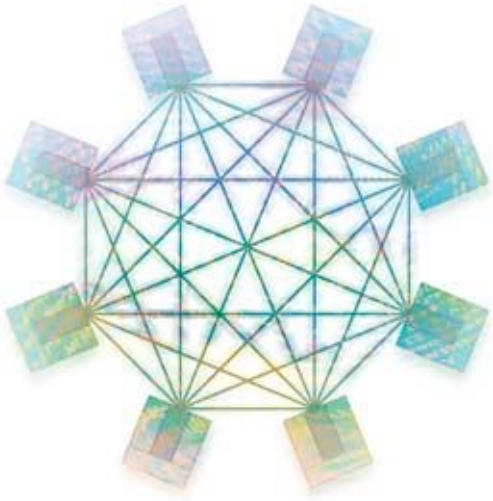
The benefits of clustered block storage



Clustered storage offers advantages in both performance and scalability, but users need to evaluate three different architectures.

By Ray Lucchesi

Today's data centers can add storage capacity to almost incomprehensible levels. However, optimal storage capacity utilization or increased capacity does not necessarily equate to increased or optimal performance. In fact, performance upgrades have traditionally been achieved only through adding storage, increasing front-end and back-end links, expanding cache capacity, and/or more efficiently distributing I/O workloads across disk drives. Some of these configuration changes can result in significant performance enhancements. However, clustered block storage technology achieves simultaneous capacity and performance improvement unattainable through normal hardware configuration alternatives. In fact, clustered block storage brings a new dimension of functionality: performance scale-out.

Clustered block storage technology provides scalable storage services with multiple interconnected, yet independent, hardware units called nodes. As storage capacity increases are warranted, data centers meet these needs by introducing another node(s) to the clustered storage system. All new nodes are interconnected and integrated with the existing clustered storage to provide a single system image of all nodes.

Nodes of a clustered block storage system often include x86-class servers consisting of multi-core processors, memory, and storage. A data center could potentially have hundreds of interconnected yet independent nodes.

Clustered block storage is often likened and compared to grid storage. While grid storage does operate as a single storage system, it usually requires geographically dispersed storage units. In contrast, clustered block storage need not be geographically dispersed. Even though IT disaster-recovery policies might mandate off-site replication of a clustered block storage system, the existence of two locations alone does not mean the storage systems are operating as a single unit. As such, all grid storage can be described as clustered block storage but not all clustered block storage systems are grid storage.

Advantages of clusters

Silverton Consulting
Strategy, Storage & Systems
P.O. Box 492|Broomfield, CO 80038-0492|
720-221-7270|info@silvertonconsulting.com

Clustered block storage's main advantage stems from the data center's ability to more closely match performance needs to current requirements. Specifically, users can purchase storage today and increase performance when necessary, just by adding more nodes. In contrast, monolithic and/or modular subsystem users must either buy more performance than needed or must forklift upgrade to another class of subsystem to satisfy additional performance requirements.

Clustered block storage technology can also be less disruptive during the introduction of new technology. Clustered storage by its very nature can be upgraded transparently to the latest technology without impacting host I/O operations. A new, higher-performance node can be introduced to the cluster, and the cluster automatically puts the new node into service. Old cluster nodes can be taken offline and removed from the cluster, again without I/O service interruption. In comparison, a monolithic storage system would not allow gradual systematic performance upgrades. Any substantive performance upgrade would most likely involve a cumbersome, disruptive forklift upgrade to another class of system.

Both the performance scalability and the gradual introduction of performance enhancing technology can result in a significant cost savings to the data center. The cost of a leading-edge storage node is far less than the purchase of excess performance often necessitated by expensive monolithic systems. Cost savings can also be realized by the lack of disruption and downtime experienced in installing a new node versus a total storage system upgrade.

