

# Polaris and Acceptance Testing



**Brian Homerding**  
Leadership Computing Facility  
Argonne National Lab

**Ben Lenard**  
Leadership Computing Facility  
Argonne National Lab

**CUG 2023**  
May 10<sup>th</sup> 2023

# Acknowledgment

- This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.
- We would like to recognize all the ALCF staff who contributed on the integration efforts and acceptance testing.
- We would like to thank the following HPE and NVIDIA personnel for their extraordinary efforts: Jon Bouvet, Carrie Breuer, Greg Cross, Lisa Giacchetti, Max Katz, Mark Juaire, Todd Letsche, and many others.

# Polaris Overview

# Polaris

- ALCF's latest computational resource

<https://www.alcf.anl.gov/polaris>

12	<b>Polaris</b> - Apollo 6500, AMD EPYC 7532 32C 2.4GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/Argonne National Laboratory United States	259,840	23,840.0	34,595.6
----	----------------------------------------------------------------------------------------------------------------------------------------------------------	---------	----------	----------

*Top500 November 2021*



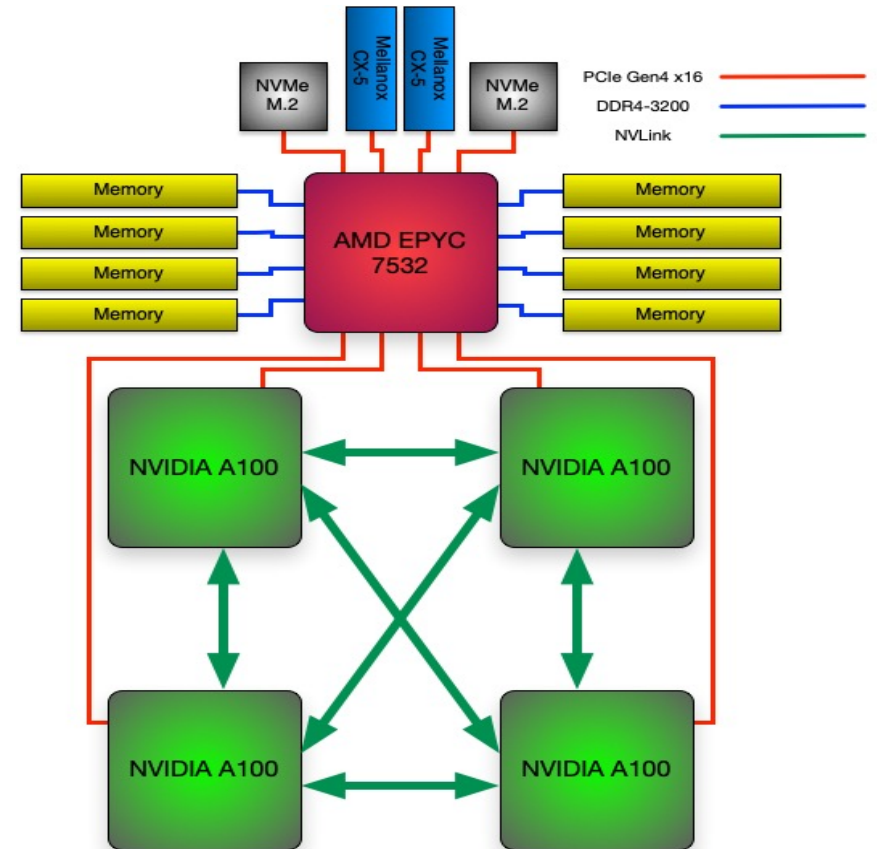
# System Overview

Polaris System Characteristics	
Compute Racks	40
Compute Node Count	560
Total GPUs	2,240
Peak DGEMM	34.6 PF
DDR Memory per Compute Node	512 GB
HBM Memory per Compute Node	160 GB
Compute Total Memory (Aggregate)	367.5 TB
Compute Total SSD Capacity	1.8 PB
Slingshot-10 Peak Injection Bandwidth	12.6 TB/s
Management Racks	2
Login Nodes	6
Quorum Nodes	3
Fabric Manager Nodes	2
Leader Nodes	3
WLM Nodes	2
Gateway Nodes	50

