Utilisation haut débit des nouvelles infrastructures réseaux de la recherche

Jérôme Bernier
Centre de Calcul de l'IN2P3/CNRS
CCIN2P3 – 29 bd du 11 Novembre 1918 – 69622 Villeurbanne Cedex
jerome.bernier@in2p3.fr

Résumé

Pour pouvoir répondre aux nouveaux projets de la recherche et à leurs besoins en très haut débit, de nouvelles infrastructures réseaux sont mises en place au niveau mondial. L'architecture projet de RENATER-4 permet, sur la base de fibres noires, de proposer des longueurs d'onde à 10Gb/s et de créer des réseaux virtuels privés à un projet. Le Centre de Calcul de l'IN2P3/CNRS utilise ces nouvelles technologies réseau à haut débit dans le cadre de plusieurs expériences.

Dans le cadre du projet LHC (Large Hadron Collider), le CCIN2P3 doit pouvoir supporter le transfert de données en continu à une vitesse de plusieurs gigabits par seconde et ceci 24 heures sur 24 et 7 jours sur 7 ! Pour répondre à ce besoin, plusieurs liaisons ethernet à 10Gb/s relient directement le CCIN2P3 au CERN et à d'autres centres de calcul européens.

Dans le cadre du projet IGTMD (Interopérabilité de Grille et Transfert Massif de Données) il est mis en place un circuit (concrètement 2x1Gb/s) entre le CCIN2P3 à Lyon et le FNAL à Chicago, afin de tester en grandeur nature les nouvelles implémentations des couches TCP qui permettront de s'affranchir des limites actuelles du protocole lors du transfert massif de données sur de longues distances.

Mots clefs

Haut débit, 10 gigabit ethernet, fibre optique, WDM, routage, TCP, RENATER, GEANT.

1 Introduction

De nombreux projets internationaux nécessitent aujourd'hui des taux de transfert de données sans commune mesure avec ceux que l'on a connus. Pour cela, de nouvelles structures réseaux sont nécessaires. Les dernières évolutions des architectures des Réseaux Nationaux de l'Enseignement et de la Recherche (NREN) sont basées sur la mise à disposition de fibres noires. L'éclairage de ces fibres avec diverses longueurs d'ondes permet de disposer de liens à très haut débit et pouvant être réservés à un projet de recherche.

Dans cet article, nous présenterons les besoins associés à deux projets internationaux que sont le LHC (Large Hadron Collider) et IGTMD (Interopérabilité de Grille et Transfert Massif de Données). Nous présenterons les solutions techniques mises en place par le Centre de Calcul de l'IN2P3[1], RENATER[2] et les autres organismes associés à ces projets. Nous parlerons des impacts sur l'organisation apportés par ces nouvelles architectures concernant la sécurité et la redondance. Nous évoquerons aussi les nécessaires optimisations à apporter tant en terme matériel que logiciel, pour pouvoir tirer pleinement partie de ces infrastructures.

2 LHC Large Hadron Collider

2.1 Description du projet

Dans le cadre du projet LHC[3], le CERN[4] prévoit de produire plus de 15 Peta octets (soit 15 000 Tera octets) par an de données. Pour traiter ces données une infrastructure informatique internationale de grille de calcul, appelée LCG[5] (LHC Computing Grid), est mise en place. Cette grille est structurée hiérarchiquement :

- un Tier0, le CERN, producteur des données réelles;
- douze Tier1s, centres de calcul importants, répartis dans le monde en Europe, Amérique et Asie sont chargés de la récupération, du traitement et de la mise à disposition de ces données;
- une centaine de Tier2s, chargés de l'analyse des données et de la simulation.

Un composant critique de cette infrastructure est le réseau privé LHCOPN[6] (LHC Optical Private Network). Ce réseau interconnecte le CERN et les onze autres Tier1s. Mis en place ces dernières années, il permet le transfert et l'échange des données à très haute vitesse. Il est basé principalement sur des liaisons à 10Gb/s fournies par les NRENs et GEANT[7] (réseau d'interconnexion des réseaux de la recherche Européens).

