La CAF
(CDF Analysis Farm)

Massimo Casarsa

Sez. INFN di Trieste

for the CDF Collaboration

IFAE 2005

Catania, 30 Marzo - 2 Aprile 2005

Talk layout

1 CAF motivation and goal

2 CAF overview from user standpoint

3 Structure and features of the CAF software

4 CAF evolution towards the GRID

5 Conclusion

CAF motivation and goal

The CDF Collaboration

Canada
McGill Univ.
Univ. of Toronto

USA
Argonne National Laboratory,IL
Brandeis Univ.,MS
Univ. of Chicago,IL
Davis UC,CA
Duke Univ.,NC
FNAL,IL
Univ. of Florida,FL
Harvard Univ.,MA
Univ. of Illinois,IL
The Johns Hopkins Univ.,MD
LBNL,CA
MIT,MA
Michigan State Univ.,MI
Univ. of Michigan,MI
Univ. of New Mexico,NM
The Ohio State Univ.,OH
Univ. of Pennsylvania,PA
Univ. of Pittsburgh,PA
Purdue Univ.,IN
Univ. of Rochester,NY
Rockefeller Univ.,NY
Rutgers Univ.,NJ
Texas A&M Univ.,TX
Texas Tech Univ.,TX
Tufts Univ.,MA
UCLA,CA
Univ. of Wisconsin,WI
Yale Univ.,CT

China
Academia Sinica,
Taiwan

Korea
KHCL

Russia
JINR, Dubna
ITEP, Moscow

Germany
Univ. Karlsruhe

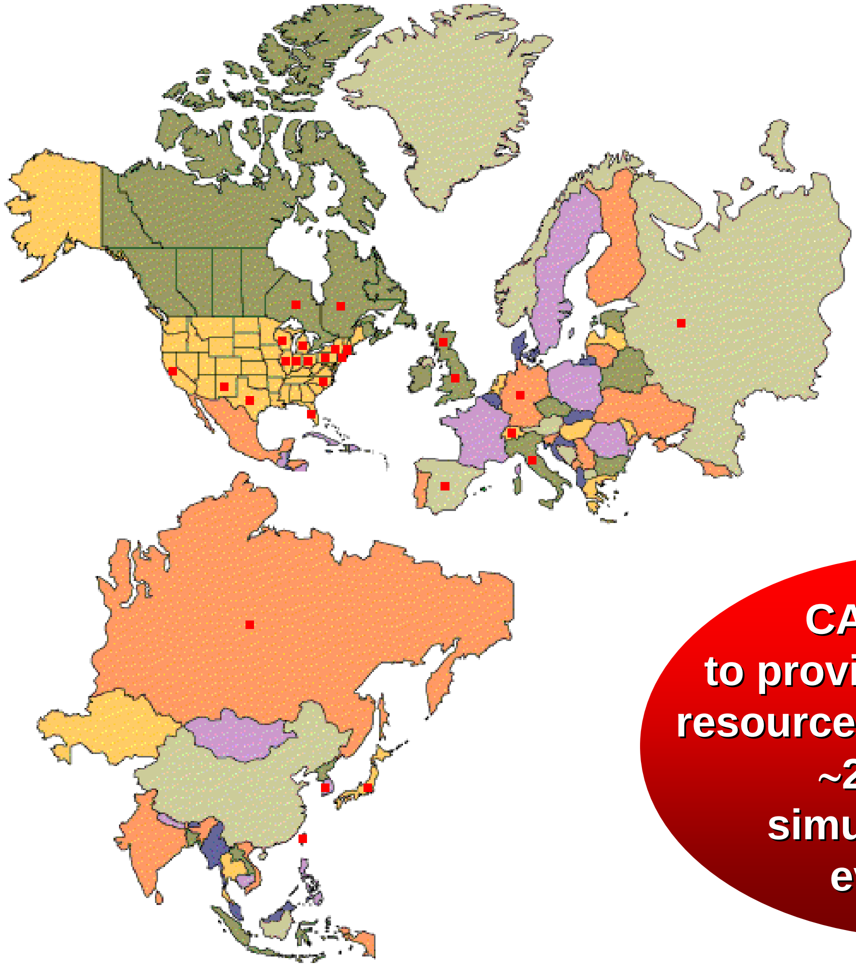
Switzerland
Univ. of Geneva

UK
Glasgow Univ.
Univ. of Liverpool
Univ. of Oxford
Univ. College London

Italy
Univ. of Bologna,INFN
Frascati, INFN
Univ. di Padova,INFN
Pisa, INFN
Univ. di Roma I,INFN
INFN-Trieste
Univ. di Udine

Spain
Univ. of Cantabria

Japan
Hiroshima Univ.
KEK
Osaka City Univ.
Univ. of Tsukuba
Waseda Univ., Tokyo

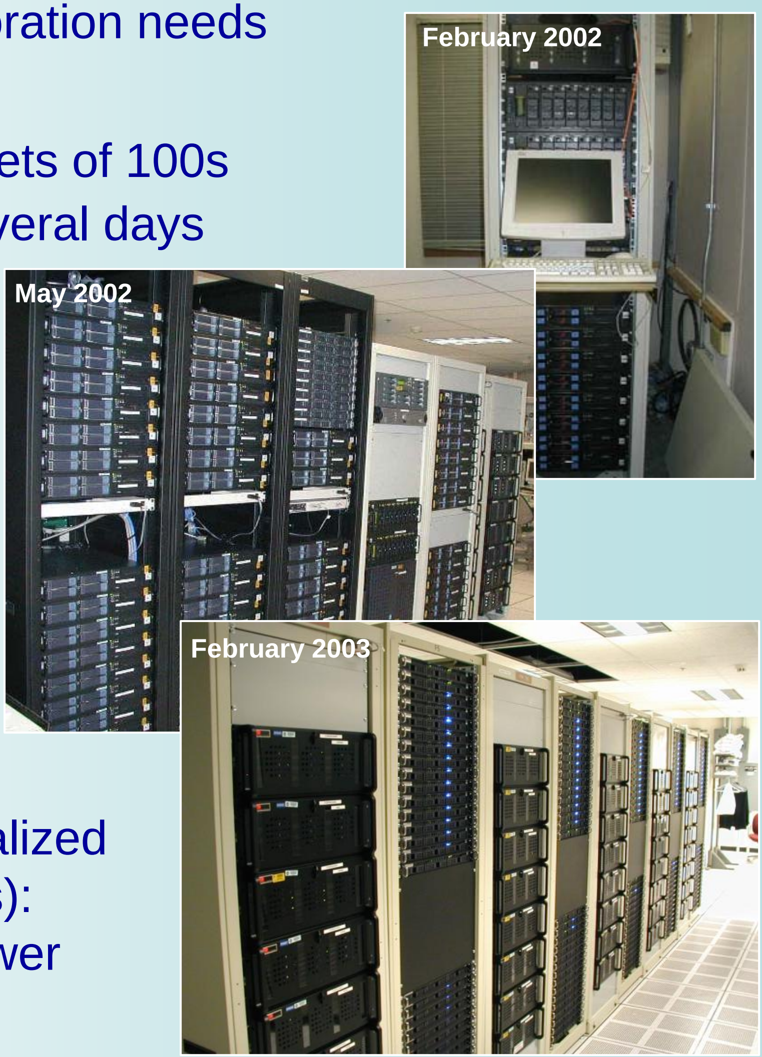


- ~600 Physicists
- 56 Institutions
- 11 Countries

CAF goal is to provide computing resources for analysis to ~200 users simultaneously every day

CAF milestones

- Assembled in 2002 to meet the Collaboration needs for computing resources:
  - ✓ data analysis: CDF produces datasets of 100s of TBs whose processing takes several days (0.1-0.5 s/event).
  - ✓ MC production: detector simulation is heavy CPU consuming (~1 s/event).
- Born as a farm localized at FNAL with the FBSNG batch manager, then migrated to Condor.
- CAF model exported and farm decentralized to many sites around the world (DCAFs): at present ~50% of CDF computing power outside Fermilab.



