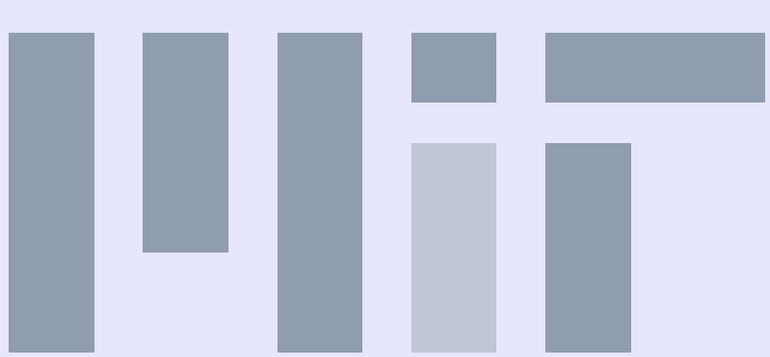
The Tier-2 in the HPRCF at Bates



Max Goncharov
for the Tier2 Team
May 10, 2023

Overview

What is the Tier-2?

- CMS HEP Tier-2 computing center (T2_US_MIT)
- CMS Heavy Ion Simulation and Analysis Facility
- LHCb Tier-2 computing center
- CLAS12 + other LNS project computing resources
- 44 racks, 24k compute cores, 17 PB storage, 100 Gb/s WAN

In light of Bates upgrade plans

- Tier-2 planning depends on the infrastructure
  - describe major subsystems
  - choices we have to consider organizing Tier-2
- What matters to our implementation of the Tier-2 (wish list)

Ultimate goal – stable Tier-2 operations

Tier-2 Components and Services

There are 3 main services our Tier-2 provides

Data transfers (this is all working for now, plan to go to 400 Gb/sec for HL-LHC)

- Data transfer nodes (xrootd doors)
- All severs are in UPS racks
- 10 Gb/s networking per node, 100 Gb/s for all of HPRCF
- Using one UPS rack, need another one soonish

