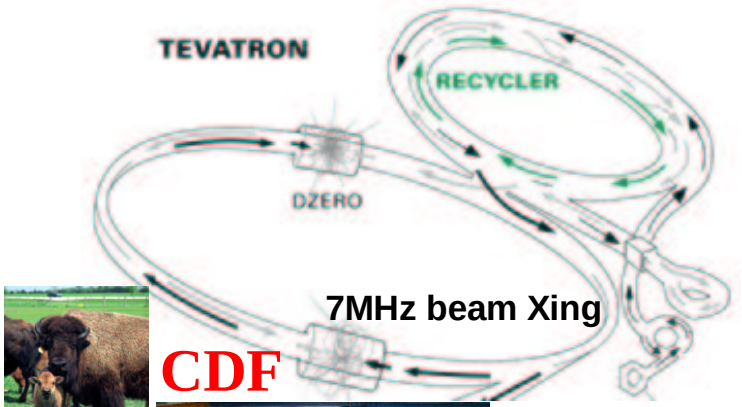
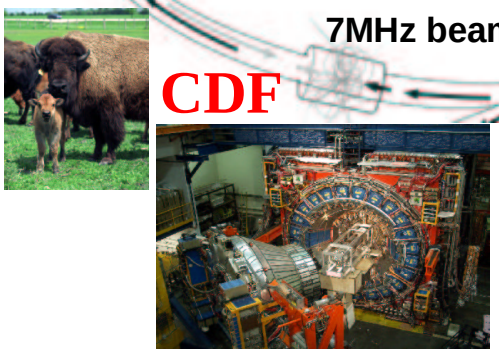
User Analysis Computing at CDF

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





0.75 Million channels

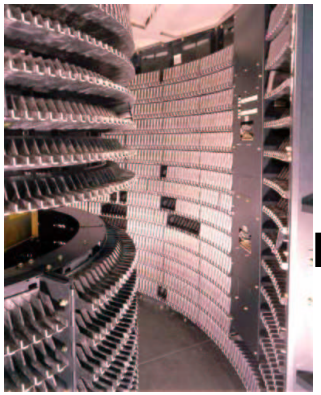
L1
↓
L2
↓
300 Hz

Level 3 Trigger
(~250 duals)

20 MB/s

75 Hz

Robotic Tape Storage



MC Recon

Production Farm
(~150 duals)

Read/write
Data

User Desktops

Data Analysis

Central Analysis Farm (CAF)
(~300 duals)

Frank Wurthwein

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
  $\rightarrow$ 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)

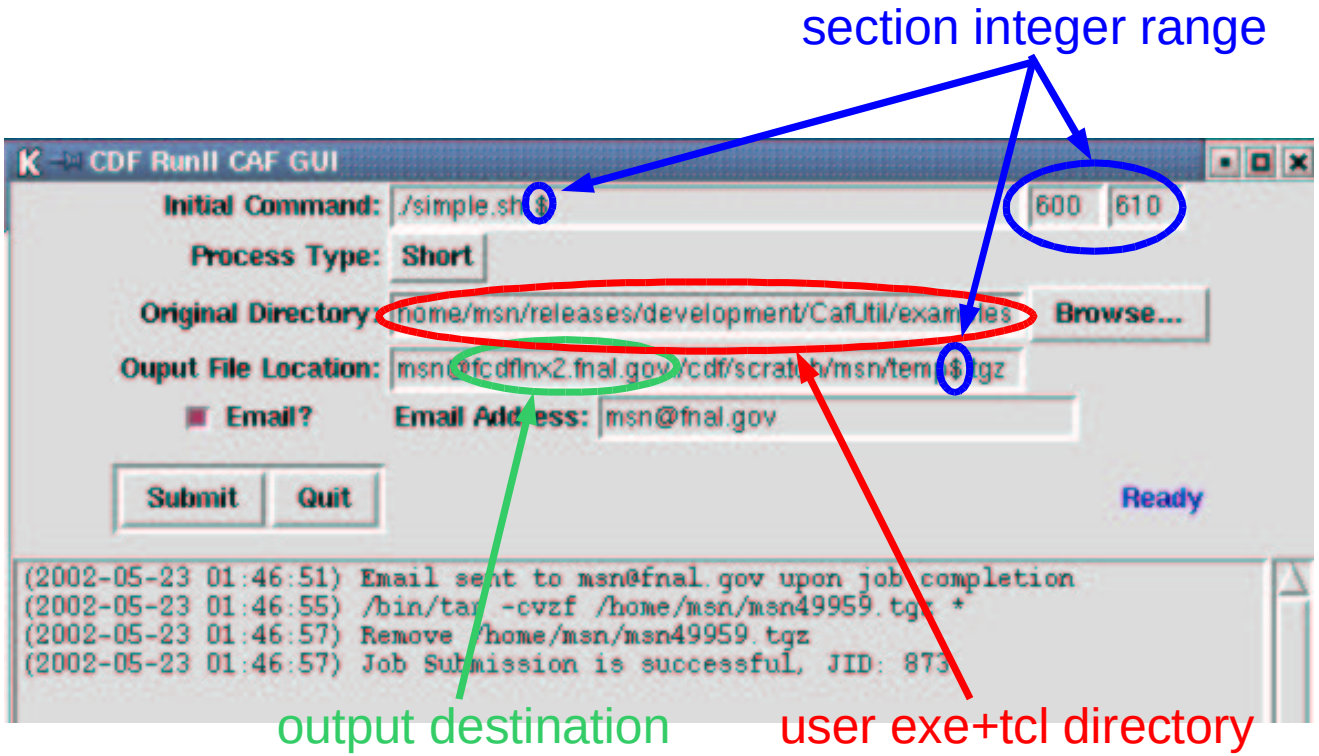
Batch Computing on "remote" cluster(s):

