Scalable Distributed Directory Implementation on Orange File System

Shuangyang Yang
Walter B. Ligon III
Elaine C. Quarles
Clemson University

CLEMSON UNIVERSITY

OrangeFS DISTRIBUTED DIRECTORIES

Outline

- Orange File System
- Motivation for Distributed Directory
- Design and Implementation
- Micro-benchmark Performance
- Summary and Conclusion
- Future Work

OrangeFS
DISTRIBUTED DIRECTORIES

What is OrangeFS?

- High performance, parallel file system
- Continuation of PVFS
- Concentration on production quality features
- Open source, GPL licensed
- More details at http://orangefs.org/

OrangeFS
DISTRIBUTED DIRECTORIES

Huge Directories

- Millions of files under the same directory
- Scalability issues
- Needed for:
  - Gene sequencing
  - Image processing
  - Data mining
  - Real-time application monitoring

OrangeFS
DISTRIBUTED DIRECTORIES

Motivation for Distributed Directories

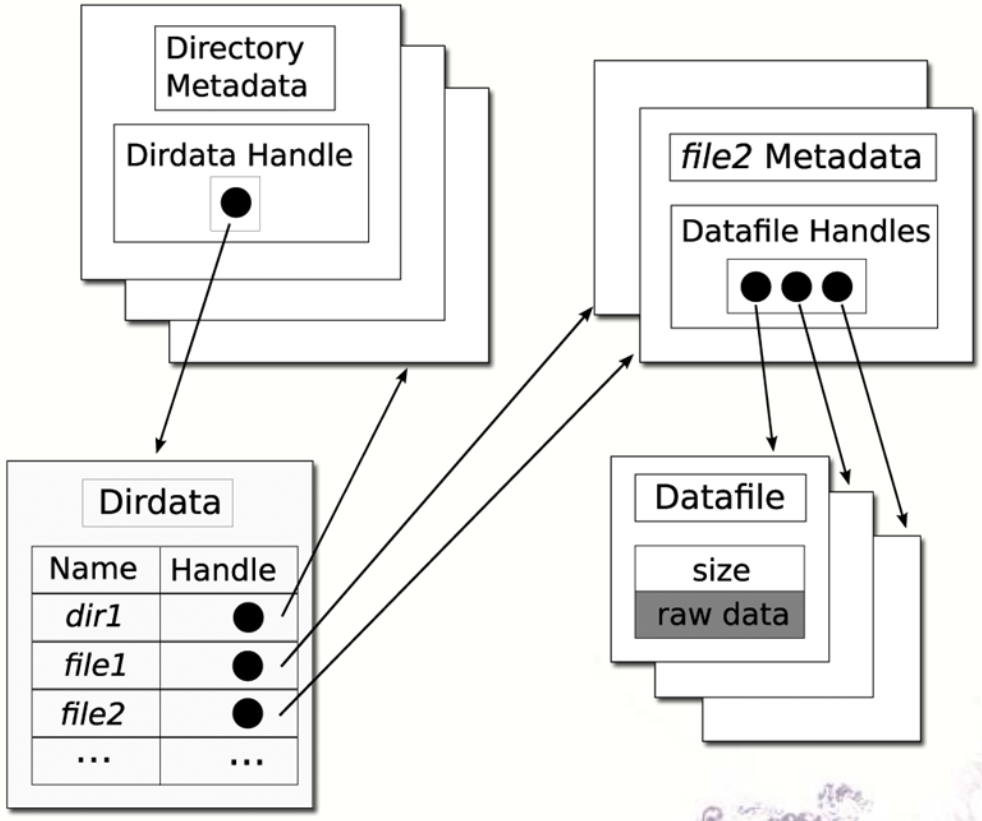
- OrangeFS stripes contents of data files among servers, BUT
- All directory entries for any given directory are held on a single server, THUS
- Creating the potential for hot spots

OrangeFS
DISTRIBUTED DIRECTORIES

OrangeFS Structure

Terminology:

- Metadata
- Datafile
- Dirdata



OrangeFS
DISTRIBUTED DIRECTORIES

Existing Techniques

- Distribute directory entries in multiple blocks or partitions on multiple servers.
- GPFS: multiple disk blocks. Extensible hashing to lookup and grow.
- GIGA+: multiple fixed-size partitions. Partition bitmap to maintain status. Extensible hashing and incremental growth.