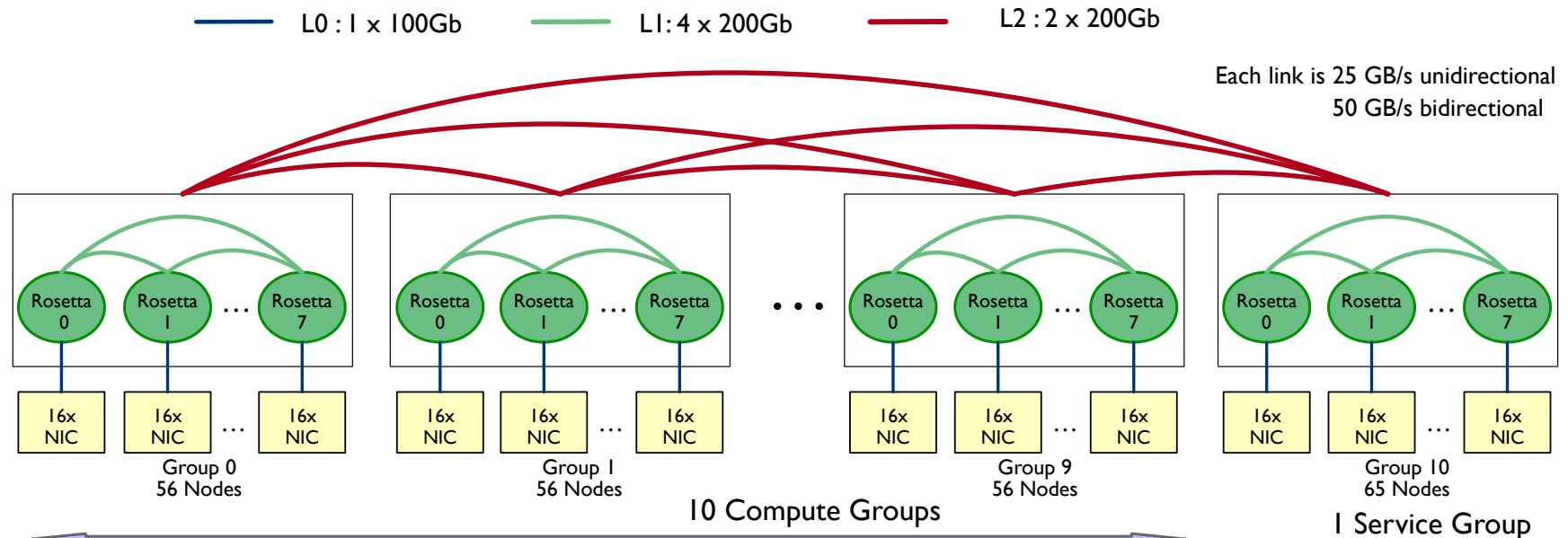
Polaris System Versions	
HPCM	1.5
CNOS	SLES15 SP2
Slingshot	1.4.0
NVIDIA Driver	450.156.00
PE	21.10
NVIDIA HPC SDK	21.9
NVIDIA CUDA	11.0

# Polaris Single Node Configuration

# of AMD EPYC 7532 CPUs	1
# of NVIDIA A100 GPUs	4
Total HBM2 Memory	160 GB
HBM2 Memory BW per GPU	1.6 TB/s
Total DDR4 Memory	512 GB
DDR4 Memory BW	204.8 GB/s
# OF NVMe SSDs	2
Total NVMe SSD Capacity	3.2 TB
# of Cassini NICs	2
Total Injection BW (w/ Cassini)	50 GB/s
PCIe Gen4 BW	64 GB/s
NVLink BW	600 GB/s
Total GPU DP Tensor Core Flops	78 TF



# Slingshot Configuration



- 11 Total dragonfly groups, 10 compute groups and 1 non-compute group
- 2 links/arc between each group
- 4 links/arc within each group (between switches of a group)
- 1 link from each NIC (100Gb with SS10, 200Gb when upgraded to SS11)

# Single AMD EPYC “MILAN” 7543P CPU Specs

Base Frequency	2.8 GHz
Max Boost Clk	3.7 GHz
# of Zen3 Cores	32
# of Threads	64
Total DDR4 Memory	512 GB
# of Memory Channels	8
DDR4 Memory BW	204.8 GB/s
Total Shared L3 Cache	256 MB
L2 Cache per Core	512 KB
L1 Cache per Core	32 KB
PCIe Gen 4	128 lanes (8 ports)
PCIe Gen4 BW	64 GB/s
TDP	225 W



# NVIDIA HGX A100 Specs

	A100 PCIe	HGX
FP64	9.7 TF	38.8 TF
FP64 Tensor Core	19.5 TF	78 TF
FP32	19.5 TF	78 TF
BF16 Tensor Core	312 TF	1.3 PF
FP16 Tensor Core	312 TF	1.3 PF
INT8 Tensor Core	624 TOPS	2496 TOPS
GPU Memory	40 GB HBM2	160 GB HBM2
GPU Memory BW	1.6 TB/s	6.4 TB/s
Interconnect	PCIe Gen4 64 GB/s	NVLink 600 GB/s
Max TDP Power	250W	400W

