# IBM Technical conference

## Buenos Aires 30 Mayo 2013

- IBM Flash System and IBM San Volume Controller

Ing. Federico Lizarralde – Storage Hardware

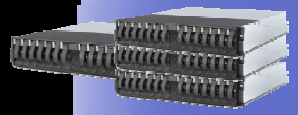
AGENDA :

- Introducción a la tecnología IBM System Flash
- Descripción de los modelos Flash , RAID VSR, modelos y equipamientos.
- Arquitectura interna de IBM System Flash
- IBM San Volumen Controller
- Virtualización en storage, características, funcionalidades
- Integración IBM San Volumen Controller e IBM System Flash.

# Familia de subsistemas de discos IBM System Storage

IBM System FLASH Technology

**Entry point**

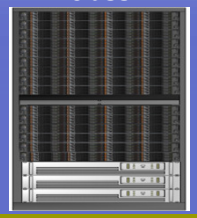


**IBM V3700**

**Unified SAN/NAS series disk for open servers**

**IBM STORWIZE V7000 / SVC**

**Foundation**
**Open Enterprise class**



**XIV Enterprise Open**

**Enterprise-class storage . Leading the industry in functionality, performance, TCO**

**DS8870**

Enterprise-class Storage Continuum

| |
|---|
| **Plataforma de administración unificada** |
| **Servicios de copia unificados** |
| **Virtualization** |
| **Compelling price points** |
| **Lider en la industria de servicios y soporte** |

**IBM System Storage Family ayuda a la innovación:**

- Simplifica **la infraestructura de storage IT permitiendo la administración a bajos costos y compejidad, mientras se incrementa la habilidad de responder ante las necesidades de los cambios.**
- **Asegurar** continuidad del negocio**, seguridad y durabilidad de los datos.**
- **Administración eficiente de la information** a travez del ciclo de vida de los datos (ILM), **relativo al valor de los negocios.**

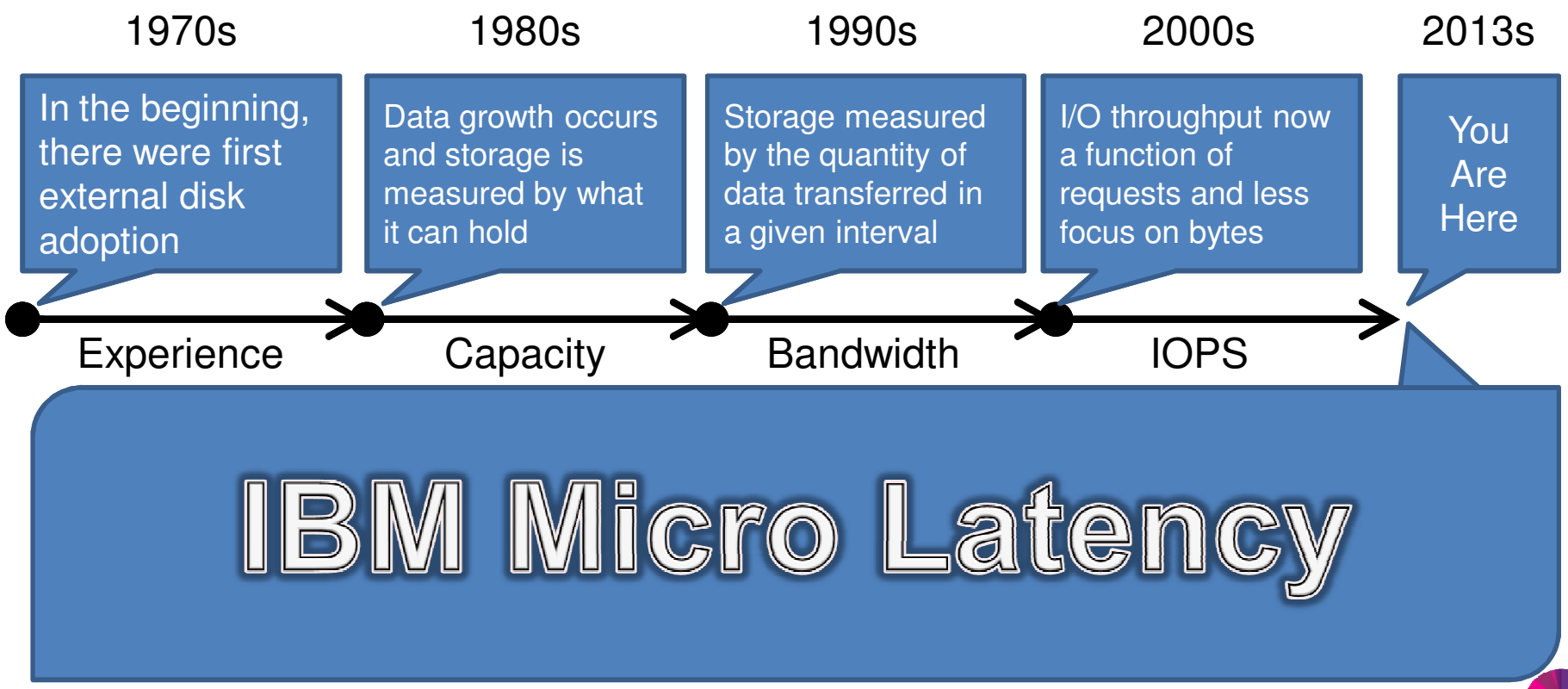
# The Flash Market is Heating Up… *Fast*

"IDC predicts that solid state storage revenue will grow from $1.7 billion in 2011 and reach $5.6 billion by 2016, resulting in a **26.8% CAGR** from 2011 to 2016.

IDC expects the amount of NAND solid state technology being shipped into the enterprise to grow an impressive **20x**."

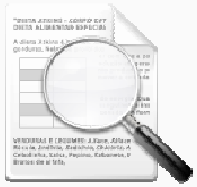
*Source: "Taking Enterprise Storage to Another Level: A Look at Flash Adoption in the Enterprise", August 2012, IDC*

# Evolution of Education in Storage Performance



# What is IBM Flash System?

## Texas Memory System
## PROVEN FLASH TECHNOLOGY LEADER



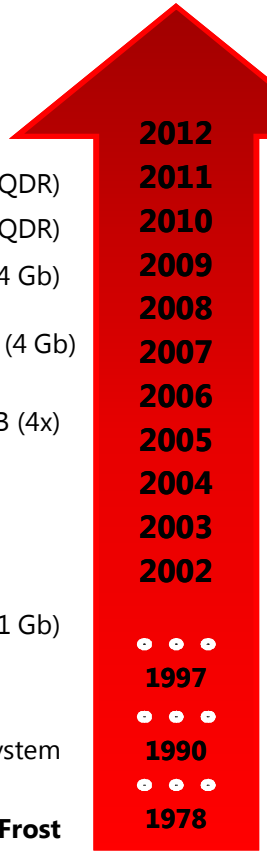
**Recognized**
by IBM Research


**34 Year History**
Deep Expertise

**30+ Patents**
Strong Flash

**800+ Customers**
In over 70 countries

# TMS (Texas System Memory ) History




**RamSan-820:** 12-24 TB eMLC Flash, 4 FC (8 Gb)/IB

**RamSan-810:** 10 TB eMLC Flash, 4 FC (8 Gb)/IB (QDR)

**RamSan-70:** 900 GB SLC Flash, PCIe x8 2.0

**RamSan-630:** 10 TB SLC Flash, 10 FC (8 Gb)/IB (QDR)

**RamSan-20:** 450 GB SLC Flash, PCIe x4

**RamSan-500:** 2 TB SLC Flash, 64 GB RAM, 8 FC (4 Gb)

**RamSan-320:** 64 GB RAM, 8 FC (2 Gb)

**RamSan-210/220:** 32 GB RAM, 4 FC (2 Gb)

**SAM-350/SAM-450:** DSP system

*Custom systems for seismic industry*

**CMPS:** custom SSD for Gulf Oil

**RamSan-720:** 5/10 TB SLC Flash, 4 FC (8 Gb)/IB

**RamSan-710:** 5 TB SLC Flash, 4 FC (8 Gb)/IB (QDR)

**RamSan-640:** 8 TB SLC Flash, 10 FC (8 Gb)/IB (QDR)

**RamSan-620:** 5 TB SLC Flash, 8 FC (4 Gb)

**RamSan-440:** 512 GB RAM, 8 FC (4 Gb)

**RamSan-400:** 128 GB RAM, 8 FC (4 Gb), 4 IB (4x)

**SAM 500:** DSP/SSD, 64 GB RAM, 15 FC (1 Gb)

**SAM-2000:** DSP system

**Company founded by Holly Frost**

# When to use IBM FLASH SYSTEM?

Consolidate application hardware and licensing

End user experience is critical to business performance

IO/GB exceeds disk economics

Cost savings against developer time

Increase scale of performance and minimize administration

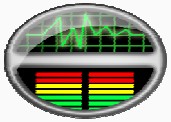
Consolidate power and rack estate

# IBM Flash System Impact

## Boost performance without re-architecting applications!

**85% Reduction**
In batch processing times

**90% Reduction**
In OLTP times



**150-200 µs Latency**

**80% Reduction**
Energy Usage

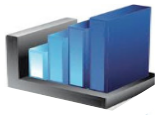
**75% Footprint Reduction**
Store one petabyte in a single floor tile. Add compression and add up to 100% more

**Enterprise Reliability**
High Availability, 2D Flash RAID™ and Variable Stripe RAID™

# IBM Flash Storage Sweet Spots: *Do More, Do it Faster!*

**OLTP Databases**

- Financial, gaming, real-time billing, trading, real-time monitoring, query acceleration (DB2/Oracle), etc.

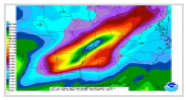


**Analytical applications (OLAP)**

- Business intelligence, batch processing, ERP systems, reporting, massive data feeds, etc.

**Virtual Infrastructures**

- VDI, Consolidated virtual infrastructures, user profiles, etc.



**HPC/Computational Applications**

- Simulation, modeling, rendering, FS metadata, scratch space, video on demand, thread efficiency, etc.

**Cloud-scale Infrastructures**

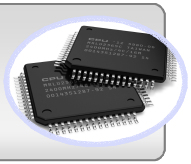
- On-demand computing, content distribution, web, caching, metadata, GPFS, active file management, etc.

