F. Chlebana
June 29 2006

Status of the CSL Upgrade

I. Bizjak, F.Chlebana, G. Guglielmo, J.Lee, T.Masubuchi,

K.McFarland, R.Sarkis, JJ.Schmidt, M.Shimojima,

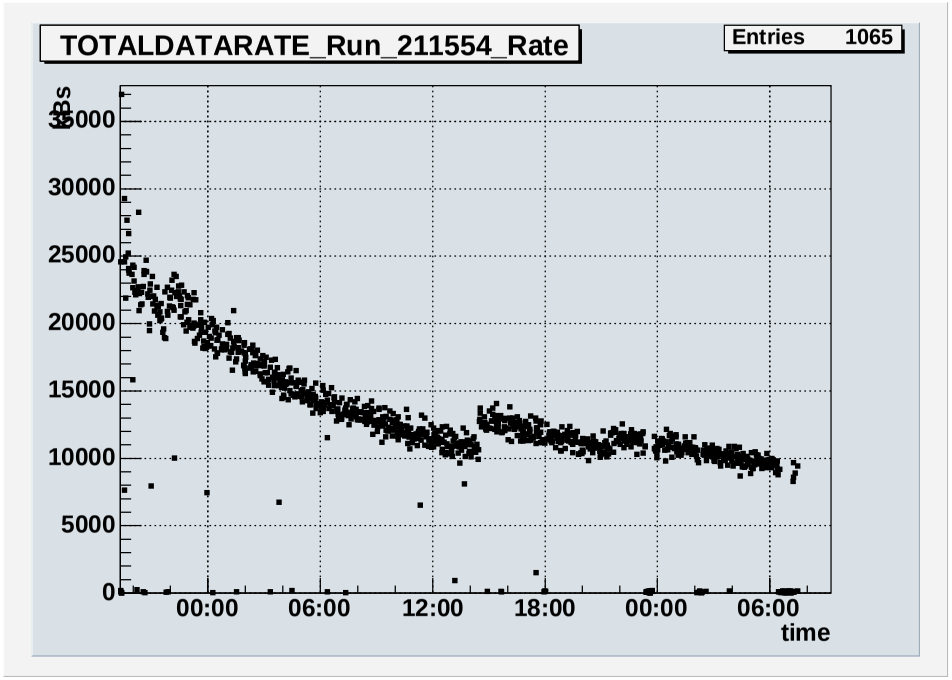
R.Snider, G.Yu, D.Zhang,

DataComm, Computing Division

Why Upgrade the CSL

- Current CSL is limited to about 24 MB/s
  → Need to be able to handle higher logging rates

Run 211554 ($\mathcal{L} = 176 \times 10^{30}$)



What logging rate will we need?

250 KB/Event
1000 Hz
L3 rejection of 4-5 times

→ 50 - 63 MB/s

Currently can write to buffer disks at higher rates (45 MB/s), limited by tape writing speed...

CSL upgrade design target is 80 MB/s

EVB Rate	1000 Hz	1000 Hz	1200 Hz
Rejection at L3	4-5x -> 250-200Hz	4x -> 250 Hz	4x -> 300 Hz
Event size	200 KB	250 KB	250 KB
Required Bandwidth	40-50 MB/s	63 MB/s	75 MB/s
Datahandling limit	80 MB/s		

Provide enough logging capacity so that the CSL is not the bottleneck in the system

- Uses Silicon Graphics and IRIX
  → Becoming more difficult to support
  → Data handling (SAM) software not ported to IRIX

- Using aging hardware
  → "end of life" of hardware approaching, difficult to support

CSL Upgrades

Retain original software design and port onto new hardware

→ Current software structure serves our purpose

→ Limited resources for any major rewrite

Increase data throughput

→ Use modular distributed logging architecture

Increased buffer capacity

→ Buffer increased from 3.7 TB to 24 TB

→ 80 MB/s → ∼7 TB/day

Improve maintainability and robustness

→ Replace aging equipment with new servers/disk arrays

→ Use Linux instead of IRIX

→ Redundant hardware and hot spares

Expandable

→ Can increase number of logger nodes and disk buffer

Hardware Status

Performance evaluated using a *prototype system* and hardware elsewhere at the lab.

