

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

# Γενική Χρήση του ARIS

Δρ. Δημήτρης Ντελλής

GRNET

ntell [at] grnet.gr

## Περιεχόμενα

- Σύνδεση στο σύστημα
- Σύστημα αρχείων
- Software Environment
  - Environment Modules
  - Διαθέσιμα πακέτα
- Batch System
- Βέλτιστες Πρακτικές Χρήσης - Συνηθισμένα Λάθη/Προβλήματα.
- Συζήτηση με ομάδες που είχαν/έχουν/ενδιαφέρονται να έχουν πρόσβαση στο σύστημα.

## ● Σύνδεση στο σύστημα

- Δύο από τους κόμβους υπηρεσιών έχουν διαμορφωθεί σε login nodes
- Πανομοιότυπη εγκατάσταση, κοινός λογαριασμοί χρηστών, κοινή πρόσβαση στο GPFS (/users, /work και /work2)
- Διεύθυνση και για τους 2 login nodes:  
**login.aris.grnet.gr**
- Σύνδεση SSH με χρήση κλειδιού **MONO**.
- Εάν δείτε Password prompt, δεν έχετε χρησιμοποιήσει το σωστό private key.
- Ζεύγος κλειδιών, public και private.
- Το public ssh key αποθηκεύεται στον ssh server (login node στην περίπτωση μας) στο Home του χρήστη.
- Το private ssh key βρίσκεται στον ssh client (π.χ. το laptop σας) και είναι **μυστικό!** Μόνο ο ιδιοκτήτης του πρέπει να έχει πρόσβαση σε αυτό.
- Το private ssh key μπορεί προαιρετικά να προστατεύεται με ένα passphrase.

## ● Λογισμικό SSH Client

- MacOS, Linux : OpenSSH, συνήθως υπάρχει εγκατεστημένο.  
Για γραφικό περιβάλλον : `ssh -X username@login.aris.grnet.gr`
  - **ssh** : SSH client, με αυτό θα συνδεθείτε
  - **ssh-keygen**: Δημιουργία, μετατροπή κλειδιών
  - **scp, sftp**: Μεταφορά αρχείων
- Windows: PuTTY
  - **PuTTY** : SSH client, με αυτό θα συνδεθείτε
  - **PuTTYgen** : Δημιουργία, μετατροπή κλειδιών
  - **PSCP, PSFTP** : Μεταφορά αρχείων
- Windows: Bitvise (δωρεάν, με πολλές γραφικές διευκολύνσεις)

- Δημιουργία ζεύγους ssh κλειδιών σε MacOS, Linux
  - **ssh-keygen -t rsa -b 2048**
  - public key: `.ssh/id_rsa.pub`
  - private key: `.ssh/id_rsa`
- Μεταφορές αρχείων πίσω στον/στους υπολογιστές σας:
  - Δεν είναι απαραίτητο να συνδεθείτε από το ARIS στους υπολογιστές σας και να δώσετε **put**.
  - Μπορεί να γίνει συνδεδόμενοι από τους υπολογιστές σας στο ARIS δίνοντας **get**.
  - Οι SSH συνδέσεις από το ARIS προς οπουδήποτε **ΔΕΝ** επιτρέπονται.

## ● X Server for windows

- Χρήσιμος για εκτέλεση διάφορων εφαρμογών όπως profilers, graphics κλπ.
- Xming X Server for Windows
- <http://sourceforge.net/projects/xming/>
- Για να δουλέψει, απαραίτητο να έχετε ενεργό το "Enable X11 forwarding"
- Το Xming πρέπει να τρέχει στο Windows PC σας πριν ξεκινήσετε κάποια γραφική εφαρμογή στο [login.aris.grnet.gr](http://login.aris.grnet.gr)

- Τι χρειάζεται να μας στείλετε για να αποκτήσετε πρόσβαση στο ARIS, μετά την έγκριση του project.
  - Επιθυμητό username - να αποφεύγονται usernames του τύπου john, dimitris κλπ.
  - Το ssh public key σας
  - Η διεύθυνση ή οι διευθύνσεις IP από όπου θα συνδέεστε, ώστε να επιτραπεί η πρόσβαση στο firewall του ARIS
  - Παραδείγματα
    - Η IP του υπολογιστή σας στο εργαστήριο σας
    - Εναλλακτικά αν δεν έχετε σταθερό IP, ολόκληρο το υποδίκτυο του εργαστηρίου σας
    - Οι IP διευθύνσεις του VPN service του ιδρύματός σας

## ● Σύστημα αρχείων GPFS

- GPFS 4.1
- 4 filesystems : /users /work /work2 και /staging
- /users
  - Περίπου 240 TB
  - Applications
  - Home directories των χρηστών
  - Δεν πρέπει να εκτελούνται (τουλάχιστον I/O intensive) jobs στο Home
  - Μακροχρόνια αποθήκευση
- /work και /work2
  - Περίπου 440 + 400 TB
  - Για κάθε χρήστη υπάρχει η μεταβλητή \$WORKDIR, που καθορίζει που θα είναι η work dir του κάθε χρήστη
  - Εδώ πρέπει να εκτελούνται τα jobs
  - Βραχυχρόνια αποθήκευση



## ● /staging

- Περίπου 150 TB
- Χώρος για μακροχρόνια αποθήκευση (πραγματική αποθήκευση σε tapes).
- Μεγάλα αρχεία : > 10 MB.
- Αν και φαίνεται να υπάρχουν τα αρχεία, πρακτικά βρίσκονται στο tape, κάθε προσπάθεια προσπέλασης ενεργοποιεί το μηχανισμό επαναφοράς από το tape : Χρονοβόρο : Βάζετε στο χώρο αυτό μόνο αρχεία που είναι για πραγματικά "αρχειοθέτηση"
- Τι μπορεί να σημαίνει αρχειοθέτηση ?  
Έγινε κάποιο post-processing των δεδομένων, βγήκαν κάποια αποτελέσματα, στάλθηκε ένα paper, και ο/οι reviewer/s ζητάνε κάτι επιπλέον, οπότε θα ξαναχρειαστούν. Αν δεν ξαναχρειαστούν, τα σβήνουμε και ελευθερώνεται ο χώρος στα tapes.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Δυνατότητα χρήσης κατόπιν αίτησης - έγκρισης. Η μεταβλητή \$ARCHIVEDIR δίνει το path για τον κάθε χρήση.

