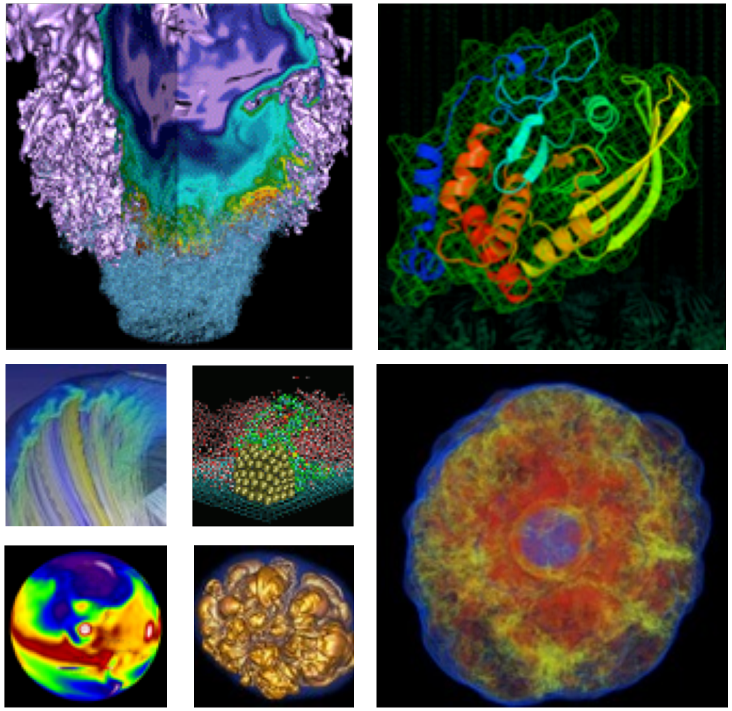
A Year in the Life of a Parallel File System



Glenn K. Lockwood, Shane Snyder, Teng Wang, Suren Byna, Philip Carns, Nicholas J. Wright

November 15, 2018

Why was my job's I/O slow?

Socrates (left) and Plato (right) contemplating I/O performance in *The School of Athens* by Raphael. 1511.

Why was my job's I/O slow?

1. You are doing something wrong
2. Another job/system task is competing with you
3. The storage system is degraded

Why was my job's I/O slow?

1. You are doing something wrong
2. Another job/system task is competing with you
3. The storage system is degraded

Most frustrating

Least studied

Our holistic approach to I/O variation

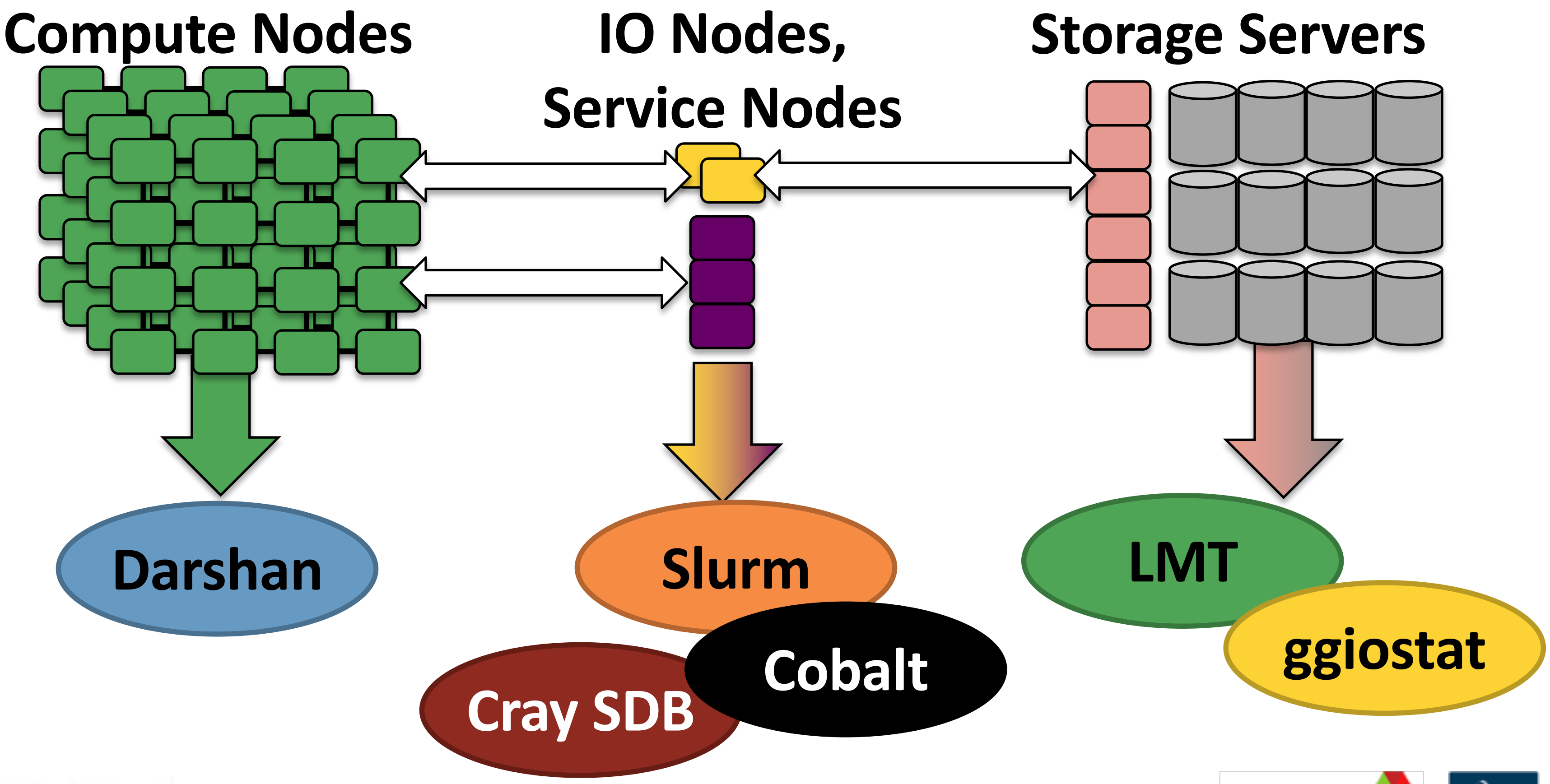
1. Measure performance variation over a year on large-scale production HPC systems
2. Collect telemetry from across the entire system
3. Quantitatively describe why I/O varies so much

1. Observing variation in the wild

App I/O Transfer Size	Shared File	File Per Process
O(1 MiB)	IOR	IOR
O(100 MiB)	VPIC BD-CATS	HACC

- **Probe I/O performance daily**
  - Jobs scaled to achieve >80% peak fs performance
  - 45 – 300 sec per probe
- **Run in diverse production environments**
  - Two DOE HPC facilities (ALCF, NERSC)
  - Three large-scale systems (Mira, Edison, Cori)
  - Two parallel file system implementations (GPFS, Lustre)
  - Five file systems (Mira gpfs1, Edison lustre[1-3], Cori lustre1)

