Laboratoire de l'Informatique du Parallélisme

École Normale Supérieure de Lyon
Unité Mixte de Recherche CNRS-INRIA-ENS LYON-UCBL n° 5668

Emulation d'un nuage réseau de grilles de calcul: eWAN

Pascale Vicat-Blanc,
Olivier Glück,
Cyril Otal,
François Echantillac

Dec 2004

Research Report N° 2004-59

École Normale Supérieure de Lyon
46 Allée d'Italie, 69364 Lyon Cedex 07, France
Téléphone : +33(0)4.72.72.80.37
Télécopieur : +33(0)4.72.72.80.80
Adresse électronique : `lip@ens-lyon.fr`

Emulation d'un nuage réseau de grilles de calcul: eWAN

Pascale Vicat-Blanc, Olivier Glück, Cyril Otal, François Echantillac

Dec 2004

Abstract

The Grid aims at expanding the cluster based parallel computing paradigm toward large scale distributed systems based on IP networks. To validate grid algorithms, evaluate their performance and to study transport and coordination protocols and grid network services, a well controlled environment is required, allowing to precisely manage the experience conditions. One of the most important issue is to emulate the potentially very high speed wide area network interconnection. This paper presents a software and hardware tool for configuring and programming a large PC cluster in a wide area network emulation instrument. High performance, fine parameter tuning and a great utilization flexibility are the main proposed features of this experimental tool. This article discusses the eWAN design principles and the first experimentations that have been done on a prototype deployed over the *Grid5000* cluster at the ENS Lyon. Some usage scenarii are also proposed.

Keywords: network emulation, grid networking, eWAN

Résumé

La grille a pour objectif d'étendre les paradigmes du calcul parallèle sur grappes d'ordinateurs fortement couplés vers des systèmes distribués géographiquement et basés sur des réseaux IP. Pour l'analyse expérimentale des algorithmes de grille, l'évaluation de leurs performances, l'étude du comportement des protocoles de transport et de coordination, la conception de services réseaux de grille nécessite un environnement bien maîtrisé permettant le contrôle précis des conditions d'expérience. Une difficulté importante est d'émuler l'interconnexion réseau longue distance, potentiellement très haut débit. Cet article présente un outil logiciel et matériel de configuration et de programmation d'un cluster de PCs en un instrument d'émulation de réseau haut débit longue distance. Un haut niveau de performance, une fine maîtrise des paramètres de communication associé à une grande flexibilité d'utilisation de l'instrument d'expérimentation est proposé. Cet article explicite les principes de conception d'eWAN et les premières expérimentations menées sur le prototype déployé sur le cluster *Grid5000* de l'ENS Lyon. Quelques scénarii d'usage sont aussi proposés.

Mots-clés: nuage réseau, émulation, réseau pour la grille, eWAN

1 Introduction

La grille a pour objectif initial d'étendre les paradigmes du calcul parallèle sur grappes d'ordinateurs fortement couplés vers des systèmes distribués géographiquement. Une grille peut aussi être utilisée comme une plate-forme d'intégration d'applications faiblement couplées - chaque composant pouvant tourner de manière indépendante sur des machines parallèles à faible latence - et pour relier des ressources de calcul, de stockage et de visualisation ainsi que des instruments [FOS 99, BER 03]. La grille en tant que nouvel outil informatique soulève de nouveaux verrous non seulement sur le plan du déploiement et de l'ingénierie mais aussi des verrous scientifiques relatifs à la performance, au facteur d'échelle, à la dynamique, à la robustesse, à la sécurité, à la flexibilité aussi bien dans les systèmes que dans les réseaux. Pour étudier ces différents aspects, le chercheur a le choix entre la modélisation, la simulation ou l'expérimentation en vraie grandeur.

En ce qui concerne les réseaux, on peut simuler leur comportement avec un simulateur classique (NS2 ou Opnet). Le simulateur exécute du code dans un environnement synthétique ; il tourne généralement sur une unique machine. Les avantages de la simulation sont un faible coût, une grande flexibilité et un contrôle total de la plate-forme expérimentale. Les limites sont la puissance et la performance du simulateur : d'une part, le simulateur ne peut pas exécuter n'importe quelle taille d'expérience ; d'autre part, le temps d'exécution d'une expérience peut s'avérer être très long. Un autre problème majeur de la simulation est relative aux modèles de trafic utilisés. On ne peut pas aisemment injecter des traces de trafic réelles dans l'expérience.

D'un autre côté, on peut déployer en grandeur nature les services réseau sur des plate-formes expérimentales réelles et évaluer par dessus de véritables applications. Dans ce type de configuration, il est parfois difficile d'obtenir une plate-forme d'une taille suffisante pour les expériences souhaitées (pour des raisons de coût par exemple) et la flexibilité est souvent limitée. Par ailleurs, les expériences sont difficiles à reproduire. Les expérimentations sur de véritables réseaux opérationnels de production ont montré leurs limites et leurs lourdeurs.

L'émulation est une approche intermédiaire dans laquelle certains éléments sont réels - les applications et les extrémités communicantes par exemple - et d'autres sont simulés - les liens longue distance par exemple. Cette approche est utilisée depuis quelques années dans le domaine des réseaux [AHN 95] et a permis l'évaluation de protocoles sur réseaux satellites par exemple. Un émulateur de lien utilise un réseau réel avec de véritables interfaces réseau et ajoute des mécanismes logiciels permettant d'introduire du délai et/ou des fautes lors de la traversée du lien. Un autre type d'émulation de réseau consiste à fournir aux applications et aux pilotes réseaux un réseau virtuel ayant les caractéristiques souhaitées par l'expérience. Un intérêt tout particulier de l'émulation est de permettre au chercheur de modifier les disciplines de services dans les routeurs ou d'explorer des liaisons à très haut produit débit-délai et d'en mesurer les effets, ce qui est en général impossible à faire sur de vrais routeurs ou de vrais réseaux. Les avantages de l'émulation sont la maîtrise d'un environnement configurable, contrôlé et reproductible, l'utilisation d'un trafic réel pour réaliser l'expérience, la possibilité d'instrumenter l'outil pour enregistrer les évènements significatifs, et enfin le déploiement des applications existantes sans aucune modification comme si elles s'exécutaient dans un environnement réel. L'émulation comporte aussi quelques inconvénients : le temps mesuré est bien un temps réel ; la vitesse de l'émulation dépend des limites du matériel de simulation utilisé ; la complexité des topologies émulées est plus limitée que dans le cadre d'un simulateur ; enfin, il est possible d'avoir des problèmes d'interaction entre les processus émulés.

