Efficient use of ARIS
Dr. Dimitris Dellis

Outline
System Details
Connect to ARIS
File Systems
Environment Modules
Batch System
Installed Software
Best Practices
Discussion-Questions

Efficient use of ARIS

Dr. Dimitris Dellis

GRNET

ntell [at] grnet.gr

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Outline

- System Details
- Connect to ARIS
- File Systems
- Software Environment
  - Environment Modules
  - Available Software Packages
- Batch System
- Best Practices, Typical problems/mistakes
- Questions/Discussion

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

ARIS islands

- Hardware
  - 426 compute nodes, E5-2680v2/64(56)G
  - 44 GPU nodes, E5-2660v3/64(56)G, 2x NVIDIA K40m
  - 16 FAT nodes, E5-4650v2/512(496)G
  - 28 FAT nodes, E5-4650v2/512(496)G, up to 80 tasks/node, with Turbo Boost/HyperThreading.
  - 18 Phi nodes, E5-2660v3/64(56)G, 2x Xeon Phi 7120P
  - 24 service nodes
- Operating System : Red Hat Enterprise Linux 6 (currently 6.8)
- File systems IBM GPFS
- Applications/Storage network : Infiniband 56 Gbps
- Management Network : Gigabit Ethernet

European Union European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE
The project is co-financed by Greece and the European Union

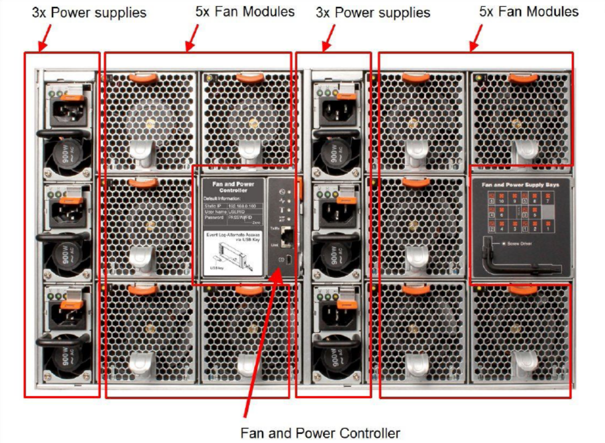
Compute Nodes (enclosures front)

European Union
European Regional Development Fund
digitalgreece Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Compute Nodes

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE
The project is co-financed by Greece and the European Union

Compute Nodes(enclosures rear)



The project is co-financed by Greece and the European Union

Compute Nodes (inside)

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Service nodes

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

- Storage System (hardware)
  - IBM GSS 26
  - 348 HDD 3 TB, 7200 RPM SAS
  - 232 HDD 4 TB, 7200 RPM SAS
  - RAID 6 Arrays (Solomon Reed 8+2p)
  - Total Disk Space 2 Petabyte
  - Usable space after RAID, about 1.4 PB
  - Performance > 12 GBytes/sec

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

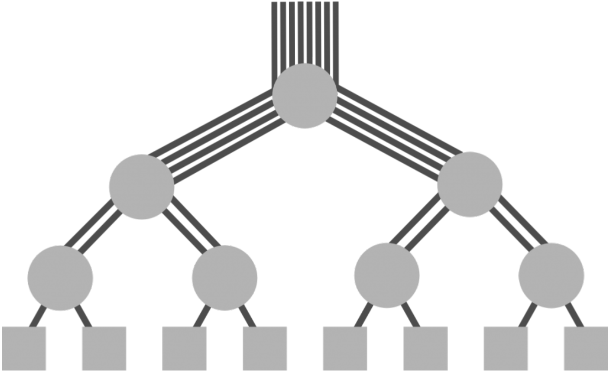
- Storage system
- Storage 1 (2015) 6 enclosures, 5 planes/enclosure, up to 12 HDD on each plane.
- Storage 2 (2016) 4 enclosures, 5 planes/enclosure, up to 12 HDD on each plane.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE
The project is co-financed by Greece and the European Union

- Interconnect network : Infiniband
  - Mellanox SX6536 648-Port Infiniband Director Switch
  - FDR 56 Gbits / sec
  - Fat tree non-blocking mode => 56 GBps All to All

