Rapport scientifique du projet de réseau régional d'expérimentations avancées sur un réseau haut débit ATM de Grenoble, C3I2 (CEA, CNET, CNRS, IMAG, INRIA)

Jean-Luc Archimbaud

Rapport scientifique du projet de réseau régional d'expérimentations avancées sur un réseau haut débit ATM de Grenoble, C3I2 (CEA, CNET, CNRS, IMAG, INRIA)

30 septembre 99 Version finale

Rédacteur principal : Jean-Luc Archimbaud CNRS/UREC (http://www.urec.cnrs.fr/jla)
coordonnateur du projet

Ce rapport établit un compte-rendu scientifique de la mise en place technique d'un réseau expérimental et décrit certaines expérimentations qui ont pu être menées avec cet outil. Il tire aussi des leçons sur la réussite et la gestion d'un tel projet, sur l'administration d'un réseau multi-partenaires et sur l'utilisation de la technologie ATM.

Le premier chapitre, synthèse du rapport, présente le projet et ses conclusions. Les expérimentations sont décrites dans le chapitre 2 et dans les annexes au chapitre 7. L'historique et l'organisation du projet est le sujet du chapitre 3, les architectures physique et logique du réseau celui du 4 et les problèmes techniques de mise en place en 5. Les conclusions techniques sur la technologie ATM sont détaillées au chapitre 6.

Des informations sont aussi disponibles en ligne : http://www.urec.fr/C3I2/.

1. Présentation et bilan de l'expérimentation C3I2

Le projet, tel que décrit dans les dossiers présentés à la région Rhône-Alpes, visait à **installer un réseau à haut débit** entre des partenaires grenoblois de la recherche pour conduire pendant un an des **expérimentations de protocoles** (ATM, IP, IPv6, …), de **services** comme la vidéoconférence et des **applications avancées** issues des laboratoires de recherche de Grenoble (calcul distribué, espace de travail réparti, …). Il devait aussi **préparer l'évolution des infrastructures et des services de réseau de campus, métropolitains et régionaux** en permettant aux ingénieurs d'acquérir un savoir-faire dans les nouvelles technologies, d'expérimenter certaines architectures et modèles d'administration des équipements. Ces objectifs n'ont pas été modifiés et on peut considérer qu'**ils ont été atteints**.

Le sigle **C3I2 est la concaténation des initiales des 5 partenaires : CEA** Grenoble (Commissariat à l'Energie Atomique), **CNET** (Centre National d'Etude des Télécommunications), **CNRS** (Centre National de la Recherche Scientifique), **IMAG** (Institut d'Informatique et de Mathématiques Appliquées de Grenoble) avec ses 2 tutelles universitaires INPG et UJF, et l'**INRIA** Rhône-Alpes (Institut National de la Recherche en Informatique). Le projet a été lancé au printemps 96 sous l'impulsion de Jean-Pierre Verjus Directeur de l'INRIA Rhône-Alpes. Il a obtenu 2 financements régionaux dans le cadre des opérations mi-lourd 96 et mi-lourd 97.

Fin 97 ont été mis en place, par l'opérateur France Télécom avec un contrat de 1 an, des liaisons **ATM à 155 Mb/s pour relier les 5 sites grenoblois** des partenaires : Polygone (CEA, CNRS), Felix Viallet (IMAG, CNRS, …), Meylan (CNET), Domaine universitaire (IMAG, CNRS, …), Montbonnot (INRIA). Après la phase de conception et la mise en place des équipements télécoms, les expérimentations avancées ont pu être menées du printemps 98 à mi 99.
Fin 98, le contrat d'un an C3I2 arrivait à échéance. Les partenaires décidèrent de maintenir avec leurs fonds propres cette infrastructure de réseau jusqu'en juillet 99, date de fin du projet et de l'arrêt des liaisons C3I2.

Ce projet a d'abord montré que **Grenoble est une très bonne métropole pour un projet de réseau métropolitain de recherche**. En effet y sont implantés différents organismes de recherche, sur plusieurs sites, certains étant même éclatés sur plusieurs pôles, comme l'IMAG par exemple. Il existe ainsi une dispersion géographique qui nécessite des liaisons télécoms entre les laboratoires. De nombreuses recherches sont menées à Grenoble sur des applications réseau, mediaspace par exemple, ou qui nécessitent un réseau haut débit, pour des calculs de simulation à distance par exemple. La variété des expérimentations prévues (cf chapitre 2) le montre. Donc les applications et les utilisations sont présentes dans la métropole. Et troisième élément, le potentiel ingénieur réseau dans chaque organisme est élevé, ce qui permet de mettre en place des solutions techniques innovantes. Tous les ingrédients nécessaires étaient donc réunis.

Mais, deux points négatifs dans le bilan

Le plus flagrant a été le **peu d'expérimentations de laboratoire** qui ont pu être faites sur ce réseau. Malgré la liste importante des projets intéressés, peu ont été menés à

terme. Une conjoncture défavorable n'a pas facilité la tâche : le calendrier de la disponibilité du réseau. Les liaisons ont été installées en décembre 97. Après de nombreux problèmes techniques de mise en place des équipements télécoms dus à la nouveauté de la technologie ATM utilisée (cf paragraphe 5), le réseau n'a été ouvert aux expérimentations qu'au printemps 98, donc peu de temps avant l'été. A la rentrée scolaire, le service devait s'arrêter fin décembre (ce n'est qu'en décembre 98 que les organismes ont décidé de poursuivre jusqu'à mi 99). Devant ce calendrier et ses incertitudes, certaines équipes de recherche ont préféré ne pas se lancer dans des installations lourdes. Mais d'autres raisons plus profondes expliquent ce peu d'engouement :

. **De nombreuses équipes de recherche n'ont pas le personnel technique nécessaire** pour mener, en un an, des expérimentations sur des équipements grandeur nature. Le chemin est long entre un prototype qui marche entre 2 stations identiques sur un réseau local et une utilisation entre 2 sites avec des matériels hétérogènes et des logiciels propriétaires. Pour résoudre ces problèmes et bien d'autres, c'est beaucoup de temps de technicien et d'ingénieur, à défaut de thésards, qui est nécessaire. Les chercheurs ne sont pas formés et n'ont pas le temps de s'investir dans ces tâches.

. Nous avions **trop d'expérimentations diverses et ponctuelles sans lien l'une avec l'autre**. Ainsi il fallait refaire tout un travail de base pour résoudre les problèmes de matériel, de logiciel, … différents pour chaque expérimentation ; et une équipe de recherche ne pouvait pas bénéficier de l'expérience d'une autre. Il n'y a pas eu de synergie entre les chercheurs autour de cet outil.

. Les produits créés dans les laboratoires tiennent compte des potentialités du réseau. Certains, comme mediaspace et calliope, ont été conçus bridés pour fonctionner avec des faibles ou moyens débits, seuls disponibles lors de leur conception. Et sans préparation C3I2 a amené des très hauts débits qui nécessitaient de modifier fortement tous les programmes pour bénéficier de cet apport de bande passante (autres formats de compression, d'affichage, autres périphériques nécessaires, …). Il est certain que si l'outil C3I2 avait été annoncé fermement plusieurs mois à l'avance, certains laboratoires auraient orienté leurs développements pour l'utiliser pleinement. Cela rejoint **le problème de la poule et de l'œuf, faut-il mettre en place l'outil avant le besoin ?**

Mais ce faible nombre d'expériences n'est pas spécifique à Grenoble, sur SAFIR (réseau expérimental national, préfiguration de Renater 2, cf chapitre 4), le bilan pourrait être identique.

L'autre point négatif est l'arrêt du service ATM C3I2 délivré par France Télécom entre les sites, **remplacé par un service SMHD** de France Télécom qui n'offre aucun avantage technique supplémentaire. C3I2 fonctionnait parfaitement avec une bonne coordination technique et dans les derniers mois transportait le trafic de production pour décharger ARAMIS sans aucun problème. Il était donc la solution technique idéale pour remplacer temporairement ARAMIS à Grenoble qui arrivait à saturation et ne pouvait plus évoluer. France Télécom n'a pas voulu maintenir le service ATM-C3I2 au tarif préférentiel « expérimental » que nous avions négocié. Les liaisons ont donc été coupées fin juillet. Pour le trafic de production, de nouvelles liaisons SMHD France Télécom vont être installées entre les sites C3I2 mi-septembre. Les sites vont néanmoins réutiliser leurs équipements télécoms C3I2 pour se connecter à ce nouveau réseau, et reconstruire l'architecture ATM de C3I2 sur les liens SMHD, ce qui est un moindre mal. Il y a dans cette transition forcée un gaspillage de temps important (mise en place d'une nouvelle organisation, négociations financières, configuration des équipements, …).

Quels apports pour la communauté ?

Ce réseau a d'abord **permis de valider certains développements de laboratoires ou de mettre en lumière divers problèmes** liés à l'utilisation des plates-formes expérimentales en vraie grandeur dans un réseau entre des sites distants. Cela a certainement permis d'adapter les développements des laboratoires et d'orienter les recherches en tenant compte des réalités. Côté services expérimentaux, **C3I2 a initié de nombreux utilisateurs à des outils un peu prospectifs comme la vidéoconférence**. La dynamique C3I2 à donnée à certains l'envie d'essayer ces outils, sans craindre de perturber le trafic de production. L'expérience de télé-enseignement par exemple (cf annexe 7.4) n'avait pas été prévue avant la mise en place de C3I2, c'est la disponibilité de ce réseau expérimental qui a donné envie aux enseignants de se lancer dans cette aventure. **Un réseau expérimental permet de donner libre court à toute expérience sans craindre de perturber les autres**.
C3I2 à Grenoble a aussi parfaitement été en phase avec la mise en place du réseau SAFIR, réseau ATM national d'expérimentation pour la recherche en France (cf chapitre 4), techniquement, contractuellement, et dans le temps. C3I2 a été relié à cette infrastructure nationale et tous les partenaires C3I2 ont pu accéder à SAFIR directement, sans surcoût. Des sites comme l'INRIA, le Polygone-CEA ou le domaine universitaire ont pu mener des expériences nationales (IPv6, vidéoconférence, …) et véhiculer une partie de leur trafic Internet sur SAFIR. Si C3I2 n'avait pas été en place, il n'est pas certain qu'une liaison SAFIR serait arrivée à Grenoble.

C'est certainement pour les ingénieurs des organismes que le projet C3I2 a été le plus bénéfique. Durant ces mois, ceux-ci ont :

. appris à se connaître, à communiquer entre eux et à **travailler en équipe**, tâche pas forcément facile lorsque l'on vient d'organismes différents. Cette habitude a déjà porté ses fruits après C3I2. Ainsi ce groupe technique a très rapidement rédigé le cahier des charges qui a abouti au choix de SMHD et travaille sereinement à la définition d'un réseau métropolitain à Grenoble.

. appris à **résoudre avec méthode les problèmes techniques complexes de réseau**. Ainsi lors des premiers problèmes de mise en place rencontrés, des techniciens très critiques sur la technologie ATM criaient immédiatement que « le réseau ATM de marchait pas », mettant en cause l'architecture de base. En fait, dans de nombreux cas, le problème venait de la périphérie : application, carte de station, … Après plusieurs erreurs de ce type, l'habitude plus scientifique de diagnostique précis des problèmes, de recherche dichotomique … s'est installée.

. **acquis une bonne connaissance technique d'ATM**, et ont déduit, après expérience, les avantages et les limites de cette technologie; analyse qui va être très utile dans la définition et la mise en place des infrastructures métropolitaines de réseau à Grenoble.

Dans un plan plus économique, le matériel acheté par les organismes sur C3I2, équipements réseau ou stations multimédia, va être réutilisé sur le réseau métropolitain SMHD qui se met en place. **L'investissement matériel aura être parfaitement rentabilisé**.

Le dernier apport est que **C3I2 a déchargé le réseau ARAMIS** d'une partie du trafic de production à un moment où il en avait bien besoin, certaines applications vitales ne pouvaient plus fonctionner sur certaines liaisons ARAMIS en particulier l'avenue Félix Viallet et le domaine universitaire.

Quels enseignements tirer ?

Le premier est que **ce type de projet nécessite de la main d'œuvre d'ingénieur**, à la fois pour la coordination générale et la mise en place des équipements télécoms mais aussi pour l'aide aux laboratoires pour les connecter, pour adapter leurs applications, ... Cette nécessité est accentuée par la limite en temps du projet induisant que tout retard est une perte d'argent et un peu moins d'expérimentation possible. Sur C3I2, la disponibilité d'un ingénieur à plein temps a fortement manqué. La lenteur du démarrage par exemple s'explique par le manque de temps disponible des ingénieurs lorsqu'il fallait résoudre un problème. Il faut donc prévoir un poste budgétaire pour un ingénieur affecté au projet.

La seconde leçon est que **un an c'est court,** qu' **il faut tenir compte du calendrier** et qu'il faudrait un planning précis de réalisation. Mais ceci est difficilement réalisable, les financements ont des difficultés à être planifiés, débloqués à l'heure et lorsque l'on monte un réseau expérimental il y a toujours des alea techniques et on découvre les problèmes en cours de route (autrement le projet ne serait pas expérimental).

Sur C3I2, il aurait certainement fallu **se focaliser sur quelques expérimentations avec la même problématique ou dans le même domaine scientifique** et travailler d'abord sur ce sujet. Le calcul distribué (à distance, parallèle, ….) aurait pu être ce noyau dur. Ainsi le réseau aurait pu créer une dynamique scientifique entre les équipes qui travaillent sur le sujet, aider à faire émerger des projets et à améliorer les compétences de ces équipes dans leur domaine scientifique. Les autres expérimentations n'auraient pas été écartées (l'outil C3I2 était trop coûteux pour être utilisé par une communauté trop restreinte) mais l'effort principal aurait porté sur un sujet.

Le dernier volet est un ensemble de leçons qui pourraient bénéficier à la réflexion sur la réalisation d'un réseau métropolitain de la recherche.
Techniquement le chapitre 6 détaille les capacités de la technologie ATM qui ont été mis en exergue sur C3I2. ATM est à utiliser uniquement en configuration simple, pour partager de la bande passante quand celle-ci est chère, donc sur le domaine public en longue ou moyenne distance. Ainsi **sur un projet métropolitain, ATM peut être utilisé aujourd'hui car disponible, mais ce n'est certainement pas la solution d'avenir, même à moyen terme**. Il serait préférable de regarder vers Ethernet haut débit, vers le multiplexage-commutation sur fibre optique (DWDM) et les nouvelles fonctionnalités de IP (la qualité de service DiffServ, la sécurité, …). Mais aujourd'hui la technologie choisie pèse en définitive peu dans la réussite d'un réseau, ce sont l'adéquation aux besoins, le prix de revient, la disponibilité, les services et l'évolutivité qui sont à étudier et à définir dans une réalisation télécom.
Dans le coût d'un réseau multi-organismes, **il ne faut pas sous estimer les ressources humaines** à la fois pour la partie administration au sens commun du terme (rédaction de documents contractuels, comptabilité, …) et aussi techniques, pour configurer, gérer et dépanner les équipements. Et en chapeau, la tâche de coordination est lourde pour que

l'ensemble converge mais aussi évolue et s'adapte rapidement. Un modèle d'organisation détaillé doit être prévu dès le départ.
D'une manière plus générale, on constate que l'Internet s'est répandu très rapidement. Les besoins et les pratiques en communication des laboratoires et des universités sont maintenant celles des entreprises et même simplement des particuliers. Il peut alors être intéressant d'étudier l'externalisation des services Internet « classiques », c'est à dire d'acheter sa connexion Internet auprès d'un prestataire, sans créer une infrastructure de réseau particulière pour la communauté. En parallèle, les ingénieurs pourront se focaliser sur les besoins spécifiques de la recherche. Même si cet argumentaire est loin de faire l'unanimité, il est certainement utile de l'envisager et d'en débattre.

Avant de conclure, **il faut remercier l'ensemble des participants au projet C3I2, en particulier les membres des différents comités, pour avoir « jouer le jeu » des expérimentations et consacré beaucoup de temps, en plus de leur travail quotidien, à C3I2**.

