

BBFTP, une alternative à FTP

Gilles Farrache, gilles.farrache@telindus.fr

Centre de Calcul de l'IN2P3/CNRS
26 Boulevard du 11 Novembre 1918
69622 Villeurbanne CEDEX

Telindus Rhône-Alpes
Parc Club Moulin à Vent
33, Rue Dr. Georges Lévy
69693 VENISSIEUX CEDEX

Résumé : Les centres de production de données de physique des hautes énergies se doivent aujourd'hui faire face au problème suivant : comment stocker et traiter des volumes de données se calculant en péta octets. Le modèle de calcul centralisé s'avérant mal approprié, il est nécessaire de s'appuyer sur des centres distribués où seront stockées, traitées et distribuées les données. Ce modèle exige alors des connexions réseaux à haut débit entre les divers centres ainsi que des outils permettant des transferts rapides.

Le but de cet article est de présenter le produit de transfert de fichier que j'ai développé pour satisfaire le besoin d'une expérience de physique (Babar) et qui depuis a été largement distribué dans le monde de la physique des hautes énergies.

1. Introduction

Dans le monde de la physique des hautes énergies (HEP) il existe un nombre restreint de centre de production de données, le CERN à Genève, DESY à Hambourg, l'accélérateur linéaire de Stanford et quelques autres. Les expériences devenant de plus en plus coûteuses et regroupant un nombre de plus en plus grand de physicien, le modèle de calcul consistant à mettre la puissance de calcul et les physiciens près du lieu de l'expérience est devenu aujourd'hui irréaliste et politiquement incorrect.

Le modèle retenu est celui de la grille de calcul, où le centre producteur de données alimentera des centres secondaires (Tiers 1) qui alimenteront alors des centres tertiaires (Tiers 2), et ainsi de suite. Dans un monde idéal, l'utilisateur final ne saurait pas où se trouvent les données, ni même où il prend la puissance de calcul pour faire son analyse. Avant de pouvoir atteindre cet état, il faut construire les premières briques dont l'une est celle qui permet de transférer les données d'un centre à un autre.

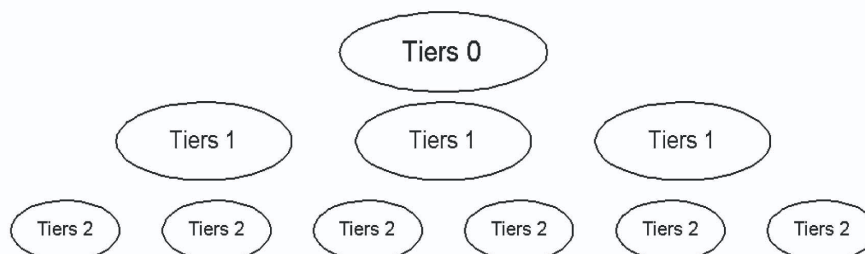


Figure 1: Hiérarchisation des centres

2. Le transfert à l'aide de bandes

Dans les expériences LEP (au CERN) la majorité des transferts de données entre Lyon et Genève se faisaient à l'aide de bandes magnétiques. Le fait de disposer du même type de robot au CERN et au centre de calcul de l'In2p3 a grandement facilité cet échange. Mais avec la fin des expériences LEP les physiciens français se sont tournés vers d'autres expériences situées aux Etats-unis comme Babar à Stanford (Californie) ou D0 au laboratoire Fermi de Chicago. Il devenait alors très difficile pour tous ces centres de se mettre d'accord sur un type de matériel de stockage de masse.

Un premier essai de transfert de données à l'aide de bandes magnétiques a été fait entre Stanford et Lyon, il s'est avéré très difficile à mettre en œuvre pour des raisons de disponibilité de main d'œuvre et de logiciel de transfert sur bandes. De plus le délai entre la disponibilité sur le Tiers 0 (Stanford) et sur le Tiers 1 (Lyon) était très important (plus d'une semaine).

3. Le transfert par le réseau à l'aide d'outils standard

3.1. La configuration réseau en 1999

En 1999 la configuration réseau entre le centre de calcul de l'in2p3 à Lyon et le SLAC (Stanford Linear Accelerator Center) était la suivante :