Financial | Government | E-Commerce | HPC | Telecom

# The world of “Flash”

ηs Lowest Latency
Very Expensive

RAM



µs Consistent Performance
Low Latency

Server Flash
SAN Shared Flash Only Array
Server SSD
Hybrid Array or SAN SSD

IBM Flash

Tier or Cache

ms Fair Latency
Expensive HDD
Unbalanced Performance

Array HDD

High-Performance Disk (SAS)

>ms Lowest Performance
Lower Cost

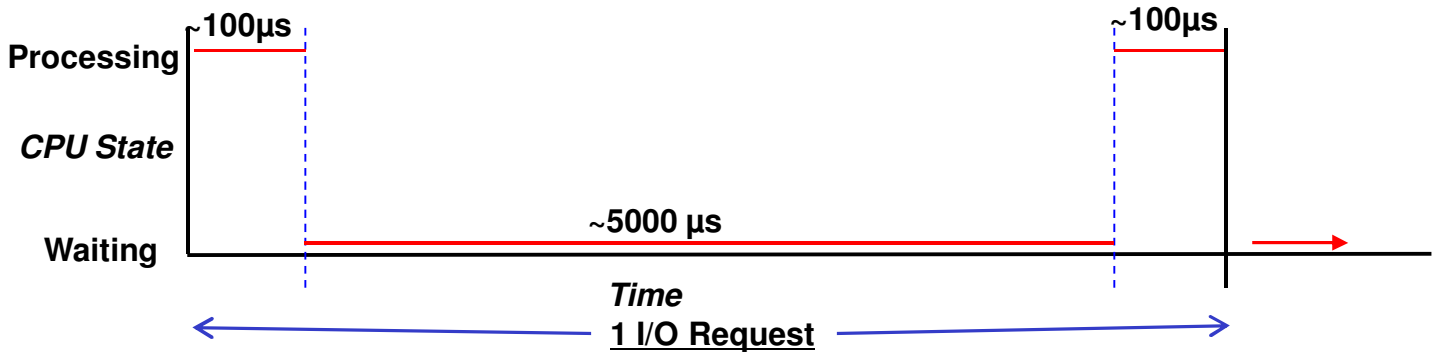
High-Capacity Disk (NLSAS)

# Flash Benefit

**I/O Serviced by Disk**

1. Issue I/O request (~ 100 µs)
2. Wait for I/O to be serviced (~ 5,000 µs)
3. Process I/O (~ 100 µs)

- Time to process 1 I/O request = 200 µs + 5,000 µs = 5,200 µs
- CPU Utilization = Wait time / Processing time = 200 / 5,200 = ~4%

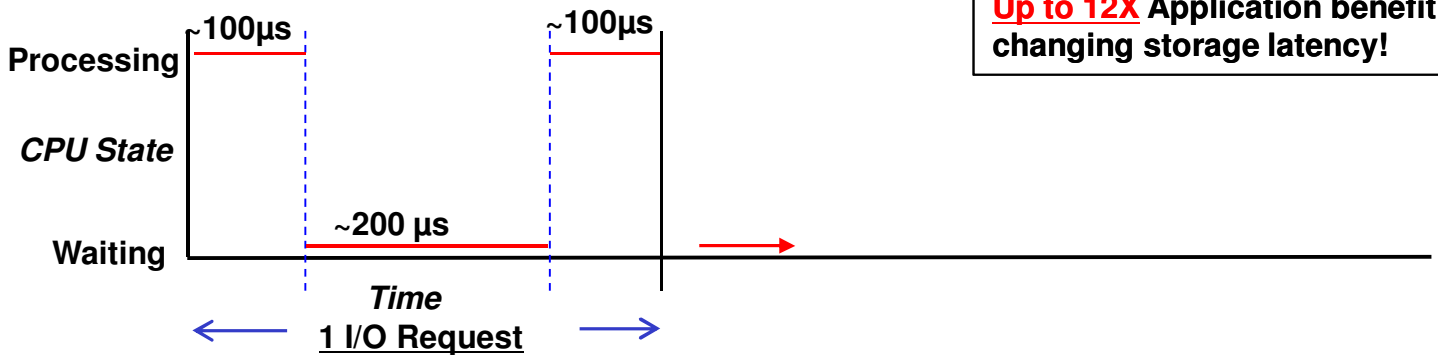


# Flash Benefit

**I/O Serviced by Flash System**

1. Issue I/O request (~ 100 µs)
2. Wait for I/O to be serviced (~ 200 µs)
3. Process I/O (~ 100 µs)

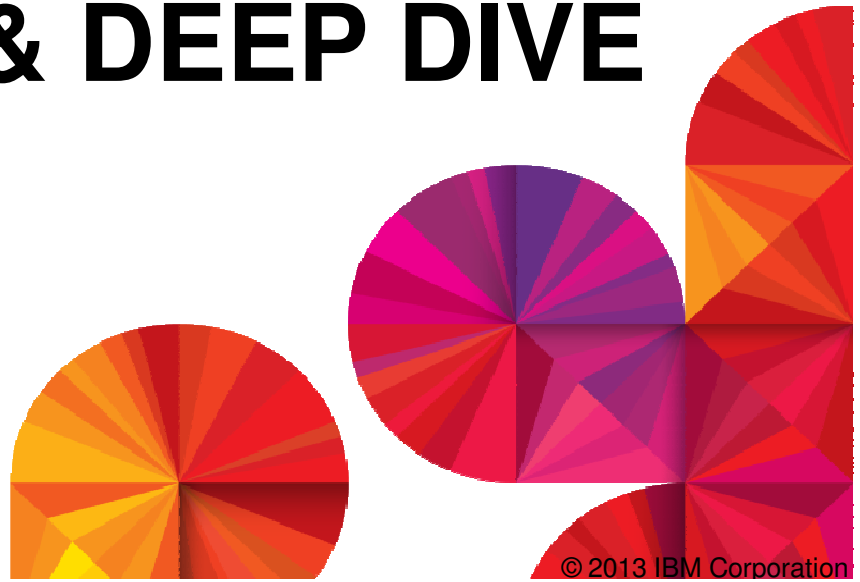
- Time to process 1 I/O request = 200 µs + 200 µs = 400 µs
- CPU Utilization = Wait time / Processing time = 200 / 400 = ~50%




**Up to 12X Application benefit by only changing storage latency!**

FlashSystem

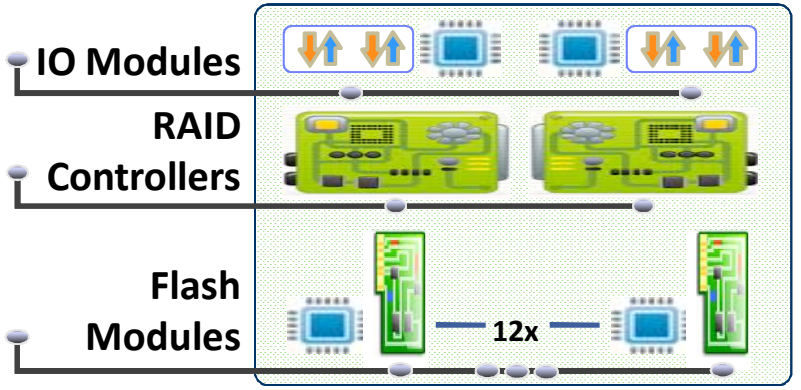
# ARCHITECTURE & DEEP DIVE



# Main Architectural Concepts

- Focus on returning low response times through high load

- Hardware only data path
  - Leverages FPGA's extensively
    - Intelligent Flash Modules
    - Field upgradable hardware logic
    - Quicker to market
    - Less expensive design cycle
  - Extremely high degree of parallelism




- Distributed computing model
  - 16 low-power PPC processors
  - Interface & Flash processors run thin RTOS
    - Not in active in data transfer
    - Responsible for garbage collection, monitoring

***"You cannot increase performance by adding lines of code."***

# IBM FlashSystem 710 / FlashSystem 810

## Speed up critical applications and make decisions faster

Accelerate **read-heavy** enterprise **storage area network (SAN)** applications…

- **Data warehouses** and online analytical processing (**OLAP**) databases
  - Sequential data collection
  - Large centralized databases
- Content delivery networks
- Rendering and video editing
- Modeling and simulation

### Extreme Performance

- **SLC** (710) / **eMLC** (810)
- 1,2,3,4,5 TB or 2,4,6,8,10 TB
- **570K (710) / 550K (810) IOPS**
- 5 GB/s (710)/ 4 GB/s (810) Bandwidth

### MicroLatency™

- **Low latency** 100/60 µs (710) and 110 / 60 µs (810) Read/Write
- **Purpose-built**, **highly parallel** design
- Maximize host **CPU efficiency** and **productivity**

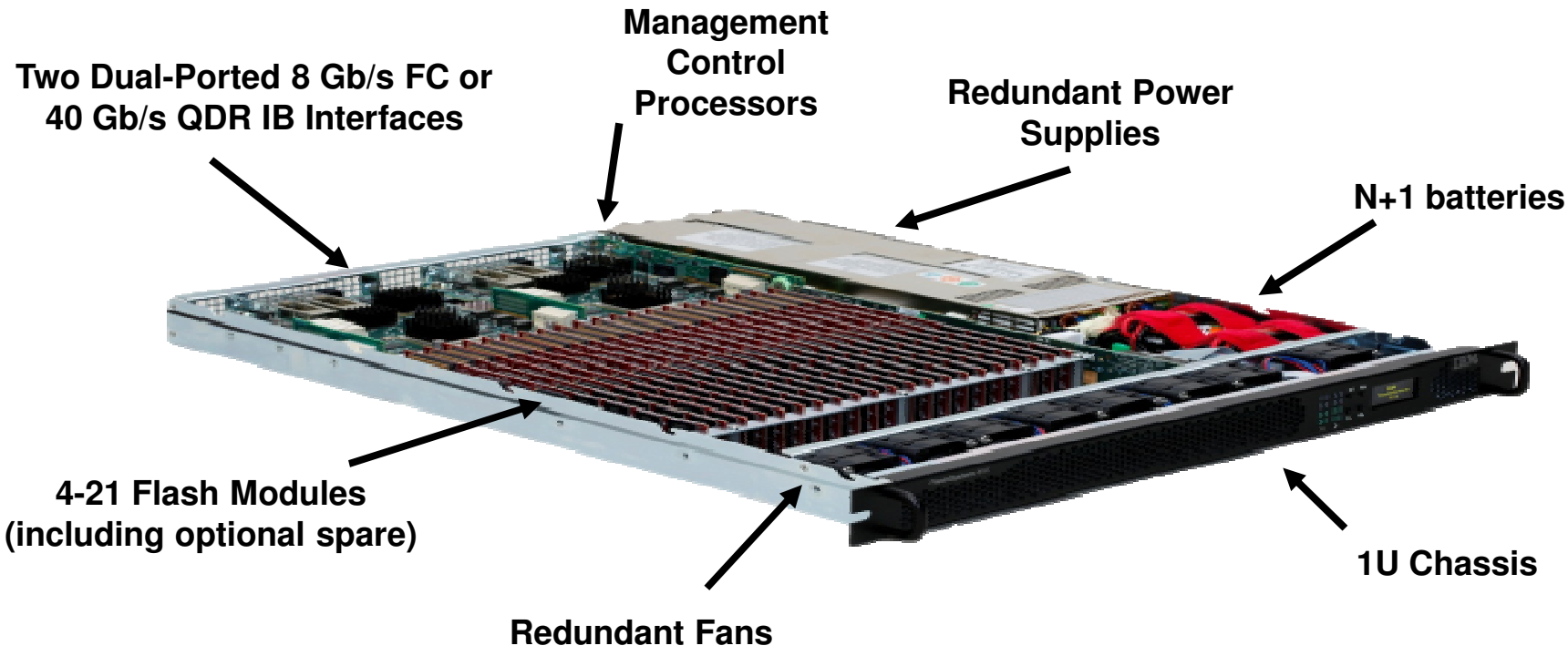
### Macro Efficiency

- **1U** form factor- minimal footprint for best of breed ROI
- Two dual-port 8 Gb **Fibre Channel** controllers or dual-port 40Gb **QDR InfiniBand** controllers
- **Low power** 450 watts (710) / 400 watts(810)
- Available hot-swapable flash modules in 720/820

