

CDF Grid, Cluster Computing, and Distributed Computing Plans



Alan Sill

Department of Physics
Texas Tech University

Dzero Southern Analysis Region Workshop
UT Arlington, Apr. 18-19, 2003

A Welcome Warning Sign

(That should exist):

Welcome to the Grid

Your Master Vision
May Have To Coexist
With Those Of Many Others!

(Have a Nice Day...)

More Caveats



The “Master Vision” presented here is simply a collection of those of some others... (thanks to those who contributed transparencies).

Errors are mine, both in presentation and emphasis.

CDF present systems: CAF, Independent Computing, and SAM

- CAF is CDF's Central Analysis Farm project, built around the idea of moving the user's job to the data location.
- SAM Stands for Sequential Access to data via Metadata. It is basically a distributed data transfer and management service: data replication is achieved by use of disk caches during file routing.
- In each case, physicists interact with the metadata catalog to achieve job control, scheduling and data or job movement.
- In addition, independent clusters and/or public resources can be used for Monte Carlo production and other tasks that produce data that can later be merged with the DFC through file import.
- CDF has been studying use of SAM for the past year with working prototypes, and is in the process of working towards merging its existing Data File Catalog into the SAM architecture.

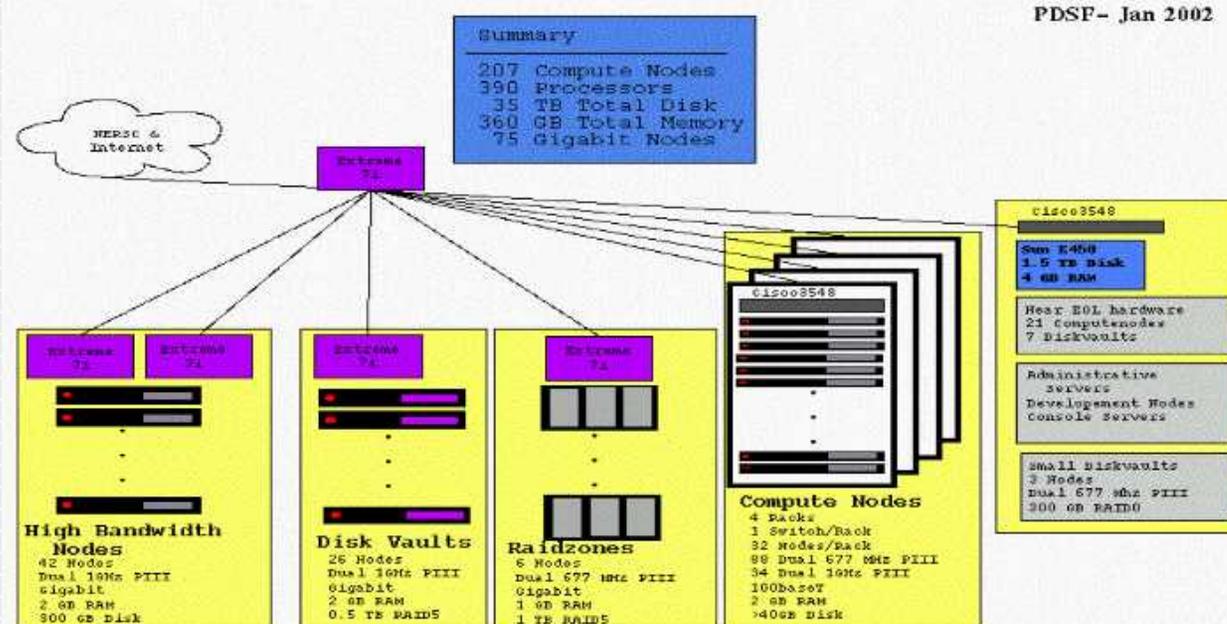
Off-Site CDF MC Production Sites

(that I'm aware of)

- IPP-Canada
 - Toronto
 - Alberta
 - McGill
- Glasgow: ScotGRID
- Universität Karlsruhe
- UC San Diego
- UIUC
- OSU
- Michigan
- LBNL
- Rutgers
- many more?... (PC husbandry is sexy)

Example: LBNL PDSF

- Initially started with use of leftover SSC hardware; expanded greatly over the years
- Shared between several experiments (CDF, ATLAS, astrophysics, etc.), many TB of disk, ~400 processors.
- Running stably for several years.