Run user jobs

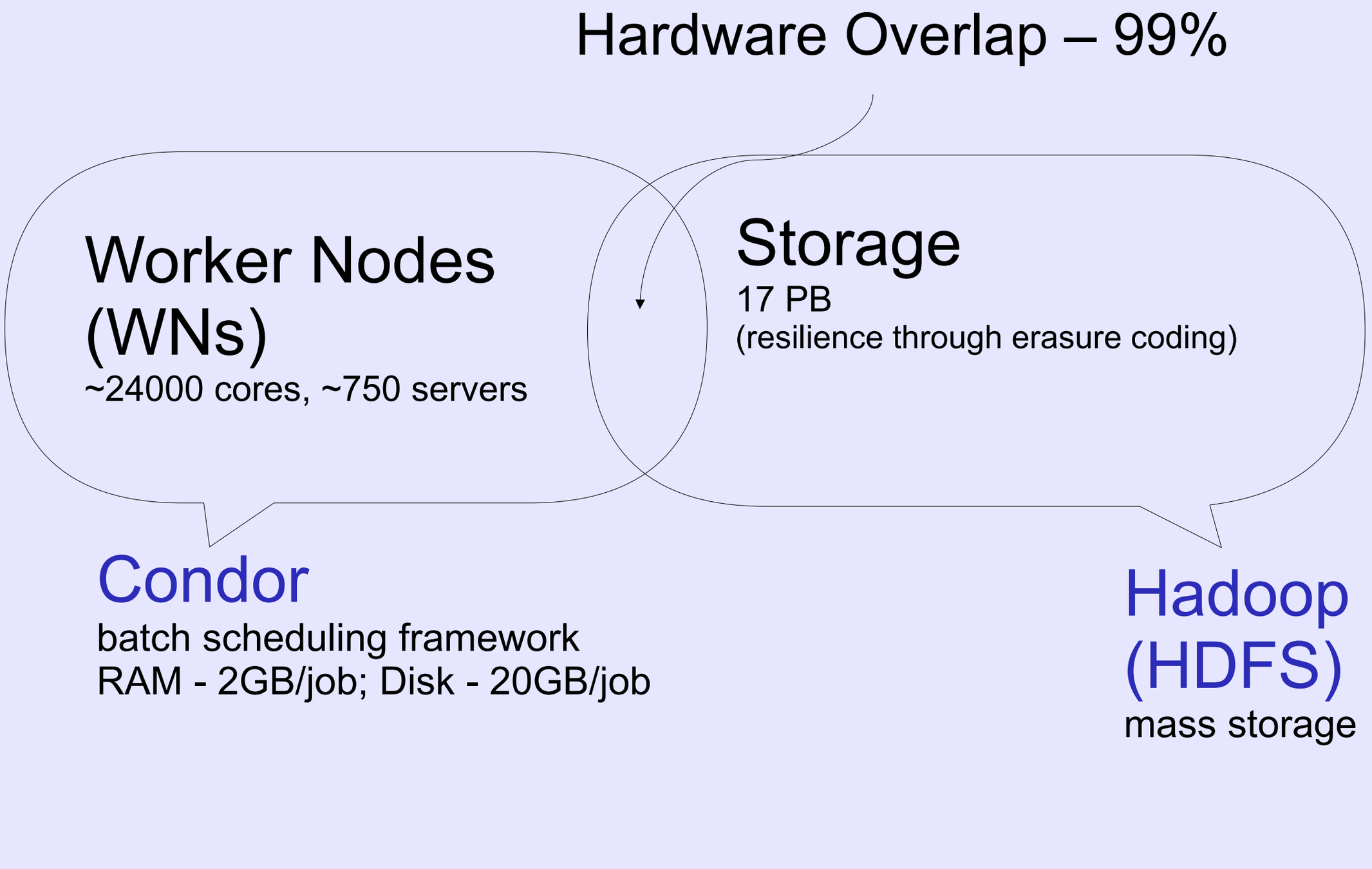
- Computing job service (condor servers, 3) in UPS racks
- Worker Nodes (WNs) in many racks, all without UPS

Storage

- Mass storage implementation uses Hadoop (HDFS)
- Conceptually all WNs carry disks that host HDFS data

Cluster Configuration

CPU/Storage Mix Model



CMS Tier-2 Community

CMS constantly probes our center

- Can we transfer data in and out (xrootd)?
  - transfers have to work (xrootd) so we appear ‘green’
  - storage has to be functional (WNs are up) so we appear ‘green’
- Can we run jobs (WNs are up)?

All sites are ranked by CMS according to availability

CMS Log Retrieval

SAM 1 day status ranking of American12Sites for 2023-Apr-01 00:00:00 to 2023-Apr-30

Site:	Bar Graph:	Status Value:	outside downtime:
T2_US_MIT		1.000	1.000
T2_US_Purdue		1.000	1.000
T2_US_Caltech		1.000	1.000
T1_US_FNAL		1.000	1.000
T2_US_Nebraska		1.000	1.000
T2_US_Wisconsin		1.000	1.000
T2_US_UCSD		0.933	0.933
T2_US_Vanderbilt		0.833	0.833
T2_US_Florida		0.767	0.793
T2_BR_UERJ		0.733	0.733
T2_BR_SPRACE		0.600	0.600

April was great for our site, no outages!