Pour l'analyse expérimentale systématique des algorithmes de grille, l'évaluation de leurs performances et du comportement des protocoles de transport, nous proposons un environnement d'émulation du réseau longue distance, bien maîtrisé et permettant le contrôle précis des conditions d'expérience. Cet outil est développé dans le cadre de la construction d'un grand instrument, appelé *Grid5000* [1], d'exploration des méthodes de programmation, de l'algorithmique et des logiciels de communication sur grille, qui regroupe une dizaine de sites dotés de centaines d'équipements réels, répartis sur le territoire français, interconnectés par un réseau très haut débit. Cette grille expérimentale est associée à un émulateur de grille de très large échelle, appelé *Data Grid Explorer* [2], basé sur une grappe qui aura, à terme, plus de 1000 processeurs. L'objectif de l'outil logiciel et matériel présenté ici est d'offrir un cœur d'émulation de réseau à haut niveau de performance présentant une fine maîtrise des paramètres associé à une grande simplicité et flexibilité d'utilisation. Afin de dégager les blocs fonctionnels d'un environnement d'émulation de réseau longue distance haut débit, nous caractérisons le nuage réseau d'une grille de calcul dans la section 2 suivante. La section 3 développe les principes de conception de l'outil. La section 4 présente les premières expérimentations menées sur le prototype Grid5000 ainsi que des propositions de scénarii d'usage. Finalement, l'état de l'art est dressé en section 5, avant les conclusions et les perspectives.

2 Caractérisation du nuage réseau d'une grille

Une grille, sur le plan de son anatomie, est une agrégation de ressources haute performances variées: entités de calcul, de stockage, de communication, de visualisation. On distingue trois niveaux d'abstraction principaux permettant le fonctionnement et l'utilisation d'une grille de calcul :

- l'infrastructure composée du nuage réseau et des ressources physiques;
- le logiciel d'administration et d'exécution appelé *middleware* ou intergiciel composé d'un ensemble de services évolués;
- les applications distribuées, exécutées sur l'infrastructure de la grille, administrées et coordonnées par le middleware et *"grillifiées"* à l'aide des services et bibliothèques fournies par le middleware.

Le cœur logiciel de la grille joue donc un rôle fondamental et central, tout comme un système d'exploitation est indispensable au bon fonctionnement d'un ordinateur et à son bon usage par les programmes d'application.

Une grille se différentie d'une grappe de calculateurs par le type d'interconnexion réseau sur lequel elle s'appuie. Alors que dans un cluster, la distance intersite est très courte, autorisant des latences de communication inférieures à la dizaine de microsecondes, dans une grille, le nuage réseau doit permettre de couvrir des distances importantes. Le type et les caractéristiques de réseau longue distance (`WAN`) sous-jacent a une incidence directe sur le type d'applications et de performances que l'on peut viser. En effet, compte-tenu que les latences ne pourront pas être inférieures à la milliseconde, et seront généralement de l'ordre de la dizaine de millisecondes, le grain de calcul des applications distribuées et parallélisées ne peut qu'être *gros*. L'avantage que présente la grille dans ces cas de figure est le facteur

[1] `http://www.grid5000.org`
[2] `http://www.lri.fr/~fci/GdX`

d'échelle, puisque le nombre de processeurs impliqués peut être potentiellement très grand, voire infini.

Le réseau longue distance introduit des problématiques d'hétérogénéité, mais aussi de performance et de sécurité à de multiples niveaux. On peut de manière simplifiée, distinguer trois types de *nuages réseau* pour l'interconnexion des ressources réparties :

- Internet, un réseau commun basique très accessible permettant de construire immédiatement une grille ;
- un réseau privé virtuel (`VPN`) dont la vocation est de limiter et de protéger l'accès aux ressources réparties ;
- un réseau privé réel, en général très haut débit pour obtenir des transferts performants, garants de la performance globale de l'environnement de grille.

Internet et plus particulièrement la technologie `TCP/IP` répond aux problèmes d'hétérogénéité et d'extensibilité, les réseaux privés virtuels à celui de la sécurité et les réseaux très haut débit à celui de la performance.

Ces trois types de réseaux sont actuellement utilisés pour l'interconnexion des ressources réparties dans le cadre des plates-formes de grille expérimentales internationales telles que EU DataGRID [VIC 03] [3] ou EGEE [4] qui s'appuient sur l'interconnexion des réseaux nationaux de la Recherche européens autour de GEANT, TeraGrid [5] qui est bâtie sur un réseau très haut débit (40Gb/s) ou EU DataTAG [6] qui interconnecte les grilles européennes et américaines par un réseau expérimental à 10Gb/s.

