CHEP 2007

glideinWMS
-
A generic pilot-based Workload Management System

by Igor Sfiligoi (FNAL)

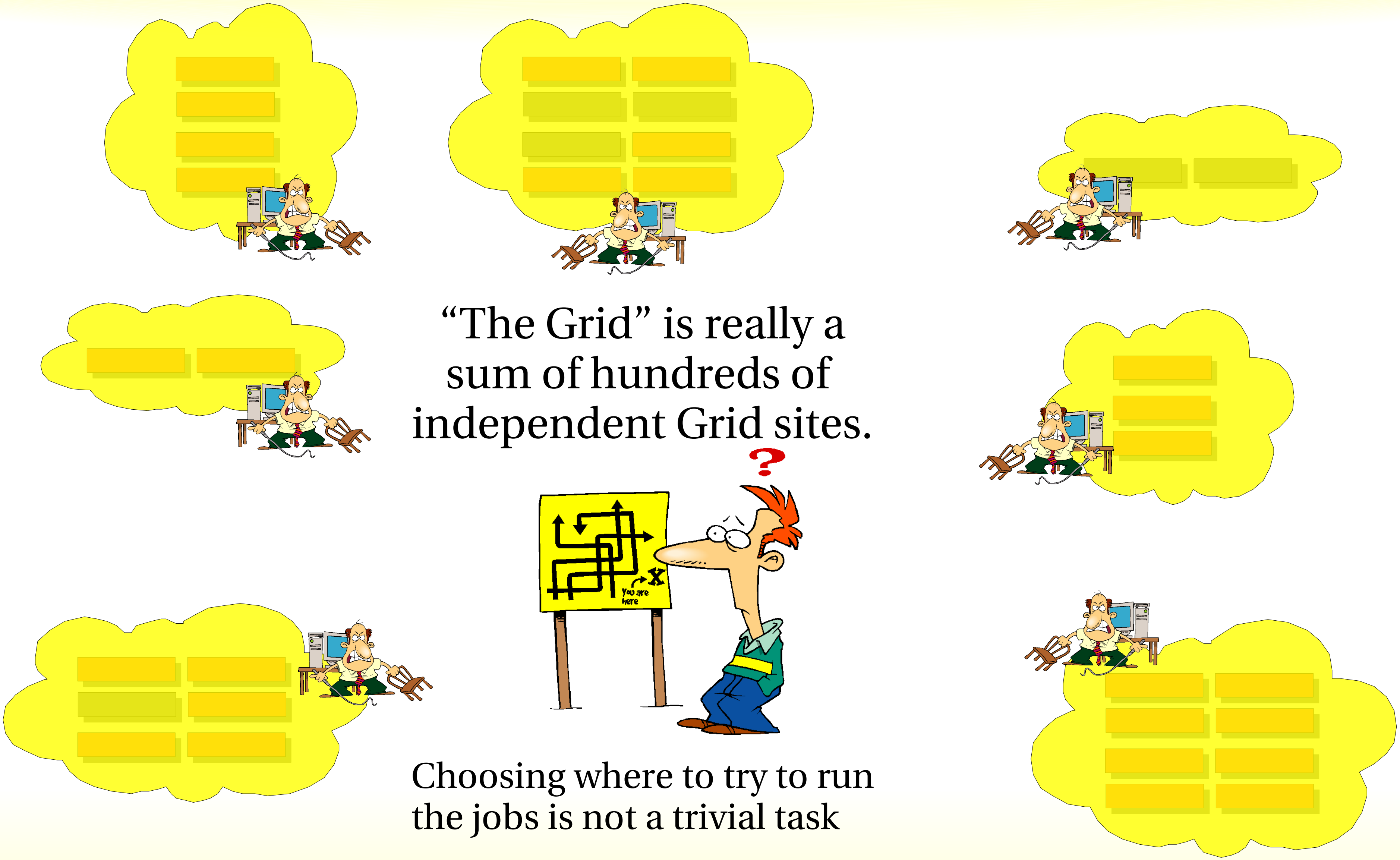
Outline

- What is glideinWMS?
- How does it work?
- How does it perform?
- Monitoring
- Conclusions

What is glideinWMS?

- A Condor glidein-based Workload Management System
- Developed by CMS for CMS, but generic enough to be used by other groups, too
  - A generalization of the CDF glidekeeper
- Available at: http://home.fnal.gov/~sfiligoi/glideinWMS/

Why do we need a WMS?

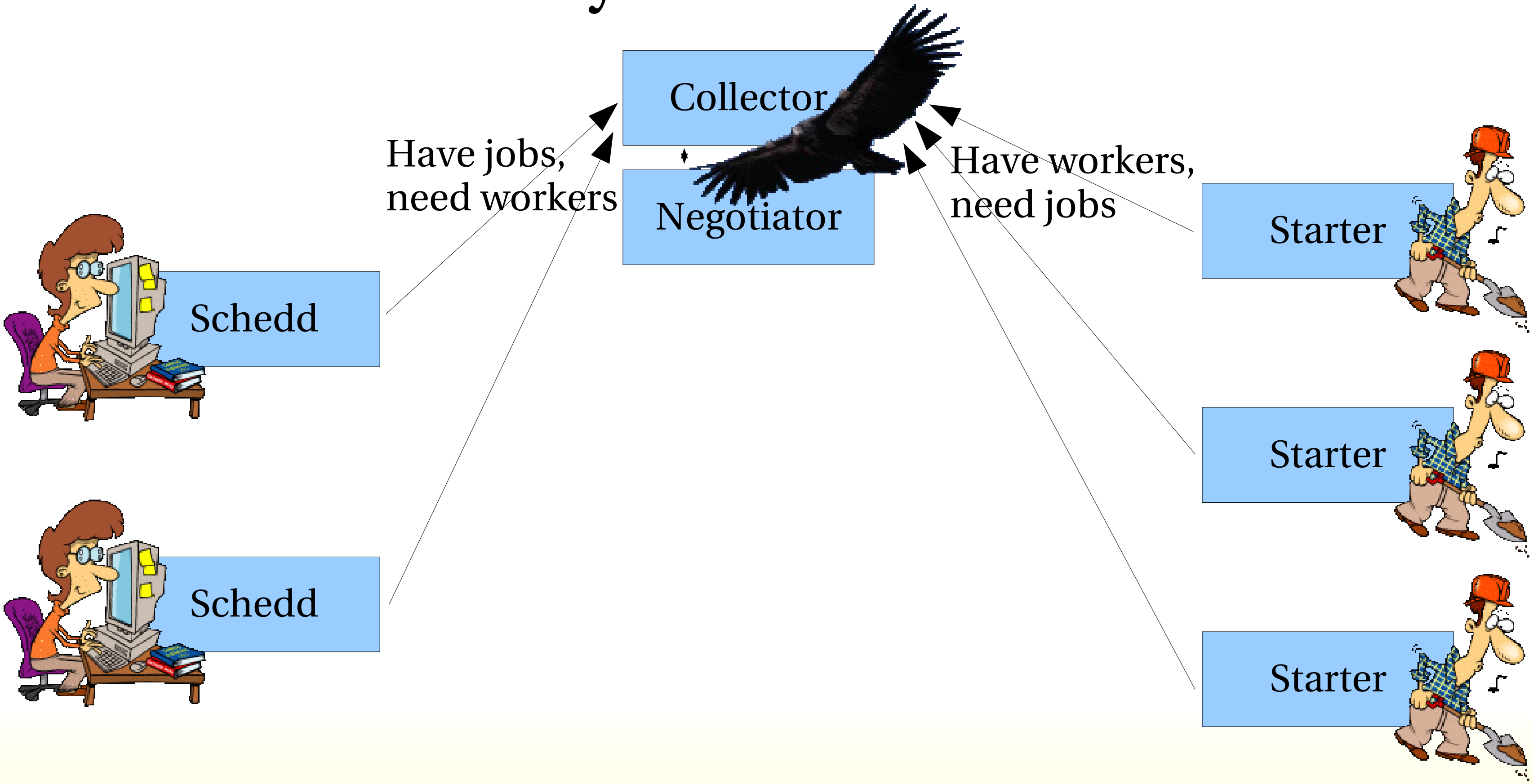


“The Grid” is really a sum of hundreds of independent Grid sites.

Choosing where to try to run the jobs is not a trivial task

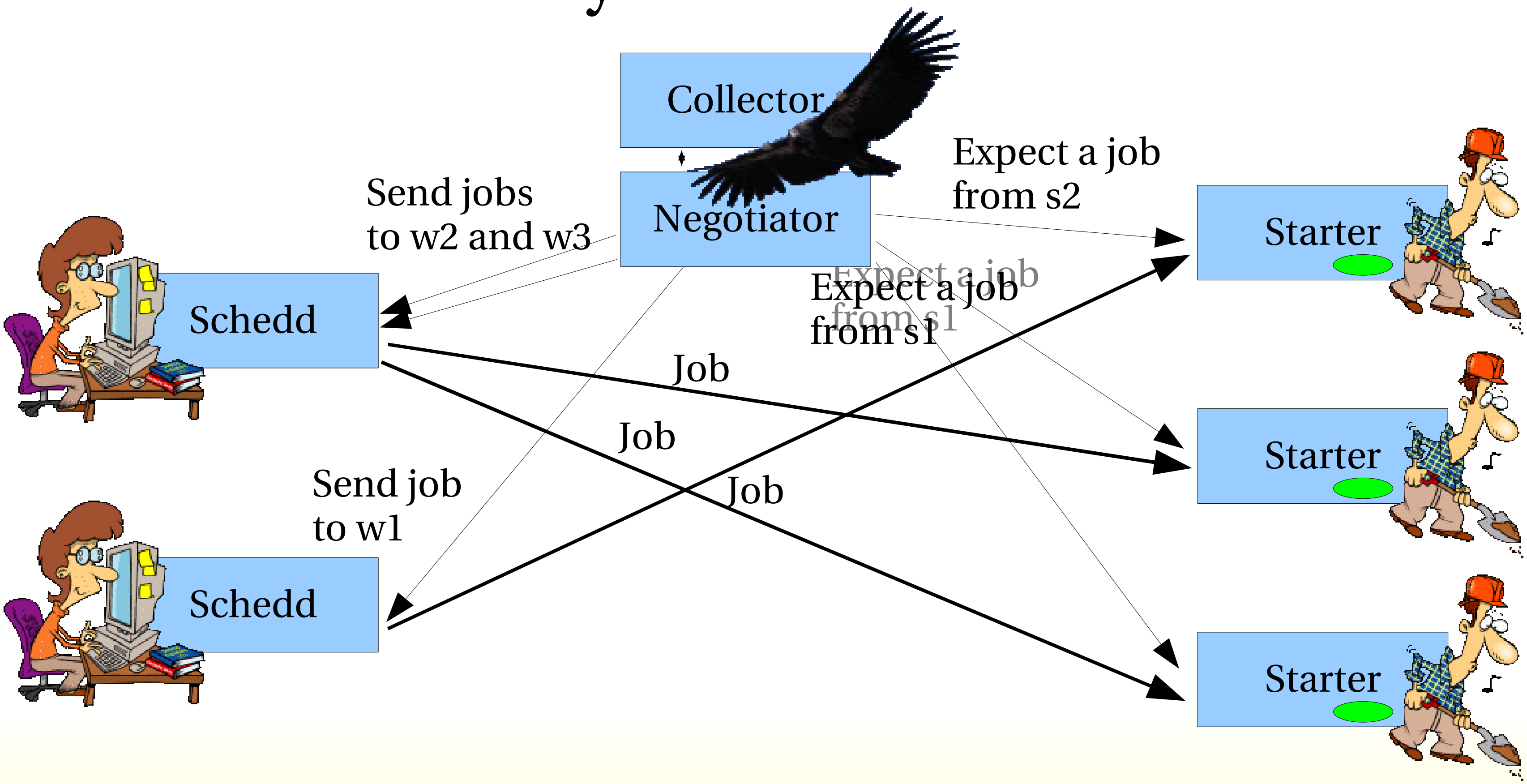
What is Condor? (1)

