IBM System z & DS8000 Technology Synergy

July 21, 2009

Robert F. Kern
IBM ATS Americas Disk Storage
E-mail: BOBKERN@US.IBM.COM

Notices

The information provided in this document is distributed "AS IS" without any warranty, either express or implied. IBM EXPRESSLY DISCLAIMS any warranties of merchantability, fitness for a particular purpose OR INFRINGEMENT.

IBM shall have no responsibility to update this information.

IBM products are warranted according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. IBM is not responsible for the performance or interoperability of any non-IBM products discussed herein.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
USA

Trademarks

The following trademarks may appear in this Paper.

AIX, AS/400, DS8000, Enterprise Storage Server, Enterprise Storage Server Specialist, ESCON, FICON, FlashCopy, Geographically Dispersed Parallel Sysplex, HyperSwap, IBM, iSeries, OS/390, RMF, System/390, S/390, Tivoli, TotalStorage, z/OS, and zSeries are trademarks of International Business Machines Corporation or Tivoli Systems Inc.

Other company, product, and service names may be trademarks or registered trademarks of their respective companies.

Abstract

System z has always had a unique relationship with its storage since the inception of S/360 in 1964. IBM's publication **The ESA/390 Principles of Operations (SA22-7201) & z Architecture Principles of Operations (SA22-7832)** defines the Extended Count Key Data architecture. Over the subsequent 45+ years many extensions have been made to the basic I/O architecture that reflects technology improvements in System z Processors, the System z Channel Architecture and the Disk and Tape Storage Subsystems. These improvements have optimized the transfer of data between the Host and storage devices to meet the needs of System z customers. In the early 1980's, with the invention of the cache disk controller the new era of Storage Virtualization began which has continued to today with numerous z/OS® software, System z hardware and DS8000® synergy items in the areas of end to end application and system performance, scalability, business continuance (availability, disaster protection etc), data replication, simplification, backup, and security. This white paper will detail many of the highlights and focus on specific System z, z/OS and DS8000 "Advanced Functions" and solutions available in the marketplace today. The IBM DS8000 has been at the forefront of innovation with the System z hardware and software platform, continuing the integration of advancements which began some 45+ years ago.

Introduction

OS/360 IBM introduced the Count Key Data (CKD) architecture which enabled the Host Operating System Software to optimize the transfer of data between the host and the spinning disks. CCW chains from key components (access methods, sort, etc.) were optimized to provide the best read and write performance. The CKD architecture was subsequently enhanced to the Extended Count Key Data (ECKD) architecture that we have today. ECKD enabled host software to communicate to the storage subsystem 'hints' relative to actions contained in the rest of the CCW chain. The Define Extent, Locate Record CCWs provided these hints.

In the early 1980s with the 3880mod 11 & 13 control units, quickly followed by the 3880 mod 21 & 23control units the concept of a cache control unit and storage virtualization was introduced. Cache enabled the control unit to actually prefetch data based on various operating system hints. This drastically improved the read performance of disk control units as data could now be read out of the controller's memory, instead of always having to access the physical disks. The invention of the cache control unit, introduced the concept of storage virtualization with the control unit managing the physical aspects of the rotating disk outboard and enabling better end to end optimized performance of all disks attached to it.

Data Facility System Managed Storage (DFSMS)

In 1987, IBM introduced two additional storage virtualization concepts that took the notion of outboard management of the physical disks attached to the control unit to a new level.

The first was the introduction of Non-volatile Storage (a write cache) which enabled the control unit to manage not just the reading from the physical disk but now also the writing to the physical disk. The name of a control unit also changed from being a control unit to being a 3990-3 disk storage controller.

The second concept came about on the MVS software side. A new subcomponent of MVS was introduced called Data Facility Systems Managed Storage (DFSMS).

DFSMS brought two key concepts to the marketplace:

1. **System Managed Storage. The client could now provide 'business policies' on how data would be managed.**
2. **Storage Groups - The formal separation of the physical storage from the logical view of the data.**

Clients could now manage their data based on the needs of the business via the SMS management policy constructs of:

Storage Class - Captured Availability & Performance attributes of the data.

Data Class – Managed special case information about the data. Fro example, please export these data sets on this specific type of tape media.

Management Class – Lifecycle Management attributes of the data. (i.e. Hold the data on tier 0 storage for 30 days, then migrate the data set to tier 1 storage for 3 months, after which it should be offloaded to tape. Expire the data after holding it for 3 years.)

ACS routines - Customer written routines to assist Allocation/Migration & Recall components of z/OS in placing data sets on specific volumes or sets of volumes in a storage group based on various attributes. (i.e. high level qualifier of the data set name, etc.)

DFSMS brought together the concepts of tiered storage and end to end data lifecycle management all policy managed by integrating System Managed Storage concepts with the access methods (DFP – Data Facilities Product) and the optional features of DFSMShsm (Hierarchical storage management) and DFSMSdss (backup/restore). With DFSMShsm and DFSMSrmm the management of inactive data (data typically stored on tape) via the Management Class business policy also was created.

The basic policy based management constructs of DFSMS have now been in place for 20+ years and have enabled z/OS customers to optimize their storage utilization (upwards of 80+%), while also managing the business requirements of data availability and performance over the lifecycle of the business data with a minimal staff.

Within z/OS, DFSMS introduced several new subcomponets that interfaced with several existing z/OS subcomponents. Gradually, all z/OS components switched their management philosophy from physically managing the data & disk storage subsystem resources to logically managing the data based on the actual customer defined data requirements. (i.e. SMS defined policies) In doing this, z/OS became aware of the entire hardware/software "Stack". The notion of the 'control unit' or disk storage subsystem was defined and managed in a few low level z/OS subcomponents including: IOS, the Asynchronous Operations Manager (AOM), the System Data Mover (SDM) and various error recovery subcomponents. The various access methods were changed to be less conscience of the physical aspects of the spinning disk and tape and more focused on reading data from and writing data to the storage subsystem as efficiently as possible. More full track and Multi-track operations were introduced along with host buffering in an attempt to insure that when possible data was in memory when required. Non volatile memory within the storage controller meant that if optimizations were successfully done, data was no longer written directly to disk. Therefore, the physical attributes of the disk could now be managed and optimized by the storage controller based on all read & write requests that it had for each device. The storage controllers cache and NVS became the real target of the hosts read and write operations. This level of indirection has subsequently lead to a number of improvements centered on providing performance hints relative to the optimization of prefetching & destaging data from and to the physical disk.

AOM coupled with changes to IOS introduced the notion of long running "asynchronous" operations outboard in the Disk and Tape storage controllers. Initially these operations centered around the management of Cache and NVS, but in 1993 they were extended to the backup function called Concurrent Copy. Concurrent Copy was integrated into the z/OS stack and exploited by DB2, DFSMShsm and DFSMSdss to provide a fast backup of various types of data, by a "Snap Shot" like function. The notion of a Point in Time (PiT) copy of data was introduced with concurrent copy. DFSMS created the "System Data Mover" (SDM) subcomponent to centralize the management of Concurrent Copy. SDM has been extended such that today it also manages host striping of data across multiple volumes and the storage based data replication features and functions.

z/OS Parallel SYSPLEX Evolution

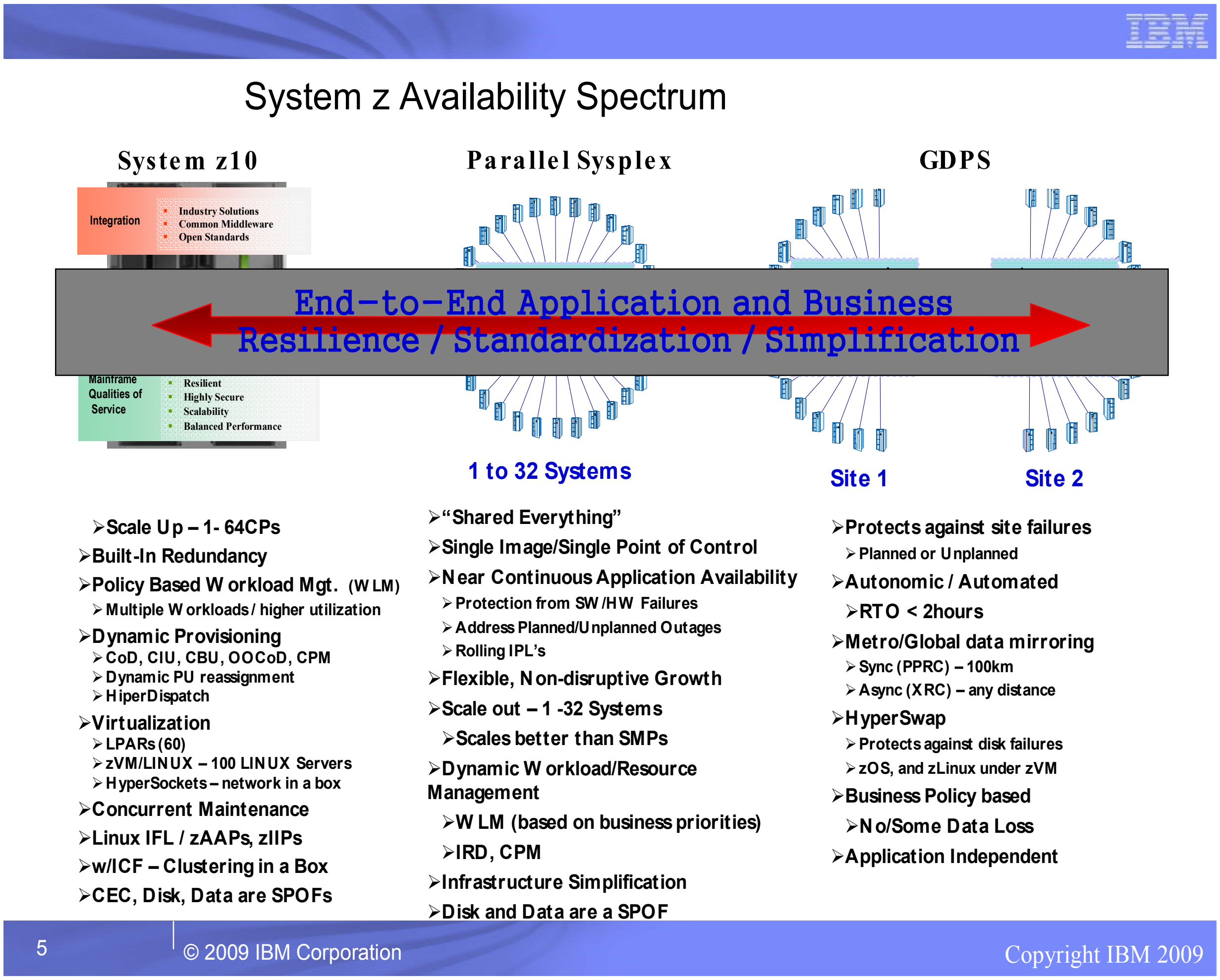
As intelligent storage controller virtualization and its related software has evolved so has the evolution of System z processors and the Parallel Sysplex. The following chart list some of the availability aspects introduced into the System z10 hardware, the zOS Parallel Sysplex Software as well as the GDPS® multi-Site resource management automation. These sorts of innovations have provided end to end application business resilience, standardization & simplification capabilities that provide clients the highest quality of service available from any system available in the marketplace today.

