Grid'5000 :
Administration d'une infrastructure distribuée et développement d'outils de déploiement et d'isolation réseau

Nicolas Niclausse - INRIA Sophia Antipolis Méditerranée - projet Aladdin Grid'5000

2 juillet 2012

Grid'5000

Présentation

Une plate-forme expérimentale pour la recherche sur les systèmes distribués et parallèles

- 7400 cœurs
- 1600 noeuds de calcul
- 26 clusters
- 10 sites en France
- 1 site au Luxembourg
- backbone 10Gbps (Renater)

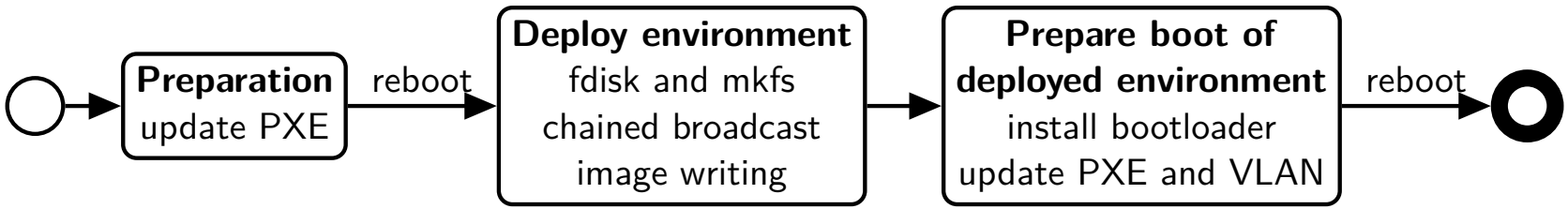
Grid'5000

Présentation

- Début du projet en 2003 (ACI GRID ; INRIA, CNRS et Universités)
- Premiers utilisateurs début 2005
- 1000 machines disponibles dès 2006
- Environ 500 utilisateurs chaque année (chercheurs, doctorants, ...)
- Système de réservation de ressources basé sur OAR (GPL)
- Ressources de type machines (cores), vlans, stockage, sous réseaux, ...
- L'utilisateur peux installer l'OS de son choix sur ses machines réservés, en quelques minutes.

Kadeploy – outil de déploiement de cluster scalable

- Fournit une infrastructure de type Cloud *Hardware-as-a-Service*
- Construit au dessus de PXE, DHCP, TFTP
- **Scalable, efficace, robuste et flexible** :
  - Broadcast d'environnement basé sur une chaîne ou sur BitTorrent
  - **255 machines déployées en 7 minutes**
- Support de plusieurs OS (Linux, Xen, *BSD, etc.)
- Interface en ligne de commande & par une API REST



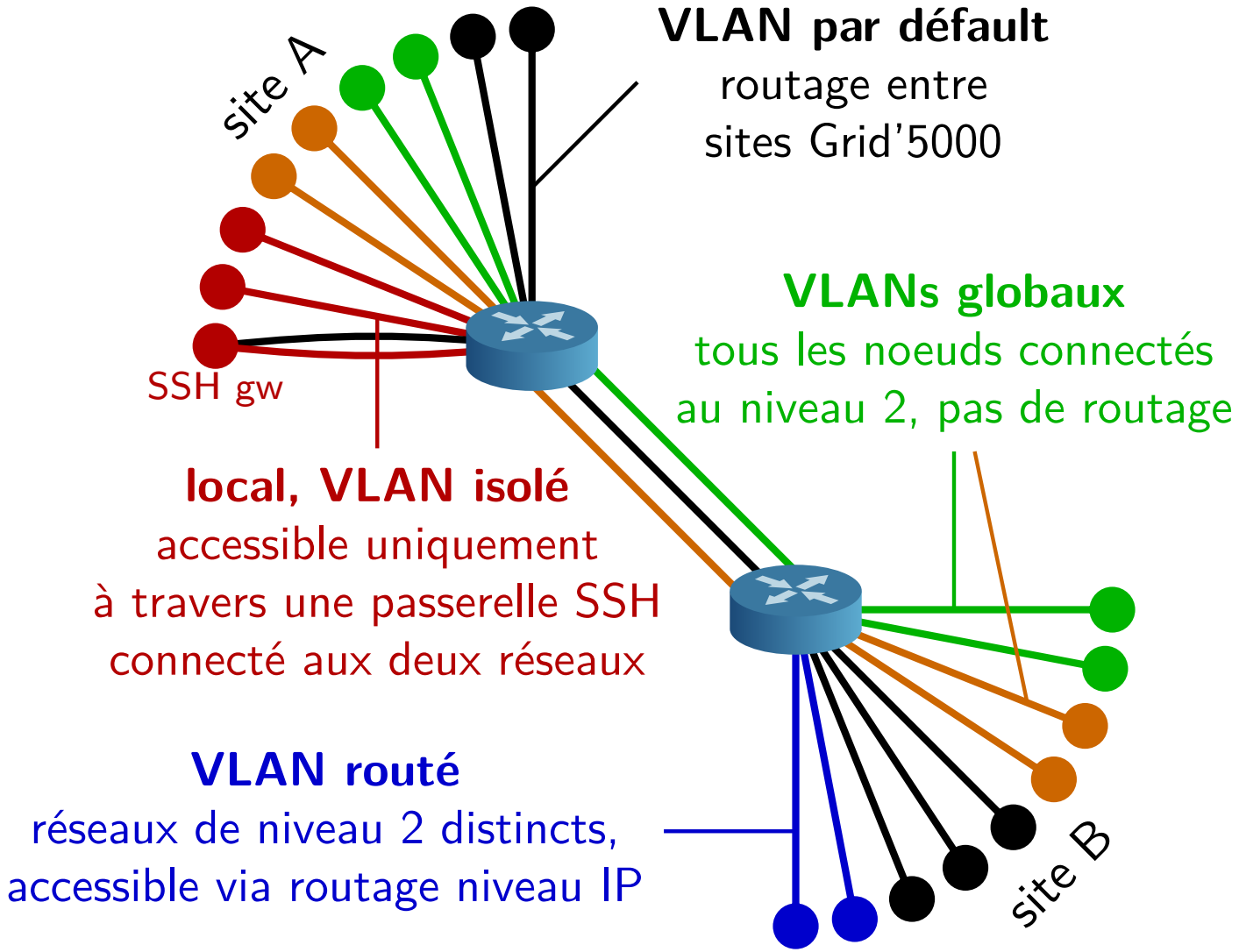
- Développé en Ruby ; Utilisation de Taktuk comme lanceur ssh et la diffusion des images.
- Disponible en licence CeCILL (GPL)
- Utilisation intensive de IPMI (reboot, console série). Assez peu fiable avec certains constructeurs : plusieurs milliers de reboots par jour $\rightsquigarrow$ gros soucis
- Testé à l'intérieur de Grid'5000 (déploiement de 3800 VM)