## Environment Modules. Τι είναι ?

- Χρησιμοποιώντας εφαρμογές που δεν προέρχονται από το σύστημα, χρειάζεται να ρυθμιστούν PATH, LD\_LIBRARY\_PATH και διάφορες άλλες μεταβλητές περιβάλλοντος για τη λειτουργία των εφαρμογών.
- Συνήθης πρακτική να ρυθμίζονται αυτές οι μεταβλητές είτε γενικά σε κάποιο σύστημα που τρέχει μερικές μόνο εφαρμογές, είτε στο .bashrc του κάθε χρήστη.
- Η κατάσταση περιπλέκεται περισσότερο με την ύπαρξη πάνω της μιας versions του ίδιου πακέτου, οι μεταβλητές των οποίων εξαρτώνται από άλλες μεταβλητές.

## 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 αφορά 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

- 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.

- Π.χ. με τα τρέχοντα defaults, τα

```
module load intel
```

και

```
module load intel/15.0.3
```

είναι ισοδύναμα.



## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Είναι σύνηθες σε συστήματα αυτού του τύπου, μετά από ειδοποίηση σε εύλογο χρονικό διάστημα πριν την ενεργοποίηση, να γίνει αλλαγή των defaults του συστήματος. Π.χ. Σε σύντομο χρονικό διάστημα θα αλλάξουν στο ARIS οι default versions. Ενδεικτικά, gnu compilers από 4.9.2 σε 6.2.0, intel compilers από 15.0.3 σε 17.0.0, cr2k από 2.6.1 σε 4.1 κλπ.
- Μετά από τέτοιες αλλαγές, και εφόσον χρησιμοποιείτε τα defaults, συνίσταται να ξανακάνετε compile τους δικούς σας κώδικες.
- Εάν χρειάζεται να χρησιμοποιείτε συγκεκριμένη version ενός πακέτου, συνίσταται να χρησιμοποιείται και η version του.

## Batch System

- Σύνηθες :
  - Έρχεται ΣΚ, τρέχει ένα run σε όλο το node, θα τελειώσει Σάββατο χαράματα. Βάλτε άλλο ένα να τρέχει ταυτόχρονα για κάποιο διάστημα, μέχρι Κυριακή πρωί, με ότι αυτό συνεπάγεται : Χρήση swar κλπ. γενικά ελλατωμένη απόδοση του συστήματος.
  - Από Κυριακή μεσημέρι μέχρι Δευτέρα πρωί το σύστημα "κάθεται".
  - Ο Χ χρήστης τρέχει αρκετά runs, ας βάλω ένα και εγώ να πάρω κάποιο μέρος του συστήματος στο επόμενο διάστημα, με ότι αυτό συνεπάγεται.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Θα πάω διακοπές 15 μέρες, βάζω όσα runs υπολογίζω για 15 μέρες (όπως πιθανότατα σκέφτονται και οι υπόλοιποι χρήστες του συστήματος).
- Πόσοι (που δεν χρησιμοποιούν Batch System) δεν αντιμετώπισαν τέτοια θέματα ?

## 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
- 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.
- Εγγυάται ότι το κάθε run θα έχει πλήρη και **αποκλειστική** πρόσβαση στους πόρους που ζήτησε, π.χ. μνήμη, cores, κλπ.
- Μπορώ να στείλω π.χ. 1000 runs, τα οποία θα εκτελεστούν χωρίς ταυτόχρονη εκτέλεση στα ίδια resources (μνήμη, cores).

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Αν κάποιος άλλος χρήστης στείλει run θα πάρει και αυτός το αναλογούν ποσοστό resources χωρίς επικάλυψη.
- Οι πόροι μπορούν να χρησιμοποιηθούν όπως θέλει ο χρήστης
  - Ένα π.χ. MPI run (Η κύρια/προτεινόμενη χρήση)
  - Πολλά σειριακά runs : Αν και μπορεί να χρησιμοποιηθεί με αυτό τον τρόπο, ένα run δεν κερδίζει κάτι από την ύπαρξη π.χ. Infiniband. Ίσως η χρήση της υποδομής Grid : [www.hellasgrid.gr](http://www.hellasgrid.gr) ταιριάζει καλύτερα σε τέτοιες εργασίες.

## 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
# To %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 # Από τις 2 επιλογές
#SBATCH --mem-per-cpu=2800M # Μνήμη ανά core # προτίνεται η πρώτη.
#SBATCH -A sept2015 # Accounting tag (θα δοθεί προφορικά αν χρειαστεί)
#SBATCH -t 1-01:00:00 # Ζητούμενος χρόνος DD-HH:MM:SS
#SBATCH -p compute # partition, compute=default στο ARIS. gpu, phi, fat, taskp
# τα εναλλακτικά.

module purge
module load gnu/4.9.2
module load intel/15.0.3
module load intelmpi/5.0.3

if [ x$SLURM_CPUS_PER_TASK == x ]; then #
    export OMP_NUM_THREADS=1 #
else # Δεν σβήνουμε αυτά εκτός αν
    export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK # ξέρουμε ΑΚΡΙΒΩΣ τι κάνουμε
fi # και τι συνέπειες μπορεί να έχει.

