TechBrief
AUGUST 2013

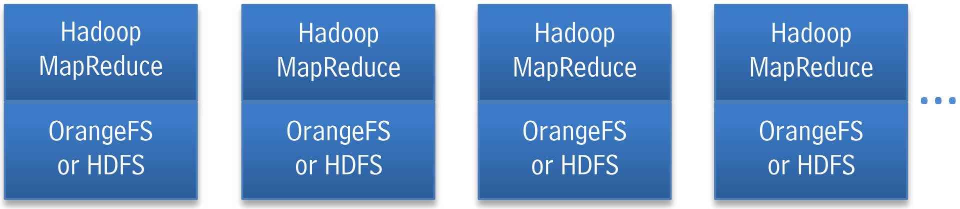
REPLACING HDFS WITH ORANGEFS

OrangeFS is the scalable software storage solution for High Performance Computing and Big Data.

OrangeFS can replace—and improve upon—HDFS as the file system for Hadoop MapReduce. This one-for-one "swap," which requires no code modifications to MapReduce, uses the abstracted file system class in MapReduce and a JNI shim to the OrangeFS client.

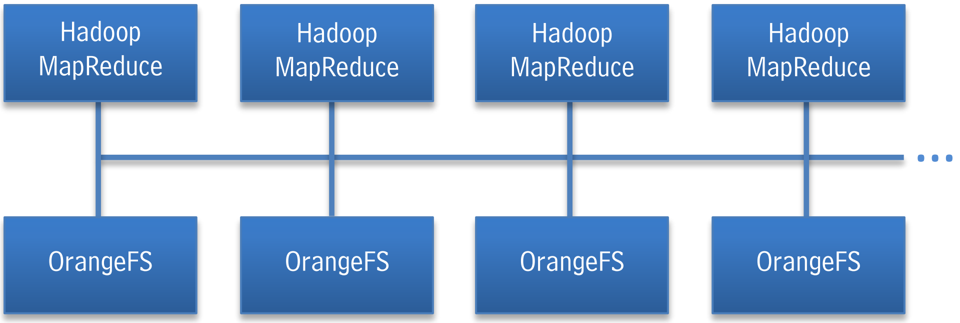
HADOOP MAPREDUCE WITH HDFS AND ORANGEFS

To test the ability of OrangeFS to replace HDFS, a full TB Terasort test was performed on 8 nodes, each running both MapReduce and the file system. The tests were performed as pictured here:

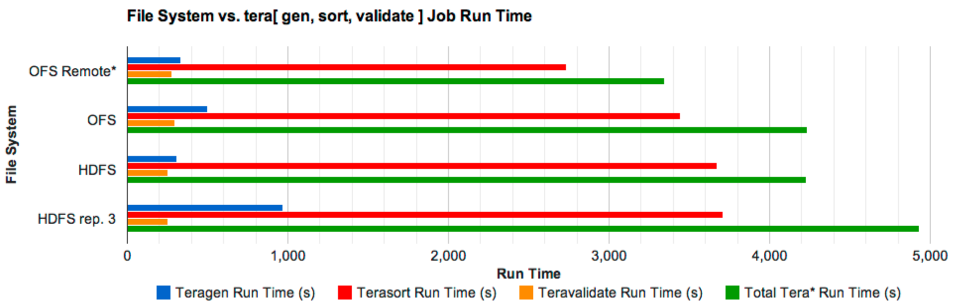


The tests were performed on 8 Dell PowerEdge R720s with Local SSDs for metadata and 12 2-TB drives for data. For the first test (above), MapReduce was run over OrangeFS and HDFS locally on the same nodes. The nodes were interconnected with 10Gb/s Ethernet.

An additional configuration, based on standard HPC configurations, was tested to see how MapReduce performs using dedicated OrangeFS servers. For this test we added an additional 8 nodes as MapReduce clients and used the same nodes from the previous tests as storage nodes only.



The Terasort test was performed in all of the configurations and the results follow:



MORE ABOUT ORANGEFS

SUPPORT

- Commercial-grade support
- Professional development team
- Automated testing on multiple disributions
- Diverse Client Access Methods: Posix, MPI, Linux VFS, FUSE (Including Mac OSX), Windows, WebDAV, Hadoop MapReduce and S3
- Kernel Module (supports standard distribution Kernels)
- Improved documentation

PERFORMANCE

- Orders of magnitude faster than NFS for MPI-IO workloads
- Excellent for large data-sets and large files
- Parallel Client for concurrent data access

HIGHLY SCALABLE STORAGE

- Supports standard Linux distributions (no custom Kernels needed)
- Supports high performing SSD for metadata
- Improved smaller-file performance
- Distributed metadata for directory entries (available version 2.9)
- Client-side caching via Direct Interface

STABILITY

- Enhanced client libraries for simplified application integration
- WebDAV support
- Capability based access control (available in version 2.9)