`http://kadeploy3.gforge.inria.fr/`

Isolation Réseau : KaVLAN

- Reconfigure les switches/routeurs pour la durée d'une expérience d'un utilisateur afin d'obtenir une **isolation réseau complète au niveau 2** :
  - Évite le pollution réseau (broadcast, connections non sollicitées)
  - Permet aux utilisateur de configurer leur propre serveur DHCP
  - Expérimentation sur les protocoles basés sur Ethernet
  - Interconnexion de machines provenant d'un autre testbed sans compromettre la sécurité de Grid'5000
- Basé sur **802.1q (VLANs)**
- Compatible avec de nombreux équipements réseau
  - Utilisation de SNMP, SSH ou telnet pour la connection aux switches
  - Support de Cisco, HP, 3Com, Extreme Networks et Brocade
- Contrôlé via la ligne de commande ou une API REST

KaVLAN - différents types de VLAN



VLAN par défaut
routage entre
sites Grid'5000

VLANs globaux
tous les noeuds connectés
au niveau 2, pas de routage

local, VLAN isolé
accessible uniquement
à travers une passerelle SSH
connecté aux deux réseaux

VLAN routé
réseaux de niveau 2 distincts,
accessible via routage niveau IP

Grid'5000 API

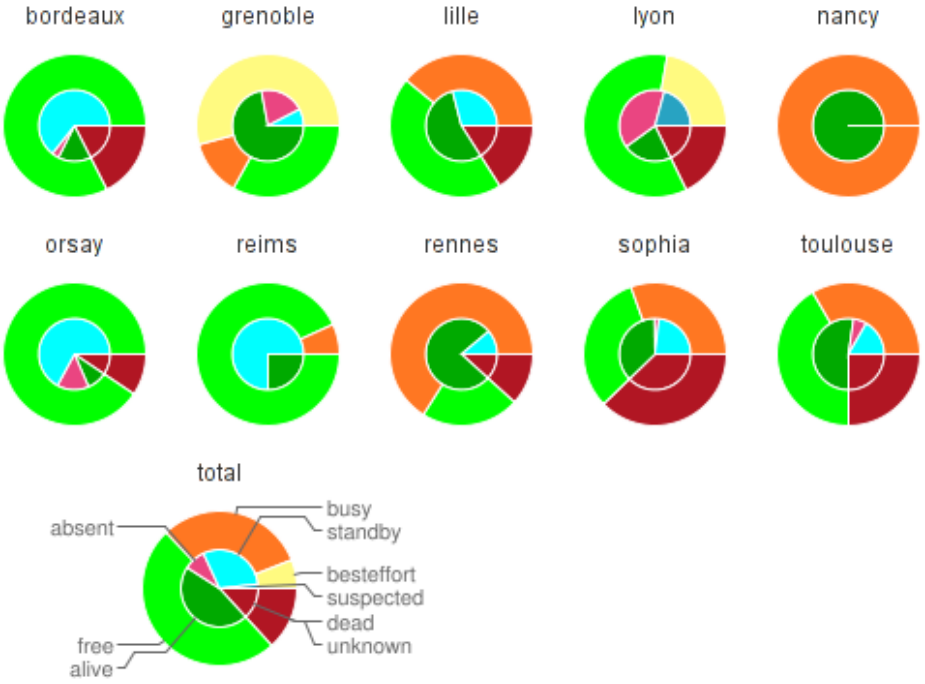
- API REST : une API pour chaque service de Grid'5000 :
  - **Reference API** : description versionnée des ressources Grid'5000
  - **Monitoring API** : état des ressources Grid'5000
  - **Metrology API** : données Ganglia (CPU, load, mémoire, etc.)
  - **Jobs API** : interface OAR
  - **Deployments API** : interface Kadeploy
  - **VLAN API** : interface KaVLAN
  - ...

Grid'5000 API : https://api.grid5000.fr/

Dashboard | Quick Start | Jobs | Metrics | Visualizations | Events | Help

Dashboard

Grid Status



Latest News

APIs on Grenoble are back. But it is still impossible to submit jobs due to local issues. #apis
Thu Feb 03 13:50:04 +0000 2011 #incident, #apis

APIs on Grenoble site are unreachable due to local maintenance #apis
Thu Feb 03 11:48:35 +0000 2011 #incident, #apis

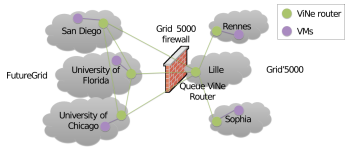
Jabber notifications are back. Sorry for the extended downtime.
Mon Oct 04 11:38:05 +0000 2010 #apis

The jabber notifications are currently down. Investigating.
Thu Sep 30 15:53:50 +0000 2010 #apis

Screencast to demo the new UI of the sid version: https://www.grid5000.fr/pub/screencasts/grid5000-web-portal.mov
Wed Sep 08 13:29:07 +0000 2010 #apis

Des résultats dans différents domaines

Cloud : Sky computing sur FutureGrid and Grid'5000



- cloud Nimbus deployé sur 450+ machines
- Grid'5000 et FutureGrid connectés via ViNe

HPC : factorisation de RSA-768

- étude de faisabilité : prouver qu'on peut le faire
- Variété de hardware $\rightsquigarrow$ comprendre les performances des algorithmes

Grid : évaluation du middleware de grille gLite

- Déploiement et configuration entièrement automatisé sur 1000 machines (9 sites, 17 clusters)

Virtualisation :

- 10240 machines virtuelles déployées sur 512 machines (4 sites)

Réseau P2P :

- 10000 clients BitTorrent instanciés sur 178 machines (analyse de performance du protocole BitTorrent)

Tendance :

- De plus en plus d'expériences sur les thèmes Cloud et BigData (Nimbus, OpenNebula, OpenStack, Hadoop, ...)

