



I D C T E C H N O L O G Y S P O T L I G H T

The Rise of Grid-Based High-Performance Computing: A Cost-Effective Approach to HPC Acquisition

May 2010

Adapted from *Worldwide Technical Computing Server 2010–2014 Forecast* by Jie Wu, IDC #222604

Sponsored by Techila Technologies

Large organizations increasingly are using high-performance computing (HPC) to shorten decision-making, product design, testing, and research and development cycles. Even smaller organizations are beginning to realize the value of computer modeling and simulation. As the economy slowly rights itself, organizations increasingly recognize that efficiency plus innovation will be the key to competitiveness. As a result, organizations are looking for new cost-effective ways to acquire HPC resources. This Technology Spotlight describes overall HPC market dynamics and, in particular, grid-based HPC offerings as one of the new approaches to access HPC capabilities. The paper also discusses Techila Technologies, an emerging supplier offering next-generation grid-based HPC solutions. Finally, this Technology Spotlight outlines some key considerations for organizations looking to implement HPC solutions.

Introduction: The Rise of HPC

As economic indicators point to the beginnings of an economic recovery, it's imperative that organizations in commercial and academic sectors not only maintain maximum operational efficiency but also again look to innovation to compete in a global market. This is especially true for industries with long, complex research and product development and design cycles. Organizations must use technology in the design and development process to reduce costs and shorten time to market.

Organizations are increasing their dependence on HPC and simulation to reduce research and design cycle times and lower development, reengineering, and material costs. In particular, HPC is considered indispensable for the following reasons:

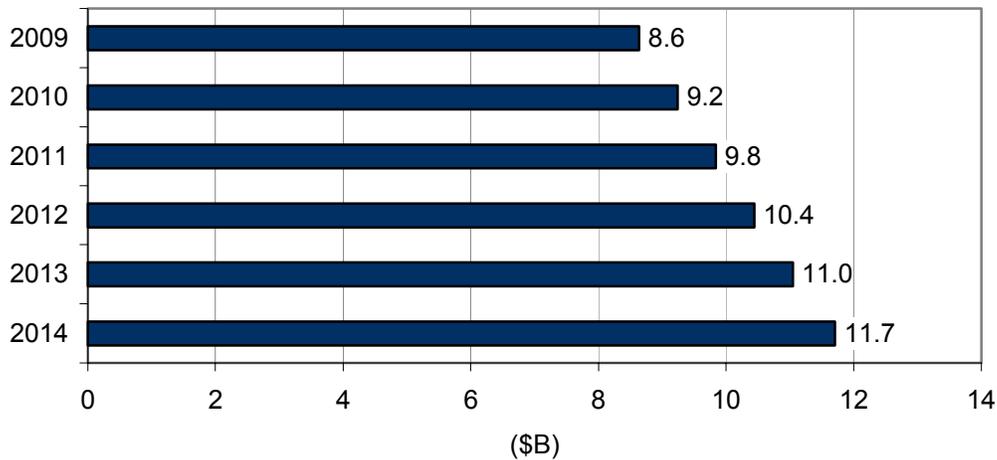
- HPC plays a key role in designing and improving many products, from aircraft to new chemicals and pharmaceuticals, autos, and even tennis racquets
- HPC is also being used to streamline processes from modeling complex financial scenarios to finding and extracting natural resources, manufacturing, planning store inventories, and weather forecasting

IDC believes that continued use of HPC is critical to success in an economy that thrives on continued innovations and improved productivities. In many cases, the traditional methods of research, product development, and business process management are being thrown out, and new methods, using more modeling and simulation with HPC, are becoming competitive tools. This technical computing can move scientific research beyond what is feasible in a physical laboratory, enables first principles-based science, and makes available research at the molecular or even nano level. In short, the concept of analysis by simulation is becoming more possible with HPC.

