

Dzero, SAM, and SAM-Grid for RunII

Lee Lueking

LCCWS

October 21, 2002

Contents:

- 1. Dzero and SAM Overview**
- 2. SAM-now: Case Studies of DZero clusters**
- 3. SAM-Grid: The Ultimate Goals**



The Dzero Experiment



- D0 Collaboration
 - ◆ 18 Countries; 76 institutions
 - ◆ 500 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, ~ 0.6 PB
- Monte Carlo Data (Run 2a)
 - ◆ 6 remote processing centers
 - ◆ Estimate ~ 300 TB in 2 years.
- Run 2b, starting 2005: > 1 PB/year





What is SAM?



- SAM is Sequential data Access via Meta-data
- Project started in 1997 to handle D0's needs for Run II data system.
- The SAM team includes:
 - ◆ ODS and D0CA: Andrew Baranovski, Diana Bonham, Lauri Loebel-Carpenter, Lee Lueking*, Carmenita Moore, Igor Terekhov, Julie Trumbo, Sinisa Veseli, Matthew Vranicar, Stephen P. White. (*project leader)
 - ◆ Emeritus: Vicky White
 - ◆ In June CDF provided: Randy J. Herber, Rob Kennedy, Art Kreymer, Jeff Tseng* . (project co-lead)
- <http://d0db.fnal.gov/sam>





