Helix Nebula Science Cloud usage by ALICE

Helix Nebula Science Cloud meeting

2018-06-14

Maarten Litmaath
CERN-IT

v1.1

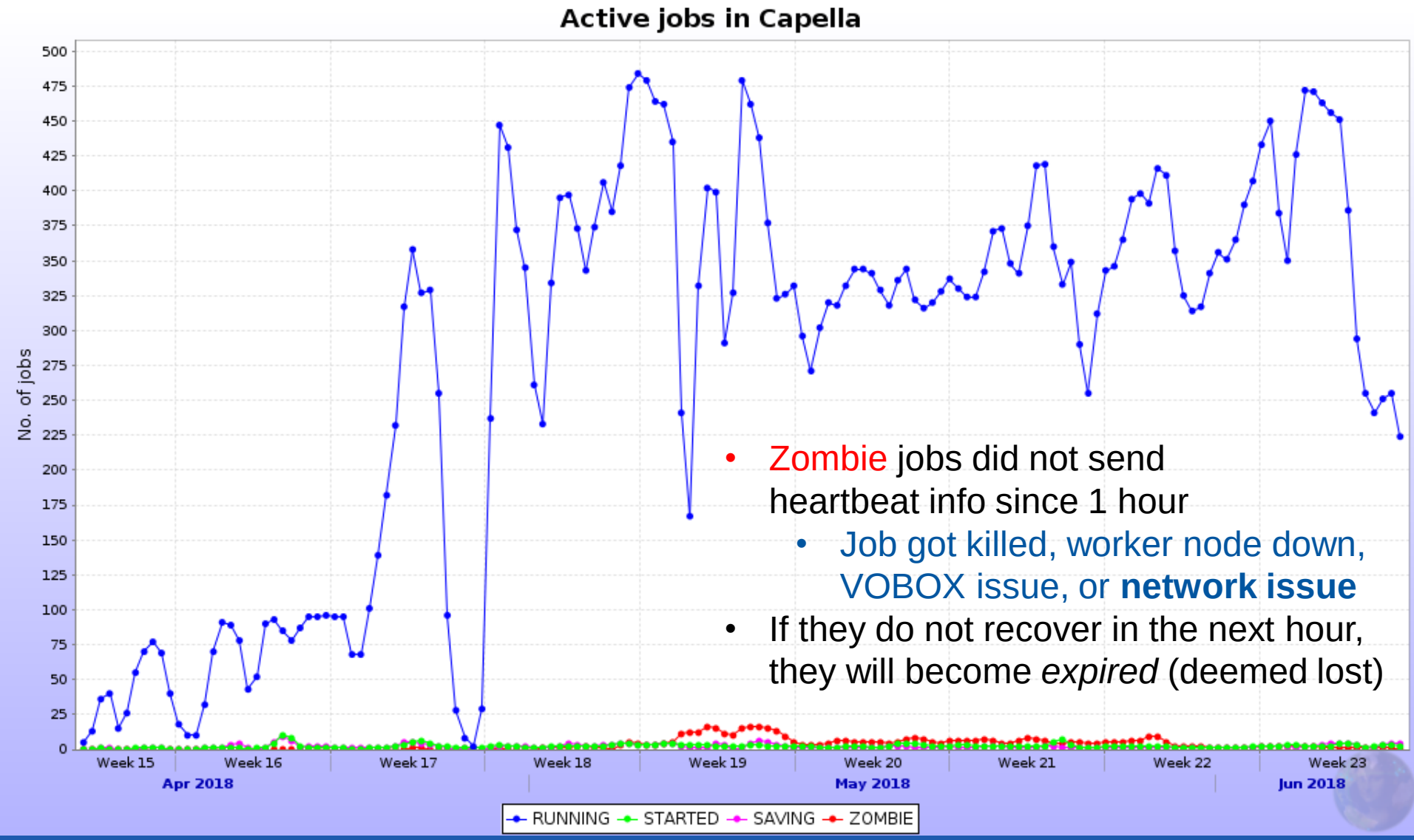
Modus operandi

- The IT batch team instantiate VMs/containers that connect to the production HTCondor pool fronted by production HTCondor Computing Element (CE) services
  - For the remote resources to match jobs, the jobs need to be submitted with an extra option to select *T-Systems* or *RHEA*
    - `+WantHNSciTsys=True`
    - `+WantHNSciRHEA=True`

