A Sysadmin's Point of View on Operations

Judith Stephen

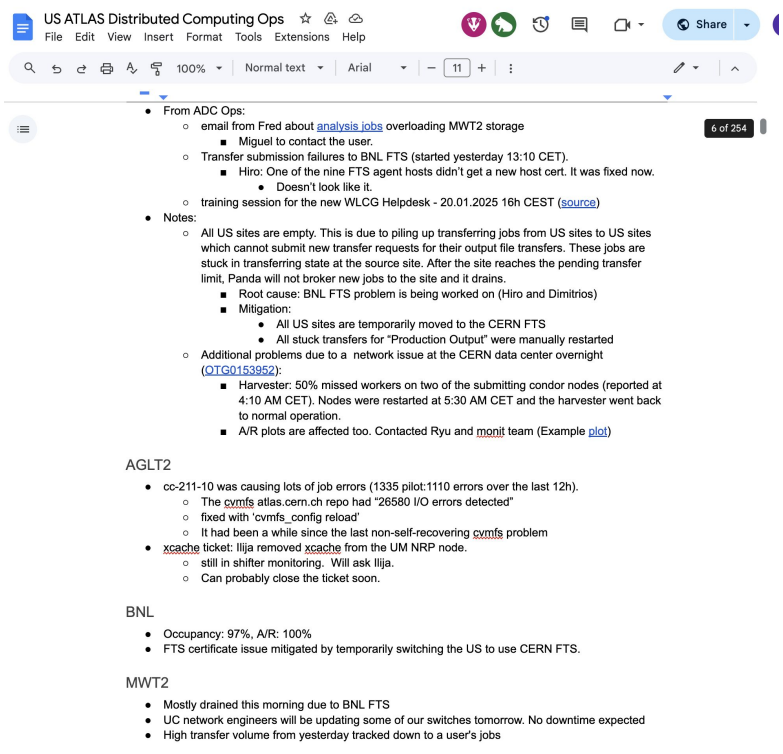
ATLAS Distributed Computing Technical Interchange Meeting @ SBU

21 January 2025

Daily USATLAS Ops Meeting

- Started in May 2024
- Bridge the gap between USATLAS site administrators and ADC
- Greatly helped in improving communication both between ADC and site admins and also between the individual sites

Daily Ops Reports



US ATLAS Distributed Computing Ops

File Edit View Insert Format Tools Extensions Help

Share

100% Normal text Arial 11

6 of 254

- From ADC Ops:
  - email from Fred about analysis jobs overloading MWT2 storage
    - Miguel to contact the user.
  - Transfer submission failures to BNL FTS (started yesterday 13:10 CET).
    - Hiro: One of the nine FTS agent hosts didn't get a new host cert. It was fixed now.
      - Doesn't look like it.
  - training session for the new WLCG Helpdesk - 20.01.2025 16h CEST (source)
- Notes:
  - All US sites are empty. This is due to piling up transferring jobs from US sites to US sites which cannot submit new transfer requests for their output file transfers. These jobs are stuck in transferring state at the source site. After the site reaches the pending transfer limit, Panda will not broker new jobs to the site and it drains.
    - Root cause: BNL FTS problem is being worked on (Hiro and Dimitrios)
    - Mitigation:
      - All US sites are temporarily moved to the CERN FTS
      - All stuck transfers for "Production Output" were manually restarted
  - Additional problems due to a network issue at the CERN data center overnight (OTG0153952):
    - Harvester: 50% missed workers on two of the submitting condor nodes (reported at 4:10 AM CET). Nodes were restarted at 5:30 AM CET and the harvester went back to normal operation.
    - A/R plots are affected too. Contacted Ryu and monit team (Example plot)

