Hanweck Associates, LLC

Accelerating Quantitative Financial Computing with CUDA and GPUs

NVIDIA GPU Technology Conference

San Jose, California

March 20, 2013

Gerald A. Hanweck, Jr., PhD

CEO, Hanweck Associates, LLC

Hanweck Associates, LLC
30 Broad St., 42nd Floor
New York, NY 10004
www.hanweckassoc.com
Tel: +1 646-414-7274

Agenda

- Why GPUs in Quant Finance?
- Overview of GPU Technology
- GPU Quant Finance Case Studies
  - 1: Real-Time Option Analytics
  - 2A: Stochastic Volatility Modeling
  - 2B: Stochastic Volatilty + Jumps Modeling
  - 3A: Large-Scale Interest-Rate Swaps Value-at-Risk (VaR)
  - 3B: Large-Scale Monte Carlo VaR
  - 3C: Large-Scale Parametric VaR
  - 4: Pricing a Basket Barrier Option
  - 5: Random Number Generation
- Concluding Remarks

Q: What Is the Biggest Problem Facing the Capital Markets Today?

A: Intraday and Real-Time Risk Management

- Increasingly complex, global structured products
- Higher correlation risk and systemic risk
- Greater regulatory requirements
- Massive grid-computing infrastructure costs
- Exploding market-data message rates

GPU Acceleration in Quant Finance

Random number generation	Trees and lattices	Equity derivatives
Path generation	Matrix algebra	Interest-rate derivatives
Payoff function acceleration	Numerical integration	Credit models
Statistical aggregation	Fourier transforms	Exotics and hybrids

Investment Banks	Asset Managers	Pension Plans
Hedge Funds	Insurance Companies	Mortgage Servicers
Prop Trading	Automated Market Making	Risk Managers

- ***10x faster*** "dollar-for-dollar" than conventional CPU computing.
- ***10x faster*** means:
  - overnight → over lunch
  - over lunch → get a cup of coffee
  - get a cup of coffee → don't blink!
- Better risk management.
- Reduce total cost of ownership and infrastructure cost explosion.

GPUs in Quant Finance

Bloomberg

Hanweck Associates, LLC

CITADEL

CRÉDIT AGRICOLE

JPMorgan

WestLB

BNP PARIBAS

nag

MUREX™

STANDARD LIFE

AON BENFIELD

The MathWorks

Wolfram Mathematica

Interactive Data

THOMSON REUTERS

SciComp Inc.

Source: NVIDIA

Hanweck Associates' Volera™ GPU-Accelerated Product Line

VOLERAFEED™

Real-time, low-latency datafeed of options implied volatilities and greeks covering global markets, powered by Hanweck Associates' Volera™ GPU-accelerated engine. VoleraFEED powers:

ISE Implied Volatility & Greeks Feed™

NYSE Technologies. ***Options Analytics™***

Premium Hosted Database™

Hosted historical and real-time "tick-level" database service of equity and options prices and analytics, with 300+ TB of data stored in an enterprise-scale cloud. "Data-and-Analytics-as-a-Service" paradigm.