Hardware has arrived and is installed

10 Servers, 3.4 GHz (dual processor) with 4 GB memory
Using Scientific Linux 4.2 (2.6 kernel)

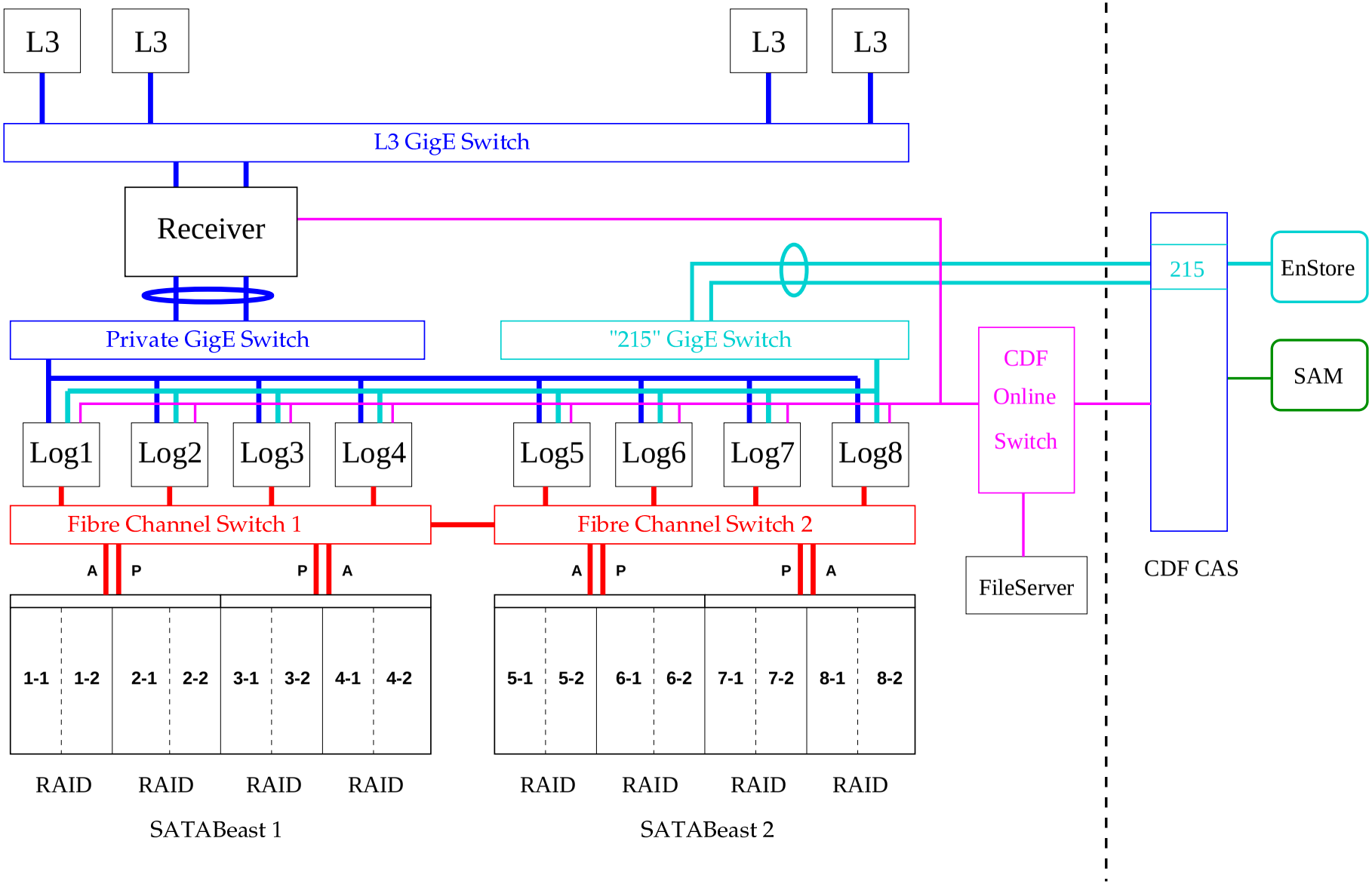
2 SATABeast disk arrays with 28 - 500 GB disks per unit (14 empty slots) $\rightarrow$ *Total of 24 TB of buffer capacity*

Data transported over Gigabit ethernet ($\sim$125 MB/s capacity) , using multiple Network Interface Cards to increase bandwidth

Servers connected to the Disk Arrays through a Fibre Channel Switch, providing a redundant path and failover protection.