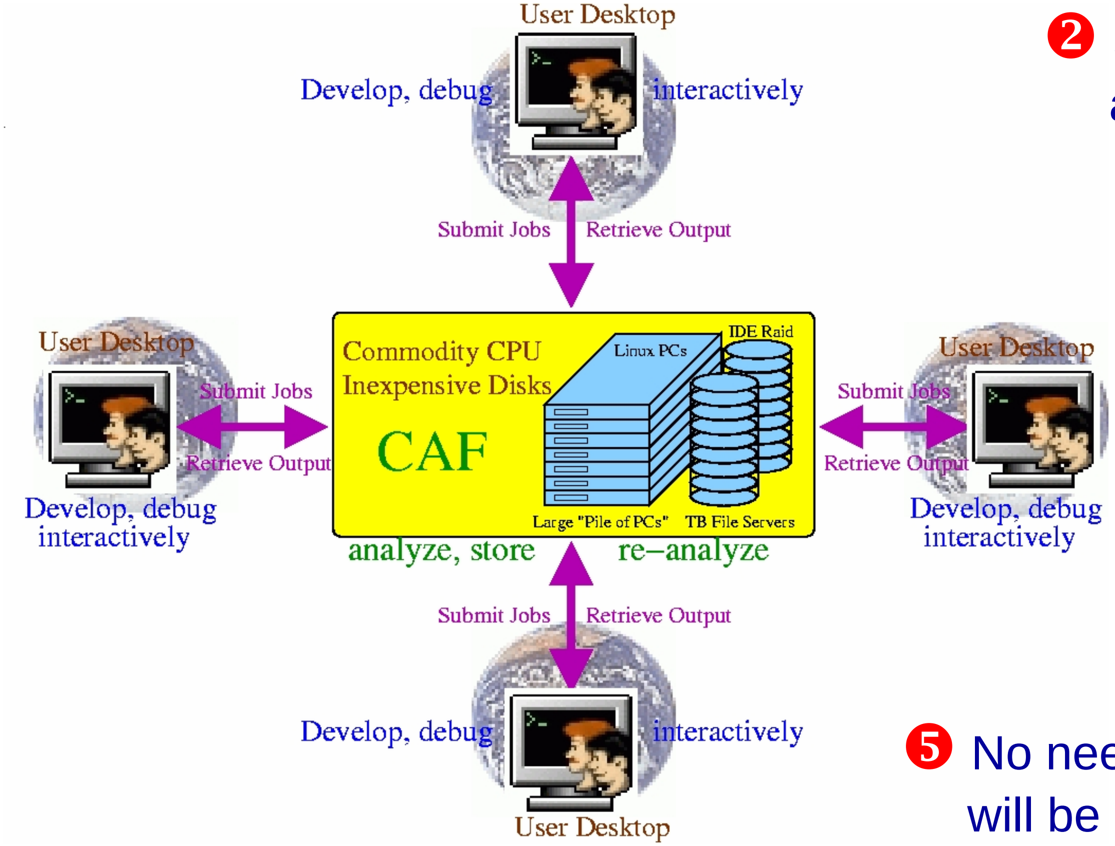
DCAFs resources around the world

Cluster name	Location	CPU [GHz]	Disk space [TB]
Original FNAL CAF	FNAL	1200	300
FNAL CondorCAF		2000	
CNAFCAF	**Bologna (Italy)**	**300**	**22**
KORCAF	KNU (South Korea)	120	0.6
ASCAF	Acc. Sinica (Taiwan)	134	3.0
SDSCCAF	San Diego (USA)	280	4.0
HEXCAF	Rutgers (USA)	100	4.0
TORCAF2	Toronto (Canada)	576	10
JPCAF	Tsukuba (Japan)	152	5.0
CANCAF	Cantabria (Spain)	52	1.5
MITCAF	Boston (USA)	110	2.0
TOTAL		**5024**	**352.1**

- Data reside at FNAL, DCAFs mostly used for MC production, recently dataset replicas allow to run also analysis jobs.

CAF philosophy

❶ Develop and debug analysis code on personal desktop or laptop.



❷ Submit jobs to CAF from anywhere in the world.

❸ Have the output delivered wherever you like.

❹ Interact with running jobs as they were local.

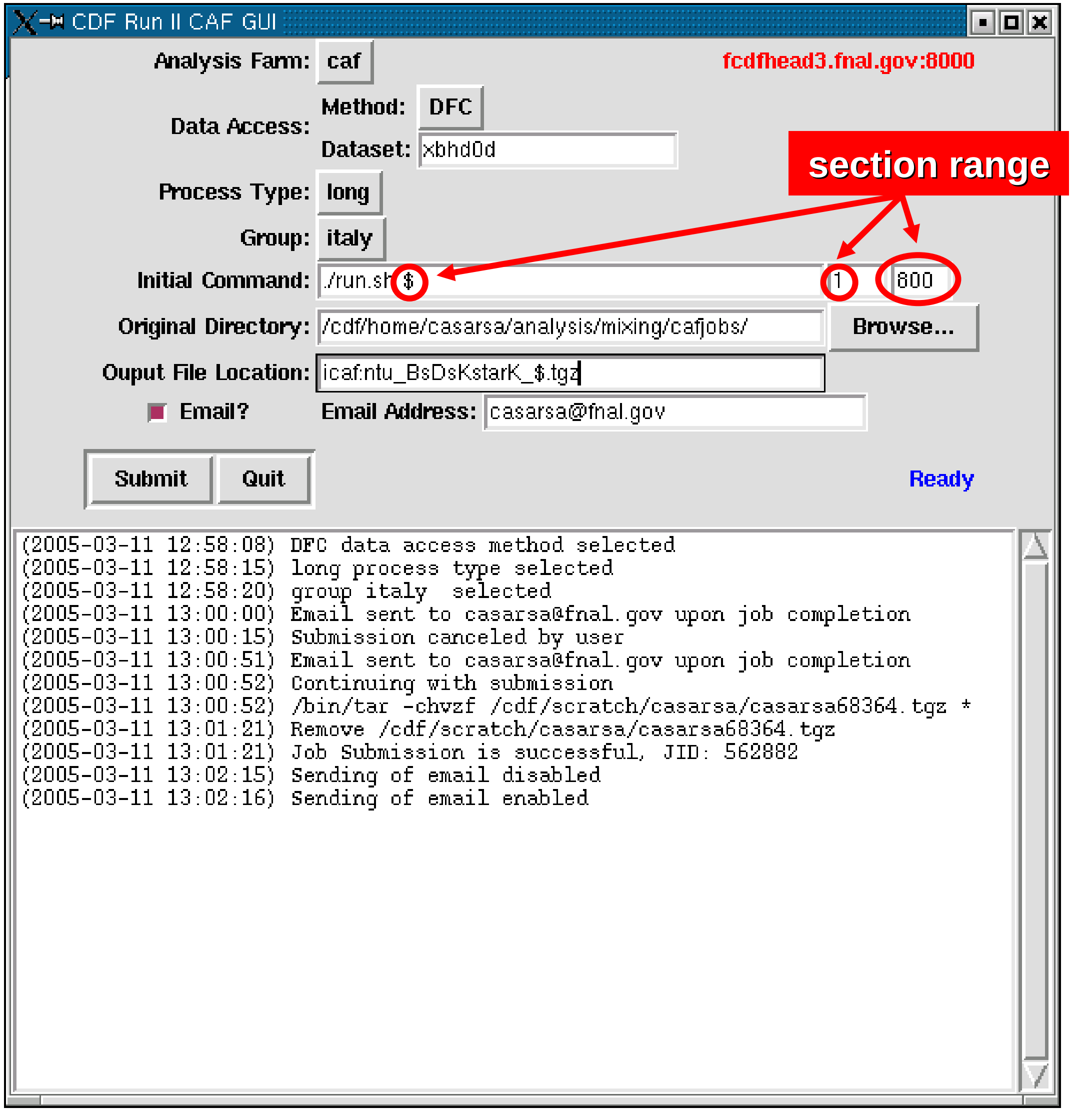
❺ No need to stay connected, will be notified at job completion.

