FusionProv: Towards a provenance-aware distributed filesystem

Chen Shou
Department of Computer Science
Illinois Institute of Technology
cshou@hawk.iit.edu

Dongfang Zhao
Department of Computer Science
Illinois Institute of Technology
dzhao8@hawk.iit.edu

Tanu Malik
Computation Institute
The University of Chicago
tanum@ci.uchicago.edu

Ioan Raicu
Department of Computer Science
Illinois Institute of Technology
iraicu@cs.iit.edu

Goal

Develop FusionProv, a distributed provenance management system based on FusionFS that offers excellent scalability and load balancing in exascale computing.

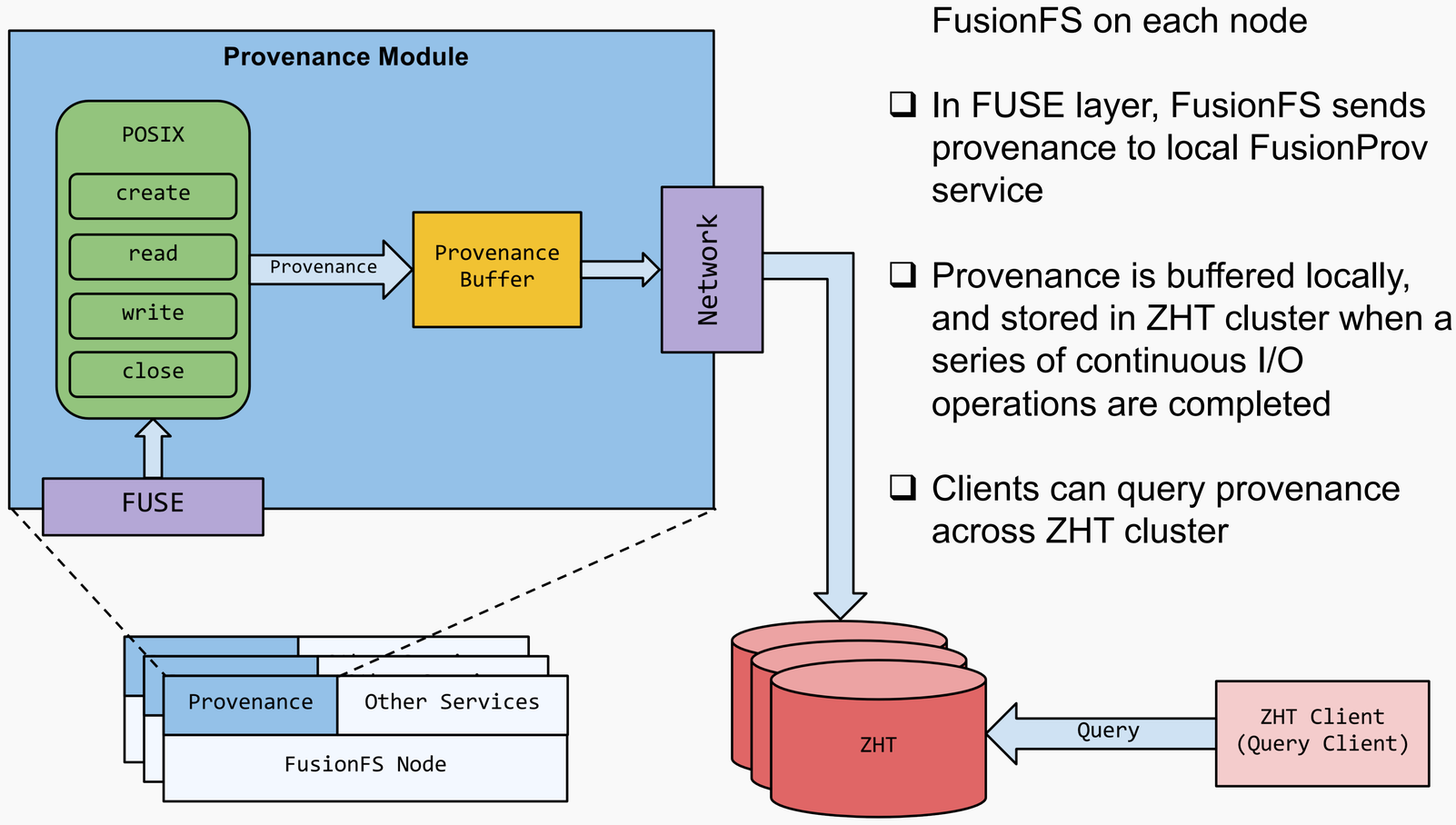
Motivation

Distributed file systems have so far proposed a central system for provenance collection, which becomes a performance bottleneck, especially for file systems meant for extreme-scales.