OrangeFS
DISTRIBUTED DIRECTORIES

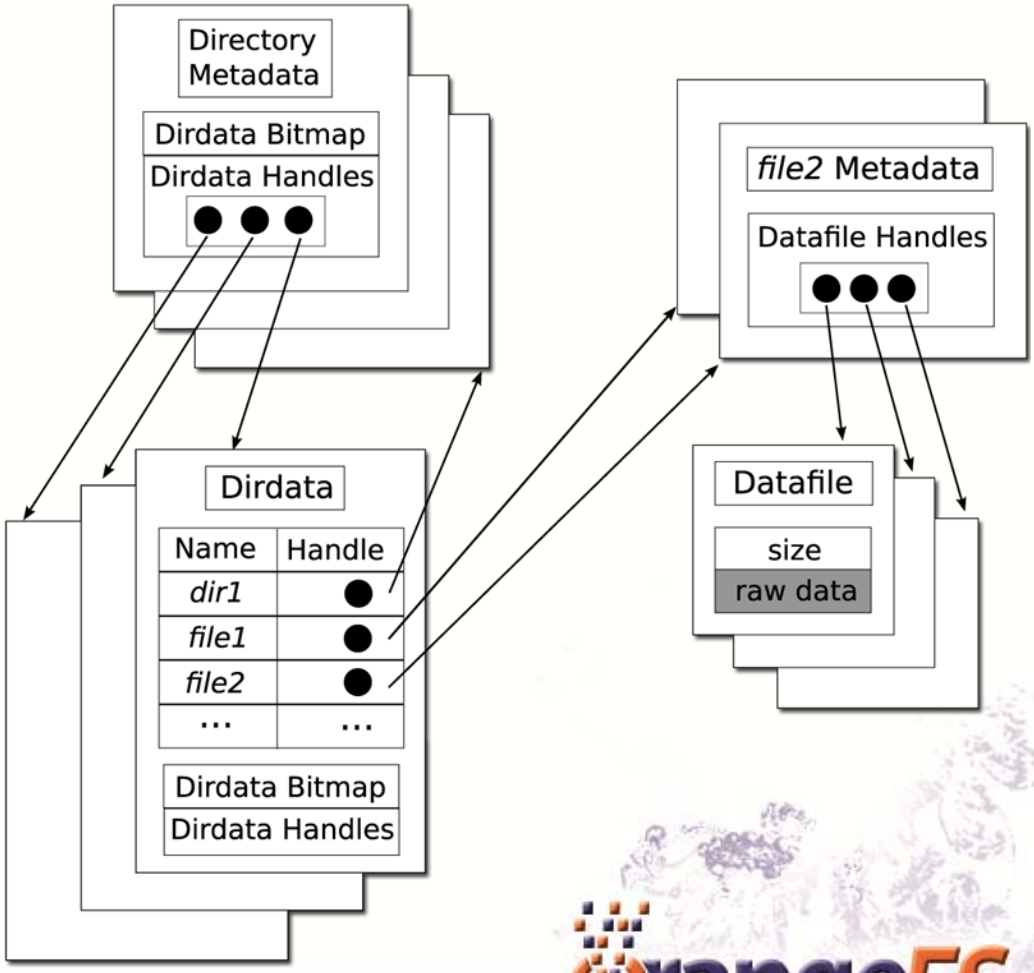
Goals of Current Work

- Distributed directory implementation based on GIGA+. Utilizing extensible hashing and partition bitmap representation.
- Different design decisions in order to be integrated into OrangeFS seamlessly.
- Evaluate performance of initial implementation