CAF user interface (CafGui)

Fill in appropriate fields:

- farm
- data access method and dataset
- process type
- group
- command
- original directory
- output location
- e-mail address

submit



section range

Analysis Farm: caf

fcdfhead3.fnal.gov:8000

Data Access: Method: DFC

Dataset: xbhd0d

Process Type: long

Group: italy

Initial Command: ./run.sh $ 1 800

Original Directory: /cdf/home/casarsa/analysis/mixing/cafjobs/

Ouput File Location: icaf:ntu_BsDsKstarK_$.tgz

■ Email? Email Address: casarsa@fnal.gov

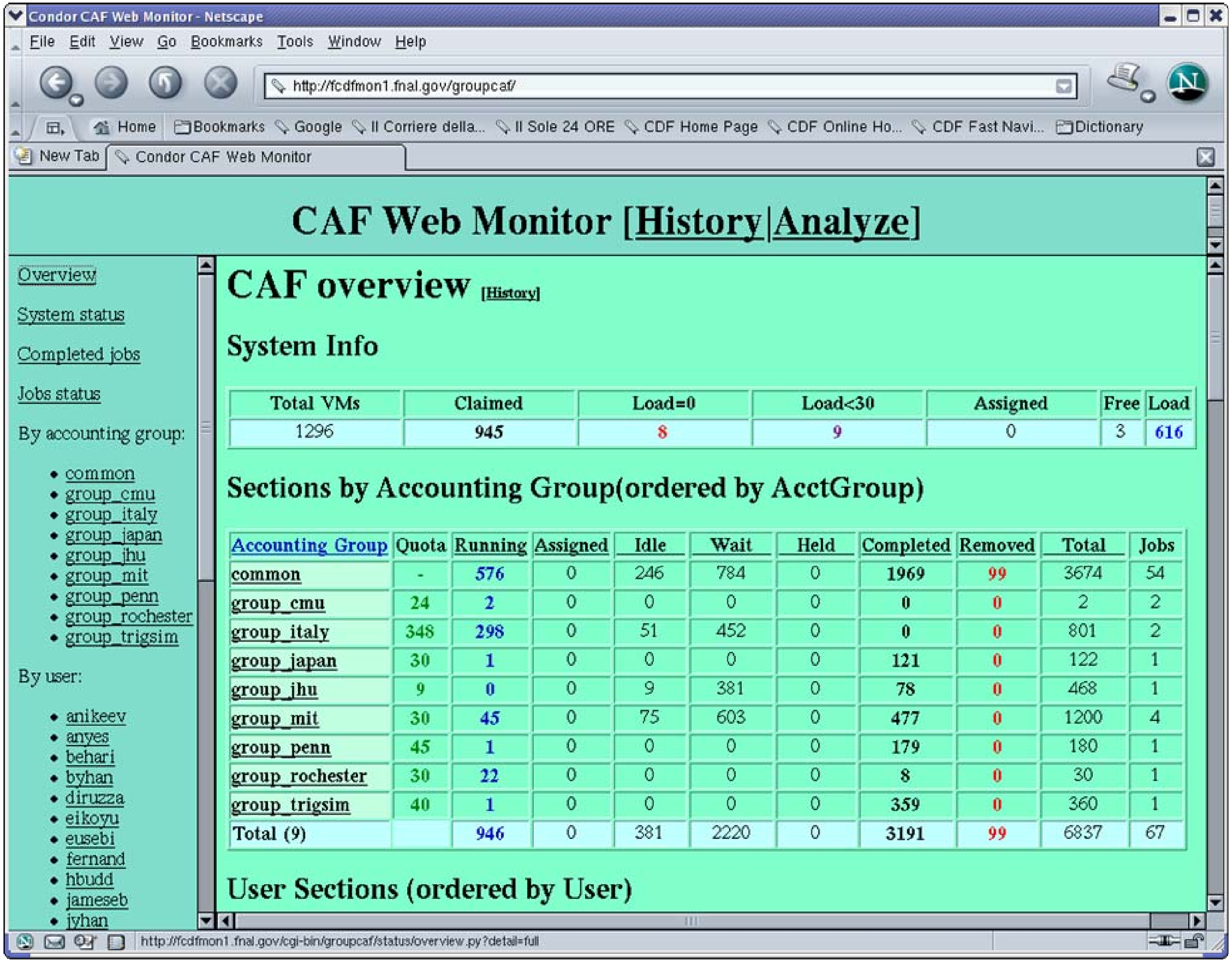
```
(2005-03-11 12:58:08) DFC data access method selected
(2005-03-11 12:58:15) long process type selected
(2005-03-11 12:58:20) group italy  selected
(2005-03-11 13:00:00) Email sent to casarsa@fnal.gov upon job completion
(2005-03-11 13:00:15) Submission canceled by user
(2005-03-11 13:00:51) Email sent to casarsa@fnal.gov upon job completion
(2005-03-11 13:00:52) Continuing with submission
(2005-03-11 13:00:52) /bin/tar -chvzf /cdf/scratch/casarsa/casarsa68364.tgz *
(2005-03-11 13:01:21) Remove /cdf/scratch/casarsa/casarsa68364.tgz
(2005-03-11 13:01:21) Job Submission is successful, JID: 562882
(2005-03-11 13:02:15) Sending of email disabled
(2005-03-11 13:02:16) Sending of email enabled
```

User interactive tools

CafMon allows users to interact with remote jobs as they were running locally:

- job management tasks:
  - ✓ kill ⇒ kill job/section;
  - ✓ hold/release ⇒ hold/release jobs;
  - ✓ chprio/chgroup ⇒ change the priority/group of a job.
- jobs and remote system interactive monitoring:
  - ✓ jobs ⇒ list submitted jobs and sections;
  - ✓ log ⇒ print out log file of a specific section;
  - ✓ top ⇒ show status of the node running a specific section;
  - ✓ ps ⇒ show the processes of a specific section;
  - ✓ dir ⇒ show the working dir of a specific section;
  - ✓ tail/head/cat ⇒ inspect any text file in section working directory.
- debug:
  - ✓ debug ⇒ attach a debugging session to a remote running process.

User web monitoring


Condor CAF Web Monitor - Netscape
File Edit View Go Bookmarks Tools Window Help
http://fcdfmon1.fnal.gov/groupcaf/
Home
Bookmarks
Google
Il Corriere della...
Il Sole 24 ORE
CDF Home Page
CDF Online Ho...
CDF Fast Navi...
Dictionary
New Tab
Condor CAF Web Monitor
http://fcdfmon1.fnal.gov/cgi-bin/groupcaf/status/overview.py?detail=full

CAF Web Monitor [History|Analyze]

- Overview
- System status
- Completed jobs
- Jobs status

By accounting group:

- common
- group_cmu
- group_italy
- group_japan
- group_jhu
- group_mit
- group_penn
- group_rochester
- group_trigsim

By user:

- anikeev
- anyes
- behari
- byhan
- diruzza
- eikoyu
- eusebi
- fernand
- hbudd
- jameseb
- jyhan

CAF overview [History]

