Future storage, I/O, and data management

4th ENES HPC WS, Toulouse

Dr. Oliver Oberst
07 April 2016

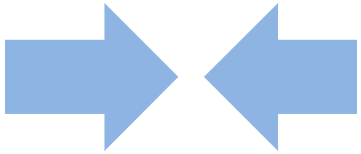
IBM

Data Centric

Big Data Driving Common Requirements

High Performance Analytics
- Unstructured data
- Primarily data mining

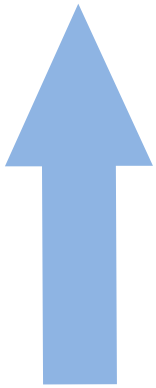
High Performance Computing
- Structured data
- Primarily scientific calculations/Simulation



Evolving requirements

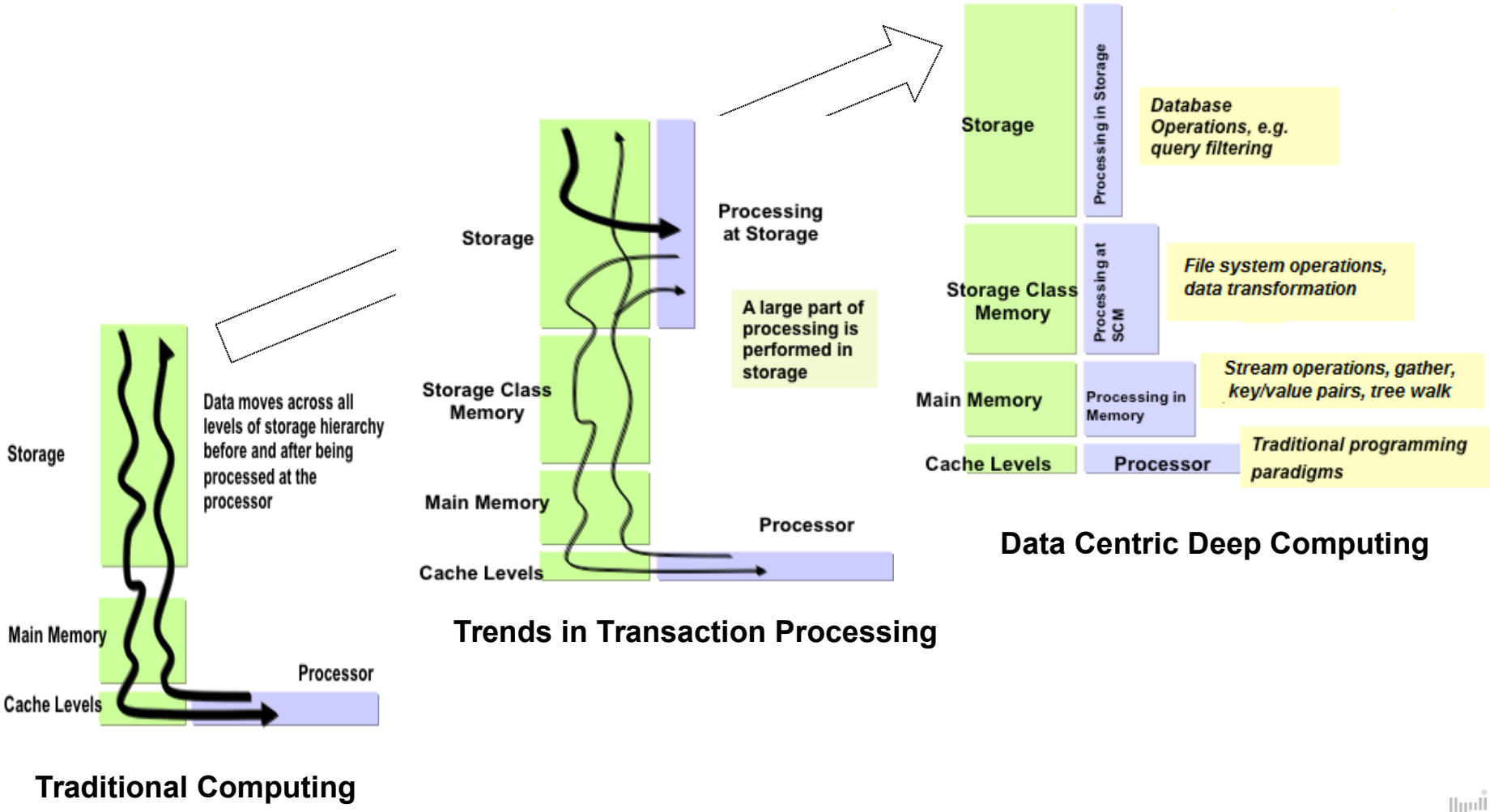
Driver: Enhanced context Improves decision making
- Incorporate modeling and simulation for better predictions
- Incorporate sensor data

Driver: Doing more with models
- Real-time decision making
- Uncertainty quantification
- Sensitivity analysis
- Metadata extraction

