Introduction to Grid'5000 (G5k)

Overview

21 octobre 2021 @CargoDay-12, Rennes

richard.randriatoamanana-at-ls2n.fr

Why do experiments[1] ?

"Beware of bugs in the above code;
I have only proved it correct, not tried it"
(Donald Knuth)

"In theory there is no difference between theory and practice. In practice there is."
(Yogi Berra)

[1] *Extract from a talk at NSFCloud in 2014 by Kate Keahey (Argonne Nat. Lab.)*

Why ?

IT Resources for Research

Carrying out "experiments" is essential in computer science today and “good experiments”[1] should fulfill the following properties.

- **Reproducibility** : same result with same input
- **Extensibility** : target comparaisons with other works
- **Applicability** : define realistic params (easy calibration, ..)
- **“Revisability”** : help to identity the reasons (object of study)

[1] *Inspired from a talk at SILECS School in 2018 given by F. Desprez (INRIA)*

What ?

Grid'5000 | Overview

- A national scientific intrument with a reconfigurable testbed infrastructure for experimental research on computer science targeting and tackling large-scale domains

  Big Compute (parallel and distributed systems – Cloud, HTC, HPC), Big Data, Datacenters, High Performance Networking.

- But it's not a grid but "Bare Metal as a servce"
- GIS created in 2012 but 15 years already...
  - a very active community (researchers, engineers, techs)
  - ±600 active users and ~120 publications per year
  - ±60 millions core hours used in 2019

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION
Liberté Égalité Fraternité

