CloudMIP & RECS

Production-grade and R&D open-source cloud platforms

http://cloudmip.univ-tlse3.fr

Pr. Jean-Marc Pierson / IRIT
Dr. François Thiebolt / IRIT
{pierson,thiebolt}@irit.fr

Photo:The CloudMIP platform
DTSI hosted | UPS Toulouse 3

Université Paul Sabatier
TOULOUSE III

IRIT
Institut de Recherche en Informatique de Toulouse

FRANCE GRILLES

Q1 2015

Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- CloudMIP use cases | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

SEPIA team & platforms

The SEPIA team (IRIT: Pr Jean-Marc Pierson / N7: Pr Daniel Hagimont) mainly focuses on GreenIT, autonomic and distributed systems.

10 permanents (4 Pr, 6 MCF, 1 Dr-engineer)

2 engineers, 1 post-doc,
2 associated researchers,
16 PhD students.

CoolEmAll (FP7),
SVC (Grand Emprunt),
SOP and Control Green (ANR).

Toulouse platforms (Pr JM Pierson):

- Grid5000-Toulouse (560),
- GridMIP (128),
- CloudMIP (256),
- RECS2.0 (72) / 1U.

Pau platform (Pr P. Congduc):

- PireCloud (128).

SSTA axis platforms (Pr MP Gleizes):

- Amilab,
- neOCampus

(number of cores)

Data Nov. 2013

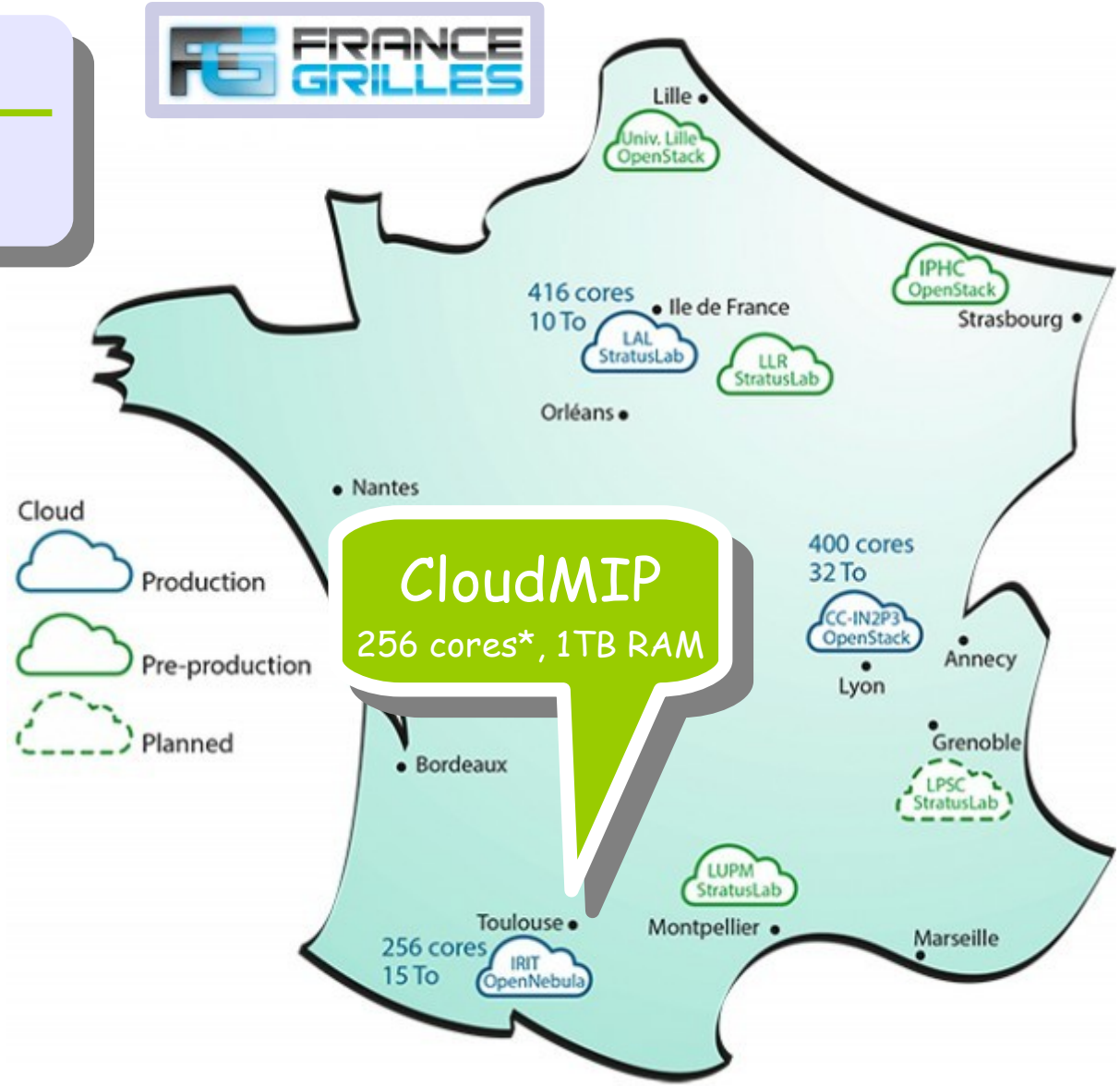
Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- CloudMIP use cases | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

France-Grilles' CloudMIP

Cloud initiative @ France-Grilles

- Inter-Cloud experimental testbed



**physical cores and 512 cores with hyper-threading activated.*

The CloudMIP platform

Facts and resources

- Funded by France-Grilles (100k€), installed in 2012,
- Who : Pr Jean-Marc Pierson (manager), Dr François Thiebolt, DTSI Network team,
- Location : Toulouse 3 university's Data Center,
- Taskforce : 1 Dr-engineer (30%)↘, 1 engineer (soon –-maybe),
- Fluids consumption annual cost (est.) : between 4k€ and 10k€,
- Status: production,

2 x Dell M1000e 16 blades

Hardware, system, middleware ...

- 2 x Dell M1000e chassis each filled with 16 blades ⇨ 256 **physical** cores, 1TB RAM, 15TB disk,
- System : Scientific Linux 6.5 x86_64,
- **OpenNebula** 4.4.1 (Cloud-Init and spice support) with **KVM** hypervisor and **CIMI*** api (DeltaCloud),
- OpenNebula GUI **Sunstone,**
- **Zabbix** monitoring.

+ power monitoring **+** seconds to launch a hundred of multi-Gb VMs **+** open-source software

** **C**loud **I**nfrastructure **M**anagement **I**nterface – DMTF standard*

The CloudMIP platform

... additional details

- Wiki → http://cloudmip.univ-tlse3.fr/
- Monitoring → http://cloudmip.univ-tlse3.fr/zabbix *login : green, passwd : cloudmip*
- GUI → http://cloudmip.univ-tlse3.fr:11000
- Deltacloud → http://cloudmip.univ-tlse3.fr:10998
- 32 blades (8 cores @ 2.4Ghz, 32GB ram, 2 x 146GB SAS 15ktpm RAID0),
  - ⇒ means upto 256 Amazon M3.medium instances (1 **physical** dedicated CPU and 4GB ram)
- OpenNebula 4.4.1 (KVM) with Qcow2 delta images to speedup deployment,
  - ⇒ a hundred of VMs in just a few seconds :)
- 1s resolution power monitoring of nodes,
- A 48TB, >500MB/s NFS server shared across G5K/MIP/Cloud,
  is a SunFire x4500 featuring SL62 and zfs filesystem,
  CloudMIP benefits of a dedicated 2 x Gigabit Ethernet LACP link.

The CloudMIP platform

... additional details

- Ways for users to access their VMs from the Internet :
  - ssh, vpn, spice display forwarding (automatic : **todo**), Sunstone GUI (**done**),
  - #1000 ports on the front-end node dedicated to routing (**done**),
  - #60 dedicated public IPs (**done**) with dynamic routing (manual : **done**, auto : **todo**).

The CloudMIP platform

- Pool of public IP to CloudMIP*

195.220.53.1
195.220.53.2
||||||||||||||||||||
195.220.53.57

subnet 195.220.53.0/26

cloudmip.univ-tlse3.fr

nfs.cloudmip.univ-tlse3.fr

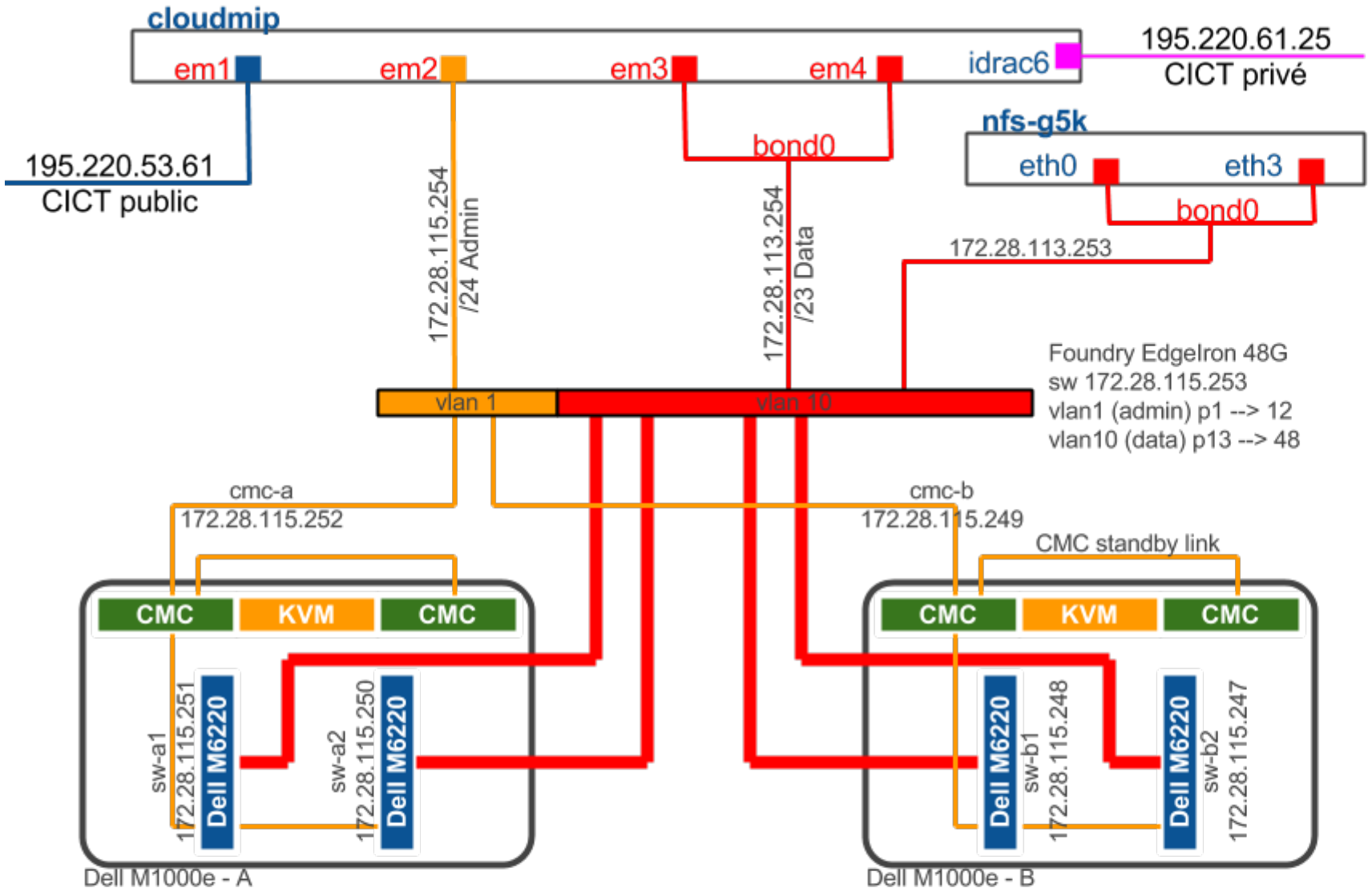
vm-112-102
vm-112-103
vm-112-104
vm-112-101

wn[1..32].cloudmip.univ-tlse3.fr

Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- CloudMIP use cases | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

