Storage expo essay

A selection of papers from exhibitors at Storage Expo 2007 the UK’s largest event dedicated to data storage. Now in its 7th year, the show features a comprehensive FREE education programme and over 90 exhibitors at the National Hall, Olympia, London from 17th-18th

Data Storage Crimes Can Cripple Your Business

Many SMB and mid-sized organisations have been left exposed by their data protection solutions -Andrew Wilson, UK Sales and Marketing Director, Hitachi Data Systems.

Data and information is increasingly vital to the success and growth of any business, whatever its size. Technology has advanced to the point where protecting your data need not be complex or cost prohibitive and a scalable and cost effective solution is now a reality for SMB and mid-market organisations.

The usual suspects associated with data protection are data loss; no back-up; failed restore; and system down time. These are all situations that companies of all sizes work hard to prevent, but in many instances customers are being sold over-priced, over-complex solutions which leave businesses exposed in one or more of the following ways:

- businesses’ applications are at risk either because the specific application is not supported or it is required to be offline in order to back-up the related data.
- businesses are left coping with expensive and unreliable backups due to poor processes, poor scalability and a reliance on tape.

- businesses face avoidable management costs because they have been sold multiple back-up and data protection products in order to accommodate an evolving spectrum of data protection needs.

Organisations, especially SMBs, are not always aware that this need not be their reality. An affordable, unified data protection solution is an option which will be good news for customers with an under-performing solution as well as those companies which faced with complexity and over-priced solutions suffer from inaction and have no data protection solution in place.

Performing back-ups without due care and attention will lead to potential data loss. Aside from the issue that information is all too often where an organisations' competitive advantage lies, there is also the much talked about issue of regulations and associated fines for non-compliance. Analyst firm, Enterprise Strategy Group claims an industry average of a 30% failure rate on back-ups and a 50% failure rate on restores. When questioned as part of the research, IT departments were not confident whether they would be able to recover all critical data, and within an acceptable time frame.

IT managers should ask themselves the following:

1. Do I currently have to stop applications to run a back-up?
2. Am I trying to manage multiple data protection tools whilst my company is experiencing growth?
3. Would it help to be able to restore individual documents and emails from Microsoft Exchange or SharePoint applications without having to recover the entire data volume?

If answering yes to any of these, businesses should question their current data protection solution and start investigating the alternative options available. Coping with the data explosion, which analysts put at anywhere from 30% to 100% data growth year-on-year for SMBs, as well as their expected business growth, requires solid foundations.

Losing 24 Hours of Data is a Real Possibility for Some Businesses

Failure to provide adequate protection for Microsoft-based environments and key applications is commonplace. Currently, IT may need to take Exchange, SQL Server and SharePoint applications offline to run their back-up routines. To prevent user downtime, IT is restricted to single nightly back-ups but in the event of a systems failure they could lose up to 24 hours worth of critical data. SMBs should not have to settle for less than a solution which offers back-up capability for live applications with no down time and consequent interruption to day-to-day business. Businesses should look for a solution which uses granular policies to simplify and automate otherwise complex back-up operations. Reliability can be further enhanced though the efficient use of disk, tape, network and CPU resources. Point and click reporting means IT no longer needs to manually check through log files to ensure that hack-ups were completed successfully.

Ten questions to ask your storage vendor or technology partner about your existing data protection infrastructure

1. How confident are you that I can recover all of my critical data and quickly?

2. How can I get affordable faster back-ups and restores?

3. Why do I currently have to stop applications to perform back-up?

4. How can I cost-effectively reduce the amount of data I need using a process which is integrated within our data protection software?

5. Can I easily identify different versions of files for recovery?

6. Can I restore individual documents and emails from MS Exchange or SharePoint applications without having to recover the entire data volume?

7. How can I lower total cost of ownership by reducing administration time?

8. When backing up our remote office(s), how can we reduce the tape autoloader problems which we currently experience?

9. How can I efficiently manage the process for making tape copies of back-up data?

10. How can I get a single application with a single graphical-user interface (GUI) for all data protection, e. g. email, back office? Positive responses to these questions will mean you are on the fight track to avoid the usual suspects associated with inadequate data protection, those of data loss; no back-up; failed restore and system downtime.

