3D Virtual Desktops that Perform

Consolidating your Desktops with NeXtScale System

Sponsored by IBM

Srini Chari, Ph.D., MBA mailto:chari@cabotpartners.com
September, 2014

Executive Summary

Scientists, engineers, and analysts routinely use High Performance Computing (HPC) or Technical Computing scale out systems that add additional compute resources to improve the performance of challenging applications across many industries. To get maximum utilization from these systems, many companies are also using these same systems to replace and consolidate end user desktops by implementing virtual desktop infrastructure (VDI) or desktop virtualization.

Desktop virtualization provides end users access to their applications and data anytime, anywhere and from any browser capable mobile device. This improves end user productivity, promotes greater collaboration and enhances security while lowering IT costs and systems management complexity. As a result, desktop virtualization is being implemented in many enterprises to manage the proliferation of compute environments and end point devices. Even graphics and video-intensive workloads that traditionally required dedicated high performance Graphics Processing Units (GPUs) in the desktop are being virtualized.

GPUs are already being used to augment or supplement server CPUs in HPC systems to achieve greater overall performance in many demanding technical computing applications. With GPU virtualization, three-dimensional applications that are video-intensive or graphics-intensive can also perform well in VDI environments. Computer Aided Design/Engineering (CAD/CAE), Radiology, Multimedia Publishing and Gaming are some key areas that can benefit from accelerated rendering of 3D scenes.

The IBM NeXtScale System provides a hyperscale (high-density and large scale out) computing platform designed for high performance for both cloud and traditional environments. With GPU capabilities provided by NVIDIA, it is an excellent platform for supporting many demanding workloads that are compute, data, video and graphics-intensive.

Other companies' product names, trademarks, or service marks are used herein for identification only and belong to their respective owner. All images and supporting data were obtained from IBM or from public sources. The information and product recommendations made by the Cabot Partners Group are based upon public information and sources and may also include personal opinions both of the Cabot Partners Group and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. The Cabot Partners Group, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your or your client's use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document. This paper was developed with IBM funding. Although the paper may utilize publicly available material from various vendors, including IBM, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

The converging worlds of HPC, Big Data and VDI

HPC is progressively becoming crucial to businesses to achieve insight faster, reduce time and cost, and gain competitive advantage

The boundaries between high performance computing (HPC) or technical computing and business computing are blurring as enterprise computing needs are becoming more data and compute intensive. Traditionally, high performance technical computing has been associated with large problems in scientific research, weather prediction, astronomy, and so on.

But, in recent years, HPC has also been increasingly used to design automobiles, simulate car crashes, improve financial risk analytics, create animation for movies, package snacks and coffee, maximize golf club striking distance, minimize bathing suit drag, and so on. In other words, HPC is progressively becoming part of business computing where the focus is to achieve insight faster, reduce time and cost, and gain competitive advantage. Additionally, there are new opportunities in High Performance Data Analysis (HPDA), such as fraud detection, anti-terrorism, emergency response analysis, etc., forecasted to reach $1.4 billion[1] in 2017 in server revenues.

As data explodes in velocity, variety, and volume, it is getting increasingly difficult to scale compute performance using enterprise class servers and storage in line with the increase in data volume and complexity. This is further triggering the rapid rise of scale out computing. Scale ‘out’ refers to increasing performance by adding more systems or resources. This is in contrast to scale ‘up’ which refers to the process of raising a single system’s performance.

GPU and desktop virtualization using scale out systems enable a new class of high performance video and graphics-intensive applications

Scale out computing enables organizations to start small and scale their systems as needed. For example, at the Universidad De Chile, earthquake prediction and astronomical modeling previously done in weeks are now done in 2 hours; and Mac Guff, the visual effects firm collaborated globally and rendered the Despicable Me movie on a relatively small cluster.

Both scale out and scale up as a replacement for individual work stations in virtual desktop infrastructure (VDI) or desktop virtualization allows collaboration across borders and can accelerate products and services delivery and innovation.

Desktop virtualization - the technology to separate the desktop environment and associated applications from the physical client device used to access it - is not a new concept. But the proliferation of compute environments and devices has changed desktop virtualization from a good-to-have to a must-have feature for today’s enterprises.