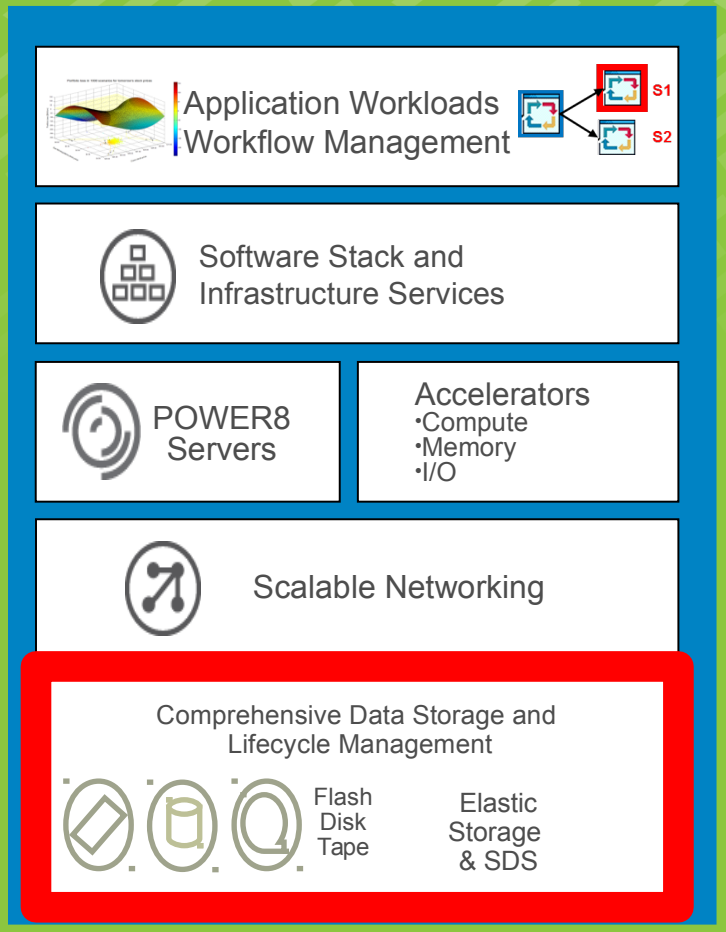


Data Centric Systems

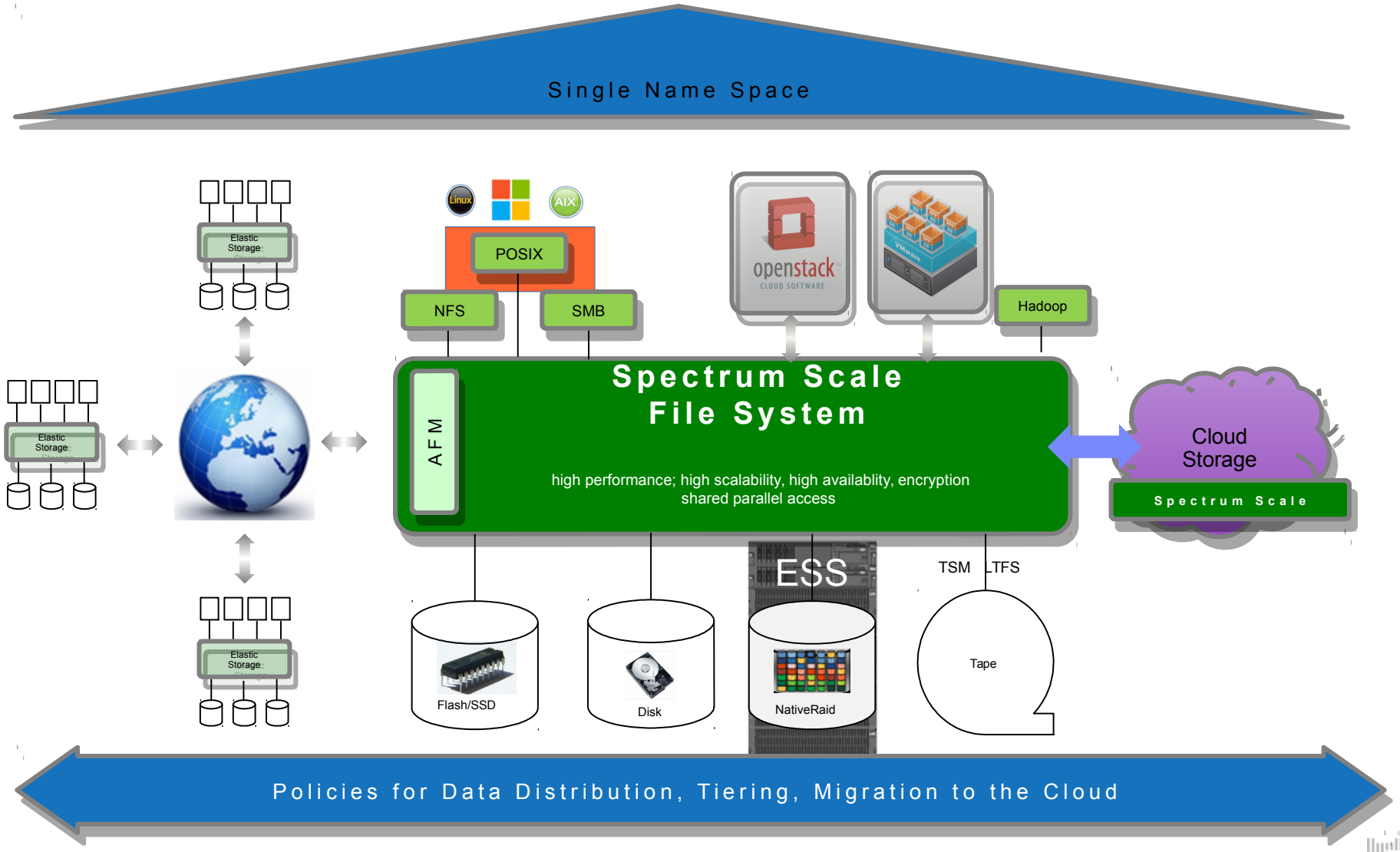
Optimized System Design for Data Centric Computing



