UL HPC School 2018

Overview & Challenges of the UL HPC facility at the EuroHPC Horizon

Prof. P. Bouvry, Dr. S. Varrette

V. Plugaru, S. Peter, H. Cartiaux & C. Parisot

June 12th, 2018, MSA 4.520

University of Luxembourg (UL), Luxembourg

Welcome to the UL HPC School 2018

https://hpc.uni.lu/hpc-school/

- **7th edition** of this training. . .
  - ↪ started in 2014
  - ↪ This one is the **long** version
    - ✓ 2-days event
    - ✓ Parallel sessions, feat. **basic & advanced** tutorials
- **Requirement**:
  - ↪ your favorite laptop with your favorite OS
    - ✓ Linux / Mac OS preferred, but Windows accepted
  - ↪ basic knowledge in Linux command line
  - ↪ ability to take notes (Markdown etc.)

Online UL HPC Tutorials

http://ulhpc-tutorials.rtfd.io/

Agenda Day 1: June 12th, 2018

Day 1	Main Track (MSA **4.520**)	Speaker
9h00	**PS1a**: Getting Started (part I: SSH)	C. Parisot
10h00	**Coffee break**	
10h30	**Overview & Challenges of the UL HPC Facility at the EuroHPC Horizon**	S. Varrette
11h45	**PS1b**: Getting Started (part II)	H. Cartiaux
12h30	**LUNCH**	
13h30	**PS2**: HPC workflow with sequential jobs	H. Cartiaux
14h30	**PS4a**: UL HPC Monitoring in practice	H. Cartiaux
15h30	**Coffee break**	
16h00	**PS5**: Parallel computations with OpenMP/MPI	S. Varrette
17h30	**PS6**: User environment and storage data management	S. Peter

Day 1	Parallel Track (MSA **4.410**)	Speaker
10h30	**Keynote in 4.520**	
13h30	**PS3a**: Advanced scheduling (SLURM, OAR)	V. Plugaru
14h30	**PS3b**: Software environment management using Easybuild	S. Peter
15h30	**Coffee break (in 4.520)**	
16h00	**PS4b**: Performance engineering - HPC debugging and profilling	V. Plugaru

PS = *Practical Session using your laptop*

Agenda Day 2: June 13th, 2018

Day 2	Main Track (MSA **4.520**)	Speaker
9h00	**PS7**: Multi-Physics workflows: test cases on CFD/MD/Chemistry applications	V. Plugaru
10h30	**Coffee break**	
11h00	**Users' session: UL HPC experiences**	
12h40	**LUNCH**	
13h30	**PS9**: [Basic + Advanced] Prototyping with Python	C. Parisot
15h30	**Coffee break**	
15h45	**PS11**: Big Data Applications	S. Varrette
17h15	**PS13**: Machine / Deep learning (Pytorch, Tensorflow, Caffe2)	S. Varrette
18h15	**Closing Keynote: Take Away Messages**	S. Varrette

Day 2	Parallel Track (MSA **4.510**)	Speaker
9h00	**PS8**: Bio-informatics workflows and applications	S. Peter
10h30	**Coffee break (in 4.520)**	
11h00	**Users' session: UL HPC experiences** (in 4.520)	
12h40	**LUNCH**	
13h30	**PS10**: Scientific computing using MATLAB	V. Plugaru
15h30	**Coffee break**	
15h45	**PS12**: R - statistical computing	A. Ginolhac
17h15	**PS14**: HPC Containers: Singularity	V. Plugaru
18h15	**Closing Keynote in 4.520**	S. Varrette

Summary

Summary

Summary

Prerequisites: Metrics

- **HPC**: **H**igh **P**erformance **C**omputing — **BD**: **B**ig **D**ata

Main HPC/BD Performance Metrics

- **Computing Capacity**: often measured in **flops** (or **flop/s**)
  - ↪ **Fl**oating **p**oint **o**perations per **s**econds (often in DP)
  - ↪ **GFlops** $= 10^9$ **TFlops** $= 10^{12}$ **PFlops** $= 10^{15}$ **EFlops** $= 10^{18}$

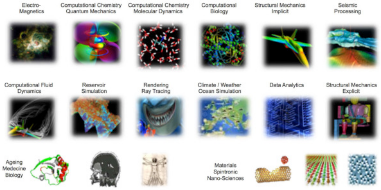
Prerequisites: Metrics

- **HPC**: **H**igh **P**erformance **C**omputing — **BD**: **B**ig **D**ata

Main HPC/BD Performance Metrics

- **Computing Capacity**: often measured in **flops** (or **flop/s**)
  - ↪ **Fl**oating **p**oint **o**perations per **s**econds (often in DP)
  - ↪ **GFlops** $= 10^9$ **TFlops** $= 10^{12}$ **PFlops** $= 10^{15}$ **EFlops** $= 10^{18}$
- **Storage Capacity**: measured in multiples of **bytes** = 8 **bits**
  - ↪ **GB** $= 10^9$ bytes **TB** $= 10^{12}$ **PB** $= 10^{15}$ **EB** $= 10^{18}$
  - ↪ **GiB** $= 1024^3$ bytes **TiB** $= 1024^4$ **PiB** $= 1024^5$ **EiB** $= 1024^6$
- Transfer rate on a medium measured in **Mb/s** or **MB/s**
- Other metrics: Sequential vs Random **R/W speed**, **IOPS** ...

Why HPC and BD ?



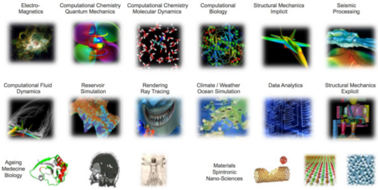
HPC: **H**igh **P**erformance **C**omputing
BD: **B**ig **D**ata

Andy Grant, Head of Big Data and HPC, Atos UK&I

To out-compete you must out-compute

Increasing competition, heightened customer expectations and shortening product development cycles are forcing the pace of acceleration across all industries.

Why HPC and BD ?



HPC: **H**igh **P**erformance **C**omputing
BD: **B**ig **D**ata

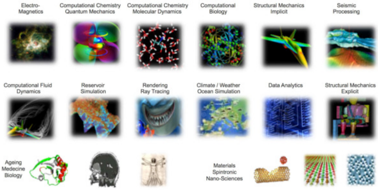
- Essential tools for **Science, Society and Industry**
  - ↪ All scientific disciplines are becoming computational today
    - ✓ requires very high computing power, handles **huge** volumes of data
- **Industry, SMEs** increasingly relying on HPC
  - ↪ to invent innovative solutions
  - ↪ ... while reducing cost & decreasing time to market

Andy Grant, Head of Big Data and HPC, Atos UK&I

To out-compete you must out-compute

Increasing competition, heightened customer expectations and shortening product development cycles are forcing the pace of acceleration across all industries.

Why HPC and BD ?



HPC: **H**igh **P**erformance **C**omputing
BD: **B**ig **D**ata

- Essential tools for **Science, Society and Industry**
  - ↪ All scientific disciplines are becoming computational today
    - ✓ requires very high computing power, handles **huge** volumes of data
- **Industry, SMEs** increasingly relying on HPC
  - ↪ to invent innovative solutions
  - ↪ ... while reducing cost & decreasing time to market
- HPC = **global race** (strategic priority) - EU takes up the challenge:
  - ↪ EuroHPC / IPCEI on HPC and Big Data (BD) Applications

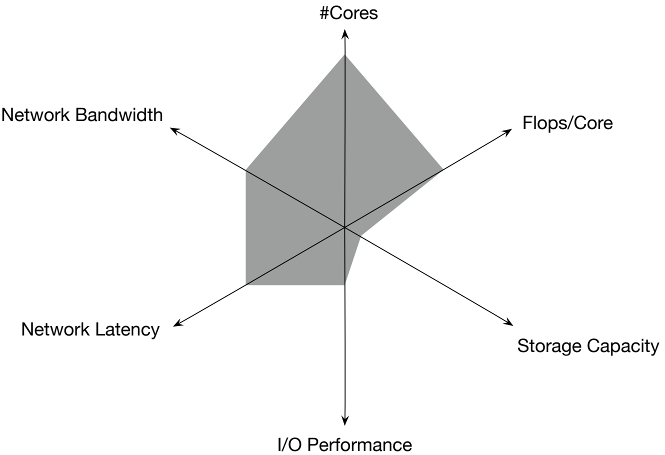
Andy Grant, Head of Big Data and HPC, Atos UK&I

To out-compete you must out-compute

Increasing competition, heightened customer expectations and shortening product development cycles are forcing the pace of acceleration across all industries.

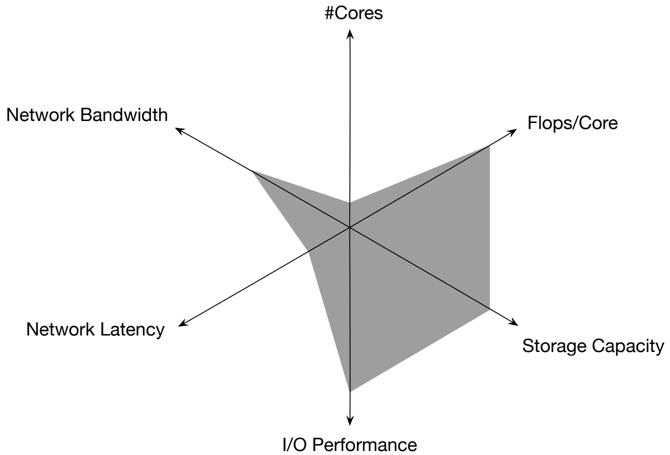
Different HPC Needs per Domains

Material Science & Engineering



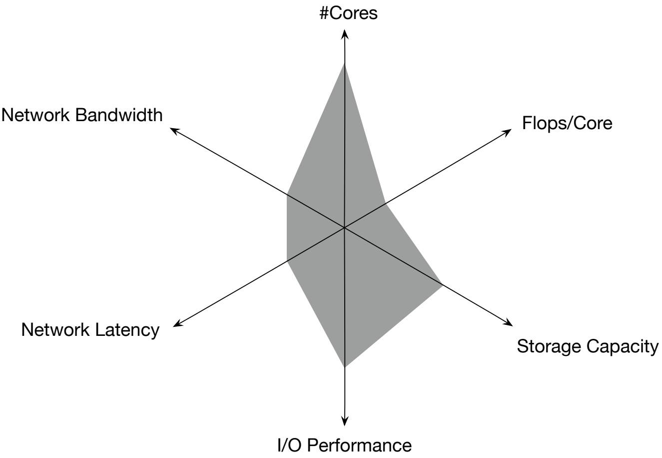
Different HPC Needs per Domains

Biomedical Industry / Life Sciences



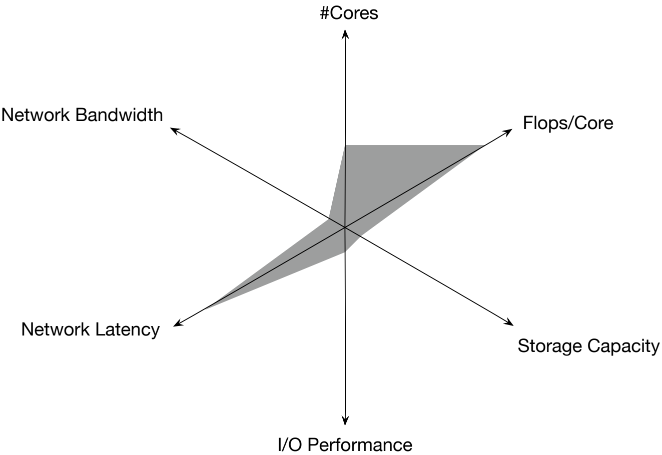
Different HPC Needs per Domains

Deep Learning / Cognitive Computing



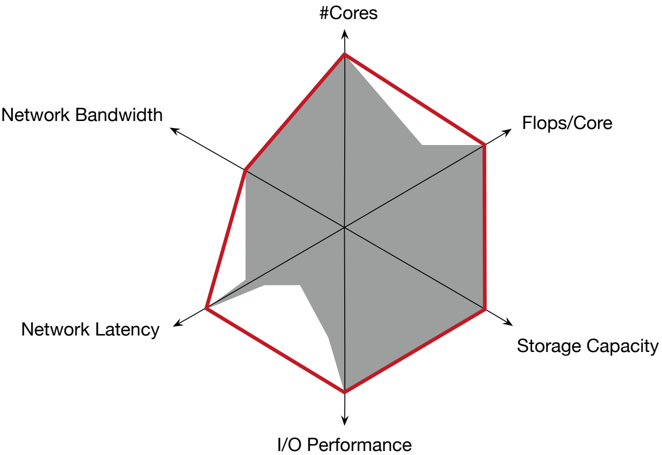
Different HPC Needs per Domains

IoT, FinTech



Different HPC Needs per Domains

ALL Research Computing Domains



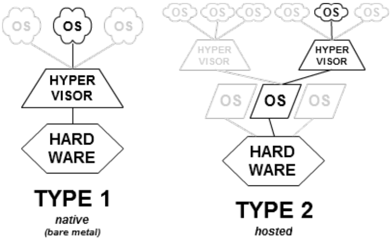
Summary

Computing for Researchers: Laptop

- **Regular PC / Local Laptop / Workstation**
  ↪ **Native OS** (Windows, Linux, Mac etc.)

