Introduction to ARIS and PRACE

Dr. Dimitris Dellis

GRNET

NTUA, 19 Dec. 2017

Outline

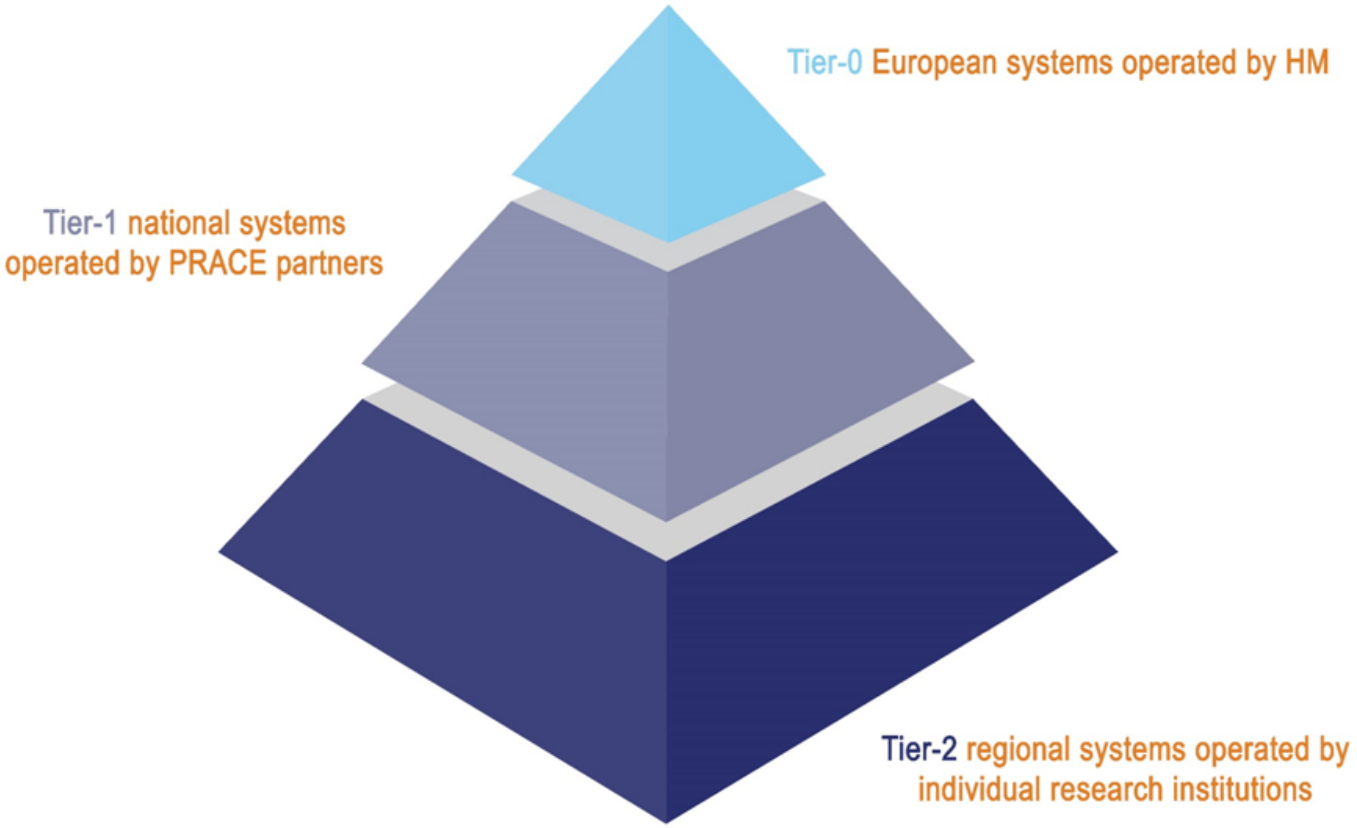
- Introduction to PRACE
- Introduction to ARIS
- Working with ARIS

European HPC Ecosystem - About PRACE

Partnership for Advanced Computing in Europe

- EU Organization
- Coordinates the development of Computational Infrastructures in Europe
- Offers access to Petaflop level machines (Tier-0)
- Much more.
- Greece is Founder member of organization - non hosting member since 2007
- Since 2015 is hosting Tier-1 system.

