TANEJA GROUP
TECHNOLOGY ANALYSTS

TECHNOLOGY IN DEPTH

The State of the Core – Engineering the Enterprise Storage Infrastructure with the IBM DS8000

April 2010

THE NET NET

Hailing from the early days of mainframes, enterprise-class storage arrays may well be one of the longest operating, consistently recognizable IT systems around. And yet if IT managers are asked today "what is the most important, core technology within your data center?" most will answer that it is their enterprise-class storage. The very heartbeats of many mission critical applications within the enterprise still depend upon such arrays, and it is little wonder that this is the case. Extensive engineering over time – often a span of decades – has taken the enterprise-class storage array into a clearly different realm of availability and performance that simply cannot be challenged.

But today, the story about why the enterprise-class array retains its place in the data center no longer stops there. In the face of the ever growing data center and never ending application demands, delivering availability and performance simply can no longer be done with a bunch of monolithic boxes that act as disinterested third parties serving up unintelligent storage blocks that are scattered across 10s if not 100s of monolithic arrays. The "core" of IT – the enterprise storage array – is being forced to change. Storage simply needs to be better coupled with the operating systems and applications that it supports, and better managed across the distributed enterprise. But not everyone realizes this in equal measure, and in this dimension, clear differentiation begins to show between the enterprise-class arrays on the market.

In this Technology Brief, we'll examine how the requirements for enterprise-class storage are changing, and are now equally fixated upon integration that enhances the performance, efficiency and value of storage features. Then with this in mind, we'll examine how one set of solutions, the IBM DS8000 series arrays, are delivering a storage foundation to help customers extend the storage capabilities of their entire infrastructure, well beyond the task of simply serving up tremendous amounts of IO.

The Marching Evolution of the Enterprise Array

Like the advancing tide of an army marching on, the flagship of the storage industry – the mighty enterprise array – seems to always continue its generational progression of adding more storage power and robustness. That power and robustness is not earned overnight – it is the product of decades of engineering over the course of that multi-generational advance. In fact, in an industry rife with startups and innovators, it is testament to the depth of this engineering behind the enterprise array that few true challengers ever arise from the

ranks of startup companies running after their own slice of the storage market.

But just like with the pounding footfalls of an army, it is challenging to navigate changes in direction with the enterprise array. In turn, there are many contentious voices that claim the enterprise array isn't doing its part in keeping up with an enterprise infrastructure that is increasing in complexity seemingly every day. Indeed, many enterprise arrays seem to take their functionality to the point of serving up enormous amounts of IO at their many ports of egress, and then do little more to optimize or enhance storage services beyond this boundary.

It is no wonder that this is seen as ineffective. Storage is with no doubt the foundational system at the very core of today's information-centric enterprise. And in today's enterprise information interactions are more complex than ever before, involving huge data sets, messaging systems, clustering, moving virtualized applications, data sets that more rapidly change in access pattern and importance, and more. A storage infrastructure that simply serves up storage IO may be fundamentally *disconnected* from the information systems of the enterprise.

You may be surrounded by evidence of this disconnect today. Experienced storage administrators well know the challenges around load balancing enterprise storage arrays with hard-to-see storage mappings between hosts, applications, and volumes. Similarly, inefficiencies or differences in host IO stacks have often stood in the way of delivering the full efficiency of an array. Good approaches for prioritizing the way storage requests are handled for mission critical applications have been few and far between. The examples abound, and the reality is, most enterprise class storage arrays have been overly fixated on increasing performance rather than leveraging integration with the information systems of the enterprise to increase the effectiveness of the storage they deliver.

The Enterprise Integrated Array

In contrast, with the right integrations, the storage core within the enterprise could be used to enhance the capabilities of all manner of systems and applications. In this digital age, digital systems are fundamentally bound to their storage, and can be limited or extended by the capabilities of that storage. Integrations could better bring foundational storage capabilities to bear – such as enabling systems to directly interact with protection and replication technologies – but could also help assure the right storage resources are allocated to the right systems without contention, and that each system understands how best to use its storage resources.

From this vantage point, we've long held that the requirements the industry holds in mind for the enterprise array must, and are, changing. While performance and robustness remain primary requirements, the enterprise array can no longer march on in isolation. Storage administrators in the information-centric enterprise must be equally concerned with how storage is

integrated with all manner of enterprise information systems. With the importance of integration in mind, let's take a look at what we see as the evolving requirements for the enterprise-class array.

Built as a Next Generation Array.
The enterprise array must remain specifically engineered to meet a unique set of performance and capacity requirements. The underlying architecture requires advanced storage management that can embrace huge numbers of disks and huge amounts of IO, under varying conditions. Moreover, with the enterprise in mind, the underlying architecture is equally engineered for extreme resiliency, extreme security, and serves up advanced, efficiency enhancing storage features. But while these may remain the fundamental grounds on top of which enterprise arrays remain evaluated; these cannot remain the only measures for a storage system that must be better integrated within an information-centric enterprise. As we see it, there are several key ways in which the enterprise should expect an array to be deeply integrated with the infrastructure, and such integrations make up the rest of our requirements for next generation enterprise storage.

Built for Integrating Availability.
While the enterprise array has long held itself apart by heightened reliability and robust options for data availability, availability can no longer stop with the array itself. Availability in the latest generation infrastructures now requires more integration with the management, communication, synchronization and failover capabilities of many different operating systems and applications. Every such system may have unique approaches to clustering, virtualization, failover and failback, testing, dependency management, and more. Availability may be limited, robustness may be lacking, or storage may even become an obstacle if the enterprise array isn't integrated with the availability architectures of such systems.

Built for Integrated Management.
Each year there are fewer enterprises that gaze upon an array with the intent to purchase only one. As the storage infrastructure becomes larger and more complex, the enterprise array must be closely paired with tools that ease management effort across multiple arrays, and harness those multiple arrays into a singular, orchestrated storage infrastructure. Moreover, the data housed on an enterprise array may be a component of Disaster Recovery (DR) strategies, data protection tasks, provisioning operations for all manner of operating systems (physical or virtual), and even information management decisions. The right management tools alongside performance and capacity metrics can provide valuable insight with which to guide administrator decisions and/or policy engines that may in turn orchestrate an entire infrastructure to keep it optimized in the face of continual change.