[1] http://cc.in2p3.fr
[2] http://www.renater.fr
[3] http://lhc.web.cern.ch/lhc/
[4] http://cern.ch/
[5] http://lcg.web.cern/lcg/
[6] http://lhcopn.web.cern.ch/lhcopn/
[7] http://www.geant2.net

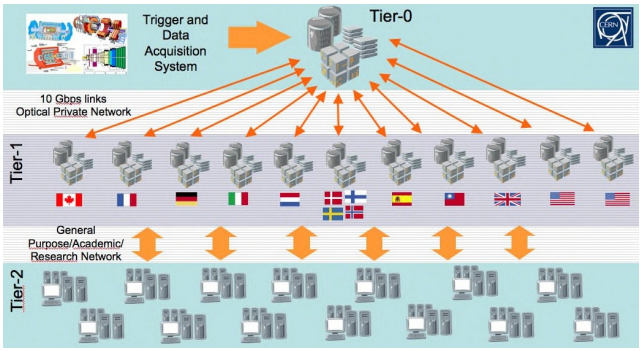


Figure 1 - *Grille hiérarchique LCG*

2.2 Le Tier1 du Centre de Calcul de l'IN2P3/CNRS

Le Centre de Calcul de l'IN2P3/CNRS est un centre de ressource Tier1 pour l'expérience LHC. Il doit donc fournir une certaine puissance de calcul et s'occuper de l'archivage d'une partie des données. Pour pouvoir répondre aux besoins de cette expérience, il est nécessaire de doubler tous les ans sa puissance de calcul et son volume de stockage disque durant la période 2006-2010. Pour donner un ordre de grandeur, ceci a nécessité en 2007 d'installer un millier de serveurs de calcul (bi-processeurs quadri-coeurs) et plus de 2 Peta octets de disques (avec la centaine de serveurs de fichiers associés).

Au niveau du réseau, le CCIN2P3 doit pouvoir rapatrier les données du CERN à une vitesse de plusieurs gigabit par seconde, et ceci 24 heures sur 24 et 7 jours sur 7. Des échanges de données entre Tier1s sont aussi nécessaires. De plus, non seulement les Tier2s Français sont raccrochés au Tier1 du CCIN2P3, mais aussi certains sites étrangers comme ceux de la Belgique, de la Roumanie, de la Chine ou du Japon. Ceci nécessite donc d'avoir aussi une très bonne connectivité avec eux.

En partant de l'estimation des transferts de données fournie par l'expérience LHC, en doublant celle-ci pour pouvoir soutenir les pics de transfert (par exemple après une certaine période d'indisponibilité), voici ci-dessous l'infrastructure réseau que nous devons mettre en place entre le CCIN2P3 et les divers autres intervenants de cette expérience.

besoins infrastructure réseau LHC	CCIN2P3
CERN	4 Gb/s
Tier1s	5 Gb/s
Tier2s Français	1 Gb/s
Tier2s Etrangers	3 Gb/s

2.3 Infrastructure réseau mise en place

Pour pouvoir répondre à ces besoins, l'infrastructure IP mutualisée est insuffisante et nous allons nous appuyer sur l'infrastructure projet de RENATER. Celle-ci, basée sur des fibres optiques noires louées, et grâce aux équipements optiques WDM installés, permet de pouvoir disposer de liens à 10Gb/s.

Figure 2 - *Infrastructure WDM de RENATER-4*

La mise à disposition de telles liaisons est aussi facilitée par le fait que le CCIN2P3 héberge le nœud Lyonnais de RENATER.

Pour la connexion Tier0-Tier1, RENATER fournit au CCIN2P3 un lien ethernet à 10Gb/s entre Lyon et le CERN. Cette liaison est réservée aux transferts du LHC. Pour des raisons de sécurité, seuls les sous-réseaux dédiés aux transferts de données peuvent utiliser ce lien. Pour le CCIN2P3, un seul sous-réseau est concerné. C'est dans ce sous-réseau que sont l'ensemble des serveurs de fichiers spécialisés dans le transfert des données. Ceci permet d'implémenter directement dans le matériel des règles de filtrage et de se protéger ainsi de l'extérieur.