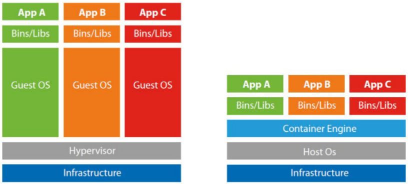
Computing for Researchers: Laptop

- **Regular PC / Local Laptop / Workstation**
  - ↪ **Native OS** (Windows, Linux, Mac etc.)
- Virtualized OS (**VM**) through an **hypervisor**
  - ↪ *Hypervisor*: core virtualization engine / environment
    - ✓ Ex: Xen, VMWare ESXi, KVM, VirtualBox
    - ✓ Non-negligible Performance loss: $\geq$ 20%



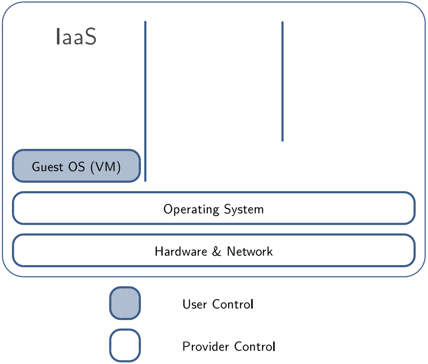
Computing for Researchers: Laptop

- **Regular PC / Local Laptop / Workstation**
  - ↪ **Native OS** (Windows, Linux, Mac etc.)
- Virtualized OS (**VM**) through an **hypervisor**
  - ↪ *Hypervisor*: core virtualization engine / environment
    - ✓ Ex: Xen, VMWare ESXi, KVM, VirtualBox
    - ✓ Non-negligible Performance loss: $\geq$ 20%
- **Container**-based Virtualization
  - ↪ similar to VMs ...
    - ✓ **yet** containers **share** the system kernel of the host with others
    - ✓ Ex: Docker, Singularity, Shifter



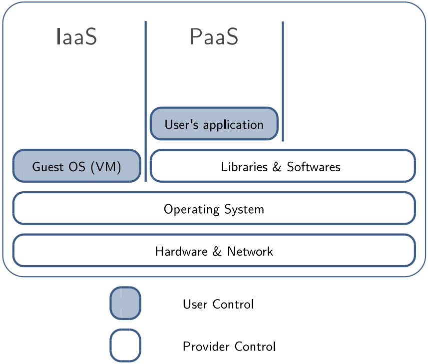
Computing for Researchers: Cloud

- **Cloud Computing**
  - ↪ access to shared *(generally virtualized)* resources
  - ↪ pay-per-use approach
  - ↪ **Infrastructure** as a Service (**I**aaS)



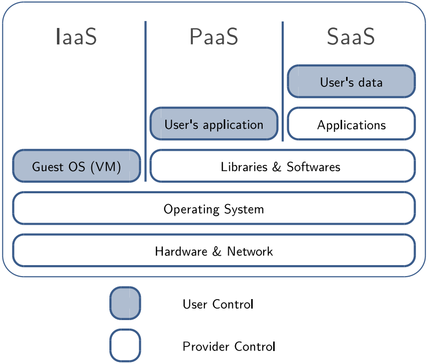
Computing for Researchers: Cloud

- **Cloud Computing**
  - ↪ access to shared *(generally virtualized)* resources
  - ↪ pay-per-use approach
  - ↪ **Platform** as a Service (**P**aaS)



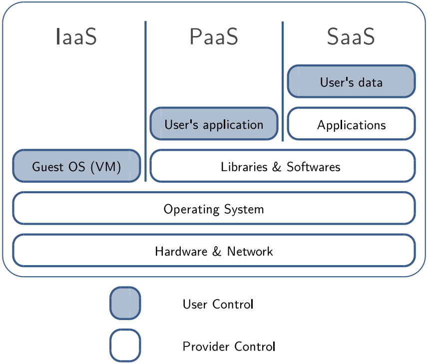
Computing for Researchers: Cloud

- **Cloud Computing**
  - ↪ access to shared *(generally virtualized)* resources
  - ↪ pay-per-use approach
  - ↪ **Software** as a Service (**S**aaS)



Computing for Researchers: Cloud

- **Cloud Computing**
  - ↪ access to shared *(generally virtualized)* resources
  - ↪ pay-per-use approach
  - ↪ **XXX** as a Service (<**X**>aaS)



Computing for Researchers: HPC

- High Performance Computing (HPC) platforms
  ↪ For **Speedup**, **Scalability** and **Faster Time to Solution**

Computing for Researchers: HPC

- High Performance Computing (HPC) platforms
  - ↪ For **Speedup**, **Scalability** and **Faster Time to Solution**

YET...

PC ≠ Cloud ≠ HPC

Computing for Researchers: HPC

- High Performance Computing (HPC) platforms
  ↪ For **Speedup**, **Scalability** and **Faster Time to Solution**

YET...

$$\textbf{PC} \neq \textbf{Cloud} \neq \textbf{HPC}$$

- HPC $\simeq$ Formula 1
  ↪ relies on ultra efficient hardware / interconnect (IB EDR...)
  ↪ ... when Cloud has to stay standard ([10] GbE etc...)
- **Does not mean the 3 approaches cannot work together**

Jobs, Tasks & Local Execution

$> ./myprog

Jobs, Tasks & Local Execution

```
$> ./myprog
```

Jobs, Tasks & Local Execution

```
$> ./myprog
$> ./myprog -n 10
```

Jobs, Tasks & Local Execution

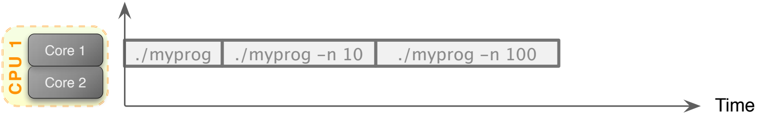
```
$> ./myprog
$> ./myprog -n 10
```

Jobs, Tasks & Local Execution

```
$> ./myprog
$> ./myprog -n 10
$> ./myprog -n 100
```

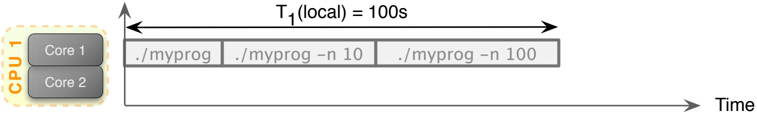
Jobs, Tasks & Local Execution

```
$> ./myprog
$> ./myprog -n 10
$> ./myprog -n 100
```

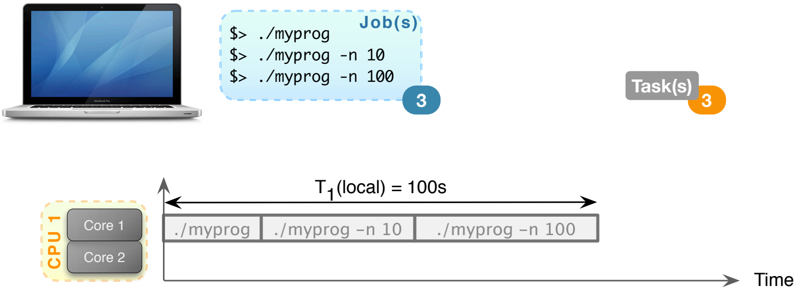


Jobs, Tasks & Local Execution

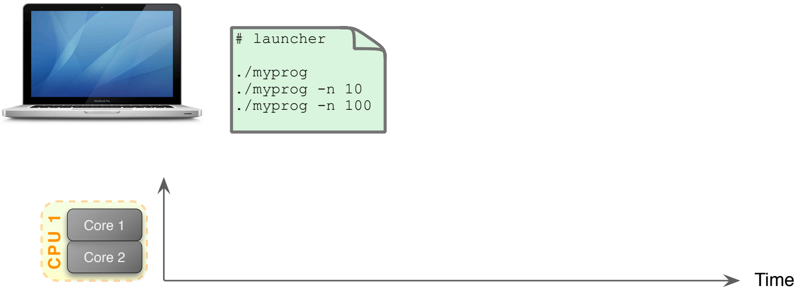
```
$> ./myprog
$> ./myprog -n 10
$> ./myprog -n 100
```



Jobs, Tasks & Local Execution



Jobs, Tasks & Local Execution



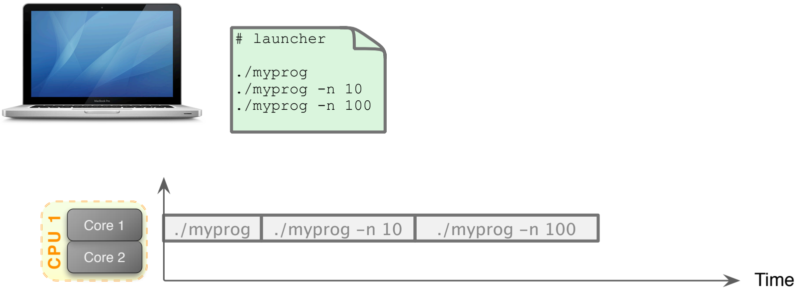
Jobs, Tasks & Local Execution

```
# launcher

./myprog
./myprog -n 10
./myprog -n 100
```

Jobs, Tasks & Local Execution

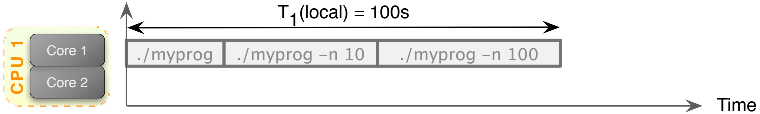
```
# launcher

./myprog
./myprog -n 10
./myprog -n 100
```

Jobs, Tasks & Local Execution

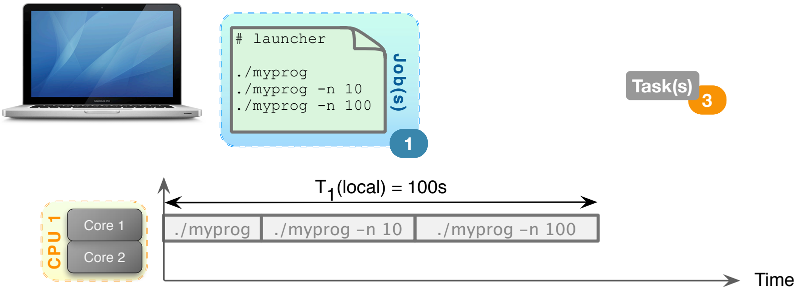


Jobs, Tasks & Local Execution

```
# launcher

./myprog
./myprog -n 10
./myprog -n 100
```



Jobs, Tasks & Local Execution



Jobs, Tasks & Local Execution

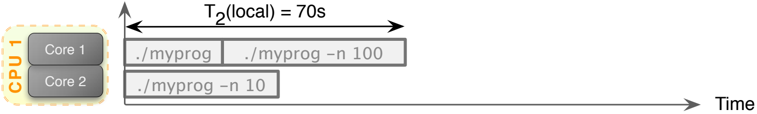
```
# launcher

./myprog
./myprog -n 10
./myprog -n 100
```

Jobs, Tasks & Local Execution

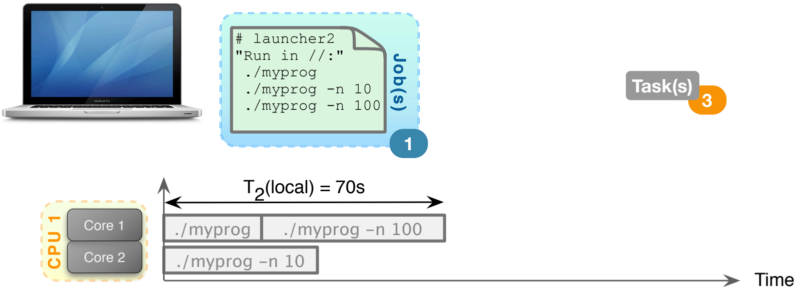
```
# launcher2
"Run in //:"
 ./myprog
 ./myprog -n 10
 ./myprog -n 100
```

Jobs, Tasks & Local Execution

```
# launcher2
"Run in //:"
 ./myprog
 ./myprog -n 10
 ./myprog -n 100
```

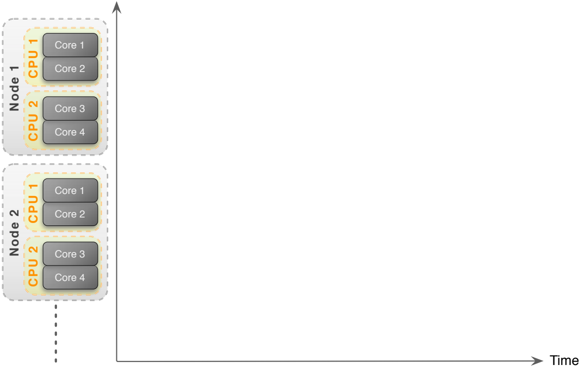


Jobs, Tasks & Local Execution

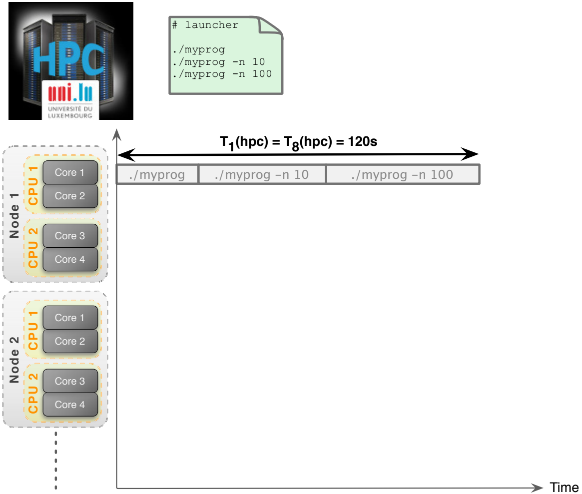


Jobs, Tasks & HPC Execution

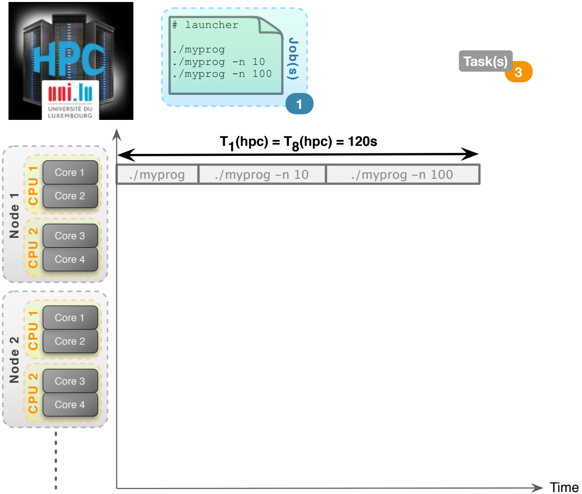
```
# launcher

./myprog
./myprog -n 10
./myprog -n 100
```



