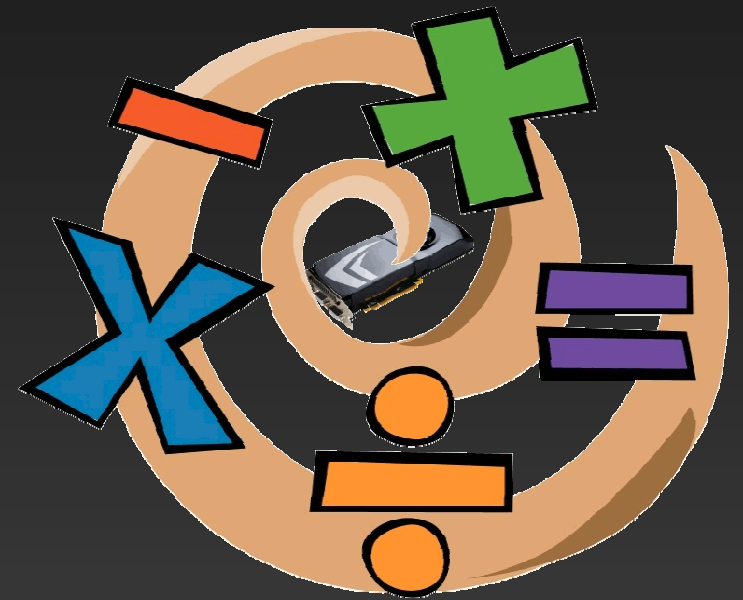


Passé, Présent et Futur des Processeurs Graphiques

David Defour

Laboratoire ELIAUS,
Université de Perpignan



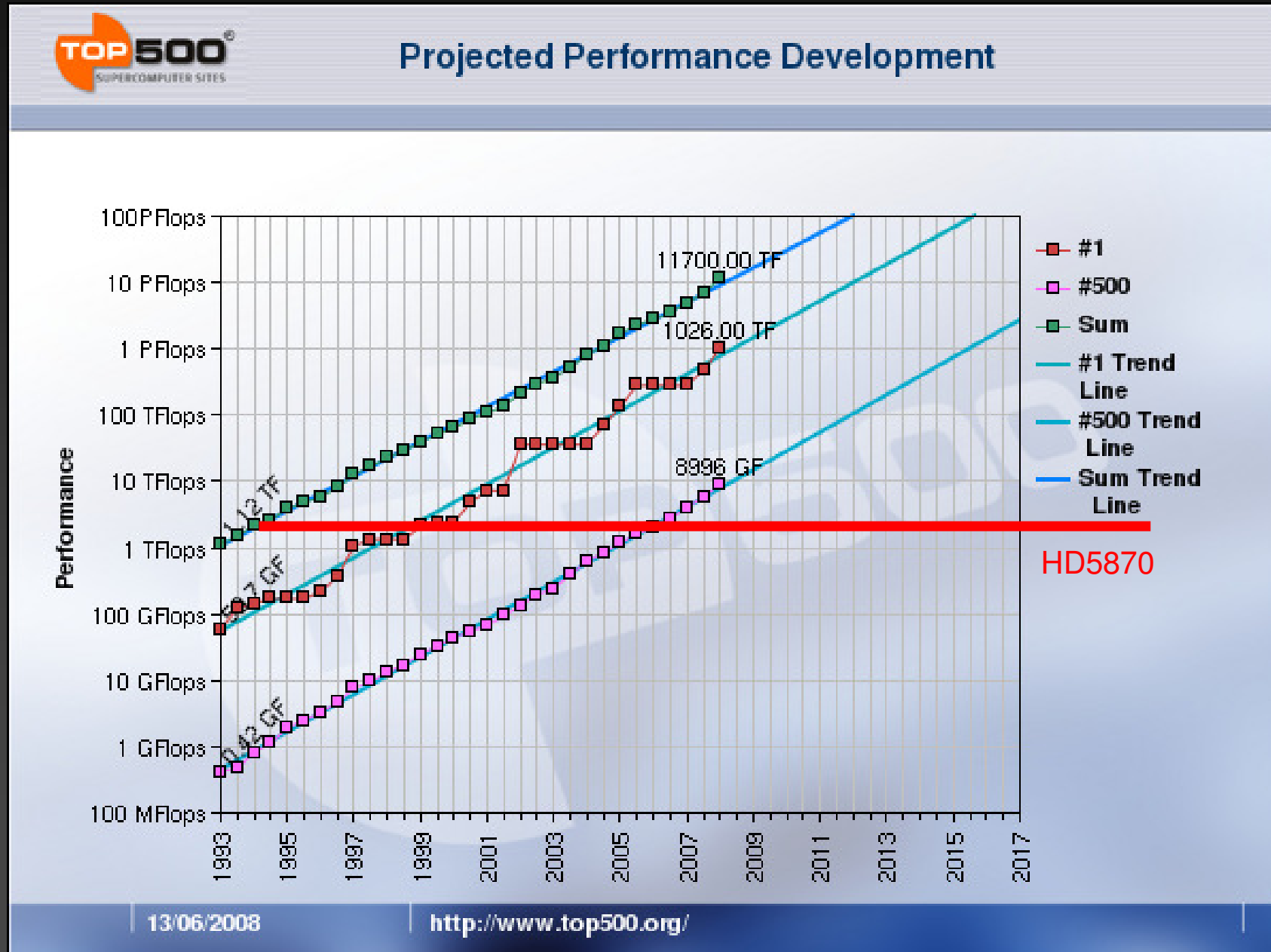
Les 10 points traités

1. Puissance de calcul
2. Précision
3. Modèle hiérarchique
4. Unités d'exécution
5. Bande passante
6. Modèle d'exécution
7. Jeux d'instruction
8. Débogage
9. Modèle de programmation
10. Consommation



1. Puissance de calcul

Plus on pédale moins fort, moins on avance plus vite



2. Précision

Ça casse pas des briques à un canard

| GPU | G80 | GT200 | Fermi |
|---|-------------------|---------------------|-----------------------------|
| Transistors | 681 million | 1.4 billion | 3.0 billion |
| CUDA Cores | 128 | 240 | 512 |
| Double Precision Floating Point Capability | None | 30 FMA ops / clock | 256 FMA ops /clock |
| Single Precision Floating Point Capability | 128 MAD ops/clock | 240 MAD ops / clock | 512 FMA ops /clock |
| Warp schedulers (per SM) | 1 | 1 | 2 |
| Special Function Units (SFUs) / SM | 2 | 2 | 4 |
| Shared Memory (per SM) | 16 KB | 16 KB | Configurable 48 KB or 16 KB |
| L1 Cache (per SM) | None | None | Configurable 16 KB or 48 KB |
| L2 Cache (per SM) | None | None | 768 KB |
| ECC Memory Support | No | No | Yes |
| Concurrent Kernels | No | No | Up to 16 |
| Load/Store Address Width | 32-bit | 32-bit | 64-bit |

Conclusion : Généralisation de la DP, IEEE 754-2008

4. Unités d'exécution

□ Hier

- PipelineS graphiqueS hyper spécialisés

□ Aujourd'hui

- PipelineS unifiéS avec encore quelques unités spécialisées (MAD, FMA, texturage, ROP, filtrage, ...)