Built for Integrated Storage.
Storage is a resource that is inextricably connected with key platforms and systems throughout the enterprise. Storage capabilities and efficiency can be enhanced

by going beyond simply serving up ports and IO, and integrating storage interactions with key operating systems and applications. Such integrations may enhance how applications request IO, and how arrays deliver IO. Other integrations may extend snapshot, availability, or even data tiering mechanisms so that they are executed by or in coordination with applications and their administrators. Such integrations may make better use of storage features and make sure that storage services such as data protection – that were previously conducted in isolation from the application – can now be delegated to the right operational teams and take place in alignment with business requirements.

Built for Adaptation.
Finally, making use of a continuing stream of storage innovations in an enterprise-class array requires an underlying architecture that is built for future capability extensions. Yesteryear's monolithic storage array that remained forever unchanged, required proprietary management interfaces, and operated off of immutable extents and static volume configurations, no longer provides the adaptability and flexibility for the use of future storage technologies – the most eagerly anticipated example may be automatic sub-volume data migration across tiers – announced in April of 2010 under the name of *Easy Tier*. The right internal underpinnings may well determine where one vendor can innovate and keep up with the competition, and where another may require entire new hardware generations.

The Array in the Core

Today, the enterprise storage array market is a largely level playing field where even sub-volume capabilities are becoming mainstream – albeit introduced in different ways by different vendors – and where arrays are becoming largely similar in core storage features. But the customer mileage still varies greatly when the integrations with systems external to a storage array are considered.

As we've surveyed the market with our requirements in mind, we believe one vendor in particular demonstrates unique differentiation in how their storage foundation is integrated into the enterprise. Specifically, IBM has taken a holistic look at how storage interoperates in the information enterprise, and has used this perspective to balance their focus on both internal architecture and external integrations simultaneously. With this in mind, let's take a look at how both of these dimensions are combined behind IBM's DS8000 series arrays to enhance how customers realize benefits from their enterprise storage.

A Building Block for the Core – the IBM DS8000 Series

More than 10 years ago, before the industry as a whole started down the path of storage controllers built out of general purpose processors and off-the-shelf componentry, IBM recognized an opportunity to leverage the unique and considerable processing and software expertise within the IBM product portfolio and apply it to the task of storage.

Specifically, IBM chose to leverage the highly differentiated IO capabilities of its Power platform (RS/6000 at the time). The Power platforms have always been held in high regard for how they harness pure gigaflops capability, cache sizes, and memory bus bandwidths – such resources that were in the day 4 to 5 times greater than that available with other industry standard processors. That effort nearly 10 years ago was widely known under its internal codename of Shark, and was responsible for design principles that still influence the underlying architecture of the DS8000 series today.

Today, in terms of pure processing, Power and other standard processors are coupled in generational games of performance leapfrog. Yet even so, IBM's shift to storage built upon Power unleashed a flexible and mature software stack, and a sophisticated hardware architecture that continue even now to demonstrate many differentiations. Paired with IBM's clustering and high availability expertise, the Power platform has enabled IBM to build a robust and highly available storage system, with all of the extensibility required to steadily advance in storage capabilities. That steady progress has kept the latest member of the DS8000 series family – the DS8700 – fully competitive in performance, features, and value while competitors have undergone complete shifts in architecture.

Since the heyday of Shark, nearly every IBM competitor has since charged to market with their own storage systems built upon off the shelf componentry, including the likes of EMC and Sun, among others. Now heralded by the industry as breakthrough accomplishments, IBM's ability to turn to standardized componentry inside of a highly available enterprise platform a decade earlier sets the tone for a deeper examination of what is today known as the DS8000 series.

The DS8700 Array

With a vision of storage at the heart of the data center, IBM today delivers their DS8700 as their flagship enterprise-class storage array. The DS8700 is scalable in multiple configurations from as little as less than one TB of storage up to over two petabytes. As a continuation of IBM's Power-driven storage architecture, the DS8700 continues to operate off of dual centralized Power-platform storage controllers (what IBM calls the CEC, or Central Electronics Complex). Dual independent systems are coupled together with high-speed local buses and integrated with individual IO handling PowerPC-based processors (up to 32) distributed throughout the storage array. But with the DS8700, IBM has taken a generational step forward in capabilities by both hardware upgrades within the processing complex and a change in the IO fabric, and consequent attachments between the processing complex and all internal componentry. On the surface, these changes are minor changes, but in terms of capabilities, they continue to keep the DS8700 well in line with competitors. But it shouldn't be overlooked that a good part of the DS8700's appeal rests in how subtle and seemingly small steps forward with features keep the DS8700 in lockstep with

competitors, while competitors undergo significant changes to bring about the same capabilities. This is in fact a key benefit of IBM's choice of architecture.

The Power in consistency

The power processor has been at work within IBM enterprise-class storage for what could be described as 5 generations now. Long ago, IBM's introduction of Power-based storage leveraged a mature and well known processing architecture while giving customers confidence that storage systems would advance in lockstep with the capabilities of what is a key technology in the IBM portfolio. In turn, Power processing advancements keep unlocking advancing memory capabilities, more processing horsepower, and more sophisticated IO. But customers have benefitted from these capabilities while also realizing the enormous stability that comes from a mature software stack that has marched steadily forward in performance and features over more than a decade.

The consistency of this Power-based architecture in fact gives IBM's DS8000 series one of the highest reliability statistics in the field (IBM claims 5 9's reliability from real-world field data), while providing IBM with a microcode upgrade record with nary a hiccup to be found. Compared to competitors' periodic shifts in architecture, and more than once running into feature introductions that delivered less than desirable results, this Power storage platform comparatively marches on into next generation capabilities without a missed step.

