Clipper Notes™
VENDOR-AGNOSTIC EXPLANATIONS AND ADVICE
FOR THE INFORMATION TECHNOLOGY BUYER

Navigating Information Technology Horizons℠

New in 2007 Report #TCG2007063 Updated May 16, 2007

Computational Grids — Server Consolidation for a Distributed, On-Demand World

Analyst: Anne MacFarland

Management Summary

The concept of a processing grid is the sound of thunder on the horizon. Those in the data center wonder: Is it a sortie by an army bent on yet more data center destruction? Is it a raid to plagiarize data center concepts for use in the distributed, converged world of data and communication networks out there beyond the firewall? Alternatively, is it a construction crew bringing an opportunity to export data center capacities and get new sources of revenue? The answer, of course, is yes to all three.

The data center – as the optimal environment for enterprise computing – has been under siege for some time by a variety of forces. Most recently, the need for real-time data over vast geographies and the need for rich media feeds for design graphics, and video for content dissemination, training, and the entertainment hooks of e-commerce have not only perforated the data center perimeter with portals and other data conduits but also made necessary the remote "edge" of repeating servers in other locations. You know you still control these remote facilities, but they likely will never come home again.

Moreover, now they want to steal your secret sauce. As all those tricks of memory and cache management were adopted by the network-attached storage (NAS) folks, so now the processing grid folks want to use data center concepts of multiprocessing and load balancing across a larger environment. They even want to appropriate the crown jewel of server consolidation that you thought was going to refresh the value of the data center to the enterprise. Will the grid have the performance characteristics comparable to those of the data center? Probably not – but there are important characteristics besides speed, like cost, particularly that of capital expenditures. There are more trucks than race cars in production and the carrying ability of trucks is more generally useful than the speed attainable by race car.

The good news is that the processing grid also provides an aftermarket for the performance and other capabilities that only a data center can provide. This gives functionality to the enterprise data center beyond that of a fully-owned, internal resource, enhancing its value.

Many have thought that with all the underlying protocols and negotiation capabilities needed to make data-center type processing possible in a larger environment, grid implementations were still not *mainstream* enough for everyday use. **But for those, particularly those in trusted environments, who have an urgent need for capacity beyond their means, for those wishing to use applications with ungainly processing needs and for those with a distributed but controlled environment they wish to optimize, grid processing's horizon may be quite close.** All enterprises will want to pay attention, for processing grids offer a significant computing option. For more details, read on.

IN THIS ISSUE

The Situation

Complex

Complex IT environments take on a power of their own. You don't want to mess with them if they're working right – but at the same time you *do* want to mess with them, because enterprise information systems are works in progress, and flexibility is the name of the game. **Much of the IT complexity – of tiered environments, of localized intelligence and point products, is part and parcel of their increasingly broad support for complex business processes once managed by humans**. You cannot simplify your way out of this problem.

Modular

Responding to the need for adaptability and flexibility, IT has gone modular. Many software products come ready to interface with large applications or management frameworks. Hardware products bristle with attributes like hot-plug modularity and auto-configuration, with more and more self-management in the offing. Aggregation and automation software products, fed to administrators and other stakeholders via dashboards, facilitate monitoring and management of all these modular devices. Application-centric products that can manage business processes across all the modules are proliferating.

Service oriented architectures leverage all these hardware developments. While complex to implement, they are popular because the economic environment for developers and the procurement practices of enterprise customers both favor the creation of many fine-grained product options. Few vendors or open standards groups want to take on the risk of trying to build a vast framework of capability single-handedly. Open standards and component models ensure more kinds of interoperability for hardware and software, and make modular development attractive. Web Services, developed for e-commerce, allow more any-to-any functionality through a service negotiation paradigm, aided by languages (XML, WSDL), directories (UDDI and others), and protocols (SOAP, *et al*).

And Geographically Distributed

Not only are enterprise IT systems complex, but they are distributed – by history and happenstance, and, recently, by the trend of corporate expansion through acquisition. The centrifugal forces keeping the assets, both human and IT, distributed are organizational, economic, political[1], and very real. E-mail, instant messaging, and extranets have lessened the need to relocate an often-reluctant workforce. In turn, they often want to keep their IT resources local for reasons of habit, latency and the persistently high cost of bulk data transfers.

