Paving the way for Exascale: Lessons learn from I/O accelerators

Jean-Thomas Acquaviva, DDN

May, 2016

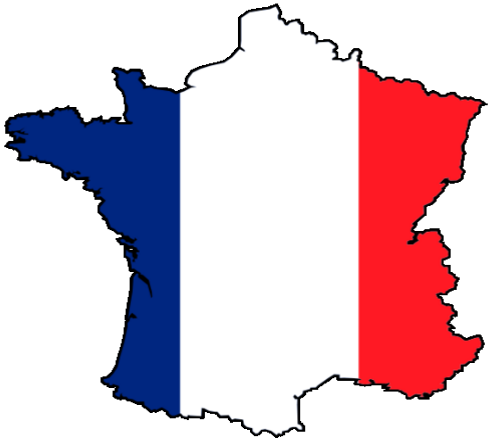
Extreme Scale Demonstrator, Prague

Corporate Status: DDN Advanced Technical Center

2

R&D centered on Emerging tech. programs, **Paris, France**

- 25+ R&D engineers



I/O Bandwidth Requirements

As seen from checkpoint restart needs

Bandwidth needs next-gen pre-Exascale systems

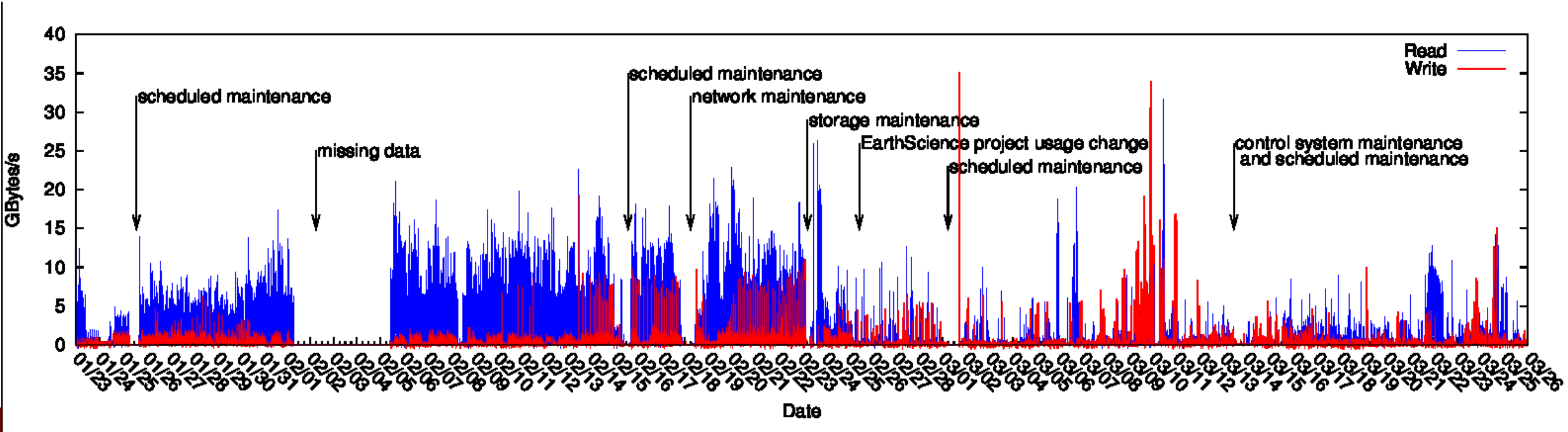
Rules of thumb:

1/ Checkpointing less than 6 minutes per hour
2/ Checkpointing means draining half of system memory

Pre-Exascale system:
4 Petabyte → bandwidth requirement 5.6 TB/s

Oakridge lab..

Irregular I/O Bandwidth Pressure



99% of the time the IO sub-system is stressed bellow 30% of its bandwidth

70% of the time the system is stress under 5% of its peak bandwidth

Argone lab.

