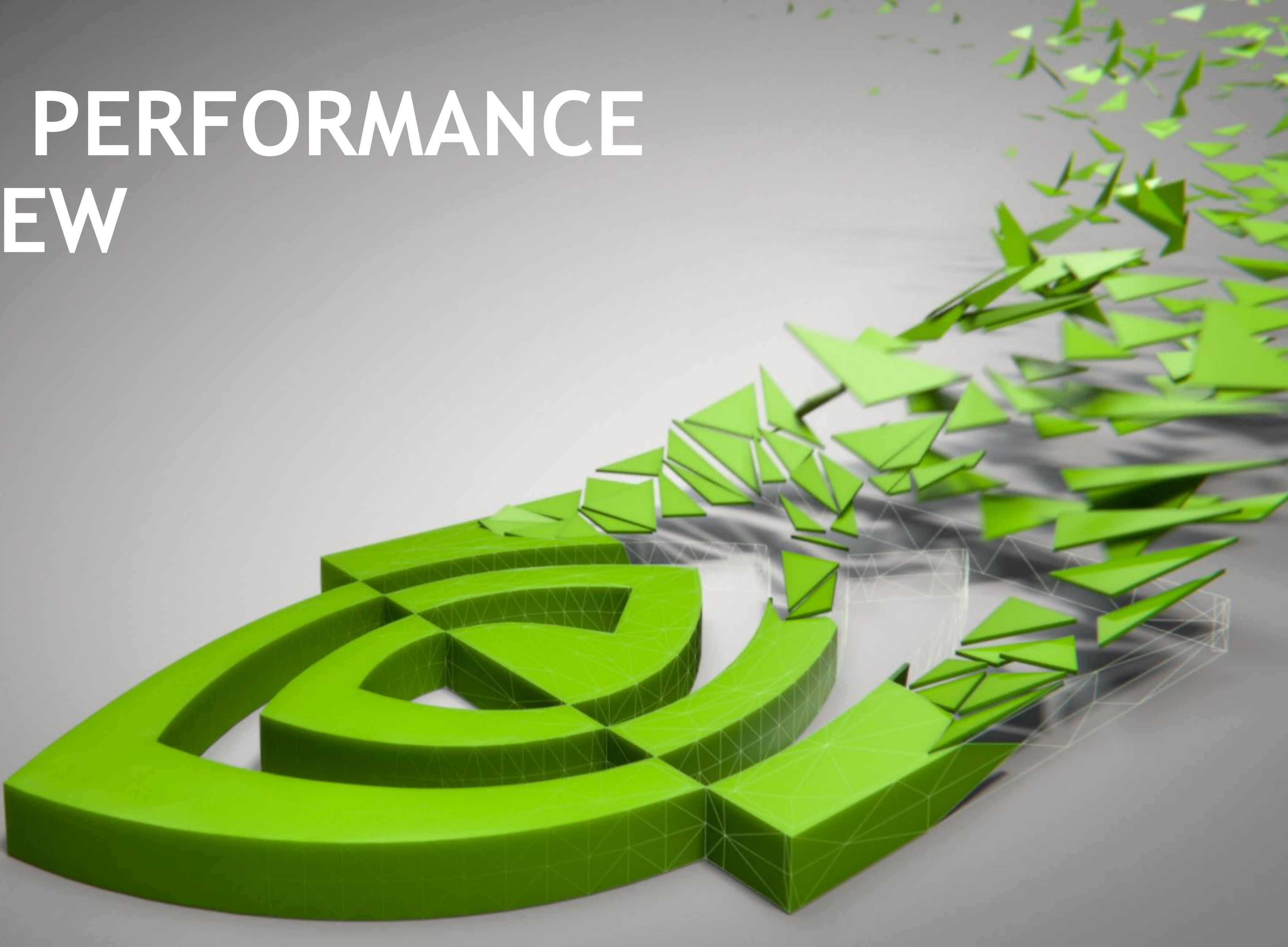
CUDA 8 PERFORMANCE OVERVIEW

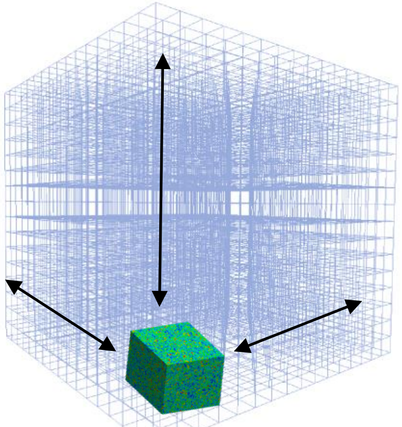
November 2016



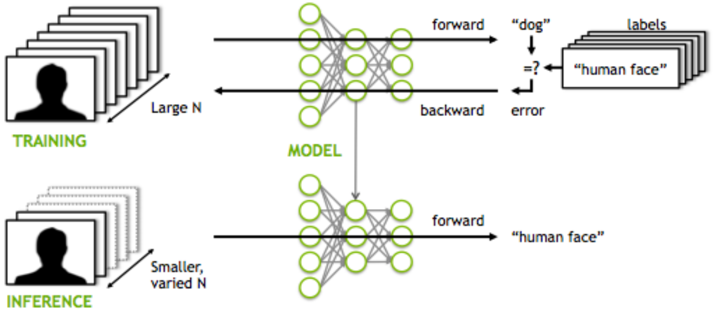
CUDA 8 PERFORMANCE HIGHLIGHTS

2X

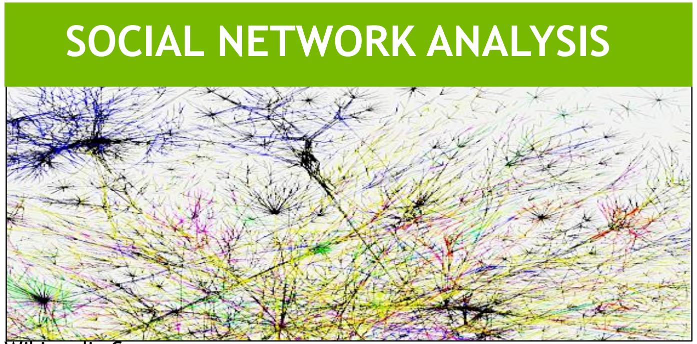
1.5-2X higher performance out-of-the-box



Solve larger problems than possible before with Unified Memory



Mixed precision support (FP16, INT8)



Wikimedia Commons

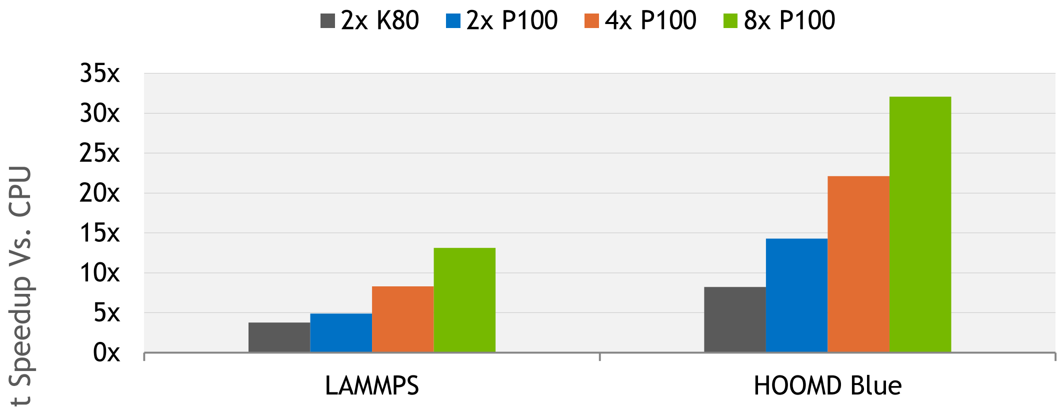
nvGRAPH: New graph analytics library

HIGH PERFORMANCE COMPUTING:

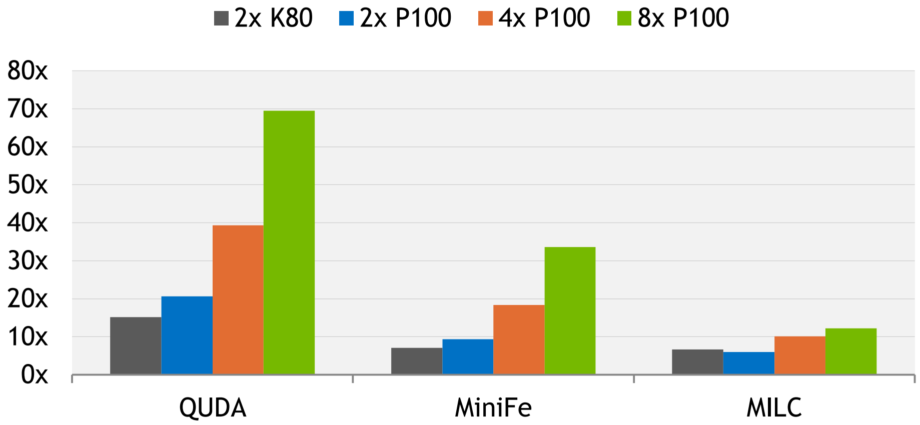
APPS WITH CUDA 8, OpenACC

> 2X SPEEDUPS ON HPC APPS WITH P100

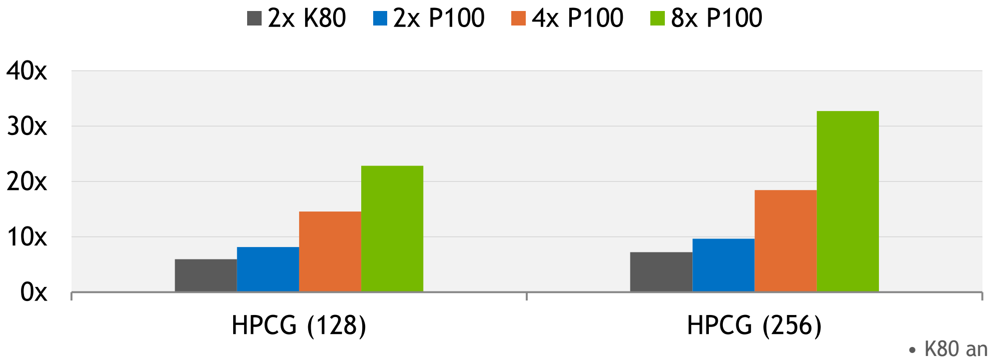
Molecular Dynamics



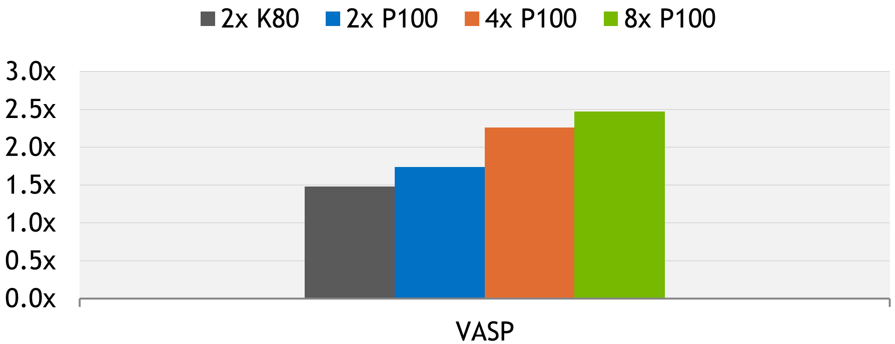
Physics



HPC Benchmarks



Quantum Chemistry

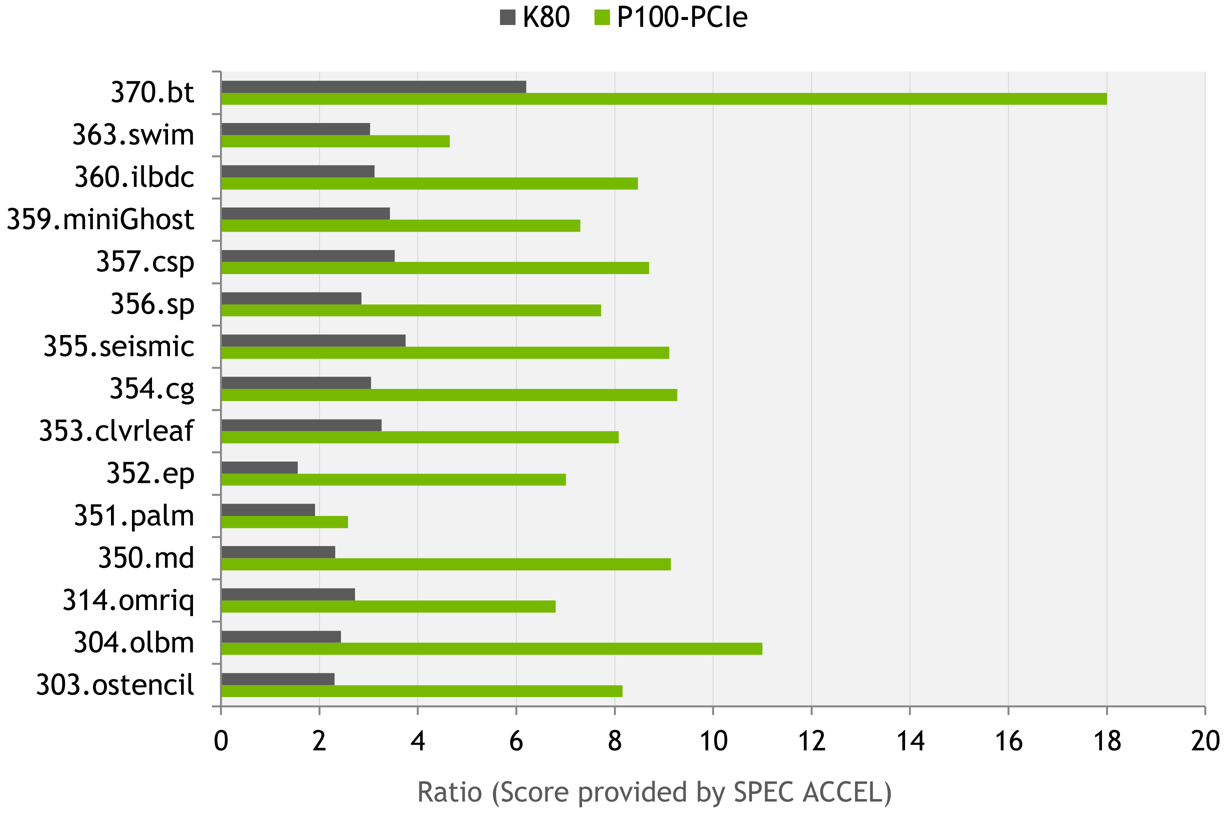


- K80 and P100 (SXM2); Base clocks; cube mesh topolgy (DGX-1)
- CUDA 8 GA (8.0.44) with r361.79 (K80) and r361.96 (P100)
- CPU System: Intel Xeon Broadwell dual socket 22-core E5-2699 v4@2.2GHz 3.6GHz Turbo with Ubuntu 14.04.3 x86-64 and 256GB memory
- Full system configurations including benchmark versions and data sets used available in the Appendix

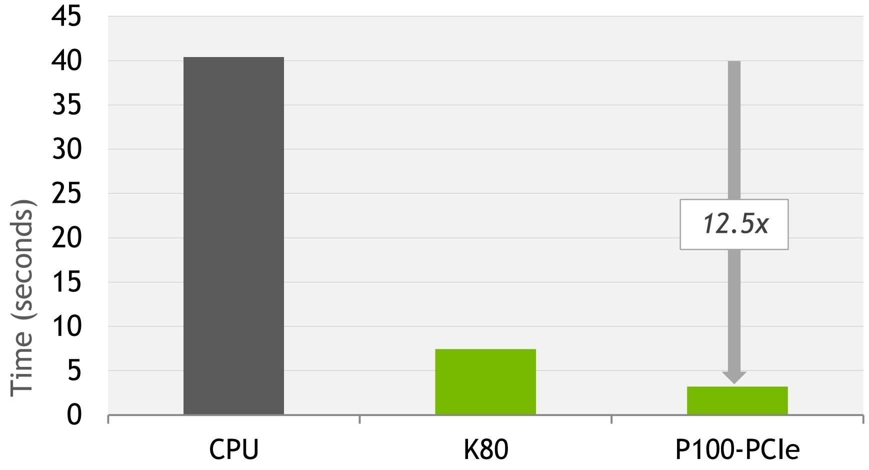
Performance may vary based on OS and software versions, and motherboard configuration

OpenACC: >2X—12X FASTER WITH P100

2.7x Speedup on SPEC ACCEL



12.5x Faster than CPU on CloverLeaf



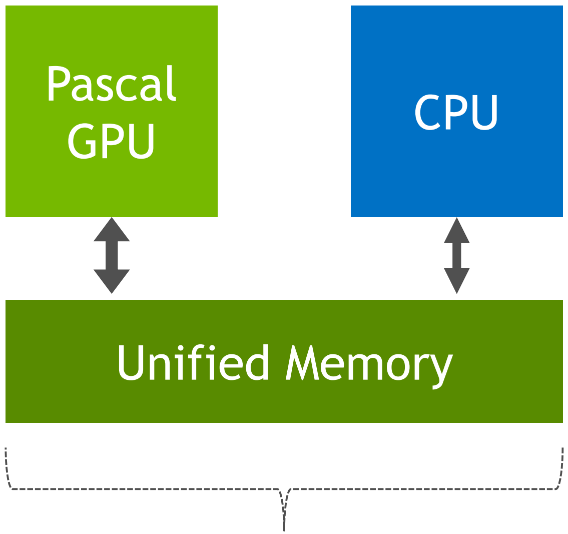
- SPEC ACCEL Base runs on 1x K80 (1-GPU), 1x P100 (PCIe) ; Base clocks and ECC ON
- CloverLeaf runs on 1x K80 (2-GPUs), 2x P100 (PCIe) ; Base clocks and ECC ON
- PGI 16.1 with CUDA 7.5 on K80 (r352) and PGI 16.7 with CUDA 8 on P100 (r361)
- CPU measurements for CloverLeaf on Intel Xeon Broadwell single-socket 20-core E5-2699 v4@2.2GHz 3.6GHz Turbo with Intel MKL 2017
- MPI versions 1.8.8 (K80), 1.10.2 (P100) and 2017.0.098 (CPU)
- Host System: Intel Xeon Haswell dual socket 16-core/socket E5-2698 v3@2.3GHz 3.6GHz Turbo with Ubuntu 14.04.3 x86-64 and 256GB memory

Performance may vary based on OS and software versions, and motherboard configuration

UNIFIED MEMORY ON PASCAL

Large datasets, Simple programming, High performance

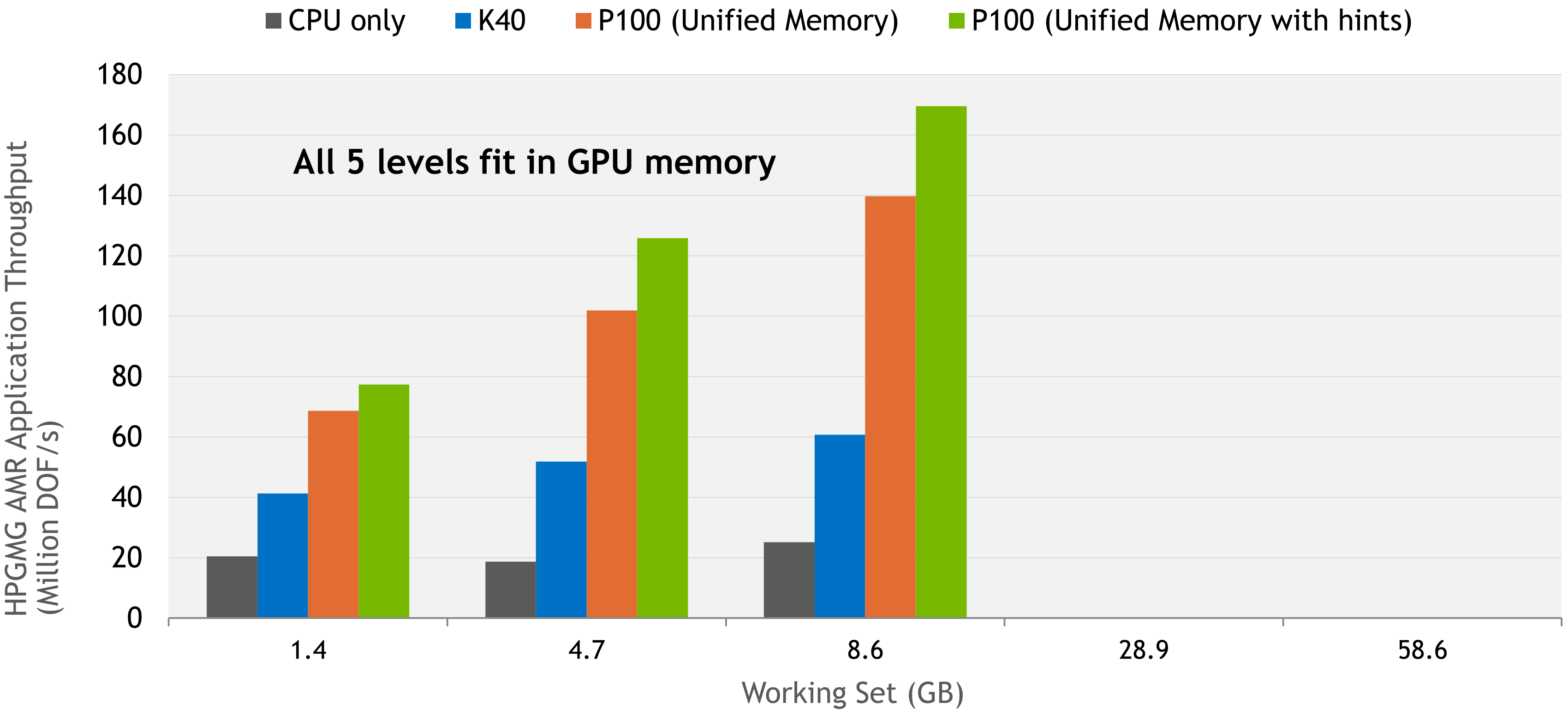
CUDA 8



Allocate Beyond
GPU Memory Size

Enable Large Data Models	Oversubscribe GPU memory Allocate up to system memory size
Higher Application Performance	Demand paging User APIs for prefetching & migration hints
Simpler Data Access	CPU/GPU Data coherence Unified memory atomic operations

