Action Concertée Incitative
[ACI]
Globalisation des Ressources
Informatiques et des Données
[GRID]

Rapport de fin de projet retenu dans le cadre de la campagne ACI GRID 2001

RAPPORT DE FIN DE PROJET

SUPPORT RESEAUX ET INTELLIGENCE POUR LA GRILLE

Projet

JE RESAM

Décembre 2003

Adresse du Laboratoire porteur du projet :

LIP/RESO (ex. RESAM)

ENS Lyon

46 Allée d'Italie

69364 Lyon Cedex 07

Action Concertée Incitative
[ACI]
Globalisation des Ressources
Informatiques et des Données
[GRID]

Rapport de fin de projet retenu dans le cadre de la campagne ACI GRID 2001

0 . Participants au projet

Faycal BOUHAFS (Ingénieur INRIA)
Benjamin GAIDIOZ (thèse soutenue le 18 décembre 2003)
Jean-Patrick GELAS (thèse soutenue le 5 décembre 2003)
Laurent LEFEVRE (Chargé de Recherches INRIA)
Moufida MAIMOUR (thèse soutenue le 25 novembre 2003)
Congduc Pham (MCF Univ Lyon 1, coordinateur du projet)
Pascale VICAT-BLANC/PRIMET (MCF ECL, détachement CR INRIA)

1 OBJECTIFS DU PROJET (résumé)
1.1 Résumé Analytique

2 MéTHODE DE TRAVAIL SUIVIE
2.1 Présentation

3 RéALISATIONS
3.1 Principe
3.2 Présentation du calendrier

4 PERSPECTIVES
4.1 Améliorations de l'existant
4.2 Directions de travaux ultérieurs

Action Concertée Incitative
[ACI]
Globalisation des Ressources
Informatiques et des Données
[GRID]

1 Objectif du projet

1.1 Résumé Analytique

Objectifs, contexte et description du projet

La globalisation des ressources informatique dans ce que l'on appelle communément une grille de calcul répond à la fois à un souci économique où l'on cherche à partager des équipements de calcul coûteux mais aussi à des soucis correspondants au passage au facteur d'échelle pour l'étude de problémes de plus en plus grand. L'idée est attrayante mais la réalisation pratique est difficile.

Parmi les nombreux problèmes soulevés par la mise en oeuvre d'une grille efficace, ceux liés plus particulièrement à l'interconnexion en réseaux des ressources informatiques et à la délivrance efficace des données font l'objet de cette ACI «Support Réseau et Intelligence pour la Grille». Dans un tel contexte de globalisation des ressources, la composante communication est trés importante pour pouvoir travailler « confortablement sur la grille », c'est-à-dire avec le moins de latence et le plus de débit possible. Partager des fichiers, transférer rapidement des programmes et des quantités de données volumineuses éventuellement sur plusieurs sites, pouvoir synchroniser rapidement et facilement des tâches exécutées sur plusieurs sites, déployer des algorithmes efficaces d'opérations collectives etc., sont autant d'exigences que les utilisateurs mettront sur une grille de calcul.

Cependant, l'infrastructure réseau traditionnelle sur laquelle repose la grille est une interconnexion de réseaux IP ou simplement l'Internet mondial et ce type d'infrastructure souffre de grandes faiblesses, en particulier au niveau de la fourniture de garantie de Qualité de Service (QoS) et du support de communications multipoints (multicast). De manière générale, haute performance, hétérogénéité et distribution sont les caractéristiques réseaux majeures d'une grille et c'est dans ces dimensions là que la jeune équipe RESO (ex-RESAM) se propose de mener des études et des expérimentations.

Pour pouvoir offrir un service réseau de grande qualité aux utilisateurs de la grille, de nouveaux protocoles adaptés aux besoins spécifiques des applications doivent être mis en oeuvre. La trés grande diversité des flux en terme de taille de message, de communications (point à point ou multicast), type de message (contrôle ou données), requiert une intelligence dans le réseau pour mieux supporter les besoins de la grille. Les axes de recherches que nous souhaitons explorer sont donc les suivants:

Environnement d'exécution actif haute-performance pour la grille (L. Lefèvre): il s'agit de développer et de déployer un environnement actif apte à supporter les performances des réseaux actuels (Gbit/s). Cet environnement doit ainsi autoriser la mise en œuvre performante de services spécifiques aux flux de données des applications de la grille.