The SAM Team and Friends

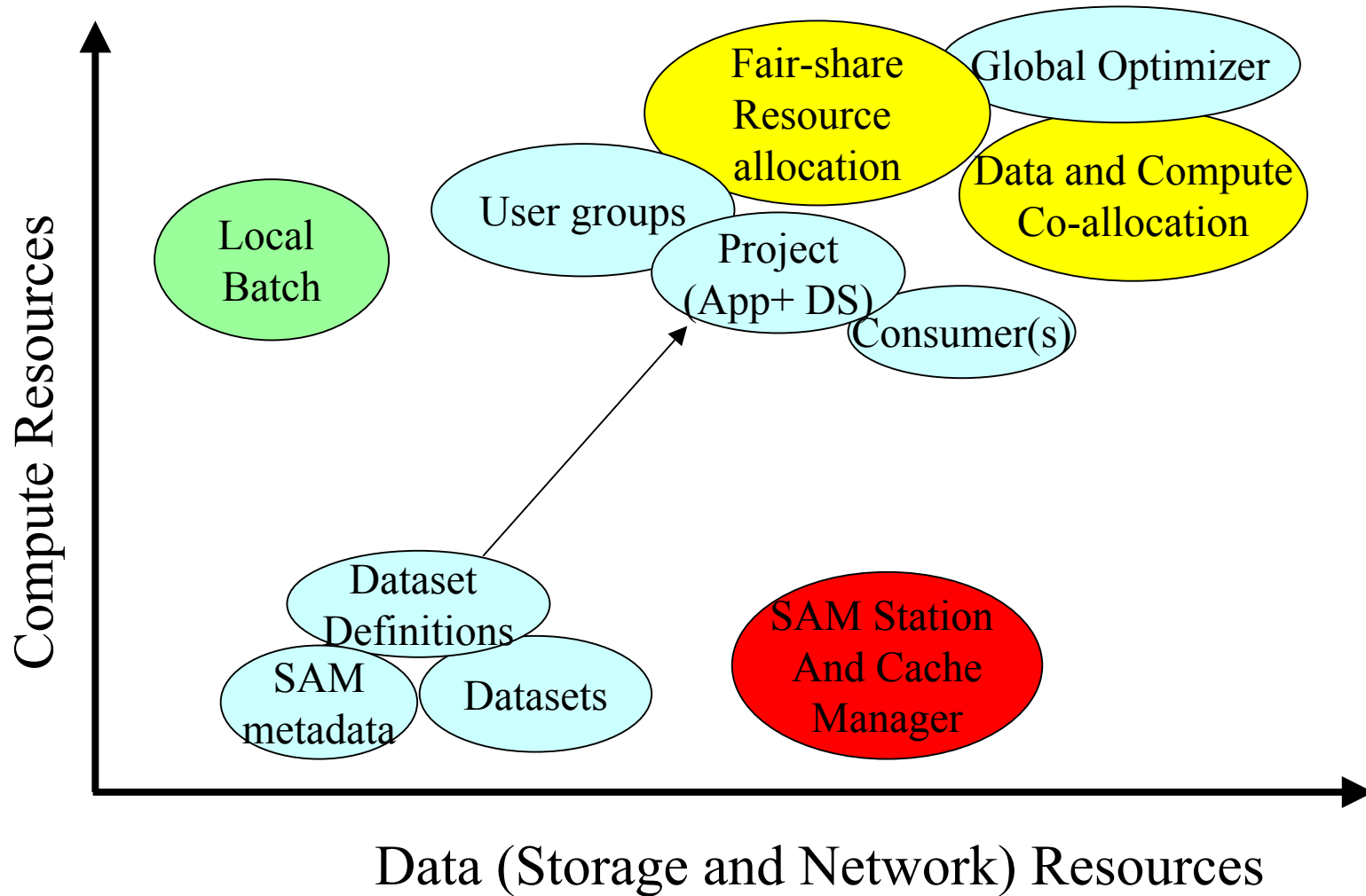


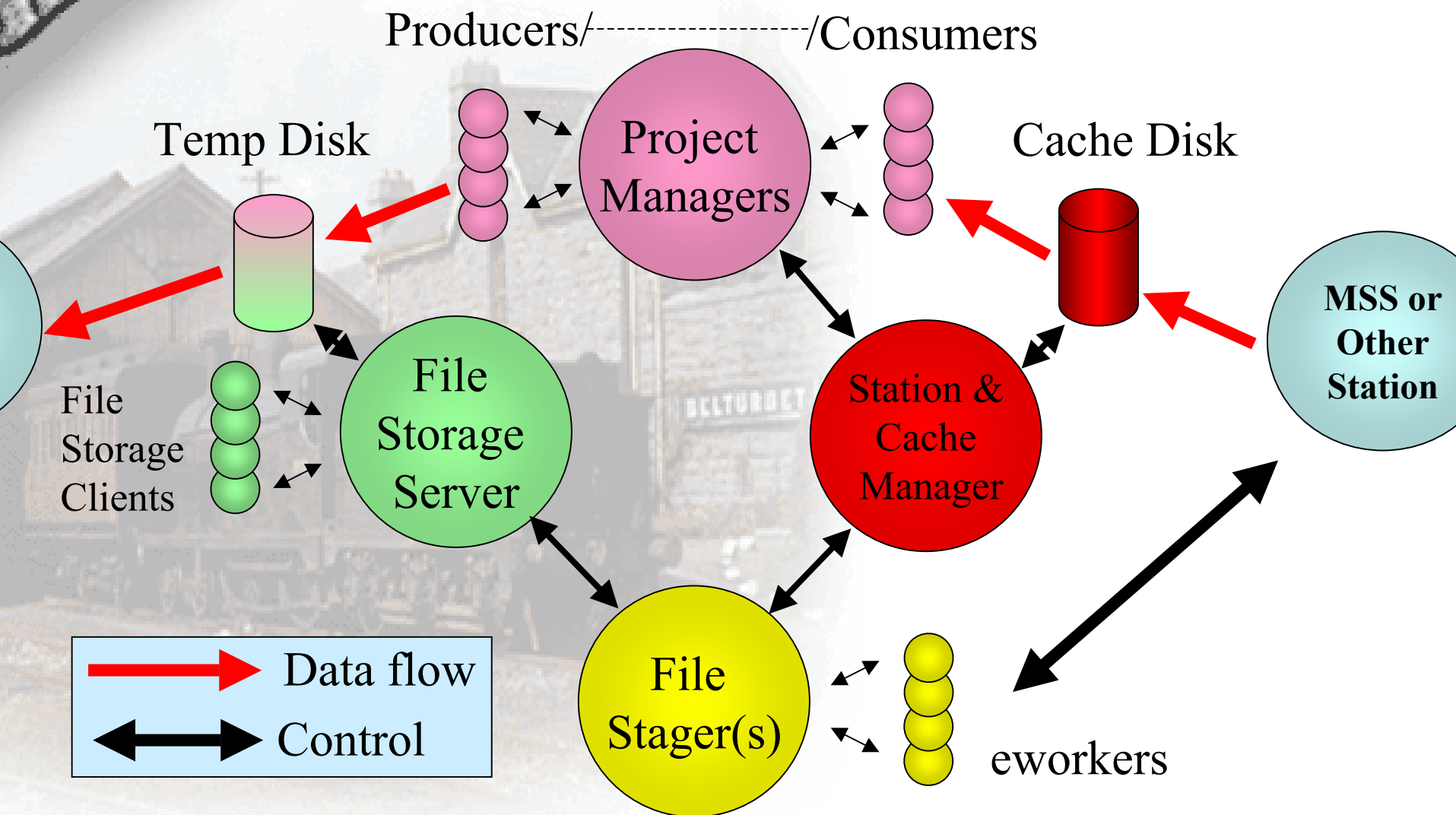
<http://d0db.fnal.gov/sam>





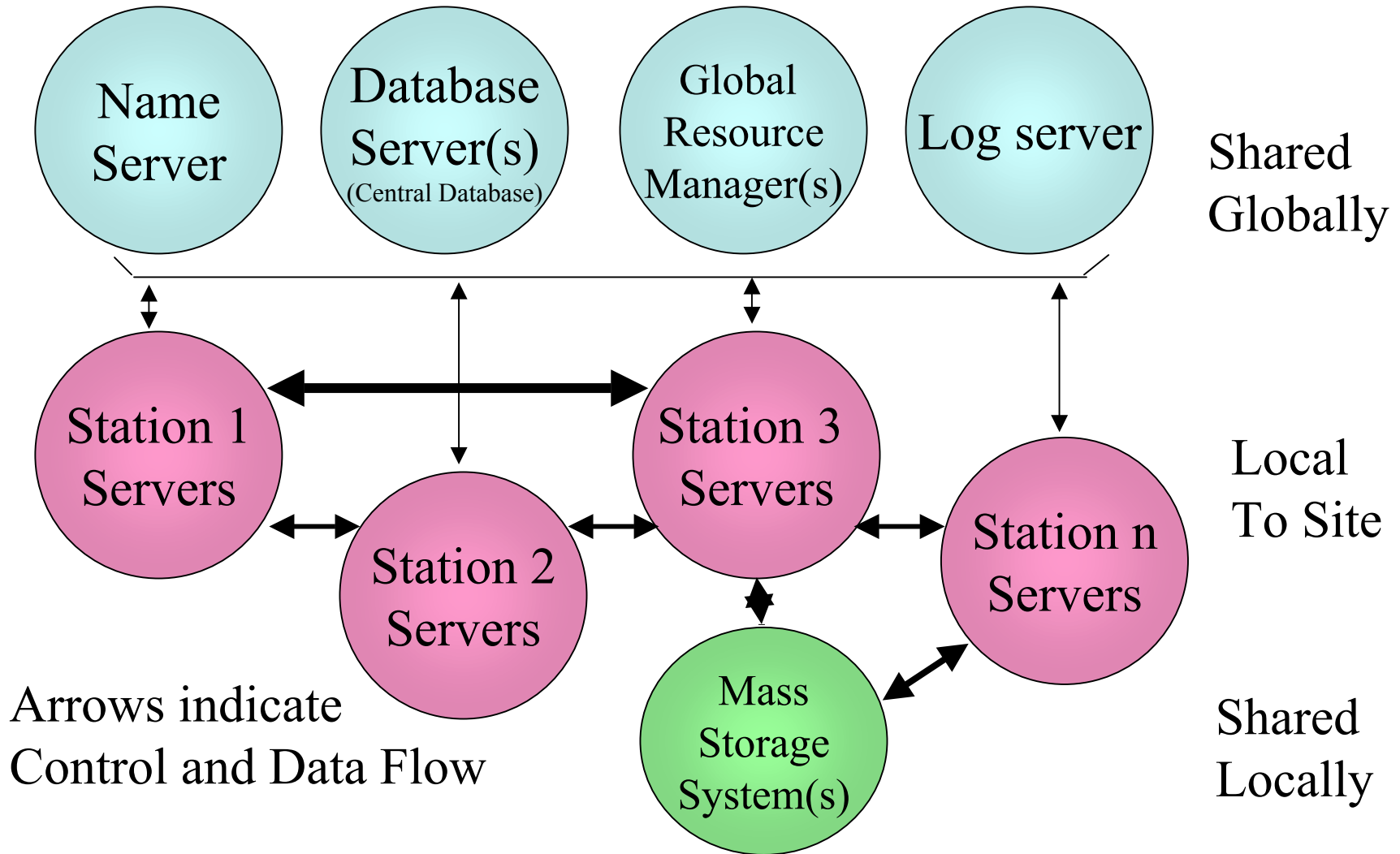
Managing Resources in SAM







SAM as a Distributed System





SAM Features



- Flexible and scalable 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 D0 users





SAM as it is now

SAM-NOW



Station Examples

Name	Location	Nodes/cpu	Cache	Use/comments
Central-analysis	FNAL	176 SMP*, SGI Origin	14 TB	Analysis & D0 code development
CAB (CA Backend)	FNAL	16 dual GHz (+ 160 dual 1.8 GHz)	1 TB	Analysis
FNAL-Farm	FNAL	100 dual mixed (+240 dual 1.8 GHz)	1.3 TB	Reconstruction
CLueD0	FNAL	17 mixed PIII, AMD. (will grow >200)	2 TB	User desktop, General analysis
Nijmegen	Nijmegen, Netherlands	1 dual 1.8 GHz gateway, 6 x dual 930MHz	1 TB	Analysis/ workers on PN
D0karlsruhe	Karlsruhe, Germany	1 dual 1.3 GHz gateway, >160 dual PIII & Xeon	3 TB NFS shared	General/Workers on PN. Shared facility
Many Others > 4 dozen	Worldwide	Mostly dual PIII, as gateway machines		MC production, gen. analysis, testing

*IRIX, all others are Linux



SAM Stations: Central Analysis and Central Analysis Backend



- Network

- Access to Enstore is through D0mino
- Intra-station file transfers “cheap” through a high speed switch

- Job Dispatch

- LSF used for Central Analysis station
- PBS used for Central Analysis Backend (CAB) Station

 **Enstore
Mass Storage**

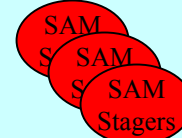
**High Speed
Switch**

**Central-analysis:
Server “D0mino”:**


SGI Origin 2000


- 176 processor
- 6 Gigabit NICs
- 45 GB Memory
- 27 TB Disk


 SAM
Station
Servers

 SAM
S
SAM
S
SAM
Stagers


**14 TB
SAM
Cache**

**Compute
Server
1**


**Compute
Server
2**


**Compute
Server
3**


...

