CMS Monitoring at the LPC

Compact Muon Solenoid

LHC Physics Center

User job monitoring at the FNAL LPC,

from the support perspective

Marguerite Tonjes, LPC Computing Support

University of Illinois, Chicago

Senior Research Specialist

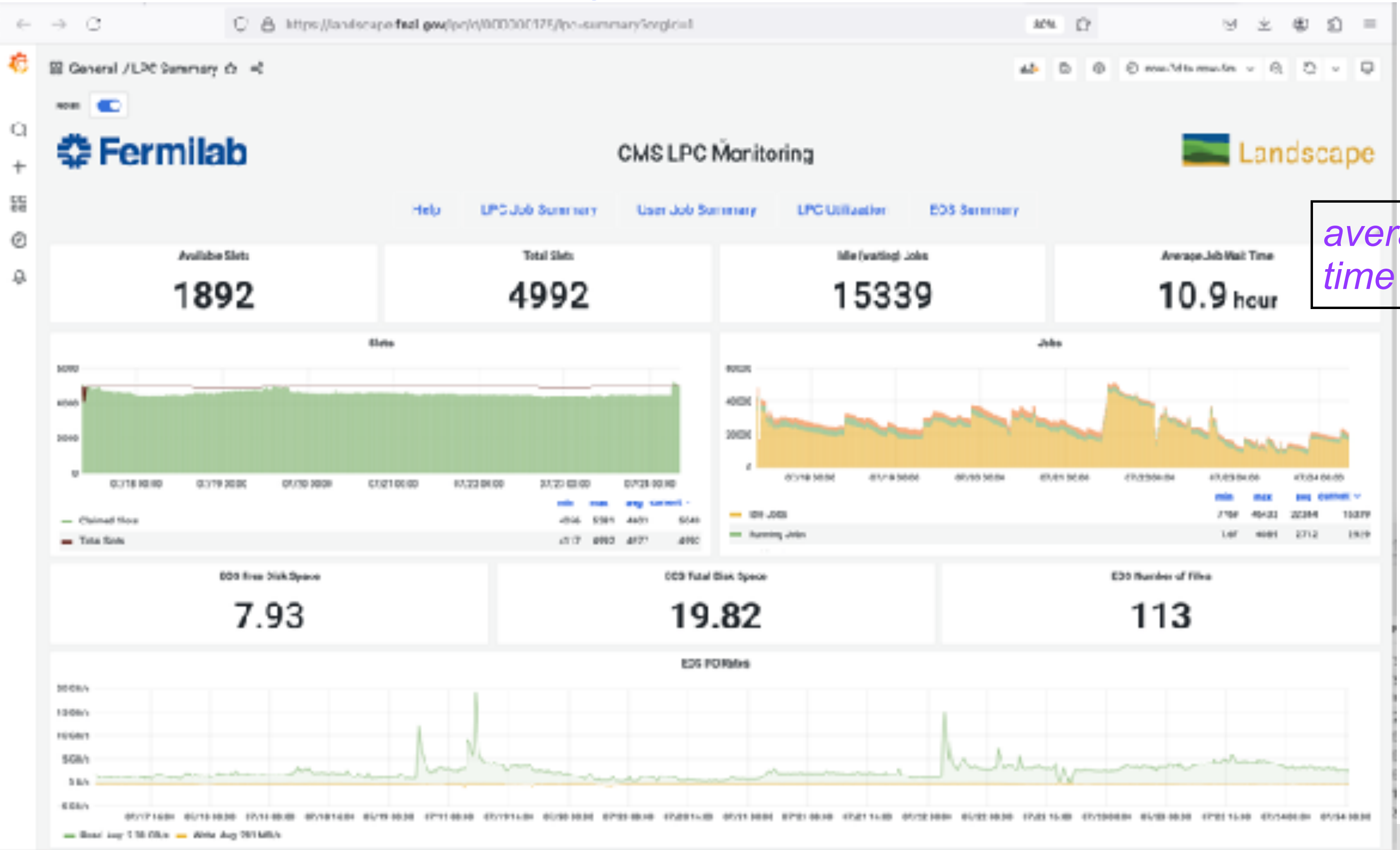
CMS LPC User jobs

- CMS as a whole has a fairly complex infrastructure, users often want to know:
  - Are my jobs running?
  - Why aren't they running?
  - Where and how did they fail?

- Jobs are submitted through local cmslpc interactive HTCondor, or through grid via CRAB (CMS Remote Analysis Builder) or via CMS Connect (HTCondor)
  - Note the CRAB and CMS Connect access the whole of the CMS computing grid (which includes opportunistic space on FNAL Tier1 as well as User Tier3
  - Only users with FNAL computing accounts access cmslpc (T3_US_FNALLPC - Tier3) HTCondor batch nodes from any source
  - CMS grid certificate used for grid job authentication (needed as well for most remote file reading), must also be stored in local database
- ***Scale***:
  - **723** unique users logged in interactively 2023-2024 *(.Xauthority)*
  - **480** unique users running cmslpc batch 2023-2024 *(HTCondor landscape)*

CMS LPC monitoring

- https://uscms.org/uscms_at_work/physics/computing/status/index.shtml collection of monitoring links useful to CMS LPC users at FNAL
- https://landscape.fnal.gov/lpc (authenticate CMS grid certificate)



General / LPC Summary

Fermilab

CMS LPC Monitoring

Landscape

Help | LPC Job Summary | User Job Summary | LPC Utilization | EOS Summary

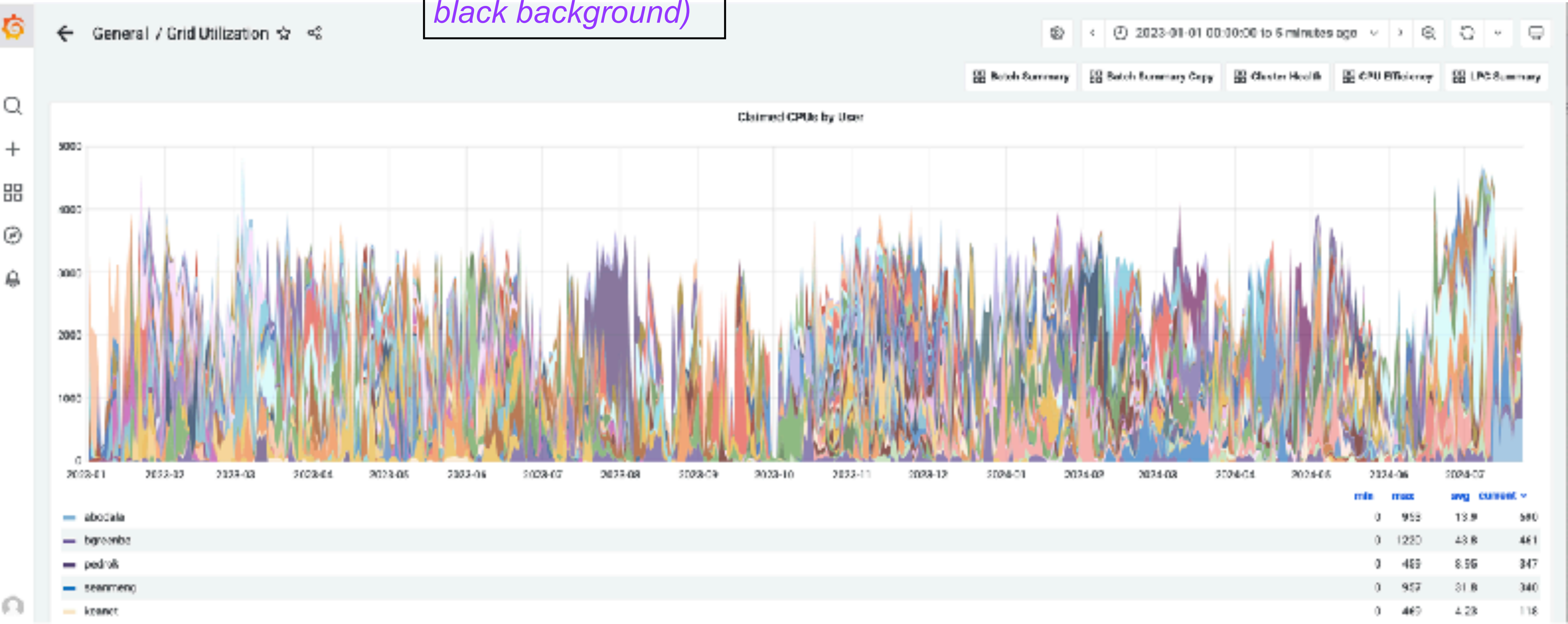
Available Slots	Total Slots	Idle (waiting) Jobs	Average Job Wait Time
1892	4992	15339	10.9 hour

EOS Free Disk Space	EOS Total Disk Space	EOS Number of Files
7.93	19.82	113

average job wait time can be large!

Claimed CPUs by User

(note - normally black background)



	min	max	avg	current
abodala	0	953	13.9	580
bgreenbe	0	1220	43.8	461
pedrolk	0	469	8.56	347
seanmeng	0	957	31.8	340
kramer	0	469	4.28	118

- From "LPC Utilization"
  - What we show management (US CMS group leads; funding)
  - *Dip due to a bug in schedulers fixed Monday*

Are my jobs running?

