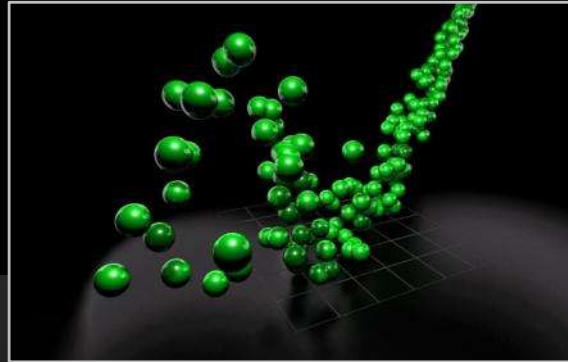
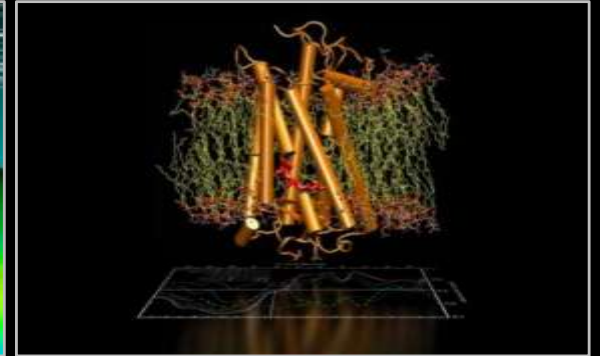


