Ecole des Mines de Nantes

Self-organization of applications and systems to optimize resources usage in virtualized data centers

Teratec

06/28 2012

Jean-Marc Menaud

Ascola team EMNantes-INRIA, LINA

INVENTEURS DU MONDE NUMÉRIQUE

ECOLE DES MINES DE NANTES

Motivations

- **Increasing popularity of Cloud Computing solutions**
  - **Data-centers (DCs) are amazingly growing**

- **DC providers have to face ressource management and energy consumption concerns**

Key characteristics

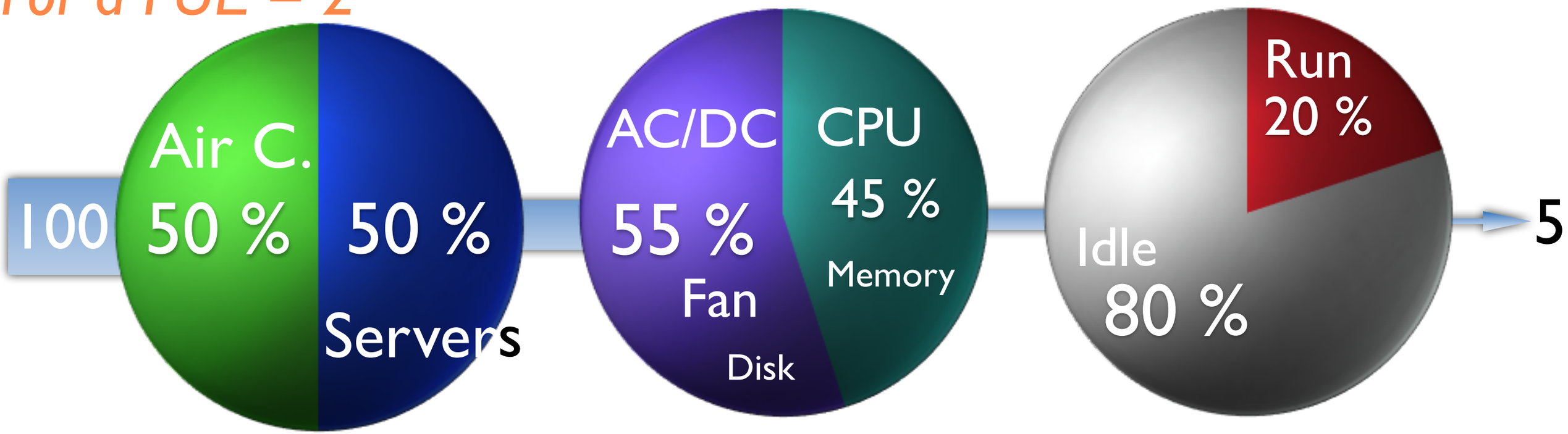
Domains of expertise

Overall objective:

Capacity planning is the process of planning, analyzing, sizing, managing and optimizing capacity to satisfy demand, swiftly and at a reasonable cost.

Capacity planning : Energy focus

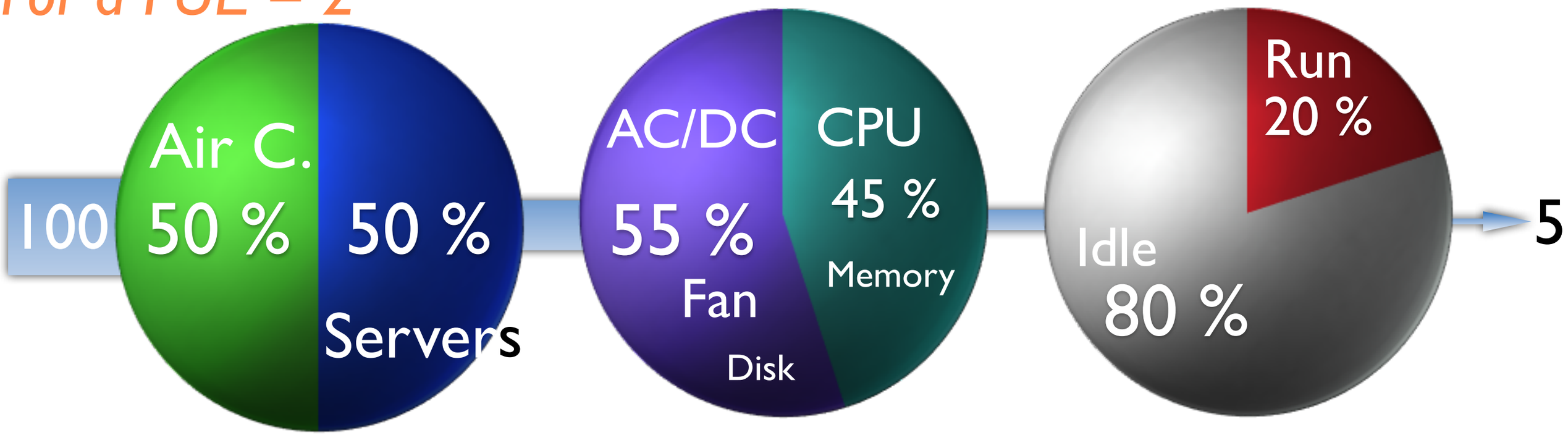
For a PUE = 2



Data center Servers CPU

Capacity planning : Energy focus

For a PUE = 2

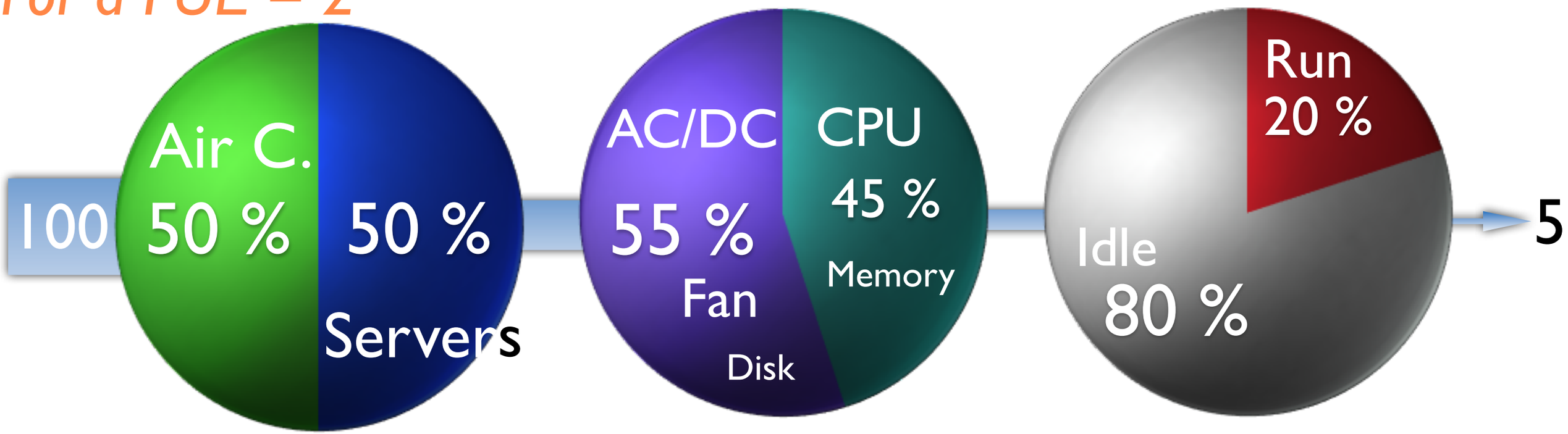


Data center Servers CPU

- **Analysis of the cost of a 2 MegaWatts DC (5000 nodes, 400w/h)**
  - PUE of 2, 0.06€/kWh => 2 120 886 €
  - A decrease of 5% enables a gain of 110K€
- **Managing DC resources finely becomes a major challenge**

Capacity planning : Energy focus

For a PUE = 2



Data center Servers CPU

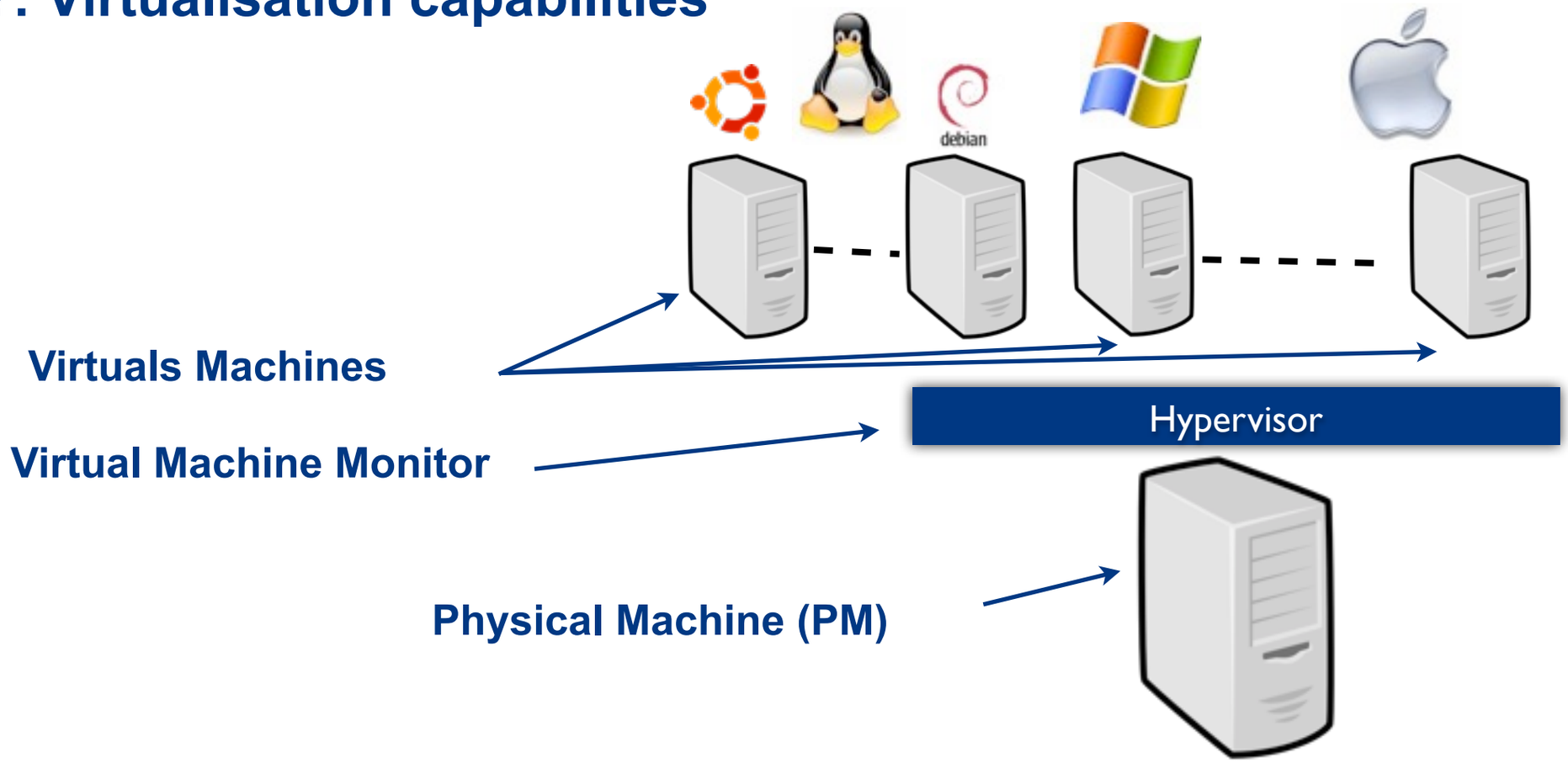
- **Analysis of the cost of a 2 MegaWatts DC (5000 nodes, 400w/h)**
  - PUE of 2, 0.06€/kWh => 2 120 886 €
  - A decrease of 5% enables a gain of 110K€

