Design Considerations for High Availability and Scalability in Blade Server Environments

What You Will Learn

Recent growth in both system consolidation and virtualization has increased the importance of high availability and scalability in data center and infrastructure design. Today, designing high availability and scalability into data center infrastructure and application deployments has become extremely important to the success of many businesses.

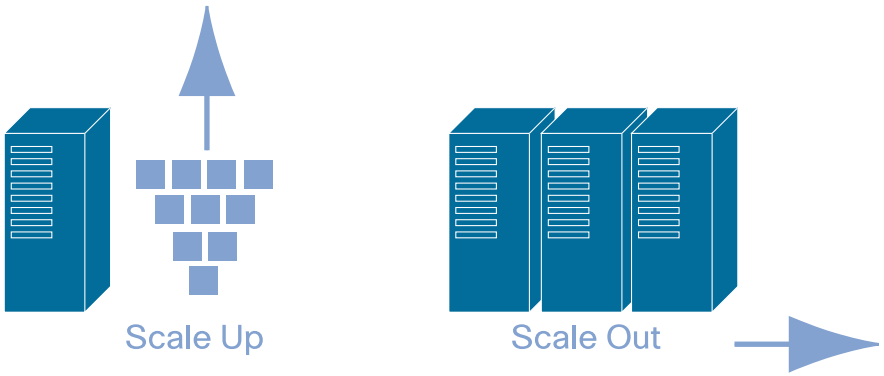
This document defines high availability and scalability and shows how implementing data center architectures with these attributes provides the application uptime needed to accomplish business processing goals. The document then shows how the Cisco® Unified Computing System platform provides an appropriate framework for environments that require high availability and scalability.

High Availability and Scalability

Compute environments are designed with high availability to help ensure a predictable degree of operational continuity during production hours. It is imperative that the uptime of business-critical applications not be compromised by unplanned equipment downtime. A corporate email service (such as Microsoft Exchange) is a common example of a business-critical application. If the system is designed properly, a single hardware failure in the compute environment should not affect the user's ability to continue using the service. Service may be degraded during the server downtime, but the application should remain operational while the hardware failure is being repaired.

In addition to keeping critical applications online, enterprises need assurance that the architecture is designed in such a way that system capacity can be increased or decreased as the needs of the business change. A scalable system is one whose performance increases (or decreases) after hardware is added or removed somewhat proportionally to the capacity added or removed. This system scalability can be achieved in several ways. In particular, resources can be added by scaling up or scaling out (Figure 1).

Figure 1. Dimensions of Scale



Vertical scaling (or scaling up) refers to the capability to add resources to a single server or node to achieve greater productivity (for example, adding more CPU or memory capacity to a single database server and achieving more transactions or queries per second than before the upgrade). Systems that provide this vertical scalability are often expensive and proprietary. In many cases, the operational costs of maintaining the system are even greater than the cost of purchasing the infrastructure. These unique and complex systems require specialized staff to support them.

Horizontal scaling (or scaling out) describes an architecture in which nodes or servers can be added to increase the capacity of an application environment. An example of this architecture is Oracle Real Application Cluster (RAC). In an Oracle RAC deployment, multiple nodes work in tandem to process user requests. By adding RAC nodes to a RAC cluster, overall capacity is increased and additional workload can be handled.

High Availability Compared to Fault Tolerance

It is important to understand the difference between fault tolerance and high availability. Fault-tolerant systems have self-healing properties, and the redundant hardware components inside each server are designed to withstand hardware failure without any downtime. Examples include mainframes or other enterprise UNIX or RISC platforms that are available from some of the larger server vendors. These platforms often have proprietary components that prevent server outages when critical hardware components fail. The processors and memory controllers in these systems are redundant and operate in a tightly coordinated way, with the redundant processor able to take over instantly in the event of a processor failure. Trade-offs with these systems include additional complexity and cost and the proprietary nature of the hardware and software used.

In contrast, high availability can be achieved through software that runs on top of an operating system, through the chosen server architecture, or even as a feature within the application itself. Applications that have high-availability properties built in eliminate the need to run additional software. High-availability systems are often designed to quickly recover from a component failure within the architecture, sometimes with a very limited amount of application downtime or with a period of degraded capacity, while the failed components are quickly resolved.