TESLA

GPU Computing



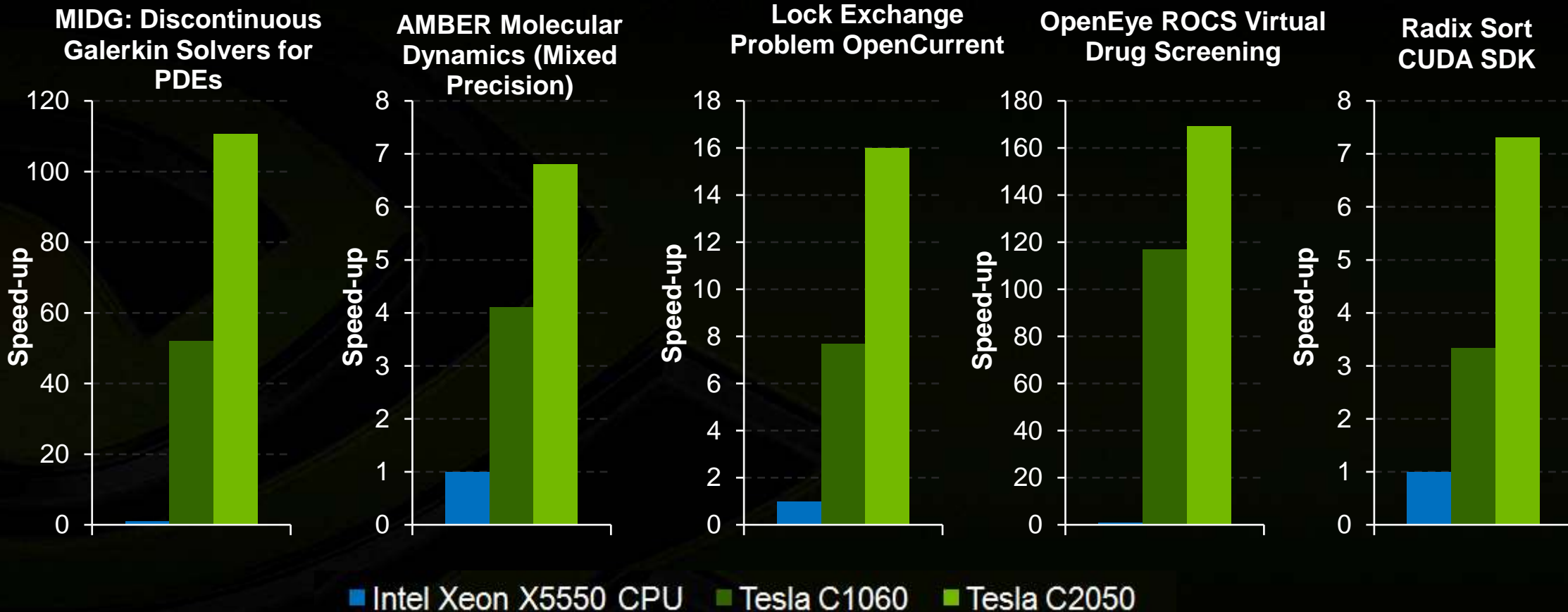
Tesla C2050 Performance Benchmarks

Tesla C-Series Workstation GPUs

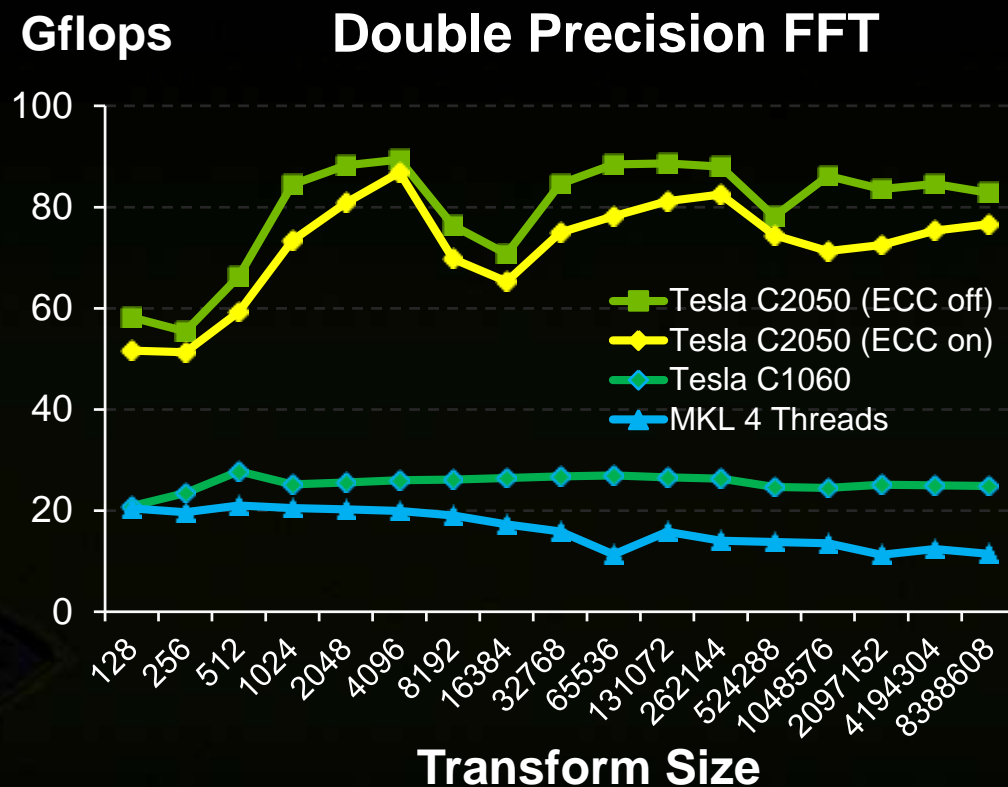
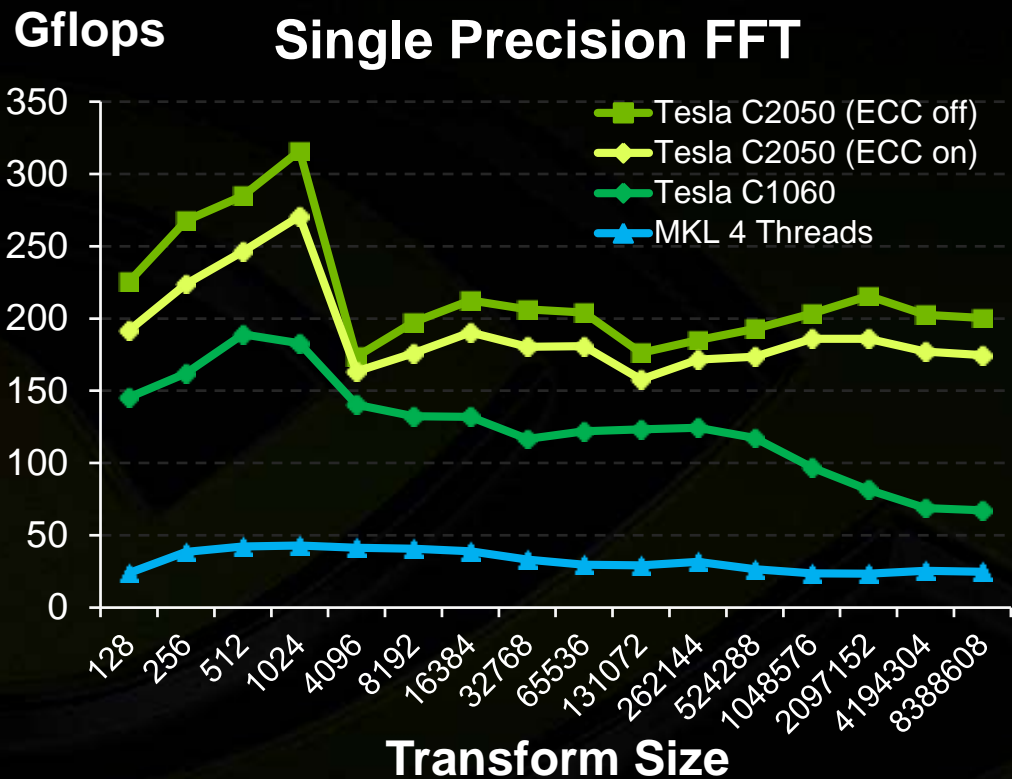


| | Tesla C1060 | Tesla C2050 | Tesla C2070 |
|---------------------------------|-------------------------------------------------|------------------------------------------------------------|-----------------------------|
| Architecture | Tesla 10-series GPU | Tesla 20-series GPU | |
| Number of Cores | 240 | 448 | |
| Caches | 16 KB Shared Memory / 8 cores | 64 KB L1 cache + Shared Memory / 32 cores, 768 KB L2 cache | |
| Floating Point Peak Performance | 933 Gigaflops (single) 78 Gigaflops (double) | 1030 Gigaflops (single) 515 Gigaflops (double) | |
| GPU Memory | 4 GB | 3 GB 2.625 GB with ECC on | 6 GB 5.25 GB with ECC on |
| Memory Bandwidth | 102 GB/s (GDDR3) | 144 GB/s (GDDR5) | |
| System I/O | PCIe x16 Gen2 | PCIe x16 Gen2 | |
| Power | 188 W (max) | 247 W (max) | 225 W (max) |
| Available | Available now | Shipping in May | Q3 2010 |