Managing DC resources finely becomes a major challenge

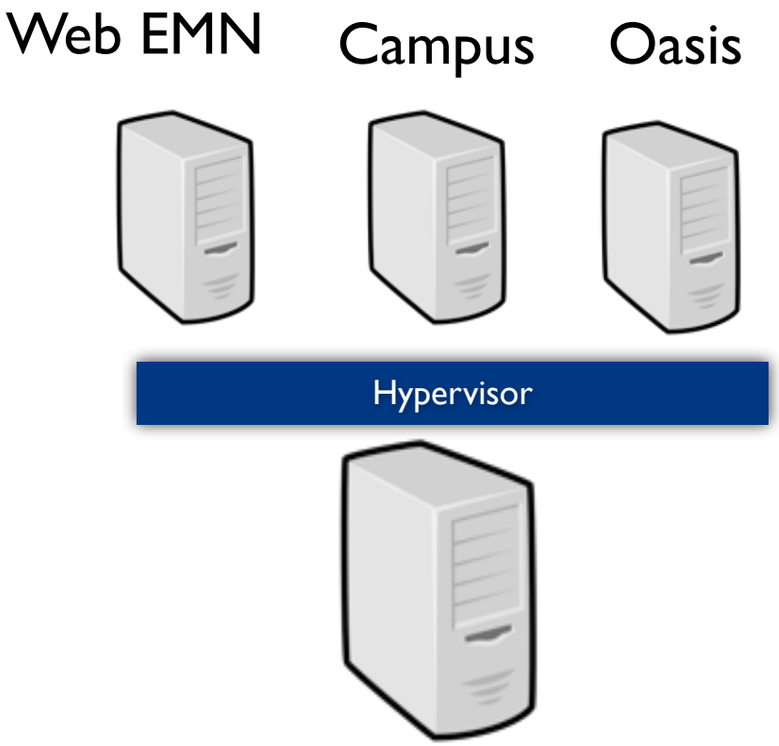
Consolidation

- **Consolidation :**
  - Consolidating to computation reduces the number of running nodes

    So ernegy consumption
  - Reduces hardware costs while providing more efficient node
- **How ?: Virtualisation capabilities**

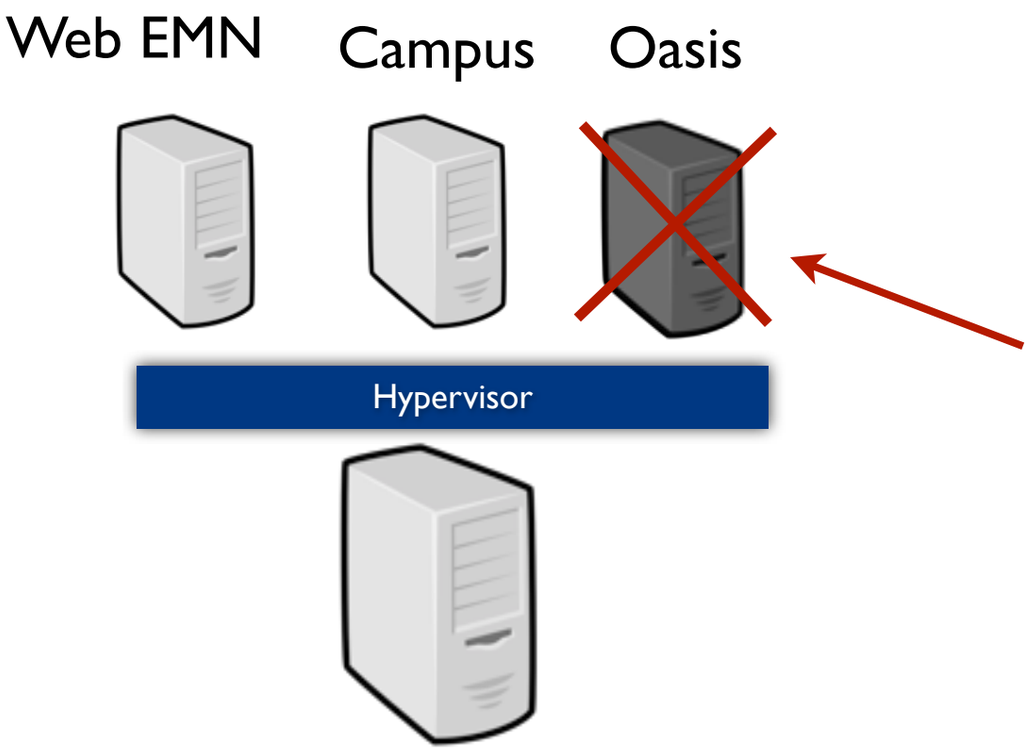


Virtualization capabilities



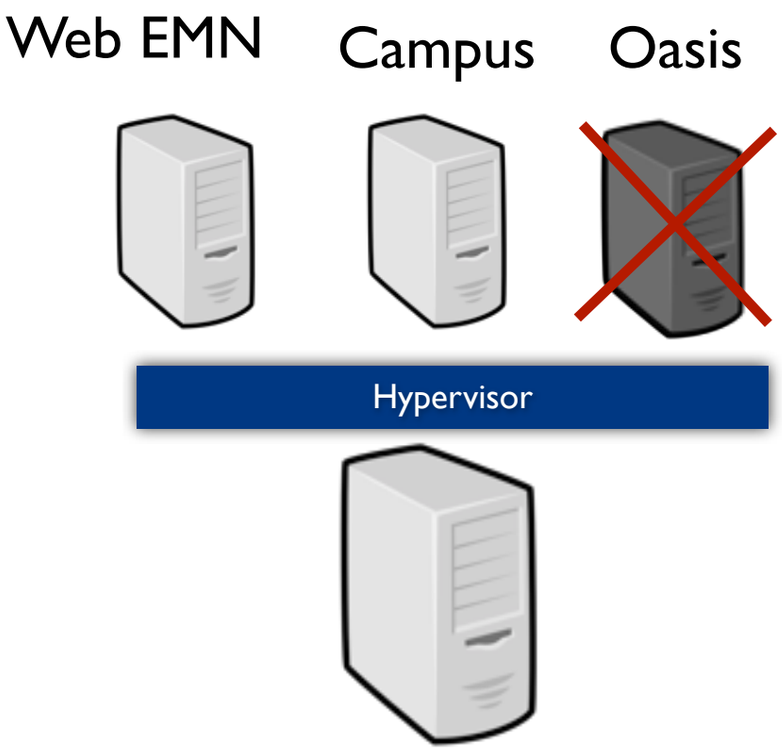
- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

Virtualization capabilities



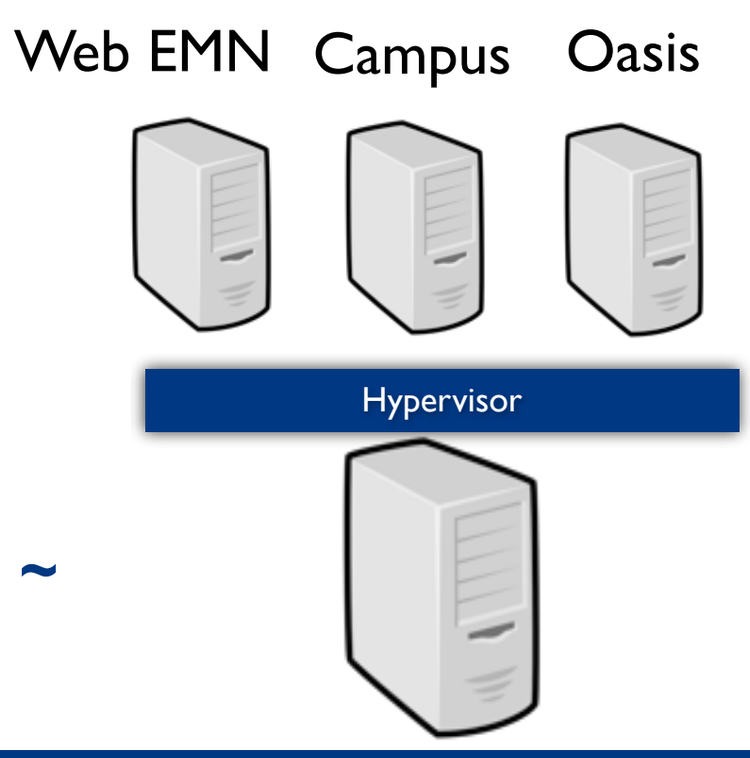
- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

Virtualization capabilities



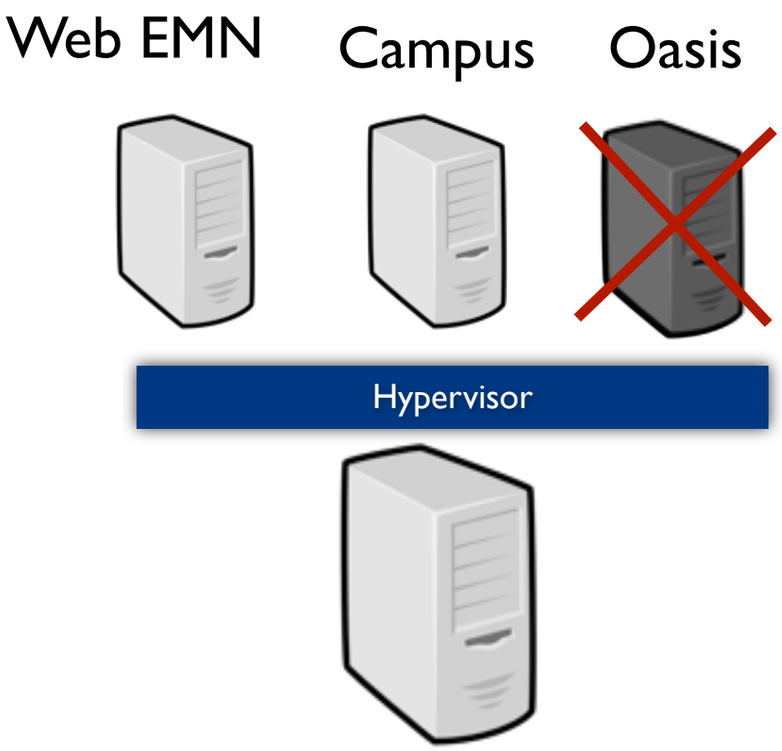
- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

Virus / Invasion / Crash



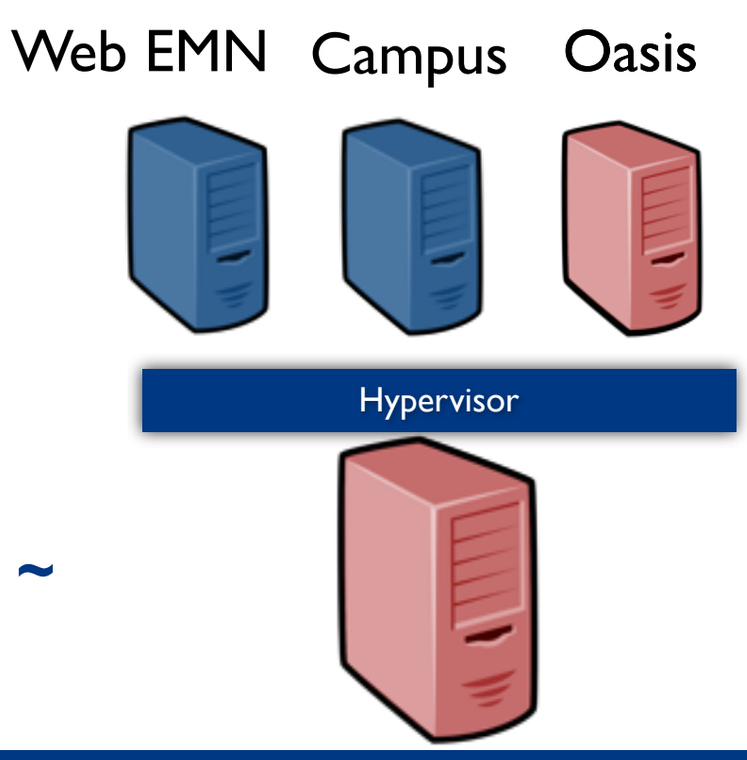
- **Live migration (load-balancing)**
- **High Availability(downtime ~ 60 ms)**