Network


cloudmip
em1
em2
em3
em4
idrac6
195.220.61.25
CICT privé
195.220.53.61
CICT public
172.28.115.254
/24 Admin
bond0
172.28.113.254
/23 Data
nfs-g5k
eth0
eth3
bond0
172.28.113.253
Foundry EdgeIron 48G
sw 172.28.115.253
vlan1 (admin) p1 --> 12
vlan10 (data) p13 --> 48
vlan 1
vlan 10
cmc-a
172.28.115.252
cmc-b
172.28.115.249
CMC standby link
CMC
KVM
CMC
sw-a1
172.28.115.251
Dell M6220
sw-a2
172.28.115.250
Dell M6220
Dell M1000e - A
CMC
KVM
CMC
Dell M6220
sw-b1
172.28.115.248
Dell M6220
sw-b2
172.28.115.247
Dell M1000e - B

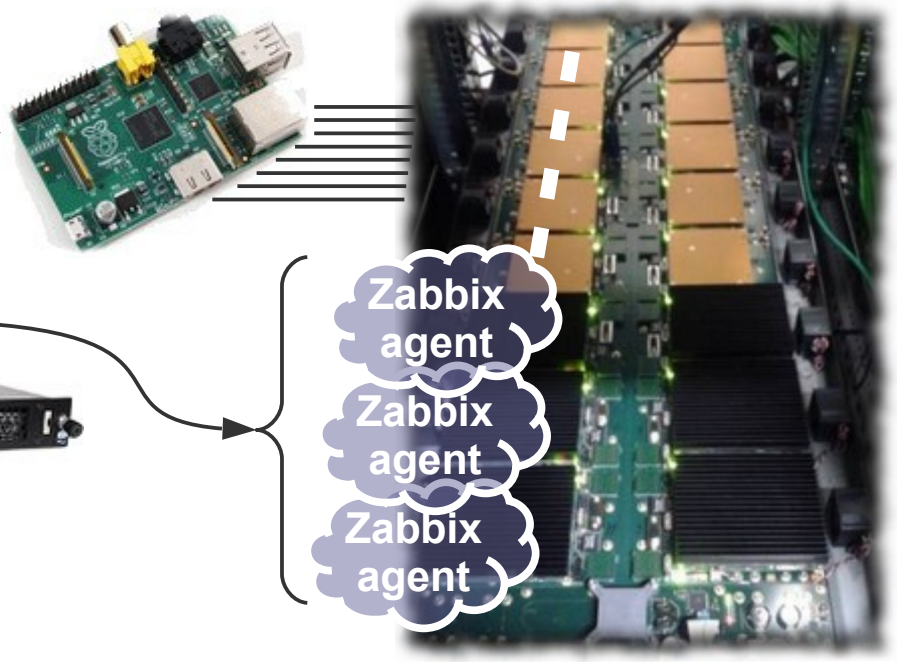
Monitoring

http://cloudmip.univ-tlse3.fr/zabbix

login : green
passwd : cloudmip

RPi + 12 x I2C T° sensors

RECS nodes



Zabbix proxy

RECS front node

Zabbix 2.0, #500 items/s,
tweaked MySQL db

Zabbix server

Httpd

CloudMIP front node

CloudMIP's worker nodes

Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- CloudMIP use cases | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

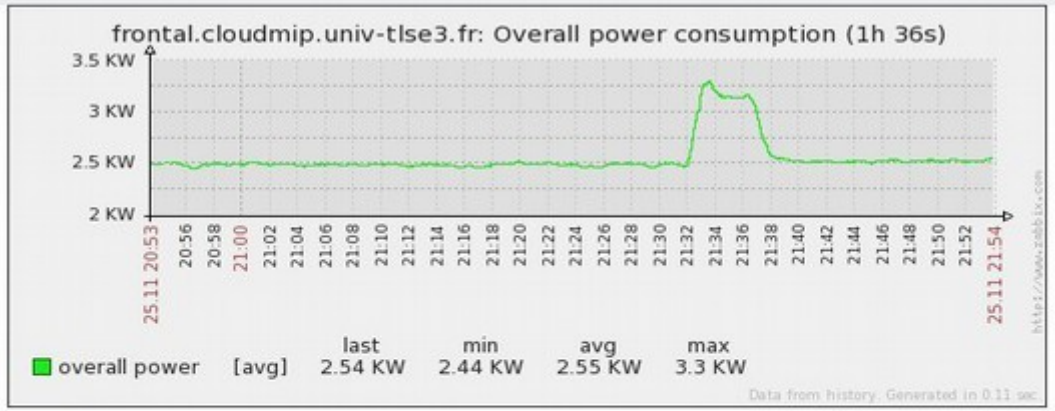
GreenIT: power metering

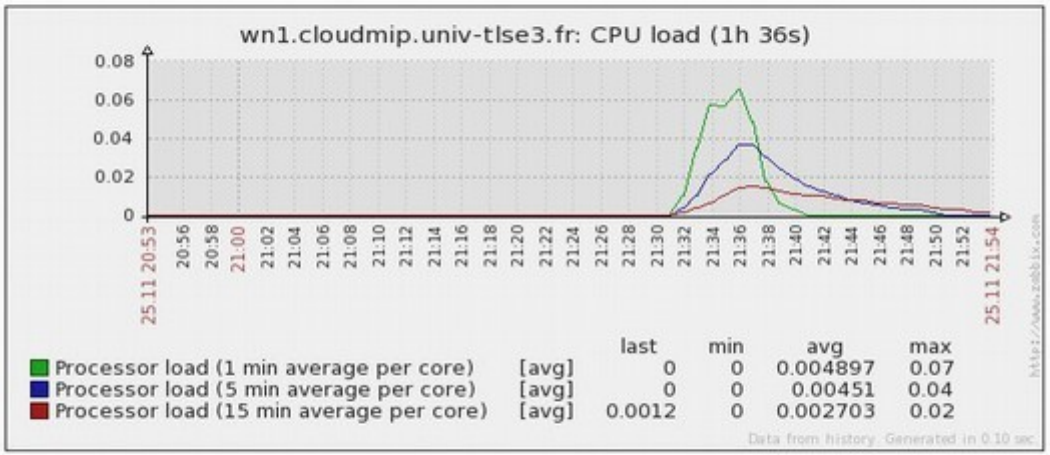
Launching 248 VMs

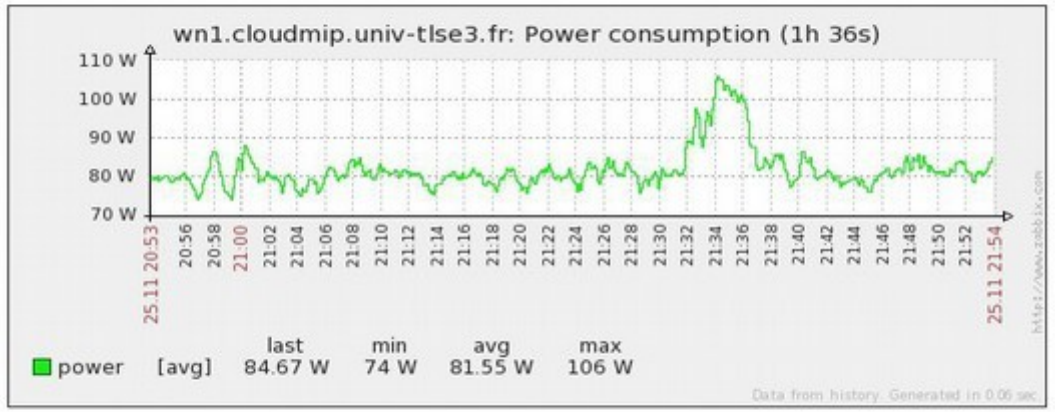
```
#> onetemplate instantiate SL64 -m 248
```

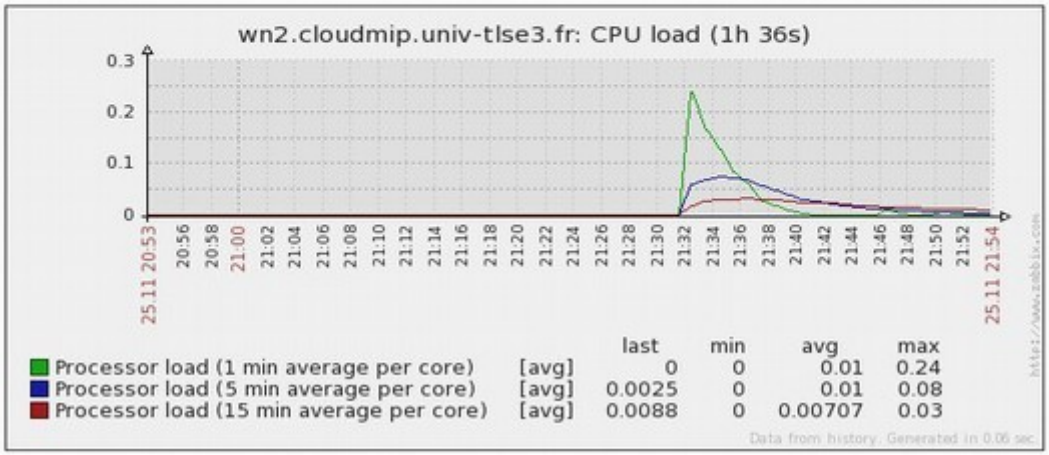
==> leads to eight VMs on each of the 31 nodes, thus each VM using one physical CPU.