Jobs, Tasks & HPC Execution

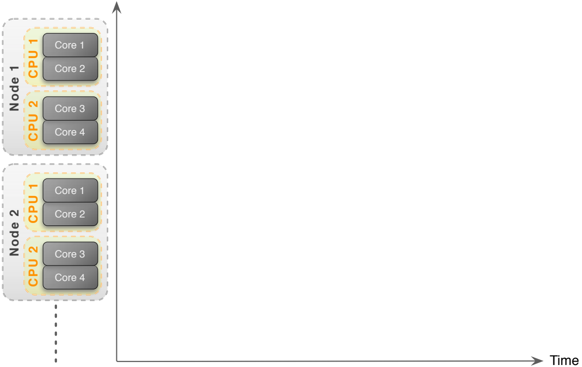


Jobs, Tasks & HPC Execution

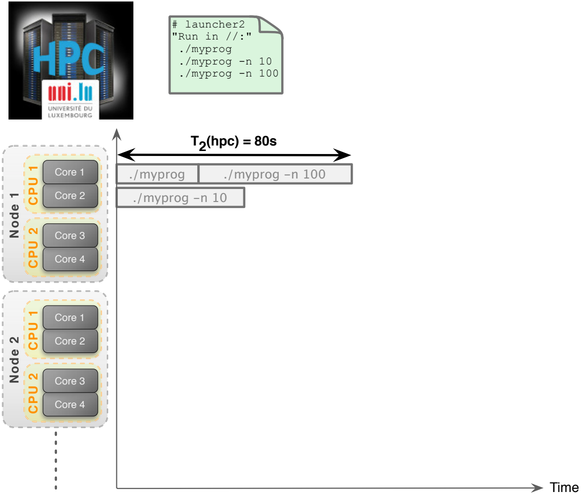


Jobs, Tasks & HPC Execution

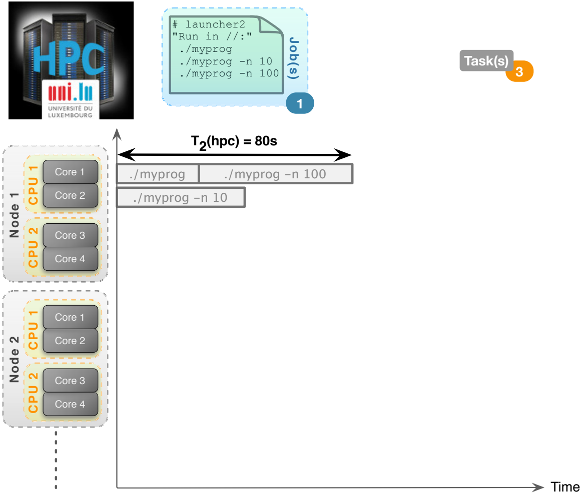
```
# launcher2
"Run in //:"
 ./myprog
 ./myprog -n 10
 ./myprog -n 100
```



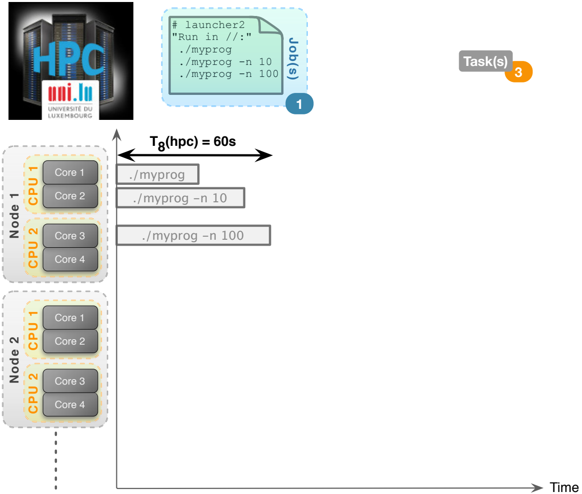
Jobs, Tasks & HPC Execution



Jobs, Tasks & HPC Execution



Jobs, Tasks & HPC Execution



Local vs. HPC Executions

Context	Local PC	HPC
Sequential	$T_1(\text{local}) = 100$	$T_1(\text{hpc}) = \mathbf{120s}$
Parallel/Distributed	$T_2(\text{local}) = 70\text{s}$	$T_2(\text{hpc}) = 80\text{s}$
		$T_8(\text{hpc}) = \mathbf{60s}$

Local vs. HPC Executions

Context	Local PC	HPC
Sequential	$T_1(\text{local}) = 100$	$T_1(\text{hpc}) = \mathbf{120s}$
Parallel/Distributed	$T_2(\text{local}) = 70s$	$T_2(\text{hpc}) = 80s$
		$T_8(\text{hpc}) = \mathbf{60s}$

- **Sequential** runs **WON'T BE FASTER** on HPC
  ↪ Reason: Processor Frequency (typically $\geq$ 3GHz vs $\geq$ 2GHz)

Local vs. HPC Executions

Context	Local PC	HPC
Sequential	$T_1(\text{local}) = 100$	$T_1(\text{hpc}) = \mathbf{120s}$
Parallel/Distributed	$T_2(\text{local}) = 70s$	$T_2(\text{hpc}) = 80s$
		$T_8(\text{hpc}) = \mathbf{60s}$

- **Sequential** runs **WON'T BE FASTER** on HPC
  - ↪ Reason: Processor Frequency (typically $\geq$ 3GHz vs $\geq$ 2GHz)

- **Parallel/Distributed** runs **DO NOT COME FOR FREE**
  - ↪ runs **will be sequential** even if you reserve $\geq$ 2 cores/nodes
  - ↪ you have to **explicitly** adapt your jobs to benefit from the multi-cores/nodes

Summary

HPC Components: [GP]CPU

CPU

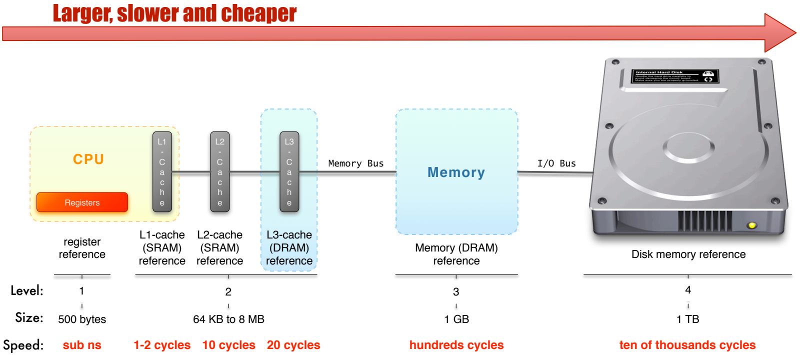
- Always multi-core
- Ex: Intel Core i7-7700K (Jan 2017) $R_{peak} \simeq$ 268.8 GFlops (DP)
  - ↪ 4 cores @ 4.2GHz (14nm, 91W, 1.75 billion transistors)
  - ↪ + integrated graphics (24 EUs) $R_{peak} \simeq$ +441.6 GFlops

GPU / GPGPU

- Always multi-core, optimized for vector processing
- Ex: Nvidia Tesla V100 (Jun 2017) $R_{peak} \simeq$ 7 TFlops (DP)
  - ↪ 5120 cores @ 1.3GHz (12nm, 250W, 21 billion transistors)
  - ↪ focus on Deep Learning workloads $R_{peak} \simeq$ 112 TFLOPS (HP)

$\simeq$ **100 Gflops** for 130$ (CPU), 214$? (GPU)

HPC Components: Local Memory

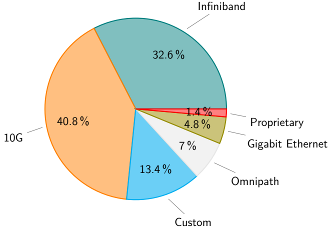


- SSD (SATA3) R/W: 550 MB/s; 100000 IOPS **450 €/TB**
- HDD (SATA3 @ 7,2 krpm) R/W: 227 MB/s; 85 IOPS **54 €/TB**

HPC Components: Interconnect

- **latency**: time to send a minimal (0 byte) message from A to B
- **bandwidth**: max amount of data communicated per unit of time

Technology	Effective Bandwidth		Latency
Gigabit Ethernet	1 Gb/s	125 MB/s	40μs to 300μs
10 Gigabit Ethernet	10 Gb/s	1.25 GB/s	4μs to 5μs
Infiniband QDR	40 Gb/s	5 GB/s	1.29μs to 2.6μs
Infiniband EDR	100 Gb/s	12.5 GB/s	0.61μs to 1.3μs
100 Gigabit Ethernet	100 Gb/s	1.25 GB/s	30μs
Intel Omnipath	100 Gb/s	12.5 GB/s	0.9μs

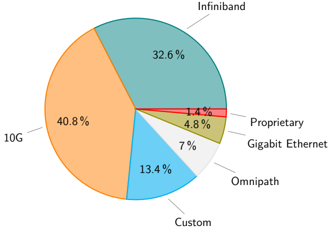


[Source : www.top500.org, Nov. 2017]

HPC Components: Interconnect

- **latency**: time to send a minimal (0 byte) message from A to B
- **bandwidth**: max amount of data communicated per unit of time

Technology	Effective Bandwidth		Latency
Gigabit Ethernet	1 Gb/s	125 MB/s	40μs to 300μs
10 Gigabit Ethernet	10 Gb/s	1.25 GB/s	4μs to 5μs
Infiniband QDR	40 Gb/s	5 GB/s	1.29μs to 2.6μs
Infiniband EDR	100 Gb/s	12.5 GB/s	0.61μs to 1.3μs
100 Gigabit Ethernet	100 Gb/s	1.25 GB/s	30μs
Intel Omnipath	100 Gb/s	12.5 GB/s	0.9μs



[Source : www.top500.org, Nov. 2017]

Network Topologies

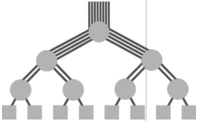
- **Direct** vs. **Indirect** interconnect
  - ↪ *direct*: each network node attaches to at least one compute node
  - ↪ *indirect*: compute nodes attached at the edge of the network only
    - ✓ many routers only connect to other routers.

Network Topologies

- **Direct** vs. **Indirect** interconnect
  - ↪ *direct*: each network node attaches to at least one compute node
  - ↪ *indirect*: compute nodes attached at the edge of the network only
    - ✓ many routers only connect to other routers.

Main HPC Topologies

- **CLOS Network / Fat-Trees** [Indirect]
  - ↪ can be fully non-blocking (1:1) or blocking (x:1)
  - ↪ typically enables **best performance**
    - ✓ Non blocking bandwidth, lowest network latency

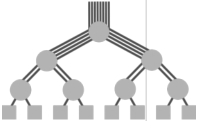


Network Topologies

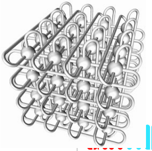
- **Direct** vs. **Indirect** interconnect
  - ↪ *direct*: each network node attaches to at least one compute node
  - ↪ *indirect*: compute nodes attached at the edge of the network only
    - ✓ many routers only connect to other routers.

Main HPC Topologies

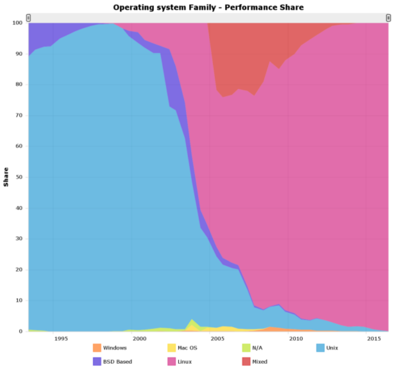
- **CLOS Network / Fat-Trees** [Indirect]
  - ↪ can be fully non-blocking (1:1) or blocking (x:1)
  - ↪ typically enables **best performance**
    - ✓ Non blocking bandwidth, lowest network latency



- **Mesh or 3D-torus** [Direct]
  - ↪ Blocking network, cost-effective for systems at scale
  - ↪ Great performance solutions for applications with locality
  - ↪ Simple expansion for future growth

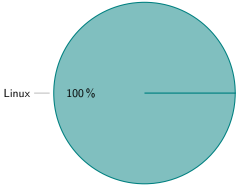


HPC Components: Operating System



- Exclusively Linux-based (**really** 100%)
- Reasons:
  - ↪ stability
  - ↪ prone to devels

[Source : www.top500.org, Nov 2017]



HPC Components: Software Stack

- **Remote connection to the platform** SSH
- **Identity Management / SSO**: LDAP, Kerberos, IPA...
- **Resource management**: job/batch scheduler
  ↪ SLURM, OAR, PBS, MOAB/Torque...
- (Automatic) **Node Deployment**:
  ↪ FAI, Kickstart, Puppet, Chef, Ansible, Kadeploy...
- (Automatic) **User Software Management**:
  ↪ Easybuild, Environment Modules, LMod
- **Platform Monitoring**:
  ↪ Nagios, Icinga, Ganglia, Foreman, Cacti, Alerta...

[Big]Data Management: Disk Encl.

- $\simeq$ **120 K€** - enclosure - 48-60 disks (4U)
  - ↪ incl. redundant (**i.e.** 2) RAID controllers (master/slave)

[Big]Data Management: FS Summary

- **File System (FS)**: Logical manner to *store, organize & access* data
  - ↪ (local) **Disk FS** : `FAT32`, `NTFS`, `HFS+`, `ext4`, `{x,z,btr}fs`...
  - ↪ **Networked FS**: `NFS`, `CIFS/SMB`, `AFP`
  - ↪ **Parallel/Distributed FS**: `SpectrumScale/GPFS`, `Lustre`
    - ✓ typical FS for HPC / HTC (High Throughput Computing)

[Big]Data Management: FS Summary

- **File System (FS)**: Logical manner to *store, organize & access* data
  - ↪ (local) **Disk FS** : `FAT32`, `NTFS`, `HFS+`, `ext4`, `{x,z,btr}fs`...
  - ↪ **Networked FS**: `NFS`, `CIFS/SMB`, `AFP`
  - ↪ **Parallel/Distributed FS**: `SpectrumScale/GPFS`, `Lustre`
    - ✓ typical FS for HPC / HTC (High Throughput Computing)

Main Characteristic of Parallel/Distributed File Systems

Capacity and Performance increase with #servers

[Big]Data Management: FS Summary

- **File System (FS)**: Logical manner to *store, organize & access* data
  - ↪ (local) **Disk FS** : `FAT32, NTFS, HFS+, ext4, {x,z,btr}fs...`
  - ↪ **Networked FS**: `NFS, CIFS/SMB, AFP`
  - ↪ **Parallel/Distributed FS**: `SpectrumScale/GPFS, Lustre`
    - ✓ typical FS for HPC / HTC (High Throughput Computing)

Main Characteristic of Parallel/Distributed File Systems

Capacity and Performance increase with #servers

Name	Type	Read* [GB/s]	Write* [GB/s]
ext4	Disk FS	0.426	0.212
nfs	Networked FS	0.381	0.090
gpfs (iris)	Parallel/Distributed FS	**11.25**	**9,46**
lustre (iris)	Parallel/Distributed FS	**12.88**	**10,07**
gpfs (gaia)	Parallel/Distributed FS	7.74	6.524
lustre (gaia)	Parallel/Distributed FS	4.5	2.956

* maximum **random** read/write, per IOZone or IOR measures, using concurrent nodes for networked FS.