Performance Summary



Standard FFT Library: cuFFT 3.1-Pre-release

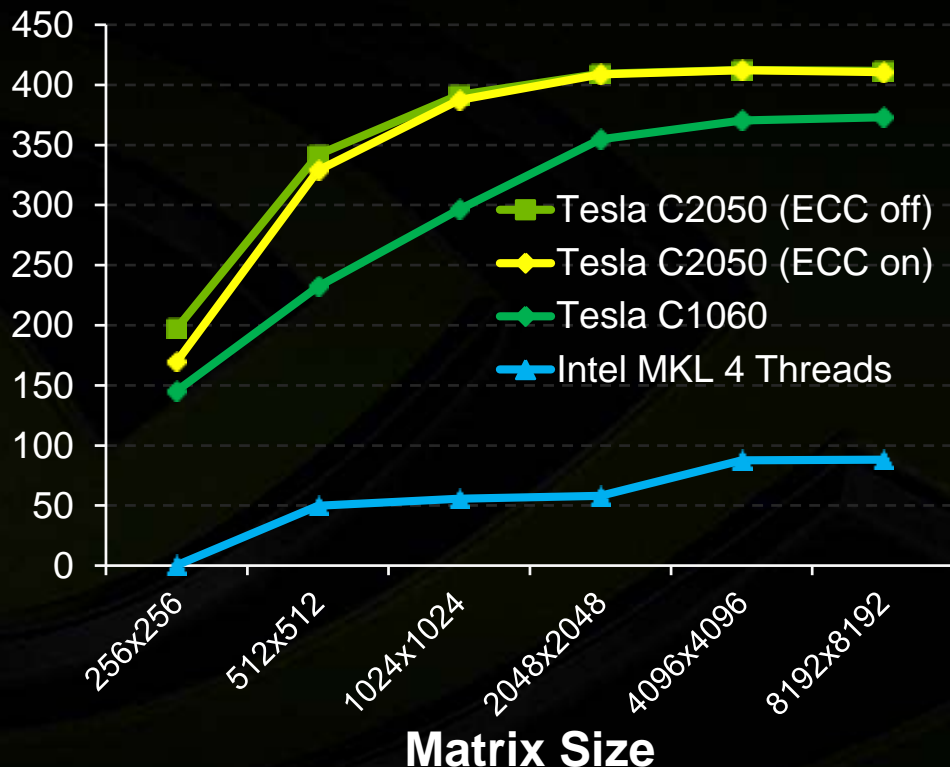


cuFFT 3.1-pre-release: NVIDIA Tesla C1060 GPU and Tesla C2050 (Fermi)
 MKL 10.1r1: Quad-Core Intel Core i7 (Nehalem) 3.2GHz

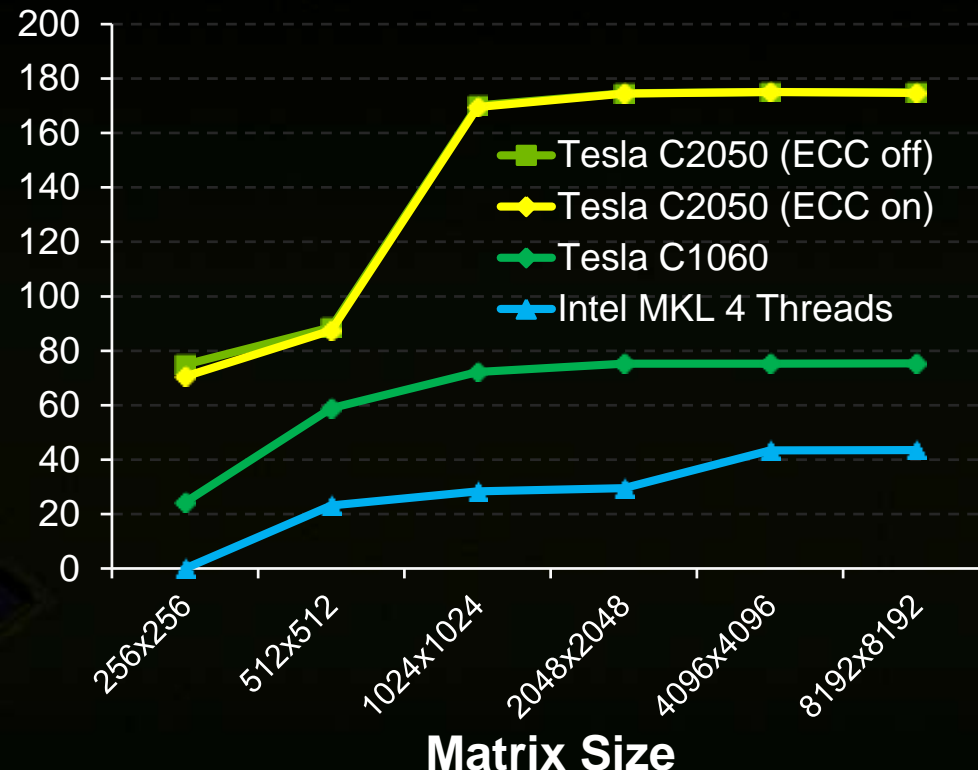
Standard BLAS Library: cuBLAS 3.1-Pre-release



Gflops Single Precision BLAS: SGEMM



Gflops Double Precision BLAS: DGEMM



cuBLAS: CUDA 3.1-pre-release: , Tesla C1060 and Tesla C2050

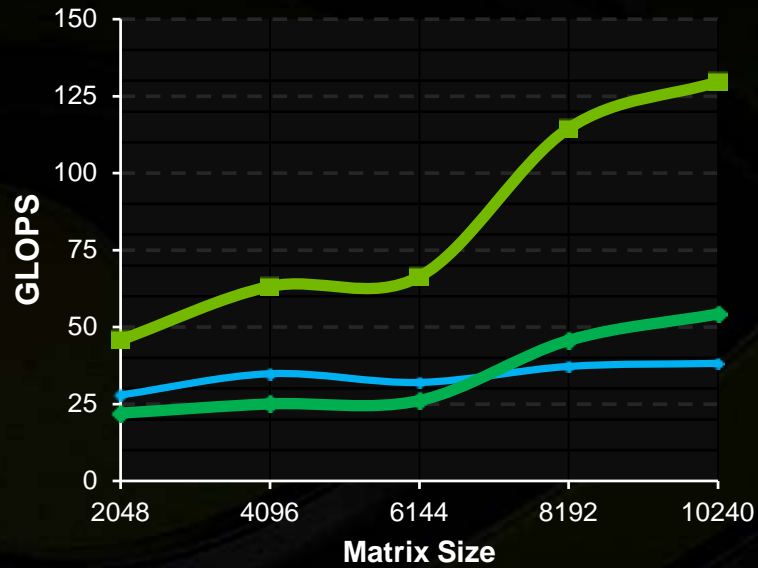
MKL 10.1r1: Intel Core2 Extreme, 3.00GHz

