# US ATLAS Ops and Monitoring

Ofer Rind, BNL

9 June, 2026

US ATLAS F2F Meeting, Madison, WI

# Operations and Monitoring - The Current State

- ATLAS ADC provides extensive monitoring of the computing grid activity
  - BigPanDA, Monit-Grafana, DDM/Rucio/FTS
  - US site-specific dashboards
  - Active, multi-level shift coverage with ticketing channels and email lists
- ESNET, perfSONAR and site network monitoring
  - With the recent addition of SciTags
- Sites have extensive local monitoring
  - Batch, storage, hardware performance and active alerting
  - SAM/Hammercloud functional testing, alerting, exclusions
- UChicago analytics db - ELK stack
  - Alarms and alerts
- Prototype site-specific LLM-based analysis of PanDA errors
  - Daily reports generated and distributed via email

Is it enough? Are there gaps? Missing correlations? Is the information easily found and presented effectively?

**What we really want is actionable information, not just data**

Brookhaven National Laboratory

# ADC Live Dashboard


Brookhaven National Laboratory

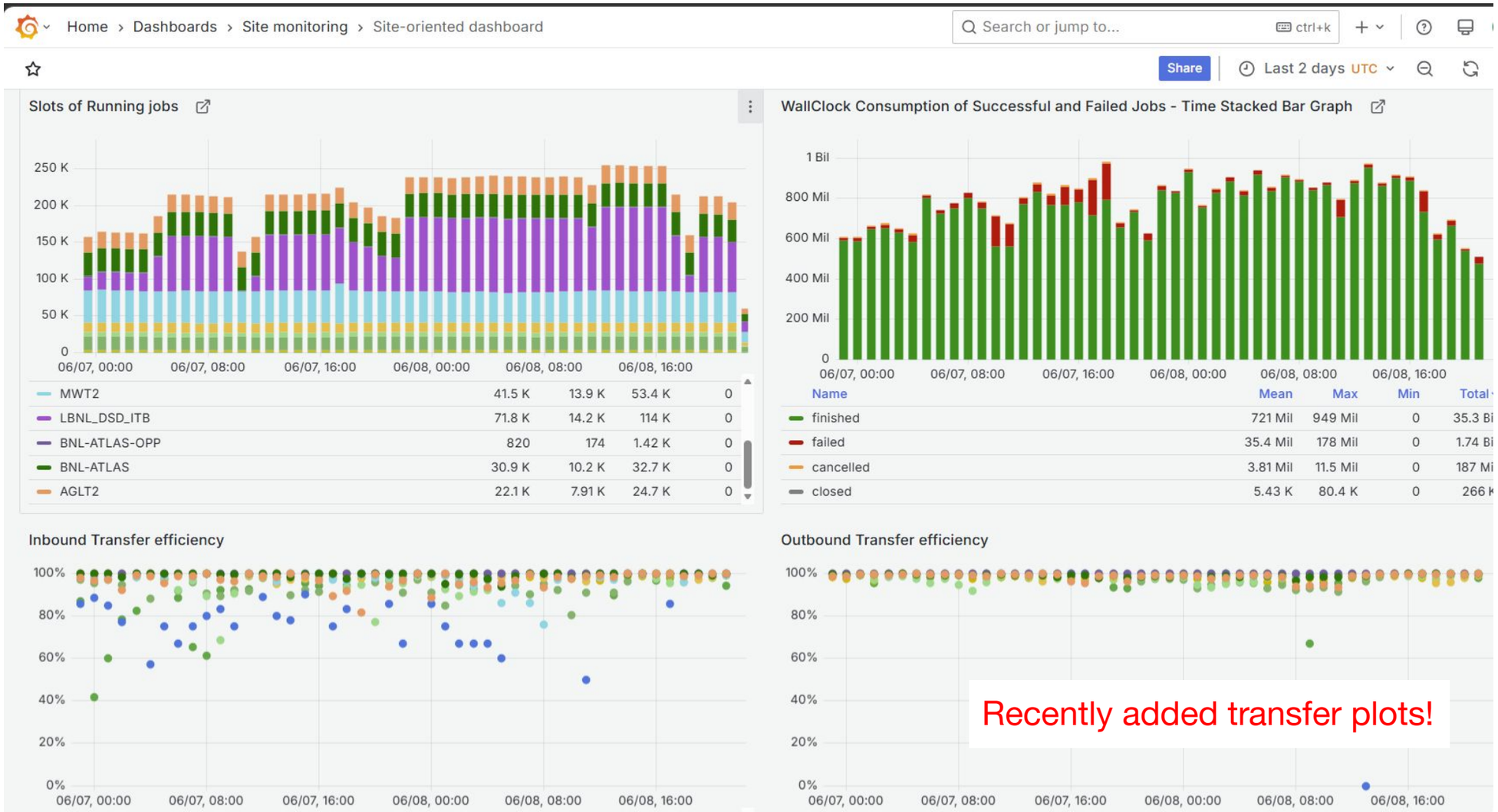
# US Site Status Boards

| SSB Overview | CRIC Downtime Status | DDM Downtime Status | DDM Transfer Efficiency | | SAM3 Site Availability | Panda Queues Status | Jobs Efficiency | | Frontier Squid Status | GGUS tickets |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | source | destinati... | | | analysis | producti... | | |
| AGLT2 | active | online | 98.0% | 96.6% | 100.0% | online | 95.0% | 99.0% | n/a | 0 |
| BNL-ATLAS | active | online | 98.6% | 99.1% | 100.0% | online | 54.0% | 100.0% | ok | 0 |
| MWT2 | active | online | 99.3% | 98.4% | 100.0% | online | 92.0% | 99.0% | ok | 0 |
