BlobSeer in the context of MapReduce applications

Diana Moise

KerData team, INRIA

1. Hadoop
   - Hadoop Core
2. BlobSeer as storage for Hadoop
   - Integrating BlobSeer with Hadoop
   - Experimental evaluation
     - Microbenchmarks
     - Experiments with Map/Reduce Applications
3. Introducing support for append in Hadoop
   - Application case
4. Conclusions

Hadoop

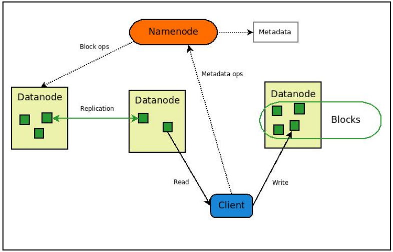
- ⋆ Yahoo!'s implementation of MapReduce
- ⋆ Open-source Java project
- ⋆ Large scale computation and data processing
- ⋆ Works on comodity hardware

Hadoop

- ⋆ Yahoo!'s implementation of MapReduce
- ⋆ Open-source Java project
- ⋆ Large scale computation and data processing
- ⋆ Works on comodity hardware

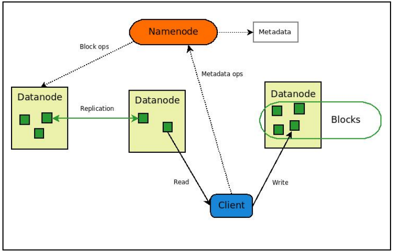
Hadoop Core

- Hadoop Distributed File System (HDFS)



Hadoop Core

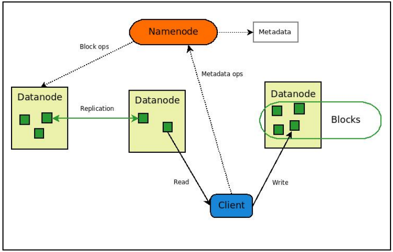
- Hadoop Distributed File System (HDFS)



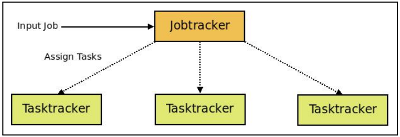
- Limitations
  1. one writer at a time
  2. no overwrites
  3. no appends

Hadoop Core

- Hadoop Distributed File System (HDFS)



- Hadoop MR framework



- Limitations
  1. one writer at a time
  2. no overwrites
  3. no appends

In-production use at...

YAHOO!
facebook
A9
amazon webservices
AOL
Bai du 百度
cloudera
IBM
last.fm
The New York Times
Ning
Adobe
Microsoft

Source: http://wiki.apache.org/hadoop/PoweredBy

Integrating BlobSeer with Hadoop

- Implementing the HDFS API for BlobSeer
  - ⋆ Exposes basic file system operations: create, read, write...
  - ⋆ Introduces support for concurrent append operations

Integrating BlobSeer with Hadoop

- Implementing the HDFS API for BlobSeer
  - ⋆ Exposes basic file system operations: create, read, write...
  - ⋆ Introduces support for concurrent append operations
- BlobSeer File System (BSFS)

Integrating BlobSeer with Hadoop

- Implementing the HDFS API for BlobSeer
  - ⋆ Exposes basic file system operations: create, read, write...
  - ⋆ Introduces support for concurrent append operations
- BlobSeer File System (BSFS)
  - ✓ File system namespace - keeps metadata, maps files to BLOBs

Integrating BlobSeer with Hadoop

- Implementing the HDFS API for BlobSeer
  - ⋆ Exposes basic file system operations: create, read, write...
  - ⋆ Introduces support for concurrent append operations
- BlobSeer File System (BSFS)
  - ✓ File system namespace - keeps metadata, maps files to BLOBs
  - ✓ Client-side buffering: data prefetching, write aggregation

Integrating BlobSeer with Hadoop

- Implementing the HDFS API for BlobSeer
  - ⋆ Exposes basic file system operations: create, read, write...
  - ⋆ Introduces support for concurrent append operations
- BlobSeer File System (BSFS)
  - ✓ File system namespace - keeps metadata, maps files to BLOBs
  - ✓ Client-side buffering: data prefetching, write aggregation
  - ✓ Exposes data layout to Hadoop, just like HDFS

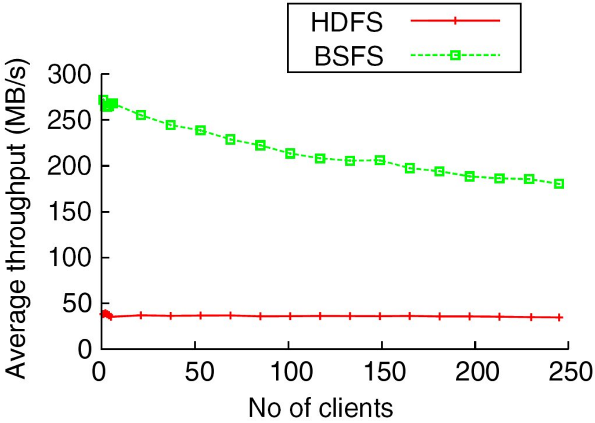
Testing and evaluation - overview and goals

- Goal
  - Measure the throughput of HDFS and BSFS
  - Evaluate the impact of replacing HDFS with BSFS
- Test scenarios
  - Microbenchmarks
    - Direct access to the file system
    - Common access patterns in Map/Reduce applications
  - Real Map/Reduce Applications
    - Distributed sort