CULA 3.1 LAPACK Library from EM Photonics



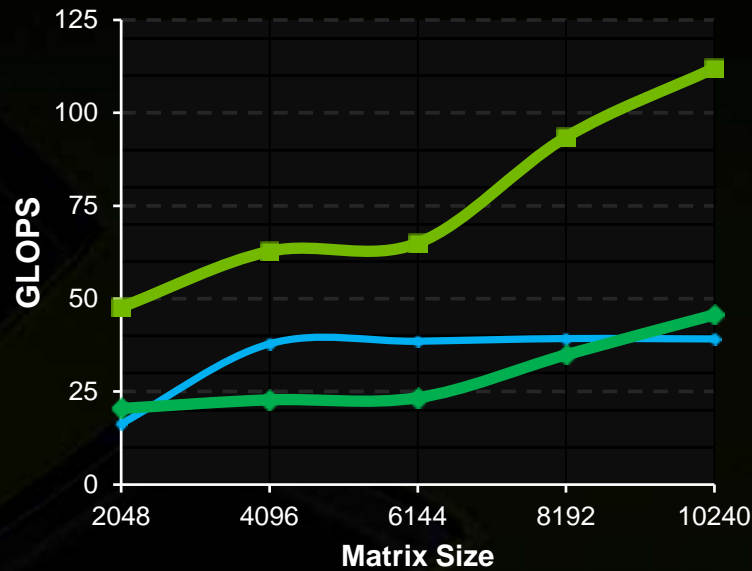
QR Decomposition (DGEQRF)

Householder method; Operation count estimated as $1.33N^3$



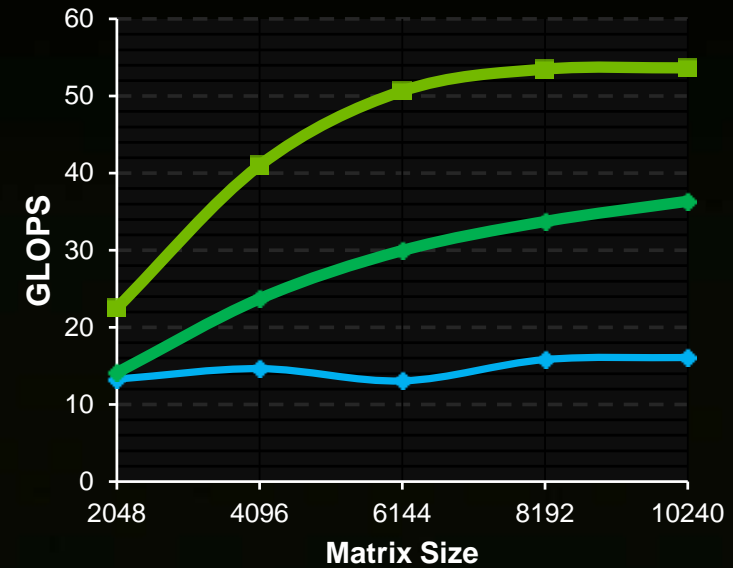
LU Decomposition (DGETRF)

Partial pivoting; Operation count estimated as $0.66N^3$



Singular Value Decomposition (DGESVD)