2. Collecting diverse data for holistic analysis



Compute Nodes

IO Nodes, Service Nodes

Storage Servers

Darshan

Slurm

Cobalt

Cray SDB

LMT

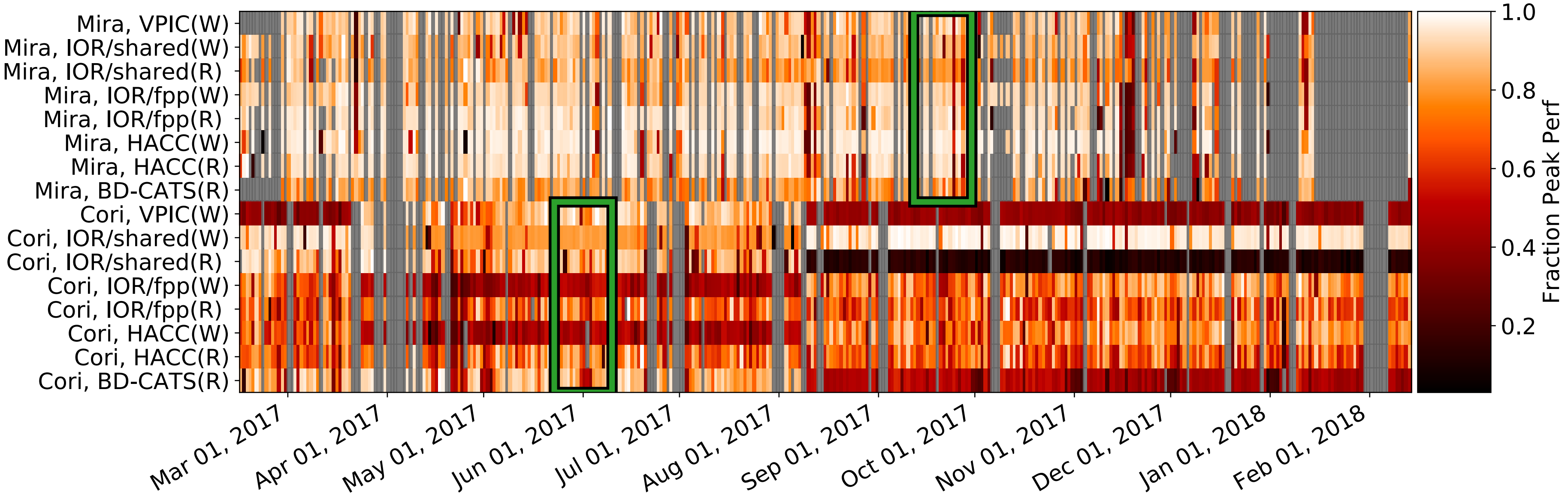
ggiostat

Year-long I/O performance dataset

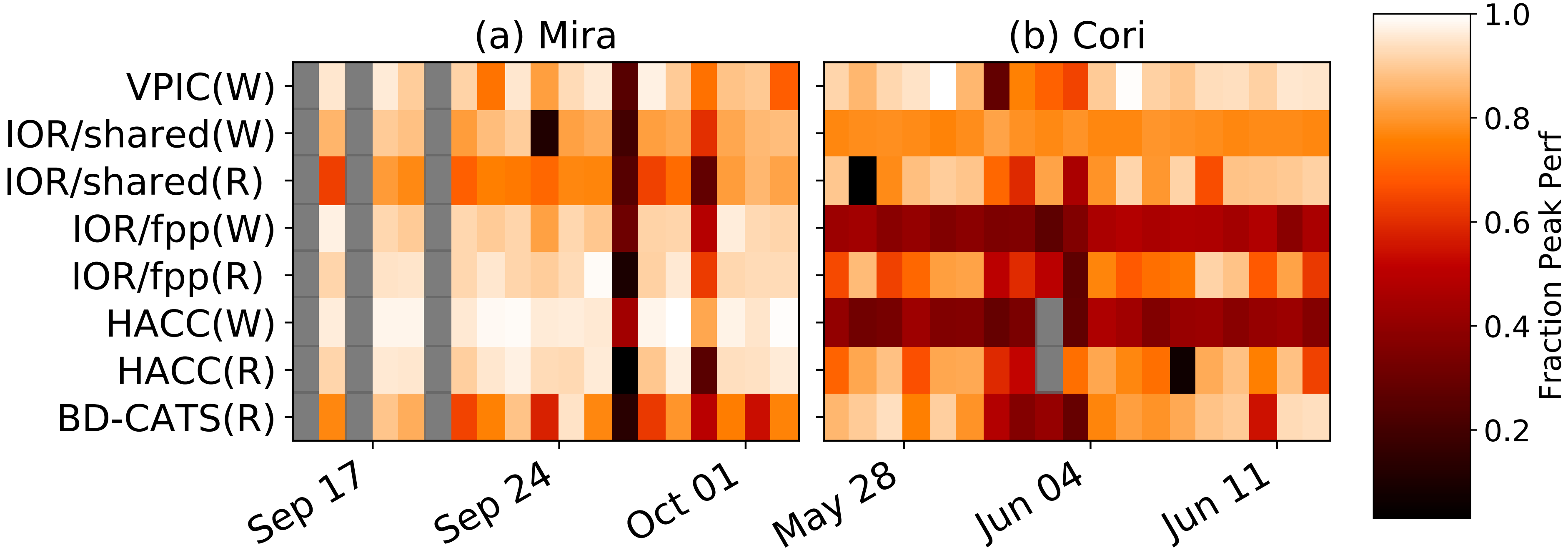
- **366 days of testing**
- **11,986 jobs run**
- **220 metrics measured per job**
  - some derived or degenerate
  - sometimes undefined