ScotGRID-Glasgow - Front View



ScotGRID-Glasgow Facts/Figures

- 59 x330 dual PIII 1GHz/2 Gbyte compute nodes
- 2 x340 dual PIII/1 GHz /2 Gbyte head nodes
- 3 x340 dual PIII/1 GHz/2 Gbyte storage nodes, each with 11 by 34 Gbytes in Raid 5
- 1 x340 dual PIII/1 GHz/0.5 Gbyte masternode
- 3 48 port Cisco 3500 series 100 bit/sec Ethernet Switch
- 1 8 port Cisco 3500 series 1000 bits/sec Ethernet Switch
- 4 16 port Equinox ELS Terminal Servers
- RedHat 7.2
- xCAT-dist-1.1.RC8.1
- OpenPBS_2_3_16
- Maui-3.0.7
- OpenAFS-1.2.2 on masternode
- RAL virtual tape access
- IP Masquerading on masternode for Internet access from compute nodes
- Intel Fortran Compiler 7.0 for Linux
- HEPiX login scripts
- gcc-2.95.2
- j2sdk-1_4_1
- ~150,000 dedicated maui processor hours
- 38 names in NIS passwd map

echGrid

- Initially
 - 1 Origin 2000 Supercomputer (56 nodes) (Irix)
 - 3 Beowulf clusters (Linux) (total 120 nodes)
 - 140 Windows IT lab machines
 - 40 Windows Math machines
- Down the road
 - Other academic and administrative computing resources on campus
 - Approximately 1,500 lab machines campus-wide
- Specific to TTU HEP
 - 2 specialized small development Linux clusters
 - Several scientific workstations
 - Ability to submit through CAF interface to TTU grid (under development)

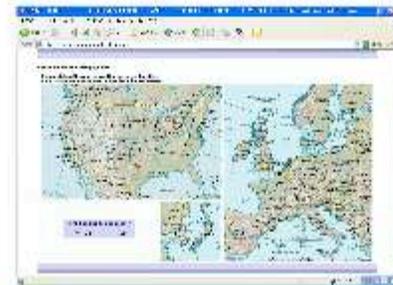


Korea



KNU node for SAM Grid

- CAF (Central Analysis Farm) at Fermilab for CDF
- At Run IIb, data size is 6 times more than now ⇒ DCAF (DeCentralized Analysis Farm)
- SAM (Sequential Access through Meta data)
 - ❑ To handle real data at Fermilab for DCAF around world
 - ❑ Gridification of DCAF via SAM Grid
 - ❑ 6 institutes (KNU, Toronto, RAL, Rutgers, Texas Tech, Glasgow) were involved in SC2002 Demo.



[Ref.] SAM Grid Home page



12 CPUS of KNU node
for SAM Grid

Karlsruhe (FZK)

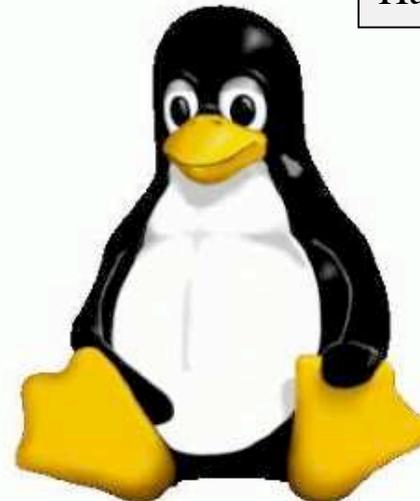


Compute Node Installation

1 kSI95 ~ 24 x 1 GHz PIII

The total CPU power at GridKa is currently ~20 kSI95

- 124 dual PIII with
 - 1 GHz or 1.26 GHz
 - 1 GB ECC RAM
 - 40 GB HDD IDE
 - 100 Mbit Ethernet
 - running Red Hat Linux 7.2
- + 130 PIV recently installed
- + ~60 PIV until April 2003



Has SAM Station

CDF GRID Project

Overall scope includes, but extends beyond, MC production

- User interacts with a standard GUI
- Job build, deployment, execution, I/O, etc. proceeds “under the hood”
- Geographical invariance: user in Liverpool may have her job execute on the CAF at FNAL, or on Toronto’s Big Mac, or elsewhere
 - but she doesn’t need to care!