Left & right singular vectors; Operation count estimated as $21N^3$

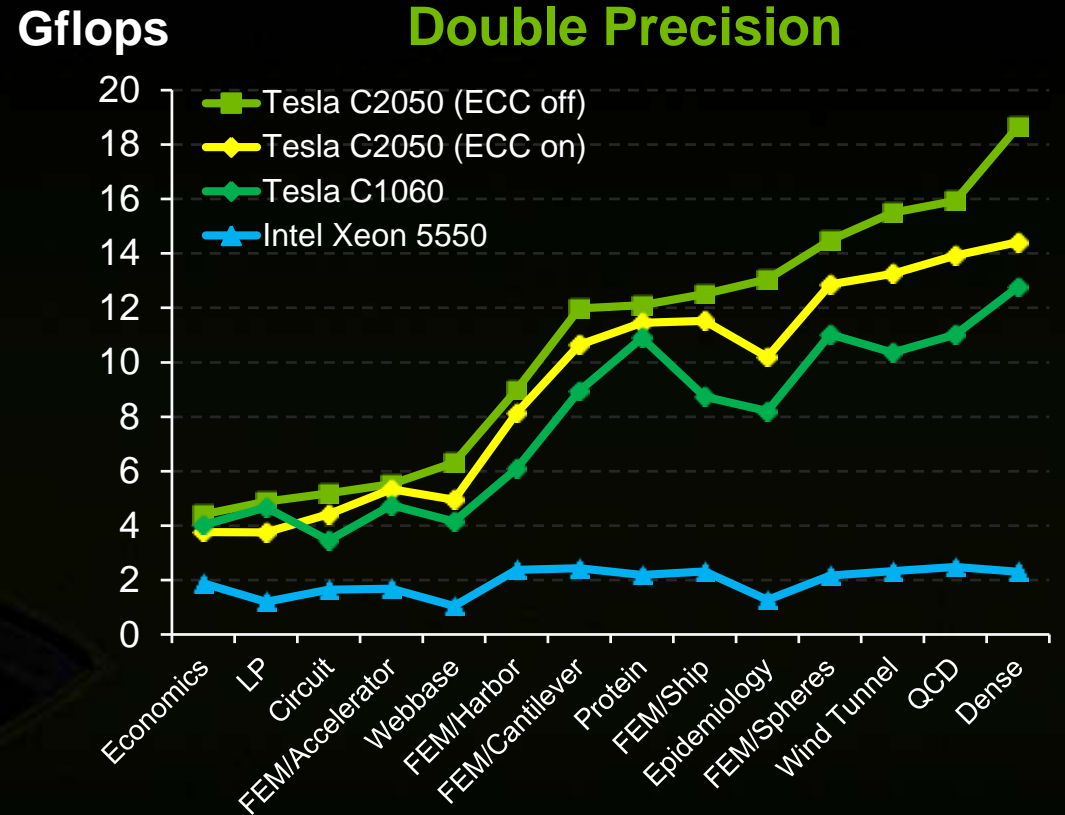
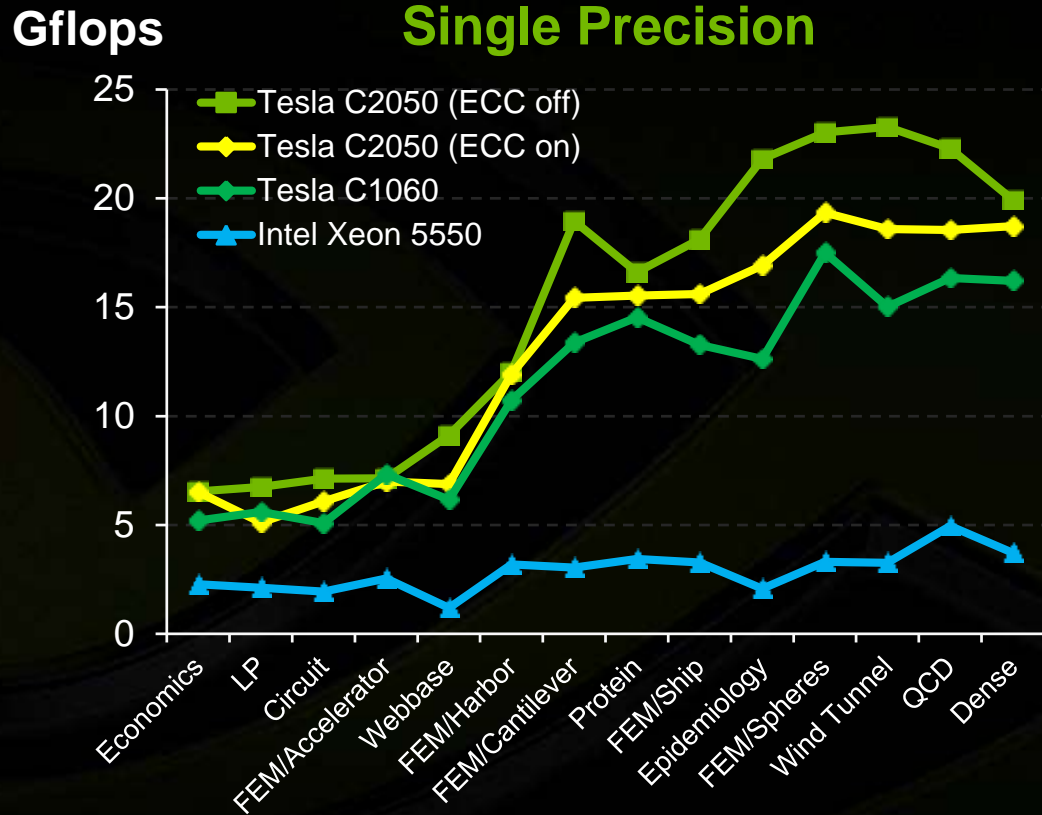