OrangeFS
DISTRIBUTED DIRECTORIES

Modified Structure

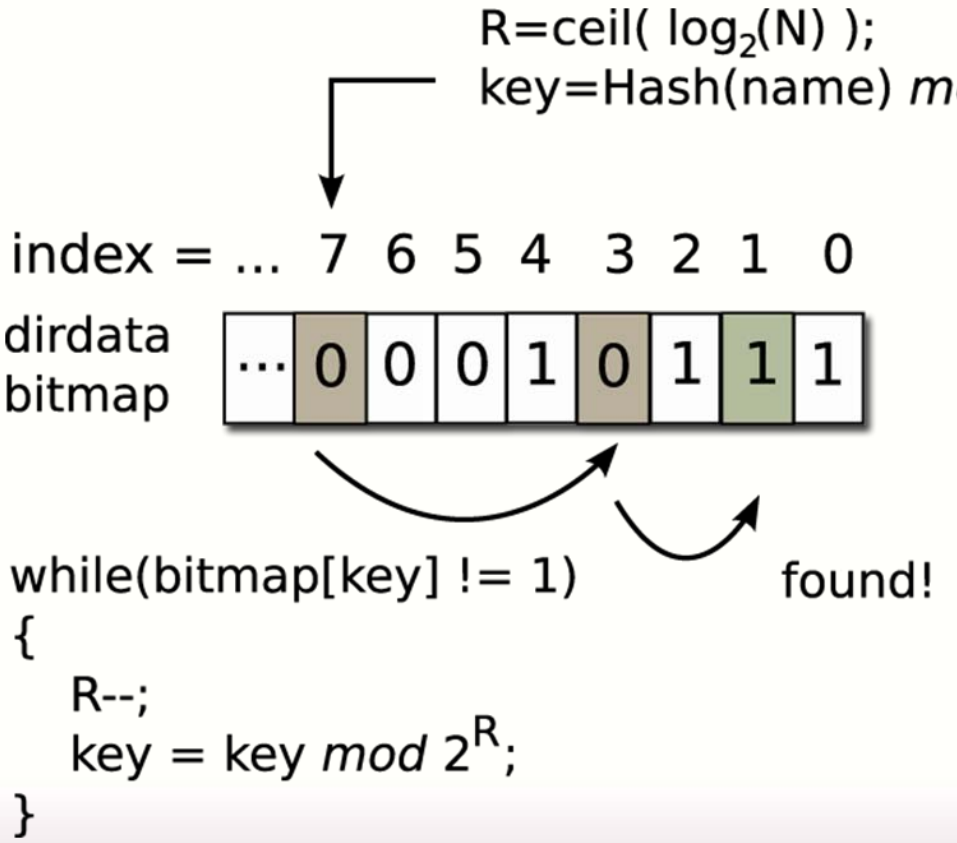
- Multiple dirdata objects represented by handles.
- Dirdata bitmap
- Stored on both metadata object and dirdata object.



