Global User Analysis Computing for CDF

CHEP'04
INTERLAKEN

Alan Sill
Texas Tech University/
Fermilab

International Conference on Computing in High Energy Physics
Interlaken, Switzerland, Sep. 27 - Oct. 1, 2004

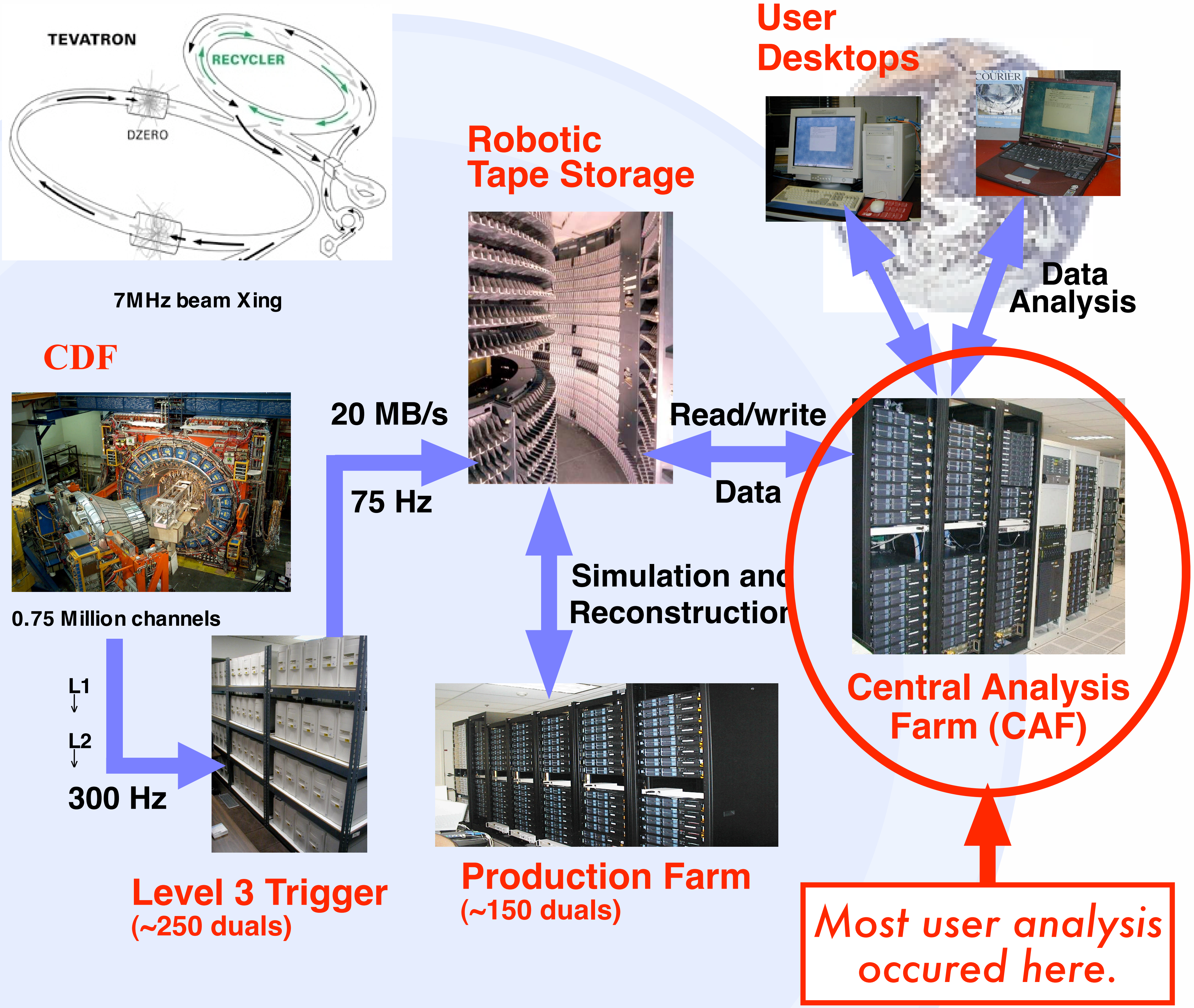
Elements of Global Computing

- It must work for everyone
  - Entire collaboration built the detector
  - Entire collaboration analyzes the data
    - => *Keep user environment simple!*
- It should be easy to adopt
  - The physics workflow should look familiar to the physicist.
- It should add or create additional resources
  - There should be some distinct advantage obtained from the computing being globally distributed.

CDF Central Computing

- Developed in 2001-2002 to respond to experiment's greatly increased need for computational and data handling resources to deal with Run II.
- Replaced Symmetrical Multi-Processing (SMP) approach with inexpensive collection of commodity Linux-based computing and file server systems.
- One of the first large-scale cluster approaches to user computing for general analysis (as opposed to farms for production of standardized jobs).
- Greatly increased cpu power & data to physicists (presently 300 TB data + 1500 cpus).

CDF Data Analysis Flow: 2002-03



Most user analysis occured here.

