

## TELECOM

## Blade-based storage subsystems

By Alan Fitzgerald

When designing a computing platform for a specific application the key design factors always include:

- CPU power
- Number and type of I/O channels
- Form-factor
- Electrical power
- Operating system

The storage requirements are usually addressed last and then viewed primarily as a capacity issue.

Storage has frequently been dismissed as “only a disk drive,” receiving little additional investigation until overall system reliability and performance are scrutinized more carefully. In fail-safe operations, a commodity disk drive (purchased at your local electronics superstore) does not meet the stringent field life and reliability requirements. Establishing the application’s storage requirements includes considering the capacity, reliability, system interface and operating system compatibility, and field life.

This article discusses various direct-attached and network-attached storage systems and the desirable characteristics that should be used to define the right solution for your application requirements.

### Storage for blade-based systems

There are two commonly accepted categories of storage for a system: direct attached and network attached. Within each category exist a range of interfaces and system characteristics that must match the available hardware, be compatible with the operating system, meet the performance demands, and match the access method of the software application.

The choice of solution based on the tradeoff of performance and capacity between an internal and external solution is dictated by the overall system demands and should be apparent when considered early in the system design stages. Figure 1 illustrates the relative position of an internal chassis-mounted disk subsystem (top), an external rack-mounted disk array (middle), and a blade-based subsystem (bottom).

### Direct attached storage (DAS)

DAS storage devices form an integral part of a CPU node and provide storage services for booting the operating system, loading the application software; and in small systems it serves as the data storage device. DAS represents the simplest and lowest cost blade storage implementation.

The DAS on a CPU node may be shared by other CPUs only if the mated CPU is configured through software to allow other systems access to the DAS disk. Sharing disk drives consumes

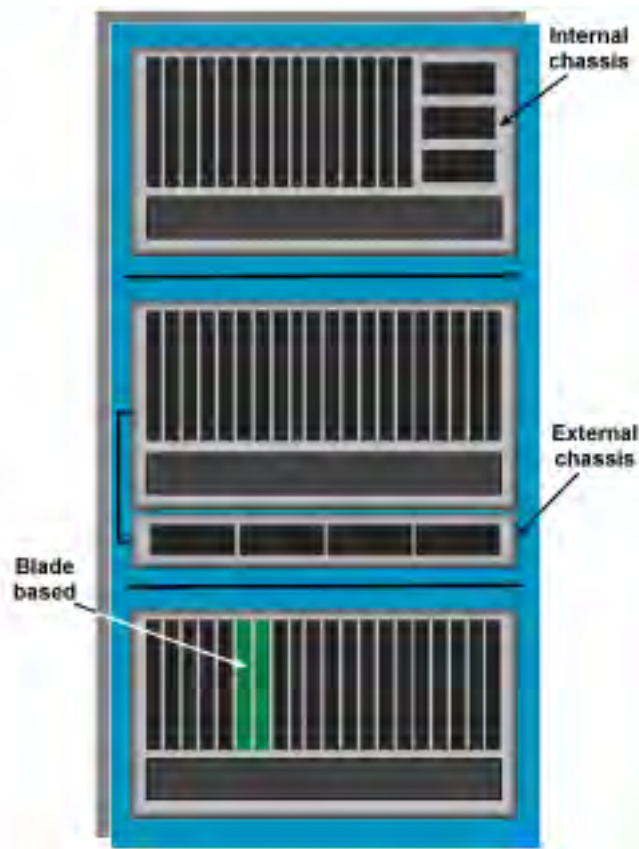


Figure 1

CPU bandwidth and may detract from the intended mission of the CPU node.

Locating the DAS device may be done in a variety of ways.

### CPU-mounted IDE disk drive

The simplest direct attached method is a 2.5-inch IDE disk drive mounted onto the CPU board. The disk is an integral component of the CPU blade and acts as the boot drive, application source, and frequently the data disk. In this configuration the reliability of the node is reduced to the life of the disk.

Using an IDE disk on the CPU board may also extend the form factor into two slots, creating additional cooling problems. High-speed processors already generate significant heat. A failure on the CPU board may not be immediately traceable to the mounted disk drive and the entire CPU blade must be replaced after taking the node offline. This increases the inventory costs by having to replace the most reliable and most expensive system component, the CPU blade, to fix a problem with the least reliable and frequently least expensive component, the disk drive.

The failover process for a CPU-based disk is to transfer processing to a full standby backup with real-time or near real-time data on the standby CPU hard disk. While the CPU-based disk drive offers the lowest cost storage, it requires significant effort and investment for replacement and is the least reliable alternative.

### Disk bays in the chassis

Hard disks in chassis drive bays function in a similar manner to a CPU-mounted disk. With simpler access for replacement, a disk bay still requires its CPU mate to be offline prior to replacement. Failover occurs to a standby CPU and disk bay.