| NET2 | active | online | 98.8% | 98.3% | 93.8% | online | 77.0% | 87.0% | n/a | 1 |
| OU_OSCER_... | active | online | 97.6% | 95.9% | 100.0% | online | 0.0% | 100.0% | n/a | 1 |
| SWT2_CPB | active | online | 96.1% | 94.4% | 100.0% | online | 64.0% | 99.0% | ok | 0 |
| TW-FTT | active | online | 97.5% | 98.8% | 100.0% | online | 98.0% | 95.0% | n/a | 0 |

| Site | | | | DT status | | | analysis | | | | | | | production | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ggus | site | ddm | fsquid | hc | status | activated | running | finished | failed | efficiency | hc | status | activated | running | finished | failed | efficiency |
| AGLT2 | US | T2 | 0 | ACTIVE | online | | no-test | online | 5930 | 2229 | 6654 | 318 | 95 | 100 | online | 4104 | 2626 | 7539 | 56 | 99 |
| **BNL-ATLAS** | US | **T1** | 0 | ACTIVE | online | OK | nodefq | online | 16895 | 9651 | 21240 | 1470 | 54 | 100 | online | 6029 | 3914 | 11366 | 10 | 100 |
| MWT2 | US | T2 | 0 | ACTIVE | online | OK | | online | 8708 | 5502 | 19998 | 820 | 92 | 100 | online | 6806 | 4998 | 17867 | 118 | 99 |
| NET2 | US | T2 | 1 | ACTIVE | online | | | online | 6116 | 3953 | 8918 | 785 | 77 | 100 | online | 3837 | 2383 | 7755 | 1199 | 87 |
| OU_OSCER_ATLAS | US | T2 | 1 | ACTIVE | online | | nodefq | online | 0 | 0 | 0 | 0 | 0 | 100 | online | 426 | 743 | 1474 | 5 | 100 |
| SWT2_CPB | US | T2 | 0 | ACTIVE | online | OK | 0 | online | 1318 | 2610 | 4414 | 2095 | 64 | 0 | online | 1498 | 2078 | 6495 | 40 | 99 |
| SWT2_GOOGLE | US | T2 | | ACTIVE | online | | | offline | 0 | 0 | 0 | 0 | 0 | multdefq | offline | 0 | 0 | 0 | 0 | 0 |
| TW-FTT | US | T2 | 0 | ACTIVE | online | | | online | 0 | 1 | 132 | 2 | 98 | 0 | online | 564 | 282 | 502 | 25 | 95 |

