this technology into their organization. Porting an application to an accelerator can take many months to a year. Customers need to plan ahead and if they start too early they might be wasting effort, time and money. But wait too long, and they might find their competitors getting a jump on them. When an application is ported to an accelerator, it can go 10 times faster, 20 times faster and sometimes 30 times faster. For example, it can make a significant difference in a simulation – it might be hours instead of days to complete a simulation. The HP Accelerator team can help gauge when various accelerator technologies are ready, where applications are already ported and which applications will port well.

## Determining which accelerator technology customers' should adopt

There are three basic technologies that HP is supporting in the HP Accelerator Program, and we are investigating all three because each has definite strengths and weaknesses. There is no "one size fits all" accelerator. The first are the General Purpose Graphical Processing Units (GPGPUs), based on the gaming industry. This industry is driving lots of technology to be put into very fast graphics to simulate reality very closely. And, these graphics processors can now do very high performance calculations as well. In this space HP supports two vendors, NVIDIA Tesla and AMD (ATI) FireStream.

Secondly, the Accelerator team is working with a range of FPGA companies. FPGA's are Field Programmable Gate Arrays, which were originally designed for circuit simulation, are now used for high performance computing because there are enough circuits in the FPGAs to do lots of computing in parallel. HP is working with Celoxica, Nallatech, XDI, DRC, and other FPGA vendors to put these FPGA's into very tight packages in HP servers. There is also a third class of accelerators from a company called ClearSpeed, who builds their own ASIC that is a wide pipe ASIC with 96 processors on it and is easily programmable in 'C' and has double-precision and full error correction. A combination of these three technologies gives HP a good range of accelerators to make available to our customers.

## Programming Accelerators

Accelerator programming methods have changed dramatically in the last few years. Programmers used to program the graphics processors, GPGPUs, in a graphics language called OpenGL and now those are programmed in a 'C-like' language. FPGA's today tend to be programmed in circuit design languages, such as VHDL. The industry is starting to see FPGA's programmed in 'C-like' languages and ClearSpeed has been programming in 'C-like' languages from the beginning.

Interfaces to the accelerators are also changing. There used to be slow interfaces such as PCI-X coming out of servers and now there are much faster interfaces. Today, it is not uncommon to see PCIe x16 Gen2 interfaces and HP will soon be plugging FPGA accelerators directly into processor sockets. Additionally, the industry is seeing an increased availability of 64-bit floating point capabilities. ClearSpeed has had 64-bit floating-point capabilities from the beginning; NVIDIA and AMD will be introducing 64-bit floating point this year in their GPUs and that capability is emerging in FPGA's as well.

There are many more application codes that have been ported today. In years past, there were a limited number of applications and now the list is growing rapidly. Accelerators are now easier to program, the accelerator interfaces are faster and the accelerators themselves are faster.

## Accelerators and HP server's next steps

To make sure that our ProLiant line and Blade System line of servers are compatible with third party accelerators, HP has been qualifying them to ensure they have the right form factor, heat, and electrical characteristics. We have incorporated faster express slots (dual PCIe G2 x16), hotter slots (over 200 watts) in PCIe slots and have enhanced the BIOS software to accommodate these new types of devices. The Accelerator team has also done benchmarks to ensure the right accelerator fits with the right server and application.

This work gives HP a body of evidence that we can use as we consult with customers and provide trusted advice to them so they can understand what to benchmark, what to try and when to try it. The Accelerator team also coordinates third party hardware and software futures and influences the industry on the appropriate accelerator technologies and the appropriate software to use. Additionally, HP promotes this as open standards that everyone can use with open programming interfaces. Finally, HP delivers all this with factory integration, and ease of use to make it easy for customers to use the accelerators, and combine them with our United Cluster portfolio.