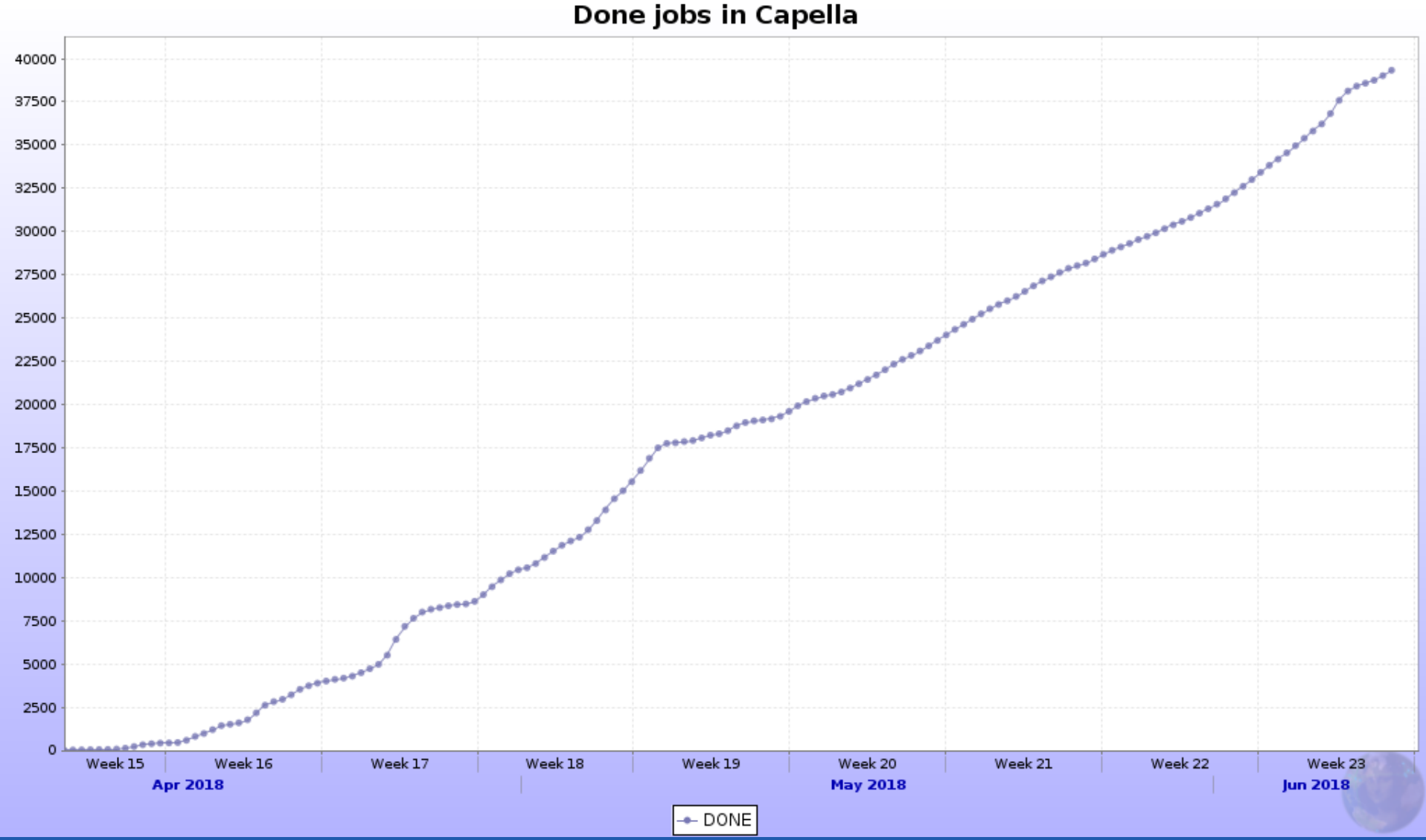
- Each destination has a dedicated ALICE VOBOX (management host) for a corresponding "virtual" site
  - *Capella* → jobs go to *T-Systems* via `ce515`
  - *Regulus* → jobs go to *RHEA* via `ce516`
  - Neither CE is currently used for other ALICE activities
    - Allowing HTCondor job numbers to be easily monitored and compared against ALICE MonALISA numbers for the virtual sites
  - As the remote resources are not in a CERN domain, each VOBOX needs *openings in the campus firewall*