Data Management

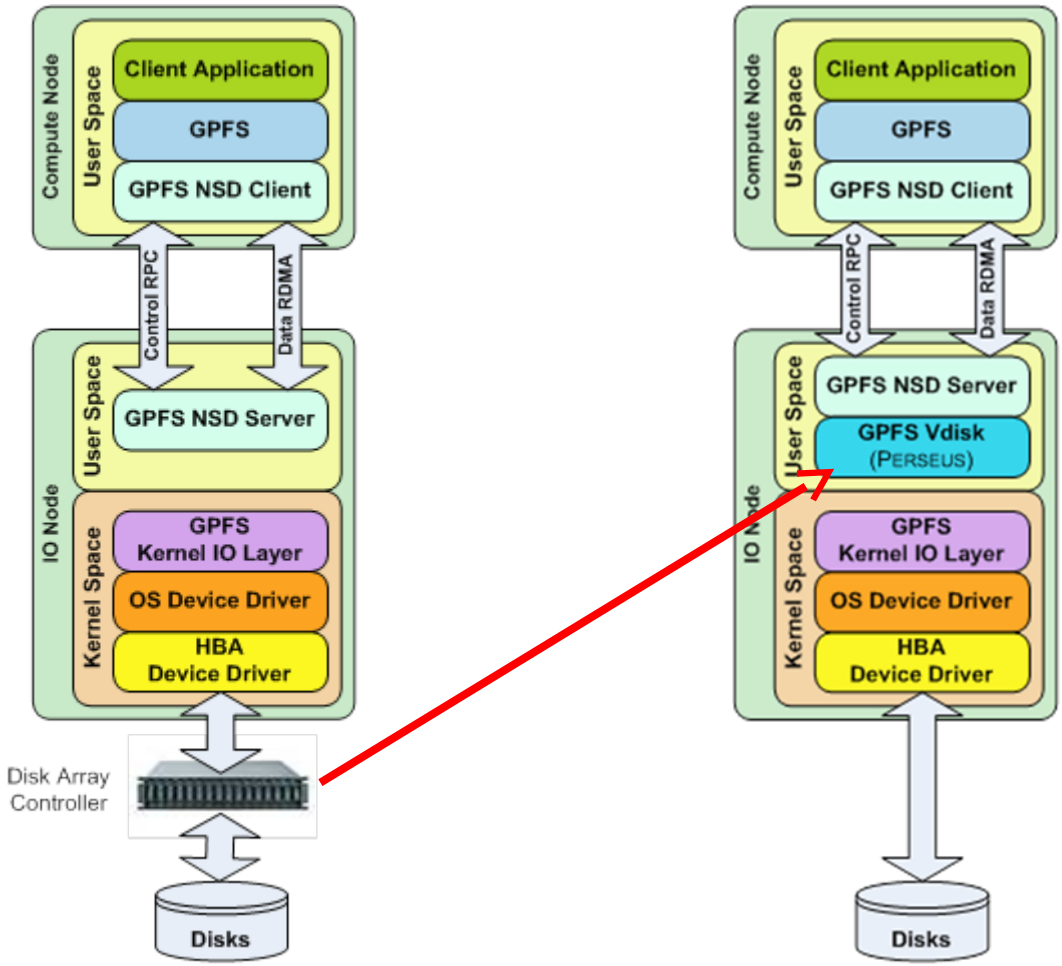


Single Name Space = Less Silos = Ease of Data Management



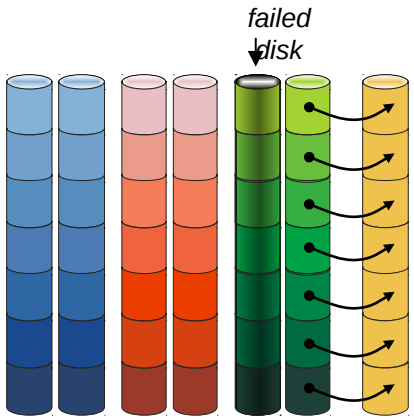
GPFS Native Raid (GNR)
used in Elastic Storage Server ESS

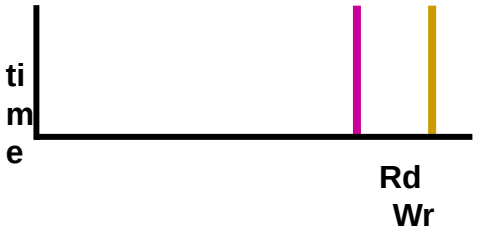
Typicall
Raid
System



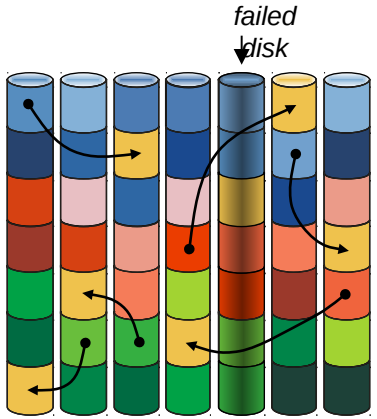
GPFS Native Raid

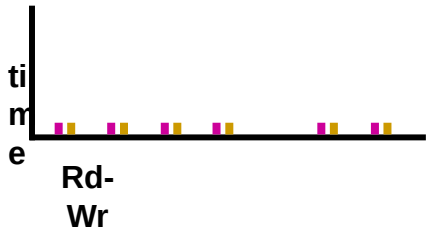
GPFS Native Raid (GNR)





Rebuild activity confined to just a few disks – slow rebuild, disrupts user programs

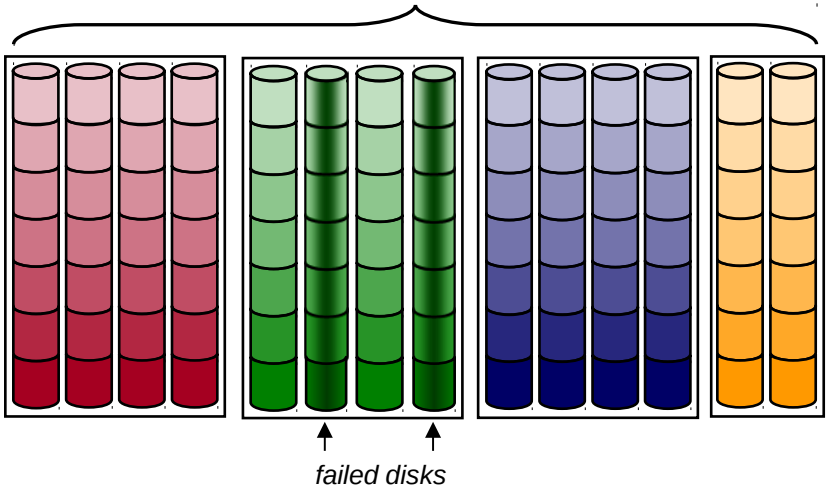




Rebuild activity spread across many disks, less disruption to user programs

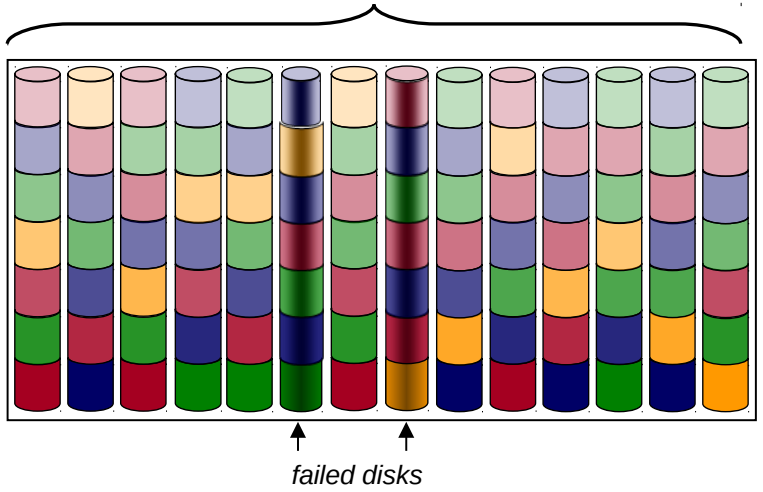
GPFS Native Raid (GNR)

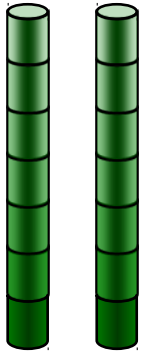
14 physical disks / 3 traditional RAID6 arrays / 2 spares



Decluster data, parity and spare

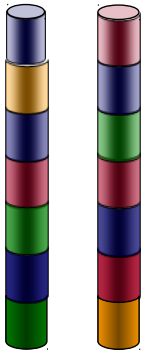
14 physical disks / 1 declustered RAID6 array / 2 spares





Number of faults per stripe		
Red	Green	Blue
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0
0	2	0

