

Silo Busting: Deploying Oracle's 10g Grid Computing on the Egenera BladeFrame

Technical Whitepaper

October 2005

ORACLE®



Silo Busting: Deploying Oracle's 10g Grid Computing on the Egenera BladeFrame

Moving application environments out of silos and into the grid computing will result in better-utilized resources and lower costs

INTRODUCTION

Traditionally, when IT is tasked with providing resources to new and existing applications, an application is treated as a single silo of hardware and software. The resources are sized to handle peak demands, configured for its specific needs and deployed. When the sizing has been done correctly, the hardware spends the majority of the time functioning at a fraction of its capabilities. If done inappropriately service levels agreements are in jeopardy or worse customers go somewhere else to satisfy their needs. It is a great deal of frustration for a business to have spent huge sums of money on large SMP machines that were sized inappropriately and to look around the data center and see other machines running idle. Since there are no rebates on unused CPU cycles it would be best to bust these silos up and consolidate the infrastructure, software, and hardware resources to better utilize computing resources. The term "Grid computing" is most often used to describe an environment that coordinates the use of multiple computers and storage devices acting as a single computer. Grid computing moves away from these silos to a more horizontal approach. This allows businesses to add resources on demand to handle peak loads, resulting in better utilization of resources, service levels, and cost savings.

Grid computing works on a virtual environment that appears as a single server despite being made up of several physical servers. While the objective of grid computing is to increase scalability, availability, and performance, this all needs to be done while keeping complexity down. Using virtualization at every level to abstract the physical to the virtual can alleviate complexity. Virtualizing allows the loss of the physical systems and allows the database or application servers to continue working.

Egenera's BladeFrame provides a framework to create a virtual environment to transform the physical resources into software so they can be pooled and centrally managed for assignment to applications as needed. The BladeFrame houses Processing Blades based on Intel processors that are diskless and have no permanent identity. This architecture allows the processing blades to be dynamically allocated and reallocated to any application. The BladeFrame can even be programmed to automatically repurpose nodes and shift computing power on the fly based on resource utilization and/or workload requirements. Through software,

Processing Blades can be added to or removed from any cluster, and attached to or disconnected from any storage volume. In short, Egenera's hardware and virtualization software were purpose-built for the grid and work with Oracle 10g grid software to create an easy-to-construct and manage grid environment. The rest of this paper explains how Egenera and Oracle work together to turn the grid vision into reality

CREATING THE "GRID READY" FRAMEWORK

It's natural when first thinking about creating a grid environment, to just pool a bunch of existing resources into a dedicated resource grid. In practice this can be very problematic for organizations that would need to move machines around a data center(s) and fight the political battles to allow this to occur. The most successful implementations create a blade farm of servers running the Linux operating system. These blades can then be added and subtracted from applications as required and provide a high density computing power and add a fine level of granularity when applying resources to meet service levels.

However, there is more to creating a "Grid Ready" environment than getting a pile of hardware and connecting them together. In a commodity blade server environment, there are pieces that need to be accumulated and configured to create a highly available grid environment. One way of determining the physical complexity is to see how many external connections are required for creating a highly available node. Physical connections for each node involves disk drives, network interface cards (NICs), network switches, Host Bus Adapters (HBA), and all the associated cables. The more external connections required, the more the complexity cascades. For example, each fully populated switch requires another downstream switch that needs to be configured and managed. One study has shown that in large data centers requiring many reconfigurations, the system administrators can spend up to 25 percent of the time managing the cables. For example, each blade should have redundant connections to the Ethernet network, Storage Area Network (SAN) fabric and High Speed Interconnect (HSI). Add in an out-of-band management network interface and that brings the total number of connections to seven for each blade. A small cluster of six computers would need 42 connections.