The Power in processing

With the latest generational step forward, the DS8700 introduced the first use of the POWER6 processor in IBM's storage products. The DS8700's 2 storage controllers can now each make use of as many as four Symmetrical Multi-Threaded POWER6 4.7GHz processors that can each act as two logical cores, for a total of 8 logical cores per storage controller. Those processor cores are each connected together with a POWER6, ring-topology, high performance processor bus (called a fabric bus), on die L2 cache, and L3 cache that is directly cascaded off of each processor's L2.

It is important to realize that IBM turned to the Power architecture for storage in order to harness the capabilities of a processor architecture that has long resided in their key enterprise systems. In turn, Power and a core set of kernel technology carried over from AIX turns the controllers within a DS8000 series array into a powerful symmetrical multi-processing (SMP) computing complex. With a sophisticated bus architecture connecting these processors, each and every processor can be turned to any given storage task, allowing IBM to utilize this aggregate pool of processing power (up to 16 logical processor cores) much more efficiently than competitors using dedicated IO processors that might be restricted to handling subsets of ports, buses, or disks. Moreover, as we'll discuss, processing power doesn't stop with the controller complex, and includes distributed PowerPC-based processors performing lower-level disk management

for each individual device adapter that attaches to back-end disks, and supporting host-facing IO on host adapters.

Beyond SMP, pure processor speeds, cache sizes, and inter-processor bandwidth improvements, the POWER6 processor is also famous for turning generational processor improvement expectations on their head – POWER6 not only doubled clock speeds but also introduced significant improvements in how instructions and cache interactions are executed. In turn, a doubling of clock rate has more than doubled processing power, and the DS8700 controllers now conduct storage interactions with the power of approximately 120 gigaflops (Linpack HPC benchmark) and with access to a claimed 300 gigabytes per second of data bandwidth on the processor bus. In comparison to these capabilities, the demands of even tremendous amounts of storage IO across 1024 disks will come far from running into processing limitations.

More Power in I/O

But the POWER6 technology behind the DS8700 controllers isn't entirely about instructions and clock cycles. The IO subsystem behind the POWER6 controllers is engineered to extremes for enterprise connectivity, from top to bottom.

In brief, that subsystem starts with changes in the IO plumbing within the POWER6 processor itself, that allows each node to redundantly connect to what has now been upgraded to an internal PCI Express bus and fabric. Then individual devices – including front-end host adapters and back-end device adapters – are plumbed into this PCI Express fabric using a multi-root switching approach that allows either storage controller to access every device in the system. Each and every tier of connectivity within this system cascades into a higher-level tier that is appropriately greater in bandwidth and lower in latency as the IO gets closer to the processors (and we cover this architecture in detail in the extended sidebar: "Pipes and Plumbing").

Moreover, as IO is processed through the system between disks and host ports, handling the IO isn't entirely a job for the storage controllers, and upgrades have been made in this sense too. Specifically, individual device adapters are responsible for handling RAID for groups of disks within the DS8000 series (64 disks per *pair* of device adapters), and these adapters also employ lightweight Power-based processors. The Power processor on these RAID adapters are up to 70% faster in clock speed with the release of the DS8700. These Power-based RAID adapters free the storage array processing complex from device level tasks, and add to the aggregate processing power within the DS8700 array, making the sum total of parts the combination of these distributed Power processors alongside the POWER6 processors within the array controllers.

Within the DS8700, this IO architecture has the effect of ensuring that each level of connectivity has more than ample bandwidth for the total of IO that might need to be processed (under all conditions – even with a controller failure), and that at the end of the chain, the POWER6

processor has more than enough internal bandwidth to process that IO. This stands in stark contrast to other architectures (x86 for example) where IO and bus limitations can introduce bottlenecks that make controller failure intolerable, or make it impossible to make full use of processing capabilities across all controllers.

Altogether, the DS8700's POWER6 processor and IO subsystem architecture has increased IO handling to 1.5 times that of the DS8300 array, and increased serviceable sequential bandwidth to 2.5 times. Among enterprise class storage, such single generation performance increases are uncommon.

The Power of insight

Moreover, the POWER6 processor provides step by step insight into instruction execution, and a myriad of resiliency and reliability features that have nothing to do with bus design, but rather focus on what might go wrong with instructions "inside" the processor – ultimately yet one more cause for IO and data integrity problems that can be the silent enemies of storage administrators on other platforms. A few examples stand out:

- The ability to trace instructions through the processor, monitor the performance of oscillators, caches, or other subcomponents to identify errors, and upon demand reroute instructions around problems (IBM calls this technology First Failure Data Capture or FFDC).
- If processor cores demonstrate errors, POWER6 can dynamically deallocate a core and redistribute instructions, caches, and everything about the instruction context to another processor; all while avoiding disruption to on-going processing.

Turning Power to resiliency

But the innovation around resiliency doesn't stop with just the POWER6 processing pipeline, and other examples abound, a few of which are:

- The tiered cache hierarchy within POWER6 is uniquely protected per tier of cache (giving specialized protection to each of L1, L2, L3, and general memory), protected with redundant copies of data stored across caches and reserved memory space, policed by hardware assisted data scrubbing, and then surrounded by algorithms that can even move data across different processor caches during extreme error conditions or failures.
- Within the bridges at the heart of the DS8700's PCIe buses, the POWER6 architecture leverages additional error handling that can monitor IO streams, and upon detected errors can deallocate controllers, reset adapters, or even failover adapters between controllers and processors. Additionally, this technology that IBM calls Extended Error Handling (EEH) is integrated into IBM system drivers and can be used by the software stack to dynamically and non-disruptively restore PCI access by resetting the bus during any bus errors.

While doubling clock speeds, quadrupling on-die caches, and similar steps represent a huge leap ahead for POWER6 over previous Power processors, IBM's innovation in

processor componentry robustness, failure avoidance, and availability set POWER6 far afield from any other standard processor in the market. POWER6's resiliency architecture is without doubt a significant contributor to IBM's 5 9's reliability claims for DS8000 series arrays.