Les autres Tier1s disposent aussi d'une liaison à 10Gb/s vers le Tier0, fournie par leur NREN pour la partie nationale et l'infrastructure fibre noire de GEANT pour arriver jusqu'au CERN. Le CCIN2P3 peut ainsi se servir de ces liens 10Gb/s pour les transferts entre Tier1s en transitant à travers le CERN.

Malheureusement, si les technologies WDM permettent de disposer de liaisons à très haut débit, elles sont beaucoup plus sensibles à la panne. Contrairement aux liens classiques SDH (Synchronous Digital Hierarchy) qui sont fournis aux clients par l'intermédiaire d'une boucle optique, une rupture de fibre entraîne la perte de la connexion. Pour

cela nous avons réfléchi à la mise en place d'un lien de secours. En collaboration avec DFN, qui a en charge le réseau Allemand de la recherche, RENATER a pu mettre en place une liaison 10Gb/s entre le CCIN2P3 et le Tier1 Allemand GRIDKA à Karlsruhe. Cette liaison chemine donc de Lyon à Paris, puis Strasbourg, Kehl et Karlsruhe. Elle est un exemple de l'utilisation de CBF (Cross Border Fiber) qui sont des fibres interconnectant des NRENs directement, sans passer par l'infrastructure transnationale de GEANT.

Ce lien CCIN2P3-GRIDKA permet donc de disposer d'un lien de secours en cas de coupure de notre liaison directe avec le CERN, à travers le lien GRIDKA-CERN. De plus ceci permet de s'échanger des données entre nos deux Tier1s. Nous utilisons le protocole de routage BGP pour gêrer dynamiquement le routage des sous-réseaux réservés aux transferts et uniquement de ceux-là.

Le LHCOPN (LHC Optical Private Network) ainsi mis en place est donc un réseau privé, réservé au transfert de données pour le LHC. Il interconnecte le CERN avec les Tier1s. L'utilisation de BGP et de liens directs entre ces derniers permet de sécuriser le réseau et d'optimiser les transferts entre Tier1s.

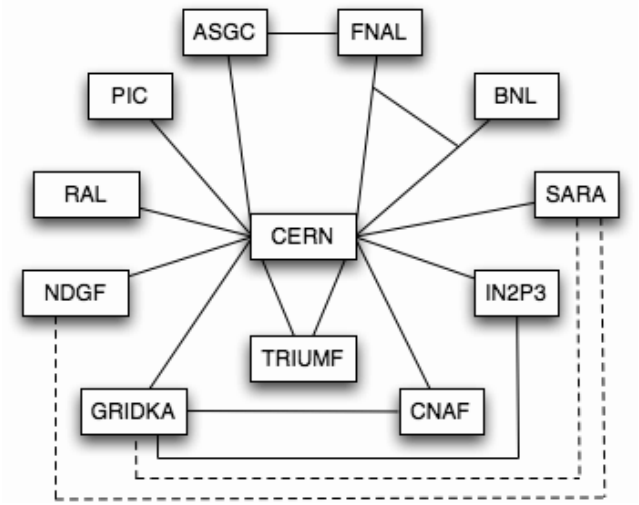


Figure 3 - *LHCOPN*

Le LHCOPN permet donc au CCIN2P3 de répondre aux besoins de transferts avec le Tier0 et les Tier1s.

Par contre, comme ce réseau est réservé, il ne faut pas que les annonces des routes des Tier1s soient utilisées par les autres sous-réseau du CCIN2P3. En effet si une machine sur un réseau classique veut atteindre une machine au CERN annoncé sur le LHCOPN, il serait routé sur le lien direct et serait automatiquement filtré. Pour résoudre cela, on peut faire du PBR (Policy Based Routing), aussi appelé routage par la source. Ainsi suivant d'où l'on provient, on utilise soit le lien direct, soit le routage par défaut sur RENATER. Une autre solution, celle mise en place actuellement, est d'avoir un deuxième routeur en sortie du site. Le premier routeur gère tous les réseaux non-LHCOPN et utilise la route par défaut fournie par RENATER. Le deuxième routeur interconnecte uniquement le sous-réseau LHCOPN et les deux liens 10Gb/s avec les Tier0-Tier1s. De plus il ne propage pas les annonces BGP.

