Contributions à la flexibilité et à l'efficacité énergétique des systèmes distribués à grande échelle

Soutenance d'Habilitation à Diriger les Recherches

Laurent Lefèvre
Inria - Equipe Avalon – Laboratoire LIP - Ecole Normale Supérieure de Lyon
laurent.lefevre@inria.fr

14 Novembre 2013, Lyon

Plan

- Introduction
  - µCV
  - Contributions/ Projets
- Flexibilité
  - Réseaux dynamiques et programmables : L'aventure Tamanoir
  - Nouveaux équipements et services : 1 focus
- Efficacité energétique
  - Mesurer et comprendre
  - Vers des ordonnanceurs et des nuages verts
  - Améliorer l'efficacité énergétique du HPC avec ou sans connaissance des services et des applications
- Conclusions et perspectives

Introduction

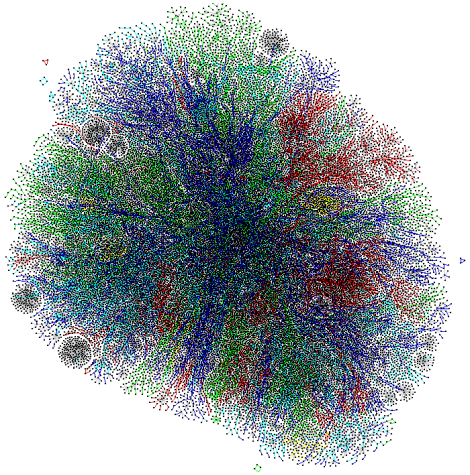
Systèmes distribués à grande échelle

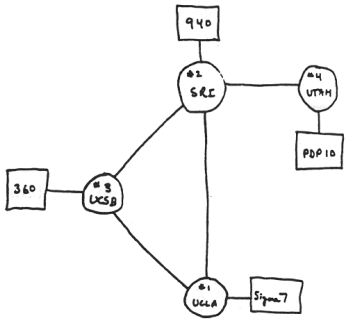
- Hétérogénéité des equipements terminaux
- Des tuyaux, et au bout de grosses infrastrucutures
  - DataCentres /HPC
  - Clouds / Grilles
  - Réseaux
  - Pas mobile
- Plus imposant, plus de services, plus d'utilisateurs

Systèmes distribués à grande échelle

- **Réseaux /Internet**
- Arpanet (1969) -> Internet (2013)
- 4 sites (69) -> 213 (81) -> 5K (86) -> 1

  Milliard de machines connectées (2013)

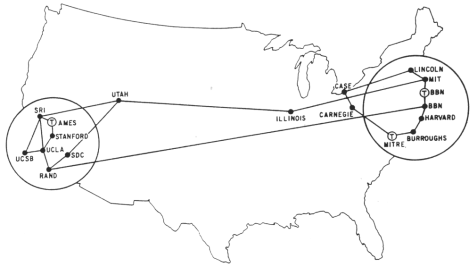




THE ARPA NETWORK

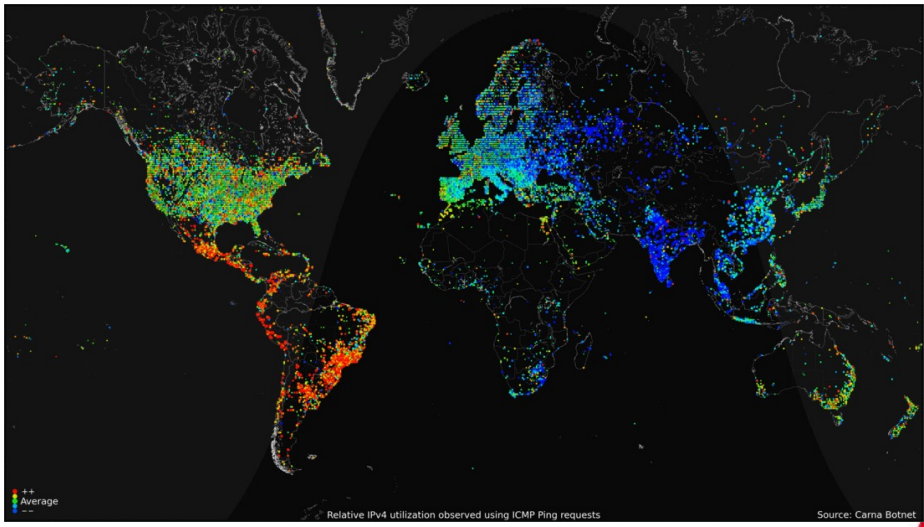
DEC 1969

4 NODES



MAP 4 September 1971

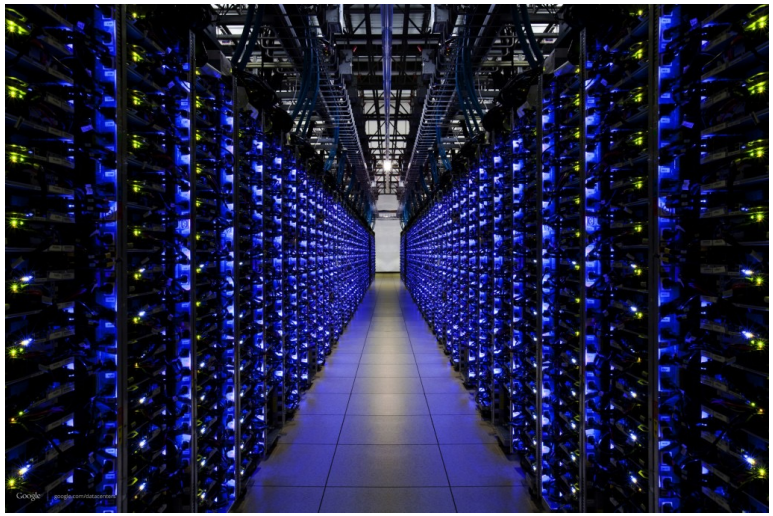
FIGURE 6.2 Drawing of 4 Node Network (Courtesy of Alex McKenzie)



Systèmes distribués à grande échelle

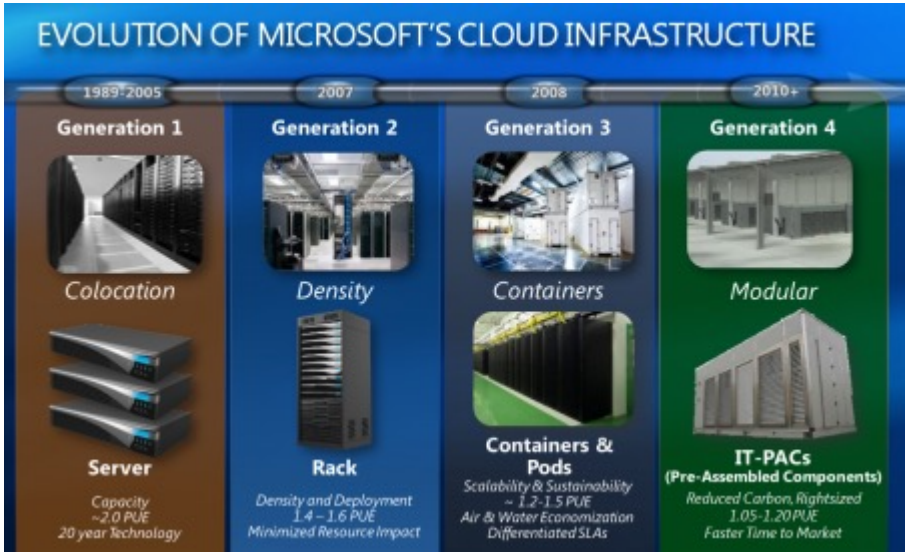
- **DataCenters**

- Google (1997 <10) -> Google (2000 - 25K) -> Google (2013 – 1M)



Systèmes distribués à grande échelle

- Clouds : des ensembles de datacenters
- Amazon, Azure…
- IaaS, Paas, HPCaaS…



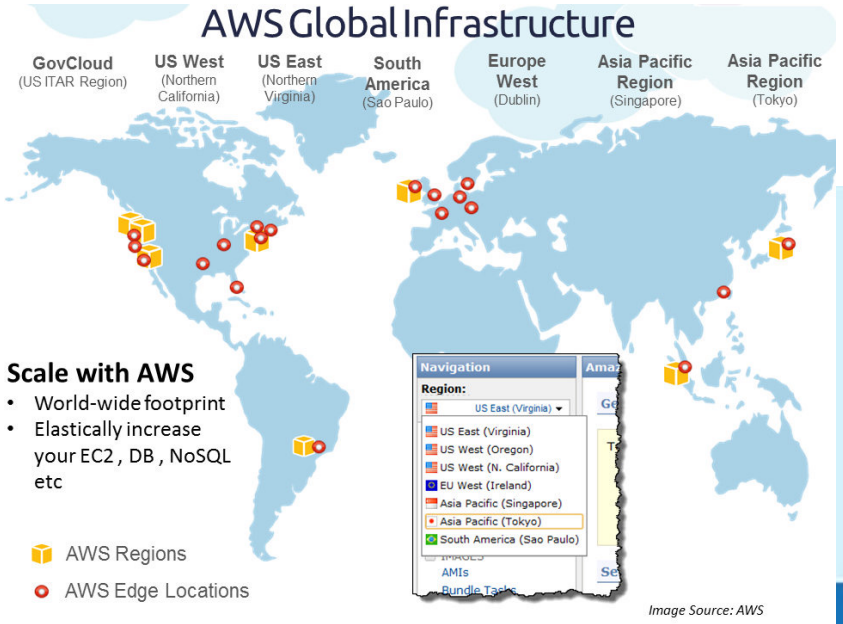
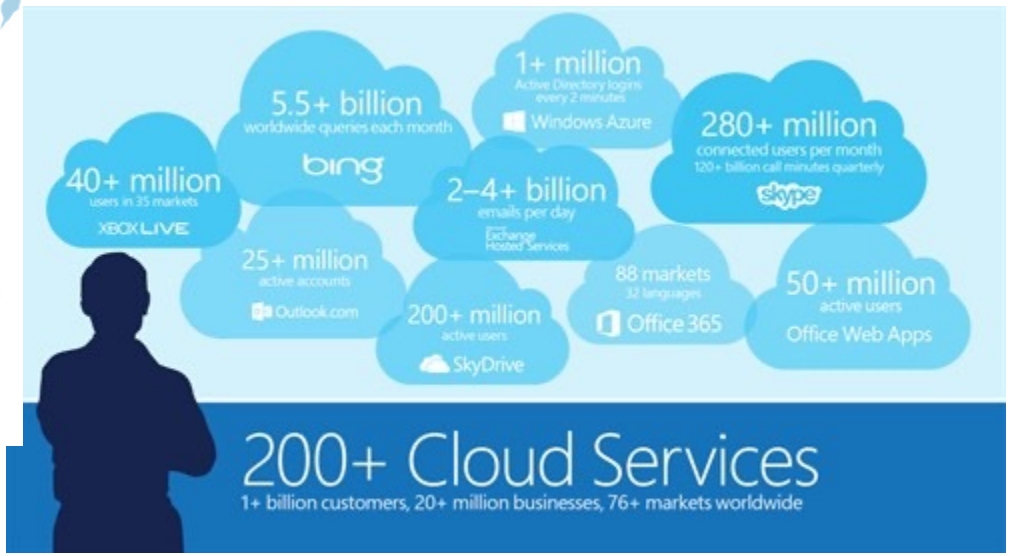


Image Source: AWS

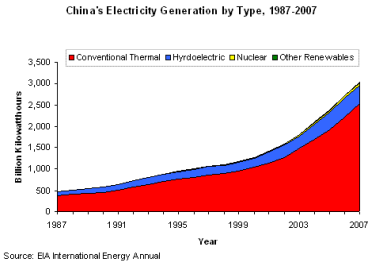


Besoins de nouveaux Services

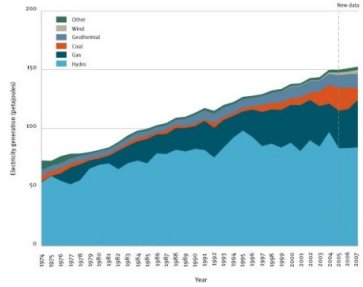
- Certains services et protocoles (terminaux) voient le jour
- D'autres restent dans les cartons : multicast, QoS, sécurité…
- Temps vers standardisation très long
- Exemple : IPv6 (proposé en 95, finalisé en 98)
- Besoin de flexibilité et intelligence dans le réseau pour : validation, déploiement expérimental, déploiement opérationnel, acceptation

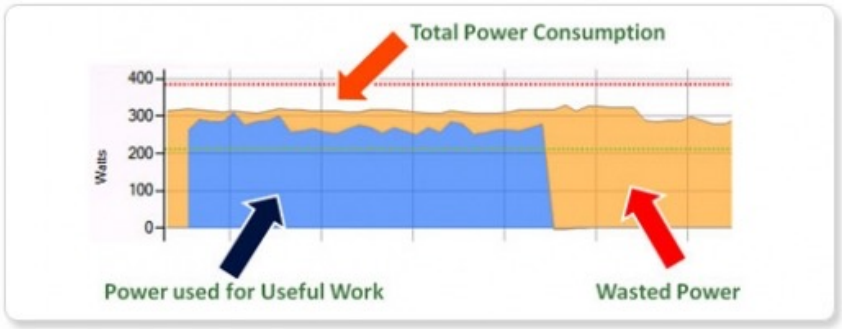
Consommation énergétique

- TICs représenteraient équivalent aviation en génération de Co2
- 10% électricité mondiale – 10% augmentation par an
- GreenIT : réduire la consommation électrique des infrastructures – CO2 dépend de la production
- Focalisation sur la phase d'usage



Source: EIA International Energy Annual





Questions abordées dans cette habilitation

- Comment ajouter de la flexibilité dans les réseaux pour évaluer, expérimenter et valider de nouveaux services et équipements ?

