DØ Data Handling Operational Experience

CHEP03

UCSD

March 24-28, 2003

Lee Lueking

Roadmap of Talk

- DØ overview
- Computing Architecture
- Operational Statistics
- Operations strategy and experience
- Challenges and Future Plans
- Summary

The DØ Experiment

- D0 Collaboration
  - 18 Countries; 76 institutions
  - 600 Physicists
- Detector Data (Run 2a end mid '04)
  - 1,000,000 Channels
  - Event size 250KB
  - Event rate 25 Hz avg.
  - Est. 2 year data totals (incl Processing and analysis): 1×10^9 events, ~1.2 PB
- Monte Carlo Data (Run 2a)
  - 6 remote processing centers
  - Estimate ~0.3 PB.
- Run 2b, starting 2005: >1PB/year
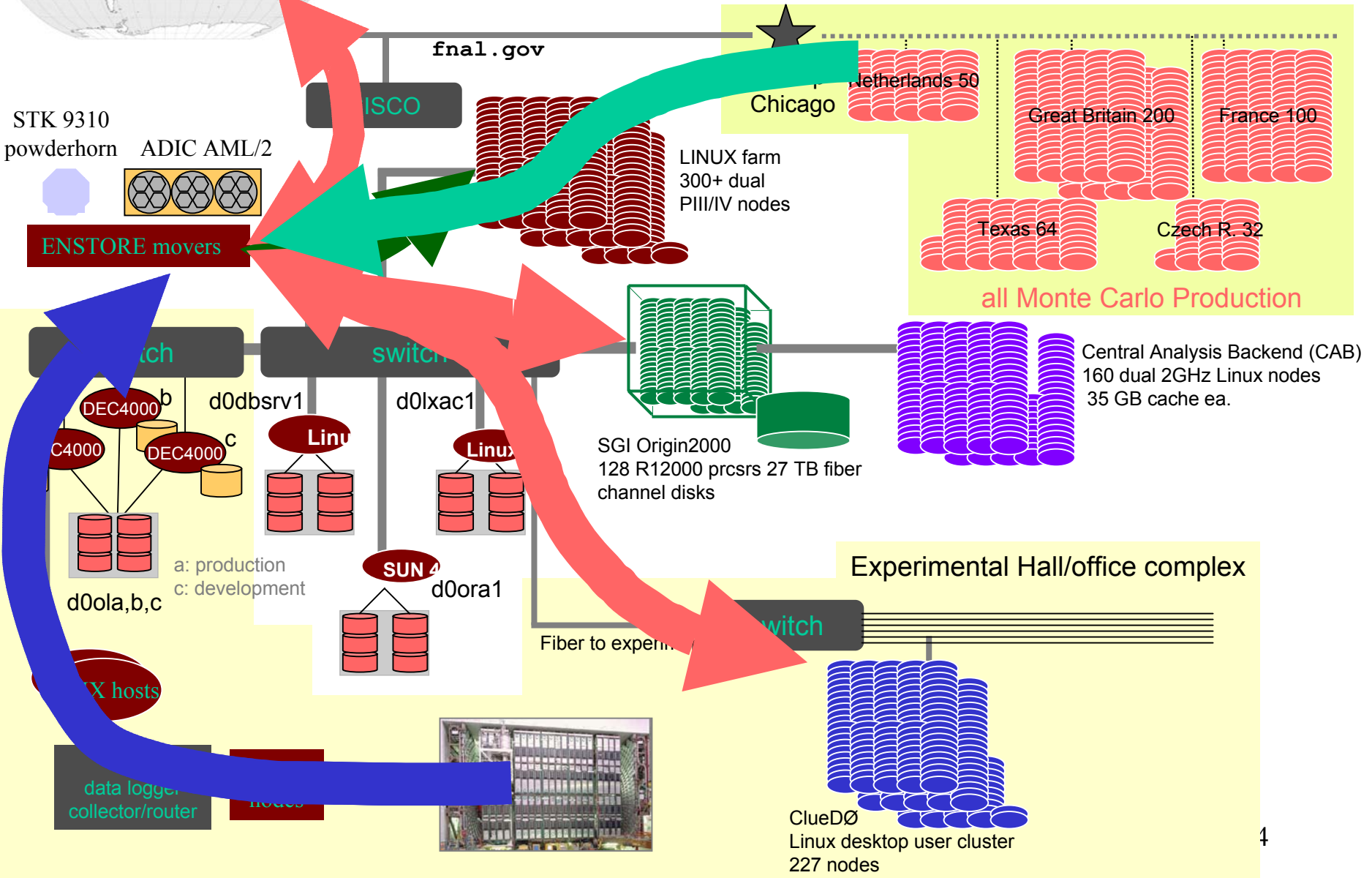
DØ Data Handling is a Successful Worldwide Effort

Thanks to the efforts of many people

- **The SAM team at FNAL: Andrew Baranovski, Diana Bonham, Lauri Carpenter, Lee Lueking, Wyatt Merritt*, Carmenita Moore, Igor Terekhov, Julie Trumbo, Sinisa Veseli, Matthew Vranicar, Stephen P. White, Kin Yip (BNL). (*project co-lead with Jeff Tseng from CDF)**

- **Major contributions from –Amber Boehnlein (D0 Offline Computing Leader), Iain Bertram (Lancaster), Chip Brock (MSU), Jianming Qian (UM), Rod Walker (IC), Vicky White (FNAL)**

- **CD Computing and Communication Fabric Dept. (CCF), in particular the Enstore Team, and Farms Support Group**

- **CD Core Computing Support (CCS) Database Support Group (DSG) and Computing and Engineering for Physics Applications (CEPA) Database Applications Group (DBS) for database support**

- **CD D0 Department, D0 Operations Team at Fermilab**

- **CAB and CLueD0 administrators and support teams**

- **Sam Station Administrators, and SAM Shifters Worldwide**

- **MC production teams: Lancaster UK, Imperial College UK, Prague CZ, U Texas Arlington, Lyon FR, Amsterdam NL,**

- **GridKa Regional Analysis Center at Karlsruhe, Germany: Daniel Wicke, Christian Schmitt, Christian Zeitnitz**

DØ computing/data handling/database architecture

SAM

fnal.gov

ISCO

STK 9310 powderhorn

ADIC AML/2

ENSTORE movers

Chicago

Netherlands 50

Great Britain 200   France 100

Texas 64   Czech R. 32

LINUX farm
300+ dual
PIII/IV nodes

all Monte Carlo Production

switch

d0dbsrv1   d0lxac1

DEC4000   b   c

EC4000   DEC4000

Linu   Linux

Central Analysis Backend (CAB)
160 dual 2GHz Linux nodes
35 GB cache ea.