High-Availability Architecture and Data Access

At the core of any enterprise data center is some amount of critical data. The servers, operating systems, and infrastructure components are in place to facilitate user access to this data.

Many methods are available to build high availability into the infrastructure. In this paper we will start at the data itself and move “out” towards the remainder of the datacenter infrastructure.

Enterprise data is often stored on either block-level storage arrays (SANs, or Storage Area Networks), or file-based arrays (also known as NAS or network-attached storage arrays). These enterprise arrays provide various levels of redundancy, so that if a single component in the array (whether a disk drive, controller, SAN or LAN access port, etc.) should fail, the data will still be accessible to the user.

The data itself will often be striped or mirrored (or both) across many physical disks inside the array, to allow continuous access to user data in the event of a physical disk drive failure. These redundancy levels (RAID levels) can vary depending on the performance needs of the applications.

To create highly available or redundant paths to this storage (and to allow multiple hosts to access these paths simultaneously), the arrays will often be connected to switches. SAN arrays typically use at least two Fibre Channel switches (Ethernet switches for NAS), so that if a switch or path fails, the surviving switches or paths can service user requests for storage.

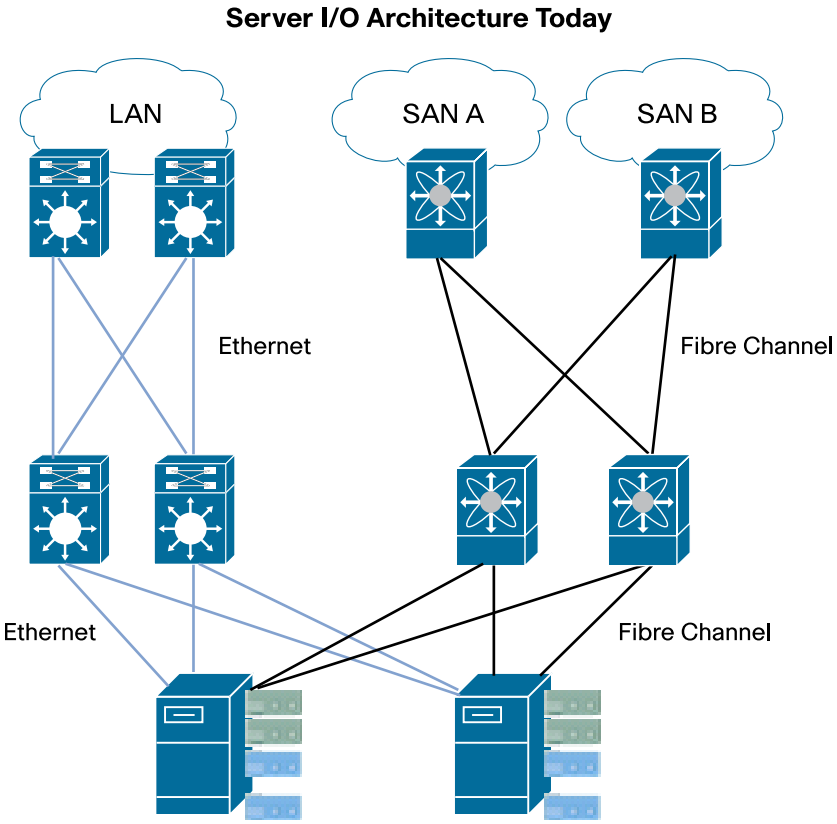
Server I/O High-Availability Architecture

Just as important as the arrays that hold the data are the servers that run the applications that ultimately access the data. In order to achieve high availability and redundancy with these servers, many architecture choices are available.

Enterprise servers based on x86 architecture are often configured with multiple connections to SANs and Ethernet networks. These redundant connections can be accomplished using dual port adapters or by using multiple (single-port) adapters, all of which depend on the space available in the server (that is, the number of I/O expansion slots) and the level of hardware redundancy required.

The servers are then tied to the infrastructure in a way that uses redundant paths, switches, and access ports for network and storage resources. This topology is sometimes referred to as a bow tie (Figure 2).

