



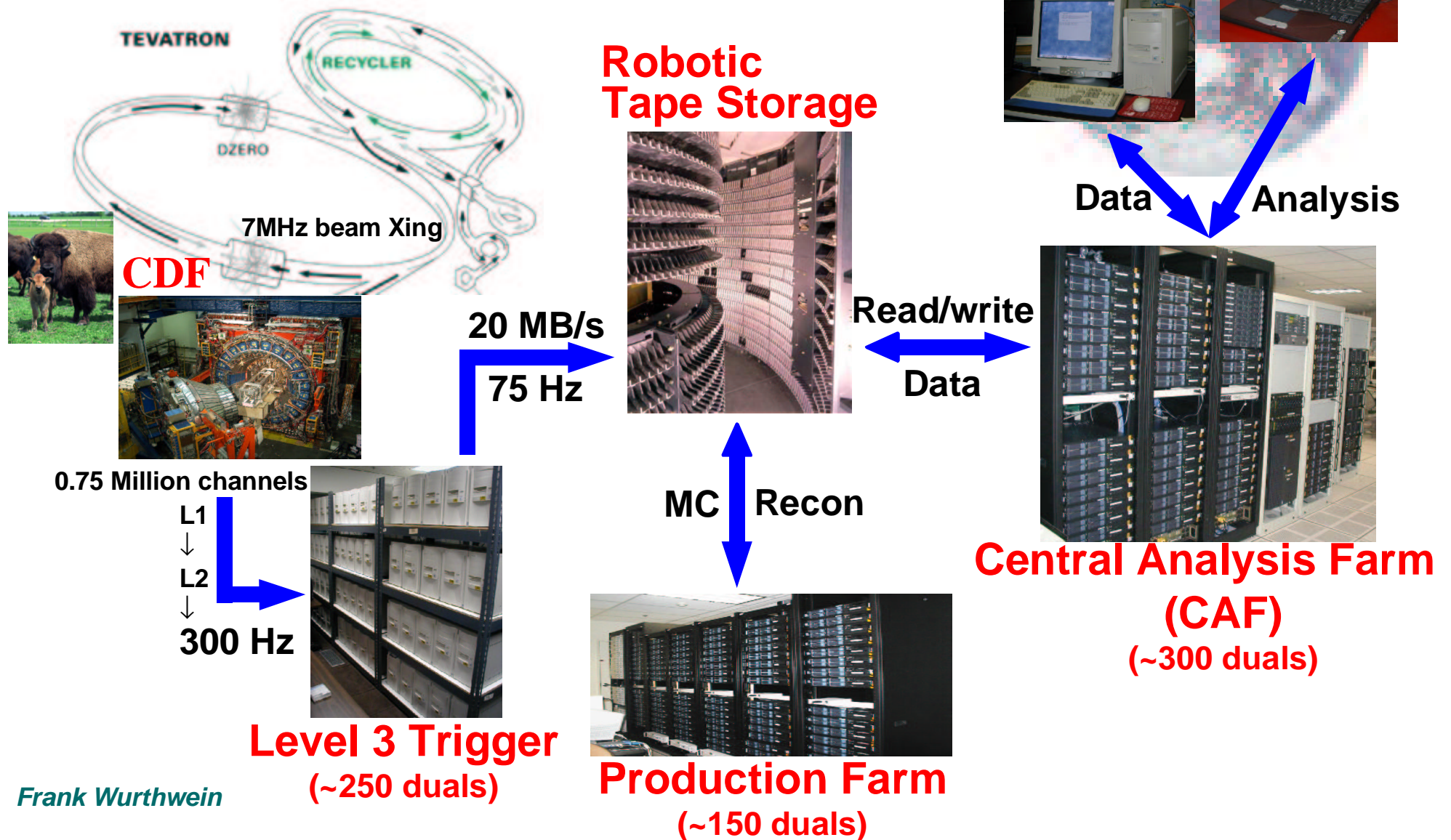
# User Analysis Computing at CDF

**Frank Wurthwein**  
*MIT/UCSD/FNAL-CD*  
for the CDF Collaboration

- **Computing Model**
- **Central Analysis Farm**
  - **User perspective**
  - **Cluster performance**
- **Future directions**



# CDF DAQ/Analysis Flow





# Data/Software Characteristics

## Data Characteristics:

- Root I/O: ~80-400 kB/event (configurable content)
- 'Standard' ntuple: 5-10 kB/event
- Typical RunIIa secondary dataset size:  $10^7$  events
- Winter03 physics: ~100 datasets adding up to ~50TB
- **Largest dataset for Winter03 physics:  $3.5e7$  evts**

## Analysis Software:

- Typical analysis jobs run @ 5 Hz on 1 GHz P3  
→ few MB/sec
- CPU rather than I/O bound (FastEthernet)



# Computing Requirements



Requirements set by goal:

200 simultaneous users to analyze secondary data set ( $10^7$  evts) in a day

Need ~700 TB of disk and ~5 THz of CPU by end of FY'05:

→ need lots of disk → need cheap disk → IDE Raid

→ need lots of CPU → commodity CPU → dual Intel/AMD



# Computing Model

## Interactive Computing on desktop:

- **Complete access to all data from desktop via dCache & rootd** (see talks by Kennedy, Litvinsev, Moibenko, Ernst)

## Batch Computing on "remote" cluster(s):

- **Binary compatible with desktop**
- **qsub, qstat, kill, ls, tail, top via command line/web**
- **Large scale parallelisation with single submission**
  - ➔ **Single summary email upon completion**
- **User scratch space inside cluster**
  - ➔ **Krb5 ticket created @ launch time**
- **Data access Winter03: 90% NFS+rootd, 10% dCache**





# Example job submission

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

- Fill in appropriate fields & submit job

section integer range

The screenshot shows the 'CDF RunII CAF GUI' window. The 'Initial Command' field is set to '/simple.sh'. The 'Process Type' is 'Short'. The 'Original Directory' is '/home/msn/releases/development/CafUtil/examples'. The 'Output File Location' is 'msn@fcdlnx2.fnal.gov/cdf/scratch/msn/temp.tgz'. The 'Email?' checkbox is checked, and the 'Email Address' is 'msn@fnal.gov'. There are 'Submit' and 'Quit' buttons. A log window at the bottom shows 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
```

output destination

user exe+tcl directory

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

# Web Monitoring of User Queues

Each user a different queue

Process type for job length

**test:** 5 mins

**short:** 2 hrs

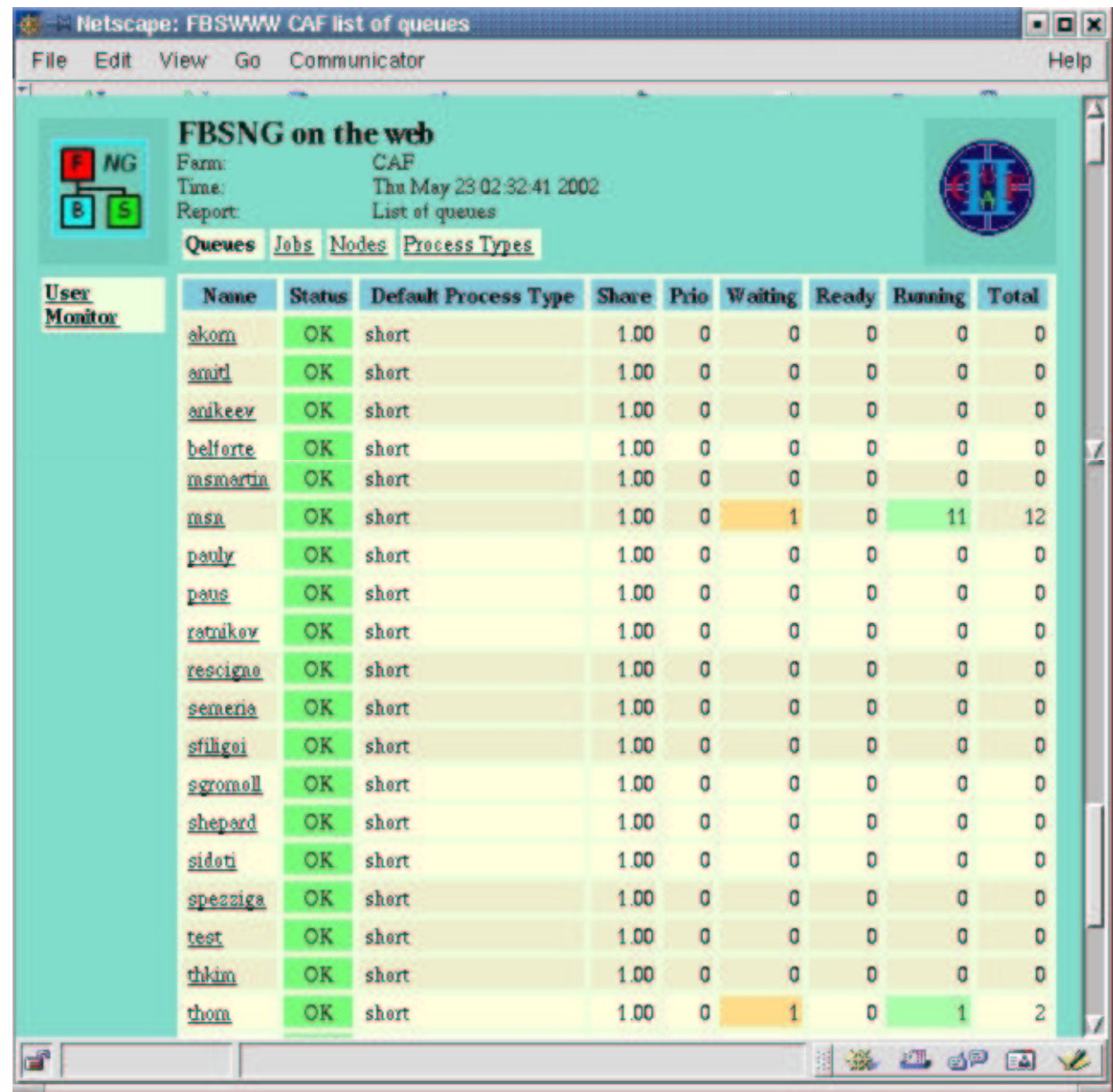
**medium:** 6 hrs

**long:** 2 days

This example:

1 job → 11 sections

(+ 1 additional section automatic for job cleanup)



Name	Status	Default Process Type	Share	Prio	Waiting	Ready	Running	Total
akorn	OK	short	1.00	0	0	0	0	0
amitl	OK	short	1.00	0	0	0	0	0
andkeev	OK	short	1.00	0	0	0	0	0
belforte	OK	short	1.00	0	0	0	0	0
msmartin	OK	short	1.00	0	0	0	0	0
msa	OK	short	1.00	0	1	0	11	12
pauly	OK	short	1.00	0	0	0	0	0
paus	OK	short	1.00	0	0	0	0	0
ratnikov	OK	short	1.00	0	0	0	0	0
rescigno	OK	short	1.00	0	0	0	0	0
semeria	OK	short	1.00	0	0	0	0	0
sfiligoi	OK	short	1.00	0	0	0	0	0
sgromoll	OK	short	1.00	0	0	0	0	0