- A widely used batch system
- Based on a fully distributed architecture



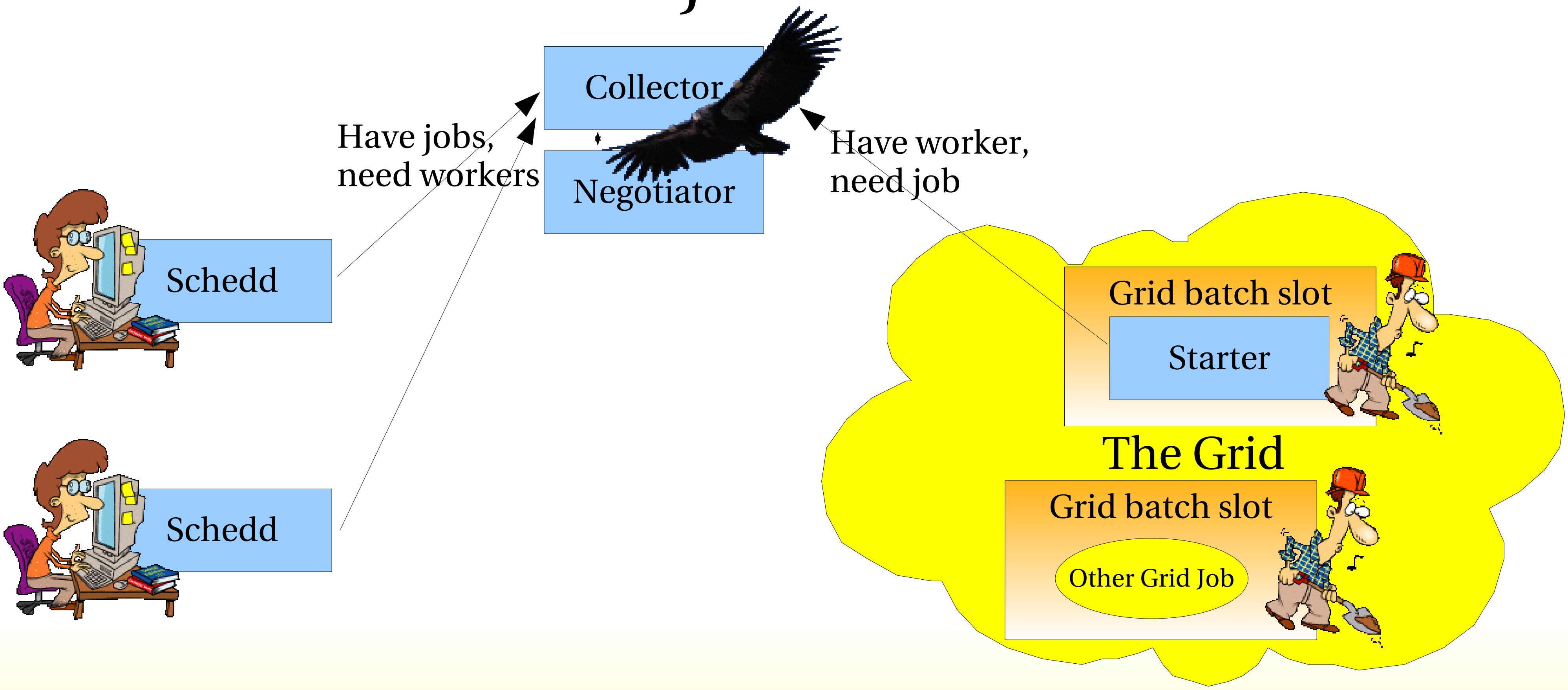
What is Condor? (2)

- A widely used batch system
- Based on a fully distributed architecture



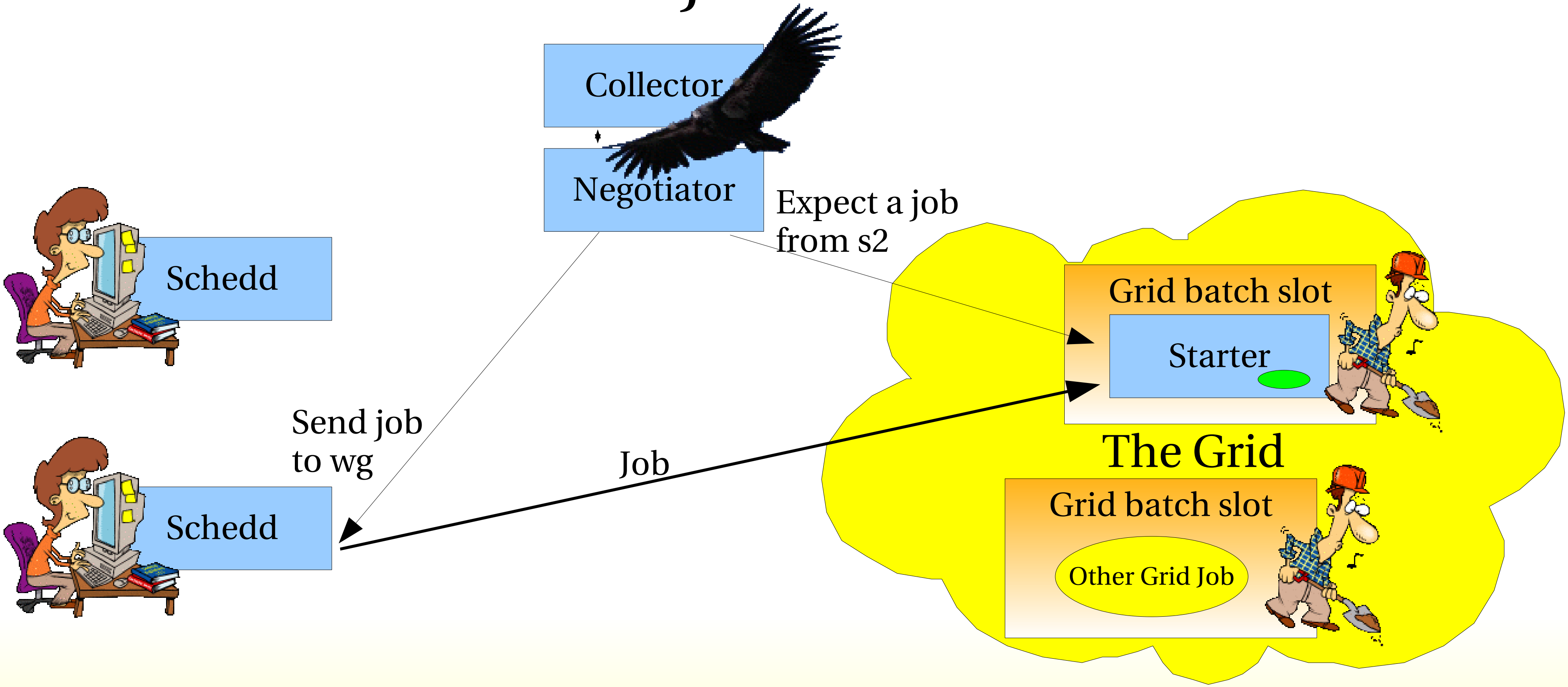
What is a glidein? (1)

- Just a regular starter
- Submitted as a Grid job



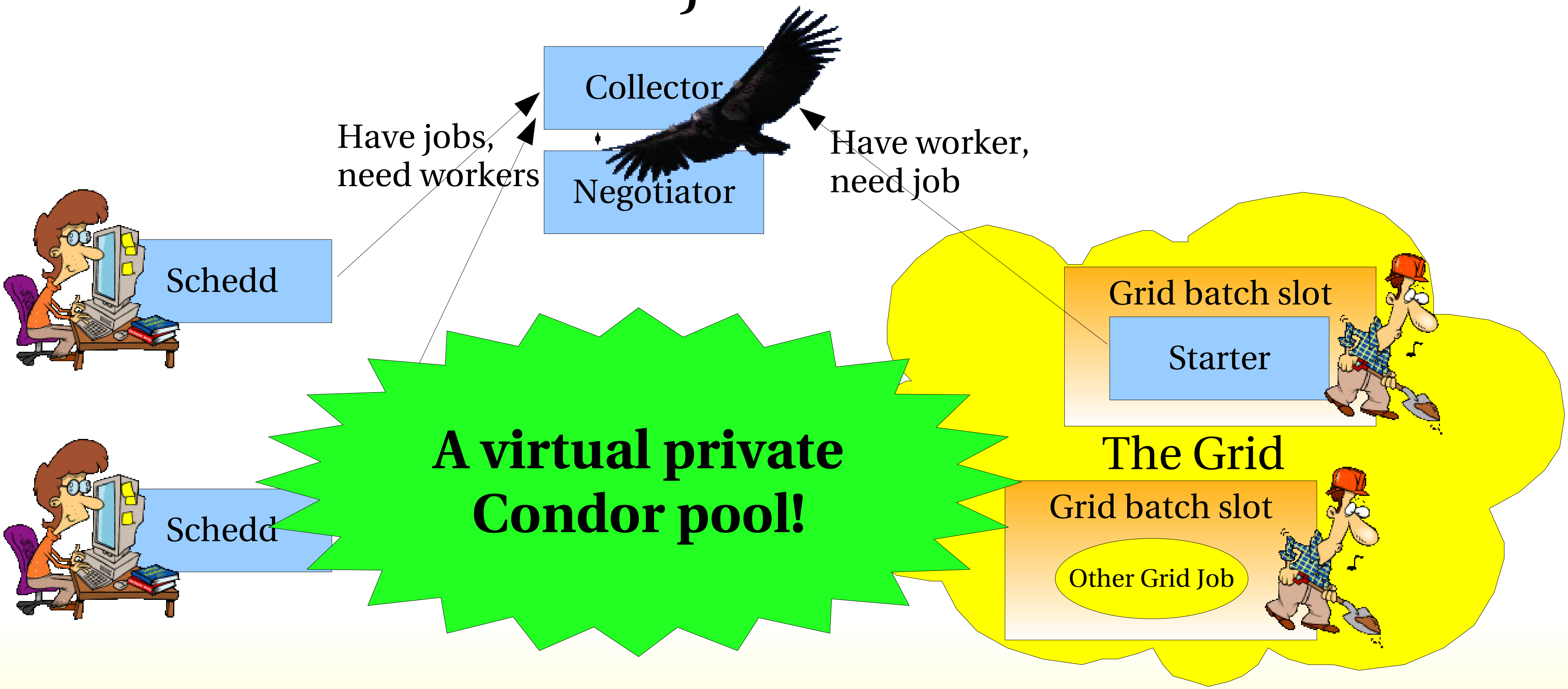
What is a glidein? (2)

- Just a regular starter
- Submitted as a Grid job



What is a glidein? (3)

- Just a regular starter
- Submitted as a Grid job



What else can a glidein do?