Figure 2. High Availability Achieved with Redundant Infrastructure and Paths



Notice that the connections are redundant and do not permit a single point of failure. If a single Ethernet or SAN path fails, there is connectivity across a surviving path. An important note is that in most cases this failover does not happen automatically and needs to be triggered by some type of multipath or failover software.

For network redundancy, this software is often available within an enterprise operating system. Many Microsoft Windows and many Linux (server) operating systems come with a bonding or network-interface-card (NIC) teaming driver. Such a driver creates an association between two network cards, allowing them to work in tandem. The driver manages load balancing of network traffic across the two links, while also enforcing failover should a physical network card, cable, or upstream network component fail.

SAN failover is often accomplished using third-party multipath software provided by the disk array vendor. Multipath software is loaded on top of an OS and provides a layer that sits between the OS and storage adapters. The multipath software closely monitors all physical paths and storage presentations while load-balancing SAN traffic across the available host bus adapters (HBAs) and paths. This software layer also accomplishes failover should an HBA, path, or upstream device fail.

Among the downsides of having so many adapters, network ports, storage ports, and cables are the power consumption and costs associated with fiber-optic transceivers and cabling. Another concern is the complexity of managing all these connections and cabling. These costs can add up significantly in a large environment with many servers. To help address these costs, the industry is adopting converged networking technology and running SAN and Ethernet networks over a single cable type. Fibre Channel over Ethernet (FCoE) is an example of this type of implementation. Using FCoE-capable adapters in a server, rather than multiple Ethernet and storage adapters, enables servers to pass both Ethernet and Fibre Channel traffic over the same media type. This approach

significantly reduces costs and management complexity, while providing redundant access to SANs and Ethernet networks.

N+1 Server Architecture: Hardware High Availability

More advanced platforms provide N+1 hardware failover. An N+1 server environment has a single spare server that is configured to provide redundancy for a group of servers. This N+1 server assumes the role of any failed server within that group when required.

N+1 failover is common for power supplies. An example is the installation of an extra power supply in a compute chassis. This power supply turns on only when it is needed to address the failure of an active power supply. This behavior provides redundancy but also saves power and cooling compared to an extra power supply that is always turned on in the compute chassis.

N+1 hardware failover (accomplished below the OS and without special applications) is much more difficult to achieve because of all the server components that make up the identity of a server. These components include:

- Storage worldwide name (WWN) addresses (which are normally fixed to a storage adapter)
- Network MAC addresses (which are normally fixed to a network adapter)
- Firmware and settings for the storage and network adapters.
- Firmware and settings for the server BIOS, including the universally unique ID (UUID)

Many available server platforms can virtualize the MAC address and WWN components of a server, but the real fulfillment of server N+1 failover lies in the abstraction of the server hardware itself, not just the identity of fixed hardware assets. Migration of the firmware components and settings to a defined N+1 failover server provides a replica of the failed server and dramatically simplifies the failover process. By extending the virtualization to the adapters themselves, a much more functional and scalable N+1 can be achieved. One complication that will need to be managed in this fluid and dynamic N+1 scenario is the association of network and storage edge properties. Maintaining the association of these edge ports and any applied policies is critical to a secure and effective failover event.

High-Availability and Scalability Software and Applications: Software High Availability

To provide additional levels of redundancy and increase overall application uptime within a local site, high-availability applications can be run on servers architected in a highly available configuration (redundant network adapters, SAN adapters, paths, switches, etc.).

These applications are often installed across two identically configured servers, and the clustering services running on the servers stay in close communication with each other. A number of clustering services are available, some built into the OS or installed on top of the OS. These applications operate in two modes:

- Active-active: Two configured servers are configured to work together to service requests and the overall application load. In the event of an application, OS, or hardware failure, the surviving node services all requests. To accomplish failover, it is important to size both servers accordingly, so that in normal operation, each server (on average) has less than 50 percent utilization. This scenario requires the use of a minimum of two physical servers, each running active licensing of both the OS and applications.
- Active-passive: This mode is similar to active-active, but the passive node is in a standby configuration. A passive node will not actively process requests until a failure condition is detected in the hardware, OS, or application of the active node. In this configuration, a passive node does not typically require an active license for the OS or the applications. The active licenses apply only to the active node in the cluster. Other considerations for the active-passive mode include the fact that the passive node will be booted and use

power and cooling resources and will require operational resources to stay up-to-date and synchronized with the primary node.