En conclusion, on peut dire que pour être compétitif, innover, être précurseur sur les usages, et faire partager son expérience aux industriels et au grand public, la communauté de la recherche a besoin de projets de ce type.

2. Les expérimentations prévues et réalisées

Des expérimentations diverses ont été effectuées sur ce réseau. Nous en décrivons un certain nombre. Cette liste n'est pas exhaustive car il n'y a pas eu véritablement de suivi précis de chaque expérimentation, ni de centralisation des résultats. Cette coordination n'a pas pu être faite faute de moyens humains. Il n'est d'ailleurs pas sur que cela aurait apporté un plus aux laboratoires et au projet. Nous ne reviendrons pas sur malheureusement le faible nombre d'expériences, le bilan en première partie de ce rapport en décrit les causes.
Ces expérimentations ont pu être menées car C3I2 a amené les débits suffisants et une qualité de service acceptable. Bien que techniquement réalisables sur le réseau de production ARAMIS alors en place, elle n'auraient pas pu être conduites sur celui-ci car ce réseau n'avait pas ces 2 avantages.

Travail collaboratif : mediaspace

Mediaspace (http://pandora.imag.fr/MediaSpace/index.html) est développé par l'équipe IIHM du laboratoire CLIPS de la fédération IMAG. C'est un outil de travail collaboratif en environnement géographiquement réparti qui vise à créer un espace de travail virtuel, avec un rapprochement visuel au moyen de caméras "de bout de couloir" et de bureaux, qui donnent autant de fenêtres sur l'écran des stations des membres du groupe. Cet outil utilise la fonction multicast IP. Sur C3I2 il a fonctionné entre 2 sites (domaine universitaire et INRIA), connectés par un tunnel IP au-dessus d'ATM. Le débit disponible a permis d'avoir une très bonne définition des images ainsi que la possibilité d'utiliser l'audio dans les communications entre membres du groupe.
Cf Annexe 7.1.

CADNET

CADNET est une plate-forme de CAO partagée en réseau et destinée à accueillir des projets de conception de systèmes et de circuits, avec une mise en commun des logiciels, ainsi qu'un partage de bases de données, de bibliothèques de cellules et de macrofonctions (IPs). La conception de circuits nécessitant un affichage de très bonne définition, le réseau doit transporter un gros volume de données en un temps très court, pour un travail interactif. Les équipes de recherche impliquées sont au CNET et dans les laboratoires TIMA et LEMO (CNRS et INPG). Cette plate-forme a fonctionnée plusieurs mois entre 3 sites de C3I2 (Polygone, av. Felix Viallet, CNET Meylan). C3I2 étant arrêté maintenant, l'équipe de recherche essaie de trouver une autre solution pour disposer de débit suffisant entre les trois pôles pour poursuivre le projet CADNET.

Calcul à distance

Il était (et est toujours) nécessaire de faire du post-traitement à la volée de résultats de simulation d'écoulements sur des stations de l'équipe MOST (http://most.hmg.inpg.fr/) du laboratoire LEGI (CNRS, INPG, UJF) sur le domaine universitaire, le traitement s'effectuant sur le Cray T3E du CEA sur le Polygone, avec utilisation de PVM et des volumes de données de l'ordre de 50 Moctets à transférer en moins de 4 minutes à chaque pas. C3I2 a permis de mettre en œuvre ce mode de travail, avec un succès mitigé (cf Annexes 7.2 et 7.5), impossible sur le réseau de production.

C3I2 étant arrêté maintenant, l'équipe de recherche essaie de trouver une autre solution pour disposer de débit suffisant entre les deux pôles pour poursuivre ce type de travail.

Calcul distribué

Cette expérimentation visait à effectuer entre plusieurs calculateurs répartis sur 2 sites (domaine universitaire et Polygone), de la programmation de machines parallèles pour le calcul haute performance, de la modélisation sur plates-formes de calculs distribués, de la visualisation et ajustement en temps réel de modélisations 3-D et des échanges importants de fichiers de résultats (128 M à 1 Goctet par pas de temps). Les partenaires impliqués était le CEA, l'IMAG, l'Observatoire de Grenoble et l'INRIA. Une partie seulement des calculateurs ont pu être interconnectés : LMC-IMAG et CEA. Cf annexe 7.3.

Expériences de télé-enseignement

Il a été réalisé plusieurs séances de télé-enseignement. Un exemple a été un cours d'apprentissage du logiciel flux-expert, logiciel de simulation de phénomènes physiques, entre un professeur dans le laboratoire Génie Atomique sur le site Polygone et plusieurs élèves de l'INPG sur un autre site avec l'utilisation de PC et de stations de travail HP et SUN multimédia ainsi que les logiciels du MBONE. Cf annexe 7.4.

Téléconférences

De nombreuses télé réunions et téléconférences ont été véhiculées par C3I2 avec différents outils tels que IP/TV (CISCO), Netmeeting (Microsoft), les outils du MBONE (http://www.urec.cnrs.fr/xcast/)… C3I2 a permis d'utiliser ces outils standards sur un réseau avec une bonne qualité de service (bande passante, délai, gigue) absente sur le réseau de production. Ainsi de nombreux laboratoires se sont équipés de stations multimédia avec ces outils pour une utilisation « en vraie grandeur », professionnelle.
Ont transitées sur ce réseau des téléconférences vraiment expérimentales comme la présentation du réseau C3I2 faite en téléconférence lors de la visite en juin 98 de M. Didier Lombard Directeur général à la Direction des Stratégies Industrielles du Ministère de l'Economie, des Finances et de l'Industrie ; une autre est décrite à l'annexe 7.4.
La diffusion des conférences mondiales ou françaises du MBONE s'est aussi faite sur C3I2 (ce réseau véhiculant le trafic de production multicast) mais de manière transparente pour les utilisateurs.

IPv6

Certainement liaisons natives IPv6 ont pu être créées, métropolitaines (entre les sites INRIA et domaine universitaire) et nationales (via SAFIR, avec Jussieu, Rocquencourt, …), connectant des routeurs et des stations IPv6 avec des codes spécifiques. Ceci a permis de tester les couches IPv6 sur un réseau dédié (natif) à moindre coût (sans infrastructure réservée). Des fonctions telles que le routage sur IPv6 ont pu être testées. Techniquement ceci a été réalisé à base de PVC ATM, démontrant que l'on peut construire un réseau « logique » indépendant sur un réseau général ATM. Ces liaisons ont fait partie de l'infrastructure nationale de test IPv6 décrite ici http://phoebe.urec.fr/G6/.

Trafic de production – routage dynamique BGP4

Durant quelques mois une partie du trafic de production entre les sites de C3I2 (et aussi de SAFIR) et ensuite l'ensemble du trafic est passé sur le réseau expérimental C3I2 avec un back up sur le réseau standard de production ARAMIS-RENATER en cas de problème. Hormis le fait que cette bascule démontre que le réseau ATM est stable et fiable, cela a permis de tester le protocole de routage dynamique BGP et le routage IP sur des réseaux maillés. Chaque site a ainsi mis en place un routage dynamique BGP pour utiliser les 2 voies de sortie possibles. Certains problèmes ont été soulevés avec BGP : le choix d'une route par rapport à l'agrégation des chemins, et le temps de convergence : impossible de descendre le temps de convergence en dessous de 2 mn (intervalle des hello BGP) car l'information de rupture de CV ATM ne remonte pas à BGP.
L'annexe 7.6 peut être consultée pour plus d'informations

Autres expérimentations qui sont restées à l'état de projet ou simplement avec un début de mise en place

Réalité virtuelle

Cette plate-forme développée par l'équipe VIS (http://www-vis.imag.fr/) du laboratoire GRAVIR (IMAG) est constituée de plusieurs stations et caméras connectées par un commutateur ATM pour créer un espace de réalité virtuelle avec un système de son spatialisé. Déjà opérationnel en local, dans une seule pièce, l'expérimentation devait viser à éclater cet espace sur 2 sites (domaine universitaire et CNET) avec une connexion directe ATM entre les 2 commutateurs ATM des 2 plates-formes. Faute de moyen cette expérimentation n'a pas pu aboutir à Grenoble. Des tests entre plusieurs commutateurs ont été fait sur le réseau ATM de l'université du Chili à Santiago. La conclusion sur l'utilisation de la technologie ATM pour une vidéoconférence de qualité est très positive, notamment en doublement d'un réseau IP : latence réduite, multicast, contrôle par la source.

Telesun

Le laboratoire LSR de l'IMAG a développé un serveur vidéo d'enseignement accessible en ATM natif. Le but était de permettre l'accès à ce serveur via C3I2. Ce projet n'a pas été poursuivi par ce laboratoire (peut-être parce que les applications développées directement sur le protocole ATM se sont avérées sans avenir).

Bibliothèque virtuelle Calliope

Calliope est une bibliothèque virtuelle (http://www.inrialpes.fr/services/bib-vir.html) répartie entre 3 organismes IMAG, INRIA et XEROX. Il avait été référencé comme projet ayant besoin de connexion spéciale C3I2. En fait Calliope fonctionne classiquement avec les logiciels standards WEB et n'a pas besoin de très haut débit. Il a donc utilisé C3I2, mais comme le reste du trafic de production classique.

Séances de travaux pratiques répartis

Une école d'ingénieurs en électronique et en télécommunication, l'ENSERG, est implantée sur 2 sites, Polygone et av Felix Viallet. Elle désirait utiliser l'infrastructure C3I2 pour monter des séances de travaux pratiques réparties sur ces 2 pôles. Le matériel avait commencé à être mis en place mais l'expérience n'a pas pu aboutir faute de moyen humain.

Utilisation d'un logiciel d'enseignement à distance

Le projet Télécabri (http://www-cabri.imag.fr/TeleCabri/) du laboratoire Leibniz (IMAG), définit un environnement d'enseignement à distance avec assistance de l'enseignant aux élèves, utilisable avec le logiciel d'enseignement de la géométrie bien

connu Cabri-géomètre. Il avait été fortement envisagé de tester cet environnement sur C3I2. Après analyse, il s'est avéré que le besoin était de connecter des élèves et des enseignants isolés, donc avec des faibles débits et les tests nécessaires était plutôt avec des réseaux bas débits comme RNIS.

Routage distribué

Ceci visait à tester une solution de routage distribué avec du matériel IPsilon, tests pouvant servir pour bâtir une nouvelle architecture de réseau distribuée pour les laboratoires IMAG. Le matériel a été installé, a fonctionné et a été utilisé un certain temps. Mais les tests approfondis de performances, fiabilité, … n'ont pas pu être menés à cause de l'indisponibilité forcée et involontaire de l'expérimentateur. Parallèlement, ce type de matériel ne s'est pas avéré une solution retenue par l'industrie comme modèle d'architecture de réseau.

Fiabilisation de sauvegardes

C3I2 devait être utilisé pour interconnecter 2 serveurs de sauvegardes distants pour les laboratoires IMAG avec secours automatique de l'un par l'autre. Un seul serveur a été finalement été acheté rendant caduque ce besoin.

Interconnexion de PABX

Un projet visait à connecter les autocommutateurs téléphoniques d'établissement par une liaison ATM (carte ATM ou carte G703 dans le commutateur). Ce test était trop délicat à mettre en œuvre (interruption du trafic téléphonique, coordination avec l'équipe d'administration du téléphone, …). Faute de moyen humain il n'a pas pu être concrétisé.

3. Genèse, chronologie, organisation

La volonté de mettre en place un réseau haut débit expérimental à Grenoble n'était pas récente. De nombreuses réflexions avaient été menées dans ce sens depuis plusieurs années, en particulier sous l'impulsion de l'IMAG et de l'INRIA. Ces idées ont été reprises dans le cadre de l'association GNI@, avant d'aboutir à C3I2. La genèse a été longue pour 2 raisons : le contour précis des partenaires intéressés et des expérimentations n'a pu émerger qu'après une large concertation; une offre acceptable techniquement et financièrement du seul opérateur qui pouvait offrir un service réseau haut débit, France Télécom, n'a pu être obtenue qu'après de longues négociations.

Initialement le projet s'est appelé **C2I2**, initiales des 4 partenaires de départ : **CEA CNRS**, **IMAG** et **INRIA**. En cours de définition, le **CNET** a manifesté son intérêt pour ce réseau et s'est joint aux premiers organismes, donnant le sigle **C3I2**.

Voici chronologiquement quelques étapes :
. Printemps 96 : définition du projet, préparation du dossier régional mi-lourd 96, premiers contacts avec l'opérateur France Télécom, ceci coordonné par Jean-Pierre Verjus.
. Automne 96 : concertation avec les collègues de Lyon pour harmoniser les 2 projets similaires présentés à la région. Accord de principe de France Télécom pour établir une proposition de service haut-débit entre 4 sites grenoblois.
. Décembre 96 : accord pour une première tranche de financement régional (mi-lourd 96).
. Hiver 96-97 : organisation des comités C2I2. Définition précise du service avec France Télécom. Arrivée du CNET dans le projet.
. Printemps 97 : préparation d'un dossier régional complémentaire pour l'opération mi-lourd 97.
. Eté 97 : consultation des différents fournisseurs d'équipements de matériel réseaux pour l'équipement des sites.
. Automne 97 : choix des équipements de site. Signature des contrats avec France Télécom.
. **Décembre 97 : ouverture du service France Télécom pour le réseau C3I2 pour une durée de 1 an**. Réception et installation des équipements de site. Accord pour la seconde tranche de financement régional (mi-lourd 97).
. Mars 98 : connexion du réseau C3I2 à SAFIR (cf chapitre suivant).
. Avril 98 : ouverture du réseau aux expérimentations avancées. Retard du à différentes causes décrites dans le chapitre 5.
. Juin 98 : bascule de certains trafics de production entre le domaine universitaire et l'INRIA Montbonnot sur C3I2, et entre Paris-Grenoble et Lyon-Grenoble sur SAFIR-C3I2.
. Sept 98 : bascule du trafic de production entre le polygone CNRS et le domaine universitaire sur C3I2.
. Oct 98 : bascule d'une partie du trafic de production entre le site Viallet (INPG, IMAG) et le domaine universitaire sur C3I2. Contacts avec France Télécom pour la poursuite éventuelle de C3I2 (le contrat se terminant le 30/11/98).
. Déc 98 : décision de poursuivre C3I2 jusqu'à juin 99, le financement des liaisons étant pris en charge par les organismes.
. Janv 99 : bascule de tout le trafic de production entre les sites sur C3I2.

. Printemps 99 : publication d'un appel d'offre pour mettre en place un réseau métropolitain de production, pour remplacer le service C3I2.
. **Juillet 99 : arrêt de C3I2**, choix de l'offre SMHD de France Télécom comme solution au réseau métropolitain de production avec un contrat de 1 an renouvelable 6 mois.
. Septembre 99 (prévu) : mise en service de la boucle SMHD métropolitaine.

Cette liste montre **qu'un tel projet nécessite du temps, 2 ans ½**, pour passer de la définition jusqu'au transport du trafic de production, preuve que le réseau offre un service stable, et que les étapes sont nombreuses mais incontournables.