HPC Components: Data Center

Definition (Data Center)

- Facility to house computer systems and associated components
  - ↪ Basic storage component: **rack** (height: 42 RU)

HPC Components: Data Center

Definition (Data Center)

- Facility to house computer systems and associated components
  ↪ Basic storage component: **rack** (height: 42 RU)

Challenges: Power (UPS, battery), Cooling, Fire protection, Security

- Power/Heat dissipation per rack:
  ↪ HPC **computing** racks: **30-120 kW**
  ↪ **Storage** racks: **15 kW**
  ↪ **Interconnect** racks: **5 kW**
- Various **Cooling** Technology
  ↪ Airflow
  ↪ Direct-Liquid Cooling, Immersion...

Power Usage Effectiveness

$$PUE = \frac{\textit{Total facility power}}{\textit{IT equipment power}}$$

Summary

Summary

High Performance Computing @ UL

- **Started in 2007**, under resp. of Prof P. Bouvry & Dr. S. Varrette
  - ↪ expert UL HPC team *(S. Varrette, V. Plugaru, S. Peter, H. Cartiaux, C. Parisot)*
  - ↪ **8,173,747€** cumulative investment in hardware

http://hpc.uni.lu

Key numbers

- 469 users
- 662 computing nodes
  - ↪ 10132 cores **346.652 TFlops**
  - ↪ 50 accelerators (+ **76.22 TFlops**)
- **9852.4 TB** storage
- 130 (+ 71) servers
- 5 sysadmins
- 2 sites: Kirchberg / Belval

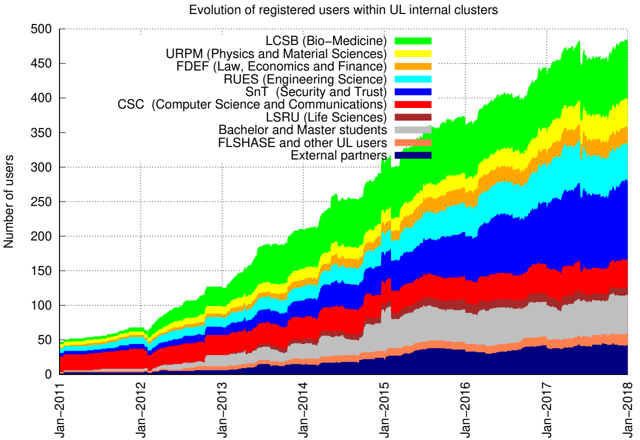
High Performance Computing @ UL

- **Enables & accelerates** scientific discovery and innovation
- **Largest facility** in Luxembourg (after GoodYear R&D Center)

Country	Institute	#Nodes	(CPU) #Cores	TFlops R_{peak}	TB (Shared) Storage
Luxembourg	**UL HPC** (Uni.lu)	**662**	**10132**	**346.652**	**9852.4**
	LIST	94	1952	83.21	494
France	LORIA (G5K), Nancy	257	2436	32.6	196
	ROMEO, Reims	174	3136	49.26	245
Belgium	NIC4, University of Liège	128	2048	32.00	20
	Université Catholique de Louvain	112	1344	13.28	120
	UGent / VSC, Gent	440	8768	275.30	1122
Germany	bwGrid, Heidelberg	140	1120	12.38	32
	bwForCluster, Ulm	444	7104	266.40	400
	bwHPC MLS&WISO, Mannheim	604	9728	371.60	420

UL HPC User Base

- **469 Active** HPC Users



UL HPC Beneficiaries

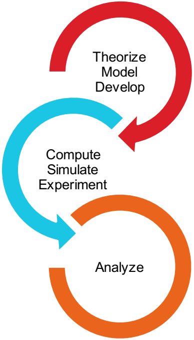
23 computational domains accelerated on UL HPC

- for the UL Faculties, Research Units and Interdisciplinary Centres
  - ↪ incl. LCSB, SnT... and now C2DH thematics
  - ↪ UL **strategic research priorities**
    - ✓ computational sciences, finance (fintech)
    - ✓ systems biomedicine, security, reliability and trust

- UL HPC feat. special systems targeting specific workloads:
  - ↪ **Machine Learning & AI**: GPU accelerators
    - ✓ 10 Tesla K40 + 16 Tesla K80 + 24 Tesla M20*: **76 GPU Tflops**
  - ↪ **BigData analytics & data driven science**: large memory systems
    - ✓ Large SMP systems with 1, 2, 3 & 4 TB RAM
  - ↪ **Scale-out workloads**: energy efficient systems
    - ✓ 90 HP Moonshot servers + 96 viridis ARM-based systems

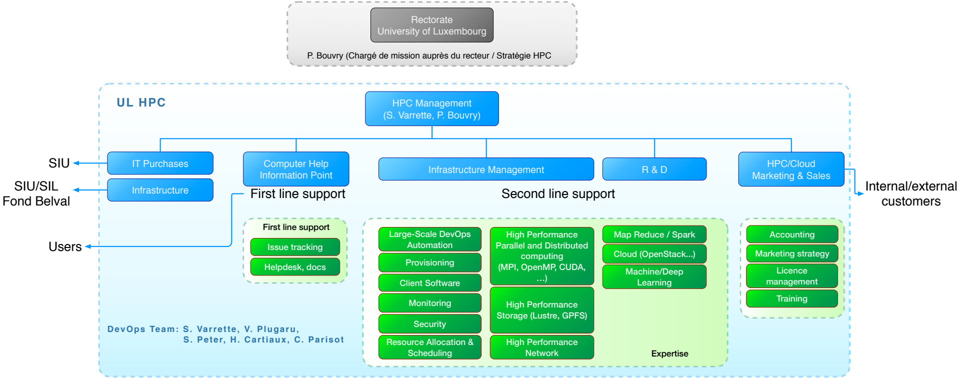
Accelerating UL Research

https://hpc.uni.lu/users/software/



- >163 **software packages** available for researchers
  - ↪ **General purpose**, statistics, optimization:
    - ✓ Matlab, Mathematica, R, Stata, CPLEX, Gurobi Optimizer. . .
  - ↪ **Bioinformatics**
    - ✓ BioPython, STAR, TopHat, Bowtie, mpiHMMER. . .
  - ↪ **Computer aided engineering**:
    - ✓ ANSYS, ABAQUS, OpenFOAM. . .
  - ↪ **Molecular dynamics**:
    - ✓ NAMD, ABINIT, Q.ESPRESSO, GROMACS. . .
  - ↪ **Visualisation**: ParaView, VisIt, VMD, XCS portal
  - ↪ Compilers, libraries, performance modeling tools
  - ↪ [Parallel] debugging tools aiding development

ULHPC Governance



UL HPC Team

Prof. Pascal Bouvry
Director of DS-CSCE, Leader of PCO Group
Senior advisor for the president as regards the HPC strategy

Sébastien Varrette, PhD
CDI, Research Scientist (CSC, FSTC)

Valentin Plugaru, MSc.
CDI, Research Associate (CSC, FSTC)

Sarah Peter, MSc.
CDD, Research Associate (LCSB)

Hyacinthe Cartiaux
CDI, Support (SIU)

Clément Parisot
CDI, Support (CSC, FSTC)

Sites / Data centers

Kirchberg

CS.43, AS. 28

Belval

Biotech I, CDC/MSA

2 sites, $\geq$ 4 server rooms

Sites / Data centers

Kirchberg

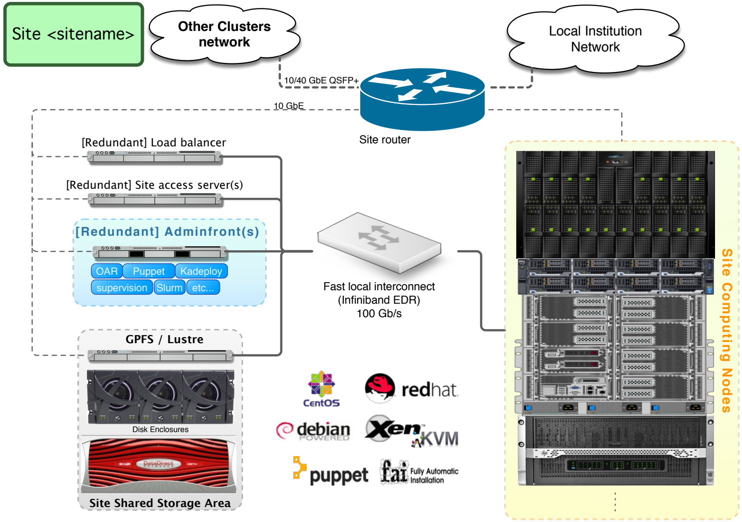
CS.43, AS. 28

Belval

Biotech I, CDC/MSA

2 sites, $\geq$ 4 server rooms

UL HPC: General cluster organization



UL HPC Computing capacity

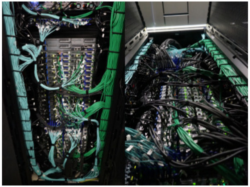
5 clusters

346.652 TFlops

662 nodes

10132 cores

34512GPU cores



UL HPC Computing Clusters

Cluster	Location	#N	#C	Rpeak	GPU Rpeak
iris	CDC S-01	168	4704	256	0
gaia	BT1	273	3440	69.296	76
chaos	Kirchberg	81	1120	14.495	0
g5k	Kirchberg	38	368	4.48	0
nyx (experimental)	BT1	102	500	2.381	0
	TOTAL:	**662**	**10132**	**346.652**	**+ 76 TFlops**

UL HPC - Detailed Computing Nodes

	Date	Vendor	Proc. Description		#N	#C	R_{peak}
iris	2017	Dell	Intel Xeon E5-2680 v4@2.4GHz	2 × 14C,128GB	108	3024	116,12 TFlops
	2018	Dell	Intel Xeon Gold 6132 @ 2.6 GHz 60	2 × 14C,128GB	60	1680	139,78 TFlops
				`iris` **TOTAL:**	**168**	**4704**	**256 TFlops**
gaia	2011	Bull	Intel Xeon L5640@2.26GHz	2 × 6C,48GB	72	864	7.811 TFlops
	2012	Dell	Intel Xeon E5-4640@2.4GHz	4 × 8C, 1TB	1	32	0.614 TFlops
	2012	Bull	Intel Xeon E7-4850@2GHz	16 × 10C,1TB	1	160	1.280 TFLops
	2013	Dell	Intel Xeon E5-2660@2.2GHz	2 × 8C,64GB	5	80	1.408 TFlops
	2013	Bull	Intel Xeon X5670@2.93GHz	2 × 6C,48GB	40	480	5.626 TFlops
	2013	Bull	Intel Xeon X5675@3.07GHz	2 × 6C,48GB	32	384	4.746 TFlops
	2014	Delta	Intel Xeon E7-8880@2.5 GHz	8 × 15C,1TB	1	120	2.4 TFlops
	2014	SGi	Intel Xeon E5-4650@2.4 GHz	16 × 10C,4TB	1	160	3.072 TFlops
	2015	Dell	Intel Xeon E5-2680@2.5 GHz	2 × 12C,128GB	28	672	26.88 TFlops
	2015	HP	Intel E3-1284Lv3, 1.8GHz	1 × 4C,32GB	90	360	10.368 TFlops
	2016	Dell	Intel Xeon E7-8867@2.5 GHz	4 × 16C,2TB	2	128	5.12 TFlops
				`gaia` **TOTAL:**	**273**	**3440**	**69.296 TFlops**
chaos	2010	HP	Intel Xeon L5640@2.26GHz	2 × 6C,24GB	32	384	3.472 TFlops
	2011	Dell	Intel Xeon L5640@2.26GHz	2 × 6C,24GB	16	192	1.736 TFlops
	2012	Dell	Intel Xeon X7560@2,26GHz	4 × 6C, 1TB	1	32	0.289 TFlops
	2012	Dell	Intel Xeon E5-2660@2.2GHz	2 × 8C,32GB	16	256	4.506 TFlops
	2012	HP	Intel Xeon E5-2660@2.2GHz	2 × 8C,32GB	16	256	4.506 TFlops
				`chaos` **TOTAL:**	**81**	**1120**	**14.495 TFlops**
g5k	2008	Dell	Intel Xeon L5335@2GHz	2 × 4C,16GB	22	176	1.408 TFlops
	2012	Dell	Intel Xeon E5-2630L@2GHz	2 × 6C,24GB	16	192	3.072 TFlops
				`granduc/petitprince` **TOTAL:**	**38**	**368**	**4.48 TFlops**

Testing cluster:

	Date	Vendor	Proc. Description		#N	#C	R_{peak}
nyx, viridis, pyro...	2012	Dell	Intel Xeon E5-2420@1.9GHz	1 × 6C,32GB	2	12	0.091 TFlops
	2013	Viridis	ARM A9 Cortex@1.1GHz	1 × 4C,4GB	96	384	0.422 TFlops
	2015	Dell	Intel Xeon E5-2630Lv2@2.4GHz	2 × 6C,32GB	2	24	0.460 TFlops
	2015	Dell	Intel Xeon E5-2660v2@2.2GHz	2 × 10C,32GB	4	80	1.408 TFlops
				`nyx/viridis` **TOTAL:**	**102**	**500**	**2.381 TFlops**

UL HPC Storage capacity



4 distributed/parallel FS
2425 disks
9852.4 TB

(incl. 1020TB for Backup)


UL HPC Shared Storage Capacities

Cluster	GPFS	Lustre	Other (NFS...)	Backup	TOTAL
iris	2284	1280	6	600	**4170** TB
gaia	960	660	0	240	**1860** TB
chaos	0	0	180	180	**360** TB
g5k	0	0	32.4	0	**32.4** TB
nyx (experimental)	0	0	242	0	**242** TB
TOTAL:	**3244**	**1940**	**3648.4**	**1020**	**9852.4 TB**

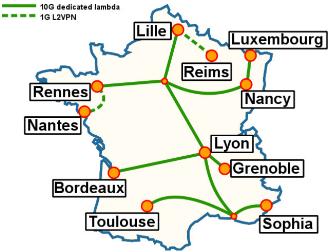
UL HPC Software Stack

- **Operating System**: **Linux** CentOS 7 (`iris`), Debian 8 (others)
- **Remote connection to the platform**: SSH
- **User SSO**: IPA, OpenLDAP
- **Resource management**: job/batch scheduler: **Slurm**(`iris`), OAR
- **(Automatic) Computing Node Deployment**:
  - ↪ FAI (Fully Automatic Installation)(`gaia`, `chaos` clusters)
  - ↪ Bright Cluster Manager (`iris`)
  - ↪ Puppet
  - ↪ Kadeploy
- **Platform Monitoring**:
  - ↪ OAR Monika/Drawgantt, Ganglia, Allinea Perf Report, SLURM
  - ↪ Icinga, NetXMS, PuppetBoard etc.
- **Commercial Softwares**:
  - ↪ ANSYS, ABAQUS, MATLAB, Intel Cluster Studio XE, Allinea DDT, Stata etc.

