Grenoble INP ENSIMAG

Cloud : performances et consommation

Plan

Éléments de performance sur un noeud

Consommation d’énergie

Pourquoi parler de performances ?

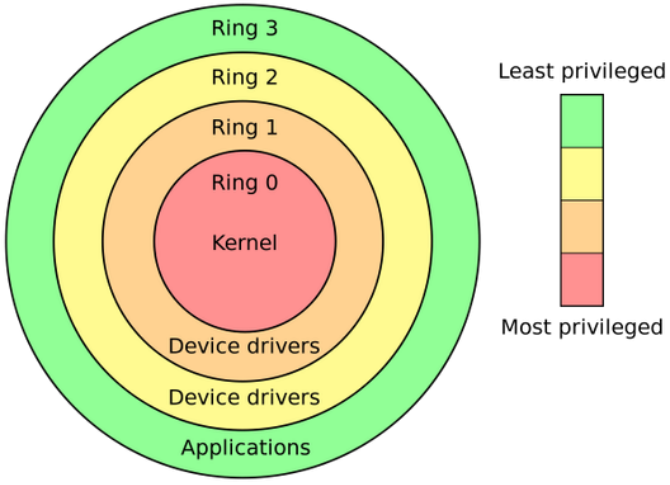
Les performances pour la cliente sont fonction de qu'elle paye

Du point de vue du fournisseur de cloud

Virtualisation

Pour le processeur

- code utilisateur : pas de surcoût
  - code exécuté directement
- code système : surcoût quant appel nécessite passage par hyperviseur
  - particulièrement si l'hyperviseur s'exécute dans un ring différent du noyau



Coût de l'accès mémoire

En conditions normales translation dans la TLB
traduction de l'adresse virtuelle de la VM jusqu'à l'adresse physique de l'hôte

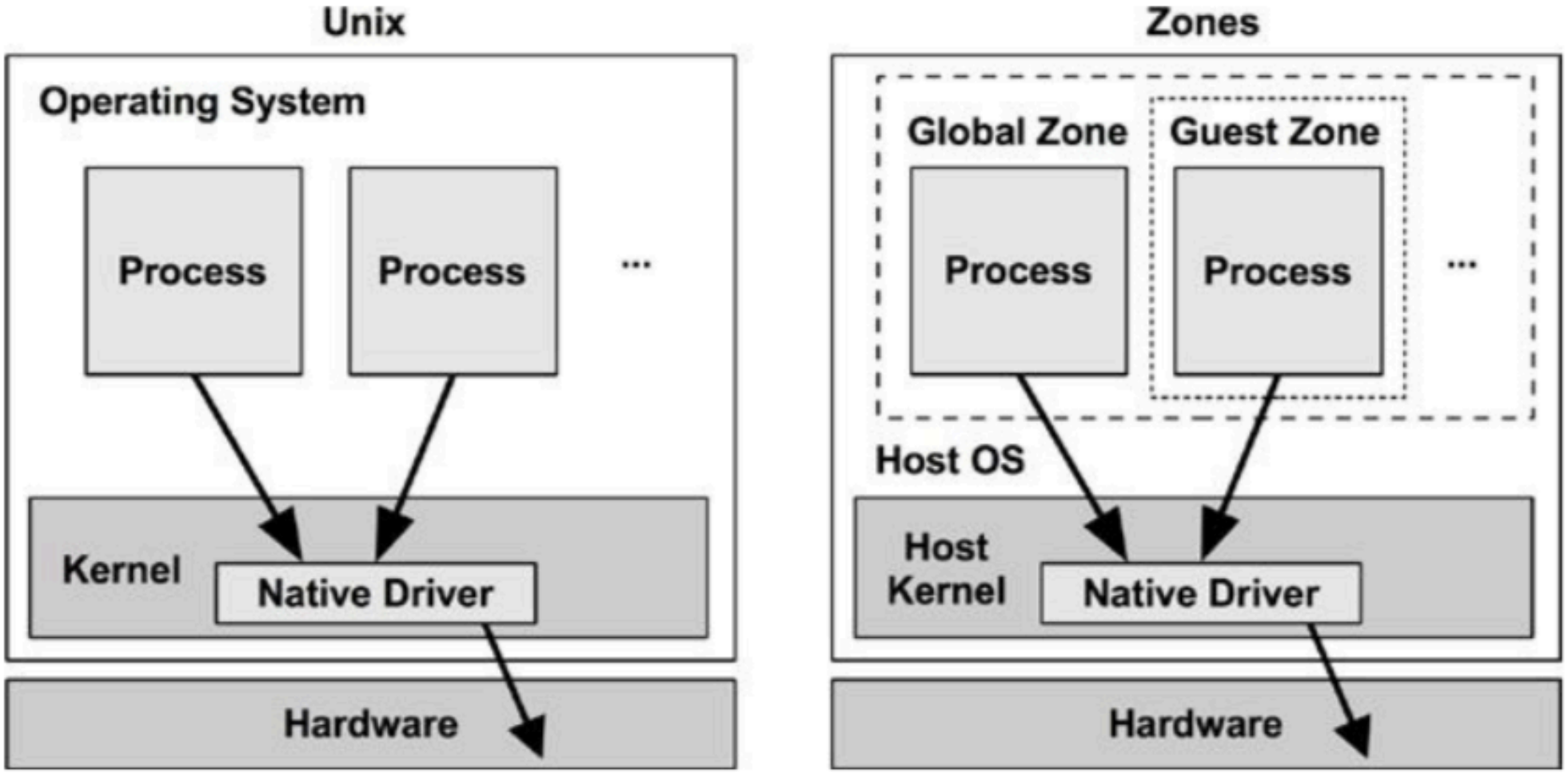
En cas de défaut de page ou de défaut de TLB, 2 étapes

