Γενική Χρήση του ARIS

Δρ. Δημήτρης Ντελλής

GRNET

ntell [at] grnet.gr

Περιεχόμενα

- Software Environment
  - Environment Modules
  - Διαθέσιμα πακέτα
- Batch System
  - Job Submission
  - Job Control
  - Accounting

Environment Modules. Τι είναι ?

- Το πακέτο Environment Modules κάνει δυναμική τροποποίηση του περιβάλλοντος χρήστη μέσω των module files.
- Κύριες μεταβλητές περιβάλλοντος που προσαρμόζονται είναι οι PATH, MANPATH, και LD_LIBRARY_PATH, αλλά και μεταβλητές περιβάλλοντος που ενδεχομένως κάθε πακέτο λογισμικού χρειάζεται.
- Κάθε module file περιέχει την πληροφορία που χρειάζεται ώστε να ρυθμίσει τις μεταβλητές περιβάλλοντος για κάποια εφαρμογή.

- Όλα τα modules θέτουν μια μεταβλητή MODULENAMEROOT. Σε modules που αναφέρονται σε βιβλιοθήκες, συνήθως τα include files βρίσκονται στην $MODULENAMEROOT/include και οι βιβλιοθήκες στην $MODULENAMEROOT/lib
- Εάν υπάρχουν εξαρτήσεις ενός πακέτου λογισμικού από άλλα τα οποία επίσης ρυθμίζονται με module file, οι εξαρτήσεις αυτές μπορούν να περιγραφούν και εφόσον το αντίστοιχο module δεν είναι ενεργό είτε το φορτώνει είτε βγάζει μήνυμα λάθους ειδοποιώντας το χρήστη ότι πρέπει πρώτα να φορτώσει τις εξαρτήσεις.
- Σε περιπτώσεις πακέτων τα οποία υπάρχουν σε πάνω από μια έκδοση, υπάρχει ένα module για κάθε έκδοση και ο administrator μπορεί να ορίσει κάποια ως default.

Environment Modules. Χρήση

- Έλεγχος πακέτων που είναι διαθέσιμα μέσω modules

  ```
  module avail
  ```

  ή

  ```
  module -l avail
  ```

- Έλεγχος ενεργών modules

  ```
  module list
  ```

- Απενεργοποίηση όλων των ενεργών modules

  ```
  module purge
  ```

- Απενεργοποίηση συγκεκριμένου module

  ```
  module unload MODULENAME
  ```

- Αλλαγή έκδοσης module

  `module switch MODULENAME/VER1 MODULENAME/VER2`

- Πληροφορίες για το τι κάνει κάποιο module

  `module whatis MODULENAME/VERSION`

- Κείμενο Βοήθειας για κάποιο module

  `module help MODULENAME/VERSION`

- Για να δείτε τι κάνει η ενεργοποίηση ενός module

  `module show MODULENAME/VERSION`

- Default version ενός module
  - Όπως θα δείτε παρακάτω, σχεδόν όλα τα πακέτα που υπάρχουν στο ARIS σε πάνω από μια version έχουν μια από αυτές επισημασμένη ως default. Στην περίπτωση αυτή, οι εντολές

    ```
    module load MODULENAME
    ```

    και

    ```
    module load MODULENAME/DEFAULTVERSION
    ```

    είναι ισοδύναμες.

Διαθέσιμα πακέτα

- Compilers/Debugers
- MPI Implementations
- Libraries
- Applications
- Debuggers/Profilers
- Graphics

Compilers

- Intel Compiler Suite : 15.0.3 (default), 16.0.0 (2 seats License)
- GNU Compiler Suite : 4.9.2(default), 4.9.3, 5.1.0, 5.2.0

Debuggers

- gdb 7.9.1
- Intel gdb 15.0.3, 16.0.0
- ddd

MPI

- Intel MPI 5.0.3 (default), 5.1.1
- OpenMPI 1.10.0, for GNU and Intel
- MVAPICH2 2.2.2a (experimental) for GNU και Intel

Σημειώσεις για τον IntelMPI

- Οι wrappers mpicc/mpicxx/mpif90 του IntelMPI χρησιμοποιούν GNU compilers
- Υπάρχουν οι αντίστοιχοι wrappers (και headers/libraries) για Intel Compilers mpiicc/mpiicpc/mpiifort.

Σημειώσεις για τον MVAPICH2

- Η χρήση του mvapich2 υποστηρίζεται MONO μέσω του srun => δεν υπάρχει mpirun, mpiexec κλπ.