Administration de Grid'5000

Historique & contraintes

Historique (2003)

- Mutualisation de sites indépendants, un sysadmin par site
- Hétérogénéité des systèmes, versions logiciels, configurations...

Nouvelles contraintes (2007 à 2011)

- Séparation de l'équipe : sysadmins / développeurs logiciel
- 5 à 6 administrateurs systèmes jeunes ingénieurs sur des contrats de 1 à 2 ans
- Traçabilité des opérations d'administration
- Traçabilité de l'évolution de la plate-forme pour garder l'expertise
- Ajout de nouveaux sites Grid'5000

Solution mise en place

Gestion de configuration des services

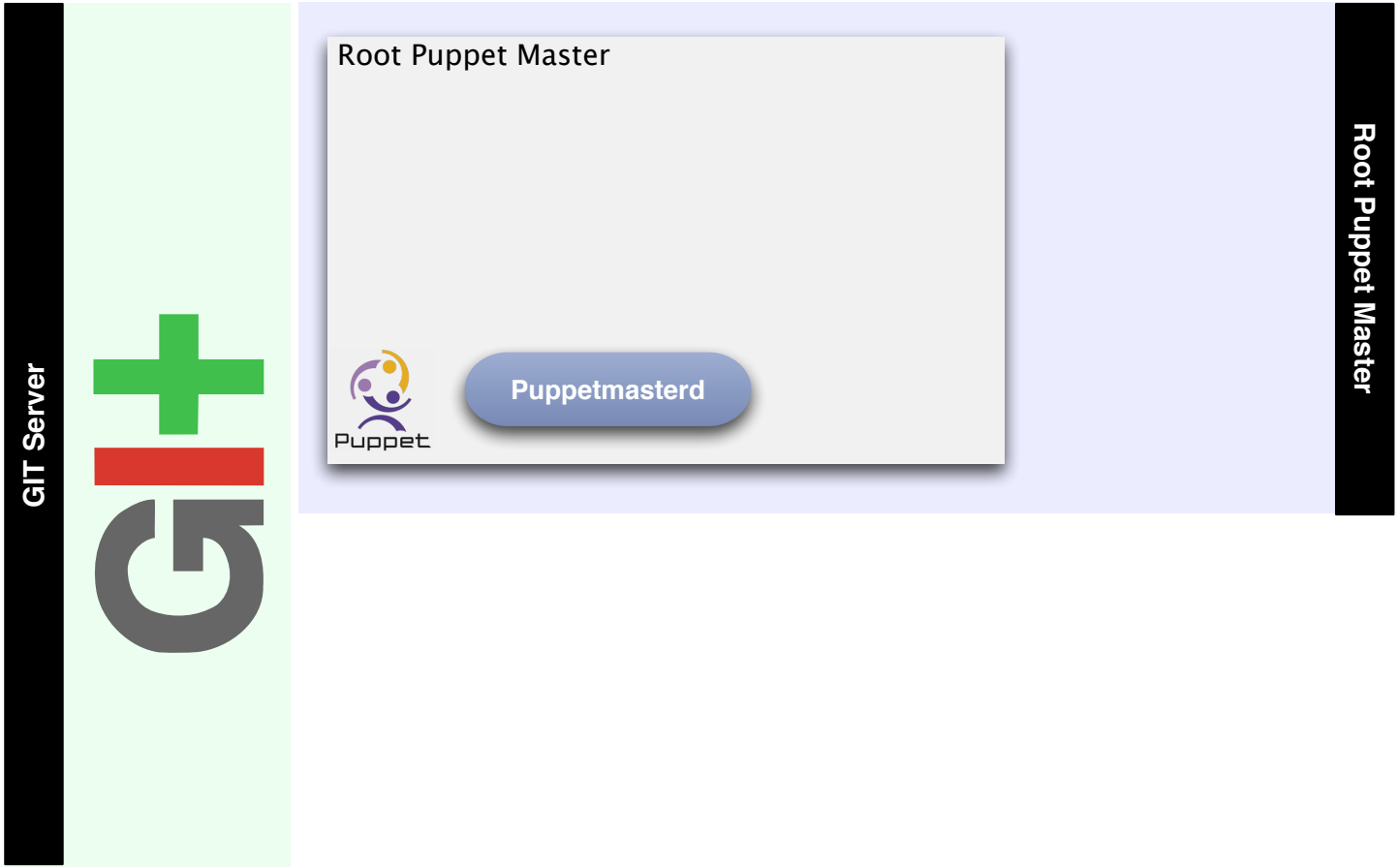
- Une VM (Xen) par service (LDAP, DHCP, DNS, Nagios, Kadeploy, OAR, MySQL, KaVLAN, ganglia, Apache, ...)
- Utilisation de puppet, git et capistrano pour la configuration
- 350 VM réparties sur les 10 sites
- Utilisation de chef pour la génération des images des OS destinées aux utilisateurs (debian)