P. Carns, K. Harms et al., *Understanding and Improving Computational Science Storage Access through Continuous Characterization, 2011*

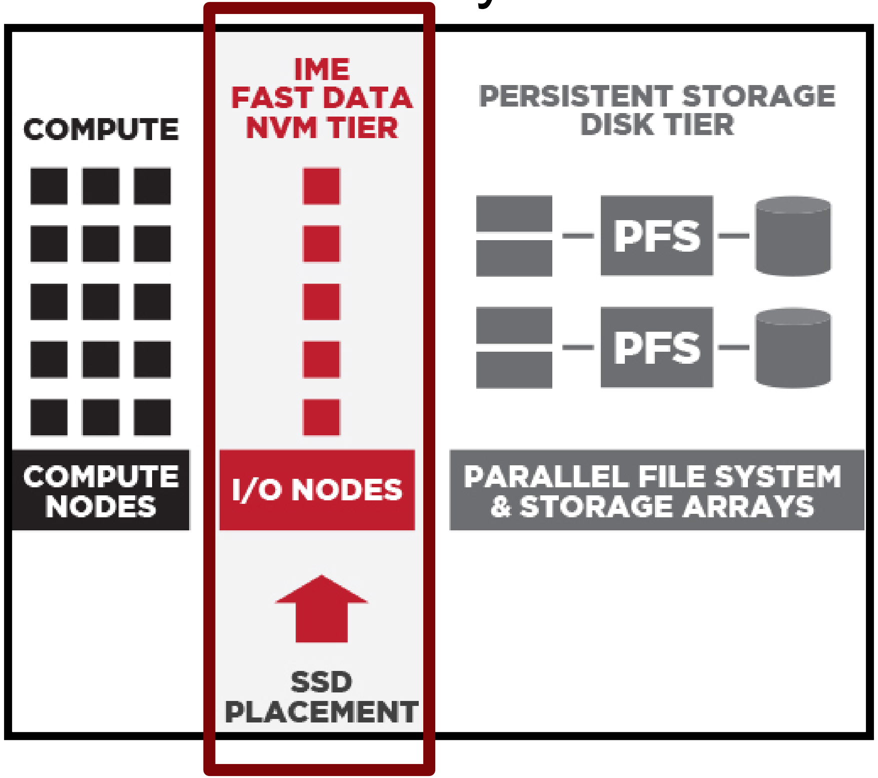
What is IME?

Distributed Virtually Shared Coherent Array of SSDs

SSD reshuffles the parameters
Latency / 40 : 4ms → 0,1 ms
Bandwidth x 3: 150 → 450 MB/s
Capacity / 8 : 8 → 1TB.