Concernant les transferts avec les Tier2s associés, au vu du grand nombre de ceux-ci, il a été décidé d'utiliser l'infrastructure IP classique existante plutôt que de faire des liens privés. Malheureusement les besoins dépassent les capacités actuelles du réseau mutualisé mis en place par RENATER qui se base sur une épine dorsale composée de liens à 2,5Gb/s. De plus, une grande partie des besoins proviennent de sites étrangers qui passeront donc par le réseau GEANT dont l'épine dorsale est composée de liens à 10Gb/s avec un point de présence en France à Paris. RENATER a donc utilisé une longueur d'onde sur la fibre Lyon-Paris allumée en 10Gb/s pour connecter le CCIN2P3 au routeur de RENATER qui est directement en face du routeur de GEANT. Notre routeur de site est interconnecté en 10Gb/s sur le commutateur RENATER de Lyon. Sur ce lien, on a un peering BGP avec le routeur de RENATER à Lyon et un autre avec le routeur de Paris. Cette configuration permet donc de disposer en théorie d'une bande passante de 10Gb/s sur RENATER (et de partir après en 2,5Gb/s sur Clermont-Ferrand) ou de poursuivre directement en 10Gb/s sur GEANT. Là aussi, cette solution pose des problèmes au niveau du routage. En effet comment décider de partir sur le peering de Paris si on veut atteindre un Tier2 étranger ? Pour cela on fait donc du PBR et dans le cas où on veut atteindre une liste précise de réseaux on forcera le passage par Paris, sinon et par défaut on passera par le routeur RENATER de Lyon. Cette liste de réseaux concerne les Tier2s étrangers ainsi que les Tier2s Français qui sont au Nord de Paris.

La figure ci-dessous synthétise la connectivité réseau du CCIN2P3 pour essayer de présenter plus simplement cela.

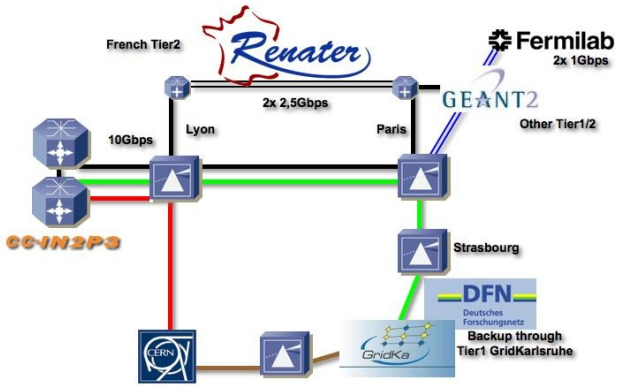


Figure 4 - *connectivité CCIN2P3*

2.4 *L'administration du LHCOPN*[8]

L'infrastructure du LHCOPN pose deux principaux problèmes pour son exploitation quotidienne : la supervision des équipements de niveau 2 et la coordination opérationnelle multi-domaines.

Au niveau organisationnel, deux structures sont impliquées:

- L'E2ECU (End to End Coordination Unit), entité de coordination sous la tutelle de GEANT responsable du bon fonctionnement des liens composant le LHCOPN. Elle est responsable de la détection de problème de liens et de la coordination des différents NRENs impliqués.

- L'ENOC[9] (EGEE Network Operating Center) qui est l'unité de support réseau du projet de grille de calcul EGEE[10] (Enabling Grid for E-scienceE). Elle est responsable de la détection de problème au niveau IP et de l'interface avec les utilisateurs de la grille. La grille LCG, dédiée a l'expérience LHC, est l'utilisateur majeur de EGEE.

Cette séparation est rendue nécessaire par le fait que des liens peuvent être utilisés par plusieurs projets, et que le client final doit pouvoir s'appuyer sur une unité bien définie pour être informé et pouvoir ouvrir un incident.