- A multi-site parallel Sysplex can be defined today up to 200km apart providing the highest levels of availability.

GDPS automation manages duplicate resources across the two sites (or two logical sites on the same data center floor) to mask all failure scenarios:

- CEC failures are managed dynamically by Sysplex concepts, eliminating a failing image as a single point of failure.
- Disk subsystem failures are masked by HyperSwap®.
- Tape subsystem failures are masked by peer to peer or Grid tape replication.
- Clock failures are addressed via multiple Sysplex Timers.
- CF structure failures are managed via CF Duplexing.
- Network Failures are addressed by the concept of persistent sessions.

To address an entire site failure, GDPS can perform a site switch to another local site or to a remote (out of region) location. In addition, GDPS has a function called GDPS/zDR that provides similar functionality for other System z Images including zLinux and zVM. For Multi-platform applications GDPS today inter-operates with GDOC (Geographically Dispersed Open Clusters) automation such that each platform can be recovered to the same point in time and the application can thus be restarted on each specific platform automatically through end to end automation that addresses the server(s), workload, data and a coordinated network switch.



System z Availability Spectrum

- Scale Up – 1- 64CPs
- Built-In Redundancy
- Policy Based Workload Mgt. (WLM)
  - Multiple Workloads/ higher utilization
- Dynamic Provisioning
  - CoD, CIU, CBU, OOCoD, CPM
  - Dynamic PU reassignment
  - HiperDispatch
- Virtualization
  - LPARs (60)
  - zVM/LINUX – 100 LINUX Servers
  - HyperSockets – network in a box
- Concurrent Maintenance
- Linux IFL / zAAPs, zIIPs
- w/ICF – Clustering in a Box
- CEC, Disk, Data are SPOFs

- "Shared Everything"
- Single Image/Single Point of Control
- Near Continuous Application Availability
  - Protection from SW/HW Failures
  - Address Planned/Unplanned Outages
  - Rolling IPL's
- Flexible, Non-disruptive Growth
- Scale out – 1 -32 Systems
  - Scales better than SMPs
- Dynamic Workload/Resource Management
  - WLM (based on business priorities)
  - IRD, CPM
- Infrastructure Simplification
- Disk and Data are a SPOF

- Protects against site failures
  - Planned or Unplanned
- Autonomic / Automated
  - RTO < 2hours
- Metro/Global data mirroring
  - Sync (PPRC) – 100km
  - Async (XRC) – any distance
- HyperSwap
  - Protects against disk failures
  - zOS, and zLinux under zVM
- Business Policy based
  - No/Some Data Loss
- Application Independent

5 © 2009 IBM Corporation Copyright IBM 2009

DS8000 and z/OS Performance

The disk storage subsystems and z/OS have created a number of synergy items over the years to simplify the configuration of I/O resources and to address end to end I/O performance.

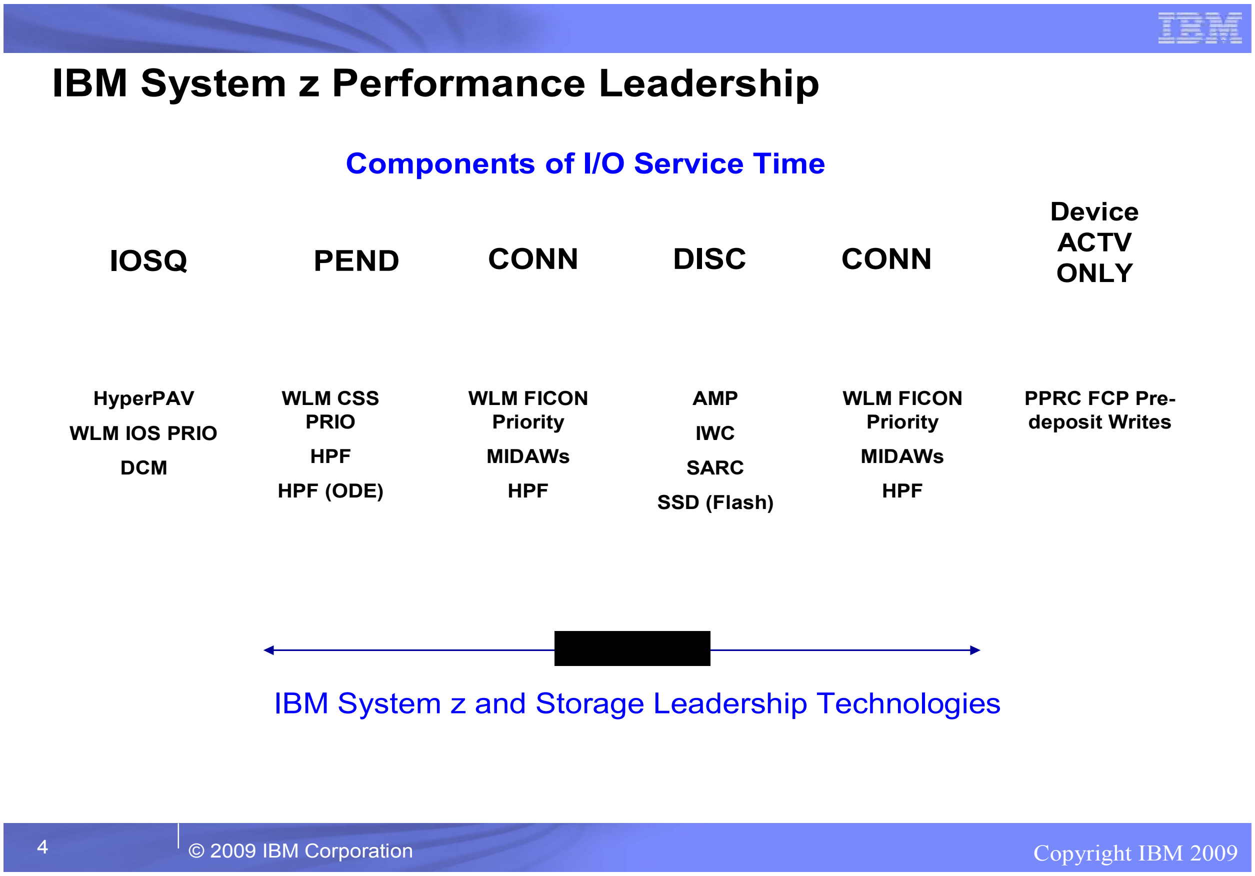
Configuration Simplification

Self Describing devices conforming to the ECKD Self Description Architecture drastically improved the configuration and management of I/O devices. The long and painful task of defining each physical device by hand was eliminated and replaced with z/OS dynamically “discovering” the attached devices, their features and functions. In addition as device features and functions change in real time, the storage subsystem working with z/OS dynamically recognizes the changes.

- During z/OS system initialization, vary online and device state transition processing, various host control blocks are dynamically built and updated with type & model, mode, and feature/function information about the various devices attached to the system. This enables SMS to assign new data set allocations based on the defined storage class criteria. This has gradually been refined over the years. Host software can now scan the host control blocks for storage subsystem and device information to find if a specific feature or function is available and as a result exploit that function. The data can be updated in real time typically through state transition interrupts. When specific static or state information for a device changes, the storage subsystem raises a device state transition interrupt. z/OS responds to this interrupt by reading the self description information for the device and updating its host control blocks. Through this mechanism, the host control blocks typically maintain the ‘latest’ information about the device, thus minimizing the host software from having to interrogate the device before various advanced functions are attempted. For example, FlashCopy® is exploited by DFSMSdss Fast Defrag if the disk/subsystem supports the function and the state of the device is such that the function can be performed. If, the disk subsystem/device does not support FlashCopy, then a normal Defrag function is performed by DFSMSdss. This is a 1st level of management. Of course, if the device state changes between when the software control block is checked and when the actual FlashCopy I/O is issued a command reject is given back by the device and DFSMSdss then performs the normal Defrag. In most cases however, having the feature/function information ‘current’ in the various host control blocks helps to streamline the checking by the software and reduces the number of I/Os issued by the host software to the device.

Performance

- Over the years there have been many System z and Storage Subsystem synergy items focused on performance. These items include replacing the original parallel channel architecture with ESCON®, and then FICON® and now with High Performance Ficon to optimize the physical protocols used to transfer data between the Host and the various tape and disk storage subsystems with full integrity checking throughout the process... Other performance synergy items available with the DS8000 Disk Storage Subsystem and System z today, include both static and dynamic Parallel Access Volumes, HyperPavs, MIDAWs (Modified Indirect Data Access Word), Adaptive Multi-System Prefetch (AMP), Adaptive Record Cache (ARC) and Intelligent Write Cache (IWC), I/O Priotity across the Ficon channel as well as within the DS8000 storage subsystem managed to the z/OS 'application level' priority by the z/OS Workload Manager (WLM) subcomponet. The following chart illustrates some of the recent z/OS and DS8000 storage subsystem performance related synergy items. As one can observe, various items have helped to reduce specific aspects of the components that make up the entire I/O service time. A specific item provides some performance improvement, but the focus of z/OS and the DS8000 has been on improving the end to end I/O time which at the end of the day is reflected back to the applications running on z/OS and the total throughput of those applications.



