IP Storage Area Networks

Storage for the Data Center of the Future

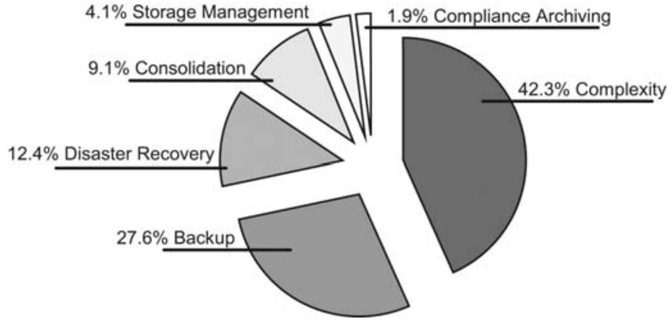
by Ellen Lary and Richard Lary
TuteLary, LLC

The Data Center of the Future is envisioned as a computing utility that will use all of the server, storage and networking elements of the data center as a dynamic set of computing resources that can be flexibly allocated to customer applications.

What is the Data Center of the Future?

The Data Center of the Future is envisioned as a computing utility that will use all of the server, storage and networking elements of the data center as a dynamic set of computing resources that can be flexibly allocated to customer applications. The ability to dynamically allocate some of these resource types, individually, in an uncoordinated fashion, already exists today. For example, the deployment of storage area networks (SANs) makes it possible to create pools of virtualized storage that can be assigned as needed to an application running on a server. Network services like load balancing, VLANs, and VPNs can also be virtualized to allow them to be dynamically assigned to applications. Server virtualization technology has recently become available from several server vendors as the last of the enabling technologies for the Data Center of the Future.

Figure 1
IT Executive Server and Storage Survey



The IT industry is devoting a lot of engineering (and marketing) talent to this concept, which has also been referred to as the computing utility, grid computing, on-demand computing, the consolidated computing center, and data center virtualization. Whatever name you choose to call it, its defining feature is the automated dynamic assignment of server, network, and storage resources to business applications.

The key management component that will be needed to fully realize the Data Center of the Future is a Data Center Resource Manager (DCRM), which performs several important policy-based functions, utilizing the various system management utilities available in the data center as its eyes and hands. The DCRM maintains knowledge of what resources are available to allocate to applications, and schedules an application for execution based on priority and the availability of appropriate resources. It then allocates available resources to the application and configures those resources to host the application. It binds the application to its required data objects – logical units or file systems – and configures the data center security mechanisms to allow access to the application's storage from and only from those servers running the application. The DCRM then starts the application and monitors the execution of the application to detect aborts or performance problems due to the failure of data center hardware components, in which case it reallocates resources so as to keep the highest-priority applications running, even if that means aborting the execution of lowest-priority applications.

Why is the Data Center of the Future Important?

The data center environment today is more complex than ever before. It is composed of hundreds of disparate components, each of which uses a different management system. Servers and storage are underutilized because dynamically allocating resources manually is just too difficult and error prone. Getting this complexity under control is critical. A January 2003 CIO Insight survey showed that 54% of IT executives polled said their systems were more complex than needed. The resultant cost of maintaining and managing that complexity consumes an average of 29% of their IT budget.