Timeline

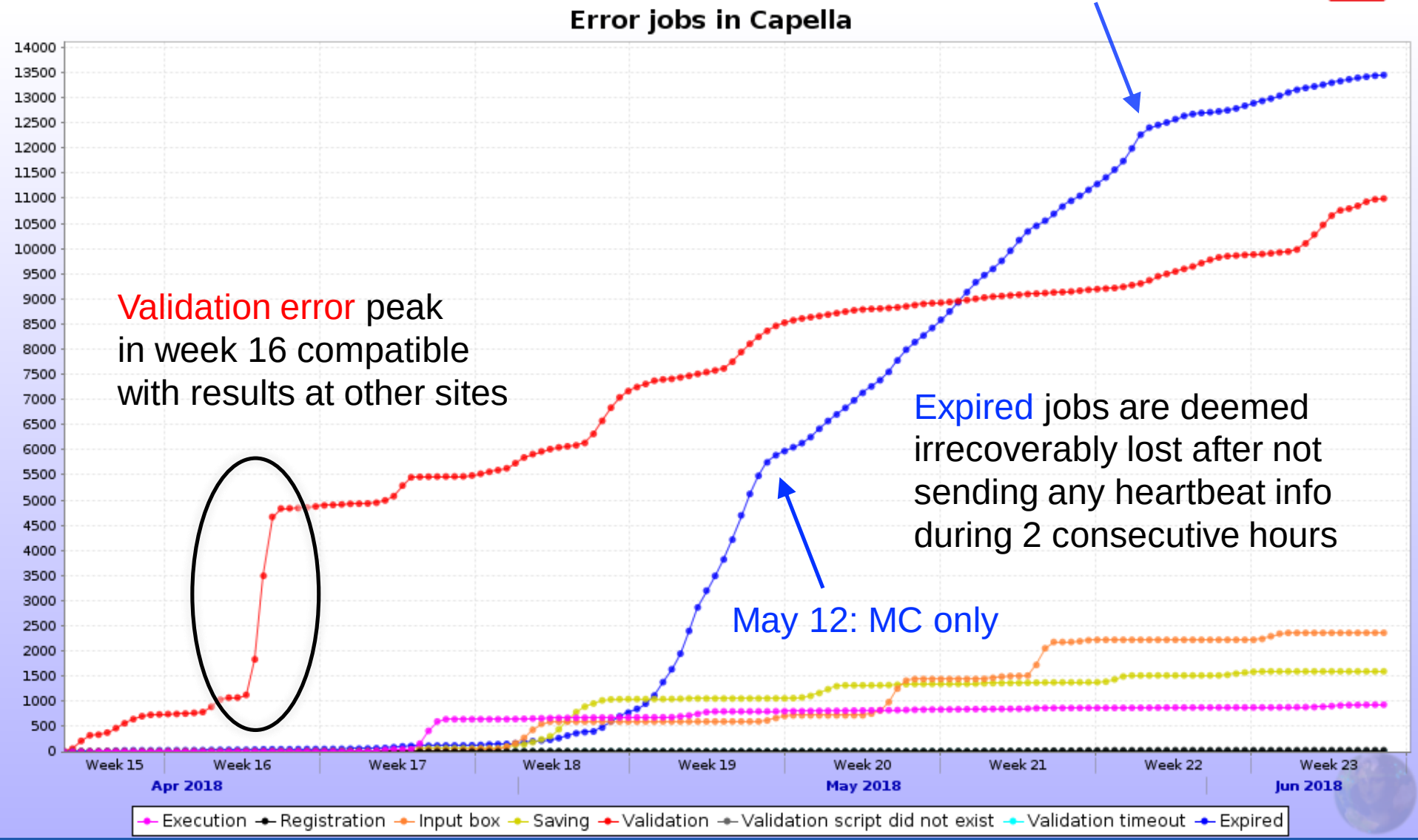
- March – early April
  - Provided the batch team with ALICE workflows to help debug cloud resource configuration issues
- April 10
  - Activity starts ramping up
  - Plots on the following pages use that as starting date
  - Job types: MC simulation, raw data reconstruction
- May 12
  - Job types restricted to MC simulation because of high failure rates possibly correlated with I/O
  - There appeared to be an instant improvement, as shown in the error jobs plot on page 6



- Zombie jobs did not send heartbeat info since 1 hour
  - Job got killed, worker node down, VOBOX issue, or **network issue**
- If they do not recover in the next hour, they will become *expired* (deemed lost)