PRACE Systems Hierarchy



PRACE Tier-0 Systems

MareNostrum: Lenovo
BSC, Barcelona, Spain

#16 on Top500

Piz Daint: Cray XC 50
CSCS
Lugano, Switzerland

#3 on Top500 Nov17

#22 on Top500

JUQUEEN: IBM
BlueGene/Q
GAUSS/FZJ
Jülich, Germany

SuperMUC: IBM
GAUSS/LRZ Garching, Germany
#44 & #45 on Top500

Hazel Hen: Cray
GAUSS/HLRS,
Stuttgart, Germany

#19 on Top500

#93 on Top500

CURIE: Bull Bullx
GENCI/CEA
Bruyères-le-Châtel, France

#14 on Top500

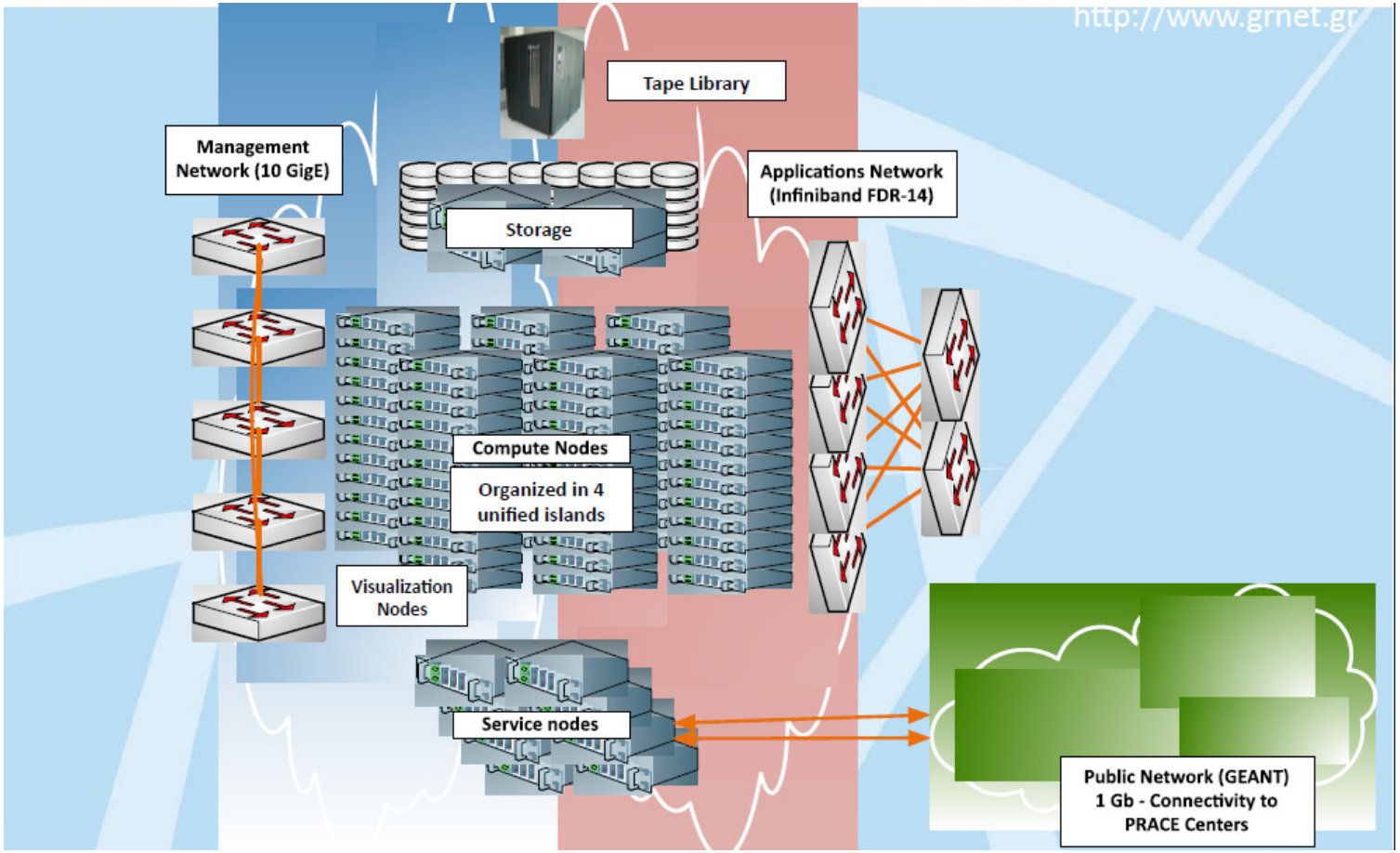
MARCONI: Lenovo
CINECA
Bologna, Italy

European HPC Tier-1 Ecosystem - PRACE

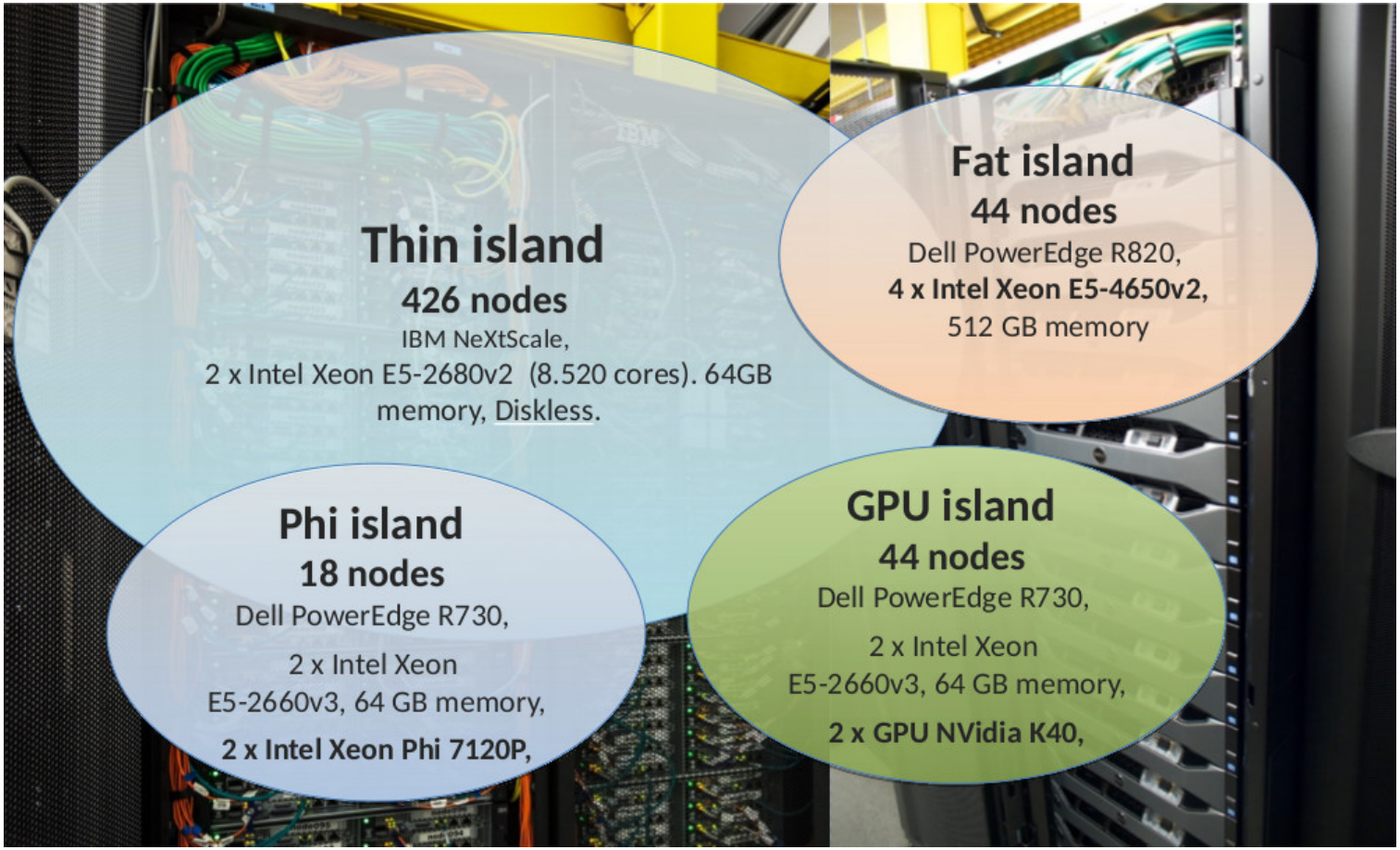
- DECI : Resources exchange program : Each Tier-1 hosting country contributes a part of compute capacity, and researchers from this country can get access to other Tier-1 systems.
- Main reasons
  - Trigger International Scientific Cooperations
  - Possibility to use resources of different type that are not available. For example, Bigger than available systems, BlueGene, Cray, KNL, etc.
  - Intermediate stage before Tier-0 access.
  - Evaluation of projects in home country.
- Countries in DECI : Cyprus, Czech Republic, Finland, Greece, Hungary, Ireland, Italy, Norway, Poland, Spain, Sweden, Netherlands, UK.
- Calls for DECI Projects every 6 months. Announced in prace (and hpc.grnet.gr) web site.

