CAF status at CDF

Frank Wurthwein
MIT/FNAL-CD

- Introduction
- Hardware
- User perspective
- Conclusions

Data/Software Characteristics

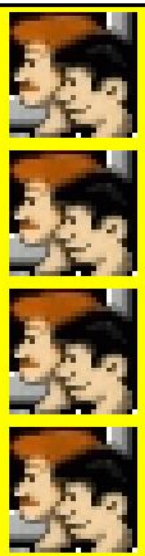
Data Characteristics:

- Root I/O sequential for raw data: ~250 kB/event
- Root I/O multi-branch for reco data: 50-100 kB/event
- 'Standard' ntuple: 5-10 kB/event
- Typical RunIIa secondary dataset size: 10^7 events

Analysis Software:

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

Past CAF Computing Model



Large SMP (128 processor SGI)
Expensive disks (FiberChannel/SCSI)

Analysis Code Development
Analysis Job Debugging
Interactive Analysis Jobs
Batch Jobs
"Other" Usage

Very expensive to expand and maintain

Bottom line:

Not enough 'bang for the buck'

Design Challenges

develop/debug interactively @ remote desktop

✓ code management & rootd

Send binary & 'sandbox' for execution on CAF

✓ kerberized gatekeeper

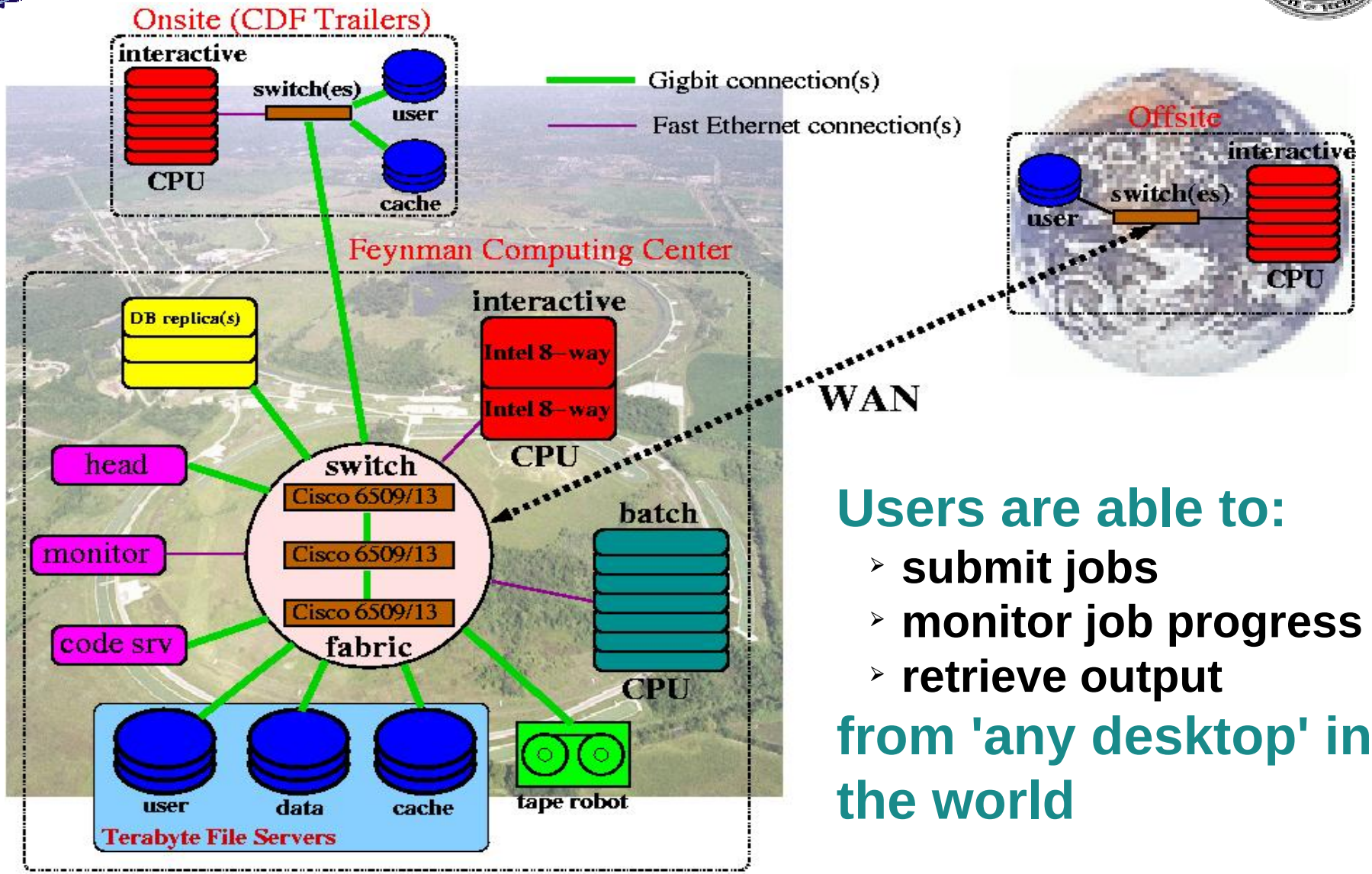
no user accounts on cluster

BUT

user access to scratch space with quotas

✓ creative use of kerberos

CAF Architecture



Users are able to:

- submit jobs
- monitor job progress
- retrieve output

from 'any desktop' in the world

CAF Milestones

- Start of CAF design — 11/01
- CAF prototype (protoCAF) assembled — 2/25/02
- Fully-functional prototype system (>99% job success) — 3/6/02
- ProtoCAF integrated into Stage1 system — 4/25/02
- Production Stage1 CAF for collaboration — 5/30/02

ProtoCAF

Stage1

Design → Production system in 6 months!

