Experience using Slurm on ARIS HPC System

Nikos Nikoloutsakos

GRNET
Greek Research and Technology Network, Greece

hpc.gnret.gr

27 September 2016

Overview

Who we are

Greek Research Technology Network

GRNET enables researchers from Greece to obtain access to the powerful national High Performance Computing system ARIS.

Advanced Research Information System

ARIS Infrastructure provides state-of-the-art supercomputing capabilities for large-scale scientific applications.

GRNET provides services to:

- Greece - Greek Academic Community
  - Greek Universities
  - Technological institutions
  - Research centers
- Europe
  - PRACE (Tier 1 system)
  - DECI
  - other EU Projects (Vi-seem, Eudat, EGI,...)

Adiministrative and Application support

Support Team HPC provides:

- Management of Infrastructure
- User Support
  - Comprehensive end-user support
  - User support in operational problems
  - Documentation
  - Educational and Training Events
- Application Support - Transfer and optimizing application
- Peer-Review support and coordination

Open Access

Peer-Review Access

The criteria for the evaluation:

- Scientific Excellence
- Impact of the proposed research
- The need for HPC resources
- Maturity and experience of the principal investigator and his/her team
- Feasibility of the project based on a technical evaluation and the availability of resources

Project Types

Preparatory-Development Projects

Execution of scalability tests, performance tests, resolve issues. Code porting, development, optimization.

- Review: Technical only
- Call: Always Open
- Access: 2-4 months

Production Projects

Projects that have the technical expertise to take advantage of available resources and are selected by the procedure of peer review

- Review: Technical-Scientific
- Periodic Call - 2 per year
- Access: 1 year

since June 2015

First pilot operational phase in June 2015

- 150 projects
- 400 Users
- 24 Organizations
- 300 software modules
- 120.000 jobs submitted,
  46M core hours (1 year)
- 25 scientific publications (up to now)
  https://hpc.grnet.gr/results-publications/

Compute Power: 180 TFlops (HPL) #465 Top500 - iteration June 2015

ARIS - Compute Nodes I

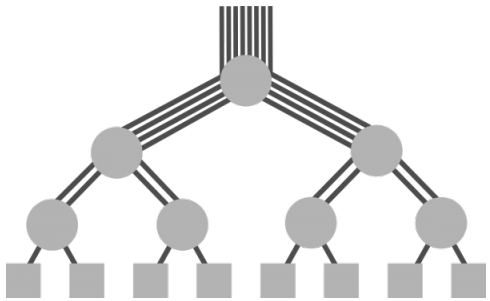
- 426 compute nodes: IBM NextScale n360 M4
- 8520 cores: 2x (Intel E5 2680v2@2.8Ghz - 10 core) per node
- 27TB total memory: 64GB memory per node (8 RDIMMS, 1866 MHz)
- Half-width, 1U systems grouped in 6U enclosures (12 nodes per enclosure)

ARIS - Compute Nodes II

- 6 Racks, 6 enclosures per rack.
- Diskless
- IBM 1PB GPFS, Tape Library IBM TS3500 6PB
- Max nominal power consumption: 162 KW (154 KW on HPL). 183 KW with air-cooling.

ARIS - Network

- Mellanox SX6536 648-Port Infiniband Director Switch
- FDR 56 Gbits / sec
- Fat tree non-blocking mode
- 450 QSFP+Optical cables
- 5 Km fabric cables



ARIS - Expansion

- 44 gpu nodes: “2 x NVIDIA Tesla k40m” accelerated nodes.
  - Dell Power Edge R730
  - 2 x Intel Xeon E5-2660v3@2.6GHz
  - 64 GB RAM
- 18 phi nodes: “2 x INTEL Xeon Phi 7120p” accelerated nodes.
  - Dell Power Edge R730
  - 2 x Intel Xeon E5-2660v3@2.6GHz
  - 64 GB RAM
- 44 fat nodes
  - Dell PowerEdge R820
  - 4x Intel Xeon E5-4650v2@2.4GHz
  - 512 GB RAM
- IBM 1PB GPFS,
- Tape Library IBM TS3500 6PB

