Future HPC Systems and Some Implications for Storage Software

Rob Ross
Mathematics and Computer Science Division
Argonne National Laboratory
rross@mcs.anl.gov

Hardware

Exascale Systems: Potential Architecture

Systems	2009	2018 (ish)	Difference
System Peak	2 Pflop/sec	1 Eflop/sec	O(1000)
Power	6 Mwatt	20 Mwatt	
System Memory	0.3 Pbytes	32-64 Pbytes	O(100)
Node Compute	125 Gflop/sec	1-15 Tflop/sec	O(10-100)
Node Memory BW	25 Gbytes/sec	2-4 Tbytes/sec	O(100)
Node Concurrency	12	O(1-10K)	O(100-1000)
Total Node Interconnect BW	3.5 Gbytes/sec	200-400 Gbytes/sec	O(100)
System Size (Nodes)	18,700	O(100,000-1M)	O(10-100)
Total Concurrency	225,000	O(1 billion)	O(10,000)
Storage	**15 Pbytes**	**500-1000 Pbytes**	**O(10-100)**
I/O	**0.2 Tbytes/sec**	**60 Tbytes/sec**	**O(100)**
MTTI	Days	O(1 day)	

From J. Dongarra, "Impact of Architecture and Technology for Extreme Scale on Software and Algorithm Design," Cross-cutting Technologies for Computing at the Exascale, February 2-5, 2010.

There will be disks.

- Storage Hierarchy is DRAM, SCM, FLASH, Disk, Tape
- Cannot manufacture enough bits via wafers vs. disks
  - SSD 10x per-bit cost, and the gap isn't closing
  - Cost of semiconductor FAB is >> cost of disk manufacturing facility
  - World-wide manufacturing capacity of semi-conductor bits is perhaps 1% the capacity of making magnetic bits
    - 500 Million disks/year (2012 est) avg 1TB => 500 Exabytes (all manufacturers)
    - 30,000 wafers/month (micron), 4TB/wafer (TLC) => 1.4 Exabytes (micron)
- … and tape doesn't go away, either
  - Still half the per-bit cost, and much less lifetime cost
  - Tape is just different
    - no power at rest
    - physical mobility
    - higher per-device bandwidth (1.5x to 2x)

