Evaluation et modélisation des communications concurrentes sur cluster HPC

Jérôme VIENNE, Maxime MARTINASSO

Laboratoire LIG-ID(UMR 5132), Grenoble, France.
BULL - HPC Échirolles, France.

3 Juin 2008

Problématique

Environnement

- Multiplication des cores au sein d'un noeud $\Rightarrow$ partage des composants entre les différentes requêtes.
- Comportement de partage des ressources : difficiles à interpréter et à prédire.

Ressource réseau

- Accès concurrent (congestion) $\Rightarrow$ Perte de performance
- Modèles existants inadaptés ou trop complexe à mettre en oeuvre

Objectif

- Développer une méthodologie afin de pouvoir modéliser le comportement de ces accés concurrents

Objectif général

Contexte BULL

- Appel d'offres
  - Ensemble de benchmarks
  - Contraintes techniques
- Elaboration de solutions de grappes

Objectif

- Déduire une architecture matérielle performante répondant à la demande
- Outils efficaces d'évaluation de performances pour l'aide au dimensionnement

$\Rightarrow$ Proposer des solutions qui ne soit ni sous-dimensionnée ni sur-dimensionnée par rapport aux besoins

Plan

LogP

Le modèle LogP est le modèle d'origine dont découlent les autres modèles de la famille LogP. Le modèle LogP définit quatre paramètres comme suit :

- L (Latency) : qui correspond à la latence réseau ;
- o (overhead) : le coût logiciel induit par le mécanisme de communication, cad la traversée des couches logicielles et préparation du message ;
- g (gap) : le temps intrinsèque entre deux envois ou réceptions de paquets ;
- P (Processors) : le nombre de processus mis en jeu.

Gigabit Ethernet

Caractéristiques

- Full duplex
- Latence : 10-20 μs
- Bande passante : 125-125 Mo/s

Myrinet

Caractéristiques

- Full duplex, RDMA
- Latence : 2 μs
- Bande passante : 500-500 Mo/s

Infiniband

Caractéristiques

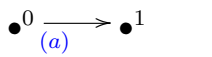
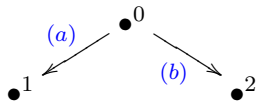
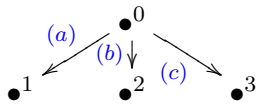
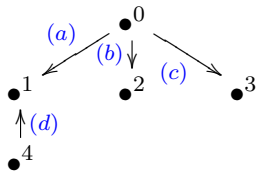
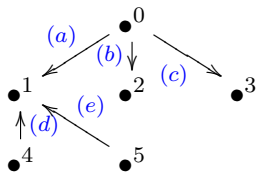
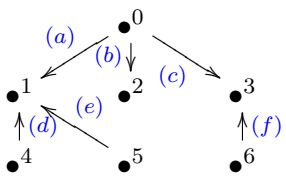
- Full duplex, RDMA
- Latence : 1,4 - 3 μs
- Bande passante (Pci Express Gen1, Infinihost III) : 1500 Mo/s U, 2700 Mo/s B
- Bande passante (Pci Express Gen2, ConnectX DDR x2) : 3600 Mo/s U, 7200 Mo/s B

Notion de pénalité

Définition

Nous appelons pénalité le rapport entre le temps d'une communication soumise à aucune concurrence avec les temps de communications soumises, quant à elles, à un contexte concurrent.

