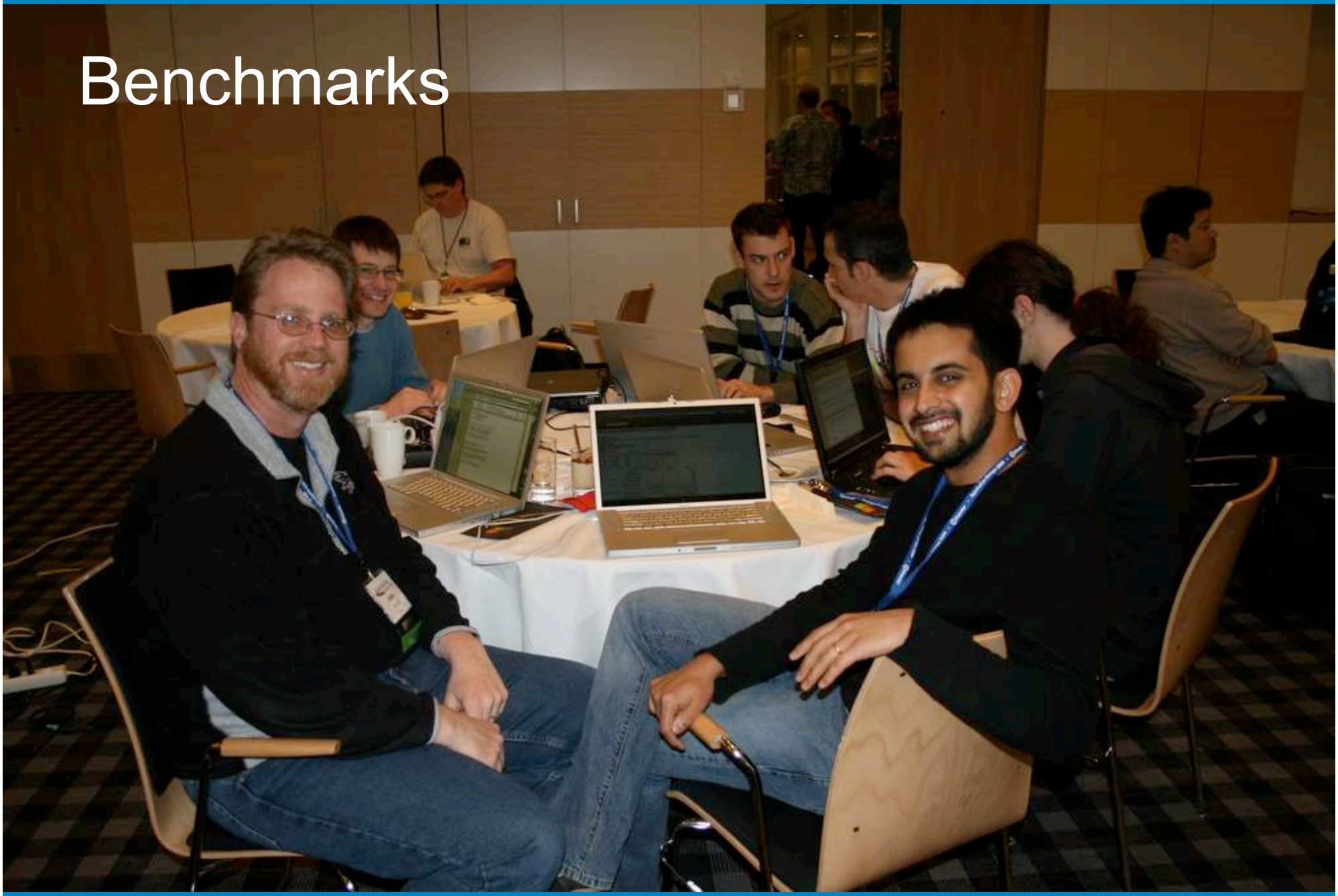
Benchmarking

Steve Loughran

Julio Guijarro

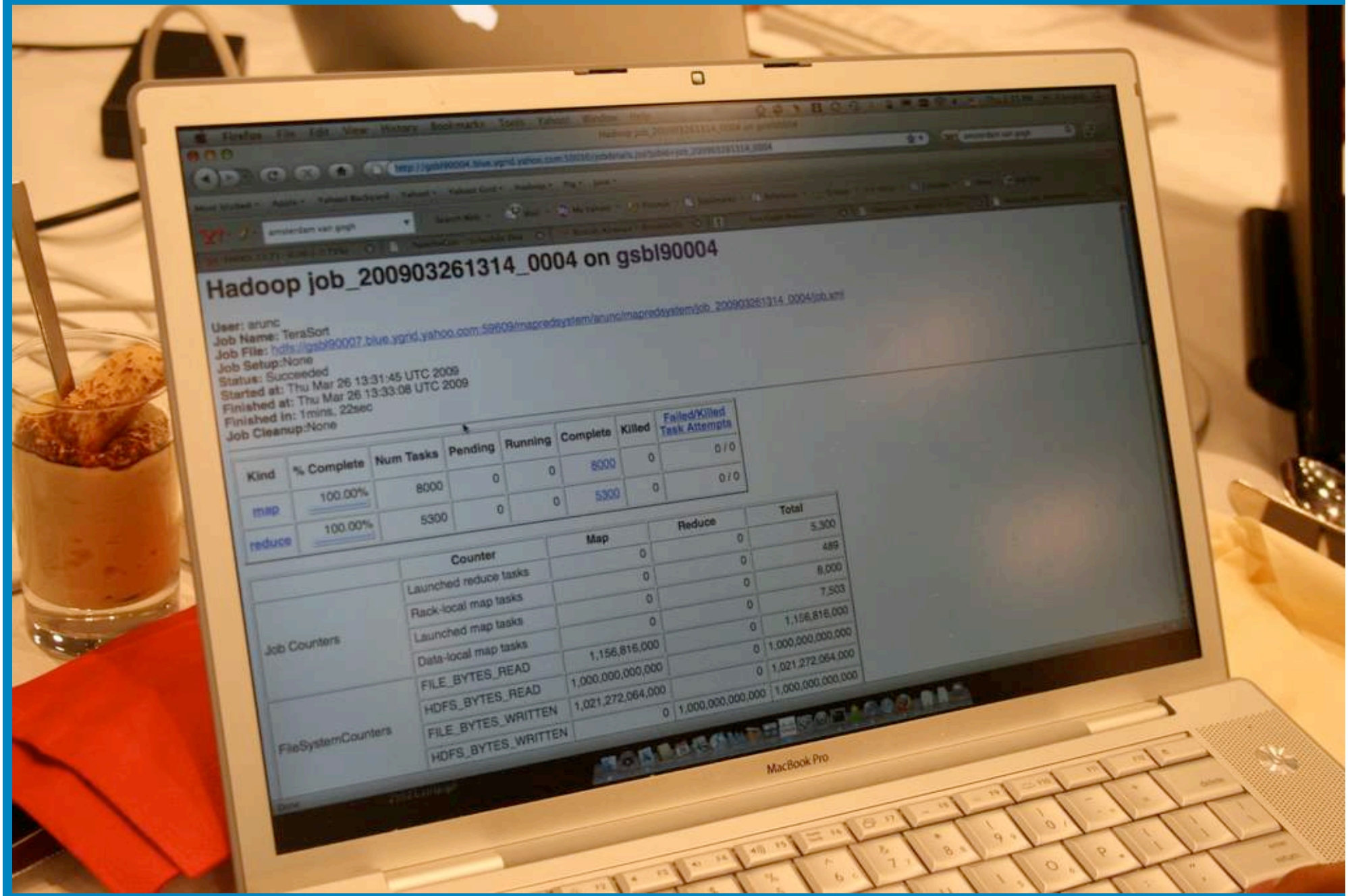
Benchmarks



Some Problems

- Estimating Hadoop performance of hardware
- Estimating Hadoop performance of a cluster
- *Designing Hadoop-ready servers*
- *Designing Hadoop-ready clusters*
- Optimising the network for Hadoop
- Optimising Hadoop/HDFS for specific applications

LABS hp


Hadoop job_200903261314_0004 on gsbl90004
User: arunc
Job Name: TeraSort
Job File: hdfs://gsbl90007.blue.ygrid.yahoo.com:59609/mapredsystem/arunc/mapredsystem/job_200903261314_0004/job.xml
Job Setup:None
Status: Succeeded
Started at: Thu Mar 26 13:31:45 UTC 2009
Finished at: Thu Mar 26 13:33:08 UTC 2009
Finished in: 1mins, 22sec
Job Cleanup:None
Kind | % Complete | Num Tasks | Pending | Running | Complete | Killed | Failed/Killed Task Attempts
map | 100.00% | 8000 | 0 | 0 | 8000 | 0 | 0 / 0
reduce | 100.00% | 5300 | 0 | 0 | 5300 | 0 | 0 / 0
Counter | Map | Reduce | Total
Job Counters | Launched reduce tasks | 0 | 0 | 5,300
Rack-local map tasks | 0 | 0 | 489
