Bring Grids Power to Internet-Users thanks to Virtualization Technologies

Virtualization Working Group - HEMERA
Yvon Jégou (MYRIADS) / Adrien Lèbre (ASCOLA)

Context

- Job scheduling strategies for clusters/grids:
  static allocation of resources / "user-intrusive"

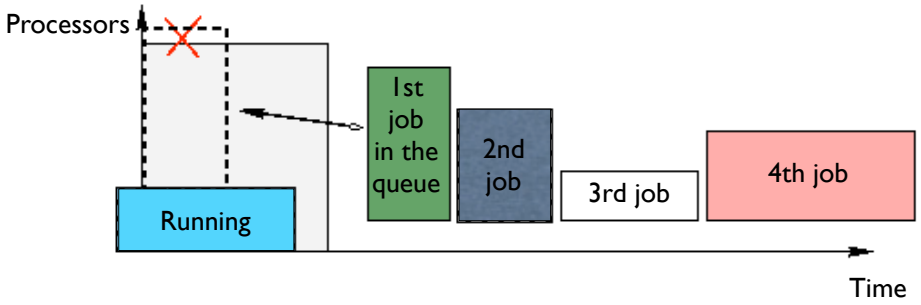
Based on user estimates (time/resources)
For a bounded amount of time
(e.g. 4 nodes for 2 hours)

Resources are reassigned at the end of the slot without considering real needs of applications
(in the worst case, running applications can be simply withdrawn from resources, i.e. G5K best effort mode)

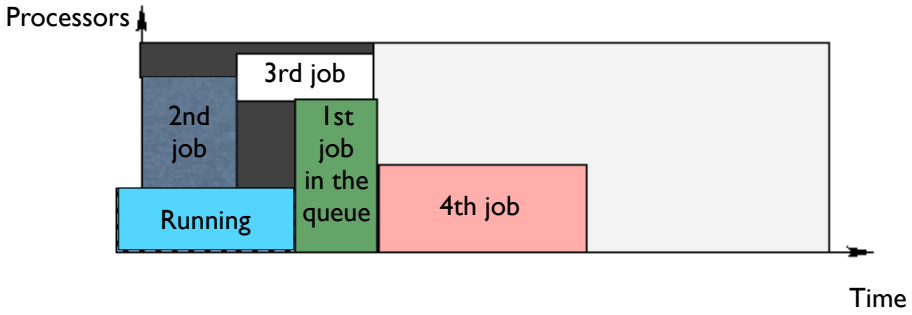
⇒ Coarse-grain exploitation of the architecture

Context

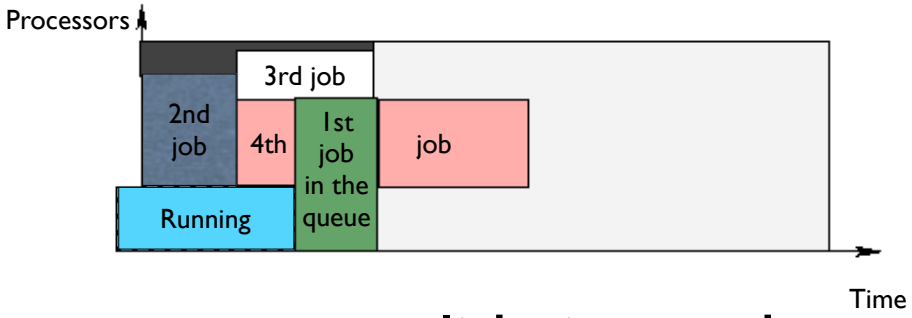
- Batch scheduler policies: closed to FCFS



Jobs arrive in the queue
and have to be scheduled.



FCFS + Easy backfilling
Jobs 2 and 3 have been backfilled.
Some resources are unused (dark areas)



Easy backfilling with preemption
The 4th job can be started without
impacting the first one.
A small piece of resources is still unused.