AGLT2

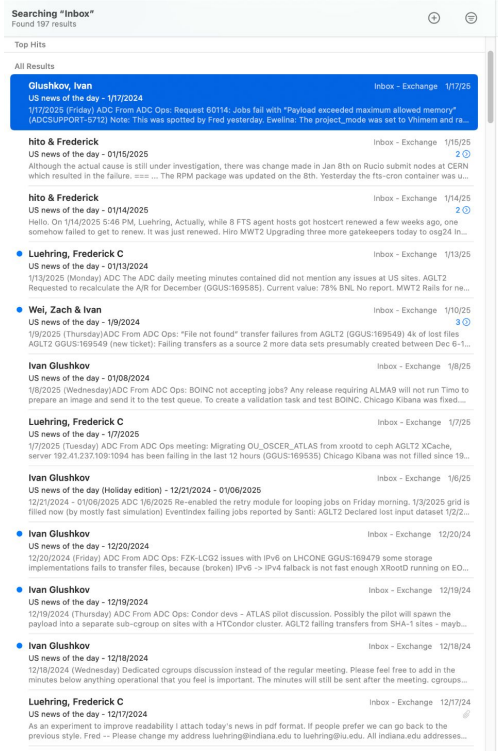
- cc-211-10 was causing lots of job errors (1335 pilot:1110 errors over the last 12h).
  - The cvmfs atlas.cern.ch repo had "26580 I/O errors detected"
  - fixed with 'cvmfs_config reload'
  - It had been a while since the last non-self-recovering cvmfs problem
- xcache ticket: Ilija removed xcache from the UM NRP node.
  - still in shifter monitoring. Will ask Ilija.
  - Can probably close the ticket soon.

BNL

- Occupancy: 97%, A/R: 100%
- FTS certificate issue mitigated by temporarily switching the US to use CERN FTS.

MWT2

- Mostly drained this morning due to BNL FTS
- UC network engineers will be updating some of our switches tomorrow. No downtime expected
- High transfer volume from yesterday tracked down to a user's jobs



Searching "Inbox"
Found 197 results

Top Hits

All Results

Glushkov, Ivan — Inbox - Exchange 1/17/25
US news of the day - 1/17/2024
1/17/2025 (Friday) ADC From ADC Ops: Request 60114: Jobs fail with "Payload exceeded maximum allowed memory" (ADCSUPPORT-5712) Note: This was spotted by Fred yesterday. Ewelina: The project_mode was set to Vhimem and ra...

hito & Frederick — Inbox - Exchange 1/15/25
US news of the day - 01/15/2025
Although the actual cause is still under investigation, there was change made in Jan 8th on Rucio submit nodes at CERN which resulted in the failure. === ... The RPM package was updated on the 8th. Yesterday the fts-cron container was u...

hito & Frederick — Inbox - Exchange 1/14/25
US news of the day - 01/14/2025
Hello. On 1/14/2025 5:46 PM, Luehring, Actually, while 8 FTS agent hosts got hostcert renewed a few weeks ago, one somehow failed to get to renew. It was just renewed. Hiro MWT2 Upgrading three more gatekeepers today to osg24 in...

Luehring, Frederick C — Inbox - Exchange 1/13/25
US news of the day - 01/13/2024
1/13/2025 (Monday) ADC The ADC daily meeting minutes contained did not mention any issues at US sites. AGLT2 Requested to recalculate the A/R for December (GGUS:169585). Current value: 78% BNL No report. MWT2 Rails for ne...

Wei, Zach & Ivan — Inbox - Exchange 1/10/25
US news of the day - 1/9/2024
1/9/2025 (Thursday)ADC From ADC Ops: "File not found" transfer failures from AGLT2 (GGUS:169549) 4k of lost files AGLT2 GGUS:169549 (new ticket): Failing transfers as a source 2 more data sets presumably created between Dec 6-1...

Ivan Glushkov — Inbox - Exchange 1/8/25
US news of the day - 01/08/2024
1/8/2025 (Wednesday)ADC From ADC Ops: BOINC not accepting jobs? Any release requiring ALMA9 will not run Timo to prepare an image and send it to the test queue. To create a validation task and test BOINC, Chicago Kibana was fixed....