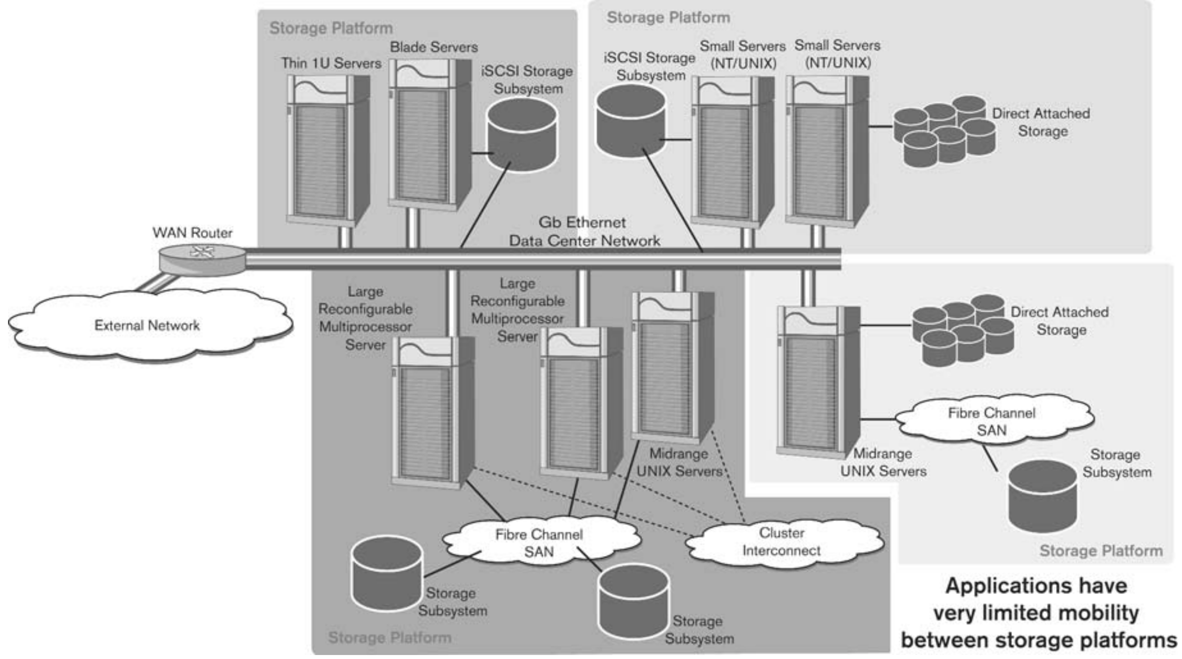
(See “Migrating to a Virtual Environment”, Inkra 2002.) A 2004 survey done by Strategic Research (Figure 1) shows that the number one problem that IT leaders are spending money to solve is complexity (42.3%), followed by backup (27.6%), disaster recovery (12.4%), consolidation (9.1%), storage management (4.1%) and compliance archiving (1.9%). Complexity is rated so high because it is a primary driver of the cost of managing the data center, and it increases the likelihood of costly management errors

Complexity is fueled by the many different computing models in the data center – simple servers, large configurable multiprocessor servers, clusters, and aggregated systems (blades or racks of thin servers) – as well as the diversity of equipment manufacturers in the server, storage and network areas. Figure 2 depicts a representation of the data center of today. The many different servers found in the data center can communicate with each other across the data center via Gigabit Ethernet, but each of them has storage that either is local to the server or on a “private” secondary network (i.e. a Fibre Channel SAN) that is not available to all other servers.

Applications have dependencies on the type of CPU and the amount of memory in their host server as well as the type and revision level of their host operating system. This has resulted in “locking down” applications to specific servers, which presents challenges if an application’s needs grow beyond its server’s capabilities or if its server becomes unavailable.

When a server running an application becomes unavailable, or a more powerful server becomes necessary, the application must be moved. This is a difficult and potentially dangerous operation. It is not easy to assure that all of the resources (processes, files, network connections, system resources, storage…) affiliated with the application are available to the new server and are correctly

Figure 2
Data Center of Today: Islands of Storage Limit Mobility of Computing Resources



Applications have very limited mobility between storage platforms

configured on that server. Today, this is a hands-on process that is done via managing the underlying infrastructure and hoping that the application comes along with it. It is definitely not quick, which is most painful after a sudden outage, as the application is unavailable during this time. Depending on what that application is, there is great potential for loss of revenue.

Given the dynamic environments that exist in most businesses today, moving an application is a frequent event. Couple this with the dependency of the application on the physical infrastructure as well as the lack of tools to manage the application in an integrated manner, and the result is a significant pain point for the IT manager.

