



Fermilab

FNAL/CERN Data Transfer Challenge

SARA

D. Petravick

Jan 20, 2004

What's Potential Performance?

Many gigabits disk to disk

The service Challenge
for Storage Elements.

Disk-to-Disk transfer
work gives upper
limits

SE's in production must
manage concurrent
local POSIX, tape as
well as GRID data
tape xfers

Transferring a TB from Caltech to CERN in 64-bit MS Windows

- ◆ Latest disk to disk over 10Gbps WAN: **4.3 Gbits/sec (536 MB/sec)**
- 8 TCP streams from CERN to Caltech; 1TB file
- ◆ 3 Supermicro Marvell SATA disk controllers + 24 SATA 7200rpm SATA disks
 - Local Disk IO – **9.6 Gbits/sec (1.2 GBytes/sec read/write, with <20% CPU utilization)**
- ◆ S2io SR 10GE NIC
 - 10 GE NIC – **7.5 Gbits/sec (memory-to-memory, with 52% CPU utilization)**
 - 2*10 GE NIC (802.3ad link aggregation) – **11.1 Gbits/sec (memory-to-memory)**
- ◆ Memory to Memory WAN data flow, and local Memory to Disk read/write flow, are not matched when combining the two operations
- ◆ Quad Opteron AMD848 2.2GHz processors with 3 AMD-8131 chipsets: 4 64-bit/133MHz PCI-X slots.
- ◆ Interrupt Affinity Filter: allows a user to change the CPU-affinity of the interrupts in a system.
- ◆ Overcome packet loss with re-connect logic.
- ◆ Proposed Internet2 Terabyte File Transfer Benchmark

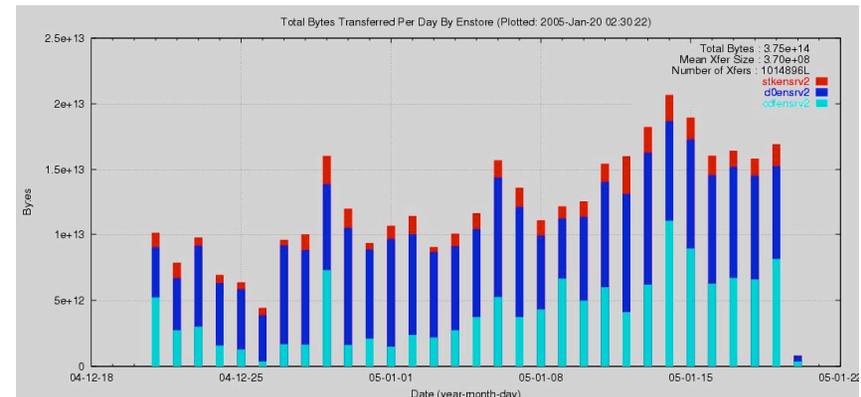
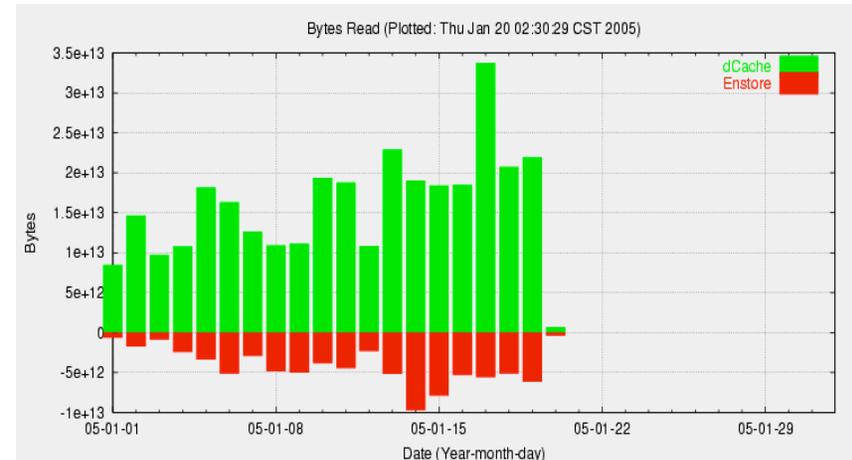
Today's Practical Storage Elements



- Very,very large commodity infrastructures have been built on LANs and used in HEP.
 - Specialized SANS are not used generally in HEP
- It must at least be the starting point for mingling advanced networks and large HENP data systems.