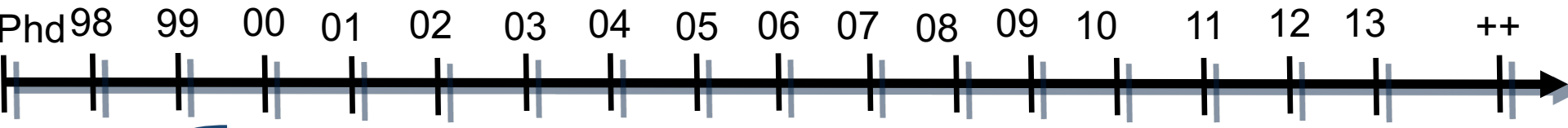
- Comment proposer des modèles et solutions logicielles efficaces pour réduire la consommation énergétique des infrastructures distribuées à grande échelle ?

µCV

- [1993-1997] Thèse sur les Systèmes à Mémoire Distribuée Partagée
- [1997] Postdoc Rice University (Texas)
- [1997-2001] Maitre de Conférences Université Claude Bernard Lyon1
- [2001-now] Chargé de Recherches Inria – Laboratoire LIP - ENS Lyon

- Co-Encadrement : 7 thèses, 1 Postdoc, 10 Ingénieurs, 13 M2, 20 M1
- Projets Nationaux, Européens, Internationaux

(Co) Encadrements



Encadrement Thèses

- J.-P. Gelas
- E. Lemoine
- D. Lopez
- N. Ayari
- M. Diouri
- L. Tsafack
- A.-C. Orgerie
- V. Villebonnet

Flexibilité

Efficacité Energétique

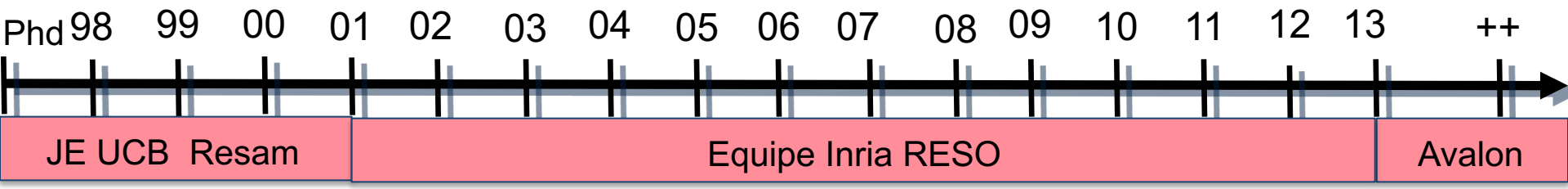
Collaborations Thèses

- P. Neira (Univ. Seville)
- F. Hasan (Univ. Otago)
- C. Herzog (IRIT)
- T. Assefa (Univ. Addis Abeba)

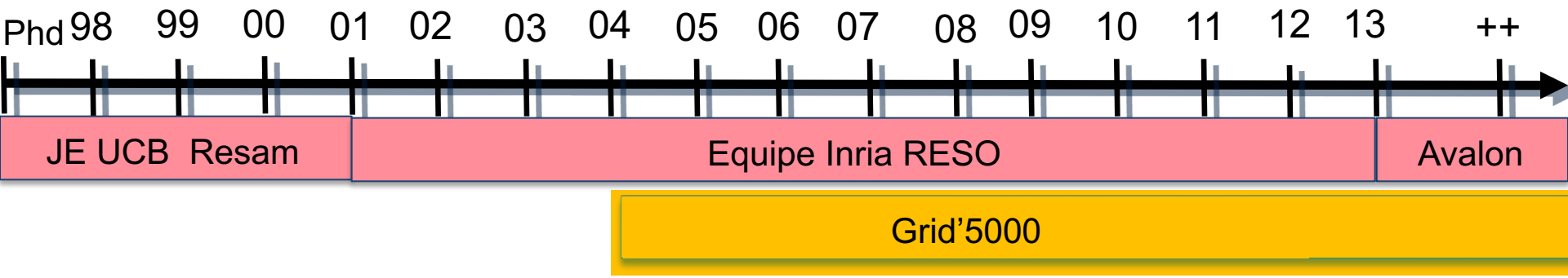
Post-Doc

- M. Dias

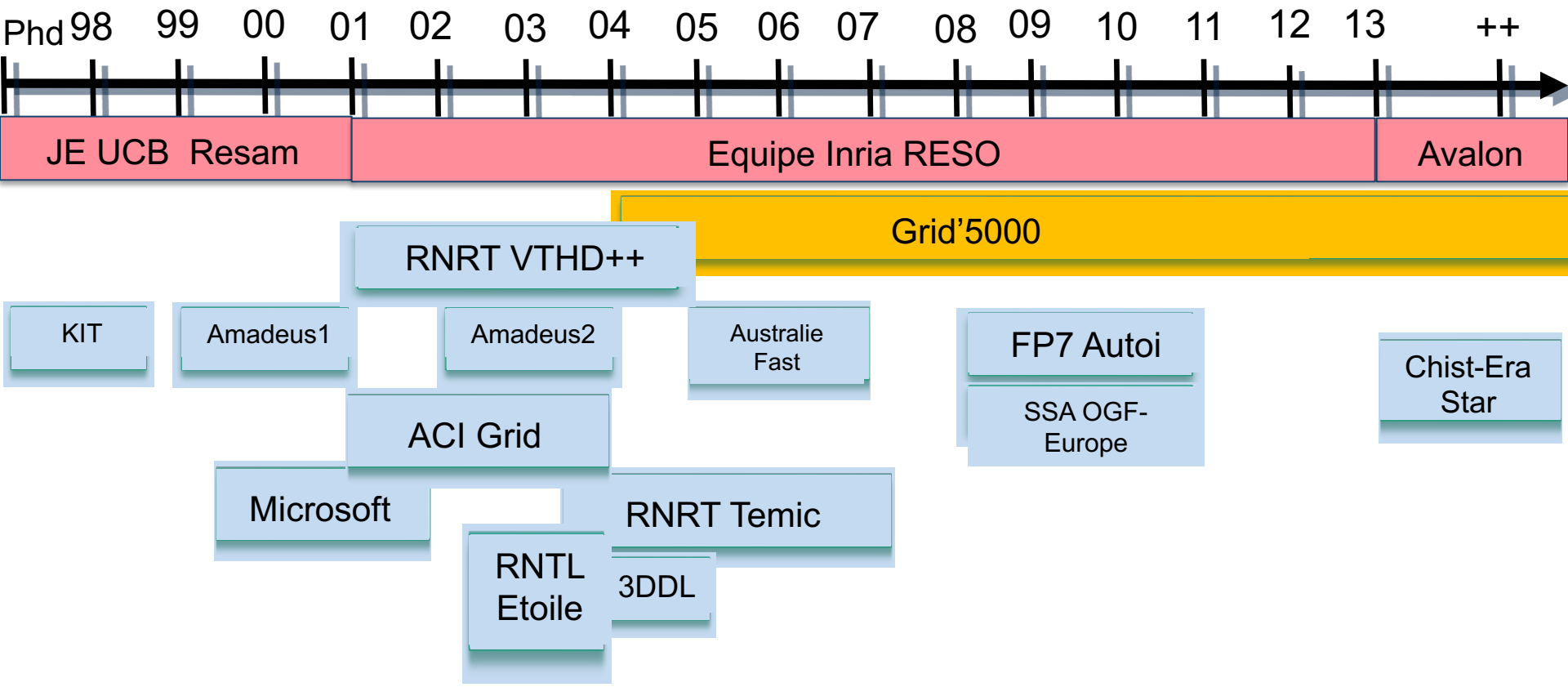
Projets de recherche



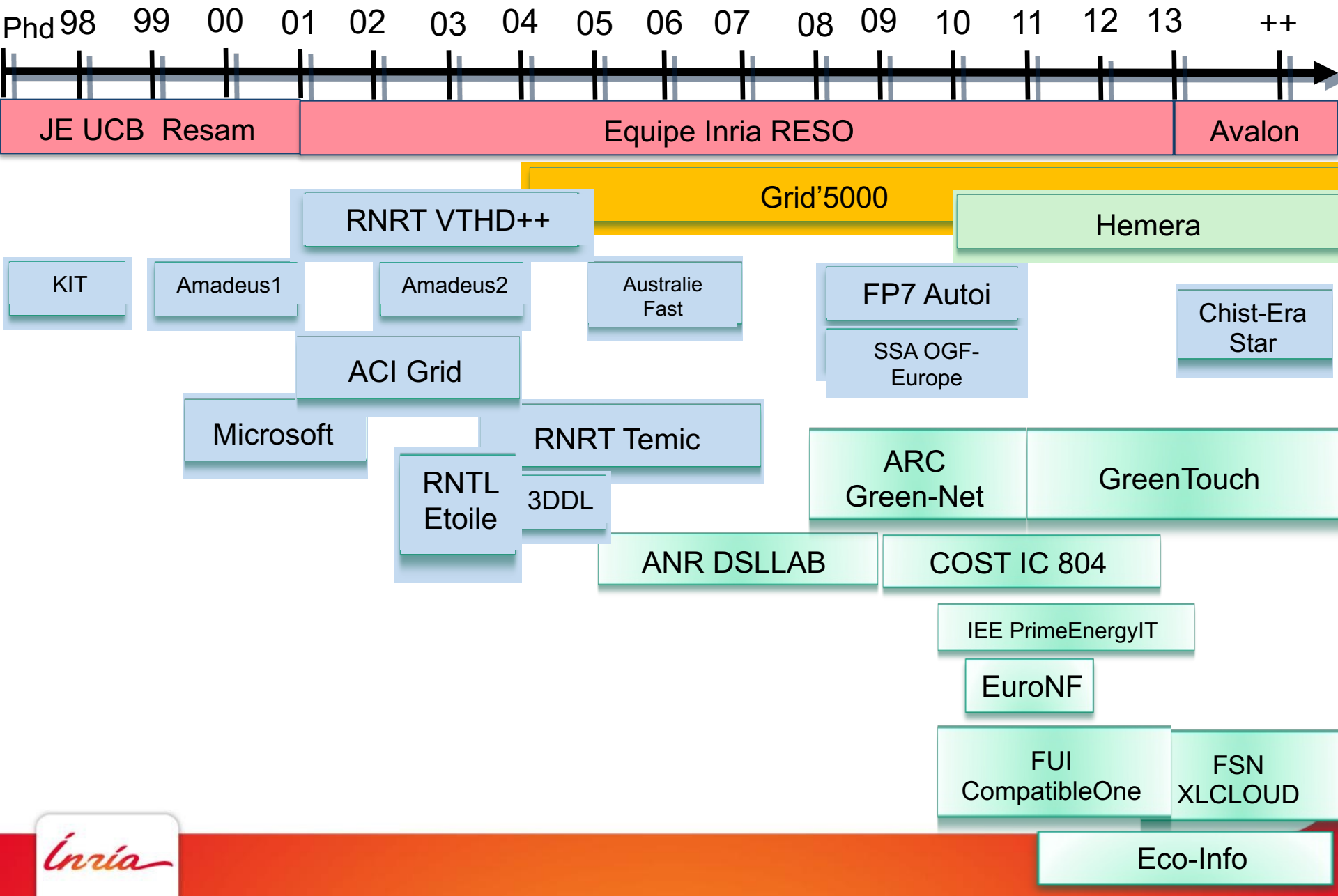
Projets de recherche (G5K)