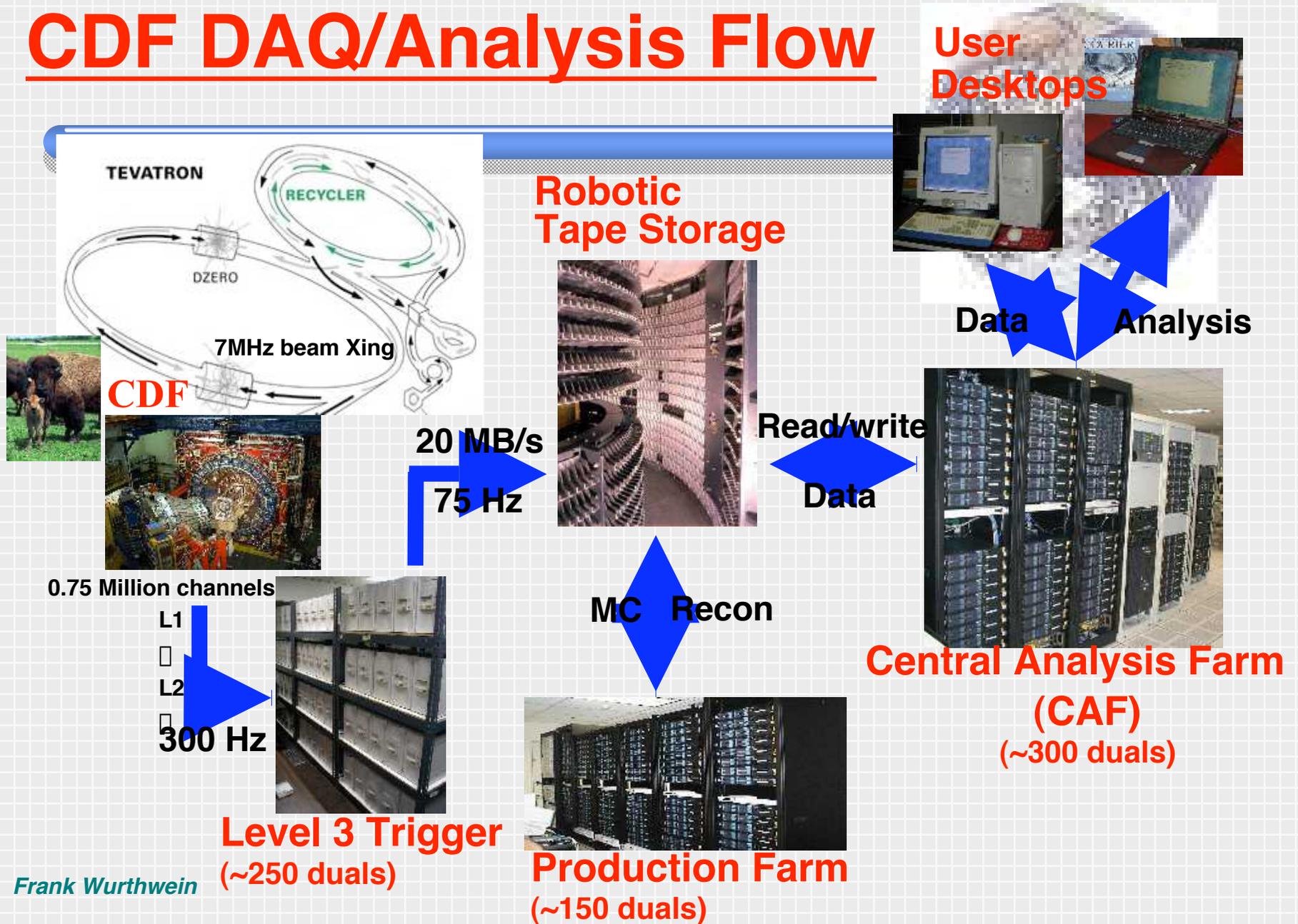
CDF GRID/CAF Software Infrastructure – Main Components:

- SAM (disk management)
- JIM (job management, based on Condor and Globus)
- CAF (interface used by a user for job submission via a GUI)
- FBSNG (batch management system)

Issues:

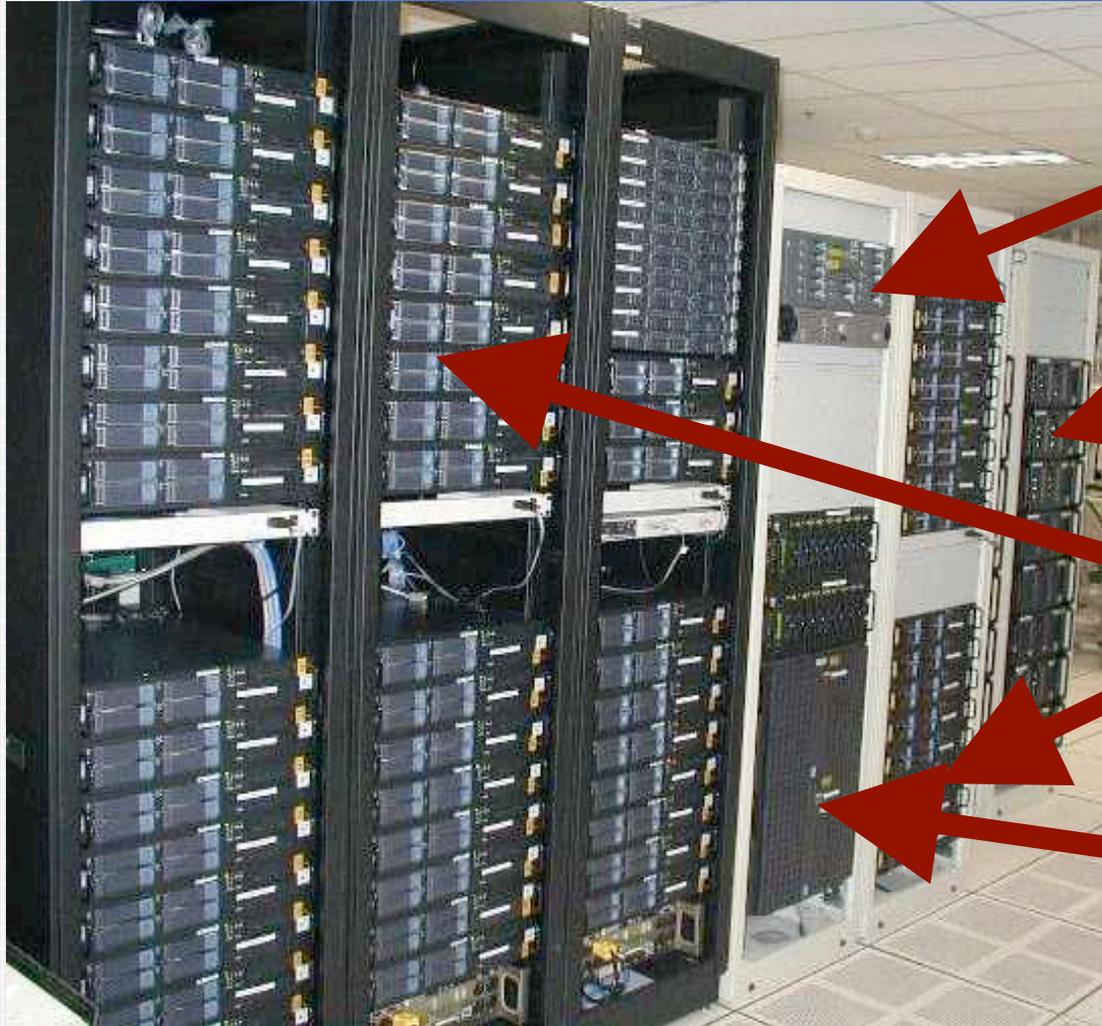
- How to deal with clusters that aren’t solely devoted to CDF computing
- Kerberization
- Compatibility of batch management systems across clusters