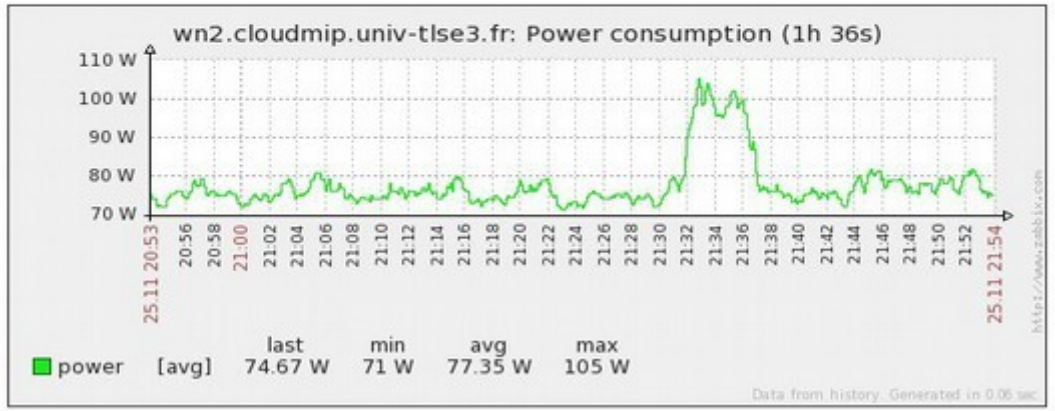
==> Max. power consumption is 3.3kw









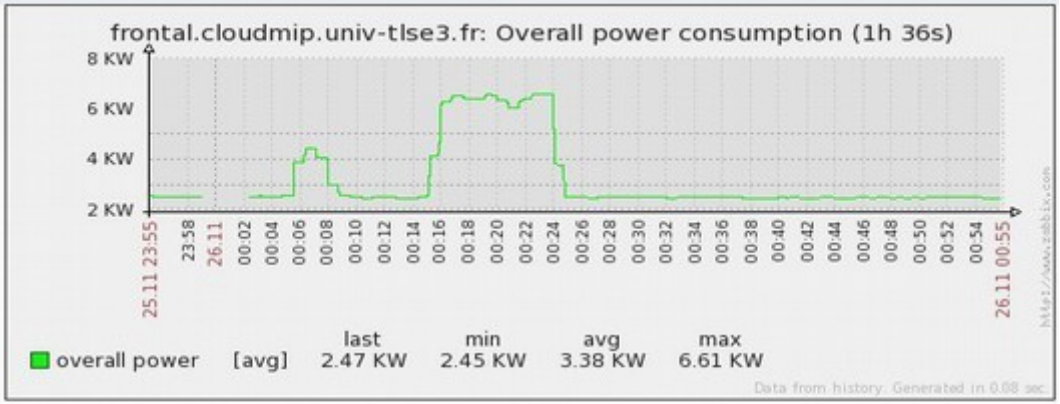


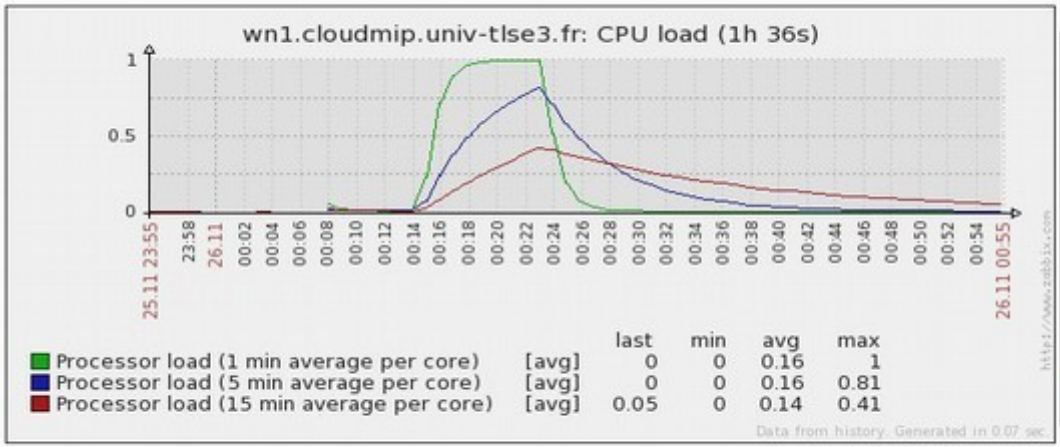
GreenIT: power metering

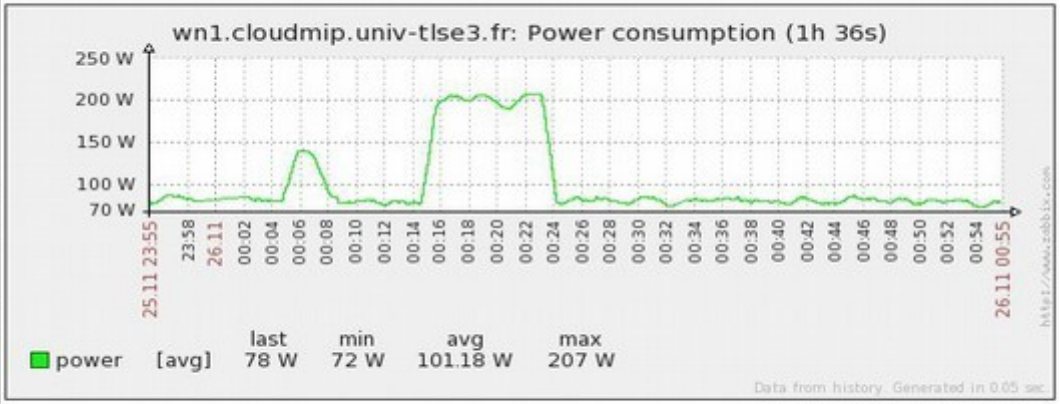
- Launching stress test on all nodes

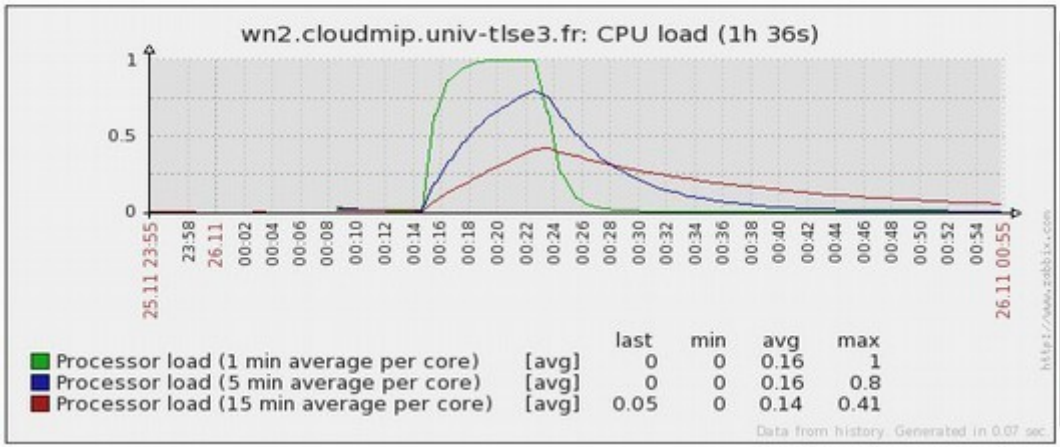
```
#> pdsh -w wn[1..32] -- openssl speed -multi 8
```

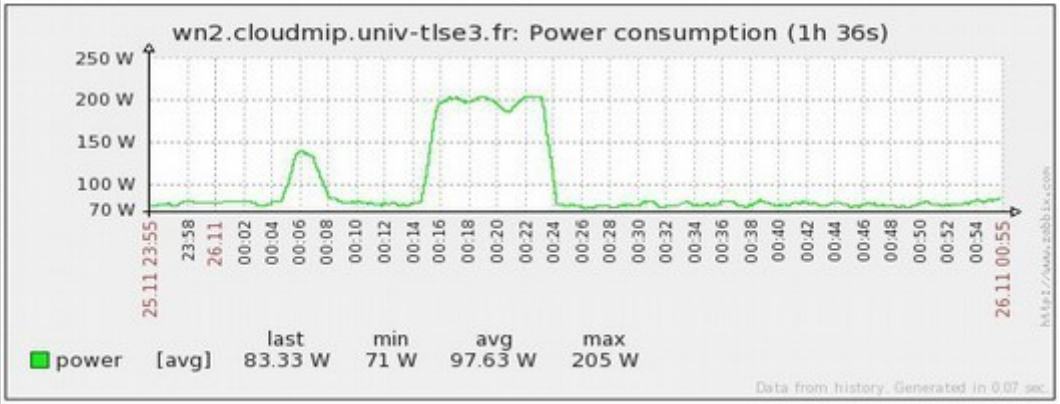
==> Peak power consumption is about 6.6kw











GreenIT: VM power metering

ECTOP | processes power consumption

ECTop is available under GPL license at https://github.com/cupertino/ectools

ectop

File Edit View Search Terminal Help

PAUSED - Thu Nov 29 09:33:12 2012
Tasks: 176, Power (W): 46.3891, CPU (%):97.1014

PID	COMMAND	%CPU v	RSS	PE_MMC2	PE_IC
5809	stress	50	30	17.5	22.9736
5810	stress	49.0385	30	17.2692	22.5318
5847	ectop	0.961538	897	5.73077	0.441801
2430	chrome	0.480769	10976	5.61538	0.2209
2145	gnome-terminal	0	3407	0	0
1711	Xorg	0	18354	0	0
7	migration/1	0	0	0	0
8	migration/1	0	0	0	0
9	ksoftirqd/1	0	0	0	0
10	watchdog/1	0	0	0	0
11	events/0	0	0	0	0
12	events/1	0	0	0	0
13	cpuset	0	0	0	0
14	khelper	0	0	0	0
15	netns	0	0	0	0
16	async/mgr	0	0	0	0
17	pm	0	0	0	0
18	sync_supers	0	0	0	0
				46.1154	46.1682

GreenIT: VM power metering

ECTOP* | processes power consumption (cont.)

Tool to estimate processes power consumption

Light weight**, several sensors (PerfCounters, CPU%, Memory, CPU temperature, …) and wattmeters (ACPI, G5K PDUs, CloudMIP, RECS, ...).

Two estimators implemented

Inverse model (PE_IC): calibration with power meter

$$P^{PID} = \frac{P^{Node} \times CPU^{PID}_{time}}{CPU^{Node}_{time}}$$

Linear model (PE_MMC2):

$$P^{PID} = \frac{(P^{Node}_{max} - P^{Node}_{min}) \times CPU^{PID}_{time}}{CPU^{Node}_{time}} + \frac{P^{Node}_{min}}{procs}$$

*** as low as Memory: 3Kb, CPU: 0.3%*

KVM hypervisor ==> each VM is a process thus leading to a possible evaluation of the power consumption of each VM!

*ongoing researches from Leandro Fontoura Cuppertino email:fontoura@irit.fr

GreenIT researches: VM aspects

[Sepia team] ongoing researches in the GreenIT field | application to virtual machine

- Deciding where to place VM, and which resources to allocate to it,
- Deciding which VM to migrate, and when,
- Deciding when to switch off or on hosts,
- Deciding how to operate the hosts (e.g. frequency scaling).