Impact de la contention suivant l'interconnect

Schema de Communication	Interconnect			
	Gigabit Eth.	Myrinet	Infiniband InfinihostIII	Infiniband ConnectX
	a = 1	a = 1	a =1	a = 1
	a = 1.5 b = 1.5	a =1.9 b = 1.9	a = 1.725 b = 1.725	a = 1.675 b = 1.675
	a = 2.25 b = 2.25 c = 2.25	a = 2.8 b = 2.8 c = 2.8	a = 2.61 b = 2.61 c = 2.61	a = 2.505 b = 2.505 c = 2.505
	a = 2.15 b = 2.15 c = 2.15 d = 1.15	a = 2.8 b = 2.8 c = 2.8 d = 1.45	a = 2.61 b = 2.61 c = 2.61 d = 1.14	a = 3.325 b = 3.305 c = 3.305 d = 1.135
	a = 4.4 b = 2.6 c = 2.6 d = 2.6 e = 2.6	a = 4.4 b = 4.2 c = 4.2 d = 2.5 e= 2.5	a = 3.66 b = 3.66 c = 3.66 d = 2.035 e = 2.035	a = 4.93 b = 4.93 c = 4.93 d = 2.5 e = 1.665
	a = 4.4 b = 2.0 c = 3.3 d = 2.6 e = 2.6 f = 1.4	a = 4.5 b = 4.5 c = 4.5 d = 2.5 e= 2.5 f = 1.3	a = 3.935 b = 3.935 c = 3.935 d = 1.995 e = 1.995 f = 1.01	a = 4.32 b = 4.33 c = 4.305 d = 2.5 e = 1.67 f = 2.715

Caractérisation des besoins applicatifs

- Pouvoir déterminer la quantité de ressource nécessaire de maniére temporelle. La dynamique des phénoménes concurrents ne peut pas être négligée pour obtenir des prédictions précises
- Caractérisation doit être indépendante de toute mesures spécifiques à une architectures matérielles. Doit être réalisé de maniére logique.

Approche par décomposition en séquences d'événements (ex. MPI : source, destination et taille)

Projection des besoins applicatifs sur une architecture abstraite

- Importante car va déterminer la façon dont les éléments vont intéragir entre eux.
- Source de la concurrences sur les ressources.
- Consiste à définir un placement des tâches sur les noeuds de la grappe abstraite

La phase de projection consiste donc à transformer les besoins logiques d'une application en besoins physiques propre à une architecture.

Prédiction des critéres de performance

- La définition de modéles prédictifs permet de déterminer les performances des applications sur une grappe donnée.
- Possède la contrainte d'être le moins dépendant possible de paramétres mesurés sur une architecture réelle

Ces modéles doivent avoir une erreur de prédiction faible.
Dans une décomposition par séquence d'événements, la prédiction de la durée d'un événement détermine la date de départ de l'événement suivant.
Cette problématique intervient dans le cadre de la propagation de l'erreur au sein d'une séquence.

Vue d'ensemble

Ensemble d'applications

Décomposition événementielle

Grappe de calcul

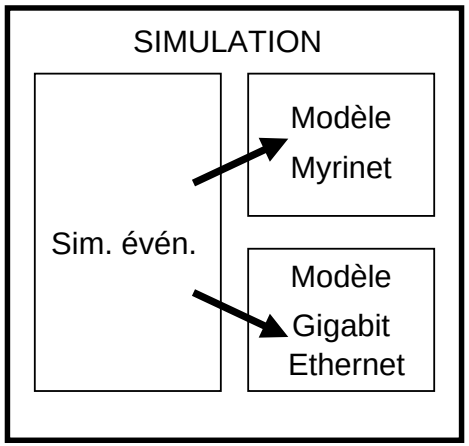
Nombre de noeuds
Nombre de processeurs par noeud