- 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



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

Web Monitoring of User Queues

Each user a different **queue**

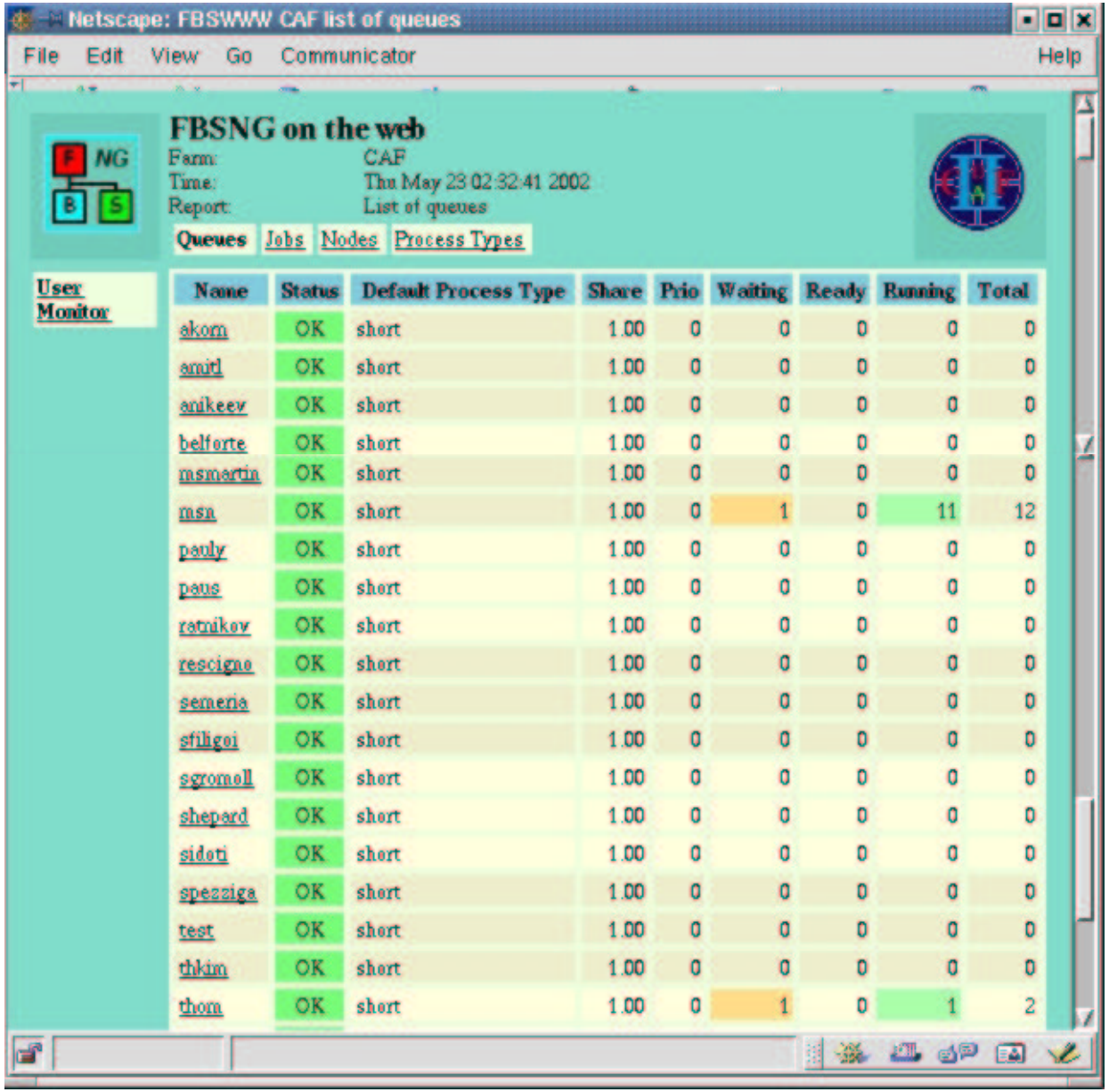
Process type for job length

- **test**: 5 mins
- **short**: 2 hrs
- **medium**: 6 hrs
- **long**: 2 days

This example:

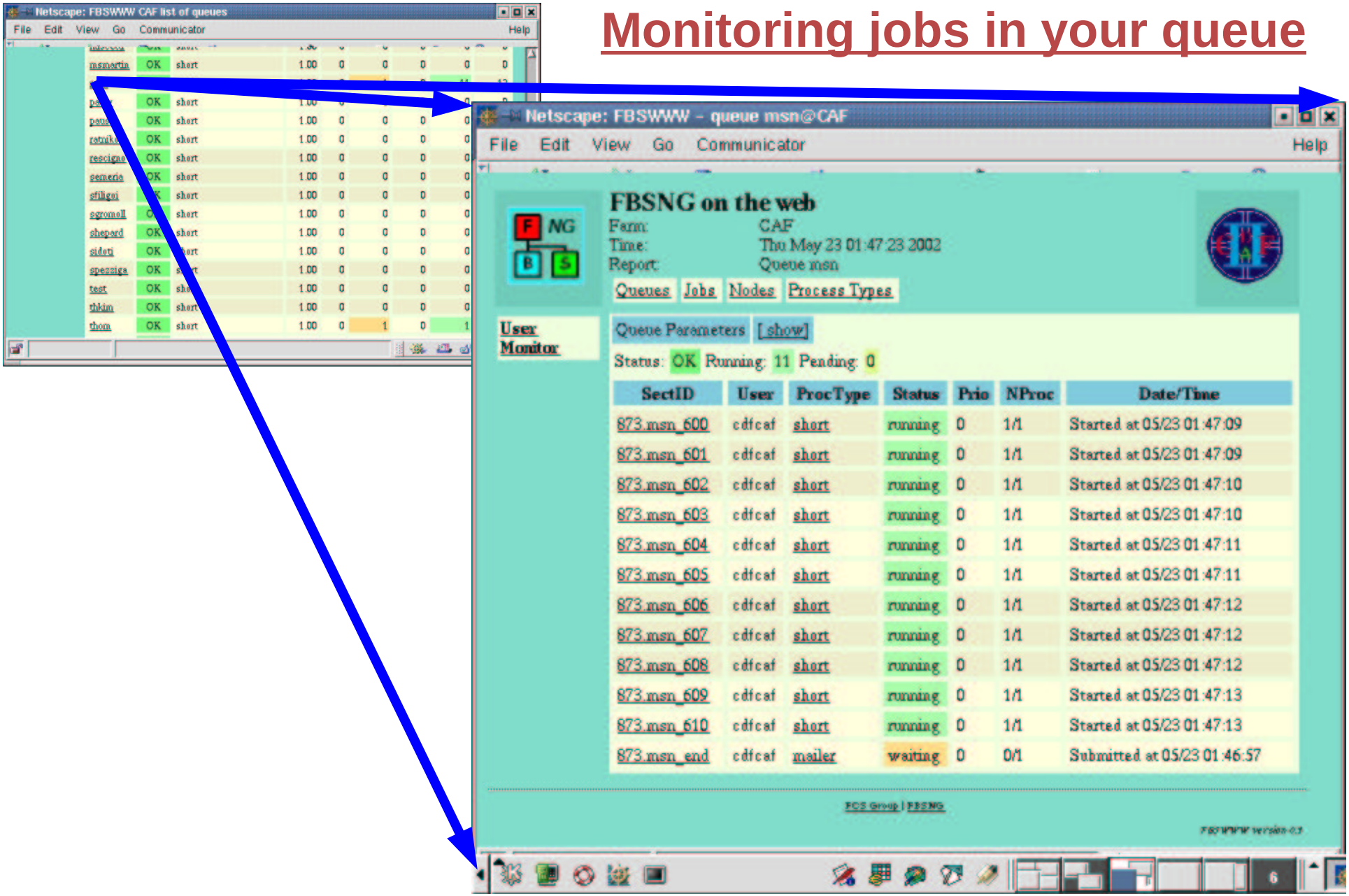
1 job → 11 sections

(+ 1 additional section automatic for job cleanup)