cnrs dépasser les frontières

Inria INVENTEURS DU MONDE NUMÉRIQUE

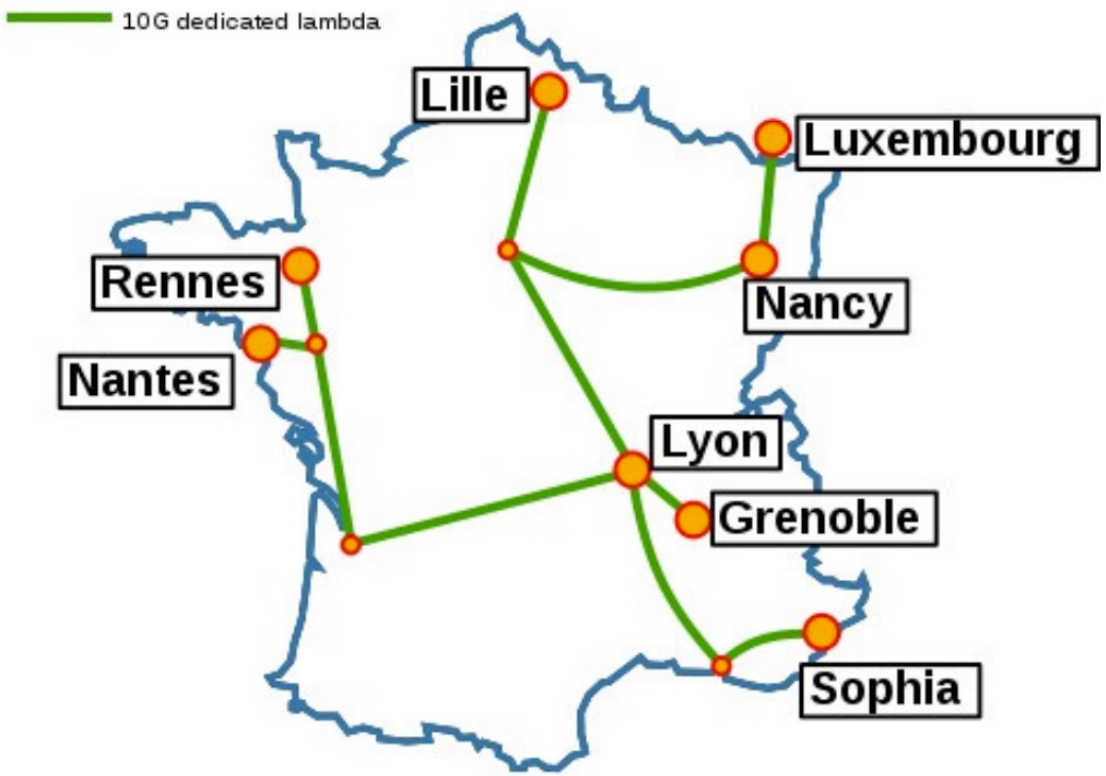
RENATER CONNECTEUR DE SAVOIRS

cat.opidor.fr/index.php/Grid%275000

Where ?

Grid'5000 | Key Features[1]

- 8 sites, 39 clusters, ±800 nodes,
- ±16000 CPU cores and ±300 GPU
- ±100 TiB RAM + 6 TiB PMEM
- R_{peak} 614.3 TFLOPS (excluding GPUs)
- 511 SSDs and 1004 HDDs on nodes (total: 1.44 PB)
- Dedicated 10-Gbps backbone network



[1] Source : https://www.grid5000.fr/w/Hardware

Where ?

Grid'5000 | Resources @ Nantes site[1]

Site	Cluster	Access Condition	Date of arrival	Nodes	CPU	Cores	Memory	Storage	Network
Sophia	uvb		2011-01-04	30	2 x Intel Xeon X5670	6 cores/CPU	96 GiB	**250 GB HDD**	1 Gbps (SR-IOV) + 40 Gbps InfiniBand
Rennes	paranoia		2014-02-21	8	2 x Intel Xeon E5-2660 v2	10 cores/CPU	128 GiB	**1 x 600 GB HDD** + 4 x 600 GB HDD	1 Gbps (SR-IOV) + 2 x 10 Gbps (SR-IOV)
Rennes	parapide		2010-01-25	17	2 x Intel Xeon X5570	4 cores/CPU	24 GiB	**500 GB HDD**	1 Gbps + 20 Gbps InfiniBand
Rennes	parapluie		2010-11-02	16	2 x AMD Opteron 6164 HE	12 cores/CPU	48 GiB	**250 GB HDD**	1 Gbps + 20 Gbps InfiniBand
Rennes	parasilo		2015-01-13	27	2 x Intel Xeon E5-2630 v3	8 cores/CPU	128 GiB	**600 GB HDD** + 4 x 600 GB HDD* + 200 GB SSD*	2 x 10 Gbps (SR-IOV)
Rennes	paravance		2015-01-13	72	2 x Intel Xeon E5-2630 v3	8 cores/CPU	128 GiB	**1 x 600 GB HDD** + 1 x 600 GB HDD	2 x 10 Gbps (SR-IOV)
Nantes	econome		2014-04-16	22	2 x Intel Xeon E5-2660	8 cores/CPU	64 GiB	**2.0 TB HDD**	10 Gbps (SR-IOV)
Nantes	ecotype		2017-10-16	48	2 x Intel Xeon E5-2630L v4	10 cores/CPU	128 GiB	**400 GB SSD**	2 x 10 Gbps (SR-IOV)
Nancy	graffiti	**production** queue	2019-06-07	13	2 x Intel Xeon Silver 4110	8 cores/CPU	128 GiB	**479 GB HDD**	10 Gbps

1312 cores
7.552 GiB Mem
±64 TB (dont 19TB SSD)

- `econome` {Dell PE C6220}
- `ecotype` {Dell PE R630}

Processors counts per families

Processor family	Grenoble	Lille	Luxembourg	Lyon	Nancy	Nantes	Rennes	Sophia	Processors total
AMD EPYC		16		10	14				**40**
AMD Opteron				28			32		**60**
Intel Xeon	88	62	28	92	612	140	248	60	**1330**
POWER8NVL	24								**24**
ThunderX2				8					**8**
Sites total	**112**	**78**	**28**	**138**	**626**	**140**	**280**	**60**	**1462**