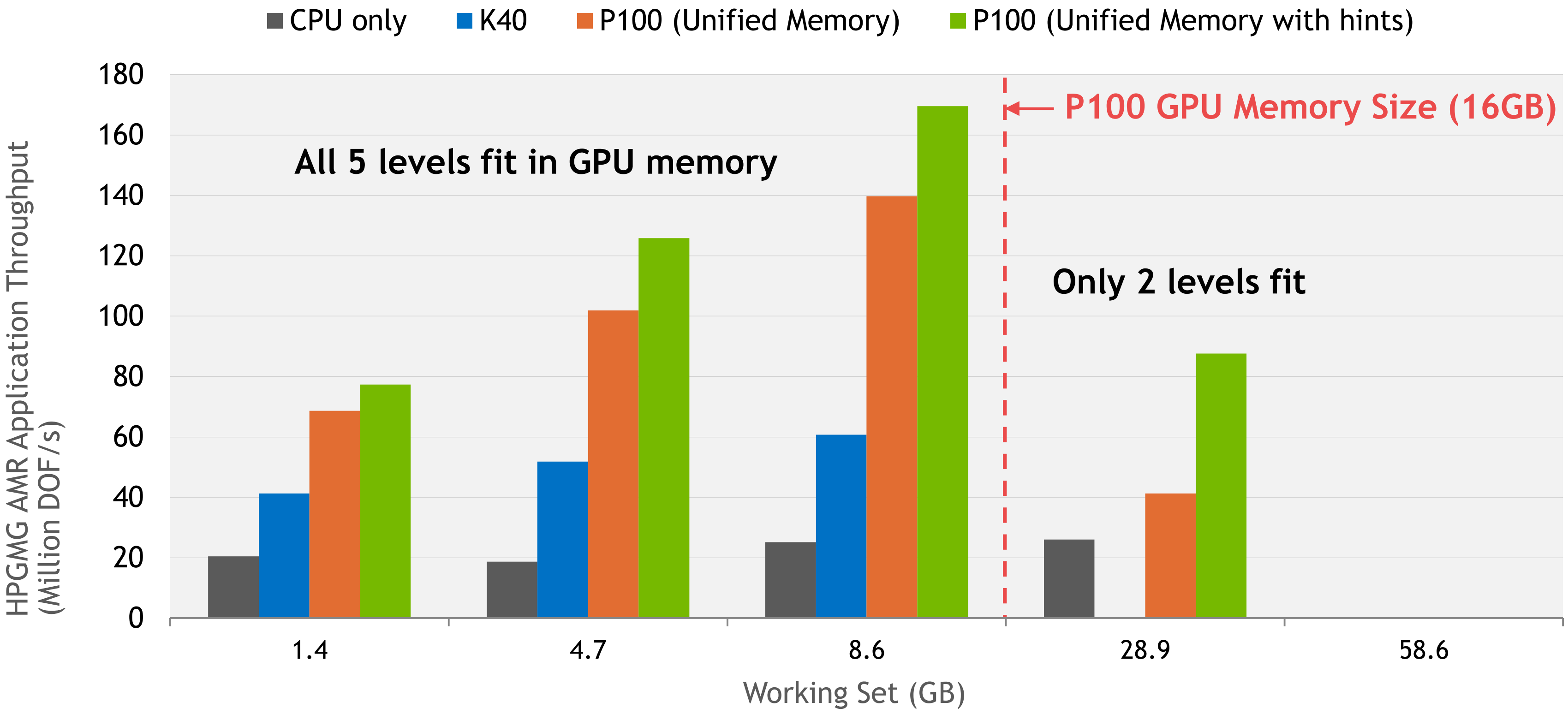
SOLVE LARGER PROBLEMS WITH HIGHER THROUGHPUT



Performance may vary based on OS and software versions, and motherboard configuration

- HPGMG AMR on 1xK40, 1xP100 (PCIe) with CUDA 8 (r361)
- CPU measurements with Intel Xeon Haswell dual socket 10-core E5-2650 v3@2.3 GHz 3.0 GHz Turbo, HT on
- Host System: Intel Xeon Haswell dual socket 16-cores E5-2630 v3@2.4GHz 3.2GHz Turbo

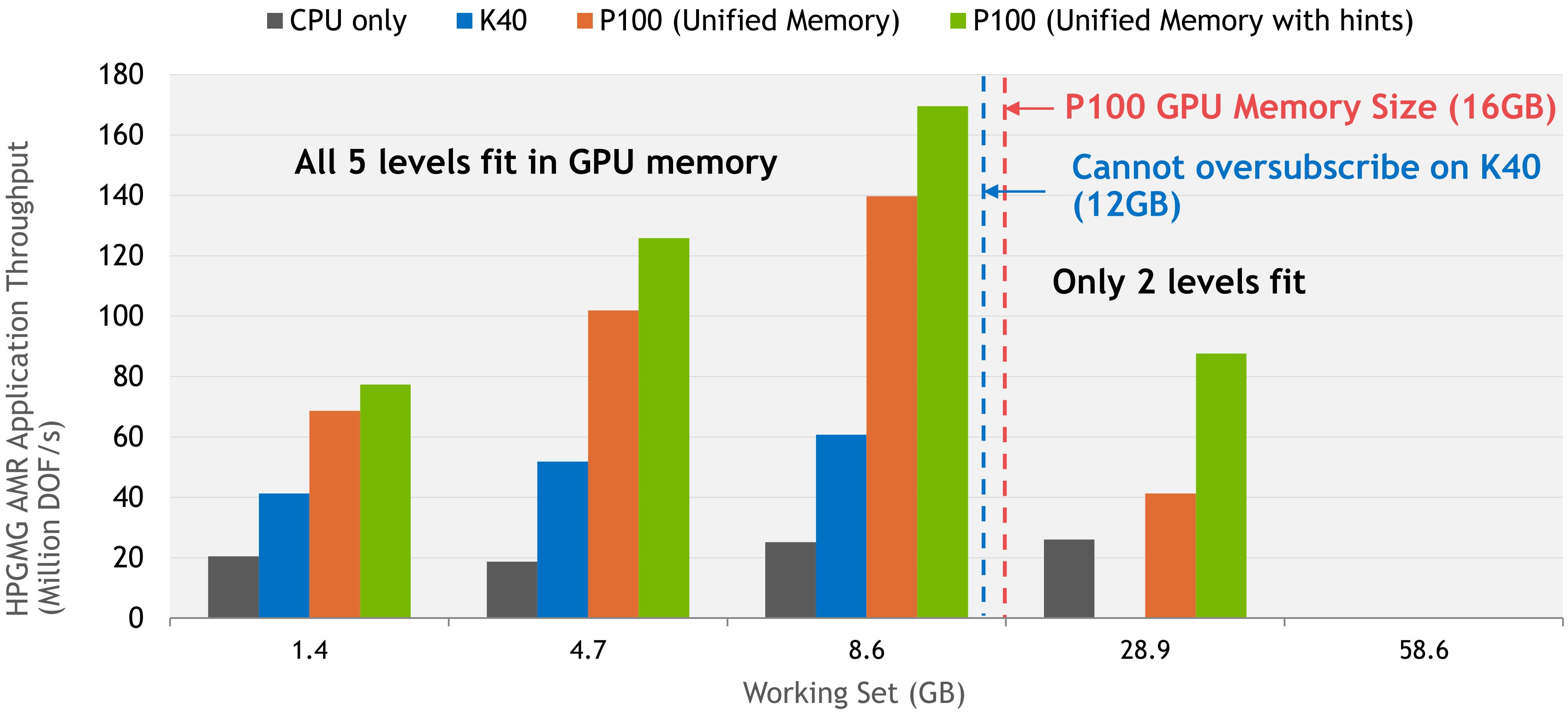
SOLVE LARGER PROBLEMS WITH HIGHER THROUGHPUT



Performance may vary based on OS and software versions, and motherboard configuration

- HPGMG AMR on 1xK40, 1xP100 (PCIe) with CUDA 8 (r361)
- CPU measurements with Intel Xeon Haswell dual socket 10-core E5-2650 v3@2.3 GHz 3.0 GHz Turbo, HT on
- Host System: Intel Xeon Haswell dual socket 16-cores E5-2630 v3@2.4GHz 3.2GHz Turbo

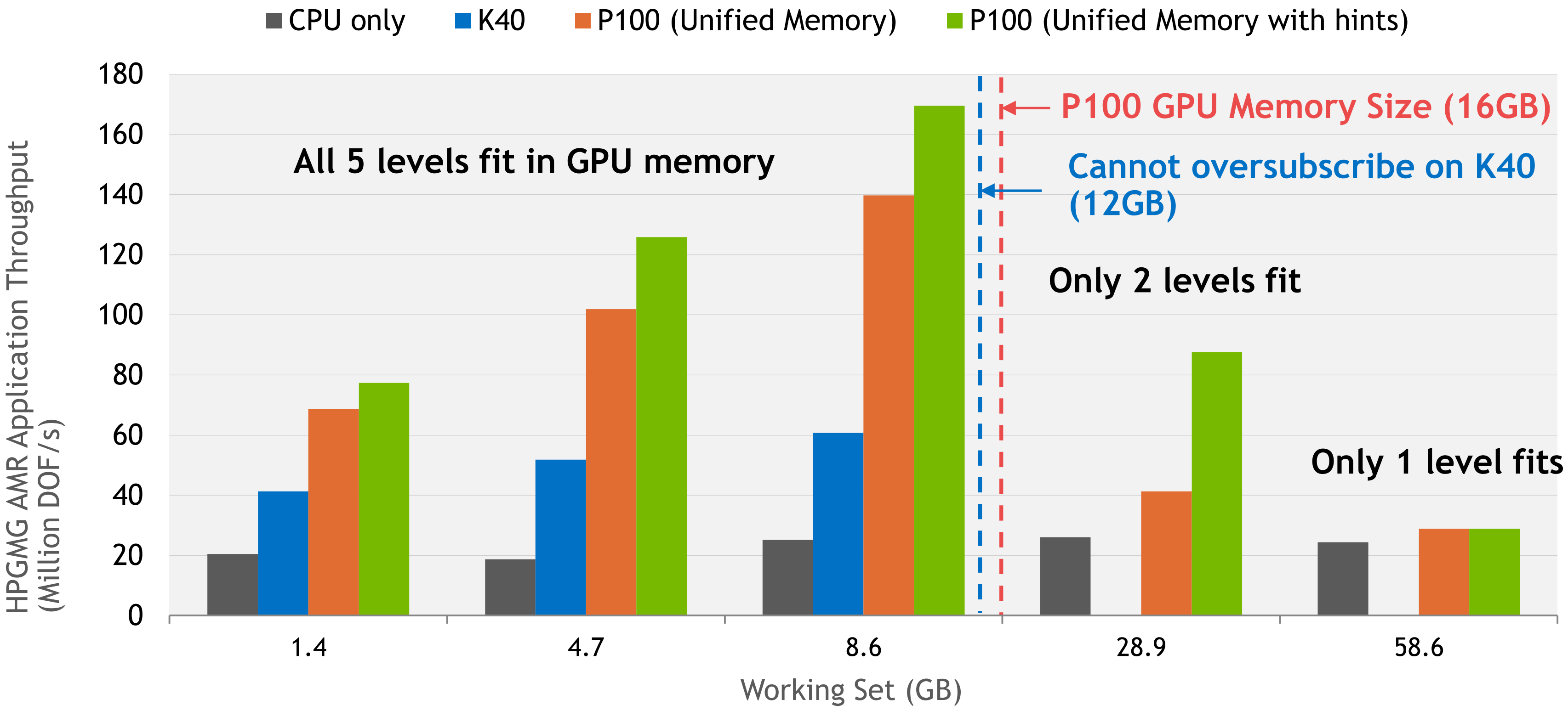
SOLVE LARGER PROBLEMS WITH HIGHER THROUGHPUT