Historically, storage controllers have often turned to special purpose processors and special purpose software stacks that could deliver high performance and high reliability. The tradeoff has always been extra design effort, and limitations when introducing new features. The Power platform has allowed IBM to harness the versatility of a general-purpose processor alongside their significant Power software expertise while simultaneously providing a processor that is better matched to storage controller requirements than most vendors could obtain with even purpose built processors.

Storage on Power

On top of the considerable hardware capabilities of the Power controllers, IBM has – throughout the life of IBM's enterprise DS series systems – continued to introduce new storage capabilities. In contrast to the competition, IBM's ability to deliver storage capabilities has almost always been rooted in the microcode. The introduction of major features – such as thin provisioned FlashCopies (IBM's snapshot equivalent), new replication capabilities, new host integrations, or even IBM's Easy Tier sub-volume, automated data tiering – have almost always been released with microcode revisions, and have rarely been constrained by the capabilities of the underlying hardware. But sometimes overlooked is how fundamental some of these changes are. In addition to a general level of feature parity with enterprise array competitors, IBM has brought to bear a number of lower level innovations within their microcode, centered around IO caching. Let's take a look.

Cache for innovation

IBM's approach with the DS8700 is fundamentally about integration from top to bottom – optimizing a connected system of storage services from the smallest hardware component to the management frameworks and applications that run on top of an infrastructure built with DS series storage. When it comes to storage specific capabilities, the very smallest yet perhaps most important component where IBM starts optimization and integration is around the cache.

To start with, Power within the DS8000 series gives IBM the horsepower to cache granularly – much more granularly in fact than the competition. In contrast to 16KB to 64KB cache slots found around the industry, the DS8700 caches data in 4KB slots - an IO size that is typical of random access patterns from enterprise workloads, but can be collected into larger aggregates for large block or sequential IO. IBM then makes that granular caching highly efficient through the use of several innovative caching algorithms at work in their microcode, each of which were brought to life through years of on-going work in IBM's research labs.

The first of these algorithms – beyond the caching innovations already at work on the POWER6 processor – is Sequential Pre-fetching in Adaptive Replacement Cache (SARC) that moves beyond the "what was recently accessed" algorithm approaches commonly found in the industry, to deliver caching that can balance its optimizations for complex mixes of both random and sequential IO.

Second, is Adaptive Multi-stream Pre-fetching (AMP)– an algorithm that peers inside individual IO sessions and can optimize caching for each application stream, even when there is a broad mix of IO from multiple applications hitting the controllers.

Finally, tied to these algorithms are a number of extensions that enhance their effectiveness, including Intelligent Write Caching (IWC) that optimizes how cached writes are moved to disk (based on a combination of both Least Recently Used and spatial write grouping) and independently manages write mirroring across caches. Leveraging distributed PowerPC-based processors on host adapters, IWC identifies and simultaneously mirrors writes across the caches within both controller nodes. By policing and only capturing writes with an intelligent host adapter, IWC frees the general cache and internal connectivity from wasting space and bandwidth by eliminating the needless mirroring of read-only data – read data is only copied and stored once. Competitors in contrast are often less efficient in cache write-outs, and often must mirror the entirety of cache (including read cache) as they cannot granularly manage write and read caching separately.

Narrower slotting and these intelligent and aggressive caching algorithms can combine to make the DS8700's caching more efficient than competitors on a gigabyte to gigabyte comparison. Moreover, SARC, AMP, and IWC create an IO optimization foundation that is further extended by a number of DS series storage integrations that tie IO handling to operating system and application-level awareness, including: prioritization of IOs in cooperation with integrated operating systems and applications (a feature IBM calls end-to-end priority); per-stream optimization of AMP caching based upon end-to-end priority; and optimization of caching through hints passed to the controllers within SCSI commands.

It has long been accepted in the industry that enterprise array caching is a critical differentiator around random IO. Perhaps more telling with the DS8000 series is the fact that these integrations carry cache effectiveness even beyond the bounds of random IO. We have seen evidence of more than 20% sequential IO performance improvement through the use of IBM's adaptive caching algorithms and database integrations. This unique IBM leg up on the competition may give the DS8000 series IO handling capabilities beyond that of other enterprise arrays using similar storage hardware.

Pipes and Plumbing

Pipes and plumbing are often touted as key differentiators between enterprise-arrays, but truth be told, in the highly competitive enterprise storage market every vendor in the market engineers their array to provide sufficient internal low latency connectivity for their high performance arrays. But if pipes and plumbing are of interest to you, read on.

When it comes to internal pipes and plumbing, the POWER6 architecture continues in IBM's steady march forward in capabilities. From the processors' shared fabric bus, a latency optimized, bi-directional ring topology GX++ bus is shared among all processors, and connects to adapter cards with external PCIe interfaces. These external PCIe connections are then networked to IO shelves that house front-end host ports and back-end disk ports. In addition, the GX++ bus within the 2 physical DS8700 controller nodes are networked together with an additional high-speed controller-to-controller link (based on IBM's RIO-G) that operates at over 8GBs. Using the multi-root architecture of PCIe Gen2 that allows PCIe resources to be controlled by multiple processing complexes, every 2GBs PCIe storage connection that attaches to the GX++ bus in one controller is effectively accessible at wire-speed to the secondary controller as well. These 2GBs 4x PCIe connections are cascaded into what is a 20GBs GX++ interface on each 2-way POWER6 processor module, using IBMs highly optimized P5IOC2 PCIe bridge that interfaces directly with the POWER6 GX++ bus interfaces, for access to more than 80GBs of total interface bandwidth in a dual 4-way controller system.