In partnership with ISE

VOLERARISK™

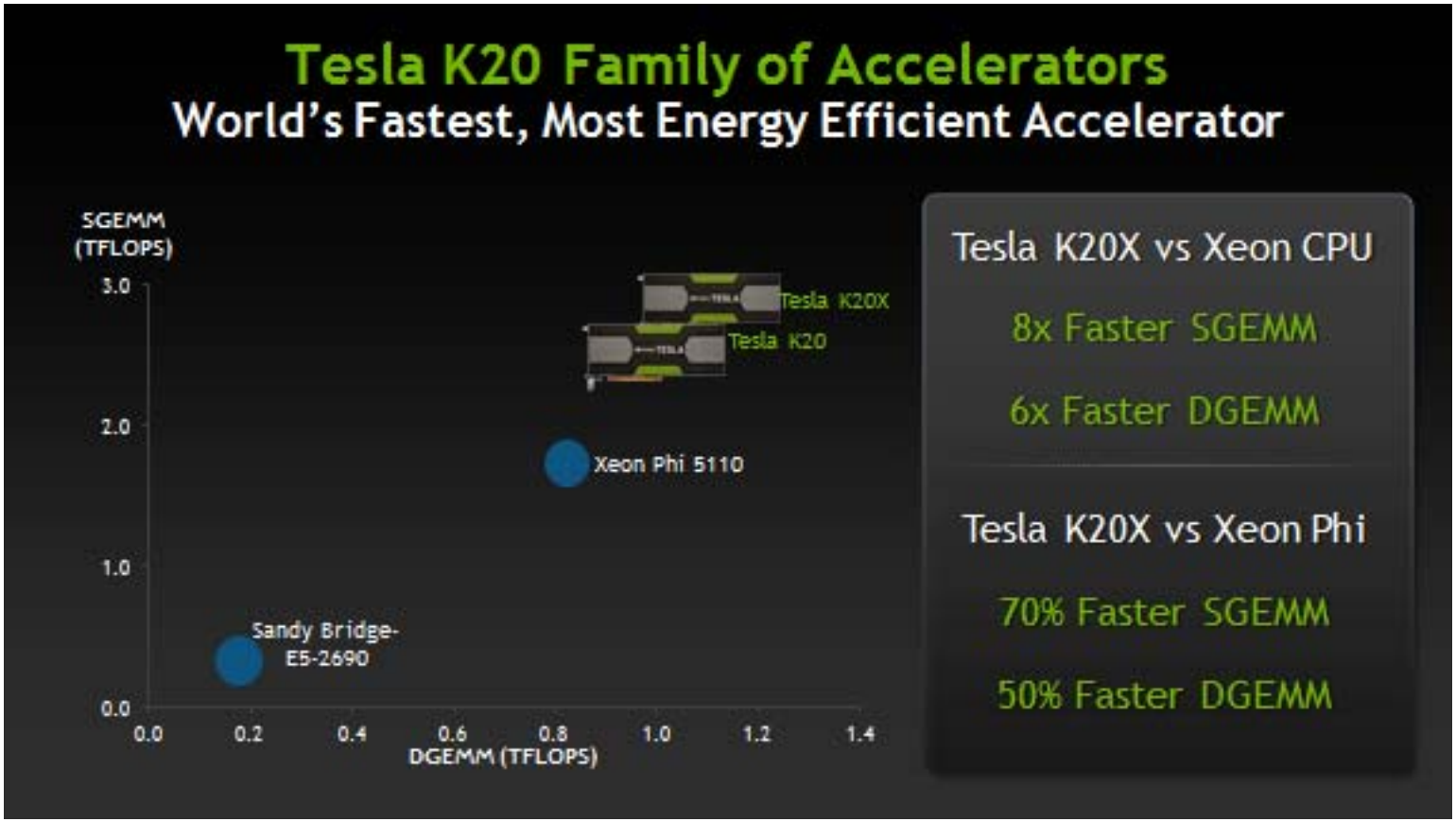
High-performance, real-time, large-portfolio pre/post-trade risk and portfolio margining, powered by Hanweck Associates' Volera™ GPU-accelerated engine.

Options Volatility Service™

Historical, end-of-day options analytics database covering more than 6,000 U.S. companies over the past 12 years.