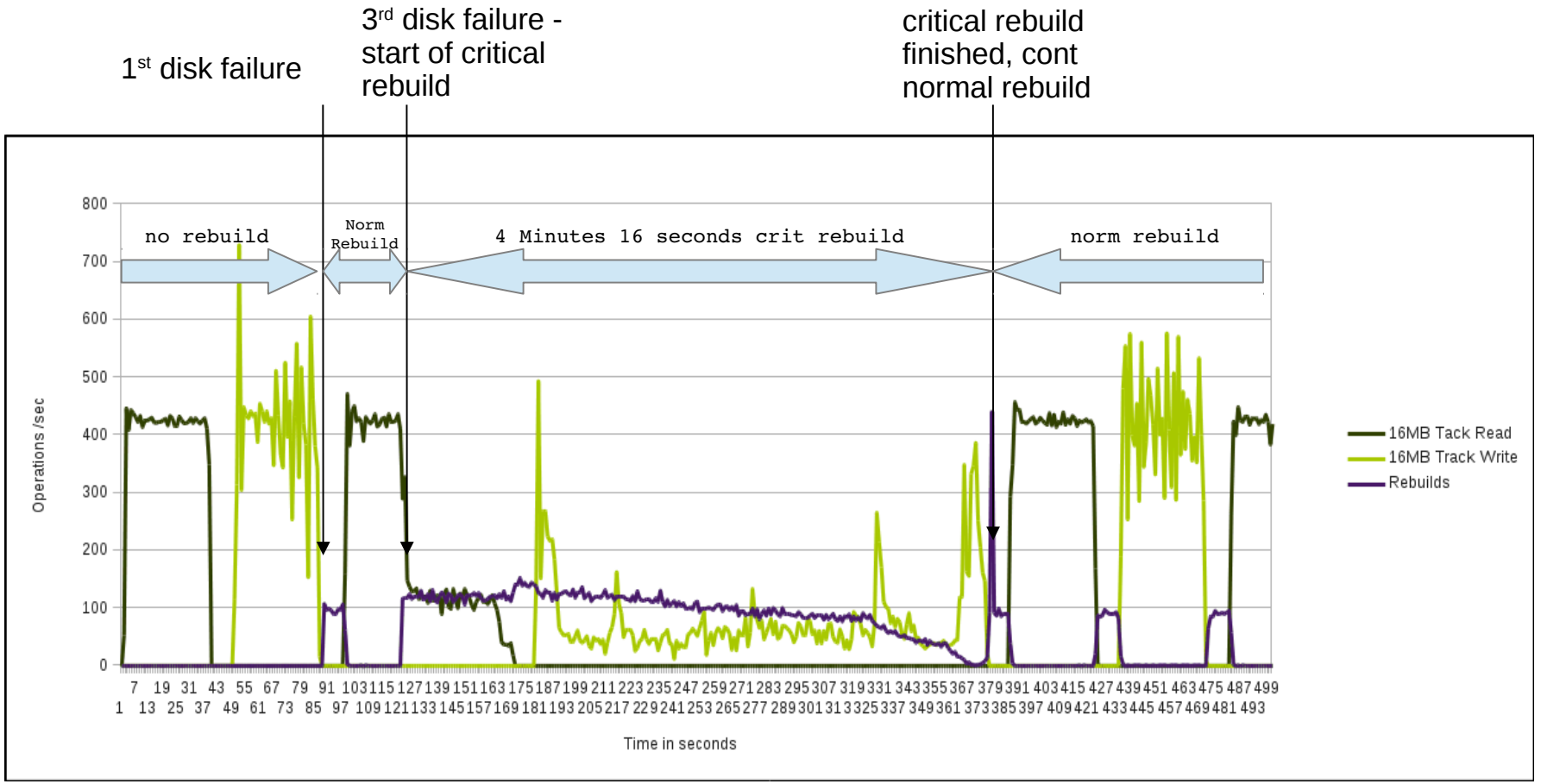
Number of stripes with 2 faults = 7



Number of faults per stripe		
Red	Green	Blue
1	0	1
0	0	1
0	1	1
2	0	0
0	1	1
1	0	1
0	1	0