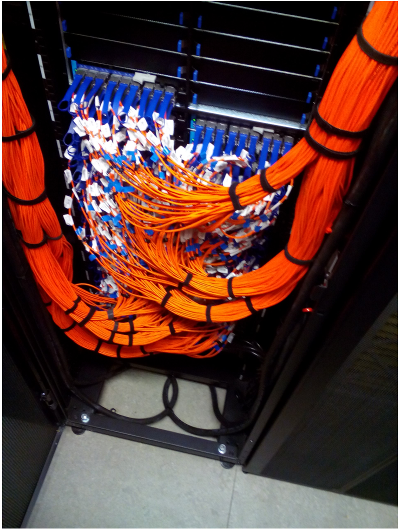
European Union
European Regional Development Fund
digitalgreece Everything is possible
NSRF 2007-2013
The project is co-financed by Greece and the European Union

Fat tree non-blocking mode (generic)



The project is co-financed by Greece and the European Union

Mellanox SX6536



European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

- Connect to ARIS
  - Two login nodes : login0[1|2].aris.grnet.gr, alias RR login.aris.grnet.gr
  - Login access ONLY on login nodes, from certain IP addressess/networks (Fixed)
  - Access is allowed ONLY via SSH, using keys - no password authentication.
  - Exactly the same installation, users, accounts, file systems with all cluster nodes

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

- SSH Clients
  - MacOS, Linux : OpenSSH, typically already installed
    Redirect graphical environment : ssh -X username@login.aris.grnet.gr
    - **ssh** : SSH client, Connect and get a shell prompt.
    - **ssh-keygen**: Create/manage keys
    - **scp, sftp**: File Transfer
  - Windows: PuTTY (Free)
    - **PuTTY** : SSH client, Connect and get a shell prompt.
    - **PuTTYgen** : Create/manage keys
    - **PSCP, PSFTP** : File Transfer
  - Windows: Bitvise (Free, with graphical interface, File transfers, capable for X11,)

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

- Create private/public key on MacOS, Linux
  - **ssh-keygen -t rsa -b 2048**
  - public key: .ssh/id_rsa.pub
  - private key: .ssh/id_rsa
- File Transfers
  - SSH Connections from ARIS to any IP is not allowed, ONLY from certain IPs to ARIS.
  - To transfer files from ARIS to your PC, is not necessary to connect from ARIS to your PC and issue **put**.
  - Instead connect from your PC and issue **get**.

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

- X Server for windows
  - Usefull for packages with graphical interfaces, data vizualization etc.
  - Xming X Server for Windows
  - http://sourceforge.net/projects/xming/
  - Make sure that X11 forwarding is enabled in your Client SSH application.
  - Xming should run before start a graphical application on ARIS

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

- File Systems : GPFS
  - GPFS 4.1
  - 4 filesystems : /users /work /work2 and /staging
  - /users
    - $\sim$ 240 TB
    - Applications
    - Users Home directories
    - Applications should NOT run here (at least the I/O intensive)
    - Long term storage (= Daily Backup)
  - /work and /work2
    - $\sim$ 440 + 400 TB
    - The $WORKDIR environment variable sets the location of each user's work dir
    - Jobs SHOULD run here
    - Short term storage, no frequent Backup

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

- /staging
  - $\sim$ 150 TB
  - Long term storage of large files, real storage on DLT Tapes.
  - Extremely slow - each access activates transfer from Tape to disk, it may take hours.

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Environment Modules. What they are ?

- Typically, in order to use applications not installed in standard system paths, one needs to adjust mainly the PATH and LD_LIBRARY_PATH environment variables.
- More variables are necessary to be set for various packages.
- Q: Why you do not install an application in standard system path ?
    - How to handle different versions of the same application ?
    - Standard system paths in large clusters usually reside in Memory.
- It is common practice to set all the variables in .bashrc on single node machines that typically run one version of few packages. For example OpenFOAM.
- It is not easy to handle different versions of the same application.
- Each user shoud care for the contents of .bashrc
- Usually one forgets what it is included there.

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

Environment Modules. What they are ?

- Environment Modules package modifies on demand (set/unset) user environment variables
- Usual variables are PATH, MANPATH and LD_LIBRARY_PATH, but also other package specific variables. For example, JAVA_HOME, LM_LICENSE_FILE for intel compilers etc.
- Each module file has the information needed in order to set the environment variables, as well as their dependencies.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE
The project is co-financed by Greece and the European Union

- If an application requires another application or library, and the corresponging module is not loaded, an error is generated.
- For packages with more than one version, one of these is marked as default by administrator.

