Monitoring I/O on Data-Intensive Clusters

Visualizing Disk Reads and Writes on Hadoop MapReduce Jobs

Thursday, July 31

Joel Ornstein — Harvey Mudd College

Joshua Long — Georgia Tech

Carson Wiens — Calvin College

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

Background

Motivation:

- I/O Intensive Jobs
  - Large amounts of scientific data

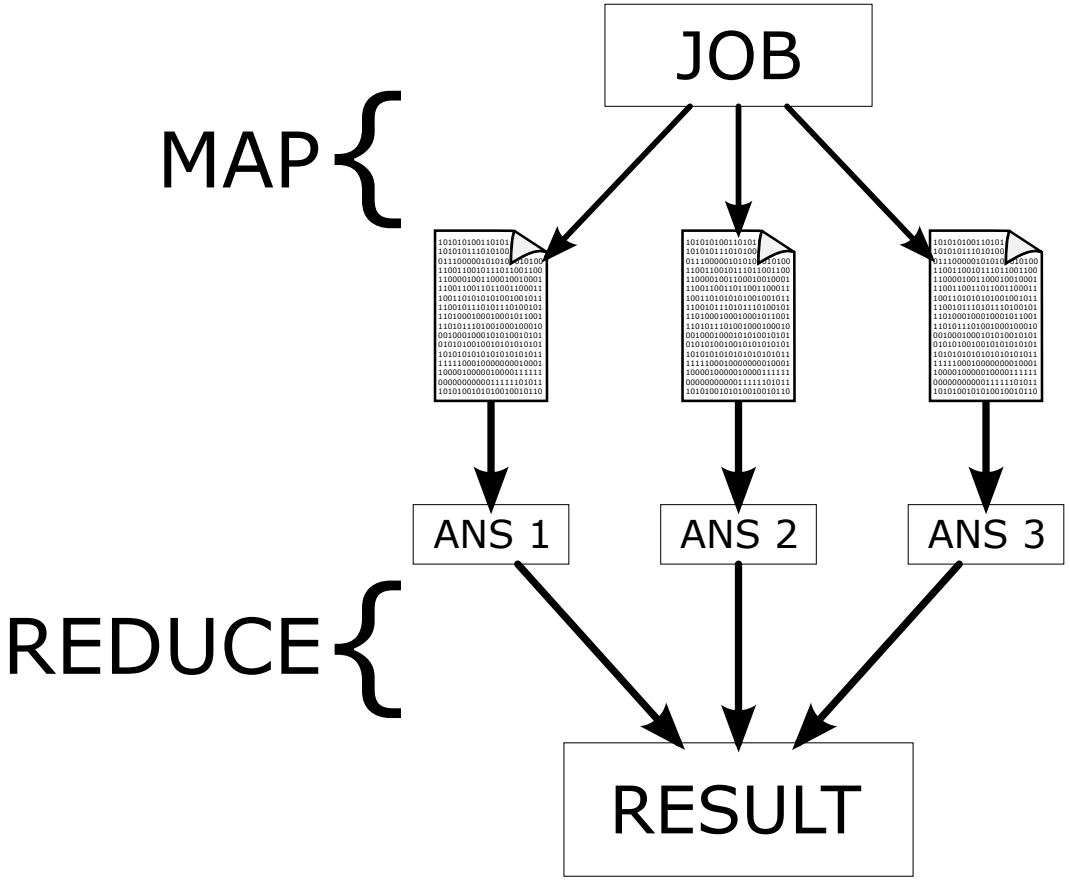
Traditional HPC

- 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



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

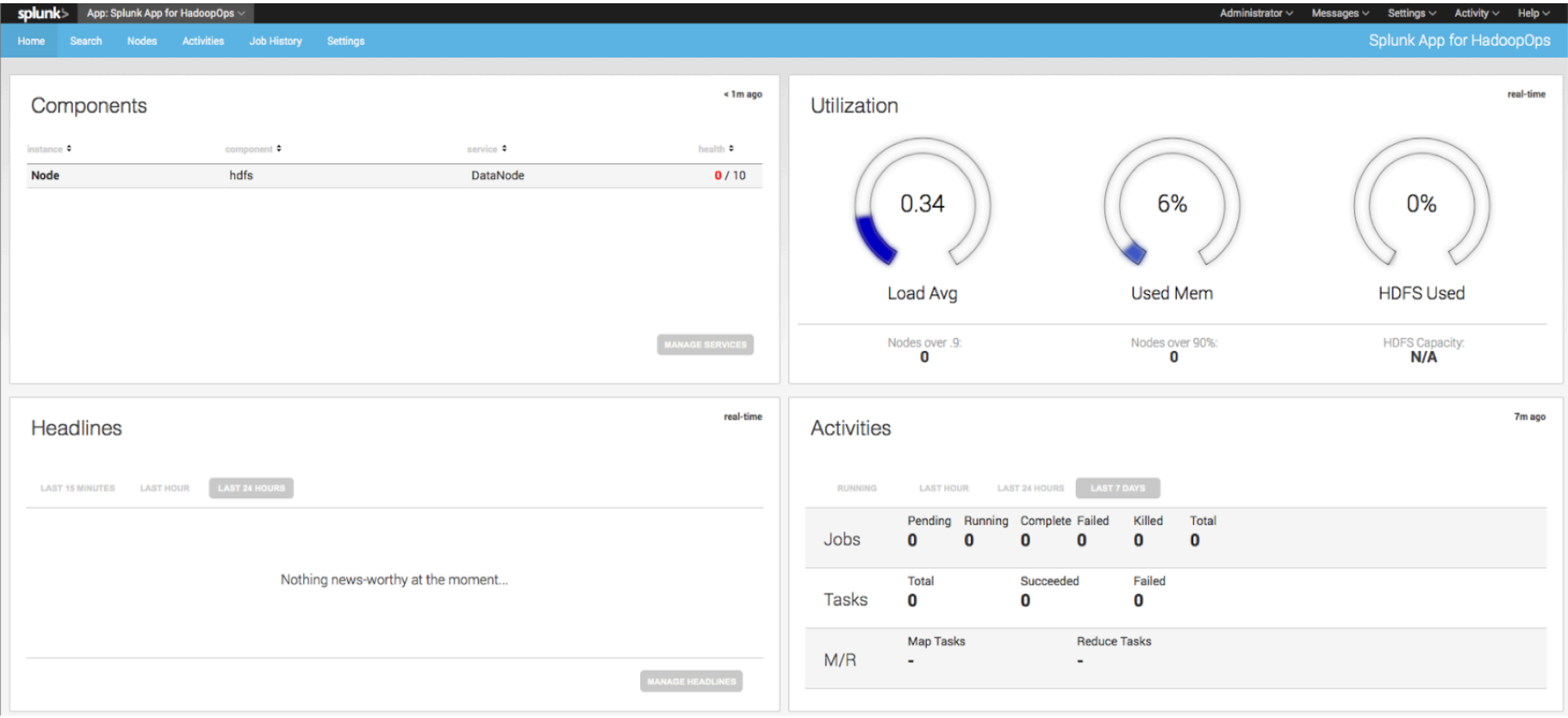
Ganglia

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

iostat

- outputs I/O statistics for devices
- command-line interface

Splunk App for HadoopOps

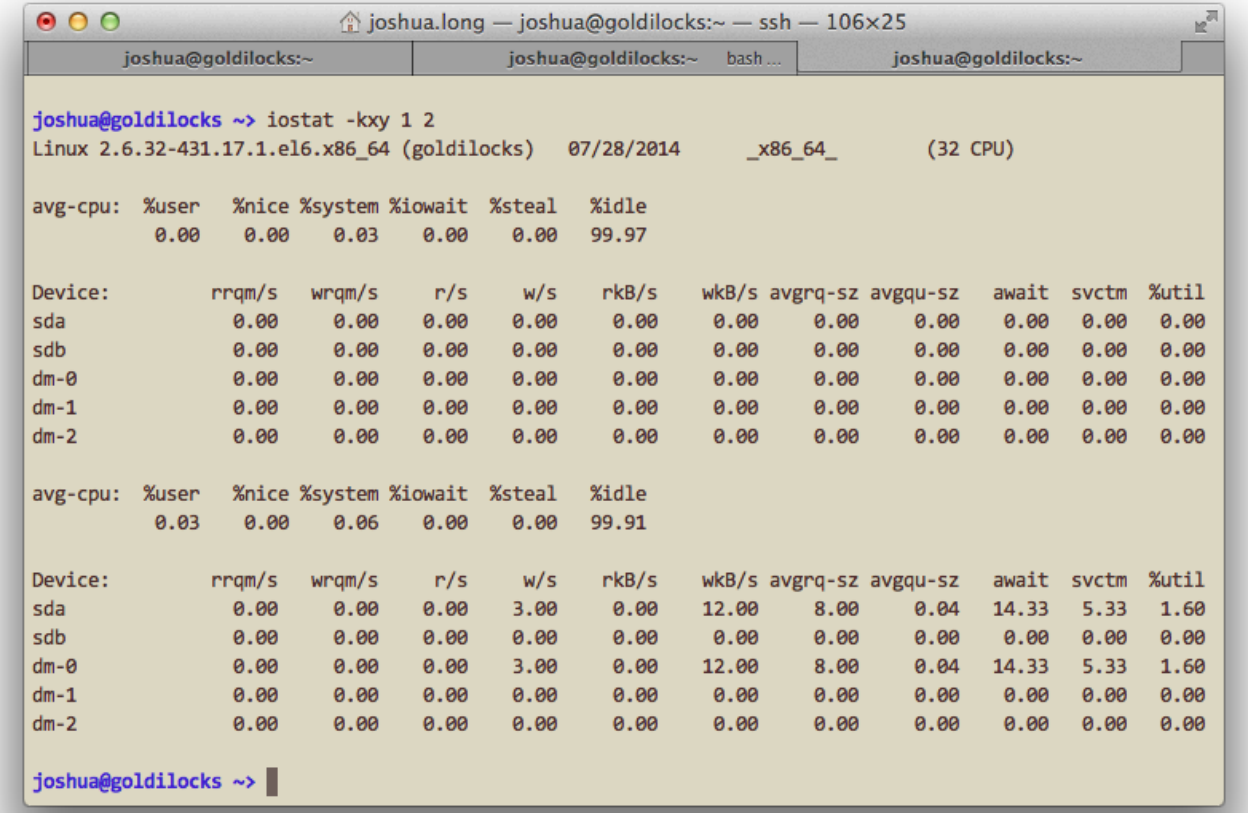


Ganglia


Ganglia
.sourceforge.net
Goldenrod Cluster Cluster Report for Tue, 22 Jul 2014 09:34:31 -0600
Get Fresh Data
Metric load_one
Last hour
Sorted descending
Physical View
Grid > Goldenrod Cluster > --Choose a Node
Overview of Goldenrod Cluster
CPUs Total: 272
Hosts up: 11
Hosts down: 0
Avg Load (15, 5, 1m):
26%,34%,31%
Localtime:
2014-07-22 09:34
Cluster Load Percentages
25-50 (72.73%)
0-25 (27.27%)
Goldenrod Cluster Cluster Load last hour
