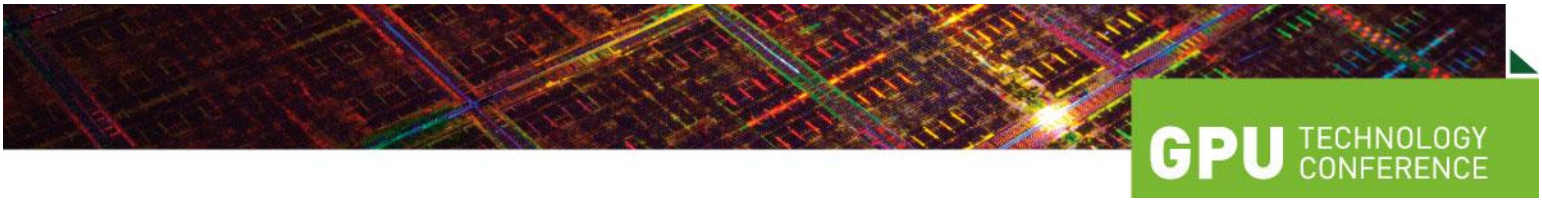

GPU Technology Conference, May 14-17, 2012
McEnery Convention Center, San Jose, California
www.gputechconf.com

Sessions on Databases, Data Mining, Business Intelligence (subject to change)

IMPORTANT: Visit http://www.gputechconf.com/page/sessions.html for the most up-to-date schedule.

S0314 - Efficient k-Nearest Neighbor Search Algorithms on GPUs
Nikos Pitsianis (Aristotle University, Greece), Xiaobai Sun (Duke University)
Day: Tuesday, 05/15 | **Time:** 4:30 pm - 4:55 pm
Topic Areas: Machine Learning & AI; Databases, Data Mining, Business Intelligence; Algorithms & Numerical Techniques
Session Level: Beginner

Come see how to select the k smallest elements from an unsorted list. We present a selection and combination of different algorithms that perform exact k-nearest neighbors search (k-NNS) on GPUs and outperform the competition. In this session we present four different selection algorithms designed to exploit differently the parallelization of the GPU according to the relative size of the corpus data set, the size of the query set and the number of neighbors sought. We show the application of Logo Retrieval with SIFT vector matching on two different GPUs, the Tesla C1060 and the Fermi GTX480.

S0219 - Efficient Top-Down Planning in Business Intelligence
Tobias Lauer (Jedox AG), Alexander Haberstroh (Jedox AG)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:25 pm
Topic Areas: Databases, Data Mining, Business Intelligence; Finance; Algorithms & Numerical Techniques
Session Level: Intermediate

In business intelligence, tasks like corporate planning or what-if analysis complement traditional reporting and analysis. One main difference is that while the latter only read data, the former require the change of possibly large numbers of existing and creation of new data records in the business model, preferably in real time. In this session, we describe the extension of an existing BI tool, Jedox OLAP, by GPU-based parallel algorithms for interactive planning scenarios. Compared to sequential in-memory algorithms, our CUDA approach yields tremendous speedups and can also cope with large amounts of data by using multiple GPUs.

S0427 - Intra-Day Risk-Management with Parallelized Algorithms on GPUs
Partha Sen (Fuzzy Logix)
Day: Tuesday, 05/15 | **Time:** 5:00 pm - 5:50 pm
Topic Areas: Databases, Data Mining, Business Intelligence; Finance; Algorithms & Numerical Techniques; Supercomputing
Session Level: Advanced

The challenge with intra-day risk management is that a very large number of calculations are required to be performed in a very short amount of time. Typically, we may be interested in calculating VaR for 100 to 1000 securities per second based on 100 million potential scenarios. The magnitude of these calculations is not Utopian but it reflects the reality of modern financial institutions and exchanges. In this presentation, we outline how the complex problem of intra-day risk management can be solved using parallelized algorithms on GPUs. The

methodology has been proven in a POC at 2 financial institutions.

S0043 - 30x Faster Regular Expressions on a GPU
David Lehavi (HP)
Day: Tuesday, 05/15 | **Time:** 5:30 pm - 5:55 pm
Topic Areas: Databases, Data Mining, Business Intelligence
Session Level: Advanced

We present a regular expression (regex) engine on a GPU. We utilize the highly parallel architecture of GPUs to accelerate such searches. We believe that previous attempts to utilize the GPU for this task did not fully tap its potential. Regex present imbalanced compute workloads which are very different from common GPU applications (CFD, CG and image processing). Hence, they can teach us general lessons on how to utilize GPUs for more general workloads. Our initial results show 30x improvement in running time relative to single threaded commercial regex engines.

S0035 - GPU Parallelization of Gibbs Sampling: Abstractions, Results, and Lessons Learned
Alireza Mahani (Sentrana)
Day: Wednesday, 05/16 | **Time:** 3:00 pm - 3:50 pm
Topic Areas: Algorithms & Numerical Techniques; Databases, Data Mining, Business Intelligence
Session Level: Intermediate

Monte-Carlo-Markov-Chain (MCMC) estimation of Hierarchical Bayesian (HB) models is not only time-consuming, but also difficult to parallelize due to its sequential (Markovian) nature. We present an abstraction of a widely-used MCMC algorithm, called Gibbs sampling. We define a taxonomy of variable blocks, and for each type of variable block we offer suitable parallelization strategies, along with their corresponding CUDA implementations. For large problems where model estimation may take several hours or days using a single-threaded software, we see speedups in the 30x-100x range, thereby reducing estimation time to a few hours. In addition to lower computation cost relative to MPI-based parallelization, the reduction in estimation time allows for a more interactive modeling experience. We offer an extensive discussion of lessons learned for the broader scientific computing field, including an analysis of tradeoffs between computation costs and development costs, implications of our tradeoff analysis for optimal software development and parallelization, and some practical tips and gotcha's for rookie GPU programmers.

S0038 - Designing Killer CUDA Applications for X86, multiGPU, and CPU+GPU
Robert Farber
Day: Thursday, 05/17 | **Time:** 4:00 pm - 4:25 pm
Topic Areas: Machine Learning & AI; Supercomputing; Databases, Data Mining, Business Intelligence; Computer Vision
Session Level: Intermediate

CUDA redefined software development with 10 to 1000-times faster GPU applications. Now a single CUDA source tree can support the x86 mass market (no GPU required) and 1/3 billion CUDA-enabled GPUs. MultiGPU and CPU+GPU apps utilize all system resources. GPUdirect, UVA, caches, prefetching, ILP (Instruction level Parallelism), automated analysis tools and more offer ease, capability, and performance. The overall impact on software investment, scalability, balance metrics, programming API, and lifecycle will be considered. Working real-time video and other examples from my book, "CUDA Application Design and Development" provide practical insight to enable augmented reality and your killer apps.