IBM System z Performance Leadership

Components of I/O Service Time

IOSQ	PEND	CONN	DISC	CONN	Device ACTV ONLY
HyperPAV WLM IOS PRIO DCM	WLM CSS PRIO HPF HPF (ODE)	WLM FICON Priority MIDAWs HPF	AMP IWC SARC SSD (Flash)	WLM FICON Priority MIDAWs HPF	PPRC FCP Pre-deposit Writes

IBM System z and Storage Leadership Technologies

4 © 2009 IBM Corporation Copyright IBM 2009

System z High Performance Ficon (zHPF) has set a new standard for optimizing the I/O protocol over the channel between the host and the storage subsystems. There have been a number of different I/O protocols introduced on various distributed platforms including SCSI, iSCSI and Fibre channel protocols, but the System z channel architecture has remained the industry standard for reliability, robustness, scalability and performance.

End to End data integrity checking, for System z includes error isolation within the channel, fabric and switches/directors. As intelligent storage controllers and software virtualization evolved, so has the evolution of System z processors and z/OS. This helps to reduce problem determination time and effort on failing components. As the storage subsystems have grown in size and functional complexity, the channel architecture has also scaled. zHPF continues to raise the bar in this arena. Exploitation of the zHPF channel protocol has been done in an end to end manner. z/OS Workload Manager (WLM) helps provide client defined workload priorities on each I/O operation, which is used by IOS to prioritize I/O over the channel, and within the DS8000. The importance of this type of end to end synergy is found in a simple example: when nightly backup operations 'run long'. The OLTP online workloads start up in the morning and a long running nightly backup can impact OLTP performance. This is dynamically minimized by the WLM I/O priority which will provide the OLTP workload, I/O priority over the backup programs I/O. This in turn minimizes the impact on the OLTP workloads when backups run "long".

Various improvements in zHPF have now been introduced with the DS8000. For example, associated with the large Read Record Set I/O chains that z/OS Global Mirror (XRC) utilizes. Optimization of this channel program has enabled better performance of channel extenders that forward Ficon frames. This in turn has provided clients a greater choice in selecting channel extender technology for greater distances. Before this innovation, clients had a limited choice of channel extenders that fully understood the CCW chain and provided the performance optimizations that were required.

Several years ago, Parallel Access Volumes (PAVs) were introduced into z/OS helping to drastically reduce IOSQ time. Static and Dynamic PAVs were managed by WLM and adjusted on WLM timer intervals. Based on what IBM learned on the performance impact of PAVs, HyperPavs were invented, enabling PAVs to be adjusted on an I/O by I/O basis dynamically within the DS8000. This finer granularity in tuning alias addresses has provided better performance, enabling more work to be completed on the platform in a specified period of time.

Another improvement has been in the area of synchronous data replication. Metro Mirror (PPRC) between the two DS8000 Storage Subsystems. Pre-Deposit Write is an optimization of the Fibre Channel Protocol (FCP) from the standard 2 round trips to a single round trip. This optimization has enabled longer synchronous distances with data replication by reducing the synchronous 'distance penalty' by 50 %. The DS8000 Synchronous Data replication distances have been extended from 103km to 300km. With fewer protocol exchanges Pre-Deposit Write also has provided better 'Link" utilization,

Numerous DS8000 disk storage subsystem caching improvements have been made over the years. z/OS and the data bases have been modified to provide various forms of hints to the disk storage subsystem on sequential processing as well as various data base I/O operations. Focus in this area has enabled clients to fully optimize the use of existing cache and NVS sizes, by optimizing what data is in cache or NVS at any point in time. A recap of recent improvements include:

2004 – ARC (Adaptive Record Cache) dynamically partitions the read cache between random and sequential portions

2007 – AMP (Adaptive Multi-Stream Pre-Fetch) manages the sequential read cache and decides what, when and how much to prefetch

2009 – IWC (Intelligent Write Cache) manages the write cache and decides what order and rate to destage

Extended Address Volumes (EAVs) is the latest architecture change required to address issues with the amount of data that can be attached to a Parallel Sysplex. Over the years z/OS and the DS8000 have introduced larger and larger volume sizes. 3390-3, 3390-9, 3390-27, 3390-54 and now EAVs. This has enabled customers to reduce the number of devices, to be managed in a sysplex. DS8000 HyperPavs have provided the additional performance improvements required to share the larger devices across the Sysplex, which have in turn has enabled more and more data types to be stored on the various size volumes. HyperPavs have also reduced the effective number of alias addresses required. Trends toward larger volumes sizes will continue. They are enabled through extensions in z/OS software, and the DS8000 Disk Storage Subsystem. But, continued end to end focus on overall performance, with capabilities like zHPF and HyperPav is important to enable the larger volumes to be used in a fast efficient manner in z/OS.

Another example of end to end z/OS and DS8000 synergy comes with the introduction of SSD Devices on z/OS. The following excerpt from: IBM RedGuide: **Ready to Access DB2 for z/OS Data on Solid-State Drives** by Jeffery Berger and Paolo Bruni, IBM Copyright 2009 provides some interesting insight into the value of the IBM end to end DB2 – z/OS – DS8000 performance focus providing a real difference for a customer's workload.

.

"Rotating disks that spin at rates of 15 000 rotations per minute typically achieve response times in the range of 4 to 8 milliseconds (for cache misses). In contrast, SSDs provide access times in tens of microseconds, rendering data access times that are more like dynamic random access memory (DRAM). However, the access time of the SSD drive itself ignores the functionality required of an enterprise storage subsystem and the effects of such function on response time. An enterprise storage subsystem provides layers of virtualization in order to shield the software from dependencies on hardware geometry. An enterprise storage subsystem needs to provide continuous availability, and it needs to enable multiple hosts to share the storage. This functionality, which is provided by the DS8000 storage server, adds to the response time of SSDs in an enterprise system, bringing it to the level of hundreds of microseconds. Furthermore, for predominant sequential accesses (cache hits), HDDs and SSDs show similar performance. Nevertheless, we must not underestimate the value of eliminating the seek and rotational delays of HDDs.

Among the lessons learned in the measurement study of DB2 for z/OS is that solid-state drives appear to cause greater stress on the channel subsystem, because SSDs enable higher levels of throughput. More improvements in the system as a whole enable solid-state drives to further realize their full potential. IBM has delivered High Performance FICON® (zHPF) to the z/OS environment to help make this happen. IBM recommends zHPF for an SSD environment. Even when the channel subsystem is not stressed, zHPF provides lower response times when accessing SSDs.

Another lesson is that disk drives are just one aspect of the I/O subsystem. I/O improvements, such as those improvements made by IBM for the z/OS environment in recent years, continue to be important. Among these improvements are high speed channels, the Modified Indirect Data Address Word (MIDAW)1 facility, Hyper Parallel Access Volumes (HyperPAVs), and Adaptive Multistream Prefetch (AMP).

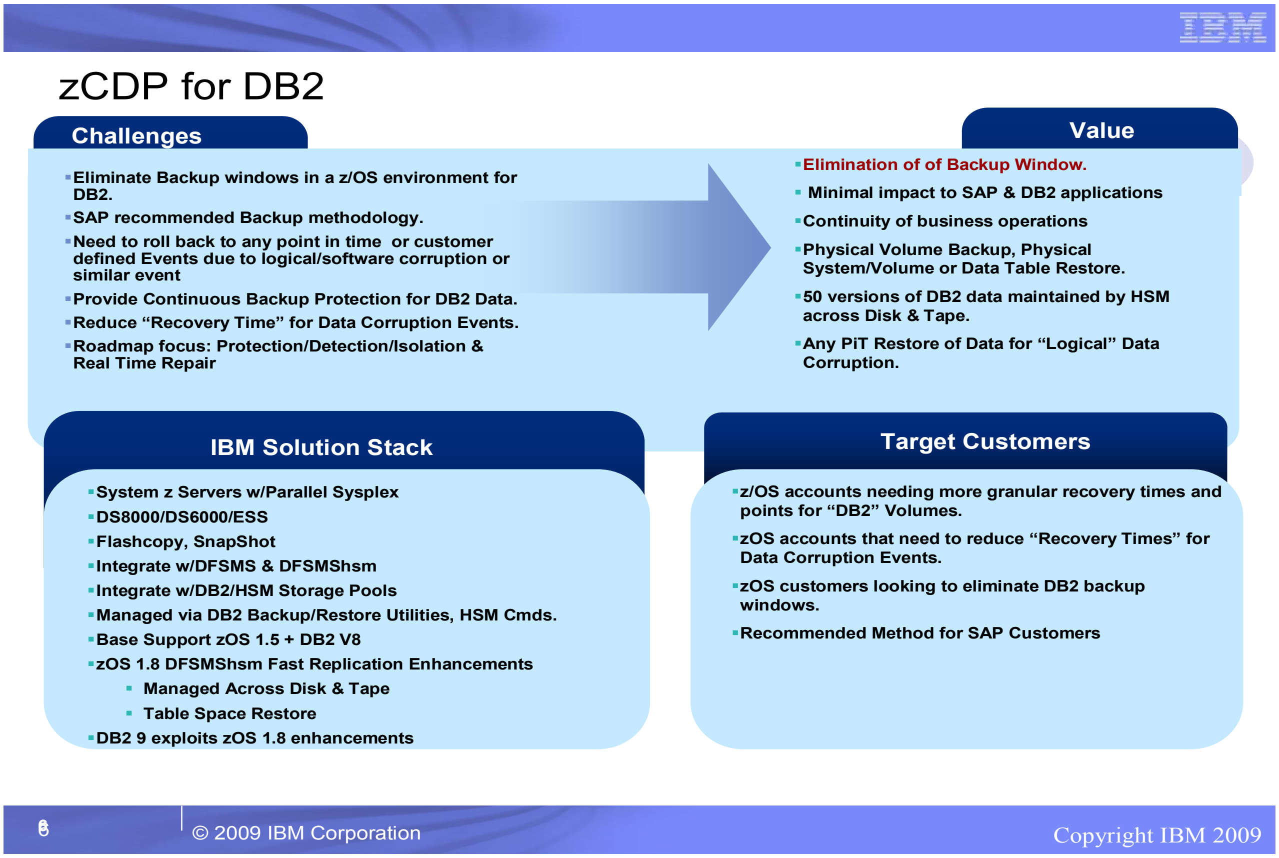
DB2 buffer pools and cache in the storage server continue to be important. It is not uncommon for an enterprise system to experience average DB2 synchronous I/O time of two milliseconds or less, and that does not count the zero I/O time for DB2 buffer hits. So, it is simplistic to argue that the busiest data sets belong on expensive fast-performing devices, such as SSD. However, data sets with low access rates do not belong on SSD either. Rather, databases with a large working set and high access rates have the most to gain from SSD performance. But large databases require large disk capacity; consequently, you cannot expect to purchase just a small amount of SSD storage and expect to realize a large performance benefit."