Virtualization capabilities



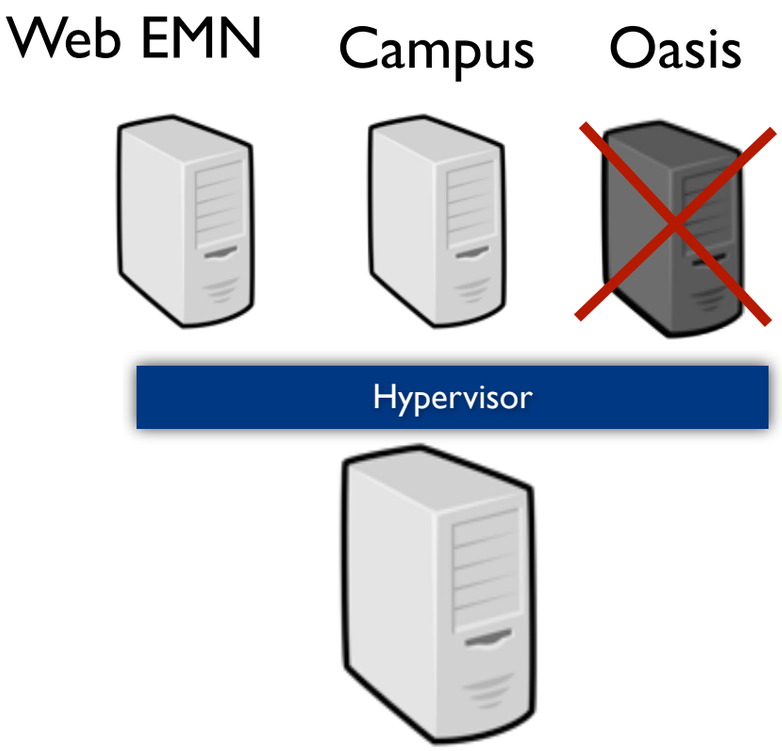
- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

Virus / Invasion / Crash



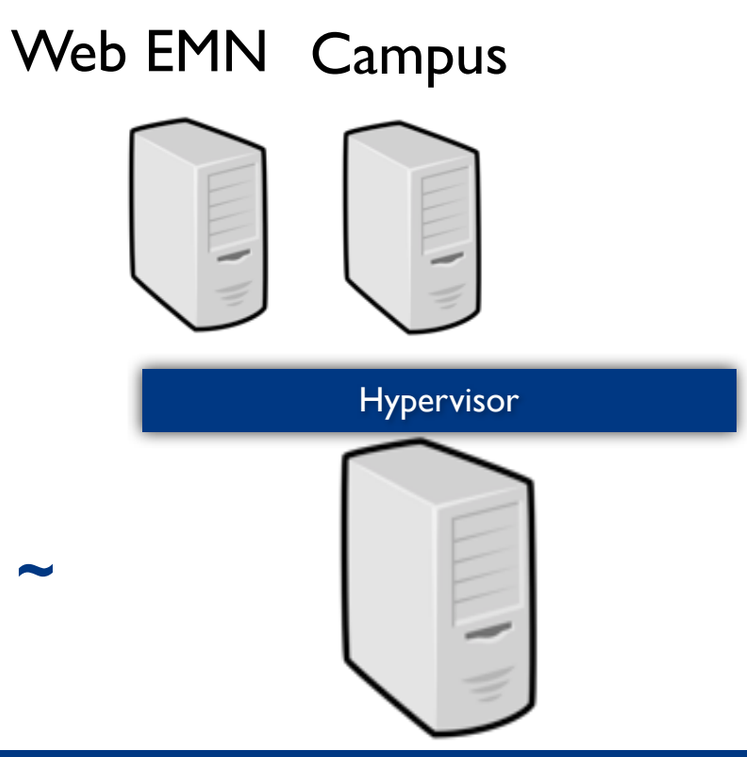
- **Live migration (load-balancing)**
- **High Availability(downtime ~ 60 ms)**

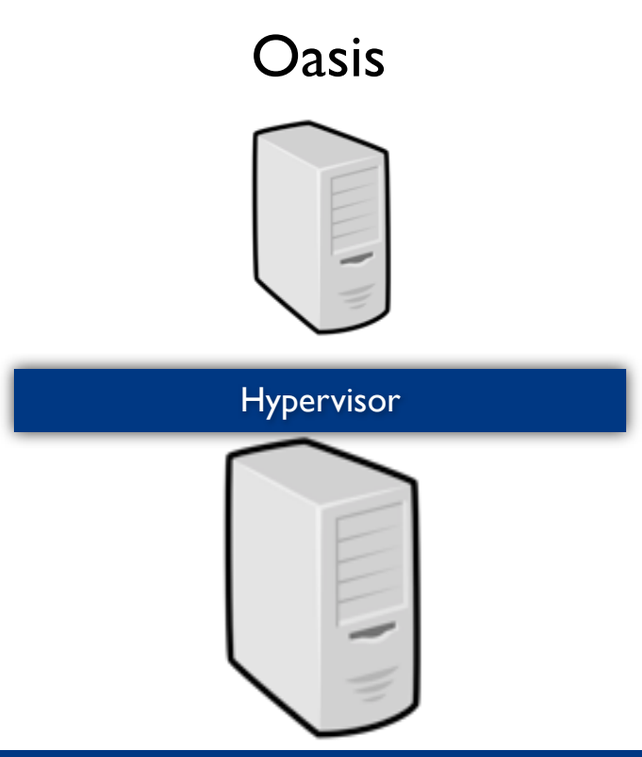
Virtualization capabilities



- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

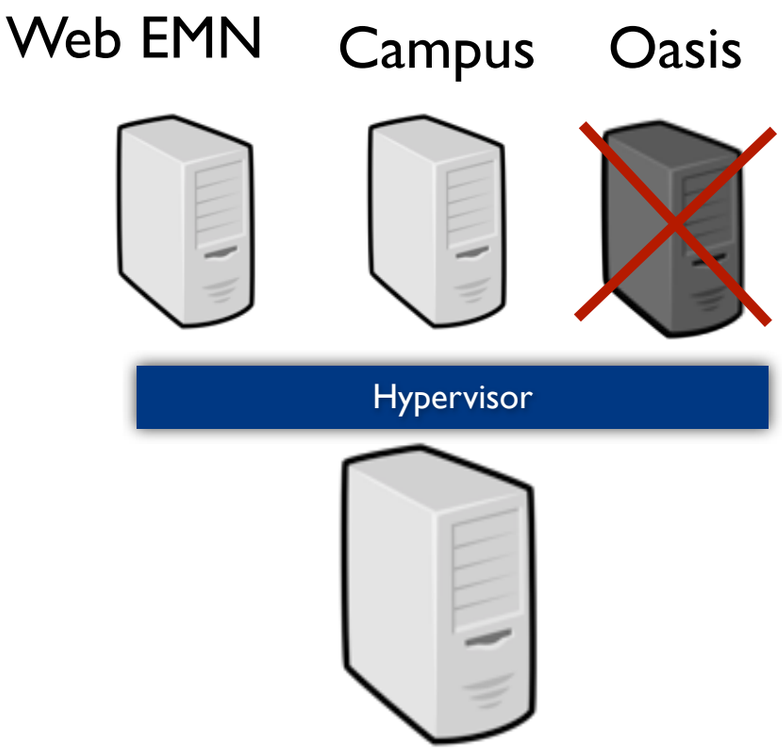
Virus / Invasion / Crash





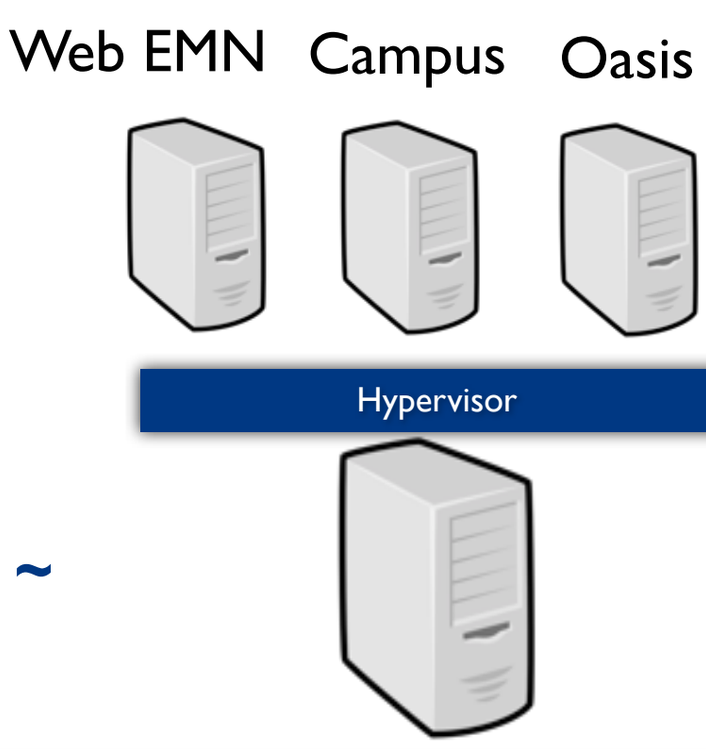
- **Live migration (load-balancing)**
- **High Availability(downtime ~ 60 ms)**

