Plateforme Bioinfo Occitanie Toulouse

Genotoul Bioinfo

SLURM cluster

(15/03/18)

Infrastructure

Environment

SLURM cluster: Infrastructure (15/03/18)

Context
=> genotoul team, users, contributions, renewal strategy, schedule

- **Infrastructure**
  => service & compute nodes, diagram
- **What does not change ?**
  => users accounts,
  => disk spaces : /home, /save, /work, /bank
  => user quotas
- **SLURM vs SGE**
  => system evolution
  => commands
  => directives
  => sample sbatch script
  => sample MPI sbatch script
- **For further with SLURM**
  => job arrays
  => job dependencies
  => format option

Genotoul Bioinfo

Context
Team

Christine Gaspin
DR INRA / Scientific animation
+33 (0)5 61 28 52 82
christine.gaspin(at)inra.fr

Christophe Klopp
IR INRA / Technical animation
+33 (0)5 61 28 50 36
christophe.klopp(at)inra.fr

Claire Hoede
IR INRA / Development and data analysis
+33 (0)5 61 28 53 05
claire.hoede(at)inra.fr

Didier Laborie
IE INRA / System administrator
+33 (0)5 61 28 54 27
didier.laborie(at)inra.fr

Jérôme Mariette
IE INRA / Development and data analysis
+33 (0)5 61 28 57 25
jerome.mariette(at)inra.fr

Marie-Stéphane Trotard
IE INRA / System administrator
+33 (0)5 61 28 52 76
marie-stephane.trotard(at)inra.fr

Céline Noirot
IE INRA / Development and data analysis
+33 (0)5 61 28 57 24
celine.noirot(at)inra.fr

Floréal Cabanettes
CDD IE France génomique / Development and data analysis
+33 (0)5 61 28 54 93
floreal.cabanettes(at)inra.fr