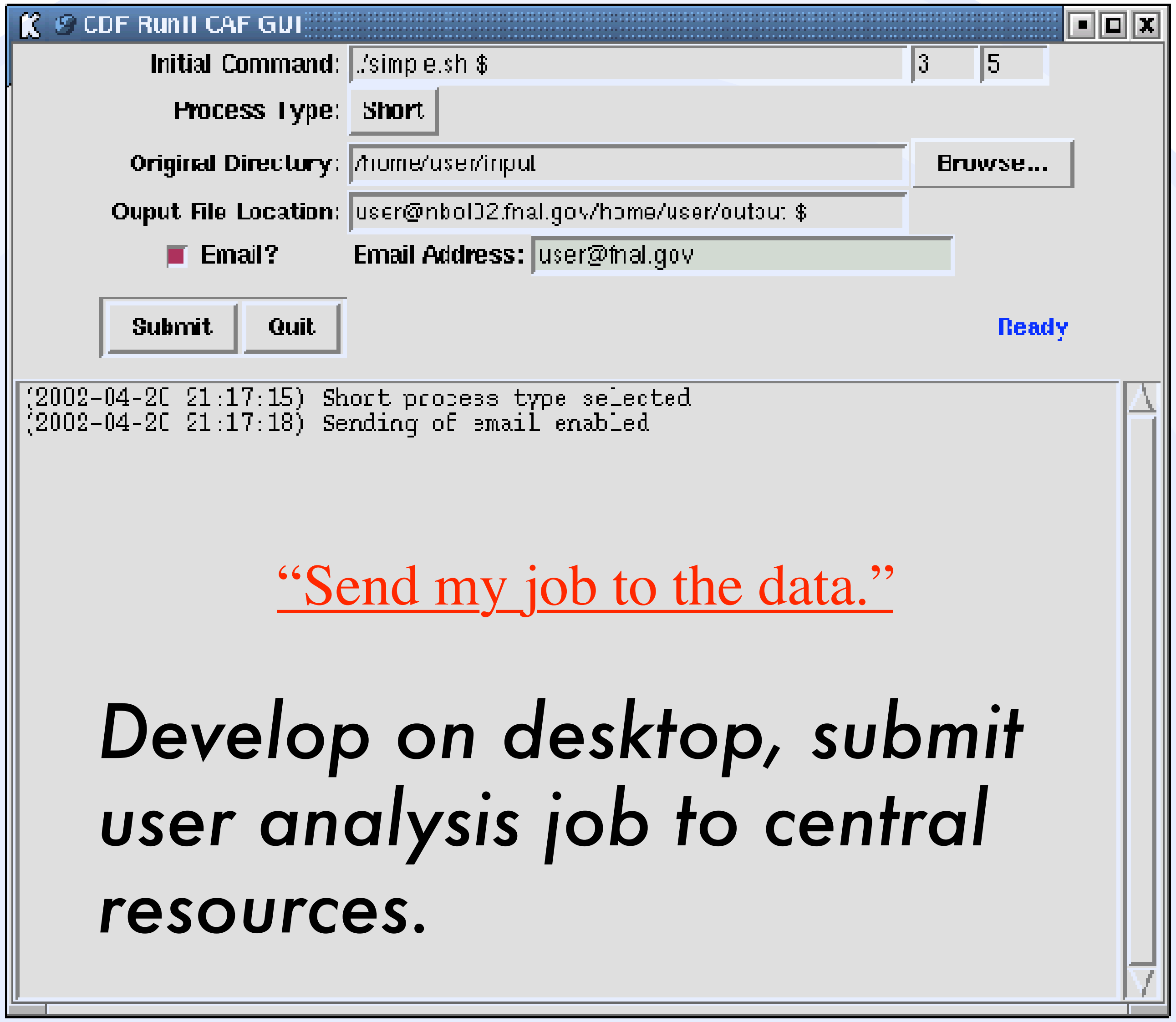
CAF part of overall analysis flow.

Users submit jobs through CAF GUI or command line to central cpu, disk and tape resources.

Queues are arranged to give priority to users on resources provided by their own institution.

CAF GUI *

* (Command line submission also possible.)



