ATIP 1st Workshop on HPC in India, Supercomputing 2009

NUMERICAL LINEAR ALGEBRA ON HYBRID ARCHITECTURES:
RECENT DEVELOPMENTS IN THE MAGMA PROJECT

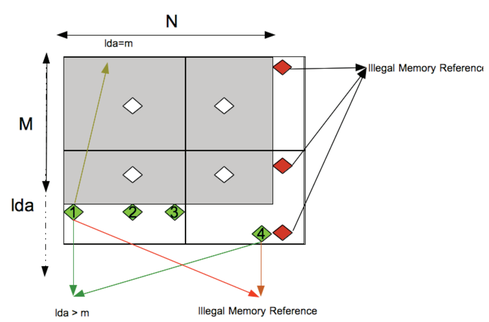
Rajib Nath
Jack Dongarra
Stanimire Tomov
Hatem Ltaief
Peng Du

INNOVATIVE COMPUTING LABORATORY
THE UNIVERSITY of TENNESSEE

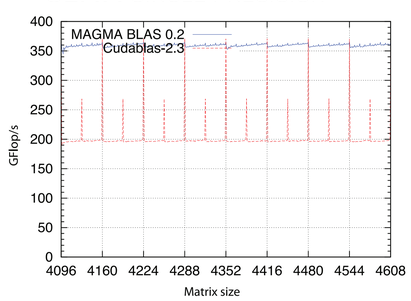
THE MATRIX ALGEBRA FOR GPU AND MULTICORE ARCHITECTURES (MAGMA) PROJECT aims to create a new generation of linear algebra libraries that achieve the fastest possible time to an accurate solution on hybrid/heterogeneous architectures, starting with current multicore+multiGPU systems. To address the complex challenges stemming from the heterogeneity of these systems, their massive parallelism, and the gap in computation vs. CPU-GPU communication speeds, MAGMA research is based on the idea that optimal software solutions will themselves have to hybridize, combining the strengths of different algorithms within a single framework.

GPU KERNEL DEVELOPMENT

Extra Flop Computing with Pointer Adjustment



GEMM SINGLE PRECISION



CPU Intel Xeon 2.33 GHz, 8 cores, s/d gemm peak 128/65 GFlop/s
GPU NVIDIA GTX280 1.33 GHz, s/d gemm peak 375/75 GFlop/s

Whenever it is Needed & Whatever it Takes

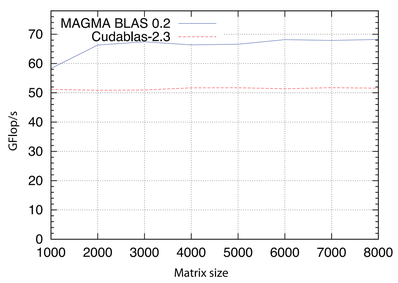
- Data Coalescing
- Memory Optimal Computation
- Computation Partition
- Enforcing Data Parallel Mode
- Hand and Auto tuning
- Faster than Existing CUDA-2.3

Auto-Tuning
The Beauty and The Beast behind MAGMA

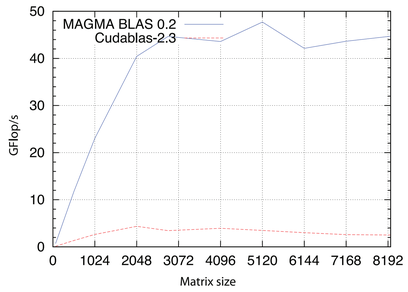
- Code Generation
- Timing
- Exhaustive Search
- Loop Transformation Techniques
- Hardware Dependent Parameters

MAGMA BLAS

DGEMM tuned for rectangular matrices
[e.g.,Matrix Size Nx32]

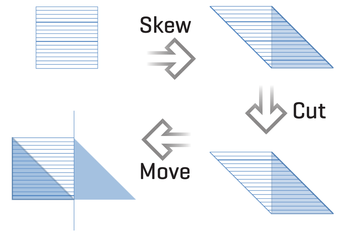


SYMV SINGLE PRECISION



CPU Intel Xeon 2.33 GHz, 8 cores, s/d gemm peak 128/65 GFlop/s **GPU** NVIDIA GTX280 1.33 GHz, s/d gemm peak 375/75 GFlop/s