Des réseaux plus flexibles, proposant des services plus sophistiqués que les services `IP` actuels et réunissant à la fois les propriétés d'extensibilité, de sécurité et de performance, basés sur la technologie optique, sont étudiés intensivement par la communauté réseau internationale [FRA 03, PRI 04]. Ces réseaux, par les services avancés qu'ils proposeraient, permettraient de créer, à la demande, des environnements de calcul adaptés aux besoins de diverses communautés d'utilisateurs (organisations virtuelles). Le groupe Grid High Performance Networking du Global Grid Forum [7] définit le concept de service réseau de grille [FER 04]. Un tel service grille, au sens OGSA (Open Grid Service Architecture) du terme, est interfacé d'une part avec le middleware de grille et d'autre part avec les services réseau sous-jacents pour offrir un support de communication efficace et intégré au calculateur virtuel aggrégé dynamiquement par une communauté. De tels services peuvent prendre en charge la gestion de la sécurité, la mesure des performances de bout en bout, la gestion de la communication de groupe, la gestion de la qualité de service pour l'ensemble d'une session de travail.

En ce qui concerne la fourniture de la qualité de service de bout en bout, le service QoSINUS, s'appuyant sur la technologie des réseaux actifs [LEF 01] en bordure de la grille a par exemple été proposé [PRI 04]. Avec ce service, les requêtes de `QoS` exprimées, via une API spécifique, sont envoyées à destination des récepteurs, membres du groupe. Les routeurs actifs situés en bordure de cœur de réseau, interceptent les requêtes de `QoS` et les traduisent dynamiquement dans la sémantique de qualité de service appropriée aux réseaux traversés.

[3] `http://www.datagrid.org`
[4] `http://www.egee.org`
[5] `http://www.teragrid.org`
[6] `http://www.datatag.org`
[7] http://www.ggf.org

Le service est déployé aux points d'accès de la grille. Une application peut programmer une qualité de service flux par flux sans avoir à savoir comment elle sera assurée et adaptée par le réseau. Par ailleurs, ce service mesure et surveille en continu les services fournis par le réseau pour allouer localement et équilibrer au mieux et dynamiquement les requêtes des flux hétérogènes. Comme les services réseaux avancés ne sont cependant pas encore réellement disponibles et déployés aujourd'hui dans un contexte multi-domaine, il est difficile de les évaluer avec de vraies applications de grille. Mais il est important de pouvoir les étudier dès aujourd'hui pour mieux comprendre leurs intérêts et leurs faiblesses dans le contexte du calcul distribué. eWAN se propose d'émuler le cœur haut débit longue distance afin de permettre l'étude de ces nouveaux services réseaux.

3 Présentation générale de eWAN

3.1 Objectifs et choix conceptuels de eWAN

eWAN est un instrument destiné à l'étude des protocoles et des logiciels haute performance pour la grille avec un très grand nombre de calculateurs interconnectés. L'émulation et le contrôle de centaines de connexions simultanées, du gigabit et du multi-gigabit ainsi que la conception et le calibrage d'outil de mesures sont les principaux problèmes de performance à résoudre. L'outil étant basé sur un cluster de PC, les communications peuvent se faire en Ethernet, en Myrinet ou en Infiniband, les débits offerts sont de l'ordre du gigabits/s, voire 10 gigabits/s. eWAN doit représenter le nuage réseau tel qu'il est vu par les extrémités communicantes de la grille. En général, comme l'illustre la figure 1, dans les modèles de grille, on distingue les domaines publics et les domaines privés des sites. eWAN a pour objectif de n'émuler que le comportement du réseau longue distance. Si la grille comprend n sites, ce nuage peut être représenté par un graphe complet en $O(n^2)$, chaque sommet du graphe représentant le routeur de bordure (ou edge). Les systèmes de mesure de performance de la grille testent par exemple régulièrement ces $\frac{n(n-1)}{2}$ liens pour pondérer les arêtes du graphe [VIC 04] selon différentes métriques.

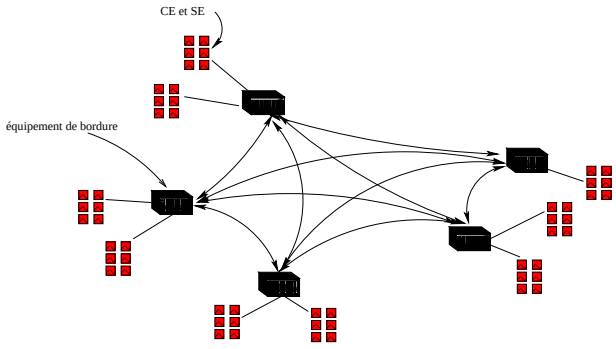

Figure 1: Modélisation d'une grille

3.2 Architecture de Ewan

Comme le montre la figure 2, eWAN émule le comportement d'un nuage réseau de grilles à partir d'un cluster de PCs interconnectés par un réseau haut-débit. Pour cela, notre outil attribue à chacun des nœuds du cluster une fonction unique: routage, émulation de liens ou génération de trafic de façon à ne pas surcharger les CPU et à distordre les performances.

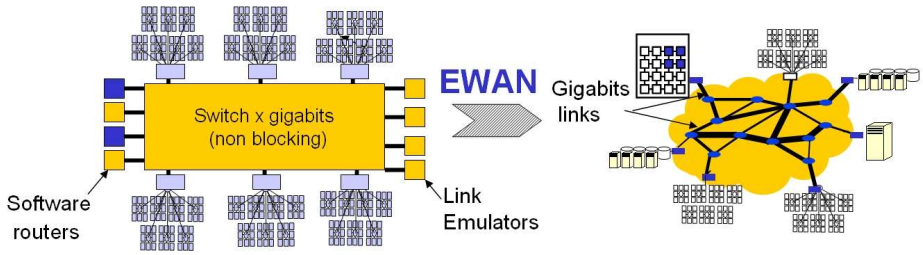

Figure 2: Architecture et principe de eWAN

Plusieurs fonctions de base ont été identifiées: 1) émulation de lien (latence et perte), 2) émulation de capacité (limitation de débit), 3) routage virtuel, 4) accès, classification et traitement différencié des paquets, 5) génération de trafic concurrent, 6) capture de trafic concurrent, 7) émission de trafic à analyser, 8) réception de trafic à analyser.