srun EXECUTABLE ARGUMENTS # Εδώ το executable και τα πιθανά arguments που παίρνει.
```



## 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

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Παραλείποντας το **--job-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.

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## SLURM Scripts

- Συμβουλευτείτε το site με το documentation του συστήματος
- <http://doc.aris.grnet.gr/scripttemplate/>
- Script generator και validator

## Χρήση **srun** για την εκτέλεση των εφαρμογών

- Οι εκδόσεις του MPI έχουν η κάθε μια ένα mpiexec/mpirun 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 πλήρη έλεγχο.
  - Το srun κάνει accounting κατανάλωσης ρεύματος, χρήση Infiniband, χρήση δίσκων, κλπ.
  - Είναι κοινός τρόπος για τις (3 προς στιγμήν) εκδόσεις MPI που υπάρχουν στο ARIS

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Η χρήση **mpirun**, **mpiexec** κλπ. δεν συνιστάται. Σε περιπτώσεις που η εφαρμογή έχει προβλήματα και 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 περισσότερες 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 των πόρων του συστήματος

```
sinfo
```

- Πληροφορίες για την τρέχουσα χρήση των πόρων συγκεκριμένου partition

```
π.χ. sinfo -p gpu
```

## 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 μια εργασία για να αρχίσει πρέπει κάποια άλλη με το ίδιο job name και χρήστη να έχει τελειώσει, στο SLURM Script εκτός των άλλων :



```
#SBATCH --dependency=singleton
```

- 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=2016-10-26T14:32:00
```

Εάν κάποια εργασία δεν τρέχει και στο nodelist/REASON εμφανίζονται τιμές εκτός από nodenames (τρέχει ήδη) ή Resources (δεν υπάρχουν resources για να ξεκινήσει) ή Priority (προηγούνται άλλα jobs), τότε λογικά έχουμε ζητήσει περισσότερους πόρους από ότι μας επιτρέπεται

- `AssocMaxNodesPerJobLimit`  
Ζητάμε περισσότερα nodes από ότι επιτρέπεται στο account μας
- `AssocMaxWallDur`  
Ζητάμε περισσότερο χρόνο από ότι επιτρέπεται στο account μας
- Διάφοροι άλλοι λόγοι που εάν από το όνομα δεν είναι αντιληπτό, ανατρέξτε 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.

<code>\$\$SLURM_NNODES</code>	# Αριθμός nodes
<code>\$\$SLURM_NTASKS</code>	# Αριθμός Tasks
<code>\$\$SLURM_NPROCS</code>	# " " "
<code>\$\$SLURM_NTASKS_PER_NODE</code>	# Αριθμός Tasks /node
<code>\$\$SLURM_TASKS_PER_NODE</code>	# " " " "
<code>\$\$SLURM_CPUS_PER_TASK</code>	# Αριθμός threads / Task
<code>\$\$SLURM_MEM_PER_NODE</code>	# Μνήμη / node (MB)

## SLURM User/Group resource limits

- Στο SLURM το κάθε account έχει κάποια όρια πόρων που μπορεί να ζητήσει/χρησιμοποιήσει. Τα όρια αυτά εφαρμόζονται σε όλους του χρήστες του account και για όλα τα partitions. Αυτά είναι :
  - Αριθμός Jobs που μπορούν να εκτελούνται ταυτόχρονα, είτε συνολικά είτε ανά partition.
  - Αριθμός Jobs που μπορούν να εκτελούνται ή να βρίσκονται σε αναμονή, είτε συνολικά είτε ανά partition.
  - Μέγιστος αριθμός cores ή nodes που μπορούν να χρησιμοποιηθούν ταυτόχρονα από jobs ενός account, είτε συνολικά είτε ανά partition.
  - Μέγιστη χρονική διάρκεια εκτέλεσης ενός Job, είτε συνολικά είτε ανά partition.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Μέγιστος αριθμός nodes που μπορεί να ζητήσει ένα Job, είτε συνολικά είτε ανά partition.
- Μέγιστος αριθμός cores που μπορεί να ζητήσει ένα Job, είτε συνολικά είτε ανά partition.
- Συνολικός αριθμός core hours στη διάρκεια ενός project, είτε συνολικά είτε ανά partition.

- Ο Scheduler στο ARIS είναι FIFO with Backfill και Fair sharing. 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.
- Fairshare
  - Παράγοντες που επηρεάζουν το priority
    - Χρόνος αναμονής
    - Μέγεθος job σε nodes

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Σχετική χρήση απο groups account π.χ. 80% production, 10% preparatory κλπ.
- Τι θα γίνει αν 4-5 χρήστες στείλουν εκατοντάδες jobs ?
- Το Fairshare αναλαμβάνει να αλλάξει τα priorities ώστε σε επίπεδο εβδομάδας κάποιος/α account να μην μονοπωλεί το σύστημα
- Όσο πιο κοντά στην κατανάλωση του budget βρίσκεται ένα account, τόσο μικραίνει το priority
- Τα jobs που χρειάζονται πολλά nodes, παίρνουν μεγαλύτερο priority.



## Εξομοίωση PBS

- Υπάρχει εγκατεστημένη η εξομοίωση του PBS/Torque. Χρήστες που είναι εξοικωμένοι στη χρήση PBS μπορούν να χρησιμοποιήσουν τα PBS scripts και εντολές, τουλάχιστον για κάποιο διάστημα εξοικίωσης με το σύστημα.
- Η εξομοίωση του PBS καλύπτει μεγάλο βαθμό περιγραφής εργασιών, αλλά όχι όλα

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## 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 2016-10-01**
- Δείτε πόσο χρόνο έχετε καταναλώσει ανά allocation  
**myreport**
- Δείτε πόσο από τον χρόνο που σας έχει δοθεί έχετε  
χρησιμοποιήσει  
**mybudget**

## Νέο υλικό.

- 44 GPU nodes, E5-2660v3/64(56)G, 2xK40m => -O3 -xCORE-AVX2/-O3 -mavx2 -mfma -march=haswell, -p gpu στο SLURM
- 44 FAT nodes, E5-4650v2/512(504)G => -O3 -xCORE-AVX-I/-O3 -mavx -march=ivybridge, -p fat στο SLURM
- 18 Phi nodes, E5-2660v3/64(56)G, 2x Phi 7120 => -O3 -xCORE-AVX-I/-O3 -mavx -march=ivybridge, -p phi στο SLURM
  - Μόνο σε Offload mode.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Τα εκτελέσιμα για GPU/Phi εφόσον υπάρχει υποστήριξη στο πακέτο έχουν την κατάληξη `_gpu` ή `_cuda` ή `_mic`. Προς στιγμή κάποια από τις *accelerated versions* υπάρχει για τις δημοφιλείς εφαρμογές για : `NAMD`, `gromacs`, `lammps`.

