Estimation du débit dans une grille de calcul

Carlos BARRIOS-HERNÁNDEZ, Yves DENNEULIN et Michel RIVEILL*

Laboratoire d'Informatique de Grenoble, Projet Mescal. 51 Av. Jean Kuntzmann 38330 Montbonnot-St Martin
Laboratoire I3S, Projet Rainbow. 930 route des Colles, BP 145, 06903 Sophia Antipolis CEDEX France

1. Introduction

Le transfert des données est un processus critique qui influe considérablement sur la performance des applications pour peu que celles-ci utilisent d'important flux de données. Pour caractériser le transfert de données en fonction des principaux éléments architecturaux de la plate-forme, deux approches sont fréquemment utilisées.
La première consiste à simuler dans un environnement contrôlé le fonctionnement supposé de l'application et la seconde utilise une plate-forme monitorée pour observer et mesurer le comportement effectif de l'application. L'objectif de ces travaux est de construire des modèles de comportement qui peuvent être ultérieurement utilisé pour élaborer des stratégies d'allocation de ressources ou planifier des ordonnancements de travaux.
Notre travail vise à compléter les travaux existant pour mieux caractériser le transfert de données en fonction des principaux éléments architecturaux de la plate-forme mais aussi de l'application, en profitant la seconde approche.
Ce document présente les travaux expérimentaux que nous avons réalisé ainsi que les informations pertinentes que nous en avons retiré en vue d'élaborer un modèle plus complet permettant de caractériser les fonctions de transferts sur une grille ou sur un réseau de cluster

2. Expérimentation

Nous définissons une infrastructure de type grille comme un ensemble de plates-formes locales de type grappe interconnectées par un réseau de communication spécifique. Une grappe étant constituée d'un ensemble de processeurs interconnectées à l'aide d'un switch. Le switch étant le point de connection de la grappe à la grille.
Le transfert de données entre deux noeux de la grille sollicite différentes ressources puisque la communication de bout en bout emprunte le chemin suivant : noeud A-lien grappe A-switch A-lien grille-switch B-lien grappe B-noeud B. Chaque étape du chemin a des propriétés différentes caractérisées par la nature des liens, la nature des équipements d'interconnexion et les différents protocoles utilisés.
Les protocoles de transfert utilisés, découpent un message de taille importante en plusieurs paquets.

2.1. Description

Nous proposons de réaliser un transfert massif de données entre deux ensembles de noeuds. Les tests ont été réalisés sur des grappes usuelles de Grid5000 dans un environnement réservé. Les expériences, qui ont fait varier la quantité de données mais aussi le nombre de processeurs impliqués dans le transfert ont été reproduites en utilisant des paires de noeuds internes à une grappe pour observer les transferts intra-grappe, et des paires de noeuds appartenant à deux grappes différentes pour observer les transferts sur la grille. *Nous avons mesuré lors de chaque expérience, pour chaque lien utilisé, le gap et la latence*.
Les grappes disponible ont des caractéristiques proches ce qui nous a permis de mettre en évidence des différences de comportement liées à la manière dont sont interconnectés les processeurs. Chaque grappe utilisée est constituée des mêmes processeurs : IBM eServer 326m constitué de 2 AMD Opteron 246 à 2 Ghz. *GDX* possède 186 noeuds et *Grillon* possède 178 noeuds.
Les interconnexions internes se font à l'aide d'un réseau Gb Ethernet via un switch qui a les mêmes caractéristiques et les liaisons entre les grappes sont fournies par le réseau Renater à 10 Gb/s pour Grid5000.

2.2. Résultats

Le graphique ci-dessus représente l'évolution du gap (somme des temps entre deux paquets consécutifs lors du transfert d'un message) en fonction du volume transféré et du nombre de paires impliquées dans l'échange. L'interval de confience sur les mésures utilisé est de 90 %.