Aucun moyen humain n'a été affecté spécifiquement à ce projet. Chaque partenaire a désigné un ou plusieurs représentants qui ont travaillé à la mise en place du réseau et des expérimentations en plus de leur charge de travail habituelle. Le projet a nécessité une coordination, à la fois décisionnelle et technique. 3 comités ont été formés :
. Un **comité de pilotage**, organe décisionnel, constitué de :

. Jean Luc Archimbaud Directeur Technique de l'UREC (Unité Réseaux du CNRS) coordonnateur du projet
. Daniel Bois Directeur du CNET Meylan
. Pierre Laforgue Directeur des moyens informatiques de l'IMAG, Directeur technique d'ARAMIS
. Jean Potier Responsable Informatique du CEA Grenoble
. Jean-Pierre Verjus Directeur de l'INRIA Rhône-Alpes
. Jacques Voiron Directeur de l'IMAG

. Un **comité technique-partenaires** qui regroupaient les chefs de projet de chaque partenaire, chargé de prendre les décisions techniques, composé de :

. Jean Luc Archimbaud (CNRS/UREC) coordinateur
. Raoul Dorge (CS/Athesa pour le CEA)
. Joseph Lecourt (CNET)
. Jean-Luc Parouty (IMAG)
. Luc Saccavini (INRIA)

. Un **comité technique-points d'accès**, réunissant les ingénieurs chargés de l'administration des équipements télécom sur les différents sites :

. Site INRIA Montbonnot : Luc Saccavini (INRIA) et Jean-Pierre Auge (INRIA)
. Site Viallet : Claire Rubat du Merac (INPG) et Jean-Luc Parouty (IMAG)
. Site domaine universitaire : Christian Lenne (CICG), Jean-Luc Parouty (IMAG) et Jacques Eudes (UJF)
. Site polygone : Raoul Dorge (CS/Athesa pour le CEA) et Daniel Guéniche (CNRS)
. Site CNET Meylan : Joseph Lecourt (CNET) et Salvador Salas (CNET)

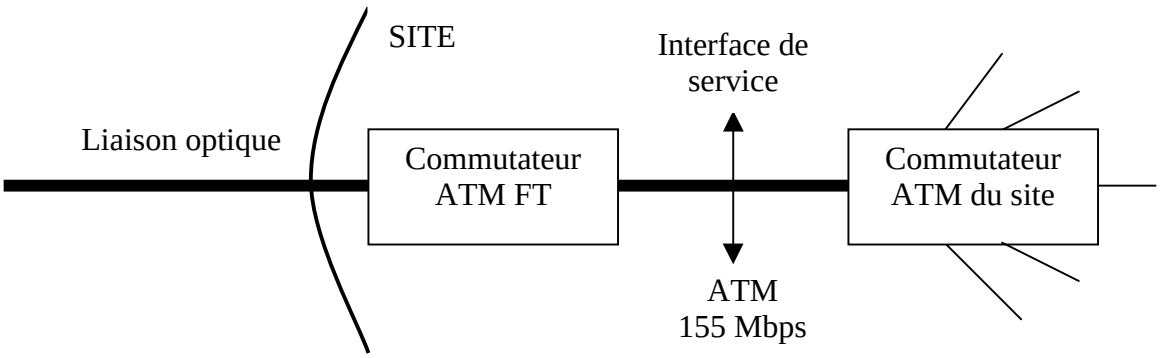
Les 2 derniers comités, très actifs, se réunissaient généralement ensemble, en moyenne toutes les deux ou trois semaines.

4. Architecture du réseau

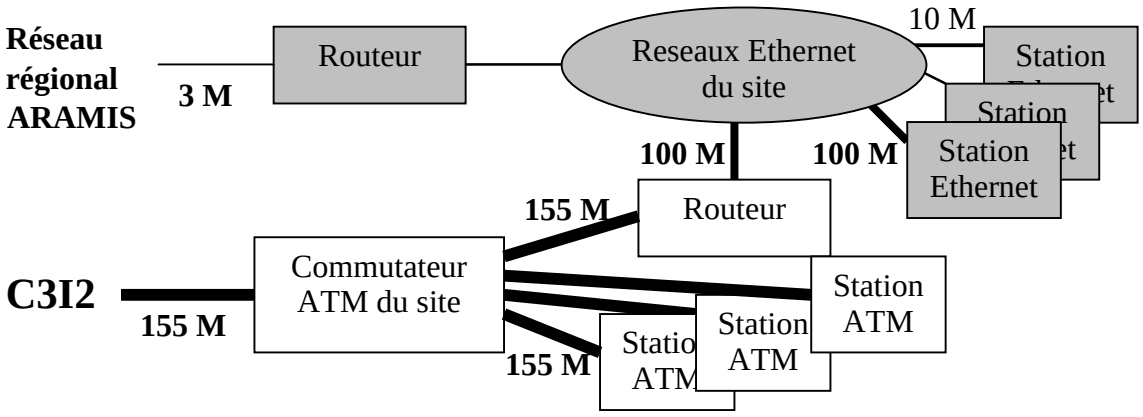
Les 5 partenaires du projet sont situés sur **5 pôles différents de Grenoble** : le domaine universitaire (avec des laboratoires de l'IMAG, du CNRS, des universités, …), les bâtiments de l'INPG avenue Félix Viallet (IMAG, CNRS, …), le polygone scientifique (CEA, CNRS, …), Montbonnot (INRIA) et Meylan (CNET).
Il faut d'abord souligner que 4 des 5 partenaires sont interconnectés, pour leur trafic de production, au travers du réseau régional ARAMIS, lui-même connecté à RENATER (donc à l'Internet) avec des prises en moyenne à 3 Mbps. Durant les expérimentations C3I2, ARAMIS a permis d'accéder à distance aux équipements C3I2 même lorsque l'accès était impossible par le réseau C3I2 et a permis aussi de tester les protocoles de routages dynamiques de IP (cf expérimentations).
Les architectures décrites ci-dessous nécessitent, pour leur compréhension, une petite connaissance d'ATM. Elles permettent de montrer les nombreuses possibilités d'architectures que l'on peut construire sur une infrastructure ATM.

Architecture physique

Le réseau C3I2 a été bâti sur l'offre de **VP (Virtual Path) ATM en mode CBR (débit constant) de l'opérateur France Télécom** pour interconnecter les 5 pôles. Actuellement ce service porte le nom d'**OMA** (Offre Multiservices sur ATM). Chaque site était desservi par une liaison optique avec un commutateur ATM, équipement d'extrémité fourni et administré par l'opérateur, auquel le site n'a pas accès. De son commutateur l'opérateur offre au client une interface de service, ATM à 155 Mbps sur media optique.



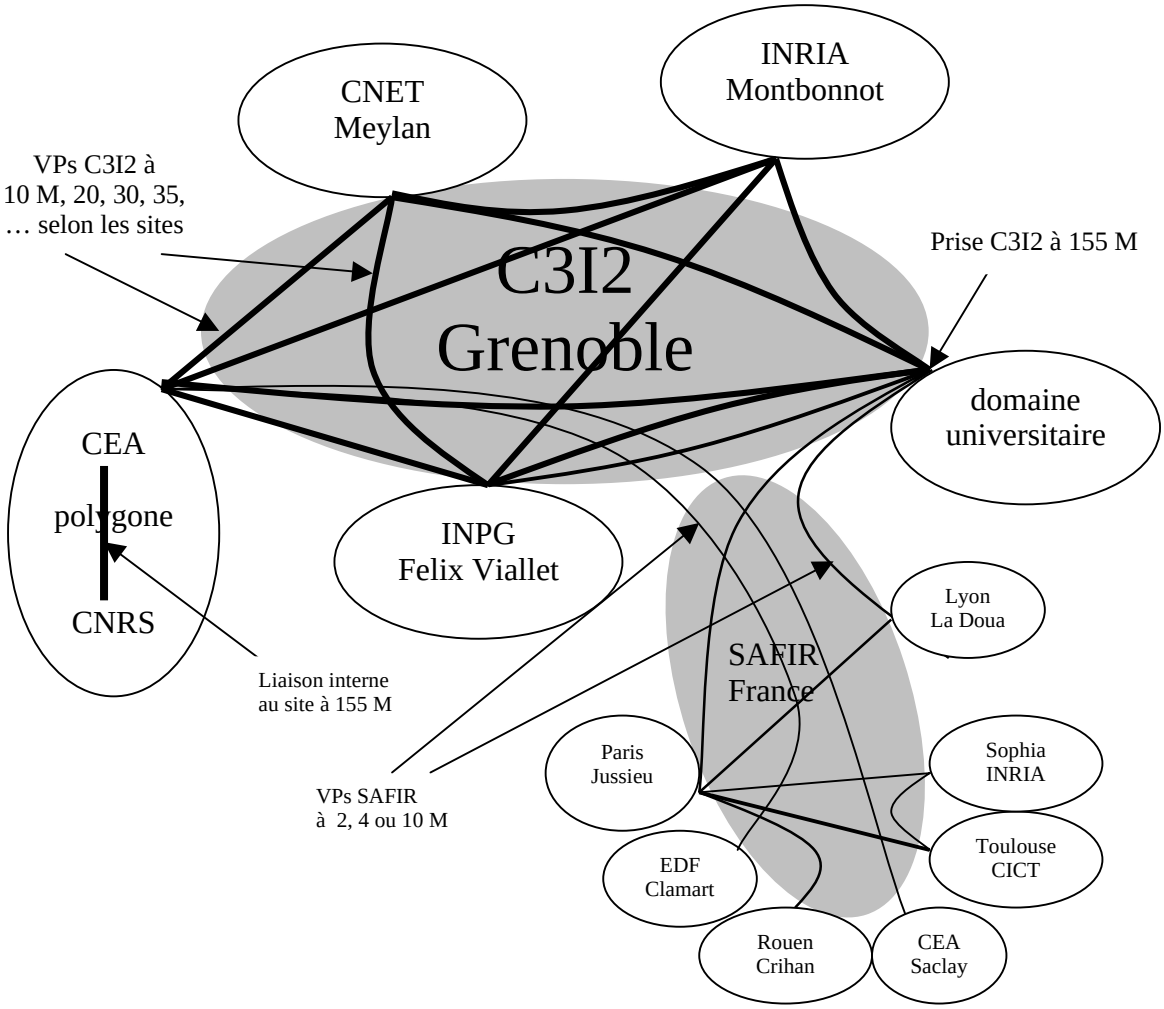
Sur cette interface de service le site connecte son propre commutateur ATM avec des stations ATM et un **routeur ATM-Ethernet permettant d'accéder au réseau de production du site** (en grisé sur le schéma suivant). Tous les sites ont en interne un réseau de production de type campus, Ethernet commuté à 100 M et parfois FDDI.



Chaque site dispose ainsi d'une bande passante de 155 Mbps vers le réseau C3I2. Celle-ci est découpée en VPs, l'équivalent de liaisons spécialisées à débit fixe et garanti, vers les autres sites.

L'architecture de réseau de VPs que nous avions construite était maillée. Chaque site était connecté aux quatre autres par au moins un VP. Les changements de configuration des VPs (nouveau VP, modifications de débit des VPs existants, …) pouvaient être demandés par télécopie à l'opérateur et étaint pris en compte dans les 24 heures, ce temps s'étant avéré largement suffisant comme délai de réaction.

Cette infrastructure était connectée au réseau national expérimental ATM SAFIR (http://www.renater.fr/Safir/safir.html), préfiguration de RENATER 2 qui raccordait Grenoble, Lyon, Paris, Toulouse, Rouen et Sophia avec un système de VPs similaire à C3I2. Concrètement, la liaison optique C3I2 qui desservait le domaine universitaire connectait aussi celui-ci à l'infrastructure SAFIR, et à travers lui tous les sites de C3I2. Du domaine universitaire partaient 2 VPs à 10 M, l'un vers Paris, l'autre vers Lyon. Ces 2x10 M étaient déduits des 155 de la prise C3I2. Toujours dans le cadre de SAFIR, le CEA Polygone avait 2 VPs à 2 et 4 M vers l'EDF Clamart et le CEA Saclay.
Sur les sites avaient été installés, entre autres, 4 commutateurs ATM LS1010 de CISCO, 2 commutateurs ASX de FORE, 4 routeurs CISCO avec une carte ATM, des commutateurs ATM-Ethernet de MADGE, ainsi que différentes stations souvent multimédia ATM ou Ethernet : PC, SUN, DEC, SGI, ...



L'architecture logique ATM

La majorité des applications qui utilisaient C3I2 sont écrites sur une pile logicielle TCP/IP. Beaucoup, qui initialement désiraient de l'ATM natif ont, pendant la mise en place du réseau, été portées sur TCP/IP. Néanmoins une connectivité uniquement IP n'était pas envisageable pour supporter certaines applications qui étaient prévues (routage distribué, interconnexion de PABX, établissement de chemins permanents pour IPv6, multicast national) mais aussi pour pouvoir facilement isoler différents flux.
Ainsi le but a été de **pouvoir établir une connexion ATM directe entre toutes les stations (ou routeurs ATM) du réseau C3I2, ainsi que du réseau SAFIR**, sans passer par un routeur IP, élément qui est un goulot d'étranglement pour les performances.
Sur l'infrastructure de VPs décrite précédemment et fournie par l'opérateur, nous avons construit notre propre réseau privé ATM par transport de notre signalisation à l'intérieur des VPs de l'opérateur ("**tunneling**"), seule méthode pour disposer d'un réseau ATM « dynamique ». Ainsi, lorsque l'on établissait une connexion ATM (concrétement un VC, Virtual Channel) entre 2 stations C3I2, les équipements de France Télécom étaient totalement transparents et ce sont les commutateurs ATM des sites qui géraient les différents protocoles de signalisation ATM.
Nous avons suivi le **plan d'adressage ATM national de SAFIR** (aussi appelé plan d'adressage RENATER 2) où un intervalle d'adresses est réservé à Grenoble. Les 10 octets 39.25.0f.00.00.00.2d.00.79.01 constituent le préfixe attribué à C3I2 par SAFIR Nous avons ensuite attribué à chaque site de C3I2 une plage d'adresses dans cet intervalle. Ainsi le domaine universitaire a eu le préfixe sur 11 octets 39.25.0f.00.00.00.2d.00.79.01.01, l'INRIA 39.25.0f.00.00.00.2d.00.79.01.02, …
Chaque site a pu ensuite utiliser les octets restants pour numéroter ses stations (une adresse ATM fait 20 octets où uniquement les 13 premiers sont configurables). Ainsi l'adressage était unique sur C3I2 mais aussi sur SAFIR et toute station ATM de C3I2 pouvait atteindre une autre station de C3I2 mais aussi de SAFIR directement en ATM (sans passer par un routeur).
Même si on considère qu'il est obligatoire de segmenter un réseau national de production par des routeurs IP entre les régions, pour avoir une architecture avec plusieurs réseaux IP, la possibilité d'avoir un chemin direct ATM entre toutes les stations en France sur ce réseau expérimental laissait une très grande liberté d'architecture. La connexion directe ATM est toujours plus efficace en performance que le passage par un routeur. Suivant les applications nous utilisions soit le chemin direct ATM, soit le chemin avec passage par un routeur IP.

Routage ATM

Sur C3I2 nous utilisions le protocole de routage dynamique **PNNI** (Private Network to Network Interface) entre les 4 commutateurs CISCO et **IISP** (Interim Inter Switch Protocol) avec un routage statique donc avec les commutateurs FORE (cf le chapitre problèmes de mise en place) et avec les équipements hors Grenoble de SAFIR. PNNI, qui évite une configuration manuelle et fastidieuse des « routes » dans chaque équipement semblait très bien fonctionner avec des équipements homogènes (CISCO chez nous) avec un nombre restreint de nœuds, sans hiérarchie (nous n'avons pas testé le multi-niveaux inutile dans notre configuration).
PNNI permet de tirer tous les bénéfices de l'architecture maillée. Si le VP direct entre 2 sites est « coupé » ou saturé, le trafic peut faire un crochet par un troisième site intermédiaire.

Architecture LANE (LAN Emulation)

Sur l'infrastructure C3I2 et SAFIR, nous avons construit l'équivalent de **2 réseaux Ethernet (LANE)**, l'un métropolitain C3I2 et l'autre national. Cette architecture permet de raccrocher des stations derrière des commutateurs ATM-Ethernet (edge device) et aussi de disposer de la fonction basique de diffusion d'Ethernet qui peut être utile pour les applications multicast.
Le commutateur CISCO sur le domaine universitaire était serveur LECS et LES pour le LANE C3I2. Les serveurs pour le LANE SAFIR étaient à Jussieu.

Architecture CLIP (Classical IP)

De la même manière que LANE nous avons bâti sur C3I2 une architecture logique CLIP, avec un sous-réseau IP pour les équipements de communications (routeurs, …) et un sous-réseau IP pour les stations. L'idée était de séparer logiquement les équipements d'interconnexion des stations utilisateurs. Le routeur CISCO du domaine universitaire eétait serveur ARP pour le CLIP C3I2. Il existait aussi un réseau CLIP SAFIR ou le serveur ARM était installé à Jussieu. Nous pouvions aussi enregistrer une station C3I2 dans le réseau CLIP de SAFIR.

