



Richard Walsh
Research Director, High-Performance Computing

HPC Directions In Financial Services

July 2008

The global market for high-performance computing (HPC) servers has experienced strong growth each year since 2003. In 2007, sales of HPC servers reached US\$11.6 billion. IDC estimates that worldwide sales of HPC systems in the financial services sector made up about US\$357 million of the 2007 total and will grow rapidly to US\$630 million in 2012. Revenue growth on the computational side has been driven almost entirely by clustered servers that provide new levels of price/performance. Clusters recently crossed the 65% market share line to become the dominant species of HPC servers. At the same time, there are a number of customer pain points in the financial services sector, including power, cooling, and space constraints; the need to converge front-office and back-office operations; and the increasing need to make data/information available with minimal delay on a global basis.

The following questions were posed by representatives of Cisco, Hewlett-Packard, Intel, and Microsoft to Richard Walsh, Research Director in IDC's High-Performance Computing practice, on behalf of the companies' HPC customers.

Q. What is HPC used for in the financial services market?

A. The financial services industry has fully integrated HPC into the operational machinery of its investment banking segment over the last 20 years. Today HPC is an indispensable resource helping this segment to price derivatives and manage its portfolio risk around the clock and around the globe. Moreover, it is moving rapidly down market into hedge funds and smaller investment institutions, as well as into neighboring disciplines charged with modeling business and economic risk, such as insurance and research and development planning. As a market, financial services continues to be a quiet but very aggressive adopter of each new wave of HPC technology, from parallel processing to cluster computing, to computational grids, to the use of processor acceleration technology. HPC resources are the tool that has given life and scope to the derivation (by Black, Scholes, and Merton) of the differential equation for pricing an option based on the value of its underlying security, its volatility, and the risk-free rate.

HPC enhances market liquidity by allowing for the risk-adjusted pricing and sale of derivative investments, and by supporting a trading forum that rapidly and globally manages and distributes market risk. Pricing derivatives is an exercise in forecasting the financial future and using that forecast to rationally value the derivative. The computational task of producing the forecast and pricing the derivative grows with the number of conditions and variables, the amount of time that must be considered, and the size of the time step. Once priced, portfolios of instruments held by the investment bank must be monitored as conditions change to determine how the risk contained in the portfolio has changed. Adjustments must often be made to limit the overall risk. This very time-sensitive process of pricing the derivative and measuring the derivative portfolio risk is precisely where HPC resources are required. It reflects both business and regulatory requirements. A financial institution's HPC capability largely governs the scope and depth

of the financial markets that it can operate in. HPC's role in trading has grown to the point where some trading operations run automatically based on targets determined through simulation alone and without transaction-by-transaction oversight from traders. As the mathematical modeling of risk in other financial services markets such as the insurance industry and business investment grows, IDC expects HPC to become an integral part of those markets' operations as well.

Q. The financial services market is sometimes referred to as an HPC proximity market. Why?

- A. Today, this is perhaps no longer a fair statement as HPC has expanded well beyond its traditional large scale, floating-point-intensive scientific and engineering workloads. Still, HPC, as it functions in the financial services industry, has a number of unique features that set it apart and deserve mention. HPC in financial services grew organically from front-office operations and the trading desk to initially include modest back-office application servers that have now become enormous HPC clusters integrated into grids at larger institutions. Unlike most HPC disciplines which are R&D oriented, HPC in financial services is a near-real-time operational activity that is taking in market data feeds constantly, integrating them into a very large number of jobs, and returning results to an impatient marketplace that can never get them too soon. In addition, the applications themselves change far more frequently than those in traditional HPC disciplines in response to new product definitions and areas of profitability. As a near-real-time operation that tracks global trading activity 24x7 around the globe, HPC in financial services also demands substantially higher standards of redundancy and reliability than are required in more traditional HPC disciplines. Space, power, cooling, and technical talent are generally much more expensive than in traditional HPC markets, so the pressure to manage total cost of ownership (TCO) is also much greater.

While the HPC done in financial services is very advanced, this is not well known because its practitioners have not wished to publicize their state-of-the-art usage for fear of conceding competitive advantage. Investment banks with large HPC operations do not publish Linpack performance results on the semi-annual TOP500 list of the world's most powerful supercomputers, but if they did, financial services companies would have entries among the top 100 systems in the world – despite the fact that applications in computational finance are not properly represented by Linpack performance. It has some very distinct numerical features, relying heavily on dependency trees and stochastic methods, including the heavy use of conditionals, and using intrinsic functions and random number generators more extensively than much of traditional HPC.

