Monitoring I/O on Data-Intensive Clusters

Visualizing Disk Reads and Writes on Hadoop MapReduce Jobs

Thursday, July 31

Joel Ornstein  Joshua Long  Carson Wiens

Mentors: Steve Senator, Tim Randles, Vaughan Clinton, Mike Mason, Graham Van Heule – HPC 3
Background

Motivation:
   – I/O Intensive Jobs
     • Large amounts of scientific data
Background

Motivation:
  – I/O Intensive Jobs
    • Large amounts of scientific data

Traditional HPC
  – Limiting factor mostly lies in processing speed
Motivation:
- I/O Intensive Jobs
  - Large amounts of scientific data

Traditional HPC
- Limiting factor mostly lies in processing speed

I/O Intensive Jobs
- Bottlenecked by read/write disk speed
- MapReduce
  - Move jobs to the data (instead of vice-versa)
MapReduce

MAP

JOB

MAP

ANS 1

REDUCE

ANS 2

RESULT

ANS 3
I/O Monitoring

Why?
- Nodes break
- Jobs run without using the specified resources
I/O Monitoring

Why?
- Nodes break
- Jobs run without using the specified resources

Deliverables
- Programs that are helpful for monitoring a Hadoop 2.3 cluster
  - Splunk App for HadoopOps
  - Ganglia
  - Other methods
I/O Monitoring

Why?
  – Nodes break
  – Jobs run without using the specified resources

Deliverables
  – Programs that are helpful for monitoring a Hadoop 2.3 cluster
    • Splunk App for HadoopOps
    • Ganglia
    • Other methods
  – Data tests
    • bonnie++
    • teragen and terasort
Environment

- 11-node CentOS cluster
  - 1 head node and 10 compute nodes

- FDR InfiniBand 56-Gb/second
  - IP over IB
  - Faster than disks can read/write

- Hadoop 2.3.0

- MRv2/YARN
  - Yet Another Resource Negotiator
  - Runs MapReduce jobs in Hadoop environment

- Java 1.6
Monitoring Tools

Splunk
- software for searching and analyzing logs
- able to generate graphs, charts, gauges, etc.
- web interface
Monitoring Tools

Splunk
- software for searching and analyzing logs
- able to generate graphs, charts, gauges, etc.
- web interface

Ganglia
- software for monitoring clusters
- generates plots from input
- web interface
Monitoring Tools

Splunk
- software for searching and analyzing logs
- able to generate graphs, charts, gauges, etc.
- web interface

Ganglia
- software for monitoring clusters
- generates plots from input
- web interface

iostat
- outputs I/O statistics for devices
- command-line interface
Splunk App for HadoopOps

Components

- Node: hdfs
- DataNode: 0 / 10

Utilization

- Load Avg: 0.34
- Used Mem: 6%
- HDFS Used: 0%

Headlines

- Nothing news-worthy at the moment...

Activities

- Jobs: Pending: 0, Running: 0, Complete: 0, Failed: 0, Killed: 0, Total: 0
- Tasks: Total: 0, Succeeded: 0, Failed: 0
- M/R: Map Tasks: -, Reduce Tasks: -
Ganglia
## iostat

```
iostat –kxy 1 2
```

```
Joshua@goldilocks ~ joshua@goldilocks:-- ssh -- 106x25

joshua@goldilocks ~> iostat -kxy 1 2
Joshua@goldilocks ~> bash...
Joshua@goldilocks ~> joshua@goldilocks ~> joshua@goldilocks ~> joshua@goldilocks ~> joshua@goldilocks ~

<table>
<thead>
<tr>
<th>Device</th>
<th>rmm/s</th>
<th>wrqm/s</th>
<th>r/s</th>
<th>w/s</th>
<th>rkb/s</th>
<th>wkb/s</th>
<th>avgrq-sz</th>
<th>avgqu-sz</th>
<th>await</th>
<th>svctm</th>
<th>%util</th>
</tr>
</thead>
<tbody>
<tr>
<td>sda</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>scb</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

```

```
### iostat

**iostat**

**iostat –kxy 1 2**

<table>
<thead>
<tr>
<th>Device</th>
<th>rrqm/s</th>
<th>wrqm/s</th>
<th>r/s</th>
<th>w/s</th>
<th>rkB/s</th>
<th>wkB/s</th>
<th>avgq-sz</th>
<th>avgqu-sz</th>
<th>await</th>
<th>svctm</th>
<th>%util</th>
</tr>
</thead>
<tbody>
<tr>
<td>sda</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>scb</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**kB read per second**
```
iostat

iostat –kxy 1 2
```

<table>
<thead>
<tr>
<th>Device</th>
<th>rmm/s</th>
<th>wrmm/s</th>
<th>r/s</th>
<th>w/s</th>
<th>rkb/s</th>
<th>wkb/s</th>
<th>avgrq-sz</th>
<th>avgqu-sz</th>
<th>await</th>
<th>svctm</th>
<th>%util</th>
</tr>
</thead>
<tbody>
<tr>
<td>sda</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>scb</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>dm-2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

kB read per second
kB written per second
Methods

Benchmarking
  – bonnie++
  – measure disk I/O

Hadoop jobs
  – teragen
  – terasort

Hadoop jobs with remote data
Methods

Local Data | Remote Data

Virtual Remote Data
Results

Bonnie++

kB/s

Time (s)

Reads  Writes
Results

Teragen Local
Results

Teragen Remote (InfiniBand)

Kb/s vs Time (s)

- Reads
- Writes
Results

Teragen Remote (1 GigE)

kB/s

Time (S)

Reads  Writes
Results

Teragen Remote (1 GigE)

kB/s

Time (S)

Local

InfiniBand (remote)

Reads

Writes
Results
Results

![Terasort Local Graph](image)

- **x-axis**: Time (s)
- **y-axis**: kB/s

The graph shows the performance of Terasort Local over time, with separate lines for Reads and Writes.
Results

Terasort Remote (InfiniBand)

kB/s

Time (s)

Reads  Writes
Conclusion

Splunk
- Splunk app for HadoopOps is not suited to Hadoop MPv2/YARN

Ganglia
- Easy to configure and to extend

Effects of network latency
- Large impact when low connectivity
- Small, but noticeable impact for reasonable connectivity
Take-Aways and Successes

Monitoring I/O is easy (with the right tools)
  – Successfully set up ganglia to monitor I/O
  – Created visuals of I/O during Hadoop jobs

Benchmark of Hadoop jobs on local data and on remote data
  – Performance suffers on data intensive jobs when data is stored remotely
Future Work

Write I/O monitoring application for Splunk

Evaluate effects of network latency with varying Hadoop parameters
  – HDFS block size

Evaluating effects of network parameters
  – Maximum transmission unit

Comparing performance on NFS to other file systems

Further examining trends in graphs
Questions?
/*Comments*/