Ces fonctions doivent être réalisées de manière logicielle à haut débit. Pour conserver une vitesse de traitement proche de celle des liens physiques, les fonctions logicielles doivent être réalisées par des processeurs différents. Si la performance des liens étudiés n'est pas proche de celle du lien, il est possible de placer plusieurs fonctions sur un même équipement.

eWAN est constitué d'une grappe de nœuds non nécessairement matériellement identiques (avec par exemple un nombre d'interfaces, une capacité mémoire ou CPU differents), reliés par un ou plusieurs commutateurs et d'un serveur qui s'occupera de configurer ces nœuds. Ces machines appartiennent donc initialement au même sous-réseau et sont accessibles entre elles au niveau 2 de la couche OSI.

Le logiciel de configuration paramétre les nœuds de la grappe afin d'émuler la topologie souhaitée, notamment en répartissant les différentes fonctions parmi eux. Ceux-ci sont *a priori* polyvalents : tous possèdent les logiciels nécessaires leur permettant d'assurer tous les rôles. Les nœuds seront organisés en différents sous-réseaux correspondants à la topologie émulée souhaitée, réalisant ainsi un réseau logiciel de niveau 3 sur une architecture matérielle de niveau 2. Plusieurs étapes sont nécessaires pour la préparation de l'instrument: 1) définition de la topologie et des caractéristiques du nuagé réseau émulé, 2) génération des scripts d'initialisation, 3) déploiement et initialisation des équipements. Les deux phases suivantes sont le lancement et l'exécution de l'expérience puis l'analyse des traces.

3.3 Définition et configuration du réseau émulé

Pour définir un nuage réseau de grille, nous proposons une architecture en étoile avec un cœur de distance nulle (collapse core). Chaque point d'accès est relié au cœur par un lien de latence spécifique. Le problème est de créer un graphe en étoile à partir d'un graphe complet pondéré, c'est à dire de résoudre un système d'équation à n inconnues à partir de la connaissance de $\frac{n(n-1)}{2}$ liens pondérés. Ce problème étant insoluble, eWAN élimine des liens non significatifs

et ne conserve que ceux ayant les poids les plus élevés. Dans le cas d'un graphe non équilibré (un des sites est très excentré), ce sont par exemple ceux du nœud le plus "éloigné" du cœur qui sont examinés en priorité.

Lorsque l'on a défini les longueurs des branches de l'étoile, on associe à chacune d'elles les fonctions-sommet correspondantes: un point d'accès, un émulateur de lien, et un routeur de cœur. Le point d'accès est le routeur qui permet aux clients d'accéder au cœur. Il est relié à un unique routeur de cœur, par un lien qui sera émulé par un nœud.

Dans le cas de l'étoile avec un cœur de distance nulle, les routeurs de cœur sont organisés en anneau unidirectionnel, sans émulateur de liens entre eux. Cette configuration permet de réaliser un cœur surdimensionné par rapport au reste de la topologie. (L'inconvénient de cette organisation est que l'on retrouve dans les tables ARP deux lignes pour une même adresse MAC : une associée à l'interface d'arrivée, et l'autre à celle de départ, differente de la première à cause de l'aspect unidirectionnel de l'anneau. Certains mécanismes de protection contre le spoofing (comme le `rp_filter` sous GNU/Linux) jettent alors les paquets reçus. C'est pourquoi EWAN les désactive (en le signalant) au moment de la configuration des nœuds.)

Une fois la topologie déterminée, la deuxième étape consiste à configurer les nœuds de la grappe afin d'émuler le nuage réseau souhaité. Les figures 3 et 4 montrent des copies d'écran de l'interface utilisateur d'EWAN.

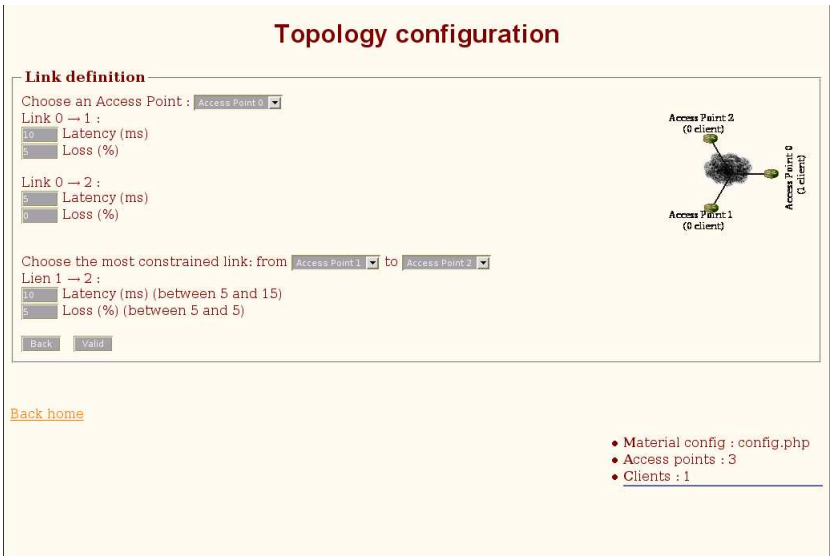

Figure 3: Phase de configuration du nuage réseau (ici, caractéristiques des liens)

Tout d'abord, les fonctions à émuler sont réparties parmi les nœuds disponibles grâce à un algorithme qui choisit les machines selon leurs caractéristiques physiques et les contraintes liées à la fonction. Par exemple, un émulateur de lien haut débit n'a besoin que de 2 interfaces, alors qu'il en faut normalement au moins 3 pour un routeur de cœur.