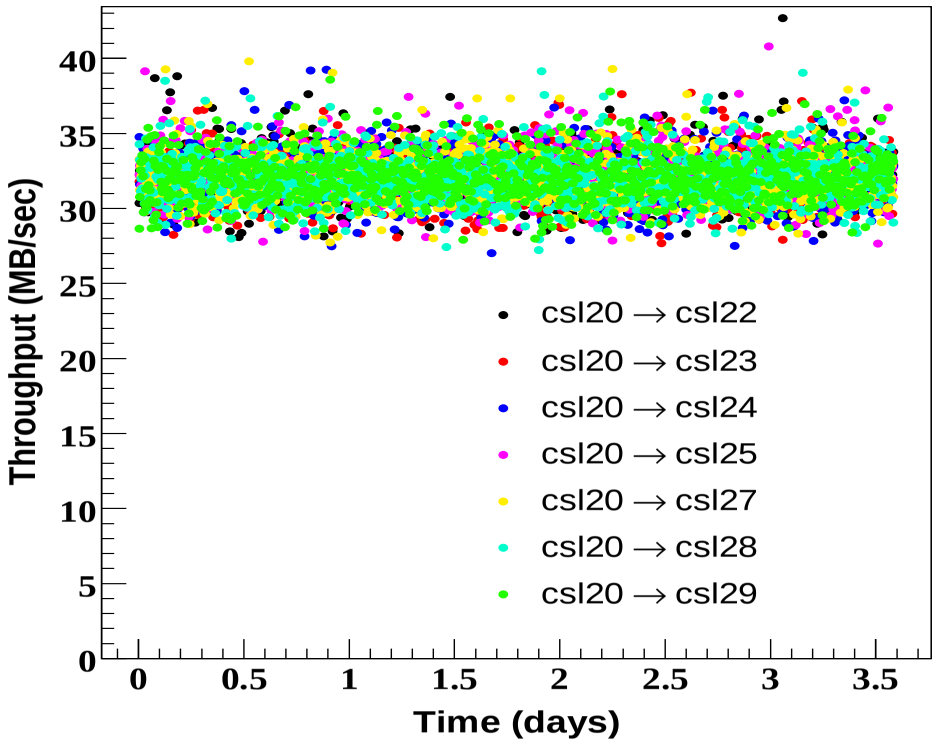
Additional equipment purchased for hot spares and to improve reliability and system monitoring

