

AVANTIX

Avancement Simulateur IQ

17/01/2024

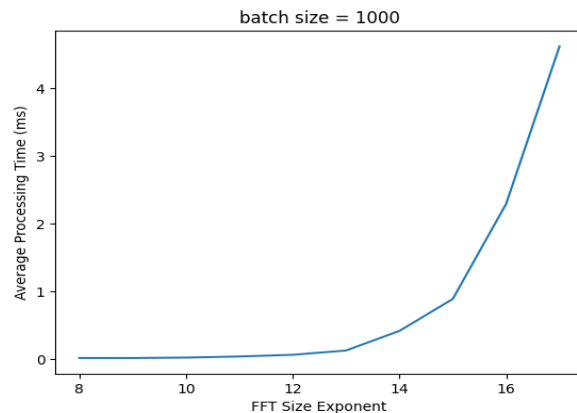
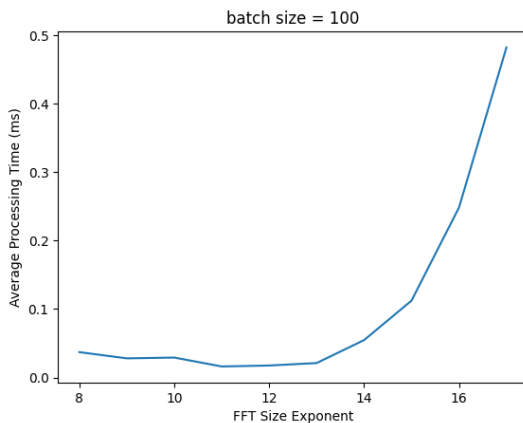
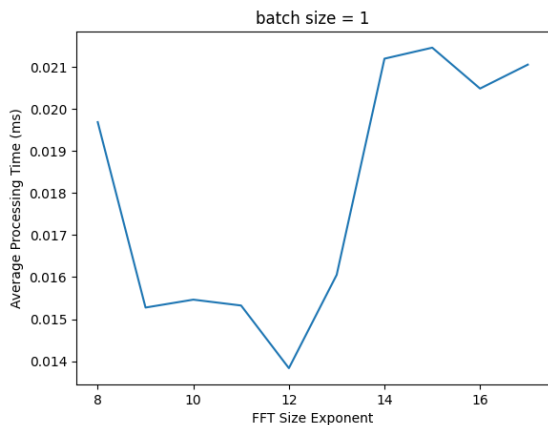
© Avantix – for internal use