Double Precision Results

Data Courtesy: EM Photonics

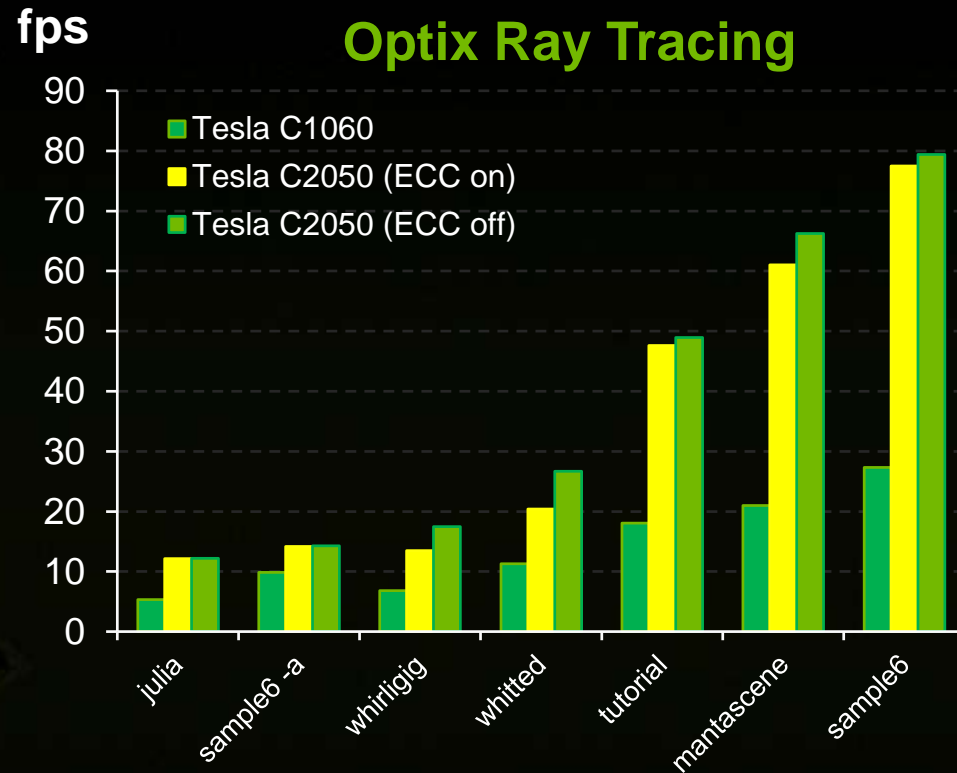
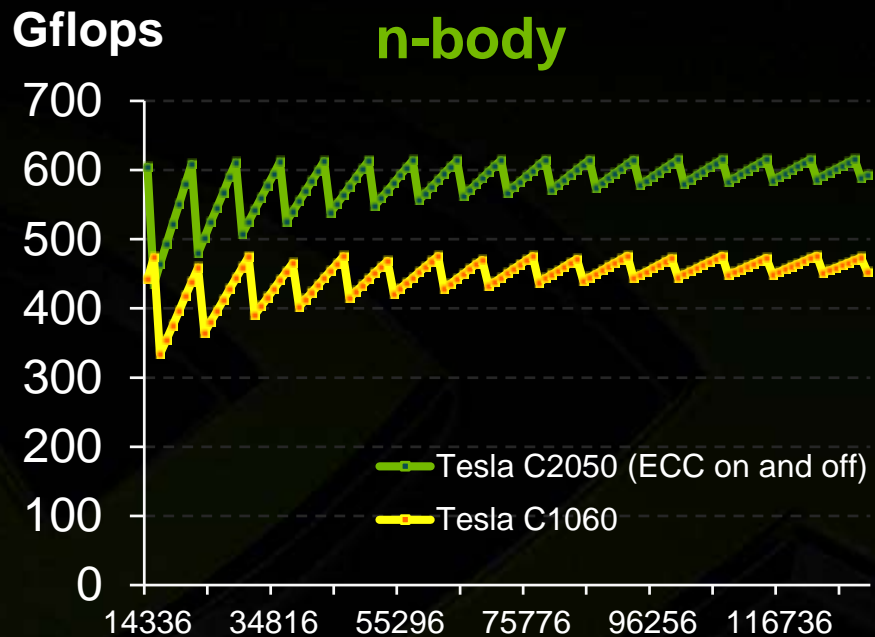


Sparse Matrix-Vector Multiplication (SpMV)



SpMv: CUDA 3.0, Tesla C1060 and Tesla C2050
MKL 10.2: Intel Xeon 5550, 2.67 GHz

N-body and Ray Tracing Performance

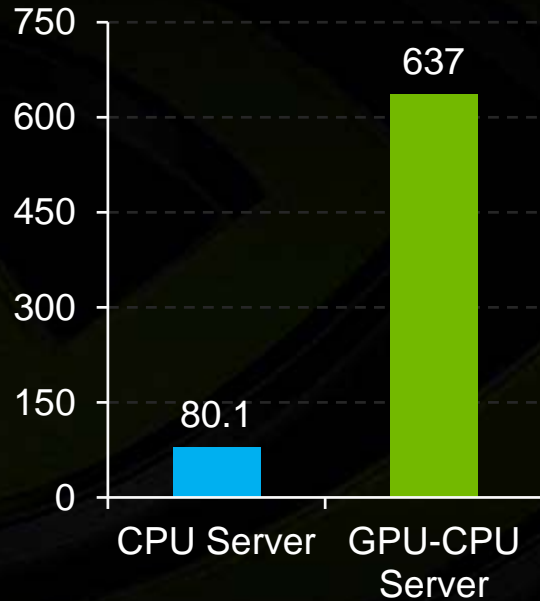


CUDA 3.0, Tesla C1060 and Tesla C2050

8x Higher Linpack

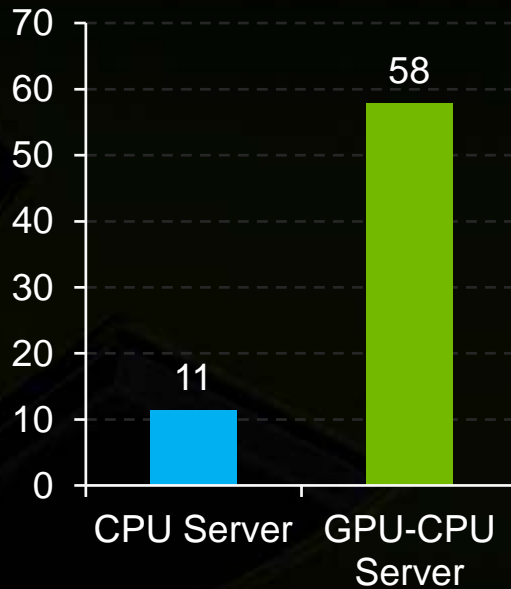
8x

Performance
Gflops



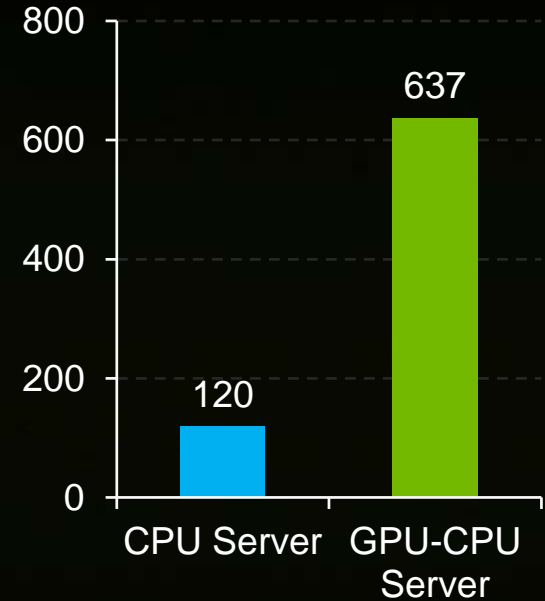
6x

Performance / \$
Gflops / \$K



5x

Performance / watt
Gflops / kwatt



CPU 1U Server: 2x Intel Xeon X5550 (Nehalem) 2.66 GHz, 48 GB memory, \$7K, 0.67 kw
GPU-CPU 1U Server: 2x Tesla C2050 + 2x Intel Xeon X5550, 48 GB memory, \$11K, 1.0 kw

“In testing our key applications, the Tesla GPUs delivered speed-ups that we had never seen before, sometimes even orders of magnitude.”



