Financial computing on GPUs

Mike Giles

mike.giles@maths.ox.ac.uk

Oxford-Man Institute for Quantitative Finance

Oxford University Mathematical Institute

TradeTech, April 22-23, 2009

Acknowledgments: G. Heber, A. Lee, A. Ng, A. Pati, X. Su, V. Sundaresh, C. Yau

and support from EPSRC, Microsoft, NVIDIA, TCS/CRL

Intel CPUs

- move to faster clock frequencies stopped due to high power consumption – big push now is to multicore chips
- current chips have up to 4 cores, each with a small SSE vector unit (4 `float` or 2 `double`)
- in next 2 years, “Westmere” likely to go up to 10 cores with AVX vectors twice as long
- technologically, many more cores are possible, but will the applications demand it, or is future direction towards low-power low-cost mobile CPUs?
- key point is that cores are general purpose, independent, able to execute different processes simultaneously

GPUs

- many-core chips (up to 240 cores on NVIDIA chips)
- simplified logic (minimal caching, no out-of-order execution, no branch prediction) means most of the chip is devoted to floating-point computation
- usually arranged as multiple units with each unit being effectively a vector unit, all cores doing the same thing at the same time, and all units executing the same program
- very high bandwidth (up to 140GB/s) to graphics memory (up to 4GB)
- not general purpose – aimed at naturally parallel applications like graphics and Monte Carlo simulations

GPU vendors

- NVIDIA: up to 30×8 cores at present
- AMD (ATI): comparable hardware, but poor software development environment at present
- IBM: Cell processor has 1 PowerPC unit plus 8 SPE vector units – relatively hard to program
- Intel: “Larrabee” GPU due out in Q1 2010, with 16-24 unit each with a vector unit – software support for first-generation product not yet clear

High-end HPC

- RoadRunner system at Los Alamos in US
  - first Petaflop supercomputer
  - IBM system based on Cell processors
- TSUBAME system at Tokyo Institute of Technology
  - 170 NVIDIA Tesla servers, each with 4 GPUs
- GENCI / CEA in France
  - Bull system with 48 NVIDIA Tesla servers
- within UK
  - Cambridge is getting a cluster with 32 Teslas
  - other universities are getting smaller clusters

Use in computational finance

- BNP Paribas has announced production use of a small cluster
  - 2 NVIDIA Tesla units (8 GPUs, each with 240 cores)
  - replacing 250 dual-core CPUs
  - factor 10x savings in power (2kW vs. 25kW)
- lots of other banks doing proof-of-concept studies
  - my impression is that IT groups are very keen; quants are concerned about effort involved
- I'm working with NAG to provide a random number generation library to simplify the task

Finance ISVs

Several ISV's now offer software based on NVIDIA's CUDA development environment:

- SciComp
- Quant Catalyst
- UnRisk
- Hanweck Associates
- Level 3 Finance
- others listed on NVIDIA CUDA website

Many of these are small, but it indicates the rapid take-up of this new technology

Programming

Big breakthrough in GPU computing has been NVIDIA's development of CUDA programming environment

- C plus some extensions and some C++ features
- host code runs on CPU, CUDA code runs on GPU
- explicit movement of data across the PCIe connection
- very straightforward for Monte Carlo applications, once you have a random number generator
- significantly harder for finite difference applications
- see example codes on my website

Programming

Next major step is development of OpenCL standard

- pushed strongly by Apple, which now has NVIDIA GPUs in its entire product range, but doesn't want to be tied to them forever
- drivers are computer games physics, MP3 encoding, HD video decoding and other multimedia applications
- based on CUDA and supported by NVIDIA, AMD, Intel, IBM and others, so developers can write their code once for all platforms
- first OpenCL compilers likely later this year
- will need to re-compile on each new platform, and maybe also re-optimise the code – auto-tuning is one of the big trends in scientific computing

My experience

- Random number generation (mrg32k3a/Normal):
  - 2000M values/sec on GTX 280
  - 70M values/sec on Xeon using Intel's VSL library
- LIBOR Monte Carlo testcase:
  - 180x speedup on GTX 280 compared to single thread on Xeon
- 3D PDE application:
  - factor 50x speedup on GTX 280 compared to single thread on Xeon
  - factor 10x speedup compared to two quad-core Xeons

GPU results are all single precision – double precision is up to 4 times slower, probably factor 2 in future.

Why GPUs will stay ahead

Technical reasons:

- SIMD cores (instead of MIMD cores) means larger proportion of chip devoted to floating point computation
- tightly-coupled fast graphics memory means much higher bandwidth

Commercial reasons:

- CPUs driven by price-sensitive office/home computing; not clear these need vastly more speed
- CPU direction may be towards low cost, low power chips for mobile and embedded applications
- GPUs driven by high-end applications – prepared to pay a premium for high performance

What is needed now?

- more libraries and program development tools to reduce programming effort
- more ISV application codes
- more education / training in parallel computing in universities
- fast development of the OpenCL standard and compilers
- continued 10x superiority in price/performance and energy efficiency relative to CPUs

Further information

LIBOR and finite difference test codes
`www.maths.ox.ac.uk/~gilesm/hpc/`

NAG parallel random number generator
(John Holden, Anthony Ng, Robert Tong)
`info@nag.co.uk`

NVIDIA's CUDA homepage
`www.nvidia.com/object/cuda_home.html`

Microprocessor Report article
`www.nvidia.com/docs/IO/47906/220401_Reprint.pdf`