# Converged Fabrics

## Converged Fabrics overview

Converged Fabrics is a solution that consolidates all communications that includes servers, storage and networks into a single fabric. These fabrics provide sufficient speed to enable consolidation and they are currently based on InfiniBand or 10GigE Ethernet. InfiniBand based converged fabric solutions are more mature and robust than the 10GigE solutions, but it is expected that 10GigE will catch up within the next year as its infrastructure matures. With InfiniBand based converged fabrics, there are multiple usages. First, there is message passing of MPI compute nodes, second there is object-based storage, and third there is InfiniBand based direct-connect storage, such as HP storage blades. Additionally, there are Fiber Channel gateways that provide connectivity to the Fiber Channel SAN storage and the Internet Gateway provides connectivity to both local and wide area networks and now storage.

## Benefits of Converged Fabrics

The biggest reason customers are interested in converged fabrics technology has to do with cost. Converged Fabrics lowers the cost of acquiring the equipment, the cost associated with managing the equipment and the costs associated with running the equipment. It is much simpler and cost effective to take three cabling systems and combine them into one. Traditionally, there would be one set of cables and one set of switches to manage. Converged Fabrics also lowers the power and space requirements to run a system. And, with fewer components, companies are decreasing their liability because there are fewer things to break, while also increasing performance. The channels can go to 10Gb, 20Gb and beyond to 40Gb. In fact, converged fabrics can be an upgrade path for faster storage. The Fiber Channel is 1Gb, 2Gb, 4Gb, and now 8Gb and beyond. The path is to use converged fabrics on faster interconnects to get faster Fiber Channel. It also allows for optimization of multi-core and virtualization so converged fabrics become a path to make those more efficient. HP is finding that the industry is really behind converged fabrics and all major networking and storage companies are participating in different ways to bring these fabrics together.

## Challenges with adopting Converged Fabrics

Converged Fabrics is new technology, InfiniBand and Ethernet are emerging technologies as well. At HP, we receive new hardware and software every quarter for the converged fabrics on InfiniBand. Right now, the industry is expecting new 10GigE Gateways to come to market and the Converged Fabrics team is seeing an emerging small market for Block Level Storage with competing protocols within the InfiniBand range.

There are three different protocols for speeding storage over InfiniBand, eventually leading to some kind of convergence. Observations show that there are larger industry investments in Fiber Channel over Ethernet, than over InfiniBand. For InfiniBand, customers are expected to mainly use the 10GigE Gateways with a smaller number of customers using block storage on them. With Converged Fabrics on Fiber Channel over Ethernet, the standards are still being set. The storage industry is working to bring this technology to market rapidly, but when they bring a new technology to market, storage consumers are very slow to adopt new storage techniques. Given this, it will take a while for Converged Fabrics in the Fiber Channel range to gain popularity and market share. The expectation is for Fiber Channel to make rapid technical progress and get to the market soon.

## Market Segments most likely to adopt Converged Fabrics

InfiniBand currently holds the market lead, especially in science and engineering. In that segment there are large HPC clusters and the expectation is that 10GigE Gateways will be added to those configurations to produce a very popular option. This will give customers very high bandwidth from external LANs into the clusters. Customers are also heavily using NAS and object protocols for storage over InfiniBand in science and engineering

Financial services is looking at using Direct Block Access Storage over InfiniBand in addition to fast messaging, as the industry pushes for storage over InfiniBand into the market. After that, Fiber Channel over Ethernet will mature and that will be a much bigger market than just HPC and Financial Services.

## Learnings from the Catalysts Converged Fabrics project

As part of the Catalysts project on Converged Fabrics, the team built a Proof of Concept of a Converged Fabrics configuration. The goal was to evaluate and compare different solutions including installation, functional stability, performance, management, cost and power and also to gain knowledge and hands-on experience with this technology. The team focused their research on InfiniBand based solutions and collaborated with several InfiniBand partners on the Proof of Concept.