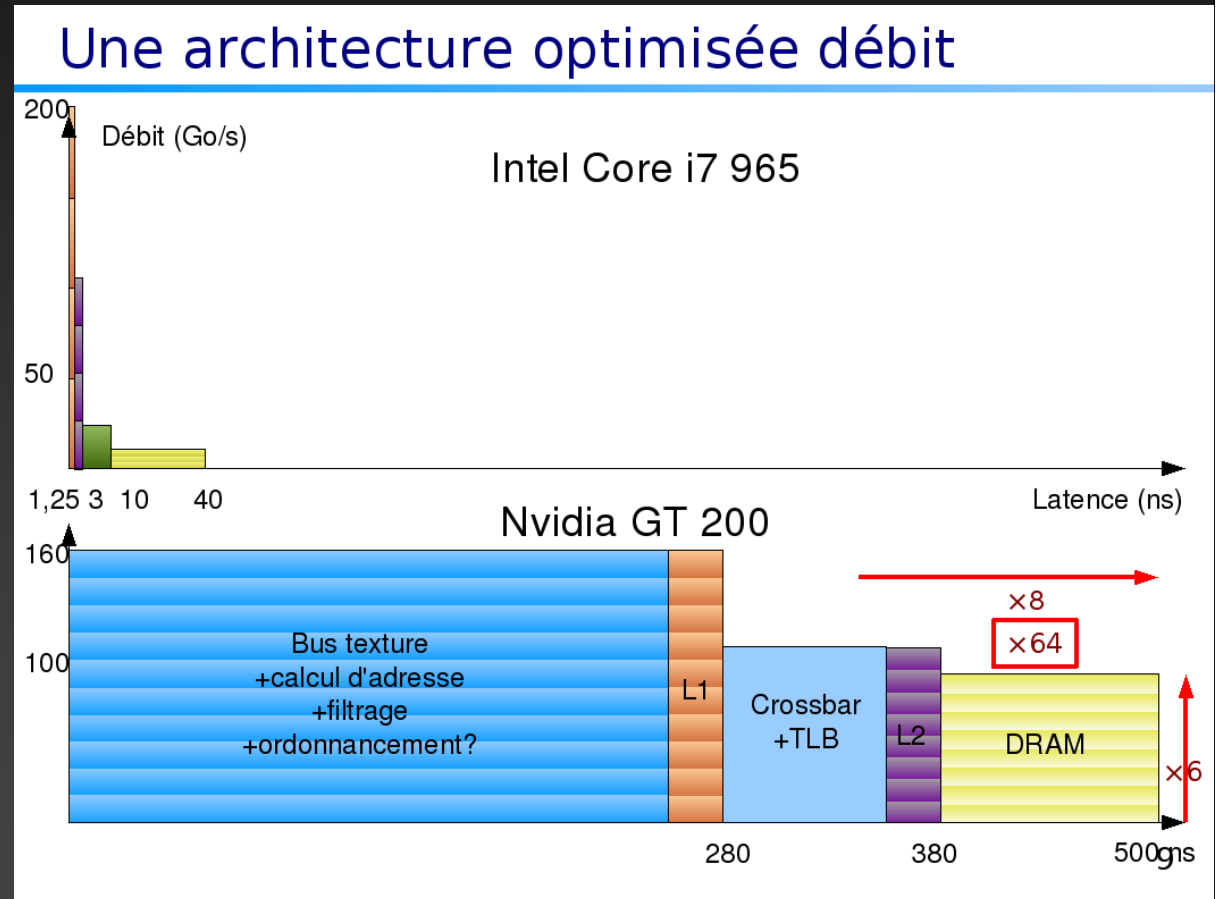
□ Demain

- Quel ratio entre chaque type de pipeline ?



5. Bande passante

- Data //
 - Peu de communication
 - Accès direct à la mémoire
- Matériel
 - Focalisé sur le débit
 - Hiérarchie mémoire 'inversée' + scratchpad
- Demain
 - Dépend du marché cible



