The functionality and advantages of a high-availability file server system

This paper discusses the benefits of deploying a JMR SHARE High-Availability File Server System. Hardware and performance considerations are discussed, as well as deploying the SHARE system as a SAN and a High-Availability Cluster System.
THE FUNCTIONALITY AND ADVANTAGES OF A HIGH-AVAILABILITY FILE SERVER SYSTEM

Background

High availability storage has seen explosive growth over the years, as users become increasingly reliant on their digital assets and the speed at which those assets can be transferred from work stations to servers, and back again. In systems where a single file server is used for user-critical assets, a server failure can have from inconvenient to catastrophic consequences. As such, mission-critical storage systems have evolved; developers and manufacturers have taken various approaches to this growing need.

In a virtualized environment or a 24-hour business cycle where keeping data and file assets safe are critical, the need for high availability becomes evident. A high-availability cluster can be used to provide critical shared data on a network, and a popular approach is to mirror all data in real time between identical storage servers. Using RAID storage servers with sufficient capacity to meet current and future needs, along with specialized operating systems providing seamless and highly fault-tolerant mirroring operations, is a very good method for guaranteeing nearly 100% up-time without excessive cost.

With conventional NAS or SAN servers, even with RAID protected storage, there are still single points of failure that can take the system down. We know disk drives fail, and RAID operations can protect against a single or sometimes even multiple drive failures; however, that doesn’t protect against controller, CPU, memory or other mother board failures, or power supply combining circuit failures. Network port failover features and power backup systems are helpful, but if the single server crashes, you’re down until it’s repaired or replaced, and sometimes last transactions are lost.

Benefits of a High-Availability Cluster

With a high-availability cluster, if a hardware fault occurs, it is detected via intelligence features and the software seamlessly (and instantly) moves storage requests to the other mirrored server. Smart systems use a cluster of two mirrored servers with real-time continuous data replication and synchronization; both servers contain identical data to offer absolute redundancy.

As an example of such a system, the JMR BlueStor™ SHARE™ system with HA Cluster powered by euroNAS® supports:

- SMB/CIFS/AFP
- Active directory authentication for CIFS and AFP
- NFS
- iSCSI, with persistent reservation iSCSI LUN and target services

Powerful cluster management functionality enables the user to define which resource should run on which server; this way, both servers are used in the most efficient way possible – providing the highest level of performance simultaneously. If one server fails, services automatically move to the other server; when the failed server comes back online, its services will automatically move back and normal operations resume as they were pre-failure.
This feature is not dependent upon services – you can have iSCSI targets split over both servers, or run iSCSI targets on one server, with CIFS/AFP services on the other. Resources created on the same cluster drive will run together on the same node.

The JMR SHARE with HA (High-Availability) Cluster is an ideal solution for IT professionals, as you gain not only high availability storage with continuous replication but also a very efficient and easy-to-use management interface. The server and cluster configuration can be completed in less than ten minutes.

Architecture

The high-availability architecture includes the following:

- **Cluster node**: This is the individual server that is a member of the cluster; all nodes are equal, and from each node the user can monitor and configure cluster services.

- **Cluster drive**: This represents a mirror of two drives on each individual server; the two drives can be a single disk, or preferably a RAID array, which are mirrored in real-time. On the drives, shares and iSCSI targets are installed. For example, each individual server contains 16 disk drives (as in the JMR SHARE), providing up to 128TB capacity and each “drive” (RAID array) can be defined to run on the server. All resources defined on this cluster drive will run on the server; this way, the load between the servers will be balanced to achieve the best possible performance.

- **Cluster resource**: This represents network share or an iSCSI target within the cluster, and may be SMB/CIFS & AFP, NFS share or an iSCSI target. Each resource has its own size and is accessible through the IP address defined for the cluster drive.

- **Corporate network**: The network used by clients for accessing cluster resources.

- **Replication network**: Internal network used for heartbeat and data replication between servers.