The first learning's from this work was to look at comparable products for converged fabrics for all interconnect vendors and that mainly included switches, Fiber Channel Internet gateways and management tools for managing the entire converged infrastructure. The team learned that in some instances, the entire solution did not reach a production level of maturity. There were issues with installation, performance and stability. When comparing converged with non-converged operations, there were some interesting observations in the performance area. For example, a single LUN showed performance degradation when accessed over the Fiber Channel gateway compared to direct-access. The degradation was in the range of 5-20% depending on the application. However, with multiple LUNs, better performance was observed with the gateway. That was because with multiple LUNs the team could take advantage of InfiniBand speed vs. Fiber Channel speed. Very good performance was observed with Ethernet Gateway and almost no degradation compared to direct access and very good performance scaling as the number of clients increased at the IB/IPoB speed. The team experimented with a mix of different application types, basically running MPI, network and storage applications simultaneously and compared that to the case when running these applications separately. In such a mix, the team was able to reach close to maximum InfiniBand bandwidth. However, they observed that MPI applications have a higher degradation when running in a mix versus running separately than for example network and storage applications.

On the cost and power front, it was observed that converged fabrics can provide substantial cost and power savings relative to the non-converged configuration. For example, in the calculation of a 300/400 node cluster, converged fabrics can provide a 30-40% cost savings and 30% power savings versus the traditional non-converged configuration. These cost and power savings increase with the size of the cluster.

## Applying the Learning's from the Catalysts Converged Fabrics project

As a result from this research the Converged Fabrics team provided recommendations for productizing some of the converged fabrics components. As an example, the Ethernet Gateway has been included in the HPC product list. The team also helped with testing and qualification of some of the gateways, for example, and worked jointly with the storage organization on certifying the Fiber Channel Gateway with the EVA8000 Storage. They also evaluated directly connected InfiniBand storage and concluded that all these products leverage excellent performance for large sequential transfers, but are often not well suited for short random transfers. In addition to the productization decisions, the team engaged with customers on how to best leverage the converged fabrics technology in their environment. There is a lot of interest in converged fabrics from Financial Services customers, but also some traditional HPC customers who are looking into ways to simplify their infrastructure and get cost and power savings by consolidating storage, networks and clusters. In many instances customers would like to consolidate their existing storage, either Fiber Channel or Ethernet based, with their InfiniBand cluster. The team is also collaborating with the Flexible Computing Services Organization at HP for delivering converged fabrics configurations that will allow InfiniBand clients to connect to both the Fiber Channel storage and external networks.

## Next steps for the Converged Fabrics project

The team is looking to continue work on converged fabrics with focus on customer collaborations. Work on emerging technologies such as an NFS-RDMA, Boot-over-IB, provisioning and virtualization with converged fabrics has started. This work has been focused on InfiniBand-based converged fabrics. The team is expecting GigE products to mature and to address cost issues over the course of the year. At that point, 10GigE solutions are likely to become interesting for HPC. To better understand which of the current 10GigE products are best suited for HPC, HP has launched an evaluation study of several 10GigE vendors and are comparing these products on performance that includes both network and MPI, maturity, standardization, storage over Ethernet, power and cost. The team is also looking at the emergence of converged fabric over Ethernet.

# Multi-core Optimization

## Introduction

The microprocessor community for many years has translated Moore's Law of transistor density into a direct doubling of single-threaded performance every eighteen months. Applications ran faster on each new processor version, and new versions were released frequently. Performance tuning of applications required minor experimentation with compilers and tuning flags.

This was a period of high productivity for application developers, since they could concentrate on product functionality and performance and minimize the time to create, tune, test, and support computer-model-unique versions. It was fun, but it could not last forever.

## Multi-core Challenges

Today, the era of single processor systems is over. The multi and many core systems world is here. We're entering a phase where taking full advantage of the power of multi-core processors is critical for customers to continue to accelerate innovation and to improve their business success. Dual-core technology is now pervasive in the industry; quad-core processors are here and about to become the new standard for server nodes and roadmaps pointing to octal-core processors are not far off.

For applications not designed to take advantage of the increased raw compute power that comes with the availability of the added cores, applications may run slower due, at least in part, because of contention issues for shared system resources. This likelihood increases as the core count increases.