shepard	OK	short	1.00	0	0	0	0	0
sidoti	OK	short	1.00	0	0	0	0	0
speziga	OK	short	1.00	0	0	0	0	0
test	OK	short	1.00	0	0	0	0	0
thkim	OK	short	1.00	0	0	0	0	0
thom	OK	short	1.00	0	1	0	1	2

## Monitoring jobs in your queue

Netscape: FBSWWW CAF list of queues

File Edit View Go Communicator Help

<u>masmactin</u>	OK	short	1.00	0	0	0	0	0	0
<u>pas</u>	OK	short	1.00	0	0	0	0	0	0
<u>pasn</u>	OK	short	1.00	0	0	0	0	0	0
<u>ratuko</u>	OK	short	1.00	0	0	0	0	0	0
<u>rescigno</u>	OK	short	1.00	0	0	0	0	0	0
<u>semeria</u>	OK	short	1.00	0	0	0	0	0	0
<u>shilgoi</u>	OK	short	1.00	0	0	0	0	0	0
<u>sgromoll</u>	OK	short	1.00	0	0	0	0	0	0
<u>shepard</u>	OK	short	1.00	0	0	0	0	0	0
<u>sidoti</u>	OK	short	1.00	0	0	0	0	0	0
<u>speaziga</u>	OK	short	1.00	0	0	0	0	0	0
<u>test</u>	OK	short	1.00	0	0	0	0	0	0
<u>thkim</u>	OK	short	1.00	0	0	0	0	0	0
<u>thom</u>	OK	short	1.00	0	1	0	0	0	1

File

User  
Moni

Netscape: FBSWWW - queue msn@CAF

File Edit View Go Communicator Help

**FBSNG on the web**

Farm: CAF  
Time: Thu May 23 01:47:23 2002  
Report: Queue msn

[Queues](#) [Jobs](#) [Nodes](#) [Process Types](#)

**User Monitor**

Queue Parameters [\[show\]](#)

Status: **OK** Running: 11 Pending: 0