Load/Procs
1-min Load
Nodes
CPUs
Running Processes
Goldenrod Cluster Cluster CPU last hour
Percent
User CPU
Nice CPU
System CPU
WAIT CPU
Idle CPU
Goldenrod Cluster Cluster Memory last hour
Bytes
Memory Used
Memory Shared
Memory Cached
Memory Buffered
Memory Swapped
Total In-Core Memory
Goldenrod Cluster Cluster Network last hour
Bytes/sec
In
Out
Show Hosts: yes no | Goldenrod Cluster load_one last hour sorted descending | Columns 4 Size small
n06.cluster load_one last hour (now 10.33)
n05.cluster load_one last hour (now 9.82)
n03.cluster load_one last hour (now 9.68)
n02.cluster load_one last hour (now 8.84)
n09.cluster load_one last hour (now 8.36)
n04.cluster load_one last hour (now 8.36)
goldilocks load_one last hour (now 10.65)
n08.cluster load_one last hour (now 7.53)
n10.cluster load_one last hour (now 5.19)
n07.cluster load_one last hour (now 3.78)
n01.cluster load_one last hour (now 2.95)
(Nodes colored by 1-minute load) | Legend
Ganglia Web Frontend version 3.1.7 Check for Updates.
Ganglia Web Backend (gmetad) version 3.1.7 Check for Updates.
Downloading and parsing ganglia's XML tree took 0.0262s.
Images created with RRDTool version 1.3.8.
Pages generated using TemplatePower version 3.0.1.

