Ecole Normale Supérieure de Lyon

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

Laurent Lefèvre

Chargé de recherche, Inria

MÉMOIRE D'HABILITATION A DIRIGER DES RECHERCHES

Spécialité : Informatique

Composition du jury :

M.	Pascal	Bouvry	Membre/Rapporteur
M.	Ken	Chen	Membre/Rapporteur
M.	Jean-Christophe	Lapayre	Membre/Rapporteur
M.	Frédéric	Desprez	Membre
Mme.	Christine	Morin	Membre
M.	Yves	Robert	Membre

Habilitation préparée au sein du Laboratoire de l'Informatique du Parallélisme (LIP)

Résumé

Les systèmes distribués à grande échelle (*Datacenters*, Grilles, *Clouds*, Réseaux) sont des acteurs incontournables dans notre société de communication et d'échanges électroniques. Ces infrastructures doivent faire face à de nombreux challenges qui limitent leur déploiement : sécurité, qualité de service, extensibilité, programmabilité, consommation électrique...

Cette habilitation retrace les travaux menés sur les domaines de la flexibilité des grands systèmes en se focalisant sur la mise en œuvre de solutions dynamiques de déploiement de services afin d'augmenter la valeur ajoutée des infrastructures existantes et de proposer de nouveaux services à valeur ajoutée.

Cette habilitation se penche aussi sur la consommation électrique de ces infrastructures et les différents moyens d'améliorer leur efficacité énergétique afin de les placer dans une perspective de développement durable.

Table des matières

La vie est un mouvement ; Plus il y a vie, plus il y a flexibilité ; Plus vous êtes fluide plus vous êtes vivant.

Albert Einstein

1 Introduction

1.1 Parcours

Cette habilitation présente les activités de recherches que j'ai menées depuis l'obtention de mon doctorat en 1997. Ces activités se sont déroulées à l'université Claude Bernard Lyon1 où j'ai été Maître de Conférences pendant 4 années (1997-2001) et au laboratoire de l'informatique du Parallélisme à l'Ecole Normale Supérieure de Lyon en tant que chargé de Recherches Inria depuis 2001.

J'ai effectué ma thèse sur les systèmes de mémoire distribuée virtuellement partagée [102] entre 1993 et 1997. Ces activités de recherche sur les systèmes distribués et le parallélisme se sont poursuivis lors de mon séjour post-doctoral à l'université de Rice (Texas, USA) en 1997.

J'ai été recruté fin 1997 comme Maitre de Conférences à l'Université Claude Bernard afin de rejoindre la jeune équipe RESAM (Réseaux haut débit et Support d'Applications Multimédia - JE 2269) dirigée par Bernard Tourancheau de l'Université Claude Bernard Lyon1. Cette équipe menait des recherches sur les protocoles et les interfaces logicielles pour les réseaux haut débit hétérogènes. Fin 1999, j'ai participé à la création de l'action INRIA RESO (Protocoles et logiciels optimisés pour réseaux très haut-débit). J'ai été Directeur de l'équipe RESAM et responsable de l'action INRIA RESO du 1/1/2001 au 1/9/2002.

En 2003, l'action INRIA RESO est devenue l'Equipe Projet INRIA RESO dans laquelle j'ai mené la plupart des activités de recherche décrites dans cette habilitation. Le projet RESO s'est naturellement arrêté après 8 années au 31 Décembre 2012.

Depuis Janvier 2013, j'ai rejoint l'équipe projet AVALON (Architecture logicielle et algorithmique pour plateformes orientées service) et je participe à la création de cette équipe menée par Christian Perez.

1.2 Évolution de la société numérique

Notre monde est fragile. Les 7 milliards d'êtres humains [51] (9 milliards prévus en 2050 d'après les prévisions des Nations unies) imposent une pression constante sur les ressources naturelles que peut nous offrir notre planète. 900 millions de personnes n'ont pas accès à l'eau potable, 1.3 milliards vivent en dessous du seuil d'extrême pauvreté alors que 2% de la population concentre 50% des richesses mondiales.

Notre société numérique est fragile. Elle repose sur un ensemble de services qui apparaissent indispensables aujourd'hui : moteurs de recherche, réseaux sociaux, transfert et stockage de con-

tenus multimédia. Les besoins évoluent en permanence et imposent de constantes adaptions aux infrastructures, équipements et services. En Juin 2012, l'ONU a reconnu que l'accès à Internet est un droit fondamental, au même titre que les droits de l'homme[151]. Internet est passé en quelques décennies d'un outil de recherche pour académiques à un outil opérationnel indispensable à notre monde moderne. En Juillet 2013, l'*Apple Store* (plate-forme de téléchargement de Apple) a mis à disposition 900 000 applications différentes et a dépassé le cap des 50 Milliards de téléchargements. Les *data centers* autrefois cantonnés à un rôle de gros calculateur pour un ensemble limité d'applications (militaire, météorologique...) se démocratisent et sont déployés à très grande échelle. Ces centres supportent des applications pour le plus grand nombre et stockent des volumes d'informations impressionnants (*cloud* et *bigdata*).

Les utilisateurs imposent aux technologies de l'information et de la communication une garantie de qualité, ils veulent que les services informatiques répondent instantanément, sans panne, sans attente. Cette société numérique repose sur une énergie abondante, ubiquitaire et bon marché.

1.3 Motivations

J'ai effectué ma thèse dans une équipe travaillant sur les systèmes (Remap), je suis devenu Maître de conférences (RESAM) puis chercheur dans un équipe impliquée dans les réseaux (RESO) et je fais maintenant partie d'une équipe travaillant sur les systèmes et les services (AVALON). C'est donc tout naturellement que mes travaux de recherche se placent à l'intersection de deux grands domaines : les réseaux et les systèmes distribués que j'ai à cœur de croiser et de mêler. Cette habilitation est le reflet de cette recherche bicéphale.

1.4 Objet d'étude : Les systèmes distribués à grande échelle

L'objet d'étude de mes travaux concerne les systèmes distribués à grande échelle. Ces systèmes regroupent les centres de données (*datacenters*) et de calcul (*HPC : High Performance Computing*) , les infrastructures physiques (Grilles) et virtualisées (*Clouds*), les grandes infrastructures de transport de données (Internet). Les infrastructures distribuées à grande échelle évoluent d'une manière hétérogène : l'Internet est apparu il y a une quarantaine d'années et supporte de plus en plus de services, la Grille a fait un passage éclair au milieu des années 2000, les *datacenters* et les *clouds* ont le vent en poupe. Ces systèmes sont la pierre angulaire de notre société numérique. Dans cette habilitation, je ne ne considère pas les systèmes distribués à grande échelle de type mobile, sans fils ou à base de capteurs qui sont hors de mon champ d'étude.

Ces objets sont de parfaits terrains d'études académiques car ils permettent des validations et expérimentations à grande échelle, ils représentent aussi des espaces d'expérimentations opérationnelles qui nous permettent de confronter nos solutions aux besoins réels des applications et des utilisateurs.

Je m'attarde tout de suite sur une plate-forme expérimentale particulière qui est le support de nombreux travaux issus de cette habilitation : la plate-forme d'expérimentation nationale Grid5000[37]. Cet outil indispensable nous permet de tester et de valider à grande échelle nos propositions scientifiques en assurant une reproductibilité des expériences. Grid5000 est aussi un formidable outil pour diffuser des logiciels et protocoles dans notre communauté afin d'analyser leur adoption.

La figure 1.1 représente une vue d'ensemble de la plate-forme d'expérimentation nationale Grid'5000[37]. Cette infrastructure distribuée à grande échelle interconnecte 10 sites répartis sur

toute la France et le Luxembourg. Chaque site possède une ou plusieurs grappes de machines interconnectées par différents types de réseaux. Les sites sont reliés entre eux par un réseau à 10 Gbits mis en place par Renater.

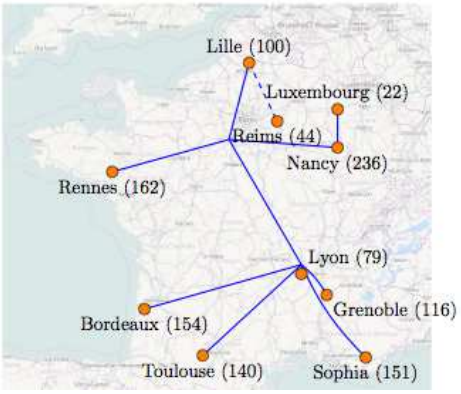


Figure 1.1: Vue d'ensemble de la plate-forme Grid'5000 (nombre de nœuds par site) - 2012 [12]

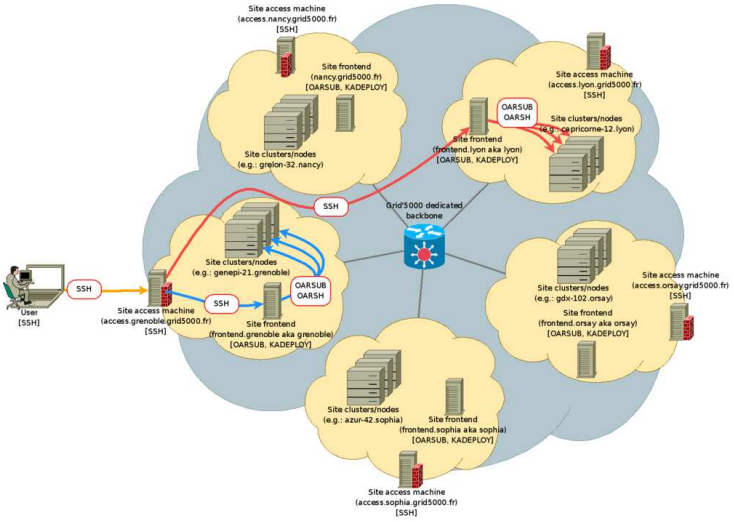


Figure 1.2: Accès à la plate-forme Grid'5000

Un utilisateur accède à Grid5000 par un point d'accès à la plate-forme (figure 1.2) puis deux modes d'utilisation sont disponibles :

- réserver des ressources pour un usage de manière exclusif (utilisation de l'environnement *oar*). Cet opération peut ainsi nécessiter une opération de déploiement d'image système (environnement *kadeploy*);
- utiliser les ressources sans réservation en mode *best effort*.

Grid5000 est l'exemple parfait d'une infrastructure distribuée à grande échelle qui nous permet de mener des développements et validations à grande échelle (plusieurs centaines de serveurs localisés sur des sites différents). Ces validations peuvent avoir lieu en mode exclusif (sans concurrence d'autres utilisateurs ni applications) et permettent ainsi une analyse très fine des observations et une reproductibilité des résultats.

1.5 Contributions

Cette habilitation ne représente pas une description complète de mes activités scientifiques. Certains travaux sont passés sous silence ou rapidement évoqués dans ce document. Mais le lecteur va pouvoir appréhender les principales questions que j'ai abordées pendant ces années de recherche : comment ajouter plus d'intelligence et réduire la consommation énergétique des grands systèmes distribués ?

Je présente ici les deux grands principes qui motivent mes recherches ainsi que les doctorants et leurs thèses que j'ai co-encadrés et qui ont contribué à l'exploration de ces domaines :

- **Lutter contre l'ossification des infrastructures et favoriser la dynamicité des solutions dans les systèmes** :
  - Mes activités sur les réseaux actifs et programmables ont commencé en 2000 avec la thèse de **Jean-Patrick Gelas** (co-encadrée avec Bernard Tourancheau, bourse MENRT, 2000-2003) sur les environnements logiciels pour les réseaux actifs supportant la haute performance. Ces travaux ont notamment donné lieu au système

Tamanoir qui a servi de socle à différents projets de recherches et d'autres explorations sur les services et les équipements flexibles.

- La thèse de **Eric Lemoine** (co-encadrée avec Cong-Duc Pham, bourse CIFRE SUN Labs, 2001-2004) a permis d'explorer une partie de ces services de flexibilité déployés dans des cartes d'interface réseaux programmables et orientés vers la haute performance pour le support de serveurs multi-processeurs.
- Dans la thèse de **Narjess Ayari** (co-encadrée avec Denis Barbaron et Pascale Primet, bourse CIFRE Orange R&D, 2005-2008), nous avons étudié la flexibilité dans les répartiteurs de charge et serveurs distribués utilisés par un opérateur réseaux.
- Dans la thèse de **Dino Lopez Pacheco** (co-encadrée avec Cong-Duc Pham, bourse Conacyt, 2005-2008), nous avons proposé une solution d'interopérabilité incrémentale pour les protocoles de transport de données reposant sur l'assistance des routeurs.

- **Lutter contre le sur-dimensionnement des systèmes distribués à grande échelle afin d'améliorer leur efficacité énergétique** :

  - Mes travaux dans le domaine de l'efficacité énergétique ont commencé avec la thèse de **Anne-Cécile Orgerie** (co-encadrée avec Isabelle Guérin Lassous, bourse MENRT, 2008-2011) par la proposition d'un modèle de réservation efficace en consommation énergétique et déployé sur des scénarios de Grille, Clouds et Réseaux.
  - Bénéficier de la connaissance des services et des applications pour proposer des réductions énergétiques conséquentes dans les infrastructures de calcul à grande échelle est le sujet abordé dans la thèse de **Mehdi Diouri** (co-encadrée avec Olivier Gluck et Isabelle Guérin Lassous, bourse MENRT, 2010-2013).
  - Dans la thèse de **Ghislain Landry Tsafack Chetsa** (co encadrée avec Jean-Marc Pierson et Patricia Stolf, bourse Hemera INRIA, 2010-2013), nous proposons un système d'optimisation de la consommation énergétique des grandes infrastructures de calcul distribués par observation des systèmes, détection de phases et applications intelligentes de leviers verts.

1.6 Organisation du document

Cette habilitation est divisée en deux grandes parties qui correspondent à mes intérêts de recherche depuis 16 ans. Elle représente un travail collectif où des doctorants, ingénieurs et étudiants ont apporté leurs contributions. J'ai aussi bénéficié de plusieurs collaborations internationales qui m'ont permis de combiner nos expertises locales avec des expertises distantes. Je cite les collaborations dans chaque chapitre ainsi que les projets de recherche et personnes associés.

La première partie intitulée "**Flexibilité des réseaux : environnements, équipements et nouveaux services**" regroupe l'ensemble de mes contributions sur la flexibilité.

Dans le chapitre 2 intitulé "**Environnements logiciels pour réseaux flexibles, actifs et autonomes**", je présente les travaux que j'ai menés et encadrés sur les environnements logiciels pour créer de la flexibilité dans les réseaux.

Dans le chapitre 3 intitulé "**Bénéficier de la flexibilité et de l'intelligence du réseau: équipements et nouveaux services**", je présente brièvement les équipements flexibles adaptés à des contraintes industrielles ou à des besoins de haute disponibilité et sécurité. Je m'attarde ensuite sur la présentation de nouveaux services et protocoles que nous avons pu étudier grâce

à la flexibilité ajoutée au sein des réseaux.

La deuxième partie du document "**Améliorer l'efficacité énergétique des infrastructures à grande échelle**" est consacrée aux activités de recherche dans le domaine de l'efficacité énergétique.

Le chapitre 4 "**Mesurer et comprendre l'usage électrique des systèmes distribués à grande échelle**" s'interroge sur la mesure et l'analyse de la consommation électrique dans les infrastructures distribuées à grande échelle.

Dans le chapitre 5 appelé "**De nouveaux composants logiciels pour gérer les ressources dans les infrastructures distribuées à grande échelle : ordonnanceurs et nuages verts**", une proposition de modèle de gestion de réservation efficace en consommation énergétique est proposée ainsi son adaptation aux Grilles et *Datacenters*. Puis nous présentons la problématique de l'efficacité énergétique dans les environnements virtualisés de type Cloud. La proposition *Green Open Cloud* est décrite ainsi que son adaptation dans deux projets de recherche.

Le chapitre 6 "**Améliorer l'efficacité énergétique des très grandes infrastructures HPC : avec ou sans connaissance des applications et des services**" présente deux approches complémentaires pour consommer moins d'énergie dans les infrastructures de calcul haute performance.

En conclusion (Chapitre 7), je dresse un rapide bilan des activités de recherche menées dans le cadre de cette habilitation et propose quelques pistes de réflexion et de travaux futurs.

Part I

Flexibilité des réseaux : environnements, équipements et nouveaux services

"Le réseau donnant vie aux robots est lui-même en voie de robotisation, car les nouveaux routeurs - ordinateurs chargés de gérer le trafic Internet - sont complètement automatisés. Ils sont aussi capables d'analyser, de corriger et de modifier le code composant les logiciels qui les font fonctionner. A force d'essayer toutes les combinaisons possibles, ils parviennent à assembler des lignes de code cohérentes, c'est-à-dire à écrire des logiciels originaux leur permettant d'augmenter leurs performances. Les autres robots peuvent ensuite télécharger ces programmes inédits et se doter de nouvelles fonctions sans que les humains en soient pleinement informés."

Yves Eudes, Le Monde, 2005 [64]

2

Environnements logiciels pour réseaux flexibles, actifs et autonomes

"Cela ne peut pas marcher, cela ne respecte pas le bout en bout !"[1] (Gordon Bell). Quand un chercheur de renom m'invective de la sorte après mon exposé ; deux possibilités s'offrent à moi : tout arrêter aussitôt et me mettre à explorer des recherches plus conventionnelles (ajouter une n-ième paramètre à TCP?) ou au contraire être attisé par ce genre de déclarations et me lancer dans cette entreprise audacieuse.

J'ai choisi la deuxième voie !

Face à l'ossification d'Internet, de nombreux services ont été proposés dans les réseaux. Mais alors que plusieurs centaines de services sont étudiés dans le monde académique et industriel, seuls quelques dizaines sont réellement déployés à l'échelle de l'Internet. La durée de validation, de déploiement et d'acceptation des nouveaux protocoles est très longue. Par exemple, le protocole IPv6 (Internet Protocol version 6) a été développé et standardisé dans les années 90. Il a été proposé dès Décembre 1995 (RFC 1883[49]) et a été finalisé en Décembre 1998 (RFC 2460[50]). Ce protocole qui devait sauver l'Internet du manque d'adresses réseaux n'est pas encore complètement déployé 15 ans après sa proposition et validation par l'IETF (Internet Engineering Task Force).

Dans ce contexte, comment ajouter de la flexibilité dans les réseaux en évaluant, validant et expérimentant de nouveaux services ?

Les réseaux actifs sont apparus à la fin des années 1990 avec la proposition de capsule active (système ANTS de David Tennehouse[172]). Reposant sur un concept en rupture : les paquets de données peuvent influencer le comportement des équipements réseaux traversés. Ce qui pouvait être considéré comme de science fiction[64] devenait réalité ! Les réseaux actifs (capsules ou nœuds configurables), programmables (P1520) et certains concepts des *Software Defined Networks* (SDN) autorisent donc la modification des équipements réseau en fonction des besoins des applications, des utilisateurs et des opérateurs.

Nos premiers travaux dans ce domaine ont commencé avec la thèse de Jean-Patrick Gelas (co-encadrée avec Bernard Tourancheau, 2000-2003) quand nous avons proposé une nouvelle

[1] "It cannot work, it is not end to end !", Gordon Bell, après mon exposé à CCGSC 2002 : 5th Cluster and Computational Grid for Scientific Computing Conference, Château de Faverges de la Tour, France.

architecture de réseaux capable de répondre à des besoins de haute performance (section 2.3). Ces propositions nous ont permis d'étudier et de valider des expérimentations extrêmes qui n'existent pas encore sur l'Internet [104, 105].

En parallèle, dans la thèse de Eric Lemoine (co-encadrée avec Cong-Duc Pham), nous avons exploré le déploiement de fonctionnalités logicielles légères dans les cartes d'interface réseaux. Ces services sont déployés pour supporter les flux utilisés par les serveurs multi processeurs (section 2.4).

Dans le cadre du projet européen FP7 Autonomic Internet (AutoI), j'étais en charge de la flexibilité et de la programmabilité des équipements réseaux. Autoi proposait de mettre en place les solutions logicielles pour la gestion des très grands réseaux de communications virtualisés déployant des services dynamiques. Avec des ingénieurs recrutés pour ce projet (Abderhaman Cheniour et Olivier Mornard), nous avons mené différents développements qui ont conduit à la maîtrise de ces réseaux virtuels à très grande échelle (section 2.5).

2.1 Internet : le succès d'un réseau bête

Dans leur article fondateur de 1984, Saltzer, Reed et Clark ont défini le principe du bout en bout (*End2End* [152]) comme un des fondements de l'architecture de l'Internet. Les auteurs annoncent que des fonctions implémentées à des niveaux bas dans un système peuvent être redondantes ou de faible valeur ajoutée par rapport au coût de ce placement au bas niveau.[2]

Pendant de nombreuses années, la communauté réseau l'a interprété comme le fait de « plutôt que d'installer l'intelligence au cœur du réseau, il faut la situer aux extrémités : les ordinateurs (équipements) au sein du réseau n'ont à exécuter que les fonctions très simples qui sont nécessaires pour les applications les plus diverses, alors que les fonctions qui sont requises par certaines applications spécifiques seulement doivent être exécutées en bordure de réseau. Ainsi, la complexité et l'intelligence du réseau sont repoussées vers ses lisières. Des réseaux simples pour des applications intelligentes. » (source wikipedia[3])

Des réseaux simples aux services uniformisés ont donc fait le succès d'Internet. L'intelligence est déployée sur les machines aux extrémités du réseau (machines terminales). Les équipements à l'intérieur du réseau traite le plus "bêtement" (et le plus rapidement) possible les paquets de données (figure 2.1).

Ce principe peut être considéré comme une des bases de la neutralité des réseaux. Les principaux services cités par l'article de Saltzer [152] concernent la tolérance aux pannes, les reprises sur erreur, sécurité avec cryptage, suppression de messages dupliqués et la gestion des accusés de réception. Le protocole de transport TCP (*Transmission Control Protocol*) développé depuis 1973 [43, 146] est le plus bel exemple d'application du principe de bout en bout. L'intelligence nécessaire au protocole (accusés de réception, vérification de l'ordre des paquets de données, retransmission, évitement de congestion...) est uniquement cantonnée sur les machines terminales. Le réseau est donc considéré comme un simple transporteur de paquets.

Pourtant ce modèle est mis-à-mal depuis plusieurs années : des équipements (*middleboxes* ([39, 126]) disposés dans le réseaux proposent un ensemble de services : Network Adress Translation (NAT), Tunnel IP, marquage de paquets, classifieur, firewalls, proxies, répartiteurs de charge.... Ces services, reposant sur des nœuds flexibles, souvent partagés par plusieurs applications, utilisateurs ou flux, apportent plus de flexibilité à l'usage des réseaux.

[2] "The principle, called the end-to-end argument, suggests that functions placed at low levels of a system may be redundant or of little value when compared with the cost of providing them at that low level. " [152]

[3] Extrait de Lawrence Lessig, L'Avenir des idées, 2005, Presses universitaires de Lyon

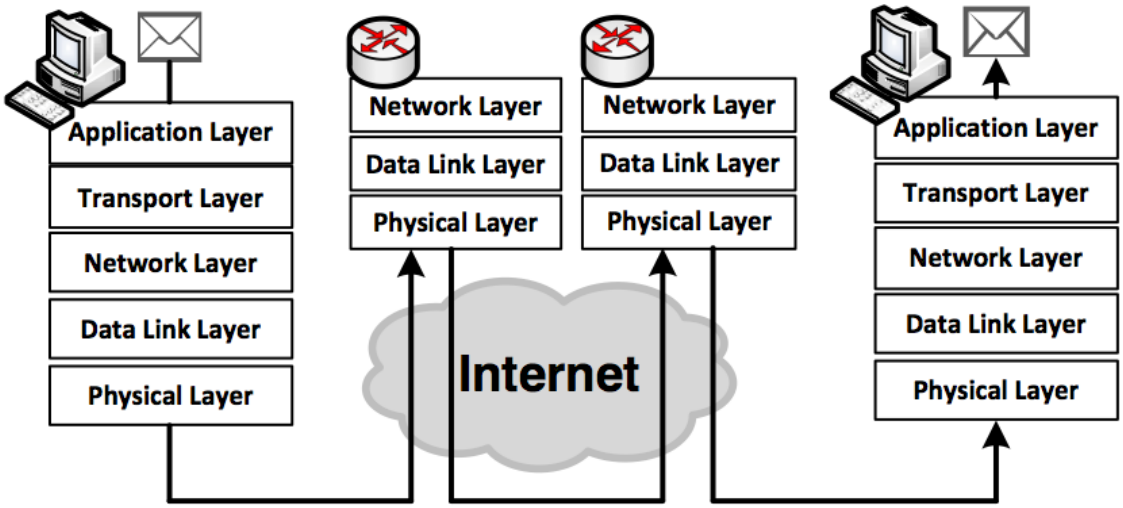


Figure 2.1: Le principe du bout en bout lors d’un échange de données entre deux machines (extrait de [126])

Si la porte est ouverte pour cette flexibilité, pourquoi ne pas aller encore plus loin et proposer des infrastructures distribuées à grande échelle reconfigurables dynamiquement en termes de services ?

2.2 Remise en cause : de “route and route” à “route, process and route”

Un réseau traditionnel (ou passif) est un réseau de transport de données qui possède un nombre restreint et fixé de services implantés dans les équipements et qui n’offre aucun moyen facile d’en ajouter. Par conséquent il est impossible de modifier dynamiquement le comportement global du réseau.

Or les opérateurs et fournisseurs de services ont un besoin crucial de flexibilité pour répondre dans un délai très court aux besoins des usagers. Des applications comme la téléphonie sur IP, la diffusion de radio ou de canaux TV, les services de cotations boursières ou encore d’achats aux enchères en ligne sur l’Internet sont largement développées sur les réseaux. Malheureusement, les services de base fournis par les protocoles existants du réseau sont mal adaptés à ces applications hétérogènes et avec des exigences particulières. Une multitude de protocoles de contrôle et de réservation de ressources (MPLS (*MultiProtocol Label Switching*), DiffServ, RSVP (*ReSerVation Protocol*),...) est disponible mais les constructeurs d’équipements ne peuvent pas intégrer tous ces nouveaux protocoles dans leurs matériels en un temps adapté aux besoins des usagers.

Au milieu des années 90, ce principe est bousculé avec de nouvelles propositions de nœuds flexibles. Deux initiatives sont apparues en parallèle en 1996-1997. Une initiative est apparue en 1997 au sein de l’IEEE sous la forme du projet P1520 (*Programmable Interface for Networks*) [24] afin de proposer le développement d’interfaces, de contrôle et de gestion pour la programmation des réseaux. Cette proposition de standardisation provient des travaux sur la signalisation dans les réseaux de télécommunications de type ATM. Parallèlement, David Tennennhouse et David Wetherall explorent et proposent le concept de réseaux actifs[160]. Un réseau actif est un réseau dans lequel les composants dans les différents plans (signalisation, supervision, données)

sont programmables dynamiquement par des entités tierces (opérateurs, fournisseurs de services, applications, usagers). Cette approche est issue des travaux sur l'insertion de services applicatifs dans le réseau Internet. Ainsi, un réseau programmable ou actif est un réseau de transport de données étendu par un environnement de programmation à l'échelle du réseau comportant un modèle de programmation des services, des mécanismes de déploiements et un Environnement d'Exécution (EE).

Un réseau actif contrairement à un réseau traditionnel n'est pas un simple support passif de paquets. Il peut être vu comme un ensemble de nœuds (routeurs) *actifs* qui réalisent des opérations personnalisées sur les flux de données qui le traversent. Il autorise les utilisateurs, les opérateurs, ou les fournisseurs de services à injecter leurs propres programmes dans les nœuds du réseau, permettant ainsi de modifier, stocker (cacher) ou rediriger le flux de données à travers le réseau. Ainsi un routeur actif embarque de nouvelles fonctionnalités qui transforme son comportement d'une approche *"route and route"* à un mode *"route process and route"* (figures 2.2 et 2.3).

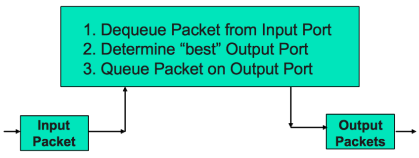


Figure 2.2: Du Modèle "route & route" ...

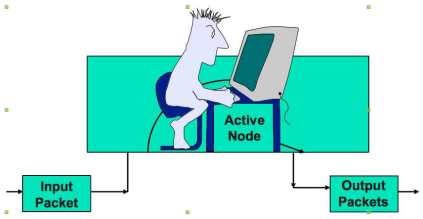


Figure 2.3: ... au Modèle "route & *process* & route" (l'homme peut être remplacé par un service)

Deux grands approches de configuration sont disponibles : par paquet (les paquets de données contiennent le code à déployer dynamiquement sur le nœud actif) ou par configuration du nœud actif (qui se charge de déployer les services demandés par les flux). Le composant principal d'un routeur actif est l'environnement d'exécution qui est chargé d'héberger les services et fonctionnalités logicielles qui sont déployés sur l'équipement.

D'un point de vue recherche et développement, ces réseaux proposent un formidable terrain d'investigation pour la mise au point grandeurs nature de nouveaux protocoles et services avant qu'ils ne soient (peut être) déployés un jour dans des équipements standards. Mais ces différents travaux font aussi apparaître très nettement les verrous qui sont la sécurité, la performance et le support de l'hétérogénéité. Bien qu'il existe de nombreuses propositions, les prototypes opérationnels proposés sont peu nombreux, peu transposables ou peu fonctionnels.

2.3 Réseaux actifs hautes performances : l'approche Tamanoir

Nos premiers travaux menés sur les réseaux actifs ont eu lieu pendant la thèse de Jean-Patrick Gelas[76] (2000-2003) que j'ai co-encadrée avec Bernard Tourancheau (Université Claude Bernard, Lyon1).

Lors de nos premières investigations sur les réseaux actifs, la plupart des systèmes disponibles (tels que ANTS [171]) permettaient l'envoi de capsules avec des bandes passantes très limitées (de l'ordre de quelques Mbits). Pourtant, dans l'équipe RESO, j'étais confronté à des demandes en bande passante et en latence d'un autre niveau. Le Gbits était la norme [82].

Nous avons donc proposé un nouveau modèle de réseaux programmables et l'avons implémenté

afin de supporter les performances des réseaux actuels. Nous pensons que le gain en performance est étroitement lié à la localisation des composantes de notre architecture. On parlera :

- d'urbanisation ou de déploiement horizontal pour traiter du positionnement d'un équipement actif ou d'un service dans le réseau;
- de classification de services ou déploiement vertical pour trouver la couche logicielle et matérielle la mieux adaptée pour l'exécution d'un service au cœur d'un nœud actif.

2.3.1 Urbaniser les services : où placer la flexibilité ?

Je n'ai jamais cru au côté utopiste des réseaux actifs où tout le monde a le droit de modifier le comportement du réseau en déployant des services ou en injectant du code dans chaque paquet de données sans contrôle ni contrainte. En reprenant le principe du bout en bout, je pense que toutes les parties d'un réseau n'ont pas besoin de flexibilité. Ainsi les réseaux de cœurs (*backbones*), largement dominés par les infrastructures optiques, sont très souvent sur-dimensionnés. Ils sont rarement à l'origine des problèmes de performances réseaux. Par contre les réseaux terminaux (dernier kilomètre), les réseaux d'accès, les réseaux métropolitains sont des points d'engorgement potentiels où la valeur ajoutée de nouveaux services peut être prépondérante.

Dans l'approche que nous suivons, seuls les équipements réseaux (routeurs, passerelles) de bordure de la couche d'accès sont dotés de capacités de traitement aptes à embarquer de la flexibilité (figure 2.4). Ces équipements (appelés nœuds) sont donc actifs, programmables et configurables . Ils peuvent être réalisés sous la forme d'un routeur avec une capacité de calcul ou avec une machine dotée de plusieurs interfaces réseaux. Ils peuvent être déployés en structuration hiérarchique, massive ou sporadique, et localisés dans des points précis de l'infrastructure réseau (point de *peering*).

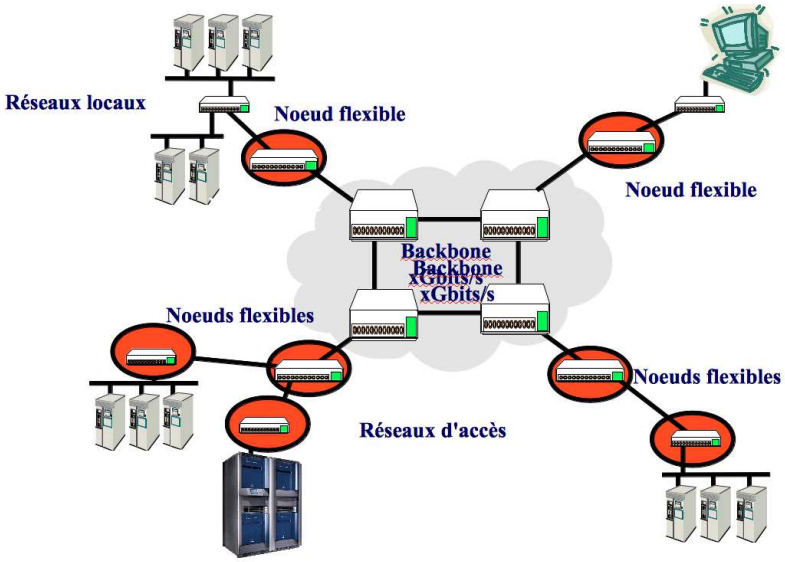


Figure 2.4: Déploiement de flexibilité dans l'architecture réseau

Les équipements réseaux grâce à leur capacité de traitement peuvent exécuter différents services. Différentes techniques de déploiement sont disponibles et nous supportons trois d'entre elles :

- services massivement déployés (Figure 2.5 - III) : le service est déployé sur tous les nœuds traversés par le chemin des données;

- déploiement spécifique : un équipement actif et programmable sur le chemin des données (ou ailleurs) est sélectionné pour recevoir un nouveau service (Figure 2.5 (I)). Ce service peut être composé de différentes fonctionnalités logicielles (A,B,C);
- services composés (figure 2.5 (II)) : les fonctionnalités du service (A,B,C) sont déployées sur différents nœuds actifs présents sur le chemin des données.

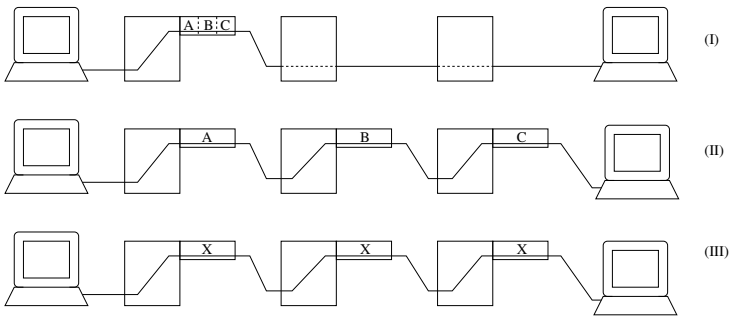


Figure 2.5: Déploiement de services dans le réseau : spécifique (I), composé (II) et massif (III)

Certains déploiements peuvent avoir une approche "*best-effort* : si le service est présent, les applications en profitent sinon tant pis. Le mode de composition de service peut poser différents problèmes de tolérance aux pannes, car il suppose une maîtrise précise de l'infrastructure réseau afin de supporter un mode *pipeliné* de traitement des paquets de données.

L'urbanisation de services peut être **anarchique** (tous les utilisateurs ont le droit de déployer), **contrôlée** (un sous-ensemble d'utilisateurs a le droit de déployer) ou **réservée** (seul l'opérateur réseau peut déployer).

2.3.2 Quels services ?

Certains travaux [74] ont montré qu'il est extrêmement complexe de réaliser des prévisions sur la quantité de ressources nécessaire au bon fonctionnement d'un service (CPU, mémoire, stockage, réseau). Dans nos travaux, nous considérons 4 classes de services [103] :

- **poids plume** : peu consommateurs de CPU, peu consommateurs de mémoire. Ce sont des services sans états, qui appliquent un traitement extrêmement léger sur les paquets de données. Ces services peuvent être exécutés au plus près des liens sur une carte d'interface réseau programmable. Nous avons étudié des fonctionnalités "poids plume" avec des propositions de services hautes performances embarqués dans une carte réseau programmable (KNET section 2.4) ou des services de support aux protocoles de transport (XCP section 3.2);
- **poids léger** : peu consommateurs de cycles CPU ces services ont besoin de mémoire pour pouvoir y stocker quelques états. Ces services peuvent être exécutés dans l'espace noyau du système qui fournit un accès total à la mémoire du système. Nous profitons du fait que nous n'avons pas encore traversé la barrière espace noyau/espace utilisateur pour conserver d'excellentes performances. Nous avons proposé des services "poids léger" de répartition de charge (section 3.1.2) ou de pare-feux à état hautement disponibles (section 3.1.3);
- **poids moyen** : qui demandent un environnement riche pour pouvoir effectuer des traitements complexes. Ces services s'exécutent dans l'espace utilisateur donc dans un espace protégé, ne risquant pas de mettre le nœud actif en péril. Le service peut accéder à toutes les ressources matérielles mises à sa disposition sur le nœud (mémoire, disques, cartes spécialisées, réseau). Nous présentons, dans le chapitre 3, un ensemble de services "poids

moyen" de réseaux logistiques (section 3.3.4) et caches web (section 3.3.4.2) ainsi que des services de support de Grille active (section 3.3);

- **poids lourd** : très consommateurs en ressources de calcul ou en mémoire. Ces services ont besoin d'une architecture distribuée pour être exécuté sans pertes de performances. La parallélisation peut intervenir à deux niveaux de granularité différents : au niveau paquet ou au niveau flux. Dans le premier cas les paquets sont distribués sur les unités de traitement qui leur appliquent le service. Dans le second cas, un flux complet est associé à une unité de traitement. Les flux sont alors traités en parallèle. Il est ensuite important d'employer un algorithme bien adapté de distribution des paquets ou des flux, en fonction de la granularité choisie. Dans le cadre de différentes collaborations, nous avons proposé des services "poids lourd" d'adaptation multimédia, et d'adaptation d'applications à la volée (chapitre 3).

2.3.3 Proposer une architecture de nœud flexible à 4 niveaux

Nous devons proposer une architecture de nœud flexible qui soit capable de répondre aux besoins hétérogènes des services considérés. Ce nœud flexible doit traiter un ensemble de flux ou de paquets de données. Il doit donc être optimisé pour offrir les meilleurs performances de traitement à ces données. Nous proposons un environnement d'exécution en couche qui autorisent le déploiement de services. Cette architecture à quatre niveaux est illustrée figure 2.6.

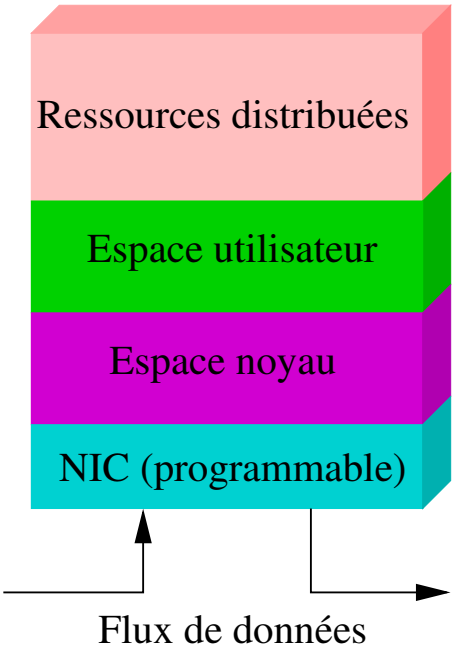


Figure 2.6: Architecture d'un nœud flexible

Les 4 couches ont chacune leurs spécificités :

- La couche **NIC** (*Network Interface Card*) embarque des services sur la carte ou le port réseau du nœud flexible. Les ressources y sont limitées et la création de services complexes. Cette couche de bas niveau permet le déploiement de *services poids plume*;
- La couche **Espace Noyau** se situe dans la zone où s'exécute le noyau du système d'exploitation du nœud flexible. Cette couche permet d'obtenir de bonnes performances de traitement bas niveau mais nécessite l'emploi de langages proches du processeur hôte (assembleur, C compilé). Dans cette couche, le paquet est dé-multiplexé par un Environnement d'Exécution

(EE) allégé appelé μ EE ce qui permet de s'affranchir des recopies coûteuses de l'espace noyau vers l'espace utilisateur. Cette couche de niveau système permet le déploiement de services de type "poids léger";

- La couche **Espace Utilisateur** autorise l'utilisation de toutes les ressources du système (mémoire de masse (disques durs), cartes dédiées (ex: compression, cryptage, GPU), cartes réseaux). C'est une couche applicative qui permet d'embarquer un environnement d'exécution utilisant des langages portables (Java, Perl). Cette couche de haut niveau permet le déploiement de services "poids moyens";

- La couche **Ressources Distribuées** permet de rattacher un ensemble de ressources de calcul ou de stockage au nœud flexible. Ces ressources hébergent des services nécessitant une utilisation intensive de certains composants matériels (CPU, mémoire, stockage). Cette couche est basée sur une infrastructure matérielle composée d'une machine frontale (*front-tend*) reliée à des machines esclaves (*back-end*). L'équipement frontal est chargé de distribuer les paquets ou les flux sur les équipements actifs situés en interne. Cette couche de haut niveau permet le déploiement de *services lourds*.

Des mécanismes de communication inter-couches sont implémentés sur le nœud flexible. Chaque couche possède ses avantages et ses inconvénients en termes de performances, sécurité, facilité de programmation, ré-utilisabilité...

La figure 2.7 présente le cheminement réalisé par un flux traversant l'architecture de nœud flexible. Plus la traversée est longue, plus la latence sera importante sur les paquets de données. Une latence supplémentaire sera induite par les services déployés. Un flux nécessitant un service "poids plume" localisé dans la couche NIC traversera le nœud flexible avec une latence réduite (Flux Ni). Les flux N et DN présentent le cheminement des paquets de données nécessitant une couche de traitement au niveau de l'espace noyau (avec ou sans la couche distribuée), alors qu'un flux U ou DU traverse la couche de l'Espace Utilisateur. Si un flux nécessite un traitement dans la couche Ressources Distribuées (flux DN et DU), les paquets traversent d'abord les couches de la machine frontale avant d'atteindre le *back end*.

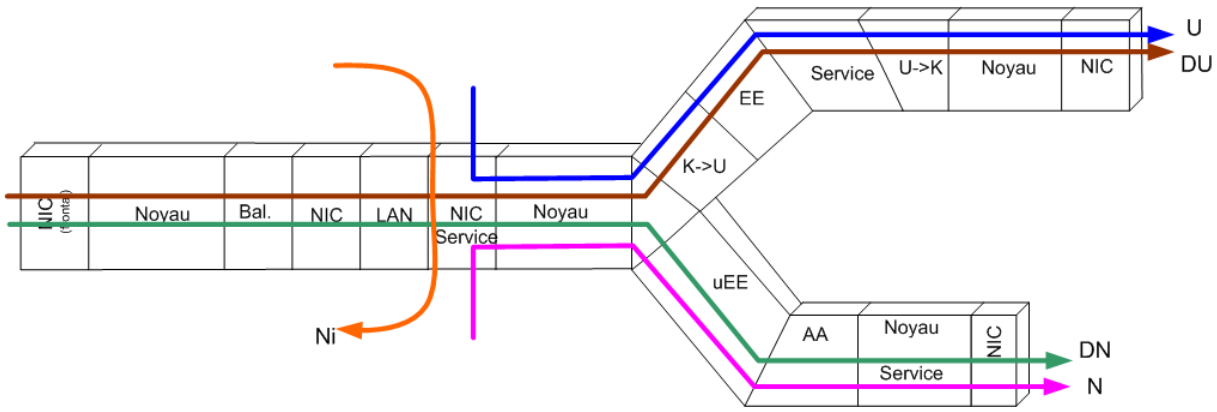


Figure 2.7: Cheminement des flux à travers les couches du nœud flexible

2.3.4 Le projet Tamanoir : supporter le déploiement de services flexibles hautes performances

Nos propositions de *nœuds flexibles actifs* orientés haute performance ont été implémentés sous la forme d'un environnement d'exécution actif appelé *Tamanoir*[4], qui adresse les problèmes de performance, d'hétérogénéité et de déploiement dynamique de services[78, 79]. Le cœur de l'environnement logiciel Tamanoir a été développé par Jean-Patrick Gelas pendant sa thèse. Le logiciel Tamanoir est déposé auprès de l'Agence de Protection des Programmes. L'environnement *Tamanoir* fournit aux utilisateurs la possibilité de déployer et de maintenir dynamiquement des nœuds actifs, appelés TAN pour *Tamanoir Active Node* (figue 2.8), distribués sur un réseau à grande échelle et hébergeant des services hétérogènes dynamiquement déployés.

2.3.4.1 Dans l'espace utilisateur

Un nœud actif Tamanoir simple ou *Tamanoir Active Node* (TAN) (figure 2.8) est constitué de trois éléments : un Environnement d'Exécution (EE), un gestionnaire local appelé *Active Node Manager* (ANM) et des services.

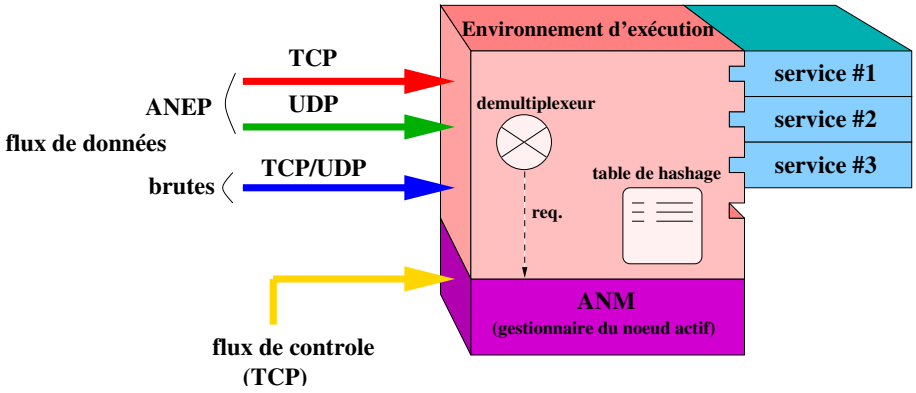


Figure 2.8: Architecture de nœud actif Tamanoir (*Tamanoir Active Node*)

- L'environnement d'exécution **EE** traite des flux de données brutes ou de format ANEP (*Active Network Encaspulation Packet* : un format de paquet pour le transport sur IP qui contient une référence au service actif [2]) avec les protocoles TCP ou UDP. Les en-têtes des paquets au format ANEP sont lus par un dé-multiplexeur qui en extrait la référence au service qui doit être appliqué sur le paquet. Si le service requis est indisponible l'EE émet une requête au gestionnaire local (ANM) afin de provoquer son installation dynamique. L'environnement d'exécution repose sur une architecture multi-processus qui permet la gestion efficace et simultanée des flux.

- Le gestionnaire de nœud actif (*ANM : Active Node Manager*) prend en charge le déploiement et la gestion des services. Il embarque un ensemble de modules permettant la surveillance du nœud TAN et des services associés.

- Les services de la couche **Espace Utilisateur** sont déployés dynamiquement et rattachés au nœud TAN. Dans cette couche les services sont programmés en Java afin de garantir leur portabilité et une facilité d'interfaçage sur de nombreuses plate formes.

[4] Le tamanoir, aussi connu sous le nom de grand fourmilier, se nourrit exclusivement de fourmis (30,000 par jour). Ce nom a été choisi en référence au projet ANTS (fourmis) [170].

Les figures 2.9, et 2.10 présentent les résultats de traitement de services "poids légers" (comptage de paquets) et "poids lourds" (compression à la volée) déployés dans l'espace utilisateur pour une infrastructure Tamanoir (sur machines bi-processeurs) déployée sur des réseaux Gbits. Les performances associées bénéficient de l'utilisation de paquets supérieurs à 8KB. Dans le cas du services lourds, la fonction de compression à la volée provoque un fort ralentissement de bande passante.

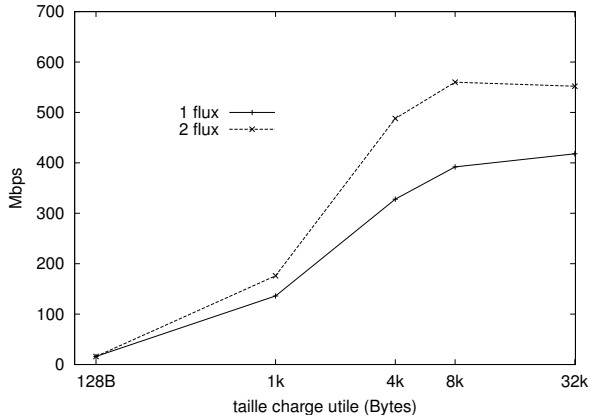


Figure 2.9: Débits obtenus sur réseau Giga Ethernet en TCP avec un service moyen

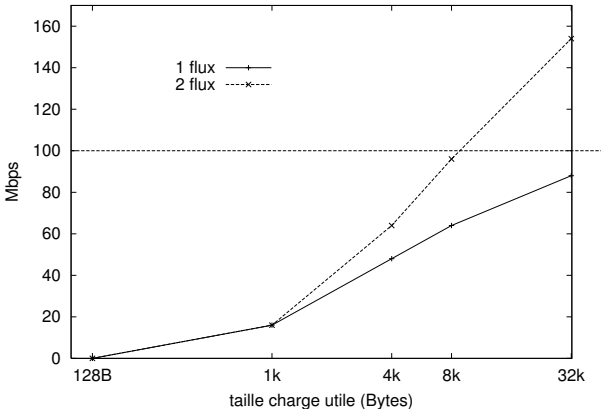


Figure 2.10: Débits obtenus sur réseau Giga Ethernet en TCP avec un service lourd

2.3.4.2 Dans l'espace noyau

Dans l'espace noyau, le nœud Tamanoir doit être capable d'intercepter des paquet s de données à la volée, de les dérouter momentanément afin d'appliquer un service et de les renvoyer vers les canaux de sortie.

Afin de pouvoir réaliser de telles opérations (filtrage de paquets, modifications de paquets, NAT) l'environnement Netfilter est disponible dans le système Linux. L'outil, *IpTables*, qui s'exécute dans l'espace utilisateur, permet de paramétrer ces trois opérations majeures. Netfilter, avec le protocole IPv4 définit un ensemble de *hooks* (accroches) qui sont des points précis dans le trajet du paquet (figure 2.11). Netfilter permet ainsi l'accroche de services (modules écrits en C) sur un hook (en entrée, en sortie, avant ou après les décisions de routage) dans l'espace noyau.

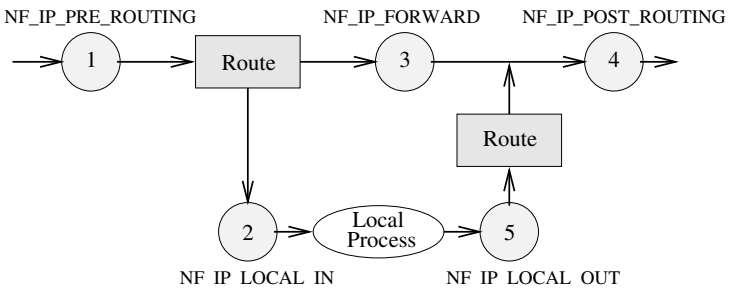


Figure 2.11: Hooks de Netfilter

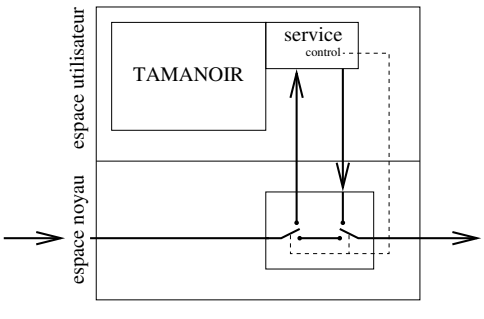


Figure 2.12: Mécanisme de communication entre le service exécuté dans l'espace utilisateur et son homologue exécuté dans l'espace noyau.

L'environnement logiciel Netfilter a été utilisé dans une partie des recherches que j'ai encadré sur les répartiteurs de charges (section 3.1.2), les pare-feux à état (section 3.1.3)). Un mécanisme de communication est déployé entre les couches Utilisateur et Noyau afin de contrôler la configuration de services noyau et d'alléger la charge de traitement dans l'espace utilisateur (figure 2.12). Dans la figure 2.13, on observe la latence ajoutée par le traitement des paquets dans l'espace utilisateur (500 premier paquets de données). Au bout de 500 paquets, le service dans l'espace utilisateur active un service de filtrage dans l'espace Noyau afin de ne plus recevoir les paquets suivants qui sont transmis vers un autre TAN.

Les mesures de latence présentées dans les figures 2.14 ont été réalisées avec deux JVM (IBM et SUN) ayant le compilateur *Just-In-Time* activé ainsi que sur une version compilée d'un TAN avec GCJ (Java compilé[5]).

Le service léger (comptage de paquets) déployé dans l'espace Noyau ajoute une latence entre 5 et 7 μs quelle que soit la taille des paquets alors le même service déployé dans l'espace utilisateur ajoute une latence plus grande d'un facteur 1000 !

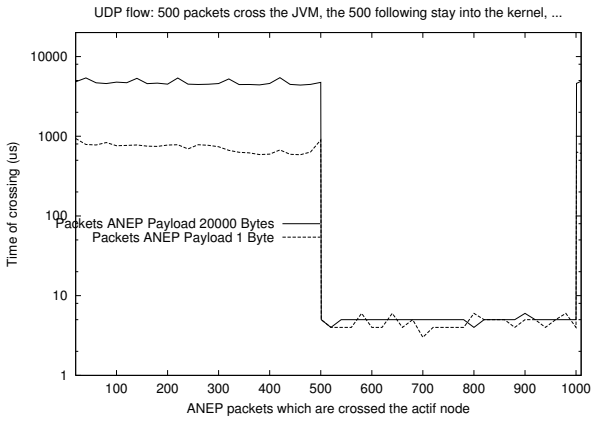


Figure 2.13: 500 paquets traversent la JVM, les 500 suivants sont transmis par le noyau directement sur l'interface de sortie.

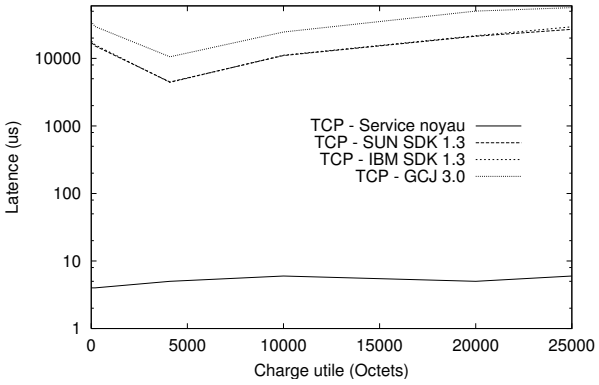


Figure 2.14: Temps de traversée d'un nœud actif Tamanoir par un paquet ANEP sur `TCP`.

2.3.4.3 Tamanoir : un équipement flexible à base de *cluster*

Nous avons observé que des services lourds déployés dans l'espace utilisateur peuvent avoir un impact très négatif sur les performances de l'infrastructure (figure 2.10). Nous proposons d'agréger un ensemble ressources distribuées à l'intérieur du nœud flexible Tamanoir afin d'effectuer les services nécessitant de la puissance de calcul ou de stockage. L'architecture de routeur cluster Tamanoir repose sur la technologie LVS (Linux Virtual Server)[177] qui est en charge de distribuer des requêtes sur une batterie de serveurs, pour distribuer la charge de traitement des paquets actifs. (figure 2.15).

Un serveur virtuel Linux (*Linux Virtual Server* : LVS) [177] est constitué d'un groupe de serveurs, appelés *backend* et d'une machine frontale (ou *frontend*). Cet ensemble de machines forme le routeur cluster virtuel qui apparaît comme une seule machine pour les clients. Chaque client croit être connecté directement au *back-end* et vice-versa. Les applications clientes et les *backend* n'ont pas le moyen de détecter qu'un *frontend* est intervenu dans la connexion.
Un serveur Linux virtuel n'est pas une grappe de machines destinée à calculer des petites parties d'un grand problème de façon coopérative. Les *backend* ne coopèrent pas, ils n'ont pas connaissance de la présence de leurs voisins.

[5]GCJ : Gnu Compiler for Java `http://gcc.gnu.org/java/?`

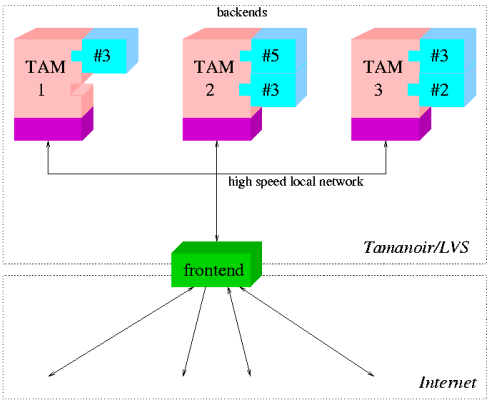


Figure 2.15: Vue logique d'un nœud Tamanoir cluster. L'EE Tamanoir de l'espace utilisateur est dupliqué sur plusieurs nœuds (backends) auxquels on accède à travers un frontal (*frontend*) qui distribue les connections

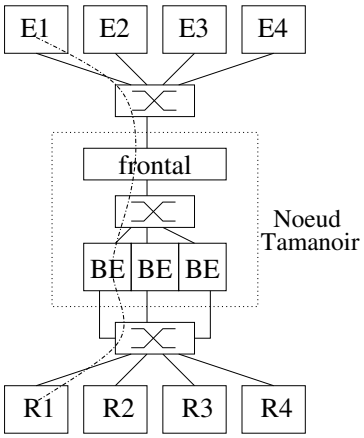


Figure 2.16: Vue réelle de la plateforme d'expérimentation Tamanoir cluster sur Myrinet.

La figure 2.16 présente une vue logique de la plateforme d'expérimentation locale. Elle met en évidence le parcours des paquets ANEP émis par un émetteur E vers un récepteur R. On constate que les paquets traversent trois commutateurs Myrinet (bande passante maximale 1.6 Gbits) [26]. C'est la machine frontale qui est en charge de sélectionner un BE.

Les figures 2.17 et 2.18 présentent les bandes passantes d'un nœud Tamanoir cluster à trois Back Ends (BE) sur un à douze flux TCP simultanés pour des services légers et lourds déployés dans l'espace Ressources Distribuées. Dans la figure 2.17, pour des paquets de 8 à 32kB plus le nombre de flux est important plus le débit agrégé se réduit. Contrairement à ce que l'on pourrait supposer, cette expérience ne met pas en évidence la limite du nœud Tamanoir mais celle des clients émetteurs. Bien que les émetteurs soient des machines performantes, bi-processeurs, le nombre de flux théorique optimal qu'elle devraient être-à-même d'émettre est de deux, soit un flux par processeur. Nous avons généralement employé trois machines pour émettre et trois machines pour recevoir. Donc le nombre de flux optimal est de six, soit deux flux par émetteur. Mais les émetteurs sont équipés d'une carte d'interface réseau unique. C'est donc dans celle-ci que des contentions apparaissent et ainsi grèvent les performances globales. Il est donc nécessaire d'avoir autant de postes émetteurs que de flux.

La figure 2.18 présente les résultats d'un service lourd appliqué par un nœud Tamanoir à 3 Back Ends (BE) sur un à douze flux TCP simultanés. Les débits agrégés bénéficient de la taille des paquets. Pour 6 ou 12 flux, les débits affichés par les courbes sont sensiblement identiques. Cela montre que pour 6 flux nous avons atteint la capacité de traitement maximum pour cette configuration de nœud actif (à 3 BE) et ce service. Soit un débit maximum de 270Mbps. On tire partie de l'architecture biprocesseurs car de 3 à 6 flux nous doublons littéralement de débit.

Nous avons déployé une infrastructure Tamanoir sur la plate-forme RNRT VTHD : 3 sites ont été mis à contribution : Grenoble (9 machines), Lyon (16 machines) et Rocquencourt (30 machines). Ces expériences nous ont permis de valider les performances de l'architecture Tamanoir sur une petite échelle autour d'un réseau de cœur au GBits. Alors que les systèmes de réseaux actifs à capsule ne fournissent quelques Mbits de bande passante, la solution obtenue avec Tamanoir présente des résultats de performance intéressants qui ouvrent la porte à différents services.