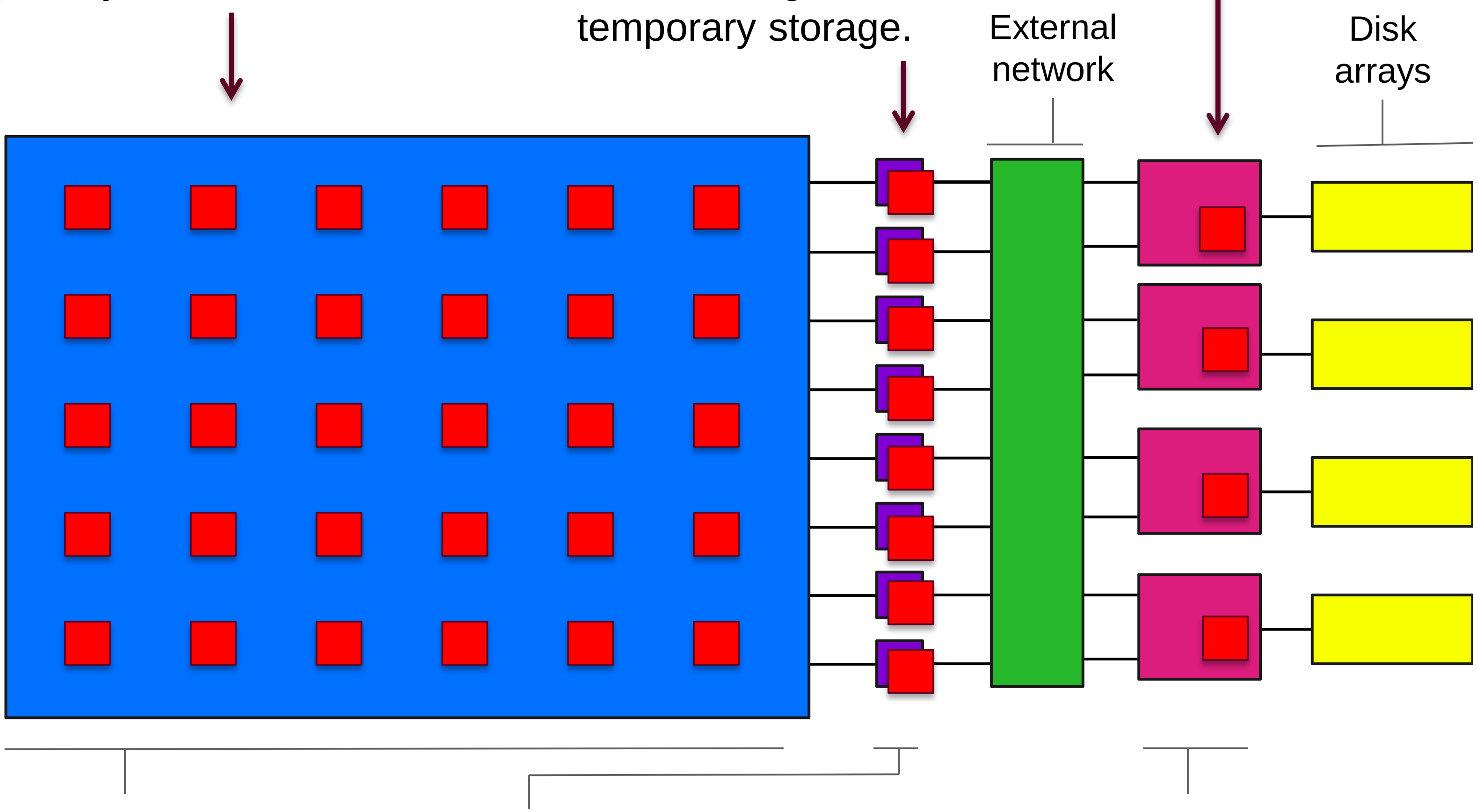
Thanks to Brent Welch (Google).

System Architecture and Nonvolatile Memory

NVM in compute nodes lets you add noise into your system network.