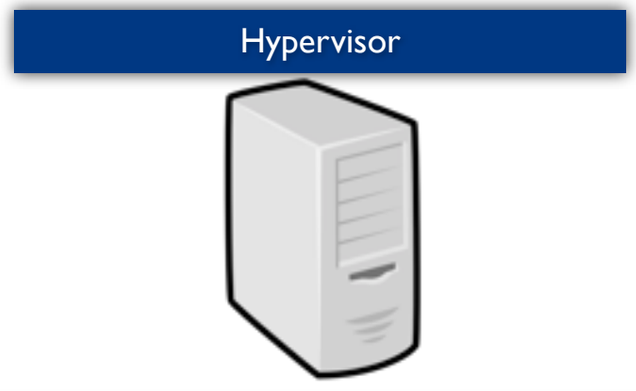
Virtualization capabilities



- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

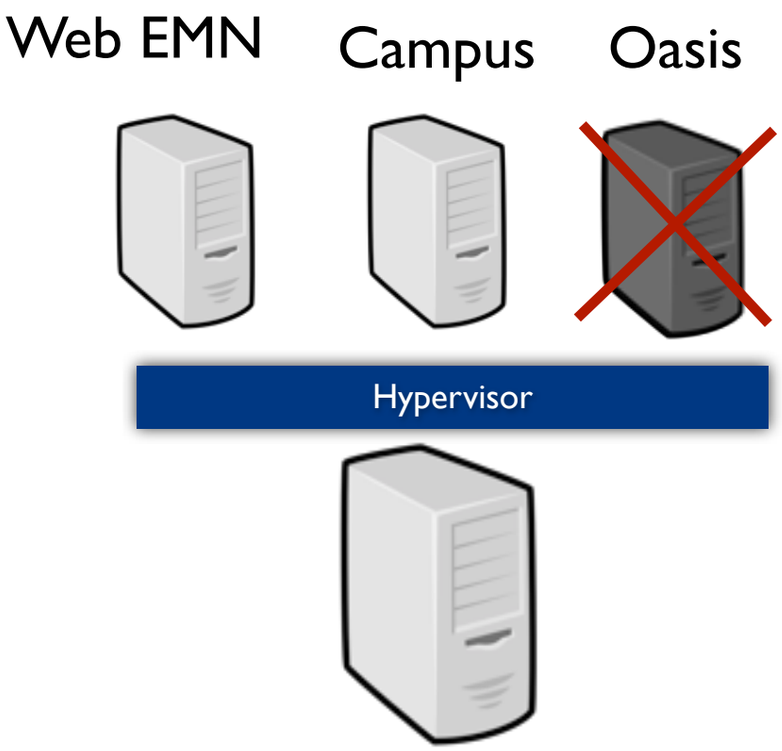
Virus / Invasion / Crash





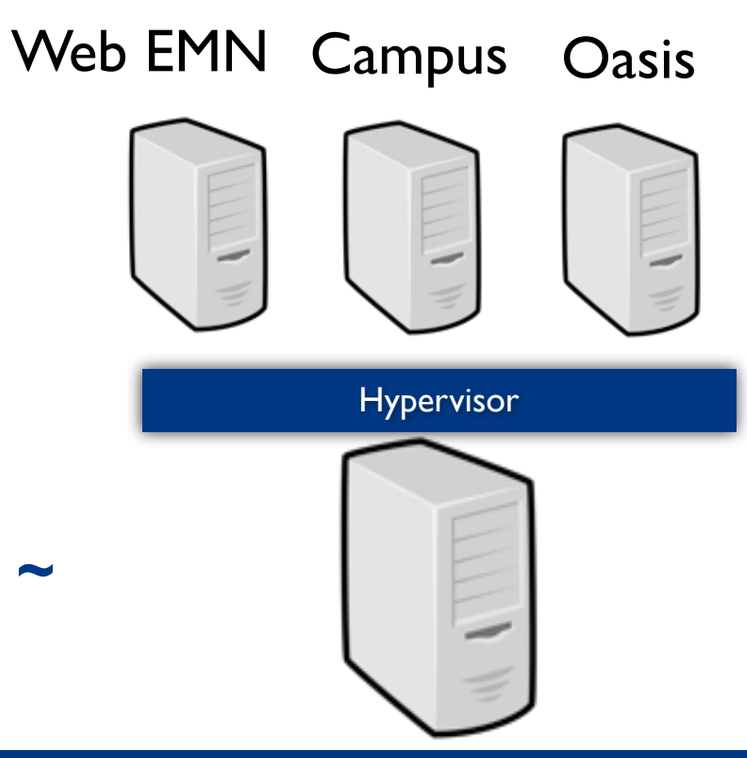
- **Live migration (load-balancing)**
- **High Availability(downtime ~ 60 ms)**