CAF Stage 1 Hardware

Code Server

File Servers

Worker Nodes

Linux 8-ways
(interactive)

Stage 1 Hardware: Workers

Workers (132 CPUs, 1U+2U rackmount):

- **16 2U Dual Athelon 1.6GHz / 512MB RAM**
- **50 1U/2U Dual P3 1.26GHz / 2GB RAM**
- **FE (11 MB/s) / 80GB job scratch each**

Stage 1 Hardware: Servers

Servers (35TB total, 16 4U rackmount):

- **2.2TB useable IDE RAID50 hot-swap**
- **Dual P3 1.4GHz / 2GB RAM**
- **SysKonnect 9843 Gigabt Ethernet card**

Stage2 Hardware

Worker nodes:

238 Dual Athlon MP2000+, 1U rackmount

1 THz of CPU power

File servers:

76 systems, 4U rackmount, dual red. Power supply
14 WD180GB in 2 RAID5 on 3ware 7500-8
2 WD40GB in RAID1 on 3ware 7000-2
1 GigE Syskonnect 9843
Dual P3 1.4GHz

150 TB disk cache

Stage1 Data Access

Static files on disk:

NFS mounted to worker nodes
remote file access via rootd

Dynamic disk cache:

dCache in front of Enstore robot

SAM on CAF

- Glasgow DCAF for tests ready.
- Could have SAM on CAF in a week if people pushed for it:
  - When to start a project
  - When to stop a project
  - How to avoid the ''Friday meltdown''

CAF: User Perspective

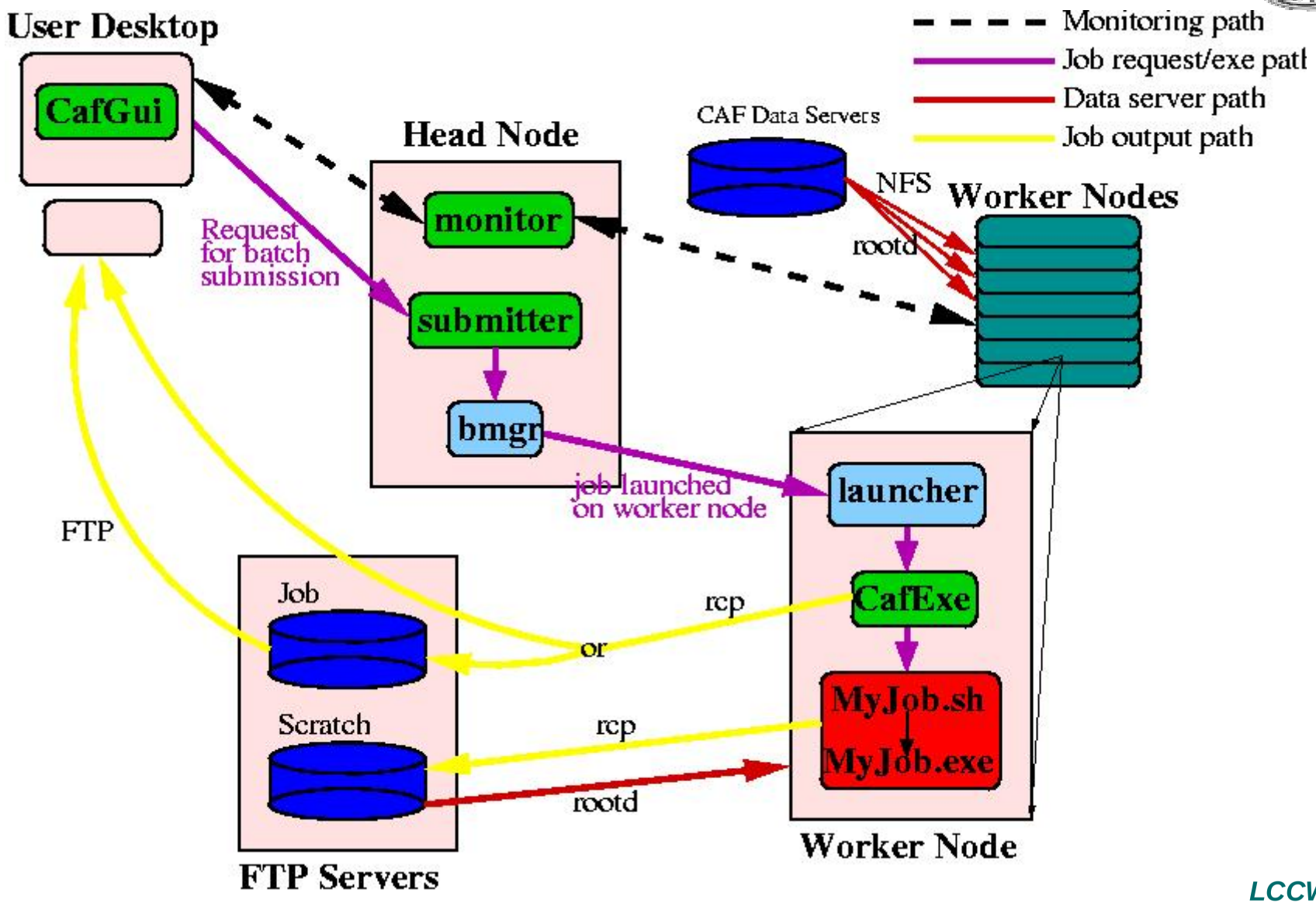
Job Related:

- Submit jobs
- Check progress of job
- Kill a job

Remote file system access:

- 'ls' in job's 'relative path'
- 'ls' in a CAF node's absolute path
- tail' of any file in job's 'relative path'

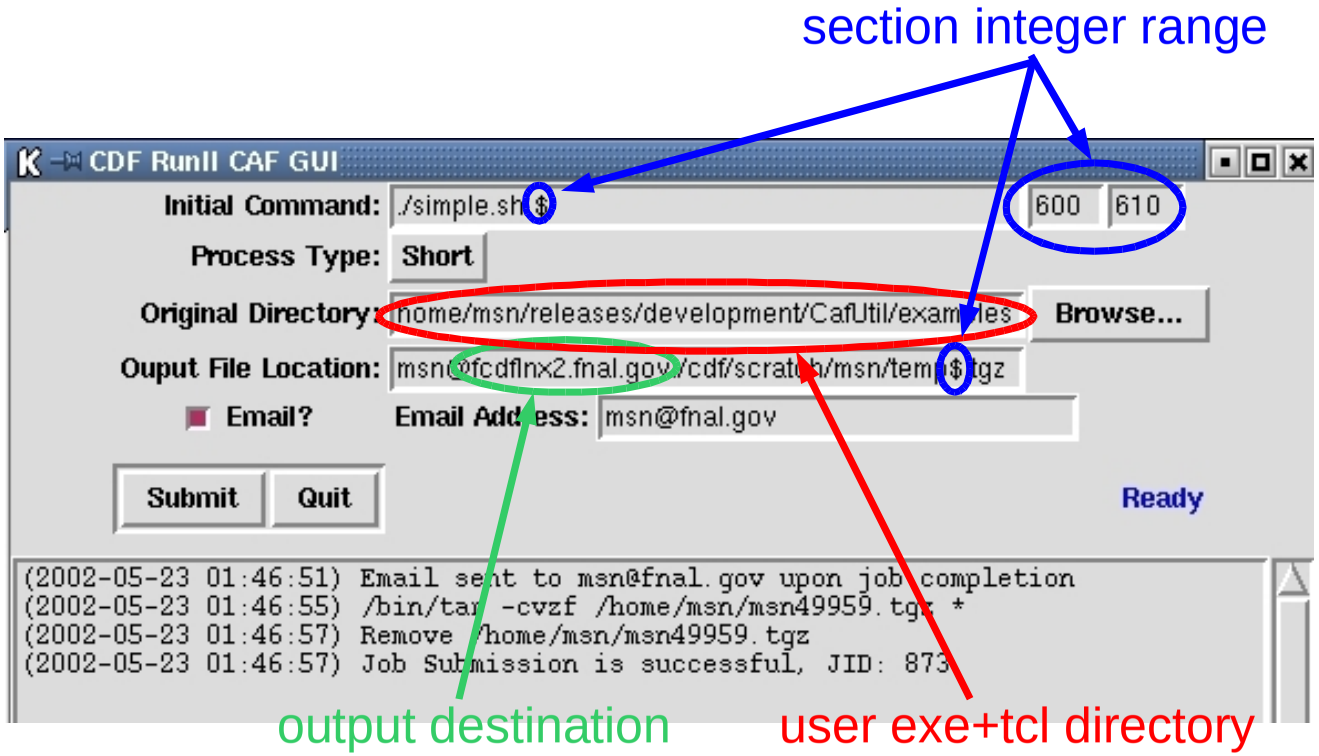
CAF Software



CAF User Interface

- 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

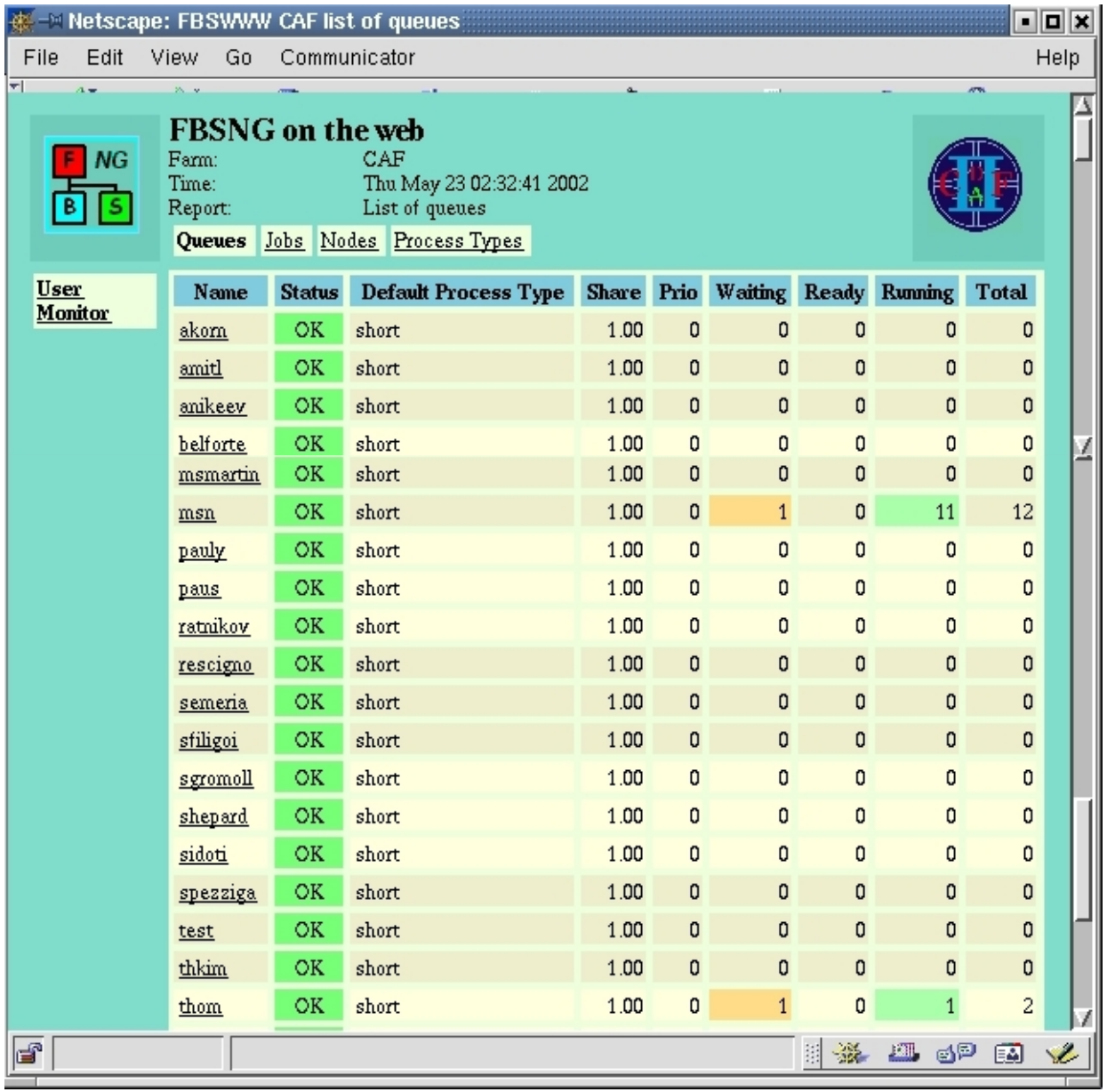
Process type for job length

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

This example:

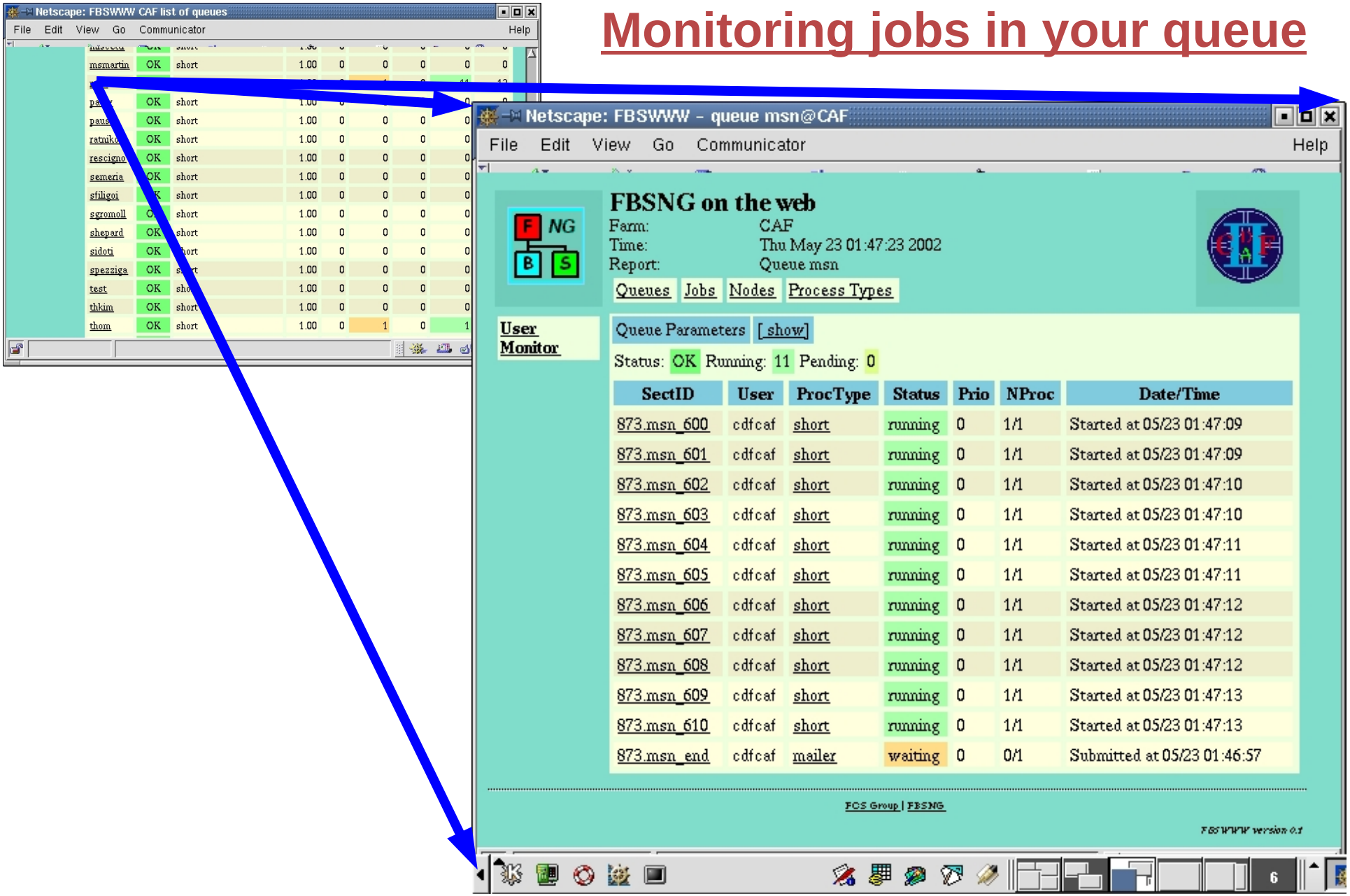
1 job → 11 sections

(+ 1 additional section automatic for job cleanup)



Name	Status	Default Process Type	Share	Prio	Waiting	Ready	Running	Total
akorn	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

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
873.msn_601	cdfcaf	short
873.msn_602	cdfcaf	short
873.msn_603	cdfcaf	short
873.msn_604	cdfcaf	short
873.msn_605	cdfcaf	short
873.msn_606	cdfcaf	short
873.msn_607	cdfcaf	short
873.msn_608	cdfcaf	short
873.msn_609	cdfcaf	short
873.msn_610	cdfcaf	short
873.msn_end	cdfcaf	mailer

Netscape: FBSWWW - section 873.msn_600 @ CAF

File Edit View Go Communicator Help

FBSNG on the web

Farm: CAF
Time: Thu May 23 01:48:13 2002
Report: Section **873.msn_600** status

Queues | Jobs | Nodes | Process Types

User Monitor

ID: 873.msn_600 User: cdfcaf
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/caflocal/v1.01/CafExe cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/cafIn/msn_%s.tgz msn@fcdflnx2.fnal.gov:/cdf/scratch/msn/temp600.tgz msn 4h cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/fbs/FBS_%s.msn_600.1.log ./simple.sh 600

Other sections:	msn_600	msn_601	msn_602	msn_603	msn_604	msn_605	msn_606	msn_607	msn_608	msn_609	msn_610	msn_end
	(running)	(running)	(running)	(running)	(running)	(running)	(running)	(running)	(running)	(running)	(running)	(waiting)