Experiments on **RECS** platform from the FP7 CoolEmAll project http://coolemall.eu

- All this with criteria of energy consumption and performance.
- At runtime,
- At system level and middleware level.

Contacts: Jean-Marc Pierson, Patricia Stolf, Damien Borgetto, Tom Guerout, Saïf ul Islam ...
{pierson, stolf, borgetto, guerout, islam}@irit.fr

Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- CloudMIP use cases | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

FG VMDirac

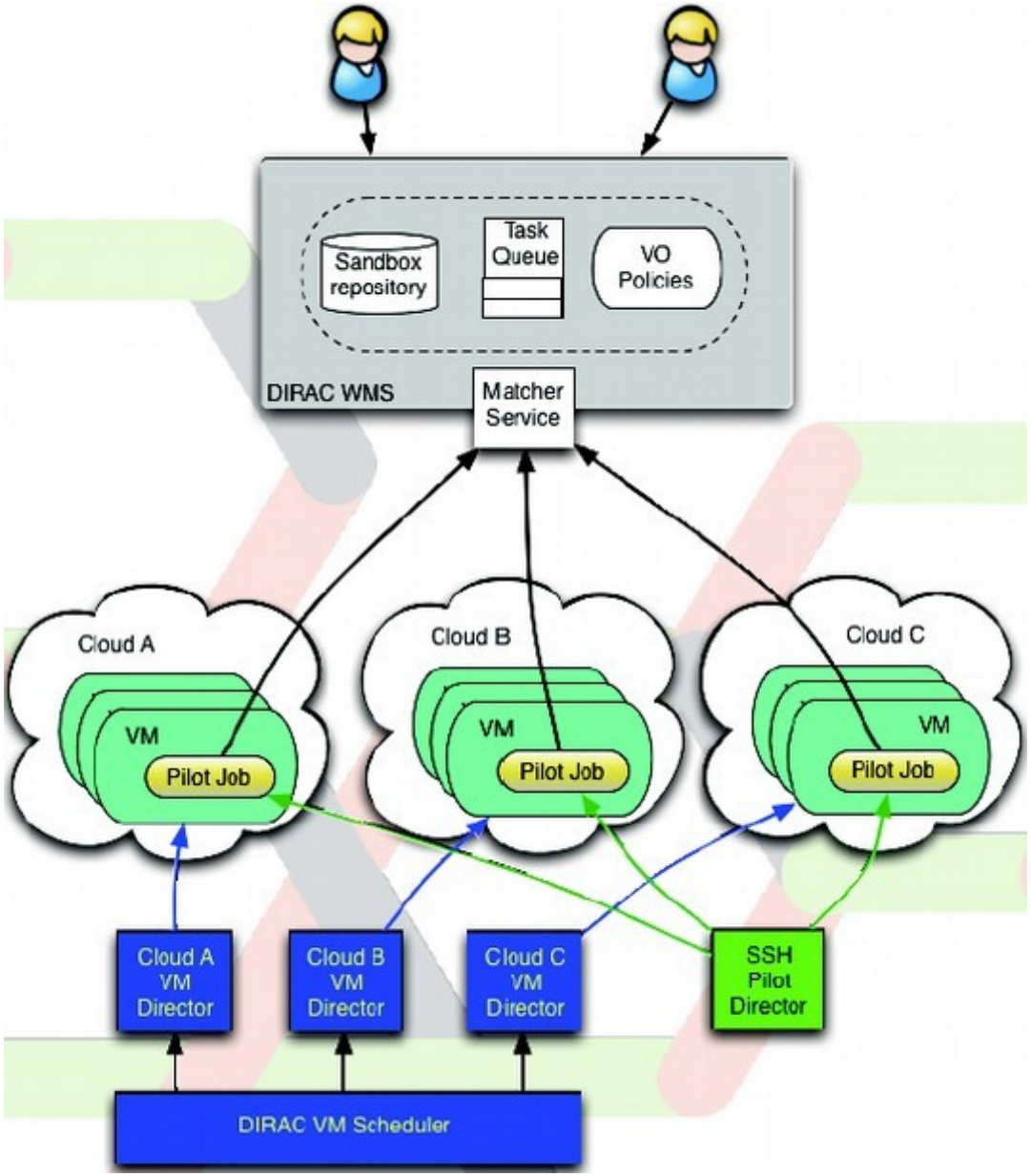
DIRAC is a framework for distributed computing written in Python and originally used for the BELLE experiment.

The VMDirac scheduler features:

- Dynamic VM spawning taking the Task Queue state into account,
- Discards VMs automatically when no more needed.

At the VMs side:

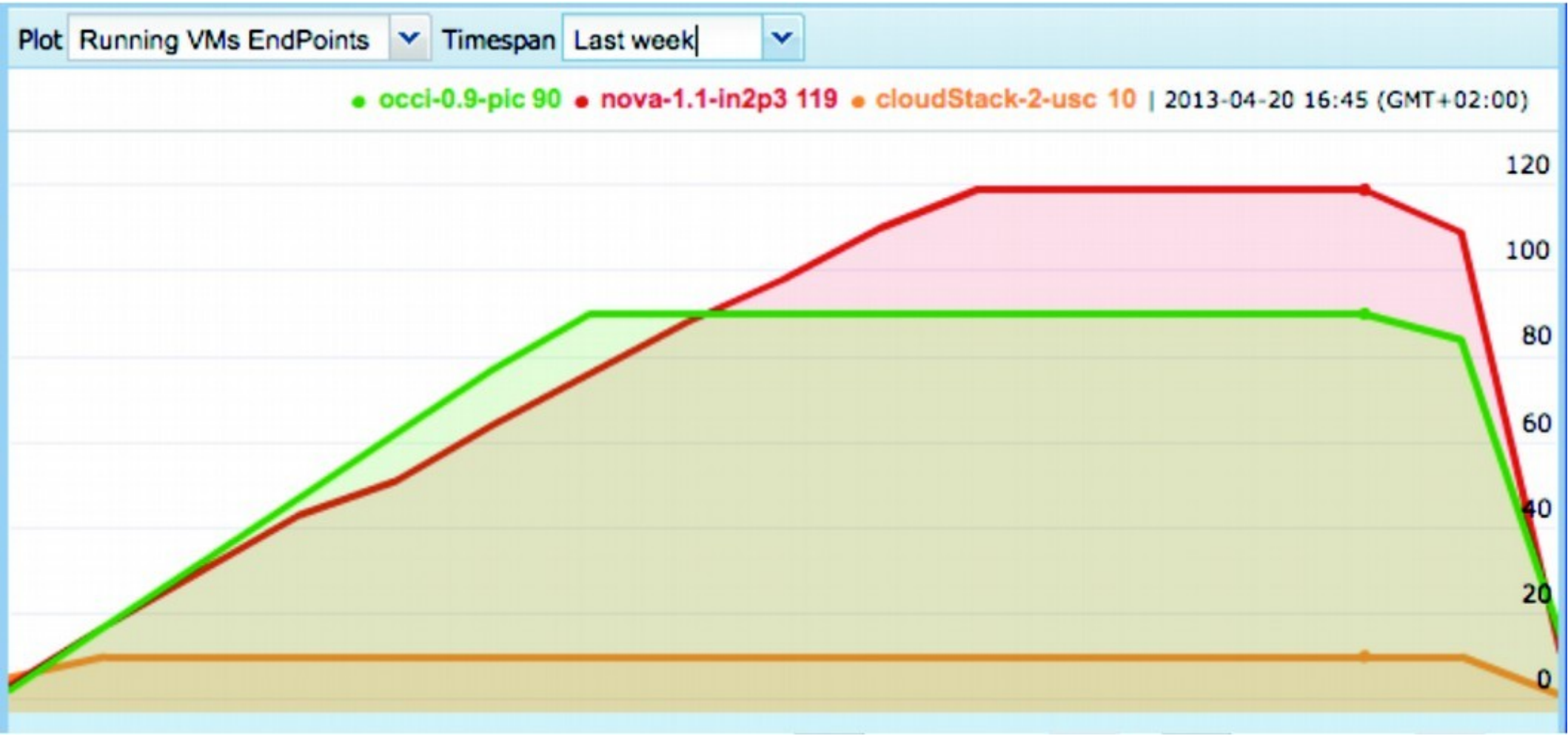
- User account and public key for SSH access,
- Internet-grade IPs (no private ones),
- At boot time, the VM start the “Pilot Job”,
- Contextualisation with .ISO image (OpenNebula – PIC), amiconfig (OpenStack –CC/IN2P3) and ad hoc image (CloudStack –USC),
- Start Job Agent and VM monitoring Agent.



Work from Andrei Tsaregorodtsev (CPPM), Victor Mendez (PIC), Victor Fernandez (USC), Mathieu Puel (CC/IN2P3).

FG VMDirac (next)

Total 250 VM slots at 3 sites:
USC, PIC, CC/IN2P3 with 3 different cloud managers
1 virtual CPU, 2GB RAM ==> up to 219 simultaneous VMs achieved.



Work from Andrei Tsaregorodtsev (CPPM), Victor Mendez (PIC), Victor Fernandez (USC), Mathieu Puel (CC/IN2P3).

FG-Cloud challenge

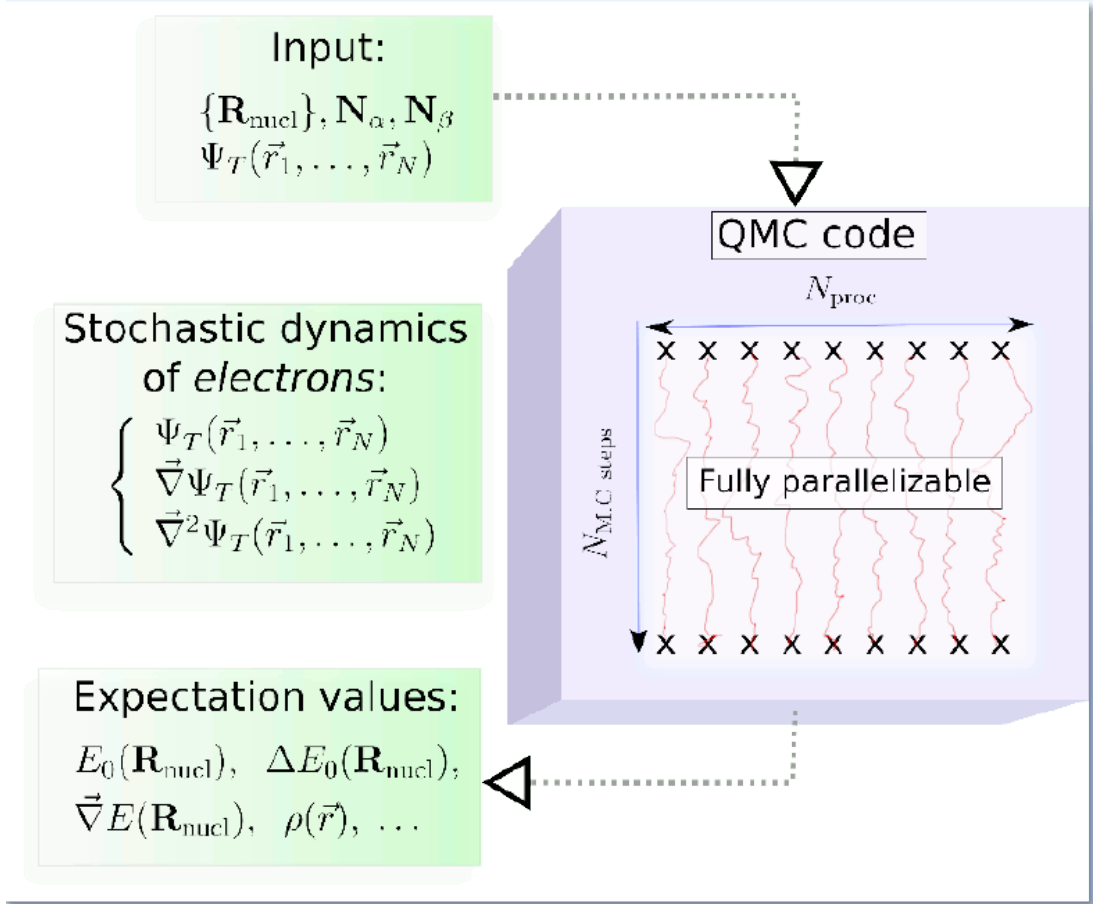
- [Dec. 14] France-Grille Cloud challenge call issued

[selected] A.Scemama (Irfamc) & G. Da Costa (Irit): **QMC**
ou Monté-Carlo Quantique appliqué à la chimie.

