BLACK-BOX LOCALIZATION OF STORAGE PROBLEMS IN PARALLEL FILE SYSTEMS

Michael P. Kasick, Priya Narasimhan (Carnegie Mellon University)

MOTIVATION

Focus: Black-box problem diagnosis for parallel file systems

- Using an automated, "production-friendly" approach
- Amenable to existing clusters, with off-the-shelf components

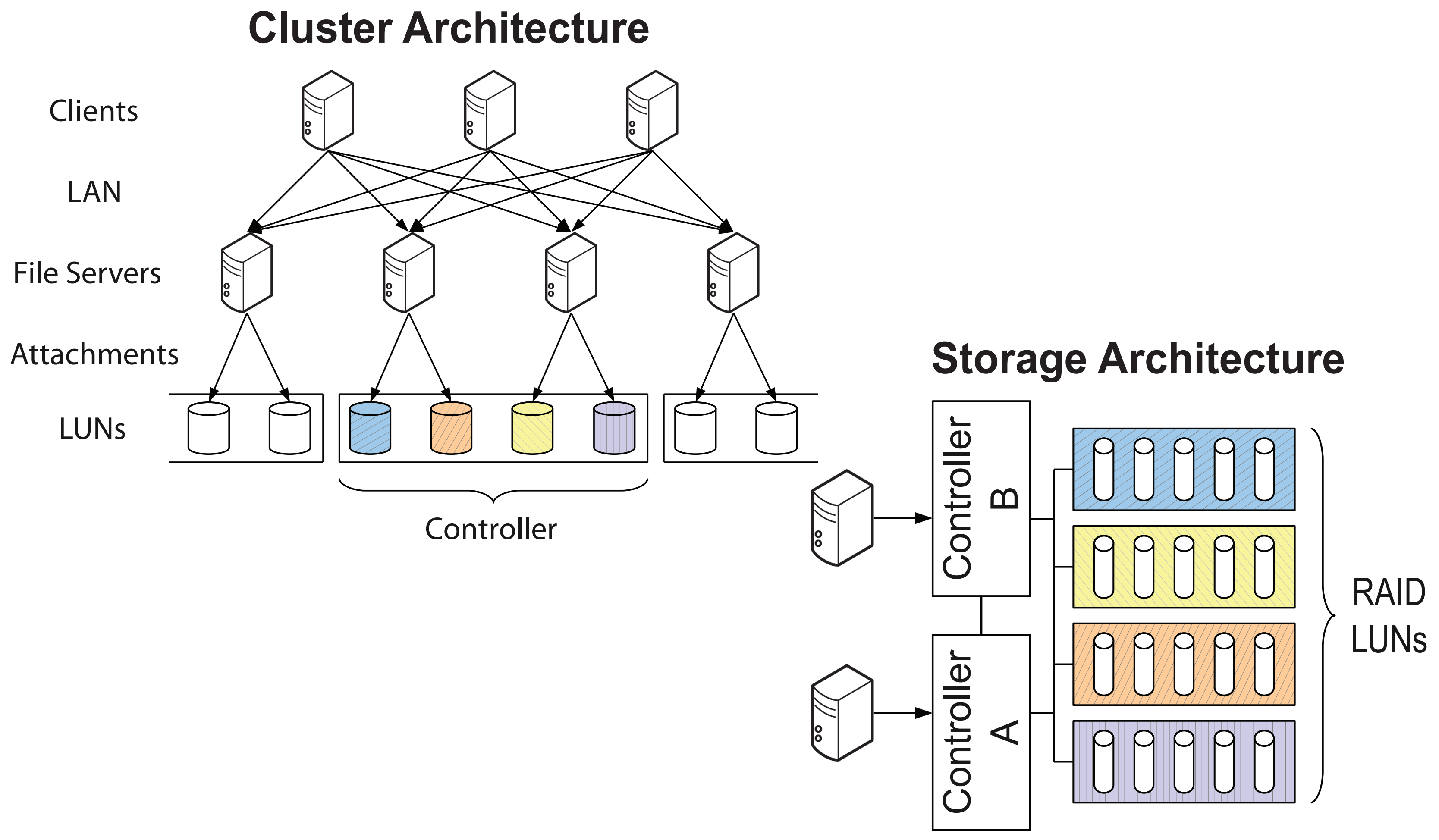
Insights:

- By design, a parallel file-system balances loads
- Components should exhibit similar performance

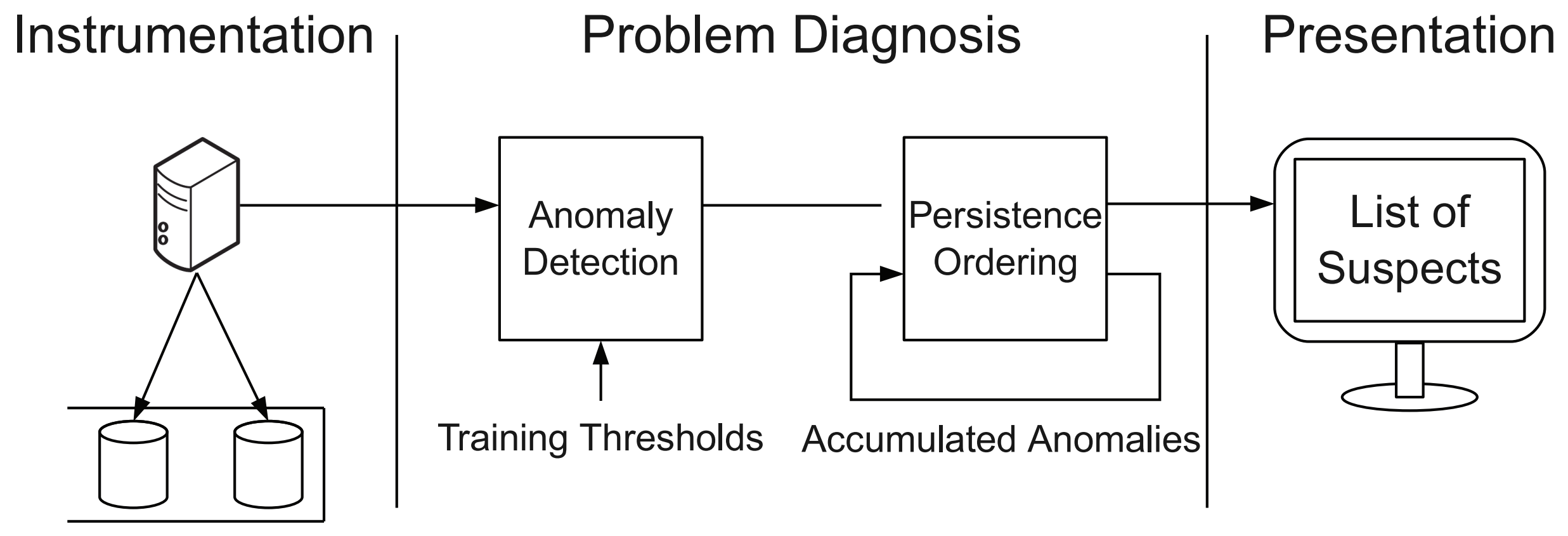
Peer-similarity hypothesis:

- Similar, fault-free components exhibit similar performance metrics
- Faulty components exhibit asymmetry in certain metrics
- Peer-comparison of metrics should identify the faulty component

ARCHITECTURE



SYNOPSIS OF APPROACH



Instrumentation:

- Sample OS-level performance metrics from each file server
- Collect samples for every LUN and network interface
- Storage metrics of interest: throughput, latency

Problem Diagnosis:

- Anomaly Detection:
  - Compares performance metrics across components
  - Identifies components that are instantaneously anomalous
- Persistence Ordering:
  - Maintains an ordered accumulation of component anomalies
  - Higher persistence implies longer-running problems

Presentation:

- Show top 100 components, most anomalous in recent history
- Calls operator attention to most problematic components

INTREPID STORAGE CLUSTER

Photo courtesy Argonne National Laboratory.