- **Cluster access IP address**: The address that enables clients to access shares and iSCSI targets. This address is shared by both HA cluster systems; if any server fails, the cluster IP address will move to the remaining server.

- **Network test IP**: This prevents storage desynchronization (“split brain”) situations which might occur when both nodes are running but one or both are network disconnected. HA Cluster nodes ping this address constantly and check if it is still able to reach the network. IP addresses on any pingable device on the network that is constantly powered (router, gateway, printer, mail server, etc.)
can be used as network test IP. If one of the nodes is unable to ping this IP, it will automatically know it cannot reach the network and push all services over to the other node. If a node cannot reach all defined test IP addresses, it will shut down services to assure data remains consistent.

Cluster Events

Within the cluster, different types of events can occur. Depending upon the type of event, transfer of services can take varying amounts of time. Initial time for servers to realize the other server has failed is about 30 seconds; after this, resources are moved quickly.

Estimated time to move ten resources on another server is less than 60 seconds. To move 30 resources usually takes 70-80 seconds. This depends on the performance of the hardware and their storage capacity. The JMR SHARE is optimized to perform these transfers as expeditiously as possible even when compared with systems at substantially higher cost.

- **Failover**: This occurs when a server detects a fault and automatically moves all resources to the other server.

- **Switchover**: Switchover occurs when a resource is manually moved from one server to another. This may be performed when node preference settings are changed, or during system management of one of the nodes.

Hardware Considerations and Performance Recommendations

The hardware used is as important as the software. Systems like the JMR SHARE is well optimized to run with euroNAS HA Cluster OS, which is a very robust Linux-based operating system that has proven extremely reliable. Using separate mirrored system disks, even that fault possibility is covered. In the JMR SHARE, a 3.5 GHz multicore CPU and 16GB high speed RAM, along with an internal 6Gb hardware SAS RAID controller (usually set to RAID 50 or RAID 60) allows each disk array to perform at about 1,800 MB/s internal transfer rates.

This is good news for prospective users, but the corporate network can be the choking point for any shared storage system. While GbE is common and multiple network ports are provided by the SHARE server, 10GbE obviously allows greater bandwidth and faster operations. The SHARE server provides dual 10GbE ports in addition to GbE, and these serve to bond not only the corporate network but also the two servers to each other (the private replication network).
A JMR SHARE HA Cluster system consists of two identical servers with HA Cluster installed on mirrored OS drives, the RAID storage drives are initialized and ready for your data, and the provided servers are tested in cluster configuration.

Limitations

There are some limitations to all systems, and the HA Cluster has a few:

- **Replication network failure**: Although the corporate network may remain operational, if the replication network which provides the synchronization “heartbeat” between the nodes fails, the cluster will not perform properly. To prevent this occurrence, we always recommend network card teaming (port failover), and provide for this with the SHARE system.

- **Powering down master node during synchronization process**: During the sync process, if the node with more current data is powered off, server replication service will cease to prevent data corruption. As such, only the server (node) with less current data may be powered off; however, if both servers are fully in sync, either of the servers may be powered off or re-booted.

- **Port failover (active backup) must be used for configurations without a switch**: If the replication network is set to switchless configuration, the only network bond that may be used is port-failover (active backup); it is the only one that does not require any switch.

Creating a SAN using JMR SHARE

As discussed earlier, a high-availability cluster is valuable for those who rely heavily on continuous access to assets and need assurance assets will never be lost due to a variety of possible disruptions.

Until recently, building a highly available and high-performance Fibre Channel storage area network was quite expensive. The JMR SHARE solution powered by euroNAS provides what were “high cost” features in a lower cost SAN that is also easy to deploy, bringing Enterprise-class SAN functionality to small and medium business operations.

Some benefits of a Fibre Channel SAN are proven reliability, high bandwidth and guaranteed order in delivery of data packets. Users have enjoyed such benefits for years, but often in systems still having a single point of failure, being the primary file server. Drawing on the technology described earlier in the networked HA Cluster, business continuity provided by a SAN cluster is now available for very reasonable cost.