### Enterprise Reliability

- **Variable Stripe RAID™** to protect against chip failure
- **Redundant power supplies** with active failover protection against single-source power issues
- **Error Correcting Code (ECC)** at chip level
- **Available integrated spare** flash card

# IBM FlashSystem 710 / FlashSystem 810 Architecture



# IBM FlashSystem 720 / FlashSystem 820

## High performance, low latency, high reliability solution to turbocharge your business

Designed for running multitenant heterogeneous (mixed workload) applications that require built-in ***high availability*** features…

- Transactional (**OLTP**) databases
- Analytical (**OLAP**) databases
- Virtualization & virtual desktop infrastructure (**VDI**)
- High performance computing (**HPC**)
- **Cloud** infrastructure, private and public

### Extreme Performance

- **SLC** (720) / **eMLC** (820)
- 6 – 12 / 12 – 24 TB
- w/ **High Availability**
- **525K (720/820) IOPS**
- 5 (720) / 4 (820) GB/s Bandwidth

### MicroLatency™

- **Low Latency** 100/25 µs (720) 110/25 µs (820) Read/Write
- **Purpose-built**, **highly parallel** design
- Maximize host **CPU efficiency** and **productivity**

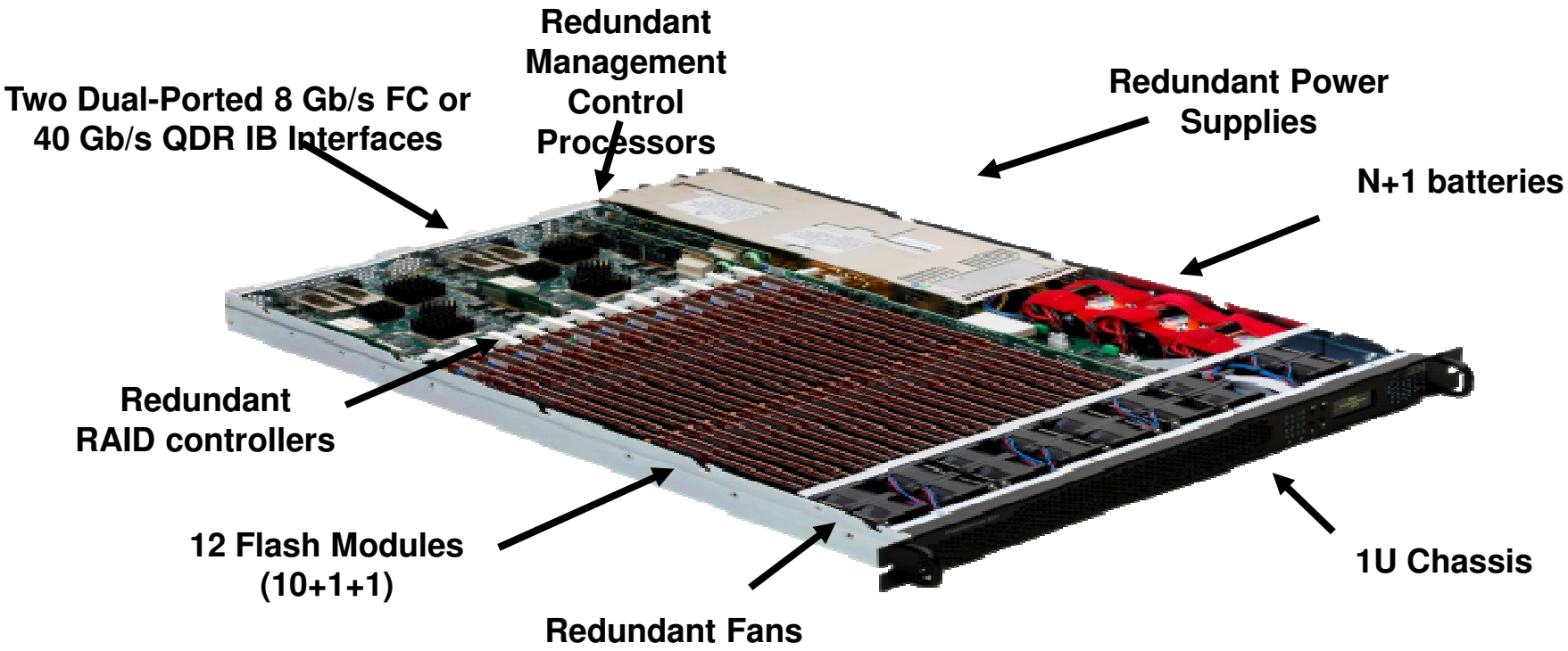
### Macro Efficiency

- **1U** form factor- minimal footprint for best of breed ROI
- Two dual-port 8 Gb **Fibre Channel** controllers or dual-port 40Gb **QDR InfiniBand** controllers
- Hot swappable flash modules
- **Low power** 500 watts (720) / 450 watts(820)

### Enterprise Reliability

- **Variable Stripe RAID™** to protect against chip failure
- **Redundancy** for power, data, and management
- **2D Flash RAID** eliminates single point of failures
- **Available integrated spare** flash card limiting down time
- **Error Correcting Code (ECC)** at chip level

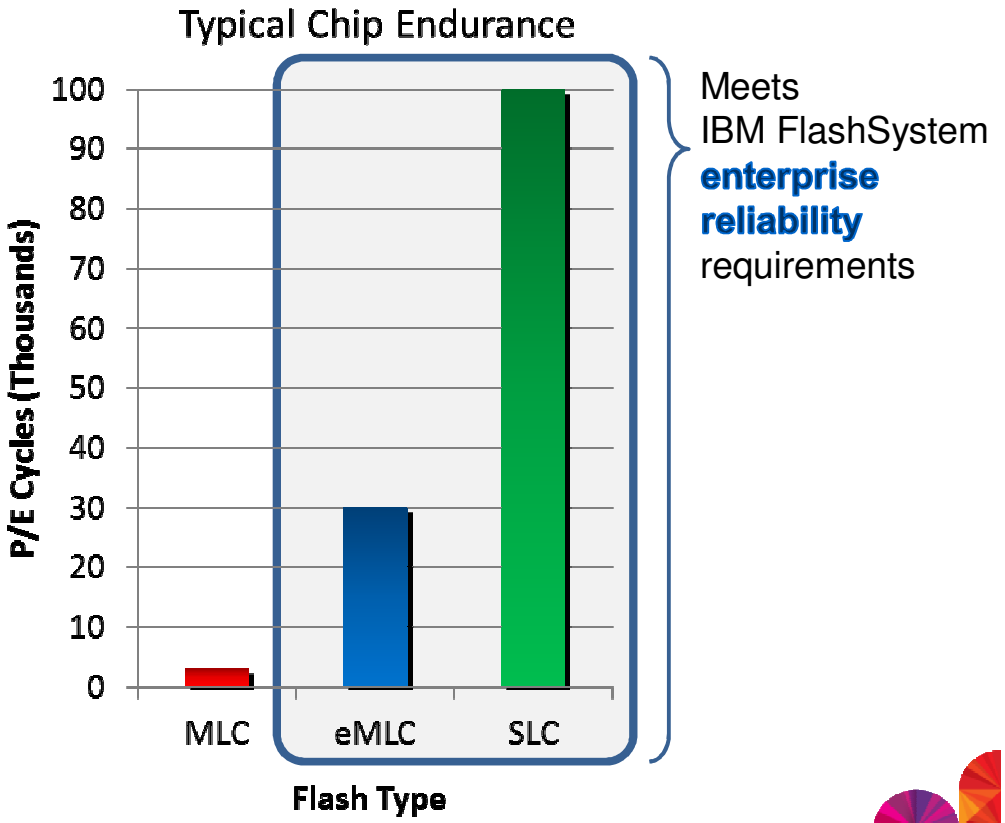
# IBM FlashSystem 720 / FlashSystem 820 Architecture



# Flash Types

$$System\ Life = \frac{Flash\ Capacity \times Flash\ Quality}{Flash\ Write\ Bandwidth}$$

- Flash type matters as P/E cycles vary
- **MLC/cMLC:** Multilevel cell
  - *traditionally consumer-grade*
- **eMLC:** Enterprise-grade version MLC
  - ***10x improvement over MLC***
- **SLC:** Single-level cell
  - ***33x improvement over MLC***
  - ***½ capacity per chip***
- **eMLC** will handle most enterprise applications workload requirements
- When addressing $ / Life, eMLC offers 10x endurance



# Interfaces

- Common TMS developed USIC code base
  - PPC for Instruction Decode (direction) and FPGA for Data Path (Flow)
  - uCOS Real-Time OS
  - Fast-path operations handled completely on card
  - Active/Active across ports
  - Active/Active across cards

**8GB FC Interface**

Same HW design for last 10+ years

Evolutionary upgrades on components (FPGAs, transceivers, uCode)

Dual-ported

Autosensing 8Gb/4Gb/2Gb

Supports Arbitrated Loop, Point-to-Point (F-port included)