Another significant pain point for the IT manager is the ongoing need to add more storage to support an application. For most businesses, growth in storage is a given. Today, depending on the specific storage product being used, the operation of adding storage can be very complex. In general, SANs make it easier to share capacity among many servers. However, the problem of assigning specific LUNs to applications and then expanding or reassigning the LUNs when more storage is needed is still hands-on and subject to errors. There are some storage products available today, the PS Series from EqualLogic is an example, that simplify these operations.

The management tools that exist today to manage storage and servers all require specialized training, and require system managers to keep high-level management policies firmly in mind at all times while slogging through low-level management functions. Because the knowledge of how to manage the data center is fragmented across network managers, storage managers, server managers, database managers, etc., data center management often suffers from poor communications between management fiefdoms. Meanwhile, IT budgets are being cut and there is not enough money to hire all the trained people required to keep a complex data center running efficiently. As a result shortcuts are taken, resulting in lowered resource utilization and increased capital and maintenance costs.

The Data Center of the Future will allow resources to be pooled – instead of managing single storage devices or single servers, a single pool of virtualized storage and multiple pools of compute resources will be available to the DCRM, which will select subsets of these pools to host an application. This pooling increases the utilization of data center resources, and also reduces the administrative effort needed to manage ongoing operations and add new applications, thus reducing both capital and operating expenses.

The Data Center of the Future will also increase application availability. Application scheduling will take failed components into account. Clustered applications will not only survive server failures, but will have their performance and redundancy restored through the automatic allocation of a replacement server to the cluster. Applications that run on a single server will still suffer an outage if that server fails, but that outage will be much shorter because the application will automatically be restarted ASAP on a replacement server. Critical applications will get priority access to data center resources, pre-empting low priority applications if necessary.

Requirements for the Data Center of the Future

Implementation of the Data Center Resource Manager (DCRM) faces several difficulties. Foremost among them is the lack of standard interfaces to the "eyes and hands" – the existing system management tools that monitor and configure data center components. Several standards bodies are cooperatively addressing this problem, and vendor-independent models exist today to describe data center components and interconnects (Oasis/DCML), to manage servers and networks (DMTF/WEBM), and to manage storage subsystems (SNIA/SMI). All of these cooperating models are designed to facilitate the execution of management operations by management utilities as well as human managers. Each set of models has a vendor community actively working to develop and test management agents for manageable components, with regular "plug-fests" to test their agents for interoperability. This is difficult, painstaking work, but the vendors have remained committed and results are starting to show.

In order for the various data center resources to communicate with each other, and for the DCRM to communicate with everything, there must be a single network to which every single data center component connects. The only current choice for this network is Ethernet. Ethernet is the only interconnect that the diverse resources have in common, and it is the industry standard bus for general communications and for management.

Given that Ethernet will connect everything, the question is whether or not a separate cluster interconnect or storage interconnect is required in the Data Center of the Future. Lets consider these two types of interconnect separately.

Cluster interconnects are used to create a low-latency, low-CPU-overhead communication path between applications running on cooperating servers; they have traditionally been expensive and proprietary. Infiniband provided the first industry standard cluster interconnect in 2000, but Infiniband did not achieve the popularity its inventors had envisioned, and so it remains a niche interconnect with a relatively high cost. As a result, it is not available in a redundant form (needed for clusters) on blade servers or thin servers. In 2002, the same technical visionaries that produced Infiniband proposed an architecture known as iWARP, which delivers the functionality of Infiniband on top of the TCP/IP networking protocol. Silicon vendors are now shipping the first iWARP interface chips, which run the iWARP protocols on Gigabit Ethernet and are priced only slightly higher than basic Gigabit Ethernet interface chips. These chips will find their way into most standalone servers and server blades in the next few years, making Ethernet the dominant cluster interface for the Data Center of the Future.