Brookhaven National Laboratory

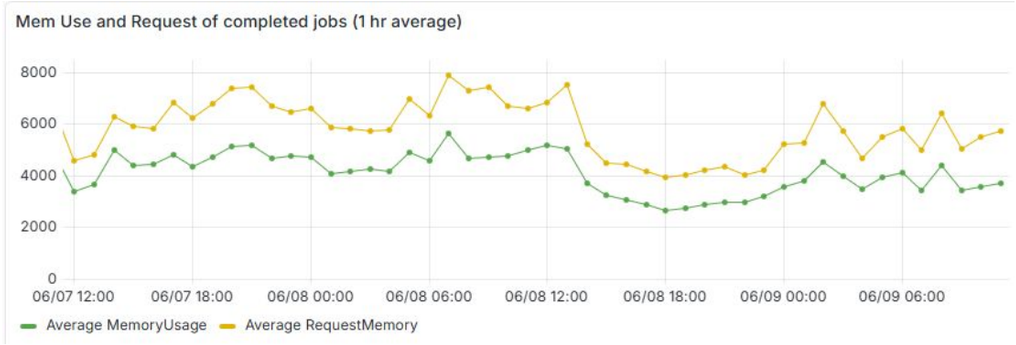
# US Site-Oriented Dashboard

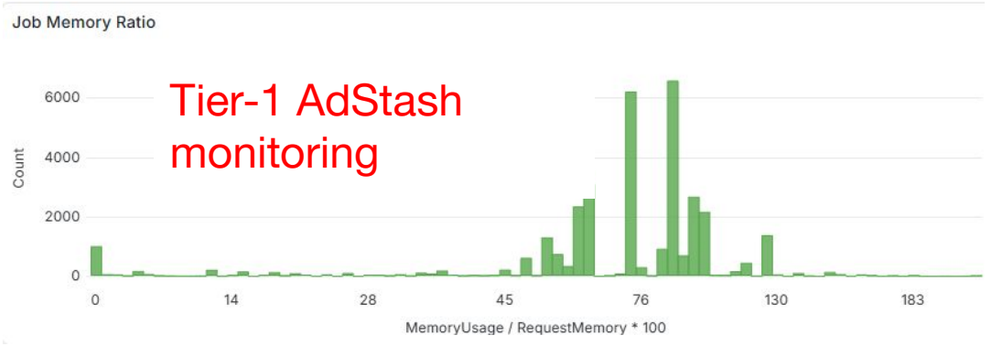



Brookhaven National Laboratory

# Site Monitoring

- Proliferation of tools (CheckMK, Nagios, Prometheus, OpenSearch,...)
  - AI tools under investigation (of course)
- Low-level information, some of it sensitive (limiting ability to share)
- Increasingly able to correlate local worker node information with PanDA monitoring, e.g. worker node performance dashboard.
- Now, with condor_chirp and AdStash, can go the other way, injecting job info into local monitoring.
  - Correlate with info such as job type, software, user, …