## Χρήση Accelerator Resources

- GPU

```
#SBATCH --partition=gpu
```

```
#SBATCH --gres=gpu:2
```

```
Variable : SLURM_JOB_GPUS=0, 1 και  
CUDA_VISIBLE_DEVICES=0, 1
```

- Xeon Phi

```
#SBATCH --partition=phi
```

```
#SBATCH --gres=mic:2
```

```
Variable : OFFLOAD_DEVICES=0, 1
```

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Διαθέσιμα πακέτα

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

## Compilers

- Εγκατεστημένοι Compilers
  - Intel 15.0.3 (default) - 17.0.0
    - module load intel (ή π.χ. intel/17.0.0)
    - icc, icpc, ifort
    - Βασικά Flags : -O3 -xCORE-AVX-I (-xAVX) ή -xCORE-AVX2
    - OpenMP : -qopenmp
  - GNU 4.9.2 (default) - 6.2.0
    - module load gnu (gnu/4.9.3, κλπ.)
    - gcc, g++, gfortran
    - Βασικά Flags : -O3 -mavx -march=ivybridge -mtune=ivybridge ή -mavx2 -mfma -march=haswell -mtune=haswell
    - OpenMP : -fopenmp
  - Για compilation για haswell χρειάζεται επιπλέον να είναι φορτωμένο το binutils/2.27
- PGI 15.5 - 16.9
- cuda 6.5.14 - 8.0.44



Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Debuggers

- gdb
- Intel gdb
- PGI debugger
- ddd

## MPI

- Intel MPI 5.0.3 (default) - 2017.0
- OpenMPI 1.8.8 - 2.0.x for GNU and Intel
- MVAPICH2 2.2.2a for GNU and 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**.

## MPI

### Εκτέλεση MPI εφαρμογών

- Οι εκδόσεις του MPI έχουν η κάθε μια ένα mpirun/mpirxexec 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 (3 προς στιγμήν) εκδόσεις MPI που υπάρχουν στο ARIS

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Σε περιπτώσεις που η εφαρμογή έχει προβλήματα και χρειαστεί να σταματήσει ίσως να παρουσιαστούν προβλήματα (zombie procs) στη χρήση του **scancel**, όταν αυτή έχει ξεκινήσει με **mpiexec/mpiirun**.
- Η χρήση **mvarich2** υποστηρίζεται **ΜΟΝΟ** με **srun**.

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Profilers

- gprof
- mpiP
- Scalasca
- Intel VTune

## Βιβλιοθήκες - Εφαρμογές

module avail για να δείτε την τρέχουσα πλήρη λίστα.

## Μοντέλα παράλληλης χρήσης : OpenMP/Threads

- Δεν αφορούν μόνο το ARIS...
- OpenMP/Threads
  - Παραλληλοποίηση σε επίπεδο loop ή task
  - Όριο : το 1 Node
  - Θετικά
    - Γρήγορη παραλληλοποίηση αλλά σε συγκεκριμένα μόνο μέρη του κώδικα

```
•  
.....  
serial code  
#omp pragma parallel  
    for (i=0;<N;i++) {  
        .....  
    }  
serial code
```

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

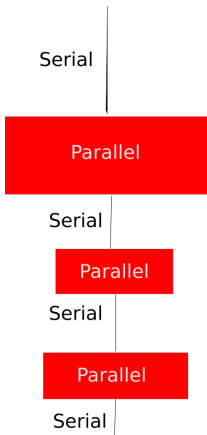
Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Μοντέλα παράλληλης χρήσης : OpenMP/Threads



Σχήμα: Διάγραμμα ροής OpenMP/Threads Παραλληλοποίησης

## Μοντέλα παράλληλης χρήσης : OpenMP/Threads

### ● Αρνητικά

- Δεν μπορεί να ξεπεράσει το scaling του ενός node.
- Δεν είναι όλος ο κώδικας παράλληλος. Εξαρτάται από τα μέρη που έχει (επαρκώς) παραλληλοποιηθεί η απόδοση του κώδικα.
- Δίνοντας τα directives δεν είναι απαραίτητο να υπάρχει καλό efficiency.
- Λόγω της σχετικής ευκολίας παραλληλοποίησης με directives, πιο επιρρεπής σε σοβαρά λάθη υπολογισμών.
- Πιθανές εξαρτήσεις μεταβλητών περιορίζουν το efficiency.
- Οι πιθανές ανάγκες για μνήμη περιορίζουν την εφαρμογή του.



## Μοντέλα παράλληλης χρήσης : MPI

- Παραλληλοποίηση σε επίπεδο υποσυστήματος υπολογισμού.
- Όριο : Το granularity του προβλήματος : Σε ποιό βαθμό μπορεί να μοιραστεί ένας υπολογισμός ?  
π.χ. για MD δεν μπορεί να ξεπερνάει τον αριθμό ατόμων (στην πράξη, 1-4 εκατοντάδες)
- Θετικά
  - Μπορεί να έχει πολύ καλό scaling.
- Αρνητικά
  - Χρειάζεται καλή γνώση του προβλήματος για καλό decomposition.

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Χρειάζεται συγκεκριμένες επικοινωνίες μεταξύ tasks, πιθανότατα και μελέτη τοπολογίας του προβλήματος : π.χ 2D/3D grid, Trees, Hybercube κλπ.
- Πιθανές εξαρτήσεις μεταβλητών επιβάλουν επικοινωνία μεταξύ tasks.
- Αυξημένες ανάγκες μνήμης εξυπηρετούνται από την κατανομή της μεταξύ των nodes.
- Σε αρκετές περιπτώσεις, οι απαιτήσεις μνήμης οδηγούν στη χρήση του.