OrangeFS
DISTRIBUTED DIRECTORIES

Lookup

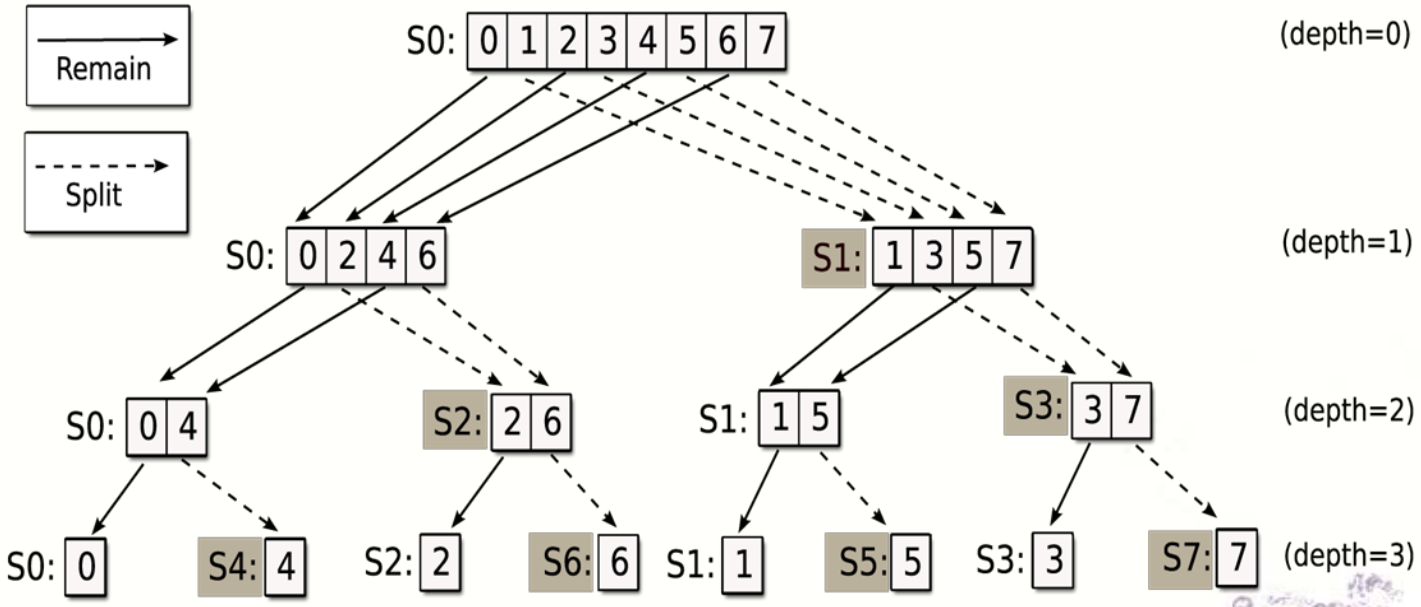
- Map a directory entry to a dirdata object.



OrangeFS
DISTRIBUTED DIRECTORIES

Split Operation

- Overloaded dirdata object can send half of its hash space to another dirdata object.



OrangeFS
DISTRIBUTED DIRECTORIES

Design Decisions

- One dirdata object on one metadata server, entries indexed by Berkeley DB
- Initial number of active dirdata objects is configurable
- Metadata object holds most up-to-date copy of dirdata bitmap

OrangeFS
DISTRIBUTED DIRECTORIES

Implementation on OrangeFS

- Directory entries are distributed among multiple metadata servers.
- Dynamic splitting is close to completion.
- The scalable distributed directory feature is available as an experimental release at

  http://www.orangefs.org/download/

OrangeFS
DISTRIBUTED DIRECTORIES

Micro-benchmark performances

- Modified version of UCAR metarates benchmark*.
- Measure throughput of file creation and removal under one directory concurrently.
- Conducted on the Palmetto Cluster housed by Clemson University.