In partnership with Interactive Data

NVIDIA GPU Performance



Source: NVIDIA

NVIDIA GPU Architecture

2,688 CUDA cores*

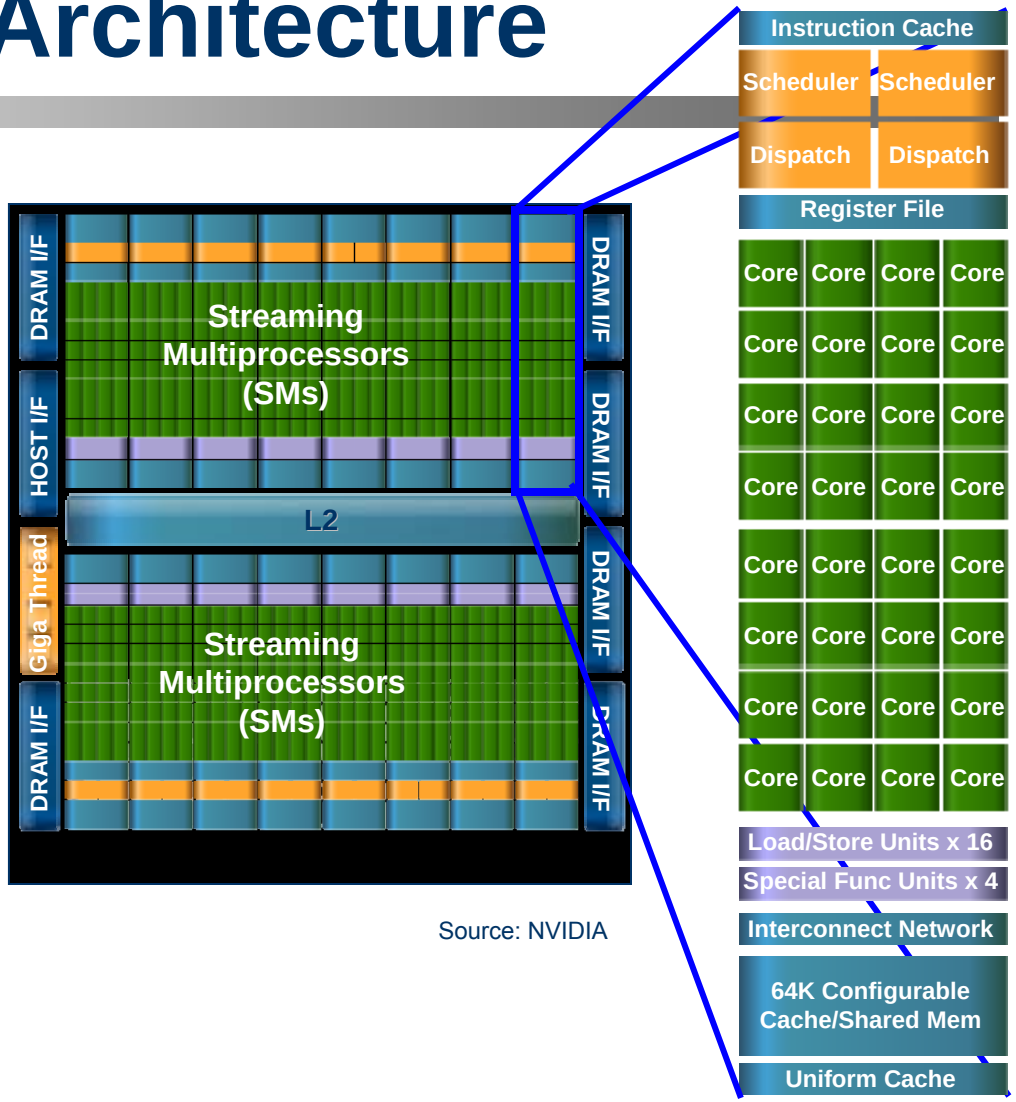
3.95 Tflops single-precision

1.31 Tflops double-precision

6 GB ECC DRAM

250 GB/sec DRAM bandwidth

64 KB of RAM for shared memory and L1 cache (configurable)



Source: NVIDIA

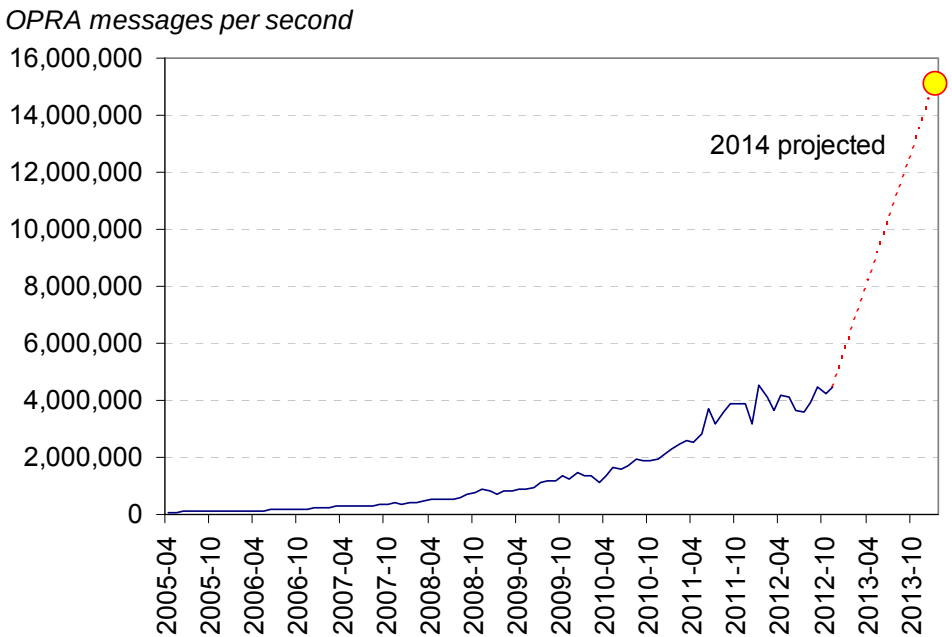
* Tesla K20X series GPU

Case Study #1: Real-Time Options Analytics

Real-time, low-latency implied volatilities and Greeks (binomial tree)