SGI Origin2000
128 R12000 prcsrs 27 TB fiber
channel disks

a: production
c: development

SUN 4   d0ora1

d0ola,b,c

IX hosts

data logger
collector/router   nodes

Fiber to experim

switch

Experimental Hall/office complex

ClueDØ
Linux desktop user cluster
227 nodes

4

SAM Data Management System

- SAM is <u>S</u>equential data <u>A</u>ccess via <u>M</u>eta-data
- Est. 1997

- Flexible and scalable distributed model
- Field hardened code
- Reliable and Fault Tolerant
- Adapters for many batch systems: LSF, PBS, Condor, FBS
- Adapters for mass storage systems: Enstore, (HPSS, and others planned)
- Adapters for Transfer Protocols: cp, rcp, scp, encp, bbftp, GridFTP.
- Useful in many cluster computing environments: SMP w. compute servers, Desktop, private network (PN), NFS shared disk,…
- Ubiquitous for DØ users

SAM Station – 1. Collection of SAM servers which manage data delivery for a node or cluster 2. The node or cluster hardware itself

Overview of DØ Data Handling

Integrated Files Consumed vs Month (DØ)

4.0 M Files Consumed
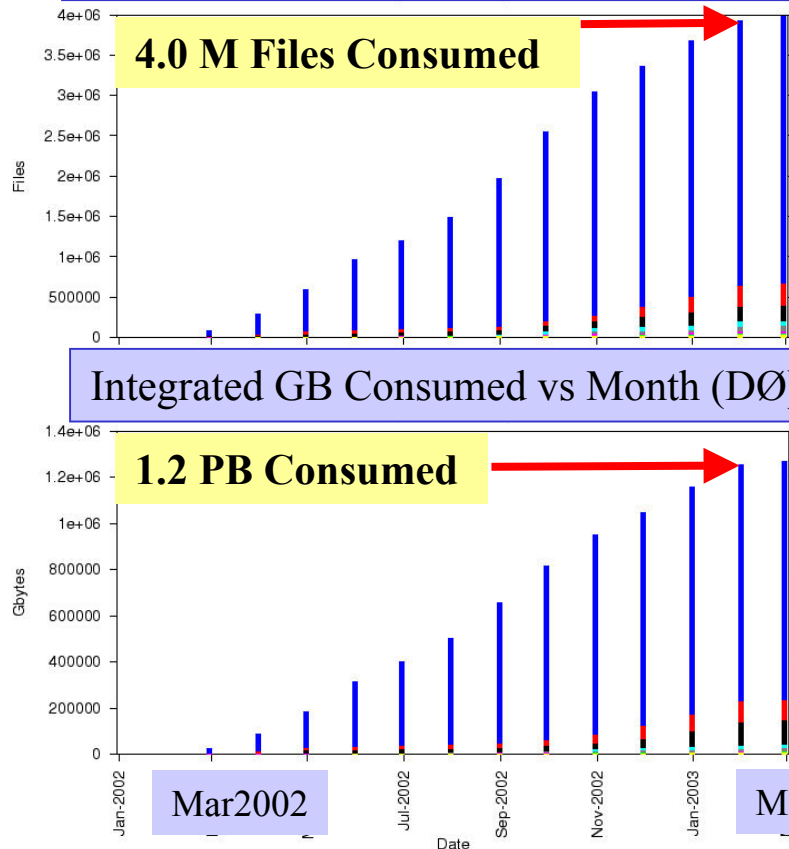
Integrated GB Consumed vs Month (DØ)

1.2 PB Consumed

Mar2002   Mar2003

Station	
central-analysis	cab
fnal-farm	clued0
	d0karlsruhe
	imperial-test
	triviaal
	other
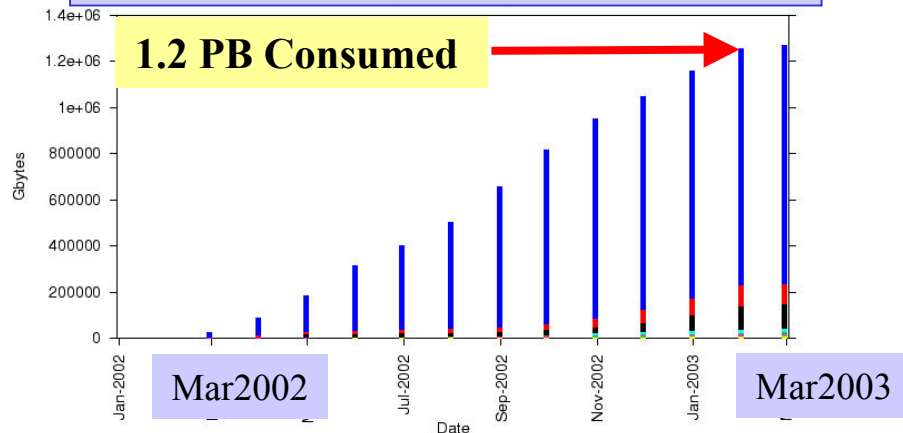
Summary of DØ Data Handling

Registered Users	600
Number of SAM Stations	56
Registered Nodes	900
Total Disk Cache	40 TB
Number Files - physical	1.2M
Number Files - virtual	0.5M
Robotic Tape Storage	305 TB

★ Regional Center
● Analysis site
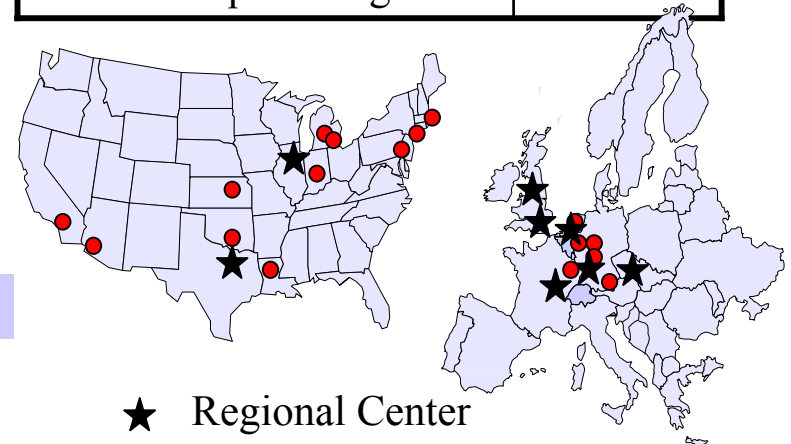
Data In and out of Enstore