The case of Grid'5000

http://www.grid5000.fr

- Large scale nation wide infrastructure
  - ↪ for large scale parallel and distributed computing research.

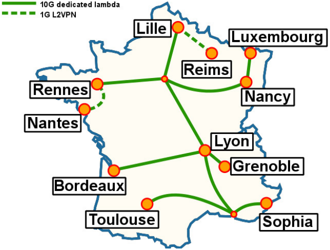


- 10 sites in France
  - ↪ **Abroad**: Luxembourg, Porto Allegre
  - ↪ Total: **7782** cores over **26** clusters
- 1-10GbE / Myrinet / Infiniband
  - ↪ **10Gb/s dedicated** between all sites
- Unique software stack
  - ↪ **kadeploy, kavlan, storage5k**

The case of Grid'5000

http://www.grid5000.fr

- Large scale nation wide infrastructure
  - ↪ for large scale parallel and distributed computing research.

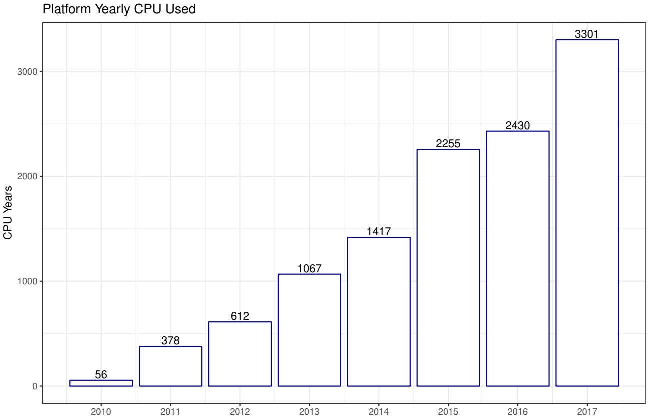


- 10 sites in France
  - ↪ **Abroad**: Luxembourg, Porto Allegre
  - ↪ Total: **7782** cores over **26** clusters
- 1-10GbE / Myrinet / Infiniband
  - ↪ **10Gb/s dedicated** between all sites
- Unique software stack
  - ↪ **kadeploy, kavlan, storage5k**

- **Out of scope for this talk**
  - ↪ General information: https://hpc.uni.lu/g5k
  - ↪ Grid'5000 website and documentation: https://www.grid5000.fr

CPU-year usage since 2010

- **CPU-hour**: *work* done by a CPU in **one hour** of wall clock time

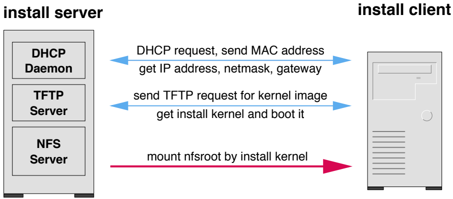


Summary

Computing nodes Management

Node deployment by FAI/Bright Manager

- Boot via network card (PXE)
  - ↪ ensure a running diskless Linux OS

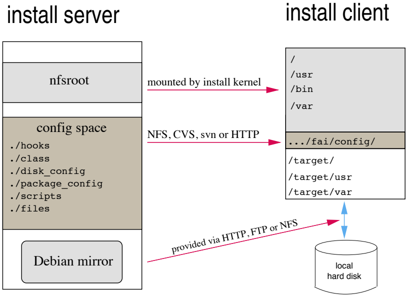


Computing nodes Management

Node deployment by FAI/Bright Manager

- Boot via network card (PXE)
  - ↪ ensure a running diskless Linux OS

- Get configuration data (NFS/other)



Computing nodes Management

Node deployment by FAI/Bright Manager

Fully Automatic Installation
Bright Computing

- Boot via network card (PXE)
  - ↪ ensure a running diskless Linux OS
- Get configuration data (NFS/other)
- Run the installation
  - ↪ partition local hard disks and create filesystems
  - ↪ install software using apt-get command
  - ↪ configure OS and additional software
  - ↪ save log files to install server, then reboot new system

Computing nodes Management

Node deployment by FAI/Bright Manager

- Boot via network card (PXE)
  - ↪ ensure a running diskless Linux OS
- Get configuration data (NFS/other)
- Run the installation
  - ↪ partition local hard disks and create filesystems
  - ↪ install software using apt-get command
  - ↪ configure OS and additional software
  - ↪ save log files to install server, then reboot new system

Average reinstallation time: $\simeq$ 500s

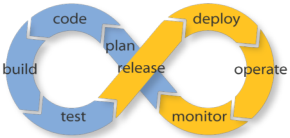
IT Serv[er|ice] Management: Puppet

Server/Service configuration by Puppet

http://puppetlabs.com

- **IT Automation** for configuration management
  - ↪ idempotent, agent/master OR stand-alone
  - ↪ cross-platform through Puppet Resource Abstraction Layer (RAL)
  - ↪ git-based workflow with r10k (**role & profiles** workflow)
  - ↪ PKI-based security (X.509)
- DevOps tool of choice for configuration management
  - ↪ Reusable modules
  - ↪ per-environment hierarchy lookup with hiera

https://forge.puppet.com/



Endless Possibilities: DevOps can create an infinite loop of release and feedback for all your code and deployment targets.

IT Serv[er|ice] Management: Puppet

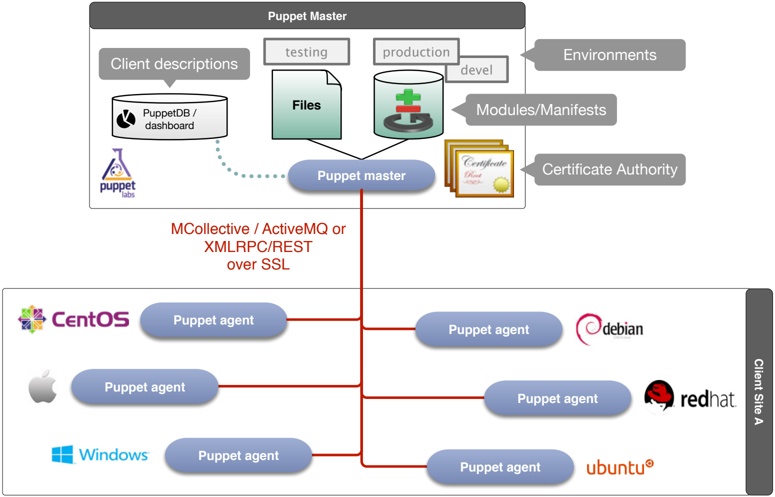
Server/Service configuration by Puppet

http://puppetlabs.com

- **IT Automation** for configuration management
  - ↪ idempotent, agent/master OR stand-alone
  - ↪ cross-platform through Puppet Resource Abstraction Layer (RAL)
  - ↪ git-based workflow with r10k (**role & profiles** workflow)
  - ↪ PKI-based security (X.509)
- DevOps tool of choice for configuration management
  - ↪ Reusable modules — https://forge.puppet.com/
  - ↪ per-environment hierarchy lookup with hiera

Average server installation/configuration time: $\simeq$ 3-6 min

General Puppet Infrastructure



Software/Modules Management

https://hpc.uni.lu/users/software/

- Based on Environment Modules / LMod
  - ↪ convenient way to dynamically change the users environment `$PATH`
  - ↪ permits to easily load software through `module` command
- Currently on UL HPC:
  - ↪ $>$ **163 software packages**, in *multiple* versions, within **18 categ.**
  - ↪ reworked software set for `iris` cluster and now deployed everywhere
    - ✓ RESIF v2.0, allowing [real] semantic versioning of released builds
  - ↪ hierarchical organization **Ex**: `toolchain/{foss,intel}`

```
$> module avail                    # List available modules
```

```
$> module load <category>/<software>[/<version>]
```

Software/Modules Management

- Key `module` variable: `$MODULEPATH` / where to look for modules
  ↪ altered with `module use <path>`. **Ex**:

```
export EASYBUILD_PREFIX=$HOME/.local/easybuild
export LOCAL_MODULES=$EASYBUILD_PREFIX/modules/all
module use $LOCAL_MODULES
```

Software/Modules Management

- Key module variable: $MODULEPATH / where to look for modules
  ↪ altered with module use <path>. **Ex**:

```
export EASYBUILD_PREFIX=$HOME/.local/easybuild
export LOCAL_MODULES=$EASYBUILD_PREFIX/modules/all
module use $LOCAL_MODULES
```

Main modules commands:

Command	Description
module avail	Lists all the modules which are available to be loaded
module spider <pattern>	Search for among available modules **(Lmod only)**
module load <mod1> [mod2...]	Load a module
module unload <module>	Unload a module
module list	List loaded modules
module purge	Unload all modules (purge)
module display <module>	Display what a module does
module use <path>	Prepend the directory to the MODULEPATH environment variable
module unuse <path>	Remove the directory from the MODULEPATH environment variable

Software/Modules Management

http://hpcugent.github.io/easybuild/

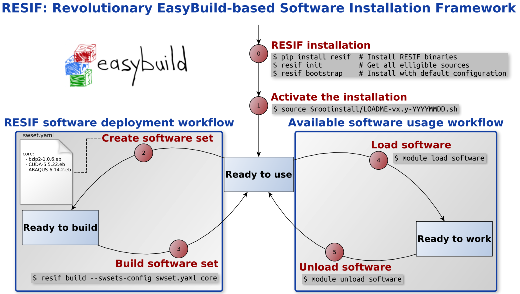
- Easybuild: open-source framework to (automatically) build scientific software
- **Why?**: *"Could you please install this software on the cluster?"*
  - ↪ Scientific software is often **difficult** to build
    - ✓ non-standard build tools / incomplete build procedures
    - ✓ hardcoded parameters and/or poor/outdated documentation
  - ↪ EasyBuild helps to facilitate this task
    - ✓ **consistent** software **build and installation** framework
    - ✓ includes testing step that helps validate builds
    - ✓ **automatically generates LMod modulefiles**

```
$> module use $LOCAL_MODULES
$> module load tools/EasyBuild
$> eb -S HPL    # Search for recipes for HPL software
$> eb HPL-2.2-intel-2017a.eb # Install HPL 2.2 w. Intel toolchain
```

Software/Modules Management

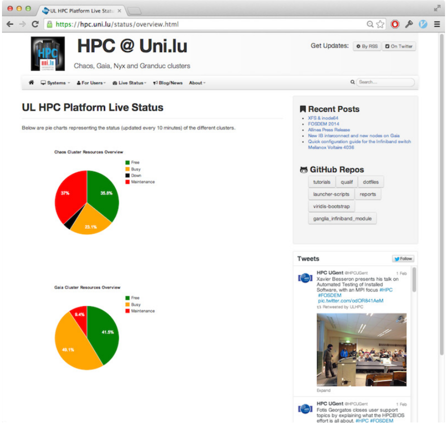
- **RESIF**: Revolutionary EasyBuild-based Software Installation Framework
  - ↪ Automatic Management of **software sets**
  - ↪ Fully automates software builds and supports all available toolchains
  - ↪ Clean (hierarchical) modules layout to facilitate its usage
  - ↪ (**incoming**) per-release module, per-ISA builds,

```
MODULEPATH=/opt/apps/resif/data/{devel,production,stable,testing}/default/modules/all/
```



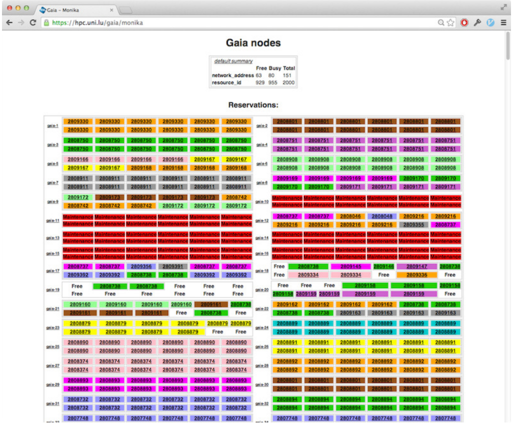
Platform Monitoring

- **General Live Status** http://hpc.uni.lu/status/overview.html/



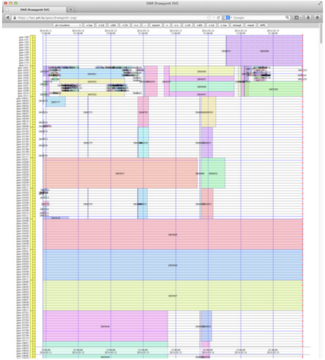
Platform Monitoring

- **Monika** http://hpc.uni.lu/{gaia,chaos,g5k}/monika



Platform Monitoring

- **Drawgantt** http://hpc.uni.lu/{gaia,chaos,g5k}/drawgantt

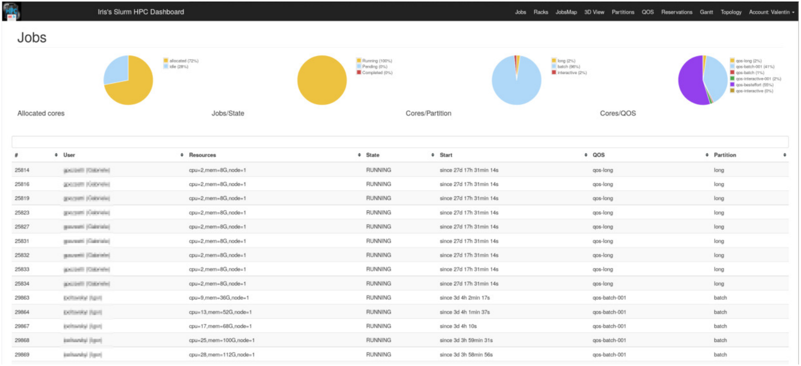


Platform Monitoring

- **Ganglia** http://hpc.uni.lu/{gaia,chaos,g5k,iris}/ganglia

Platform Monitoring

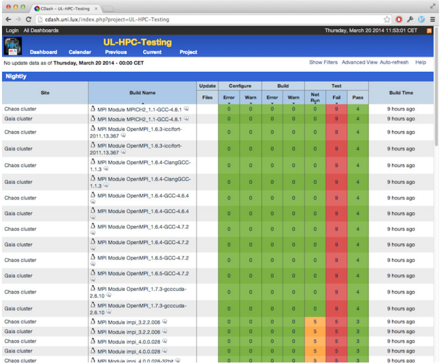
- **SLURM-Web** http://hpc.uni.lu/iris/slurm/