Placement

Round–robin
Utilisateur

Modèle réseau

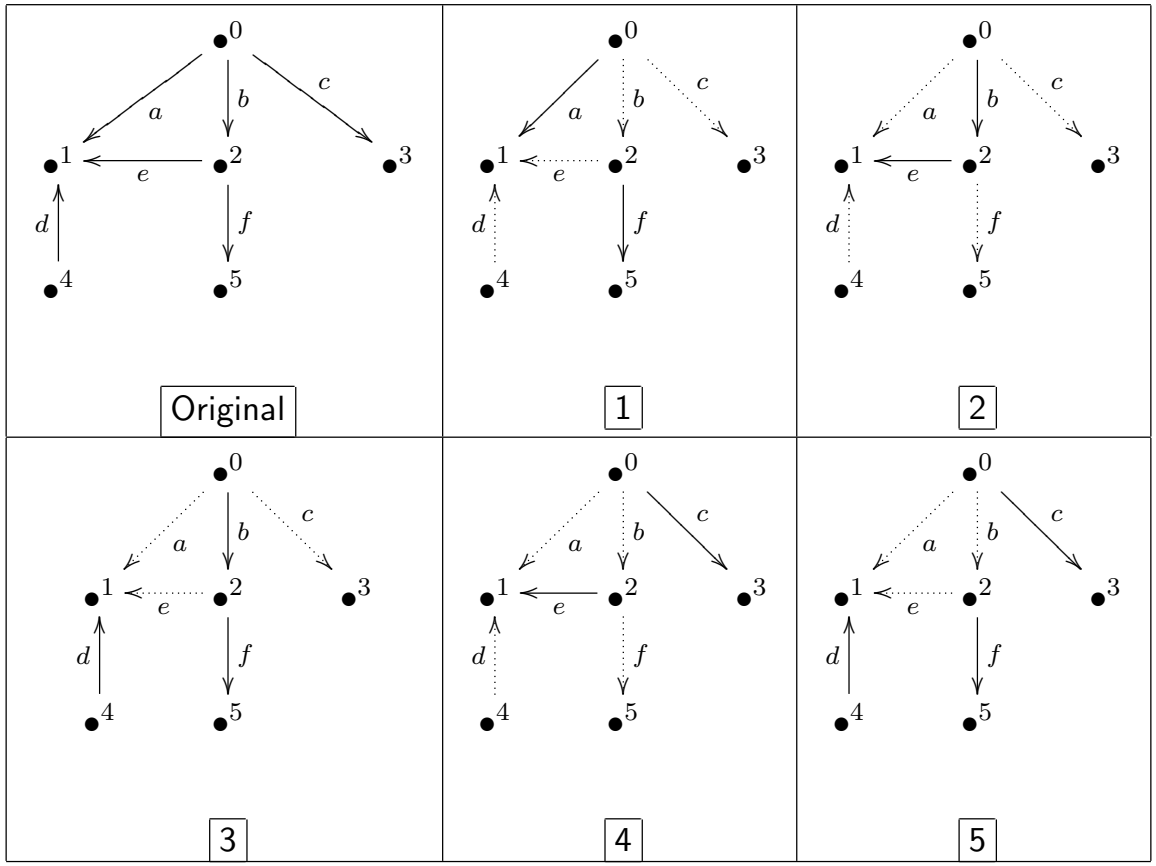
Type: Myrinet, Gigabit Ethernet
Paramètres: latence, bande passante,...



Résultats

Temps par événement
Temps par tâche
Conflits
...

Modèle Myrinet



	Communications					
	a	b	c	d	e	f
Somme	1	2	2	2	2	3
Minimum	1	1	1	2	2	2
Pénalité	5	5	5	2.5	2.5	2.5

Jérôme VIENNE, *Maxime* MARTINASSO

Traçage de l'application

Taille du Pb	Moy. du nb événs. calculs / coms	Temps moy. [s] calculs / coms	Volume moy. des coms. [Go]
12500	3084 / 6683	546 / 104	9.8
20500	5375 / 10901	2416 / 275	26.5

TAB.: Statistiques par tâche sur le graphe *HPL*, grappe Helios, réseau *Myrinet*

