Argonne NATIONAL LABORATORY

Overcoming Roadblocks to Exascale Storage

Rob Ross, Pete Beckman, Phil Carns, Jason Cope, Kevin Harms, Kamil Iskra, Dries Kimpe, Rob Latham, Rusty Lusk, Tom Peterka, Katherine Riley, Seung Woo Son, Rajeev Thakur, Venkat Vishwanath, and Justin Wozniak

Mathematics and Computer Science Division

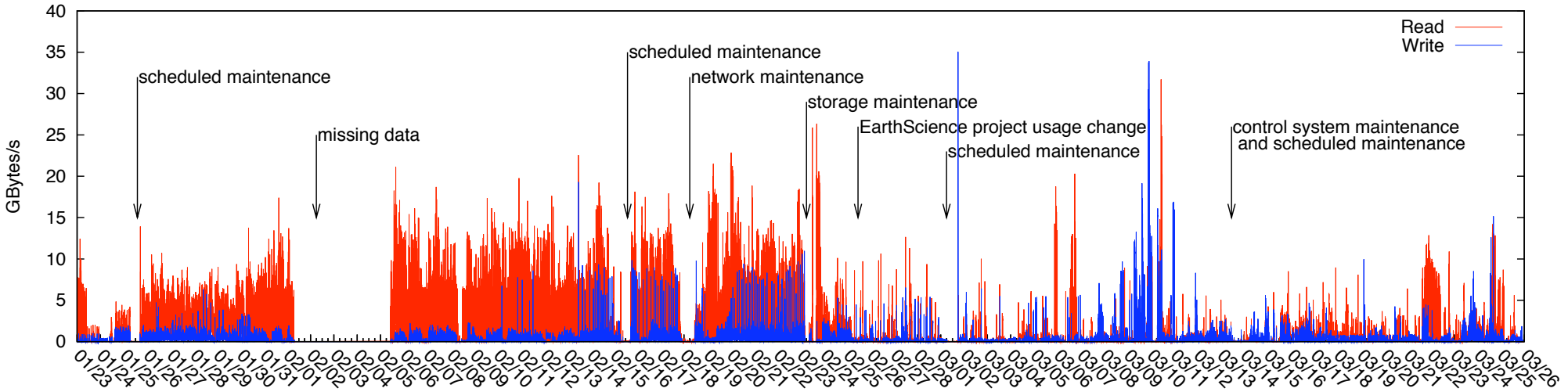
Argonne National Laboratory

rross@mcs.anl.gov

U.S. DEPARTMENT OF ENERGY

The Obvious Challenges

- **Bulk data movement.** Gotta reduce synchronization in the I/O path, help users perform in situ analysis to reduce I/O overall, and use in-system storage resources to best drive I/O to external storage.
- **Reliability.** Must hide failures when possible and degrade gracefully when not. See Internet services storage solutions for ideas.

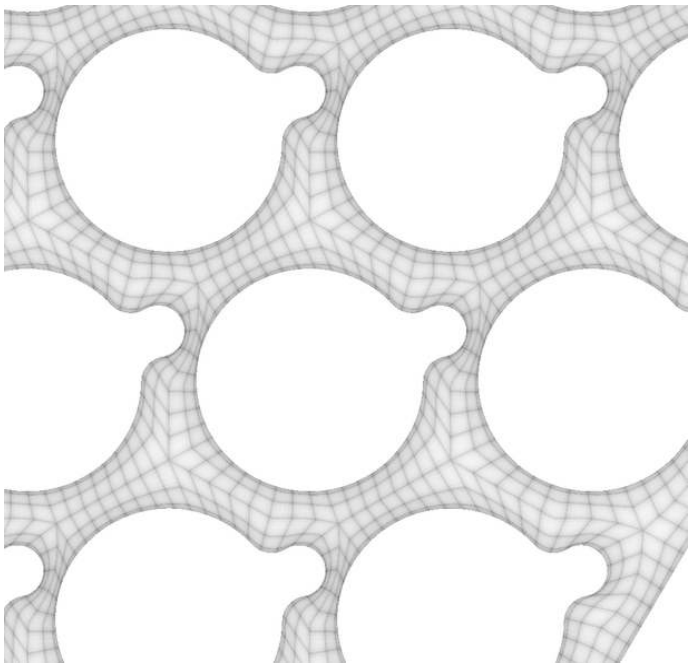


Aggregate I/O throughput on BG/P storage servers at one minute intervals. Lots of bandwidth is left on the floor.

P. Carns et al. Understanding and improving computational science storage access through continuous characterization. In Proceedings of 27th IEEE Conference on Mass Storage Systems and Technologies (MSST 2011), May 2011.

Challenge: Inertia (or Data Models and Semantics)

- The connection (interface) between applications and storage should reflect the models used in science codes
  - (e.g., structured and unstructured meshes, particles, varying levels of fidelity)
- POSIX sucks as an I/O model for computational science
  - Neither convenient nor easy to reach high performance for real workloads
  - Serves as a continual distraction from more productive development
- Need to present an alternative storage model
  - Object storage (i.e., OSD) is a nice start
  - Need more understanding of name spaces
  - Must consider analysis use cases as well
- Temptation to "support POSIX" will remain and taint most efforts.