Adressage IP

Sur C3I2, nous utilisions un réseau de classe C officiel, segmenté en 4 sous-réseaux : CLIP et LANE pour les équipements réseaux (commutateurs, routeurs) et séparément pour les stations. Nous avions aussi une plage d'adresse réservée dans les 2 réseaux de classe C du CLIP et du LANE nationaux de SAFIR.

Routage IP

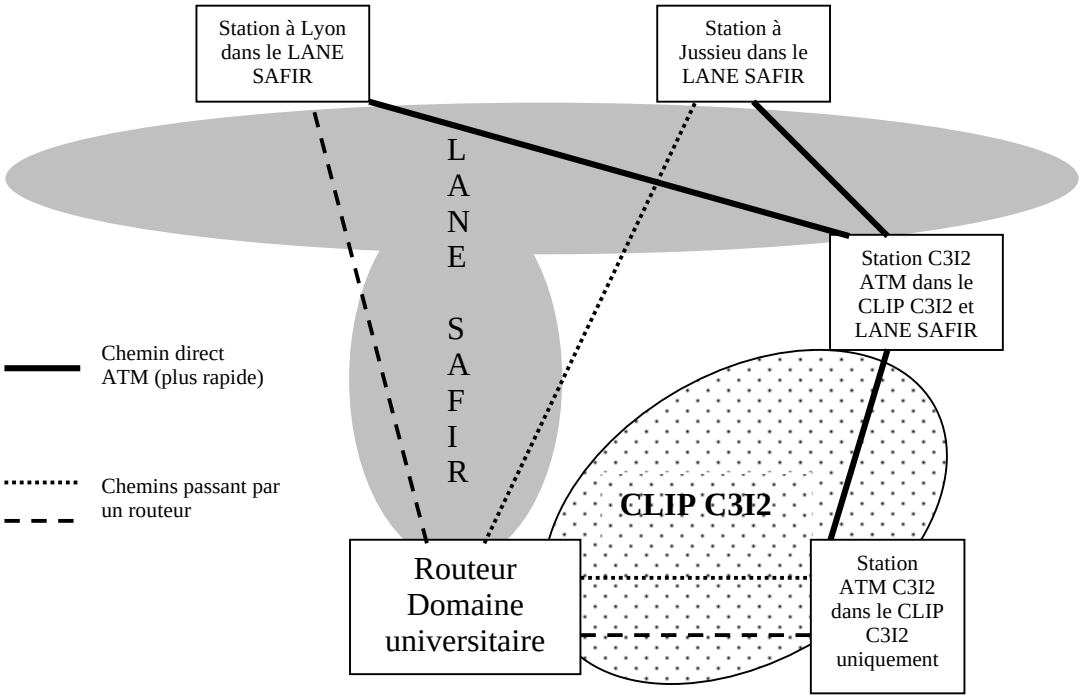
Le routeur ATM du domaine universitaire assurait le routage entre les sous-réseaux IP de C3I2 (LANE et CLIP) et avec les réseaux SAFIR. Le réseau IP de C3I2 était annoncé sur l'Internet par le routeur du domaine universitaire connecté sur ARAMIS. Au cœur de C3I2 nous utilisions le routage IP statique avec une route par défaut, mais pour transporter du trafic de production nous utilisons BGP (cf chapitre expérimentations).

Connexion des stations

Initialement, comme le voudrait une bonne méthodologie expérimentale, nous désirions séparer physiquement les réseaux de production et ce réseau expérimental. Mais les sites n'ayant pas en interne un réseau ATM, pour connecter des stations d'expérimentation en ATM il fallait une fibre optique entre la station et le commutateur d'entrée, denrée rare sur les sites. Donc peu de stations étaient directement connectées sur ce réseau expérimental. Beaucoup étaient sur le réseau Ethernet de production et nous utilisions le routage statique IP par station dans les routeurs sur le chemin pour orienter le trafic vers C3I2 ou vers ARAMIS. Avec ce type de raccordement, nous vérifions le débit disponible sur l'ensemble du chemin en particulier sur les réseaux de production, pour garantir un minimum de bande passante.

Avec ces différentes architectures, nous avions toutes les variantes possibles de connectivité avec ATM, en ajustant simplement la configuration des stations. Ainsi une station ATM (ou un routeur C3I2) pouvait appartenir au réseau CLIP de C3I2, au réseau CLIP de SAFIR, au LANE de C3I2 ou au LANE de SAFIR, à plusieurs de ces réseaux, voire à tous. Le fait d'appartenir par exemple au LANE SAFIR permettait de ne pas avoir à traverser un routeur pour accéder aux stations de SAFIR.

Exemple de connectivité :



Administration

Chaque ingénieur de site administrait son routeur, son commutateur et ses stations. Les réunions régulières des comités techniques permettaient de se coordonner et plusieurs listes de diffusion électroniques étaient très actives. Un outil très simple (à base de ping avec petits et gros paquets et traceroute) vérifiait régulièrement, toutes les 5 minutes, l'accessibilité IP des équipements télécom et de certaines stations. En cas d'anomalie (accessibilité ou routage), un message électronique était envoyé (à travers ARAMIS évidemment) aux administrateurs de sites concernés. L'outil MRTG avec un accès Web installé sur le site INRIA permait d'avoir des relevés de trafic et de charge de tous les commutateurs et routeurs de C3I2.

5. Les problèmes de mise en place

Il y a eu de très nombreux problèmes de mise en place, phénomène normal et prévisible vues les nouvelles technologies utilisées et les versions expérimentales des logiciels fournis par les constructeurs sur les équipements de communication. Ces erreurs ont malheureusement beaucoup retardées l'ouverture du réseau aux expérimentations avancées.
Ce chapitre fait un court résumé de chaque problème et essaie d'en tirer une leçon pour la mise en place d'un réseau, au risque de formuler des lapalissades.

Où 155 Mbps théoriques deviennent 134 Mbps utiles

Notre contrat auprès de France Télécom stipulait un accès à 155 Mbps. Dans un premier temps, nous avions réfléchi à une architecture avec des VPs à hauteur de 155 Mbps par site. Mais lorsque nous avons demandé à l'opérateur de la mettre en place, nous avons eu les précisions sur ce qu'indique 155 M., C'est le débit physique de l'accès. ATM n'est pas le protocole transporté directement sur la fibre optique, c'est le protocole SDH (Synchronous Data Hierarchy) qui transporte les cellules ATM. Ainsi aux 155 Mbps, il faut ôter la bande passante utilisée par les entêtes SDH, ce qui ramène la bande passante à 149 Mbps. D'autre part, une règle d'ingénierie retenue dans le réseau ATM de France Télécom, impose de réserver une marge de manœuvre de 10 % de cette bande passante pour l'exploitation et la bonne marche des commutateurs de l'opérateur. Il est donc resté 134 M bps par site pour définir les différents VPs.
Sur tous les types de réseau, il est obligatoire de connaître le débit utile offert par le cœur du réseau et aux points d'accès, et pas uniquement le débit théorique qui est mis en avant commercialement.

Shaping (lissage)

Lorsque l'on crée plusieurs VPs (entités logiques) avec un débit fixé, maximal, sur une interface physique à 155 Mbps, comme c'était le cas des accès C3I2, il est nécessaire que ces maxima de débit soient respectés pour chaque VP individuellement. Il ne faut pas émettre plus de 30 Mbps dans un VP à 30 Mbps, même si le coupleur permet d'écouler 155 Mbps. Pour faire respecter ce contrat, comme tout opérateur, France Télécom installe une fonction de contrôle (policing) sur ses équipements qui élimine les cellules ATM « au dessus » de la bande passante de chaque VP. Sur le site, il faut que le commutateur implémente une fonction de lissage (shaping) du trafic qui permette d'absorber les rafales venant des autres interfaces à 155 Mbps dans des buffers pour les émettre avec un certain espacement de manière à arriver à un débit inférieur au débit du VP. Dans notre cas, cette fonction est appelée shaping multi-VPs sur un même port physique. Cette fonctionnalité avait été un critère de choix pour nos commutateurs.
Malheureusement, malgré les promesses du constructeur et la documentation, les premières versions de logiciel livrées assuraient incorrectement cette fonction de shaping. Nous avons mis du temps a nous en rendre compte et à attribuer des mauvais fonctionnements à cette déficience. Ainsi, les datagrammes pings standards (de 64 octets) étaient transportés correctement sur le réseau mais pas les datagrammes de 1000 octets. La raison était que la fonction de contrôle de notre opérateur acceptait des rafales de trois cellules (une cellule contient 48 octets de données utiles) au dessus du débit du VP, mais pas plus. Les « pings standards » étaient donc acceptés par le réseau mais pas

les pings de plus de 48 x 3 octets. Un mauvais shaping se manifeste principalement par le fait que des petits paquets IP sont transférés sans problème alors que les gros ne passent pas.
Avec l'aide de France Télécom qui, en temps réel, pouvait nous indiquer si des cellules ATM sont rejetées par ses équipements à cause d'un mauvais shaping de notre part, au bout d'un certain temps, nous avons pu mettre au point une méthode pour vérifier que cette fonctionnalité de lissage était correctement assurée par nos équipements.
Il est très difficile de diagnostiquer un mauvais shaping dans un équipement ATM et plus généralement de reconnaître une mauvaise fonction de base (physique) dans un élément d'un réseau.

Utilisation du VP 0 et standardisation

Pour les expérimentations, nous avions installé des équipements de routage distribué Ipsilon sur 2 sites distants avec un commutateur ATM Ipsilon réservé à ces équipements, sur un des sites. Ces routeurs Ipsilon nécessitent un « tunneling » de VP entre les commutateurs CISCO d'entrée de site. Ils utilisent le VP 0 de manière tout à fait propriétaire, sans rapport avec les standards habituels de l'ATM Forum qui réservent ce VP pour la signalisation UNI, ... Il a fallu déclaré sur les commutateurs CISCO un paramétrage non documenté, que uniquement les développeurs de CISCO ont été capables de nous fournir. Lorsque l'on a essayé de savoir qui de CISCO ou de Ipsilon ne respectait pas les standards, on a eu un débat de spécialistes sans arriver à une conclusion claire pour le commun des mortels. En simplifiant, Ipsilon argumentait que rien dans les standards ATM Forum n'obligeait à réserver le VP 0 pour la signalisation, ce qui semble exact, et CISCO que tout le monde utilise le VP 0 pour cette fonction, ce qui est aussi exact (sauf le constructeur Ipsilon). CISCO a montré de la bonne volonté pour résoudre ce problème car notre cahier des charges spécifiait que les commutateurs ATM devaient permettre l'interconnexion de routeurs Isilon.
La leçon semble double : **il est préférable d'implémenter les standards comme « le plus grand nombre » et il est très difficile de prouver qu'un constructeur ne respecte pas les standards.**

PNNI

PNNI (Private Network to Network Interface) est un protocole de routage dynamique sur ATM. Il fonctionne parfaitement entre les commutateurs CISCO, par contre entre CISCO et FORE il n'est actuellement pas interopérable. Le diagnostique de ce problème a été long car les tables de routage ATM sur les commutateurs étaient correctement mises à jour, mais certaines applications entre les sites équipés de commutateurs FORE et les autres de CISCO fonctionnaient à certains moments et pas dans d'autres, de « manière un peu aléatoire », apparemment sans logique. Le problème venait d'un refus d'ouverture de circuit virtuel, qui demandait une signalisation PNNI, dans le sens CISCO vers FORE, alors que dans l'autre sens elle était acceptée. Ainsi lorsqu'un CV bidirectionnel était déjà ouvert dans le sens FORE vers CISCO, le trafic qui pouvait utiliser ce CV était écoulé sans problème. S'il y avait besoin d'une ouverture de CV dans le sens CISCO vers FORE (les CV se ferment au bout d'un certain temps d'inactivité) alors la connexion ne s'établissait pas et les données n'étaient pas transmises.
Le problème était du à un bug CISCO qui positionnait mal un « information element » dans le CALL SETUP d'ouverture de CV lors de l'utilisation de VP différent de zéro pour la signalisation PNNI, ce qui était notre cas (avec le tunneling de VP), cas rare

dans le monde. Le problème a été contourné par du routage statique (IISP) entre CISCO et FORE.
La leçon est qu'une très bonne connaissance des standards et un analyseur de protocole sont obligatoires pour résoudre ce type de problèmes.

Performances décevantes avec le super calculateur Cray T3E

Plusieurs expérimentations nécessitaient l'accès à un super calculateur Cray T3E équipé d'une carte ATM. Dans notre cas, une station à distance d'abord connectée sur un réseau Ethernet 10 M avait de très bons résultats lors des transferts avec le T3E via C3I2. Quand nous l'avons connectée sur un Ethernet à 100 M, pensant augmenter les débits de transfert obtenus, les résultats sont devenus catastrophiques, bien inférieurs aux précédents. D'autre part, les tests de performance effectués localement, sans transiter par le réseau C3I2, montraient des résultats convenables. Longtemps, nous avons pensé à un problème de shaping dans le réseau, ce qui n'était pas le cas. Une fois encore, il a fallu avoir recours à un outil d'analyse ATM pour diagnostiquer ce dysfonctionnement. Les communications expérimentales vers le Cray T3E traversant plusieurs routeurs, la taille des paquets de données IP est beaucoup plus petite que lors d'une connexion directe, pour se prémunir de toute fragmentation intermédiaire (MTU de 512 octets contre 9180). La solution pour contourner ce problème aurait été d'établir une connexion directe avec le supercalculateur Cray T3E, mais cette configuration allait à l'encontre de la politique de sécurité du CEA en matière de communications informatiques. L'annexe 7.5 détaille la mise en évidence des problèmes de la carte ATM du Cray.
Une démarche logique pour éliminer un élément défaillant ne peut se substituer à des outils d'analyse.

Mauvais fonctionnement des couches TCP/IP sur les OS

Nous avons aussi mis en cause le réseau ATM de manière injustifiée dans un autre cas. Des tests de ttcp (outil de mesure de performance TCP) entre 2 stations de 2 sites distants, donnaient des résultats corrects lorsqu'on effectuait un transfert. Mais lorsqu'on lançait deux transferts en parallèle la somme des débits obtenus étaient loin des performances précédentes dans un sens, alors que dans l'autre les débits obtenus étaient normaux. En fait, des deux cotés, les stations n'étaient pas identiques en terme d'Operating System. Dans un des OS les couches TCP/IP savaient mal gérer 2 sessions TCP en parallèles. Moralité **: il faut avoir des stations identiques en terme de matériel et de logiciel pour faire des tests de performance et valider une infrastructure de réseau**.

Interprétation de mauvaises performances TCP

Au fil des tests de performances de TCP sur C3I2, il est apparu clairement qu'il faut être très prudent dans l'interprétation des résultats et hormis la qualité des logiciels sur les stations et les performances des stations il faut aussi tenir compte des équipements traversés qui peuvent imposer des tailles de segments plus ou moins grands ou fragmenter les segments, et du paramétrage de TCP sur les stations comme la taille des fenêtres d'émission et de réception (qui sur certains OS demande de refaire un noyau). Moralité : **il faut tenir compte de l'ensemble de la chaîne dans des mesures de performance.**

6. ATM était il un bon choix ?

Le but initial de C3I2 était d'avoir une infrastructure métropolitaine haut-débit sur lequel nous aurions pu tester différentes technologies de réseau. Pour ce faire nous aurions désiré disposer de fibres optiques réservées, entre chaque site de Grenoble. Le seul opérateur, France Télécom n'a pas voulu nous fournir ces liens et nous a proposé son offre ATM. Nous avons négocié pour avoir le minimum de service assuré par l'opérateur (pas de routeur opérateur par exemple) pour pouvoir avoir le maximum de liberté de tests dans les protocoles réseau. C'est ainsi que nous avons abouti sur **l'offre VP ATM de France Télécom qui était la seule possibilité pour avoir de tels débits à Grenoble, donc ATM était incontournable**. C'est un service ATM de base, un VP étant simplement l'équivalent d'une liaison spécialisée point à point. Cela nous a laissé la possibilité de construire autant de réseaux logiques que nécessaires. Ce service minimal s'est avéré un bon choix pour valider des solutions ATM et mener des expérimentations.