Ampere 7nm

A100 PCIe

HGX A100 4-GPU

# Node Local Storage

- Each compute node has two NVMe SSDs
  - 1.6 TB each / 3.2 TB total
- Similar to Theta, ALCF provides no specific software for using SSDs
- Each volume will be mounted as an ext4/xfs volume that is user accessible
- Users access SSD via standard POSIX APIs
- Data is destroyed when the job ends so any data users wish to keep must be moved to Grand or Eagle

# Acceptance Testing

# Acceptance Phases

- Requirements
  - Defect categorization
  - Root cause analysis of all failures
  - Every run shall obtain a correct result
  - Approval of Argonne Technical Representative
- Functional & Performance Tests (ATP-FP)
  - Applications
  - System
  - Performance
- Stability (ATP-S)
  - 95% Availability
  - 90% Utilization
  - 24 hours of no job failures related to system software or hardware
  - Mix of job sizes, applications, runtimes

# ATP-FP Applications

- Requirements
  - Defect categorization
  - Root cause analysis of all failures
  - Every run shall obtain a correct result
  - Approval of Argonne Technical Representative
- Functional & Performance Tests (ATP-FP)
  - **Applications**
  - System
  - Performance
- Stability (ATP-S)
  - 95% Availability
  - 90% Utilization
  - 24 hours of no job failures related to system software or hardware
  - Mix of job sizes, applications, runtimes

- QMCPACK
- LAMMPS
- NekBench/NekBone
- HACC
- CosmicTagger
- Uno
- HPL-AI
- OvO
- SOLLVE VV

# ATP-FP System

- Requirements
  - Defect categorization
  - Root cause analysis of all failures
  - Every run shall obtain a correct result
  - Approval of Argonne Technical Representative
- Functional & Performance Tests (ATP-FP)
  - Applications
  - **System**
  - Performance
- Stability (ATP-S)
  - 95% Availability
  - 90% Utilization
  - 24 hours of no job failures related to system software or hardware
  - Mix of job sizes, applications, runtimes

- BOM Validation
- Cold/Warm Boot
- RAS
  - Logging
  - System Metrics
  - Fault Tolerance
  - NHC
- Software Environment
  - CNOS
  - Compilers
  - Debuggers
  - Profilers / Tuners
  - Data/Learning Frameworks
- IOR
- System Security

# ATP-FP Performance

- Requirements
  - Defect categorization
  - Root cause analysis of all failures
  - Every run shall obtain a correct result
  - Approval of Argonne Technical Representative
- Functional & Performance Tests (ATP-FP)
  - Applications
  - System
  - **Performance**
- Stability (ATP-S)
  - 95% Availability
  - 90% Utilization
  - 24 hours of no job failures related to system software or hardware
  - Mix of job sizes, applications, runtimes

- HPL
- DGEMM
- STREAM
- MPI
- IOR
- IO-500

# Polaris Results



# Functional & Performance

# Acceptance Test Checklist (ATC)

- ATC provides the details of the specific test cases that were run for the Acceptance Functional and Performance period
  - Application and Benchmark cases used for Stability test as well
- Defines 131 test cases
- Defines tests with both functional and performance criteria
- Performance projections were made using ThetaGPU as a baseline
  - Identical NVIDIA A100 40GB GPU
  - AMD CPUs
  - Eight GPUs per node instead of 4
  - Fat Tree Infiniband network instead of Slingshot
- ATC provided documentation of projections and FOMs along with information on how to build and run
- ATC allowed for 30% margin of error on performance projections

# Application and Benchmark Descriptions

- QMCPACK
  - Open source quantum Monte Carlo package for *ab initio* electronic structure calculations
  - Weak scaling from 1 to 560 nodes, MPI + OpenMP, C++
- LAMMPS
  - Classical molecular dynamics code with a focus on materials modeling
  - Weak scaling from 1 to 560 nodes, MPI + Kokkos, C++
- NekBench
  - Proxy application for NekRS (previously NEK5000)
    - Navier Stokes solver based on the spectral element method
  - Strong scaling from 32 to 512 nodes, MPI + OCCA, C++
- HACC
  - Extreme-scale cosmological simulation code
  - Weak scaling from 1 to 560 nodes, MPI + CUDA, C++
- Cosmic Tagger
  - Remove background particles by applying semantic segmentation on full detector images from the SBND detector via deep learning
  - Weak scaling from 1 to 512, Python + PyTorch + mpi4py, using Nvidia container
- Uno
  - Predict drug response to cure cancer cells via machine/deep learning
  - Weak scaling on 1 node with 1 and 7 instances (via MIG) on a single GPU, Python + Keras + Tensorflow
- OvO
  - Collection of OpenMP Offloading test functions for C++ and Fortran
- SOLLVE VV
  - OpenMP Validation and Verification project is a suite of test cases to validate conformance and correctness for an OpenMP 4.5/5.0
  - C/C++