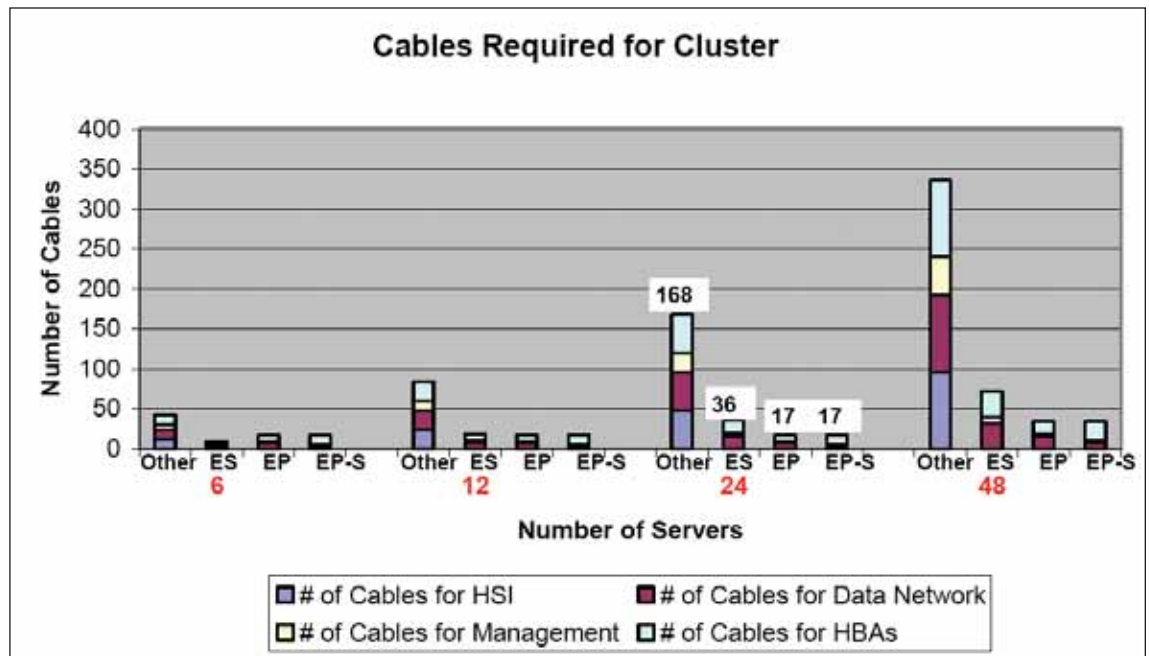
Egenera makes this easier by abstracting the physical resources and putting most of the elements needed to create a Grid environment in one box or BladeFrame. The Egenera BladeFrame provides the individual blades (pBlades) a high speed low latency interconnect via the backplane, access to the SAN and NAS resources, and access to high speed Ethernet networks. All the hardware (processors, memory, storage, network connections,...) that is contained within the BladeFrame allows the virtual assembly of servers and clusters called a Processing Area Network (PAN). Leveraging the virtualization and the high-speed backplane, a grid ready environment requires only a fraction of the number of physical complexities as one that would be created from another blade environment. To provide the same grid

ready environment on the Egenera BladeFrame, the number of connections for a six-node cluster would be anywhere from four to seventeen depending on the model and projected amount of resources being used.

Reducing physical connections and virtualizing these resources to the pServers make a "lights out" data center a reality

The BladeFrame system is configured and managed through the PAN Manager software, which provides a single control point for allocating and monitoring both physical and logical resources. For example, using PAN Manager software to associate individual Processing Blades with IP network and storage capacity creates pServers. Also, multiple Logical PANs (LPAN) can be created on a single BladeFrame to allocate physically distinct, secure resources to enterprise divisions or individual customers. Pan Manager is also used to configure the high availability features built into the architecture.

This chart shows the number of connections used to create a cluster of servers ranging from six to 48 nodes. This chart does not take into account the cables and configurations required for downstream network devices. As more commodity servers are required, the number of connections needed by those servers is exploding. However, those required by the Egenera BladeFrames are only marginally growing whenever the number of servers required expands beyond the size of the BladeFrame.



Egenera has two different types of BladeFrames; the six-node (ES) and the 24-node; the same diskless processing blades are used in either model. Since each pBlade has only processors and memory and use no direct connections, they are both hot-swappable and hot-pluggable. The diskless nature of the pBlades also lets them boot up in whatever kernel desired. In the time it takes to reboot a pBlade, the configuration could change from a 32-bit application server to a 64-bit database