System Info

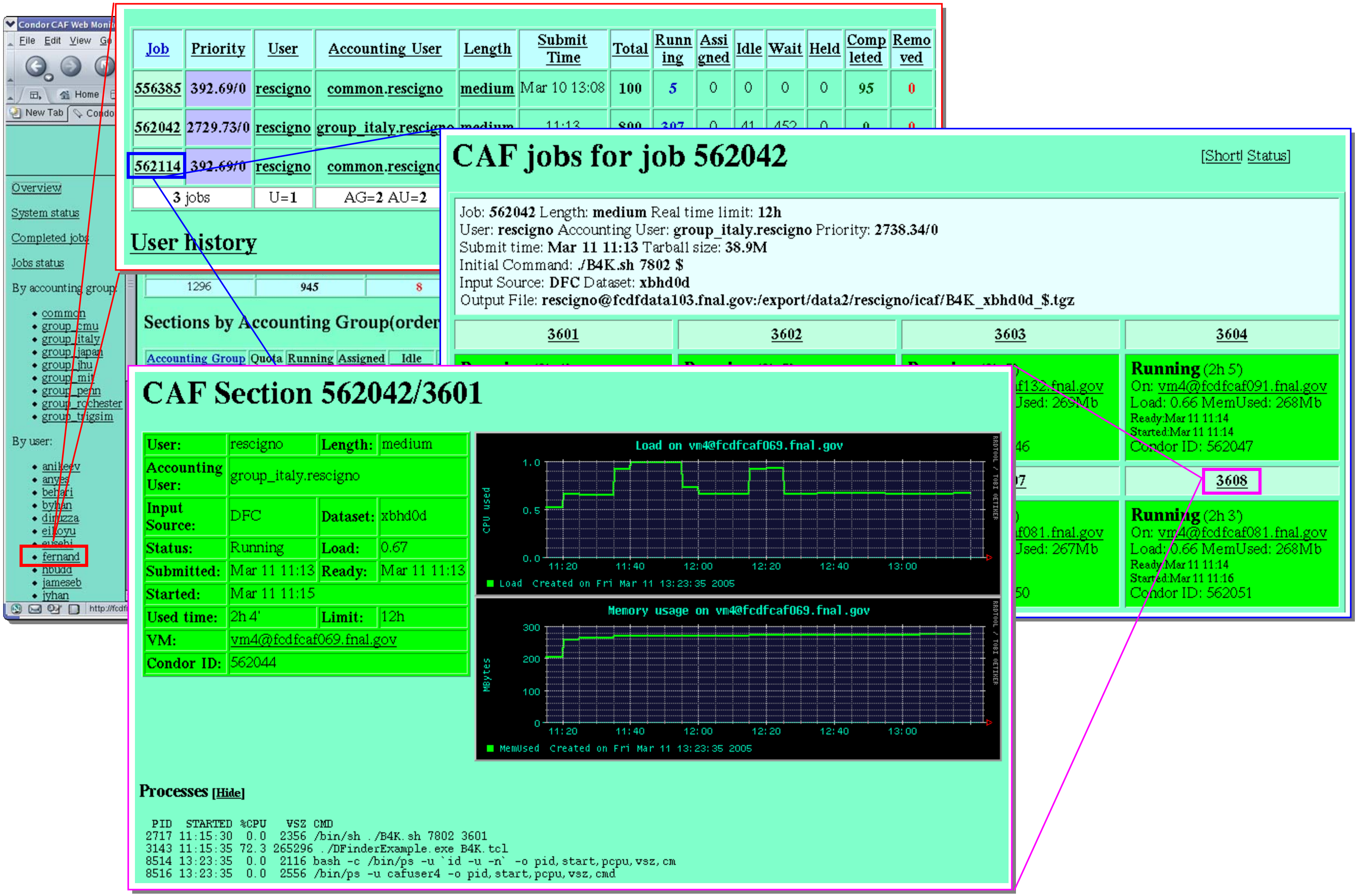
Total VMs	Claimed	Load=0	Load<30	Assigned	Free	Load
1296	945	8	9	0	3	616

Sections by Accounting Group(ordered by AcctGroup)

Accounting Group	Quota	Running	Assigned	Idle	Wait	Held	Completed	Removed	Total	Jobs
common	-	576	0	246	784	0	1969	99	3674	54
group_cmu	24	2	0	0	0	0	0	0	2	2
group_italy	348	298	0	51	452	0	0	0	801	2
group_japan	30	1	0	0	0	0	121	0	122	1
group_jhu	9	0	0	9	381	0	78	0	468	1
group_mit	30	45	0	75	603	0	477	0	1200	4
group_penn	45	1	0	0	0	0	179	0	180	1
group_rochester	30	22	0	0	0	0	8	0	30	1
group_trigsim	40	1	0	0	0	0	359	0	360	1
Total (9)		946	0	381	2220	0	3191	99	6837	67

User Sections (ordered by User)

User web monitoring: job/section status



Job	Priority	User	Accounting User	Length	Submit Time	Total	Running	Assigned	Idle	Wait	Held	Completed	Removed
556385	392.69/0	rescigno	common.rescigno	medium	Mar 10 13:08	100	5	0	0	0	0	95	0
562042	2729.73/0	rescigno	group_italy.rescigno	medium	11:13	800	307	0	41	452	0	0	0
562114	392.69/0	rescigno	common.rescigno										
3 jobs		U=1	AG=2 AU=2										

User history

1296	945	8

Sections by Accounting Group(order

Accounting Group	Quota	Running	Assigned	Idle

CAF jobs for job 562042

[Short] Status]

Job: **562042** Length: **medium** Real time limit: **12h**
User: **rescigno** Accounting User: **group_italy.rescigno** Priority: **2738.34/0**
Submit time: **Mar 11 11:13** Tarball size: **38.9M**
Initial Command: **./B4K.sh 7802 $**
Input Source: **DFC** Dataset: **xbhd0d**
Output File: **rescigno@fcdfdata103.fnal.gov:/export/data2/rescigno/icaf/B4K_xbhd0d_$.tgz**

3601	3602	3603	3604
		f132.fnal.gov Used: 269Mb 46	**Running** (2h 5') On: vm4@fcdfcaf091.fnal.gov Load: 0.66 MemUsed: 268Mb Ready:Mar 11 11:14 Started:Mar 11 11:14 Condor ID: 562047
		07	3608
		f081.fnal.gov Used: 267Mb 50	**Running** (2h 3') On: vm4@fcdfcaf081.fnal.gov Load: 0.66 MemUsed: 268Mb Ready:Mar 11 11:14 Started:Mar 11 11:16 Condor ID: 562051

CAF Section 562042/3601

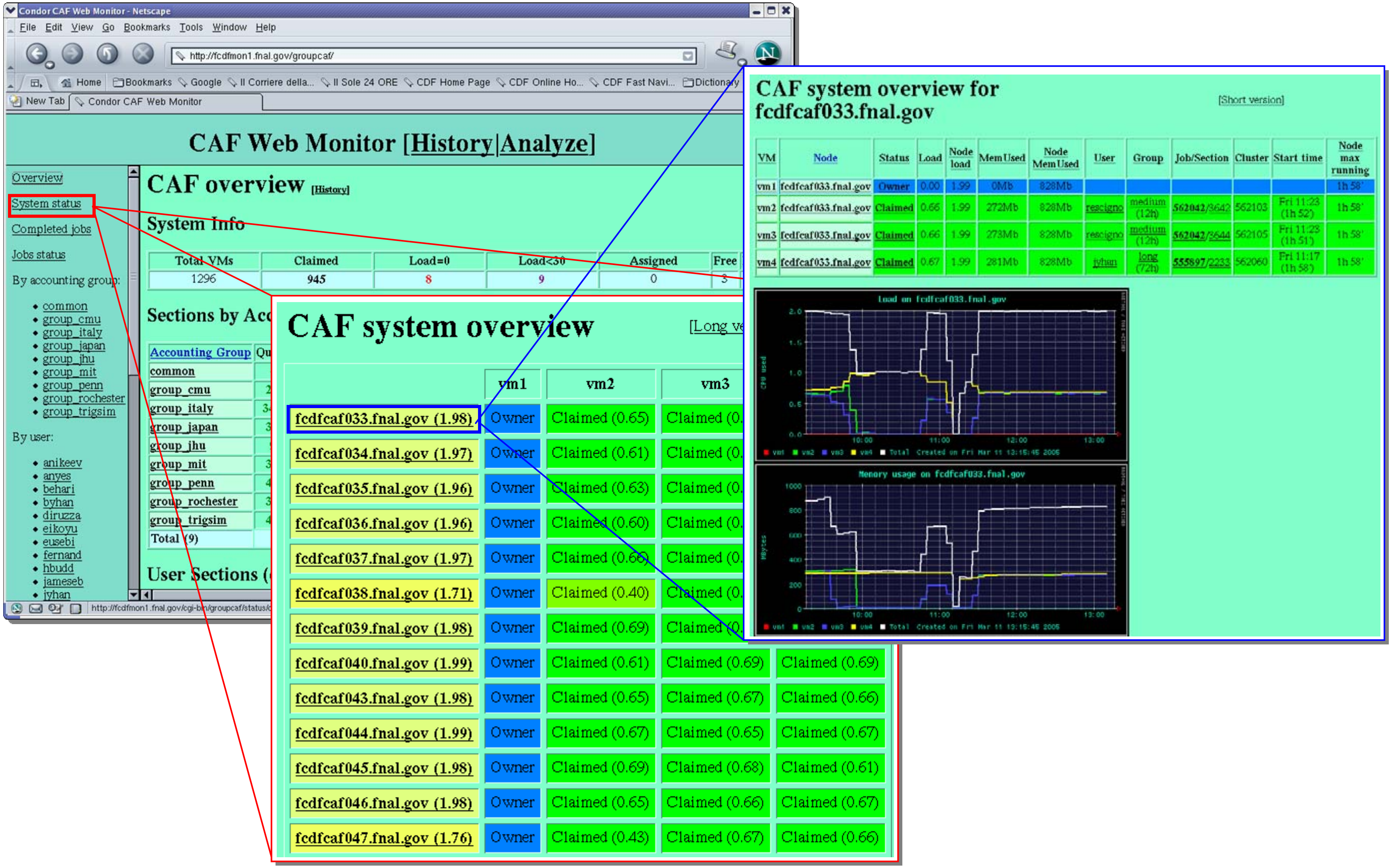
User:	rescigno	Length:	medium
Accounting User:	group_italy.rescigno		
Input Source:	DFC	Dataset:	xbhd0d
Status:	Running	Load:	0.67
Submitted:	Mar 11 11:13	Ready:	Mar 11 11:13
Started:	Mar 11 11:15		
Used time:	2h 4'	Limit:	12h
VM:	vm4@fcdfcaf069.fnal.gov		
Condor ID:	562044		