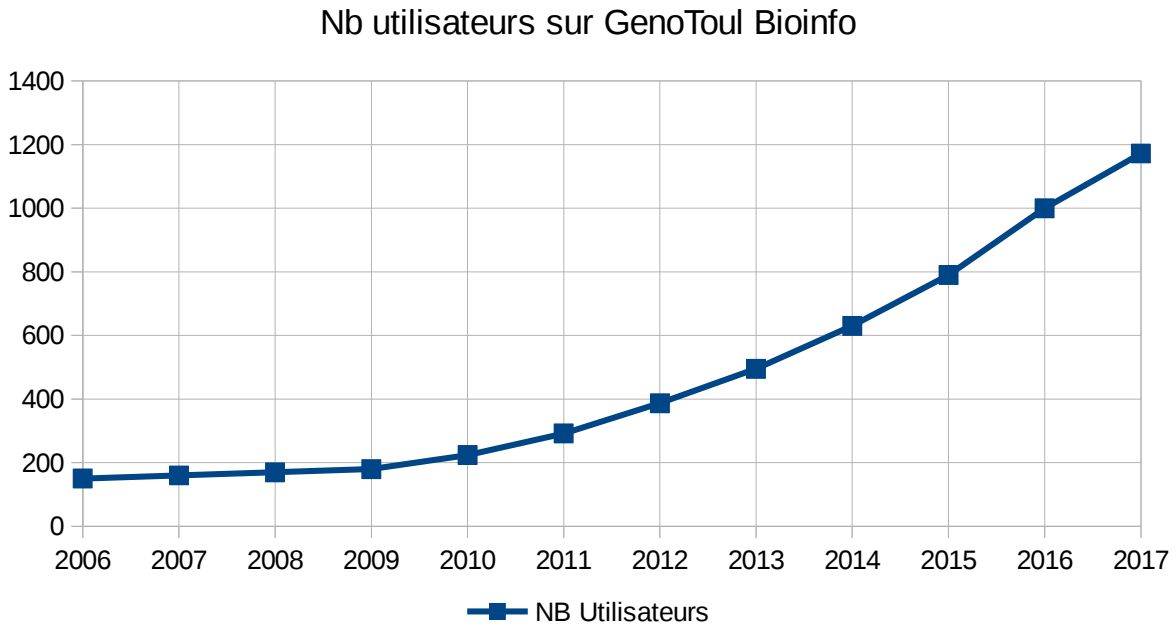
GenoToul Bioinfo

Context
users, contributions

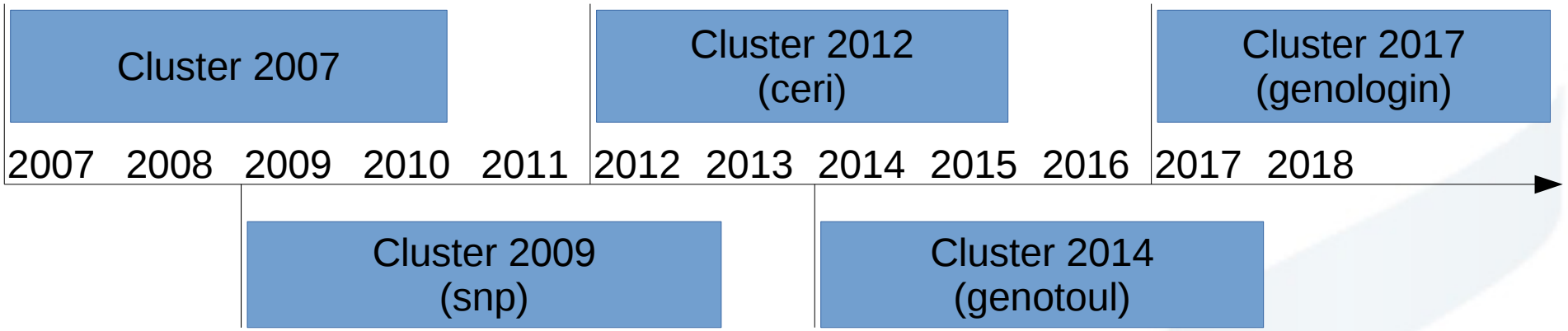
- 1200 user accounts
  (50% INRA, 50 % OCCITANIE)
- 845 cluster users
  (galaxy users included)
- 1800 annual support tickets

La Région Occitanie Pyrénées - Méditerranée

INRA SCIENCE & IMPACT



Context
renewal strategy



- Overlapping clusters enabling to keep the service active and to renew the machines
- But this time we have to change the job scheduler

Context
schedule

Step 1 (07/17 – 10/17)

- Compute capacity reduction (to make room)
- End of use 34 CERI nodes

Step 2 (11/17 – end of 2018)

- 2 compute clusters cohabitation
- Progressive migration of applications to SLURM

Step 3 (2019)

- End of use genotoul / SGE cluster

Infrastructure
service & compute nodes

Service nodes

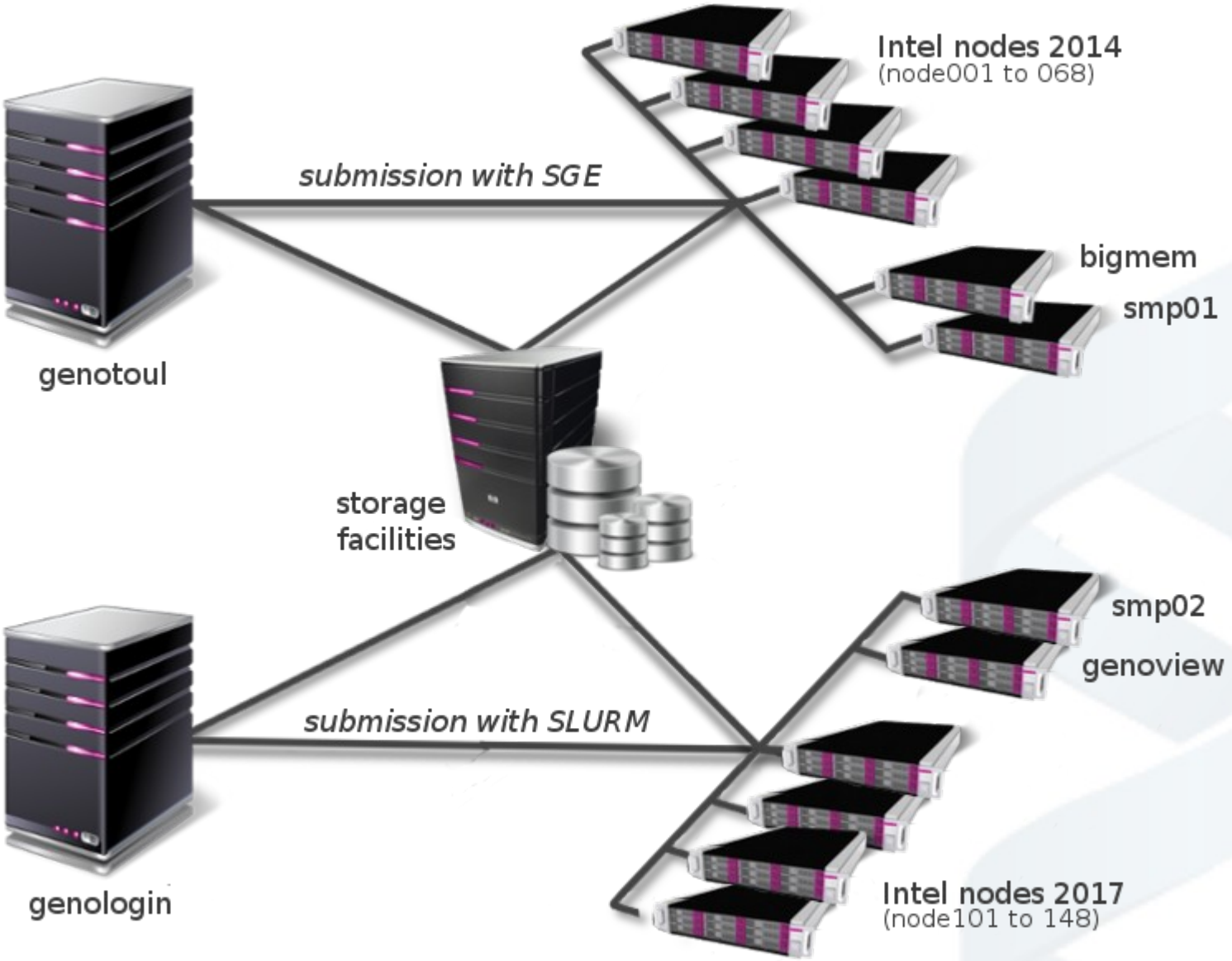
- 2 login nodes : genologin1&2 * (32 cores, 128 GB RAM)
- 1 admin node : genobull (20 cores, 128 GB RAM)