Source : Sylvain Collange©

6. Modèle d'exécution

□ Interface logiciel d'accès au matériel

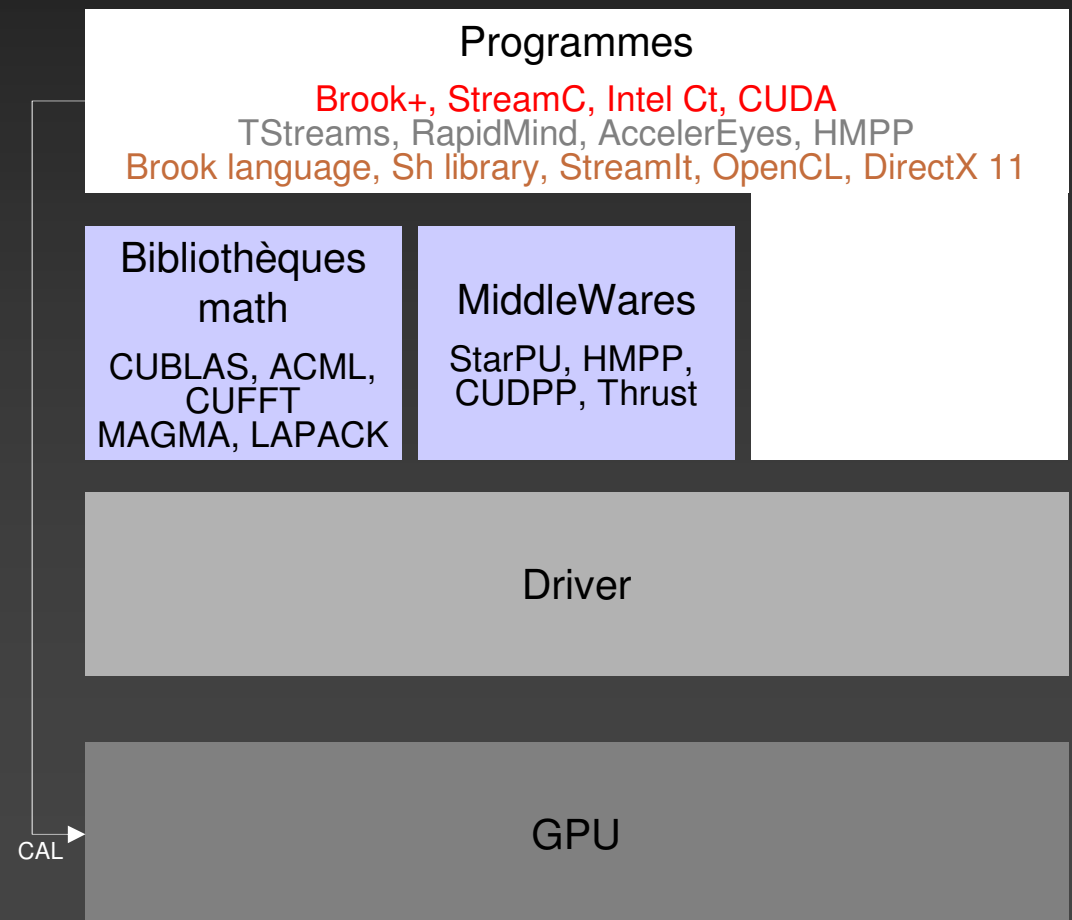
1. Langages de haut niveau (CUDA, OpenCL)
2. Compilation : pseudo-code
3. Interprétation : assembleur