Chassis mounted drive bays consume much needed chassis space and require loose cables to be routed to the bay for signal and power. Selection of the location of the disk bay in the chassis must consider heat dissipation to supply proper cooling.

### Mass storage module (MSM)

A mass storage module performs similar functions to the chassis-mounted drive bays by keeping the disks off of the CPU blade and inside the chassis. An advantage of an MSM is the elimination of some cabling. An MSM blade installs in a chassis slot with storage management handled by the CPU.

Typically, the disks used on an MSM are 3.5-inch SCSI drives that consume two slots. Cabling comes through a rear transition module from a host or host-mounted PCI mezzanine connector (PMC). The host may provide mirroring software and may allow disk hot swapping. Hot swapping a failed disk in a mirrored disk array allows the CPU to continue to operate without causing a failover to a standby CPU node.

In this configuration, the CPU manages the mirroring, failure processing, and data rebuild after the failed disk is replaced. The bandwidth consumed by these tasks may be more effectively used for the application and generally causes lower performance and unexpected losses of system bandwidth, which is usually unacceptable.

### External rack-mounted disk array

An external RAID chassis provides a number of advantages in reliability, serviceability, capacity, and performance. An external rack-mounted disk array manages RAID and fits in the space of a 19-inch wide 2U, 3U, or 4U chassis. The disk array provides self-contained cooling, redundant power supplies, and host interface. A failed drive can be replaced without taking down the CPU node and the disk array's internal controllers manage the RAID functions.

This solution provides high capacity and performance, but at a significantly higher price than a CPU-mounted, internal chassis, or MSM blade storage devices. The benefits include hot swappable disks, continued operation with a single disk failure, and automatic rebuild with disk replacement. Price is the most significant drawback to this approach.

When required capacity is 60 Gbytes or less, the cost and performance of a disk array can be difficult to justify. Add in the cost of certification compliance, such as NEBs, and rack-mounted disk arrays become a significant portion of the overall solution price.

RAID disk arrays support DAS, NAS, and SAN architectures, which provides flexibility in the attachment method. The ability to connect to the system fabric provides additional benefits for high-availability systems by sharing storage resources and reassignment of storage resources when a node fails.

### Blade-based storage subsystem

A blade-based storage subsystem integrates disk drives and the RAID controller technology onto a single slot blade. The application of RAID controller technology to a blade blends the advantages of an in-chassis MSM with the redundant reliability of RAID 1 (mirrored redundancy) separation of the storage management from the CPU, and the onboard support for hot-swappable disks.

As shown in Figure 2, a blade-based storage subsystem (CompactPCI 2.x 6U) may also connect to the PCI bus in a CompactPCI chassis and provide the host controller, eliminating the expense of a PMC adapter. Even though the direction of blade-based systems, especially in telecommunications, is away from high-speed parallel buses, today most systems incorporate the PCI bus and will continue to support PCI connectivity for the foreseeable future.

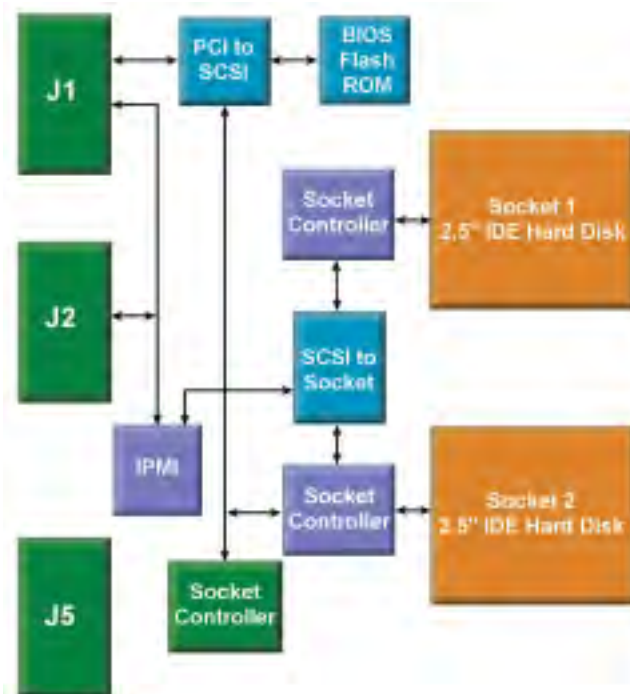


Figure 2

Like an external rack-mounted disk array, the RAID management resides in the storage subsystem. A single disk failure does not interfere with the CPU node. Hot-swappable disk drives provide non-stop operation and automatic rebuilding of data without system intervention. Agency certification is simplified by containing all functions on a blade within the chassis as standard system components.