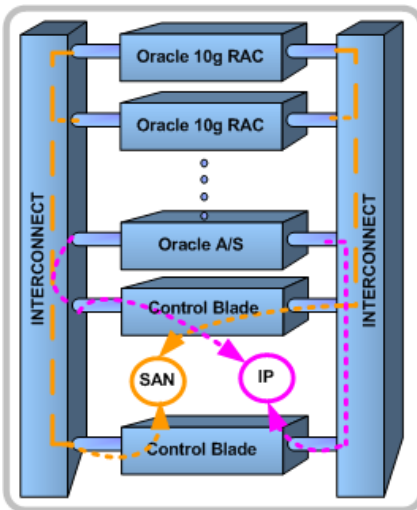
node. The flexibility of the Intel® Xeon™ processor further adds the flexibility to run both 64-bit and 32-bit application code simultaneously under the same operating system. To transform pBlades into useful servers (pServers) each pBlade will be configured with external Ethernet and storage using the BladeFrame's two "Control Blades" (cBlade). Each cBlade contains both multiple 1GB Ethernet and 2GB Fibre Channel connections that provide both redundancy and higher bandwidth. Each pBlade is capable of up to 8 cores. That means a BladeFrame is capable of applying 192 CPUs toward supporting a variety of applications.

The ultimate advantage for virtualization is to apply servers and storage where the most work needs to be performed. The following virtual elements comprise grid infrastructure and create an environment where Oracle's grid computing can be maximized:

- Server Virtualization
- Network Virtualization
- Storage Virtualization

Server Virtualization

Server virtualization, at the hardware level, will be handled by horizontally scaling the two- or four-processor Egenera pBlades using the latest Intel Xeon processors and an Intel chipset for memory and I/O control. The virtualization will allow the pServers to be repurposed as a database or application server depending on the workload or needs. The PAN combines the processing and networking in a single chassis and replaces physical server components with virtual, software-based entities to virtualize data center infrastructure. Specifically, the BladeFrame consolidates processors and memory while virtualizing IP and storage networking, clustering, load balancing, hardware failover and secure partitioning all under the control of integrated management software. We expect even more powerful BladeFrame server virtualization when Intel brings processors to market with virtualization technology support native to the processor.



Oracle Software using the Processing Area Network

When the pServers are being used as a database node, Oracle uses Real Application Clusters (RAC) to cluster the database pServers together over a shared cached architecture using the BladeFrame's backplane as the high speed interconnect. Because all nodes access the same database, the failure of one instance will not cause loss of access to the database. On 10g, Oracle supplies the required clusterware called Cluster Ready Services (CRS) that provides concurrent access to the same storage and the same set of data files from all nodes in the cluster. CRS also provides a platform for Services on RAC that is used to virtualize the usage of the pServers and maximize the value of the cluster's processing resources.

Services are a facility within Oracle Database 10g new automatic workload management. With Services, different types of workloads that make up an application can be managed. Using Services, RAC connections to the database instances will be transparent to the applications on which physical pServers are being used and requires no changes to application code. In addition, Services provides scalability, load-balancing, and failover capabilities. To optimally use Services in a grid environment, all the pServers will be capable of running any of the instances for any database(s). If a failure occurs another pServer can be immediately provisioned to take its place. Each Service, depending on its processing requirements, can be assigned to one or more instances for normal startup (preferred). In addition, a secondary instance can be added to provide HA capabilities should the primary instances become unavailable. The Services can be both measured and monitored to ensure that service level agreements are being met. Utilizing services is key to the Oracle grid computing environment allowing all applications to have access to the same excess capacity and provide a faster Mean Time to Recover (MTTR).

Services combined with other facilities in the automated workload management are one of the keys to Oracle grid computing