(robotic tape storage) Daily Feb 14 to Mar 15

enstore
Incoming(+) / Outgoing(-)
MBytes Transferred Per Day

1.3 TB incoming

2.5 TB outgoing

Stations:
- fnal-farm
- ral-analysis
- datalogger-d0olc
- central-router
- cab
- clued0
- umdzero
- other
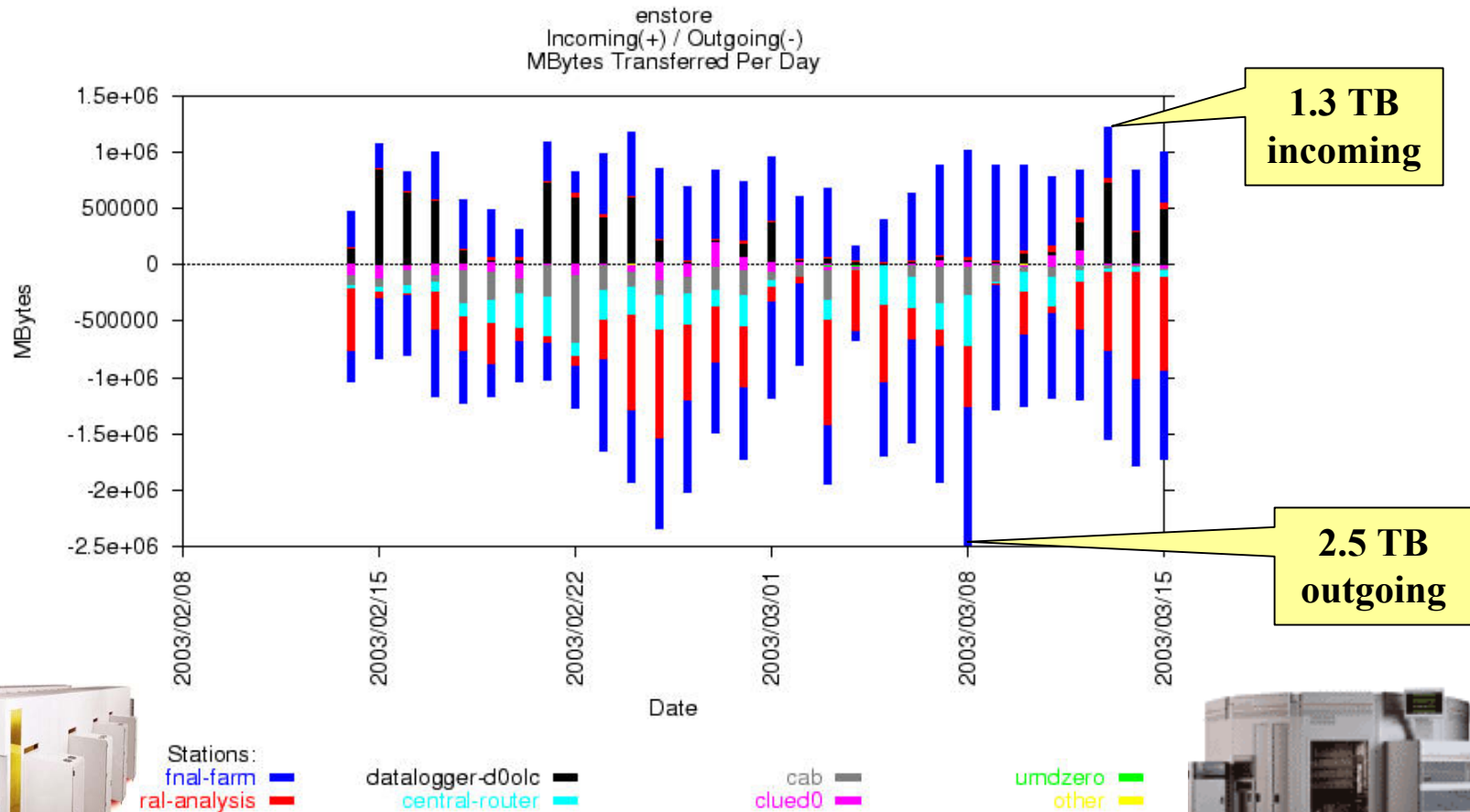
Enstore Talk, Cat. 3 Tuesday

DØ SAM Station Summary

Name	Location	Nodes/cpu	Cache	Use/comments
Central-analysis	FNAL	128 SMP*, SGI Origin 2000	14 TB	Analysis & D0 code development
CAB (CA Backend)	FNAL	16 dual 1 GHz + 160 dual 1.8 GHz	6.2 TB	Analysis and general purpose
FNAL-Farm	FNAL	100 dual 0.5-1.0 GHz +240 dual 1.8 GHz	3.2 TB	Reconstruction
CLueD0	FNAL	50 mixed PIII, AMD. (may grow >200)	2 TB	User desktop, General analysis
D0karlsruhe (GridKa)	Karlsruhe, Germany	1 dual 1.3 GHz gateway, >160 dual PIII & Xeon	3 TB NFS shared	General/Workers on PN. Shared facility
Nijmegen	Nijmegen, Netherlands	1 dual 1.8 GHz gateway, 6 x dual 930MHz	1 TB	Analysis/ workers on PN
Many Others > 4 dozen	Worldwide	Mostly dual PIII, Xeon, and AMD XP		MC production, gen. analysis, testing

*IRIX, all others are Linux

Station Stats: GB Consumed
Daily Feb 14 – Mar 15
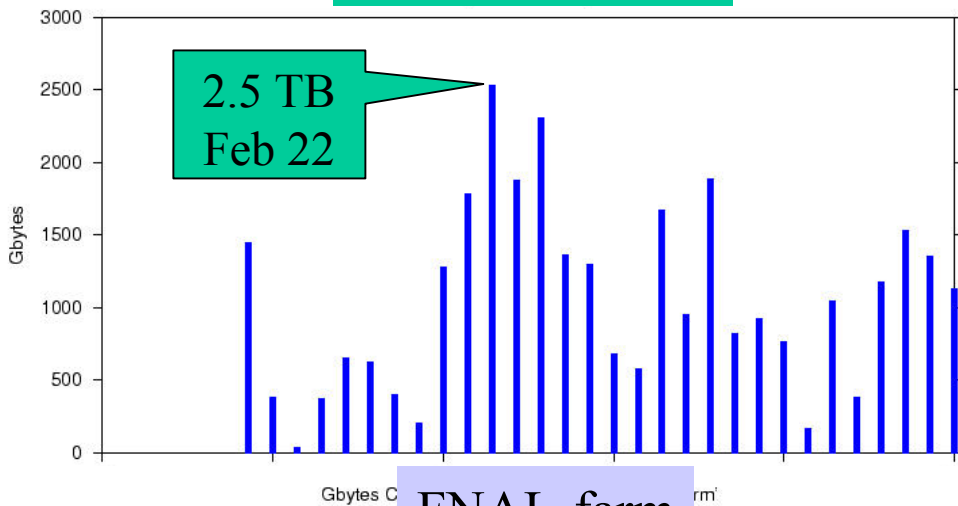
Central-Analysis

2.5 TB
Feb 22

ClueD0

270 GB
Feb 17

FNAL-farm

1.1 TB
Mar 6

CAB

>1.6 TB
Feb 28

Central-Analysis

1 TB
Feb 22

Delivered t[...]