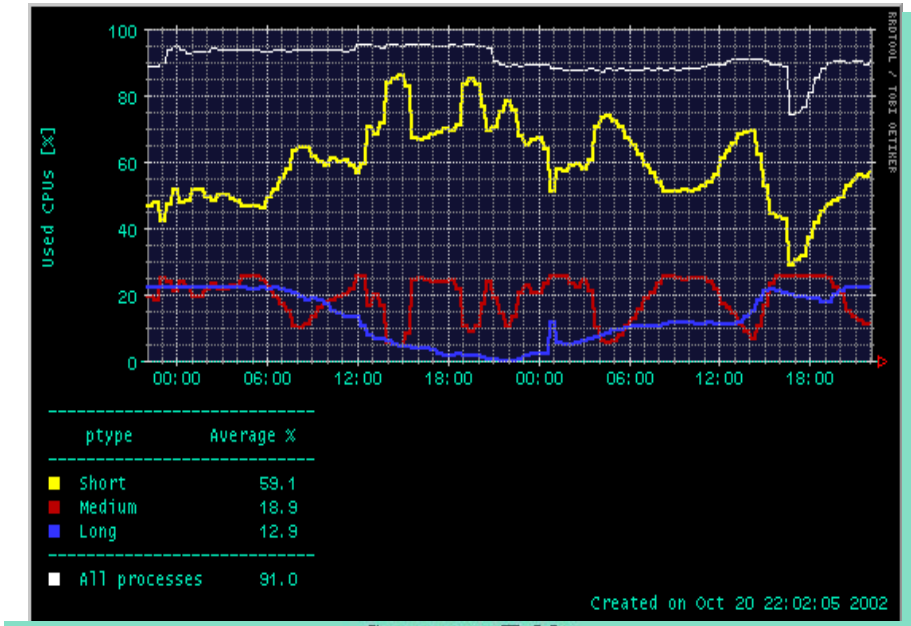
Processes

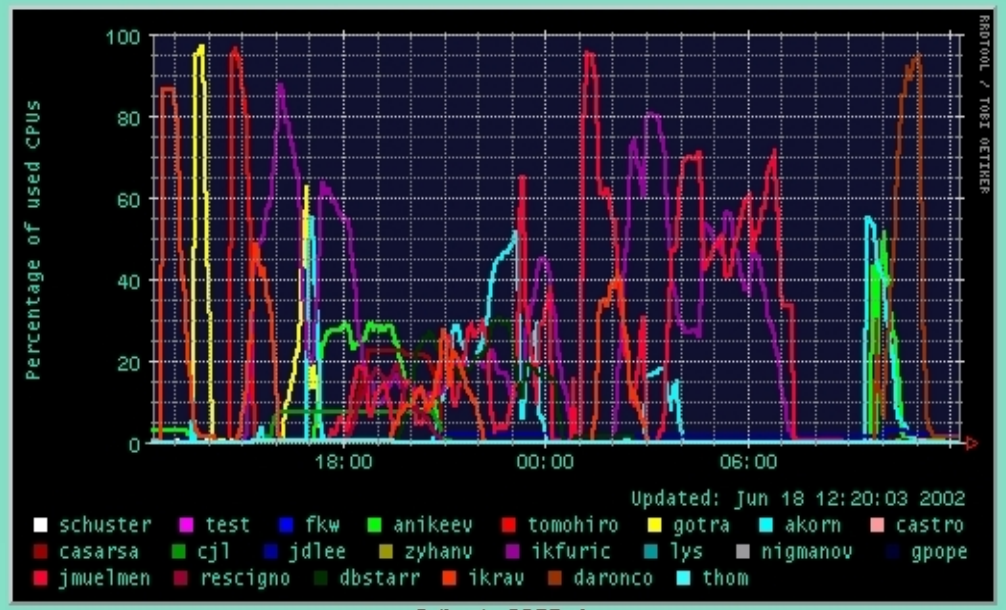
Process #	Node	Status	CPU Time	PID	Command
1	fcdfcaf057	running	0	6931	CafExe cdfcaf@fcdfhead1.fnal.gov:/home/cdfcaf/v1.01/submitter/cafIn/msn_%s.tgz msn@fcdflnx2.fna
			0	6940	simple.sh 600
			0	7221	sleep 120

FCS Group | FBSNG

FBSWWW version 0.1