Another unique feature of HPC in financial services is that it is embedded in and coupled to a more traditional enterprise-oriented front-office infrastructure that is dominated by Microsoft-Intel and Sun desktop systems. Compared to traditional HPC environments, this front-to-back office integration requirement adds complexity to an already pressured and dynamic environment, but has produced interesting cross-pollination and innovation across the enterprise-technical computing boundary. The optimal management of front-office workflows is as important as the optimal use of the best available back-office resources. At this boundary, standard desktop tools such as Microsoft's Excel are being connected to back-office resources often running HPC's dominant operating system, Linux, but also (more often than is typical in HPC in some other market segments) Windows and Solaris. The financial services industry strongly supports much of the technology transfer between the distinct worlds of enterprise and high-performance computing.

Q. What are the issues in delivering "back office" computational power to trading desks today? What is the outlook for the future?

- A. The adoption of clusters as back-office workhorses in financial services added a third tier to a two-tiered, trading desk and applications server model. The further separation

between the front- and back-office environments was stimulated by the predominance of the Linux operating system on clusters (instead of Solaris or Windows), the unique skills required to create distributed parallel applications and run these clusters, and the lack of connection between standard desktop interfaces and the back-end Linux clusters. The current trend is toward consolidation and integration on several fronts. Windows and Solaris are still strong on the desktop in financial services and have found a place on more back-office clusters. Microsoft's cluster OS offering has matured (latest version: Windows HPC Server 2008), and Microsoft is strongly supporting the movement of HPC workflows between Windows application front-ends such as Excel and back-end clusters. The use of distributed desktop systems as *ad hoc* loosely coupled clusters has added momentum to this reintegration trend. Linux is still very strong in the back office, but is now, while better integrated, facing competition from Windows in buildouts from the desktop.

The success of distributed desktop clustering within asset classes and at particular trading desks has increased the financial industry's comfort level with and understanding of the value of pooling resources. This and power, cooling, space, and support pressures have led investment banks to seek savings and increased utilization through server virtualization and grid computing architectures with several clusters placed in different geographic locations, bound together by a secure low latency fabric, and managed by grid-ready middleware. Consolidating back-end resources that serve different asset classes in several locations outside of high rent districts saves money and minimizes the risk of resource blackouts. However, virtualized, distributed grid architectures create new challenges of resource allocation, new system stand-up, dynamic provisioning, and end-to-end latency management. Looking briefly at latency management, by removing the slowest step in the trade (the human being) from the trading model, automatic trading has made profitability dependent on transactional latency which must now be consistently in the millisecond range. When operational economies combined this requirement with a tiered, distributed backend computational grid, correct job placement and network latency management can determine profitability.

In this context, the role of the network as the integrating component is elevated and the potential value in virtualizing all of the basic components it binds to together – the compute server, storage server, and network capacity – grows. The completely virtualized resource implied is a utility computing model, and the future of HPC in the financial services segment will tend in that direction.

Q. What kind of HPC systems are likely to be deployed in this market in the next three to four years? How well will the numerical methods used in financial services today map onto these systems?

A. The system architectural requirements of financial services applications are reasonably well matched to current design trends in the HPC industry, which has moved rapidly to cluster systems (with blades becoming more popular), multicore processors, low latency memory subsystems, faster interconnects, and acceleration technology. Financial applications are less constrained than more traditional HPC codes by the per-core bandwidth limitations of multicore processors. Branch prediction, intrinsic function performance, and instruction issue rates are generally more important. Clusters of nodes, whether more tightly coupled or distributed, can service the substantial bulk-parallel workloads that make up an important part of the financial services job mix. The arrival of multicore processors has made nodes "fatter" and allowed them to run more parallel jobs within individual nodes of clusters. This approach to parallelism is often preferred because failures on a single node require only that the job be terminated and rescheduled elsewhere. Cleanup after failure in a distributed parallel processing environment using MPI is far messier. Similarly, financial services applications performance within the node has been improved by accelerators such as general-purpose graphical processing units (GPGPUs) and field programmable gate arrays (FPGAs). These applications will also be candidates for acceleration using the on-chip

technologies expected from Intel and AMD in 2009. HPC practitioners in the financial industry have been early adopters of acceleration technology, and IDC expects this pattern to continue.

Taking a look at systems as a whole, the trend toward more densely packed clusters built from blades and having greater node counts will continue. Space, power, and cooling are often at a great premium in the office towers where computational finance is done. The modularity, ease of replacement, and power and space efficiency of blades suits this market well. Vendors have taken note and had much success selling blades into this market. Larger firms have trading operations in multiple locations around the globe and have been binding their distributed cluster resources into computational grids for some time. This approach offers TCO benefits across trading desks and asset classes where time-dependent variation in these workloads makes resource-sharing work. At the workgroup end of the HPC market, the small footprint, easy-to-use cluster server products that have recently been introduced by HP and others should attract buyers in financial services. Their highly engineered and integrated enclosures make their standards-based cluster components substantially easier to manage.