2.5 TB
Feb 22
Consumed

ClueD0

150 GB
Feb 17

270 GB
Feb 17
Consumed

FNAL-farm

1.2 TB
Mar 6

1.1 TB
Mar 6
Consumed

CAB

600 GB
Feb 28

1.6 TB
Feb 28
Consumed
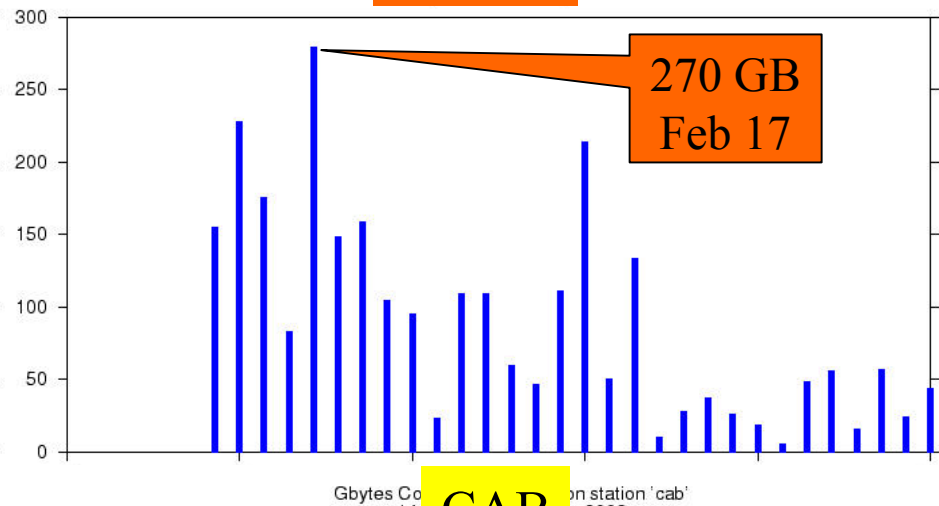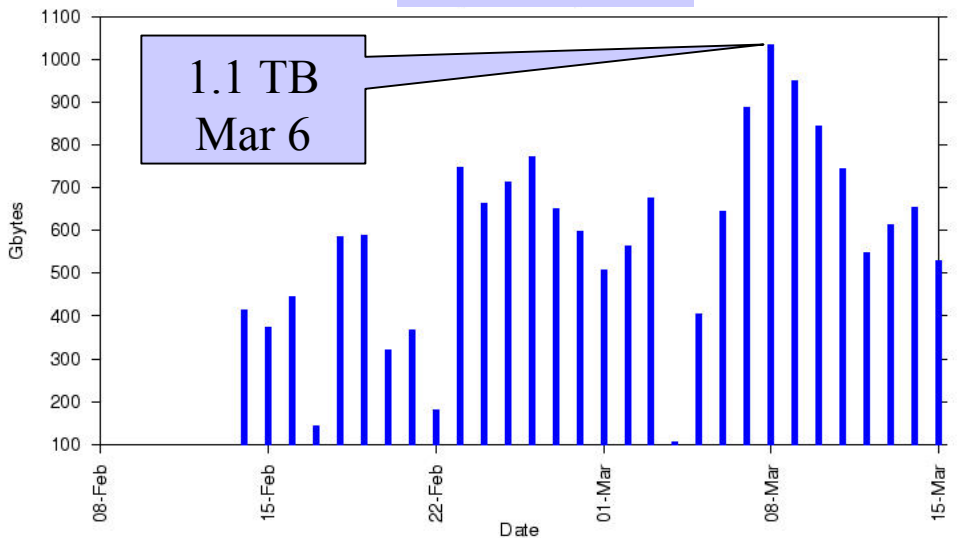