Solution mise en place

Architecture Puppet

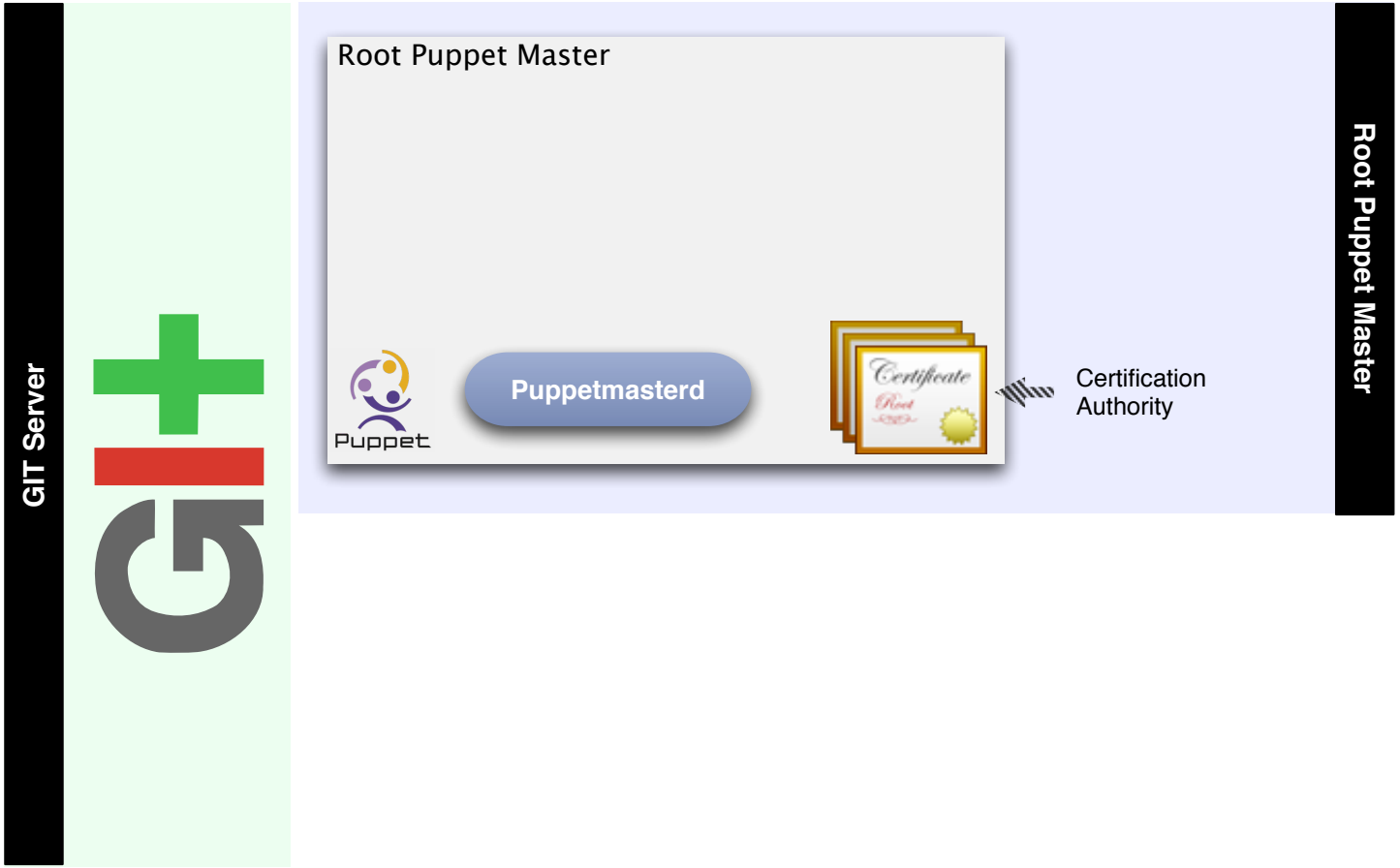
Solution mise en place

Architecture Puppet



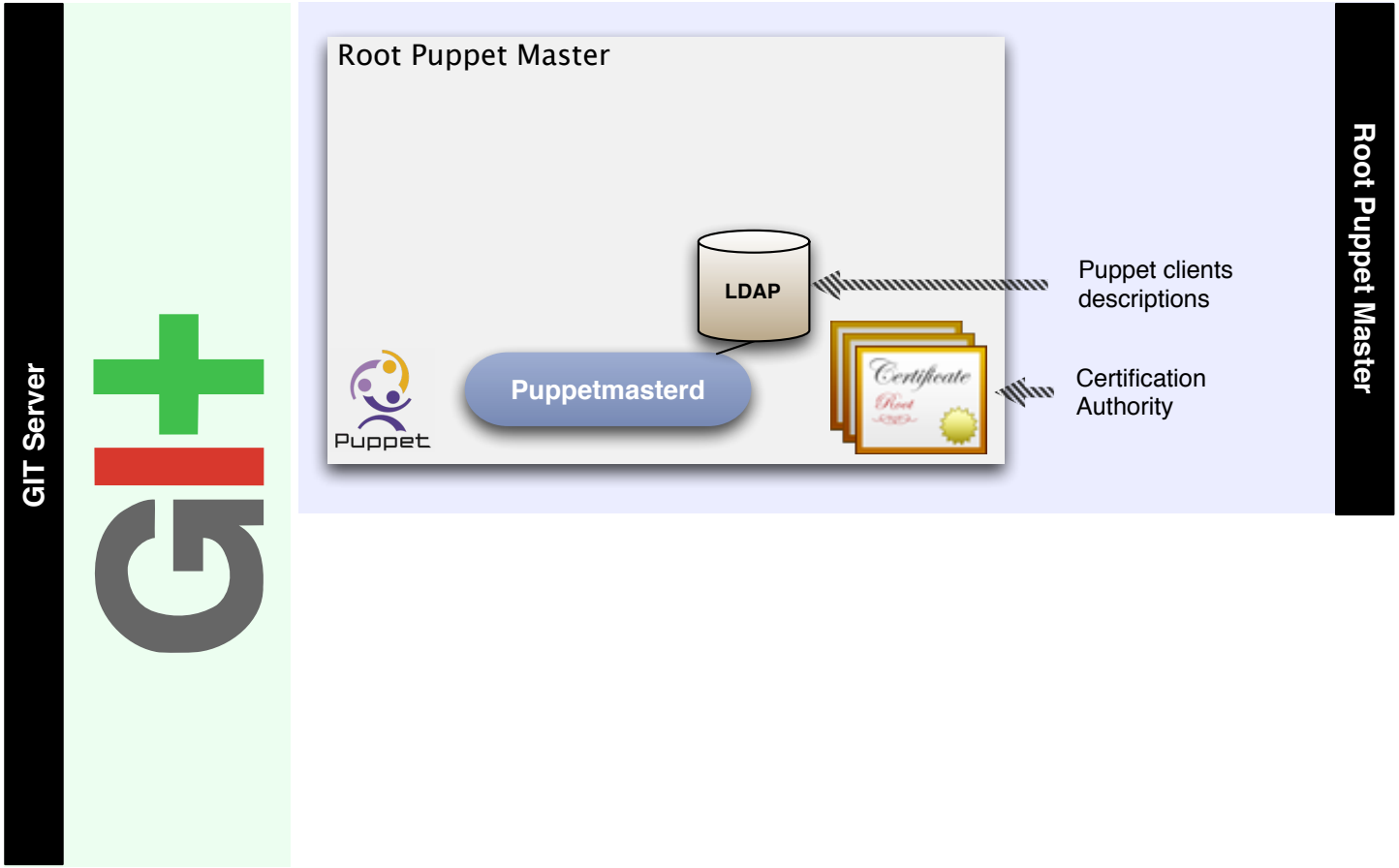
Solution mise en place

