zeqL

CC BY NC SA

Geste de savoir

PCI Express : petit tour d'horizon

24 mars 2021

Table des matières

La version 5.0 du PCI Express a été ratifiée en juin 2019 avec des puces qui devraient arriver en nombre d'ici quelques mois et années mais surtout la ratification du PCI Express 6.0 qui devrait avoir lieu cette année, le PCI Express va de plus en plus faire parler de lui. Petit tour d'horizon hardware du PCI Express qui devient progressivement un protocole de communication majeur dans les cartes électroniques!

1. PCI Express 5.0

Le PCI Express 5.0 offre tout simplement un doublement du débit par rapport à la précédente génération. On passe donc de 16GT/s à 32GT/s. Les GigaTransfers/s correspondent au débit brut qui est transmis, le débit utile est légèrement inférieur à cause de l'utilisation de technique de codage. En l'occurence nous restons sur un codage 128b/130b présent depuis le PCI Express 3.0.

Génération	Codage	Débit brut	Débit utile	x1	x16
PCIe 1.0	8b/10b	2.5 GT/s	2.0 Gbps	0.25 Go/s	4.0 Go/s
PCIe 2.0	8b/10b	5.0 GT/s	4.0 Gbps	0.50 Go/s	8.0 Go/s
PCIe 3.0	128b/130b	8.0 GT/s	7.87 Gbps	0.985 Go/s	15.754 Go/s
PCIe 4.0	128b/130b	16.0 GT/s	15.75 Gbps	1.969 Go/s	31.508 Go/s (252 Gbps)
PCIe 5.0	128b/130b	32.0 GT/s	31.51 Gbps	3.938 Go/s	63.015 Go/s (504 Gbps)

Au niveau électrique le signal est en NRZ, il n'y a que deux états possibles: 0 ou 1. La fréquence du signal électrique (aussi appelée fréquence de Nyquist) est de 16GHz. On commence à atteindre les limites techniques (ou plutôt physiques) en terme de matériaux et d'atténuation du signal pour les cartes électroniques.
C'est pour cela que le PCI Express 6.0 passe sur un signal modulé en PAM4, avec 4 états

possibles. La fréquence de Nyquist reste la même: 16GHz mais le signal est moins résistant à des perturbations.

Les spécifications électriques du PCI Express 5.0 ont été adaptées à cette fréquence de signal plus élevée (et surtout d'atténuation plus importante) pour pouvoir réaliser des cartes de type carte mère où les ports sont un peu éloignés du processeur. Néanmoins l'utilisation de retimer (des puces qui permettent de renvoyer un signal atténué) risque d'être la norme pour les cartes mères de grande taille ou alors il y aura un mix de ports en 5.0 et en 4.0 (les plus éloignés du processeur).

Des connecteurs améliorés, mais rétro-compatibles, seront probablement nécessaires pour utiliser cette nouvelle génération sauf à avoir une faible distance entre le contrôleur et le connecteur.

2. Topologie

Commençons par voir la topologie d'un bus PCI Express pour expliquer les différents éléments:

C'est une architecture top-down avec un hôte unique.

Le point le plus important est le principe du *root complex* qui relie tous les périphériques au processeur et à sa mémoire (l'hôte). Rappelons que même si le processeur a de la mémoire embarquée (cache) le traitement de données passe d'abord par un échange dans la mémoire RAM. Les périphériques PCI Express vont venir écrire les données à traiter dans la mémoire RAM.

Les périphériques sont appelés *endpoint*. Même si on peut relier tous les périphériques directement sur le *root complex*, on peut utiliser des *switch* pour augmenter le nombre de connexions ainsi que des *bridges* vers d'autres protocoles (PCI, Ethernet, USB, etc.)

Une liaison PCI express est point à point avec un port d'*upstream* (qui remonte vers le *root complex*) et un de *downstream* (qui descend du *root complex*).

2.1. Protocole

Le PCI Express même si on l'appelle bus a un fonctionnement similaire à l'Ethernet avec des données transmises par paquets. Les périphériques ont des adresses pour communiquer entre eux. Il fonctionne sur les trois premiers niveaux du modèle OSI.

Je vous invite à aller voir le lien *Down to the TLP : How PCI express devices talk* donné à la fin de l'article qui détaille le protocole de communication.

3. Ca switch !

Si on l'on peut comparer le PCI Express à l'Ethernet sur certains points, le fonctionnement par paquets, la possibilité d'utiliser des switch, il y a néanmoins une différence fondamentale: l'hôte et le *root complex*. Ce qui en fait un protocole avec un maître et des esclaves, ce qui n'est pas le cas de l'Ethernet.