The choice of a storage interconnect is a bit more complex. Fibre Channel is currently the dominant storage interconnect in data centers, and has a substantial installed base, so any data center strategy must take it into account. However, Fibre Channel suffers from some of the same problems as Infiniband – it is generally not available on blade servers, and is an expensive add-on to a thin server, especially if a redundant connection is required. In addition, the work involved in maintaining a Fibre Channel network that spans the data center, and reliably upgrading it in synchrony with updates to the data center's Ethernet network, is daunting. These factors all lead one to look at employing Fibre Channel as a storage interconnect, but leaving the smaller hosts connected to Ethernet and employing gateway appliances (iSCSI to Fibre Channel translators) to connect the two networks. This does work, and presents a way to incorporate existing Fibre Channel devices into the new data center model, but it is an expensive and unwieldy solution for several reasons:

- The connections between servers and storage must be able to handle storage traffic for any distribution of applications among the servers. This implies that the gateways provided must be capable of handling the full I/O capabilities of the storage network. Current iSCSI gateways are not designed for this – they operate on the assumption that most Fibre Channel traffic stays on Fibre Channel, with only specific applications accessing storage through the gateway.
- The protocol translation performed in the Fibre Channel to IP gateway adds significant packet latency. Every I/O operation incurs this packet latency 2-4 times, and the resulting additional I/O latency can exceed the I/O latency of the storage subsystem. This reduces the performance of latency sensitive applications.
- The security models for iSCSI and Fibre Channel are completely different. Device access in Fibre Channel is controlled by who you are (specifically, by the World Wide Name of your Fibre Channel HBA), whereas device access in iSCSI is controlled by what you know (specifically, by a secure password-based authentication scheme similar to that used in secure Internet commerce). The incompatible security models make effective storage security difficult to achieve through a gateway in a dynamic application execution environment.

From these arguments, it can be seen that to maximize flexibility and minimize security and management issues, the sole interconnect of the Data Center of the Future should be Ethernet. Of course, no one is going to throw perfectly good Fibre Channel storage away to realize a new data center model, even if that model promises operational savings. An interim solution is to initially restrict those applications with data on Fibre Channel storage to the set of servers with direct Fibre Channel connections. As Fibre Channel storage is retired, migrate the data of those applications to Ethernet-attached storage and place the application servers into the general resource pool, replacing their basic Ethernet adapters with iSCSI HBAs if needed for extra performance.

It should be noted that while the Data Center of the Future will increase utilization of servers and storage, it requires some extra provisioning of network resources so that applications can get high bandwidth, low latency access to each other and their storage no matter where in the data center they are hosted. Fortunately, network resources, especially Ethernet components, are relatively inexpensive compared to servers and storage.

Storage for the Data Center of the Future

Storage in the Data Center of the Future must directly connect to the pervasive data center network of choice, Ethernet/iSCSI, but it must also support the storage features required by enterprise applications, or it cannot be part of a universal storage pool. This implies that the storage system must support all leading operating systems and cluster architectures; must be highly available and reliable through hardware redundancy, hot swappable components, and online firmware upgrades; and must support extensive snapshots and disaster tolerant mirroring.

In addition, the storage system must be fully virtualized. The DCRM is a policy engine and does not perform any storage virtualization functions itself; it must rely on the storage system to do this. Virtualization features required by the DCRM are:

- The storage system must allow flexible, automatable allocation of storage.
- The storage system must be able to expand "infinitely" in capacity, performance and network connectivity as the needs of the applications grows.
- The storage system must spread all virtual devices as widely as possible across the available physical devices, so that the I/O load is distributed evenly across the physical devices regardless of which applications are running at any given time.

Storage virtualization is provided today in several ways. Some storage subsystems consist of a storage controller that virtualizes the disks in back of the controller. This works very well until the capacity or performance requirements of the data center exceed what a single subsystem can provide; at that point you can add another subsystem, but you cannot manage the storage across those two subsystems to handle the dynamic application loads of the Data Center of the Future. The two subsystems must be treated as two separate storage pools, and applications must be manually partitioned between them. Another way to provide virtualization is through the use of a central virtualization server, with non-virtual storage subsystems behind it. This technique was all the rage in 2003, but has since fallen somewhat from favor for several reasons:

- A central virtualization server represents a significant investment, especially at the beginning of its deployment when the amount of storage it virtualizes is small.
- A central virtualization server actually adds complexity to some aspects of storage management, as it still requires data center personnel to manage the storage systems behind it as well as the virtualization server itself. Adding new capacity to the storage system becomes a multi-step process of upgrading or adding a storage subsystem, managing it to create new RAID sets, and then managing the virtualization server to recognize the new RAID sets, add them to the storage pool, and redistribute data.
- The increasingly enterprise-critical features of snapshots and disaster tolerant mirroring do not run as well on virtualization servers as they do on storage subsystems that implement virtualization natively. This is because the storage subsystems generally incorporate a highly reliable write-back cache that is used to accelerate these features, whereas virtualization servers do not incorporate such a cache. Data center managers are understandably reluctant to give up performance on these highly desirable features to gain system-wide virtualization.

There is, however, one way of building a virtual storage system that avoids the limitations of the virtualizing subsystem, and the cost and performance penalties of the central virtualization engine. That way is to build a virtualizing storage subsystem that can stand on its own but can also automatically cooperate with other, similar subsystems to dynamically balance capacity and performance across the set of subsystems comprising a storage system. In this way, you get the simplicity and performance advantages of the single virtualizing storage subsystem combined with the expandability of the central virtualization engine without a large initial investment – you can start with a single subsystem, and simply add more subsystems as your capacity or performance needs require it.

This is the architecture that EqualLogic has implemented in its PS Series family of products.

Evolving to the Data Center of the Future

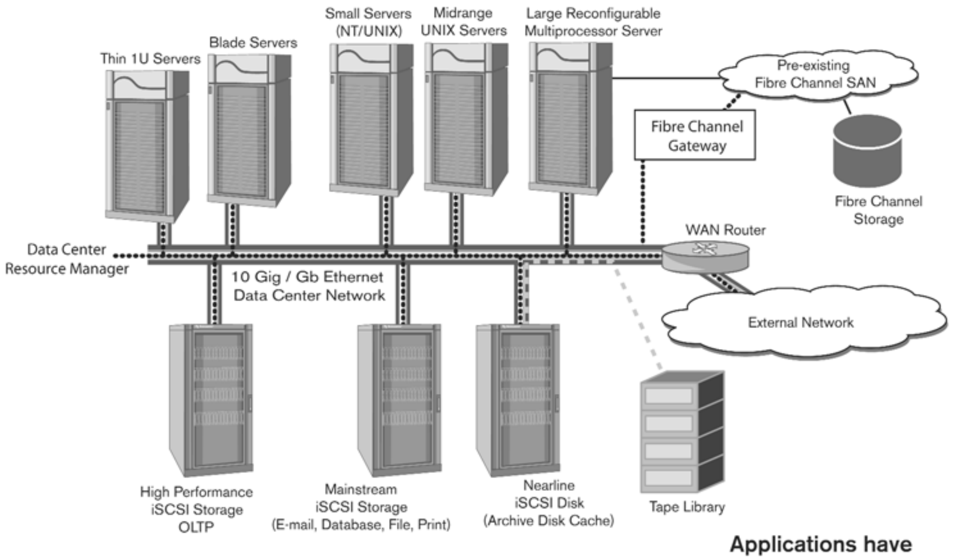
As mentioned previously, one of the primary pain points that the IT manager must cope with in today's environment is the movement of an application to a new server as well as the reassignment of the storage to that server. While application movement can result from addition of a new server, it is more frequently required on an emergency basis due to an unexpected server availability problem. Since reassignment of an application to a new server is essentially a hands-on operation it is not only slow (a real issue in an emergency), it is also prone to human error and, finally, the operation must be repeated (doubling the downtime and opportunity for errors) when the original server is repaired and becomes available.

Figure 3 depicts a generalized instantiation of the Data Center of the Future. Comparing it to the data center of today (in Figure 2), the primary benefit provided by the ubiquitous Ethernet based network becomes clearer: all storage, including the legacy Fibre Channel SAN behind the gateway, is potentially accessible to all servers. Less obvious from the figure, but equally important, application movement is simplified as a result of the iSCSI security model. When an application is moved to a new server, as soon as the new application instance is provided the appropriate authentication passwords it can access its application storage. Contrast this simple action with what is required today in a Fibre Channel network when an application is moved to a new server: Fibre Channel switches and the storage system must be reconfigured to allow access from the new application site, using vendor-specific zoning and LUN masking techniques, and then reconfigured back when the original server is reinstated.