Initial Command: ./simple.sh $ 3 5

Process Type: Short

Original Directory: /home/user/input

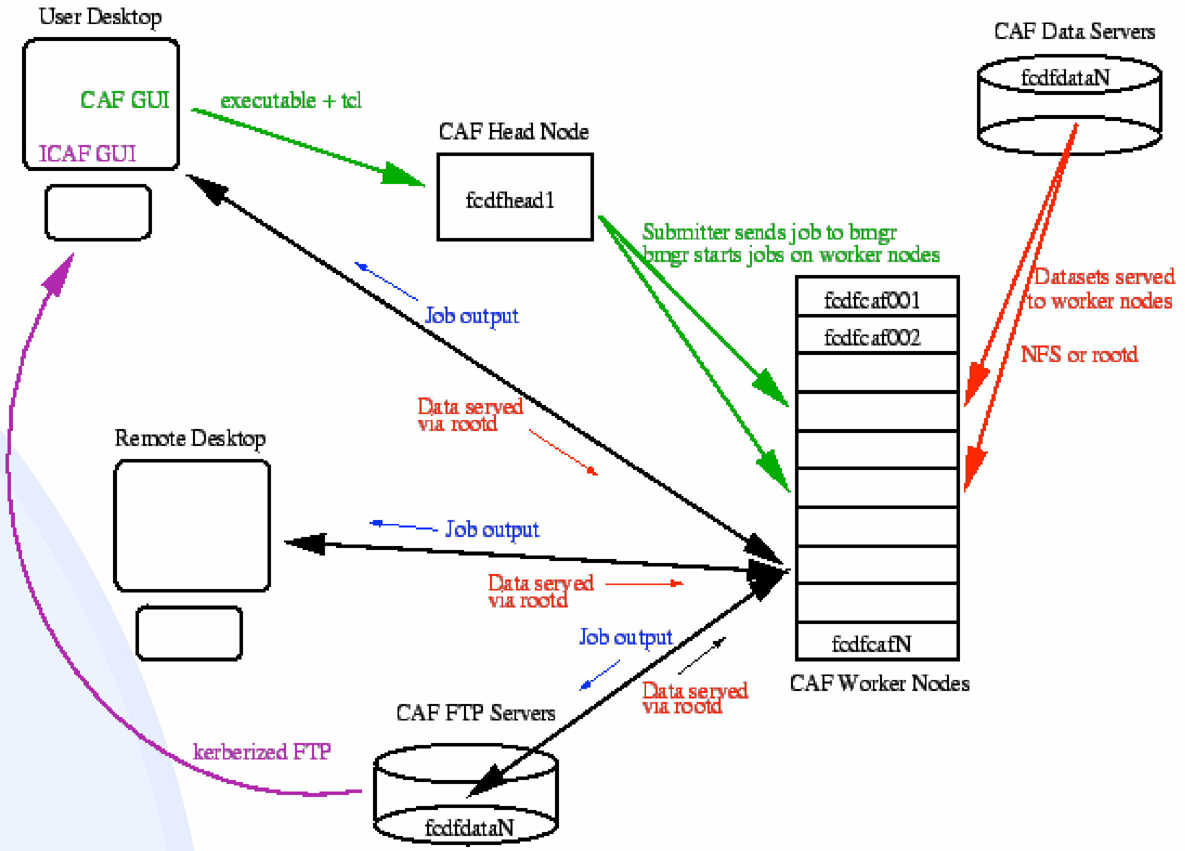
Ouput File Location: user@nbol02.fnal.gov/home/user/output $

■ Email? Email Address: user@fnal.gov

```
(2002-04-20 21:17:15) Short process type selected
(2002-04-20 21:17:18) Sending of email enabled
```

“Send my job to the data.”

Develop on desktop, submit user analysis job to central resources.



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

Most CDF physicists are familiar with the operation of this system.

Environment on a CAF

- Basic CDF software exists on CAF.
- Authentication via Kerberos
  - All jobs run as cafuser with authentication of actual user through special principal
  - Database, data handling remote user ID passed on through lookup of actual user via special principal (important for monitoring)
- User's entire environment can come over in a tarball - no need to pre-register or submit only certain jobs.
- Job returns results to user via secure ftp/rcp controlled by user script and principal.

Elements of Global Computing

- It must work for everyone
  - Entire collaboration built the detector
  - Entire collaboration analyzes the data
    - => *Keep user environment simple!*
- It should be
  - The physics o the physicist.
- It should ac ources
  - There should be some distinct advantage obtained from the computing being globally distributed.

The user's job context is transported to the execution site as a sandboxed tarball => no surprises should occur when job runs.

Sociological factors to build a CAF

- All users in CDF have an account if they request one. (Currently 768 registered users.)
- Institutions get priority on use of portions of the CAF that they contribute
  - Exclusive queues
  - Priority using disks, etc.
- Common hardware choices made by central admins.
- All leftover time available to everyone.
- Result: 100% utilization!

This is a very efficient, grid-like way to use resources!!