Cost x 10 $ 0,05/Gbit → $0.04

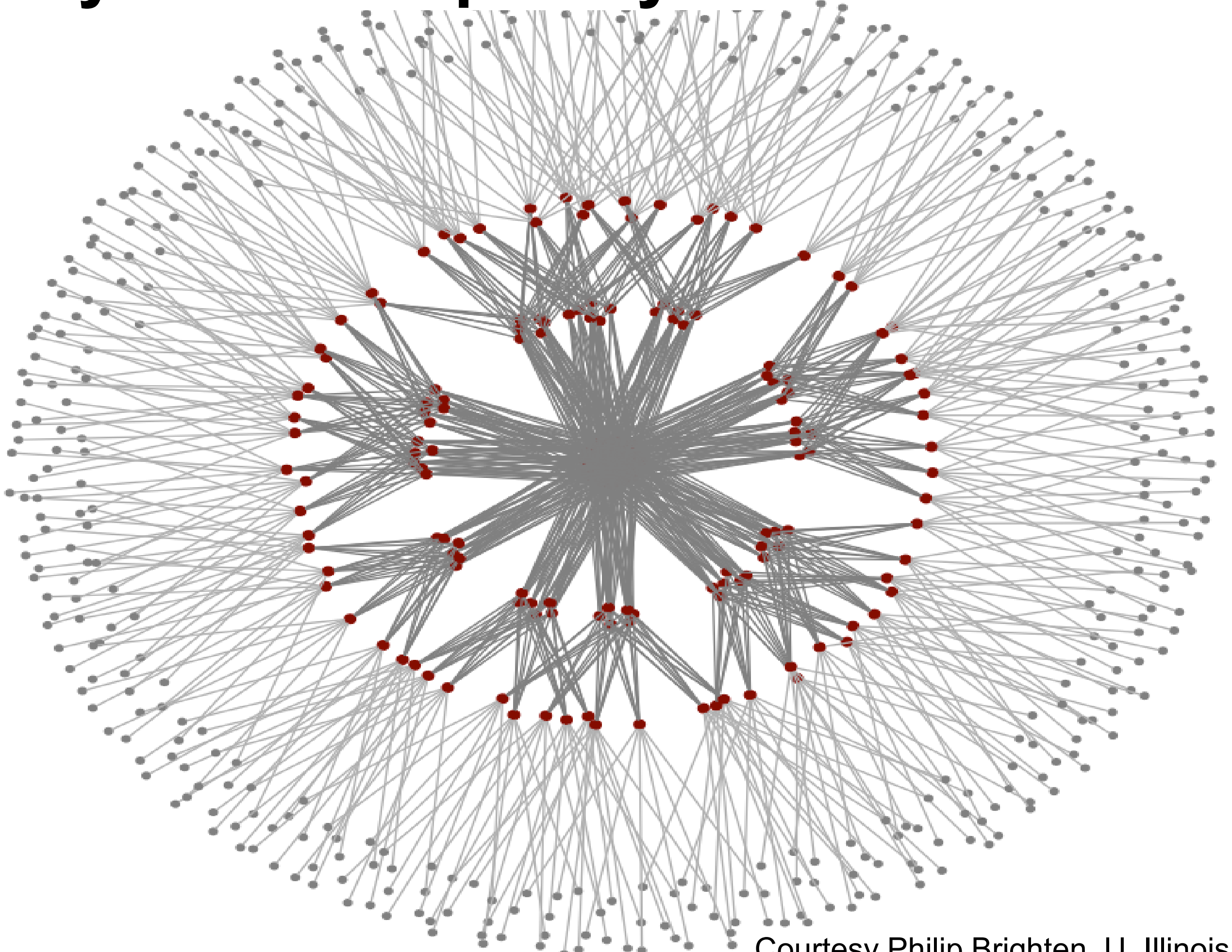
What can we do with a costly high bandwidth