FUTURE PROOF ARCHITECTURE

- Open Source
- Modular object based design
- Supports multiple network architectures (Ethernet, IB, Portals, Myrinet)
- Stateless architecture
- Hardware independent

SIMPLE YET POWERFUL

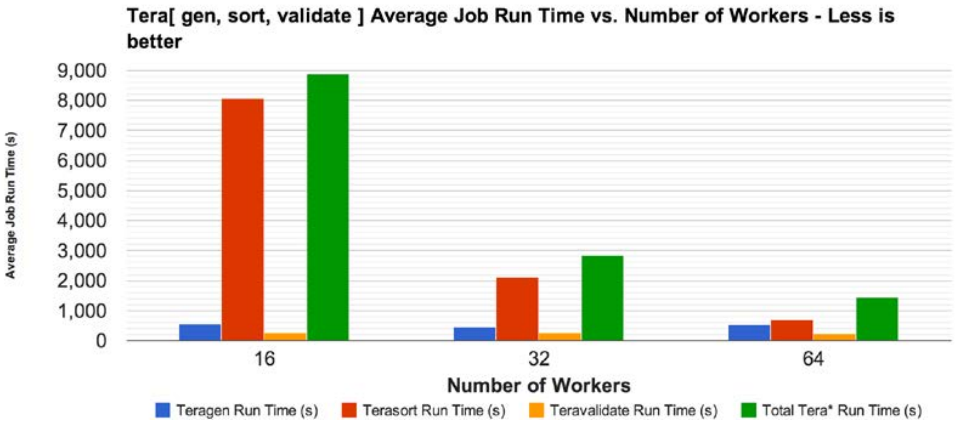
- User-level implementation
- Painless deployment

As noted in the chart on page 1:

- The overall measurement of the Terasort operation, the green (bottom) line for each group, showed HDFS with replication set to 3 performing slower than the others. OrangeFS and HDFS with no replication performed nearly the same (within 0.2%), and the MapReduce using dedicated OrangeFS servers performed 25% faster than the local configurations.
- Each stage of the operation is graphed, in order, as Teragen, Terasort, Teravalidate, followed by the complete operation time.
- These tests demonstrate that local MapReduce over OrangeFS, through the JNI shim, offers similar performance on the same hardware to HDFS, while still providing all the advantages of a general purpose, scale-out file system.
- An observed MapReduce client node had problems and died during one of the remote tests, and since the file system was separated from the MapReduce Clients, OrangeFS was unaffected, and the task was reassigned to another node. Client code is generally more susceptible to faults than hardened server code, so running in this separated mode can add potential stability.

HADOOP MAPREDUCE OVER REMOTE OVERCOMMITTED ORANGEFS

In this test we took the concept of MapReduce over OrangeFS and performed tests that would overcommit the storage nodes to see how well it scales with more MapReduce clients than storage nodes. We used the same Terasort test and performed it with an increasing number of clients. The servers were the same as the storage server nodes in the previous tests. The clients were similar, except they only had a single SAS drive for intermediate data, thus increasing the time for intermediate work over the previous tests. SSDs would help alleviate the slowdown caused by the single disk vs. an array of disks.



SUMMARY

Dedicated OrangeFS servers increase MapReduce performance by about 25% and provide good results when clients significantly overcommit the storage servers (4 to 1 in our tests). While providing these improvements as a good general purpose file system, OrangeFS can also provide an excellent concurrent working file system for other applications running MapReduce code.

MORE ABOUT ORANGEFS

IN ADDITION TO HADOOP, YOU CAN USE IT FOR…

- **HPC Parallel Jobs** OrangeFS interfaces with MPI, allowing HPC parallel jobs to use the OrangeFS system seamlessly.
- **Big-Data Aggregate Storage** For years, PVFS has excelled in handling extremely large data sets. OrangeFS is ideal for the growing demands of commercial big data, where the ability to capture, manage and process huge data sets within tolerable times has spawned an entirely new field in information management.
- **Guest Virtualization** Guest VMs can use this shared, high performance file system to achieve I/O similar to a dedicated system. Traditional NFS or CIFS file systems with this ability fall short of OrangeFS flexibility and performance.
- **Capacity and Throughput** OrangeFS handles heavy medium-to-large-file I/O capacity without the costs and difficulty of local, directly attached or other network file systems. Add flexibility when you need to use storage quickly or shuffle it among hosts or file systems. Capacity is simply a matter of scaling out—adding systems as growth dictates. More servers equal more throughput.

SUPPORTS MANY CLIENT INTERFACES…

- Kernel module (VFS); no kernel patches necessary
- Windows Client support
- MPI-IO interface
- Shared standard I/O library
- WebDAV and S3 support (Web Pack)
- FUSE support for Mac
- Direct Interface (bypasses Linux Kernel)
- Hadoop (via JNI)