Ces structures s'appuient sur des outils comme perfSONAR[11] qui permet de collecter des métriques de bas niveau dans un environnement multi-domaine. Un lien étant souvent composé de multiples segments terminés par des éléments optiques et/ou de niveau ethernet, chacun géré par un entité administrative distincte. L'ENOC a développé des outils permettant d'évaluer l'état du réseau LHCOPN au niveau IP et BGP qui permettent de donner des informations au niveau des différents sites LCG (cf http://ccenoc.in2p3.fr).

2.5 *Le Tier2 de GRIF*[12]

La France possède un Tier2 un peu particulier: GRIF (Grille de Recherche d'Ile de France). Ce projet vise à mettre en place une ressource unique vue de la grille, tout en s'appuyant physiquement sur plusieurs laboratoires répartis en région parisienne. Ces laboratoires étant le LAL (IN2P3, Orsay), le DAPNIA (CEA, Saclay), l'IPNO (IN2P3, Orsay), le LLR (Ecole Polytechnique, Palaiseau) et le LPNHE (IN2P3, Jussieu).

Afin de s'assurer de la transparence au niveau de l'accès et de la gestion des données réparties sur des sites géographiquement lointains, une infrastructure réseau privée et à haut débit est mise en place. Ainsi des liens à 10Gb/s sont déployés entre ces laboratoires. Ces liens utilisent les mêmes techniques de longueur d'onde sur des fibres et sont rendus disponible grâce à divers intervenants: fibres noires privées (sur le campus de Paris-Sud), réseau projet Île-de-France de RENATER (sur les fibres noires reliant les points de présence Parisiens), réseau métropolitain SAPHIR (Réseau Haut Débit du Plateau de Saclay).

3 IGTMD Interopérabilité des Grilles de calcul et Transferts Massifs de Données

3.1 Description du projet

Le projet Franco-Américain IGTMD a pour but de réaliser concrètement l'intéropérabilité de la grille de calcul Européenne EGEE[13] (Enabling Grid for E-scienceE) avec la grille Américaine OSG[14] (Open Science Grid). Le projet s'attache particulièrement à la problématique du transfert de données très volumineuses sur de très longues distances. Il s'appuie sur deux centres de calcul, le CCIN2P3 à Lyon et le FNAL[15] (Fermi National Accelerator Laboratory) à Chicago. Un outil important pour cette expérience est la mise à disposition de deux liens gigabit ethernet entre ces deux sites. La mise en place de ces liaisons a nécessité la collaboration étroite entre les réseaux européens RENATER et GEANT et américains ESNET[16] et INTERNET2[17]. La difficulté principale fut ici de coordonner les efforts d'intervenants nombreux et distincts, utilisant de plus des techniques différentes. En effet, la fourniture de ces deux liens s'est faite au moyen d'équipements gigabit ethernet, 10 gigabits ethernet et SDH. Le plan de travail ci-dessous montre la complexité technique du cheminement.

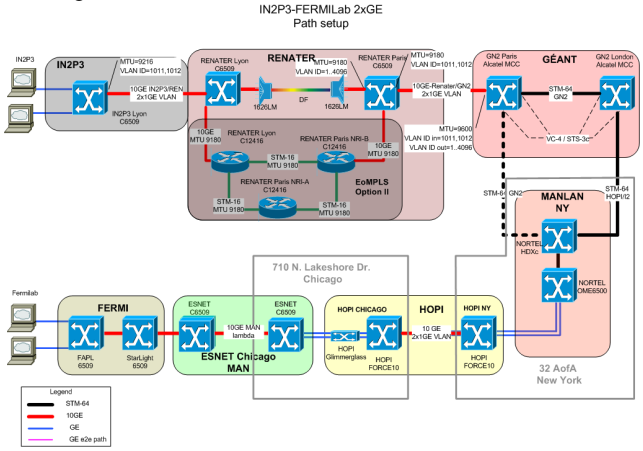


Figure 5 - Etude des liens pour le projet IGTMD

[8] Pour plus d'informations se référer à l'article de Guillaume Cessieux et Mathieu Goutelle JRES2007

[9] http://egee-sa2.web.cern.ch/egee-sa2/ENOC.html

[10] http://www.eu-egee.org/

[11] http://www.perfsonar.net/

[12] http://www.grif.fr

[13] http://www.eu-egee.org/

[14] http://www.opensciencegrid.org/

[15] http://www.fnal.gov/

[16] http://www.es.net/

[17] http://www.internet2.edu/

Les liaisons sont fournies aux extrémités sous la forme de deux VLANs sur une connexion 10 gigabit ethernet.