NVM in I/O nodes provides a fast staging area and region for temporary storage.

NVM in storage nodes serves as a PFS accelerator.



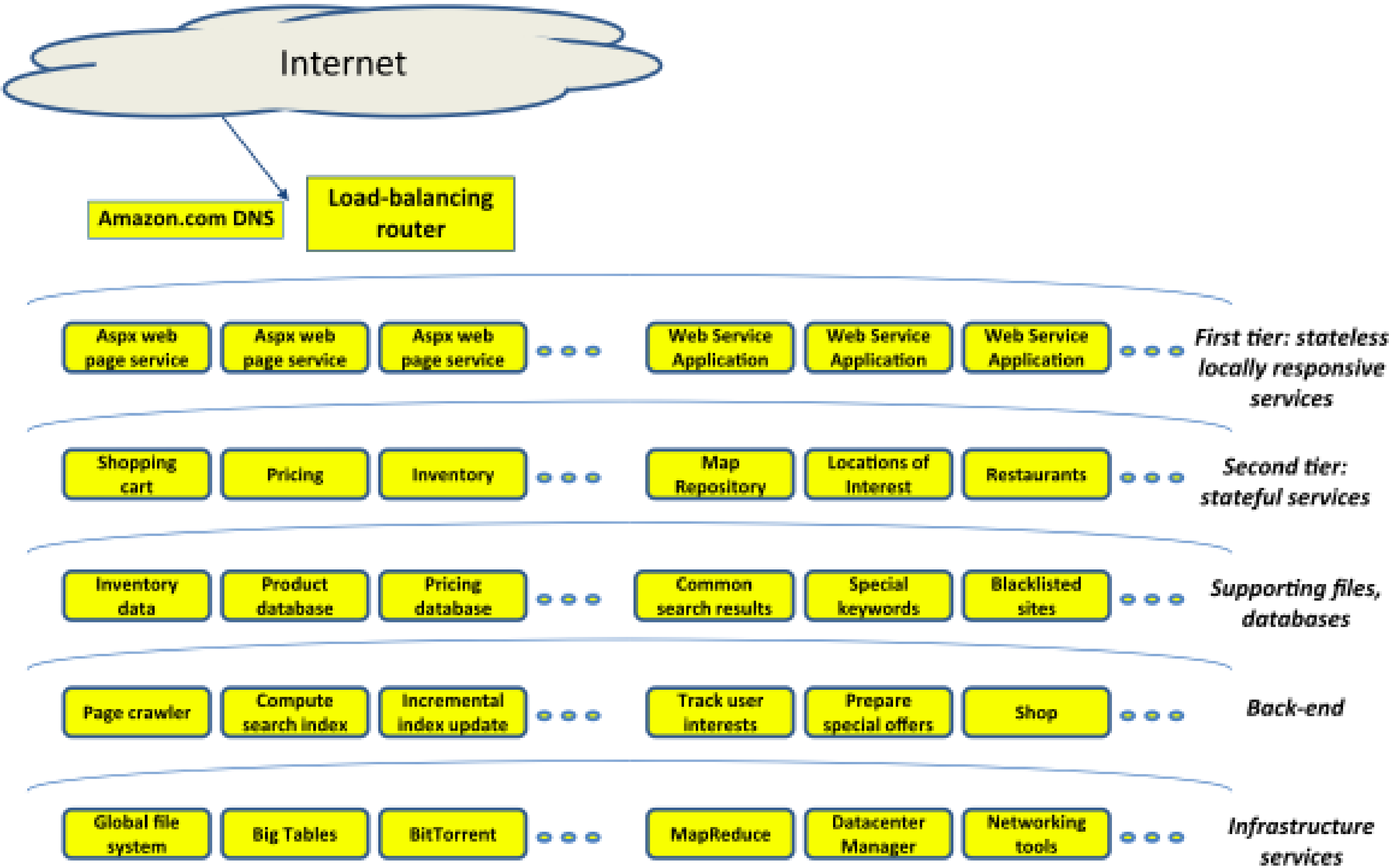
Compute nodes run application processes.