Based on this demand, IDC estimates that the HPC server market will show healthy growth in the next five years (2010–2014), expanding at a CAGR of 6.3% to reach \$11.7 billion in revenue by 2014 (see Figure 1).

Figure 1

Worldwide Technical Computer Market Revenue, 2009–2014



Source: IDC, 2010

A groundbreaking IDC study, conducted for the Council on Competitiveness, found that 97% of the firms that had adopted HPC said that they could no longer compete or survive without it. This continued realization of the value of HPC will drive the market, especially in the mid-to-low end of the market segment. Solutions that help users gain more performance and efficiency out of their existing resources will be appealing, particularly as organizations are forced to quickly respond to pent-up demand for innovative products as the economy transitions out of the downturn.

The major pain points in HPC today are often associated with high acquisition costs and associated power, cooling, and storage demands; system management complexity; the training of staff to maintain and support these often proprietary environments; as well as a lack of ease of use from a user standpoint. Software challenges also will rise as organizations strive to apply these resources to solving many corporate problems. Therefore, organizations seek solutions that provide the much-needed processing power to solve complex problems while addressing issues such as better price/performance, energy efficiency, ease of use, software scaling, I/O performance, and lower total cost of ownership.

From Conventional HPC to Grid

IDC uses the terms "technical computing" and "high-performance computing" to encompass the entire market for computer servers used by scientists, engineers, analysts, and other groups using computationally intensive modeling and simulation applications. Technical servers range from small servers costing less than \$5,000 to large-capability machines valued in the hundreds of millions of dollars. In addition to scientific and engineering applications, technical computing includes related market/application areas, including economic analysis, financial analysis, animation, server-based gaming, digital content creation and management, business intelligence modeling, and homeland security database applications. These areas are included in the technical computing market based on a combination of historical development, application type, computational intensity, and associations with traditional technical markets.

Many organizations have used HPC to create digital models of products or processes and to evaluate and improve their design by manipulating these computer models. The transition from the physical test to digital prototyping has brought significant benefits to organizations, including faster product to market and the ability to do more and better science faster, all at a much lower cost compared with using conventional methods.

Parallel Computing to Clusters and Grids: HPC Evolves

Parallel computing was originally touted as a way to capture virtually unlimited computing power at a very low cost by linking tens of thousands of inexpensive processors to solve a single problem. As networking technology improved, the concepts of clustered systems and grid computing emerged. In general, both clusters and grids use independent computer systems that could operate on their own outside the system with minimal additional modification; within the system, those independent computers can be built on a heterogeneous platform, and they are connected via standard interconnect technologies. However, in a clustered environment all components are exclusively dedicated and managed as part of the HPC system, with all resources known and fixed, and a dedicated interconnect used between the nodes. A cluster system is located at a centralized site and is owned by a user or user group.

Grids are virtual systems in that the computing resources within a grid could be geographically dispersed and owned by multiple owners and the resources being used are transparent to the end users. IDC believes that grid computing in HPC will rise as many organizations look for a lower-cost approach to access HPC capability. This concept is not new; organizations have been extending distributed computing to share and combine resources in an attempt to meet capacity and capability computing needs, especially in virtual computing systems that can span organizations and geographies.

As the increase of enterprise applications and virtualization causes "server sprawl," enterprises are combining their desire for HPC with their drive to optimize resources within their organizations. The new wave of grid-HPC adoption is also driven by improved capabilities in grid management software, in particular, the much-enhanced security and workload management features.

The Vision for Grid-Based HPC

The vision for grid computing is that organizations can harness much of the spare CPU cycles of existing computers, from single PCs to more powerful resources such as supercomputers, to create a universal source of pervasive and dependable computing power that will support scientific, technical, and other large-scale applications. Grid-based HPC also provides a viable solution to address the rising cost of power and cooling as well as real estate investment — all major issues faced by many HPC sites today. With a grid-based HPC approach, the incremental cost of power and cooling the grid is comparatively free because organizations already are paying for and using those resources. The need to acquire additional real estate is also eliminated. All of this helps create a more efficient and green IT environment.

