Utilisation de GP-GPU[1] pour le calcul à hautes performances

David HILL

David.Hill@univ-bpclermont.fr

[1]GENERAL-PURPOSE COMPUTING ON GRAPHICS PROCESSING UNITS

Architecture “CUDA” Connexion d’une machine hôte et des cartes GPGPU multicoeurs

COMPUTE UNIFIED DEVICE ARCHITECTURE

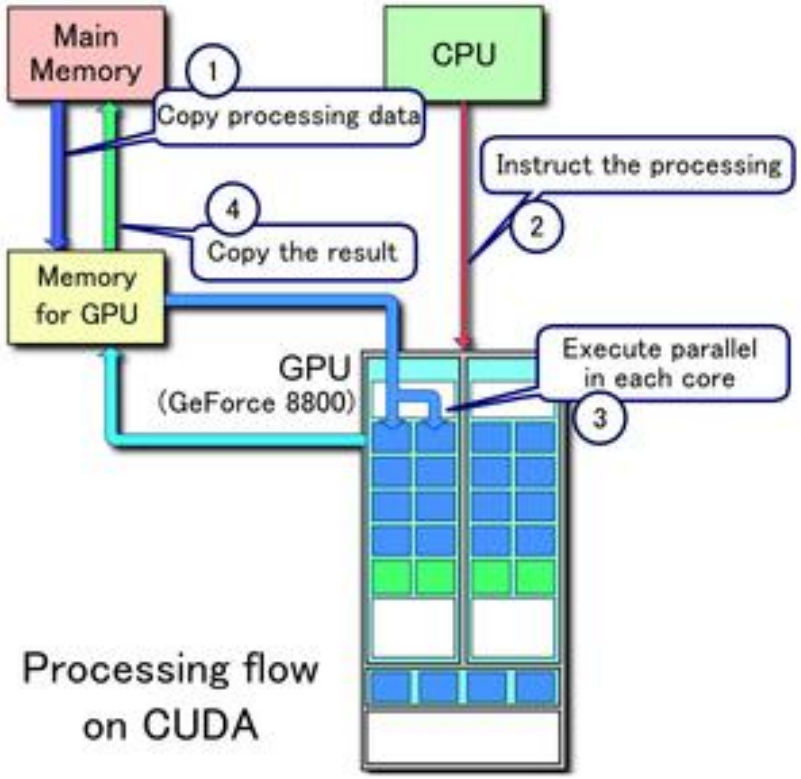
Quelques avantages / GPU classiques

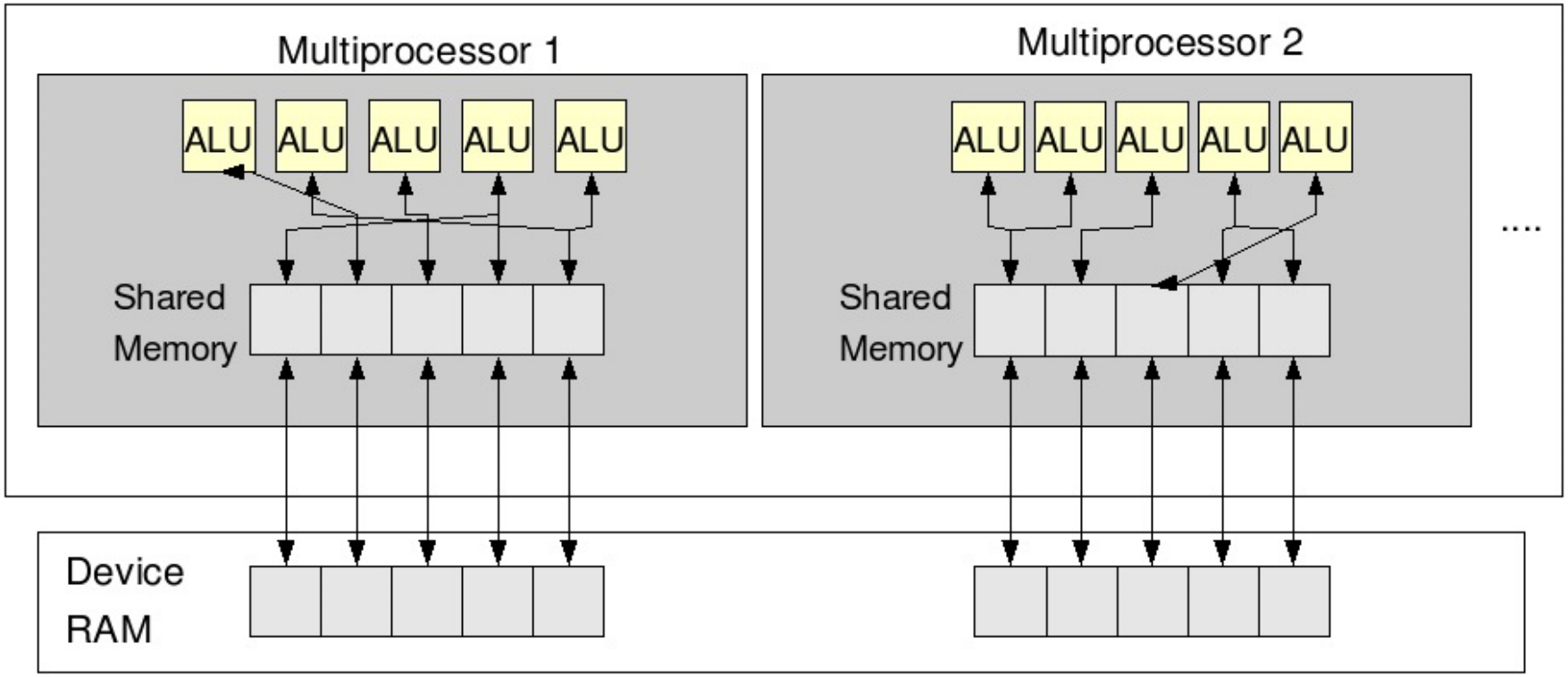
- **Mémoire partagée rapide accessible de tous les threads (processus parallèles)**
- **Cette mémoire peut être utilisée comme un cache utilisateur**
- **Téléchargement rapide des codes et données (PCI 16X actuellement)**
- **Support complet des opérations classiques**