3. Estimation et analyse du débit

Depuis le modèle LogGP on sait que les caractéristiques d'un réseau N qui permet le transfert d'un

* barrios@imag.fr, yves.denneulin@imag.fr, riveill@unice.fr

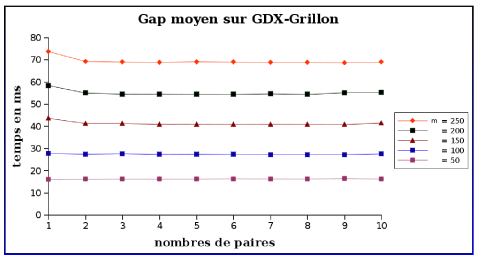

FIG. 1 – Evolution du gap lors de transferts sur la grille

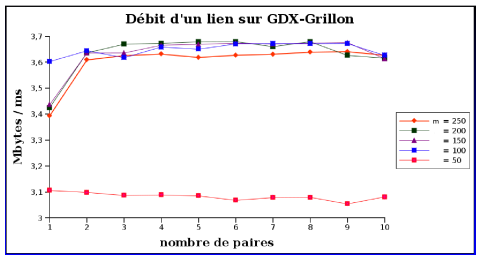

FIG. 2 – Evolution du débit d'un lien

message de m bytes entre P processeurs peut être décrit par l'ensemble des paramètres :

$$N(m)_P = (L(m), O(m), g(m), P) \quad (1)$$

L est la latence, O la moyenne entre le temps d'envoie et de réception du message (overhaed), g le gap pour l'envoie des paquets constituants un message.

A partir des expériences réalisées et de ce modèle élaboré pour des paquets TCP ou UDP, nous proposons de l'adapter à l'échange de grands volumes de données effectués simultanéement sur plusieurs liens.

Nous pouvons en particulier, calculer le débit moyen de chaque lien en fonction du nombre de paires impliquées dans le transfert.

En utilisant la relation qui existe entre la quantité des données transférées m et la somme des temps d'attente entre paquets d'un même message $g(m)$, on peut calcule le débit de chaque lien par la formule :

$$B = \frac{m}{g(m)} \quad (2)$$

La figure suivante, représente l'évolution du débit moyen de chaque lien en fonction du nombre de paires impliquées dans l'échange :

4. Discussion

Les résultats expérimentaux obtenus permettent d'analyser la performance de communication, dans le cadre du transfert haut débit intensif. L'analyse conduit à faire évaluer le modèle analytique, dans la manière d'utiliser le gap par rapport à la taille des messages, d'introduire une nouvelle manière d'utiliser l'overhead dans des architectures hétérogènes, d'interpréter la latence pour des échanges intensifs afin de mieux prédire la performance du transfert de bout en bout pour des échanges volumineux de données impliquants des nombreux nodes.

Nous tenons à attirer l'attention sur la difficulté d'obtenir des mesures fiables. La multiplicité des noeuds impliqués dans le transfert impose un cadre strict pour que les variations des valeurs de chaque paramètre permettent bien d'idendifier les éléments du modèle en cours de construction.

Bibliographie

1. Alexandrov, A., Ionescu, M., Schauser, K. et Scheiman, C. *LogGP : Incorporating Long Messages into the LogP Model*. Proceedings in 7th Annual ACM Syposium on Parallel Algorithms and Architectures, Santa Barbara, California, U.S.A. 1995.
2. Barchet-Estefanel, L., *LaPIe : Communications Collectives Adaptées aux Grilles de Calcul*. PhD. Thesis, Institut National Polytechnique de Grenoble, Grenoble, France, 2005.
3. Foster, I. et Kesselman, C. *The Grid 2 : Blueprint for a Future Computing Infrastructure*. Morgan Kaufmann Publishers, U.S.A. 2004.
4. GGF Network Measurements Working Group. *A Hierarchy of Network Performance Characteristics for Grid Applications and Services*. Internet Document, http ://nmwg.internet2.edu/docs/nmwg-measurements-v14.pdf . 2003.
5. Grid 5000 Project http ://www.grid5000.fr
6. Kielmann, T., Bal, H., et Verstoep, K. *Fast Measurement of LogP Parameters for Message Passing Platforms*. Lecture Notes In Computer Science ; Vol. 1800, Proceedings of the 15th IPDPS 2000 Workshops on Parallel and Distributed Processing, Springer - Verlag, U.K., 2000.
7. R. Jain, *The Art of Computer Systems Performance Analysis : Techniques for Experimental Design, Measurement, Simulation, and Modeling* Wiley- Interscience, New York, NY, April 1991, ISBN :0471503361.