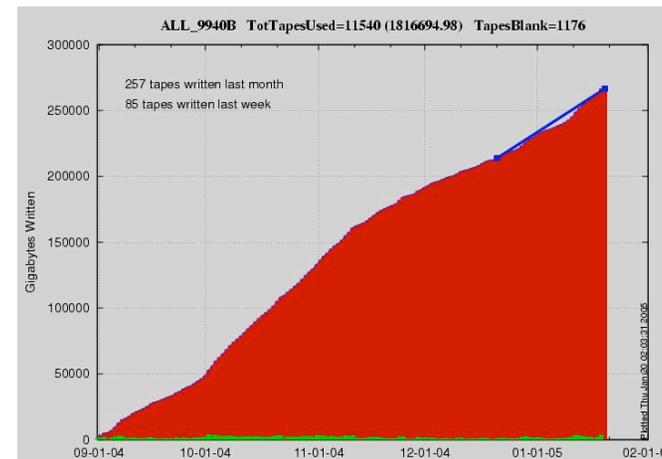
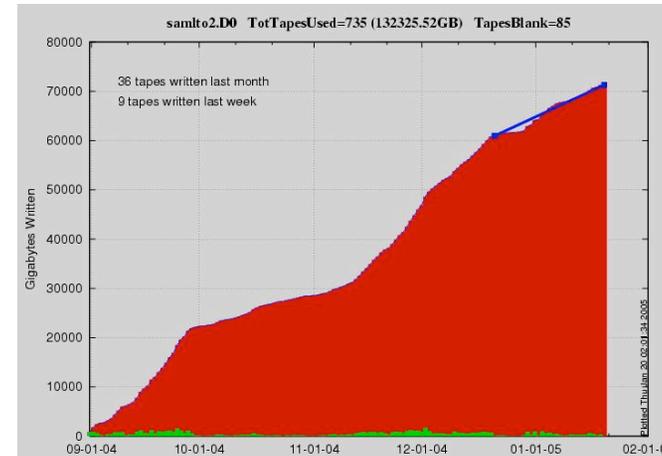
Run II

- All experiments at FNAL use Enstore
 - STK and LTO tape in production
- CDF use dCache, D0 do not.
- Local tape and CDF posix IO are routinely at 10's of TB/day.



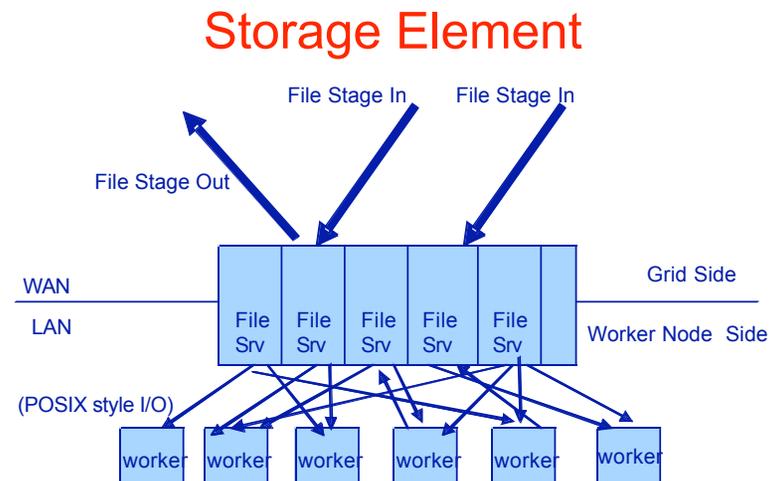
What we get from tape (aside from expense and bother)

- Basis of custodial store.
 - Separation of roles on (true) delete
 - Write protect tab used
- Expandable with low risk
 - Could we really have commissioned 1 PB of disk since Sept just in time (and every time)?
 - Perhaps in the near future.
- Its hard to see how File systems, by themselves provide a disk based custodial store.