Architecture Puppet



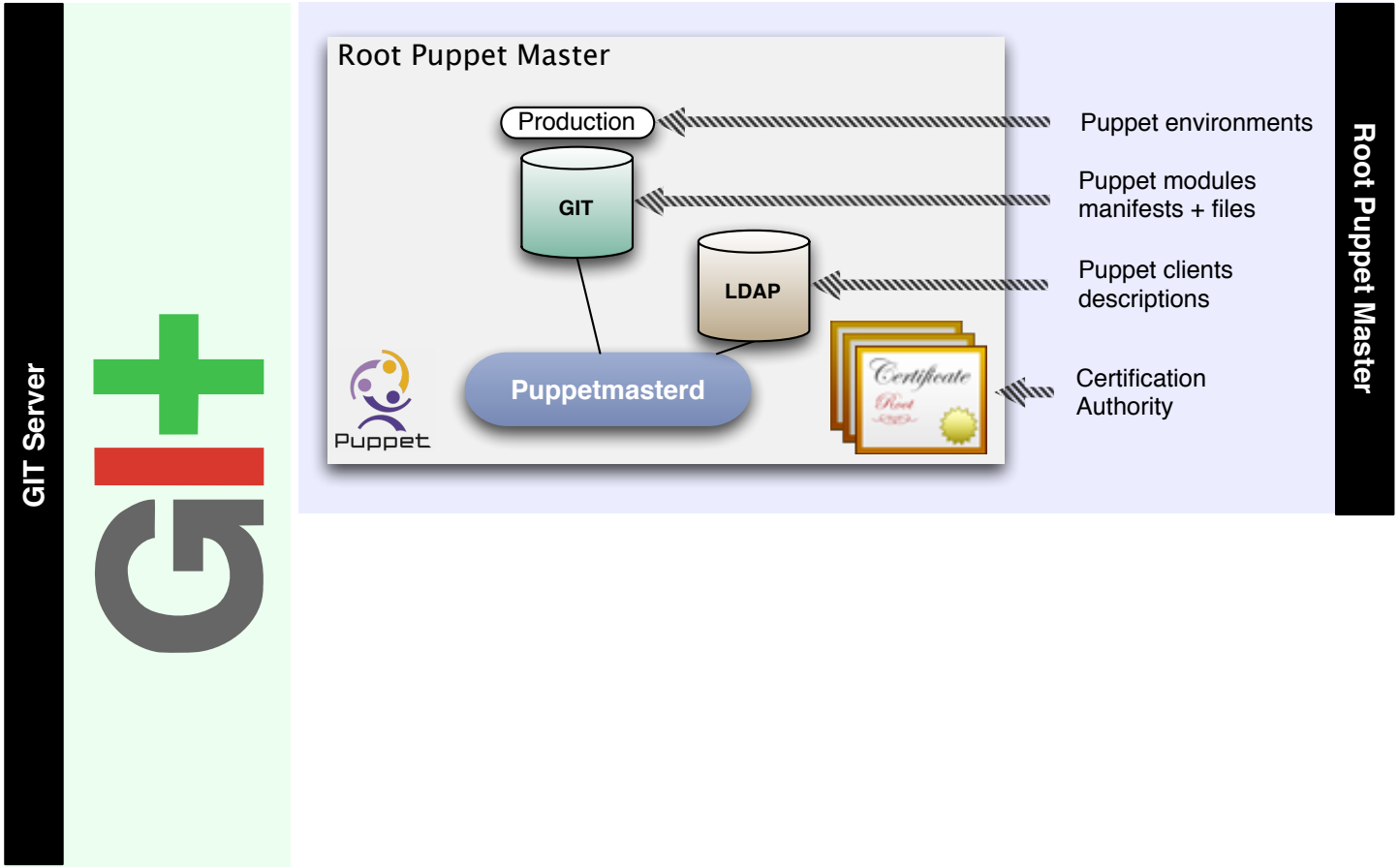
Solution mise en place

Architecture Puppet



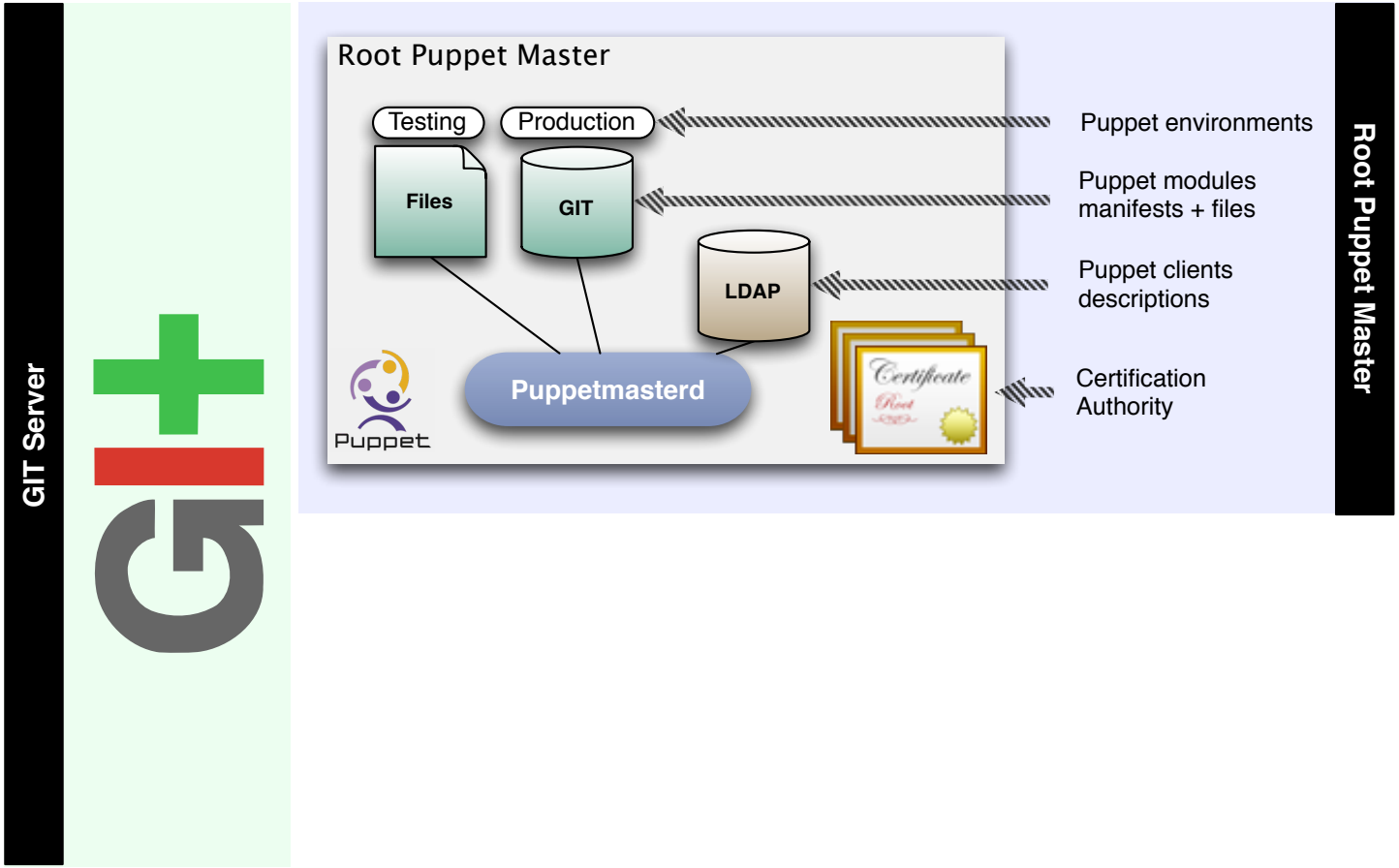
Solution mise en place

Architecture Puppet