J'ai encadré le stage de fin d'études INSA et de Pierpaolo Giacomin sur des techniques d'équilibrage dans les routeurs *clusters*. Ces travaux sont à la base des activités que j'ai menées par la suite avec Narjess Ayari sur les répartiteurs de charge tolérants aux pannes (section 3.1.2).

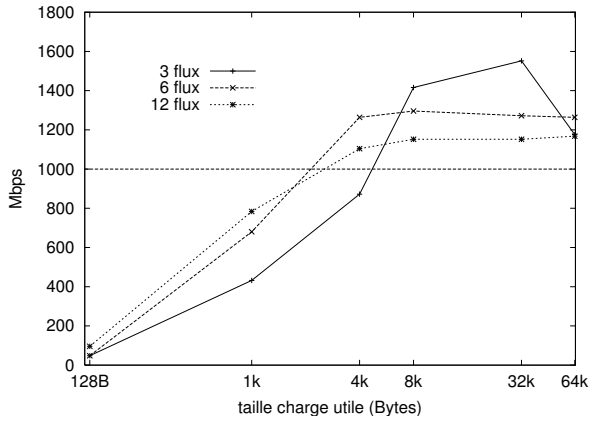


Figure 2.17: Débits sur réseau Myrinet avec un Tamanoir cluster à 3 BE appliquant un service léger.

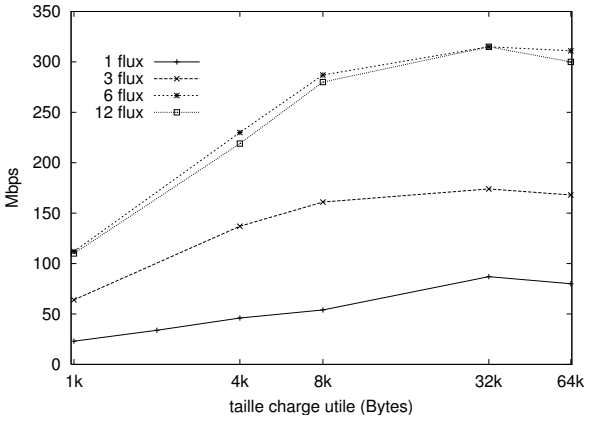


Figure 2.18: Débits sur réseau Myrinet avec un Tamanoir/3BE appliquant un service lourd.

Lorsque l'on déploie un ensemble de ressources distribuées, il convient d'utiliser efficacement les ressources nécessaires et de mettre en veille les ressources non-utilisées. Cette technique est actuellement en cours de développement dans le cadre de grappes de machines embarquant des services de passerelles maison virtualisées (*Virtual Home Gateway section* 7.2.1). Les travaux de [86] clament qu'une répartition de charge circulaire est suffisante dans les routeurs clusters actifs car ils ne déploient que des services homogènes en consommation de ressource. Tamanoir, avec son architecture à 4 niveaux, permet le déploiement de services très hétérogènes auxquels il faut pouvoir fournir suffisamment de ressources. La durée et le volume de ressources consommées par un service ne peuvent être connus à l'avance et la prédiction de l'impact d'un service ne peut être utilisée dans les réseaux programmables [73]. Nous avons donc exploré des solutions dynamiques en proposant la politique d'équilibrage FBSb (*Feedback stream based*) qui répartit efficacement et dynamiquement les flux de données et les services nécessaires sur les infrastructures distribuées d'un routeur cluster.

LVS ne fournit qu'un ensemble limité de stratégies d'équilibrage de charge utilisées par la machine frontale pour répartir les connections sur les machines esclaves (*back-end* figure 2.16) suffisantes pour des flux homogènes : ordonnancement circulaire (*round robin*) ou moins utilisé (*Least Connected*) quand le frontal choisit la machine esclave qui a reçu le moins de connections.

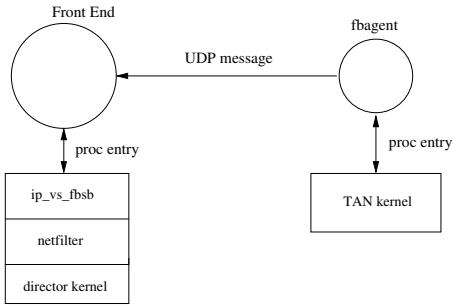


Figure 2.19: Politique d'équilibrage *FBSb*

Nous proposons une nouvelle stratégie basée sur le concept de boucle de rétroaction. L'architecture FBSb (figure 2.19) repose sur :

- un ensemble d'agents qui collectent la charge des machine escales. La charge *CPU* est

évaluée par lecture des compteurs systèmes et renvoyée au composant collecteur en UDP;

- un collecteur de charge frontal qui gère une table de charge des machines esclaves, récupère des informations venant des agents et prend des décisions d'équilibrage appropriées. Le collecteur incrémente virtuellement la charge d'une machine esclave choisie après l'attribution d'un flux, cette charge est ensuite remise à jour avec les véritables valeurs de charge collectées. Ceci évite de sélectionner très rapidement la même machine esclave quand de nombreux nouveaux flux arrivent en même temps sur le routeur cluster.

Nous évaluons cette politique sur un routeur cluster Tamanoir composé de 4 machines connectées avec 12 machines clients sur un réseau Myrinet (taille des paquets : *8192* octets) . En reprenant la classification de services Tamanoir, nous évaluons *FBSb* avec deux type de services : service lourd (*L: cryptage 3-DES*) qui sature 1 *CPU* avec un occurrence, service moyen (*M : analyse de paquet*) qui sature 1 *CPU* avec 4 occurrences.

- *20**: 1 flux requiert un service lourd, les 11 autres flux utilisent des services moyens;
- *30**: suite d'appels à 1 service lourd et 3 services moyens se répète 4 fois.

Déploiment	Services
20*	LMM MMM MMM MMM ("presque homogène")
30*	LMM LMM LMM LMM ("très hétérogène")

Figure 2.20: Scénario de déploiement de services

Chaque service dure en moyenne 1 (*01) ou 2 secondes (*02). Nous avons comparé la politique *FBSb* avec "l'ordonnancement circulaire" (*Round Robin RR*) et la politique du "moins utilisé" (*Least Connected LL*) (table 2.21) :

	FBSb				RR				LC			
TEST	MAX	AVG	MIN	SDEV	MAX	AVG	MIN	SDEV	MAX	AVG	MIN	SDEV
201	38.22	20.20	15.98	3.31	43.25	20.13	15.51	4.30	38.50	20.31	15.43	3.75
202	33.31	20.04	15.73	3.09	37.49	20.12	15.41	3.93	36.44	20.17	15.47	3.58
301	50.83	24.63	15.47	8.87	90.31	26.93	16.12	14.87	64.01	25.60	15.40	10.43
302	50.96	23.29	15.38	8.78	92.90	25.53	15.92	13.61	57.83	25.07	15.68	10.52

Figure 2.21: Comparaison des 3 politiques : FBSb, RR et LC en temps minimum(s), temps maximum(s), temps moyen(s) et déviation standard

On peut observer que *RR* et *LLC* fournissent des résultats efficaces dans les scénarios quasi homogènes (20*). Quand la distribution de services devient fortement hétérogène, la politique *FBSb* démontre des qualités de meilleur équilibrage de charge et garantissent ainsi une meilleure qualité de service au flux traités par le routeur cluster.

2.4 KNET : la flexibilité et la performance dans les interfaces réseaux programmables

Dans la thèse de Eric Lemoine[109] (CIFRE SUN Labs-INRIA, 2001-2004) que j'ai co-encadrée avec Cong-Duc Pham, nous avons étudié la possibilité d'apporter encore plus d'intelligence et de flexibilité dans les réseaux afin d'augmenter la performance des systèmes de communication. Un des résultats de cette thèse a été la proposition de l'architecture réseau KNET [110] qui se

situe dans la couche **NIC** de notre architecture de nœud flexible .

Nous avons abordé la question suivante : **Comment ajouter de nouvelles fonctionnalités dans les cartes réseaux pour s'adapter aux spécificités des serveurs (machines multi-processeurs, multi-cœurs) afin d'augmenter les performances des protocoles de communications ?**

Un serveur doit traiter en entrée un volume de données. Le débit utile fourni par le serveur en sortie augmente en fonction du débit d'entrée (figure 2.22). Il existe un point de saturation (*MLFRR : Maximum Loss Free Receive Rate*) au delà duquel le serveur est saturé [124]. A partir de là, les performances du serveur s'écroulent et celui-ci n'est plus capable de renvoyer un service utile.

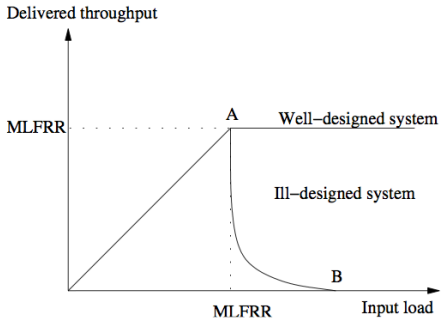


Figure 2.22: Effondrement des performances d'un serveur

Les cartes réseaux actuelles effectuent de nombreuses opérations et services : calcul de *checksum*, cryptage, hébergement de protocoles de transport (*TOE : TCP Offloading*). Dans les serveurs de calculs ou de données (type serveur web), un grand nombre de connexions doit être gérée par le serveur. Ces connexions sont établies entre un des processeurs de la machine serveur et un processeur distant sur une machine cliente.

Le système KNET ajoute de la flexibilité dans le traitement des paquets de données en réception sur la carte d'interface réseau d'un serveur [110]. Les serveurs étant multi processeurs et multi cœurs; il est crucial de délivrer le plus vite possible un paquet de données au processeur qui est concerné. L'approche choisie consiste à effectuer une classification des paquets au plus tôt dès leur arrivée sur la carte d'interface à l'aide d'une infrastructure en anneau par processeur présent sur le serveur. Cette solution permet plus de robustesse et de performances qu'un tri dans le noyau de la machine hôte.

KNET a été développé et déployé sur une carte d'interface programmable Myrinet [26], qui embarque différentes ressources (processeur, mémoire..), ouvertes et programmables. S'il est facile de concevoir rapidement et facilement un service sur une carte réseau programmable, la conception d'un service capable de supporter la très haute performance est beaucoup plus complexe.

La figure 2.23 illustre l'architecture de l'approche KNET. Les paquets de données venant du réseau sont directement traités par les services KNET présents sur la carte Myrinet qui les dirigent vers des files d'attente séparées et distinctes. Chaque file est utilisée pour alimenter en paquets de données les processeurs de la machine SMP. Les performances du système KNET dépassent les autres solutions (telles que NAPI[124] sous Linux) (figure 2.24).

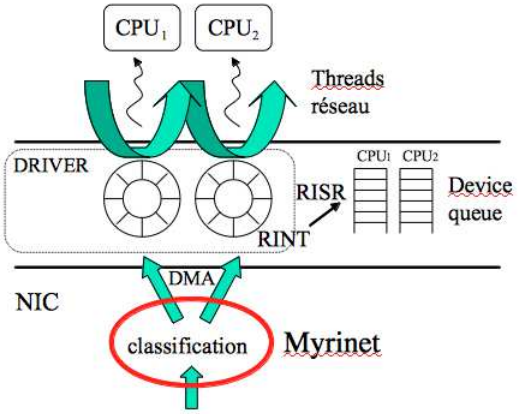


Figure 2.23: Architecture KNET

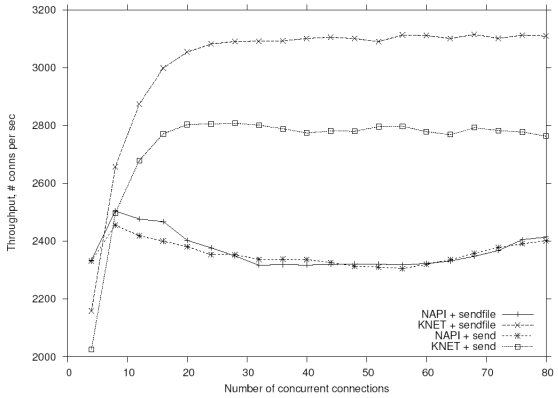


Figure 2.24: Performances de l'architecture KNET sur Carte réseau Myrinet

2.5 Réseaux autonomes à grande échelle : l'approche *AutoI*

Je me suis impliqué dans le montage et la réalisation du projet européen FP7 *Autonomic Internet (AutoI, 2008-2010)* mené par *University College of London* qui regroupe un ensemble de partenaires européens académiques (LIP6, INRIA, Université de Patras, Université de Passau, Université Polytechnique de Barcelone) et industriels (TSSG, Gingko Networks, Ucopia, Hitachi). *AutoI* adresse la problématique de la création de réseaux autonomes pour la validation de services réseaux à grande échelle. Les réseaux autonomes supportent les mêmes spécificités que le calcul autonome (*Autonomic Computing*) : auto-configuration, auto-défense, auto-réparation...

2.5.1 L'architecture *Autonomic Internet*

Autoi propose une architecture complexe qui permet aux opérateurs réseaux de déployer un ensemble d'infrastructures virtuelles aptes à supporter de nouveaux services. Différents plans orientés service (orchestration, connaissance, mise à disponibilité) sont déployés en soutien. Nos travaux se sont focalisés sur le plan de mise à disponibilité (*Service Enablers Plane*) (figure 2.25).

Nos contributions à cette architecture sont doubles : fournir un ensemble de solutions flexibles de type *programmatic enablers* autorisant la programmation à la volée d'équipements réseaux virtuels, déployer et valider à grande échelle des ensembles de réseaux virtuels programmables afin de vérifier la faisabilité et la maîtrise de ces plate-formes. Le premier challenge découle directement de nos activités en réseaux programmables. Le deuxième aspect (découlant de nos activités sur la maitrise de grandes infrastructures distribuées (section 2.6) a nécessité de nouvelles expertises et réalisations logicielles.

2.5.2 "*Programmatic enablers*"

Deux composants essentiels de l'architecture AutoI sont proposés pour la mise en œuvre de réseaux autonomes :

- VCPI (développé par l'Université de Passau) qui fournit la mise en œuvre de routeurs et de liens virtuels permettant de créer l'infrastructure réseaux;
- ANPI (développé par l'équipe INRIA RESO) qui fournit l'infrastructure logicielle pour le déploiement de services autonomes (figure 2.26).

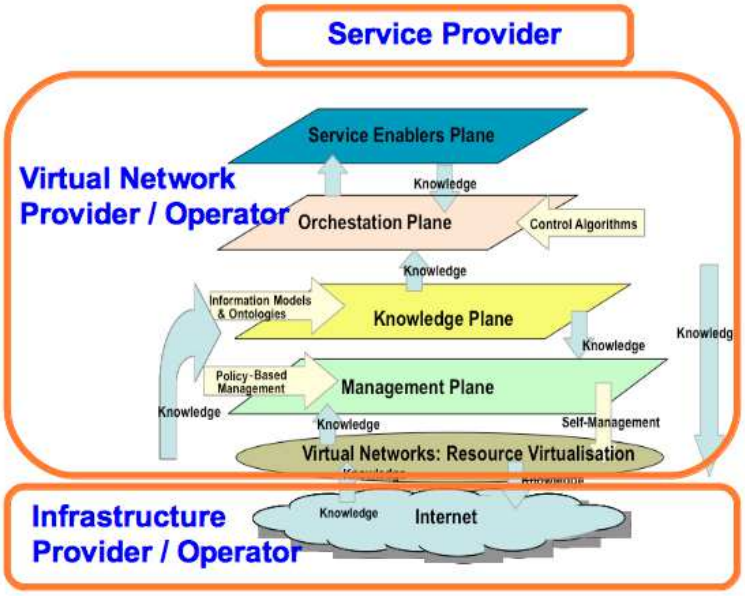


Figure 2.25: Architecture *Autonomic Internet*

ANPI est déployé comme un environnement d'exécution et supporte les fonctionnalités autonomiques (auto- déploiement, auto-configuration, auto-contrôle) des services. ANPI est aussi utilisé pour le déploiement de l'infrastructure virtuelle (réseaux et liens) sur les machines physiques.

Grâce à ANPI, différentes implémentations de services réseaux ont pu être testés et validés : *roaming* autonome de clients mobiles, adaptation des équipements réseaux au changement de topologies, tolérance aux pannes...

2.5.3 Orchestration du déploiement de réseaux virtuels à grande échelle

La difficulté dans la manipulation de systèmes distribués à grande échelle réside dans la mise en place d'expérimentations. Après nos expériences sur la plate-forme VTHD++ et la difficulté de mettre en place des expériences à grande échelle; nous avons décidé d'utiliser la plate-forme d'expérimentation nationale Grid5000 [37].

Dans *AutoI*, nous avons proposé l'environnement OVNI (*Orchestrated Virtual Network Interface*) corrélé avec la plate forme d'expérimentation Grid5000, c'est un des premiers environnements à autoriser le déploiement d'infrastructures réseaux virtuelles automatisé. OVNI permet la mise en œuvre de topologies réseaux classiques (*pipeline*, arbres, anneaux, graphe complet...etc) et libres. Ainsi la figure 2.27 présente le déploiement d'un réseaux de 150 routeurs virtuels en moins de 5 minutes sur Grid5000 ! Après cette opération, le routage et les routeurs sont opérationnels et la phase de déploiement de services peut commencer. L'infrastructure OVNI a été mise à disposition des autres partenaires européens du projet qui ont pu valider les plans de connaissance et de gestion autonomes à grande échelle.

Sur la plate-forme Grid5000, dans le cadre du stage de Pablo Pazos nous avons proposé l'environnement LSCAN (Large Scale Autonomic Networks) qui est une adaptation de l'environnement Nagios afin de manipuler graphiquement des nœuds autonomes à grande échelle sur la plate-forme Grid5000 (figure 2.28).

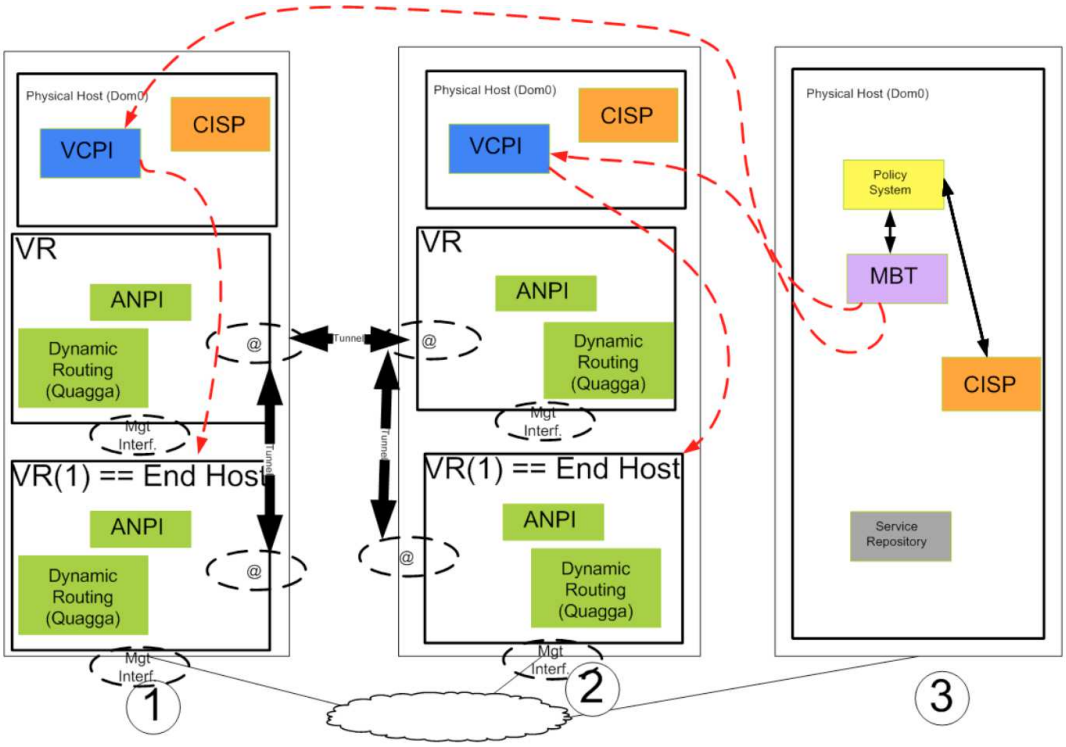
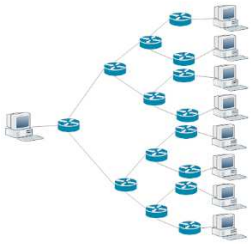


Figure 2.26: Autonomic Network Programmable Interface



Deployment of VR	10PM * 1VR = 10 VR	10PM * 3VR = 30 VR	50PM * 1VR = 50 VR	50PM * 3VR = 150 VR
Lyon	55s	3m18s	1m38s	3m42s
Bordeaux	57s	3m7s	1m14s	4m41s

Figure 2.27: Orchestrated Virtual Network Interface

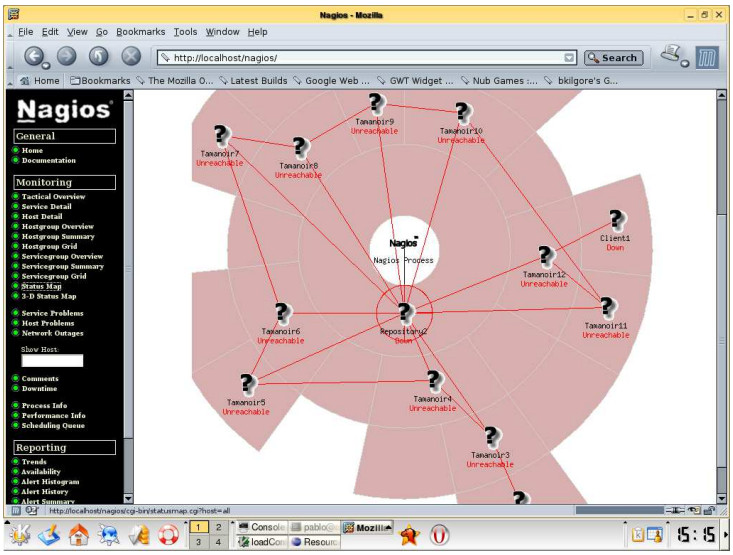


Figure 2.28: L*SCAN : Large Scale Autonomous Networks*

2.6 Conclusion

Tamanoir s'est révélé être une plate-forme de développement et des test facile à maîtriser pour évaluer différents services réseaux ou applicatifs. Afin de compléter cet environnement nous lui avons adjoint un ensemble d'outils pour gérer des expériences de flexibilité à grande échelle. Toujours dans la famille des fourmiliers, l'environnement Echidne[6] permet de déployer des infrastructures distribuées de machines virtuelles terminales qui sont coordonnées par des gestionnaires répartis afin d'émuler un grand nombre d'applications clientes.

L'environnement Pangolin[7] a été développé dans le cadre du projet VTHD++ afin de maîtriser une infrastructure de nœuds actifs Tamanoir à grande échelle. Basé sur l'outil Mapcenter [29], il permet le contrôle d'un ensemble de services et équipements actifs à travers un service web.

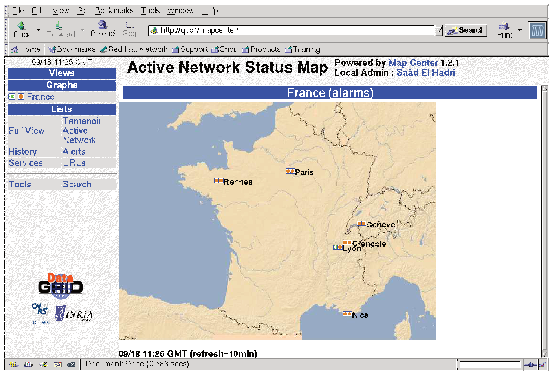


Figure 2.29: Visualisation du déploiement de nœuds actifs avec Pangolin

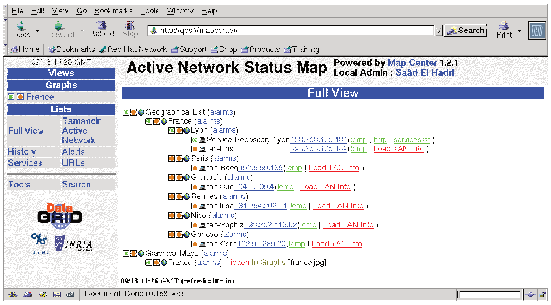


Figure 2.30: Gestion des nœuds actifs

La suite logicielle Tamanoir a servi comme composant à de nombreux travaux étudiants que j'ai encadré. Tamanoir a été au cœur de nombreux projets de recherche : RNRT VTHD++, RNTL e-Toile, RNRT Temic. Cette plate-forme ouverte et performante a permis de valider des services différenciés (voir chapitre 3) et de proposer des équipements réseaux flexibles (section 3.4).

6 Echidné : Mammifère ovipare, fouisseur et insectivore, couvert de piquants et dont le museau porte un bec corné. L'échidné vit en Australie, en Tasmanie et en Nouvelle-Guinée et pond des œufs. (source `wikitionary.org`)

7 Pangolin : Mammifère édenté d'Afrique et d'Asie du sud dont le corps est presqu'entièrement recouvert d'écailles (ostéodermes) larges, plates et triangulaires, dont la taille varie de 60 cm à plus d'un mètre pour le pangolin géant. Animal nocturne très craintif qui se roule en boule dès qu'il se sent menacé. Il se sert de ses griffes pour percer des trous dans les fourmilières et aspire ensuite les fourmis avec sa longue langue effilée. (source `wikitionary.org`)

Se mettre ensemble, c'est un début, rester ensemble, c'est du progrès et travailler ensemble, c'est le succès.

Henry Ford

3

Bénéficier de la flexibilité et de l'intelligence du réseau : équipements et nouveaux services

"Welcome back - You are again roaming on the Telstra network."[1]*. Il fait une chaleur d'enfer, je suis dans la vallée des vents (Kata Tjuta) en plein milieu du désert rouge de l'Australie. Nous sommes en 2005, et la 3G n'a pas encore envahi les réseaux mobiles, la couverture voix est encore faible dans certaines régions. Alors que j'aperçois quelques kangourous qui somnolent en toute tranquillité sous un eucalyptus; voir le téléphone mobile vibrer et recevoir ce SMS en plein milieu du désert rouge australien a un coté hallucinant et décalé. Je suis en plein bush, à plus de 500 kms de la plus grande ville (Alice Springs 20000 habitants), mais l'opérateur historique australien (Telstra) a jugé rentable de couvrir une partie de ce désert pour ses éventuels clients. En effet, à 60 kms de moi, se trouve Ayers Rock (Uluru), le plus célèbre monolithe rouge d'Australie qui attire 400000 visiteurs par an et autant d'utilisateurs de téléphonie mobile potentiels. Quand la rentabilité rentre en ligne de compte les opérateurs réseaux sont donc prêts à déployer des solutions complexes.*

Les services numériques (gratuits ou payants) sont partout. Le monde de l'Internet s'est transformé en un monde de services multimédia, hétérogènes et de grande envergure. Les services ont besoin de flexibilité pour s'adapter aux nouvelles contraintes imposées par les utilisateurs et leurs applications : plus de performances, de la robustesse, de la tolérance aux pannes, la prise en compte de la consommation énergétique, de l'adaptation pour les infrastructures hétérogènes (du datacenter au *smartphone*)...

De nombreux services pourraient bénéficier de la flexibilité des réseaux (figure 3.1). Ces services peuvent être cantonnés au plan de gestion et de contrôle (*monitoring*), fournir de nouvelles fonctionnalités (multicast fiable, *QoS*, transport de données...) ou concerner le plan de données (adaptation de flux, stockage à la volée...).

Dans le cadre de nos recherches sur la flexibilité dans les réseaux, nous avons validé nos propositions de réseaux flexibles et dynamiques en explorant différents équipements et services réseaux.

Nous présentons brièvement les travaux menés dans la conception d'équipements de flexibilité à contrainte industrielle (Projet RNRT Temic) et des équipements de haute disponibilité

[1] "Re-bienvenue - Vous êtes de nouveau en train d'utiliser le réseau Telstra", SMS, Valley of the winds, Uluru - Kata Tjuta National Park, Northern Territoty, 29 Juin 2005

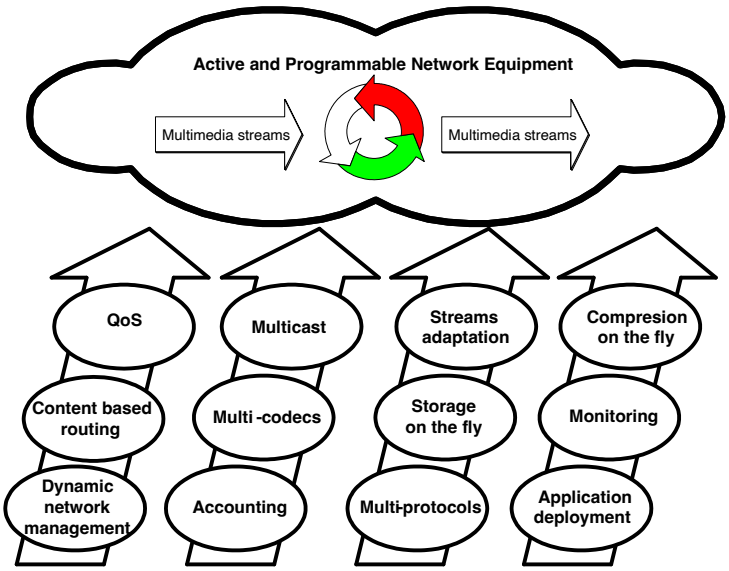


Figure 3.1: Services pouvant bénéficier de la flexibilité

pour héberger des services réseaux d'un opérateur télécom (dans le cadre de la thèse CIFRE de Narjess Ayari avec Orange R&D) ou de sécurité (pare-feu à état, en collaboration avec Pablo Neira Ayuso de l'Université de Séville).

Ce chapitre présente 3 exemples de services que nous avons étudiés :

- **Les protocoles à assistance de routeurs** : Dans la thèse de Dino Lopez Pacheco (2005-2008), co-encadrée avec Cong-Duc Pham, nous avons adressé le problème de l'interopérabilité de certains protocoles de transport avancés. Nous nous sommes focalisés sur le protocole XCP (eXtensible Control Protocol [94]) qui repose sur l'assistance des routeurs pour recevoir des informations sur l'état du réseau et ainsi décider du volume d'émissions des données. Ce protocole prometteur n'est fonctionnel que :
  - si la totalité des infrastructures réseaux est compatible et renvoie des informations;
  - si la totalité des flux transportés dans les réseaux sont es flux XCP (plus TCP).

  Partant de ce constat irréaliste, nous avons proposé XCP-i [115], qui reprend les principes fondamentaux de XCP en termes d'estimation de la bande passante et propose en outre des mécanismes d'interopérabilité afin de supporter des infrastructures non XCP et d'assurer l'équité lors de la cohabitation de flux XCP et TCP (section 3.2). Ces nouvelles fonctionnalités garantissent ainsi la possibilité de déployer incrémentalement des infrastructures distribuées à grande échelle.

- **La Grille Active** : J'ai proposé le concept de Grille Active (*Active Grid* [98]) qui intègre le rapprochement d'infrastructures réseaux flexibles avec des Grilles de calcul distribuées (section 3.3).

  Cette proposition [30, 81] a fédéré une partie des recherches des membres de l'équipe RESO et a abouti à des validations dans différents projets de recherche : le projet RNTL Etoile[167, 14], le projet RNRT VTHD++, le projet PAI FAST[101]. Dans ce contexte j'ai initié au cours du temps quatre collaborations internationales principales qui ont donné lieu à différents résultats : avec l'équipe du Professeur Micah Beck (Université du Tennesse, Knoxville, USA) sur l'intégration conjointe de réseaux actifs et logistiques [16], l'équipe du Professeur Alex Galis (University College of London, UK) sur la gestion d'infrastructures

de Grille active basée sur des politiques [72] , l'équipe du Professeur Joan Serrat (UPC, Barcelone, Espagne) pour la surveillance de Grille avec la thèse de Edgar Magana [118, 117], l'équipe de Professeur Paul Roe (Queensland university of Technology, Brisbane, Australie) sur la flexibilité dans les grilles actives [101].

- **Les services d'adaptation** : en support des activités de recherche et développement de la PME Lyonnaise 3DDL ("3 Degrés de Liberté") et en lien avec nos activités dans le projet RNRT Temic, nous avons étudié un ensemble de services d'adaptation de contenus à l'intérieur du réseau. Nous supportons à la fois l'adaptation d'applications à la volée pour faire face à l'hétérogénéité d'équipements terminaux ainsi que de l'adaptation de flux multimédia pour faire face aux conditions réseaux variables.

3.1 Imaginer de nouveaux équipements flexibles pour l'Internet

Notre étude de la flexibilité dans les réseaux nous permet de proposer différents équipements (routeurs, répartiteurs de charges et pare-feux) capables d'être déployé dans des contextes réseaux particuliers et répondant à différents manques dans les infrastructures actuelles.

Dans la première partie, directement inspirée des travaux sur les réseaux actifs et l'environnement Tamanoir, nous proposons une adaptation de notre architecture à un contexte plus industriel afin de supporter différentes technologies et environnement réseaux. Ces travaux ont eu lieu dans le cadre du projet RNRT Temic en collaboration avec Jean-Patrick Gelas et les ingénieurs Martine Chaudier et Pierre Bozonnet.

La deuxième partie concerne la proposition de solutions logicielles pour des équipements réseau distribués ayant besoin de haute disponibilité. Cette recherche a été menée en collaboration avec Orange / France Télécom R & D lors du co-encadrement de la thèse CIFRE de Narjess Ayari (avec Denis Barbaron et Pascale Primet, 2005-2008).

La troisième partie est focalisée sur nos travaux dans le domaine de la haute disponibilité pour des pare-feu à états . Cette recherche a été menée dans le cadre d'une collaboration que j'ai établie avec l'Université de Séville (Espagne). J'ai participé activement aux activités scientifiques doctorales menées par Pablo Neira Ayuso que j'avais encadré pendant son stage de DEA.

3.1.1 Vers un équipement réseau autonome à contexte industriel

En collaboration avec différents partenaires académiques (LIFC, GRTC) et industriels (Société SWI), nous avons travaillé dans le cadre du projet RNRT Temic[33] à la conception [44] et au déploiement d'équipements réseaux programmables pour répondre à certains scénarios industriels de maintenance et de surveillance multi-capteurs. Ces scénarios mettaient plusieurs contraintes sur les équipements: facilité de déploiement, auto-configuration, robustesse en milieu industriel et facilité de démontage à la fin des expériences.

Un des scénarios envisagé concerne la surveillance de sites industriels à grande échelle sans connection informatique. Un site distant de pompage d'eau doit être surveillé.(Figure 3.2) Un nœud réseau autonome, déployé sur le site, embarque un ensemble de services réseaux configurables et personnalisables afin d'assurer la surveillance, du site, la remontée d'alertes, la collecte d'informations statistiques sur les équipements industriels...etc.. Ce nœud réseau plongé dans un contexte industriel doit être auto configurable, personnalisable à distance, robuste et résistant à l'environnement.

Nous avons proposé une adaptation de l'environnement générique Tamanoir[77] afin qu'il soit déployable sur un équipement à ressources limitées. Cette implémentation a été réalisée sur des solutions légères de la société Bearstech [17]. Afin de proposer une solution facilement transportable et robuste nous avons utilisé un un petit boitier aluminium avec carte mère intégrée

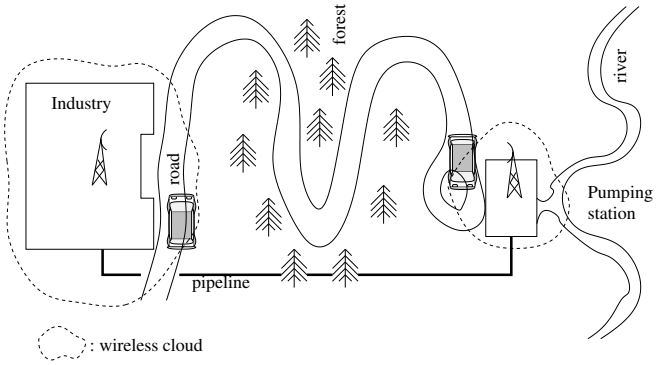


Figure 3.2: Déploiement sur le terrain d'un équipement réseau autonome

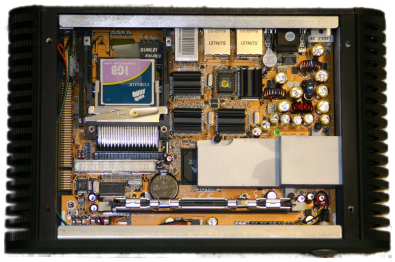


Figure 3.3: Vue interne de l'équipement IAN2

(VIA C3 CPU 1GHz, 256MB RAM, 3 ports Giga Ethernet). Ce boitier ne contient aucune pièce mécanique en mouvement, (pas de ventilateur, ni de disque dur mécanique). La figure 3.3 présente une vue interne du boitier avec son système de refroidissement passif. Le système d'exploitation, le système de fichiers et les environnements d'exécution sont stockés sur une carte mémoire.

Ainsi nous avons conçu l'architecture IAN2 (**I**ndustrial **A**utonomic **N**etwork **N**ode) qui supporte des fonctions réseaux de routage et de transmission à l'aide d'interfaces réseaux filaires et sans fils. Les capacités limités en termes de ressources (CPU, mémoire, stockage) sont mises à disposition des services autonomes embarqués.

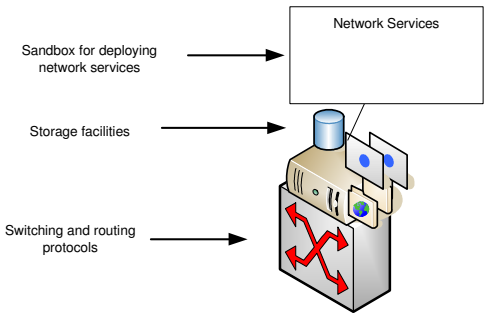


Figure 3.4: IAN2 : Equipment Réseau Autonomique Industriel

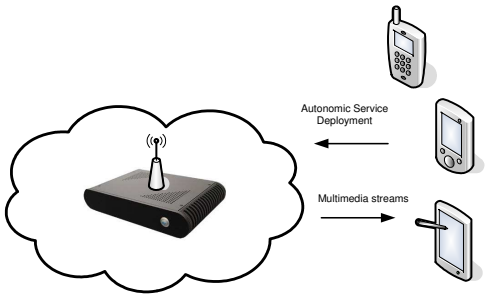


Figure 3.5: Déploiement de services autonomes à partir de nœuds mobiles

L'architecture du nœud IAN2 repose sur environnement d'exécution $Tamanoir^{embedded}$. Les contraintes matérielles (faibles ressources) et industrielles (robustesse, sécurité) nous ont amené à simplifier les couches architecturales et logicielles de l'environnement. Les classes logicielles inutilisées et certaines fonctionnalités (ressources distribuées, NIC) ont été supprimées afin de réduire l'empreinte de l'environnement et de faciliter sa maintenance et sa portabilité (Figure 3.4).

Nous avons aussi ajouté la possibilité du déploiement de services à partir de terminaux mobiles (PDA, smatphones) qui transportent les services et échangent des données avec les nœuds IAN2. Pendant cette étape, les nœuds mobiles se comportent comme des *repositories* poussant les nouvelles fonctionnalités vers les nœuds autonomes (Figure 3.5).

3.1.2 Haute disponibilité dans les serveurs distribués

J'ai co-encadré la thèse CIFRE de Narjess Ayari avec l'entreprise Orange France Télécom R&D (avec Denis Barbaron et Pascale Primet, 2005-2008). Cette thèse a été l'occasion de bénéficier

du point de vue en interne de l'opérateur télécom et d'intégrer ses besoins et ses attentes par rapport à la flexibilité dans les infrastructures distribuées à grande échelle. Ces travaux ont porté sur la haute disponibilité dans les serveurs multi-machines afin d'améliorer les profits des opérateurs. Ces travaux ont notamment fait l'objet d'un brevet avec France Télécom [6].

Les serveurs de services distribués sont constitués de grappes de machines qui répondent aux requêtes des clients transportées par différents flux de communications lors de sessions. Ces serveurs peuvent être des points de pannes uniques potentiels (*Single Point of Failure*). Ils sont cruciaux pour les opérateurs réseaux afin de fournir un ensemble de services à leurs clients (serveurs vocaux, transferts de données)

Avec Narjess Ayari, nous avons travaillé sur la proposition logicielle d'une solution de tolérance aux pannes et de haute disponibilité afin de garantir l'intégrité des sessions (voix, transfert) lors de pannes d'équipement chez les opérateurs réseaux (Figure 3.6). Le but est de maximiser le profit de l'opérateur en limitant les impacts des pannes sur les sessions en cours [7] Une architecture d'équipement assurant la réplication active des équipements a été proposée. Ces équipements se trouvent sur le chemin des données et garantissent une reprise sur erreur transparente pour les clients sans compromission pour les autres sessions en cours dans l'équipement. Ces travaux ont donné lieu à diverses validations dans les contextes de services d'un opérateur réseau [4, 8, 5].

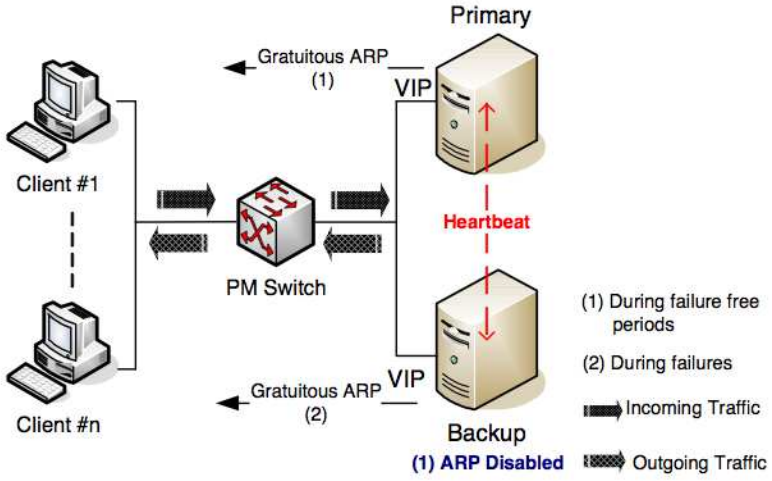


Figure 3.6: Réplication active pour haute disponibilité (service de voix)

3.1.3 FT-FW : un pare-feu à états tolérant aux pannes

Les pare-feu sont un élément clé de l'infrastructure Internet afin de protéger les utilisateurs et les services réseau contre les attaquants. Un pare-feu (logiciel ou matériel) sépare plusieurs segments de réseau et applique une politique de filtrage. Cette politique de filtrage détermine les paquets de données autorisés à entrer et à quitter un segment de réseau donné. Il existe deux grandes familles de pare feu : les pare-feu sans état (*stateless firewall*) qui examinent chaque paquet de données le traversant et le soumet à une liste de règles Les pare-feu à états (*stateful firewall*) vérifient en plus le bien fondé de l'arrivée des paquets par rapport à une connexion en cours (par exemple TCP : figure 3.7).

Du point de vue de tolérance aux pannes , les pare-feu introduisent un point de défaillance unique dans le schéma du réseau.

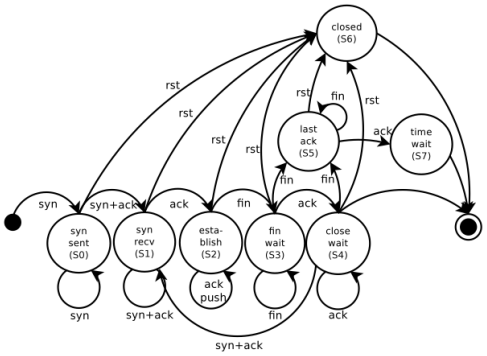


Figure 3.7: Etats d'une connexion TCP

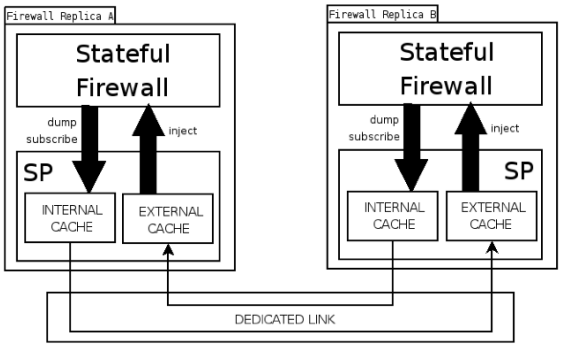


Figure 3.8: Architecture FT-FW

Suite au stage de DEA de Pablo Neira Ayuso que j'avais encadré, j'ai établi une collaboration avec l'Université de Séville où Pablo effectuait sa thèse dans l'équipe de Rafael Gasca. Dans la lignée de nos travaux sur la haute disponibilité des serveurs distribués (section 3.1.2), nous avons étudié la proposition d'un équipement flexible de type "pare-feu à état prenant en compte des caractéristiques de haute disponibilité : l'architecture FT-FW (*Fault Tolerant Stateful Fire-Wall*) qui combine des infrastructures matérielles et logicielles (figure 3.8) [127, 10].

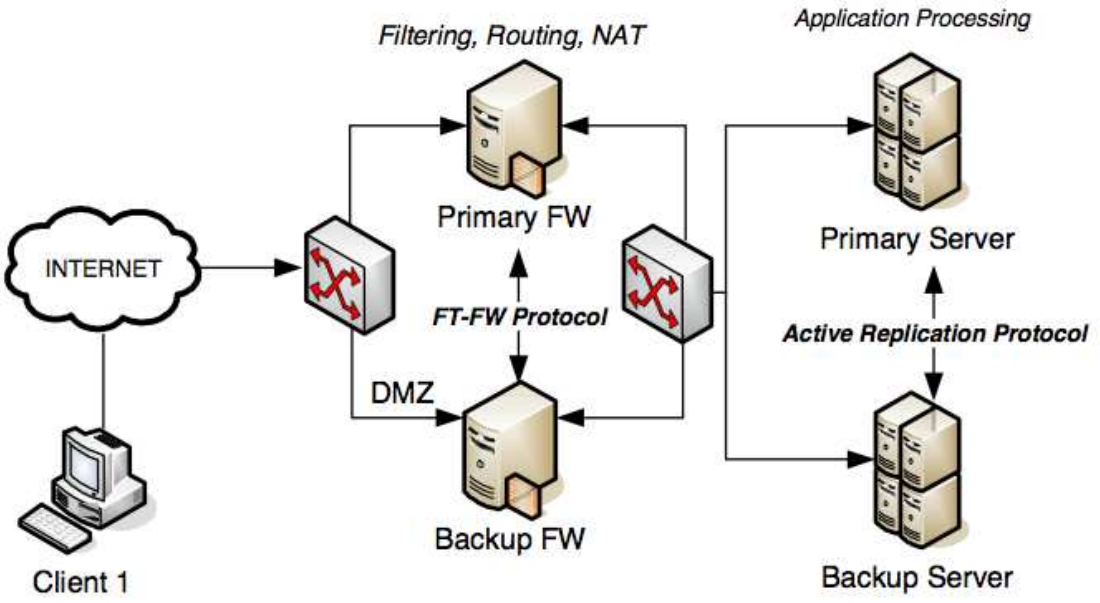


Figure 3.9: Architecture de haute disponibilité sécurisée

Nos propositions sur la réplication active ont été combinés avec l'architecture de pare-feu à état hautement disponible afin de fournir une infrastructure sécurisée tolérante aux pannes (figure 3.9) [9].

Nous avons proposé la librairie SNE (*Stateful Network Equipment*) qui permet de construire des équipements à état. Cette librairie contient des ajouts au noyau linux ainsi qu'un démon logiciel pour sa configuration [128]. Ces travaux ont été poursuivis à l'Université de Séville avec

notamment l'implémentation de l'environnement de gestion des tables de connexion *conntrack*. Plusieurs sociétés de pare-feu tels que Vyatta, 6WIND, EdenWall, Astaro AG et la solution open source Firewall Builder 4 ont utilisé l'implémentation de l'architecture FT-FW dans différentes solutions commerciales.

3.2 XCP-i : Un protocole de transport interopérable et extensible

Dans la thèse de Dino Lopez Pacheco que j'ai co-encadrée avec Cong-Duc Pham (2005-2008), nous nous sommes intéressés à la conception d'un protocole de transport inter-opérable avec les infrastructures distribuées à grande échelle et qui bénéficie de l'intelligence du réseau pour prendre des décisions d'utilisation des ressources réseaux robustes, équitables et efficaces [114, 113, 115].

3.2.1 Les protocoles "Explicit Rate Notfication" à assistance de routeurs

TCP (*Transport Control Protocol*) est le protocole de transport phare de l'Internet. Il existe depuis plus de 40 ans et permet d'échanger des données fiables entre terminaux hétérogènes. TCP est un véritable protocole de transport de bout en bout; la complexité du protocole est déployé sur les machines terminales impliquées dans l'échange de données. De part son aspect générique, TCP est confronté à de nombreuses limitations notamment en termes de performances.Malgré certaines optimisations et réglages; TCP peut être pénalisé par la lenteur de la phase d'évitement de congestion où la fenêtre de congestion n'augmente que d'un paquet par aller-retour (figure 3.11 (a)).

Différents travaux montrent que le simple fait de renvoyer une information condensée sur l'état du trafic est suffisant pour améliorer les performances des protocoles de transport de données [164, 175]. Par exemple, la notification explicite de congestion (ECN « Explicit Congestion Notification » [147, 34]) est une extension aux protocoles TCP et IP qui permet aux routeurs de signaler une congestion dans le réseau avant qu'une perte de paquets ne se produise. Un ensemble de solutions appelées protocoles ERN (*Explicit Rate Notification*) reposent sur le fait que les équipements réseaux puissent être capables d'envoyer une information sur l'état du trafic les traversant (par exemple : MAXNET [174, 120], JetMax [176], XCP [94]).

3.2.1.1 Le protocole XCP : un protocole ERN

XCP (*eXplicit Control Protocol*) proposé par Dina Katabi [94] est un protocole qui utilise l'assistance des routeurs pour informer l'émetteur des conditions de congestion du réseau. Ainsi l'émetteur de données peut déterminer la taille optimale de sa fenêtre de congestion et maximiser de cette façon l'utilisation des liens et le niveau d'équité.

XCP repose sur une coopération entre machines terminales (émetteur et récepteur) et routeurs XCP :

- L'émetteur : émet des paquets de données qui contiennent un en-tête rempli avec la taille de la fenêtre de congestion, l'estimation du temps aller-retour et une valeur appelée $feedback$ qui indique à l'émetteur un incrément (si elle est positive) ou un décrément (si elle est négative) à appliquer à sa fenêtre de congestion. Le champ $feedback$ est le seul qui peut être modifié par les routeurs XCP en fonction des valeur des deux autres champs. À la réception de l'accusé de réception, l'émetteur met à jour la taille de sa fenêtre de congestion en tenant compte du $feedback$ calculé par les routeurs XCP.

- Le récepteur : renvoie à l'émetteur les données contenues dans l'en-tête lors de la phase d'acquittement des paquets.

- Les routeurs XCP : utilisent un contrôleur d'efficacité (qui maximise l'utilisation de la bande passante sans perte de paquets) et un contrôleur d'équité (qui se base sur la différence entre trafic entrant et capacité du lien de sortie). Pour une description complète du mécanisme XCP, le lecteur pourra se reporter à [94].

  XCP utilise des routeurs spécialisés qui permettent de signaler et d'informer de manière très précise sur l'état de congestion dans le réseau permettant ainsi à la source de déterminer la taille optimale à donner à la fenêtre de congestion. De cette manière, XCP présente une très bonne stabilité tout en récupérant très rapidement le débit disponible maximisant ainsi l'utilisation des liens haut-débit. De plus, le contrôle de XCP instaure une équité forte entre les flux XCP.

  Dans une deuxième étape, le contrôleur d'équité traduit la valeur $feedback$ (qui peut être assimilée à une valeur agrégée positive ou négative), en une valeur $feedback$ par paquet (qui sera ensuite placée dans l'en-tête du paquet de données) en suivant des règles d'équité similaires aux règles de TCP (processus AIMD).

Il faut noter qu'il n'y a pas d'états par flux conservés par le routeur XCP pour exécuter toutes ces opérations. En effet, comme les paquets de données d'un flux donné portent dans leur en-tête la valeur actuelle de la fenêtre de congestion et le RTT, il est possible de calculer pour chaque flux le nombre de paquets envoyés par fenêtre de congestion afin d'assigner la bande passante disponible de manière proportionnelle.

Dans la figure 3.10, on observe des flux XCP qui traversent un goulot d'étranglement à 45 Mbps [94]. Le flux 1 s'approprie très rapidement la totalité de la bande passante disponible. A l'arrivée des flux concurrents, la bande passante est partagée de manière équitable entre tous les flux. C'est un donc un exemple parfait de réactivité et d'équité où la flexibilité présente dans le réseau permet une utilisation optimale des ressources.

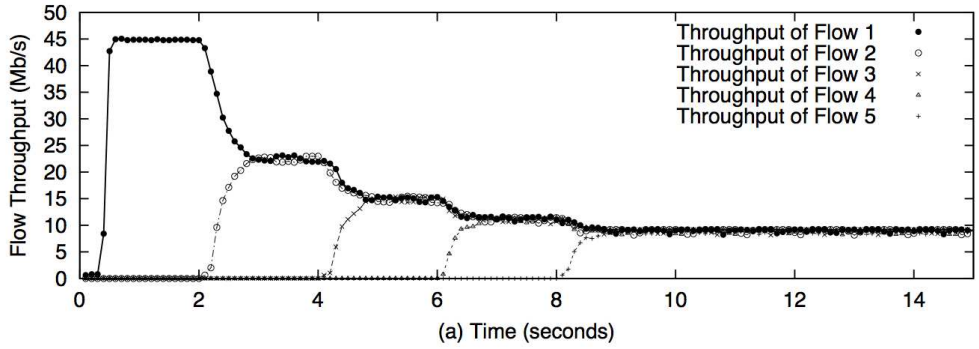


Figure 3.10: Performance et équité entre les flux XCP (extrait de [94])

XCP semble être une solution très prometteuse sur les réseaux haut-débit et plusieurs études ont montré analytiquement ses performances [116], ont proposé des améliorations pour le rendre plus robuste face aux pertes sur le chemin de retour [112] ou encore ont mené des études expérimentales sur des implémentations réelles [178].

Pourquoi ne trouve-t'on pas des protocoles ERN partout dans l'Internet ?

3.2.1.2 Limites de XCP

La plupart des travaux disponibles dans la littérature sont uniquement validés "en laboratoire": les simulations et implémentations reposent sur des infrastructures matérielles exclusivement composées d'équipements XCP et les données sont transportées dans des flux XCP. Ces études

mentionnent le manque d'interopérabilité lors de la présence de routeurs non-XCP entre la source et le récepteur mais ne proposent pas de solution !

Par exemple, lors d'une simulation sur *ns*, un flux TCP New Reno sur le scénario (a) obtient la bande passante au bout de 8 secondes. Alors que le même scénario basé sur des routeurs XCP et des flux XCP permet aux flux d'obtenir en quelques *ms* la bande passante maximale et le débit reste au maximum pendant toute la durée de l'expérience. Par contre, lors d'un scénario où des routeurs non-xcp se trouvent sur le chemin des flux XCP, on observe une grande instabilité en termes de débit (Figure 3.12 - scénario (c)). Pendant la simulation d'une minute, sur le scénario (a) TCP a envoyé 215 Mo, XCP sur le scénario (b) a envoyé 223 Mo, et XCP sur le scénario (c) a seulement envoyé 52 Mo !

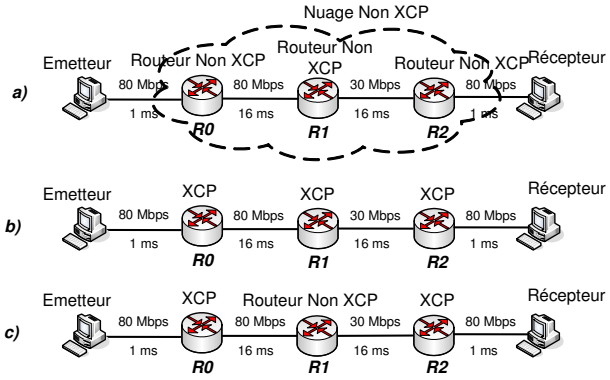


Figure 3.11: (a) scénario pour TCP, (b) et (c) scénarios pour XCP.

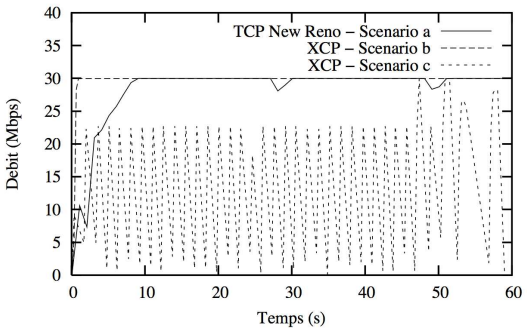


Figure 3.12: Évolution du débit pour les scénarios a,b et c.

Cette forte dépendance de XCP envers des routeurs spécialisés limite considérablement l'intérêt de XCP !

Peut-on sortir le protocole XCP des laboratoires afin de le confronter à la réalité des échanges réseaux sur l'Internet ?

3.2.2 Architecture (simplifiée) de XCP-i

Dans [94], les auteurs proposent un modèle de déploiement incrémental basé sur la création de nuages XCP avec des routeurs de bordure qui traduisent les flux non-XCP vers XCP. Cependant cette idée très complexe n'a pas été développée ni validée.

Nous avons donc proposé une extension de XCP, appelée XCP-i (XCP inter-opérable), qui permet de déployer des infrastructures XCP autour de matériels non-XCP et en concurrence avec des flux TCP. Plutôt que de proposer brutalement une nouvelle solution architecturale qui aurait un impact limité sur la communauté réseau, nous avons choisi une approche incrémentale:

- respecter la philosophie XCP en gardant les mêmes mécanismes d'estimation de bande passante et de notification. Les nouvelles fonctionnalités apportées par XCP-i n'augmentent que légèrement la complexité du protocole XCP.

- autoriser un déploiement incrémental de la solution XCP-i petit à petit dans des réseaux existants (non XCP (figure 3.11 (c)) et en concurrence avec les flux et protocoles utilisés classiquement dans l'Internet (TCP, UDP... figure 3.18).

Le protocole XCP-i détecte sur le chemin du réseau chaque ensemble (nuage) d'équipements non XCP (grâce à un jeu de compteurs TTL) et localise l'équipement XCP-i le plus proche de ce nuage. Chaque nuage détecté est donc "remplacé" par un router XCP-i virtuel instantié sur le routeur le plus proche du nuage (Figure 3.13). Chaque routeur virtuel a comme fonctionnalité

d'estimer la bande passante disponible sur le chemin dans le nuage par des mécanismes légers d'estimation. Les routeurs XCP-iv renvoient cette information de manière transparente aux clients XCP.

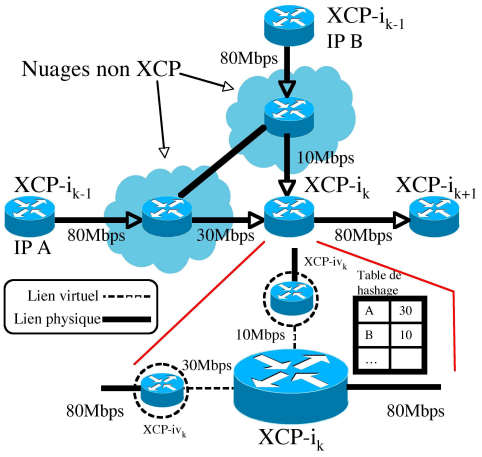


Figure 3.13: Routeur XCP-i avec un routeur virtuel par nuage non-XCP.

3.2.3 Interopérabilité avec des équipements non XCP

Notre modèle de XCP-i a été développé sur une extension du modèle *ns* de XCP de Katabi. Nous supposons que l'estimation de bande passante disponible renvoie la valeur correcte à la fin de chaque intervalle de contrôle XCP. Nous avons aussi étudié les cas où l'estimation de bande passante peut être erronée, cette étude ne sera pas mentionnée dans ce document (voir [115]). Nous considérons aussi les cas où les machines terminales sont directement reliées à des routeurs non XCP, dans ce cas les fonctionnalités de XCP-peuvent être déployées dans la carte d'interface réseau des machines terminales en respectant le développement que nous avons proposé avec KNET (section 2.4).