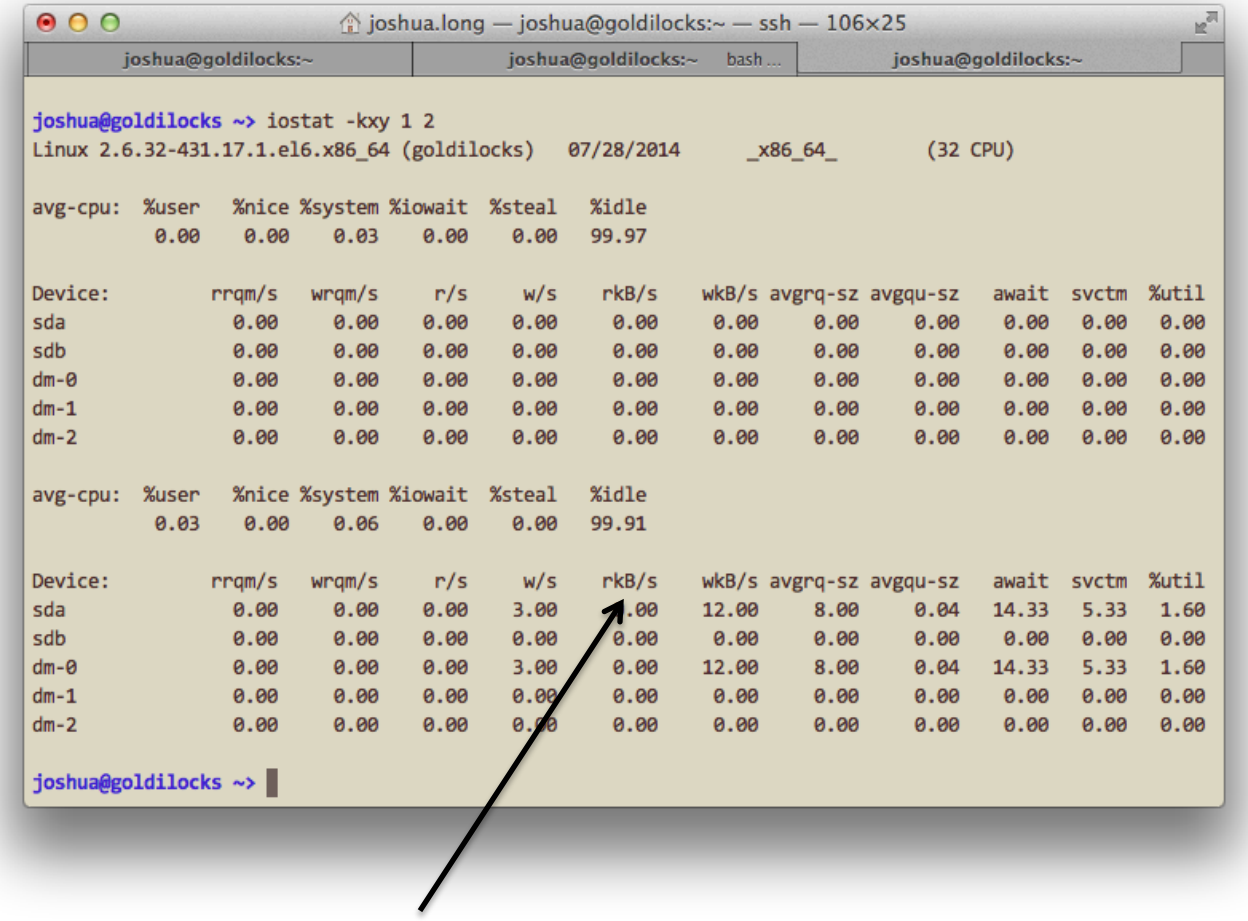
iostat

iostat –kxy 1 2



iostat

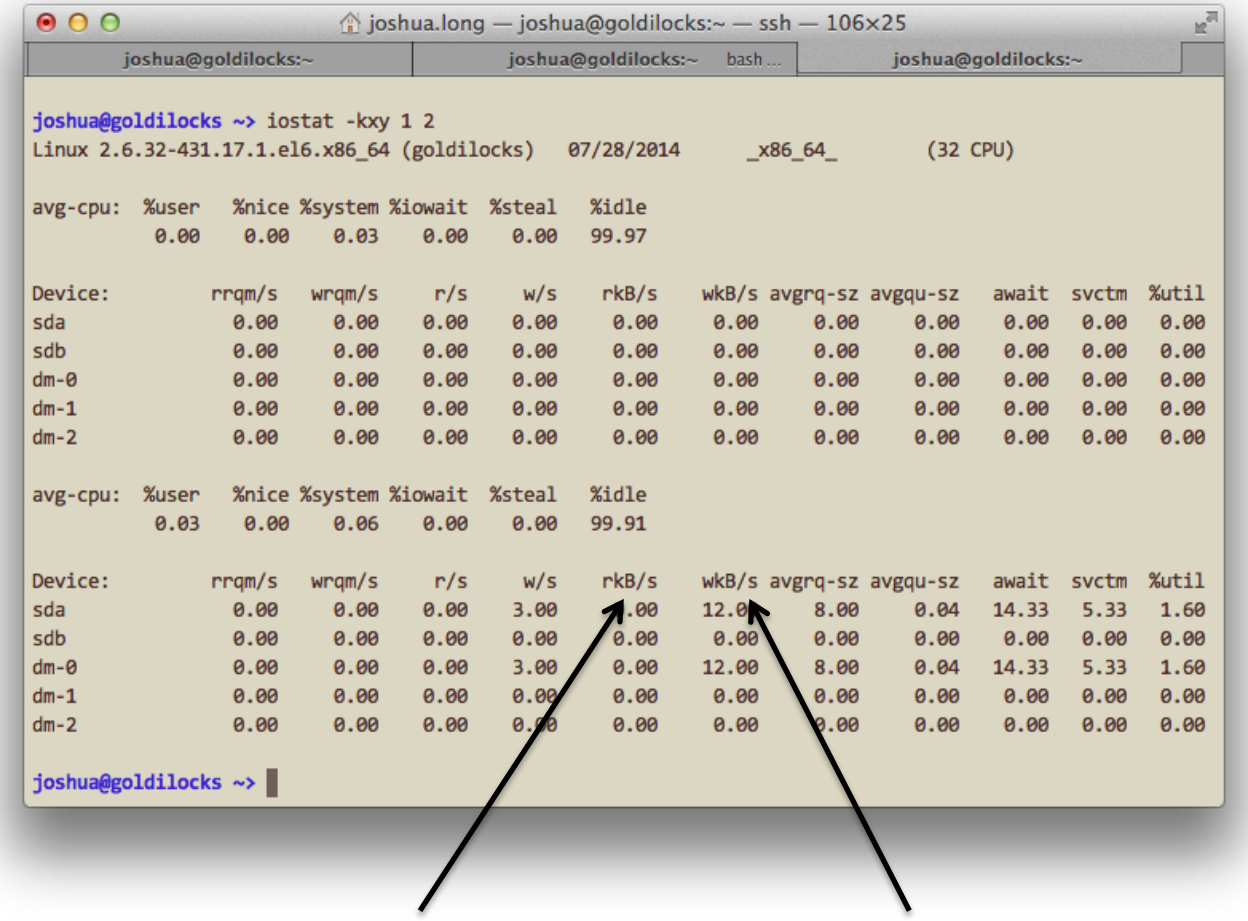
iostat –kxy 1 2



kB read per second

iostat

iostat –kxy 1 2



kB read per second

kB written per second

