Date and Location ----------

Thursday, April 07, 2005
4:00 PM - 6:00 PM

Attendees ----------

Apologies from FZK

Subject ----------

Subject was a general technical post-mortem of SC2, focusing on problems see by the sites.

BNL

Host performance slow
Kernel issues on disk I/O performance
Also long-haul network delay
Per-file rate 7MB/s 10 or 15 streams
Try to apply IN2P3

Need to gather what
- kernel version
- Patches
- Tuning information

FNAL

Possible to use radiantservice alias for gridftp ?
JAMES: This is an issue with castor SRM software - we need to raise it with the software developers first
Is it worthwhile to add a component into SC3 where we try and move data into as well as out of the radiant cluster
JAMES: Definitely - since we don't want to do it first off when the experiment start staging from tape

FZK

Apologies

IN2P3

Nothing particular
Good rate from Thursday
Added two machines more
One problem - handling of the cache disk on the machine

Need to move towards a consistent database of site configurations

BNL: Need to see what happens when we move up the stack

NL : What kernel params were tuned ? - Laurent knows?

INFN

Didn’t use the wiki, but collected the observations in a document - sent to the list

Observations

- Transfer queue size to 10000 from 1000, the throughput through iperf increases and was more stable. Relevant for large RTT
- Tuning of kernel params Re Read/Write buffers
  - Min 1Mb/ Max 32M/ Default 16M. This was needed for optimum single stream iperf performance.
- Window size
  - Optimum minimum value is 3MB (6MB for the kernel) for INFN
  - Gridftp auto-tuning feature will help for this
- Had network performance issues.
  - Were not able to understand the reasons
  - Was asymmetric issues - only INFN to CERN, so didn't affect the SC traffic
  - Some issues on oplapro nodes due to monitoring traffic affecting errors in the counters
- Saw performance issues related to scaling
  - 35MB/s at the worst scenario. This should load the link at full load
  - Would like to test to see how the aggregate changes with number of gridftp sessions - BOOK A SLOT
- Issues with I/O performance which is dependant on amount of space on disk

Mark: NL have seem that if the disk is about 70%, you get decrease in throughput

Andrew: fragmentation problems as well

- Giuseppe will write a document summarising the load-balancing used for the SC

RAL

Network arrived late in the day. 2 x 1Gb link. Couldn't achieve more than 75-85 Mb/s

Had problem with UDP about 750Mb/s - above that packet loss is high

UKLIGHT concluded that the link had been underprovisioned

Link reprovisioned - now can get 1Gb/s

DANTONG: you can see performance problems with agregate links

Mark: How is the agregation done

Andrew: IPv4 XOR'ing of IP addresses. Complications due to another aggregate 2x1Gb at RAL end.

Mark: We saw similar problems with out etherbundling

Andrew: would like to go to SRM. Also trying to get the gridftp servers to get connections from both production and UKLIGHT network.

James: We need to see how we do this multiple connection to the storage cluster

Andrew: We want

NL

Mark - 950MB iperf single node

Aggregate 200MB/s across hosts with directed transfers to specific disks -not tuned hosts. Saw hosts crashing and bad performance once started doing

radiant transfers. Was due to buffer cache again - became more stable. Saw that data was kept 50 seconds in memory before being written to disk. After tuning only kept in memory for 5 seconds. Flushd was waking up every 2.5 seconds

Had to schedule transfers across disks as well as nodes.
Saw some movement of transfers across to a single node. This killed the aggregate performance

Problem with scheduling agent - saw db was down.
James: Saw problems with central db service
Also, don't want to put FS and host knowledge into scheduler - this should be a site issue to put SRM, etc...

Mark: Create single FS on nodes as a response

Best performance with single stream transfers.

Cert got revoked before end of lifetime and stopped jobs running

Bug in radiant that needs a grid-cert for job submission, when there was ones in myproxy - should be passed forward to glite team

General

Mark: What is the layout and configurations of sites. What is amount of data for SC3. What is the ramp-up for the production phase.

James: We need to work on this as well. Perhaps a future phonecon (2 weeks time?) should be dedicated to this issue?