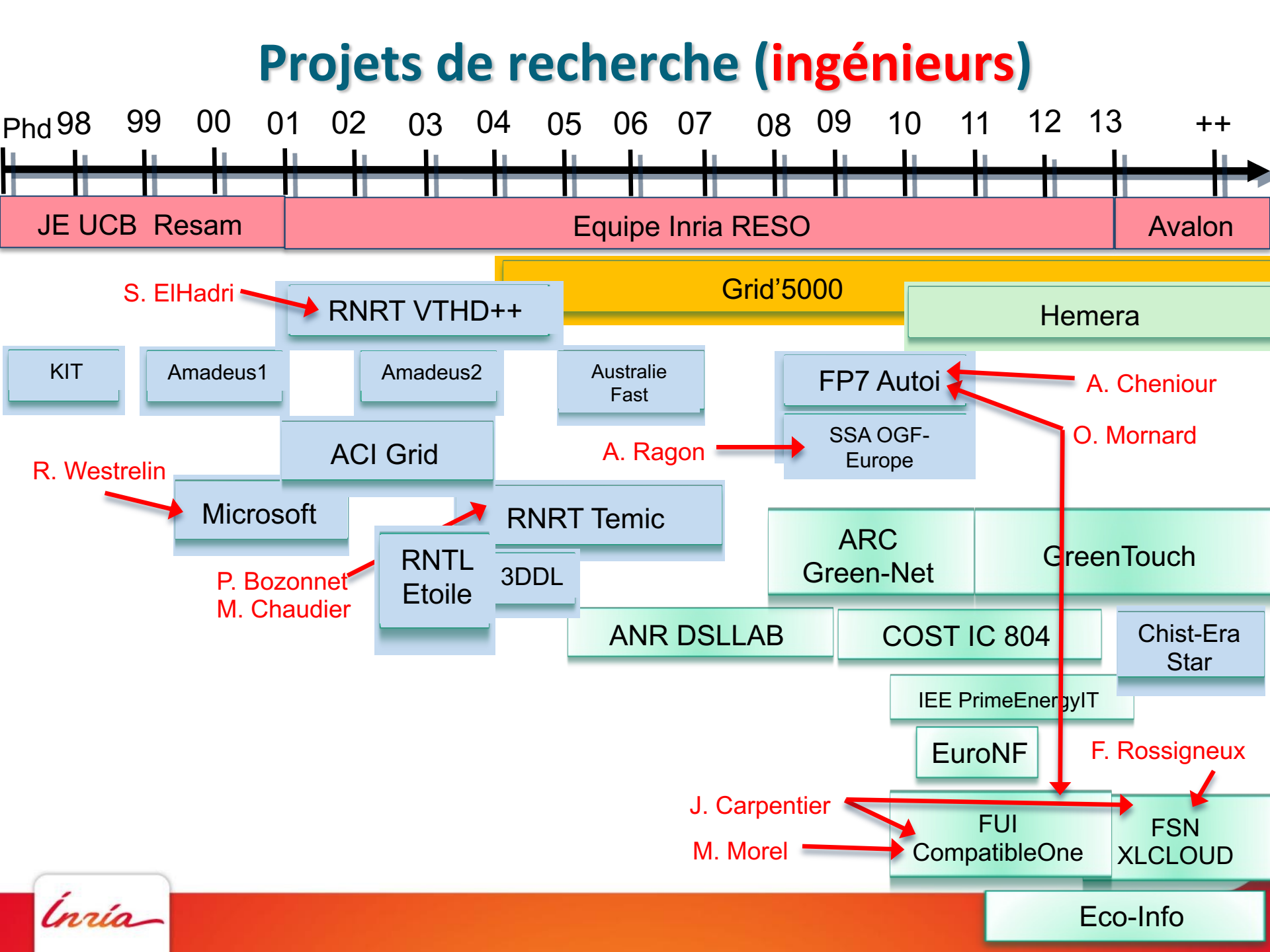
Projets de recherche (flexibilité)



Projets de recherche (EE)



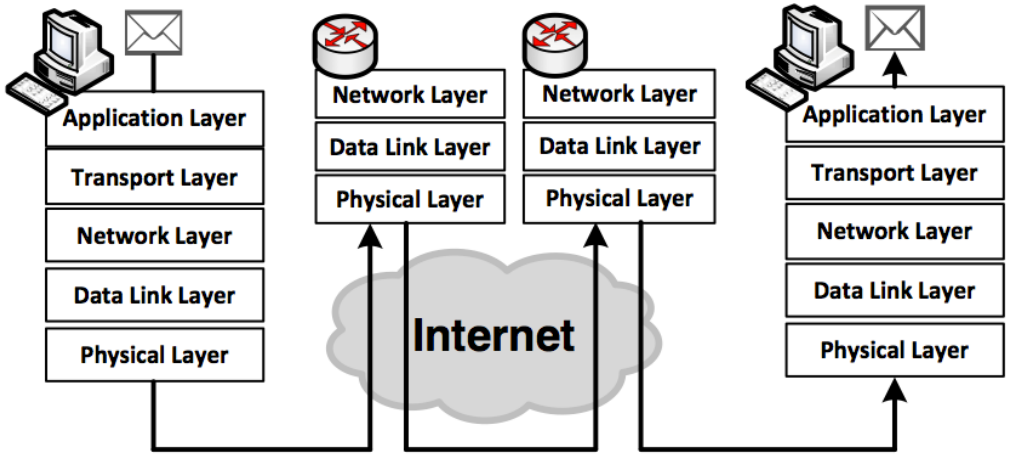
Projets de recherche (ingénieurs)



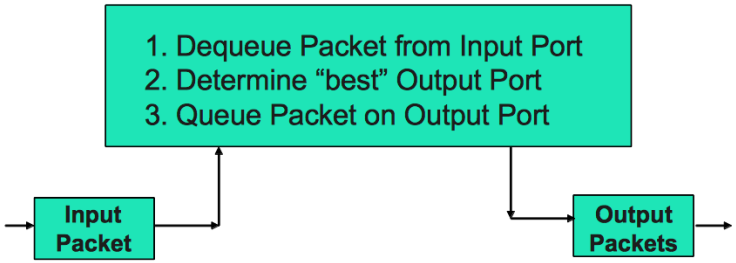
Flexibilité dans les systèmes distribués à grande échelle

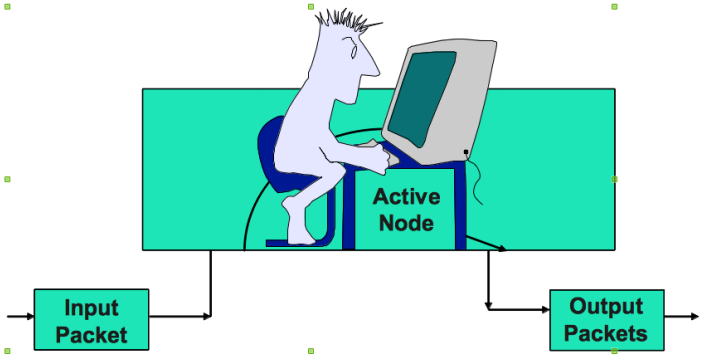
Internet

- Le succès d'un réseau bête
- Le principe du bout en bout : *End2End (Saltzer84)* -> on urbanise l'intelligence aux extrémités : des réseaux simples pour des applications intelligentes



- De *route & route* à *route & process & route* - explorer les limites du E2E et proposer des solutions alternatives

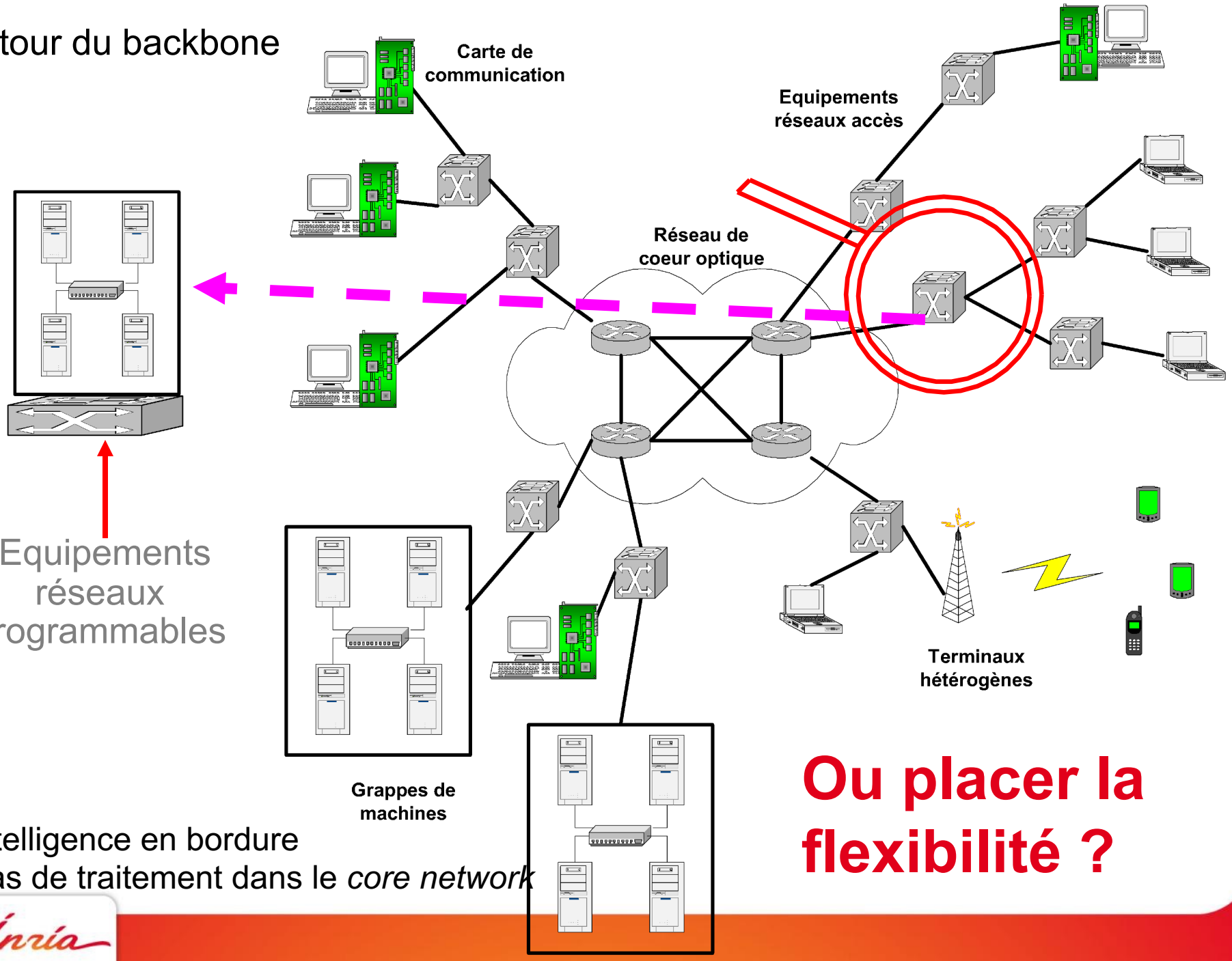




4 vagues de flexibilité

- ***Réseaux pour tous*** : réseaux actifs et programmables
- ***Réseaux sans intervention humaine*** : réseaux autonomes
- ***Infrastructures à valeur ajoutée*** : réseaux virtuels
- ***Réseaux flexibles pour maitriser l'efficacité énergétique***

Autour du backbone



- Intelligence en bordure
- Pas de traitement dans le *core network*

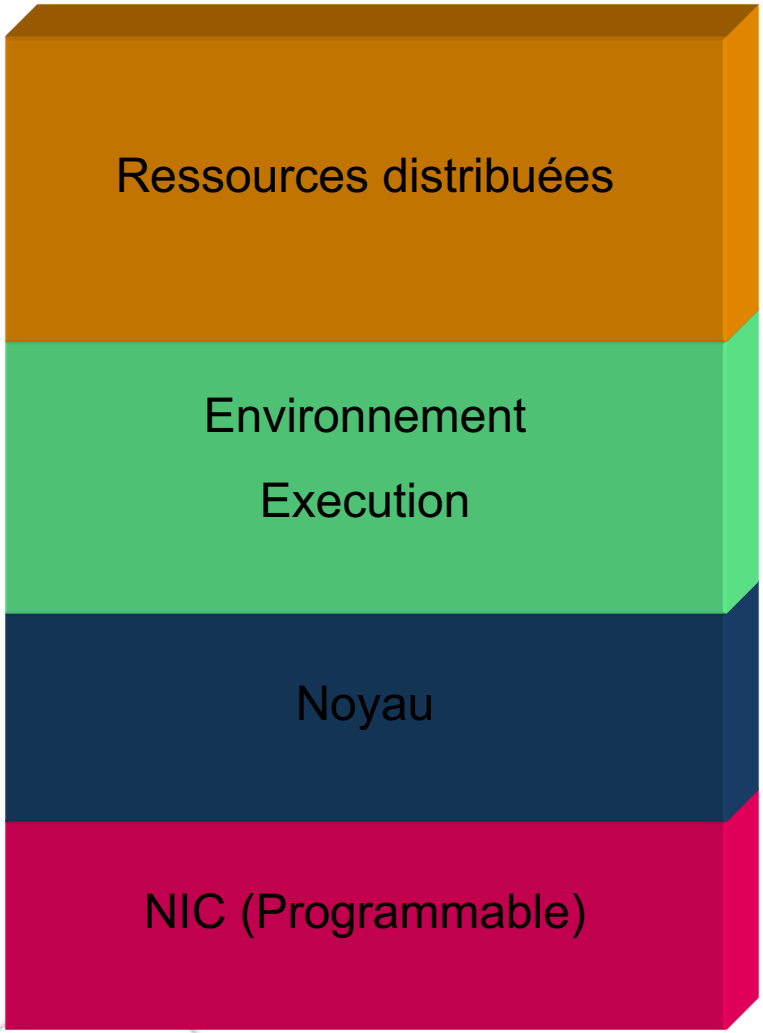
Ou placer la flexibilité ?

Objectifs

- 1er système de réseaux actifs : ANTS (D. Tennehouse)
- Démonstration de la programmation par paquets
- Performances limitées (qq Mbits)
- Explorer les différentes alternatives pour fournir de la haute performance dans les Réseaux Actifs
- Fournir un EE portable et simple (Java) avec des couches adaptées aux besoins des services
- Développer / expérimenter une gamme variée de services actifs (multicast fiable, QoS, adaptation dynamique de flux..)