CDF DAQ/Analysis Flow





CAF Hardware



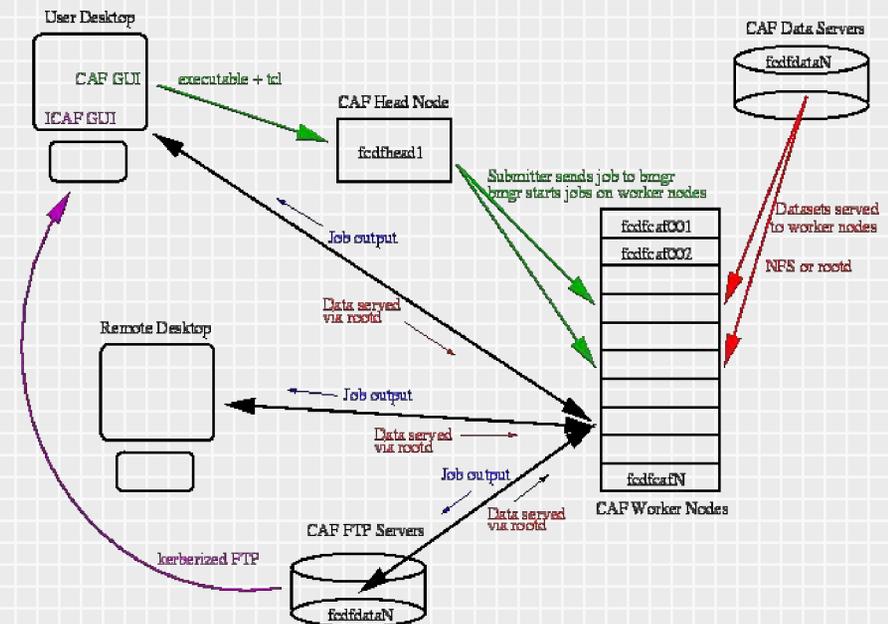
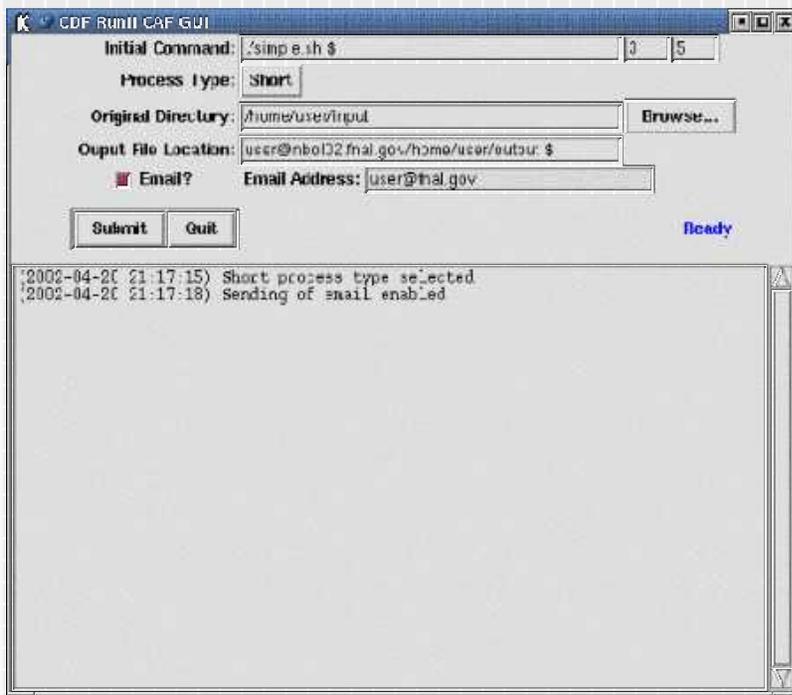
Code Server

File Servers

Worker Nodes

Linux 8-ways
(interactive)

CDF CAF Model & GUI



“Send my job to the data.”