Backup Functionality Improvements

z/OS and the DS8000 Storage Subsystem synergy have also provided improvements in reducing the time it takes to perform backups. In 1993, IBM invented the Concurrent Copy function. Integration of DB2, DFSMShsm, DFSMSdss and z/OS helped to reduce the backup windows of DB2 data and DFSMShsm CDSs. (HSM control data sets). Backup operations are typically done ten times more than Restore operations in typical IT environments and as the trend to 24x7 processed started to emerge, Concurrent Copy was the first attempt to address the impact of backup windows. Concurrent Copy was then followed by FlashCopy which was introduced first on the IBM System Storage Subsystems. DS8000 FlashCopy has enabled a fast efficient method to clone volumes with or without a full background copy being done. Functions like Consistent FlashCopy and combination of the Metro Mirror "Data Freeze" coupled with FlashCopy have enabled fast efficient Point in Time backups. In many cases, the actual backing up of data to tape now occurs in the background or is offloaded to a different processor set from production. The Point in Time nature of these backups have reduced the strong requirement that the applications and operations teams work closely together in understanding all of the dependent data requirements for an application or set of applications as input to the backup/restore process. Today, many customers create a PiT version of all data attached to the Sysplex for backup, disaster recovery testing, system clones, application testing etc.

A recent improvement, through synergy of DB2, z/OS (DFSMShsm Fast Replication), DFSMS and the DS8000 FlashCopy function is zCDP for DB2. With DB2 9, the DB2 System Backup utility interfaces with DFSMShsm Fast Replication to FlashCopy first the data base tables, and then the data base logs to SMS Copy Pools. HSM then maintains up to 50 versions of the DB2 Backups across disk and tape, while the customer manages the backup versions via the standard SMS construct of Management Class (MC). This function outlined in the following chart has *eliminated the need for DB2 backup windows*. The DB2 System Backup utility manages the consistency of the DB2 data and as a result, the entire DB2 backup operation can now occur in the background without stopping or even quiescing the DB2 subsystem.

- In the 1H/2009, the DS8000 also introduced a new function called Remote Pair FlashCopy which enables zCDP for DB2 to operate in a GDPS/PPRC with HyperSwap® environment without causing HyperSwap to become disabled if both the production source volumes as well as the DFSMShsm Fast Replication target volumes are all PPRCed and managed under GDPS/PPRC HyperSwap.



zCDP for DB2

Challenges

- Eliminate Backup windows in a z/OS environment for DB2.
- SAP recommended Backup methodology.
- Need to roll back to any point in time or customer defined Events due to logical/software corruption or similar event
- Provide Continuous Backup Protection for DB2 Data.
- Reduce "Recovery Time" for Data Corruption Events.
- Roadmap focus: Protection/Detection/Isolation & Real Time Repair

Value

- Elimination of of Backup Window.
- Minimal impact to SAP & DB2 applications
- Continuity of business operations
- Physical Volume Backup, Physical System/Volume or Data Table Restore.
- 50 versions of DB2 data maintained by HSM across Disk & Tape.
- Any PiT Restore of Data for "Logical" Data Corruption.

IBM Solution Stack

- System z Servers w/Parallel Sysplex
- DS8000/DS6000/ESS
- Flashcopy, SnapShot
- Integrate w/DFSMS & DFSMShsm
- Integrate w/DB2/HSM Storage Pools
- Managed via DB2 Backup/Restore Utilities, HSM Cmds.
- Base Support zOS 1.5 + DB2 V8
- zOS 1.8 DFSMShsm Fast Replication Enhancements
  - Managed Across Disk & Tape
  - Table Space Restore
- DB2 9 exploits zOS 1.8 enhancements

Target Customers

- z/OS accounts needing more granular recovery times and points for "DB2" Volumes.
- zOS accounts that need to reduce "Recovery Times" for Data Corruption Events.
- zOS customers looking to eliminate DB2 backup windows.
- Recommended Method for SAP Customers

Backup windows will continue to be a focus area for z/OS and the DS8000 & Tape Storage Subsystems.

Other examples of FlashCopy exploitation are found between the DS8000 Storage Subsystem and DFSMSdss, as well as within client Batch scheduling as follows:

DFSMSdss Fast Defrag – Exploits the FlashCopy Extent level function to move extents around as part of the DSS DEFRAG function. This synergy has drastically reduced the time it takes to perform volume Defrags on z/OS.

Fast Restart on Batch failures. Several customers now FlashCopy all data before batch, then use the FlashCopy Fast Reverse restore function to restore the production volumes to the state they were in before the batch run on a batch failure. This enables a fast, automated Batch restart capability.

Clients also exploit FlashCopy to provide a Fast Restart point within Batch processing. A FlashCopy is taken just before the start of Batch or during various points within the Batch process.. If a failure is encountered a fast FlashCopy Reverse Restore operation is done and the batch process is quickly restarted.

Storage Based Data Replication

In 1995, IBM announced storage based data replication functions. Extended Distance Remote Copy (XRC now z/OS GM), Peer to Peer Remote Copy (PPRC, now Metro Mirror)) and the data migration solution of Extended Distance PPRC (PPRC-XD, now Global Copy) became the basis for the various Parallel Sysplex “data” portions of the High Availability and Disaster Recovery solutions exploited today by many clients. Storage based data replication, introduced multiple copies of data, created without the knowledge of the application(s), data base(s) and most z/OS components. As a result, the storage based data replication techniques now owned the responsibility of data integrity and cross volume/cross storage subsystem data consistency issues. The concept of “Consistency Groups”, initially introduced with XRC in 1995 was created to manage dependent I/O sequences originally coded into ‘software’ (applications, data bases, file systems and the operating system) to help manage data consistency issues of data stored on disk across power failures. The “PPRC Data Freeze” concept provided consistency group processing for synchronous data replication techniques and was introduced into zOS in 1997. Here again, z/OS software and storage controller hardware synergy combined to enable and manage multiple copies of data created, outside of the host software with full data integrity and cross volume/cross storage subsystems data consistency.

Storage Based Data Replication functions along with numerous features built on the functions and various combinations of functions has provided the basis for many improvements for high availability (ex. HyperSwap), disaster recovery (GDPS & GDOC) and data backup protection across multiple sites. This of course also includes improvements in the simple backup, exploiting FlashCopies of data, Point in Time FlashCopies of sets of volumes as well as Consistent FlashCopies of volumes. Integration of many of these functions has occurred to improve backups as shown in the zCDP for DB2 solution previously mentioned.

The following three charts outline the various Storage Based Data Replication functions available today. The first charts also lists data replication functions originally invented and shipped with the high end IBM System Storage and now also available with various IBM mid-range storage products.

The DS8000 has introduced several recent Data Replication features built to assist various client configurations. These include:

- **Extended Distance Ficon** – Support provides additional client's choice for long distance channel extenders as the standard frame forwarding channel extender technology can now obtain similar performance with the channel extenders that internally optimize the Read Record Set CCW chain I/O.

-**Offloading the XRC SDM Mips to zIIP Engines** – When the XRC System Data Mover host address space is dispatched by z/OS, it can now be dispatched on general purpose CPs or on zIIP engines.

- XRC Multiple Reader Support – Multiple XRC Readers can now be used to offload updates for a single XRC session on an LSS. With larger volume support (ex. 3390-9. 3390-27, 3390-54 & EAVs) this support helps to optimize the data transfer between the DS8000 and the System Data Mover.

- GDPS/MzGM HyperSwap Incremental Resync Support – On a HyperSwap operation, the XRC session that was previous running between the source and a remote volume is now switched to the HyperSwap target and the XRC session is then reconnected and resync'ed from hardware bitmaps to the remote volume.

- Remote Pair FlashCopy – When one does a FlashCopy from a Metro Mirror (MM) primary volume to a FlashCopy target that is also a MM primary volume, no data is transferred to the MM target volume. Instead the FlashCopy command is sent to the Metro Mirror target volume and executed as part of the synchronous MM operation against both the MM Primary and target volumes. This enables HyperSwap to stay enabled throughout this operation when the volumes involved are also in a HyperSwap configuration.

All of the new DS8000 features were developed in response to client requirements. IBM spends a great deal of time working with and listening to clients. Learning how and why technology is being used to meet various clients' business requirements is an ongoing process. This is a great source of innovation. IBM would encourage any client or potential client faced with a business challenge in this area to talk with their local IBM team. This can result in various solutions, like those mentioned throughout this paper.