…and not very insightful at a glance

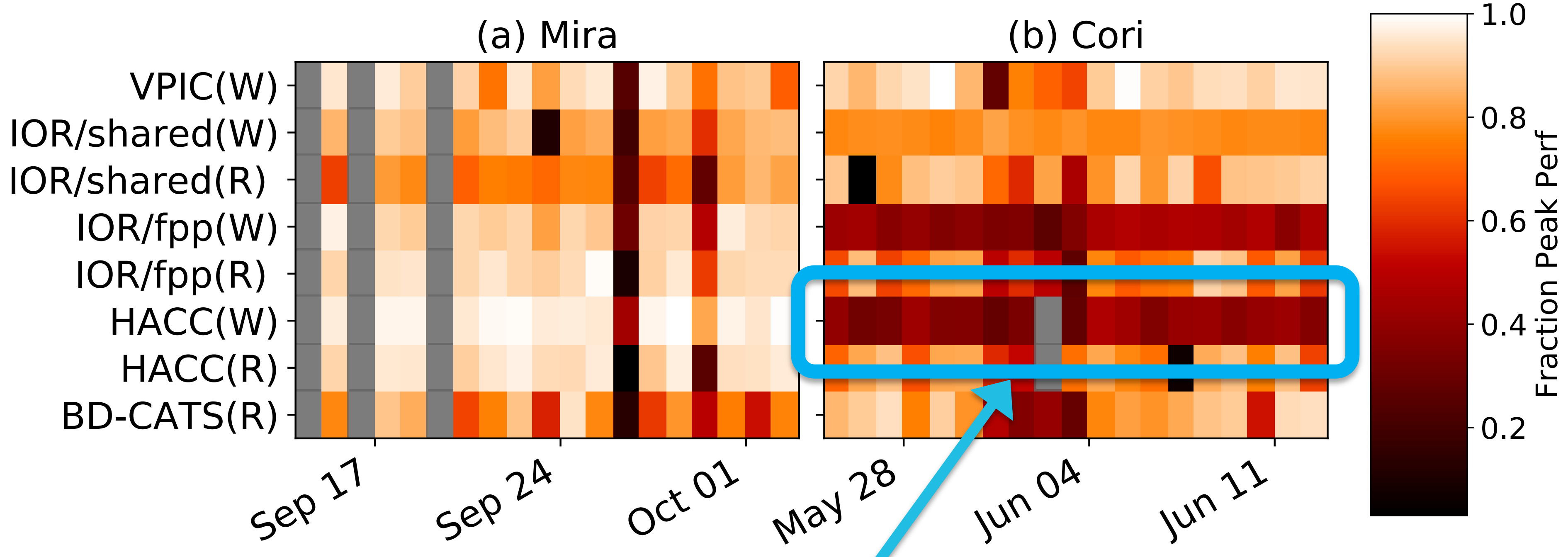
I/O performance variation in production



Two flavors of I/O performance variation

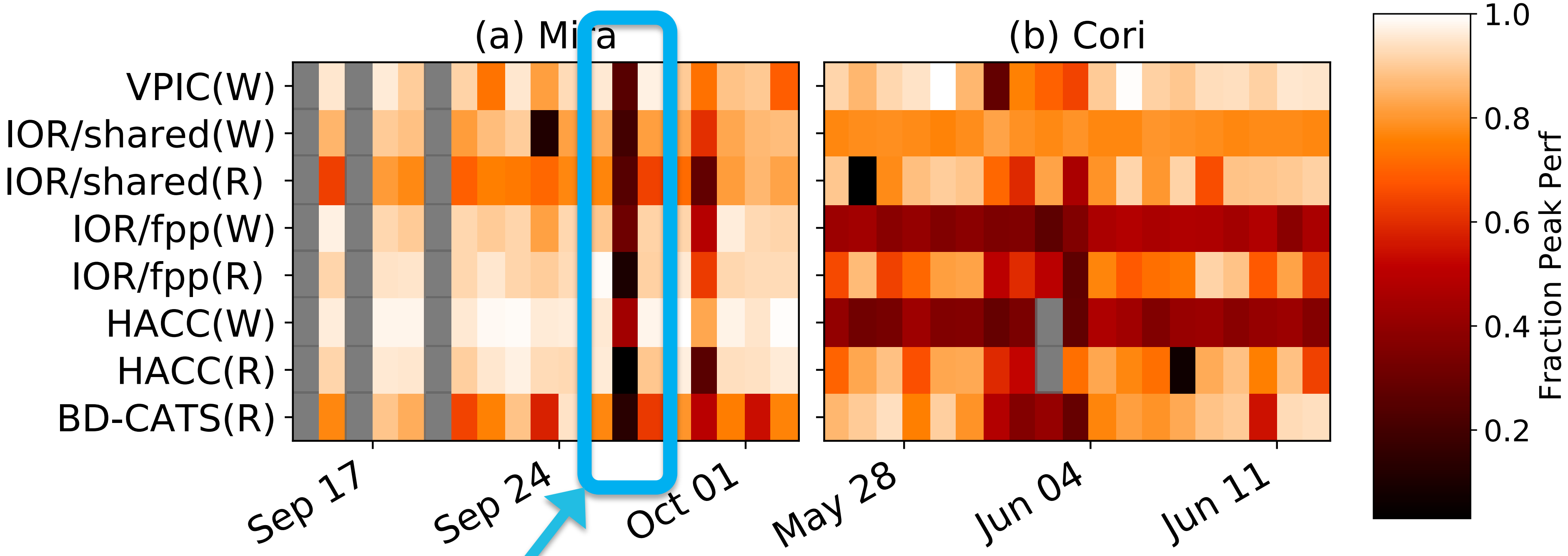


Performance varies over the long term



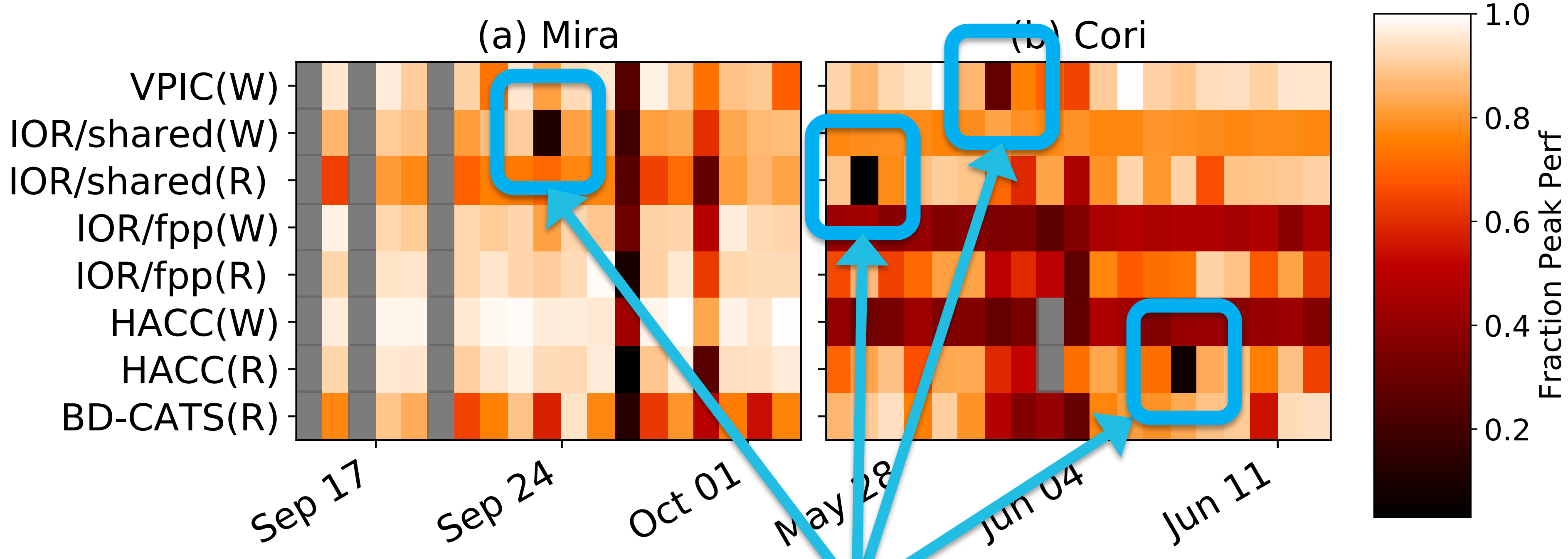
Systematic, long-term problem for one I/O pattern