Sur C3I2, **le service ATM a été très peu utilisé de bout en bout**, c'est à dire de station à station. En effet, les cartes ATM étaient et restent très chères par rapport à l'équivalent en Ethernet 100 M. De plus, les sites ne sont pas équipés de réseau ATM interne donc la connexion d'une station nécessitait une fibre optique dédiée entre le commutateur ATM d'entrée et la station souvent dans un autre bâtiment. Beaucoup d'expérimentations qui au départ désiraient de l'ATM natif (car ATM était le synonyme de haut débit) se sont révélées avoir uniquement besoin de bande passante et fonctionner parfaitement sur IP, via Ethernet puis ATM, avec la bande passante désirée. Ainsi, nous n'avons pas pu utiliser les fonctionalités d'ATM qui ont trait à la qualité de service (hormis la bande passante garantie). Cela nécessite des applications en ATM natif, mais aussi des équipements ATM qui peuvent implémenter les mécanismes nécessaires pour supporter ces fonctionnalités sur des VC, ce qui ne semblait pas être disponible.

Par contre **certaines fonctions d'ATM sont très pratiques et ont permis d'avoir une infrastructure souple, facilement segmentable, pour véhiculer des flux divers qui ne se perturbent pas**; problématique des opérateurs de réseaux longues distance partagés entre plusieurs utilisateurs indépendants. Ainsi ces fonctions basiques d'ATM qui a été conçu initialement pour ces opérateurs de WAN, ont été très utilisées sur ce réseau métropolitain. Ainsi :

. La possibilité d'avoir plusieurs VPs entre chaque site permettait d'attribuer facilement des bandes passantes réservées de différents débits entre les sites, mais aussi pour certaines applications ou certains sous-réseaux IP, c'est à dire certains groupes d'utilisateurs. Il suffisait dans ce cas de configurer correctement le routage ATM avec les numéros de VP associés. Cela permettait aussi de séparer le trafic expérimental et de production sans que l'un perturbe l'autre.

. La facilité sur une même infrastructure ATM de créer autant de réseaux logiques IP (Classical IP) ou Ethernet (LANE) que l'on veut permet de créer des réseaux virtuels entre communautés ou applications très facilement. Ainsi une station peut basculer d'un réseau logique à l'autre simplement avec une ou deux commandes, voire appartenir à plusieurs réseaux logiques. Dans la même idée, l'utilisation de PVC a permis de créer un réseau natif IPv6 entre certains équipements, c'est à dire un réseau avec un autre protocole que IPv4 (version IP actuelle). Cela peut-être nécessaire en particulier pour la transition IPv4 vers IPv6.

. Le protocole de routage ATM PNNI sur ce réseau maillé permet d'avoir une utilisation optimale, performante et robuste des liens, en adaptant le routage à la charge et à la disponibilité des liaisons de manière tout à fait transparente.
. L'utilisation de LANE avec une couverture régionale ou nationale, crée l'équivalent d'un réseau Ethernet, utilisable par les applications multicast. On dispose ainsi « en natif » d'une infrastructure de diffusion sans besoin de tunnels souvent difficiles à mettre en place comme sur les routeurs « classiques » Internet. Avec cette architecture de niveau 2, il est facile de monter des vidéoconférences pour une utilisation expérimentale ou pour un petit nombre de stations. C'est aussi une alternative aux tunnels IP pour créer l'équivalent d'un MBONE régional pour interconnecter les routeurs des sites et transporter des flux multicast entre ces sites.
. ATM peut permettre d'interconnecter des PABX très facilement (sous réserve des problèmes de synchronisation sur AAL5), sans que le trafic de données ne perturbe celui de la voix.

Ces utilisations très diverses du réseau auraient été très difficile à réaliser avec un réseau constitué de routeurs IP et de liaisons spécialisées classiques.

Par contre, **ATM est très loin d'être parfait** :
. Le coût des équipements (5 KF minimum pour une carte de station lors des expérimentations) a fait que l'utilisation d'ATM s'est limitée au coeur du réseau et ne s'est pas étendue à l'intérieur des sites, ce qui est un très bon choix économique.
. La technologie ATM par sa complexité nous a posé de grosses difficultés pour résoudre les problèmes de mise en place du réseau et était problématique en exploitation lorsque quelque chose « ne marchait pas ». Il était alors très difficile d'identifier où se situait le problème : lien physique, VP, routage IP, routage ATM, signalisation, shaping, serveur ARP, serveur LES-LECS, ... ? Il est nécessaire d'avoir une bonne méthodologie et un réseau d'experts.
. Une lacune flagrante d'un réseau élargi ATM est l'unicité des serveurs ARP d'un côté pour Classical IP et LECS, LES, … de l'autre pour LANE. Cela implique que le réseau dépend de la bonne marche et de l'accessibilité d'un seul équipement. On est ainsi conduit à réduire la taille des réseaux IP et à créer plusieurs réseaux IP, forçant le passage à travers un ou plusieurs routeurs pour communiquer, ce qui réduit les performances par rapport à un chemin direct ATM. Ainsi, il aurait été plus performant d'avoir un seul réseau IP national pour avoir toujours des connexions directes ATM, sans passage par un routeur. Cela n'est pas raisonnable. On aurait fait dépendre le trafic interne C3I2 par exemple d'un seul serveur, certainement externe. La solution a ce problème peut être MPOA ou MPLS mais cela n'était pas encore disponible avec des équipements hétérogènes, comme c'était notre cas sur C3I2.
. La solution mise en place avec un maillage complet des n sites a l'inconvénient de ne pas supporter une mise à l'échelle puisqu'en moyenne le débit effectif vers chaque site est de 134/(n-1) Mbps.

En conclusion on pourrait dire que la technologie ATM est une bonne solution de réseau métropolitain ou national quand la bande passante n'est pas illimitée, quand on veut créer des réseaux indépendants (pour garantir des bandes passantes, créer des réseaux virtuels, mixer plusieurs protocoles, le téléphone et les données, …), mais que sa complexité fait qu'elle doit être utilisée de la manière la plus basique possible et qu'il est inutile d'offrir un service ATM aux stations des utilisateurs.

7. Annexes

7.1 Mediaspace

Joelle Coutaz laboratoire CLIPS-IMAG

FICHE SYNTHETIQUE MEDIASPACE
CoMedi : Outil de communication informelle entre personnes géographiquement réparties.
Partenaires de départ : UJF, France Télécom-CNET
Autres partenaires potentiels : INRIA, tous les laboratoires de l'IMAG
Laboratoire-projet : laboratoires CLIPS (équipe IIHM) et GRAVIR (équipe PRIMA)
Participants : Joelle Coutaz (responsable), Jim Crowley
Sites : domaine universitaire (bat C et GETA) et INRIA Montbonnot
Protocoles : IP/ATM

MOTIVATIONS :
La proximité physique favorise les coopérations. Or, le travail à distance prenant une place grandissante, la communication en face à face doit être supportée par des techniques multi-média. Les systèmes de vidéoconférence et les mediaspaces sont deux exemples de support à la communication interpersonnelle médiatisée. Ces outils visent l'un et l'autre à abolir les distances entre les interlocuteurs. Alors que la vidéoconférence s'appuie sur un protocole social conventionnel, le mediaspace, par sa disponibilité permanente et son éventail de connexions ("jeter un coup d'oeil", "ouvrir une vue sur la cafétéria de l'entreprise", "envoyer un message", etc.), vise à créer un espace virtuel partagé qui se superpose à l'espace physique. Ainsi parle-t-on de bureau virtuel, lieu de communication informelle, renfort de la coopération au sein d'un groupe par une conscience partagée de l'activité commune. C'est à la communication informelle, au support à la conscience de groupe et à sa fonction duale, la protection de l'espace privé, que le mediaspace CoMedi tente de répondre.

DESCRIPTION DES SERVICES
Une mosaïque déformable montre dans chaque loge, les informations personnelles publiées des utilisateurs distants. Elle sert de véhicule à la conscience de groupe. Toute scène vidéo privée peut ne pas être publiée et, si elle est publiée, elle peut être filtrée par un traitement numérique : poster, stores vénitiens, codage par eigen-space. Ce dernier filtre, fondé sur la technique de l'Analyse en Composantes Principales, reconstitue l'image source en la dépouillant des "pixels non exportables" (correspondant par exemple au stylo que l'on mâche ou à la présence de visiteurs que l'on ne veut pas montrer).

CoMedi permet de glisser sans rupture de la communication informelle (par simple coup d'oeil sur la mosaïque) à la communication formelle (visiophonie). S'il en a les droits, l'utilisateur peut appeler un collègue distant et enclencher le système de suivi du visage par vision par ordinateur. Il peut alors se lever, la caméra le suit et, éloigné de quelques mètres de sa station, il peut encore contrôler son mediaspace à la parole. La télévisite d'un bureau distant (partage d'espace physique) se pratique selon la technique de la fenêtre virtuelle ou de la fovea.

DESCRIPTION TECHNIQUE

CoMedi est réalisé en Java pour les composantes IHM du système et la communication (multicast) entre les stations utilisateur. C et C++ ont été retenus pour les algorithmes de vision. Dans sa version actuelle, les télévisites par fenêtre virtuelle et par fovea ne sont pas intégrées au mediaspace. De même, le filtre par ACP fait l'objet d'un démonstrateur distinct.

COMEDI ET C3I2

L'utilisation des services ATM est transparente tant du point de vue technique que du point de vue utilisateur : 1 image toutes les 5 ou 10 secondes constitue un taux de rafraîchissement suffisant pour animer les scènes vidéo de la mosaïque dont la fonction, on le rappelle, est de favoriser la conscience de groupe. Nous regrettons l'absence de main d’oeuvre pour quantifier les performances attendues du réseau pour les tâches de télévisite ou de visiophonie, ou pour un mediaspace comportant plus de 10 utilisateurs.

7.2 Calcul à distance

Patrick Begou laboratoire LEGI

BUT DE NOTRE EXPERIMENTATION :

Jumeler un code de simulation d'écoulements turbulents (Couche de mélange, Jet, Sillage...) s'exécutant en parallèle sur le T3E du CEA sur le site Polygone avec sa partie Post-processing et visualisation sur une "ferme" de RISC6000 (interconnectées
à 100 Mbits) sur le site du domaine universitaire. L'outil de visualisation est AVS Express.
Cela pourrait permettre de modifier au vol la configuration de l'écoulement et de suivre son développement en temps réel.

ETAT D'AVANCEMENT :

Le code de simulation de couches de mélange est opérationnel sur T3E et en production.
Une partie des outils de post-processing ont été ré-écrits en C++/conception objet pour pouvoir s'intégrer au mieux avec l'outil de visualisation graphique.
La partie communication T3E/RISC6000 a été testée et validée : un processus sur les stations attend le démarrage du calcul et assure la réception des données expédiées depuis le T3E.
Dans cette première phase ce n'est pas le code de simulation qui expédie les données, mais un processus qui accède à des fichiers existants.
Une fibre optique supplémentaire a été posée aux frais de notre équipe de recherche, pour pouvoir établir une connexion directe de nos RISC6000 aux routeurs du CICG.

RESULTAT DES PREMIERS TESTS

Les premiers tests ont été réalisés avec une connexion en Ethernet 100Mb/s jusqu'au CICG, puis ATM du CICG au T3E. Des difficultés ont été rencontrées avec l'interface ATM CRAY et les performances de transferts sont faibles.
R. Dorge (CS/ATHESA pour le CEA) précise:

> Les pietres performances sont imputables au stack TCP/IP de cobea qui semble limitée à 2000 paquets / s (chiffre déterminé de manière empirique). Au delà de ce rythme, les performances s'écroulent. Lors d'un transfert FTP thor -> cobea, les paquets TCP font 512 octets. Le débit maximum sera observé sur un VP shappé à 2000 * 512 * 8 = 8 Mb/s.

Le phénomène de saturation ramène le débit théorique de 8Mb/s à 4.9 Kb/s (1600 fois plus lent) du RISC6000 vers le T3E et à 200 Kb/s du T3E vers le RISC6000.
Un second problème qui a été rencontré est un problème de sécurité. Les processus communiquent au travers de la librairie PVM. Celle ci utilise le protocole "rsh" qui n'est pas autorisé sur le T3E par mesure de sécurité. Chaque test a donc du faire l'objet d'une demande d'ouverture de ce protocole sur une durée limitée.
Enfin, l'utilisation du Korn-shell et de PVM sur T3E semblent soulever des problèmes de validation des environnements de travail. Le contournement de ces problèmes impose de démarrer PVM depuis le T3E.

SUITE DES TRAVAUX

Actuellement la partie développement de l'intégration des outils de post-processing dans l'outil de visualisation graphique continue avec la mise au point du code de simulation de JET en version parallèle (sous PVM) sur notre réseau de stations.

De nouveaux tests devraient prochainement être entrepris côté performances du réseau (information à confirmer).

7.3 Calculs parallèles

Philippe Augerat laboratoire LMC-IMAG

1 Problématique

L'objectif était d'utiliser une ligne ATM haut débit pour coupler des calculs parallèles entre le CRAY T3E du CEA Grenoble sur le site Polygone et la plate-forme parallèle du projet APACHE composée d'un IBM SP, de PCs sous SOLARIS et d'une SPARC bi-processeur sur le site du domaine universitaire.

2 Environnement

2.1 Environnement matériel

Dans la première phase, le réseau IP de la plate-forme APACHE et le réseau ATM auquel est connecté le Cray ont été reliés via une série de routeurs et commutateurs, dans cet ordre.
- les machines du projet APACHE sont connectées à un commutateur 100 mégabits.
- l'Ipsilon, commutateur IP capable, grâce à un mécanisme de routage dynamique sous ATM, de router 100 Mb/s de données sans perte de débit, selon une étude des services communs informatique de l'IMAG.
- un commutateur Ethernet 100 M
- le CISCO 7206, routeur effectuant le passage d'IP/Ether100 en IP/ATM.
- un commutateur ATM
- un routeur ATM du CEA donne accès au CRAY

Le débit réservé côté ATM était de 30 Mb/s et la machine choisie côté APACHE pour la première série de tests était une Sparc bi-processeur capable de délivrer d'après nos mesures 50 Mb/s de données sur le réseau de la plate-forme APACHE.

2.2 Environnement logiciel

Le logiciel ATHAPASCAN développé par le projet APACHE est basé sur des librairies de threads et de "passage de messages standards" dont MPI (Message Passing Interface). Différents choix de bibliothèques MPI sont possibles pour faire fonctionner ATHAPASCAN sur des plates-formes hétérogènes, mais aucune n'était disponible sur le CRAY. La bibliothèque MPI présente sur le CRAY ne permet pas d'adresser à la fois les noeuds internes du CRAY et les noeuds de la plate-forme APACHE. Pour cette raison, il a été proposé de travailler dans un premier temps en PVM, car la bibliothèque PVM CRAY s'interface correctement avec la bibliothèque PVM domaine public. Dans un second temps, il s'agissait de voir comment encapsuler MPI dans PVM pour permettre la compilation, l'exécution et la communication en ATHAPASCAN sur les deux plates-formes.

3 Expérimentation

3.1 Mesures brutes

La première étape du projet a été de mesurer des débits avec des outils standards comme l'outil de benchmark netperf. Malheureusement, les différences d'encodage entre le

CRAY et les machines UNIX actuelles font que les développeurs de netperf n'ont pas encore délivré de version fonctionnant entre les deux mondes. Les autres outils de mesure classique ont présenté le même problème. La mesure que nous avons fait a donc été la mesure d'envois de message via PVM qui permet une conversion de données entre machines hétérogènes. A titre de comparaison, nous avons fait des mesures de transferts de fichiers par FTP ainsi que des mesures sur les réseaux d'exploitation de l'IMAG.

