514-82
S37/ [illegible]

The Cornerstone of Data Warehousing for Government Applications

Doug Kenbeek and Jack Rothschild
EMC Corporation
Hopkinton, MA 01748-9103
rothschild_jack@emc.com
http://www.emc.com
1-508-435-1000
1-800-424-EMC2[1] (in North America)

Abstract

The purpose of this paper is to discuss data warehousing storage issues and the impact of EMC open storage technology for meeting the myriad of challenges government organizations face when building Decision Support/Data Warehouse systems.

Introduction

Most technology advisors in government believe that data warehousing is a perfect match with government agencies. The reason is because data warehouses work best for large organizations with mission-critical data distributed on a variety of heterogeneous systems – as is often found with federal, state and local government agencies. Although slow to jump on the data warehousing bandwagon, agencies have begun developing full-blown data warehouses.

Most data warehousing planners focus their efforts on four foundation pieces - or cornerstones - of a data warehouse: (1) the operational data and its acquisition, transformation and integration into a data pool, (2) the database management system and associated servers for managing the data pool, (3) the client DSS applications, and (4) the storage system where the information resides.

One of these cornerstones if planned incorrectly will cause enormous waste and frustration and can make the entire DSS susceptible to collapse. Yet it is the one cornerstone that usually gets the least amount of thought and planning. The hidden

[1] EMC^2, ICDA, and Symmetrix are registered trademarks, and EMC, Centriplex, and SRDF are trademarks of EMC Corporation. Other trademarks are the property of their respective owners.

This paper is being distributed by EMC Corporation for informational purposes only. EMC Corporation does not warrant that this document is free from errors. No contract is implied or allowed.

Abbreviations used: DW - Data Warehouse, Data Warehousing; DSS - Decision Support System; OLTP - On-Line Transaction Processing cornerstone is the storage system that physically manages the movement, placement, backup, and restoration of data.

Potential problems associated with data storage are acute because the DW places greater stress on the storage system in terms of data volume and seek functions than operational data from business process systems. And the value of all that data is entirely dependent on the protection and speed of data movement provided by the storage system. If it doesn't work well - the DSS is compromised.

A discussion of open storage technology and its impact on DW environments must take into account other drivers of information technology change. Trends in servers, hardware, software, data management, networking, geographical distribution of systems, I/O management, failure rates, procurement strategies, and disaster recovery are all important to consider when trying to understand the benefits of EMC open storage technology.

This paper briefly recaps the history of computing and storage, reviews some current trends, and then progresses to the problems associated with storage in today's expanding data warehousing operations. It concludes with a description of the key storage shortcomings inherent in DW environments and the EMC open storage features that can overcome both long- and short-term challenges when managing Decision Support/Data Warehousing implementations.

Information Technology Recap

There are three distinct phases in information processing: the automation of labor intensive tasks, online transaction processing, and data warehousing. In essence, these represent a transition from CPU-centric computing to data-centric computing to information-centric computing. This evolution parallels the transition from batch computing to "realtime" processing to the distribution of information and empowerment of knowledge workers. In many ways they are synonymous and represent similar challenges. Each phase addresses a business's return on investment and produces its own technology challenges.

Significant trends are related to these phases.

- Generation of data is increasing with the expansion of OLTP and DW.
- Computing is transitioning from a CPU-centric to an information-centric orientation.
- Management challenges are increasing exponentially with increased demand for information.
- Information is now the key to service enhancements for all government organizations, and is increasing the pressures and demands on suppliers and implementors.

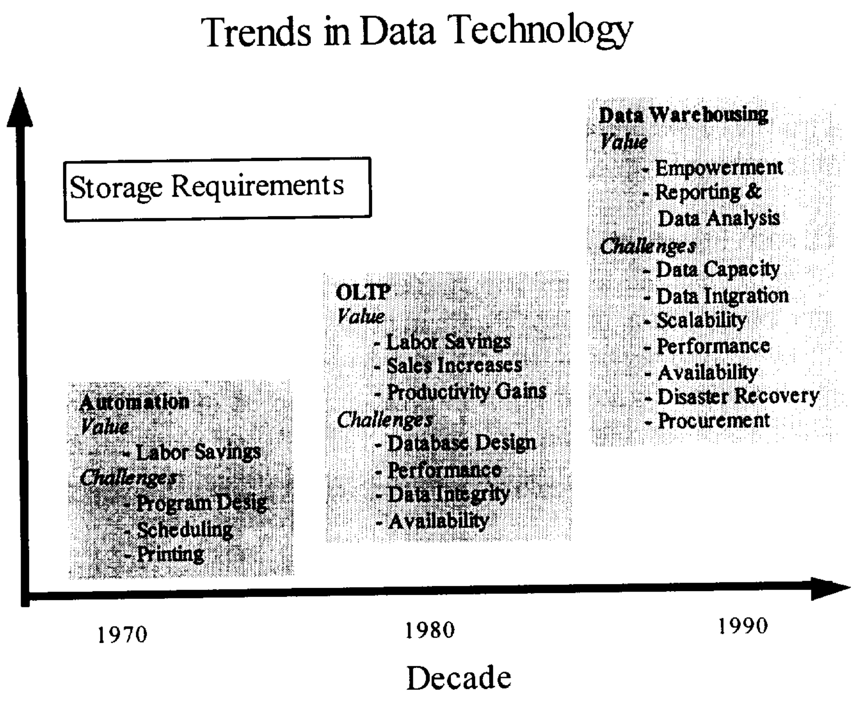

Figure I - Trends in Data Technology

The first phase, batch-oriented data processing, is not relevant to DW and not part of this discussion.