European Union
European Regional Development Fund
digitalgreece Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Environment Modules. How to use

- List of available modules

  ```
  module avail
  ```

  or

  ```
  module -l avail
  ```

- List active modules

  ```
  module list
  ```

- Deactivate any loaded module

  ```
  module purge
  ```

- Deactivate certain module

  ```
  module unload MODULENAME
  ```

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

- Version change of a module

```
module switch MODULENAME/VER1 MODULENAME/VER2
```

- Information about a module

```
module whatis MODULENAME/VERSION
```

- Module Help text

```
module help MODULENAME/VERSION
```

- To see what a module load does in environment :

```
module show MODULENAME/VERSION
```

European Union European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

Environment Modules. How to use

- Explore available packages : Live

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Batch System

- It is common to :
  - Before weekend we have a running job that use all the cores of a workstation, that it is expected to finish sometime at Saturday. Start another job, both will run till sometime at Sunday. In general degradation of performance.
  - Sunday to Monday machine is idle.
  - User X runs many jobs, I also start few to get some slots of machine power.
  - Next 15 days on leave for vacation. Start 20 jobs all together, sometime they will finish.
    Probably many users have the same idea.
  - Is this situation something that you faced sometime ?

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Batch System

- What is Batch System/WorkLoad Manager
  - A Batch system controls the use of available resources in order to :
    - All users use a portion of machine power
    - Ensure that each job will use the machine when the requested resources are free.
    - Control the job. After submission user can logout.
- ARIS Batch System : SLURM, PBS emulation is supported.

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

When a job is submitted to a batch system :

- Users describe the required resources i.e. number of cores, memory, execution time, probably start date.
- Batch system prioritize the jobs according to the requested resources.
- When the available resources fullfil the requested resources, resources are allocated to the job and the job starts execution.
- Batch system guantee that each job will have exclusive access on the allocated resources (cores, memory etc.)
- A user may submit for example 1000 jobs. They will be executed without overlap of resources.

European Union European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

SLURM Scripts

A Slurm Script (as well other workload managers scripts) describes the required resources as well as what to do in real execution.

Job requirements are flags to lines starting with #SBATCH

There are two ways to describe each requirement : for example

`--nodes=10` and `-N 10`. Other statements (not starting with #SBATCH) are either shell commands used for job execution.

European Union European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

SLURM Scripts

```
#!/bin/bash
#SBATCH --job-name=”test”    # Job Name
#SBATCH --error=job.err.%j   # Filename for stderr redirection
#SBATCH --output=job.out.%j  # Filename for stdout redirection
                             # %j is the value of JobID
#SBATCH --nodes=200          # Number of nodes
#SBATCH --ntasks=400         # Number of Tasks (i.e. processess)
#SBATCH --ntasks-per-node=2  # Tasks / node
#SBATCH --cpus-per-task=10   # Threads / Task
#SBATCH --mem=56G            # Memory per node
#SBATCH --mem-per-cpu=2800M  # Memory per core
#SBATCH --account=pa1704099  # Accounting tag (each project has one)
#SBATCH -t 1-01:00:00        # Requested Time DD-HH:MM:SS
#SBATCH -p compute           # partition, compute=default on ARIS. gpu, phi, fat, taskp

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

if [ x$SLURM_CPUS_PER_TASK == x ]; then             #
  export OMP_NUM_THREADS=1                          # We NEVER remove these statements.
else                                                # f we don’t know what we do
  export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK       # and its consequences.
fi                                                  #

srun EXECUTABLE ARGUMENTS   # Executable and arguments.
```

SLURM Scripts

- Previous slide script is a complete job requirements description.
- One SLURM Script may be also complete if ommit few of the SBATCH directives
  - Specifying only `--nodes` without `--ntasks` : system is able to calculate the number of tasks - it knows how many cores are available on each node.
  - Specifying `--ntasks` without `--nodes` : System is able to calculate the number of nodes needed to run ntasks.
  - The presence of `--account` is mandatory. If it is not specified, the job is rejected on submission.

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

SLURM Scripts

- Visit the ARIS documentation site for more details
- http://doc.aris.grnet.gr/scripttemplate/
- Script generator and validator

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Use of **srun** to run applications

- Each MPI flavour has an mpirun/mpiexec etc.
- On ARIS the use of srun is suggested with any type of executable, serial or parallel.
- Some reasons
  - srun spawns the executables on each node, knows the state of each task, etc.
  - srun logs in accounting power consumption, I/O, network usage etc.
  - srun propagade all the environment variables to all nodes. With ssh (that is the underlying protocol for other wrappers (mpirun) it is not guaranteed that the environment variables are the same to all tasks.

The project is co-financed by Greece and the European Union

Working with SLURM

- Submit a Job

```
sbatch SLURM_JobScript.sh
Submitted batch job 15242
```

- List of jobs

```
squeue
```

- Detailed list of jobs

```
squeue -o ”%.8i %.9P %.10j %.10u %.8T %.5C
%.4D %.6m %.10l %.10M %.10L %.16R”
```

(**man squeue** for details).