Optimisation du transport TCP pour la grille (P. Vicat-Blanc/Primet): il s'agit d'explorer des solutions simples et efficaces pour l'amélioration du protocole de transport TCP pour les transferts de données dans une grille et en particulier des transferts de données massifs (plusieurs giga-octets).L'objectif est d'optimiser le débit utile offert aux applications.

Optimisation du multipoint fiable pour la grille (C. Pham): il s'agit de fournir un support communication de multipoint fiable de données qui permet de réduire la latence des communication de diffusion sur la grille.

2 MéTHODE

2.1 Présentation

Pour l'instant, les grilles sont essentiellement « académiques », avec des industriels bien sûr mais financées par des fonds de recherche publics, et il faut distinguer ces grilles des grilles "industrielles". En effet, si dans le premier cas le débit des accès est grand, commençant au Gbits/s et pouvant aller jusqu'à plusieurs Gbits/s, les compagnies privées concernées par la grille n'ont souvent que des débits beaucoup plus faibles : de quelques centaines de Kbit/s jusqu'à 1 ou 2 Mbit/s le plus souvent. Seules les trés grandes compagnies pourraient se permettre d'avoir des accès de l'ordre de 34Mbits/s à 155Mbits/s (OC-3). La différence est énorme ! C'est dans ce contexte générale de grilles (et non pas uniquement des grilles à trés haut-débit) que nos travaux s'inscrivent.

Les études dans cette ACI peuvent être séquencées en 3 parties :

1. Evaluer et caractériser les besoins de la grille
2. Proposition et évaluation de nouveaux mécanismes
3. Prototypage, expérimentations

Notre équipe participe à plusieurs projets européens et nationaux dans lesquels nous avons la possibilité d'expérimenter les solutions proposées. Nous utilisons par conséquent à la fois l'évaluation analytique, la simulation et le prototypage pour valider nos propositions.

3 RéALISATIONS

3.1 Principe

Problématique :

Optimisations de TCP pour la grille

Actuellement les interconnexions de grille sont effectuées avec des liens de plusieurs Gbits/s. Le protocole TCP (Transmission Control Protocol) de l'Internet ne fonctionne pas de manière satisfaisante dans un environnement où le débit est grand, et où bien souvent le temps aller-retour est grand. Les problèmes très souvent rencontrés sont une sous-utilisation de la bande passante disponible et une très mauvaise réponse aux congestions qui peuvent se produire dans le réseau. Ces problèmes sont importants dans les grilles de calcul car TCP est utilisé de manière intensive pour transporter de très grande quantité de données. Nos travaux cherchent donc à optimiser ce protocole pour des environnements très haut-débit.

Environnements d'exécution actifs

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 noeuds (routeurs) actifs qui réalisent des opérations personnalisées sur les flux de données qui le traversent, et qui autorise les utilisateurs, les opérateurs, ou les fournisseurs de services à injecter leurs propres programmes dans les n\oe{}uds du réseau, permettant ainsi de modifier, stocker (cacher) ou rediriger le flux de données à travers le réseau. Les réseaux actifs mettent ainsi en oeuvre des services réseaux plus complexes que de la transmission simple de paquets de données, ce qui a pour effet de consommer plus de cycles CPU ou de mémoire dans les équipements réseaux.

Les performances des technologies de transmissions de données et de commutations ne cessent de progresser. Les débits supportés par les liens ne cessent d'augmenter. Il existe aujourd'hui sur le marché des routeurs supportant plusieurs liens à 10 Gbps .

La courbe de progression réalisée dans le domaine des technologies réseaux suit une pente supérieure à celle de la progression realisée dans le domaine des processeurs, même si les progrès de ceux-ci suivent toujours la fameuse loi de Moore (doublement des performances des circuits integrés tous les 18 mois). Georges Gilder a énoncé une loi identique qui indique que la bande passante des réseaux augmente trois fois plus vite que la puissance des ordinateurs. Les calculs d'Eric Schmidt (CEO de Google, ex-CEO de Novell) et les études de Probe Research montrent que la bande passante double en fait tous les ans depuis 1997 (cf. Wired, juillet 2002).