Profilers

- gprof
- mpiP
- Scalasca
- Intel VTune

Βιβλιοθήκες

```
atlas/3.11.34
boost/1.58.0
cgnslib/3.2.1/intel
elpa/2015.05.001/intel
fftw/2.1.5
fftw/3.3.4/avx
fftw/3.3.4/sse2
flame/5.0/gnu
flame/5.0/intel
glpk/4.55
gsl/1.16/gnu
hdf5/1.8.12/gnu
hdf5/1.8.12/intel
hdf5/1.8.15/gnu
hdf5/1.8.15/intel
jasper/1.900.1
libint/1.1.5
libjpeg-turbo/1.4.1
libsmm/gnu
libsmm/intel
libxc/2.2.2
med/3.0.8/intel
metis/5.1.0
netcdf/3.6.3/intel
netcdf/4.1.3/gnu
netcdf/4.1.3/intel
netcdf-c/4.3.3.1/gnu
netcdf-c/4.3.3.1/intel
netcdf-combined/4.3.3.1/intel
netcdf-fortran/4.4.2/gnu
netcdf-fortran/4.4.2/intel
openblas/0.2.14/gnu/int4
openblas/0.2.14/gnu/int8
openblas/0.2.14/intel/int4
openblas/0.2.14/intel/int8
parmetis/4.0.3/gnu
parmetis/4.0.3/intel
pnetcdf/1.6.1/gnu
pnetcdf/1.6.1/intel
scalapack/2.0.2/gnu
scalapack/2.0.2/intel
szip/2.1
udunits2/2.2.19
voro++/0.4.6
```

Εφαρμογές

```
abinit/7.10.4
bigdft/1.7.6
cdo/1.7.0
code_saturne/4.0.1/intel
cp2k/2.6.1
cpmd/4.1
dlpoly/2.20
dlpoly/4.07
gamess-US/2014R1
gopenmol/3.00
gromacs/4.5.7
gromacs/4.6.7
gromacs/5.0.5
gromacs/5.0.6
gromacs/5.1
lammps/15May15
mdynamix/5.2.7
molden/5.2
molekel/5.4.0
mpqc/2.3.1
namd/2.10/hybrid/memopt
namd/2.10/hybrid/normal
namd/2.10/purempi/memopt
namd/2.10/purempi/normal
ncarg/6.3.0
ncview/2.1.5
nwchem/6.5
octave/4.0.0
octopus/4.1.2
openbabel/2.3.2
openmd/2.2
paraview/4.3
qhull/2012.1
quantum-espresso/5.2.0
R/3.2.1
towhee/7.1.0
vmd/1.9.2
wrf/3.4.1/hybrid
wrf/3.4.1/purempi
wrf/3.7/hybrid
wrf/3.7/purempi
wrf-chem/3.7
wrf-chem/3.7-hybrid
```

Εφαρμογές

- Κάποιες εφαρμογές που απαιτούν κάποιου είδους άδεια, υπάρχουν εγκατεστημένες, για τους σκοπους της εκπαίδευσης είναι προσβάσιμες, αλλά υπο κανονικές συνθήκες θα έχουν πρόσβαση μόνο όσοι προσκομίσουν την άδεια που έχουν. π.χ. cpmd, dl_poly, gamessUS.
- Εκτός από τα πακέτα που είναι διαθέσιμα μέσω modules υπάρχουν και πακέτα που ενδεχομένως είναι χρήσιμα από το σύστημα.
  - Gnuplot
  - Grace
  - Gimp

Μερικές σημειώσεις για compilers/MPI/Βιβλιοθήκες/Εφαρμογές

- Οι βιβλιοθήκες μεταγλωτίστηκαν με gnu/4.9.2 και intel/15.0.3 όπου υπάρχει λόγος.
- Η ανάγκη για πολλαπλές versions (gnu/Intel) βιβλιοθηκών προέρχεται από το γεγονός ότι περιέχουν Fortran 90 modules τα οποία δεν είναι χρησιμοποιήσιμα από άλλο compiler.
- Οι εφαρμογές μεταγλωτίστηκαν με gnu/4.9.2 ή και intel/15.0.3, intelmpi/5.0.3
- Όσες βιβλιοθήκες δεν περιέχουν F90 modules και η ταχύτητά τους δεν επηρεάζεται σημαντικά από τους compilers μεταγλωτίστηκαν με GNU/4.9.2