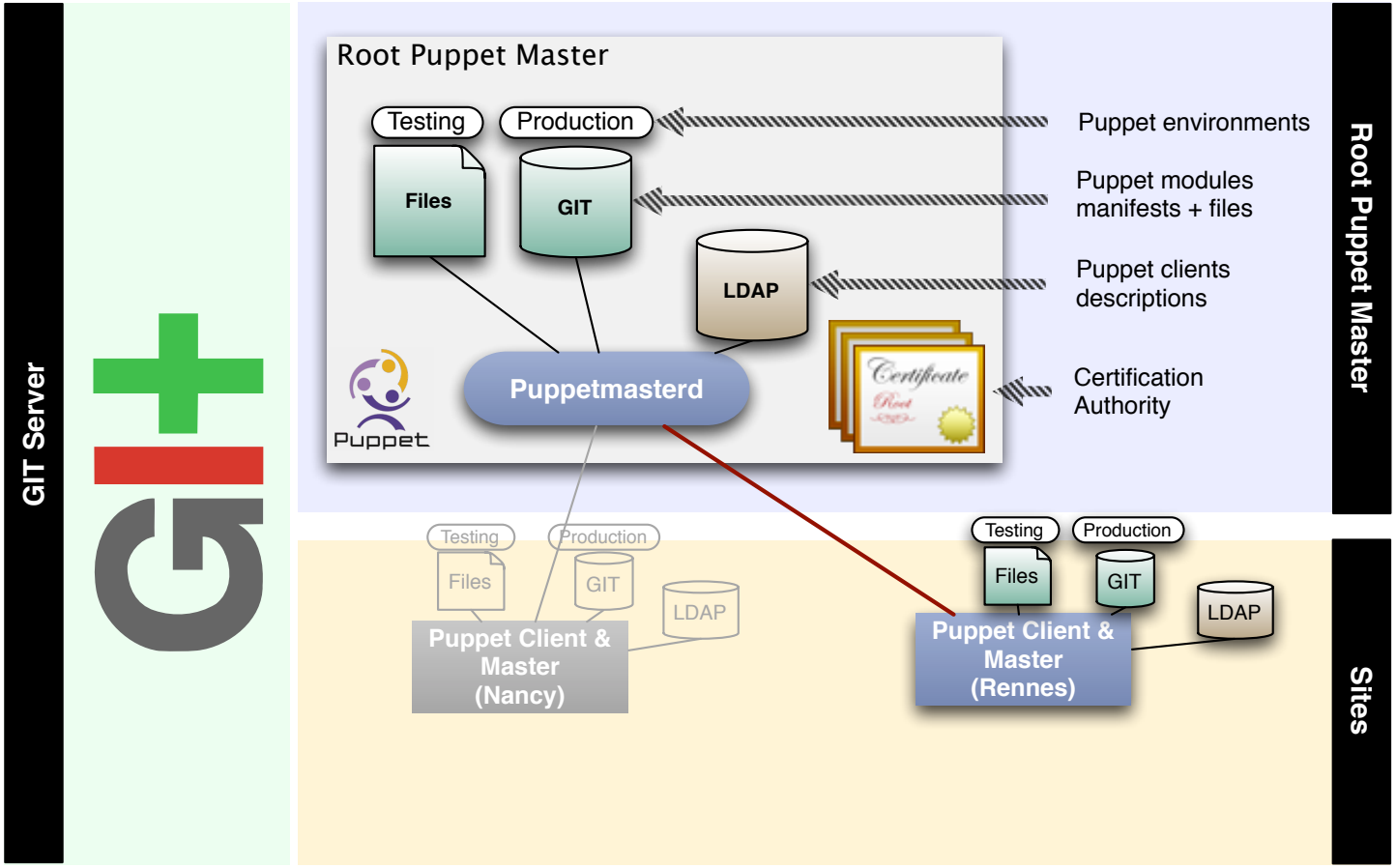
Solution mise en place

Architecture Puppet



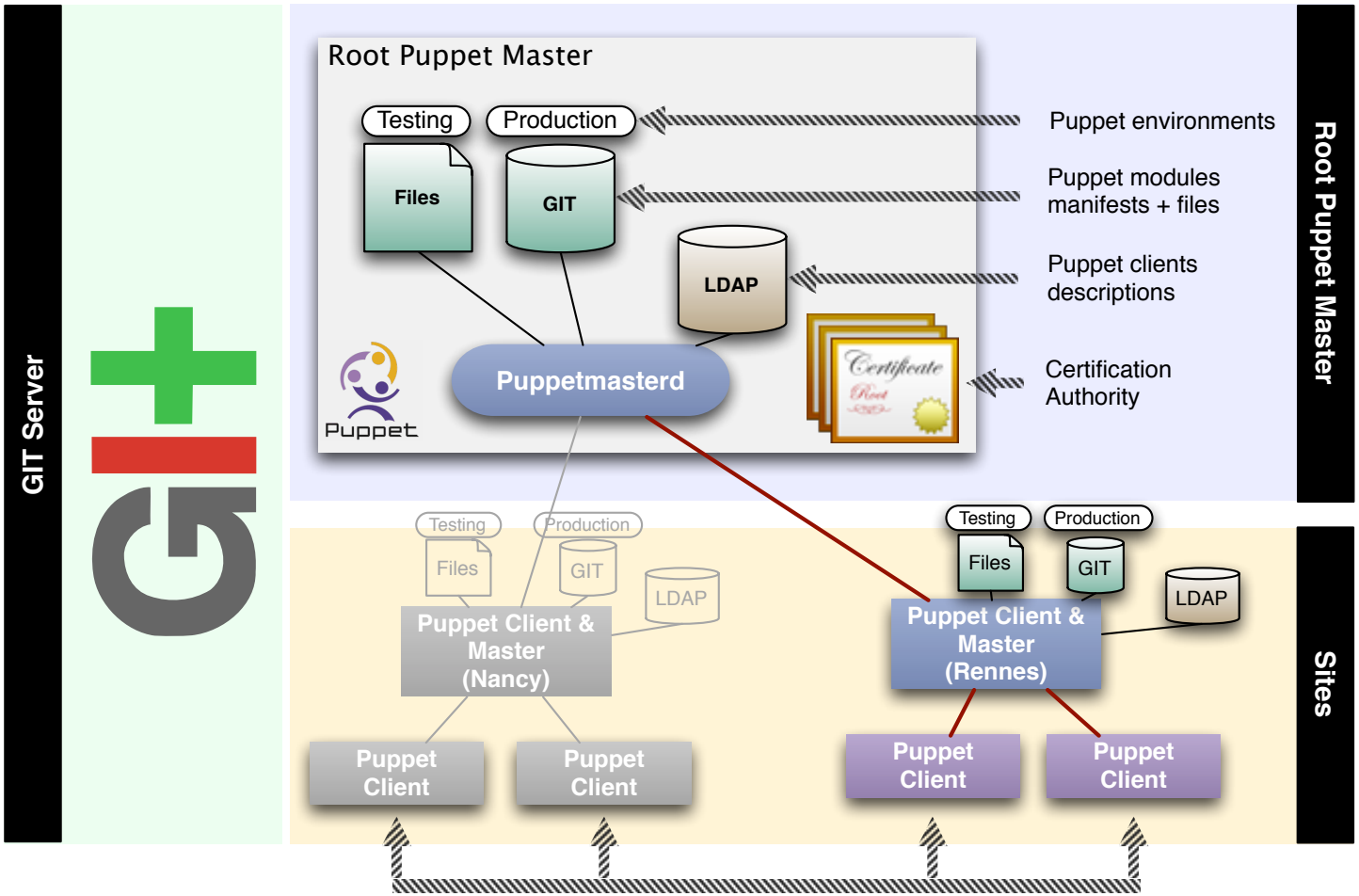
Solution mise en place

Architecture Puppet



Solution mise en place

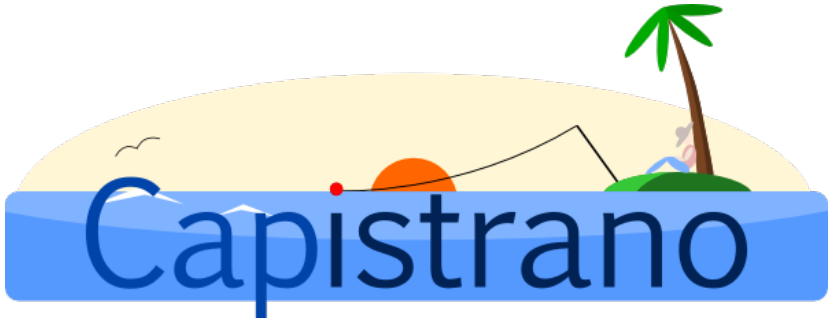
Architecture Puppet



Common services and configurations : bindg5k, proxyg5k, mysqlg5k, adming5k, xeng5k, oarg5k, kadeployg5k, etc...