Σχήμα: Διάγραμμα ροής MPI Παραλληλοποίησης

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Μοντέλα παράλληλης χρήσης : Hybrid

- Υβριδικό μοντέλο παραλληλοποίησης
  - Αρκετά μεγάλα προβλήματα
  - Πρώτα διαμοιρασμός σε tasks μέσω του μοντέλου MPI
  - Σε κάθε ένα από τα tasks, εφαρμογή μοντέλου OpenMP/Threads

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

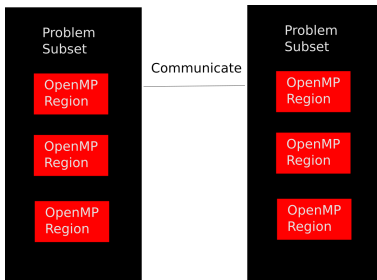
Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές



Σχήμα: Διάγραμμα ροής Υβριδικής MPI/OpenMP/Threads  
Παραλληλοποίησης

## Μοντέλα παράλληλης χρήσης

- Ποιό είναι το καλύτερο ?
  - Εξαρτάται από τον αλγόριθμο
  - ...και κυρίως από τα data.
  - Ο ίδιος αλγόριθμος μπορεί να έχει πολύ διαφορετικό efficiency με διαφορετικά data.
  - Για διάφορους αλγορίθμους υπάρχουν οι "χρυσοί" κανόνες τι είναι καλύτερο, αλλά :
  - Μετρήστε προσεκτικά την απόδοση συναρτήσει δεδομένων και αριθμού nodes/cores καθώς και άλλων λεπτομερειών του προβλήματός σας πριν αποφασίστε.

## Βέλτιστες Πρακτικές

- Τα nodes του ARIS διαθέτουν :
  - Thin, GPU, Phi nodes :20 cores και 64 GB RAM Διαθέσιμα για jobs τα 56 GB.
  - Fat nodes : 40 cores και 512 GB Ram, διαθέσιμα για jobs τα 496 GB.
  - Fat nodes **taskp** partition : 40 physical cores, 80 virtual cores, 512 GB Ram.
- Χρησιμοποιήστε κατά το δυνατόν πλήρως όλα τα cores των nodes, π.χ. 20 cores/node στα thin nodes.

```
--tasks-per-node=20
```

```
--cpus-per-task=1
```

ή

```
--tasks-per-node=2
```

```
--cpus-per-task=10
```

...

ή άλλους συνδιασμούς tasks/threads με γινόμενο 20.

- Εάν χρειάζεστε μη πλήρη nodes χρησιμοποιήστε αναλογικά τη διαθέσιμη μνήμη : 10 cores/node => 28 GB
- Σε περίπτωση που χρειάζεται RAM πάνω από 2.8 GB/core, μπορεί να ζητηθούν λιγότερα cores/node με ταυτόχρονη αύξηση της μνήμης / task, π.χ.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

```
--tasks-per-node=18  
--cpus-per-task=1  
--mem-per-task=3.1G
```

- Η καλύτερα χρησιμοποιήστε fat nodes.



## Βέλτιστες Πρακτικές

- Σε περίπτωση που οι απαιτήσεις μνήμης δεν είναι ίδιες για όλα τα process, χρησιμοποιήστε τη μεταβλητή για συνολική μνήμη / node.

```
--tasks-per-node=20
```

```
--cpus-per-task=1
```

```
--mem=56G
```

## Βέλτιστες Πρακτικές

- Εάν για κάποιο λόγο χρειάζεται αριθμός cores όχι πολλαπλάσιο του 20, συνήθως δυνάμεις του 2 (256, 512, κλπ.)
  - Χρησιμοποιήστε το μικρότερο δυνατό αριθμό nodes.

cores	Nodes	tasks/node	Αχρησιμοποίητα cores
64	4	20	16 σε 1 node
128	7	20	12 σε 1 node
256	13	20	4 σε 1 node
512	26	20	8 σε 1 node

- Σύνθετος λάθος που μεταφέρεται από τη χρήση συστημάτων με 12 ή 16 cores

cores	Nodes	tasks/node	Αχρησιμοποίητα cores
64	4	16	4 cores/node σε 4 nodes = 16
90	6	15	5 cores/node σε 6 nodes = 30
128	8	16	4 cores/node σε 8 nodes = 32
480	40	12	8 cores/node σε 40 nodes = 320
512	32	16	4 cores/node σε 32 nodes = 128

## Βέλτιστες Πρακτικές

- Αρκετά πακέτα διαθέτουν ρυθμίσεις για τα όρια μνήμης στο input τους. Φροντίστε να είναι σε συμφωνία με τα όρια μνήμης που ζητούνται από το SLURM.
- Για jobs που έχουν μεγάλο I/O, χρησιμοποιήστε το χώρο σας στην \$WORKDIR.
- Εάν έχετε το δικό σας κώδικα και κάνετε μεταγλώτιση, χρησιμοποιήστε τα κατάλληλα για το σύστημα compiler flags.
- Χρησιμοποιήστε κατά το δυνατόν τις διαθέσιμες Μαθηματικές βιβλιοθήκες που υπάρχουν στο σύστημα και είναι βελτιστοποιημένες για αυτό.