SectID	User	ProcType	Status	Prio	NProc	Date/Time
<a href="#">873.msn_600</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:09
<a href="#">873.msn_601</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:09
<a href="#">873.msn_602</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:10
<a href="#">873.msn_603</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:10
<a href="#">873.msn_604</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:11
<a href="#">873.msn_605</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:11
<a href="#">873.msn_606</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:12
<a href="#">873.msn_607</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:12
<a href="#">873.msn_608</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:12
<a href="#">873.msn_609</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:13
<a href="#">873.msn_610</a>	cdlcaf	<a href="#">short</a>	running	0	1/1	Started at 05/23 01:47:13
<a href="#">873.msn_end</a>	cdlcaf	<a href="#">mailer</a>	waiting	0	0/1	Submitted at 05/23 01:46:57

FCS Group | FBSNG

FBSWWW version 0.3



## Monitoring sections of your job

**FBSNG on the web**  
Farm: CAF  
Time: Thu May 23 01:47:23 2002  
Report: Queue msn

**User Monitor**

SectID	User	ProcType
873.msn_601	cdcfat	short
873.msn_602	cdcfat	short
873.msn_603	cdcfat	short
873.msn_604	cdcfat	short
873.msn_605	cdcfat	short
873.msn_606	cdcfat	short
873.msn_607	cdcfat	short
873.msn_608	cdcfat	short
873.msn_609	cdcfat	short
873.msn_610	cdcfat	short
873.msn_end	cdcfat	mailer

**FBSNG on the web**  
Farm: CAF  
Time: Thu May 23 01:48:13 2002  
Report: Section 873.msn\_600 status

**User Monitor**

ID: 873.msn\_600 User: cdcfat  
Queue: msn Process Type: short  
NProc: 1 Status: **running**  
Need: 0 Depends:  
Submitted: 05/23 01:46:57 Started: 05/23 01:47:09  
CPU time limit: 2h00m  
Proc Rsrc: cpu:100 disk:15 Sect Rsrc:

Command: /fbsng/caflcal/v1.01/CafExe cdcfat@fcdthead1.fnal.gov/home/cdcfat/v1.01/submitter/cafln/msn\_%s.tgz msn@fcdflx2.fnal.gov/cdf/scratch/msn/temp600.tgz msn 4h  
cdcfat@fcdthead1.fnal.gov/home/cdcfat/v1.01/submitter/fbs/FBS\_%s.msn\_600.1.log /simple.sh 600

Other sections: msn\_600 (running) msn\_601 (running) msn\_602 (running) msn\_603 (running) msn\_604 (running) msn\_605 (running) msn\_606 (running) msn\_607 (running) msn\_608 (running) msn\_609 (running) msn\_610 (running) msn\_end (waiting)

