INRIA – UPB ASSOCIATED TEAM

XtreemFS vs HadoopFS evaluation in the context of map reduce applications

Eliana Tirsa

Agenda

- Motivation
- Hadoop + XtreemFS
- Hadoop + HDFS/XtreemFS
- Testing XtreemFS vs HDFS
- XtreemFS/HDFS Randomtextgenerator
- XtreemFS/HDFS Distributed Grep
- XtreemFS/HDFS Distributed Sort
- BlobSeer/HDFS Distributed Sort
- Conclusions

Motivation

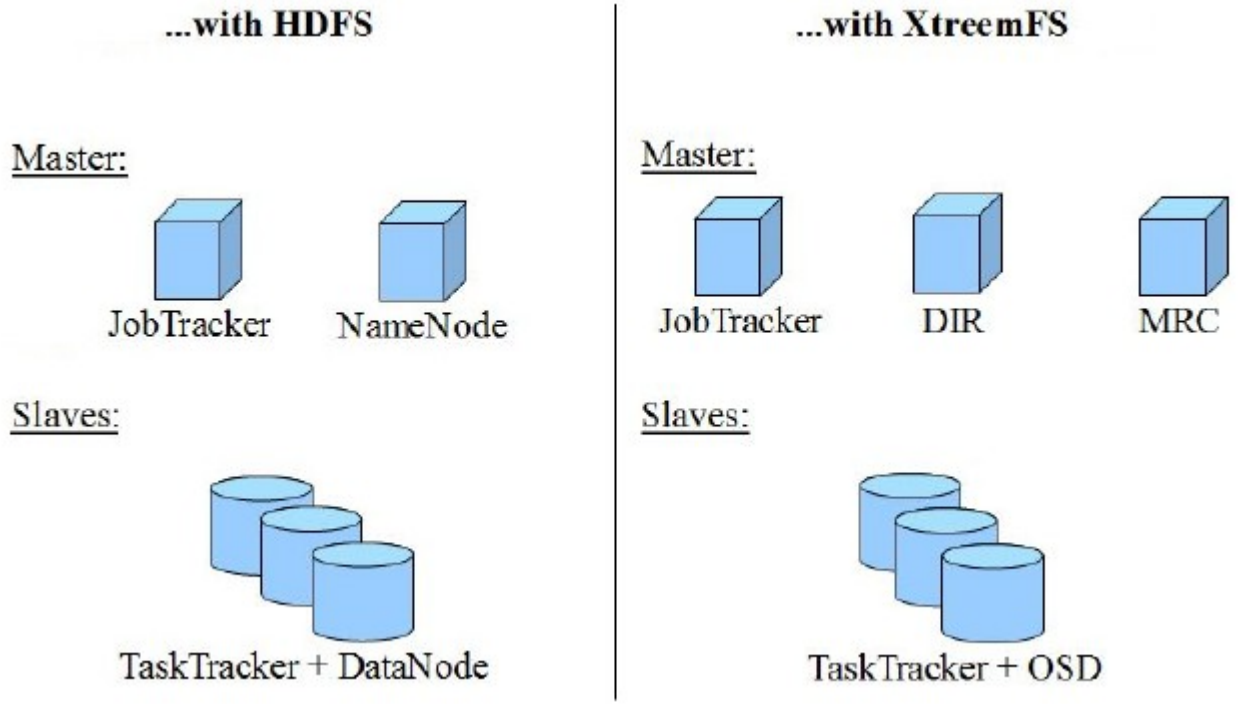
- Asociated team topic
  - XtreemOS + BlobSeer + Nimbus
- XtreemOS + BlobSeer = ?
  - Checkpoint-restart applicatios
  - Map-reduce applicatios
  - VM storage (in Nimbus or XtreemOS as a IAAS)
  - File storage
- Map-reduce in XtreemOS
  - Hadoop + XtreemFS (experimental driver)
  - XOS SAGA
  - Hadoop + BlobSeer

Hadoop + XtreemFS

- Experimental driver
  - Initially, all that worked was:
    - bin/hadoop dfs -ls
  - Resolved bugs
    - File creation
      - Incorrect return type of some methods
      - Path creation (-touchz /path/file)
  - Implemented methods
    - In order to obtain basic functionality