Performance varies over the short term



Transient bad I/O day for all jobs

Performance also experiences transient losses



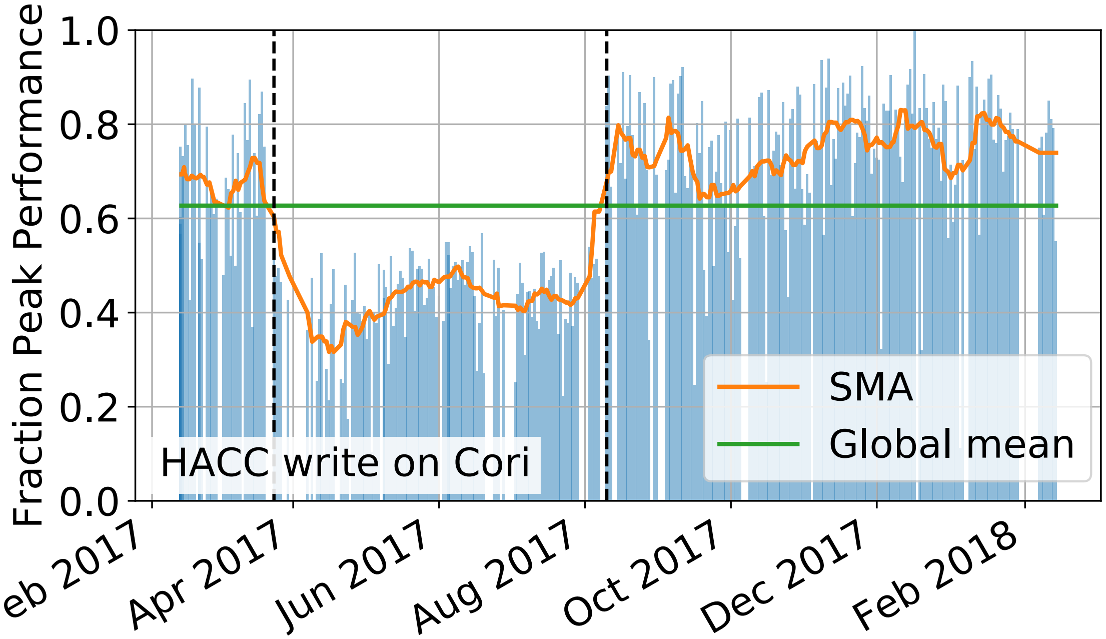
Transient I/O problems

Again: Why was my job's I/O so slow?

- **Could be:**
  - Long-term systematic problems
  - Short-term transient problems
- **The next questions:**
  - What causes long-term, systematic problems?
  - What causes short-term transient problems?
- **Our approach:**
  - Separate problems over these two time scales
  - Independently classify causes of longer-term and shorter-term variation

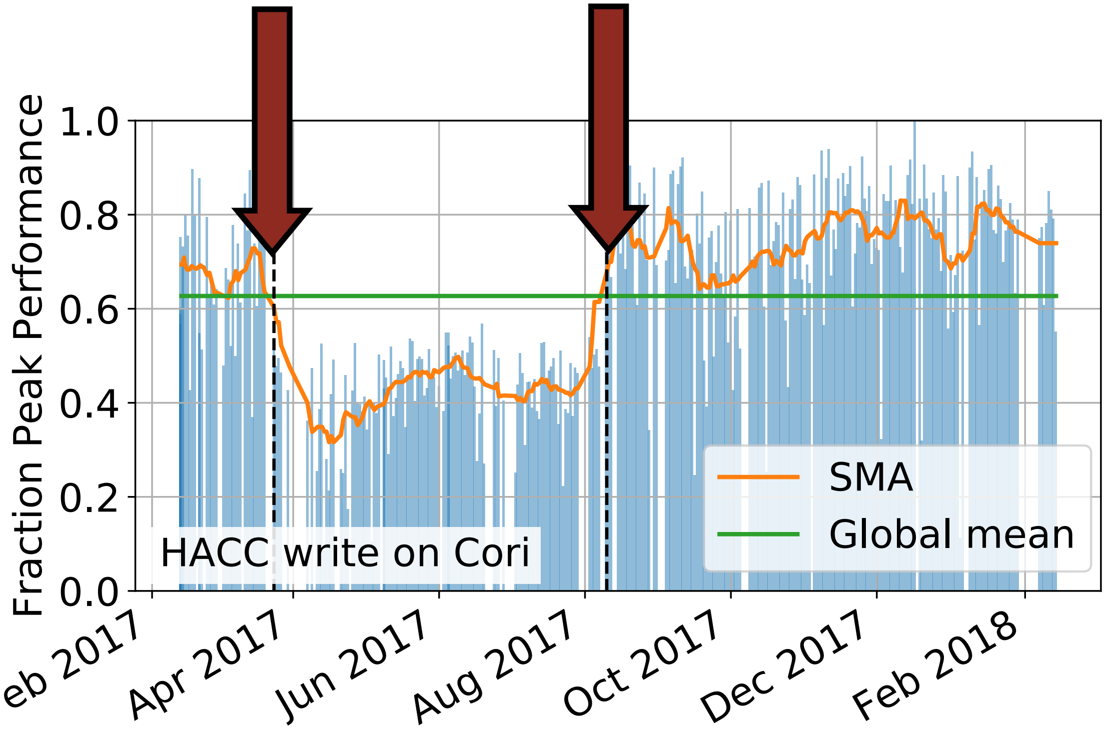
Separating short-term from long-term

- **Goal: Numerically distinguish time-dependent variation**
- Simple moving averages (SMAs) from financial market technical analysis
- Where short-term average performance diverges from overall average



Quantitatively bound long-term problems

- **Goal: Numerically distinguish time-dependent variation**
- Simple moving averages (SMAs) from financial market technical analysis
- Where short-term average performance diverges from overall average
- Example: Bug in a specific file system client version