Une première campagne de mesures a été faite. Voici les résultats numériques.

navajo (SPARC) - cobea (CRAY) via le réseau d'expérimentation C3I2 :
netperf : non disponible
ftp : 1200 Kbytes/s
pvm : 350 Kbytes/s

navajo - cobea via le réseau d'exploitation ARAMIS :
netperf : non disponible
ftp : 150 Kbytes/s
pvm : 70 Kbytes/s

PC SOLARIS - PC SOLARIS via le réseau d'expérimentation APACHE :
netperf : 3 Mbytes/s
ftp : 1500 kbytes/s
pvm : 2 Mbytes/s

navajo - curie (DIGITAL) (réseau d'exploitation IMAG) :
netperf : non disponible
ftp : 200 Kbytes/s
pvm : 100 Kbytes/s

A la lecture de ces résultats, on peut souligner deux problèmes : la très faible bande passante maximale obtenue sur le réseau C3I2, le mauvais comportement de PVM lorsque les machines qui communiquent n'utilisent pas le même encodage des messages.
Les problèmes de bande passante étant confirmés par les mesures d'autres projets et la mise en cause de la carte réseau ATM du CRAY, nous avons repris le projet en septembre avec deux directions : de nouvelles mesures après les changements matériels effectués, l'utilisation du produit PACX-MPI de l'Université de Stuttgart permettant de faire communiquer le CRAY et l'IBM SP au-dessus de MPI.

3.2 Nouvelles mesures après évolution de l'environnement

Par rapport à l'environnement de départ, la carte ATM du CRAY a été changée, le routeur Ipsilon a été remplacé par un routeur CISCO et la machine SPARC bi-processeur a subi un upgrade système pour passer en SOLARIS 2.6. Les nouvelles mesures permettent de noter une amélioration des performances puisque le débit mesuré entre le CRAY et la Sparc avec PVM passe à 400 Kbytes/s au lieu de 350 Kbytes/s. Ces chiffres restent malgré tout très faibles comparés au potentiel de débit de la station SPARC mesuré par netperf à 6000 Kbytes/s sur la plate-forme APACHE face à une seconde station Sparc bi-processeur.

3.3 Portage Athapascan

La seconde étape devait voir le portage d'Athapascan et son utilisation pour l'évaluation d'une exécution parallèle entre les sites Cray et Apache. Pour cela, il était nécessaire de disposer de bibliothèques MPI interopérables. De telles bibliothèques ne sont pas disponibles. Nous avons pensé nous rabattre sur le logiciel PACX-MPI permettant de faire communiquer un CRAY T3E et un IBM SP tournant chacun leur propre MPI.
Cette expérience n'a pu être menée à bien faute de temps. En effet, la version actuelle de ce logiciel ne fonctionne plus qu'entre deux CRAYs. Une version plus ancienne fonctionnait sur AIX mais pas sur SOLARIS. Il aurait été nécessaire de faire une mise à jour et un portage Cray, IBM SP et stations Unix.

4 Conclusion

En conclusion, ce projet a montré les difficultés de couplage des deux plates-formes parallèles, puisque des difficultés matérielles (couplage réseau lent) et logicielles (absence de briques logicielles de base sur le CRAY) se sont cumulées.
Par ailleurs, l'efficacité est restée très en dessous de l'efficacité espérée du fait de goulot d'étranglement au niveau des interconnexions de réseaux mais aussi du surcoût considérable de prise en compte de l'hétérogénéité des processeurs.

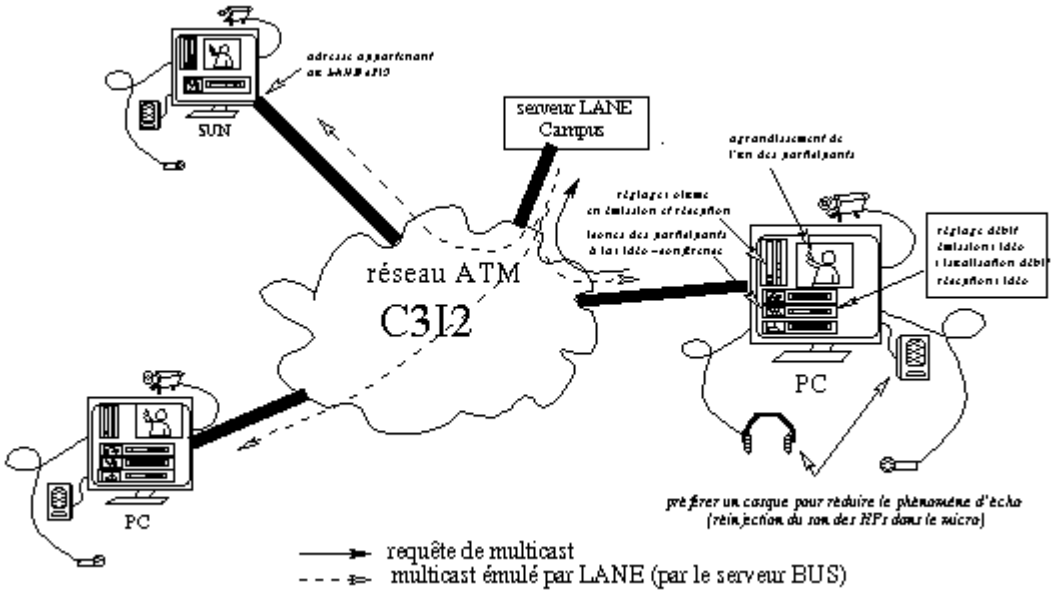
7.4 Visioconférences et télé-enseignement

Daniel Guéniche CNRS

Visioconférences

De très nombreuses visioconférences ont eu lieu sur C3I2 avec des matériels très divers, stations Unix sous différents systèmes, PC sous Linux, NT ou Win-95. Voici la description d'une plate-forme utilisée par le CNRS.
Un PC-Pentium Pro200 avec deux disques internes de 2 Go (l'un abrite NT, l'autre Linux) a été utilisé. Quatre choix s'offrent au démarrage : Windows NT 4.0, Linux 2.0.30 natif, Linux 2.0.29 avec noyau ATM 0.31 et multicast, Linux avec noyau IPV6.

Un kit vidéo a été installé. D'autres sites ont équipé également un de leur PC d'une carte d'acquisition audio/vidéo et d'une caméra. Ce qui nous a permis d'effectuer des expériences de vidéoconférence.
Voici ci-dessous un de ces tests entre le CNRS (PC/NT), le CEA (PC/W95) et l'INRIA (SUN) :
. Schéma :



. Logiciels utilisés :
. partie vidéo : vic (VIdeo Conferencing tool)
. partie audio : vat full duplex (Visual Audio Tool)
. Protocole : LANE pour supporter le multicast nativement

Voici quelques constats :
. Pour l'audio le *full duplex* est indispensable ;
. La qualité sonore laisse tout de même à désirer : ronflements, échos importants, hachures ;
. Dans les outils du domaine public comme vic ou vat on dispose de fonctionalités et indications intéressantes : plusieurs types d'encodage proposés, visualisation du débit en émission et en réception, réglage du débit d'émission, visualisation du taux de pertes, statistiques, etc.
. Parmi les codages proposés par vic, on a pu nettement en différencier deux :

. H261 (norme CCITT) :
. Médiocre définition de l'image
. S'accommode d'un faible débit : 128Kb/s
. Les mouvements restent lissés même si le correspondant bouge

. nv (variante de RTPv1 Xerox PARC Network Vidéo) :
. Très bonne définition même avec image de 20x20 cm (SCIF)
. Nécessite un débit élevé : entre 1 et 3Mb/s
. Lissage des mouvements s'ils ne sont pas trop rapides, sinon les mouvements intermédiaires sont remplacés par des carrés noirs mettant 2 à 3 secondes à se remplir

. Windows 95 ne convient pas à ce type d'application (blocages, instabilité), mais Windows NT a un comportement satisfaisant ;
. La taille de la fenêtre de visualisation de l'image vidéo a une influence directe sur le CPU qui oscille autour de 90% avec un *Pentium* 200 pour une fenêtre de 15x15 cm. Si on agrandit davantage la fenêtre, le son devient haché.
Un autre compte-rendu d'expérimentation est disponible ici : www.polycnrs-gre.fr/FrPages/T-sem.html

Télé-enseignement

Toute expérimentation dans ce domaine demande la participation très active d'enseignants. Sur le Polygone, au laboratoire de Génie Atomique, Philippe Masse, enseignant-chercheur a été prêt à soutenir le projet d'une retransmission simultanée de son cours à des étudiants se trouvant à distance ou d'une télé-assistance sur un logiciel graphique qu'il a commercialisé (Flux Expert), ou d'échanges autour de ce logiciel (simulation de phénomènes physiques) avec des chercheurs distants. Les trois applications nécessitaient l'exportation d'une fenêtre X11.
Une recherche documentaire sur les comptes-rendus d'expériences similaires montrait que l'enthousiasme était pondéré : tantôt les étudiants regrettaient de ne pas voir le professeur ou de se sentir isolés, tantôt c'était le professeur qui se plaignait de la complexité des outils, de la plus lourde préparation de ses interventions, de ne pas « sentir » la salle distante…
Après une réflexion sur ces documents et de nombreux échanges avec des techniciens du domaine et des enseignants, les objectifs ont été les suivants :

. Limiter pour le professeur les contraintes : ajouter un assistant dans la salle enseignant et un dans la salle étudiants ;
. « Rapprocher » cette salle distante : le professeur doit pouvoir l'entendre, la voir ;
. L'étudiant distant doit sentir qu'il fait partie intégrante du cours : il doit pouvoir intervenir comme s'il était en local (micro, logiciels interactifs temps réel), et surtout entendre parfaitement le professeur (le son doit être parfait, même au détriment de la qualité vidéo) ;
. Limiter le coût afin que l'opération soit abordable : recourir prioritairement aux logiciels du domaine public, écrire si besoin des applications, trouver le matériel offrant le meilleur rapport qualité/prix ;
. Utiliser le moins de bande passante possible dans l'espoir de rendre l'opération réalisable sur un réseau Ethernet (commuté) : l'usage du multicast peut-être ici déterminant, ainsi que l'utilisation d'une bonne compression, et la maîtrise du débit d'émission ;
. L'opération ne doit pas être liée à une marque de station, un système d'exploitation : là encore les logiciels du domaine public sont précieux.

Le choix des logiciels a été le suivant : pour l'exportation de fenêtres X11, un seul convenait : nv (Network Video), pour la vidéo vic (VIdeo Conferencing tool) et pour l'audio : vat (Visual Audio Tool) ont été retenus. Un outil de dialogue en *muticast* a été développé par 2 stagiaires : Forum**,** dont la particularité est d'être simple : il n'accepte que du texte et tient en une fenêtre munie d'un ascenseur qui peut être très petite.

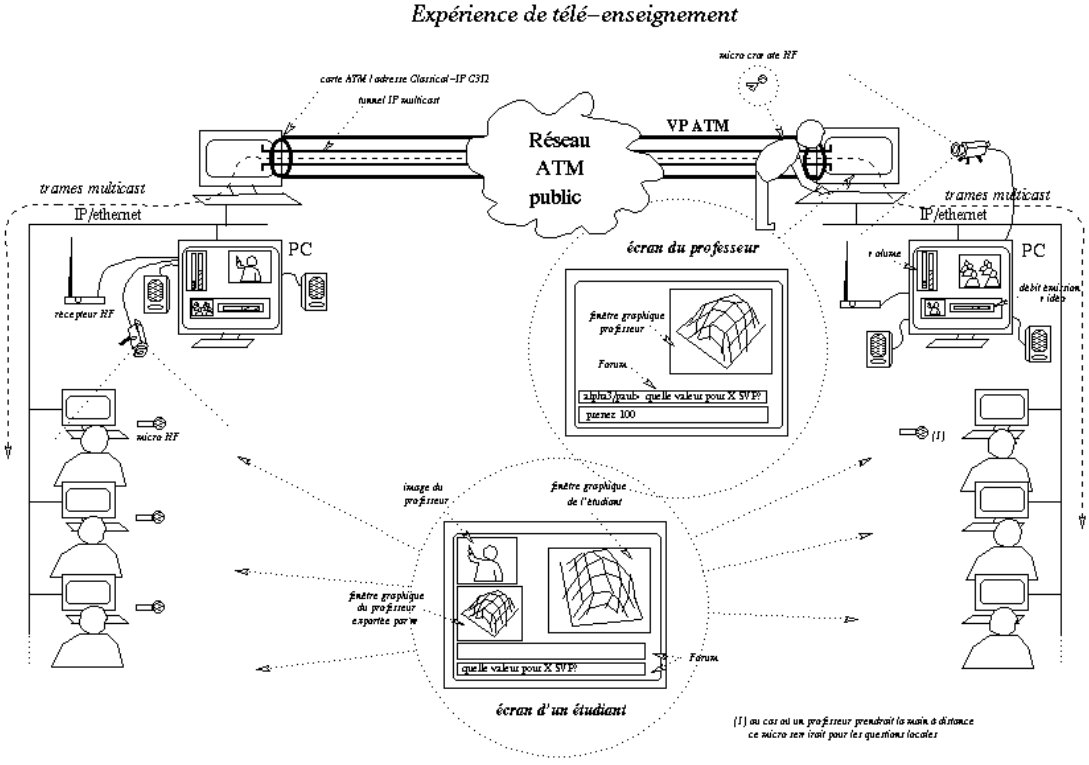
L'installation s'est déroulée ainsi :

. Mise en place du *multicast* sur le noyau du serveur côté GA ;
. Installation et configuration des cartes ATM ;
. Installation du logiciel mrouted sur les serveurs et configuration d'un tunnel « privé » (n'empruntant pas le réseau Mbone –le réseau mondial multicast) entre les deux serveurs sur les 2 sites;
. Installation sur les deux PCs des kits vidéo et de la connectique audio.

Dans la salle où est installé le professeur (appelée locale), ainsi que quelques étudiants :
. Le professeur dispose d'un micro-cravate HF (Haute Fréquence –sans fil : un récepteur est raccordé au PC)
. Une caméra couvre avec son grand angle sur petits déplacements
. Deux haut-parleurs retransmettent la voix de ceux qui prennent la parole en salle étudiants

Dans la salle étudiants (appelée distante)**:**

. 4 micros HF ont été achetés dans un premier temps (~1 pour 4 étudiants)
. Le champ de la caméra couvre la salle
. Deux haut-parleurs retransmettent la voix du professeur



Voici le compte-rendu du premier essai (juin 98) :

Le professeur a commencé son cours pour la salle "locale" : création d'une image 3D à l'aide du logiciel Flux Expert. Les élèves "cobayes" recevaient la copie du graphique que le professeur obtenait, et devaient selon ses instructions réaliser le même graphique (les élèves dans la salle distante se contentaient de suivre, puisque Flux Expert n'était pas installé sur les stations distantes).

Malheureusement le professeur découvrait en même temps que ses élèves les outils multicast, et de plus les élèves ne connaissaient pas du tout Flux expert. Une grande confusion s'est alors peu à peu installée : les élèves se perdaient dans les diverses fenêtres de leur écran et ne savaient plus où ils devaient entrer les instructions.

Un autre professeur à distance a alors pris la main : il a exporté sa fenêtre et surlignait des extraits du dialogue graphique qu'il commentait (il n'y avait aucun décalage avec la réception de sa voix).
Ses explications ont été suivies avec un grand intérêt.
Nous avons passé ensuite 2 heures autour d'une table pour commenter cette expérience :
. Le professeur distant s'est plaint de l'écho qu'il entendait depuis notre salle : ceci était dû au fait que notre micro (nous n'avions pas encore reçu le micro-cravate) était près du haut-parleur et donc ré-émettait sa voix. En coupant notre micro, le phénomène a disparu, mais n'est il pas finalement souhaitable d'entendre un tout petit peu notre voix à distance ?
. Le cafouillage qui a eu lieu durant la première partie s'est produit localement, et ne mettait donc pas en cause le télé-enseignement.
. La réussite de la prise en main distante a par contre démontré :

. qu'un cours était parfaitement possible depuis une salle distante ;
. la réversibilité du système (en local il a suffit de tourner la caméra vers la salle).

