Évaluation multicritère de topologies réseau pour le calcul haute performance

Matthieu Perotin, Jean-Olivier Gerphagnon

Atos Bull, HPC, 38160 Échirolles, France
{matthieu.perotin,jean-olivier.gerphagnon}@atos.net

Mots-clés : *aide à la décision, optimisation multicritère, topologies réseau, interconnect, calcul haute performance*

1 Introduction

Le domaine des *interconnects* haute performance est particulièrement dynamique aujourd'hui. Si la technologie Infiniband a dominé le marché pendant plusieurs années, la concurrence s'est intensifiée en 2015 avec l'annonce de l'arrivée des technologies Bull eXascale Interconnect (BXI) [2] et d'Intel Omni-Path (OPA) [1]. Malgré une bataille industrielle portant sur les technologies, les topologies réseau n'évoluent pratiquement pas, et ce malgré une littérature abondante. Cet article propose de se focaliser sur les aspects topologiques seuls en fournissant une démarche d'aide à la décision portant sur trois critères d'évaluation.

2 Topologies pour le HPC

De nombreuses topologies ont été proposées afin d'interconnecter les équipements composant un super-calculateur. Classiquement, les topologies en étoile utilisés par ethernet, sont exclues car elles créent de trop nombreux points de congestion. À chaque instant, un très grand nombre de communications parallèles sont émises sur un super-calculateur. Les réseaux en *mesh* ou en tore ont été historiquement utilisés. Néanmoins, au sein des ordinateurs les plus puissants de la planète [1], ce sont les topologies de type *fat-tree* [4],[5] qui sont les plus répendues.

Le *fat-tree* canonique a une bande passante de bissection relative (BBW) de 1. Cela signifie que quel que soit l'ensemble de paires de nœuds communiquant les uns avec les autres, il existe un routage permettant à ces communications de se réaliser sans congestion. Le *fat-tree* est ainsi perçu comme une topologie « performante ». Pour autant, elle est particulièrement onéreuse. Plusieurs nouvelles propositions ont été formulées, telles le DragonFly [3].

La réponse à un appel d'offre est un problème complexe, qui nécessite notamment de déterminer un nombre de nœuds ainsi que la façon de les interconnecter. La puissance en FLOPs qu'un calculateur doit être capable d'atteindre, duquel dérive directement le nombre de nœuds (ie. du nombre et du type de cœurs de calcul), est souvent exprimée sous la forme d'une borne minimale à atteindre. La puissance du calculateur étant utilisée comme élément de comparaison entre les offres, il peut être considéré comme un critère à maximiser. Cet article se concentre sur la partie topologie, où en utilisant le nombre de nœuds comme une contrainte, il s'agit d'évaluer la pertinence d'une topologie à l'aide des critères suivants :

— **Le prix de la topologie**, à minimiser, est une combinaison linéaire du nombre de nœuds, du nombre de switches et du nombre de liens optiques (câbles longs). Lorsque l'on souhaite comparer des topologies ayant un nombre de nœuds différents, il est pertinent de le ramener au prix par nœud ;

1. https ://www.top500.org/

— **La performance**, à maximiser, est évaluée par la bande passante de bissection ;
— **La résilience**, à maximiser, est généralement évaluée par le nombre de routes qui existent entre deux destinations.

3 Évaluation d'une topologie

Minimiser le prix, tout en maximisant performance et résilience est un problème multicritère à l'énoncé classique. Pour ce qui concerne les topologies HPC, certains critères supportent bien une ϵ-contrainte. C'est par exemple le cas de la performance, que l'on ne souhaite généralement pas voir descendre en dessous de 50%. Cette contrainte est cependant beaucoup trop impactante et mène à la necessité de préciser le problème, en distingant :

— Les îlots : sous-ensemble de la topologie, contenant des nœuds à une distance topologique maximale de 3, et devant avoir une BBW spécifique (souvent 100%) ;
— La topologie : interconnexion des îlots, ne devant rarement avoir plus de 50% de BBW (hormis cas particuliers).

Une fois l'ϵ-contrainte appliquée, les solutions sont ordonnées par ordre lexicographique.

4 Conclusions et perspectives

Dans le cadre des réponses aux appels d'offre concernant les super-calculateurs, le choix de la topologie utilisée afin d'interconnecter les équipements est un facteur clé du prix et des performances de l'ensemble. Ce choix, loin d'être trivial, se ramène à un problème d'optimisation multicritère, résolu par l'ϵ-contrainte et ordre lexicographique. Néanmoins, un raisonnement sur le graphe topologique uniquement, fait abstraction de plusieurs éléments contribuant à la performance atteinte par une application sur un cluster : la confrontation simultanée de l'algorithme de routage, du placement de l'application sur les nœuds, ainsi que les patterns de communications utilisés par l'application. L'élargissement du champs de cette étude en prenant en compte ces nouvelles perspectives est en cours.

Références

[1] Mark S. Birrittella, Mark Debbage, Ram Huggahalli, James Kunz, Tom Lovett, Todd Rimmer, Keith D. Underwood, and Robert C. Zak. Intel® omni-path architecture : Enabling scalable, high performance fabrics. In *23rd IEEE Annual Symposium on High-Performance Interconnects, HOTI 2015, Santa Clara, CA, USA, August 26-28, 2015*, pages 1–9, 2015.

[2] Saïd Derradji, Thibaut Palfer-Sollier, Jean-Pierre Panziera, Axel Poudes, and François Wellenreiter Atos. The bxi interconnect architecture. In *Proceedings of the 2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*, HOTI '15, pages 18–25, Washington, DC, USA, 2015. IEEE Computer Society.

[3] John Kim, Wiliam J. Dally, Steve Scott, and Dennis Abts. Technology-driven, highly-scalable dragonfly topology. In *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ISCA '08, pages 77–88, Washington, DC, USA, 2008. IEEE Computer Society.

[4] Charles E. Leiserson. Fat-trees : Universal networks for hardware-efficient supercomputing. *IEEE Trans. Comput.*, 34(10) :892–901, October 1985.

[5] Eitan Zahavi. D-mod-k routing providing non-blocking traffic for shift permutations on real life fat trees. Technical report, Technion Israel Institute of Technology, 2010.