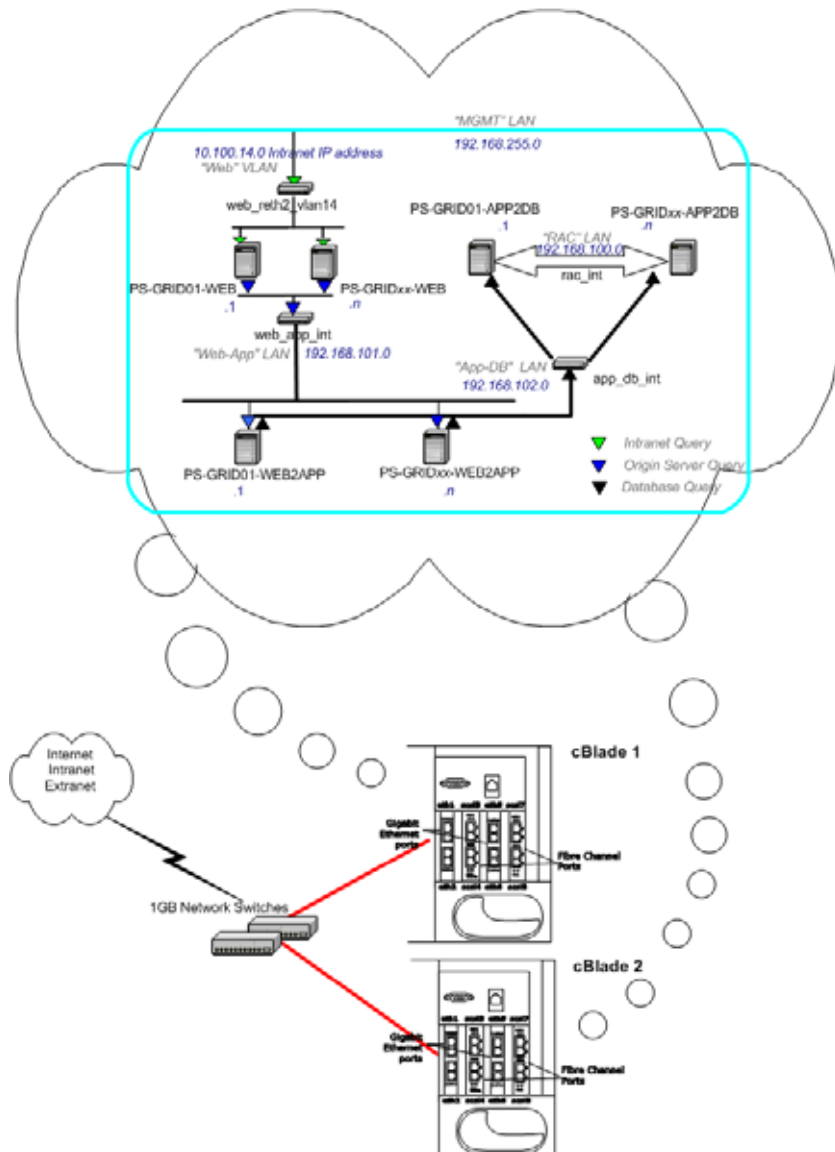
In addition to acting as RAC nodes, pServers can be repurposed to act as application servers. The Oracle Application Server 10g has a number of features that are designed to take advantage of grid computing. Oracle Containers for J2EE 10g introduces many performance enhancements that enable applications to satisfy the appropriate service level agreements. Some of these features include clustering with faster and more flexible state replication, transparent database to application server notifications, and faster data source registration to connection pools. Oracle Application Server also has a Dynamic Resource Manager to scale up or scale out applications while using computing resources optimally. The Dynamic Resource Manager interprets the resource management policies specified and routes requests based on these policies. Similar to the database “Services” facility, if an application become resource constrained, the Dynamic Resource Manager can shut down idle processes; shift capacity from other Applications that do not need them; start up new Application Server instances; or add capacity on demand. The Dynamic Resource Manager optimizes resource allocation and offloads application administrators from having to manually carry out resource balancing tasks.

Network Virtualization

Network virtualization allows the devices connected to both the virtual and physical switches to be dynamically rearranged in order to create virtual local area networks (VLANs) and assign servers to those segments. Virtualization can be both handled by the VLANs created through both external switches to the Egenera BladeFrame and internally created virtual vSwitches. The vSwitches can be created to provide communication between the pServers and also to provide VLAN-capable redundant connection to external network devices and routers.

Since all the pServers can communicate across the low latency high speed backplane and the PAN allows for the easy creation of virtual vSwitches, a three tier network architecture containing all the internal communications between servers can occur without leaving the BladeFrame. The same infrastructure that provides the high-speed interconnects for the grid’s database nodes also provide for the communication between the application server and database. That means that the following conceptual architecture could be created within the confines of the BladeFrame without any new cables or additional NICs:

A three tier network can be created in the confines of the BladeFrame and use the same high bandwidth and low latency backplane to communicate between all nodes in all the tiers



In non-virtualized environments the individual blades should be configured with multiple NIC cards to provide for failover and aggregation. Communications in Egenera's virtualized environment do not require this because the individual NIC cards are actually Virtual Ethernet interfaces (vEths) and the Virtual Switches (vSwitches) are software switches that connect the pServers and broadcast domains to external networks. These components, combined with the VLAN capabilities and individual subnets, provide highly available, private, and secure channels to both external networks and internal pServers.