- Make sanity checks before fetching any job
- Discover and publish batch slot characteristics:
  - OS version
  - CPU model, available RAM and disk
  - Availability of certain software
- Importing VO specific software
- Prepare the environment for the user jobs
  - Possibly putting the VO software in the path
- etc.

Why using glideins? (1)

- For people already using Condor
  - An easy way to extend the pool
    - Or to create one from scratch
  - Can hide all the grid stuff from user jobs
    - Can even run standard universe jobs on the Grid!
- For people just wanting to use the Grid (even if not Condor fans)
  - Protect user jobs from many obvious errors
    - A dead glidein will not pull a user job
  - Simplifies resource selection
    - A glidein can detect what is available on the worker and user jobs get sent only to complying workers
    - No guessing involved, job sent after resource acquired

Why using glideins? (2)

- Get all the advantages of a local batch system
  - Locally set priorities between different users
    - Including group quotas
    - Or even priorities between jobs of the same user
  - Reliable, real time monitoring
  - Reliable file transfer
    - Full file encryption supported, too
- While still running on the Grid!

Any weak points will be presented at the end

How do I submit a glidein?

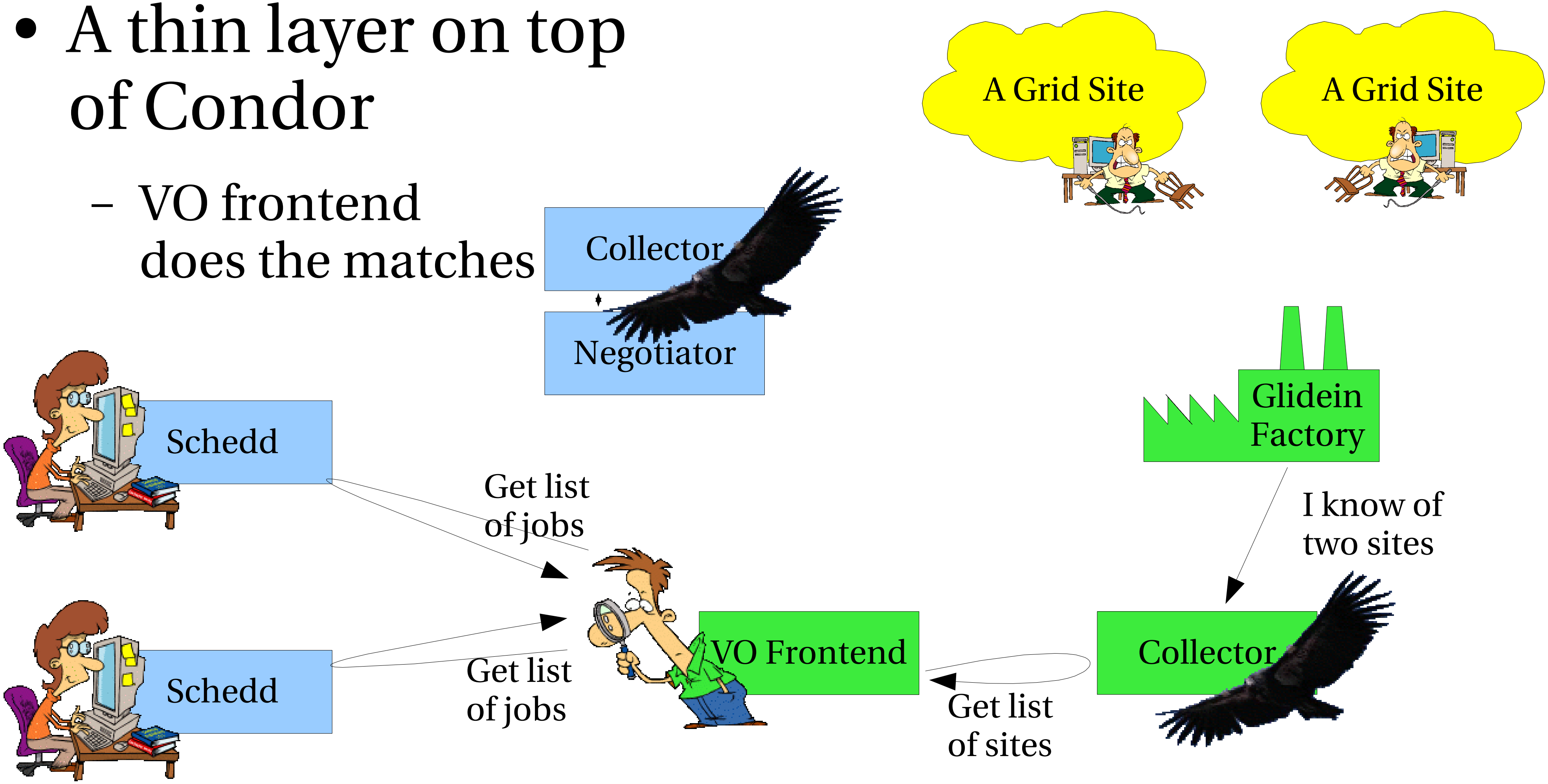
- Condor provides condor_glidein
  - Simple command line tool
  - **Useful when you have just a few jobs**
    - Will submit a single glidein per invocation
- Install a glideinWMS instance
  - Needs more resources and some initial effort to set it up
  - **Setup once, glideins will be launched as needed**
    - Will look for jobs that need resources
    - Submit glideins as needed to sites that seem to match at least an idle job

glideinWMS

How does it work?

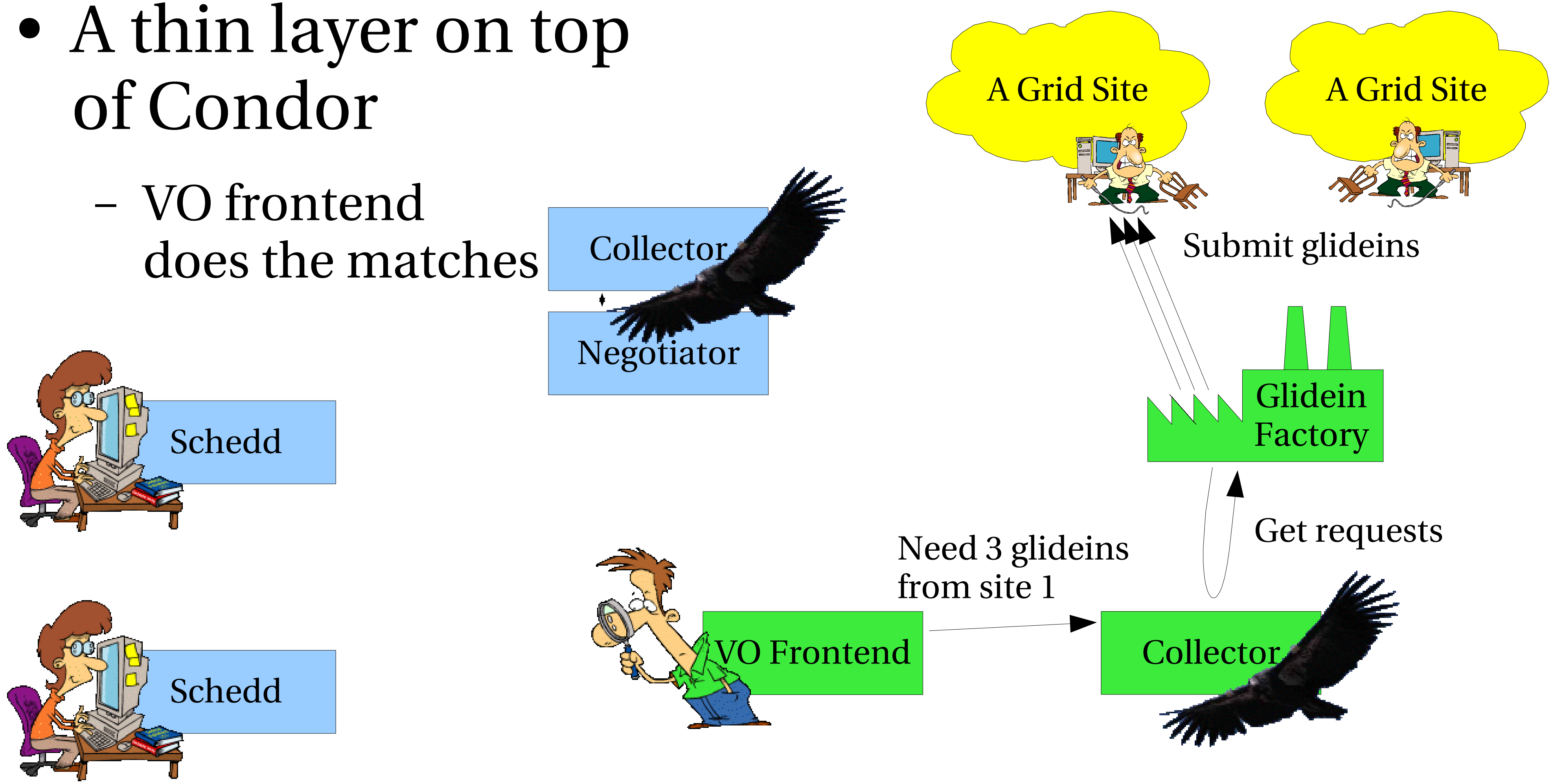
glideinWMS overview (1)

- A thin layer on top of Condor
  - VO frontend does the matches



glideinWMS overview (2)