Ces deux liens privés gigabit, dédiés et indépendants, reliant Lyon à Chicago, permettent de tester en grandeur réel les transferts de données à haut débit sur de très longues distances. Nous avons décidé d'utiliser un des liens pour faire passer du trafic de production. Ceci permet de comparer le comportement de nos applications de test de transfert soit sur un lien vide, soit en compétition avec un trafic réel.

3.2 Transferts à très hauts débits

La quasi-totalité des applications de transferts de données se base sur le protocole TCP/IP. En effet la couche transport TCP s'assure que toutes les données de l'utilisateur sont bien reçues et dans le bon ordre, grâce à la numérotation de tous les paquets émis et un acquittement de ceux-ci par le récepteur. Si le protocole est bien décrit dans les RFCs[18], son implémentation est laissée à la discrétion des programmeurs systèmes. L'implémentation la plus répandue dans les systèmes Unix s'appelle RENO. Datant de 1990 elle utilise pour moduler son débit un algorithme appelé AIMD pour Additive Increase – Multiplicative Decrease. Celui-ci permet de faire varier le nombre de paquets transmis par l'émetteur suivant la bande passante disponible, en s'appuyant sur la détection des paquets perdus. Après une phase d'accroissement exponentielle du débit de l'émission des paquets appelée « Slowstart », on passe en cas de perte d'un paquet, à un mode appelé « Congestion Avoidance » qui consiste en une division par deux du débit d'émission puis à un accroissement linéaire de celui-ci (voir Figure 6). Ce mode de régulation n'est plus adapté aux réseaux à grande vitesse et à longue distance disponibles aujourd'hui. On appelle LFN (Long Fat Network) un réseau ayant un important débit associé à un temps de transit très long. Aujourd'hui, une liaison offrant un gigabit par seconde de débit sur une distance transatlantique rentre dans cette dénomination; ou bien une liaison 10 gigabits entre deux pays Européens. Pour pouvoir optimiser l'utilisation de ces réseaux il faut faire attention à de nombreux paramètres dont nous allons essayer de donner un panorama.

- Taille de la fenêtre TCP: c'est le nombre maximal d'octets que l'on peut envoyer avant d'attendre un acquittement de bonne réception. Avec l'implémentation standard, elle est de 64 kilooctets, c'est à dire qu'après avoir atteint ce volume, l'émetteur s'arrête d'envoyer des données sur le réseau, en attente de l'acquittement des premières ! Avec un réseau LFN, cela veut dire que l'on enverra un burst de données, puis que l'on passera la grande majorité du temps à attendre. En moyenne, on ne pourra donc pas dépasser quelques megabits par seconde de transfert sur une liaison transatlantique. Le RFC1323 introduit un moyen appelé "Window Scaling" qui permet de passer cette fenêtre jusqu'à une valeur proche de un gigaoctet. L'utilisation de cette possibilité doit être validée par les deux extrémités du transfert lors de la mise en place de la connexion TCP.

- Acquittement: par défaut, si un paquet est perdu, l'émetteur ne recevant pas d'acquittement de celui-ci va le réémettre; ainsi que tous les paquets suivants qu'il avait d'ailleurs déjà émis et qui sont probablement, eux, arrivés sans problèmes ! Le RFC 2018 permet de mettre en place un système appelé SACKS pour "Selective Acknowledgements" qui permet d'acquitter un ensemble de paquets disjoints, et donc de signaler les paquets manquants. L'émetteur n'a alors plus qu'à réémettre les paquets perdus et seulement ceux là. Là aussi, l'utilisation de cette possibilité doit être validée par les deux extrémités du transfert lors de la mise en place de la connexion TCP.