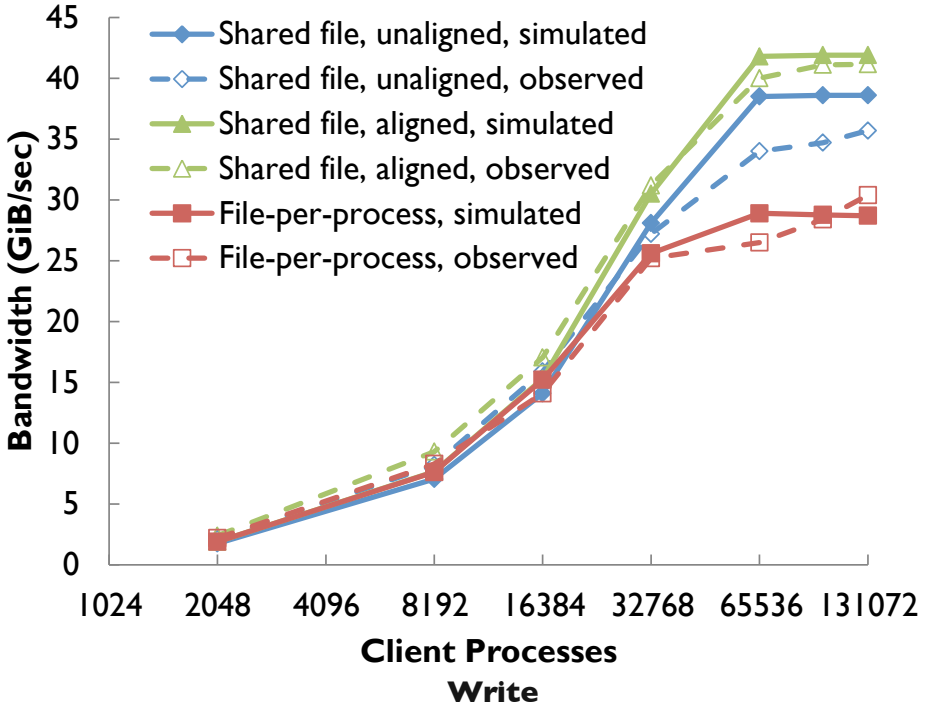


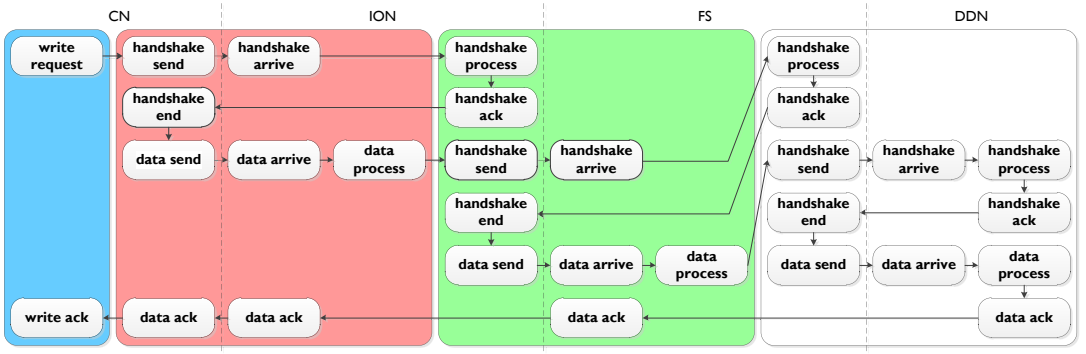
Cross-section of spectral element mesh used in large eddy simulation of 217-pin reactor subassembly.

Image from P. Fischer (ANL).

Challenge: Understanding

- **Applications.** Need more information on current and future application I/O needs. Information should guide design. See Darshan, IPM-IO, ScalaTrace.
- **Storage Designs.** Exascale storage will be complex mix of heterogeneous components. Need tools to help us explore the HW/SW design space quickly and efficiently before we start building, while we continue development. See CODES, HECIOS, IMPIOUS, PFSsim.





Current CODES model includes simulation of I/O protocols on all major BG/P components. First goal is to validate with BG/P data.

N. Liu et al. Modeling a Leadership-scale Storage System. In Proceedings of the 9th International Conference on Parallel Processing and Applied Mathematics (PPAM 2011). September 2011 (paper to appear).

Acknowledgment

This work was supported by the Office of Advanced Scientific Computing Research, Office of Science, U.S. Dept. of Energy, under Contract DE-AC02-06CH11357.

Our work is only possible with the help of our many collaborators, including:

- Alok Choudhary, Kui Gao, Wei-keng Liao, Arifa Nisar (NWU)
- Kwan-Liu Ma, Hongfeng Yu (UC Davis)
- Lee Ward (SNL)
- Gary Grider, James Nunez (LANL)
- Steve Poole, Terry Jones (ORNL)
- Yutaka Ishikawa, Kazuki Ohta (University of Tokyo)
- Javier Blas, Florin Isaila (University Carlos III of Madrid)