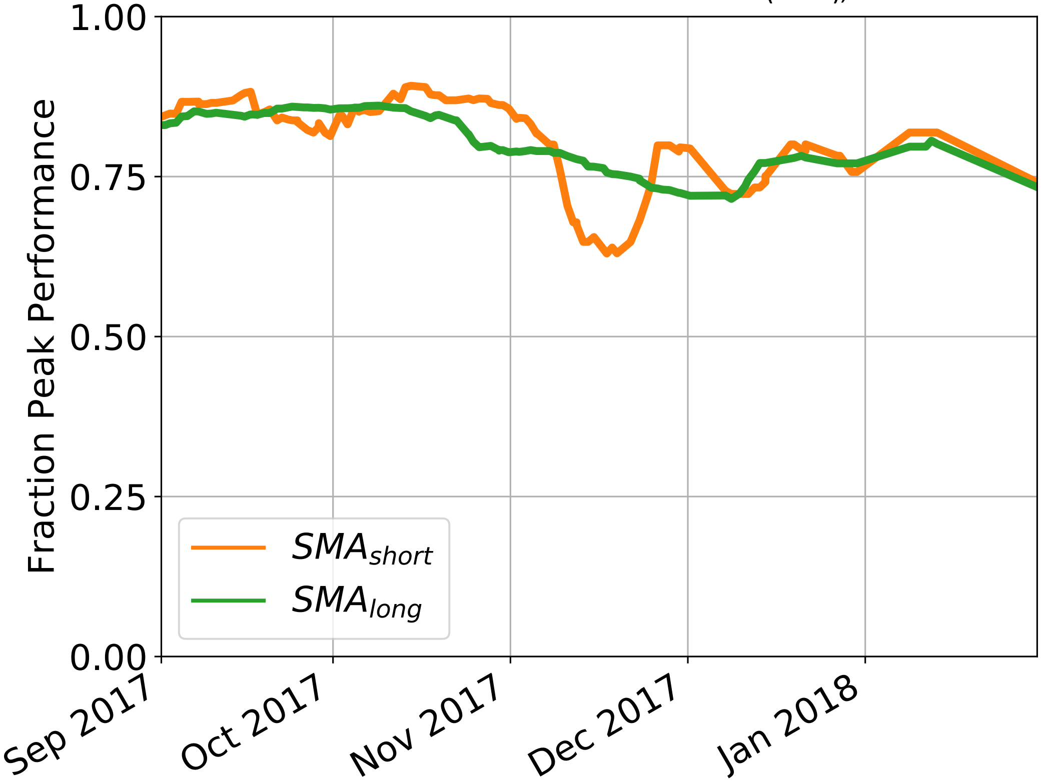


Separating short-term from long-term variation

Goal: Contextualize transient variation happening during long-term variation

- Two SMAs at different time windows (e.g., 14 days and 49 days)

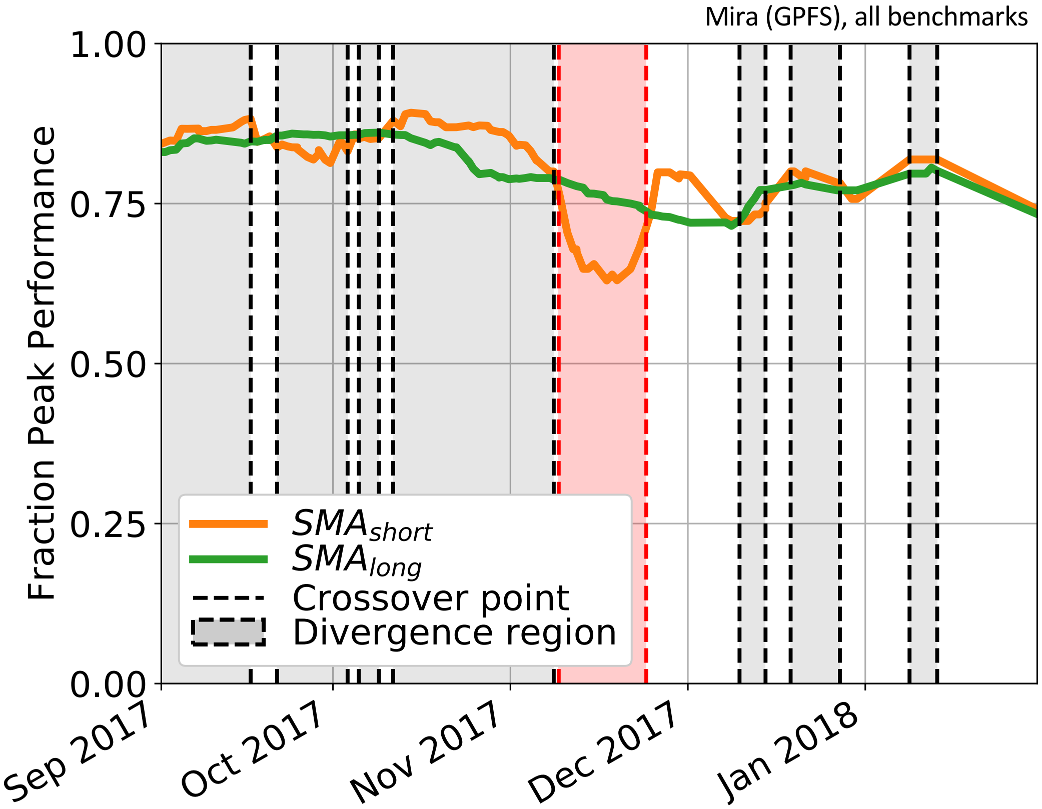
Mira (GPFS), all benchmarks



Separating short-term from long-term variation

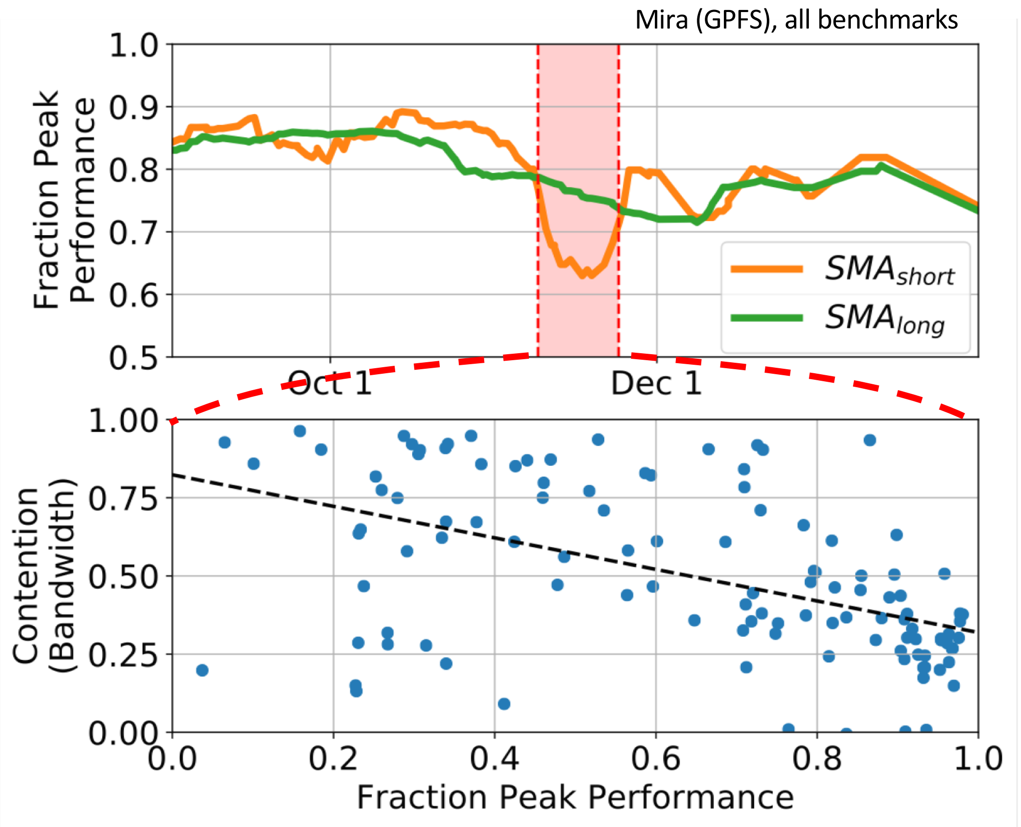
Goal: Contextualize transient variation happening during long-term variation