Compute nodes

- 1 visualization node : genoview (32 cores, 128GB, Nvidia K40)
- **48 compute nodes : [101 à 148] * (32 cores, 256 GB)**
- 1 SMP node: genosmp02 (48 cores, 1536GB RAM, 22TB HD)
- Low latency & high bandwidth interconnection (56GB/s)

SLURM cluster : 1584 cores / 3168 threads / 51 TFlops

Infrastructure



What does not change

- **Users accounts**

All of your genotoul linux accounts are available on SLURM cluster
=> use **genologin** server **instead of genotoul,**

- **Disk spaces**

All of your directories (/home, /save, /work) are the same.
=>you don't have to copy anything between **genotoul** and **genologin**

- **User quotas**

All of group/personnal quotas are the same :
100MB for /home directory (configuration files only)
250GB for /save directory (backuped)
1TB for /work directory (tempory disk space)
100,000H annual calculation time (500H for private user)

SLURM (vs SGE)
System evolution

SLURM (before: SGE)

- Simple Linux Utility for Resource management
- Adopted by the academic community
- Supported by IT providers
- New features

CentOS-7 (before: CentOS-6.5)

- Community ENTerprise Operating System
- Supported by IBM Spectrum Scale
- Cgroups (Control Groups) compatible

SLURM (vs SGE)
Commands (1/2)

Job submission

[BATCH]

- **sbatch** (qsub) : submit a batch job to slurm (default workq partition()
- **sarray** (qarray) : submit a batch job-array to slurm
- **scancel** (qdel) : kill the specified job

[INTERACTIVE]

- **srun [--pty bash]** (qrsh) : submit an interactive session with a compute node (default workq partition).
- **runVisuSession.sh** (qlogin) : submit a TurboVNC / VirtualGL session with the graphical node (interq partition). Just for graphics jobs.

SLURM (vs SGE)
Commands (2/2)

Job holding

- **scontrol** hold (qhold) : job hold
- **scrontol** release (qrls) : job release

Job monitoring

- **sinfo** (qhost) : display nodes, partitions, reservations
- **squeue** (qstat) : display jobs and state
- **scontrol show** : get informations on jobs, nodes, partitions
- **sstat** (qstat -j) : show status of running jobs
- **sview** (qmon) : graphical user interface
- **sacct** (qacct) : display accounting data

SLURM (vs SGE)
Directives (1/2)

SLURM		SGE
-p workq	#partition name	**-q workq**
--time=00:10:00	#job time limit	**-l h_ rt=00:10:00**
-J testjob	#job name	**-N testjob**
-o output.out	#output file	**-o output.out**
-e error.out	#error file name	**-e error.out**
--mem=8G or --mem-per-cpu	#memory size	**-l mem=8G, h_vmem=10G**

SLURM (vs SGE)
Directives (2/2)

SLURM		SGE
--cpus-per-task=4	#ncpu on the same node	**-l parallel_smp 4**
--mail-type=[events]	#event notification	**-m bea**
--mail-user=[address]	#default LDAP account's	**-M**
--export=[ALL|NONE|variables]	#copy environment	**-V**
--workdir=[dir_name]	#working directory	**-wd**

