Interex
HPWORLD ‘96
August 9, 1996
Anaheim, California

Paper Number 3002

BEYOND DATABASE ACQUISITION:
How Disk Storage Optimizes Return on Your Database Investment

By James W. Baker
Product Marketing Manager

IPL Systems, Incorporated
124 Acton Street
Maynard, Massachusetts 01754
U.S.A.

508-461-1000
800-475-3678
http://www.iplsys.com

BEYOND DBMS ACQUISITION:
How Disk Storage Optimizes Return on Your Database Investment

ABSTRACT:

As companies spend significant dollars on Database Management System software packages (DBMSs) to streamline business functions and support effective decision-making, certain factors within the computing environment must be considered if the database application is to live up to and exceed expectations once it is installed. This paper will explain the role of disk storage in helping users exploit the potential of their database and maximize business return on the database investment.

By the end of this presentation, you will know the following "how to-s:"

How to evaluate storage architectures for optimal database performance.
How to use storage to leverage the application processing power of your CPU.
How to use storage as a database tuning mechanism.
How to evaluate storage vendors.

INTRODUCTION:

Virtually every computer installation is either using or considering the purchase of a database management application to streamline business functions and support effective decision-making. While selecting the right database software is a large part of the purchase decision, it is by no means all of it. Other factors within the computing environment must be considered if the database is to live up to and exceed expectations once it is installed. Key among these factors is the disk storage subsystem chosen to store and protect information residing in the database. **Does its architecture support or deny users control over where to place database objects, as recommended by their database vendor? Does it enhance or pose a bottleneck to performance, and why? Will it support changes in a company's IS model, or frustrate the ability to adapt to new business conditions? Is it able to grow as the company grows? Will it allow the CPU to spend its cycles on user applications rather than on I/O-related system functions? Does the vendor have a solution that will *backup* large databases, say 1 terabyte or more, in less than 4 hours? Better yet, will the solution *restore* 1 terabyte databases in less than 4 hours?**

Recognizing that an effective database environment requires more than the right database software, this presentation will take a systems point of view to educate DBAs (Database Administrators), MIS managers, and CIOs in how disk storage can optimize business return on the database investment. At the close of the session, attendees will receive a checklist to help them evaluate competing storage architectures and shape their database environment for the highest levels of availability, performance, portability, scaleability, and investment protection.

AVOIDING HOST/SERVER UPGRADES:

Most host/server vendors will not tell you this, but perhaps your next upgrade is not as imminent or as necessary as it might seem. Server vendors are in the business of selling systems. An upgrade for them is, as Martha Stewart calls it, "a very good thing." It is very good for them because you are spending more money with them, you are reaffirming your satisfaction with their products and services, you are likely to be buying an upgrade without forcing them to compete (hence you'll likely pay more), and you likely are not forcing them to benchmark or at least set expectation levels for performance.

Before spending money on that upgrade, you owe it to yourself and to your organization to take a good strong look at the match between your I/O strategy and the server host. **Is it a funnel or not?** Hopefully you will be able to ask some meaningful questions to find out once this presentation is complete. It is my premise that selecting optimal storage is a low cost way to enhance the performance of database management system installations and acquire necessary capacity at the same time.

The money associated with a host/server upgrade (such as moving from a 2-way to a 4-way server, or from a 4-way to an 8-way) is significant. So are the likely tertiary expenses: host downtime, additional wiring, spending more valuable floor space, additional training for your staff, maintenance dollars once the warranty period is over, changes to system layout such as config files, and revised operating procedures. **Once the system upgrade is installed, is your installation any less vulnerable to the failure of the most mechanical part of your system, the disk drives on which you would be placing your valuable data?** I say, NO!! **Would your organization be better off spending that money on its disk storage which very likely involves simply adding an I/O processor card and common 110 volt electrical power already located in your facility?** I say, YES!! But listen to the rest of this presentation to see if I really do persuade you.