The Drive for Server Consolidation

In the current spate of doing more with less, enterprises have found that the processing capacity of servers, particularly the scattered, small, limited-application servers, is underutilized. Processing capacity, unlike storage capacity, is a function of time, and the idle cycles of a server, like water from a dripping faucet, represents an unrecoverable asset. Co-location, where it is possible, is only a partial solution. It provides a single point of physical management, but does not, by itself, allow a single aggregated functionality to load-balance the processing of multiple applications. Consolidation onto larger servers with the partitioning ability to run multiple workloads is one better answer, but there are also remedies for the distributed environment.

Clustering across a dedicated network is an alternative. Failover between clustered servers, though not simple, provides a full transfer of functionality between one server and another. Today's supercomputers are not monoliths, but large clusters of relatively prosaic 2- and 4-processor servers. Physically distributed processing has been around for a while in UNIX remote procedure calls and the mainframe's Geographically Dispersed Parallel Sysplex (GDPS). Years ago, limited bandwidth and processing speed limited the usefulness of distributed processing. Now, where bandwidth is plentiful, distributed hardware structures like Grid can be leveraged for greater system resilience.

The processing assets that are distributed may not be as "fashionable" as the "carrier-class" data centers, but their aggregated capacity is nothing to sneeze at, particularly if one can use this compute capacity across geographies, platforms, and even organizations. This is the promise of an emerging computing initiative, the processing grid. **The processing aggre-**

[1] Dispersed assets make the IT environment harder to attack physically and increase resilience to disaster.

gation of the grid is a distributed kind of server consolidation, not competing with centralization, but truly complementary to it.

The Grid for Distributed Consolidation

A grid is an inherently robust architecture. An IT grid can be as simple as a server farm, SAN, or content distribution network, where an aggregation of assets is available to multiple applications. In Web server farms, a request is assigned to the least busy server. In processing grids, a workload is split and distributed to multiple servers or PCs, processed and then re-aggregated. **The participating computers are generally wholly dedicated to the grid for the duration of the grid workload operation[2], but not necessarily co-located or even owned by a single organization.**

The first large implementations of computing grids have been to aggregate computing capability beyond the budget of any one organization for massive parallel processing of scientific and technical workloads in academic and governmental environments. Pharmaceutical and life science organizations have implemented grids for protein folding visualization. Engineering design is another area with sporadic, compute-intensive tasks ripe for grid deployment. Other grid implementations include use of idle PC cycles for benevolent purposes (cancer research, the search for intelligent life, etc.), and for broadcast of static content. These are all huge applications but limited, inherently parallelizable workloads in environments that do not have the full-blown commercial needs for precipitated security, tightly controlled asset allocation, and charge-back.

Building a commercial processing grid involves combining, in a multiplexed (non-batch-driven) environment, elements of distributed processing, failover, and the aggregating and integrating elements developed by organizations like *Globus.org*,[3] including security, the quality of service measurability needed for charge-back, and the ability to invoke processing as a limited-time service engagement.

The grid initiative represented by Globus intends to support servers from different vendors, because that is what most large enterprises have in their IT environments. Most significantly, this grid goes beyond the bounded, pre-enumerated clusters of server farms and supercomputers, to allow enterprises to extend their IT capabilities on a short term basis – and to develop an aftermarket for their own IT capabilities. By making processing a vendible product, this robust co-processing grid will give a hedging capability to enterprises against the cost of over-provisioning and the risks of under-provisioning that are inherent in any business (or IT) model.

Extending the Spectrum of Computing

All computing can be characterized as an aggregation of scheduled processes and subroutines, differing in degrees of physical separation and connectivity. Performance varies considerably, and latency over long distances will be an issue until we can get around the speed of light. Traditionally, closer has been important and the enterprise data center the ideal performance engine. Outsourcing to someone else's Internet data center kept that performance, while adding only access latency, but outsourcing has other risks of security, reliability (quality of service), and corporate persistence. **Today, for many enterprise workloads, the issues of scope, risk and cost may be as important as performance.** In the scope-risk-cost spectrum of how computing is done, the processing grid sits at the far end from the enterprise data center, with outsourced computing somewhere between.

