OAK RIDGE National Laboratory

PolKA Server: A Distributed Policy Engine for Lustre (An Update)

Chris Brumgard

brumgardcd@ornl.gov

LUG 2023

ORNL is managed by UT-Battelle LLC for the US Department of Energy

U.S. DEPARTMENT OF ENERGY

Problem (As a reminder)

- At ORNL, filesystems are becoming increasingly complex to accommodate the needs for faster storage as well as larger storage but at the same time $$$ matters.
  - Tiers
    - Users are terrible at managing the spatial and temporal placement of their data.
  - Unified namespace
    - Tricks them into thinking that their storage is equally accessible.
      - They will forget to purge and migrate.
- Beyond just tiering, admins want an easy and reliable way to implement different policies for different users/groups.
  - Telemetry and querying.

Enter PoliMOR

- A Distributed, Extensible, and Automated Policy Engine for Lustre.

- Purposes:
  - Migration, Purging, Data collection, and telemetry

- What do we mean by Distributed?
  - Agent-Oriented Microservices
  - Distributed Messaging queue
  - Fault Tolerance & Scalable

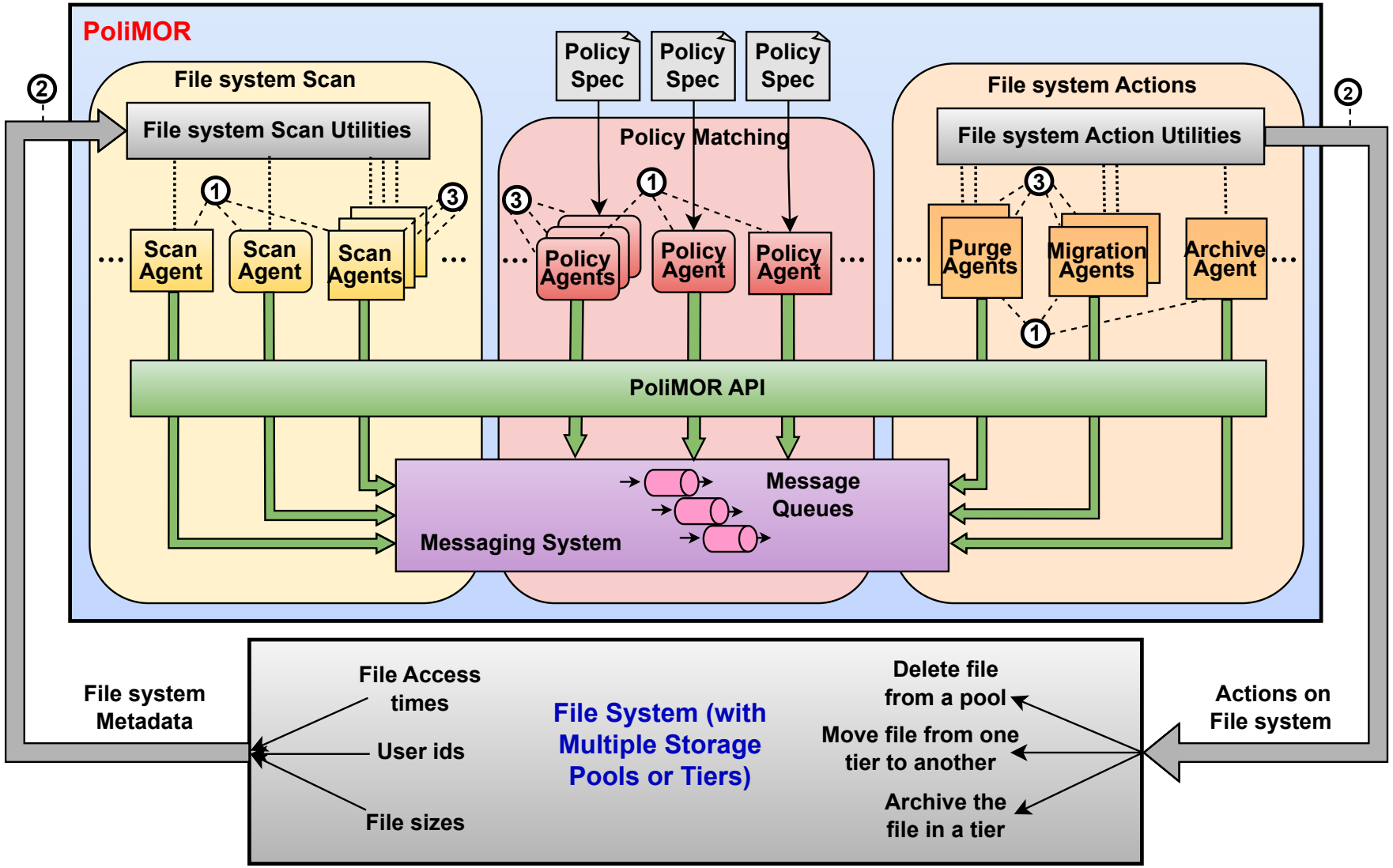
Enter PoliMOR

- A Distributed, Extensible, and Automated Policy Engine for Lustre.

- What do we mean by Extensible?
  - New agents can be added for functionality.

- What do we mean by Automated?
  - No intervention by the users or admins.
  - Define a set of invariants within a policy to be maintained.

