Hadoop on HTC

Daniel Traynor, QMUL
GRidPP 40 10/04/2018

Overview

- This talk is about deploying Hadoop and integrating within a HTC cluster. I try and extend to how this might be “gridified”. Technical details available offline.
- NOT about how Hadoop/ hdfs / mapreduce / etc… work.
- NOT how to develop Hadoop applications or where they might be used in HEP/astro.

Hadoop

- Hadoop has become a fixture in “Big data” analysis.
- Increasing number “solutions” assume Hadoop environment: mapreduce, Hive, Pig, Spark, HDBase, etc…
- Need to build dedicated clusters for Hadoop. Consisting of compute servers with several tens of TB of storage each. e.g. CERN has several such Hadoop clusters.
- Hadoop not secure unless integrated with kerberos.

SPARK

- Spark, an alternative to MapReduce, is effectively a Swiss Army Knife for data processing in Hadoop, delivering SQL access, streaming data processing, graph and NoSQL processing, machine learning, Graph applications, Pattern recognitionand much more right out of the box.
- Scala, Python, Java, SQL, R.
- Can be used with or without YARN to process data in HDFS or other filesystems. 3rd party integration with HPC (slurm) and Jupyter notebooks.
- Can be used Interactively or in batch.
- IO: Local or network filesystems, object storage such as Amazon S3 or Ceph, relational database systems and NoSQL stores, messaging systems, etc…
- Spark is widely expected to ultimately supplant MapReduce and the other SQL-on-Hadoop engines as the predominant processing framework for the Hadoop platform.

	Hadoop	HTC
Scheduling	YARN	HTCondor, SGE, SLURM
Storage	HDFS, direct attached, immutable.	dpm, dcache, ceph, Lustre. Network attached
compute	compute moved to the data.	data moved to the compute
interconnect	Ethernet	Ethernet
Services	several dedicated daemons required	system services + scheduler

Hadoop on HTC

- Hadoop has become a fixture in “Big data” analysis but is difficult to use in HPC/HTC cluster environments. Prefer not to run multiple clusters.

- Several attempts to integrate the two have been made.

  - Most common approach is to create a user space Hadoop cluster on the fly using a network file systems with in a multi node/core batch job. These clusters are deleted one the batch job is finished.

  - There also exists a “hybrid” approach to replace HDFS with Lustre/GPFS/other plugin and replace YARN with HTC scheduler (SLURM/PBS/other).

Method

- Run benchmarks on the following setups.

1. Set up a standard dedicated persistent Hadoop cluster using local disks for HDFS.

2. Setup a Hadoop cluster in userspace. i.e. Create a Hadoop cluster on demand with HDFS over the network (using Lustre scratch space) from a batch job.

3. (Attempt to) Set up a “hybrid” cluster. i.e. Replace HDFS with Lustre plugin, Replace Yarn with SLURM plugin.

Setup

3 HPE DL60, E5 2650 V3, 128 GB RAM. 10 Gb/s ethernet.
1 name node, 2 data nodes. 3 x 1TB disks on data node.
Hadoop 2.8, no replication.

Tests

TestDFSIO: Read and write benchmarks for HDFS (IOtest)

yarn jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.8.3-tests.jar TestDFSIO -D test.build.data=mytestdfsio -write -nrFiles 1000 -fileSize 1000

yarn jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.8.3-tests.jar TestDFSIO -D test.build.data=mytestdfsio -read -nrFiles 1000 -fileSize 1000

TeraSort: Sort data as fast as possible (classic hadoop test)

yarn jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.3.jar teragen -Dmapreduce.jobs.maps=1000 100000000 random-data

yarn jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.3.jar terasort random-data sorted-data

yarn jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.3.jar teravalidate sorted-data report

Calculate PI (CPU intensive)

yarn jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.3.jar pi 32 100000

Results

	DAS	NAS	Lustre
TestDFSIO Write	51m	50m	
TestDFSIO Read	83m	128m	
Teragen	1m5s	1m11s	
Terasort	4m12s	4m5s	
Pi	29s	27s	

DDSIO - 1TB, tests disk IO

terasort/gen (10G) data in memory

PI calculation cpu dominated

Conclusions

- Dedicated clusters easy to build but issues of resource usage.
- A usrespace Hadoop cluster on demand a doable solution on our HTC clusters could work via the grid.
- Hybrid solutions not yet working, issues about longtwerm support
- How to get data into and out of a Hadoop cluster? Integration with existing data stores.
- We tried Hadoop on HTC. What about HTC on Hadoop?

Links

- http://iopscience.iop.org/article/10.1088/1742-6596/898/7/072034/pdf
- https://www.nextplatform.com/2015/07/07/bringing-hpc-and-hadoop-under-the-same-cluster-umbrella/. https://arxiv.org/ftp/arxiv/papers/1506/1506.08907.pdf
- https://events.static.linuxfound.org/sites/events/files/slides/ApacheCon2015-Devaraj-Kavali.pdf. https://github.com/intel-hpdd/scheduling-connector-for-hadoop. https://github.com/intel-hpdd/lustre-connector-for-hadoop.
- https://github.com/LLNL/magpie
- https://arxiv.org/pdf/1602.00345.pdf