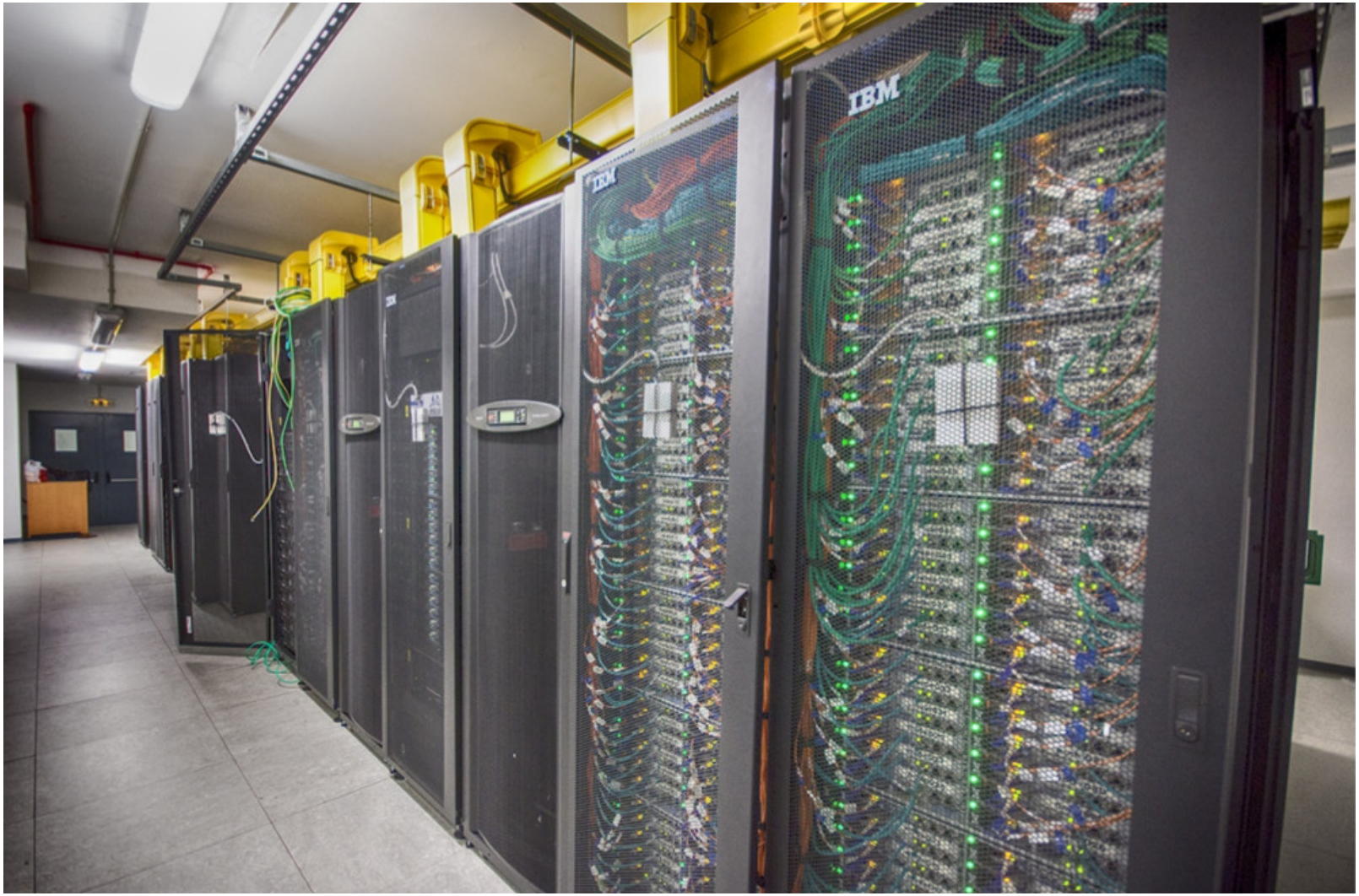
Introduction to ARIS

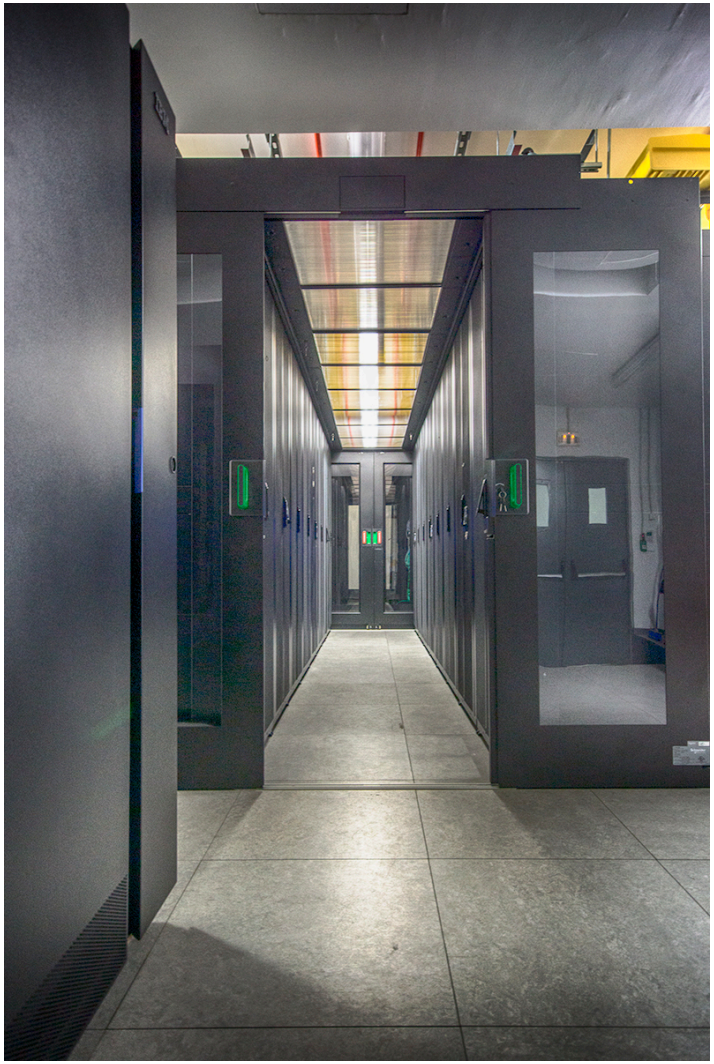
System Organization

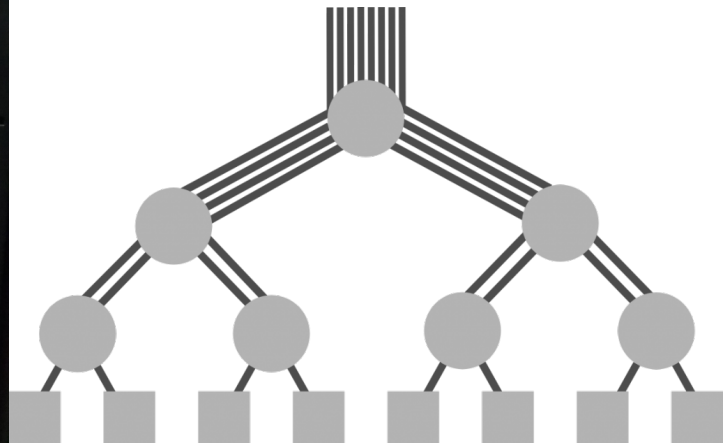


Compute Nodes Organization

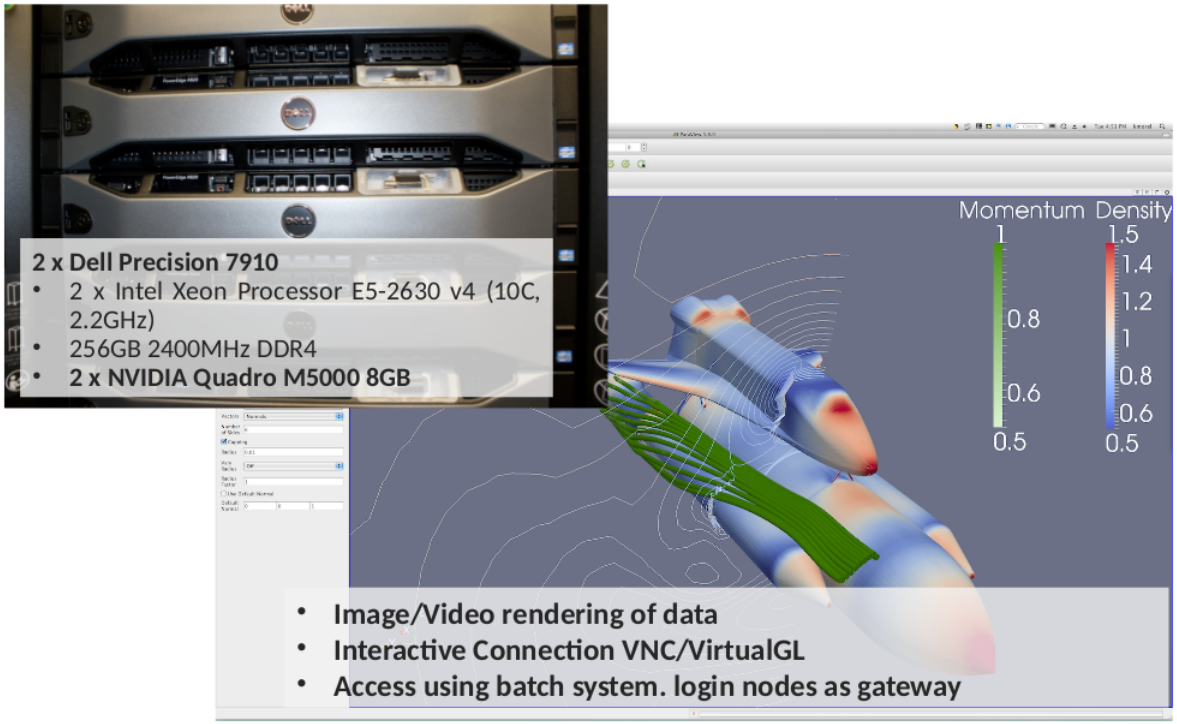



IBM
IBM

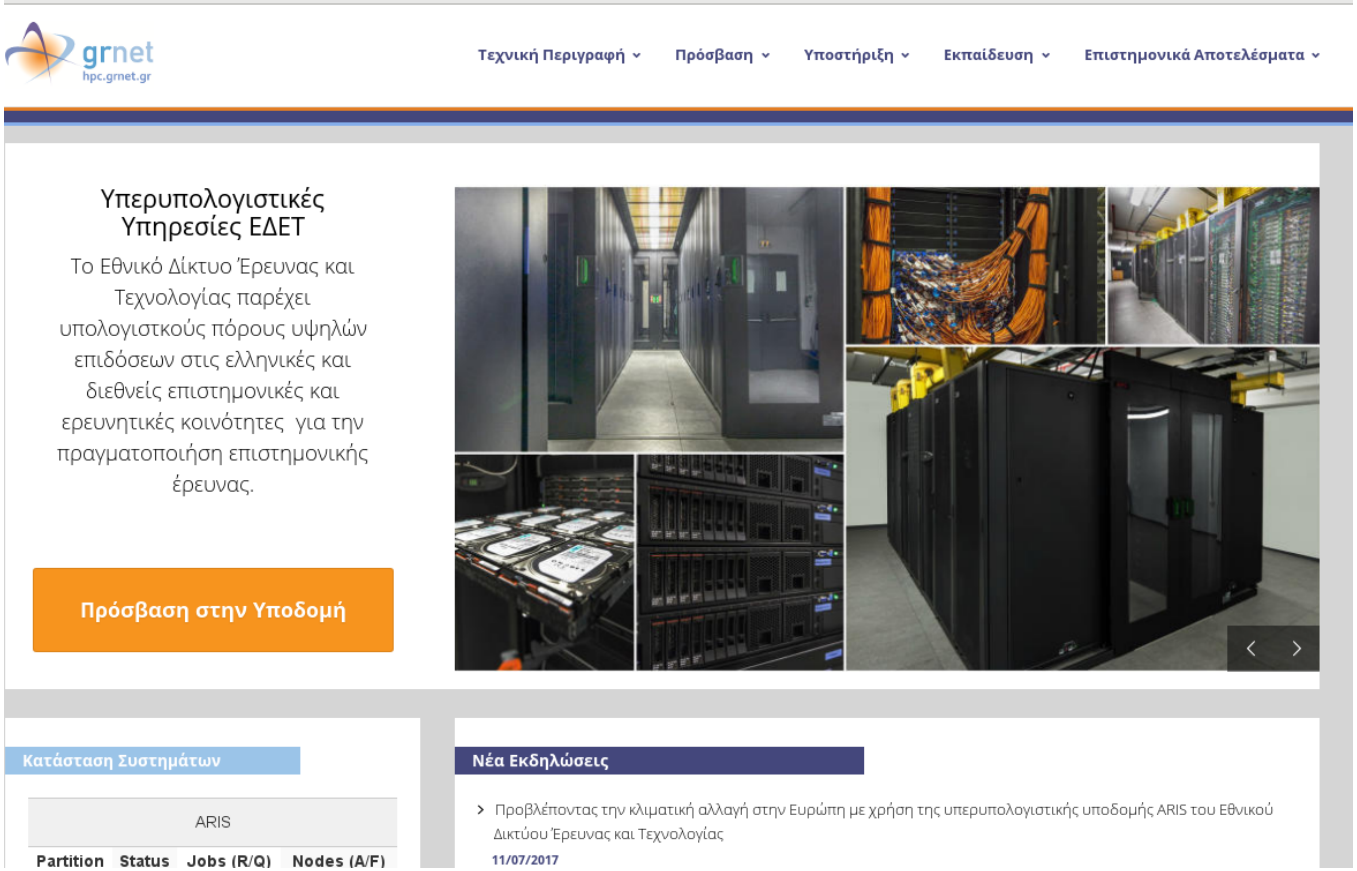




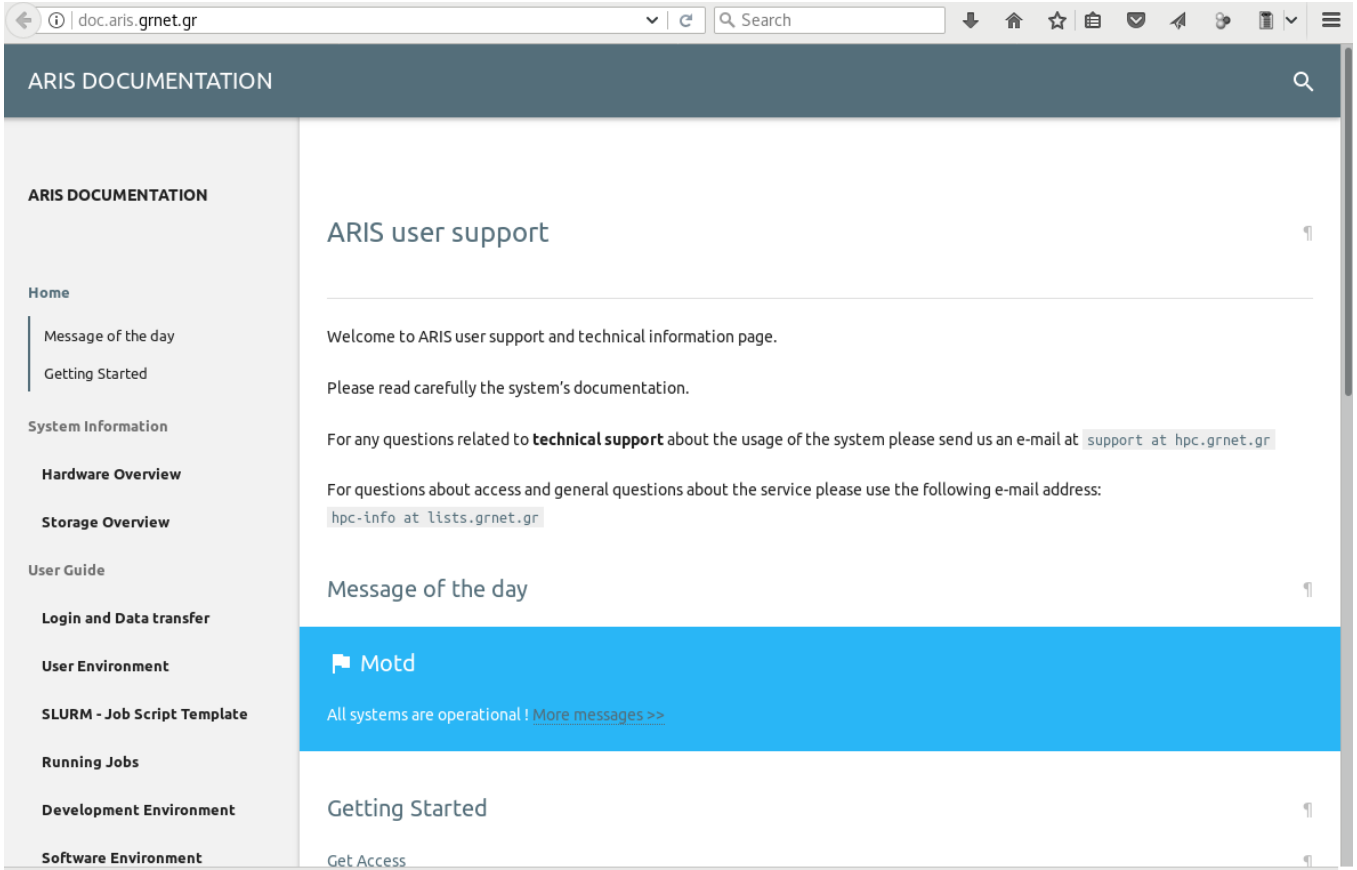
Visualization Nodes



Information for Access, News etc. (mainly in Greek)