1. virtuel vers physique de la VM par l'OS de la VM
2. physique VM vers physique de l'hôte par l'hyperviseur

Certains processeurs ont des caches spécifiques pour cela (*nested paging*, EPT/NPT)

Circuit des I/O

Pas de surcoût, circuit identique



Coût du partage

Les machines physiques sont partagées entre les machines virtuelles

Impact sur les performances

1. Cache CPU peut être flushé plus souvent
2. CPU interrompu pour gérer les interruptions des processus des autres machines virtuelles
3. Ressources partagées (réseau, I/O)

politique de l'hébergeur s'applique

Partage des ressources entre VM

Exemple de Joyent

Resource	Priority	Limit
CPU	FSS	caps
Memory capacity	rcapd/zoneadmd	VM limit
File system I/O	ZFS I/O throttling	—
File system capacity	—	ZFS quotas, file system limits
Disk I/O	see file system I/O	—
Network I/O	flow priority	bandwidth limits

CPU

Client décide s'il borne ce qu'il veut utiliser ou utilise toute la machine (bursting)
- risque de coût induit
- si une autre VM gourmande arrive il sera pénalisé

politique de partage : fair-share scheduler
- exprimé en terme de pourcentage de ressource utilisée
  - lié à la quantité de mémoire payée

Mémoire

Physique
- taille maximale fixée par le système en fonction du contrat
- modification du démon de gestion des pages libres
  - pour swapper en avance les pages quand dépassement

Virtuelle
- théoriquement illimitée
- en pratique `malloc` renvoie erreur si 2*mémoire physique

Stockage

Espace disponible

1. volume virtuel de taille fixée
2. si FS partagé utilisation de quotas

Volume d'I/O

- problème de la perturbation des autres machines
- *throttling* contrôlé par la couche d'I/O
  - locale au noeud
  - injecte des délais artificiels pour ralentir les processus utilisateurs