I/O forwarding nodes (or I/O gateways) shuffle data between compute nodes and external resources, including storage.

Storage nodes run the parallel file system.

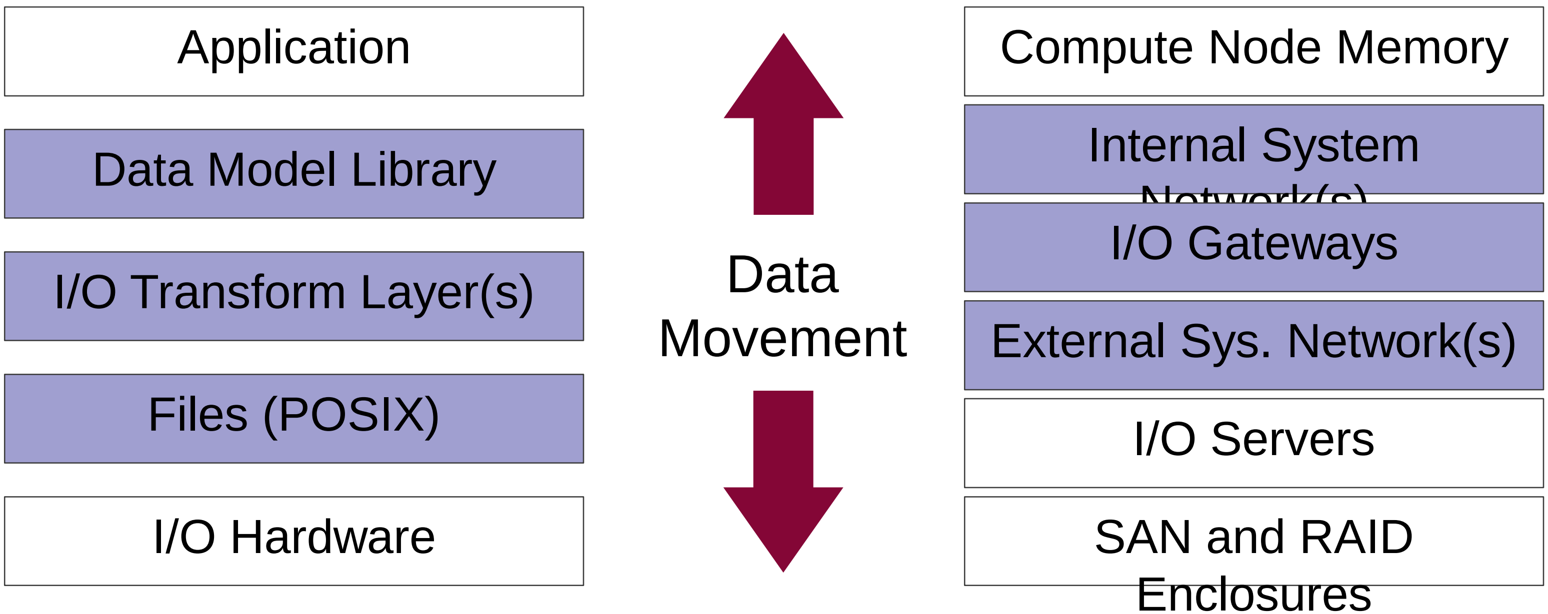
Software and Integration

Another Community: Integration, Scale, and State



From K. Birman, Guide to Reliable Distributed Systems, 2012.

The PFS is integrating, and must integrate with, many other components.



Logical (data model) view of data access.

Physical (hardware) view of data access.

- Not shown: other system services that the PFS must interact with (e.g., resource management and reliability services)

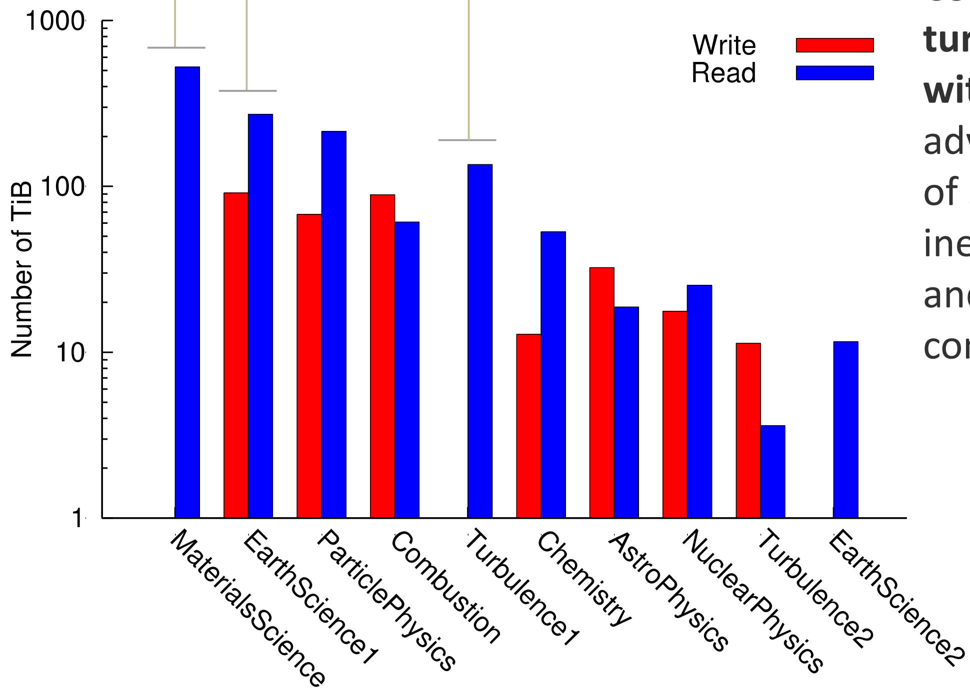
Big Data

Big Data on Leadership Platforms: It's Happening

Matching large scale simulations of dense suspensions with empirical measurements to better understand properties of complex materials such as concrete.

Processing large-scale seismographic datasets to develop a 3D velocity model used in developing earthquake hazard maps.

Comparing simulations of turbulent mixing of fluids with experimental data to advance our understanding of supernovae explosions, inertial confinement fusion, and supersonic combustion.



Top 10 data producer/consumers instrumented with Darshan over the month of July, 2011 on Intrepid BG/P system at Argonne. Surprisingly, three of the top producer/consumers almost exclusively read existing data.