Luehring, Frederick C — Inbox - Exchange 1/7/25
US news of the day - 1/7/2025
1/7/2025 (Tuesday) ADC From ADC Ops meeting: Migrating OU_OSCER_ATLAS from xrootd to ceph AGLT2 XCache, server 192.41.237.109:1094 has been failing in the last 12 hours (GGUS:169535) Chicago Kibana was not filled since 19...

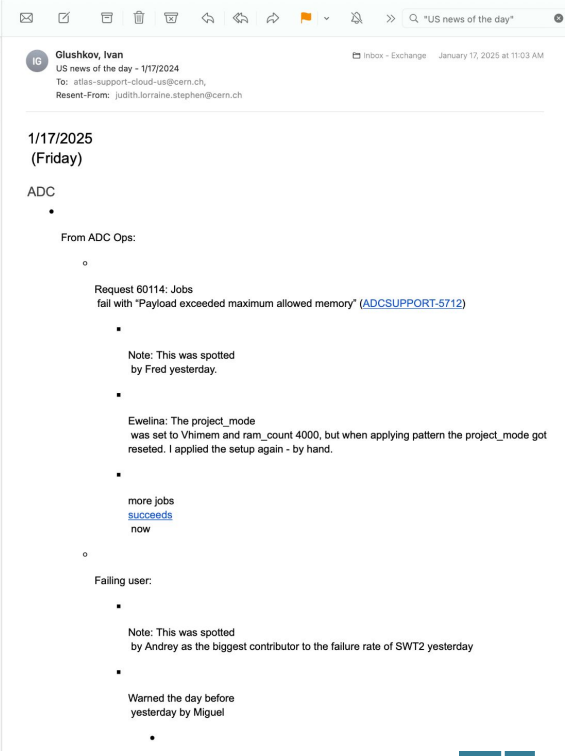
Ivan Glushkov — Inbox - Exchange 1/6/25
US news of the day (Holiday edition) - 12/21/2024 - 01/06/2025
12/21/2024 - 01/06/2025 ADC 1/6/2025 Re-enabled the retry module for looping jobs on Friday morning. 1/3/2025 grid is filled now (by mostly fast simulation) EventIndex failing jobs reported by Santi: AGLT2 Declared lost input dataset 1/2/2...

Ivan Glushkov — Inbox - Exchange 12/20/24
US news of the day - 12/20/2024
12/20/2024 (Friday) ADC From ADC Ops: FZK-LCG2 issues with IPv6 on LHCONE GGUS:169479 some storage implementations fails to transfer files, because (broken) IPv6 -> IPv4 falback is not fast enough XRootD running on EO...

Ivan Glushkov — Inbox - Exchange 12/19/24
US news of the day - 12/19/2024
12/19/2024 (Thursday) ADC From ADC Ops: Condor devs - ATLAS pilot discussion. Possibly the pilot will spawn the payload into a separate sub-cgroup on sites with a HTCondor cluster. AGLT2 failing transfers from SHA-1 sites - mayb...

Ivan Glushkov — Inbox - Exchange 12/18/24
US news of the day - 12/18/2024
12/18/2024 (Wednesday) Dedicated cgroups discussion instead of the regular meeting. Please feel free to add in the minutes below anything operational that you feel is important. The minutes will still be sent after the meeting. cgroups...

Luehring, Frederick C — Inbox - Exchange 12/17/24
US news of the day - 12/17/2024
As an experiment to improve readability I attach today's news in pdf format. If people prefer we can go back to the previous style. Fred -- Please change my address luehring@indiana.edu to luehring@iu.edu. All indiana.edu addresses...