To extend protection beyond the local site, disaster recovery process, policies, and procedures are often used. This protection is set up to prepare for recovery after a natural or human-induced disaster. IT organizations often spend a percentage of their yearly budgets (3 to 6 percent) on disaster recovery in order to avoid larger losses that can occur if a disaster stops business operation. Flexible server architectures allow organizations to reduce the time needed to achieve disaster recovery objectives.

Cisco Unified Computing System: A Scalable and Highly Available Compute Platform

Cisco Unified Computing System is an integrated computing solution with an architecture that provides a deep foundation for running applications that require high availability and scalability. This platform is an x86-compatible server platform that incorporates embedded and holistic management, a unified fabric for I/O and management, and optimizations for virtualization throughout the solution.

Cisco UCS 6140XP 40-Port Fabric Interconnect

A pair of Cisco UCS 6140XP 40-Port Fabric Interconnects (Figure 3) are at the center of the Cisco Unified Computing System. The fabric interconnects provide an integrated access layer for the many chassis of server blades that can be connected. The fabric interconnects also provide a single point of connectivity to storage networks, Ethernet networks, and management networks. Each fabric interconnect contains a highly available boot process. Redundant boot blocks within the hardware hold the active boot software and a backup version. If the fabric interconnect fails to boot its active software version, it will boot on the backup version.

Figure 3. Cisco UCS 6140XP 40-Port Fabric Interconnect

These main solution components contain the Cisco UCS Manager software, which provides a single point for complete server management, including provisioning of servers and I/O connectivity through the use of service profiles. These service profiles are server definitions, including I/O and hardware profiles, with the goal of completely abstracting components of a server that would normally require independent management, including management of low-level firmware settings and identifiers associated with a server BIOS, storage HBAs, and network adapters. These service profiles also include edge-port associations for SAN and network connectivity, so that all applied security and policies remain intact if the service profile moves to a new blade.

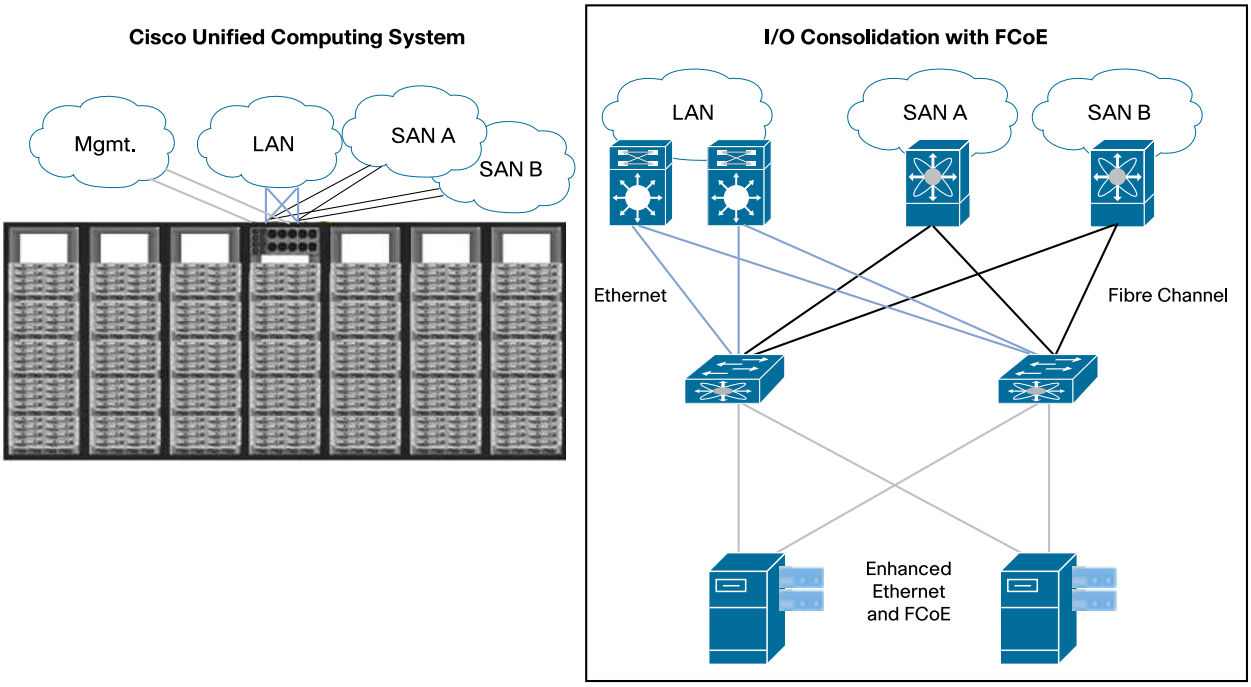
All cabling in the solution is based on Cisco unified fabric or 10 Gigabit Ethernet using FCoE. Unified fabric cabling can be copper or fiber optic (depending on distance requirements), and Ethernet, Fibre Channel, and management traffic can be carried over the same cabling.