Virtualization of Graphics Processing Units (GPUs) is one leading emerging technology that is being increasingly adopted by companies in the manufacturing, higher education, healthcare, multimedia, gaming and architecture, and engineering industries. These organizations tend to require three-dimensional, video-intensive or gaming or graphics-intensive applications which only perform well in a VDI environment, that has some kind of processing offload in place, such as with GPU virtualization.

NeXtScale with NVIDIA GPUs is an excellent platform for HPC, Big Data and VDI workloads

The IBM NeXtScale System provides a hyperscale (high-density and large scale out) computing platform designed for high performance for both cloud and traditional environments. With GPU capabilities provided by NVIDIA, it is an excellent platform for supporting the convergence of HPC, Big Data and VDI.

[1] IDC Revenue for HPDA market http://www.scientificcomputing.com/blogs/2014/03/high-performance-data-analysis-big-data-meets-hpc

NeXtScale brings dense computing to business

The NeXtScale hyper-scale out architecture provides flexibility, simplicity and scalability for all environments

NeXtScale brings high performance computing power to business computing. It is an x86 offering that debuted in 2013 and introduces a new category of dense computing. The building blocks include dense chassis and half-wide compute, storage, and GPU nodes. The configuration is built for a standard rack platform with a 1U half-wide architecture.

The system extends the IBM System x solution – including the x3100, x3250, x3530, x3550, x3630, x3650, x3690, x3750 and x3850 – to large scale systems. The guiding design principles are flexibility, simplicity, and scalability. The system directly targets the general-purpose server market, delivering what IBM calls "scale for everyone". Customers can start small or large with NeXtScale; if they choose to start small; they can easily and rapidly expand the architecture as their needs increase.

GPU virtualization key for enabling three dimensional VDI

Desktop virtualization provides access anytime, anywhere and from any device

One of the dominant trends in business computing today is the increasing adoption of virtual desktop infrastructure (VDI) as an answer to the desktop dilemma. On one hand, IT managers are under pressure to control costs and ensure compliance, manageability and security. On the other hand, end users increasingly require the freedom and flexibility to access their applications and data from multiple devices and locations. This dilemma – end-user freedom versus the need for IT control – can drive up costs, impact security, and overwhelm IT resources. Desktop virtualization addresses this dilemma by decoupling the desktop, applications, and the operating system from the endpoint device and by managing them as centralized services. Organizations can thus apply policies and quickly enable or disable users from a centralized console.

One of the barriers to a wider adoption of desktop virtualization in technical computing is that until recently, VDI couldn't deliver the type of graphics performance that end users get from a PC. In VDI environments, graphics are typically rendered in the server CPU and then delivered to the end user. But rendering graphics is a parallel computing problem, not a serial computing problem that CPUs can handle, so graphics performance has been subpar.

NVIDIA's pioneering virtualized GPUs on scale out systems provide excellent 3D graphics performance to remote users

NVIDIA recently changed that with low latency access to virtualized GPU resources. GPU virtualization allows portions of the GPU to be assigned to different applications or users, supporting cost effective, high performance remote 3D or graphics acceleration. In virtual and remote desktop environments, it is difficult to render and deliver complex graphics to endpoints with adequate performance. The standard GPU was originally developed to offload processing calculations from the CPU for graphics-intensive applications.

NVIDIA introduced the first virtual GPU in 2012 to help solve that problem, reducing lag time when delivering graphics to remote users and providing the same performance they would get from a PC at competitive cost points. By adding the virtualized GPU to the server, graphics circumvent the CPU and go directly to the virtual GPU, then out to end users. This means that graphics-intensive applications such as computer-aided design/engineering (CAD/CAE) can be delivered to remote users using VDI without sacrificing performance. The ability to share the GPU resources, as can also be done with CPU, memory and networking resources, results in high performance, cost effective solution infrastructure for 3D VDI environments.

How GPUs help

GPUs enhance performance by massive parallelism and superior floating point features

GPUs augment or supplement the server CPUs to achieve greater overall performance and/or efficiency by leveraging massive parallelism, superior floating point capability and on card memory. The sequential part of the application runs on the CPU. The computationally-intensive part runs on the GPU:

- Applications run faster because they can use the high-performance of parallel cores on the GPU
- Many codes and algorithms in HPC and other applications benefit from parallel floating point calculations
- GPU's are able to handle many general purpose scientific and engineering applications.