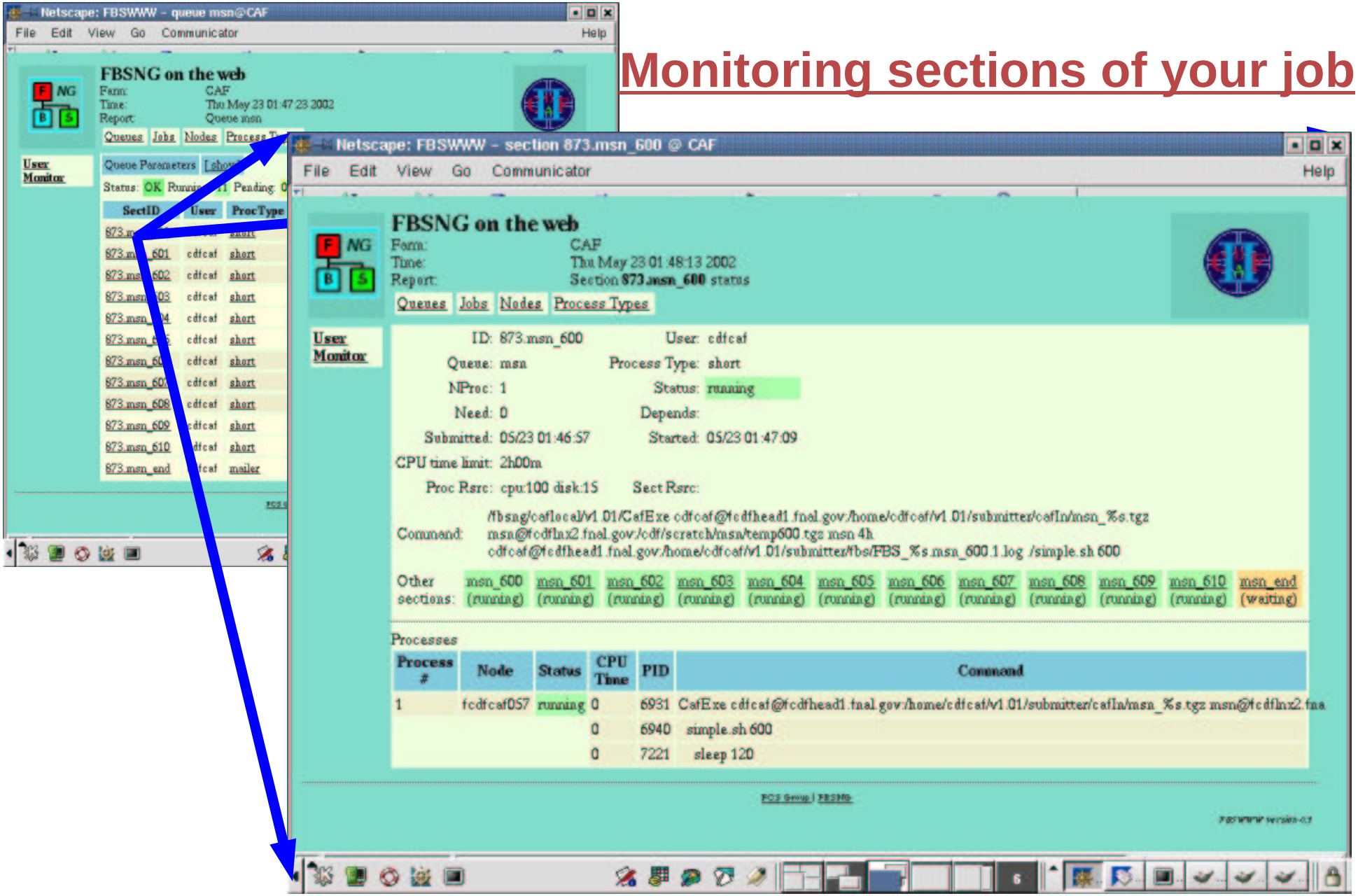


Name	Status	Default Process Type	Share	Prio	Waiting	Ready	Running	Total
akom	OK	short	1.00	0	0	0	0	0
amitl	OK	short	1.00	0	0	0	0	0
anikeev	OK	short	1.00	0	0	0	0	0
belforte	OK	short	1.00	0	0	0	0	0
msmartin	OK	short	1.00	0	0	0	0	0
msn	OK	short	1.00	0	1	0	11	12
pauly	OK	short	1.00	0	0	0	0	0
paus	OK	short	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
spezziga	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



Monitoring sections of your job



Detailed life of a job

At Submission:

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

At Launch:

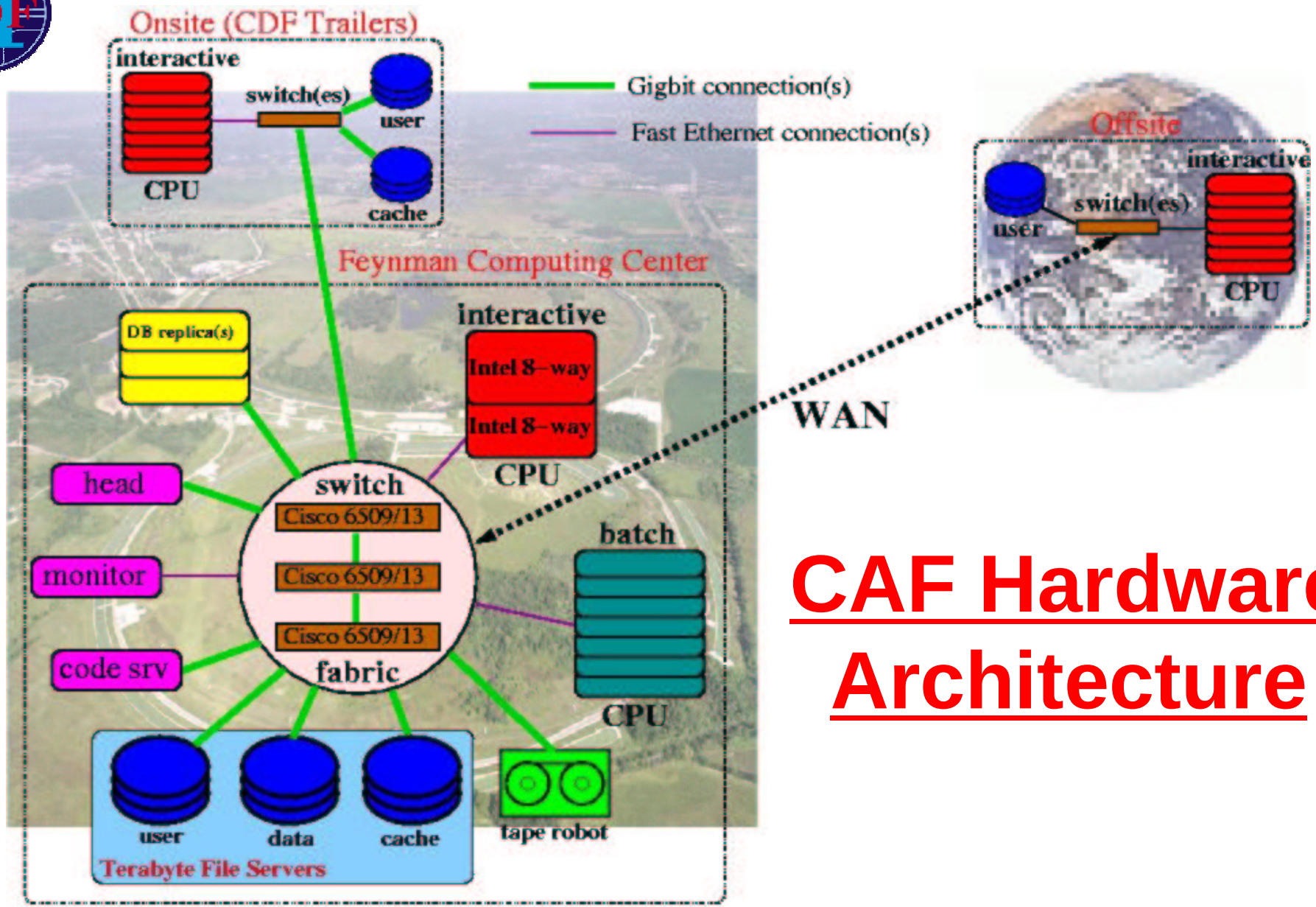
- CafExe is launched
  - Fcp sandbox
  - Untar onto local disk
  - Get user krb5 ticket
  - Start user shell script

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

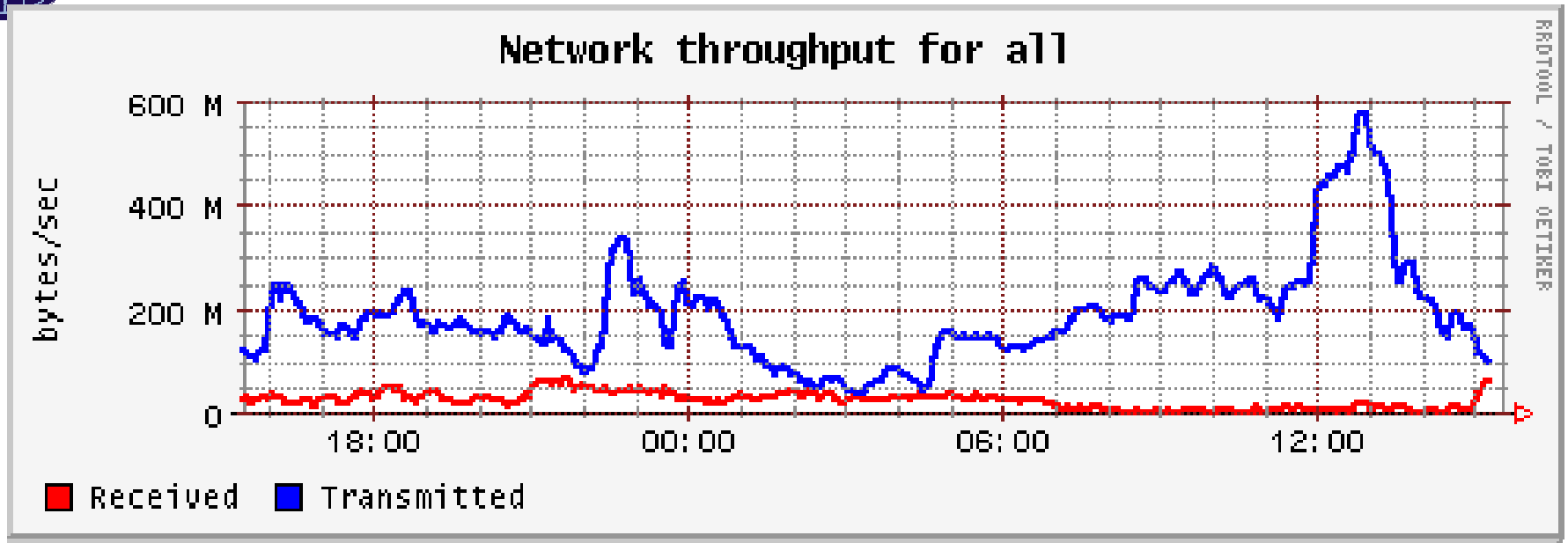
Linux 8-ways
(interactive)