## Βέλτιστες Πρακτικές

- Εάν για κάποιο λόγο πρέπει να χρησιμοποιήσετε `mpirun`, χρησιμοποιήστε το χωρίς τα συνήθη `-np`, `-machinefile` κλπ. Συμβαίνει όταν χρησιμοποιούνται, να μην αλλάζει ταυτόχρονα ο αριθμός των `tasks` στο `SLURM` και ο αριθμός των `tasks` στο `mpirun -np` π.χ.

```
#SBATCH --nodes=10
```

```
#SBATCH --ntasks=200
```

```
mpirun -np 8
```

**Δεσμεύετε (και χρεώνεστε) για 200 cores ενώ χρησιμοποιείτε μόλις 8.**

## Βέλτιστες Πρακτικές

- Εάν η εφαρμογή σας χρησιμοποιεί OpenMP :
  - Φροντίστε ώστε να δίνετε τα σωστά threads/task στο SLURM.
  - Κοινά λάθη :
    - Δεν θέτουμε τη μεταβλητή  
OMP\_NUM\_THREADS=\$SLURM\_CPUS\_PER\_TASK
    - Για όσο χρόνο το job μας τρέχει μόνο του στο node, μπορεί να χρησιμοποιεί όλα τα cores. Εάν έρθει και άλλο job στο node, τότε το load του node θα ανέβει πάνω από 20 και το performance των jobs εξαρτάται κατά πολύ από τα υπόλοιπα jobs στο node.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Με Hybrid MPI/OpenMP εφαρμογές, αν δεν θέσουμε τη μεταβλητή OMP\_NUM\_THREADS και χρησιμοποιούμε π.χ. 20 tasks/node, τότε το load του node γίνεται  $20 \times 20 = 400$ , με αποτέλεσμα ελλειπόμενο performance.
- Στο script template υπάρχει κώδικας που μας προστατεύει από αυτό.
- Παραδόξως, είναι το σημείο που αφαιρείται πολύ συχνά στα υποβαλλόμενα scripts, ακόμα πιο συχνά και από το job-name.....

## Βέλτιστες Πρακτικές

- Εξερευνήστε την εφαρμογή σας για πιθανές λεπτομέρειες που αφορούν τις επιδόσεις, ειδικά εάν υπάρχει αρκετό I/O.
- Παραδείγματα : quilting στο wrf, Scratch space και direct/semidirect μέθοδοι σε εφαρμογές quantum mechanics.
- Μετρήστε τις επιδόσεις της εφαρμογής εφόσον είναι υβριδική (MPI/OpenMP) για το input σας με διάφορους συνδιασμούς MPI Tasks/Threads per Task (gromacs, namd, lammmps, Quantum Espresso, ....)

## Βέλτιστες Πρακτικές

- Μάθετε ή εξερευνήστε την απόδοση/κλιμάκωση της εφαρμογής σας συναρτήσει του μεγέθους/χαρακτηριστικών των δεδομένων σας. Χρησιμοποιήστε τόσα resources όσα χρειάζονται ώστε να υπάρχει καλό efficiency.



Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

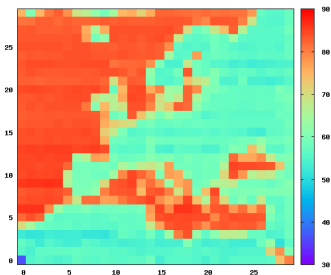
Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Βέλτιστες Πρακτικές



Σχήμα: WRF: Ποσοστό του  
συνολικού χρόνου που  
καταναλώνεται σε MPI κλήσεις.

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

## Βέλτιστες Πρακτικές

- Αν δεν θυμίζει κάτι....

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

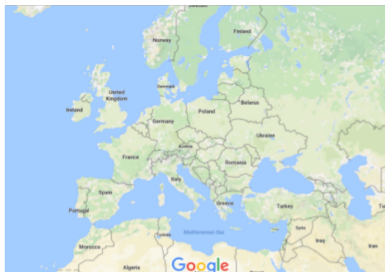
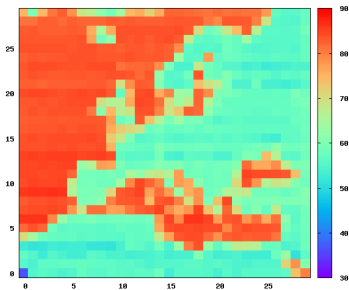
Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

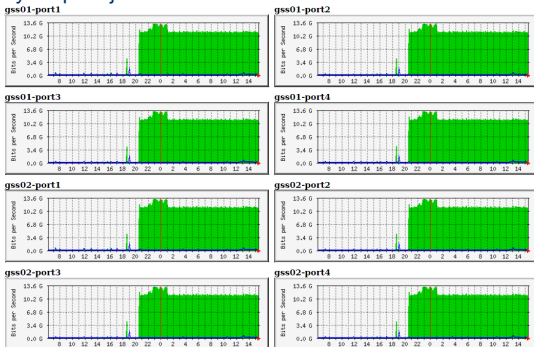
Βέλτιστες  
Πρακτικές

## Βέλτιστες Πρακτικές



## Βέλτιστες Πρακτικές

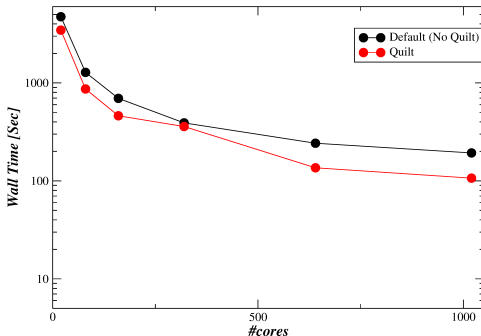
- Παράδειγμα Βαριάς χρήσης SCRATCH : Διάβασμα από files με ρυθμό 12.6 GBytes/s για 2 ημέρες = 2.12 PBytes για 1 job των 100 cores!!!!.



- Με αλλαγή μόνο ενός flag στο input το I/O γίνεται φυσιολογικό.