Les différentes expérimentations de nœuds actifs logiciels menés jusqu'à présent ont limité leur support à des flux à quelques Mbit/s. Afin de proposer des solutions de réseaux actifs réalistes, il est primordial d'implémenter des environnements actifs aptes à supporter

les bandes passantes des réseaux d'accès actuels. Ils pourront ensuite donc être déployés sur des architectures de grilles de calcul à haute vitesse.

Multicast fiable

Le multicast IP fournit au niveau réseau un support efficace pour un grand nombre d'applications: dissémination de données, simulation interactive distribuée, vidéoconférence et applications coopératives. Certaines de ces applications, en plus de l'efficacité du routage, nécessitent une grande fiabilité dans la délivrance des données. C'est le cas pour les calculs de type grille. Le probléme de la fiabilité en point à point est bien maîtrisé et de bonnes solutions ont été déployées. Par contre l'assurance de la fiabilité dans le contexte du multicast est un probléme plus ardu et les solutions sont moins évidentes, surtout sur des réseaux étendus. Par exemple, la fiabilité nécessite des messages de contrôle tels que les acquittements (positifs ou négatifs) qui vont remonter jusqu'à la source. Ce trafic peut-être extrêmement pénalisant pour la communication multicast : baisse de débit, augmentation de la latence de bout en bout, perte de synchronisation des récepteur... Le défi est donc de garantir la fiabilité tout en assurant le passage à l'échelle avec un grand nombre de récepteurs géographiquement distribués comme cela est le cas pour une grille de calcul.

Une autre propriété souhaitable, et non des moindres, est la co-habitation des flux multicast avec les autres flux unicast gérés essentiellement par TCP. C'est ce que l'on nomme communément le contrôle de congestion. Dans les versions actuellement déployées de TCP par exemple ce contrôle est une combinaison de la phase de slow-start avec une phase de congestion avoidance. Ce mécanisme permet aux entités sources des connexions TCP de partager relativement équitablement la bande passante disponible. Dans le cas du multicast, il est ardu de concevoir des mécanismes de contrôle de congestion qui offrent à la fois une bonne utilisation du réseau pour tous les récepteurs, et une compatibilité avec les flux TCP.

Résultats :

Optimisations de TCP (Resp. Pascale Vicat-Blanc/Primet)

Ces travaux concernent l'étude des protocoles de transport trés haute performance sur réseaux gigabit longue distance. Nous étudions finement les nouvelles propositions de l'IETF basées sur une modification de l'algorithme AIMD du protocole TCP. Une analyse des limitations dues aux traitements locaux a été faite. Une série d'expérimentations dans un environnement gigabit réel avec un émulateur de latence (NistNet) optimisé a été conduite. En mettant ces travaux sur les protocoles longues distance en perspective avec les approches utilisées dans les réseaux courte distance, nous avons eu l'idée d'explorer l'approche de contrôle de congestion par contrôle de flot «file à file» (Network of Queue, NOQ), qui lève un certain nombre des limitations de TCP dans un cadre trés haut débit et longue latence. Cette approche, courante dans les réseaux haute performance tels que Myrinet, un peu étudiée dans le cadre des réseaux ATM, n'a jamais été explorée dans un contexte IP. Les premiers résultats obtenus sont trés prometteurs.

Environnement actifs (Resp. Laurent Lefévre)