Storage System Model

- Whole files are moved to and from the SE over the GRID w/ grid interfaces.
 - Large bandwidth*delay
 - Grid interfaces (SRM, gFTP)
- Local access by WN's is Posix
 - Files are accessed bit by bit.
- dCache SE's can have more structure (not detailed in figure)
 - Can support tape.
 - Can move data through firewalls and NAT devices.
 - Can add nodes without recycling the system
 - Can carry on if nodes fail.



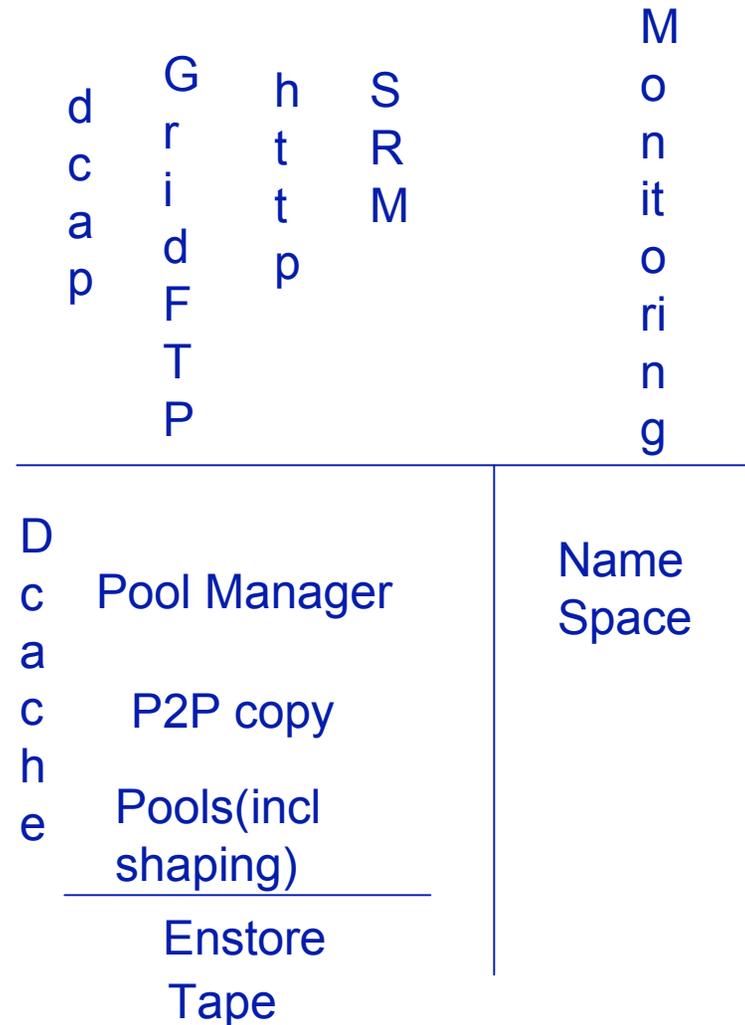
1/10/2005

Don Petravick -- Fermilab

12

SE's at FNAL incl T-1

- Enstore (tape)
 - Scalable, flexible tape IO (STK and LTO)
- dCache (disk)
 - DESY/FNAL
- PNFS Name space
 - DESY (FNAL some Mods)