Methods

Benchmarking

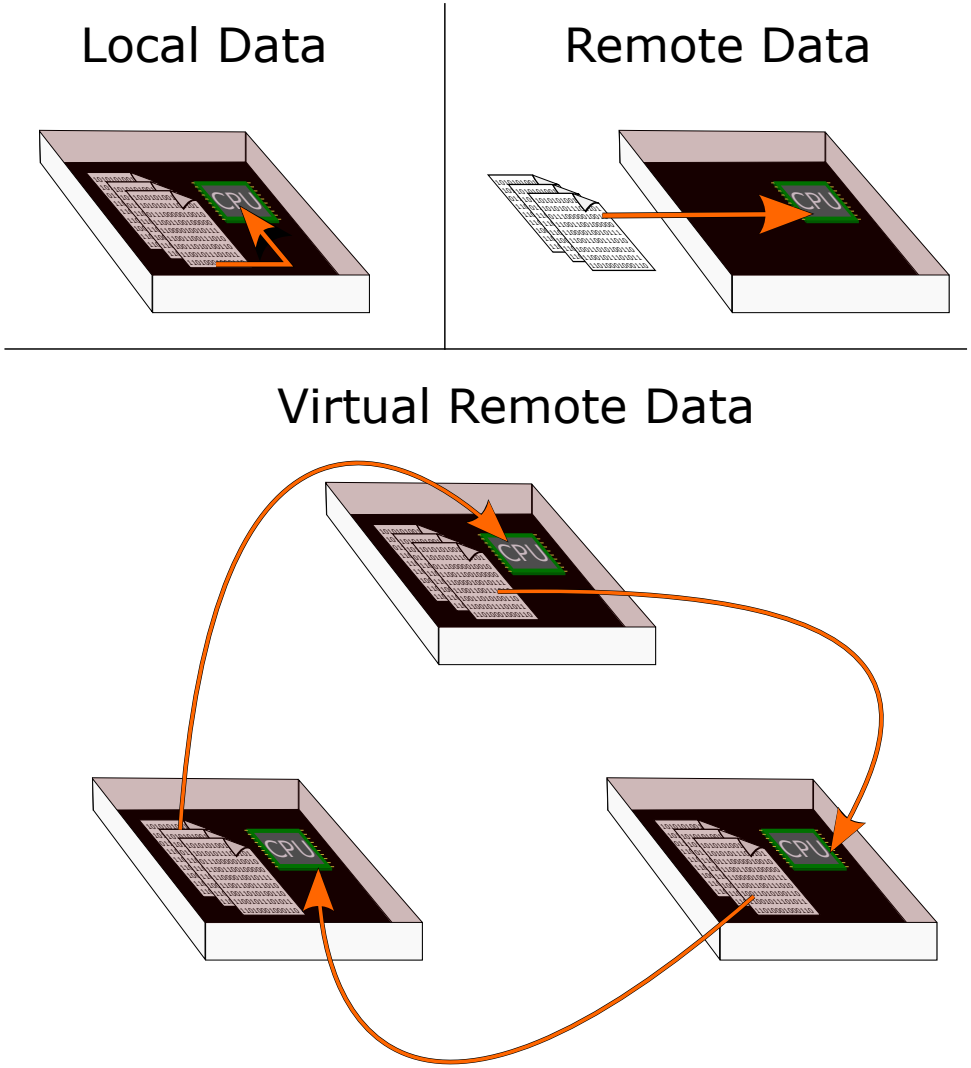
- bonnie++
- measure disk I/O

Hadoop jobs

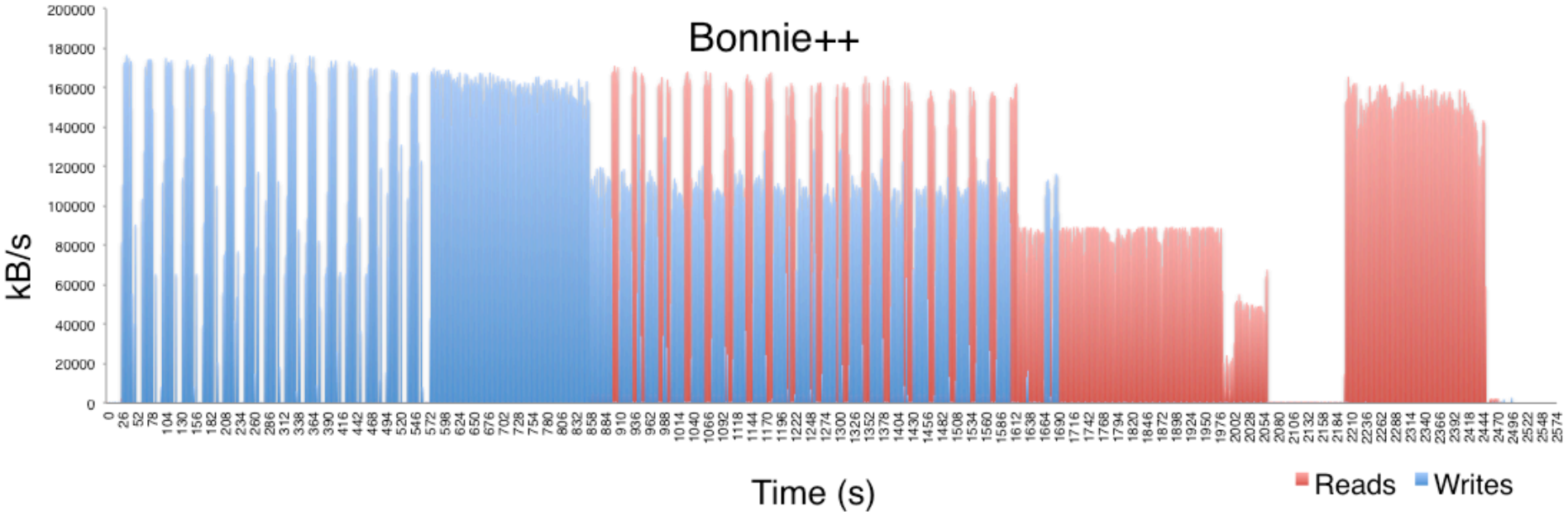
- teragen
- terasort

Hadoop jobs with remote data

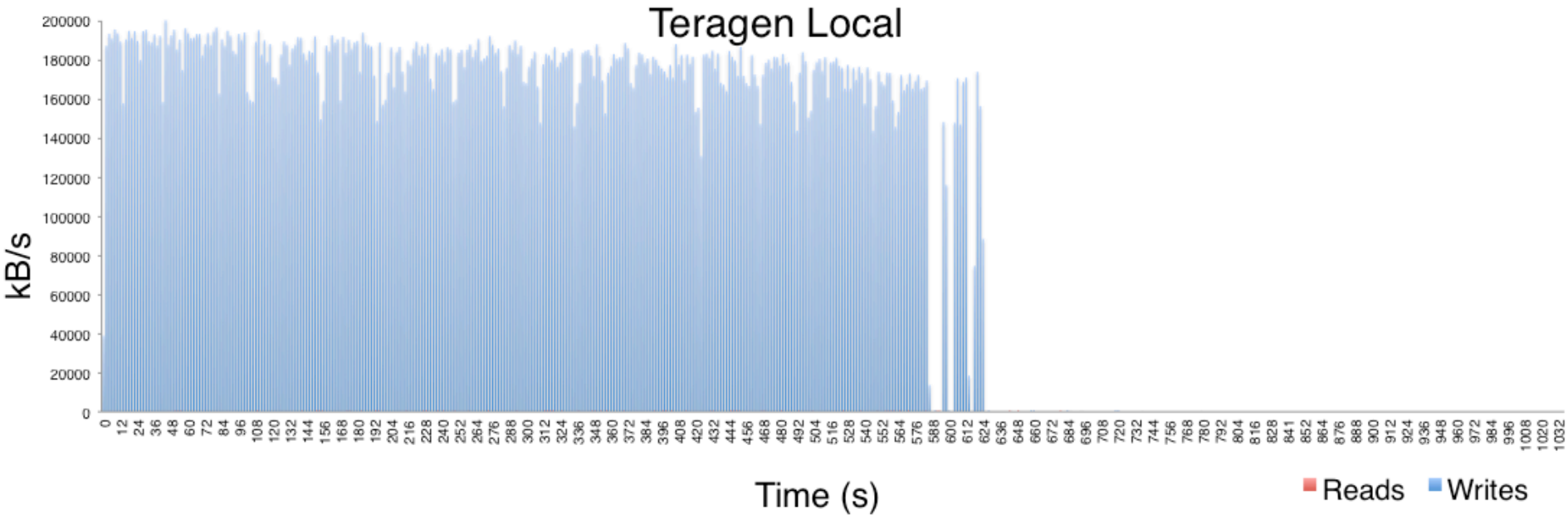