Accelerator cores

Accelerator model
AMD Radeon Instinct MI50 32GB
Intel Xeon Phi 7120P
Nvidia A100-PCIE-40GB
Nvidia GeForce GTX 1080 Ti
Nvidia GeForce GTX 980
Nvidia GeForce RTX 2080 Ti
Nvidia Quadro RTX 6000
Nvidia Tesla K40m
Nvidia Tesla M2075
Nvidia Tesla P100-PCIE-16GB
Nvidia Tesla P100-SXM2-16GB
Nvidia Tesla T4
Nvidia Tesla V100-PCIE-32GB
Nvidia Tesla V100-SXM2-32GB

[1] *Source : https://www.grid5000.fr/w/Nantes:Hardware*

How ?

Grid'5000 | An experiment's outline

"reserve your physical server resource on-fly"

- Discovering resources, selecting resources and submitting jobs
- Reconfiguring the resources to meet experimental needs
- Monitoring experiments by extracting and analyzing data
- Controlling experiments, automation, reproducible research

How ?

Grid'5000 | Software Stack[1]

- Isolated network, access using **SSH**

- Tasks/Resources Management: **OAR**
- System Reconfiguration: **Kadeploy**
- Network Configuration: **Kavlan**
- Monitoring: **Kaspied, Kwapi, Kwollect (grafana), OAR/{Monika,DrawGantt}** ...

oar.imag.fr

kadeploy.gitlabpages.inria.fr

- All in One: **Grid'5000 API**

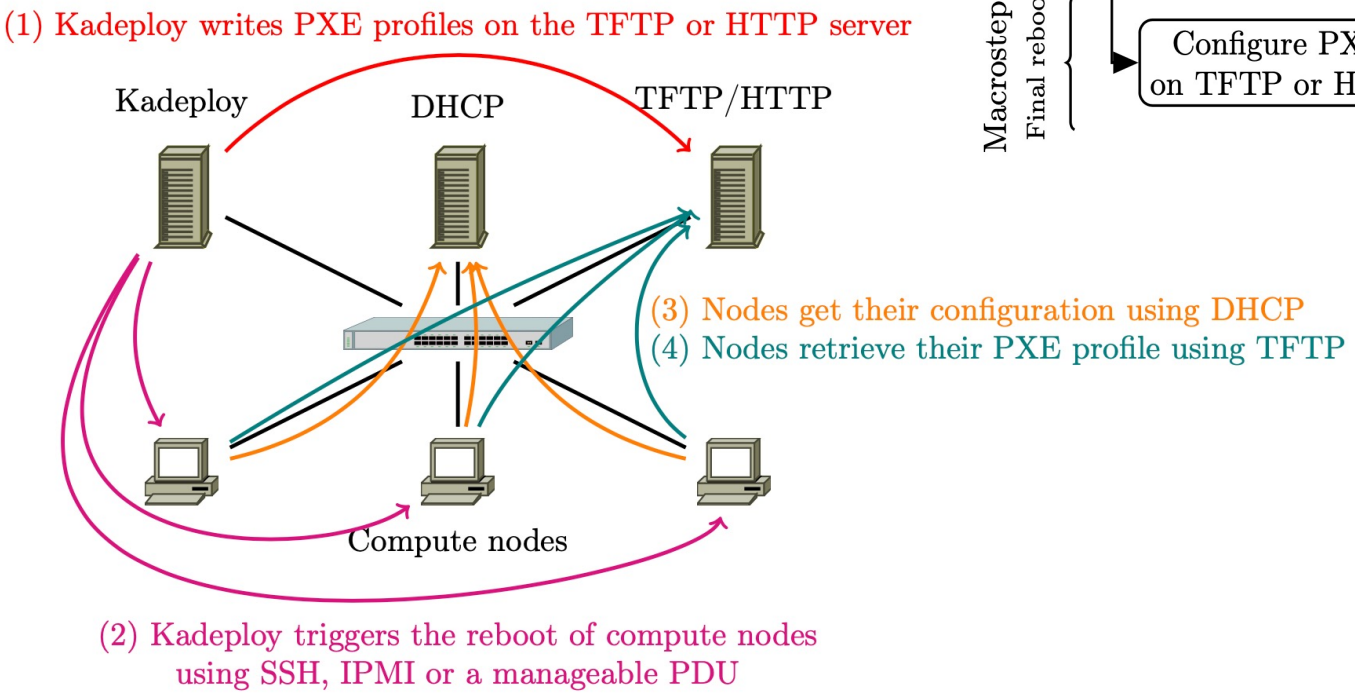
[1] Source: https://www.grid5000.fr/w/Getting_Started