Nous avons validé l'approche incrémentale de XCP-i dans différents scénarios. Deux d'entre eux sont décrits ici à titre illustratif :

- **Déploiement incrémental autour de nuages non-XCP** : le premier scénario sur lequel XCP-i a été expérimenté concerne un déploiement symétrique dans des points de *peering* du réseau (figure 3.14) où deux nuages non-XCP sont connectés par un routeur XCP-i. Les résultats de simulation montrent que nous n'observons aucune perte de paquets. Les routeurs XCP-i virtuels dans R1 et R2 estiment la bande passante disponible dans le nuage non-XCP et calcule ainsi la valeur de $feedback$ optimale. Ces résultats montrent que XCP-i est capable de supporter efficacement des flux hautes performances dans des réseaux hétérogènes même si son déploiement se limite à quelques endroits stratégiques du réseau.

- **Fusion autour de plusieurs nuages non-XCP** : le scénario de fusion repose sur une topologie où un équipement XCP-i est confronté à deux nuages non-XCP (Figure 3.16). Le routeur XCP-i R1 doit générer un routeur virtuel XCP-iv pour chaque lien connecté à un nuage non-XCP. Nous validons ainsi la capacité du protocole XCP à garantir l'équité entre 2 flux agrégés. La figure 3.17 démontre que XCP-i réussit à maintenir une équité des flux avec les émetteurs i et j qui obtiennent respectivement 280Mbits/s et 100Mbits/s.

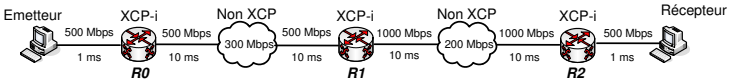


Figure 3.14: Déploiement incrémental sur des points de *peering*

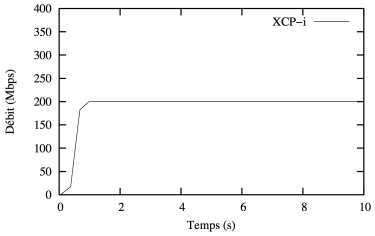


Figure 3.15: Débit observé

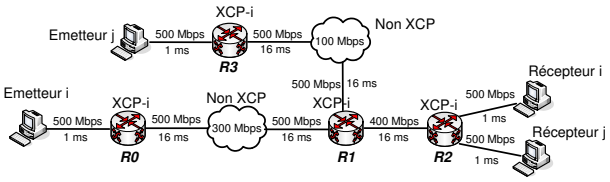


Figure 3.16: 2 files d'attente non-XCP partageant un chemin XCP

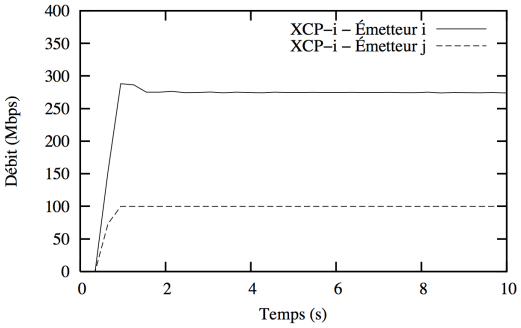


Figure 3.17: Débit dans le scénario de Fusion

3.2.4 Équité entre flux

Nous avons complété notre intégration incrémentale de XCP en étudiant et en proposant des solutions pour assurer l'interopérabilité des flux XCP face aux flux TCP.

Dans la figure 3.18, nous observons une simulation *ns* d'un flux XCP qui utilise la bande passante disponible (1 Gbits). A la seconde 10, deux flux TCP traversent le réseau et se mettent en compétition pour partager la bande passante. Le flux XCP ne participe pas à cette compétition et ne peut bénéficier que d'une bande passante très faible.

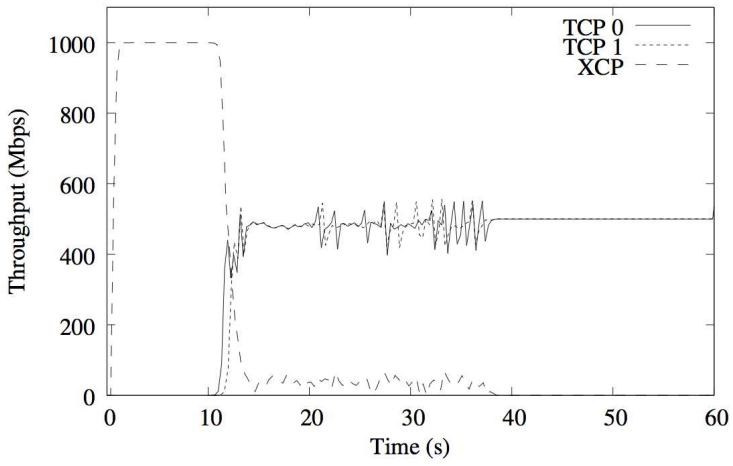


Figure 3.18: Absence d'équité entre un flux XCP et deux flux TCP

Nous avons ainsi mené des études sur l'équité entre flux XCP et TCP dans divers scénarios [113, 114] et proposé une solution d'équité entre des flux ERN (type XCP) et des flux E2E (type TCP).

Nous proposons un mécanisme d'équité qui déploie deux fonctionnalités principales :

1. Estimation des ressources nécessaires aux flux XCP : nous souhaitons connaitre le nombre de flux actifs XCP et TCP traversant un routeur XCP. Cette approche peut être coûteuse en temps de calcul, nous avons choisi une approche où le routeur XCP estime le nombre de flux actifs en utilisant des Bloom Filters [25], en adaptant une partie des mécanismes proposés dans SRED (*Stabilized Random Early Detection*[142]) qui évite les congestions réseaux. Partant de cette estimation, le routeur peut calculer les ressources nécessaires à chaque flux en prenant en compte la capacité du lien de sortie (information déjà connue par le routeur).

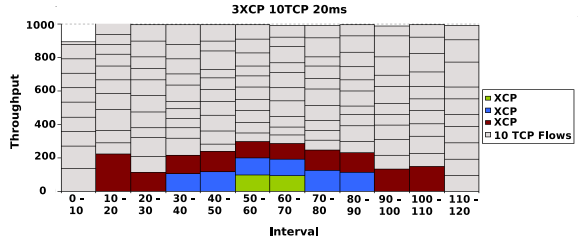
2. Limitation de la consommation de ressources des protocoles de bout en bout : en cas de confrontation, les flux XCP se contentent de la bande passante restante non utilisée par les flux TCP. Notre approche est donc de limiter l'usage de ressources des ces deniers afin de garantir une certaine équité. Nous exécutons donc un service intelligent de destruction de paquets sur les flux TCP quand la capacité du lien est utilisée au dessus d'un certain seuil. Une approche simple, avec peu d'état et peu coûteuse en CPU est proposée :

   (a) si les ressources prises par les flux TCP dépassent notre estimation $->$ destruction des paquets avec une faible probabilité;

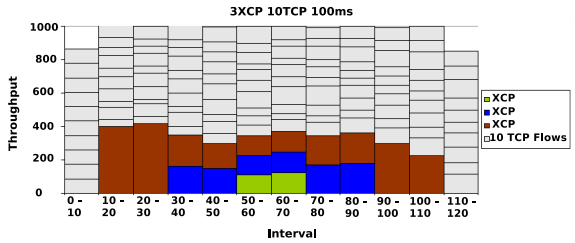
   (b) si les flux TCP ne réduisent par leur débit de transmission $->$ augmentation de la probabilité de destruction;

   (c) si les flux TCP ont moins de ressources qu'estimées $->$ réduction de la probabilité de destruction de paquets;

Nous évaluons l'impact de notre mécanisme d'équité : des flux XCP au milieu de flux TCP (figure 3.19) et vice versa (figure 3.20). On observe que notre mécanisme arrive à garantir de la bande passante aux flux XCP. Notre mécanisme est quelque fois un peu trop agressif envers les flux TCP, mais ceux ci récupèrent rapidement la bande passante disponible et contraignent les flux XCP à respecter l'équilibre.

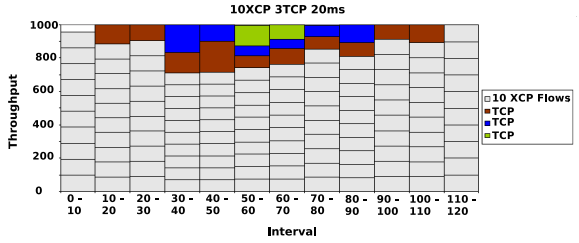


(a) 20 ms RTT

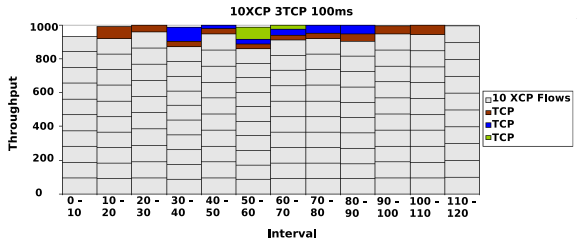


(b) 100 ms RTT

Figure 3.19: 3 flux XCP apparaissent au milieu de 10 flux TCP



(a) 20 ms RTT



(b) 100 ms RTT

Figure 3.20: 3 flux TCP apparaissent au milieu de 10 flux XCP

3.3 Supporter des calculs distribués à grande échelle : La Grille Active

Par définition, la Grille de calcul (*Grid Computing* [70]) est une infrastructure de ressources géographiquement distribuées qui peut être utilisée de manière structurée afin de résoudre un problème donné. Au début des années 2000, le concept de Grille s'est imposé comme le modèle de référence des infrastructures distribuées à grande échelle. De nombreux intergiciels de Grille ont été développés et mis à disposition de la communauté scientifique dont les plus connus sont Globus[69], Condor[161] ou Legion[85]. La Grille promettait de révolutionner la manière dont nous utilisions les ressources de calcul pour de nouvelles classes d'applications. Ce fut le cas. La Grille ne sera détrônée qu'avec l'arrivée du Cloud et de ses possibilités de facturation, déploiement dynamique et virtualisation. Malgré les difficultés de complexité des environnements, déploiement, configuration, sécurité, gestion de l'hétérogénéité, problèmes des communications, de nombreux chercheurs ont utilisé le modèle de Grille comme scénario d'infrastructure distribuée à grande échelle. C'est mon cas, mais j'ai abordé les problèmes de la grille par un angle différent.

3.3.1 Proposition d'une architecture de Grille Active

Une Grille Active [98, 81] est une infrastructure de composants matériels et logiciels dynamiquement programmables et re-configurables (*AAC Applications Aware Components*): répartie sur les réseaux de la Grille afin de supporter un ensemble de services dynamiques adaptés aux besoins des applications et des gestionnaires de Grille. Afin de respecter l'urbanisation proposée en section 2.5 et de garantir des performances aux infrastructures, les nœuds AAC ne sont pas disposés dans les réseaux de cœur (*backbone*) mais sont cantonnés en périphérie (réseaux d'accès, réseaux de centres de données). Ils sont ainsi gérables et configurables par les gestionnaires de Grille. Les nœuds AAC prennent en charge les communications et les services pour un ensemble de nœuds de la Grille.

Pour supporter la plupart des configurations d'applications distribuées, l'architecture de Grile Active est adaptable aux deux principales configurations de Grilles de calcul (figure 3.21) :

- Support de calcul multi-grappes (*meta-computing*): des passerelles réseaux programmables sont disposées autour de grappes de calcul sur différents sites. Les nœuds AAC peuvent traiter tous les flux de données entrant ou sortant d'une grappe de machines. Les nœuds communiquent entre eux et peuvent échanger des services.
- Calcul global et Pair-à-Pair (figure 3.21) : dans cette configuration, les nœuds AAC sont associés à une ou plusieurs ressources de calcul de la Grille. Des topologies hiérarchiques de nœuds AAC sont ainsi déployées aux points d'hétérogénéité du réseau.

Dans ces deux configurations, les nœuds AAC gèrent les opérations et les flux concernant les ressources de calcul et peuvent ainsi supporter des services de collecte de résultats, caches, synchronisation de nœuds, sauvegarde de points de reprise...

Notre proposition a été d'implémenter le concept de Grille Active grâce à la technologie des Réseaux Actifs et de le valider avec la technologie Tamanoir que nous avons développé (section 2.3). Suite à cette proposition, différents services de Grille Active ont pu être étudiés et validés dans l'équipe RESO : la qualité de services active dans les grilles (thèse de B. Gaidioz [71] encadrée par Pascale Primet), le *multicast* fiable (thèse de Moufida Maimour[119] encadrée par Cong-Duc Pham) et le transport de données hautes performances (thèse de Jean-Patrick Gelas[76] que j'ai co-encadrée).

3.3.2 Flexibilité avec des Services Web

Lors d'un séjour longue durée au sein de la Queensland University of Technology (Brisbane, Australie), j'ai établi une collaboration avec l'équipe du professeur Paul Roe (Programme d'Action

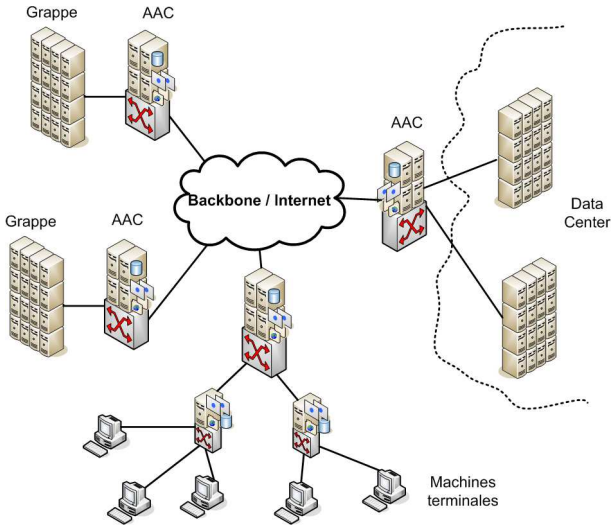


Figure 3.21: Architecture de Grille active supportant des infrastructures multi-grappe et du calcul P2P

Intégrée FAST). Cette collaboration avait pour but d'explorer l'ajout de flexibilité au sein de Grille active grâce aux solutions de type WebServices (basées sur sur Open Grid Service Infrastructure, en utilisant WSRF (*Web Service Resource Framework*) [63]. Ces travaux ont été menés avec Chien-Jon Soon dont j'ai co-encadré le stage avec Paul Roe [101].

Nous avons proposé l'architecture WeSPNI (*Web Services based on Programmable Networks Infrastructure*) qui est dérivée de l'architecture de Grille Active tout en intégrant les propriétés et facilités des services web dans une infrastructure flexible [159].

3.3.2.1 Des services réseaux flexibles exposés

Nous ajoutons aux services réseaux programmables présentés en section 2.3.2, la composition de services réseaux grâce à l'appel de services à partir des services réseaux programmables. Ainsi dans WeSPNI, les services contiennent des opérations de code local combinées à des invocations de services web (figure 3.22).

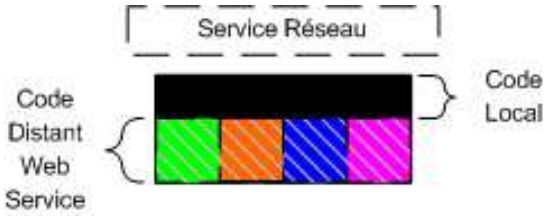


Figure 3.22: Un service réseau

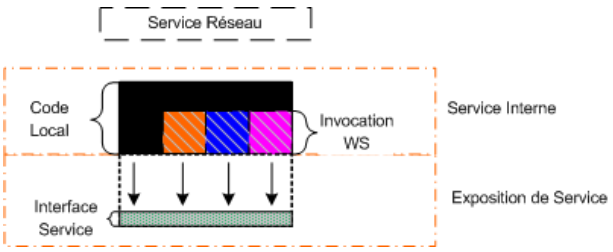


Figure 3.23: Exposition des interfaces de l'infrastructure réseau

L'approche WeSPNI permet aussi de mettre facilement à disposition les services déployés dans le réseau afin que d'autres applications et utilisateurs puissent en bénéficier. Les Services Web sont des composants logiciels qui se conforment aux standards et peuvent ainsi être utilisés de manière externe.

L'exposition des capacités du service (figure 3.23) est obtenue en utilisant une description de l'interface du service (Web Services Description Language (WSDL)) et d'un schéma XML

référencé par la description.

Nous proposons une exposition simplifiée basée sur le triplé { *input, processing, output* } :

- *Input* : les nœuds actifs travaillent au niveau paquet de donnée. Les paramètres d'entrée peuvent donc se retrouver au niveau de l'entête ou de la charge utile du paquet. Ces paramètres contiennent des références vers les services qui doivent etre déployés sur les nœuds ainsi quelques variables d'états. *input* = { *packet header* || *packet payload* }
- *Processing* : contient la description des fonctionnalités appliquées par le nœud programmable. Elle est fondée sur une exécution en série d'un ensemble d'opérations locales ou de fonctions invoquées par les services Web et disponibles sur des ressources distantes (Figure 3.23). *processing* = { $service_1$ & $service_i$ & $service_n$ }
- *Output* : renseigne le niveau de modification effectuée par le service réseau : destruction d'un paquet de données, duplication, diffusion, modification de l'entête ou de la charge utile. *output* = { *modified header* || *modified payload* }

3.3.2.2 Flexibilité du plan de contrôle

Avec l'apparition du *WS-Management* [62], les Services Web permettent une gestion flexible de serveurs et d'équipements distants. Les passerelles et équipements réseaux programmables peuvent ainsi être facilement gérées et surveillées.

Dans l'infrastructure WeSPNI, les services sont gérés par l'intermédiaire de services web. Ainsi, l'utilisateur peut contrôler leur déploiement, leur disponibilité, leur configuration et leur surveillance à partir de n'importe quel type de plate-forme.

Cette infrastructure distribuée est contrôlée pour les opérations de déploiement des différents services avec l'environnement Pangolin (voir section 2.6)(3.24)

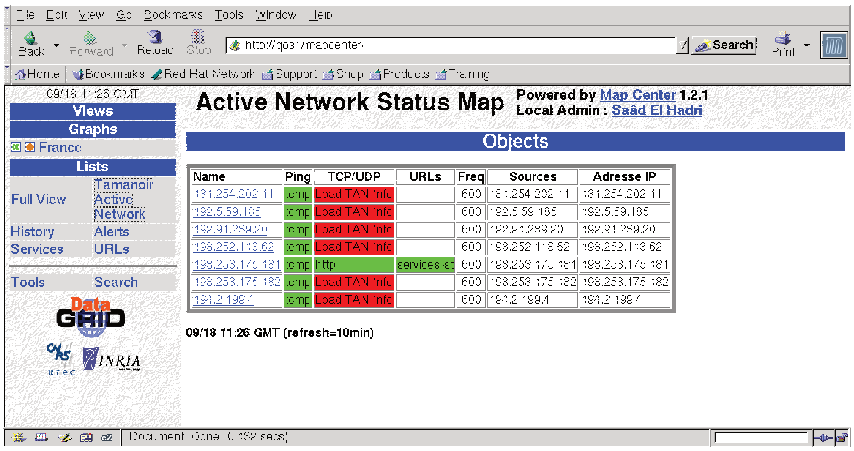


Figure 3.24: Ensemble de services contrôlés avec l'environnement Pangolin

3.3.2.3 Flexibilité du plan de données

L'architecture WeSPNI propose différentes fonctionnalités programmables liées au plan de données:

- Déploiement de fonctionnalités actives à l'intérieur du réseau : les Sevices Web génèrent un coût d'usage non négligeable et consomment des ressources (CPU, mémoire, stockage). Nous proposons d'orchestrer et de déployer certains services Web dans les équipements réseaux programmables (Figure 3.25). Ainsi un service réseau WeSPNI déployé sur une machine terminale (figure 3.25) peut profiter d'un ensemble de services web dynamiquement déployés sur des nœuds réseaux distants.

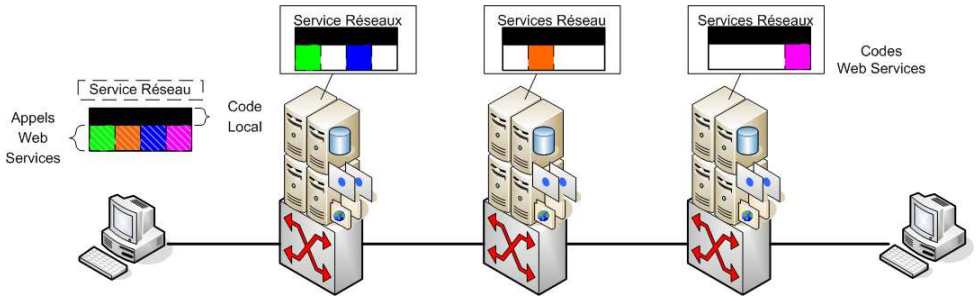


Figure 3.25: Ressources réseaux embarquant des Services Web

- Services flexibles à valeur ajoutée sur des infrastructures réseaux non programmables ou à ressources limitées : la disponibilité d'équipements programmables n'est pas toujours garantie. Avec la solution WeSPNI, nous avons étudié la possibilité de disposer de routage adaptatif dans le réseau afin d'autoriser des flux nécessitant certains services à traverser des serveurs distants, non présents sur le chemin de données (technique proche des *Content Delivery Networks*, mais basée sur les fonctionnalités et non les données) (figure 3.26). Ainsi les nœuds réseaux peuvent avoir des fonctionnalités de programmation limitées; les fonctions coûteuses en ressources demeurant sur les serveurs distants qui se chargent de les exécuter.

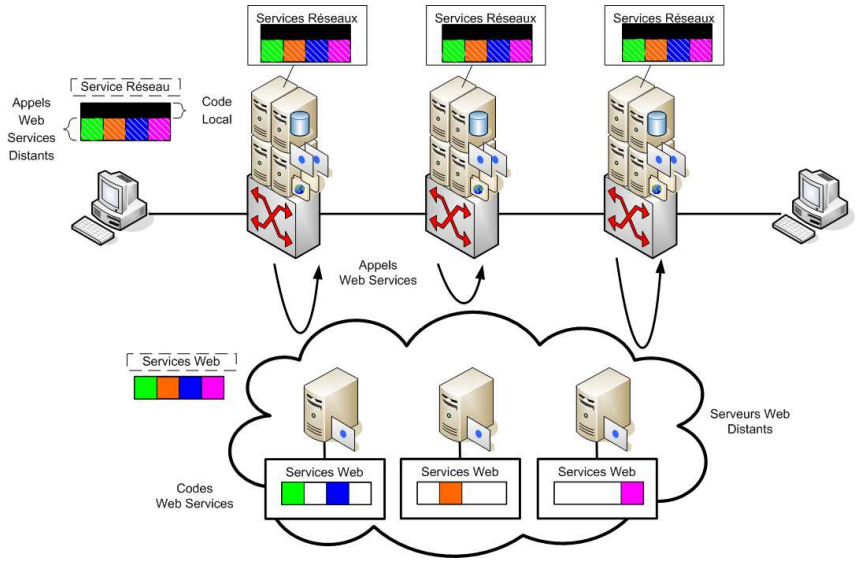


Figure 3.26: Le routage adaptatif envoie les flux de données vers des Services Web déployés sur des serveurs distants

Bien sur cette flexibilité extrême peut avoir un coût en termes de performance et invoquer des services distants peut aussi imposer des contraintes en termes de fiabilité et de disponibilité.

3.3.3 Une Grille active extrême

Afin de supporter l'exploration spatiale des planètes lointaines comme Mars; Vint Cerf, un des pionniers de la conception de l'Internet et de TCP/IP a proposé dès 1998 le concept d'Internet interplanétaire .

A cause de nombreuses contraintes dues au déploiement dans l'Espace, les protocoles de transport, le routage, le nommage des ressources doivent être modifiés afin de proposer plus de robustesse et de la tolérance face aux mauvaises conditions de transport de données. Les

protocoles de transport comme TCP/IP ne peuvent être utilisés à cause des trop grandes latences de communication.

En parallèle la communauté réseau travaille sur des architectures réseaux de type DTN (*Delay* ou *Disruption Tolerant Networks*) qui résistent aux longs délais ou aux coupures. Ainsi à un instant donné, un chemin de bout en bout n'existe peut-être pas entre 2 nœuds du réseau. Mais dans le futur une partie du réseau pourra être disponible pour faire transiter les données. Dans les réseaux tolérant aux coupures, les paquets de données sont encapsulés dans des *containers* logiciels (*bundle*) qui assurent une protection des données. Les bundles sont stockés au fur et à mesure de leur progression dans les différents équipements réseaux traversés (passerelles, routeurs, relais). Différentes infrastructures DTN ont été déployées : DakNet [145] (connectivité dans les pays en voie de développement), seaweb [148] (communications sous marines) ou DieselNet [53] (réseau DTN déployé sur des bus de transport d'un campus). Nous avons aussi utilisé la technologie DTN dans la gestion de réseaux de grande envergure dont des parties sont éteintes en vue d'économiser de l'énergie [162].

Différents travaux proposent de bénéficier de la flexibilité de la technologie DTN comme protocole de base des réseaux interplanétaires [36]. Le logiciel ION *(Interplanetary Overlay Network)* a été proposé comme implémentation de l'architecture DTN telle que décrite dans le RFC 4838[42].

Profitant de notre expérience en réseaux flexibles, nous avons proposé une adaptation de l'architecture *Interplanetary Internet* [36] sur une architecture de Grille active. Le scénario retenu concerne le déploiement d'un ensemble de ressources de calculs (type Grille ou Cloud) nécessaire à l'exploration et la conquête de planètes lointaines. Les nouveaux moyens de transport spatiaux sont capables de transporter des infrastructures de plus en plus volumineuses, et permettent donc d'embarquer des ressources de calcul et de stockage. Mais on peut penser que les moyens de calculs et de stockage les plus importants demeurent sur Terre pour des facilités de fourniture énergétique, de maintenance, de robustesse et de coût. (Fig. 3.27). Les communications entre plate-forme sur la planète distante et plate-forme sur Terre sont indispensables et nombreuses afin d'accéder à la disponibilité des ressources de calcul et de stockage.

Nous proposons une infrastructure distribuée à grande échelle flexible capable de supporter de très longs délais et des coupures réseaux tout en autorisant le déploiement dynamique de nouveaux services. Ce concept a recueilli un accueil favorable auprès de la communauté scientifique impliquée dans la recherche spatiale [105].

Comme dans une Grille Active, des passerelles de flexibilité (*Programmable Network Gateways - PNG*) assurent une utilisation transparente de la plate-forme malgré les déconnections et pertes de réseaux. Embarquant des technologies de DTN, elle supportent un ensemble de services dynamiquement déployés en fonction des besoins des applications. Par exemple, Si les nœuds récepteurs sont déconnectés, les passerelles sont autorisées à stocker ou rediriger les messages venant des émetteurs en attente de la reprise de la connexion avec les machines réceptrices. Cette démarche est transparente pour les émetteurs, les PNG s'occupent d'assurer la fiabilité de la transmission qui re-démarre dès la connection rétablie.

L'approche Grille interplanétaire a pu être validée en partie pendant une démonstration à la conférence SuperComputing 2007 lors d'une collaboration avec des chercheurs du *Space Robotic Laboratory* (Université de Tohoku, Japon). Les chercheurs japonais développent un robot explorateur autonome en forme d'araignée, capable de se déplacer par un mouvement de pattes ou de transformer ses pattes en roues afin de parcourir de plus longues distances (Robot LEON : *Lunar Exploration Omni directional Netbot* [150]). Pour les besoins de la démonstration, le robot novateur de type arachnéen était à la fois présent sous forme physique dans le laboratoire de Tohoku mais aussi comme un avatar logiciel exécuté dans une machine présente au Japon. Le contrôle du robot avait lieu à distance depuis le stand INRIA aux USA en utilisant une plate-forme logicielle incorporant certains composants logiciels de la Grille Inter-Planétaire.

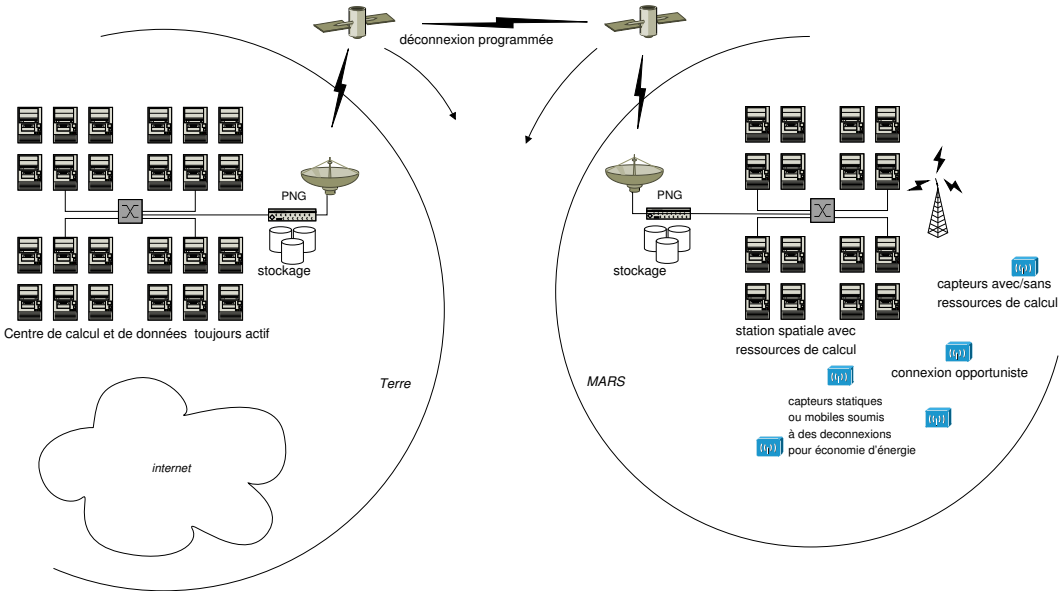


Figure 3.27: Grille inter-planétaire entre la Terre et Mars

Figure 3.28: Démonstration SC07 : Robot réel (localisé au Japon)

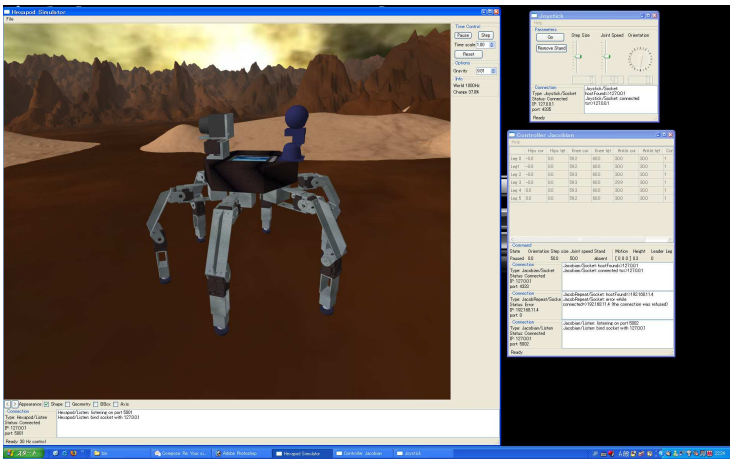


Figure 3.29: Démonstration SC07 : Robot avatar et interface de manipulation (localisé aux USA)

Notre approche nous semblait presque être de la science fiction. Mais une expérience similaire (sans déploiement flexible) a été récemment menée de manière concrète par la NASA et l'Agence Spatiale Européenne, en novembre 2012, lorsqu'un spationaute présent dans la station spatiale a contrôlé un robot déployé en Allemagne en utilisant la technologie DTN[2].

3.3.4 Ajouter du stockage intelligent dans le réseau

Nous avons vu dans le scénario de Grille Inter-Planétaire (section 3.3.3), que les équipements réseaux peuvent embarquer une zone de stockage afin de sauvegarder temporairement des données pour les retransmettre quand les conditions réseaux sont rétablies. Cette approche qui permet de disposer de zones de stockage dans le réseau ouvre la porte à différents services à valeur ajoutée. Cette section présente certains des services que nous avons étudiés et validés.

3.3.4.1 Réseaux logistiques

A la recherche d'une solution de stockage intelligente et légère, nous avons collaboré pendant plusieurs années avec l'équipe de Micah Beck (Université du Tennessee à Knoxville, USA) qui propose une solution de réseaux logistiques. J'ai invité le chercheur Alessandro Bassi à rejoindre l'équipe RESO et pendant deux ans j'ai exploré avec lui et Jean-Patrick Gelas le concept de réseau actif logistique dans des contextes de Grille. Après sa thèse, Jean-Patrick Gelas est parti en séjour post-doctoral pendant un an à l'Université du Tennessee pour continuer ses recherches sur le stockage logistique dans le réseau.

IBP[18] est une suite logicielle permettant le partage de ressources distribuées de stockage à travers un réseau. IBP expose ces ressources à l'aide d'un système de nomage qui peut être utilisé par toute application pour une période de temps donnée. La proposition *exNode* (*externalNode*), en se basant sur le concept de *inode*, agrège des grands volumes d'allocations de stockage sur l'Internet.

Nous avons intégré la suite logicielle Tamanoir avec l'environnement de stockage IBP afin de supporter le principe de Cache Actif Logistique (*Active Logistical Cache*) [15, 16]. Un cache actif logistique permet le support d'un sous-ensemble de services qui nécessitent d'accéder à des zones de stockage (figure 3.1).

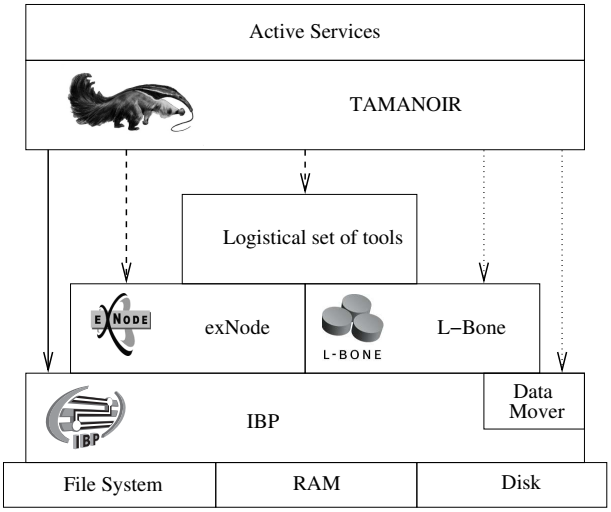


Figure 3.30: Tamanoir-IBP

[2]NASA, ESA Use Experimental Interplanetary Internet to Test Robot From International Space Station : `http://www.nasa.gov/home/hqnews/2012/nov/HQ_12-391_DTN.html`

3.3.4.2 Les caches revisités : Caches Web coopératifs et intelligents

Dans le cadre du stage de DEA de Sidali Guebli en collaboration avec le laboratoire LIRIS de l'INSA de Lyon (Jean-Marc Pierson), nous avons revisité la technologie des caches web en les améliorant à l'aide de technologie de réseaux flexibles. Ces caches actifs permettent ainsi la gestion des services dans une grille pervasive [107].

Nos travaux ont porté sur les *proxies* caches coopératifs qui sont déployés à l'intérieur du réseau. Un *proxy* cache répond à des requêtes et tente de les satisfaire en fournissant du contenu stocké localement plutôt que d'accéder à des serveurs de contenus distants. Cette approche tente d'améliorer les performances en termes de réactivité et d'efficacité pour éviter de traverser de longues distances réseaux. Pour obtenir de meilleures performances avec moins de ressources on fait appel à la coopération entre différents *proxies* cache qui échangent des données afin de répondre le plus favorablement possible aux requêtes des utilisateurs. Les *proxies* caches sont distribués de manière hiérarchique ou géographique dans l'infrastructure réseau. Afin de proposer une solution efficace en limitant le nombre d'échanges pour trouver un objet, nous avons choisi une architecture de caches à 2 niveaux. Notre approche a consisté à utiliser le concept de Grille active à base de passerelles pour étendre une fonctionnalité de cache sur l'ensemble des nœuds disposant de capacité de stockage. La flexibilité des réseaux actifs permet la création rapide d'un service de cache web figure. Les machines terminales grâce aux fonctionnalités de Tamanoir sont capables de dialoguer en direct sur les ports du nœud actif qui agit en tant que *proxy*[107].

3.4 Adaptation de contenus dans le réseau

3.4.1 Adaptation d'applications à la volée : supporter le déploiement de jeux java sur des plate-formes mobiles

Nous avons validé la proposition de flexibilité dans le réseau pour le support du déploiement de jeux sur terminaux hétérogènes. Ces travaux ont été menés avec Aweni Saroukou et Jean-Marc Pierson dans le cadre d'un collaboration avec la PME lyonnaise 3DDL (3 degrés de liberté). Cette société, spécialiste du développement de jeux pour plate-formes et terminaux mobiles était confrontée à des contraintes d'hétérogénéité en termes de parc de terminaux mobiles. Comment concevoir des jeux et applications mobiles adaptés aux spécificités techniques des terminaux : taille d'écran, nombre de couleurs, puissance, langage..etc... ? La solution technique choisie par cette société est le développement de jeux en Java.

Le format Mobile Information Device Profile (MIDP), conçu pour les téléphones mobiles correspond à un ensemble d'API JavaME qui définit la façon dont les applications se connectent à l'interface des terminaux mobiles. Les application appelées portables appelées MIDlets ressemblent à des *applets* ou *servlets*. Ces applications prennent en compte les spécificités des mobiles et contiennent une archive Java (JAR) et un descripteur de l'application Java (JAD). Les fournisseurs d'applications utilisent notamment le fichier JAD pour inclure des informations temporaires : logos, publicités... Les spécificités des interfaces imposent aux concepteurs d'applications de créer différents versions de la même application pour chaque modèle de téléphone.

Les fichiers JAD sont donc spécifiques à chaque application et à chaque terminal mobile. Lors du téléchargement d'un jeu, une zone de donnée est créée et les fichiers JAD et JAR y sont copiés (figure 3.31). C'est la manière la plus simple de garantir la relation entre l'utilisateur, son terminal mobile et la version de jeux demandée.

Pour éviter cette duplication de données et les transferts nécessaires à la gestion de l'application, nous avons donc proposé de modifier dynamiquement les *containers* JAD à la volée à l'intérieur du réseau entre le serveur de jeux et les terminaux légers [100, 108] . Cette flexibilité est ajoutée grâce au support d'une infrastructure de réseaux programmables (Tamanoir - section 2.3). Cette

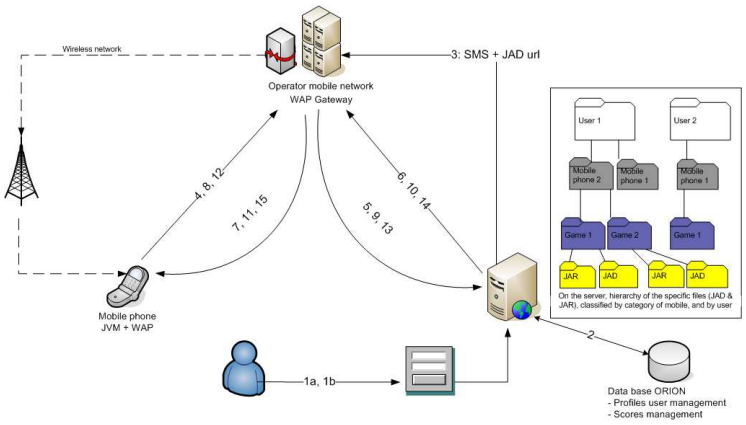


Figure 3.31: Déploiement de Jeux sans support de flexibilité

approche (figure 3.32) met en relation le terminal mobile avec une passerelle Tamanoir qui assure le téléchargement, l'adaptation à la volée (pour respecter les spécificités matérielle du terminal) et la personnalisation (logos, publicités..) de l'application. Cette adaptation est transparente pour l'utilisateur et l'opérateur réseaux, les données résultantes de cette adaptation ne sont pas stockées par le fournisseur d'applications.

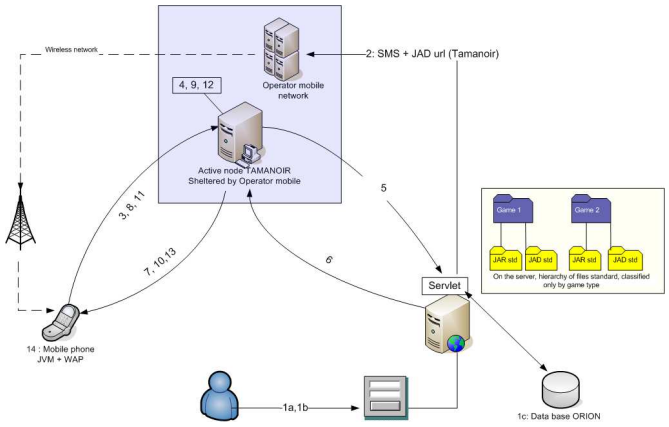


Figure 3.32: Déploiement de jeux avec adaptation par Tamanoir

Nous avons ainsi développé un service léger, déployé dans le réseau et assurant le transcodage à la volée. Avant déploiement en grandeur réelle, nous avons évalué cette approche en émulation à l'aide de serveurs et clients hétérogènes. Différentes expériences ont montré que le temps de traitement imposé par Tamanoir demeure négligeable dans le déploiement des applications. Nous avons aussi montré l'impact positif du déploiement de services d'adaptation dans le réseau avec un temps de développement d'applications réduit, une économie de bande passante avec les serveurs des opérateurs et une amélioration des temps de téléchargement des applications. Notre approche a remis en cause la manière de travailler de la société 3DDL, en leur permettant de répondre plus rapidement aux besoins de leurs clients et de mieux maîtriser leur infrastructure de développement et de déploiement.

3.4.2 Adaptation de flux multimédia pour réception sur terminaux hétérogènes

Les nœuds actifs conçus dans le projet RNRT Temic (IAN2 - section 3.4) en plus de répondre à différentes spécificités industrielles, doivent aussi supporter de l'adaptation de contenus à l'intérieur du réseau. Nous validons ainsi la flexibilité multimédia à l'aide de services type "poids lourd" afin de supporter une variété de clients hétérogènes. Ces travaux ont été menées avec

Martine Chaudier et Pierre Bozonnet (ingénieurs dans projet Temic) et Jean-Francois Rolland (Collaboration avec la société 3DDL).

Par défaut, l'hétérogénéité des machines clientes doit être supportée par le serveur qui délivre des flux de plusieurs qualités; les clients s'abonnent au flux correspondant. Cette approche impose plus de contraintes (calcul, utilisation réseau ou stockage de plusieurs flux) au niveau du serveur qui prend en charge la délivrance du bon flux. L'adaptation peut aussi être confiée aux équipements réseaux (type passerelle) afin que l'adaptation se fasse au plus près des terminaux ou des points d'engorgement réseau.

Dans le projet Temic, un ensemble de capteurs (caméras, capteurs de température, détecteurs d'alarmes) génère des flux multimédia collectés et éventuellement stockés ou archivés en qualité maximale par la passerelle active (figure 3.33). Lorsqu'un utilisateur armé de son *smartphone* ou *PDA* vient collecter ces données, celles ci sont distribuées avec une adaptation à la volée des flux vidéos pour les adapter aux contraintes réseaux (bande passante limitée) ou aux contraintes de l'équipement mobile (résolution d'écran faible, nombre de couleurs..) (figure 3.34).

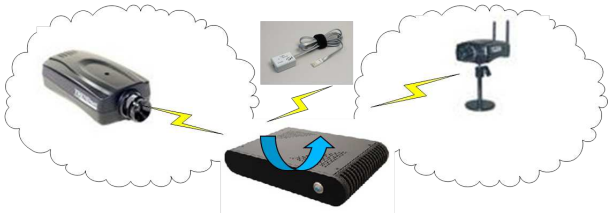


Figure 3.33: Collecte de flux multimédia

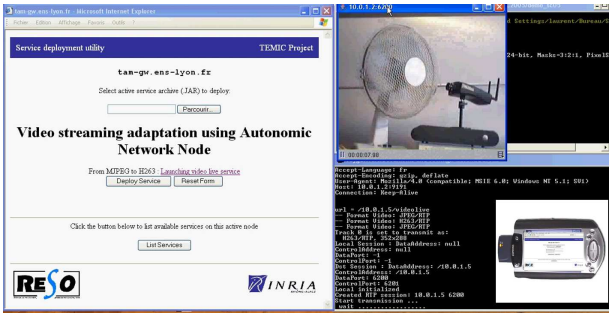


Figure 3.34: Adaptation à la volée et distribution à un terminal mobile

3.5 Conclusion

Ce chapitre présente certains équipements et services que nous avons étudiés dans nos travaux sur la flexibilité. Ils font apparaitre les bénéfices apportés par la flexibilité réseau en termes de nouvelles fonctionnalités : interopérabilité (XCP-i), support d'infrastructures hétérogènes (Grille Active), adaptations dans les réseaux pour les applications et le flux.

Grâce aux propositions et développements réalisés dans la thèse de Dino Lopez Pacheco, XCP est passé d'un stade "protocole jouet de laboratoire" à un protocole capable d'affronter les contraintes et la flexibilité d'infrastructures distribuées à grande échelle. Nous avons proposé une approche en douceur, garantissant une inter-opérabilité avec l'existant (protocole XCP, équipements non XCP et flux TCP). Nos solutions ont aussi été développées en imposant le maximum de légèreté en termes de consommation CPU et le non usage d'états par flux afin de garantir de bonnes performances aux flux de données.

Nous avons montré que la flexibilité que nous proposons est capable de supporter des infrastructures à venir (Grille InterPlanétaire) mais aussi de répondre à des contraintes industrielles actuelles (thèse de Narjess Ayari avec l'opérateur Orange et déploiement de jeux avec la PME 3DDL).

Les services que nous avons développés, répondent à des besoins précis et sont optimisés pour supporter finement les requêtes des applications. Nous suivons donc une démarche de flexibilité maitrisée où seul un sous ensemble défini d'utilisateurs ou d'applications est autorisé à déployer de nouvelles fonctionnalités.

Part II

Améliorer l'efficacité énergétique des infrastructures à grande échelle

Si le problème a une solution, il ne sert à rien de s'inquiéter. Mais s'il n'en a pas, alors s'inquiéter ne change rien.

Proverbe Tibétain

4

Mesurer et comprendre l'usage électrique des systèmes distribués à grande échelle

Alors que dans le monde, les personnes privées d'électricité et de lumière représentent 1,5 milliard d'individus, soit un quart de la population; on observe de nombreux gaspillages d'électricité associés aux équipements électronique grand public. Par exemple, selon la deuxième édition du baromètre AFP-Powermetrix (publiée le 10 juillet 2013), « les appareils en veille représentent 11 % de la facture d'électricité des Français, soit 86 euros par foyer et près de 2 milliards d'euros au total chaque année ».

Les infrastructures distribuées à grande échelle sont de gros consommateurs d'énergie électrique et sont pointés du doigt [84, 83]. La consommation des *datacenters* est estimée à 2% de la consommation électrique mondiale. Les plus gros centres de calcul cités dans le Top500 [1] affichent des consommation de plusieurs MWatts. Ces infrastructures n'échappent pas au gaspillage. Des centres de données sont sur-dimensionnés, sur-refroidis; des infrastructures réseaux sont redondantes afin d'assurer une qualité de réponse et de service proche de la perfection.

Mon intérêt pour l'efficacité énergétique des infrastructures distribuées à grande échelle est partie d'un constat très simple. En 2007, j'observais une grappe de machines (*cluster*) de la plate-forme *Grid5000* [37] : une machine imposante, ventilée, bruyante dans une salle climatisée. Je me demandais si cette machine était bien dimensionnée face aux besoins des applications et des services qui devaient s'exécuter dessus. Et surtout je me demandais quelle pouvait bien être la consommation électrique de cette plate-forme, son usage effectif, son coût financier. A cette époque, peu d'activités de recherche étaient menées sur ce domaine. Aux Etats Unis, l'Université de *Virginia Tech* avait lancé le projet GreenDestiny [67]. La conception de cette grappe de machines consommant l'équivalent d'un sèche-cheveux rendait sceptique une partie de la communauté scientifique. Certes, les machines consommaient peu d'électricité, mais le "service rendu" était bien faible en termes de puissance de calcul ou de stockage.

Décidés à participer à cette aventure, il nous fallait d'abord prendre certaines compétences. En partant du constat que "l'on ne comprend bien que ce que l'on mesure bien", nous avons développé une méthodologie expérimentale d'observation à la fois de l'usage de nos infrastructures distribuées à grande échelle mais aussi de leur comportement en termes de consommation électrique. Nous avons ainsi proposé la mise en place d'une infrastructure de mesures électriques

[1] TOP500 Supercomputer Sites : `http://top500.org/`

à grande échelle originale et unique dans sa dimension. Cette infrastructure a été utilisée dans le cadre des différentes thèses que j'ai co-encadrées sur ce domaine.

Avec Anne-Cécile Orgerie (thèse co-encadrée avec Isabelle Guérin Lassous, 2008-2011), nous avons défriché le terrain de la mesure énergétique à moyenne échelle. Nous nous sommes focalisés sur notre objet d'étude : la plate-forme Grid5000 et la mise en place d'une infrastructure logicielle et matérielle de collecte, traitement et mise à disposition de mesures énergétiques.

Ces activités ont notamment été menées dans le cadre des projets ARC INRIA Green-Net (2008-2010), du projet Européen IEE PrimeEnergyIT (2010-2012).

Dans la thèse de Mehdi Diouri (co-encadrée avec Olivier Gluck, 2010-2013), nous avons poussé nos explorations dans la compréhension des consommations énergétiques des composants des systèmes distribués à grande échelle.

Nous avons ainsi pu contredire diverses contre-vérités et analyser finement des applications et services déployés sur des systèmes distribués à grande échelle [137, 58, 56]. Ces activités ont été menées dans le cadre de l'action Européene COST IC 804 (2009-2013) ainsi que dans le laboratoire commun INRIA-Argonne National Laboratory (depuis 2010).

4.1 De l'usage des infrastructures à grande échelle à l'usage électrique

Avec Anne-Cécile Orgerie (doctorante MENRT co-encadrée avec Isabelle Guérin Lassous), nous avons mené une étude poussée de la consommation énergétique d'un infrastructure distribuée à grande échelle.

Notre première interrogation a été de comprendre si les plate-formes de calcul sont vraiment sous utilisées. Nous nous somme focalisés sur la plateforme expérimentale nationale Grid5000. Nous avons collecté et analysé les traces de réservations de nœuds de Grid'5000 pour chaque site durant deux années complètes (2007[132] et 2008[134]). Le tableau 4.1 donne des valeurs moyennes par site pour 2008 : le nombre de réservations sur l'année, le nombre de nœuds disponibles à la fin de l'année (certaines nœuds étant rajoutés en cours d'année), le nombre moyen de ressources par réservation, la durée moyenne d'une réservation et le pourcentage de temps d'utilisation sur l'ensemble de l'année.

Site	Nb de réservations	Nb de nœuds	Nb moyen de nœuds par réservation	Durée moyenne d'une réservation	Utilisation
Bordeaux	356222	650	7.44	2473.38 s	53.20 %
Lille	344538	618	8.11	3154.58 s	72.89 %
Lyon	138217	322	4.39	3723.55 s	69.27 %
Nancy	74592	574	14.63	8912.82 s	60.08 %
Orsay	92862	684	14.58	6246.07s	57.82 %
Rennes	58843	714	27.32	7069.33 s	64.58 %
Sophia	58142	568	22.14	8767.35 s	81.51 %
Toulouse	166191	434	6.29	2211.80 s	61.67 %

Table 4.1: Utilisation des différents sites de Grid'5000 sur l'année 2008

On peut constater que les ressources sont fortement sollicitées. Tous les sites sont à plus de 50% d'utilisation en 2008. Les pourcentages d'utilisation varient également beaucoup. En moyenne, sur l'ensemble des sites, la plate-forme a été utilisée à 40% pour l'année 2007 et à 65% pour l'année 2008. On peut constater que ces résultats sont très disparates d'un site à l'autre. De plus, les valeurs élevées des écarts types, pour chaque site, soulignent la disparité à la fois des durées des réservations et de leur nombre de ressources.

La figure 4.1 présente l'utilisation des différents sites de Grid5000 par semaine sur l'année 2007. La ligne en rouge (Jobs) indique le nombre total de jobs par semaine. On observe une

grande disparité parmi les sites, mais aussi une utilisation non constante des ressources de la plate-forme. On observe des périodes de pics d'activité où la plate-forme est utilisée dans son ensemble à plus de 95 % pendant plusieurs semaines d'affiliée. Néanmoins, des contraintes temporelles (jour/nuit, période de vacances), géographiques (intérêt des utilisateurs pour leur propre site), professionnelles (*deadline* avant les conférences importantes) font varier l'usage de la plate-forme. Alors que pendant certaines semaines, l'usage est intensif, il existe des périodes d'accalmie où le nombre de ressources est sur-évalué par rapport aux besoins réels des utilisateurs et de leurs applications.

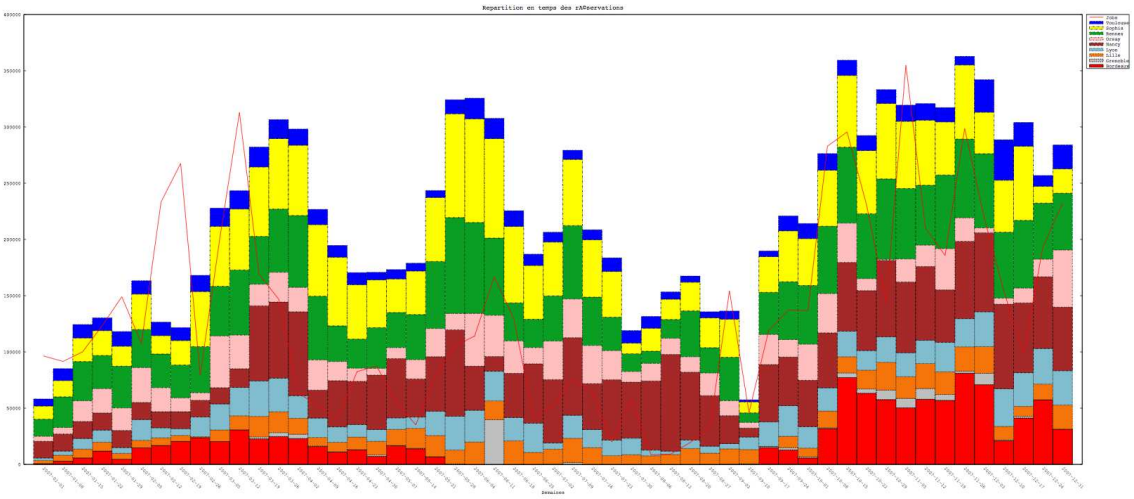


Figure 4.1: Utilisation des différents sites de Grid5000 par semaine sur l'année 2007

L'utilisation d'une plateforme comme GriD5000 est donc constituée de pics et de creux en termes de charge. Dans une optique de réduction de la consommation énergétique, dimensionner finement le nombre de ressources disponibles en fonction des besoins des applications et des utilisateurs est donc une approche intéressante.

4.2 Maîtriser les équipements de mesure et offrir de nouveaux services aux utilisateurs

4.2.1 Dans la jungle des wattmètres

De nombreux systèmes de mesures électriques de référence tels que Powerpack[75] considèrent qu'une mesure unique sur un serveur de l'infrastructure suffit à étalonner l'ensemble des mesures et comportements électriques pour la machine. Nous avons opté pour une approche différente en fondant nos travaux expérimentaux sur une mesure de l'infrastructure complète afin de prendre en compte les contraintes d'environnement (positionnement des machines dans l'infrastructure des salles machines, hétérogénéité). Notre approche se révélera très utile lors de l'analyse de certains comportements électriques (section 4.4).

Une fois l'usage analysé, comment mesurer la consommation électrique des ressources ?

La première difficulté a été de mesurer l'utilisation électrique d'un ensemble de composants informatiques (serveurs, équipements de stockage, équipements réseaux...). Au début de nos travaux dans ce domaine, nous sommes partis à la recherche d'un équipement de mesure de con-

sommation électrique de type wattmètre capable de répondre à nos besoins. Un wattmètre est un instrument de mesure de la puissance électrique consommée dans un circuit. Nous désirions un équipement capable de mesurer la consommation sur une prise électrique externe. La plupart des wattmètres disponibles à cette époque étaient des équipements de type PDU (*Power Distribution Unit*) pour une grappe complète de machines ou par prise électrique mais avec une mesure touts les dizaines de secondes. Ce genre d'équipement, très utile pour détecter les problèmes électrique au sein d'une salle machine et pour remonter des alertes, ne correspondait pas vraiment à nos besoins. Je me suis mis en relation avec la PME Valentinoise OmegaWatt[2] qui a adapté un équipement existant pour répondre à nos besoins. Nous avons donc pu disposer d'un wattmètre (Figure 4.2) capable de mesurer la consommation de 6 prises électriques et de remonter une information de puissance électrique instantanée toutes les secondes. En fait, la puissance mesurée est une puissance instantanée moyennée (3900 points de mesures par seconde). Ce fut le début de nos travaux dans ce domaine en nous permettant de mesurer finement et d'analyser la consommation électrique de ressources informatiques (section 4.3).

Nos travaux nous ont amené par la suite à considérer et expérimenter d'autres équipements de mesures[56] : **externes** Omegawatt v2 (boitiers de 25 kg pour 48 ports de mesures figure 4.3), Zimmer (fig. 4.4), des PDUs EATON (Fig. 4.5), Raritan, Schleifenbauer, **internes** (déployés à l'intérieur des serveurs) tels que PowerMon (fig. 4.6), NI, DCM (table 4.2) et **embarqués** (disponibles sur la carte mère fournie par les constructeurs de machines). Certains de ces équipements (hors PDUs et capteurs embarqués) sont décrits dans la table 4.2.

Figure 4.2: Wattmètre Omegawatt v1

Figure 4.3: Wattmètre Omegawatt v2

Figure 4.4: Wattmètre Zimmer

4.2.2 Showwatts :une suite logicielle pour les chercheurs en efficacité énergétique

Pour manipuler cet ensemble d'équipements de mesures, nous avons proposé différents outils et *frameworks* logiciels afin d'aider une large classe d'utilisateurs à mener des campagnes de mesures de consommation électrique. La suite **Showwatts** regroupe ces outils (de collecte, pré-traitement et archivage, exposition de traces) et a été développée dans le cadre de différents projets (ARC Green-Net, Projet Européen PrimeEnergyIT) avec les principales contributions

[2] `http://www.omegawatt.fr/`

Figure 4.5: PDU Eaton

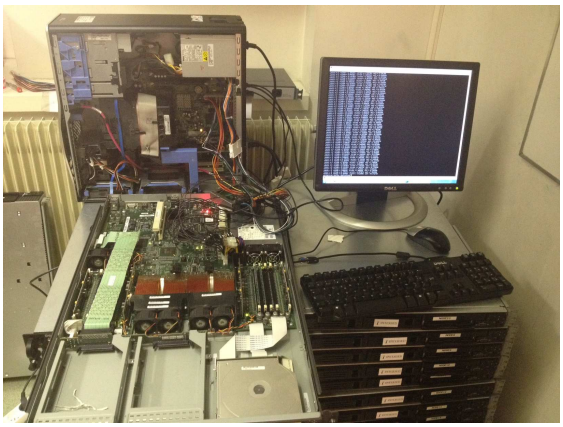


Figure 4.6: Déploiement de Wattmètre Powermon sur site

	Externe			Interne		
Wattmètre	OmegaWatt	WattsUp	LMG 450	PowerMon2	NI	DCM
Compagnie	OmegaWatt[a]	WattsUp?[b]	Zimmer[c]	RENCI iLab[d]	National Instruments[e]	Universitat Jaume I
# Canaux	6	1	4	8	32	12
Branchement	Prise électrique	Prise électrique	Prise électrique	Lignes ATX- (3.3 V, 5 V, 12 V)[f]	Ligne 12 V ATX	Ligne 12 V ATX
Power nature	Moyennée	Instantanée/moyennée	Instantanée	Instantanée	Instantanée	Instantanée
Microcontroleur	-	-	-	Atmel ATmega16	NI9205 NIcDAQ-9178	*Microchip* PIC 18
Capteurs puissance	-	-	-	*Analog Devices ADM1191* resistors	*LEM HXS 20-NP* transducers	*LEM HXS 20-NP* transducers
Fréquence de mesure par seconde par canal (max)	1	1	300	1024	1000	28
Précision	$< \pm 1\%$	$< \pm 1.5\%$	0.1%	$\pm 5\%$	$\pm 1\%$	$\pm 1\%$
Interface	RS232	USB	RS232	USB	USB	RS232
Prix d'achat	600 Euros	200 Euros	11000 Euros	125 Euros	2700 Euros	Non comercialisé

[a] OmegaWatt: `http://www.omegawatt.fr/`
[b] WattsUp: `https://www.wattsupmeters.com/`
[c] Zimmer: `http://www.zes.com`
[d] PowerMon2: `http://ilab.renci.org/powermon`
[e] NI: `http://www.ni.com/`
[f] Les lignes 3.3 V et 5 V permettent la mesure de puissance de certains composants (GPUs, cartes réseaux, etc.). Les lignes 12 V permettent la mesure de la consommation électrique du processeur et des ventilateurs.

