IBM Cloud & Smarter Infrastructure
Visibility. Control. Automation.

IBM TSM User Forum

IBM Tivoli Storage Manager

Trends und Kundenreferenzen

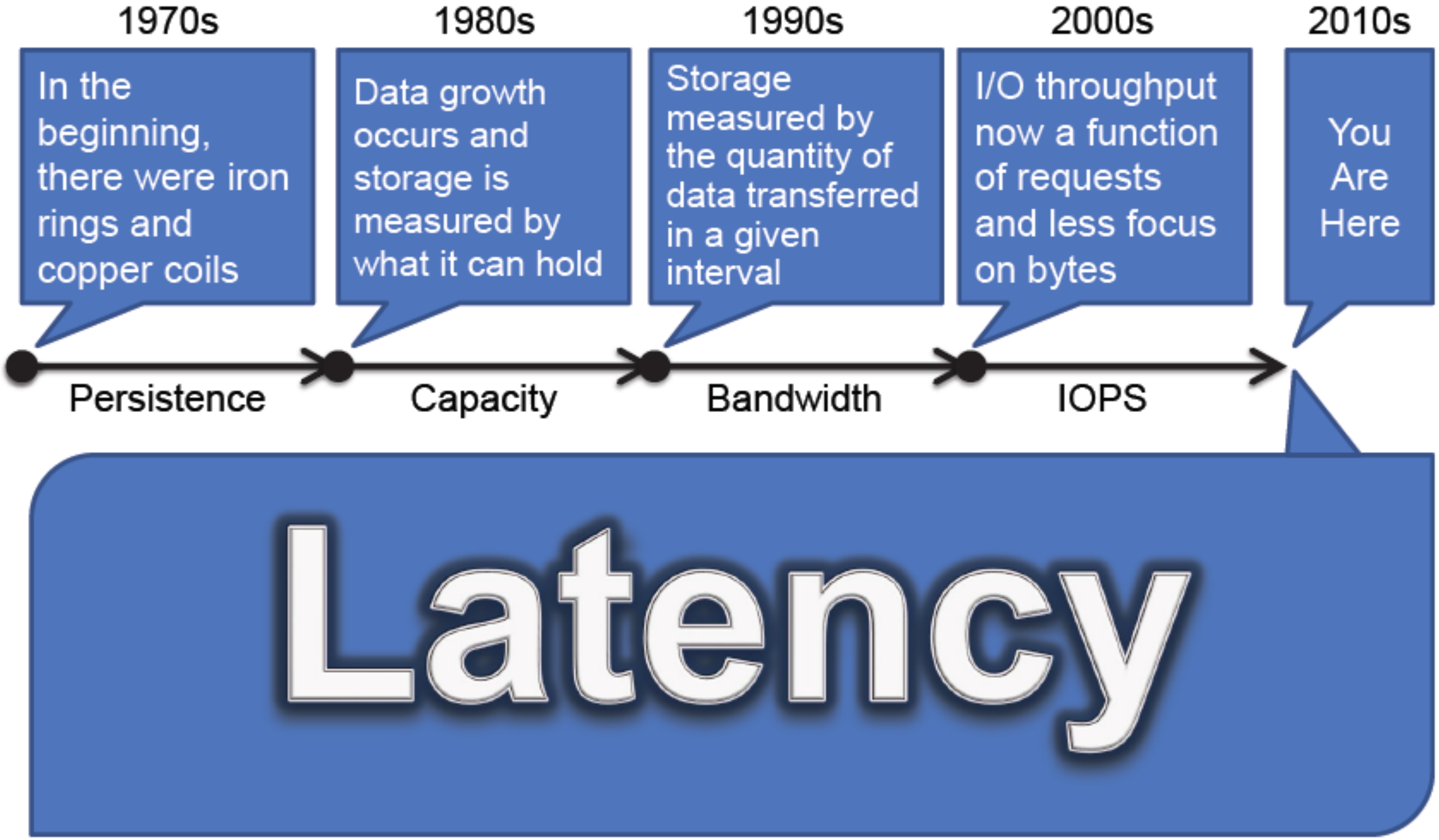
Manuel Schweiger – Senior IT Specialist

Nov. 2013

IBM FlashSystem

Das Ende des klassischen Disksystems?

Evolution of Education in Storage Performance



1970s

In the beginning, there were iron rings and copper coils

Persistence

1980s

Data growth occurs and storage is measured by what it can hold

Capacity

1990s

Storage measured by the quantity of data transferred in a given interval

Bandwidth

2000s

I/O throughput now a function of requests and less focus on bytes

IOPS

2010s

You Are Here

Latency

Reality of Storage Performance

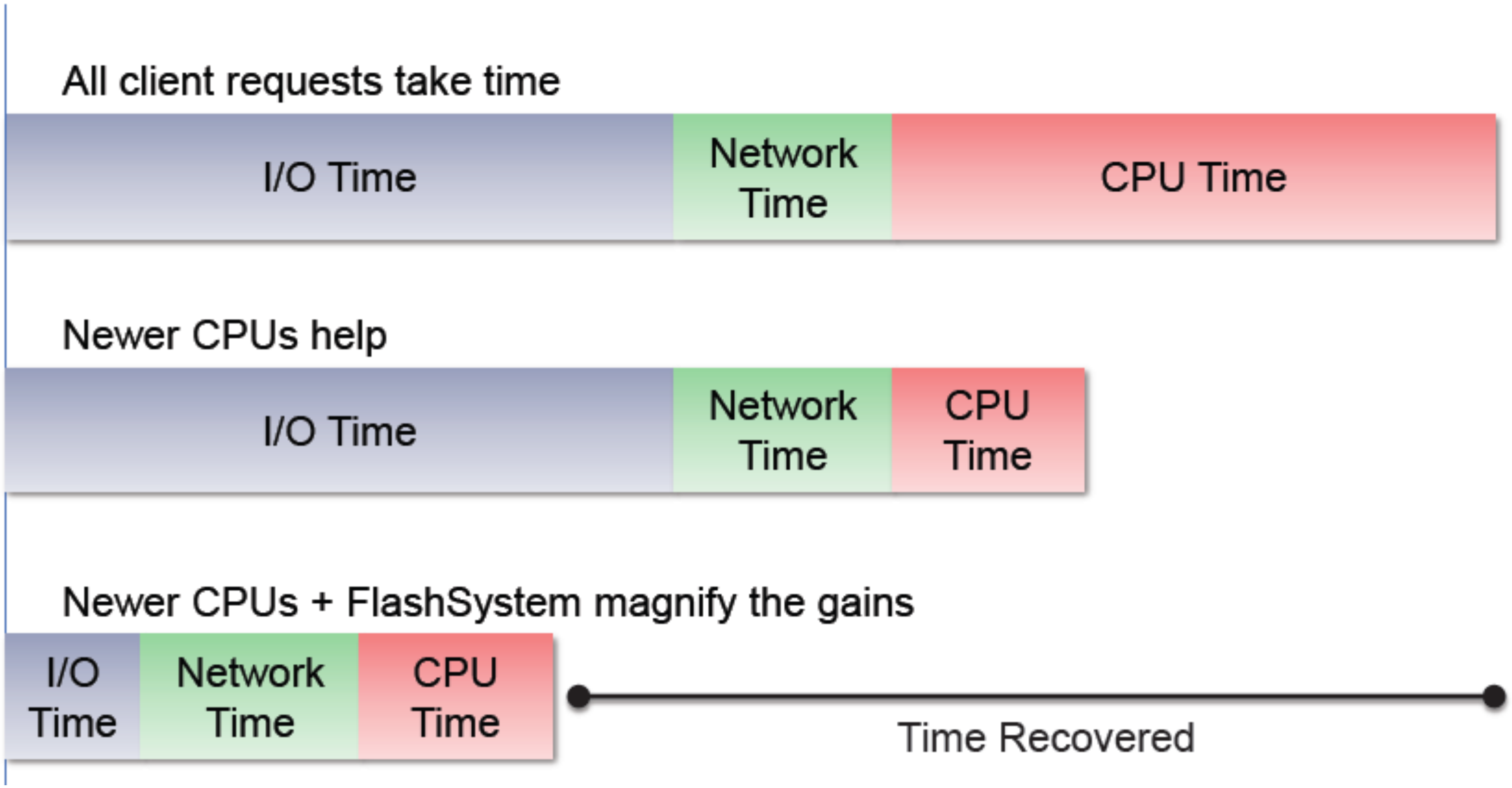
- The core responsibility of storage is to give/take data at the request of the processor (user)
- In consequence, the only impact storage has to the application performance is the amount of **time** the processors must **wait**
- Little's Law defines that there are only two ways to increase performance:
  - Change the app (Q)
  - Lower the response time (t)