□ Pro

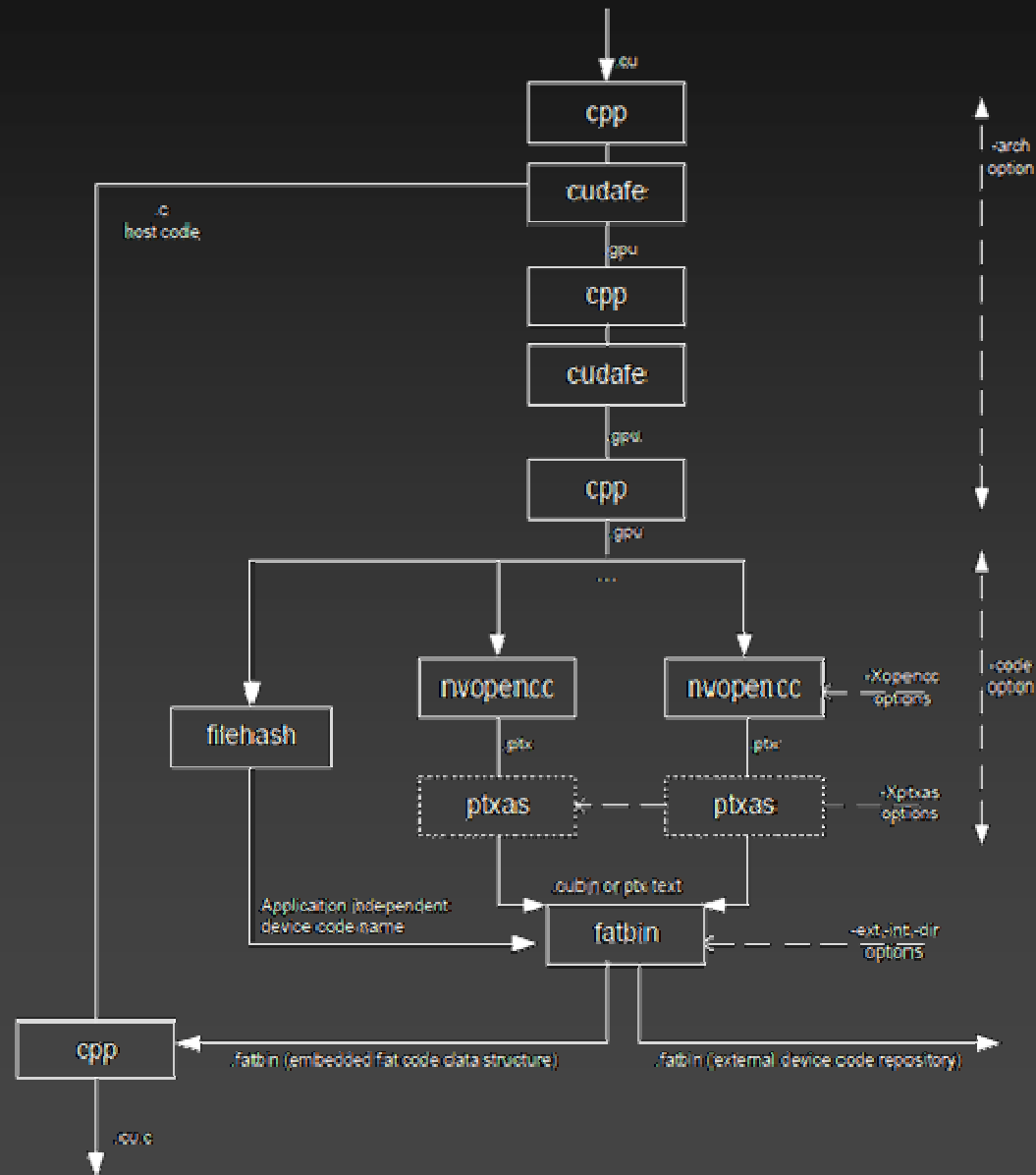
- Optimisation, transparence lors des changements architecturaux ou microarchitecturaux