- cmslpc HTCondor or CMS connect: many users just use command line checks like `condor_q`
  - https://uscms.org/uscms_at_work/computing/setup/batch_troubleshoot.shtml#Troubleshooting describes useful troubleshooting techniques for cmslpc
  - Unfortunately recent HTCondor software update removed CPU Time and Memory reporting
  - Landscape User Batch Summary (example)
- CRAB command line: `crab status`

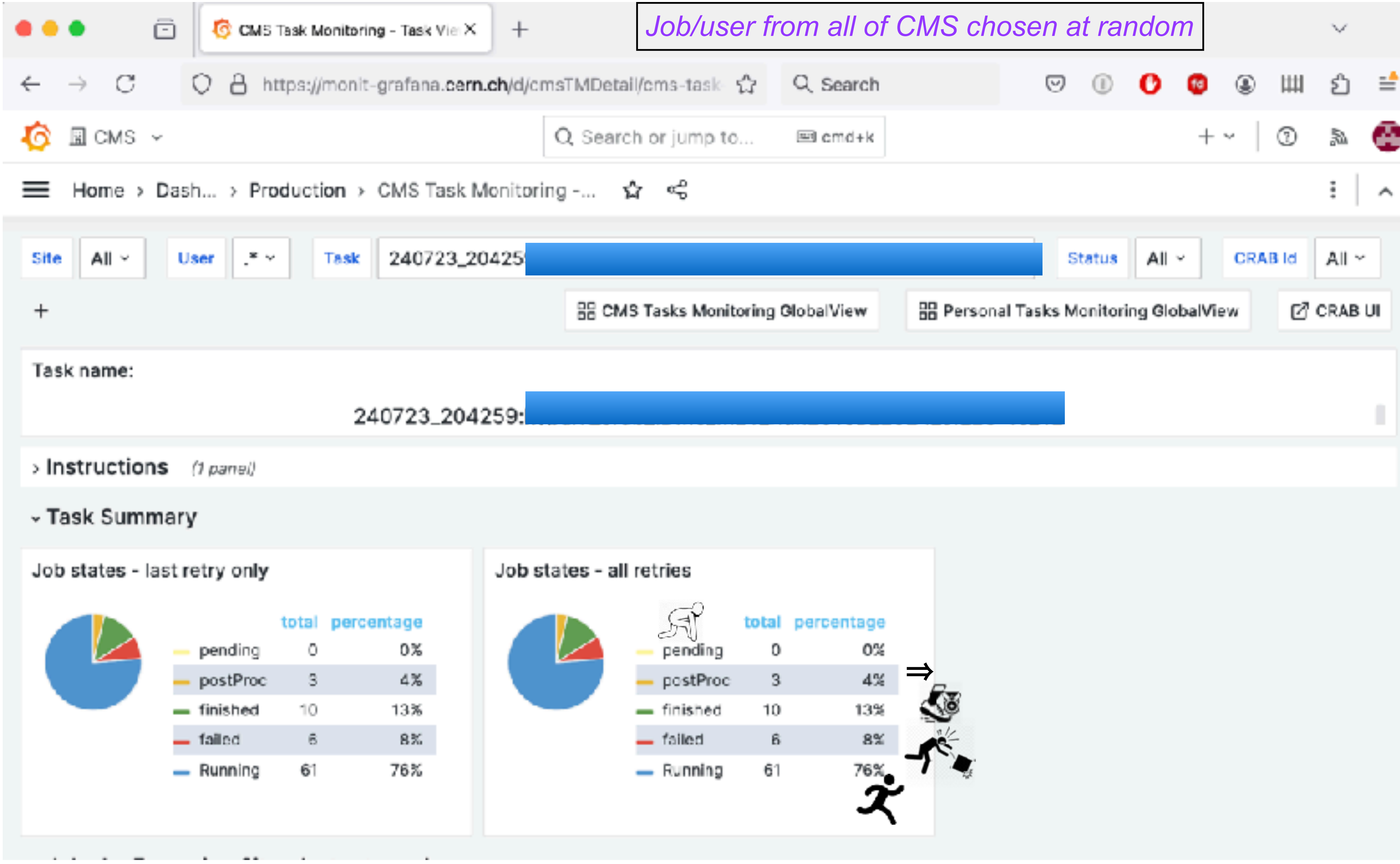
```
crab status -d crabsubmit/crab_cmsdas_minbias_test0
CRAB project directory:     /uscms_data/d3/username/cmsdas/CMSSW_13_0_13_mcgen/src/crabsubmit/
crab_cmsdas_minbias_test0
Task name:             231110_212908:username_crab_cmsdas_minbias_test0
Grid scheduler - Task Worker:   crab3@vocms0198.cern.ch - crab-prod-tw01
Status on the CRAB server:  SUBMITTED
Task URL to use for HELP:   https://cmsweb.cern.ch/crabserver/ui/task/
231110_212908%3Ausername_crab_cmsdas_minbias_test0
Dashboard monitoring URL:   https://monit-grafana.cern.ch/d/cmsTMDetail/cms-task-monitoring-task-view?
orgId=11&var-user=username&var-task=231110_212908%3Ausername_crab_cmsdas_minbias_test0&from=1699648148000&to=now
Status on the scheduler:    SUBMITTED

Jobs status:                     idle                100.0% (10/10)

No publication information available yet
Log file is /uscms_data/d3/username/cmsdas/CMSSW_13_0_13_mcgen/src/crabsubmit/crab_cmsdas_minbias_test0/crab.log
```

not yet running

Example of CRAB dashboard



Job/user from all of CMS chosen at random

CMS Task Monitoring - Task Vie
https://monit-grafana.cern.ch/d/cmsTMDetail/cms-task
Search
CMS
Search or jump to...
cmd+k
Home › Dash... › Production › CMS Task Monitoring -...
CMS Tasks Monitoring GlobalView
Personal Tasks Monitoring GlobalView
CRAB UI
⇒

Site: All | User: .* | Task: 240723_20425[redacted] | Status: All | CRAB Id: All

Task name:

240723_204259:[redacted]

› Instructions (1 panel)

⌄ Task Summary

Job states - last retry only

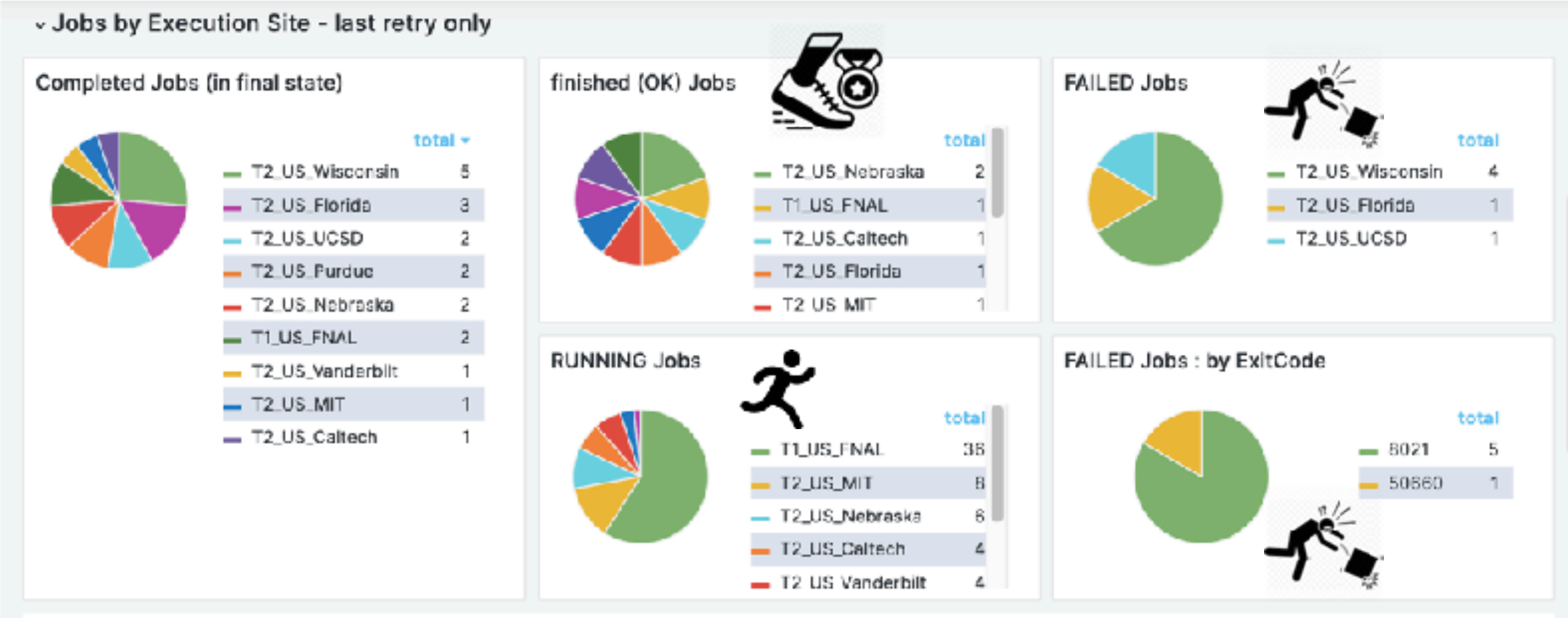
	total	percentage
🟨 pending	0	0%
🟧 postProc	3	4%
🟩 finished	10	13%
🟥 failed	6	8%
🟦 Running	61	76%

Job states - all retries

	total	percentage
🟨 pending	0	0%
🟧 postProc	3	4%
🟩 finished	10	13%
🟥 failed	6	8%
🟦 Running	61	76%

CRAB user dashboard 2

- Depending on the error code, CRAB will automatically retry jobs. This particular analysis went to US sites.