Methods


Local Data
Remote Data
CPU
CPU
Virtual Remote Data
CPU
CPU
CPU

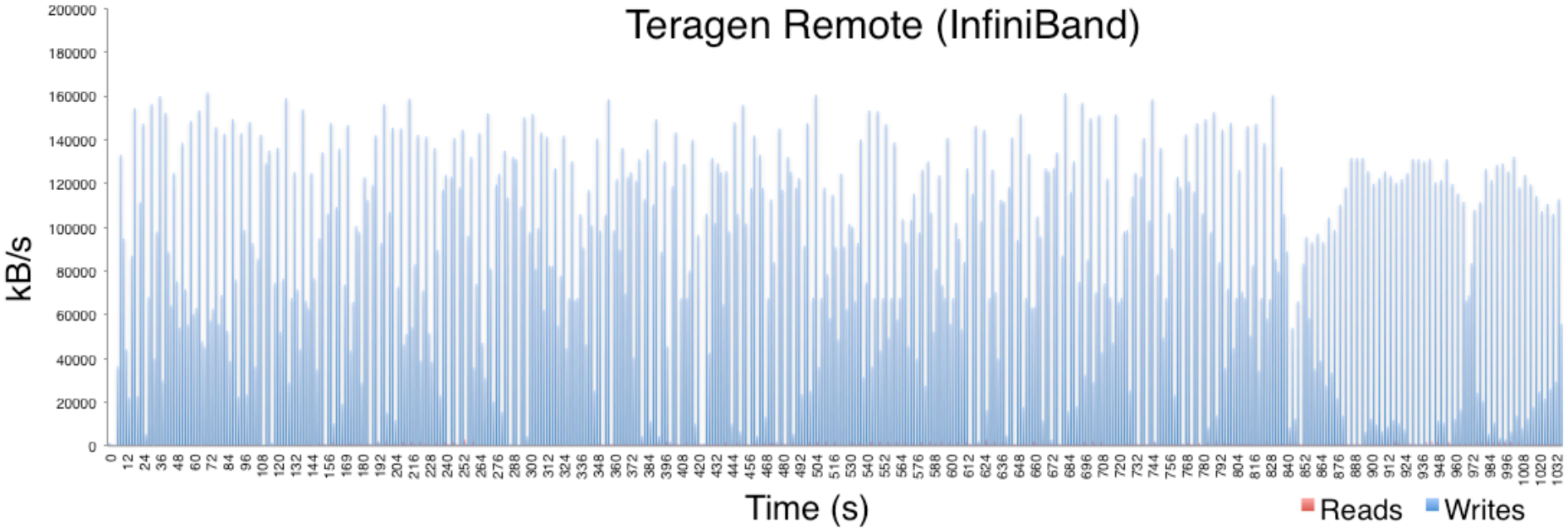
Results


Bonnie++
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes

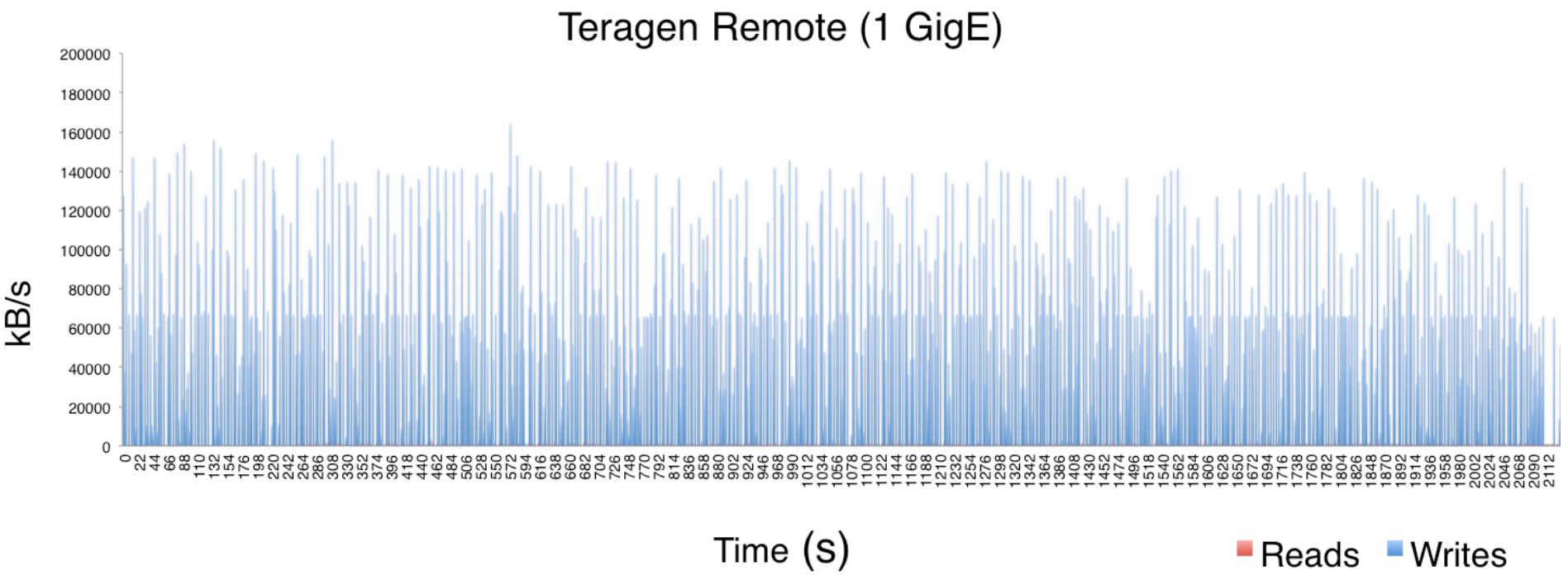