- User submits job, which is tarred and sent to CAF cluster
- Results packed up and sent back to or picked up by user



Example CAF job submission

- Compile, build, debug analysis job on 'desktop'

section integer range

- Fill in appropriate fields & submit job

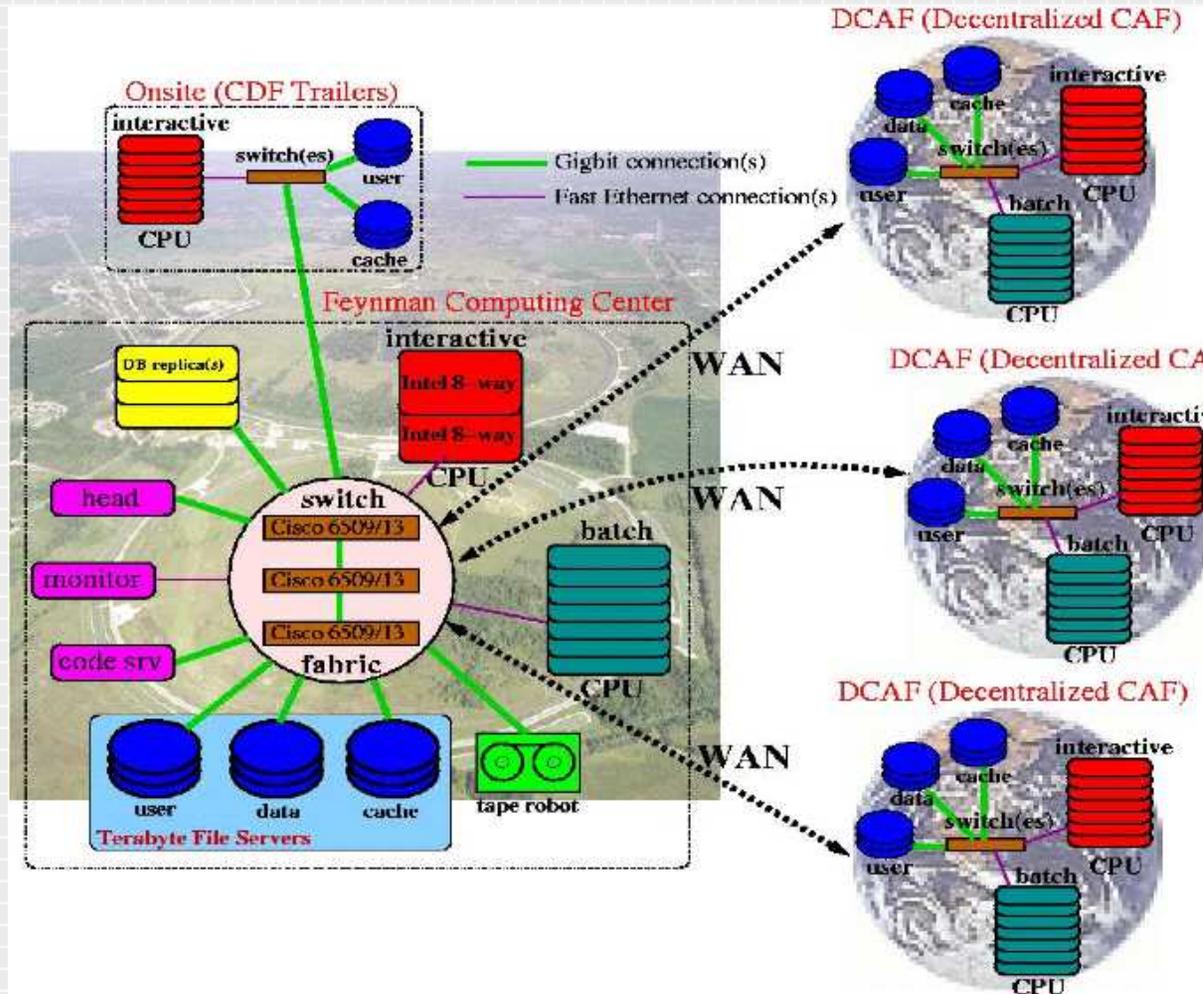
The screenshot shows the 'CDF RunII CAF GUI' window. The 'Initial Command' field contains `./simple.sh`. The 'Process Type' is set to 'Short'. The 'Original Directory' field contains `/home/msn/releases/development/CafUtil/examples` and is circled in red. The 'Output File Location' field contains `msn@cdfnrx2.fnal.gov/cdf/scratch/msn/temp$.tgz` and is circled in green. The 'Email?' checkbox is checked, and the 'Email Address' field contains `msn@fnal.gov`. A blue circle highlights the '600 610' range in the 'section integer range' field. A blue arrow points from the 'section integer range' text to the '600 610' field. A red arrow points from the 'user exe+tcl directory' text to the 'Original Directory' field. A green arrow points from the 'output destination' text to the 'Output File Location' field. The 'Submit' and 'Quit' buttons are visible. The status bar shows 'Ready'. A terminal window at the bottom displays the following output:

```
(2002-05-23 01:46:51) Email sent to msn@fnal.gov upon job completion
(2002-05-23 01:46:55) /bin/tar -cvzf /home/msn/msn49959.tgz *
(2002-05-23 01:46:57) Remove /home/msn/msn49959.tgz
(2002-05-23 01:46:57) Job Submission is successful, JID: 873
```