Performance may vary based on OS and software versions, and motherboard configuration

- HPGMG AMR on 1xK40, 1xP100 (PCIe) with CUDA 8 (r361)
- CPU measurements with Intel Xeon Haswell dual socket 10-core E5-2650 v3@2.3 GHz 3.0 GHz Turbo, HT on
- Host System: Intel Xeon Haswell dual socket 16-cores E5-2630 v3@2.4GHz 3.2GHz Turbo

SOLVE LARGER PROBLEMS WITH HIGHER THROUGHPUT

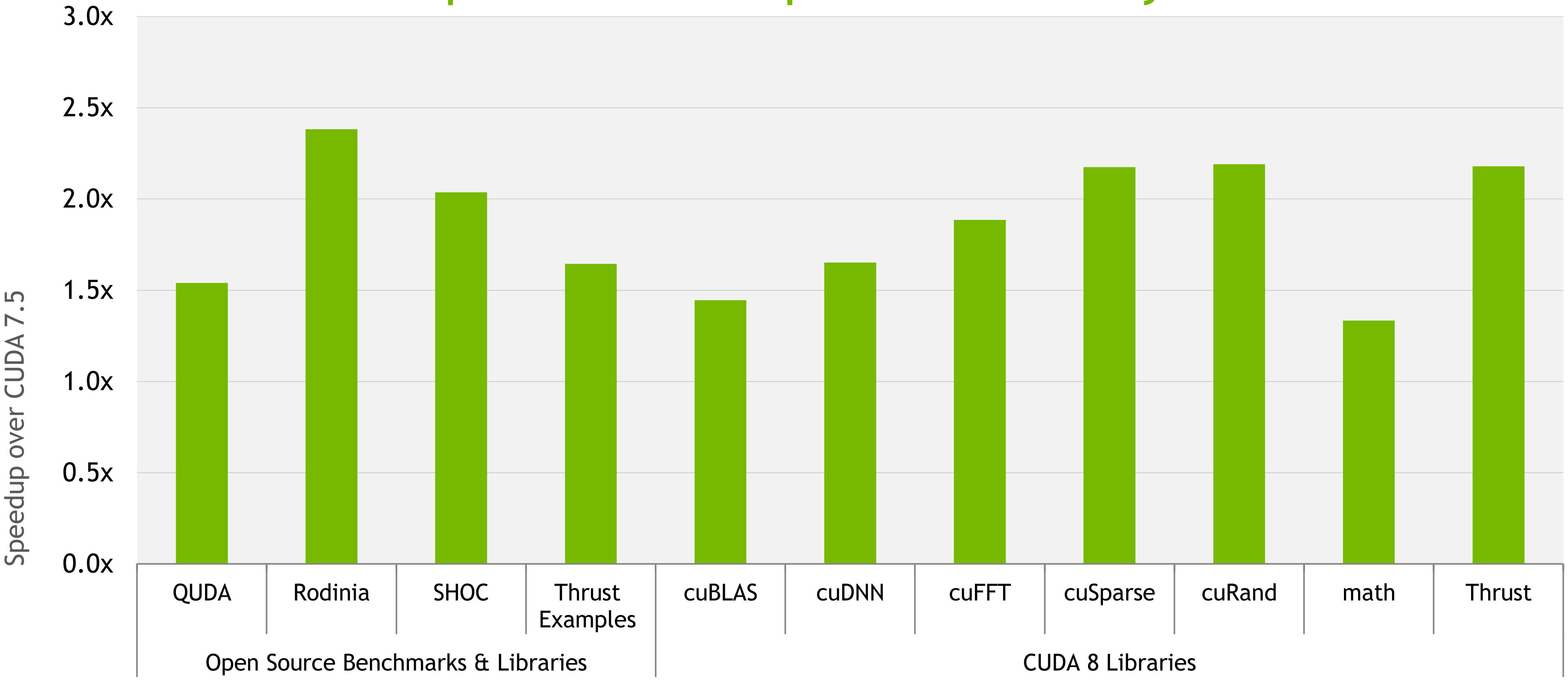


Performance may vary based on OS and software versions, and motherboard configuration

- HPGMG AMR on 1xK40, 1xP100 (PCIe) with CUDA 8 (r361)
- CPU measurements with Intel Xeon Haswell dual socket 10-core E5-2650 v3@2.3 GHz 3.0 GHz Turbo, HT on
- Host System: Intel Xeon Haswell dual socket 16-cores E5-2630 v3@2.4GHz 3.2GHz Turbo

CUDA 8 NVCC: > 2X FASTER COMPILE TIMES

Improved Developer Productivity



- Average total compile times (per translation unit)
- Host system: Intel Core i7-3930K 6-cores @ 3.2GHz
- CentOS x86_64 Linux release 7.1.1503 (Core) with GCC 4.8.3 20140911
- GPU target architecture sm_52

Performance may vary based on OS and software versions, and motherboard configuration

GPU ACCELERATED LIBRARIES:

GRAPH ANALYTICS

nvGRAPH

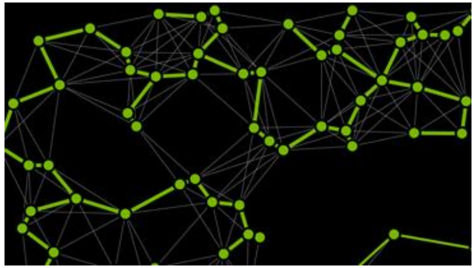
GPU Accelerated Graph Analytics

Parallel Library for Interactive and High Throughput Graph Analytics

Solve graphs with up to 2.5 Billion edges on a single GPU (Tesla M40)

Includes — PageRank, Single Source Shortest Path and Single Source Widest Path algorithms

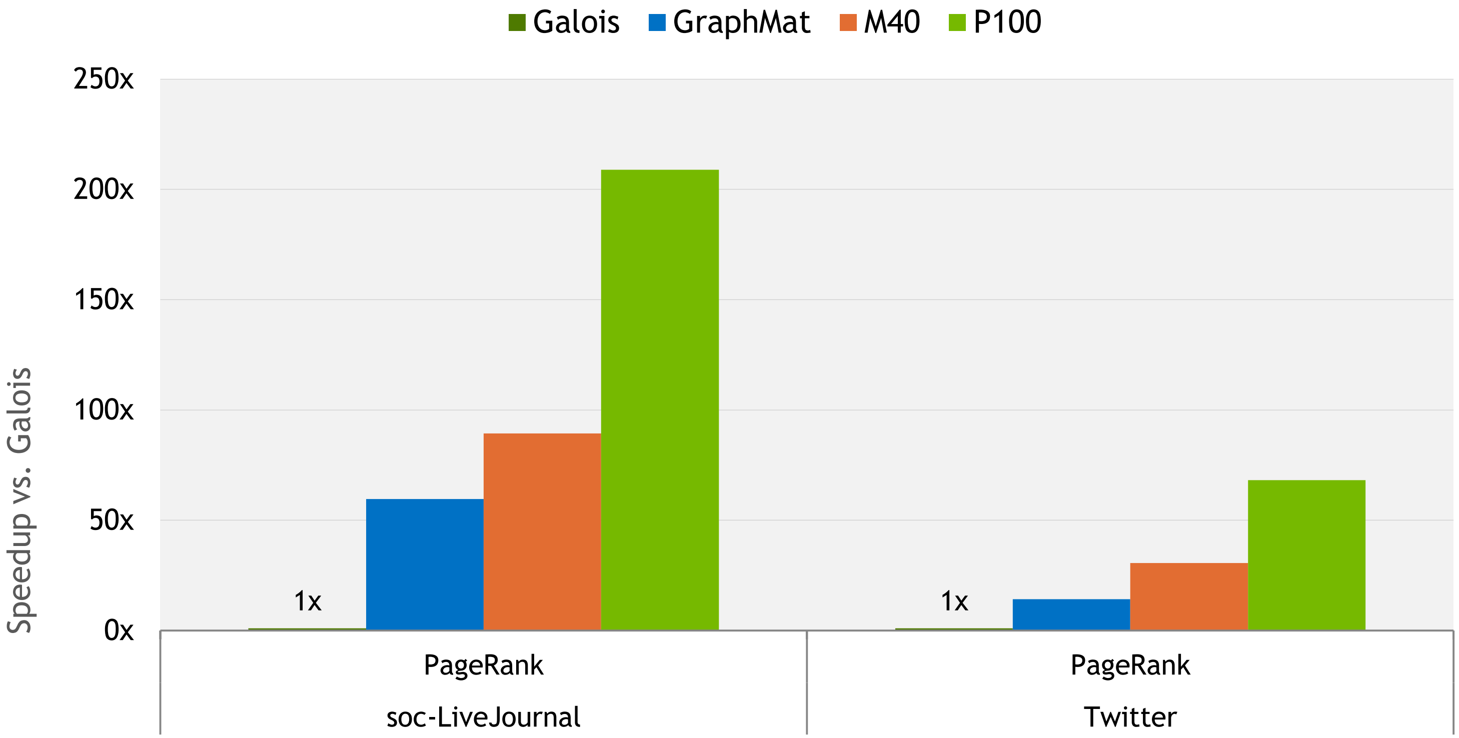
Semi-ring SPMV operations provides building blocks for graph traversal algorithms



PageRank	Single Source Shortest Path	Single Source Widest Path
Search	Robotic Path Planning	IP Routing
Recommendation Engines	Power Network Planning	Chip Design / EDA
Social Ad Placement	Logistics & Supply Chain Planning	Traffic sensitive routing

https://developer.nvidia.com/nvgraph

> 200X SPEEDUP ON PAGERANK VS GALOIS

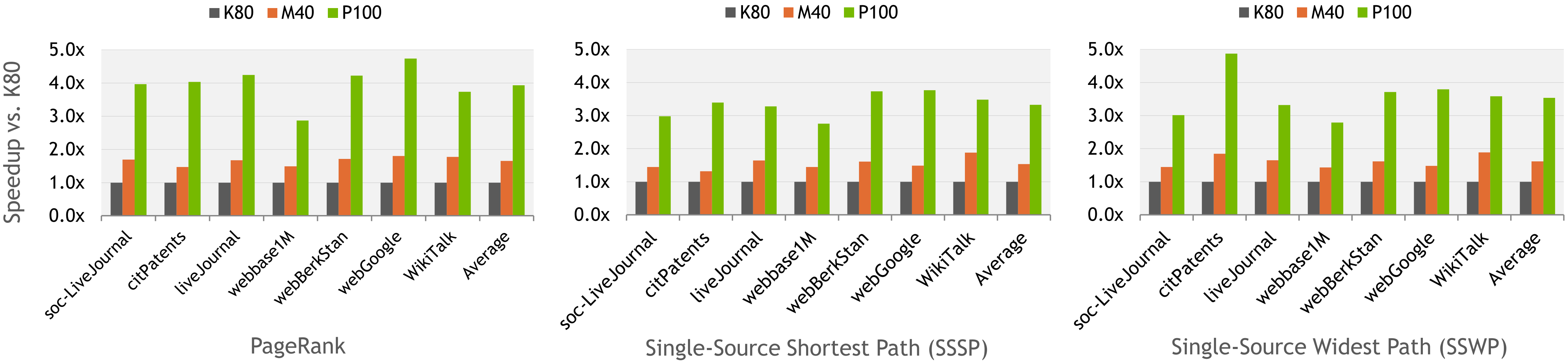


- nvGRAPH on M40 (ECC ON, r352), P100 (r361), Base clocks, input and output data on device
- GraphMat, Galois (v2.3) on Intel Xeon Broadwell dual-socket 22-core/socket E5-2699 v4 @ 2.22GHz, 3.6GHz Turbo
- Comparing Average Time per Iteration (ms) for PageRank
- Host System: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