CAF Utilization





Built using RRDTool

Summary Table

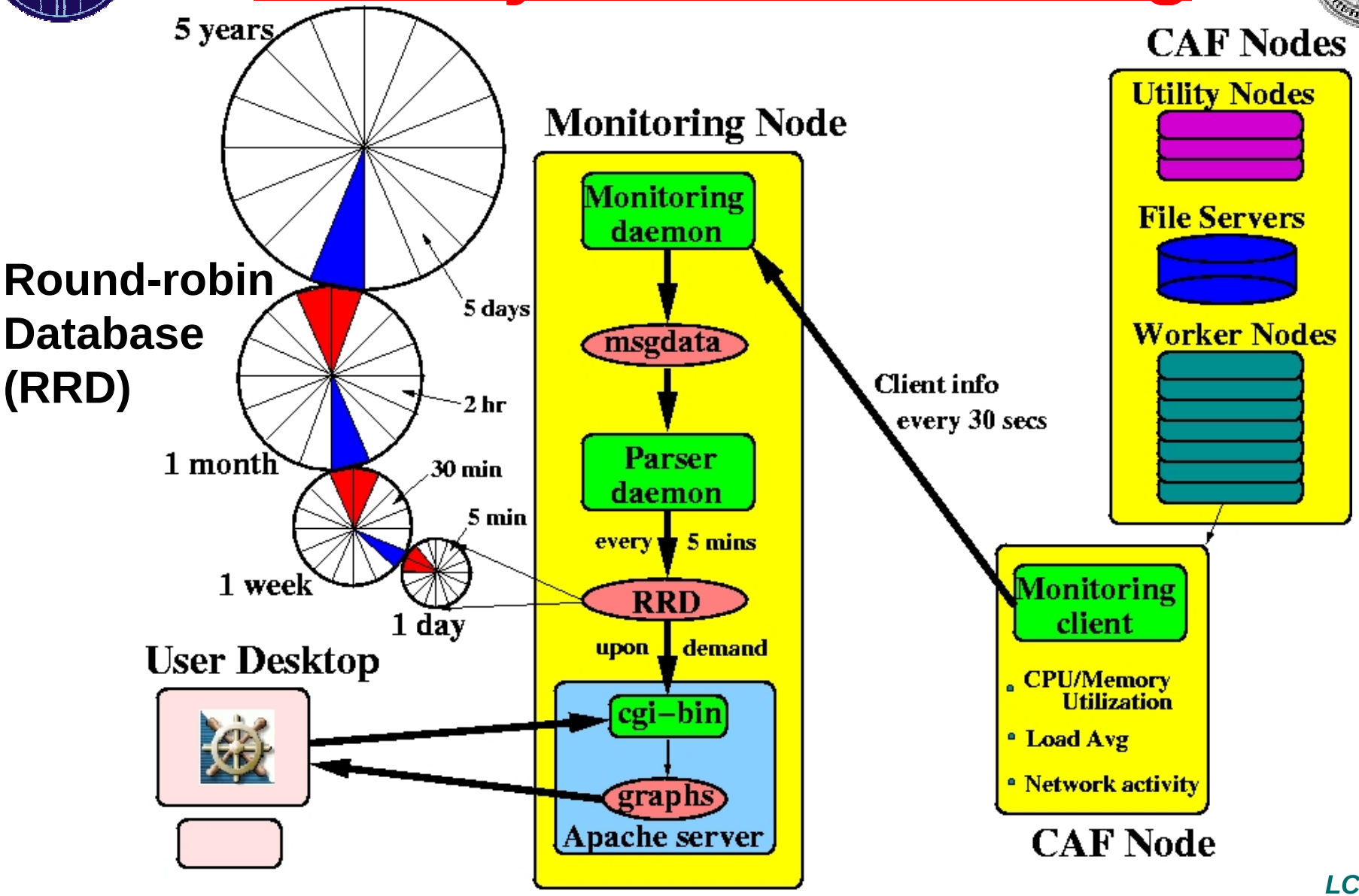
	Short	Medium	Long	All Types
Running sections	98	7	21	126
Pending sections	0	70	66	136
Waiting time [hh:mm] (24h average):				
per job	0:52	0:15	0:00	0:33
per section	4:14	3:29	2:44	3:29
Running time [hh:mm] (24h average)	0:27	4:34	4:08	3:03