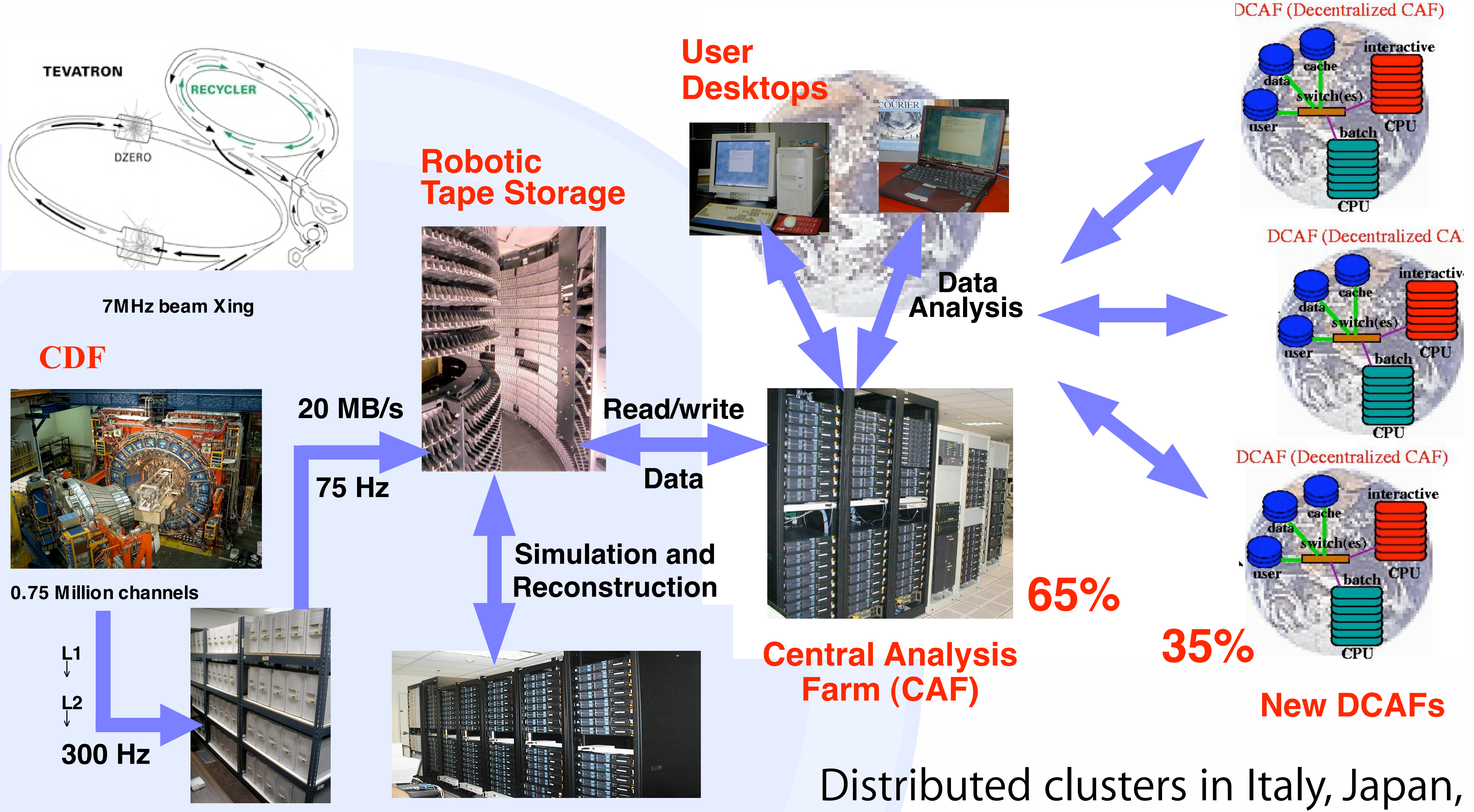
Elements of Global Computing

- It must work fo
  - Entire collabor
  - Entire collabor
    - => *Keep use*

To go further, we took the basic approach of extending the familiar CAF interface used on the central system to remote clusters in locations throughout the world.

- It should be easy to adopt
  - The physics workflow should look familiar to the physicist.
- It should add or create additional resources
  - There should be some distinct advantage obtained from the computing being globally distributed.

CDF Data Analysis Flow: 2004



7MHz beam Xing

CDF

0.75 Million channels

L1

L2

300 Hz

20 MB/s

75 Hz

Robotic Tape Storage

User Desktops

Data Analysis

Read/write

Data

Simulation and Reconstruction

65%

35%

Level 3 Trigger (~250 duals)

Production Farm (~150 duals)

Central Analysis Farm (CAF)

New DCAFs

Distributed clusters in Italy, Japan, Taiwan, Spain, Korea, Germany, UK, US, and Canada (more coming).

What works now

- Cluster technology (CAF = "CDF analysis farm") extended to remote sites (DCAFs = Decentralized CDF Analysis Farms). Batch systems: presently have choice of FBSNG Fermilab-written system or Condor (see poster 390).
- SAM data handling system:

  Several talks in this conference: see papers 38, 373, 455, 460, 462, 500, and posters 113, 335, 451, 468, 481