But power and cooling outages push us down →

Outages

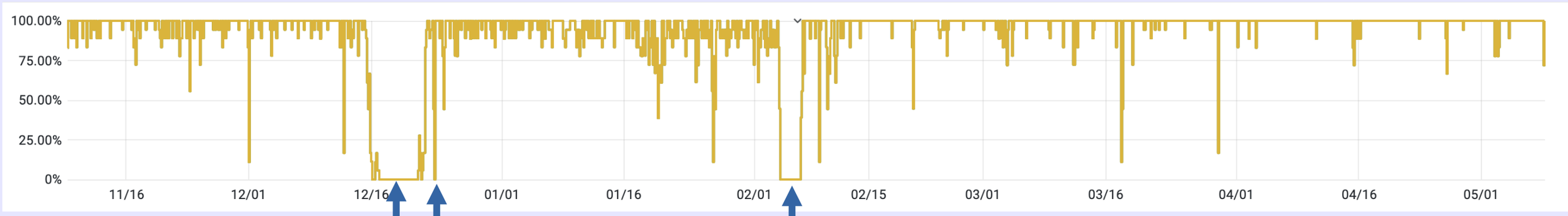
Our performance over the last six month

CMS Log Retrieval

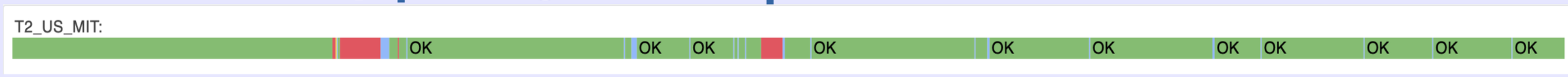
SiteReadiness 1 day status ranking of American12Sites for 2022-Nov-01 00:00:00 to 2023-May-08

Site:	Bar Graph:	Status Value:	outside downtime:
T2_US_Wisconsin		0.989	0.989
T2_US_Nebraska		0.984	0.984
T1_US_FNAL		0.957	0.957
T2_US_Caltech		0.936	0.936
T2_US_UCSD		0.893	0.893
T2_US_MIT		0.834	0.834
T2_US_Vanderbilt		0.818	0.841
T2_US_Purdue		0.786	0.786
T2_US_Florida		0.781	0.793
T2_BR_SPRACE		0.770	0.787
T2_BR_UERJ		0.481	0.495

Outages affect our ranking, and put pressure on Mike and the rest of the team to bring center online and recover damaged hardware.