AVANTIX

Benchmark fonctionnel

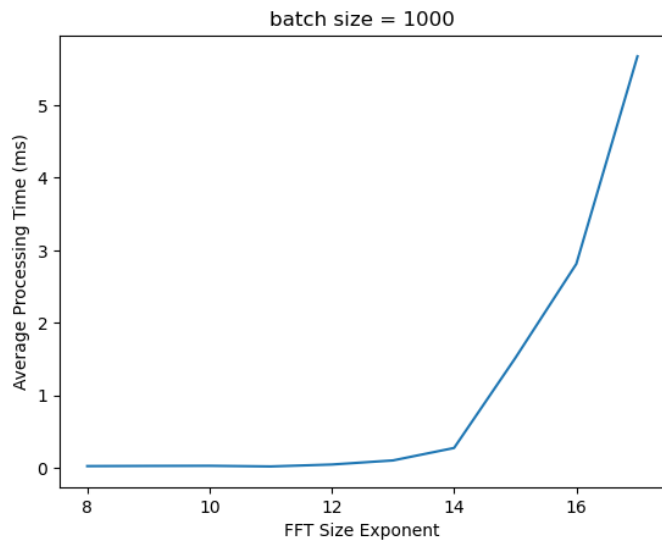
Benchmark Pytorch FFT

FFT par batches



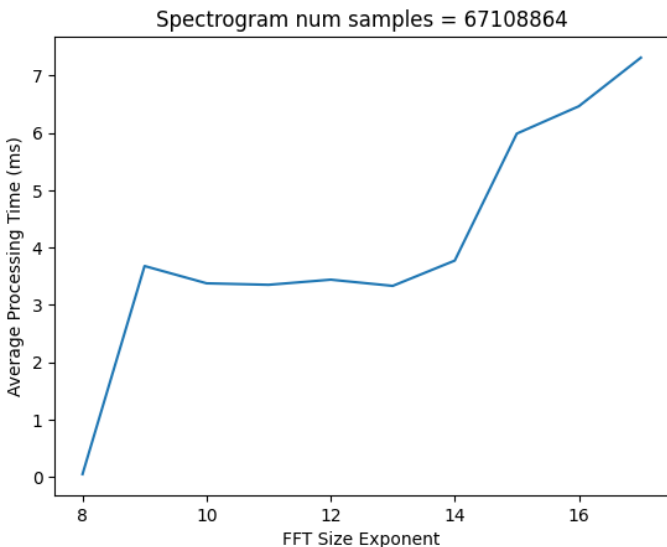
Benchmark Cupy FFT

FFT par batchs : num ech = N FFT * bach size

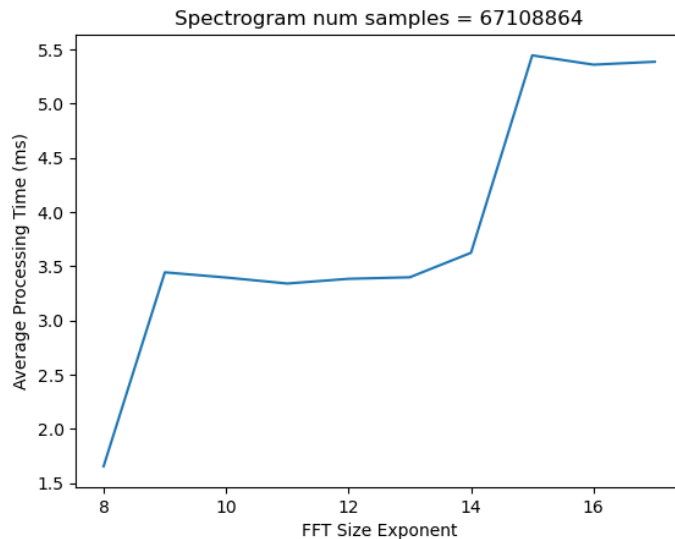


Benchmark Pytorch vs cupy FFT

Spectrogramme sans fenêtrage sur nombre d'échantillons fixe



Torch



Cupy

- Sensiblement iso-perf avec avantage pytorch pour l'intégration (prefetching)
- Relative invariance de la taille de FFT

Benchmark YOLO

Yolo v8 nano 640*640

- Batch_size = 1000 ==> 3,4 ms / image dont 0,6ms uniquement d'inférence
- Impact quantification FP16 non perceptible sur V100 à tester sur H100 : [problème similaire observé ailleurs](#)
- Potentiel rapport 3 en se ramenant en mono-canal au lieu de RGB
- Investiguer le BFLOAT16 https://en.wikipedia.org/wiki/Bfloat16_floating-point_format

Cas d'usage:

- 15 batch d'images à traiter / seconde @3GHz mono-voie pour 1 résolution
- 1 batch ~ 6,5 Gb de données à transférer au GPU par RDMA
- 2 ordres de grandeur à gagner
- Besoin d'affiner le benchmark

Benchmark Pytorch

Sur serveur DESPINA

Pytorch 2.1.2

CUDA 12.1

Méthode `%timeit -n 1 -r 1`

Modèle ~300k paramètres

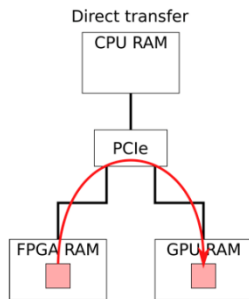
AVANTIX

Transferts FPGA GPU

XDMA ou RDMA

Direct communication between FPGA and GPU using Frame Based DMA (FDMA) 10/2019

- S'appuyer sur le GPUDirect (RDMA) de NVIDIA pour définir IP FPGA
- 1 DMA par voie ou entrelacement sur un seul DMA
- Timestamping pour consolider la gestion des paquets perdus

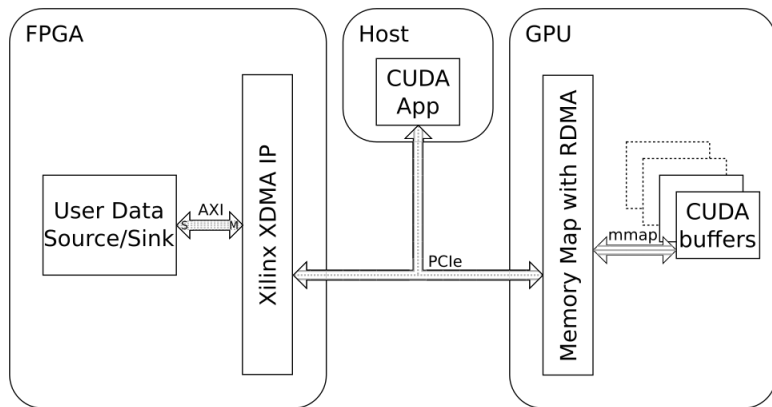


- Mapping GPU RAM Linux IO-mem address space
- Permet un ré-adressage auto des transferts

XDMA ou RDMA

IP classiques sur FPGA ultrascale +:

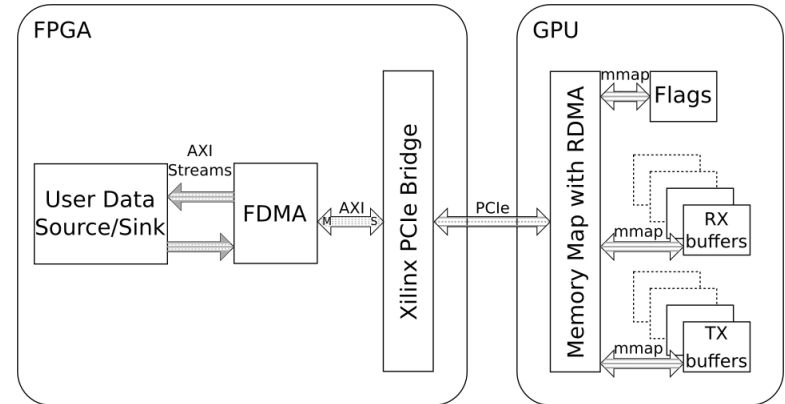
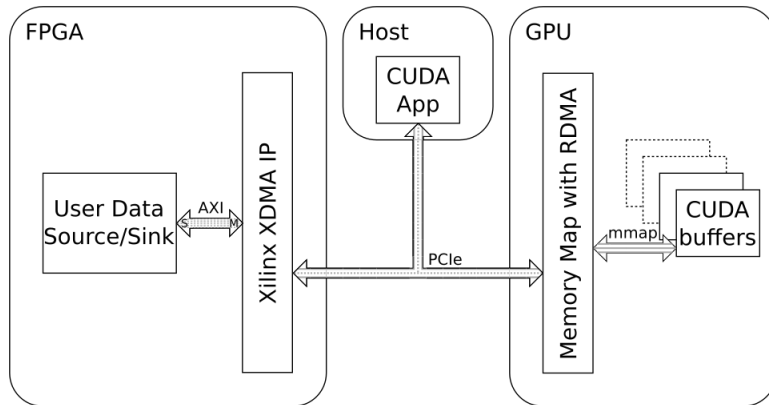
- XDMA ou QDMA, les 2 ont été utilisées extensivement chez ADE EW
- Une appli sur le CPU initie chaque transfert en direct
- Cette appli doit suivre les messages
- ETH Zurich a mis en place une gestion continue du flux (FDMA)



XDMA ou RDMA

IP classiques sur FPGA ultrascale +:

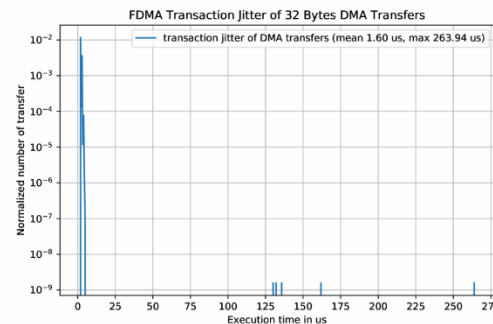
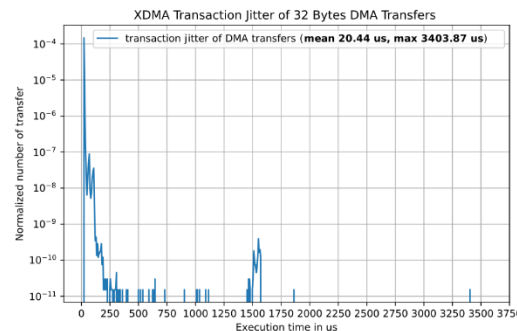
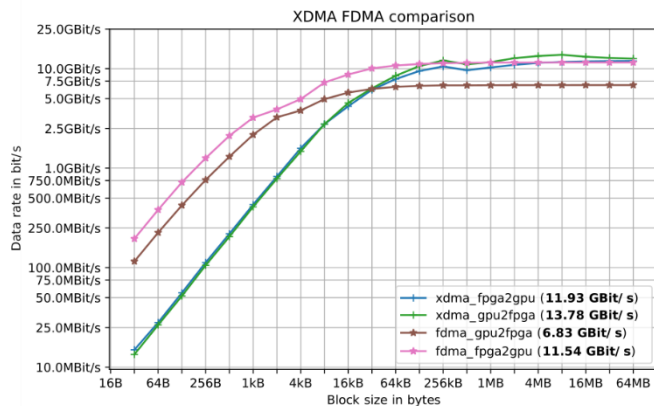
- XDMA ou QDMA, les 2 ont été utilisées extensivement chez ADE EW
- Une appli sur le CPU initie chaque transfert en direct
- Cette appli doit suivre les messages
- ETH Zurich a mis en place une gestion sans intervention du CPU après init (FDMA)



XDMA ou RDMA

Comparaison des performances sur du PCIe gen2x4 (16Gbps)

- Jitter prohibitif sur XDMA



AVANTIX

MERCI