□ Cons

- Effet boîte noire



7. ISA et Modèle de compilation CUDA



8. Débogage

□ Hier

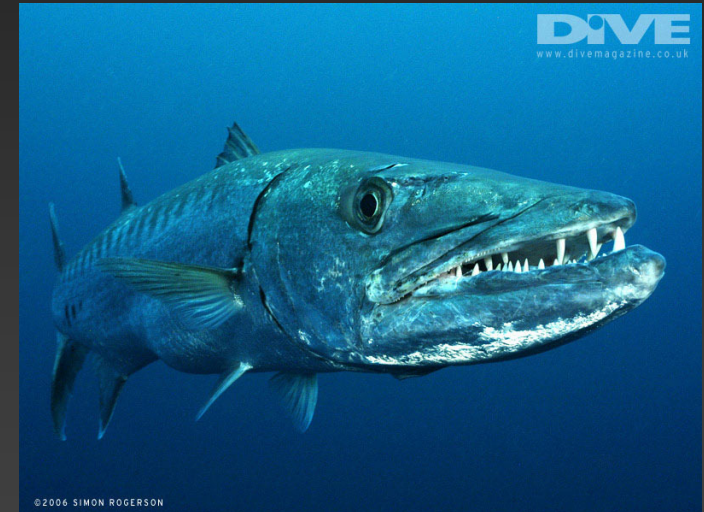
- Pas possible

□ Aujourd'hui

- Decuda, GPUsim, Barra, compteurs

□ Demain

- Intégration dans gdb



9. Programmation

□ Hier

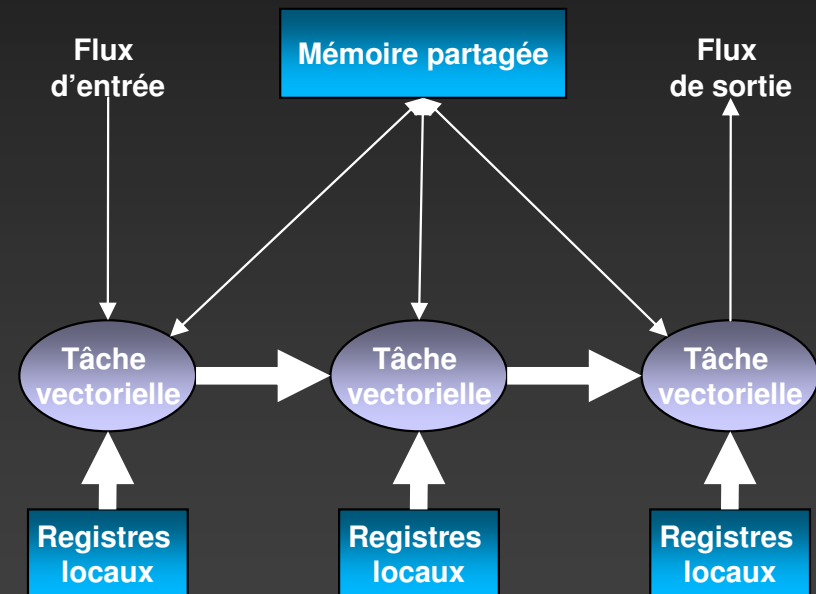
- OpenGL, Direct X

□ Aujourd'hui

- Langages dédiés

□ Demain

- Outils automatiques ou
- Langages dédiés



Le pourquoi du langage dédié

(N'engage que moi)

□ Flux

- Notion de 1:1 entre Input : Output
- I/O : opérations pipelinées
 - Matériel cache les latences
 - Lectures/écritures groupées
- ⇒ Meilleure utilisation des bus
- ⇒ Exhibe naturellement du parallélisme de donnée

□ Mémoire

- Découpage en mémoires locales (registres, ...) et globales (communication, stockage, ...)
- ⇒ Plus grand nombre de processeurs vectoriels pour un coût raisonnable

Flux et tâches

- Tâches vectorielles (kernel)
 - Fonctions à appliquer sur chaque donnée
 - Map, gather/scatter, reduce
 - Utilisation de structures
 - Kernel pipelinée
connexion par les données
 - Gestion des dépendances inter-kernel :
 - Exécution dans le désordre par le scheduler de kernel
 - ⇒ Minimise l'utilisation des registres locaux
 - ⇒ Concentre naturellement le calcul dans des fonctions spécialisées

10. Consommation

□ Hier

- Pas de problème

□ Aujourd'hui

- GPU(GTX280):
4 SP GFLOP / Watt
- CPU(core i7 960):
0.8 SP GFLOP / Watt

□ Demain

- Power gating,...
- Green HPC (*Green500*)



Prospectives :

L'erreur est humaine mais un véritable désastre nécessite un ordinateur.

Numerical Precision How Much is Enough - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.scientificcomputing.com/article-hpc-Numerical-Precision-How-Much-is-Enough-063009.aspx

Wikipedia (en)

Most Visited Bienvenue David (70) Yahoo! France Homepage of David... DB BAHN - Corresp... Petites annonces gr...

(Non lus : 1... David Defo... asr-forum - ... asr-forum - ... GPGPU.org ... Tag: Numer... Numerical ... Fisher Price... The Journal ... (Non lus : 1...)