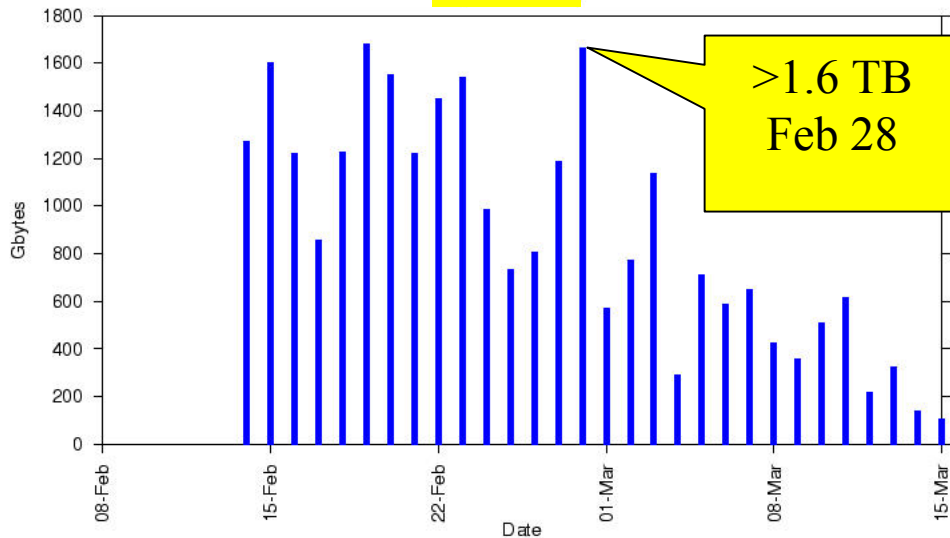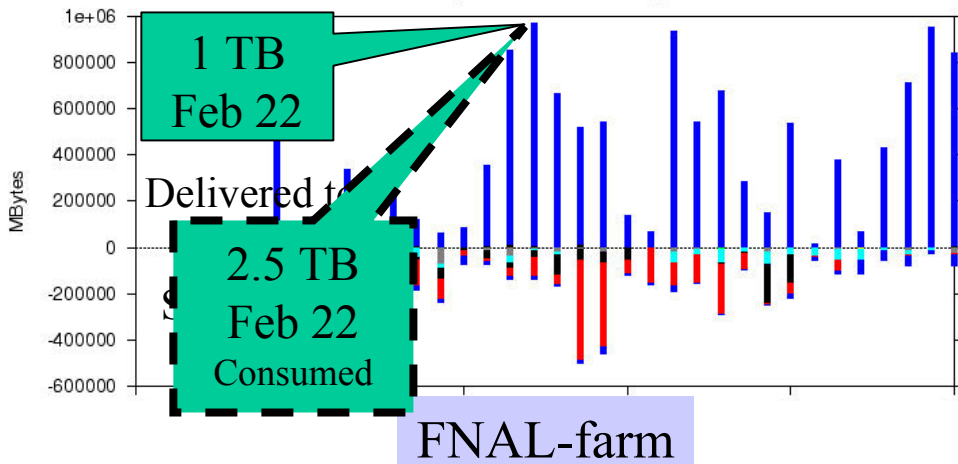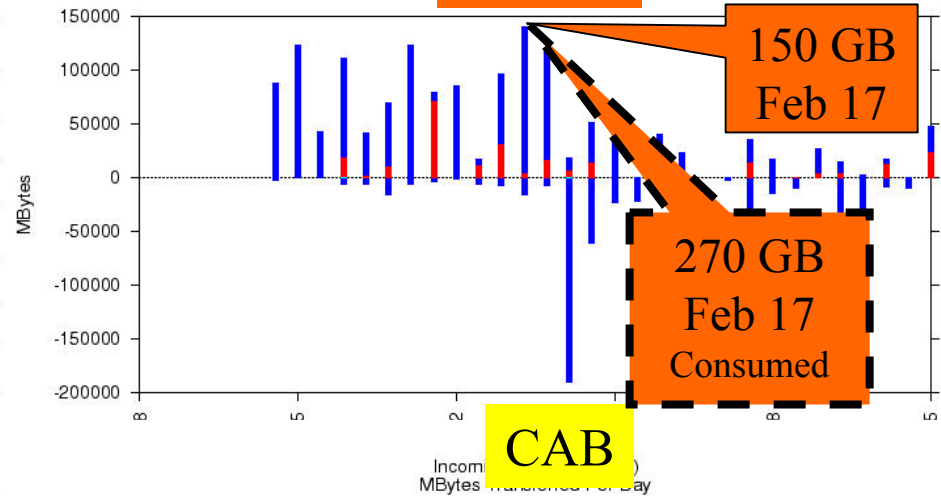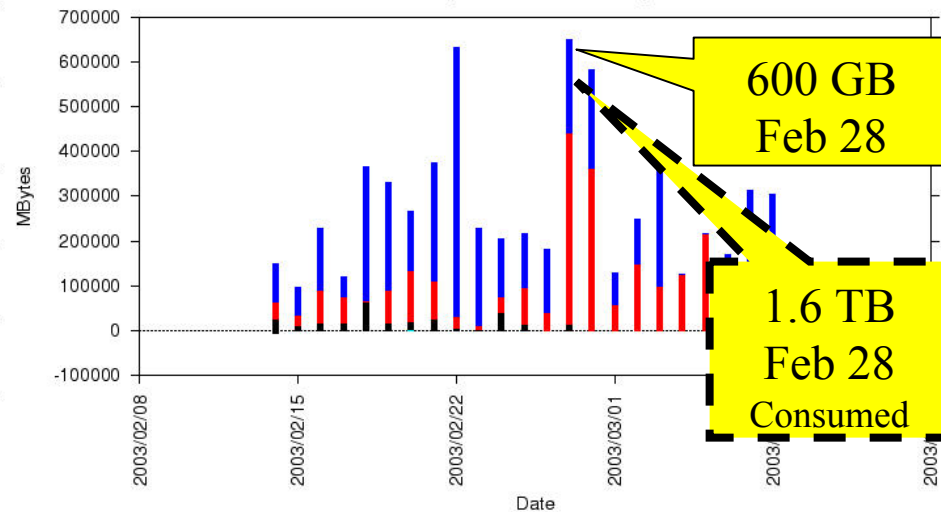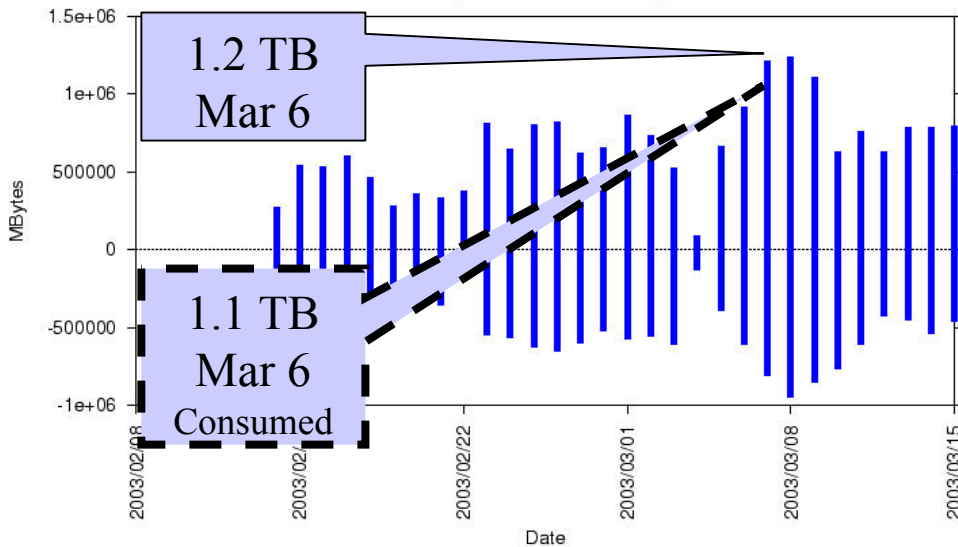
Stations:
enstore
clued0