Plan

Éléments de performance sur un noeud

Consommation d'énergie

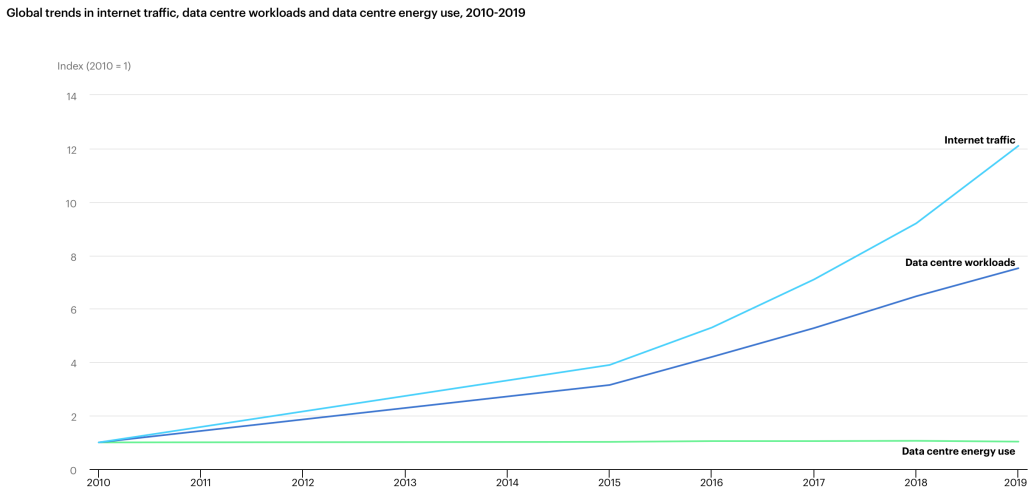
Peu de données sur les datacenters mais travaux sur les centres de calcul

Quelques repères…

2020 : 400 TWh de consommation cumulée pour les data centers et les réseaux mondiaux

- 2% de la consommation électrique mondiale
- ~2% des émissions de gaz à effet de serre (~aviation)

Pas de prise en compte du coût environnemental de production des équipements



https://www.iea.org/reports/data-centres-and-data-transmission-networks

Le Green500

Classement des machines les plus puissantes et les moins gourmandes

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	301	**MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect, Preferred Networks Preferred Networks Japan	1,664	2,181.2	55	39.379
2	291	**SSC-21 Scalable Module** - Apollo 6500 Gen10 plus, AMD EPYC 7543 32C 2.8GHz, NVIDIA A100 80GB, Infiniband HDR200, HPE Samsung Electronics South Korea	16,704	2,274.1	103	33.983
3	295	**Tethys** - NVIDIA DGX A100 Liquid Cooled Prototype, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 80GB, Infiniband HDR, Nvidia NVIDIA Corporation United States	19,840	2,255.0	72	31.538
4	280	**Wilkes-3** - PowerEdge XE8545, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 80GB, Infiniband HDR200 dual rail, DELL EMC University of Cambridge United Kingdom	26,880	2,287.0	74	30.797
5	30	**HiPerGator AI** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR, Nvidia University of Florida United States	138,880	17,200.0	583	29.521

Grenoble INP ENSIMAG

Progression vs 2021

39 -> 62 : +63%

Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	301	**MN-3** - MN-Core Server, Xeon Platinum 8260M 24C 2.4GHz, Preferred Networks MN-Core, MN-Core DirectConnect, Preferred Networks Preferred Networks Japan	1,664	2,181.2	55	39.379
2	291	**SSC-21 Scalable Module** - Apollo 6500 Gen10 plus, AMD EPYC 7543 32C 2.8GHz, NVIDIA A100 80GB, Infiniband HDR200, HPE Samsung Electronics South Korea	16,704	2,274.1	103	33.983
3	295	**Tethys** - NVIDIA DGX A100 Liquid Cooled Prototype, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100 80GB, Infiniband HDR, Nvidia NVIDIA Corporation United States	19,840	2,255.0	72	31.538
4	280	**Wilkes-3** - PowerEdge XE8545, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 80GB, Infiniband HDR200 dual rail, DELL EMC University of Cambridge United Kingdom	26,880	2,287.0	74	30.797
5	30	**HiPerGator AI** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Infiniband HDR, Nvidia University of Florida United States	138,880	17,200.0	583	29.521