Number of stripes with 2 faults = 1

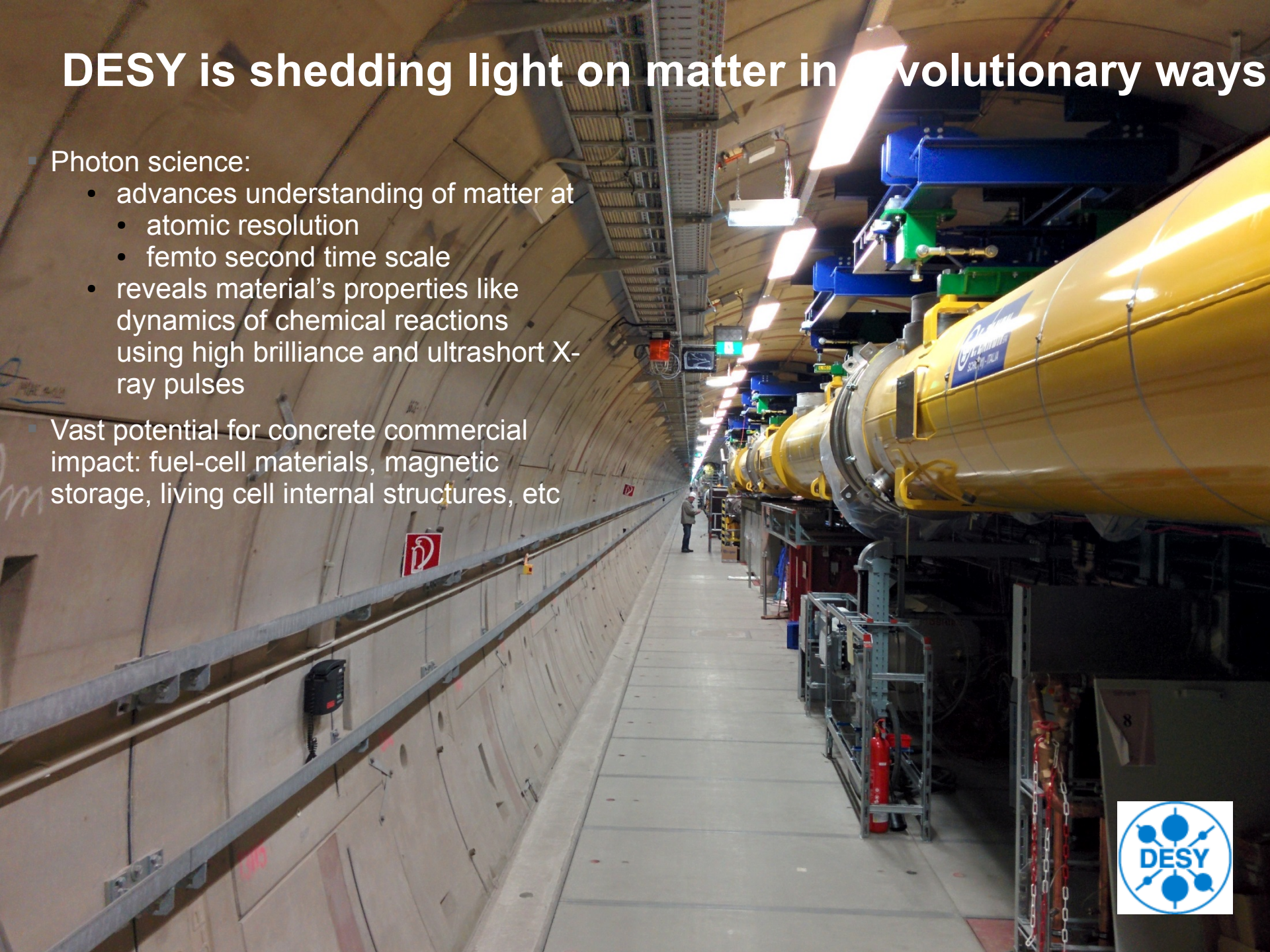
Rebuild Benchmark



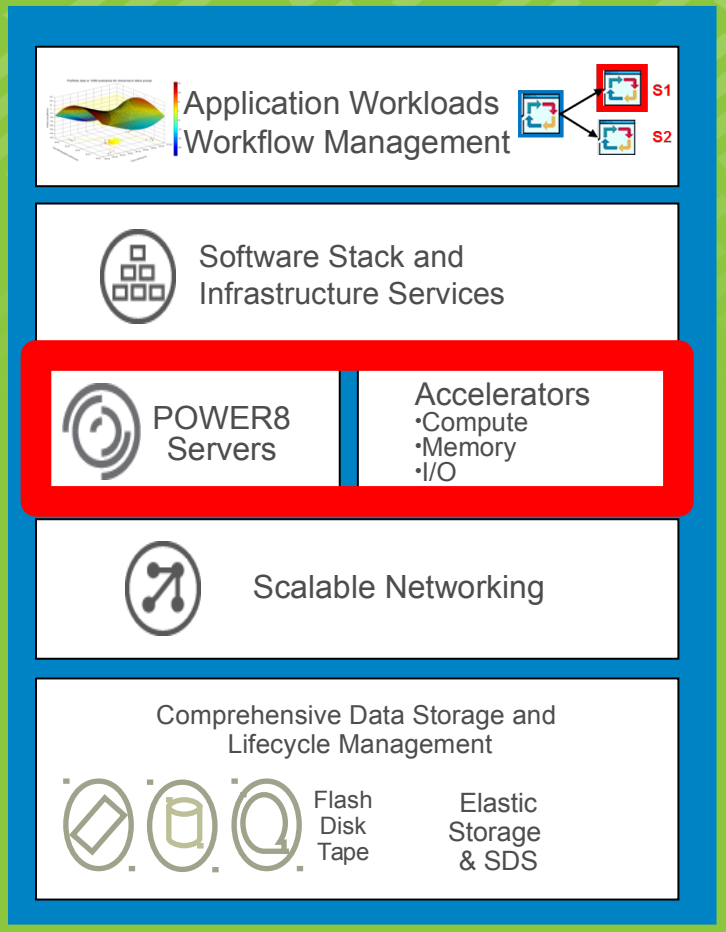
During the short critical rebuild time the impact on workload was higher, but as soon as we were back to double parity (+2P) the impact to the end user workload was less than 5%

DESY is shedding light on matter in revolutionary ways

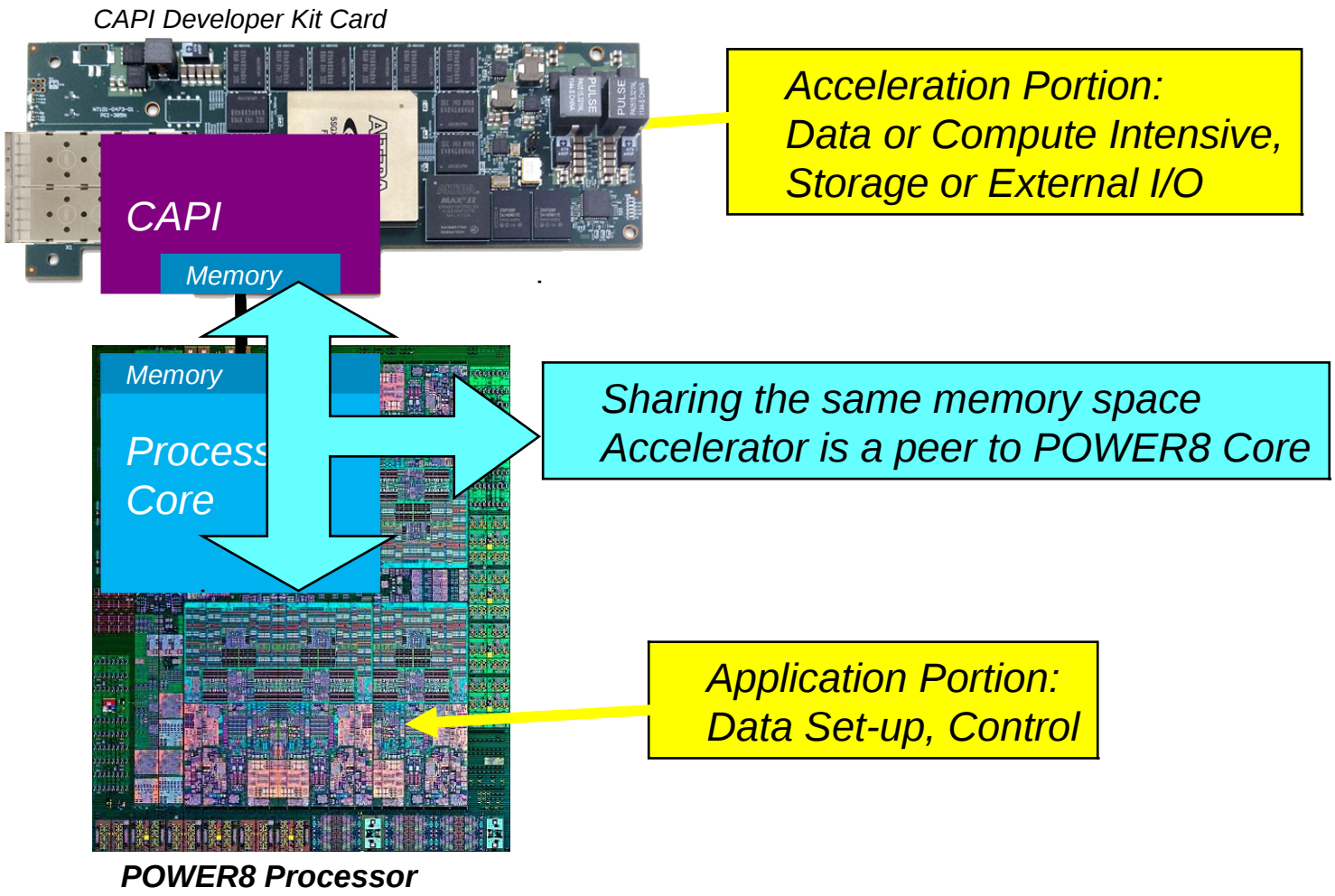
- Photon science:
  - advances understanding of matter at
    - atomic resolution
    - femto second time scale
  - reveals material's properties like dynamics of chemical reactions using high brilliance and ultrashort X-ray pulses
- Vast potential for concrete commercial impact: fuel-cell materials, magnetic storage, living cell internal structures, etc