# Applications

- Acceptance Test Report contains details of all test case results and metrics against targets
- All application tests had correctness checks and all returned correct results

Application	Configs	Passed?	Notes
QMCPACK	7	Yes	~90% of target, issue with cusolver that is pending evaluation after CUDA Driver upgrade
LAMMPS	11	Yes	Single node test case below target
NekBench	5	Yes	Strong scaling test
HACC	11	Yes	
Cosmic Tagger	5	Partial	3 test cases below performance targets, 1 test case (512 nodes) fails to initialize. Currently under investigation.
Uno	2	Yes	

# Benchmarks

Application	Configs	Passed?	Notes
HPL-AI	1	Yes	#8 on November 21 HPL-AI list
HPL	1	Yes	77% of peak at 26.6 PF
GEMM	2	Yes	All GPUs within 95% of peak
STREAM	1	Yes	All GPUs within 95% of peak
MPI	3	Yes	Global injection and bisection bandwidth > 13 TB/s
IOR	6	Partial	Segmentation faults when running > 16 ranks per node. Performance of single-shared-file on read below requirements.
OvO	1	Yes	NVIDIA compilers 79% successful completion
SOLVVE VV	1	Yes	4.5 >90% success rate

# System

- Variety of operational test cases to validate all aspects of Polaris

Application	Tests	Passed?	Notes
BOM Validation	2	Yes	
Cold/Warm Reboot	3	No	Full System Reboot exceed required time
Software Environment	10	Yes	Issue with STAT tool, gdb4hpc working
PBS and Batch Jobs	21	Yes	
Compute Node OS	7	Yes	
RAS / power redundancy	11	Yes	
Network	6	Yes	

# Exceptions

- All exceptions for test cases documented in Acceptance Test Report
- Argonne elected to not execute 14 ATC test cases during ATP-FP
  - MPICH & Network
    - (2) tests deferred for Slingshot-11 testing
    - (1) occurred during acceptance but no record captured
  - IO-500
    - test deferred waiting on remaining gateway nodes to be connected
  - CNOS
    - (2) tests deferred until post acceptance because additional configuration required
    - (2) tests not needed as running standard SLES Linux rather than COS
  - Cold/Warm reboot tests
    - (2) login and gateway not done as too disruptive
      - These node types rebooted many times as part of Polaris standup
  - WLM tests
    - (1) requires additional work outside of Polaris project
    - (1) test unable to implement
  - RAS
    - (1) test deferred until post acceptance because additional configuration required
    - (1) RDHX failure occurred during bring up but no record captured

# Stability



# Overview of Stability

- Demonstrate the reliability of the system
  - Reproducibly generate correct results
  - Manage hardware and software errors

- Requirements

- 21 contiguous days
- 95% availability
- 90% load
- 24 hours of no job failures related to system software or hardware

Availability defined as:  $\frac{\sum_i^N (S_i - D_i)}{\sum_i^N S_i}$

$S_i$  is number of schedulable hours for node  $i$

$D_i$  is the number of hours of downtime for node  $i$

- Variety of tests from the FP phase
  - 8 different applications/benchmarks
  - 78 different problem configurations

- QMCPACK
- LAMMPS
- NekBench/NekBone
- HACC
- CosmicTagger
- Uno
- BabelStream
- D/SGEMM