In the globally scaled-out financial industry, where profit is dependent on the timeliness and accuracy of a buyers' bid, reducing the end-to-end latency of feeds streaming out of back-end HPC resources is vital. So, low latency interconnects are not only required to support message passing interface (MPI) parallel applications sending messages between fat nodes, they are also needed to minimize the travel time to the desktop from potentially any node in a globally distributed grid. InfiniBand (IB) interconnects currently offer a compelling combination of high bandwidth (20Gbps today, 40Gbps by year's end), low latency (~1 μ s point-to-point), high message rates (greater than 20 million per adapter), and multipurpose functionality compared to alternatives. InfiniBand offers the prospect of integrating data delivered from storage devices, clustered servers, external feeds, and via inline data modification engines. In addition, new long-haul fiber and IB interconnect products offer the prospect of more tightly coupled grids. Networks that can respond to bandwidth and message rate spikes also add value in the financial services space. IDC has forecast a CAGR of over 35% for IB-related revenue through 2011. Its adoption in the financial services segment will contribute significantly to this. Still, 1 and 10 Gigabit Ethernet (GE) is adequate for many applications in computational finance and will continue to be used. The mixed IB and GE fabrics implied will need to be seamlessly integrated.

IDC expects that someone in the financial services industry will silently construct a single or grid-like HPC cluster system that can perform a Linpack petaflop (10^{15} calculations per second) in the next 24 months. IDC also notes, however, that the financial industry's vision of utility computing may blur the distinction between a petascale *system* and petascale *capability*. The end-to-end requirements of HPC resources in financial services have made this segment a leader in the push for utility computing, a vision in which application and security standards allow financial workloads to be bundled up like so many Internet packets and directed to the most available and cheapest computational resource, perhaps repurposed from the bare metal up with the required software stack. The push toward integration is already under way as middleware vendors seek more complete and modular software offerings between the operating systems and the application, and network equipment vendors rethink the value they can add to the utility picture. Whatever the time table and extent of HPC system integration, the financial services quants are eager to test the natural parallel scalability of Monte Carlo methods on future petascale, multicore processor pools.

Q. How important are industry-standard technologies/solutions and easy-to-use GUIs?

A. The use of standards-based server and storage hardware resources augmented with reliability and TCO management features will continue to dominate in the financial

services market. At larger firms with the largest HPC resources, the shift to automatic trading will change the emphasis and requirements somewhat. Likely hardware additions that have yet to become standardized include the use of accelerators on the processor side and SAIDs (Sealed Arrays of Identical Disks) on the storage side. The trading desk has been central to the adoption of HPC in financial services, and it will continue to be the cockpit for engaging the market and computational resources in play behind the scenes. The trading desk's preference for the Windows operating system and Excel spreadsheets will likely lead to the proliferation of these tools on back-end servers. Still, back-end resources tied to particular asset classes will continue to merge and serve multiple front-end sub-markets. To some extent, this consolidation has had a disruptive effect on the old model of asset/application-specific servers connected to GUIs designed for that sub-market. IDC expects that as the back ends converge, front-end GUIs managing workflows will again stabilize. Smaller firms will need support for coupling their desktops to the HPC cluster resources they will add to their back offices. This market's operational orientation and party-counter-party standardization requirements will drive it in the direction of utility computing, in which application-specific workloads are packaged and distributed to run at the most available and lowest-cost resource.

ABOUT THIS ANALYST

Richard Walsh, Research Director in IDC's High-Performance and Technical Computing Group, is charged with providing technical depth and direction to IDC's hardware and software research in the HPC market space. Mr. Walsh has more than 20 years of study and experience with HPC hardware and software component technologies, applications support and performance optimization, parallel programming, computational research, teaching, HPC system acquisition, installation, and acceptance. He has direct HPC experience in the areas of computational chemistry and finance, custom vector and MPP systems, and commodity cluster specification, acquisition, installation, and support.

Those interested in more information can contact the following company representatives:

Simon Lim
silim@cisco.com
65-6317-5588
www.cisco.com

Dennis Ang
dennis.ang@hp.com
65-6727-6074
www.hp.com

Dave Chee
dave.chee@intel.com
65-6213-1000
www.intel.com

Deepak Setty
hpcinfo@microsoft.com
www.microsoft.com/hpc

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 65-6829-7757 or gmsap@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.

IDC Asia/Pacific, 80 Anson Road, #38-00, Singapore 079970 P.65.6226.0330 F.65.6220.6116 www.idc.com.

Copyright 2008 IDC. Reproduction is forbidden unless authorized. All rights reserved.