Puis les nœuds sont répartis dans des sous réseaux afin de representer l'architecture de la topologie virtuelle. Un réseau de contrôle utilisant des interfaces virtuelles est conservé : il regroupe tous les nœuds et permet au serveur de configuration de s'adresser à n'importe quel nœud directement. Ce réseau n'est pas utilisé durant l'expérimentation, mais seulement aux moments des changements de configuration ; il n'a donc aucune influence sur les performances

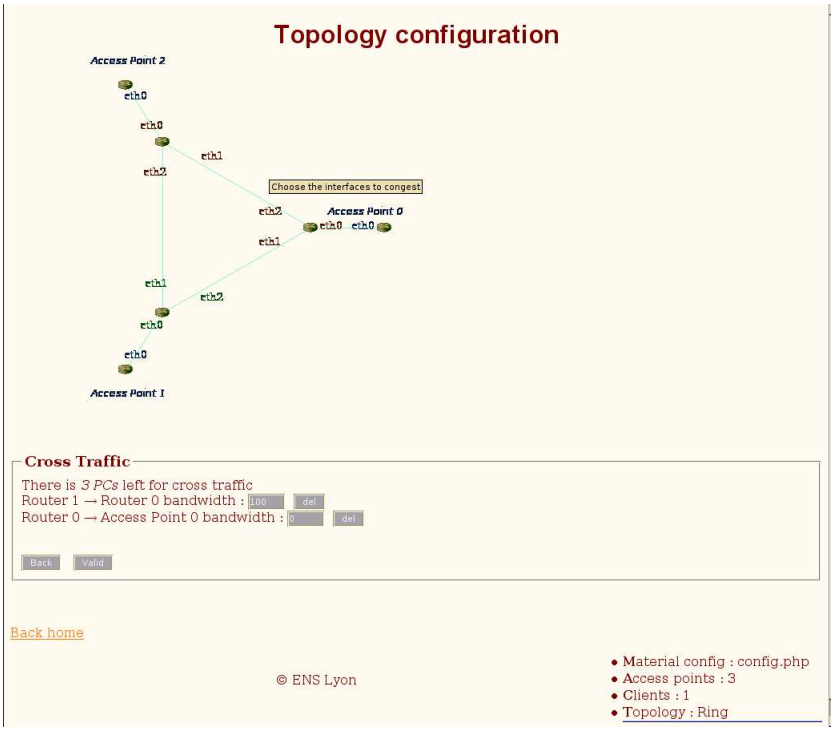

Figure 4: Mise en place d'un trafic perturbateur

de la grappe.

Les tables de routage sont calculées statiquement par l'algorithme de plus court chemin (Dijkstra) utilisé par OSPF : étant donné qu'une topologie n'est définie que pour une expérience donnée et ne doit pas subir de changement en cours d'expérience, il n'est pas nécessaire de mettre en place un routage dynamique. Un routage dynamique pourrait être assez aisément ajouté si cet aspect devait être étudié.

Enfin, dans la dernière étape, le serveur crée les scripts de configuration pour chaque nœud utilisé, avant de les déployer et de les exécuter dans la grappe qui sera ainsi totalemment configurée. La figure 5 montre sur un exemple simple les fonctions d'émulation attribuées aux machines du cluster ainsi qu'une partie des scripts de configuration générés par eWAN.

4 Implantation, expérimentations et résultats

4.1 Implantation d'eWAN

Un prototype du logiciel eWAN a été développé dans l'environnement Linux. Le logiciel est réalisé en PHP et utilise les bibliothèques MySQL, XML et GD. Ce logiciel active les outils d'émulation logicielle de latence NIST Net [CAR 03] et Netem, l'outil GtrcNet1 [KOD 03] d'émulation matérielle installé dans la grappe (cf. section 5). Pour le conditionnement de trafic la commande `tc` de Linux est utilisée. `Iperf` est utilisé comme générateur de trafic. Les expérimentations sont menées sur la grappe *Grid5000* de l'ENS Lyon. Ce cluster, totalisant actuellement 180 processeurs, comprend 24 serveurs bi-Xeon à 2GHz interconnectées par un

Figure 5: Phase de déploiement de la configuration choisie pour le nuage réseau

réseau Myrinet gigabit, une grappe de 12 machines Sun Fire V60x monoprocesseur à 3GHz dotées chacune de 2 Go de mémoire et de 3 interfaces réseau Ethernet Gigabit, interconnectées par un commutateur Ethernet Gigabit Foundry FES X448, une grappe de 60 nœuds biprocesseur Opteron à 2 GHz dotés de 2 Go de mémoire, d'un disque IDE de 80 Go et de deux interfaces réseau Ethernet Gigabit.

4.2 Premières expérimentations

Les figures 6 et 7 montrent les expérimentations menées sur les 12 machines à 3GHz et dotées de trois interfaces réseaux. Une étude comparitive des solutions NIST Net et Netem proposé dans le noyau Linux montre des performances comparables quant au taux d'utilisation du CPU pour l'émulation de latence (figure 6). Ce pourcentage est de l'ordre de 33% sur des machines dédiées. Par ailleurs, que ce soit NIST Net ou Netem, le coût CPU de l'emulation de delai n'augmente pas en fonction de la latence. Il augmente plutôt en fonction du débit du flux,c'est-à-dire en fonction du nombre de paquets à traiter par seconde. Pour cette expérience, le débit UDP est de 900Mbps (debit maximum d'environ 960Mbps) et le coût CPU est inférieur à 40%, on peut donc considérer qu'on est capable d'émuler n'importe quel delai usuel d'un nuage réseau de grille en maintenant un débit au Gigabit.

Par ailleurs, la figure 7 montre que les paramètres des émulateurs ont une importante influence sur la qualité de l'émulation. Ici, le paramètre *burst* de la discipline de service (`qdisc tbf`) permet d'émuler une limitation de débit. Ce paramètre correspond à la taille du tampon `tbf` : plus il est petit plus la limitation sera rapide mais moins elle supportera des rafales longues. La courbe claire montre le comportement dans le cas d'une taille de tampon définie à 1.54 Moctets alors que la courbe sombre est relative à une taille de 15.4 Moctets.