low latency technology?



Dealing with System Complexity

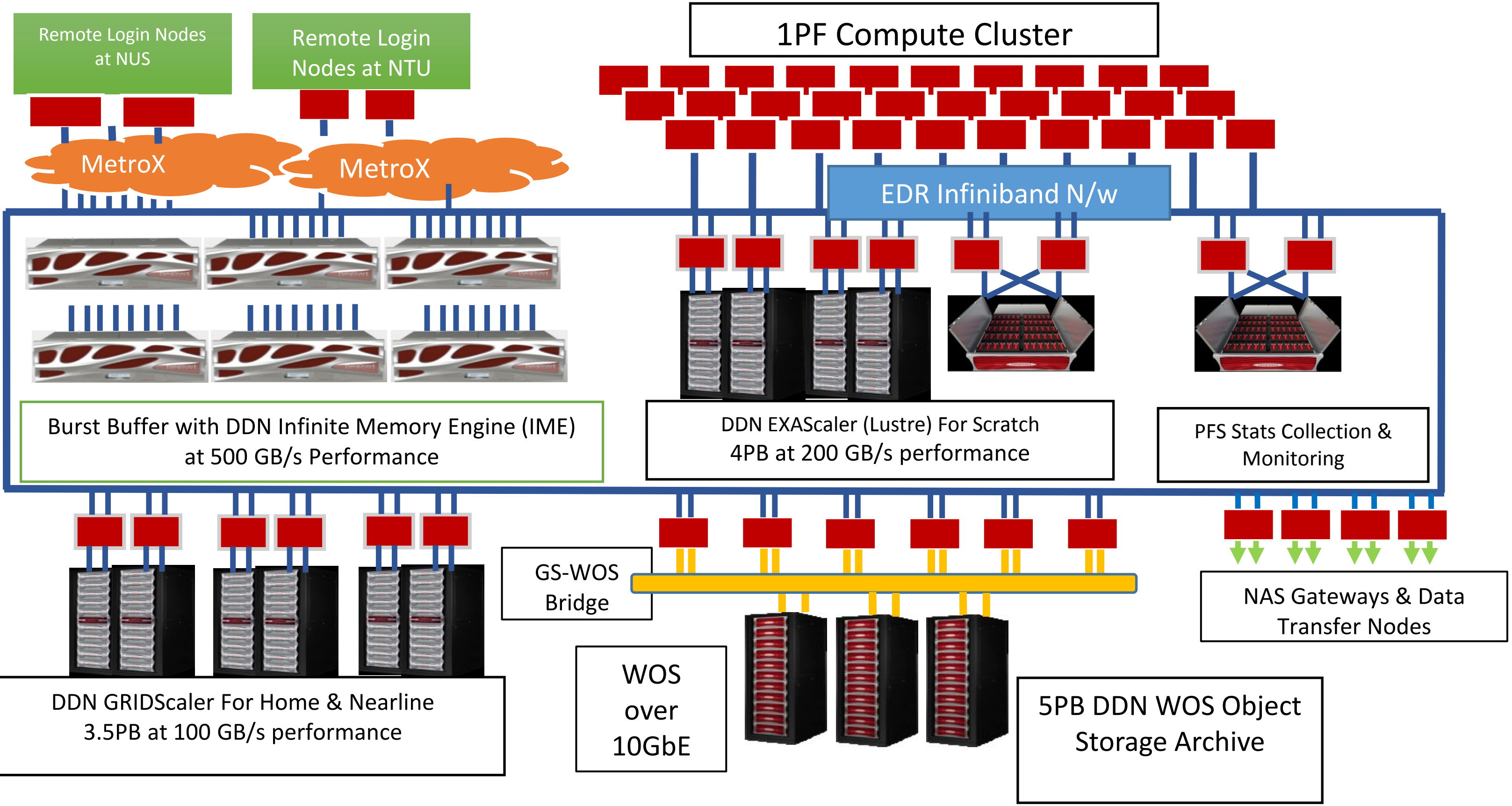
lustre?