https://hpc.grnet.gr/

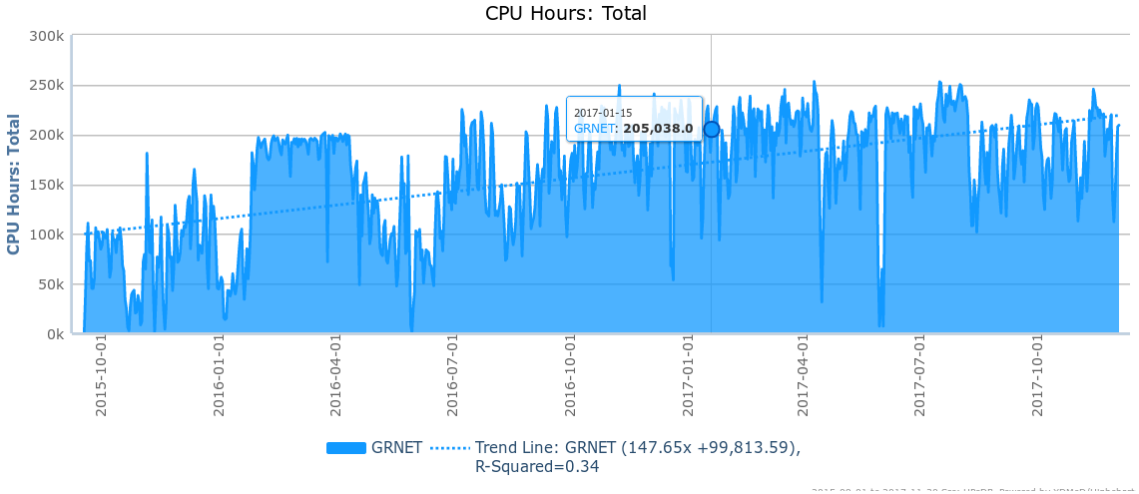


System Documentation (Only in English) : http://doc.aris.grnet.gr/



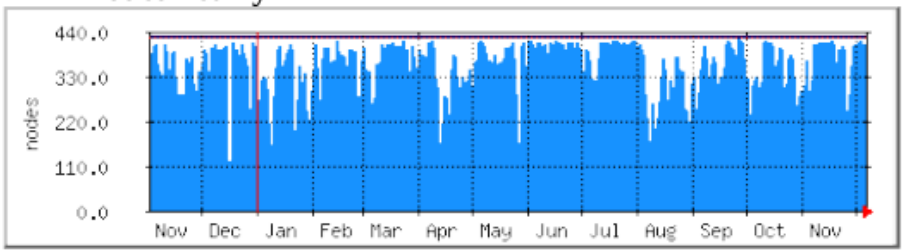
System Usage

- System Capability :
  8520 / 11520 core Years / year before/afrer Aug 2016.
- Allocated up to now : $\sim$ 26,000 core Years
- In 4 Production calls, always open preparatory call, contribution in DECI, VI-SEEM, SoHPC, etc.
- 3rd and 4th production calls in progress, 5th just closed.
- Consumed up to now : $\sim$ 15,000 core Years



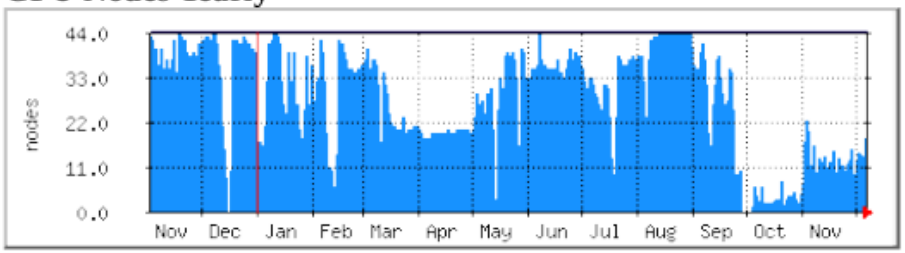
System Usage

Thin Nodes Yearly



	Max	Average	Current
Nodes Allocated	419 Nodes (98.4%)	344 Nodes (80.8%)	401 Nodes (94.1%)
Total Nodes	426 Nodes (100.0%)	426 Nodes (100.0%)	426 Nodes (100.0%)