Achieving Compliance with Email Archiving: Dave Hunt, CEO, C2C Systems

Driven by government and industry regulatory demands and also good business governance, compliance means different things to organisations in different industries, countries and sectors, but there is a common theme– copying emails to a secure archive.

Your organisation may already face dozens of different and often conflicting regulatory requirements already, and is probably expecting more to come.

Internal email policies are more important than ever too. There has been plenty of publicity of court cases involved with discrimination and inappropriate use of email. Your organisation needs to protect itself against employee misuse of email. The ability to analyse, find or remove email will reduce the risk of litigation.

It’s a myth that there are ‘ compliant’ solutions you can just buy off the shelf. It’s up to your organisation to translate the applicable regulatory requirements into processes and find out which IT solutions can help you reach your goals. Only then can they be combined in a way to help you with proof and accountability in instances of specific legal actions around the new compliance laws and provide the tools you need for good business governance and ethical audits of business practices.

The key to meeting regulatory and e-policy requirements is to finding an archiving solution that is flexible to your needs, yet is built from the core to maintain e-mail integrity–it may prove insufficient to copy the email into another ‘ compliance’ system as the messaging integrity could be broken.

Do not underestimate email integrity, compliance regulations almost certainly require that any retrieved email record can be reproduced, viewed, and manipulated in the same manner as the original. When time comes for regulatory audits, you won’t want e-mails challenged for lack of authenticity.

It’s also important to understand why back-up of email isn’t enough to meet regulatory requirements. The fast indexing and search for retrieval of email is inherent to true archiving solutions. Should you be required to track down email, you should also expect a restricted time-frame for searching amongst possibly millions of messages and their contents. Email back-up just doesn’t allow for this to happen- true archiving solutions are built for the writing away and retrieval of high volumes of email, maintaining full indexes and audit trails which would stand up in a court of law. Another point to remember is that searching and retrieving messages within a prescribed time-frame is virtually impossibly to do manually; without a fully flexible, well-managed system your lawyers will lose valuable preparation time while the IT department track down the required email.

What about storage? Compliance indisputably means storing more email. The key for the storage manager is to do this within available resources: to work out how the need to save and manage more and more email can be fitted in with the storage infrastructure and strategy already undertaken–and do so without blowing the budget on these often unforeseen storage expenses.

The obvious approach is to make the most of the storage infrastructure you have by ensuring the appropriate email can be archived to the most appropriate media for optimal cost and accessibility.

It’s important to integrate the archival process with the storage management software or direct storage media that you have in place–archiving is a process, it’s no good setting up the archiving solution then discovering it doesn’t work with the storage software you already have. It is also essential not to waste space–compress an email so you make the most of the storage capacity that you have, not only in the archive but on the email servers too.

Follow these guidelines when you’re looking for a solution, and you’ll find that you don’t have to abandon some good common sense and the worlds of storage and email archiving for compliance really don’t have to be a million miles apart.

Whether you need to minimise compliance issues, reduce risk, or simply optimise performance, email archiving should be top of your IT priority list.

The Data Centre Goes Green: Aloke Guha, COPAN Systems

Walk into any Fortune 1000 data centre and you’ll find huge racks of servers, enormous arrays of storage disks, and extensive cooling systems that hum and drone constantly–all of them burning up kilowatts and racking up growing energy bills. To call the situation an energy crisis is no exaggeration–and many companies are feeling the pinch. One of the biggest contributing factors to the data centre energy crisis is the phenomenal expansion in data storage. As any business moves forward, it generates and amasses more and more information–and rarely deletes it. Legal regulations have added even more long-term storage requirements, and companies tend to respond the only way they know how; by adding more spinning disks. The installed storage capacity for the average Fortune 1000 customer has doubled every 10 months since 2005, and if that rate continues, the average storage requirement only four years from now may be more than eight petabytes (8, 000 TB). As a result, the number of always-on disks–along with their cooling requirements–is spiraling out of control.