Your CFO will be very appreciative if your organization precludes or at least postpones a significant host/server upgrade by procuring a storage subsystem instead. The dollars are usually far less as is the business disruption required by the installation. We need not mention that the performance levels being sought are being achieved by the simple installation of appropriate I/O rather than a full-fledged host/server upgrade.

STORAGE IS NEEDED ANYWAY; WHY NOT BUY STORAGE THAT IS PROTECTED?

The need for storage is growing because of many different dynamics:

- Growth of on-line databases (OLTP and OLAP)
- Growth of Data Warehousing applications (Will we ever throw anything away again?)
- Storing multiple data types (beyond traditional ASCII) such as full motion video, audio, HTML files, images, et. al.
- Many, many, many more users, users, users and their data, data, data

Clearly storage is a high growth market just from a raw capacity need point of view. However, the importance of that data is increasing at the same time. Should that data be unavailable for any length of time, it will be felt throughout the organization, the organization’s customer chain, and the organization’s supply chain. Thus it makes sense to protect it as much as possible as cost effectively as possible. Hence users are turning to the technology known as RAID, Redundant Array of Inexpensive Disks. RAID is a relatively recent technology that protects data and supports continuous data access even in the event of a disk drive failure. RAID uses controller cache and multiple inexpensive disk drives in creative ways to address three storage concerns: reliability, performance, and cost.

A number of RAID implementations can be used to protect data and speed access to it. Each of these implementations is defined as a RAID “level” with current levels ranging from 0 through 5, including combinations thereof. In the business world the two predominant levels are RAID 1 and RAID 5 because they offer the highest amount of protection. RAID 1 is called "mirroring," whereby data is duplicated across two different sets of identical hardware. RAID 5 is called "parity and data striping," whereby data and rebuild information (parity) are spread across multiple drives to offer redundancy using less than two times the amount of hardware that mirroring requires. In effect there is information about every drive on every drive so that the database application can remain up in spite of a drive failure.

Controller cache is used by various storage vendors in some very creative ways. Remembering that I/O transfers between the host/server and the storage subsystem are one of the slowest elements in a database installation, the storage vendors tend to use controller cache to:

- Mask the relatively slow speeds associated with rotating storage (7 or so millisecond average access times) (Whereas transfers between CPU and controller cache are at electronic speeds of microseconds/nanoseconds)
- OverWRITE controller cache on a Least Recently Used basis -- that is, data least recently used will be the first to be overwritten so that the likelihood of other data will be already located there
- Update and calculate parity in cache before committing the data and its parity to rotating disk storage (Cuts the number of WRITE operations in a RAID 5 application)
- Pin physical disk sectors into cache so that referenced data will already reside in the high speed cache rather than requiring a physical READ to slow speed rotating storage
- Depending on the application (heavily READ or WRITE dominated), partition the cache for an optimal mix given the demands of the application
- Use cache like a solid state disk so that so-called "hot" logical files can be found in the controller's cache
- Enable an "early release" of the host/server -- if the controller accepts data in its WRITE cache, the controller signals the host that the WRITE is committed even before the data is written down to rotating storage. In this case it is obvious that the WRITE cache must be battery backedup because the application is counting on the disk WRITE to have been accomplished

Not every vendor uses his controller cache in the same way, so it is important to discern how your storage vendor takes advantage of this valuable resource. Ask the vendor's Systems Engineer.

RAID 5 AND SYSTEM LEVEL PERFORMANCE:

Performance is an onion. Solve one bottleneck and there is another one right behind it. The next gain may be large or small, but the one after that might just be a "big one." Database users understand this phenomenon probably better than anyone in the industry. Look at the following example.

RAID 5 has historically been anathema to database applications. The reason? **Presumptive Parity and Data Striping in a RAID 5 Environment.**

RAID Level 5 is defined by the RAID Advisory Board (RAB) as parity and data striping across all the drives in a RAID array. Although well intentioned from a reliability point of view, this scheme plays havoc with virtually every database management system, regardless of vendor. In RAID 5 the controller splits up the data and “stripes” it across all the drives in the array. It likewise stripes the parity across all the drives in the array. If the data being striped is an index and also the table to which it points, it is quite likely that both elements could wind up on the same physical spindle. When this happens, the drive thrashes itself performing the data and parity updates. Any simultaneity hoped for in a multi-drive array is thwarted. Performance suffers. Users wait.