- Located at Argonne National Laboratory
- 128 file servers, 1152 LUNs across 16 controllers
- 11,520 total disks (4.5 PB)

OBSERVED PROBLEMS

Lost Attachments:

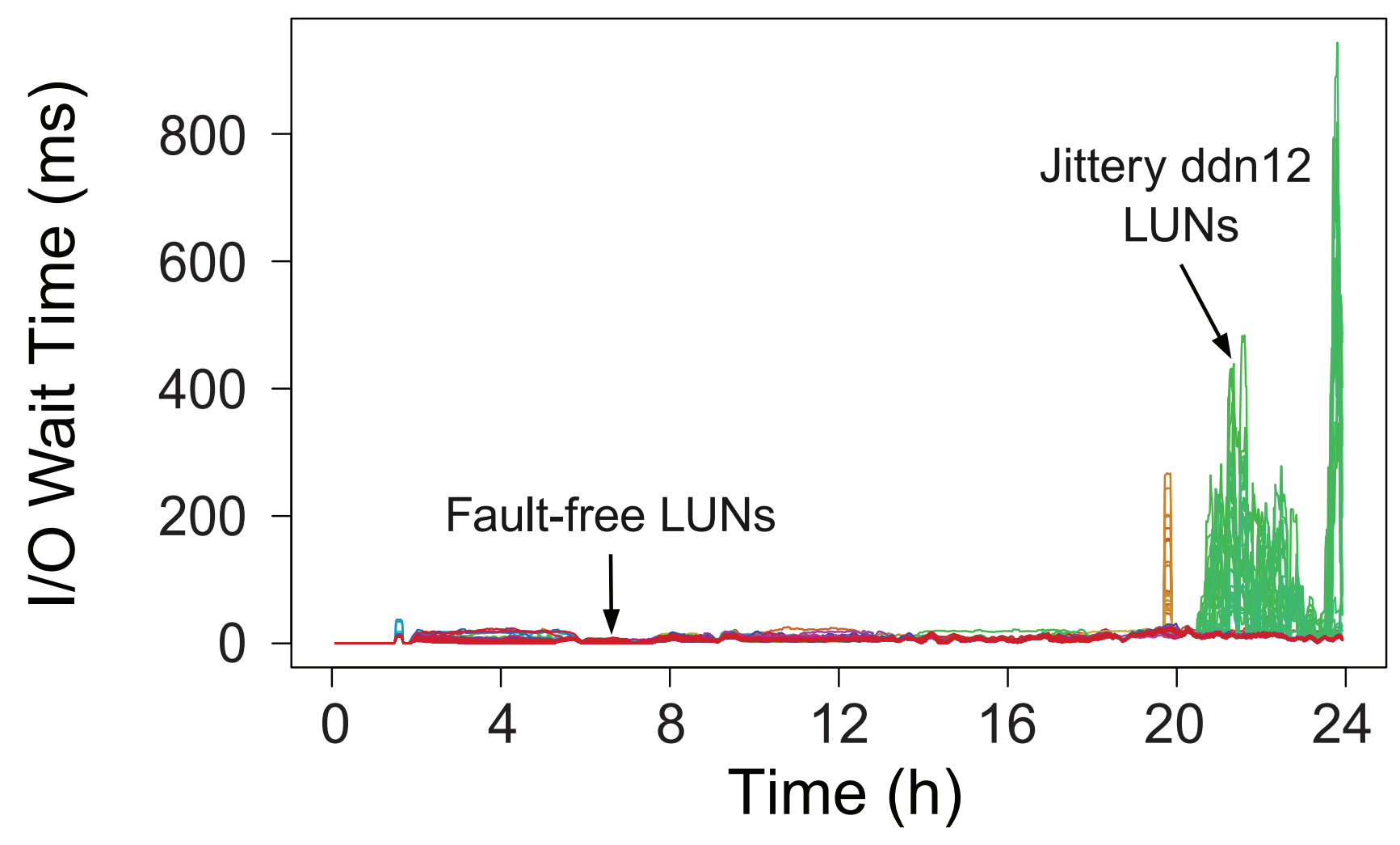
- Server stops routing I/O for one or more LUNs
- Storage controllers failures (5 incidents)
- File server failures (3)
- Missing resources after reboot (3)
- Misconfigured cache coherency (2)

Cascaded Failure:

- Controller performs 71 "LUN resets"
- Delays I/O responses up to 103 seconds
- Three file servers timeout and refuse further I/O
- Diagnosed in 39 minutes, undiscovered for 50 days