Load on vm4@fcdfcaf069.fnal.gov
CPU used
1.0
0.5
0.0
11:20
11:40
12:00
12:20
12:40
13:00
Load Created on Fri Mar 11 13:23:35 2005
RRDTOOL / TOBI OETIKER
Memory usage on vm4@fcdfcaf069.fnal.gov
MBytes
300
200
100
0
11:20
11:40
12:00
12:20
12:40
13:00
MemUsed Created on Fri Mar 11 13:23:35 2005
RRDTOOL / TOBI OETIKER

Processes [Hide]

```
  PID  STARTED %CPU   VSZ CMD
 2717 11:15:30  0.0  2356 /bin/sh ./B4K.sh 7802 3601
 3143 11:15:35 72.3 265296 ./DFinderExample.exe B4K.tcl
 8514 13:23:35  0.0  2116 bash -c /bin/ps -u `id -u -n` -o pid,start,pcpu,vsz,cm
 8516 13:23:35  0.0  2556 /bin/ps -u cafuser4 -o pid,start,pcpu,vsz,cmd
```

User web monitoring: worker node status


Condor CAF Web Monitor - Netscape
File Edit View Go Bookmarks Tools Window Help
http://fcdfmon1.fnal.gov/groupcaf/
Home
Bookmarks
Google
Il Corriere della...
Il Sole 24 ORE
CDF Home Page
CDF Online Ho...
CDF Fast Navi...
Dictionary
New Tab
Condor CAF Web Monitor
http://fcdfmon1.fnal.gov/cgi-bin/groupcaf/status/c
Load on fcdfcaf033.fnal.gov
CPU used
2.0
1.5
1.0
0.5
0.0
10:00
11:00
12:00
13:00
vm1 vm2 vm3 vm4 Total Created on Fri Mar 11 13:15:45 2005
Memory usage on fcdfcaf033.fnal.gov
MBytes
1000
800
600
400
200
0
10:00
11:00
12:00
13:00
vm1 vm2 vm3 vm4 Total Created on Fri Mar 11 13:15:45 2005

CAF Web Monitor [History|Analyze]

Overview

System status

Completed jobs

Jobs status

By accounting group:

- common
- group_cmu
- group_italy
- group_japan
- group_jhu
- group_mit
- group_penn
- group_rochester
- group_trigsim

By user:

- anikeev
- anyes
- behari
- byhan
- diruzza
- eikoyu
- eusebi
- fernand
- hbudd
- jameseb
- jyhan

CAF overview [History]

System Info

Total VMs	Claimed	Load=0	Load<30	Assigned	Free
1296	945	8	9	0	3

Sections by Acc

Accounting Group	Qu
common	
group_cmu	2
group_italy	3
group_japan	3
group_jhu	
group_mit	3
group_penn	4
group_rochester	3
group_trigsim	4
Total (9)	

User Sections (

CAF system overview

[Long ve

	vm1	vm2	vm3	
fcdfcaf033.fnal.gov (1.98)	Owner	Claimed (0.65)	Claimed (0.	
fcdfcaf034.fnal.gov (1.97)	Owner	Claimed (0.61)	Claimed (0.	
fcdfcaf035.fnal.gov (1.96)	Owner	Claimed (0.63)	Claimed (0.	
fcdfcaf036.fnal.gov (1.96)	Owner	Claimed (0.60)	Claimed (0.	
fcdfcaf037.fnal.gov (1.97)	Owner	Claimed (0.66)	Claimed (0.	
fcdfcaf038.fnal.gov (1.71)	Owner	Claimed (0.40)	Claimed (0.	
fcdfcaf039.fnal.gov (1.98)	Owner	Claimed (0.69)	Claimed (0.	
fcdfcaf040.fnal.gov (1.99)	Owner	Claimed (0.61)	Claimed (0.69)	Claimed (0.69)
fcdfcaf043.fnal.gov (1.98)	Owner	Claimed (0.65)	Claimed (0.67)	Claimed (0.66)
fcdfcaf044.fnal.gov (1.99)	Owner	Claimed (0.67)	Claimed (0.65)	Claimed (0.67)
fcdfcaf045.fnal.gov (1.98)	Owner	Claimed (0.69)	Claimed (0.68)	Claimed (0.61)
fcdfcaf046.fnal.gov (1.98)	Owner	Claimed (0.65)	Claimed (0.66)	Claimed (0.67)
fcdfcaf047.fnal.gov (1.76)	Owner	Claimed (0.43)	Claimed (0.67)	Claimed (0.66)

CAF system overview for fcdfcaf033.fnal.gov

[Short version]

VM	Node	Status	Load	Node load	MemUsed	Node MemUsed	User	Group	Job/Section	Cluster	Start time	Node max running
vm1	fcdfcaf033.fnal.gov	Owner	0.00	1.99	0Mb	828Mb						1h 58'
vm2	fcdfcaf033.fnal.gov	Claimed	0.66	1.99	272Mb	828Mb	rescigno	medium (12h)	562042/3642	562103	Fri 11:23 (1h 52')	1h 58'
vm3	fcdfcaf033.fnal.gov	Claimed	0.66	1.99	273Mb	828Mb	rescigno	medium (12h)	562042/3644	562105	Fri 11:23 (1h 51')	1h 58'
vm4	fcdfcaf033.fnal.gov	Claimed	0.67	1.99	281Mb	828Mb	jyhan	long (72h)	555897/2233	562060	Fri 11:17 (1h 58')	1h 58'

Farm implementation

- Hardware:

need lots of CPUs	⇒	commodity CPUs	⇒	dual Intel/AMD;
need lots of disk	⇒	cheap disk	⇒	IDE RAID 50 arrays.

- Nodes software:

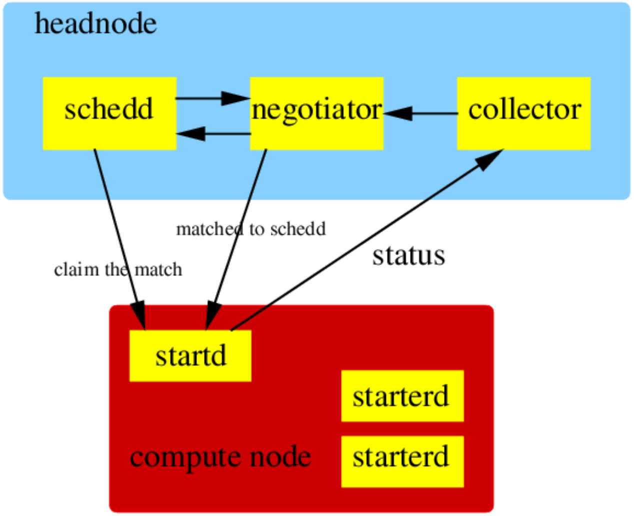
  ✓ operating system: Linux;
  ✓ have access to CDF software.

- Batch manager: Condor

  ✓ six virtual-machines (VM) per node;
  ✓ higher priority for groups on institutions' proprietary hardware;
  ✓ process type for job length:

test:	2 h (max 4 sections),
short:	6 h,
medium:	12 h,
long:	72 h.