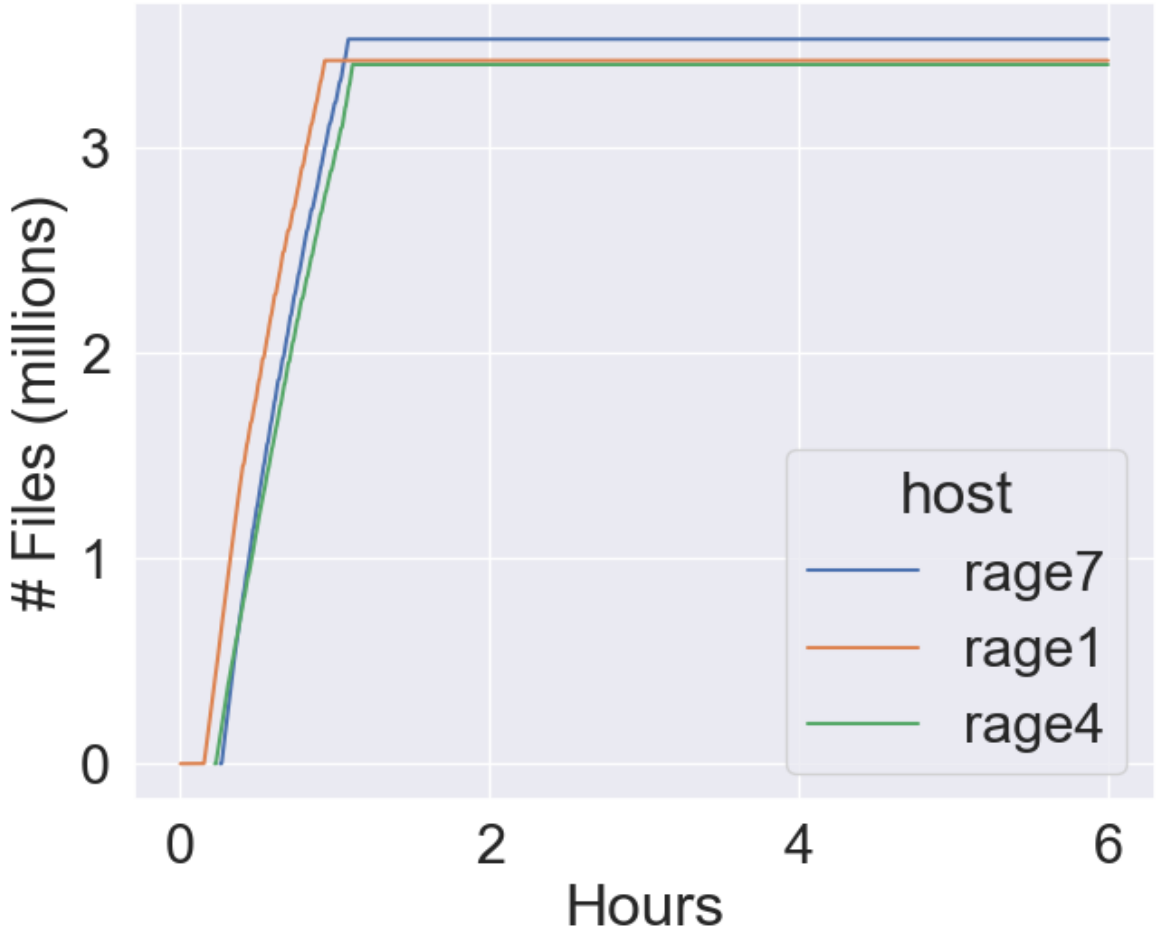
PoliMOR diagram

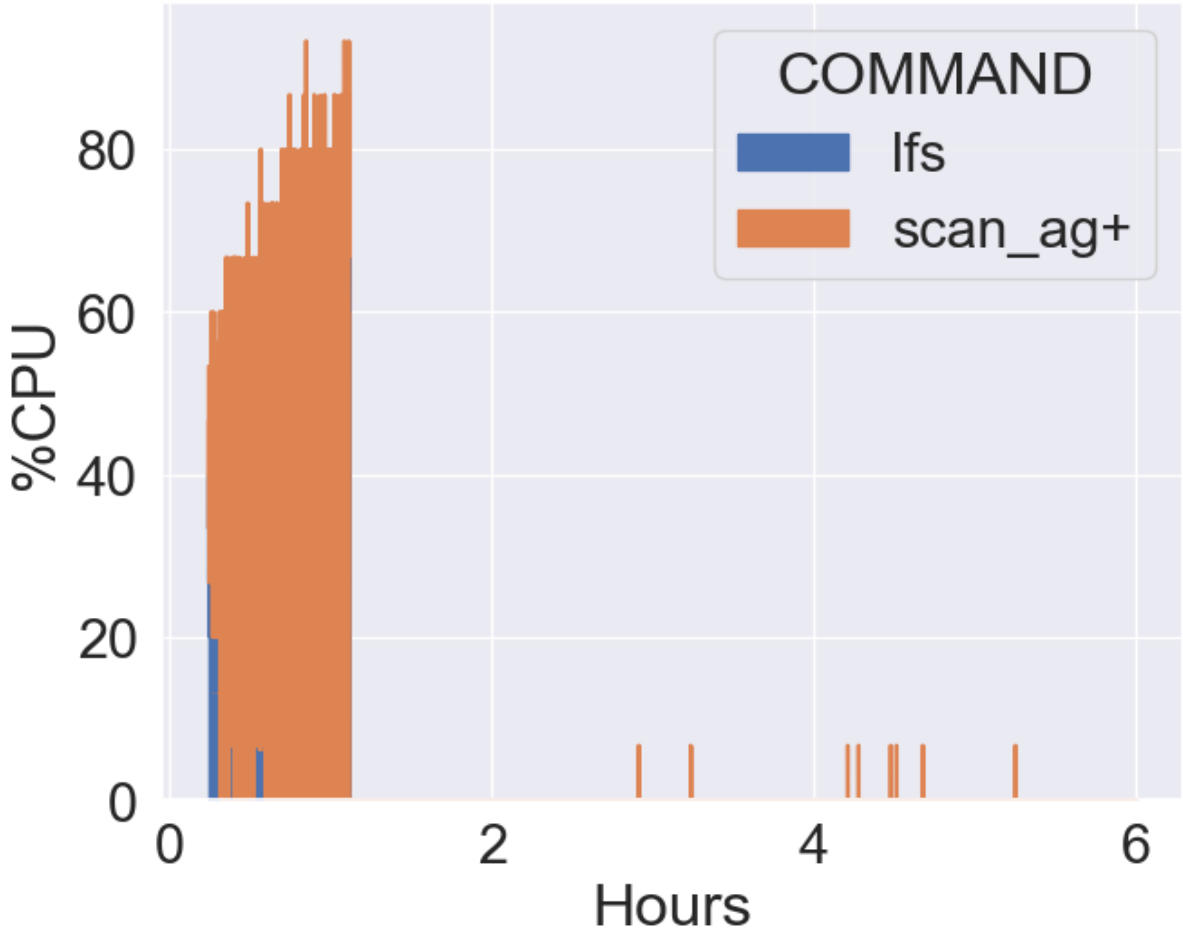


Performance Test Setup

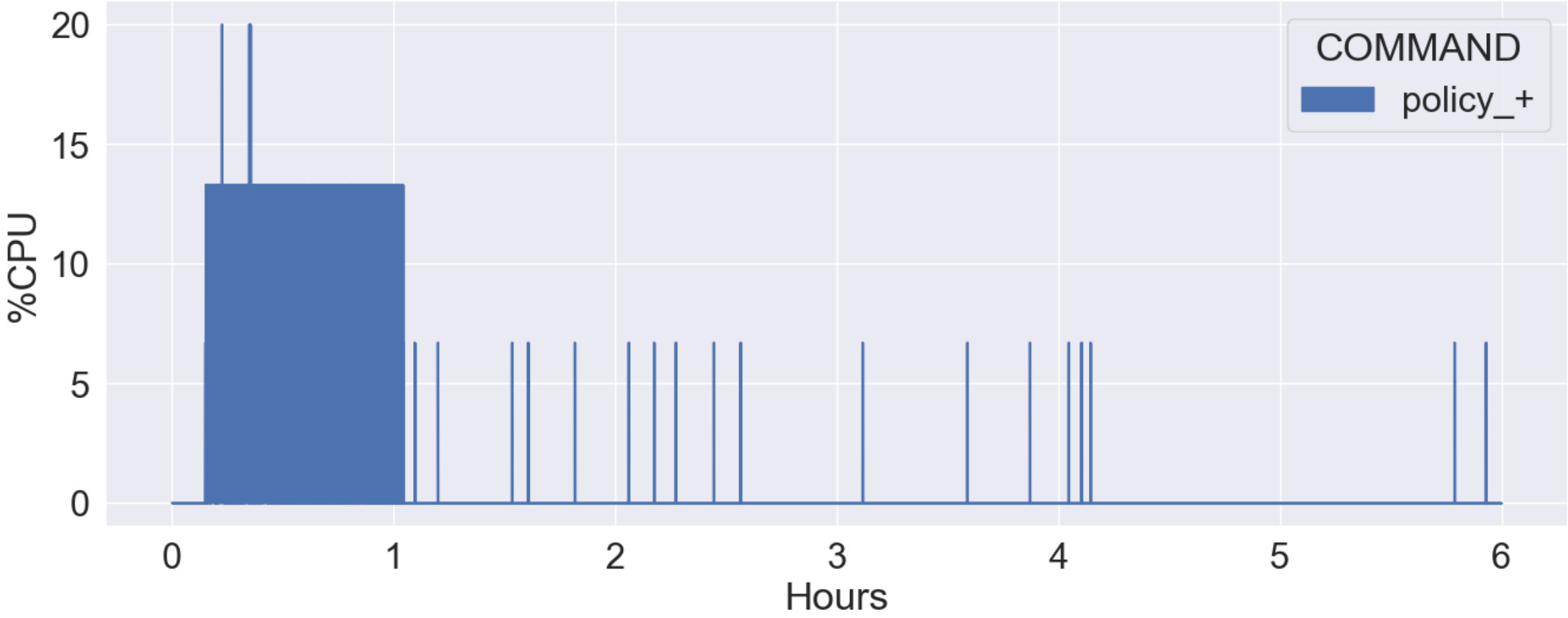
- Testbed
  - 11 nodes AMD EPYC 7351 16-core processor, 126 GB of RAM.
  - 6 OSSes and 4 MDS.
  - EDR Infiniband.
- ~10 million files spread across 100 project directories.
  - 80% on the capacity tier.
  - 20% on the performance tier.
  - Random timestamping for migration and purging.
  - File size used bimodal distribution found on production file system.
- 3 Scan agents, 2 policy agents, 3 purge agents, and 3 migration agents.