Done jobs in Capella
40000
37500
35000
32500
30000
27500
25000
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
Week 15
Week 16
Week 17
Week 18
Week 19
Week 20
Week 21
Week 22
Week 23
Apr 2018
May 2018
Jun 2018
DONE

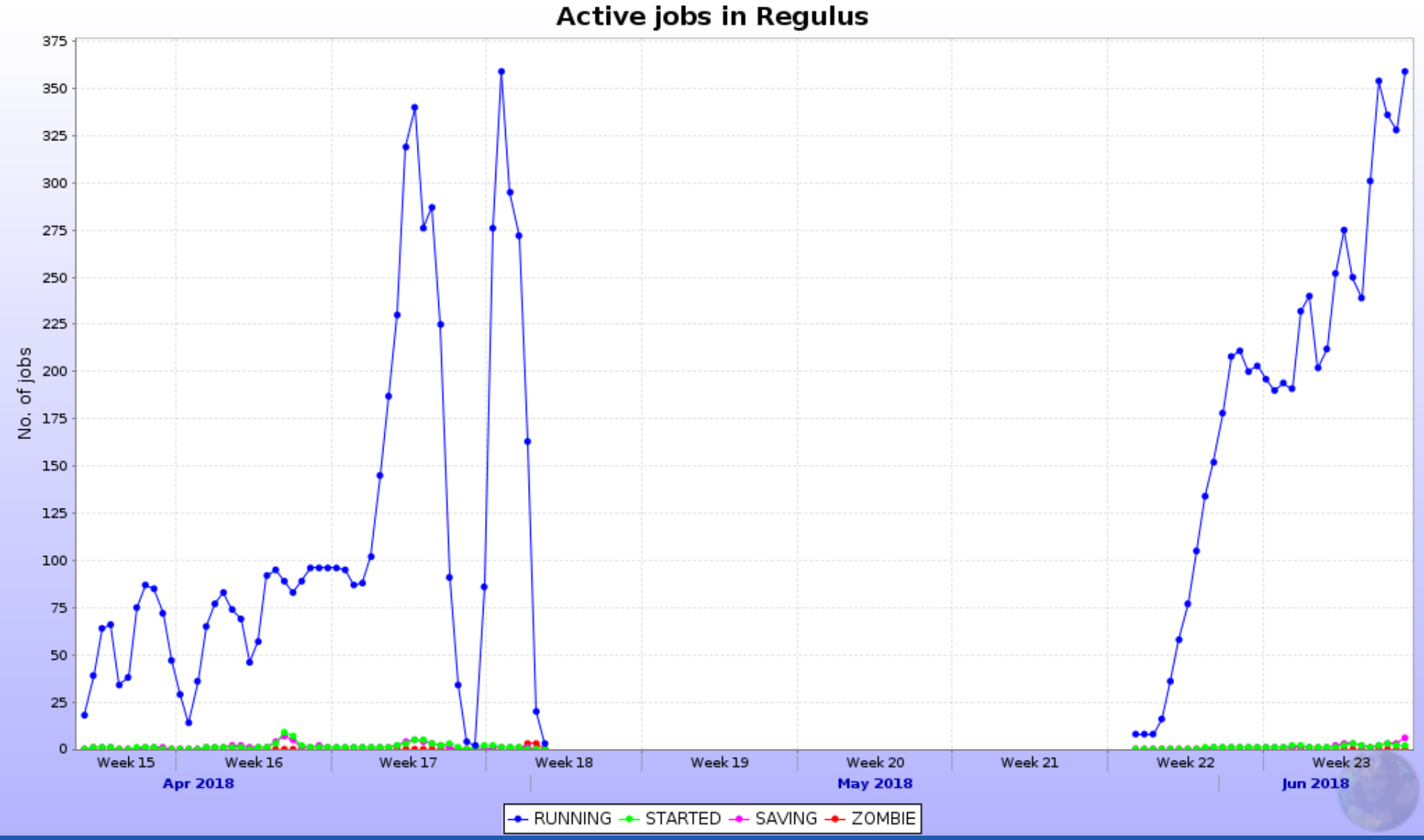
May 29: NAT MTU (?) change by batch team