(Hardware) Failures

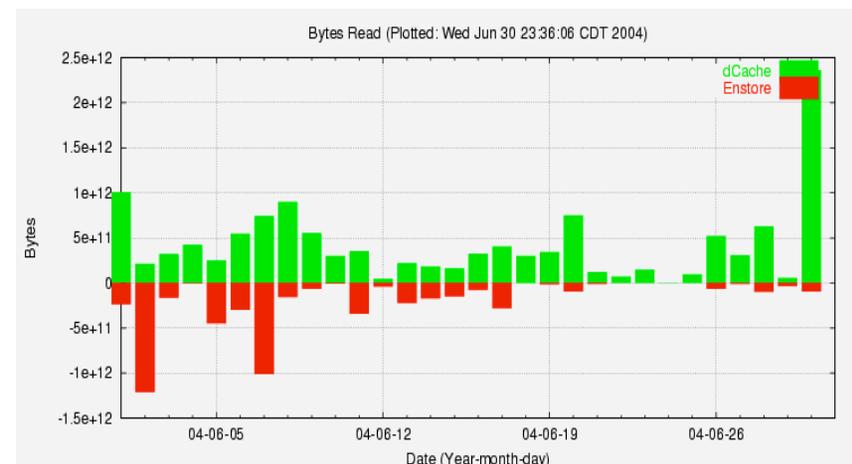
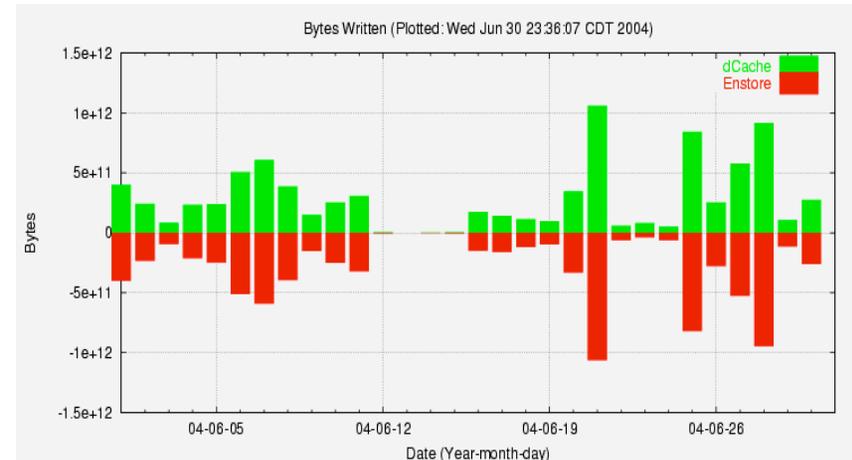
- “personal” disks to obtain capacity.
- See MTBF, infant mortals, etc. (and servers).
- Mitigations:
 - Maintain functionality to Posix users
 - Dcap features, and SE ability to muster replicas
 - Replicas come from
 - Lower level store (typically tape, but could be some other system.
 - Replicas which happened to be in the system
 - A resilient dCache deployment.(overt creation and maintenance of duplicates.
 - Target application -- Storage system physically co-resident with a computing farm.

CMS Challenge Goals

- Concentrate on full system level issues rather than optimizing single components.
- Given specialized tests, see how the integrated framework performs.
 - specialized set ups for specialized tests
- A production system was used at FNAL for the challenge
 - Service challenge co-existed with normal uses of the system.
 - S.C. Could not impede or interfere greatly with normal work.
 - Able to study feature interaction and scaling in the most realistic environment.

June 05 -- Prior to Challenge

- Routine production at the FNAL Tier 1 center.
 - Green – netwrk movement
 - Red - Tape movement
 - The two plots are not scaled the same
 - FNAL was OC-12 in June.
- Upper plot - Ingest
 - up to 2.5 TB/day
- Lower plot – Reads
 - up to 1.1 TB/day