Sporadic IO traffic
leads to difficult routing



Courtesy Philip Brighten, U. Illinois

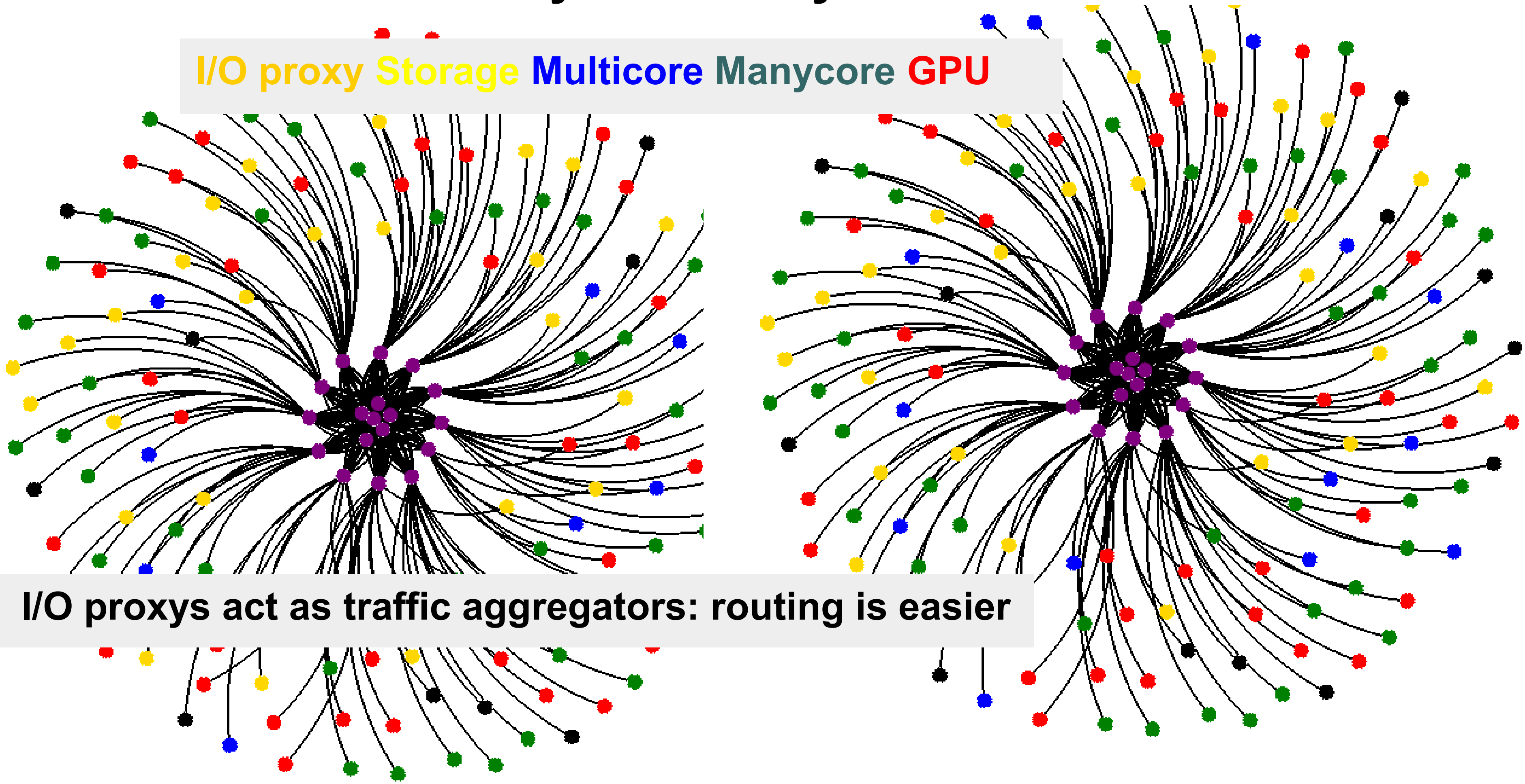
State of the Art Storage Architecture