## Βέλτιστες Πρακτικές

- Παράδειγμα performance/scaling WRF με και χωρίς quilting



## Βέλτιστες Πρακτικές

- Εάν η εφαρμογή σας έχει διαδικασία save/restart χρησιμοποιήστε τη. Αντί για jobs της π.χ. 1 εβδομάδας, προτιμήστε 7 jobs της 1 ημέρας χρησιμοποιώντας τα dependencies του SLURM. Βασικό πρόβλημα στα Hexascale συστήματα.
- Στο πρώτο call σε ορισμένους χρήστες δόθηκε η δυνατότητα να τρέχουν jobs των 7 ή 15 ημερών λόγω αδυναμίας save/restart.  
Ποσοστό jobs που τελειώσαν κανονικά αρκετά μικρό.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Αποφύγετε μή υποχρεωτικές παραμέτρους στο input που ρυθμίζουν το μοίρασμα των υπολογισμών σε cores, δημιουργία grid ή τη μέθοδο που θα χρησιμοποιηθεί αν υπάρχει η δυνατότητα να δίνονται δυναμικά σε run time, κλπ. π.χ. `NPROC_X/Y` στο WRF, `processors` ή `pair_style lj/cut/gpu` vs `pair_style lj/cut` και `-sf gpu` στο LAMMPS.

## Βέλτιστες Πρακτικές

- Αποφεύγετε να βάζετε μεταβλητές περιβάλλοντος στα `.bashrc` κλπ. Ειδικά όταν υπάρχουν πάνω από 1 εκδόσεις ενός πακέτου καλό είναι να ρυθμίζετε το περιβάλλον μέσω των `modules` ή και `scripts` του πακέτου. Παράδειγμα OpenFOAM.  
`module load openfoam/3.0.1`  
`source $FOAM_BASHRC` αντί να βάλετε στο `.bashrc` όλες τις μεταβλητές που θέτει το `$FOAM_BASHRC` μιας έκδοσης.



## Βέλτιστες Πρακτικές

- Εάν τα job σας αποτελούνται από πολλά σειριακά tasks, συγκεντρώστε τα κατά το δυνατόν σε 20άδες για τα compute ή 40άδες/80άδες για τα fat/taskr.
- Εάν τα παράλληλα jobs έχουν μικρή διάρκεια π.χ. 30 λεπτά, δώστε στις απαιτήσεις χρόνου χρονικό διάστημα λίγο παραπάνω.
  - Συχνή κακή τακτική : Στέλνουμε π.χ. 50 jobs τα οποία χρειάζονται 5 λεπτά το καθένα.
  - Εάν στα job descriptions ζητήσουμε π.χ. 10 λεπτά και μας επιτρέπεται να τρέχουμε έως 10 jobs ταυτόχρονα, το σύστημα θα τα προγραμματίσει να τρέξουν, εφόσον υπάρχουν ελεύθερα resources, σε < 1 ώρα.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Πολύ συχνά οι χρήστες βάζουν το μέγιστο όριο χρόνου στα requirements, π.χ. 24 h.
- Στο παραπάνω παράδειγμα το σύστημα θα προγραμματίσει να τα τρέξει σε 5 μέρες.
- Η κατάσταση για τον προγραμματισμό της εκτέλεσης περιπλέκεται ακόμα περισσότερο όταν το σύστημα έχει πολλά jobs που περιμένουν να τρέξουν.

## Βέλτιστες Πρακτικές

- Στατιστικά Σεπτεμβρίου 2016
  - Το 68.5% των jobs χρειάστηκε για να τελειώσει λιγότερο από το 5% του χρόνου που ζήτησε
  - Το 3.5% των jobs μεταξύ 5 και 10 %.
  - Το 13% πάνω από 50%
- Στατιστικά Μαΐου 2016
  - Το 46% των jobs χρειάστηκε για να τελειώσει λιγότερο από το 5% του χρόνου που ζήτησε
  - Το 7% των jobs μεταξύ 5 και 10 %.
  - Το 15% πάνω από 50%

## Βέλτιστες Πρακτικές

- Παράδειγμα για το πόσο μπορεί να βελτιωθεί η ταχύτητα ενός κώδικα χρησιμοποιώντας διαφορετικούς compilers, flags, βιβλιοθήκες : Matrix-Matrix Multiplication

## Πολλαπλασιασμός Πινάκων : Ορισμοί

$$A \cdot B = C$$

$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1N} \\ A_{21} & A_{22} & \dots & A_{2N} \\ \dots & \dots & \dots & \dots \\ A_{M1} & A_{M2} & \dots & A_{MN} \end{bmatrix} \times \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1K} \\ B_{21} & B_{22} & \dots & B_{2K} \\ \dots & \dots & \dots & \dots \\ B_{N1} & B_{N2} & \dots & B_{NK} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1K} \\ C_{21} & C_{22} & \dots & C_{2K} \\ \dots & \dots & \dots & \dots \\ C_{M1} & C_{M2} & \dots & C_{MK} \end{bmatrix}$$

- Για κάθε στοιχείο του πίνακα C απαιτούνται N πολλαπλασιασμοί και N προσθέσεις.
- Συνολικά  $N \times M \times K$  πολλαπλασιασμοί και  $N \times M \times K$  προσθέσεις,  $= 2 \times N \times M \times K$  πράξεις κινητής υποδιαστολής.
- Για  $M = N = K = 1000$ , χρειάζονται  $2 \times 10^9$  πράξεις. Εάν μια μηχανή κάνει αυτές τις πράξεις σε 1 δευτερόλεπτο, τότε η απόδοσή της είναι 2 GFlops.

## Πολλαπλασιασμός Πινάκων

- Τα nodes του ARIS έχουν επίδοση
  - Thin nodes (compute) : 422 GFlops=21 GFlops/core
  - GPU/Phi nodes : 632 GFlops=31.6 GFlops/core (χωρίς χρήση Accelerators)
  - Fat nodes: 644 GFlops=16.1 GFlops/core
- Να το δούμε στην πράξη : Κώδικας σε Fortran,  $N=M=K$

```
do i = 1, N
  do j = 1, N
    c(i,j) = 0.0000000000000000
    do ij = 1, N
      c(i,j) = c(i,j) + a(i,ij) * b(ij,j)
    enddo
  enddo
enddo
```

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

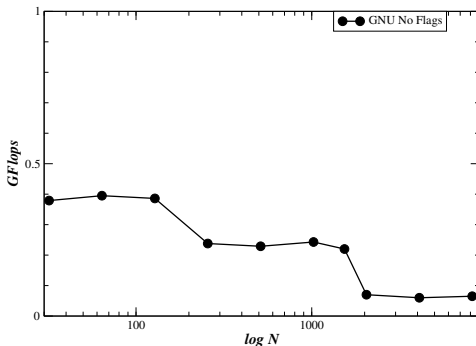
Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