SLURM
Sample sbatch script

```
# !/bin/bash
#SBATCH --time=00:10:00 #job time limit
#SBATCH -J testjob      #job name
#SBATCH -o output.out   #output file name
#SBATCH -e error.out    #error file name
#SBATCH --mem=8G        #memory reservation
#SBATCH --cpus-per-task=4       #ncpu on the same node
#SBATCH --mail-type=BEGIN,END,FAIL (email address is LDAP account's)
#Purge any previous modules
module purge
#Load the application
module load bioinfo/ncbi-blast-2.2.29+
# My command lines I want to run on the cluster
blastall ...
```

SLURM

Sample MPI sbatch script

```
# !/bin/bash
#SBATCH -J mpi_job       #job name
#SBATCH --nodes=2        #2 different nodes
#SBATCH --tasks-per-node=4          #4 tasks per node
#SBATCH --cpus-per-task=2        #2 cpu per task
#SBATCH --time=00:10:00 #job time limit

cd $SLURM_SUBMIT_DIR
module purge
module load compiler/intel-2018.0.128 mpi/openmpi-1.8.8-intel2018.0.128
mpirun -n $SLURM_NTASKS -npernode $SLURM_NTASKS_PER_NODE ./hello_world
```

SLURM
Job arrays

sbatch -a | array=<indexes>

Submit a job array, multiple jobs to be executed with identical parameters.

Multiple valued may be specified using a comma separated list and/or a range of values with a « - » separator.

Example :

--array=1-10

--array=0,6,16-32

--array=0-16:4 a step of 4

--array=1-10%2 a maximum of 2 simultaneously ruuning task

Variable	Correspondance
SLURM_ARRAY_TASK_ID	Job array ID (index) number
SLURM_ARRAY_JOB_ID	Job array's master job ID number
SLURM_ARRAY_TASK_MAX	Job array's maximum ID (index) number
SLURM_ARRAY_TASK_MIN	Job array's minimum ID (index) number
SLURM_ARRAY_TASK_COUNT	total number of tasks in a job array

SLURM
Job dependencies

sbatch -d | --dependency=<dependency_list>

Defer the start of this job until the specified dependencies have been satisfied completed.

<dependency_list> is on the form <type :jobID[:jobID][,type :jobID[:jobID]]>

Example :

Sbatch --dependency=afterok:6265 HELLO.job

Type	Correspondance
after	this job can begin execution after the specified jobs have begun execution
afterany	this job can begin execution after the specified jobs have terminated
afterok	This job can begin execution after the specified jobs have successfully executed (ran to completion with an exit code of zero)
afternotok	This job can begin execution after the specified jobs have terminated in some failed state (non-zero exit code, node failure, timed out, etc)

SLURM

--format option

Some commands (like **sacct** and **squeue**) give the possibility to **tune output format** :

Example :

```
sacct --format=jobid%-13,user%-15,uid,jobname%-15,state%20,exitcode,Derivedexitcode,nodelist% -X –
job 6969
```

```
        JobID            User    UID         JobName                State ExitCode DerivedExitCode        NodeList
------------ -------------- ------ -------------- -------------------- -------- -------------- ---------------
6969          root                0 toto                        COMPLETED      0:0              0:0   node[101-102]
```

```
squeue --format="%10i %12u %12j %.8M %.8l %.10Q %10P %10q %10r %11v %12T %D %R" -S "T"
```

```
JOBID      USER         NAME            TIME TIME_LIM   PRIORITY PARTITION  QOS        REASON     RESERVATION STATE        NODES NODELIST(REASON)
6612       root         bash           16:09 4-00:00:00          1 workq      normal     None       (null)      RUNNING      2 node[101-102]
6542       dgorecki     TurboVNC    1-06:27:44 UNLIMITE          1 interq     normal     None       (null)      RUNNING      1 genoview
```

SLURM cluster : Environment (15/03/18)

New cluster -> New organisation -> New practices -> New services

- **New environment**
  - => user access, Environment modules package
  - => SLURM vs SGE
  - => software installation, search/find software
- **Software usage**
  - => run a soft, module command options
  - => create your own module
- **Help**
  - => software help, other help
- **New services**
  - => monitoring, account information, saccount_info command
- **Best pratices**
  - => security, shared resources

New environment
User access

Access to the platform: via a command line SSH connection (putty or MobaXterm for Windows)

New frontal/login servers: genologin1 & 2

New hostname for the connection: genologin.toulouse.inra.fr

Example

$ ssh <login>@genologin.toulouse.inra.fr

From March 16th, new accounts will be oriented to SLURM cluster.

New environment
Environment modules

The **Environment Modules package** provides for the dynamic modification of a user's environment via modulefiles.

module command alter or set shell environment:

- add command in your PATH
- define specific environment variable
- add path to dependencies
- add path to specific librairies

Modules can be loaded and unloaded dynamically.