To many market observers, grids symbolize the evolution of computing architecture into a utility-like source of computing power that companies can use or purchase as needed. To the average IT manager running the datacenter at a research lab, grids represent a way to improve performance on certain applications, especially embarrassingly parallel applications, by distributing bits of the application to otherwise idle servers and desktops sitting within his or her controlled domain. To those in need of technical computing, grids are a way to get the horsepower they need without breaking the bank. To the IT executive, grids are a price/performance play that harnesses spare horsepower from mostly existing, commodity-level hardware components.

Because grid HPC is a dynamic complex of systems in which component resources can change at any time in any geographic location, it can be viewed as a virtual supercomputer. From the user's perspective, resources are virtualized so that the specific resources used to run a program or to store data are transparent, but they can be brokered by a grid utility, which also can handle accounting and chargeback. In addition, the physical structure of the grid is often defined by the organizational structure, so resources can be added or removed from the grid at any time. Such changes can be regularly scheduled events, the results of long-term planning or contracts, or virtually random occurrences.

IDC has already seen early adoption of grid HPC in many industries, including economic and financial modeling, bioscience, chemical engineering, CAD, CAE, digital content creation and distribution, electronic design and analysis, geosciences and engineering, general research, national defense, weather forecasting, and academic pursuits. However, the environments created by provisioning grids are still very new operational concepts. The functional goal of the environment is to shield users from the complexity of the architecture by providing them with convenient abstractions of the underlying functional, interfacial, software, and hardware resources.

Architecturally, it is useful to think of provisioning grids as consisting of larger versions of traditional three-tier application architectures — a user-interface layer, a middleware layer, and a resource layer. The user-interface layer typically consists of desktop applications or Web-based portals. The resource layer typically consists of HPC servers, storage, applications, and data content. One of the most critical components of grid HPC is the middleware because it provides a layer of abstraction between the user interface and the computational, data, or application resources, thus decoupling the workflow from any dependencies upon a specific interface or resource.

Technical grid-based computing strategies use system software, middleware, and networking technologies to combine independent computers into logically unified systems. The major distinguishing feature of grids is that they are configured from components that are at least nominally owned and/or managed by independent individuals or organizations. Advantages of this approach are increased utilization of computing resources, access to specialized computer systems, cost sharing, and improved management.

Grid computing does pose its own set of challenges, however:

- Perhaps the most challenging issue for HPC grid adoption resides in data security. Many HPC users consider their HPC applications important IP of their organizations, and therefore they are reluctant to send those jobs to the external resources. Because of such concerns, most of the configurations in HPC grids today take the form of "private clouds" — a variation of public grids. Those private clouds refer to the HPC resources within a company's firewall, thus making all resources in the cloud compliant with the corporate security policy.
- Another major concern in HPC grid adoption is the capability of the resources in the grids. Many HPC applications have specific requirements on the hardware such as systems with large memories, faster processors, and fast interconnect. Many existing computing resources will not have the capability required by the HPC workloads. Therefore, HPC users often would like to acquire a dedicated HPC system and have full control over it.
- Lastly, in a grid that truly takes advantage of an organization's unused processing power, there is the issue of individual node owners turning off their workstations. As a result, the maximum and minimum performance range of an individual grid would range widely, making it difficult to control service-level agreements (SLAs) based on specific performance targets. Datacenter managers who are being asked for 99% uptime, as well as grid providers, would have to build significant overcapacity to meet these targets.