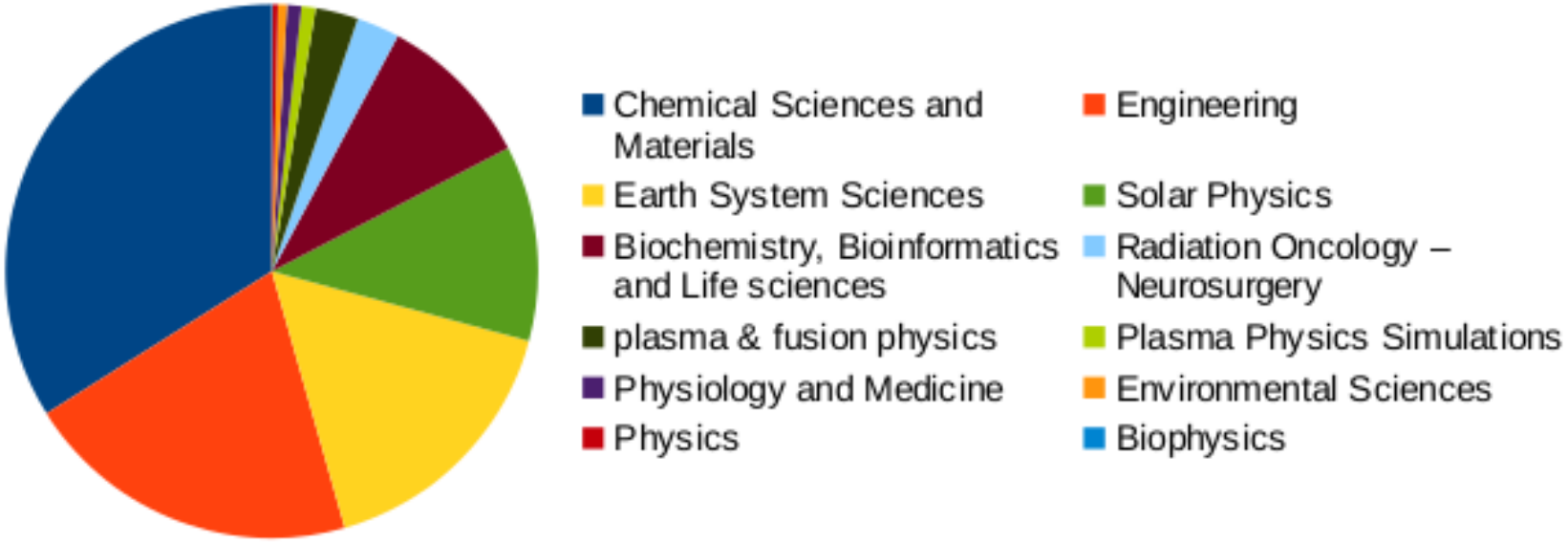
GPU Nodes Yearly



	Max	Average	Current
Nodes Allocated	44 Nodes (100.0%)	27 Nodes (61.4%)	18 Nodes (40.9%)
Total Nodes	44 Nodes (100.0%)	44 Nodes (100.0%)	44 Nodes (100.0%)

System Usage

3rd production call Scientific Fields allocations:



System Usage

3rd production call Institutes allocation:

- Picture is usually different between calls : 12 months allocation, calls every 6 months.

Scientific Results

Publications

grnet hpc.grnet.gr

Τεχνική Περιγραφή | Πρόσβαση | Υποστήριξη | Εκπαίδευση | Διεθνείς Συνεργασίες | Επιστημονικά Αποτελέσματα

Δημοσιεύσεις

search

Types
- article
- inbook
- all types

Calls
- Publications per Call
- pr001
- pr002
- pr003
- preparatory

Years
- 2017
- 2016

Affiliations
- aristotle university of thessaloniki
- athens university of economics

57 results

2017

[57] **Nesterov-based Alternating Optimization for Nonnegative Tensor Factorization: Algorithm and Parallel Implementation** (A. P. Liavas, G. Kostoulas, G. Lourakis, K. Huang and N. D. Sidiropoulos), *In IEEE Transactions on Signal Processing*, volume PP, 2017. [details] [pdf] [doi]

[56] **Neutron-star Radius Constraints from GW170817 and Future Detections** (Andreas Bauswein, Oliver Just, Hans-Thomas Janka and Nikolaos Stergioulas), *In The Astrophysical Journal Letters*, volume 850, 2017. [details] [pdf]

[55] **The influence of the solid to plasma phase transition on the generation of plasma instabilities** (Kaselouris, E. and Dimitriou, V. and Fitilis, I. and Skoulakis, A. and Koundourakis, G. and Clark, E. L. and Bakarezos, M. and Nikolos, I. K. and Papadogiannis, N. A. and Tatarakis, M.), *In Nature Communications*, volume 8, 2017. [details] [pdf] [doi]

[54] **Molecular Simulations of Free and Graphite Capped Polyethylene Films: Estimation of the Interfacial Free Energies** (Sgouros, A. P., Vogiatzis, G. G., Kritikos, G., Boziki, A., Nikolakopoulou, A., Liveris, D. and Theodorou, D. N.), *In Macromolecules*, volume 50, 2017. [details] [pdf] [doi]

[53] **Exploring the interactions of irbesartan and irbesartan–2-hydroxypropyl-β-cyclodextrin complex with model membranes** (Adamantia S. Liossi, Dimitrios Ntountaniotis, Tahsin F. Kellici, Maria V. Chatziathanasiadou, Grigorios Megariotis, Maria Mania, Johanna Becker-Baldus, Manfred Kriechbaum, Andraž Krajnc, Eirini Christodoulou, Clemens Glaubitz, Michael Rappolt, Heinz Amenitsch, Gregor Mali, Doros N. Theodorou, Georgia Valsami, Marinos Pitsikalis, Hermis Iatrou, Andreas G. Tzakos and Thomas Mavromoustakos), *In Biochimica et Biophysica Acta (BBA) - Biomembranes*, volume 1859, 2017. [details] [pdf] [doi]

[52] **Implementation of a two-way coupled atmosphere-ocean wave modeling system for assessing air-sea interaction over the Mediterranean Sea** (George Varlas, Petros Katsafados, Anastasios Papadopoulos and Gerasimos Korres), *In Atmospheric Research*, 2017. [details] [pdf] [doi]

[51] **Monte Carlo and experimental determination of correction factors for gamma knife perfexion small field dosimetry measurements** (E Zoros, A Moutsatsos, E P Pappas, E Georgiou, G Kollias, P Karaiskos and E Pantelis), *In Physics in Medicine and Biology*, volume 62, 2017. [details] [pdf]

[50] **Near Real-Time Aerosol Predictions During the First Citizen Observatory Campaign in Greece** (Athanasopoulou, E. and Spyrou, C. and Apostolopoulou, S. and Papageorgiou, S. and Amiridis, V. and Gerasopoulos, E.), *Chapter in (Karacostas*

Contact

hpc-info@lists.grnet.gr	General Information
hpc-access@lists.grnet.gr	Access Information, reports etc.
support@hpc.grnet.gr	User support
events.hpc.grnet.gr	Events announcement, registration etc.

Access Policy

Access Policy to ARIS
(and other European Systems)

Access Policy, Project Types

- Basic Targets
  - Efficient use of System, maximize scientific production given the resources.
  - Maximize the impact of research projects.

- Production : Periodic call for Production projects (every 6 months, for 1 year). Need to pass both technical and scientific review.
- Preparatory : Open call for projects in order to verify scaling, fit on HPC system etc. Duration 2 months. Only techical review and very basic scientific review.
- Development : Development / modification of Parallel applications. Basic technical and scientific review. Duration 4 months.

Access Policy, Review process

- Call Announchement.
- Call open for $\sim$ 1 month. Applications
- Technical Review
- Reviewers assignment, Scientific Review.
- Summarize technical and scientific reviews, accept or reject.
- Allocation of resources (may be different than what requested, usually less core hours but not only)
- Results announcement, sign AUP, start of project.
- Periodic check of activity.
- Final Report, Results dissemination.
- Follow up : Inform for any publication with results from project.

Notes on applications I

- Read carefully the goals of call announcement and preqrequisities.
- Technical description has the same weight as scientific description.
- Carefully calculate the requested resources.
- Describe the social, scientific etc. impact of your research.
- Describe your team's background in scientific field but also in the use this type of systems.