The most promising solution to this growing problem is a completely new storage architecture called MAID, or Massive Array of Idle Disks. COPAN Systems developed the enhanced MAID architecture in 2002 to answer the most pressing data storage question: What’s the best way to handle the largest amount of stored data–the 70% to 90% that’s fixed, long-term, and infrequently accessed (also referred to as persistent data)? The answer: create a system that is specifically built to handle persistent data in the most efficient way possible.

As the name suggests, MAID comprises arrays of disks that are powered down–i. e.: idle–whenever they’re not needed, and spun up quickly when they are. The system allows fast and cost-effective access to persistent data while significantly reducing the energy required to store it. In addition, because MAID generates less heat than always-on systems, the disks themselves have a longer effective life, and they can be packaged more densely.

As a solution to the data centre crisis, MAID addresses all aspects of the problem. First, it radically diminishes energy requirements by putting an end to unnecessary disk-spinning and by minimising the need to cool overworked machines. Second, MAID helps minimise expensive data centre space requirements by safely putting more disks into a smaller area. While some traditional storage vendors have tried to reduce their energy footprint by tightly packing higher capacity disks, the resulting arrays have called for even more power and more cooling, and failure rates have generally increased. In comparison, MAID results in longer disk life and higher reliability–both important considerations for persistent data.

MAID also answers the corporate need to practice better environmental stewardship. For some companies, that means rising to new green standards as part of a broader commitment to social responsibility; for others it simply means minimising the costs and risks of volatile energy pricing and unstable supply chains.

For companies that have green standards to meet–as well as those that just want to save money–the energy savings that result from moving persistent data to enhanced-MAID architecture are definitely impressive. Moving 50% to 75% of the data in a typical Fortune 1000 data centre from SATA and/or Fibre Channel to an enhanced MAID platform can easily result in $470, 000 to $640, 000 USD savings (that's 240, 000 [pounds sterling] to 330, 000 [pounds sterling] GBP) in just the first year. Four years later the annual savings are expected to be $5. 7 to $7. 8 million (2. 9 [pounds sterling] to 4 million [pounds sterling] GBP).

With data storage growing at a phenomenal rate and showing no signs of a slow-down, and with data centre energy consumption climbing out of control, it's clear that the data centre's energy crisis is not going away on its own. It's up to conscientious IT departments to react–and to react quickly. The enhanced MAID platform delivers the most practical, effective, and thorough solution to averting a potential disaster by storing the majority of an organisation’s data safely, responsibly, and with less energy consumption.

Altogether, making IT greener is an important goal. Saving millions–not to mention tens of millions–in data centre energy costs is a great way to get there.

Protected Data Lifecycle Management (pDLM): Patrick Dowling, BridgeHead Software

Given massive data growth across all industries, Information Lifecycle Management or ILM has become accepted as a critical business goal many organizations hope to achieve over time. Most organizations recognize that they cannot simply continue to store and then blindly manage data of all types on primary storage. That data which has immediate relevance to active business processes merits a place on high-performance/high-availability primary storage. It also warrants special attention with frequent or continuous data protection and business continuance processes. Most data, however, is not immediately relevant to ongoing operations and does not need to be highly available or immediately recovered in response to failures or disaster. The 2005 ILM Audit by eMedia and BridgeHead Software suggested that 80% of data had not been accessed within the last 90 days and at least 60% will not be accessed ever again. Obviously, this data does not need to be stored on the most expensive storage technology or consume expensive equipment and operational costs of continuous or frequent data protection/recovery infrastructure. Clearly, as the odds that information is not going to be accessed increase over time, the underlying data should be migrated to progressively less expensive media.

Based on the answers to the following questions different storage policies will be selected:

- The question is how does an ITorganization determine which data should be migrated?
- If data is to be migrated to secondary storage, how should it be stored, protected, or secured?
- Is the data required by law or corporate practices to be available for long periods of time?

The difficulty in implementing ILM is that it requires the entire organization to be disciplined, systematically classifying information so that IT management of the underlying data can be clearly defined and automated. Few organizations have even remotely reached this level of information management. In the meantime data is still growing exponentially and IT has had to develop an alternative approach. This process is called Data Lifecycle Management (DLM). DLM is an automated approach towards optimizing the placement and data management techniques used for data throughout its lifecycle. DLM operates on what is already known about data from its attributes and textual or other analytically induced content. From the resulting data classification, policies can be created to automate the repositioning of data and to correctly apply other data management rules for creating the appropriate number of spare data copies to ensure data protection, business continuance, long-term retention, and compliance.

