FOUNTAINHEAD

GPU Supercomputing for Finance

A quiet revolution on Wall Street

Microsoft, New York, 24 January 2011

Microsoft®

FOUNTAINHEAD

Outline for the Talk

The talk is in two parts:

A. An overview of GPU supercomputing technology
   - The power, and promise, of GPU supercomputing.
   - Future trends to watch.

B. Application areas for GPU technology in finance
   - Where and how GPU is being used now.
   - Future trends to watch.

FOUNTAINHEAD

Outline - Part A

The power, and promise, of GPU supercomputing:

- GPUs versus CPUs.
- GPU as a co-processor for the CPU.
- Why GPUs? Why now?
- Speed, Speed & Speed.
- Data, Data & Data.
- Real-time, Real-time & Real-time.
- 10x, 10x, 10x, 10x, 10x.
- Cloud+GPU.

FOUNTAINHEAD

Outline - Part B

Application areas for GPU technology in finance:

- GPUs versus CPUs
- Questions & answers

FOUNTAINHEAD

Talk Admin

- Talk materials available online & extra resources.
- Vendors in the room: Microsoft (obviously!), Nvidia & others.
- Questions and answers ~ as we go along, and at the end.
- 10 day post-talk discussion forum.
- On-the-spot prize draw for the best question of the session.
- On-the-spot prize draw for handing in the questionnaire.
- Future events:
    - HPC / GPU enthusiast meetup 6pm tonight.
    - 3-day training course, 28-30 March 2011: “Microsoft HPC for Finance”.
    - Additional “brown bag” lunchtime talks.

FOUNTAINHEAD

Part A

The power, and promise, of GPU supercomputing.

FOUNTAINHEAD

The Rise of the GPU

- Demo - Pong (circa 1972).
- Demo - Bioshock (circa 2010).

FOUNTAINHEAD

GPU Supercomputing

- GPU (Graphics Processing Unit).
- Massively parallel.
- Fantastic floating point performance (~1TFlop SP, ~500MFlop DP). IEEE floating point.
- For comparison, CPU ~100MFlop DP.
- GPU memory ~6GB. ECC memory.
- Speedup x10, x100, x300 … x1000 in exceptional cases.

FOUNTAINHEAD

Why GPUs?

- More computational power than CPUs.
- CPUs are getting faster, GPUs are getting faster quicker!
- Scalable (many cards can fit in one PC).
- More power efficient.
- Far cheaper on a $ per GFlop basis.

FOUNTAINHEAD

Demo: GPU parallelism

GPUs are massively parallel:

- Demo - Mythbusters.
- Key takeaways:
  - GPU as co-processor to the CPU.
  - Sequential versus parallel tasks.
  - GPU is a massively parallel processor.

FOUNTAINHEAD

Why now?

- GPGPU (General Purpose GPU) programming has been around for a while. Folding@Home added GPU in 2006.
- Previously difficult to program:
    - Low level bit twiddlers only.
- Now, tools are much, much better.
- C/C++ tool chain. Debugger and IDE.
- Choice of APIs:
    - CUDA (Nvidia) / Nsight (IDE)
    - DirectCompute (part of DirectX, Microsoft)
    - OpenCL (Khronos)
    - ATI Stream (AMD)

FOUNTAINHEAD

Speed, Speed & Speed

- In our business, time really is money!
- Three most important things about trading: speed x 3.
- Nothing so impresses traders as ***SPEED.***
- Imagine pricing & structuring deals quicker than others.
- Imagine risk going from overnight to real-time.
- Low latency development (iterate quicker, faster, better).
- 24hrs @
  - x10 speedup --> 2.4 hours
  - x100 speedup --> 15 minutes
  - x1000 speedup --> 90 seconds

FOUNTAINHEAD

Data, Data & Data

- Storage is (almost) free.
- Bandwidth is (almost) free (i.e. moving data around is free).
- Data growth in the enterprise and in finance increasing.
- Electronic & high-frequency trading is exploding.
- Making sense of the data -- how to convert data into actionable knowledge?
- Number crunching and compless event processing.
- Visualization.