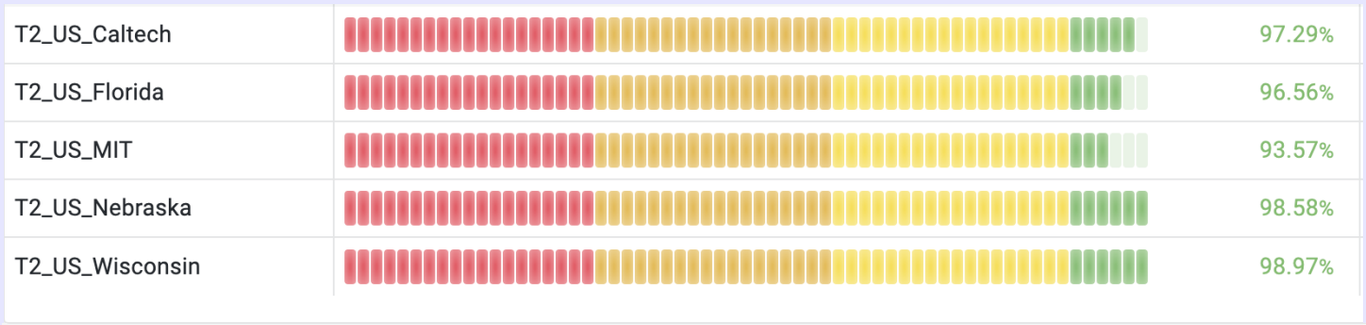


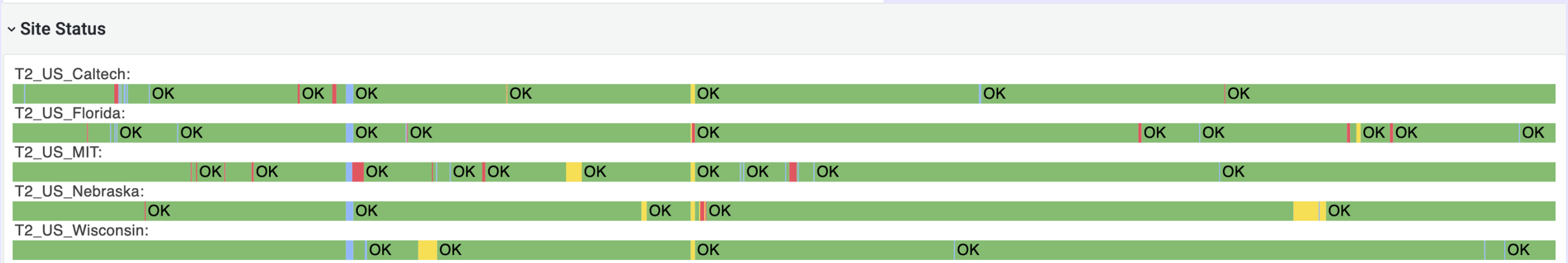
Outages



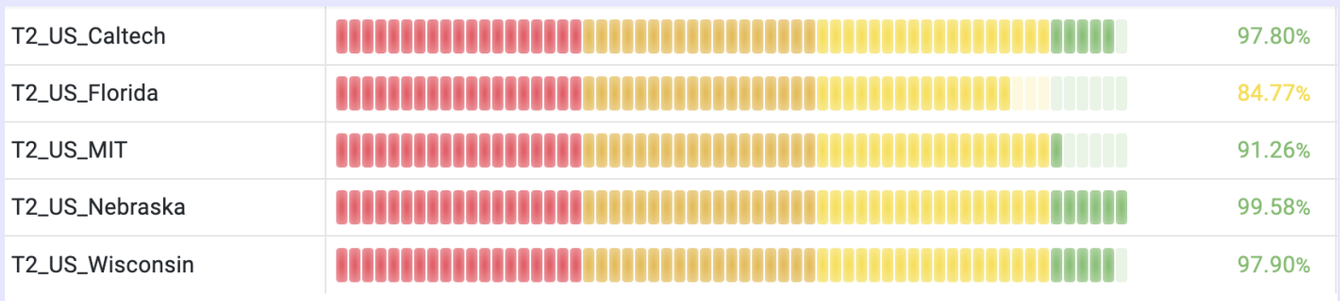
US Tier-2 Sites

May 1/22 – Oct 31/22

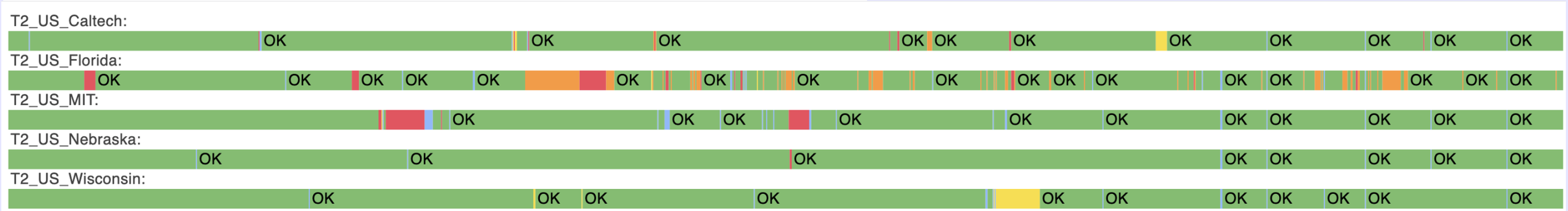




Nov 1/22 – May/23



Tier-2 at MIT is okay, but we want to do better.