**Compute
Server
N**


**CAB: Tested with 16 dual 1 GHz PIII Backend nodes.
Have 160 additional dual AMD XP 1800 burning in**

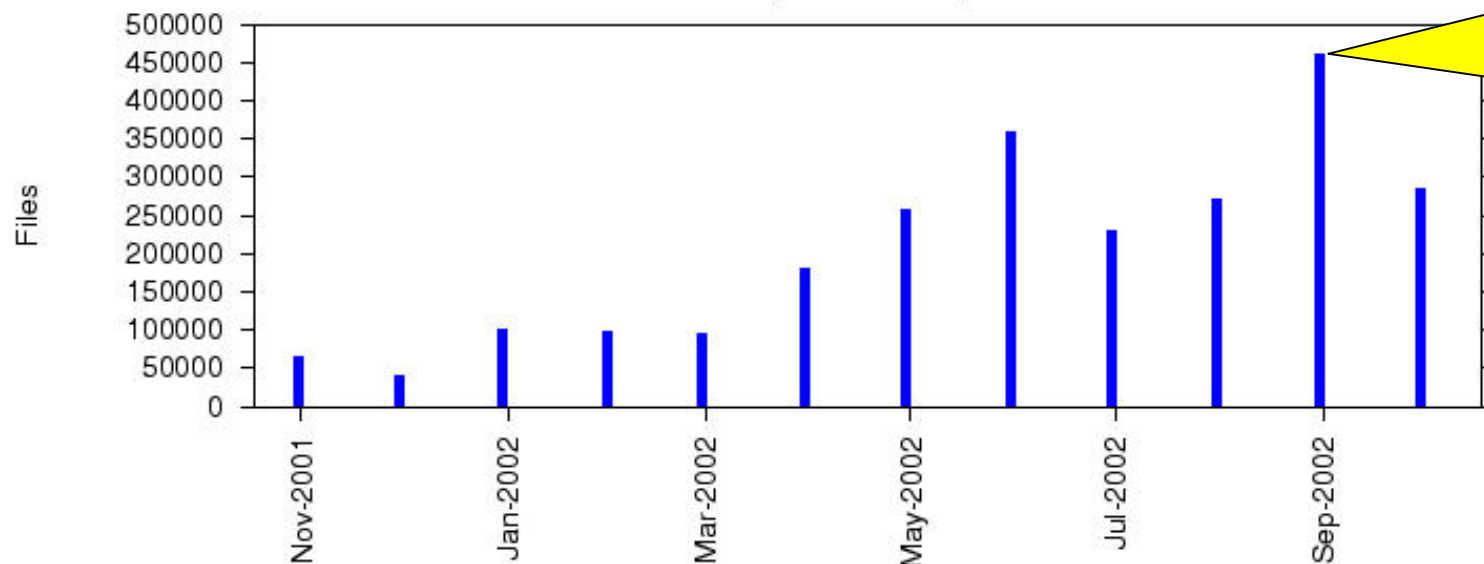
<http://d0db.fnal.gov/sam>

now

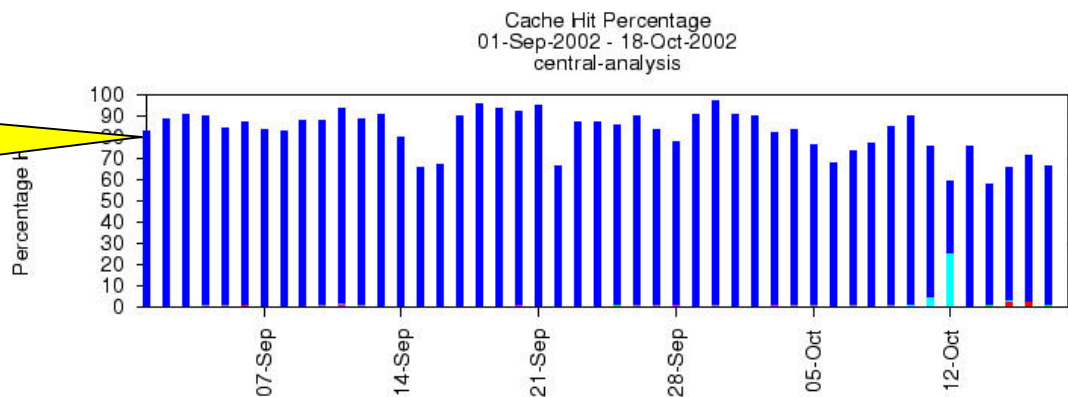


Central-Analysis Stats

Files Consumed on station 'central-analysis'
Year ending 18-Oct-2002
(D0 Production)



In the last month, close to 80% of requested files were already in the cache





SAM Station: Distributed Reconstruction Farm (Fnal-farm)

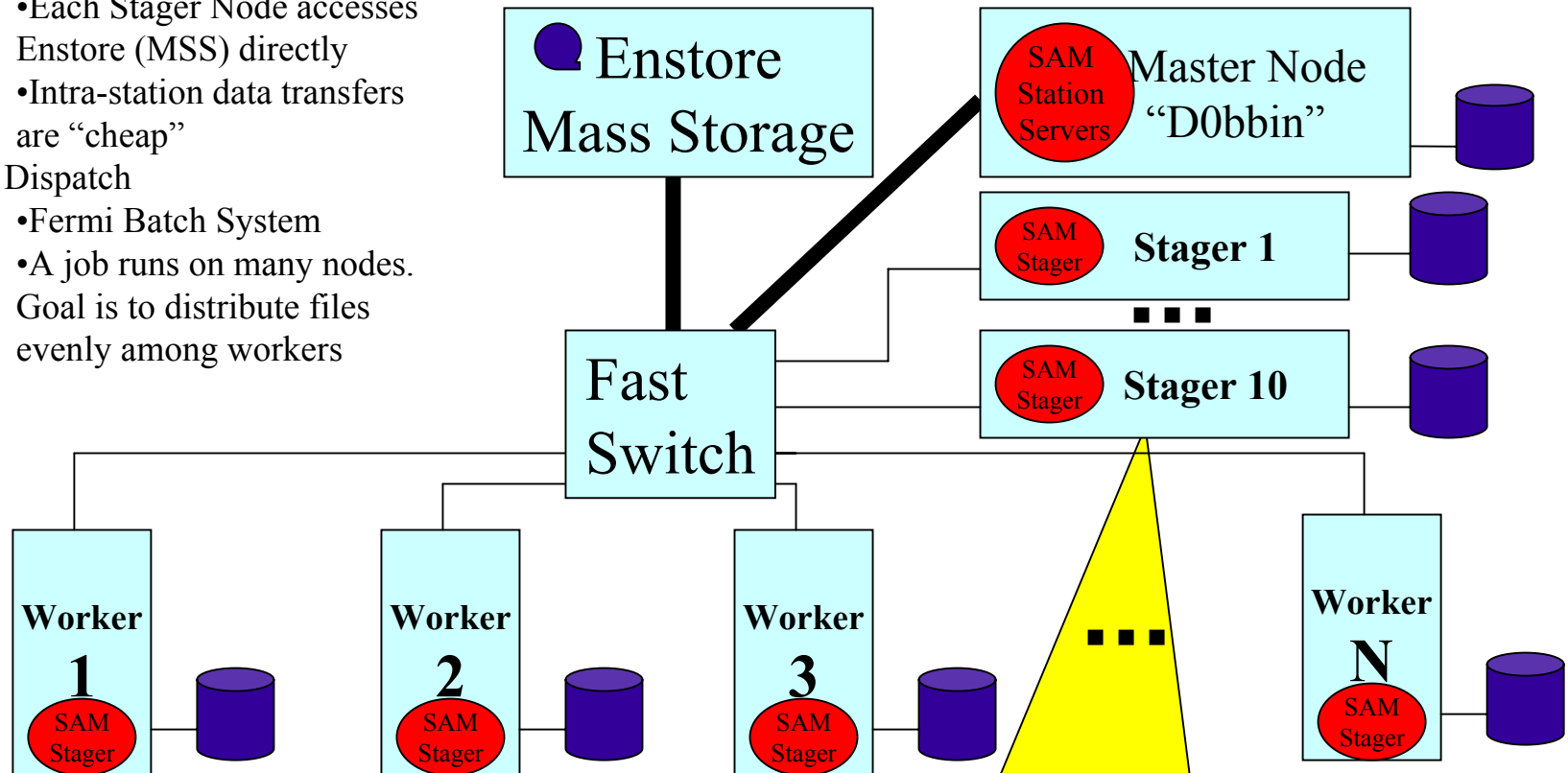


- Network

- Each Stager Node accesses Enstore (MSS) directly
- Intra-station data transfers are “cheap”

- Job Dispatch

- Fermi Batch System
- A job runs on many nodes.
Goal is to distribute files evenly among workers



Data is directed through Stager nodes and replicated to the workers for processing.



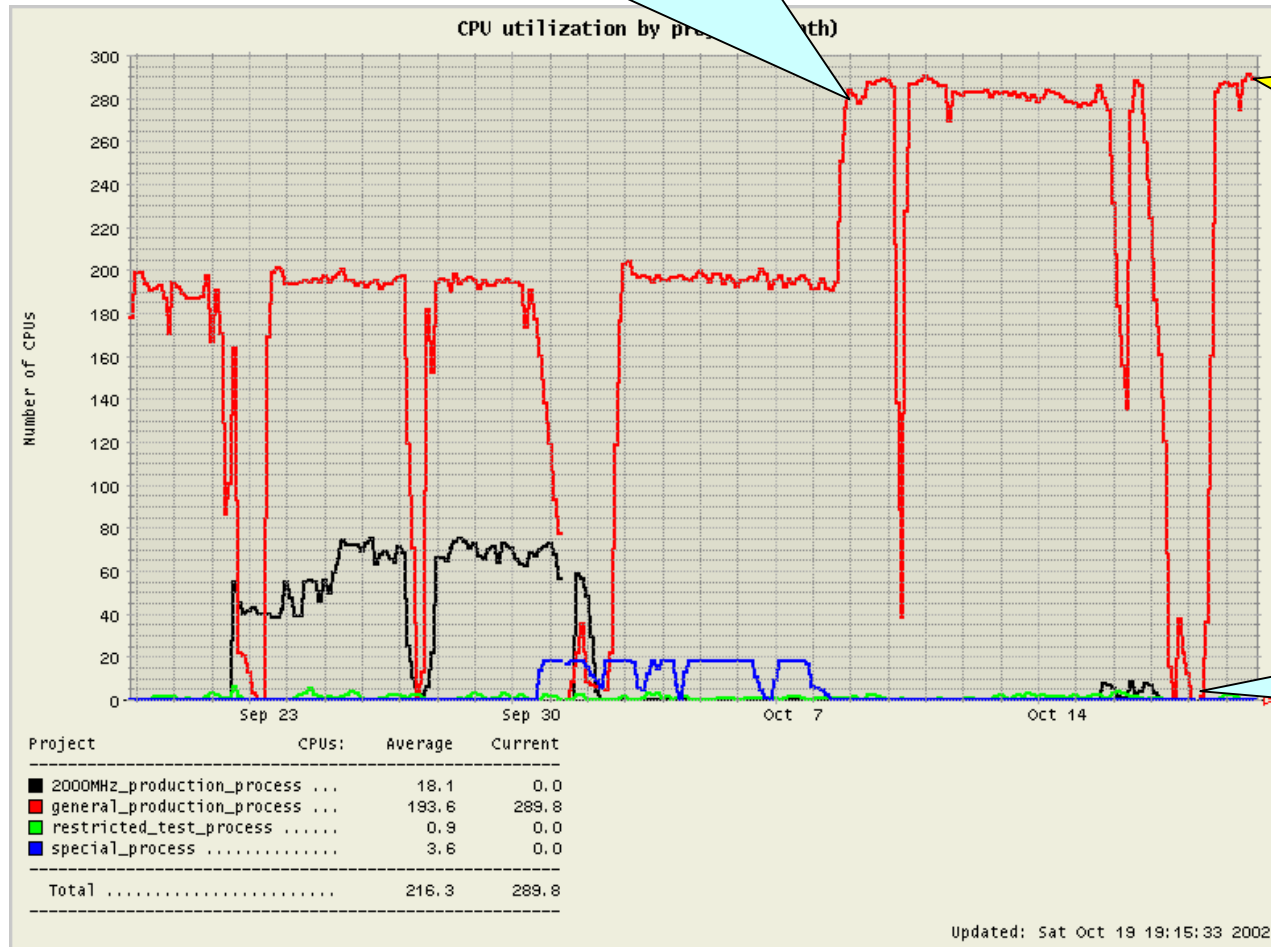


FNAL-Farm Stats



Some of the new nodes are added to burn-in with d0reco

Using 300 production cpu's (150 nodes)