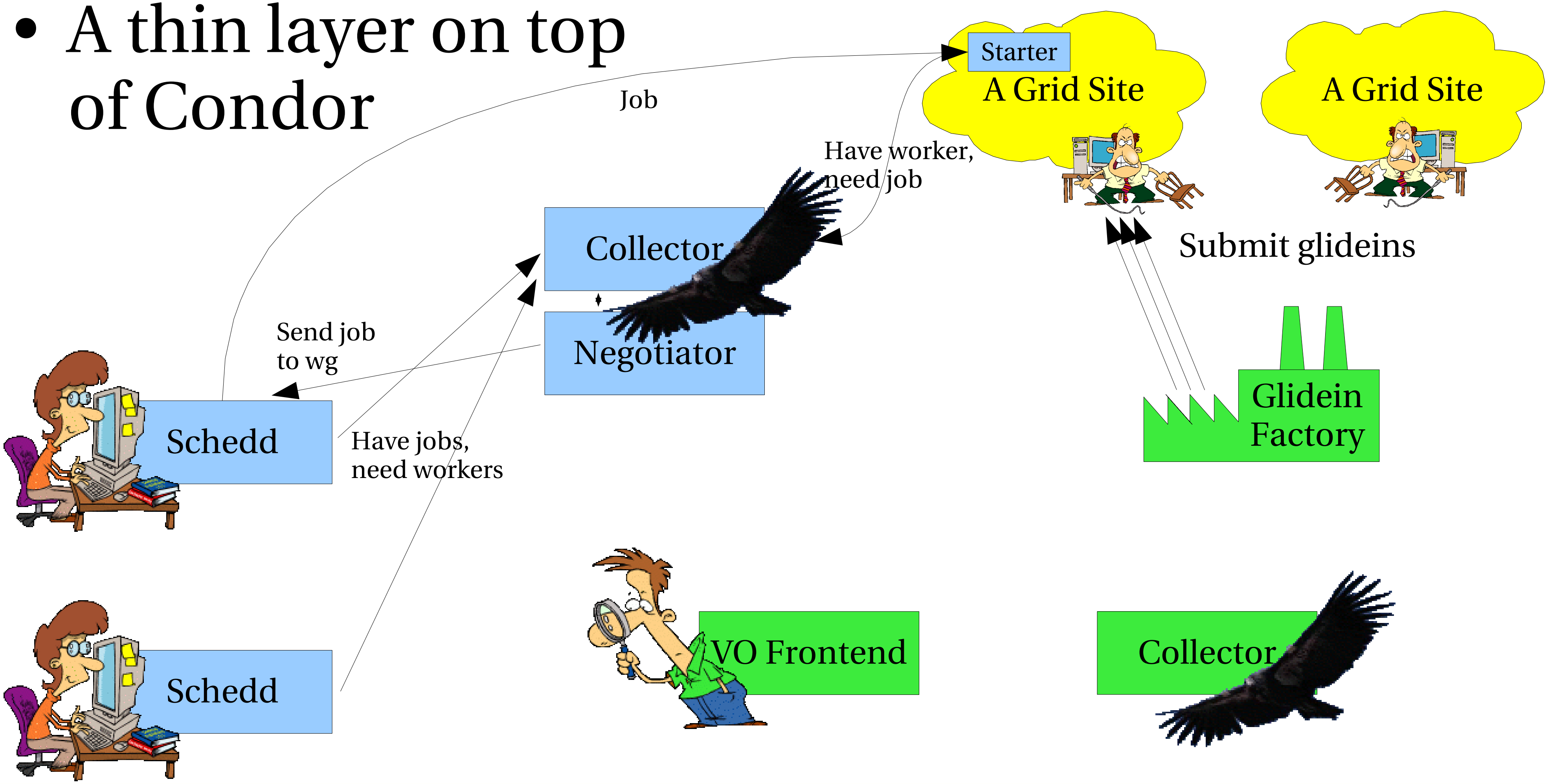
- A thin layer on top of Condor
  - VO frontend does the matches



More details at http://cd-docdb.fnal.gov/cgi-bin/ShowDocument?docid=2048

glideinWMS overview (3)

- A thin layer on top of Condor



glideinWMS details(1)

- Matchmaking done on two levels
  - **VO frontend** matches glideins to sites that claim to support at least one job waiting in the queue
  - The **condor negotiator** matches glidein starters to the jobs waiting in the queue
- The condor negotiator has the final word
  - If a site was lying about its capabilities, the starter will not be matched and will exit within minutes
  - The job that is sent to a starter might not be the one for which the glidein was submitted for

glideinWMS details(2)

- The WMS logic is to keep constant pressure on the Grid sites
  - As long as there are waiting jobs that could be run on a site, it tries to keep a steady number of idle glideins in the site queues
- The VO frontend drives the WMS
  - Deciding how much pressure to put on different sites
  - The glidein factories will submit the glideins, following the orders from the VO frontend

glideinWMS details(3)

- Communication between processes based on Condor ClassAds
  - For each site, a Glidein Factory publishes:
    - CE attributes
    - list of parameters it accepts
  - Each VO Frontend replies a ClassAd containing:
    - The target site
    - VO parameters (a subset of the above)
    - Number of idle glideins to keep in the queue
- Using a standard Condor collector

glideinWMS details(4)

- Condor-G used for glidein submission
- The list of sites a factory serves is a configuration parameter
  - – Can be set manually, fine tuning each and every site characteristics
  - – Easy to script
    - Just a standard XML file
    - The installation script can use the CRONUS information system, and ReSS information system will be added soon
  - – Or can be paired with a Condor-G matchmaker, like ReSS and CRONUS

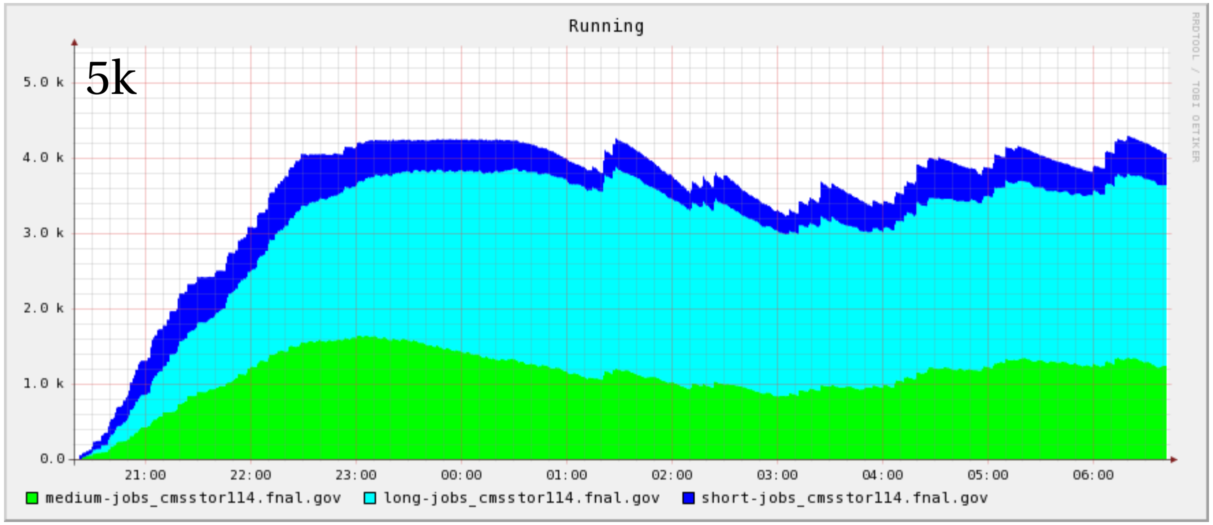
glideinWMS

How does it perform?

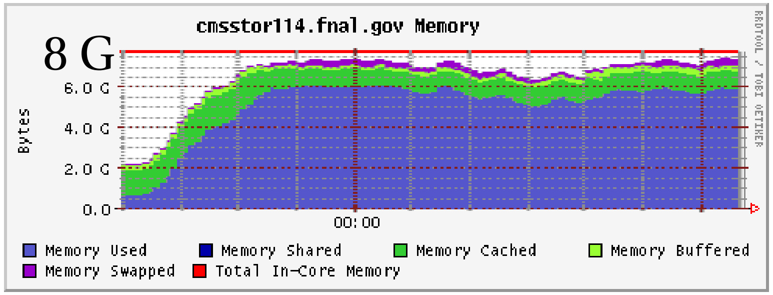
Do glideins scale? (1)

- Synthetic tests, using a single submit machine, scaled well with ~4000 running jobs
  - Memory a major limiting factor

See also
Talk #216



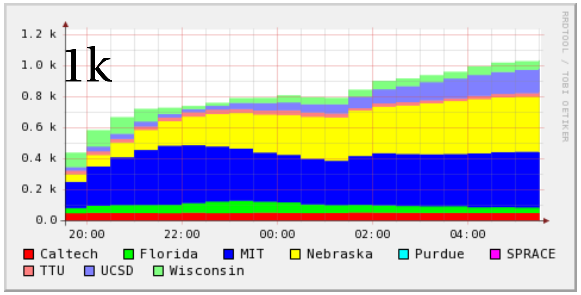
Ignore
the colors



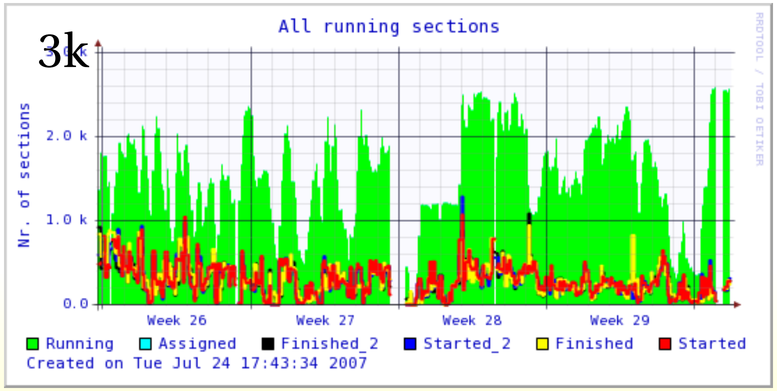
- Further scalability can be obtained by using multiple submission machines

Do glideins scale? (2)

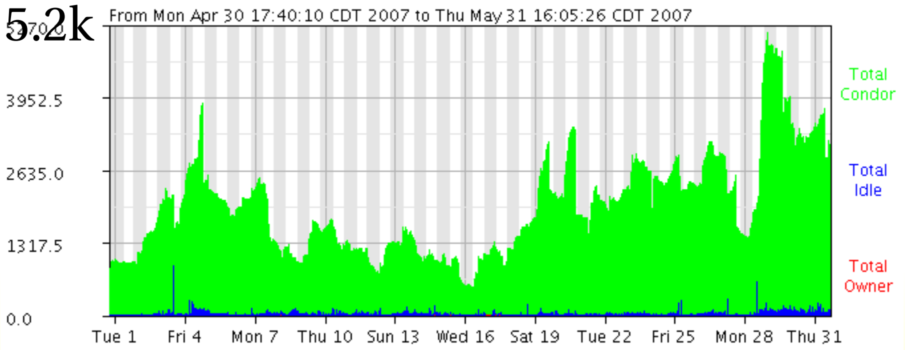
- glideinWMS-based CMS MC production up to 1k jobs in parallel



- CDF GlideCAF up to 2.5k

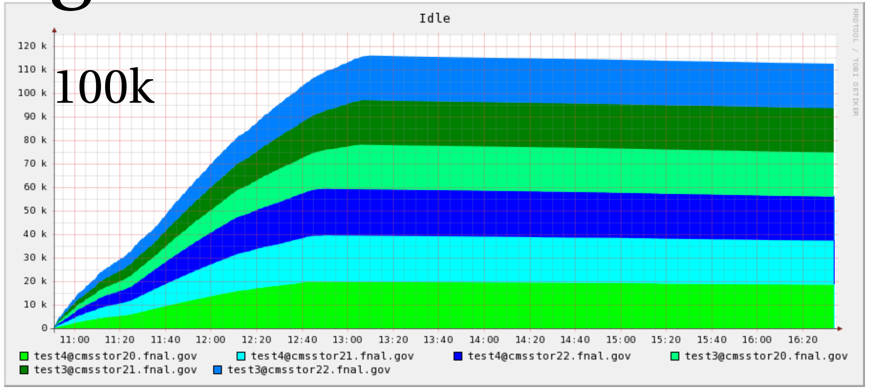


- ATLAS Cronus up to 5k

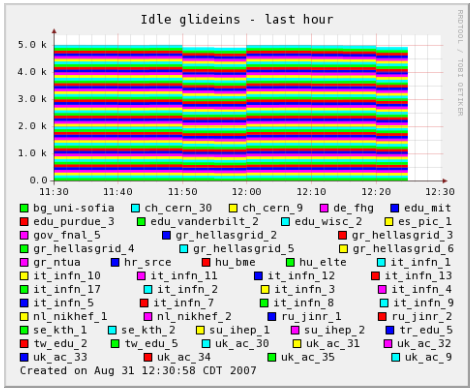


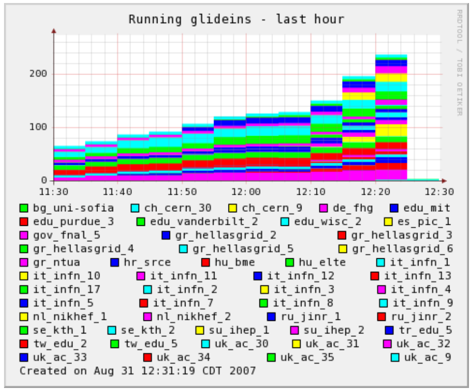
Does glideinWMS scale?