L'aventure Tamanoir : une architecture adaptée pour des services hétérogènes

Thèse Jean-Patrick Gelas

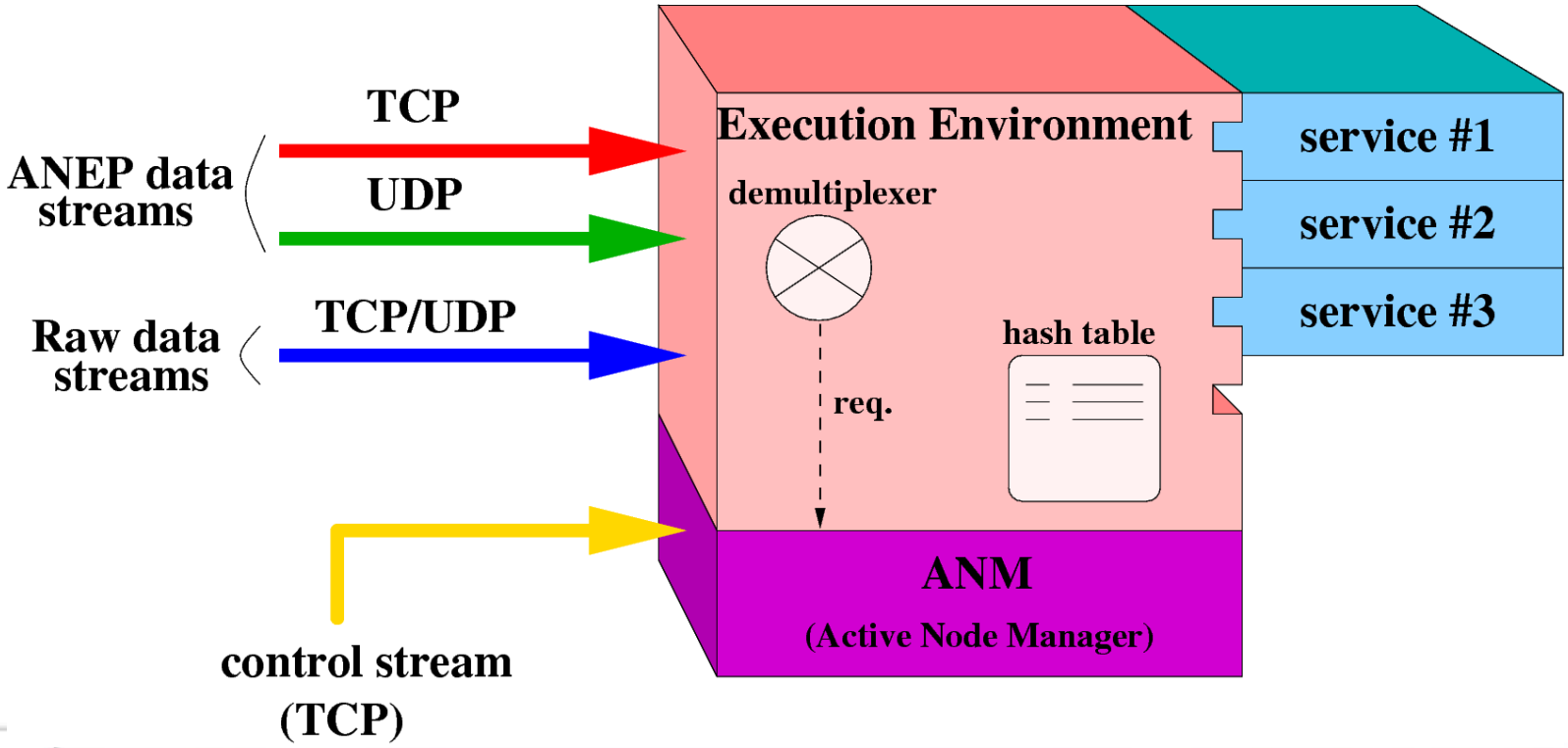


- **Poids lourd** : transcodage multimédia, adaptation d'applications, compression
- **Poids moyen** : caches web, multicast fiable
- **Poids léger** : répartition de charge, pare-feux
- **Poids plume** : support serveurs HP, protocoles de transport

Thèse Eric Lemoine

Un nœud actif Tamanoir : TAN

- EE / services en Java (compilation / JIT…)
- Traitement optimisé (limite les copies)
- Format de données ANEP
- Multi services, multi threadé
- Multi protocoles (tcp,udp)
- Déploiement dynamique de services



Fléxibilité: brique de base pour d'autres projets

- Environnement actif Tamanoir : seul routeur logiciel actif apte à supporter des débits Gbits
- Déploiement et validation sur plate-forme locale et longue distance

- Support de différents projets (RNRT VTHD++, RNTL Etoile, RNRT Temic)
- Déployé et utilisé par partenaires académiques (LAAS, Univ. Vannes, INSA) et contexte "industriel" (3DDL)

Thèse Narjess Ayari

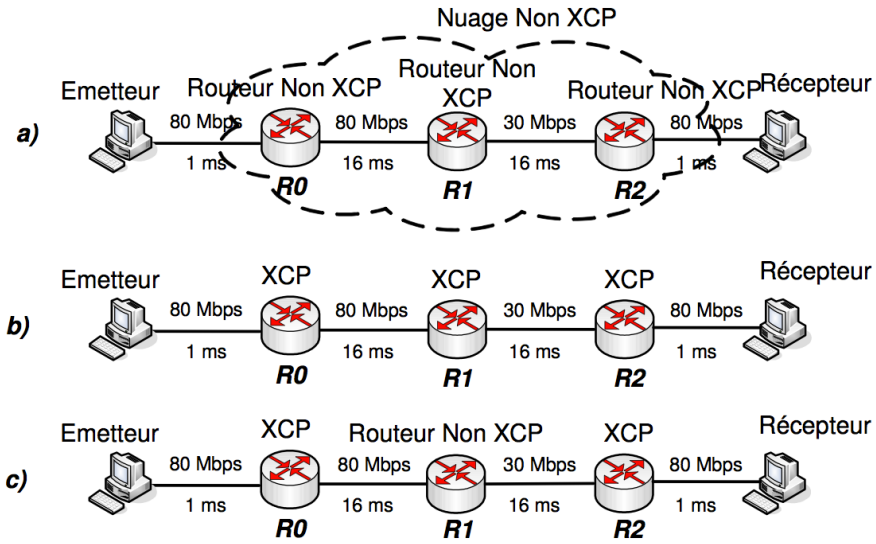
- Flexibilité ré-utilisée dans différents équipements : haute disponibilté, pare-feux à états

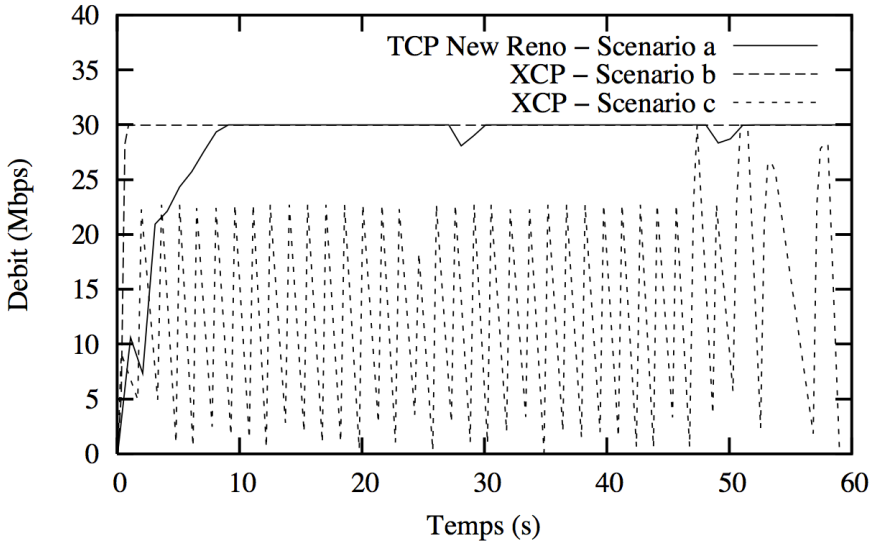
- 1 focus : protocoles de transport

Un service interopérable : XCP-i

Thèse Dino Lopez Pacheco

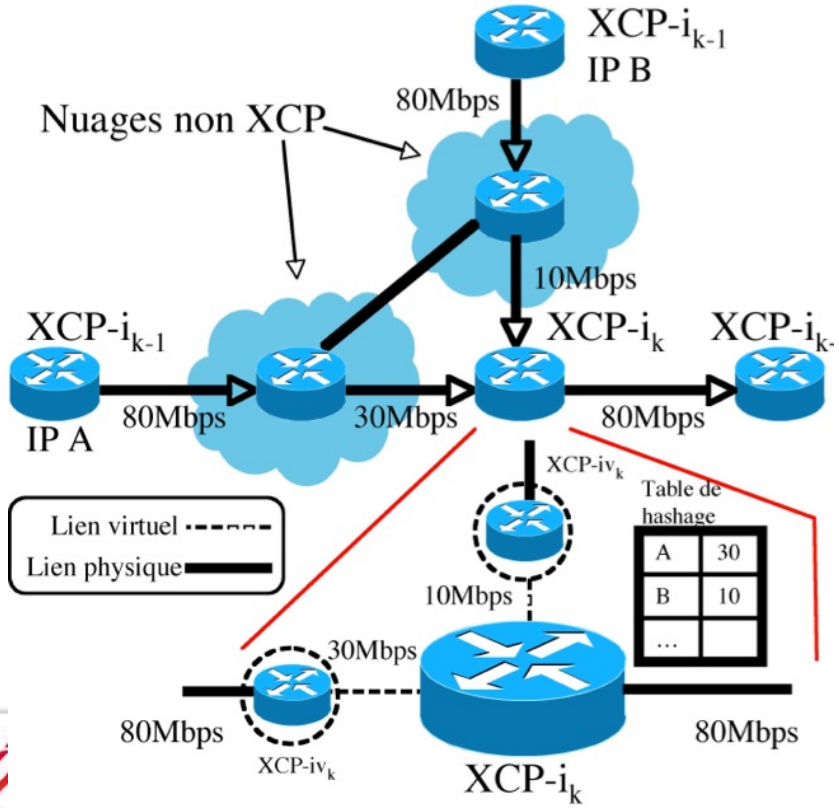
- TCP ne récupère pas toute la bande passante
- Proposition de XCP (eXtensible Control Protocol) (Katabi02)
- Protocole de transport à assistance de routeurs => renvoie une info sur l'état du chemin à l'émetteur
- Flexibilité légère dans les équipements réseaux

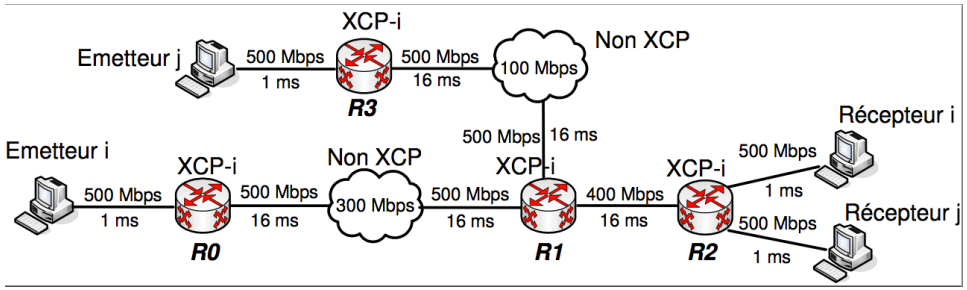


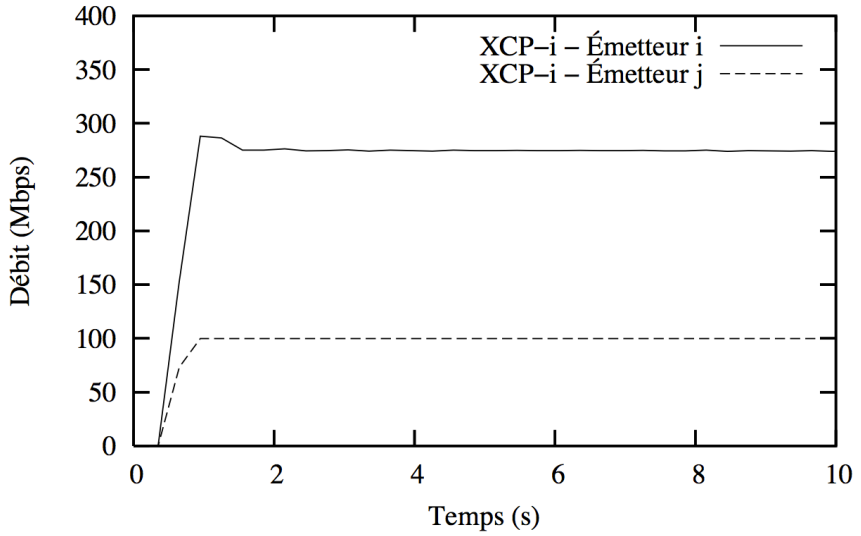


Un service interopérable : XCP-i

- Sortir ce protocoles du laboratoire?
- XCP-i : remplacer un nuage non-XCP par un routeur virtuel situé aux abords du nuage sur le chemin des données
- Gérer l'interopérabilité avec des équipements non-XCP et l'équité entre flux