Phase two, OLTP, is now widely deployed with a trend towards client/server implementations and more widely distributing the users and data. This is where most RDBMS systems are now active as organizations continue to move operations online. Most legacy and second generation online applications are moving to this type of implementation. The databases in these environments are growing fast, with more than 10GB the norm and many growing to hundreds of gigabytes.

Data Warehousing

Phase three, the latest information processing trend, requires information managers to adopt a concept known as data warehousing. DW promises employee empowerment and creates the demand for a broad range of historical information presented in a useful format. So in addition to demanding access to critical operational data, end users also are seeking historical information to accomplish key job functions; analyze program impact and effectiveness, trends analysis, improve citizen services, and help identify and reduce fraud or other inefficiencies. This information is dominated by standard forms of textual or image data, but increasingly can include voice and video data. More complex data types, due to their large sizes, greatly increase the demands on the storage and communications systems.

Operational data, from which the data warehouse is constructed, is typically transported to a database in a centralized repository (data warehouse) where it may again be distributed to organizational servers. The operational data is scrubbed for inconsistencies and converged to eliminate duplication in the DW. All this movement, storage, and cleansing of data requires a high level of storage system performance and integrity.

Data warehousing creates historical data from operational data. Decision Support Systems gather DW data or summaries of it and transform it into easy to understand information. Since operational data is OLTP-oriented information gathered from the applications that run day-to-day operations, the DW database is systematically updated so operational data is represented to a known point in time. The speed of the DW update is dependent on the performance of the storage system. The more frequent and more voluminous the updates, the more critical the Decision Support Systems and the storage system.

The data quality, access time, and window of availability are of extreme importance since DSSs depend on the DW to produce their information. The DW storage system affects the integrity of data in the DW, the speed that the DW data can be accessed, the availability of the DW itself to the server system, and the efficiency of the updates to the DW.

EMC's family of Integrated Cached Disk Array (ICDA®) and high performance backup solutions directly address the storage requirements of a modern data warehouse while preserving the ability to choose best-of-breed technologies for other DW components.

Open System Server Dependence on MIPS

Storage subsystems for open systems have predominately followed a server manufacturer, CPU-centric model. The storage system being provided with the server. While it is often possible to substitute controllers and disk drives to increase capacity and/or performance, the limiting characteristics of these storage subsystems has not changed substantially.

CPU performance has considerably outpaced server I/O performance in the open systems arena as illustrated in Figure II.