- Synthetic tests with single VM frontend and single glidein factory
  - 6 submission points with 100k queued jobs



  - 50 grid sites

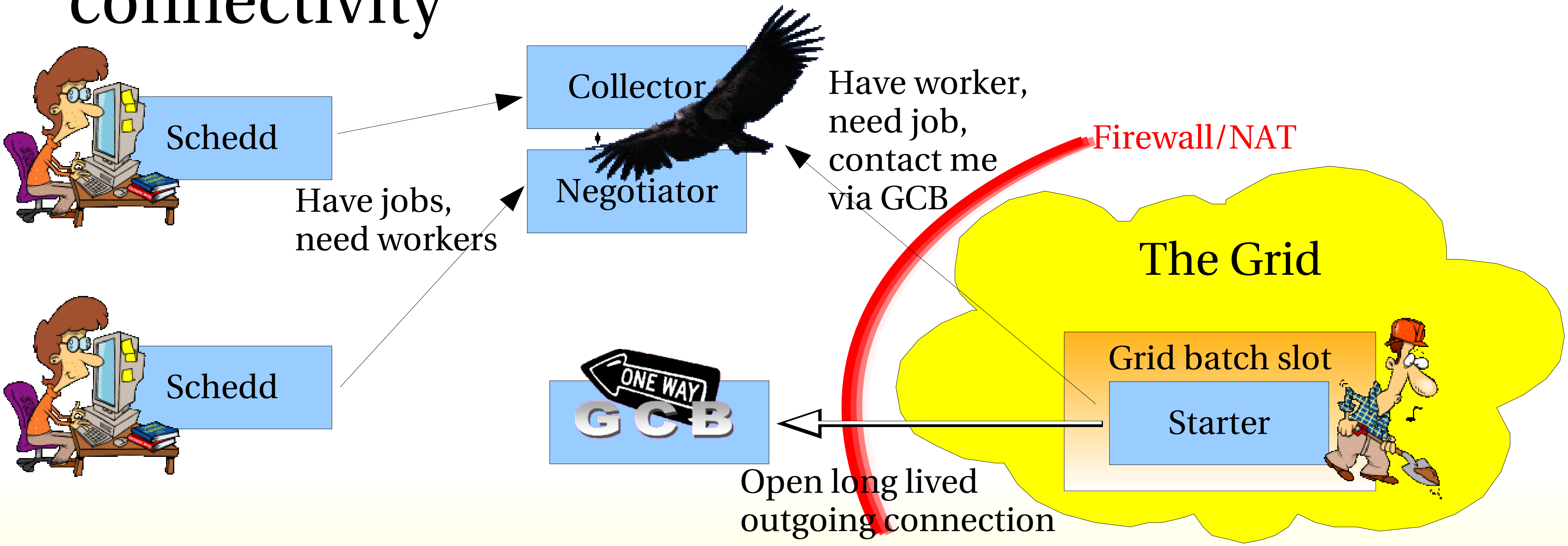




- Further scalability by using multiple VO frontends and multiple glidein factories

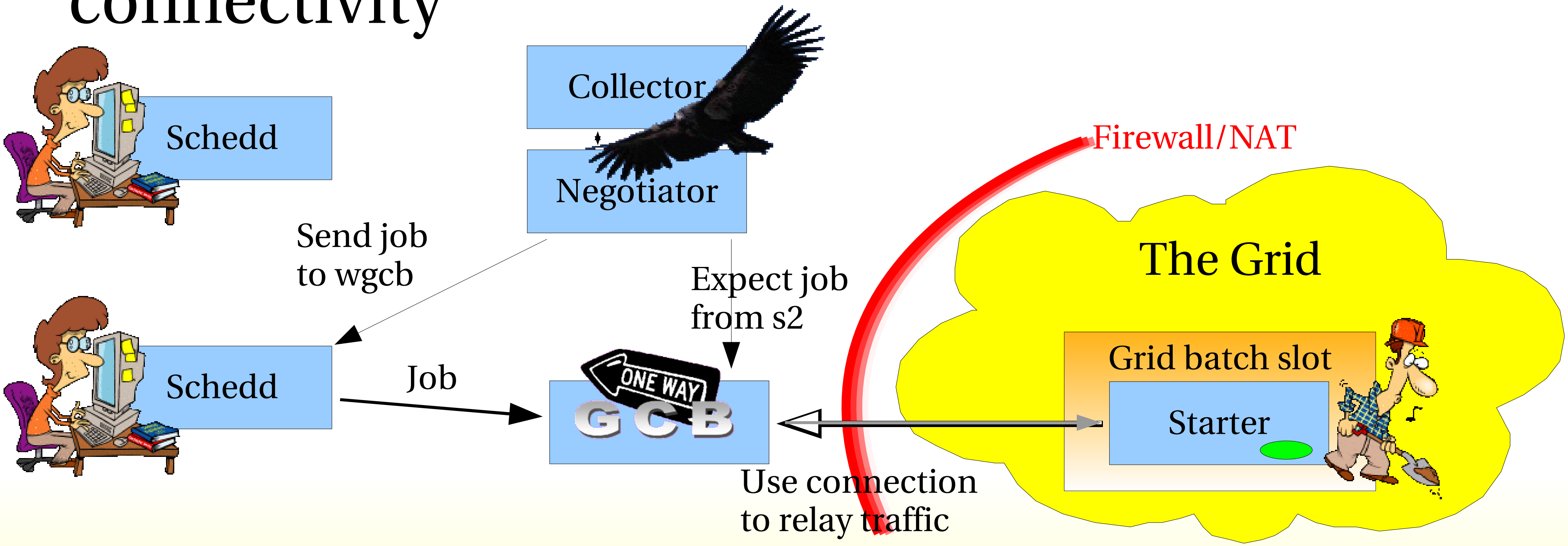
Do glideins really work over WAN? (1)

- Yes, but it needs GCB to work
  - A Condor proxy server
- The only requirement is that there is outgoing connectivity



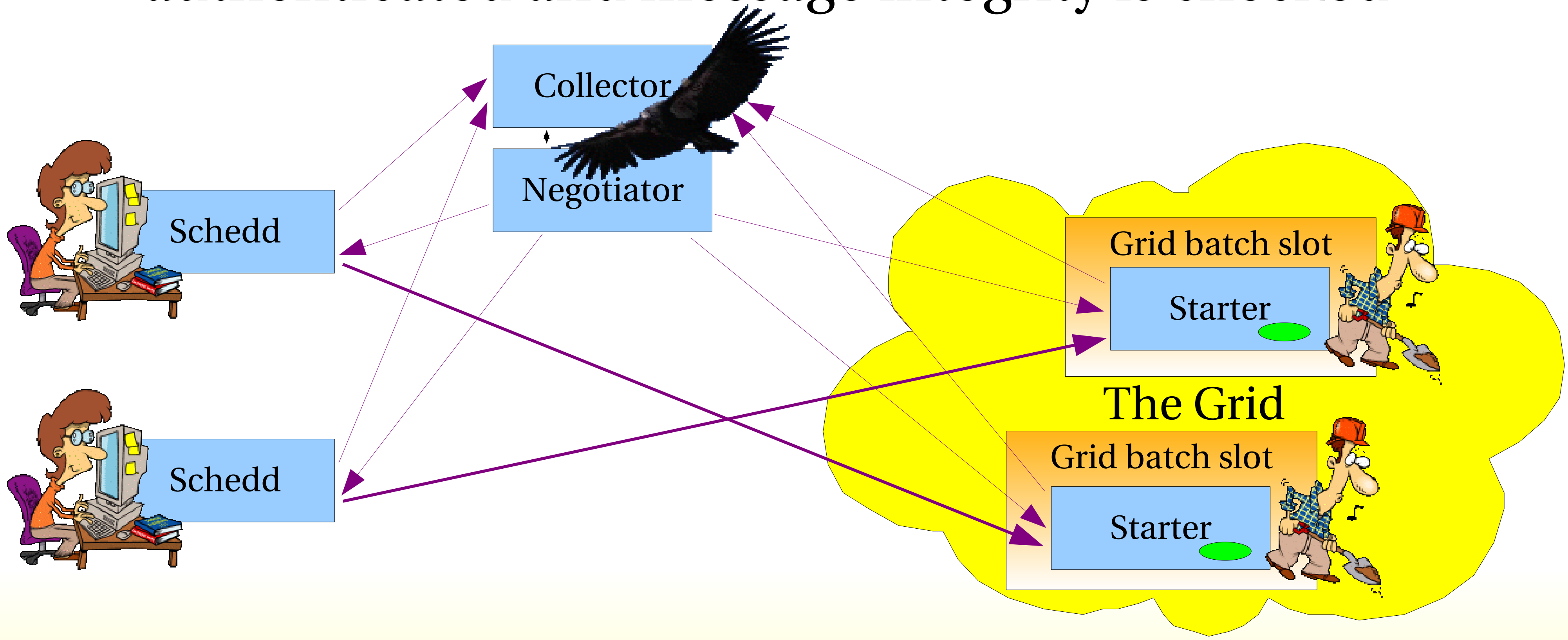
Do glideins really work over WAN? (2)

- Yes, but it needs GCB to work
  - A Condor proxy server
- The only requirement is that there is outgoing connectivity



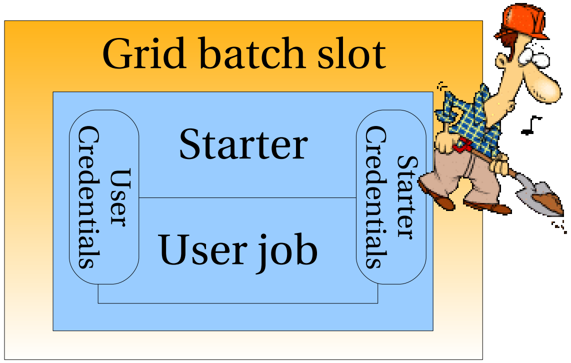
What about security? (1)

- glideinWMS glideins use GSI for authentication
  - All daemon to daemon communication is fully authenticated and message integrity is checked

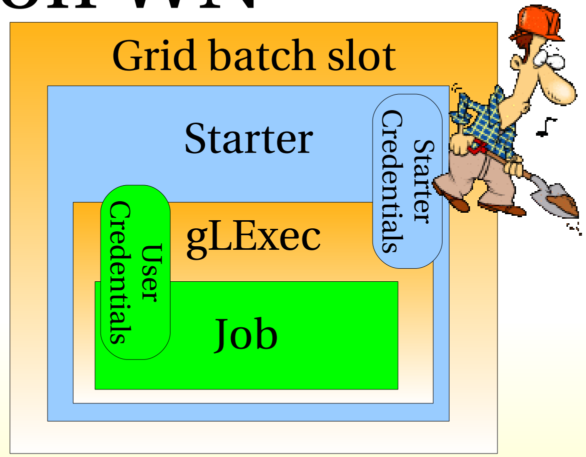


What about security? (2)

