Compas'2023 : Parallélisme / Architecture/ Système
LISTIC/USMB - Annecy France, 4-7 juillet 2023

Interruptions en espace utilisateur pour améliorer la réactivité des communications en calcul haute-performance

Charles Goedefroit*

Centre Inria de l'université de Bordeaux,
200 avenue de la Vieille Tour
33405 Talence CEDEX - France
charles.goedefroit@inria.fr

Résumé
Les supercalculateurs modernes sont composés de plusieurs noeuds de calcul reliés par un réseau haute-performance. Ce type de réseaux a une latence de l'ordre de la microseconde, ce qui nous contraint à effectuer les communications directement à partir de l'*espace utilisateur* plutôt que par le noyau, qui utilise des appels système ayant un coût non négligeable à cette échelle. Le *polling*, qui consiste à interroger régulièrement la carte réseau, permet de faire progresser les requêtes en cours, mais celui-ci est effectué par un thread applicatif, ce qui consomme donc des ressources CPU et affecte la réactivité des communications. Pour éviter cela, l'utilisation d'*interruptions* est envisageable, mais celles-ci sont gérées par le noyau et utilisent donc des appels système que l'on cherche à éviter.
Un nouveau mécanisme d'*interruptions en espace utilisateur* arrive sur les nouveaux processeurs *Intel Xeon Sapphire Rapids*, permettant l'utilisation d'interruptions complètement en *espace utilisateur*, ce qui permet de réaliser des communications sans recourir au *polling*.
Ces nouvelles interruptions nous apportent donc une alternative efficace au *polling*, qui ouvre la voie à de nouveaux mécanismes pour faire progresser les communications au moment où les données sont disponibles. Nous proposons de modifier la bibliothèque de communication *NewMadeleine* – développée par l'équipe *TADaaM* d'*Inria* – pour qu'elle puisse prendre en charge la progression des communications grâce à des *interruptions en espace utilisateur*.
Dans un premier temps, nous modifions la gestion des paquets réseau pour permettre la progression des requêtes en cours à partir d'un handler d'interruptions et nous réalisons un prototype de driver *Uintr* pour les *communications inter-processus* en mémoire partagée. Nous comparons ensuite les performances de ce driver avec celui qui effectue du *polling* et celui qui utilise un traitement de signal.
Dans un second temps, nous nous concentrons sur les *communications inter-noeuds* en utilisant *BXI* (Bull eXascale Interconnect), un réseau haute-performance développé par *Atos* qui permet l'offloading des tâches réseau qui se déroulent sur la carte sans intervention du CPU. Notre objectif à terme est que ces cartes réseau puissent déclencher elles-mêmes des *interruptions en espace utilisateur* pour remonter les événements réseaux à un driver *Uintr* de *NewMadeleine*.

Mots-clés : *interruptions en espace utilisateur*, *réseau haut-performance*, *communications inter-processus*, *OS bypass*

---

*. Stage effectué sous la direction d'Alexandre Denis (Inria) et Grégoire Pichon (Atos)