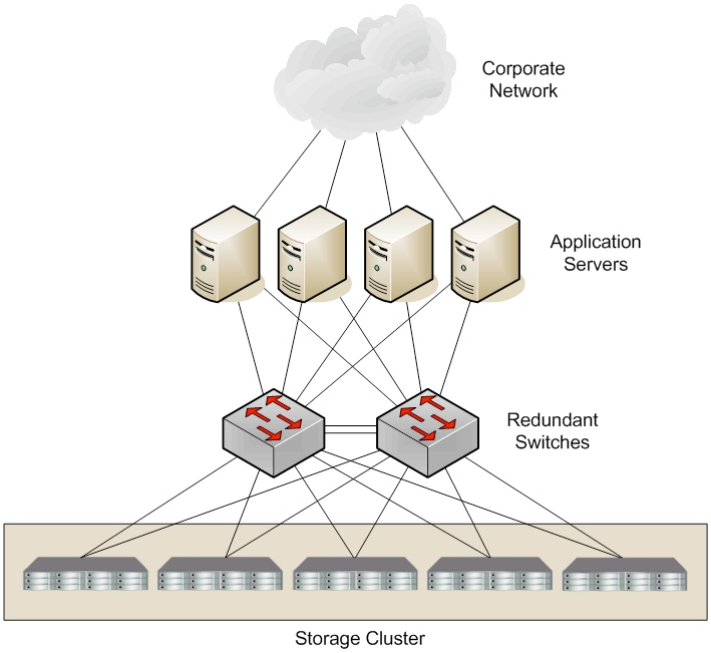
Performance scalability of clustered block storage is particularly impressive in throughput performance. Introduction of new nodes results in significant throughput service gains exceeding traditional monolithic system incremental advances. As such, clustered block storage is particularly well-suited for applications that demand high data streaming workloads with large capacities. In fact, with the recent emergence of rich media digital content and the low cost and high throughput advantages of clustered block storage, this technology will become more commonplace.

Downsides to clustered storage

Data redundancy overhead for most node-oriented clustered block storage systems is significant and depends on mirrored data. One architecture (discussed below) has reduced redundancy overhead but not to the substantially lower redundancy level of RAID 5 (25% or less). However, "stacked controller" products (see below) can be configured with RAID 5 as well as data mirroring and as such do not share this disadvantage.

In contrast to the significant throughput performance advantage of clustered block storage technology, response-time performance of some clustered block storage is at a disadvantage. Reading and writing of small blocks of data generates substantial I/O overhead for clustered block storage. Such overhead can result in considerably longer response time to a particular I/O request in clustered block storage than with a more traditional storage subsystem. As such, clustered block storage is not well-suited for transaction-oriented, small block, randomly accessed I/O workloads. Once again, "stacked control-

ler" clustered storage proves the exception and maintains very good response time for these workloads.



In a typical "all nodes =" architecture, the storage cluster consists of independent storage nodes.

Another, often-overemphasized, disadvantage of clustered block storage technology is the reliability of node hardware. As most clustered storage architectures use x86 servers, a node could conceivably fail more frequently than special-purpose hardware. However, some clustered storage nodes are based on rugged-ized, highly reliable x86 hardware components that meet or exceed any reliability requirements. In addition, the redundancy requirements of clustered block storage technology would mitigate this concern. A failed node does take considerable time to rebuild and is inoperative for that time, but operations would only be catastrophically impacted with multiple, concurrent node failures.