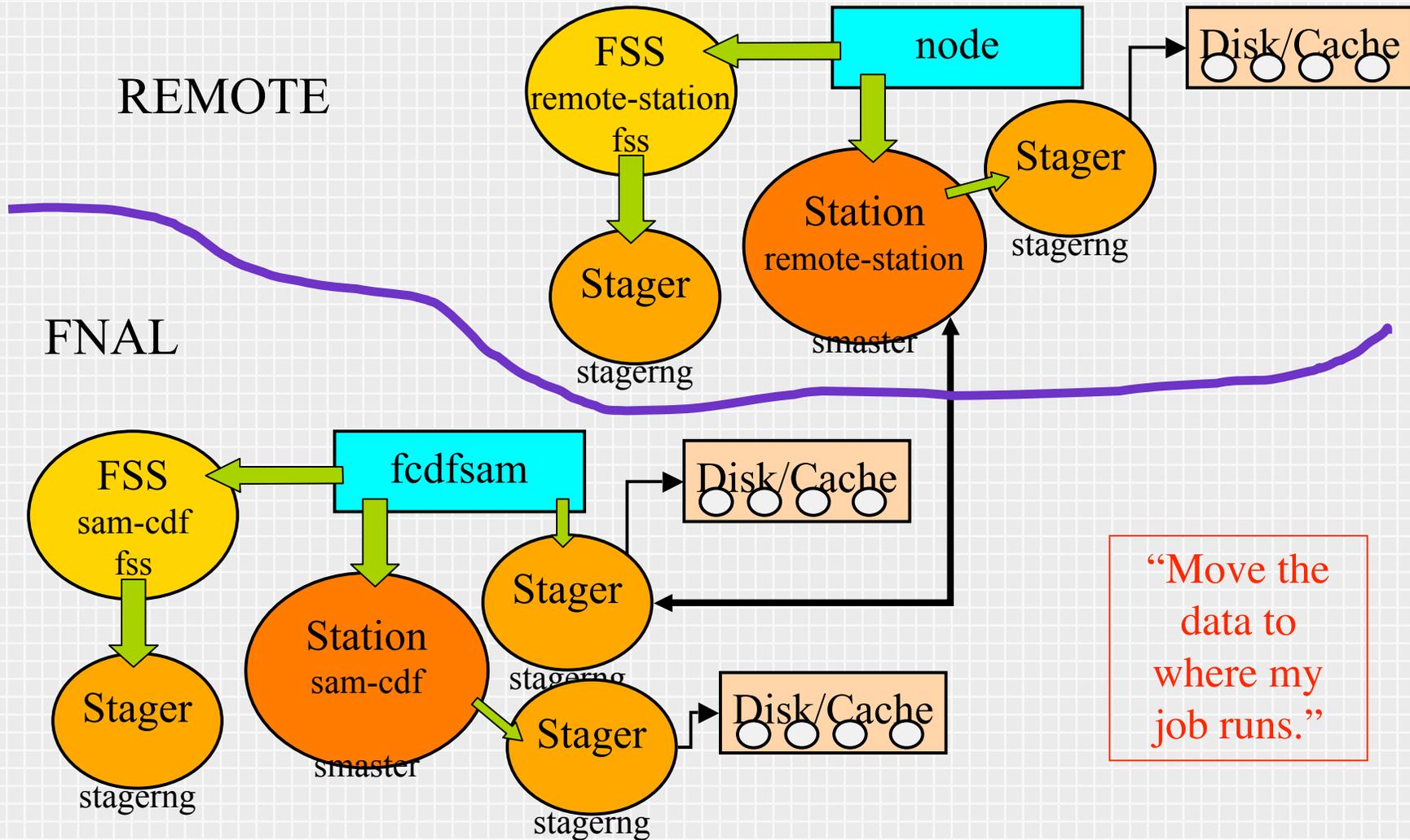
- Retrieve output using kerberized FTP tools
... or write output directly to 'desktop'!