⇒ consolidation and preemption to finely exploit distributed resources

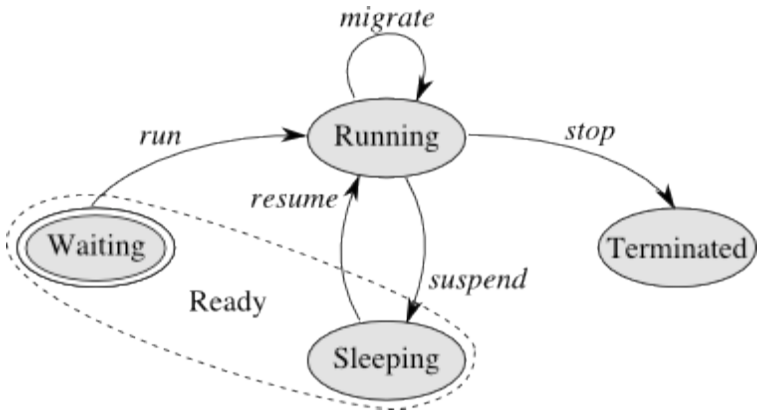
Consolidation and Preemption

- Few schedulers include preemption mechanisms based on checkpointing solutions:
  -  Strongly middleware/OS dependent
  -  Still not consider application resource changes
- SSI approaches include both consolidation and preemption of processes:
  -  Strongly middleware/OS dependent
  -  SSI developments are tedious (most of them have been give up)
- Exploit all VM capabilities
  (start/stop - suspend/resume - migrate)

Cluster-Wide Context Switch

- General idea: manipulate *vJobs* instead of jobs
  (by encapsulating each submitted job in one or several VMs)

- In a similar way of usual processes,
  each vjob is in a particular state:



- A cluster-wide context switch (a set of VM context switches)
  enables to efficiently rebalance the cluster according to the:
  scheduler objectives / available resources / waiting vjobs queue
  (elasticity) [VTDC 2010]