The fabric interconnects provide up to 40 FCoE-enabled ports connecting into the compute chassis containing the server blades. This domain of server management scales out to a total of 320 servers, all with a single point of management for all aspects of server and I/O provisioning and monitoring.

To provide high availability for system components, the fabric interconnects can be installed in pairs, providing redundant I/O fabrics for the connected servers while also providing redundant access to premise Ethernet, SAN, and management networks. Each fabric interconnect also contains dedicated network ports for clustering directly

between the two highly available fabric interconnects. The Cisco UCS Manager software that runs inside the fabric interconnects contains extensive logic for managing heartbeat, problem detection and remediation, and replication of configuration information (Figure 4).

Figure 4. I/O Consolidation with FCoE



Cisco UCS 2104XP Fabric Extender

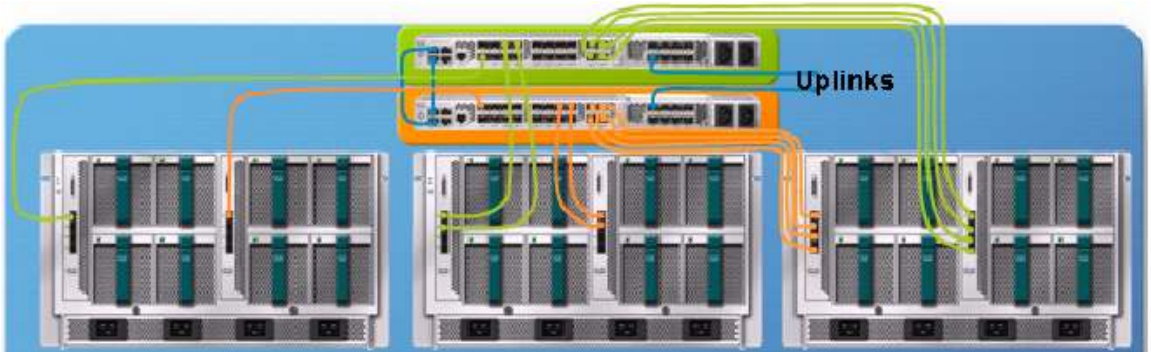
Each Cisco UCS 5100 Series Blade Server Chassis contains a pair of redundant Cisco UCS 2104XP Fabric Extender modules (Figure 5). These modules provide consolidated I/O connectivity (network and Fibre Channel through FCoE) to the fabric interconnects, and provide access to isolated I/O fabrics for all blades within a blade chassis. Like the fabric interconnects, the fabric extenders have a highly available boot process. Redundant boot blocks allow each fabric extender to hold an active and a backup firmware image. If the fabric extender cannot boot its active firmware image, it will boot the backup version.

Figure 5. Cisco UCS 2104XP Fabric Extender

Connectivity from these fabric extenders to the fabric interconnects is typically made with low-latency, copper Twinax cabling running at 10-gigabit speeds. This cabling provides unification of all I/O, replacing the bundle of redundant SAN, network, and management cabling often used in traditional enterprise servers: running to the rack patch panels, in trays under the data center floors, to network and SAN row patch panels, and into the appropriate network and SAN access ports.

