

## Key features

- ➔ **Install and configure a large number of nodes**
  - ▶ Install several clusters in one shot
  - ▶ Support for concurrent deployments
  - ▶ Control several clusters from a single client
- ➔ **Manage a library of pre-configured system images**
  - ▶ User-provided images
  - ▶ Visibility of images (shared, private)
- ➔ **Reliability of the installation process**
  - ▶ Customizable workflow engine
  - ▶ Windowed operations
  - ▶ Escalation of low-level remote commands
- ➔ **Hardware compatibility**
  - ▶ Built on top of PXE, DHCP, TFTP/HTTP
  - ▶ Remote operations based on SSH
  - ▶ Customizable remote low-level operations (IPMI, ...)
- ➔ **Software compatibility**
  - ▶ Support any operating system (Linux, \*BSD, Windows, ...)
  - ▶ Integration with batch scheduler and net. isolation tools
  - ▶ Remote control API

## Scalability

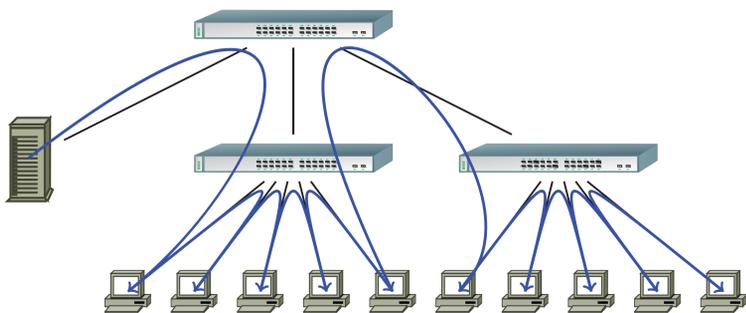
### ➔ System image broadcast

Goal : send a big file on thousands of nodes

Challenges : avoid network bottlenecks, saturation of links

#### ▶ Several alternatives available

Chain, Tree, Bittorrent, ...



- ▶ **Default alternative : Topology-aware chain broadcast**
  - Parallel tree-based initialization of the chain
  - Saturation of full-duplex network in both directions
  - Efficient on networks composed of hierarchy of switches

### ➔ Parallel operations

Goal : executing commands on thousands of nodes

Challenges : avoid client overloading, gather commands outputs

- ▶ Based on **TakTuk** (<http://taktuk.gforge.inria.fr>)
- ▶ **Hierarchical connections** between nodes
- ▶ Adaptive work-stealing algorithm
- ▶ **Auto-propagation** mechanism

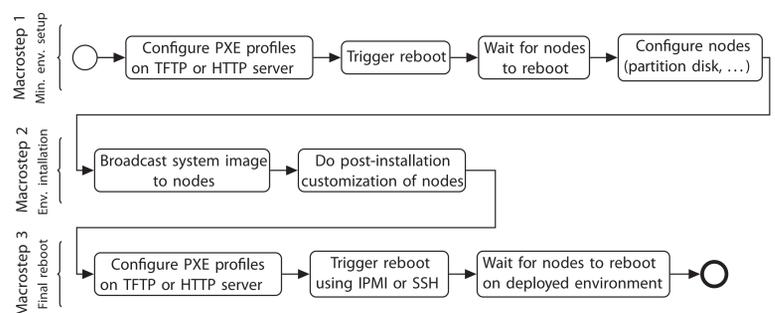
## Reliability

### ➔ Reliable workflow engine

Goal : manage the installation process

Challenges : handle hw. network failures, customization

- ▶ Engine based on **event automata**
- ▶ **Fallback methods** in case of failure
- ▶ Timeouts and retries at every step
- ▶ A typical workflow example:



### ➔ Reliable reboot and power operations

Goal : trigger remote reboot and power on/off on nodes

Challenges : reliability, compatibility

- ▶ **Compatibility** with remote hardware managements protocols
- ▶ **Escalation** of several level of administrator defined commands
- ▶ Managing groups of nodes (e.g. PDU reboots)
- ▶ **Windowed operations** (DHCP flood, electrical hazards, ...)

## Evaluation

- ▶ **Key software on Grid'5000 since 2004**
- ▶ **25 clusters on 10 sites**
- ▶ 620 users, **170 000 deployments**
- ▶ about **10 mins** to deploy 130 nodes
- ▶ **Virtualized infrastructure**
- ▶ **4000 VMs** dispatched on **635 physical nodes**
- ▶ **3838 nodes** in a single shot in **less than 1 hour**

## Software suite

- ▶ **Management of images**
- ▶ User custom images
- ▶ **Rights management**
- ▶ Compatibility with batch scheduler
- ▶ **Statistics collection**
- ▶ Identify hardware issues, ...
- ▶ **Frontends to low-level tools**
- ▶ Reboot and power on/off operations, serial consoles
- ▶ **DEB and RPM packages**
- ▶ **Actively developed** since 2009