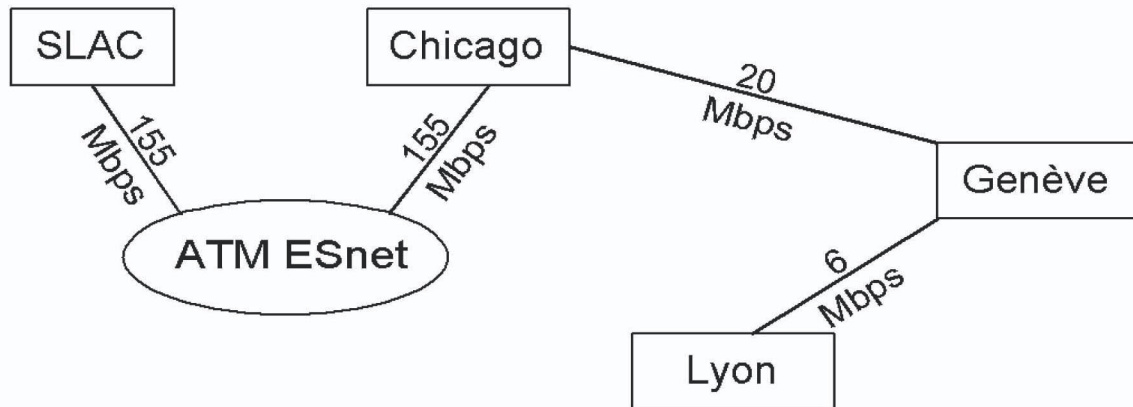


Figure 2 : Configuration réseau en 1999

3.2 Les tests de transfert à l'aide de ftp

Des tests de transfert ont été faits pour voir si cette méthode pouvait remplacer la méthode de transfert par bande. Les résultats ont été relativement décevants. En effet comme le montre les figures 3 et 4 sur un transfert, le débit atteint est de l'ordre de 80 Koctets par seconde, et si l'on multiplie les flux, on obtient un débit cumulé de 450 Koctets par seconde pour 15 flux, chacun d'eux ayant un débit de 30 Koctets par seconde.

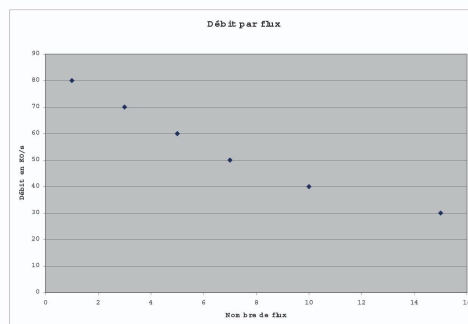


Figure 3: Débit par flux

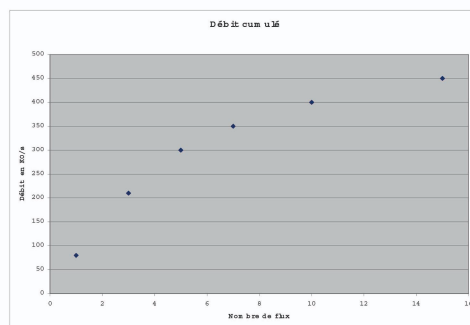


Figure 4 : Débit cumulé

Hélas, la taille des fichiers de l'expérience Babar est de l'ordre de deux Gigaoctets et donc le transfert dans le cas d'un flux unique durait plus de sept heures, et dans le cas de quinze transferts en parallèle de plus de dix-neuf heures ! Par contre si l'on arrivait à transférer le fichier à l'aide de quinze flux parallèles on pouvait espérer une durée de transfert d'environ une heure trente.

Un autre problème se posait alors, comment automatiser, sécuriser et surtout rendre moins vulnérable au piratage ces transferts. En effet si les données ne sont pas confidentielles les mots de passe et les noms de comptes circulent en clair sur le réseau.

4. Les principes de base de BBFTP

4.1 Les pré-requis

BBFTP devait permettre de transférer le plus rapidement possible entre SLAC et Lyon des fichiers de plus de deux Gigaoctets, avec reprise d'erreur et sans transmission du mot de passe en clair sur le réseau. Ce logiciel devait, de plus, s'exécuter sur pratiquement toutes les plates-formes Unix utilisées dans la collaboration Babar (SunOS, Linux, DEC Unix, AIX, HP-UX...).

4.2 Les choix

Au regard des distances entre Lyon et la Californie il a semblé judicieux d'utiliser des fenêtres TCP supérieures à 64 Koctets (RFC 1323). De plus au vu des tests effectués avec ftp, utiliser plusieurs flux TCP pour ce transfert pourrait améliorer les performances. Enfin comme les données de l'expérience sont compressibles avec un facteur deux, pourquoi ne pas les compresser au vol ?

4.3 L'architecture

Dans le cas de la méthode de connexion standard, c'est une architecture classique client serveur. Le serveur écoute sur un port donné, après une connexion et une authentification, établit un dialogue sur une connexion de contrôle. Lorsque le client et le serveur se sont mis d'accord sur un transfert (nombre de flux, options ...), ils démarrent des processus qui vont réellement effectuer ce transfert, chacun des processus ne transférant qu'une tranche de fichier. Si cela a été requis chaque processus compressera (pour l'envoyeur) et décompressera (pour le receveur) au vol.