Software issues



SAM Station: Distributed Analysis Cluster (ClueD0)

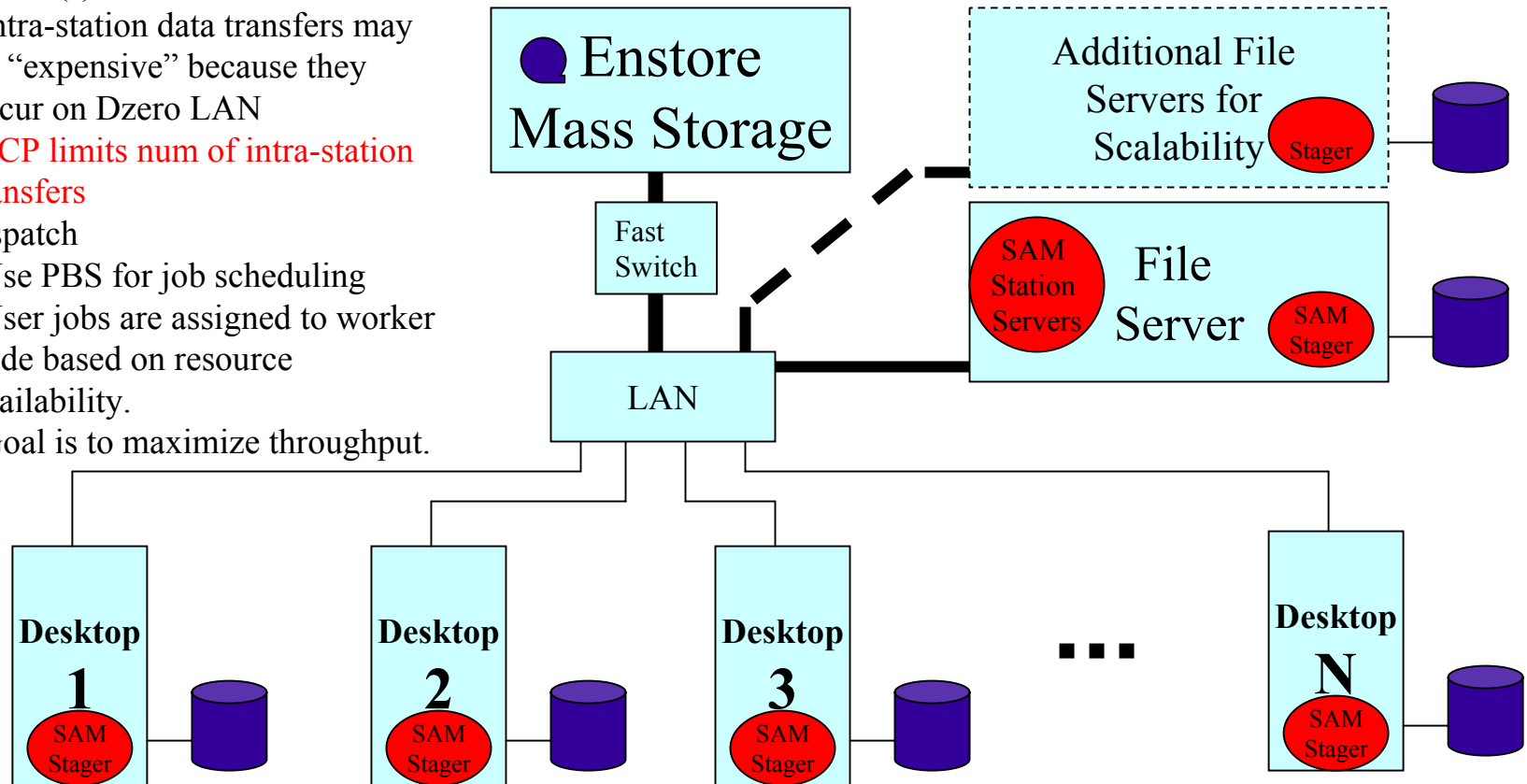


- Network

- Access to MSS limited to File Server(s)
- Intra-station data transfers may be “expensive” because they occur on Dzero LAN
- **FCP limits num of intra-station transfers**

- Job Dispatch

- Use PBS for job scheduling
- User jobs are assigned to worker node based on resource availability.
- Goal is to maximize throughput.



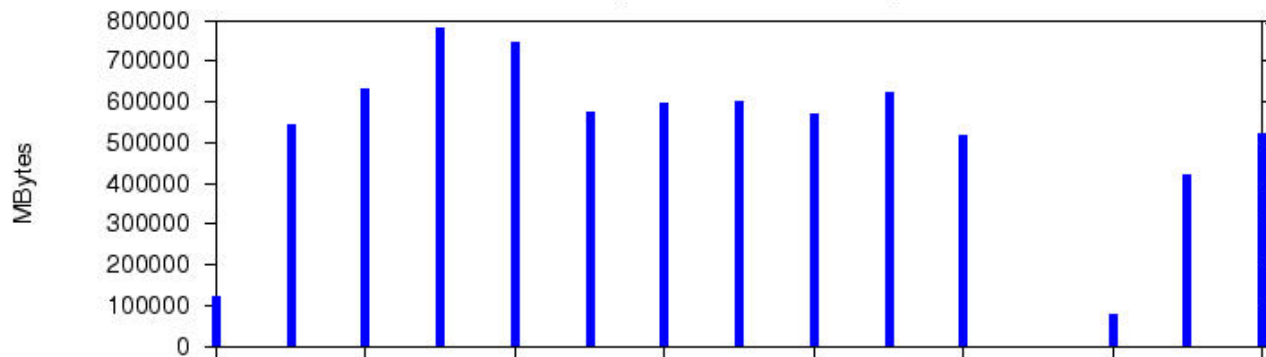


CLueD0 Testing



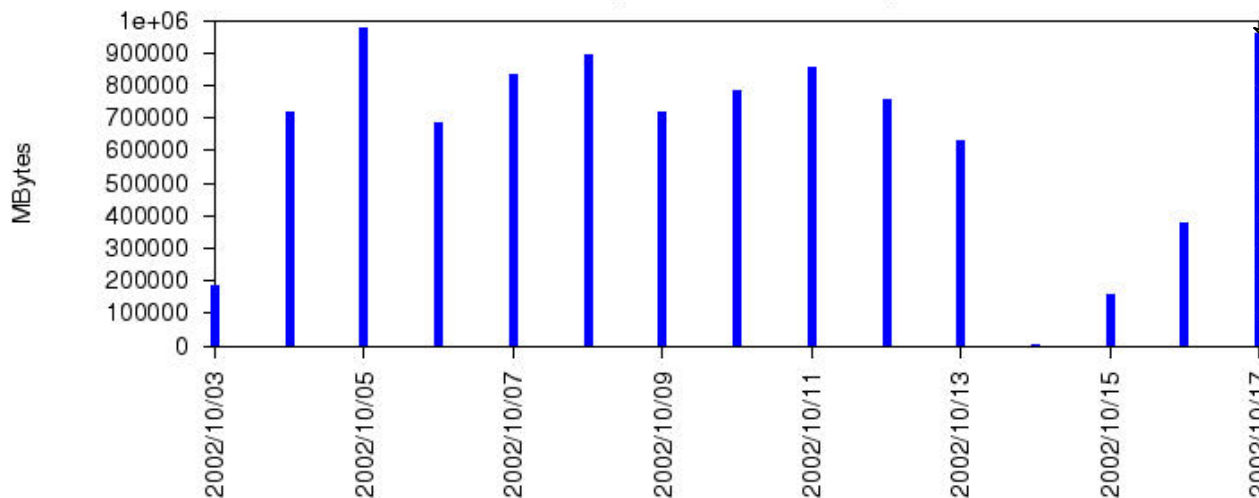
The test harness simulates
real operation, failures too!