- However, starter does not run as root!
  - Without help the user and the starter have to run under the same account!
    - The malicious user job can use starter privileges
    - In this scenario the glidein should only run jobs from the factory user



- OSG is starting to deploy gLExec on WN
  - Allows starter to start user job under appropriate UID
  - See gLExec talks #43 and #94



Any other drawbacks?(1)

- Condor uses a lot of resources
  - Be prepared to budget 1.5Mb of RAM per running process on the submit node
  - Possibly distribute job submission over multiple nodes
- GCB is still in active development phase
  - Production version stable only to ~600 running jobs per GCB/schedd pair
    - Need to deploy many of them to scale
    - Development team is promising much higher scalability

See also Talk #216

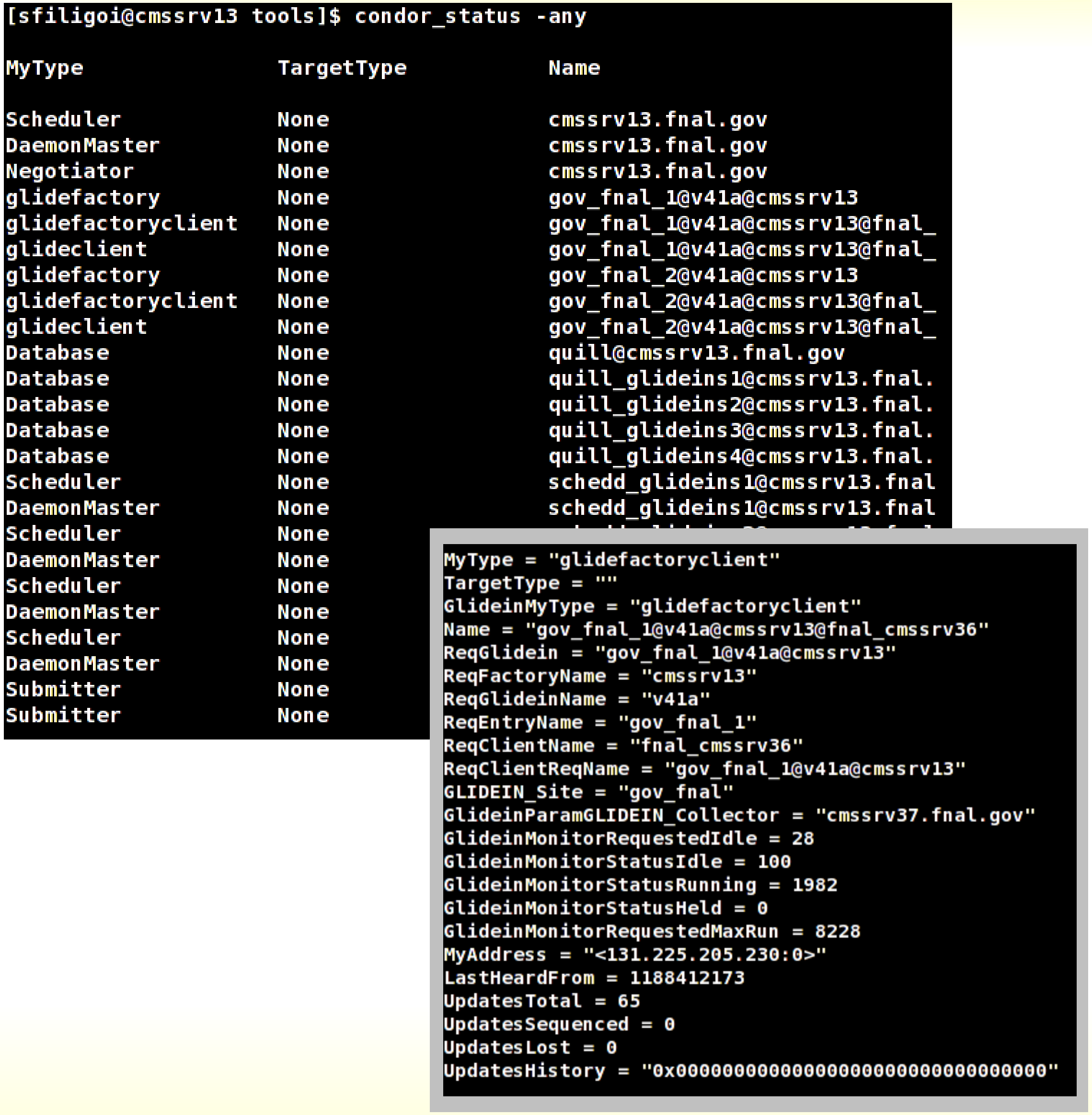
Any other drawbacks? (2)

- Glidein factory load scales with number of Grid sites
  - Budget one node every few dozen sites for current version
    - or pair it with Condor-G matchmakers, like ReSS
  - Work in progress to reduce the load
- If anything goes wrong with the setup, the debugging can be challenging
  - Glidein log files are returned only when the job finishes
    - May not get them back, if it never ends
    - This is a fundamental Grid limitation, not much that can be done about it