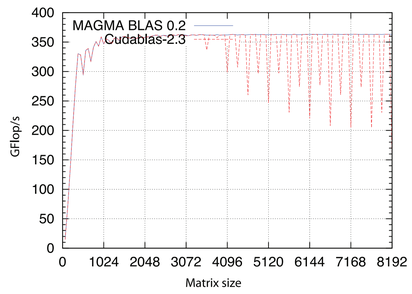
- GEMM Tuned for Rectangular Matrices
- SYRK, GEMV, and SYMV
- TRSM of High Parallelism/Performance
  (Trade-off for Numerical Stability)



Reordering the Computation
Circular Loop Skewing

Problem with GPU Memory Layout
Dips in Single Precision

SGEMM(AB^T) with Circular Loop Skewing

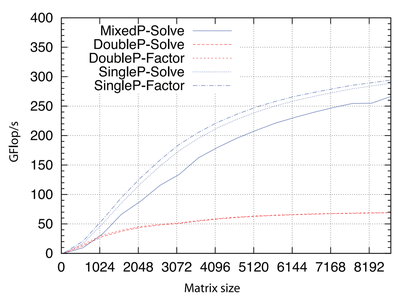


CPU Intel Xeon 2.33 GHz, 8 cores, s/d gemm peak 128/65 GFlop/s
GPU NVIDIA GTX280 1.33 GHz, s/d gemm peak 375/75 GFlop/s

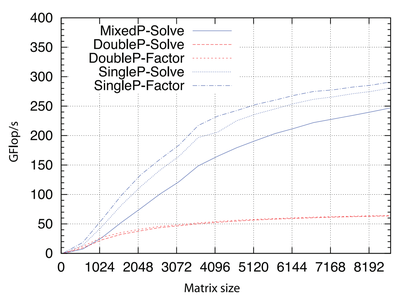
MIXED PRECISION ITERATIVE REFINEMENT

EXPLOITING SINGLE PRECISION ARITHMETIC TO GET DOUBLE PRECISION ACCURACY (1 GPU AND 1 CPU)

Solving A x = b using LU Factorization



Solving A x = b using Cholesky Factorization



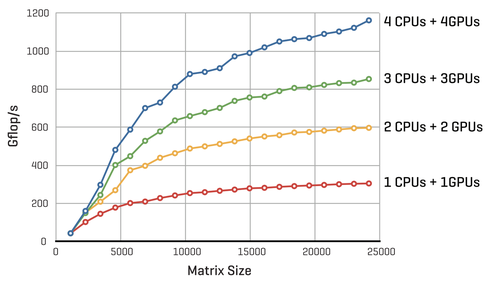
CPU Intel Xeon 2.33 GHz, 8 cores, s/d gemm peak 128/65 GFlop/s **GPU** NVIDIA GTX280 1.33 GHz, s/d gemm peak 375/75 GFlop/s

SCALABLE HYBRID CHOLESKY

MULTI GPU AND MULTI CORE ALGORITHM

- Block Data Layout
- PLASMA Scheduler
- Memory Optimal
- Optimized Communication
- Hybrid & GPU Kernels

Cholesky Factorization SINGLE PRECISION



CPU AMD Opteron 1.8GHz, 4 cores **GPU** Tesla C1070 1.44GHz, 4 GPUs

CURRENT AND FUTURE FOCUS

- Multicore and Multi GPU Algorithms
- Resource Sharing
- Load Balancing
- Self-Adjusting Scheduler
- Performance Modeling
- Auto-Tuning in Hybrid Framework
- Computation with Limited GPU Memory

FIND OUT MORE AT http://icl.eecs.utk.edu/magma/

A COLLABORATION OF
THE UNIVERSITY of TENNESSEE
Berkeley UNIVERSITY OF CALIFORNIA
University of Colorado Denver

WITH SUPPORT FROM
The MathWorks™
Microsoft
NVIDIA

SPONSORED BY
NSF
CITR CENTER FOR INFORMATION TECHNOLOGY RESEARCH