> 4X SPEEDUPS WITH P100

Using Different Algorithms in nvGRAPH



- nvGRAPH on K80, M40, P100, ECC ON, Base clocks, input and output data on device
- GraphMat, Galois (v2.3) on Intel Xeon Broadwell dual-socket 44-core E5-2699 v4 @ 2.22GHz, 3.6GHz Turbo
- Comparing Average Time per Iteration (ms) for PageRank and Total Solver Time (ms) for SSSP and SSWP
- Host System: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

GPU ACCELERATED LIBRARIES:

FAST FOURIER TRANSFORMS

cuFFT

Complete Fast Fourier Transforms Library

Complete Multi-Dimensional FFT Library

Simple “drop-in” replacement of a CPU FFTW library

Real and complex, single- and double-precision data types

Includes 1D, 2D and 3D batched transforms

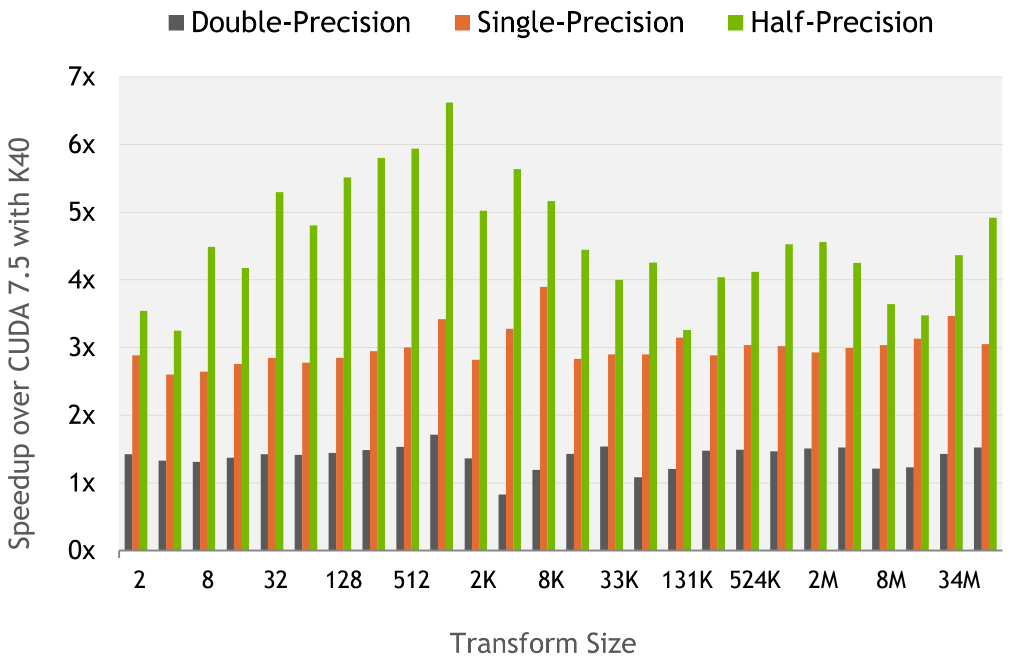
Support for half-precision (FP16) data types

Supports flexible input and output data layouts

XT interface now supports up to 8 GPUs

https://developer.nvidia.com/cufft

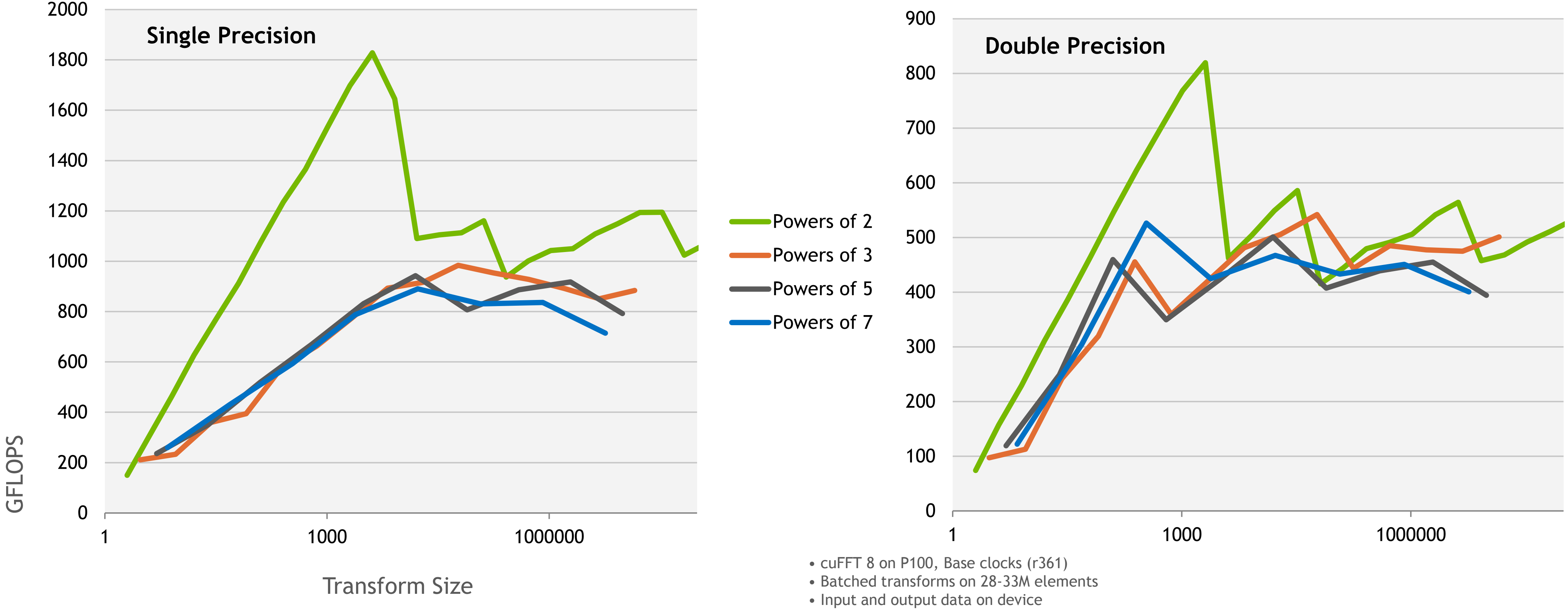
> 6x Speedup with Half-Precision on P100



- Speedup of P100 with CUDA 8 vs. K40m with CUDA 7.5
- cuFFT 7.5 on K40m, Base clocks, ECC on (r352)
- cuFFT 8.0 on P100, Base clocks, ECC on (r361)
- 1D Complex, Batched transforms on 28-33M elements
- Input and output data on device
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

cuFFT: > 1800 GFLOPS SINGLE PRECISION

1D Complex, Batched FFTs



Single Precision

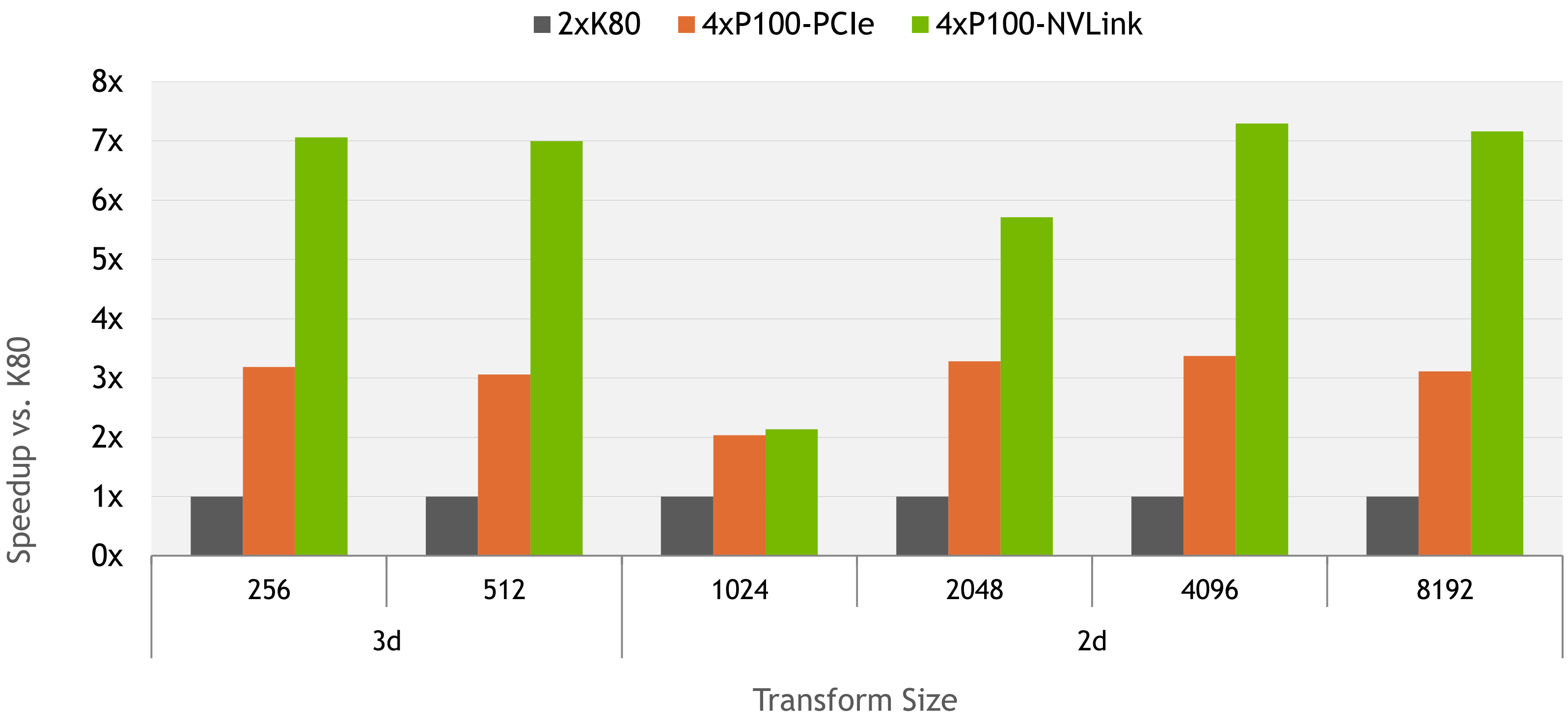
Double Precision

- cuFFT 8 on P100, Base clocks (r361)
- Batched transforms on 28-33M elements
- Input and output data on device
- Excludes time to create cuFFT "plans"
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3@ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

cuFFT-XT: > 7X IMPROVEMENTS WITH NVLINK

2D and 3D Complex FFTs



- cuFFT 7.5 on 2xK80m, ECC ON, Base clocks (r352)
- cuFFT 8 on 4xP100 with PCIe and NVLink (DGX-1), Base clocks (r361)
- Input and output data on device
- Excludes time to create cuFFT "plans"
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3@ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