Hadoop + HDFS/XtreemFS

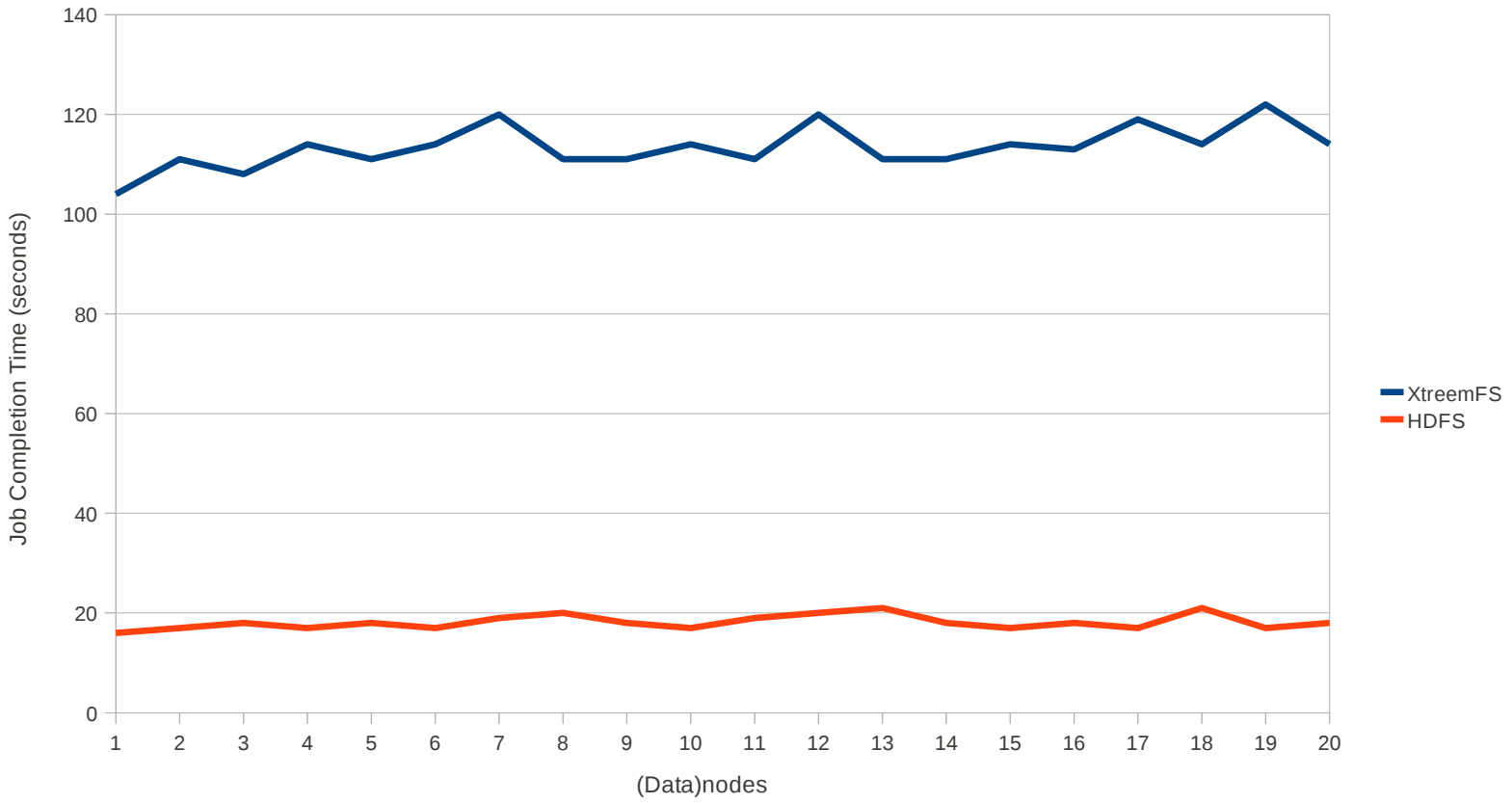
- Hadoop ...