Virtualization capabilities



- **Isolation (security between VM)**
- **suspend/resume/reboot (maintenance)**

Virus / Invasion / Crash

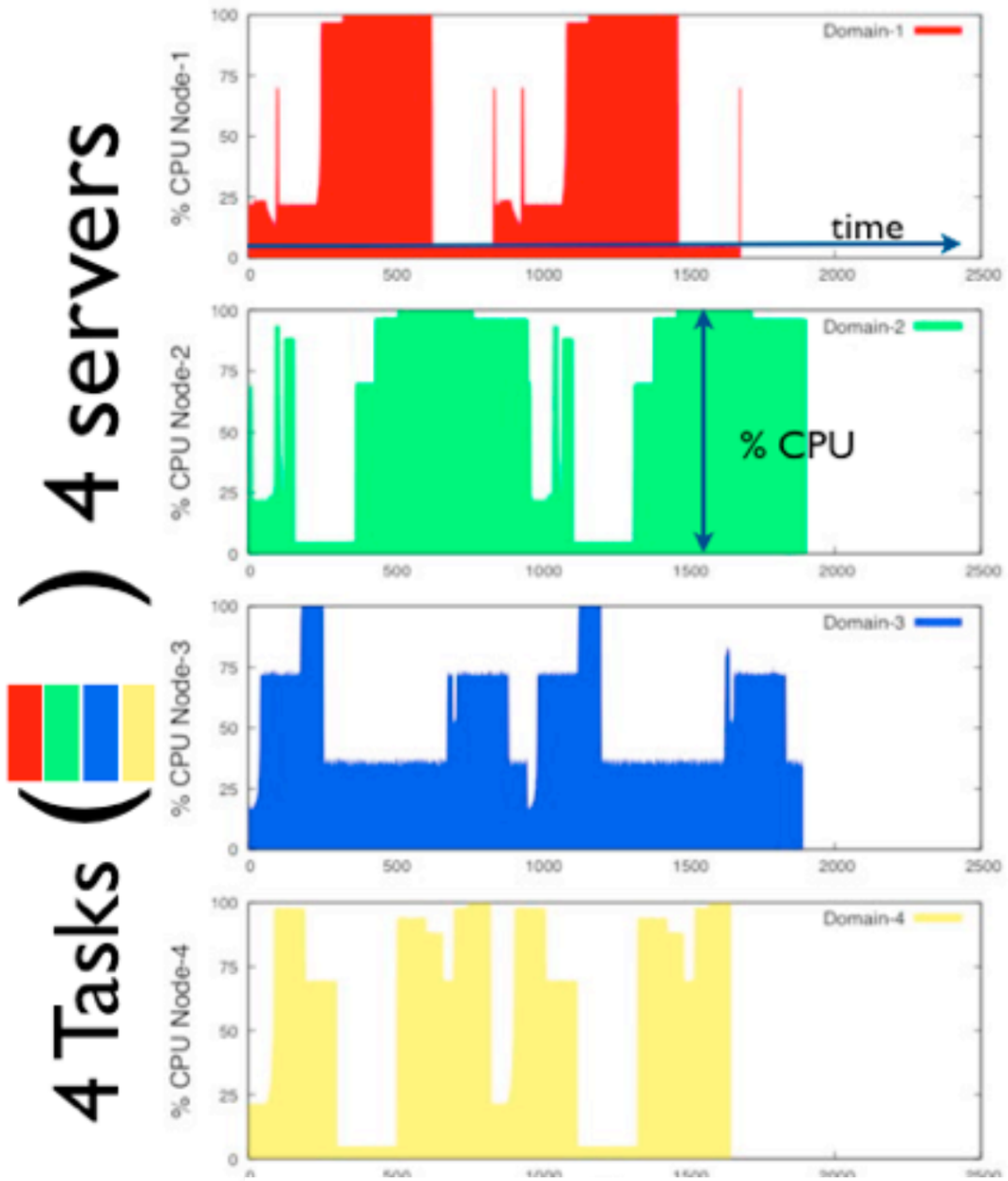


- **Live migration (load-balancing)**
- **High Availability(downtime ~ 60 ms)**

2>Virtual machine manager for clusters

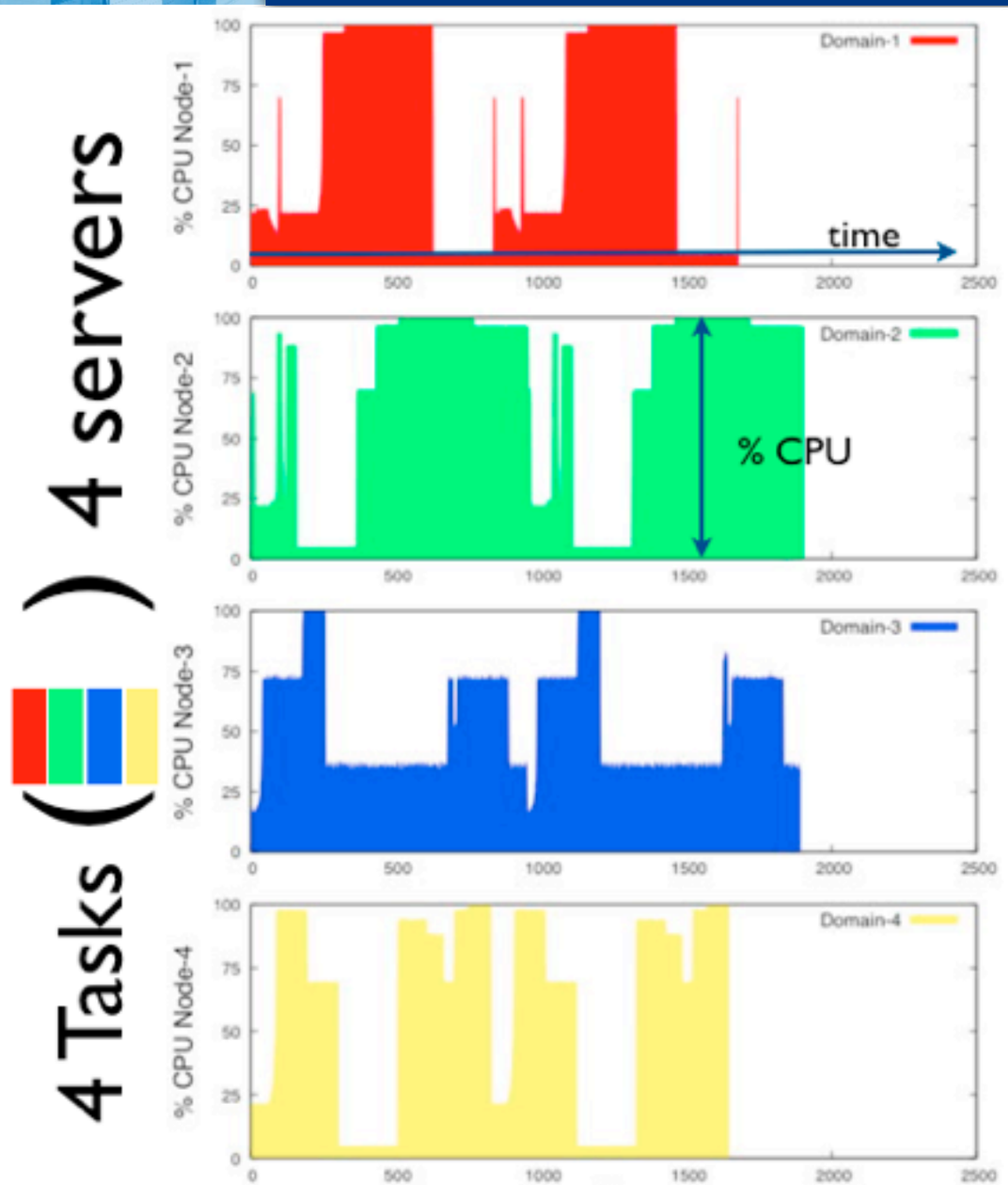
1
2
3
4
5
6
7

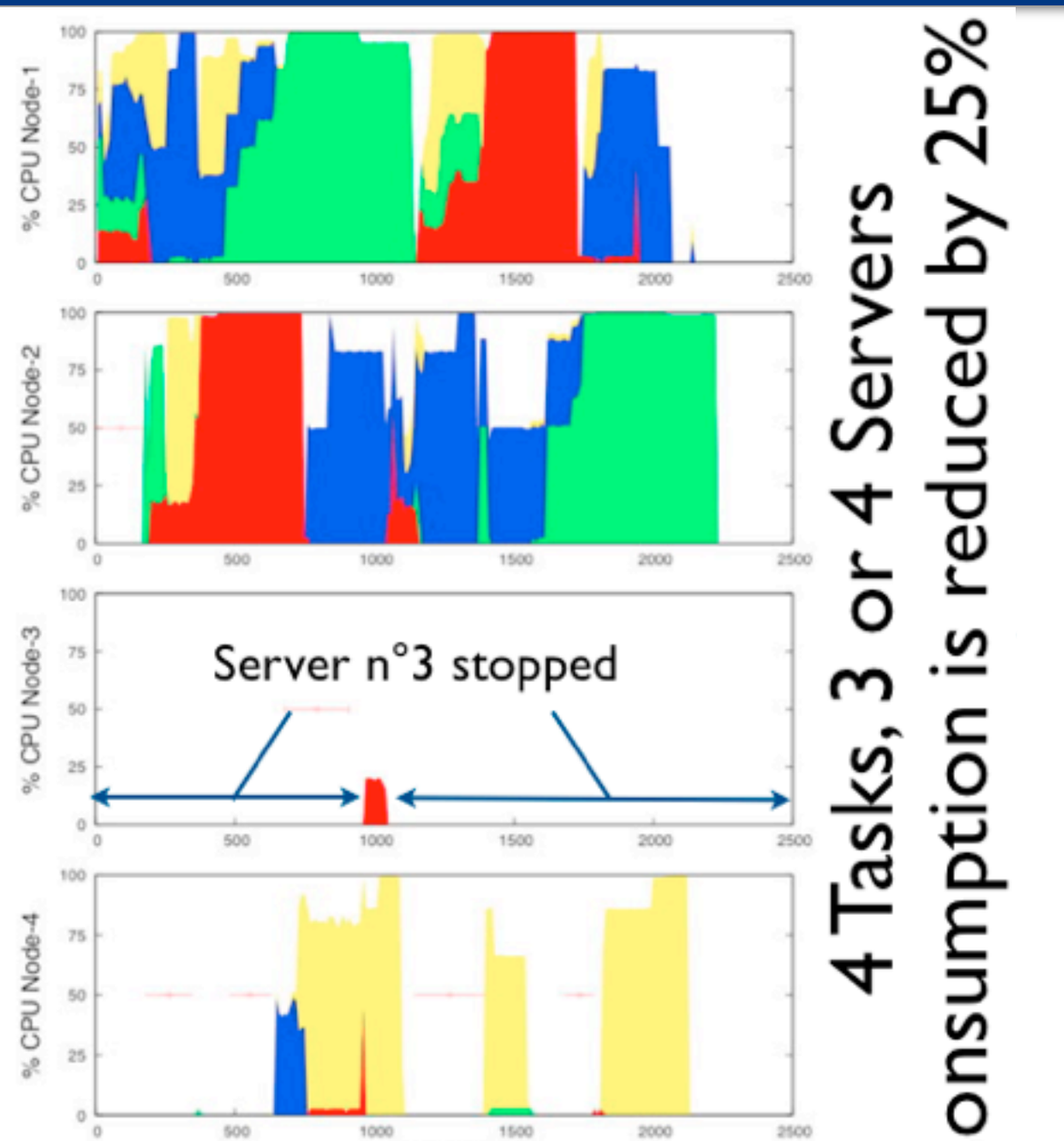
btrPlace: Principles



[XHPC'06]

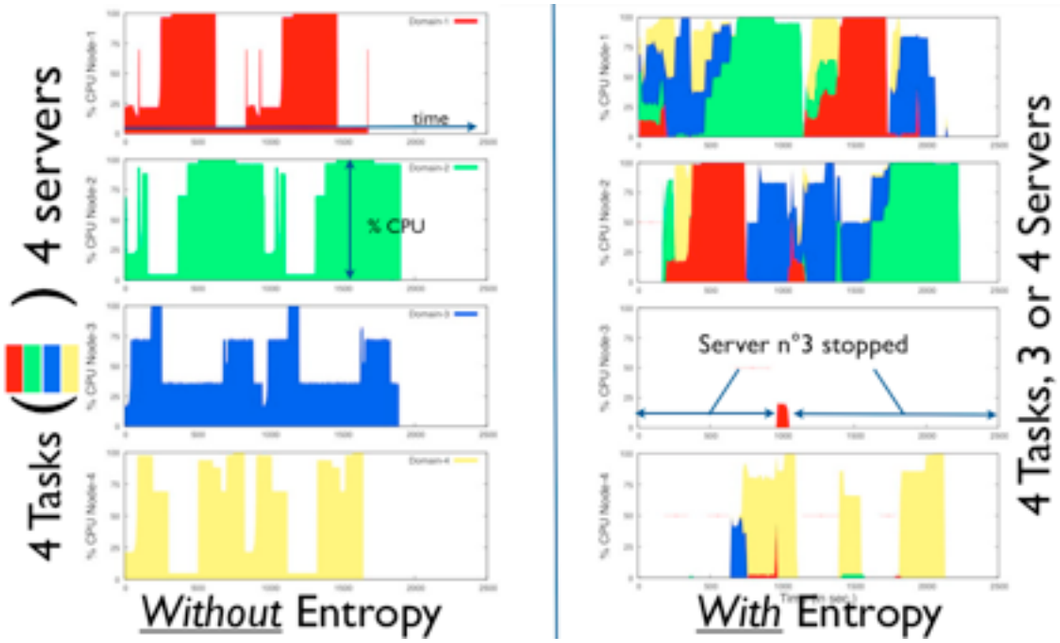
btrPlace: Optimizing the placement of virtual servers



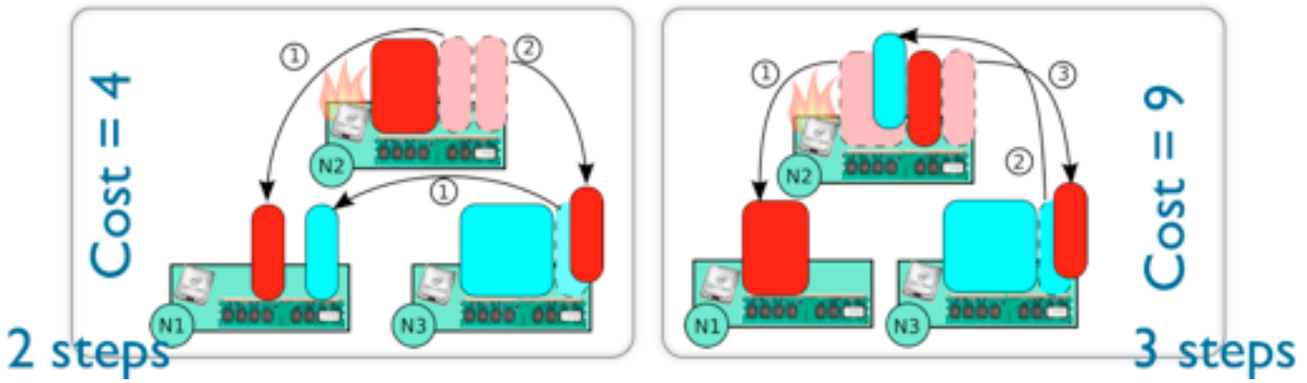


[XHPC'06]

btrPlace: Optimizing the placement of virtual servers



- **Determine an efficient reconfiguration plan (thanks to a cost function)**

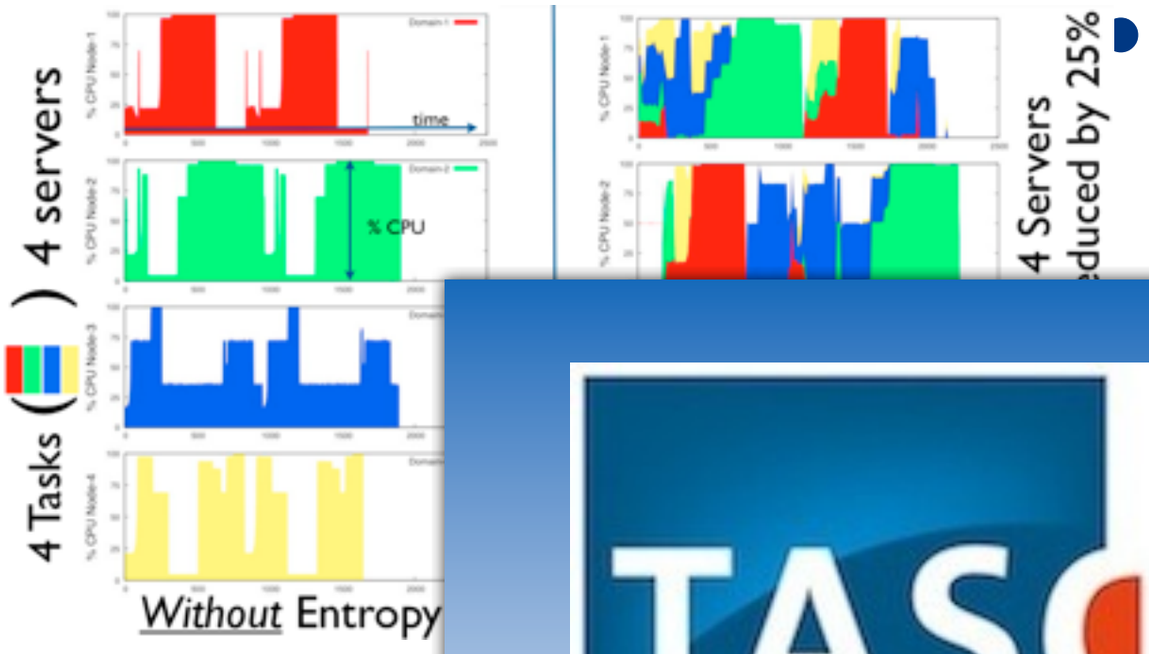


- **Administration and Application placement constraints must be considered**

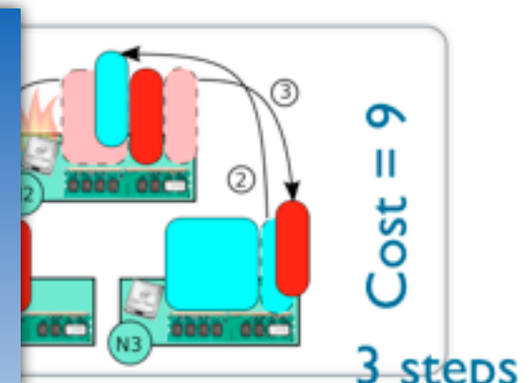
- **ban({VM1, VM2}, {N1, N2})**
  - prevents a set of VMs from being hosted on a given set of nodes
- **fence({VM1, VM2}, {N1, N2})**
  - forces a set of VMs to be hosted on a set of nodes
- **spread({VM1, VM2})**
  - ensures that the specified VMs are never hosted on the same node at the same time
- **latency({VM1, VM2}, {{N1, N2}, {N3, N4}})**
  - forces a set of VMs to be hosted on a single group of nodes.

[VEE'09,CFSE'11]

btrPlace: Optimizing the placement of virtual servers



- **Determine an efficient reconfiguration plan (thanks to a cost function)**



- **Adminis**
  Applicat
  constraints must be considered

a given set of nodes

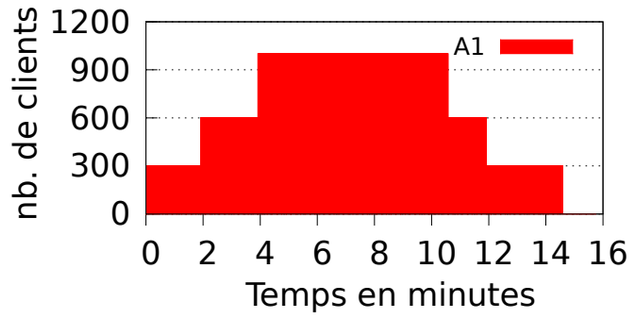
nodes

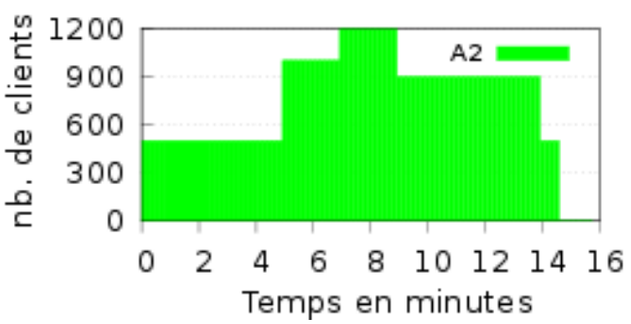
- **spread({VM1, VM2})**
  - ensures that the specified VMs are never hosted on the same node at the same time
- **latency({VM1, VM2}, {{N1, N2}, {N3, N4}})**
  - forces a set of VMs to be hosted on a single group of nodes.