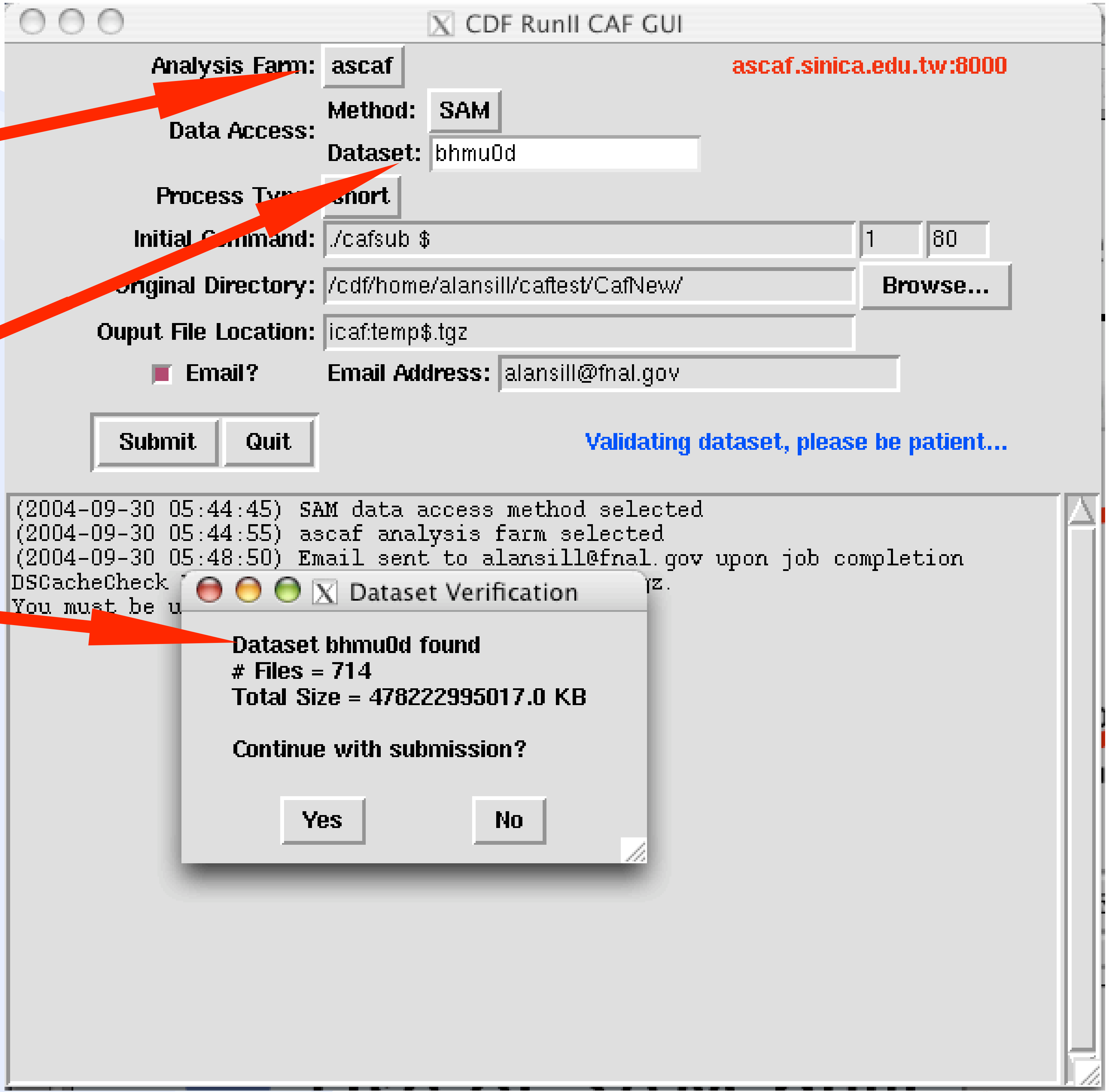
  - Pinning of defined datasets at remote locations, or use built-in cache management features.
  - Works also on systems with a non-CAF architecture.
  - Migrating to new version that can more easily recover from real-world problems from handling large datasets.

Other standard components

- Ganglia hardware monitoring.
- Bandwidth monitoring by Fermilab network team.
- Batch system user monitoring for both FBSNG and Condor.
- Bookkeeping of pinned datasets via SAM.
- New database query caching system for calibrations.
- MonALISA monitoring soon of both SAM and DCAF.
- Have experimented with "resilient dCache" at some sites.

Modifications to the CAF GUI

- Allows selection of CAF (not automatic yet).
- User specifies data set.
- Use of SAM features built in.
- Same interface can be used for grid.



CDF RunII CAF GUI

Analysis Farm: ascaf ascaf.sinica.edu.tw:8000

Data Access: Method: SAM

Dataset: bhmu0d

Process Type: short

Initial Command: ./cafsub $ 1 80

Original Directory: /cdf/home/alansill/caftest/CafNew/ Browse...

Ouput File Location: icaf:temp$.tgz

■ Email? Email Address: alansill@fnal.gov

Submit Quit Validating dataset, please be patient...

```
(2004-09-30 05:44:45) SAM data access method selected
(2004-09-30 05:44:55) ascaf analysis farm selected
(2004-09-30 05:48:50) Email sent to alansill@fnal.gov upon job completion
DSCacheCheck                                                 gz.
You must be u
```