Modules are useful in managing different versions of applications.

New environment
SLURM vs SGE cluster

Genotoul/SGE	Genologin/SLURM
Default environment (current version software) and, sometimes, use of « module » command	Always use of « module » command
Current versions difficult to maintain (dependencies)	No maintenance. Dependencies managed by loading modules
Linux CentOS 6.5: installation with old librairies version	Linux CentOS 7.3: new libraries → new installations
Random help for use on cluster	Help to use on cluster with example directory and « How_to_use » file

New environment
Software installation

SGE to SLURM cluster: new installations

- Some software already installed on SLURM cluster (~140 with versions/ + 840 without versions on SGE)
- Software already installed on SGE **but not** on SLURM cluster or new software:

  on demand by filling form on our website (Ask for/Software installation)

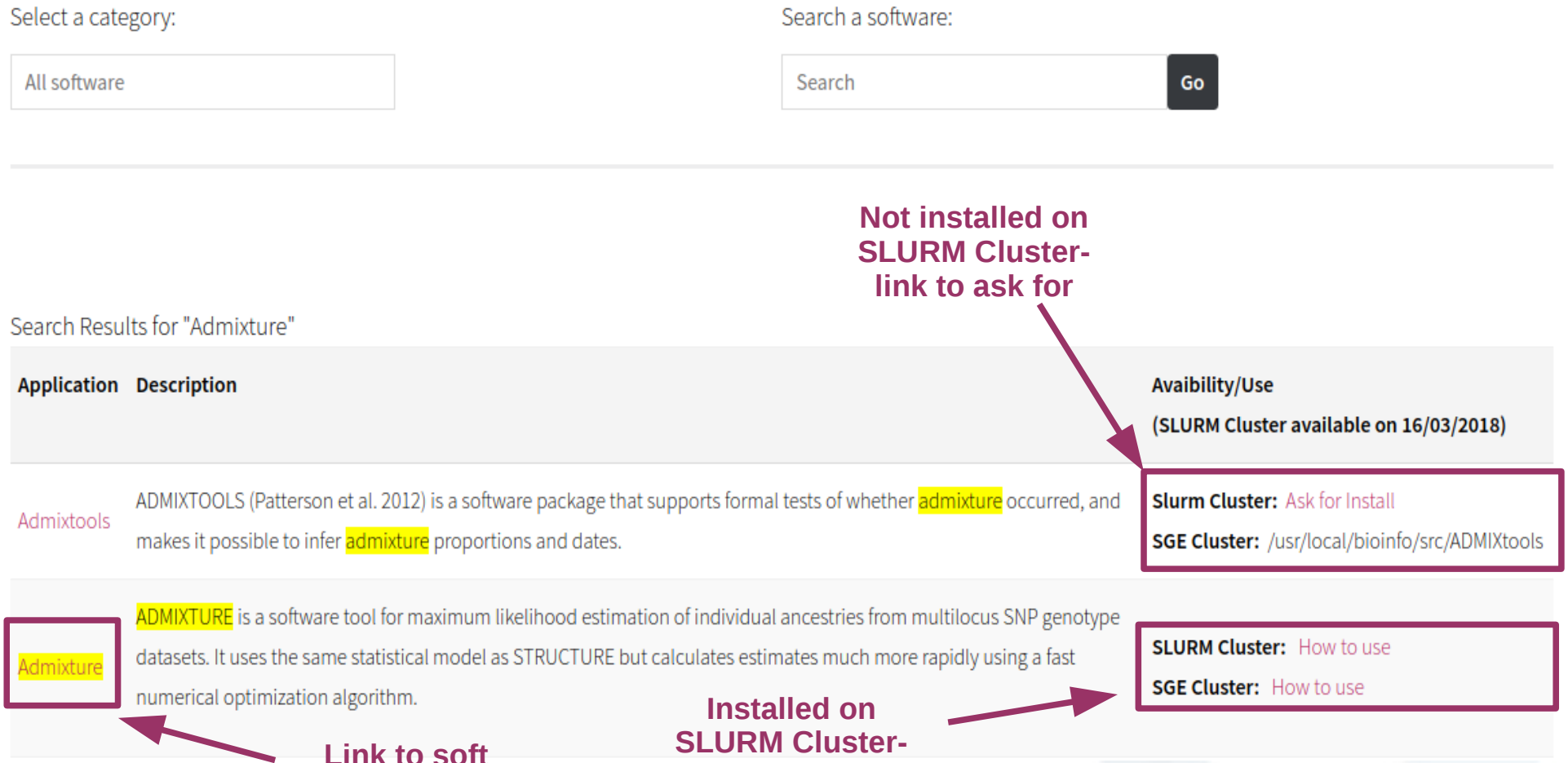
  http://bioinfo.genotoul.fr/index.php/ask-for/install-soft/