Exemple de flot de traitement sur une architecture CUDA

Cartes NVIDIA

- **GeoForce…**
- **Quadro…**
- **Tesla**





Les GPU compatibles CUDA implémentent un multi-processeur, chacun dispose de plusieurs ALUs (arithmetic logic unit) capable d'exécuter le même code à chaque cycle d'horloge (jeu d'instruction) mais sur des données différentes. Chaque ALU peut lire et écrire la mémoire partagée et communiquer avec la mémoire RAM

Manavski and Valle BMC Bioinformatics 2008 9(suppl 2):s10 doi:10.1186/1471-2105-9-s2-s10

Cartes Tesla
Le début d'une petite révolution en HPC – 2008…

- **Cartes NVIDIA**
  - **~1 Teraflops en simple précision**
  - **80 Gigaflops en double précision**
  - **240 coeurs !!!**
  - **300 W**
- **Chassis 1 U**
  - **4 GPUs Tesla**
  - **3.7 Teraflops SP**
- **Clusters …**

Extraits de Vidéo aux conférences sur les supercalculateurs…

Google video : personal supercomputer

Caractéristiques techniques (1/2)

Configuration	Model	# of GPUs	Core clock in MHz (each)	Shaders	
				Thread Processors (total)	Clock in MHz (each)
GPU Computing Processor[1]	C870	1	600	128	1350
Deskside Supercomputer[1]	D870	2	600	256	1350
GPU Computing Server[1]	S870	4	600	512	1350
C1060 Computing Processor[2]	C1060	1	602	240	1300
S1070 1U GPU Computing Server[2,3]	S1070	4	602	960	1500

Caractéristiques techniques (1/2)

Configuration	Model	# of GPUs	Core clock in MHz (each)	Shaders	
				Thread Processors (total)	Clock in MHz (each)
GPU Computing Processor[1]	C870	1	600	128	1350
Deskside Supercomputer[1]	D870	2	600	256	1350
GPU Computing Server[1]	S870	4	600	512	1350
C1060 Computing Processor[2]	C1060	1	602	240	1300
S1070 1U GPU Computing Server[2,3]	S1070	4	602	960	1500

Caractéristiques techniques (2/2)

Configuration	Memory					Processing Power (GigaFLOPS, total)
	Bandwidth max (GB/s)	Bus type	Bus width (bit, each GPU)	Total size (MiB)	Clock (MHz)	
GPU Computing Processor[1]	77	GDDR3	384	1536	1600	519
Deskside Supercomputer[1]	154	GDDR3	384	3072	1600	1037
GPU Computing Server[1]	307	GDDR3	384	6144	1600	2074
C1060 Computing Processor[2]	102	GDDR3	512	4096	1600	936
S1070 1U GPU Computing Server[2,3]	410	GDDR3	512	16384	1600	4320

