



White Paper

Advantages of Running Oracle® Real Application Clusters
on the Egenera® BladeFrame® System

Executive Summary

Over the past decade, the use of clustering has grown in response to the need for high performance computing (HPC), for high availability (HA), and for load balancing. Database applications, however, have typically been run on single servers, often requiring expensive, high-capacity Symmetric Multiprocessing (SMP) systems. These systems also represent a single point of failure for one of the most critical applications in a datacenter.

Innovations by Oracle with its Real Application Clusters database server provide a more reliable and cost-effective alternative—a scalable, highly available database on Storage Area Network (SAN)-connected x86 clusters. The Egenera Processing Area Network (PAN) architecture implemented on the Egenera BladeFrame system provides many unique advantages for this clustered application.

Oracle's Cache Fusion technology allows clustered network servers to share data blocks and reduce time-intensive reads and writes. The high-speed, low-latency switched fabric incorporated into the BladeFrame performs these data transfers at speeds greater than Gigabit Ethernet networks.

Cache Fusion technology implements a "shared cache" architecture, also known as a "shared disk" or "shared everything" architecture, where each server in the cluster has full and complete access to all data in the database. This type of cluster is very scalable as each server can operate autonomously on any part of the database. The inherent nature of the hot-swappable Egenera Processing Blade™ modules in the BladeFrame, as well as the system's centralized management, are tailored to adding devices on demand and supporting scalability.

Highly available databases offer continuous operation through server hardware and software failures. Egenera PAN Manager™ software is purpose-built to monitor for failures and return services to operation by reinstantiating failed servers automatically.

The management of complex clusters is simplified with PAN Manager software and the BladeFrame design. The assignment of multiple SAN disks to single servers and the manipulation of disk partitions are managed by PAN Manager. Creation of multiple network segments is also easily managed and implemented.

The high cost of clustering servers is due to the many SAN adapters and network interface cards (NIC) required to interconnect them. These adapters and external switches are virtualized and made redundant in the BladeFrame, reducing hardware and operational costs dramatically.

The Egenera BladeFrame provides distinct, cost-effective extensions to the innovations offered by Oracle Real Application Clusters.

Background

Oracle Real Application Clusters

Oracle has provided a scalable and highly available database server in its Real Application Clusters option. Both of these attributes are important to the enterprise user. Scalability provides a way to increase the processing capacity of an existing database. Being highly available keeps the database open, even through hardware and software failures.

Oracle Real Application Clusters is run on a cluster of servers. Each server has access to the same database storage devices and is connected to other services via normal network protocols. The portion of the database running on a single server is called an instance. Each instance in the database has read/write access to all disks used by the database. These disks contain all pertinent client data as well as logs of transactions and processes from each server.

Cache Fusion is the name of the Oracle technology used to mediate access to storage devices and allow inter-server transfers of data blocks along network connections. This intercommunication allows every distinct instance to share data in its cache with every other distinct instance, eliminating the need for either costly disk 'pings' or application and disk partitioning.

High availability is built-in by design. During an instance failure, of either a hardware or software nature, any work uncommitted is replayed on any other instance and all remote database connections are moved to an alternate instance for continued processing. This ability to replay in-flight transactions on any other instance allows continuous database processing as long as one instance remains available.

Egenera Processing Area Network (PAN) Architecture

The Egenera PAN architecture is a group of autonomous, stateless Processing Blades (pBlade™) redundantly connected via a high-speed, low-latency switched fabric. This technology is similar to a SAN in that it isolates compute components to allow versatility in their application. The PAN is controlled by a redundant set of Egenera Control Blade™ modules (cBlade™) which provide external network and SAN connectivity as well as configuration management through PAN Manager software. The architecture offers advantages in rapid deployment, simplified management, efficient processor utilization, high reliability, high availability and scalability.

The PAN architecture is implemented on a BladeFrame to virtualize connections among pBlades and to external SAN resources. The pBlades are coupled to disks on the SAN to create virtual servers called pServers. These pServers are connected via a high-speed, low-latency switched fabric along the Egenera BladePlane™. The relationships of disk and network connections on a pServer are maintained by PAN Manager software.