Courtesy Philip Brighten, U. Illinois

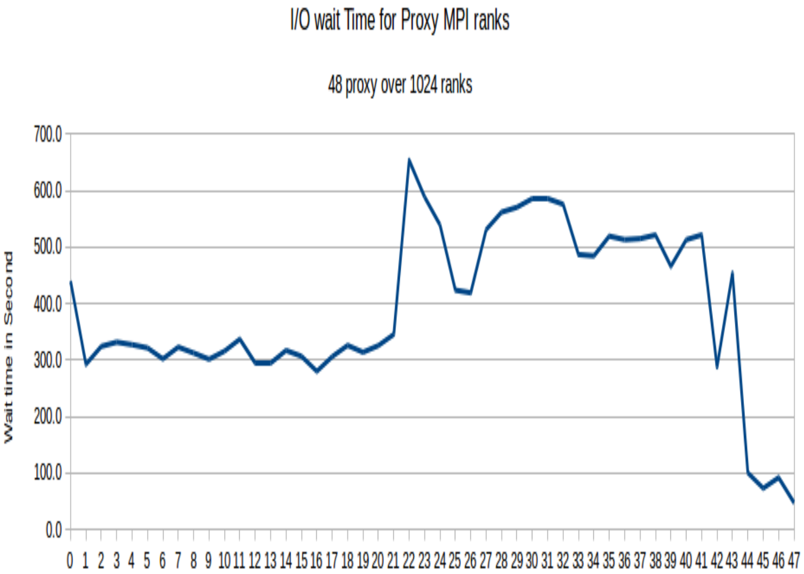
Exascale as a System of Systems

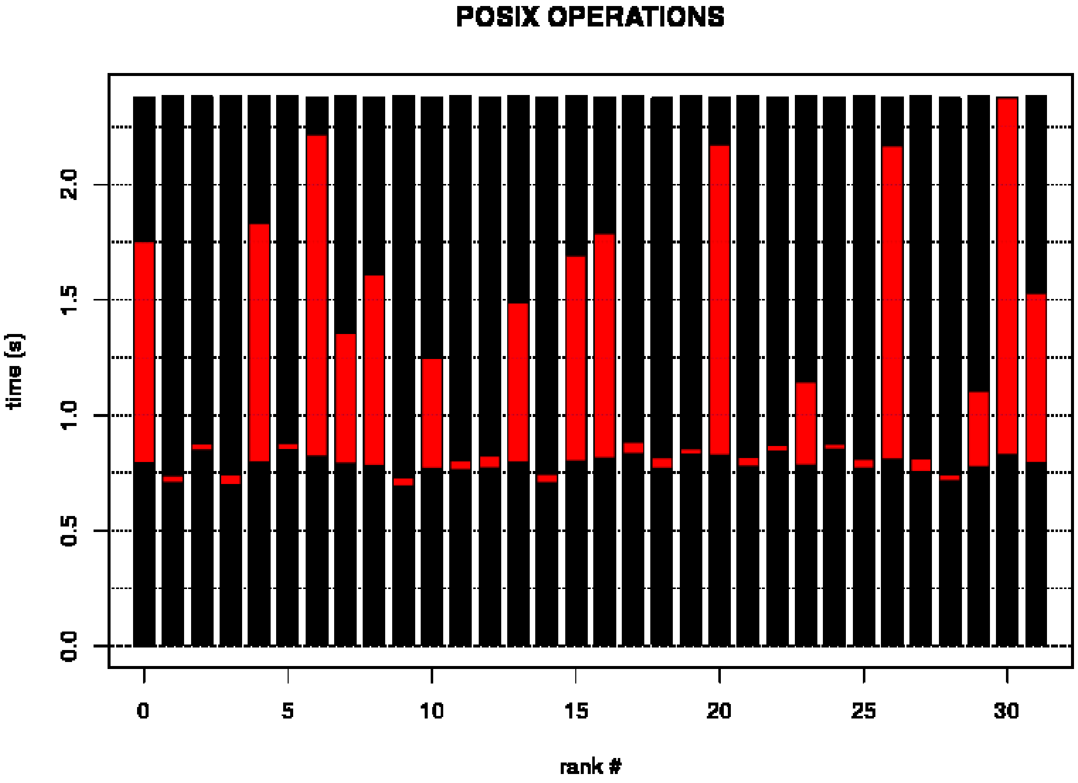
I/O proxy Storage Multicore Manycore GPU