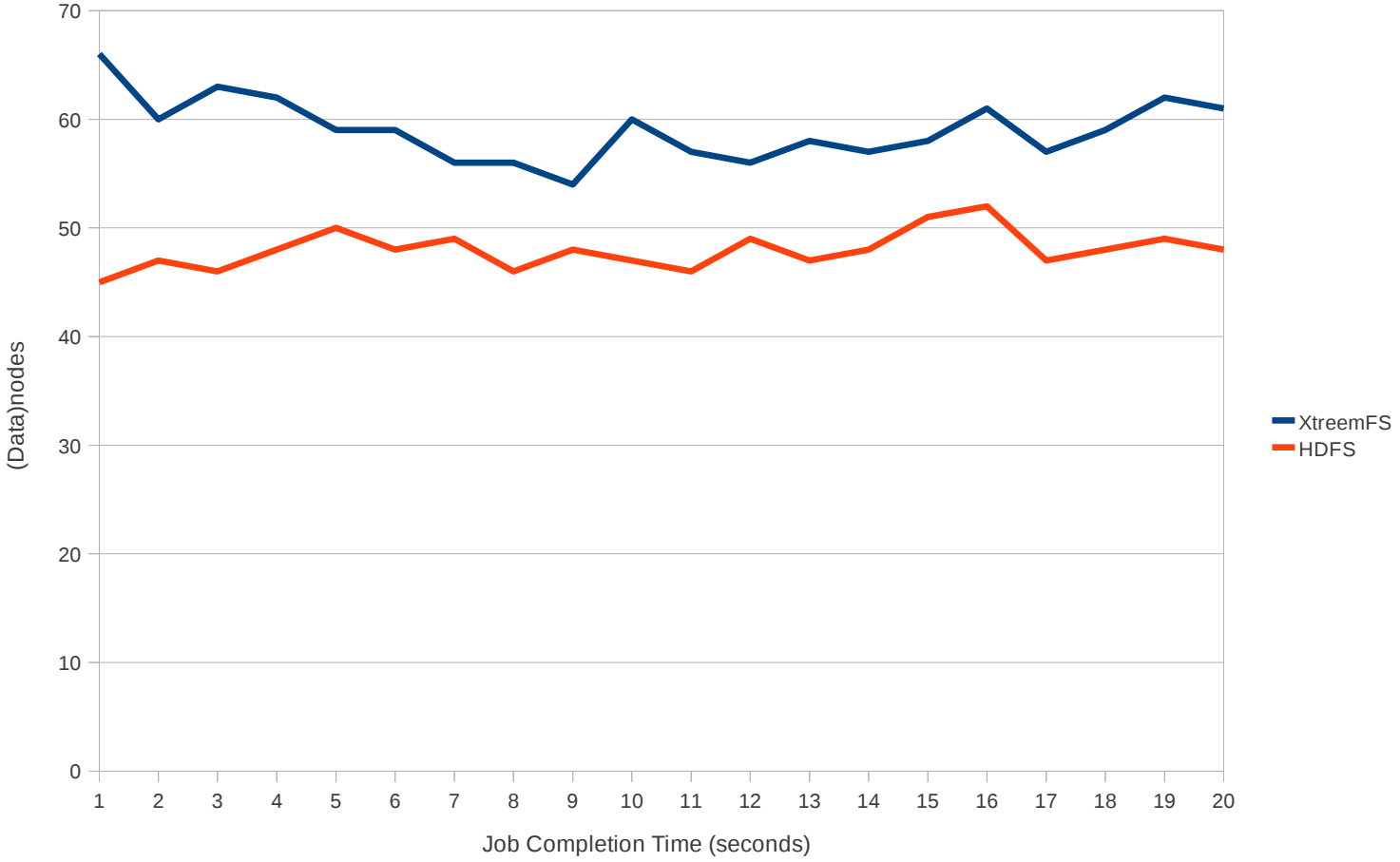
Testing XtreemFS vs HDFS

- Completion time of map reduce applications running on hadoop
  - Randomtextgenerator
  - Distributed grep
  - Distributed sort
- N clients
- 1 mapper/client/node
- Processed data: N * 64MB (64MB/client)
- 1 <= N <=20