OrangeFS
DISTRIBUTED DIRECTORIES

* http://www.cisl.ucar.edu/css/software/metarates/

Palmetto Cluster

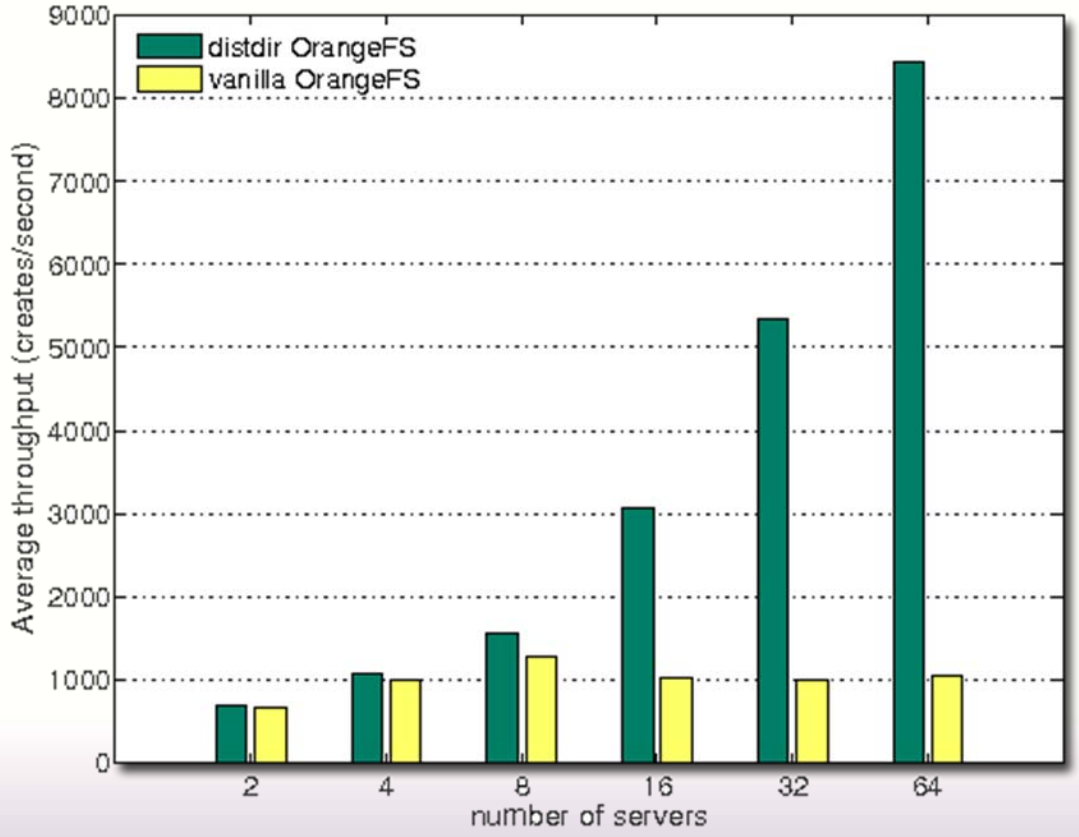
- 1541 nodes, 8 cores.
- 12/16 GB memory.
- Myrinet 10G interconnect.
- High throughput storage.
- More at

http://citi.clemson.edu/palmetto

OrangeFS
DISTRIBUTED DIRECTORIES

File Creation

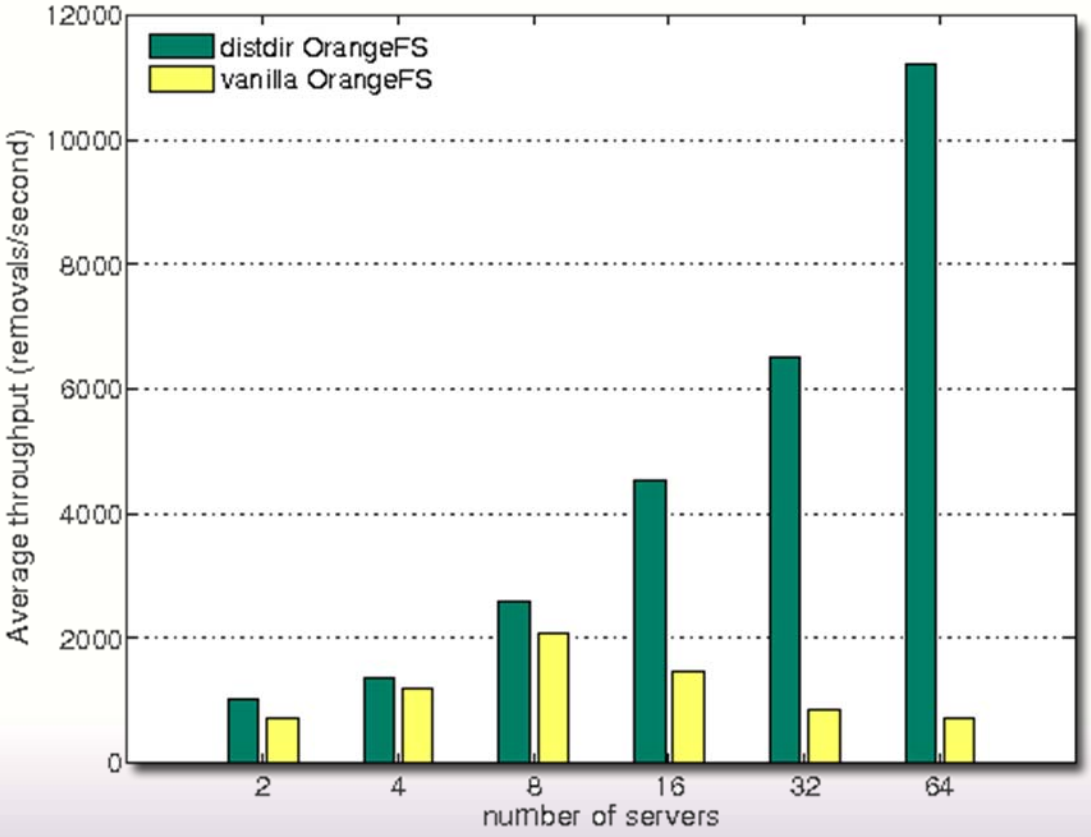
- Average throughput can reach 8000+ creations/second with 64 servers, 128 clients.