⌄ Jobs by Execution Site - last retry only

Completed Jobs (in final state)

	total ▾
T2_US_Wisconsin	5
T2_US_Florida	3
T2_US_UCSD	2
T2_US_Purdue	2
T2_US_Nebraska	2
T1_US_FNAL	2
T2_US_Vanderbilt	1
T2_US_MIT	1
T2_US_Caltech	1

finished (OK) Jobs

	total
T2_US_Nebraska	2
T1_US_FNAL	1
T2_US_Caltech	1
T2_US_Florida	1
T2_US_MIT	1

FAILED Jobs

	total
T2_US_Wisconsin	4
T2_US_Florida	1
T2_US_UCSD	1

RUNNING Jobs

	total
T1_US_FNAL	36
T2_US_MIT	8
T2_US_Nebraska	6
T2_US_Caltech	4
T2_US_Vanderbilt	4

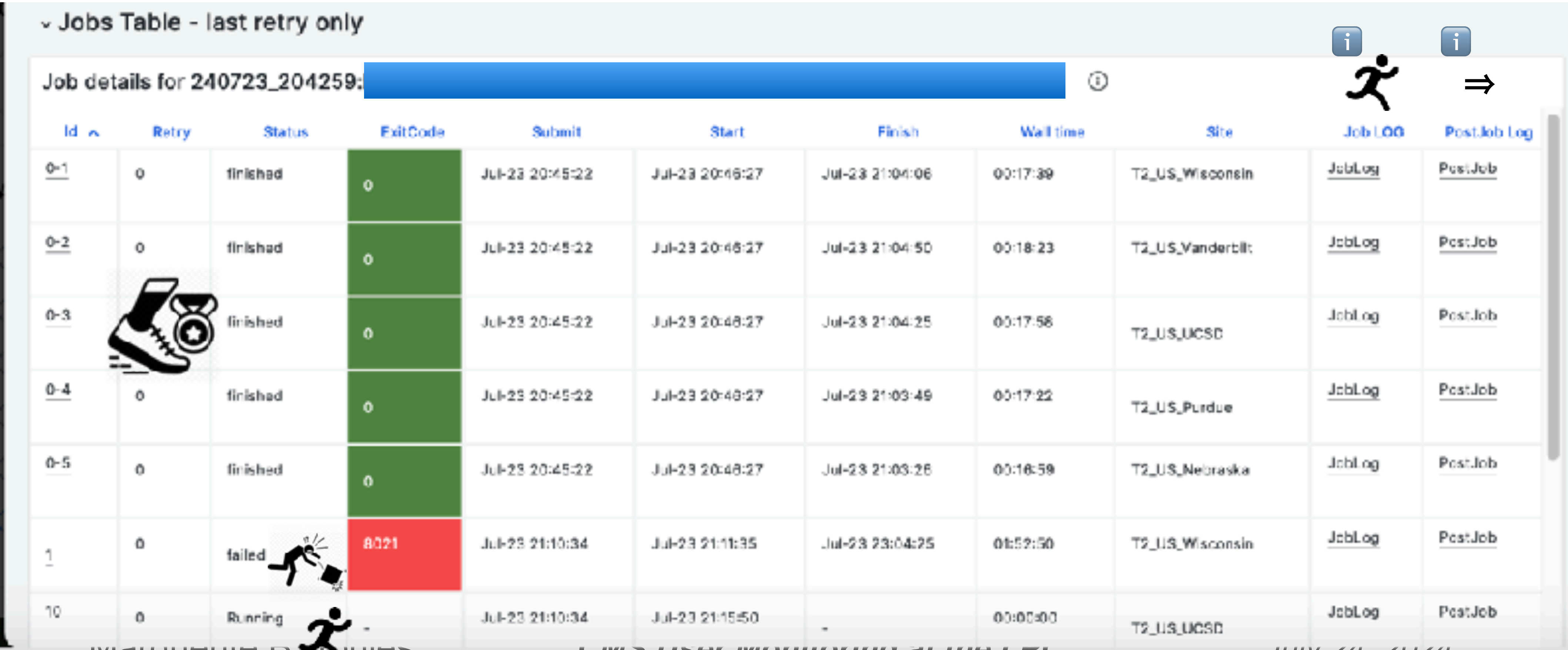
FAILED Jobs : by ExitCode

	total
8021	5
50660	1

- Where did the job process?

CRAB user dashboard list

- User can click on JobLog to look at analysis stdout/stderr. ExitCodes have a reference *(8021: file read error)*
- PostJob is the transfer: handled centrally by FTS *(File Transfer Service)* from a temporary location on the remote processing site to the final file destination

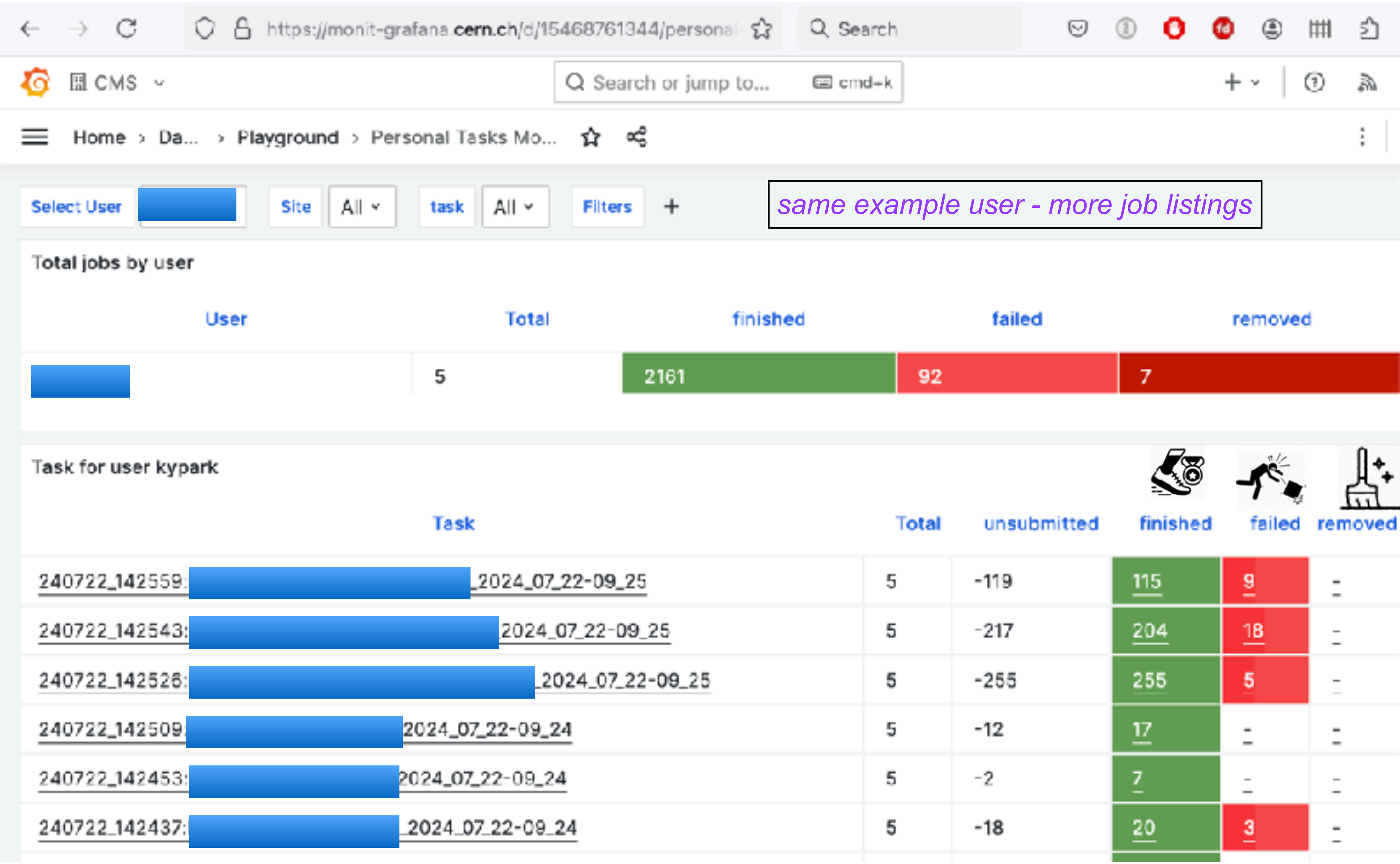


⌄ Jobs Table - last retry only

Job details for 240723_204259:[redacted]

Id ⌃	Retry	Status	ExitCode	Submit	Start	Finish	Wall time	Site	Job LOG	PostJob Log
0-1	0	finished	0	Jul-23 20:45:22	Jul-23 20:46:27	Jul-23 21:04:06	00:17:39	T2_US_Wisconsin	JobLog	PostJob
0-2	0	finished	0	Jul-23 20:45:22	Jul-23 20:46:27	Jul-23 21:04:50	00:18:23	T2_US_Vanderbilt	JobLog	PostJob
0-3	0	finished	0	Jul-23 20:45:22	Jul-23 20:46:27	Jul-23 21:04:25	00:17:58	T2_US_UCSD	JobLog	PostJob
0-4	0	finished	0	Jul-23 20:45:22	Jul-23 20:46:27	Jul-23 21:03:49	00:17:22	T2_US_Purdue	JobLog	PostJob
0-5	0	finished	0	Jul-23 20:45:22	Jul-23 20:46:27	Jul-23 21:03:26	00:16:59	T2_US_Nebraska	JobLog	PostJob
1	0	failed	8021	Jul-23 21:10:34	Jul-23 21:11:35	Jul-23 23:04:25	01:52:50	T2_US_Wisconsin	JobLog	PostJob
10	0	Running	-	Jul-23 21:10:34	Jul-23 21:15:50	-	00:00:00	T2_US_UCSD	JobLog	PostJob