#	User	Resources	State	Start	QOS	Partition
25814	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25816	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25819	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25823	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25827	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25831	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25832	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25833	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
25834	[illegible]	cpu=2,mem=8G,node=1	RUNNING	since 27d 17h 31min 14s	qos-long	long
29863	[illegible]	cpu=9,mem=36G,node=1	RUNNING	since 3d 4h 2min 17s	qos-batch-001	batch
29864	[illegible]	cpu=13,mem=52G,node=1	RUNNING	since 3d 4h 1min 37s	qos-batch-001	batch
29867	[illegible]	cpu=17,mem=68G,node=1	RUNNING	since 3d 4h 10s	qos-batch-001	batch
29868	[illegible]	cpu=25,mem=100G,node=1	RUNNING	since 3d 3h 59min 31s	qos-batch-001	batch
29869	[illegible]	cpu=28,mem=112G,node=1	RUNNING	since 3d 3h 58min 56s	qos-batch-001	batch

Platform Monitoring

- **CDash** http://cdash.uni.lu/



CDash - UL-HPC-Testing
cdash.uni.lux/index.php?project=UL-HPC-Testing
Login All Dashboards Thursday, March 20 2014 11:53:01 CET
UL-HPC-Testing
Dashboard Calendar Previous Current Project
No update data as of **Thursday, March 20 2014 - 00:00 CET**
Show Filters Advanced View Auto-refresh Help

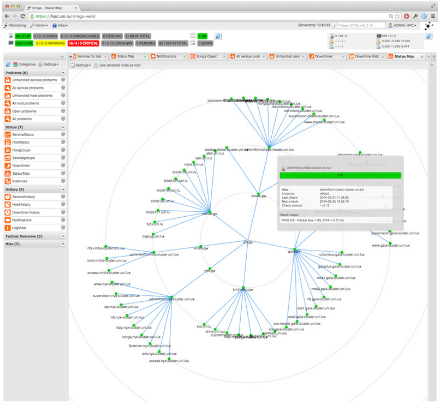
Nightly

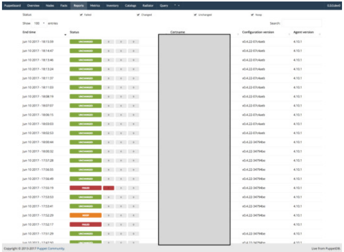
Site	Build Name	Update	Configure		Build		Test			Build Time
		Files	Error	Warn	Error	Warn	Not Run	Fail	Pass	
Chaos cluster	MPI Module MPICH2_1.1-GCC-4.8.1		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module MPICH2_1.1-GCC-4.8.1		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module OpenMPI_1.6.3-iccifort-2011.13.367		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module OpenMPI_1.6.3-iccifort-2011.13.367		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module OpenMPI_1.6.4-ClangGCC-1.1.3		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module OpenMPI_1.6.4-ClangGCC-1.1.3		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module OpenMPI_1.6.4-GCC-4.6.4		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module OpenMPI_1.6.4-GCC-4.6.4		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module OpenMPI_1.6.4-GCC-4.7.2		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module OpenMPI_1.6.4-GCC-4.7.2		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module OpenMPI_1.6.5-GCC-4.7.2		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module OpenMPI_1.6.5-GCC-4.7.2		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module OpenMPI_1.7.3-gcccuda-2.6.10		0	0	0	0	0	9	4	9 hours ago
Gaia cluster	MPI Module OpenMPI_1.7.3-gcccuda-2.6.10		0	0	0	0	0	9	4	9 hours ago
Chaos cluster	MPI Module impi_3.2.2.006		0	0	0	0	5	5	3	9 hours ago
Gaia cluster	MPI Module impi_3.2.2.006		0	0	0	0	5	5	3	9 hours ago
Chaos cluster	MPI Module impi_4.0.0.028		0	0	0	0	5	5	3	9 hours ago
Gaia cluster	MPI Module impi_4.0.0.028		0	0	0	0	5	5	3	9 hours ago
Chaos cluster	MPI Module impi_4.0.0.028-32bit		0	0	0	0	5	5	3	9 hours ago

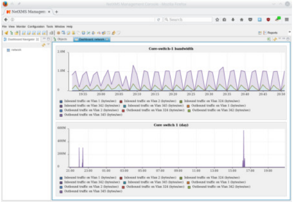
Platform Monitoring

- **Internal Monitoring**

Icinga / Puppet / NetXMS (networking)



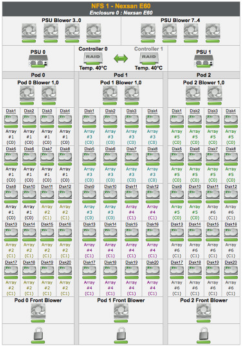


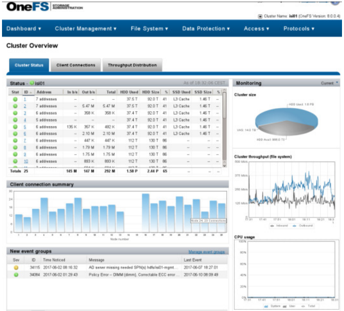


Platform Monitoring

- **Internal Monitoring**

[Disk] Enclosure status

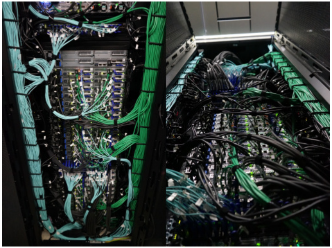




Summary

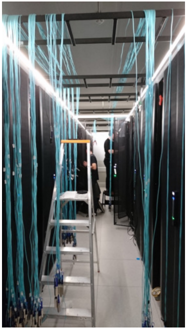
Past Year Achievements / Technical

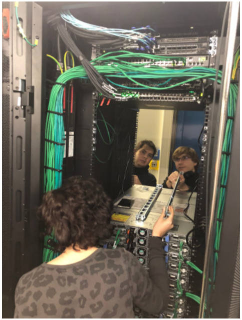
- **2017/2018: Installation of the new `iris` cluster**
  - ↪ RFP 160020: **168 nodes**, 4707 cores, **256 TFlops**
    - ✓ Dell C6320, 128 GB RAM
    - ✓ 108 x 2 Intel Xeon E5-2680v4@2.4 GHz [2x14c] (**2017**)
    - ✓ 60 x 2 Intel Xeon Gold 6132@2.6 GHz [2x14x] (**Q1 2018**)

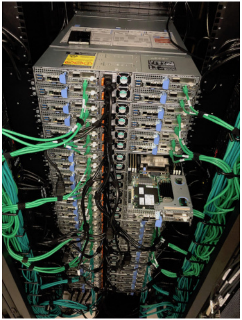


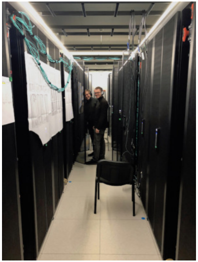
Past Year Achievements / Technical

- **Feb. 2018: Moving `iris` cluster**
  ↪ CDC S-01 → S-02







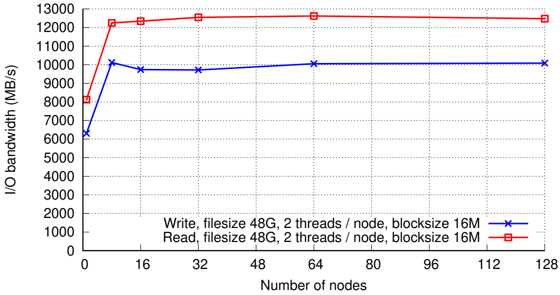


Past Year Achievements / Technical

- **April 2018:** `iris` **storage**
  - ↪ RFP 160019: **SpectrumsScale GPFS** (DDN GridScaler) **2284 TB**
    - ✓ Initial deployment June 2017 `/mnt/irisgpfs`
    - ✓ Extension done in April 2018
  - ↪ RFP 170035: **Lustre** (DDN Exascaler) **1280 TB**
    - ✓ Initial deployment April 2018 `/mnt/lscratch`

Past Year Achievements / Technical

- **iris Storage Performances**: Lustre
  - ↪ Self Encrypting Disks (SED)-based storage



Past Year Achievements / Technical

- Beyond `iris` setup, we **introduced several new elements**
  - ↪ Migration to **SLURM**
  - ↪ Consolidation of the High Availability (HA) setup
  - ↪ **Improved system automation** (Puppet 4.x / hiera)

Past Year Achievements / Technical

- Beyond `iris` setup, we **introduced several new elements**
  - ↪ Migration to **SLURM**
  - ↪ Consolidation of the High Availability (HA) setup
  - ↪ **Improved system automation** (Puppet 4.x / hiera)
- **Continuous OS / software modules / security Upgrade**
  - ↪ Migration to **Debian 8** on `gaia` and `chaos`
  - ↪ **RESIF v2**, updated software sets
    - ✓ 2018a toolchain & co. released **after** the school
  - ↪ **Meltdown/Spectre** processor vulnerability **mitigation**

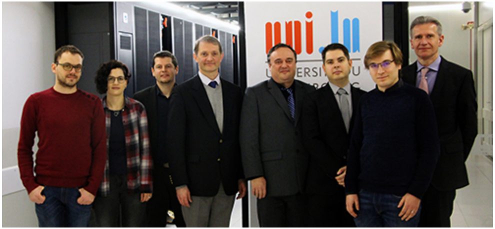
Past Year Achievements / Technical

- Beyond `iris` setup, we **introduced several new elements**
  - ↪ Migration to **SLURM**
  - ↪ Consolidation of the High Availability (HA) setup
  - ↪ **Improved system automation** (Puppet 4.x / hiera)
- **Continuous OS / software modules / security Upgrade**
  - ↪ Migration to **Debian 8** on `gaia` and `chaos`
  - ↪ **RESIF v2**, updated software sets
    - ✓ 2018a toolchain & co. released **after** the school
  - ↪ **Meltdown/Spectre** processor vulnerability **mitigation**
- **Storage**:
  - ↪ **Lustre `iris`**: Performance issue, pending resolutions

- **GPFS saturated on gaia cluster (99% usage)!!**
  - ↪ project quota frozen to the current usage (hard limit +10G)
  - ↪ **Action list: `$WORK` move to Lustre / project to Isilon**

Past Year Achievements

- (Oct. 2017) **Luxembourg 25th country to join PRACE**
  - ↪ Official representatives for Luxembourg from UL
    - ✓ Delegate: Prof. Pascal Bouvry
    - ✓ Advisor: Dr. Sebastien Varrette
  - ↪ Press release (Visit of S. Bogaerts, managing director PRACE)
    - ✓ "Luxembourg joins the prestigious European network PRACE"



Summary

Summary

General Guidelines

- The UL HPC is a ***shared*** resource
  - ↪ hundreds of users may be logged on at one time
  - ↪ hundreds of jobs may be running on all compute nodes,
- All users must practice ***good citizenship***
  - ↪ limit activities that may impact the system for other users.
  - ↪ **Do not abuse the shared filesystems**
    - ✓ Avoid too many simultaneous file transfers
    - ✓ regularly clean your directories from useless files
  - ↪ **Do not run programs on the login nodes**
  - ↪ Plan large scale experiments during night-time or week-ends
    - ✓ **no more than 120 cores** during working day and working hours

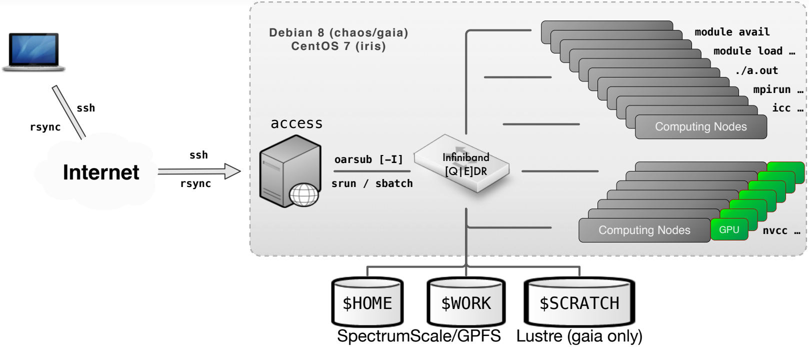
General Guidelines

- The UL HPC is a ***shared*** resource
  - ↪ hundreds of users may be logged on at one time
  - ↪ hundreds of jobs may be running on all compute nodes,
- All users must practice ***good citizenship***
  - ↪ limit activities that may impact the system for other users.
  - ↪ **Do not abuse the shared filesystems**
    - ✓ Avoid too many simultaneous file transfers
    - ✓ regularly clean your directories from useless files
  - ↪ **Do not run programs on the login nodes**
  - ↪ Plan large scale experiments during night-time or week-ends
    - ✓ **no more than 120 cores** during working day and working hours

- For **ALL** publications having results produced using the UL HPC
  - ↪ Acknowledge / cite the UL HPC facility (using official banner)
  - ↪ Tag your publication upon registration on ORBiLu.