Efficacité energétique dans les systèmes distribués à grande échelle

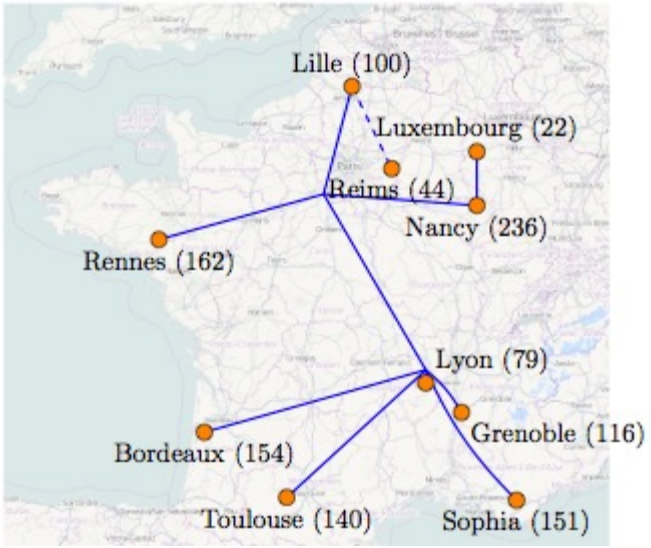
Vers des systèmes distribués efficaces en énergie

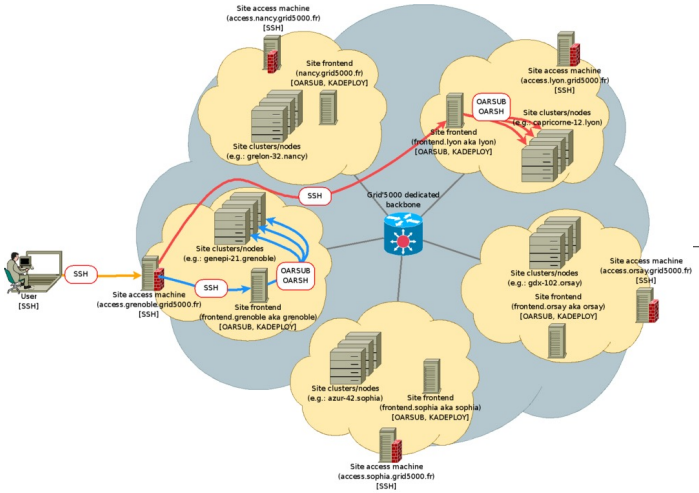
Comment diminuer la consommation énergétique sans impacter les performances ?

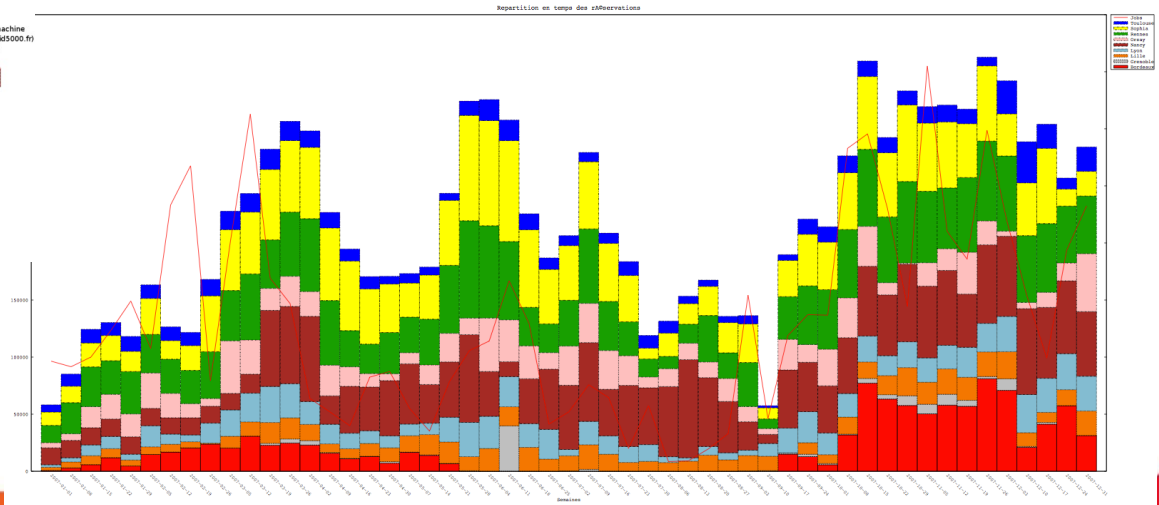
- Comprendre l'usage des systèmes et la consommation énergétique associée
- Etre capable de mesurer et de proposer des métriques utiles
- Concevoir des modèles et des solutions logicielles efficaces en consommation énergétique
- Proposer des composants logiciels génériques dérivables sur différents scenario (Grids, Clouds, Networks)
- Simuler et valider expérimentalement à grande échelle
- Aider les utilisateurs à faire les bons choix

Notre objet d'étude : Grid5000

- 7000 cœurs de calcul / 10 sites
- Réseau 10 Gbits
- Usage exclusif avec réservation / mode best effort
- Usage pics/creux







Mesurer et comprendre la consommation électrique des TICs

Mesurer des TICs ?

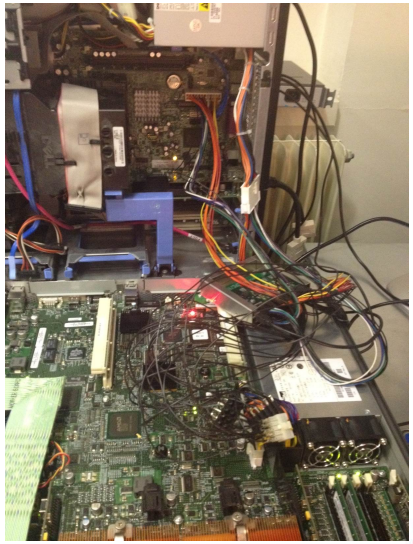
Wh, co2, joules ?

C'est quoi une tonne de co2 ?

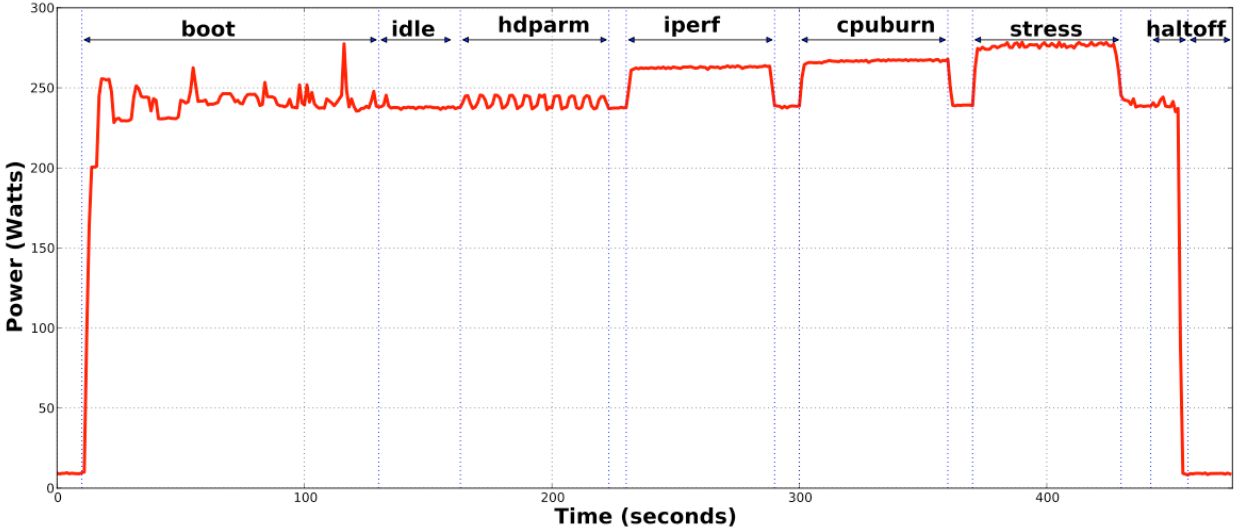
©Dave Ames, Cohasset High School

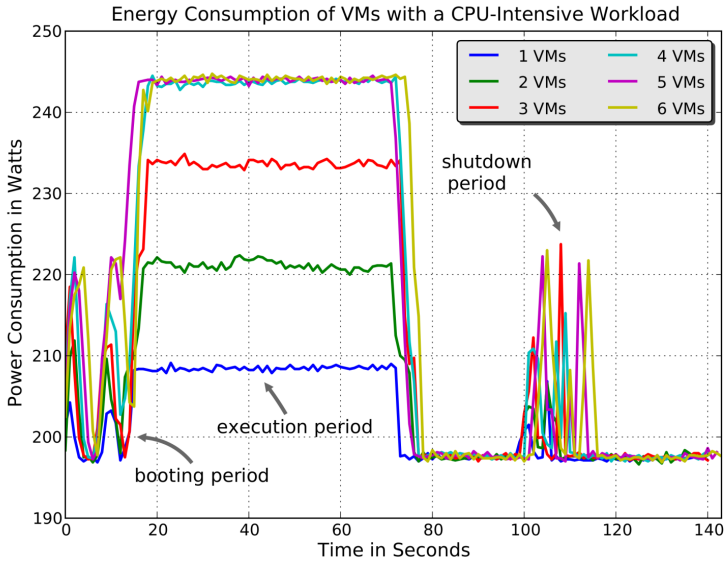
Mesurer ce que l'on comprend : des watts !

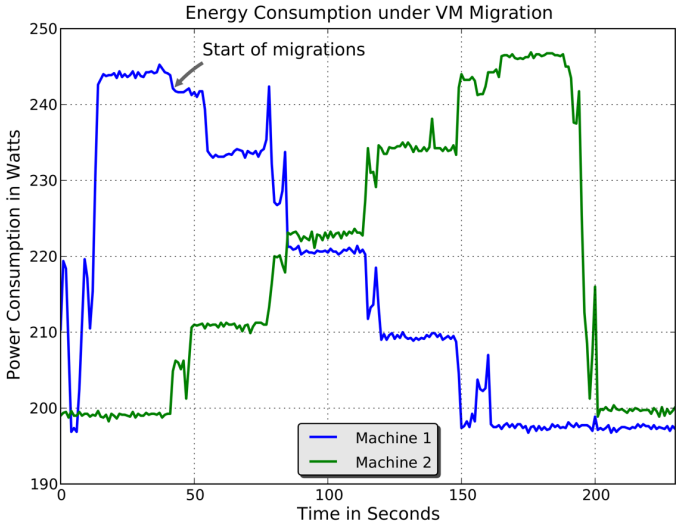
Capteurs externes, internes, intégrés, compteurs de performances



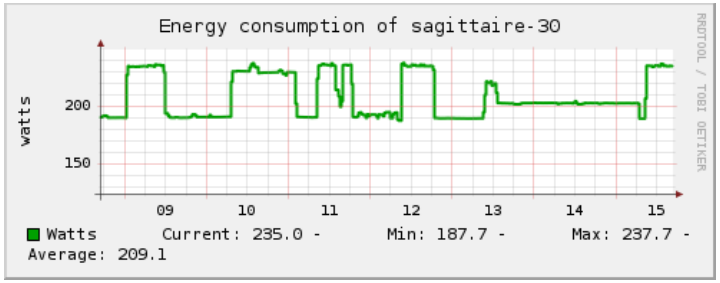
Profiler des infrastructures physiques et virtuelles