. Sur certains écrans, le fait de passer sur une autre fenêtre, rendait la fenêtre "d'explications" illisible : le gestionnaire de fenêtre CDE est trop gourmand en ressources, le problème est très atténué avec *openwin* (gestionnaire de fenêtre natif de SUN). Par ailleurs les cartes graphiques standard 8 plans limitent la table de couleurs, il faudrait essayer de faire en sorte que les logiciels utilisés partagent une même table de couleurs au lieu de définir leur propre table.

. Les professeurs ont été unanimes à affirmer qu'ils consacraient 90 % de leur temps en cours à prendre en aparté un ou plusieurs élèves. Comment faire en télé-enseignement ?

La solution d'équiper chaque salle de 3 casques HF a semblé satisfaisante :

Les étudiants en difficulté pourraient ainsi –sans gêner les autres– bénéficier d'une explication supplémentaire (avec en parallèle une diffusion sur une adresse multicast différente).

Réflexions sur cette expérience :

. L'attention est accrue lorsqu'il s'agit de suivre un cours distant. Ceci sans doute parce qu'on sait qu'on ne pourra pas immédiatement interrompre l'intervenant si on "décroche". Si une voiture a de mauvais freins, il faut :

. redoubler d'attention
. réduire la vitesse

. Le professeur doit donc parler lentement, soigner l'articulation et marquer des pauses fréquentes pour tenir compte de la réactivité déphasée à distance par le décalage de la réception de sa voix.

. Un logiciel (*unicast*) est à l'étude : l'étudiant pourra colorer (en vert, orange ou rouge) une case de son écran. Cette case sera exportée en temps réel dans un

tableau figurant dans l'écran du professeur. A tout moment celui-ci saura donc si ses élèves (locaux ou distants) suivent facilement, difficilement ou sont bloqués.

. L'enseignement et le logiciel doivent faire un pas l’un vers l’autre. Les outils doivent être perfectionnés, mais ils ne pourront jamais rendre transparent un enseignement à distance. Aussi l'enseignement lui-même doit-il s'adapter à cette technologie.

7.5 Performances IP sur ATM entre le CrayT3E et une IBM Risc.

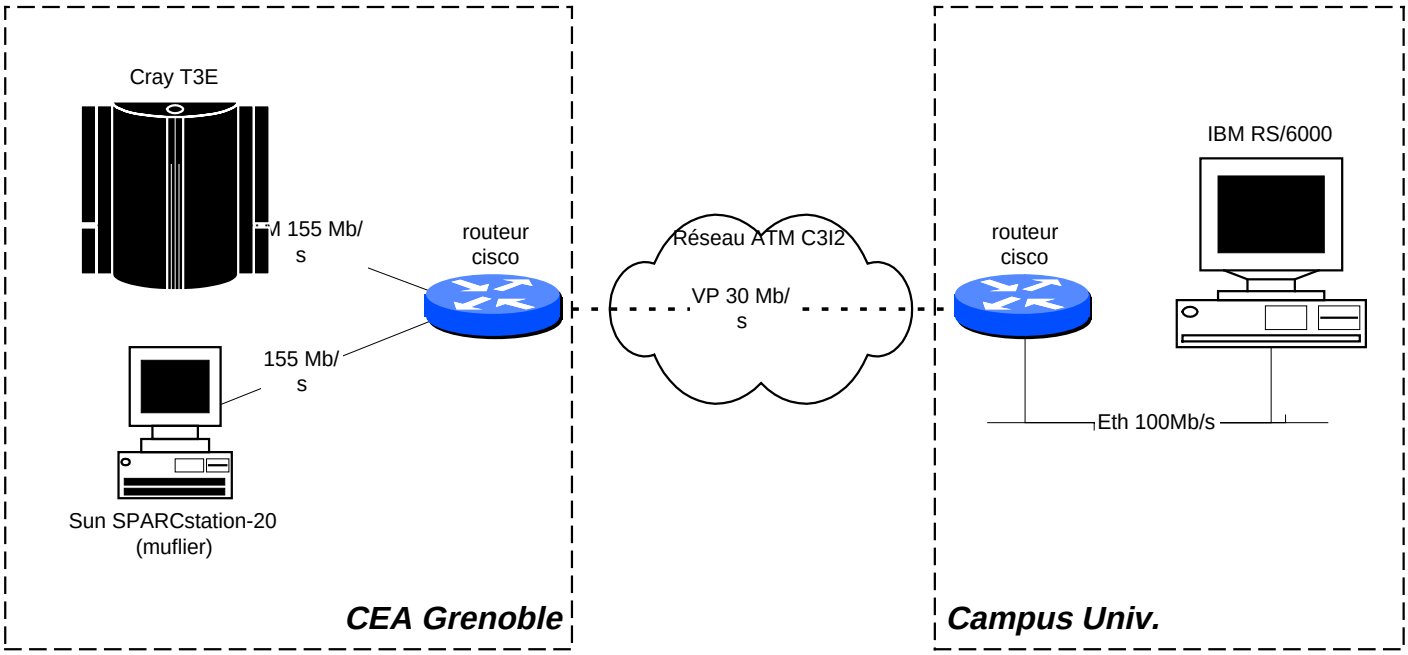
Raoul Dorge (CS/ATHESA pour le CEA)

INTRODUCTION

Nombres de projets ont été définis pour véhiculer du flux expérimental afin de tirer profit des artères haut-débits du réseau C3I2. Parmi ces expérimentations, le LEGI (Laboratoire des Ecoulements Géophysiques de Grenoble) s'est montré intéressé pour "faire du PVM hétérogène" entre un supercalculateur Cray T3E et une BM Risc6000. Les performances décevantes obtenues ont nécessité un audit réseau. Ce rapport met en évidence que les mauvaises performances sont imputables au Cray T3E.

ARCHITECTURE LOGIQUE

Le schéma suivant résume la topologie mis en œuvre.



PERFORMANCES MESUREES

Les performances obtenues par différents transferts FTP sont les suivantes :
Cray T3E ←→ IBM Risc = quelques Ko/s (5 Ko/s dans un sens à 250 Ko/s dans l'autre sens)
Sun SS20 ←→ IBM Risc = plusieurs 100Ko/s

AUDIT RESEAU

Un analyseur ATM a été positionné juste devant le Cray T3E. La capture suivante a été faite lors d'un transfert FTP dans le sens IBM → Cray T3E. L'extrait ci-dessous a pour but de montré le comportement TCP du Cray T3E.

Une trame est codée de la manière suivante :
N° trame - [TCP] cray (port source) > @ ibm (port destin.) N° d'acquittement TCP
ou
N° trame - [TCP] ibm (port source) > @ cray (port destin.) N° séquence TCP] (longueur)

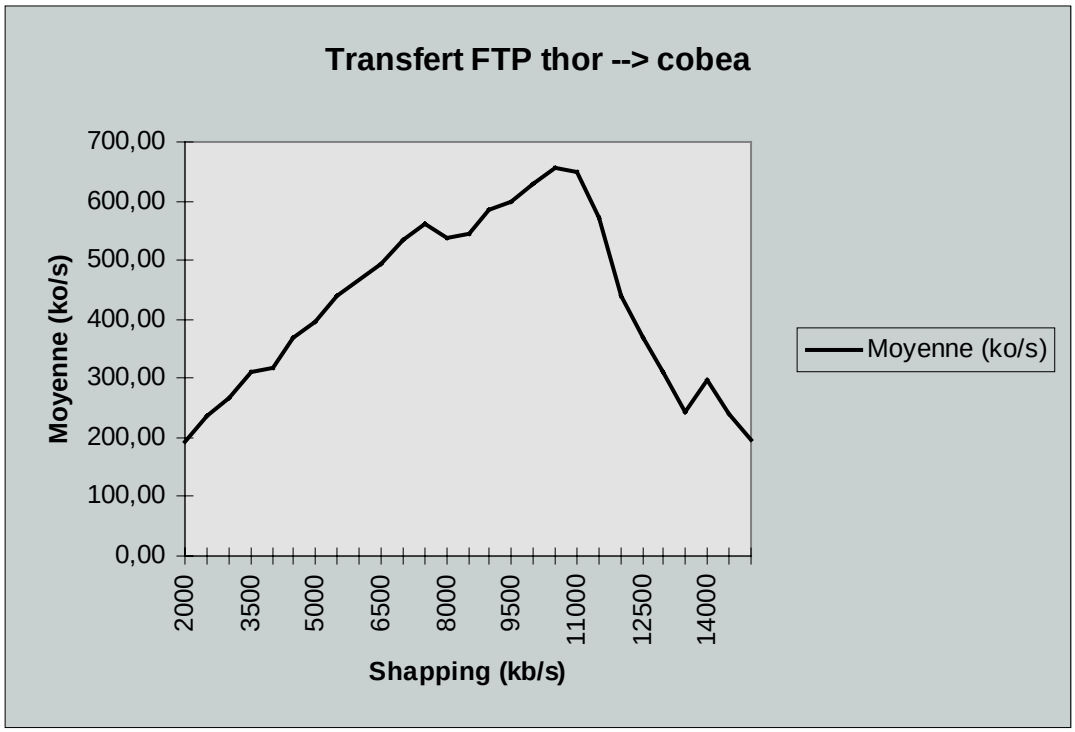
```
0000000086 - [TCP] ibm(4340) > cray(20) 12801(512)
0000000087 - [TCP] ibm(4340) > cray(20) 13313(512)
0000000088 - [TCP] ibm(4340) > cray(20) 13825(512)
0000000089 - [TCP] ibm(4340) > cray(20) 14337(512)
0000000090 - [TCP] ibm(4340) > cray(20) 14849(512)
0000000091 - [TCP] ibm(4340) > cray(20) 15361(512)
0000000092 - [TCP] cray(20) > ibm(4340) ack 15361
0000000093 - [TCP] ibm(4340) > cray(20) 15873(512)
0000000094 - [TCP] ibm(4340) > cray(20) 16385(512)
0000000095 - [TCP] ibm(4340) > cray(20) 16897(512)
0000000096 - [TCP] ibm(4340) > cray(20) 17409(512)
0000000097 - [TCP] ibm(4340) > cray(20) 17921(512)
0000000098 - [TCP] cray(20) > ibm(4340) ack 15361
0000000099 - [TCP] cray(20) > ibm(4340) ack 15361
0000000100 - [TCP] cray(20) > ibm(4340) ack 15361
0000000101 - [TCP] cray(20) > ibm(4340) ack 15361
0000000102 - [TCP] ibm(4340) > cray(20) 15361(512)
0000000103 - [TCP] ibm(4340) > cray(20) 18433(512)
0000000104 - [TCP] cray(20) > ibm(4340) ack 17921
0000000105 - [TCP] cray(20) > ibm(4340) ack 17921
0000000106 - [TCP] cray(20) > ibm(4340) ack 17921
0000000107 - [TCP] ibm(4340) > cray(20) 18945(512)
0000000108 - [TCP] ibm(4340) > cray(20) 19457(512)
0000000109 - [TCP] cray(20) > ibm(4340) ack 17921
0000000110 - [TCP] cray(20) > ibm(4340) ack 17921
0000000111 - [TCP] ibm(4340) > cray(20) 17921(512)
0000000112 - [TCP] cray(20) > ibm(4340) ack 19969
0000000113 - [TCP] cray(20) > ibm(4340) ack 19969
0000000114 - [TCP] ibm(4340) > cray(20) 19969(512)
0000000115 - [TCP] ibm(4340) > cray(20) 20481(512)
0000000116 - [TCP] cray(20) > ibm(4340) ack 20993
0000000117 - [TCP] ibm(4340) > cray(20) 20993(512)
0000000118 - [TCP] ibm(4340) > cray(20) 21505(512)
0000000119 - [TCP] ibm(4340) > cray(20) 22017(512)
0000000120 - [TCP] cray(20) > ibm(4340) ack 22529
0000000121 - [TCP] ibm(4340) > cray(20) 22529(512)
```

Les trames 86, 87, 88, 89,90 et 91 constituent le flux FTP normal de l'IBM vers le T3E.
La trame 92 du Cray T3E acquitte jusqu'à la trame 90 inclus.
Puis l'IBM poursuit son émission par les trames 93, 94, 95, 96, 97.
Ensuite, les choses se gâtent ... Le Cray T3E retransmet 4 fois (trames 98, 99, 100, 101) son acquittement 92. Il exige ainsi la retransmision de la trame 91. Or cette trame 91 n'a pas été perdu par le réseau puisqu'elle a été capturé par l'analyseur.
L'IBM retransmet en 102 la trame demandée, puis poursuit son émission FTP (trame 103).
Ensuite le scénario se reproduit ... Le Cray T3E retransmet 3 fois (trames 104, 105, 106) un acquittement pour exiger la retransmission de la trame 97. Or la trame 97 est passé auparavant dans le buffer de l'analyseur.
D'une manière générale, on constate que le Cray T3E exige la retransmission de trames valides.

MISE EN EVIDENCE D'UN EFFET BOULE DE NEIGE

Il est possible sur les commutateurs ATM cisco de limiter le débit ATM (shaping). Ainsi, le shaping du VP CEA-Campus a été modifié pour offrir différent débit entre le

CEA et le Domaine Universitaire. A chaque fois, un transfert FTP a été mesuré depuis l'IBM vers le Cray T3E. Les résultats de ces mesures apparaissent sous forme de graphe ci-dessous :



Il apparaît que les performances s'effondrent lorsque la valeur du shaping devient > 11 Mb/s.

INTERPRETATION DES RESULTATS

Ce phénomène ne se reproduit pas lorsque les tests sont effectués avec une machine distante directement raccordée en ATM. Dans ce cas, la négociation MSS (Maximum Size Segment) permet d'échanger des trames IP de 9000 octets alors qu'entre le Cray T3E et l'IBM, la négociation MMS s'établit à 512 octets (MTU du Cray T3E = 9180, MTU de l'IBM = 1500).
Ainsi, le Cray T3E semble dans l'incapacité de traiter plus de 1400 trames IP par seconde (1400 * 512 = 700 Ko/s = débit maximum mesuré).

La solution pour contourner ce problème et obtenir des performances en rapport avec le débit des VP aurait été d'établir une connexion ATM directe entre le Cray T3E et l'IBM, sans passer par un routeur qui limite la taille des datagrammes IP. Le rapport d'efficacité avec et sans routeur s'établit ainsi : 9180/512 = 17,9.

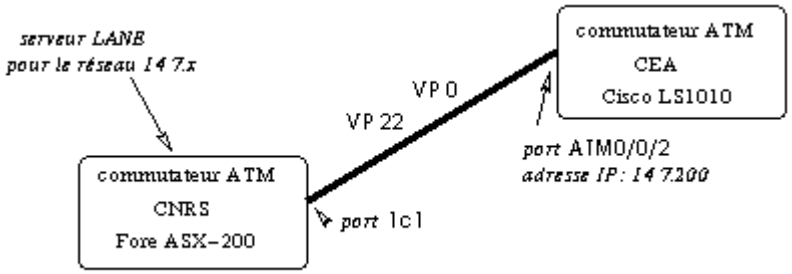
7.6 Acquisition d’un savoir-faire en ATM-WAN et en routage dynamique BGP4

Daniel Guéniche CNRS et Jean-Pierre Augé INRIA

De nombreuses expérimentations sur les protocoles ATM, IP ainsi que sur le routage ont été menées par les différents ingénieurs des organismes à la fois pour acquérir un savoir-faire dans ces domaines mais aussi pour construire les briques de base de l’infrastructure et permettre aux applications de fonctionner. Ci-dessous nous ne décrirons que 3 expérimentations ou mises en oeuvre : le tunnelling de la signalisation ATM, des exemples de tests de performance (débit, shaping) et le routage dynamique IP mis en service.

Tunneling de la signalisation ATM

Les normes de l'ATM Forum réservent pour la signalisation UNI les 31 premiers circuits virtuels (VCs) du canal 0 (VP0), exemple le VC5 pour la propagation d'appels, le 16 pour ILMI,... Or sur C3I2, France Télécom se réserve ce canal 0 pour la signalisation entre ses équipements et nous attribue des canaux différents (exemple : le VP 211 entre le Domaine universitaire et l’INRIA). La question se posait donc de savoir s’il serait possible entre deux sites –surtout équipés en constructeur différents– de véhiculer la signalisation entre nos équipements (pour établir des circuits virtuels dynamiques SVCs) au travers d’un VP autre que 0.
Un test a d’abord été effectué entre les commutateurs du CEA (équipé d’un CISCO) et du CNRS (équipé d’un FORE) sur le site Polygone avant d’être déployé sur C3I2.
Voici la description de cette première expérience :
Le but était d’établir des SVCs via le VP22.