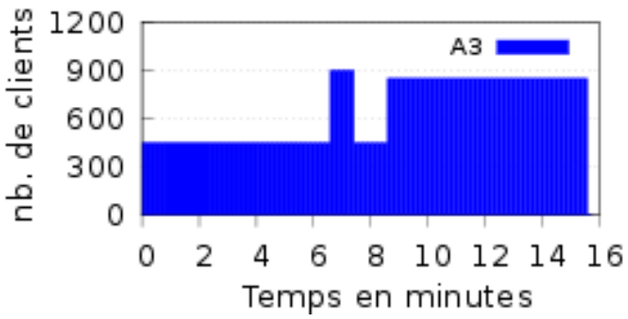
[VEE'09,CFSE'11]

Evaluations

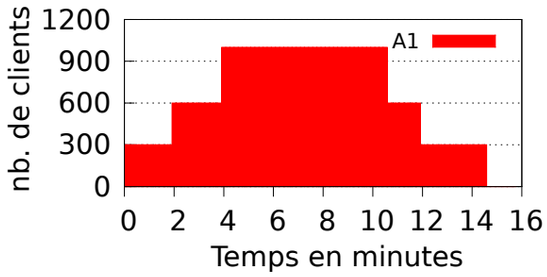
- **RUBiS : The three tiers of each instance of RUBiS are deployed as 7 VMs (2,3,2)**
- **3 applications**
- **21 nodes**

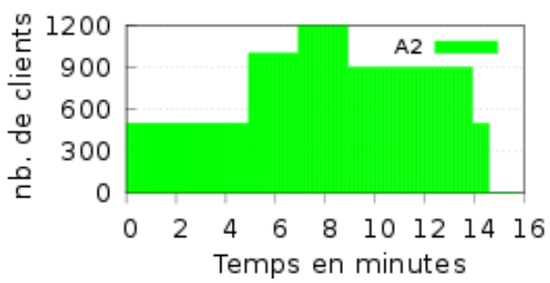


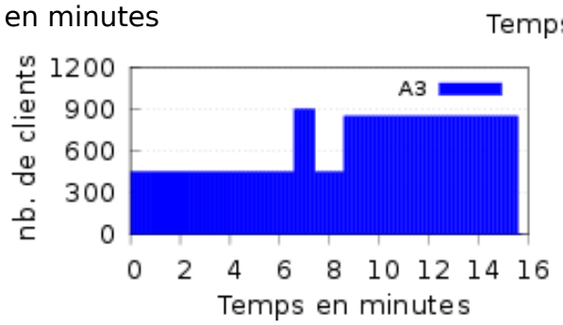


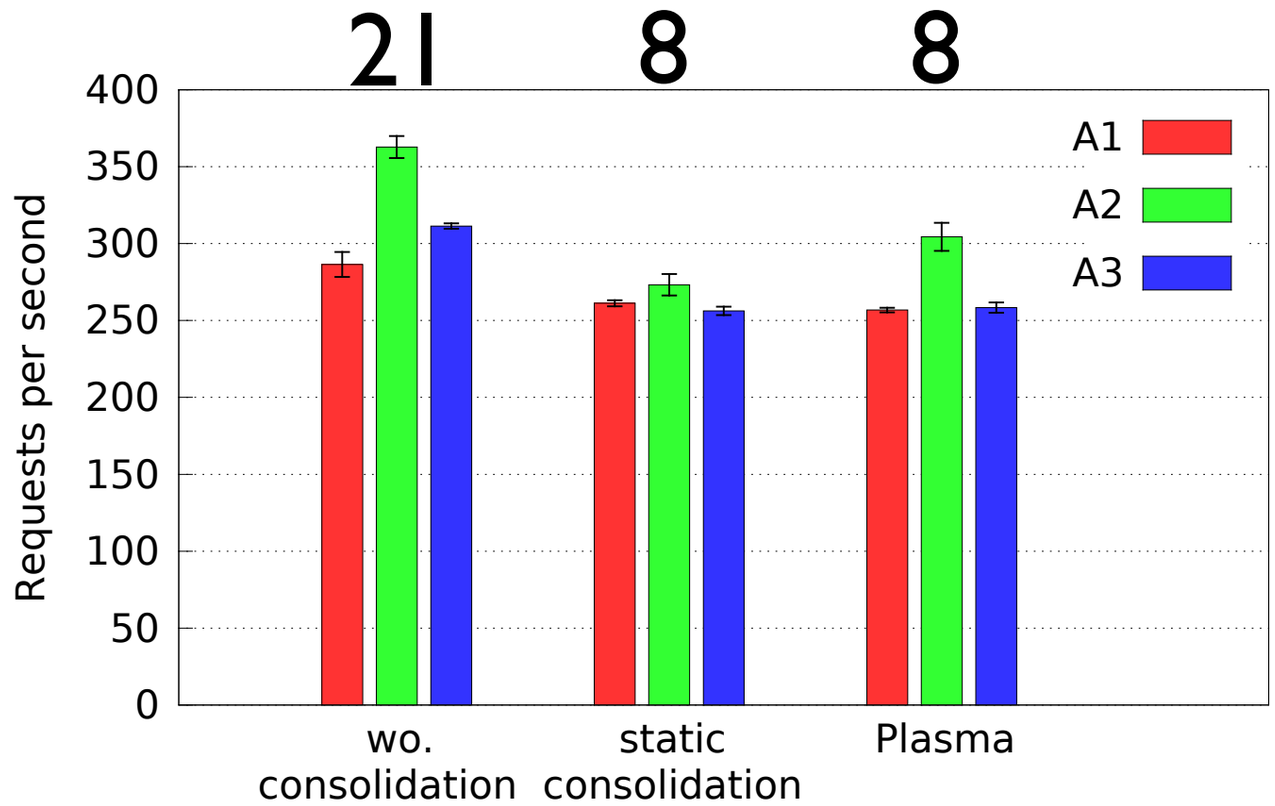


RUBiS Benchmark : Load spikes





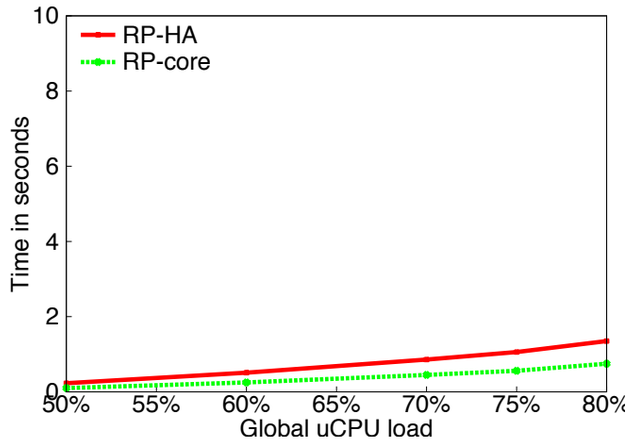


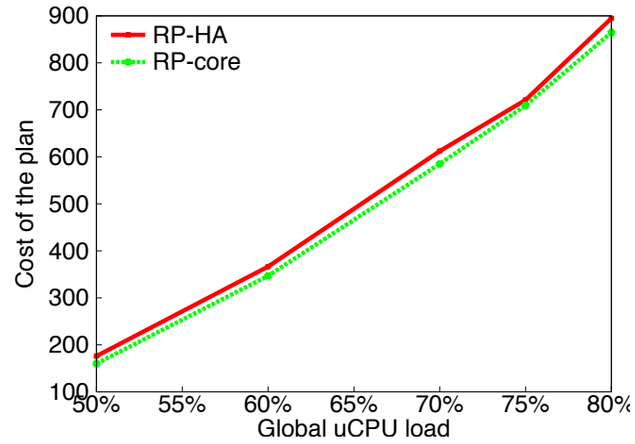


- **Improvement wrt. static consolidation (14.7% vs. 17.7%)**
- **About 12 reconfigurations (29 secs) per execution**
- **Longest reconfiguration: 10 migrations in 89 seconds**

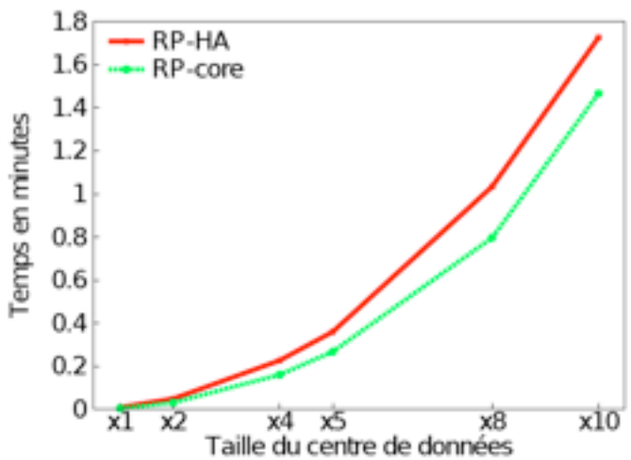
Impact of the global uCPU demand

- **Impact of placement constraints is not significant**

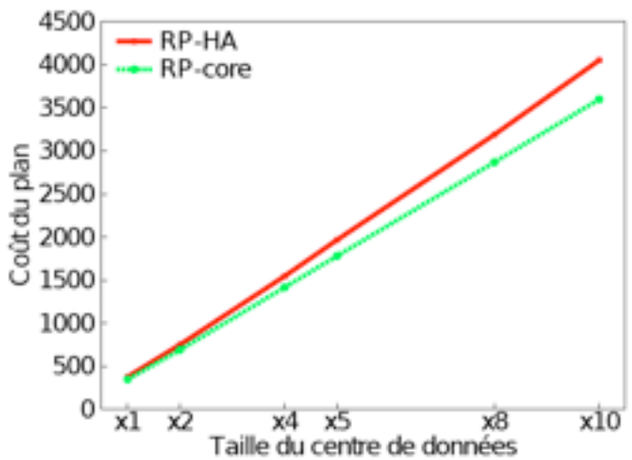




- **In practice**
  - place 1117 candidates VMs on 1980 nodes with 600 spread + 200 latency
  - schedule 475 actions