IBM Copy Services Technologies

FlashCopy

- Point in time copy
- DS8K FCSE
- Available on:
  - DS8000, DS6000, SAN Volume Controller
  - DS4000/DS5000
  - N Series
  - XIV

Metro Mirror

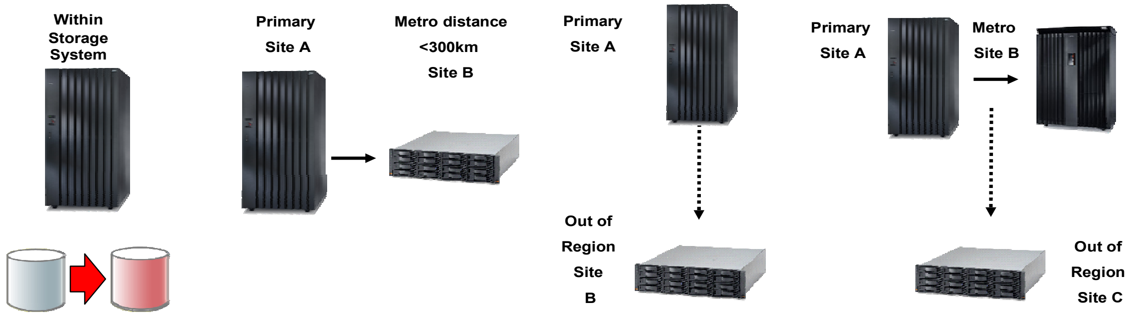
- Synchronous mirroring
- DS8K HyperSwap
- Available on:
  - DS8000, DS6000, ESS
  - SAN Volume Controller
  - DS4000/DS5000
  - N Series
  - XIV

Global Mirror

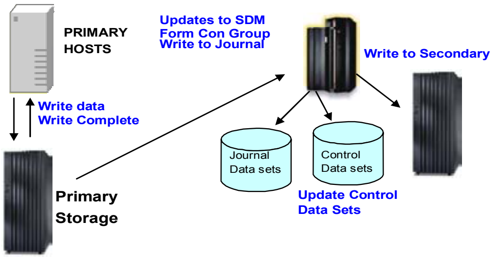
- Asynchronous mirroring
- Available on:
  - DS8000, DS6000
  - SAN Volume Controller
  - DS4000/DS5000
  - N Series

Metro / Global Mirror

- Three site synchronous and asynchronous mirroring
- Available on:
  - DS8000
  - N Series



IBM 2 Site z/OS Global Mirror (XRC)

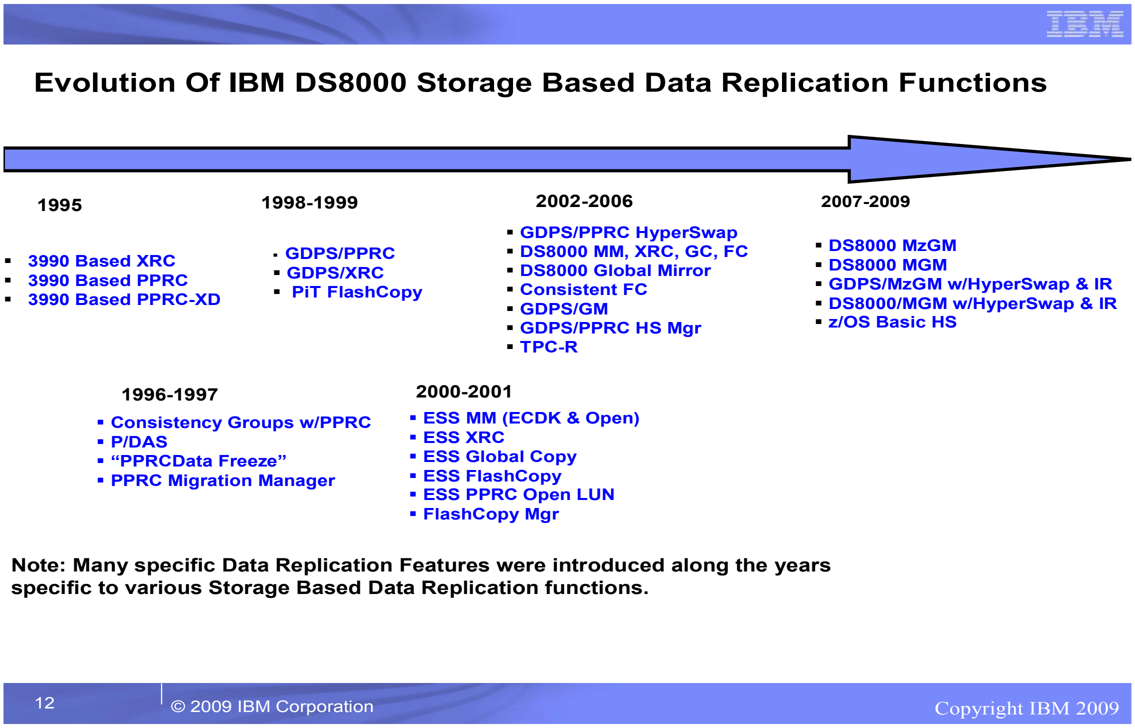


Continues to Provide:

- **Premium performance & scalability**
  - Data moved by System Data Mover (SDM) address space(s) running on System z (14 SDMs/Lpar).
  - Supports heterogeneous disk subsystems
- **Unlimited distances**
- **Time consistent data at the recovery site**
  - RPO within seconds
- **Supports zOS and zLinux data**
- **200+ installations worldwide**
- **Real & historical Monitor**
- **Proven, Scalable, Repeatable, Auditable..**

Z/OS Global Mirror (XRC) clients include:

- **Major EMEA Banks in Germany, Scotland, Italy, Turkey, Greece, Spain**
- **Major US Banks/Brokerages/Insurance Co's**
- **Major Banks in Taiwan, Japan, China, Thailand, Korea**



Evolution Of IBM DS8000 Storage Based Data Replication Functions

1995

- 3990 Based XRC
- 3990 Based PPRC
- 3990 Based PPRC-XD

1996-1997

- Consistency Groups w/PPRC
- P/DAS
- "PPRCData Freeze"
- PPRC Migration Manager

1998-1999

- GDPS/PPRC
- GDPS/XRC
- PiT FlashCopy

2000-2001

- ESS MM (ECDK & Open)
- ESS XRC
- ESS Global Copy
- ESS FlashCopy
- ESS PPRC Open LUN
- FlashCopy Mgr

2002-2006

- GDPS/PPRC HyperSwap
- DS8000 MM, XRC, GC, FC
- DS8000 Global Mirror
- Consistent FC
- GDPS/GM
- GDPS/PPRC HS Mgr
- TPC-R

2007-2009

- DS8000 MzGM
- DS8000 MGM
- GDPS/MzGM w/HyperSwap & IR
- DS8000/MGM w/HyperSwap & IR
- z/OS Basic HS

Note: Many specific Data Replication Features were introduced along the years specific to various Storage Based Data Replication functions.

12 © 2009 IBM Corporation Copyright IBM 2009

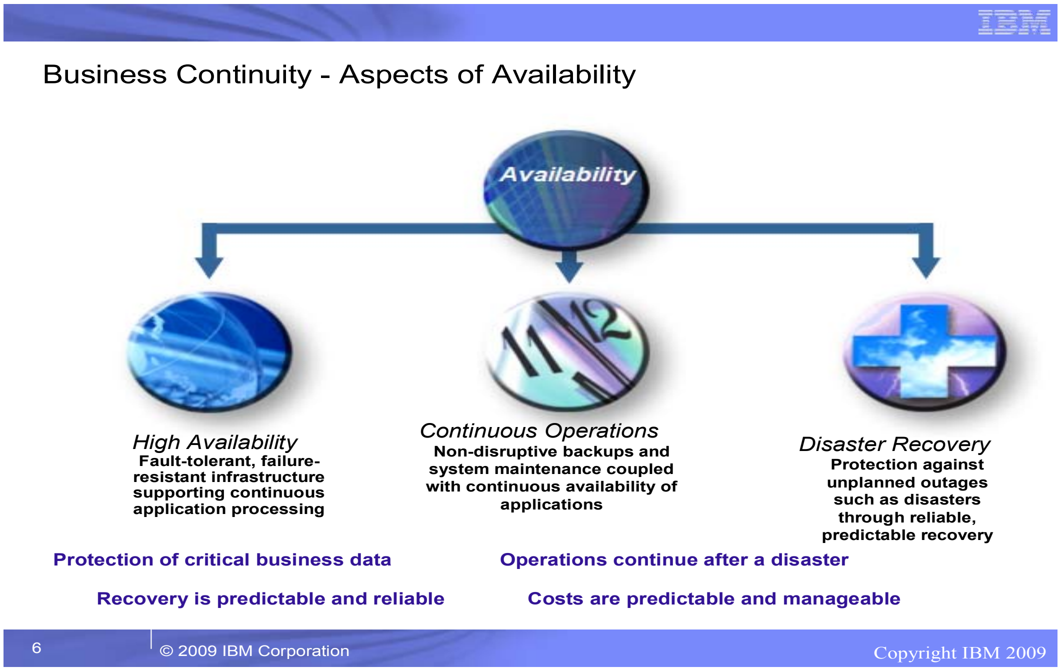
GDPS

GDPS (Geographically Disperses Parallel Sysplex) shipped originally in 1998 and introduced the concept of multi-site IT Infrastructure resource management, for the Sysplex. GDPS extended the Parallel Sysplex management to an end to end "server, workload, data with a coordinated network switch solution" while providing increased business continuity concepts for clients.

GDPS is Storage Vendor independent as all major storage vendors on the System z platform can participate in solutions using their implementation of the IBM DS8000 Disk Storage Subsystem data replication architected solutions of Metro Mirror, FlashCopy and zGM (XRC). New features and functions are developed with the IBM Systems Storage team on the DS8000. IBM sells the Host to Storage Subsystem architecture to the other storage vendors, who implement the feature/function on their disk subsystems. Depending on the specific feature/function there generally is some time where the feature/function is only available on the DS8000. One should consult with each storage vendor to understand when they will support any existing or new storage subsystem enhancement.

In addition, the GDPS automation inter-operates with all major system automation packages available for System z.

IT Infrastructure Availability can be broken down into three pieces; High Availability, Continuous Operations and Disaster/Recovery. Each brings unique client requirements to the table when addressing Business Continuity. Through an understanding of the client business requirements in this arena, IBM can help to tailor the right solution at the right cost point for any IT infrastructure.