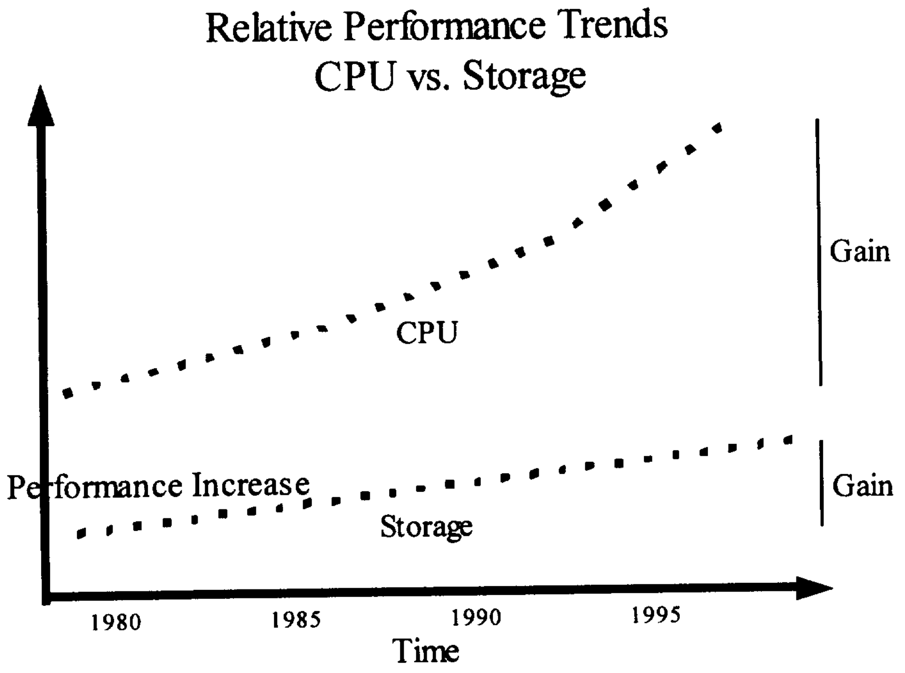

Figure II - Relative Performance Trends: CPU vs. Storage

Servers/Databases — CPU Focus

Generic and portable operating systems, industry hype, and database developments have contributed to the problem. Simply, storage subsystems have not been a focus of server hardware architectures. Open systems DW servers suffer from limited storage expansion capabilities, and often, increased storage requirements force customers to upgrade CPU types and cabinetry for increased data capacity. Many times, I/O communication channels are overloaded in fear of using up additional CPU or memory expansion slots. DW database performance is a function of both CPU and I/O performance, so high performance open storage improves overall performance and off-loads the CPU. This way the entire system is better utilized and more in balance, in many cases eliminating or deferring server upgrades.

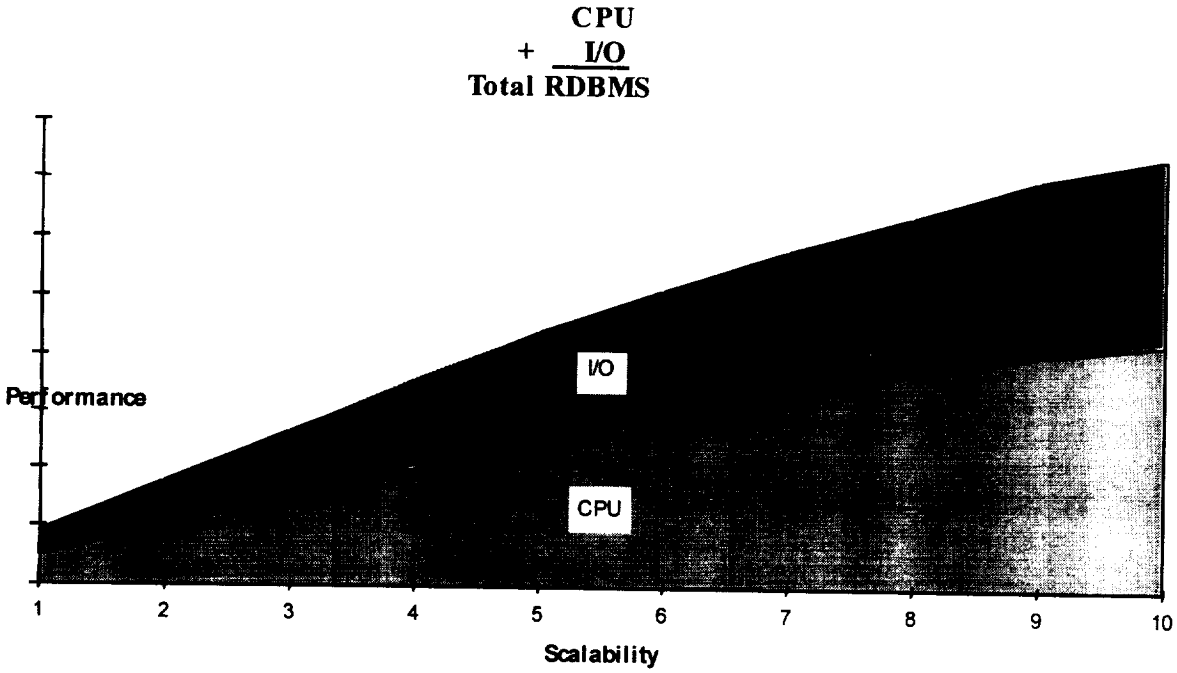

Figure III - RDBMS Performance

EMC accomplishes dramatic throughput improvements for updates and queries by providing large amounts of onboard cache memory in an intelligent, microprocessor-based storage system. By focusing on the storage component, EMC has optimized the performance and management of this unseen but critical DW component. EMC's ICDA system manages the disk drives, controllers and diagnostics as a coordinated system, relieving the CPU from these overhead tasks and delivering the highest levels of performance.

Open system hardware is made for generic use. Open systems servers may be used as communications gateways, clients, X-windows servers, firewalls, video servers, and OLTP or DW database servers. These applications range from CPU-intensive to disk-intensive environments. The majority of servers utilize the same software and hardware technology and are not primarily designed for storage-intensive applications. Server research and development expenditures bear out this fact, with the majority of investment dedicated to chip, operating system, and CPU design.

Server storage is off-the-shelf technology and "bolted on" using standard components and connections. There are a range of connectivity standards and device types, the most common being SCSI (Small Computer Systems Interface). Typically, an I/O card plugs into the server system bus and supports multiple SCSI disks per card. The storage system is built into the server cabinet, but not in a very integrated way. The disks operate independently or if there is an intelligent controller providing some coordination, it utilizes CPU cycles. Sharing of storage to boost utilization is rare between homogeneous servers and not supported between heterogeneous servers.