OrangeFS
DISTRIBUTED DIRECTORIES

File Removal

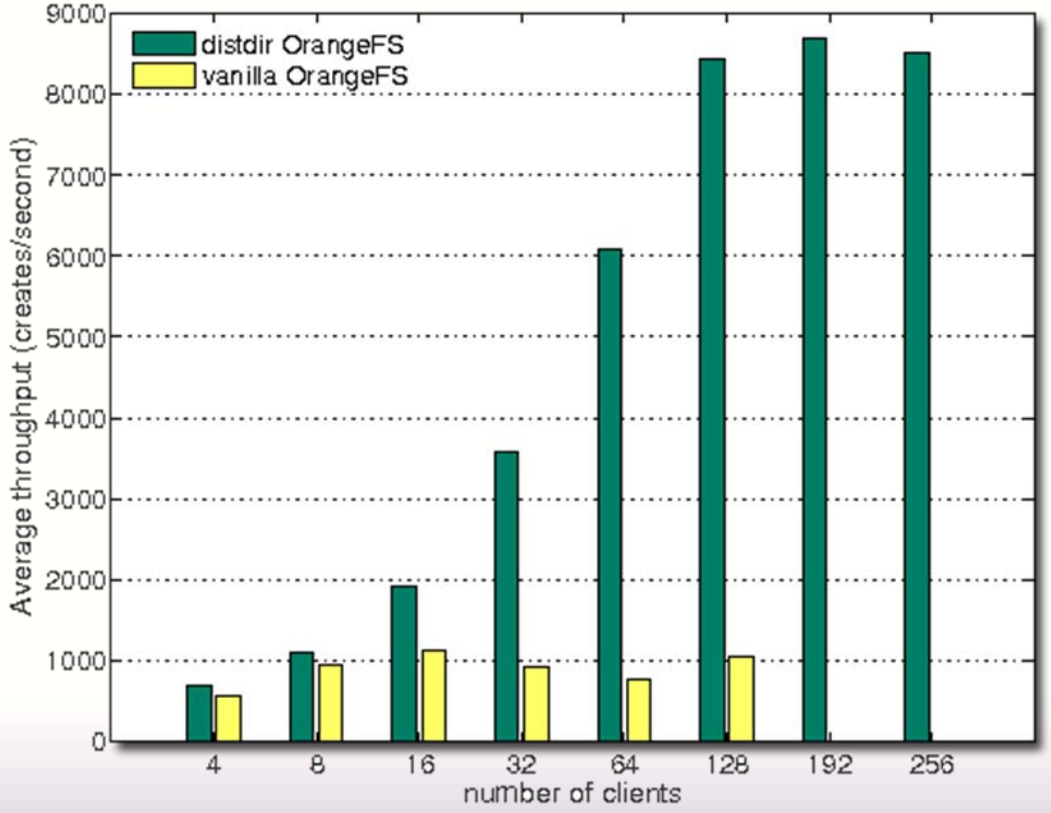
- Average throughput can reach 11000+ removals/second with 64 servers, 128 clients.