This fabric extender architecture enables simplified I/O scaling as shown in Figure 6. I/O can be scaled out within a blade chassis simply by adding connections from the chassis fabric extenders to the fabric interconnects. This approach allows the compute environment to easily scale for additional compute capacity (with fewer I/O cables) or for additional I/O capacity (with more I/O cables). Most significantly, I/O can be scaled without the need to make physical changes to mezzanine cards or switches found in many blade server environments.

Figure 6. In-Chassis I/O Simplification



Cisco UCS 5108 Blade Server Chassis

The Cisco UCS 5108 Blade Server Chassis (Figure 7) can be installed without affecting other system components, providing the capability to add chassis quickly and easily at any time. The blade chassis holds up to eight blade servers, up to two fabric extenders, and up to four power supplies. These power supplies can be configured to support non-redundant, N+1, N+N, or grid-redundant configurations. Each chassis has been streamlined and does not require any SAN, LAN, or management modules as are normally seen in a typical blade chassis. These functions are either embedded or used from within the Cisco UCS 6140XP fabric interconnects. Therefore, adding blade chassis will not create more points of management, as all device management remains centralized. When a new chassis of blade servers is installed, all the components are quickly discovered and made available for device management within Cisco UCS Manager.

Figure 7. Cisco UCS 5108 Blade Server Chassis

Cisco UCS B-Series Blade Servers

The form factor of the Cisco B-Series Blade Servers enables simple scaling out of additional servers within a chassis (Figure 8). After a new blade is inserted into a chassis, the blade is quickly discovered and made available for centralized device management within Cisco UCS Manager. The onboard service processor (baseboard management controller [BMC]) is a critical component for all management and discovery. Like the fabric

interconnects and fabric extenders, all blades have a highly available boot process. If a blade's BMC or BIOS fails to boot the active firmware version, the backup version will be used.

Figure 8. Cisco UCS B-Series Blade Servers

Cisco UCS Network Adapters

To complete the highly available architecture of the Cisco Unified Computing System, each server uses a dual-port 10-gigabit network adapter with redundant, unified I/O connectivity to each fabric extender (Figure 9). This redundant fabric connectivity extends all the way to the top of the solution, where connection to the external management, SAN, and network is made. This design is crucial to helping ensure that a single failure of a component within one I/O fabric does not affect server I/O within the solution.

Figure 9. Cisco UCS Network Adapters

These adapters can be replaced or upgraded in the field, and like many components of the Cisco Unified Computing System, the network adapters are designed with a highly available boot process. The components store both active and backup firmware versions, helping ensure that if an adapter cannot initiate its active firmware version, the backup version will be used.

- The Cisco UCS 82598KR-CI 10 Gigabit Ethernet Adapter provides redundant 10 Gigabit Ethernet (Intel 82598) connectivity to a blade server.
- The Cisco UCS M71KR-Q QLogic and M71KR-E Emulex Converged Network Adapters (CNAs) support both Fibre Channel and Ethernet using FCoE. These adapters provide redundant 10 Gigabit Ethernet (Intel 82598) and Fibre Channel connectivity based on either QLogic or Emulex chip sets.
- The Cisco UCS M81KR Virtual Interface Card (VIC) can instantiate and scale up to 128 PCIe-compliant Ethernet and Fibre Channel devices for use by the host OS or hypervisor. Through complete management by Cisco UCS Manager, all I/O devices configured in a service profile are automatically programmed onto the adapter when the service profile is applied to the blade. This network adapter also can present highly available virtual Ethernet interfaces to the OS running on a server. These virtual interfaces offer below-the-OS failover and, in the event of a single component or path failure, can be automatically failed over to the surviving I/O fabric in the system. This method of failover removes the complexity of using multipath or bonding drivers running within the OS.

Conclusion

Today's data center compute environments require a highly available and scalable architecture to support the mission-critical applications of the enterprise.

The Cisco Unified Computing System meets these needs by providing an x86-compatible compute solution with a simplified, scalable architecture designed with high-availability components throughout the solution, starting with the server and extending to both storage and Ethernet networks.

For More Information

For more information about Cisco Unified Computing, visit http://www.cisco.com/en/US/products/ps10265/index.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV
Amsterdam, The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.

Printed in USA C11-553711-00 08/09