Scope

The enterprise data center, with its extensions, has an organizational scope, while outsourcing an application-centric scope, though, of course, these overlap considerably. **The grid negotiates computing in much smaller increments by workload, or fraction thereof. A grid is something that enterprises will use for applications or data centers, as a starting point.** ,Grid processing, in the form of *Software as a Service (SAAS)*, has already led to more granular pricing schemes and to lower software licensing fees for applications not often used.

Cost and Risk

The enterprise data center is a wholly-

[2] With partitions and virtual machines, multiprocessing can be supported. Habits of management and lack of administrative trust in autonomics and business policy-based automation can limit its adoption.

[3] Globus is a group that comprises an international assortment of hardware and software vendors (including the big players), academic institutions, governments, and others.

owned (or leased), expenditure – an intensive, long-term commitment, whose value to the organization depends on the quality and timing of the IT strategy underlying its evolution. External sources of risk are minimized, but internal risks remain. Outsourcing offloads some of the financial risks, regularizes costs, and indemnifies tenants against some bad technology decisions, while limiting flexibility and independence and adding some new risks. **The grid approach can give capability without capital expenditures, and considerable dynamic flexibility – but it involves amplified security and data integrity measures, and a loss of predictable availability of those capabilities (depending on the service contract).**

To cope with rapid changes in the marketplace, enterprises need the full spectrum of computing options. Committing to a single strategy, even the old standard of asset ownership, is in itself a risk. Technology improvements (device self-management and business - process - based security and management) will change the parameters of computing as well. As the value of a business to its customers, and the role of partnerships in that value, continues to change, flexibility is crucial. **"Make or buy" must be joined by "rent" and even by "borrow" in the decision-making spectrum.**

Complicating Factors

The Need for many-to-many integration

A growing need for integration between applications is both driving and complicating the development of computing grids. The need for a standard method of many-to-many integration between the hardware *and* software components of IT solutions, urgent in many enterprise environments, is crucial, both for the development of a commercially viable processing grid and for supporting the blizzard of portals, applet, and mashups that serve information to different end-user roles within an enterprise.. The time-and labor-intensive practices of *enterprise application integration (EAI)* and standard *Application Protocol Interfaces (APIs)* stifle the adaptability of the processing grid, something that is its most attractive feature. EAI and APIs enable a guarantee of interoperability between a product and each of its supported platforms, but they are still one-to-one integration processes. To get to the universality of a many-to-many interface, people are turning to more modular, deterministic forms of programming. This modular approach allows software features to be deployed to match the needs of the users, not just as a very expensive *more* or a disappointingly inadequate *less*. With the commonality of XML and its variants, new capabilities can be added, in all-modular environments, without disturbing existing software dependencies. This new discipline in software development, embodied in SOA, makes grid architecture more and more appropriate for enterprises and service providers of all sizes.

An education analogy might be to the apprentice system (EAI), providing the most complete knowledge transfer but a limited scope, the trade school (APIs), which educates many to a narrowly-defined capability, and the university (XML), which seeks to prepare a diverse set of students for a diverse and changing range of employment opportunities. The value, scope, and efficacy of university education have been debated for centuries, and there are no absolute answers. The same is probably true for application and hardware integration.

Enduring issues of Security and Control

In IT as elsewhere, security has always been a matter of foundations (insured persistence), walls (isolation), gateways (authentication, etc.), and, importantly, the ability to track an incursion to learn how it works and, if possible, its source. A global file system and strong authentication of all grid participants and workloads is essential. Isolation on the host server side can be enforced by extensions to partitioning capabilities. Isolation of the migrant workload can be enabled by the envelopes of virtual machines and other similar products. Maintaining the boundary between the two requires integration of host-based and workload-based security features. Comprehensive approaches to security now support the control features needed to assuage the qualms of users of an open, non-membership grid will be a challenge.