Table 4.2: Spécifications des Wattmètres utilisés dans nos travaux (hors PDU et capteurs embarqués)

de Anne-Cécile Orgerie, Jean-Patrick Gelas, Marcos Dias de Asuncao et Olivier Mornard.

Un ensemble de démons collecte les informations de puissance électrique de chacune des ressources observées et stocke ces informations dans différentes bases de données. Ainsi le site de Lyon de la plate-forme Grid5000 est surveillé par un ensemble de 138 capteurs wattmètres de type Omewgatt qui génèrent des fichiers de traces et les met à disposition sous différents formats (Figure 4.7).

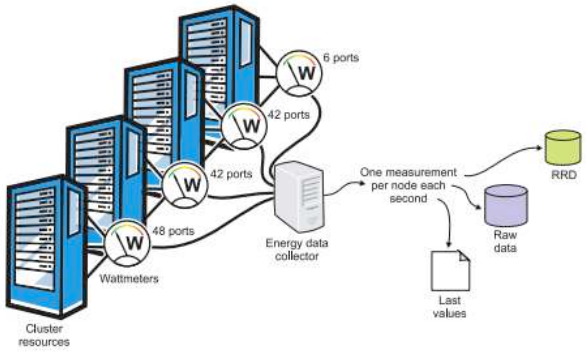


Figure 4.7: Infrastructure de collecte du site de Grid5000 à Lyon

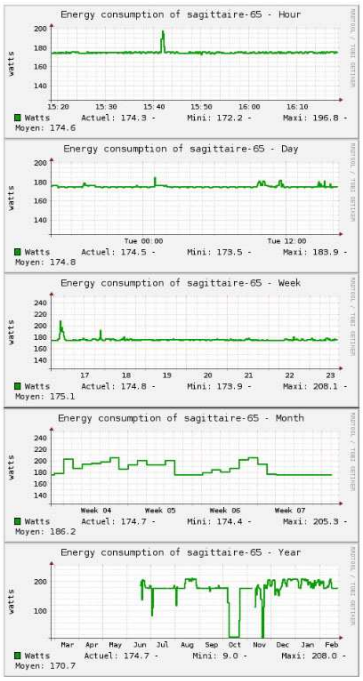


Figure 4.8: Consommation énergétique d'un nœud de calcul de plate-forme Sagittaire

L'utilisateur peut connaitre la consommation d'une ressource de calcul ou d'un équipement réseau en consultant des mesures sur la dernière heure, le jour, la semaine, le mois ou l'année (Figure 4.8).

Un tableau de bord pour un site complet est ainsi proposé aux utilisateurs de la plate-forme Grid5000 (Figure 4.9) et les utilisateurs peuvent extraire un ensemble de logs d'énergie ainsi que les graphes pour une expérimentation donnée.

La figure 4.10 présente le tableau de bord pour l'étude de la consommation instantanée de nœuds de calculs et d'équipements réseau. L'utilisateur sélectionne les équipements à surveiller et peut vérifier en direct leurs consommations électrique en watts ainsi que le coût en électricité (en euros).

Cette infrastructure matérielle et logicielle de collecte nous a permis de mettre en place l'architecture *Green_Grid5000* [52] disponible pour les utilisateurs du projet Green-Net puis étendue à tous les utilisateurs de Grid5000. Dans l'action européenne COST IC804[47], avec Anne-Cécile Orgerie et Marcos Dias de Asuncao, nous avons aussi mis à disposition un ensemble de traces énergétiques ainsi que d'usage de la plate-forme Grid5000 : le projet *"ICT Energy Logs"*.

Ce *repository* permet à d'autres chercheurs du domaine de mener des simulations et de valider leurs modèles de réduction d'énergie en utilisant un jeu de traces référence.

Figure 4.9: Tableau de bord (site web) de consommation du site Grid5000 de Lyon

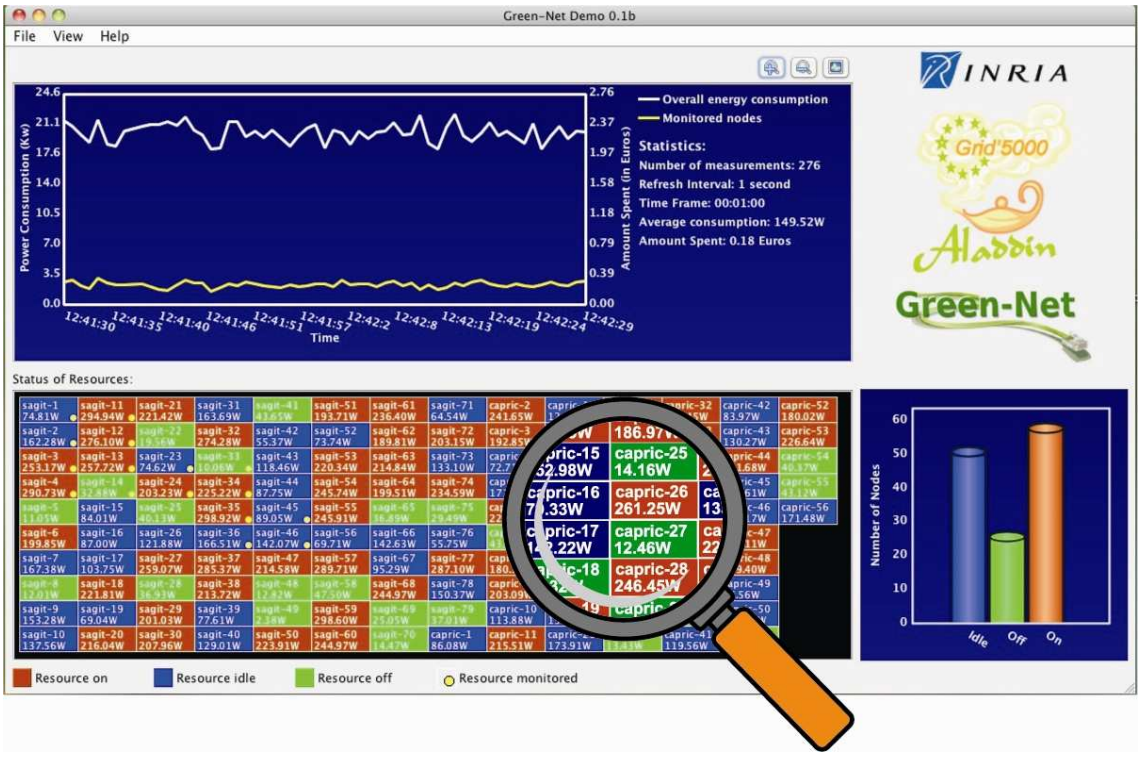


Figure 4.10: Tableau de bord (application java) pour étude de la consommation instantanée d'un grand ensemble de nœuds de calcul

4.3 D'une compréhension locale à une compréhension globale

4.3.1 Profiler des infrastructures physiques

Cette infrastructure de mesures[52] nous a permis d'évaluer différents scénarios et contextes. Notre première tâche a été de comprendre le comportement électrique des ressources informatiques considérées : serveurs de calcul, terminaux légers et équipement réseaux.

Dans les centres de données, les matériels informatiques (serveurs, stockage, réseau) sont confrontés à deux types de consommation d'énergie :

- **Consommation statique** qui provient des composants matériels (ventilateurs , disque dur, réseaux ...). Cette consommation statique correspond à la consommation d'électricité des appareils en mode inoccupé (sans services ni applications).
- **Consommation dynamique** qui résulte de l'utilisation des ressources (mémoire, CPU, entreéessorties) par les applications et les services.

4.3.1.1 Profilage énergétique d'un serveur

La première mesure menée concerne la compréhension de la consommation d'un équipement informatique de type nœud de calcul, en profilant la consommation énergétique en fonction de la charge en termes de services, d'applications, du volume de données traitées...

Pour la suite de nos travaux, nous considérons les bancs d'essai suivants qui utilisent intensément une ressource spécifique (processeur, disque dur et mémoire RAM) :

- `idle` : Le serveur est allumé et n'exécute que le système d'exploitation. C'est un état d'inactivité.
- `iperf`[3] : Cet outil génère des communications intensives sur le réseau dans l'optique de mesurer son débit. Il peut être configuré avec un trafic TCP ou UDP entre un serveur et un client. Dans nos mesures, nous utilisons cet outil avec un trafic TCP.
- `hdparm`[4] : Cette application fournit une interface en ligne de commande pour divers noyaux supportés par le sous-système SATA/PATA/SAS *libATA* de Linux. Nous l'utilisons pour accéder intensivement au disque dur.
- `cpuburn`[5] : Ce banc d'essai exécute un nombre maximal de calculs en virgule flottante en vérifiant les résultats renvoyés pour garantir une utilisation CPU maximale.
- `burnMMX`[6] : Ce programme, inclus dans le paquet de `cpuburn`, utilise intensivement le cache et la mémoire.

La figure 4.11 illustre la consommation électrique en termes de puissance électrique (watts) d'un serveur de calcul datant de 2006 soumis à plusieurs bancs d'essais qui placent le serveur dans différentes phases. Ce serveur consomme une dizaine de watts en mode veille. En phase de démarrage ("*boot*"), on observe une consommation en pic lorsque le serveur vérifie les bancs mémoire et démarre les différents composants mécaniques (ventilateurs, disques durs..etc). Pendant cette période la consommation atteint un pic de 320 Watts. En phase d'extinction , un pic de consommation moindre est aussi observé ("*turn off*"). Lorsque que le système d'exploitation est chargé et que la machine n'exécute aucune application (phase "*idle*"), on observe une puissance consommée de l'ordre de 190W (85% de la consommation totale). C'est ce que nous définissons comme la consommation statique d'un serveur (mode `idle`).

[3] `iperf` : `http://iperf.fr`
[4] `hdparm` : `http://linux.die.net/man/8/hdparm`
[5] `cpuburn` : `http://manpages.ubuntu.com/manpages/precise/man1/cpuburn.1.html`
[6] `burnMMX`: `http://pl.digipedia.org/man/doc/view/burnMMX.1`

Lors de l'exécution de *benchmarks* tels que des accès disques intensifs (*disk access*), des calculs et communications intensives (*communicating and computing* phase de `cpuburn` et `iperf`), on observe que la puissance électrique varie en fonction de la charge de travail. La part de la consommation dynamique (hors "*boot*") par rapport à la puissance totale consommée par la machine est vraiment faible (une trentaine de watts pour ce type de serveur).

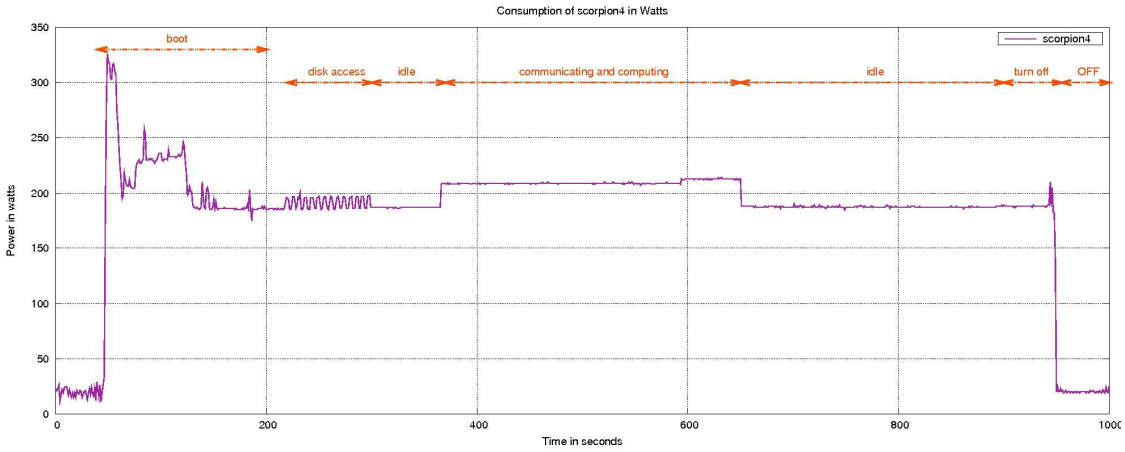


Figure 4.11: Consommation électrique d'un serveur de calcul (HP Proliant 85 G2 - 2.2GHz, 2 duo core CPUs par nœud) - Mesures avec Wattmètre OmegaWatt

On considère que la part de la consommation statique diminue dans les nouvelles générations de serveurs. Nous avons voulu vérifier ce point et si ce profilage énergétique de serveur est encore d'actualité avec un serveur plus récent (Machine PoweEdge DELL R610) (Figure 4.12). Ce serveur consomme aussi une dizaine de watts en mode veille (*shutdown*). La phase de démarrage (*cold booting*) ainsi que celle de réveil après hibernation (*wake up after hibernation*) ont le même profil avec des pics de consommation élevée. Ce serveur a une consommation statique (idle) de 90 W, alors que l'on peut observer des pics de consommation importants jusqu'à 190 W lors de calculs numériques intensifs (*cpuburn*). Les phases d'accès mémoire et réseaux génèrent une consommation de l'ordre de 120 W. Cette mesure nous confirme que la part de la consommation statique a été fortement diminuée (de l'ordre de 50%) et que la charge de calcul est un point de consommation électrique important pour ce type de serveur.

L'utilisation d'un wattmètre Zimmer (table 4.2), nous permet de mesurer différentes métriques électriques : la puissance apparente (VA : Voltampère), la puissance réactive (VAR : Voltampère réactif) et la puissance active (Watts) ainsi que le facteur de puissance (Phi en degrés) (Figure 4.12). Pour la suite de nos travaux, nous nous sommes uniquement focalisés sur la puissance active dont l'unité est le Watt (Tension (volt) * Intensité (ampère)) ainsi que l'énergie électrique consommée dont l'unité est le WattHeure (Wh) (Puissance * temps).

4.3.1.2 Profil d'un équipement léger

Dans le cadre du projet ANR DSLLAB, nous avons participé au déploiement et à la conception d'une plate-forme expérimentale destinée à l'évaluation des performances d'infrastructures réparties autour de liens ADSL[65]. Une quarantaine de nœuds légers de type IAN2 (section 3.4) ont été répartis en France et hébergés par des particuliers.

Ces nœuds légers étant hébergés par des volontaires, nous souhaitions pouvoir leur donner une estimation précise du coût électrique pendant les différents phases applicatives. La figure 4.13 présente le profil électrique de 6 nœuds de la plate-forme. Une machine consomme aux alentours de 1.5W quand elle est en veille (coût de la carte réseaux en attente de réveil en mode *wake on lan*) et de 9 à 10 W quand la machine est disponible sans exécuter d'application. Lorsque

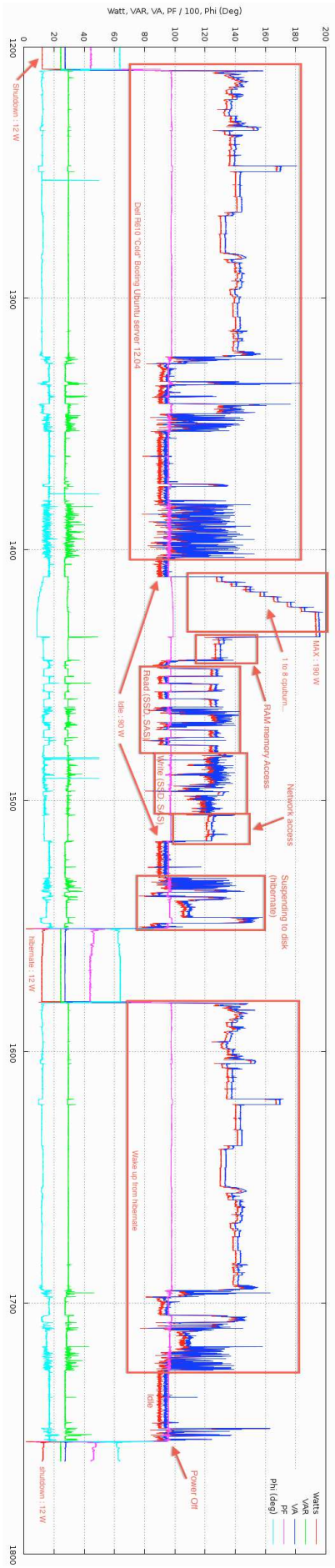


Figure 4.12: Consommation électrique d'un serveur de calcul DellR610 (Wattmètre Zimmer)

que la machine effectue des opérations de manière intensive (type `cpuburn`), elle consomme de 13 à 14W. Comme pour les serveurs de calcul on observe un pic de consommation significatif en phase de démarrage. Nous avons mis en place une politique de mise en veille des machines lors de leur non utilisation. Ces mesures nous ont permis de quantifier un coût d'hébergement de cette machine de l'ordre de 3.60 euros par an (pour une machine utilisée 8 heures par jour, 5 jours par semaine, 11 mois par an).

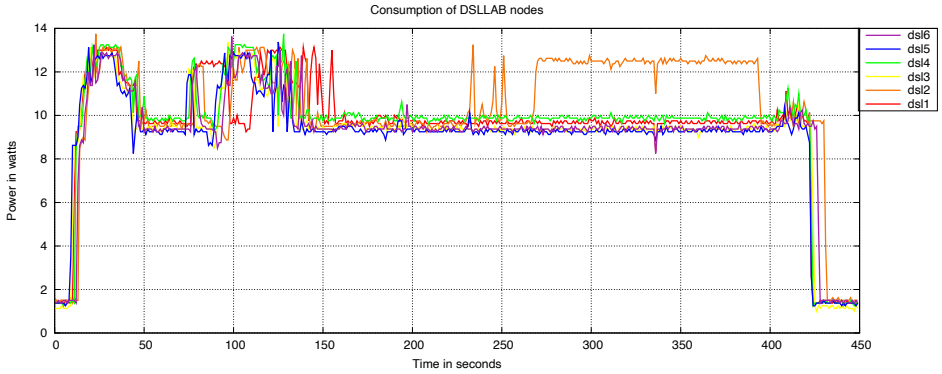


Figure 4.13: Consommation électrique des nœuds de la plate-forme DSLLAB[65]

4.3.2 Profiler des infrastructures virtuelles

En instantiant des solutions de virtualisation sur des infrastructures physiques, l'utilisateur peut manipuler des machines virtuelles qui hébergent les applications et les services. La plate-forme *Green_Grid5000* nous permet ainsi de mesurer une infrastructure déployant des machines virtuelles [138]. La figure 4.14 montre le surcoût associé au déploiement de machines virtuelles de calcul intensif (`cpuburn`).

Une opération de migration est une des facilités importantes qu'offre la virtualisation. Même si cette solution doit être utilisée avec précaution [168], elle permet facilement d'équilibrer la charge dans les infrastructures distribuées à grande échelle. Le profil énergétique d'une migration (Figure 4.15) nous apprend que cette opération a un coût énergétique non négligeable : 4 machines virtuelles (512 MO) déployées sur la machine 1, commencent à migrer à la seconde 40 vers la machine 2. Cette migration provoque une augmentation de la consommation de la machine réceptrice. Mais la machine émettrice continue elle aussi à consommer de l'énergie pendant la phase de migration. Des pics électriques sont aussi observables lors de chaque opération.

4.3.3 Profiler des applications et des services

Il est fondamental de mesurer la consommation électrique d'une infrastructure physique afin de comprendre les impacts de la charge applicative sur la puissance électrique. Mais nous poussons plus loin nos mesures, en nous focalisant sur la mesure de la consommation des applications et des services considérés par certains de nos travaux. Ainsi dans le cadre de la thèse de Mehdi Diouri[55], nous avons mené des campagnes de mesures sur différents services de tolérance aux pannes et de distribution de données nécessaire aux infrastructures distribuées à grande échelle (voir section 6.1).

La figure 4.16 permet de comparer les coûts énergétiques de bancs d'essais sur un nœud de la plate-forme Grid5000 lyonnaise Taurus (figure 4.16). On observe que le coût de `iperf` n'augmente pas avec le nombre de processus, alors que le nombre de processus `cpuburn` a une incidence forte sur la consommation électrique.

En traçant la consommation énergétique des bancs d'essai (en joules) à l'aide de différents wattmètres, on se rend compte de l'importance du calcul intensif (`cpuburn`) et de l'utilisation

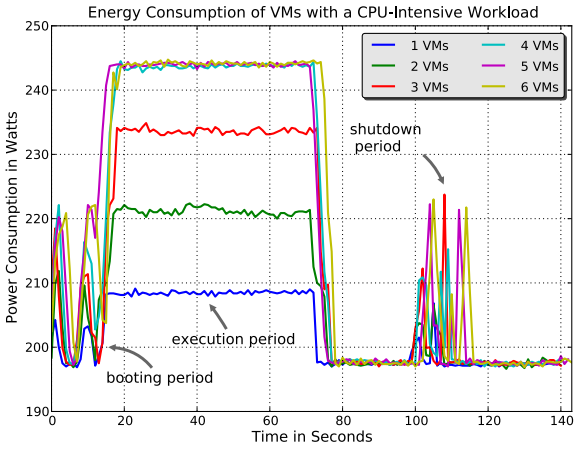


Figure 4.14: Profil énergétique de machines virtuelles déployées sur un serveur

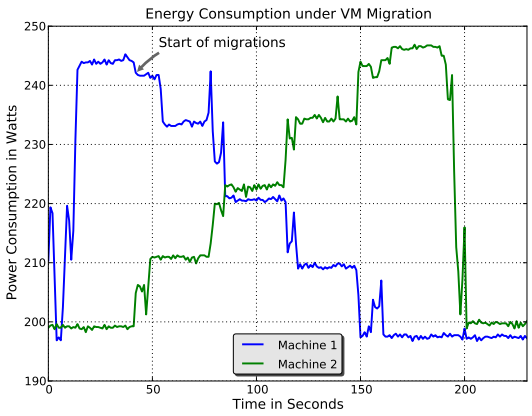


Figure 4.15: Profil énergétique d'un ensemble de machines virtuelles migrant entre deux serveurs

mémoire (`burnMMX`) (figure 4.17).

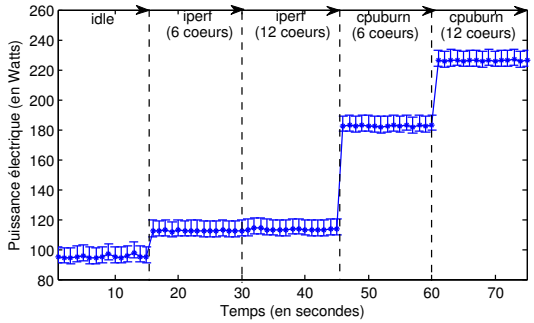


Figure 4.16: Consommation électrique de bancs d'essais sur 3 serveurs de la plate-forme Taurus

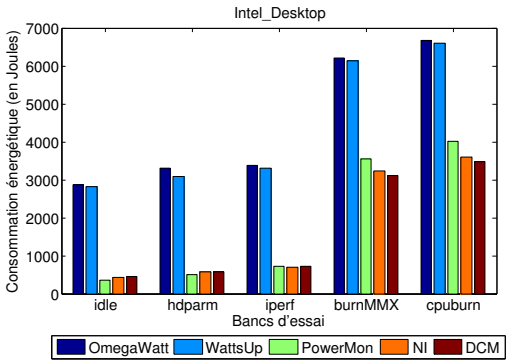


Figure 4.17: Consommation de bancs d'essais (joules) pendant 60 secondes sur 1 serveur de Taurus

4.4 Démystifier et analyser certains usages électriques des TICS

La communauté des systèmes distribués s'est récemment emparée des problématiques énergétiques et différentes erreurs ou imprécisions circulent dans la littérature. Grâce à notre arsenal de capteurs énergétiques hétérogènes, nous avons voulu explorer différents "mythes" associés à la consommation énergétique des infrastructures distribuées à grande échelle. Je présente ici deux exemples illustratifs (plus de mythes sont disponibles dans [137, 58, 56]).

4.4.1 Mon wattmètre est le bon, je suis sûr de ce que je mesure !

En 2012, lors d'un meeting de l'action européenne COST IC 804 sur l'efficacité énergétique, je suis impressionné par les travaux d'une équipe comparant des implémentations d'applications distribuées à quelques % près en termes de consommation énergétique. En discutant avec l'un des chercheurs, sur leur manière de mesurer la consommation électrique, j'appris que l'équipement de mesure était un simple wattmètre du commerce avec écran LCD disposé sur la prise d'un

serveur d'une salle machine. Afin d'accéder à distance aux mesures, une *webcam* était orientée sur le wattmètre et renvoyait une vidéo à un doctorant dans son bureau qui recopiait en direct les mesures observées !

Nous nous sommes rendus compte de l'écart de certaines mesures observées avec nos wattmètres (figure 4.17). Nous avons poursuivi cette étude dans le cadre de l'action européenne COST IC804. J'ai mis en place une collaboration avec Manuel Dolz et Enrique Quintana-Orti de Université de Jaume1 (Castellon, Espagne). Lors de visites croisées entre Manuel Dolz et Mehdi Diouri, nous avons mené une des premières études s'interrogeant sur la pertinence de mesures de certains wattmètres [56].

4.4.1.1 Quelle précision ?

Si on compare des profils de puissance électrique lors d'un banc d'essai; on observe une différence de mesure entre les wattmètres externes (omegawatt et WattsUp) et internes (powermon, NI et DCM). Les wattmètres externes branchés sur la prise électrique du serveur enregistrent la consommation électrique de la totalité de la machine alors que les wattmètres internes branchés après le bloc d'alimentation ne mesurent qu'un sous ensemble des composants. Malgré cette différence en valeur absolue, certains profils "semblent" visuellement proches (figure 4.18). On retrouve la stabilité en termes de puissance électrique du banc d'essai `cpuburn`. Mais dans le cas d'un *benchmark* mélangeant différentes phases (type `hdparm`), on observe une plus grande disparité entre les mesures. C'est aussi le cas lors d'une mesures en mode `idle` (figure 4.19). Les wattmètres ne sont donc pas identiques en termes de précision et leur mode de mesure (instantanée, instantanée moyennée) influence les résultats collectés.

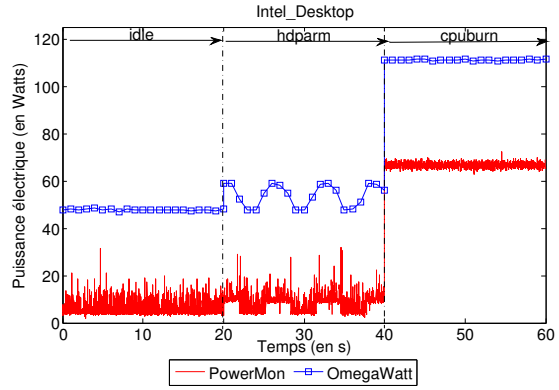


Figure 4.18: Profils de puissance en mode `idle` et avec les bancs d'essai `cpuburn` et `hdparm` avec des wattmètres externes et internes.

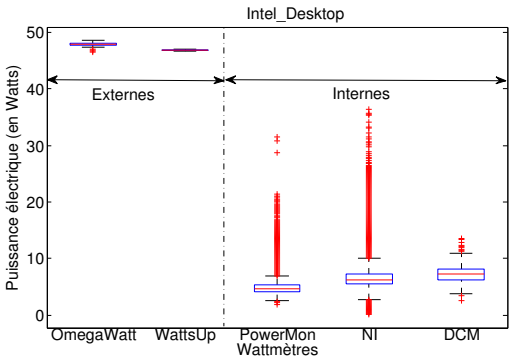


Figure 4.19: Variabilité dans les mesures de puissance active pendant 60 secondes lorsque la machine est en mode `idle`

4.4.1.2 Quelle est la bonne fréquence de mesure ?

Des mesures toutes les quelques secondes suffisent pour comprendre le comportement électrique d'une application régulière de calcul intensif (comme `cpuburn` figures 4.20 (a)). Plusieurs mesures par seconde peuvent perturber la compréhension de la consommation électrique. A l'inverse, pour des applications mélangeant des phases de calcul et d'entrée sorties (comme `hdparm` figure 4.20 (b)), une fréquence de plusieurs mesures par seconde semble indispensable pour évaluer finement le comportement de l'application.

Bien mesurer le comportement électrique d'une application demande donc une maîtrise assez fine des wattmètres et des paramètres nécessaires à la prise de mesure.

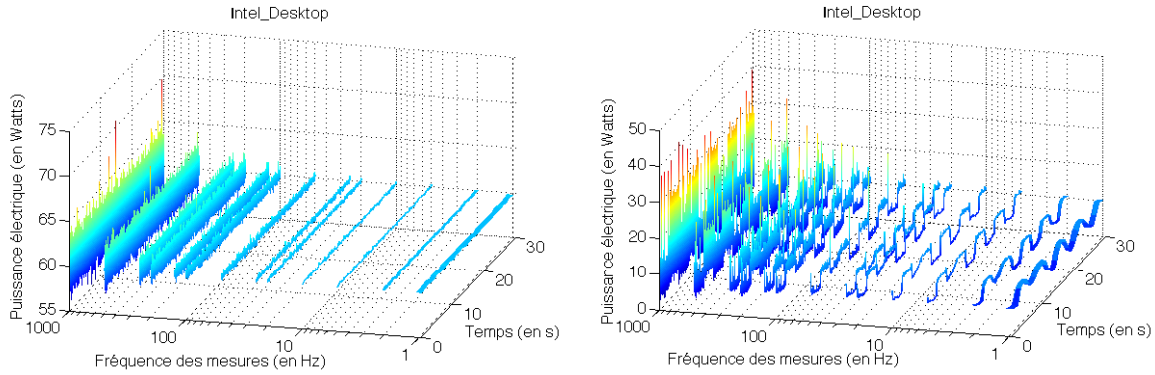


Figure 4.20: Variation de la fréquence de mesure sur *benchmarks* `cpuburn` (a) et `hdparm` (b)

4.4.2 Homogénéité en performance == homogénéité énergétique ?

De nombreux travaux [157, 95, 75] sur l'efficacité énergétique des grands systèmes distribués assument que des nœuds de calcul homogènes en performances sont aussi homogènes en consommation électrique. Par exemple l'article sur le framework Powerpack[75], qui fait référence dans le domaine considère qu'une mesure unique sur un nœud d'un data center est suffisante et que le profil obtenu peut être répercuté et appliqué comme modèle sur l'ensemble des nœuds de même catégorie. Ainsi les environnements logiciels et les simulations sont plus faciles à construire en respectant cette hypothèse [95].

Mais nous avions l'intuition (Figure 4.16) que cette hypothèse, plaisante pour les modèles énergétiques, n'est pas tout a fait conforme à la réalité observée [137, 58] Pendant la thèse de Mehdi Diouri, nous avons poursuivi nos investigations en menant des campagnes de mesures et d'analyse à grande échelle sur des infrastructures de la plate-forme Grid5000.

La figure 4.21 présente le profil temporel de puissance électrique obtenu sur les 60 nœuds de calcul de la grappe *Sagittaire*. Ce résultat permet l'observation de profils identiques en fonction des bancs d'essais choisis (consommation haute pour `cpuburn`, oscillations et consommation basse pour les accès disques générés par `hdparm`). Mais, cette figure montre aussi de grandes différences de consommation entre des nœuds aux architectures et performances identiques (en flops) pour la même exécution d'un banc d'essai. Lors de `cpuburn`, le nœud le moins consommateur affiche une puissance électrique d'environ 225 W tandis que le plus consommateur nécessite environ 275 W , soit un écart de 22%.

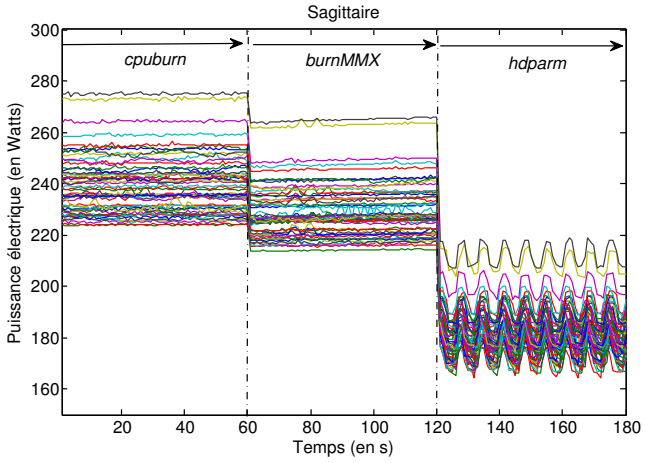


Figure 4.21: Puissance électrique des 60 nœuds identiques de la grappe *Sagittaire* éxécutant 3 bancs d'essais

Nous avons mis en place une véritable enquête (avec de nombreuses séries de mesures internes et externes) afin de trouver les causes de cette hétérogénéité en termes de consommation électrique. Dans la figure 4.21 nous remarquons que la différence de consommation entre les nœuds est la même quel que soit le banc d'essai exécuté. Nous nous sommes donc penchés sur la consommation des nœuds en mode `idle` où l'on retrouve cette même différence (figure 4.22). Cette disparité provient donc des équipements et non des applications. La grappe la plus ancienne (*sagittaire*) affiche une différence de 22% alors que la plus récente (*Taurus*) affiche des variations de 5% (figure 4.23).

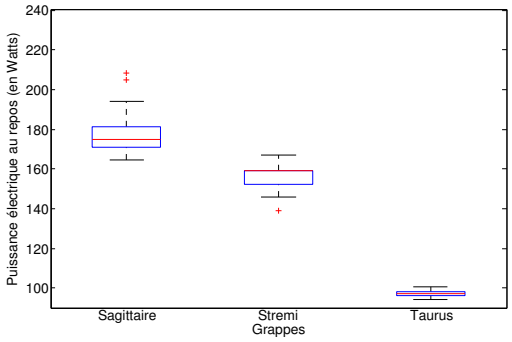


Figure 4.22: Puissance électrique au repos des nœuds issus de trois différentes grappes homogènes (Wattmètre Omegawatt)

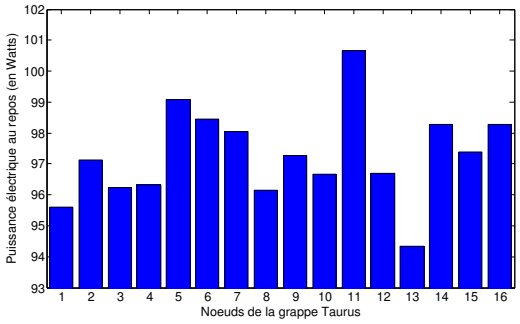


Figure 4.23: Puissance électrique des nœuds de la grappe Taurus en mode `idle`

Nous avons donc mené des mesures composant par composant (disque, cartes réseaux...) en modifiant l'infrastructure physique des nœuds pour comparaison. La figure 4.24 présente une mesure interne (*wattmètre Powermon*) de deux nœuds de plate-forme *sagittaire* effectuant les mêmes bancs d'essais. Cette mesure de puissance ne porte que sur les processeurs et les ventilateurs de la machine. On observe une différence de l'ordre de 38W quel que soit l'état de la machine : `idle`; `cpuburn`, `burnMMX`, `hdparm`. Cette différence est très proche de celle observée lors d'une mesure externe sur les nœuds les plus extrêmes de la plate forme sagittaire (figure 4.22).

Les responsables sont donc démasqués : les processeurs (différence de consommation dès la production des puces) et les ventilateurs (usure mécanique) sont parmi les plus gros générateurs d'hétérogénéité. Nous avons pris en compte cette différence lors de la création de nos environnements logiciels d'efficacité énergétique gérant les ressources (*Clouds Verts* - section 5.3). Par exemple, un placement de tâches sur des nœuds identiques en performance n'a pas le même impact énergétique à cause de l'hétérogénéité en consommation énergétique.

4.5 Conclusion

L'énergie est un métrique à part. Nous avons investi dans la mesure énergétique afin d'être capable d'analyser les comportements des infrastructures distribuées à grande échelle, de mettre en place de nouveaux modèles et d'offrir des environnements logiciels aux utilisateurs.

La création d'une plate forme de mesure énergétique et des environnements logiciels nécessaires à sa maîtrise a été consommatrice de temps et en moyens humains. Mais c'est l'étape obligée pour arriver à comprendre et profiler les applications, les services et les équipements. Dans le

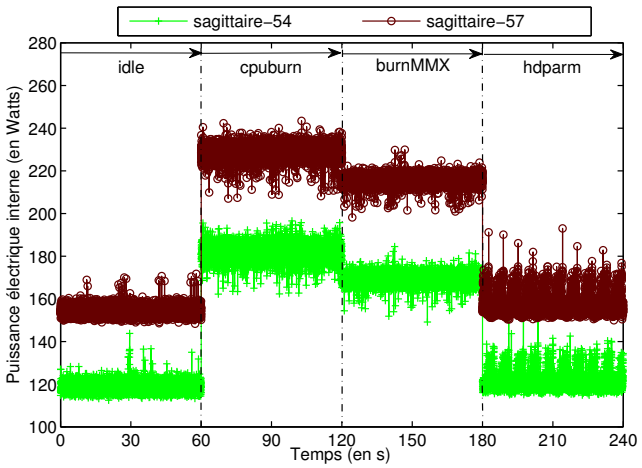


Figure 4.24: Puissance électrique des processeurs et des ventilateurs associés des deux nœuds spécifiques de la grappe *Sagittaire* pendant l'exécution de différents bancs d'essai

cadre de nos travaux, nous avons ainsi proposé la première solution de collecte d'informations à moyenne échelle (150 capteurs répartis sur différents sites). Cette infrastructure matérielle et logicielle est unique dans notre communauté scientifique. De nombreux travaux dans l'efficacité énergétique se contentent de quelques mesures sur un faible nombre de ressources. Puis ces mesures sont généralisés et appliquées à grande échelle dans les modèles proposés. Nous avons démontré que la mesure intégrale à fréquence raisonnable est un passage obligé pour éviter les écueils non maitrisés de l'hétérogénéité énergétique.

La suite logicielle Showwatts qui permet la mesure, la collecte et la mise à disposition de mesures énergétiques est un développement logiciel collectif qui s'est étalé sur différents projets et travaux de recherches. Cette suite s'enrichit en permanence lors de l'ajout de nouveaux capteurs électriques ou lors d'adaptations pour supporter de nouveaux scénarios.

Dans le cadre de la thèse de Anne-Cécile Orgerie, une étude complète de l'usage de Grid5000 et sa consommation électrique a permis de jeter les bases de nos travaux suivants. Cette étude a démontré l'usage en dents de scie d'une infrastructure distribuée à grande échelle comme Grid5000 et laisse entrevoir la possibilité d'optimisation et de gains énergétiques conséquents.

La thèse de Mehdi Diouri a permis de poursuivre certains travaux sur la mesure électrique en clarifiant notamment les aspects précision et hétérogénéité. Nous luttons contre certaines idées reçues : tous les équipements de mesure ne sont pas comparables en termes de qualité, précision et fréquence. Choisir le bon équipement et appliquer les bons réglages dépend de ce que l'on cherche à observer : des alertes et consommation cumulée pour les administrateurs de infrastructures distribuées à grande échelle, des profils pour étudier le comportement électrique des applications ou des valeurs instantanées pour observer des "picos" phénomènes. L'homogénéité en performances ($flops$) ne se retrouve pas en homogénéité en puissance électrique. Les machines vieillissent, les alimentations sont moins efficaces..etc.. toutes ces considérations peuvent remettre en cause les modèles théoriques énergétiques que l'on trouve dans la littérature. Ainsi, les gestionnaires de ressources[56] et les ordonnanceurs (section 5) de tâches doivent en tenir compte pour proposer des solutions réellement efficaces en consommation énergétique.

L'approche que nous proposons et qui est mise à disposition des utilisateurs de la plate-forme nationale Grid5000 permet d'envisager de nouveaux domaines tels que le *Green Programming* (analyse et optimisation de logiciels avec un objectif de réduction énergétique). La mesure de la

consommation énergétique à grande échelle demeure encore un domaine largement ouvert pour la recherche et le développement auquel s'intéressent les constructeurs de grandes infrastructures distribuées. Je souhaite poursuivre dans ce domaine dans les années futures en favorisant les collaborations entre académie et industrie. Nos travaux sur la mesure énergétique sont focalisés sur les ressources informatiques : serveurs de calcul, baies de stockage, équipements réseaux. Une autre perspective est de corréler et combiner ces mesures avec des mesures sur l'infrastructure matérielle complète d'une salle machine : climatisation, groupes électrogènes de secours, éclairage...

L'étape préliminaire et indispensable de la mesure énergétique étant franchie, il faut donc maintenant proposer de nouveaux *frameworks* logiciels adossés aux infrastructures à grande échelle et capables de réduire le consommation énergétique mais tout en gardant la même qualité de service, d'expérimentation et d'usage.

Professeur Thibault et Mlle Hortense, Les Nuls

5

De nouveaux composants logiciels pour gérer les ressources dans les infrastructures distribuées à grande échelle : ordonnanceurs et nuages verts

Un ordonnanceur dans un système distribué à grande échelle est un composant logiciel responsable du placement des tâches de travail en sélectionnant un ensemble de ressources (physiques ou virtuelles) appropriées dans le temps et répondant aux contraintes exprimées. L'ordonnanceur est une pièce maîtresse des infrastructures distribuées à grande échelle et un nombre important de recherches académiques portent sur l'optimisation de ce composant pour répondre à des besoins de performances, d'équilibrage de charge, de qualité de service et, plus récemment, de consommation énergétique.

Les travaux présentés dans ce chapitre ont eu lieu dans le cadre de la thèse de Anne-Cécile Orgerie (co-encadrée avec Isabelle Guérin Lassous). Nous avons proposé une nouvelle architecture de gestionnaire de ressources appelée ERIDIS (*Energy-efficient Reservation Infrastructure for large-scale DIstributed Systems*) dont le but est de lutter contre le sur-dimensionnement en nombre de ressources des infrastructures distribuées à grande échelle. ERIDIS propose un gestionnaire de ressources et un ordonnanceur basés sur la réservation de ressources dans un système distribué. ERIDIS conseille les utilisateurs dans leurs choix de placement de réservations afin de les encourager à favoriser l'agrégation de réservations dans le temps et dans l'espace.

Le modèle ERIDIS a été instancié et validé dans un contexte de Grilles et de *datacenters* avec la proposition EARI (*Energy Aware Reservation Infrastructure*) qui permet d'ordonnancer les réservations (section 5.2). Ces travaux ont été supportés par l'Action de Recherche Collaborative INRIA Green-Net, que j'ai dirigée entre 2008 et 2010 et qui a permis de structurer nos recherches et de donner l'impulsion (financière et scientifique) nécessaire à notre équipe (et à d'autres) pour se lancer sur ce sujet.

Nous avons aussi exploré les mécanismes de réduction énergétique dans les infrastructures distribuées à grande échelle de type Cloud. Reposant sur la modélisation ERIDIS, la proposition *Green Open Cloud* (section 5.3.1) autorise une surveillance fine de la consommation énergétique des infrastructures physiques et virtuelles au sein du Cloud afin de supporter de nouvelles fonctionnalités : ordonnancement éco-efficace, facturation à l'usage (énergétique), gestion des ressources... Ces travaux sont validés par différents développements logiciels menés

avec une petite équipe d'ingénieurs (Julien Carpentier, Maxime, Morel, Olivier Mornard et François Rossigneux) dans des projets de grande envergue FUI CompatibleOne (section 5.3.2) et FSN XLCLOUD (section 5.3.3) pour lesquels nous validons nos propositions dans des contextes académiques et industriels.

5.1 ERIDIS : une infrastructure de réservation efficace en consommation énergétique pour les systèmes distribués à grande échelle

Les mécanismes de réservation en avance sont largement utilisés dans les systèmes distribués à grande échelle [156, 41, 143] car ils garantissent aux utilisateurs une certaine qualité de service, en incluant les délais et les contraintes matérielles et logicielles. Une réservation est un triplet comprenant une durée, un *deadline* et les besoins logiciels et matériels. La durée de chaque réservation étant connue lors de la soumission, cela permet une gestion des ressources plus souple et prévisible. Cette approche rend ainsi la tâche des ordonnanceurs de réservations et de *jobs* plus facile. C'est cette spécificité que nous exploitons afin de proposer un environnement de réservation efficace en consommation énergétique adapté aux exigences des infrastructures distribuées à grande échelle (Datacenters, Grilles, Clouds et Réseaux) : le modèle ERIDIS (*Energy-efficient Reservation Infrastructure for large-scale DIstributed Systems*) [140].

L'objectif principal de ERIDIS est de lutter contre le sur-dimensionnement des infrastructures distribuées à grande échelle. De nombreuses ressources (calcul, stockage, réseaux) sont disponibles dans ces infrastructures afin de satisfaire des requêtes futures ou hypothétiques. ERIDIS propose de dimensionner finement le nombre de ressources alimentées afin de réduire la consommation énergétique tout en garantissant la même qualité de fonctionnement aux utilisateurs et leurs applications. Pour atteindre ce but, ERIDIS favorise au maximum l'agrégation sur deux plans :

- dans le temps : les réservations sont groupées les unes à la suite des autres afin de favoriser l'apparition de périodes creuses propices à l'extinction de ressources;
- dans l'espace : des réservations sont agrégées sur un sous ensemble de ressources physiques afin de limiter le nombre de ressources alimentées à un instant donné.

ERIDIS est une approche au "niveau gestionnaire de ressources". Elle met en œuvre des techniques d'extinction et d'allumage de ressources couplées à des modules de prédiction d'usage.

Le gestionnaire ERIDIS est constitué d'un ensemble de composants logiciels (figure 5.1) : un contrôleur d'admission de réservations (section 5.1.2) qui dialogue avec l'utilisateur, un ordonnanceur de réservations qui place les réservations sur les ressources physiques, des modules de prédiction afin d'anticiper l'usage des ressources (section 5.1.4), un gestionnaire de ressources qui gère l'allumage et l'extinction des infrastructures (section 5.1.3), un gestionnaire de politiques qui s'assure que les décisions prises sont bien en conformité avec les choix des utilisateurs et des administrateurs.

5.1.1 Le modèle de réservation

Chaque gestionnaire ERIDIS maintient un agenda (figure 5.1) qui contient l'ensemble des réservations futures (et passées) sur une ressource donnée. L'agenda contient aussi l'état de la ressource : réservée, libre, allumée, en mode veille, extinction, démarrage, arrêt... Un exemple d'agenda est illustré figure 5.2. Cet agenda contient deux réservations qui n'utilisent pas la totalité de la ressource. La capacité maximum peut être en nombre de cœurs pour un serveur, en bande passante pour un lien réseau ou en espace de stockage pour un disque. L'agenda permet de prévoir la prochaine extinction de la ressource entre les deux réservations ainsi que d'anticiper l'heure à laquelle la ressource devra être rallumée afin d'être disponible à temps.

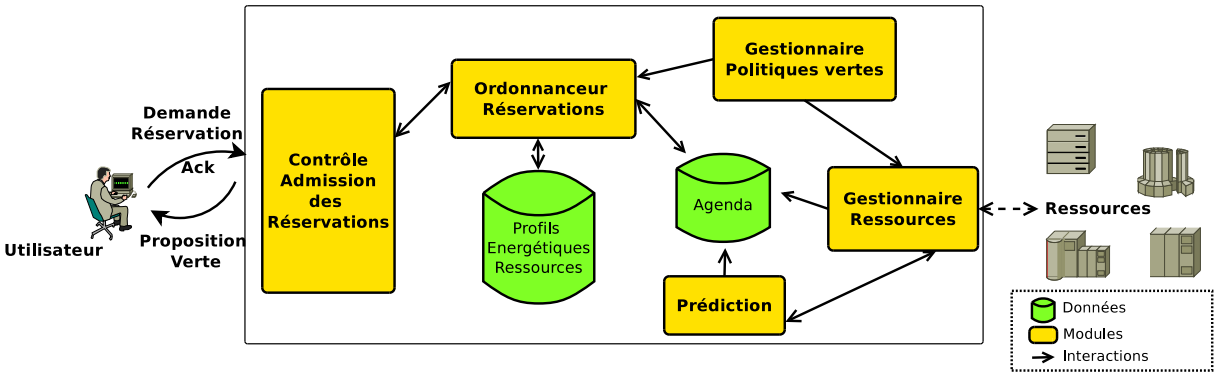


Figure 5.1: Architecture du gestionnaire ERIDIS

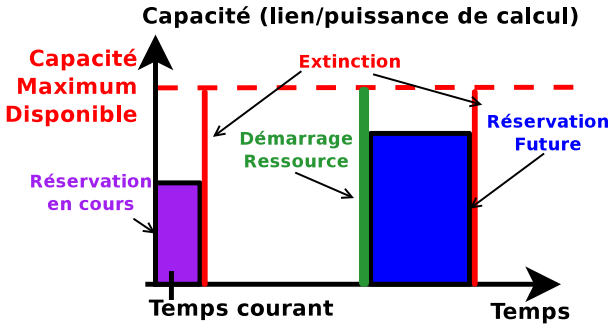


Figure 5.2: Agenda d'une ressource

5.1.2 Gestion des réservations

ERIDIS met en œuvre un ensemble d'algorithmes d'ordonnancement des réservations afin de répondre aux besoins d'efficacité énergétique. Mais ERIDIS est aussi un système en lien avec l'utilisateur et qui permet de l'impliquer dans une prise de décision consciente de l'usage des ressources. Le gestionnaire d'ERIDIS agit comme un contrôleur d'admission et vérifie la validité des requêtes de réservations de l'utilisateur (figure 5.1) :

- en fonction des politiques d'efficacité énergétique ou de performances appliquées par l'administrateur de l'infrastructure distribuée à grande échelle;
- en fonction des politiques vertes choisies par l'utilisateur (qui peut accepter de déplacer sa requête dans le temps et dans l'espace).

5.1.3 Attention avant d'éteindre les ressources inutiles !

Le modèle ERIDIS propose d'éteindre les ressources inutilisées entre des réservations agrégées dans le temps et dans l'espace. Afin de prendre des décisions d'allumage et d'extinction, nous devons prédire la prochaine réservation pour ne pas désactiver des ressources qui vont être utilisées dans un futur proche. Effectivement, ceci consommerait plus d'énergie que si on avait laissé les nœuds allumés. On a observé qu'allumer une machine provoque un pic de consommation électrique dont la durée et l'amplitude ne peuvent être négligées (figure 4.11).

C'est pourquoi on cherche le temps minimum (appelé T_s) tel que l'on gagne de l'énergie à éteindre le nœud plutôt qu'à le laisser allumé [135, 136]. Cette définition est illustrée par la figure 5.3. La courbe du haut montre l'énergie consommée si on éteint le nœud et la courbe du bas celle consommée si on laisse le nœud allumé en mode `idle`.

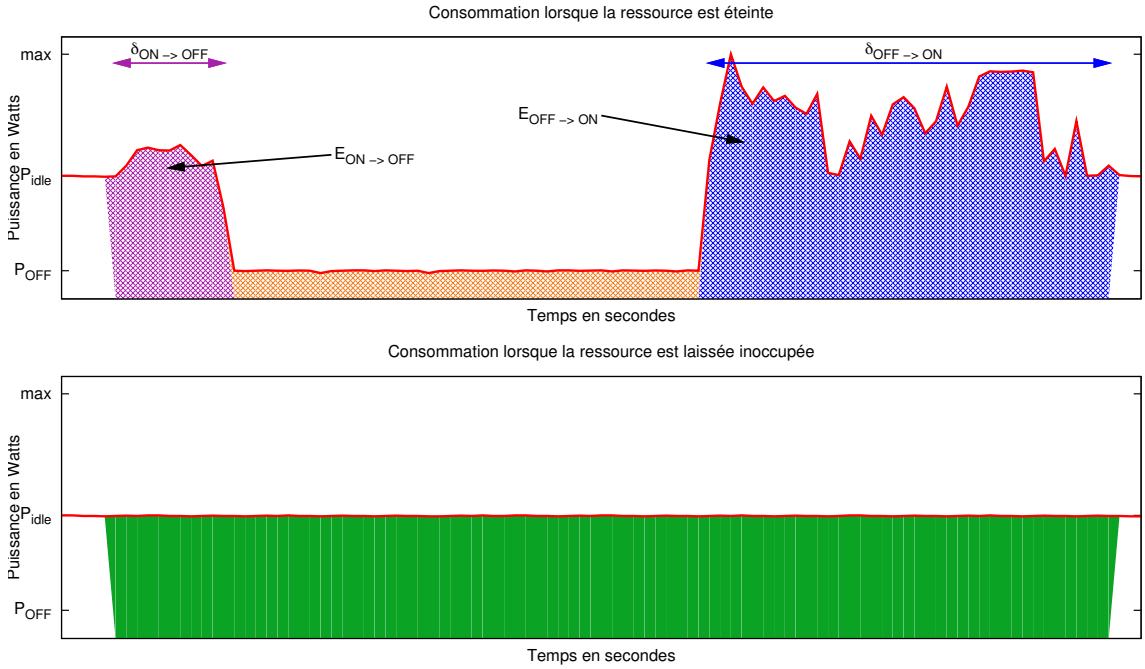


Figure 5.3: Profil d'un serveur qui s'éteint ou reste allumé en mode `idle`

5.1.4 Prédire l'usage des infrastructures distribuées à grande échelle

Si on pouvait prédire parfaitement l'usage des systèmes distribués à grande échelle, le dimensionnement dynamique du nombre de ressources nécessaires pour répondre à la demande serait simple. Mais l'usage d'une infrastructure est constitué de pics et de creux qui sont variables dans le temps. Par exemple, les infrastructures de calcul haute performance font souvent face à des usages soutenus mais on peut néanmoins observer des périodes d'inactivité entre chaque lancement de campagne de calcul ou lors d'opérations d'entrées sorties et de transfert de données. Les *datacenters* de *clouds* peuvent faire face à des usages plus variables avec des périodes de faibles activités qui peuvent être significatives.

À la fin de chaque réservation, lorsque les ressources sont libérées, on exécute des algorithmes de prédiction pour savoir quand va survenir la prochaine réservation. Si cette réservation prédite est imminente, on laisse le nombre nécessaire de ressources allumées (nombre également prédit). Sinon, on éteint les machines.

ERIDIS a permis la mise en place d'une collaboration avec l'université de Séville, nous avons accueilli pour quelques semaines Alejandro Fernandes Montes (en thèse de doctorat) pour étudier de nouveaux modèles de prédictions d'usage [68].

ERIDIS a été adapté sous trois formes en fonction des contraintes retenues pour les différents types d'infrastructures distribuées à grande échelle : **EARI (*Energy Aware Reservation Infrastructure*)** pour ordonnancer les tâches de calcul dans les Grilles, **GOC (*Green Open Cloud*)** pour l'efficacité énergétique dans les Clouds et **HERMES (*High-level Energy-aware Reservation Model for End-to-end networkS*)** pour la réservation de bande passante et l'orchestration de solutions eco-efficaces dans les réseaux de grande taille (non présenté dans cette habilitation, se reporter à [141, 139]).

5.2 EARI un ordonnanceur de réservations pour les centres de données et les Grilles à la recherche d'usage en dents de scie

Le modèle ERIDIS est le composant de base des travaux de la thèse de Anne-Cécile Orgerie [131]. Sur ce modèle, l'instance EARI (*Energy Aware Reservation Infrastructure*), adaptée aux Grilles de calcul et *data centers*, a été proposée.

5.2.1 L'architecture d'EARI

L'architecture d'EARI repose sur trois idées principales, héritées du modèles ERIDIS :

- éteindre (mettre en veille) les nœuds de calcul inutilisés, puis allumer (réveiller) les machines nécessaires ;
- prédire les réservations pour ne pas éteindre des ressources qui pourraient servir très peu de temps après avoir été libérées ;
- favoriser des agrégations de réservations dans le temps et dans l'espace pour éviter des cycles d'allumage/extinction trop fréquents et lutter contre le sur-dimensionnent en nombre de ressources allumées.

Notre architecture est illustrée par la figure 5.4 avec une présentation des composants et de leurs interactions. Elle met en œuvre un portail qui permet le dialogue avec les utilisateurs, les système de gestion de ressources et un ensemble de capteurs de consommation électrique qui remontent leurs mesures au gestionnaire EARI [136].

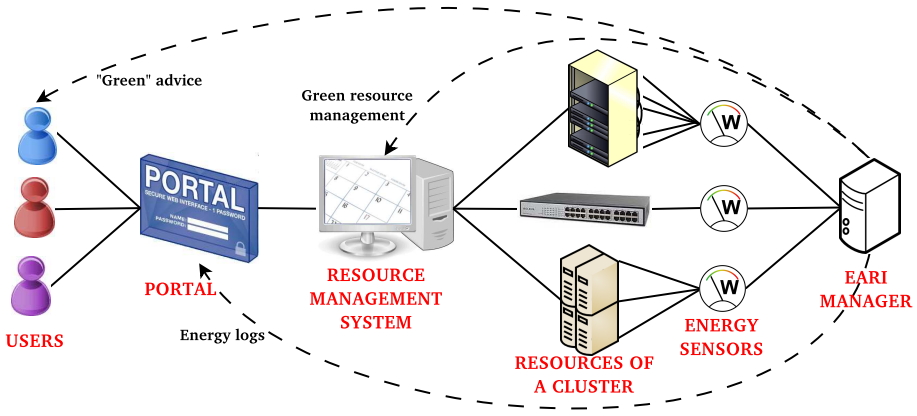


Figure 5.4: Architecture globale de l'infrastructure EARI

Les appareils de mesure de consommation collectent en temps réel la consommation électrique des nœuds, et mettent ces données à disposition des utilisateurs *via* le portail. De plus, l'infrastructure EARI donne des conseils pour sensibiliser les utilisateurs afin qu'ils placent des réservations dans l'agenda pour consommer moins d'énergie. Finalement, EARI indique au gestionnaire de ressources quand celui-ci doit éteindre ou allumer des nœuds.

5.2.2 Exploiter les leviers verts : algorithmes d'allumage et d'extinction de machines

EARI repose sur des algorithmes de prédiction pour anticiper les réservations imminentes et ainsi consommer moins d'énergie. Les prédictions concernent :

- la prochaine réservation (durée, nombre de ressources et date de début) pour savoir s'il faut éteindre ou non des ressources libérées à la fin d'une réservation (voir figure 5.2);

- la date de début de la prochaine période creuse qui est proposée comme solution lors de la soumission d'une réservation.

Nos algorithmes de prédictions sont basés sur des valeurs moyennes (temps entre deux réservations, nombre de ressources par réservation, durée d'une réservation) calculées sur un historique des dernières réservations. Par exemple, l'algorithme de prédiction de la prochaine réservation calcule une moyenne des temps entre les dates de débuts des réservations ayant eu lieu juste avant d'appeler cet algorithme. Ainsi, dans la section 5.2.3, pour valider nos algorithmes nous avons utilisé les temps entre les dates de début des six dernières réservations. Ce nombre de réservations prises en compte peut sembler faible, mais sur les exemples de traces fournis par Grid'5000, nous avons constaté qu'augmenter ce nombre n'améliore pas le pourcentage de bonnes réponses de notre algorithme. En effet, notre algorithme de prédiction de la prochaine réservation prend la bonne décision (de maintenir une ressource allumée ou de l'éteindre à la fin d'une réservation) dans 70 % des cas en moyenne sur l'ensemble des traces de Grid'5000 étudiées [135]. Ceci conduit à des économies d'énergies non négligeables comparé à la solution qui consisterait à laisser les ressources allumées un laps de temps fixé après la fin d'une réservation [135].

5.2.3 Validation expérimentale d'EARI

Afin d'évaluer EARI, nous avons utilisé les traces d'usage de Grid'5000 en les calibrant avec les mesures de consommation électrique observées. Nous avons simulé des *replays* de ces traces en appliquant différentes politiques énergétiques pour modéliser les différents comportements possibles. Nous pouvons ainsi estimer les gains en termes de réduction de consommation énergétique apportés par EARI.