"US news of the day"

Glushkov, Ivan
US news of the day - 1/17/2024
To: atlas-support-cloud-us@cern.ch,
Resent-From: judith.lorraine.stephen@cern.ch

Inbox - Exchange January 17, 2025 at 11:03 AM

1/17/2025
(Friday)

ADC

- From ADC Ops:
  - Request 60114: Jobs fail with "Payload exceeded maximum allowed memory" (ADCSUPPORT-5712)
    - Note: This was spotted by Fred yesterday.
    - Ewelina: The project_mode was set to Vhimem and ram_count 4000, but when applying pattern the project_mode got reseted. I applied the setup again - by hand.
    - more jobs succeeds now
  - Failing user:
    - Note: This was spotted by Andrey as the biggest contributor to the failure rate of SWT2 yesterday
    - Warned the day before yesterday by Miguel

Impact of Job Mixture on Site Operations

- Different job campaigns have different resource requirements
  - CE/Harvester schedules CPU, memory, disk space
  - Site issues are often caused by heavy disk I/O, high network utilization, etc.
- If we know what's coming, it is easier for us as site administrators to diagnose issues faster and report back to ADC

Job Mixture Example

June 26th, 2024

9:29 AM **jlstephen** @Fred Luehring how do i get the job batch information again for the pileup jobs? we now have a ticket against us

9:30 AM **Fred Luehring** Look for the word pile on the panda monitor page for MWT2...

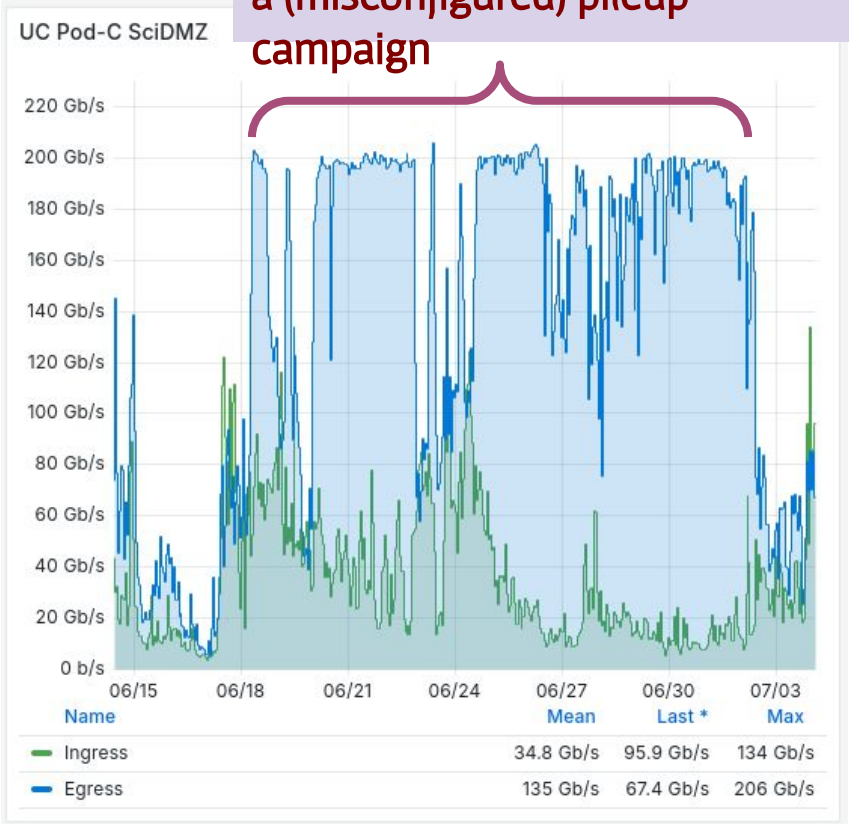
9:38 AM **jlstephen** thanks

i'm not sure how to reply to this ticket. most of the jobs in the linked https://bigpanda.cern.ch/jobs/?computingsite=MWT2&jobstatus=failed&days=2&piloterrorc[...]imit=100&date_from=2024-06-24T14:23&date_to=2024-06-26T14:23 are pile jobs, and we already know we can't do anything to increase our wan bandwidth from uc to iu and uiuc

9:53 AM **Fred Luehring** Let me investigate a bit.