FOUNTAINHEAD

Real-time, Real-time & Real-time

- Everything is moving to real-time.
- Everything is moving towards continuous time.
- Everything is moving towards mobility (anywhere, anytime).
- Data & processing power must be brought together.
- GPUs co-locate data with number crunching.

FOUNTAINHEAD

10x, 10x, 10x, 10x, 10x

- Rule of 10 (10x better or/and 10x cheaper)
- 10x speedup.
- 10x cheaper.
- 10x less space.
- 10x less power.
- 10x less cooling (related to power).

FOUNTAINHEAD

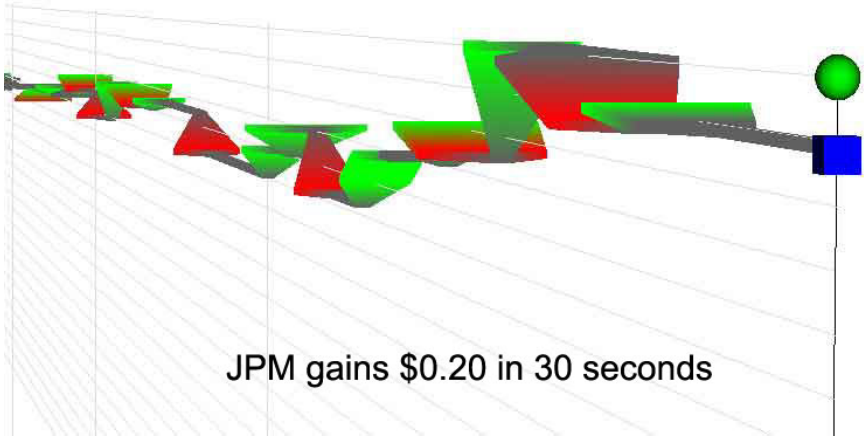
GPUs Getting Faster

- GPUs are getting faster (512 cores).
- CPUs are getting faster (4 cores).
- But GPUs have the advantage.
- Gap between GPUs and CPUs is widening for pure number crunching.
- But GPU is not a replacement for CPU. Rather, the GPU is a co-processor that augments the CPU with massive amounts of parallel number crunching capability.

FOUNTAINHEAD

Part B

Application areas for GPU technology in finance.



FOUNTAINHEAD

Applications in Finance

- Pricing, especially of complex assets. (Hanweck.)
- Risk analysis. (Overnight to real-time risk.)
- Algorithmic trading. (Pre/post trade analysis.)
- High-frequency trading. (Complex event processing).
- Tick data. (Added-value data feeds in real-time).
- Data mining. (Machine learning.)
- Trading strategy prospecting. (Backtesting too.)
- Data visualization. (Making sense of it all.)
- … anything that needs to crunch numbers, or process vast amounts of data … fast!

FOUNTAINHEAD

Applications - Added-Value Data

Data visualization is in its infancy in finance:

- Data sets are so large, and the velocity of data is so great these days, how are we to make sense of it all.
- Human vision linked to the human brain is still the best information processor and pattern recognition system we have!
- Demo: STOC.

FOUNTAINHEAD

Applications - Visualization

Data visualization is in its infancy in finance:

- Data sets are so large, and the velocity of data is so great these days, how are we to make sense of it all.
- Human vision linked to the human brain is still the best information processor and pattern recognition system we have!
- Demo: STOC.

FOUNTAINHEAD

A Cautionary Warning

Apophenia

Apophenia is the experience of seeing meaningful patterns or connections in random or meaningless data. The term was coined in 1958 by Klaus Conrad, who defined it as the "unmotivated seeing of connections" accompanied by a "specific experience of an abnormal meaningfulness".

FOUNTAINHEAD

Questions & Answers

FOUNTAINHEAD

Speaker: Andrew Sheppard

Andrew Sheppard is a financial consultant with extensive experience in quantitative financial analysis, trading-desk software development, and technical management. Most recently, from 2006 to 2010, Andrew worked at a New York multi-strategy hedge fund. He also was the manager of an innovative software company based in London that was owned by the hedge fund but run independently. For more than two years, Andrew has been an active developer of GPU (CUDA) massively parallel software in C/C++ for real-time financial trading and risk. Andrew is also the author of the forthcoming book "Programming GPUs", to be published by O'Reilly (www.oreilly.com).