Scan Agent Characteristics

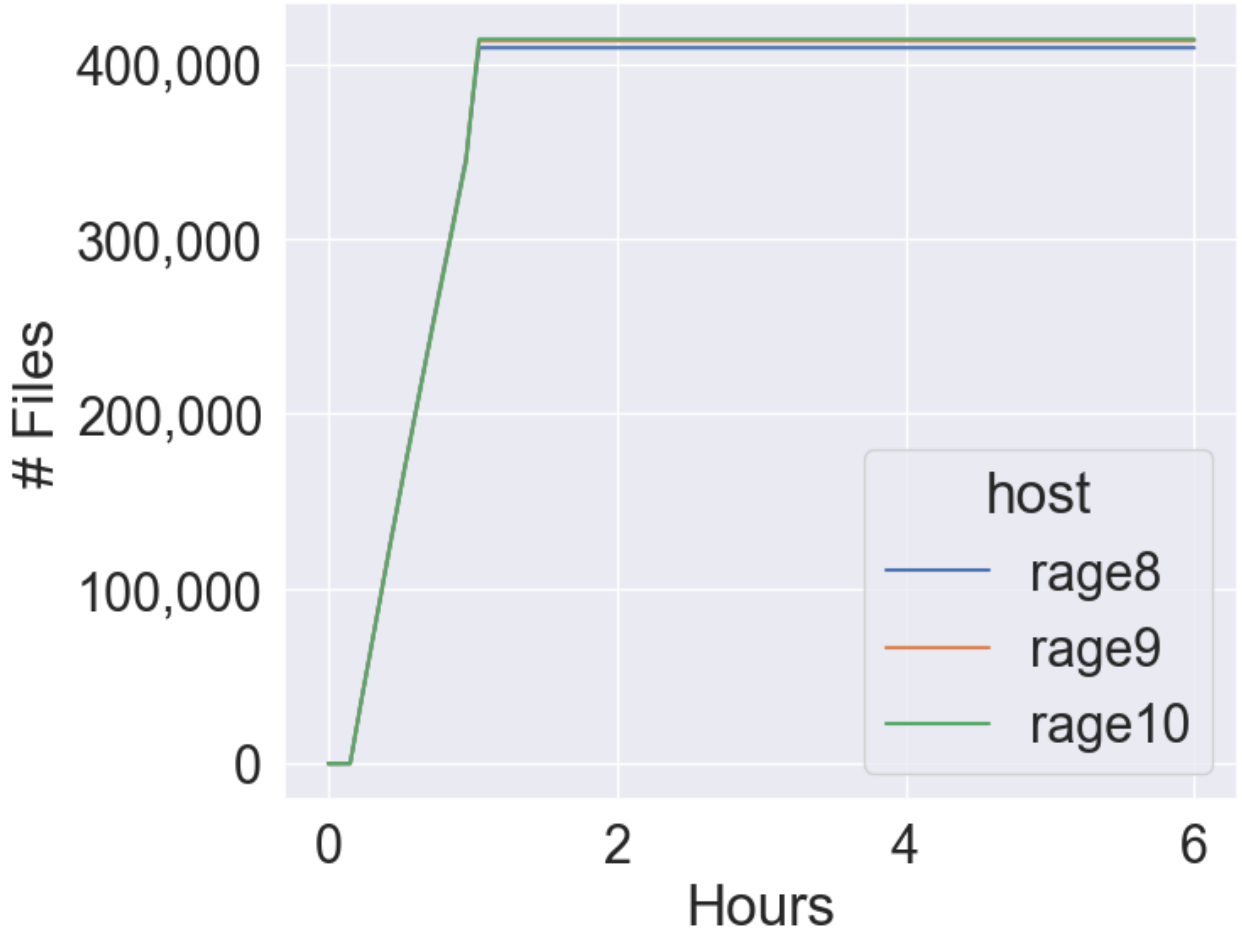


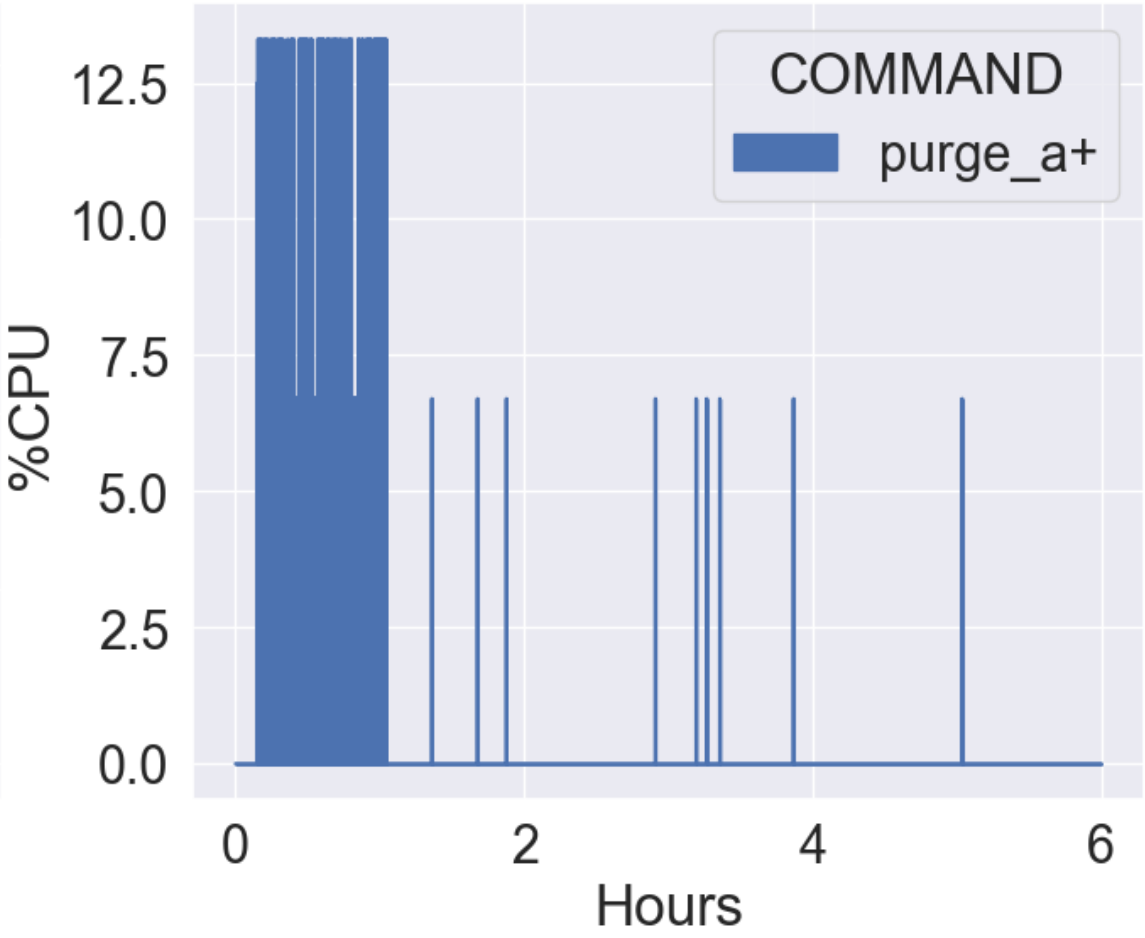


Policy Agent Characteristics



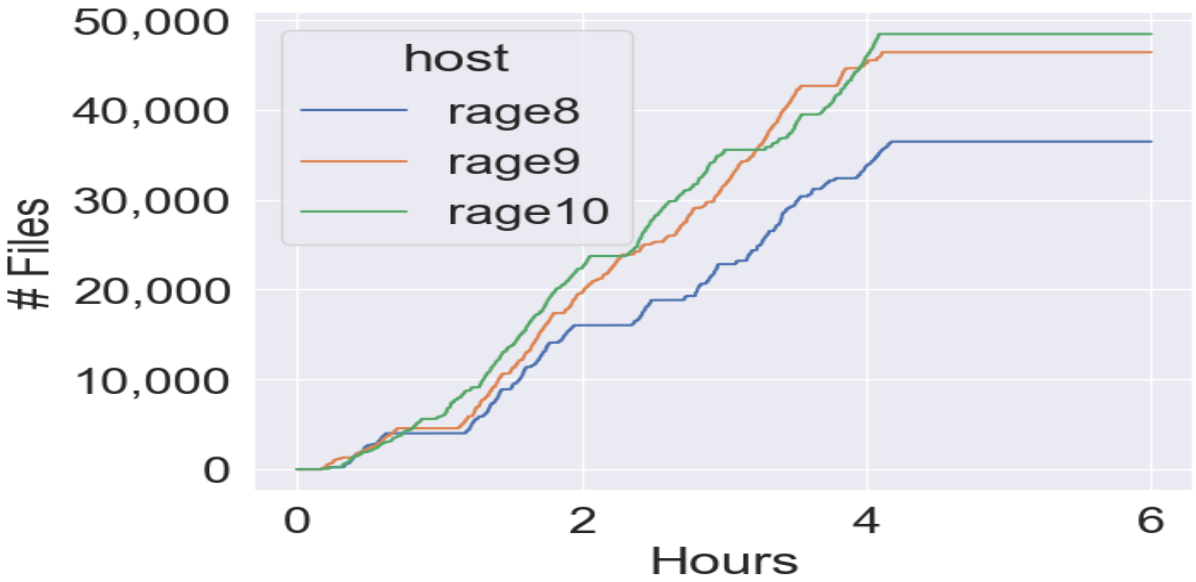
Purge Agent Characteristics

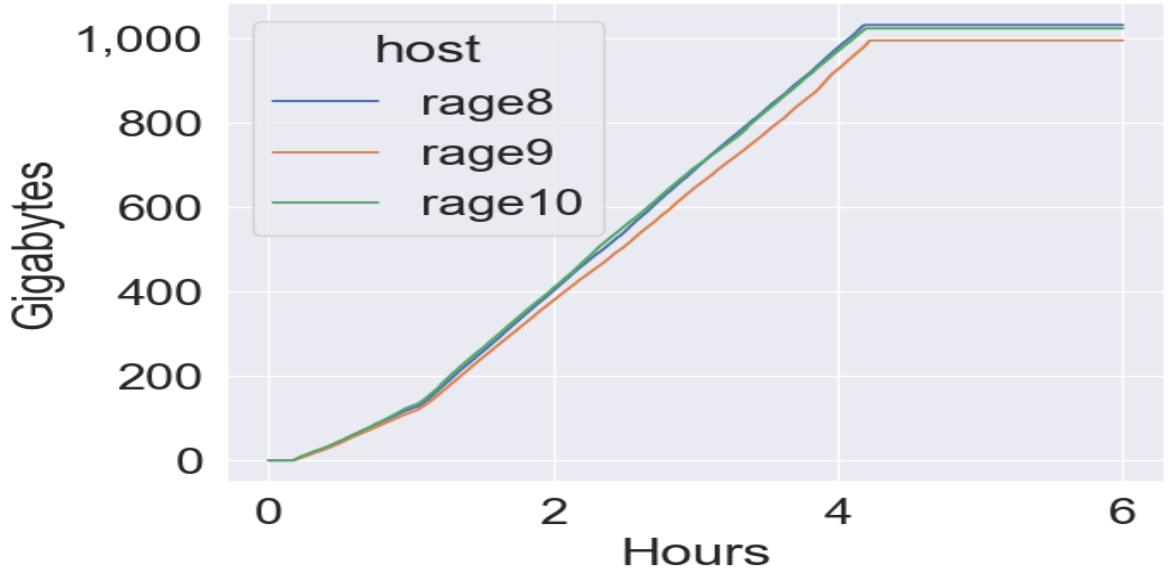