Pour mener à bien nos simulations, nous proposons six politiques énergétiques :

- *user* : on satisfait toujours la demande de l'utilisateur, c'est-à-dire que l'on place sa réservation à la date qu'il demande si celle-ci est possible (ressources suffisantes) ou à la date possible la plus proche;
- *fully-green* : on sélectionne la solution qui consomme le moins d'énergie (tout en nécessitant le moins d'allumages et d'extinctions de machines) parmi les choix proposés par EARI;
- *25%-green* : pour 25 % des soumissions prises au hasard, on applique la politique *fully-green* et pour le reste on applique la politique *user*;
- *50%-green* : pour 50 % des soumissions prises au hasard, on applique la politique *fully-green* et pour le reste on applique la politique *user*;
- *75%-green* : pour 75 % des soumissions prises au hasard, on applique la politique *fully-green* et pour le reste on applique la politique *user*;
- *deadlined* : on utilise la politique *fully-green* si cela ne retarde pas la réservation de plus de 24 heures par rapport à la demande de l'utilisateur, sinon on utilise la politique *user*.

Ces politiques simulent le comportement des utilisateurs : certains sont plus enclins à décaler leurs réservations (sans borne maximum pour les politiques **-green* et avec 24h de délai maximum pour la politique *deadlined*) si cela permet d'économiser de l'énergie.

Nous avons donc rejoué les traces d'utilisation de Grid'5000 sur les différents sites de la plate-forme. Nous présentons ici les mesures de Bordeaux et Lyon. La figure 5.5 présente la consommation énergétique en utilisant EARI comparée à la consommation actuelle, c'est-à-dire lorsque tous les nœuds restent allumés en permanence même lorsqu'ils sont inactifs (cette consommation correspond à 100% sur la figure). Des mesures électriques nous permettent le connaître la puissance électrique consommée par les machines de Lyon (190 Watts en moyenne).

On observe une diminution forte de la consommation statique des serveurs de calculs, nous présentons donc les résultats pour trois consommations différentes : P_{idle} = 100, 145 et 190 Watts.

Le graphe de la figure 5.5 présente donc pour trois P_{idle} différents les résultats obtenus en utilisant nos 6 politiques introduites dans la section précédente et un T_s (temps minimum avant extinction) initialisé à 240 secondes. La borne inférieure idéale est représentée par le seuil "all glued" (figure 5.5). Elle représente le cas optimal (in-atteignable) en consommation électrique où l'on pourrait coller toutes les réservations de l'année les unes à la suite des autres.

La première constatation est que la politique d'allumage et d'extinction proposée par EARI apporte une réduction énergétique conséquente. 25% à 40% de l'énergie peut être économisée sur le site de Lyon.

Dans le cas actuel (P_{idle} = 190 Watts), nous voyons sur la figure 5.5 que le gain d'énergie réalisé en utilisant la politique *fully-green* à la place de la politique *user* (politique actuelle) passe de 75 à 73% soit un gain supplémentaire de 2 %, qui représente 3 300 kWh[1] dans le cas de Lyon. La politique *fully-green* permet des réductions énergétiques très proches des valeurs optimales ("*all glued*").

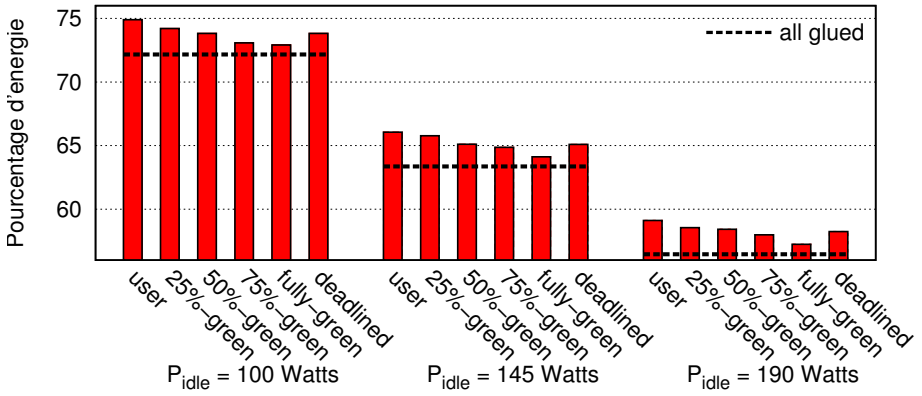


Figure 5.5: Pourcentage d'énergie consommée pour Lyon avec EARI et T_s = 240 s

Nous observons que pour le cas du site de Lyon, la politique *fully-green* décale 99 % des réservations (par rapport au nombre total de réservations). C'est normal puisque notre politique ne décale pas seulement les réservations qui peuvent être placées avant ou après une autre, mais aussi les réservations qui ne peuvent plus avoir lieu au moment demandé du fait d'autres déplacements. Nous observons que les réservations ne sont pas retardées de plus de 15 heures en moyenne mais que cela permet d'obtenir un gain en énergie important.

Nous avons analysé l'impact de T_s sur la consommation électrique. Nous pouvons augmenter T_s pour accroître la réactivité : les ressources sont alors plus longtemps allumées après une réservation et donc peuvent répondre immédiatement à une requête surprise (non prédite) (on économise le temps d'un démarrage des machines). Ces expérimentations sont illustrées sur la figure 5.6 pour Bordeaux. Nous avons fixé P_{idle} à 100 Watts et nous avons fait varier T_s entre 120 et 420 secondes par incrément de 60. Nous constatons sur la figure 5.6 que les consommations des six politiques restent du même ordre. De plus, nous observons que T_s n'a pas un grand impact sur la consommation globale, une grande valeur de T_s peut en effet compenser d'éventuelles erreurs de prédiction.

On a vu que la politique *user* est toujours celle qui consomme le plus. Ainsi ces résultats montrent qu'EARI mène à des économies d'énergie significatives pour les infrastructures distribuées à grande échelle. Pour le site de Lyon, en 2007, l'économie est de 73 800 kWh (aux alentours de 10 000 euros au tarif EDF pour les particuliers) sur l'énergie consommée uniquement par

[1]La consommation annuelle d'électricité d'un ménage moyen français (hors chauffage, eau chaude et cuisson) est d'environ 3 000 kWh selon l'ADEME (`http://www.ademe.fr/particuliers/fiches/reseau/rub2.htm`).

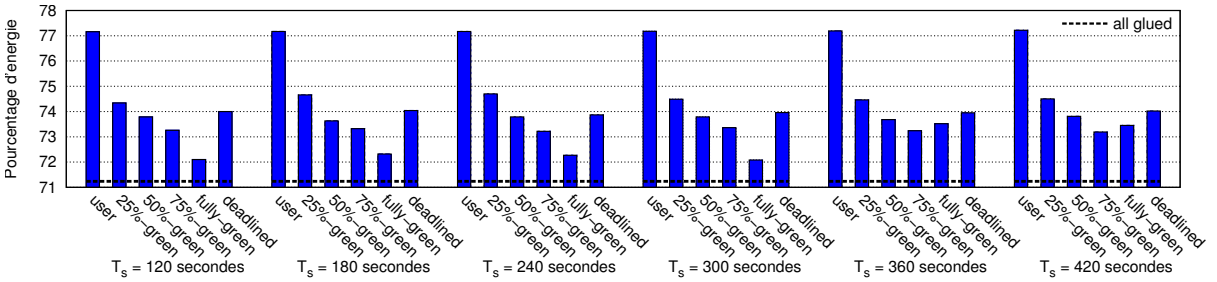


Figure 5.6: Pourcentage d'énergie consommée pour le site Grid'5000 de Bordeaux avec EARI et $P_{idle} = 100$ W

les nœuds eux-mêmes (sans prendre en compte la climatisation et les équipements réseaux) en utilisant EARI et la politique *fully-green*. Pour l'ensemble des nœuds de la plate-forme en 2007, les économies d'énergies qui auraient été réalisées en utilisant EARI et la politique *fully-green* correspondent à 52 % de la consommation actuelle.

5.3 Efficacité énergétique et Cloud

Les Grilles de calcul avaient beaucoup promis : la gestion de l'hétérogénéité des infrastructures, une facilité de développement, le support de nouvelles applications. Mais la complexité de déploiement et d'usage des grilles a eu raison de ce modèle tombé en désuétude au profit du Cloud.

Le *Cloud computing* (Informatique en nuage, informatique dématérialisée) est l'accès via le réseau, à la demande et en libre-service, à des ressources informatiques partagées configurables (définition du NIST[2]). La paternité du Cloud Computing revient à John Mac Carty qui dès 1961 évoquait ce modèle d'*utlity computing"* (offre d'un mélange de ressources de calculs, stockage et réseau comme un service tarifé) et son futur impact dans l'industrie.[3]

Amazon[4] a été une des premières entreprises à proposer des solutions de cloud pour les particuliers. Malgré les problèmes de sécurité" et de confidentialité, l'usage et le stockage dans le Cloud deviennent une réalité avec de nombreux services disponibles (dropbox[59], justcloud[93], sugarsync[158]). Le gouvernement français a lancé le projet Andromède pour l'obtention d'un cloud sécurisé pour l'administration française. Deux projets de 75 millions d'euros chacun sont supportés par cette initiative : Numergy[5] et Cloudwatt[6].

Une partie de la communauté scientifique s'est engouffrée dans le domaine du Cloud pour les challenges qu'il représente en termes de dimensionnement, sécurité, virtualisation, maîtrise de performance, conception architecturale... La virtualisation au sein d'un Cloud permet la mise en œuvre de politiques de consolidation qui peuvent être favorables en termes de consommation énergétique. On assiste ainsi depuis quelques années à un certains nombre de travaux sur l'utilisation de technique de virtualisation pour réduire la consommation énergétique des appli-

[2] *Five myths of Cloud Computing - White Paper HP* : `http://www.hp.com/hpinfo/newsroom/press_kits/2011/HPDiscover2011/DISCOVER_5_Myths_of_Cloud_Computing.pdf`

[3] "*If computers of the kind I have advocated become the computers of the future, then computing may someday be organized as a public utility just as the telephone system is a public utility....The computer utility could become the basis of a new and important industry*" : John Mac Carty - discours MIT Centennial, 1961 - Architects of the Information Society, Thirty-Five Years of the Laboratory for Computer Science at MIT, H. Abelso ed.

[4] Amazon Web Services : `http://aws.amazon.com/fr/`

[5] Numergy : `https://www.numergy.com/`

[6] Cloudwatt : `https://www.cloudwatt.com/`

cations et des services dans le cloud [19]. De nombreux chercheurs se retrouvent dans cette vague appelée "Cloud Vert" afin d'appliquer leurs modèles et optimisations sur d'autres scénarios et contextes.

Nous avons été parmi les premiers à surfer sur la vague du cloud vert : avec Anne-Cécile Orgerie (thèse co-encadrée avec Isabelle Guérin Lassous), nous avons proposé une adaptation du modèle ERIDIS qui a débouché sur l'architecture *Green Open Cloud*.

Nous poursuivons ces travaux en relation avec des partenaires industriels et académiques dans le cadre des projets CompatibleOne et XLCLOUD. CompatibleOne nous permet d'architecturer et de mettre en place des infrastructures de collecte d'information énergétique au sein d'un *broker* de *cloud open source*. Dans le cadre du projet XLCLOUD, nous poursuivons cette approche en injectant de l'efficacité énergétique au sein d'infrastructures de clouds dédiées à l'hébergement d'applications et de services de calcul hautes performances.

5.3.1 La proposition Green Open Cloud

Si le Cloud vert et le placement efficace en énergie de machines virtuelles sont des sujets de recherche à la mode en 2013; dès 2009, en nous fondant sur les travaux menés avec Anne-Cécile Orgerie sur l'ordonnancement vert dans les infrastructures distribuées à grande échelle (ERIDIS), nous avons mis à disposition une des premières propositions de nuage vert appelée *Green Open Cloud* [133, 106]

En généralisant l'architecture ERIDIS (section 5.1) à une infrastructure dynamique et virtualisée de type Cloud, nous définissons un environnement énergie-sensible adapté aux Clouds. L'approche GOC (*Green Open Cloud*) repose sur le concept de "garantir la même qualité d'usage mais avec moins d'énergie" en supportant le re-dimensionnement des ressources physiques (nombre de serveurs alimentés), des modèles de prédiction, d'équilibrage et le placement optimisé de machines virtuelles.

L'architecture Green Open Cloud propose l'infrastructure matérielle et logicielle suivante (figure 5.7) :

- des capteurs de consommation électriques (type wattmètre) fournissent des mesures précises et fréquentes de la puissance électrique consommée par les ressources physiques du cloud (serveurs, baies de stockage, équipements réseaux);
- un collecteur de traces d'énergie qui récupère et met en forme l'ensemble des informations de consommation électrique de la plate-forme;
- un *proxy* de confiance qui répond à la place des nœuds éteints du cloud;
- un gestionnaire de ressource et ordonnanceur éco-sensible

L'architecture GOC supporte des facilités d'allumage et d'extinction des ressources physiques (calcul, réseau et stockage). Elle dispose de modules de prédiction de l'usage de l'infrastructure afin d'anticiper l'allumage des ressources qui risquent d'être utilisées dans un futur proche. GOC utilise des facilités d'agrégation de tâches de calcul en déployant des solutions de migrations de machines virtuelles. Des politiques vertes permettent à l'utilisateur d'exprimer ses contraintes en termes d'efficacité énergétique. Les gestionnaire de ressources GOC (figure 5.8) assure ces fonctionnalités.

Lorsque des nœuds du Cloud sont éteints, ils ne répondent plus aux gestionnaire de ressources qui peut alors les considérer comme défaillants. GOC déploie un proxy de délégation de confiance qui assure la présence réseaux des nœuds éteints. Avant extinction d'un nœud, GOC envoie un ensemble de services de base (service *heartbeat*, réponse au *ping*) dans une machine virtuelle qui se déploie sur le *proxy* [48, 129, 169]. Cette solution résout aussi des problèmes de sécurité

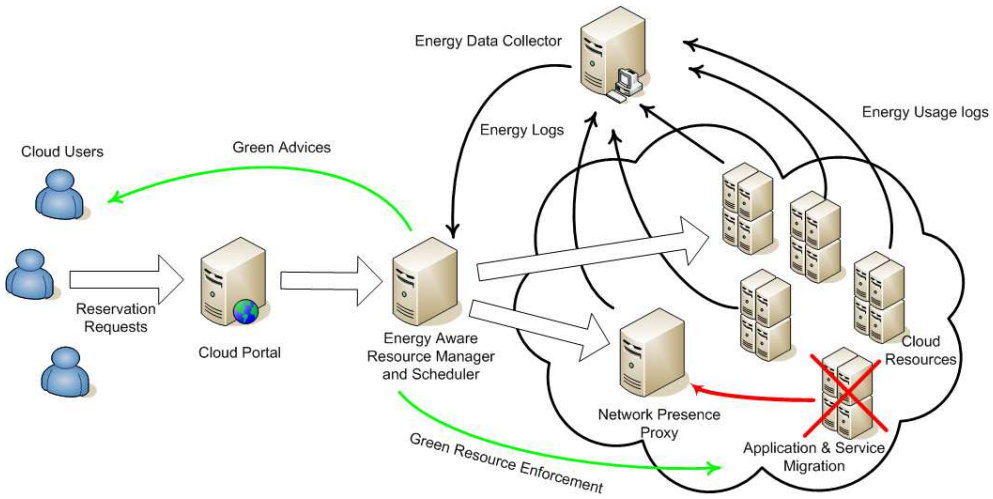


Figure 5.7: Infrastructure de Nuage Vert (GOC)

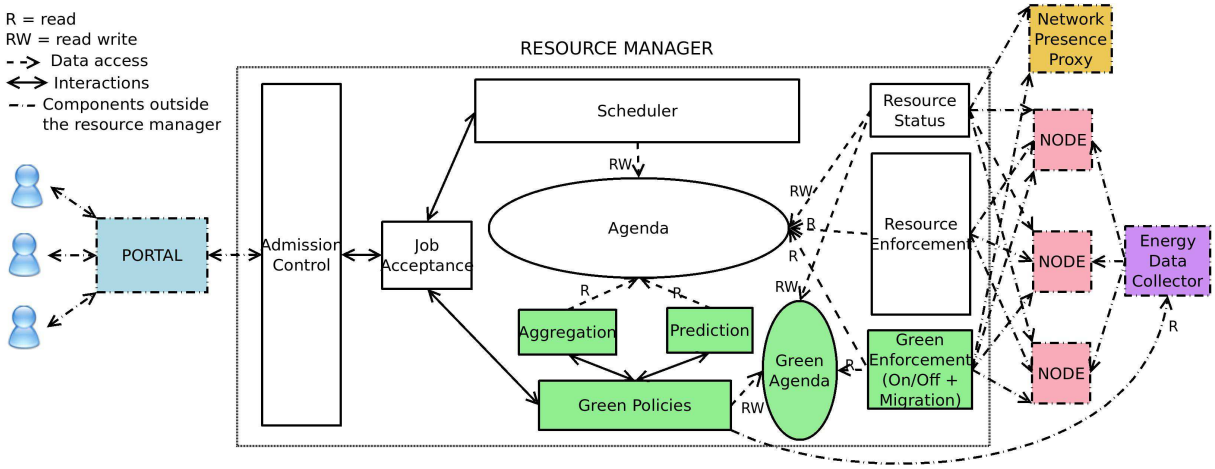


Figure 5.8: Architecture du gestionnaire de ressources du Green Open Cloud

en évitant le réveil de nœuds qui pourraient se faire passer pour des ressources actuellement éteintes.

Afin d'illustrer le fonctionnement de GOC nous présentons 3 exemples de réduction énergétique avec GOC sur une mini infrastructure de 2 nœuds de cloud (type HP Proliant 85 G2 Servers (2.2 GHz, 2 dual core CPUs par node avec XenServer 5). Pour plus d'expériences et validations, le lecteur pourra se reporter à [106].

Chaque ressource est capable d'héberger 7 machines virtuelles (machine à 8 cœurs). Les machines virtuelles hébergent une tâche de calcul intensif (simulés par `cpuburn`). L'arrivée des tâches de calcul est la suivante;

- $t = 10$: 3 jobs de 120 secondes et 3 jobs de 20 s chacun;;
- $t = 130$: 1 job de 180 s;
- $t = 310$: 8 jobs de 60 s chacun;
- $t = 370$: dans l'ordre : 5 jobs de 120 s , 3 jobs de 20 s et 1 job de 120 s.

Les scénarios et profils énergétiques associés sont :

- Déploiement de machines virtuelles sur Cloud sans intervention de GOC avec ordonnancement circulaire (*round robin*) : Les tâches de calcul sont déployées sur les 2 nœuds les uns après les autres. La figure 5.10 présente le profil énergétique des deux nœuds du Cloud (watts). On peut observer qu'une machine virtuelle avec `cpuburn` coûte une dizaine de watts. Pendant la période 300-400 secondes, on peut observer que la 5 ème tâche ne provoque pas de puissance consommée supplémentaire car la machine Cloud2 a atteint sa consommation maximale.

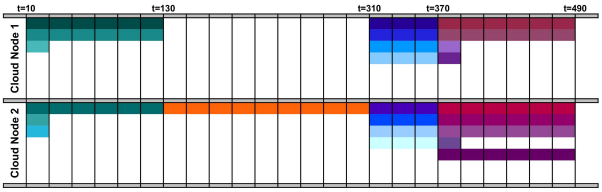


Figure 5.9: Gantt de déploiement des tâches avec ordonnancement circulaire (scenario basique)

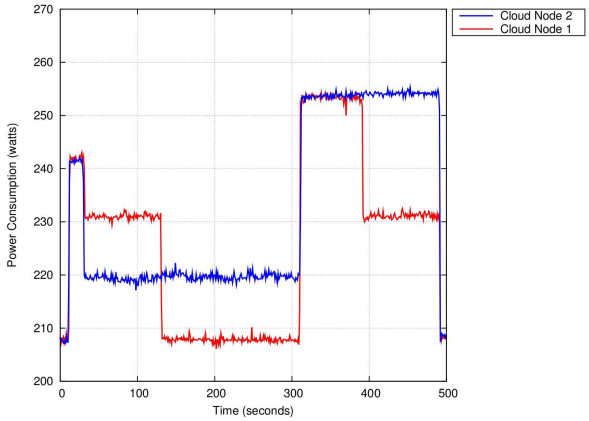


Figure 5.10: Profil énergétique associé

- Déploiement de machines virtuelles sur Cloud avec GOC (scénarion Green) : Ce scénario respecte l'ordonnancement circulaire pour les petites tâches (sec 10-20). Par contre, pour les tâches longues telles que le *job* commençant à t=30, ceux-ci sont ré-alloués sur le nœuds allumés afin de créer une condition propice à l'extinction des nœuds inutilisés. Ainsi le nœud Cloud 2 est éteint de t=40 à t=200. On observe le pic de puissance due à la migration de la machine virtuelle au temps t=35 et t=395. au temps, t=200 on peut observer le pic de consommation résultant de l'allumage de la machine Cloud 2 par le module de prédiction de GOC afin d'anticiper l'arrivée de jobs au temps t=310. Ce pic de consommation à l'allumage (ventilateurs, tests mémoire...) est compensé par le temps passé en mode veille (avec une puissance consommée de 20 Watts).

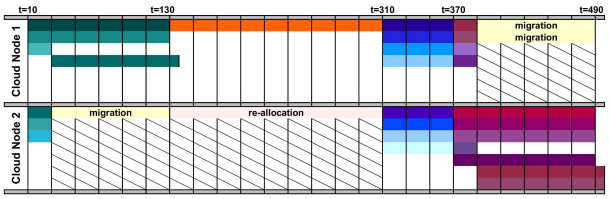


Figure 5.11: Gantt de déploiement des tâches avec ordonnancement circulaire (scenario Green)

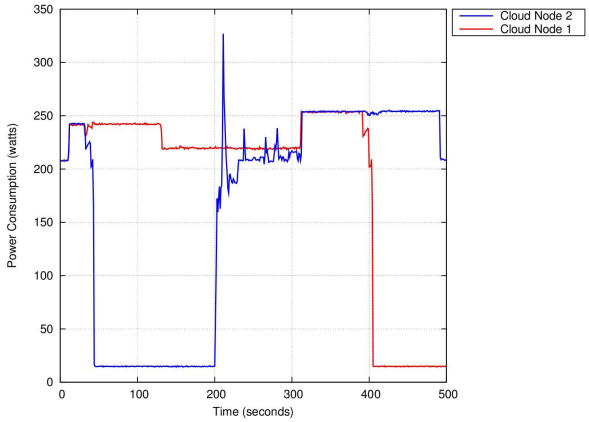


Figure 5.12: Profil énergétique associé

- Déploiement de machines virtuelles sur Cloud avec GOC (scénario GreenD) avec ordonnancement glouton : les jobs sont placés sur les machines en utilisant au maximum les

ressources disponibles sur chaque nœud. Cette politique est la plus efficace en consommation énergétique car elle agrège au maximum les machines virtuelles sur un sous ensemble restreint de ressources physiques. Ainsi plus de ressources peuvent être éteintes ou celles-ci peuvent être placées en mode veille plus longtemps (Figure 5.14).

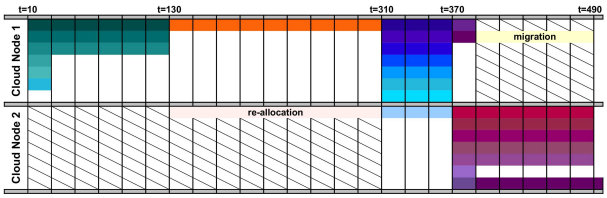


Figure 5.13: Gantt de déploiement des tâches avec dés-équilibrage (scenario GreenD)

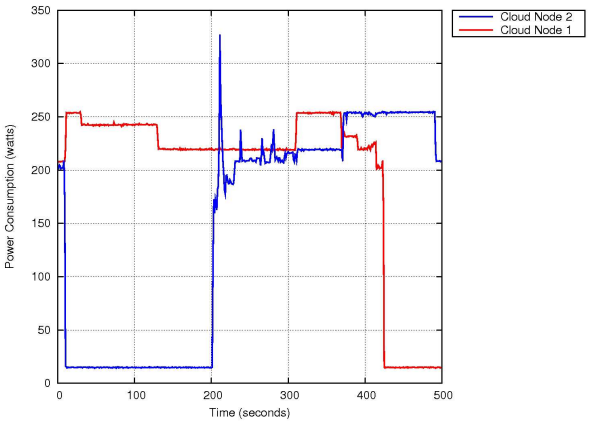


Figure 5.14: Profil énergétique associé

On peut ainsi observer que le scénario GreenD qui favorise l'agrégation des machines virtuelles, l'extinction des nœuds inutilisés et qui respecte un ordonnancement "glouton" est celui qui consomme le moins d'électricité. Le non-équilibrage de charge qui est une idée pourchassée depuis longtemps par la communauté HPC est finalement une bonne chose en termes d'efficacité énergétique. Bien sûr ce déséquilibre qui provoque une agrégation de machines virtuelles sur une même ressource physique peut avoir un impact négatif en termes de performances (due à la compétition des machines virtuelles pour une ressource partagée).

De plus, cet exemple illustratif devrait prendre en considération le coût de migration d'une machine virtuelle en utilisant un système comme Entropy [88]. Il faut aller plus loin que ces validations simplifiés et confronter le modèle GOC à des réalités de clouds plus opérationnelles. C'est cette direction que nous suivons ; le modèle de Green Open Cloud sert de composant de base à différents projets de recherche académiques ou collaborations industrielles auxquels je participe.

5.3.2 *Broker* de Nuage Vert

5.3.2.1 L'approche CompatibleOne

Dans le projet FUI CompatibleOne, nous étudions la conception d'un "broker" de cloud (qui dialogue avec le client et peut fédérer un ensemble de fournisseurs d'infrastructures Cloud) avec prise en compte de l'usage électrique [99]. Ces travaux sont menés avec l'aide d'une petite équipe d'ingénieurs que j'ai recrutés et dirigés dans le cadre du projet (Julien Carpentier, Maxime Morel et Olivier Mornard)

CompatibleOne[7] est un projet supporté par le Fonds Unique Interministériel et qui regroupe des partenaires industriels (Bull (Leader) Activeon, CityPassenger, Enovance, Eureka, Mandriva, Nexedi, Nuxeo, Prologue et Xwiki) et académiques (INRIA et l'Institut Telecom). Les participants développent et intègrent une implémentation complète des 3 services principaux du Cloud : IaaS (*Infrastructure as a Service*), PaaS (*Platform as a Service*), SaaS (*Software as a Service*). Le but principal du projet CompatibleOne est de proposer une architecture de *broker* de cloud open source autorisant l'interaction entre les systèmes de clouds existants.

[7] http://www.compatibleone.org/

Ainsi, à terme, CompatibleOne fournit la capacité de fédérer les ressources du Cloud entre différents fournisseurs de services. Cela résulte en une architecture complexe présentée figure 5.15 (plus d'informations [40]).

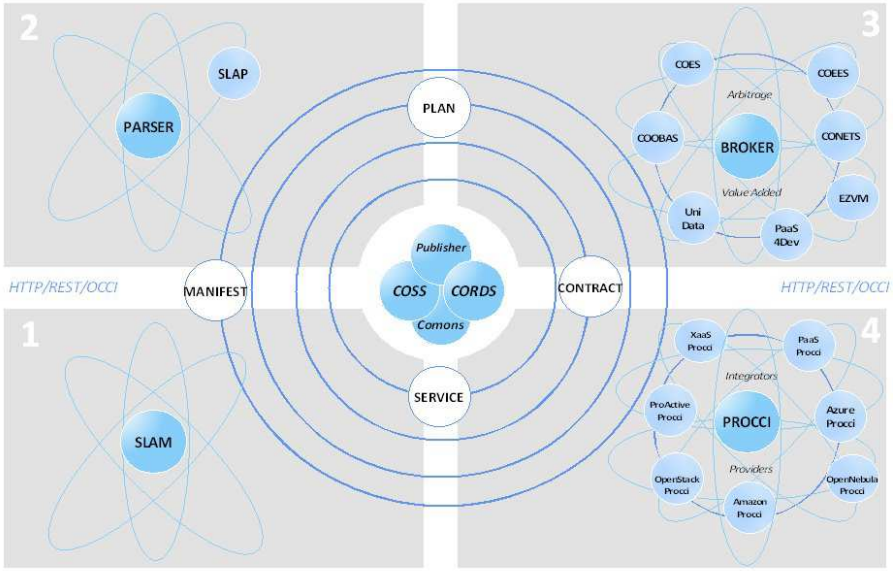


Figure 5.15: Architecture globale de CompatibleOne

5.3.2.2 Le module COEES : CompatibleOne Energy Efficiency Services

Dans ce projet nous avons proposé le module COEES (CompatibleOne Energy Efficiency Service), une instanciation du modèle Green Cloud adaptée aux contraintes du *broker*. Les services fournis par le module COEES agissent sur le cœur du système en proposant au *broker* d'intégrer des contraintes énergétiques. Ces contraintes régulent le placement et le déplacement des machines virtuelles en intégrant le point de vue de l'énergie. Des données énergétiques sont également proposées au reste du système afin de pouvoir être intégrées à différentes problématiques, comme par exemple la facturation. L'ensemble des données du module est injecté dans le système de monitoring (COMONS) de Compatible One.

Pour la collecte de ces informations, COEES repose sur un ensemble de sondes physiques (externes et internes) ainsi que des sondes logicielles.

Une partie des sondes externes de la plateforme GreenGrid5000 (section 4.2) est complétée par des sondes internes logicielles spécifiques à DELL (iDRAC [97] accédées par l'interface IPMI (Figures 5.16 et 5.17). Leur précision est faible (de l'ordre de 7 Watts) et la fréquence de mesure est limitée à 1 mesure de puissance moyennée toutes les 5 secondes. L'avantage est que ces équipements sont disponibles dans les serveurs des partenaires du projet.

Le module COEES est conçu pour surveiller à la fois la consommation électrique des ressources physiques du Cloud et la consommation des machines virtuelles (5.18).

Techniquement, nous avons adapté l'environnement de monitoring *ganglia* [121] afin de manipuler les flux de données énergétiques collectés dans le Cloud. Ganglia [121], est une solution logicielle libre de l'Université de Californie qui propose un système de *monitoring* distribué pour le calcul haute performance. Il est basé sur une hiérarchie de fédération de grappes de machines (figure 5.19) et utilise des technologies logicielles standard telles que XML pour la représentation de données et les représentations graphiques *Round Robin Database tool* (*RRDtool* identique à l'infrastructure ShowWatts (section 4.10).

Figure 5.16: Carte IPMI

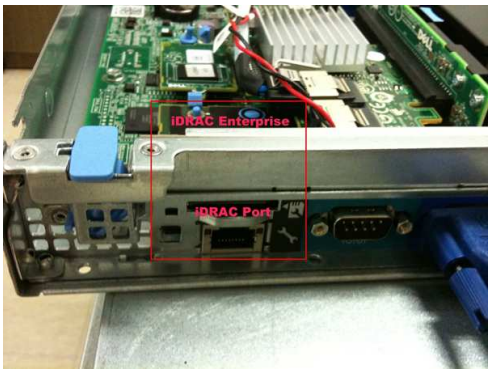


Figure 5.17: Déploiement d'un capteur avec IPMI dans un serveur

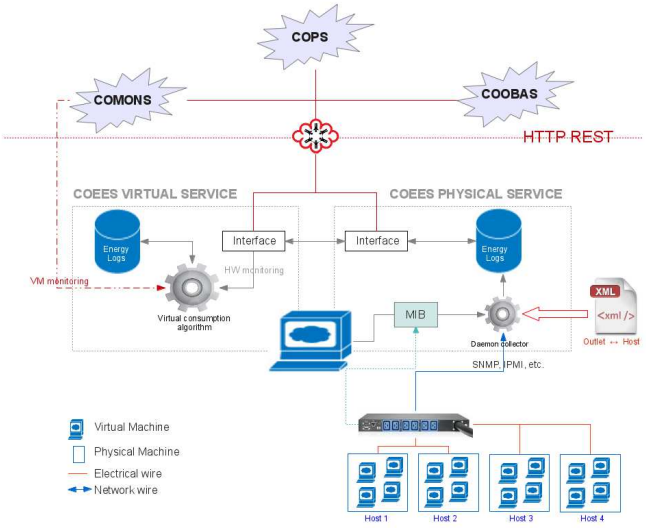


Figure 5.18: Architecture des modules de gestion énergétique de CompatibleOne

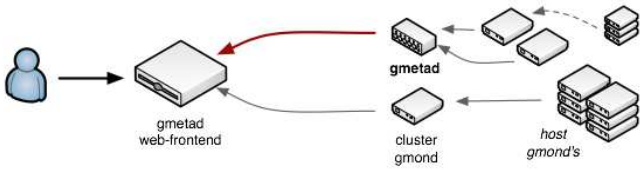


Figure 5.19: Ganglia Monitoring System architecture

Le modules COEES mesure les consommations électriques des ressources physiques du nuage avec les infrastructures de mesure électrique disponibles. Dans la figure 5.20 on peut ainsi comparer la différence de précision lors d'une campagne de mesures entre un wattmètre externe dédié (Omegawatt)) et un capteur interne (iDrAC/IPMI). COEES permet aussi la mesure des opérations de consolidation supportées par le Cloud. Dans la figure 5.21, on peut ainsi observer un déplacement de 5 machines virtuelles présentes sur un seul serveur physique vers 5 serveurs physiques différents. Grâce à COEES, les concepteurs de gestionnaires de tâches et de ressources peuvent mesurer et comparer l'impact énergétique des solutions activées (optimisations, leviers verts) dans le Cloud.

Figure 5.20: Mesures cumulées d'une grappe de machines avec wattmètre omegawatt et IPMI

Figure 5.21: Rapport ganglia sur un déploiement de machines virtuelles sur une grappe de machines

5.3.3 Nuage vert dans des scénarios de HPC Cloud

5.3.3.1 L'approche XLCLOUD

Depuis 2012, je suis membre du projet FSN XLCLOUD qui se focalise sur la fourniture d'un environnement de Cloud basé sur Openstack afin de supporter les contraintes et les besoins d'applications de calcul intensif. Je mène ces travaux avec François Rossigneux (ingénieur que j'ai recruté dans XCLOUD) et Jean-Patrick Gelas.

Le projet XLCLOUD est supporté par le Fonds de Solidarité Numérique et mené par BULL. Il inclut des partenaires académiques (Institut Telecom, INRIA) et industriels (Bull SAS , Serviware, AMG.net, Artemis, R2SM, Silkan, EISTI, ATEME, CEA List). L'objectif de ce projet est de pouvoir efficacement supporter dans des infrastructures virtualisées un ensemble varié d'applications de calcul intensif. Dans ce projet notre contribution porte sur la gestion efficace en énergie de l'infrastructure (figure 5.22) avec l'ajout de nouveaux services de facturation, d'ordonnancement, de réservation.

Alors que le module COEES de CompatibleOne est conçu de manière indépendante et interfacé avec Ganglia, l'approche choisie dans XLCLOUD est d'intégrer de nouveaux modules dans l'architecture OpenStack [8] (Figure 5.22). Nos contributions s'insèrent donc dans la branche logicielle de OpenStack en lien avec les équipes développant ce produit *open source*.

5.3.3.2 Kwapi : gestionnaire de mesures électriques dans Openstack

Nous avons proposé le gestionnaire KWAPI (*kilo-watt Application Programming Interface*) qui est responsable de la collecte de mesures énergétiques et la remontée d'informations aux différents

[8]OpenStack : Open source software for building private and public clouds : `http://www.openstack.org/`

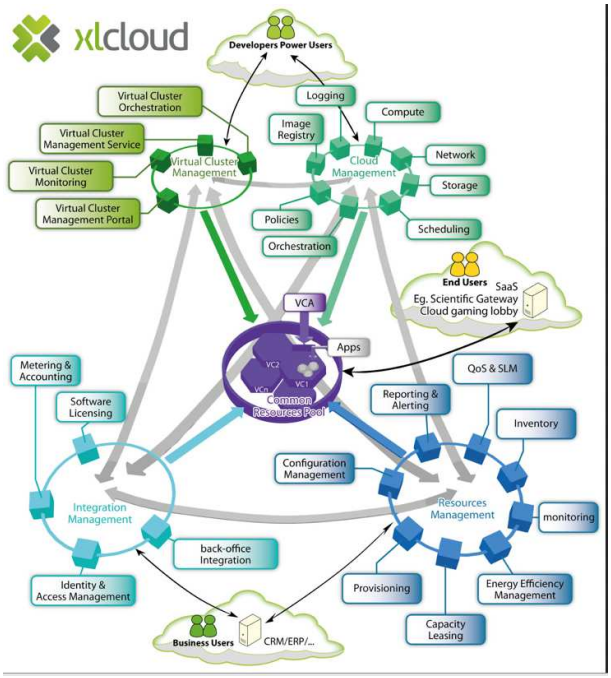


Figure 5.22: Architecture de XLCLOUD

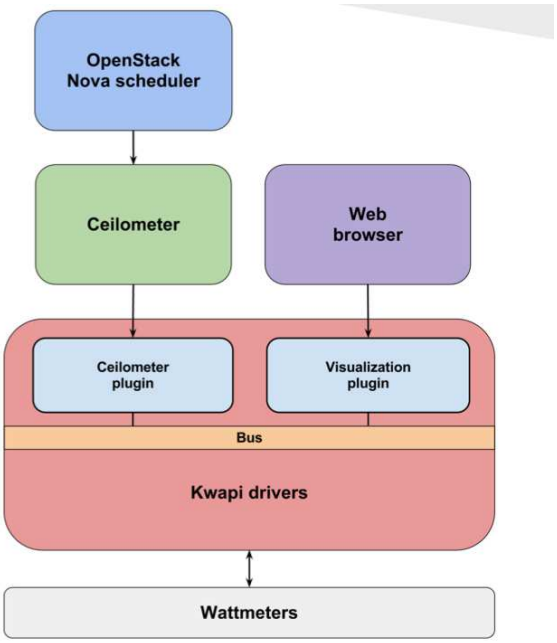


Figure 5.23: Architecture de Kwapi

composants de OpenStack.

Kwapi déploie un ensemble de *drivers* logiciels interfacés avec les capteurs énergétiques (figure 5.24). Nous attachons *kwapi* à l'infrastructure de comptage et de métrique de OpenStack appelée *Ceilometer*[9] (figure 5.23). L'expertise nécessaire à la mise en œuvre du module COEES dans CompatibleOne a été rentabilisée dans la réalisation de *Kwapi*.

Kwapi fournit un ensemble de modules afin de supporter les opérations d'ordonnancement et de facturation à l'usage d'OpenStack. Nous avons vu que des machines homogènes en performance (flops) n'utilisent pas forcément la même énergie (section 4.4.2). Kwapi favorise donc l'utilisation des machines les plus efficaces en énergie en maintenant un indice de performance par machine physique (flops/watt). La calibration de performance est réalisée une seule fois à l'allumage de la ressource physique. La mesure énergétique est réalisée de manière régulière car elle peut varier avec le temps.

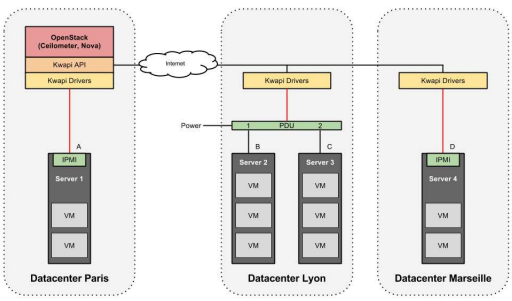


Figure 5.24: Architecture XLCLOUD avec Kwapi

Figure 5.25: Mesures énergétiques dans OpenStack avec Kwapi

[9] `https://launchpad.net/ceilometer`

5.3.3.3 Climate : Réservation dans le nuage

Nous proposons à la communauté OpenStack une nouvelle manière d'utiliser une infrastructure de Cloud en autorisant la réservation de ressources physiques. Cette approche peut nous permettre d'ordonnancer les réservations sur les bonnes ressources (les plus efficaces en consommation électrique) au bon moment (lorsque l'énergie est peu chère ou la plate-forme est sous-utilisée). Nous étudions différentes stratégies de réservation : *immédiate* (les ressources sont réservées instantanément), *planifiée* (les ressources sont réservées à une date précise pour une durée donnée), *best-effort* (les ressources sont réservées dès que disponible). Bien sûr, ce modèle de réservation pourra favoriser une facturation adaptée en fonction des contraintes exprimées par l'utilisateur. Notre approche dans XCLOUD respecte le modèle ERIDIS afin de favoriser l'extinction des ressources inutiles en les éteignant et les rallumant de manière maitrisée pour de ne pas mettre en péril l'infrastructure des *datacenters* tout en garantissant une bonne qualité de service aux utilisateurs du cloud.

5.4 Conclusion

Le modèle ERIDIS proposé dans la thèse de Anne-Cécile Orgerie a permis la mise en place de différentes composants logiciels pour réduire l'énergie dans les infrastructures distribuées à grande échelle.

Dans la plate-forme Grid5000, l'environnement logiciel EARI a démontré que des politiques vertes de placement de réservations peuvent réduire la consommation électrique de manière très significative. Les résultats sur EARI ont été suffisamment convaincants pour que l'équipe de développement de la plate-forme Grid5000 propose une version simplifiée d'économie d'énergie au sein du système de réservation OAR[10]. Ces fonctionnalités d'économie sont donc maintenant disponibles en mode opérationnel pour les administrateurs et les utilisateurs de Grid5000.

Le modèle Green Open Cloud proposé dès 2009 avec Anne-Cécile Orgerie a permis d'ouvrir la voie à des infrastructures de Cloud éco-sensibles et efficaces en consommation électrique. Ce modèle a été validé à l'aide de solutions d'émulations et de *replay* de traces car aucun modèle d'usage de Cloud significatif n'était disponible.

Nous contournons cette difficulté, encore présente aujourd'hui, en contribuant à la source à la création d'intergiciels de cloud comme OpenStack (projet XLCLOUD). La mesure énergétique et l'efficacité dans les clouds représentent un domaine nouveau qui commence à être exploré dans les infrastructures de clouds. Nous souhaitons poursuivre notre implication avec des industriels dans ce domaine afin de proposer des solutions déployables dans des infrastructures de production.

La remontée d'informations de consommation électrique dans un cloud doit néanmoins être réalisée en respectant certaines règles. En effet, l'exposition de mesures électriques peut aussi constituer une source d'information utilisable à mauvais escient contre les utilisateurs de Clouds. Dans l'action européenne COST IC804, j'ai établi une collaboration avec l'Université de Vienne (Helmut Hlavacs et Thomas Treuner, Autriche), pour étudier les risques potentiels associés à la mise à disposition de mesures énergétiques. Nous avons étudié les possibilités de détection et de reconnaissance d'applications dans des environnements virtualisés par simple connaissance de la consommation énergétique. Nous avons montré que certaines applications ont une empreinte d'usage des ressources suffisamment marquée et détectable en termes de consommation électrique qu'elle permet de les reconnaitre [91]. Ceci peut constituer un risque à prendre en

[10] `https://www.grid5000.fr/mediawiki/index.php/OAR2`

compte dans le cas de Clouds partagés.

Nos propositions permettent la création de nouveaux services de Cloud que nous étudions actuellement :

- la facturation à l'usage (et non plus au temps) afin de facturer précisément l'usage énergétique des ressources. Ceci doit encourager les utilisateurs à développer des applications moins gourmandes en ressource ;
- un ordonnancement réellement efficace en énergie (en prenant en compte l'hétérogénéité électrique des ressources de calcul qui n'est jamais prise en compte dans la littérature [19]) ;
- la réservation de ressources permettant de planifier et gérer plus efficacement les ressources nécessaires. Cette approche nous permet ainsi d'appliquer des solutions fournies par le modèle ERIDIS.

"Notre maison brûle et nous regardons ailleurs."
J Chirac, Sommet pour le développement durable, Afrique du Sud, 2002

6

Améliorer l'efficacité énergétique des très grandes infrastructures HPC : avec ou sans connaissance des applications et des services

La "course à l'*exascale*" est lancée, avec pour objectif de construire une machine capable de supporter l'exaflops (10^{18} opérations flottantes par seconde) pour l'horizon 2018-2020. Cette course entraine dans son sillage la communauté HPC (*High Performance Computing*) en prenant les mêmes arguments que dans l'automobile : "Ce que l'on étudie et utilise en Formule 1 se retrouvera demain dans les voitures de base". Ce que l'on prépare pour l'*exascale* se retrouvera demain dans les machines de calculs de monsieur Tout-le-monde.

Actuellement, la plus grosse machine du Top500 [1] est la machine chinoise Tianhe2 (classement de Juin 2013) qui embarque 3 millions de cœurs de calcul pour une puissance de calcul de 33 Pflops et qui nécessite une puissance électrique de 17 MW. Parallèlement, la machine la plus efficace en consommation énergétique relevée par le classement Green500[2] est la machine de CINECA avec une efficacité énergétique de 3.2 GFlops par watt consommé (pour un total de 30KW) [154, 66]. La machine Tianhe2 n'est classée que 32 ème du Green500 avec une efficacité énergétique de 1.9 GFlops par watt.

Aux USA, le DARPA (Defense Advanced Research Projects Agency) a mis comme contrainte forte le fait de réaliser une machine exascale avec une consommation maximum de 20MW. Si l'on veut soutenir cette course à la puissance tout en restant "mesuré" en termes de consommation électrique, il va donc falloir construire des systèmes avec une efficacité de 50 GigaFlops par watt soit un grain en efficacité entre 15 et 25 !

Les améliorations des composants matériels permettront de gagner en efficacité énergétique. Mais les logiciels ont aussi leur part de potentiel d'efficacité dans cette course. Améliorer les environnements, les services et les applications déployés sur des infrastructures de calcul hautes performances est donc un réel challenge que j'explore depuis plusieurs années.

J'adresse ce problème sous deux angles différents en fonction des hypothèses prises en compte.

Dans la thèse de Mehdi Diouri [55] (co-encadrée avec Olivier Gluck, 2010-2013), nous considérons que nous sommes capables d'analyser finement les applications et les services HPC afin de déterminer les composants des services les plus coûteux en termes de consommation

[1] Classement des 500 plus gros super-ordinateurs dans le monde : `http://top500.org`

[2] Classement des 500 super-ordinateurs les plus efficaces en consommation énergétique dans le monde : `http://green500.org`

énergétique. Dans cette thèse, nous nous focalisons sur deux types de services nécessaires au HPC : la tolérance aux pannes et la diffusion de données. Certaines de ces activités de recherche ont lieu en collaboration avec Franck Cappello dans le cadre du laboratoire commun INRIA-ANL sur les architectures Petaflops[3]. Je participe aussi à l'initiative Européenne *European Exascale Software Initiative* [4] en tant qu'expert dans le groupe de travail sur l'énergie et la performance des infrastructures exascale.

Dans la thèse de Ghislain Landry Tsafack Chetsa [45] (co-encadrée avec les chercheurs de l'IRIT à Toulouse : Jean-Marc Pierson et Patrica Stolf, 2010-2013), nous prenons le parti de dire que les applications de calcul haute performance sont trop complexes et difficiles à maintenir pour être modifiées. Nous préférons surveiller le système dans sa globalité pour détecter des phases d'accalmie, intensives ou hétérogènes dans l'utilisation des ressources afin de proposer l'application d'un ensemble de leviers verts aux moments opportuns. Ces activités de recherche ont lieu dans le cadre de l'action d'envergure Hemera [5].

Ces deux approches menées en parallèle explorent des points de vue différents. Mais ces thèses ont un point commun : elles ne s'intéressent pas à l'optimisation logicielle des applications et à l'efficacité énergétique que l'on pourrait essayer d'en retirer.

6.1 Avec connaissance des applications et des services : services efficaces en énergie dans l'exascale

Dans le cadre de la thèse de Mehdi Diouri (co-encadrée avec Olivier Gluck)[55], nous avons adressé le problème de l'efficacité énergétique des infrastructures distribuées à grande échelle de type exascale en nous concentrant sur les services applicatifs génériques. Un service applicatif est "un composant logiciel permettant de réaliser une fonctionnalité donnée pour la bonne exécution de l'application, au service de cette dernière et qui peut être utilisé avec une large gamme d'applications de calcul haute performance. Comme exemples de services on peut citer la tolérance aux pannes [38], la visualisation de données d'expérience [1], les échanges de données massifs [3], la supervision des ressources à grande échelle avec la manipulation d'énormes volumes de traces [179] La consommation de ces services devient non négligeable à l'échelle exaflopique et leur caractère générique nous pousse à les étudier.

La méthodologie proposée repose sur 4 étapes distinctes : l'analyse et le découpage des services associés aux applications de calcul hautes performances, la calibration énergétique des opérations de base des services sur un ensemble ciblé de scénario et d'architectures du système distribué à grande échelle, l'estimation de la consommation énergétique dans des conditions non mesurées et l'aide aux utilisateurs afin qu'ils choisissent les versions des opérations correspondantes à leur besoin et efficaces en consommation énergétique.

L'approche suivie avec Mehdi est basée sur deux composants principaux : un calibrateur de consommation électrique et un estimateur apte à conseiller les utilisateurs dans leurs choix de services (figure 6.1).

La méthodologie est validée étape par étape sur les services applications de tolérance aux pannes (exécuté en arrière plan, parallèlement à l'application) et de diffusion de données (impliquée dans l'enchainement des étapes applicatives) qui exhibent des besoins et des comportements différents.

[3]INRIA-Illinois-ANL Joint Laboratory for Petascale Computing : `http://jointlab.ncsa.illinois.edu/`
[4]European Exascale Software Initiative - EESI2 : `http://www.eesi-project.eu/`
[5]Action d'envergure Hemera : `https://www.grid5000.fr/Hemera`

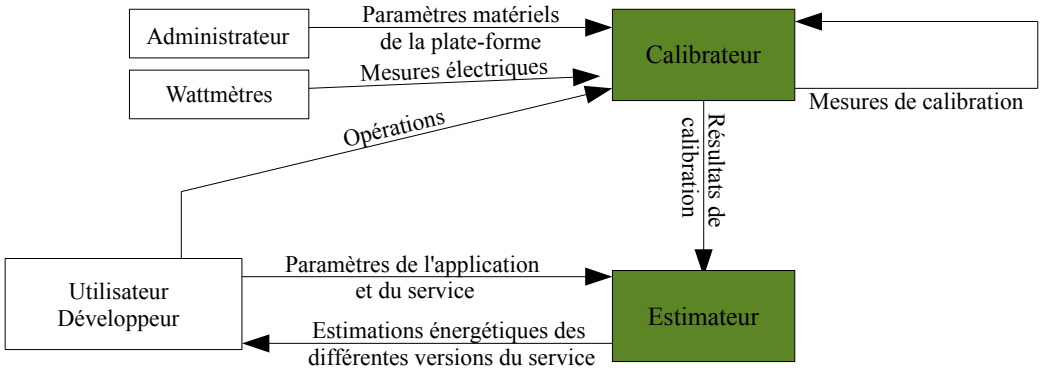


Figure 6.1: Méthodologie de réduction d'énergie avec connaissance des services

6.1.1 Découpage des services HPC en opérations

La première étape de notre méthodologie consiste à identifier les différentes opérations que nous retrouvons dans les différentes versions d'un service applicatif. Une opération est une tâche que le service applicatif peut devoir effectuer plusieurs fois pendant l'exécution d'une application et elle peut se décomposer en une ou plusieurs phases. Cette étape requiert une connaissance précise du service. Chaque opération doit pouvoir être évaluée.

Par exemple dans le cas de services de tolérance aux pannes ; les deux grandes familles de protocoles (coordonnés et non coordonnés) ont été étudiés. Les protocoles coordonnées [32] reposent sur une synchronisation de tous les processus de l'application avant la sauvegarde d'un point de reprise (*checkpoint*). En cas de panne, tous les processus doivent redémarrer au dernier point de reprise. Les protocoles non coordonnés [31] reposent sur l'enregistrement des messages envoyés lors de l'exécution. En contre partie, lors d'un redémarrage après panne, seuls les processus défaillants recommencent à partir du dernier point de sauvegarde. Ces services sont fondés sur les 4 opérations suivantes : sauvegarde des points de reprise (sauvegarde d'une image instantanée de l'état courant de l'application, enregistrement de messages (sauvegarde sur un disque ou en mémoire), coordination (synchronisation de processus), attente (lors de barrière de reprise sur erreur).

Dans le cas de services de diffusion de données de type MPI (*Message Passing Interface* [125]) nous détectons et différencions les 4 opérations réseaux suivantes : *Scatter* (découpage des messages et envoi de *"1 vers n"*), *AllGather* (échange de données entre tous les nœuds de *"n vers n"*), *CopyPrivate* (copie d'un message d'un processus vers tous les processus d'un même nœud), *Pipeline* (découpage d'un message et transmission en mode *pipeline*).

6.1.2 Calibration de la consommation énergétique des opérations

Pour chacune des opérations identifiées et pour chaque type de nœuds de la grappe, nous calibrons le surcoût moyen de puissance électrique. Les figures 6.2 et 6.3 montrent un exemple de calibration électrique sur la plate-forme Taurus (Site de Grid5000 Lyon). Grâce à une maîtrise des équipements de mesures (wattmètres section 4.2), les opérations sont mesurées en termes de surcoût de puissance électrique moyenné. Cette mesure nécessite une phase de calibration assez fine afin d'amplifier l'empreinte électrique de chaque opération (en augmentant le nombre de messages échangés, la taille des données, le volume de calcul). Ceci permet d'extraire la consommation électrique de chaque opération par rapport au bruit électrique résiduel (consommation des ressources matérielles, consommation du système d'exploitation. . .).

Cette calibration est effectuée dans différents contextes matériels; par exemple en faisant varier le nombre de cœurs utilisés sur le serveur (figures 6.2 et 6.3) mais aussi en faisant varier les paramètres applicatifs tels que la taille d'un point de reprise (figure 6.4) ou le nombre de

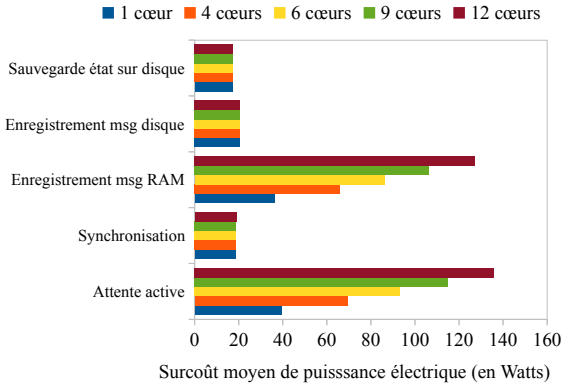


Figure 6.2: Surcoût moyen de puissance électrique (opérations des protocoles de tolérance aux pannes) - site Grid5000 Lyon/Taurus

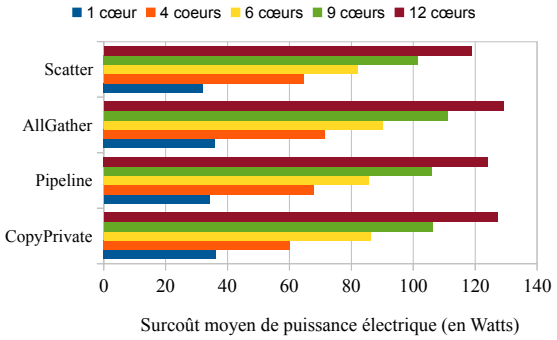


Figure 6.3: Surcoût moyen de puissance électrique (opérations des algorithmes de diffusion de données) - site Grid5000 Lyon/Taurus

nœuds impliqués dans une opération de synchronisation (figure 6.5).

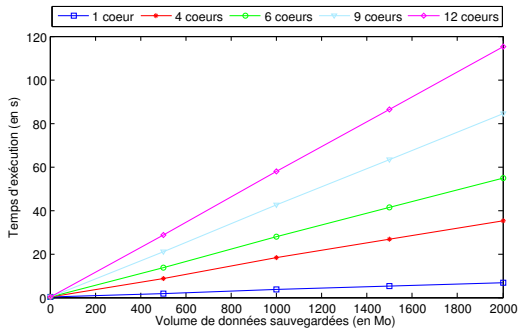


Figure 6.4: Calibration de la sauvegarde de points de reprise sur disque dur local

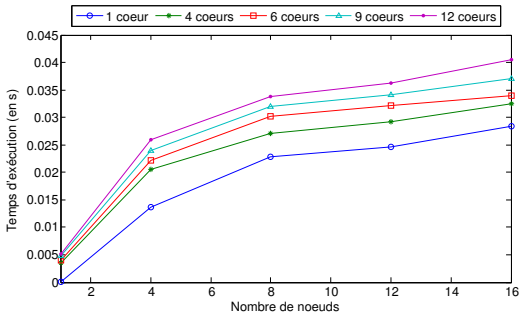


Figure 6.5: Calibration du temps de synchronisation de la plate-forme expérimentale

6.1.3 Estimation de la consommation électrique

La calibration n'est possible que sur une famille de machines. Chaque changement d'architecture matérielle impose une nouvelle calibration. Mais nous avons aussi observé (section 4.4.2) que les nœuds d'une même plate-forme peuvent avoir un comportement électrique très différent (voir section 4.4.2). Il faut donc associer à cette calibration un ensemble de mesures de modes `idle` sur les machines de la grappe.

La calibration est coûteuse en temps humain, ainsi le nombre de points de mesures est limité (nombre de cœurs utilisés, taille des données échangées, volume de données sauvegardées). Il manque donc volontairement des scénarios dans la calibration. Par exemple, dans la figure 6.2, la surcoût moyen électrique d'une attente active lorsque l'on utilise 8 cœurs n'est pas connu pour les machines de la plateforme Taurus. Mais la mesure est connue pour 6 et 9 cœurs. Dans la figure 6.5, nous ne connaissons pas le temps d'exécution d'une synchronisation avec 10 nœuds; mais on dispose d'une mesure avec 8 et 12 nœuds.

Le rôle de l'estimateur est donc de prédire la consommation des opérations de base des services applicatifs avant leur exécution à grande échelle :

- sur des machines dont on possède un ensemble de mesures de calibration ainsi que la consommation en mode `idle`;
- sur des configurations (nombre de cœurs, taille des données) non mesurées pendant la phase de calibration.

L'estimateur propose une prédiction de performances sur la durée des opérations ainsi que sur le surcoût moyen de puissance électrique associé à chaque opération. Ainsi l'estimateur est capable de proposer une valeur énergétique pour chaque opération lors de son utilisation dans un service associé à une application.

Dans le cas d'un service de tolérances aux pannes, nous étudions 4 applications : CM1 une application de météorologie [6] avec une résolution 2400x2400x40 exécutée sur 144 processus de la grappe Taurus et 3 bancs d'essais NAS[11] en classe D (SP, BT, et EP) exécutés sur 144 processus (i.e., 12 nœuds de 12 cœurs par nœud) de la grappe Taurus. Sur la figure 6.6, nous affichons la consommation énergétique proposée par notre estimateur pour chaque opération d'un service de tolérance aux pannes associé à chaque application de calcul haute performance considérée sur une plate-forme cible. Nous observons que la consommation énergétique des opérations n'est pas la même en fonction des applications car cette consommation dépend du volume de données traité. Par exemple, la consommation énergétique de l'enregistrement de messages sur RAM avec l'application SP est 10 fois plus importante que celle avec l'application CM1. Ceci est dû au fait que CM1 échange beaucoup moins de messages que l'application SP.

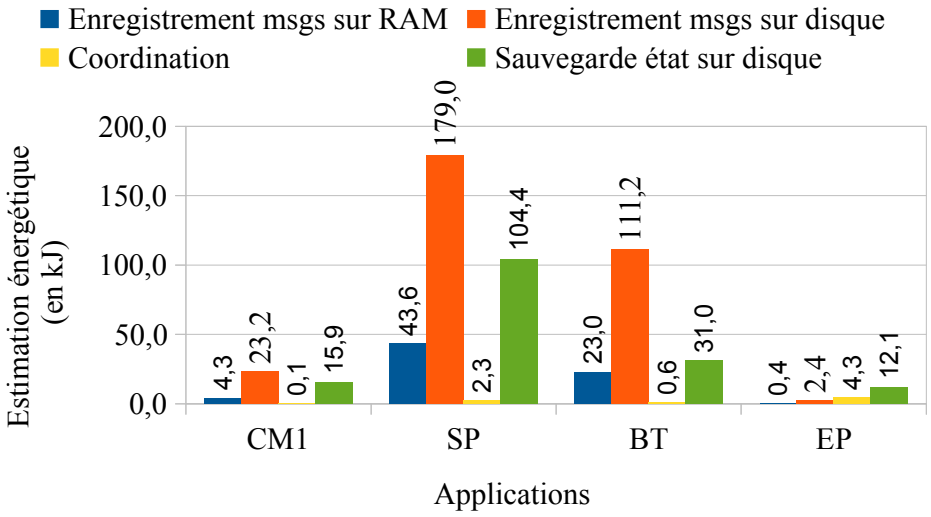


Figure 6.6: Consommations estimées (en kJoules) pour les opérations de tolérance aux pannes

Nous évaluons la précision de cette estimation (figure 6.7). Nous observons des imprécisions sur les opérations d'enregistrements de messages sur RAM et de coordinations car ce sont les opérations les plus courtes, ce qui peut générer quelques imprécisions lors de la phase de calibration. L'estimation s'avère précise avec des écarts d'au maximum 7% et avec un écart moyen de 4.9%..