Aha! You might suggest using Logical Volume Manager from the operating system or from a set of utilities on top of the operating system. You might naively believe that because you have defined logical volumes at the operating system level that the end result will be that different spindles hold the indexes, tables, and redo logs. That’s a definite DOUBTFUL since the disk array’s controller will preemptively stripe data and parity and the problem of co-located indexes and tables described above returns.

In order for you to be sure that LVM has separated the database objects per your request, the controller has to be smart enough to override the classical RAID 5 definition. Parity must still be striped so that REBUILDs can occur upon drive failure. Yet data has to be addressable for user data placement. One vendor, IPL Systems, Inc. whom I represent, has achieved this technique and has asked for trademark protection under the term Database RAID. Users can assign Logical Volumes to SCSI addresses and LUNs and be assured that the chosen objects will reside on the chosen physical spindles. No other vendor that we are aware of can offer this kind of flexibility in a low cost RAID 5 environment. RAID 1 mirroring, yes, but not RAID 5.

I/O’s RELATIONSHIP TO DATABASE PERFORMANCE:

In their book Tuning ORACLE, Oracle Press, 1995, authors Michael J. Corey, Michael Abbey, and Daniel J. Dechichio, Jr., page 62, make the following tuning suggestions as they relate exclusively to I/O:

- Create separate tablespaces for heavily accessed tables and their indexes and put them on separate disks.
- Never put application or user objects in the system tablespace.
- By knowing how your users will be accessing the data, you can plan your data distribution better.
- Place objects that are most often referenced simultaneously and frequently on separate disks.
- Stripe large objects over multiple disks.
- Create user-defined rollback tablespaces to hold rollback segments.
- Put rollback segments in at least two tablespaces and interleaf(sic) their order in the initialization parameter file.
- Create at least one tablespace for the exclusive use of temporary segments and assign users this tablespace as their temporary tablespace.
- Put you redo logs on a disk that has a low incidence of reads and writes.
- Distribute your I/O evenly over disk controllers.
- Identify and reduce disk hot spots.
- Properly size your tables, indexes, and tablespaces.
- Monitor the space allocated and used by your tables and indexes and make adjustments when necessary.

Your storage vendor's architecture should have the tools to help address some or most of these issues. All the database tuning tips listed above are attempts to get the I/O pattern spread out as much as possible over the disk resource in order to 1) gain the simultaneity that comes from multiple arms reading and writing data on multiple spindles such that overlapped I/O can occur and 2) mask the rotational latency, head settling, etc. mechanical aspects of using disks by means of effectively using controller cache. Ask your storage vendor these questions:

Do you have tools to help me locate and identify hot spots on my disks?

Do you have tools to help me cool down hot spots on my disks such as forcing certain sectors to be always located in controller cache?

Are these tools easy to use? Are they GUI (Graphical User Interface) based? Show me.

Can I use high performance disks such as solid state disks or treat unused cache as solid state disks for hot files that can benefit from quick CPU to controller cache transfers?

COST AVOIDANCE WITH REGARD TO DATABASE CONSULTANTS AND TUNERS:

Have you ever worried that your database's tuning is the best it will ever be on the very first day and that performance will be eroded from that point on? There's a very highly paid cottage industry out there called "Database Consultants" or "Database Tuners". They exist because there is a drastic need to up database performance and keep it up. Of course there are many reasons to use the experts and they have much to offer in terms of their expertise. Yet step one, I believe, is to examine the relationship between your CPU and its I/O characteristics. The throughputs associated with a well tuned disk array system can have immediate and long term payoffs. Save the phone calls to the highly paid experts for the really serious issues associated with tough database tuning problems. A big win could be as simple as buying the right hardware storage product.