Stations:
enstore   central-router   uta-hep   d0karlsruhe
central-analysis   princeton-d0   imperial-test   umdzero

FNAL-farm Station and CAB CPU Utilization

Feb 14 – March 15

600 CPUs
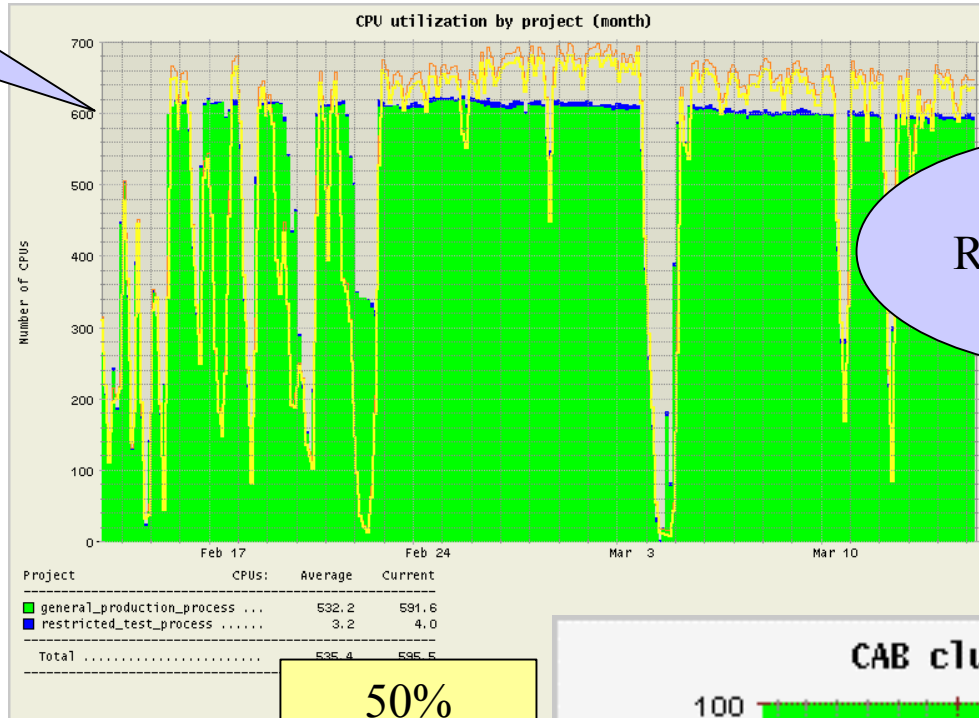
FNAL-farm Reconstruction Farm

Also, see CLued0 Talk in Section 3, Monday

CAB Usage will increase dramatically in the coming months

50% Utilization
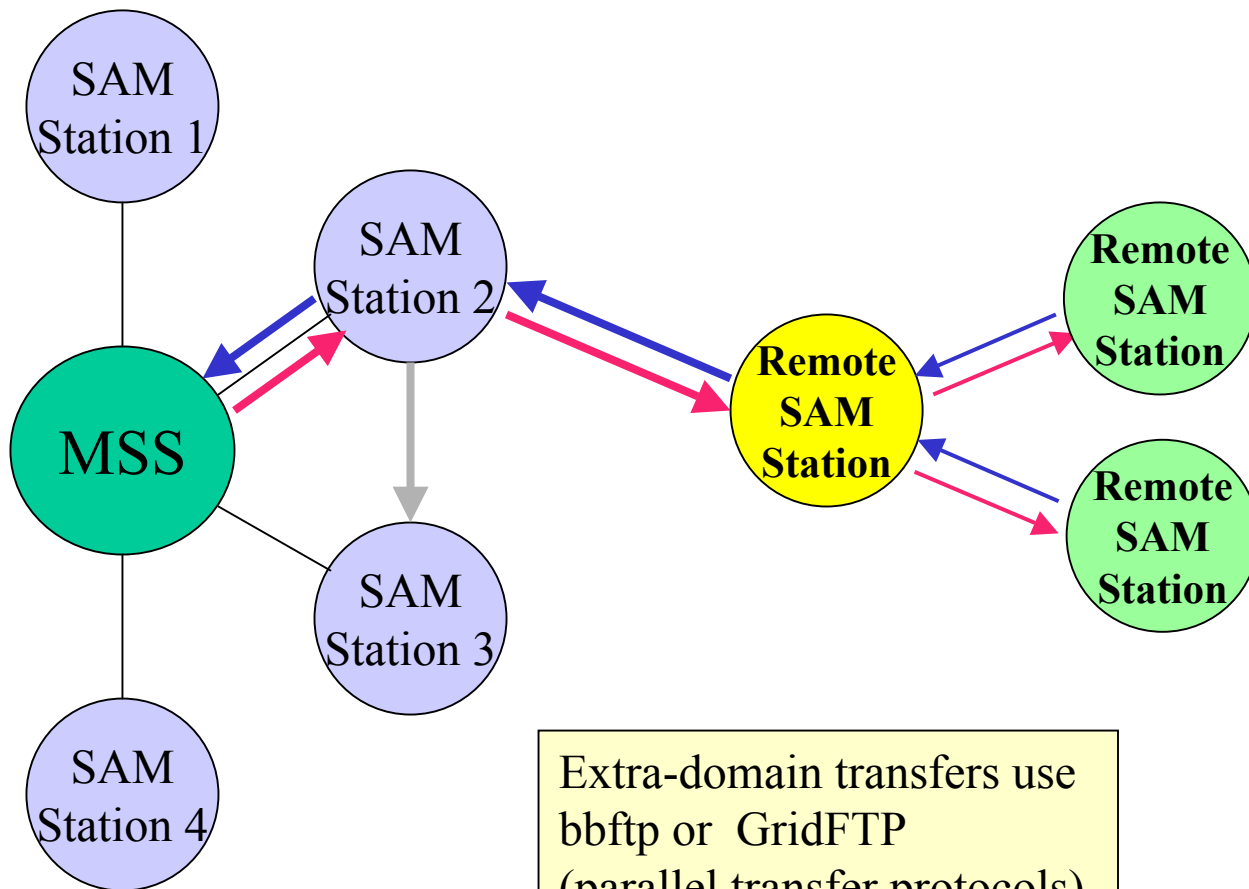
Central-Analysis Backend Compute Servers

CPU utilization by project (month)

Project	CPUs:	Average	Current
general_production_process ...		532.2	591.6
restricted_test_process		3.2	4.0
Total		535.4	595.5

CAB cluster CPU last month

User CPU   Nice CPU   System CPU   Idle CPU
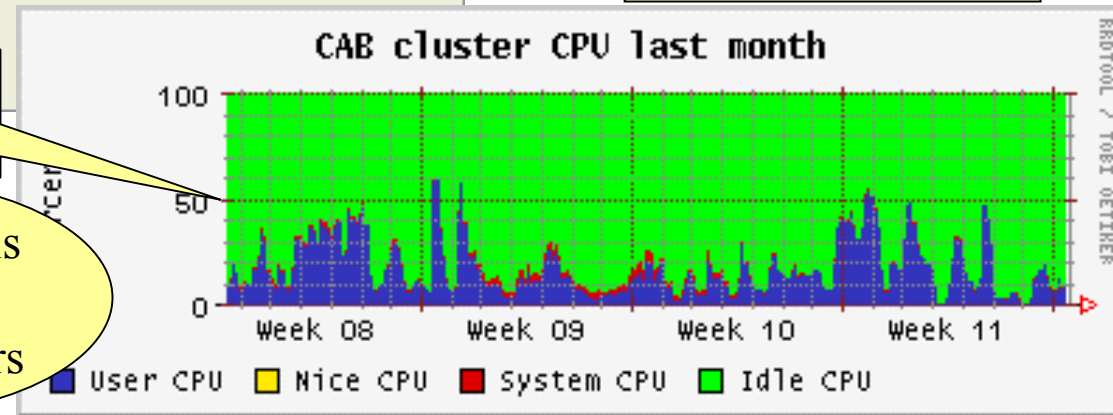
Data to and from Remote Sites

Station Configuration
- Replica location
  - Prefer
  - Avoid
- Forwarding
  - File stores can be forwarded through other stations
- Routing
 - Routes for file transfers are configurable

SAM Station 1

SAM Station 2

MSS

SAM Station 3

SAM Station 4

Remote SAM Station

Remote SAM Station

Remote SAM Station

Extra-domain transfers use bbftp or GridFTP (parallel transfer protocols)