Remember: IOPS & BW (aka Rate)are products of the application pressure and the response time of the storage.

Little's Law

$$\frac{Q}{t} = Rate$$

Applications see time waiting, not IOPS



All client requests take time

I/O Time

Network Time

CPU Time

Newer CPUs help

I/O Time

Network Time

CPU Time

Newer CPUs + FlashSystem magnify the gains

I/O Time

Network Time

CPU Time

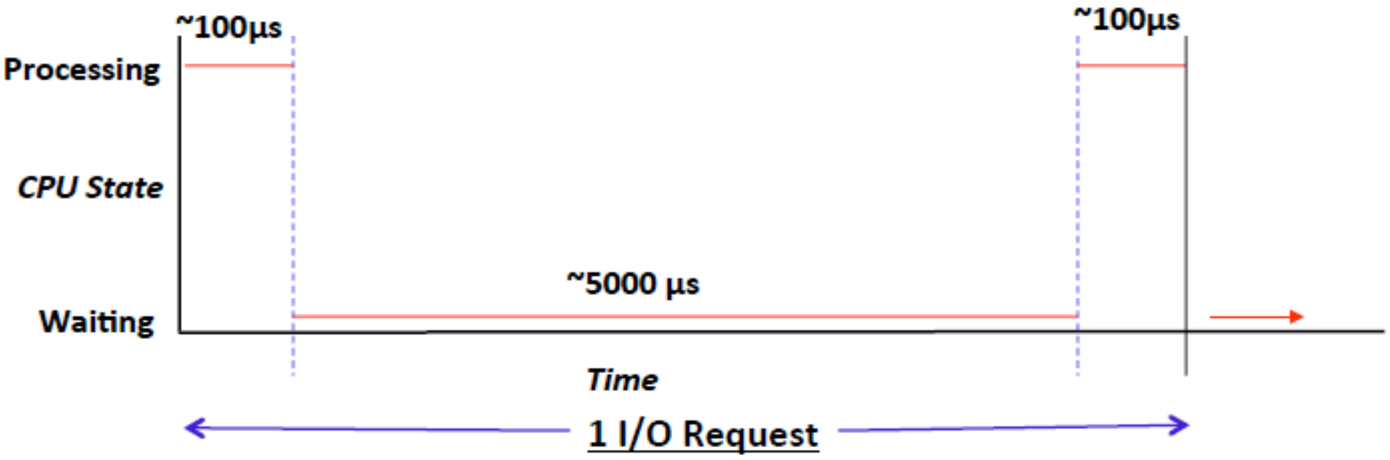
Time Recovered

FlashSystem 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%

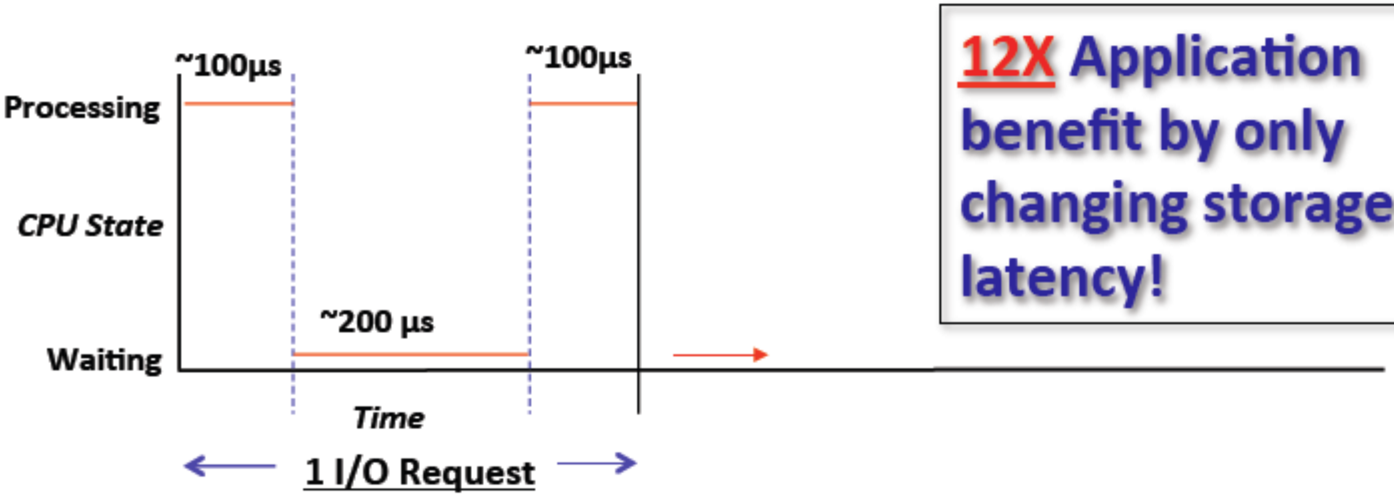


FlashSystem Benefit

I/O Serviced by FlashSystem

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

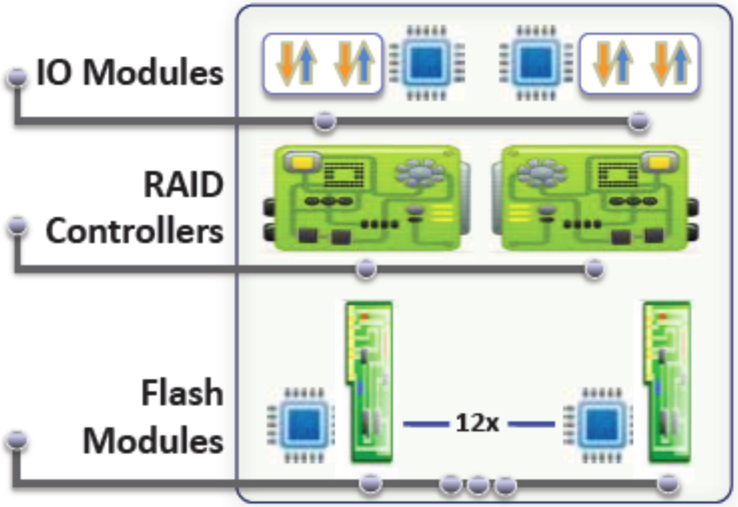
- Time to process 1 I/O request = 200 µs + 200 µs = 400 µs
- CPU Utilization = Wait time / Processing time = 200 / 400 = ~50%



12X Application benefit by only changing storage latency!