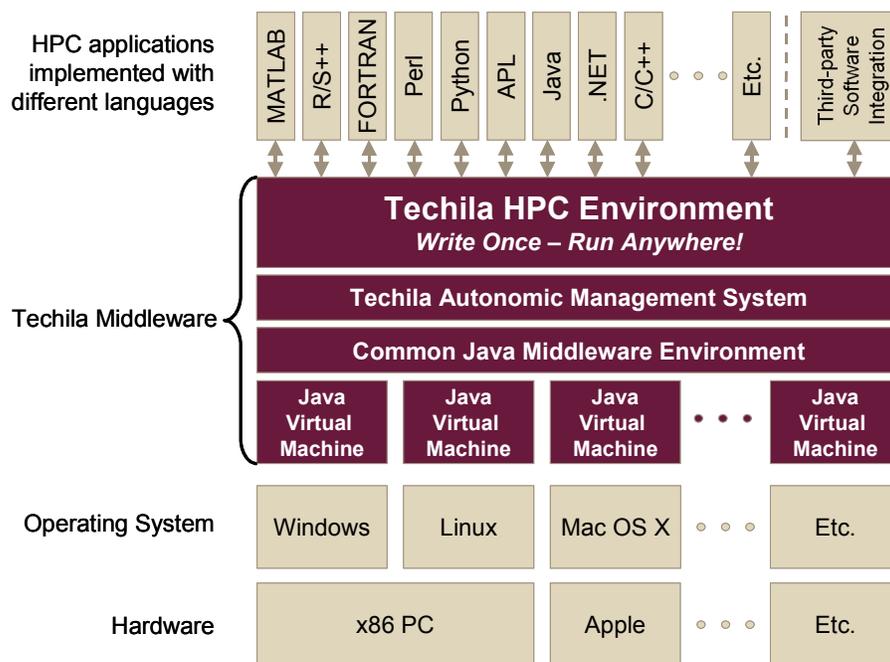
Considering Techila Technologies

Techila Technologies, founded in 2006, is a privately held provider of HPC solutions based in Tampere, Finland. The key solution Techila offers is a middleware stack that distributes HPC applications to an organization's existing computing resources and manages workloads to achieve results in an optimal time. The software stack helps create a virtual HPC environment within an organization's existing IT infrastructure, which enables users with computationally demanding workloads to access the collective resources and get to the results faster. Applications benefiting from Techila's solution include modeling, simulation, optimization, and data analyses in a broad range of industries such as finance, bioscience, universities, and academia.

Techila's flagship product is Techila Grid, which enables organizations to create HPC infrastructure without major capital investments and operating expenses required by new HPC hardware. Figure 2 illustrates the architecture of the Techila Grid.

Figure 2

The Techila High-Performance Computing (HPC) Architecture



Source: Techila Technologies, 2010

The solution consists of two components: Linux-based server software and middleware installed on computers that have idle capacity. The Linux-based server software component acts as a "gateway" to the Techila computing environment. It carries on responsibilities including distributing the computing tasks to the network of Worker computers, managing available resources, and optimizing the execution of tasks.

Currently, the supported hardware platform that runs the Techila Server software is based on IBM System x technology. The middleware runs on the lowest possible priority on Worker computers, resulting in no impact on local application performance. Workers are benchmarked, and Server assigns work to the Workers based on continuously monitored processing availability. Workers can

be Windows, Macintosh, Linux, or Unix based, enabling Techila Grid to create a heterogeneous virtual supercomputing environment.

The solution also offers scalability. If there is any change to the workloads in the grid, Techila Grid will readjust available hardware resources to provide the computing needs to the existing workloads. The software supports applications written in various environments, including MATLAB, FORTRAN, Java, Perl, C/C++, APL, R, .NET, and Python. Techila Grid also provides application programming interfaces (APIs) for third-party software integration.

Techila Grid offers a comprehensive set of security features to protect the security of distributed processes, as well as the computers in the grid. Security measures include certificate-based authentication, SSL-encrypted connections and traceability of jobs, Worker execution policies, and user group rights management.