Notes on applications II

- Describe and give reference to the software you plan to use. In case of multi-method packages describe which methods etc. of package you'll use.
- Describe the problem size of your research.
- Carefully describe why you need an HPC system.
  - Describe the scaling of your application as function of data size/methods etc.
  - The fact that an application is highly scalable does not imply that the same happens with your data.
  - Describe how the code is parallelized (MPI/OpenMP/Hybrid/Other)

Notes on applications III

- Detailed description of application performance with your data on other machines (MachineName, CPU type, Memory etc.)

Application Example

Run Type	no.Runs	Steps/Run	Time/step	no.cores	Total Core Hours
1	20	1000	1s	100	$= 20 \times 1000 \times 1 \times 100/3600$=555.5
2	10	1000000	0.001 s	1000	$= 10 \times 1000000 \times$ $0.001 \times 1000/36000 = 2777$
..					3332.5

You have limited scaling data ?
Apply for a preparatory project to obtain.
Some reasons that may result in reject

- Request memory per core more than the node memory
- Request cores per node more than the maximum available
- You ask for commercial software requiring license that either you don't have or it is locked to certain machines.

Connect to ARIS

- login nodes : Accessibe from Internet, ONLY from certain IPs/Networks.
- SSH ONLY using keys
- compute nodes : Only for running jobs, not directly accessible, no access to internet.
- SSH connections from login to everywhere are not allowed.
- Files transfers : Instead $ARIS \longrightarrow PC$ *put*, $PC \longrightarrow ARIS$ *get*.
- SSH software for Windows : bitvise, mobaXterm, putty.

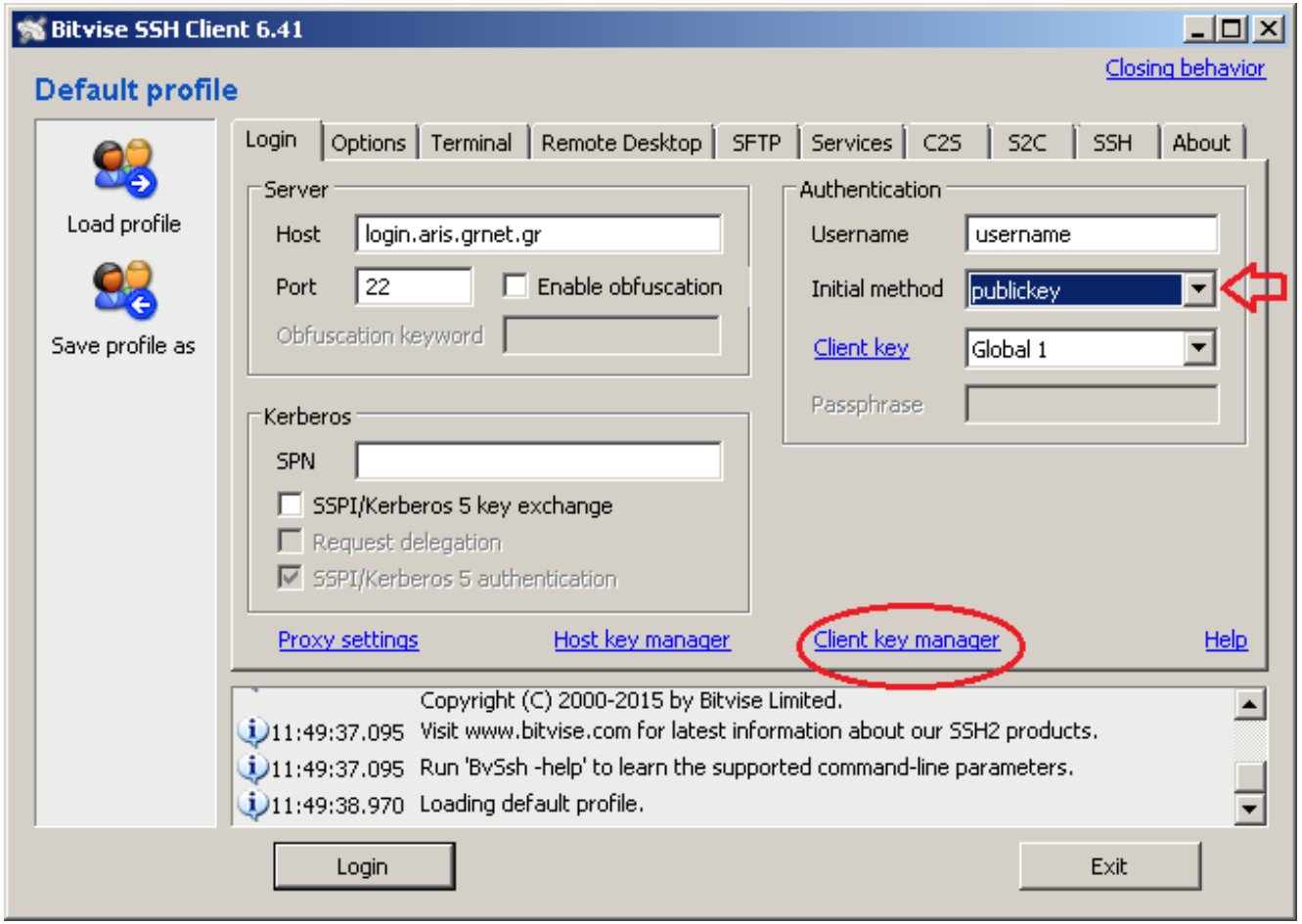
Connect to ARIS

- Your username/private key will be announced.
- Save your private key somewhere
- Linux/Mac Users
  - ssh -i PATHTOprivkey user@login.aris.grnet.gr