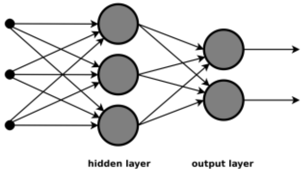
DENSE AND SPARSE LINEAR ALGEBRA

Applications

cuBLAS

BASIC LINEAR ALGEBRA SUBPROGRAMS

DEEP LEARNING

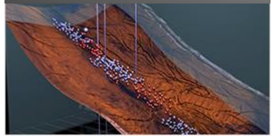


cuSPARSE

COMPUTATIONAL FLUID DYNAMICS

CAD/CAM/CAE

SEISMIC EXPLORATION

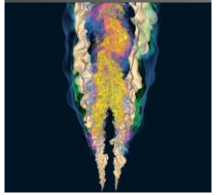


RECOMMENDATION ENGINES

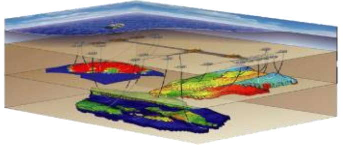
NLP

cuSOLVER

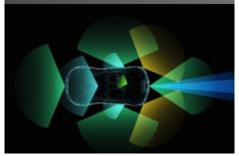
COMBUSTION SIMULATION



OIL & GAS WELL MODELS



COMPUTER VISION



cuBLAS

Dense Linear Algebra on GPUs

Complete BLAS Library Plus Extensions

Supports all 152 standard routines for single, double, complex, and double complex

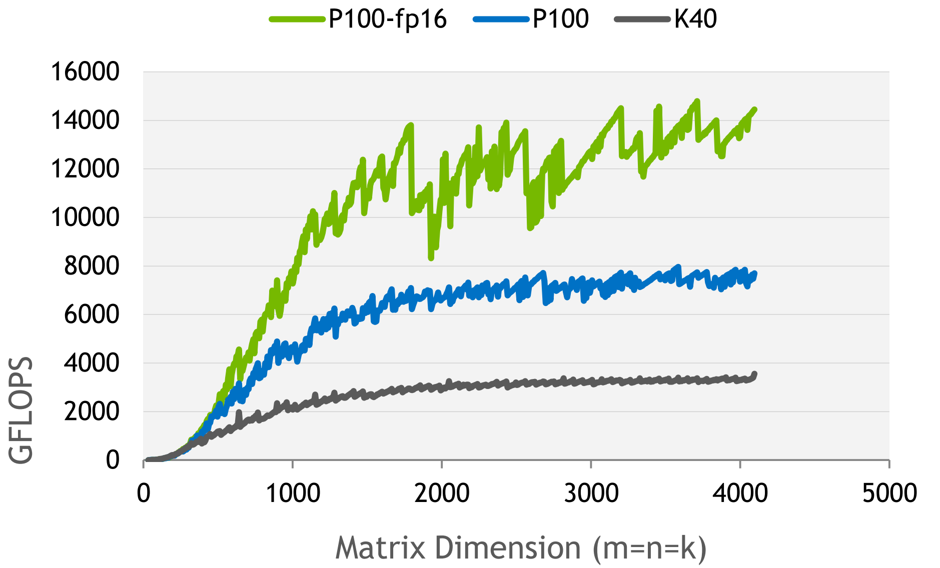
Supports half-precision (FP16) and integer (INT8) matrix multiplication operations

Batched routines for higher performance on small problem sizes

Host and device-callable interface

XT interface supports distributed computations across multiple GPUs

> 4x Faster GEMM Performance with FP16 on P100

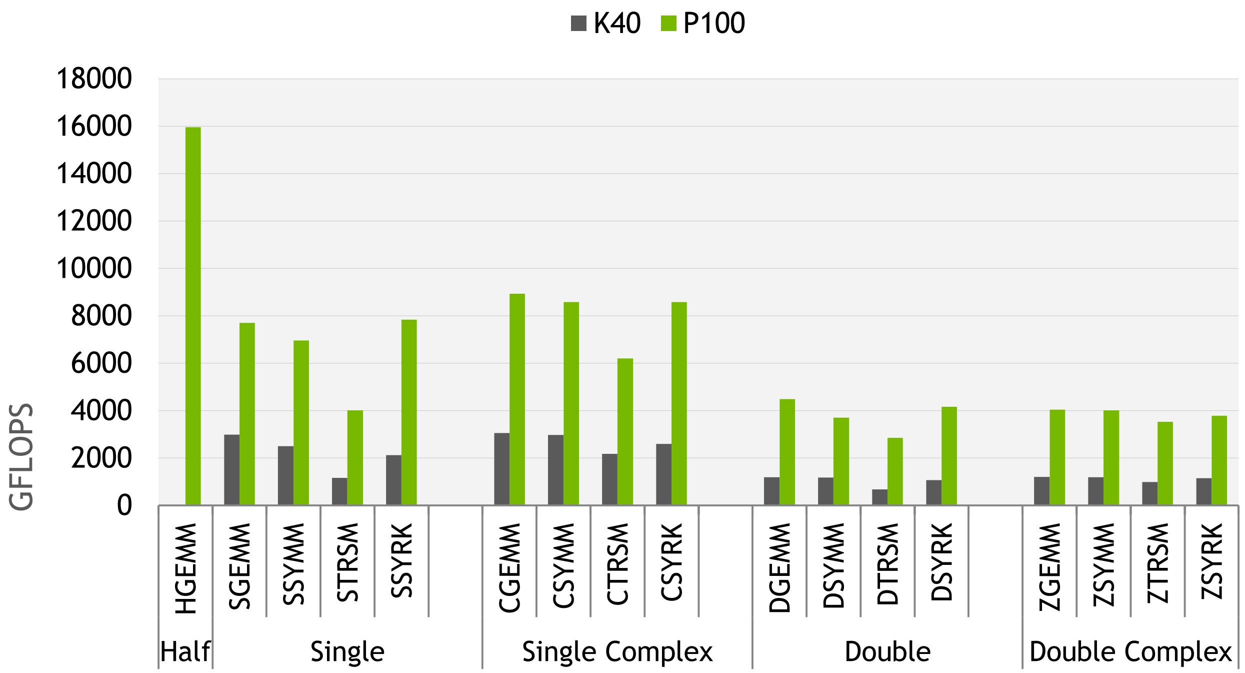


- Comparing GEMM performance on K40m (FP32) and P100 (FP32 and FP16)
- cuBLAS 8 on P100, Base clocks (r361)
- cuBLAS 8 on P40, Base clocks (r367)
- cuBLAS 7.5 on K40m, Base clocks, ECC ON (r352)
- Input and output data on device
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3@ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory
- m=n=k=4096

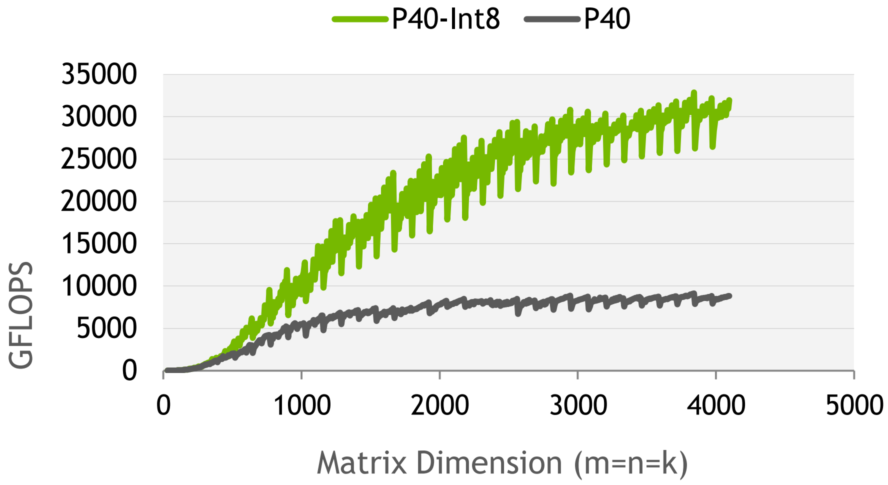
https://developer.nvidia.com/cublas

cuBLAS: > 8 TFLOPS SINGLE PRECISION

16 TFLOPS FP16 GEMM Performance



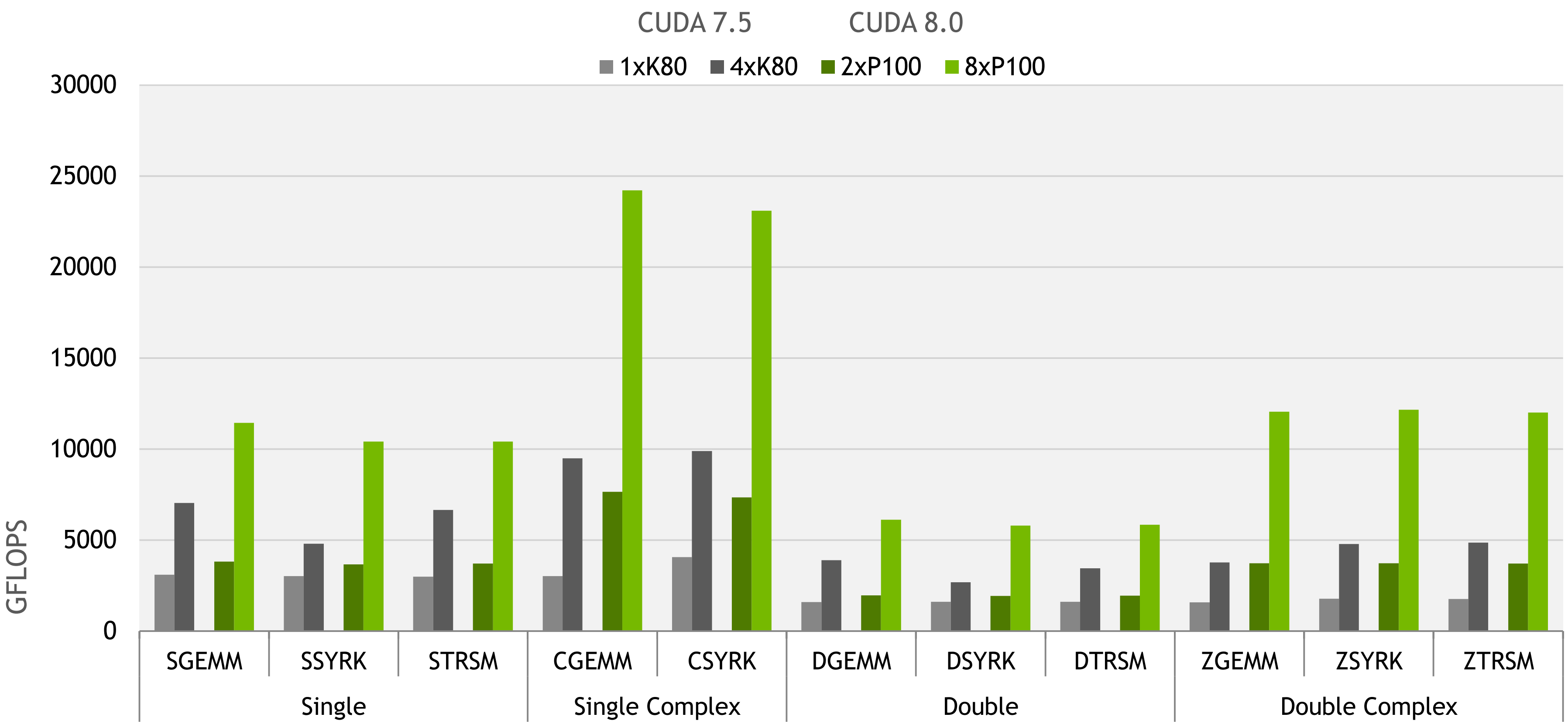
32 TFLOPS INT8 GEMM Performance



- cuBLAS 8 on P100 (r361) and P40 (r367) ; Base clocks
- cuBLAS 7.5 on K40m ; Base clocks, ECC ON (r352)
- Input and output data on device
- m=n=k=4096, transpose=no, side=right, fill=lower
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3@ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

cuBLAS-XT: > 24 TFLOPS ON A SINGLE NODE



- cuBLAS 8 on P100 (r361); Base clocks
- cuBLAS 7.5 on K80 ; Base clocks, ECC ON (r352)
- 1xK80 indicates 2-GPUs (or one K80 board)
- Input and output data on device
- m=n=k=4096, transpose=no, side=right, fill=lower
- Host system: Intel Xeon Haswell dual-socket 22-core E5-2699 v4@ 2.2GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 256GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

cuSPARSE

Sparse Linear Algebra on GPUs

Optimized Sparse Matrix Library

- Optimized sparse linear algebra BLAS routines for matrix-vector, matrix-matrix, triangular solve
- Support for variety of formats (CSR, COO, block variants)
- Incomplete-LU and Cholesky preconditioners
- Support for half-precision (fp16) sparse matrix-vector operations