Hanweck Associates' VoleraFEED™ Real-Time Options Analytics Engine

- real-time, low-latency implied volatilities and Greeks
- U.S. OPRA universe:
  - 530,000 options on 3,800 stocks
  - 2012: 4.6 million msg/sec peak
  - 2014: 15.1 million msg/sec peak*
- 128-step CRR binomial tree
  - discrete dividends (escrowed)
  - discount and borrow curves
  - bid/ask/mid implieds & mid Greeks



Average Chain Calculation
20 milliseconds**

* OPRA Jan 2014 projection
** 4 NVIDIA Kepler K10s

Case Study #2A: Stochastic Volatility Modeling

Price options under the Heston* stochastic volatility model:

- European-style call and put options
- Solution involves numerical integration of complex-valued integrands for each distinct strike and expiry
- Simpson's rule with dynamic integration ranges
- Hardware: NVIDIA C2070 GPU vs. Intel Xeon E5640 (2.67 GHz)

Option Pricing under Stochastic Volatility:

2,000 option pricings per second

(70x faster than a single CPU core)

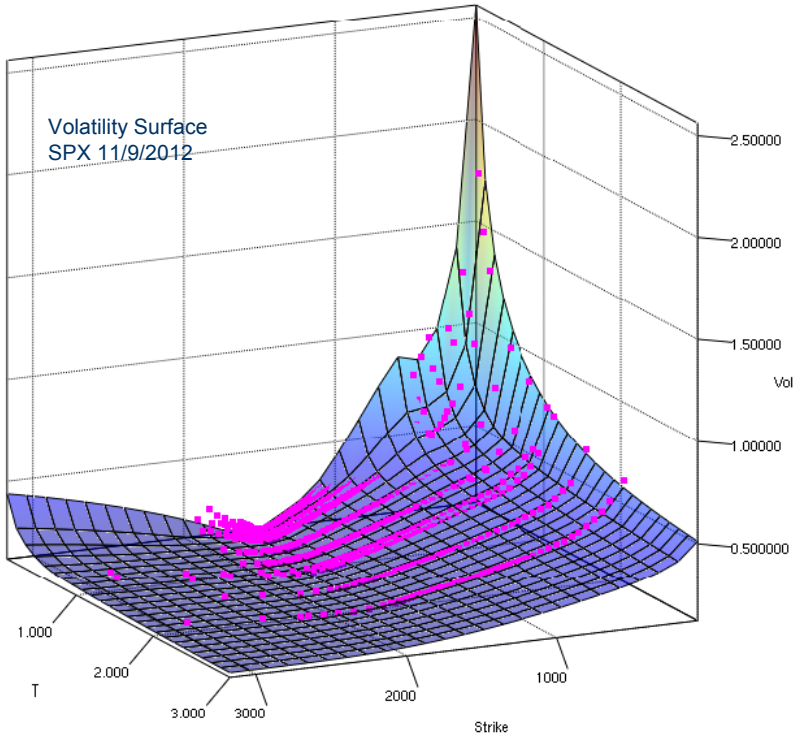
* Heston, Steven L. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options," *The Review of Financial Studies, 6(2), 1993, pp. 327-343.*

Case Study #2B: Stochastic Volatility+Jumps Modeling

Price options under the Bates* stochastic volatility+jumps model

- European-style call and put options
- Solution involves FFT integration of the characteristic function for each expiry across range of strikes**
- NVIDIA C2090 GPU, cuFFT 4.0

Option Pricing under Stochastic Volatility + Jumps:
1,200 expiry pricings per second (double precision, 2^15 nodes)



* Bates, David S. "Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options," *The Review of Financial Studies,* 9(1), 1996, pp. 69-107.
** Carr, Peter *et al*. "Option Valuation Using the Fast Fourier Transform," *Journal of Computational Finance,* 2(4), 1999, pp. 61-73.

Case Study #3A:
Large-Scale Interest-Rate Swaps Risk

Calculate Value-at-Risk (VaR) for a large-scale portfolio of interest-rates swaps (IRS):

- 30,000 distinct IRS positions.
- 1,300 Monte Carlo paths representing yield-curve shocks.
- Full cash-flow and day-count revaluation in each path.
- Calculation of VaR and expected shortfall.
- Hardware: 1 NVIDIA C2090 GPU w/ 8-core Xeon host server.