With the focus on hardware, and increasing processor counts, there is increasing need to understand the new complexities of application design, debug, and optimization in multi-core systems. In order to take advantage of the additional processing power that multi-core systems offer, new parallel code development languages, development environments and parallel execution environments are needed that allow the applications to change as well. Exploiting the power of multi-core processors will be critical for customers to improve their business success.

## How did we get here?

The introduction of dual-core processors, then quad-core processors traditionally gets the blame for making it more difficult and complex to program applications. However, they also get the credit for keeping power and cooling requirements at reasonable levels. The real issue is to balance system power, cooling, I/O, memory and cache. To meet new system balance requirements for power and cooling, the clock speeds declined: the more cores per processor chip, the lower the clock frequency. Moore's Law continues, but the additional transistors are used to implement more cores and larger caches.

As a result of this, the problem moved from a HW problem of making things run faster with faster clock cycles to a software problem – and how to use all the additional cores (raw compute power) now available on the chip to improve performance. Unfortunately, this created coding problems for application developers. A multi-core processor can do more work than a single-core processor, so the total amount of work, in compute jobs per month, increases on multi-core-based servers. But without taking into consideration the multi-core nature of today's systems, the performance of an individual application will not increase – it is likely to run more slowly as the number of cores increases, due to the combination of lower clock rates and competition for memory bandwidth and cache. And it's not just the applications. Sending all your communications interrupts for example to one core could overwhelm that core possibly slowing down the rest of the system. It's that system balance thing again.

## Solving the application performance problem

There is no easy solution to application performance. Serial (non-parallel) applications in many cases cannot become parallelized without considerable work and time, possibly even complete rethinking of basic algorithms. Some HPC applications are parallel, and some are highly scalable and can run faster if it is possible to allocate more cores to their execution. But other parallel applications are not very scalable, with the same performance barrier as serial applications.

The best way to make progress is to understand how an application uses system resources. With this knowledge, both developers and users of applications can improve performance. It's important to

look at the resource usage at the server level, not just the processor level. Much of the available data comes from the processor developers, but to understand application performance, the complete server must be analyzed. Important resources include – memory bandwidth per core, I/O bandwidth per core, network bandwidth per core, amount of memory per core, and amount of cache per core.

Shared caches are also a complicating factor in application performance. New x86-64 processors share cache among two or more cores. As a result, it is not possible to know the amount of cache being used by one core at any one time. Shared caches can be both friend and foe to code performance. With analysis and work, application developers can take advantage of this low latency opportunity. But many codes are tuned for some minimal amount of cache per core, and application performance will suffer if less is available. Erratic application runtimes will be one symptom.

One of the solutions to application performance is to parallelize more code. A roadblock to developing multi-threaded programs exists: the uncertainty and confusion about the rules that must be followed by both users of, and compilers for, the C++ and C languages. As a result, it is much harder to implement shared memory parallel programming. New languages may be necessary, languages that make it easier for average programmers to create efficient parallel applications, thereby reducing time to solution. HP's work on PGAS (Parallel Global Address Space) languages like HP UPC (Unified Parallel C) and HP SHMEM addresses this problem.

HP is leading an effort to address this issue by specifying how multi-threaded C++ programs may interact through shared memory. HP has developed a proposal for the upcoming revision to the C++ standard. This proposal supports both a simple model that requires no understanding of hardware or compiler optimizations.

Job management software tools can also be implemented to aid performance. An application will run faster if it is intelligently scheduled onto servers that satisfy the code's resource requirements. HP works with its partners like Platform Computing, open-source tools like SLURM, and HP software products like HP-MPI to design and implement this functionality.

## What is HP doing for Multi-core?

To help customers address Multi-core issues, HP has created the Multi-core Optimization Program and the Multi-core Optimization Program Toolkit. The toolkit serves as a clearinghouse for information and resources to help with building, testing, tuning and deploying industry-standard multi-core applications. Nearly two dozen partners already contribute, with more coming on-stream.