Batch System

- Τι είναι ένα Batch System
  - Ένα Batch System ελέγχει την πρόσβαση στους διαθέσιμους υπολογιστικούς πόρους ώστε όλοι οι χρήστες να μπορούν να χρησιμοποιούν το σύστημα - Συνήθως σε ένα σύστημα υπάρχει μεγαλύτερη ζήτηση για πόρους από τους διαθέσιμους.
  - Δίνει τη δυνατότητα στο χρήστη να προδιαγράψει μια υπολογιστική εργασία (Job) , να την υποβάλει στο σύστημα και να αποσυνδεθεί από αυτό.
  - Η εργασία θα εκτελεστεί όταν υπάρχουν πόροι (cores, nodes, μνήμη) και χρόνος
- Διαδεδομένα Batch systems
- SLURM, PBS/Torque, LSF, LoadLeveler, SGE/OGE.

- ARIS Batch System : SLURM, υποστηρίζεται PBS emulation

Όταν μια εργασία υποβάλεται σε ένα Batch system :

- Περιγράφονται οι πόροι που χρειάζεται το σύστημα (π.χ. cores, nodes, μνήμη, χρόνος εκτέλεσης
- Το σύστημα κατάγράφει τους πόρους που ζητήθηκαν
- Όταν βρεθούν οι διαθέσιμοι πόροι, ξεκινάει η εκτέλεση της εργασίας
- Οι πόροι μπορούν να χρησιμοποιηθούν όπως θέλει ο χρήστης

SLURM Scripts

Ένα SLURM Script περιγράφει τους πόρους που χρειάζεται για να τρέξει η εργασία, όπως επίσης τις εντολές εκτέλεσης της εργασίας.

Παρατηρήστε τους 2 τρόπους που μπορούν να περιγραφούν οι απαιτήσεις της εργασίας π.χ.

```
--nodes=200
```

και

```
-A sept2015
```

.

SLURM Scripts

```
#!/bin/bash
#SBATCH --job-name="testSlurm"   # Όνομα για διαχωρισμό μεταξύ jobs
#SBATCH --error=job.err.%j    # Filename για το stderr
#SBATCH --output=job.out.%j   # Filename για το stdout
                              # Το %j παίρνει την τιμή του JobID
#SBATCH --nodes=200           # Αριθμός nodes
#SBATCH --ntasks=400          # Αριθμός MPI Tasks
#SBATCH --ntasks-per-node=2   # Αριθμός MPI Tasks / node
#SBATCH --cpus-per-task=10    # Αριθμός Threads / MPI Task
#SBATCH --mem=56G             # Μνήμη ανά node
#SBATCH --mem-per-cpu=2800M   # Μνήμη ανά core
#SBATCH -A sept2015           # Accounting tag (sept2015 για
                              #  όλους στο training)
#SBATCH -t 01:00:00           # Ζητούμενος χρόνος HH:MM:SS
#SBATCH -p compute            # partition, default στο ARIS.
module purge
module load gnu
module load intel
module load intelmpi
export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK

srun EXECUTABLE ARGUMENTS
```

SLURM Scripts

- Το script του προηγούμενου slide είναι η πλήρης περιγραφή μιας εργασίας.
- Μπορεί να υποβληθεί εργασία και με λιγότερα από τα #SBATCH directives
  - Δίνοντας μόνο το --nodes χωρίς το --ntasks το σύστημα μπορεί να υπολογίσει πόσα tasks θα χρησιμοποιήσει
  - Αντίστοιχα, δίνοντας μόνο το --ntasks το σύστημα μπορεί να υπολογίσει πόσα nodes χρειάζεται.
  - Τα υποχρεωτικά που σχετίζονται με τον αριθμό των cores που θα χρησιμοποιήσει μια εργασία είναι ένα από τα παραπάνω
  - Παραλείποντας το --name, το σύστημα το θέτει ίδιο με το όνομα του script.
  - Παραλείποντας το --output το σύστημα το θέτει σε slurm-JOB_ID.out
  - Υποχρεωτική είναι η χρήση του --account (ή -A)
  - Θέτοντας όλες τις μεταβλητές έχετε πλήρη έλεγχο του τι πόρους ζητάτε από το σύστημα.