To address availability concerns, many server vendors have implemented RAID 0 (disk striping) and RAID 1 (disk mirroring) functionality in their systems. Although this helps increase performance and reliability respectively, there are still many points of failure within their storage subsystems and these features are again using valuable CPU cycles. The failure of a power supply, fan, SCSI channel, or controller may cause an entire bank of devices to fail. The use of RAID 1 technology within the server is questionable since many other components are unprotected and they can cause the mirrored disks to fail.

Database developers have created portable products that perform across many hardware architectures. Consequently, performance is software-oriented and highly dependent upon the speed of the processors. I/O bottlenecks and performance degradation are often addressed through CPU upgrades. CPU upgrades frequently cause a rippling effect in the remainder of the server system resulting in device and controller upgrades, downtime, and hardware incompatibilities. The reliance on MIPS for database performance has kept server emphasis on CPU technology and placed storage technology on the back burner.

Data warehousing is by nature both a storage-intensive and storage-expansive application area. Although open systems servers and databases tend to be CPU-oriented, EMC has developed the ICDA system to elegantly integrate intelligent storage algorithms with high quality storage hardware. This enhances the performance, scalability, availability, and reliability of the DW storage component. And these storage systems easily connect to every major open systems server without the need for special devices or drivers. EMC enables open systems DW servers and databases to scale and support small, medium, and large data warehouses effectively.

Distribution of DW Data

Decision Support Systems rely on the decentralization of information to the DSS user, but widely distributed applications and hardware have brought about difficult challenges in infrastructure configurations, availability, and systems management. With ever-improving communications bandwidth and technology, DSS applications running on intelligent clients can be distributed to the end user while maintaining data warehouses in a centralized or nearly centralized state. Single DW servers or small groups of DW servers provide many data security, availability, and operational advantages over a widely distributed DW scenario. There are good reasons for centralizing DW data while maintaining a highly distributed DSS environment.

In a distributed environment, storage devices are usually purchased independently for multiple server types at multiple sites. Server upgrades and consolidation may necessitate that storage devices be abandoned at worst or physically reconfigured at best. This disruptive cycle usually requires field engineers, OS administrators, and application experts to "qualify" new configurations and is both a labor- and training-intensive effort. For the DW implementation, it is therefore more effective to minimize the number of data warehouses, centralizing the data as much as possible. This could provide dramatic impact for megacenters and other government consolidations currently under way. This increases utilization of the storage investment. For sites with multiple data warehouses or other co-located systems, EMC supports the ultimate in storage flexibility — hot re-allocation of storage devices to heterogeneous servers. The result is extremely high storage utilization.

Storage management operations differ among CPU server types within a vendor's range and usually differ among vendors. There are different routines for mounting, unmounting, striping, and mirroring devices. Methods differ in the way bad spots are mapped to existing and mirrored pairs, in the way failed mirrors are swapped out, and in the routines used to resynchronize the devices. In a mixed server environment, different disk storage subsystems require extensive training, configuration, logistical, and operational knowledge resulting in labor-intensive and error-prone operations. Databases that replicate full or partial data warehouses require storage management knowledge for every type of server involved. Obviously, the greater the DW distribution the greater the potential overhead. EMC uses a single management scheme regardless of the server attached, even if multiple concurrent open systems servers are running on a single ICDA system. This simplifies the storage management challenge even with distributed data warehouses.

DW Data Availability

Some argue that the DW data is not "business critical" and so should not be considered for protection. We assume that an organization's investment in the DSS/DW is substantial, both in dollars and human resources, and that the DW data itself is key to at least one aspect of a firm's management. Consequently it is important enough to protect.

Data warehouses need or will need to store large amounts of data, so the high number of storage devices required in either centralized or distributed servers results in higher error and failure rates. The number of components that comprise a system are directly proportional to the failures rates experienced.

Standard Server Storage

To avoid failures and associated downtime in the DW, many servers mirror their storage (RAID 1) requiring a like spare for each primary drive. However, a failure of a disk controller can also cause entire I/O channels to become unavailable. To avoid this, server vendors require that all mirrored channels reside on separate controllers. Storage devices must be load-balanced between the channels, that is, every other device on a similar channel are primary with the remaining devices mirrored spares. This scheme works fine until a controller fails, causing all I/O to be achieved on a single controller, reducing performance. In extreme cases, servers may use dual-port controllers to continue mirroring spares in case of a controller failure. All of these methods require intimate server expertise and prove to be a cumbersome and administrative-intensive solution.

It should be noted that server mirroring of DW storage subsystems of this size helps reduce some downtime, but additional storage devices, controllers, and channels can greatly increase the failure rate and further increase administration requirements. To minimize this type of DW storage failure and to simplify administration, EMC efficiently packages all the necessary DW storage components into a storage cabinet. A single ICDA system includes all disks, from 35 GB to 1.1 terabytes of storage, as well as duplexed fans, power supplies, backup batteries, and controller boards.

Electro-mechanical disks tend not to fail from one second to the next, rather, over a period of time. Monitoring storage devices and their associated components for errors usually requires manual filtering of device logs. EMC again outpaces server storage with extensive automatic diagnostics, reporting, and self-correcting capabilities. DW failures are detected and corrected in time to prevent loss of data.