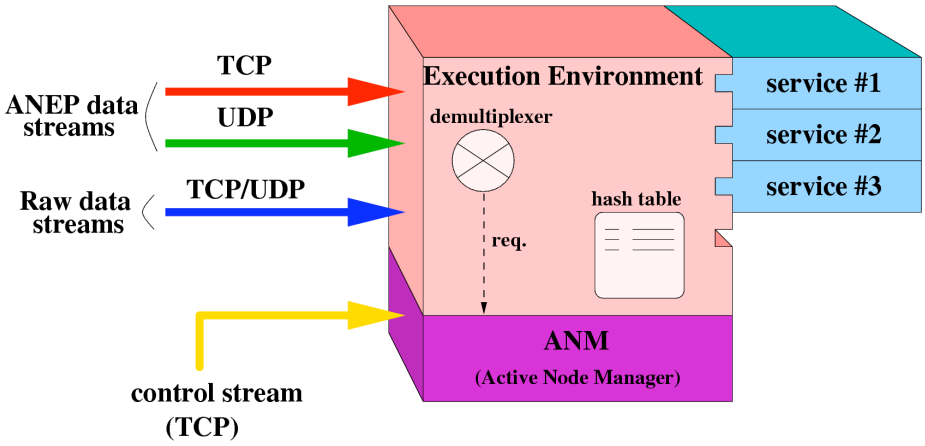
Les travaux menés dans le cadre de cette ACI-GRID ont permis la réalisation et le déploiement de l'environnment de réseaux actifs Tamanoir. L'environnement Tamanoir fournit aux utilisateurs la possibilité de déployer et de maintenir dynamiquement des routeurs actifs, appelés TAN (Tamanoir Active Node), distribués sur un réseau à grande échelle. Chaque TAN supporte des connexions multiples ainsi que différents services au même moment (approche multithreadée). Les services sont déployés dynamiquement entre les noeuds actifs en suivant la progression des flux de données ou à partir de dépots de services.

L'architecture d'un noeud actif Tamanoir repose sur un modèle distribué (grappe de routeurs) apte à supporter les demandes des flux transportés. Le modèle Tamanoir remet en cause la conception des routeurs en instaurant une classification des traitements à effectuer dans les équipements. L'architecture repose sur une répartition en couches en fonction des classes de services embarqués :

- services réseaux légers (forwarding, marquage de paquets, ajouts d'options,…) qui nécessitent peu de traitement sur les paquets de données et sont déployés dans le noyau ou sur les cartes d'interface programmables;

- services actifs moyens (QoS, agrégation, filtrage,…) qui nécessitent une intelligence et des ressources de haut niveau (services Java, mémorisation d'états…) mais peu de puissance de calcul embarqués dans l'environnement d'exécution (espace utilisateur);

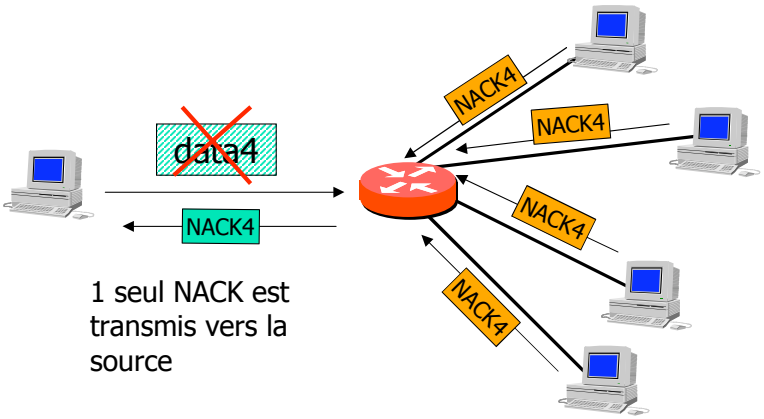
- services actifs lourds (compression à la volée, filtrage) déployés et distribués sur une architecture parallèle (grappe de routeurs) afin de bénéficier d'une forte puissance de calcul extensible.



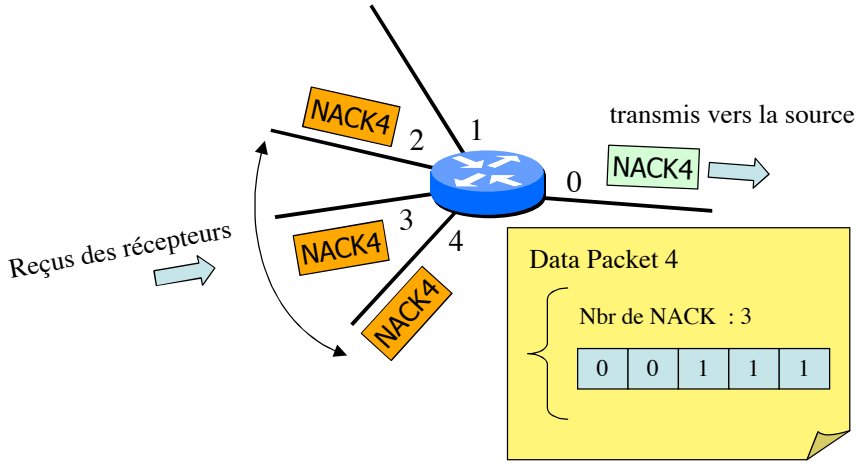
Les services bénéficient ainsi de l'architecture entièrement distribuée de Tamanoir (multithreading, équilibrage de charge). Cette conception permet d'appréhender de futurs développements matériels et logiciels pour une mise en oeuvre dans des équipements industriels. Différentes expérimentations de Tamanor ont été menées sur des plate-formes locales ou longue distance dans le cadre des projets RNRT VTHD++ et RNTL Etoile.