Cela implique que l'on ne peut pas, nativement, connecter deux hôtes sur le même bus PCI Express même avec un switch. Néanmoins cette possibilité a été envisagée via l'utilisation d'un Non Transparent Bridge (NTB).

Le NTB isole électriquement et logiquement les différents bus PCI Express. Pour assurer la communications entre les différents réseaux, une translation d'adresses est effectuée, chaque côté du bridge ayant son propre registre d'adresses.

Deux méthodes de translation sont possible: translation directe avec offset ou alors via une table de correspondance (*lookup table*).

Des registres supplémentaires peuvent être prévus pour assurer la communication entre les deux hôtes (CPU) de status, d'interruptions, etc.

Les NTB permettent de développer des systèmes multi-host et de développer des plateformes de calcul intensif utilisant le PCI Express comme bus de communication. Bus PCI Express géré nativement par les CPU, ce qui évite les latences dues à des changement de protocole (par exemple utilisation de l'Ethernet, qui imposera une translation PCI Express -> Ethernet -> PCI Express).

Dans cet exemple de calcul intensif, les CPU sont des facteurs limitant. Pour optimiser l'utilisation des GPU, les switch PCIe permettent à chaque CPU d'utiliser les quatre GPU via le PCI Express. Les données sont écrites sur des SSD connectés en PCI Express.

4. Allons plus loin !

Un goulot d'étranglement peut être identifié avec le *root complex*, le CPU et la mémoire RAM. En effet, si deux périphériques (*endpoint*) souhaitent discuter entre eux, la communication devra passer par le *root complex* (et surtout la RAM), même si celui-ci n'a pas grand chose à voir avec l'échange.

NVidia a développé la technologie GPUDirect qui consiste en un DMA entre deux *endpoints* sans passer par la mémoire RAM. Il y a toujours une communication avec le *root complex* pour l'organisation de la transaction mais les données circuleront directement entre les *deux endpoints*. Le GPUDirect s'utilise avec un GPU mais l'autre périphérique n'est pas forcément un GPU, cela peut être une carte d'acquisition vidéo, un adaptateur Ethernet, etc. il doit néanmoins être compatible GPUDirect.
AMD a une technologie similaire avec le DirectGMA. Ces deux technologies sont disponibles sur les versions professionnelles des cartes graphiques.

4.1. Direct Memory Access

Première possibilité offerte par le GPUDirect, les accès DMA entre les périphériques. Dans l'exemple ci-dessus une carte graphique qui souhaite envoyer un flux vidéo sur le réseau. Sans GPUDirect, la transaction PCI Express se fera en deux étapes. Premièrement le GPU copie ses données en RAM (dans un espace alloué par la MMU), puis le CPU vient copier les données du GPU dans un second espace en RAM, dédié à la carte réseau, enfin la carte réseau vient récupérer les données dans cet espace mémoire.
Le GPUDirect permet de réaliser un accès DMA entre le GPU et la carte réseau en n'utilisant qu'un seul espace mémoire en RAM. Le CPU est déchargé de la copie des données et n'est donc plus un goulot d'étranglement.

4.2. Peer-to-Peer communication

Une autre possibilité est la communication entre les mémoires de deux GPU.

— Direct Access: un GPU peut venir lire ou écrire dans la RAM d'un autre GPU (registre)
— Direct Copy: un GPU peut réaliser une copie de ses données en RAM via un accès DMA vers un autre GPU

L'utilisation d'un switch permet à ces transferts de données de ne pas passer par le *root complex*, d'autres transactions PCI Express peuvent donc avoir lieu sans risque de saturer la bande passante. Néanmoins il y a toujours des accès avec le CPU pour initier les transactions mais le débit est fortement réduit.

J'espère que ce billet sur le PCI Express bas niveau vous aura donné envie de vous y intéresser un peu plus. Si vous souhaitez creuser un peu plus voici quelques liens:

— Down to the TLP : How PCI express devices talk
— Practical introduction to PCI Express with FPGAs
— Non-Transparent Bridging Simplified
— NVIDIA GPUDirect Technology Overview

A noter que Nvidia développe de manière très agressive ses différentes technologies logicielles se basant sur le PCI Express qui sert de lien de communication pour les cartes graphiques. De nombreux SDK ciblant des domaines différents existes, mais se basant plus ou moins sur les même technologies:

— Rivermax (Streaming)
— Clara (médical)

Liste des abréviations

NRZ Non Return to Zero. 1