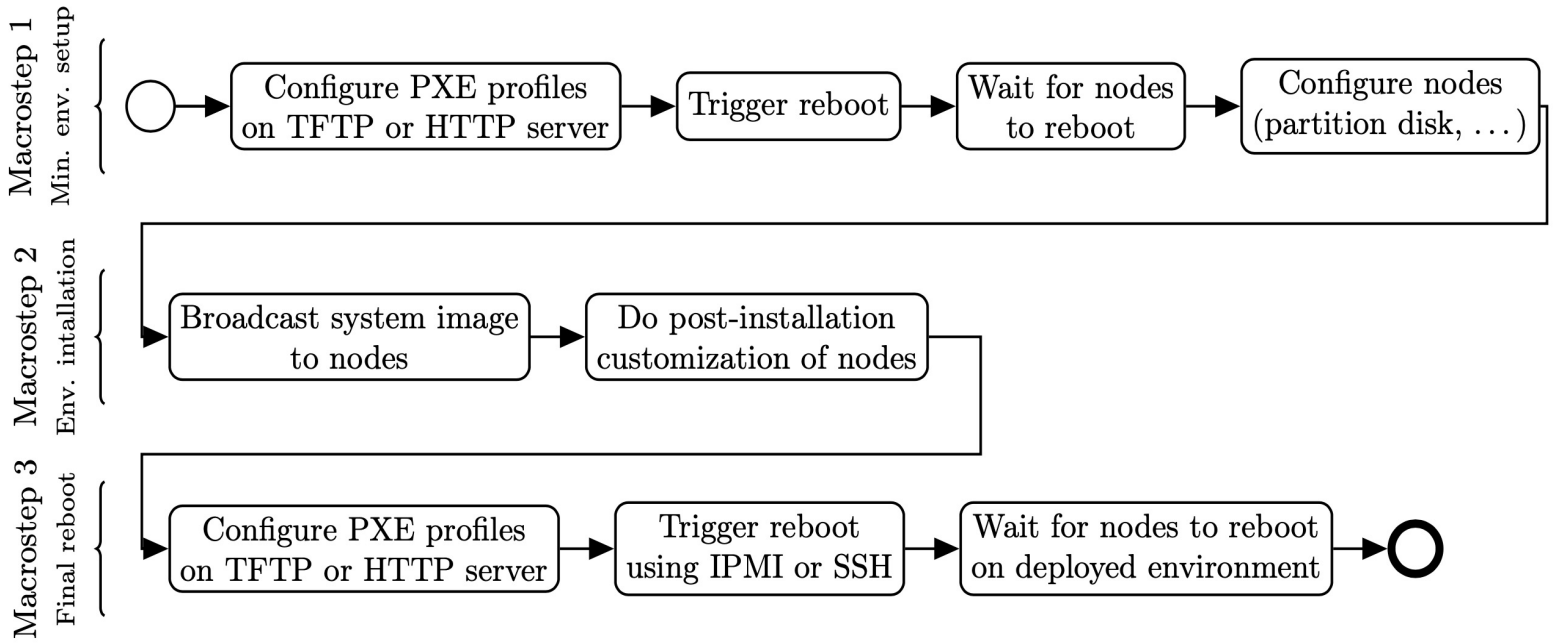
https://hal.inria.fr/hal-00710638
https://www.grid5000.fr/w/Advanced_Kadeploy