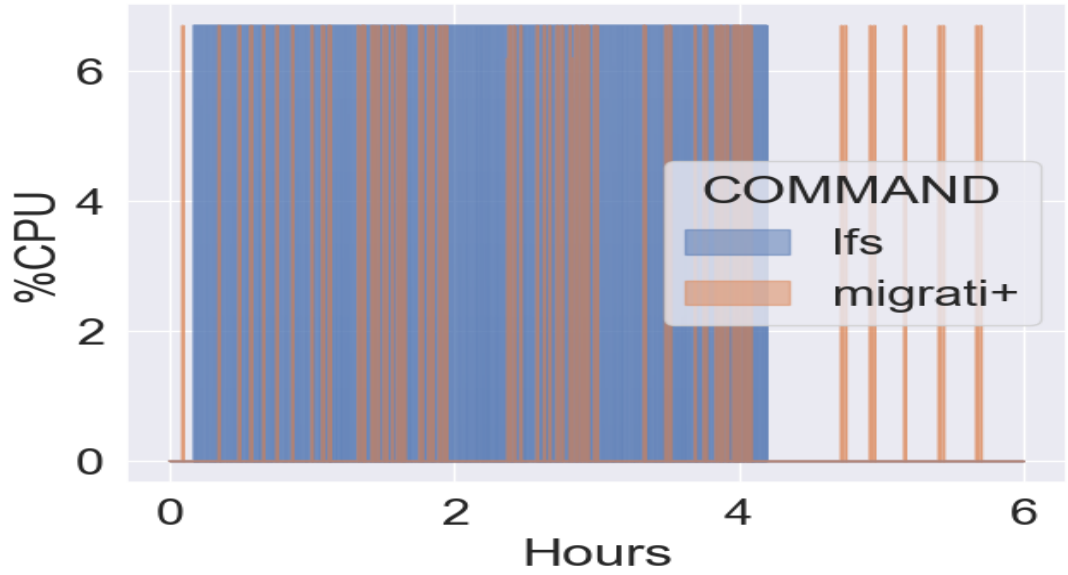


OAK RIDGE National Laboratory

Migration Agent Characteristics

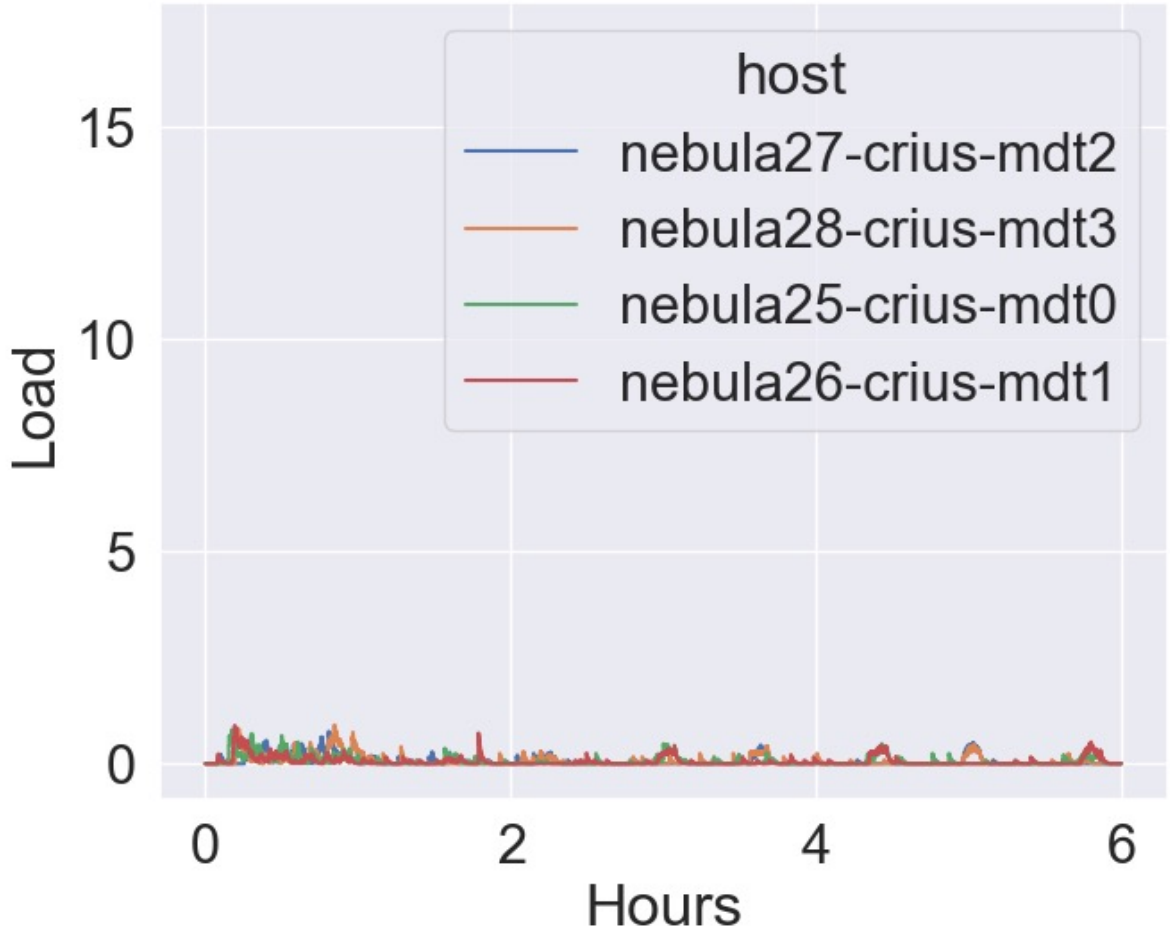


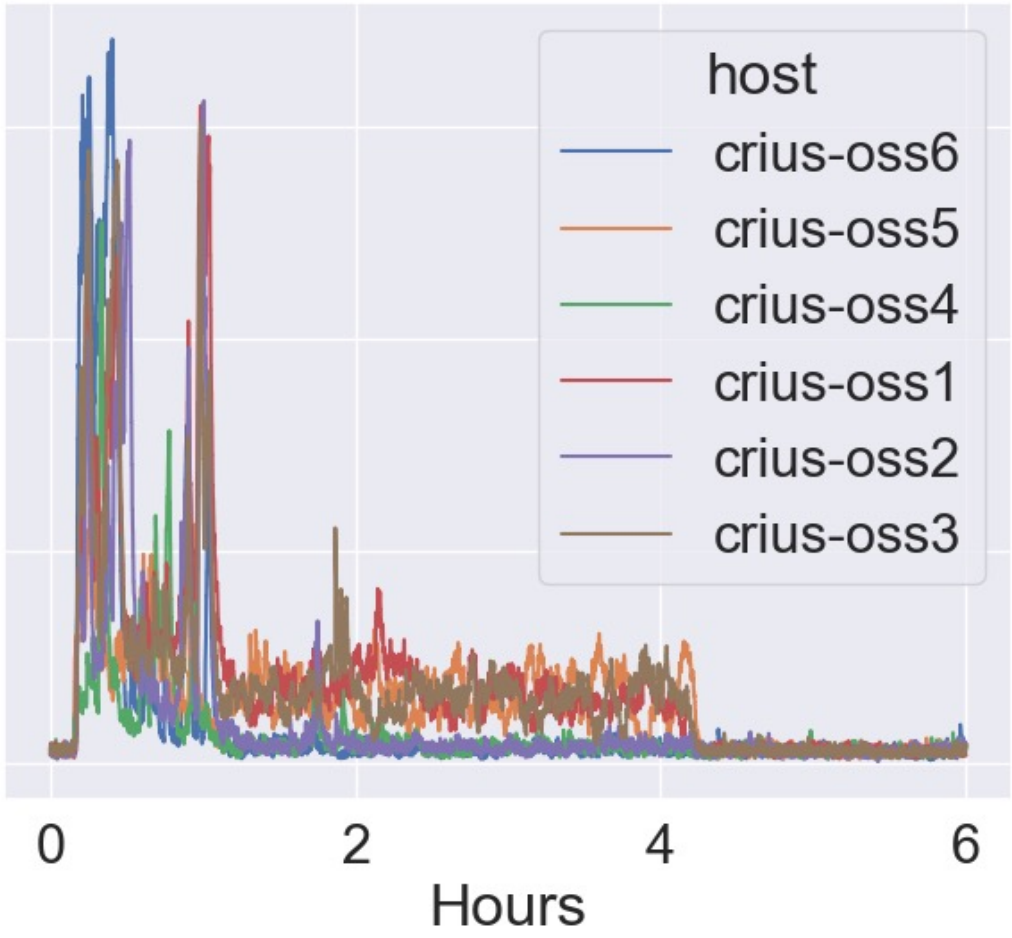




OAK RIDGE National Laboratory

Impact on MDS and OSS