I/O proxys act as traffic aggregators: routing is easier

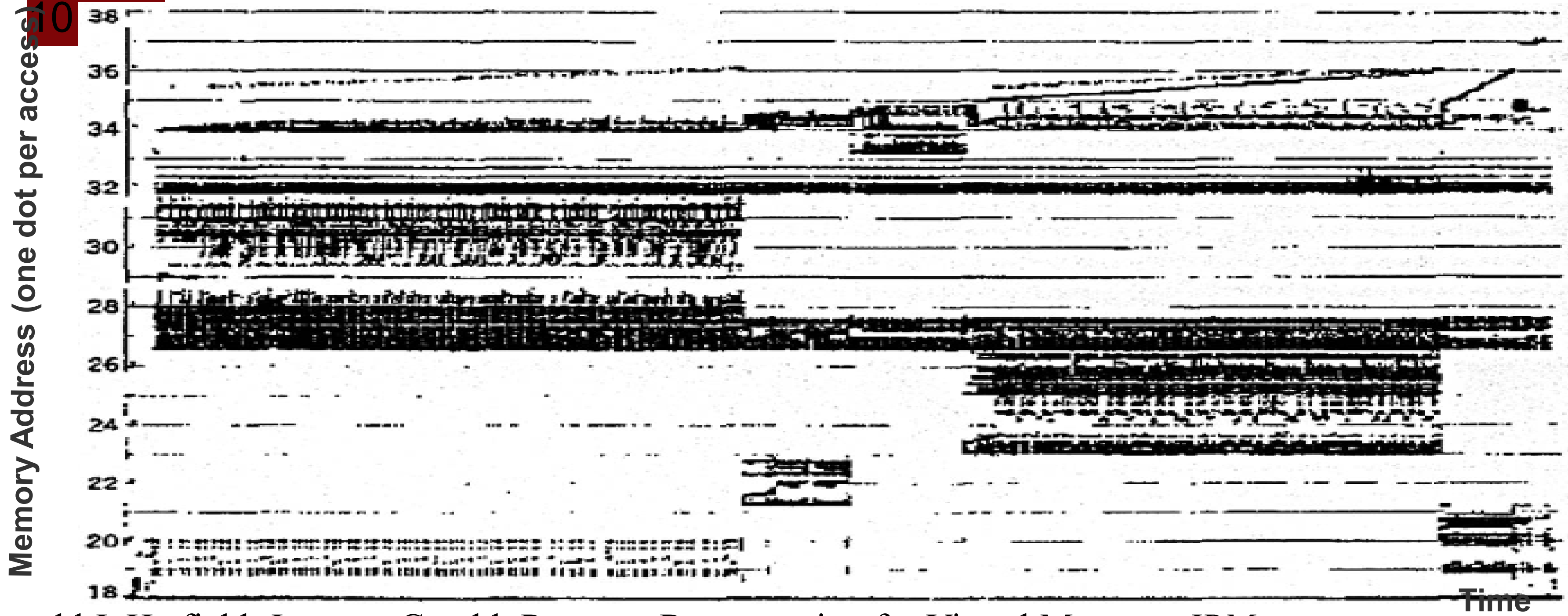
From Monitoring to Orchestration





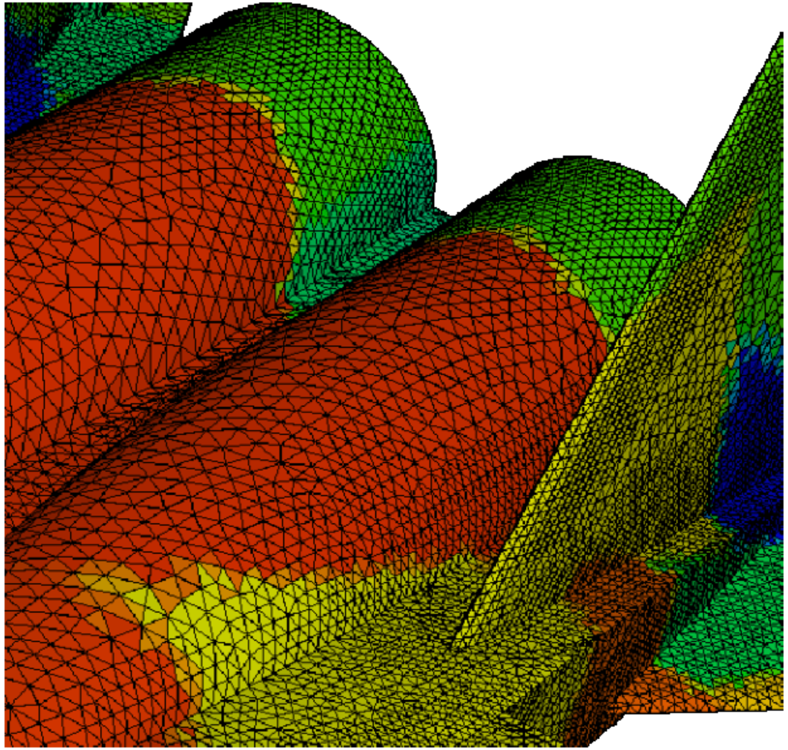
DDN DIO-pro 2016

Temporal and Spatial Patterns...