Results


Teragen Local
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes

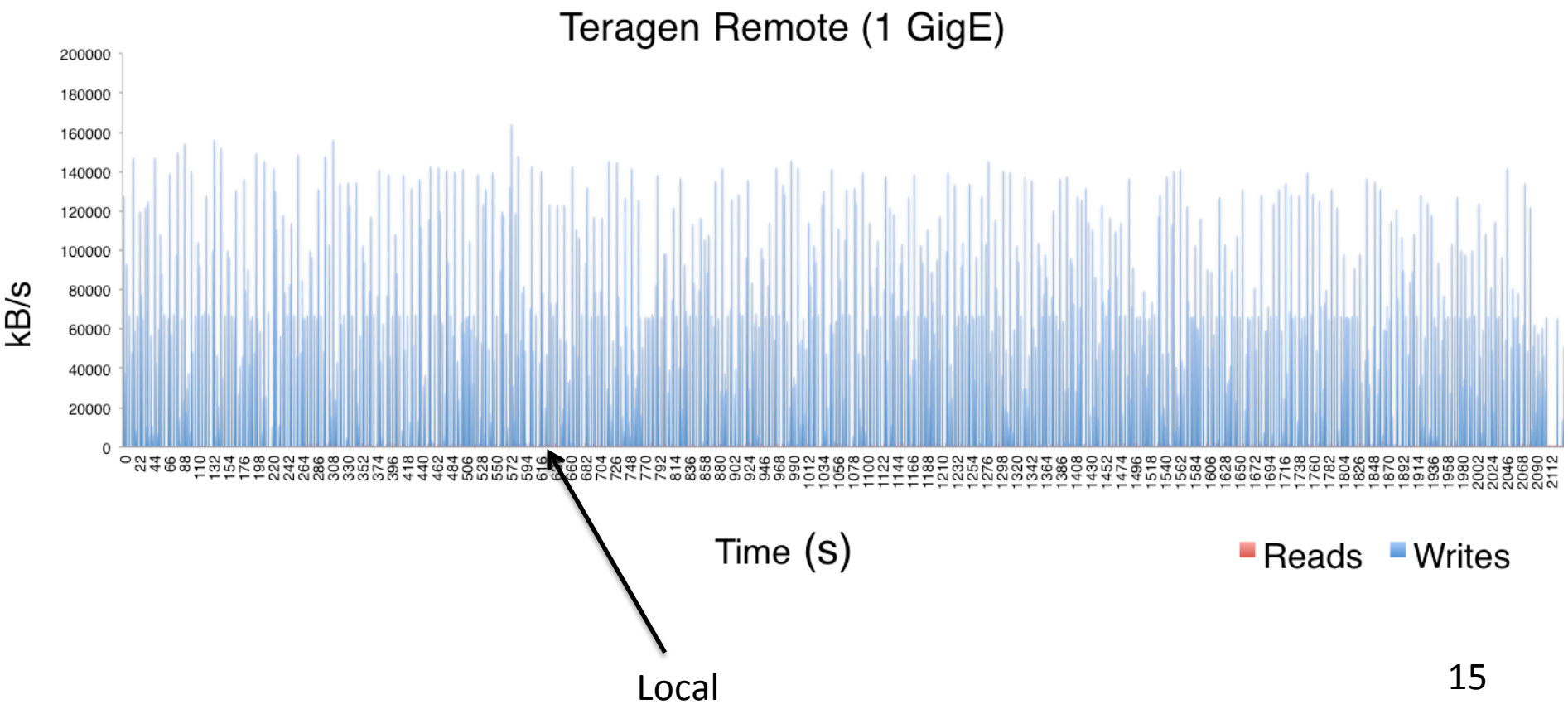
Results


Teragen Remote (InfiniBand)
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes

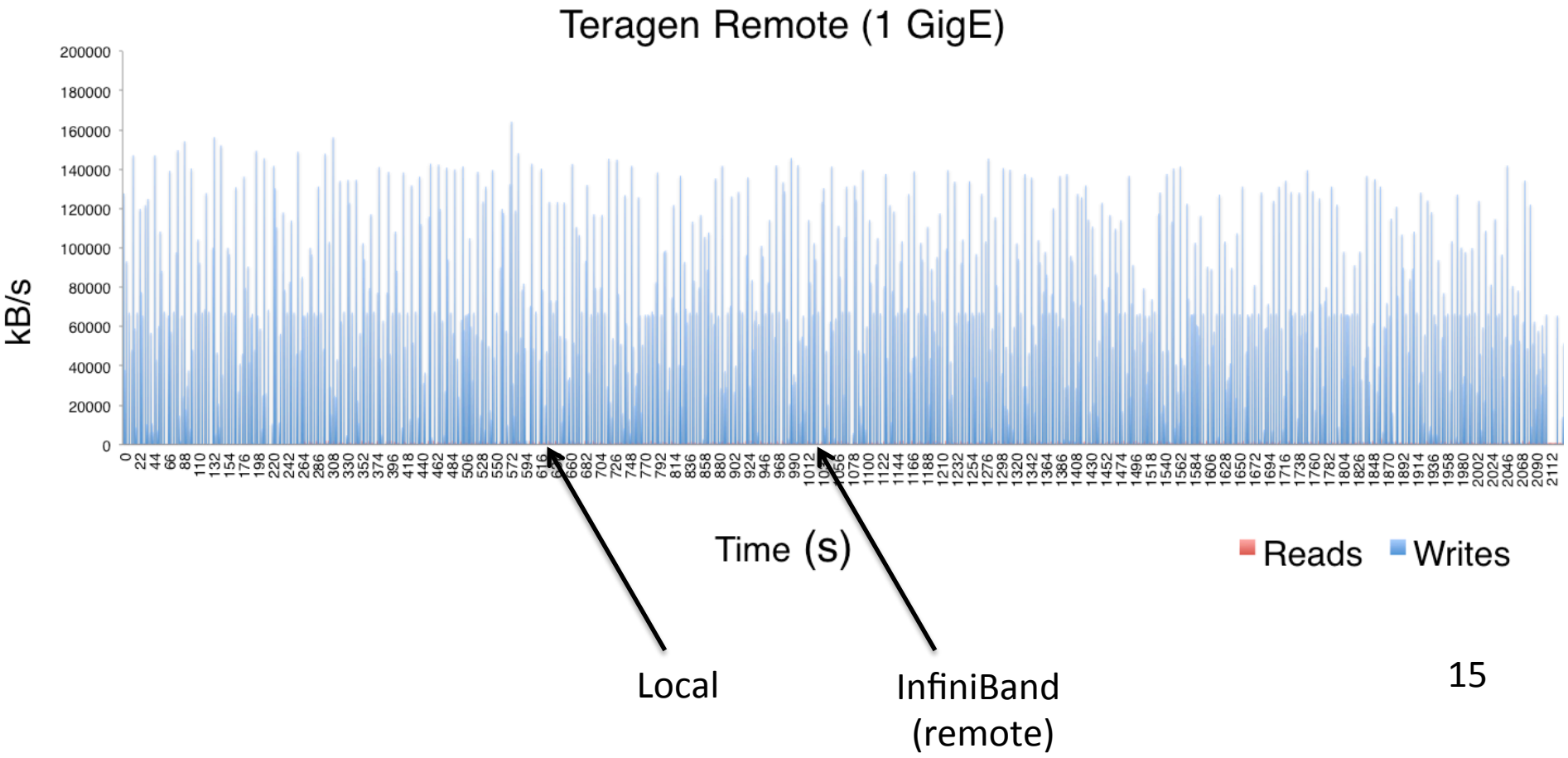
Results


Teragen Remote (1 GigE)
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
0 22 44 66 88 110 132 154 176 198 220 242 264 286 308 330 352 374 396 418 440 462 484 506 528 550 572 594 616 638 660 682 704 726 748 770 792 814 836 858 880 902 924 946 968 990 1012 1034 1056 1078 1100 1122 1144 1166 1188 1210 1232 1254 1276 1298 1320 1342 1364 1386 1408 1430 1452 1474 1496 1518 1540 1562 1584 1606 1628 1650 1672 1694 1716 1738 1760 1782 1804 1826 1848 1870 1892 1914 1936 1958 1980 2002 2024 2046 2068 2090 2112
Time (s)
Reads
Writes

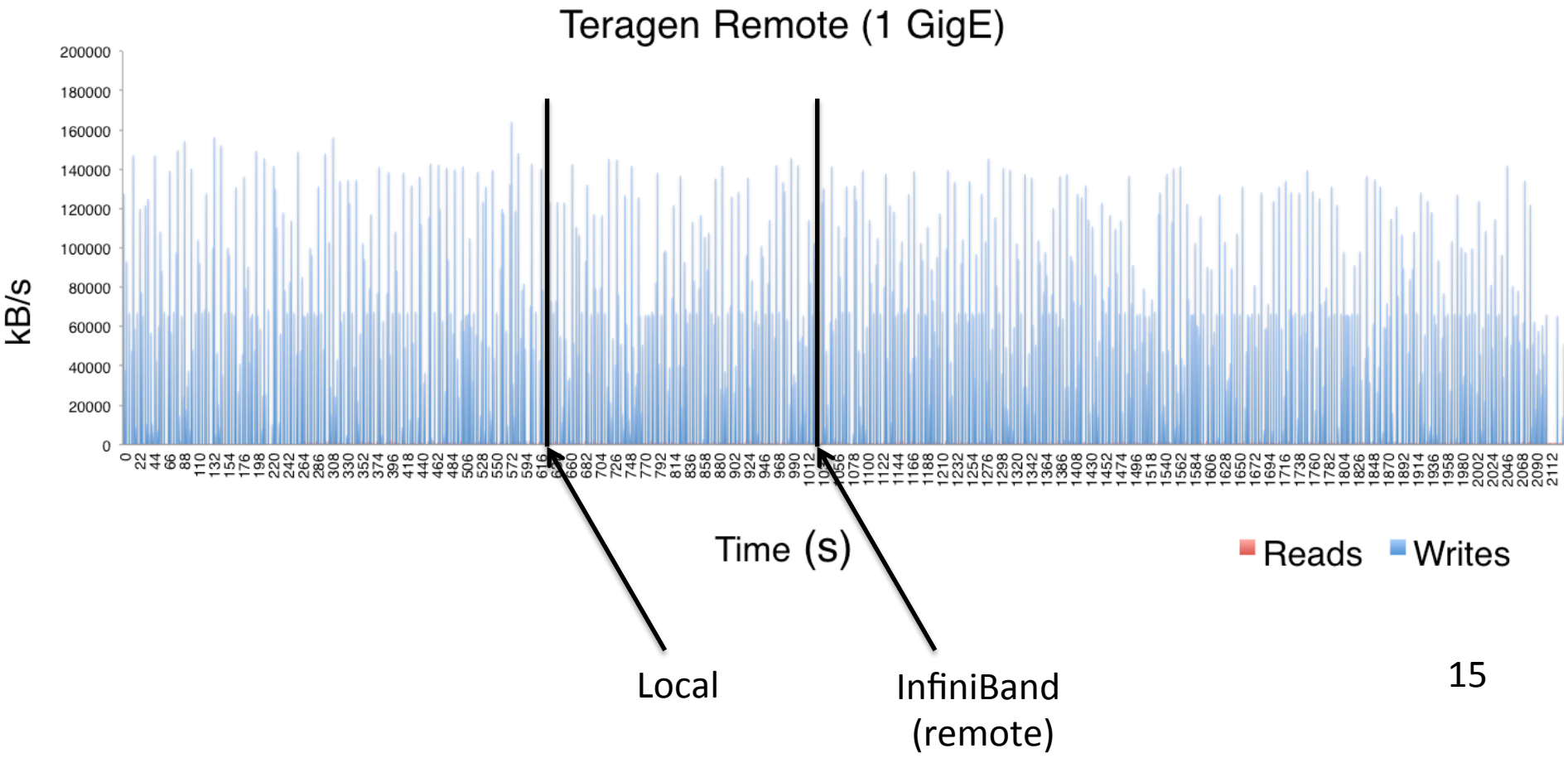
Results


Teragen Remote (1 GigE)
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes
Local