Quantum Monte Carlo (QMC) methods :

- require a small amount of memory (~100MB per core),
- Single core independent tasks. Communications are mandatory only at the initialization and the finalization stages,
- The initialization and finalization times don't depend on the length of the run,
- have a much better scaling than standard methods with the size of the chemical system but require a large amount of CPU time.

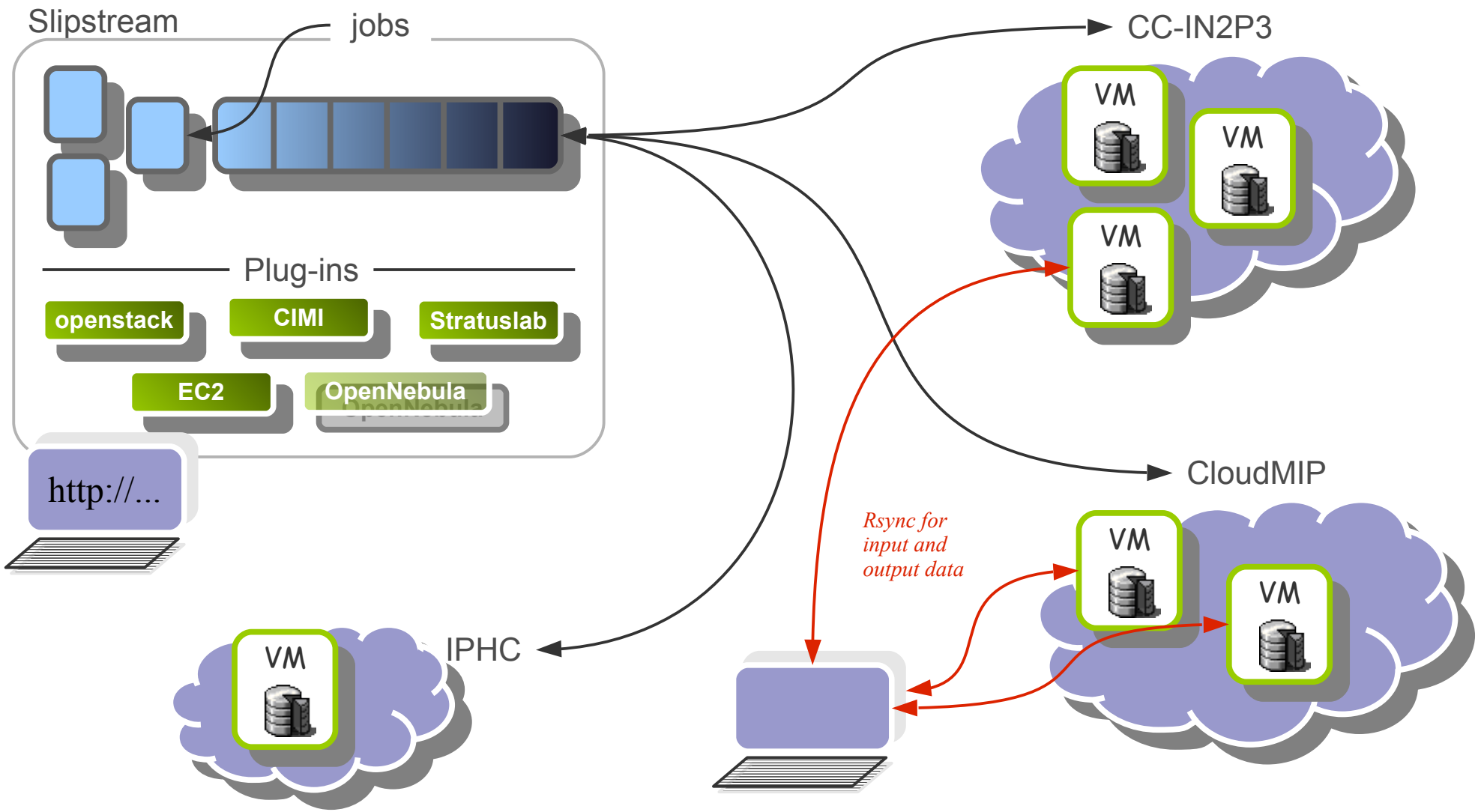
Welcome to the Cloud!



Anthonny Scemama / IRFAMC – UPS Toulouse 3
Quantum Monte-Carlo applied to chemistry
http://irpf90.ups-tlse.fr/files/sc11.pdf

FG-Cloud challenge (cont.)

FG Cloud challenge | architecture overview



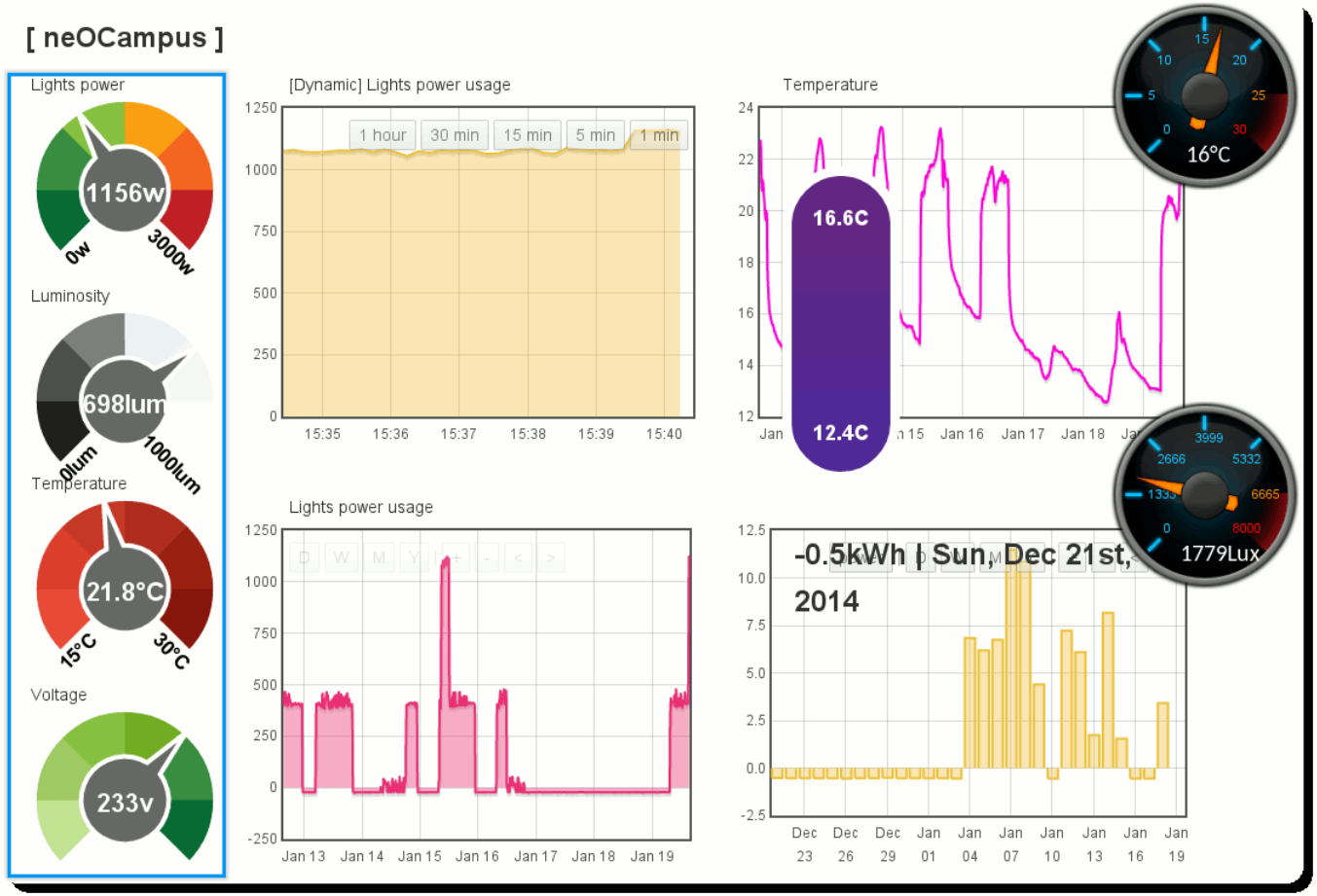
CloudMIP usage (cont.)

CloudMIP use case | alternate stuff

- neOCampus operation http://neocampus.univ-tlse3.fr/neocampus/main

  Campus-wide intelligent ambiant sensors / actuators (IoT)

- Virtual Arena (teach.)
- Inter-Cloud QOS (Saragosse University)
- ...



Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- Usage | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

CoolEmAll / RECS

European FP7 project that tackles Energy efficiency in Data Centers.

CoolEmAll testbed

- PSNC (Pozan),
- HLRS (Stuttgart),
- IRIT (Toulouse).

CoolEmAll consortium

- PSNC,
- HLRS,
- IRIT,
- IREC,
- Christmann Gmbh,
- 451 Research Ltd,
- ATOS.