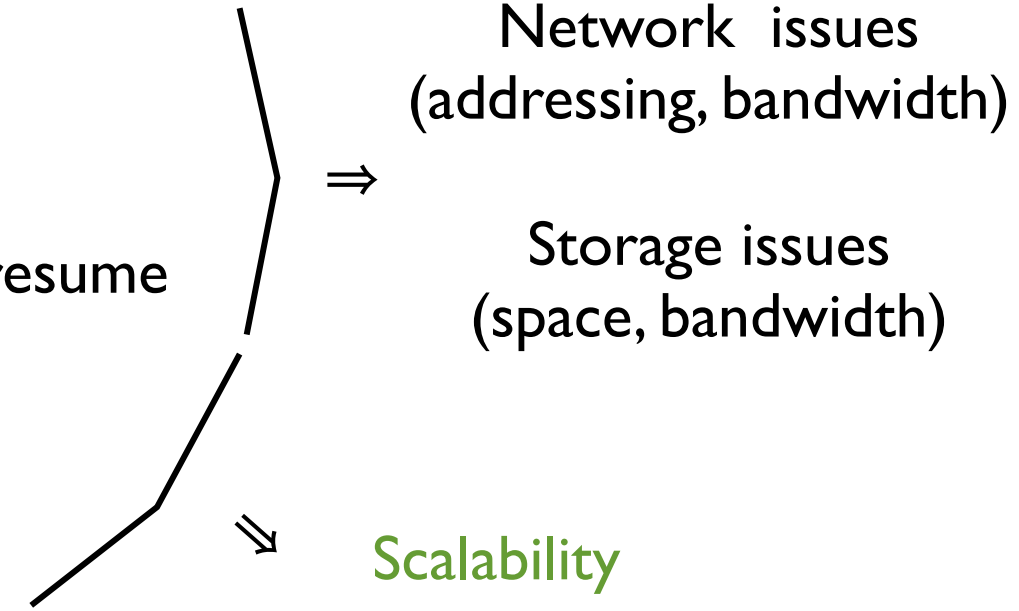
Grid-Wide Context switch

- Management of *vjobs* through the whole infrastructure
- Multi-sites

Live migration

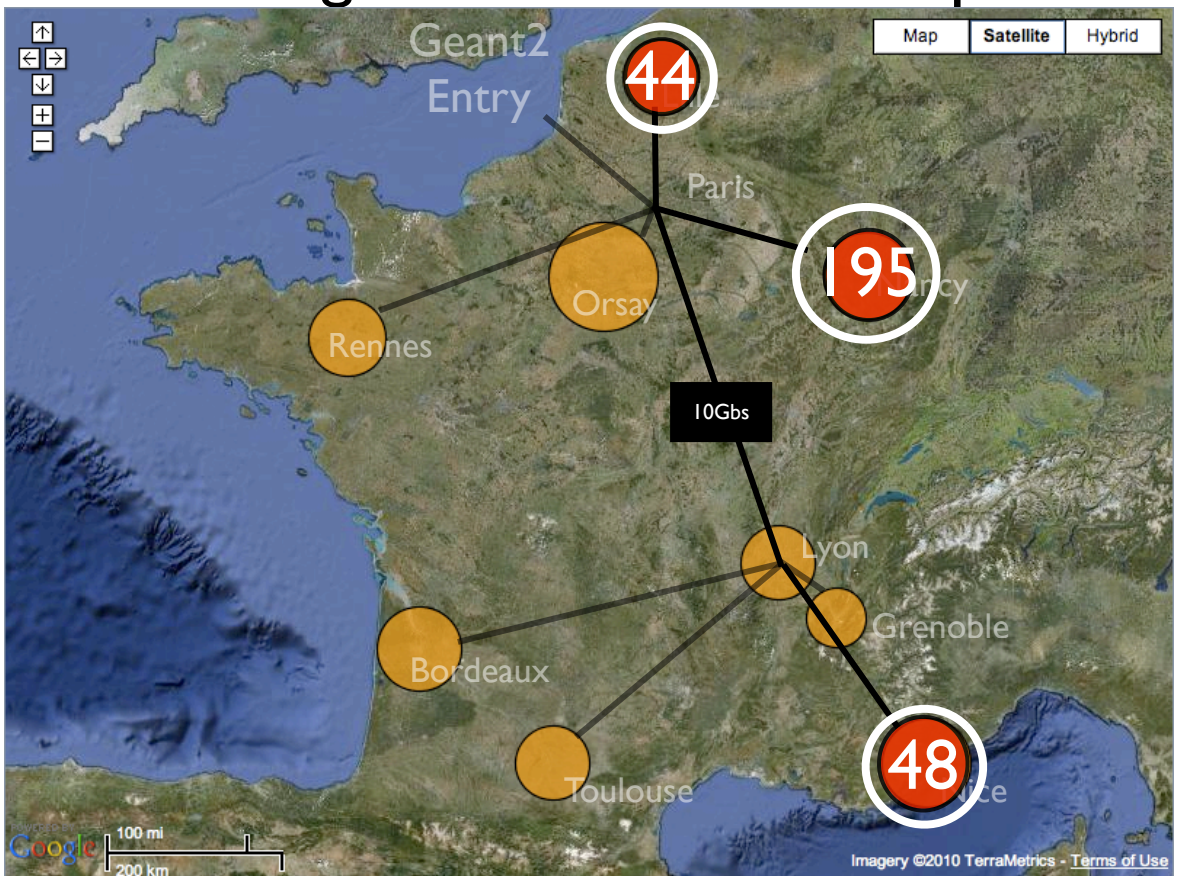
Local suspend / remote resume

Periodic checkpoint



Grid'5000 - Cloud Study Example

- Deploy Nimbus on a large number of nodes spread over 3 sites



280 VMMs, 1600 virtual CPUs and more than 2TBytes of RAM (3 sites, static assignment) [Riteau - G5K - Spring School 2010]

Grid-Wide Context switch

Mid-term objective: the whole grid

1500 VMMs / 3000 VMMs (2 VMs per node) / ....

Dynamic placement of VMs according to maintenance operations, failures, scheduling policies, etc.

Work in progress

Integration between Entropy (autonomic cluster-wide management / Ascola) and Saline (grid-wide fault tolerant system / Myriads)

From the Grid to the Desktop

- Interconnecting each desktop to the Grid
- Launch your *vjob* on your desktop and run it somewhere in the infrastructure (on the Grid ? on another desktop ?)

Network issues
(addressing, bandwidth, external services)

Security issues
(external connections)

Animation

- Since 2009: listing/promoting and pooling research and development activities done in the context of the Grid'5000 around virtualization
- A wiki page on G5K, a mailing list (virtualisation@lists.grid5000.fr)
- A first JTE (June 2010 ASR/Mines - 15 pers - 2 international talks)
- Several WIPs:

  Elasticity concerns at application level (ANR SelfXL)

  Energy issues (TUNe, Entropy, ...)

  DSL to manipulate a large number of VMs across the Grid and through a simple shell interface [Pottier-DAIS 2010]
  Next target: 10 000
- CloudForHPC (COST proposal in progress)

Bring Grids Power to Internet-Users thanks to Virtualization Technologies

Virtualization Working Group - HEMERA
Yvon Jégou (MYRIADS) / Adrien Lèbre (ASCOLA)