Solution mise en place

Capistrano



- Outil de déploiement d'applications (Ruby on Rails à l'origine)
- Description dans des Capfiles de tâches, de rôles et de dépendances entre ces tâches
- Intégration poussée de SSH et d'un lanceur parallèle
- `http://github.com/capistrano/capistrano`

Solution mise en place

Le dépôt GIT Puppet

Dépôt GIT

```
.
├── Capfile
├── Gemfile
├── config
├── modules
│   ├── apache
│   ├── frontendg5k
│   ├── oar
│   ├── oarg5k
│   ├── openssl
│   └── puppetg5k
│       ├── Capfile
│       ├── files
│       ├── lib
│       ├── manifests
│       └── templates
└── nagios
```

- Un Capfile principal : namespaces modules, git
- Le répertoire des modules Puppet
- Des modules génériques propres à un service
- Des modules spécifiques à Grid'5000
- Un Capfile par module : namespaces puppet, bind nagios, ...
- 83 modules actuellement

Retour d'expérience

Migration

- Migration progressive depuis septembre 2009
- L'hétérogénéité des systèmes a compliqué l'écriture des recettes au départ
- Prise en compte de spécificités de chaque sites

Retour d'expérience

Évolution de la plate-forme

Choix d'une distribution unique

- Utilisation de la même version de Debian sur tous les serveurs

Grâce à Puppet

- Les spécificités des sites se sont révélées peu à peu
- Un travail d'uniformisation peut donc se faire
- La plate-forme est plus stable et plus maintenable

Retour d'expérience

Adoption par les administrateurs systèmes

Points favorables

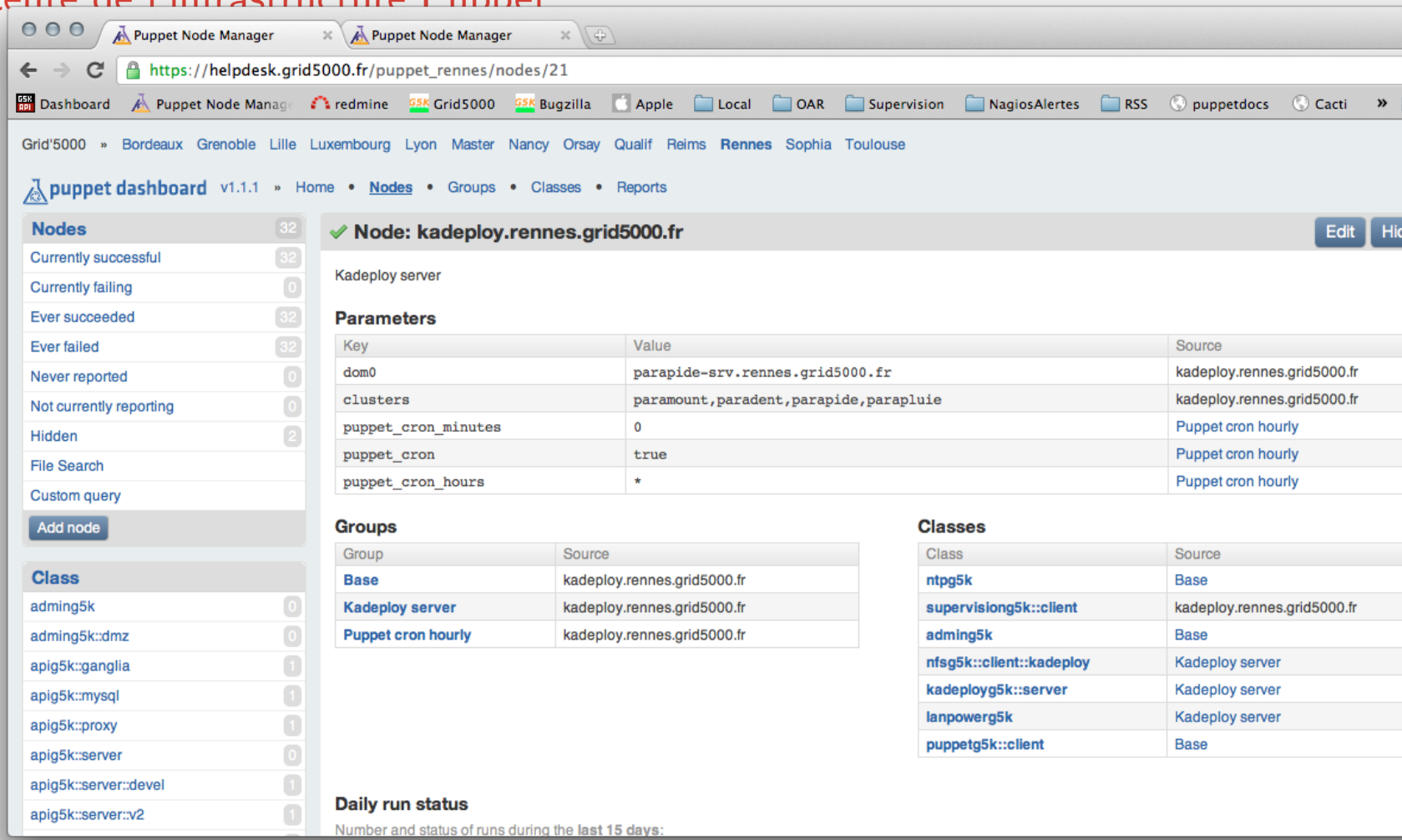
- Environnement technique de travail stimulant
- Prise de conscience de la pérennité du travail réalisé
- Évite de devoir maintenir des documentations d'installation parfois complexes
- La traçabilité, important dans une équipe géographiquement distribué avec un tel turn-over

Point important

- Rester consciencieux : on peut tout "casser" assez facilement

Retour d'expérience

Évolution récente de l'infrastructure Puppet



Grid'5000

- David Margery, Directeur Technique
- Pascal Morillon
- Emmanuel Jeanvoine
- Frédéric Desprez, Directeur Scientifique
- Lucas Nussbaum
- Olivier Richard
- et beaucoup d'autres
- `https://www.grid5000.fr/`