ARIS - Managment

- 14 support nodes, NextScale x3650 M4 2 x E5-2640v2
- 2x Managment Nodes, 2x Login Nodes, 10x service nodes
- Monitoring software xCAT, Nagios, Ganglia, BMS (Business Management System) Dell OpenManage, MRTG
- Scheduler SLURM 14.11.8
- XDMoD, UMGMT (User Managment Tool) in house

Partition Queues

ONE cluster "ARIS"

Partition	Description	Nodes
compute	Thin nodes	426
gpu	GPU nodes	44
phi	PHI nodes	18
fat	FAT nodes	24
taskp	Serial queue	20

Default timelimit 2 days

Configurtion parameters I

Consumable Resources

- `SelectTypeParameters= CR_CORE_MEMORY`
- Shared mode - unless user specifies `--exclusive`

Resource Limits

- AccountingStorageEnforce = associations,limits,safe

Generic Resource (GRES) Scheduling

- `GresTypes = gpu,mic`
- mic offload mode only

Configurtion parameters II

MpiDefault = pmi2

Supports MPI implementation being used on system: IntelImpi,OpenMPI, mvapich2

The larger the job, the greater its job size priority.

`PriorityFavorSmall=NO`

Accounting Gather

- `AcctGatherEnergyType=acct_gather_energy/ipmi`
- `AcctGatherInfinibandType=acct_gather_infiniband/ofed`
- `JobAcctGatherType = jobacct_gather/linux`

Priority Flags I

Multifactor Priority

- PriorityType= priority/multifactor
- PriorityWeightAge = 100
- PriorityWeightFairShare = 1000
- PriorityWeightJobSize = 1000
- PriorityWeightPartition = 0
- `PriorityDecayHalfLife = 00:00:00`
- `PriorityUsageResetPeriod = WEEKLY`
- `PriorityMaxAge = 30-00:00:00`
- `PriorityWeightQOS=0`