Quality of Service

Promise to CMS for the Tier-2

- Outages/failures during business hours are ‘immediate’ (within ~hour)
- Otherwise service is best effort based
- Consistent with our operation model as long as we have run the Tier-2

Outages - Communications

- E-mail by main responsible (Jim?) to all HPRCF people (Bates): downtime incident and later once it is save to switch back on
- Mike is hands-on first responder
- Max communicates further about Tier-2 specific actions in coordination with Mike, and Chad and Qier who can help with some key services
- Tools: slack channel, cleo ticket system

Outage Specific

Outages

- Power outage: short (UPS stays up)
  - Main services intact (UPS), workers need powering
  - Recover ‘broken’ workers carrying the largest disk space first
- Power outage: extended (all down)
  - Services (UPS racks) first, then workers
  - Recover large storage nodes next
- Cooling outage (UPS racks stay up)
  - Worker nodes need to shut down (automatic?)
- Network outage
  - All services including worker nodes are powered, but cannot do work
  - Hopefully the network problem can be found and fixed (sometimes driving along the fiber running to Bates helps)

Mitigating Outage Impact

How to mitigate impact of outages

Power outage / Cooling Outage

- All worker nodes go down
- Batch system (condor) and storage go down
- Data transfer servers (xrootd) are in UPS rack and stay up

Can Condor remain available?

What about storage?

Keep Condor Alive

To improve condor availability to protect against outages: what if some WNs stay online and available?

Put some CPUs into a different physical location. Network latency does not seem to be an issue.

Will run with significantly reduced center, but availability will not suffer from Condor probes failing.

CMS Log Retrieval

SAM 1 day status ranking of American12Sites for 2023-Apr-01 00:00:00 to 2023-Apr-30

Site:	Bar Graph:	Status Value:	outside downtime:
T2_US_MIT		1.000	1.000
T2_US_Purdue		1.000	1.000
T2_US_Caltech		1.000	1.000
T1_US_FNAL		1.000	1.000
T2_US_Nebraska		1.000	1.000
T2_US_Wisconsin		1.000	1.000
T2_US_UCSD		0.933	0.933
T2_US_Vanderbilt		0.833	0.833
T2_US_Florida		0.767	0.793
T2_BR_UERJ		0.733	0.733
T2_BR_SPRACE		0.600	0.600

Condor stays alive

Cluster Configuration

CPU/Storage Mix Model

Hardware Overlap – 99%

Worker Nodes (WNs)
~24000 cores, ~750 servers

Storage
17 PB
(resilience through erasure coding)

Condor
batch scheduling framework
RAM – 2 GB/job; Disk – 20 GB/job

Hadoop (HDFS)
mass storage

Power and cooling outages:

Worker Nodes go down → Storage becomes unavailable

Keep Storage Alive

Storage Models

Current model

Primary
(constant writes and reads)

2U	2U	2U	2U	2U	2U
2U	2U	2U	2U	2U	2U
2U	2U	2U	2U	2U	2U

Spread disks to as many controllers as possible

Cannot stay up during outages

Facilitates parallel reads/writes

High maintenance

Secondary
(store data, access it occasionally)

JBOD

JBOD

Put disks in big JBOD boxes

Can sit in few racks with UPS power

Can stay up during outages

Somewhat decreased performance

Worker nodes can be repaired on less urgent track

Mitigating Outage Impact

At the moment we use HDFS

- 17 PB of disk space
- erasure coding
- non-posix file system
- no geolocations
- 1 Gb/s connection is enough

Community starting to use other technologies: CephFS

- erasure coding
- posix compliant
- geolocation
- 10 Gb/s is a must

With geolocation we could split storage in different physical locations, like we did for batch system (condor)

Non-posix FS not supported by CMS any more → we have to maintain plugins

Conclusions

All racks on UPS power

- Continue with the current Condor/HDFS mix model
- 10 Gb connection for Wns

Limited UPS availability

- We would ask for one row of racks on UPS
- Break away from Condor/HDFS mix model
  - 1U nodes packed with CPUs
  - need fewer racks
- Put storage in UPS racks
  - can stay up in case of power outage
  - can stay up during cooling outage
  - 10 Gb connection for storage nodes

Need 1 more UPS rack soon (expand xrootd servers, 10 Gb/s)