**Processes**

Process #	Node	Status	CPU Time	PID	Command
1	fcdcfat057	running	0	6931	CafExe cdcfat@fcdthead1.fnal.gov/home/cdcfat/v1.01/submitter/cafln/msn_%s.tgz msn@fcdflx2.fnal.gov
			0	6940	simple.sh 600
			0	7221	sleep 120

FBS Group | FBSNG  
FBSWWW version 0.1



# Detailed life of a job

## At Submission:

- Krb5 authenticate
- Check output
- Store exe "sandbox"
- Submit to local batch
- Return JID to user

## At Launch:

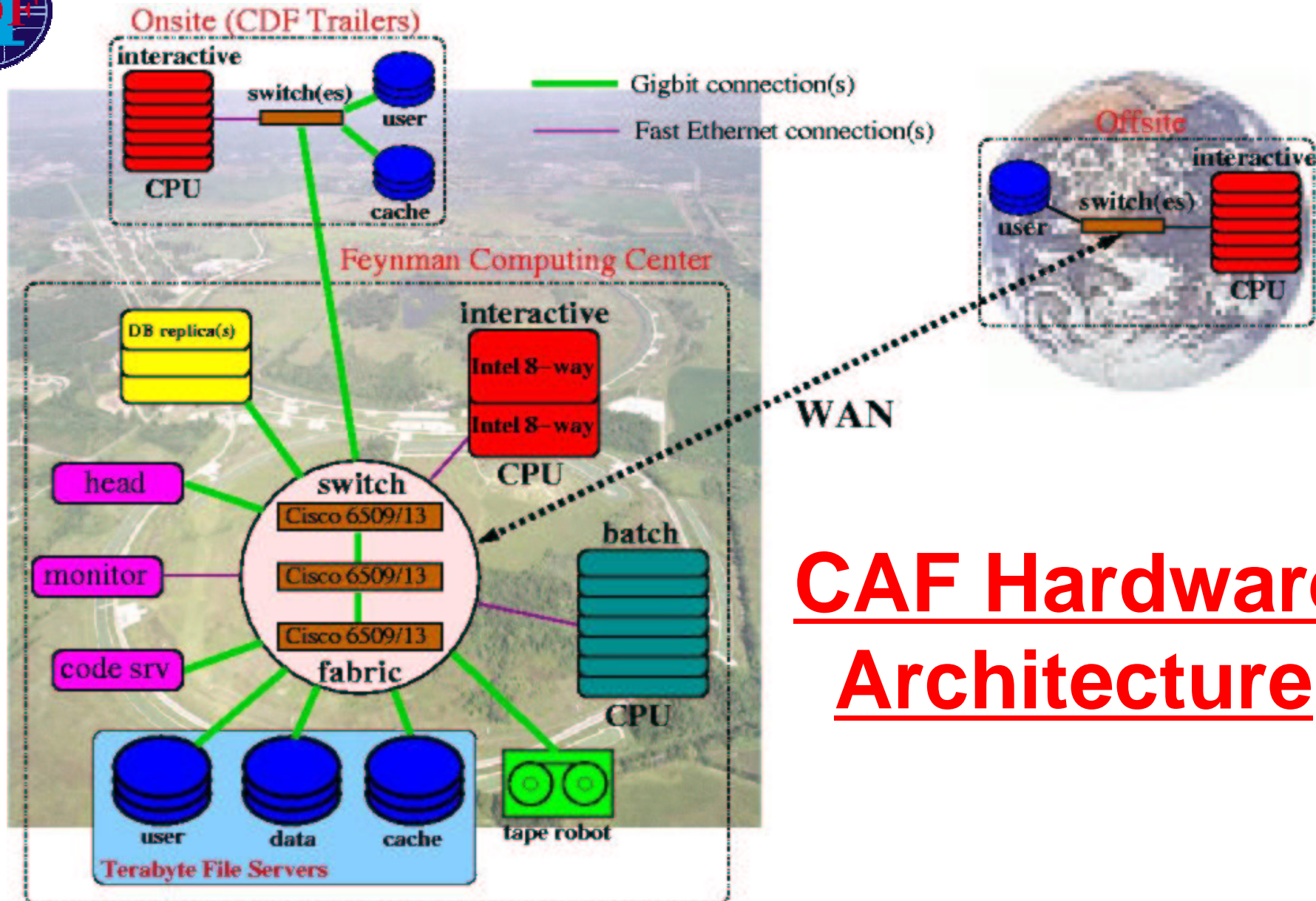
- CafExe is launched
  - Fcp sandbox
  - Untar onto local disk
  - Get user krb5 ticket
  - Start user shell script

## After termination:

- User shell completes
- CafExe tars up "sandbox"
- CafExe rcp sandbox to output
- CafExe store log
- CafExe completes

## At job completion:

- Mailer parse all logs
- Mailer send summary email



# CAF Hardware Architecture





# CAF Hardware



Code Server

File Servers

Worker Nodes

Linux 8-ways  
(interactive)





# Hardware: Servers



**Servers (~180TB total, 92 4U servers):**

IDE RAID50 hot-swap

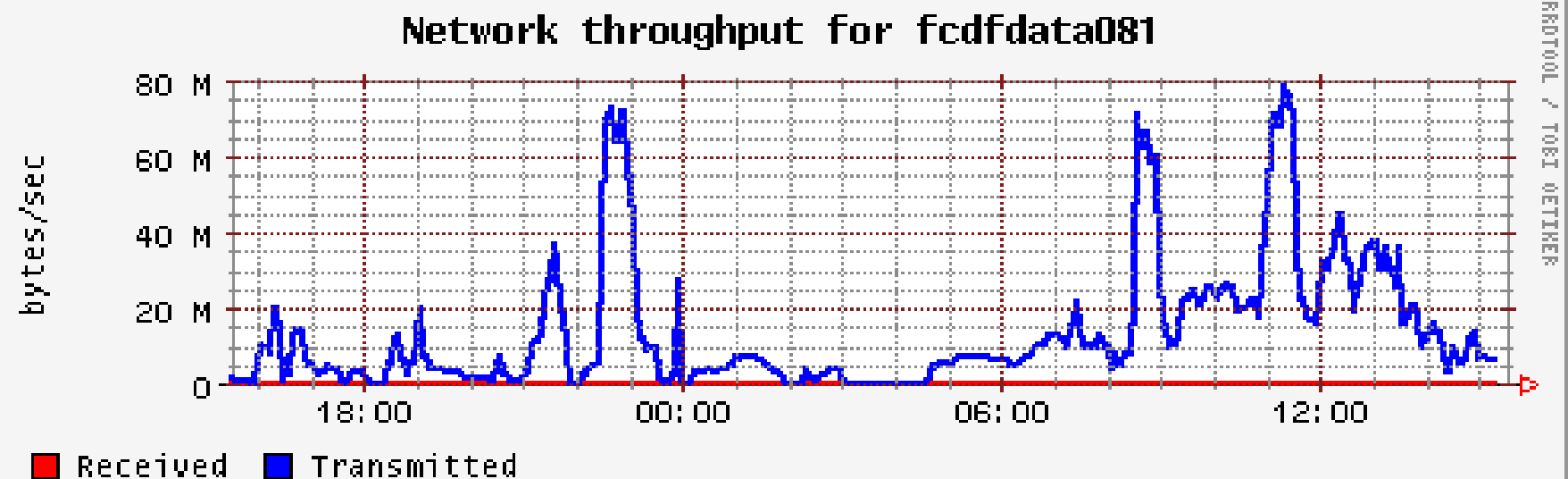
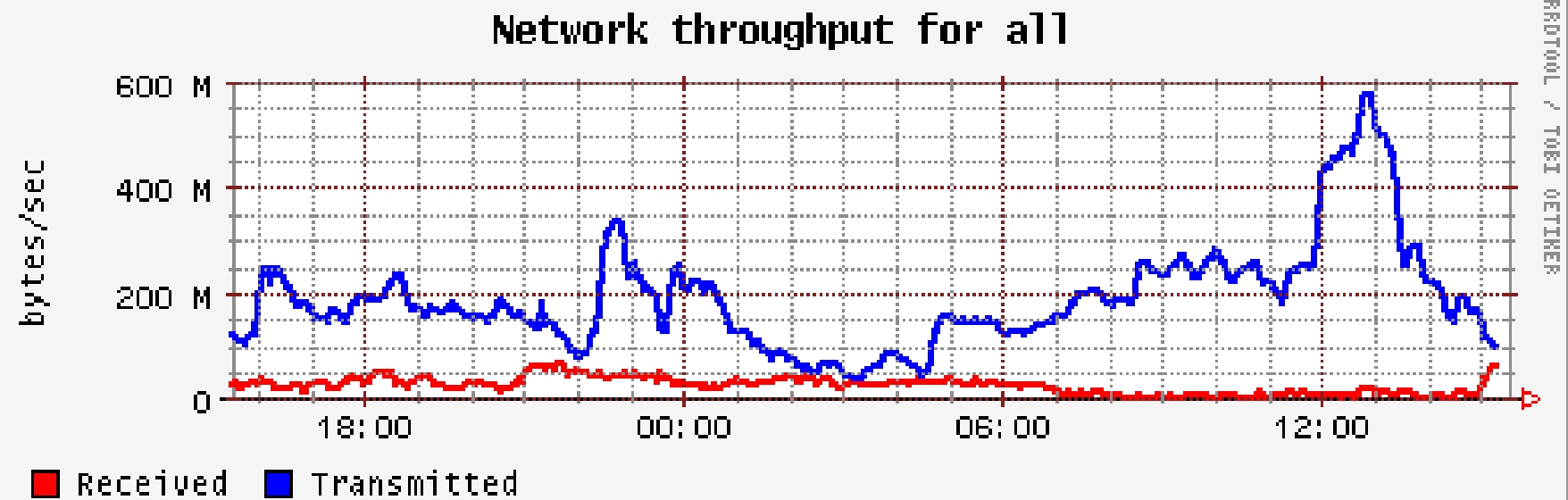