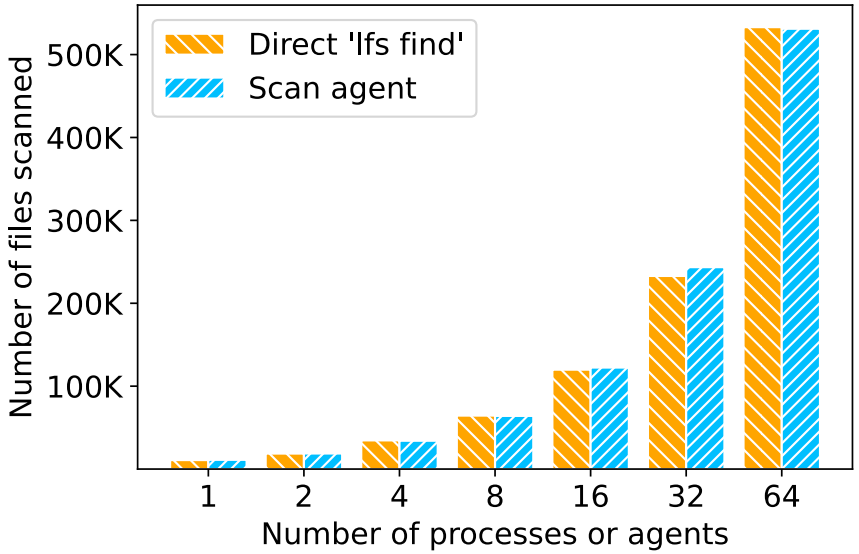


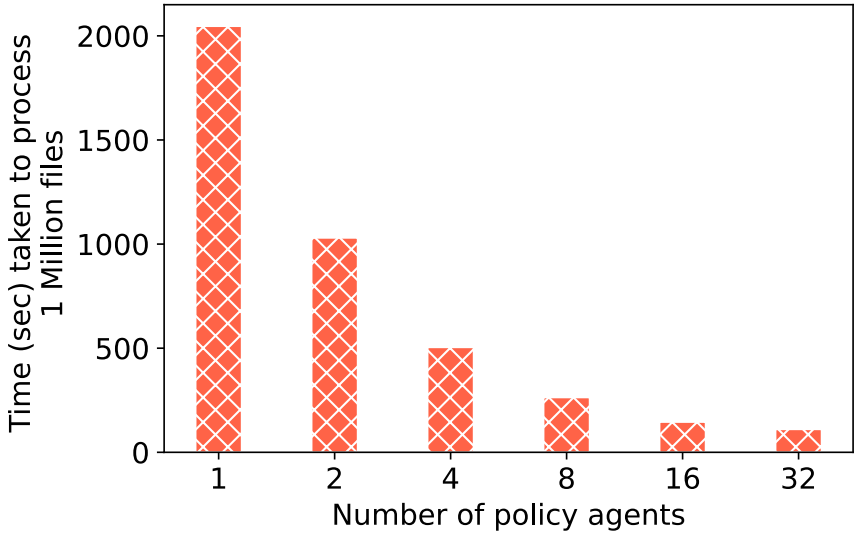


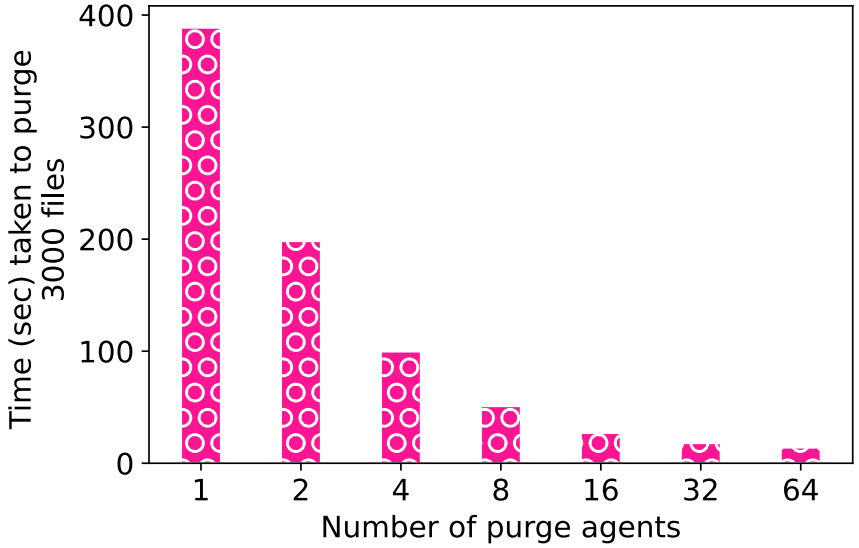
Scalability Results

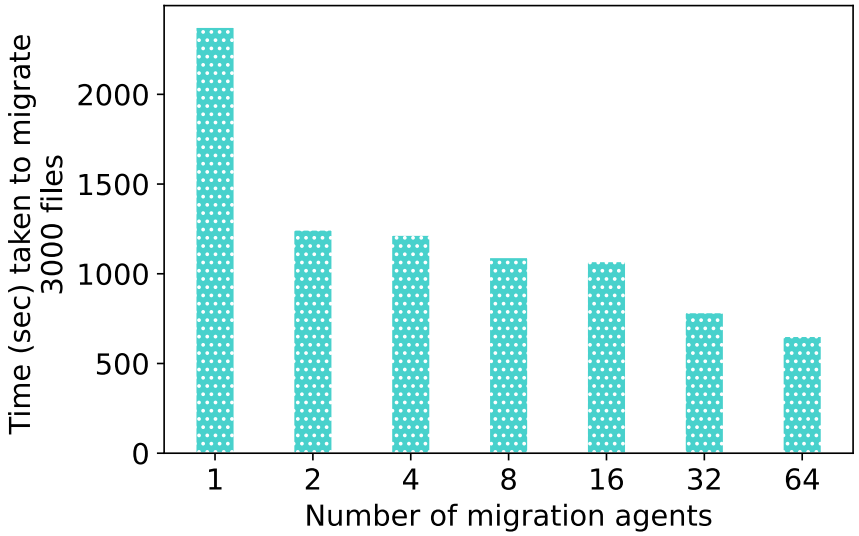
- Used 8 nodes of the testbed.
- Deployed 1 to 64 agents across the nodes.
- Tested each agent type in isolation.

Scalability Results









Future work

- Still in active development.
  - Reducing the scan work.
  - Productionizing.
  - Currently deploying to Orion.
- More complex policies.
  - Decomposing complex actions into simpler rules.
- Non-Lustre agents.
  - HPSS
  - Edge

Acknowledgments

- Thanks to my fellow developers Anjus George, Ketan Maheshwari, Rick Mohr, James Simmons.
- This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.