1. On crée le VP22 dans les 2 sens (exemple de syntaxe côté Fore) :
   configuration vpc> new 1c1 22 orig -shapeovpi -reserved 500
   configuration vpc> show 1c1

Input		Output							
Port	VPI	Port	VPI	MaxBW	BW	MaxVCs	VCs	UPC	Prot
1C1	0	terminate		N/A	0.0K	511	4	N/A	pvc
1C1	22	terminate		499.9K	0.0K	511	0	N/A	pvc
originate		1C1	0	N/A	0.0K	511	4	N/A	pvc

originate	1C1	22	499.9K	0.0K	511	0	N/A	pvc

2. On demande à ce que le VP22 bénéficie de la signalisation UNI :

configuration uni30> new 1c1 22

configuration vcc> show 1c1

Input Port	VPI	VCI	Output Port	VPI	VCI	UPC	Protocol
1C1	0	5	1CTL	0	46		uni30
1C1	0	16	1CTL	0	75		uni30
1C1	22	5	1CTL	0	137		uni30
1C1	22	16	1CTL	0	138		uni30

3. Il ne reste qu'à créer une route statique entre les commutateur du CEA et du CNRS :

Commutateur CNRS :

configuration nsap route> new 0x39250f000000250004230101000603e5b3c01001c1 22

Commutateur CEA :

atm route 47.0005.80ff.e100.0000.f21a.32b8 ATM0/0/2.22

4. Essais de *ping*s → les VCs se créent :

configuration vcc> show 1c1

Input Port	VPI	VCI	Output Port	VPI	VCI	UPC	Protocol
1C1	0	5	1CTL	0	46		uni30
1C1	0	16	1CTL	0	75		uni30
1C1	22	5	1CTL	0	137		uni30
1C1	22	16	1CTL	0	138		uni30
1C1	22	41	1CTL	0	419		uni30
1C1	22	43	1CTL	0	420		uni30

Dès les premiers *ping*s échangés, des circuits virtuels dynamiques se sont créés : le test s'est donc révélé concluant.

Tests du shaping (lissage)

Les VPs mis à notre disposition par l'opérateur sont à débit fixe. Celui-ci fait la police, c'est à dire n'accepte que les trames avec un flux inférieur au début des VPs mis en place. Il faut donc que chaque commutateur de site mette en œuvre une fonction de lissage du trafic (shaping) qui vient du site avant de l'envoyer sur C3I2.

Dans le test suivant, prunie*r* est une machine de l'INRIA connectée au commutateur ATM Cisco de l'INRIA qui communique avec une machine du CNRS. Le VP entre les 2 sites est de 33 Mb/s.

1. Le shaping sur les 2 commutateurs (INRIA et CNRS) est fixé à 33 Mb/s (condition normale) :

prunier% /bin/ttcp -t -s -p 9 -n 5000 172.20.5.2 (essai sur 5000 paquets)
ttcp-t: buflen=8192, nbuf=5000, align=16384/0, port=9 tcp -> 172.20.5.2
ttcp-t: socket
ttcp-t: connect
ttcp-t: 40960000 bytes in 11.31 real seconds = **27.63 Mbit/sec** +++
ttcp-t: 40960000 bytes in 2.52 CPU seconds = 123.79 Mbit/cpu sec
ttcp-t: 5000 I/O calls, msec/call = 2.32, calls/sec = 442.06
ttcp-t: 0.1user 2.3sys 0:11real 22% 0i+0d 0maxrss 0+0pf 1676+124csw

Le débit obtenu est d'environ 28 Mb/s, ce qui correspondant au débit maximum que l'on peut atteindre (les entêtes des protocoles utilisent l'autre partie de la bande passante)

2. Les fonctions de shaping ne sont pas validées sur les commutateurs :
. Le « connect tcp » passe très rarement... :

```
prunier% /bin/ttcp -t -s -p 9 -n 1 172.20.5.2
ttcp-t: buflen=8192, nbuf=1, align=16384/0, port=9  tcp  -> 172.20.5.2
ttcp-t: socket
```

. La limite de pointes qui dépassent le débit souscrit (*bursts*) acceptée par France Télécom se situe aux environs de 800 bytes :

```
prunier% ping -s 172.20.5.2 900
PING 172.20.5.2: 900 data bytes
908 bytes from 172.20.5.2: icmp_seq=0. time=3. ms
908 bytes from 172.20.5.2: icmp_seq=1. time=2. ms
908 bytes from 172.20.5.2: icmp_seq=2. time=2. ms
908 bytes from 172.20.5.2: icmp_seq=3. time=2. Ms
        ← surpression du shaping... plus rien ne passe ...
908 bytes from 172.20.5.2: icmp_seq=14. time=2. Ms
        → rétablissement du shaping
908 bytes from 172.20.5.2: icmp_seq=15. time=2. ms
908 bytes from 172.20.5.2: icmp_seq=16. time=2. ms
```

3. Shaping fixé à 34 Mb/s sur les équipements (au lieu de 33 Mb/s) :

```
prunier% /bin/ttcp -t -s -p 9 -n 100 172.20.5.2
ttcp-t: buflen=8192, nbuf=100, align=16384/0, port=9  tcp  -> 172.20.5.2
ttcp-t: socket
ttcp-t: connect
ttcp-t: 819200 bytes in 9.03 real seconds = 0.69 Mbit/sec +++
ttcp-t: 819200 bytes in 0.06 CPU seconds = 110.10 Mbit/cpu sec
ttcp-t: 100 I/O calls, msec/call = 92.44, calls/sec = 11.08
ttcp-t: 0.0user 0.0sys 0:09real 0% 0i+0d 0maxrss 0+0pf 63+3csw
```

Le débit est ridicule (0.69 Mb/s) : le shaping est obligatoire est doit être correctement dimensionné.

Tests de débits effectifs

Une fois les équipements réseaux installés sur les sites et correctement configurés (adressage, shaping, routage, …), des tests de performances ont été réalisés entre tous les sites, dans tous les sens, pour couvrir toutes les combinaisons. Les outils utilisés ont été bien naturellement *ping* mais aussi *ttcp* et *netperf*. Voici un exemple de trace avec un VP a 33 Mbps :

```
azalee% ttcp -t -s -p9 195.221.230.97 < BIG  (1)
ttcp-t: nbuf=1024, buflen=1024, port=9
ttcp-t: 0.0user 0.2sys 0:02real 8% 0i+0d 0maxrss 0+0pf 0+0csw
ttcp-t: 8192000 bytes in 0.220000 CPU seconds = 36363.636364 KB/cpu sec
ttcp-t: 8192000 bytes in 2.750955 real seconds = 2908.081012 KB/sec
→ 29.2 Mb/s
```

azalee% ftp frene.c3i2.imag.fr
ftp> put test3M
3202712 bytes sent in 0,45 seconds (**2,4e+03 Kbytes/s**)
→ 19.2 Mb/s

Utilisation du routage dynamique BGP-4 pour sécuriser le transport du trafic de production sur C3I2 par un back-up sur ARAMIS

Lorsque le réseau d'expérimentation C3I2 a été estimé stable, que les débits mesurés ont été ceux attendus, le passage de certains trafics d'exploitation non vitaux y a été envisagé (*news*, *multicast*, cache *web*, ...). Cela tombait d'autant mieux que la lenteur des échanges avec le domaine universitaire via le réseau régional ARAMIS commençait à poser des problèmes.
Le site CNRS, avec ses communications avec le domaine universitaire, servira d'exemple dans ce qui suit.
Ce site n'étant pas équipé de routeur ATM, le logiciel gated a été installé sur un serveur (serveur-CNRS) doté d'une interface Ethernet et d'une interface ATM. Gated est un logiciel qui émule le fonctionnement d'un routeur sur une station UNIX. gated a été configuré pour que :

. Sa route par défaut soit le routeur connecté à ARAMIS

. Il annonce les réseaux du CNRS au domaine universitaire via BGP (Border Gateway Protocol) à travers C3I2

. Il accepte les annonces de certains réseaux du domaine universitaire via C3I2

Sur le site CNRS la route par défaut des laboratoires n'est plus le routeur ARAMIS, mais ce serveur avec gated. Ce serveur (grâce aux annonces lui parvenant via C3I2), sait qu'il peut joindre le domaine universitaire via son interface ATM (donc C3I2). Pour les autres réseaux il renvoie les datagrammes sur le routeur ARAMIS qui est sa route par défaut.

Exemple de traces :

1. Depuis le CNRS (labs est une machine de ce site, horus du domaine universitaire) :

labs > traceroute horus.imag.fr
1 serveur-CNRS.cnrs-grenoble.fr (177.143.1.13) 4 ms 3 ms 2 ms
2 campusr.c3i2.imag.fr (195.221.230.97) 3 ms 4 ms 3 ms
3 r-imag.grenet.fr (193.54.185.123) 5 ms 5 ms 5 ms
4 horus.imag.fr (129.88.38.2) 5 ms 5ms 6 ms

2. Depuis le domaine universitaire

horus.imag.fr> traceroute labs.cnrs-grenoble.fr
1 imag-campus (129.88.38.254) 4 ms 2 ms 2 ms
2 r-cicg-atm.grenet.fr (193.54.185.120) 2 ms 2 ms 2 ms
3 serveur-CNRS.c3i2.imag.fr (195.221.230.109) 2 ms 3 ms 2 ms
4 labs.cnrs-grenoble.fr (177.143.1.26) 5 ms 9 ms 4 ms

3. Vers un laboratoire du site (crtbtl est une station sur le même site que labs, le CNRS):

labs > traceroute crtbt1
1 serveur-CNRS.cnrs-grenoble.fr (177.143.1.13) 3 ms 3 ms 3 ms
2 Rt-BT.cnrs-grenoble.fr (177.143.1.50) 5 ms 4 ms 4 ms
3 crtbt.cnrs-grenoble.fr (177.143.50.10) 5 ms 4 ms 4 ms

4 crtbt1.cnrs-grenoble.fr (177.143.49.2) 5 ms 4 ms 5 ms

et aussitôt après :

labs > traceroute crtbt1
1 Rt-BT.cnrs-grenoble.fr (177.143.1.50) 5 ms 5 ms 4 ms
2 crtbt.cnrs-grenoble.fr (177.143.50.10) 4 ms 11 ms 4 ms
3 crtbt1.cnrs-grenoble.fr (177.143.49.2) 5 ms 10 ms 5 ms

On ne passe plus par le serveur avec gated, car la route directe a été apprise par ICMP redirect

4. Vers un autre laboratoire de France :

labs > traceroute www.jussieu.fr
2 serveur-CNRS.cnrs-grenoble.fr (177.143.1.13) 4 ms 4 ms 4 ms
3 Rt-CNRS (177.143.1.1) 4 ms 4 ms 4 ms
4 ft-aramis.cnrs-grenoble.fr (193.55.52.1) 6 ms 6 ms 6 ms
5 grenoble.aramis.ft.net (194.199.224.49) 19 ms 13 ms 12 ms
6 grenoble.renater.ft.net (194.199.224.114) 12 ms 12 ms 12 ms
7 stamand1.renater.ft.net (195.220.180.5) 27 ms 24 ms 23 ms
....

5. Trace d'une rupture puis de la remontée du lien C3I2 entre le domaine universitaire et le CNRS

Vérification que l'on passe bien par C3I2 pour joindre le domaine universitaire (spectro est une machine du domaine universitaire)

serveur-CNRS % traceroute spectro.ujf-grenoble.fr
1 campusr.c3i2.imag.fr (195.221.230.97) 1 ms 1 ms 1 ms
2 r-ujf.grenet.fr (193.54.185.124) 5 ms 4 ms 3 ms
3 phy-gate.ujf-grenoble.fr (193.54.238.6) 3 ms 3 ms 5 ms
4 spectro.ujf-grenoble.fr (193.54.234.59) 3 ms 2 ms 3 ms

Traces de gated.log :

Jul 17 15:45:24 BGP RECV 195.221.230.97+179 -> 195.221.230.109+2875
Jul 17 15:45:24 BGP RECV message type 4 (KeepAlive) length 19
Jul 17 15:46:24 (gated reçoit les hello bgp émis par le domaine universitaire)

On débranche la fibre optique reliant le CNRS à C3I2

Jul 17 15:46:24 BGP SEND 195.221.230.109+2875 -> 195.221.230.97+179
Jul 17 15:46:24 BGP SEND message type 4 (KeepAlive) length 19
Jul 17 15:47:24
Jul 17 15:47:24 BGP SEND 195.221.230.109+2875 -> 195.221.230.97+179
Jul 17 15:47:24 BGP SEND message type 4 (KeepAlive) length 19
Jul 17 15:47:24 BGP SEND 195.221.230.109+2875 -> 195.221.230.97+179
Jul 17 15:47:24 BGP SEND message type 4 (KeepAlive) length 19
Jul 17 15:47:24

Gated ne reçoit plus de hello du domaine universitaire

Jul 17 15:48:24 bgp_traffic_timeout: holdtime expired for 195.221.230.97 (External AS 1942)
Jul 17 15:48:24 NOTIFICATION sent to 195.221.230.97 (External AS 1942): code 4 (Hold Timer Expired Error) data
Jul 17 15:48:24

Jul 17 15:48:24 BGP SEND 195.221.230.109+2875 -> 195.221.230.97+179
Jul 17 15:48:24 BGP SEND message type 3 (Notification) length 21
Jul 17 15:48:24 BGP SEND Notification code 4 (Hold Timer Expired Error) subcode 0

Gated bascule le trafic sur ARAMIS (il a fallu ~2mn)

```
serveur-CNRS % traceroute spectro.ujf-grenoble.fr
 1  commutateur-interne (177.143.240.1)  4 ms *  4 ms
 2  rt-CNRS (177.143.1.1)  3 ms  2 ms  2 ms
 3  ft-aramis (193.55.52.1)  7 ms  4 ms  5 ms
 4  grenoble.aramis.ft.net (194.199.224.49)  27 ms  12 ms  21 ms
 5  cicg-grenoble.aramis.ft.net (194.199.224.122)  22 ms 15 ms 16 ms
 6  aramis.grenet.fr (193.54.184.1)  25 ms  25 ms  21 ms
 7  r-ujf.grenet.fr (193.54.185.124)  36 ms  20 ms  36 ms
 8  phy-gate.ujf-grenoble.fr (193.54.238.6)  21 ms  17 ms  29 ms
 9  spectro.ujf-grenoble.fr (193.54.234.59)  29 ms  23 ms  20 ms
```

On rebranche les fibres : on rebascule sur C3I2 après ~2mn

```
serveur-CNRS % netstat -rn (extrait):
Destination      Gateway          Flags      Refs   Use  Interface
193.54.232       195.221.230.97   UG         0      0  qab1 ← via C3I2
193.54.232.33    177.143.1.1      UGHD       0      172  tu0  ← via ARAMIS
193.54.233       195.221.230.97   UG         0      0  qab1 ← via C3I2
```

En cet instant, toutes les routes (comme 193.54.232.33) n'ont pas encore basculés.

Enfin :

```
serveur-CNRS % traceroute spectro.ujf-grenoble.fr
 1  campusr.c3i2.imag.fr (195.221.230.97)  1 ms  1 ms  1 ms
 2  r-ujf.grenet.fr (193.54.185.124)  3 ms  5 ms  3 ms
 3  phy-gate.ujf-grenoble.fr (193.54.238.6)  3 ms  4 ms  3 ms
 4   spectro.ujf-grenoble.fr (193.54.234.59)  2 ms  2 ms  3 ms
```

Ce mécanisme peut ainsi être utilisé pour connecter un site à 2 réseaux, un accès en back-up de l'autre, ce qui est obligatoire lorsque l'on dispose de 2 réseaux mais aussi en phase de migration d'un réseau vers un autre.