For communication outside the BladeFrame, each cBlade has multiple Gigabit Ethernet ports on multiple cards. Correctly cabling these ports to multiple redundant external switches helps provide a highly available environment. This

allows for the BladeFrame to lose a cBlade, multiple cards, multiple cables, switches, and/or ports and the individual pServers can continue to work.

Since all network communication between pServers is on this high speed fabric, communications between the business logic software and the database are at better than twice the speed of normal networks between servers. Thus, transactional calls between the application server and database are more robust and quicker than physically separated servers.

Prior to virtualization, the network would need to be reconfigured when new servers or applications are added. With Egenera's use of virtualization, the network can be changed without visiting the data center to move cables and configuration can be handled via scripts and updated instantaneously.

Storage Virtualization

The Egenera PAN provides storage virtualization to the operating system. Oracle's ASM virtualizes the storage to the database.

Storage virtualization will allow the software to be provided to the servers where it is needed when it is needed. Using virtualization, additional storage nodes can be added to a shared network storage environment using multiple switches and zoning to enable higher availability, better utilization, lower cost, and centralized management. The two predominate storage technologies are Storage Area Networks (SAN) and Network Attached Storage (NAS). SAN is a network that uses the SCSI protocol to transfer data, while NAS uses the Network Filesystem protocol (NFS) to provide storage capabilities. In a grid environment, both NAS and SAN capabilities reduce the complexity associated with managing large amounts of storage devices. Furthermore, administrators have greater flexibility in allocating storage space. Using both NAS and SAN storage allows pServers to be instantaneously provisioned with boot kernels, application software, and database datafiles.

Network Attached Storage (NAS)

NAS provides storage capabilities by allowing the pServers to NFS mount the filesystems over the network. NAS enables the grid by letting database and system administrators mount and remount filesystems to different servers and permit different servers to share the same binaries. RAC can also be used over a NAS environment. To validate the NAS configuration, the Oracle Application Standard Benchmark (OASB) for Oracle's E-Business Suite was performed on an Egenera BladeFrame ES and EMC's NS700 NAS filer. In this validation, two of the pServers were used to create a two-node RAC database, three pServers acted as the application server tier, and one pServer was used in the Web tier using Oracle's Web Cache as way of load balancing against the application servers. The benchmark was able to easily sustain a load of over 2,400 mixed-workload users and maintain an average CPU utilization of 45% on the application servers and 35% on the database servers. An impressive average response time for the business

transactions was .21 seconds. While technically possible to boot the pServers over a NAS, it is not advisable. At minimum the pServers should boot from a SAN and, provided that I/O requirements are compatible with the NAS filer being used, all Oracle functionality can be used on NAS storage.

Storage Area Network (SAN)

The SAN is an extended link between the server and storage that allows the SCSI protocol to be used over longer distances. Despite being diskless, the BladeFrame was designed to complement SAN storage and utilizes some advanced SAN features like multipathing, path grouping, and persistent binding. All this complexity is configured at the two cBlades. This means the individual pServers transparently leverage the interfaces to the SAN storage through the cBlades, which are configured via the PAN Manager software. The BladeFrame can be connected to direct (or arbitrated loop), single-switch, and multi-switch fabric connections. Leveraging multipath allows storage administrators to connect the BladeFrame to multiple redundant SAN networks for high availability and performance. The host bus adapter on each cBlade connects to the storage environment through SAN switches so that two of the HBA adapters on each cBlade see the same disk storage. Multipath distributes the I/O requests among the eight paths from the two cBlades to the SAN storage. This increases the bandwidth of the I/O and the number of I/O operations per second. Also, if one cBlade fails or reboots, the other can fulfill any I/O requests.

Automatic Storage Management (ASM)

Storage Virtualization within the database is handled using ASM. Before ASM, DBAs needed to manage files and drives individually. Storage for the database would require expensive storage devices capable of different RAID levels for performance and redundancy. With ASM, IT can leverage inexpensive storage devices like purchasing just-a-bunch-of-disks (JBOD) and have the ASM instance manage the redundancy and striping capabilities. Disk groups can be created consisting of disks and their assigned files. By combining the use of Oracle Managed Files (OMF) and ASM, DBAs no longer need to specify file names and locations for the physical database files when creating and maintaining the database. ASM solves the problem of poor storage utilization by letting administrators create disk groups based on the performance and redundancy of the disks and making them appear as one homogeneous set. These sets of disks can be created to handle a multitude of different storage capabilities. For example, for highly active datafiles, the SAN device handles the redundancy and striping. For less active datafiles and flashback recovery areas, the redundancy and striping are handled by the ASM instance. Another benefit is that ASM does not need to be used exclusively. A single database can have datafiles that exist as regular files and other datafiles managed by ASM. This allows for using both NAS and SAN storage in a single database.