- Two SMAs at different time windows (e.g., 14 days and 49 days)
- Crossover points indicate short behavior == long behavior
- Divergence regions where short behavior diverges from long behavior

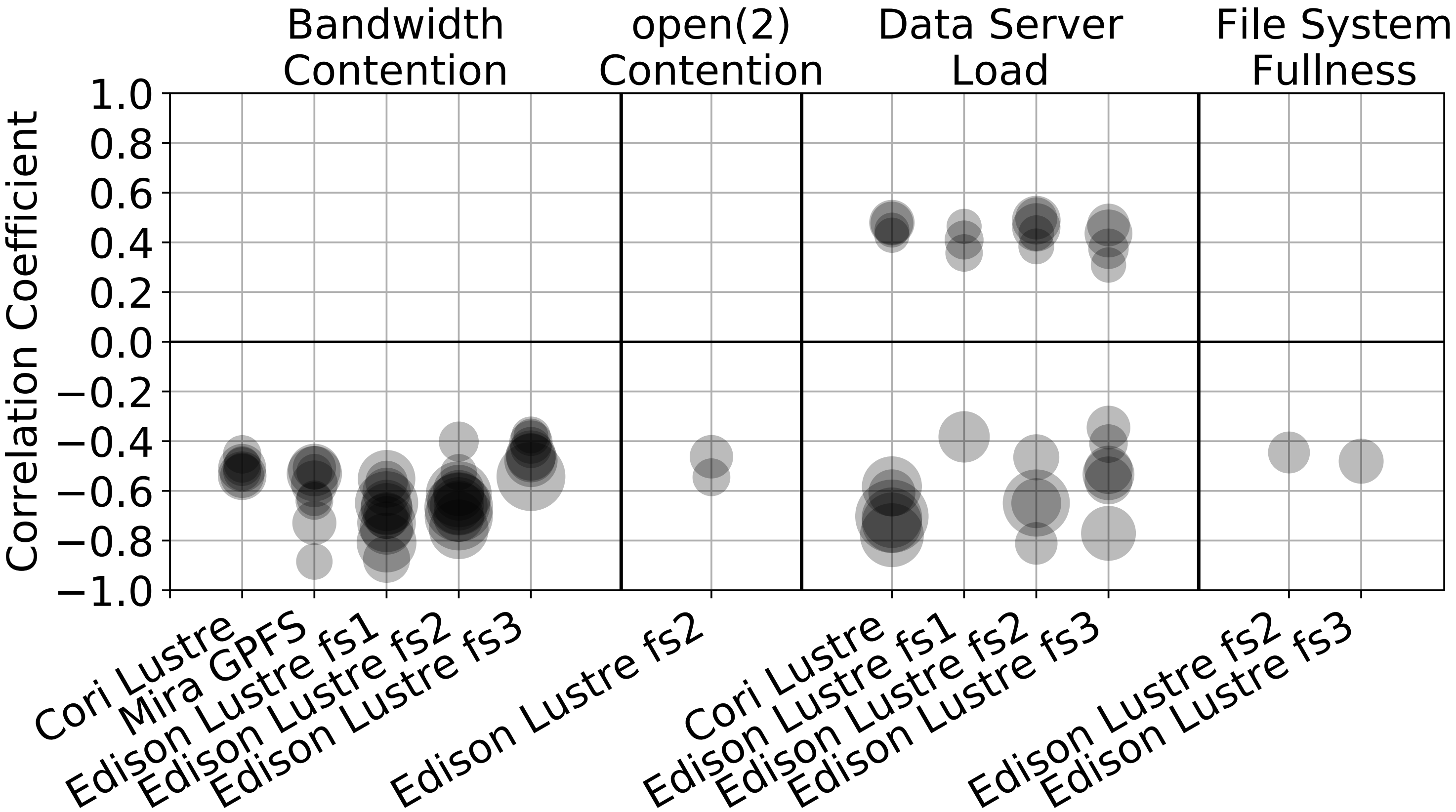


What causes divergence regions?

- Capitalize on widely ranging performance (and all 219 other metrics)
- Correlate performance in this region with other metrics
  - Bandwidth contention
  - IOPS contention
  - Data server CPU load
  - ...