Updated: Oct 20 22:00:03 2002

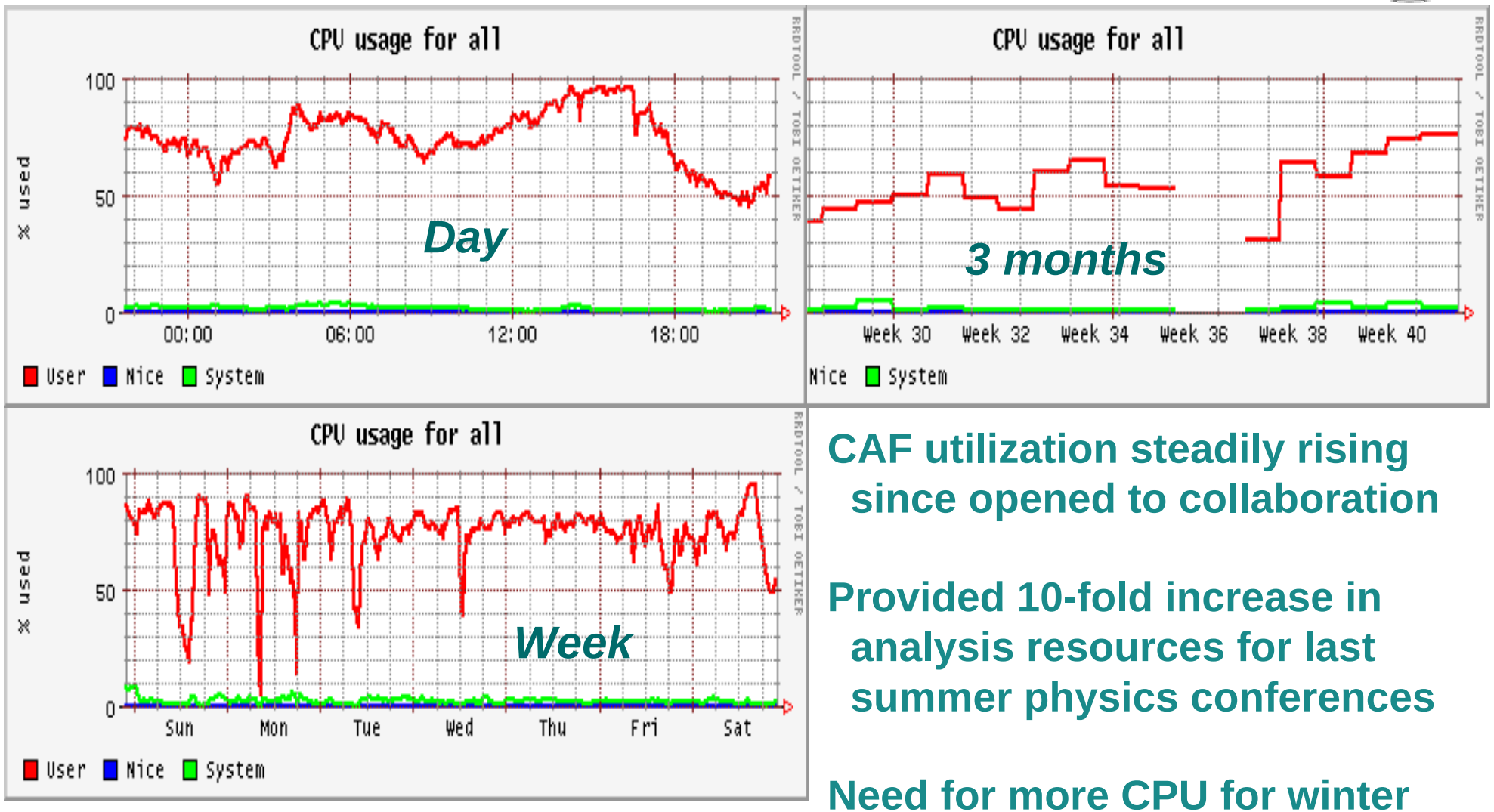
CAF in active use by CDF collaboration

- **300 CAF Users (queues) to date**
- **Several dozen simultaneous users in a typical 24 hr period**