DØ Karlsruhe Station at GridKa

Monthly Thumbnail Data Moved to GridKa

1.2 TB in Nov 2002

Cumulative Thumbnail Data Moved to GridKa

5.5 TB since June 2002
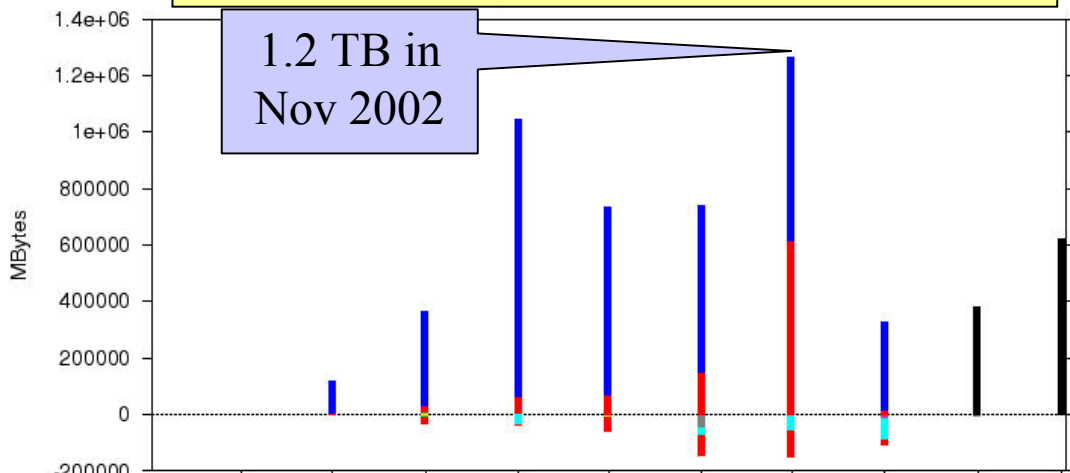
The GridKa SAM Station uses shared cache config. with workers on a private network

This is our first Regional Analysis Center (RAC). See DØ RAC Concept talk, Category 1, Tuesday.
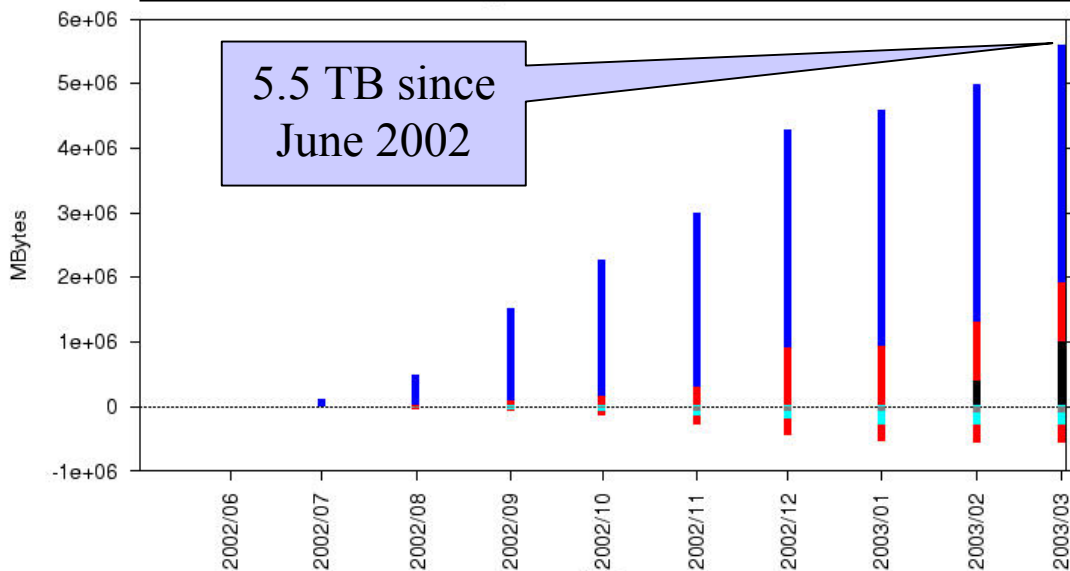
- Resource Overview:
  - **Compute: 95 x dual PIII 1.2GHz, 68 x dual Xeon 2.2 GHz. D0 requested 6%. (updates in April)**
  - **Storage: D0 has 5.2 TB cache. Use of % of ~100TB MSS. (updates in April)**
  - **Network: 100Mb connection available to users.**
  - **Configuration: SAM w/ shared disk cache, private network, firewall restrictions, OpenPBS, Redhat 7.2, k 2.418, D0 software installed.**

SAM Shift Stats

Weekly SAM Problem Resolution

Problems Resolved (y-axis: 0, 5, 10, 15, 20, 25, 30)

Legend: ■ Expert ■ Shifter

X-axis (Date): 11/25/2002, 12/9/2002, 12/23/2002, 1/6/2003, 1/20/2003, 2/3/2003, 2/17/2003

Shifter Time Zones

Shifters (y-axis: 0, 2, 4, 6, 8, 10, 12)

X-axis (Time Zone): GMT +5.5, GMT +1, GMT, GMT -3, GMT -5, GMT -6

Overview
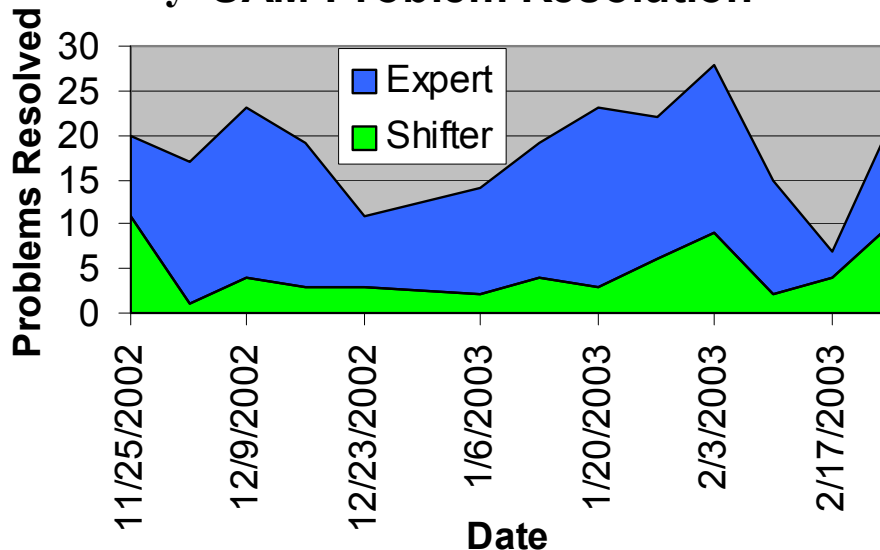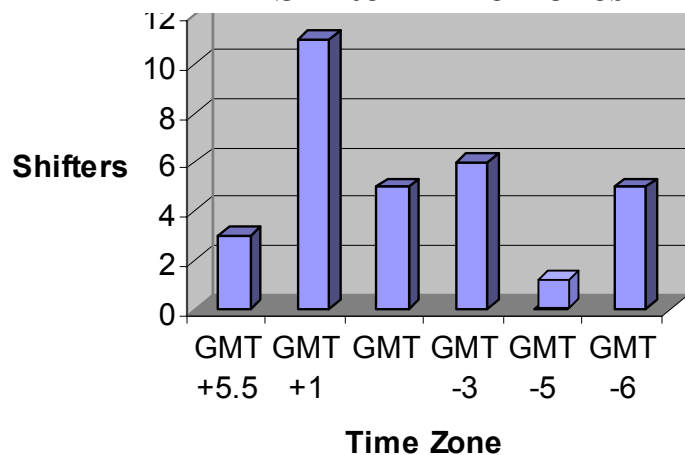
- In operation since summer 2001
- Kin Yip (BNL) is current shift coordinator
- 27 general shifters from 5 time zones.
- 7 expert shifters at FNAL
- Experts still carry much of the load.
- Problems range from simple user questions to installation issues, hardware, network, bugs...