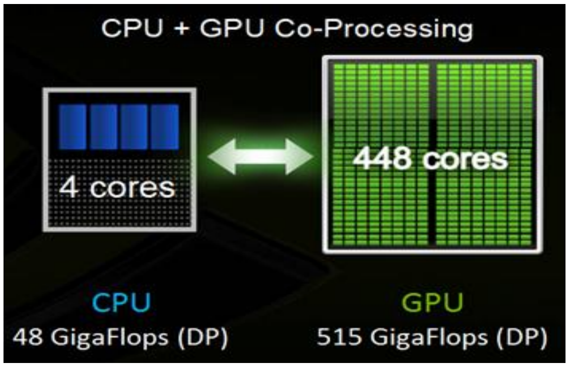

Figure 1: x86 CPU + GPU work together in a heterogeneous computing model

NeXtScale accelerates many workloads

NeXtScale provides a single architecture for high performance compute, storage and graphics acceleration

The NeXtScale System can accommodate NVIDIA Tesla GPUs or Intel Xeon Phi co-processors to achieve extreme acceleration.

NeXtScale includes a half wide, 1U tall PCI Native Expansion (NeX) Tray with two full height full length PCI slots for a range of NVIDIA Tesla and GRID GPUs and Intel Xeon Phi co-processors. Users can expect to see significant benefit from the single architecture for compute, storage, and graphics acceleration and sweet spot GPU density with up to 2 per node.

The usual NeXtScale System benefits will also apply – a simple, light chassis that is designed for 'front-of-rack' servicing, tool-less access to servers and server removal without touching its power. The compute, storage and PCI-GPU/GPGPU components are designed to swap easily, mix and match in standard configurations. All are compatible with standard racks. Storage and GPU or co-processing expansion units make upgrading easy without any unique mid-plane dependencies. All Power and LEDs are forward facing. Networking cables and Switches are front facing and direct to system with no proprietary switching. All switching is done at the top of the rack.

Figure 2: PCIe Native Expansion (NeX), 1/2 Wide, 2U Tall

Some illustrative use cases

The NeXtScale System, with its ability to handle graphics intensive workloads, can be deployed across a wide range of applications and industries where it is important to both process big data efficiently and to visualize results remotely. 3D graphics and resource intensive applications are common in manufacturing, construction, media and entertainment, electronics, healthcare, oil and gas, and many others. Here is a cross-section of use cases:

NeXtScale with GPUs is ideal for many applications and industries that need to efficiently process big data and visualize results remotely

Engineering design: Collaboration is critical to product development today. Design activities requiring access to 3D data can be spread around the globe, and this raises security, bandwidth, and performance concerns. Therefore, engineering organizations are already working to deploy remote workflow for a centralized engineering HPC/private cloud. The user is remote from both the compute resources and the data which are managed by a central server (Figure 3). The files reside at the HPC resource leading to better efficiency, data security, enhanced collaboration, and data access. Engineering enterprises can now conduct a full simulation via remote access using graphics servers co-located with the central computing and data storage, all in NeXtScale. IBM has in fact designed several Application Ready[2] solutions for prominent engineering applications including ANSYS, MSC Software and Abaqus.