The toolkit sorts multi-core issues into categories such as application development and compilers, performance and tuning tools, job scheduling and libraries, and more. The toolkit web site is subdivided according to these categories, and within each category users will find links to HP, partner, and open source products and tools. There is also a section on best practices for each category, offering white papers, efficiency processes, demos and multi-core 'do's and don'ts.' Topics include power consumption, application performance on HP servers that use multi-core processors, measuring power against performance as cores are added, and more.

## Learning's from the Multi-core Catalyst project

HP has been independently and with our partners studying and benchmarking a cross-section of applications that can provide information that applies to broad application sets. For instance, the team has completed characterization work in the areas of application energy, job scheduling and performance analysis. HP has collaborated with technology partners for decades, measuring and improving application performance. HP partners include well established industry giants such as Intel to those that specialize in specific technology such as multi-core debugging with TotalView to emerging technologies such as those from Acumem that provide performance tuning via an application finger print for multi-core systems. The multi-core work has allowed the team to develop numerous white papers to help customers and will be expanded to broader industries.

## Multi-core Next steps

Future plans for HP and our multi-core partners include a reference model for the development environment, research into non-traditional uses for multi-core systems, investigation of what new development tools and capabilities will be required in a many-core world, and dedicated cores designed for better security, ease of management and other mission-critical roles.

# Power & Cooling

## Introduction

Power and cooling in today's data centre environments contributes significantly to TCO. Power & cooling cost are therefore becoming more important in the selection of cluster hardware & software for HPC applications. We are moving into an era of "performance per watt", rather than just "performance."

## What is HP doing for Power and Cooling?

Today's generation of HP's industry standard server blades come with inbuilt power saving and cooling capabilities. For example, HP Thermal logic technologies. These include HP PARSEC architecture for centralized cooling, Active Cool fans for efficient cooling and other capabilities. Processors support CPU frequency scaling to permit trade-off of performance and heat dissipation, informed by embedded sensors that provide vital information on thermal and power status of the system. Data centre cooling technologies like DSC (Dynamic Smart Cooling) help in further cost reduction.

## Challenges in HPC:

The challenge is to leverage some of the capabilities provided by server blades in HPC environments (at the cluster level) to improve operational efficiency and lower TCO, without significant degradation of performance. The overall goal is to increase computational throughput per dollar per watt. Efforts are being made to understand if the cluster resource managers and schedulers (like Platform LSF) can be integrated into an intelligent active power and cooling management schema.

For instance, cluster resource managers could be made to choose efficient blades to run jobs based on sensor information and HP PARSEC architecture in a pool of nodes. There is also an opportunity to scale up the frequency of the processor when the scheduler allocates jobs to the nodes and scale down when the job completes. This gives the flexibility of running jobs with full CPU power, which is the necessity for HPC applications and save power when the nodes are sitting idle.

HP has also contributed the open source "power save module" of SLURM resource manager which could be used in future versions of HP's XC cluster management software stack.

## Power & Cooling next Steps

Future plans of HP include study and research in the area of hibernation of server nodes in HPC environments. This involves an understanding of the impact of power cycles on hardware components and demonstrating net positive impact on total operational cost.

# Dense Computing

## Introduction

The high cost of power and floor space within a data center is a large and growing concern for the computing industry. Aggregate electricity use for servers, cooling and infrastructure doubled over the period 2000 to 2005 to 1.2% of total US energy consumption. As energy costs climb past \$0.10 per kilowatt-hour, new data center solutions are required which address power conservation, cooling, and space requirements without sacrificing performance or cost.

Power and floor space costs are a particular concern for High Performance Computing workloads. Large cluster and server farm configurations make power costs a significant line item in the budget. Enabling performance growth, particularly now that process technology improvements no longer translate directly into higher processor speeds, can require a substantial investment in additional power and data center real estate.