Future CAF Directions



Comparison with SAM



CDF SAM Station Status



SAM At A Glance

CDF Production Environment
This page generated on 17 Apr 2003 22:51:56



SAM Stations:

Monitor Level: Normal

Station	Host	Version	Last Update
cdf-chicago	no smaster/fss		
cdf-ekoka	ekpcdf2.physik.uni-karlsruhe.de	v4_2_1_39	17 Apr 2003 02:55:40
cdf-enstore			
cdf-fnal-glasgow			
cdf-fnal-oxford			
cdf-fnal-uc1			
cdf-fakka	cdf.fzk.de	v4_2_1_3	
cdf-glasgow	edfa.ph.gla.ac.uk	v4_2_1_4	
cdf-glasgow.2	lf7.ph.gla.ac.uk	v4_2_1_4	
cdf-glasgow-fnal	no smaster/fss		
cdf-glasgow-fnal.2	nglas09.fnal.gov	v4_2_1_4	
cdf-knu	no smaster/fss		
cdf-oxford	matrix.physics.ox.ac.uk	v4_2_1_2	
cdf-rai	no smaster/fss		
cdf-rdk-fnal-1	no smaster/fss		
cdf-rutgers	hexcaf.rutgers.edu	v4_2_1_21	06 Jan 2003 19:27:18
cdf-sag	fcdfdata016.fnal.gov	v4_2_1_33	06 Apr 2003 17:36:05
cdf-san1			
cdf-scotgrid	no smaster/fss		
cdf-scotgrid.2	ifl.ph.gla.ac.uk	v4_2_1_42_cdf_scotgrid	11 Apr 2003 17:08:27
cdf-test			
cdf-toronto	chrome.physics.utoronto.ca	v4_2_1_21	12 Nov 2002 14:32:16
cdf-trieste	pccdf2.ta.infn.it	v4_2_1_41	15 Apr 2003 08:01:08
cdf-trieste-1			
cdf-ttu	testwulf.hpec.ttu.edu	v4_2_1_21	11 Apr 2003 13:28:49
cdf-ttu-phys	atarfire.phys.ttu.edu	v4_2_1_21	10 Apr 2003 20:58:26
cdf-tufts	tubapt.phy.tufts.edu	v4_2_0_3	04 Apr 2003 10:41:28
cdf-uc1	no smaster/fss		
cdf-ucsd	no smaster/fss		
cdf-walton_a			
cdfyale	cdf1.physics.yale.edu	unknown	unknown
nedf8			
robk-1			
samtrieste-1	no smaster/fss		

CDF has SAM stations at Fermilab, TTU, Rutgers, UK (3 locations), Karlsruhe, Korea, Italy, and Toronto. Other present locations in testing stages or inactive.

We are actively involved in developing and deploying SAM for CDF!

Main CDF SAM features to date

- Manual routing of SAM data analysis jobs to remote execution sites works!
- Routing of a SAM analysis job to the station that caches the maximum number of files requested by the job also works.
- Monitoring remote job routing works.
- Monitoring the status of SAM jobs on both grid and non-grid enabled stations works.
- Installation of new station software works, but must be tuned manually.
- Much borrowed from D0, but much independent development work going on too!

CDF SAM/Grid Organization: a Collaborative Effort

- We hold daily and weekly meetings to coordinate efforts on the CDF/Dzero Grid and SAM projects.
- Participants are from UK institutions, TTU, Karlsruhe (Germany), INFN (Italy), Korea, and other US institutions.
- Recently have strong interest from Finland and Canada.
- Participation is by OpenH323 video.
- We discuss operations, design, and implementation.
- The real pressure comes from trying SAM and Grid on data coming from the experiment now (so this is not a theoretical exercise!)
- Opportunity for other group participation is high.

Some Personal Observations

- This is a VERY disparate and distributed set of resources.
- Our ability to control (and even categorize) these resources is very limited.
- Our ability to *specify the terms* for interconnection with our resources (database, data handling, even job submission for running on our nodes), however, is perfect.
- The right context for this is *service definition* (what are the services we provide and how to connect to them, etc.).
- A minimal set of standards is crucial.
- Monitoring is crucial.

SAM+CAF: Towards the Grid

- Neither model fully implements the negotiation, standards preference, distributable nature or full set of protocols needed to be considered “grid-enabled.”
- Even DCAF (“De-Centralized Analysis Farm”) plus SAM would not be sufficient by itself. (Too much manual intervention for job handling.)
- Authentication, authorization, data transfer, monitoring and database problems.
=> Need a fully Grid-aware approach!

Reminder: What is a Grid?

- Grid computing, of course, consists of *standards* and *protocols* for linking up clusters of computers.
- Basic idea is to provide methods for access to distributed resources (data sets, cpu, databases, etc.).

Foster (2002):

“ ... a Grid is a system that:

- coordinates resources that are not [otherwise] subject to centralized control
 - using standard, open, general-purpose protocols and interfaces
 - to deliver nontrivial qualities of service
- Some of these goals can be achieved without full grid resources!!

A general list of issues for grids:

- Authentication:
 - Individuals, Hosts and Services must each authenticate themselves using flexible but verifiable methods. For example, in the US Physics Grid, projects are supported by the DOE Science Grid SciDAC project, which provides a centralized Certificate Authority and advice and help on developing the necessary policies. This CA is trusted by the European Data Grid. The issue of Certificate Authority cross-acceptance is under intense study as one of the defining features of generalized grid computing.
- Authorization:
 - This issue should be distinguished from authentication, which establishes identity, not just permission to utilize a given resource. The short term interoperable solution for authorization is LDAP. The EDG Local Center Authorization Service (LCAS), Virtual Organization Management Service (VOMS) and Globus Community Authorization Service (CAS) are being considered as longer term interoperable authorization solutions. Note that authorization can be handled locally once authentication has been assured.