IDC sees the following benefits in the Techila Grid offering:

- **Savings on new hardware spending.** Techila Grid allows companies to leverage their existing hardware capability and offer users computing resources on demand, resulting in higher resource utilization and operation efficiency.
- **Savings on power, cooling, and floor space.** Without new hardware acquisition, there is no additional power and cooling consumption from the computing resources and the need for expanded facility space is also eliminated.
- **Ease of system management.** Techila Grid is architected based on the IBM Autonomic Computing Manifesto to manage the entire distributed environment; it reduces the complexity of managing the actual HPC system by offering self-configuration, self-healing, self-protection, and self-optimization capabilities for the virtual HPC system.

Use Cases

At Tampere University of Technology in Finland, scientists were able to cut the execution time of a full-scale simulation on brain research from seven years running on a single high-end PC to six days running on a Techila Grid-enabled virtual computing environment — without sacrificing accuracy.

Many other applications have been tested and have benefited from Techila Grid, including material research for developing superconductors, risk analysis, exotic options pricing, and medical imaging research. Early results have shown a positive outcome from the technology adoption, and IDC expects to see more applications tested in the future. As one user from a test site reported: "We have been setting tough requirements for the Techila solution. It has been a pleasure to contribute to developing the usability of the Techila solution...[which] has completely changed the rhythm of research work...."

In 2009, the company won the prestigious InnoFinland prize for its innovative, environmentally friendly approach to HPC.

Challenges

Techila does face market challenges, however. First, as an emerging company in a high-profile market, Techila must work hard to establish its technology as a key alternative to the conventional approach to HPC resources. The company faces stiff competition from companies offering similar solutions. Cloud/grid computing is overheated today, with almost all major OEMs in HPC offering some kind of cloud or on-demand computing service. In addition, middleware vendors have also jumped on the bandwagon, offering various flavors of cloud/grid management software. The battle is on, and Techila will have to fight strategically in order to win in the market.

From a technology standpoint, two critical concerns in the grid arena are security and data transport. The company must convincingly demonstrate the effectiveness of its security benefits and the efficiency of how Techila Grid moves distributed data through the grid without sacrificing capacity. Economically, Techila will need to develop an attractive pricing model to further differentiate its offering.

Conclusion and Essential Guidance

As the economy begins to slowly turn, enterprises will once again look for innovation as a source for delivering flexible and adaptable IT solutions, particularly in the HPC arena. HPC is proving itself to be more than just a way to design or model technical and scientific solutions; many organizations now consider it indispensable for their corporate survival and competitiveness.

As a result, organizations are looking for the best of both worlds: economically and environmentally friendly approaches to access HPC resources. One such new avenue is to take advantage of the existing processing power already residing within a company. For organizations looking at grid HPC, IDC suggests the following:

- Identify the organization's requirements for HPC, including applications and desired results
- Evaluate various HPC service offerings out on the market, including public/private cloud, on-demand computing, etc., and pick the delivery system that best fits the application requirements as well as the organization's processing, overhead, and environmental needs
- Understand the organization's tolerance levels for performance (SLAs and QoS) because grid-based HPC may cause capacity fluctuations based on client idle time
- Be crystal clear about security and throughput requirements when selecting an HPC provider, and demand customer references

As organizations look to leading-edge HPC solutions, they won't forget the lessons learned during the past tumultuous years. Efficiency is still the number 1 requirement of corporate finance, and effective resource utilization is the mantra of IT executives. Added to this is the growing need for green computing in which organizations strive to reduce their overall energy consumption. To the extent that Techila Technologies can overcome the challenges described in this paper, the company has a significant opportunity for success in this important market.

ABOUT THIS PUBLICATION

This publication was produced by IDC Go-to-Market Services. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Go-to-Market Services makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

COPYRIGHT AND RESTRICTIONS

Any IDC information or reference to IDC that is to be used in advertising, press releases, or promotional materials requires prior written approval from IDC. For permission requests, contact the GMS information line at 508-988-7610 or gms@idc.com. Translation and/or localization of this document requires an additional license from IDC.

For more information on IDC, visit www.idc.com. For more information on IDC GMS, visit www.idc.com/gms.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com