Priority Flags II

Fair Tree Fairshare

- `PriorityFlags = FAIR_TREE`
- `PriorityCalcPeriod = 02:00:00`

Backfill Scheduling

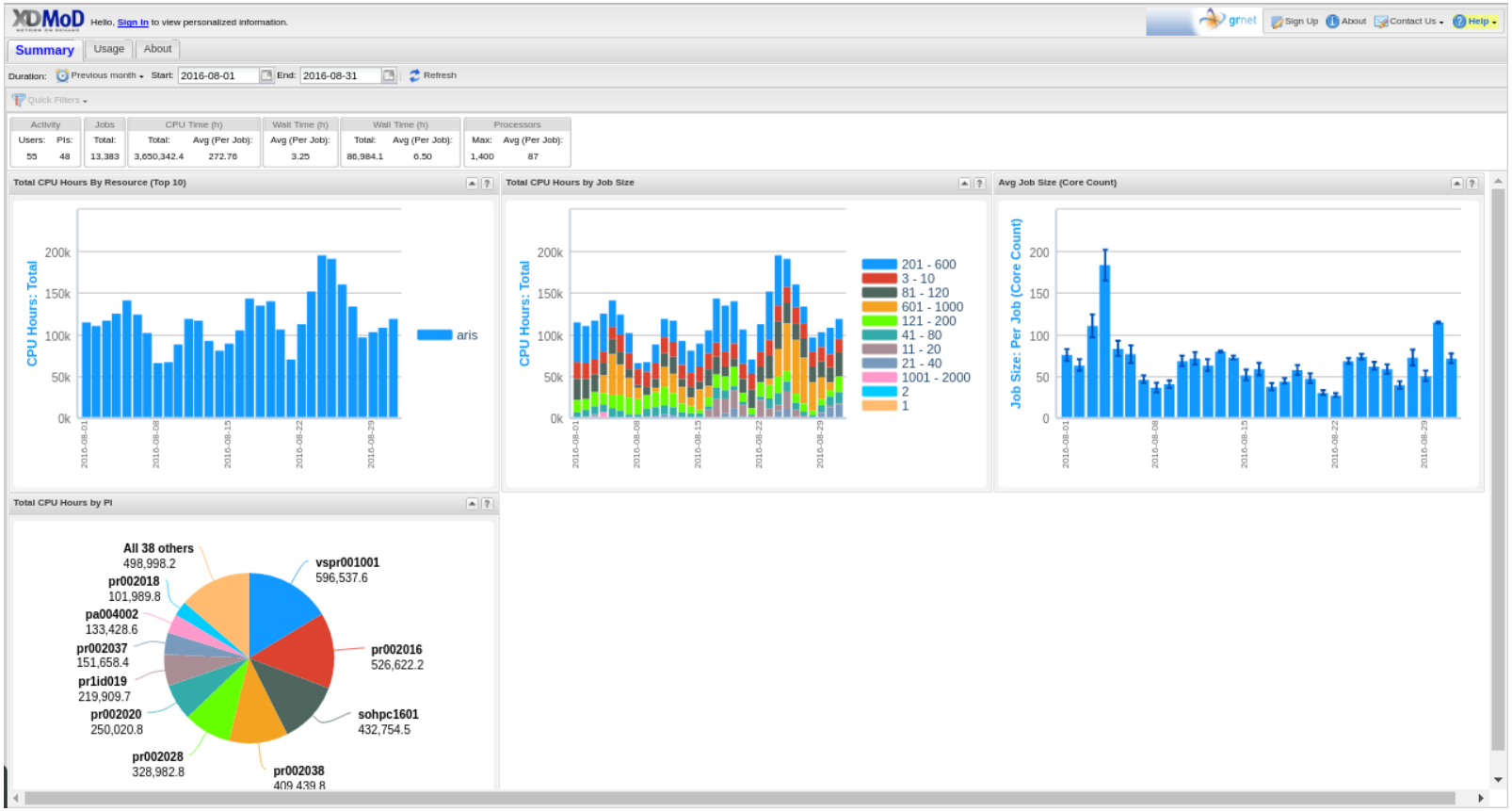
- `SchedulerType= sched/backfill`

$mybudget

```
==========================================================================
Core Hours Allocation Information for account : testproj
==========================================================================
Allocated Core Hours              :        10000000.00
Project Consumed  Core Hours      :         3410968.00
User    Consumed  Core Hours      :             523.00
Percentage of Project Consumed    :              34.11
Percentage of User    Consumed    :               0.01
Account limits (Job,Node,Core)    :     0    0      0
==========================================================================
```

$myreport

```
Time reported in CPU Hours
--------------------------------------------------------------------------------
  Cluster         Account     Login     Proper Name       Used     Energy
--------- --------------- --------- --------------- ---------- ----------
     aris     testproj_ nikolout+ Nikos Nikolout+        371        384
```