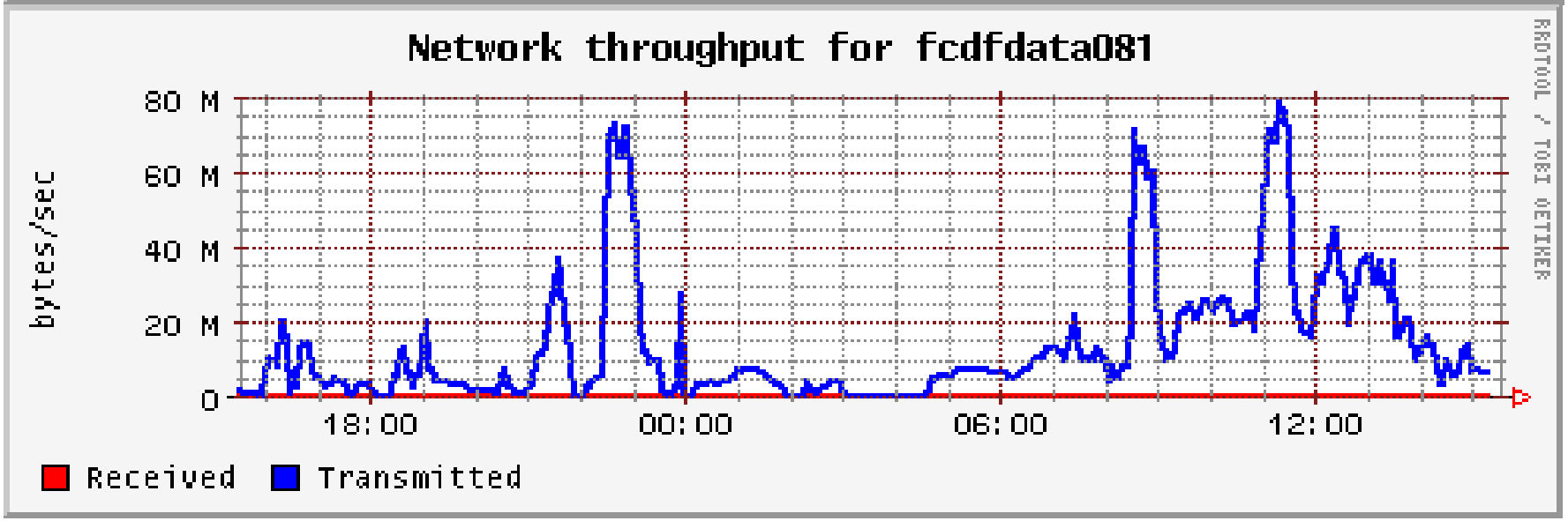
Hardware: Servers

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

- IDE RAID50 hot-swap
- Dual P3 1.4GHz / 2GB RAM
- SysKonnect 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