clued0
Incoming(+) / Outgoing(-)
MBytes Transferred Per Day



Almost
800GB were
brought into
the cluster in
one day

clued0
Intra-Station
MBytes Transferred Per Day



Intra-
station
transfers
were as
high as 1
TB per day

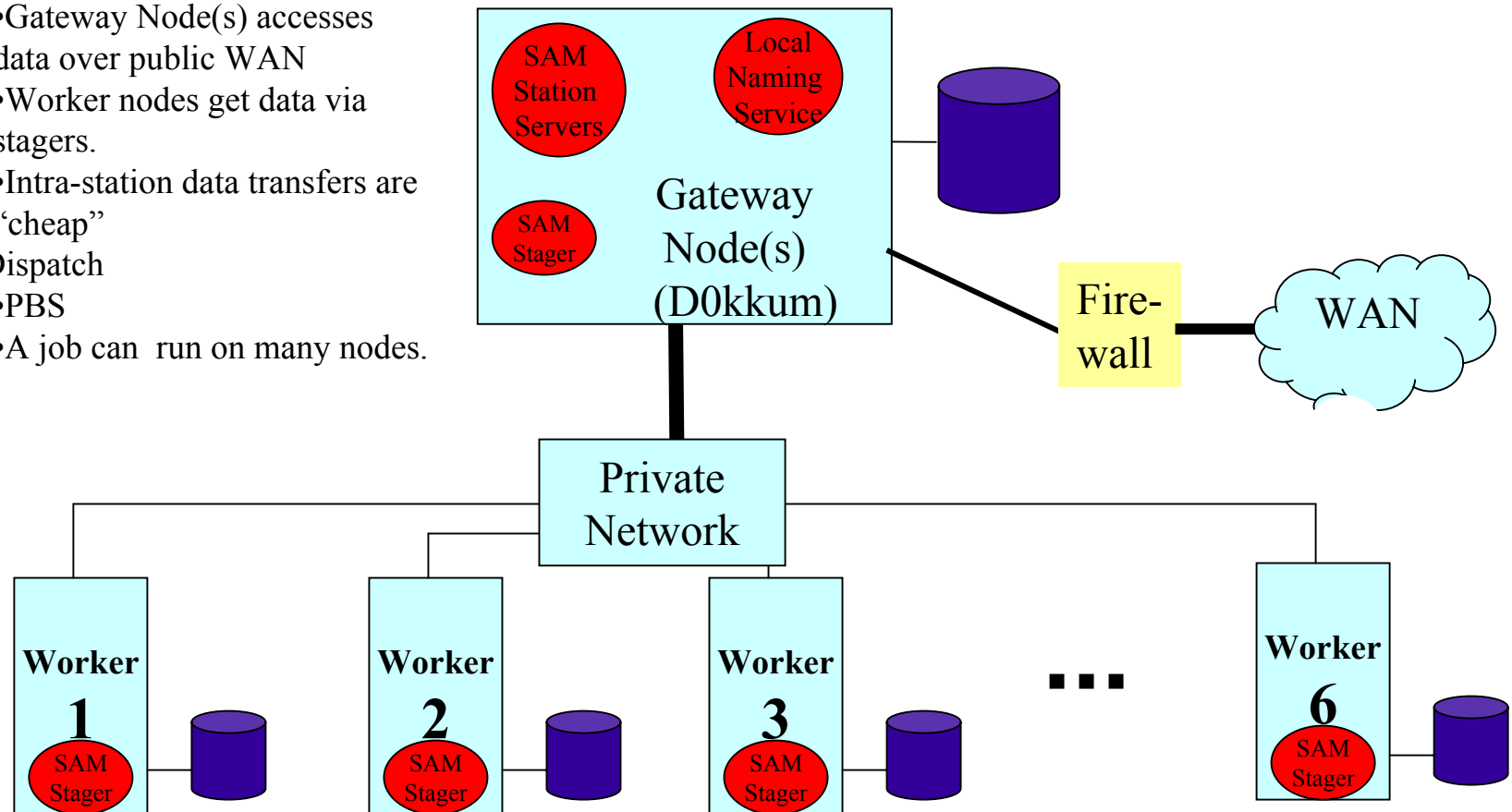




SAM Station: Dzero Distributed Cache Farm on VPN (Nijmegen)

- Network

- Gateway Node(s) accesses data over public WAN
- Worker nodes get data via stagers.
- Intra-station data transfers are “cheap”
- Job Dispatch
- PBS
- A job can run on many nodes.





Nijmegen farm: use



- Upgrade from previous farm server to present one:
 - ◆ 2 TB of disk space
 - ◆ ~ 1 TB for SAM cache
 - ◆ ~ 1 TB for software, (private) data files
- Aim: aid graduate students in doing physics analysis
 - ◆ At present, 4 students
 - ◆ Use nodes for batch job submission
- Will also use this for code development
 - ◆ (Wouldn't need SAM for this)
- Possibly: setup prototype Regional Analysis Center
 - ◆ SAM cache sufficiently big
 - ◆ Depends on availability of student(s) to help with setup

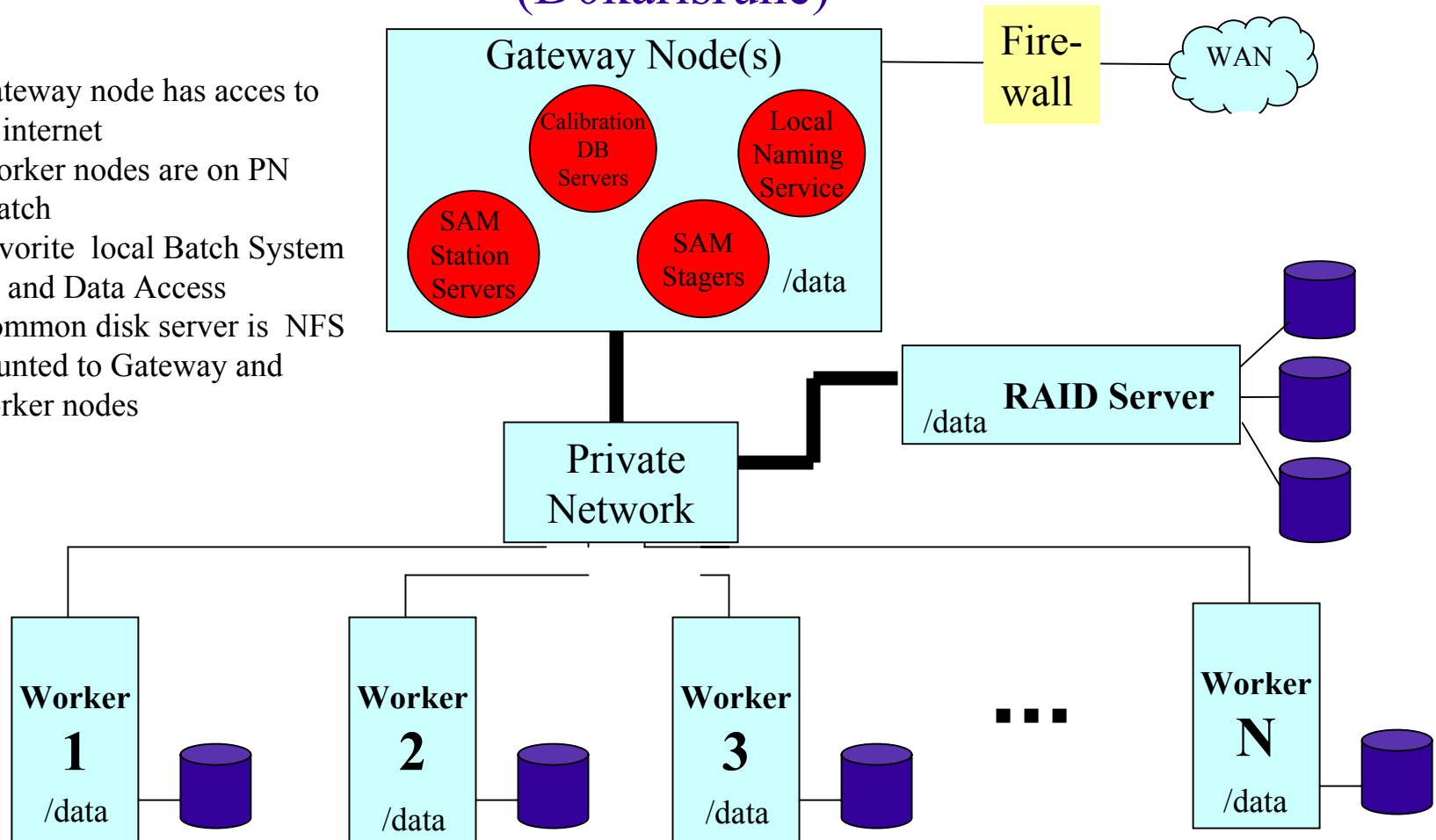
The Nijmegen station is used for analysis



SAM Station: Shared Cache Configuration w/ VPN (D0karlsruhe)



- Network
 - Gateway node has access to the internet
 - Worker nodes are on PN
- Job Dispatch
 - Favorite local Batch System
- Software and Data Access
 - Common disk server is NFS mounted to Gateway and Worker nodes



