ERM
ENTERPRISE RISK
MANAGEMENT SYMPOSIUM

# GP GPU in Risk Management

Iouri Karpov, Prudential Financial

# From Avatar to GP GPU

**GPU – Graphics Processing Unit**

- Every video game console, computer, and smart phone has one
- Consumer GPU technology driven primarily by video gaming and ever improving HD standard

**GP GPU – General Purpose GPU Computing**

- GPU's are designed for efficient matrix computations
- Offload numerically intensive and parallelizable calculations to GPU
- Used heavily in scientific computing, finance, cryptography

**High Performance GPU Cards**

- Nvidia
  - Tesla cards are prevalent in high performance computing
- AMD Radeon (formerly ATI)
  - Not as popular in HPC, but offers similar hardware benefits

ERM
ENTERPRISE RISK MANAGEMENT SYMPOSIUM

# HPC Needs in Insurance ERM/ALM

**Stochastic Liability Valuation**

- Market consistent pricing
- Statutory reserves

**ALM/Hedging**

- Liability greeks
- Hedge strategy impact/optimization

**Forecasting and Stress-Testing**

- Reserve, capital, and liquidity projections

**Nested Stochastic / Stochastic-On-Stochastic Projections**

- Required in forecasts and ALM strategy

ERM
ENTERPRISE RISK MANAGEMENT SYMPOSIUM

# Why GP GPU?

**Massive Parallelization**

- Typical insurance valuation involves running a significant number of policies with similar payoff definition across multiple economic scenarios

**Technological Alternative to Inforce Reduction/Clustering and LSMC**

**Large Number of Policies = Better Efficiency**

- Monte Carlo pricing convergence and efficiency can be dramatically improved using variance reduction techniques such as independent and importance sampling

**Reduced Runtime and Reduced Costs/Footprint**

- Possible 10x (or greater) reduction in cost while achieving same runtime
- Portability – most “everyday” analytics can be run using only 1-2 cards on a single desktop

ERM
ENTERPRISE RISK MANAGEMENT SYMPOSIUM

# Practical Considerations

**Most Efficient and Flexible Solution Achieved Using CUDA C**

- Full control and efficiency in throughput optimization
- Optimal use of GPU memory architecture

**Generic, Modular Design**

- In practical CUDA C++ implementation, can be achieved using templates
- Device agnostic benefit logic. Same payoff code runs on CPU or GPU. GPU nuances handled in the architecture, not the core policy implementation

**Scaling it Up**

- Interface, and organizing multiple parameter and data inputs
- GPU cluster management
  - Parallelized execution across nodes and devices
  - Job scheduler

ERM
ENTERPRISE RISK MANAGEMENT SYMPOSIUM