Implémentation du support SVM pour l'IOMMU AMD virtuelle de QEMU

Niveau: dernière année de cycle Ingénieur ou Master 2 (Bac+5)
Durée: 5 à 6 mois
Lieu: division R&D d'Eviden à Échirolles (Grenoble) - Isère
Équipe: BXI Low Level
Contact: dl-bxi-sw-ll@eviden.com

Contexte

Le calcul haute performance

Eviden, à travers sa filiale Bull, est l'un des acteurs majeurs dans la course vers le calcul haute performance Exascale. Le supercalculateur Leonardo fabriqué par Bull se hisse à la quatrième place du Top500 dans son classement de septembre 2023. Certains de ces supercalculateurs ont la chance de pouvoir embarquer le réseau d'interconnexion haute performance BXI, également conçu et fabriquée par Eviden/Bull. Ces réseaux d'interconnexions sont composés de plusieurs centaines de cartes réseau appelées Network Interface Controller (NIC) ainsi que de switchs à plusieurs niveaux qui forment ensemble la topologie réseau. Les objectifs de ces réseaux d'interconnexion haute performance sont double :

- Pouvoir traiter les tâches de communication réseau rapidement et en parallèle des phases de calcul ;
- La mise à l'échelle de la performance sur des milliers de noeuds communicants entre eux.

La dernière génération du réseau BXI permet d'obtenir un débit utile de 100Gb/s avec une latence pouvant descendre sous la micro seconde. Avec les nouvelles générations de processeur et l'arrivée de la cinquième génération du bus PCIe, ce débit ne suffit plus pour rivaliser avec la vitesse des processeurs. Dans l'optique d'offrir une solution à la hauteur des derniers calculateurs, la troisième génération de la technologie BXI est en cours de conception.

Technologie IOMMU

Les plateformes actuelles intègrent une isolation mémoire à l'échelle des périphériques PCIe, appelée Input Output Memory Management Unit (IOMMU). Grâce à cette technologie, un périphérique PCIe n'aura accès qu'aux zones mémoire explicitement enregistrées au préalable par le système d'exploitation. L'isolation des périphériques d'I/O permet d'éviter tout débordement mémoire, ce qui augmente drastiquement la sécurité et la stabilité du système hôte. La future génération du NIC BXI devra cohabiter avec l'IOMMU de la plateforme hôte pour accéder à la mémoire du noyau ou de l'utilisateur. Pour des questions de performance, dans la plupart des cas, il n'est pas acceptable de devoir enregistrer chaque buffer de transmission ou de réception auprès de l'IOMMU avant de pouvoir y accéder depuis le NIC. De plus, au sein d'une application, les buffers réseaux sont localisés dans de l'espace utilisateur et sont identifiés par leurs adresses virtuelles, ce qui complexifie un peu plus leurs préparations. Les nouvelles générations des IOMMU apportent une solution à cette problématique avec la technologie Shared Virtual Addressing (SVA) chez Linux, aussi appelée Shared Virtual Memory (SVM) dans la norme PCIe. Cette technologie permet aux périphériques PCIe d'accéder à l'ensemble de l'espace

d’adressage d’un processus utilisateur à travers l’IOMMU. Ainsi, après avoir autorisé le NIC à accéder à l’espace d’adressage du processus, il n’est plus nécessaire d’enregistrer les buffers applicatifs avant leur utilisation. Cette technologie est cruciale pour l’architecture de BXI.

Émulateur BXI sous Qemu

Le logiciel libre Qemu permet d’exécuter un ou plusieurs systèmes d’exploitation (et leurs applications) isolés dans des machines virtuelles sur une même machine physique. Il embarque des versions émulées de la plupart des périphériques PCI courants : son, USB et réseau. Les systèmes d’exploitation invités partagent ainsi les ressources de la machine physique de façon relativement invisible. Qemu peut également être utilisé pour des besoins de recherche et développement sur des composants matériel. L’équipe BXI Low Level, qui s’occupe du pilote Linux pour le projet BXI, a utilisé cette technologie pour développer un émulateur de la carte réseau. Cet émulateur permet de travailler sur les couches logicielles (driver, bibliothèques exposées aux utilisateurs) sans attendre la disponibilité du matériel et donc de prototyper rapidement de nouvelles idées.

Qemu est également capable d’émuler des IOMMU de différents fabricants (Intel, AMD, ARM et bientôt RISC-V) et ainsi permet à l'équipe de travailler facilement avec cette technologie dans les machines virtuelles de développement BXI. Cependant en l'état actuel des choses, seule l'IOMMU virtuelle d'Intel implémente la technologie SVM indispensable à la solution BXI. La carte BXI devant être compatible avec plusieurs plateformes (aarch64, x86) et plusieurs fabricants d'IOMMU, la situation à ce jour ne permet qu'une mise en place partielle des fonctionnalités voulues.

Objectif du stage

Le futur stagiaire se verra proposer la réalisation d’un prototype d’implémentation de la technologie SVM dans l’IOMMU virtuelle AMD de Qemu. La méthode de travail sera basée sur des cycles itératif constitués d’étapes de design, de développement et de test. Le prototype sera finalement mis en application sur un cas d’usage réel qui fonctionne d’ores et déjà dans l'émulateur BXI. Si le prototype est concluant, le stagiaire pourra proposer son implémentation sous forme de patchs à la communauté de Qemu. Pour arriver à l'objectif proposé, le stagiaire devra étudier la spécification PCIe Gen 5 ainsi que celle de l'IOMMU AMD.

Profil recherché

Le profil idéal doit avoir de bonnes connaissances sur les points suivants :

- Architecture des ordinateurs
- Langage C
- Fonctionnement de la mémoire virtuelle
- Gestion de versions de code (Git)
- Debug (gdb)

Des connaissances sur les points suivants sont appréciées :

- Noyau Linux
- Virtualisation

Mots clés

QEMU | Virtualisation | PCIe | IOMMU | Mémoire virtuelle

Bibliographie

Eviden / High-Performance Computing Solutions
https://eviden.com/solutions/advanced-computing/high-performance-computing/

AMD I/O Virtualization Technology (IOMMU) Specification, Octobre 2023
https://www.amd.com/content/dam/amd/en/documents/processor-tech-docs/specifications/48882_IOMMU.pdf

QEMU / Developer Information / Internal QEMU APIs /The memory API
https://www.qemu.org/docs/master/devel/memory.html

The Linux Kernel / CPU Architectures / x86-specific Documentation / Shared Virtual Addressing (SVA) with ENQCMD
https://www.kernel.org/doc/html/next/x86/sva.html

LWN.net / Shared Virtual Addressing for the IOMMU
https://lwn.net/Articles/747230/