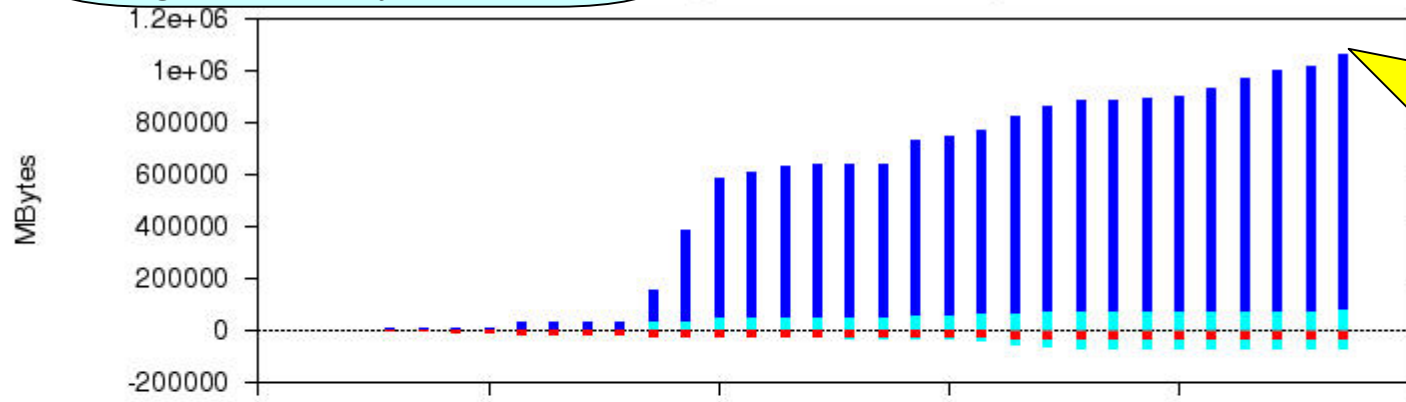


Data Transferred to D0Karlsruhe



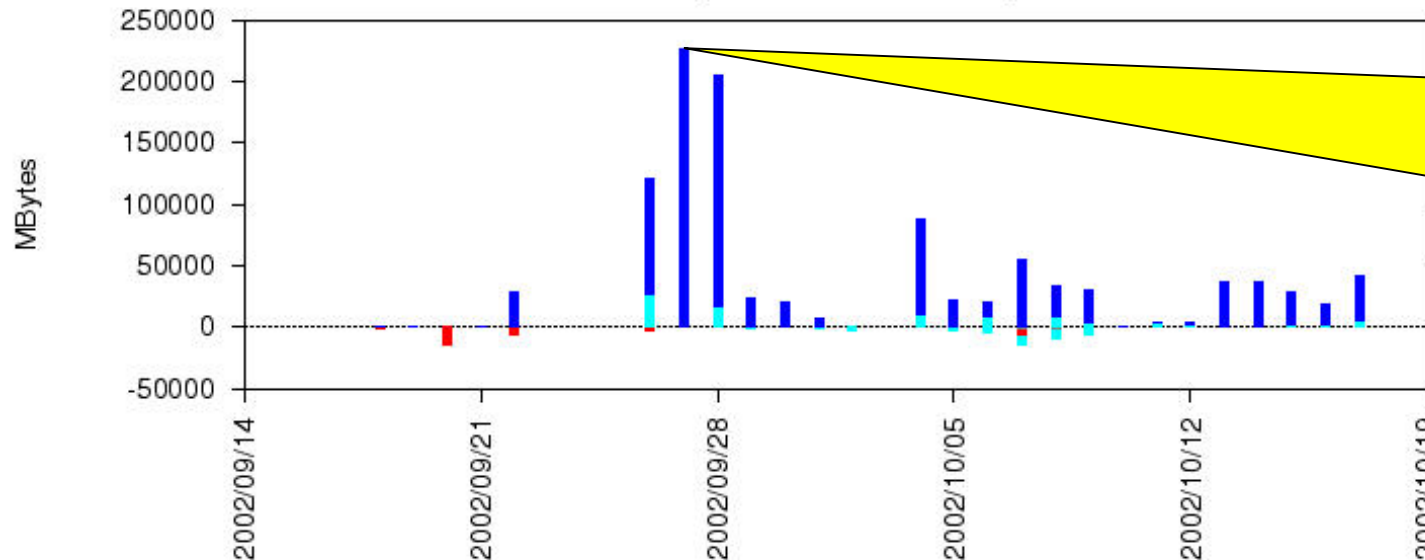
This is our first prototype
Regional Analysis Center!

d0karlsruhe
Incoming(+) / Outgoing(-)
MBytes Transferred Per Day



Over 1TB of
Dzero
Thumbnail data
pulled to the
Karlsruhe station
in the last 30
days

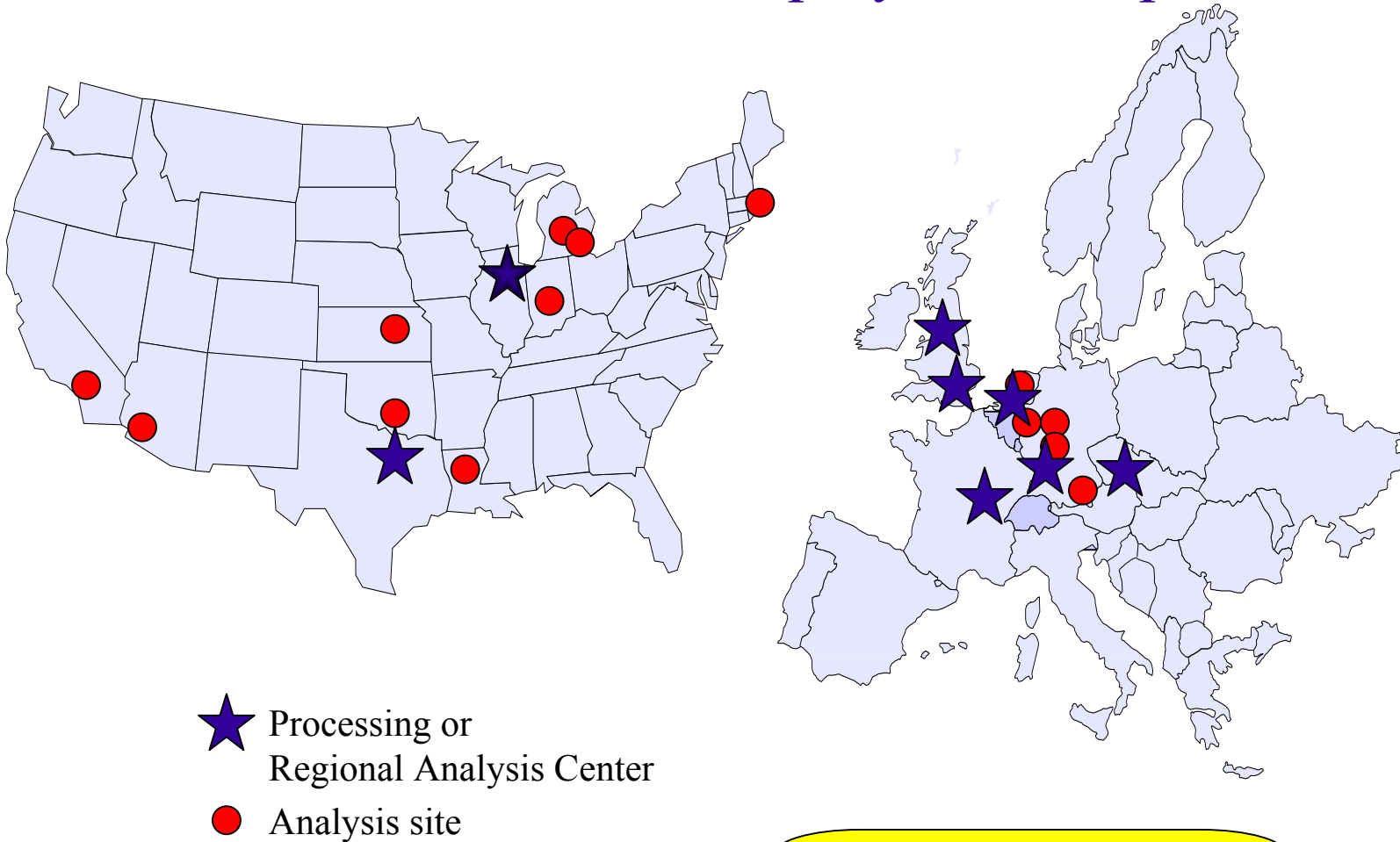
d0karlsruhe
Incoming(+) / Outgoing(-)
MBytes Transferred Per Day



Over 220 GB of
Thumbnail data
pulled to the
Karlsruhe station
in one day.