Overall CRAB dashboard for example



https://monit-grafana.cern.ch/d/15468761344/personal
Search
CMS
Search or jump to...
cmd+k
Home › Da... › Playground › Personal Tasks Mo...
Select User
Site
All
task
All
Filters
+

same example user - more job listings

Select User [redacted]

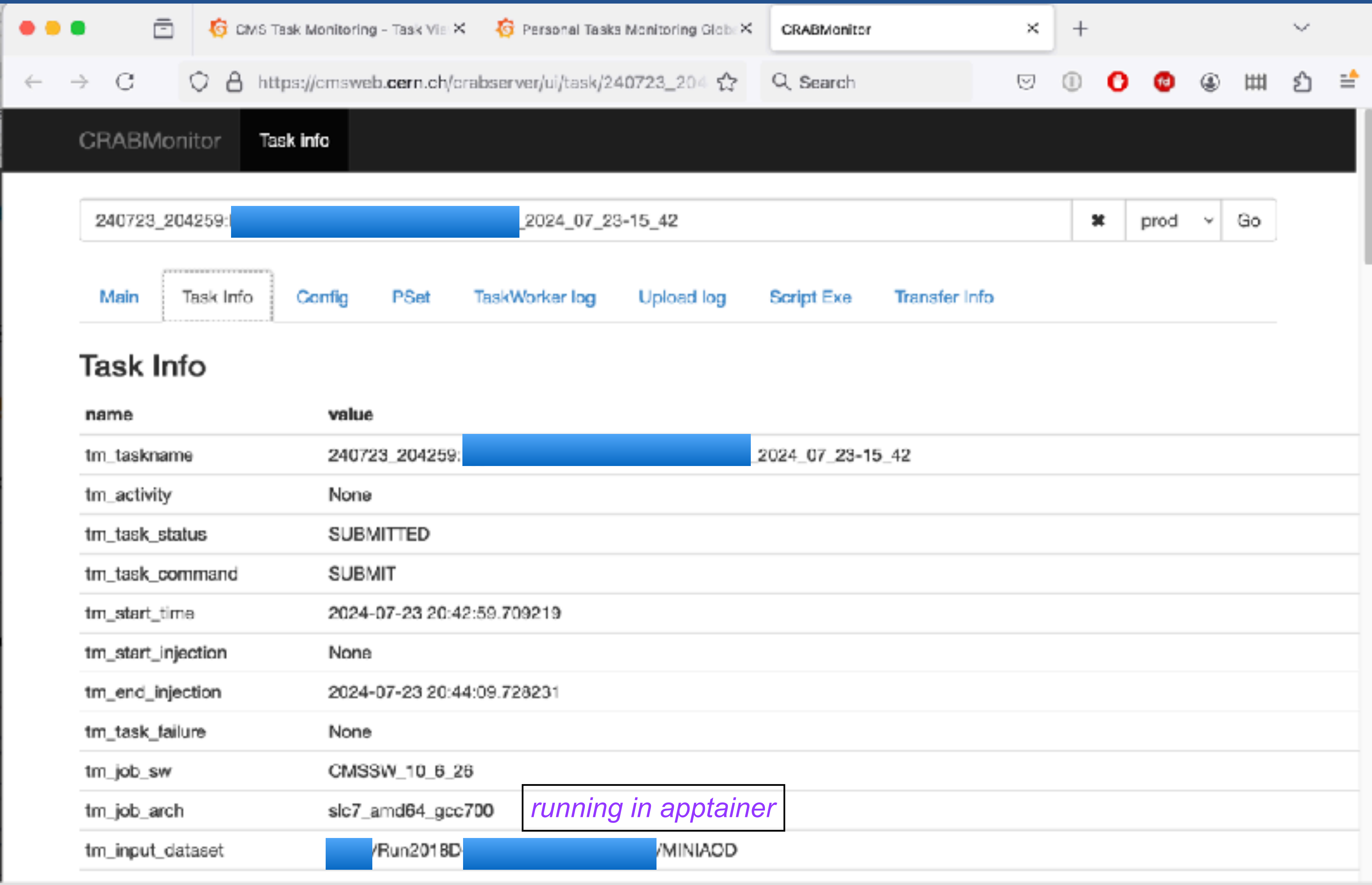
Total jobs by user

User	Total	finished	failed	removed
[redacted]	5	2161	92	7

Task for user kypark

Task	Total	unsubmitted	finished	failed	removed
240722_142559:[redacted]_2024_07_22-09_25	5	-119	115	9	-
240722_142543:[redacted]2024_07_22-09_25	5	-217	204	18	-
240722_142526:[redacted]_2024_07_22-09_25	5	-255	255	5	-
240722_142509[redacted]2024_07_22-09_24	5	-12	17	-	-
240722_142453:[redacted]2024_07_22-09_24	5	-2	7	-	-
240722_142437:[redacted]_2024_07_22-09_24	5	-18	20	3	-

CRAB UI (check job specifics)



240723_204259:[redacted]_2024_07_23-15_42

Task Info

name	value
tm_taskname	240723_204259:[redacted]_2024_07_23-15_42
tm_activity	None
tm_task_status	SUBMITTED
tm_task_command	SUBMIT
tm_start_time	2024-07-23 20:42:59.709219
tm_start_injection	None
tm_end_injection	2024-07-23 20:44:09.728231
tm_task_failure	None
tm_job_sw	CMSSW_10_6_26
tm_job_arch	slc7_amd64_gcc700
tm_input_dataset	[redacted]/Run2018D-[redacted]/MINIAOD

running in apptainer

Why have my jobs failed HTCondor?

- Failed to run or failed to finish properly?

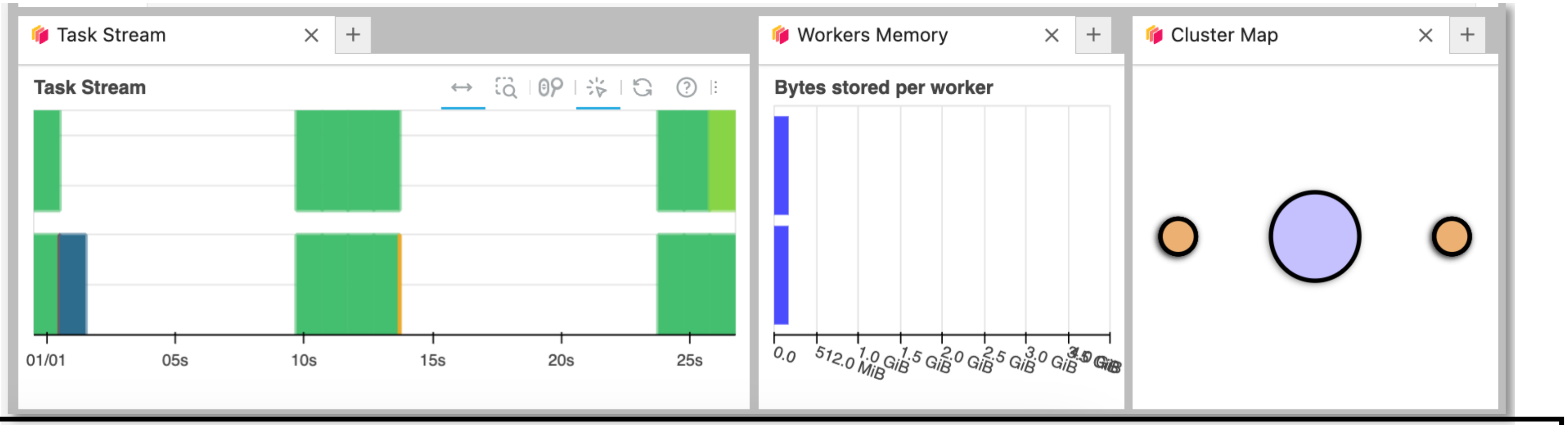
  - So many possibilities:
    - Asked for too much Memory per job and few groups of slots on machines available
    - Crashed in remote dataset
    - Went over limits (48 hours; 40GB local space, etc.)
    - Failed to transfer output (they're over quota; rarer disk error at the LPC; problem with FTS/transfer from grid site)
    - Failed to read some/all files (problem at site; user(s) read too many times in parallel same file and overloaded network; parsed input wrong; file on tape and not disk, ...)
    - User error
    - etc...
  - Note that CRAB produces error codes for various cases which can be used to debug many problems *(not full number range for each)*
    - 1 - 512 (Unix); 7000 - 9000 (CMSSW exit codes); 10000 - 19999 (environment setup); 50000 - 59999 (executable); 60000 - 69999 (staging out); 70000 - 79999 (WMAgent - job transfer); 80000-99999 (CRAB and other: only 6)

HTCondor troubleshooting

- Typically we tell people to check `condor_q -better-analyze` and `.stdout`/`.stderr`
  - General recommendation to put the following in .stdout:

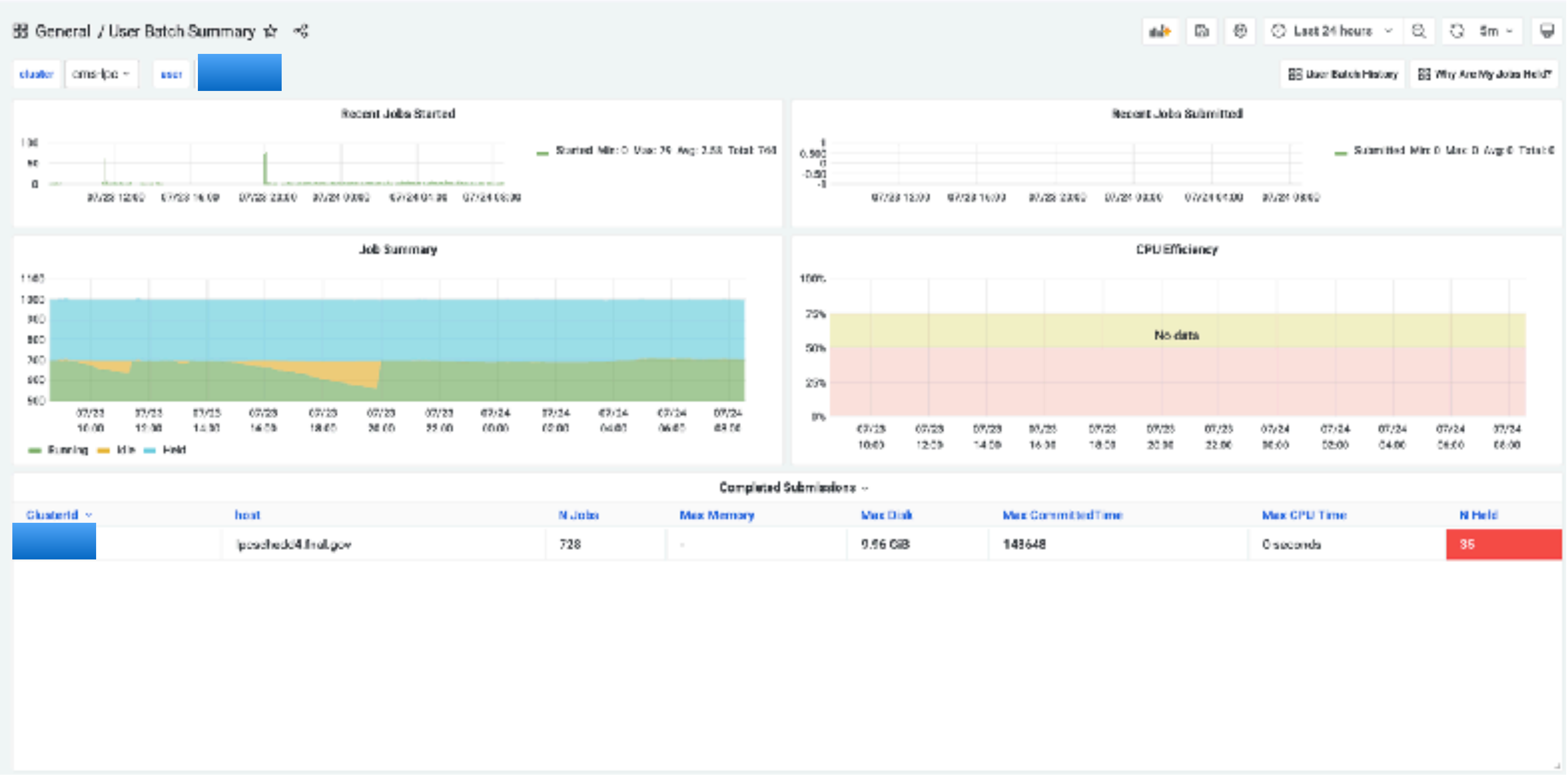
```
echo "Starting job on " `date` #Date/time of start of job
echo "Running on: `uname -a`" #Condor job is running on this node
echo "System software: `cat /etc/redhat-release`" #Operating System job is running in
```

- Some users developed their own job management scripts
- The coffea (columnar analysis) workflow has its own job monitoring for dask (coffea-casa; Purdue)



example from: Dmitry Kondratyev (Purdue University (US)) (from LPC tutorial on dask)

CMS LPC Landscape User job



cluster cms-lpc

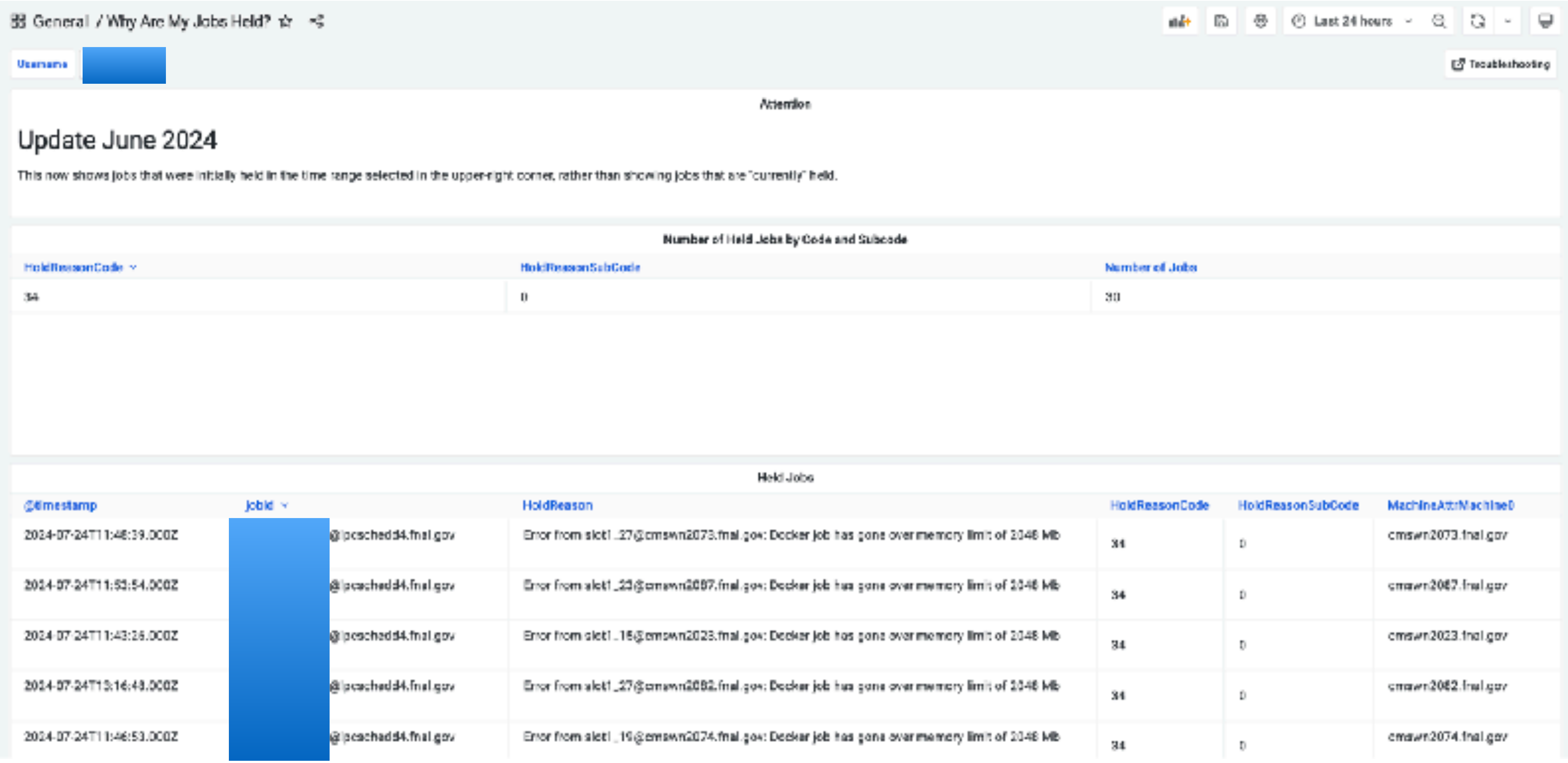
user [redacted]

Completed Submissions

ClusterId	host	N Jobs	Max Memory	Max Disk	Max CommittedTime	Max CPU Time	N Held
[redacted]	lpcschedd4.fnal.gov	728	-	9.56 GiB	143648	0 seconds	35

memory and CPU time missing: HTCondor bug :(

Too much memory



[redacted]

Attention

Update June 2024

This now shows jobs that were initially held in the time range selected in the upper-right corner, rather than showing jobs that are "currently" held.