Business Continuity - Aspects of Availability

Availability

High Availability
Fault-tolerant, failure-resistant infrastructure supporting continuous application processing

Continuous Operations
Non-disruptive backups and system maintenance coupled with continuous availability of applications

Disaster Recovery
Protection against unplanned outages such as disasters through reliable, predictable recovery

Protection of critical business data

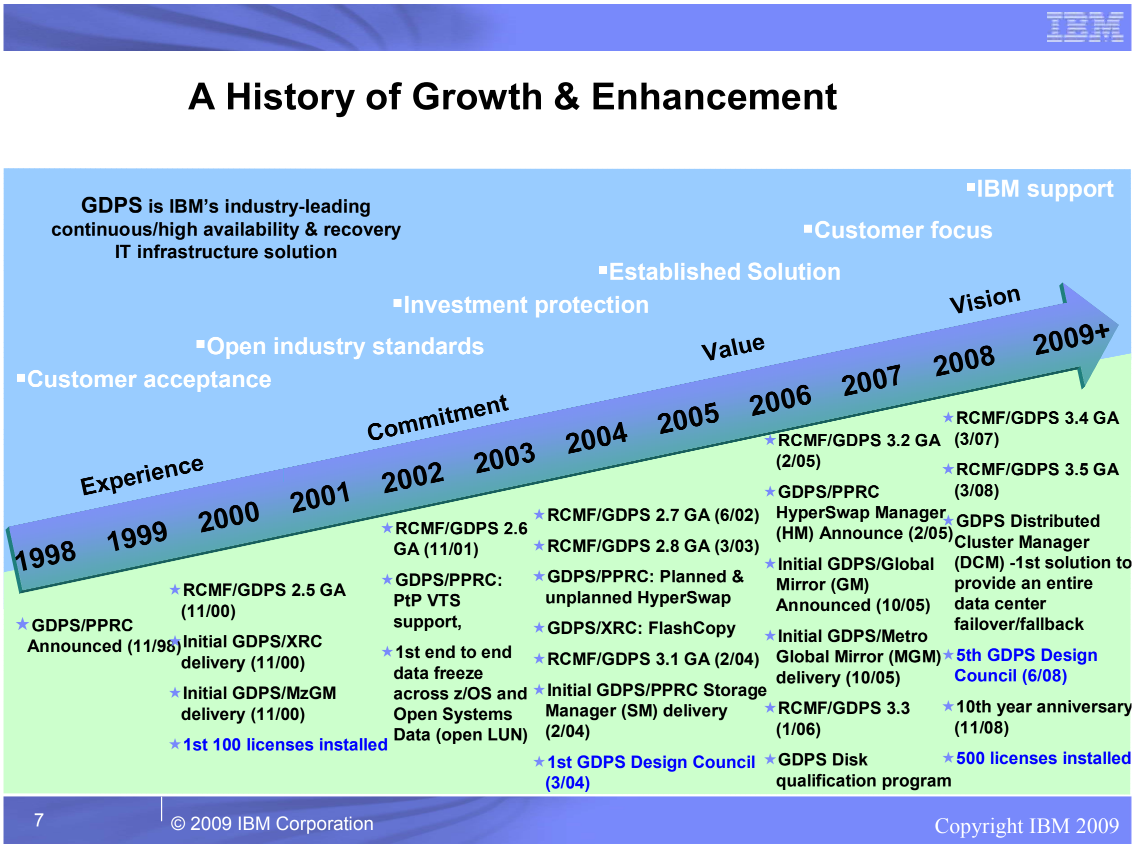
Operations continue after a disaster

Recovery is predictable and reliable

Costs are predictable and manageable

6 © 2009 IBM Corporation Copyright IBM 2009

The following chart detail the history of GDPS over the last 11+ years.



A History of Growth & Enhancement

GDPS is IBM's industry-leading continuous/high availability & recovery IT infrastructure solution

- Customer acceptance
- Open industry standards
- Investment protection
- Established Solution
- Customer focus
- IBM support

Experience
Commitment
Value
Vision
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009+

- GDPS/PPRC Announced (11/98)

- RCMF/GDPS 2.5 GA (11/00)
- Initial GDPS/XRC delivery (11/00)
- Initial GDPS/MzGM delivery (11/00)
- 1st 100 licenses installed

- RCMF/GDPS 2.6 GA (11/01)
- GDPS/PPRC: PtP VTS support,
- 1st end to end data freeze across z/OS and Open Systems Data (open LUN)

- RCMF/GDPS 2.7 GA (6/02)
- RCMF/GDPS 2.8 GA (3/03)
- GDPS/PPRC: Planned & unplanned HyperSwap
- GDPS/XRC: FlashCopy
- RCMF/GDPS 3.1 GA (2/04)
- Initial GDPS/PPRC Storage Manager (SM) delivery (2/04)
- 1st GDPS Design Council (3/04)

- RCMF/GDPS 3.2 GA (2/05)
- GDPS/PPRC HyperSwap Manager (HM) Announce (2/05)
- Initial GDPS/Global Mirror (GM) Announced (10/05)
- Initial GDPS/Metro Global Mirror (MGM) delivery (10/05)
- RCMF/GDPS 3.3 (1/06)
- GDPS Disk qualification program

- RCMF/GDPS 3.4 GA (3/07)
- RCMF/GDPS 3.5 GA (3/08)
- GDPS Distributed Cluster Manager (DCM) -1st solution to provide an entire data center failover/fallback
- 5th GDPS Design Council (6/08)
- 10th year anniversary (11/08)
- 500 licenses installed

7 © 2009 IBM Corporation

Copyright IBM 2009

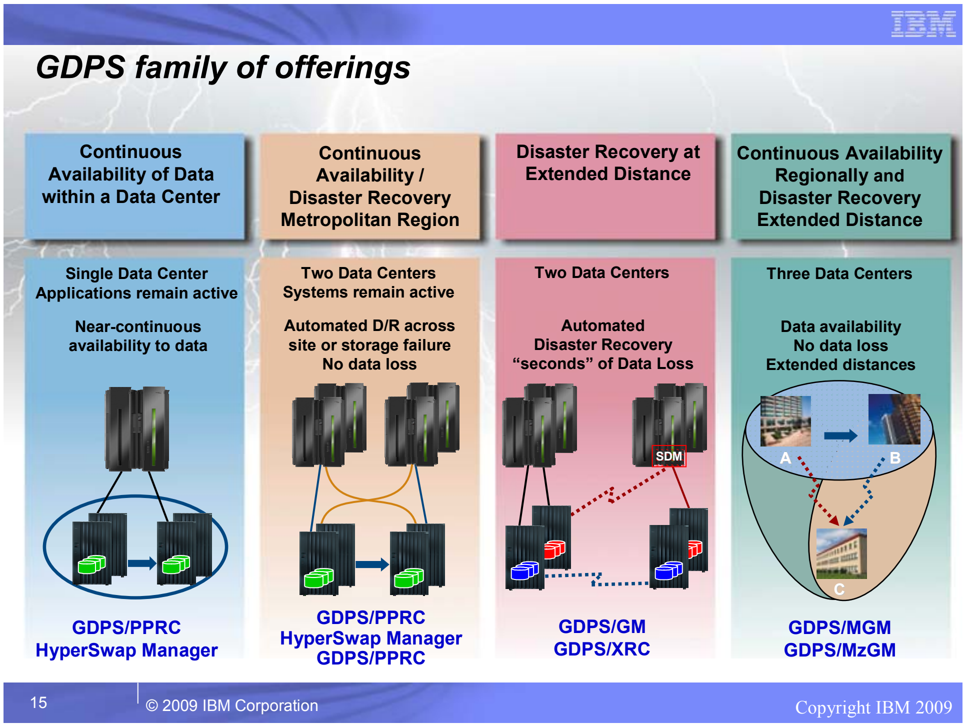
Relative to Business Resiliency/Business Continuity, IBM's Flagship products are GDPS & GDOC. The following two charts illustrate the various solutions that IBM has in these arenas.

GDPS provides the GDPS HyperSwap manager, a low cost entry level solution focused on providing the HyperSwap availability solution on the same data center floor or across two local area data centers up to 200km with parallel Sysplex. GDPS HyperSwap Manager actually provides the HyperSwap function for z/OS, zLinux and zVM volumes, masking storage subsystem failures.

GDPS/PPRC HyperSwap is the Full Function version of HyperSwap Manager, which one can easily upgrade to. In addition to masking disk subsystem failures the full function version, exploits parallel SYSPLEX to mask CEC failures, persistent sessions to coordinate a network switch, CF Duplexing and VTS PtP and TS7700 Grid to mask tape subsystem failures. Finally, if the failures turn into a disaster scenario, GDPS can provide a complete end to end site failover/fall back capability for both planned and unplanned site switches. One mouse Click and the Server, Data, workload and a coordinated Network switch occurs, everything recovered, the SYSPLEX IPL'ed, data bases restarted as well as the applications. No more are skilled personnel required in the event of a disaster.

GDPS/GM (System z & Open Systems data) & GDPS/XRC (z/OS & zLinux only) provide site failover/failback (FO/FB) out of region exploiting IBM's Global Mirror and XRC data replication technologies.

GDPS/MzGM and GDPS/MGM provide high availability locally and out of region D/R protection, fully automated, proven, auditable, and in the case of PPRC and XRC storage vendor independent !