Like any new technology advance, the Data Center of the Future is likely to enter the market in stages. From a server perspective, primitive DCRM-like solutions already exist today for uniform pools of blade servers or thin servers running a restricted set of applications, and for dynamic partitioning of a single large multiprocessor server among many applications. Single-vendor solutions that integrate server and storage virtualization will follow; and finally, enabled by management standards, multi-vendor solutions will appear. Customers will limit deployments of early versions of this technology to avoid vendor lock-in and to make sure they pick the right long-term solution.

However, just because the full realization of the Data Center of the Future may be several years away does not mean that customers should not start planning for it, and reaping some of its benefits, today. In the area of storage, it makes a lot of sense to meet today's new storage requirements with products designed to integrate easily into the future data center architecture. Storage products that interface to Ethernet via iSCSI can be used today by any server in the data center, as every server has Ethernet connectivity; many servers even have unused Ethernet interfaces on their motherboard. Buying Ethernet storage today will save on the need to add expensive gateways between Fibre Channel and Ethernet later on, and will allow immediate improvement in application availability because movement of applications is simplified. Similarly, buying storage that is already virtualized and scalable today, and has the enterprise features you will need to run mission-critical applications, will save on the need to add virtualization and enterprise features later on. It even makes sense to buy storage with features you may need in the future but not need in today’s deployment, like disaster tolerant mirroring, provided the extra cost of those features today is low enough – it could cost lots more for an upgrade later.

Figure 3
Data Center of the Future: Fluid Mobility of Computing Resources


Small Servers (NT/UNIX)
Midrange UNIX Servers
Large Reconfigurable Multiprocessor Server
Blade Servers
Thin 1U Servers
Pre-existing Fibre Channel SAN
Fibre Channel Gateway
Fibre Channel Storage
WAN Router
Data Center Resource Manager
10 Gig / Gb Ethernet Data Center Network
External Network
High Performance iSCSI Storage OLTP
Mainstream iSCSI Storage (E-mail, Database, File, Print)
Nearline iSCSI Disk (Archive Disk Cache)
Tape Library

Applications have fluid movement between storage platforms

As the final section of this paper will demonstrate, EqualLogic’s PS Series family of IP SAN solutions provides full virtualization and a very high degree of scalability at a cost comparable to competitive departmental storage today, and includes an extensive list of enterprise storage features at zero additional cost. In addition, the automated features included with the EqualLogic products make adding new storage capacity a simplified operation TODAY thus effectively eliminating one of the pain points faced by IT manager in today’s data center. These advantages make PS storage arrays a compelling choice for anyone looking at incremental storage today that will fit seamlessly into the architecture of the Data Center of the Future.

The Storage Solution for the Data Center of the Future: PS Arrays from EqualLogic

The ideal storage consolidation product for the Data Center of the Future must meet all of the requirements given above. EqualLogic PS arrays provide all of those features, and do so in such a way as to greatly simplify initial and ongoing management effort. Consider the following attributes of the PS array system:

1. The PS array system runs across a standard Gigabit Ethernet network using the industry standard iSCSI protocol.
2. High availability is guaranteed. All disks, control modules, network interfaces, power supplies and cooling fans are redundant and hot swappable. A mirrored write-back cache with 72-hour battery backup is provided. Array software is online upgradeable with zero downtime.
3. The PS array comes with snapshot creation and management features, integrated Web-based telnet and SNMP management interfaces and automatic asynchronous replication of selected volumes to another local or remote PS array for rapid disaster recovery. These features are all included in the base product – no extra software licenses, or license expenses, are required.
4. The PS array system utilizes advanced virtualization technology to hide the gritty details of storage provisioning. To create a logical volume, a manager or management utility need only specify the size of the volume; the PS system handles the details of allocating the volume to specific disks and controllers. To expand a logical volume, a manager or management utility need only specify the amount of capacity to add; again, the PS system handles the details of allocation. This makes it easier for human managers or automated management tools to allocate storage, as they do not have to map logical volumes onto individual physical disks or individual array controllers.
5. A PS array system is expanded simply by plugging another PS controller and associated drives into the network. This new controller automatically joins the existing controllers and adds its capacity, performance, and network bandwidth to the PS system. When a new controller is added, the PS system automatically reallocates logical volumes to physical storage resources in such a way that capacity and performance remain balanced among the physical disks and arrays. This means the capacity and performance of an added PS array is quickly and automatically available to all applications, as needed – no manual storage management operations required!
6. Every PS array in a PS sub-system can present the entirety of storage in the PS array system. Storage Allocation is done in such a way as to balance storage capacity and performance across the physical disks and arrays available to the PS system, eliminating physical hot spots, and the ability to access any storage through any PS array reduces network hot spots as well.
7. Secure CHAP authentication and access control for iSCSI and management interfaces are provided, allowing application migration without reconfiguration of the PS array system.
8. The PS array system supports all leading operating systems and host applications, providing access using iSCSI-compliant drivers and host adapters. See the EqualLogic web site for the most updated lists of supported systems.

Available today: a self-managing storage array that provides enterprise-class information availability combined with industry standard Gigabit Ethernet SAN connectivity and iSCSI standard protocol, and the result is a network storage solution that can communicate dynamically with any of the compute resources in the IT data center.

Future enhancements to the EqualLogic PS Series include support for low latency (15,000 RPM) disk drives, including the automatic placement of frequently-accessed data on these faster drives; synchronous data replication for “zero transaction loss” disaster tolerance; and 10 Gigabit Ethernet support. And, in keeping with EqualLogic’s “no extra charges” policy, the future array firmware upgrades that support these features will be available to all PS Series customers at no cost. Contact your EqualLogic representative to get the PS Series product roadmap, available under nondisclosure.

About the Authors

Ellen Lary has a Ph.D in Operations Research with a specialization in database technology. She was a database researcher and database research manager for Bell Laboratories, Digital Equipment, and Cincom, where she focused on turning leading edge technology into shippable products. In 1994, she rejoined Digital Equipment as the Array Controller Engineering Manager. Under her leadership, Digital's StorageWorks RAID controller product line grew from a proprietary point product to a complete subsystem family with multiplatform support. She was appointed Vice President and General Manager of the Storage Product Division in 1996, and subsequently grew the business from $1.2B to $1.9B over a 20-month period. After the Compaq acquisition of Digital, she was appointed Vice President of Business Critical Storage for Compaq. She left Compaq in June 1999 and formed Tutelary, LLC, a storage consulting company, in February 2000.

Richard Lary has been in the computer industry for 35 years. He started out as a software engineer at Digital Equipment Corporation, building operating systems and compilers for the PDP-8 and PDP-11 computers. In 1975, he served as a member of the core team that defined the VAX computer architecture and then the implementation team for the first VAX computer. After joining Digital's Storage Business Unit in 1978, Richie was a key architect and implementor for the Digital Storage Architecture and a key implementor of several of the hardware and firmware components of the associated product family. He became Digital's Storage Architect in 1990, Digital's Storage Technical Director in 1994, and Compaq Computer Corporation's Storage Technical Director in 1998. As Technical Director, he was responsible for technical oversight of the entire corporate storage architecture and product line. Richie holds 28 patents for his work in processor and storage system architecture and design, and was awarded a Lifetime Achievement Award at the Server I/O Conference in January 2000, in recognition of his contributions. He is now a member of Tutelary, LLC.

For more information regarding the EqualLogic and the PS Series, please visit www.equallogic.com or contact us at 888-579-9762 ext. 7792

TuteLary, LLC
1650 Summit Point Court
Colorado Springs, CO 80919
tel: 719-264-1990