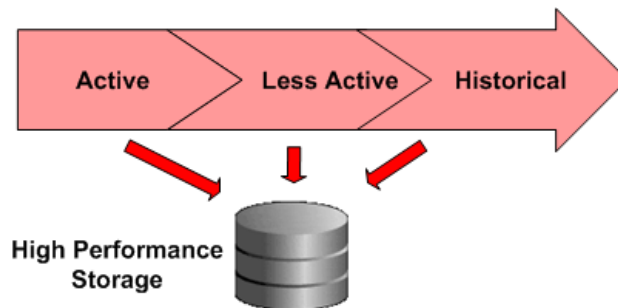
Essentially, ASM acts like a volume manager built into the database kernel. Just like a logical volume manager, that spreads files across multiple disks, ASM spreads the extents across the disks in a disk group. Using ASM also eliminates the complexity of adding disks or LUNs to the database node by automatically redistributing the load over the new disk while the database is running. In addition to distributing for performance, ASM can protect against data loss by creating mirrored failure groups.

Information Lifecycle Management (ILM)

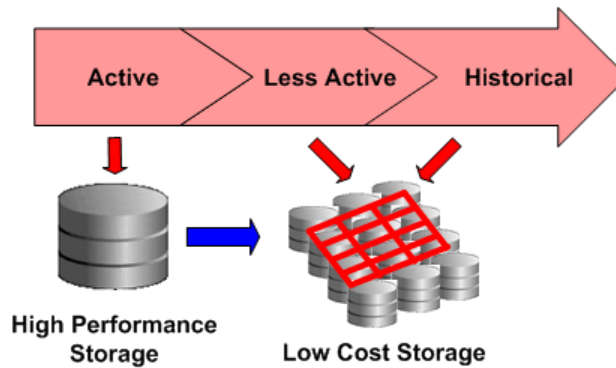
Information Lifecycle Management is becoming increasingly important, as regulations require more information to be retained for long periods of time

In addition to making sure that storage virtualization provides storage where it is needed when it is needed; it also needs to be the right type of storage. The grid environment is also responsible for storing vast quantities of data, for the lowest cost, and meeting the new regulatory requirements for data retention and protection. Information Lifecycle Management (ILM) is concerned with everything that happens to data during its lifetime. ILM, through the use of virtualization, seeks to ensure that data is stored on the correct media to satisfy the need to have data immediately available and various regulatory requirements.

In many IT data centers all information is treated the same. Whether the data is recent and active or is not active and only available for historical reasons, all the data is stored on the same expensive high-performance storage facility. What is needed is a way to leverage virtualization to assign lower-cost storage to the appropriate storage type as it moves through the data lifecycle.