THE BACKUP/RESTORE IMPERATIVE:

No storage vendor is worthy of the name unless it has a strong backup and recovery story. Simply providing RAID protection is necessary but not sufficient. There are still disasters waiting to happen -- the natural kind -- such as tornadoes, hurricanes, and earthquakes, and -- the man made kind -- such as operators fat fingering in the wrong date for a group delete. Murphy has a company ID at every installation in the world!! Worse, he works *over* time *all* the time!!

Disasters happen. It is how we deal with them that is important. In his book, ORACLE BACKUP and RECOVERY HANDBOOK, Oracle Press, author Rama Velpuri quotes the following very scary statistic on pages 225-226. "To summarize, the survey showed that of the total system downtime, 95 percent was due to unscheduled outages, and the other 5 percent of the time, the systems were down due to maintenance. In addition, the average Mean Time Between Failures (MTBF) is calculated for the sample surveyed. This gives the mean time elapsed between two consecutive failures, which is calculated to be 102 days. The average Mean Time to Recover (MTTR) when a failure occurs is estimated to be 17 hours and 53 minutes." This is not a theoretical statistic, but a measured one as the Oracle Worldwide Support organization conducted a "system outage" survey and a "Down System and Recovery" survey of 30 companies running mission critical applications with an average 100 GB database during 1994 and 1995.

With on-line databases increasing in size, and companies needing information around the clock, the time available to perform the backup function is shrinking. As a result, traditional tape devices and save-while-active techniques are quickly becoming both impractical and obsolete. The challenge is clear: how to backup and restore the database without severely impacting business productivity. Yet, the need for data to be accurate and protected has never been greater.

The real danger is that too many companies take the risk of not backing up critical database information because they cannot afford the downtime and lost productivity. And these are companies with very large databases, in the range of 50 gigabytes to multiple terabytes of data. Strategic Research Corporation of Santa Barbara, California, an independent research firm, projects that as much as 85% of storage on UNIX servers is unprotected. Wow!!

Ask your storage vendor, what is your recommended backup solution? What is your restore solution? How fast will it backup or restore a 1 terabyte database?

THIRD PARTY STORAGE VENDORS VS. USING THE STORAGE OFFERING FROM THE HOST/SERVER VENDOR:

Perhaps your choice of a storage vendor is the same as your choice of host/server vendor. That choice may be right for your organization, but at least consider the following carefully before coming to that conclusion. Remember back -- is your client server system vendor the same as your legacy host vendor? For many folks the answer to this question is "NO" -- graphically illustrating that no purchase decision is ever permanent. We all buy the "best of breed" product at a single instant in time and "best" changes all too quickly. In short, never say never when considering whether you will always have the same suppliers -- particularly in this industry.

Systems vendors are working hard on, and investing heavily in, their servers so that they can have the "hottest box on the street" label even if just for a week or two. Competition in the server arena is ferocious. **When it becomes time for the systems vendor to choose between the priorities of spending money on R&D for the next generation server vs. the next generation storage, which do you think it will be?** Don't think that such trade-off decisions are not made at big companies -- these are the 90s and those prioritization decisions have to be made regardless of the size of this firm. Even if your client server environment is based on one vendor right now, that choice may change during the next round of acquisitions. In that case, **do you think the storage from the system vendor will be able to be unplugged from today's server and then replugged into the competitor's server once the first vendor's equipment has been swapped out? Even if the former systems vendor does allow it, how much support do you think you will get or how timely will that support be?** Contrast your answers to the same questions with the answers you might receive from a storage vendor whose product works on all the popular new servers and architectures.