SAM Problem Resolution

Percentage (y-axis: 0%, 20%, 40%, 60%, 80%, 100%)

Legend: ■ Expert ■ Shifter

X-axis (Date): 11/25/2002, 12/9/2002, 12/23/2002, 1/6/2003, 1/20/2003, 2/3/2003, 2/17/2003
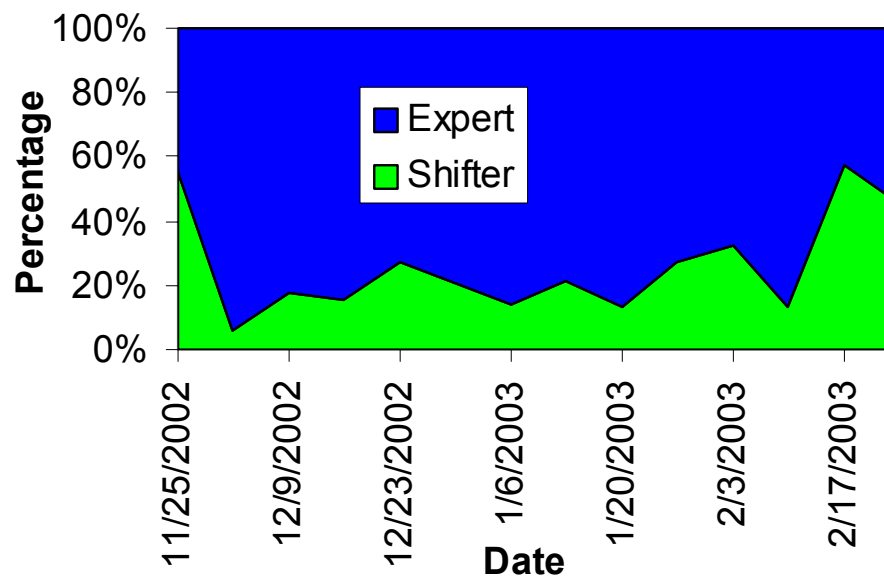
SAM At A Glance

D0 Production Environment
This page generated on 16 Mar 2003 13:29:47
[Other Sam Diagnostics](#)

SAM Stations:

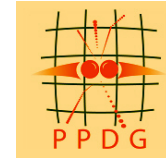|		Host	Version	Up Since
Monitor Level: Critical				
🟢	datalogger-d0olb	d0olb-gb-1.fnal.gov	v4_2_1_12	07 Mar 2003 17:13:43
🟢	datalogger-d0olc	d0olc-gb-1.fnal.gov	v4_2_1_12	07 Mar 2003 17:13:09
Monitor Level: High				
🟢	central-analysis	d0mino.fnal.gov	v4_2_1_24	14 Mar 2003 16:47:29
🟢	fnal-farm	d0bbin.fnal.gov	v4_2_1_31	05 Mar 2003 11:54:19
Monitor Level: Normal				
🟢	cab	d0mino-sam.fnal.gov	v4_2_1_40	14 Mar 2003 13:48:19
🔴	cab-test	no smaster/fss		
🟢	ccin2p3-analysis	ccd0.in2p3.fr	v4_2_1_24	10 Mar 2003 09:53:54
🟢	ccin2p3-production	ccd0.in2p3.fr	v4_2_1_24	10 Mar 2003 09:53:57
🟢	central-router	d0mino.fnal.gov	v4_2_1_39	04 Mar 2003 23:04:46
🟢	cinvestav-station	higgs.fis.cinvestav.mx	v4_2_1_31	24 Feb 2003 11:59:47
🟢	clued0	flotsam-clued0.fnal.gov	v4_2_1_40	07 Mar 2003 11:43:26
🟢	common-archive	d0ora1.fnal.gov	v4_2_1_20	25 Feb 2003 08:30:34
🔴	csuf_hep2-station	no smaster/fss		
🔴	d0-ires			
🟢	d0-umich	dzero2.engin.umich.edu	v4_2_1_35	12 Mar 2003 08:33:51
🟢	d0indiana	d0ds.physics.indiana.edu	v4_2_1_24	04 Mar 2003 16:39:45
🔴	d0kahrlsruhe			
🟢	d0karlsruhe	d0.fzk.de	v4_2_1_39	10 Mar 2003 05:13:19
🔴	d0nevis-station	no smaster/fss		
🟢	hoeve	schuur.nikhef.nl	v4_2_0_3	05 Feb 2003 05:49:28
🟡	imperial-lesc	saturn.icpc.doc.ic.ac.uk	unknown	unknown

Challenges

- Getting SAM to meet the needs of DØ in the many configurations is and has been an enormous challenge. Some examples include…
  - File corruption issues. Solved with CRC.
  - Preemptive distributed caching is prone to race conditions and log jams. These have been solved.
  - Private networks sometimes require "border" naming services. This is understood.
  - Additional simplicity and generality are provided in the NFS shared cache configuration, at the price of scalability (star configuration). This works.
  - Installation procedures for the station servers have been quite complex. They are improving and we plan to soon have "push button" and even "opportunistic deployment" installs.
  - Lots of details with opening ports on firewalls, OS configurations, registration of new hardware, and so on.
  - Username clashing issues. Moving to GSI and Grid Certificates.
  - Interoperability with many MSS.
  - Network attached files. Sometimes, the file does not need to move to the user.

Stay Tuned for SAM-Grid
The best is yet to come…

Summary

- The DØ Data Handling operation is a complex system involving a worldwide network of infrastructure and support.

- SAM provides flexible data management solutions for many hardware configurations, including clusters in private networks, shared NFS cache, and distributed cache. It also provides configurable data routing throughout the install base.

- The software is stable and provides reliable data delivery and management to production systems at FNAL and worldwide. Many challenging problems have been overcome to achieve this goal.

- Support is provided through a small group of experts at FNAL, and a network of shifters throughout the world. Many tools are provided to monitor the system, detect and diagnose problems.

- The system is continually being improved, and additional features are planed as the system moves beyond data handling to complete Grid functionality in the SAM-Grid project (a.k.a. SAM + JIM).