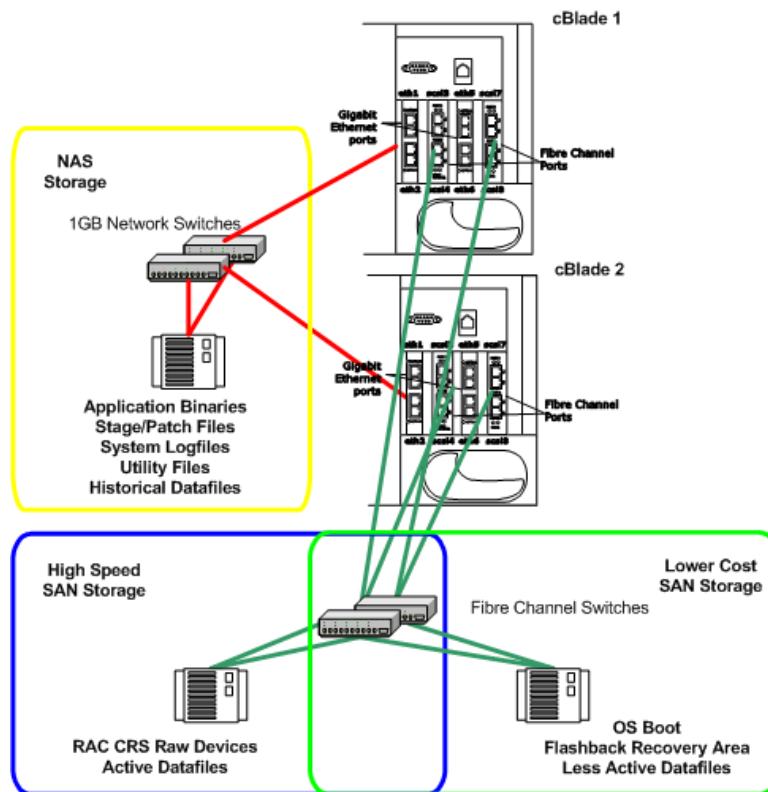
In ILM, storage can be divided up into multiple tiers depending on performance, cost or history. Using ASM, each tier can be assigned a disk group and data can be partitioned across the disk group. When data needs to move to the next tier, Oracle database's online operations, partition moves, or copying of the tablespace can be used to facilitate the transfer to the next disk group.



Another important aspect of the ILM strategy is backup and recovery. As the database continues to expand, the time and storage required to perform backups can start to increase. Oracle 10g Recovery Manager (RMAN) can be used to incrementally back up and create an image copy of the database into an online recovery area that can be available to “flashback” the database. At regular intervals, either the full or incremental backups can be moved to tape for offsite storage. During restoration, RMAN advises which backupset will be needed based on the catalog that can be stored in a controlfile or RMAN repository.

Putting It All Together

A traditional environment would require each server to have a SAN adapter and a Gigabit network adapter. Doubling all of these adapters would provide redundancy. Also, a series of network and SAN switches would be required to connect the servers completely. The BladeFrame allows all these storage options to be used together both virtually and redundantly. The BladePlane interconnects all pBlades to the cBlades, dramatically reducing the need for external connections. All required network and SAN connections for 24 pServers can be redundantly configured with eight physical connections from the cBlades. While the PAN allows the storage to be virtualized to the pServers, ASM allows the storage to be virtualized to the database. The diagram below shows a conceptual view on a configuration that utilizes all the storage options. Of course this can also be turned into a SAN-only or a mainly NAS solution.



MANAGING THE FRAMEWORK

It is critical to have tools to easily and efficiently manage the virtual environment. Oracle and Egenera both provide tools that provide this capability. Oracle's Grid Control provides a central point for automating provisioning, monitoring the operating system and databases, ASM instance monitoring, adding disks and disk groups, and doing this across all the BladeFrames. Egenera's PAN Manager provides access to the BladeFrame and allows creating virtual networks and provisioning the pBlades into pServers.

Oracle's Grid Control

Oracle's Grid Control can automate the installation, configuration, and cloning of the database and application servers across multiple pServers. The framework provides the ability for administrators to create, configure, deploy and utilize the pServers with new instances of the database and application servers. Administrators are also able to apply patches and upgrade existing pServers. In short, Grid Control enables grid administrators to manage the grid environment through a Web browser and maintain the system's lifecycle, front to back and from any location on the network.

Grid Control provides an extensible framework from which monitoring and administrative capabilities can be customized

Grid Control is installed from a single instance of Oracle's application server and uses agents deployed to the pServers to maintain communication with the individual nodes. When the Management Server piece of Grid Control is maintained on the pServer, the high availability features of PAN Manager can be used to create cold failover capabilities to keep downtime to just minutes as the Management Server is transparently brought up on another pServer. The Egenera HA capabilities will assume the IP address and run the shutdown and startup scripts, allowing the administrators to move the server between pServers upon failover or at a click of the mouse.

Egenera's PAN Manager

PAN Manager software features a Command Line Interface (CLI) and a browser-based GUI; roles-based privileges; and remote, lights-out access to permit dynamic reconfiguration, expansion and upgrades. Moreover, out-of-the-box integration with leading enterprise management consoles is provided, allowing customers to leverage BladeFrame resources from within an existing data center infrastructure.

Through software-based resource allocation, the BladeFrame enables a system administrator to add or remove virtual servers from a cluster on-the-fly, a process that is both time-consuming and disruptive with legacy platforms. Egenera PAN Manager software, which virtualizes 80 percent of server hardware components, replaces error-prone, physical activities with point-and-click or scripted commands to speed deployment, lower management costs and eliminate the need for complex clustering software.