Summary

Compute Nodes Environment

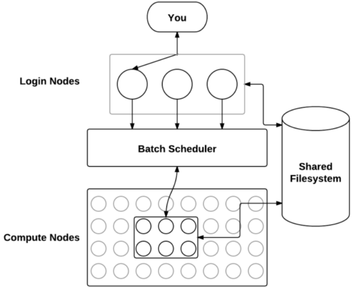


- *OS*: Debian 8 / CentOS 7
- **Storage usage**: `df-ulhpc`
- **Env. modules**: `modules`
  - ↪ **Not** available on frontends
  - ↪ ***Only*** on compute nodes

Directory	Max size	Max #files	Backup
$HOME (chaos,gaia)	100 GB	1.000.000	YES
$HOME (iris)	500 GB	1.000.000	YES
$WORK (excl. iris)	3 TB		NO
$SCRATCH	per request		NO

Resource and Job Management Systems

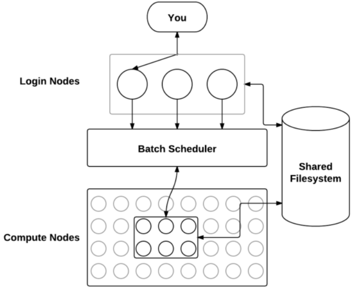
- **Resource and Job Management System (RJMS)**
  - ↪ *Glue* for a parallel computer to execute parallel jobs
  - ↪ **Goal**: satisfy users demands for computation
    - ✓ assign resources to user jobs with an efficient manner



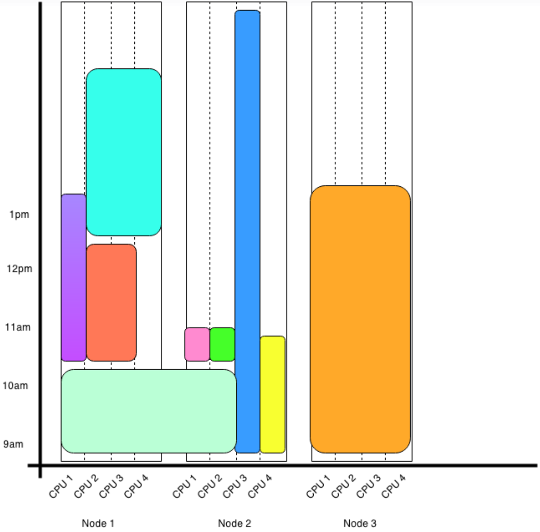
Resource and Job Management Systems

- **Resource and Job Management System (RJMS)**
  - ↪ *Glue* for a parallel computer to execute parallel jobs
  - ↪ **Goal**: satisfy users demands for computation
    - ✓ assign resources to user jobs with an efficient manner

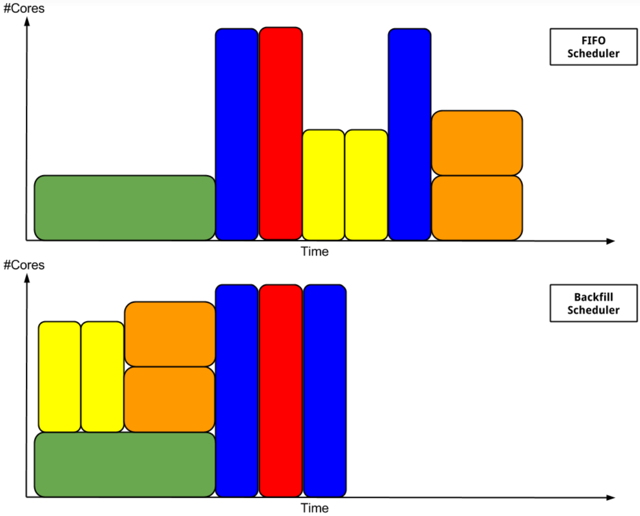
- **HPC Resources**:
  - ↪ Nodes (typically a unique IP address)
    - ✓ Sockets / Cores / Hyperthreads
    - ✓ Memory
    - ✓ Interconnect/switch resources
  - ↪ Generic resources (e.g. GPUs)
  - ↪ Licenses
- **Strategic Position**
  - ↪ Direct/constant knowledge of resources
  - ↪ Launch and otherwise manage jobs



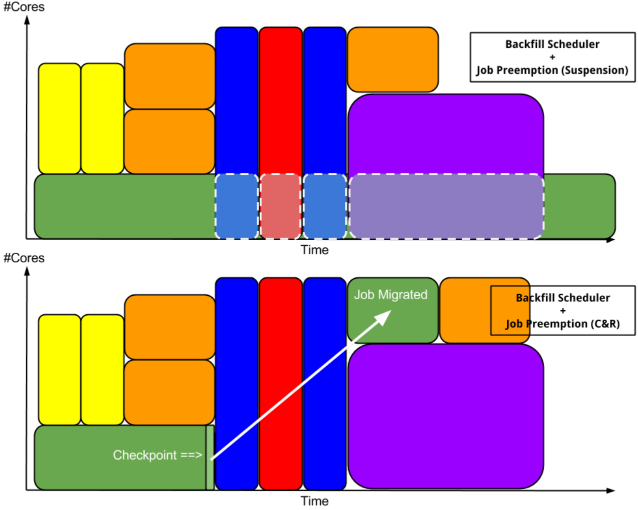
Job Scheduling



Job Scheduling (backfilling)



Job Scheduling (suspension & requeue)



List of the Main Job Schedulers

Name	Company	Version*
SLURM	SchedMD	17.11.5
LSF	IBM	10.1
OpenLava	LSF Fork	2.2
MOAB/Torque	Adaptative Computing	6.1
PBS	Altair	13.0
OAR (PBS Fork)	LIG	2.5.78
Oracle Grid Engine (formely SGE)	Oracle	

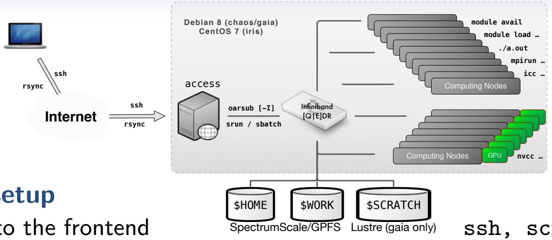
**: As of June. 2018*

List of the Main Job Schedulers

Name	Company	Version*
SLURM	SchedMD	17.11.5
LSF	IBM	10.1
OpenLava	LSF Fork	2.2
MOAB/Torque	Adaptative Computing	6.1
PBS	Altair	13.0
OAR (PBS Fork)	LIG	2.5.78
Oracle Grid Engine (formely SGE)	Oracle	

**: As of June. 2018*

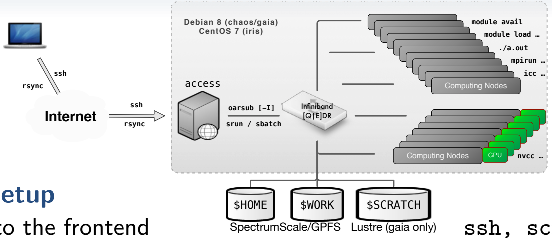
Typical Workflow on UL HPC resources



- **Preliminary setup**
  1. Connect to the frontend `ssh, screen`
  2. Synchronize you code `scp/rsync/svn/git`
  3. Reserve a few interactive resources `oarsub -I [...]`

  or, on iris: `srun -p interactive [...]`
  - ✓ *(eventually)* build your program `gcc/icc/mpicc/nvcc..`
  - ✓ Test on small size problem `mpirun/srun/python/sh...`
  - ✓ Prepare a launcher script `<launcher>.{sh|py}`

Typical Workflow on UL HPC resources



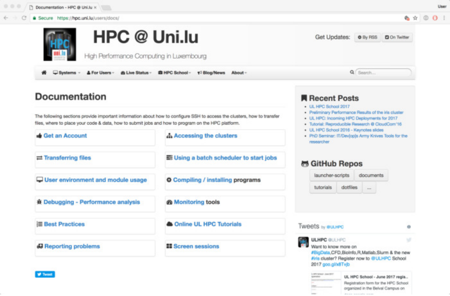
- **Preliminary setup**
  1. Connect to the frontend `ssh, screen`
  2. Synchronize you code `scp/rsync/svn/git`
  3. Reserve a few interactive resources `oarsub -I [...]`
     or, on iris: `srun -p interactive [...]`
     - ✓ *(eventually)* build your program `gcc/icc/mpicc/nvcc..`
     - ✓ Test on small size problem `mpirun/srun/python/sh...`
     - ✓ Prepare a launcher script `<launcher>.{sh|py}`
- **Real Experiment**
  1. Reserve passive resources `oarsub [...] <launcher>`
     or, on iris: `sbatch -p {batch|long} [...] <launcher>`
  2. Grab the results `scp/rsync/svn/git ...`

Summary

Documentation

http://hpc.uni.lu/users/getting_started.html

... aka the **rtf$_{\text{ine}}$m** paradigm



http://hpc.uni.lu

Reference Documentation

http://hpc.uni.lu/docs/

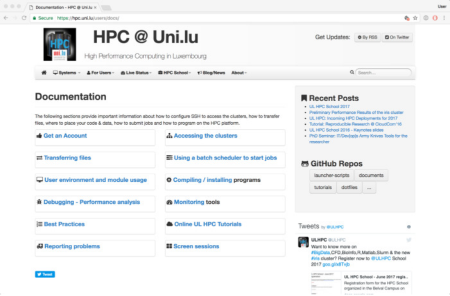
Online Tutorials

http://ulhpc-tutorials.rtfd.io/

Documentation

http://hpc.uni.lu/users/getting_started.html

... aka the **rtf$_{ine}$m** paradigm



http://hpc.uni.lu

Reference Documentation

http://hpc.uni.lu/docs/

Online Tutorials

http://ulhpc-tutorials.rtfd.io/

- UL HPC Ticketing System
  - ↪ https://hpc-tracker.uni.lu/
  - ↪ merge service.uni.lu ?
- Ask other users hpc-users@uni.lu
  - ↪ . . . or us hpc-sysadmins@uni.lu

Reporting Problems

https://hpc.uni.lu/users/docs/report_pbs.html

- **First checks**
  1. My issue is probably documented — see User Doc
  2. An event is on-going — cf mail from `hpc-platform@uni.lu`
  3. check the state of your nodes
     - ✓ `{ oarsub -C <jobid> | ssh <node>}; htop` — *on active jobs*
     - ✓ `{ oarsub -f -j <jobid> }` — *post-mortem*
     - ✓ **iris**: `scontrol show job <jobid>` OR `sacct --job <jobid> -l`
     - ✓ Ganglia on your node(s) — https://hpc.uni.lu/status/ganglia.html

Reporting Problems

- **First checks** — https://hpc.uni.lu/users/docs/report_pbs.html
  1. My issue is probably documented — see User Doc
  2. An event is on-going — cf mail from `hpc-platform@uni.lu`
  3. check the state of your nodes
     - ✓ `{ oarsub -C <jobid> | ssh <node>}; htop` — *on active jobs*
     - ✓ `{ oarsub -f -j <jobid> }` — *post-mortem*
     - ✓ **iris**: `scontrol show job <jobid>` OR `sacct --job <jobid> -l`
     - ✓ Ganglia on your node(s) — https://hpc.uni.lu/status/ganglia.html

- **ONLY NOW**, consider the following depending on the severity:
  - ↪ Open an new issue on `http://hpc-tracker.uni.lu` (**preferred**)
  - ↪ Mail (only now) us — `hpc-sysadmins@uni.lu`
  - ↪ **Ask the help of other users** — `hpc-users@uni.lu`

Reporting Problems

https://hpc.uni.lu/users/docs/report_pbs.html

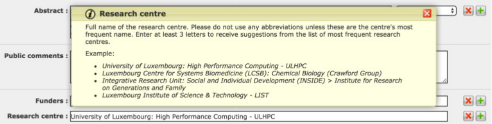
- **First checks**
  1. My issue is probably documented — see User Doc
  2. An event is on-going — cf mail from `hpc-platform@uni.lu`
  3. check the state of your nodes
     - ✓ `{ oarsub -C <jobid> | ssh <node>}; htop` — *on active jobs*
     - ✓ `{ oarsub -f -j <jobid> }` — *post-mortem*
     - ✓ **iris**: `scontrol show job <jobid>` OR `sacct --job <jobid> -l`
     - ✓ Ganglia on your node(s) — https://hpc.uni.lu/status/ganglia.html

- **ONLY NOW**, consider the following depending on the severity:
  - ↪ Open an new issue on `http://hpc-tracker.uni.lu` (**preferred**)
  - ↪ Mail (only now) us `hpc-sysadmins@uni.lu`
  - ↪ **Ask the help of other users** `hpc-users@uni.lu`

- In all cases: **Carefully describe the problem and the context**
  - ↪ Guidelines

Reporting Obtained Results

- In your **scientific publications**: *as per Acceptable Use Policy (AUP)*
  - ↪ **acknowledge** your usage of the UL HPC platform
  - ↪ *(if possible)* **cite** the UL HPC paper `\cite{VBCG_HPCS14}`
- **More importantly**: add ULHPC Tag on your ORBi[lu] publication




```
@InProceedings{VBCG_HPCS14,
   author =       {S. Varrette and P. Bouvry and H. Cartiaux and F. Georgatos},
   title =        {Management of an Academic HPC Cluster: The UL Experience},
   booktitle =    {Proc. of the 2014 Intl. Conf. on High Performance Computing \& Simulation (HPCS 2014)},
   year =         {2014},
   pages =        {959--967},
   month =        {July},
   address =      {Bologna, Italy},
   publisher =    {IEEE},
 }
```

Summary

HPC International State of Affairs

Global race toward Exascale Technology

IDC-Projected Exascale Investment Levels (In Addition to System Purchases)

U.S.
- $1 to $2 billion a year in R&D (including NRE)
- Investments by both governments & vendors
- Plans are to purchase multiple exascale systems

EU
- About 5 billion euros in total
- Investments in multiple exascale and pre-exascale systems
- Investments mostly by country governments with a little from the EU

China
- Over $1billion a year in R&D
- Investments by both governments & vendors
- Plans are to purchases multiple exascale systems each year
- Already investing in 3 pre-exascale systems by 2017/18

Japan
- Planned investment of just over $1billion* (over 5 years) for both the R&D and purchase of 1 exascale system
- To be followed by a number of smaller systems ~$100M to $150M each
- Creating a new processor and a new software environment

** Note that this includes both the system and R&D*

HPC International State of Affairs

Global race toward Exascale Technology

IDC-Projected Exascale Dates and Suppliers

U.S.
- Sustained ES: 2023
- Peak ES: 2021
- Vendors: U.S.
- Processors: U.S.
- Initiatives: NSCI/ECP
- Cost: $300-500M per system, plus heavy R&D investments

EU
- Sustained ES: 2023-24
- Peak ES: 2021
- Vendors: U.S., Europe
- Processors: U.S., ARM
- Initiatives: PRACE, ETP4HPC
- Cost: $300-$350 per system, plus heavy R&D investments

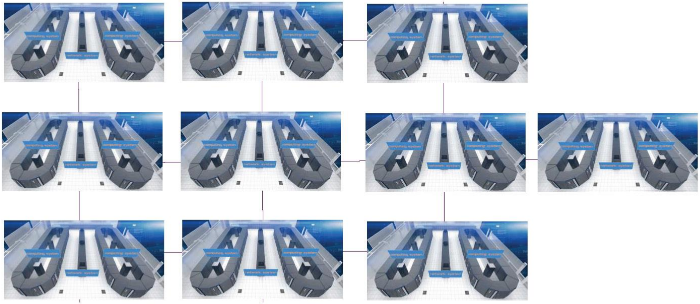
China
- Sustained ES: 2023
- Peak ES: ~~2020~~ 2019…
- Vendors: Chinese
- Processors: Chinese (plus ~~U.S.~~?)
- 13th 5-Year Plan
- Cost: $350-500M per system, plus heavy R&D

Japan
- Sustained ES: 2023-24
- Peak ES: Not planned
- Vendors: Japanese
- Processors: Japanese
- Cost: $600-850M, this includes both 1 system and the R&D costs…will also do many smaller size systems

Exascale Feasibility

We Can Build an Exascale System Today?