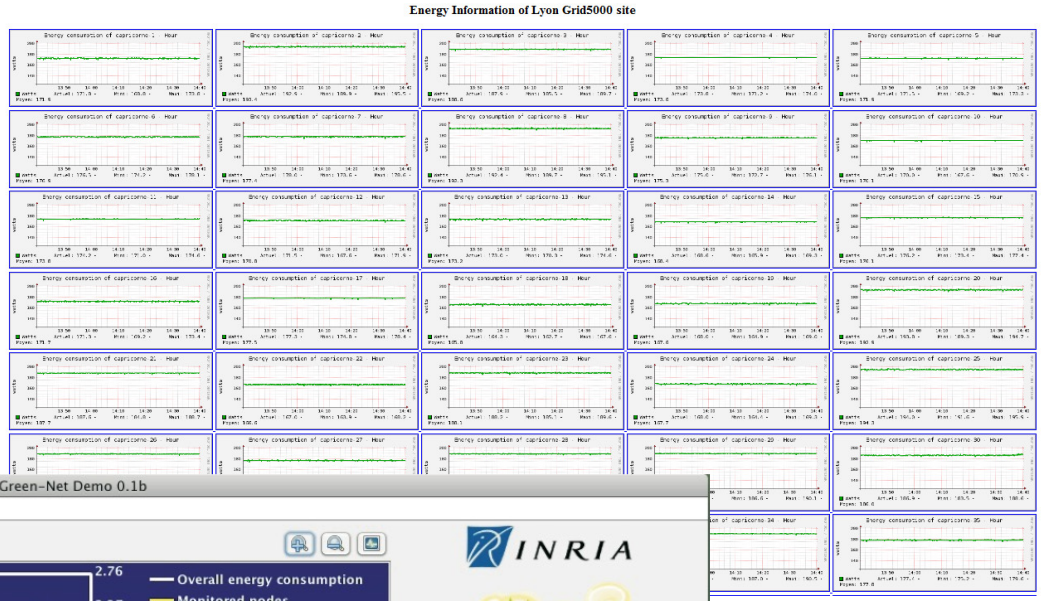


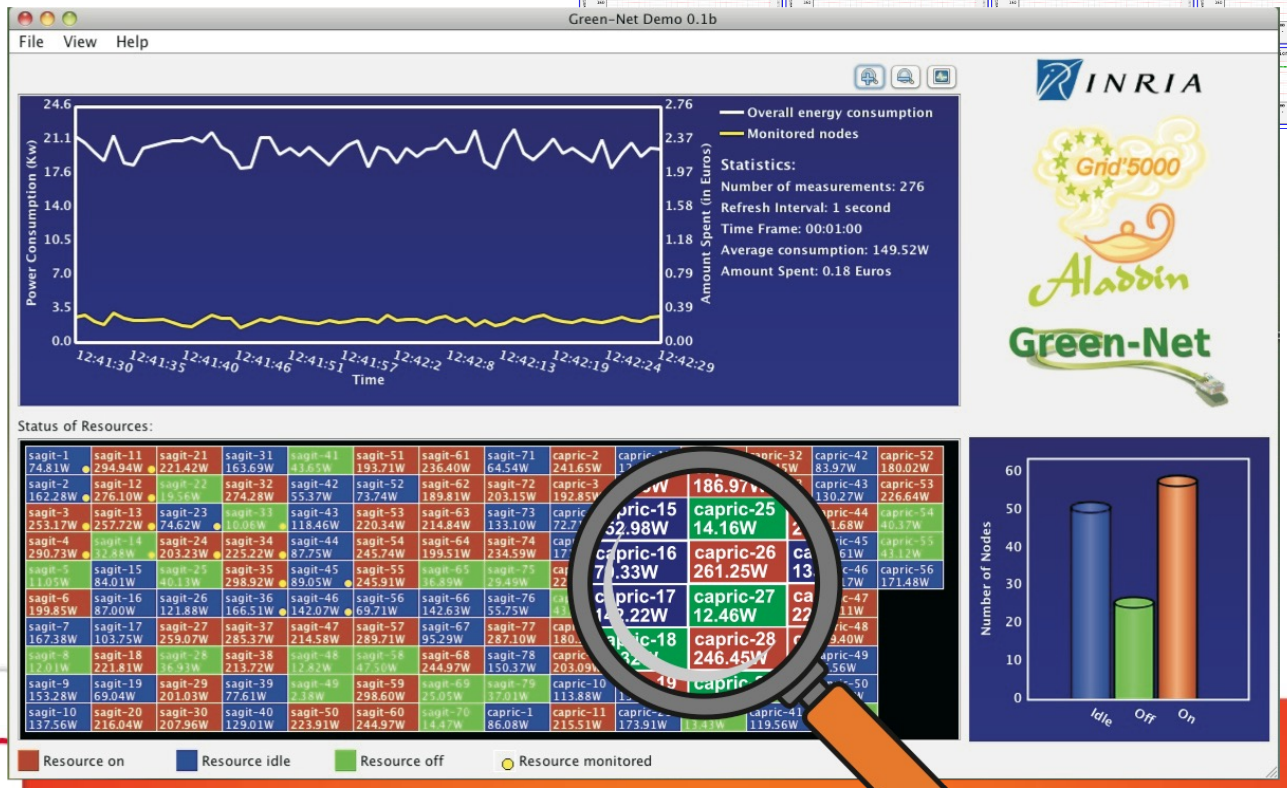




Donner des infos (utiles) à l'utilisateur



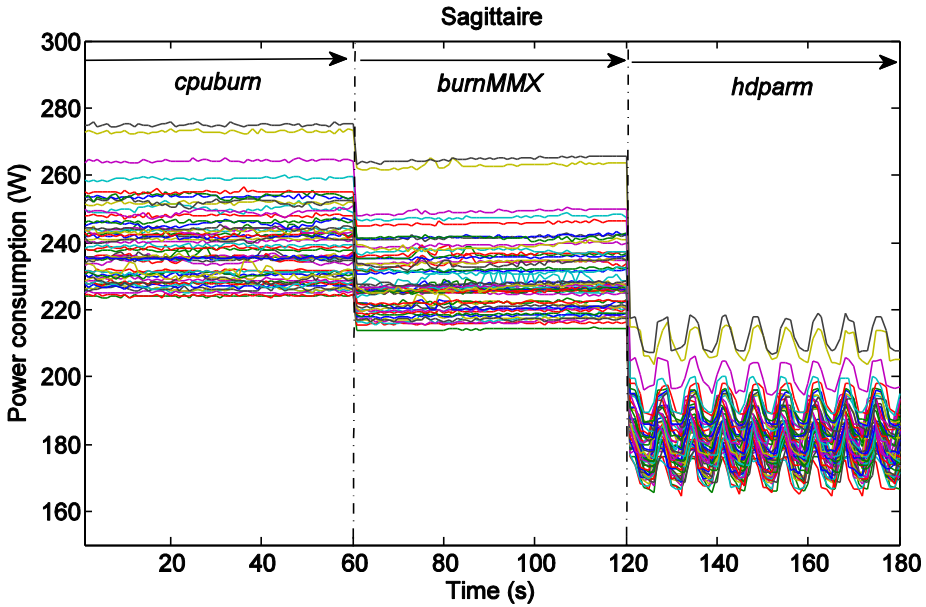


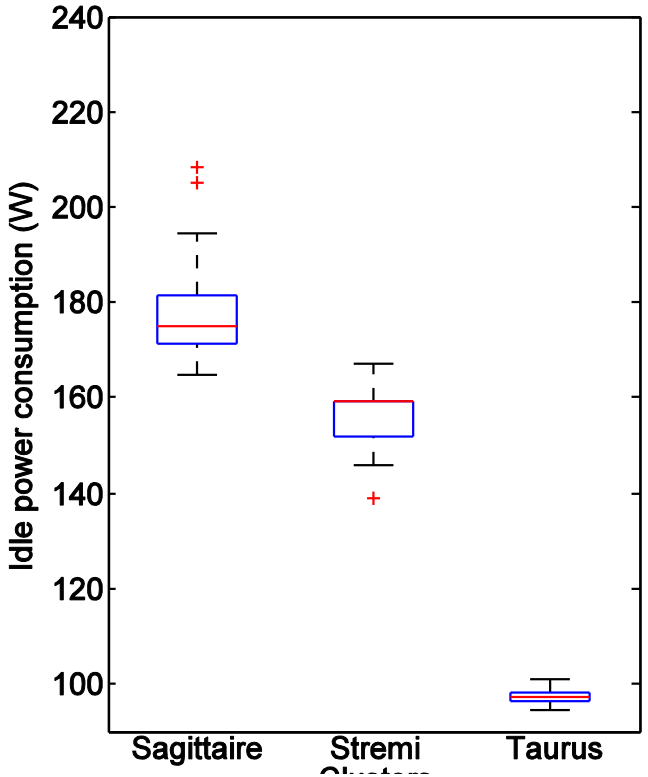


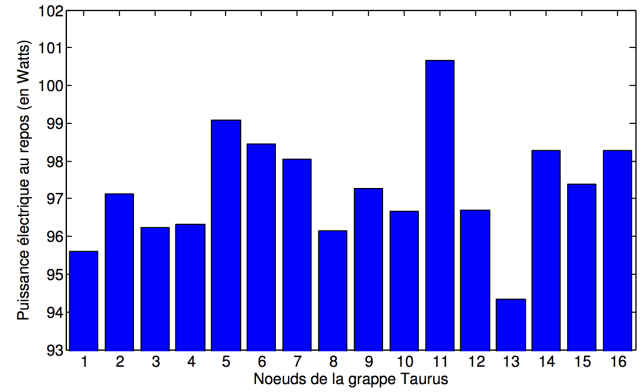
Thèse Mehdi Diouri

Chasse aux mythes : l'homogénéité énergétique

- Mêmes Flops mais flops par watt différents
- Coût Idle
- CPU : responsable principal
- A prendre en compte pour faire des ordonnanceurs verts !







Systèmes à base de réservation et nuages verts

Favoriser des systèmes à base de réservations

Chaque usage est basé sur une réservation (ressources, durée, temps limite) :

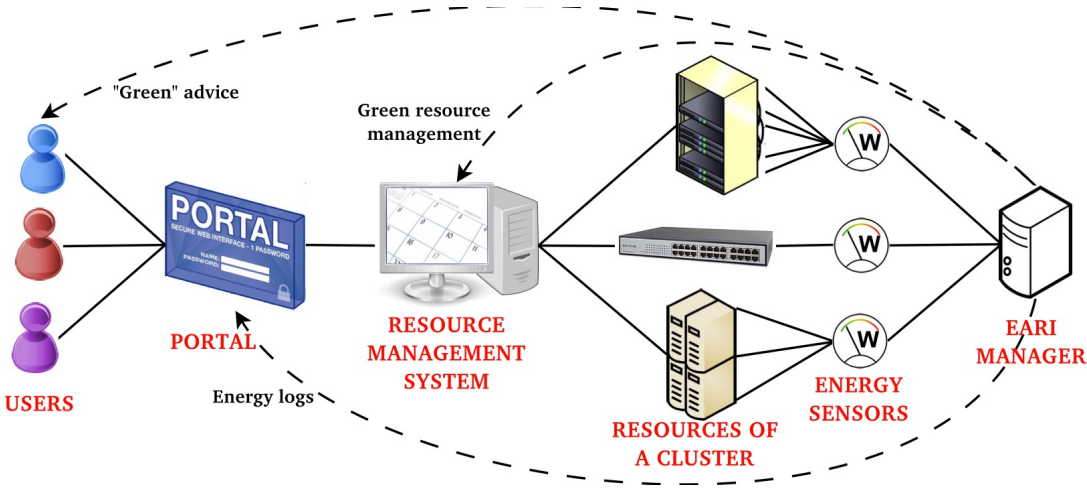
- Réserver des CPUs en HPC
- Réserver des machines physiques ou virtuelles dans les Clouds
- Réserver de la bande passante dans le transport de données

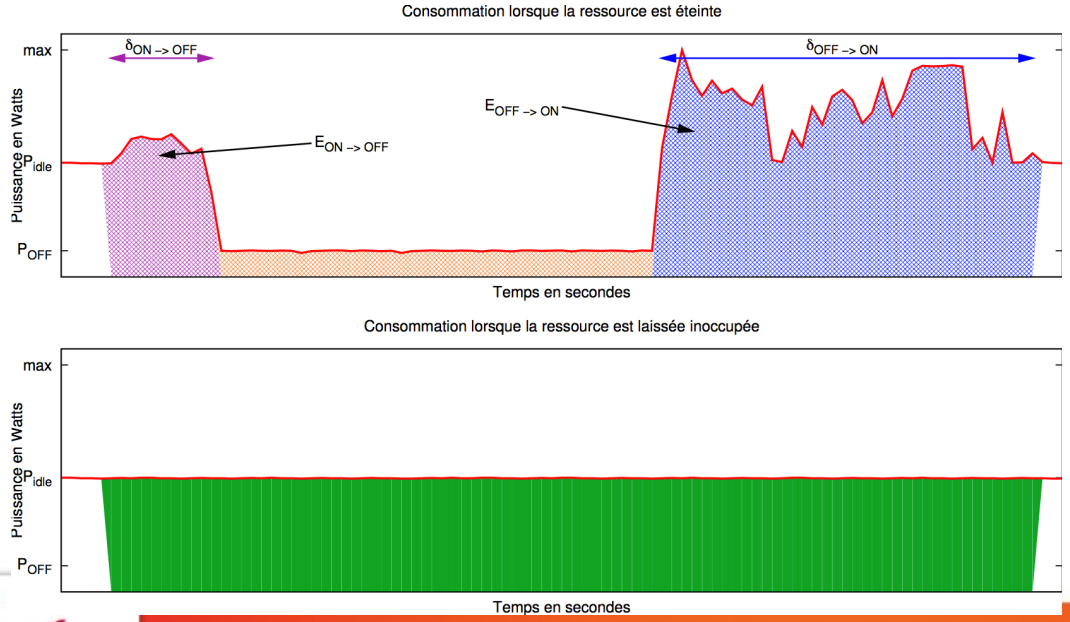
- Leviers:
  - Trouver et alimenter le nombre optimal de ressources :
    - HPC et Grilles : allumer / éteindre des ressources physiques (serveurs)
    - Clouds : allumer/éteindre des VMs
    - Réseaux : allumer/éteindre des ports réseaux, liens, routeurs…
  - Adapter la vitesse et la consommation aux besoins des applications, services et des utilisateurs :
    - HPC: dvfs
    - Clouds : tuning, capping
    - Réseaux : adaptation de bande passante (ALR)
- Favoriser des aggrégations :dans le temps, dans l'espace