Building Blocks

- **FusionFS**: a distributed file system designed for extreme-scales as the host filesystem
  - Distributed metadata and data management
  - Data locality and data indexing
  - POSIX interface
  - Scales up to 8K nodes
- **ZHT**: a zero-hop distributed hashtable as the storage system of provenance
  - Scales up to 8K nodes
  - Reliable persistent storage
  - Light-weighted

FusionProv Architecture



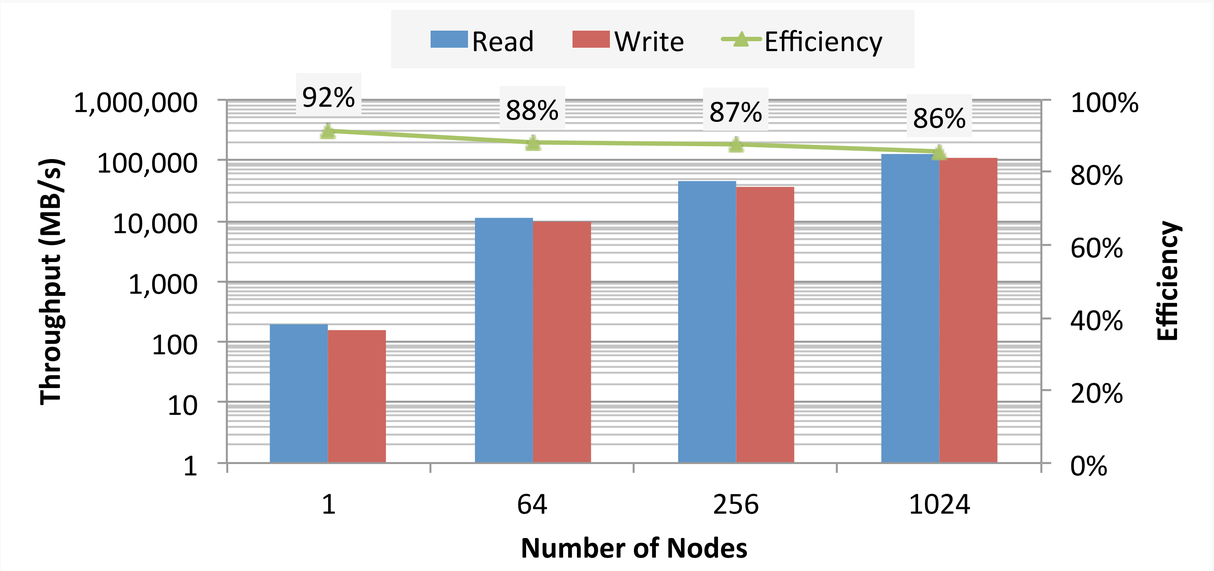
- ❑ FusionProv coexsits with FusionFS on each node
- ❑ In FUSE layer, FusionFS sends provenance to local FusionProv service
- ❑ Provenance is buffered locally, and stored in ZHT cluster when a series of continuous I/O operations are completed
- ❑ Clients can query provenance across ZHT cluster