Using the shared IO enabled by this multi-rooted PCIe and GX++ bus architecture the full external IO attachments behind either storage controller can be routed across either node without introducing unnecessary latency or bandwidth bottlenecks, and this bus is leveraged during normal operations for tasks such as synchronous write caching. In turn, the full aggregate of system processors are able to be applied to any task on demand, and instantaneously as well as non-disruptively assume control of IO taking place on the other node during any failure. In contrast, competitive approaches to inter-node connectivity often restrict controller-to-controller cooperation, reduce aggregate processor horsepower for any particular IO task, and can introduce massive latency and bottlenecks for any re-routed IO. Moreover, the PCIe connections attached directly into these high speed, low latency GX++ buses can reduce latency compared to the multiple hops associated with controller bridge to controller bridge to processor bus links found in some competitive architectures (as an example, with Ethernet handling on some platforms, direct GX++ interfaces have demonstrated an ability to reduce latency by a factor of more than 4, to less than 25%).

With the introduction of DS8700, 8 IO enclosures with 8 slots each house all front-end host ports and back-end drive ports, and are redundantly connected to both controllers over 4x PCIe links. 8 4-port host adapter cards in each enclosure serve up a combination of 128 front-end host ports and 128 back-end disk ports. Each of the 128 front-end ports can be software configured as 1Gb, 2Gb, or 4Gb FC or FICON ports. On the back-end, 4 port FC adapters are redundantly attached to 16 drive, fully switched disk drive shelves that house dual-ported FC-AL drives. It is important to note, these 4 port

device adapters within the DS8700 use a specially enhanced Power-based RAID processor to perform RAID 5, 6, or 10 for each port and each of those ports are responsible for a single 8 drive disk group. Those individual processors add to the aggregate of entire system IO processing capability.

As we've stated elsewhere, within the DS8000 family, the introduction of next generation technology often happens transparently to customers. This remains the case here, except in terms of performance, where these invisible internal architecture changes have vaulted the performance of the DS8700 1 ½ to 2 ½ times beyond the capabilities of the DS8300, depending on workload. Not bad for what are otherwise transparent changes.

Built from the DS8700: The Storage Foundation

While the capabilities of the IBM DS8700 are clearly enterprise-class, as we've discussed, enterprise requirements for block storage are changing. Today, it takes more than storage IO to support a complex enterprise. While the real changes are afoot in how the array is integrated with other enterprise systems, in the case of the DS8700, such integrations are almost universally built upon three key storage capabilities. While the enterprise-class arrays on the market today stand in near feature parity with each other, we'll next take a look at these three foundational capabilities that make up the pillars on top of which other IBM integrations stand. Then we'll turn to look more specifically at a number of DS8000 integrations that extend enterprise storage capabilities well beyond the boundaries of the array.

The Virtual Volume

Within the DS8000 series, IBM has now long used a virtualized volume construct underneath what they call their logical volume – with the logical volume being the IBM term for what appears to host systems as a LUN or System z volume. The construction of a logical volume starts with on-disk stripes of 64KB (typically) across all active drives in an 8-device drive group (IBM calls these 8 drive groups a *Rank*, and they may be configured as RAID 5, 6 or 10, with different levels of sparing, as desired). Then 1GB groupings (or for the Mainframe reader, the capacity of a 3390 Model 1 device) of stripes across a rank are gathered into what IBM calls an *Extent*. That extent is the building block with which the logical volume is constructed. A single logical volume in turn may be made up of many 1GB extents that are distributed across many different groupings of disks. But IBM hasn't stopped here.

Recognizing that volume virtualization opens a door upon opportunities to better manage logical storage volumes, IBM has begun leveraging IO statistics with new performance analytics to guide optimized volume placement. Statistics for the extents that make up a volume can be sampled over a period of time, and displayed as a "heat map" that shows where excessive IO is pushing the limits of specific disk ranks.

The Extensible Next Generation Block Architecture

We have labeled the volume architecture at work within the IBM DS8000 an Extensible Next Generation Block Architecture. This architecture has integrated the enterprise-array's traditional high horsepower scattered volume slices with larger data aggregates that pave the way for efficient future interaction with data. Today, this architecture is the foundation on top of which IBM is delivering dynamic sub-volume storage optimization that appears to be a clearly differentiated match for enterprise customers. For those familiar with "analysis paralysis," when it comes to optimizing systems, the administrator needs to have enough detail to make their actions have impact, but not so much detail as to get bogged down by the data – 1GB units of analysis look to be a good working size for this purpose. Moreover, when it comes to data movement, these 1GB extents have carefully maintained a balance between moving data granularly enough to have real cost and performance impacts, but not moving data so granularly as to impact more important system processes. In our view, those characteristics are prerequisites for making data optimization real for the enterprise. As we've discussed in the sidebar: *Easy Tier,* page 17, volume virtualization is likely to play an important role in enhancing future IBM storage capabilities.

Under the banner of IBM Easy Tier, this functionality allows administrators to peer into IO patterns in order to harness the full performance and capacity potential of their entire set of disk media, including SSD as well as every tier of rotational disk. With a heat map of unusually "hot" or "cold" volume extents in hand, administrators can identify the ideal placement of new volumes, manually relocate existing volumes, or configure the automated policies that will dynamically tier sub-LUN data across SSD and rotational disks (see Easy Tier sidebar, page 17). IBM's sub-volume virtualization approach combined with Easy Tier allows administrators to spread out IO across a system, flexibly use the entire capacity and performance of each pool of disks, and intermix different types of volumes within a single pool (e.g. space efficient snapshot volumes and full volumes) to share the same disk pools. This results in utilization efficiencies not found in systems with unshared pools.

Fundamentally, the way the DS8000 interacts with storage volumes is designed top to bottom to enable the right granularity in the right layers of the storage system, without introducing unnecessary constraints. The interaction starts with a finer level of cache granularity than a disk stripe. At the next layer, on disk, the architecture has less granularity that better aligns IO with disk characteristics, while distributing IO requests across disks. Then virtualized volume extents introduce even larger data aggregates that further distribute IO across spindles, and are more practical units of interaction (too much granularity could have the effect of encouraging users to manage data placement off of randomized or non-recurrent IO patterns). The resulting system is IO optimized, and has created a

foundation on top of which flexible data management and tiering can take place, without creating excessive load on the array.