The control issues of grid computing have been addressed by operating system-based approaches and by the service paradigm of SOA. You need both the capability of allowing dynamic participation and a common method of advertising, negotiating and evoking the exchange of processing services to coordinate a distributed grid. For membership grids, where all the participants are known or knowable, a larger network or fabric version of an operating system will fulfill the needs. The more radical long-term vision is of an unbounded heterogeneous grid, enabled by grid protocols and Web services and arbitrated by market forces. The grid may be more threatening to the data center in the price pressure it exerts than in its conceptual "otherness."

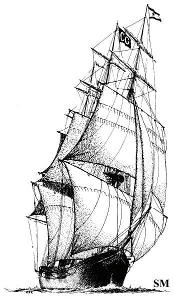
The Perils and Opportunities of "On Demand"

The world of "on demand" computing capabilities, and particularly on-demand pricing for these capabilities is here. It is attractive to the consumers of computing because it allows them to utilize capabilities and applications opportunistically, without having to plan. The unpredictability of this opportunism makes provision of "on-demand" a challenge. **Grid computing is one of the less perilous ways to provision capacity to meet that challenge.**

Conclusion

The processing grid is a natural evolution, given the technology assets now available. We have the engines (processor speed), and we have the connectivity (bandwidth). We have expanding measurability at the device level, and we are developing the automated, policy-driven systems to monitor, manage and alert, using all the data gleaned from that device measurement. We also have the tools to aggregate IT capacities and capabilities to deploy them creatively. The cost of N+1 redundancy grows cheaper when N is many. It is only question of what kinds of *bricks* an organization will use, and how much it will leverage grids beyond its borders. **Grids (processing and other) are the way to optimize larger computing environments, and the cost-effectiveness of larger computing environments can drive the cost of computing down.**

It is no longer a matter of *when*. Even if your enterprise is not doing grids itself, it is probably benefiting from them through its outsourcing and software-as-a-service relationships. Consider how your organization can leverage them most wisely.



About The Clipper Group, Inc.

The Clipper Group, Inc., is an independent consulting firm specializing in acquisition decisions and strategic advice regarding complex, enterprise-class information technologies. Our team of industry professionals averages more than 25 years of real-world experience. A team of staff consultants augments our capabilities, with significant experience across a broad spectrum of applications and environments.

- ***The Clipper Group can be reached at 781-235-0085 and found on the web at www.clipper.com.***

About the Author

Anne MacFarland **is Director of Data Strategies and Information Solutions for The Clipper Group.** Ms. MacFarland specializes in strategic business solutions offered by enterprise systems, software, and storage vendors, in trends in enterprise systems and networks, and in explaining these trends and the underlying technologies in simple business terms. She joined The Clipper Group after a long career in library systems, business archives, consulting, research, and freelance writing. Ms. MacFarland earned a Bachelor of Arts degree from Cornell University, where she was a College Scholar, and a Masters of Library Science from Southern Connecticut State University.

- ***Reach Anne MacFarland via e-mail at Anne.MacFarland@clipper.com or at 781-235-0085 Ext. 128. (Please dial "128" when you hear the automated attendant.)***

Regarding Trademarks and Service Marks

The Clipper Group Navigator, **The Clipper Group Explorer**, **The Clipper Group Observer**, **The Clipper Group** *Captain's Log*, **The Clipper Group Voyager**, **Clipper Notes**, and *"clipper.com"* are trademarks of The Clipper Group, Inc., and the clipper ship drawings, *"Navigating Information Technology Horizons"*, and *"teraproductivity"* are service marks of The Clipper Group, Inc. The Clipper Group, Inc., reserves all rights regarding its trademarks and service marks. All other trademarks, etc., belong to their respective owners.

Disclosure

Officers and/or employees of The Clipper Group may own as individuals, directly or indirectly, shares in one or more companies discussed in this bulletin. Company policy prohibits any officer or employee from holding more than one percent of the outstanding shares of any company covered by The Clipper Group. The Clipper Group, Inc., has no such equity holdings.

Regarding the Information in this Issue

The Clipper Group believes the information included in this report to be accurate. Data has been received from a variety of sources, which we believe to be reliable, including manufacturers, distributors, or users of the products discussed herein. The Clipper Group, Inc., cannot be held responsible for any consequential damages resulting from the application of information or opinions contained in this report.