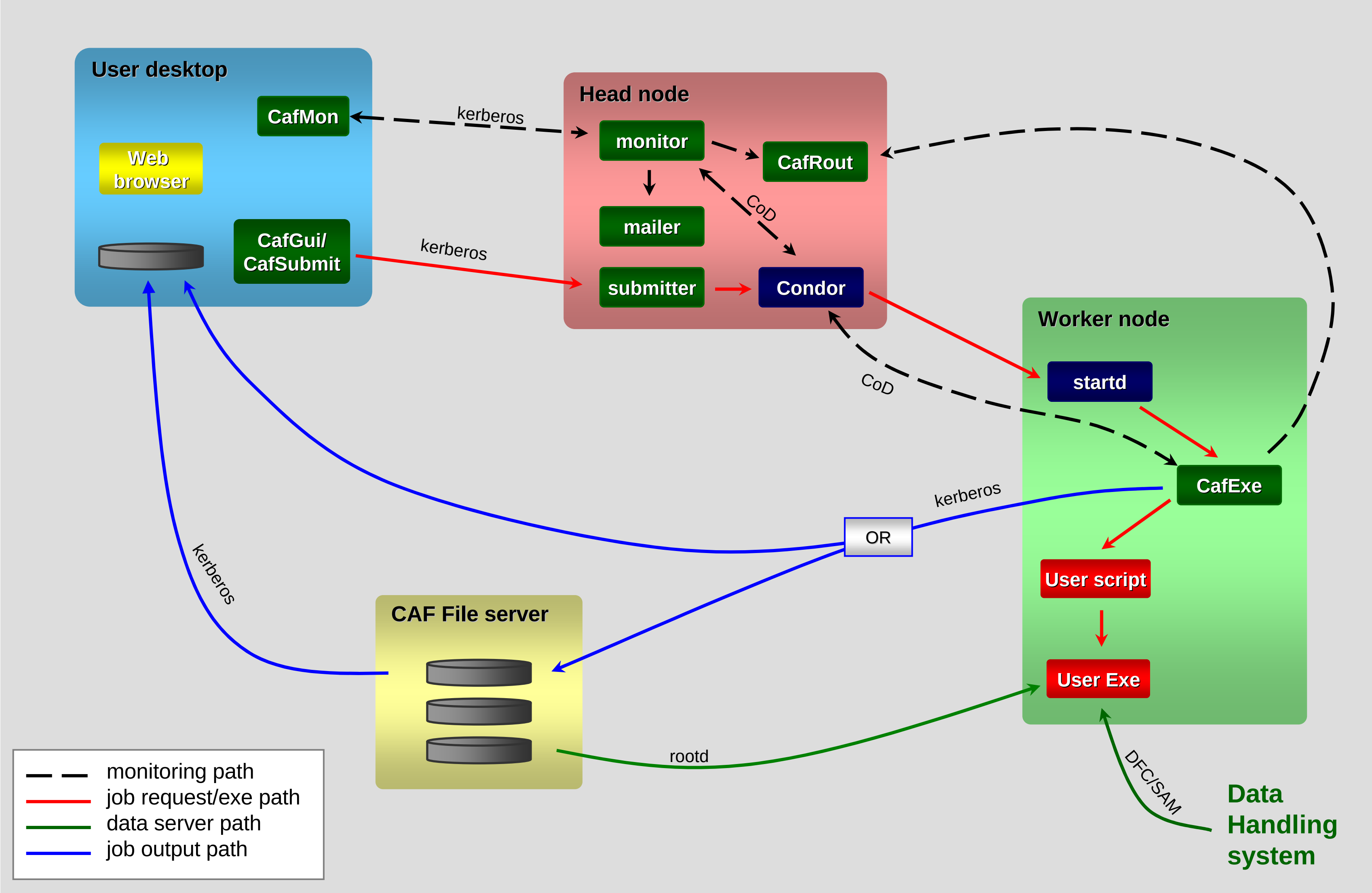
Condor configuration:



CAF software overview

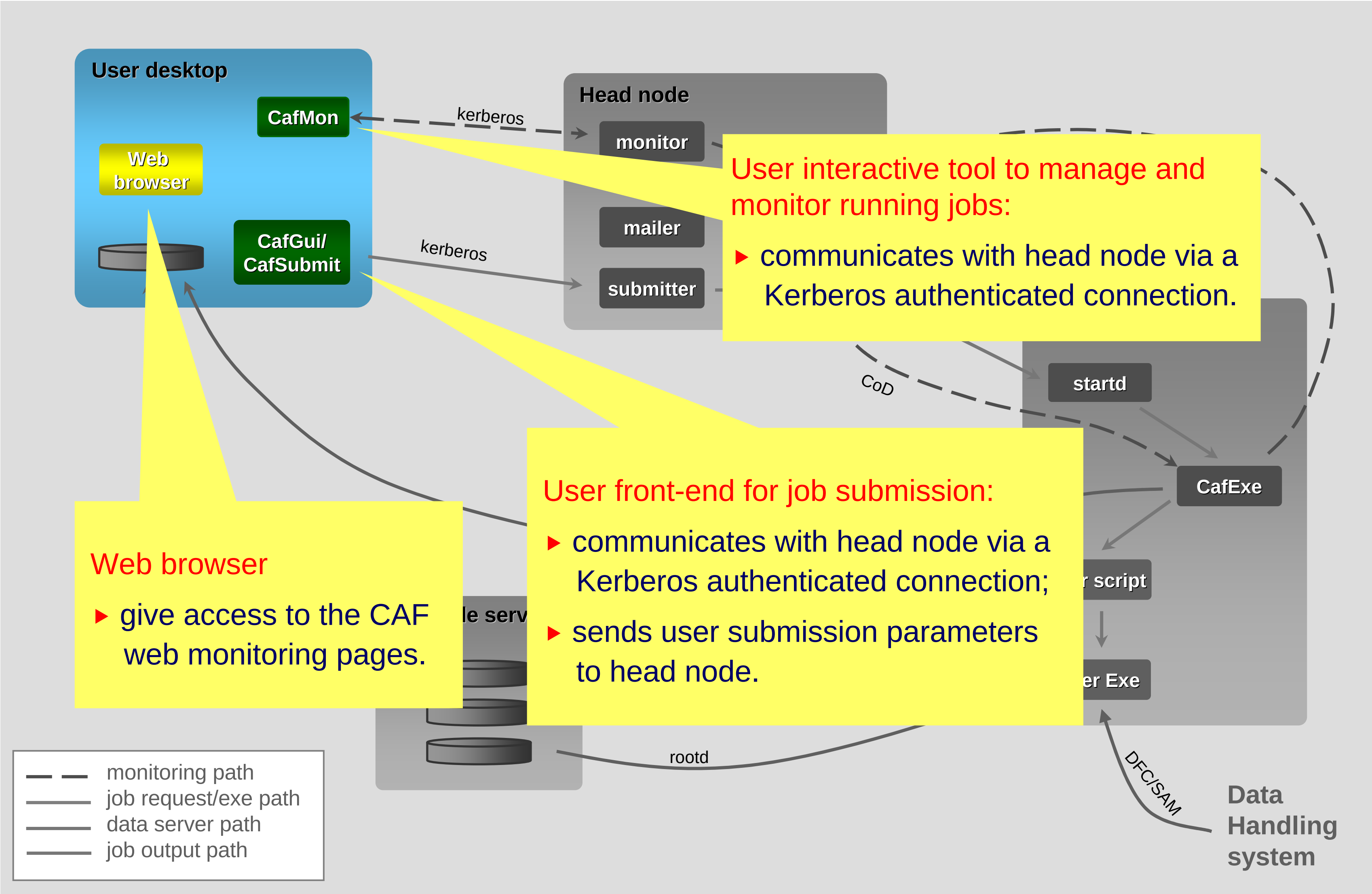
- Design goal:
  - ✓ Give user access to CAF resources from anywhere in the world.
- Design constraints/desirables:
  - ✓ Fermilab computing security policy → Kerberos;
  - ✓ administrative ease → no user accounts
    → non-interactive batch, jobs run as generic users (one for each VM);
  - ✓ user identity → unique privileges for batch jobs, disk space;
  - ✓ large scale parallelization with single submission (Condor dagman).
- Result:
  - ✓ very user-friendly software;
  - ✓ user provides a shell script and an executable + all needed files;
  - ✓ everything is tarred up and sent to CAF;
  - ✓ user is notified when his job is completed.