The shift from faster cores to multiple cores also has an unintended consequence of limiting the memory and I/O bandwidth available to each core. While transistor counts continue to increase, following Moore's Law, resulting in a regular increase in the number of cores per die, memory and IO bus frequencies are not increasing at the same rate. This results in a new class of processors with more raw compute capacity but less memory and IO bandwidth per FLOP than the previous generation of chip. For memory and IO intensive workloads, this can result in lower application performance per chip despite a doubling of the number of cores.

Severs and solutions are needed that can meet the need for growing performance within the constraints of current process technology trends and with substantially lower operating costs. The historic focus on optimizing performance must now shift to a focus on optimizing performance per watt (ops/watt) and performance per square foot (ops/sq ft). These solutions must simultaneously push the boundaries of raw performance, to enable continued growth for HPC workloads within the power and floor space constraints of existing data centers.

The Dense Computing project is researching novel system architectures and new rules for server design to provide substantial improvements in ops/watt and ops/sq ft.

## Dense Computing Solutions

Current server roadmaps are already responding to the need for lower power in the data center. To provide a substantial gain over these planned improvements, new architectures and lower power processors are needed.

A new class of processor is becoming available that is optimized for low power consumption at full utilization. High end embedded processors are becoming more capable, driven by the needs of multi-layer switches, network processors, game consoles and ultra-mobile PCs. With these processors, server designs with dramatically higher performance per watt can be achieved, despite having less compute capacity per core than high-end server CPUs.

In addition to being optimized for low power utilization, high-end embedded System-on-a-Chip (SoC) processors integrate the memory controller and key I/O components directly onto the processor chip. These processors contain the functions typically performed by the North Bridge (processor to memory) and South Bridge (processor and memory to IO) chip components in a traditional architecture. Use of these processors allows server designs without additional chipsets, and, by using the embedded IO components, without PCIe slots. Elimination of the chipset components and IO slots provides a substantial reduction in power and board area on the server mother board. This board area is now free to be used for additional processors and memory, increasing the compute density of server design.

The integrated memory controllers on this class of processor, coupled with lower core counts per chip compared to high-end server processors, provide substantially higher memory bandwidth per op than the latest multi-core CPUs. Integrated IO components also provide higher IO bandwidth per op ratios, with the added performance advantage of lower IO latencies due to the physical proximity of memory and IO on the processor chip. While current trends in process technology will continue to drive all classes of processor to increasing numbers of cores with limited corresponding increases in memory

and IO performance, SoC processors have and will maintain an advantage in memory and IO bandwidth per op that can benefit memory and IO intensive applications.

## Enabling Technologies

Typical server designs are optimized to serve a broad range of workloads well. Server economies of scale demand that a given design be reused as much as possible. HP's c-Class Blade infrastructure breaks this rule, offering many different types of blades (x86, IA64, storage, etc) in a single, common architecture, enabling a special purpose blade focused on HPC application requirements. The advanced thermal management features of the c-Class allow extremely dense server configurations within the thermal design envelope of the c-Class chassis. The c-Class management infrastructure (Onboard Administrator, iLO) allows integrated chassis management (thermal sensors, automatic fan speed adjustment) and system management (console and administration) for a new specialized blade type.

## Dense Computing Challenges

The difficulties in designing a server that is fundamentally better in the desired metrics of operations per watt (ops/watt) and operations per square foot of floor space (ops/sq ft) than the main stream server roadmap are compounded by several factors:

**CPU suitability:** Most low-power processors are designed for commodity applications where double precision floating point, advanced architecture features, and server-class reliability are not required.

**Design issues:** Building motherboards with larger socket counts places significant constraints on the design. Increased chip counts, memory modules, and IO components all add power, complexity and cost, in addition to contributing to airflow and cooling issues.

**Memory and IO issues:** In addition to the memory and IO bandwidth issues discussed above, many HPC workloads are sensitive to IO latency and the amount of available system memory. The system must balance the amount of memory available to a processor, expressed as a number of DIMM sockets, against providing additional compute resources (more processors) in the same physical space on the system board.. Similarly, IO slots provide flexibility but consume significant board space and power. Embedded networking eliminates slot issues but typically supports Ethernet, not InfiniBand, which provides very low latency to support a broader HPC application set.