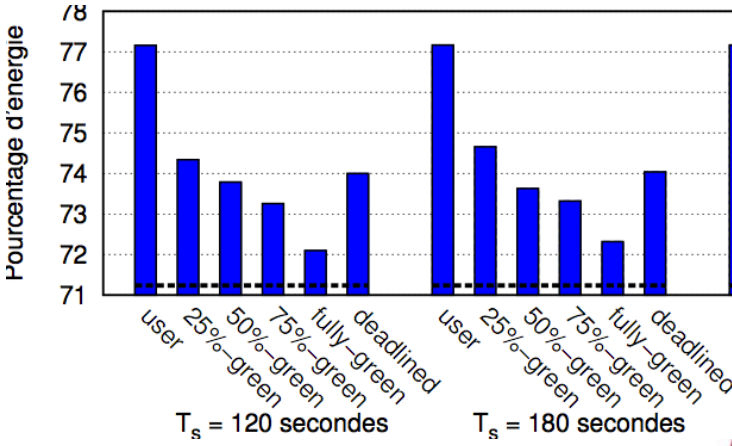
Thèse Anne-Cécile Orgerie

ERIDIS/EARI

- Extinction de ressources inutiles
- Prédiction de réservations
- Placement de taches et de réservations

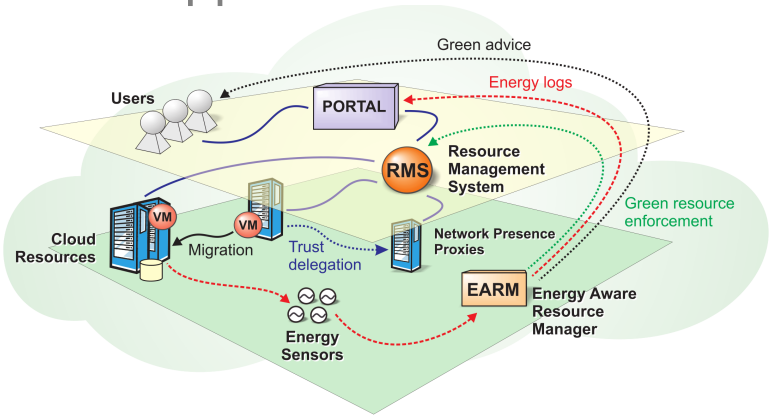


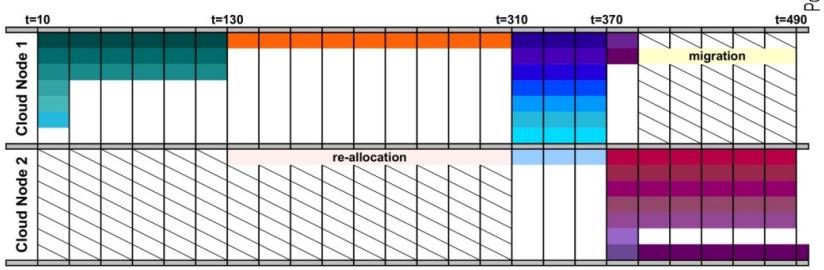


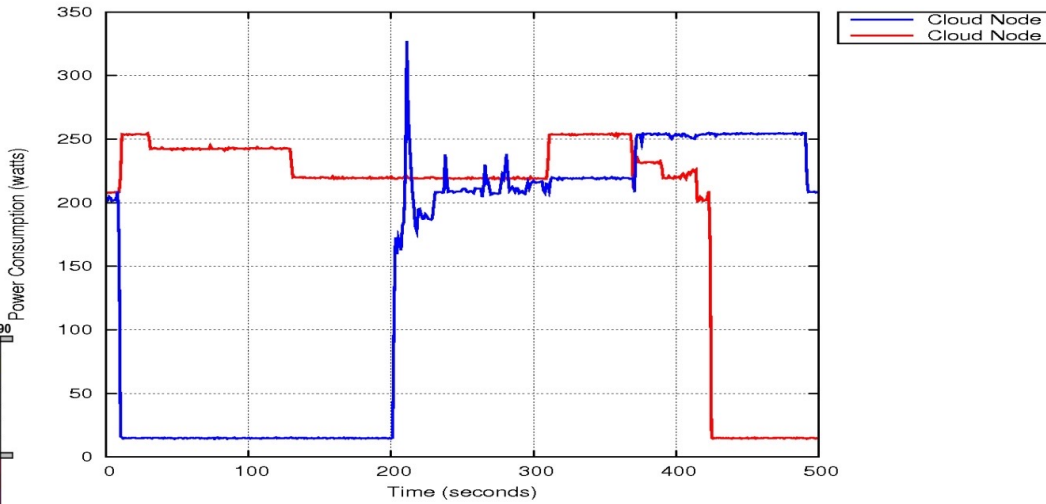


Nuage Vert / Green Cloud

- Mesurer l'usage énergétique des infrastructures de Cloud
- Prendre en compte la consommation pour proposer des composants logiciels économes
- Proposition de l'architecture Green Open Cloud
- Aider les utilisateurs à exprimer leurs contraintes Green
- Aider les gestionnaires à intégrer les couts environnementaux
- Supporter de nouveaux leviers : migration, *tuning*, *capping*







Efficacité energétique dans les infrastructures HPC : avec ou sans connaissance des services et des applications

Vers des infrastructures HPC Exascale

Exascale à l'horizon 2020
1er Top500: Tianeh2 : 3M cœurs, 30-50 Pflops, 17 MW
Eviter le mur des 100 MW → 20 MW

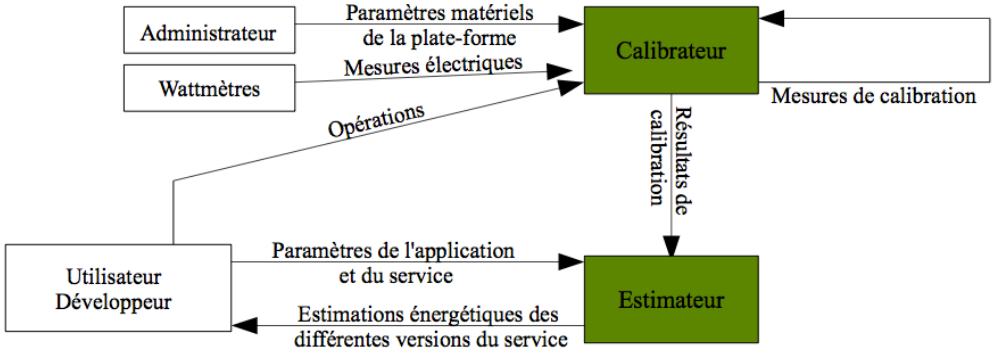
Green500 : meilleurs : 3 Gflops par watt (Tianhe2 1.9Gf/W)
Pour exaflops : 50 Gflops /W

2 approches matérielles : accélérateurs (GPUs) et LPP
(low power processors) (ex: projet MontBlanc)

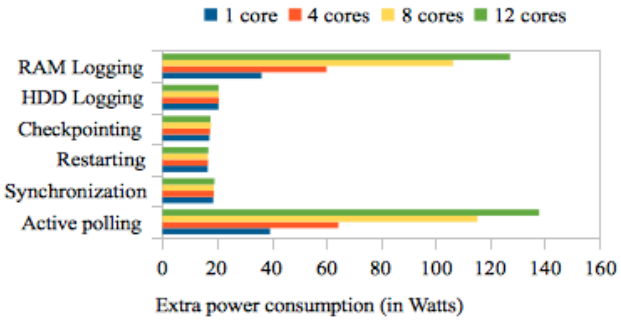
Focus sur applications et services pour exascale

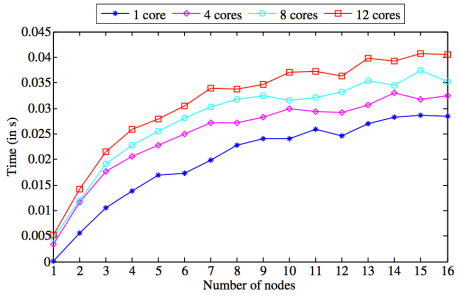
Thèse Mehdi Diouri

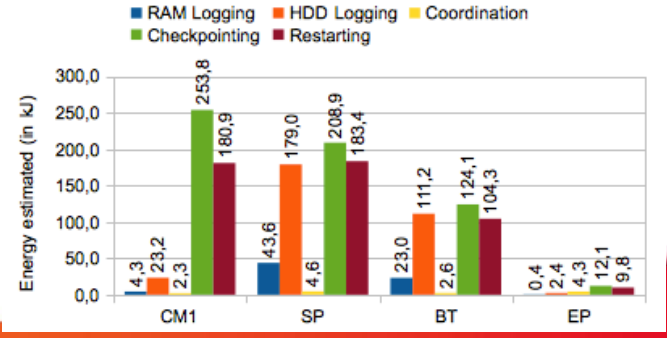
EE HPC avec connaissance des services



- Exemple services exascale : résilience et diffusion
- 4 étapes
- **Découpage** du service en opérations
- **Calibration** de la consommation énergétique des opérations
- **Estimation** de la consommation électrique
- **Aider** utilisateur à choisir le bon service en fonction des contraintes des applications



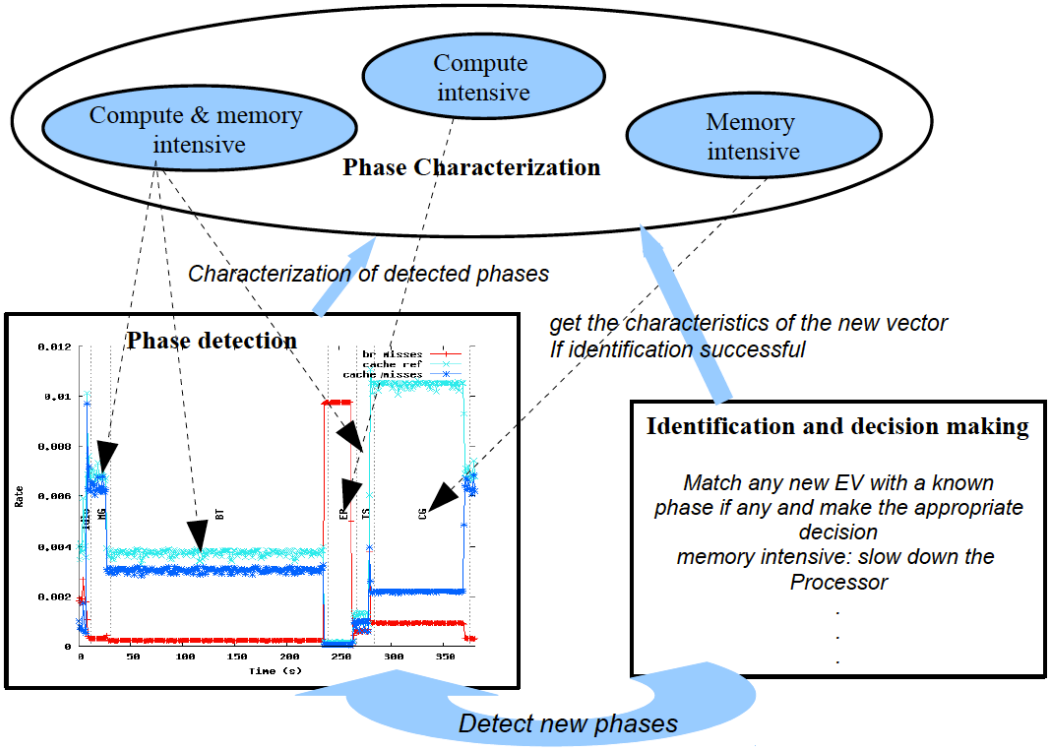




Thèse Landry Tsafack

Et si le logiciel est trop complexe ?