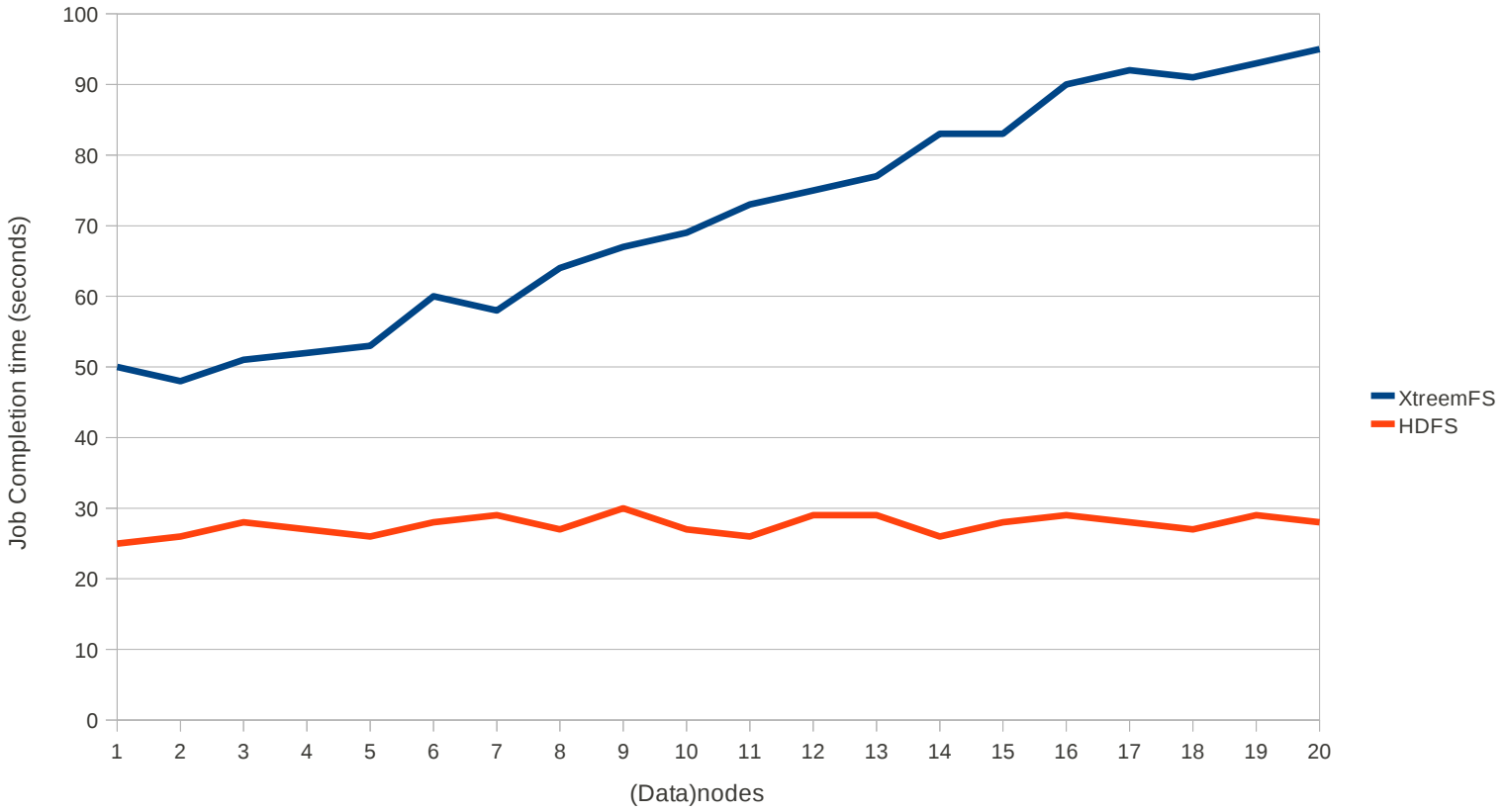
XtreemFS/HDFS Randomtextgenerator



XFS/HDFS Distributed Grep

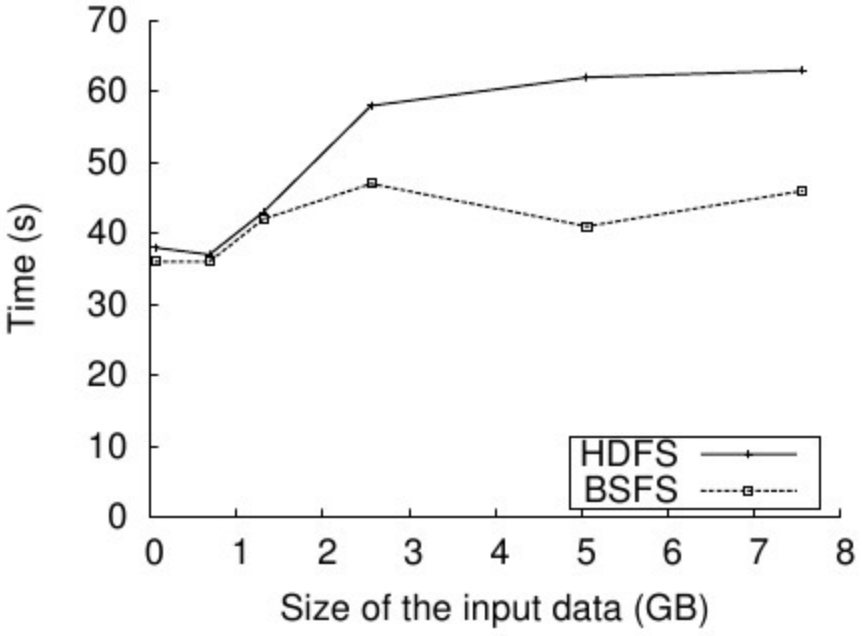


XFS/HDFS Distibuted Sort



BlobSeerFS/HDFS Distributed Sort Completion Time

*from "*BlobSeer: Next Generation Data Management for Large Scale Infrastructures*", B. Nicolae et al, 2010



Conclusions

- With regard to job completion time, HDFS outperforms XtreemFS
- Indirectly, Blobseer outperforms XtreemFS in this context

- Future direction:
  - More (bigger) tests
  - Comparative evaluation using Microbenchmarks
  - BlobSeer vs XtreemFS on XtreemFS terms (data storage over WAN)

- Thank You !