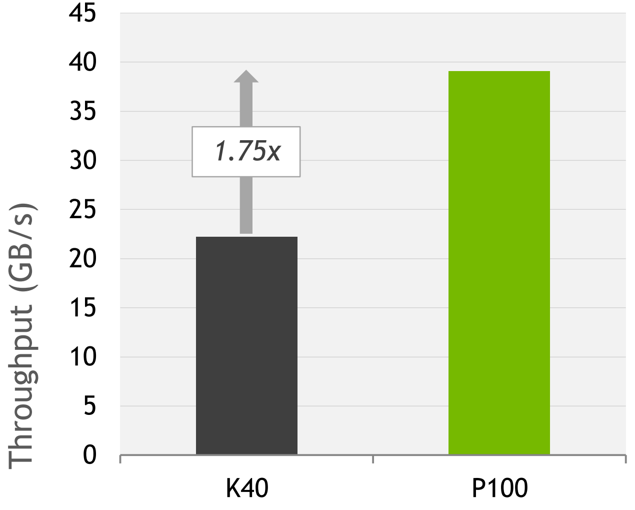
https://developer.nvidia.com/cusparse

GEMVI – Dense Matrix X Sparse Vector

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \alpha \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} \begin{bmatrix} - \\ 2.0 \\ - \\ 4.0 \end{bmatrix} + \beta \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

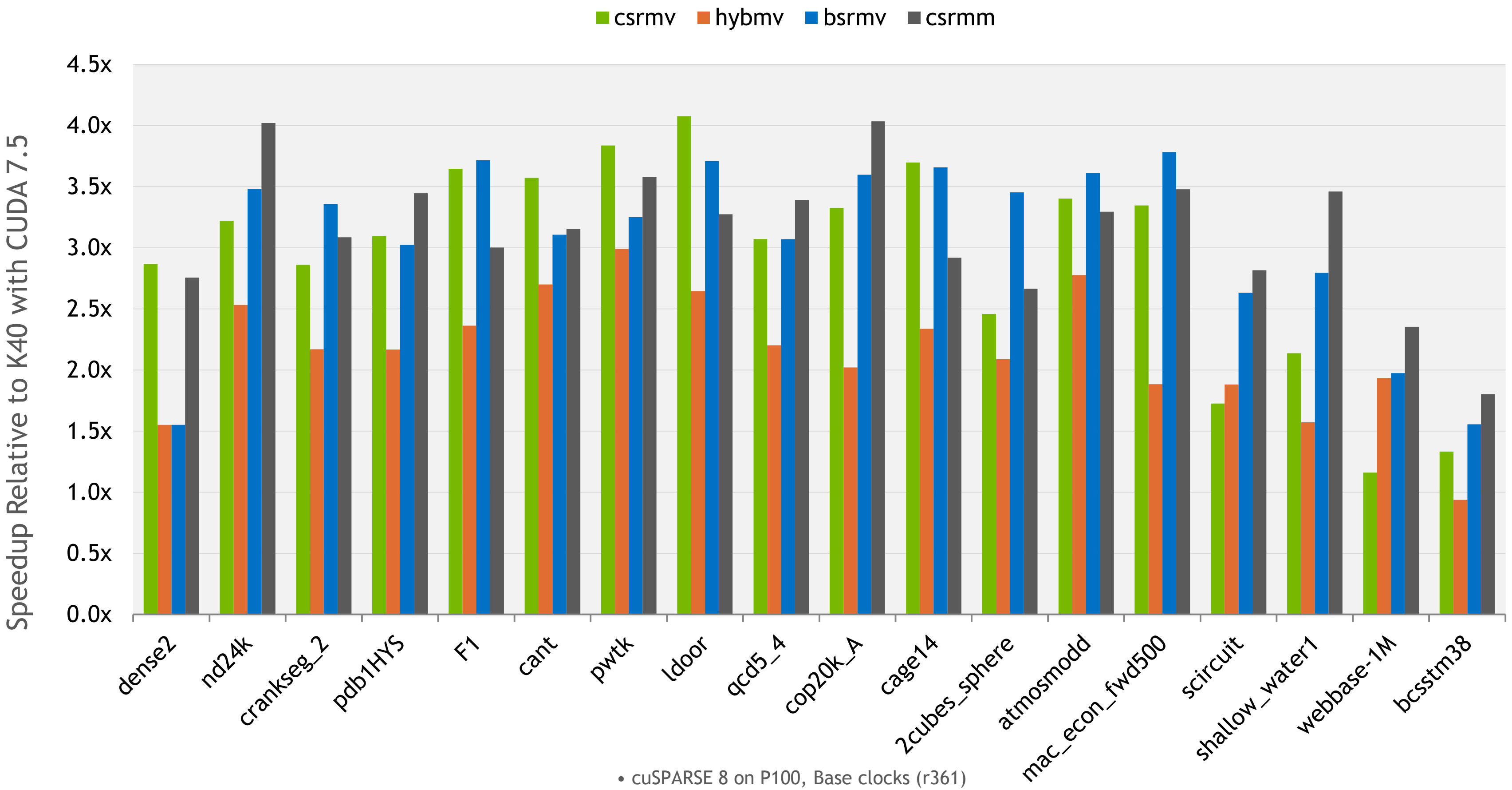
Used in language modeling and auto-encoders for recommender systems

Improved GEMVI Performance with P100



- cuSPARSE 8 on P100, Base clocks (r361)
- cuSPARSE 7.5 on K40m, Base clocks, ECC ON (r352)
- Input and output data on device
- Dense matrices with 1e6 columns and 1e3 rows; Sparse vectors with less than 100 non-zeros out of 1e6 locations
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

cuSPARSE: > 4X FASTER WITH P100



- cuSPARSE 8 on P100, Base clocks (r361)
- cuSPARSE 7.5 on K40m, Base clocks, ECC ON (r352)
- Input and output data on device; Average speedup across different data types
- Matrices obtained from: http://www.cise.ufl.edu/research/sparse/matrices/
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

cuSOLVER

Linear Solver Library

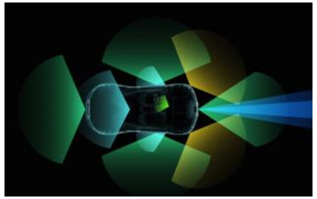
Library for Dense and Sparse Direct Solvers

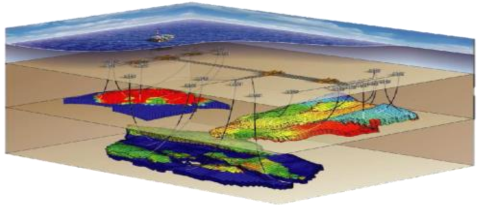
Supports Dense Cholesky, LU, (batched) QR, SVD and Eigenvalue solvers (new in CUDA 8)

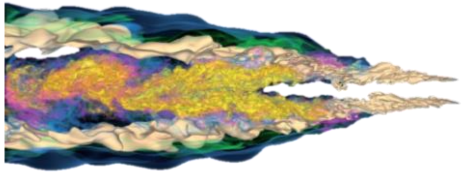
Sparse direct solvers & Eigensolvers

Includes a sparse refactorization solver for solving sequences of matrices with a shared sparsity pattern

Used in a variety of applications such as circuit simulation and computational fluid dynamics



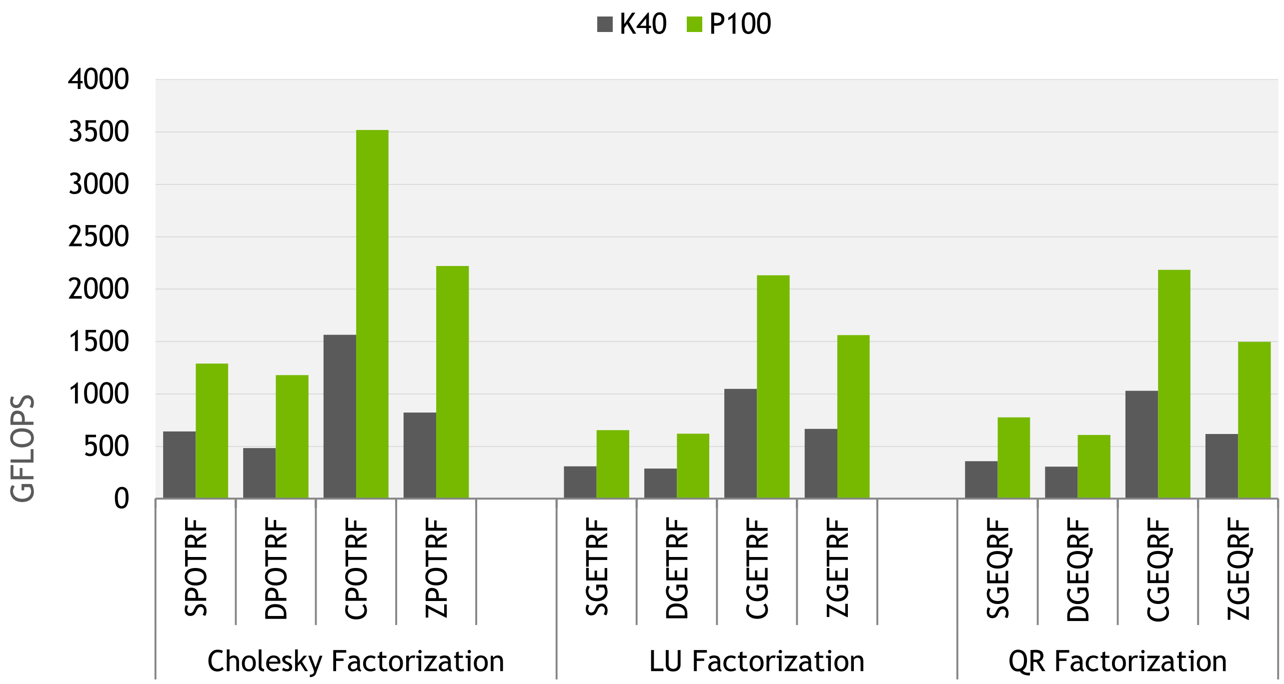


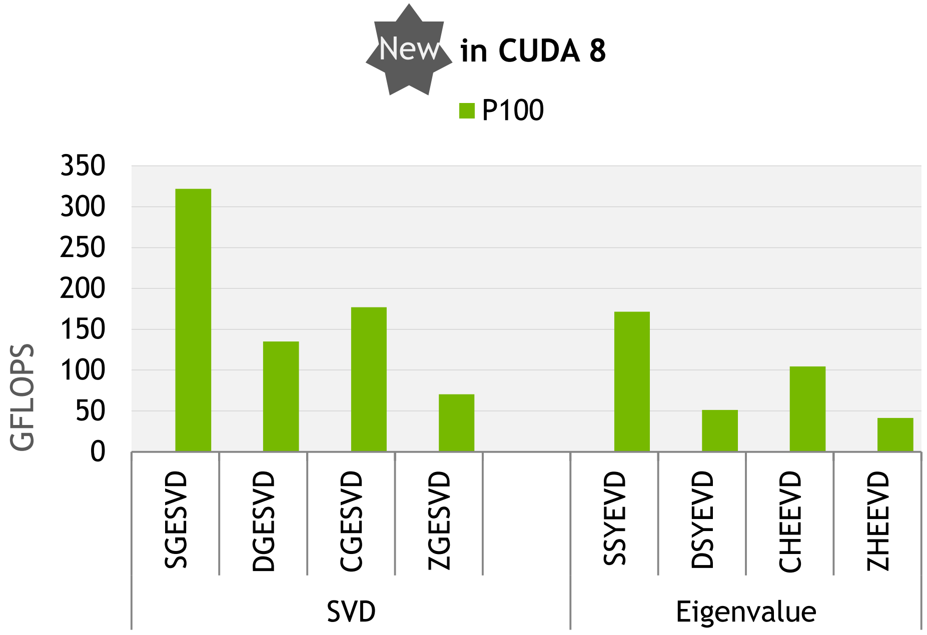


Sample Applications
- Computer Vision
- CFD
- Newton's method
- Chemical Kinetics
- Chemistry
- ODEs
- Circuit Simulation