(a) Temps de résolution



(b) Coût des solutions

VM Management : next challenges

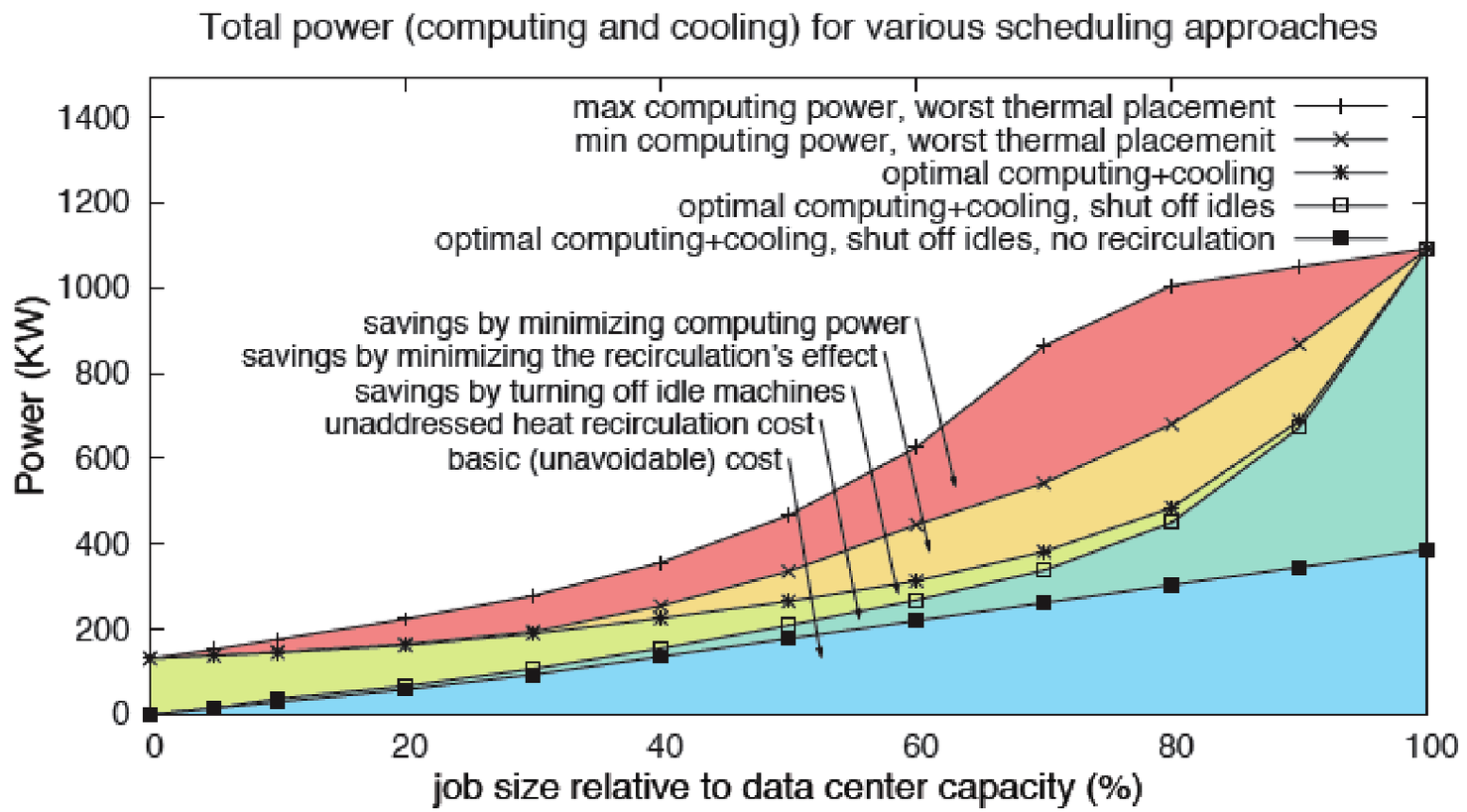
Constraints : hard/soft, flexibility, language ...

Placement : more abstraction, more optimization ...

Energy :

Thermal LoadBalancing

Sustainable Cloud



Distributed model : scalability, reactivity, fault tolerence ...

[CIT'09,DAIS'10, Cloud'10, ICAS'12]

VM Management : next challenges

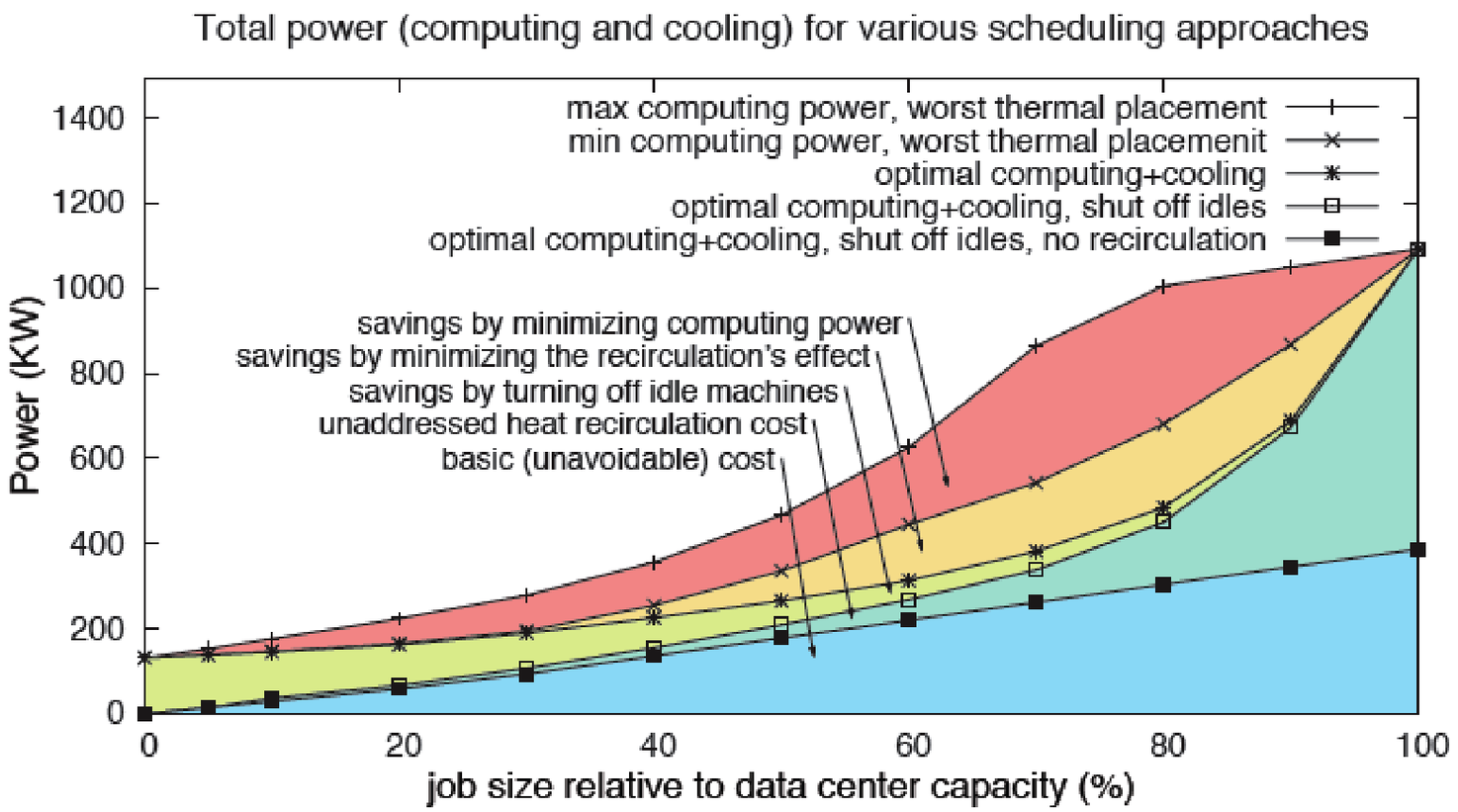
Constraints : hard/soft, flexibility, language ...

Placement : more abstraction, more optimization ...

Energy :

Thermal LoadBalancing

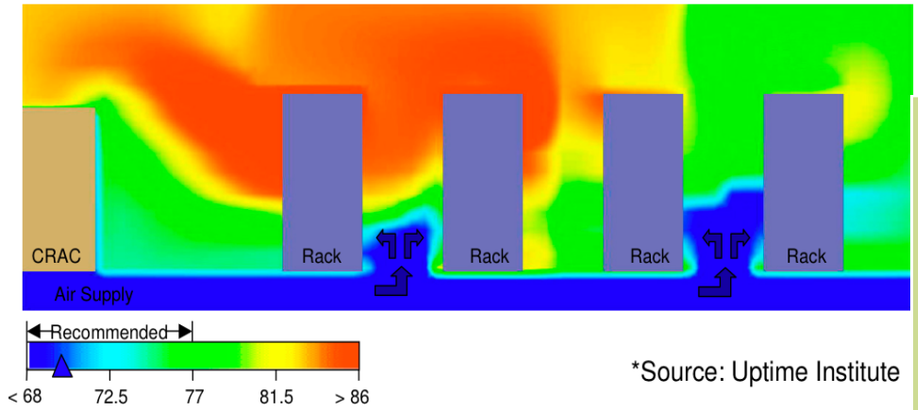
Sustainable Cloud



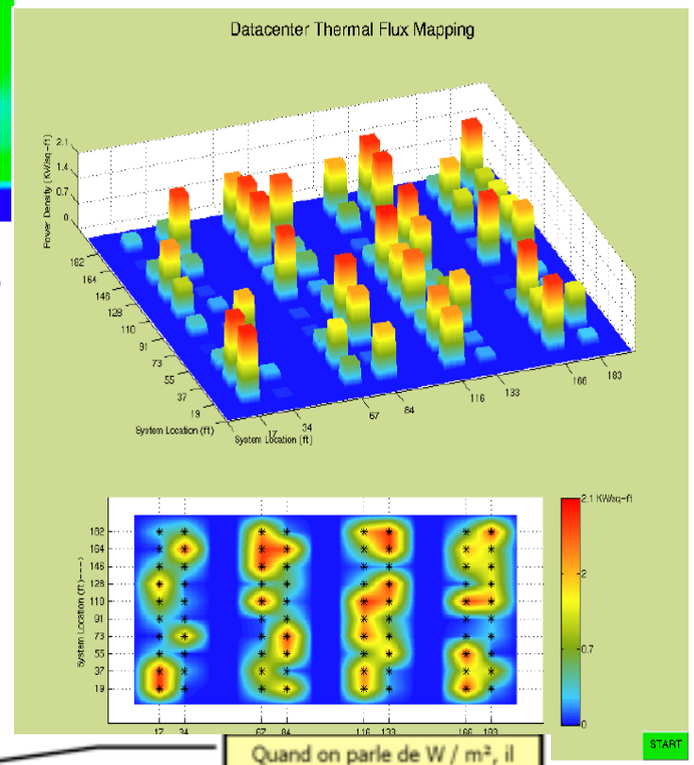
Distributed model : scalability, reactivity, fault tolerence ...

[CIT'09,DAIS'10, Cloud'10, ICAS'12]

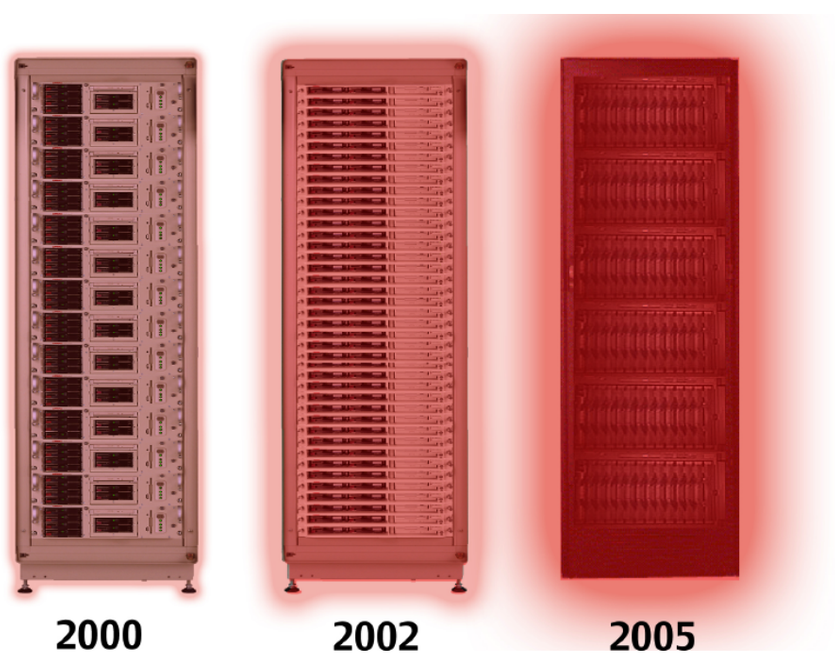
Des solutions ? L'équilibrage de charge thermique