Large-Scale, IRS VaR:

10 seconds

(vs. 45 minutes on a CPU-based compute grid)

Case Study #3B:
Large-Scale Monte Carlo Risk

System for real-time risk monitoring of large portfolios of listed options:

- 350,000 distinct options representing the listed universe.
- 10,000 Monte Carlo paths generated from factor shocks (2,500 factors) on 3,500 underlying stocks and indices.
- Hundreds of large portfolios.
- Full binomial-tree revaluation of each option in each path.
- Calculation of VaR and expected shortfall under multiple correlation scenarios.
- Hardware: 24 NVIDIA C2050 GPUs w/ 8-core Xeon host servers.

Large-Scale, Full-Revaluation Monte Carlo VaR:

< 1 minute

(hundreds of times faster than a single CPU core)

Case Study #3C:
Large-Scale Parametric VaR

System developed for a large investment bank to evaluate parametric factor VaR on millions of private client portfolios, with aggregation across accounts, advisors, offices and regions:

- 1.25 million portfolios
- 2,000 factors covering 400,000 global assets
- Hardware: 12 NVIDIA C2050 GPUs w/ 8-core Xeon host server

Large-Scale Parametric Factor VaR:
2 minutes
(hundreds of times faster than a single CPU core)

Case Study #4: Basket Barrier-Option (Monte Carlo)

Valuing a basket barrier-option

Monte Carlo simulation of a multi-factor, local-volatility model for pricing lookback structures:

- 4 underlying assets
- 100,000 MC paths
- 750 steps per path

	CPU[1] Time (sec)	GPU[2] Time (sec)
Stage 1: RNG	20.2	0.139
Stage 2: Path Gen	23.5	0.181
Stage 3: Payoffs	8.5	0.223
Stage 4: Stats	9.3	0.061
Total	61.9	0.604
Performance Gain		**102x faster**

Realistic "dollar-for-dollar"[3] performance gain: **12x faster**

1. One core of Intel Xeon L5640 @ 2.26GHz
2. One NVIDIA Fermi C2070 GPU
3. Performance adjusted for: core/GPU density, amortized hardware costs, power/cooling costs, etc.

Case Study #5: Random Number Generation

Implementation of a GPU-parallel Monte Carlo simulation and random-number generator for a major investment bank:

- Monte Carlo simulation of a multi-factor, local-volatility model for pricing lookback structures
- Implementation of an efficient GPU-parallel random-number generator*
- Hardware: 1 NVIDIA C2070 GPU w/ 8-core Xeon host server

GPU-Parallel Monte Carlo:
2.5 billion normal random numbers per second
(200x faster than a single CPU core)

* L'Ecuyer, Pierre; Richard Simard; E. Jack Chen and W. David Kelton, "An Object-Oriented Random-Number Package with Many Long Streams and Substreams," *Working Paper, December 2000*.

Random-Number Generation

Large base of existing GPU code and resources:

NVIDIA's cuRAND RNG library

http://developer.nvidia.com/curand

- L'Ecuyer (MRG32k3a), MTGP Mersenne Twister, XORWOW PRNG and Sobol QRNG

NVIDIA's CUDA SDK sample code:

- Niederreiter, Sobol QRNGs, Mersenne Twister
- Monte Carlo examples

GPU Gems 3 and *GPU Computing Gems* (Emerald Edition)

GPU Gems 3 is available online: http://developer.nvidia.com/object/gpu-gems-3.html

- Tausworth, Sobol and L'Ecuyer (MRG32k3a)
- Monte Carlo examples (*GPU Gems 3*)

Concluding Remarks

Real-time and intra-day risk management is a major problem facing the financial industry today... but it is pushing conventional computing to its limits.

GPUs *are* the way forward. Major financial institutions are using them for quant finance.

Performance gains of *more than 10x – dollar for dollar* – are achievable in practice in many common use cases, which is generally sufficient to offset the costs of new development.

GPU programming in general – and CUDA in particular – push developers to parallelize their code.

Parallelizing quant finance is critical if quant finance software is to take advantage of the advances in many-core hardware.

This presentation has been prepared for the exclusive use of the direct recipient. No part of this presentation may be copied or redistributed without the express written consent of the author. Opinions and estimates constitute the author's judgment as of the date of this material and are subject to change without notice. Information has been obtained from sources believed to be reliable, but the author does not warrant its completeness or accuracy. Past performance is not indicative of future results. Securities, financial instruments or strategies mentioned herein may not be suitable for all investors. The recipient of this report must make its own independent decisions regarding any strategies, securities or financial instruments discussed. This material is not intended as an offer or solicitation for the purchase or sale of any financial instrument.

Additional information is available upon request.