Kadeploy3

https://gitlab.inria.fr/grid5000/kadeploy

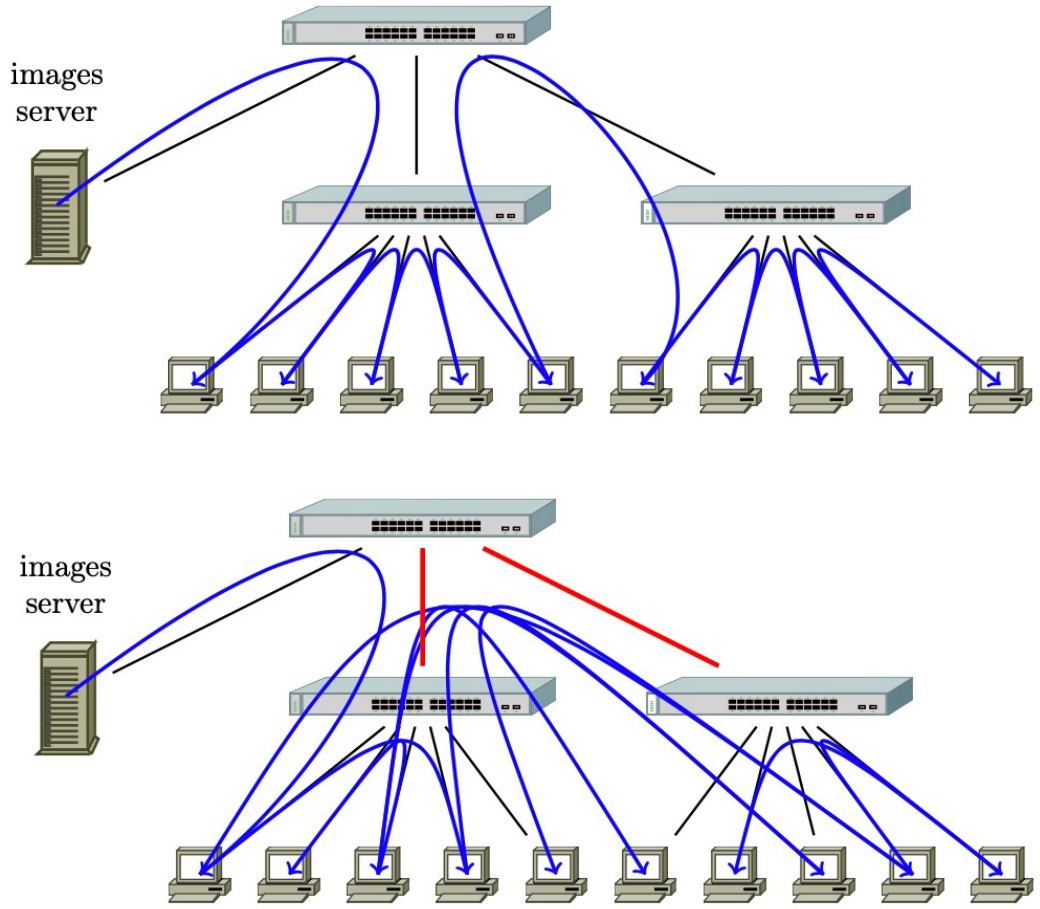
A scalable, efficient and reliable deployment system (cluster provisioning solution) for cluster and grid computing on OS like *Linux, *BSD, Windows or Solaris*.