The FlashCopy

Another key DS8000 capability, one that also spans almost every other IBM storage platform, is IBM's comprehensive FlashCopy. Between FlashCopy, or the upgrade FlashCopy SE (for Space Efficient), the DS8700 administrator has a full range of cloning and snapshot capabilities. FlashCopy supports full size copies of logical volumes (sometimes called clones), while FlashCopy SE supports space-efficient (thin provisioned) copies of logical volumes (sometimes called snapshots). Such point-in-time FlashCopies include a long list of potential options and configurations, including incremental delta-only FlashCopies, FlashCopy consistency groups, integration with host-side frameworks like Microsoft's Volume Shadow Copy Services (VSS), remote synchronization and execution, and multiple live FlashCopies for any given FlashCopy configuration. Moreover, it should not be overlooked that FlashCopies across multiple IBM platforms can be managed comprehensively with IBM management tools, or host-specific interfaces like FlashCopy Manager.

The Encrypted Disk

Third, with an eye set steadily upon ever increasing security, compliance, and regulatory requirements, the DS8000 series is the product of one of the few efforts to successfully introduce a scalable enterprise architecture for granular and full storage system encryption.

Just like the rest of IBM's data encryption offerings that also encompass tape, file storage, and data transmission solutions, the DS8000 series with Full Disk Encrypting drives depends upon Tivoli Key Lifecycle Manager (TKLM) to serve up a key repository that can handle the key management requirements of even the biggest enterprises. Using a combination of encryption technologies, TKLM stores the high-speed symmetric keys used in data systems behind industry standard layers of Kerberos and asymmetric keys for authentication and access control.

At time of purchase, customers can choose to configure a DS8000 series with IBM's Full Disk Encryption drives and enable encryption. Because of the potential security compromise inherent in a partially encrypted system that dynamically moves drive extents, the first generation of this architecture requires all or nothing encryption for the entire array, and enabling encryption after the fact is a destructive operation. Following enablement, each drive in the array is encrypted with symmetric keys stored within TKLM.

Individual keys are never persistently stored on the DS8000, but are stored, managed, and protected by the TKLM servers, using a combination of the organization's asymmetric keys and drive signatures for access. Using the DS8000 series 256 bit RSA keys, an individual encryption processor on each drive encrypts data transparent to applications and with zero impact on IO and latency. Even during drive failures, disk disposal becomes

instantaneous with no erasure required, and customers are always fully protected from lost or stolen media.

Total Storage

With no doubt, storage features vary by vendor, and every enterprise class array may bring some unique in-array features to the table. IBM has in the past referred to various DS systems as Total Storage. While that moniker has fallen by the wayside, these features keep the DS8700 marching along as a total package designed for IBM's poster picture of the information infrastructure – an infrastructure that is highly reliable, efficient, and secure, and built to serve as a foundation for layering on other "information-centric" enterprise services. Next we'll turn to look at how DS8000 specific storage services are integrated "beyond-the-box" – both horizontally, with other arrays, and vertically, with systems and applications – and can extend the efficiency and the availability of the entire information infrastructure.

Storage Beyond the Box – The Integrated Information Infrastructure

Turning the lens of our examination outward from the DS8000 internals, IBM long ago realized that delivering enterprise storage goes well beyond array architecture, and is equally about how the building block of the DS8000 allows the enterprise to better leverage the storage underneath their key operating systems and applications. In many cases, IBM has achieved this through efficiency improvements delivered by integration – enhancing the management or performance of storage behind key operating systems and applications. But just as often, integrations also extend the capabilities of key operating systems and applications, perhaps by bringing new DR architectures to market, or optimizing how data is moved between systems.

Management that crosses arrays

The first and perhaps most important step in delivering storage beyond single arrays, is making sure that multiple arrays can be managed together in an orchestrated whole. In this sense, the DS8000 series delivers, and can turn centralized or distributed DS8000 arrays into a commonly managed infrastructure.

IBM has surrounded the DS8000 series with management services to fit any style of infrastructure. At the most basic level of GUI management, the DS8000 series comes with a built-in DS Storage Manager GUI, a command-line interface, SMI-S compliant API, as well as FlashCopy management and mirroring.

Moving beyond this basic level of management, the DS Storage Manager GUI is integrated with multiple optional Tivoli Storage Productivity Center (TPC) packages. Packaged with the DS8000 by default is the Basic Edition of TPC (packaged as licensed software, but upgradeable to a fully pre-packaged appliance). TPC Basic Edition can aggregate access to multiple DS arrays anywhere across an enterprise, display those DS arrays on a topology in the context of the rest of the SAN they are connected to,

monitor overall SAN and DS array health and capacity utilization, and increase the efficiencies of management tasks.

At the next level, is TPC Standard Edition (TPC SE) that adds end-to-end insight into the performance, configuration, and efficiency of a storage infrastructure from the host file system and databases down to the fabric, and down to individual DS spindles within the array. To optimize performance based on end-to-end analysis, TPC SE can drive whole volume relocation tasks that can be executed – dynamically and non-disruptively, or at scheduled times – and will soon comprehend IBM's Easy Tier sub-volume data tiering technology for the automatic movement of groups of extents to higher or lower cost groups of disks (see sidebar: *Easy Tier*, page 17*)*. Just as importantly though, TPC SE's insight into what is happening on the array can be an important management tool for root cause analysis and on-going storage system care and feeding – IBM references DS8000 series customers who have demonstrated 25% improvements in mean-time-to-repair by using TPC SE.

Finally, several of these storage management tools can also be extended in capabilities by host-side agents or software packages that further extend DS management functionality. For example, an assortment of lightweight "near-agentless" 5MB Storage Resource Agents (SRAs) and heavier Java-based agents can enhance performance reporting, and FlashCopy Manager software on specific hosts can enable localized management and application integration of FlashCopy with the host and application based snapshot mechanisms from Oracle, DB2, SAP, Microsoft SQL Server and Exchange applications on that host.
In addition to the TPC suite versions we've identified, is one more – TPC "R" version (TPC for Replication) provides specialized management interfaces and a policy engine for replication including internal FlashCopies and remote mirroring features for two and three-site disaster recovery failover scenarios. These replication services are in fact the second important ingredient in further extending storage services beyond single arrays.