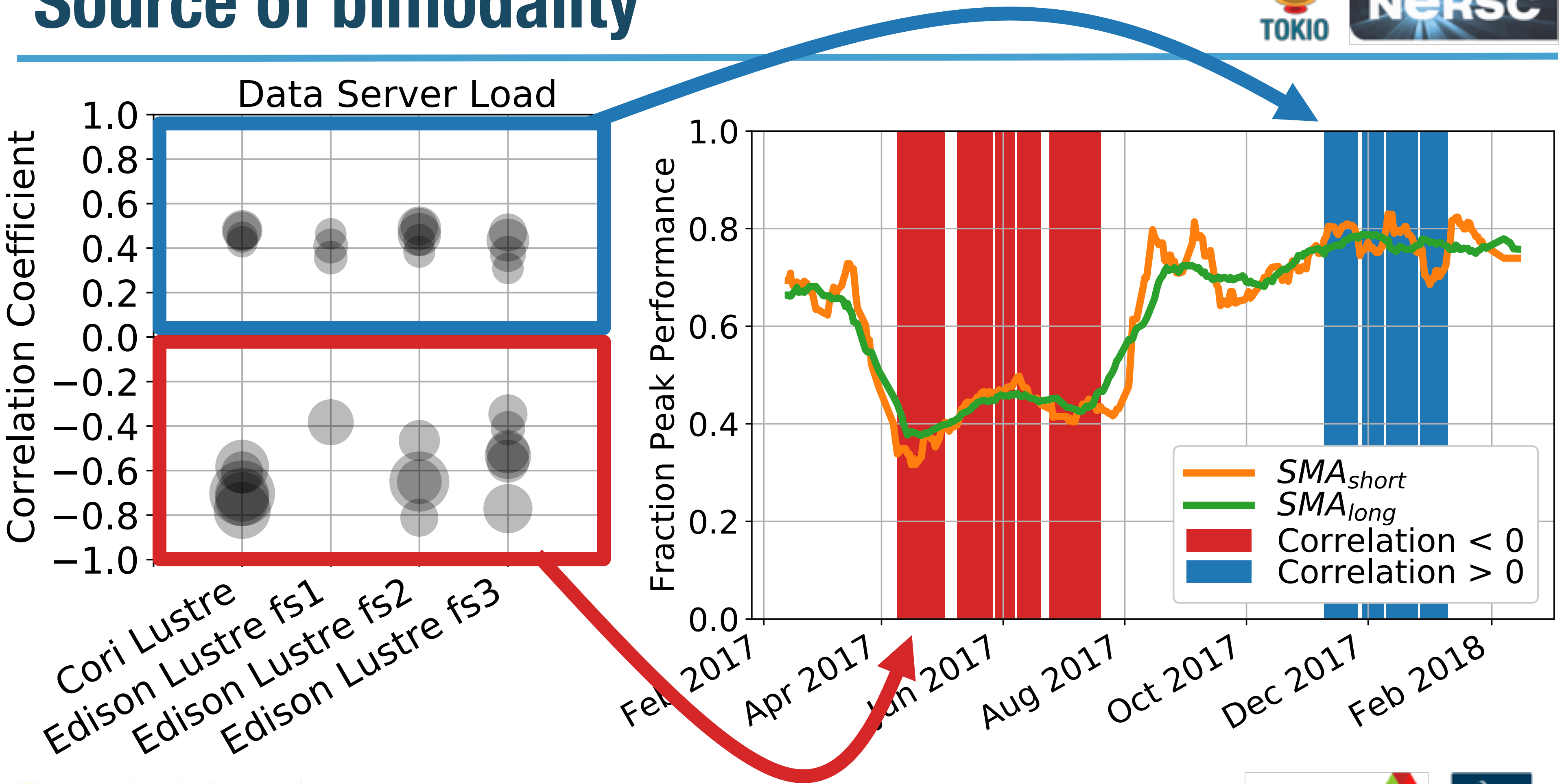
What causes short-term variation over a year?



Each spot is correlation within a single divergence region with p-value < 10^{-5}

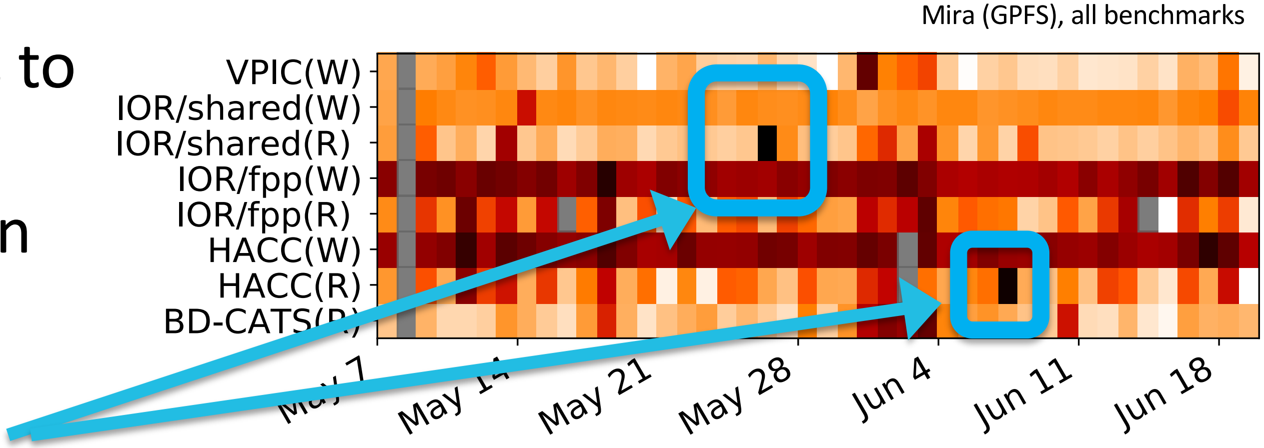
Dot radius ∝ -log(p-value)

Source of bimodality



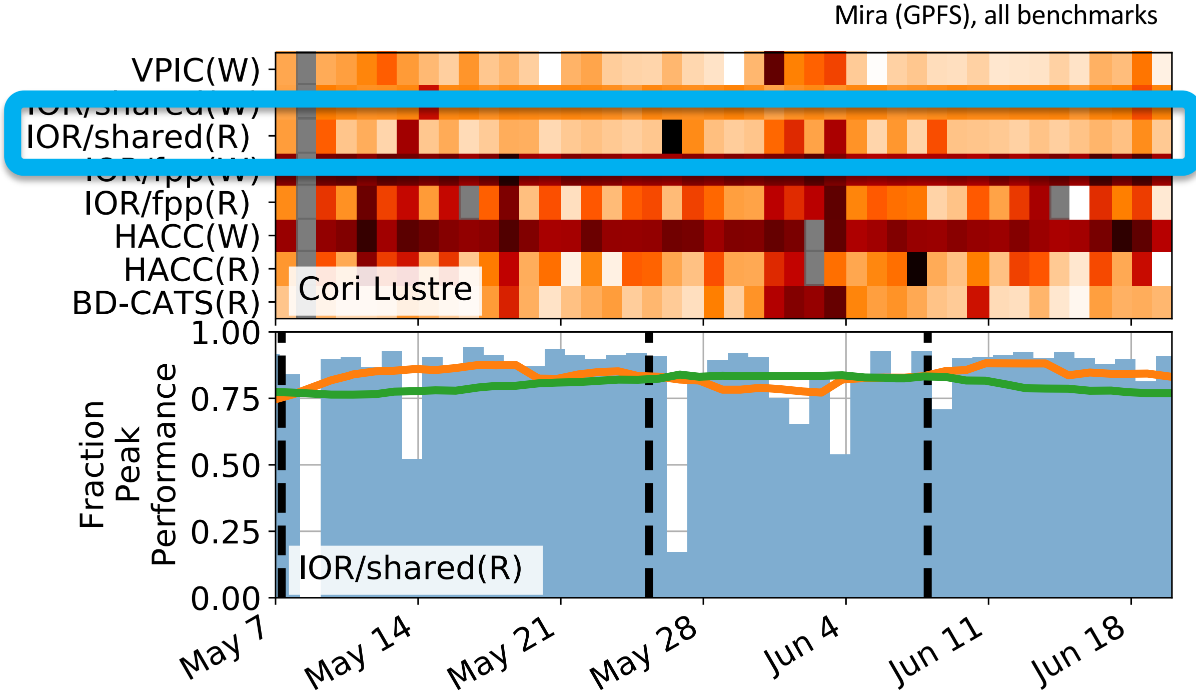
Identifying sources of transient variation

- Partitioning allows us to classify short-term performance variation
- Can't correlate truly transient variation though