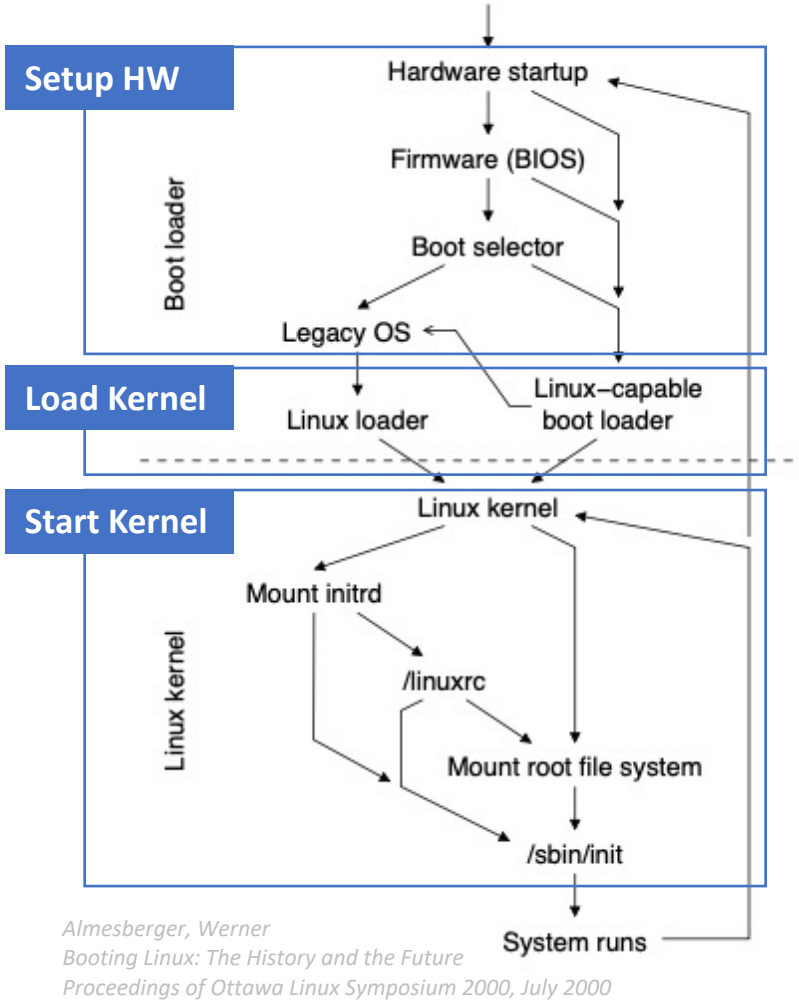
1. **Minimal environment setup** . The nodes reboot into a trusted minimal environment that contains all the tools required for the deployment (partitioning tools, archive management,...) and the required partitioning is performed.
2. **Environment installation** . The environment is broadcast to all the nodes and extracted on the disks. Some post-installations operations can also be performed.
3. **Reboot** on the deployed environment.

Built for scalability



- TakTuk : a model of hierarchical connection for parallel and execution and reporting
- Scalable file distribution approaches :
  - tree-based, chain-based and BitTorrent-based
- Windowed operations
  - Loop of “100 reboots & wait 10”