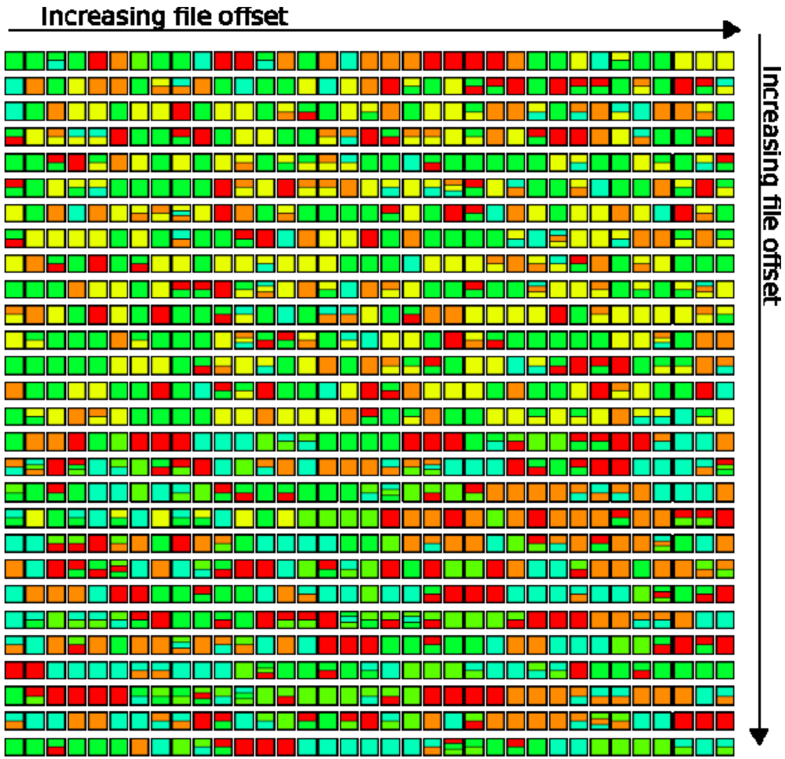


Donald J. Hatfield, Jeanette Gerald: Program Restructuring for Virtual Memory. *IBM Systems Journal,* 10 (3): 168-192 (1971)

Temporal and Spatial Patterns... are here to stay



(a) Decomposed mesh



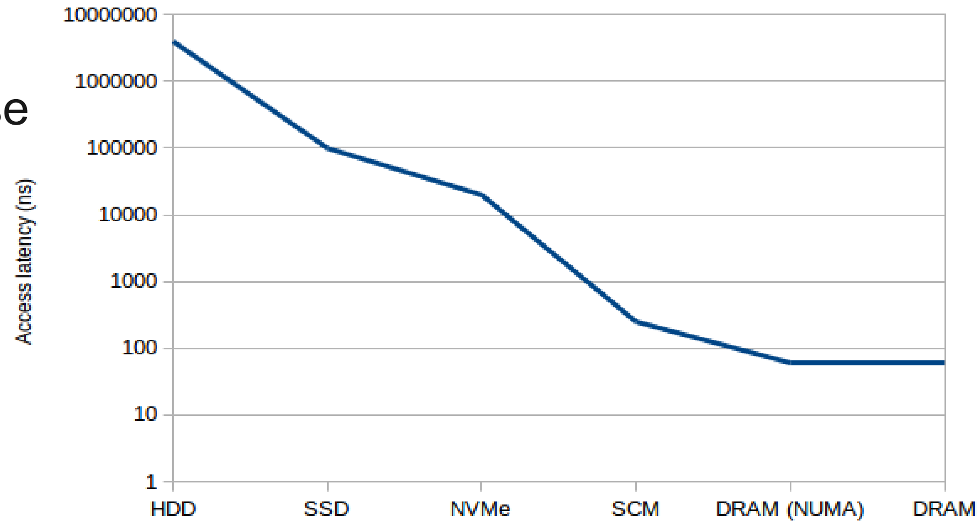
(b) File mapping

Source: Storage Models: Past, Present, and Future. Dres Kimpe et Robert Ross, Argonne National Laboratory

Conclusion: Storage Evolution

Storage getting closer to the CPU

- Mechanically same needs will arise
  - Tools convergence



- **Access latency put pressure on the software design**

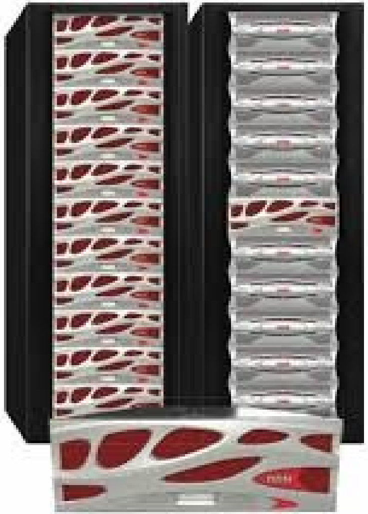
  → window of opportunity to drastic redesign

Early IME I/O Accelerator feed-back

- Harnessing distributed HW resources
  - → From Fault tolerance to QoS
- Hierarchical storage
  - → Narrowing the gap between Storage and Process

→ **Inter-Operable**

- → Software only solution are versatile
- → System wide profiling
- → Data policies
- → Orchestration by job scheduler



System of Systems: Think Out of the Box!

Gracias

ありがとう

Merci !

Thank you !

谢谢

Grazie

спасибо