“ Protected Data Lifecycle Management (pDLM)” takes DLM one step further and is a more comprehensive and disciplined approach to managing the data lifecycle. The goals of pDLM include:

1. Protect data throughout its lifecycle–whether online or in the archive. Traditional HSM products may relocate data to less expensive storage, but they still require routine backup of the repository and therefore do not save much in the way of storage management costs. With pDLM, the archive is written with multiple copies potentially to multiple media types and locations, automatically backing itself up and providing rapid accessibility for disaster recovery scenarios.

2. Secure data that is copied into an archive–prevent unauthorized access, encrypted, and placed on a secure medium such as WORM.

3. Manage data retention and destruction–automatically select what needs to be retained and apply a retention policy to ensure the data is both accessible during its lifecycle and that all instances of it are immediately destroyed upon expiration. Data is not only an asset, but after its useful lifecycle, can be a liability.

4. Assimilate data for corporate governance–most data particularly end-user data is not truly under corporate control. A DLM archive should provide index and searching on both attributes and contents to give the organization the ability to rapidly find the underlying information assets within the archive.

5. Guarantee data authenticity–keeping data secure in a non-editable, non deletable environment with proof of authenticity via digital hashing algorithms applied upon retrieval.

6. Ensure regulatory compliance–regulated data requires set levels of retention, accessibility, access control, and authentication. If removable media is used as it often may be to meet certain regulatory requirements, the physical location of media should be managed. Also, regulations often call for retention far beyond the lifetime of the media, requiting a reliable strategy for data migration over time to updated media.

pDLM begins with the automated analysis of data structures to identify data that should be copied or repositioned to a secondary storage archive. The data can involve any number of formats and applications from raw, unstructured user files and email to generic databases and specialized applications. Since the automated decision to move data off primary storage may not always be correct, the DLM system must be able to provide accessibility to the data in the event it is needed. Archiving products usually make this possible at various levels including transparent access via stubbing, placeholders within the file system, database application, or alternative access through a specialized interface to the archive repository. Either way, if data that has been repositioned to alternative storage is needed, it should be easy, if not transparent, to access.

The pDLM model intelligently integrates archiving with the critical functions of backup. The process highlights the distinction between archiving and backup and the need for both technologies to address different business problems. The purpose of backup is to create copies of the online environment that can be recovered rapidly in the event of failure or data loss. Backup is oriented towards storing and moving large amounts of data and it does not purport to make data in backup save sets immediately available. The purpose of archiving is to provide an alternate, secure place for data that must be kept for long periods of time while providing a granular level of management over data that backup does not. Not only can each data entity put in the archive be retained, migrated, and stored according to its own rules, but the archive ensures that the data can be quickly located and restored. With pDLM, archived data does not need to be backed up routinely because the archive consists of multiple repository copies, some of which can be removed or located offsite alongside backup tapes.

The differences between backup and archiving are not stressed here to discredit either approach, but rather to emphasize the importance of both. This is why pDLM is fundamentally different to both traditional HSM utilities and data classification products, pDLM integrates data protection, business continuance, and disaster recovery strategies into the long-term retention and management of data as its lifecycle requirements cause it to be copied into and subsequently repositioned entirely to a secondary storage archive. It does this by allowing archives to be defined as multiple copies on multiple media types and it uses a distributed architecture to allow these copies to be written and managed at different network locations. In conclusion, pDLM represents the full integration of archiving with other vital storage management processes into a single enterprise-wide facility for ensuring that data is available for both operational and disaster recovery, that it is protected and compliantly retained for suitable periods, and that the most cost effective storage technology can be leveraged to minimize storage and storage management costs.

Clustered Storage Will Break The NAS and SAN Stranglehold: Philip Crocker, Director of EMEA Marketing, Isilon Systems

The rise of large unstructured data files is unstoppable. From high definition video on demand, to clinical data for medical research, many organisations are increasingly forced to work with data capacities that are reaching into the realm of Petabytes.