While Ethernet is the most popular network protocol and is low-cost and fairly easy to set-up, it is not designed for transferring block data in a networked storage environment.
Ethernet’s way of handling data collisions, when more than one computer tries to transmit data simultaneously, can reduce the network’s efficiency due to retransmissions.

Fibre Channel with its asynchronous design helps assure that even under very heavy load conditions, collisions are handled efficiently to maintain the highest possible throughput.

**What Makes It Better?**

iSCSI was developed as a cost-effective alternative to Fibre Channel; it is built on an underlying TCP/IP protocol and is normally implemented as a software initiator that incurs substantial processing overhead. By contrast, Fibre Channel commands have been directly implemented into HBAs (Host Bust Adapters), making it more efficient and less CPU consuming. It can be more suitable for data transfers.

One big advantage of Fibre Channel is constant data throughput. While Fibre Channel consistently performs at very close to its line speed, e.g., 8Gb/s or 16Gb/s, collision management on Ethernet makes it difficult to achieve such consistent performance.

The JMR SHARE SAN Cluster is a native Fibre Channel product, designed to provide critical shared data on SAN workloads. When using the SAN Cluster, your network applications will continue to work: No matter if a server fails completely, clients will still have access their data. Using only “backup,” it takes considerable time to recover and manual intervention is usually required to get server and data back on line.

With the SAN Cluster, should a fault occur, it is detected through intelligent features (SSM, or Smart SAN Mirror) and the software automatically moves storage requests to the other server mirrored in the cluster, the same as in HA Cluster discussed earlier. All the benefits of HA Cluster are available in the SHARE SAN Cluster, just using Fibre Channel as the protocol and transport. Using the SAN Cluster will provide:

- Highest redundancy for mission critical applications
- Highest I/O performance thanks to Fibre Channel technology
- Advanced data integrity protection against data corruption
- Multipath I/O technology for even higher redundancy
- Synchronous data mirroring using native Fibre Channel protocol
- Automatic failover
- Failback with fast synchronization
- FC port management: Allocate FC targets to single or multiple ports
VMware vSphere® Storage APIs

SAN Cluster supports VMware VSphere Storage APIs – Array Integration (VAAI). SAN Cluster communicates with VMWare ESXI and enables it to offload storage operations to the cluster, which reduces resource overhead on the ESXI hosts, improving performance for storage intensive operations.

SAN Cluster Components

To achieve high availability, the JMR SHARE SAN Cluster components include:

- **Cluster node**: An individual member server of the cluster.

- **Active node**: The server providing storage to the clients; it replicates data to the passive node and assures the data is consistent on both nodes.

- **Passive node**: Contains the same data as the active node and stays in standby mode as long as the other node is active. In the event of an active node failure, it will automatically take over and continue to provide storage to the clients.

- **Cluster drive**: This is generally a RAID set of drives in each server, which are mirrored in real time. On each drive, Fibre Channel targets are installed. Up to 50 cluster drives are supported.

- **Corporate connection**: The FC connection used by the clients for accessing Fibre Channel cluster targets. More than one FC port can be used for this connection.

- **Replication connection**: Internal Fibre Channel connection used for data replication between the nodes; you can also use more than one FC port for this connection.

- **MPIO/ALUA (Asymmetric Logical Unit Access)**: Supports and delivers high availability by establishing multiple sessions from a client to each node of the SAN Cluster through Fibre Channel. If a device in the path fails, I/O requests will be automatically redirected to an alternate path for continued application availability. In the SHARE SAN Cluster system, the active node will be automatically recognized as active and the passive node will be automatically recognized as a standby node to the client.

The Cluster Events and Hardware Considerations for the SAN Cluster are the same as described earlier for the HA Cluster, except for the SAN Cluster, we provide dual-port FC HBAs at 8Gb or 16Gb/s and use SAN Cluster OS instead of HA Cluster OS. In all cases, mirrored system disks are provided and all the enterprise features of the systems are retained.