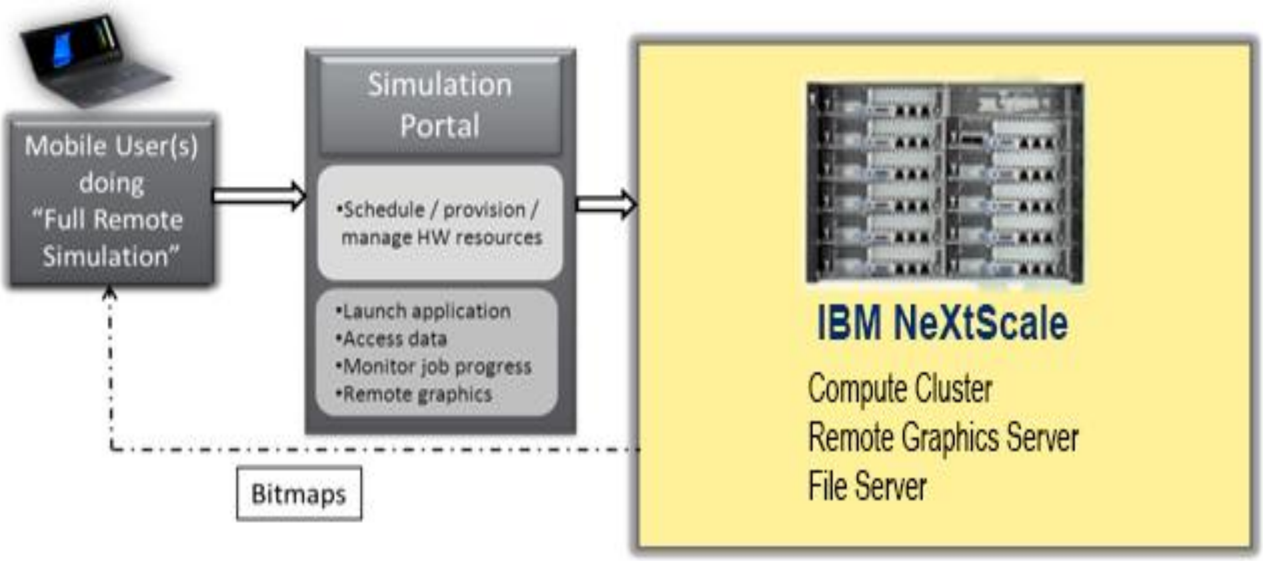

Figure 3: Engineering design through remote interactive access

[2] http://www.ibm.com/systems/platformcomputing/solutions/appready.html

NeXtScale with NVIDIA GPU co-processors makes it possible to virtualize graphics-intensive workloads from remote locations; deepening user collaboration for many high-value applications

Radiology: Modern radiology is now almost entirely digital. Present day approaches inevitably involve long sequences of multiple images. The current "big four" in terms of data footprint are CR (Computed Radiography), CT (Computed Tomography), MRI (Magnetic Resonance Imaging) and Digital Mammography. A single CT study can routinely involve 10,000 or more images. Image processing has made it possible for radiologists to perform advanced cardiac and vascular image manipulation, and volume rendering remotely from the desktop. Such collaborative work across locations has great potential for improvement through GPU-based graphics processing and VDI.

Publishing: Creation and publishing of high-value content is a very collaborative effort between writers, graphics and animation artists, digital content editors, magazine editors and others. This is similar to how legendary Hollywood film production crews work but without the need to be at the same physical location for extended periods of time. Publishing teams must develop new content to delight audiences across multiple channels, venues and devices. For this, they must collaborate and share images, videos, text and other content in real-time to meet stringent production deadlines. With GPU virtualization, all team members can work from home or in the field using a laptop or tablet. Further, stateless (non-persistent) virtual desktops provide more streamlined IT management and reduce costs and complexity.

Summing up

The convergence of HPC and desktop virtualization has the potential to deliver powerful compute and storage capabilities to business and technical computing environments. Additionally, businesses get significant benefits from the centralization of IT resources, applications and data in a cloud. The one barrier to a wider adoption was the limited graphics processing capability available until now, at competitive cost points, in 3D VDI environments. NeXtScale System, with NVIDIA GPU co-processors, overcomes that limitation by making it possible to virtualize graphics intensive workloads. This opens up VDI to a wide range of applications ranging from engineering, media and entertainment, distance learning to healthcare applications where knowledge workers need to collaborate from remote locations.

For More Information

For more information on the NeXtScale System, visit:
http://www.ibm.com/systems/x/hardware/highdensity/nextscalem5/index.html

Cabot Partners is a collaborative consultancy and an independent IT analyst firm. We specialize in advising technology companies and their clients on how to build and grow a customer base, how to achieve desired revenue and profitability results, and how to make effective use of emerging technologies including HPC, Cloud Computing, and Analytics. To find out more, please go to www.cabotpartners.com.