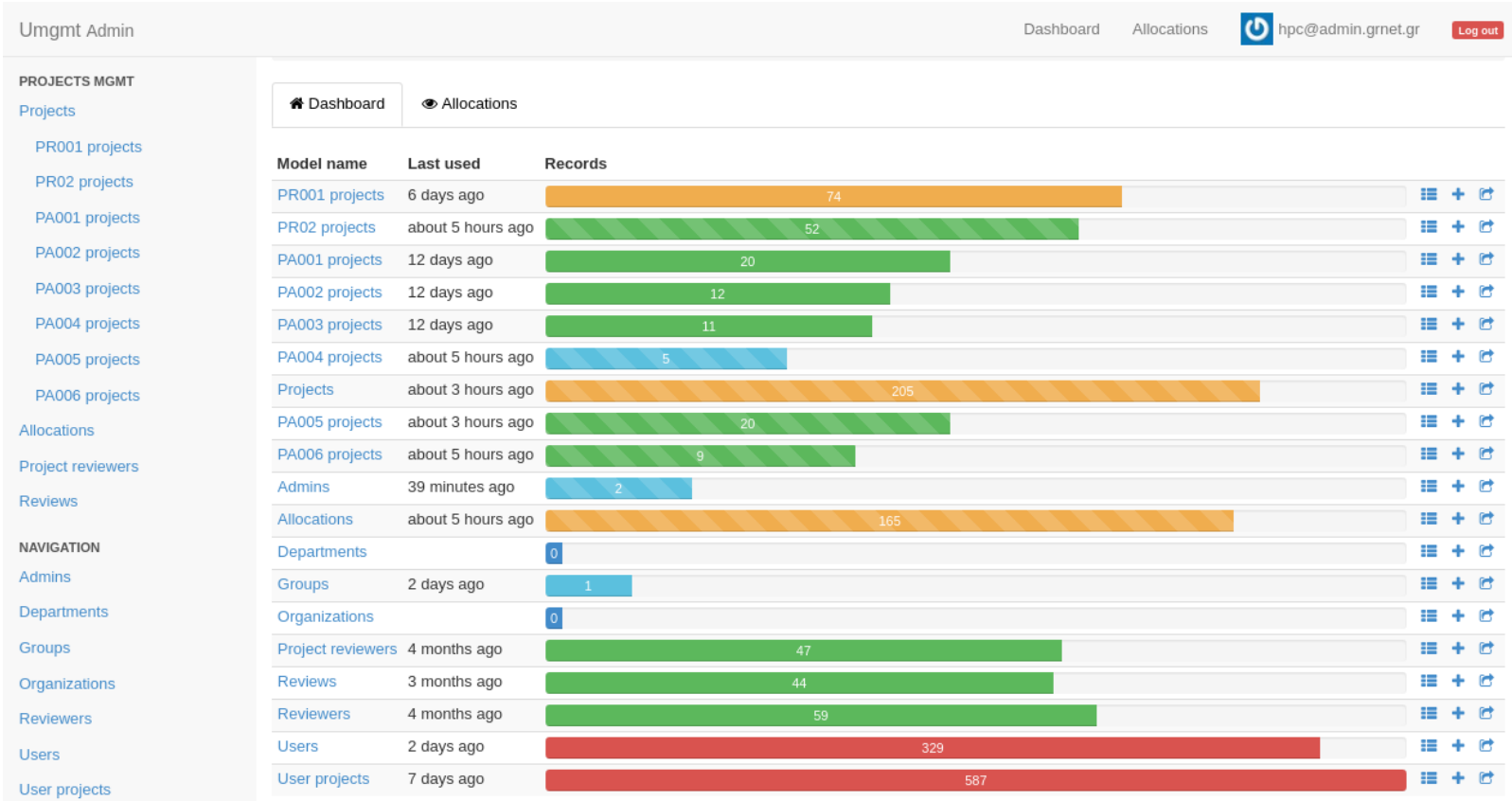
UMGMT I

*U*sers Management Tool

- Tool to manage project proposals and user access on the system.
- Associate project proposals to slurm accounting information
- Keep Track start end dates per project, Extensions: core hours-access period
- Project status , send alert emails to users
- Statistics consumed core-hours(%) per project