**40GB QDR IB Interface**

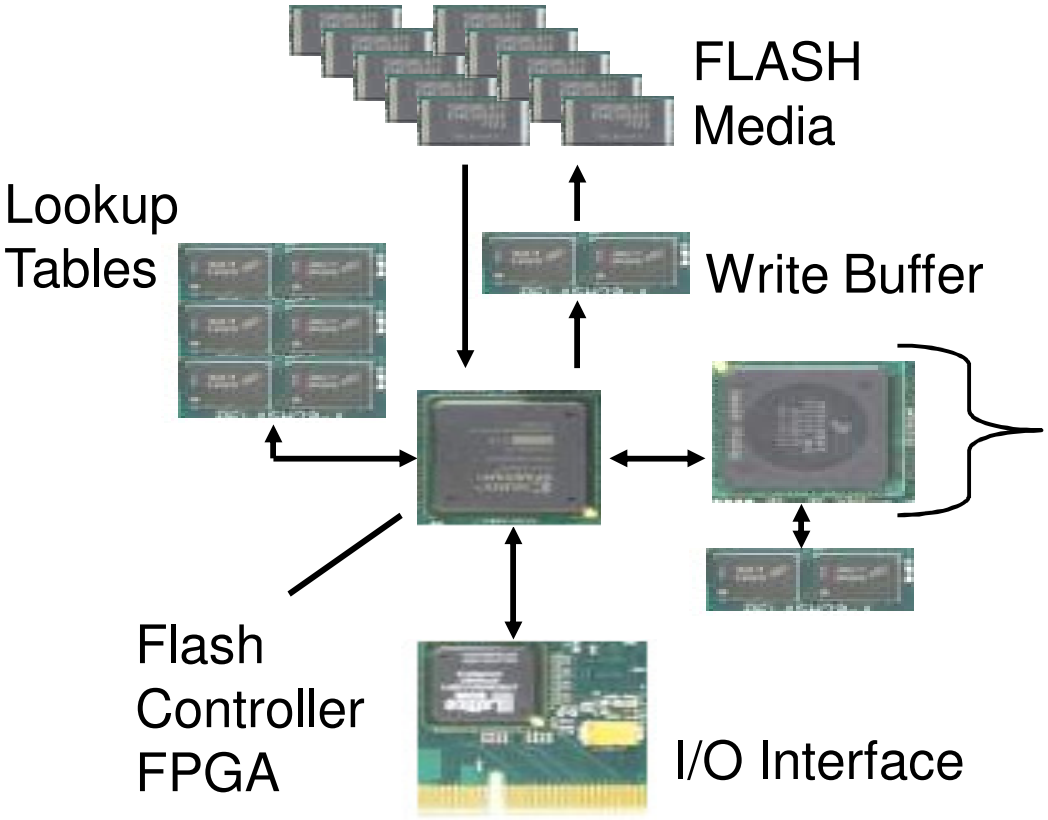
Same HW design for last 3+ years

Melanox Protocol Driver

Dual-ported

Supports SRP over RDMA

# Series-7™ Flash Controller Design

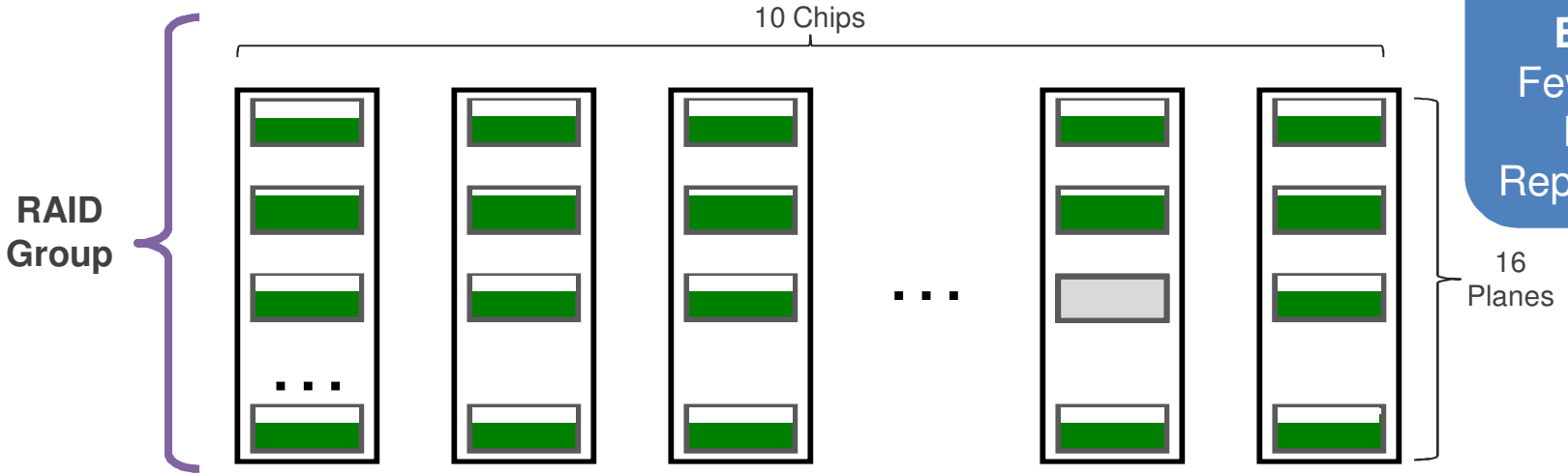



CPU – Write setup, Garbage collection, Error handling

***Out of the data path activities***

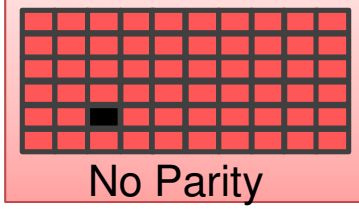
# Variable Stripe RAID™ (VSR)

- Patented Variable Stripe RAID allows RAID stripe sizes to vary.
- If one die fails in a ten-chip stripe, only the failed die is bypassed, and then data is restriped across the remaining nine chips. **No system rebuild needed!**
- VSR reduces maintenance intervals caused by Flash failures

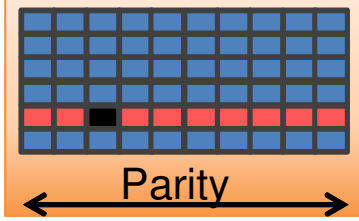



**Unique Business Benefit:** Fewer Flash Module Replacements

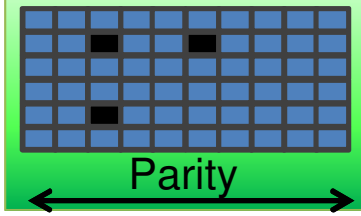
# Variable Stripe RAID™ (VSR)




Form Factor SSD
- Flash failure = Disk failure
- Requires top-level RAID
- Relatively frequent hot-swaps




Enterprise Flash Drive or Memory Module
- Flash failure = Degraded state within module
- Performance impact on RAID set
- Hot-swap to resolve

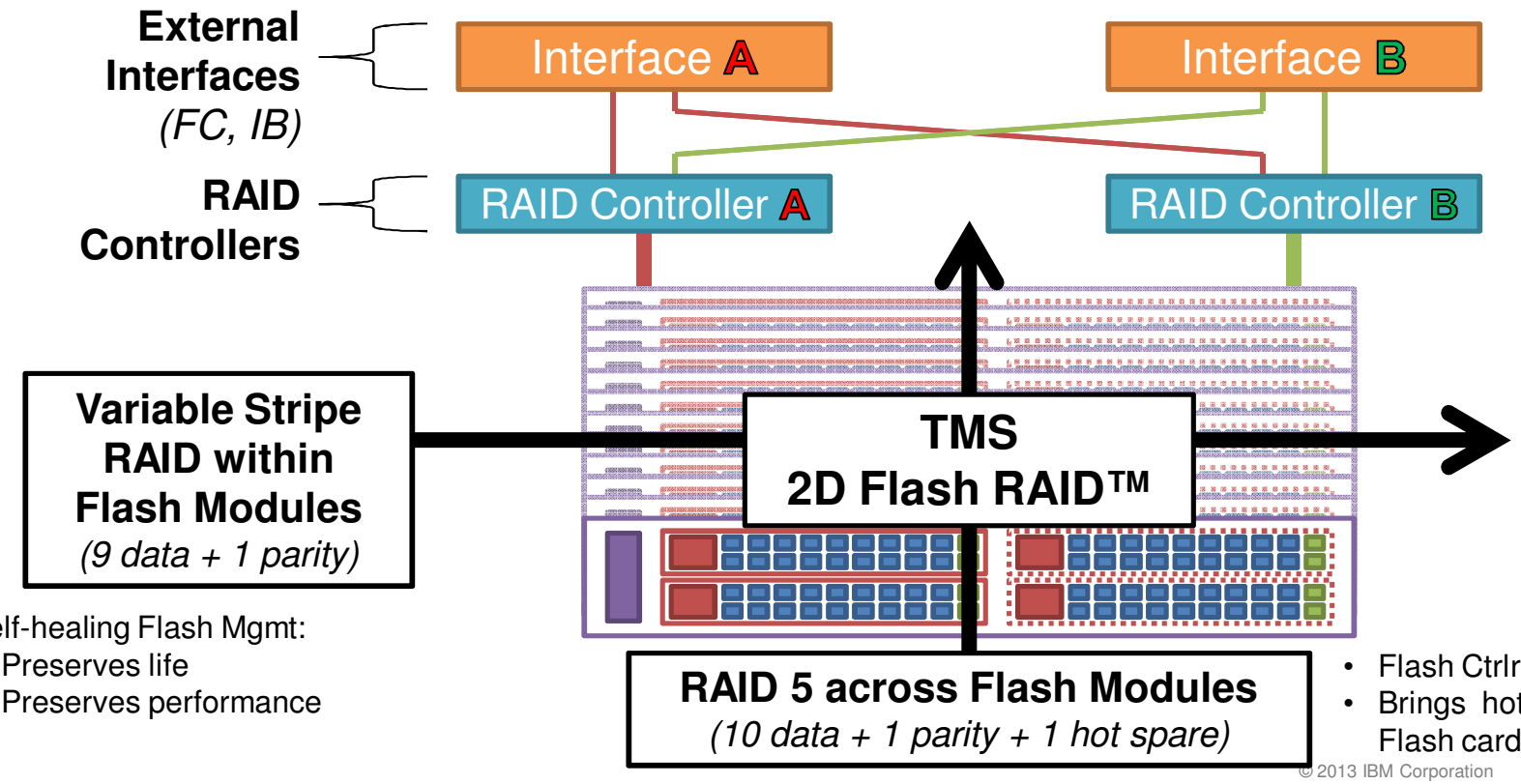



FlashSystem with Variable Stripe RAID
- Preserves Flash life
- Preserves performance
- Re-parity data in microseconds

Value of
Variable Stripe RAID™

Less maintenance touches while still preserving the **life**, **protection**, and **performance** of the Day-1 experience

# 2D Flash RAID™ (FlashSystem 720/820)



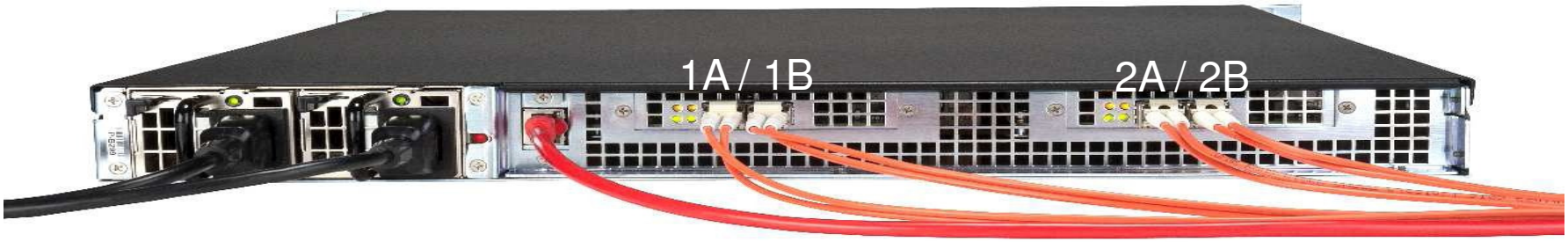

Self-healing Flash Mgmt:
- Preserves life
- Preserves performance

- Flash Ctrlr Protection
- Brings hot-swap to Flash cards

# Supported FC Attachment Modes

- **Point-to-Point:** IBM Flash supports point-to-point (N-port to N-port) topology for Fibre Channel. The most popular way to attach a RamSan is in a switched fabric topology (F-port). The switched fabric topology implies that there is a FC storage network switch between the host and the IBM Flash.
- **Arbitrated Loop:** IBM Flash supports older arbitrated loop (NL-port) topology for Fibre Channel if needed. Using this topology, IBM Flash can be attached directly to up to four host servers in half-duplex.