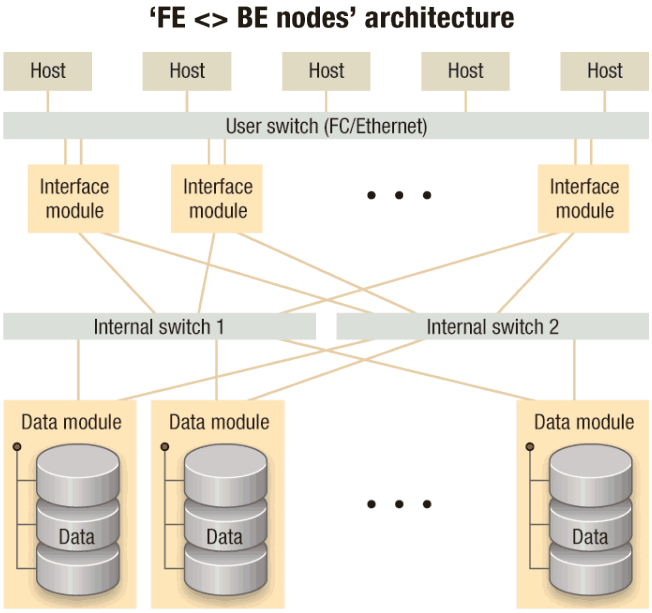
Vendors/products/architectures

Only a few vendors support Fibre Channel-based clustered block storage. Examples include XIV (recently acquired by IBM), Digi-Data, Fujitsu, and 3PAR. Most vendors have instead supported an iSCSI block interface. Examples include CleverSafe, Dell EqualLogic, Intransa, LeftHand Networks, Pivot3, and Sun (the Thumper product line). It is unclear why clustered storage has been so successful in the iSCSI market, but presumably it reflects a willingness to adopt the latest technology to economically manage data growth.

Basically, there are three architecture alternatives for clustered block storage:

- The all nodes are the same ("all nodes =") approach consists of nodes all doing the same work, both front-end I/O and back-end storage support. Examples of the "all nodes =" clustered block storage architecture include Sun's Thumper, LeftHand Networks' SAN/iQ, and Pivot3's RAIGE.
- The front-end and back-end nodes are different ("FE <> BE nodes") approach consists of specialized front-end nodes that attach to the host servicing iSCSI and/or Fibre Channel protocols and back-end nodes supporting data/storage services. Examples of the "FE <> BE nodes" approach include XIV/IBM's Nextra, Intransa's StorStac, and CleverSafe's DsNET.

- The approach that involves connecting four or more controllers together ("stacked controllers") consists of two or more dual-controller pairs stacked together to form a single storage subsystem. Examples of stacked controller configurations include DigiData's STORM, 3PAR's InServ, Fujitsu's Eternus 8000 model 2100, and the Dell EqualLogic PS Series subsystems.



In a typical "FE <> BE nodes" architecture, Interface Modules are front-end nodes and Data Modules are back-end nodes

All nodes = and FE <> BE nodes

The all nodes = and FE <> BE nodes are similar architectures and hereinafter are referred to as non-controller approaches. Both of these approaches support x86-class servers. The all nodes = approach also has direct-attached SATA storage bundled in each cluster node. In the FE <> BE nodes approach only the BE nodes contain storage, while the FE nodes typically only contain limited storage but have high-performance processors for better I/O servicing.

Under both approaches, I/O can be handled completely by the node that processes the incoming I/O but, usually, other nodes participate to service I/O. Also, most of these systems support dynamic load-balancing where I/O can be optimally handled by any I/O servicing node in the system.

The cluster interconnect for the non-controller approaches is based on Gigabit Ethernet, typically using a proprietary protocol to pass data/messages encapsulated within standard TCP/IP packets. Two Gigabit Ethernet ports are supplied with each node for cluster interconnection. However, XIV/IBM's FE <> BE nodes support 10GbE interconnect between racks of cluster nodes, providing additional bandwidth to support inter-rack communications.

Each node in an all node = configuration has two GbE ports for host I/O and two GbE ports for cluster interconnection. In contrast, each FE in the FE <> BE node approach has either two GbE or four Fibre Channel ports for host I/O and two GbE ports for cluster interconnection; each BE node only has two GbE ports for cluster interconnection.

All vendors with a non-controller approach base their software on variants of a Linux kernel with special storage application software that supplies block storage services.

Some vendors allow users to select the hardware from standard server vendors. The clustered storage vendors then supply the proper hardware drivers to support the storage system.

Most non-controller vendors quote performance in throughput numbers. Many claim almost wire speed for iSCSI per GbE connection (~120MBps), but this data must come out of cache. However, EqualLogic, a FE <> BE node vendor, claims disk-level performance, quoting ~150MBps across two GbE links.