Connect together 10 Sunway TaihuLight systems



Require 150 MW of power, programming for 100 M threads, and $2.7B price tag 22

European HPC strategy

- EU HPC strategy initiated in 2012
  - ↪ implementation within H2020 program

European HPC strategy


IMPORTANT PROJECT
OF COMMON
EUROPEAN INTEREST
(IPCEI)

ON
HIGH PERFORMANCE COMPUTING
AND
BIG DATA ENABLED APPLICATIONS
(IPCEI-HPC-BDA)

European Strategic Positioning Paper

Luxembourg, France, Italy (& Spain)
November 2015

- EU HPC strategy initiated in 2012
  - ↪ implementation within H2020 program
- More recently:
  - ↪ IPCEI on HPC and Big Data (BD) Applications (Nov. 2015)
    - ✓ Luxembourg (leader), France, Italy & Spain
    - ✓ Testbed around Personalized Medicine, Smart Space, Industry 4.0, Smart Manufacturing, New Materials, FinTech, Smart City. . .

European HPC strategy

DECLARATION

Cooperation framework on High Performance Computing

Bundesrepublik Deutschland

and

República Portuguesa

and

République française

and

Reino de España

and

Repubblica Italiana

and

Grand-Duché de Luxembourg

and

Koninkrijk der Nederlanden

The signing Member States agree to work together towards making available across the EU an integrated world-class high performance computing (HPC) infrastructure, which in combination with European data and network infrastructures would upraise Europe's scientific capabilities and industrial competitiveness.

- EU HPC strategy initiated in 2012
  - ↪ implementation within H2020 program
- More recently:
  - ↪ IPCEI on HPC and Big Data (BD) Applications
    - ✓ Luxembourg (leader), France, Italy & Spain
    - ✓ Testbed around Personalized Medicine, Smart Space, Industry 4.0, Smart Manufacturing, New Materials, FinTech, Smart City...
- **Latest advances**:
  - ↪ EU Member States sign EuroHPC (Mar. 2017)
    - ✓ common effort to create/grow European supercomputing ecosystem
    - ✓ Federation of national/regional HPC centers (see also PRACE2)
  - ↪ EU Objective with EuroHPC:
    - ✓ EuroHPC JU effectively operational starting **Jan 1st, 2019**
    - ✓ 2-3 **Pre-exascale** systems by 2021, **2 exascale** systems by 2023

EU HPC Strategy Implementation

- **European Technology Platform (ETP) for HPC**
  - ↪ Industry-led forum feat. HPC stakeholders
  - ↪ Providing EU framework to define HPC research priorities/actions
    - ✓ UL (P. Bouvry, S. Varrette, V.Plugaru) part of ETP4HPC (2016-)
    - ✓ See Strategic Research Agenda, 2017 European HPC Handbook...

EU HPC Strategy Implementation

- **European Technology Platform (ETP) for HPC**
  - ↪ Industry-led forum feat. HPC stakeholders
  - ↪ Providing EU framework to define HPC research priorities/actions
    - ✓ UL (P. Bouvry, S. Varrette, V.Plugaru) part of ETP4HPC (2016-)
    - ✓ See Strategic Research Agenda, 2017 European HPC Handbook...
- EU **COST Actions**, for instance:
  - ↪ NESUS: Network for Sustainable Ultrascale Computing
  - ↪ cHiPSet: High-Performance Modelling and Simulation for BDA

EU HPC Strategy Implementation

- **European Technology Platform (ETP) for HPC**
  - ↪ Industry-led forum feat. HPC stakeholders
  - ↪ Providing EU framework to define HPC research priorities/actions
    - ✓ UL (P. Bouvry, S. Varrette, V.Plugaru) part of ETP4HPC (2016-)
    - ✓ See Strategic Research Agenda, 2017 European HPC Handbook...

- EU **COST Actions**, for instance:
  - ↪ NESUS: Network for Sustainable Ultrascale Computing
  - ↪ cHiPSet: High-Performance Modelling and Simulation for BDA

- **PRACE** - Partnership for Advanced Computing in Europe
  - ↪ Non-profit association, 25 member countries, now entering PRACE2
  - ↪ Providing access to **Five EU Tier-0** compute & data resources
  - ↪ Luxembourg 25th country to join (Oct. 17th, 2017)
    - ✓ Official Delegate/Advisor (P. Bouvry/S. Varrette) from UL

EU HPC Strategy Implementation

- **European High-Performance Computing Joint Undertaking**
  - ↪ EuroHPC JU effectively operational starting **Jan 1st, 2019**
    - ✓ administrative management from Luxembourg
  - ↪ Public and private members
    - ✓ EC, 14 MS, representatives from supercomputing/BD stakeholders
    - ✓ Governing Board (public members)
    - ✓ Industrial & Scientific Advisory Board (private members)
  - ↪ EU Objective with EuroHPC:
    - ✓ 2-3 **Pre-exascale** systems by 2021, **2 exascale** systems by 2023
    - ✓ Pending decision on hosting countries

EU HPC Strategy Implementation

- **European High-Performance Computing Joint Undertaking**
  - ↪ EuroHPC JU effectively operational starting **Jan 1st, 2019**
    - ✓ administrative management from Luxembourg
  - ↪ Public and private members
    - ✓ EC, 14 MS, representatives from supercomputing/BD stakeholders
    - ✓ Governing Board (public members)
    - ✓ Industrial & Scientific Advisory Board (private members)
  - ↪ EU Objective with EuroHPC:
    - ✓ 2-3 **Pre-exascale** systems by 2021, **2 exascale** systems by 2023
    - ✓ Pending decision on hosting countries

EuroHPC Budget: 2 × 486 M€

EU HPC Strategy Implementation

- **European High-Performance Computing Joint Undertaking**
  - ↪ EuroHPC JU effectively operational starting **Jan 1st, 2019**
    - ✓ administrative management from Luxembourg
  - ↪ Public and private members
    - ✓ EC, 14 MS, representatives from supercomputing/BD stakeholders
    - ✓ Governing Board (public members)
    - ✓ Industrial & Scientific Advisory Board (private members)
  - ↪ EU Objective with EuroHPC:
    - ✓ 2-3 **Pre-exascale** systems by 2021, **2 exascale** systems by 2023
    - ✓ Pending decision on hosting countries

EuroHPC Budget: 2 × 486 M€

- **European Processor Initiative (EPI)**
  - ↪ Initial plan vs current plan...
  - ↪ **120 M€** via Framework Partnership Agreement (FPA)

New Trends in HPC

- **Continued scaling** of scientific, industrial & financial applications
  - ↪ . . . well beyond Exascale
- New trends changing the landscape for HPC
  - ↪ Emergence of **Big Data analytics**
  - ↪ Emergence of (**Hyperscale**) **Cloud Computing**
  - ↪ **Data intensive Internet of Things (IoT)** applications
  - ↪ **Deep learning & cognitive computing** paradigms

[Source : EuroLab-4-HPC]


This study was carried out for RIKEN by
IDC
Analyze the Future

Special Study

Analysis of the Characteristics and Development Trends of the
Next-Generation of Supercomputers in Foreign Countries

Earl C. Joseph, Ph.D. — Robert Sorensen
Steve Conway — Kevin Monroe

[Source : IDC RIKEN report, 2016]

Summary

Conclusion

- **Luxembourg government priority on HPC**
  - ↪ sustained by University of Luxembourg HPC developments
    - ✓ started in 2007, under resp. of Prof P. Bouvry & Dr. S. Varrette
    - ✓ expert UL HPC team *(S. Varrette, V. Plugaru, S. Peter, H. Cartiaux, C. Parisot)*
  - ↪ UL HPC (as of 2018): **346.652 TFlops, 9852.4TB (shared)**
  - ↪ consolidate and extend Europe efforts on HPC/Big Data

Conclusion

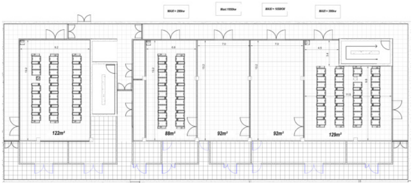
- **Luxembourg government priority on HPC**
  - ↪ sustained by University of Luxembourg HPC developments
    - ✓ started in 2007, under resp. of Prof P. Bouvry & Dr. S. Varrette
    - ✓ expert UL HPC team *(S. Varrette, V. Plugaru, S. Peter, H. Cartiaux, C. Parisot)*
  - ↪ UL HPC (as of 2018): **346.652 TFlops, 9852.4TB (shared)**
  - ↪ consolidate and extend Europe efforts on HPC/Big Data

Several On-going Strategic HPC efforts in Europe...

- **... in which UL *(HPC)* is involved ...**
  - ↪ ETP4HPC, EU COST Action NESUS etc.
  - ↪ PRACE - Official representative for Luxembourg from UL
    - ✓ Delegate: Prof. Pascal Bouvry
    - ✓ Advisor: Dr. Sebastien Varrette
  - ↪ EuroHPC / IPCEI on HPC and Big Data (BD) Applications
  - ↪ **National HPC-BD Competence Center**

Incoming Milestones

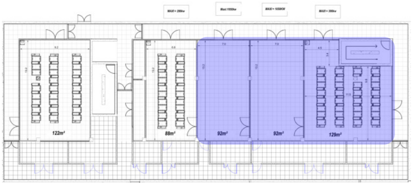
- MSA **CDC S-02** as the **new UL HPC Data Center (DC)**



- $\simeq$ **1050kW** per **HPC** room
  - ↪ Direct Liquid Cooling (DLC)
- $\simeq$ **300kW** per **storage** room
  - ↪ rooms 1, 2 & 5
  - ↪ Air-flow storage / HPC

Incoming Milestones

- MSA **CDC S-02** as the **new UL HPC Data Center (DC)**



- $\simeq$ **1050kW** per **HPC** room
  - ↪ Direct Liquid Cooling (DLC)
- $\simeq$ **300kW** per **storage** room
  - ↪ rooms 1, 2 & 5
  - ↪ Air-flow storage / HPC

- `iris` cluster deployed in CDC S-02-005
  - ↪ incl. GPFS/Lustre storage
- pending national commitment for having next-gen HPC in CDC
  - ↪ **DLC** rooms CDC S-02-{003,004} ready

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law
  - ↪ RFP 180027 (**release date**: July 2018)
    - ✓ `iris` **extension with GPU and `bigmem` nodes**

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law
  - ↪ RFP 180027 (**release date**: July 2018)
    - ✓ `iris` **extension with GPU and `bigmem` nodes**
  - ↪ **GDPR**: extending work in progress @ LCSB

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law
  - ↪ RFP 180027 (**release date**: July 2018)
    - ✓ `iris` **extension with GPU and** `bigmem` **nodes**
  - ↪ **GDPR**: extending work in progress @ LCSB
  - ↪ Reference platform research article update

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law
  - ↪ RFP 180027 (**release date**: July 2018)
    - ✓ `iris` **extension with GPU and** `bigmem` **nodes**
  - ↪ **GDPR**: extending work in progress @ LCSB
  - ↪ Reference platform research article update
  - ↪ **Luxembourg HPC-BD Competence Center** (exp. mid 2019)
    - ✓ National commitment for having next-gen HPC in CDC S-02

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law
  - ↪ RFP 180027 (**release date**: July 2018)
    - ✓ `iris` **extension with GPU and `bigmem` nodes**
  - ↪ **GDPR**: extending work in progress @ LCSB
  - ↪ Reference platform research article update
  - ↪ **Luxembourg HPC-BD Competence Center** (exp. mid 2019)
    - ✓ National commitment for having next-gen HPC in CDC S-02
  - ↪ EU: entering **PRACE2**

Incoming Milestones

- **Short term actions**
  - ↪ Official **Research computing @ UL & abroad** structure
    - ✓ pending HR positioning with the new law
  - ↪ RFP 180027 (**release date**: July 2018)
    - ✓ `iris` **extension with GPU and `bigmem` nodes**
  - ↪ **GDPR**: extending work in progress @ LCSB
  - ↪ Reference platform research article update
  - ↪ **Luxembourg HPC-BD Competence Center** (exp. mid 2019)
    - ✓ National commitment for having next-gen HPC in CDC S-02
  - ↪ EU: entering **PRACE2**
  - ↪ **International Cooperation**
    - ✓ EuroHPC
    - ✓ Thailand: NSTDA / Nectec / Biotec

Questions?

http://hpc.uni.lu

High Performance Computing @ uni.lu

Prof. Pascal Bouvry
Dr. Sebastien Varrette
Valentin Plugaru
Sarah Peter
Hyacinthe Cartiaux
Clement Parisot

University of Luxembourg, Belval Campus:
Maison du Nombre, 4th floor
2, avenue de l'Université
L-4365 Esch-sur-Alzette
mail: hpc@uni.lu