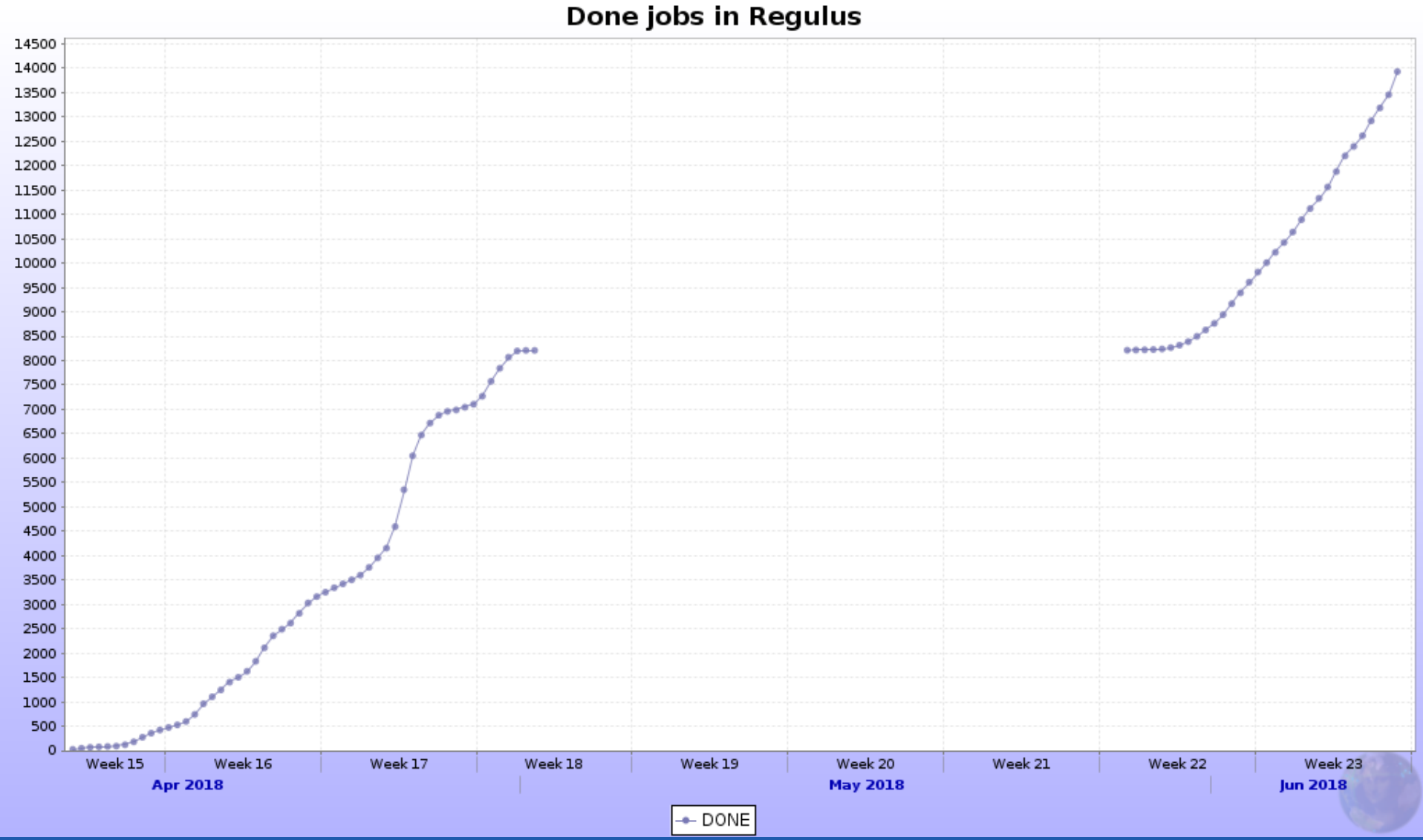


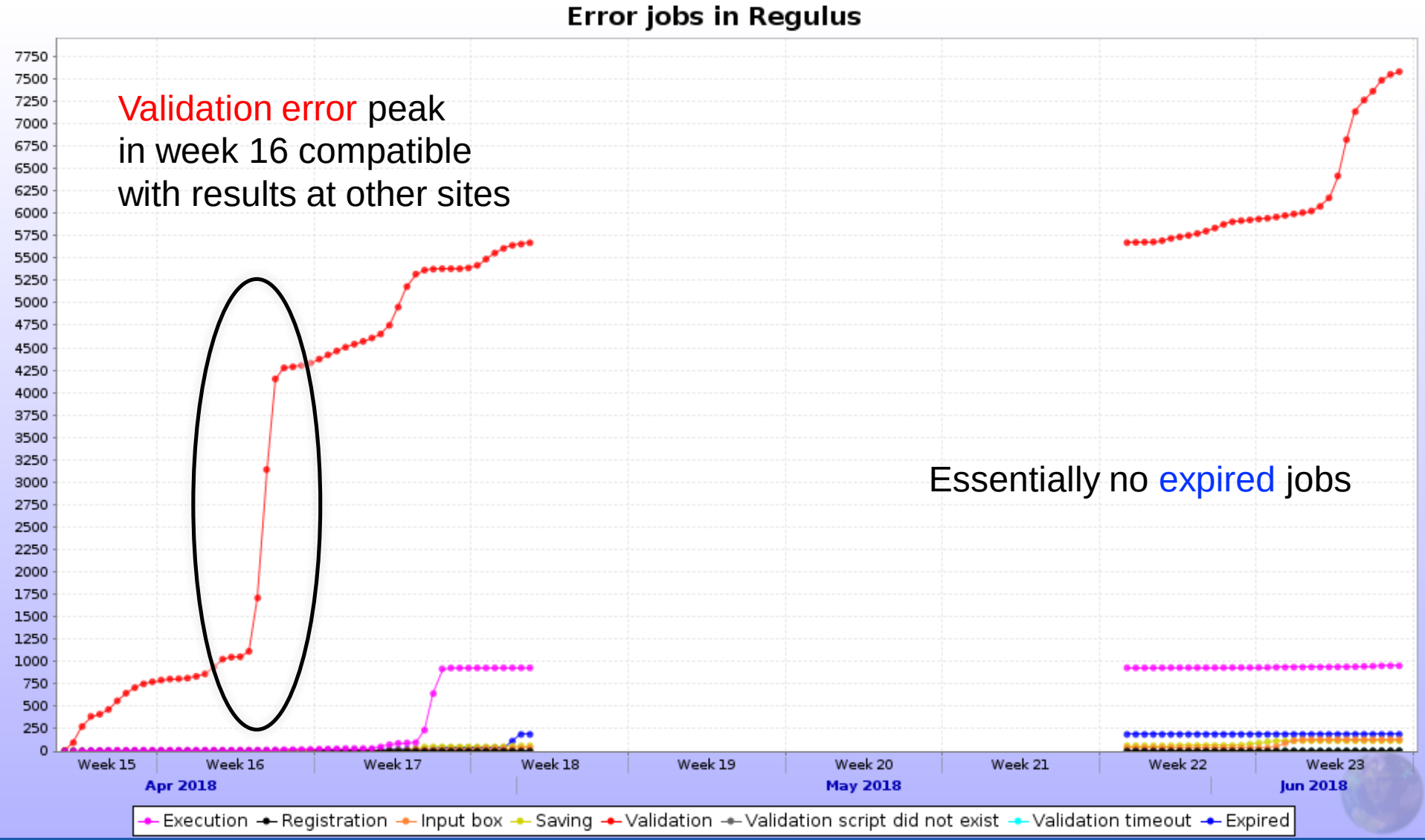
Validation error peak in week 16 compatible with results at other sites

Expired jobs are deemed irrecoverably lost after not sending any heartbeat info during 2 consecutive hours