Ease of Configuration

PAN Manager software enables users to easily configure pServers with all required connections. During creation of a pServer, the user can attach more than 256 SAN disks and up to 30 individual network segments to that pServer. All connections are done virtually through PAN Manager software, reducing the physical layout of wiring and switching. New pServers can be configured as needed and added to the Oracle cluster without closing the database.

The Oracle 10g Real Application Clusters database environment can use many disk partitions defined by the database's file set requirements. Partitioning of these disks is performed centrally from PAN Manager software.

Database

Software loading can also be performed across all instances from PAN Manager. Moreover, boot control and instance status can be managed. Monitor software is available to determine the health of the database service and of the instance server. A series of failover policies can be configured in response to the monitor's events. These capabilities ensure that the database remains highly reliable and accessible.

**Egenera's Hardware and Oracle's Software
provide a fully integrated grid ready
infrastructure.**

CONCLUSION

The goal for modern IT enterprises is to optimize IT resources. IT can achieve this by using the concepts of data center virtualization and transform physical resources into virtual software components that can be pooled and centrally managed for applications as required, and using “grid computing” to virtualize the applications. The BladeFrame’s virtualization allows the pServers to participate in any of the tiers on demand. Oracle software takes advantage of the pServers in all the tiers: Web Cache Clusters at the Web tier, Application Server clustering at the apps tier, and RAC for the database tier.

Oracle's grid database technology is an excellent fit on Egenera's BladeFrame. Each of the pServers appears to be a standalone Linux server to the Oracle software; however, these individual machines are tightly linked through the Egenera backplane and software system to provide a level of computing purpose-built to enhance clustered applications. The high-speed backplane offers more than double the network communication of gigabit Ethernet between pServers. This facilitates all interactions between pServers, either for RAC's cache fusion or the sending and retrieving of data between databases and application servers and external network-administrated storage devices. Full hardware control from a central console allows ease in scaling and maintaining servers.

The Egenera Blade Frame enables data centers to run any application, on any resource, at any time, automatically. The BladeFrame minimizes complexity and optimizes responsiveness through flexible, secure resource sharing. With the Egenera BladeFrame system and Oracle 10g, you can simplify the data center with confidence and agility. Combined, Egenera and Oracle provide a fully integrated infrastructure, both hardware and software. This solution provides performance and sophistication since it was purpose-built for the enterprise grid environment.

The PAN architecture combines processing and networking into a single chassis and replaces physical server components with virtual, software-based entities to virtualize data center infrastructure. Specifically, the BladeFrame consolidates processors and memory while virtualizing IP and storage networking, clustering, load balancing, hardware failover and secure partitioning all under the control of integrated management software.

Oracle and Egenera's Accelerator Service

Oracle and Egenera have jointly tested and developed best practices for the deployment of Oracle 10g databases on Egenera's BladeFrame systems. This new joint Accelerator service provides customers with the capability to have Oracle and Egenera professionals use field-proven best practices to rapidly deploy Oracle databases and applications on Egenera's BladeFrame systems for a fixed price, in a short, fixed period of time.

Purchasing this Accelerator Services provide customers the following benefits:

- Integrated delivery approach using jointly pre-defined objectives as well as post deployment testing and acceptance with your IT staff
- Avoid resource impact of integrating and customizing Oracle and Egenera proven solutions for your business
- Rapid deployment from Oracle and Egenera Professional Services, shortening time to productivity
- Customized, business specific repeatable process—where your IT staff is trained alongside Oracle and Egenera experts.



Silo Busting: Deploying Oracle's Grid Computing on the Egenera BladeFrame
April 2005
Author: Eric Evans
Contributing Authors: Joe Gorski

Oracle Corporation
World Headquarters
500 Oracle Parkway
Redwood Shores, CA 94065
U.S.A.

Worldwide Inquiries:
Phone: +1.650.506.7000
Fax: +1.650.506.7200
oracle.com

Copyright © 2005, Oracle. All rights reserved.

This document is provided for information purposes only and the contents hereof are subject to change without notice.

This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle, JD Edwards, and PeopleSoft are registered trademarks of Oracle Corporation and/or its affiliates. Other names may be trademarks of their respective owners.