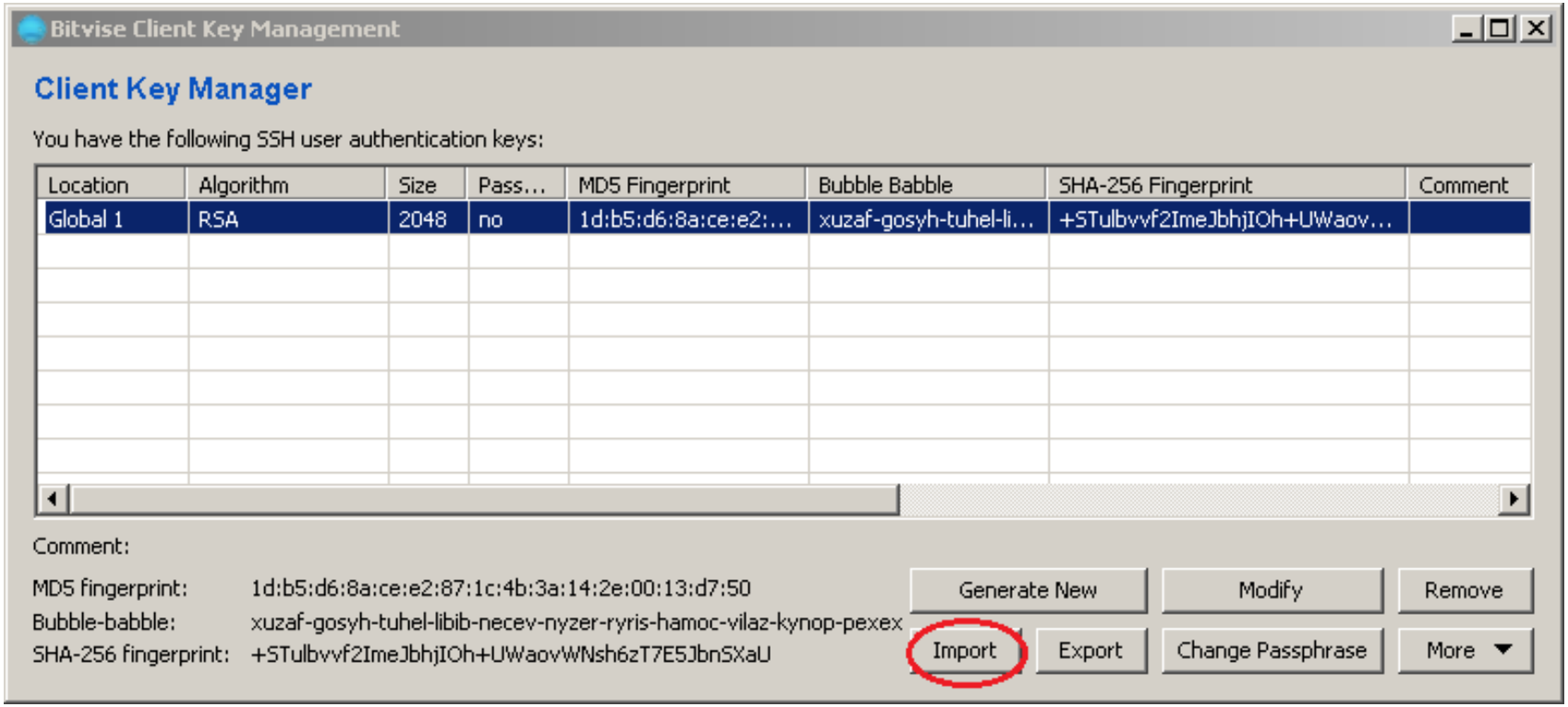
Connect to ARIS

- For Windows users
  - Putty
  - http://www.putty.org/
    - `puttygen id_rsa -o private_key_putty`
    - In SSH → Auth section of putty session configuration select the converted to putty format private key : `private_key_putty`
  - Bitvise
    - https://www.bitvise.com/ssh-client-download
    - Set Host to login.aris.grnet.gr
    - Set your Username
    - Set Initial method to publickey
    - Import your private key (Click “Client key manager”)
    - Click Import
  - mobaXterm https://mobaxterm.mobatek.net/download-home-edition.html

Connect to ARIS



Connect to ARIS



Connect to ARIS

- If everything is correct you'll get a prompt login01 or login02
- You are connected to ARIS login nodes

Environment Modules

- What they are ?
- Dynamic modification of some environment variables, mainly - but not only - `PATH` and `LD_LIBRARY_PATH`
- Easy way to switch between versions

Environment Modules

- What modules are available

  `module avail`

  or

  `module -l avail`

- List active modules

  `module list`

- Deactivate all active modules

  `module purge`

Environment Modules

- Deactivation of a certain module

```
module unload MODULENAME
```

- Switch module version

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

Development Tools

- Available compilers : GNU, Intel, PGI, Sun(Oracle)
- Available MPI Flavors : IntelMPI, OpenMPI, MVapiCH.
- Best Compiler flags, more flags may be needed
- GNU : -O3 -mavx -march=ivybridge
- Intel : -O3 -xCORE-AVX-I
- PGI : -O4 -tp=sandybridge
- MPI :
  - IntelMPI (Intel): mpiicc, mpiicpc, mpiifort
  - OpenMPI(gnu/intel/pgi) : mpicc, mpicxx, mpif90

SLURM Scripts

- A Slurm Script describes the required resources as well as the commands, to run a job
- Script generator and validator http://doc.aris.grnet.gr/scripttemplate/

SLURM Scripts

```
#!/bin/bash
#SBATCH --job-name=”testSlurm” # JobName
#SBATCH --error=job.err.%j     # Filename : stderr
#SBATCH --output=job.out.%j    # Filename : stdout
                               # %j value of JobID
#SBATCH --nodes=2              # Number of nodes
#SBATCH --ntasks=4             # Number of (usually MPI) Tasks
#SBATCH --ntasks-per-node=2    # Number of Tasks / node
#SBATCH --cpus-per-task=10     # Number of Threads / MPI Task
#SBATCH --mem=56G              # Memory per node  #  One of these 2 specs
#SBATCH --mem-per-cpu=2800M    # Memory per core  #
#SBATCH -A ptc                 # Accounting tag # ptc for training
#SBATCH -t 1-01:00:00          # Requsted  DD-HH:MM:SS
#SBATCH -p compute             # partition, one of compute, gpu, phi, fat, taskp, short
module purge
module load gnu/4.9.2
module load intel/15.0.3
module load intelmpi/5.0.3
if [ x$SLURM_CPUS_PER_TASK == x ]; then            #
  export OMP_NUM_THREADS=1                         #
else                                               # Never delete these lines
  export OMP_NUM_THREADS=$SLURM_CPUS_PER_TASK      # unless you exactly know what you do
fi                                                 #
srun EXECUTABLE ARGUMENTS   # Executable and possible arguments
```

SLURM Scripts

- DO NOT use the typical mpirun/mpiexec(.hydra). Use srun for SLURM.
- You may omit some of requirements if the rest can define the required resources
- Examples : You may omit ntasks, requires nnodes, ntasks-per-node, cpus-per-task to be defined. System can calculate how many tasks to use
- Especially for hybrid MPI/OpenMP applications DO NOT delete the piece of code that checks if you set correctly threads/tasks : A common mistake in production runs.
- Required time is mandatory. If you omit it, either job will never run (default for ARIS) or will use the default maximum wall time (2 days for aris)

Communicating with SLURM

- Job Submission

```
sbatch SLURM_JobScript.sh
Submitted batch job 123456
```

- Job List

```
squeue
```

- Detailed Job List

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

Communicating with SLURM

- Job Cancel

```
scancel JobID
```

- Send KILL signal (instead of the default TERM) to a job

```
scancel -s KILL JobID
```

- Estimation of job start time that is queued due to not available resources

```
squeue --start
```

- Information for the resources status.

```
sinfo
```

SLURM User/Group resource limits

- Each account has certain resource limits.
  - Maximum number of running Jobs
  - Maximum number of Jobs in queue
  - Maximum number of concurrently used cores and/or nodes
  - Maximum Wall time duration of a job
  - Maximum consumable Core Hours for project duration (=Budget).

Accelerators with SLURM

- GPU
  #SBATCH –partition=gpu
  #SBATCH –gres=gpu:2
  Variable : `SLURM_JOB_GPUS=0,1` και
  `CUDA_VISIBLE_DEVICES=0,1`
- Xeon Phi (Coprocessors => Offload ONLY)
  #SBATCH –partition=phi
  #SBATCH –gres=mic:2
  Variable : `OFFLOAD_DEVICES=0,1`

Hands On

- Connect to training System
- Examine the available environment modules, load module, check what changes it implies in environment.
- Purge modules, check for mpicc, load intelmpi - recheck for mpicc.
- Create a Slurm Script with script generator, modify it, use simple commands like date, hostname etc. submit it, check status, see stdout, stderr upon completion (not in queue).