RECS platform

Facts and resources

- European FP7 project CoolEmAll (Q4-2011 → Q1-2013),
- Who : Pr Jean-Marc Pierson (manager), Dr François Thiebolt,
- Location : IRIT's Data Center,
- Taskforce : 1 Dr-engineer,
- Status: production,

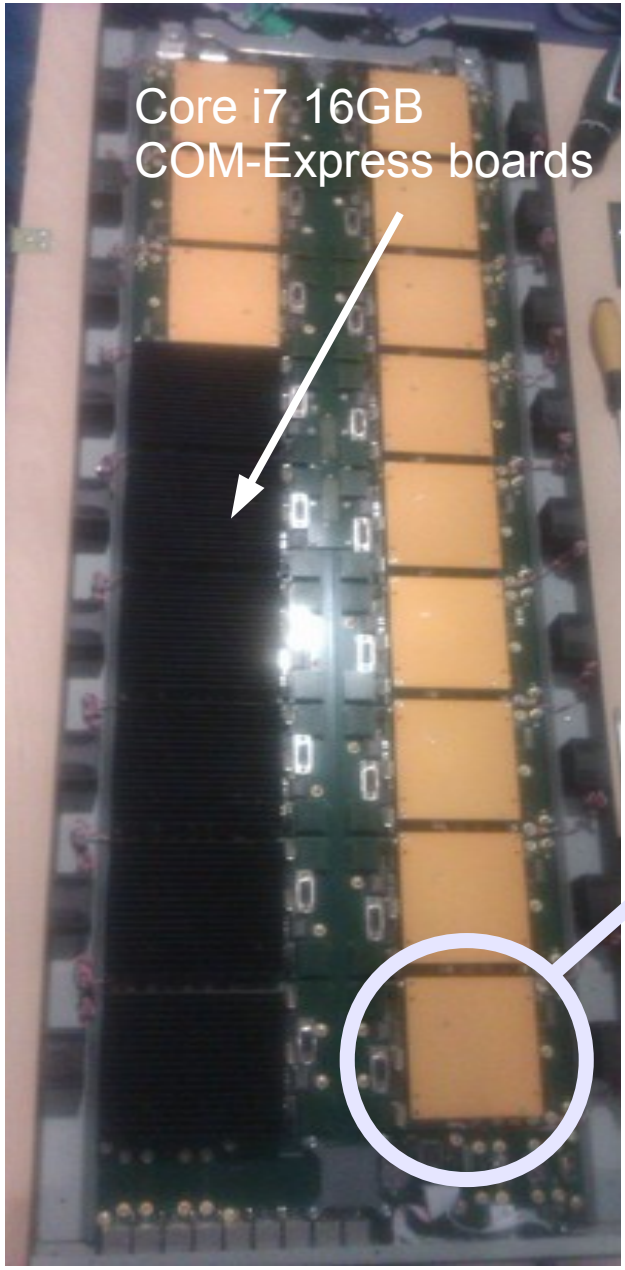
Christman's RECS 2.0

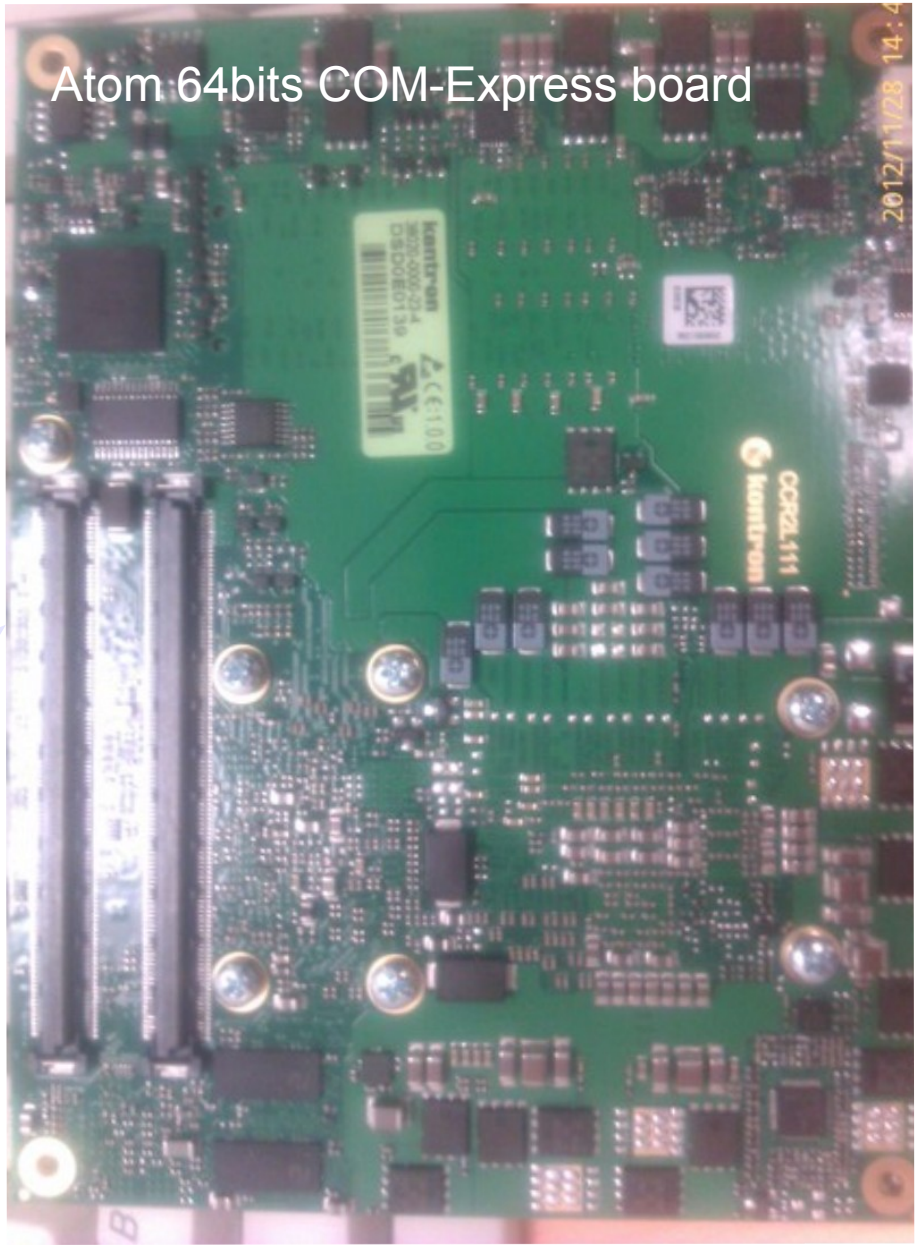
Hardware, system, middleware ...

- 18 nodes in a 1U chassis (6 x i7/4cores/16GB + 12 x Atom/2cores/2GB)
- System : Scientific Linux 6.4 x86_64,
- **OpenNebula** 4.4.1 (Cloud-Init and spice support) with **KVM** hypervisor,
- **NFSRoot read-only** + per-nodes deltas → one single image,
- **Zabbix** monitoring,
- Extended **temperature** monitoring (Raspberry Pi + 12 sensors → Zabbix)

GreenIT : RECS platform

RECS2.0: up to **72** CPU with **288GB** ram in **1U** chassis!





GreenIT : RECS platform

... external 12v 1080w power supply connection detail

GreenIT : RECS platform

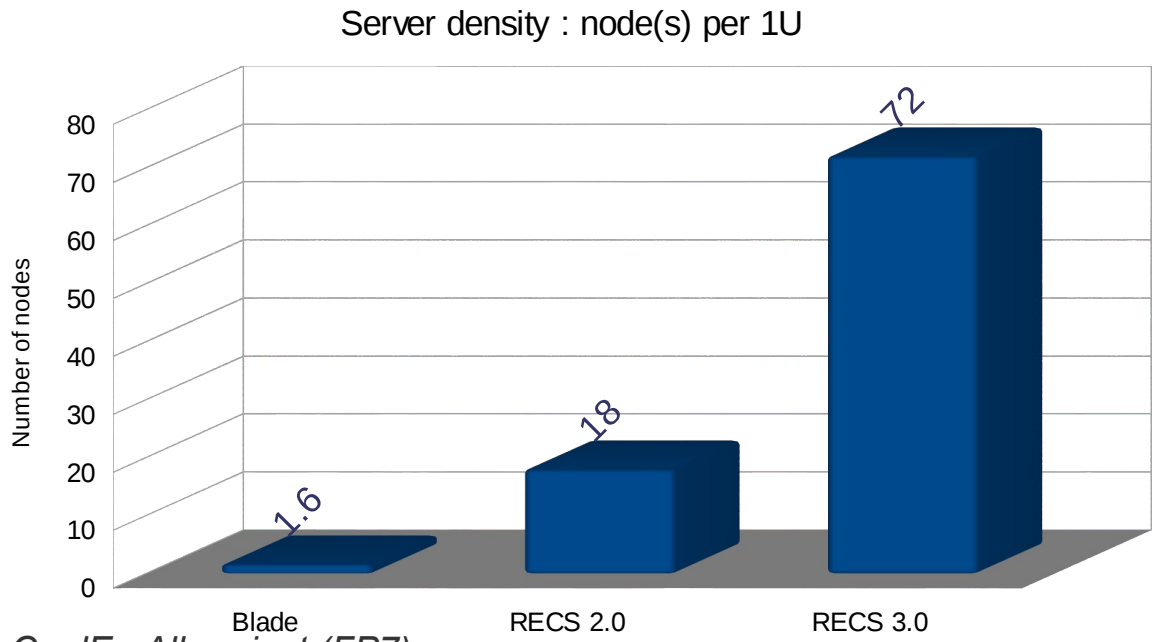
RECS : extreme density servers from Christmann Gmbh*

RECS 2.0 : upto 18 nodes* (Atom, i7, ARM ...) to fit in a standard 1U rack !

***COM-Express boards**

And RECS 3.0 : upto 72 ULP-COM boards!