- Job Cancel

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

```
scancel JobID
```

- Some applications include signal handling. When we send scancel to a job, tasks receive a SIGHUP. If for any design reason they ignore or handle it as something else, the job will leave zombie procs on nodes. In this case on should specify that scancel has to send SIGKILL to tasks

```
scancel -s KILL JobID
```

- Estimation of job start time

```
squeue --start
```

- Info about resources usage

```
sinfo
```

- Info about resources usage of certain partition.

```
π.χ. sinfo -p gpu
```

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE
The project is co-financed by Greece and the European Union

SLURM jobs dependency

- When start of a job requires that another job is finished, we should add (among other directives)

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

  or

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

- If we require that we run ONLY one instance of a job with a certain job-name,

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

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE
The project is co-financed by Greece and the European Union

SLURM jobs dependency

- If a job should start at a certain date/time :
  - Start at next 16:00

    ```
    #SBATCH --begin=16:00
    ```

  - Start at certain date and time:

    ```
    #SBATCH --begin=2017-04-06T16:32:00
    ```

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

If a job is not running and in the nodelist/REASON column appear values other than nodename (Already Running), or Resources (No available resources), or Priority (other jobs should run before this job), or Dependency (job not running due to requested dependcy),
it is possible that we ask for more resources than the allowed to our account

- `AssocMaxNodesPerJobLimit`
  We ask more than allowed nodes
- `AssocMaxWallDur`
  We ask for more wall time than allowed (2 days)
- Other reasons starting with Assoc, that are self explained.

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

SLURM User/Group resource limits

- There are various SLURM limits. Each account has a certain set of limits
  - Number of Jobs that may run concurrently
  - Number of Jobs in Queue - independent of state (running/waiting)
  - Maximum number of cores, nodes etc. that all jobs of an account may use concurrently
  - Maximum execution time - 2 days for all
  - Maximum number of Core Hours that a project can use on ARIS.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

SLURM User/Group resource limits

- Calculation of Core Hours : If a job requests 20 tasks (=cores) then 20 cores are allocated to job. The budget consumption of this job is 20 CoreHours for 1 hour wall time, no matter if the job really use the allocated resources.

The project is co-financed by Greece and the European Union

Use of Accelerated partitions

- GPU

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

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

- Xeon Phi

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

  Variable : `OFFLOAD_DEVICES=0,1`

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Installed Software

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

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- ARIS nodes have :
  - Thin, GPU, Phi nodes :20 cores and 64 GB RAM. Available for jobs 56 GB.
  - Fat nodes : 40 cores and 512 GB Ram, available for jobs 496 GB.
  - Fat nodes **taskp** partition : 40 physical cores, 80 virtual cores, 512 GB Ram, available for jobs 496 GB.
- Use if possible all node cores, for example 20 cores/node on thin nodes.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

```
--tasks-per-node=20
--cpus-per-task=1
or
--tasks-per-node=2
--cpus-per-task=10
...
```

or any other combination threads/task with product *tasks* $\times$ *threads* $= 20$.

- If your runs require less than availab cores of a node, use the corresponging memory if possible : 10 cores/node => 28G (fair) but not 56G (leave room for other users with lower memory requirements to use the unused cores)

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE
The project is co-financed by Greece and the European Union

- If you need more than 2.8 GB/core on thin nodes or aggressive memory/node more than 56 GB, you could ask less cores/node.

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

- Consider to move to fat nodes for efficiency reasons

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- If (it is usual in some applications) task 0 needs (much) more memory than other tasks, use the per node memory directive :

```
--tasks-per-node=20
--cpus-per-task=1
--mem=56G
```

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- If your application need number of tasks that are not multiple of 20 (or whatever in general), usually powers of 2 number of tasks (128, 256, 512 etc.)
  - Use the minimum number of nodes :

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

  - Typical mistake that comes from the use of 8/12/16 cores/node systems :

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

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

Best Practices

- Many packages contain in their input files variables for memory use. Try to be in aggreement with what is requested from Batch system.
- If your job is I/O intensive, avoid to use your HOME directory for runs. Instead use your $WORKDDIR,
- If you have your own code, use the suggested compilers and compiler flags to obtain the optimum performance.
- Use if possible the precompiled/optimized for system math - I/O libraries available on the system. If you need something that is missing, just ask at suport[at]hpc.grnet.gr to install it.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- If for any reason you have to use mpirun for MPI execution, use it without any argument about tasks, nodes etc. -np, -machinefile. It is frequent, when you use them to have different job requirements (tasks) and -np tasks, for example,

```
#SBATCH --nodes=10
#SBATCH --ntasks=200
mpirun -np 8
```