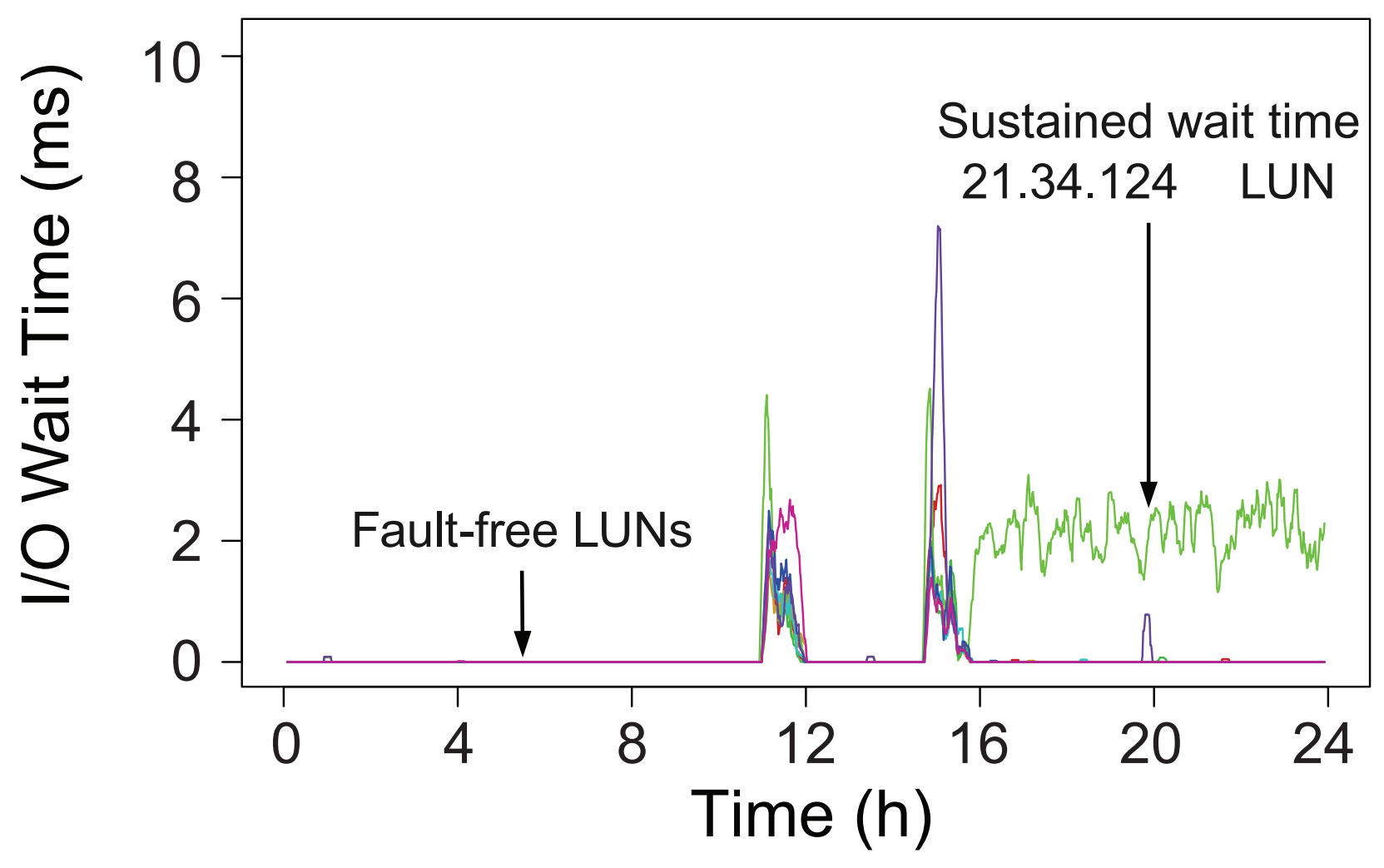
Drawer Errors (4 incidents):

- I/O errors on many disks within a single drawer
- Become very frequent, add consider jitter to I/Os



Single LUN Events (40 incidents):

- LUN exhibits considerable I/O wait time
- Durations up to 11 days, in absence of any workload



RESULTS

- Diagnosed problems in a real-world cluster
- With latencies (1-22 hours) comparable to methods currently used by operators
- Even identified problems that operators' methods miss

Intel Science & Technology Center for Cloud Computing
Carnegie Mellon University
Georgia Tech
intel
Princeton University
UC Berkeley