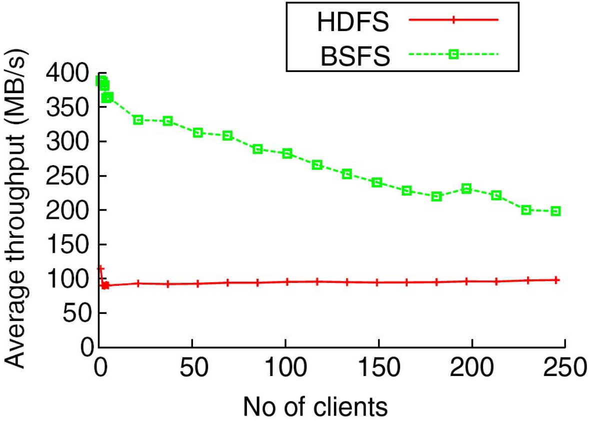
Setup

- 270 nodes from the same cluster on Grid'5000
- HDFS:
  - one namenode on a dedicated machine
  - one datanode on each cluster node
- BSFS:
  - one vmanager, one pmanager, one namespace manager
  - 20 metadata providers
  - providers on the rest of the nodes

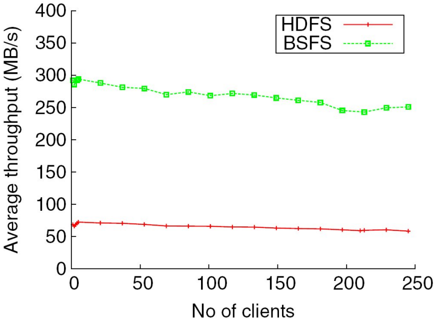
Concurrent clients writing to different files



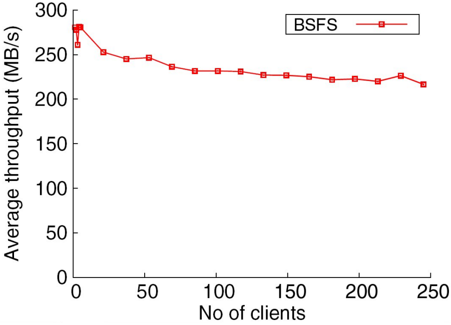
Concurrent clients reading from different files



Concurrent clients reading parts from the same file

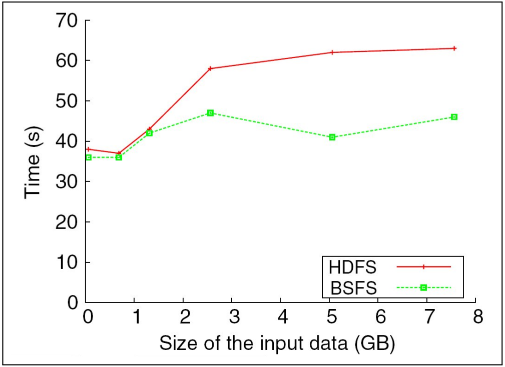


Concurrent clients appending data to the same file



Distributed sort

- Sorts key-value pairs
- Both read and write intensive

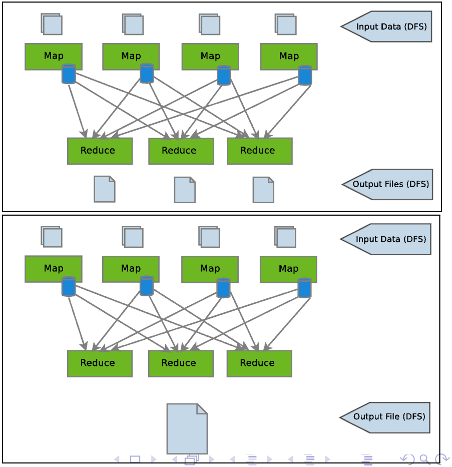


Modifying Hadoop to use appends

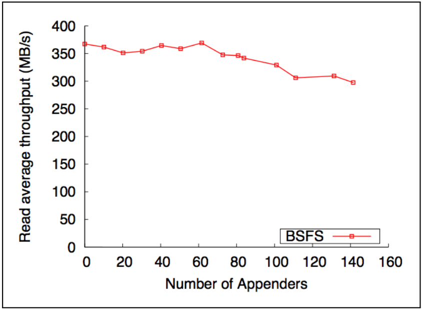
- Append implemented at the file system level

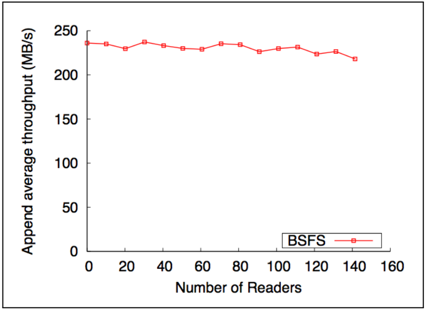
Modifying Hadoop to use appends

- Append implemented at the file system level
- Modify reducer code in Hadoop to append the output to a single file



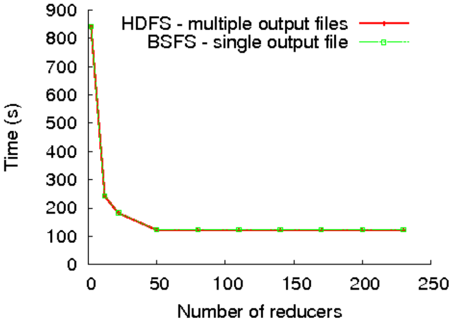
Concurrent reads and appends to the same file





Data join - Results

- Similar to outer join from the database context
- Merge two input files based on common keys
- 6.3 GB of output



Conclusions

- BSFS improves Hadoop's throughput
- Support for append
- Work in progress
  - ⋆ Intermediate data management
    - Store map output to BSFS
    - Resume computation in case of failures
  - ⋆ Pipeline MapReduce applications
    - Schedule mappers as soon as splits are produced
    - Application study: Pig

Thank you!