Just so I have good understanding of the current situtation.

WAN maxed out largely due to a (misconfigured) pileup campaign

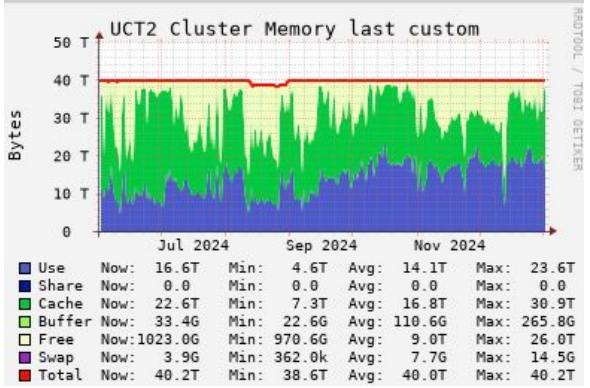


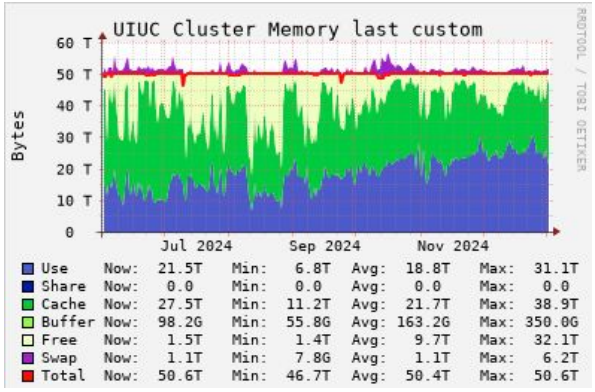
Evolving Queue Requirements

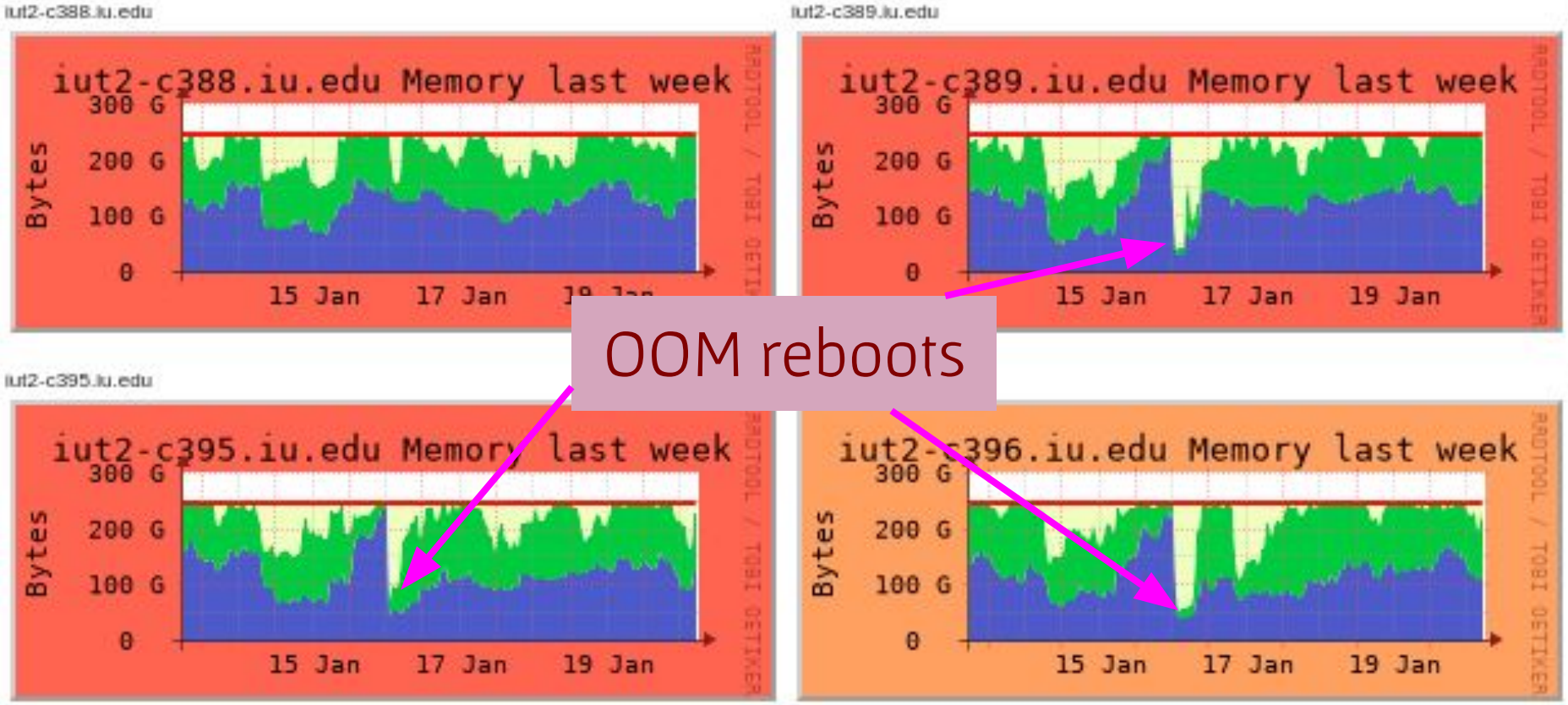
- Optimizing memory utilization at our site to allow more VHIMEM jobs to run
  - 2 GB/core has been the recommendation for a while, is this still appropriate for future compute purchases?
- If jobs exceed their requested memory, what does ADC want us to do? As an HTCondor site we can…
  - Let the job run to completion
  - Hard kill job immediately after it exceeds its request via cgroups
  - Hard kill the job after it goes over some multiple of its request
- Is there a feature we want the HTCondor team to implement?