You allocate (and **you are charged for**) 200 cores while you really use just 8.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- If your application is Hybrid MPI/OpenMP
  - Try to correctly describe tasks/threads in slurm script.
  - Common mistakes:
    - We ommit the variable OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK
    - If you run alone on a node, you run may use all cores. If another job is landed on the same node, the node load increase to values higher than the number of cores => degradation of performance, increased power consumption, higher temperatures etc.
    - With hybrid applications if we ommit the OMP_NUM_THREADS variable, using 20 tasks may lead to a node load of 400 on a 20 cores node => degraded performance.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE
The project is co-financed by Greece and the European Union

- In script template at http://doc.aris.grnet.gr/scripttemplate/, there is code that protect us from such errors.
- Surprisingly, this piece of code it is frequently removed from submitted scripts

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- Learn or explore how your application performance is affected by the size/characteristics of your input. The fact that someone published that X application has high performance/scaling doesn't mean that you'll get similar performance with your data.
- Use the maximum resources (cores, nodes etc) that yield the best performance/resources ratio, at least an efficiency of 60%.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- If ypur application gives you the chance to use save/restart procedure, use it. Instead of a week job (if it is allowed) use 7 jobs of 24h with save restart, probably using the dependencies feature of slurm.
- There are projects that consumed almost 4 millions CoreHours using this procedure.
- Some stats : For some projects with no chance for save/restart => Wall time 15 or 30 days, only a few percent of jobs was completed.
- If your application input file has variables related to number of cores/nodes, avoid it, if possible.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- Avoid to set any variable in .bashrc etc. Especially if more than one versions of the package are available (example : OpenFOAM). Use the corresponding environment modules.

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE
The project is co-financed by Greece and the European Union

Best Practices

- If your jobs contain many serial jobs (like R, octave) pack them if possible in bundles of 20, 40, 80 - depending on partition.
- Use time requirements that match your expected times. Frequent bad practice :
  - We submit for example 50 jobs that really need 10 minutes each.
  - If we ask for each 24 hours and our limits allow 10 concurrently running jobs, the system will schedule to run them (depending on resources availability) in the next 5 days.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

- If we ask for example 6 minutes for each and there are available resources, the jobs will be scheduled to run within next 1 hour.
- In the case system has many waiting jobs, scheduling is more complicated.

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- September 2015
  - 68.5% of jobs completed in less than 5% of the requested time
  - 3.5% of jobs between 5 and 10%.
  - 13% more than 50%
- May 2016
  - 46% of jobs completed in less than 5% of the requested time
  - 7% of jobs between 5 and 10 %.
  - 15% more than 50%
- March 2017
  - 52.62% of jobs completed in less than 5% of the requested time.
  - 9.07% of jobs between 5 and 10%.
  - 19.93% more than 50%.

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

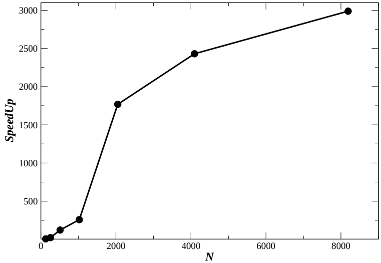
Conclusions after some Training events :

- Users that follow the instructions (slowly increasing percentage) : 13 -> 15 -> 19.9% of jobs.
- Users who do not care (and probably face delays) almost constant.
- Try to be in the first case.

European Union European Regional Development Fund digitalgreece Everything is possible Operational Programme "Digital Convergence" NSRF 2007-2013 PRACE GR
The project is co-financed by Greece and the European Union

Best Practices

- What may mean for Performance : I follow the instructions ?
- Matrix - Matrix Multiplication Performance, ”I do not care about details” compilation vs Optimum compilers/flags/libraries :



- Following Best Practice guides, I had a speed up of ~ 3000 times.

European Union
European Regional Development Fund
digitalgreece Everything is possible Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union

Discussion on the software you plan to use

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
The project is co-financed by Greece and the European Union

Efficient use of ARIS

Dr. Dimitris Dellis

Outline

System Details

Connect to ARIS

File Systems

Environment Modules

Batch System

Installed Software

Best Practices

Discussion-Questions

Questions ?

European Union
European Regional Development Fund
digitalgreece
Everything is possible
Operational Programme "Digital Convergence"
NSRF 2007-2013
PRACE GR
The project is co-financed by Greece and the European Union