Architecture Overview



Running hardware stress tests on the new hardware

Network Tests



Sending data from receiver node to *7 logger nodes*

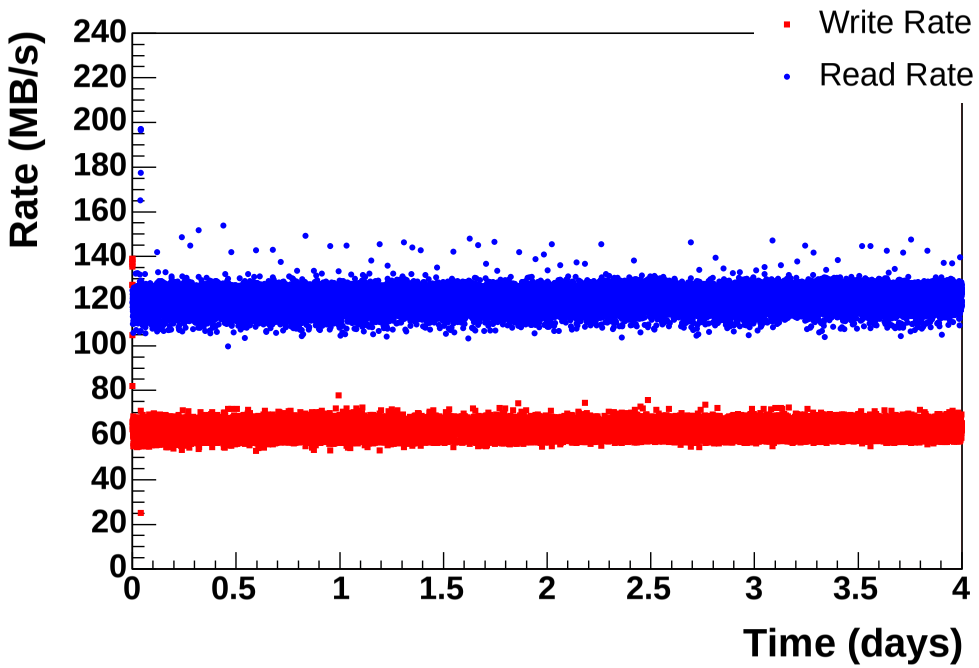
Bandwidth is equally shared

Using two GigE links *bonded* together

Total network rate: $7 \times 32 = 224$ MB/s

Disk IO Tests

Have been running long term tests in various configurations



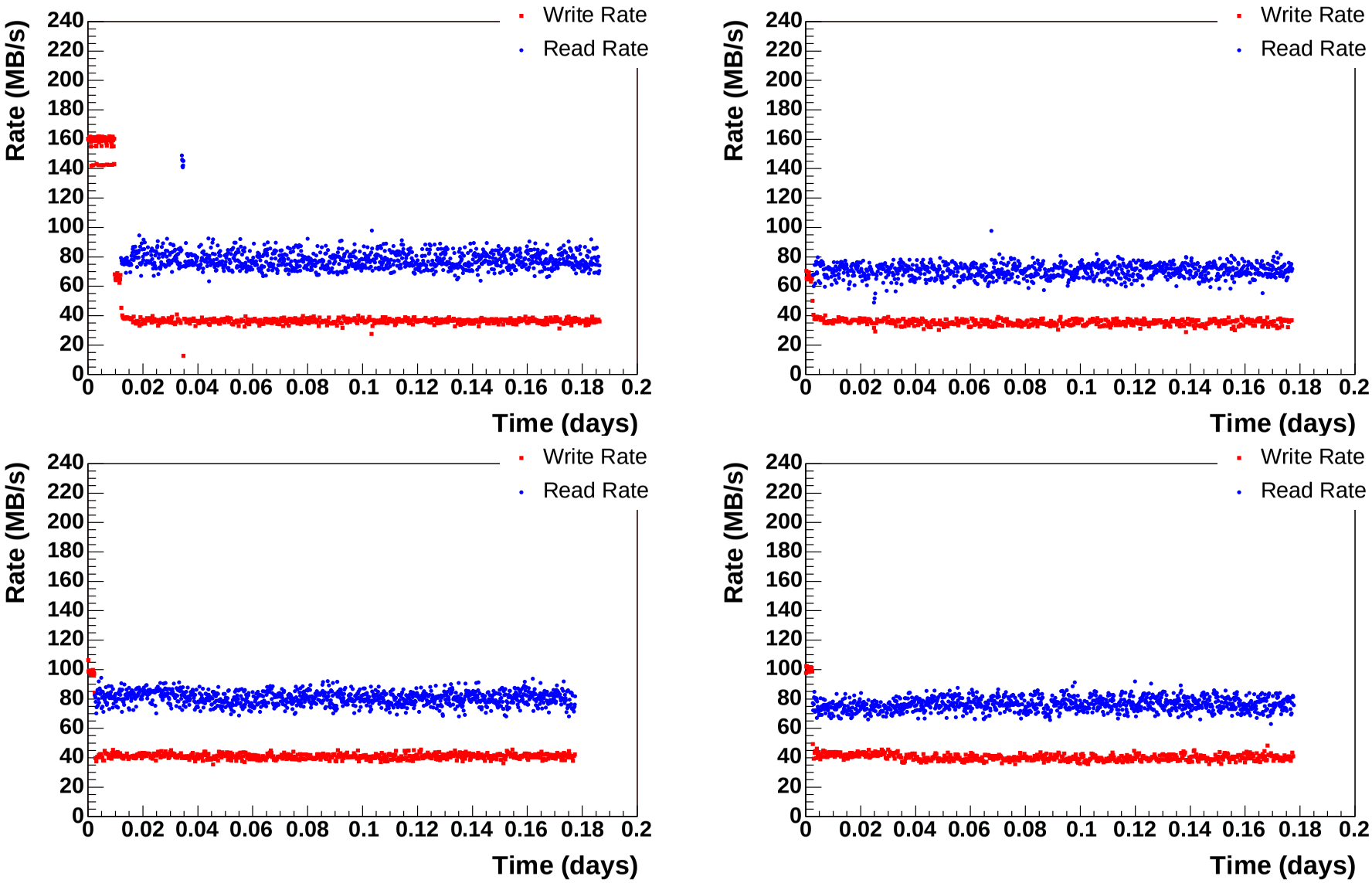
Using 4 servers to write to 2 SATABeast through 4 controllers

Servers are accessing the disk arrays through the Fibre Channel Switch

Tests running on/off for about two weeks

So far looks good!