Challenge SS Hardware

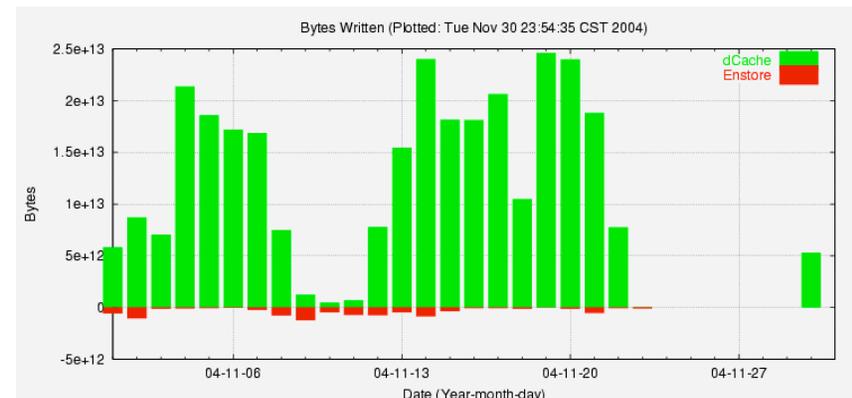
- CERN-side
 - Nine nodes were deployed for this test.
 - 3 disks/node, 10 files/disk (270 files)
 - Gigabyte files.
 - not supporting other production
 - CASTOR not present.
- FNAL-side
 - dCache Storage system software
 - 32 dCache pools on 22 nodes (3 TB total)
 - Same disk served other production pools.
 - Volatile pools (No tape), automatically cleared many times/day
 - Production users and tape accessing the disks.

Challenge Network Hardware

- Network setup:
 - CMS dCache nodes at FNAL are...
 - Each 1 Gbit connected to a 6509 which is...
 - Connected via 10 GB LAN_PHY to CalTech Starlight POP which is...
 - No special fire wall device
 - OC192 connected to CERN...
- Notes:
 - The test flows were substantially the only flows on this wide area link.
 - FNAL dCache nodes are on a local production LAN.
 - Large MTU's were not used.

Results (throughput)

- November graph
- GridFTP protocol
- Throughput:
 - Many consecutive days of > 10 TB/day
 - 25 TB/day peak (~3 Gbit sustained)
 - Gap – SC 2004
 - “bandwidth challenge”



Results (TCP)

- Previous tests on specialized hardware concentrated on large TCP buffers and a few streams.
- This was not made to work in the challenge.
 - hundreds of user transfers per node, large TCP buffers quickly exhausted memory and caused machines to crash.
- The design used:
 - 2 MB TCP buffers and 20 parallel streams for each transfer.
 - 2 MB/sec/stream giving rates of 40 MB/s for each file.
 - Tuned with FNAL production netflow analyzer.
- Expectations for this hardware:
 - The current set of hardware combined with the dCache IO yields a maximum rate between 50-60 MB/s for each node.
 - Unit tests using optimized C code can achieve 70-80 MB/s for each node.
 - Therefore, the 40 MB/s per file transfer was deemed acceptable at this point of development
- Much was tried and learned,
 - This challenge provided more rate than some R&E networks currently carry.
 - Did not achieve the best performance levels seen in unit tests
 - Did not investigate whole parameter space
 - E.g. Did not use large MTU's

Additional Performance Comments

- Reliable running was achieved after several week so of debugging.
- Typically
 - 150 SRMCP's (copies) in queue, each requesting 3-10 files.
 - 6 active gridFTP's/pool
 - 250 – 350 MB/sec.
- Transfers were submitted; sometimes many dozens would depend on a single node at CERN.
 - This dramatically affected performance, but not reliability.
 - dCache would have used pool-to-pool copy to optimize.
 - It would be beneficial to test with a full storage system at CERN

Results (Functionality)