Number of Held Jobs by Code and Subcode

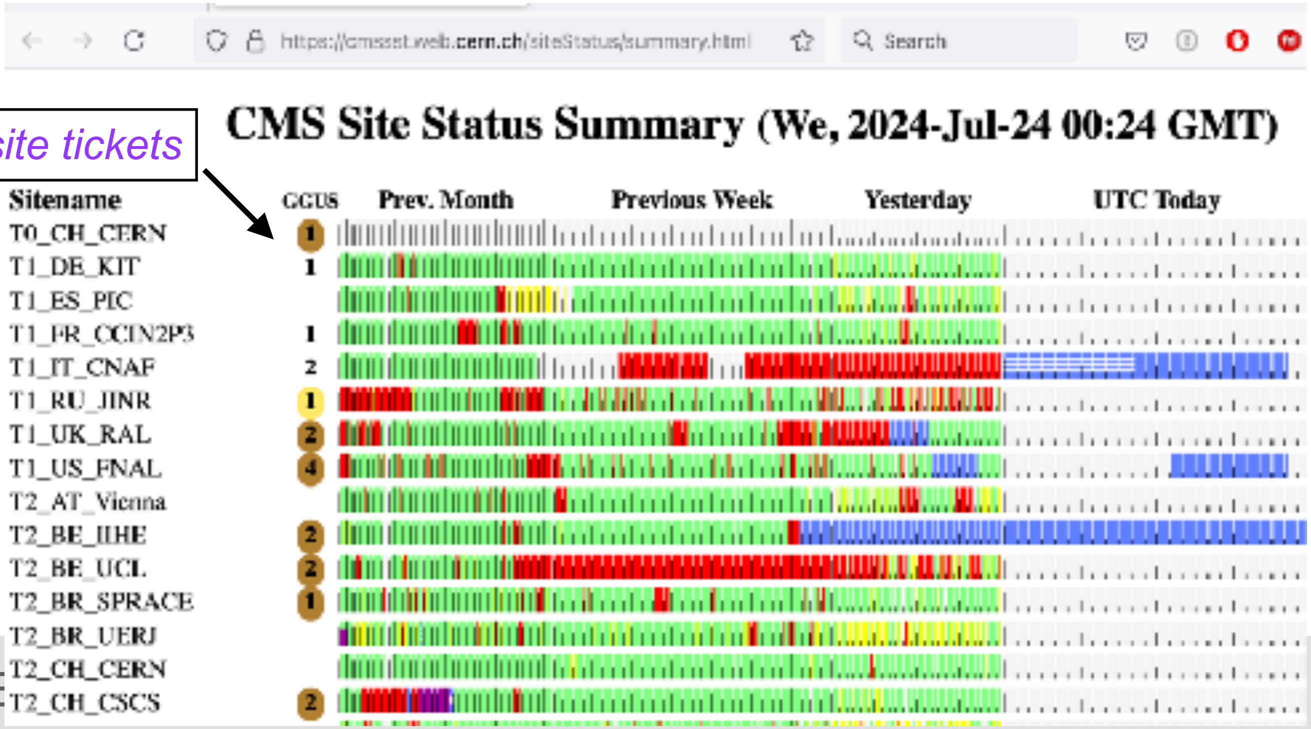
HoldReasonCode	HoldReasonSubCode	Number of Jobs
34	0	30

Held Jobs

@timestamp	JobId	HoldReason	HoldReasonCode	HoldReasonSubCode	MachineAttrMachine0
2024-07-24T11:48:39.000Z	[redacted]@lpcschedd4.fnal.gov	Error from slot1_27@cmswn2073.fnal.gov: Docker job has gone over memory limit of 2048 Mb	34	0	cmswn2073.fnal.gov
2024-07-24T11:53:54.000Z	[redacted]@lpcschedd4.fnal.gov	Error from slot1_23@cmswn2087.fnal.gov: Docker job has gone over memory limit of 2048 Mb	34	0	cmswn2087.fnal.gov
2024-07-24T11:43:26.000Z	[redacted]@lpcschedd4.fnal.gov	Error from slot1_15@cmswn2023.fnal.gov: Docker job has gone over memory limit of 2048 Mb	34	0	cmswn2023.fnal.gov
2024-07-24T13:16:48.000Z	[redacted]@lpcschedd4.fnal.gov	Error from slot1_27@cmswn2082.fnal.gov: Docker job has gone over memory limit of 2048 Mb	34	0	cmswn2082.fnal.gov
2024-07-24T11:46:53.000Z	[redacted]@lpcschedd4.fnal.gov	Error from slot1_19@cmswn2074.fnal.gov: Docker job has gone over memory limit of 2048 Mb	34	0	cmswn2074.fnal.gov

Frequently: Cannot Access Data

- Note many users use xrootd: which gives **A**ny Data, **A**ny Time, **A**nywhere
  - Important factor is that software needs fallback built in (CMSSW)
- We have a Data Aggregation Service (web based and command line) to locate data/Monte Carlo and understand what and where it is and if it's valid
- Can check the [CMS Site Status Board]() to check outages (sites are clickable) - also tells you status of remote compute resources



https://cmssst.web.cern.ch/siteStatus/summary.html

also number of site tickets

CMS Site Status Summary (We, 2024-Jul-24 00:24 GMT)

Sitename	GGUS	Prev. Month	Previous Week	Yesterday	UTC Today
T0_CH_CERN	1				
T1_DE_KIT	1				
T1_ES_PIC					
T1_FR_CCIN2P3	1				
T1_IT_CNAF	2				
T1_RU_JINR	1				
T1_UK_RAL	2				
T1_US_FNAL	4				
T2_AT_Vienna					
T2_BE_IIHE	2				
T2_BE_UCL	2				
T2_BR_SPRACE	1				
T2_BR_UERJ					
T2_CH_CERN					
T2_CH_CSCS	2				

Disk & node health

- I find users don't show me they have checked disk/node health in troubleshooting
  - They like to open tickets and blame sites
  - Maybe I solve too much too quickly so they don't check themselves - also a lot of things are very clearly documented, this is a ***big help*** not to be underestimated

- FNAL EOS has a landscape link
- Only if one of the 50 (Alma8) or 50 (Alma9) interactive nodes is problematic do users check SSI metrics (requires FNAL VPN or onsite fgz)
- Worker nodes have been so stable and monitored well I haven't checked their status to support tickets in a couple years! Same with NFS *(users don't have access to those checks)*
- Quota: command line (`quota -s`; `eosquota`)

Other monitoring

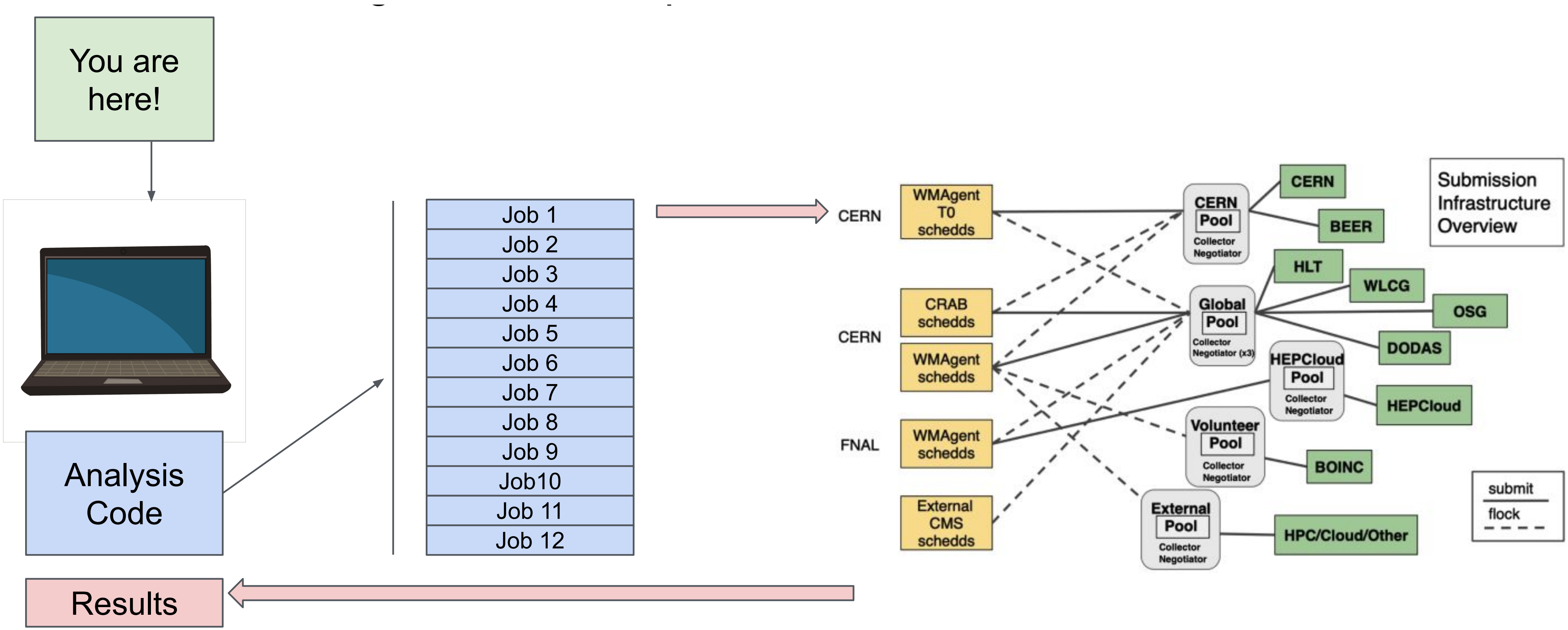
- We have a community mailing list as well as community chat (LPC gethelp web page) and people will frequently just ask "this thing broke with this error, is FNAL EOS disk working?"
  - No seriously, a high number of "yes I also have this error" shows a problem, just like a high number of tickets shows it
  - This has led to tickets (strongly encouraged)
  - This has led to community instruction on how to access data, fixing problems of misunderstanding old analysis recipes, etc.
- CMS LPC emails a user list for outages and I also post updates on the chat (*a lot of emails were @fnal.gov for users that dropped off when forwarding stopped*)

Conclusion/Questions?

- I only focused on user monitoring and did not cover a lot more of CMS central production
- There are many more tools available to understand all of these systems (CRAB servers; database servers; xrootd file servers; FTS servers; etc...) not covered here

most icons from IconScout.com

Backup slide 1: CMS grid workflow



You are here!