*Source: Uptime Institute



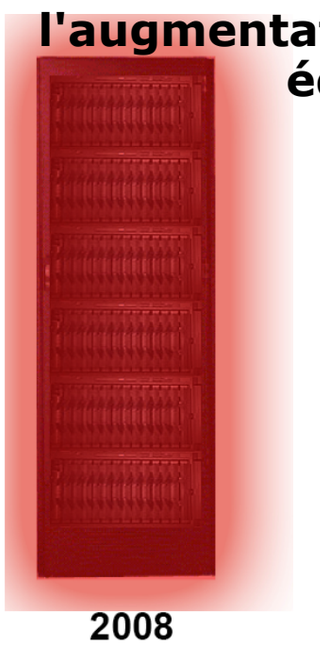
Le refroidissement ne suit pas l'augmentation de puissance des équipements



2000
28x2U Serveurs
2kW dissipation thermique

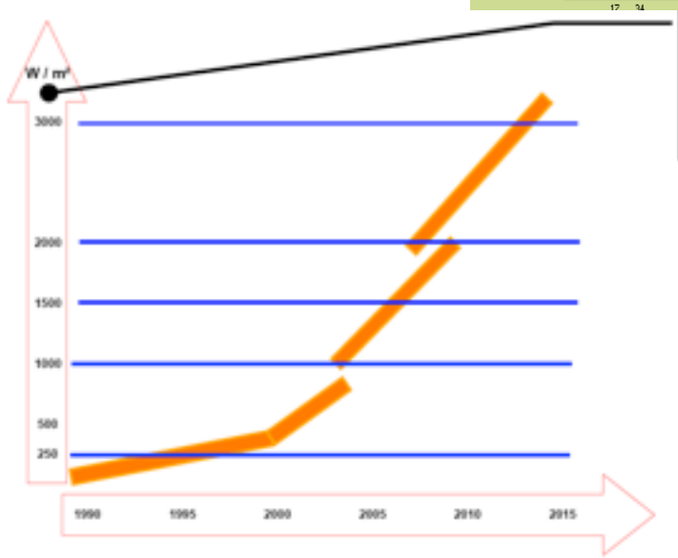
2002
42x1U Serveurs
6kW dissipation thermique

2005
6 BladeCenters
24kW dissipation thermique



2008
6 BladeCenters
30kW dissipation thermique

Source: Emerson Network Power/Liebert



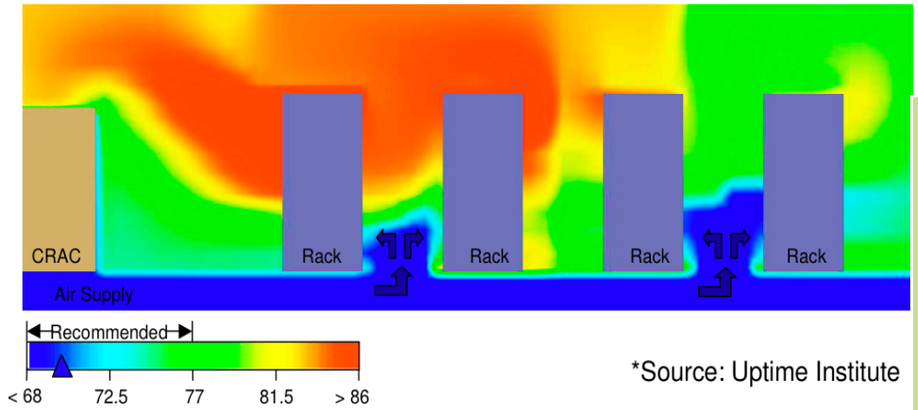
Quand on parle de W / m², il s'agit de la puissance en courant HQ disponible dans la salle, ramenée à la surface de la salle.

Une densité de 3000 W / m² peut correspondre à une salle équipée en totalité de baies de 10 kw, ou à un mix entre des baies de 4 kw,10 kW jusqu'à 30 kW

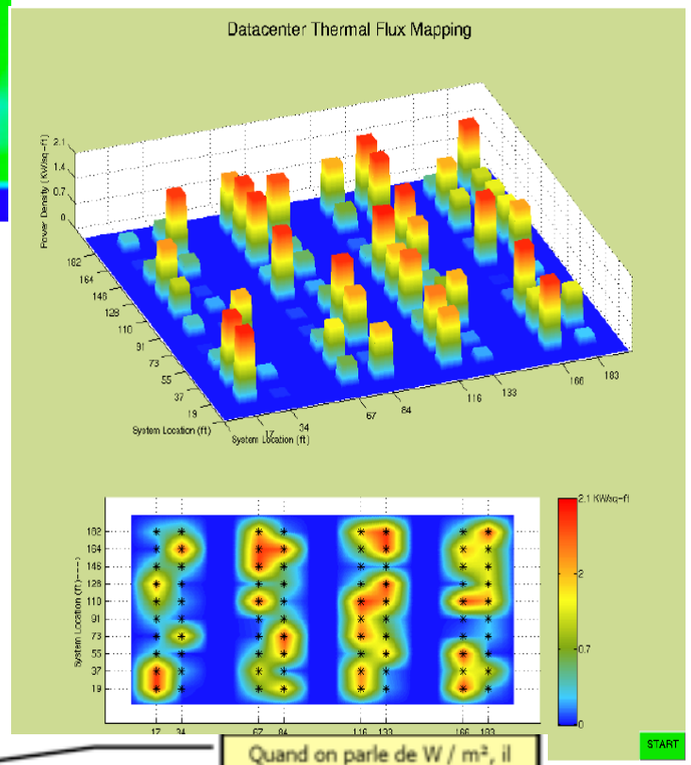
Sun

Des solutions ? L'équilibrage de charge thermique

Confinement et monter en T°

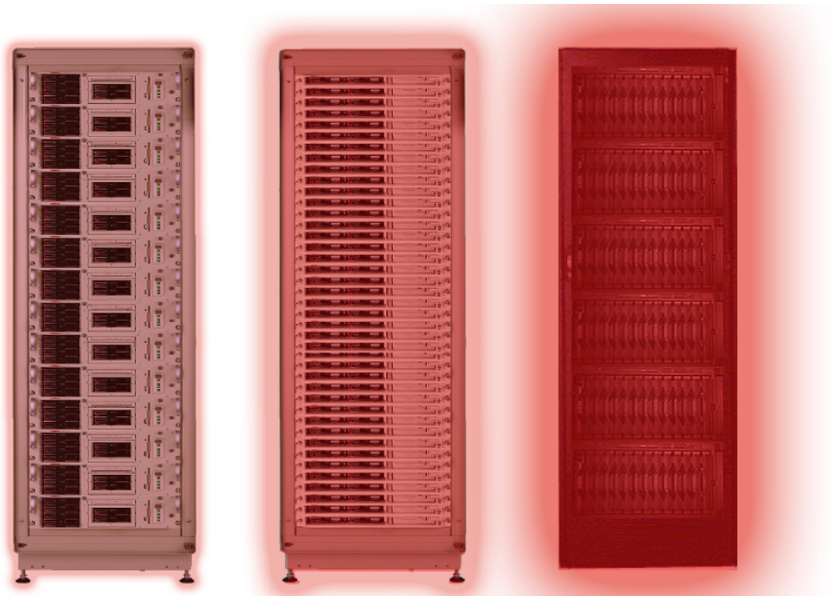


*Source: Uptime Institute



Sun

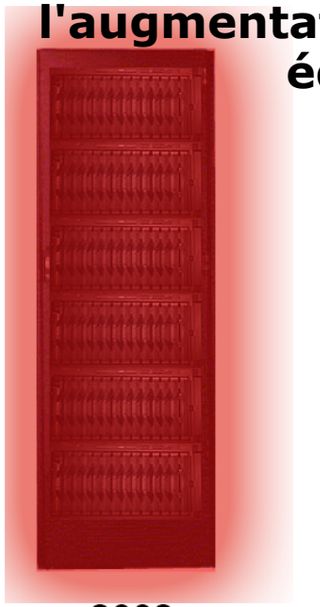
Le refroidissement ne suit pas l'augmentation de puissance des équipements



2000
28x2U Serveurs
2kW dissipation thermique

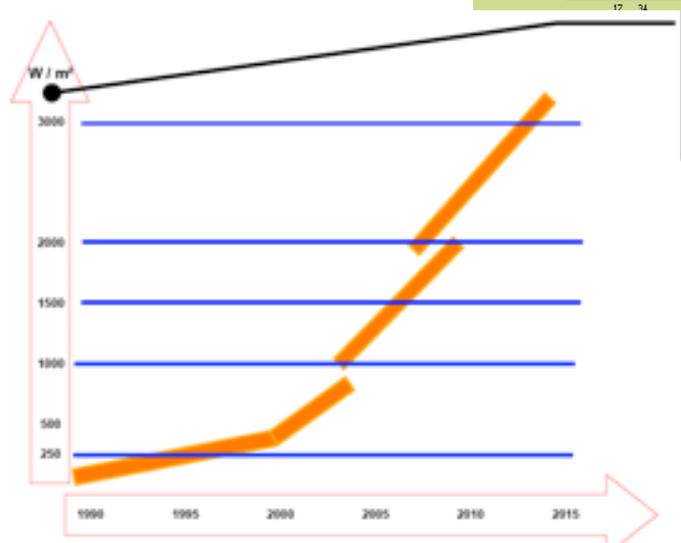
2002
42x1U Serveurs
6kW dissipation thermique

2005
6 BladeCenters
24kW dissipation thermique



2008
6 BladeCenters
30kW dissipation thermique

Source: Emerson Network Power/Liebert



Quand on parle de W / m², il s'agit de la puissance en courant HQ disponible dans la salle, ramenée à la surface de la salle.

Une densité de 3000 W / m² peut correspondre à une salle équipée en totalité de baies de 10 kw, ou à un mix entre des baies de 4 kw,10 kW jusqu'à 30 kW

Thermal Load Balancing : Heat and CRAC

- **Each server specifies a maximum temperature of inlet air Tmax, at a constant rate applied to CARC.**

  Tmax(j) max inlet temp for server j

- **Heat linear model.**
  - The server i will heat the server j I(i, j) °C by W consumed.

    Ci = power consumption for server i
    1) Timp(j) = Σ (server i) Ci * I (i, j)
  - The inlet temperature for server j is

    2) Tin(j) = Tcrac+ Timp(j) < Tmax(j)
  - For the server j the CRAC must be

    3) Tcrac < Tmax(j) - ∑(serveur i) Ci * I(i,j)
  - For multiple servers, the CRAC must be

    4) Tcrac = min((Tmax(j) - ∑(serveur i) Ci * I(i,j)) for all j)
  - we need to maximise Tcrac
- **First results ...**

Cool-IT

- **Parteners :**
  - **Bull SAS**, Splitted-Desktop Systems (SDS), INRIA, CEA, AVOB, ATRIUM DATA, SINOVIA, WILLELEC, EURODECISION
- **Main Focus**
  - Cooling system, transfer and transport of heat dissipated

    Water cooling inside the server
  - Electrical system, integration of uninterruptible power supply (UPS)

    By using supra-capacity
  - Collection of energy data

    system, power etc.
  - Control of power management

    Controling server (DVFS) and task placement
  - Test and evaluation
- **Control of power management**
  - EuroDecision : Task placement with processor frequency variability

Conclusion

btrCloud

- btrPlace : Configurable consolidation manager
- PPC appraoch

  scalable to datacenter with up to 2000 nodes/ 4000 VMs
  placement constraints do not impact the solving process
- Constraints and optimize

  Server and CRAC

Futures works

- new placement constraints for new concerns (currently 10 constraints)
- improvment of the scalability using partitioning
- Hard/soft placement constraints.

Questions ?

Jean-Marc.Menaud@mines-nantes.fr