Multicast fiable (Resp. Congduc Pham)

Afin d'améliorer les performance du multicast, nous avons étudié les solutions utilisant l'assistance des routeurs, ces solutions peuvent être appelées solutions actives lorsqu'un environnement d'exécution actif est utilisé, ce qui est généralement le cas pour le prototypage des solutions. Ces protocoles utilisent les routeurs dans l'infrastructure réseau pour effectuer des traitements sur les flux (que nous appelerons service actif). Par exemple, des services d'aggrégation des messages de contrôle peuvent être mis en place pour éviter l'implosion des messages de contrôle au niveau de la source. La figure suivante illustre ce cas.



L'implémentation d'un tel service peut se faire trés simplement, et le plus important est de maintenir le coût d'un service actif à un niveau trés bas. La figure suivante montre schématiquement comment un tel service peut-être implémenté avec des structures de données simples.



Nous avons donc proposé des services actifs visant à réduire les délais de recouvrement et à améliorer le contrôle de congestion. Ces services sont :

1/ l'élection dynamique d'un retransmetteur: un routeur actif d'assistance peut élire un récepteur pour retransmettre un paquet perdu par un de ces voisins.

2/ la détection rapide des pertes dans les routeurs: un routeur actif d'assistance peut générer un NAK vers la source s'il détecte une perte de séquence dans le flux des paquets.

3/ l'aggrégation des RTTs: les routeurs actifs d'assistance participent à l'aggrégation des RTTs par segment afin de générer une valeur de RTT plus précise permettant une régulation du débit plus fluide par la source (contrôle de congestion).

4/ le partionnement des récepteurs en sous-groupes pour améliorer la gestion des groupes hétérogénes.

Un protocole de multicast fiable actif appelé DyRAM a été proposé. Il utilise les services actifs décrits ci-dessous pour réduire les latences de recouvrements pour les applications distribuées. Dans un premier temps, nous avons évalué les apports de nos propositions de maniére analytique et par simulation. Ensuite, nous avons implémenté la solution DyRAM en utilisant l'environnement Tamanoir et réalisé des tests sur la plate-forme VTHD.

Indicateurs :

Nos travaux ont été publiés dans de nombreuses conférences internationales et dans des revues de grande renommée scientifique. Les prototypes sont distribués dans le cadre du projet e-Toile, et de nombreuses expérimentations ont été effectuées.

Une démonstration des résultats de cette ACI a été effectuée à IPDPS 03 (21 avril 2003). Une démonstration liée aux thémes de cette ACI a été faite lors des démonstrations du projet RNTL e-Toile.

Des logiciels ont été développés : Environnement actifs Tamanoir, librairie pour le multicast fiable MFTP.

Le site WWW du projet est
http://www710.univ-lyon1.fr/~cpham/PROJETS/ACIgrid.htm
il recense toutes les publications, présentations et liens relatifs à cette ACI.

3.2 Présentation du calendrier

Année 1&2

T0+6	T0+12	T0+18	T0+24
Evaluation et caractérisation des besoins grille. Identification des requis.	Prototypage, test et déploiement sur une plate-forme de test en local	Expérimentation et validation sur une plate-forme locale.	Plate-forme de démonstration

4 PERSPECTIVES - Directions de travaux ultérieurs

Nous continuons ces travaux sur plusieurs pistes nouvelles : techniques de monitoring réseaux pour estimer la bande passante disponible sur les liens haut-débit des nouvelles grilles de calcul haute-performance, étude de TCP dans des environnements à bande passante très variable telle que ceux qui pourront être rencontrés sur les grilles dédiées avec réservation de ressources, déport de traitements dans les cartes d'interface pour construire des routeurs actifs haute-performance.