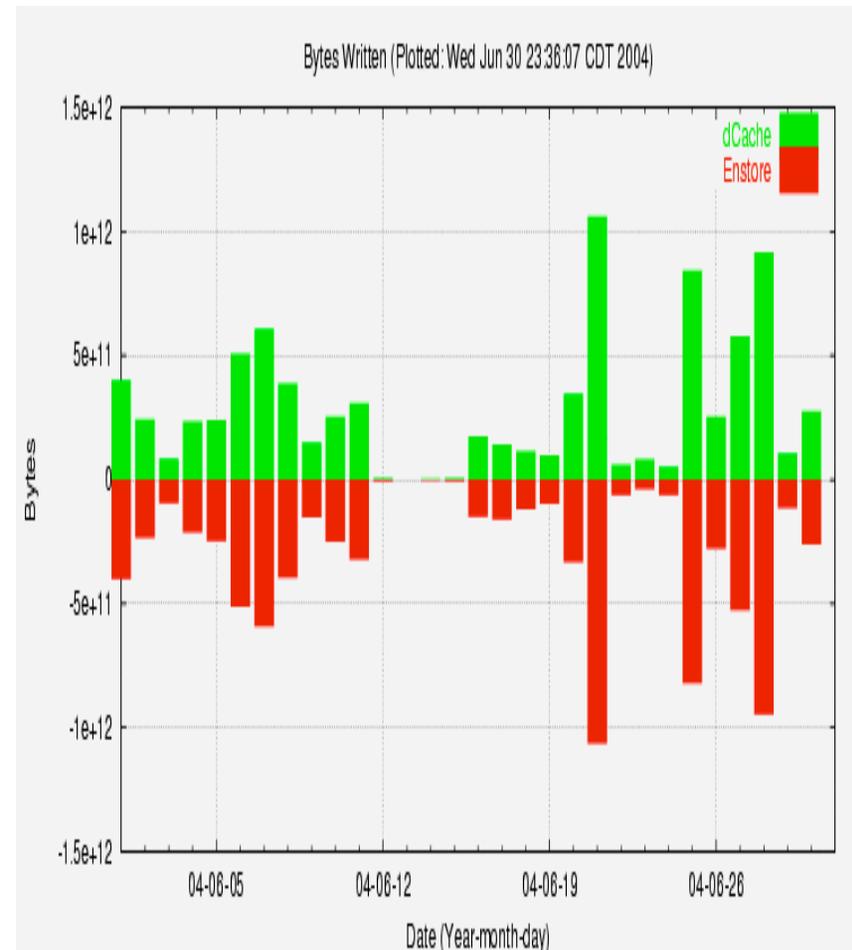
- FNAL-side SRM mastered the transfers.
 - CERN side provided just GridFTP functionality
 - Result: belief that pull is better than push.
- Many bugs were identified and fixed over the 1st weeks of the challenge. This was the real goal of the challenge.
 - Properly clean up when xfers were killed
 - Developed a simple system view to understand transfers.
 - Modular updates - Make SRM its own .jar
 - Applying priorities properly (service challenge uses v.s. production use)
 - Regulate # of concurrent FTP's independently of # of local accesses (data movement resources are different)
 - Preserve state across crashes, power-downs, failure of pool nodes, Use preserved state to recover where possible.
 - Properly scaling monitoring (SPY)
 - Configuration issues (gridmap files, corrupted certificates).

Grid FTP server transfers

- As a separate test, a special gridftp only script was written.
 - explicitly matched files at CERN with pools here at FNAL to optimize throughput.
- Only 1 file from each disk at CERN was used, (3 per node), possibly leading to some memory caching at CERN.
- Results:
 - Files written to memory at FNAL, provided a rate of 500 MB/s.
 - Files were then written to dCache pool disks (by the GLOBUS COG gridftp Client , not through the dCache), and the rate was 400 MB/s.

50 MB/sec Tape Challenge

- News to the CMS T1 center
 - Seems feasible
- Questions about its definition
 - How long will ingested data be retained?
- 50 MB/sec \approx 5 TB/day
 - FNAL T1 have had many days of 1 TB/day ingest for over a year
 - June production input plot is shown



CMS T1 integration needs

- Continue production
- Integrate to US T2 centers w a SC style
 - Florida first.
 - OSG context

Directions

- Commissioning of resilient dcache for the CMS T-1 (plan -- organize 1 PB of disk on its farm)
- Work with OSG to provide manage SE in the US (dcache and DMR)
- Add accounting
- Add integration with dynamically provisioned “pipes” (lambda station)
- Incorporate the best WAN integration knowledge.
 - Investigate gap between SC and best performance

Summary

- The challenge was a valuable system integration exercise.
- Demonstrated the usability of our systems at throughput which are a factor of 10 greater than prior CMS production use.
 - Achieved routine movements of 20 TB/day
 - Run II production is > 30 TB/day (on LAN)
- The Service challenge found problems in a fairly realistic deployment.
 - Some problems would have likely been missed in a less integrated test.
 - Many of these problems are resolved.