Storage that crosses arrays

At the heart of orchestrating and leveraging the value of multiple IBM DS8000 arrays, sits IBM's replication and FlashCopy offerings, and the integrations around these technologies paints the clearest picture of how the IBM DS8000 series stands out from the competition. The variety, depth, and sophistication of features built by combining mirroring, TPC for Replication automation, FlashCopy, and host-side intelligence enable a range of data protection, clustering, and disaster recovery implementations across a breadth of operating systems and applications that is simply unmatched by competitors.

The DS8000 has a complete array of replication services that fill nearly every slot on any enterprise's wish list, and further, can also be integrated together to deliver well beyond the norm capabilities. A review of these capabilities starts with Global Copy – IBM's asynchronous replication service between DS arrays that can copy data

anywhere, irrespective of latency and promises best effort updates for near real time synchronization.

Next on the list is Metro Mirror. Metro Mirror performs volume replication over high speed, low latency connections (up to 300km) synchronously. In turn, Metro Mirror can be easily integrated into a variety of different architectures for long distance high availability or for distance sharing/clustering of resources, and can support various IBM solutions for near-instant failover, across every IBM platform.

And the final replication technology is Global Mirror. In contrast to Metro Mirror's synchronous replication, Global Mirror is asynchronous. But Global Mirror moves beyond the simple data movement of Global Copy by integrating FlashCopies to mark consistent points in time at the remote location. Moreover, Global Mirror stays close to real time, because FlashCopies are not the source of replication. Rather a FlashCopy *instruction* is passed from the master site to the destination site and creates a FlashCopy at the remote site at a point in time at which the replicated data is known to be consistent. In contrast to the competition, this integration between Mirroring and remote FlashCopy execution creates an asynchronous mirroring architecture that is tolerant of out-of-order data. Through that integration, remote replication sites can accept and track out-of-order data, and use FlashCopy intelligence to identify what data has to be received between one FlashCopy and the next to create a next known good consistency point. This makes data at the remote site less susceptible to replication

Easy Tier – a demonstration of extensibility and the optimization of solid-state disk

The virtualized volume layout within the IBM DS8000 series is representative of what the Taneja Group calls Extensible Next Generation Block Architectures, and is a testament to how the right storage architecture can be extended with new technology without reinventing the wheel. The DS8000 virtual volume was implemented more than 6 years ago, yet IBM has just now (in 2010) launched sub-volume data tiering – called Easy Tier – that will dynamically and automatically redistribute the 1GB extents within DS8000 disk pools. Operating across administrator identified pools of storage – including any selected type of rotational disk, as well as SSD disks where customers have it – Easy Tier will automatically reposition data based on IO patterns and disk characteristics with no further administrative interaction. Administrators will be able to use out of the box templates or create their own policies – selecting specific volumes or entire pools, setting performance sampling windows, scheduling movement intervals, and more – to keep volumes within all selected pools automatically optimized. But Easy Tier will also be accompanied by an API that can trigger periodic migrations, allow applications to provide storage tiering suggestions, or enable custom user integrations. That is a sophisticated range of functionality, but just as with most other DS8000 features, Easy Tier will be enabled with a microcode upgrade. The architecture within the DS8000 demonstrates how storage arrays can be extended with new technology. Competitors too often replace hardware instead.

latency, and easier to resynch after outages. With each of these tools, IBM has enabled sophisticated support for consistency groups and support for flexible remote storage system configurations, sophisticated failover and fail-back architectures, including the use of space-efficient FlashCopies.

While these tools are powerful on their own, they can also be combined. One example is a synchronous nearby (300km) Metro Mirror that then in a cascade fashion asynchronously replicates data to a tertiary Global Mirror somewhere far away while issuing FlashCopy instructions to mark known-good data recovery points.

Storage Above the Box

With these management tools and replication technologies in mind, it is clear that multiple DS8000 series storage arrays across an enterprise can be elegantly orchestrated together for a total storage foundation that is more powerful than just the sum of its parts. But the IBM DS8000 story is about building an Information Infrastructure doesn't stop with solely these integrations *across* the storage foundation. The story behind the DS8000 series is equal parts about how the storage infrastructure is integrated *above* the storage layer. While some other vendors stop their host integration efforts at the design of multipath drivers, IBM has a comprehensive suite of multipath technology, but goes several steps beyond. While competitors may leverage a varying and limited subset of some of these capabilities with their own storage, it is IBM's focus on integrating storage that has driven the innovation behind these features, and the full set of these features working together (only possible on an IBM storage infrastructure) that paints a picture of a storage infrastructure uniquely integrated with the rest of the infrastructure. Let's look at several examples:

OS and application acceleration. We have previously discussed cooperative caching (the SCSI commands with which applications like DB2 on AIX can help optimize caching) as a unique, performance accelerating application integration. In addition, IBM has leveraged new SCSI command standards to pass IO prioritization flags between the DS8000 and some IBM operating systems. The combination of these technologies are enabling the DS8700 to manage QoS across selected sets of IO by allocating resources to specific IOs for specific operating systems *and* applications, and then interactively handle congestion by controlling how unnecessary demands for IO from lower priority systems are treated during peak periods of demand. Finally, when it comes to the Mainframe, it is well known that no other storage vendor accelerates performance like IBM, with a comprehensive set of technologies to accelerate data transfer (like Hyper Parallel Access Volumes, or HyperPAV) and drive latency out of individual connections (with z/OS High Performance FICON, or zHPF). These integrations alongside the DS8700's processing power and bus architecture are behind IBM's statements of up to 2.5x performance improvements over previous DS8000 arrays when supporting real world workloads. We suspect that deeper OS and

application integration has only just begun, and we are looking forward to seeing which domains of the consolidated, increasingly virtualized infrastructure IBM chooses to accelerate next.