ReusIts & Evaluation

Provenance Collection

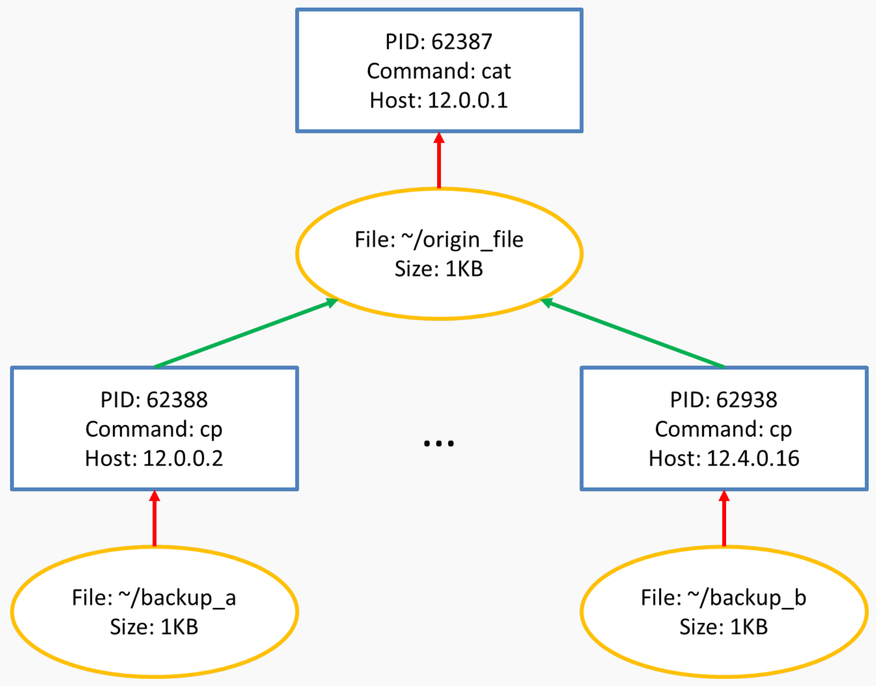
❑ Single node throughput with I/O blocks size from 16KB to 128KB

❑ Aggregate throughput at different scales

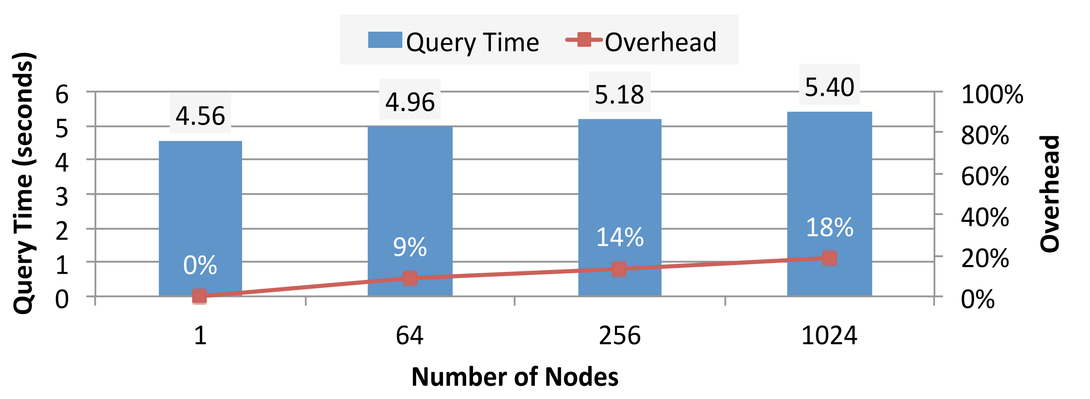


Provenance Query

❑ Provenance query sample



❑ Query executing time at different scales



Conclusion

- ❑ ***Excellent scalability*** up to 1K nodes
- ❑ ***Negligible overhead*** introduced by provenance capture
- ❑ ***Almost constant time*** to query provenance at larger scale (with the same workload)

References

[1] Dongfang Zhao and Ioan Raicu. Distributed File Systems for Exascale Computing (poster). ACM/IEEE Supercomputing, Salt Lake City, UT, 2012.

[2] Tonglin Li, Xiaobing Zhou, Kevin Brandstatter, Dongfang Zhao, Ke Wang, Anupam Rajendran, Zhao Zhang, and Ioan Raicu. ZHT: A Light-weight Reliable Persistent Dynamic Scalable Zero-hop Distributed Hash Table. IEEE IPDPS, Boston, MA, 2013, to appear.

Acknowledgement

- This work is supported by NSF grant OCI-1054974.
- Thanks to Xian-He Sun (IIT) for providing the access to the HEC cluster.
- Special thanks to Tonglin Li (IIT) and Xiaobing Zhou (IIT) for insightful discussions.