Launched map tasks | 0 | 0 | 8,000
Data-local map tasks | 0 | 0 | 7,503
FileSystemCounters | FILE_BYTES_READ | 1,156,816,000 | 0 | 1,156,816,000
HDFS_BYTES_READ | 1,000,000,000,000 | 0 | 1,000,000,000,000
FILE_BYTES_WRITTEN | 1,021,272,064,000 | 0 | 1,021,272,064,000
HDFS_BYTES_WRITTEN | 0 | 1,000,000,000,000 | 1,000,000,000,000
MacBook Pro

Recent customer request

"They want data for
Hadoop Sort
for 100GB."

LABS hp

Terasort: what else?

- PageRank: CPU intensive, small (static) input dataset
- *Something that stresses RAM and CPU*
- *Something that seeks in the files?*

Test Datasets

- Wikipedia: 5-10 TB of XML data with changes; user relationships have to be inferred
- SpamAssassin: 70+ GB of SPAM
- Physics? Something Small?

Network Measurement

What to add to Hadoop/Avro/Thrift to monitor network traffic -and relate to specific jobs?

LABS hp

Predicting performance

Can an MR job on small datasets predict performance on full size datasets?

What extra instrumentation can help?

Hardware Q

What should a Hadoop-ready server look like?

What about a rack?

Or a container?

LABS hp