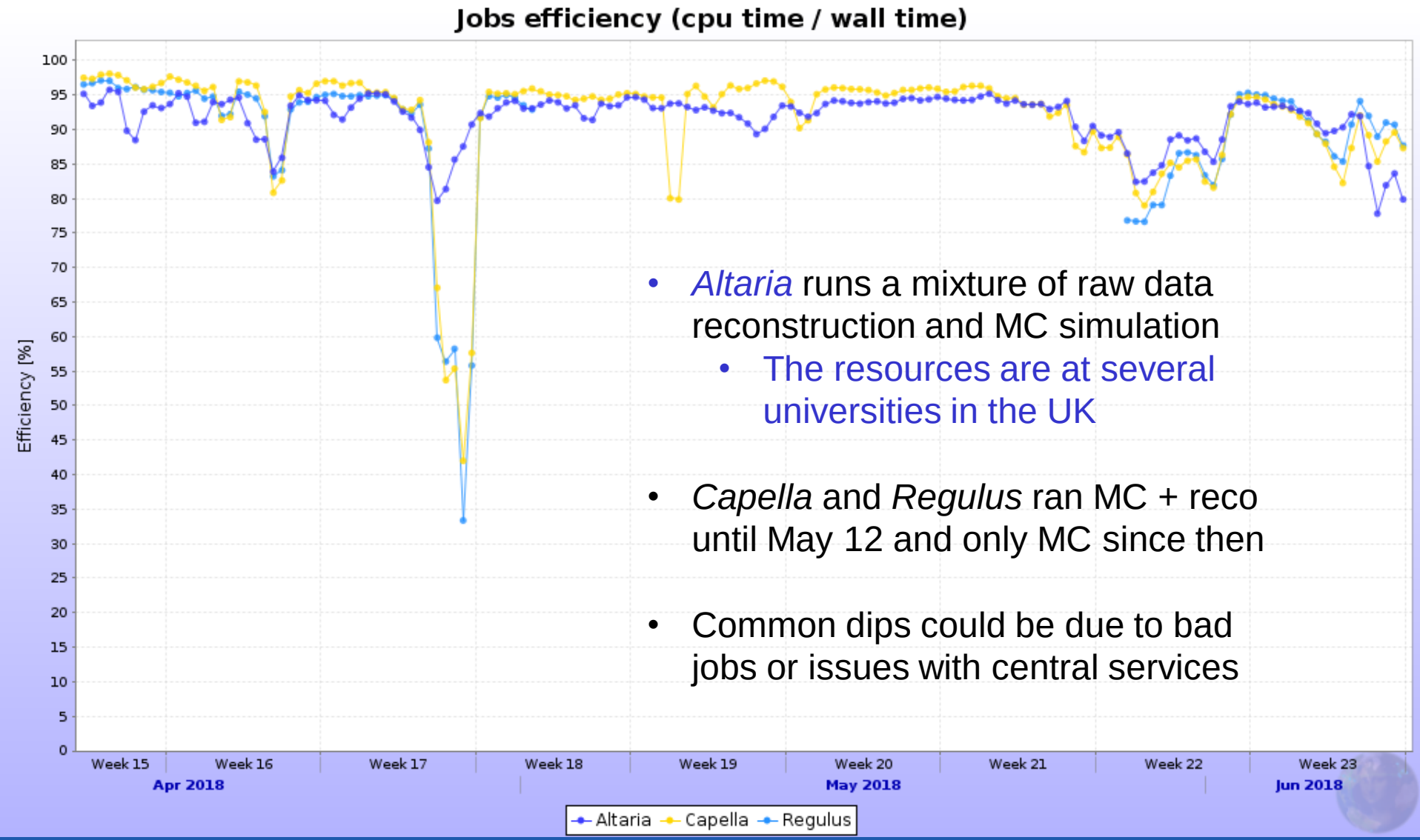
May 12: MC only


Active jobs in Regulus
No. of jobs
375
350
325
300
275
250
225
200
175
150
125
100
75
50
25
0
Week 15
Week 16
Week 17
Week 18
Week 19
Week 20
Week 21
Week 22
Week 23
Apr 2018
May 2018
Jun 2018
RUNNING
STARTED
SAVING
ZOMBIE


Done jobs in Regulus
14500
14000
13500
13000
12500
12000
11500
11000
10500
10000
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Week 15
Week 16
Week 17
Week 18
Week 19
Week 20
Week 21
Week 22
Week 23
Apr 2018
May 2018
Jun 2018
DONE