I/O

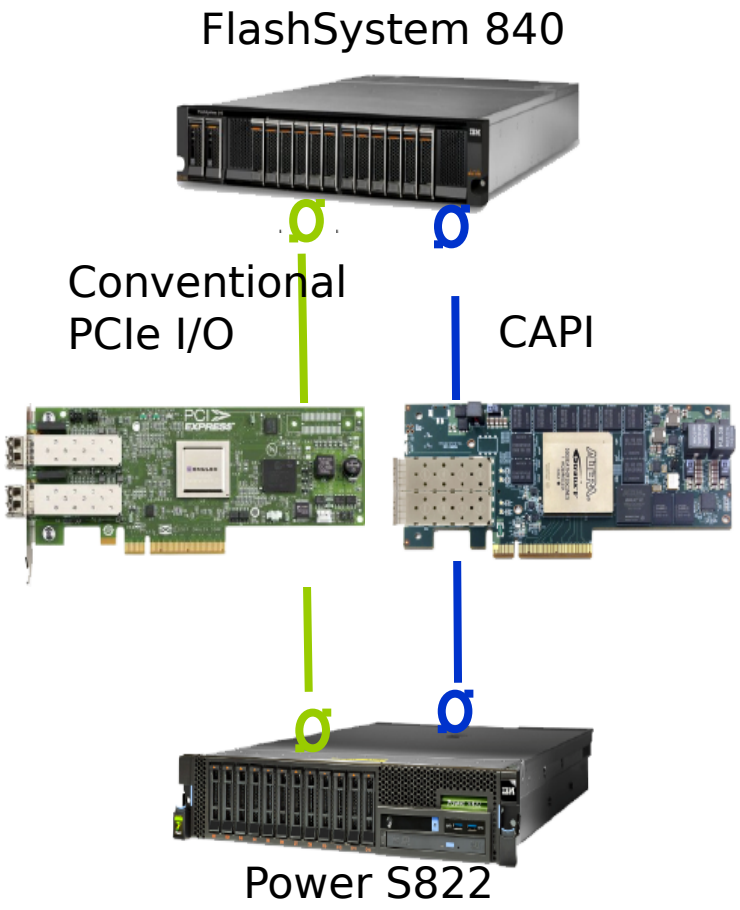

Application Workloads
Workflow Management
S1
S2
Software Stack and
Infrastructure Services
POWER8
Servers
Accelerators
•Compute
•Memory
•I/O
Scalable Networking
Comprehensive Data Storage and
Lifecycle Management
Flash
Disk
Tape
Elastic
Storage
& SDS

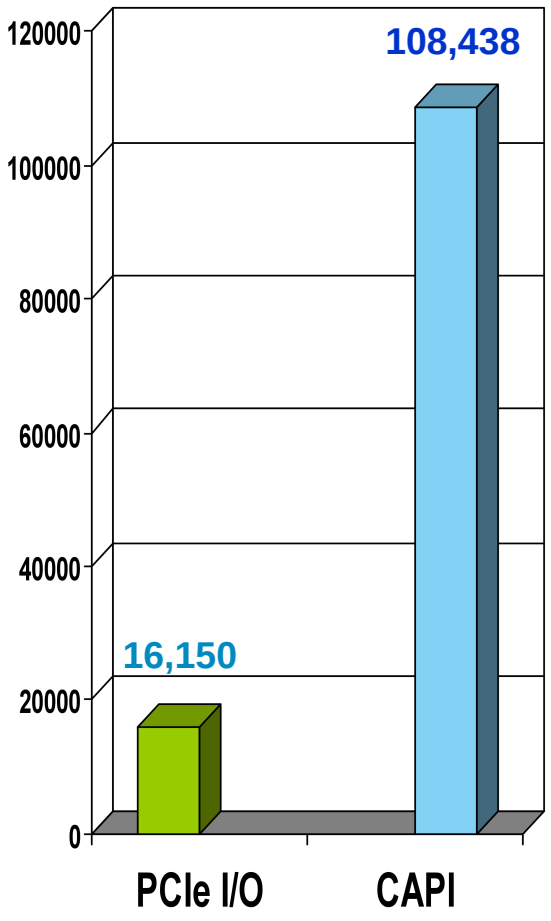
OpenPower - How CAPI Works



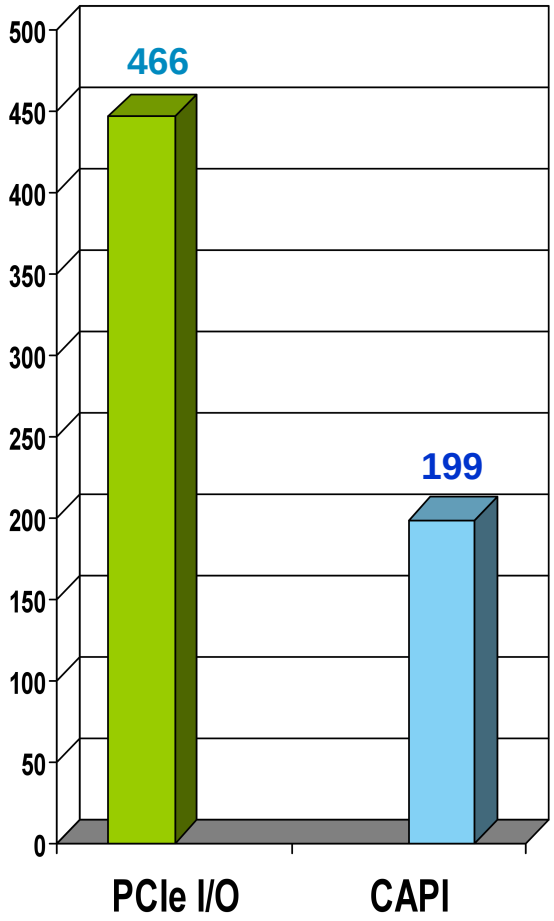
Demonstrating the Value of CAPI Attachment for DataCentric Workloads

Identical hardware with
2 different paths to data





IOPs per HW Thread



Latency (us)

IBM Accelerated GZIP Compression

What it is:

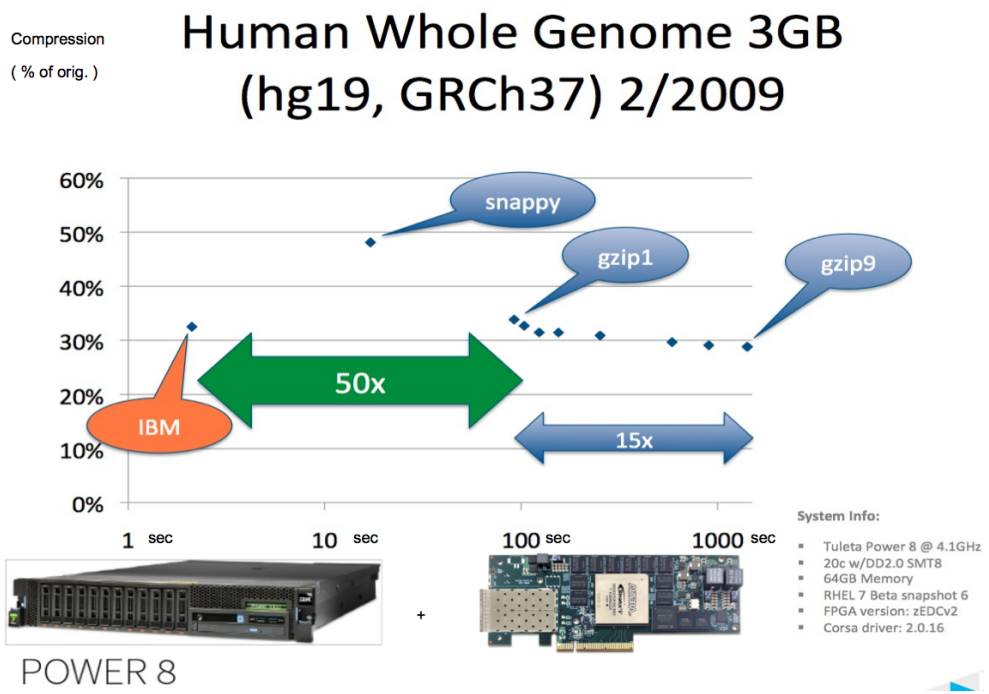
- An FPGA-based low-latency GZIP Compressor & Decompressor with.

Results:

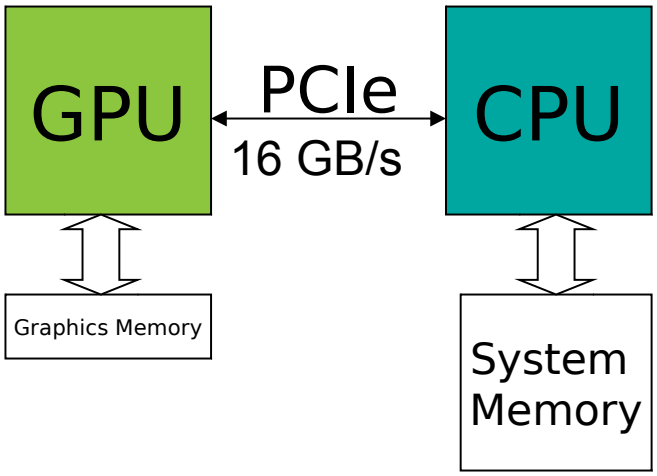
- **Single-thread** througput of ~2GB/s and a compression rate significantly better than low-CPU overhead compressors like snappy

Source:

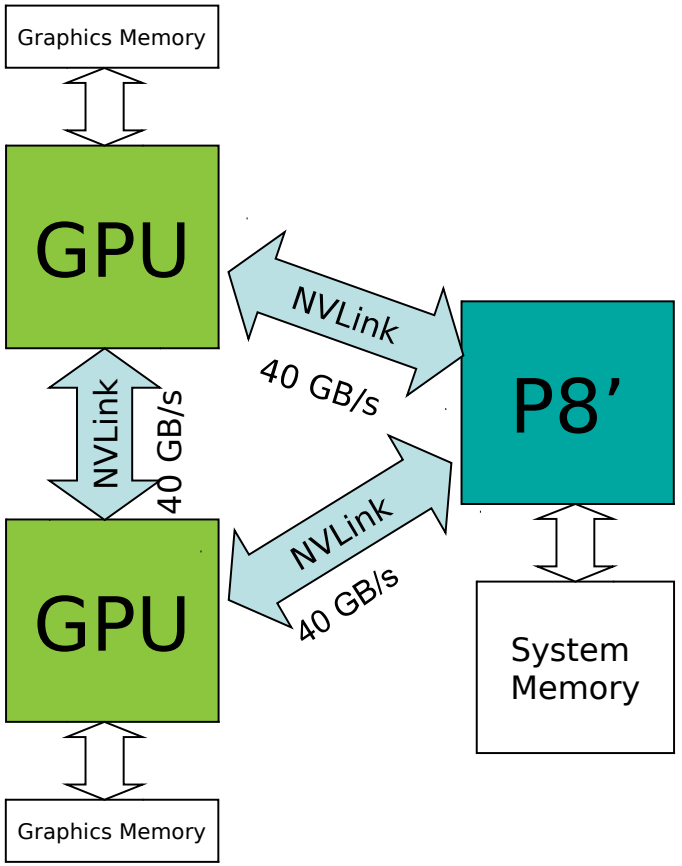
- Non-published results



2.5x Faster CPU-GPU Connection via NVLink

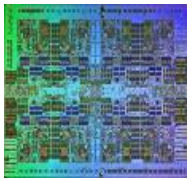
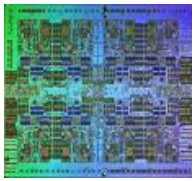
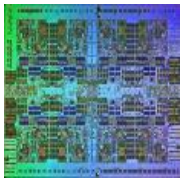
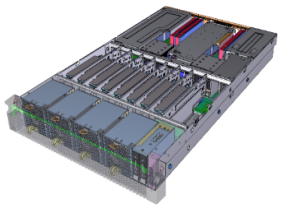
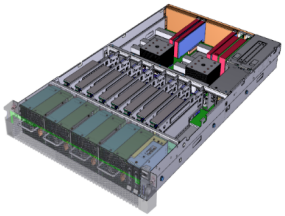
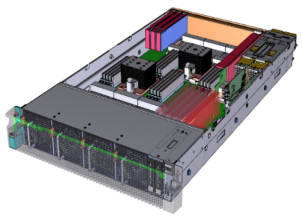


GPUs Bottlenecked by PCIe Bandwidth From CPU-System Memory



NVLink Enables Fast Unified Memory Access between CPU & GPU Memories

IBM OpenPOWER-based HPC Roadmap

	2015	2016	2017
Mellanox Interconnect Technology	Connect-IB FDR Infiniband PCIe Gen3	ConnectX-4 EDR Infiniband CAPI over PCIe Gen3	ConnectX-5 Next-Gen Infiniband Enhanced CAPI over PCIe Gen4
NVIDIA GPUs	Kepler PCIe Gen3	Pascal NVLink	Volta Enhanced NVLink
IBM CPUs	POWER8  OpenPower CAPI Interface	POWER8+  NVLink	POWER9  Enhanced CAPI & NVLink
IBM Nodes			