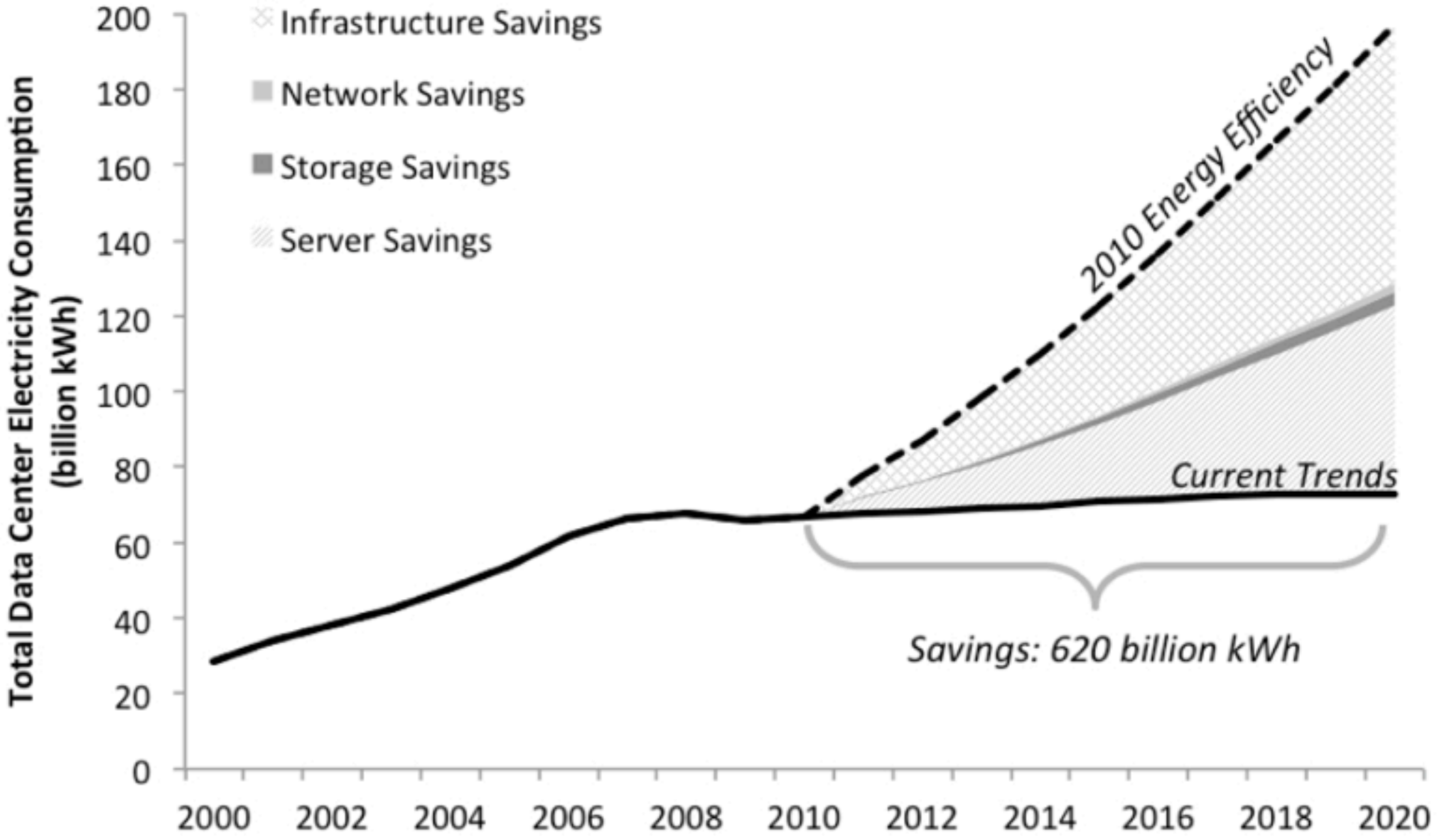
Leçon du Green500

Clés pour la baisse de la consommation

- processeurs économes : les GPUs ont un très bon rendement énergétique
  - mais ne sont pas des processeurs généralistes