- BIC-TCP: avec une couche TCP standard, en cas de perte d'un paquet, on divise par deux le débit d'émission. Puis on augmente celui-ci très lentement et de façon linéaire. Ainsi, après la perte d'un paquet, il faudra environ une demi-heure pour atteindre de nouveau le débit maximal constaté sur une liaison transatlantique ! D'autres implémentations de la couche TCP sont étudiées pour remplacer cet algorithme AIMD. Les personnes intéressées pourront suivre les travaux du groupe de travail PFLDnet[19] (Protocols for Fast Long-Distance networks) qui étudie ces nouveaux développements. Par exemple, depuis le noyau Linux 2.6, l'implémentation BIC-TCP est le nouvel algorithme par défaut. BIC, pour Binary Increase Control, permet lors de la perte d'un paquet, de se rapprocher très rapidement de la vitesse maximale que l'on avait atteinte, puis de venir "tangenter" au plus près de celle-ci. Ainsi on optimise l'utilisation de la bande passante disponible. De plus, il n'est nécessaire d'implémenter ce type de nouveau comportement qu'au niveau de la couche TCP de l'émetteur.

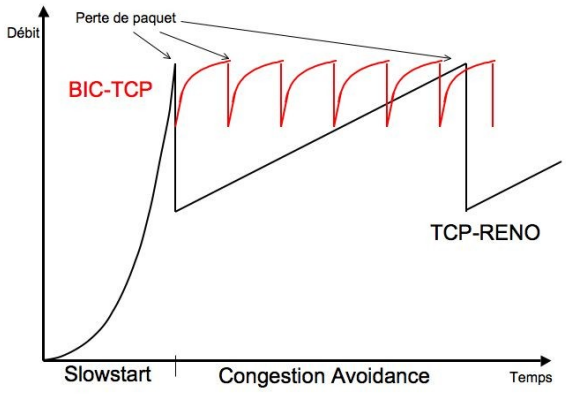


Figure 6 - Différentes implémentations de TCP

- Jumboframe: par défaut un paquet ethernet a une taille voisine de 1500 octets. L'utilisation d'une trame de taille supérieure, par exemple jusqu'a 9 kilooctets, permet

[18] http://www.ietf.org/rfc.html

[19] http://wil.cs.caltech.edu/pfldnet2007/

d'obtenir des débits supérieurs. Dans les faits, ceci est la conséquence que le traitement de plus grandes trames permet de minimiser le nombre d'interruptions système nécessaires. Actuellement, la puissance CPU n'est pas un problème, et comme utiliser cette solution demande qu'elle soit implémentée aussi bien par les extrémités responsables du transfert que par tous les commutateurs et routeurs intermédiaires, cette possibilité est très rarement utilisée.

- Applications multi-streams: si un flux TCP est limité par le protocole et son implémentation, il est envisageable d'utiliser plusieurs flux en parallèle. Le transfert massif de données demande la mise en place d'outils particuliers. Il a été développé de nombreuses applications qui permettent de transférer un fichier en utilisant simultanément plusieurs flux TCP en parallèle et ainsi d'optimiser l'utilisation des liaisons à haut débit.

Nos tests réseau ont montré que, pour un temps de transit d'environ 200ms avec les Etats-Unis, là où un noyau standard ne permet pas de dépasser 4Mb/s, un système optimisé permet d'atteindre 200Mb/s et un système utilisant BIC-TCP 400Mb/s ! Et ceci avec seulement un seul flux TCP. L'utilisation d'une application multi-stream permet de saturer une liaison gigabit transatlantique. Il est donc très intéressant que les systèmes d'exploitation récents utilisent de plus en plus par défaut ces nouvelles possibilités.

Mais il reste encore un large champ de développement avant de savoir utiliser pleinement les liaisons très haut débit. Par exemple, l'outil PSPacer[20] (Precise Software Spacer) insère des trames ethernet vides entre les trames contenant les données en cours de transfert. Ces trames vides sont jetées par le premier commutateur traversé et permettent d'avoir un trafic stable et sans à-coups. Avec cette méthode nous avons pu obtenir 900Mb/s de taux de transfert entre Lyon et le Japon avec une seule machine connectée en gigabit ethernet, et cela sur le réseau mutualisé RENATER/GEANT !