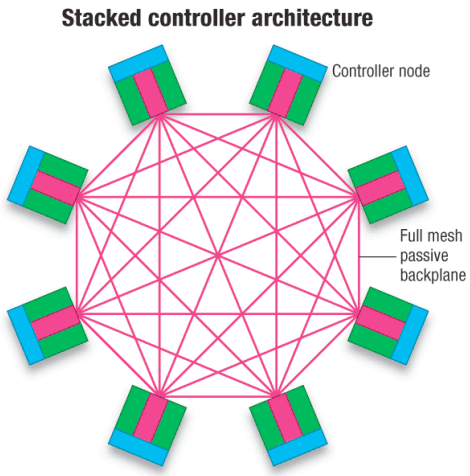
In an all nodes = configuration, both capacity and performance must be purchased at the same time and scaled together. Each server can have different amounts of storage according to the all nodes = vendors, but for redundancy reasons this is usually not the case. In contrast, the FE <> BE nodes approach allows users to easily scale front-end performance independently from capacity and back-end performance.

In non-controller approaches, all data must be mirrored across nodes, dividing effective storage capacity in half. How-ever, CleverSafe, a FE <> BE node vendor, has substantially reduced this redundancy overhead by assigning a factor to determine the number of BE servers that must survive to recover data. For example, in a four-FE and 12-BE configuration set to eight-BE node survival, the system could continue to reconstruct all data provided eight BE nodes survived, an overhead of only 50% of effective capacity.

Under the non-controller approaches, host multi-path software detects an I/O timeout and re-issues the I/O across another path. Block storage segments (e.g., LUNs) are then moved to one of the surviving nodes. However, Intransa, a FE <> BE node vendor, does not require host multi-path support for a node swap-out–a unique feature not offered by other clustered block storage vendors.

Stacked controllers

The 'stacked controllers' approach usually includes special-purpose controller hardware. (Exception: 3PAR's controllers are based on standard x86 servers.) Stacked controller products dedicate a portion of the cluster's storage pool to be accessed by each pair of controllers.



In a typical stacked controller architecture, a mesh backplane is the cluster interconnect

Many stacked controller products support dynamic front-end load balancing. Any single controller can service the I/O operation while a controller elsewhere in the cluster supplies the data for I/Os.

Stacked controller configurations have FE <> BE node-like scalability, allowing users to scale front-end performance independent from back-end performance. Stacked controller architectures support drive expansion frames connected via Fibre Channel or SAS/SATA for disk storage and allow the number of drive frames to also scale independently.

As for the cluster interconnect, stacked controller vendors use different interfaces. For example, EqualLogic's architecture looks like a non-controller approach and uses out-board GbE switches. 3PAR and Fujitsu use an inboard mesh backplane. And DigiData just uses Fibre Channel. Stacked controller products use either four GbE ports or eight Fibre Channel front-end ports with four cluster interconnect links per controller pair.

Most stacked controller vendors have their own proprietary kernel with specialized soft-ware to supply block storage services. However, 3PAR's product is based on a Linux kernel.

The performance of stacked controller products supposedly meets or exceeds monolithic and/or modular storage subsystem performance in throughput and I/Os per second. Stacked controller subsystems do not have the response time disadvantages of non-controller clustered storage.

In contrast to the non-controller approaches, stacked controller vendors provide standard RAID 5 as well as mirrored data redundancy. This allows users to select the redundancy level needed by the application rather than having a product-dictated redundancy level.

For stacked controller vendors, a single controller failure is dealt with much the same as a non-controller front-end node failure and depends on multi-path host software. Back-end failures for stackable controllers involve individual drive failures and are dealt with via normal RAID drive rebuilds.

Clustered block storage has been around for years in the iSCSI space and has also been available via stacked controllers for Fibre Channel storage. What's changed today is the non-controller approach to Fibre Channel storage symbolized most recently by XIV. In addition, some of the historic data redundancy limitations to non-controller clustered storage are beginning to be relaxed by CleverSafe and others. Together these two trends foretell the emergence of a new architectural alternative for both iSCSI and Fibre Chan-nel storage that will ultimately lead to much less expensive and more massively scalable storage.

About the author

Ray Lucchesi *is president of Silverton Consulting, a Storage, Strategy & Systems consulting services company, based in the USA offering products and services to the data storage community.*

mailto:info@silvertonconsulting.com

http://www.silvertonconsulting.com