Open systems server storage suffers from inconsistent management utilities and limited fault detection/correction operations. EMC overcomes these DW challenges, ensuring the delivery of DSS information.

DW Updates/Backup/Recovery

Government departments and agencies like Defense, Intelligence and Secretary of State are now operating longer and increasingly on a global basis. Data warehousing/decision support systems are following this trend. DSS applications are beginning to drive a constant demand for online information over flexible work hours, increasing storage requirements and distribution of DW data over multiple time zones. This has the effect of significantly decreasing both the DW update window and the archival window for DW database managers.

Since most data warehouses are read only, backup can be viewed as disaster protection. In data warehousing, data corruption or deletion is caused principally by a programming error or by actual physical damage to the storage system. Programming errors can occur during an update or during a database modification, a frequent occurrence at many DW sites. Updates to the data warehouse use DW server resources, slowing DSS queries, and may require shutdown of the database. A large data warehouse, a frequently updated DW, or a combination of these requires a high performance storage system to minimize update time. Unless the DW is offline for an extended period, high performance backup or online backup protects DW information from deletion or corruption.

For DW operations that run 7 x 24, online backups appear to be a solution, however they can create serious server performance degradation hindering productivity and workflow. Complicating the challenge, open systems DW databases (RDBMS) environments offer minimal archival utilities, usually limited to files recognized by UNIX® file structures. Database tables are treated as a single large volume that restricts the granularity of the restore. Restoring a single row of one table would require restoration of the entire database. Conversely, database-supplied archival methods ignore operating system files.

Information managers are often required to maintain multiple archive schedules and utilities for each DW operating system and database.

EMC storage systems support up to 32 concurrent channels and intelligent writes to disk, speeding DW updates. In addition, the EMC Data Manager backup system provides high speed backup at up to 78GB per hour, with support for RDBMS integrated online backup due in 1996. Additionally, the ICDA system's local mirroring augments data protection without a performance penalty, and EMC's unique remote mirroring feature is a reliable disaster recovery method. EMC solutions enable database managers quickly update the DW or quickly recover in the event of a disaster or data loss.

Data Centralization

Although data warehousing is different in many ways than OLTP systems, it is likely to benefit from one OLTP trend — the recentralization of data.

As many agencies downsized or consolidated their organizations and distributed computing, management difficulties increased, downtime increased, productivity decreased, and departmental computing hungered for empowerment. Soon they realized the burden of the operational realities, and that they were not prepared and did not have available the tools or expertise to manage their own environments. Distributed management tools for decentralized data were and still are in their infancy.

This may be the reason for an interesting and significant trend taking place in the industry. Government organizations continue to crave more and more information, as can be seen in the move to add decision support/data warehouse applications, yet are returning the management of distributed systems back to IT. To effectively manage this distributed data, IT is centralizing the management of storage while maintaining distributed applications. This tends to increase productivity and lower procurement and operational costs. Figure IV illustrates the trend to centralize data centers found in the commercial market. Although slower in its initial implementation of DW, government entities are expected to see similar results.

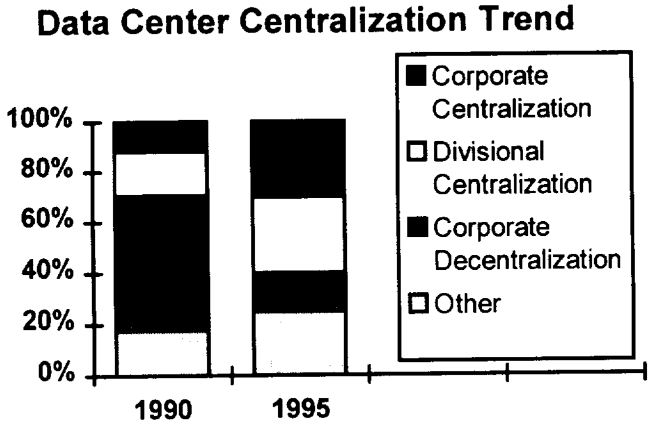

Figure IV - Data Center Centralization Trend

In less than five years, a major paradigm shift has taken place. The challenges mentioned here have database managers and the information industry evolving architectures that massively or regionally centralize data. In 1990, more than fifty percent of all companies were creating distributed architectures. In five years, less than fifteen percent of the industry is planning to continue decentralizing corporate information. More than two-thirds of industry is implementing strategies that centralize information in some fashion. This trend is reflected in numerous government agencies from defense megacenters to civilian data centers.

The trend is more than just a consolidation strategy. We are seeing a major change in the procurement, investment, and management strategy of storage. The industry is moving from a CPU-centric to an information-centric mind-set and data warehousing is leading the charge. Automating manual tasks and online operations no longer supply the competitive edge to businesses. More and better information, the promise of data warehousing, is the key to successful ventures, products, services, productivity, and roll-outs.