Core FlashSystem Concepts

- FlashSystem is hardware-only block storage devices that follows open-standard SCSI-3 protocol
- FlashSystem provides a hardware-only data path
  - Custom FPGA-based data movement decreases latency vs. software
- Lower latency on standard SAN interfaces vs. competitors
  - Either on DAS (PCIe cards) or SAN!
- Distributed out-of-data-path CPU processing



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

IBM FlashSystem 710 / FlashSystem 810

Speed up critical applications and make

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)
- 1-5 TB or 2-10 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 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)
- 5,10, 20 TB w/ **High Availability** (6,12, 24 TB non HA)
- **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**

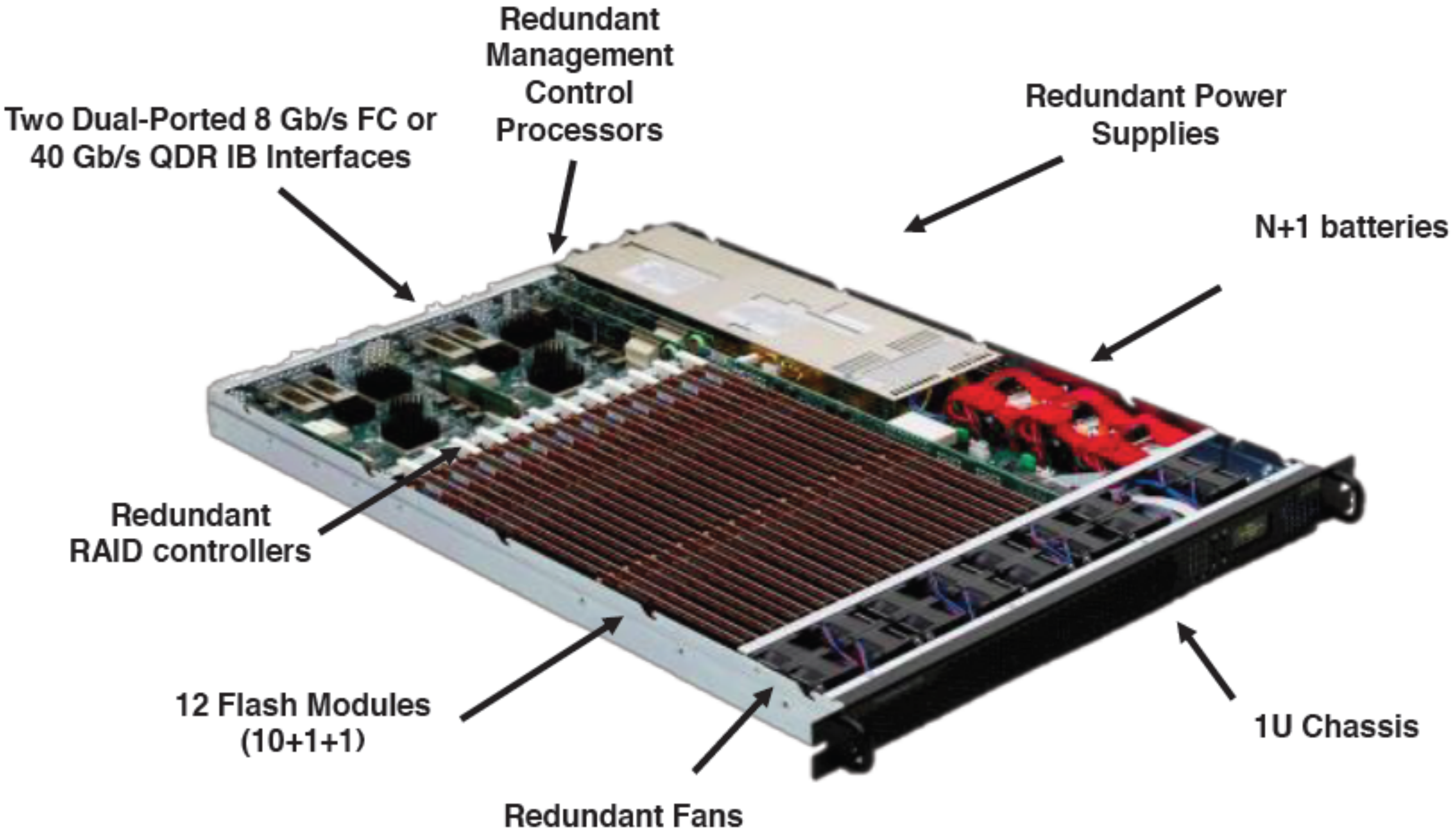
Macro Efficiency

- **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

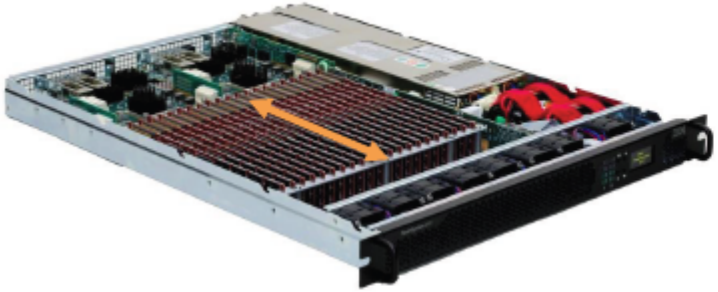
- **Variable Stripe RAID™** to protect against chip failure
- **Redundancy** for power, data, and management
- **2D Flash RAID** eliminates single point of failures
- **Available integrated spare** flash card limiting down time
- **Error Correcting Code (ECC)** at chip level

IBM FlashSystem 720 / FlashSystem 820 Architecture


Redundant Management Control Processors
Two Dual-Ported 8 Gb/s FC or 40 Gb/s QDR IB Interfaces
Redundant Power Supplies
N+1 batteries
Redundant RAID controllers
12 Flash Modules (10+1+1)
Redundant Fans
1U Chassis

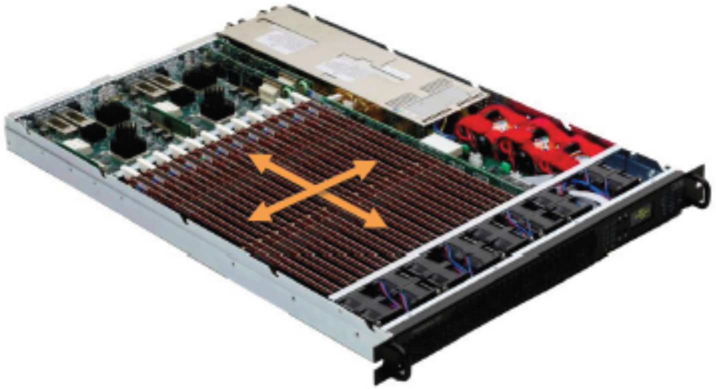
Key Differences Between FlashSystem x10 and x20

FlashSystem 710/810



1D RAID across Flash chips
Incremental Capacities
No Flash Hot-Swap
5TB/10TB Max Capacity