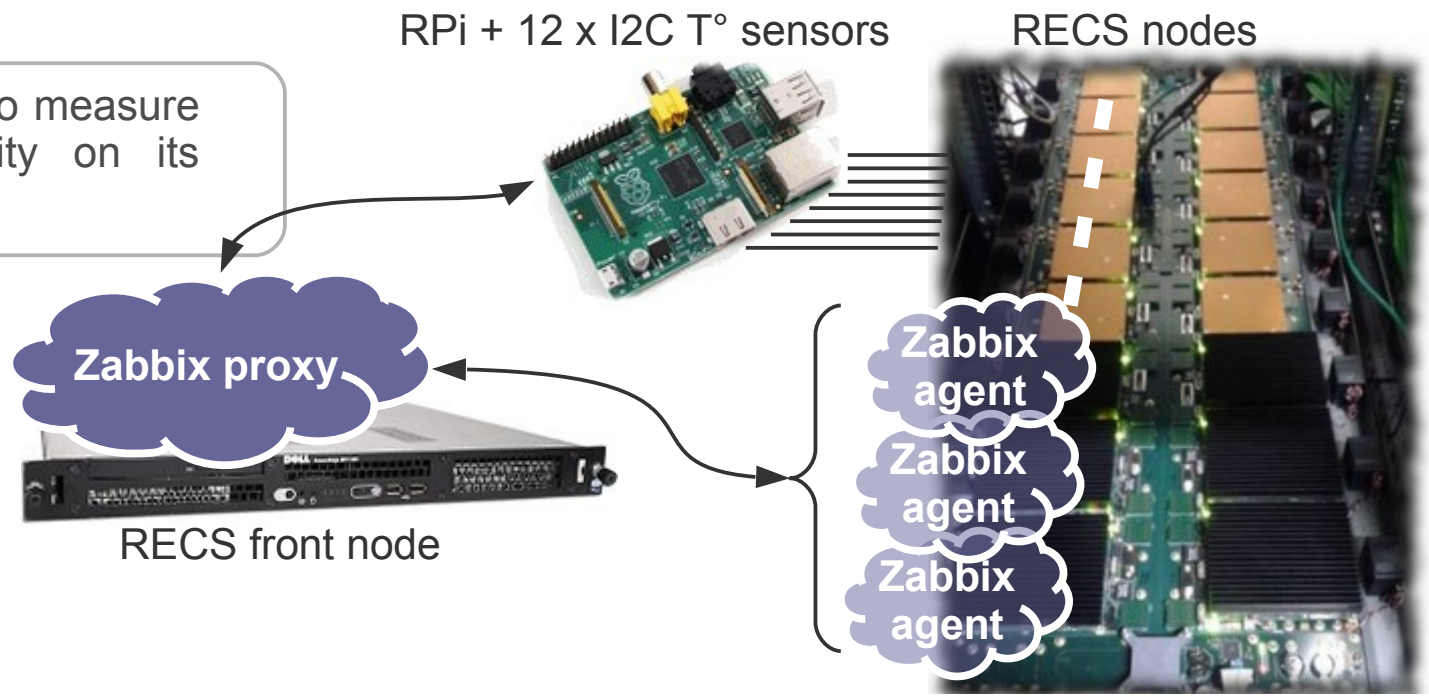
CHRISTMANN
INFORMATIONSTECHNIK + MEDIEN

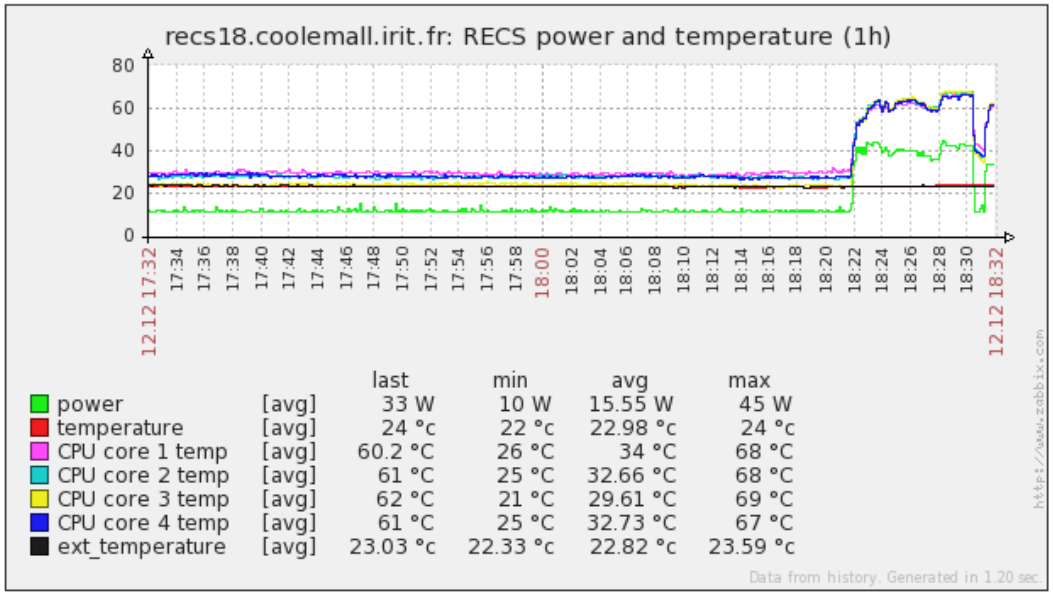


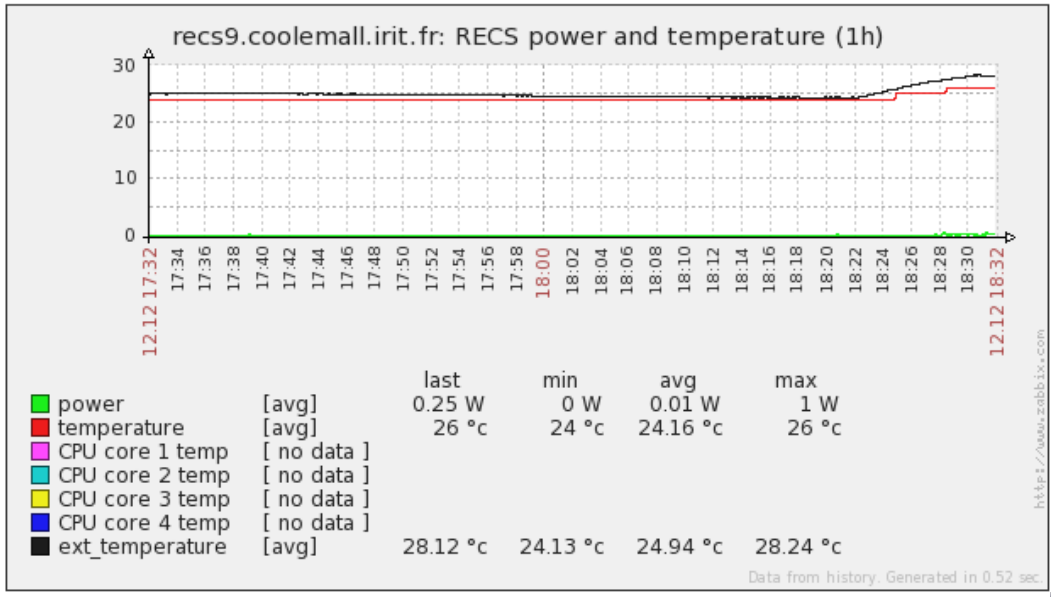
**http://www.christmann.info/* *Member of the CoolEmAll project (FP7)*

RECS temperature monitoring

Added temperature sensors to measure impact of one node activity on its neighbours.







RECS 3.0

Up to 72 x ULP-COM* modules per 1U rack!

- E.g. 72 x Apalis T30 (Tegra 3 – 4 x ARM9, 2GB DRR3) ==> **288** cores / 1U!

**now standardized as SMARC format.*

Cebit 2013 - http://www.silicon.de/41580841/cebit-christman-packt-72-arm-server-in-ein-u/

Plan

- The SEPIA team,
- The CloudMIP platform | overview,
- Inside CloudMIP | network, monitoring …,
- GreenIT | power metering @ node/VM level, ongoing researches,
- CloudMIP use cases | FG-VMDirac, FG-Cloud challenge, others ...
- [R&D] The RECS platform (FP7 CoolEmAll) | overview,
- What's next ?

What's next ?

[CloudMIP] ongoing work ...

- Move to CentOS7 and OpenStack,
- Test for NFSROOT + deltas on hosts,
- Aggregate unsed nodes' storage to a shared FS (Ceph, glusterFS …),
- Setup a Virtual Desktop Infrastructure (VDI),
- Integreate VM power consumption to OpenStack,
- ...

Questions ?

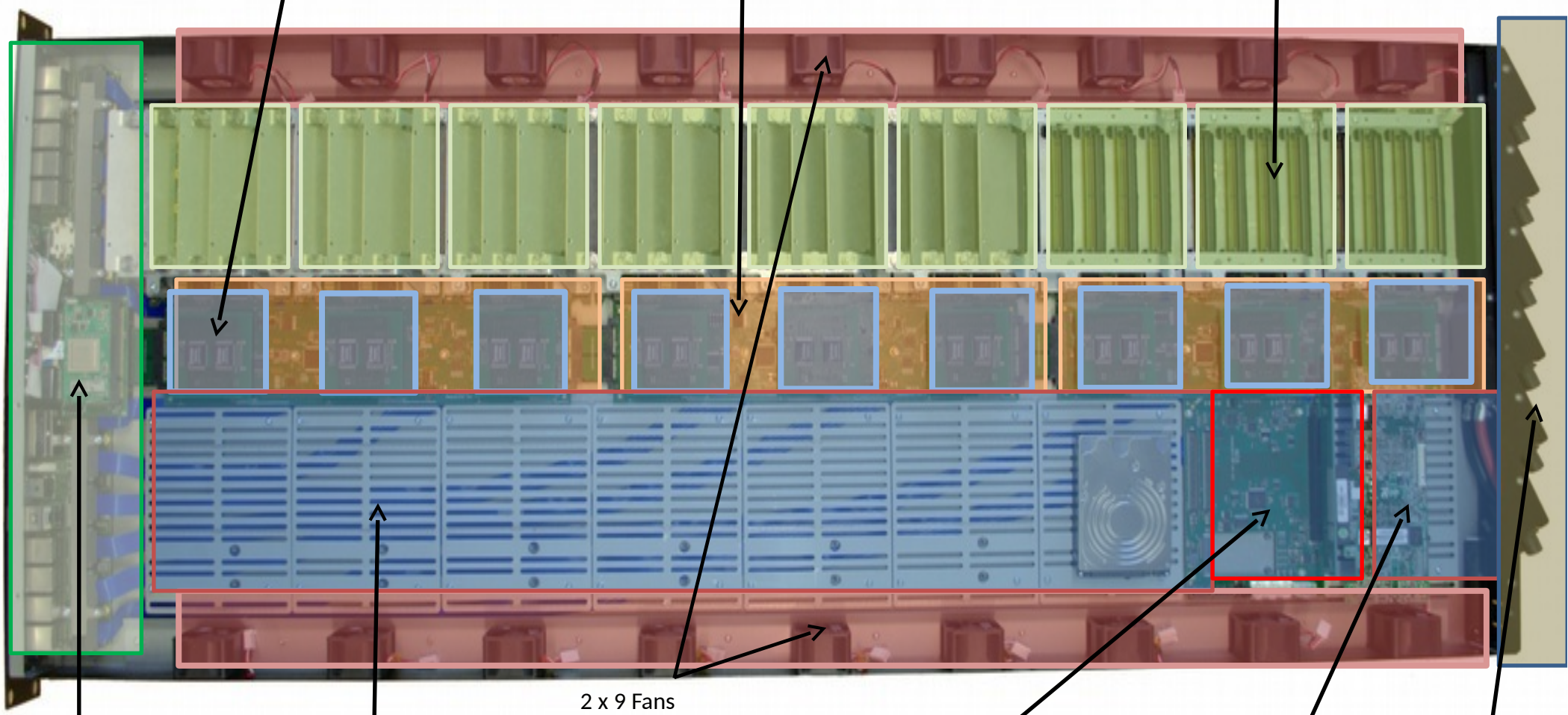
Want to collaborate with France-Grilles ? http://www.france-grilles.fr/

To benefit from our expertise, to access powerful platforms and to share yours in innovative academic / business workflows!