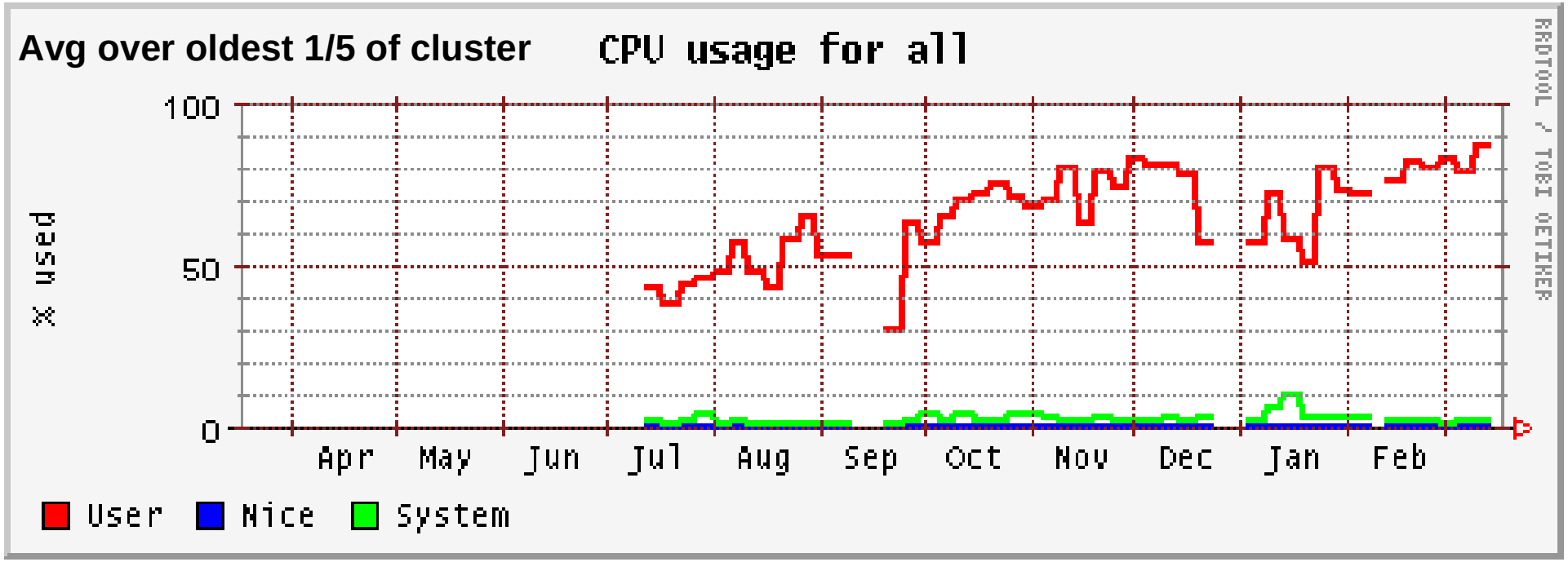
CAF utilization

User perspective:

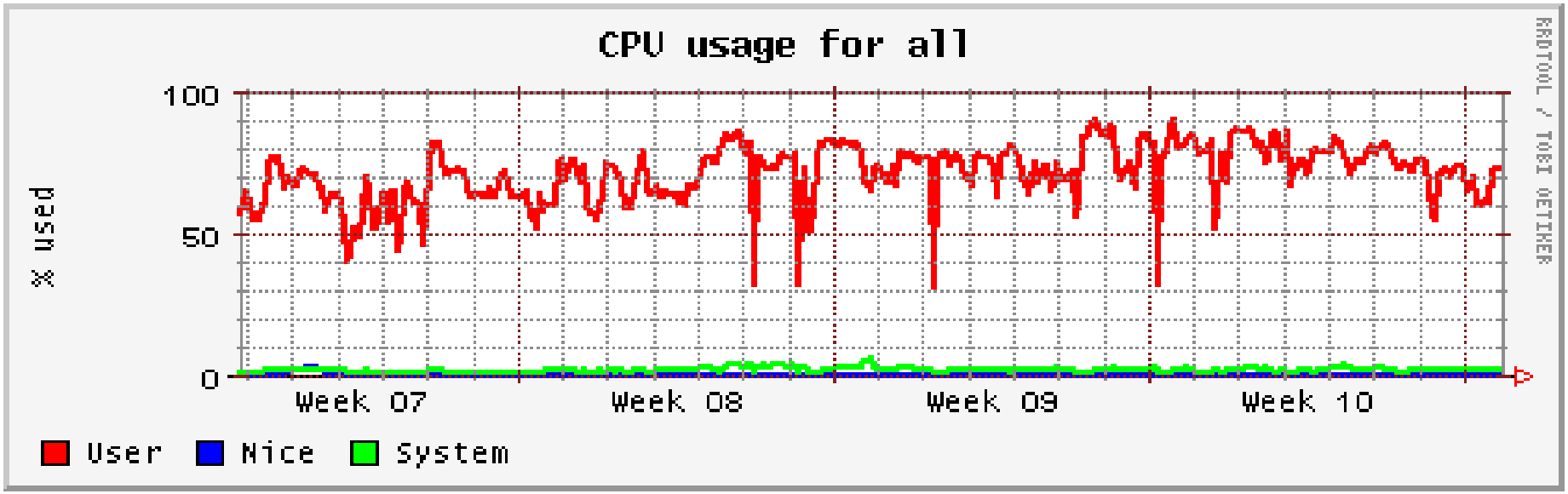
- 10,000 jobs launched/day
- 400 users total
- 100 users per day

System perspective:

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

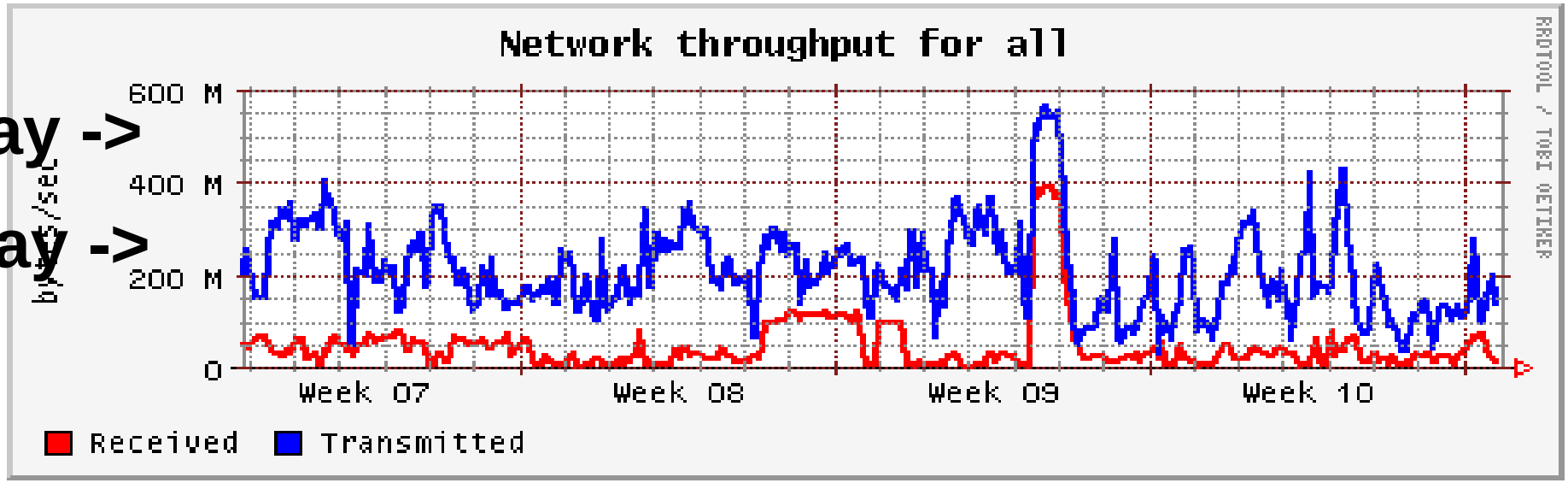


CAF utilization last month

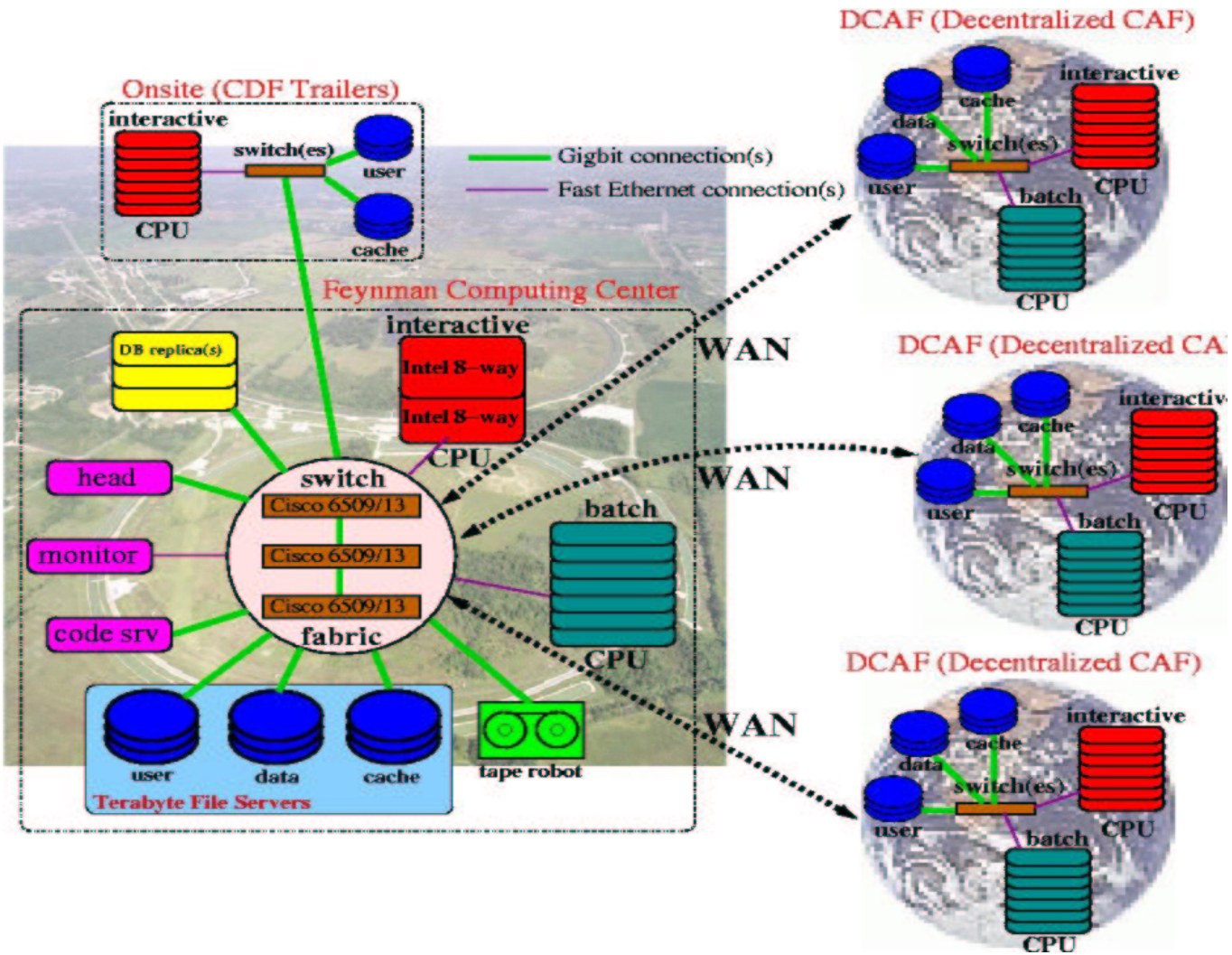


TB/day ->

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.