Clustering of Instances on pServers

Cache Fusion

Oracle's Cache Fusion technology provides an elegant clustering solution to yield a loosely coupled clustered database. The design allows for instances of the database cluster to operate autonomously and to communicate all clustering information through normal network connections. The "shared cache" nature of the cluster makes it very scalable. New servers, and hence processing power, can be added to an existing cluster with limited network coordination and virtually no interruption of existing service.

Scalable BladeFrame

The BladeFrame was designed with expansion in mind. Each autonomous pBlade is hot-swappable and upgradeable. Thus, new pServers can be configured and added to a cluster through PAN Manager software with no rewiring. As more powerful Processing Blades become available, they can be swapped into the BladeFrame for an immediate performance increase.

Minimal Disk Locking

Inter-instance communication is used to reduce contention for the same data blocks by more than one instance at a time. Since all instances share the same data disks, simultaneous requests for rows in a data block may be made by more than one instance. Cache Fusion negotiates between instances so that once a block is read it can be passed along the network to the next instance requiring access. This reduces the need for the first instance to write the data block, and for the second instance to read the data block. Reductions in disk access have been shown to increase database performance.

Egenera BladePlane

The BladeFrame incorporates a high-speed, low-latency, inter-blade switching fabric, called the BladePlane, to link its pServers. The BladePlane virtualizes and optimizes traditional network connections, providing an excellent media for the Cache Fusion interconnect and enabling inter-pServer communications significantly faster than Gigabit Ethernet traffic. Communications along these internal connections are isolated and redundant, providing security and robustness.

Highly Available Database

Oracle Failover Process

Every instance server in an Oracle Real Application Clusters deployment acts as a backup to every other instance server. Oracle's Cache Fusion technology allows multiple instances to share disk resources and communicate work status along the cluster's network segment. This results in a system where any instance can look up and finish the processing of any transaction not committed due to an instance failure. The database remains open since all remaining instances continue processing after a failed instance is removed from the cluster. As long as one instance stays online, the database remains available.

PAN Architecture Extends Oracle Failover

The BladeFrame and its PAN Manager software were specifically designed to handle server hardware and software failures. Since a pServer is built from a virtual linking of SAN disk resources, virtual network connections and a stateless pBlade, it can be reconfigured and restarted automatically. Thus, Oracle failover handles the processing missed by the failed server while PAN Manager software rebuilds and reintroduces the lost instance, returning the cluster to full strength automatically.

Managing a Clustered Database

Reduced Connection Hardware

The Oracle Real Application Clusters database environment requires that each server instance have SAN storage and fast network connections for good Cache Fusion performance. Thus, a traditional server configuration would require a SAN adapter, a NIC and a Gigabit network adapter. Doubling of 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 provides all of these connections 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. These external cBlade connections are configured during BladeFrame installation.

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 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.

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.

Conclusions

The BladeFrame's advantages in running the Oracle Real Application Clusters database server are many. Very fast inter-instance communication along the BladePlane increases the performance of Cache Fusion technology. Built-in pServer hardware failover complements Real Application Clusters' transparent application failover by reinstantiating a failed pServer quickly and automatically. Management of a complex clustered environment is simplified by PAN Manager software. Finally, the hardware cost of constructing the fast network and SAN-enabled cluster is almost completely eliminated as required connections are virtualized within the PAN architecture.

The Egenera BladeFrame naturally extends the benefits of Oracle Real Application Clusters like no other platform presently available.



Corporate Headquarters
Egenera, Inc.
165 Forest Street
Marlboro, MA 01752
U.S.A.
Phone: 508-858-2600
Fax: 508-481-3114
www.egenera.com

European Headquarters
Egenera Ltd.
Venture House
Arlington Square
Bracknell, Berkshire RG12 1WA
United Kingdom
Phone: +44 (0)1344 475237
Fax: +44 (0)8703 305946
www.egenera.com

Asia Pacific Headquarters
Egenera K.K.
Shinjuku NS Bldg. 6F,
2-4-1 Nishishinjuku,
Shinjuku-ku
Tokyo 163-0806 Japan
Phone: +81-3-5321-7157
Fax: +81-3-5321-7158
www.egenera.com