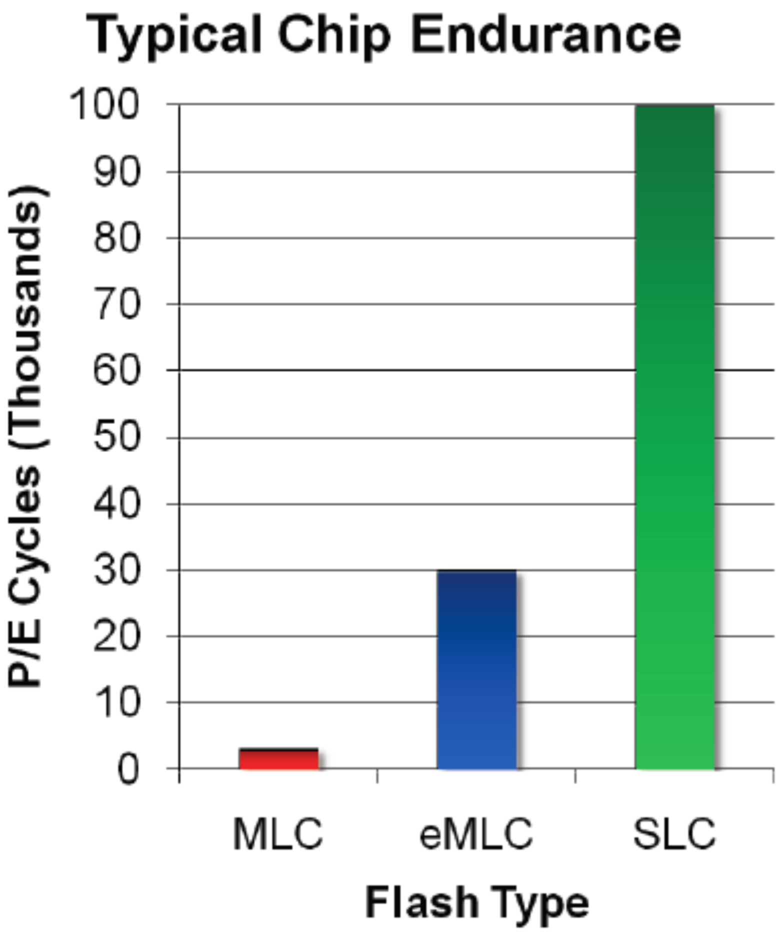
FlashSystem 720/820



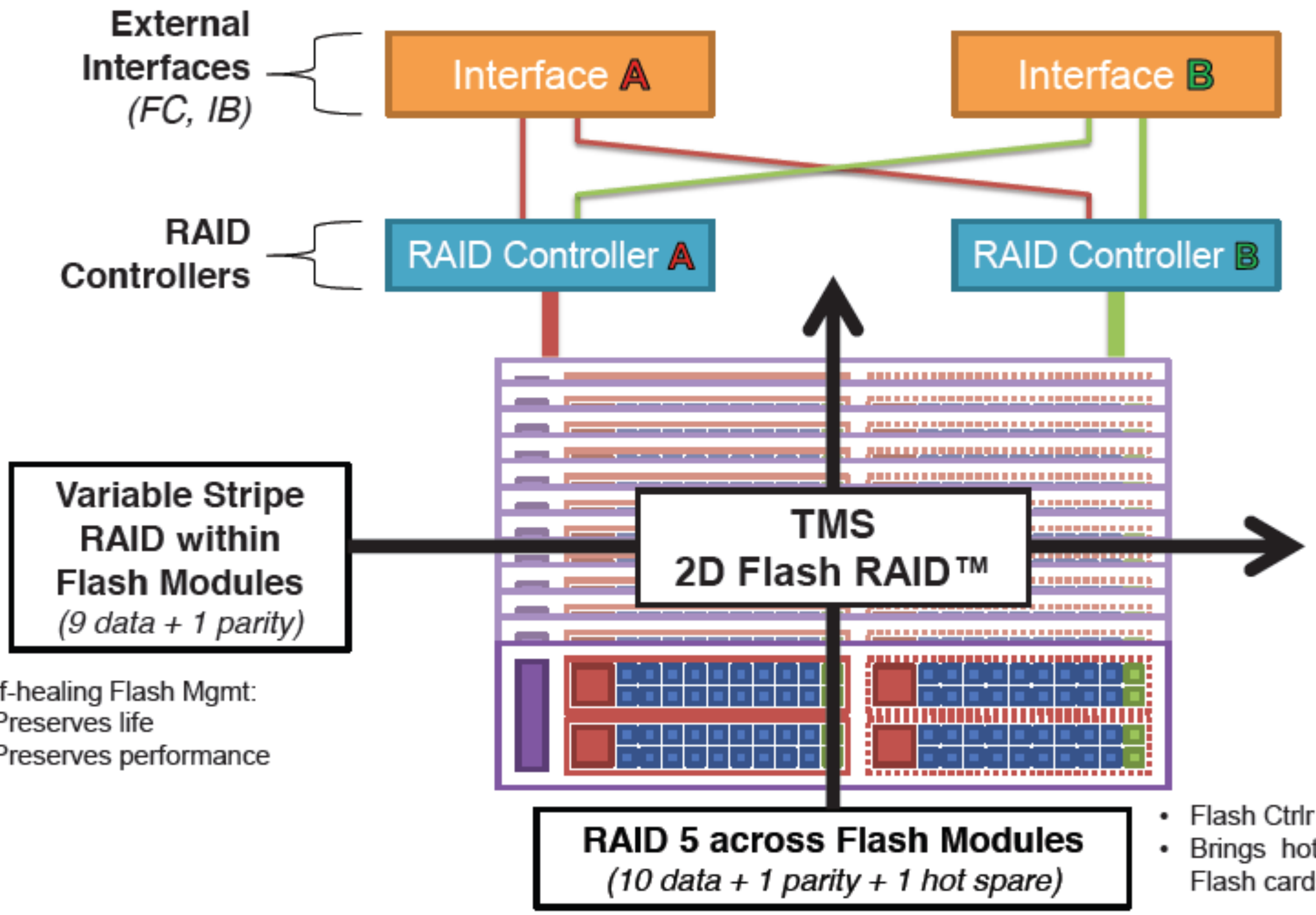
2D RAID across Flash chips
& Flash Modules
Flash Module Hot-Swap
10TB/20TB Max Capacity

Flash Quality – Components Matter

- Choose flash type based on workload profile.
- The number of Program & Erase (P/E) cycles that a given device can sustain varies with the type of technology.
- Consumer-grade Flash is multi-level cell (MLC).
- Enterprise-grade MLC (or eMLC) offers a **10x** improvement over MLC.
- Single-level cell (SLC) offers a **33x** improvement over MLC.
- eMLC flash media will handle workload profiles that most enterprise applications require.
- TMS technologies like Variable Stripe RAID™ lengthen system life by improving endurance of both eMLC and SLC.