Scientific COMPUTING INFORMATION TECHNOLOGY FOR SCIENCE

Search Popular Searches: lms, visualization, chemistry, statistics, hpc

INFORMATICS HPC DATA ANALYSIS DATA SOLUTIONS LIMS GUIDE MULTIMEDIA NEWSLETTERS HOME

JOB SEARCH WHITE PAPERS SUBSCRIBE DIGITAL LIBRARY ADVERTISE EDITORIAL CONTACT ABOUT US

US

HPC

- Clusters
- Grids
- Servers
- Supercomputers
- Workstations

Home > HPC > Numerical Precision: How Much is Enough?

Numerical Precision: How Much is Enough?

As we approach ever-larger and more complex problems, scientists will need to consider this question

Rob Farber

The advent of petascale computers and teraflop-per-board graphics processors has raised the question of "how do we know that anything we compute is correct?" Numerical errors can quickly accumulate when performing a trillion to thousands of trillions of floating-point operations per second due to approximations, rounding, truncation errors and other concerns.

This problem confronts every scientist and applications developer because of the speed of current computational hardware. In asking for even more capable computational systems, we might be caught by the adage, "beware of what you ask for because you might get it."

While not many people will gain access to a petascale supercomputer in the next few years, GPU computing is becoming ever more ubiquitous. NVIDIA states they now have an installed base of over 100 million CUDA-capable graphics processors. With teraflop-per-board capability, graphics processors have dramatically increased the computational capabilities available for scientific calculations — and at a commodity price-point. A challenge with GPU computing is that peak performance for the current generation of hardware can only be achieved when using single-precision, 32-bit, floating-point arithmetic. The use of double-precision, 64-bit arithmetic will result in a significant decrease in performance. As a result, anyone planning to use the current generation of graphics processors for scientific computation must consider the question "how important is single-precision compared to double-precision (64-bit) arithmetic for my application?" Happily, some GPUs have limited 64-bit floating-point capability, which opens the possibility of performing multi-precision calculations where the 64-bit operations are only used sparingly when the additional precision might make a difference — say to calculate a sum of a vector or similar such operations.

Numerical accuracy is one of those opaque areas of scientific computing that people try to solve by using the hammer of 64-bit arithmetic to fix the problem. Generally, the problem is caused by thinking that more bits of precision are better and acknowledging that, unfortunately, any number of bits of precision is never really quite enough. The very real fear behind this thinking is that too low a precision can introduce non-physical artifacts into physical simulations, cause important criticality phenomena to be missed, or result in the application exhibiting other undesirable or pathological behavior.

Compatibility with legacy software is also a significant problem. For these applications, producing the exact same result as other computers is a paramount concern when evaluating newer hardware and compilers.

Alistair Rendell provided some nice examples at his SciDAC talk last summer "Build Fast, Reliable, and Adaptive Software for Computational Science." In his talk, he discussed Rump's example, a well-know demonstration that illustrates how increasing floating-point precision does not necessarily equate to greater accuracy. As can be seen

Most Viewed Content

- Bird-Eating Fanged Frog Found in Greater Mekong
- Solar Wind Hits Earth like Fire Hose
- A Trillion Triangles: New code cracks ancient math problem
- Black Holes, Ripples and Galaxies, Oh My!
- Amazing, Interactive, Panoramic, 360-view of Entire Night Sky Unveiled
- Ancient, Bizarre Ghostshark a New Species
- Superheavy Element 114 Confirmed: A Stepping Stone to the Island of Stability
- NASA Maps Provide New Way of Seeing the Moon
- Supervolcano Rosetta Stone Discovered in Alps
- Nullarbor Fireball Cameras Find Rare Meteorite

SCIENTIFIC COMPUTING Webcast Series

Extreme Data Storage: Next-gen Sequencing Solutions

Sponsored by: hp

ON DEMAND NOW

REGISTER FREE

SITE SPONSORS

STARLIMS

Agilent Technologies

Waiting for m1.2mdn.net...