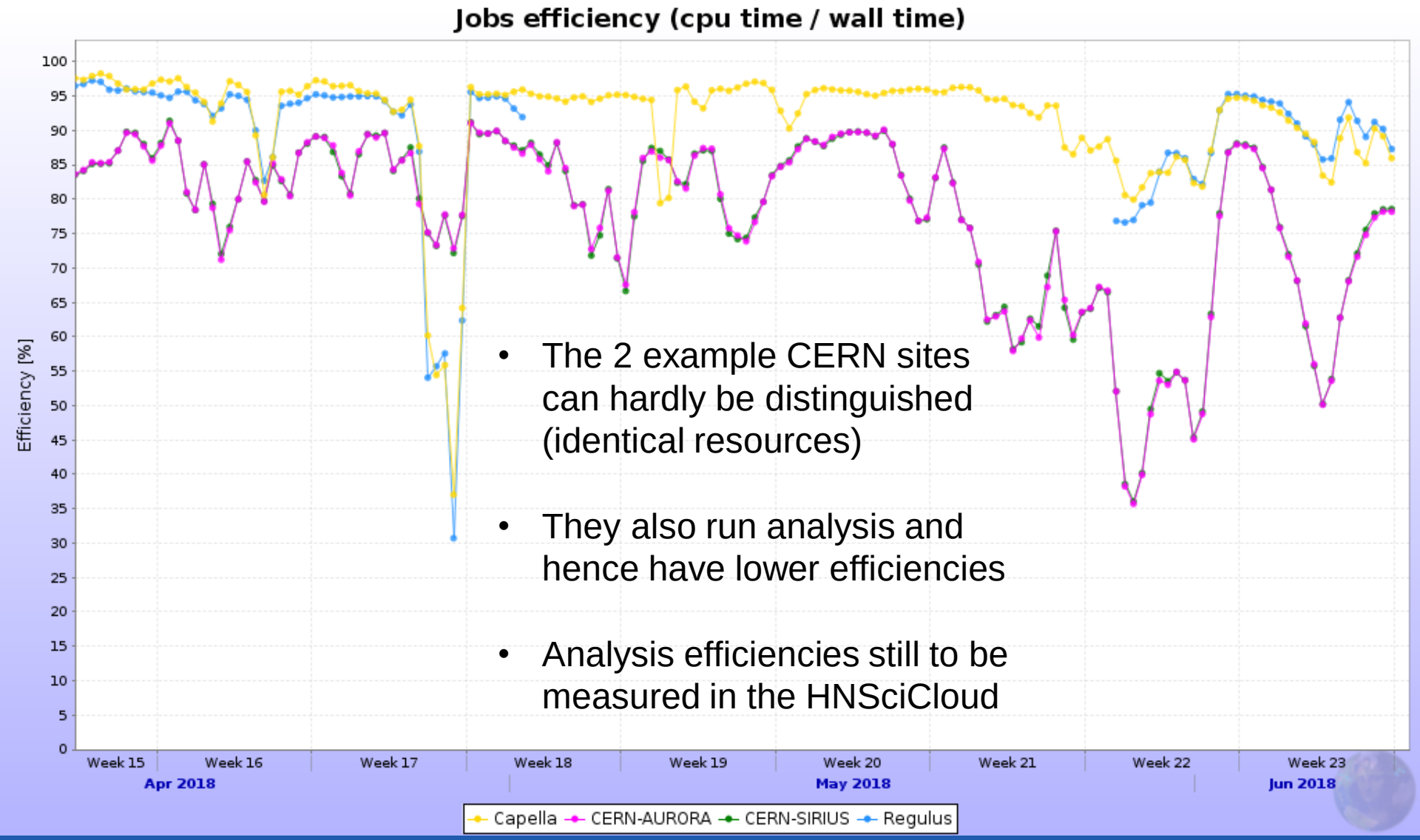

Error jobs in Regulus
Validation error peak
in week 16 compatible
with results at other sites
Essentially no expired jobs
7750
7500
7250
7000
6750
6500
6250
6000
5750
5500
5250
5000
4750
4500
4250
4000
3750
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
Week 15
Week 16
Week 17
Week 18
Week 19
Week 20
Week 21
Week 22
Week 23
Apr 2018
May 2018
Jun 2018
Execution
Registration
Input box
Saving
Validation
Validation script did not exist
Validation timeout
Expired

Cloud site job efficiencies



- *Altaria* runs a mixture of raw data reconstruction and MC simulation
  - The resources are at several universities in the UK
- *Capella* and *Regulus* ran MC + reco until May 12 and only MC since then
- Common dips could be due to bad jobs or issues with central services

HNSciCloud & CERN efficiencies



- The 2 example CERN sites can hardly be distinguished (identical resources)
- They also run analysis and hence have lower efficiencies
- Analysis efficiencies still to be measured in the HNSciCloud

Conclusions & outlook

- ALICE were able to start making use of HNSciCloud resources quite quickly
  - Cloud details were handled by the batch team
- MC simulation workloads are a good match
  - Still to be fixed: expired jobs in T-Systems
- Success rates and efficiencies of other job types are still to be measured
- I/O-intensive jobs have some concerns
  - Network performance between cloud resources and remote Storage Elements (SEs)
  - To measure network bandwidth, latency and reliability, the VOBOXes should be in the cloud, not at CERN
    - Important for matching jobs to virtual sites and to SEs they should use
  - Use of S3 cloud storage would require major development
  - XRootD-compatible cloud storage could work