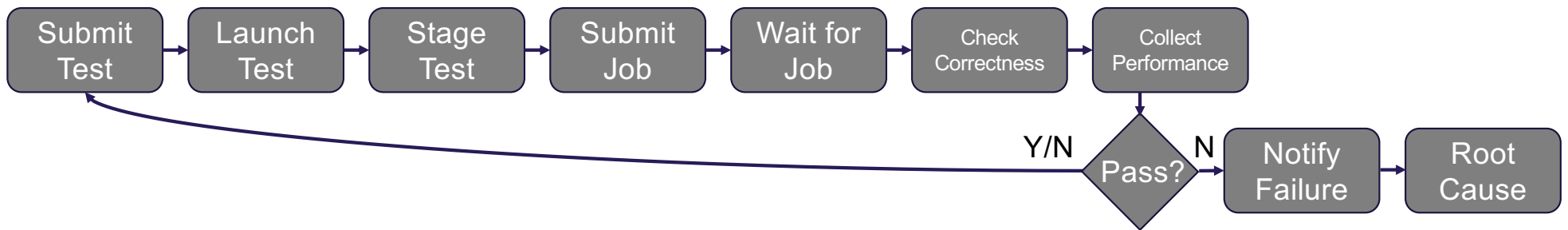
# Polaris Test Harness

# Test Harness

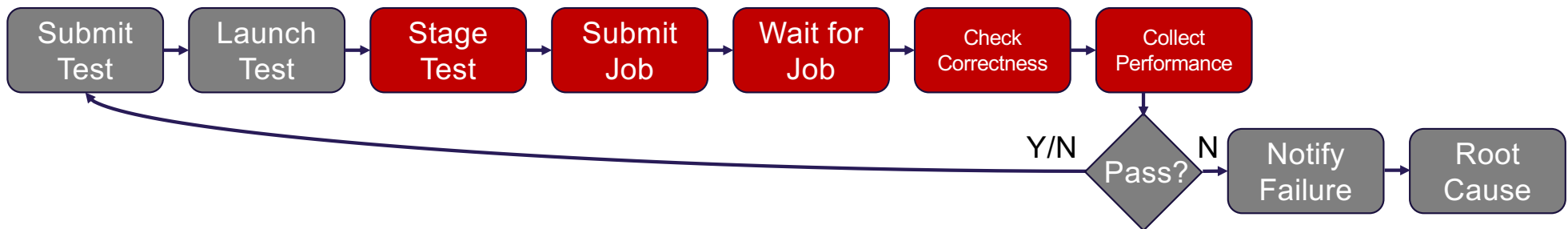
- Automated system to manage the requirements of the stability acceptance testing
  - Continuous job submission
  - Test correctness checking
  - Test performance collection
  - Failure reporting

App/Benchmark	Configurations	Scale (Nodes)
CosmicTagger	8	1 – 128
HACC	20	1 – 560
LAMMPS	22	1 – 560
NekBench	10	32 – 512
QMCPack	10	2 – 560
Uno	1	1
BabelStream	5	1
D/SGEMM	2	1

# Test Harness Pipeline and Software

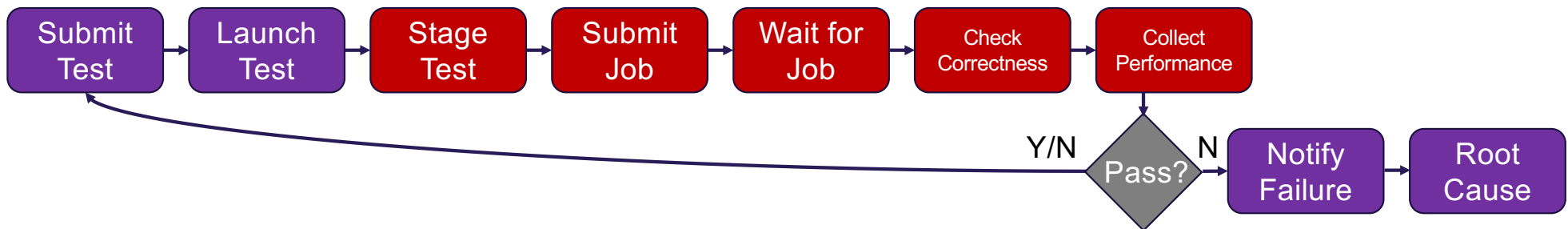


# Test Harness Pipeline and Software



- **ReFrame** is a high-level framework for writing regression tests for HPC systems which provides abstractions for writing sanity and performance checks. Our test harness utilizes ReFrame to:
  - Stage tests
  - Submit jobs
  - Check correctness
  - Collect performance

# Test Harness Pipeline and Software



- **ReFrame** is a high-level framework for writing regression tests for HPC systems which provides abstractions for writing sanity and performance checks. Our test harness utilizes ReFrame to:
  - Stage tests
  - Submit jobs
  - Check correctness
  - Collect performance
- **Jenkins** is an open source automation server to enable CI/CD while offering a vast library of plugins. Our test harness utilizes Jenkins to:
  - Continuously create tests
  - Provide web console for root cause analysis
  - Enable/disable tests
  - Notify on failure

Jenkins is going to shut down  
Shut down reason: ATP-S Complete

### Project HAC\_C002\_Node

Full project name: Polaris/HAC\_C002\_Node  
HAC\_C 2 node test

Job Owners

- Workspace
- Recent Changes
- Disk Usage
  - Job
  - All builds
  - Locked builds
  - All workspaces
  - Slave workspaces
  - Non-slave workspaces

MTTR	Last 7 Days
	Last 30 Day
	All Time
MTTF	Last 7 Days
	Last 30 Day
	All Time
Standard Deviation	Last 7 Days
	Last 30 Day
	All Time

- Up
- Status
- Changes
- Workspace
- Build Now
- Configure
- Delete Project
- Rebuild Last
- Favorite
- Manage Ownership
- Dependency Graph
- Job Config History
- Rename
- Jenkins Lint
- Set Next Build Number

Build History trend

Filter builds...