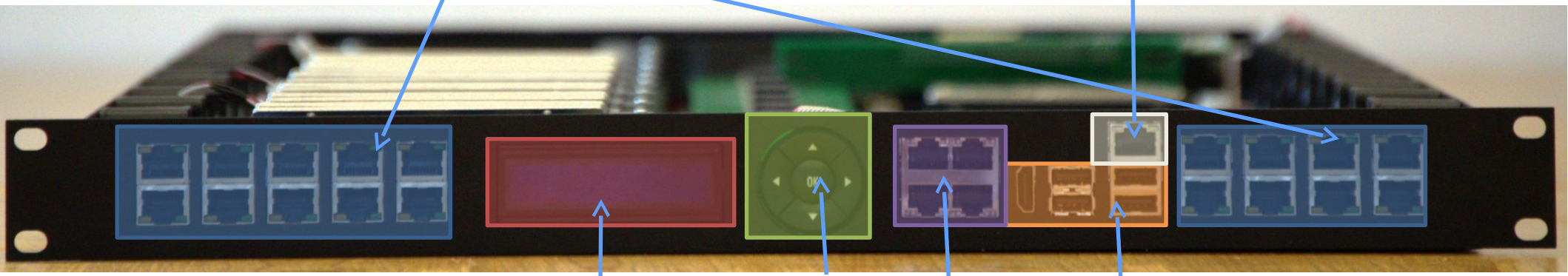
The END

END

RECS® | Box Compute Unit 3.0 (Codename Arneb)



RECS® | Box Compute Unit 3.0 (Codename Arneb)

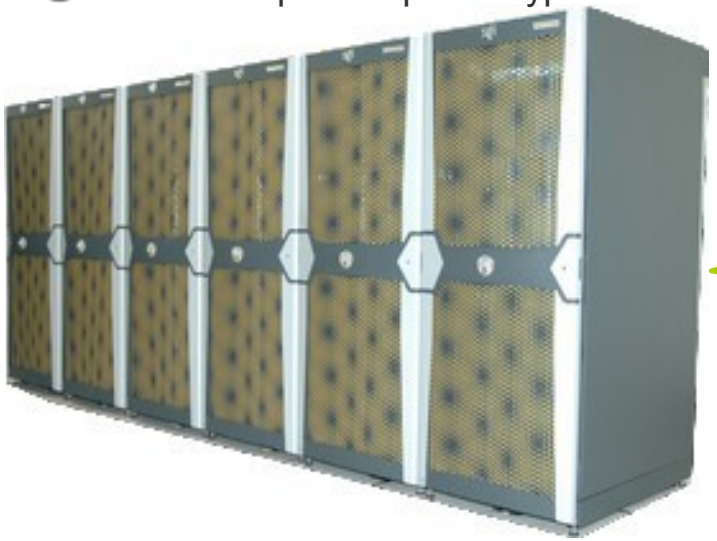


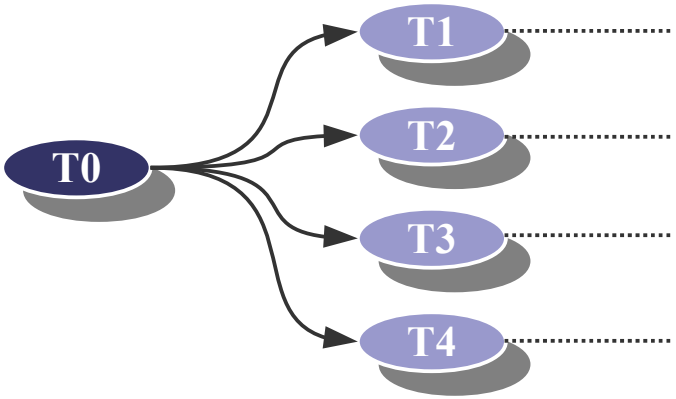
Embarrassingly parallel workload

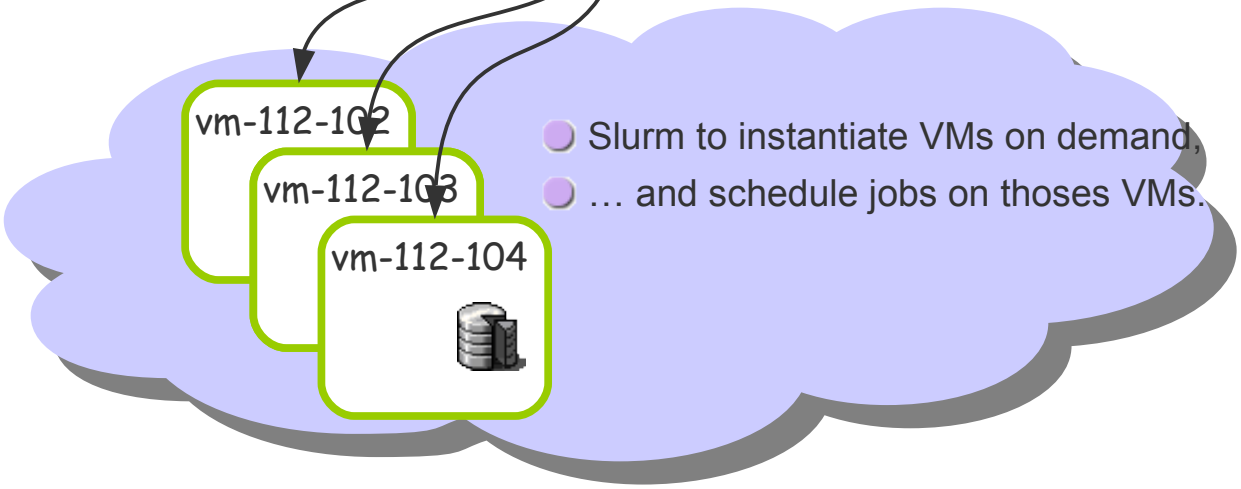
… a bit further …

- Offloading embarrassingly parallel workload from the Hyperion supercomputer to CloudMIP.

- CALMIP supercomputer Hyperion





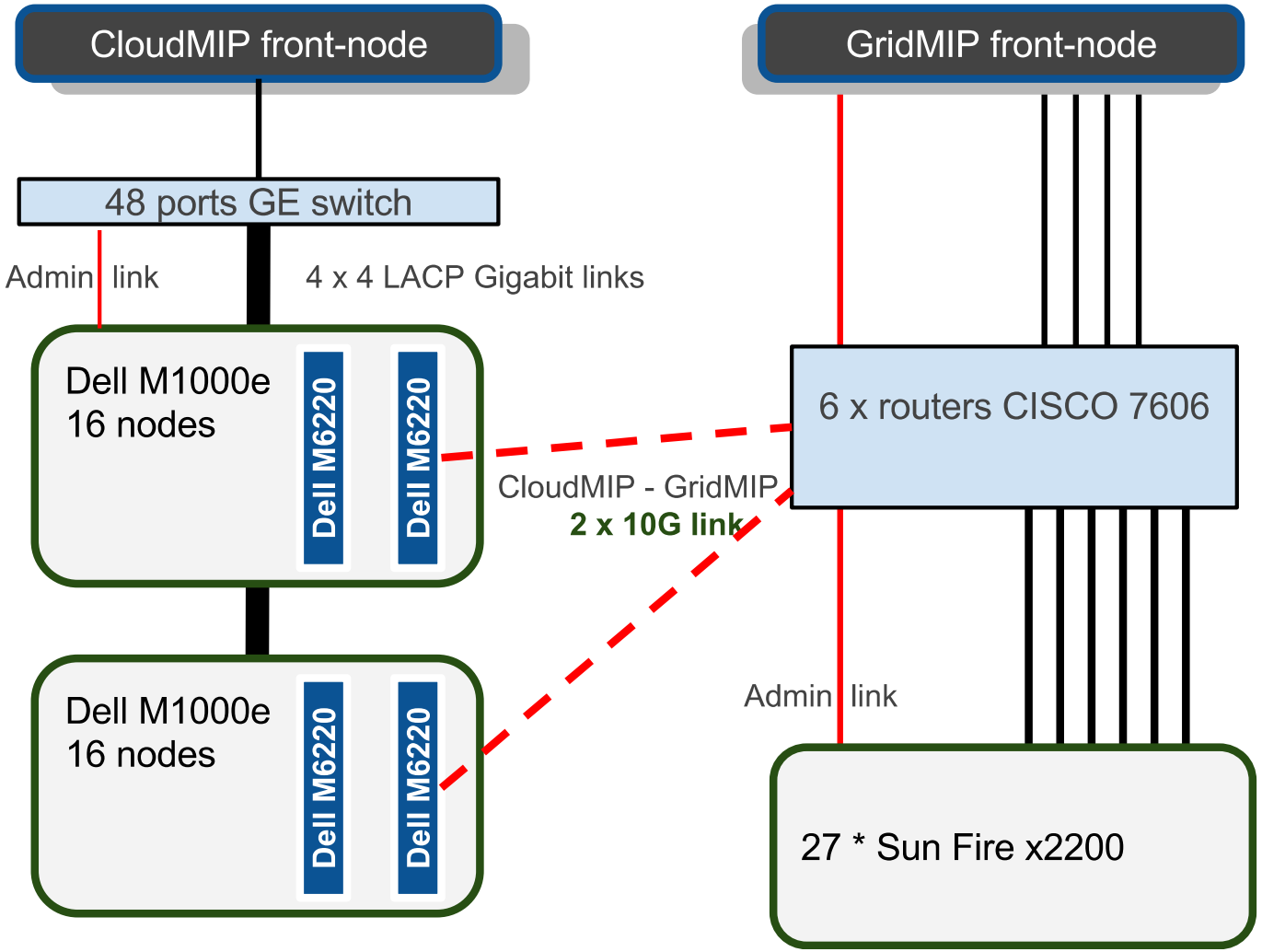


- Slurm to instantiate VMs on demand,
- … and schedule jobs on thoses VMs.

CloudMIP ⟺ GridMIP

GridMIP (network experiments platform) and CloudMIP **interconnect**

- 2 x 10G links on both GridMIP and CloudMIP backbones (est.cost in on way).



GreenIT: VM power metering

Near future ...

- Ectop : the per-process power consumption accounting (Leandro's work),
- Ectop calibration and integration in CloudMIP,
- Availability of a per-VM power metering through Zabbix,
- Integration of the per-VM power metering in Sunstone (OpenNebula GUI)
  → enables a per-user view of VMs along with their power consumption,
- VM power metering data availability in an Inter-Cloud configuration ?

Others stuff

Various improvements, new features etc ...

- **[High]** Cloud-Init (contextualisation) and CIMI api (DMTF VM mgt.) tests,
- **[High]** Improve platform visibility through the Wiki,
- **[Important]** to broaden the range of users,
- Improve Zabbix responsiveness through a migration from apache to nginx,
- Enable VNC direct connection to VMs through Sunstone,
- Add template-based SPICE connexion to a VM (re. IPtables hooks in ONE),
- Add template-based Public IP redirect to a VM,
- Activate VOMS authentication (Inter-Cloud France-Grilles),
- VMDirac plug-ins integration to OpenNebula,
- [Low] Upgrade to new SL65.