GDPS family of offerings

Continuous Availability of Data within a Data Center	Continuous Availability / Disaster Recovery Metropolitan Region	Disaster Recovery at Extended Distance	Continuous Availability Regionally and Disaster Recovery Extended Distance
Single Data Center Applications remain active Near-continuous availability to data	Two Data Centers Systems remain active Automated D/R across site or storage failure No data loss	Two Data Centers Automated Disaster Recovery "seconds" of Data Loss	Three Data Centers Data availability No data loss Extended distances
GDPS/PPRC HyperSwap Manager	GDPS/PPRC HyperSwap Manager GDPS/PPRC	GDPS/GM GDPS/XRC	GDPS/MGM GDPS/MzGM

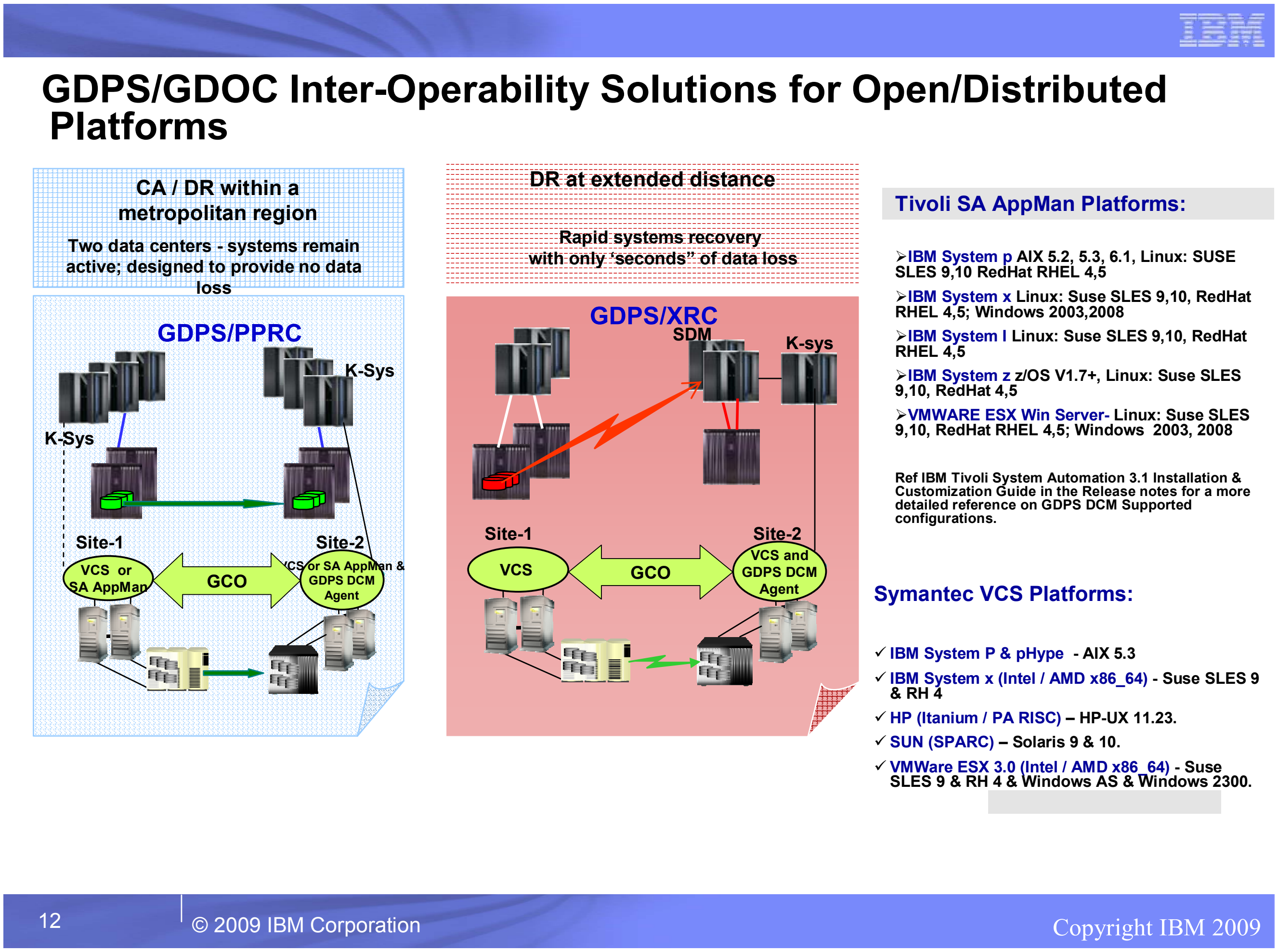
15 © 2009 IBM Corporation Copyright IBM 2009

The GDPS System z umbrella also includes the ability for GDPS automation to inter-operate with System p, x, i (Linux), Windows, HP and Sun through inter-operability code with Tivoli SA AppMan and the Symantec Vertias Cluster Server Solutions. Now, a single button can yield a coordinated site failover/fall back of all of the customers systems. (ex. System z (z/OS, zLinux, zVM) coordinated with say System p AIX systems). The disk replication functions can be managed separately with GDPS and GDOC or together.

GDPS is build upon the IBM DS8000 Storage based data replication architecture for FlashCopy, Metro Mirror, z/OS Global Mirror and Global Mirror. As new features and functions are implemented in the DS8000, GDPS automation is modified to exploit those features and functions. In addition, GDPS supports various DS8000 base box features used in conjunction with the various advanced functions.

IBM DS8000 Metro Mirror and Global Mirror support a function known as 'Open Lun Support', such that through an ECKD device address, GDPS automation is able to manage the Metro Mirror and/or Global Mirror distributed systems Luns. This is also true for Metro Global Mirror configurations. With the Open Lun support, GDPS can provide a single restart point across the platforms. More systems and data replication alternatives will continue to be addressed in this space. This is especially important for customers that have Multi-Platform Applications where transactions are for example initially received by a Windows system, then routed to say an AIX system and then to the backend z/OS System. Each system may save data and as a result to recover the application, all three platforms must be recovered to the same point in time. GDPS inter-operability with Tivoli AppMan and/or Symantec Veritas Cluster Server can provide such a solution for clients.

This function is also important at a number of our accounts especially with applications like SAP where the user interfaces are typically on non System z platforms and the backend data base runs on z/OS. In some cases clients have moved the applications pieces that were running on non-System z platforms to zLinux, but many clients are not willing to introduce change and risk associated with any change to critical production applications that have been running for some time. This functionality is important for application(s) that span multiple platforms. All data is recovered to a single point in time enabling each platform's data base to perform a data base Restart operation instead of a data base recover operation when a site switch occurs. The data base restart process manages all "in flight" and "in doubt" transactions, which in turn permits the application(s) pieces spread across the different platforms to resume processing from the recovered point in time forward. GDPS automation when combined with the GDOC automation can inter-operate across the enterprise to provide a real complete Business Solution for our clients in the area of Business Continuity. This critical business function is made possible by the DS8000 'open Lun support'.



GDPS/GDOC Inter-Operability Solutions for Open/Distributed Platforms

CA / DR within a metropolitan region
Two data centers - systems remain active; designed to provide no data loss
GDPS/PPRC
K-Sys
K-Sys
Site-1
Site-2
VCS or SA AppMan
GCO
VCS or SA AppMan & GDPS DCM Agent
DR at extended distance
Rapid systems recovery with only 'seconds" of data loss
GDPS/XRC
SDM
K-sys
Site-1
Site-2
VCS
GCO
VCS and GDPS DCM Agent

Tivoli SA AppMan Platforms:

- ➢IBM System p AIX 5.2, 5.3, 6.1, Linux: SUSE SLES 9,10 RedHat RHEL 4,5
- ➢IBM System x Linux: Suse SLES 9,10, RedHat RHEL 4,5; Windows 2003,2008
- ➢IBM System I Linux: Suse SLES 9,10, RedHat RHEL 4,5
- ➢IBM System z z/OS V1.7+, Linux: Suse SLES 9,10, RedHat 4,5
- ➢VMWARE ESX Win Server- Linux: Suse SLES 9,10, RedHat RHEL 4,5; Windows 2003, 2008

Ref IBM Tivoli System Automation 3.1 Installation & Customization Guide in the Release notes for a more detailed reference on GDPS DCM Supported configurations.

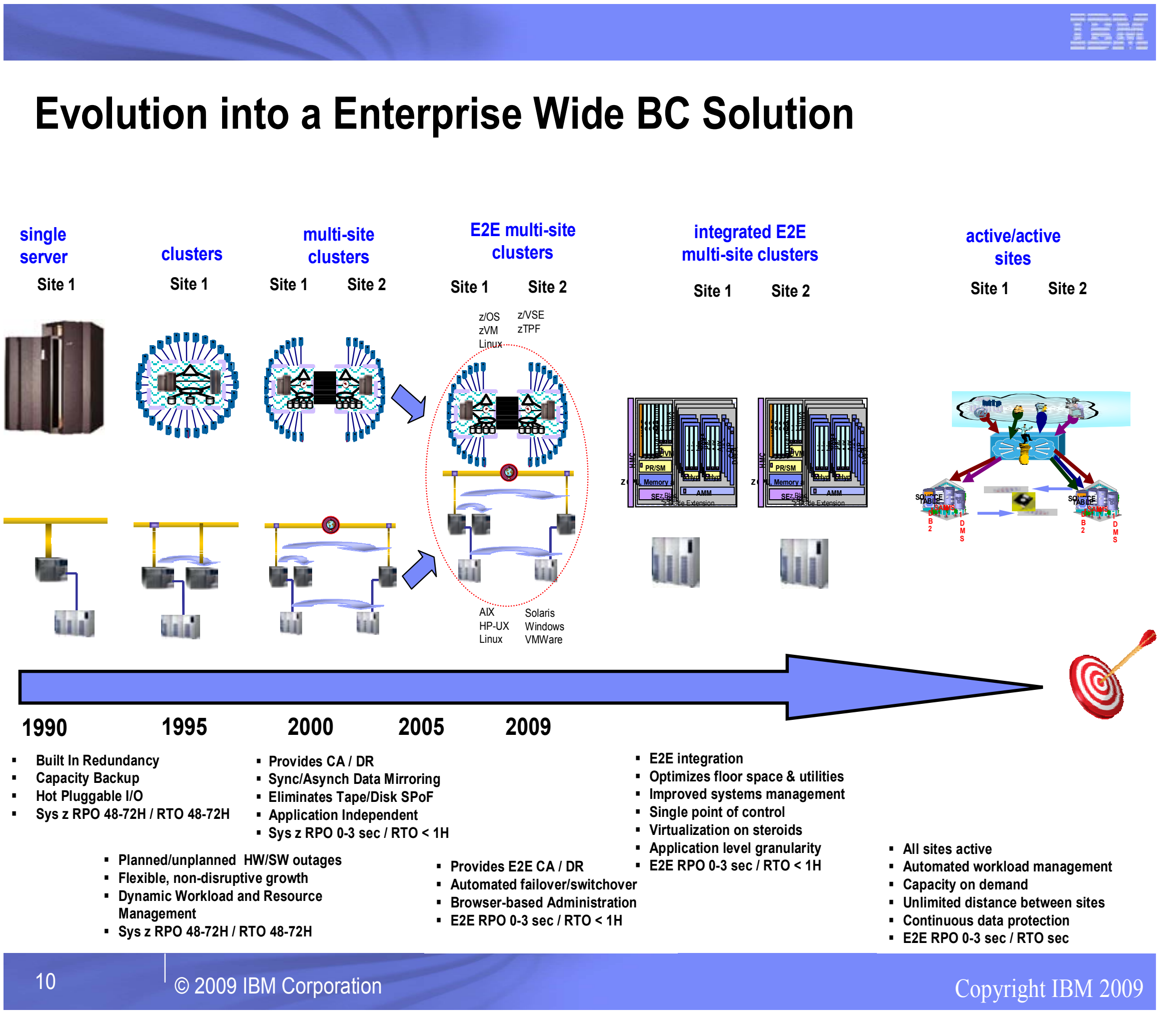
Symantec VCS Platforms:

- ✓ IBM System P & pHype - AIX 5.3
- ✓ IBM System x (Intel / AMD x86_64) - Suse SLES 9 & RH 4
- ✓ HP (Itanium / PA RISC) – HP-UX 11.23.
- ✓ SUN (SPARC) – Solaris 9 & 10.
- ✓ VMWare ESX 3.0 (Intel / AMD x86_64) - Suse SLES 9 & RH 4 & Windows AS & Windows 2300.