- Applications ont une utilisation non régulière des ressources
- Analyse ADN live de l'exécution des applications et services
- Détection des phases
- Caractérisation des applications
- Applications de leviers verts



Phase label	Possible reconfiguration decisions
compute intensive	switch off memory banks; send disks to sleep; scale the processor up; put NICs into LPI mode
memory intensive	scale the processor down; decrease disks or send them to sleep; switch on memory banks
mixed	switch on memory banks; scale the processor up send disks to sleep; put NICs into LPI mode
communication intensive	switch off memory banks; scale the processor down switch on disks
I/O intensive	switch on memory banks; scale the processor down; increase disks, increase disks (if needed)

Conclusion et Perspectives

Conclusions

- Œuvre collective : merci !
- Modèles, algorithmes, simulations, validations expérimentales
- Proposition logicielles

- Flexibilité
  - Participation aux 4 vagues de flexibilité des réseaux
  - Proposition de plates-formes expérimentales

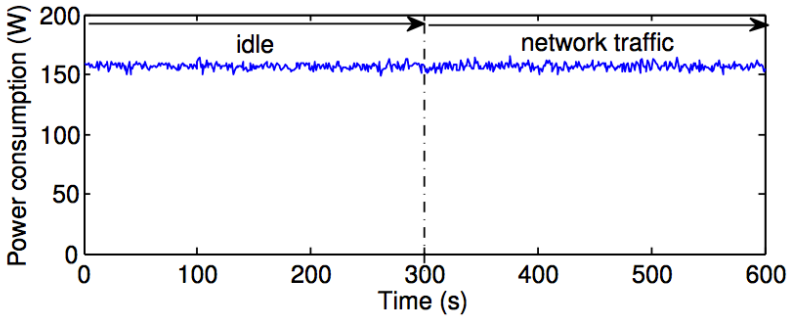
- Efficacité Energétique
  - Modéliser et mesurer la composante énergétique
  - Proposer des systèmes logiciels d'adaptation et éco-efficaces
  - Animation de la communauté GreenIT

Perspectives

- Contribuer au facteur 1000 dans les réseaux
- Consommer mieux dans les infrastructures distribuées à grande échelle
- Proportionnalité énergétique
- Vers le développement durable

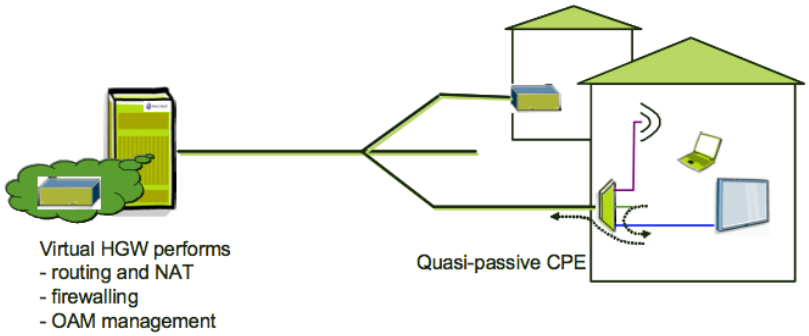
Réseaux flexibles et efficaces en consommation énergétique

- Réseaux : utilisation énergétique plate – non proportionnelle à l'usage
- GreenTouch : réduire la consommation électrique des réseaux d'un facteur 1000 à l'horizon 2015 tout en supportant l'explosion du trafic et la QoS
- Contributions à tous les niveaux: combinaison de matériel et logiciel – sans fil, optique, cœur et extrémités
- *Focus sur les équipements terminaux (10W*box) – virtualisation des services*
- *Virtual Home Gateway*



(A) 2015-2020 NETWORK FORECAST: DEVICE DENSITY AND ENERGY REQUIREMENTS IN THE BUSINESS-AS-USUAL CASE (BAU). EXAMPLE BASED ON THE ITALIAN NETWORK.

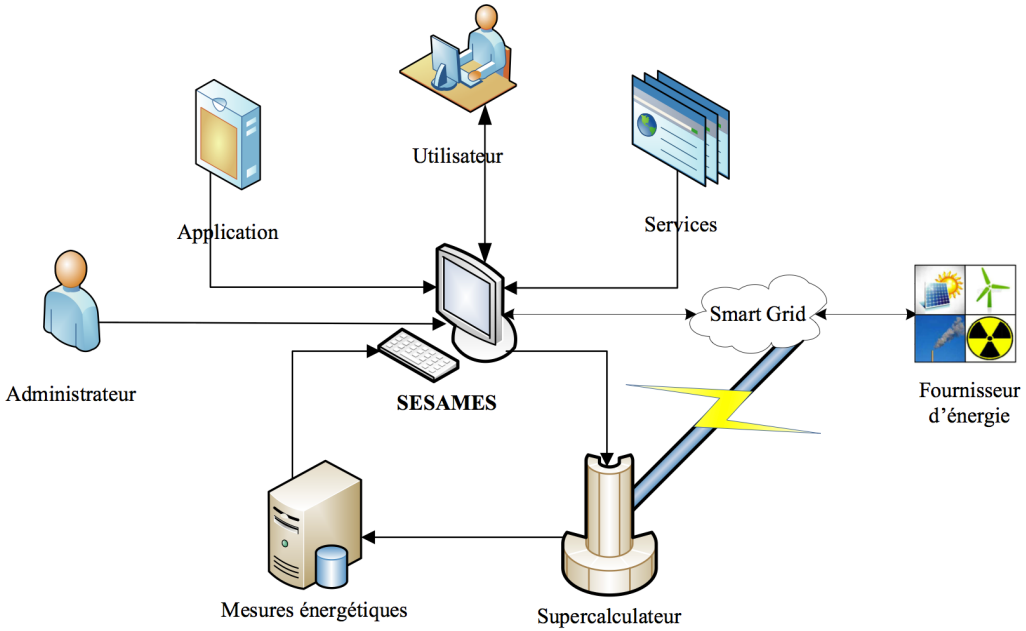
	power consumption [W]	*number of devices* [#]	*overall consumption* [GWh/year]
Home	10	17,500,000	1,533
Access	1,280	27,344	307
Metro/Transport	6,000	1,750	92
Core	10,000	175	15
	Overall network consumption		1,947



GreenTouch™

Consommer mieux

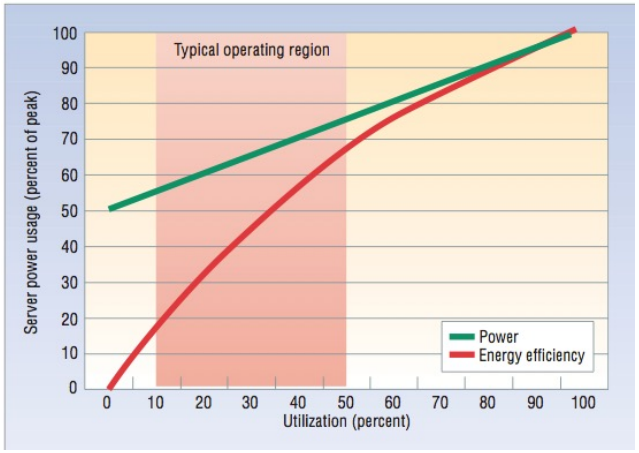
- DataCentre : un client « électrique » à part
- Premières expériences avec SESAMES
- Favoriser le dialogue entre fournisseur électrique, utilisateur, admin
- Prise en compte d'autres contraintes/métriques : production d'énergie, disponibilité, cout financier, pollution…

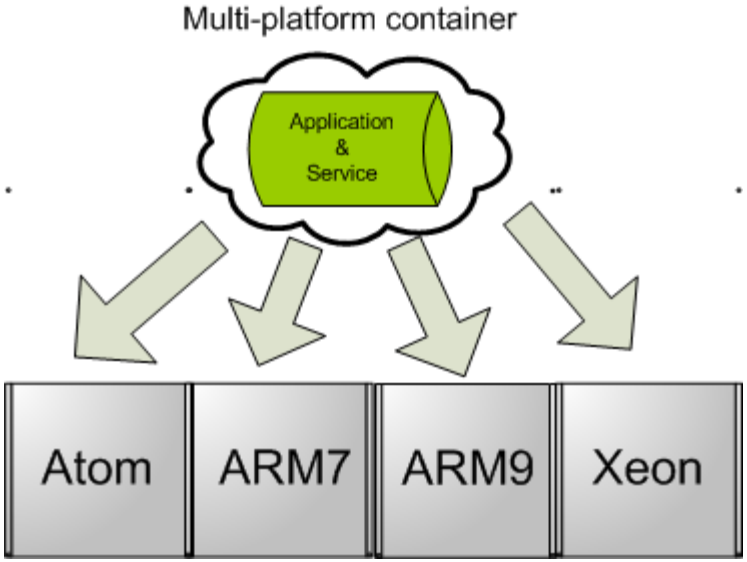


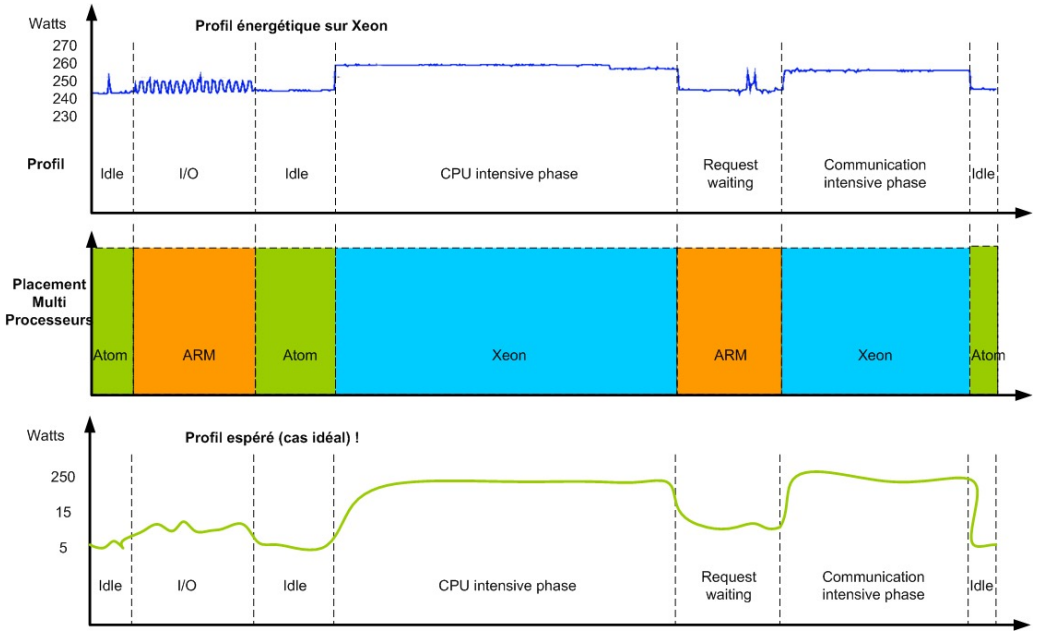
Thèse Violaine Villebonnet

A la recherche de la proportionnalité énergétique

- Exploiter les périodes creuses dans la vie du système
- Combiner des architectures CPU hétérogènes adaptées aux besoins des applications

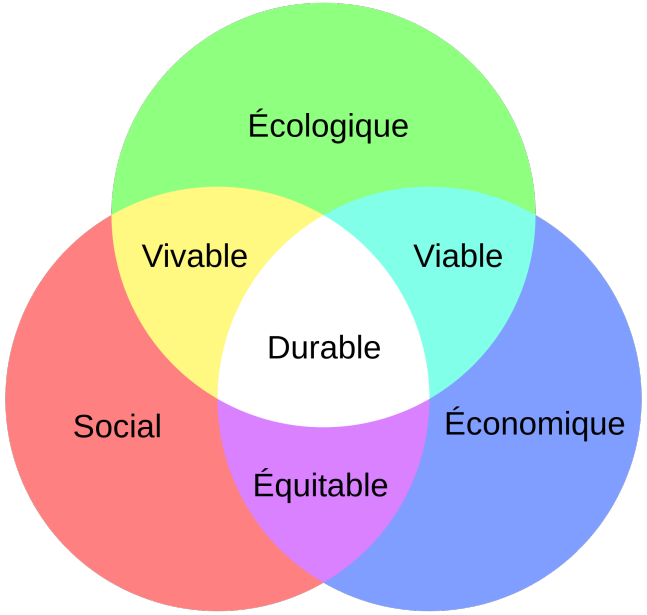


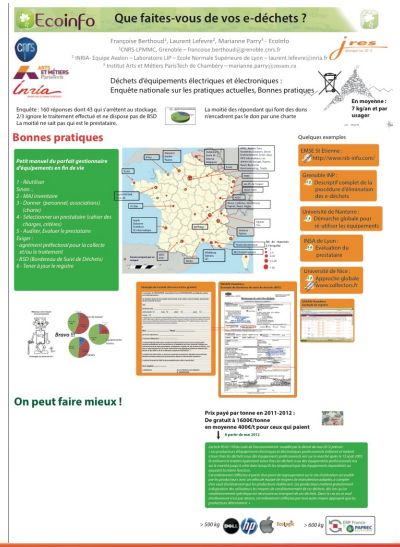




Incursion dans le développement durable

- Supporter un autre usage des systèmes distribués à grande échelle
  - Vers la fin des ressources infinies toujours disponibles ?
- Prendre en compte le cycle de vie des systèmes pour des modèles plus réalistes : production, transport, usage, recyclage





EcoInfo cnrs INS2I

Remerciements

- Jean-Patrick Gelas
- Olivier Gluck
- Christian Perez
- Eddy Caron
- Frédéric Desprez
- Gilles Fedak
- Paulo Goncalves
- Thomas Begin
- Isabelle Guérin Lassous
- Bernard Tourancheau
- Cong-Duc Pham
- Marcos Dias de Asuncao
- Christina Herzog
- Teferi Assefa
- Mehdi Diouri
- Ghislain Landry Tsafack Chetsa
- Anne-Cécile Orgerie
- Narjess Ayari
- Dino Lopez Pacheco
- Eric Lemoine
- Pablo Neira Ayuso
- Edgar Magana
- Jean-Christophe Mignot
- Francois Rossigneux
- Julien Carpentier
- Maxime Morel
- Abderhaman Cheniour
- Olivier Mornard
- Augustin Ragon
- Pierre Bozonnet
- Martine Chaudier
- Saad ElHadri
- Roland Westrelin
- Roya Golchay
- Sidali Guebli
- Alice Bonhomme
- Walid el Dahabi
- Pablo Pazos
- Chien Jon Soon
- Grégoire Locqueneux
- Aweni Saroukou