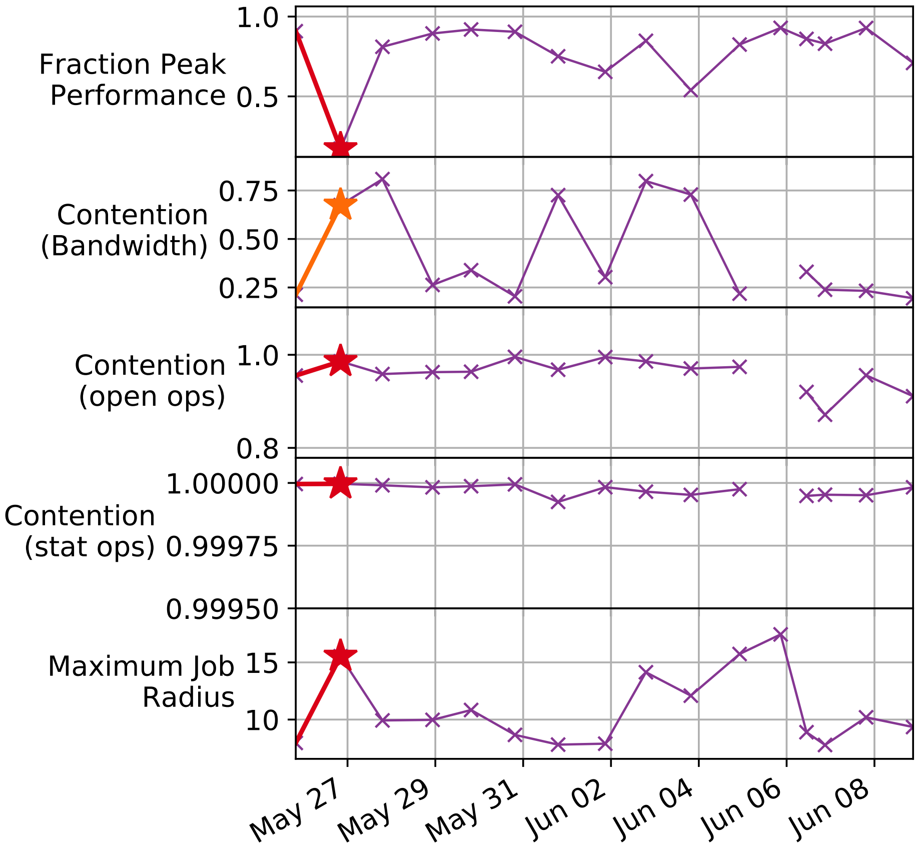
Identifying sources of transient variation

- Confidently classifying transients is statistically impossible
- Classifying in aggregate *is* possible!
- If we observe a possible relationship…
  - One time? Maybe coincidence
  - Many times? Maybe *not* a coincidence



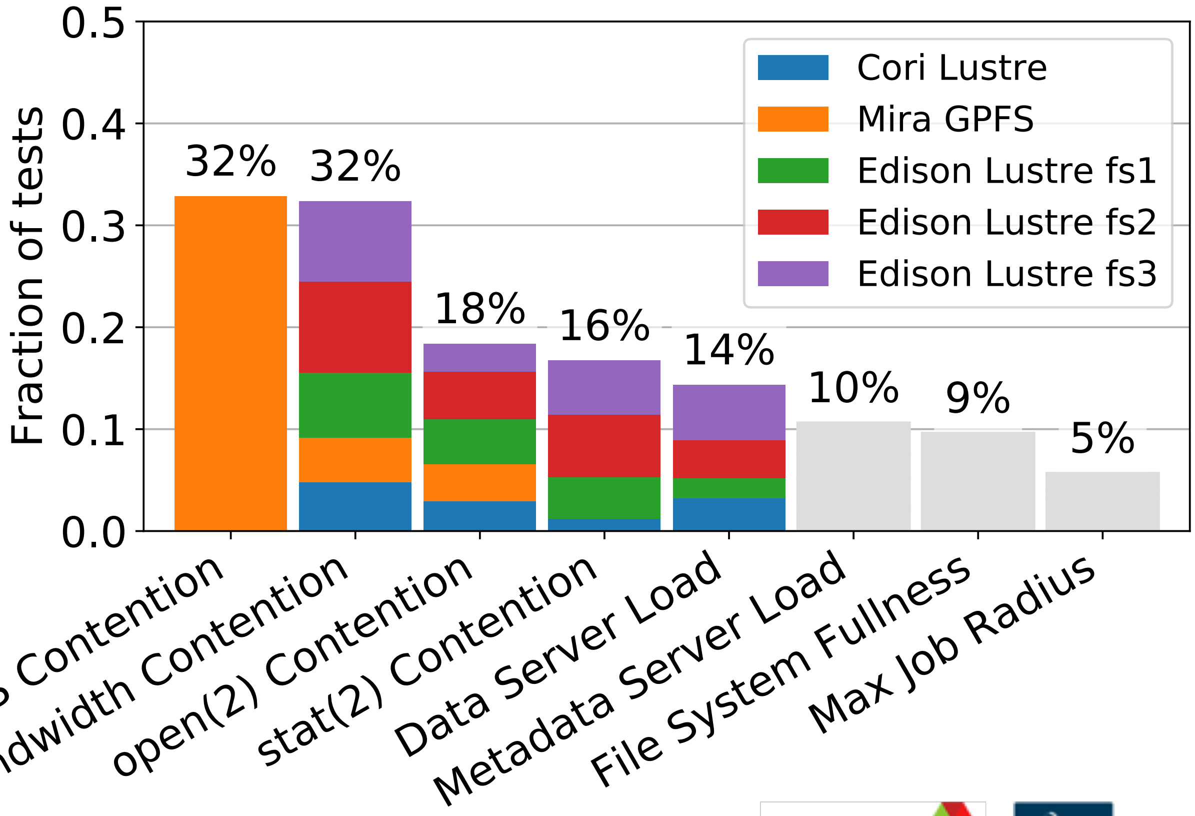
Identifying sources of transient variation

1. Identify jobs affected by transient issues
2. Define divergence regions
3. Classify jobs based on region, calculate p-values
4. Repeat for all transients and, calculate aggregate p-values



Sources of transient variation in practice

- #1 source is resource contention
- Other factors implicated but too rare to meet $p < 10^{-5}$
- 16% of anomalies defy classification



Overall findings

- **Baseline performance and variability change over time**
  - Patches & updates
  - Sustained bandwidth contention from scientific campaigns
- **Partitioning performance in time yields more insight**
  - Can classify short-term and transient variation
  - Quantifies effects of contention and suggests avenues for system architecture optimization
- **We can learn things from other fields of study**

Try this at home!

Reproducibility (code + year-long dataset):

https://www.nersc.gov/research-and-development/tokio/a-year-in-the-life-of-a-parallel-file-system/ (or see the paper appendix)

pytokio Framework:

https://github.com/nersc/pytokio

This material is based upon work supported by the U.S. Department of Energy, Office of Science, under contracts DE-AC02-05CH11231 and DE-AC02-06CH11357. This research used resources and data generated from resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.