There have been two significant changes in the procurement and investment of hardware and software in the open systems market. Open architectures, such as UNIX, enabled the customer to purchase hardware independently of the manufacturer, protecting software investments. Relational database products enabled the buyer to further protect the information investment, procuring hardware and software solutions independent of the information.

The industry is now recognizing that an open storage strategy, adopted as an autonomous entity, is a natural extension of the open systems model. The storage subsystems should be procured, maintained, and upgraded independently of the CPU, operating system, and database in much the same vein that a network is not dependent upon database or hardware vendors. EMC has pioneered this model in the large data center and now is offering the same advantages for open system servers.

Information is becoming the primary focus of organizations and will become the single focus prior to this century's conclusion. The drive toward DW is evidence of this inevitability. A robust storage architecture is fundamental to the availability and management of this information. Builders of data warehouses who recognize this paradigm change and implement intelligent storage management strategies make their organizations more competitive.

Intelligent Open Storage Solutions for the DW

As discussed, data warehousing has some storage attributes in common with operational systems, but it also has its own unique requirements.

The following characteristics are necessary for DW storage architectures to achieve the benefits of an information-centric strategy.

- High Data Integrity Performance
- Open Architecture
- Scalable/Very Large Capacity
- Continuous Availability
- Intelligent Management
- Disaster Recovery

High Data Integrity Performance

In the past, economically protecting data and performance have been mutually exclusive goals. The use of RAID 1 (disk mirroring) technology provides consistent performance, but requires twice the amount of disks. Other RAID implementations provide data protection with only one extra disk for each four or five operational disks, but are weaker performers. EMC ICDA systems offer industry-leading RAID 1 performance and RAID-S, a high performance RAID 5 implementation.

EMC open storage systems have implemented very large caches (up to four gigabytes) that contain recently used data as well as buffering for the latency of writes to multiple devices; excellent for DW updates. This cache is nonvolatile, as a power loss would be catastrophic resulting in lost data. EMC systems also provide the ability to multiplex I/O, that is, convert synchronous requests from servers into parallel reads and writes further increasing performance. This use of RAID technology combined with large nonvolatile caches and parallel I/O, provides a high performance, high availability DW storage environment.

Many databases require the use of extensive server memory to mimic I/O caching. This is undesirable as the operating system, applications, and network are also competing for limited memory resources. This caching memory must also be duplicated for each individual database server. EMC storage systems remove these I/O constraints and have the ability to increase performance substantially with a proper configuration. DW database tests have shown from 1.5 to 4.0 times performance gains depending upon operation (load, create, index, join, scan, transaction, check) on the ICDA system.

Open Architecture

An open storage architecture is relative to the storage system design and the technology used in open systems servers. It allows for ease of connectivity with multiple servers and protects the DW storage investment. EMC uses a modular approach, utilizing "best-of-breed" standard components. This lets the DW storage system adopt technology improvements in line with industry standards and trends. Storage systems based upon proprietary interfaces isolate and limit the DW as well as add cost through specialized management and shortened life cycles.

In most cases, open systems servers use SCSI (Small Computer System Interface) as both the interface and device standard. SCSI interfaces allow servers to be attached through standard SIC (SCSI Interface Cards) controllers. IBM® has created a new interface called SSA, but it has not been adopted as an industry standard at this time and could be a lock-in strategy for customers. A competing industry-driven standard is based on a fiber interface, but is still in the development stages. EMC is tracking both technologies closely and will integrate based upon market demand.

EMC's open architecture is a proven design with thousands of customers. It is ideal for DW/DSS implementations because it is flexible enough to address the unknown twists and turns the DW is likely to take as it grows and matures.

Scalable/Very Large Capacity

A DW storage solution should allow for simultaneous connectivity of servers accessing channels to all storage devices. EMC ICDA systems are highly configurable, allowing dedicated or shared access to devices without rewiring, cabling, manual switches, or removal of drives. This is the difference in implementing a DW storage system as a logical information center versus a physical configuration. The storage system is separate from the server to accommodate multiple attachments and remove dependencies on the server vendor's design. EMC systems' physical assimilation of storage includes attachments for multiple servers each with multiple channels. The DW can grow in servers or storage and be accommodated by the ICDA without the need to trade-out existing storage.