Analysis Code

Job 1

Job 2

Job 3

Job 4

Job 5

Job 6

Job 7

Job 8

Job 9

Job10

Job 11

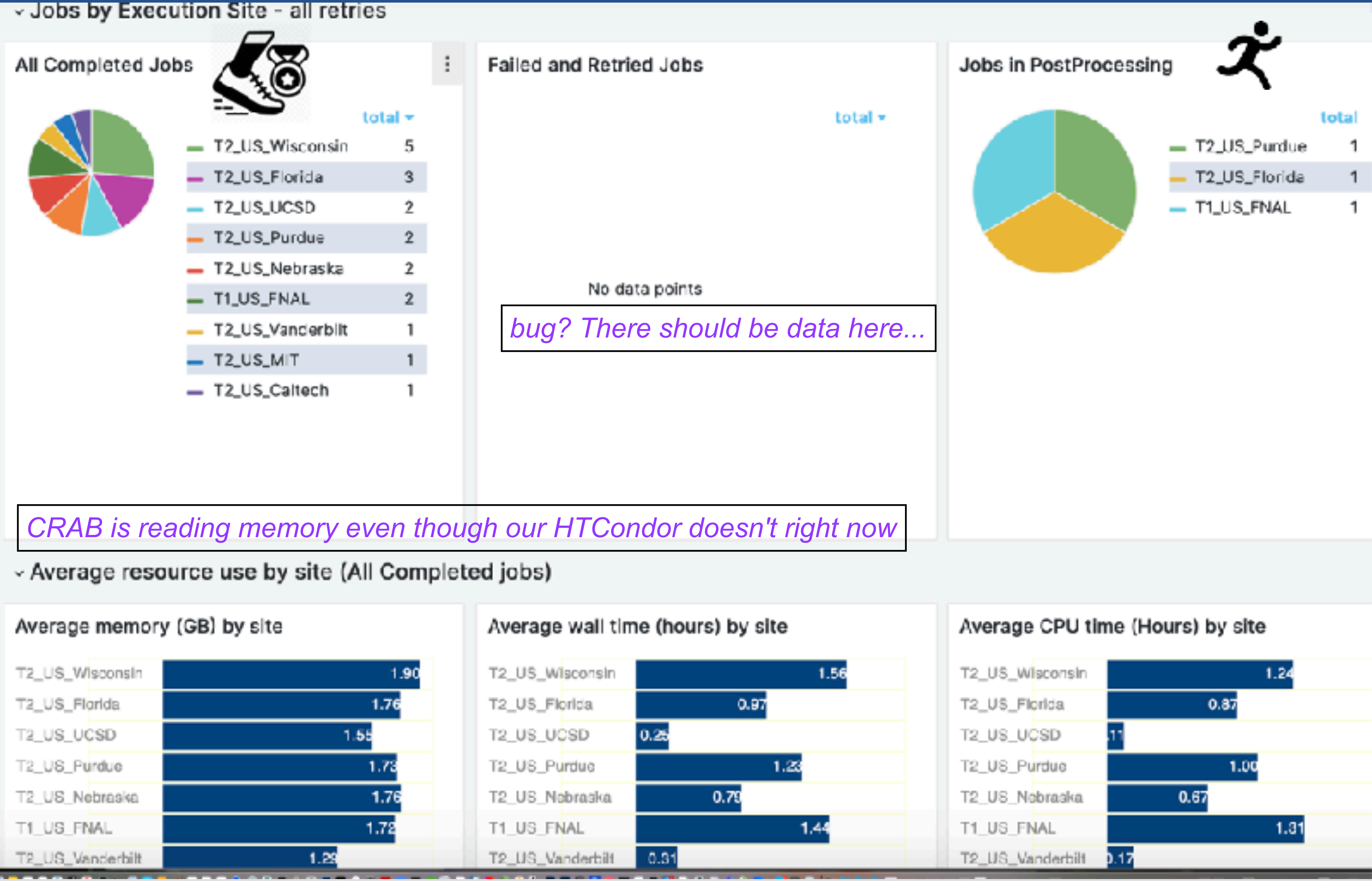
Job 12

Results

CERN
WMAgent T0 schedds
CERN
CRAB schedds
WMAgent schedds
FNAL
WMAgent schedds
External CMS schedds
CERN Pool Collector Negotiator
Global Pool Collector Negotiator (x3)
HEPCloud Pool Collector Negotiator
Volunteer Pool Collector Negotiator
External Pool Collector Negotiator
CERN
BEER
HLT
WLCG
OSG
DODAS
HEPCloud
BOINC
HPC/Cloud/Other
Submission Infrastructure Overview
submit
flock

taken from Andrew Melo's presentation to USCMS PURSUE interns 2024

CRAB user dashboard 3



⌄ Jobs by Execution Site - all retries

All Completed Jobs

total ▾

T2_US_Wisconsin	5
T2_US_Florida	3
T2_US_UCSD	2
T2_US_Purdue	2
T2_US_Nebraska	2
T1_US_FNAL	2
T2_US_Vanderbilt	1
T2_US_MIT	1
T2_US_Caltech	1

Failed and Retried Jobs

total ▾

No data points

bug? There should be data here...

Jobs in PostProcessing

total

T2_US_Purdue	1
T2_US_Florida	1
T1_US_FNAL	1

CRAB is reading memory even though our HTCondor doesn't right now

⌄ Average resource use by site (All Completed jobs)

Average memory (GB) by site

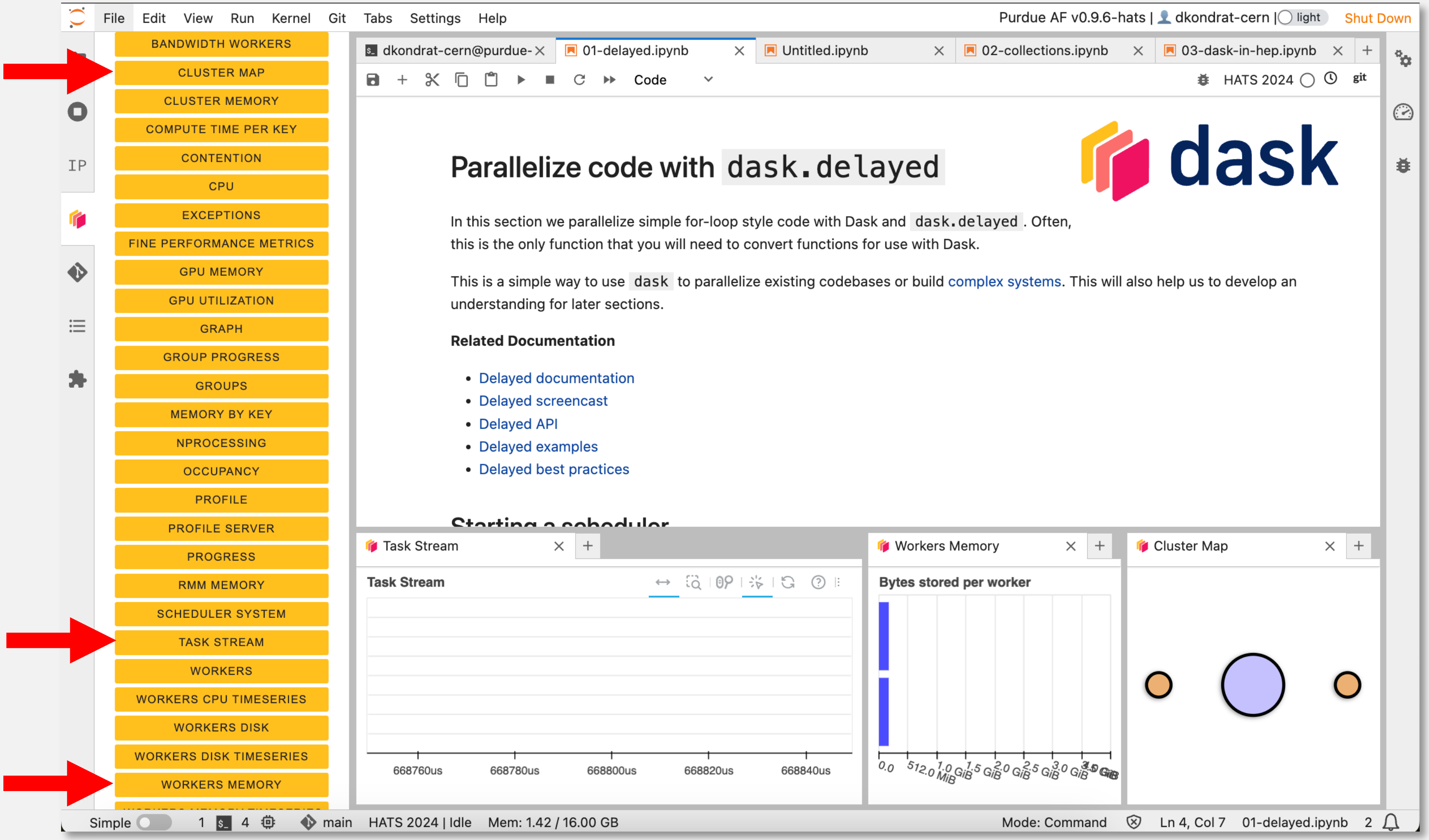
Average wall time (hours) by site

Average CPU time (Hours) by site

Example of coffea dask monitoring

- At Purdue Tier2, for batch jobs: Dmitry Kondratyev (Purdue University (US)) (from LPC tutorial on dask)

- You can drag-and-drop panels to place them side by side with other tabs.