4 logger nodes writing/reading to/from one SATABeast



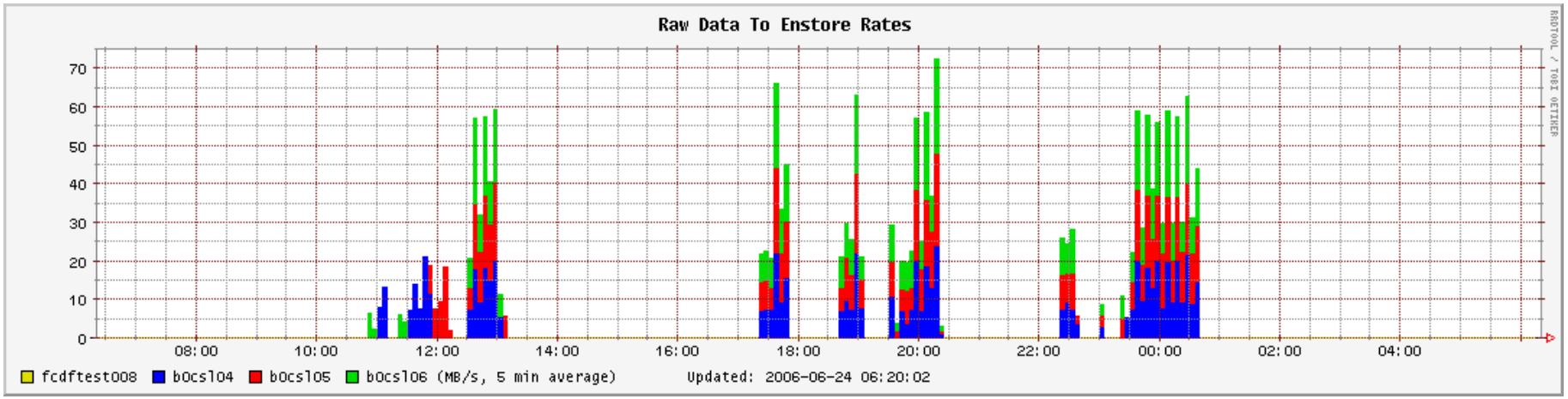
Total Disk Bandwidth using 8 loggers: > 300 MB/s

Integration Tests

Started full chain integration testing using *prototype system*

L3Sender → CSL Receiver → Distributed to 3 logger nodes → written to disk → Three separate Stager processes transfered data to Feynmann

→ Worked through details regarding naming conventions, location of files, logger/stager handshake etc...
→ Also sent events to a dummy consumer!



Average rate of about 40 MB/s - limited by L3Sender

Commissioning

Just yesterday we dismantled the prototype system so we can move development to the new system.

Start testing under more realistic conditions

→ Optimize performance

→ Work on reliability

→ Develop operating procedures

→ Test disaster recover

L3Sender (Sends events from data file to Receiver)

→ Verify data integrity

→ Ensure right number of events end up in the right place

One Subfarm Testing

→ Designate one L3 Output node for CSL testing

→ Run high rate data tests

→ Cosmic run and demonstrate data can be processed offline

Calibrations

→ Functionality of Calibration CSL and Data CSL combined

→ Routinely use new CSL for calibrations

End of store testing with all L3 output nodes

→ Demonstrate we can run reliably

Downstream logging software same as used at $D\emptyset$

New tape drives in Feynman being commissioned - started now, will take one month

→ *Could write to old tape drives but prefer not to...*

Switching back and forth between the old and new CSL should be straight forward.

→ Switch destination in L3 Output node

→ File naming convention is data driven

→ File location is stored in a database

Summary

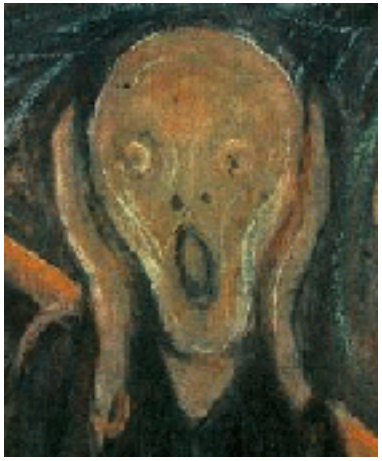
Equipment purchased and installed, currently being stress tested...
Additional equipment to improve reliability has been ordered

First integration test was successful!

CSL Logging software *compiled and run on new hardware*
L3Sender → Receiver → 4 logger nodes

So far hardware performing as expected and *well above* the designed target of 80 MB/s

Disk buffering:	$>$ 300 MB/s
Network:	$\sim$ 220 MB/s



→ *Next major goal is to run with 8 logger nodes while getting real data from L3*

Target delivery date: Sept 30th