Dataset Verification

Dataset bhmu0d found
Files = 714
Total Size = 478222995017.0 KB

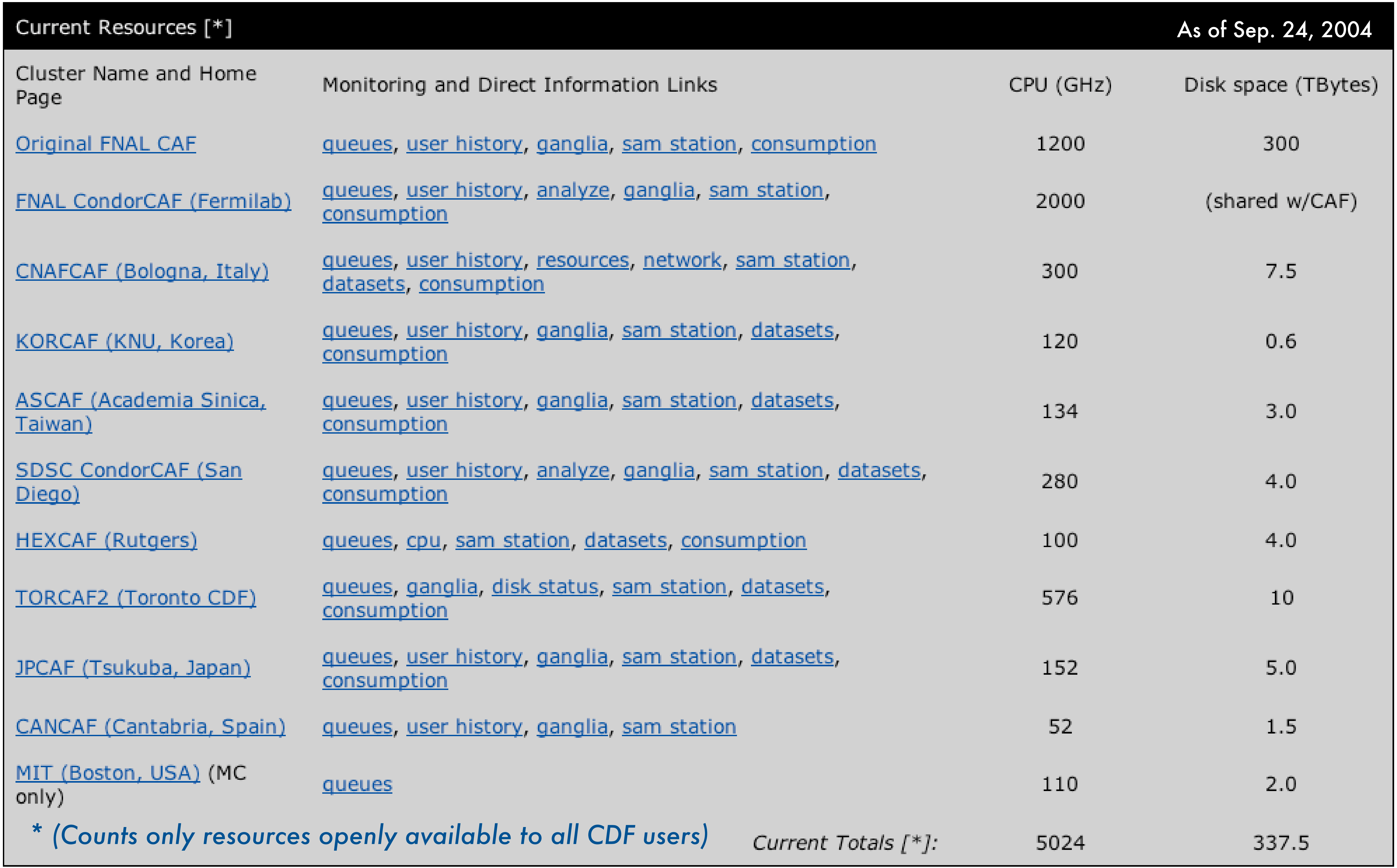
Continue with submission?

Yes No

Functionality for Users (9/2004)

Feature	Status
Simple self-contained sandbox	Yes
Runs arbitrary user code	Yes
Automatic identity management	Yes
Network delivery of results	Yes
Input and output data handling	Yes
Batch system priority management	Yes
Automatic choice of farm	Not yet!
Negotiation of resources	Not yet!

Current CDF Dedicated Resources:



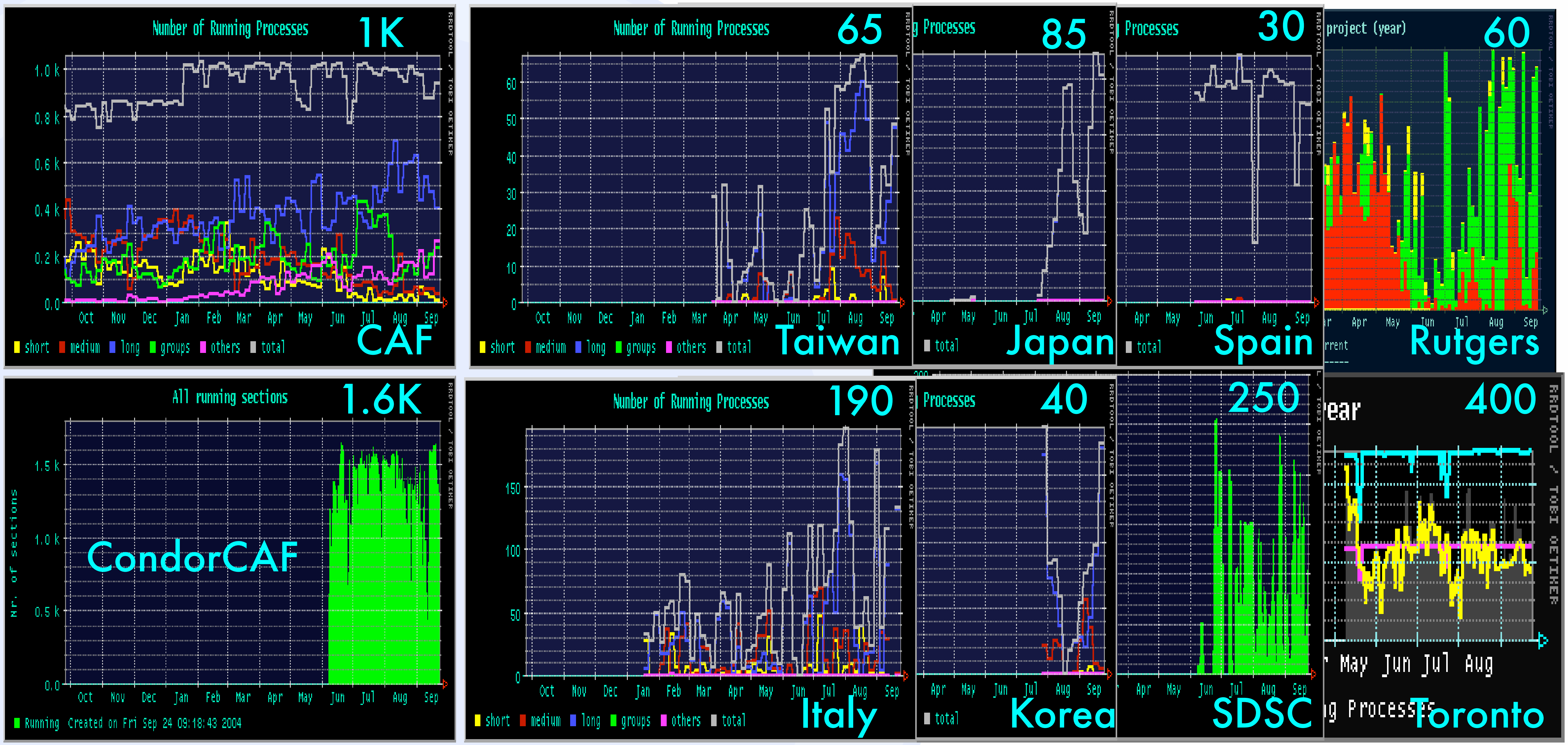
Current Resources [*] — As of Sep. 24, 2004

Cluster Name and Home Page	Monitoring and Direct Information Links	CPU (GHz)	Disk space (TBytes)
Original FNAL CAF	queues, user history, ganglia, sam station, consumption	1200	300
FNAL CondorCAF (Fermilab)	queues, user history, analyze, ganglia, sam station, consumption	2000	(shared w/CAF)
CNAFCAF (Bologna, Italy)	queues, user history, resources, network, sam station, datasets, consumption	300	7.5
KORCAF (KNU, Korea)	queues, user history, ganglia, sam station, datasets, consumption	120	0.6
ASCAF (Academia Sinica, Taiwan)	queues, user history, ganglia, sam station, datasets, consumption	134	3.0
SDSC CondorCAF (San Diego)	queues, user history, analyze, ganglia, sam station, datasets, consumption	280	4.0
HEXCAF (Rutgers)	queues, cpu, sam station, datasets, consumption	100	4.0
TORCAF2 (Toronto CDF)	queues, ganglia, disk status, sam station, datasets, consumption	576	10
JPCAF (Tsukuba, Japan)	queues, user history, ganglia, sam station, datasets, consumption	152	5.0
CANCAF (Cantabria, Spain)	queues, user history, ganglia, sam station	52	1.5
MIT (Boston, USA) (MC only)	queues	110	2.0
	Current Totals [*]:	5024	337.5