A general list of issues for grids:

- **Resource Discovery:**
 - Provides methods to locate suitable resources on an automatic basis. The Grid Laboratory Uniform Environment “Glue Schema” sub-project (<http://www.hicb.org/glue/glue-schema/schema.htm>) is an example of information specifications that can be used for resource discovery.
- **Job Scheduling:**
 - The Globus Resource Allocation Manager (GRAM) is presently the standard protocol for grid job scheduling and dispatch in the EU and US high energy and nuclear physics grid projects. Job dispatch to EU and US sites through GRAM has been demonstrated in test mode by the ATLAS-PPDG effort. Other approaches have been proposed by localized grids.
- **Job Management:**
 - Examples: Condor-G, ClassAds, the EDG WP1 Resource Broker. The collaboration between Condor Project, Globus, EDG WP1 and PPDG is working towards a more common standard implementation, and hopes to make progress within the next six months.

A general list of issues for grids:

- Monitoring and Information Services:
 - Recent work done for the SC2002 conference demo has been able to demonstrate monitoring capability across widely distributed sites. In general, the infrastructure developed for monitoring should be at least as well developed as those developed for resource discovery and job submission.
- Data Transfer
 - Most high energy physics jobs require data movement capability that includes robust high speed file transfer. At present, the preferred tool is GridFTP, implemented via a publish/subscribe mechanism. Study of parallelized multi-socket protocol variants is also making progress.
- Databases
 - Databases are a crucially important but neglected area of grid development that is just beginning to get the attention that is required to enable highly distributed processing to proceed efficiently. This field should mature rapidly in the near future.

This sounds like a big list of topics, but:

- We're beginning to make progress!
 - SAM + GridFTP is being adopted by CDF on an experimental basis for remote file transfer.
 - For example: we have implemented 2 SAM stations, one in the Physics Department and one at the High Performance Computing Center, at TTU, 2 in Karlsruhe, 1 in Toronto, 1 in Italy, several in the UK, 1 in Korea, etc.
 - Have achieved >30 Mbit/second transfer rates from Fermilab to TTU into the HPCC station via SAM; even faster rates to Karlsruhe, the UK, and Toronto.
 - Development underway to interconnect SAM with TTU commercial grid.

Present Projects

- More complete documentation! (Software packages, human design specs, policies all need enhancement).
- Better job description language (soon).
- Improved metadata schema (soon).
- JIM / SAM / CAF integration (see SC2002).
- New brokering algorithms.
- More robust installation scripts.
- Merging DFC into SAM schema (db tables).

The SuperComputing 2002 Demo

Participating institutions included:

- CDF:
 - Texas Tech University, Texas;
 - Rutgers State University, New Jersey;
 - University of Toronto, Canada;
 - Rutherford Appleton Lab, UK;
 - Kyungpook National University, Korea.
- DZero:
 - UT Arlington, Texas;
 - Michigan State University, Michigan;
 - University of Michigan, Michigan;
 - Imperial College, UK;

SC2002 Demo! (Nov 16-22, 2002)



SC2002 Monitoring (Rutgers)

Combined Monitoring → JIM → FBS

Projects submitted from sumeggs.fnal.gov

see projects that have been switched with a resource. Information becomes available about the execution site, the station and the project's process/owner details.

Global Job ID	Owner	Status	Type	Execution Site	Local ID	Local Status	Station	Universe	Experience
patil_sameggs.fnal.gov_211241_20274_0	patil	Removed	cat	FNAL	Unknown	No data from server	cdf-fnl	prd	cdf
terakhov_sameggs.fnal.gov_132145_18831_0	terakhov	Held	sam_analysis	IC	Unknown	No data from server	imperial-test	dev	d0
ratnikov_sameggs.fnal.gov_0133348_18950_0	ratnikov	Running	cat	RUTGERS	441	running	cdf-rutgers	prd	cdf

March 24, 2003 F.Ratnikov: GRID Based Monitoring on the Rutgers CDF Analysis Farm 13/16

Needs and Plans

- Reliable, routine execution of metadata-driven, locally distributed Monte Carlo and real data analysis jobs with basic brokering.
- Scheduling criteria for data-intensive jobs, with fully automated or user-controllable job handling – data handling interaction.
- Hierarchical caching and distribution for both physics data and database metadata interactions.
- Integration with non-high energy physics grids.
- Automatic matching of jobs to both database and data.
- Fully distributed monitoring of jobs, data flow, database use, and other information services.

Conclusions

- We are implementing automated grid-enabled mechanisms for high energy physics data analysis as part of a new effort in Grid Computing for CDF. We have successfully operated a prototype of this system and are beginning to involve students and faculty in its configuration, installation, development and use.
- This project has the goal of integrating the CAF, independent cluster computing, and SAM resources with grid technologies to enable fully distributed computing both DZero and CDF.
- This has to coexist with a wide variety of distributed resources that **ALREADY EXIST** (and in many cases are shared with other experiments) => generalization helps, standardization crucial!
- This will be our first step towards creating a general capability in high-profile, high-volume scientific data analysis for CDF.