glideinWMS

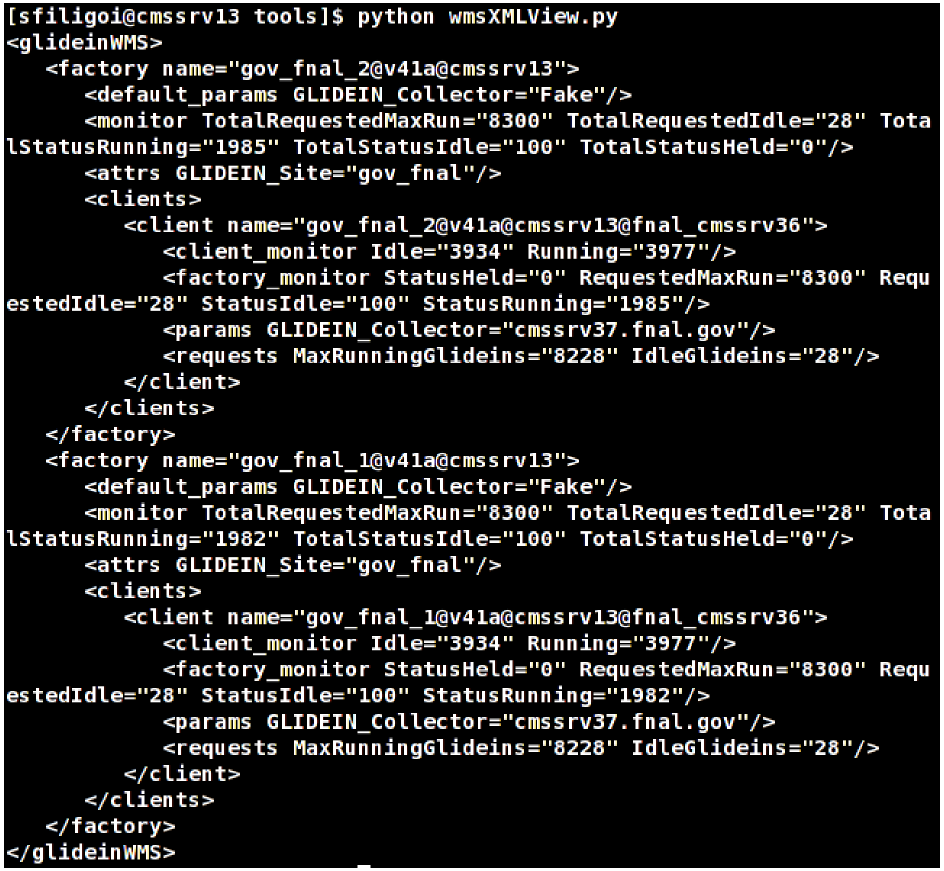
Monitoring

Condor collector monitoring



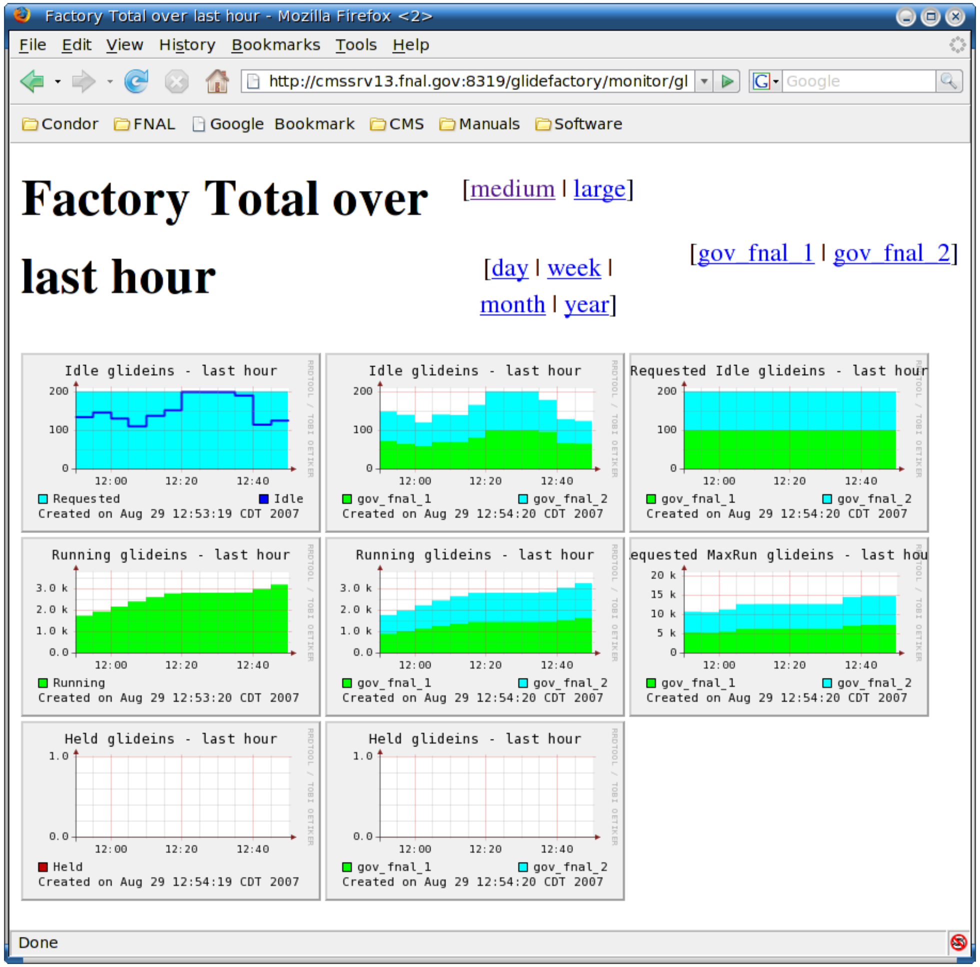
```
[sfiligoi@cmssrv13 tools]$ condor_status -any

MyType               TargetType           Name

Scheduler            None                 cmssrv13.fnal.gov
DaemonMaster         None                 cmssrv13.fnal.gov
Negotiator           None                 cmssrv13.fnal.gov
glidefactory         None                 gov_fnal_1@v41a@cmssrv13
glidefactoryclient   None                 gov_fnal_1@v41a@cmssrv13@fnal_
glideclient          None                 gov_fnal_1@v41a@cmssrv13@fnal_
glidefactory         None                 gov_fnal_2@v41a@cmssrv13
glidefactoryclient   None                 gov_fnal_2@v41a@cmssrv13@fnal_
glideclient          None                 gov_fnal_2@v41a@cmssrv13@fnal_
Database             None                 quill@cmssrv13.fnal.gov
Database             None                 quill_glideins1@cmssrv13.fnal.
Database             None                 quill_glideins2@cmssrv13.fnal.
Database             None                 quill_glideins3@cmssrv13.fnal.
Database             None                 quill_glideins4@cmssrv13.fnal.
Scheduler            None                 schedd_glideins1@cmssrv13.fnal
DaemonMaster         None                 schedd_glideins1@cmssrv13.fnal
Scheduler            None
DaemonMaster         None
Scheduler            None
DaemonMaster         None
Scheduler            None
DaemonMaster         None
Submitter            None
Submitter            None
```

```
MyType = "glidefactoryclient"
TargetType = ""
GlideinMyType = "glidefactoryclient"
Name = "gov_fnal_1@v41a@cmssrv13@fnal_cmssrv36"
ReqGlidein = "gov_fnal_1@v41a@cmssrv13"
ReqFactoryName = "cmssrv13"
ReqGlideinName = "v41a"
ReqEntryName = "gov_fnal_1"
ReqClientName = "fnal_cmssrv36"
ReqClientReqName = "gov_fnal_1@v41a@cmssrv13"
GLIDEIN_Site = "gov_fnal"
GlideinParamGLIDEIN_Collector = "cmssrv37.fnal.gov"
GlideinMonitorRequestedIdle = 28
GlideinMonitorStatusIdle = 100
GlideinMonitorStatusRunning = 1982
GlideinMonitorStatusHeld = 0
GlideinMonitorRequestedMaxRun = 8228
MyAddress = "<131.225.205.230:0>"
LastHeardFrom = 1188412173
UpdatesTotal = 65
UpdatesSequenced = 0
UpdatesLost = 0
UpdatesHistory = "0x00000000000000000000000000000000"
```



```
[sfiligoi@cmssrv13 tools]$ python wmsXMLView.py
<glideinWMS>
   <factory name="gov_fnal_2@v41a@cmssrv13">
      <default_params GLIDEIN_Collector="Fake"/>
      <monitor TotalRequestedMaxRun="8300" TotalRequestedIdle="28" Tota
lStatusRunning="1985" TotalStatusIdle="100" TotalStatusHeld="0"/>
      <attrs GLIDEIN_Site="gov_fnal"/>
      <clients>
         <client name="gov_fnal_2@v41a@cmssrv13@fnal_cmssrv36">
            <client_monitor Idle="3934" Running="3977"/>
            <factory_monitor StatusHeld="0" RequestedMaxRun="8300" Requ
estedIdle="28" StatusIdle="100" StatusRunning="1985"/>
            <params GLIDEIN_Collector="cmssrv37.fnal.gov"/>
            <requests MaxRunningGlideins="8228" IdleGlideins="28"/>
         </client>
      </clients>
   </factory>
   <factory name="gov_fnal_1@v41a@cmssrv13">
      <default_params GLIDEIN_Collector="Fake"/>
      <monitor TotalRequestedMaxRun="8300" TotalRequestedIdle="28" Tota
lStatusRunning="1982" TotalStatusIdle="100" TotalStatusHeld="0"/>
      <attrs GLIDEIN_Site="gov_fnal"/>
      <clients>
         <client name="gov_fnal_1@v41a@cmssrv13@fnal_cmssrv36">
            <client_monitor Idle="3934" Running="3977"/>
            <factory_monitor StatusHeld="0" RequestedMaxRun="8300" Requ
estedIdle="28" StatusIdle="100" StatusRunning="1982"/>
            <params GLIDEIN_Collector="cmssrv37.fnal.gov"/>
            <requests MaxRunningGlideins="8228" IdleGlideins="28"/>
         </client>
      </clients>
   </factory>
</glideinWMS>
```

Status Web monitoring



Factory Total over last hour - Mozilla Firefox <2>

http://cmssrv13.fnal.gov:8319/glidefactory/monitor/gl

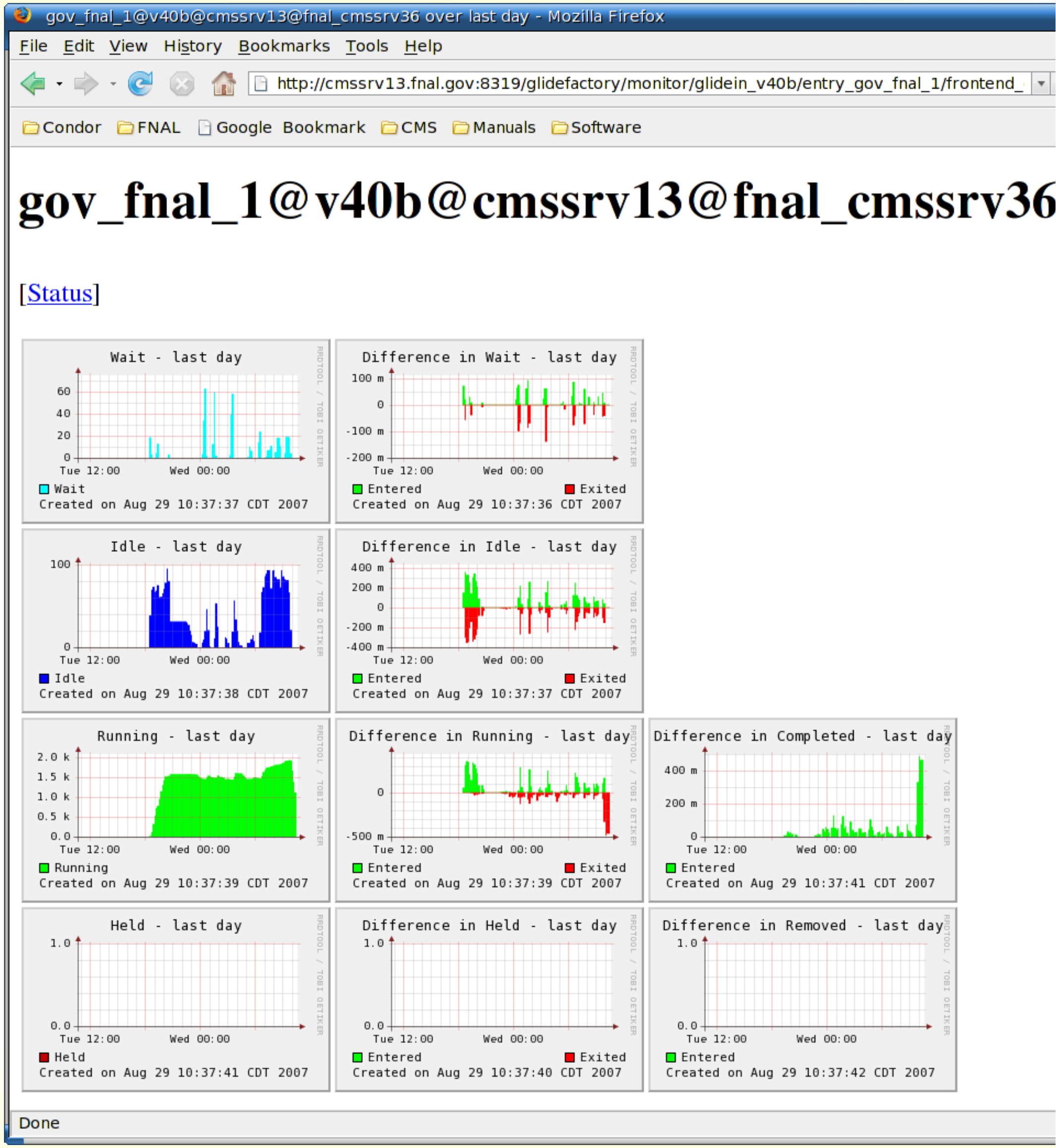
Factory Total over last hour

[medium | large]

[day | week | month | year]

[gov_fnal_1 | gov_fnal_2]

File Edit View History Bookmarks Tools Help
Google
Condor FNAL Google Bookmark CMS Manuals Software
Idle glideins - last hour
200
100
0
12:00 12:20 12:40
Requested
Idle
Created on Aug 29 12:53:19 CDT 2007
Idle glideins - last hour
gov_fnal_1
gov_fnal_2
Created on Aug 29 12:54:20 CDT 2007
Requested Idle glideins - last hour
Created on Aug 29 12:54:20 CDT 2007
Running glideins - last hour
3.0 k
2.0 k
1.0 k
0.0
Running
Created on Aug 29 12:53:20 CDT 2007
Running glideins - last hour
Created on Aug 29 12:54:20 CDT 2007
equested MaxRun glideins - last hou
20 k
15 k
10 k
5 k
0
Created on Aug 29 12:54:20 CDT 2007
Held glideins - last hour
1.0
0.0
Held
Created on Aug 29 12:54:19 CDT 2007
Held glideins - last hour
Created on Aug 29 12:54:20 CDT 2007
RRDTOOL / TOBI OETIKER
Done