Many big name host/server companies are really OEMing another company's product. Hence, the "just one vendor for both CPU and storage" story may indeed be a myth. And, relationships change depending on product cycles and renegotiated business agreements. This year's RAID product may be from one vendor, last year's from another, and next year's from a third. Where is the consistency in that? The host/server vendor has to relearn the support of differing products, sparing provisions differ, documentation sets fall out of currency, and perhaps the once cooperative relationship between server vendor and storage vendor is strained due to the presence of a new OEM on the block. Hopefully none of these issues will get out of hand at your expense. Ask your host/server vendor, **do you design and manufacture the storage subsystem?**

ARCHITECTURES ASIDE, CHOOSE A RAID VENDOR WITH WHOM YOU CAN WORK:

If you remember nothing more from this presentation, please remember this: Choose a storage vendor with whom you can work. Be sure to feel comfortable. **Is your storage vendor small enough to work with you?** If the storage vendor is too big, the staff -- both marketing and engineering -- likely is busy satisfying the 5 big OEM clients as close to equally as humanly possible. As a result, no one OEM can be given a preferential design over the other because the other 4 will be upset. This means that truly innovative ideas are shrugged off for the sake of keeping parity among the 5 big OEM customers. Smaller customers' requests for innovation (even if they are brilliant!) simply will not happen in this environment. Don't laugh -- it happens.

Is your storage vendor large enough to work with you? A storage vendor that is too small does not have the wherewithal to make meaningful design changes that will affect performance in your business or organization. There must be hundreds of RAID vendors in the marketplace. Some could be considered systems integrators because they have little or no "value add." They do not design a single thing. Rather they buy drives from one vendor, controllers from another, and packaging from a third. Be sure to consider whether the vendor is large enough to bring innovation to the party. The other indicator to consider regarding vendor size is its maintenance and repair organization in your geography of interest. If you are a US based only firm, the fact that a small vendor does not have worldwide repair presence is not meaningful. If you're a multinational, then it is drastically important and worthy of your investigation who will be the support organization overseas.

Is your storage vendor in the cookie cutting business? If so, design changes you might recommend or need are patently ignored because it would upset the cookie cutting assembly line. A way to test this flexibility is ask for a timeline to implement a certain needed feature. A reasonable answer will tell you a tremendous amount about the vendor. P.S. Be sure the timeline quote comes from someone with authority in the Engineering rather than Marketing Department.

Does your storage vendor offer full 24 x 7 coverage for service and support? This is important only if you need it now or will need it soon. Many organizations are discovering that full 24 x 7 coverage is in their immediate future It's always 10:00 in the morning somewhere in the world!

CONCLUSION:

Is your vendor and his storage architecture solution-oriented rather than peripheral-oriented? The "net-net" of this discussion is summed up in the answer to this question. Clearly, in a technical market space such as storage, it is all too easy to fall back on "feeds and speeds" and "specsmanship." If your storage vendor knows the issues of working with databases or your particular application level software and he or she has the attitude to want to work collaboratively for your mutual benefit, then you have found the right storage vendor. The two of you will speak the same language. Each of you can learn from the other. The storage vendor can make design tradeoffs if the customer needs are known and articulated. The good storage vendor will be willing to bend the architecture to fit or explain how the architecture is accommodating enough to handle particular nuances that your need structure presents.

However, there's a lot of history out there. Peripheral-oriented vendors die slowly, but they are truly dying off. They will still be calling on you. Just tell them you want to work with a vendor that understands your problems and is willing to collaborate to solve them. Look for the solution-oriented storage vendors. There are more of them out there every day and as Martha Stewart says "that too is a good thing."

BEYOND DATABASE ACQUISITION:
How Disk Storage Optimizes Return on Your Database Investment
3002-9

Storage Architecture Checklist: **Vendor A** **Vendor B** **Vendor C**

1. Does storage architecture scale with database capacity?
   Cache growth (from ___ to ____MB/controller) _______ _______ ______
   Additional controller(s)? _______ _______ ______
   Additional data paths between Host and storage? _______ _______ ______
   Other? ____________________ _______ _______ ______

2. When using RAID 5, can data striping be separated from parity striping to enable user control over data placement? (Yes or No) _______ _______ ______

3. Can multiple RAID levels be configured simultaneously in different arrays within the storage subsystem in order to optimize performance? (Yes or No) _______ _______ ______

4. Which of the following business-oriented RAID levels are supported?
   JBOD (Just a Bunch of Disks) _______ _______ ______
   Level 0 (Data Striping without Protection) _______ _______ ______
   Level 1 (Mirroring) _______ _______ ______
   Level 0/1 (Data Striping on each Mirror Image) _______ _______ ______
   Level 3 (Data Striping with Dedicated Parity Drive) _______ _______ ______
   Level 5 (Data and Parity Striping) _______ _______ ______
   Database RAID (Parity Striped but not Data) _______ _______ ______
   Other: _____________________ _______ _______ ______

Boldface type indicates most popular approaches.