Par exemple dans le cas d'un fichier de 100 octets à transférer à l'aide de 10 flux chaque processus transfèrera 10 octets, le premier processus transfèrera de l'octet 1 à l'octet 9, le second de l'octet 10 à l'octet 19

5. Les performances obtenues

Sur la configuration réseau de 1999 les débits utiles étaient compris entre 7 et 20 Mbps pour un fichier. Pour un fichier standard de Babar, le débit moyen était de 12 Mbps en utilisant la compression.

Hélas il s'est avéré, lors du passage de la liaison Lyon/Genève de 6 Mbps à 34 Mbps, que pour des liaisons de débit supérieur ou égal à 34 Mbps la compression se révélait pénalisante. Avec compression sur un fichier standard de Babar le débit utilisateur était de l'ordre de 20 Mbps alors que sans compression, il atteignait les 28 Mbps.

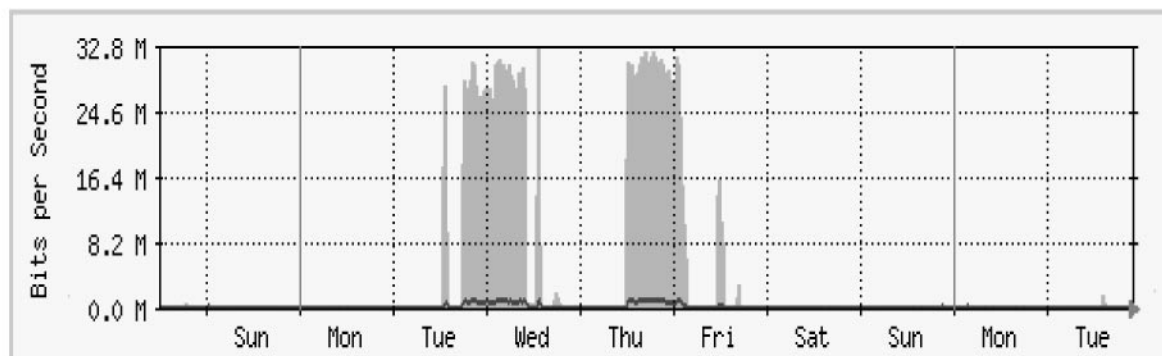


Figure 5 : Saturation de la liaison Lyon/Genève à l'aide de BBFTP

6. La sécurisation de la connexion

6.1 La méthode standard

Le serveur doit écouter sur un port, et, pour être utilisé par tous les utilisateurs, être exécuté en mode privilégié.

Lorsque le serveur reçoit une connexion, il génère une paire de clefs RSA et envoie sur la connexion la clef publique. Le client sur réception de cette clef, crypte le nom de compte et le mot de passe avec cette clef avant de les transmettre. Le serveur n'a plus qu'à décrypter ces données à l'aide de sa clef privée, puis à effectuer l'authentification.

Cet échange étant le seul qui nécessite réellement une protection les clefs ne sont plus utilisées lors des échanges suivants (tant que la connexion de contrôle reste établie).

6.2 La méthode à travers SSH

Cette méthode, basée sur une idée de Tim Adye, consiste à démarrer un processus SSH, sur la machine cliente, qui va s'occuper de toute la partie contrôle d'identification, puis qui va démarrer le serveur sur la machine distante. En effet avec cette méthode, le serveur n'écoute plus sur un port, il suffit que l'exécutable du serveur se trouve sur la machine distante et le client va la démarrer grâce à la connexion SSH.

Cette méthode présente l'avantage de ne rien avoir à installer en mode privilégié sur la machine distante.

6.3 Les commandes possibles

BBFTP étant un outil pour faire des transferts massifs entre des points relativement bien déterminés, ce produit ne s'utilise pas comme ftp en interactif. L'utilisateur doit construire un fichier de commandes. Ces commandes sont de deux sortes, les commandes d'action (get, put, mget, mput, cd, lcd, mkdir) ou des commandes de comportement (utilise une fenêtre TCP de xx Koctets, conserve la date de création du fichier...).

Le but de cet article n'étant pas de détailler toutes les possibilités du produit, le lecteur, s'il est intéressé, se reportera à l'URL <http://ccweb.in2p3.fr/bbftp/>

8. Conclusion

Avec l'arrivée des hauts débits 34 Mbps, 155 Mbps et plus on s'aperçoit que les applications réseaux sont, en réseau longue distance, incapables d'utiliser toute la bande passante offerte. Elles sont parfaites quand elles sont utilisées par des milliers d'utilisateurs qui vont se partager la bande passante équitablement.

La multiplication des flux TCP pour une application, si elle semble capable d'être efficace jusqu'à 155 Mbps, n'a jamais été testée sur du Gigabit (les machines capables de soutenir des débits réels de l'ordre du Gigabit étant encore rare).

Un axe de recherche intéressant serait de transformer la couche transport de BBFTP qui s'appuie sur TCP pour qu'elle utilise UDP, ce qui lui permettrait peut-être, d'être plus efficace.