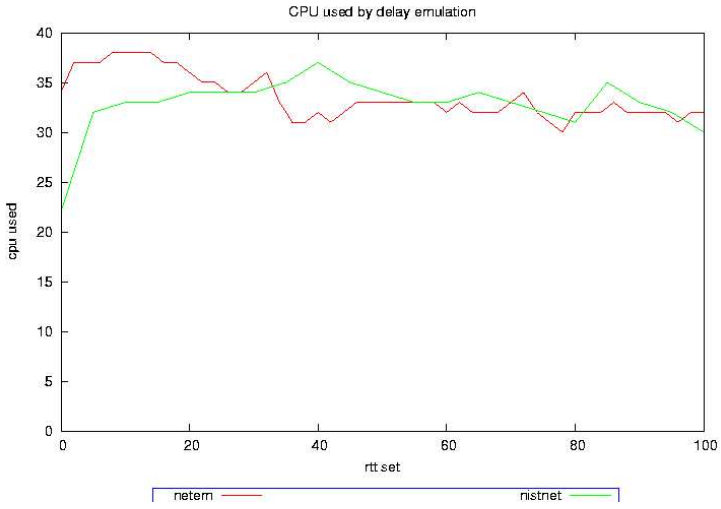

Figure 6: Pourcentage d'utilisation du CPU avec NIST Net et Netmem en fonction du délai

Dans ce deuxième cas, la limitation de trafic, n'est pas effective immédiatement (retard de 350ms).

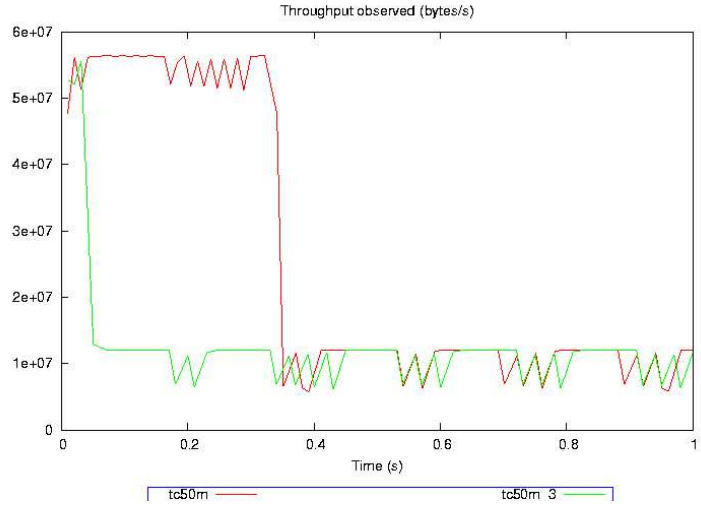

Figure 7: Paramétrage de `tc` dans eWAN

4.3 Scénarii d'utilisation de eWAN

Nous proposons d'utiliser eWAN pour trois types d'expériences différentes: 1) la validation et l'évaluation de performance d'algorithmes distribués sur une grille, 2) la validation de services réseaux de grille et 3) l'expérimentation d'outils et de services réseau classiques.

4.3.1 Optimisation du placement de données

Prenons un exemple du premier type. Supposons un utilisateur souhaitant valider un algorithme de placement de données sur la grille. Cet algorithme est basé sur l'évaluation des distances intersites et cherche à minimiser les temps d'accès ou les temps de transfert des données entre les nœuds de calcul et les nœuds de stockage. On a n sites de calcul et m sites de stockage. On connait les latences de tous les liens de ce graphe a $n.m$ sommets. Le chercheur, pour vérifier son algorithme, cherche à exécuter un benchmark de migration de données. Il souhaite évaluer trois types de configuration WAN: le mode équilibré (tous les sites équidistants), le mode déséquilibré à un pôle (un seul site excentré), le mode déséquilibré à deux pôles. Il doit pouvoir reproduire les mêmes conditions expérimentales pour chacune de ces trois expériences.

4.3.2 Expérimentation d'un service réseau de grille

Le service QOSINUS présenté précédemment a été validé expérimentalement dans le contexte du projet e-Toile [PRI 03]. De plus amples investigations avec des services réseaux différents et des applications variées sont nécessaires. L'approche émulation haut débit et à large echelle proposée par eWAN, permet de concevoir des scénarii d'expérimentation appropriés.

4.3.3 Validation d'un outil de mesure du réseau

Dans un contexte de réseaux longue distance haut débit, de nombreuses recherches sont menées pour proposer des outils de mesure de la performance d'une liaison entre deux extrémités. Les méthodes de mesures de débit disponible et les méthodes d'évaluation de la capacité d'un chemin sont de plus en plus étudiées. Par exemple, l'outil `Trace-Rate` [GOU 04] propose une méthode de découverte saut par saut de la capacité grâce à la technique *Packet Pair* et à une analyse fine de la distribution des mesures. Cette méthode a été validée en simulation, puis implantée dans Linux et évaluée expérimentalement sur le réseau expérimental à haut produit débit-délai DataTAG [MAR 04]. L'utilisation de l'émulateur eWAN permettra de confronter extensivement cette méthode à celles étudiées précédemment pour définir ses limites et ses réelles perspectives d'utilisation.

5 Emulation: état de l'art

eWAN s'insère dans un contexte très actif de développement de nouveaux outils expérimentaux de validation des protocoles et des applications distribuées.