Both SAN and NAS technologies are widely established but were never designed to deal with terabyte sized files or high throughput access requirements that are emerging in the unstructured data market. The comparatively simple storage requirement of enterprise applications such as Microsoft Exchange, SAP and Oracle are much more the preserve of these technologies.

Clustered storage in contrast was designed from the ground up to deal with large files and capacities that scale into the Petabyte range and beyond. However, size is not the only factor, the overriding issue is the same as it has always been, namely cost.

Under the NAS and SAN ideology, the bigger it gets the more costly it is to implement, manage and fix in the event of the inevitable problems and there will always be problems! Hard disks fail, switches falter, data gets corrupted, irrespective of vendor, these issues happen and incur substantial costs.

Clustered storage has evolved from the clustered server space where instead of a massive mainframe processing all the application traffic, hundreds and in the case of advocates like Google–thousands of servers share the workload of requests, processes and delivery. Failures still happen but with clustering, the rest of the working servers just keep on ticking.

In a clustered storage environment, each node is independent and intelligent. No one node stores all the file data and each cluster has copies or parity bits to allow it to reassemble any file in the event of the loss of any individual disk, node or for ultra secure systems–several nodes. Nodes connect together like individual bricks in a wall, scaling from tens to thousands of terabytes.

However, lets not get carried away, clustered storage is not the magic bullet for everything. For database access, email servers and ERP systems, traditional SAN or NAS storage is a better option. For small data capacities under 2 terabytes, the upfront cost of clustered storage is probably slightly more expensive, not reaching its full potential until you get into the 10-15 TB region. These truisms may change over the next few years as a number of hybrid solutions currently only at the drawing board reach the market–but these are still someway off.

Clustered storage is also new technology and although it has some heavy weight customer such as Kodak EasyShare, Pratt ; Witney and MySpace. com, like any ‘ new kid on the block’, it has to have been around a while for certain more conservative customers to take the plunge … but the water is looking increasingly inviting.

It’s still early days for clustered storage but the take up has so far been phenomenal and the need within the market is growing even faster, for many industry watchers, it’s not a case of if clustered storage will eclipse NAS and SAN–its simply a case of when!

Data virtualisation: The Basis of File Area Networking (FAN): Paul Phillips, Country Manager UK, Brocade

Storage requirements continue to grow, making it almost impossible for companies to cope. A host of factors contribute to the enormous increase in network storage requirements. These include, among others, increased integration of WANs (Wide Area Networks) or the creation and duplication of gigantic quantities of data. In light of these factors, it is crucial that companies find new ways of efficiently organising their data and their environment. Problems in managing enormous amounts of data are not scenarios to be dealt with in the future: administrators already face concrete problems. For example, distributed storage capacities are not put to good use. A survey by the Gartner Group reveals that a mere 30-40 per cent of storage capacities in distributed networks are actually used. Set against 80-per-cent usage of storage capacities in mainframe environments, that is an unsatisfactory ratio. To that must be added the higher cost of distributed storage capacities. Another Gartner study revealed that mass storage and the associated administrative costs gobble up as much as 75 per cent of a company’s IT budget and that for every pound invested in storage hardware, another 3. 50 [pounds sterling] is spent on administration. In view of the fact that storage requirements are rising by some 100 per cent annually, companies cannot afford to solve the problem merely by adding more hardware or hiring more system administrators to manage the additional burden caused by managing distributed storage systems. In addition, maintaining high availability of distributed data throughout the entire company is incredibly complex. What is required is a well designed and central network storage architecture; something that was previously not possible in distributed mass storage systems in Windows environments.