MWT2 Memory Utilization

- On average, we have the extra memory to allow jobs to exceed the request
  - Doing nothing makes sense if jobs only exceed their request by a small amount
  - We currently kill after the job goes 3x over its request
- So how do we use all of the extra memory most effectively?






iut2-c388.iu.edu
iut2-c388.iu.edu Memory last week
Bytes
300 G
200 G
100 G
0
15 Jan
17 Jan
iut2-c389.iu.edu
iut2-c389.iu.edu Memory last week
Bytes
300 G
200 G
100 G
0
15 Jan
17 Jan
19 Jan
OOM reboots
iut2-c395.iu.edu
iut2-c395.iu.edu Memory last week
Bytes
300 G
200 G
100 G
0
15 Jan
17 Jan
19 Jan
iut2-c396.iu.edu Memory last week
Bytes
300 G
200 G
100 G
0
15 Jan
17 Jan
19 Jan

Site Communication to ADC

- When we're debugging issues, we have:
  - Ganglia
  - Nagios
  - Grafana plots for networking, CE queues, storage, etc.
- How can we better communicate our site status to ADC so they can have an idea of what's going on on the site side?

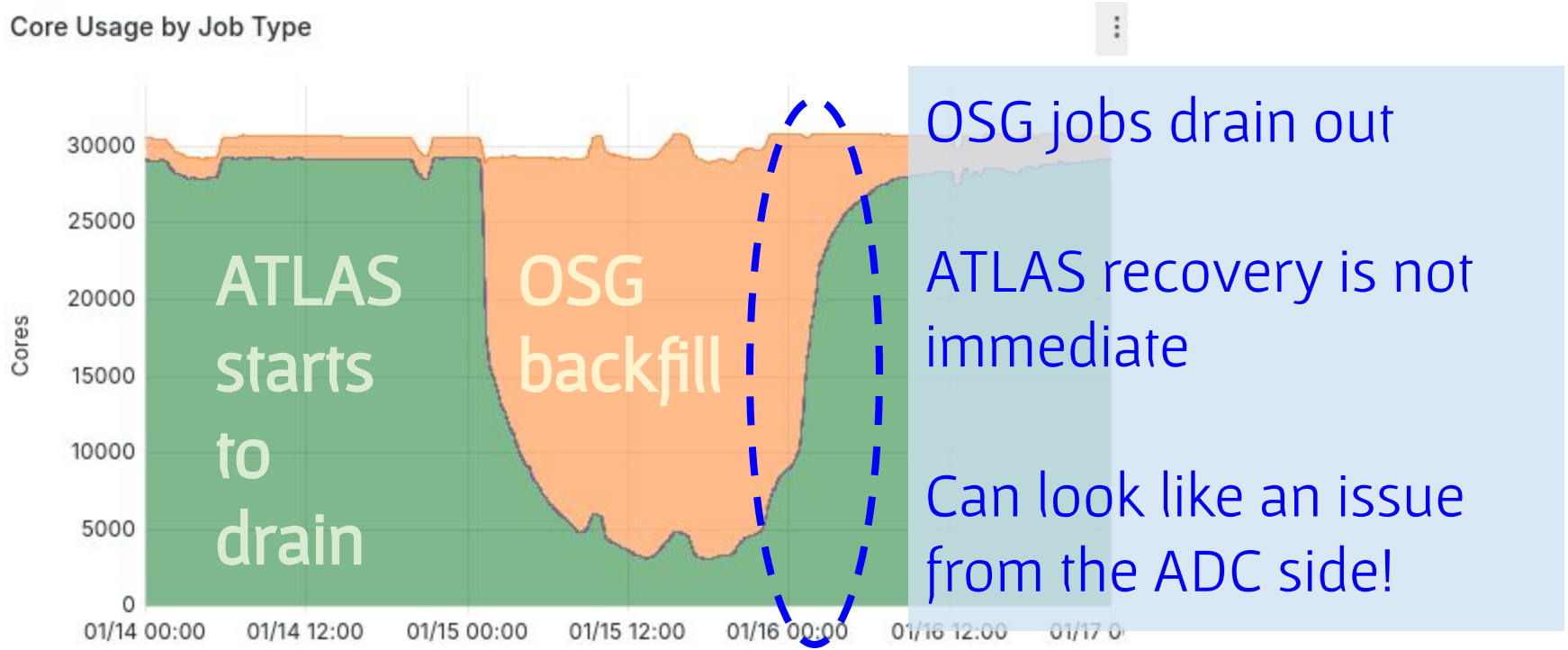
Recent Examples: Site Partial Downtime

- Currently we can only declare a CE or SE downtime
  - We were just discussing this recently because we had a ~few week partial downtime for one of our sites
  - No good mechanism to communicate the partial outage upstream
  - Right now: Tell Ivan that 15K cores are offline for a few weeks
    - No GKs are down
    - No Storage is down
  - We can only report binary service statuses to ADC

Recent Examples: Site Drainage

- ATLAS is not our only customer
  - When we are drained of ATLAS jobs, we fill our site opportunistically with jobs from OSG
  - To ADC, this appears that the site is not filling when the drain condition is gone from the ATLAS side
  - We communicate this to ADC via Ivan on the Ops call
- Is there anything else as a site we can provide that would give more queue visibility to ADC?

Example Drain/Refill Event



Job Diagnostics

- A lot of effort has gone into diagnosing issues such as stuck CVMFS, zombie pilots, misbehaving workloads, etc.
- How to make it easier for site admins to track down job issues to report back to ADC?

Inter-site Communications

- Recently, a number of US sites have been active in the BNL Mattermost server in addition to the daily meetings
- This has been a powerful tool for us to leverage to work together to identify (US) ATLAS-wide problems
- We should continue to use this tool

Thank you!