Χρήση **srun** για την εκτέλεση των εφαρμογών

- Οι εκδόσεις του MPI έχουν η κάθε μια ένα mpirun/mpiexec κλπ.
- Προτείνεται να χρησιμοποιείται το srun για την εκτέλεση παράλληλων εργασιών.
- Κάποιοι από τους λόγους
  - Το srun ξεκινάει τα εκτελέσιμα σε όλους τους κόμβους οπότε έχει πιο πλήρη έλεγχο.
  - Το srun κάνει accounting κατανάλωσης ρεύματος, χρήση Infiniband, χρήση δίσκων, κλπ.
  - Είναι κοινός τρόπος για τις (2 προς στιγμήν) εκδόσεις MPI που υπάρχουν στο ARIS
  - Σε κάποιες περιπτώσεις, χρησιμοποιώντας mpiexec.hydra με Intelmpi έχουμε κάπως πιο γρήγορη εκτέλεση των εφαρμογών.
  - Αντιθέτως, σε περιπτώσεις που η εφαρμογή έχει προβλήματα και χρειαστεί να σταματήσει ίσως να παρουσιαστούν προβλήματα (zombie procs) στη χρήση του scancel.

Επικοινωνία με το SLURM

- Υποβολή εργασίας
  ```
  sbatch SLURM_JobScript.sh
  Submitted batch job 15242
  ```
- Κατάλογος εργασιών
  ```
  squeue
  ```
- Κατάλογος εργασιών με περισότερες λεπτομέρειες
  ```
  squeue -o "%.8i %.9P %.10j %.10u %.8T %.5C
  %.4D %.6m %.10l %.10M %.10L %.16R"
  ```
- Ακύρωση εργασίας
  ```
  scancel JobID
  ```
- Σε κάποιες περιπτώσεις που τα εκτελέσιμα δεν τερματίζονται άμεσα παίρνοντας SIGHUP από το SLURM
  ```
  scancel -s KILL JobID
  ```
- Εκτίμηση του πότε θα αρχίσει η εκτέλεση των εργασιών που είναι σε αναμονή για πόρους
  ```
  squeue --start
  ```
- Πληροφορίες για την τρέχουσα χρήση
  ```
  sinfo
  ```

SLURM jobs dependency

- Εάν μια εργασία για να αρχίσει πρέπει κάποια άλλη να έχει ήδη αρχίσει ή τελειώσει, στο SLURM Script εκτός των άλλων :

  ```
  #SBATCH --dependency=after:Job_ID
  ```

  ή

  ```
  #SBATCH --dependency=afterok:Job_ID
  ```

  αντίστοιχα
- Εάν πρέπει μια εργασία να ξεκινήσει κάποιο συγκεκριμένο χρονικό διάστημα, στο SLURM Script εκτός των άλλων :
  - Έναρξη στις 16:00

    ```
    #SBATCH --begin=16:00
    ```
  - Έναρξη συγκεκριμένη ημέρα και ώρα :

    ```
    #SBATCH --begin=2015-09-14T16:30:00
    ```

Εάν κάποια εργασία δεν τρέχει και στο nodelist/REASON εμφανίζονται τιμές εκτός από nodenames ή Resources, τότε έχουμε ζητήσει περισότερους πόρους από ότι μας επιτρέπεται

- `AssocMaxNodesPerJobLimit`
  Ζητάμε περισότερα nodes από ότι επιτρέπεται στο account μας
- `AssocMaxWallDur`
  Ζητάμε περισότερο χρόνο από ότι επιτρέπεται στο account μας
- Διάφοροι άλλοι λόγοι που εάν από το όνομα δεν είναι αντιληπτό, ανατρέξετε στο documentation του SLURM.

SLURM Environment Variables

Όταν ξεκινάει η εργασία το SLURM βάζει κάποιες μεταβλητές που σχετίζονται με αυτή, και ενδεχομένως είναι χρήσιμες στον χρήστη.

```
$SLURM_NNODES                 # Αριθμός nodes
$SLURM_NTASKS                 # Αριθμός Tasks
$SLURM_NPROCS                 #  ”   ”    ”
$SLURM_NTASKS_PER_NODE        # Αριθμός Tasks /node
$SLURM_TASKS_PER_NODE         #  ”   ”    ”    ”
$SLURM_CPUS_PER_TASK          # Αριθμός threads / Task
$SLURM_MEM_PER_NODE           # Μνήμη / node (MB)
```

SLURM User/Group resource limits

- Στο SLURM κάθε χρήστης έχει κάποια όρια πόρων που μπορεί να ζητήσει/χρησιμοποιήσει. Αυτά είναι :
  - Αριμός Jobs που μπορούν να εκτελούνται ταυτόχρονα
  - Αριθμός Jobs που μπορούν να εκτελούνται ή να βρίσκονται σε αναμονή
  - Μέγιστη χρονική διάρκεια εκτέλεσης ενός Job
  - Μέγιστος αριθμός nodes που μπορεί να ζητήσει ένα Job
  - Μέγιστος αριθμός cores που μπορεί να ζητήσει ένα Job
  - Συνολικός αριθμός core hours στη διάρκεια ενός project.