Installation and update

- Default installation on Genologin/Slurm from March 16, 2018 (and, exceptionally, on demand, on Genotoul/SGE cluster)
- Reminder : updates only upon user request

New environment
Search/Find a soft (Web)

Website (Resources/Software): http://bioinfo.genotoul.fr/index.php/resources-2/softwares/



New environment
Search/Find a soft (CLI)

New installation paths

same path as SGE cluster but new storage spaces

Bioinfo -> /usr/local/bioinfo/src/

Compilers → /tools/compilers

Libraries → /tools/librairies

Others system tools → /tools/others_tools

Languages (Python, R , Java..) → /tools

Useful scripts → /tools/bin (sarray, squota_cpu, saccount_info…). In user's default PATH.

Commands

- with ls command: **$ ls /usr/local/bioinfo/src|grep -i <soft_name>**

- with module command:

$ module avail: display all available software installed on the cluster

$ module avail <category/soft_name>: display available versions for a specific application (with category in bioinfo,compiler,mpi or system) (case sensitive)

$ module avail -t 2>&1 | grep -i <soft_name>: display available versions for a specific application (case insensitive)

New environment
Search examples

$ module avail bioinfo/cutadapt

$----------/tools/share/Modules ----------

bioinfo/cutadapt-1.14-python-2.7.2 bioinfo/cutadapt-1.14-python-3.4.3

$ module avail -t 2>&1 | grep -i blast

bioinfo/blast-2.2.26

bioinfo/ncbi-blast-2.2.29+

bioinfo/ncbi-blast-2.6.0+

Software usage
Run a soft

Run a software

To run a software you need to load the corresponding module.

$ module load <modulename>

To run a software with others software dependencies, you need to load all required modules.

Best practices

Check modules already loaded : **$ module list**

Purge modules already loaded if not needed :

$ module purge (all modules)

$ module unload modulename (only one module)

Software usage
Usage examples

Use Bismark_v0.19.0

$ module load bioinfo/Bismark_v0.19.0

Need bowtie or bowtie2 and samtools, so :

$ module load bioinfo/bowtie2-2.3.3.1

$ module load bioinfo/samtools-1.4

$ module load bioinfo/Bismark_v0.19.0

$ which bismark

/usr/local/bioinfo/src/Bismark/Bismark_v0.19.0/bismark

$ bismark --help

Use Python-2.7.2

$ module load system/Python-2.7.2

$ which python

/tools/python/2.7.2/bin/python

$ python --help

Software usage
Module command

The basic command to use is **module**:

module : (no arguments) print usage instructions

module avail : list available software module

module load modulename : add a module to your environment

module unload modulename : unload remove a module

module purge : remove all modules

module show modulename : show what changes a module will make to your environment

module help modulename : path to the "How_to_use_SLURM_<soft_name>" file

For more documentation, see the Environment Module website :

http://modules.sourceforge.net/

Software usage

Create your own default environment

Create your modulefile

Basic example : ~/save/my_own_module :

```
#%Module1.0#####################################################################
##

module load bioinfo/bowtie2-2.2.9
prepend-path PATH /usr/local/bioinfo/src/MultAlin/multalin.5.4.1/
setenv MULTALIN /usr/local/bioinfo/src/MultAlin/multalin.5.4.1/
```

Load your module

$ module load -f ~/save/my_own_module

Verify your environment

$ which bowtie2

/usr/local/bioinfo/src/bowtie/bowtie2-2.2.9/bowtie2

Unload one module

$ module unload bioinfo/bowtie2-2.2.9

Unload all module and specific variable

$ module purge

Help
Software help (1/2)

Software documentation

- official software documentation in the installation folder /usr/local/bioinfo/src/<soft_name>/<soft_version>

- our website Software page: link to software website.

Use on SLURM cluster

- "How_to_use_SLURM_<soft_name>" file:

  software installation directory /usr/local/bioinfo/src/<soft_name>

  our website Software page (Avaibility/Use column, click on SLURM cluster link).