Bio-Informatics and Life Sciences

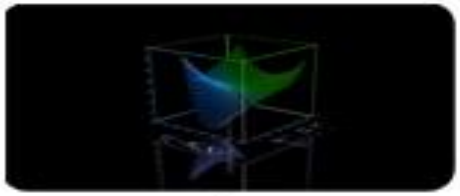
Computational Chemistry

Computational Electromagnetics and Electrodynamics

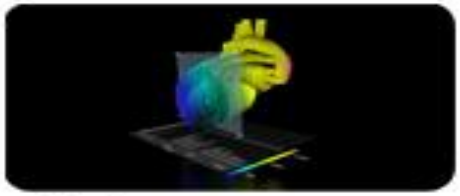
Computational Finance

Computational Fluid Dynamics

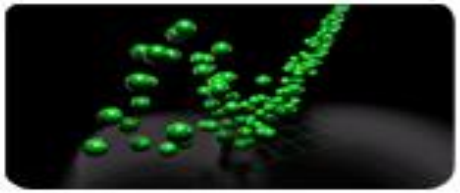
Imaging and Computer Vision



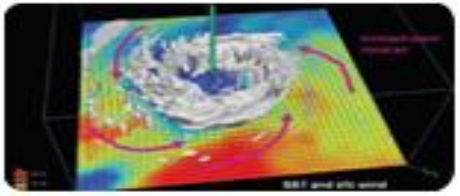
MATLAB Acceleration



Medical Imaging



Molecular Dynamics



Weather, Atmospheric, Ocean Modeling, and Space Sciences

Quelques applications…

Intérêts de tous les constructeurs majeurs…

1. IBM
2. HP
3. SGI
4. DELL
5. CRAY

…

Quelques références pour la bioinformatique

Schatz, M.C., Trapnell, C., Delcher, A.L., Varshney, A. High-throughput Sequence Alignment Using Graphics Processing Units. BMC Bioinformatics 8:474, (2007)

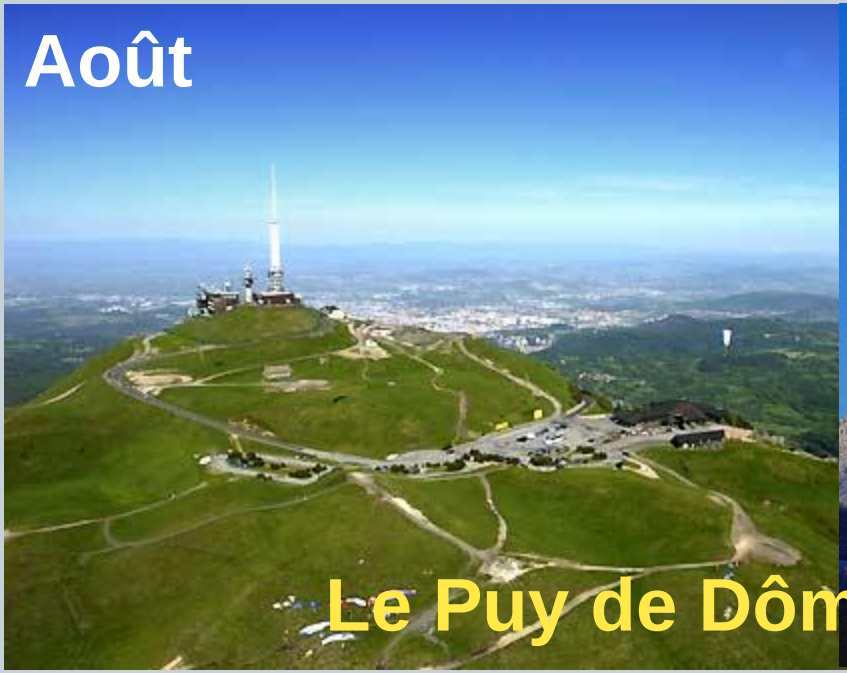
Svetlin A. Manavski, Giorgio Valle (2008). "CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment".
BMC Bioinformatics 9 (Suppl 2):S10.

Cole Trapnell, Michael C. Schatz, Optimizing data intensive GPGPU computations for DNA sequence alignment, Parallel Computing, Volume 35, Issues 8-9, August-September 2009, Pages 429-440, ISSN 0167-8191.

Limitations…

1. Sous ensemble du C (sans recursion, sans pointeurs de fonction, …
2. Pas de “texture rendering”
3. Déviation du standard IEEE 754 pour la double precision
4. En simple précision : denormals and signalling nans ne sont pas implémentés, seulement 2 modes d’arrondis ieee.
5. La bande passante du bus et la latence entre la CPU et le GPU peuvent être un goulet d’étranglement.
6. Les threads doivent être lancés par paquets de 32 pour de meilleures performances. Les branchements n’ont pas trop d’impacts si les 32 threads prennent le même chemin d’exécution
7. Les GPU CUDA ne sont actuellement disponibles que chez NVIDIA

Questions ?



Le Puy de Dôme - Clermont-Ferrand