6.1.4 Aider les utilisateurs à prendre les bons choix

L'ambition de notre approche est d'aider l'utilisateur d'un système distribué à grande échelle à prendre les bonnes décisions en qui concerne l'efficacité énergétique des services applicatifs. Pour un besoin donné, différentes versions d'un même service applicatif existent et différentes

[6] Cloud Model 1 : http://www.mmm.ucar.edu/people/bryan/cm1

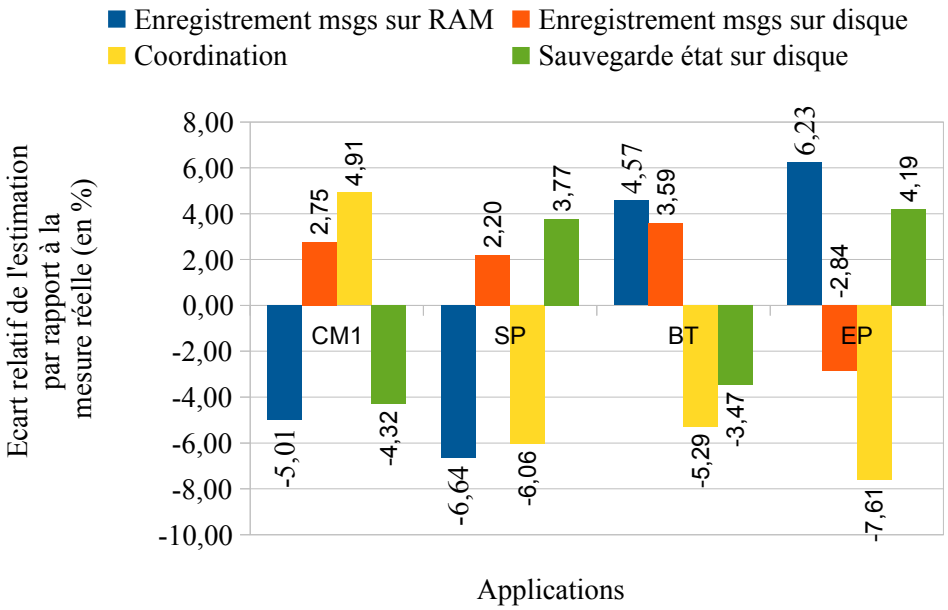


Figure 6.7: Écarts (en %) entre consommations estimées et mesures pour les opérations de tolérance aux pannes

implémentations d'une même opération sont également disponibles. Cette aide à la décision est utile pour le développeur qui est en train de mettre au point un service applicatif et désire avoir une estimation de sa consommation électrique dans différents contextes applicatifs (figure 6.2). Mais elle est aussi utile pour l'utilisateur de système HPC qui est confronté à plusieurs implémentations d'un service applicatif et peut ainsi choisir le service le plus approprié en fonction de ses contraintes de performances ou d'énergie.

Par exemple, pour chaque application et pour chaque protocole de tolérance aux pannes, nous estimons la consommation énergétique en nous appuyant sur les différentes opérations identifiées. Nous obtenons la consommation énergétique totale de chaque protocole de tolérance aux pannes en additionnant les consommations énergétiques des opérations considérées dans chaque protocole :

- Protocole coordonné = Sauvegarde État + Coordination(s)

- Protocole Non Coordonné = Sauvegarde État + Enregistrement des Messages (RAM ou Disque)

Dans le cas simpliste d'une unique sauvegarde des points de reprise, le protocole qui consomme le moins d'énergie est le protocole coordonné pour les applications CM1, SP et BT, car les valeurs énergétiques pour l'enregistrement de messages (RAM et disque) sont supérieures aux valeurs énergétiques de la coordination (figure 6.6). Pour EP, le protocole le moins consommateur est le protocole non coordonné (avec enregistrement des messages sur RAM ou disque).

Dans le cas plus probable où plusieurs *checkpoints* sont réalisés pendant l'exécution de l'application, il faudra analyser le nombre de sauvegarde face aux opérations de coordinations et proposer un choix adapté aux utilisateurs.

6.2 Sans connaissance des applications et des services : en analysant l'utilisation des ressources des systèmes

Dans la thèse de Ghislain Landry Tsafack Chetsa[45] (co-encadrée avec Jean-Marc Pierson et Patricia Stolf du laboratoire IRIT à Toulouse, 2010-2013), nous avons aussi pris le parti ambitieux de proposer des infrastructures logicielles automatiques d'amélioration énergétique sans connaissance *a priori* des applications et des services exécutés. L'approche repose sur une observation d'un système distribué (type ensemble de serveurs) en mesurant l'utilisation des ressources par les compteurs de performances associés. Cette mesure permet de diviser la vie du système en différentes phases correspondantes au comportement : calcul intensif, utilisation mémoire intensive, entrées/sorties... L'enjeu est de détecter au plus tôt (en cours de phase) le type de comportement du système et d'appliquer ou d'activer une solution de réduction énergétique (leviers verts).

L'approche proposée repose donc sur trois étapes distinctes : la détection de phases, la caractérisation et reconnaissance des phases détectées et l'application de leviers verts en vue d'une réduction énergétique (Figure 6.8) [163].

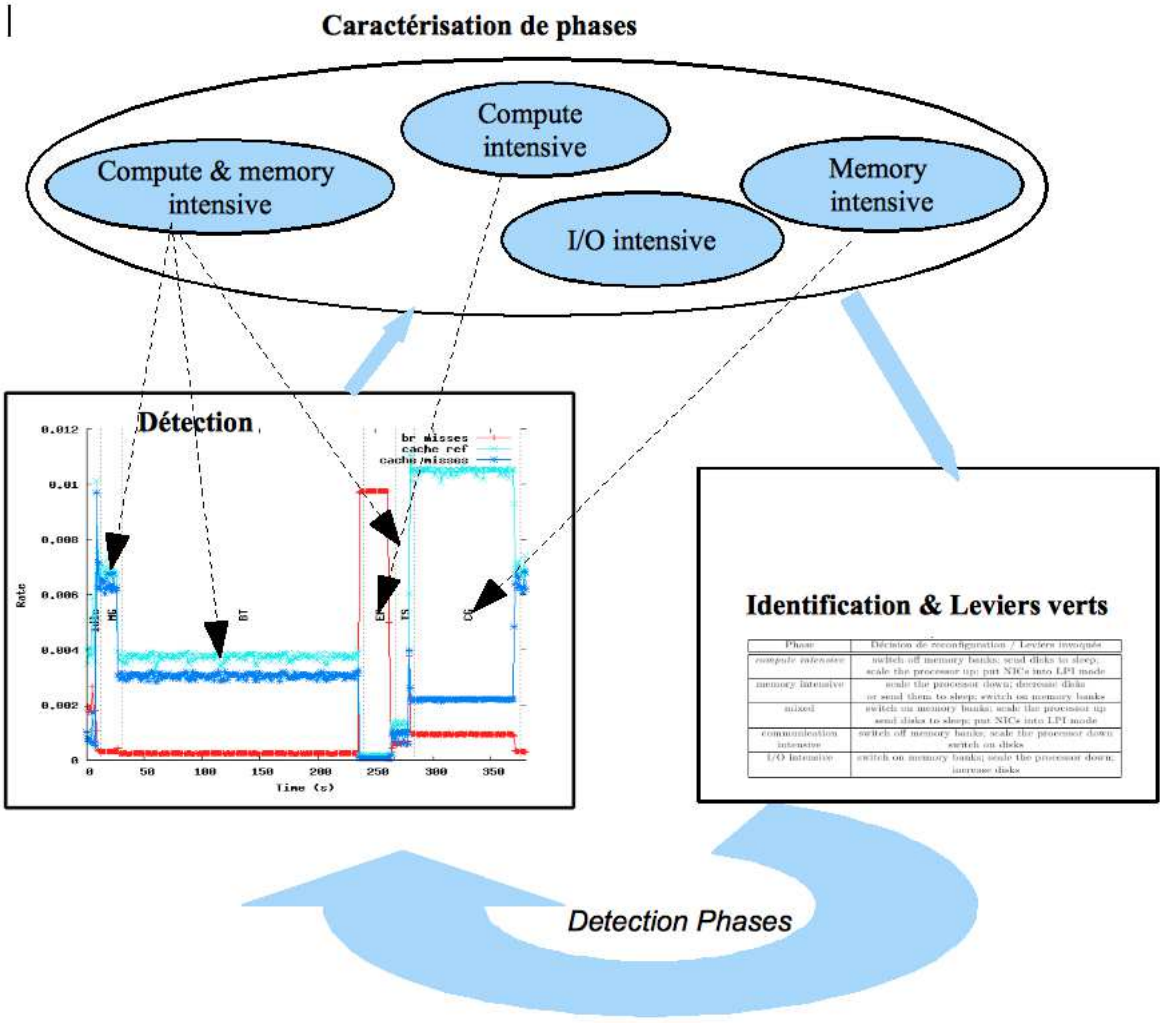


Figure 6.8: Détection, caractérisation, reconfiguration

6.2.1 Détection de phases

Notre approche générale est fondée sur la reconnaissance partielle des phases au cours de l'exécution du système. Une phase de système représente le comportement d'exécution du système pendant une certaine durée. Les valeurs à l'intérieur d'une phase ne divergent pas entre deux étapes de plus d'un seuil donné.

Cette étape observe l'utilisation des ressources de la machine en analysant les compteurs de performances [46]. Par exemple, la figure 6.9 présente un système d'une infrastructure distribuée à grande échelle dont la charge du processeur est mesurée. Si le seuil est fixé à 5%, alors nous pouvons détecter facilement 4 phases (lignes verticales dans la figure, de 0 à 2 secondes, de 3 à 6, de 7 à 9 et de 10 à 13 secondes). Chacune de ces phases est caractérisée par une durée, et les valeurs de la charge au cours de la phase. Si le seuil est fixé à 50%, nous pouvons détecter 2 phases (de 0 à 9 secondes, 10 à 13 secondes). On peut aussi utiliser les compteurs d'énergie afin de détecter des phases d'une application scientifique (figure 6.10).

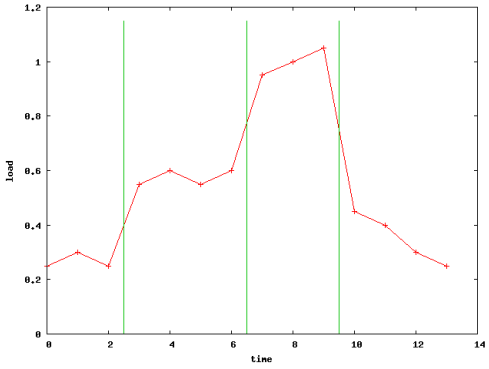


Figure 6.9: Détection de phases en se basant sur la charge du processeur

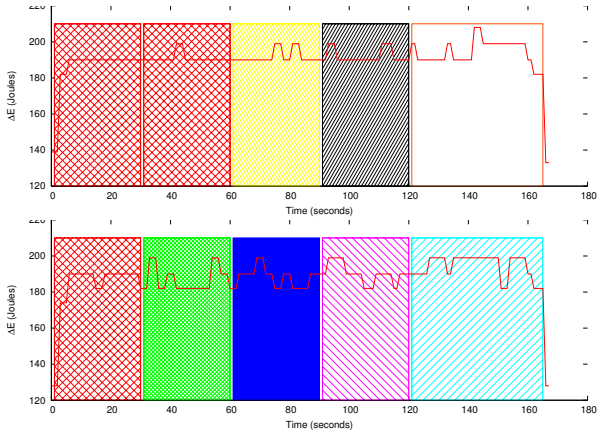


Figure 6.10: Exemple de détection de phases sur deux serveurs exécutant l'application NAS LU

Avoir une seule donnée surveillée n'est en général pas suffisant pour détecter les phases qui peuvent être utiles par la suite. Comprendre le comportement d'un serveur est de plus en plus difficile et dépend du point de vue suivi (figure 6.11).

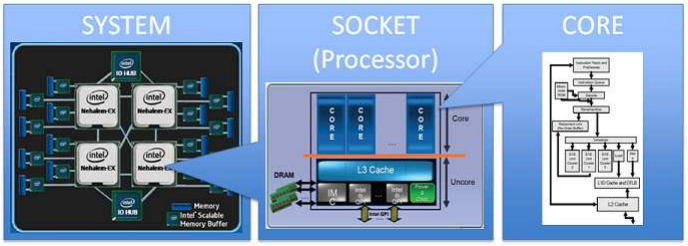


Figure 6.11: Complexité d'un serveur multi-cœurs actuel [173]

L'approche utilisant les compteurs de performances pour évaluer les applications est en vogue dans le monde HPC[111]. Chaque processeur a son propre jeu de compteurs, mais les grands constructeurs comme Intel[7] ou AMD[8] fournissent des interfaces pour accéder aux compteurs. Des environnement comme PAPI (Performance Application Programming Interface)[35] permettent d'exposer de manière uniforme ces compteurs de performance. Plus récemment avec l'apparition

[7]Intel Software Developer Manual, volume 3
[8]AMD Bios and Kernel Developer Guide

du Cloud, des compteurs de performances pour environnements virtualisés et supportés par les hyperviseurs sont disponibles [153].

Notre approche considère donc un ensemble de compteurs surveillés et représentatifs de l'utilisation des ressources du système (Figure 6.1) comme un moyen simplificateur d'observer les activités en cours au sein d'un système multi-cœurs.

PERF_COUNT_HW_CPU_CYCLES	**PERF_COUNT_HW_INSTRUCTIONS**
PERF_COUNT_HW_CACHE_REFERENCES	PERF_COUNT_HW_STALLED_CYCLES_FRONTEND
PERF_COUNT_HW_CACHE_MISSES	PERF_COUNT_HW_BUS_CYCLES
PERF_COUNT_HW_BRANCH_INSTRUCTIONS	**PERF_COUNT_HW_BRANCH_MISSES**
netSENTbyte	**netSENTpkt**
netRECVbyte	netRECVpkt
Write I/O	Read I/O

Table 6.1: Compteurs de performance

Afin de manipuler les phases du système nous avons choisi la représentation sous forme de vecteur d'exécution (dérivé de *Power Vector* [92]). Chaque vecteur inclus différentes métriques: les taux d'accès aux compteurs de performances sélectionnés, le volume de données échangés sur le réseau, le nombre d'écritures et lectures sur disque. La fréquence de mesure et de manipulation des vecteurs est de l'ordre de quelques secondes afin de ne pas perturber le système par des actions trop intrusives. Ces vecteurs nous permettent de calculer une distance (*Manhattan*) entre eux afin de détecter des changements de phase. Ainsi des changements de consommation de ressources pourront indiquer à notre système des changements d'états.

6.2.2 Caractérisation des phases d'un système

Cette partie permet de caractériser les phases observées dans le système afin de déterminer les modes de reconfiguration appropriés. Nous avons proposé cinq catégories de charge représentative de la vie d'un système HPC : *idle* (peu d'activité système), *compute intensive* (activité CPU intense), *memory intensive* (utilisation soutenue de la mémoire), *mixed / compute and memory intensive* (utilisation intensive mixte), *I/O intensive* (utilisation du système de stockage), *communication intensive* (utilisation intensive du réseau).

La principale question adressée ici est de trouver la bonne méthode pour caractériser les phases d'un système. Par exemple, l'état `idle` peut être difficile à caractériser en fonction d'un compteur de performance (figure 6.12).

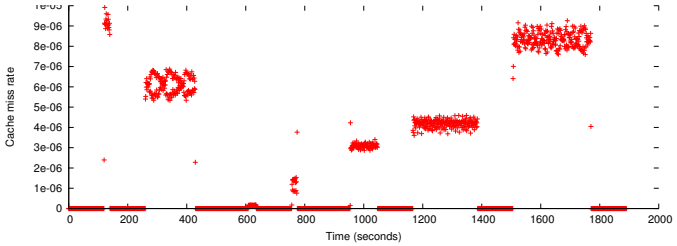


Figure 6.12: Etat `idle` caractérisé par le compteur *cache_miss*

Différents algorithmes de caractérisation ont été proposés, étudiés et comparés :

- LLCRIR (*Last Level Cache References per Instruction Ratio*) : permet de calculer la sensibilité au cache du système. Si le système est très sensible à l'utilisation du cache, son comportement sera plutôt orienté vers un mode *memory intensive*. Cette approche permet de proposer les règles suivantes en fonction de la sensibilité (Table 6.2) mais elle ne permet pas de détecter les modes de communications et d'entrée sorties.

- PCA (compteurs) : en partant d'une approche à base d'analyse en composant principale qui vise à identifier des corrélations entre les données, cette approche permet de sélectionner

Phase	Ordre de grandeur de LLCRIR
compute intensive	$\leq 10^{-4}$
memory intensive	$\geq 10^{-2}$
mixed (memory & compute intensive)	10^{-3}

Table 6.2: Règles de caractérisation pour algorithme LLCRIR (références LLC)

un sous-ensemble de compteurs de performance les moins significatifs. Cette approche permet de caractériser par opposition les phases observées.

Compteurs sélectionnés par PCA (Compteurs)	**Phase**
cache_ref & cache_misses & branch_misses or branch_ins	compute-intensive
no IO related sensor	communication IO intensive
branch_misses & hardware_ins or branch_ins	mixed
hardware_ins & cache_ref or cache_misses	memory-intensive

Table 6.3: Règles de caractérisation pour algorithme PCA (compteurs)

De la même manière que LLCRIR, cette approche ne permet pas de détecter les modes `idle` et caractérise mal les modes d'entrées sortie et de communication réseaux.

- PCA : basé sur l'Analyse en Composantes Principales, cette approche extrait les caractéristiques des charges de travail ayant les mêmes comportements prédominants. Elle permet de détecter comment les compteurs de performances sont corrélés avec les composantes principales. Cette approche fonctionne avec les 5 modes de caractérisation proposés.

La table 6.4 présente les résultats de caractérisation avec les différents algorithmes ainsi qu'une comparaison de deux politiques d'identification.

6.2.3 Identification des phases et application de leviers verts

Dans notre approche il convient d'identifier au plus tôt les phases détectées afin d'appliquer des solutions de reconfiguration du système avec des leviers verts. Nous procédons à une reconnaissance partielle de la phase en prenant comme hypothèse que l'application du levier sur le reste de la phase sera bénéfique sur l'énergie consommée par le système. Cette étape d'identification et les choix de reconfiguration sont listés dans la table 6.6. Ainsi, par exemple, dans le cas d'une phase d'inoccupation (`idle`), le système décide de ralentir la plupart des ressources (CPU, disque et réseaux).

Dans la thèse de Landry nous utilisons principalement des leviers verts issus de la famille (*slowdown*) afin d'adapter la performance des ressources aux besoins du système. Nous bénéficions principalement de DVFS (*Dynamic Voltage and Frequency Scaling*) qui est un des outils les plus utilisés par la communauté scientifique pour adapter la consommation énergétique au besoins réels des applications et des services. Ce levier vert permet de faire varier la fréquence ou le voltage des processeurs afin de s'adapter à la charge de processeur. Cette facilité technologique permet la création d'un ensemble d'outils (gouverneurs) adaptant automatiquement la fréquence à la charge. D'autres leviers existent en fonction des ressources visées et sont applicables : par exemple la parallélisation de l'utilisation des canaux mémoire[61] des modes de micro endormissement dans les cartes réseaux (LPI), de l'adaptation de la bande passante face à la charge (ALR[21] [22]) ...

Bancs d'essai	Caractérisation			Politiques de choix	
	LLCRIR	**PCA**	**Compteurs**	*Majorité*	*PCA puis LLCRIR*
IDLE	mem	idle	mem	mem	idle
FT	mixt	mixt	mem	mixt	mixt
SCP	mixt	IO netRecv transmit mem	mem	com	com
BT	mixt	mixt	mixt	mixt	mixt
IDLE	mem	netRecv transmit	mem	mem	com
CG	mem	mem	mem	mem	mem
IDLE	mem	idle	mem	mem	idle
EP	cpu	cpu	cpu	cpu	cpu
IDLE	mem	idle	mem	mem	idle
UA	mixt	mixt	mixt	mixt	mixt
IDLE	mem	idle	mem	mem	idle
MG	mixt	–	mixt	mixt	mixt
IDLE	mem	netRecv mem	mem	mem	com
SP	mixt	cpu	mixt	mixt	cpu
IDLE	mem	idle	mem	mem	idle
FT	mixt	mixt	mixt	mixt	mixt
SCP	mixt	cpu	mem	cpu	cpu
BT	mixt	mixt	mixt	mixt	mixt
IDLE	mem	netRecv mem	mem	mem	com
CG	mem	mem	cpu	mem	cpu
IDLE	mem	netRecv	mem	mem	com
EP	cpu	cpu	cpu	cpu	cpu
IDLE	mem	idle	mem	mem	idle
UA	mixt	mixt	mixt	mixt	mixt
IDLE	mem	netRecv mem	mem	mem	com
MG	mixt	netRecv transmit mixt	mixt	mixt	com
IDLE	mem	netRecv mem	mem	mem	com
SP	mixt	cpu	mixt	mixt	cpu
EP	mixt	cpu	cpu	cpu	cpu
SCP	mixt	IO transmit mem	mixt	mixt	com
BT	mixt	mixt	mixt	mixt	mixt
IDLE	mem	idle	mem	mem	idle
CG	mem	mem	mem	mem	mem
EP	cpu	cpu	cpu	cpu	cpu
IDLE	mem	idle	mem	mem	idle
UA	mixt	mixt	mixt	mixt	mixt
MG	mixt	–	mixt	mixt	mixt
SP	mixt	–	mixt	mixt	mixt
IDLE	mem	idle	mem	mem	idle

Table 6.4: Comparatifs des algorithmes de caractérisation et des politiques de choix lors de l'exécution d'une suite de bancs d'essai sur un serveur

Table 6.5: *

Légende :		
	cpu	compute intensive
	mem	memory intensive
	mixt	memory et compute intensive
	netRecv	receive intensive (communication)
	transmit	transmit intensive (communication)
	idle	système inoccupé (pas d'application)
	com	communication intensive
	–	pas de résultat

Phase	Décision de reconfiguration / Leviers invoqués
idle	scale the processor down, put NICS into LPI mode, send disk to sleep
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

Table 6.6: Identification de phases et leviers verts associés

6.2.4 MREEF : Multi-Resource Energy Efficient Framework

L'approche proposée avec Landry a été implémentée dans le système logiciel MREEF (*Multi-Resource Energy Efficient Framework*) et validé sur différents systèmes de la plate-forme Grid5000 sur une large gamme de scénarios applicatifs.

Nous illustrons ici un résultat avec différentes applications et bancs d'essais. Nous avons utilisé les bancs d'essais de type NAS [11] : Lower-Upper Guass-Seidel solver (LU), Block Tri-diagonal solve (BT), Conjugate Gradient, Embarrassingly Parallel (EP), Integer Sort (IS), Unstructured Adaptive mesh (UA), Scalar Penta-diagonal solver (SP), and Multi-Grid (MG) qui exhibent des comportements réguliers et des applications réelles telles que MDS (*Molecular Dynamic Simulation*) [23] et WRF-ARW (*Advanced Research Weather Research and Forecasting*) [155] qui ont un comportement hétérogène au cours du temps. Dans la table 6.7, nous présentons la comparaison énergétique et l'impact sur le temps lorsque le système MREEF applique un levier vert unique (DVFS). Le mode performance est celui où le système distribué à grande échelle n'applique aucun levier de réduction énergétique. Les gains sur l'application WRF-ARW peuvent être de 15% en énergie sans impact sur le temps d'exécution. Les gains sur les applications régulières des bancs d'essais sont plus modestes !

Application	Energie (Normalisée)		Temps (Normalisé)	
	Performance	MREEF	Performance	MREEF
WRF-ARW	100	85	100	100
BT	100	97	100	97
SP	100	94	100	100
LU	100	95	100	99

Table 6.7: Comparaisons en énergie et temps pour WRF-ARW, BT, SP, LU avec un unique levier (15 nœuds)

Lorsque l'on observe les gains en énergie sur deux applications réelles qui exhibent des comportements diversifiés dans le temps, on peut escompter des meilleurs résultats avec l'utilisation d'une plus grande palette de leviers. La table 6.8 présente la comparaison des applications en temps et en énergie en utilisant un levier unique (CPU (DVFS)) ou en combinant celui-ci avec du ralentissement réseau (diminution de bande passante type ALR) et mise en veille de disque. L'application des leviers est gérée par le système MREEF contrairement au mode statique (Performance). Ainsi, l'environnement MREEF est capable d'obtenir des réductions énergétiques jusqu'à 25% (MDS) avec 4% de temps en plus.

Application	Energie (Normalisée)			Temps (Normalisé)	
	Performance	MREEF		Performance	MREEF
		Mono Leviers	Multi Leviers		Multi Leviers
MDS	100	81	75	100	96
WRF-ARW	100	87	82	100	96

Table 6.8: Comparaison en énergie et temps pour MDS et WRF-ARW avec de multiples leviers (25 nœuds)

6.3 Conclusion

Améliorer l'efficacité des infrastructures distribuées à grande échelle orientées HPC est un des nombreux défis explorés par la communauté de recherche sur le parallélisme et les systèmes distribués. La recherche de la performance ne peut être le seul critère pris en compte car l'enveloppe énergétique des systèmes distribués à grande échelle est un des principaux facteurs limitant leur développement. La course à l'exascale permet d'étudier des scénarios innovants où le facteur d'échelle impose la création de nouvelles solutions logicielles pour extraire de l'efficacité énergétique.

Dans leurs deux thèses que j'ai co-encadrées en parallèle sur ce sujet, Mehdi et Landry ont proposé des solutions innovantes qui se positionnent dans la création de systèmes exascales efficaces en consommation électrique. Leur approche commune repose sur une maîtrise fine de la consommation énergétique des ressources d'un système distribué à grande échelle dans différents scénarios d'usage.

Les travaux avec Mehdi Diouri sont parmi les premiers à s'attaquer à la mesure d'opérations de services applicatifs afin d'estimer ces opérations dans d'autres contextes matériels ou applicatifs. Ces calibrations et estimations permettent de conseiller plus finement le développeur de services pour les systèmes distribués à grande échelle. La contrainte de cette approche est qu'elle nécessite une compréhension très précise des services et de leur relation avec les applications. Les logiciels doivent être *open source* et maitrisés afin de supporter le découpage en opérations de base proposé par cette méthode. Mais si cette contrainte est levée, cette approche ouvre la porte à des réductions énergétiques conséquentes.

Les résultats de la thèse de Ghislain Landry Tsafack Chetsa démontrent l'efficacité du système d'optimisation énergétique fondé sur la détection et la reconnaissance de phases avant d'appliquer des leviers verts de manière coordonnée. Avec la proposition de l'environnement logiciel MREEF, les systèmes distribués à grande échelle sont analysés en direct et les décisions d'optimisations énergétiques prises sans la connaissance de l'utilisateur ni de l'administrateur. Les solutions à base de multi leviers verts n'en sont qu'à leur début. Seuls quelques leviers sont actuellement disponibles en mode opérationnel (changement de fréquence des CPUs(dvfs), parallélisation de l'accès à la mémoire, variation de la bande passante des cartes réseaux), mais d'autres leviers apparaissent et donnent encore plus d'impact à la méthode proposée.

Les thèses de Mehdi et Landry se terminent au moment de la rédaction de cette habilitation, mais la route vers un *exascale* plus respectueux de l'environnement ou tout au moins sensible et optimisé en consommation électrique est encore longue. Les deux thèses défendues dans ce cadre ouvrent la porte à bien d'autres modèles et approches.

Aller vite a ses avantages. Aller lentement a aussi ses avantages.

Proverbe Africain

7 Conclusions et perspectives

7.1 Bilan de cette habilitation

Les technologies de l'information et de la communication sont devenues indispensables à notre société moderne. Parmi celles-ci, les systèmes distribués à grande échelle constituent une famille à part sur laquelle repose de nombreuses infrastructures et usages. Les infrastructures distribuées à grande échelle évoluent : les *data centers* sont plus imposants, les équipements terminaux plus hétérogènes (moins de PC, plus de tablettes et d'équipements mobiles légers), les services et les nouveaux usages sont massifs (réseaux sociaux, *BigData*, Cloud).

Après des années de travail sur la performance des systèmes distribués et parallèles, je me suis rendu compte que ce n'est ni le seul critère ni le plus le plus important dans notre société numérique actuelle. J'espère avoir démontré dans cette habilitation que les infrastructures distribuées à grande échelle ont aussi besoin de flexibilité et d'efficacité énergétique pour atteindre les enjeux futurs.

7.1.1 ... sur la flexibilité

La question de savoir si on doit respecter le modèle de bout-en-bout a finalement peu d'importance. De nouveaux besoins apparaissent dans les réseaux et ces besoins ne peuvent pas être uniquement couverts par les services et protocoles localisés sur les machines terminales :

- Les *souris* et les *éléphants* continuent de se battre pour les ressources dans les réseaux[87]. Les souris (petit trafic) ont besoin d'une latence faible alors que les éléphants (trafic long) sont plutôt sensibles à une bonne bande passante. Les services de flexibilité doivent tenir compte de ces différences pour proposer des réglages réseaux adaptés.

- De nombreux travaux n'hésitent pas à remettre en cause la suprématie de TCP et à proposer des approches innovantes pour les réseaux à grande échelle [144] De nouveaux protocoles apparaissent en permanence pour améliorer la configuration des *middleboxes* (XSP[96]);

- La communauté scientifique a besoin de plate-formes (*planetlab*, *grid*5000) et d'équipements ouverts [28] supportant la flexibilité afin d'expérimenter rapidement et à une échelle "significative" de nouveaux protocoles pour les valider en vraie grandeur avant de se lancer dans des activités de normalisation et de standardisation.

J'ai participé à 3 vagues de flexibilité qui ont agité la communauté réseaux ces dernières années :

- Les **réseaux pour tous**: lorsque le concept (un peu utopiste) de réseaux actifs est apparu, la programmation d'équipements réseaux par paquets semblait prometteuse. Même si je n'adhérais pas complètement à cette ouverture totale des équipements, j'ai vu dans cette approche un formidable moyen pour mettre en place des plate-formes et des solutions logicielles afin de tester et valider très rapidement de nouveaux protocoles et services réseaux à grande échelle.

  En proposant un des premiers environnements d'exécution actifs capable de supporter les performances des réseaux Gbits (Tamanoir) nous avons surmonté certaines limitations des réseaux actifs. L'environnement Tamanoir a notamment été déployé dans des équipements et validé avec différents services réseaux nécessitant l'assistance de routeurs.

  Mais l'engouement pour le concept de réseau actif ouvert pour tous est retombé. Comment assurer une sécurité des infrastructures et des services réseaux si chaque utilisateur est capable d'injecter son propre code dans les équipements du réseau ? Comment garantir qu'un code malicieux ne va pas prendre toutes les ressources disponibles ? Comment assurer l'équité ? De plus, en autorisant des services lourds dans le réseau, comment, dimensionner les équipements afin qu'ils soient à même d'anticiper de futures demandes ?

- Les **réseaux sans intervention humaine** : L'être humain étant souvent la source des problèmes (sécurité, fiabilité, réactivité), le concept de réseaux autonomes a été évalué par les chercheurs. Le réseaux autonomes permettent d'exhiber des fonctionnalités d'auto gestion intéressantes pour les réseaux à très grande échelle. Nos contribuons à ce domaine notamment dans le cadre du projet Autonomic Internet, ont permis de valider différents concepts de programmabilité en mode autonome.

- Les **infrastructures à valeur ajoutée (mesurée)** : avec l'apparition de la virtualisation réseau[149], le besoin de flexibilité est toujours présent. Par exemple, les architectures de type Cloud ont besoin de performance et flexibilité[123]. Une nouvelle solution de flexibilité prend de l'ampleur avec le développement des réseaux logiciels (*SDN : Software Defined Networks*) avec notamment la technologie Openflow [122, 130]. Les réseaux SDN essaient de ne pas reproduire les erreurs des réseaux actifs en offrant la possibilité à l'opérateur réseau de maîtriser son infrastructure et de configurer les équipements. Donc la flexibilité est bien disponible, mais seulement pour l'opérateur réseau !

  Nous avons contribué à ce domaine notamment avec la mise en œuvre d'infrastructures virtuelles à grande échelle et le développement d'équipements embarquant des services spécifiques (répartiteurs de charge, pare-feux) et dans les protocoles de transport à assistance de réseaux.

Depuis 2010, je participe à une 4 ème vague : la **flexibilité maitrisée pour l'efficacité énergétique dans les services et équipements réseaux**. Je représente l'INRIA au sein du consortium GreenTouch qui se focalise sur la réduction de la consommation énergétique des réseaux de communications d'un facteur 1000 à l'horizon 2015. Cette initiative réunit une cinquantaine d'institutions académiques et de groupes industriels dans le monde. Impliqué à différents niveaux (*executive board*, *co-working group chair* sur les réseaux filaires), je poursuis mes activités logicielles dans les réseaux en utilisant la flexibilité des services et des équipements pour réduire la consommation énergétique tout en garantissant la même qualité de service (voir section 7.2.1). Sur cette lancée, je participe au projet européen CHIST-ERA STAR (*SwiTching And trAnsmission*, 2012-2015), mené par l'Université de Leeds (avec Cambridge, l'INRIA et AGH) qui propose l'intégration de nouveaux équipements de commutation hautes performances et basse consommation avec des piles logicielles aux fonctionnalités adaptées.

7.1.2 ..et l'efficacité énergétique

En peu de temps, l'énergie est devenue le principal facteur limitant la création et le déploiement d'infrastructures distribuées à grande échelle. La prise en compte de l'efficacité énergétique commence à devenir un processus standard dans les recherches sur ces systèmes. Mais lorsque nous avons débuté nos travaux dans ce domaine en 2008, ce n'était pas le cas et nous avons dû défricher différents champs d'investigation.

Mes contributions dans le vaste domaine de l'efficacité énergétique sont focalisées sur :

- **Modéliser et mesurer la composante énergétique** : L'efficacité énergétique n'est pas qu'une simple "chasse aux watts". C'est d'abord la prise en compte d'une nouvelle métrique (puissance électrique ou énergie) comme étant un facteur d'optimisation prépondérant. L'efficacité énergétique a besoin de modèles énergétiques des ressources utilisées dans les infrastructures distribuées à grande échelle. Mais ces modèles doivent être corroborés avec des mesures de la consommation fréquentes et précises afin de garantir le bien-fondé des propositions.

- **Proposer des systèmes logiciels d'adaptation** : Dans nos travaux nous avons pris le parti de n'explorer que des approches avec leviers verts. Nous avons naturellement proposé nos premières contributions avec des possibilités d'extinction des ressources (type: *shutdown*) qui permettent des gains importants mais dépendent d'un usage en dent de scie favorisant des périodes d'inactivité assez conséquentes. Ces travaux adressent la partie statique de la consommation électrique. Plus récemment, nous mettons en œuvre des logiciels d'optimisation énergétique reposant sur les leviers de ralentissement (*slowdown*). Cette dernière approche nécessite une maîtrise plus fine des possibilités de reconfiguration des infrastructures mais permet d'économiser de l'énergie sur la partie dynamique de la consommation.

- **Animation de la communauté Green IT** : Dans l'action européenne COST IC804 (ou j'étais *Working Group Leader* sur les adaptations énergétiques dans les systèmes distribués à grande échelle) et dans l'action d'envergure INRIA Hemera (où je co dirige le défi scientifique sur le profilage énergétique des applications à grande échelle), je participe à la mise en place et à l'animation d'une communauté de chercheurs focalisés sur l'efficacité énergétique dans les systèmes distribués à grande échelle. Les équipes se connaissent maintenant de manière croisée ; c'est un gage de succès potentiel pour de futurs projets nationaux et européens dans ce domaine.

7.1.3 Contributions

Les doctorant(e)s et ingénieurs que j'ai encadrés et les projets auxquels j'ai participé ont permis de faire avancer ma compréhension dans les domaines de la flexibilité et de l'efficacité énergétique. Ces deux domaines suscitent controverses et questionnements car il peuvent être ressentis comme un frein à la course à la performance.

Grâce aux thèses co-encadrées de Jean-Patrick Gelas, Eric Lemoine, Dino Lopez Pacheco, Narjess Ayari et la collaboration avec Pablo Neira Ayuso, je peux explorer de nombreuses facettes de la flexibilité dans les infrastructures réseaux. Ces recherches innovantes bousculent les idées reçues. La validation des architectures proposées par des suites logicielles, des équipements et nouveaux services ont un impact sur ces domaines.

Les thèses menées par Anne-Cécile Orgerie, Mehdi Diouri et Landry Tsafack que j'ai co-encadrées et la collaboration avec Christina Herzog me permettent d'appréhender le domaine de l'efficacité énergétique. Ces travaux sont focalisés sur l'optimisation de la consommation

énergétique en phase d'usage des équipements. Les modèles et architectures proposés sont ambitieux et difficiles à mettre en œuvre mais validés dans les scénarios les plus représentatifs des systèmes distribués à grande échelle actuels (*DataCenters*, *Clouds*, Réseaux). Ils ouvrent la porte à des optimisations et possibilités de réduction énergétique conséquents.

Les systèmes distribués à grande échelle expérimentaux tels que ceux manipulés dans cette habilitation sont des terrains d'expérimentation formidables avant de passer en mode production et opérationnel. J'ai suivi et encouragé une démarche de recherche appliquée tout au long de ces années. Une fois l'architecture ou les modèles construits, j'ai toujours encouragé les doctorants à implémenter des prototypes de leurs solutions afin de les confronter aux réalités des infrastructures distribuées à grande échelle. Cette approche est coûteuse en temps et en moyens humains mais elle représente une validation concrète et réaliste de nos propositions et reste pour moi une véritable aventure technologique et humaine.

7.2 Quelques perspectives scientifiques

Je sélectionne quatre pistes de recherches que je souhaite aborder à court et moyen terme :

7.2.1 Quand la flexibilité contribue au "facteur 1000" dans les réseaux

Depuis 2011, je travaille avec Teferi Assefa (doctorant à l'Université de Addis Abbeba, Ethiopie) et son encadrant Mulugeta Libsie sur le domaine de la virtualisation des passerelles maison (*VHGW : Virtual Home Gateway*) [80]. Cette recherche est menée en étroite collaboration avec Jean-Patrick Gelas et s'inscrit dans le cadre du consortium GreenTouch[1] .

L'initiative GreenTouch, lancée depuis 2010, se focalise uniquement sur les réseaux de communications (filaires, sans fils, optiques et cuivre) afin de proposer un ensemble de solutions matérielles et logicielles pour obtenir une réduction de la consommation énergétique de l'ensemble de ces équipement par un facteur 1000 (**mille !**) à l'horizon 2015.

Quand on observe une infrastructure distribuée à grande échelle de type réseau (figure 7.3), on constate que la plus importante partie de la consommation électrique des réseaux de communication filaires concerne les équipements des derniers kilomètres (réseaux et équipements d'accès et de maison)

	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

Figure 7.1: Consommation électrique et volume d'équipements réseaux pour un opérateur italien (extrait de [27])

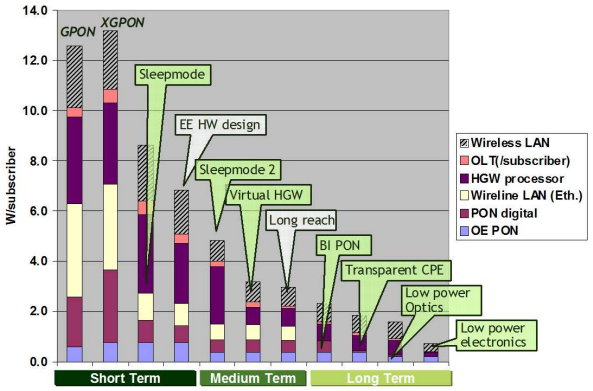


Figure 7.2: *Roadmap* du groupe de recherche Réseaux filaires de GreenTouch

[1] http://www.greentouch.org

Une passerelle maison (*Home Gateway*) est un équipement de bordure embarquant des fonctionnalités de calcul, de stockage et de communication (par exemple : boitier ADSL, serveur frontal...). Cet équipement fournit un service de connexion avec l'opérateur réseau et assure un ensemble de services de plus en plus poussés (voix, musique, télévision, vidéo à la demande, jeux, magnétoscope numérique...). Ces services requièrent la plupart du temps un allumage constant de l'équipement et nécessitent une puissance électrique de plus en plus grande. Ces équipements terminaux ont une consommation raisonnable à l'échelle d'un foyer (de 15 à 30 watts par équipement soit une quarantaine d'euros par an en France) mais leur nombre les rend prépondérants dans la consommation électrique totale d'un infrastructure réseau (figure [27]),

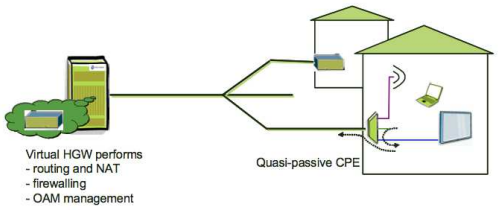


Figure 7.3: Une passerelle maison virtualisée

Partant de ce constat, nous travaillons dans GreenTouch à définir un fil directeur pour nos recherches afin de contribuer au facteur 1000 [166]. C'est dans ce contexte que j'ai initié une activité de recherche sur la virtualisaiton des équipements réseaux de type passerelle. Cette approche provoque la virtualisation des services présents dans une HGW et leur migration et mise à disposition dans un autre point du réseau partagé (*datacenters* de l'opérateur, regroupement régional...). Ces travaux sont menés en lien avec d'autres partenaires de GreenTouch (IMEC et Bell Labs) qui étudient la fourniture d'un équipement remplaçant la passerelle (*Quasi Passive CPE* avec une consommation de quelques milliwatts).

Une architecture de virtualisation de passerelle maison a été proposée et le premier prototype de VHGW mono-machine a été implémenté et validé. Les premières mesures de performances assurent une réduction énergétique d'un facteur 300 tout en garantissant de bonnes performances dans les services réseaux (équité, latence). La figure 7.4 présente ainsi la virtualisation mono-machine de 100 passerelles maison type ADSL (20 Mbps) qui se partagent équitablement le trafic. Cette activité de recherche et développement [165] a été démontrée en 2012 (figure 7.5); c'est une des rares activités très orientée logiciel et financièrement supportée par le consortium GreenTouch.

De la même manière que nous étions confrontés aux problèmes de performances dans les équipements programmables et flexibles travaillant sur le flux des données; la virtualisation de passerelles maison impose une mise en œuvre efficace de technique de *containers* et de virtualisation afin de garantir une qualité de service satisfaisante pour les équipements virtualisés.

Afin d'évaluer complètement les gains énergétiques de cette approche, nous travaillons sur la modélisation de la consommation totale d'une infrastructure de *VHGW* afin de ne pas se focaliser uniquement sur l'usage des équipements IT mais également de prendre en compte la totalité des infrastructures mises en jeux : data-centres, refroidissement, équipements, trafic réseaux. Seule cette modélisation nous garantira un gain réel sur l'approche proposée. Ces aspects combinés à des développements de grande échelle sont en cours d'investigation. Nous sommes au début de cette activité de recherche que je souhaite poursuivre jusqu'à la validation finale de l'approche GreenTouch en 2015. Les perspectives à court terme concernent donc la virtualisation de passerelles maison à grande échelle tout en respectant des niveaux de qualités de services pour les utilisateurs finaux.

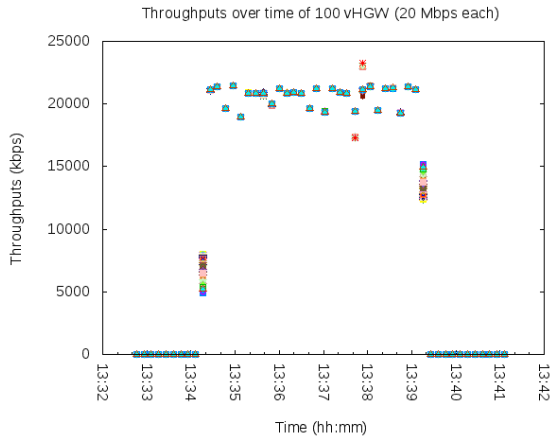


Figure 7.4: Bande passante et équité pour 100 vHGW utilsant 20 Mbps de bande passante

Figure 7.5: Démonstration de **VHGW** sur le stand GreenTouch (Conférence Telecommunications Industry Association, Dallas, Juin 2012)

7.2.2 Vers des infrastructures distribuées à grande échelle à consommation proportionnelle en énergie

Les infrastructures distribuées à grande échelle doivent faire face à un grand nombre de services et d'applications présentant un comportement hétérogène en termes d'utilisation des ressources. Par exemple, les services cloud Web peuvent attendre des demandes externes avant d'utiliser les ressources ; les applications de calcul scientifique doivent faire face à des phases de consommation CPU intensive (et consommant de l'électricité) combinées à des attentes d'entrées sorties ou des temps d'inactivité (en raison de déploiement). Certaines études montrent que en moyenne, beaucoup de centres de données ne sont pas utilisés plus de 20% de leurs capacité.

Nous avons vu que la famille des leviers verts (*slowdown*) permet de re-dimensionner la fréquence des ressources (par exemple *cpu* avec $dvfs$) en fonction de la charge de travail. Mais ces leviers agissent sur la consommation dynamique et ont une portée limitée car la consommation statique des infrastructures est encore importante.

Des travaux ont posé les bases de la conception d'équipements qui pourraient exposer une consommation énergétique proportionnelle à la charge de travail sur les ressources[13] (Figure7.6). Par exemple, en cas d'inactivité, l'équipement doit consommer une très faible quantité d'énergie. Lorsque l'appareil est fortement sollicité (100 % de la consommation), il peut utiliser le maximum de puissance électrique disponible (statique + dynamique). Entre ces deux états, l'équipement doit consommer une fraction linéaire de sa consommation maximum.(figure 7.6). Cette conception doit bénéficier de l'appui de nouveaux équipements associés à un logiciel adapté.

Créer un système dont la consommation énergétique est proportionnelle à la charge de travail est un des Graals du Green-IT.

Nous avons vu dans nos travaux sur l'efficacité énergétique dans le calcul haute performance (chapitre 6) qu'un système distribué à grande échelle traverse des phases très hétérogènes en termes de consommation de ressources (`idle`, calcul, entrées/sorties, communications réseaux. . .). Ces phases provoquent des consommations électriques très différentes et ne nécessitent pas toutes les mêmes ressources. Par exemple, une machine parallèle multi cœurs alimentée pour un simple mode `idle` provoque du gaspillage. Il faut donc concevoir des infrastructures distribuées à grande échelle disposant de ressources matérielles hétérogènes dupliquées et capables de placer les charges de travail à la volée sur les bons composants;

La figure 7.7 illustre l'approche (idéale) que je souhaite explorer : un système exhibant de multiples phases hétérogènes en besoins de ressources (a) qui provoquent une consommation

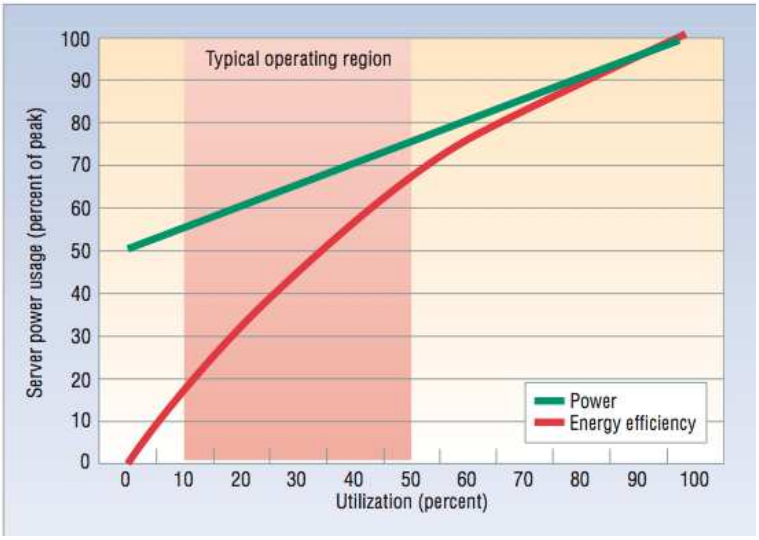


Figure 7.6: Proportionnalité énergétique des serveurs [13]

forte provenant de la partie statique (240 W) et dynamique (20 W). Le système dispose de ressources de calcul hétérogènes (Atom, ARM, Xeon) adaptés pour différents besoins. Chaque changement de phase majeur dans l'application provoque des migrations de la tâche de calcul vers le processeurs adapté (b). A un instant donné, seul le processeur répondant aux besoins du système est alimenté, les autres CPU sont en mode veille.

De nombreux problèmes associés restent ouverts : démarrage et extinction rapide de machines en prenant en compte les pics de puissance électrique éventuels (voir chapitre 5), maîtrise fine des leviers verts, gestion de la migration des données et des calculs sur des architectures hétérogènes multi cœurs, mesures électriques fines.

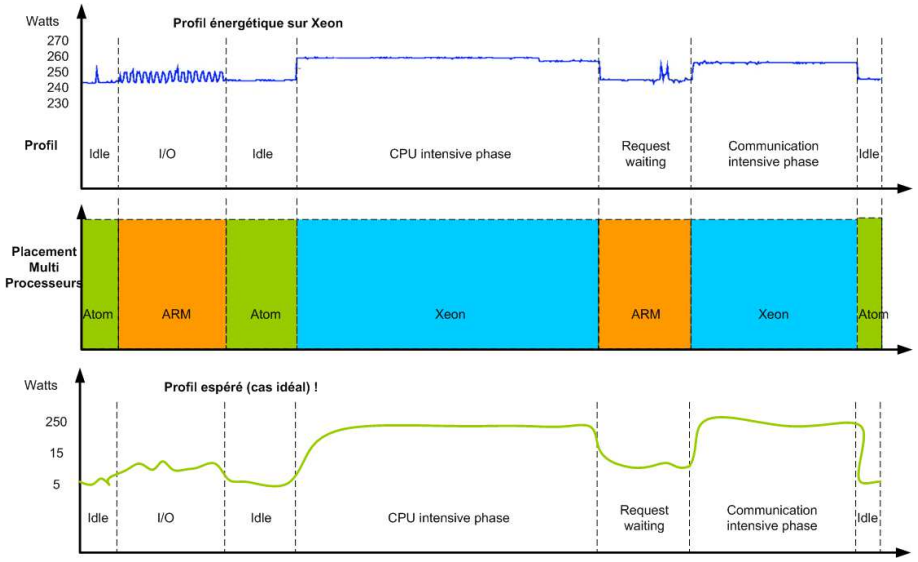


Figure 7.7: Vers une consommation énergétique proportionnele

Si nous arrivions à construire ce modèle sur toutes les ressources d'une infrastructure distribuée à grande échelle (calcul, stockage, réseaux), une des quêtes du Graal en GreenIT serait terminée ! Les leviers verts d'extinction de ressources (*shutdown*) deviendraient inutiles. Les équipements ne consommeraient que l'énergie nécessaire à leur charge de travail. Il demeur-

erait néanmoins indispensable de poursuivre des travaux en optimisation énergétique et en amélioration de l'efficacité énergétique (flops par watt, octet par watt...).

La chemin est encore long : il faut construire ce modèle, développer et valider des solutions de migration, mesurer les gains observés, les pertes... un travail de recherche complet qui a débuté fin 2013 par le co-encadrement du doctorat de Violaine VilleBonnet (doctorante ENS Lyon-IRIT) avec Jean-Marc Pierson.

7.2.3 Lier flexibilité et efficacité énergétique

Pendant de nombreuses années le principal critère pris en compte a été la performance et la qualité de service : plus de Hz, plus de bande passante, plus de mémoire... Cette course effrénée est en train de s'inverser en prenant en compte un développement plus harmonieux et plus en rapport avec les besoins des applications actuelles. Il existe d'autres critères d'efficacité à prendre en compte. Dans la thèse de Mehdi Diouri (co-encadrée avec Olivier Gluck) nous avons ainsi étudié la possibilité de considérer différents objectifs afin d'optimiser la consommation des infrastructures distribuées à grande échelle. Nous proposons une approche parallèle au "consommer moins" : le "consommer mieux." Le système SESAMES (figure 7.8) permet ainsi de choisir différents critères : performance, énergie, pollution, coût financier...

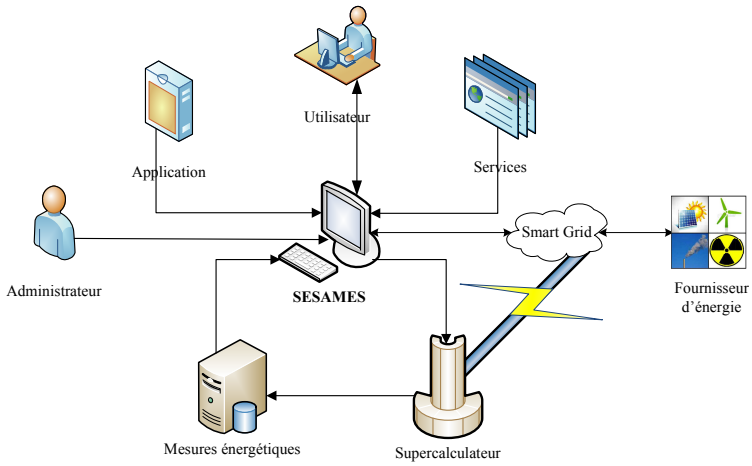


Figure 7.8: Le système de négociation énergétique multi-critères SESAMES [54, 57]

Reposant sur une infrastructure de type *smart grid*, le système permet une communication entre le fournisseur électrique et le système distribué à grande échelle. Il rajoute ainsi plus de flexibilité dans les choix des systèmes d'ordonnancement et de gestion de ressources qui peuvent prendre en compte la variabilité du tarif électrique, la production temporaire d'énergie verte (renouvelable)... Cette approche ouvre la porte à de nouvelles études et réalisations pratiques que je souhaite poursuivre dans le futur afin de considérer des objectifs multi-critères mélangeant flexibilité des infrastructures (*smart grid*) et efficacité énergétique.

7.2.4 Une incursion dans le développement durable

Mes travaux portent principalement sur l'efficacité énergétique pendant la phase d'usage des systèmes distribués à grande échelle. C'est une petite partie du GreenIT. Je souhaite inscrire certains de mes travaux futurs dans une démarche plus globale de développement durable défini comme "un développement qui répond aux besoins des générations du présent sans compromettre la capacité des générations futures à répondre aux leurs". Comme présenté dans la figure 7.9, le développement implique de nombreux aspects écologiques, sociaux et économiques.

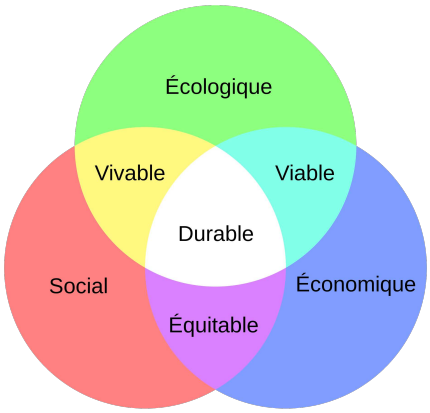


Figure 7.9: Développement durable (A. Villain source wikipédia)

Depuis 2010, je suis membre actif du Groupement de Service du CNRS Ecoinfo[2] qui regroupe un ensemble de chercheurs et d'ingénieurs spécialisés dans la formation et les propositions sur le GreenIT. Cette activité me permet d'acquérir de nouvelles connaissances dans le domaine de l'efficacité énergétique [60] et d'appliquer mon expertise à d'autres scénarios [20].

7.2.4.1 Supporter un autre usage des systèmes distribués à grande échelle?

Le principe fondamental que j'ai suivi dans les travaux de cette habilitation est de proposer de nouvelles fonctionnalités, services et outils tout en respectant au maximum la qualité d'usage des infrastructures distribuées à grande échelle. La qualité de service et l'utilisation ne doivent pas être impactées afin de favoriser le déploiement et l'adoption à grande échelle de solutions de flexibilité ou d'efficacité énergétique.

Mais ce postulat peut être considéré maintenant comme un frein au développement de nouvelles solutions. Il faut replacer l'humain (l'utilisateur, le concepteur de services) au centre des problématiques. Comme première étape, l'utilisateur doit être éco-sensibilisé à son usage des des systèmes distribués à grande échelle. C'est ce que nous avons proposé dans les solutions d'ordonnancement verts ou d'efficacité énergétique dans les clouds et le HPC. Il est temps maintenant de commencer une deuxième phase où l'utilisateur doit accepter un usage différent des ressources informatiques.

Nous travaillons déjà dans ce sens avec le module Climate du projet XLCLOUD (section 5.3.3.3) qui propose d'éviter l'utilisation à la volée des ressources physiques du Cloud en favorisant au maximum la réservation de ressources. Cette approche permet ainsi de dimensionner finement les ressources disponibles et d'éviter la sur-abondance de ressources de calcul et de stockage inutilement alimentées en énergie. Il convient maintenant d'explorer de nouveaux modèles et solutions d'efficacité énergétique en prenant en compte l'hypothèse d'un usage différent.

7.2.4.2 Placer l'efficacité énergétique au cœur de la société

Le Green-IT repose sur des principes d'utilisation raisonnée ou tout au moins éco-consciente. Paradoxalement, ce domaine se retrouve considéré comme un moteur de croissance et d'innovation. Je collabore actuellement avec Christina Herzog (doctorante IRIT à Toulouse) et Jean-Marc Pierson sur la mise en œuvre de solutions d'échanges entre le monde industriel et académique sur le domaine du Green dans les systèmes distribués à grande échelle [89, 90]. Pour des raisons

[2]Ecoinfo : `http://ecoinfo.cnrs.fr/`

financières, d'image ou environnementale, le monde industriel et les entreprises commencent à prendre en compte l'usage énergétique dans leurs modes de production[3]. Les entreprises sont ainsi classées en fonction de leurs efforts pour adopter des politiques plus respectueuses de l'environnement. Les industriels appliquent les recommandations de bonnes pratiques dans le développement de leurs infrastructures distribuées à grande échelle [20]) Nous étudions les facilités d'adoption du Green IT par les entreprises et la manière d'améliorer les collaborations industrio-académiques dans ce domaine.

7.2.4.3 Prendre en compte le cycle de vie des systèmes distribués à grande échelle afin de proposer des solutions valides

Mes travaux portent sur la phase d'usage des technologies de l'information et de la communication. Les métriques que j'ai considéré sont des métriques de puissance (watts) et d'énergie (joules) car elle sont indépendantes des lieux de production et d'usage. Pour moi, **le système distribué à grande échelle qui a le moins d'impact est celui qui évite le sur dimensionnement et utilise une énergie adaptée à la charge de travail !** Avec l'apparition de la vitualisation et des techniques de refroidissement plus efficaces (*freecooling*), on pourrait s'attendre à observer une décroissance de la consommation dans certaines infrastructures distribuées à grande échelle. Il n'en est rien; les centres de données sont par exemple à l'origine de pollutions de plus en plus importantes (Co2) en raison du mix énergétique fortement carboné des régions où ils sont majoritairement implantés.