* *(Counts only resources openly available to all CDF users)*

Utilization

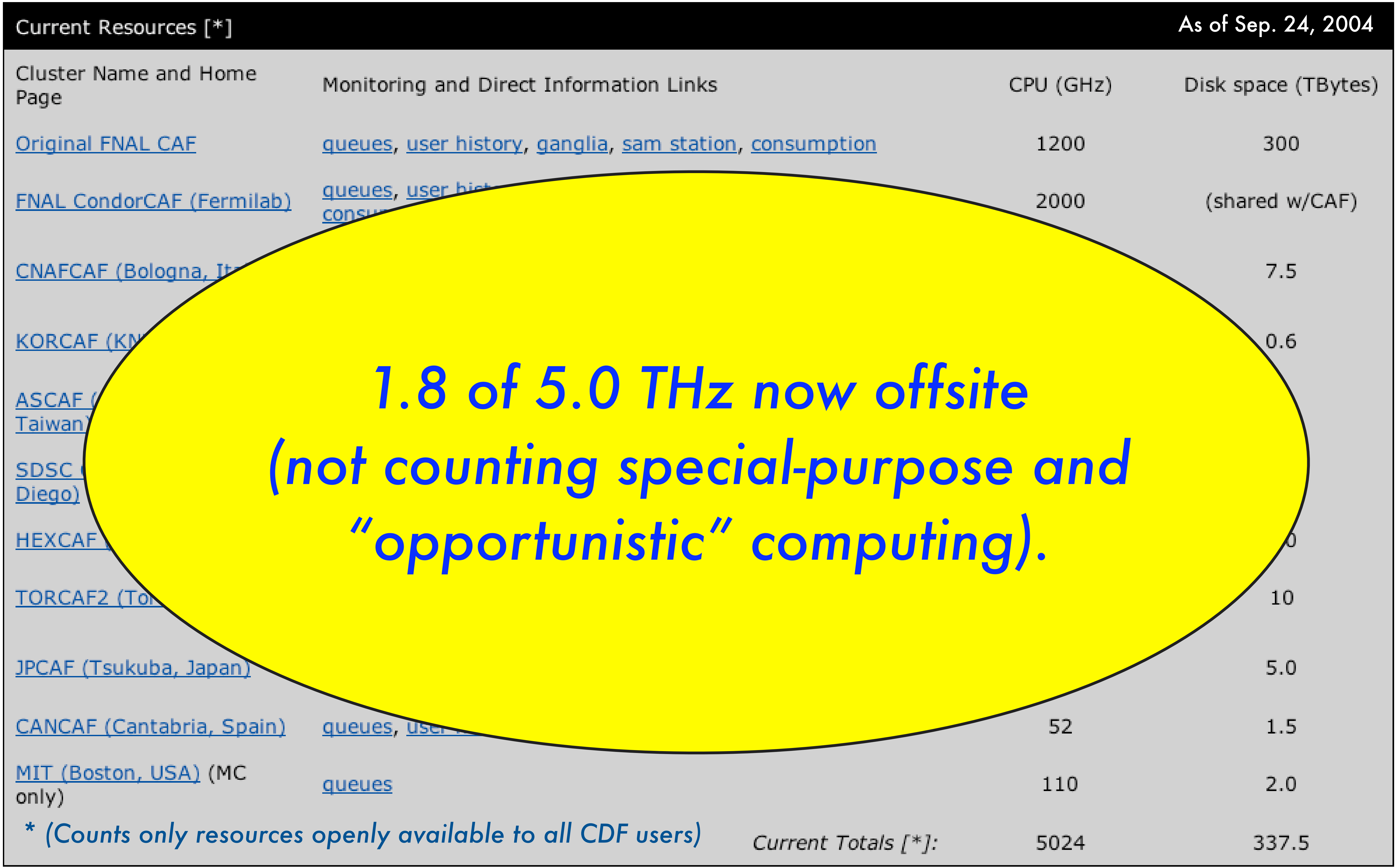
- Experiment-wide global aggregate monitoring not done yet (in progress)
- Off-site new CAFs show usage pattern ramping up, driven by conference seasons



Central

Off-Site

Current CDF Dedicated Resources:



Current Resources [*]

As of Sep. 24, 2004

Cluster Name and Home Page	Monitoring and Direct Information Links	CPU (GHz)	Disk space (TBytes)
Original FNAL CAF	queues, user history, ganglia, sam station, consumption	1200	300
FNAL CondorCAF (Fermilab)	queues, user hist consu	2000	(shared w/CAF)
CNAFCAF (Bologna, It			7.5
KORCAF (KN			0.6
ASCAF (Taiwan)			
SDSC (Diego)			
HEXCAF (			0
TORCAF2 (To			10
JPCAF (Tsukuba, Japan)			5.0
CANCAF (Cantabria, Spain)	queues, use	52	1.5
MIT (Boston, USA) (MC only)	queues	110	2.0
* *(Counts only resources openly available to all CDF users)*	*Current Totals [*]:*	5024	337.5

1.8 of 5.0 THz now offsite (not counting special-purpose and "opportunistic" computing).

How did we achieve this?

- Series of workshops oriented to installation of cluster (DCAF) software and data handling (SAM) pieces, as well as other components (monitoring, DB handling, etc.) beginning Jan. 2004. (3 held so far.)
- Rigorous application of "it's got to work" mentality and focus on ability to function in our environment.
- Postponement of features not yet truly needed.
- Clear, "live" frequently updated installation manuals.
- Focus on operations while development is going on. (Daily 15-min op's meeting, weekly 1-hr summary.)

Next steps

- Follow through on near-term technology deployments
  - Distributed database cache (paper 204)
  - SRM-dCache (paper 460)
  - Aggregated global monitoring (poster 484)
- Continue and complete evaluation of grid directions
  - JIM (poster 293), Grid3 (paper 171), LCG and Condor "glide-in" technologies currently under evaluation.
  - Decision on future directions within next few months.
  - See also paper 192 on OSG directions.

Conclusions

- CDF has successully deployed a global computing environment for user analysis based on a simple clustering and submission protocol, with a large number of registered and active users.
- A large portion (35%) of the total cpu resources of the experiment are now provided offsite.
- Aggressive trimming of the software suite has allowed us to bring up and use a large number of computational resources in short order.
- Plans are in progress to build in more grid-like protocols to provide a bridge to the future.