Boot Sequence Matters…



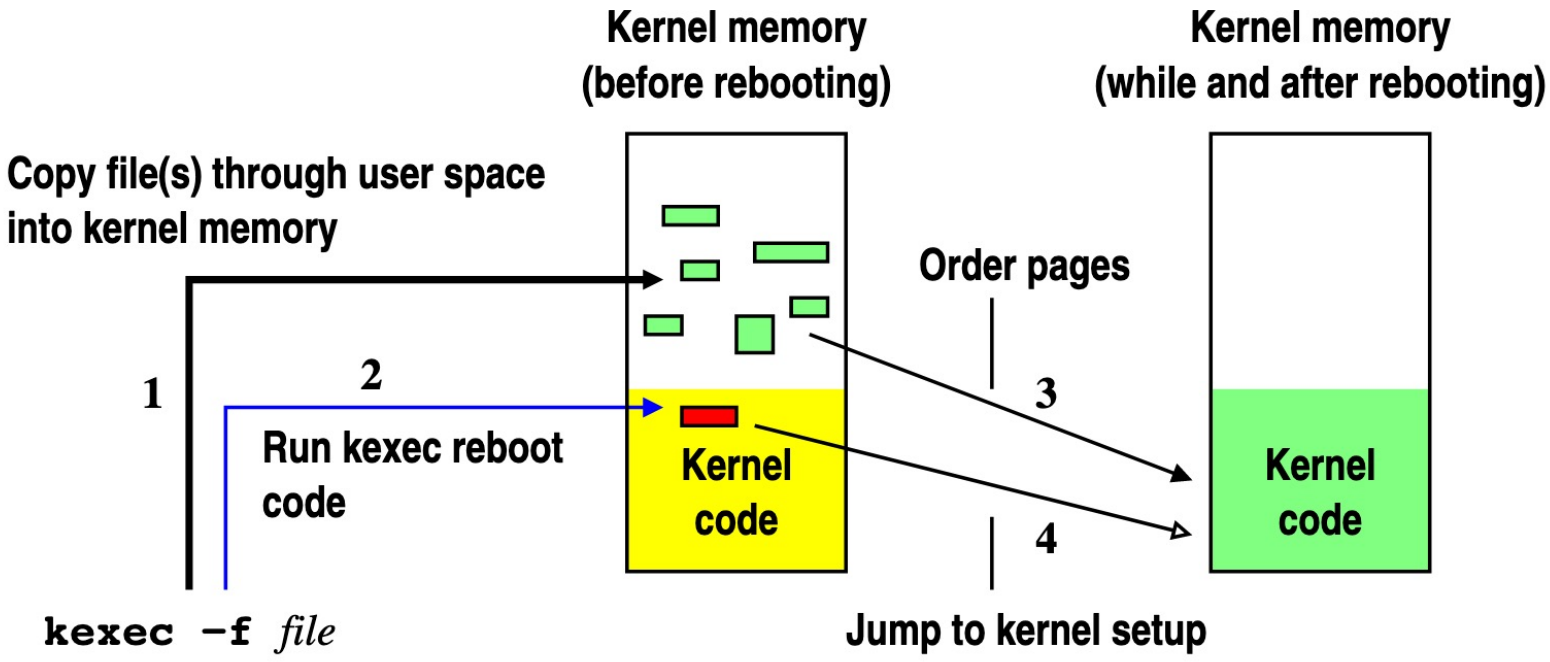
Almesberger, Werner
Booting Linux: The History and the Future
Proceedings of Ottawa Linux Symposium 2000, July 2000

By Eric Biederman.