2D Flash RAID™ (FlashSystem 720/820)

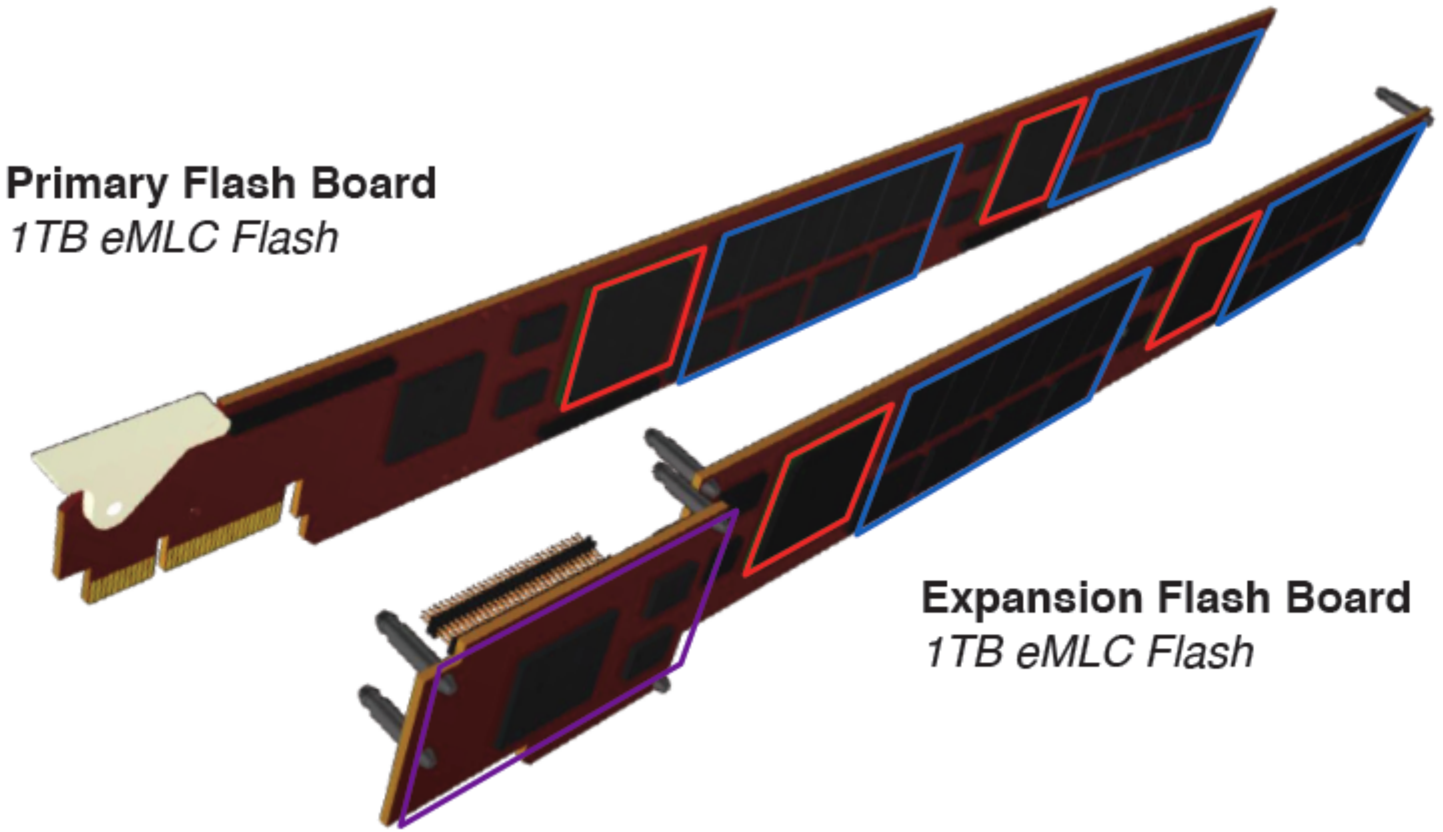

External Interfaces (FC, IB)
Interface A
Interface B
RAID Controllers
RAID Controller A
RAID Controller B
Variable Stripe RAID within Flash Modules (9 data + 1 parity)
TMS 2D Flash RAID™
RAID 5 across Flash Modules (10 data + 1 parity + 1 hot spare)

Self-healing Flash Mgmt:

- Preserves life
- Preserves performance

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

Flash Module

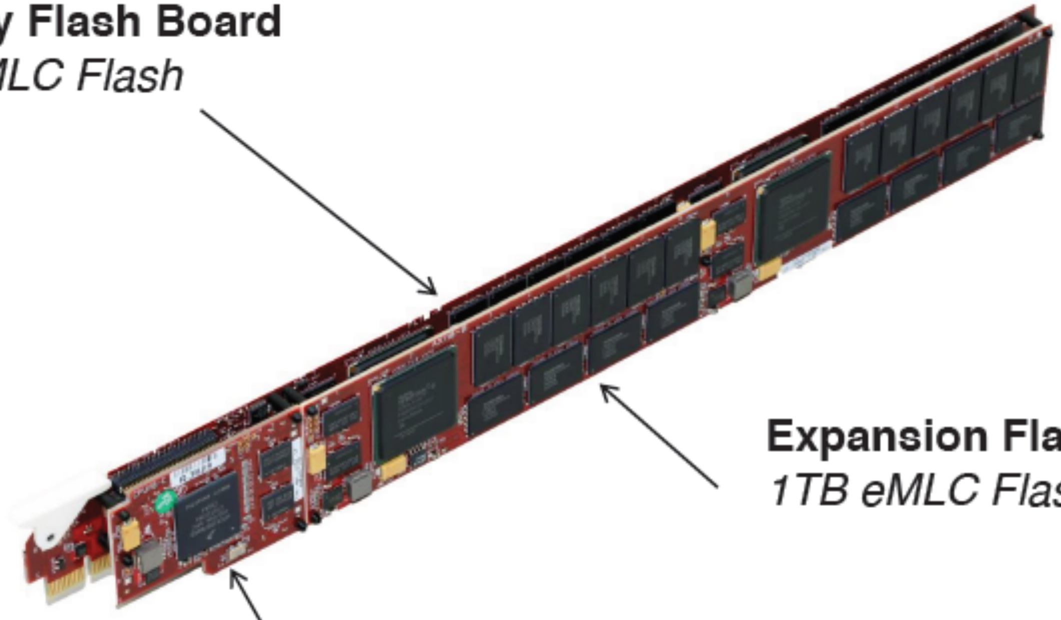


🟥 **Series-7 Flash Controller™**
2 per Board
4 per Module

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

🟪 **Gateway Interface**
Dual ports to backplane

Flash Module Live

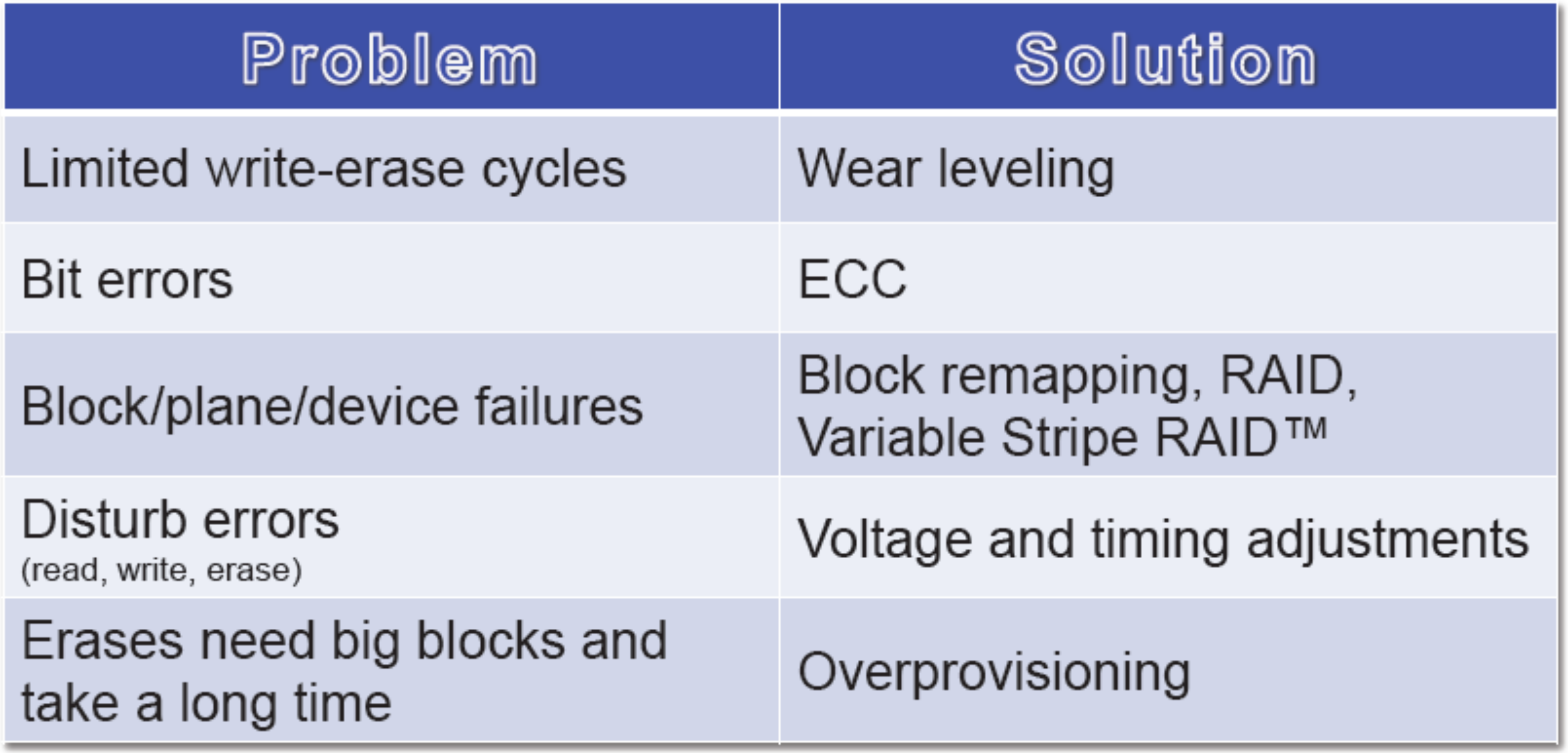


Primary Flash Board
1TB eMLC Flash

Expansion Flash Board
1TB eMLC Flash

Gateway Interface & PPC CPU

Flash Nuance Protection



Problem	Solution
Limited write-erase cycles	Wear leveling
Bit errors	ECC
Block/plane/device failures	Block remapping, RAID, Variable Stripe RAID™
Disturb errors (read, write, erase)	Voltage and timing adjustments
Erases need big blocks and take a long time	Overprovisioning

Understanding SPC-1

- Metrics driving performance:
  Queue / Avg. Time p/IO = IOPS

Queue: Queue Depth

Limited control in server

Application dependent

Avg. Time p/IO: Response Time

Latency from array

Function of how fast data is R/W

- **200K IOPS** equals:
  - $2.97M HP 3PAR 10K V800
    612 Threads
  - $2.62M in HP EVA w/SSD
    198 Threads
  - $490K Kaminario K2-D (DRAM)
    87 Threads
  - **$400K RamSan (Flash)**
    68 Threads

RamSan delivers maximum work per CPU and highest application efficiency.

Performance Scenario: Oracle RAC, 4 Nodes

Enterprise Array, No Flash

2 million queries
12.25 minutes to complete

16K Total IOPS
4K per RAC Node

```
[oracle]$ time ./spawn_50.sh

real     12m15.434s
user     0m5.464s
sys      0m4.031s
```

FlashSystem

2 million queries
1.3 minutes to complete

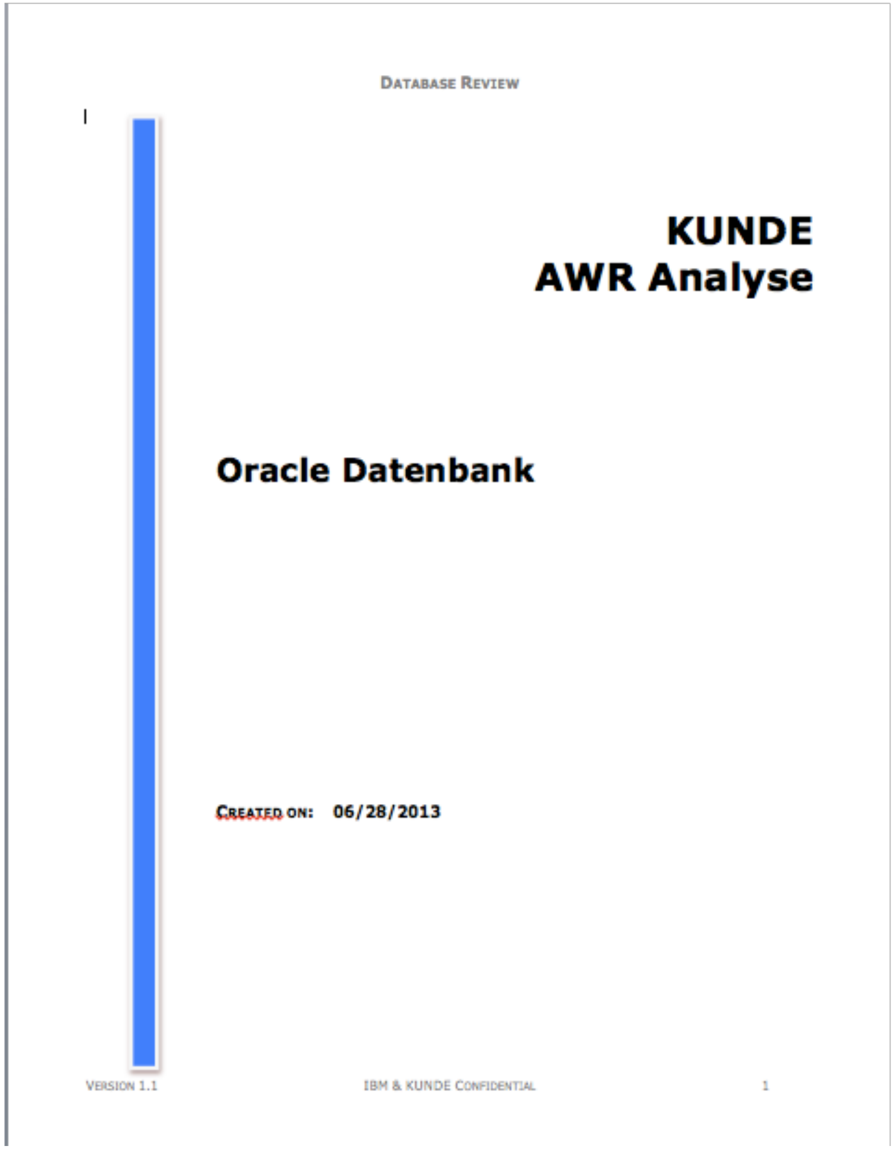
160K Total IOPS
40K per RAC Node

```
[oracle]$ time ./spawn_50.sh

real     1m19.838s
user     0m4.439s
sys      0m3.215s
```

A factor of over 9x improvement!