Nous pouvons distinguer deux grandes catégories d'émulation dans les réseaux [FAL 99]: l'émulation de réseaux qui permet aux composants simulés de communiquer avec les protocoles implantés dans le monde réel et l'émulation dans un environnement logiciel, une extension de l'émulation de réseaux, qui permet d'exécuter directement dans un simulateur un protocole réel. Cette classification débouche sur deux types d'émulateurs :

- les émulateurs de liens qui permettent d'émuler un nuage réseau comme un ensemble de liens émulés : Hitbox [AHN 95], Ohio Network Emulator [ALL 97], Dummynet [8] [RIZ 97,

[8] `http://info.iet.unipi.it/~luigi/ip_dummynet`

RIZ 98], NIST Net [9] [CAR 03], Netem [10], GtrcNET-1 [11] [KOD 03]

- les émulateurs de réseaux virtuels qui permettent de simuler/émuler un nuage réseau en temps réel et d'y injecter du trafic réel : VINT/nse [BRE 00, FAL 99], IP-TNE [SIM 03], Virtual Routers [BAU 02, BAU 03], MicroGrid/MaSSF [12] [LIU 03], ModelNet [VAH 02], PlanetLab [CHU 03], Emulab/Netbed [13] [WHI 02], WAN in LAB [14].

eWAN appartient à cette deuxième catégorie mais s'appuie sur des émulateurs de liens existants pour introduire de la latence, des pertes ou des dupplications à l'intérieur du nuage réseau émulé.

L'émulation peut se faire à diffférents niveaux : au niveau de la couche Transport, Réseau ou même Liaison de données. Les émulateurs de liens n'utilisent généralement aucun support matériel spécifique pour introduire du délai ou des pertes : ils se contentent d'intercepter les paquets et les stockent dans des files d'attente pour leur appliquer le traitement logiciel adequate (selon les régles spécifiées par l'utilisateur). C'est le cas par exemple de Dummynet (tourne sur FreeBSD) et NIST Net (sur Linux) qui sont largement répandus. Les limitations de ces émulateurs logiciels sont principalement leur performance (émuler un grand délai à très haut débit nécessite un processeur rapide et beaucoup de mémoire) et leur précision (granularité du timer, nécessité d'un OS temps réel pour éviter que des tâches périodiques ne s'exécutent en retard). Ainsi, pour des raisons de performances, certains émulateurs de liens matériels commencent à voir le jour. GtrcNET-1, développée à l'AIST, est une boîte noire basée sur un FPGA permettant de connecter 4 ports Gigabit Ethernet, le FPGA pouvant être programmé pour faire de l'émulation de lien, de la génération de trafic ou du monitoring. GtrcNET-1 permet d'émuler une latence de 134 milli-secondes par port (soit un délai de 268 ms sur le lien) en maintenant un débit de 1 Gbit/s. Nous avons récemment intégré deux équipements de ce type dans eWAN pour réaliser l'émulation de lien et ainsi pouvoir faire des comparaisons avec NIST Net ou Netem qui sont des émulateurs de liens logiciels. Un émulateur matériel du même type est développé par une équipe du *Technology Transfer Group* au CERN [15].

Au-delà de l'émulation de liens, certains projets proposent des outils permettant d'émuler un réseau virtuel. VINT/nse, basé sur le simulateur `ns`, introduit du traitement d'événements en temps-réel : une interface entre le trafic réel du réseau et le simulateur `ns` capture les paquets du réseau, les injecte dans `ns` qui lui-même les ré-injecte après traitement dans le réseau via des *raw sockets*. ModelNet permet de générer une topologie réseau complète mais nécessite des modifications dans le noyau FreeBSD. Tout comme eWAN , ModelNet assigne certaines fonctions aux différents nœuds de l'émulateur (nœuds de bordure qui exécutent l'application, routeurs de cœur qui émulent le réseau virtuel, ...). Netbed, un descendant d'Emulab, utilise Dummynet, `ns` et des Vlans pour fournir un environnement réseau configurable. L'utilisateur peut définir via une interface Web une topologie virtuelle et les charactéristiques des noeuds réseau. IP-TNE est un émulateur réseau qui s'appuie sur de la simulation parallèle pour être scalable : tout s'exécute sur la même machine dans un environnement temps-réel. Les *Virtual*

[9] `http://snad.ncsl.nist.gov/itg/nistnet`
[10] `http://developer.osdl.org/shemminger/netem`
[11] `http://www.gtrc.aist.go.jp/gnet/gnet1e.html`
[12] `http://www-csag.ucsd.edu/projects/grid/microgrid.html`
[13] `http://www.emulab.net`
[14] `http://netlab.caltech.edu`
[15] `http://www.cern.ch/ttdb/Technologies/networkemulator`

Routers émulent un réseau à l'aide de routeurs virtuels implantés par des processus linux en espace utilisateur ; les paquets IP transitent à travers les routeurs virtuels qui eux-même communiquent entre eux via UDP ou IPC s'ils sont sur un même nœud.

Aux USA, les projets Emulab ou WAN in LAB visent la création d'émualteurs dans le but d'étudier principalement des problématiques réseau avec peu de nœuds de calcul. La particularité de eWAN est de s'appuyer sur un cluster possédant plusieurs dizaines de PCs standards pour construire un émulateur d'un nuage réseau de grille haute performance et très flexible : chaque nœud du cluster se voit attribuer, selon ses propres caractéristiques (quantité de mémoire, nombre et vitesse des processeurs, nombre et performance des interfaces réseau), une unique fonction permettant l'émulation globale du nuage réseau (émulateur de lien, routeur d'accès au nuage, routeur de cœur, générateur de trafic perturbateur, ...).

6 Conclusions et perspectives

Cet article a présenté un outil matériel et logiciel eWAN pour l'exploration et la validation expérimentale de nouvelles solutions de contrôle, de nouveaux services de grille et pour l'amélioration des performances des communications dans la grille. L'évaluation expérimentale doit en effet compléter l'analyse et la simulation qui atteignent leurs limites lorsque l'on atteint des échelles de performance et d'entités communicantes très importante.

La conception d'expériences sur cet émulateur est un axe que nous souhaitons explorer dans la suite de ces travaux.

Les résultats obtenus sur cette plate-forme d'émulation pourront le cas échéant être confrontés aux valeurs effectives obtenues sur les plate-formes expérimentales auxquelles la communauté grille est raccordée au niveau français, au niveau européen et international [16].