gov_fnal_1@v40b@cmssrv13@fnal_cmssrv36 over last day - Mozilla Firefox

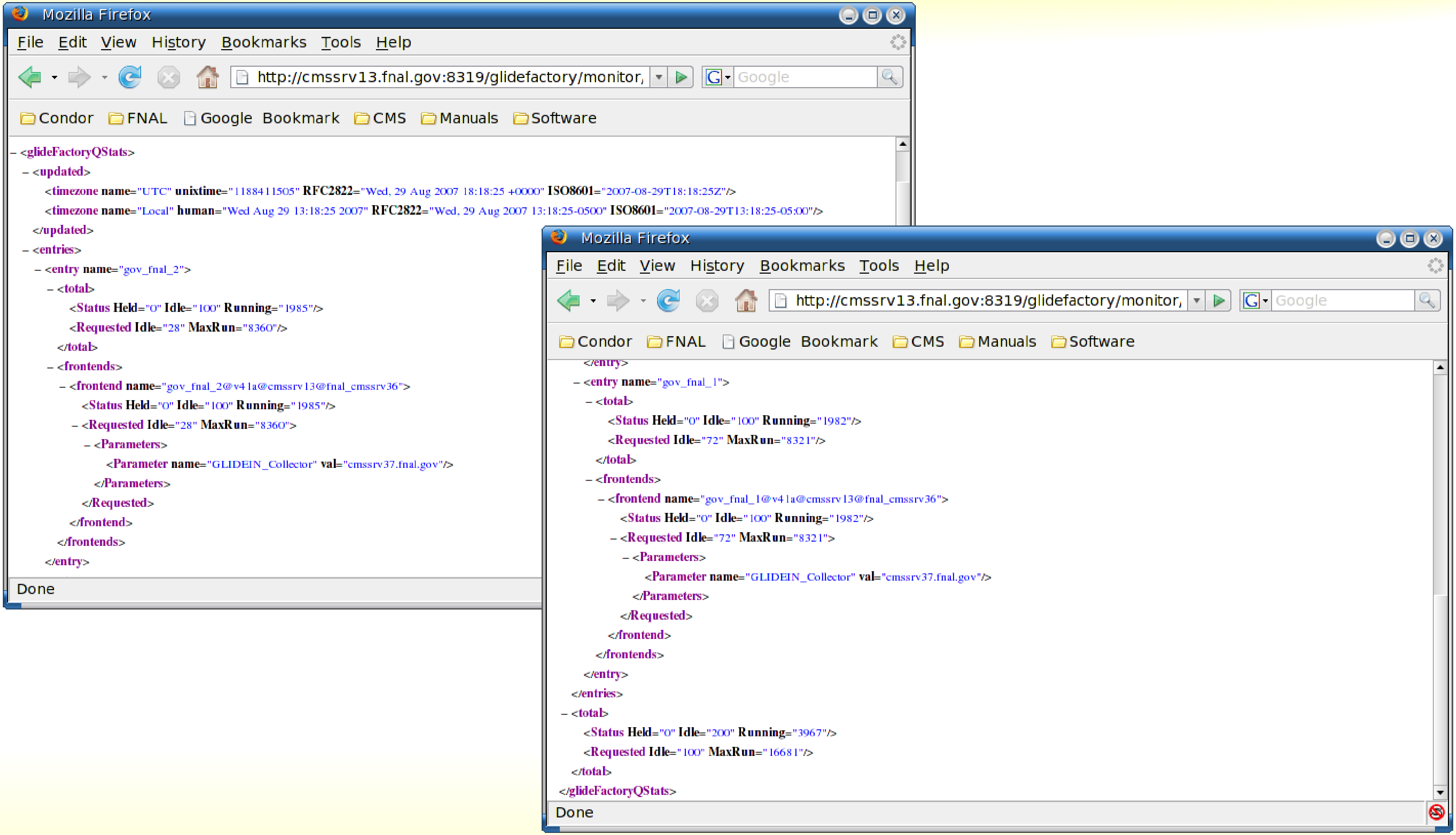
http://cmssrv13.fnal.gov:8319/glidefactory/monitor/glidein_v40b/entry_gov_fnal_1/frontend_

gov_fnal_1@v40b@cmssrv13@fnal_cmssrv36

[Status]

File Edit View History Bookmarks Tools Help
Condor FNAL Google Bookmark CMS Manuals Software
Wait - last day
60
40
20
0
Tue 12:00 Wed 00:00
Wait
Created on Aug 29 10:37:37 CDT 2007
Difference in Wait - last day
100 m
0
-100 m
-200 m
Entered
Exited
Created on Aug 29 10:37:36 CDT 2007
Idle - last day
100
0
Idle
Created on Aug 29 10:37:38 CDT 2007
Difference in Idle - last day
400 m
200 m
0
-200 m
-400 m
Created on Aug 29 10:37:37 CDT 2007
Running - last day
2.0 k
1.5 k
1.0 k
0.5 k
0.0
Running
Created on Aug 29 10:37:39 CDT 2007
Difference in Running - last day
0
-500 m
Created on Aug 29 10:37:39 CDT 2007
Difference in Completed - last day
400 m
200 m
0
Entered
Created on Aug 29 10:37:41 CDT 2007
Held - last day
1.0
0.0
Held
Created on Aug 29 10:37:41 CDT 2007
Difference in Held - last day
Created on Aug 29 10:37:40 CDT 2007
Difference in Removed - last day
Created on Aug 29 10:37:42 CDT 2007
RRDTOOL / TOBI OETIKER
Done

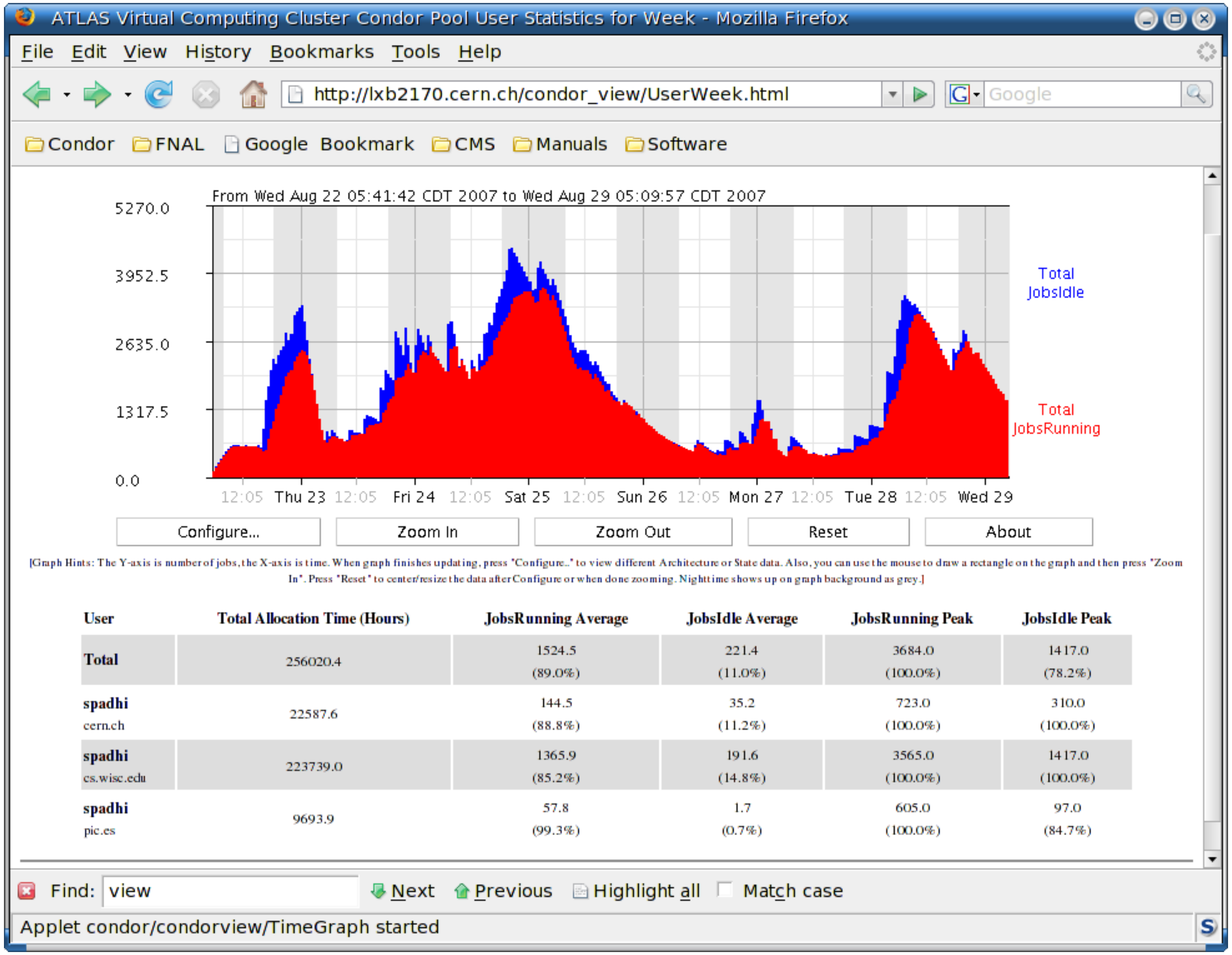
Status XML monitoring


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://cmssrv13.fnal.gov:8319/glidefactory/monitor,
Google
Condor FNAL Google Bookmark CMS Manuals Software
Done

```
– <glideFactoryQStats>
  – <updated>
      <timezone name="UTC" unixtime="1188411505" RFC2822="Wed, 29 Aug 2007 18:18:25 +0000" ISO8601="2007-08-29T18:18:25Z"/>
      <timezone name="Local" human="Wed Aug 29 13:18:25 2007" RFC2822="Wed, 29 Aug 2007 13:18:25-0500" ISO8601="2007-08-29T13:18:25-05:00"/>
    </updated>
  – <entries>
    – <entry name="gov_fnal_2">
      – <total>
          <Status Held="0" Idle="100" Running="1985"/>
          <Requested Idle="28" MaxRun="8360"/>
        </total>
      – <frontends>
        – <frontend name="gov_fnal_2@v41a@cmssrv13@fnal_cmssrv36">
            <Status Held="0" Idle="100" Running="1985"/>
          – <Requested Idle="28" MaxRun="8360">
            – <Parameters>
                <Parameter name="GLIDEIN_Collector" val="cmssrv37.fnal.gov"/>
              </Parameters>
            </Requested>
          </frontend>
        </frontends>
      </entry>
```

```
      </entry>
    – <entry name="gov_fnal_1">
      – <total>
          <Status Held="0" Idle="100" Running="1982"/>
          <Requested Idle="72" MaxRun="8321"/>
        </total>
      – <frontends>
        – <frontend name="gov_fnal_1@v41a@cmssrv13@fnal_cmssrv36">
            <Status Held="0" Idle="100" Running="1982"/>
          – <Requested Idle="72" MaxRun="8321">
            – <Parameters>
                <Parameter name="GLIDEIN_Collector" val="cmssrv37.fnal.gov"/>
              </Parameters>
            </Requested>
          </frontend>
        </frontends>
      </entry>
    </entries>
  – <total>
      <Status Held="0" Idle="200" Running="3967"/>
      <Requested Idle="100" MaxRun="16681"/>
    </total>
  </glideFactoryQStats>
```

CondorView Monitoring



ATLAS Virtual Computing Cluster Condor Pool User Statistics for Week - Mozilla Firefox

http://lxb2170.cern.ch/condor_view/UserWeek.html

[Graph Hints: The Y-axis is number of jobs, the X-axis is time. When graph finishes updating, press "Configure.." to view different Architecture or State data. Also, you can use the mouse to draw a rectangle on the graph and then press "Zoom In". Press "Reset" to center/resize the data after Configure or when done zooming. Nighttime shows up on graph background as grey.]

User	Total Allocation Time (Hours)	JobsRunning Average	JobsIdle Average	JobsRunning Peak	JobsIdle Peak
Total	256020.4	1524.5 (89.0%)	221.4 (11.0%)	3684.0 (100.0%)	1417.0 (78.2%)
spadhi cern.ch	22587.6	144.5 (88.8%)	35.2 (11.2%)	723.0 (100.0%)	310.0 (100.0%)
spadhi cs.wisc.edu	223739.0	1365.9 (85.2%)	191.6 (14.8%)	3565.0 (100.0%)	1417.0 (100.0%)
spadhi pic.es	9693.9	57.8 (99.3%)	1.7 (0.7%)	605.0 (100.0%)	97.0 (84.7%)

Find: view

Applet condor/condorview/TimeGraph started

Standard Condor tool, not glideinWMS specific

This one is actually from the CRONUS site

Conclusions

Condor Glideins

- Can shield user jobs from the Grid
- Give you total control over your jobs
- Allow you to have more control over the jobs scheduling

GlideinWMS

- An automatic way to create glidein pools on the fly
- Needs some initial effort, but then it operates on its own

glideinWMS home page

http://home.fnal.gov/~sfiligoi/glideinWMS/

sfiligoi@fnal.gov