Performance prediction with Oracle AWR report



DATABASE REVIEW

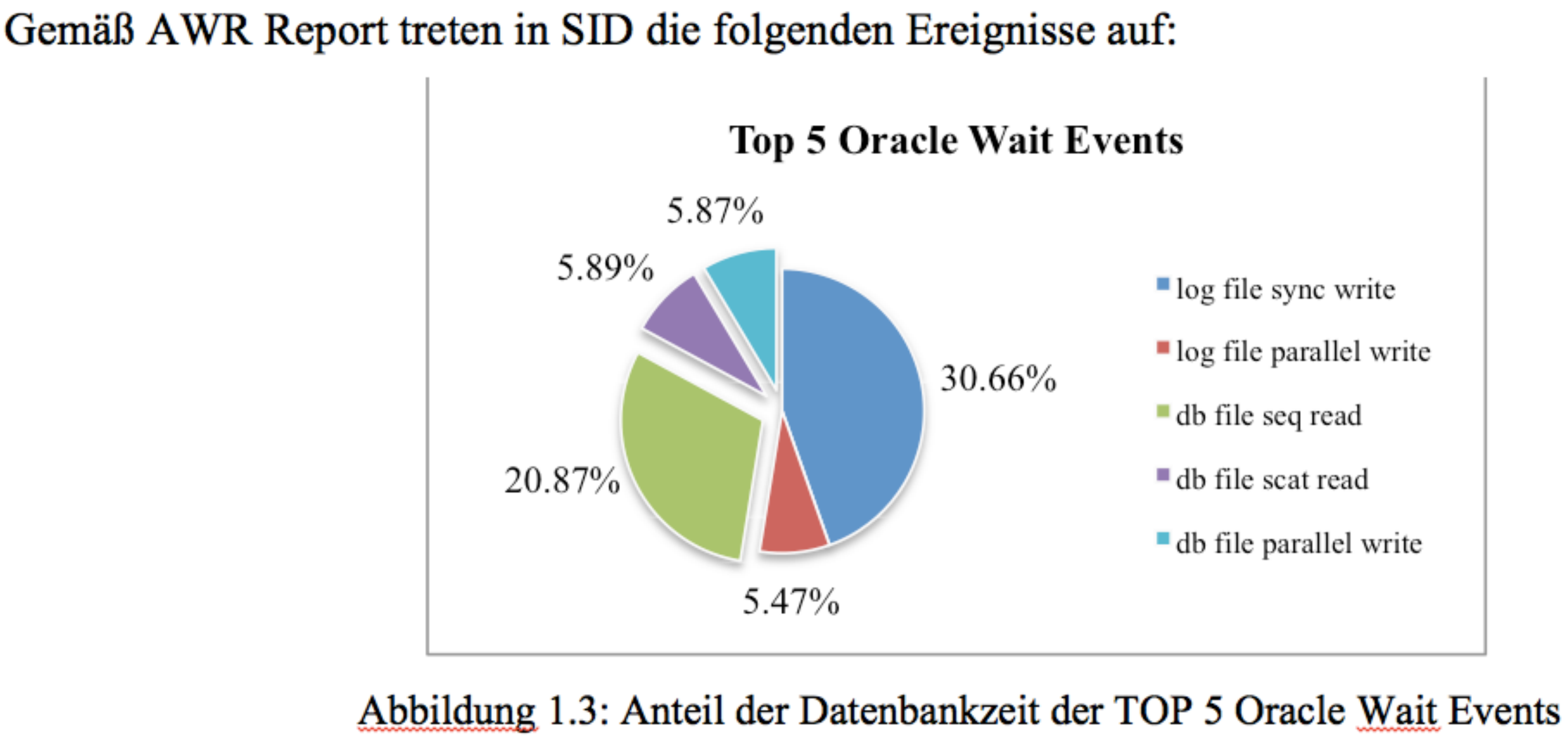
KUNDE
AWR Analyse

Oracle Datenbank

CREATED ON: 06/28/2013

VERSION 1.1 IBM & KUNDE CONFIDENTIAL 1

Performance prediction with Oracle AWR report



Gemäß AWR Report treten in SID die folgenden Ereignisse auf:

Top 5 Oracle Wait Events

Abbildung 1.3: Anteil der Datenbankzeit der TOP 5 Oracle Wait Events

Performance prediction with Oracle AWR report

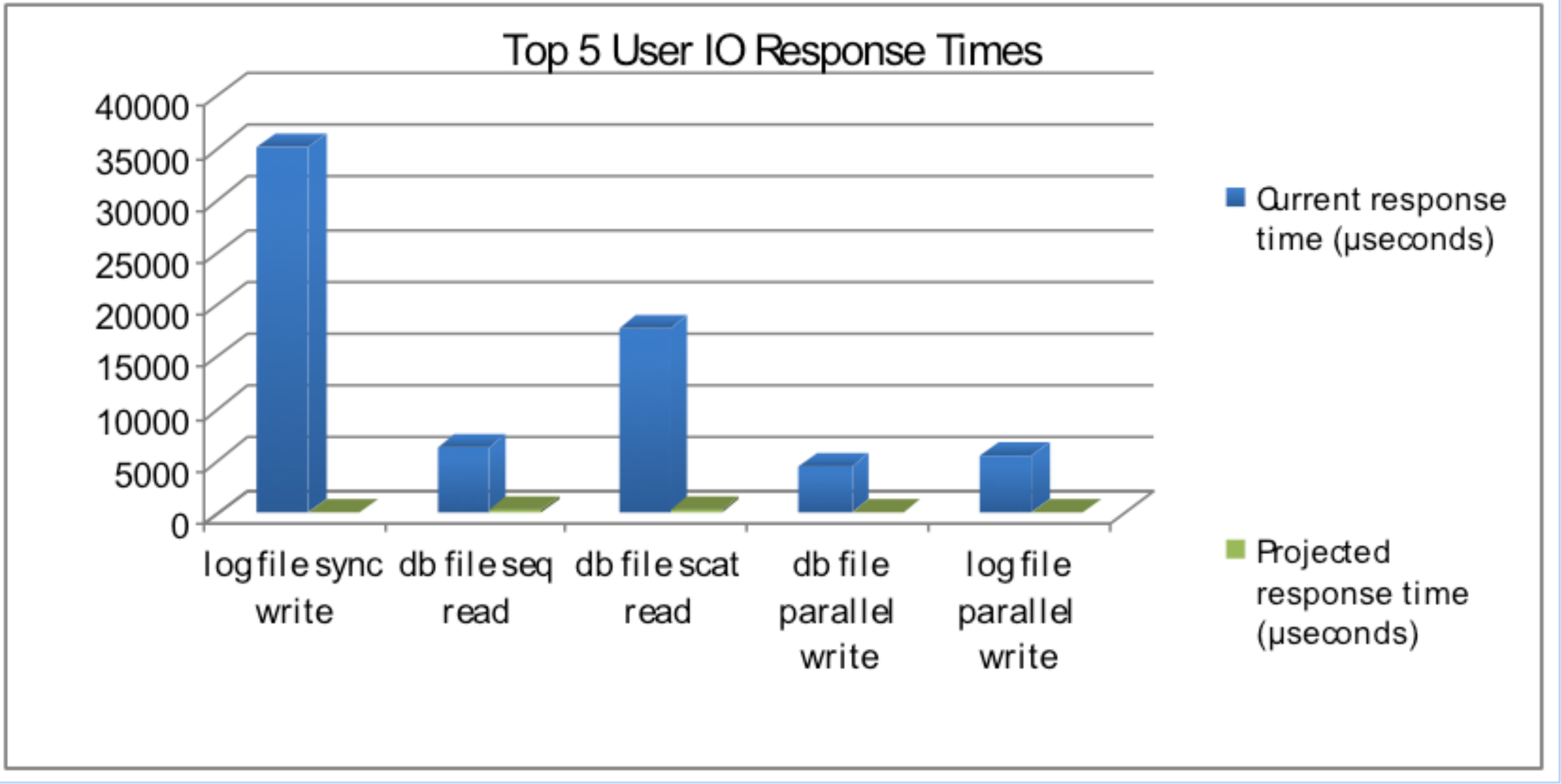
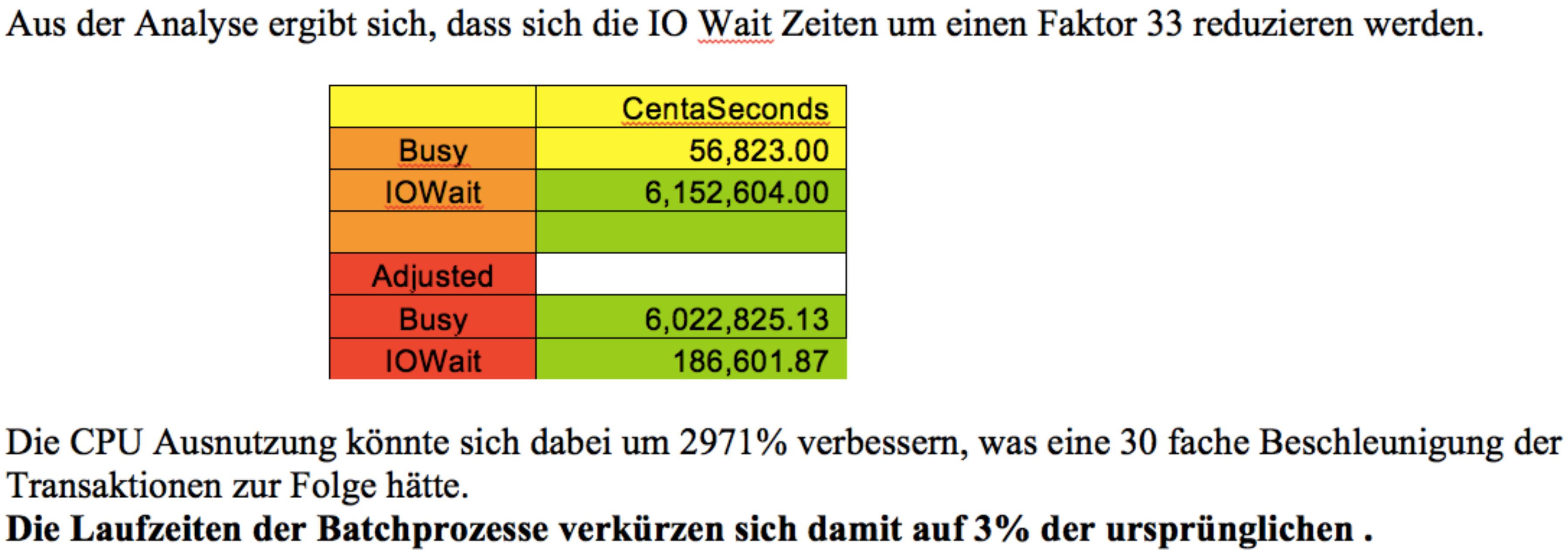