Satoshi Matsuoka

Professor
Tokyo Institute of Technology

“Future computing architectures will be hybrid systems with parallel-core GPUs working in tandem with multi-core CPUs”

Jack Dongarra

Professor, University of Tennessee
Author of Linpack



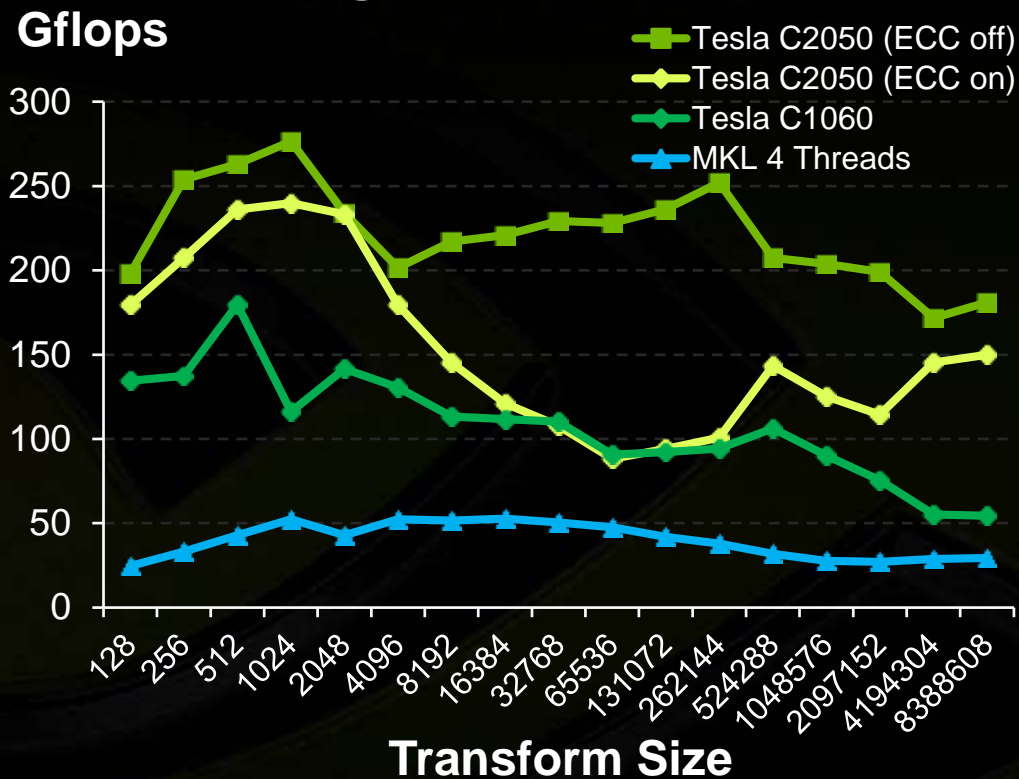
Backup Slides

CUDA 3.0
cuFFT and cuBLAS

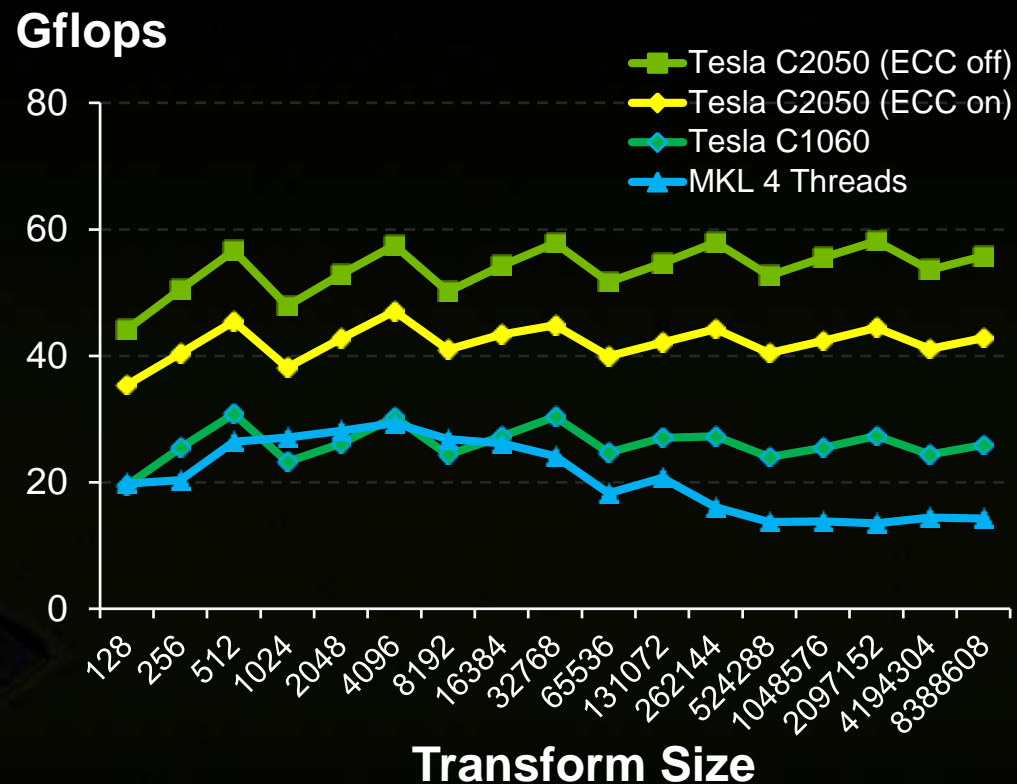
Standard FFT Library: cuFFT 3.0



Single Precision FFT