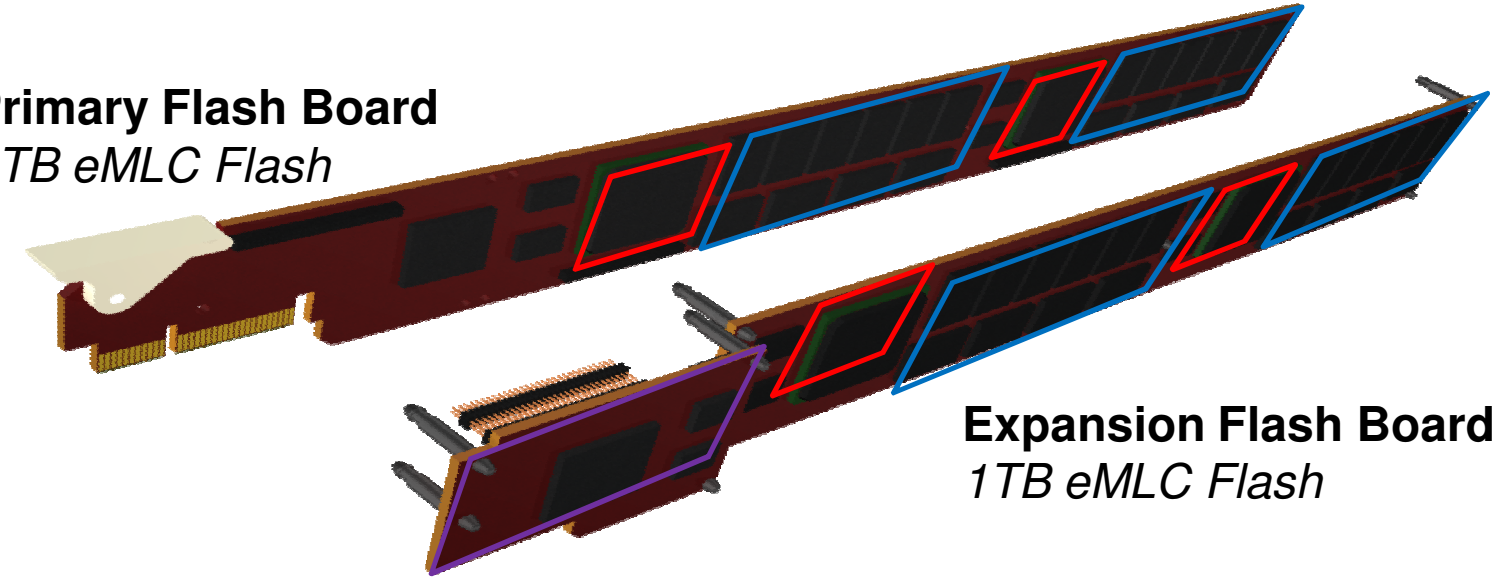
# Switch Zoning and Cabling




- When connecting to redundant fabrics, connect the A-ports to one fabric and the B-ports to the other fabric.
- Zoning is best deployed in a 1:1 port ratio with 1 server's HBA port for 1 IBM Flash FC port. A server HBA port serving I/O to multiple FC ports within the same IBM Flash results in excessive paths with no real benefit, since the server's single HBA port is the limiting factor for performance.
- Zone dual-HBA hosts to ports 1A and 2B, or 2A and 1B respectively, to ensure your IBM Flash config. has no SPOFs.

# Supported FC Attachment Modes - GUI

# Flash Module



**Primary Flash Board**
*1TB eMLC Flash*

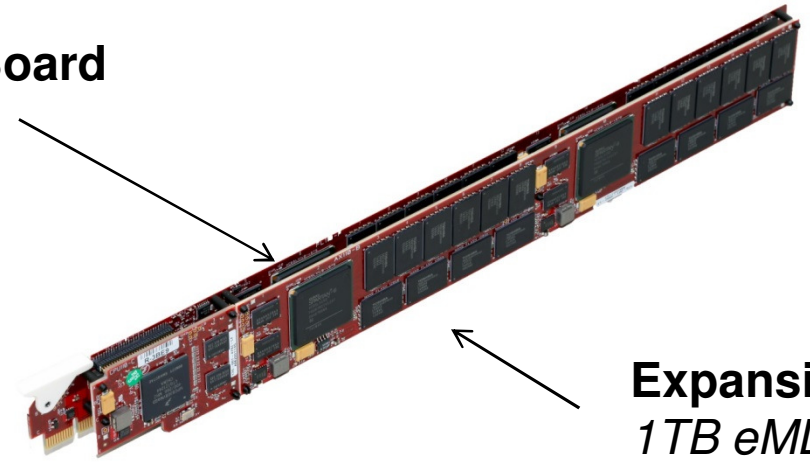
**Expansion Flash Board**
*1TB eMLC Flash*

**Series-7 Flash Controller™**
*2 per Board*
*4 per Module*

**eMLC Flash Chips**
*20 per Flash Controller*
*40 per Board, 80 per Module*

**Gateway Interface**
*Dual ports to backplane*

# Flash Module Live

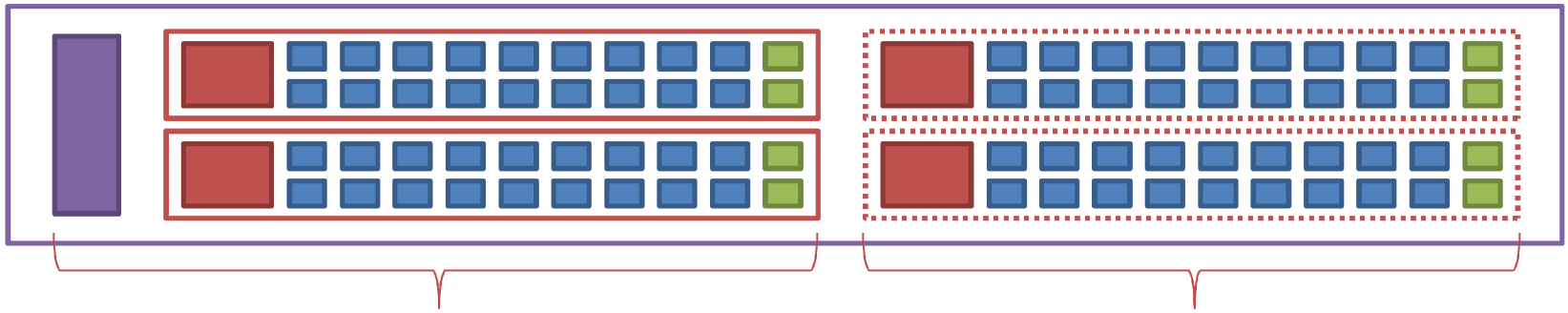


**Primary Flash Board**
*1TB eMLC Flash*

**Expansion Flash Board**
*1TB eMLC Flash*

**Gateway Interface & PPC CPU**

# Flash Module Architecture



**Primary Board**

**Expansion Board**
*(optional)*

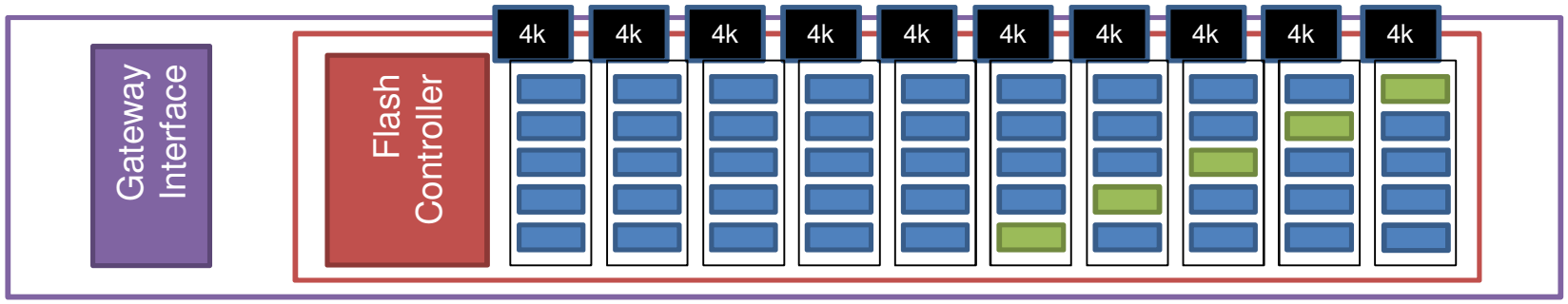
**Series-7 Flash Controller™**
*2 per Board*
*2 or 4 per Module*

**Gateway Interface & Control PPC**
*Dual ports to backplane*

**Flash Chips**
*20 per Flash Controller*
*40 per Board*
*40 or 80 per Module*

data  XOR parity

# Flash Module Data Layout




- Each flash controller operates independently of other controllers
- Data buffered by DRAM into Flashcard
- Data striped across Flash lanes at 4KB
  - Flash at 9+1 RAID initially, subject to change through VSR
- Parity is rotated across lanes in different Plane stripes
- All writes done sequentially within Flash at 8KB or higher coalescing
- Flash controller maintains write ordering and layout

# Four Layers of Data Correction

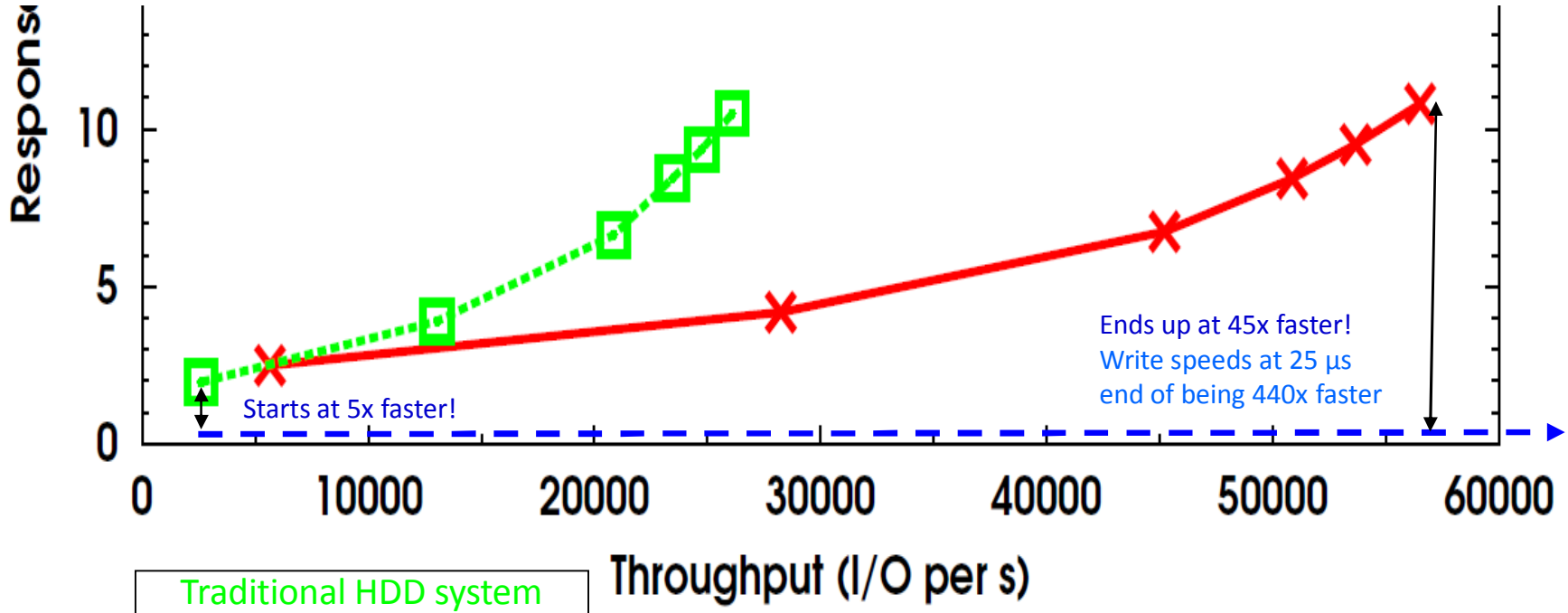
| Layer | | Protection |
|---|---|---|
| System-level RAID 5* managed by centralized RAID controllers | | Module failure |
| Module-level RAID 5 managed by each module across its chips | | Page failure |
| **Module-level Variable Stripe RAID™** chips | managed by each module across its | Sub-chip, chip or multi-chip failure |
| Chip-level ECC ( Error code correction ) managed by each module using its chips | | Bit and block errors |

**2D Flash RAID™**

* IBM System Flash 720/820 Only

IBM System Flash-720/820 introduce **System-Level RAID 5** across Flash modules, plus the other mechanisms found on all RamSan Flash storage systems.