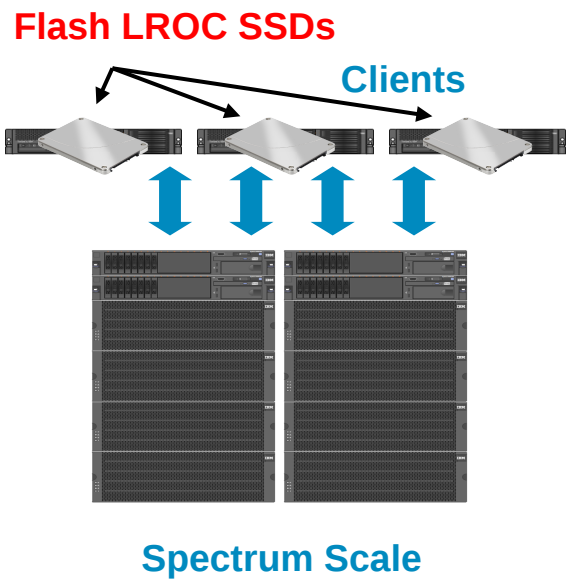
Thank you

- Data growth leads to evolving requirements for the infrastructure architecture
- Unified file name space across different access methods and across different storage tiers eases data management
- Use Accelerators (FPGAs) to improve I/O
- Improve bandwidth between CPU and accelerators

Backup

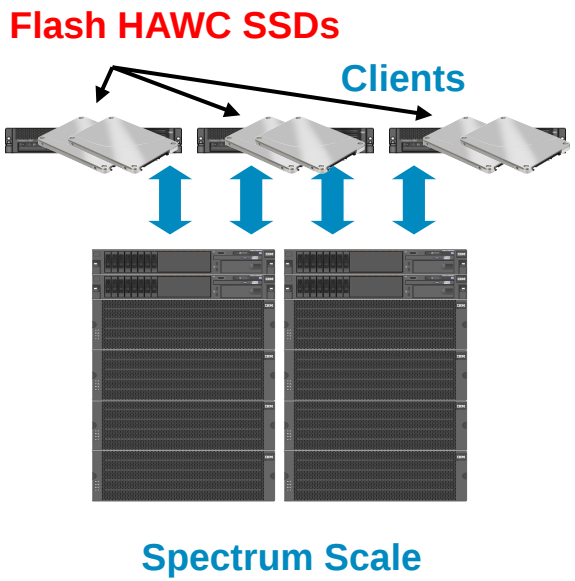
Flash Local Read Only Cache (LROC)

- Inexpensive SSDs placed directly in Client nodes
- Accelerates I/O performance up to **6x** by reducing the amount of time CPUs wait for data
- Also decreases the overall load on the network, benefitting performance across the board
- Improves application performance while maintaining all the manageability benefits of shared storage
- Cache consistency ensured by standard tokens
- Data is protected by checksum and verified on read
- IBM Spectrum Scale handles the flash cache automatically so data is transparently available to your application with very low latency and no code changes

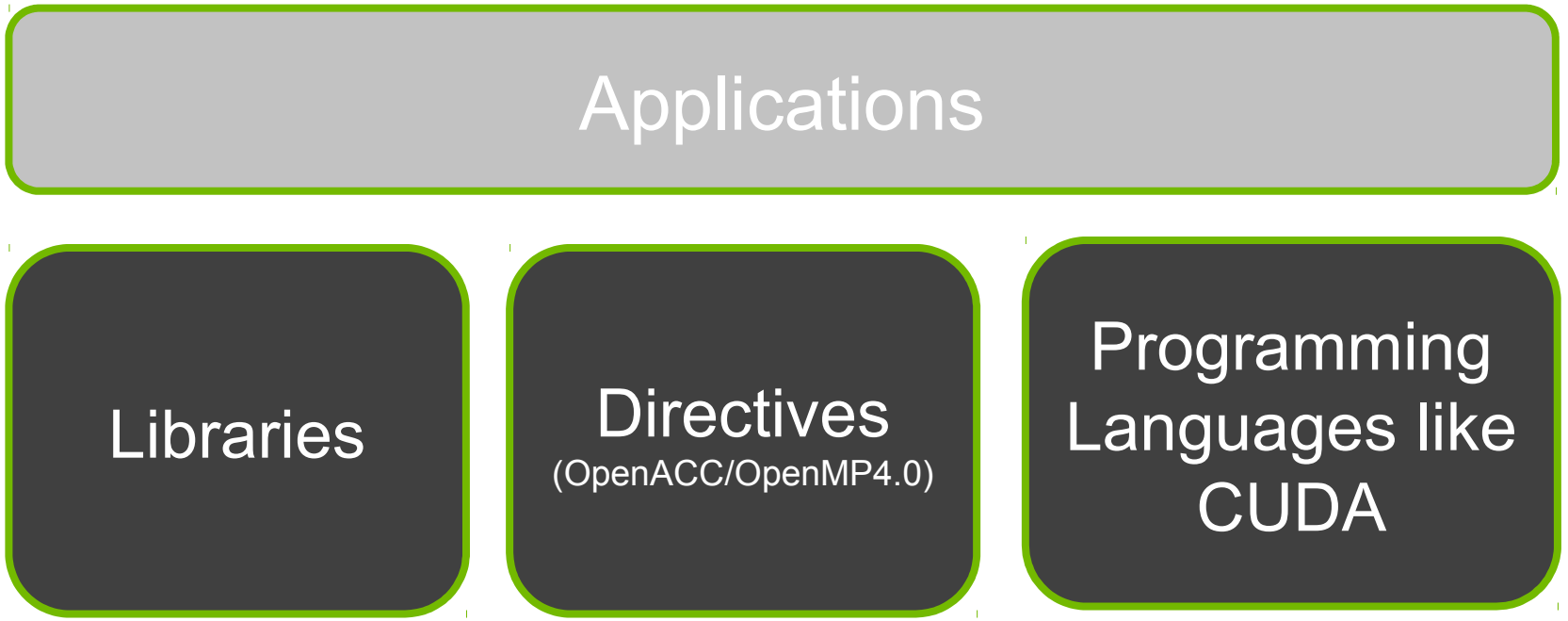


High Available Write Cache (HAWC)

- Inexpensive SSDs placed directly in Client nodes
- Reduces the latency of small write requests by initially hardening data in a non-volatile fast storage device prior to writing it back to the backend storage system.
- Applications that exhibit this type of write behavior include VMs, databases, and log generation.
- In general speedups should be seen in any environment that either currently lacks fast storage or has very limited (and non-scalable) amounts of fast storage.



3 Ways to Accelerate Applications



US & UK Research Establishments Select OpenPOWER-Based Supercomputers

IBM, Mellanox, and NVIDIA awarded $325M U.S. Department of Energy's CORAL Supercomputers

CORAL: Leadership Class Supercomputers

5X – 10X HIGHER APP PERF THAN CURRENT SYSTEMS

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

IBM & UK's STFC Partner for Big Data & Cognitive Computing Research in £313M Partnership

Science & Technology Facilities Council

HM Government

OpenPOWER™

IBM Research

IBM Watson

Different Solutions for Different Parts of the Cube