“I am very much in favor of renaming it ‘high-efficiency computing’ instead of ‘high-performance computing since we really are the community that knows how to design fast algorithms, we know how to make really efficient implementations, how to efficiently parallelize algorithms – which is always necessary – and how to choose the right hardware to do it.”

Importance des mesures prises



Leviers utilisés

Développement des machines virtuelles
  diminution du nombre total de serveurs

Développement de très grands data centers
  plus d'investissements sur l'infrastructure externe

Hausse de la température
  développement du free cooling

Amélioration de la gestion des ressources

Coût du refroidissement

Pendant longtemps typiquement facteur de 1,5 (jusqu'à 2)
  maintenant 1,1 à 1,3
    sur de grandes architectures ordonnanceur peut choisir le site le plus frais
      local
      global

Construction de data center dans des zones froides
  Avec volonté de produire son énergie (géothermie)

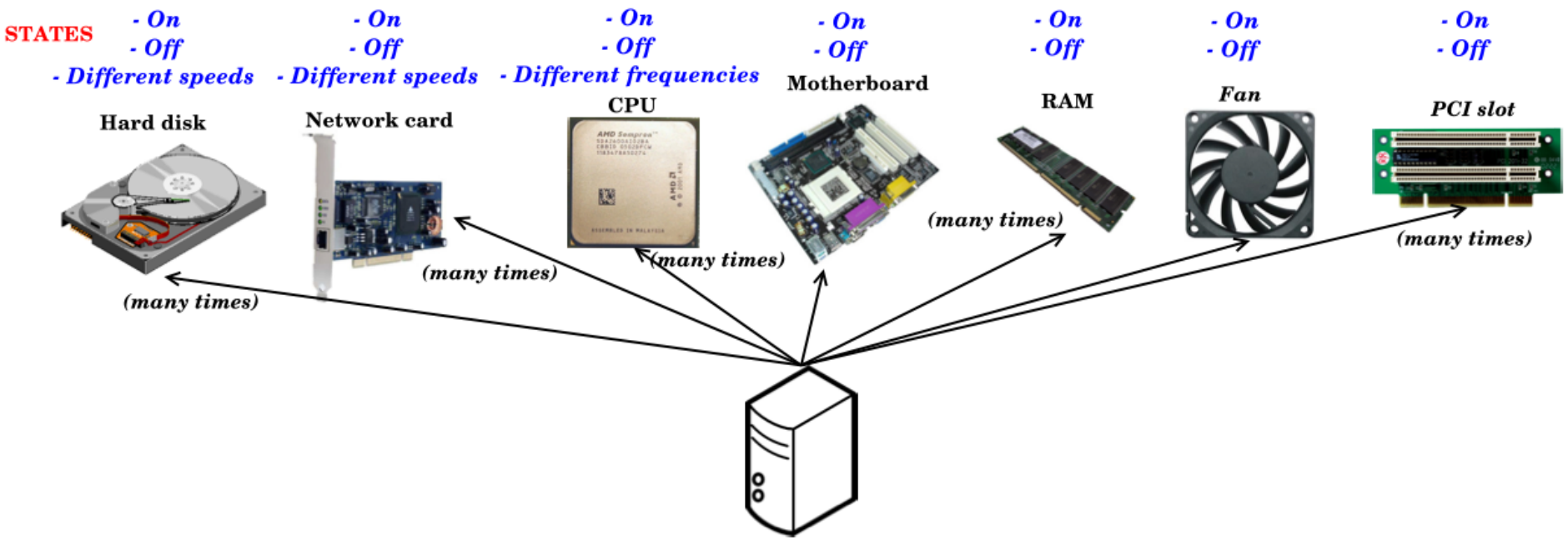
Consommation

La consommation d'énergie est le coût le plus élevé pour une infrastructure