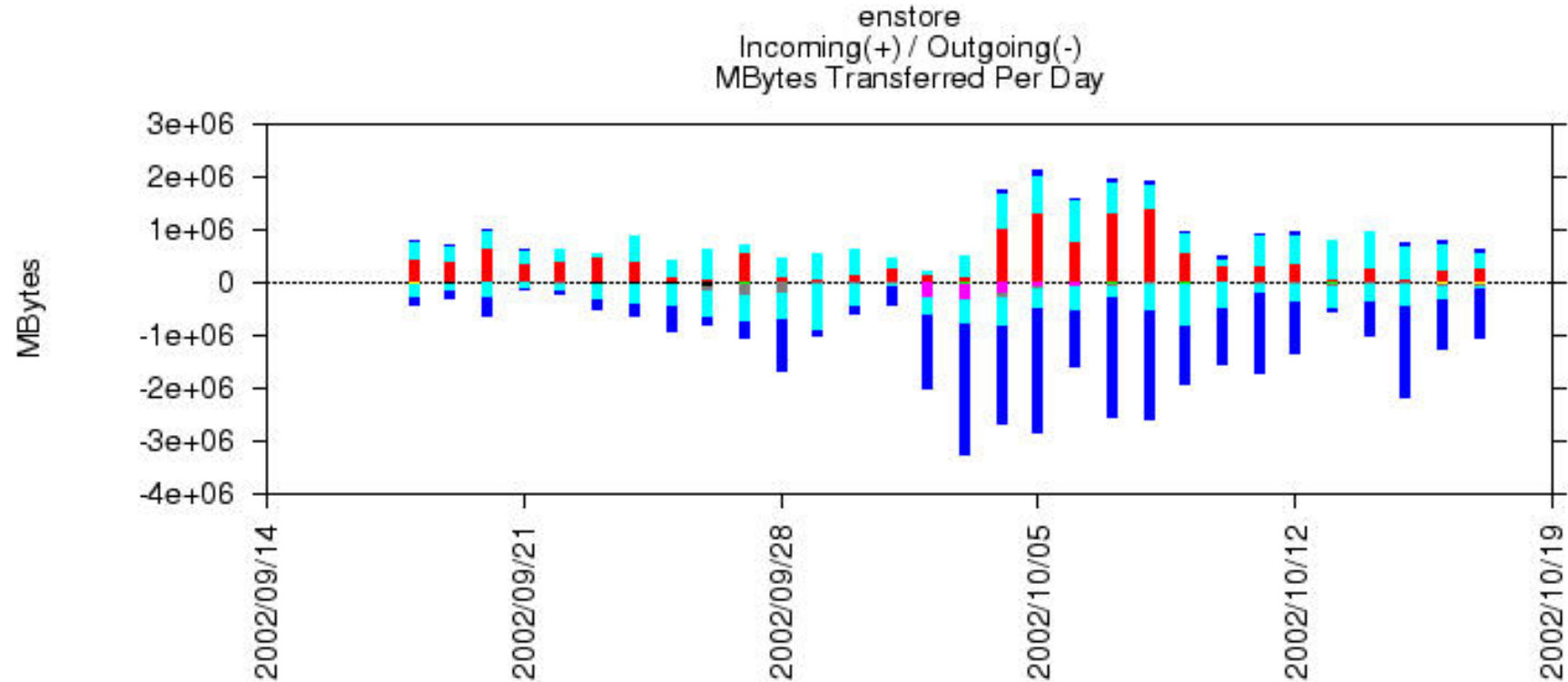
Dzero SAM Deployment Map



Shown are the most active station sites



Data In and out of Enstore



Stations:

central-analysis	datalogger-d0olc	imperial-test	hoeve
fnal-farm	d0karlsruhe	cab	other



The
Holy
Grail



SAM and the Grid

SAM-Grid



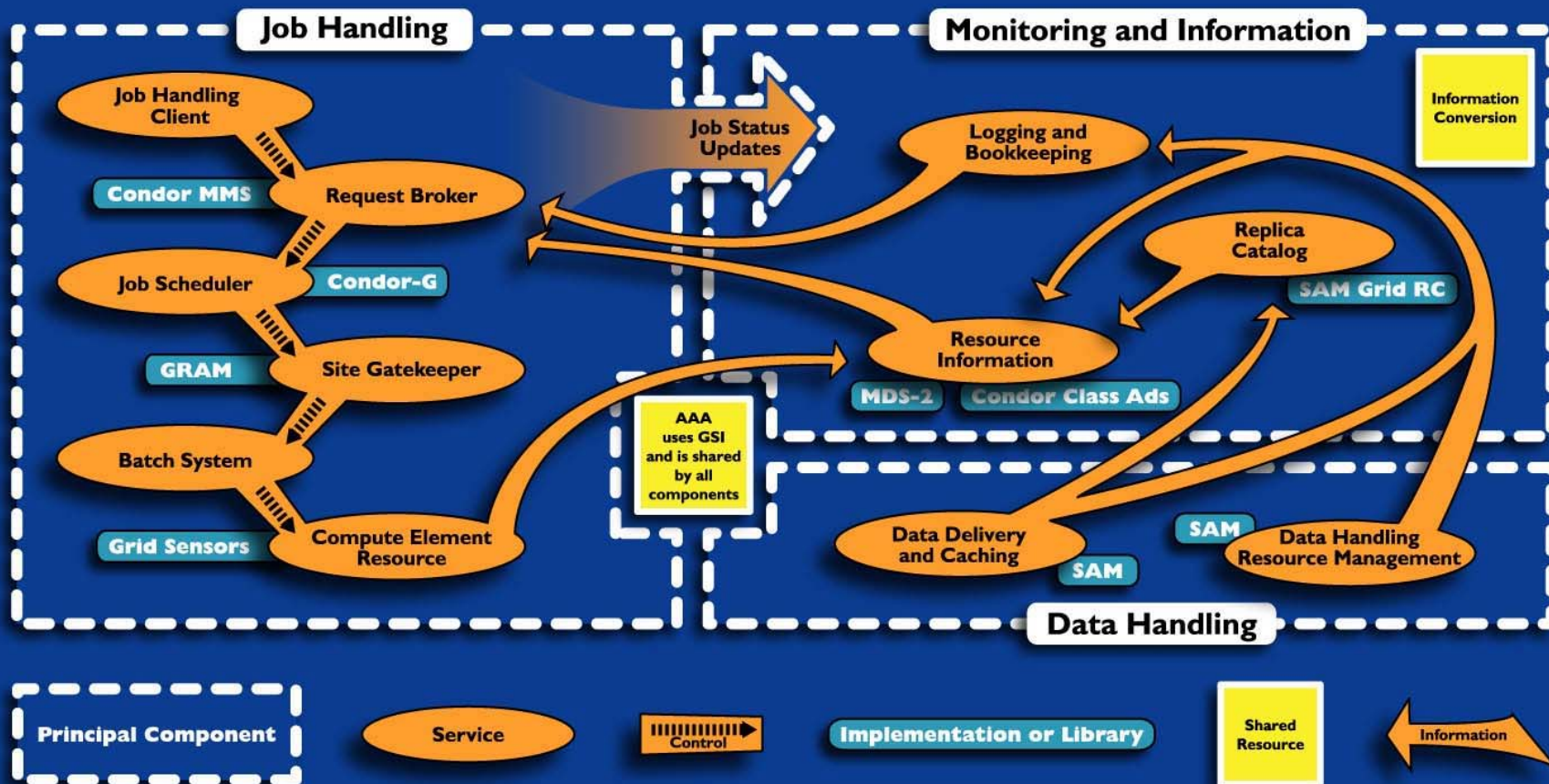


What is SAM-Grid?

- Project to include Job and Information Management (JIM) with the SAM Data Management System
- Project started in 2001 as part of the PPDG collaboration to handle D0's expanded needs. Architecture design in Spring 2002.
- Current SAM-Grid team includes:
 - ◆ Andrew Baranovski, Gabriele Garzoglio, Hannu Koutaniemi, Lee Lueking, Siddharth Patil, Abhishek Rana, Dane Skow, Igor Terekhov*, Rod Walker (Imperial College), Jae Yu (U. Texas Arlington). *Team Leader
 - ◆ Collaboration with U. Wisconsin Condor team.
- **<http://www-d0.fnal.gov/computing/grid>**