Dual P3 1.4GHz / 2GB RAM

SysKonnnect 9843 Gigabit Ethernet card





## Server I/O example (3/13/03)







# Hardware: Workers



**Workers (600 CPUs, 1U+2U rackmount):**

- 16 2U Dual Athlon 1.6GHz / 512MB RAM
- 48 1U/2U Dual P3 1.26GHz / 2GB RAM
- 236 1U Dual Athlon 1.8GHz / 2GB RAM
- FE (11 MB/s) / 80GB job scratch each





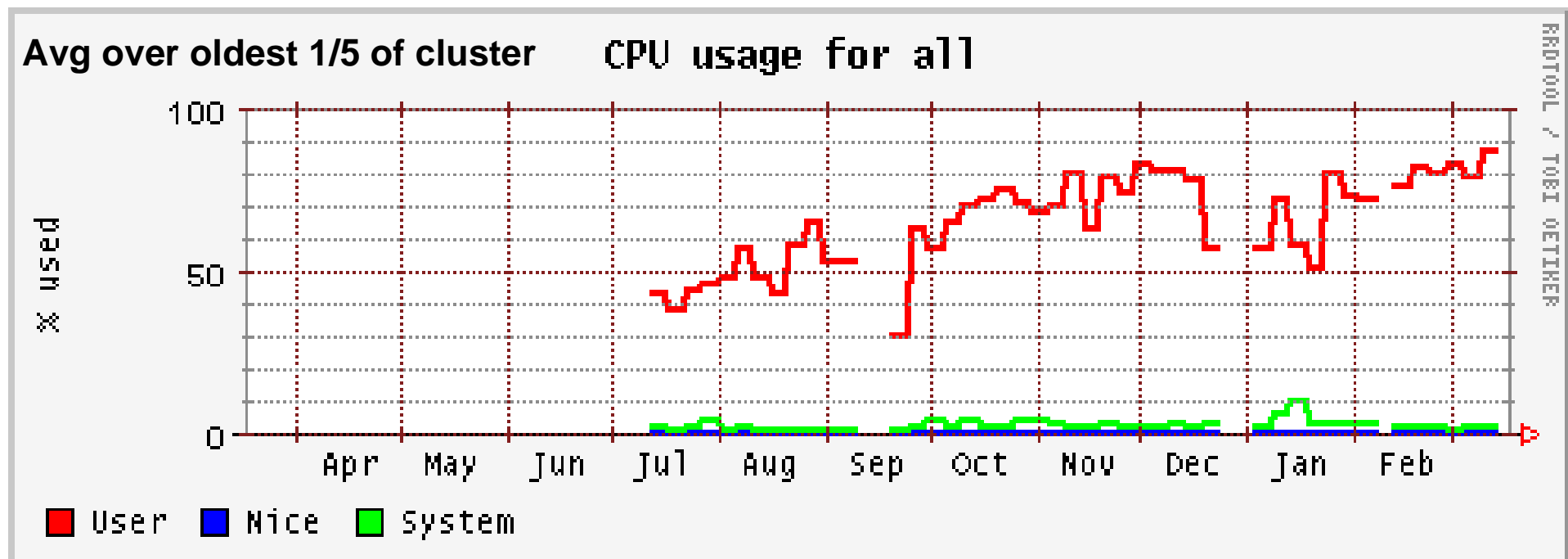
# CAF utilization

## User perspective:

- 10,000 jobs launched/day
- 400 users total
- 100 users per day

## System perspective:

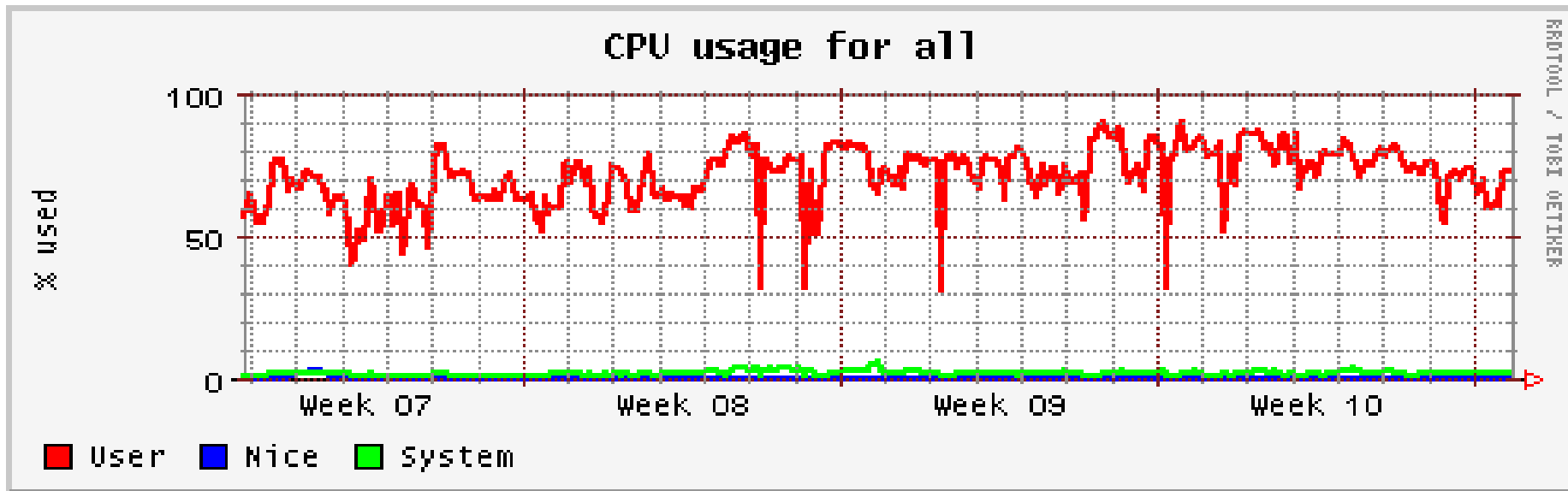
- Up to 90% avg CPU utilization
- 200-600MB/sec I/O
- Failure rate ~1/2000
- Avg uptime of WN = 60days



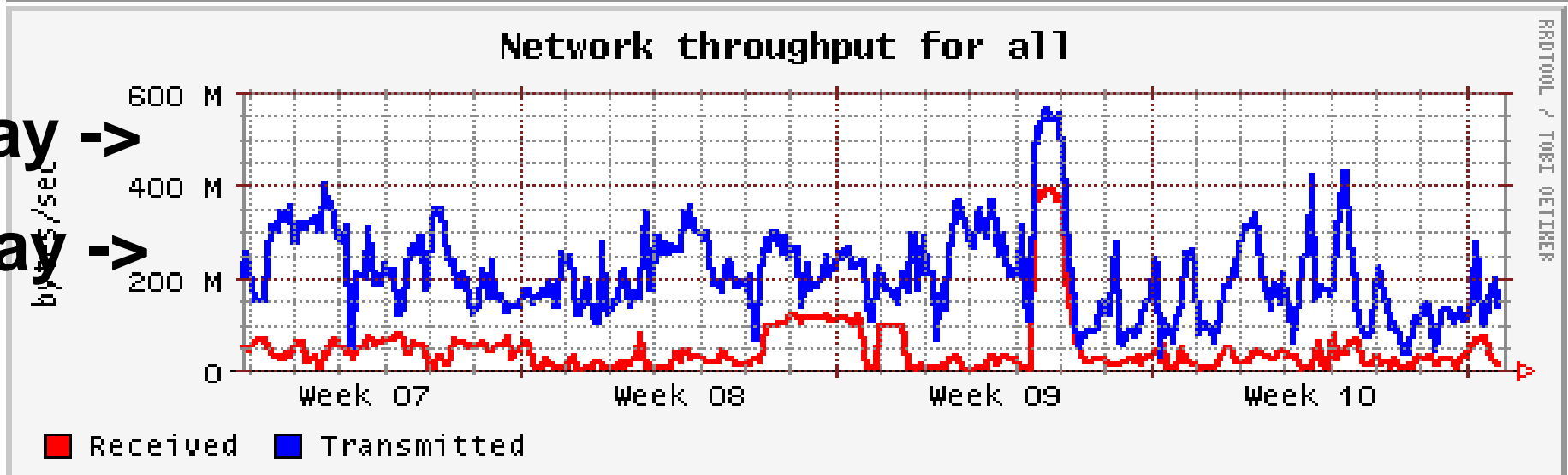




# CAF utilization last month

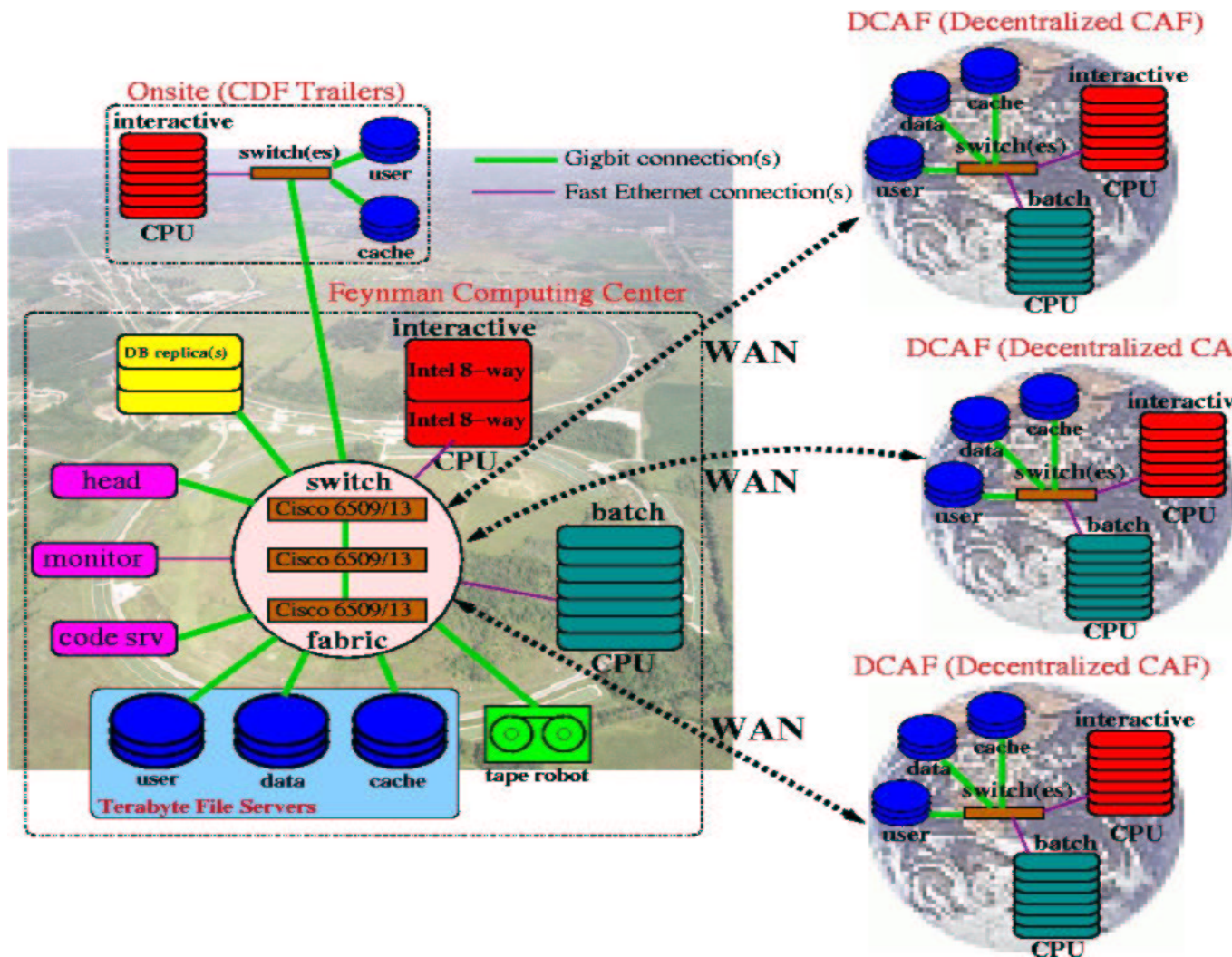


TB/day ->  
TB/day ->





# Future Directions



See talks by:

- Stonjek
- Ratnikov
- Terekhov
- Garzogli



## **Conclusion**

**User analysis computing based on commodity PC's**

**180TB disk space      1THz batch CPU**

**Focus on building strong infrastructure**

**up to 600MB/sec I/O      99.95% reliability**

**that has been deployed as part of CDF grid  
"proof of principle" for SC2002 demo.**