Component	Peak power	Count	Total	Percentage
CPU	40 W	2	80 W	37.6 %
Memory	9 W	4	36 W	16.9 %
Disk	12 W	1	12 W	5.6 %
PCI slots	25 W	2	50 W	23.5 %
Motherboard	25 W	1	25 W	11.7 %
Fan	10 W	1	10 W	4.7 %
System total			213 W	

Table I. Component peak power breakdown for a typical server .

composants d'un noeud



Adaptation de la consommation

Dynamic voltage and frequency scaling
- permet d'ajuster fréquence pour diminuer la consommation
  - P-state, sous Linux en utilisant cpufreq
  - c-state état inactif, plus élevé sommeil plus profond

Difficulté de déterminer quant il vaut mieux laisser actifs
- nécessite de connaître le futur

Extinction de composants mémoire
- gestion de la mémoire adaptée aux composants
- impacte les performances!

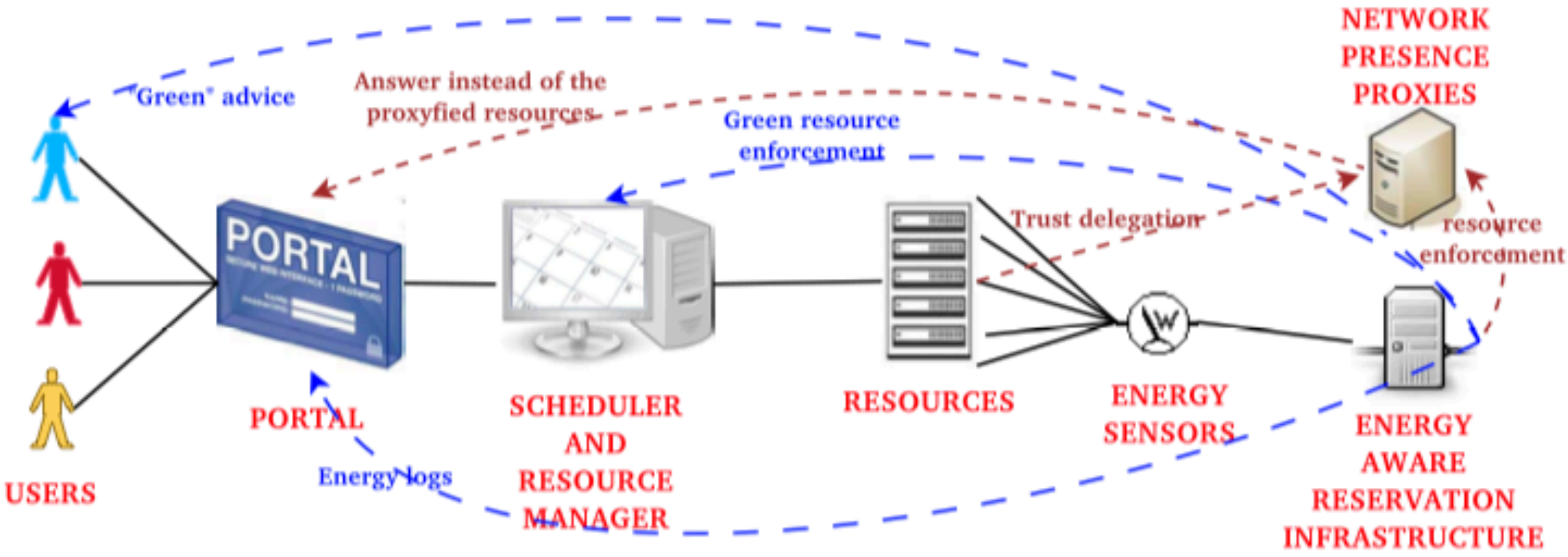
Work load consolidation

Minimisation du nombre de machines allumées
- Construction du plan d'ordonnancement avec ce critère
  - Apprentissage automatique
- Suivant les applis peut être contreproductif
  - Utiliser plus de noeuds lents consomme moins que des rapides
- Minimiser aussi le nombre de démarrages de machines