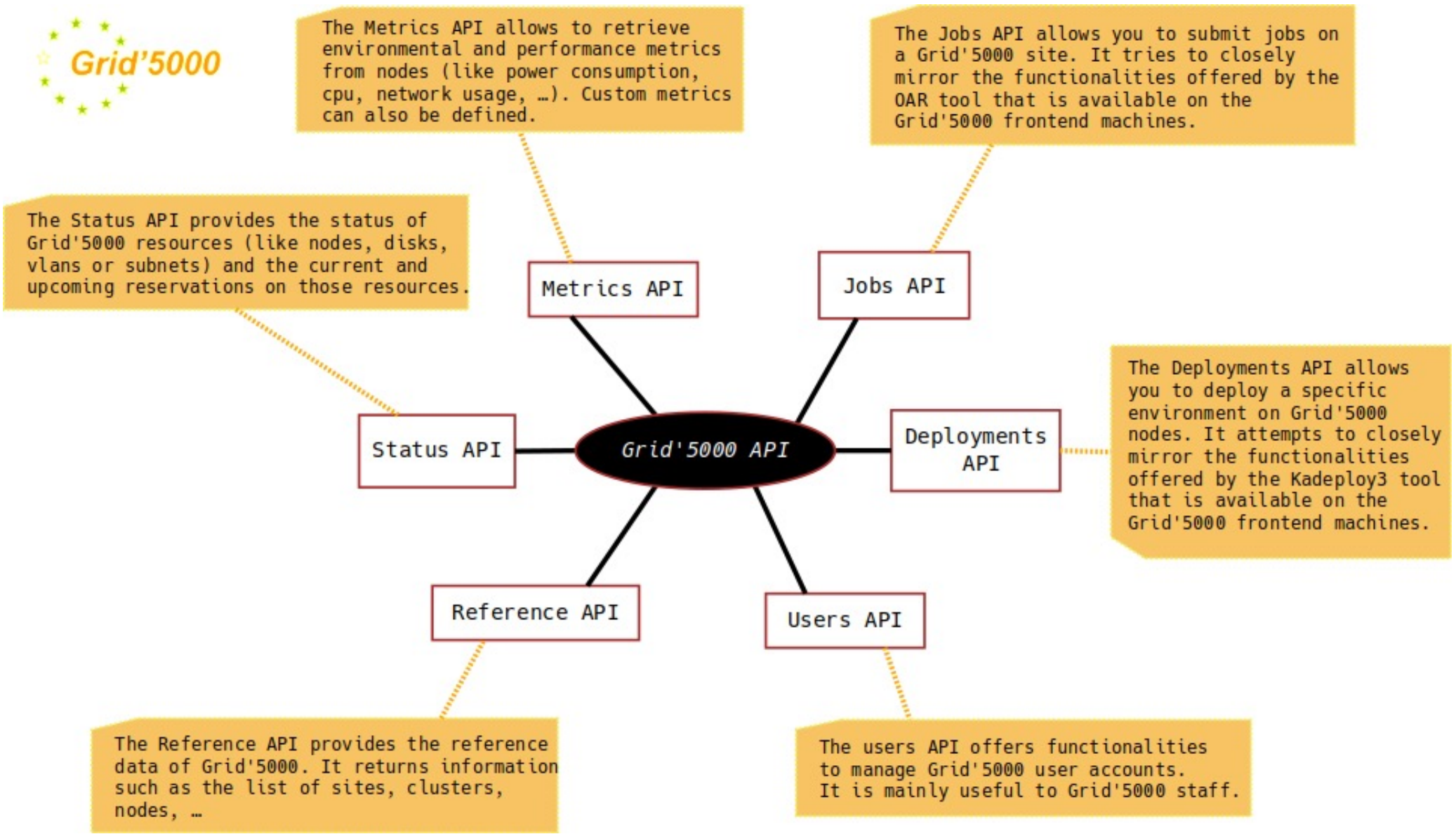
kexec

"a system call that implements the ability to shutdown your current kernel, and to start another kernel. It is like a reboot but it is indepedent of the system firmware. And like a reboot the you can start any kernel with it not just Linux."

Configuration help text in Linux-2.6.17



API


Grid'5000
Grid'5000 API
Metrics API
The Metrics API allows to retrieve environmental and performance metrics from nodes (like power consumption, cpu, network usage, …). Custom metrics can also be defined.
Jobs API
The Jobs API allows you to submit jobs on a Grid'5000 site. It tries to closely mirror the functionalities offered by the OAR tool that is available on the Grid'5000 frontend machines.
Status API
The Status API provides the status of Grid'5000 resources (like nodes, disks, vlans or subnets) and the current and upcoming reservations on those resources.
Deployments API
The Deployments API allows you to deploy a specific environment on Grid'5000 nodes. It attempts to closely mirror the functionalities offered by the Kadeploy3 tool that is available on the Grid'5000 frontend machines.
Reference API
The Reference API provides the reference data of Grid'5000. It returns information such as the list of sites, clusters, nodes, …
Users API
The users API offers functionalities to manage Grid'5000 user accounts. It is mainly useful to Grid'5000 staff.

How ?

Grid'5000 | Demo time[1] !

https://gitlab.in2p3.fr/resinfo-cargo/cargoday-12

[1] Source: https://www.grid5000.fr/w/Getting_Started