References

[AHN 95] Ahn J. S., Danzig P. B., Liu Z.Yan L., Evaluation of TCP Vegas: Emulation and Experiment, *ACM SIGCOMM Computer Communication Review*, 25, 4, October 1995, 185-205.

[ALL 97] Allman M., Caldwell A.Ostermann S., ONE: The Ohio Network Emulator, Technical Report TR-19972, August 1997, Ohio University Computer Science.

[BAU 02] Baumgartner F., Braun T.Bhargava B., Virtual Routers: A Tool for Emulating IP Routers, *In the 27th IEEE Conference on Local Computer Networks (LCN 2002)*, Tampa, USA, November 6-8 2002.

[BAU 03] Baumgartner F., Braun T., Kurt E.Weyland A., Virtual Routers: A Tool for Networking Research and Education, *ACM SIGCOMM Computer Communication Review*, 33, 3, July 2003, 127-135.

[BER 03] Berman F., Fox G.Hey A. J., *Grid Computing: Making The Global Infrastructure a Reality*, 2003, ISBN: 0-470-85319-0.

[16] EGEE, GRANDE, GARDEN, PlanetLab

[BRE 00] Breslau L., Estrin D., Fall K., Floyd S., Heidemann J., Helmy A., Huang P., McCanne S., Varadhan K., Xu Y.Yu H., Advances in Network Simulation, *IEEE Computer*, 33, 5, May 2000, 59-67.

[CAR 03] Carson M.Santay D., NIST Net: a Linux-based Network Emulation Tool, *ACM SIGCOMM Computer Communication Review*, 33, 3, July 2003, 111-126.

[CHU 03] Chun B., Culler D., Roscoe T., Bavier A., Peterson L., Wawrzoniak M.Bowman M., PlanetLab: An Overlay Testbed for Broad-Coverage Services, *ACM SIGCOMM Computer Communication Review*, 33, 3, July 2003, 3-12.

[FAL 99] Fall K., Network Emulation in the VINT/NS Simulator, *In Proceedings of the fourth IEEE Symposium on Computers and Communications*, Red Sea, Egypt, July 1999, 244-250.

[FER 04] Ferrari T., Travostino F.al., Grid Network Services, work in progress, , *https://forge.gridforum.org/projects/ghpn-rg/document/draft-ggf-ghpn-netservices-1/en/1*, 2004.

[FOS 99] Foster Y.Kesselman C., *The Grid - Blueprint for a New Computing Infrastructure*, 1999, ISBN: 1-55860-475-8.

[FRA 03] Franck Bonnassieux Mathieu Goutelle P. P., Network Services - final report, Rapport de recherche, 2003, European DataGrid project.

[GOU 04] Goutelle M.Primet P., Trace-Rate, a non-intrusive method for measuring the hop by hop performances of a path, *In Proceedings of the 2004 International Conference on Communications*, Paris, France, 2004, IEEE Communication Society.

[KOD 03] Kodama Y., Kudoh T., Takano T., Sato H., Tatebe O.Sekiguchi S., GNET-1: Gigabit Ethernet Network Testbed, *In Proceedings of the IEEE International Conference Cluster 2004*, San Diego, California, USA, September 20-23 2003.

[LEF 01] Lefèvre L., Pham C., Primet P., Tourancheau B., Gaidioz B., Gelas J.Maimour M., Active Networking Support for the Grid, Ian W. Marshall Scott Nettles N. W., , *IFIP-TC6 Third International Working Conference on Active Networks, IWAN 2001*, 2207 *Lecture Notes in Computer Science*, 2001, 16-33, ISBN: 3-540-42678-7.

[LIU 03] Liu X.Chien A., Traffic-based Load Balance for Scalable Network Emulation, *In Proceedings of the ACM Conference on High Performance Computing and Networking (SC2003)*, Phoenix, Arizona, November 2003.

[MAR 04] Martin-Flatin J. P.Vicat-Blanc Primet P. E., *High Performance Networks and Services for Grid : the IST DataTAG project experience*, Elsevier, dec 2004.

[PRI 03] Primet P. V.-B., Chanussot F., Blanchet C., Lacorne N.d'Anfray. P., E-Toile: High performance Grid Middleware, *In IEEE International Cluster Conference. Grid Demo session*, 2003.

[PRI 04] Primet P. V.-B.Chanussot F., Network Quality of Service in Grid environments: the QoSinus approach, *In Proceedings of the IEEE International Broadnet Conference, GridNets Workshop.*, 2004.

[RIZ 97] Rizzo L., Dummynet: A Simple Approach to the Evaluation of Network Protocols, *ACM SIGCOMM Computer Communication Review*, 27, 1, January 1997, 31-41.

[RIZ 98] Rizzo L., Dummynet and Forward Error Correction, *In Proceedings of the USENIX 1998 Annual Technical Conference*, New Orleans, USA, June 15-19 1998.

[SIM 03] Simmonds R.Unger B. W., Towards Scalable Network Emulation, *Computer Communications*, 26, 3, February 2003, 264-277, Elsevier Science.

[VAH 02] Vahdat A., Yocum K., Walsh K., Mahadevan P., Kostic D., Chase J.Becker D., Scalability and Accuracy in a Large-Scale Network Emulator, *In Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI)*, December 2002.

[VIC 03] Vicat-Blanc Primet P., High Performance Grid Networking in the DataGrid Project, *special issue Future Generation Computer Systems*, 1, jan 2003, Elsevier.

[VIC 04] Vicat-Blanc Primet P., Bonnassieux F.Harakaly R., Network monitoring in the DataGRID project, *International Journal of High Performance Computer Applications*, 1, august 2004.

[WHI 02] White B., Lepreau J., Stoller L., Ricci R., Guruprasad S., Newbold M., Hibler M., Barb C.Joglekar A., An Integrated Experimental Environment for Distributed Systems and Networks, *In Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI)*, December 2002, 255-270.