SAM-Grid Architecture



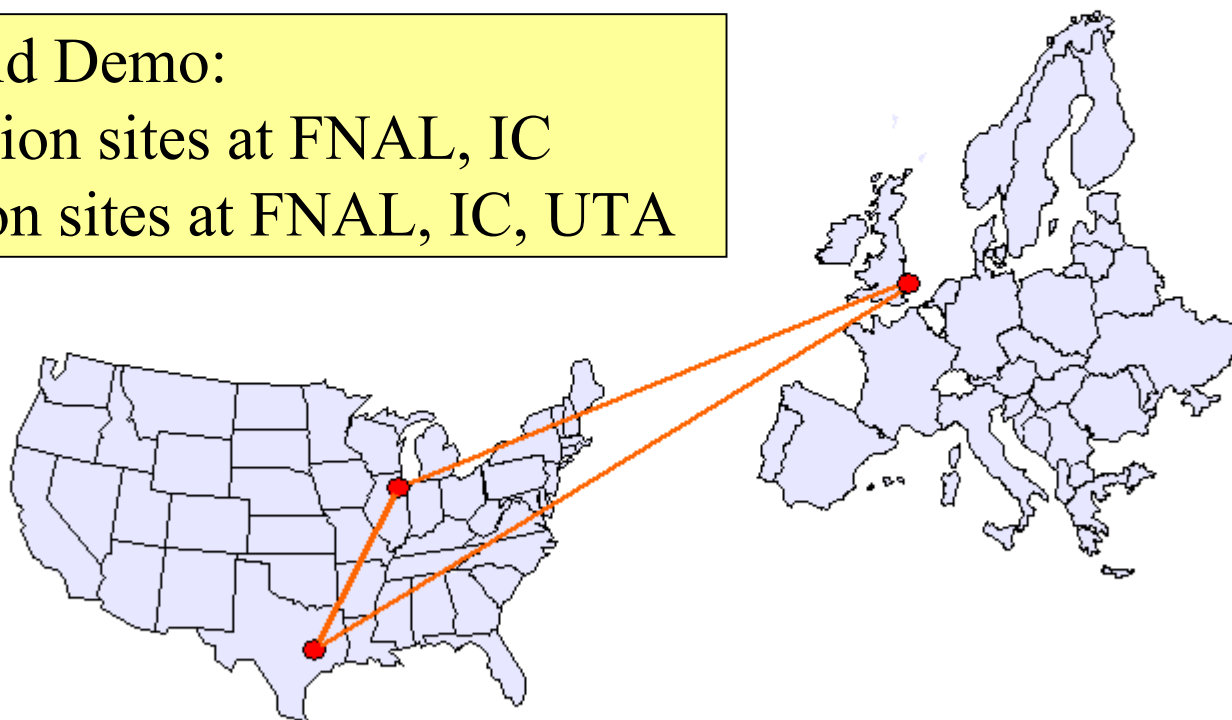
SAM GRID INFORMATION & MONITORING SYSTEM

Launching the Monitoring System:

Please click at the map to monitor the execution sites.
Click [here](#) to get information about the submission sites.

SAM-Grid Demo:

- Submission sites at FNAL, IC
- Execution sites at FNAL, IC, UTA



SAM Grid Monitoring System

Wed Oct 9 23:58:10 CDT 2002



Monitoring at the FNAL Site

[View Authorized Grid Users](#)

Please click on a station's name to get its Server-Version and Start-time.
For stations that are grid-enabled, the Cluster Details can be viewed through the available link.

Station Name	Universe	Grid-enabled	Projects	Disks	Groups	Experiment
samadams	dev	Yes	0	3	9	d0
sammy	dev	Yes	0	2	4	d0
sameggs	dev	Yes	0	1	7	d0
central-analysis	prd	No	15	53	9	d0
droidd	dev	No				d0
fnal-farm	prd	No	17	169	1	d0
cdf-glasgow-fnal	dev	No	0	2	1	cdf
cdf-glasgow-fnal	prd	No	0	8	1	cdf

Legend

SAM Grid Monitoring System

Oct 17 9:49:44 CDT 2002



Monitoring at the IC Site

Please click on a station's name to get its Server-Version and Start-time.
stations that are grid-enabled, the Cluster Details can be viewed
through the available link.

[View Authorized Grid Use](#)

Station Name	Universe	Grid-enabled	Projects	Disks	Groups	Experiment
imperial-test	dev	Yes	● 2	● 2	● 2	d0
imperial-test	prd	Yes	● 1	● 6	● 1	d0

Legend

Sam Grid Projects

Projects at: *imperial-test - dev*



Sam Project Id	Total Files	Locked	Given	Delivery Errors	Wanted	Local Owner	Group
terekhov_sammy.fnal.gov_165135_18786_0	16	0	0	0	16	sam	grid
patil_sameggs.fnal.gov_215953_15454_0	15	15	0	0	15	sam	grid
patil_sameggs.fnal.gov_220241_15491_0	15	15	0	0	15	sam	grid
patil_sameggs.fnal.gov_221141_15556_0	15	15	0	0	15	sam	grid
patil_sameggs.fnal.gov_222030_17095_0	15	15	0	0	15	sam	grid
patil_sameggs.fnal.gov_223138_17608_0	15	15	0	0	15	sam	grid
sam_64476	15	15	0	0	15	sam	grid
terekhov_sameggs.fnal.gov_234839_20084_0	10	3	2	0	8	sam	grid

Sam Grid Authorized Users



Authorized Grid Users at: IC Site

Global Id Subject	Local Id	Certificate Authority	User Type
C=FR, O=CNRS, OU=LAPP, CN=Dominique Boutigny/Email=boutigny@in2p3.fr	collngdj		
O=Grid, O=UKHEP, OU=hep.ph.ic.ac.uk, CN=Rod Walker	walker	UK-HEP	Person
O=Grid, O=UKHEP, OU=hep.ph.ic.ac.uk, CN=Philip Lewis	pl297	UK-HEP	Person
O=Grid, O=UKHEP, OU=hep.ph.ic.ac.uk, CN=Dr D J Colling	collngdj	UK-HEP	Person
O=Grid, O=UKHEP, OU=hep.ud.ac.uk, CN=Ben West	mcprod	UK-HEP	Person
O=doesciencegrid.org, OU=People, CN=Gabriele Garzoglio 762243	sam	doesciencegrid	Person
O=Grid, O=UKHEP, OU=hep.ph.ic.ac.uk, CN=Alex Howard	howard	UK-HEP	Person
O=Grid, O=Globus, OU=hep.ph.ic.ac.uk, CN=cas/sampc.hep.ph.ic.ac.uk	sam	Globus	Services
O=doesciencegrid.org, OU=People, CN=Warren Matthews 837082	condor	doesciencegrid	Person
O=Grid, O=UKHEP, CN=host/fb00.hep.ph.ic.ac.uk	gdmp	UK-HEP	Services
O=doesciencegrid.org, OU=People, CN=Tomasz Wlodek 50053	sam	doesciencegrid	Person
O=Grid, O=Globus, CN=Jaehoon Yu	sam	Globus	Person
O=doesciencegrid.org, OU=People, CN=Jaehoon Yu 520999	sam	doesciencegrid	Person
O=doesciencegrid.org, OU=People, CN=Siddharth Patil 966454	sam	doesciencegrid	Person
O=doesciencegrid.org, OU=People, CN=Abhishek Rana 891895	sam	doesciencegrid	Person
O=doesciencegrid.org, OU=People, CN=Mateusz Tkaczyk 347443	sam	doesciencegrid	Person
O=Grid, O=Globus, OU=fnal.gov, CN=SAM Run II Sequential Access	sam	Globus	Person
O=doesciencegrid.org, OU=People, CN=Hannu Koutaniemi 10449	sam	doesciencegrid	Person
O=doesciencegrid.org, OU=People, CN=Andrew Baranovski 232305	sam	doesciencegrid	Person
O=doesciencegrid.org, OU=People, CN=Igor V Terekhov 444282	sam	doesciencegrid	Person



Additional Stops on the Quest for the Grail





The steps in getting to SAM-Grid



- JIM Project
 - ◆ Job Management
 - ◆ Job Description Language
 - ◆ Information Service
 - ◆ Testbed prototype deployment includes
 - ◆ Resource advertisement: ClassAd
 - ◆ Gatekeeper and local scheduler: GRAM (Globus Resource Allocation Manager)
 - ◆ Monitoring: MDS (Monitoring and Discovery Service)
 - ◆ Submission sites: Grid Client (Condor-G)
- Grid Security (AAA) using GSI
 - ◆ May include kerberos cross authentication
 - ◆ Have GridFTP working as a sam transfer protocol. Latest bbftp also has security plug-in feature.
 - ◆ Need VO maintenance and User-level certificate authentication and authorization.
 - ◆ Other policy details of grid job submission





Other planned steps for SAM, SAM-Grid, and DZero



- dCache integration for rate adapting and remote station file serving.
- Understand the modularization of the data handling and storage interfaces
- Generalized HSM Adapters to employ:
 - ◆ HPSS, and other general MSS.
 - ◆ Network attached files (file url)
 - ◆ SRM interface
 - ◆ Use of additional dCache features
- D0 Run Time Environment will allow running on resources not tailored to D0 (no D0 installation).
- Site Autonomous SAM station and site resource management (general decentralization of SAM)
- Opportunistic SAM service deployment





Summary



- DZero is in the midst of rapid growth for data taking, MC production, processing and analysis.
- The SAM Data Handling model has proven to provide a flexible system for efficient utilization of DZero clusters under many uses and in many configurations.
- SAM-Grid (SAM and JIM) promises to provide capability for over-all job, data, and information management in compute and storage regimes.





Acknowledgments

- SAM Team, Dzero, CDF, ODS members
- D0 Task Force
- ISD, Enstore and dCache support and operation
- OSS, Farms Group
- ODS, database support
- DCD, Networking Group
- SAM shifters
- SAM station admins
- Many DZero and CDF users and contributors