D'autres métriques telles que le rejet des gaz à effet de serre (comme le CO2) peuvent aussi être mesurées et prises en compte. De plus, l'analyse de cycle de vie des équipements informatiques montrent qu'il existe d'autres phases très importantes, avec par exemple :

- la production et le transport : Alors que l'industrie alimentaire propose des alternatives plus respectueuses de l'environnement (mais aussi plus coûteuse) telles que le bio; la production des technologies de l'information et de la communication n'obéit pas à cette règle. Les équipements électroniques actuels nécessitent des métaux (étain, indium...), minerais (lithium...) et terres rares (telles que cérium, prometheum, samarium, ou lutécium) très souvent récoltés dans des conditions sociales et environnementales désastreuses. Il n'existe malheureusement pas de technologies de l'information et de la communication propres [60]. **Le système distribué à grande échelle ayant le moins d'impact est celui qui est éco-concu ou celui que l'on ne construit pas !**

- la destruction et le recyclage : En 2011, la fabrication des téléphones portables, *smartphones*, ordinateurs et des tablettes numériques a nécessité 320 tonnes d'or (8% de la production mondiale) et 7.500 tonnes d'argent. De ces volumes, moins de 15% seront recyclés ; le reste est donc perdu. Pourtant, on compte 300 grammes d'or dans 1 tonne de téléphones mobiles alors que l'on extrait que 5 grammes dans 1 tonne de minerai. **Le système distribué à grande échelle ayant le moins d'impact est celui que l'on recycle !**

Les technologies de l'information et de la communication ont aussi des impacts sur l'épuisement des ressources, les prélèvements d'eau, la production de déchets, sans parler des problèmes sociétaux et géopolitiques. Afin de proposer des solutions GreenIT complètes il faut donc prendre en compte ce cyle de vie des équipements en modélisant de manière complète les différents cycles et leurs impacts. Je souhaite explorer ces aspects dans le futur et ainsi établir certains ponts avec d'autres domaines de recherche pour croiser des expertises complémentaires.

[3] http://www.greenpeace.org/international/en/publications/Campaign-reports/Climate-Reports/Cool-IT-Leaderboard/

7.3 Bilan personnel

Le métier de chercheur est formidable!

Deux événements majeurs ont déclenché en moi la passion des nouvelles technologies de l'information et de la communication. Le film *Wargames* que j'ai dévoré à 13 ans, armé de mon TO7, plongé dans des lignes de code d'assembleur : j'ai su que je voulais en faire mon métier. Mon arrivée à l'ENS en tant qu'étudiant de Maîtrise et les cours donnés par Michel Cosnard et Yves Robert : j'ai su que je voulais être chercheur en informatique.

"L'habilitation à diriger des recherches sanctionne la reconnaissance du haut niveau scientifique du candidat, du caractère original de sa démarche dans un domaine de la science, son aptitude à maîtriser une stratégie de recherche dans un domaine scientifique ou technologique suffisamment large et de sa capacité à encadrer des jeunes chercheurs."[4] J'espère avoir répondu à ces critères. 16 ans depuis l'obtention de ma thèse ! C'est vrai que j'ai pris mon temps pour écrire cette habilitation. 4 années en tant que Maitre de Conférences, des projets à gérer, des communautés à animer, des évènements à organiser, des thèses à co-encadrer (sans habilitation), une vie de famille à équilibrer sont autant d'arguments qui ont allongé cette durée.

Mais quand je regarde le travail accompli, les résultats obtenus, les compétences mises en œuvre, les personnes que j'ai encadré et avec qui j'ai collaboré pendant ces années, je me rends compte que je suis fier d'avoir mené ce travail collectif. J'ai sans doute aussi eu de la chance. Tous les doctorants que j'ai co-encadrés ont fait preuve de créativité, d'autonomie et d'engouement pour les activités de recherche que nous avons exploré.

Cette habilitation est l'occasion de se pencher sur ce rôle de passage de témoin au cours des années. On entend souvent dire que tous les 10 ans (3 thèses) les chercheurs ont tendance à recommencer sous une autre forme les mêmes activités et thèmes de recherche. Pour ma part je n'ai pas observé ce phénomène et je suis prêt à partir vers de nouvelles aventures scientifiques.

[4] Arrêté paru au Journal Officiel du 29 novembre 1988, modifié le 13 février 1992

Bibliographie

[1] James Ahrens et Becky Springmeyer David Rogers and : Visualization and data analysis at the exascale - a white paper for the national nuclear security administration (nnsa) accelerated strategic computing (asc) exascale environment planning process. Rapport technique LLNL-TR-474731, Lawrence Livermore National Laboratory, juillet 2010. `https://asc.llnl.gov/exascale/exascale-vdaWG.pdf`. 90

[2] D.Scott Alexander, Bob Braden, Carl A.Gunter, Alden W.Jackson, Angelos D.Keromytis, Gary J.Minden et David Wetherall : Active network encapsulation protocol (anep). RFC Draft, Category : Experimental, http://www.cis.upenn.edu/switchware/ANEP/, juillet 1997. 17

[3] Giovanni Aloisio et Sandro Fiore : Towards exascale distributed data management. *Int. J. High Perform. Comput. Appl.*, 23(4):398–400, novembre 2009. 90

[4] Narjess Ayari, Denis Barbaron et Laurent Lefèvre : Evaluating session aware admission control strategies for improving the profitability of service providers. *In The 3rd IEEE Workshop on Enabling the Future Service-Oriented Internet: Towards Socially-Aware Networks - Held in conjunction with IEEE GLOBECOM 2009*, Honolulu, USA, décembre 2009. 33

[5] Narjess Ayari, Denis Barbaron, Laurent Lefèvre et Pascale Vicat-Blanc Primet : A session aware admission control scheme for next generation ip services. *In Fifth annual IEEE Consumer Communications and Networking Conference (IEEE CCNC 2008)*, Las Vegas, USA, janvier 2008. 33

[6] Narjess Ayari, Denis Barbaron et Laurent Lefèvre : Procédés de gestion de sessions multi-flux. France Telecom R&D Patent, juin 2007. 33

[7] Narjess Ayari, Laurent Lefèvre et Denis Barbaron : Design and evaluation of a session-aware admission-control framework for improving service providers profitability. *Journal of Internet Engineering - Special issue on Service Oriented Infrastructures*, 4(1), décembre 2010. 33

[8] Narjess Ayari, Laurent Lefèvre et Denis Barbaron : On improving the reliability of internet services through active replication. *In PDCAT 2008 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies*, pages 259–262, Dunedin, New Zealand, décembre 2008. 33

[9] Narjess Ayari, Pablo Neira Ayuso, Laurent Lefèvre et Denis Barbaron : Towards a dependable architecture for highly available internet services. *In ARES'08 : The Third International Conference on Availability, Reliability and Security*, pages 422–427, Barcelona, Spain, mars 2008. 34

[10] Pablo Neira Ayuso, Rafael M. Gasca et Laurent Lefevre : FT-FW: A cluster-based fault-tolerant architecture for stateful firewalls. *Computers and Security*, 31(4):524–539, juin 2012. 34

[11] D. H. Bailey, E. Barszcz, J. T. Barton, D. S. Browning, R. L. Carter, L. Dagum, R. A. Fatoohi, P. O. Frederickson, T. A. Lasinski, R. S. Schreiber, H. D. Simon, V. Venkatakrishnan et S. K. Weeratunga : The nas parallel benchmarks : summary and preliminary results. *In Proceedings of the 1991 ACM/IEEE conference on Supercomputing*, Supercomputing '91, pages 158–165, New York, NY, USA, 1991. ACM. 93, 100

[12] Daniel Balouek, Alexandra Carpen Amarie, Ghislain Charrier, Frédéric Desprez, Emmanuel Jeannot, Emmanuel Jeanvoine, Adrien Lèbre, David Margery, Nicolas Niclausse, Lucas Nussbaum, Olivier Richard, Christian Pérez, Flavien Quesnel, Cyril Rohr et Luc Sarzyniec : Adding Virtualization Capabilities to Grid'5000. Research Report RR-8026, INRIA, juillet 2012. 3

[13] L.A. Barroso et U. Holzle : The case for energy-proportional computing. *Computer*, 40(12):33–37, 2007. 108, 109

[14] Alessandro Bassi, Micah Beck, Fabien Chanussot, Jean-Patrick Gelas, Robert Harakaly, Laurent Lefèvre, Terry Moore, James Plank et Pascale Vicat-Blanc Primet : Active and logistical networking for grid computing: the e-toile architecture. *The International Journal of Future Generation Computer Systems (FGCS) - Grid Computing: Theory, Methods and Applications*, 21(1):199–208, janvier 2005. Elsevier B.V (ed),ISSN 0167-739X. 30

[15] Alessandro Bassi, Micah Beck, Jean-Patrick Gelas et Laurent Lefèvre : Logistical storage in active networking: a promising framework for network services. *In* Hamid R.Arabnia et Youngsong Mun, éditeurs : *International Conference on Internet Computing 2002 (IC'2002)*, volume 2, pages 209–216, Las Vegas, Nevada, USA, juin 2002. CSREA Press. ISBN: 1-892512-36-x. 47

[16] Alessandro Bassi, Jean-Patrick Gelas et Laurent Lefèvre : Tamanoir-IBP : Adding Storage to Active Networks. *In Fourth Annual International Workshop on Active Middleware Services (AMS 2002), 11th IEEE International Symposium on High Performance Distributed Computing*, pages 27–34, Edinburgh, Scotland, juillet 2002. IEEE Computer Society. ISBN 0-7695-1721-8. 30, 47

[17] Bearstech company. `http://bearstech.com`, 2005. 31

[18] M. Beck, T. Moore, J. Plank et M. Swany : Logistical networking : sharing more than the wires. *In* C. S. Raghavendra S. Hariri, C. A. Lee, éditeur : *Active Middleware Services, Ninth IEEE International Symposium on High Performance Distributed Computing*, pages 140–154, Pittsburgh, Pennsylvania, USA, aug 2000. Kluwer Academic Publishers. ISBN 0-7923-7973-X. 47

[19] Anton Beloglazov, Rajkumar Buyya, Young Choon Lee et Albert Y. Zomaya : A taxonomy and survey of energy-efficient data centers and cloud computing systems. *Advances in Computers*, 82:47–111, 2011. 79, 88

[20] Francoise Berthou, Robert Ferret et Laurent Lefèvre : Consommation énergétique des datacentres : quand la politique européenne s'en mêle. *In JRES 2011 : 9 ème Journées Réseaux de l'enseignement supérieur et de la recherche*, Toulouse, France, novembre 2011. 111, 112

[21] A.P. Bianzino, C. Chaudet, D. Rossi et J. Rougier : A survey of green networking research. *Communications Surveys Tutorials, IEEE*, 14(1):3–20, 2012. 98

[22] Kashif Bilal, SameeU. Khan, SajjadA. Madani, Khizar Hayat, MajidI. Khan, Nasro Min-Allah, Joanna Kolodziej, Lizhe Wang, Sherali Zeadally et Dan Chen : A survey on green communications using adaptive link rate. *Cluster Computing*, pages 1–15, 2012. 98

[23] Kurt Binder, Jürgen Horbach, Walter Kob, Wolfgang Paul et Fathollah Varnik : Molecular dynamics simulations. *Journal of Physics: Condensed Matter*, 16(5):S429, 2004. 100

[24] J. Biswas, J.F Huard, A. Lazar, K. Lim, S. Mahjoub, L.F Pau, M. Suzuki, S. Torstensson, W. Weiugo et S. Weinstein : Application programming interfaces for networks. *IEEE P1520 WG Draft White Paper*, jan 1999. 11

[25] Burton H. Bloom : Space/Time Trade-Offs In Hash Coding With Allowable Errors. *Communications of the ACM*, 13(7):422–426, 1970. 40

[26] Nanette Boden, Danny Cohen, Robert Felderman, Alan Kulawik, Charles Seitz, Jakov Seizovic et Wen-King Su : Myrinet : a gigabit per second local area network. *EEE-Micro*, 1995. 20, 23

[27] R. Bolla, F. Davoli, R. Bruschi, K. Christensen, F. Cucchietti et S. Singh : The potential impact of green technologies in next-generation wireline networks: Is there room for energy saving optimization? *Communications Magazine, IEEE*, 49(8):80 –86, août 2011. 106, 107

[28] Raffaele Bolla et Roberto Bruschi : An open-source platform for distributed linux software routers. *Computer Communications*, 36(4):396 – 410, 2013. 103

[29] Franck Bonnassieux, Robert Harakaly et Pascale Primet : Mapcenter: an open grid status visualization tool. *In ISCA 15th International Conference on parallel and distributed computing systems*, Louisville, Kentucky, USA, septembre 2002. http://mapcenter.in2p3.fr. 27

[30] Faycal Bouhafs, Jean-Patrick Gelas, Laurent Lefèvre, Moufida Maimour, Cong-Duc Pham, Pascale Vicat-Blanc Primet et Bernard Tourancheau : Designing and evaluating an active grid architecture. *The International Journal of Future Generation Computer Systems (FGCS) - Grid Computing: Theory, Methods and Applications*, 21(2): 315–330, février 2005. 30

[31] Aurelien Bouteiller, George Bosilca et Jack Dongarra : Redesigning the message logging model for high performance. *Concurr. Comput. : Pract. Exper.*, 22(16):2196–2211, novembre 2010. 91

[32] B. Bouteiller, P. Lemarinier, K. Krawezik et F. Capello : Coordinated checkpoint versus message log for fault tolerant mpi. *In Cluster Computing, 2003. Proceedings. 2003 IEEE International Conference on*, pages 242–250, 2003. 91

[33] M. Brahma, M. Chaudier, E. Garcia, J.P. Gelas, H. Guyennet, F. Hantz, L. Lefèvre, P. Lorenz et H. Tobiet : TEMIC: a New Cooperative Platform for Industrial Tele-Maintenance. *In DFMA06 : International Conference on Distributed Framework for Multimedia Applications*, Penang, Malaysia, mai 2006. 31

[34] Bob Briscoe : Tunnelling of explicit congestion notification. Rapport technique RFC6040, Internet Engineering Task Force, novembre 2010. 35

[35] S. Browne, J Dongarra, N. Garner, G. Ho et P. Mucci : A portable programming interface for performance evaluation on modern processors. *The International Journal of High Performance Computing Applications*, 14:189–204, 2000. 96

[36] S. Burleigh, A. Hooke, L. Torgerson, K. Fall, V. Cerf, B. Durst, K. Scott et H. Weiss : Delay tolerant networking: an approach to interplanetary internet. *IEEE Communications Magazine*, pages 128–136, juin 2003. 45

[37] F. Cappello, F. Desprez, M. Dayde, E. Jeannot, Y. Jegou, S. Lanteri, N. Melab, R. Namyst, P. Primet, O. Richard, E. Caron, J. Leduc et G. Mornet : Grid'5000: A large scale, reconfigurable, controlable and monitorable grid platform. *In 6th IEEE/ACM International Workshop on Grid Computing, Grid'2005*, Seattle, Washington, USA, novembre 2005. 2, 25, 53

[38] Franck Cappello, Al Geist, Bill Gropp, Laxmikant Kale, Bill Kramer et Marc Snir : Toward exascale resilience. *Int. J. High Perform. Comput. Appl.*, 23(4):374–388, novembre 2009. 90

[39] B. Carpenter et S. Brim : Middleboxes: Taxonomy and Issues. RFC 3234 (Informational), février 2002. 10

[40] Julien Carpentier, Jean-Patrick Gelas, Laurent Lefevre, Maxime Morel, Olivier Mornard et Jean-Pierre Laisne : Compatibleone: Designing an energy efficient open source cloud broker. *In The 2nd International Conference on Cloud and Green Computing (CGC 2012)*, Xiangtan, China, novembre 2012. 83

[41] C. Castillo, G.N. Rouskas et K. Harfoush : Efficient resource management using advance reservations for heterogeneous Grids. *In International Symposium on Parallel and Distributed Processing (IPDPS)*, pages 1–12, 2008. 72

[42] V. Cerf, S. Burleigh, A. Hooke, L. Torgerson, R. Durst, K. Scott, K. Fall et H. Weiss : RFC 4838, Delay-Tolerant Networking Architecture. *IRTF DTN Research Group*, 2007. 45

[43] V. Cerf, Y. Dalal et C. Sunshine : Specification of Internet Transmission Control Program. RFC 675, décembre. 10

[44] Martine Chaudier, Jean-Patrick Gelas et Laurent Lefèvre : Towards the design of an autonomic network node. *In IWAN2005 : Seventh Annual International Working Conference on Active and Programmable Networks*, Nice, France, novembre 2005. 31

[45] Ghislain Landry Tsafack Chetsa : *Profilage du système et leviers verts pour les infrastructures distribuées à grande échelle.* Thèse de doctorat d'informatique, Laboratoire LIP - ENS Lyon, Lyon, France, décembre 2013. 90, 95

[46] G.L. Tsafack Chetsa, L. Lefèvre, J.M. Pierson, P. Stolf et G. Da Costa : Exploiting performance counters to predict and improve energy performance of HPC systems. *Future Generation Computer Systems*, 2013. 96

[47] Action Européenne COST IC 804 : "Energy Effciency in Large Scale Distributed Systems". http://www.cost804.org, 2012. 58

[48] Georges Da-Costa, Jean-Patrick Gelas, Yiannis Georgiou, Laurent Lefèvre, Anne-Cécile Orgerie, Jean-Marc Pierson, Olivier Richard et Kamal Sharma : The green-net framework: Energy efficiency in large scale distributed systems. *In HPPAC 2009 : High Performance Power Aware Computing Workshop in conjunction with IPDPS 2009*, Rome, Italy, mai 2009. 79

[49] S. Deering et R. Hinden : Internet Protocol, Version 6 (IPv6) Specification. RFC 1883, décembre 1995. 9

[50] S. Deering et R. Hinden : *RFC 2460 Internet Protocol, Version 6 (IPv6) Specification.* Internet Engineering Task Force, décembre 1998. 9

[51] UNFPA Fonds des Nations Unies pour la population : Etat de la population mondiale 2011. Rapport technique, 2011. http://foweb.unfpa.org/SWP2011/reports/FR-SWOP2011.pdf. 1

[52] Marcos Dias de Assuncao, Jean-Patrick Gelas, Laurent Lefèvre et Anne-Cécile Orgerie : The green grid5000: Instrumenting a grid with energy sensors. *In 5th International Workshop on Distributed Cooperative Laboratories: Instrumenting the Grid (INGRID 2010)*, Poznan, Poland, mai 2010. 58, 60

[53] UMass DieselNet :`http://prisms.cs.umass.edu/dome/umassdieselnet`. 45

[54] Mohammed Diouri, Olivier Glück et Laurent Lefevre : Smart energy management for greener supercomputing. *ERCIM News*, 92, janvier 2013. 110

[55] Mohammed El Mehdi DIOURI : *Efficacité énergétique dans le calcul très haute performance : application à la tolérance aux pannes et à la diffusion de données.* Thèse de doctorat d'informatique, Laboratoire LIP - ENS Lyon, Lyon, France, septembre 2013. 63, 89, 90

[56] Mohammed El Mehdi Diouri, Manuel F. Dolz, Olivier Glück, Laurent Lefèvre, Pedro Alonso, Sandra Catalán, Rafael Mayo et Enrique S. Quintana-Ortí : Solving some Mysteries in Power Monitoring of Servers: Take Care of your Wattmeters! *In Energy Efficiency in Large Scale Distributed Systems conference (EE-LSDS)*, Vienne, Autriche, avril 2013. 54, 56, 64, 65, 68

[57] Mohammed El Mehdi Diouri, Olivier Glück et Laurent Lefèvre : SESAMES: a Smart-Grid Based Framework for Consuming Less and Better in Extreme-Scale Infrastructures. *In GreenCom 2013 : The 2013 IEEE International Conference on Green Computing and Communications, Beijing, China*, août 2013. 110

[58] Mohammed El Mehdi Diouri, Olivier Glück et Laurent Lefèvre : Your Cluster is not Power Homogeneous: Take Care when Designing Green Schedulers! *In 4^{th} IEEE International Green Computing Conference (IGCC), Arlington, VA USA*, juin 2013. 54, 64, 66

[59] Dropbox. http://www.dropbox.com. 78

[60] Groupe Ecoinfo : *Impacts Ecologiques des Technologies de l'Information et de la Communication : les faces cachées de l'immatérialité.* EDP Sciences, 2012. 111, 112

[61] Karthik Elangovan, Ivan Rodero, Manish Parashar, Francesc Guim et Isaac Hernandez : Adaptive memory power management techniques for hpc workloads. *In 18th International Conference on High Performance Computing, HiPC 2011*, pages 1–11, Bengaluru, India, décembre 2011. IEEE. 98

[62] Akhil Arora et al. : (WS-Management) Specification , décembre 2008. DMTF - Distributed Management Task Force - Document Number: DSP0226 - Final Standard. 43

[63] Karl Czajkowski et al. : The WS-Resource Framework, mai 2004. http://www.globus.org/wsrf/. 42

[64] Yves Eudes : Mon ami robot. *Le Monde*, 2005. Article paru dans l'édition du 02.08.05. 9

[65] Gilles Fedak, Jean-Patrick Gelas, Thomas Hérault, Victor Iniesta, Derrick Kondo, Laurent Lefèvre, Paul Malecot, Lucas Nussbaum, Ala Rezmerita et Olivier Richard : DSL-Lab: a Low-Power Lightweight Platform to Experiment on Domestic Broadband Internet. *In International Symposium on Parallel and Distributed Computing (ISPDC 2010)*, Istanbul, Turkey, juillet 2010. 61, 63

[66] Wu-chun Feng et Kirk Cameron : The green500 list: Encouraging sustainable supercomputing. *Computer*, 40(12):50–55, décembre 2007. 89

[67] Wu-chun Feng, Xizhou Feng et Rong Ge : Green supercomputing comes of age. *IT Professional*, 10(1):17–23, janvier 2008. 53

[68] A. Fernández-Montes, L. Gonzalez-Abril, Juan A. Ortega et Laurent Lefèvre : Smart scheduling for saving energy in grid computing. *Expert Syst. Appl.*, 39(10):9443–9450, août 2012. 74

[69] Ian Foster et Carl Kesselman : Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications*, 11:115–128, 1996. 41

[70] Ian Foster et Carl Kesselman : High-throughput resource management. *In The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, 1998. 41

[71] Benjamin Gaidioz : *Traitements différenciés et marquage adaptatif de paquets pour l'amélioration du transport des flux hétérogènes dans l'Internet.* Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, décembre 2003. 41

[72] Alex Galis, Jean-Patrick Gelas, Laurent Lefèvre et Kun Yang : Active network approach to grid management & services. *In Workshop on Innovative Solutions for Grid Computing - ICCS 2003 Conference*, pages 1103–1113, Melbourne, Australia, juin 2003. LNCS 2658, ISBN 3-540-40195-4. 31

[73] Virginie Galtier, Kevin Mills et Yannick Carlinet : Modeling cpu demand in heterogeneous active networks. *In 2002 DARPA Active Networks Conference and Exposition (DANCE'02)*, San Francisco, CA, mai 2002. 21

[74] Virginie Galtier, Kevin L. Mills, Yannick Carlinet, Stefan Leigh et Andrew Rukhin : Expressing meaningful processing requirements among heterogeneous nodes in an active network. *In Workshop on Software and Performance*, pages 20–28, 2000. 14

[75] Rong Ge, Xizhou Feng, Shuaiwen Song, Hung-Ching Chang, Dong Li et Kirk W. Cameron : Powerpack: Energy profiling and analysis of high-performance systems and applications. *IEEE Trans. Parallel Distrib. Syst.*, 21(5):658–671, 2010. 55, 66

[76] Jean-Patrick Gelas : *Vers la conception d'une architecture de réseaux actifs apte à supporter les débits des réseaux gigabits.* Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, décembre 2003. 12, 41

[77] Jean-Patrick Gelas, Saad El Hadri et Laurent Lefèvre : Towards the design of an high performance active node. *Parallel Processing Letters journal*, 13(2), juin 2003. 31

[78] Jean-Patrick Gelas et Laurent Lefèvre : Tamanoir: A high performance active network framework. *In* C. S. Raghavendra S. Hariri, C. A. Lee, éditeur : *Active Middleware Services, Ninth IEEE International Symposium on High Performance Distributed Computing*,

pages 105–114, Pittsburgh, Pennsylvania, USA, aug 2000. Kluwer Academic Publishers. ISBN 0-7923-7973-X. 17

[79] Jean-Patrick Gelas et Laurent Lefèvre : Mixing high performance and portability for the design of active network framework with java. *In 3rd International Workshop on Java for Parallel and Distributed Computing, International Parallel and Distributed Processing Symposium (IPDPS 2001)*, San Fransisco, USA, avril 2001. 17

[80] Jean-Patrick Gelas, Laurent Lefevre, Teferi Assefa et Mulugeta Libsie : Virtualizing home gateways for large scale energy reduction in wireline networks. *In Electronic Goes Green 2012 (EGG)*, Berlin, Germany, septembre 2012. 106

[81] Jean-Patrick Gelas et Laurent Lefèvre : Towards the design of an active grid. *In* Lecture Notes in Computer Science, éditeur : *Computational Science - ICCS 2002*, volume 2230, pages 578–587, Amsterdam, The Netherlands, avril 2002. ISBN 3-540-43593-X. 30, 41

[82] P. Geoffray, L. Lefèvre, C. Pham, L. Prylli, O. Reymann, B. Tourancheau et R. Westrelin : High-speed LANs: New environments for parallel and distributed applications. *In ACM/IFIP EuroPar'99*, numéro 1685 de LNCS, pages 633–642, Toulouse, France, août 1999. Springer-Verlag. 12

[83] Make IT Green: Cloud Computing and its Contribution to Climate Change. Greenpeace report, 2010. 53

[84] How dirty is your data? Greenpeace report, 2011. 53

[85] Andrew Grimshaw, Adam Ferrari, Fritz Knabe et Marty Humphrey : Legion: An operating system for wide-area computing. *IEEE Computer*, 32(5):29–37, 1999. 41

[86] Jiani Guo, Fang Chen, Laxmi Bhuyan et Raj Kumar : A cluster-based active router architecture supporting video/audio stream transcoding service. *In International Parallel and Distributed Processing Symposium (IPDPS'03)*, Nice, France, avril 2003. 21

[87] Liang Guo et I. Matta : The war between mice and elephants. *In Network Protocols, 2001. Ninth International Conference on*, pages 180–188, 2001. 103

[88] Fabien Hermenier, Xavier Lorca, Jean-Marc Menaud, Gilles Muller et Julia L. Lawall : Entropy: a consolidation manager for clusters. *In Proceedings of the 5th International Conference on Virtual Execution Environments, VEE 2009*, pages 41–50. ACM, mars 2009. 82

[89] Christina Herzog, Laurent Lefevre et Jean-Marc Pierson : Green it for innovation and innovation for green it: The virtuous circle. *In* Magda David Hercheui, Diane Whitehouse, William McIver Jr. et Jackie Phahlamohlaka, éditeurs : *ICT Critical Infrastructures and Society, 10th IFIP TC9 International Conference on Human Choice and Computers (HCC)*, numéro 386, pages 79–89, Amsterdam, septembre 2012. Springer. 111

[90] Christina Herzog, Laurent Lefevre et Jean-Marc Pierson : Link between academia and industry for green it. *In ICT4S Conference : ICT for Sustainability*, pages 259–264, Zurich, Switzerland, février 2013. 111

[91] Helmut Hlavacs, Thomas Treutner, Jean-Patrick Gelas, Laurent Lefevre et Anne-Cecile Orgerie : Energy consumption side-channel attack at virtual machines in a cloud. *In International Conference on Cloud and Green Computing (CGC 2011)*, Sydney, Australia, décembre 2011. 87

[92] Canturk Isci et Margaret Martonosi : Runtime power monitoring in high-end processors: Methodology and empirical data. *In Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 36, Washington, DC, USA, 2003. IEEE Computer Society. 97

[93] Justcloud. http://www.justcloud.com. 78

[94] D. Katabi, M. Handley et C. Rohrs : Congestion control for high bandwidth-delay product networks. *In ACM SIGCOMM*, 2002. 30, 35, 36, 37

[95] Kyong Hoon Kim, Rajkumar Buyya et Jong Kim : Power aware scheduling of bag-of-tasks applications with deadline constraints on dvs-enabled clusters. *In Seventh IEEE International Symposium on Cluster Computing and the Grid (CCGrid 2007), 14-17 May 2007, Rio de Janeiro, Brazil*, pages 541–548, 2007. 66

[96] Ezra Kissel et Martin Swany : The extensible session protocol: A protocol for future internet architectures. Rapport technique, 2012. UNIVERSITY OF DELAWARE TECHNICAL REPORT `http://damsl.cs.indiana.edu/projects/phoebus/kissel_xsp.pdf`. 103

[97] Brad Lawrence et John Jenne : Dell energy smart architecture and power management adoption. Rapport technique, Dell, novembre 2011. 83

[98] L. Lefèvre, C. Pham, P. Vicat-Blanc Primet, B. Tourancheau, B. Gaidioz, J.P. Gelas et M. Maimour : Active networking support for the grid. *In* Noaki Wakamiya Ian W. Marshall, Scott Nettles, éditeur : *IFIP-TC6 Third International Working Conference on Active Networks, IWAN 2001*, volume 2207 de *Lecture Notes in Computer Science*, pages 16–33, octobre 2001. ISBN: 3-540-42678-7. 30, 41

[99] Laurent Lefevre, Olivier Mornard, Jean-Patrick Gelas et Maxime Morel : Monitoring energy consumption in clouds: the compatibleone experience. Poster in Cloud and Green Computing conference (CGC 2011), Sydney, Australia, décembre 2011. 82

[100] Laurent Lefèvre et Jean-Marc Pierson : Just in time entertainment deployment on mobile platforms. *In ICIW'06 : International Conference on Internet and Web Applications and Services*, Guadeloupe, French Caribbean, février 2006. 48

[101] Laurent Lefèvre et Paul Roe : Improving the flexibility of active grids through web services. *In* Australian Computer Society, éditeur : *Fourth Australasian Symposium on Grid Computing and e-Research (AusGrid2006)*, volume 28, pages 3–8, Hobart, Australia, janvier 2006. ISBN 1-920-68236-8. 30, 31, 42

[102] Laurent Lefèvre : *Design of a Parallel Programming Environment based on a Distributed Shared Memory System - Conception et Mise en Oeuvre d'un Environnement de Programmation Parallèle fondé sur un Système de Mémoire Distribuée Virtuellement Partagée.* Thèse de doctorat, Ecole Normale Supérieure de Lyon, France, janvier 1997. 1

[103] Laurent Lefèvre : Heavy and lightweight dynamic network services : challenges and experiments for designing intelligent solutions in evolvable next generation networks. *In* IEEE Society, éditeur : *Workshop on Autonomic Communication for Evolvable Next Generation Networks - The 7th International Symposium on Autonomous Decentralized Systems*, pages 738–743, Chengdu, Jiuzhaigou, China, avril 2005. ISBN : 0-7803-8963-8. 14

[104] Laurent Lefèvre et Jean-Patrick Gelas : Active Web : active networking support for web transport. *In ANTA 2003 : The second International Workshop on Active Networks Technologies and Applications*, pages 147–156, Osaka, Japan, mai 2003. 10

[105] Laurent Lefèvre et Jean-Patrick Gelas : Towards interplanetary grids. *In Workshop on "Next Generation Communication Infrastructure for Deep-Space Communications" held in conjunction with the Second International Conference on Space Mission Challenges for Information Technology (SMC-IT)*, Pasadena, California, juillet 2006. 10, 45

[106] Laurent Lefèvre et Anne-Cécile Orgerie : Designing and evaluating an energy efficient cloud. *The Journal of SuperComputing*, 51(3):352–373, mars 2010. 79, 80

[107] Laurent Lefèvre, Jean-Marc Pierson et Sidali Guebli : Collaborative web caching with active networks. *In International Working Conference on active networks (IWAN2003)*, volume LNCS 2982, pages 80–91, Kyoto, Japan, décembre 2003. 48

[108] Laurent Lefèvre et Aweni Saroukou : Active network support for deployment of java-based games on mobile platforms. *In* IEEE Computer Society, éditeur : *The First International Conference on Distributed Frameworks for Multimedia Applications (DFMA'2005)*, pages 88–95, Besancon, France, février 2005. 48

[109] Eric Lemoine : *Nouvelles fonctions dans les interfaces de communication pour l'augmentation des performances réseau des machines multi-processeur.* Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, juillet 2004. 22

[110] Eric Lemoine, Cong-Duc Pham et Laurent Lefèvre : Packet Classification in the NIC for Improved SMP-based Internet Servers. *In IEEE Proceedings of the International Conference on Networking (ICN 2004)*, Guadeloupe, French Caribbean, février 2004. 22, 23

[111] Min Yeol Lim, Allan Porterfield et Robert Fowler : Softpower: fine-grain power estimations using performance counters. *In Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, HPDC '10, pages 308–311, New York, NY, USA, 2010. ACM. 96

[112] D. M. Lopez Pacheco et C. Pham : Robust Transport Protocol for Dynamic High-Speed Networks: enhancing the XCP approach. *In IEEE MICC-ICON*, 2005. 36

[113] Dino M. Lopez Pacheco, Laurent Lefèvre et Cong-Duc Pham : Fairness issues when transferring large volume of data on high speed networks with router-assisted transport protocols. *In High Speed Networks Workshop 2007, in conjunction with IEEE INFOCOM 2007*, Anchorage, Alaska, USA, mai 2007. 35, 39

[114] Dino M. Lopez Pacheco, Laurent Lefèvre et Cong-Duc Pham : Lightweight Fairness Solutions for XCP and TCP Cohabitation. *In IFIP/TC6 Networking 2008*, pages 715–726, Singapore, mai 2008. 35, 39

[115] Dino M. Lopez Pacheco, Cong-Duc Pham et Laurent Lefèvre : XCP-i : eXplicit Control Protocol for heterogeneous inter-networking of high-speed networks. *In Globecom 2006*, San Francisco, California, USA, novembre 2006. 30, 35, 38

[116] S. H. Low, L. Andrew et B. Wydrowsk : Understanding XCP: Equilibrium and Fairness. *In IEEE Infocom*, 2005. 36

[117] Edgar Magana, Laurent Lefèvre, Masum Hasan et Joan Serrat : Snmp-based monitoring agents and heuristic scheduling for large scale grids. *In Grid computing, high-performAnce and Distributed Applications (GADA'07)*, Vilamoura, Algarve, Portugal, novembre 2007. 31

[118] Edgar Magana, Laurent Lefèvre et Joan Serrat : Autonomic management architecture for flexible grid services deployment based on policies. *In Architecture of Computing Systems - ARCS 2007*, volume 4415, pages 157–170, ETH, Zurich, Switzerland, mars 2007. Springer Berlin / Heidelberg. 31

[119] Moufida Maimour : *Design, analysis and validation of router-assisted reliable multicast protocols in wide area networks*. Thèse de doctorat d'informatique, Université Claude Bernard Lyon1 - Laboratoire LIP - ENS Lyon, Lyon, France, décembre 2003. 41

[120] Bartek Wydrowskilete Martin Suchara, Ryan Witt : TCP MaxNet: Implementation and Experiments on the WAN in Lab. *In IEEE International Conference on Networks*, novembre 2005. 35

[121] Matthew L. Massie, Brent N. Chun et David E. Culler : The ganglia distributed monitoring system: design, implementation, and experience. *Parallel Computing*, 30(5-6):817–840, 2004. 83

[122] Nick McKeown, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker et Jonathan Turner : Openflow: enabling innovation in campus networks. *SIGCOMM Comput. Commun. Rev.*, 38(2):69–74, mars 2008. 104

[123] Yiduo Mei, Ling Liu, Xing Pu, Sankaran Sivathanu et Xiaoshe Dong : Performance analysis of network i/o workloads in virtualized data centers. *IEEE T. Services Computing*, 6(1):48–63, 2013. 104

[124] Jeffrey C. Mogul et K. K. Ramakrishnan : Eliminating receive livelock in an interrupt-driven kernel. *In Proceedings of the 1996 annual conference on USENIX Annual Technical Conference*, ATEC '96, pages 9–9, Berkeley, CA, USA, 1996. USENIX Association. 23

[125] Message Passing Interface Forum : MPI : A Message-Passing Interface Standard. juin 1995. 91

[126] Andreas Müller : *Analysis and Control of Middleboxes on the Internet*. Dissertation, Network Architectures and Services, Department of Computer Science - Technische Universitat Munchen, Munich, Allemagne, juillet 2013. 10, 11

[127] Pablo Neira Ayuso, Rafael M. Gasca et Laurent Lefèvre : Demystifying cluster-based fault-tolerant firewalls. *IEEE Internet Computing : Special Issue on Unwanted Traffic*, 13(6):30–37, novembre 2009. 34

[128] Pablo Neira Ayuso, Rafael M. Gasca et Laurent Lefèvre : Communicating between the kernel and user-space in linux using netlink sockets. *Journal Software: Practice and Experience*, 40(9):797–810, août 2010. 34

[129] B. Nordman et K. Christensen : Proxying: The next step in reducing it energy use. *Computer*, 43(1):91–93, 2010. 79

[130] Open networking foundation. https://www.opennetworking.org/. 104

[131] Anne-Cécile Orgerie : *An Energy-Efficient Reservation Framework for Large-Scale Distributed Systems*. Thèse de doctorat, École Normale Supérieure de Lyon - France, septembre 2011. 75

[132] Anne-Cécile Orgerie et Laurent Lefèvre : A year in the life of a large-scale experimental distributed system: usage of the Grid'5000 platform in 2007. Research Report 6965, INRIA, avril 2009. 54

[133] Anne-Cécile Orgerie et Laurent Lefèvre : When Clouds become Green: the Green Open Cloud Architecture. *In Parco2009 : International Conference on Parallel Computing*, pages 228–237, Lyon, France, septembre 2009. 79

[134] Anne-Cécile Orgerie et Laurent Lefèvre : A year in the life of a large-scale experimental distributed system: usage of the Grid'5000 platform in 2008. Research Report 7481, INRIA, décembre 2010. 54

[135] Anne-Cécile Orgerie, Laurent Lefèvre et Jean-Patrick Gelas : Chasing gaps between bursts : Towards energy efficient large scale experimental grids. *In PDCAT 2008 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies*, pages 381–389, Dunedin, New Zealand, décembre 2008. 73, 76

[136] Anne-Cécile Orgerie, Laurent Lefèvre et Jean-Patrick Gelas : Save watts in your grid: Green strategies for energy-aware framework in large scale distributed systems. *In ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems*, pages 171–178, Melbourne, Australia, décembre 2008. 73, 75

[137] Anne-Cécile Orgerie, Laurent Lefèvre et Jean-Patrick Gelas : Demystifying energy consumption in grids and clouds. *In The Work in Progress in Green Computing (WIPGC) Workshop, in conjunction with the first IEEE sponsored International Green Computing Conference*, pages 335–342, Chicago, USA, août 2010. 54, 64, 66

[138] Anne-Cécile Orgerie, Marcos Dias de Assuncao et Laurent Lefèvre : *Energy Aware Clouds*, chapitre "Grids, Clouds and Virtualization" - Massimo Cafaro and Giovanni Aloisio editors, pages 145–170. Springer Book, octobre 2010. ISBN : 978-0-85729-048-9. 63

[139] Anne-Cécile Orgerie et Laurent Lefèvre : Energy-efficient overlay for data transfers in private networks. *In IEEE International Conference on Networks (ICON 2011)*, Singapore, décembre 2011. 74

[140] Anne-Cécile Orgerie et Laurent Lefèvre : Eridis: Energy-efficient reservation infrastructure for large-scale distributed systems. *Parallel Processing Letters*, 21:133–154, juin 2011. 72

[141] Anne-Cécile Orgerie, Laurent Lefèvre et Isabelle Guérin-Lassous : On the energy efficiency of centralized and decentralized management for reservation-based networks. *In IEEE Global Communications Conference (GLOBECOM 2011)*, Houston, USA, décembre 2011. IEEE Computer Society Press. 74

[142] Teunis J. Ott, T. V. Lakshman et Larry H. Wong : SRED: Stabilized RED. *In INFOCOM*, pages 1346–1355, 1999. 40

[143] C. Palansuriya, M. Buchli, K. Kavoussanakis, A. Patil, C. Tziouvaras, A. Trew, A. Simpson et R. Baxter : End-to-End Bandwidth Allocation and Reservation for Grid applications. *In Conference on Broadband Communications, Networks and Systems (BROADNETS)*, pages 1–9, 2006. 72

[144] Subharthi Paul, Jianli Pan et Raj Jain : Architectures for the future networks and the next generation internet: A survey. *Comput. Commun.*, 34(1):2–42, janvier 2011. 103

[145] A. Pentland, R. Fletcher et A. Hasson : Daknet: rethinking connectivity in developing nations. *Computer*, 37(1):78–83, 2004. 45

[146] J. Postel : Transmission Control Protocol. RFC 793 (Standard), sep 1981. Updated by RFC 3168. 10

[147] K. Ramakrishnan, S. Floyd et D. Black : The Addition of Explicit Congestion Notification (ECN) to IP. RFC 3168 (Proposed Standard), septembre 2001. 35

[148] Joseph Rice : SeaWeb acoustic communication and navigation network. *In Proceedings of the International Conference on Underwater Acoustic Measurements: Technologies and Results. Heraklion, Greece.*, juin 2005. 45

[149] Scot Rixner : Network virtualization: Breaking the performance barrier. *Queue*, 6(1): 37:36–37:ff, janvier 2008. 104

[150] E. Rohmer, G. Reina, G. Ishigami, Keiji Nagatani et Kazuya Yoshida : Action planner of hybrid leg-wheel robots for lunar and planetary exploration. *In Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3902–3907, 2008. 45

[151] Frank La Rue : Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, . Rapport technique, United Nations - Human Rights Council, mai 2011. 2

[152] J. H. Saltzer, D. P. Reed et D. D. Clark : End-to-end arguments in system design. *ACM Trans. Comput. Syst.*, 2(4):277–288, 1984. 10

[153] Benjamin Serebrin et Daniel Hecht : Virtualizing performance counters. *In Euro-Par Workshops (1)*, pages 223–233, 2011. 97

[154] S. Sharma, Chung-Hsing Hsu et Wu-chun Feng : Making a case for a green500 list. *In Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International*, page 8 pp., avril 2006. 89

[155] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang et J. G. Powers : A description of the advanced research WRF version 2. *NCAR Tech Note*, NCAR/TN-468+STR, 2005. 100

[156] B. Sotomayor, R.S. Montero, I.M. Llorente et I. Foster : Resource Leasing and the Art of Suspending Virtual Machines. *In Conference on High Performance Computing and Communications (HPCC)*, pages 59–68, 2009. 72

[157] M. Steinder, I. Whalley, J.E. Hanson et J.O. Kephart : Coordinated management of power usage and runtime performance. *In Network Operations and Management Symposium, 2008. NOMS 2008. IEEE*, pages 387–394, 2008. 66

[158] Sugarsync. http://www.sugarsync.com. 78

[159] Nishikado Takashi, Koizumi Minoru et Oochi Hideo : Large-scale high-quality communication service solution using active network technology. *In Hitachi Review*, volume 49, décembre 2000. 42

[160] David Tennenhouse et David Wetherall : Towards an active network architecture. *Computer Communications Review*, 26(2):5–18, avril 1996. 11

[161] Douglas Thain, Todd Tannenbaum et Miron Livny : Distributed computing in practice: the condor experience. *Concurrency - Practice and Experience*, 17(2-4):323–356, 2005. 41

[162] Ghislain Landry Tsafack, Laurent Lefevre et Jean Patrick Gelas : On applying dtns to a delay constrained scenario in wired networks. *In International Workshop on Opportunistic and Delay/disruption-Tolerant Networking in conjunction with the 14th International Symposium on Wireless Personal Multimedia Communications*, Brest, France, octobre 2011. 45

[163] Ghislain Landry Tsafack Chetsa, Laurent Lefèvre et Patricia Stolf : A Three Step Blind Approach for Improving HPC Systems' Energy Performance . *In Energy Efficiency in Large Scale Distributed Systems conference (EE-LSDS), Vienna*. Springer, avril 2013. 95

[164] Nedeljko Vasic, Srinidhi Kuntimaddi et Dejan Kostic : One bit is enough: a framework for deploying explicit feedback congestion control protocols. *In Proceedings of the First international conference on COMmunication Systems And NETworks*, COMSNETS'09, pages 503–511, Piscataway, NJ, USA, 2009. IEEE Press. 35

[165] P. Vetter, T. Ayhan, K. Kanonakis, B. Lannoo, K.L. Lee, L. Lefevre, C. Monney, F. Saliou et X. Yin : Towards energy efficient wireline networks, an update from greentouch. *In CLEO-PR & OECC 2013 PS : The 18th OptoElectronics and Communications Conference*, Kyoto, Japan, juillet 2013. 107

[166] P. Vetter, L. Lefevre, L. Gasca, K. Kanonakis, L. Kazovsky, A. Lee, C. Monney, X. Qiu, F. Saliou et A. Wonfor : Research roadmap for green wireline access. *In Workshop on Green Communications and Networking during IEEE ICC'12*, pages 5941–5945, Ottawa, Canada, juin 2012. IEEE. 107

[167] P. Vicat-Blanc Primet, P. d'Anfray, C. Blanchet, L. Bobelin, O. Boudeville, F. Chanussot, V.D. Cung, Y. Denneulin, A. Djerrah, M. Drabik, Y. Jegou, P.N. Hyunh, N. Lacorne, D. Lacoste, B. Le Cun, L. Lefèvre, D. Mateo, R. Metery, Y. Meurdesoif, J.-C. Mignot, A. Pajot, C.D. Pham, R. Revire, G. Romier et I. Touche : Réflexions et propositions pour une grille haute performance. Rapport technique, Scientific Report of RNTL e-Toile Project, 2003. http://www.urec.cnrs.fr/etoile/RS-etoile.pdf. 30

[168] William Voorsluys, James Broberg, Srikumar Venugopal et Rajkumar Buyya : Cost of virtual machine live migration in clouds: A performance evaluation. *In Proceedings of the 1st International Conference on Cloud Computing*, CloudCom '09, pages 254–265, Berlin, Heidelberg, 2009. Springer-Verlag. 63

[169] P. Werstein : An experimental network proxy for power managed end nodes. *In Advanced Information Networking and Applications - Workshops, 2008. AINAW 2008. 22nd International Conference on*, pages 668–674, 2008. 79

[170] David Wetherall : Active network vision and reality: lessons from a capsule-based system. *Operating Systems Review*, 34(5):64–79, décembre 1999. 17

[171] David Wetherall, John Guttag et David Tennenhouse : ANTS : a toolkit for building and dynamically deploying network protocols. *In IEEE OPENARCH '98*, April 1998. 12

[172] David J. Wetherall : Ants : Network services without the red tape. *Computer*, 32(4):42, April 1999. 9

[173] Thomas Willhalm : Intel® performance counter monitor - a better way to measure cpu utilization. http://software.intel.com/en-us/articles/intel-performance-counter-monitor-a-better-way-to-measure-cpu-utilization, août 2012. 96

[174] B. Wydrowski et M. Zukerman : Maxnet: a new network congestion control architecture for max-min fairness. *In Communications, 2003. ICC '03. IEEE International Conference on*, volume 1, pages 132–136 vol.1, 2003. 35

[175] Yong Xia, L. Subramanian, I. Stoica et S. Kalyanaraman : One more bit is enough. *Networking, IEEE/ACM Transactions on*, 16(6):1281–1294, 2008. 35

[176] Derek Leonard Yueping Zhang et Dmitri Loguinov : JetMax: Scalable Max-Min Congestion Control for High-Speed Heterogeneous Networks. *In INFOCOM*, avril 2006. 35

[177] Wensong Zhang : Linux Virtual Server for Scalable Network Services. *In Ottawa Linux Symposium*, 2000. http://www.linuxvirtualserver.org. 19

[178] Y. Zhang et T. Henderson : An Implementation and Experimental Study of the eXplicit Control Protocol (XCP). *In IEEE Infocom*, 2005. 36

[179] Lei Zhu, Jianhua Gu, Tianhai Zhao et Yunlan Wang : Research on log pre-processing for exascale system using sparse representation. *Grid and Pervasive Computing*, 7861:336–347, 2013. 90

Annexes

A

CV Détaillé

CURRICULUM VITÆ

Laurent Lefèvre

Date de naissance : 23 décembre 1970
Nationalité française

STATUT

Chargé de recherches 1 ère classe INRIA
Equipe INRIA Avalon
Laboratoire LIP
Ecole Normale Supérieure de Lyon
– 46, allée d'Italie - 69364 LYON Cedex 07
Tel : +33(0)4 72 72 82 28 - Fax : +33 (0)4 72 72 80 80
Mobile : +33 (0) 6 62 50 26 08
e-mail : laurent.lefevre@ens-lyon.fr
`http://perso.ens-lyon.fr/laurent.lefevre`

EXPERIENCE PROFESSIONNELLE

Depuis Nov 2001	**Chargé de recherches INRIA** Depuis 2013 : Membre de l'équipe INRIA Avalon 2003-2012 : Membre de l'équipe INRIA RESO 2001-2002 : Responsable de l'action INRIA RESO 2001-2002 Directeur du Laboratoire RESAM
Sept 1997 à Oct 2001	**Maître de Conférence** Université Claude Bernard Lyon 1, France. Laboratoire RESAM Membre de l'action INRIA RESO. Co-responsable du DESS Réseaux - Université Claude Bernard Lyon 1
Fev. 1997 à Aout 1997	**Post-doc**. Rice University Houston, Texas.
Sept. 1993 à Janvier 1997	**Thèse** en informatique. Ecole Normale Supérieure de Lyon (ENS), France.
Juin. 1995 à Déc 1996	**Formateur SUN**. Administration/ Utilisation Solaris SUN Formation France
Nov. 1995 à Sept. 1996	**Service National**.

CURSUS UNIVERSITAIRE

Juin 1993	**D.E.A. d'Informatique Fondamentale**. Ecole Normale Supérieure de Lyon

1 Responsabilités collectives

- Membre du CNU (Conseil National des Universités) - Section 27 - Informatique (2009-2011)
- Membre du comité de direction du Pole Systèmes du GdR ASR (Architecture, Système et Réseaux) (depuis 2013)
- Membre de *executive board* et Working group chair (2009-2011) du consortium GreenTouch
- Working Group Chair et représentant national dans l'Action Européeene IC804 sur l'efficacité énergétique dans les systèmes distribués à grande échelle (2009-2013)
- Responsable des Stages M2 Recherche Informatique, ENS-Lyon (depuis 2007)
- Directeur du laboratoire RESAM, jeune équipe JE 2269 de l'université Claude Bernard du 1/1/2001 au 1/9/2002.
- Responsable de l'action INRIA RESO du 1/1/2001 au 1/9/2002.
- Membre de commissions de spécialistes :
  - Membre titulaire de la commission de spécialistes de l'Université Lumière Lyon2, Section 27 (2005-2008)
  - Membre titulaire de la commission de spécialistes de l'Université Antilles-Guyane, Section 27 (2013, 2005-2008)
  - Membre suppléant de la commission de spécialistes de l'Université Jean-Monnet, Saint-Etienne, Section 27 (2005-2008)
- Editor associé de IEEE Transactions on Cloud Computing (TCC) depuis 2013
- Membre de Steering Comittee
  - IEEE TCSC Technical Area in Green Computing , depuis 2010
  - CCGrid : IEEE/ACM International Symposium on Cluster Cloud and Grid Computing (2004-2013)
  - ICPS2007 : IEEE International Conference on Pervasive Services (2006-2008)
  - IWAN06 : Eight International Workshop on Programmable and Active Networks, Paris, France, 25-29 Septembre 2006 pendant la conférence Autonomic Networking 2006

2 Encadrement d'activités de recherche

2.1 Doctorants : sujets de thèse et position actuelle

1. **Mohammed el Mehdi Diouri** : "**Energy efficiency in HPC : application on fault tolerance and data distribution**", co-dirigé avec Olivier Gluck et Isabelle Guerin Lassous (2010-2013) - Actuellement enseignant chercheur et administrateur du Groupe IGA Casablanca (Ecole d'ingénieurs et de management, Maroc).
2. **Ghislain Landry Tsafack Chetsa** : "**Energy profiling and Green Leverages**", Hemera/ALADDIN, co dirigé avec Jean-Marc Pierson et Patricia Stolf (IRIT, Toulouse) (2010-2013) - Actuellement ATER à l'Université Claude Bernard Lyon1.
3. **Anne-Cécile Orgerie** : "**Energy aware large scale systems for efficient communication and computing**", co-dirigée avec Isabelle Guerin Lassous (Univ. of Lyon) (2008-2011) - Actuellement Chargée de Recherche CNRS dans le laboratoire IRISA (Rennes)
4. **Ayari Narjess** : "Contribution to session aware frameworks for next generation Internet services" - Thèse CIFRE avec France Telecom Research and Development, co-dirigée avec Denis

Barbaron et Pascale Primet (2005-2008) - Actuellement Ingénieure chez Orange

5. **Dino Martin Lopez Pacheco** : "**Contributions aux protocoles de transport hautes performances assistés par les routeurs**" (2006-2008) - Co-encadrement avec Cong-Duc Pham (Université de Pau) - Actuellement Maitre de Conférences à l'Université de Nice
6. **Eric Lemoine** : "**Déploiement dynamique de services sur cartes réseaux programmables**" - (Mai 2001 - Juillet 2004) - Co-Encadrement avec C.D. Pham - Actuellement Ingénieur dans la société CamptoCamp
7. **Jean-Patrick Gelas** : "**Vers la conception d'une architecture de réseaux actifs apte à supporter les débits des réseaux gigabits**" - Co-dirigé avec B. Tourancheau (2000 - 2003) - Actuellement Maitre de Conférences à l'Université Claude Bernard, Lyon1.

2.2 PostDoc :

– **Marcos Dias de Asuncao** : "**Energy aware large scale systems for efficient communication and computing**", dans l'action ARC GREEN-NET project (2009-2010) et dans PrimeEnergyIT European Project (2010-2011) - Actuellement Ingénieur chez IBM Brésil (Sao Paulo)

2.3 Recrutement d' Ingénieurs experts INRIA

1. Francois Rossigneux : "Energy efficiency in OpenStack based Clouds for HPC as a Service", Ingénieur Expert INRIA, Projet FSN XLCLOUD, depuis le 1/1/2013
2. Julien Carpentier : "Energy efficient monitoring and exposing" Ingenieur Expert INRIA, FUI Compatible One Project, 1/1/2012-31/12/2012
3. Maxime Morel : "Energy efficient Open Source Cloud" Ingenieur Expert INRIA, FUI Compatible One Project, 15/9/2011-15/9/2012
4. Olivier Mornard : "Green Cloud" Ingenieur Expert INRIA, FUI Compatible One Project, 1/6/2011-31/5/2012
5. Olivier Mornard : "Service Enablers in Virtual Networks in Autonomic Internet" Ingenieur Expert INRIA, Autonomic Internet Project, 1/1/2010-31/6/2010
6. Abderhaman Cheniour : "Service Enablers Plane : Deployment of autonomic services in Autonomic Internet", Ingenieur Expert INRIA, Autonomic Internet Project, 15/7/2008-15/7/2009
7. Augustin Ragon : "Project management and dissemination activities", Ingenieur Expert INRIA, OGF-Europe European Project, 2009-2010
8. Pierre Bozonnet : "Network adaptation and large scale deployment", Ingénieur Expert INRIA, Projet RNRT Temic, 1/3/2006-31/7/2006
9. Jean-Patrick Gelas : "Lightweight network functionnalities and network support", Ingénieur Expert INRIA, Projet RNRT Temic, 17/1/2005-17/1/2006
10. Martine Chaudier : "Multimedia and adaptive programmable network services", Ingénieur Expert INRIA, Projet RNRT Temic, 15/3/2005-15/3/2006
11. Saad El Hadri : "Experimenting and testing high performance active router around VTHD backbone", Ingénieur Expert INRIA, Projet RNRT VTHD++, 1/4/2002-15/11/2003
12. Roland Westrelin : "High performance communications libraries on top of Windows 2000", Ingénieur Expert INRIA, Projet Microsoft, 2002

2.4 Stage de DEA/Master

- Tsafack Chetsa Ghislain Landry : "Green Internet networks", M2 IFI Hanoi, Vietnam, 1/4/2010-1/12/2010, co-dirigé avec Jean-Patrick Gelas
- Roya Golchay : "Energy reduction in Internet Networks", Master 2 - PFE INSA, 1/3/2010-15/7/2010, co-dirigé avec Olivier Gluck
- Barbara Walter : "Etude de faisabilité de la mise en oeuvre d'une architecture visant à proposer des solutions d'ordonnancement basées sur les technologies green dans un cadre de serveurs d'applications sur la grille", Master 2 - PFE INSA, 1/2/2010-15/9/2010, co-dirigé avec Eddy Caron
- Sylvère Tanaugh Ekponon : "Towards Green HPC nodes", M2 CCI, 03/2009-1/10/2009, co-dirigé avec Jean-Patrick Gelas et Anne-Cécile Orgerie
- Alejandro Fernandez : *"Prediction Models for Energy Efficiency in Large Scale Distributed Systems"*, Universit" de Seville, Espagne, Novembre 2008
- Anne-Cécile Orgerie : *"Energy aware large scale systems for efficient communication and computing"*, Master Recherche, 15/10/2007-15/7/2008
- Damien Nicolet : *"Large scale autonomous networking"*, Master Recherche, 6/2/2006-31/7/2006
- Jean-Paul Corvo : *"Lightweight network software functionalities inside large scale platforms : impact on Grid infrastructure"*, DEA, DIF, ENS Lyon, 1/2/04-15/7/04
- Pablo Neira : *"High availability and fault tolerant techniques for networking equipments"*, PFE-DEA, INSA Lyon, 1/12/03-31/7/04
- SidaLi Guebli : *"Designing collaborative cache with high performance active nodes"*, DEA (1/2/2003 - 15/7/2003), co-encadrement avec J.M. Pierson LISI INSA
- Julien Laganier : *"Sécurité et SUCv"* - DEA Informatique Fondamentale de Lyon - Février à Juillet 2002
- Jean-Patrick Gelas : *"Flux intelligents haute performance"* - DEA Informatique Fondamentale de Lyon - Février à Juillet 2000 -
- Alice Bonhomme : *" Cohérence sur groupes pour un partage efficace des données - Application aux réseaux de communication"* - DEA Informatique de Lyon - Mars à Juillet. 1998

2.5 Stages de Licence - Maîtrise

- Julien Delaborde : "Simulating energy consumption in large scale networks", M1 ENS-Lyon, 1/3/2010-15/6/2010
- Walid El Dahabi : *Network experiments on the DSLLAB platform (DSLLAB project)* (24/4/07-30/9/07)
- Anne-Cécile Orgerie : *"High speed transport protocols based on XCP (eXplicit Control Protocol)"*, Stage L3, ENS Lyon, (5/6/2006-22/7/06).
- Pablo Pazos : *"Large scale programmable network deployment on Grid5000"*, Master informatique Lyon1 Univ. (10/5/06-10/8/06).
- Chien-Jon Soon : *"Active Networks and Web Services"*, PLAS Group, Queensland University of Technology, Brisbane, Australie (28/7/2005-30/10/2005).
- Grégoire Locqueneux : *""Interfacage de reseaux programmables avec middleware de Grille"*, Maîtrise Informatique (1/12/2003-29/2/2004).
- Aweni Saroukou : *"Support dynamique de réseaux pour le déploiement de jeux Java sur téléphones*

portables", Maîtrise Informatique (1/12/2003-29/2/2004).

- Pierpaolo Giacomin : *"Efficient load balancing scheduling for cluster-based active routers"*, Stage INSA (1/4/2003 - 31/7/2003).
- J. Guilloux : *"Support des OS dans les routeurs actifs logiciels"*, Maîtrise Informatique (1/12/2001-28/2/2002).
- L. ElMalih *"Multimédia et transport"*, Maîtrise Informatique(1/12/2001-28/2/2002).
- E. Degoute : *"Conception et mise en oeuvre d'un environnement de MDVP sur réseau à capacité d'adressage SCI"*, Maîtrise Informatique, 1999.
- R. Herilier "*Editeur cooperatif*", Maîtrise Informatique, 1999.
- J.P. Gelas : "*Jeux distribués sans serveur sur réseaux longue distance*", Maîtrise Informatique, 1998.
- A. Said Ahmed : "*Réalisation d'un éditeur coopératif distribué*", Maîtrise Informatique, 1998.
- H. Tubert : "*Transport de flux video MPEG sur Réseaux Haut Débit*", Maîtrise Informatique, 1998.
- S. Oranger "*Outils pour le Grid Computing*", Maîtrise Informatique, 1999.
- F. Goffinet "*Gestion de cluster* , Maîtrise Informatique, 1999.
- Vincent Vanackere : "*Développement d'une MDVP au dessus d'un réseau Myrinet*", Magistère ENS Lyon (1997-1998)

3 Mobilité

- Séjours Longue durée : chercheur invité à l'Université d'Otago, Dunedin Nouvelle Zélande (Juillet-Septembre 2007), équipe de Z. Huang - chercheur invité au Queensland University of Technology, Brisbane, Australie, (Mai-Octobre 2005), équipe de P. Roe
- Séjour Post-doctoral à Rice University, Houston USA dans l'équipe "Système MDVP" du Professeur W. Zwaenepoel sur "la conception et la mise en oeuvre de nouveaux systèmes de MDVP persistants", Avril à Septembre 1997
- Séjours de recherche de plusieurs semaines : Université de Melbourne, Australie (2003), Université de Linz (1999-2000), Université d'Oujda, Maroc (1998), Université de Loughborough, Angleterre (1996), TUM Munich, Allemagne (1995)

4 Valorisation et transfert technologique : participation à des projets de recherche

4.1 Projets internationaux

- 2010 - : **Projet GreenTouch** : sur l'efficacité énergétique dans les réseaux de communications - Membre de l'*executive Board*- Co-Chair du groupe de travail sur les réseaux filaires - Représentant pour l'INRIA dans ce consortium ;
- 2010-2013 : **Projet ANR-JST FP3C Franco-Japonais** : sur "Framework and Programming for Post Petascale Computing" - Activités sur l'efficacité énergétique avec le laboratoire IRIT, Toulouse
- **Programme d'Actions Integrees Fast** : responsable de ce projet sur "Proposition d'une architecture de Web Services fondée sur les réseaux programmables" avec Queensland University of Technology, Brisbane, Australie (2005-2006)

- **Projet NSF-INRIA** : responsable de ce projet sur "Active Grid middleware" avec l'équipe de Craig Lee - AeroSpace Organization, Los Angeles, USA (2004-2006)
- 1998 : **Projet KIT (Keep in Touch)** : Participant à ce projet en collaboration avec l'Université d'Oujda (Maroc).