Obtention de la matrice de communication logique

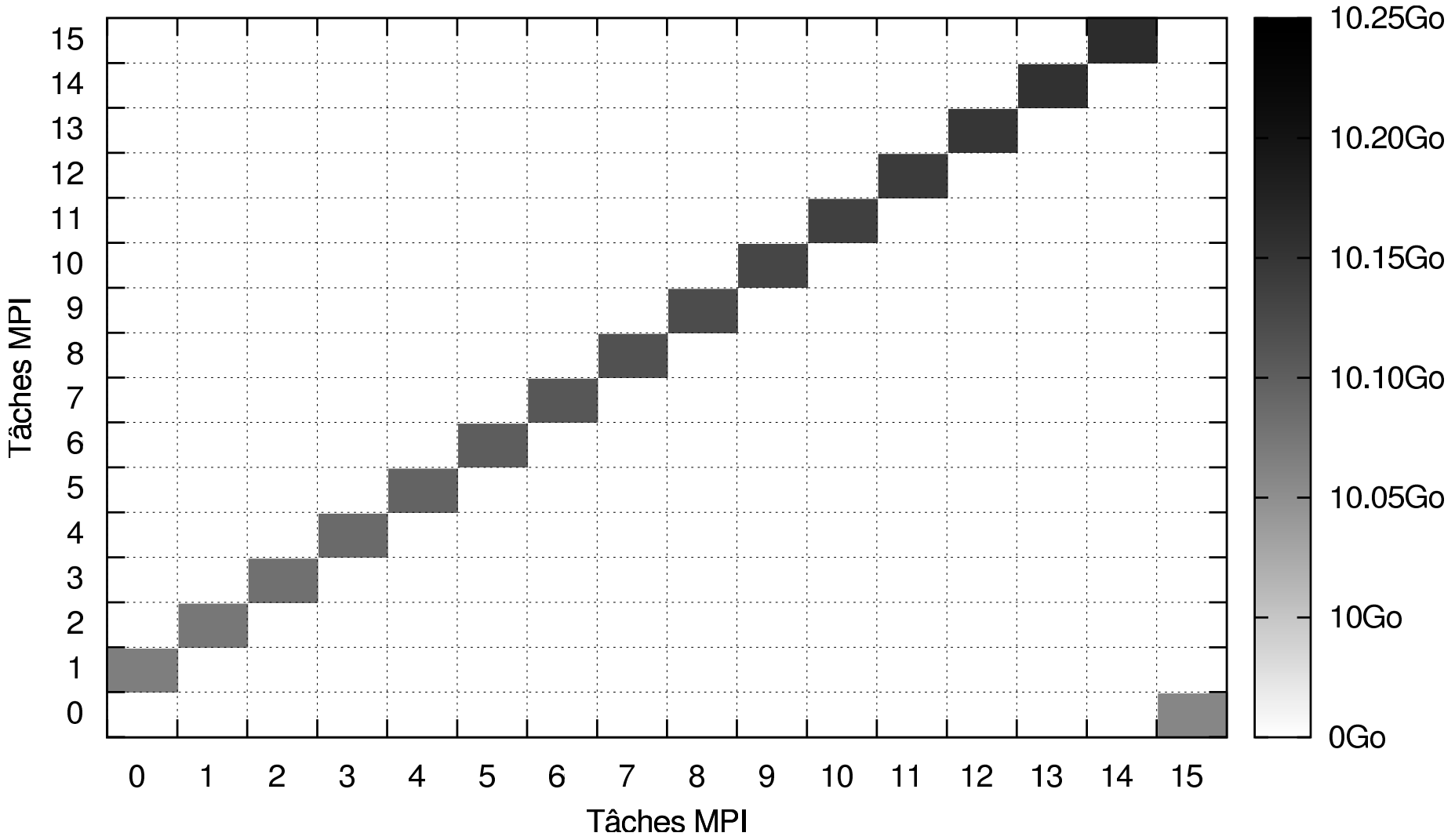
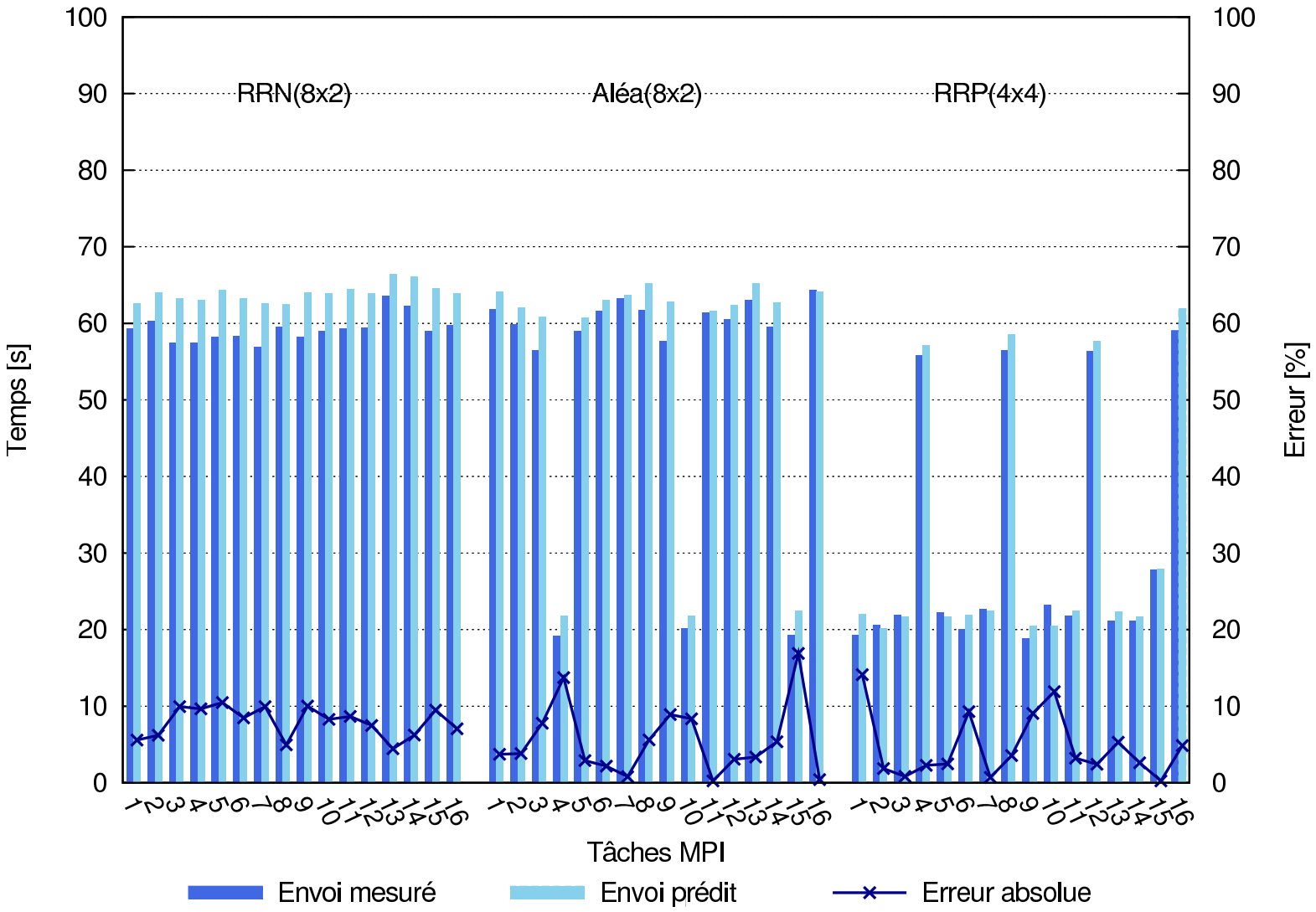


FIG.: Matrice de communication, 16 tâches, taille 12500

Résultat


RRN(8x2)
Aléa(8x2)
RRP(4x4)
Temps [s]
Erreur [%]
Tâches MPI
Envoi mesuré
Envoi prédit
Erreur absolue

Conclusion

- Outil permet de prendre en compte la concurrence réseau
- Travail au niveau de la congestion mémoire encore à faire
- Modélisation de l'Infiniband en cours