# IBM FLASH SYSTEM Performance





# STORAGE VIRTUALIZATION
# IBM SAN Volume Controller 6.3
## Industry-leading storage virtualization




# Virtualize Storage to Increase Utilization by up to 30%

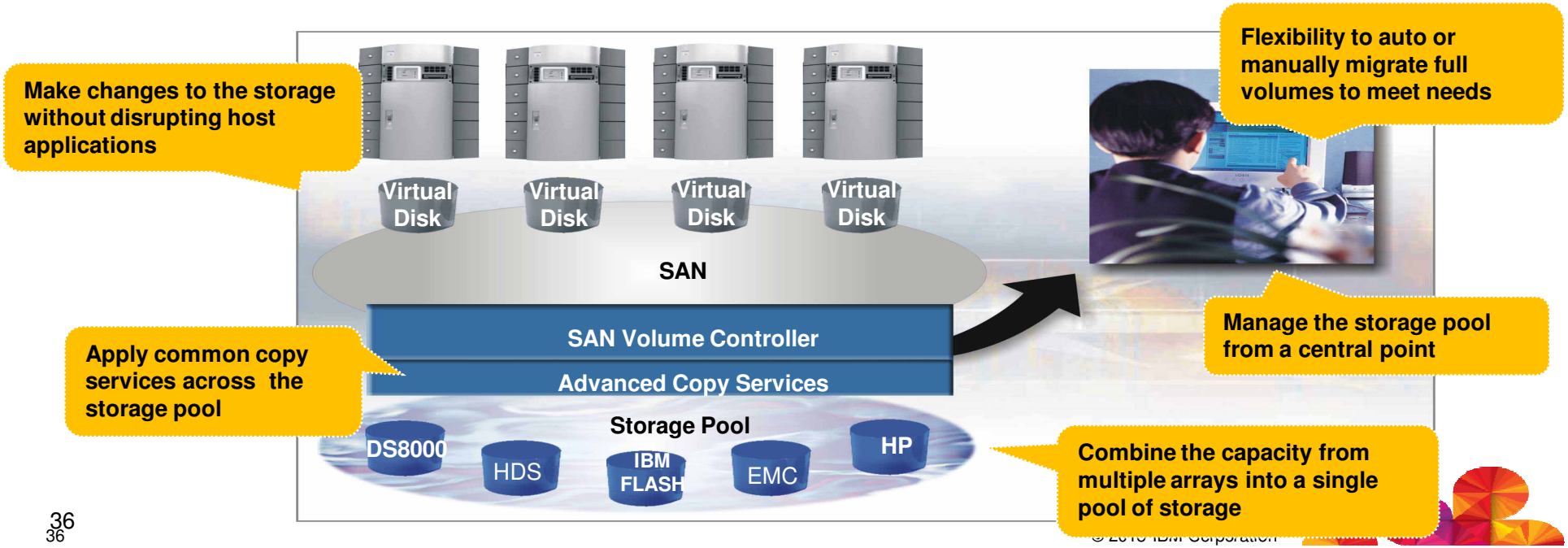
IBM

Store more with what's on the floor

Virtualize existing storage with IBM SAN Volume Controller

- **Increase usable capacity and flexibility** without complexity

*Over 25,000 IBM storage virtualization engines delivered!*

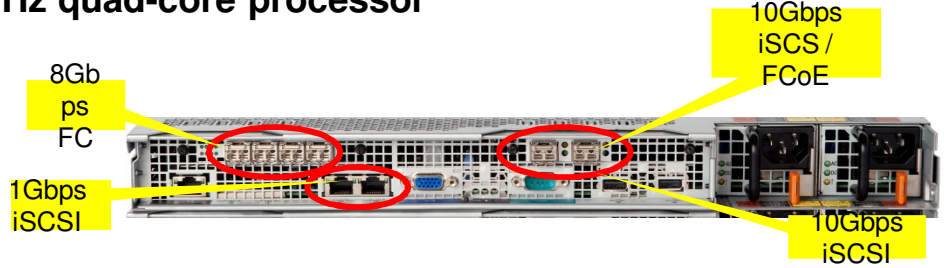




36

# IBM System Storage SAN Volume Controller

- **SVC 2145-CG8 engine based on IBM System x3550 M3 server (1U)**
  - **Intel® Xeon® 5600 (Westmere) 2.53 GHz quad-core processor**
  - **24GB of cache**
  - **Four 8Gbps FC ports**

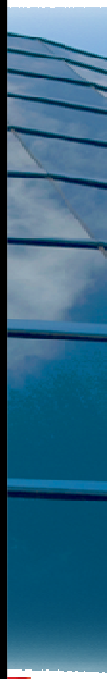



- Supports up to four internal 146GB SSDs
- Optional 10 Gb Ethernet support for 10 Gb iSCSI host attachment and future FCoE support
  - Two 10Gb ports per storage engine
  - Storage engine supports 10Gb iSCSI / FCoE or internal SSDs *but not both*
  - Factory install or nondisruptive field upgrade (field upgrades ship starting August 1)
- Engines may be intermixed in pairs with other engines in SVC clusters
- Mixing engine types in a cluster results in volume throughput characteristics of the engine type in that I/O group
- Cluster non-disruptive upgrade capability may be used to replace older engines with new CG8 engines
- Replaces the SVC 2145-CF8 engine

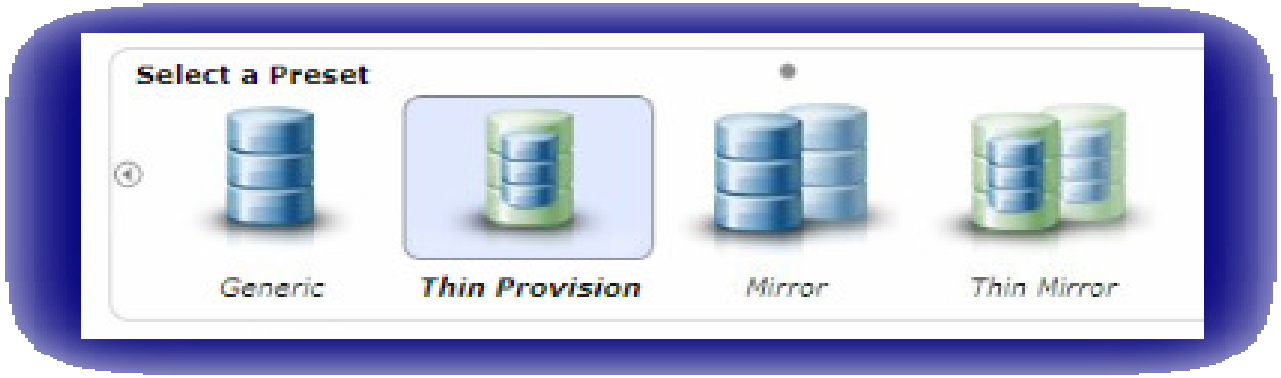
# Improve storage efficiency at every step

- Provision Storage
- Migrate Data
- Protect Data
- Tune Performance

*Advanced capabilities are built-in and ready to use*

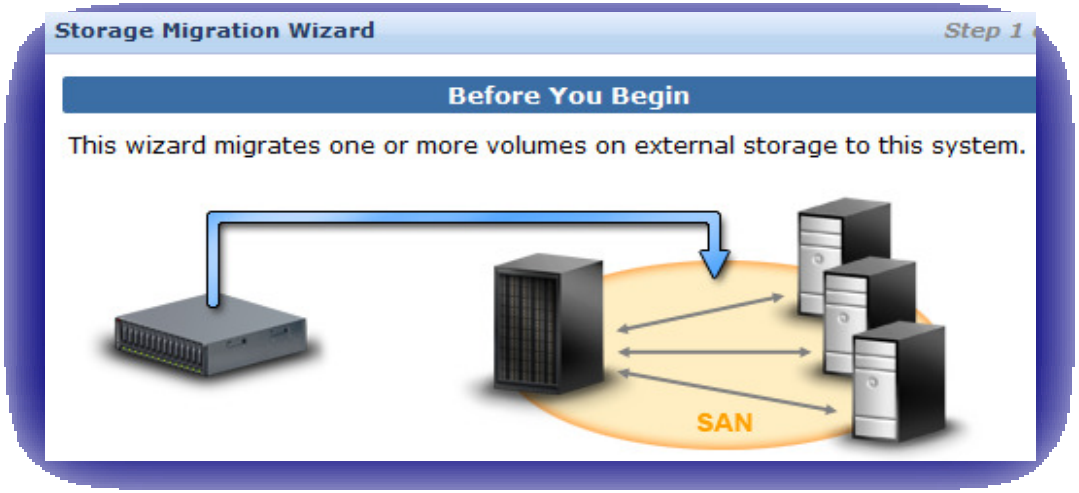


# Efficient storage provisioning



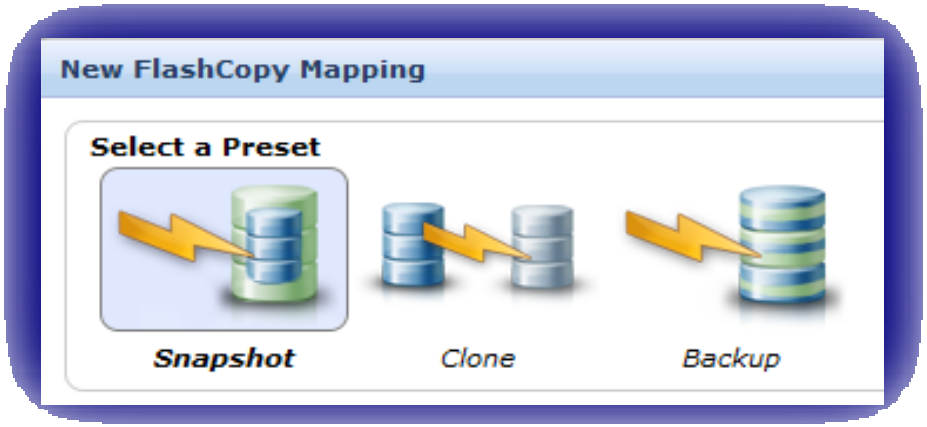

- **100% virtualized** storage enables flexible provisioning from a central storage pool
- High-performance **thin provisioning** frees up pre-allocated space
- **Multiple internal mirroring** options – RAID levels – increases flexibility