- Αντίστοιχα ένα group χρηστών μπορεί να έχει όρια που αφορούν όλους τους χρήστες είτε ως όρια, π.χ. Max Walltime, είτε ως άθροισμα π.χ. core hours όλων των χρηστών του group.

- Όρια χρήσης για την πρακτική εξάσκηση :

```
Max Nodes                    4
Max Cores                   80
Max WallTime              1:00:00
Total Core Hours/user       1200
```

- Ο Scheduler στο ARIS είναι FIFO with Backfill. Αυτό σημαίνει
  - Το job που υποβλήθηκε πρώτο θα εκτελεστεί πρώτο
  - Από τη στιγμή που ξεκινάει η εκτέλεση, η εργασία θα τελειώσει το αργότερο μετά από όσο χρόνο ζητήθηκε στο SLURM script.
  - Εάν το σύστημα έχει μεν ελευθερους πόρους (cores/nodes/memory) αλλά δεν είναι αρκετοί για να τρέξει το πρώτο στη σειρά από τα queued, τα επόμενα jobs θα περιμένουν
  - εκτός...

- Κάποιο από τα επόμενα jobs ζητάει πόρους που υπάρχουν, και ο χρόνος εκτέλεσης που ζητάει είναι μικρότερος από τον πιο κοντινό αναμενόμενο χρόνο τέλους των jobs που εκτελούνται. Αυτό το job θα παρακάμψει τη σειρά, και θα εκτελεστεί πρώτο χωρίς να προκαλέσει καμιά καθυστέρηση σε άλλα jobs.
- Έτσι το σύστημα έχει τη μεγαλύτερη δυνατή χρήση.
- Ζητήστε λίγο παραπάνω από όσο χρόνο υπολογίζετε ότι χρειάζεται η εργασία σας και όχι το μέγιστο που μπορείτε

Εξομοίωση PBS

- Υπάρχει εγκατηστημένη η εξομοίωση του PBS/Torque. Χρήστες που είναι εξοικοιωμένοι στη χρήση PBS μπορούν να χρησιμοποιήσουν τα PBS scripts και εντολές.
- Η εξοιμοίωση του PBS καλύπτει μεγάλο βαθμό περιγραφής εργασιών, αλλά όχι όλα

SLURM

sbatch
squeue
scancel

```
#!/bin/sh
#SBATCH --mem-per-cpu=2G
#SBATCH -t 1:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=20
#SBATCH -A sept2015
#SBATCH -p compute
....
```

PBS

qsub
qstat
qdel

```
#!/bin/sh
#PBS -l pvmem=2G
#PBS -l walltime=1:00:00
#PBS -l nodes=1:ppn=20
#PBS -A sept2015
#PBS -q compute
....
```

Accounting

- Δείτε τα jobs σας το τρέχον 24ωρο
  sacct
- Δείτε τα jobs σας τον τελευταίο μήνα
  sacct -S 2015-08-14
- Δείτε πόσο χρόνο (και ενέργεια σε Wh εφόσον χρησιμοποιείτε srun) έχετε καταναλώσει το τελευταίο εξάμηνο
  myreport
- Δείτε πόσο από τον χρόνο που σας έχει δοθεί έχετε χρησιμοποιήσει
  mybudget

Ερωτήσεις ?

Θέματα για την πρακτική εξάσκηση

- Εξερευνήστε τα modules, load/switch/unload
- Ετοιμάσετε SLURM job scripts, δοκιμάστε υποβολή, ακύρωση, εξαρτήσεις και γενικά ότι ταιριάζει στην καθημερινή απασχόλησή σας.
- Δοκιμάστε λάθη στο SLURM script και δείτε τη συμπεριφορά του συστήματος.
- Σημείωση : Επειδή υπάρχει μόνο 2 seats άδεια για Intel Compilers, προτιμήστε τους GNU compilers σε όλες τις πρακτικές προς αποφυγή καθυστερήσεων.
- Εφόσον δεν γνωρίζετε κάποια απο τις εφαρμογές που είναι εγκατεστημένες, ή δεν έχετε κάποιο μικρό κώδικα να κάνετε compile, κάνετε δοκιμές με srun hostname