4 Infrastructure du CCIN2P3

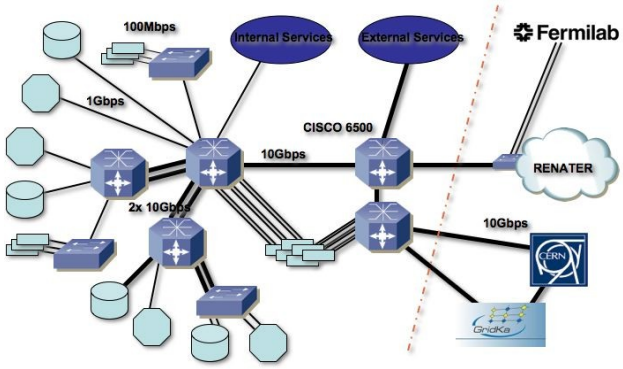
Pour conclure la figure ci-dessous présente l'infrastructure mise en place au CCIN2P3 pour pouvoir assurer les transferts de données à très haute vitesse.

Figure 7 - Infrastructure réseau du CCIN2P3

Les constituants sont :

- un routeur principal : il est connecté en 10Gb/s sur RENATER. Il a deux peerings BGP à Lyon et Paris et est chargé de tous les routages par défaut. Il accueille les 2 liaisons privées gigabit ethernet avec le FNAL;

- un routeur spécialisé pour les réseaux du LHCOPN: sur celui-ci arrivent les deux liaisons 10Gb/s vers le CERN et GRIDKA. Il est chargé du routage dans le réseau



privé LHCOPN et n'annonce qu'un sous-réseau du CCIN2P3. C'est sur ce routeur et dans ce sous-réseau que sont connectées les machines spécialisées dans le transfert de données;

- un réseau local dont l'armature est composée de un où plusieurs liens 10 gigabits ethernet;

- un ensemble de machines dédiées aux transferts: actuellement au nombre d'une trentaine, ces machines disposent de deux interfaces gigabit ethernet, une dans le réseau privé LHCOPN et l'autre vers nos systèmes de gestion hiérarchique de données. Ces machines gèrent un important volume disque, de l'ordre de 200 Tera octets, qui sert de tampon. Pour des raisons de performance et de compatibilité, elles utilisent encore principalement un système d'exploitation Unix propriétaire. Majoritairement sous SOLARIS, elles disposent des options de SACKS et de « Window Scaling » mais pas encore d'une couche TCP modifiée. Nous travaillons de plus en plus à intégrer des machines avec un système d'exploitation Linux pour pouvoir profiter des derniers perfectionnements réseau, mais aussi bien les performances liées au matériel que notre besoin d'avoir un service extrêmement stable font que ceci se fait très progressivement;

- des applications de transferts de données spécialisées: nous utilisons des applications qui permettent de gérer la taille de la fenêtre TCP et le nombre de flux utilisés en parallèle. De plus, comme les valeurs optimales de ces paramètres ne sont pas les mêmes pour des transferts avec les sites Européen ou avec des sites Américains, par exemple, il est aussi nécessaire de disposer d'une surcouche logicielle qui permet de mémoriser ceux-ci et des les utiliser à bon escient. Pour ceux que cela intéresse, voici une liste non exhaustive des applications utilisées: bbftp[21], SRB[22], iRODS[23], DCACHE[24], FTS[25]...

[20] http://www.gridmpi.org/gridtcp.en.jsp

[21] http://doc.in2p3.fr/bbftp/

[22] http://www.sdsc.edu/srb/index.php/

[23] http://irods.sdsc.edu/

[24] http://www.dcache.org/

[25] https://twiki.cern.ch/twiki/bin/view/EGEE/FTS

Nous voyons que la disponibilité de tuyaux à très haut débit n'est pas suffisant pour permettre de transférer de très gros volumes de données. En revanche, c'est une condition nécessaire et nous tenons à remercier RENATER pour la mise en place de ces liens, sans lesquels nous ne pourrions pas participer à ces projets internationaux.