Parallelize code with `dask.delayed`

In this section we parallelize simple for-loop style code with Dask and `dask.delayed`. Often, this is the only function that you will need to convert functions for use with Dask.

This is a simple way to use `dask` to parallelize existing codebases or build complex systems. This will also help us to develop an understanding for later sections.

Related Documentation

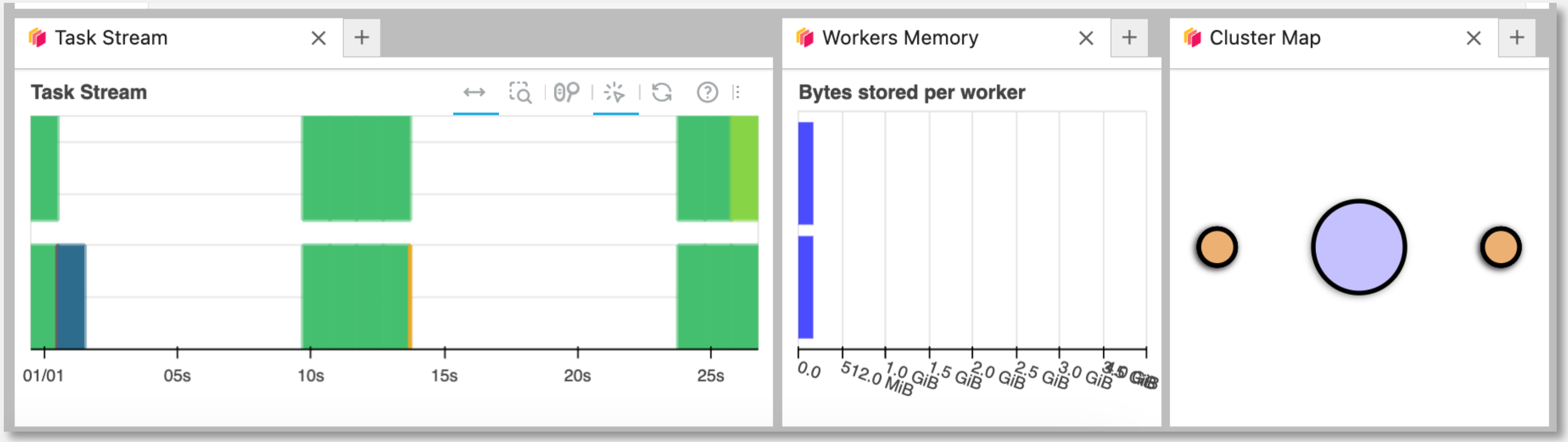
- Delayed documentation
- Delayed screencast
- Delayed API
- Delayed examples
- Delayed best practices

Starting a scheduler

LPC Coffea Dask tutorial

- At Purdue Tier2, for batch jobs: Dmitry Kondratyev (Purdue University (US)) (from LPC tutorial on dask)

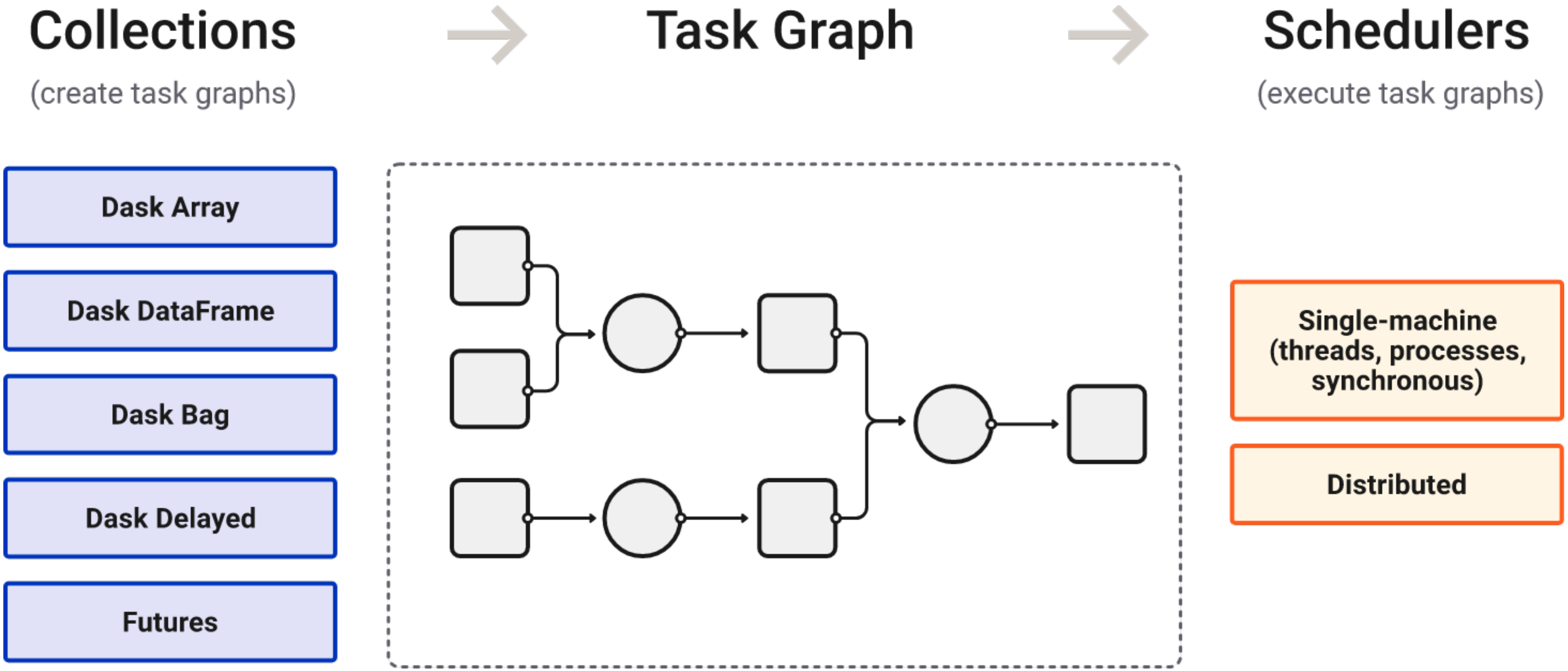
- Monitor the execution in the dashboard panels:



- **These tasks now run in a distributed mode on remote machines!**

So what does that set of words really mean?

Coffea Dask



Collections → Task Graph → Schedulers

(create task graphs)

(execute task graphs)

Dask Array

Dask DataFrame

Dask Bag

Dask Delayed

Futures

Single-machine (threads, processes, synchronous)

Distributed

- You use collections to write straightforward python
- That code generates an abstract, declarative, description of your analysis
  - It can then be executed by anything that implements the collection's array interface!
  - This makes analysis code extremely portable for tradeoff in underlying complexity
- I hope to dig into this complexity enough so you can reason about task graphs

Columnar Analysis

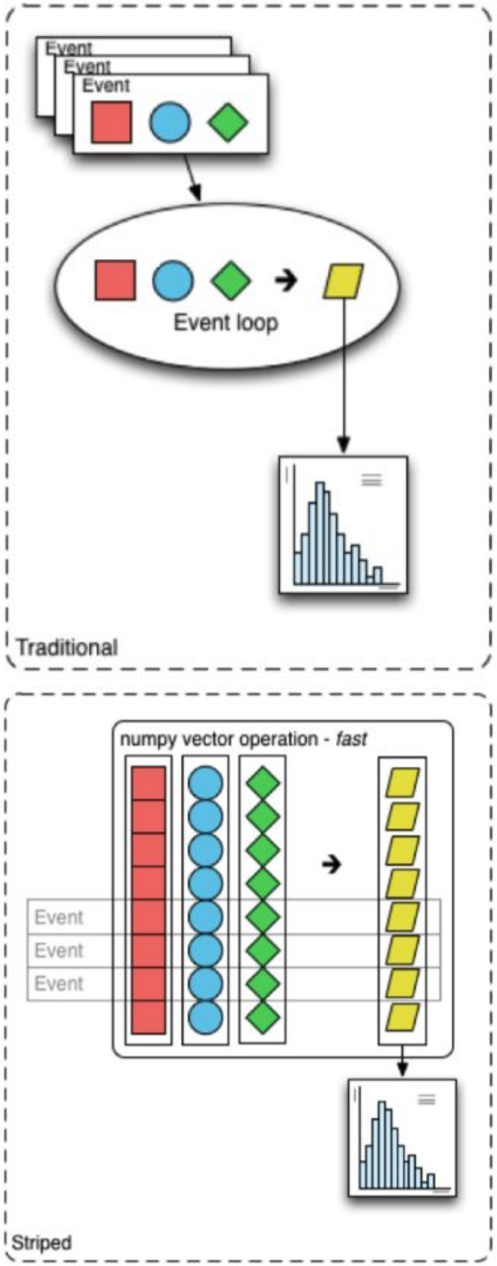
In a “traditional” analysis, each event is processed one-at-a-time, meaning:

- An event is loaded and the appropriate values are extracted
- Some computation happens over this single event
- The temporary space is cleaned and the next event is loaded

In a “columnar” analysis, whole *batches* of events are processed at once, meaning

- 100s or even 1000s of events are loaded at once
- A bulk computation is done over the whole batch
- The temporary space is cleaned and the next event is loaded

The lower overhead for this method is particularly attractive for Data Scientists and is being embraced by CMS


Event
Event
Event
Event loop
Traditional
numpy vector operation - *fast*
Event
Event
Event
Striped