- a basic « example_on_cluster » directory in the software installation directory

  /usr/local/bioinfo/src/<soft_name>/example_on_cluster

Help
Software help (2/2)

HOW TO USE ON SLURM CLUSTER

```
SOFT : samtools
------
Site du soft: http://samtools.sourceforge.net
------------

LICENSE:
--------

The MIT/Expat License

See software documentation for more informations.

Location: /usr/local/bioinfo/src/samtools
---------
```

Software informations

```
Load binaries and environment:
------------------------------

-> Version v0.1.19

module load bioinfo/samtools-0.1.19

-> Version 1.3.1

module load bioinfo/samtools-1.3.1

-> Version v1.4

module load bioinfo/samtools-1.4

or use absolute path
```

Usage and versions

```
Example directory for use on cluster:
-------------------------------------

/usr/local/bioinfo/src/samtools/example_on_cluster

To submit:

sbatch test_v1.4.sh
```

Example

Help
Help example

With Admixture on our website Software page

With Bowtie in command line

$ ls /usr/local/bioinfo/src/bowtie/

bowtie-1.2.1.1 bowtie-1.2.1.1-linux-x86_64.zip bowtie2-2.2.9 bowtie2-2.3.3.1
bowtie2-2.3.3.1-linux-x86_64.zip example_on_cluster How_to_use_SLURM_bowtie

$ ls /usr/local/bioinfo/src/bowtie/example_on_cluster/

errot.txt example lambda_virus.1.bt2 lambda_virus.2.bt2 lambda_virus.3.bt2
lambda_virus.4.bt2 lambda_virus.rev.1.bt2 lambda_virus.rev.2.bt2 output.txt
test_v2-2.2.9.sh

Help
Other help

- **Find "How_to_use_SLURM_<soft_name>" file path**

$ module help bioinfo/bowtie2-2.2.9

---------- Module Specific Help for 'bioinfo/bowtie2-2.2.9' ----------

See How_to_use file: /usr/local/bioinfo/src/bowtie/How_to_use_SLURM_bowtie

- **Browse all "How_to_use_SLURM_<soft_name>" files** (in your web browser)

http://vm-genobiotoul.toulouse.inra.fr/How_to_Softs/

- **Useful scripts** (already in your default path or /tools/bin)

saccount_info, **sq_long** or **sq_debug** (squeue long format), **sa_debug** (sacct long format), **sarray** (wrapper for sbatch to create a job array that will execute in parallel SHELL lines from a file), **squota_cpu** to see your CPU time limit.

- **Updated FAQ:** http://bioinfo.genotoul.fr/index.php/faq/

New services
Cluster monitoring

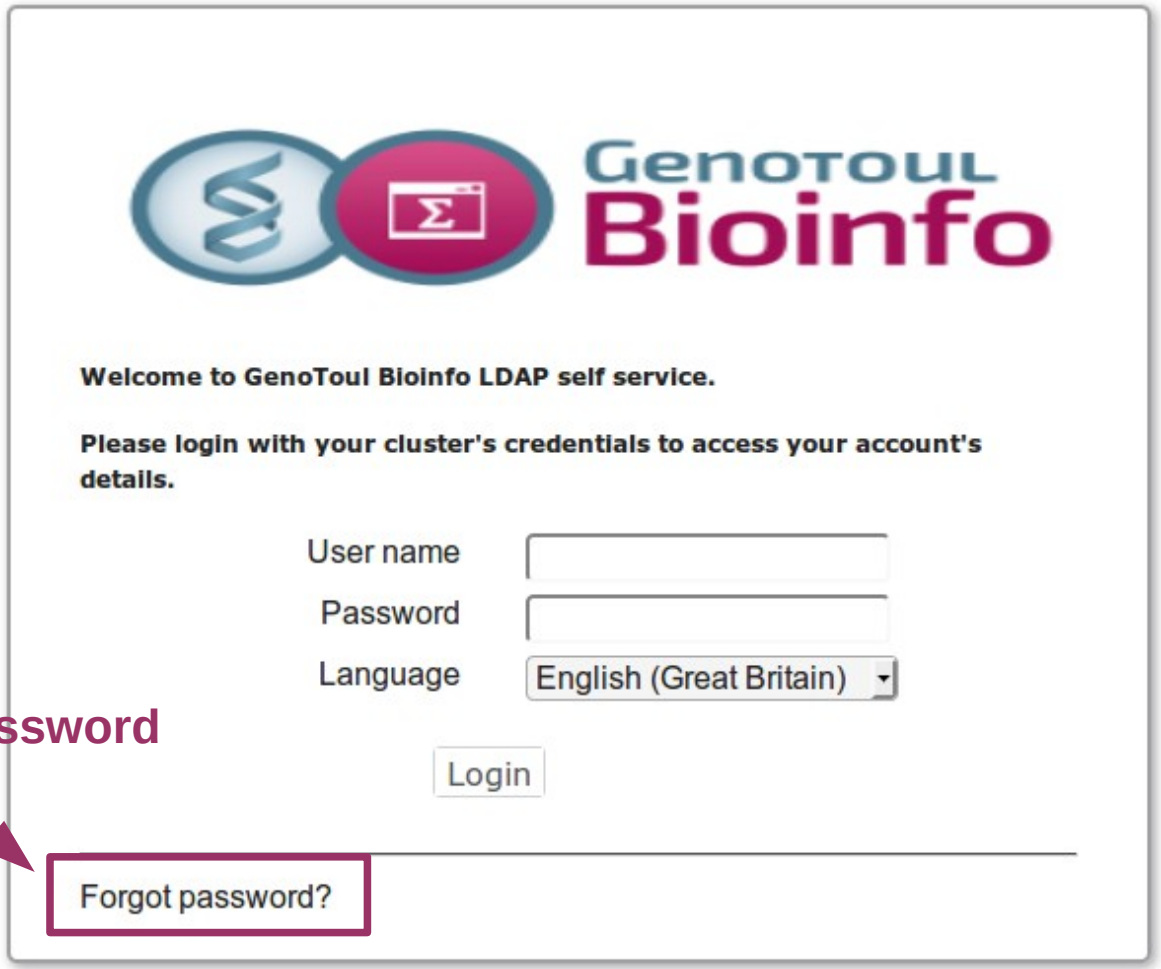
Grafana → http://monitoring.bioinfo.genotoul.fr