**Target workloads:** A significant advantage over general purpose servers can be found if a server design targets a specific set of workloads. For those workloads, the processor and design must support the key applications, operating systems and development tool chain.

## Dense Computing Solution Benefits

Dense Computing server designs will drive down data-center total cost of ownership by directly reducing power and cooling costs and increasing data center space efficiency. SoC-based solutions that integrate IO components also eliminate NIC or HBA costs. Elimination of the complexity of additional chipset components and IO slots also provides fewer components and potential points of failure compared to commodity servers.

SoC solutions also provide significant advantages for scale-out workloads that are sensitive to IO or memory bandwidth. Integrated IO components can provide twice the IO bandwidth per op of traditional processors. Memory controllers on high-end SoC processors can provide three to four times the memory bandwidth per op of a quad-core CPU.

The Dense Computing project has worked closely with customers to validate the Dense Computing solution by modeling the performance of specific HPC customer workloads. Initial results show that workload focused server designs yield significant improvements in both ops/watt and ops/sq ft over the standard sever roadmap and can maintain that advantage over time. HP strongly believes that this line of research is leading to fundamentally new architectures that will delivery breakthrough performance for HPC applications while delivering on the promise of high performance with low power in less space. This research promises significant advances in ops/watt and ops/sq ft for the HPC data center, with additional benefits in reduced cooling costs, simpler infrastructures, higher reliability and higher raw performance for memory and IO constrained workloads. Continued

research, working closely with customers and partners, will allow the growing performance needs of HPC environments to be met with dramatically lower total cost of ownership.

## Next Steps

Additional work is planned to broaden the set of scale-out customer workloads evaluated on the current prototype design. Research is also planned on new specialized server designs focused on optimizing ops/watt and ops/sq ft for new workloads and deployment environments.

## For more information

### HP High Performance Computing

[www.hp.com/go/hpc](http://www.hp.com/go/hpc)  
[www.hp.com/go/catalysts](http://www.hp.com/go/catalysts)

### Grid

[www.hp.com/go/grid](http://www.hp.com/go/grid)

### Visualization

<http://www.hp.com/go/visualization>  
[http://www.hp.com/techservers/hpccn/sci\\_vis/index.html](http://www.hp.com/techservers/hpccn/sci_vis/index.html)  
[http://www.hp.com/techservers/hpccn/hpccollaboration/ADCatalyst/Parallel\\_Compositing\\_Library.html](http://www.hp.com/techservers/hpccn/hpccollaboration/ADCatalyst/Parallel_Compositing_Library.html)  
<http://sourceforge.net/projects/paracomp/>

### Accelerators

<http://www.hp.com/go/accelerators>

### Converged Fabrics

[http://www.hp.com/techservers/hpccn/hpccollaboration/ADCatalyst/Converged\\_Fabrics.html](http://www.hp.com/techservers/hpccn/hpccollaboration/ADCatalyst/Converged_Fabrics.html)

### Multi-core Optimization

<http://www.hp.com/go/multi-core>

### Power & Cooling

<http://www.hp.com/go/catalysts>  
<http://www.hp.com/go/bladesystem/thermallogic>  
<http://www.hp.com/go/dsc>  
[https://computing.llnl.gov/linux/slurm/power\\_save.html](https://computing.llnl.gov/linux/slurm/power_save.html)

### Dense Computing

[http://www.hp.com/techservers/hpccn/hpccollaboration/ADCatalyst/Dense\\_Computing.html](http://www.hp.com/techservers/hpccn/hpccollaboration/ADCatalyst/Dense_Computing.html)

© Copyright 2008 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

Linux is a U.S. registered trademark of Linus Torvalds. Microsoft and Windows are U.S. registered trademarks of Microsoft Corporation. UNIX is a registered trademark of The Open Group.

4AA2-1263ENW, September 2008