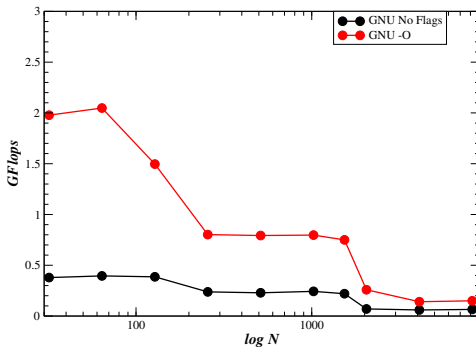
## Πολλαπλασιασμός Πινάκων

- GNU Compilers



## Πολλαπλασιασμός Πινάκων

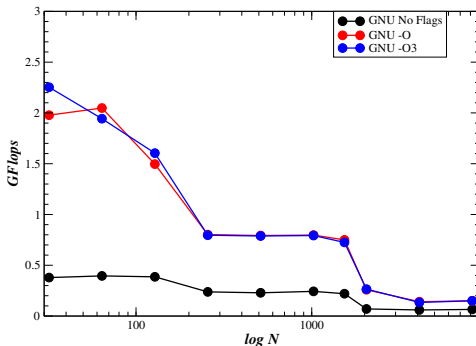
### ● GNU Compilers





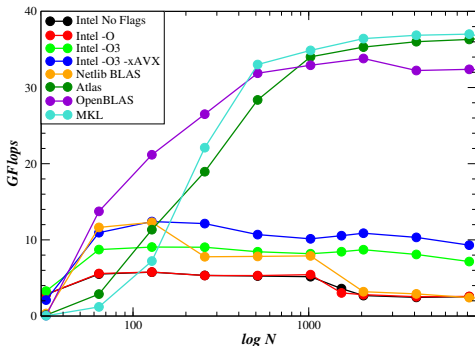
## Πολλαπλασιασμός Πινάκων

### ● GNU Compilers



## Πολλαπλασιασμός Πινάκων

- Intel Compilers + single core libraries



Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

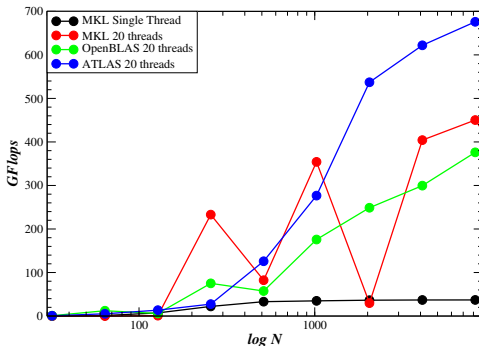
Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

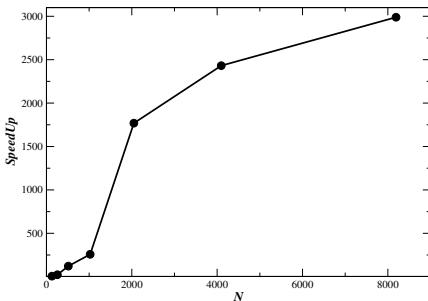
## Πολλαπλασιασμός Πινάκων

### ● Multithreaded Libraries



## Πολλαπλασιασμός Πινάκων

- Συνοψίζοντας :



- Ακολουθώντας πρακτικές βέλτιστης χρήσης, επιταχύνθηκε ο πολλαπλασιασμός πινάκων έως ~ 3000 φορές σε ένα μόνο node.

## Πολλαπλασιασμός Πινάκων

- Τελειώσαμε εδώ φτάνοντας πολύ κοντά στη θεωρητική επίδοση σε μεγάλα  $N$  ?
- Σε κάποιες περιπτώσεις εφαρμογών, γίνεται μεγάλος αριθμός πολλαπλασιαμών πινάκων μικρών διαστάσεων, π.χ.  $A(11,7) \times B(3,11)$ . Οι optimized BLAS συνήθως δεν έχουν ικανοποιητική επίδοση.
- **LibSMM** (Small Matrices Multiplication) developed by Cray Inc.

## Γενική Χρήση του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

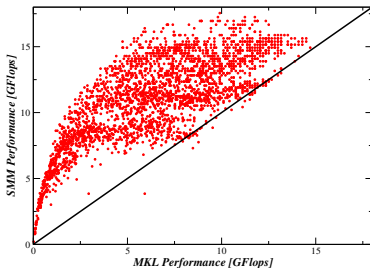
Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

- Επιλέγοντας compiler, flags και κάποια BLAS αναφοράς, βρίσκει την επίδοση για N/M/K στο εύρος 1-24, και αφού κάνει αρκετές χρονομετρήσεις αποφασίζει για κάθε συνδυασμό N/M/K αν θα χρησιμοποιήσει το δικό της κώδικα ή την βιβλιοθήκη αναφοράς.
- Σε χρήση απο τις εγκατεστημένες versions του cp2k.

## Πολλαπλασιασμός Πινάκων

- Αντικατάσταση είτε του κώδικα είτε της κλήσης `dgemm`  
`call dgemm('N','N',N,N,N,ONE,a,lda,b,ldb,ONE,c,ldc)`  
με `call smm_dnn(M,N,K,A,B,C)`.



Σχήμα: Σύγκριση απόδοσης libSMM vs MKL για  
 $1 \leq N/M/K \leq 24$ .

Γενική Χρήση  
του ARIS

Δρ. Δημήτρης  
Ντελλής

Περιεχόμενα

Σύνδεση στο  
σύστημα

Σύστημα  
αρχείων

Environment  
Modules

Batch System

Διαθέσιμα  
πακέτα

Μοντέλα  
Παράλληλης  
Χρήσης

Βέλτιστες  
Πρακτικές

# Ερωτήσεις ?