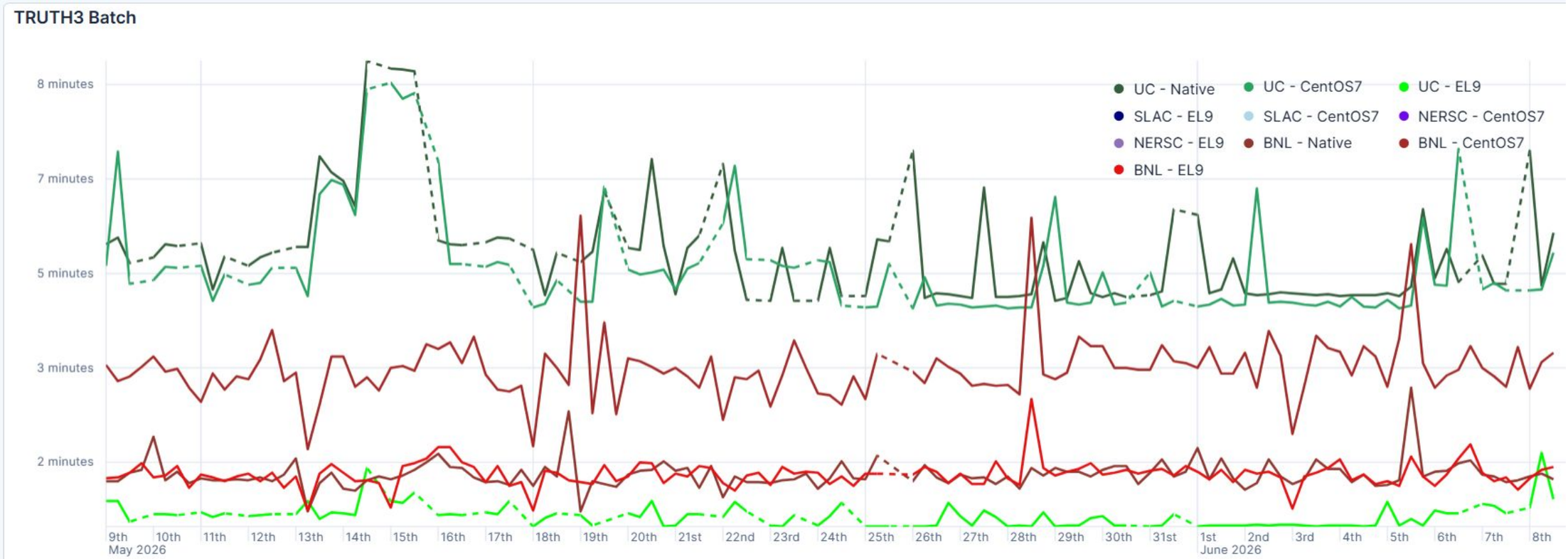



Brookhaven National Laboratory

# Centralized Analytics

- Ilija has been collecting a lot of data at UC for a long time
- Can we make better use of this data operationally?
- Example: benchmarking for monitoring on AFs (Giordon, Qi Bin)




Brookhaven National Laboratory

# AI Monitoring:  LLM analysis of PanDA logs

## [SITE ERROR]

BNL Example
Source: Kaushik De

**Error Count:** 2

**Error Code:** pilot:1361 (piloterrorcode)

### Description

Remote file could not be opened

### Category

Network and Communication Errors

### Diagnosis

Jobs failed to open input data files from the remote XRootD storage at BNL (root://dcgftp.usatlas.bnl.gov). Multiple EVNT pool files from the mc16_13TeV dataset were inaccessible.

### Classification

**Site Error** — This is an infrastructure issue related to storage access and network connectivity at the BNL computing facility, not a user job or payload problem.

## Pilot Job Analysis: PandaID 7170512481

### Node Information

- **Node ID:** `slot1_27@acas0616.usatlas.bnl.gov`
- **Site:** BNL-ATLAS
- **Queue:** BNL

Brookhaven National Laboratory

# LLM analysis of PanDA logs

## Root Cause

**Pilot timeout due to excessive runtime.** The pilot process (PID 4014815) exceeded its configured time floor and was forcefully terminated while still in the stage-out phase.

### Key Error Message

WARNING | the pilot has run out of time (timefloor=3600 has been passed)

## Critical Events

### 1. Payload Process Premature Exit

- Payload process (PID 4076474) exited with `exit_code=0` at `01:49:49`
- However, pilot continued attempting stage-out for ~3 minutes
- Multiple monitor loop warnings: *"aborting job monitor tasks since payload process 4076474 is not running"* (loops #1196–#1231)

### 2. Pilot Exit Code

Pilot exit status: 1
no translation to shell exit code for error code 1361

Exit code **1361** suggests a time-related shutdown (non-standard error code translation failure).

### 3. Memory & Resource Pressure

- Pilot memory usage spike: **7.1 GB** (controller cgroup) at `01:50:48`
- 46–47 concurrent pilot processes running on node
- No OOM events detected, but high memory contention evident

## Component Failure

### Pilot Daemon / Time Management

- The pilot's internal time floor (3600 seconds = 1 hour) was breached
- Stage-out operations (RuCIO/davs uploads) completed successfully but pilot hit timeout boundary
- Job transitioned to `DONE_FINAL` state before final cleanup

## Actionable Recommendation

### For Site Admin:

Increase the pilot time floor allocation at BNL. Currently set to **3600 seconds**, this is insufficient for jobs with large output files (10+ GB AOD files).

**Action:**

- Review `/etc/condor_config.local` or harvester configuration to increase `PILOT_TIMELIMIT` or equivalent parameter to **5400–7200 seconds** (90–120 minutes)
- Monitor if stage-out duration + payload execution routinely exceeds 1 hour for BNL physics workflows
- Alternatively, optimize stage-out parallelism to reduce total pilot runtime

Brookhaven National Laboratory

# Discussion - Future of US ATLAS Ops and Monitoring

We have extensive monitoring at multiple levels, but is what we have optimized for efficient operations?

- Is there information that's missing?  Are there parts we can consolidate? What can we improve?
  - Site-level monitoring correlation with PanDA, DDM, Network Monitoring, etc.
  - Greater use of analytics information
- Agentic Ops?
  - Build upon and improve current LLM log analysis
  - Expand OpenClaw/MCP tools to multiple sites
  - Cross-site anomaly detection
- What should be the work plan for 6 months, 12 months, LS3?
  - How to coordinate effort both across US ATLAS and with ADC?
  - Who can work on this?