The solution in the FAN: open virtualisation

In a perfect world, a business could simply stem–or at least control–the flow of critical data, but that is a rather utopian view. More realistic is to accept that the solution must lie in lowering the cost of storage systems. Open virtualisation is the best storage management technology to do this because it lets administrators order data logically, in other words, independent of storage location. “ Open” here means that the virtualisation solution relies on existing standardised platforms (e. g. Microsoft Distributed File System, Domain Name System) and can manage existing and future resources regardless of the system manufacturer or technology. This means it can simply make the data available to users and applications, while at the same time separating the logical and physical data components. So what does virtualisation mean for files? Rs fast aim is to shield users from the complexity of the physical storage architecture and, secondly, to enable administrators to manage this physical level without hindering user access to the data. As far as the storage itself is concerned, virtualisation is the centralisation of distributed storage systems that can then be considered and managed as individual units. Virtualisation involves grouping files from several heterogeneous storage system types (DAS, DAS+SAN, NAS) and providers. Data virtualisation gives administrators more flexibility in handling data and as well as allowing for integrated management of logical and physical data components.

Creation and management of the logical level

The logical level is often referred to as a ‘ name space’. The ‘ name space’ is what users see: a summarised display of the storage systems in the network presented as a single file system. It does not take into account the storage location or the number of physical devices making up the system. There may be hundreds of name spaces in a company that can all be grouped into one company-wide name space. Creating and managing name spaces is just as important as the physical administration of storage resources. In a purely Windows environment, creating and managing name spaces is a difficult process and is prone to errors. Most organisations prefer not to take advantage of the name space architecture simply because most system administrators believe that the problems in setting up and managing a logical name space cancel out its benefits. Data virtualisation makes it easy to create and maintain a name space of any size. Suitable solutions offer convenient functions for creating, using, managing and, if needed, reconfiguring the name space according to the user’s and organisation’s requirements.

The name space shows users the released data based strictly on function and department, regardless of its physical location. In this way, storage architectures and the respective location structures are hidden from the user. Actual changes to the physical structure of folders (migrations etc.) do not lead to errors. Thanks to the logical level, folders remain in place for users and applications, and access is still assured. This separation of the display of data from their physical locations is the fast step towards effectively managing files. Once the data is referenced in the logical name space, direct access is provided between the client and the device holding the data. The direct data path avoids the additional effort that would be created by placing the logical name space between user and physical data. Became the data is referenced just once, the performance of the file system does not suffer. Bottlenecks simply no longer occur when an out-of-band data virtualisation solution is employed.

Distributed File System (DFS): the service for virtual storage solutions

DFS is Microsoft's platform upon which other manufacturers can base their storage solutions. Because it is a component of the Microsoft Server, providers can easily develop storage virtualisation solutions using it as the basis because since Windows 95, Windows clients can be integrated in a DFS construct. The DFS service enables providers like Brocade to develop, construct and manage DFS structures; and develop solutions for central storage problems, such as disaster recovery, data migration, server consolidation, user storage management, the addition or configuration of storage capacities, or for optimising available storage capacities.

In fact, DFS, data virtualisation and corresponding file management applications can solve a range of the most urgent storage management problems in large organisations. This includes tasks such as consolidating and standardising management of heterogeneous NAS devices, ongoing management of user data, data migrations, server and storage system consolidation, and–as far as possible–guaranteeing uninterrupted data availability in the event of breakdowns.

DFS-based storage platforms such as StorageX from Brocade make it easy to use DFS and add the most important management functions to it. When creating DFS structures (name spaces), StorageX saves days if not weeks in implementation time and administration time. This is thanks to an important component–the Active Sharefinder. StorageX supports DFS management in operation–creating enterprise name spaces, data migration services, replication services, disaster recovery management, environment reporting and monitoring, load balancing, management of NetApp Filer structures and integration of Active Directory. Data virtualisation optimises the administration and user-friendliness of Microsoft DFS and enables the central administration of distributed roots and logical visualisation of a complete distributed file system. There are also powerful graphics tools for configuring DFS servers, a user-friendly interface and fully developed management tools for monitoring, troubleshooting and administering distributed DFS roots. Sophisticated reporting functions and web-based administration facilitate management and monitoring of DFS networks.

Many storage problems can be solved with structured virtualisation (e. g. DFS) and the management tools in the File Area Network that build on it. File storage virtualisation gives administrators more room for manoeuvre since it lets them independently scale and manage logical and physical storage levels. In this context, File Area Networking offers a whole new level of quality in reliably solving storage management problems in companies.