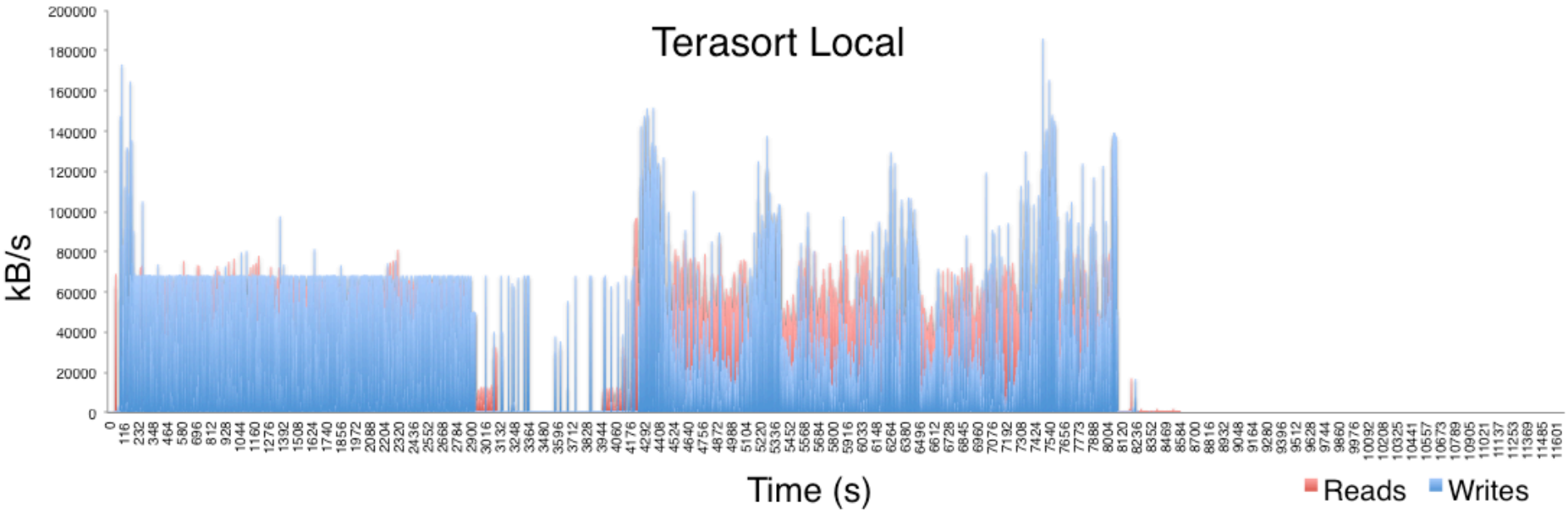
Results


Teragen Remote (1 GigE)
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes
Local
InfiniBand
(remote)

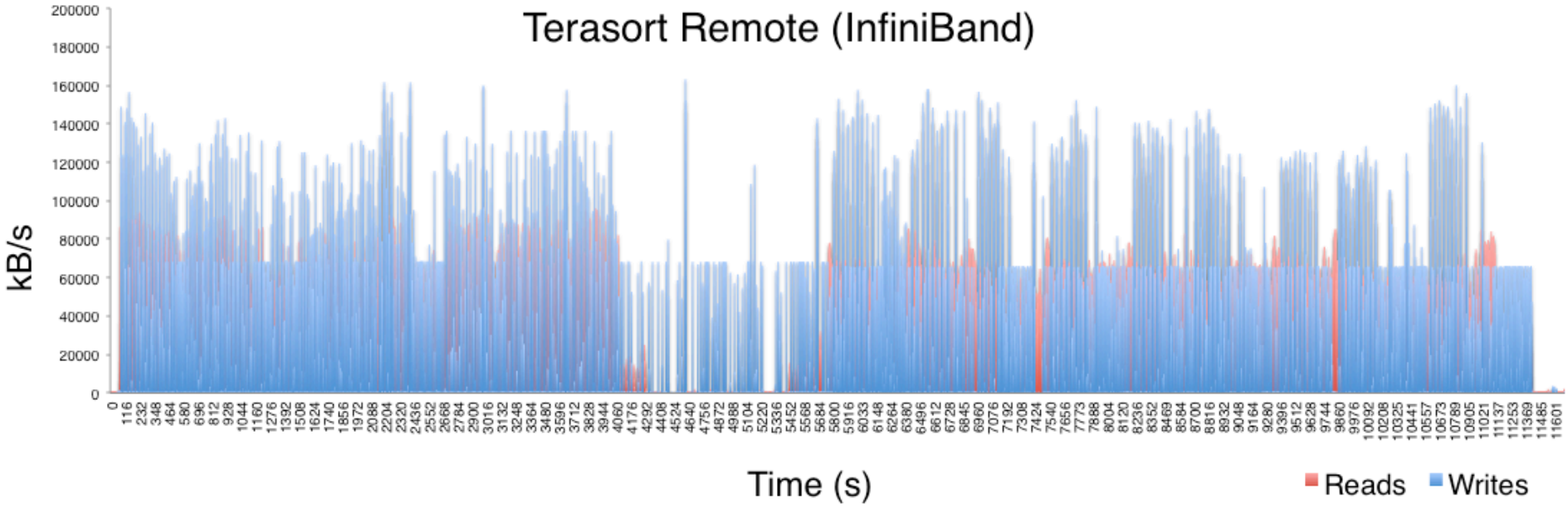
Results


Teragen Remote (1 GigE)
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes
Local
InfiniBand
(remote)

Results


Terasort Local
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
Time (s)
Reads
Writes

Results


Terasort Remote (InfiniBand)
kB/s
200000
180000
160000
140000
120000
100000
80000
60000
40000
20000
0
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*/