#1324	(pending—Jenkins is about to shut down)
#1323	Mar 23, 2022, 3:26 PM 387 KB
#1322	Mar 23, 2022, 3:07 PM 387 KB

Jenkins is going to shut down  
Shut down reason: ATP-S Complete

### Polaris AT


Folder name: Polaris  
Polaris AT

Folder Owners

S	W	Name ↓	Last Success	Last Failure	Last Duration	Built On	Fav
✓	⚙️	BabelStream_134217728	12 days - #9219	15 days - #7747	2 min 35 sec	JenkinsAT on Polaris-login1	☆
✓	⚙️	BabelStream_268435456	12 days - #9155	15 days - #7724	2 min 5 sec	JenkinsAT on Polaris-login2	☆
✓	⚙️	BabelStream_33554432	12 days - #9332	15 days - #7837	2 min 27 sec	JenkinsAT on Polaris-login2	☆
✓	⚙️	BabelStream_536870912	12 days - #8935	15 days - #7566	2 min 13 sec	JenkinsAT on Polaris-login2	☆
✓	⚙️	BabelStream_67108864	12 days - #9262	15 days - #7759	2 min 58 sec	JenkinsAT on Polaris-login1	☆
✓	⚙️	CosmicTagger_001	12 days - #2331	15 days - #1867	9 min 36 sec	JenkinsAT on Polaris-login2	☆
✓	⚙️	CosmicTagger_002	12 days - #1458	N/A	10 min	JenkinsAT on Polaris-login2	☆
✓	⚙️	CosmicTagger_004	12 days - #989	N/A	40 min	JenkinsAT on Polaris-login2	☆
✓	⚙️	CosmicTagger_008	12 days - #604	N/A	53 min	JenkinsAT on Polaris-login2	☆
✓	⚙️	CosmicTagger_016	12 days - #407	15 days - #317	3 hr 19 min	JenkinsAT on Polaris-login2	☆

- Build Queue (78)
- Build Executor Status

# Failure Root Causing – First Step

 jenkins APP 11:49 AM  
Polaris AT » LAMMPS\_256\_Node\_Long - #65 Failure after 19 hr (Open)

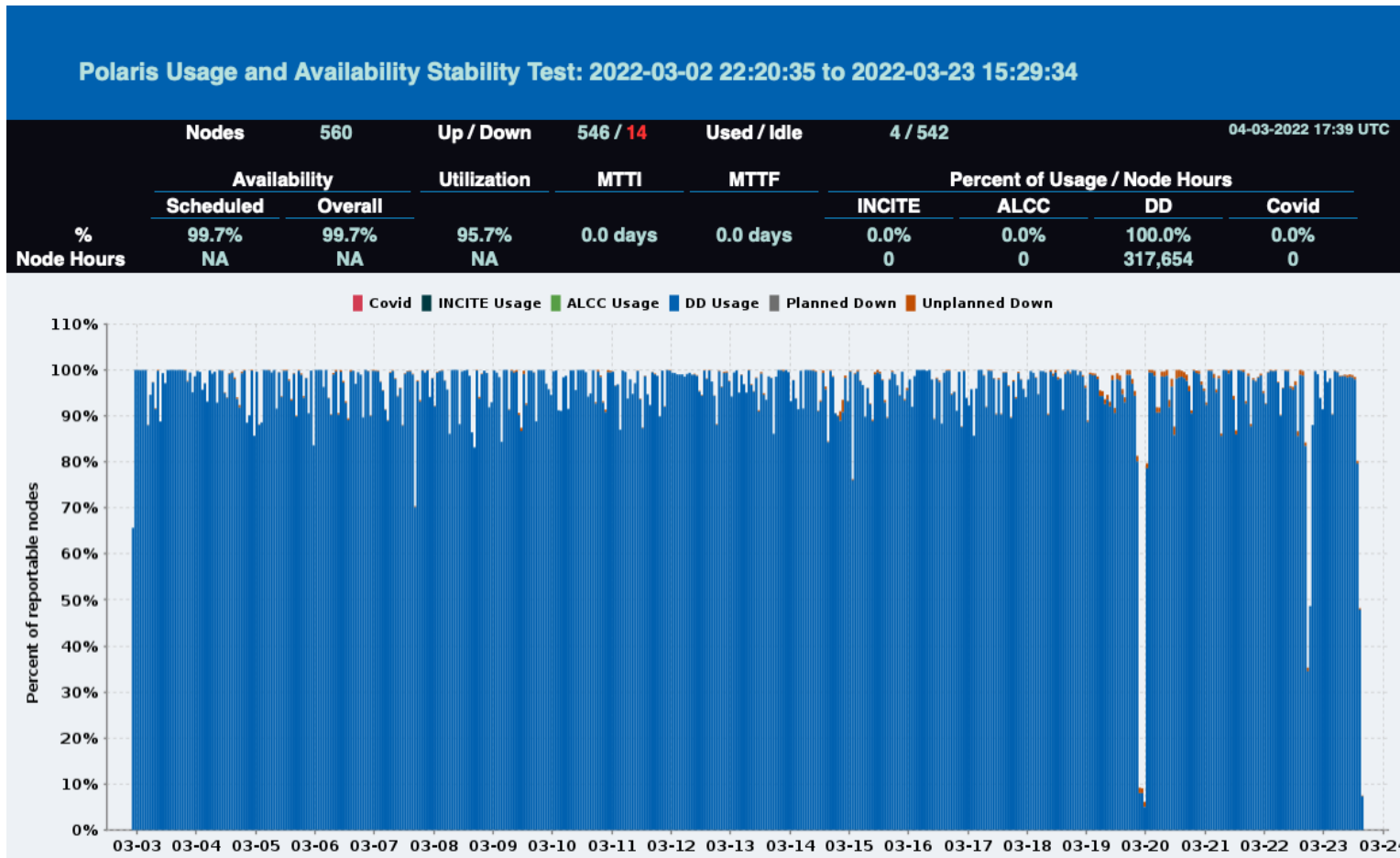
## On Failure:

- stdout & stderr are captured in console
- Jenkins notifies via Slack
- Initial investigator checks output through web console
- Additional investigation proceeds as needed

```
Dashboard » Polaris AT » LAMMPS_256_Node_Long » #65
16:49:15 MPII_wait(202)..... MPI_wait(request=0x7fff60397fc8, status=0x1) failed
16:49:15 aborting job:
16:49:15 Fatal error in PMPI_Wait: Other MPI error, error stack:
16:49:15 MPII_wait(202)..... MPI_wait(request=0x7fff60397fc8, status=0x1) failed
16:49:15 MPII_wait(93).....
16:49:15 MPII_wait_impl(41).....
16:49:15 MPII_progress_wait(186).....
16:49:15 MPII_progress_test(80).....
16:49:15 MPII_OFI_handle_cq_error(1059): OFI poll failed (ofi_events.c:1061:MPII_OFI_handle_cq_error:Input/output error - transport retry counter exceeded)
16:49:15 Kokkos::Cuda ERROR: Failed to call Kokkos::Cuda::finalize()
16:49:15 x3106c0s7b0n0.hsn0.cm.polaris.alcf.anl.gov: rank 943 exited with code 255
16:49:15 MPICH ERROR [Rank 32] [job id 56b585be-1d0d-4d71-af5c-305175540ac8] [Wed Mar 23 16:45:56 2022] [x3006c0s31b1n0] - Abort(137003663) (rank 32 in comm 0): Fatal error in PMPI_Wait: Other MPI error, error stack:
16:49:15 MPII_wait(202)..... MPI_wait(request=0x7ffda89c0468, status=0x1) failed
16:49:15 MPII_wait(93).....
16:49:15 MPII_wait_impl(41).....
16:49:15 MPII_progress_wait(186).....
16:49:15 MPII_progress_test(80).....
16:49:15 MPII_OFI_handle_cq_error(1059): OFI poll failed (ofi_events.c:1061:MPII_OFI_handle_cq_error:Input/output error - transport retry counter exceeded)
16:49:15 aborting job:
16:49:15 Fatal error in PMPI_Wait: Other MPI error, error stack:
16:49:15 MPII_wait(202)..... MPI_wait(request=0x7ffda89c0468, status=0x1) failed
16:49:15 MPII_wait(93).....
16:49:15 MPII_wait_impl(41).....
16:49:15 MPII_progress_wait(186).....
16:49:15 MPII_progress_test(80).....
16:49:15 MPII_OFI_handle_cq_error(1059): OFI poll failed (ofi_events.c:1061:MPII_OFI_handle_cq_error:Input/output error - transport retry counter exceeded)
16:49:15 Kokkos::Cuda ERROR: Failed to call Kokkos::Cuda::finalize()
16:49:15 + echo lammps/256_128_128_256_6300_Steps/65
16:49:15 POST BUILD TASK : SUCCESS
16:49:15 END OF POST BUILD TASK : 0
16:49:15 Started calculate disk usage of build
16:49:15 Finished Calculation of disk usage of build in 0 seconds
16:49:15 Started calculate disk usage of workspace
16:49:15 Finished Calculation of disk usage of workspace in 0 seconds
16:49:16 [Slack Notifications] found #64 as previous completed, non-aborted build
16:49:16 [Slack Notifications] will send OnEveryFailureNotification because build matches and user preferences allow it
16:49:16 Finished: FAILURE
```



# Usage and Availability Report



# Stability Results

Stability was completed in March.