(or our website : Resources/Monitoring)

New services

Account information and password change

Self Service → http://selfservice.bioinfo.genotoul.fr



Change your password (every year)

New services
saccount_info command

saccount_info <login>

- account expiration date and last password change date
- your primary Linux group
- your secondary Linux groups if you have any
- status of your Linux primary group in Slurm (contributors, inraregion or others)
- your groups' members
- some Slurm limitations of your account : cpu and memory limit, CPU Time ...

Best pratices
Security

One user = one account

You are responsable of the damage caused by your login.

Default permissions directories

- home: drwxr-x—x : **R**ead, **W**rite, e**X**ecution for the owner, **R**ead and e**X**ecution for the group members, e**X**ecution for all.

- save and work: drwxr-x--- : **R**ead, **W**rite, e**X**ecution for user, **R**ead and **E**xecution for your group members, no permissions for all.

To change permissions: **chmod** command

Best pratices
Resources

Cluster is a shared resource, so ... think about others

- try to adapt requested resources to your needs.

- DO NOT run treatments on frontal servers:

Why ?

- overloading frontal servers slow down everyone.
- overloading frontal servers can crash frontal servers and block everyone.
- more time for the administrators to answer support requests.

Check your process on frontal servers : **$ pstree -u <login>**

Any treatment launched on the servers "genologin" will be immediately killed by the system administrators

Support

- **Bioinfo genotoul website :**

http://bioinfo.genotoul.fr/

- **Bioinfo Genotoul Chart**

http://bioinfo.genotoul.fr/wp-content/uploads/ChartPFBioinfoGenoToul.pdf

- **FAQ**

http://bioinfo.genotoul.fr/index.php/faq/

- **Support**

Mail: support.bioinfo.genotoul@inra.fr

Fill form (best for us): http://bioinfo.genotoul.fr/index.php/ask-for/support/

End of Presentation

Thanks for your attention !

Further informations
Partitions (queues)

Each job is submitted to a specific partition (the default one is the workq).

Each partition has a different priority considering the maximum time of execution allowed.

Partitions (queues)	Access	Priority	Max time	Max slots
workq	everyone	100	4 days (96h)	3072
unlimitq	everyone	1	180 days	500
interq (runVisusession.sh)	on demand		1 day (24h)	32
smpq	on demand		180 days	96
wflowq	specific software		180 days	3072

Further informations

More SLURM directives (+)

--depend=[state:job_id]	**#partition name**	**(-hold_jid)**
--nodelist=[nodes]	**#host preferrence**	**(-l hostname)**
--array=[array_spec]	**#job arrays**	**(-t)**
--begin=[datetime]	**#begin time**	**(-a)**
--exclusive or shared	**#resource sharing**	**(-l exclusive)**

Further informations

SLURM variables (+)

$SLURM_JOBID	**#jobID**	**($JOB_ID)**
$SLURM_SUBMIT_DIR	**#submit directory**	**($SGE_O_WORKDIR)**
$SLURM_SUBMIT_HOST	**#submit host**	**($SGE_O_HOST)**
$SLURM_NODELIST	**#node list**	**($PE_HOSTFILE)**
$SLURM_ARRAY_TASK_ID	**#job array index**	**($SGE_TASK_ID)**

$SLURM_NNODES	**(#SBATCH -N)**
$SLURM_NTASKS	**(#SBATCH -n)**
$SLURM_NTASKS_PER_NODE	**(#SBATCH –task-per-node)**
$SLURM_CPUS_PER_TASK	**(#SBATCH -c)**