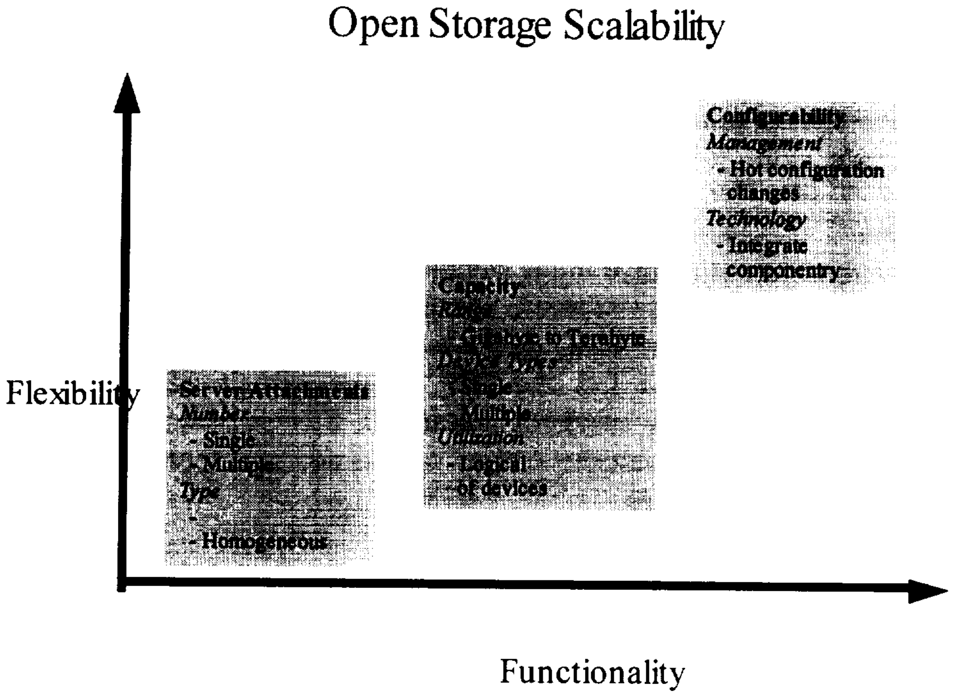

Figure V - Open Storage Scalability

EMC scalability/capacity features include:

- The ability to increase server attachments,
- The ability to increase the number of channels per server,
- The ability to increase the number of devices per channel,
- The ability to sustain DW performance while upsizing the storage configuration, and
- The ability to grow the DW to hundreds of gigabytes in a single system.

EMC's separation of storage from the servers and its terabyte capacity gives the DW managers the ability to logically assign, switch, remove, share, and manage storage independent of the servers. This functionality has significant positive implications when faced with server configuration limitations, upgrades of servers, server additions, and server consolidations. The storage investment is maintained independently of the server investment and configuration changes can be accomplished reliably and easily.

Continuous Availability

Large DW storage systems can be prone to failure as previously discussed. Consolidating storage systems into single or multiple storage environments does not preclude the need for continuous availability, rather, it forces the issue. A single component failure of a consolidated system without redundant components can cause increased downtime as the failure may effect all devices. A single fan failure in a cabinet serving multiple servers could cause all server applications to become unavailable.

EMC ICDA systems use redundant components to handle the entire load based upon a sibling failure. For example, one power supply can carry the entire load of a system in case of a failure. This is also true of fans and controllers and even buses within an EMC storage system.

Availability not only means operational, but sustained performance. Open system servers' I/O subsystems have mirrored components to some degree, but suffer from economies of scale. They must duplicate each I/O subsystem with multiple matched pairs. EMC consolidated storage requires that only a few matched pairs are necessary for an entire ICDA system.

EMC's RAID implementations discussed previously protect access to the DW by preventing the halt of a DW server due to a failed disk and also enable the recovery of data from the failed drive. The EMC advantage is that the ICDA is operating as a system, not using valuable server CPU resources to manage the RAID and other availability features.

In addition, EMC storage systems can allocate "hot" spares in case of a failure. This spare can be allocated as a replacement device for any failed storage component. Hot spares practically eliminate the vulnerability of a hard failure by narrowing the time window of repairing the faulty device.

EMC's continuous availability features utilize modular technology to both repair and upgrade systems. All components are field serviceable and cause minimal disruption. Continuous availability and storage consolidation increase DW access, permitting volumes and databases to be reallocated to other servers in case of a server failure. EMC's high availability architecture delivers information availability as an economical added value of open storage consolidation.

Intelligent Storage Management

Storage system monitoring, detection, and reporting, combined with collaborative support and management standards are an integral part of EMC storage products. In this way, DW storage problems are not catastrophic as redundant systems or intelligent algorithms recover the failed component. EMC open storage systems also include online access from a 7 ¥ 24 support organization to monitor and diagnose problems instantaneously with minimal disruption.

In addition EMC provides a centralized management console for configuring and managing the definition of attachments, channels, physical and logical groupings, and RAID levels as needed. This results in simplified, proactive storage management.

Backup/Disaster Recovery

As mentioned earlier, backup windows for offline archivals are rapidly decreasing and performance degradation for online backups inhibit further data storage expansion. Offline windows can be expanded with increased I/O performance and the throughput of online archives improved. The increased performance of EMC open storage systems may in itself suffice for increased backup demands.

For environments requiring massive data recovery, the EMC Data Manager closely couples the backup/recovery system with the ICDA system. This high performance product automates online data archiving transparent to the application and minimizes operations, training, and skill sets of administrative staff. It also provides security of all distributed data by consolidating and managing it as a single logical entity.

EMC has a unique feature — Symmetrix Remote Data Facility (SRDF™) that duplicates disk information transparently to a second local or remote location to provide continuous business operations in the event of a storage center disaster. This is accomplished with a robust fiber communication interface that supports sustained high performance data transfer over T3 lines.