Double Precision FFT

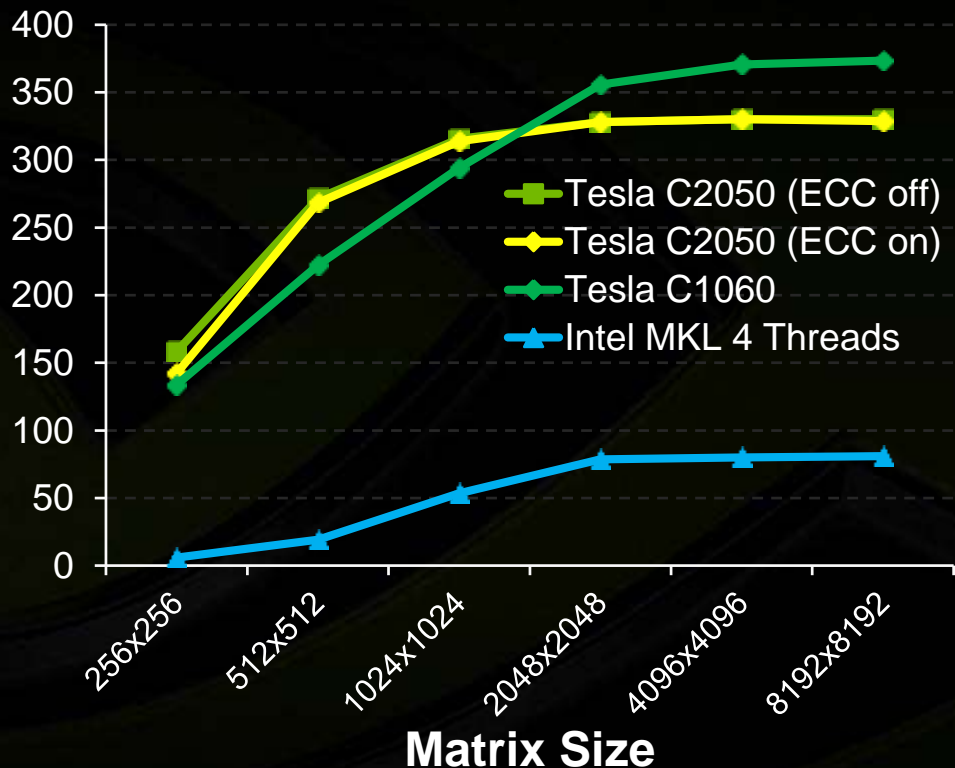


cuFFT 3.0: NVIDIA Tesla C1060 GPU and Tesla C2050
MKL 10.2.3: Quad-Core Intel Xeon 5550, 2.67 GHz

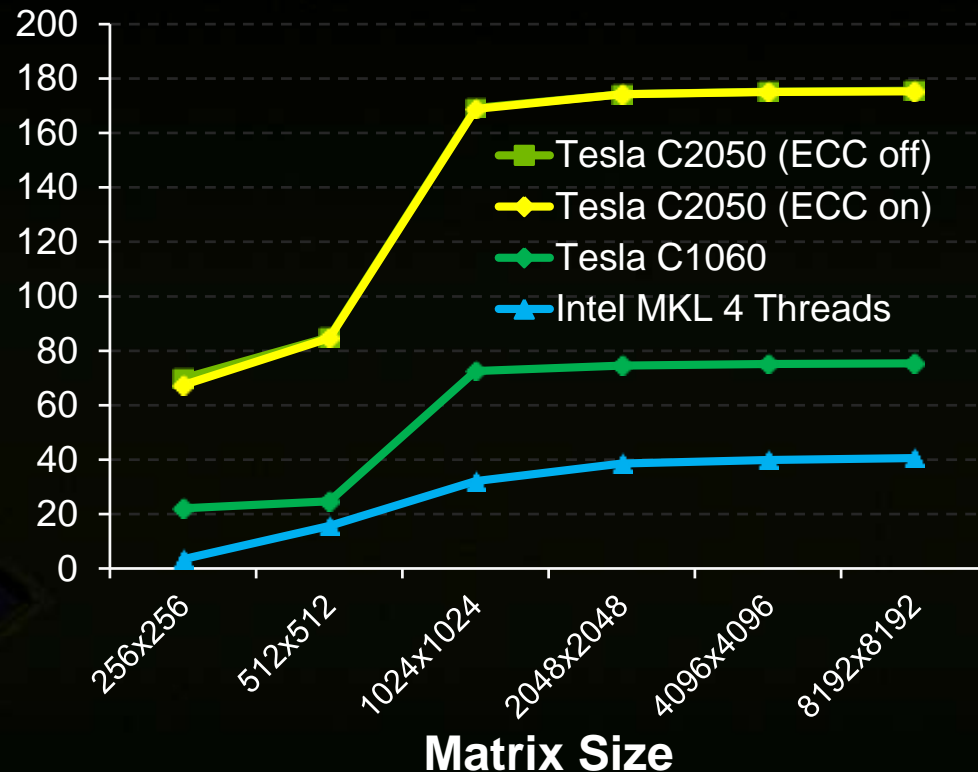
Standard BLAS Library: cuBLAS 3.0



Gflops Single Precision BLAS: SGEMM



Gflops Double Precision BLAS: DGEMM



cuBLAS: CUDA 3.0, Tesla C1060 and Tesla C2050

MKL 10.2.3: Quad-Core Intel Xeon 5550, 2.67 GHz