Algorithmes clairvoyants : doivent connaître le futur et/ou être capable de le prévoir
- Peuvent aussi demander des informations aux utilisateurs

Green-NET

Solution pour une meilleure utilisation des ressources sur des grappes de calcul



Grenoble INP ENSIMAG

Green-NET

Implication des utilisateurs

- souplesse dans les demandes de moyens
- acceptation d'une baisse de performances pour être plus green

Table 1: Percentages of energy gain vs. performance degradation (execution time).

Method	Energy/Performance		
	HDD Spin	CPU Freq	HDD Spin + CPU Freq
EP	2.5 / 0	10.3 / -18.9	12.2 / -20.5
SP	1.6 / 0.3	8.5 / -1.3	10.2 / -1.5
BT	2 / -0.4	9 / -5.4	10.4 / -5.5
LU	2.2 / 0.2	9.5 / -7.6	11.5 / -10.8
CG	2 / -0.13	8.2 / -1.4	10 / -3.1
IS	1.4 / 1.5	6.4 / -1.5	10 / -7.2
MG	1.2 / -1.1	8.2 / -0.5	9.8 / -3.4
Overall	1.8 / 0.05	8.5 / -5.2	10.5 / -7.4

Green-NET

Ordonnanceur green

- utilise les travaux déjà soumis et les traces précédentes pour prédire les périodes d'inactivité
  - extinction anticipée des noeuds
- économie d'énergie de 44% en moyenne sur 4 gros sites

Virtualisation et consommation

Deux niveaux de décision

- Macro (ensemble de machines) : placement des VMs sur les noeuds physiques pour minimiser le nombre de noeuds actifs
  - migration possible au sein d'un datacenter

- Micro (une machine) : consommation limite à ne pas dépasser sinon actions prises
  - ralentissement (ou pause) d'une VM
    - extinction de coeurs
  - diminution de la fréquence du processeur
  - interactions possibles avec le système invité

Consommation des réseaux

Difficulté d'avoir une valeur globale

- Deux estimations
  - Internet à 100Mbps -> 1% consommation pays développé
  - Internet à 1Gbps -> 4% consommation

	power consumption (W)	number of devices	overall consumption (GWh/year)
Home	10	17,500,000	1,533
Access	1,280	27,344	307
Metro/Transport	6,000	1,750	92
Core	10,000	175	15
	Overall network consumption		1,947

Bilan et leçons

Big is beautiful
- faire le maximum de calcul dans des datacentre
- pour les fournisseurs être frugal est aussi être économe
  - SaaS plus optimisable que IaaS

Poids des transferts
- difficile à évaluer de manière individuelle

Restriction des usages
- thème très délicat
- qui décide et comment ?
- des réflexions internes aux communautés qui commencent à prendre (IA)

Quid des externalités positives ?
- Enjeu : prolonger la durée de vie des équipements (rôle du logiciel)

Le futur

Quels gains sont encore possibles ?

- plus l'architecture est grande, meilleures sont les possibilités
  - car vue plus globale sur les ressources et les travaux
  - mais sont aussi des SPOFs

Changement d'architectures

- ARM vs Intel vs AMD
  - ARM chez AWS
    - pour le support des micro services en containers
    - avec un processeur conçu in-house, jusqu'à 45% d'économie
    - et pas de symetric multithreading
- quantique ?