Abbildung 2.2: Erwartete SID Wait time Verbesserungen mit IBM FlashSystem

Performance prediction with Oracle AWR report



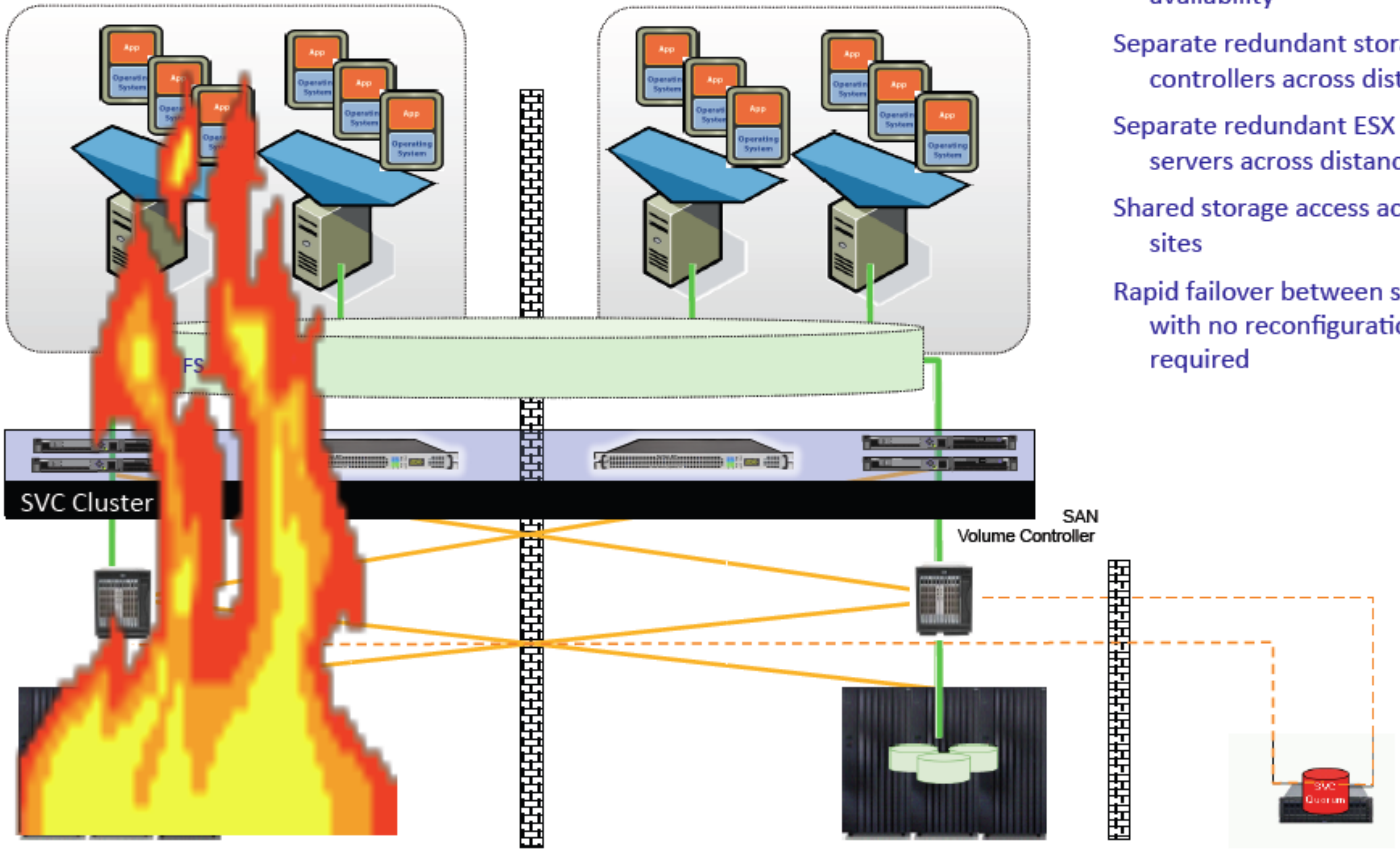
Aus der Analyse ergibt sich, dass sich die IO Wait Zeiten um einen Faktor 33 reduzieren werden.

	CentaSeconds
Busy	56,823.00
IOWait	6,152,604.00
Adjusted	
Busy	6,022,825.13
IOWait	186,601.87

Die CPU Ausnutzung könnte sich dabei um 2971% verbessern, was eine 30 fache Beschleunigung der Transaktionen zur Folge hätte.

Die Laufzeiten der Batchprozesse verkürzen sich damit auf 3% der ursprünglichen .

IBM San Volumes Controller – adding functionality again



Cost effective multi-site high availability

Separate redundant storage controllers across distance

Separate redundant ESX servers across distance

Shared storage access across sites

Rapid failover between sites with no reconfiguration required

Vielen Dank für Ihre Aufmerksamkeit

Haben Sie noch Fragen ?