### Fabric attached (NAS or SAN)

While direct attached storage ties the functionality of a CPU node to the operation of the storage device, either the single disk or RAID-based disk, attachment of the storage subsystem to the system fabric enables another level of high-availability management and potential savings when used as a shared resource. Current architectures split storage between block devices (SAN) and file structured storage (NAS) subsystems.

There are a number of choices for fabric standards for the system designer. However, only a few choices are available with hardware and software support for storage. The PICMG 2.16 specification defines a CompactPCI backplane connection that supports Ethernet, which is becoming widely available from many chassis and backplane vendors.

This advance in backplane connectivity allows the use of existing storage management software and simpler blade-level hot swapping. Another attribute for storage management is the PICMG 2.9 IPMI standard for communication of field replaceable unit (FRU) and health status to a network controller. These features enhance the system designer's ability to achieve high-availability goals.

Network connection to an external rack-mounted storage provides many of the same features of shared resources and simplified management. Again, as stated for DAS connection, the external subsystem achieves higher capacity and performance than the blade-based subsystem, with the obvious additional expense, large footprint, and operating environment concerns.

Ethernet provides the simplicity afforded by years of field experience in LAN and IP protocol and is found in every modern operating system. Development to make high-availability extensions to storage in this environment promises to make the combination of an Ethernet-connected IP network storage subsystem very attractive for a wide range of telecommunications applications.

Storage subsystems designed with redundant network ports, blade hot swapping, disk hot swapping, and 1 Gbyte/sec performance allows the system designer to implement a high-availability strategy for management of the system storage. These features coupled with IPMI for network resource health monitoring brings the storage subsystem into compliance with the HA demands, and to the level of HA performance found in many network I/O blades.

Functional operation of the network subsystem operations (NAS or SAN), becomes the choice of the system and application designer. Either choice achieves the benefits of network attachment.

### Current blade-based implementations

Current PICMG 2.x CompactPCI chassis, such as the Force

Centellis CO 21000 pictured in Figure 3, provide adequate space to incorporate two 2.5-inch hard disk drives on a single slot 6U blade. The front panel size is adequate to allow openings large enough for the drives to be removed without unplugging the blade. Figure 3 also shows the Adtron SC6M with a RAID 1 SCSI controller and two hot-swappable disk drives.

Disk controllers available for blade storage provide a flexible host controller interface for connection to standard interfaces including IDE, SCSI, Ethernet, and USB. On the disk side the controller manages the disk array as RAID 0 (interleaved with 2x disk capacity), RAID 1, and just-a-bunch-of-disks (JBOD independent logical disks). Depending on the system requirements, the system designer selects the best-fit configuration.

When configured for RAID 1, the disk controller manages the power and signal controls of each socket allowing removal, replacement and rebuilding of a disk without system intervention. The disk controller communicates with the onboard IPMI controller to report blade health to the chassis alarm card.

In the event of a disk failure, the storage blade continues to operate at full performance using the operational disk. An IPMI message informs the chassis alarm card of a disk failure. After network notification, a field service technician arrives to install a new disk through the front panel. During the exchange of the failed disk with a new disk, the system continues to operate with full performance. Once the failed disk is replaced the disk controller automatically completes a full data rebuild and restores the mirror without CPU intervention.

### Conclusion

Scaling the physical size of a storage subsystem to fit within a blade-based chassis provides significant cost and integration advantages over other internal and external storage subsystems. Blade-