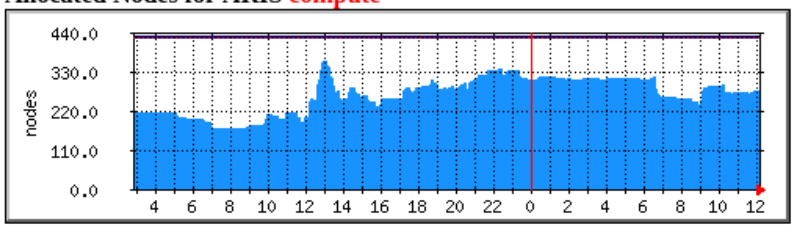
in development: Ruby on Rails

UMGMT II

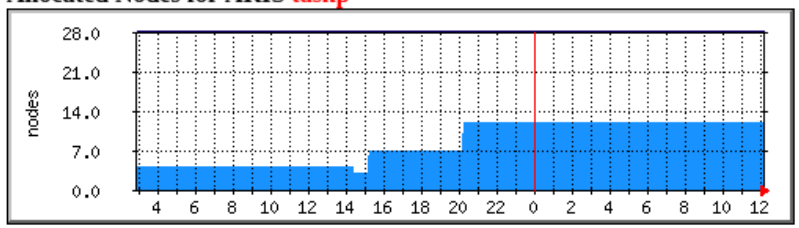

Umgmt Admin
Dashboard
Allocations
hpc@admin.grnet.gr
Log out
PROJECTS MGMT
Projects
PR001 projects
PR02 projects
PA001 projects
PA002 projects
PA003 projects
PA004 projects
PA005 projects
PA006 projects
Allocations
Project reviewers
Reviews
NAVIGATION
Admins
Departments
Groups
Organizations
Reviewers
Users
User projects
Dashboard
Allocations
Model name
Last used
Records
PR001 projects 6 days ago 74
PR02 projects about 5 hours ago 52
PA001 projects 12 days ago 20
PA002 projects 12 days ago 12
PA003 projects 12 days ago 11
PA004 projects about 5 hours ago 5
Projects about 3 hours ago 205
PA005 projects about 3 hours ago 20
PA006 projects about 5 hours ago 9
Admins 39 minutes ago 2
Allocations about 5 hours ago 165
Departments 0
Groups 2 days ago 1
Organizations 0
Project reviewers 4 months ago 47
Reviews 3 months ago 44
Reviewers 4 months ago 59
Users 2 days ago 329
User projects 7 days ago 587

SLURM - MRTG

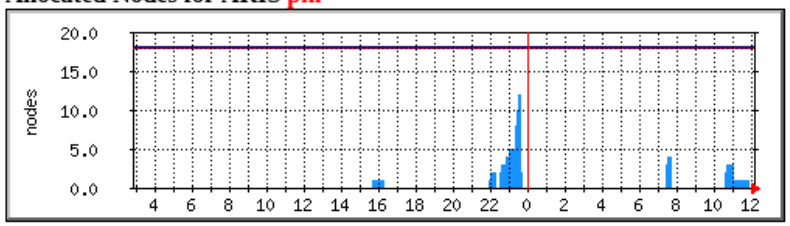
Allocated Nodes for ARIS compute



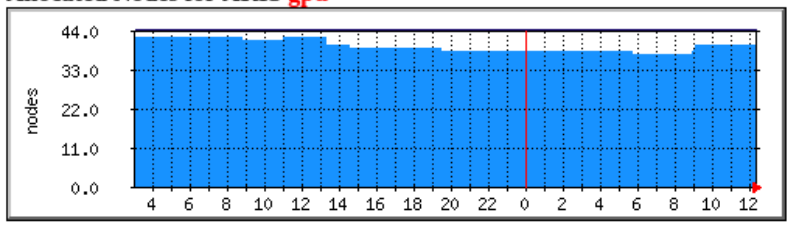
Allocated Nodes for ARIS taskp



Allocated Nodes for ARIS phi



Allocated Nodes for ARIS gpu



Slurm Script Template

Helps users prepare batch job scripts for Slurm at ARIS.

Job name:	jobname
Total number of tasks **(across all nodes)**:	20
Total number of nodes:	1
Tasks per node:	20
Threads per task:	1
Memory per node:	56 GB
Walltime: (Hours:Minutes:Seconds)	01 HH 00 MM 00 SS
Partition:	compute
Account:	pr0000

Acknowledgment BYU Job Script Generator
https://github.com/BYUHPC/BYUJobScriptGenerator

Issues

- Problem:
  Reservation (daily) had 20 nodes , 15 where active , 5 where active by same user but for other job
  1 node (from 15) died, unable to reschedule.

Feature request

- More verbose error messages:
  Users could figure why a job is rejected.
  More information about which limit violated
- MPI Task 0: may need more memory
  Ability to specify less processes on first node.
- Allocation per GRES(gpu,mic) not only cpu ch

What's Next

- upgrade to version 16

Thank you !