4.2 Projet Européens

- 2013-2015 : **Projet STAR (SwiTching And tRansmission)** , CHIST-ERA (European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net) : Responsable de l'activité sur les couches logicielles réseaux optimisées pour l'efficacité énergétique
- 2010-2012 : **Projet PrimeEnergyIT**, Intelligent Energy Europe : Reponsable de l'activité sur la consommation des solutions de stockage
- 2009-2013 : **Action Européenne COST Action IC0804** sur l'Efficacité Energétique dans les systèmes distribués à grande échelle - Représentant Français - Responsable de l'animation d'un groupe de travail
- 2010-2012 : **Projet SPEC de EuroNF JRA.S.1.44** sur "Security and Privacy Concerns in Energy Efficient Computing"
- 2008-2010 : **Projet Européen "Autonomic Internet"** (ICT-2007.1.1 The Network of the Future) : membre et responsable de workpackage - Encadrement d'un ingénieur expert INRIA
- 2008-2010 : **Project Européen "OGF-Europe"** (SSA) : membre de ce projet qui favorise la standardisation dans les Grilles de calcul
- 1998-2000 puis 2001-2003 : **Programme d'Actions Intégrées Amadeus** : Responsable de mise en place d'une action bilatérale franco-autrichienne avec l'Université de Linz, Autriche (équipe de J. Volkert) sur le thème "Déport de services dans des cartes réseaux programmables" - Programme accepté et financé par le Ministère des Affaires Étrangères ;
- 1995-1996 : **Projet européen EUROTOPS** Participant à ce projet qui visait le développement d'une architecture massivement parallèle (dans le cadre du Laboratoire des Hautes Performances en Calcul) et de l'environnement logiciel nécessaire à son utilisation
- 1996 : **Projet Européen Esprit NATHAN** (Nanotechnological and Holographic Methods for Real-Time Pattern Recognition) suivie d'une collaboration avec l'Université de Loughborough (Grande-Bretagne), .
- 1996-1997 : **PAI Procope** avec l'Institut für Informatik München de l'Université de Munich (Allemagne) dans une collaboration avec l'équipe de parallélisme dans le cadre de la réalisation d'outils logiciels communs

4.3 Projets Nationaux

- **Projet Grid5000** : membre et participant de ce projet depuis 2004. Responsable du site de Lyon. Membre du comité de direction.
- **Projet FSN XLCloud Project** : Responsable de l'activité "Efficacité énergétique dans une infrastructure OpenStack" - Recrutement d'un ingénieur expert INRIA (2012-2014)
- **Projet FUI CompatibleOne** : Responsable de la tache "Efficacité Énergétique et Gestion dans le Cloud" - Recrutement de trois ingénieurs expert INRIA (2010-2012)
- **Groupement de Service Eco-info** : membre de ce projet depuis 2010

- **Action d'Envergure Inria Hemera** : membre de ce projet depuis 2010 - Responsable de Groupe de travail et d'un défi scientifique sur le profilage énergétique des applications à grande échelle
- **Action de Recherche Collaborative INRIA GREEN-NET** : leader de ce projet sur l'économie d'énergie dans les systèmes distribués à grande échelle (2008-2010)
- **Projet ANR "Jeunes Chercheurs" DSLLAB** : membre et participant de ce projet (2005-2008)
- **ACI Masse de Données GridExplorer (GDX)** : membre et participant (2004-2008)
- **Projet RNRT Temic** : Responsable de l'implication de l'action RESO dans ce projet qui vise à valider les approches hétérogènes supportées par les réseaux actifs à l'aide d'une application de télé-maintenance réseau. Recrutement et encadrement de trois ingénieur expert INRIA (2003-2006).
- **Projet RNRT VTHD++** : Responsable de l'activité sur les réseaux actifs hautes performances autour du backbone VTHD. Recrutement et encadrement d'un ingénieur expert (2001-2005).
- **Projet RNTL Etoile** : Participant aux sous-projet sur l'aspect Grille Active en lien avec les besoins des applications et middlewares (2002-2004).
- **Projet Région** : Co-reponsable du projet "Fédération Lyonnaise de Calcul Scientifique Haute Performance" avec diverses équipes de recherche Lyonnaise impliquées, financé par la Région Rhône-Alpes 2000-2002 et 2003-2005
- **Membre de ACI-Grid Jeune Equipe** : travaux menés sur les support réseaux pour les Grilles de calcul (2001-2003).
- **ASPROnet** : membre de cette Action Specifique 47 du CNRS sur la Programmabilité des Réseaux et des Services (2002)
- **Action Bio-Informatique** : Co-responsable sur le thème "Cluster et Grid Computing pour l'évolution des génômes de mammifères" avec le Laboratoire de Biométrie et de Biologie Évolutive de l'Université Claude Bernard dans le but de développer et mettre en oeuvre des compétences de calcul intensif appliqué au génome, financé par l'université Claude Bernard, 2000-2001
- **Action Incitative INRIA** : membre de "Rescapa : Support logiciel pour reseaux a capacite d'adressage" (1998-1999)

4.4 Collaborations industrielles

- **Hitachi Europe** : Organisation commune de la conférence IWAN2005 et colaboration dans le projet Autonomic Internet
- **France Telecom R&D** : Co-encadrement d'une thèse CIFRE avec FTR&D Lannion (2005-2008)
- **Support à l'incubation d'une jeune entreprise** - Société 3DDL. Collaboration sur l'apport de solutions de réseaux actifs pour le déploiement d'applications mobiles sur téléphones. Projet financé par la Région Rhone-Alpes en partenariat avec LIRIS, INSA Lyon, 2003-2005
- **SUNLAbs Europe** : participant à la collaboration RESO-SUN. Co-encadrement d'un doctorant CIFRE (2001-2004)
- **EDF/DER** : Co-responsable de cette étude contractuelle sur les réseaux haut-débit et la gestion de clusters en ce qui concerne le partage d'objets distribués à l'aide de réseaux haute-performance, 1999-2003.
- **Microsoft** Responsable pour l'action RESO de ce projet en lien avec d'autres équipes INRIA Rhône-Alpes : développement et exploration d'outils de communications haute performance adaptés à des environnements Windows NT, 2000-2002. Recrutement et encadrement d'un ingénieur

expert.

- **ANVAR** : Co-responsable sur le thème "Architecture Active Hétérogène pour les Réseaux de l'Internet du Futur" - mise en oeuvre d'une plate-forme de d'expérimentation de réseaux actifs, financé par l'ANVAR, 2000-2002
- **MDS** : Responsable de la mise en place un fonds d'expertise avec la société MDS International concernant l'étude de protocoles asymétriques pour réseaux hertziens, 1999-2000. Mise en place d'une plate-forme d'expérimentation réseaux satellites.

5 Organisation d'évènements scientifiques

Cette liste ne couvre que la période 2008-2013, pour une liste complète ainsi que pour la liste des évènements où je suis membre de Comités de Programmes se référer à http ://perso.ens-lyon.fr/laurent.lefevre

- **Co General chair de ICPP 2013 : 42th International Conference on Parallel Processing**, Lyon, October 1-4, 2013
- **Co-Program chair de CGC 2013 : Third International Conference on Cloud and Green Computing**, Karlsruhe, Germany, 30 Septembre - 2 Octobre, 2013
- **Co-Workshop chair de ExtremeGreen 2013 : Extreme Green & Energy Efficiency in Large Scale Distributed Systems** , Delft, The Netherlands, Mai 2013
- **Program Vice-Chair de track "Cluster, Grid and Cloud Computing" dans 12th IEEE International Conference on Scalable Computing and Communications (ScalCom 2012)**, Changzhou, China, 17-20 Décembre , 2012
- **Program Chair de IEEE International Conference on Green Computing and Communications (GreenCom 2012)**, Besancon, France, 20-23 Novembre, 2012
- **Co-organisateur des Entretiens Jacques Cartier** : Colloquium on "Towards ecological and energy efficient Information and Communication Technology - Vers des Technologies de l'Information écologiques et efficaces en consommation énergétique", Lyon, France, 19-20 Novembre 2012
- **Co PC Chair de Cloud&Grid 2012 : The Second International Workshop on Cloud and Grid Interoperability**, Gwangju, Corée, 6-8 Septembre 2012
- **Co PC Chair de 14th IEEE International Conference on High Performance Computing and Communications (HPCC-2012)**, Liverpool, UK, 25-27 Juin 2012
- **Co-organisateur des Rencontres INRIA Industrie** : Sciences Numériques et efficacité énergétique - Numerical Sciences and energy efficiency, INRIA Montbonnot, France, 8 Mars 2012
- **Co-organisateur des Entretiens Jacques Cartier** : Colloquium on "Information and Communications Technologies : Are they Green ?", Montreal, Canada, 3-4 Octobre 2011
- **Workshop co-chair de SUNSET2011 : Workshop on Sustainable Networking during Networking 2011 conference**, Valencia, Espagne,13 Mai, 2011
- **Program vice-chair de GreenCom 2010 : The 2010 IEEE / ACM International Conference on Green Computing and Communications**, Hangzhou, Chine, 18-20 Décembre 2010
- **Program chair de Grid 2010 : The 11th IEEE / ACM International Conference on Grid Computing**, Bruxelles, 25-29 Octobre 2010
- **Co-Organization de INTECH seminary day** on "Green IT, Green by IT and sustainable development", INRIA, Grenoble, 8 Juin 2010
- **Co-Track Chair de "Track 8. Grid and scalable computing"** de 2nd International Conference

on Computer Science and its Applications (CSA 2009), Jeju Island, Korea, 10-12 Décembre 2009

- **Program Comittee Co-chair de GADA2009 : Fourth International Conference on Grid computing, high-performAnce and Distributed Applications"**, Vilamoura, Algarve, Portugal, 2-6 Novembre 2009
- **Workshop co-chair et organisateur du workshop E2GC2 : Energy Efficient Grids, Clouds and Clusters**, during the IEEE Grid2009 conference, Banff, Canada, 13-15 Octobre 2009
- **Co-organisateur local de Parco 2009 : International Conference on Parallel computing**, Lyon, France, 1-4 Septembre, 2009
- **Co General Chair de 11th IEEE International Conference on High Performance Computing and Communications (HPCC-09)**, Korea University, Seoul, Corée, Juin 2009
- **Workshop Co-Chair de HPPAC 2009 : The Fifth Workshop on High-Performance, Power-Aware Computing**, pendant la conférence IPDPS2009, Rome, Italie, 25 Mai 2009
- **Tutorial et workshop chair de PDCAT'08 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies**, Dunedin, Nouvelle Zélande, Décembre 2008
- **General chair de CCGrid 2008 : 8th IEEE/ACM International Symposium on Cluster Computing and the Grid, Lyon**, France, Mai 2008

6 Jury de thèses :

- Remigiusz Modrzejewski : "Distribution and Storage in Networks", Université de Nice, Octobre 2013
- Anton Beloglazov : "Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers", University of Melbourne, Australie, Mars 2013 - Rapporteur
- Marco Guazzone : "Power and Performance Management in Cloud Computing Systems", University of Torino, Italy, Février 2012 - Rapporteur
- A. Sivagami : "Energy efficient latency optimized data gathering framework for wireless sensor networks", Anna Univeristy, Chennai, Inde, Aout 2011 - Rapporteur
- Anthony Mouraud : "Approche distribuée pour la simulation événementielle de réseaux de neurones impulsionnels", Université Antilles Guyane, France, Juin 2009 - Examinateur
- Lakshmi Priya : "A network layer Grid to support application-awareness", Anna University, Chennai, Inde, Juin 2009 - Rapporteur
- Benjamin Quetier : ""EMUGRID : études des mécanismes de virtualisation pour l'émulation conforme de grilles à grande échelle", Université Paris XI, France, Septembre 2008 - Examinateur
- Frank Chiang : "Self-Adaptability, Resilience and Vulnerability in Autonomic Communications with Biology-inspired Strategies", University of Technology, Sydney, Australie, Juillet 2008 - Rapporteur
- Sylvain Martin : "WASP - Lightweight Programmable Ephemeral State on Routers to Support End-to-End Applications", University of Liège, Belgique, Octobre 2007
- Bruno Volckaert : "Architectures and algorithms for network and service aware Grid resource management", Université de Gent, Belgique, Mai 2006 - Rapporteur
- Edgar Magana : "Heuristic algorithm for scheduling computational resources in policy-based

grid network management", Universitat Politechnica de Catalunya, Barcelone, Espagne, Avril 2005 - Rapporteur

– Eric Garcia : "Une plate-forme pour le développement pour applications coopératives multimédia intégrant la gestion de la qualité de service", Université de Franche-Comté, Besancon, Nov. 30, 2001 - Examinateur

7 Publications 1994-2013

7.1 Brevet

– Narjess Ayari, Denis Barbaron and Laurent Lefèvre : *"Procédés de gestion de sessions multi-flux"*, France Telecom R&D Patent. , Juin 2007

7.2 Publications d'actes de conférences

– Zhiyi Huang, Zhiwei Xu, Laurent Lefèvre, Hong Shen, John Hine, and Yi Pan. *Special Issue on Emerging Research in Parallel and Distributed Computing, Journal of Supercomputing,* volume 51, Mars 2010

– Zhiyi Huang, Zhiwei Xu, Nathan Rountree, Laurent Lefevre, Hong Shen, John Hine, and Yi Pan, editors. *Proceedings of PDCAT 2008 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies*, IEEE Computer Society, Dunedin, New Zealand, Décembre 2008

– Thierry Priol, Laurent Lefèvre, and Rajkumar Buyya. Proceedings of CCGrid2008 : Eight IEEE International Symposium on Cluster Computing and Grid. IEEE, Lyon, France, Mai 2008

– Laurent Lefèvre et Jean-Marc Pierson. *Special issue from International Conference on Pervasive Services - Journal of System and Software*. 2007. Volume 80, Issue 12, Pages 1939-2076, Décembre 2007

– Lionel Brunie, Salim Hariri, Laurent Lefèvre et Jean-Marc Pierson. *Proceedings of ICPS2006 : International Conference on Pervasive Services*. IEEE, Lyon, France, Juin 2006.

– David Hutchison, Spiros Denazis, Laurent Lefèvre et Gary Minden. *Proceedings of IWAN2005 : Seventh Annual International Working Conference on Active and Programmable Networks*. Nice, France, Novembre 2005.

– Bernard Tourancheau, Laurent Lefèvre et Cong-Duc Pham. *Proceedings of MUG 2000 : First Myrinet User Group Conference*. Lyon, France, Septembre 2000.

7.3 Contributions à des chapitres de livres

1. Robert Basmadjian, Georges Da Costa, Ghislain Landry Tsafack Chetsa, Laurent Lefevre, Ariel Oleksiak, Jean-Marc Pierson. "*Energy Aware Approaches for HPC Systems*", **High-Performance Computing on Complex Environments**. à paraitre 2014
2. Laurent Lefèvre and Jean-Marc Pierson. "*Environmental Impact of Networking Infrastructures*", Chapter 1, pages 1-16, **Green Networking Book**, Wiley, Francine Krief edition, 2012
3. Marcos Dias de Assuncao and Laurent Lefevre. "*State of the Art on Technology and Practices for Improving the Energy Efficiency of Data Storage*", volume 87, Chapitre du livre **Green and sustainable computing in Advances in Computers**, pages 89-124. Elsevier, Ali Hurson and Atif Memon edition, 2012
4. Anne-Cécile Orgerie and Laurent Lefèvre. "*Energy-Efficient Reservation Infrastructure for Grids, Clouds and Networks*", Chapitre du livre **Energy Efficient Distributed Computing Systems**,

Wiley Series on Parallel and Distributed Computing, John Wiley & Sons. Pages 133-162, (ISBN : 978-0-470-90875-4), Aout 2012

5. Anne-Cécile Orgerie and Laurent Lefèvre. "*Energy-efficient data transfers in large-scale distributed systems*", Chapitre du livre **Handbook of Energy-Aware and Green Computing**, Chapman & Hall, 2011
6. Anne-Cécile Orgerie, Marcos Dias de Assuncao, and Laurent Lefèvre. "*Energy Aware Clouds*", Chapitre du livre "**Grids, Clouds and Virtualization**", M. Cafaro and G. Aloisio (Eds.), pages 145-170. Springer Book, Octobre 2010
7. Laurent Lefèvre et Jean-Patrick Gelas *"High Performance Execution Environments"* - Chapitre 14 du livre "**Programmable Networks and their Management**" - ISBN 1-58053-745-6 ; A. Galis, S. Denazis, C. Brou, C. Klein (ed), publié par Artech House Books, pages 291-321, UK, Mai 2004

7.4 Revues internationales

1. Anne-Cécile Orgerie, Marcos Dias de Assunção and Laurent Lefèvre "*"A Survey on Techniques for Improving the Energy Efficiency of Large Scale Distributed Systems*", **ACM Computing Surveys**, à paraître en 2014
2. G.L. Tsafack Chetsa, L. Lefevre, J.M. Pierson, P. Stolf, G. Da Costa. "*Exploiting Performance Counters to Predict and Improve Energy Performance of HPC Systems*", **The International Journal of Future Generation Computer Systems (FGCS)** - Aout 2013
3. Mohammed Diouri, Ghislain Tsafack Chetsa, Olivier Glück, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf and Georges Da Costa. « *Energy efficiency in HPC with and without knowledge on applications and services* », **International Journal of High Performance Computing Applications**, Sage Publisher, 27, Pages : 232-243, Aout 2013
4. Mohammed Diouri, Olivier Glück, and Laurent Lefevre. "*Smart energy management for greener supercomputing*", **ERCIM News**, 92, Janvier 2013
5. Alejandro Fernández-Montes, Luis González Abril, Juan Antonio Ortega, Laurent Lefèvre. "*Smart scheduling for saving energy in grid computing*", **Expert Systems with Applications**, Volume 39, Issue 10, Pages 9443-9450, Aout 2012
6. Pablo Neira Ayuso, Rafael M. Gasca, and Laurent Lefevre. "*FT-FW : A cluster-based fault-tolerant architecture for stateful firewalls*", **Computers & Security**, 31(4) :524-539, Juin 2012.
7. J. Rubio-Loyola, A. Galis, A. Astorga, J. Serrat, L. Lefevre, A. Fischer, A. Paler, and H. de Meer. "*Scalable Service Deployment on Software Defined Networks*", I**EEE Communications Magazine**, Vol. 49 Issue 12, ISSN : 0163-6804. pp 84-93, Décembre 2011.
8. Anne-Cécile Orgerie and Laurent Lefèvre. "*ERIDIS : Energy-efficient Reservation Infrastructure for large-scale DIstributed Systems*", **Parallel Processing Letters**, Volume 21, Issue 2, pages 133-154, Juin 2011
9. Anne-Cécile Orgerie, Laurent Lefèvre, and Isabelle Guérin-Lassous. "*Energy-Efficient Bandwidth Reservation for Bulk Data Transfers in Dedicated Wired Networks*", **The Journal of SuperComputing**, Special issue on Green Networks, 28 pages, Mars 2011.
10. Narjess Ayari, Denis Barbaron, and Laurent Lefèvre." *Design and Evaluation of a Session-Aware Admission-Control Framework for Improving Service Providers Profitability*", **Journal of Internet Engineering** - Special issue on Service Oriented Infrastructures, 4(1), Décembre 2010
11. Pablo Neira Ayuso, Rafael M. Gasca, and Laurent Lefèvre. "*Communicating between the kernel and user-space in Linux using Netlink sockets*". **Journal Software : Practice and Experience**, 40(9) :797-810, Aout 2010

12. Laurent Lefèvre and Anne-Cécile Orgerie. "*Designing and Evaluating an Energy Efficient Cloud*". **Journal of SuperComputing**, Volume 51, Number 3, pages 352-373, Mars 2010.
13. Pablo Neira Ayuso, Rafael M. Gasca, and Laurent Lefèvre. "*Demystifying Cluster-Based Fault-Tolerant Firewalls*", **IEEE Internet Computing** : Special Issue on Unwanted Traffic, 13(6) :30-37, Novembre 2009
14. Laurent Lefevre and Anne-Cecile Orgerie. "*Towards Energy Aware Reservation Infrastructure for Large-Scale Experimental Distributed Systems*", **Parallel Processing Letters** - Special Issue on Clusters and Computational Grids for Scientific Computing, Volume 19, Issue 3, pages 419-433, Septembre 2009.
15. Narjess Ayari, Denis Barbaron, Laurent Lefevre, and Pascale Vicat-Blanc Primet. *"Fault Tolerance for Highly Available Internet Services : Concepts, Approaches, and Issues"*, **IEEE Communications Surveys and Tutorials**, Volume 10, Number 2, Juillet 2008
16. Laurent Lefèvre and Jean-Patrick Gelas. *"IAN2 : Industrial Autonomic Network Node architecture for supporting personnalized network services in the industrial context"*, **The International Journal of Future Generation Computer Systems (FGCS)** - Grid Computing : Theory, Methods and Applications, Volume 24, Issue 1, pages 58-65, Janvier 2008
17. F. Bouhafs, J.P. Gelas, L. Lefèvre, M. Maimour, C. Pham, P. Primet, B. Tourancheau. *"Designing and Evaluating An Active Grid Architecture"*, **The International Journal of Future Generation Computer Systems (FGCS)** - Grid Computing : Theory, Methods and Applications, Elsevier editors, Volume 21, issue 2, pages 315-330, Février 2005
18. Alessandro Bassi, Micah Beck, Fabien Chanussot, Jean-Patrick Gelas, Robert Harakaly, Laurent Lefèvre, Terry Moore, James Plank, Pascale Primet. *"Active and Logistical Networking for Grid Computing : the e-Toile Architecture"*, **The International Journal of Future Generation Computer Systems (FGCS)** - Grid Computing : Theory, Methods and Application, Elsevier editors, Volume 21, issue 1, pages 199-208, ISSN 0167-739X, Janvier 2005
19. Jean-Patrick Gelas, Saad El Hadri et Laurent Lefèvre. *"Towards the Design of an High Performance Active Node"*, **Parallel Processing Letters journal**, Vol. 13, No. 2, pp. 149-167, Juin 2003
20. Laurent Lefèvre. *"Parallel programming on top of DSM Systems : An Experimental Study"*, **Parallel Computing** - Environments and Tools for Parallel Scientific Computing III, volume 23 (1-2), pages 235-249. Avril 1997.

7.5 Revues nationales ou francophones

1. Francoise Berthoud, Bernard Boutherin, Romaric David, Robert Ferret, Laurent Lefevre. "*Réduire la consommation électrique des centres de données*", **La Revue Durable**, Dossier "Les Technologies de l'Information et de la Comumunication et l'Impératif de la Sobriété", Publication CERIN, Aout 2013
2. Anne-Cécile Orgerie, Laurent Lefèvre, and Jean-Patrick Gelas "*Etudier l'usage pour économiser l'énergie dans les Systèmes Distribués à Grande Echelle : l'Approche EARI*", **TSI : Technique et Science Informatiques**,Volume 30, Number 5, pages 515-538, May 2011
3. Jean-Patrick Gelas, Laurent Lefèvre. *"Flexibilité et performance dans les routeurs actifs logiciels pour un support efficace des services déployés sur des réseaux gigabits"*, **Annales des Télécoms - Numéro spécial sur les réseaux actifs** - Vol. 59, n 5-6, pages 645-685, Juin 2004

7.6 Conférences d'audience internationale avec comité de lecture et publication des actes

2013

1. Mohammed el Mehdi Diouri, Olivier Gluck, Laurent Lefevre and Jean-Christophe Mignot. "*Energy Estimation for MPI Broadcasting Algorithms in Large Scale HPC Systems*", **EuroMPI 2013 : 20th European MPI Users' Group Meeting**, Madrid, Espagne, 15-18 Septembre 2013
2. Ghislain Landry Tsafack Chetsa, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf, Georges Da Costa. "*A User Friendly Phase Detection Methodology for HPC Systems' Analysis*", **GreenCom 2013 : The 2013 IEEE International Conference on Green Computing and Communications**, Pékin, Chine, 20-23 Aout, 2013
3. Mohammed El Mehdi Diouri, Olivier Gluck and Laurent Lefevre. "*SESAMES : a Smart-Grid Based Framework for Consuming Less and Better in Extreme-Scale Infrastructures*", **GreenCom 2013 : The 2013 IEEE International Conference on Green Computing and Communications**, Pékin, Chine, 20-23 Aout, 2013
4. P. Vetter, T. Ayhan, K. Kanonakis, B. Lannoo, K.L. Lee, L. Lefevre, C. Monney, F. Saliou, X. Yin. "*Towards Energy Efficient Wireline Networks, an Update from GreenTouch*", CLEO-PR & OECC 2013/PS : **The 18th OptoElectronics and Communications Conference**, Kyoto, Japan, 30 Juin - 4 Juillet, 2013
5. Mohammed el Mehdi Diouri, Olivier Gluck, Laurent Lefevre and Jean-Christophe Mignot. "*Your Cluster is not Power Homogeneous : Take Care when Designing Green Schedulers !*", **IGCC2013 : International Green Computing Conference**, Arlington, USA, 27-29 Juin, 2013
6. Mohammed Diouri, Olivier Glück, Laurent Lefevre, and Franck Cappello. "*ECOFIT : A Framework to Estimate Energy Consumption of Fault Tolerance Protocols during HPC executions*", **CCGrid2013, the 13th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing**, Delft, the Netherlands, 13-16 Mai 2013
7. Ghislain Landry Tsafack, Laurent Lefevre, Patricia Stolf. "*A Three Step Blind Approach for Improving HPC Systems Energy Performance* ", **EE-LSDS 2013 : Energy Efficiency in Large Scale Distributed Systems conference** , Vienna, Austria, 22-24 Avril 2013
8. Mohammed Diouri, Manuel Dolz, Olivier Glück, Laurent Lefevre, Pedro Alonso, Sandra Catalan, Rafael Mayo, Enrique Quintan-Orti. "*Solving some Mysteries in Power Monitoring of Servers : Take Care of your Wattmeters ! *", **EE-LSDS 2013 : Energy Efficiency in Large Scale Distributed Systems conference** , Vienna, Austria, 22-24 Avril 2013

2012

9. Ghislain Landry Tsafack, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf, and Georges Da Costa. "*A runtime framework for energy efficient HPC systems without a priori knowledge of applications*", **ICPADS 2012 : 18th International Conference on Parallel and Distributed Systems**, Singapoure, Décembre 2012
10. Ghislain Landry Tsafack, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf, and Georges Da Costa. "*Beyond CPU Frequency Scaling for a Fine-grained Energy Control of HPC Systems*", **SBACPAD 2012 : 24th International Symposium on Computer Architecture and High Performance Computing**, New York City, USA, Octobre 2012
11. Jean-Patrick Gelas, Laurent Lefevre, Teferi Assefa and Mulugeta Libsie . "*Virtualizing Home Gateways for Large Scale Energy Reduction in Wireline Networks*", **Electronic Goes Green 2012 (EGG)**, Berlin, Allemagne, Septembre 2012

12. Christina Herzog, Laurent Lefevre and Jean-Marc Pierson. "*Green IT for Innovation and Innovation for Green IT : The virtuous circle*", **Human Choice and Computers (HCC10) International Conference**, Amsterdam, Septembre 2012
13. M. Diouri, O. Gluck, and L. Lefevre. "*Towards a novel Smart and Energy-Aware Service-Oriented Manager for Extreme-Scale applications*", **First Workshop for Power Grid-Friendly Computing (PGFC'12)**, co-located with IEEE IGCC'12, San Jose, USA, Juin 2012
14. M. Diouri, O. Gluck, L. Lefevre, and F. Cappello. "*Energy considerations in Checkpointing and Fault Tolerance protocols*", **2nd Workshop on Fault-Tolerance for HPC at Extreme Scale (FTXS 2012)**, Boston, USA, Juin 2012
15. P. Vetter, L. Lefevre, L. Gasca, K. Kanonakis, L. Kazovsky, A. Lee, C. Monney, X. Qiu, F. Saliou, and A. Wonfor. "*Research Roadmap for Green Wireline Access*", **Workshop on Green Communications and Networking during IEEE ICC'12**, Ottawa, Canada, Juin 2012.
16. Ghislain Landry Tsafack, Laurent Lefevre, Jean-Marc Pierson, Patricia Stolf, and Georges Da Costa. "*DNA-inspired Scheme for Building the Energy Profile of HPC Systems*", **1st International Workshop on Energy-Efficient Data Centres**, Madrid, Spain, Mai 2012.

2011

17. Helmut Hlavacs, Thomas Treutner, Jean-Patrick Gelas, Laurent Lefevre, and Anne-Cecile Orgerie. "*Energy consumption side-channel attack at Virtual Machines in a Cloud*", **International Conference on Cloud and Green Computing (CGC 2011)**, Sydney, Australie, Décembre 2011
18. Anne-Cécile Orgerie, Laurent Lefèvre, and Isabelle Guérin-Lassous."*On the Energy Efficiency of Centralized and Decentralized Management for Reservation-Based Networks*", **IEEE Global Communications Conference (GLOBECOM 2011)**, Houston, USA, Décembre 2011
19. Anne-Cécile Orgerie and Laurent Lefèvre."*Energy-Efficient Overlay for Data Transfers in Private Networks*", **IEEE International Conference on Networks (ICON 2011)**, Singapour, Décembre 2011
20. Ghislain Landry Tsafack, Laurent Lefevre, and Jean Patrick Gelas. "*On Applying DTNs to a Delay Constrained Scenario in Wired Networks*", **International Workshop on Opportunistic and Delay/disruption-Tolerant Networking** in conjunction with the 14th International Symposium on Wireless Personal Multimedia Communications, Brest, France, Octobre 2011
21. Anne-Cécile Orgerie, Laurent Lefèvre, Isabelle Guérin-Lassous an Dino Lopez Pacheco. "*ECOFEN : an End-to-end energy Cost mOdel and simulator For Evaluating power consumption in large-scale Networks*", **Sustalnet 2011 : First International Workshop on Sustainable Internet and Internet for Sustainability**, Lucca, Italie, Juin 2011
22. Anne-Cécile Orgerie and Laurent Lefèvre. "*Energy-Efficient Framework for Networks of Large-Scale Distributed Systems*", **ISPA 2011 : The 9th IEEE International Symposium on Parallel and Distributed Processing with Applications**, Busan, Corée, Mai 2011
23. Mohammed el Mehdi Diouri, Olivier Gluck, and Laurent Lefevre. "*Vers des machines exaflopiques vertes*", **Renpar 20 : Rencontres francophones du Parallélisme,** Mai 2011
24. Alex Galis, Stuart Clayman, Laurent Lefevre, Andreas Fischer, Hermann de Meer, Javier Rubio-Loyola, Joan Serrat, and Steven Davy. "T*owards In-Network Clouds in Future Internet*", **The Future Internet - Future Internet Assembly 2011 : Achievements and Technological Promises,** volume 6656, pages 19-33, Lecture Notes in Computer Science, Springer, Mai 2011

2010

25. Marcos Dias de Assuncao, Anne-Cécile Orgerie, and Laurent Lefèvre. "*An Analysis of Power Consumption Logs from a Monitored Grid Site*", **IEEE/ACM International Conference on Green Computing and Communications (GreenCom-2010)**, Hangzhou, Chine, Décembre 2010
26. J. Rubio-Loyola, A. Astorga, J. Serrat, W. K. Chai, L. Mamatas, A. Galis, S. Clayman, A. Cheniour, L. Lefevre, O. Mornard, A. Fischer, A. Paler, and H. de Meer. "*Platforms and Software Systems for an Autonomic Internet*", **IEEE Globecom 2010 - Next Generation Networking Symposium,** Miami, USA, Décembre 2010.
27. Anne-Cécile Orgerie, Laurent Lefèvre, and Jean-Patrick Gelas. "*Demystifying Energy Consumption in Grids and Clouds*", **The Work in Progress in Green Computing (WIPGC) Workshop**, in conjunction with the first IEEE sponsored International Green Computing Conference, Chicago, USA, Aout 2010
28. Gilles Fedak, Jean-Patrick Gelas, Thomas Hérault, Victor Iniesta, Derrick Kondo, Laurent Lefèvre, Paul Malecot, Lucas Nussbaum, Ala Rezmerita, and Olivier Richard."*DSL-Lab : a Low-Power Lightweight Platform to Experiment on Domestic Broadband Internet*", **International Symposium on Parallel and Distributed Computing (ISPDC 2010)**, Istanbul, Turquie, Juillet 2010.
29. Marcos Dias de Assuncao, Jean-Patrick Gelas, Laurent Lefèvre, and Anne-Cécile Orgerie. "*The Green Grid5000 : Instrumenting a Grid with Energy Sensors*", **5th International Workshop on Distributed Cooperative Laboratories : Instrumenting the Grid (INGRID 2010),** Poznan, Pologne, May 2010
30. Georges Da-Costa, Marcos Dias de Assuncao, Jean-Patrick Gelas, Yiannis Georgiou, Laurent Lefèvre, Anne-Cécile Orgerie, Jean-Marc Pierson, Olivier Richard, and Amal Sayah. "*Multi-facet approach to reduce energy consumption in clouds and grids : The GREEN-NET Framework*", **e-Energy 2010 : First International Conference on Energy-Efficient Computing and Networking,** Passau, Allemagne, Avril 2010.
31. J. Rubio-Loyola, A. Astorga, J. Serrat, L. Lefevre, A. Cheniour, D. Muldowney, S. Davy, A. Galis, L. Mamatas, S. Clayman, D. Macedo, Z. Movahedi,, G. Pujolle, A. Fischer, and H. de Meer. "*Manageability of Future Internet Virtual Networks from a Practical Viewpoint*", **FIA Book : Towards the Future Internet - Emerging Trends from European Research**, Valencia, Espagne, Avril 2010.

2009

32. Narjess Ayari, Denis Barbaron, and Laurent Lefèvre. "*Evaluating Session Aware Admission Control Strategies for Improving the Profitability of Service Providers*", **The 3rd IEEE Workshop on Enabling the Future Service-Oriented Internet : Towards Socially-Aware Networks** - Held in conjunction with IEEE GLOBECOM 2009, Honolulu, USA, Décembre 2009
33. Laurent Lefèvre and Anne-Cécile Orgerie. "*When Clouds become Green : the Green Open Cloud Architecture*", **Parco2009 : International Conference on Parallel Computing**, Lyon, France, Septembre 2009
34. Georges Da-Costa, Jean-Patrick Gelas, Yiannis Georgiou, Kamal Sharma, Laurent Lefevre, Anne-Cecile Orgerie, Jean-Marc Pierson, and Olivier Richard. "*The GREEN-NET Framework : Energy Efficiency in Large Scale Distributed Systems*", **HPPAC 2009 : High Performance Power Aware Computing Workshop** dans le cadre de IPDPS 2009, Rome, Italie, Mai 2009
35. Alex Galis, Spyros Denazis, Alessandro Bassi, Pierpaolo Giacomin, Andreas Berl, Andreas Fischer, Herman de Meer, John Srassner, Steven Davy, Daniel Macedo, Guy Pujolle, Javier R. Loyola, Joan Serrat, Laurent Lefevre, and Abderhaman Cheniour. "*Management Architecture and Systems for Future Internet Networks*", **Future Internet Assembly Book**, Prague, Mars 2009

2008

36. Anne-Cecile Orgerie, Laurent Lefevre, and Jean-Patrick Gelas. *"Save Watts in your Grid : Green Strategies for Energy-Aware Framework in Large Scale Distributed Systems"*, **ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems**, Melbourne, Australie, Décembre 2008
37. Pablo Neira Ayuso, Laurent Lefevre, and Rafael M. Gasca. *"hFT-FW : Hybrid Fault-Tolerance for Cluste-based Stateful Firewalls"*, **ICPADS 2008 : The 14th IEEE International Conference on Parallel and Distributed Systems**, Melbourne, Australie, Décembre 2008
38. Anne-Cecile Orgerie, Laurent Lefevre, and Jean-Patrick Gelas. *"Chasing Gaps between Bursts : Towards Energy Efficient Large Scale Experimental Grids"*, **PDCAT 2008 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies**, Dunedin, Nouvelle Zélande, Décembre 2008
39. Narjess Ayari, Laurent Lefevre, and Denis Barbaron. *"On improving the Reliability of Internet Services through Active Replication"*, **PDCAT 2008 : The Ninth International Conference on Parallel and Distributed Computing, Applications and Technologies**, Dunedin, Nouvelle Zélande, Décembre 2008
40. Pablo Neira Ayuso, Leonardo Maccari, Laurent Lefevre, and Rafael M. Gasca. *"Stateful Firewalling for Wireless Mesh Networks"*, **NTMS 2008 : The second IFIP International Conference on New Technologies, Mobility and Security**, Tanger, Maroc, Novembre 2008
41. Pablo Neira Ayuso, Laurent Lefevre, and Rafael M. Gasca. *"Multiprimary support for the Availability of Cluster-based Stateful Firewalls using FT-FW"*, **ESORICS 2008 : 13th European Symposium on Research in Computer Security**, Malaga, Espagne, Octobre 2008
42. Lawrence Cheng, Alex Galis, Bertrand Mathieu, Kerry Jean, Roel Ocampo, Lefteris Mamatas, Javier R. Loyola, Joan Serrat, Andreas Berl, Hermann de Meer, Steven Davy, Zeinab Movahedi, and Laurent Lefevre. *"Self-organising Management Overlays for Future Internet Services"*, **MACE 2008 : 3rd IEEE International Workshop on Modelling Autonomic Communications Environments** dans le cadre de ManWeek 2008 : 4th International Week on Management of Networks and Services, Samos Island, Grèce, Septembre 2008
43. Narjess Ayari, Denis Barbaron, Laurent Lefevre, and Pascale Vicat-Blanc Primet. *"On improving the Reliability of Cluster based Voice over IP Services"*, FastAbstract : **DSN 2008 : The 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks**, Anchorage, Alaska, USA, Juin 2008
44. Dino M. Lopez Pacheco, Laurent Lefevre, and Cong-Duc Pham. *"Lightweight Fairness Solutions for XCP and TCP Cohabitation"*, **IFIP/TC6 Networking 2008**, Singapoure, Mai 2008
45. Syed Hasan, Laurent Lefevre, Zhiyi Huang, and Paul Werstein. *"Cross Layer Protocol Support for Live Streaming Media"*, **AINA-08 : The IEEE 22nd International Conference on Advanced Information Networking and Applications**, Okinawa, Japon, Mars 2008
46. Narjess Ayari, Pablo Neira Ayuso, Laurent Lefevre, and Denis Barbaron. *"Towards a Dependable Architecture for Highly Available Internet Services"*, **ARES'08 : The Third International Conference on Availability, Reliability and Security**, Barcelone, Espagne, Mars 2008
47. Pablo Neira Ayuso, Rafael M. Gasca, and Laurent Lefevre. *"FT-FW : Efficient Connection Failover in Cluster-based Stateful Firewall"*, **PDP2008 : 16th Euromicro International Conference on Parallel, Distributed and network-based Processing**, Toulouse, France, Février 2008
48. Syed Hasan, Laurent Lefevre, Zhiyi Huang, and Paul Werstein. *"Supporting Large Scale eResearch Infrastructures with Adapted Live Streaming Capabilities"*, **6th Australasian Symposium on Grid Computing and e-Research**, Wollongong, Australie, Janvier 2008

49. Narjess Ayari, Denis Barbaron, Laurent Lefevre, and Pascale Vicat-Blanc Primet. *"A Session Aware Admission Control Scheme for next Generation IP Services"*, **Fifth annual IEEE Consumer Communications and Networking Conference (IEEE CCNC 2008)**, Las Vegas, USA, Janvier 2008

2007

50. Narjess Ayari, Denis Barbaron, Laurent Lefevre, and Pascale Vicat-Blanc Primet. *"SARA : A Session Aware Infrastructure for High Performance Next Generation Cluster-based Servers"*, **ATNAC 2007 : Australasian Telecommunication Networks and Applications Conference**, Christchurch, Nouvelle Zélande, Décembre 2007
51. Edgar Magaña, Laurent Lefevre, Masum Hasan, and Joan Serrat. *"SNMP-based Monitoring Agents and Heuristic Scheduling for large scale Grids"*, **Grid computing, high-performAnce and Distributed Applications (GADA'07)**, Vilamoura, Algarve, Portugal, Novembre 2007
52. Narjess Ayari, Denis Barbaron, Laurent Lefèvre, and Pascale Primet. *"Implementation of an Active Replication based Framework for Highly Available Services"*, **NetFilter Workshop** 2007, Karlsruhe, Allemagne, Septembre 2007
53. Dino M. Lopez Pacheco, Laurent Lefevre, and Cong-Duc Pham. *"Fairness issues when transferring large volume of data on high speed networks with router-assisted transport protocols"*, **High Speed Networks worhshop 2007**, dans le cadre de IEEE INFOCOM 2007, Anchorage, Alaska, USA, Mai 2007
54. Narjess Ayari, Denis Barbaron, Laurent Lefevre et Pascale Vicat-Blanc Primet. *"Session Awareness issues for next-generation cluster-based network load balancing frameworks"*, **AICCSA07 : ACS/IEEE International Conference on Computer Systems and Applications**, Jordanie, Mai 2007
55. Narjess Ayari, Denis Barbaron, Laurent Lefevre et Pascale Vicat-Blanc Primet. *"T2CP-AR : A system for Transparent TCP Active Replication"*, **AINA-07 : The IEEE 21st International Conference on Advanced Information Networking and Applications**, Niagara Falls, Canada, Mai 2007
56. Edgar Magana, Laurent Lefèvre et Joan Serrat. *"Autonomic Management Architecture for Flexible Grid Services Deployment Based on Policies"*, **ARCS 2007 : Architecture of Computing Systems**, Zurich, Suisse, Mars 2007

2006

57. Dino Lopez Pacheco, Cong-Duc Pham et Laurent Lefèvre. *"XCP-i : eXplicit t Control Protocol for heterogeneous inter-networking of high-speed networks"*, **Globecom 2006**, San Francisco, Californie, USA , Novembre 2006
58. Laurent Lefèvre et Jean-Patrick Gelas. *"Towards interplanetary Grids"*, **Next Generation Communication Infrastructure for Deep-Space Communications Workshop** pendant la conférence Second International Conference on Space Mission Challenges for Information Technology (SMC-IT), Pasadena, Californie,17-21 Juillet 2006
59. M. Brahma, M. Chaudier, E. Garcia, J.P. Gelas, H. Guyennet, F. Hantz, L. Lefèvre, P. Lorenz, H. Tobiet. *"TEMIC : a New Cooperative Platform for Industrial Tele-Maintenance"*, **DFMA06 : International Conference on Distributed Framework for Multimedia Applications**, Penang, Malaisie, Mai 14-17, 2006
60. Pablo Neira Ayuso, Laurent Lefèvre et Rafael M. Gasca. *"High Availability support for the design of stateful networking equipments"*, **ARES'06 : The First International Conference on Availability, Reliability and Security**, Vienna, Autriche, 20-22 Avril 2006

61. Laurent Lefèvre et Jean-Marc Pierson. *"Just in time Entertainment deployment on mobile platforms"*, **ICIW'06 : International Conference on Internet and Web Applications and Services**, Guadeloupe, 23-25 Février 2006
62. Laurent Lefèvre et Paul Roe. *"Improving the flexibility of Active Grids through Web Services"*, **4th Australasian Symposium on Grid Computing and e-Research**, Hobart, Australie, 16-19 Janvier 2006

2005

63. Martine Chaudier, Jean-Patrick Gelas et Laurent Lefèvre. *"Towards the design of an autonomic network node"*, **IWAN2005 : Seventh Annual International Working Conference on Active and Programmable Networks**, Nice, France, 21-23 Novembre 2005
64. Narjess Ayari, Denis Barbaron, Laurent Lefèvre, Pascale Primet. *"A Survey on High Availability Mechanisms for IP Services"*, **HAPCW2005 : High Availability and Performance Computing Workshop**, Santa Fe, New Mexico, USA, 11 Octobre 2005
65. Laurent Lefèvre. *"Heavy and lightweight dynamic network services : challenges and experiments for designing intelligent solutions in evolvable next generation networks"*, **Workshop sur Autonomic Communication for Evolvable Next Generation Networks** - The 7th International Symposium on Autonomous Decentralized Systems, Chengdu, Chine, 4-8 Avril 2005
66. Dieter Kranzlmuller, Laurent Lefèvre. *"A Record and Replay mechanism on programmable network card"*, **The IASTED International Conference on Parallel and Distributed Computing and Networks (PDCN 2005)**, Innsbruck, Autriche, 15-17 Février 2005
67. Laurent Lefèvre, Aweni Saroukou. *"Active network support for deployment of Java-based games on mobile platforms"*, **The First International Conference on Distributed Frameworks for Multimedia Applications (DFMA'2005)**, Besancon, France, 6-9 Février 2005

2004

68. Participant au papier commun *"Network services and traffic engineering methods for supporting applications on the VTHD experimental gigabit network"*, P. Cinquin, Y. Devillers, A. Gravey, E. Larreur, **Annals of telecommunications,** Vol. 59 n 11-12, Novembre 2004
69. Laurent Lefèvre, Dieter Kranzlmuller, Martin Maurer. *"Incremental Monitoring on Programmable Network Interface Cards"*, **PDPTA'04 : The 2004 International Conference on Parallel and Distributed Processing Techniques and Applications**, Las Vegas, Nevada, USA, 21-24 Juin, 2004
70. Alessandro Bassi, Micah Beck, Fabien Chanussot, Jean-Patrick Gelas, Robert Harakaly, Laurent Lefèvre, Terry Moore, James Plank, Pascale Primet. *"Active and Logistical Networking for Grid Computing : the e-Toile Architecture"*, **First International Workshop on Active and Programmable Grids Architectures and Components - APGAC'04** dans le cadre de ICCS 2004, Kraków, Pologne, 7-9 Juin 2004
71. E. Lemoine, C. Pham, L. Lefèvre. *"Packet Classification in the NIC for Improved SMP-based Internet Servers"*, **IEEE Proceedings of the International Conference on Networking (ICN 2004)**, Guadeloupe, 29 Février 2004

2003

72. Laurent Lefèvre, Jean-Marc Pierson, SidAli Guebli. *"Deployment of collaborative Web Caching with Active Networks"*, **International Working Conference on Active Networks, IWAN 2003**, LNCS 2982, pp 80-91, Kyoto, Japon, 9-12 Décembre 2003

73. Alessandro Bassi, Jean-Patrick Gelas, Laurent Lefèvre. *"A Sustainable Framework for Multimedia Data Streaming"*, **International Working Conference on Active Networks, IWAN 2003**, LNCS 2982, pp 68-79, Kyoto, Japon, 9-12 Décembre 2003
74. Alex Galis et Laurent Lefèvre. *"Programmable and Active Networks : a network infrastructure for next generation GRIDs"*, **Parco2003, Parallel Computing 2003 conference, Mini Symposium on Grid Computing**, Dresden University of Technology, Allemagne, 2-4 Septembre 2003
75. Laurent Lefèvre et Alice Bonhomme. *"Towards the Hierarchical Group Consistency for DSM systems : an efficient way to share data objects"*, **Parco2003, Parallel Computing 2003 conference**, Elsevier, pages 55-62, Dresden University of Technology, Allemagne, 2-4 Septembre 2003
76. Alex Galis, Jean-Patrick Gelas, Laurent Lefèvre, Kun Yang. *"Active Network Approach to Grid Management & Services in Workshop on Innovative Solutions for Grid Computing"*, **International Conference on Computational Science (ICCS 2003)**, LNCS 2658, ISBN 3-540-40195-4, pages 1103-1113, Melbourne, Australie, 2-4 Juin 2003
77. Jean-Patrick Gelas, Saad El Hadri, Laurent Lefèvre. *"Tamanoir : a software active node supporting gigabit networks"*, **ANTA 2003 : The second International Workshop on Active Networks Technologies and Applications**, pages 159-168, Osaka, Japon, 28-30 Mai 2003
78. Laurent Lefèvre et Jean-Patrick Gelas. *"Active Web : active networking support for web transport"*, **ANTA 2003 : The second International Workshop on Active Networks Technologies and Applications**, pages 147-156, Osaka, Japon, 28-30 Mai 2003

2002

79. L. Brunie, L. Favory, J.P. Gelas, L. Lefèvre, A. Mostefaoui, et F. Nait-Abdesselam. *"Sirsale : Integrated video databases management tools"*, T. Zhang J.R. Smith, S. Panchanathan, editor, **ITCOM2002 : Multimedia Networks and Management Systems**, volume 4862, SPIE. ISBN : 0-8194-4641-6., Boston, USA, Aout 2002
80. Alessandro Bassi, Jean-Patrick Gelas, et Laurent Lefèvre. *"Tamanoir-IBP : Adding storage to active networks"*, **Active Middleware Services,** IEEE computer society. ISBN : 0-7695-1721-8, pages 27-34, Edinburgh, Ecosse, Juillet 2002
81. Alessandro Bassi, Micah Beck, Jean-Patrick Gelas, et Laurent Lefèvre. *"Logistical storage in active networking : a promising framework for network services"*, Hamid Arabnia and Youngsong Mun, editors, **International Conference on Internet Computing, IC'2002,** volume 2, pages 209-216, Las Vegas, Nevada, USA, Juin 2002
82. Jean-Patrick Gelas, Jérôme Guilloux, et Laurent Lefèvre. *"Using os filtering capabilities for the improvement of software active routers"*, **International conference on parallel and distributed processing techniques and aplications (PDPTA 2002)**, volume 4, pages 1665-1671, ISBN : 1-892512-90-4., Las Vegas, Nevada, USA, Juin 2002
83. Marc Herbert, Pascale Primet, Bernard Tourancheau, et Laurent Lefèvre. *"A scalable and fully distributed architecture for ethernet switching"*, **Workshop on High Performance Switching and Routing (HPSR 2002)**, Kobe, Japon, Mai 2002
84. Laurent Lefèvre et Jean-Patrick Gelas. *"Towards the design of an active grid"*, Lecture Notes in Computer Science, editor, **International Conference on Computational Science (ICCS 2002)**, volume 2230, pages 578-587, Amsterdam, The Netherlands, Avril 2002
85. Laurent Lefèvre et Roland Westrelin. *"High performance communications libraries for windows 2000 : from a developer standpoint"*, **International conference on parallel and distributed processing techniques and aplications (PDPTA 2002)**, volume 4, pages 1665-1671, Las Vegas, Nevada, USA, Juin 2002

2001

86. L. Lefèvre, C. Pham, P. Primet, B. Tourancheau, B. Gaidioz, J.P. Gelas, et M. Maimour. *" Active networking support for the grid"*, Noaki Wakamiya Ian W. Marshall, Scott Nettles, editor, **IWAN 2001 : IFIP-TC6 Third International Working Conference on Active Networks**, volume 2207 of Lecture Notes in Computer Science, pages 16-33, Philadelphie, USA, Octobre 2001
87. Jean-Patrick Gelas et Laurent Lefèvre. *"Mixing high performance and portability for the design of active network framework with java"*, **3rd International Workshop on Java for Parallel and Distributed Computing, International Parallel and Distributed Processing Symposium (IPDPS 2001)**, San Fransisco, CA, USA, Avril 2001

2000

88. Jean-Patrick Gelas and Laurent Lefèvre. *"Tamanoir : A high performance active network framework"*, C. S. Raghavendra S. Hariri, C. A. Lee, editor, **Active Middleware Services, Ninth IEEE International Symposium on High Performance Distributed Computing**, pages 105-114, Kluwer Academic Publishers. ISBN 0-7923-7973-X, Pittsburgh, Pennsylvania, USA, Aout 2000
89. Laurent Lefèvre et Olivier Reyann. *"Combining low-latency communication protocols with multithreading for high performance DSM systems on clusters"*, **8th Euromicro Workshop on Parallel and Distributed Processing**, pages 333-340, IEEE Computer Society Press, Rhodes, Grèce, Janvier 2000

1999

90. Lionel Brunie, Laurent Lefèvre, et Olivier Reymann. *"High performance distributed objects for cluster computing"*, **1st IEEE International Workshop on Cluster Computing (IWCC '99)**, pages 229-236, IEEE Computer Society Press, Melbourne, Australie, Décembre 1999
91. P. Geoffray, L. Lefèvre, C. Pham, L. Prylli, O. Reymann, B. Tourancheau, et R. Westrelin. *"High-speed LANs : New environments for parallel and distributed applications"*, **ACM/IFIP EuroPar'99,** number 1685 in LNCS, pages 633-642, Springer-Verlag, Toulouse, France, Aout 1999

1998

92. Alice Bonhomme et Laurent Lefèvre. *"How to combine strong availability with weak replication of objects ?"*, **ECOOP98 : 12th Conference on Object Oriented Programming : Workshop on Mobility and Replication,** Bruxelles, Belgique, Juillet 1998

1996

93. Lionel Brunie et Laurent Lefèvre. *"A DSM-based Structural Programming Environment for Distributed and Parallel Processing"*, IEEE Computer Society Press, editor, **HiPC '96 : 3rd International Conference on High Performance Computing,** pages 469-474, Trivandrum, Inde, Décembre 1996
94. Lionel Brunie et Laurent Lefèvre. *"New propositions to improve the efficiency and scalability of DSM systems"*, **1996 IEEE Second International Conference on Algorithms & Architectures for Parallel Processing - ICA3PP '96**, pages 356-364, In IEEE, editor, Singapoure, Juin 1996
95. Lionel Brunie, Laurent Lefèvre, et Olivier Reymann. *Integration of Performance Evaluation Facilities into Distributed Shared Memory based Programming Environments*. **TDP'96 - Telecommunication, Distribution, Parallelism**, pages 279-294, La Londe Les Maures, France, Juin 1996

96. Lionel Brunie, Laurent Lefèvre, et Olivier Reymann. *Monitoring and performance evaluation of distributed shared memory applications*, **Second International Conference on Massively Parallel Computing Systems**, pages 382-389, Ischia, Italie, Mai 1996
97. Lionel Brunie, Laurent Lefèvre, et Olivier Reymann. *Execution Analysis of DSM Applications : A Distributed and Scalable Approach*. ACM Press, editor, **SPDT'96 : SIGMETRICS Symposium on Parallel and Distributed Tools**, pages 51-60, Philadelphia, Pennsylvania, USA, Mai 1996
98. Lionel Brunie, Laurent Lefèvre, et Olivier Reymann. *DOSMOS+ : A scalable Distributed Shared Memory environment including monitoring facilities*, **Parallel Programming Environments for High Performance Computing**, pages 165-168, Alpe d'Huez, France, Avril 1996

1995

99. S.A. Amin et D.J. Evans et Laurent Lefèvre. *PVM implementation for low-level image processing systolic array designs*, **EuroPVM95 Conference**, Lyon, France, Mai 1995
100. Henri-Pierre Charles, Laurent Lefèvre, et Serge Miguet. *An optimized and load-balanced portable parallel zbuffer*. In Robert F. Erbacher Eds SPIE Proceedings, Georges G. Grinstein, editor, **IST SPIE Symposium on Electronic Imaging**, volume Vol 2410 Visual Data Exploration and Analysis II, pages 394-403, IST SPIE., San Jose, USA, Février 1995

1994

101. L. Brunie et L. Lefèvre. *DOSMOS : A distributed shared memory based on P.V.M*, **First european PVM users group meeting**, Universita di Roma, Italie, Octobre 1994
102. L. Brunie, S. Chaumette, M. Cosnard, F. Desprez, L. Lefèvre, M. Loi, B. Tourancheau, X. Vigouroux, and M. Pourzandi. *"The LHPC Programming Environment"*, Proceedings of **Environments and tools for parallel scientific computing**, 1994

7.7 Conférences d'audience nationale avec comité de lecture et publication des actes

1. Dino Lopez Pacheco, Cong-Duc Pham et Laurent Lefèvre. *"XCP-i : eXplicit Control Protocol pour l'interconnexion de réseaux haut-débit hétérogènes"*, **CFIP 2006 : Colloque Francophone sur l'Ingénierie des Protocoles**, 30 Octobre - 3 Novembre 2006, Tozeur, Tunisie
2. L. Brunie et L. Lefèvre. *"Modèle de mémoire distribuée-partagée pour machine massivement parallèle"*, **RenPar'6 : 6 èmes rencontres francophones du parallélisme**, Ecole normale Supérieure de Lyon, France, Juin 1994

7.8 Rencontres nationales

1. Francoise Berthoud, Robert Ferret, and Laurent Lefèvre. "*Consommation énergétique des datacentres : quand la politique Européenne s'en mêle*", **JRES 2011 : 9 ème Journées Réseaux de l'enseignement supérieur et de la recherche**, Toulouse, France, Novembre 2011
2. Anne-Cecile Orgerie, Laurent Lefevre and Jean-Patrick Gelas. "E*conomies d'Energie dans les Systèmes Distribués à Grande Echelle : l'Approche EARI*", **JDIR 09 : Journees Doctorales en Informatique et Reseaux**, Belfort, France, Février 2009
3. Jean-Patrick Gelas et Laurent Lefèvre. *"Performance et dynamicité dans les réseaux : l'approche Tamanoir"*, **JDIR 2002**, pages 81-90, Toulouse, France, Mars 2002
4. Lionel Brunie, Laurent Lefèvre, et Olivier Reymann. *"Extensibilité et systèmes de mémoire distribuée virtuellement partagée*, **MPR '96 : Journées de Recherche sur La Mémoire Partagée Répartie**, Bordeaux, France, Mai 1996

5. Lefèvre Laurent. *"Une mémoire distribuée-partagée pour machine massivement parallèle"*, ETCA / ARCUEIL, editor, **Journées 1995 du Site Expérimental en Hyperparallelisme**, volume 2, pages 69-102, Janvier 1995

7.9 Interviews et articles de vulgarisation

1. **Blog "Web developpement durable"** : "Internet & Data Center : bilan énergétique positif ou négatif ?", 17 Juillet 2013
2. **Journal La Vie**, "Using Internet with less energy", 13 Juin 2013
3. **Journal Libération**, "Datacenters : la donnée écolo", 15 Avril 2013
4. **20 ans Inria Rhone Alpes**, Laurent Lefevre et Bel Dumé, "L'émergence du Green-IT, pour une informatique plus verte" - Article et interview, 15 Novembre 2012
5. **Brève Inria ; "Le saviez-vous ?"**, Laurent Lefevre et Christophe Castro, "Le Green IT se réfugie au Nord !", 10 Mai 2012
6. **Les Echos,** "Comment se chauffer avec les « data centers »", Avril 2012
7. **Journal Okapi**, Contributions à l'article "Internet Géant énergivore", Avril 2011
8. **01 Informatique Business et Technologies**, "La programmation eco responsable en chantier, Dossier développement durable ", Numéro 2054, pages 44-45, 30 Septembre, 2010
9. **La Recherche**, "Le défi de l'informatique verte ", Numéro 444, pages 64-66, Septembre 2010
10. **RCF Isère,** Interview radio : "L'empreinte écologique de l'Internet", Magazine EJDG (Ecole de Journalisme) - 25 Mars, 2010
11. **Radio Méditerranée Internationale**, Interview radio : "Le développement numérique durable", "Magazine du Développement Durable", 25 Février, 2010
12. **Pour la Science**, "Le coût écologique d'Internet", Dossier , Numéro 66, pages 40-41, Janvier-Mars 2010