# Efficient data migration



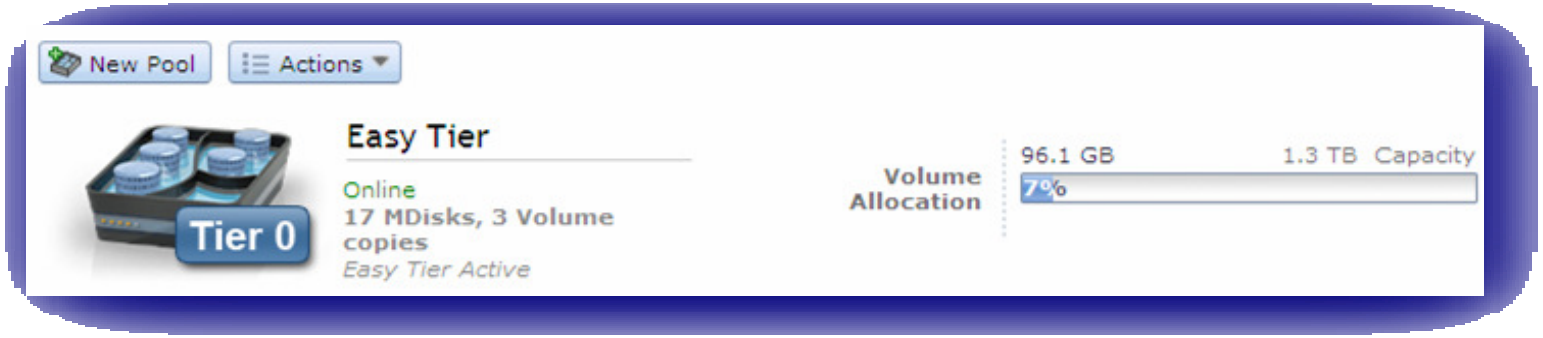

- **Migrate data online,** with no performance impact
- Migrate to thin provisioned volumes to **reclaim space**
- **One-time data migration** speeds deployment

# Efficient data protection




- **Space-efficient FlashCopy™** (snapshots) make data protection fast and easy
- Reverse FlashCopy enables **restores in minutes**
- Optional Metro and Global Mirror enables fast reliable **alternate site failover**

# Efficient tuning

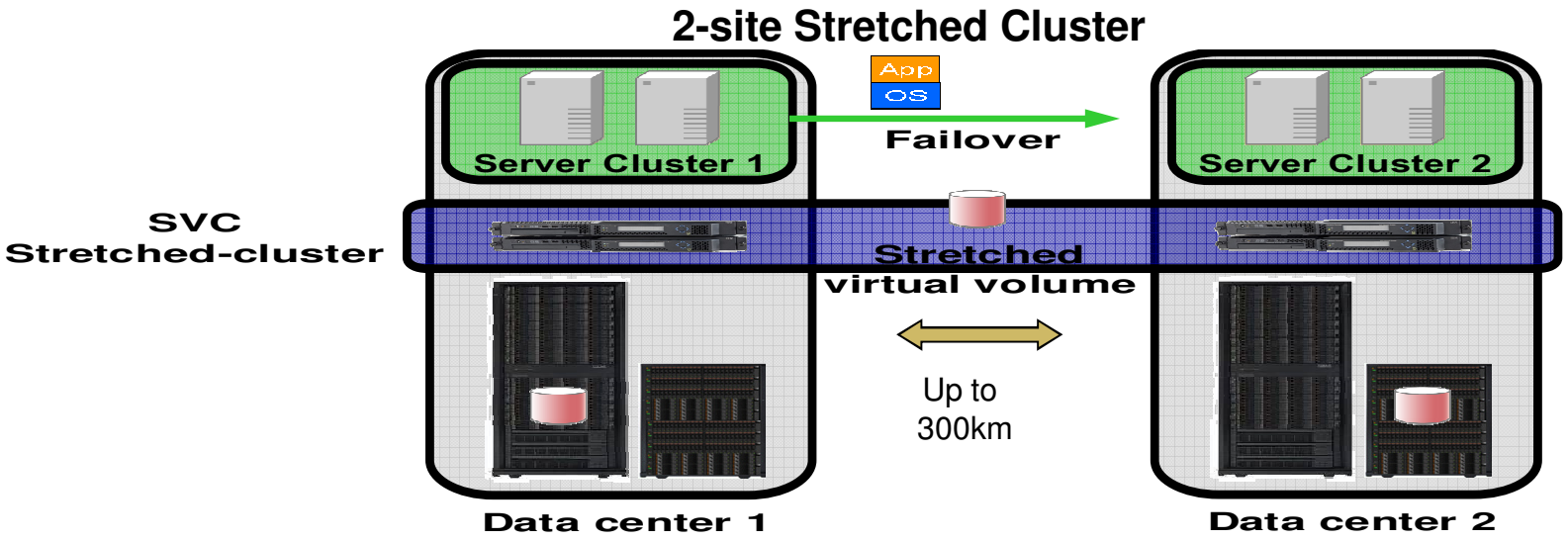



- IBM Easy Tier optimizes the use of solid-state drives
- **Boost performance** while **lowering per transaction costs** with efficient use of solid state storage
- Simple to set up – simply turn it on and it **runs automatically**
- **Eliminate performance hotspots** with help from Tivoli Storage Productivity Center

## Virtual Disk Mirroring

- **SVC stores two copies of a virtual disk, usually on separate disk systems**
  - SVC maintains both copies in sync and writes to both copies
- **If disk supporting one copy fails, SVC provides continuous data access by using other copy**
  - Copies are automatically resynchronized after repair
- **Intended to protect critical data against failure of a disk system or disk array**
  - A local high availability function, not a disaster recovery function
- **Copies can be split**
  - Either copy can continue as production copy
- **Either or both copies may be space-efficient**

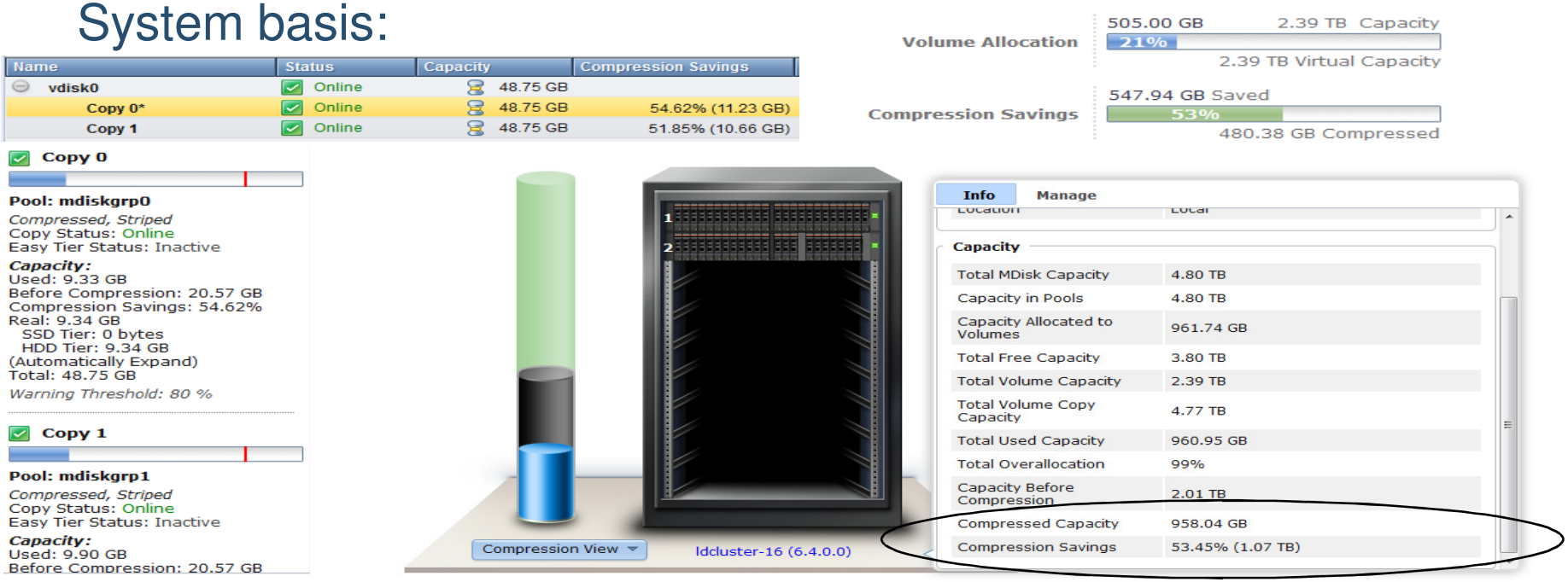
# Improved Data Protection with SVC Enhanced Stretched Cluster




- ✓Improve availability, load-balance, and deliver real-time remote data access by distributing applications and their data across multiple sites.
- ✓Seamless server / storage failover when used in conjunction with server or hypervisor clustering (such as VMware or PowerVM)
- ✓Up to 300km between sites (3x other vendor)

# Real-time Compression – GUI Support

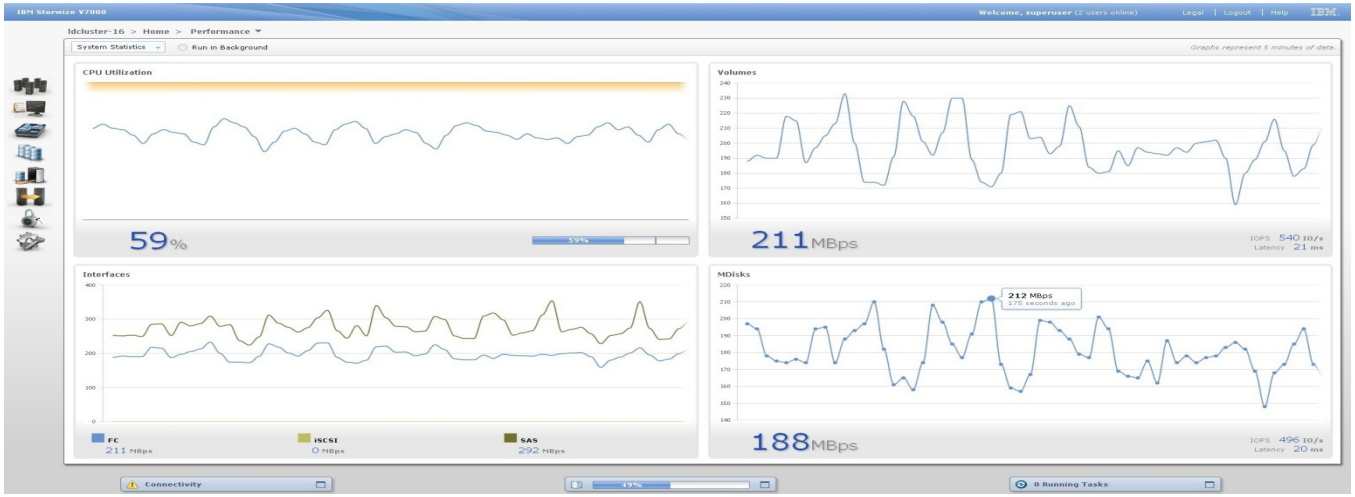
- GUI Displays Compression Savings on a Volume, Pool and System basis:



# Breakthrough Graphical User Interface Eliminates Complexity

- Auto-discovery and presets **slash system setup time**
- Easy navigation aids **speed administrator task selection and completion**
- Realistic graphics **simplify capacity and event management**

# Real Time Performance Statistics

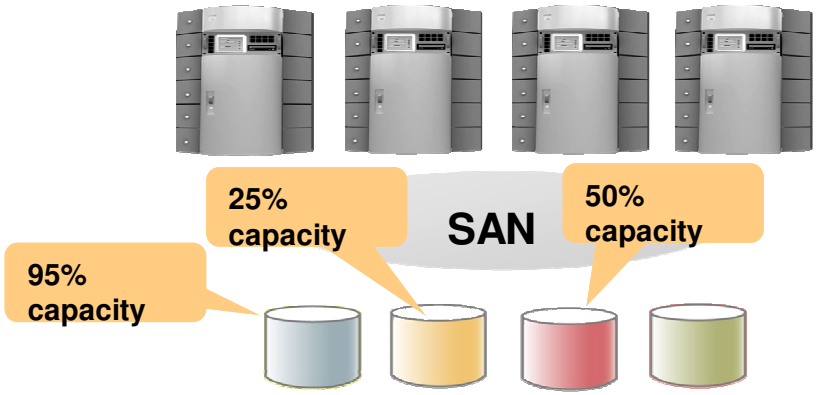



- Gathers system level performance statistics (CPU utilization; port utilization and I/O rates; volume and MDisk I/O rate, bandwidth, latency) in real time with sampling rates down to **5 sec**.
- Provides a snapshot view for **immediate monitoring** with 5-minutes of performance history
- Get “immediate” monitoring during environmental changes
- Troubleshoot sudden drops in performance
- Pair up with **TPC** for complete performance solutions

# Infrastructure Simplification with SAN Volume Controller

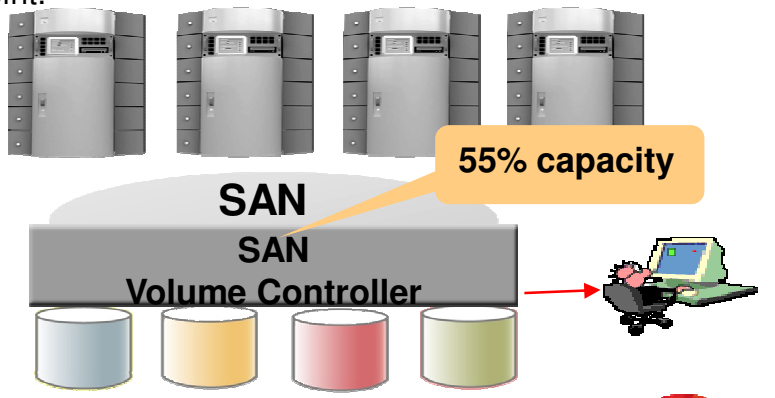
**Traditional SAN**

- Capacity is isolated in SAN islands
- Multiple management points
- Poor capacity utilization
- Capacity is purchased for, and owned by individual processors

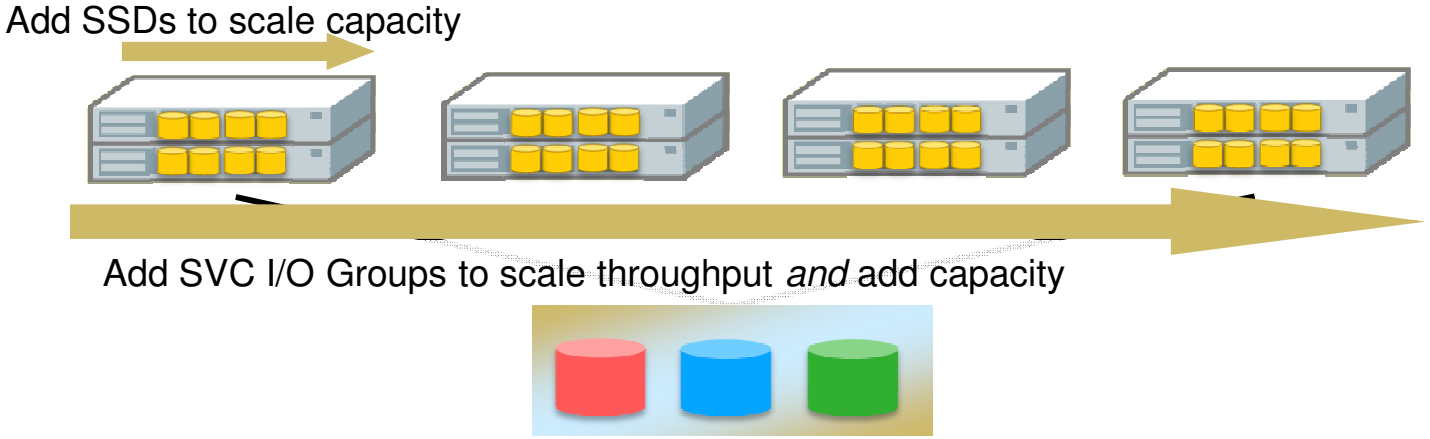



**SAN Volume Controller**

- Combines capacity into a single pool
- Uses storage assets more efficiently
- Single management point
- Capacity purchases can be deferred until the physical capacity of the SAN reaches a trigger point.



# SVC Scale-Out SSD Implementation



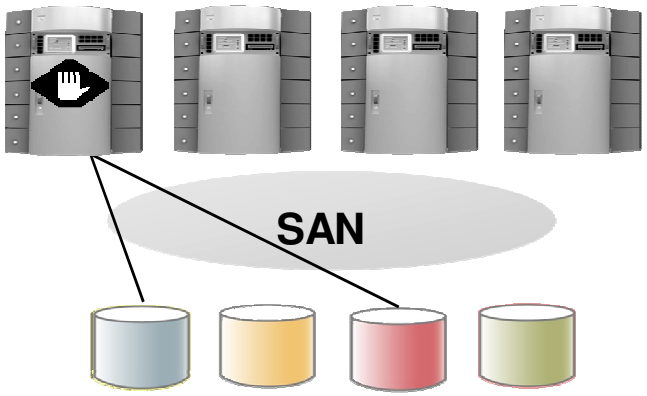

- **Add SSDs to SVC engines for more capacity**
  - SSDs may be added without disruption to engines
- **Add SVC engines for more capacity and throughput**
  - Additional engines provide more processing power, more bandwidth, more SAN attachments
  - SVC designed to deliver maximum I/O capability of SSDs
  - Up to 50,000 read IOPS per SSD
  - Up to 200,000 read IOPS per SVC I/O Group
  - Up to 800,000 read IOPS per SVC cluster


49

# Non-disruptive Data Migration with SAN Volume Controller

**Traditional SAN**

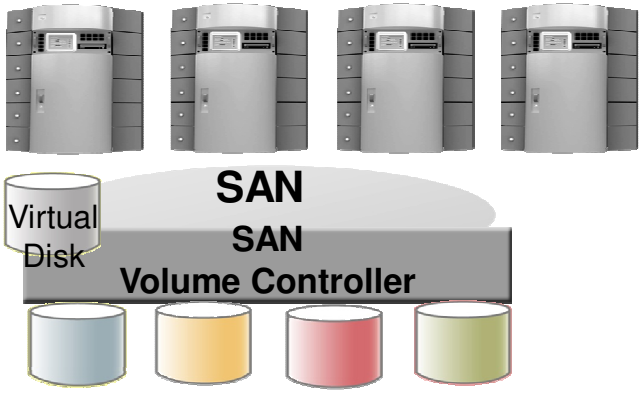
1. Stop applications
2. Move data
3. Re-establish host connections
4. Restart applications




**SAN Volume Controller**

1. Move data

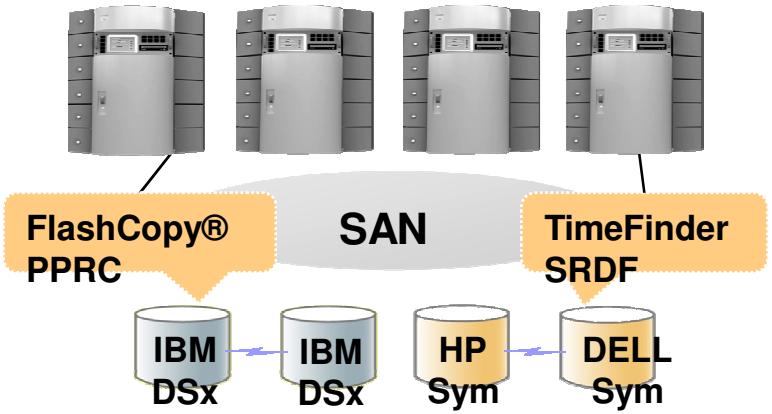
   Host systems and applications are not affected.



# Business Continuity with SAN Volume Controller

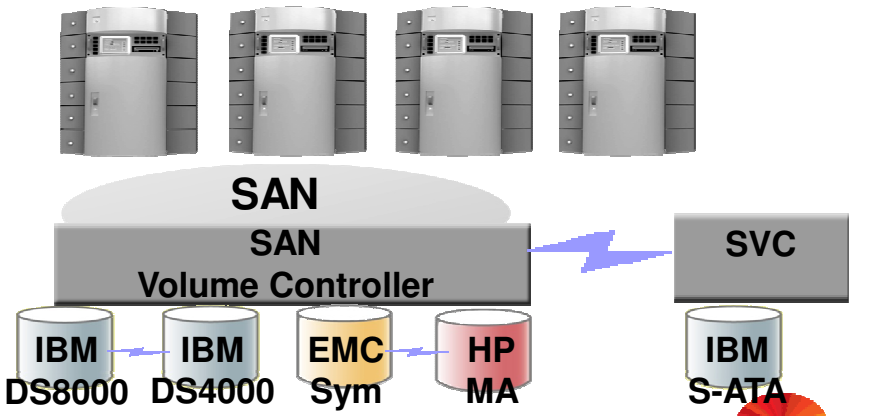
**Traditional SAN**

- Replication APIs differ by vendor
- Replication destination must be the same as the source
- Different multipath drivers for each array
- Lower-cost disks offer primitive, or no replication services




**SAN Volume Controller**

- Common replication API, SAN-wide, that does not change as storage hardware changes
- Common multipath driver for all arrays
- Replication targets can be on lower-cost disks, reducing the overall cost of exploiting replication services



# Introducing the IBM FlashSystem Solution

Extreme Performance with Enterprise Capabilities

Delivered with IBM On-site Integration

## IBM FlashSystem 820

Extreme Performance with IBM MicroLatency™
All Flash 20 TB RAIDed data capacity
Macro Efficiency 1U form factor
Variable Stripe RAID and 2-D RAID for Enterprise Reliability

## SAN Volume Controller

Business Continuity with Copy Services
Flash Copy for Backup & Optimal Workload Availability
$/TB Value with Thin Provisioning & Real Time Compression
Drive Storage Efficiency with Easy Tier

**High performance enterprise class featured solution**
**Scalable to 1.5M IOPs for large scale enterprise systems performance**

# Muchas Gracias !!!!