**Start:** Wed Mar 2 22:20:35 UTC 2022

**End:** Wed Mar 23 15:29:34 UTC 2022

## High level results:

- 99,381 Jobs run
- 146 Failures
- 99.7% Availability achieved
- 95.7% Utilization achieved
- 6 Distinct 24+ hour periods without job failures due to system

# Stability Results

Stability was completed in March.

**Start:** Wed Mar 2 22:20:35 UTC 2022

**End:** Wed Mar 23 15:29:34 UTC 2022

## High level results:

- 99,381 Jobs run
- 146 Failures
- 99.7% Availability achieved
- 95.7% Utilization achieved
- 6 Distinct 24+ hour periods without job failures due to system

## Requirements:

- 21 contiguous days
- 95% availability
- 90% load
- 24 hour no failures

# Stability Results

Stability was completed in March.

**Start:** Wed Mar 2 22:20:35 UTC 2022

**End:** Wed Mar 23 15:29:34 UTC 2022

## High level results:

- 99,381 Jobs run
- 146 Failures
- 99.7% Availability achieved
- 95.7% Utilization achieved
- 6 Distinct 24+ hour periods without job failures due to system

## Requirements:

- ✓ • 21 contiguous days
  - Ended early on Argonne's discretion as metrics were met for full 21 day period
- 95% availability
- 90% load
- 24 hour no failures

# Stability Results

Stability was completed in March.

**Start:** Wed Mar 2 22:20:35 UTC 2022

**End:** Wed Mar 23 15:29:34 UTC 2022

## High level results:

- 99,381 Jobs run
- 146 Failures
- 99.7% Availability achieved
- 95.7% Utilization achieved
- 6 Distinct 24+ hour periods without job failures due to system

## Requirements:

- ✓ • 21 contiguous days
  - Ended early on Argonne's discretion as metrics were met for full 21 day period
- ✓ • 95% availability
  - 99.7% achieved
- 90% load
- 24 hour no failures

# Stability Results

Stability was completed in March.

**Start:** Wed Mar 2 22:20:35 UTC 2022

**End:** Wed Mar 23 15:29:34 UTC 2022

## High level results:

- 99,381 Jobs run
- 146 Failures
- 99.7% Availability achieved
- 95.7% Utilization achieved
- 6 Distinct 24+ hour periods without job failures due to system

## Requirements:

- ✓ • 21 contiguous days
  - Ended early on Argonne's discretion as metrics were met for full 21 day period
- ✓ • 95% availability
  - 99.7% achieved
- ✓ • 90% load
  - 95.7% achieved
- 24 hour no failures

# Stability Results

Stability was completed in March.

**Start:** Wed Mar 2 22:20:35 UTC 2022

**End:** Wed Mar 23 15:29:34 UTC 2022

## High level results:

- 99,381 Jobs run
- 146 Failures
- 99.7% Availability achieved
- 95.7% Utilization achieved
- 6 Distinct 24+ hour periods without job failures due to system

## Requirements:

- ✓ • 21 contiguous days
  - Ended early on Argonne's discretion as metrics were met for full 21 day period
- ✓ • 95% availability
  - 99.7% achieved
- ✓ • 90% load
  - 95.7% achieved
- ✓ • 24 hour no failures
  - 6 distinct 24+ hour periods

# Failures List

App/Benchmark	Runs	Failures
BabelStream	45903	3
CosmicTagger	6359	1
D/SGEMM	16404	0
HACC	6989	13
LAMMPS	9598	36
NekBench	1837	12
QMCPack	2947	21
Uno	374	1

Failure Type	Count
Hardware	32
Human Error	2
Performance	8
Walltime	43
Software	1
Nek	1



# Polaris Upgrade



# Polaris Upgrade

- Polaris began life with AMD “Rome” CPUs.
  - Initial delivery and acceptance was completed with the “Rome” CPUs.
- Polaris design was to use AMD “Milan” processors.
  - After initial acceptance testing, Polaris was upgraded to the “Milan” CPUs.
- The test harness enabled the easy shakeout of failing nodes during the upgrade.
- A second three day stability acceptance was completed for the CPU upgrade.

# Thank you