CAF software



— — monitoring path
🟥 job request/exe path
🟩 data server path
🟦 job output path

CAF software: user desktop



User interactive tool to manage and monitor running jobs:

- communicates with head node via a Kerberos authenticated connection.

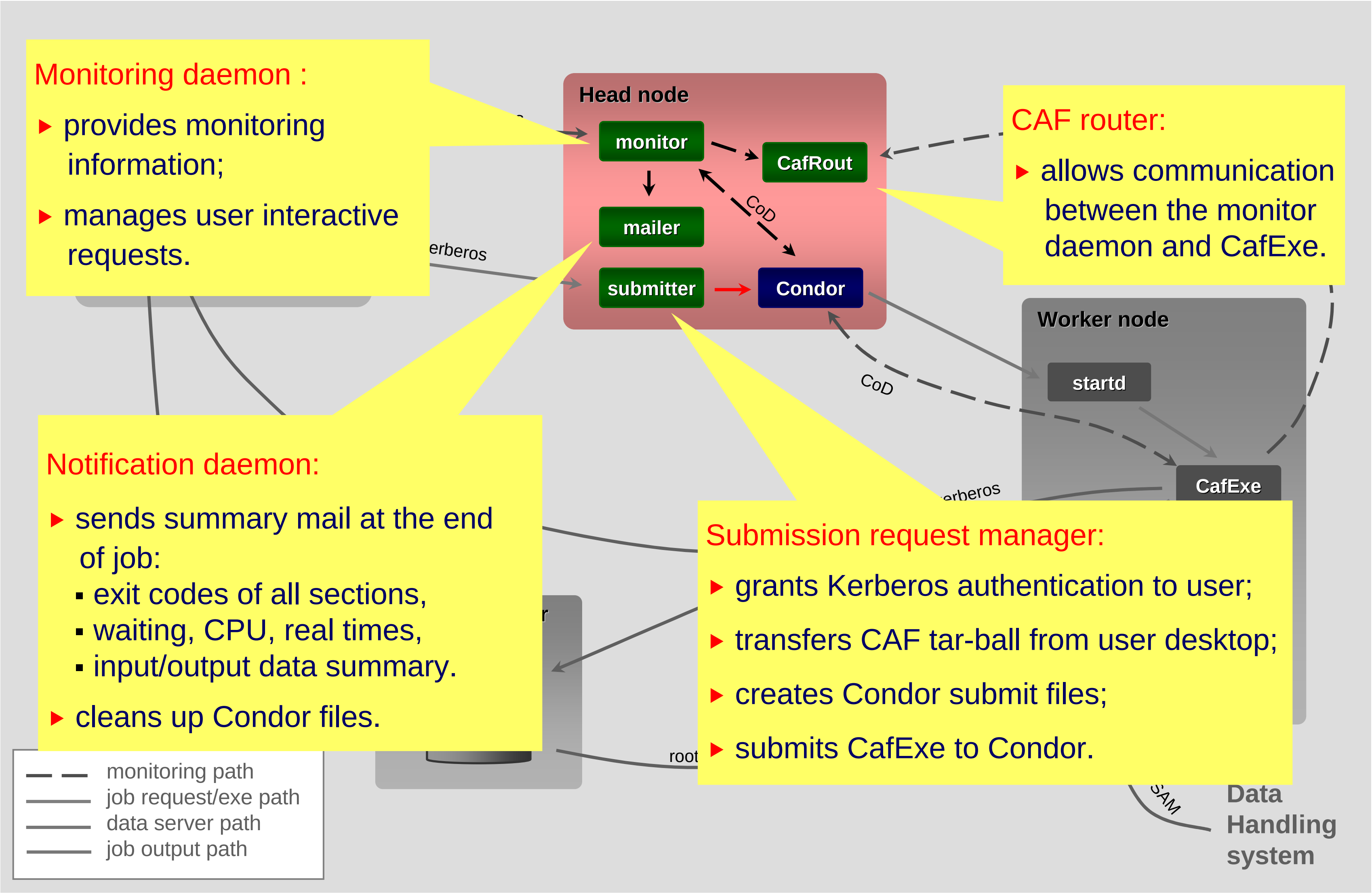
User front-end for job submission:

- communicates with head node via a Kerberos authenticated connection;
- sends user submission parameters to head node.

Web browser

- give access to the CAF web monitoring pages.

CAF software: head node



Monitoring daemon :

- provides monitoring information;
- manages user interactive requests.

CAF router:

- allows communication between the monitor daemon and CafExe.

Notification daemon:

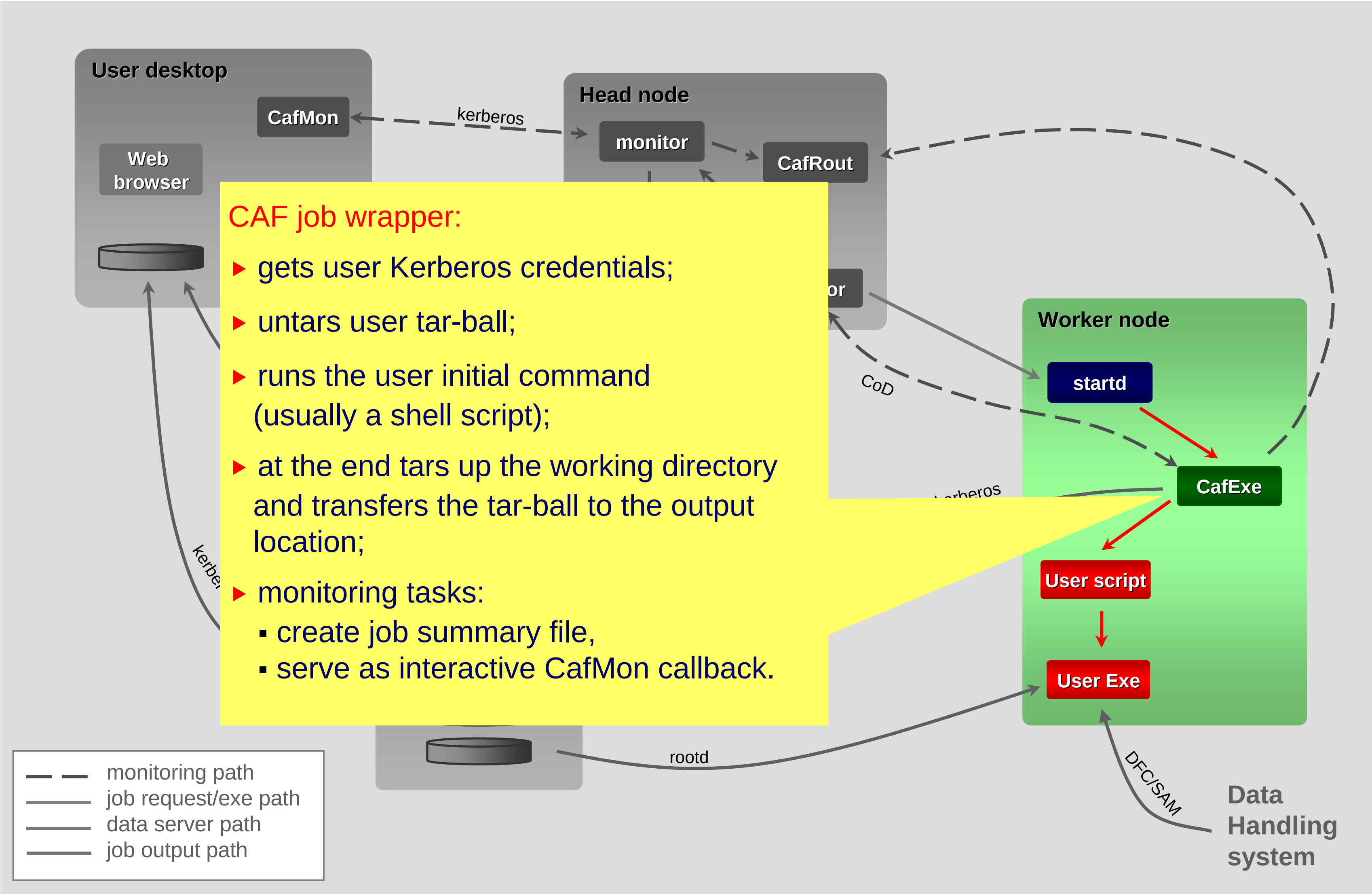
- sends summary mail at the end of job:
  - exit codes of all sections,
  - waiting, CPU, real times,
  - input/output data summary.
- cleans up Condor files.