Availability integration. Another example revolves around how DS8700 and the availability architectures of different operating systems and applications are integrated. The examples stand out around System i. DS8000 series arrays have unique integrations with System i through the Copy Services Toolkit. This toolkit integrates FlashCopy functionality, and synchronizes all of the DS Series replication capabilities with the System i availability and clustering approaches. This includes availability-specific integrations that work together with a set of already sophisticated set of more general System i storage integrations such as such as support for LPARs and Virtual IO Server partitions, multi-pathing, boot from SAN, and i Application Services. These technologies all have a role to play in a wide range of flexible and sophisticated extended distance clustering and failover / fail-back solutions that are possible through System i and DS8000 series integrations. These take place through LUN level switching capabilities, PowerHA for i (previously HASM), Remote Mirror and Copy (RMC) and Cross Site Mirroring (XSM) technologies that work together with DS8000 Metro Mirror, Global Mirror, and FlashCopy. On AIX, availability includes integration of PowerHA SystemMirror (previously HACMP/XD) with the full range of DS8000 mirroring, including Metro Mirror.

Next Generation Media. With SSD, next generation media is upon us. But how effectively this storage media can be used may well be determined by how well arrays are able to move beyond simply serving up IO and provide access optimization for different storage systems. For example, leveraging the SSD behind mainframes will require a platform that can provide best in class access, and full support for volume virtualization. For the IBM DS8000, such integrations include Hyper-Parallel Access Volumes (HyperPAV) and z/OS Hyper Performance FICON (zHPF). These technologies virtualize and parallelize connections and exponentially improve mainframe *connect* performance. Once the latency of connect is optimized, SSD disk can be leveraged to further optimize disk time, and maximize the benefit of the low latency SSD media. Other integrations – such as DB2's SSD awareness – can place specific data sets or table spaces on SSD media. Moreover, extending SSD benefits in a consolidated infrastructure may come down to just providing pure access versatility, where examples include the ability to attach FC rather than FICON to System z Linux (potentially making SSD volumes easier to access and share), or leverage SSDs behind System i attach (both of which are unique to IBM).

Integration with the Information Infrastructure. It shouldn't be missed that the integrations between the DS8000 series and the rest of the infrastructure do not stop at just these technologies and applications. IBM's integrations around DS8000 also include all of the storage management tools in IBM's arsenal, and

numerous other storage technologies – a few of which include offerings like Tivoli Storage Manager data protection offerings and IBM's SAN Volume Controller (SVC). SVC can leverage common tools and technologies between SVC and the DS8000 series (and other IBM and non-IBM arrays) to orchestrate and automate provisioning, perform volume movement or total migrations, provide enhanced virtual server services (such as integration with VMware's Site Recovery Manager) and serve up consolidated FlashCopy and mirroring functionality even among heterogeneous arrays across what can be transformed into a unified, scale-out storage infrastructure. The previously mentioned Tivoli Storage FlashCopy Manager, when combined with DS8000 series or SVC FlashCopy, is deeply integrated with other TSM data protection technologies. This includes automatically moving FlashCopies to offline storage (backup media) while populating the TSM recovery management (backup) catalog, with no further host involvement. And these are but a few of the examples within the suite of intersecting IBM storage and information management tools, that can include compliance gateways combined with encryption, IBM's Content Collector classification technology, HSM and System Storage Archive Manager, and much more. Each of these provides unique value add when placed on top of the IBM DS8000 series storage foundation and integrated into an end-to-end, managed information infrastructure. With the continued introduction of new DS8000 series features like IBM's Easy Tier automated data tiering and encryption, we only expect to see more integrations with each of these products.

Taneja Group Opinion

The DS8000 continues to demonstrate IBM's expertise in creating unique efficiencies and capabilities through strategic integrations among the whole portfolio of IBM technologies. This is in fact why so many IBM customers are known as IBM shops. System z is often touted as the reason why IBM shops are IBM shops, but that statement doesn't even scratch the surface. IBM shops are IBM shops because of well-engineered, deeply integrated capabilities. The integrations around DS8000 are important testimony to this, and drive home how the DS8000 not only delivers value in packaged capabilities, but also increases that value as it is deployed and integrated with the customer's infrastructure.

While the market may sometimes miss it, IBM has long been focused on integration. In fact, with a focus on integration, they may have been one of the earliest heralds ringing in the message that the enterprise storage array as an ironclad monolith is dead. Moreover, while a number of platforms in the IBM stable have monolith roots – like the infamous mainframe of yesteryear – they have in fact long departed from those single box, inflexible roots. System z is surrounded by logical partitions and virtualization technologies. System i has an LPAR and virtualized IO architecture that can support complex, scale-out-like compute architectures with considerable efficiency. Storage systems across the portfolio are built from standardized componentry, and functionality is extended outside of single storage systems not only

with integration, but also with IBM's market leading SVC virtualization engine, also built on top of IBM's industry standard componentry.

It takes only passing observation to note that there is power in this shift to standard componentry (inclusive of the Power processors that are standard-like for IBM) and the multiple layers of virtualization found within IBM solutions. Purely as proofs of concept, IBM lab exercises often show off the hypothetical capabilities of these architectures, and they can be incredible – simple hardware changes can deliver capabilities that are considered far in the future for other vendors. One such example: IBM's SVC-based demonstrations with some solid state media types that have churned out more than 1 million IOPS; seemingly with little more effort involved than the drop of a hat.

Such feats for us are demonstrations of the merits behind these architectures, and show how IBM is laying a foundation for future innovation as technology trends change. While SSD integration and storage tiering are important first steps, we suspect the future of IBM storage is steeped in the potential for unlocking data optimization and blistering performance like never before.

With this in mind, it is no wonder that the IBM DS8000 series continues to be held in high regard by IBM customers, but we think irrespective of the predominant vendor in an enterprise data center, the DS8700 deserves serious consideration.