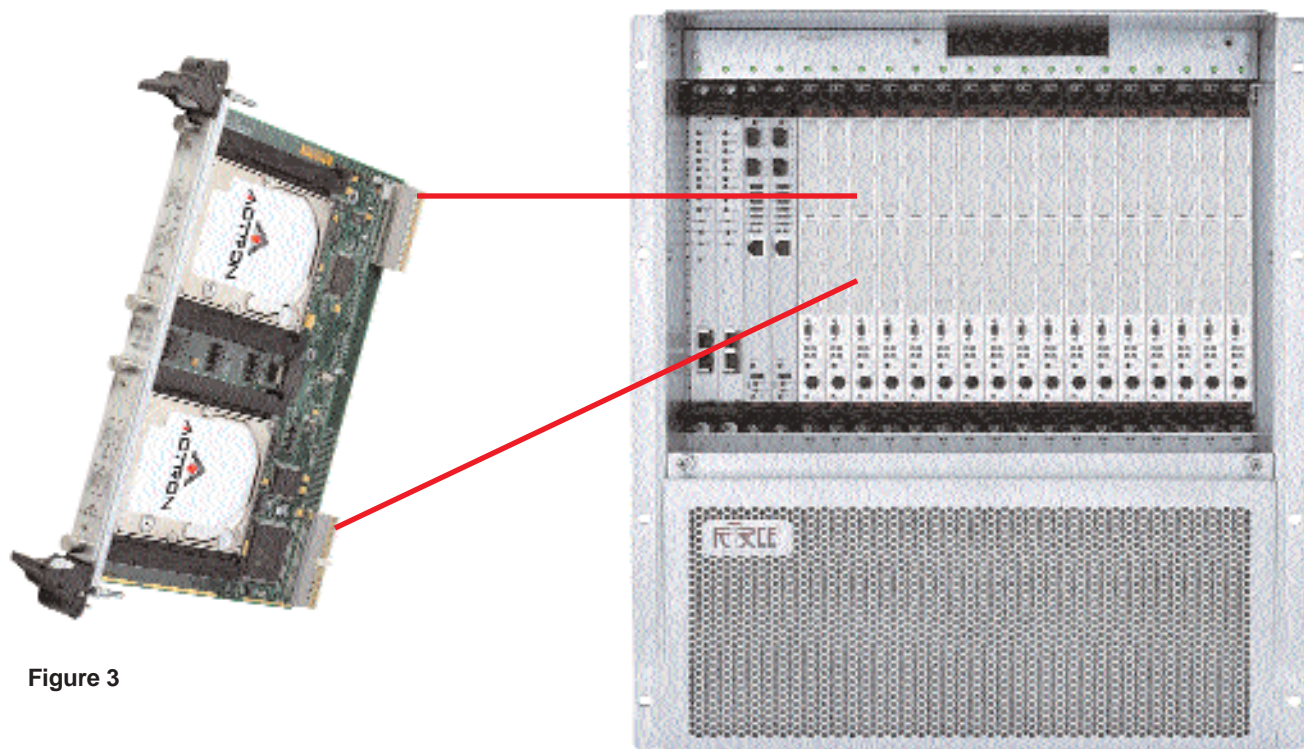


Figure 3

based storage subsystems using onboard storage management processing enhance the overall system reliability by providing disk mirroring, disk hot swapping, and IPMI communications to an alarm card. State-of-the-art blade storage technology incorporates storage management intelligence on the blade and offers significant system level advantages over simple mass storage modules (MSM), and material cost advantages over external storage subsystems.

Figure 4 illustrates the relative cost of CPU-mounted or MSM, blade-based, and external storage chassis. The cost-benefit of the blade-based storage subsystem in relation to the added reliability is significant when compared with an external storage subsystem.

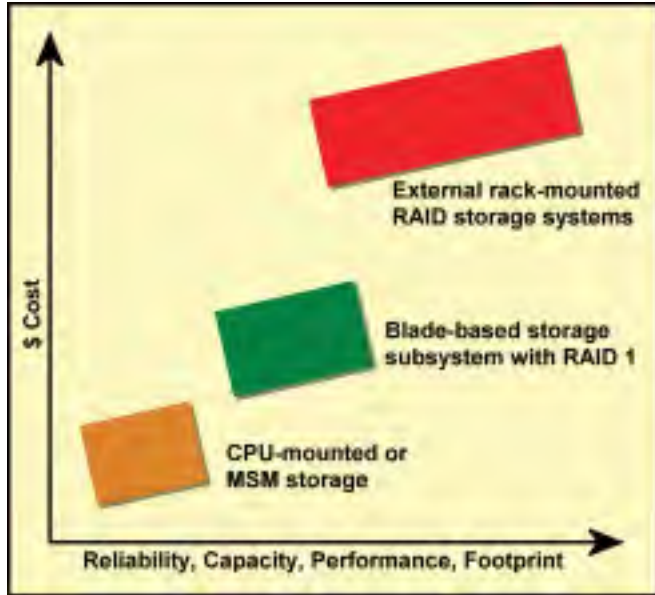


Figure 4

The next step to integration in a fabric-based system is to connect the blade storage through the system network. This further enhances the system integrity and fulfills the high-availability strategy. As blade-based storage subsystems evolve, the current storage attachment architecture will move away from the direct attached storage (DAS) to network attached storage (NAS and SAN) architectures with all the subsequent benefits of HA achievement.



*Alan Fitzgerald, the founder of Adtron, is the President and Chief Technology Officer. Alan is an electrical engineer by training, with an MSE from Arizona State University. His background includes engineering positions with Motorola, Hamilton Test Systems, and a position as Chief Engineer for Audiometer Systems. Alan is responsible for leadership of the technical, product strategy, and development aspects of Adtron's operations.*

For more information, contact Alan at:

Alan Fitzgerald  
**ADTRON**  
3710 E. University Drive, Suite 5  
Phoenix, AZ 85034  
Tel: 602-735-0300 • Fax: 602-735-0359  
E-mail: [afitzgerald@adtron.com](mailto:afitzgerald@adtron.com)  
Web site: [www.adtron.com](http://www.adtron.com)