5. Does RAID storage run on most popular UNIX hosts?
   Hewlett Packard (HP-UX) _______ _______ ______
   Sun (Sun OS and Solaris) _______ _______ ______
   DEC Alpha (Digital UNIX) _______ _______ ______
   IBM RS 6000 (AIX) _______ _______ ______
   Data General (DG-UX) (Unixware) _______ _______ ______
   Other: _____________________ _______ _______ ______

6. Does RAID storage run on Novell Netware servers? _______ _______ ______

7. Does RAID storage run on NT?
   Intel based _______ _______ ______
   DEC Alpha based _______ _______ ______
   HP NT based _______ _______ ______
   Other: ____________________ _______ _______ ______

8. Is controller cache general purpose (GP) or tailorable (T) by installation?
   _______ _______ ______

9. Can storage subsystems be dual ported such that drives are shared between different hosts? (Yes or No)
   _______ _______ ______

Storage Architecture Checklist (continued): **Vendor A** **Vendor B** **Vendor C**

10. What is bandwidth between Host and controller of the RAID array? If not available now, when will the feature be available?

Fast and Narrow SCSI (10 MB/s) _______ _______ ______
Fast and Wide SCSI (20 MB/s) _______ _______ ______
Ultra SCSI (40 MB/s) _______ _______ ______
SSA _______ _______ ______
Fiber channel _______ _______ ______

11. What is bandwidth between controller of the RAID array and its target drives? If not available now, when will the feature be available?

Fast and Narrow SCSI (10 MB/s) _______ _______ ______
Fast and Wide SCSI (20 MB/s) _______ _______ ______
Ultra SCSI (40 MB/s) _______ _______ ______
SSA _______ _______ ______
Fiber channel _______ _______ ______

12. How many drives are controlled by each RAID controller?

Minimum # of drives _______ _______ ______
Maximum # of drives _______ _______ ______
Most Common # of drives _______ _______ ______

13. Which components are "hot pluggable"?

Controller _______ _______ ______
Fans _______ _______ ______
Power supplies _______ _______ ______
Disk drives _______ _______ ______

14. Using the vendor's recommended tape backup strategy, how long does it take to backup a 1 TB database?

_______ _______ ______

15. Using the vendor's recommended tape backup strategy, how long does it take to restore a 1 TB database?

_______ _______ ______

16. Can the RAID vendor's cache be used as a solid state device for various hot files?

_______ _______ ______

17. Does the RAID vendor offer Graphical User Interface based tools for controlling *local* storage?

_______ _______ ______

18. Does the RAID vendor offer Graphical User Interface based tools for controlling *remote* storage?

_______ _______ ______

19. Does the RAID vendor offer SNMP compliant agents for alarm reporting to email, the supervisory console, or a beeper?

_______ _______ ______

Storage Architecture Checklist (continued): Vendor A Vendor B Vendor C

20. Who is the OEM/manufacturer of your RAID storage product?

_______ _______ ______

21. Give at least one current reference in my industry. Contact person, address, and phone.
Vendor A:
Vendor B:
Vendor C:

22. Where is the nearest office that will support me from a sales point of view?
Vendor A:
Vendor B:
Vendor C:

23. Where is the nearest office that will support me from a repair point of view?
Vendor A:
Vendor B:
Vendor C:

24. Will the RAID vendor work with me on specific product capabilities that address the unique needs of my business or industry?

_______ _______ ______

25. If yes to question 24, what is the timeline to availability as quoted by vendor’s Engineering Department?

_______ _______ ______