Submission request manager:

- grants Kerberos authentication to user;
- transfers CAF tar-ball from user desktop;
- creates Condor submit files;
- submits CafExe to Condor.

— —	monitoring path
——	job request/exe path
——	data server path
——	job output path

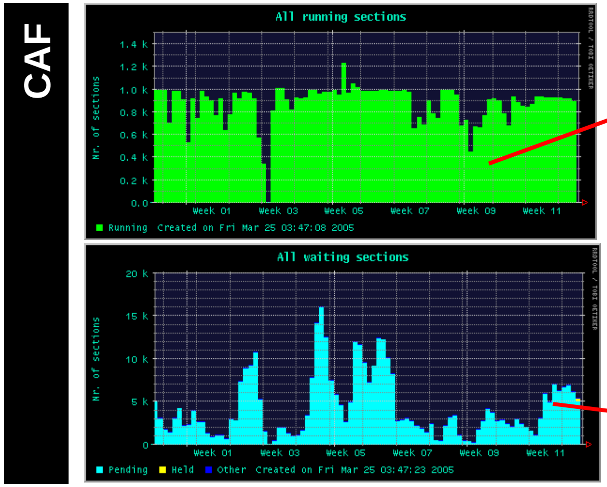
CAF software: worker node



CAF job wrapper:

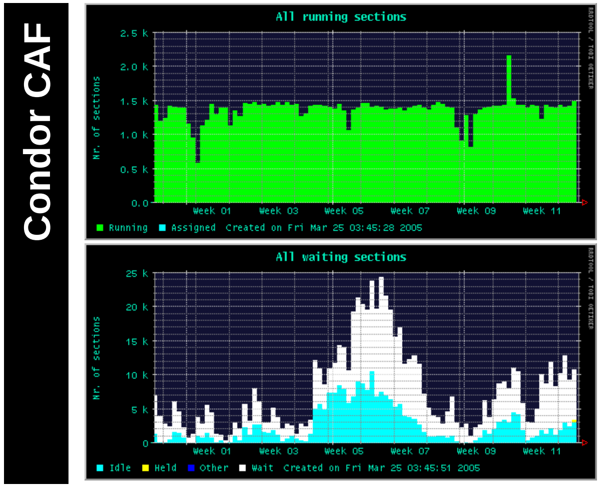
- gets user Kerberos credentials;
- untars user tar-ball;
- runs the user initial command (usually a shell script);
- at the end tars up the working directory and transfers the tar-ball to the output location;
- monitoring tasks:
  - create job summary file,
  - serve as interactive CafMon callback.

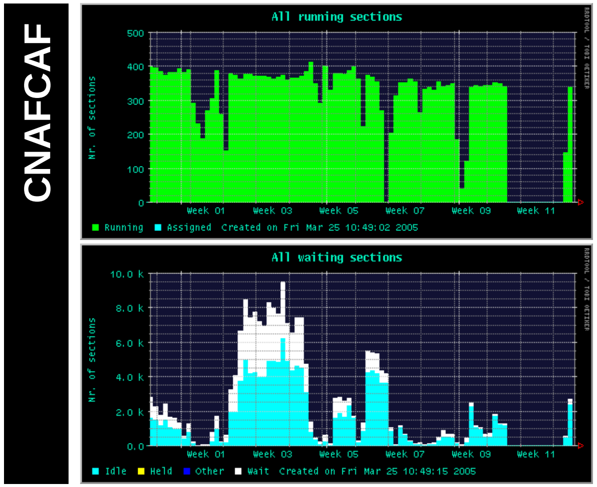
CAF and DCAF utilization

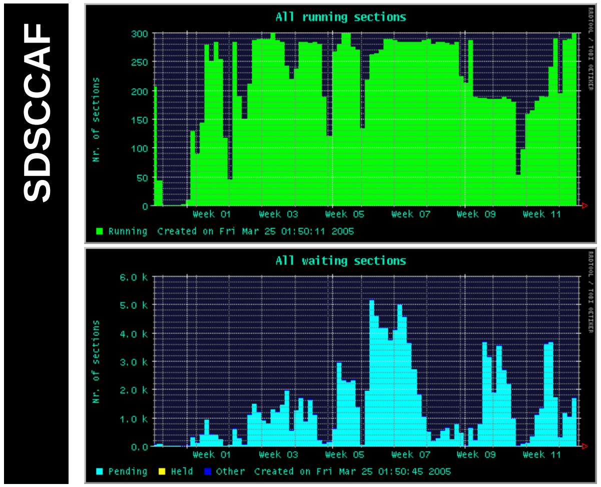

CAF

running sections

pending sections


Condor CAF


CNAFCAF


SDSCCAF

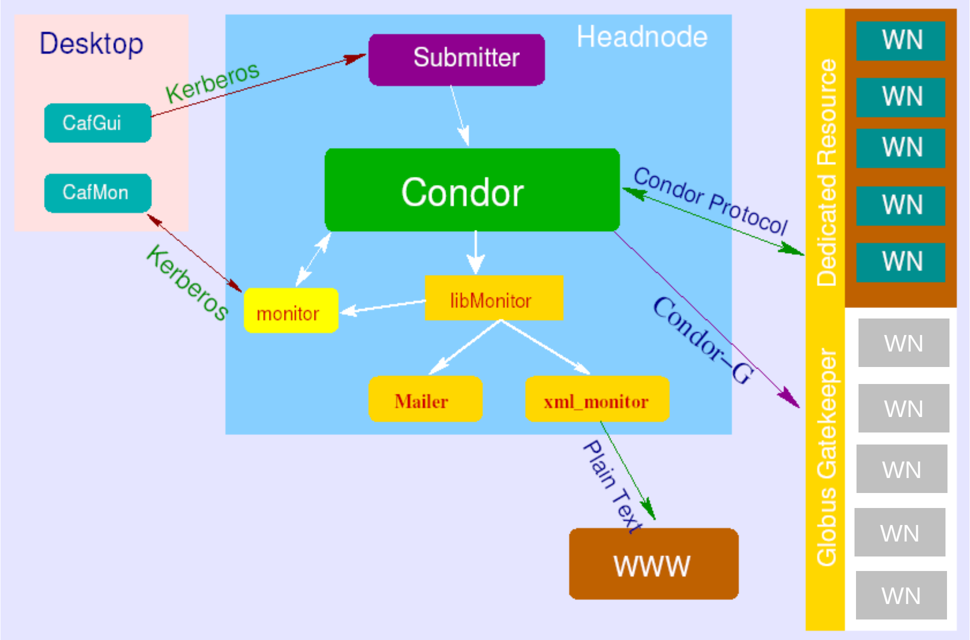
- The CDF increasing volume of data pushes for more and more computer power in the next future ⇒ GRID may offer plenty of resources, at least until LHC turns on.

Toward CafGRID

- First step: **GlideCAF at CNAF Tier1**
  - ✓ based on condor_glidin;
  - ✓ first working prototype.

condor_glideins are Condor-G jobs which bypass the Resource Broker and reach the GRID site Gatekeeper directly:

- the GateKeeper distributes the jobs to the WNs;
- the WNs install/run Condor on the fly once the Condor-G jobs start;
- the WNs become a part of the Condor pool when the Condor daemons start.



- Final goal:
  - ✓ modify CAF software in order to allow jobs submission to GRID;
  - ✓ integrate DCAFs into the GRID as Computer Elements is desirable, but not yet designed.

Conclusion

- The need for computing resources has been steadily urging the CDF Central Analysis Facility to evolve from an old-fashioned farm, originally localized at FNAL, to a GRID oriented structure, distributed worldwide (~50% of CDF computing power already outside FNAL).

- GRID represents an abundant reservoir of computing resources to fulfill CDF analysis needs.

- First step has been done to integrate CDF DCAFs into GRID, at CNAF a working prototype already exploits Tier1 resources:
  - ✓ users may submit MC jobs,
  - ✓ preliminary tests on data analysis jobs.