OrangeFS
DISTRIBUTED DIRECTORIES

Creation with Variable Clients

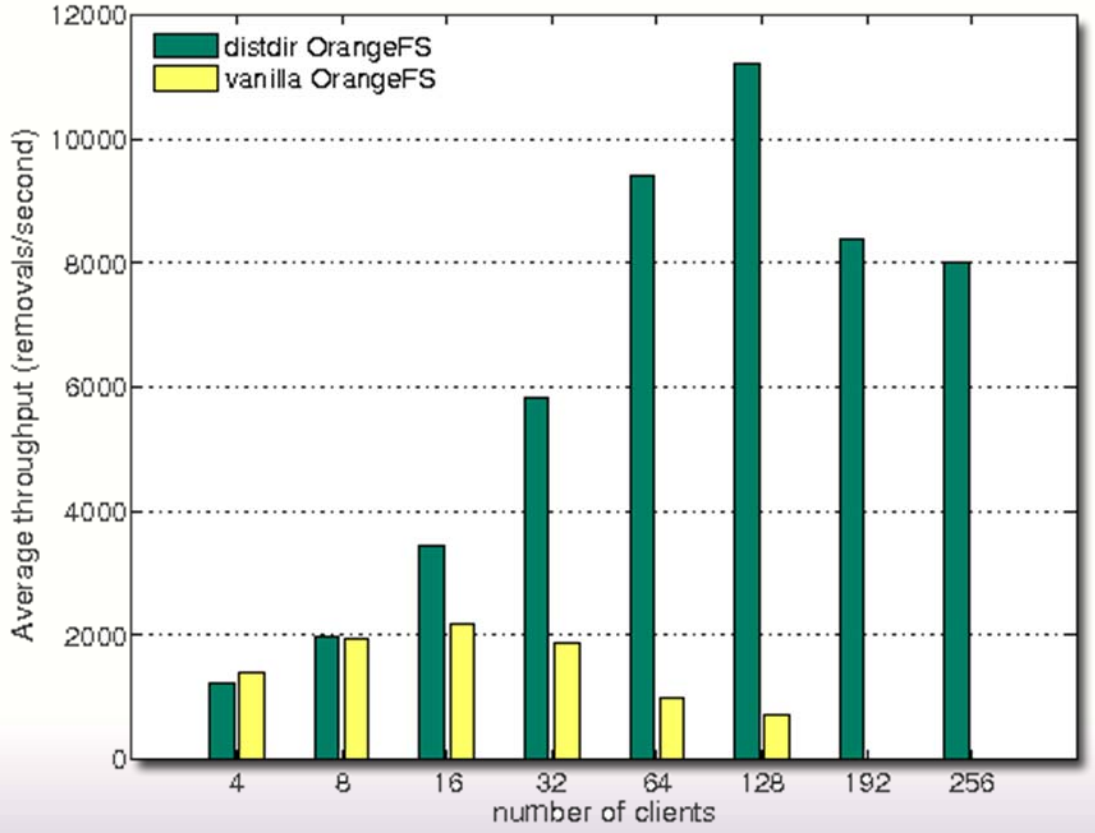
- Maintain high throughput when saturated on 64 servers.



OrangeFS
DISTRIBUTED DIRECTORIES

Removal with Variable Clients

- Maintain high throughput when saturated on 64 servers.



OrangeFS
DISTRIBUTED DIRECTORIES

Summary and Conclusion

- The distributed directory implementation shows great scalability in creating and removing large numbers of files by multiple clients concurrently.
- Can reach 8000+ file creations/second and 11000+ file removals/second with 64 servers on Palmetto Cluster.

OrangeFS
DISTRIBUTED DIRECTORIES

Future Work

- Finish implementation of splitting functionality and get ready for public release.
- Optimize performance, particularly using collective communication.
- Thoroughly evaluate performance of throughput, scalability, overheads and etc.

OrangeFS
DISTRIBUTED DIRECTORIES

Acknowledgement