12 © 2009 IBM Corporation Copyright IBM 2009

In addition to the GDPS HyperSwap solutions, IBM also introduced z/OS Basic HyperSwap in 2008 providing high availability on the same data center floor for z/OS customers.

The next chart outlines the evolution from a single server into an Enterprise Wide Business Continuity Solution. Single Servers, became clustered servers, clustered servers then spanned physical sites. This was then extended to end to end multi-site heterogeneous clusters, followed by integrated end to end multi-site clusters. The emerging trend is now toward multiple Active/Active Sites at distance.



Evolution into a Enterprise Wide BC Solution

- Built In Redundancy
- Capacity Backup
- Hot Pluggable I/O
- Sys z RPO 48-72H / RTO 48-72H

- Provides CA / DR
- Sync/Asynch Data Mirroring
- Eliminates Tape/Disk SPoF
- Application Independent
- Sys z RPO 0-3 sec / RTO < 1H

- E2E integration
- Optimizes floor space & utilities
- Improved systems management
- Single point of control
- Virtualization on steroids
- Application level granularity
- E2E RPO 0-3 sec / RTO < 1H

- Planned/unplanned HW/SW outages
- Flexible, non-disruptive growth
- Dynamic Workload and Resource Management
- Sys z RPO 48-72H / RTO 48-72H

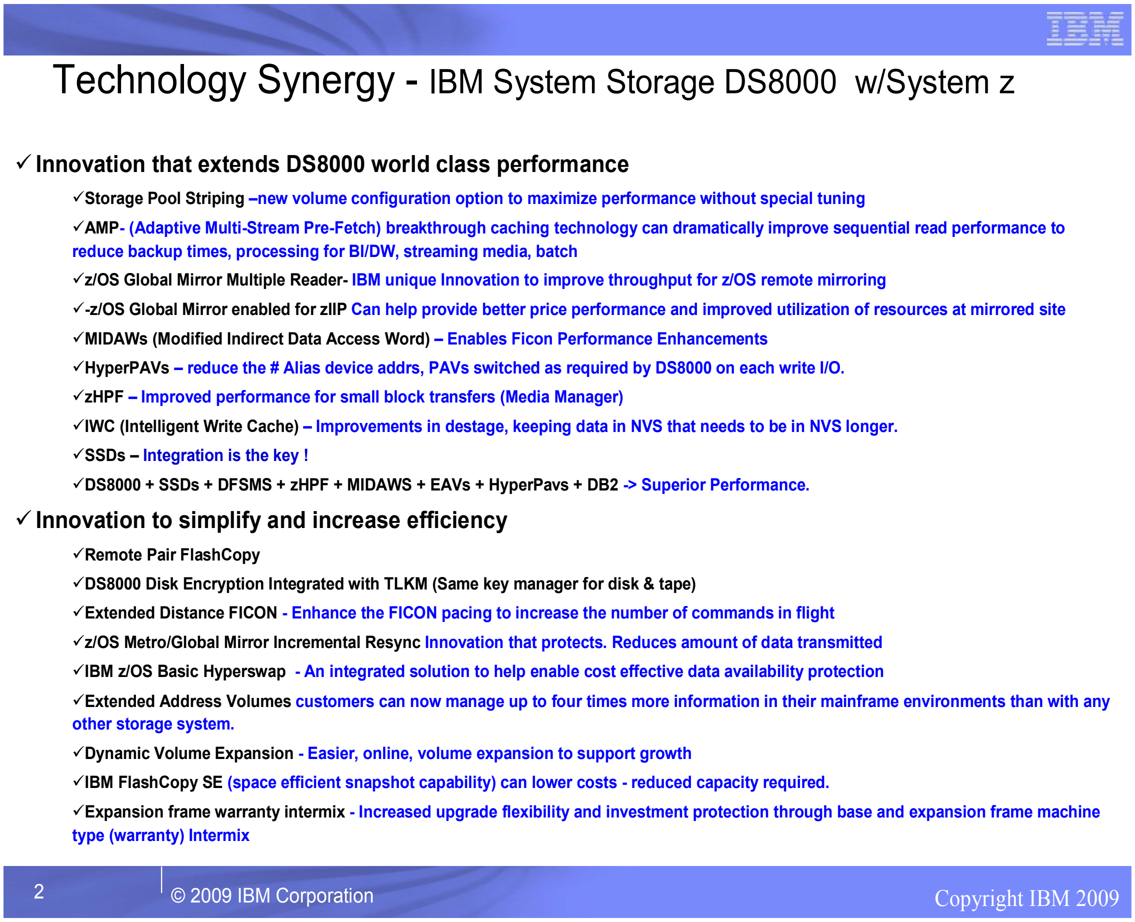
- Provides E2E CA / DR
- Automated failover/switchover
- Browser-based Administration
- E2E RPO 0-3 sec / RTO < 1H

- All sites active
- Automated workload management
- Capacity on demand
- Unlimited distance between sites
- Continuous data protection
- E2E RPO 0-3 sec / RTO sec

10 © 2009 IBM Corporation Copyright IBM 2009

Forty Five Plus years of IBM System Storage and System z synergy history brings us to the present day. Functions being developed today are placed in the appropriate location across the End to End (Network/Server/Channel/Fabric/Storage Subsystem) stack to optimize performance, provide higher levels of security, minimize the movement of data, and provide higher levels of availability and disaster recovery protection. Further, in many cases these synergy items have been extended to incorporate various distributed systems in the Enterprise yielding solutions to manage the Dynamic IT Infrastructure while providing multiple qualities of service levels to the clients of the IT organization.

The following chart is a point in time picture of some of the current IBM DS8000 System Storage and System z synergy items. Most of these items have already been discussed throughout this paper. The specific items on the list will continue to change and evolve in the future as new client requirements are addressed meeting the future demands placed on IT. As one can see, over the past 45+ years, the synergy between IBM System Storage, represented today by the IBM DS8000 and System z, and in particular the z/OS operating system has evolved to provide the highest levels of performance, availability, security, and new functions required by today's IT organizations, which in turn have become the key "manufacturing arm" of our clients business, turning raw data into the vital information that each client requires to manage their business.



Technology Synergy - IBM System Storage DS8000 w/System z

✓ **Innovation that extends DS8000 world class performance**

- ✓Storage Pool Striping –new volume configuration option to maximize performance without special tuning
- ✓AMP- (Adaptive Multi-Stream Pre-Fetch) breakthrough caching technology can dramatically improve sequential read performance to reduce backup times, processing for BI/DW, streaming media, batch
- ✓z/OS Global Mirror Multiple Reader- IBM unique Innovation to improve throughput for z/OS remote mirroring
- ✓-z/OS Global Mirror enabled for zIIP Can help provide better price performance and improved utilization of resources at mirrored site
- ✓MIDAWs (Modified Indirect Data Access Word) – Enables Ficon Performance Enhancements
- ✓HyperPAVs – reduce the # Alias device addrs, PAVs switched as required by DS8000 on each write I/O.
- ✓zHPF – Improved performance for small block transfers (Media Manager)
- ✓IWC (Intelligent Write Cache) – Improvements in destage, keeping data in NVS that needs to be in NVS longer.
- ✓SSDs – Integration is the key !
- ✓DS8000 + SSDs + DFSMS + zHPF + MIDAWS + EAVs + HyperPavs + DB2 -> Superior Performance.

✓ **Innovation to simplify and increase efficiency**

- ✓Remote Pair FlashCopy
- ✓DS8000 Disk Encryption Integrated with TLKM (Same key manager for disk & tape)
- ✓Extended Distance FICON - Enhance the FICON pacing to increase the number of commands in flight
- ✓z/OS Metro/Global Mirror Incremental Resync Innovation that protects. Reduces amount of data transmitted
- ✓IBM z/OS Basic Hyperswap - An integrated solution to help enable cost effective data availability protection
- ✓Extended Address Volumes customers can now manage up to four times more information in their mainframe environments than with any other storage system.
- ✓Dynamic Volume Expansion - Easier, online, volume expansion to support growth
- ✓IBM FlashCopy SE (space efficient snapshot capability) can lower costs - reduced capacity required.
- ✓Expansion frame warranty intermix - Increased upgrade flexibility and investment protection through base and expansion frame machine type (warranty) Intermix

2 | © 2009 IBM Corporation Copyright IBM 2009

Author

Bob Kern - IBM Advanced Technical Support America's (bobkern@us.ibm.com). Mr. Kern is an IBM Master Inventor & Executive IT Architect. He has 36 years experience in large system design and development and holds numerous patents in Storage related topics. For the last 28 years, Bob has specialized in disk device support and is a recognized expert in continuous availability, disaster recovery and real time disk mirroring. He created the DFSMS/MVS subcomponents for Asynchronous Operations Manager and the System Data Mover. Bob was named in 2004 a Master Inventor by the IBM Systems & Technology Group and is one of the inventors of Concurrent Copy, PPRC, XRC, GDPS and zCDP solutions. He continues to focus in the Disk Storage Architecture area on HW/SW solutions focused on Continuous Availability, and Data Replication. He is a member of the GDPS core architecture team and the GDPS Customer Design Council with focus on storage related topics.