CAF System Monitoring



CPU Utilization

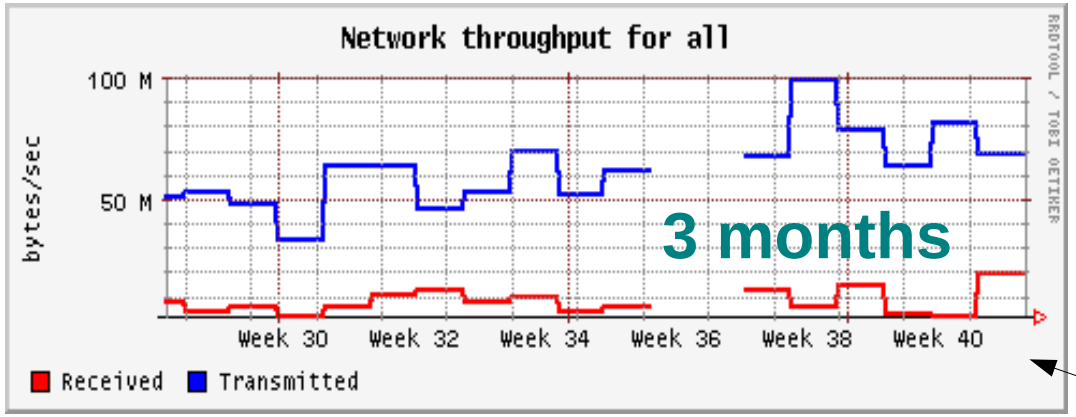


CAF utilization steadily rising since opened to collaboration

Provided 10-fold increase in analysis resources for last summer physics conferences

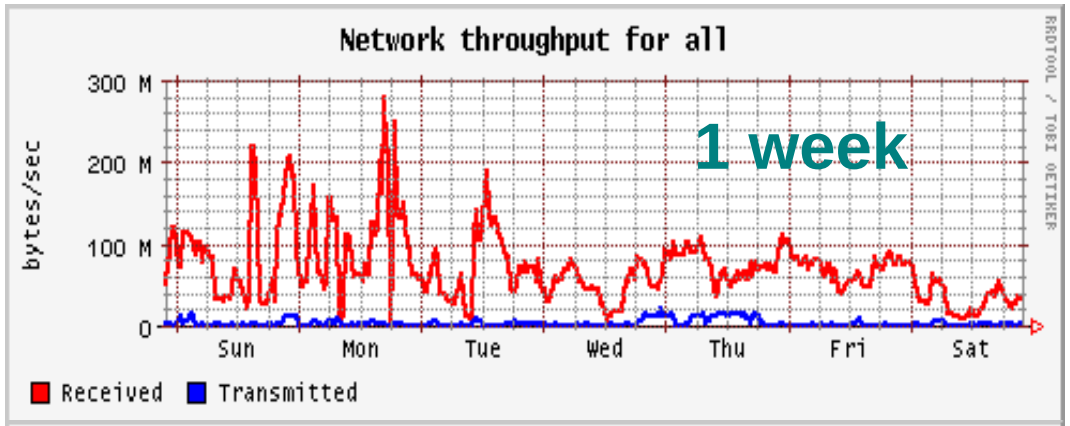
Need for more CPU for winter

Data Processing

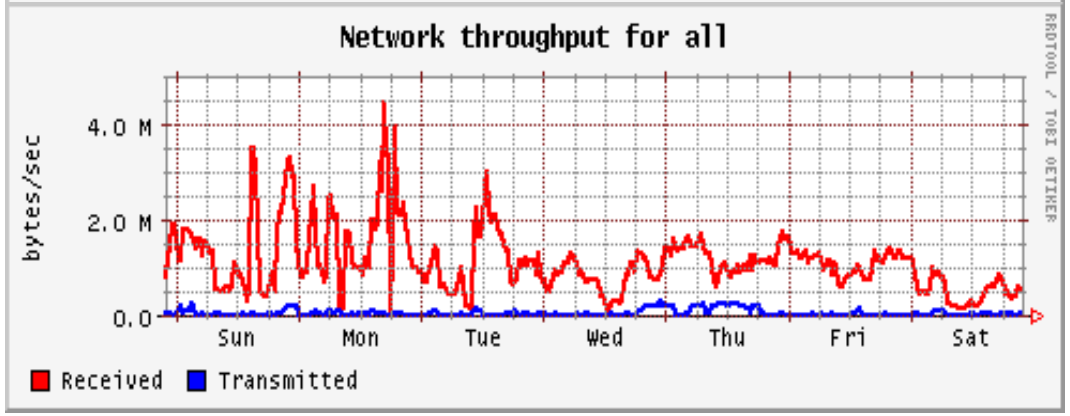


File Server

Aggregate I/O
4-8TB/day



Aggregate I/O



Worker Node

Average I/O
1-2MB/sec @
~80%CPU util.

Work in Progress

Stage2 upgrade: 1THz CPU & 150TB disk

SAM → framework for global data handling/distribution

"DCAF" → remote "replicas" of CAF

Central login pool @ FNAL

JIM & CAF

CAF Summary

Distributed Desk-to-Farm Computing Model

Production system under heavy use:

- **Single farm at FNAL**
  - 4-8TB/day processed by user applications
  - Average CPU utilization of 80%
- **Many users all over the world**
  - 300 total users
  - typical: 30 users per day share 130 CPU's
  - Regularly several 1000 jobs queued
- **Connected to tape via large cache**
- **Currently updating to 1THz & 150TB**

Afterthought

Significant fraction of CAF owned by Universities.

Disk Space:

- **O(10) Universities own ~25TB disk space**

Compute power:

- **O(10) Universities own O(100) CPU's**

All the hardware located in FCC, and maintained as part of CAF.