Summary

Government entities are seeing the continued expansion of OLTP and the emergence of data warehousing. This is forcing a rapid transition from a CPU-centric to an information-centric information infrastructure. Dramatic decreases in storage device costs coupled with greatly increased demand for information has quadrupled storage server requirements and is enabling IT staffs to build scalable open systems data warehouses. Storage is a key technology of the data warehouse and therefore a critical element in its successful implementation.

Standard, server-supplied storage technology has failed to keep pace with DW requirements. This is partly due to server vendors' MIPS-centric development efforts and the generic design of open system servers. Decreased availability, poor management tools, inconsistent information, and end-user management delusion are forcing companies to consolidate information or recentralize data. The data warehouse adds to the problem by creating one or more additional data pools. Deployments of open system servers utilizing RDBMS DW software are confronted with the same problem. Keeping pace with the information demand while retaining the current investment in open systems technology is a major challenge.

EMC's solution to the DW storage problem is achieved by implementing a strategy that decouples the storage system from the server. This storage consolidation strategy gives the DW the flexibility to expand as the business requires — procuring servers, memory, and storage in a cost-effective manner while providing continuous access to the data pool from multiple servers. This increases DSS effectiveness and enables the implementation of a comprehensive storage management scheme.

Optimal information management is the key to competitive business strategies and EMC's open storage architecture fulfills the requirements necessary for successfully adapting data warehousing to this environment. Data centers that rely on existing server storage systems will find it difficult to cost-effectively manage their information. Open storage systems are not a trend in the industry or in data warehousing, but a major computing paradigm shift.

Business Value

EMC's intelligent storage systems are a DW advantage for organizations because they enable open systems topologies that offer the advantage of inexpensive server MIPS, ideal for the DW. An ICDA system does this by removing the storage limitations (performance, capacity, scalability, reliability, and manageability) that have previously hindered DW implementations.

Decoupling storage is investment protection and storage optimization lowers DW costs. High availability storage increases the reliability of the DSS applications and increases the DW ROI. Multiserver support and high availability deliver more information fulfilling the promise of the DW — competitive advantage.

Conclusion

Intelligent open storage is a cornerstone of every DW environment — a foundation technology. It improves the DW implementation through high capacity delivery of DSS information, more reliable information access and better storage management. A superior storage system addresses the key, hidden storage issues discussed and delivers solid business value.

The following open storage checklist provides a basis for evaluating DW storage products. Important DSS/DW-dependent applications require a check-off in the advantageous column. In addition, the service, upgradability, and storage reputation of the supplier should be heavily weighed.

Data Warehousing - EMC Open Storage Checklist

Requirement:	**Desirability:** **Limited**	**Acceptable**	**(EMC)** **Advantageous**
Host/Server Support Type	Single Homogenous	Multiple Homogenous	Multiple Heterogeneous
Device Sharing	None	Multiple Homogenous Hosts	Multiple Concurrent Heterogeneous Hosts
Number of Hosts Supported	Less than 4	4 to 15	Greater than 15
Platform Support	Single Vendor	Sun®, HP9000®, IBM/RS/6000®	Sun, HP9000, IBM/RS/6000, DEC® Alpha, Sequent®, Pyramid, SGI, Compaq®, AT&T/GIS®, IBM/SP2®
SCSI Channels Supported	Less than 4	4 to 24	Greater than 24
Maximum Storage Capacity	Less than 100GB	101 to 256GB	Greater than 256GB
RAID Support	RAID 0	RAID 0 & 1	RAID 0, 1,& 5
High Availability Features	No High Availability Features	Duplexed Fans Duplexed Power Supplies Alternate SCSI Path Fault-Resilient Cache	Duplexed Fans Duplexed Power Supplies Duplexed Controllers Alternate SCSI Path Fault-Resilient Cache RAID 1,5 Support
Dynamic Spares	Unavailable	Single Spare	Multiple Spares
Field-Replaceable Components	Unavailable	Selected Components	All Components
Online Swappable Components	None	Selected Components	All Components
Maximum Cache Size	Less than 256MB	256 to 512MB	Greater than 512MB
Maintenance & Diagnostics Support	Error messages Field Service	Error messages Field Service Remote online support	Onboard diagnostic processors Auto error reporting to service provider Remote online support Field Service Self-maintenance Option
Proactive Support Features	No Fault Detection or Reporting	Fault Detection and Reporting to Local Location	Automatic Fault Detection and Reporting to Remote Location
Device Assignments	Hard-wired	Physical Assignment	Logical Assignment Physical Assignment
Operating System Support	Single Support	UNIX® Novell®	UNIX Novell NT® OS/400® Mainframe
Disaster Recovery Support	High Speed Backup	Remote Mirroring High Speed Backup	Remote Mirroring Hierarchical Storage High Speed Backup
RDBMS Support	None	Oracle® Informix® Sybase®	Oracle Informix Sybase MS-SQL® DB2®