https://developer.nvidia.com/cusolver

DENSE PERFORMANCE: > 2X FASTER





- cuSOLVER 8 on P100, Base clocks (r361)
- cuSOLVER 7.5 on K40m, Base clocks, ECC ON (r352)
- M=N=4096
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

AmgX

Algebraic Multi-Grid Solvers

Flexible Solver Composition System

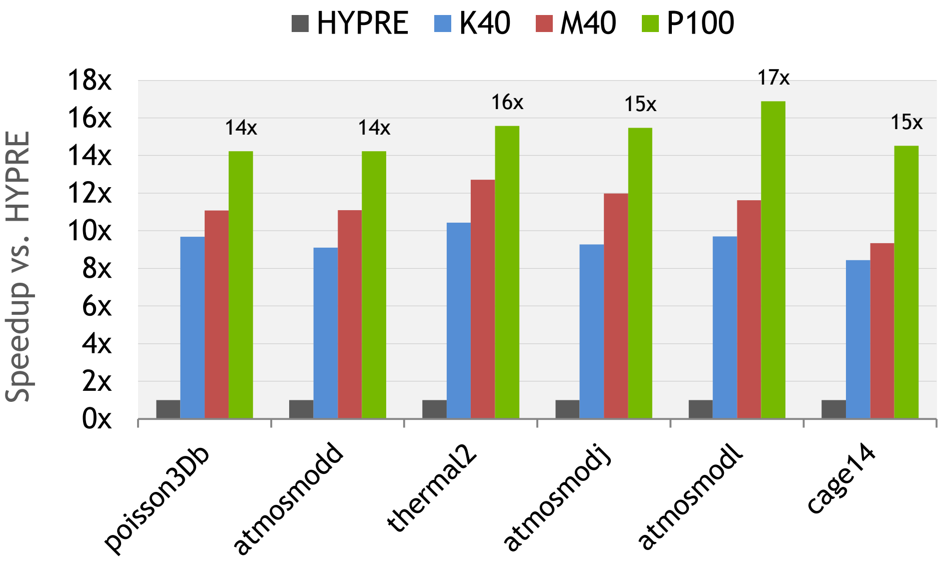
Easy construction of complex nested solvers and pre-conditioners

Flexible and simple high level C API that abstracts parallelism and GPU implementation

Includes Ruge-Steuben, un-smoothed aggregation, Krylov methods and different smoother algorithms

https://developer.nvidia.com/amgx

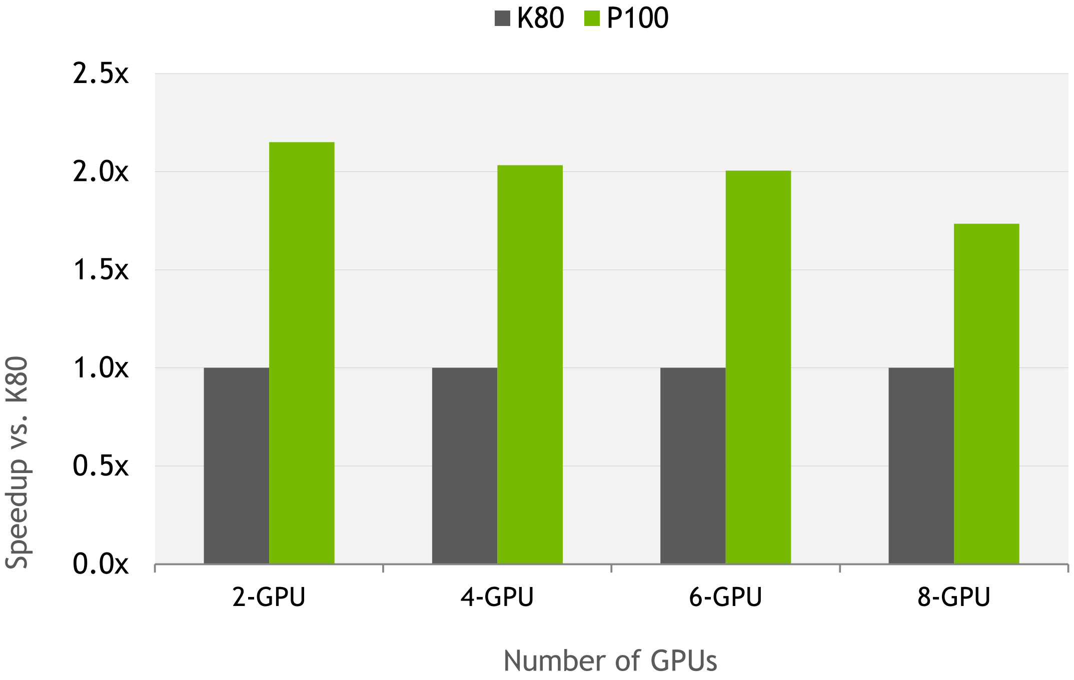
> 15x Speedup vs HYPRE



- Florida Matrix Collection; Total Time to Solution
- HYPRE AMG Package (http://acts.nersc.gov/hypre) on Intel Xeon E5-2697 v4@2.3GHz, 3.6GHz Turbo, Hyperthreading off
- AmgX on K40, M40, P100 (SXM2); Base clocks
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

> 2X FASTER SOLUTION TIMES WITH P100

SPE10 Using Multi-GPUs



- AmgX on K80 and M40 with CUDA 7.5, and P100 (PCIe) with CUDA 8; Base clocks, ECC ON
- Society of Petroleum Engineers 10th comparative simulation model (SPE10)
- Matrix size varied with 100 x Number of GPUs x 200 x 50 for SPE10
- Time to solution includes both setup and solve times

Performance may vary based on OS and software versions, and motherboard configuration

GPU ACCELERATED LIBRARIES:

PARALLEL PRIMITIVES, IMAGE PROCESSING AND RNGs

THRUST

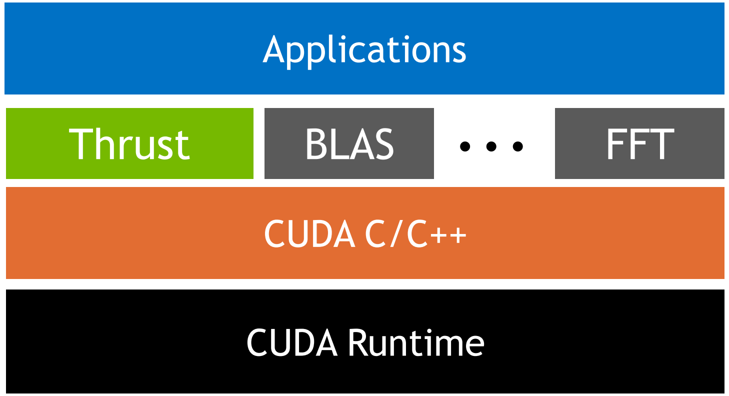
Parallel Primitives and Data Structures

Templated Library of Algorithms and Data Structures

GPU-accelerted scan, sort, transform and reduce

Improved developer productivity via C++ template library, allowing developer to focus on high-level tasks

Library simplifies memory management and data transfer


Applications

Thrust BLAS • • • FFT

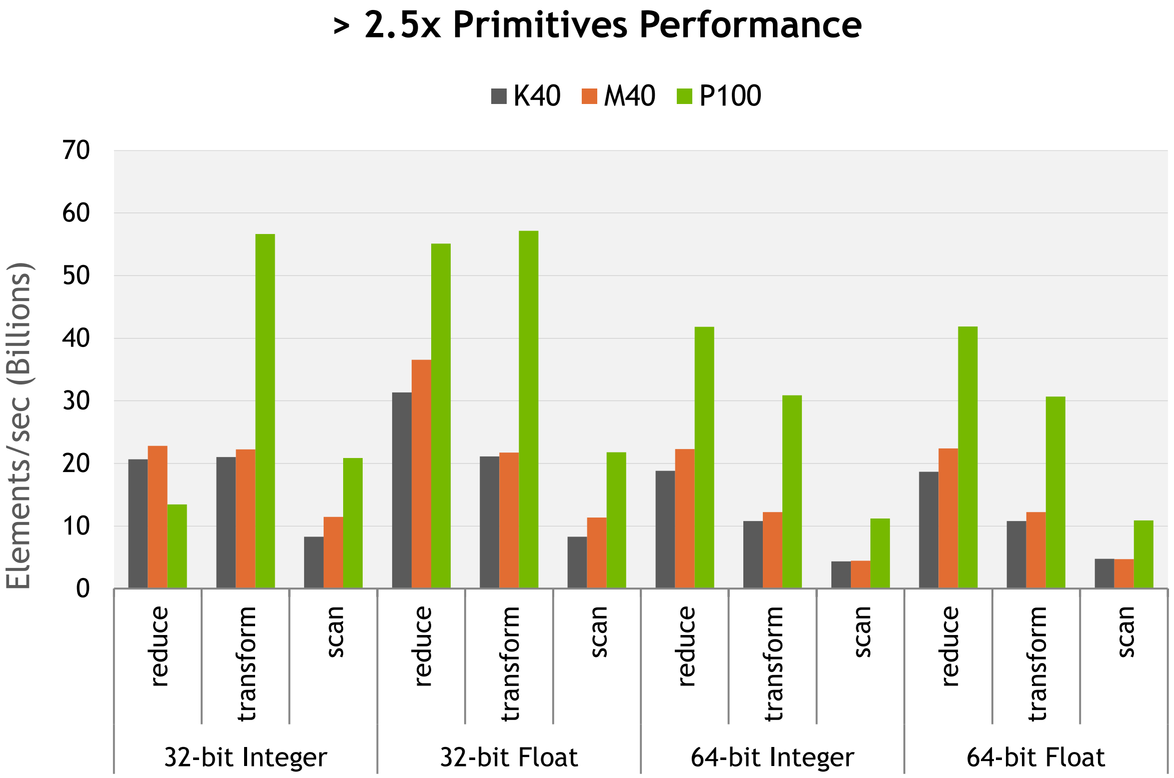
CUDA C/C++

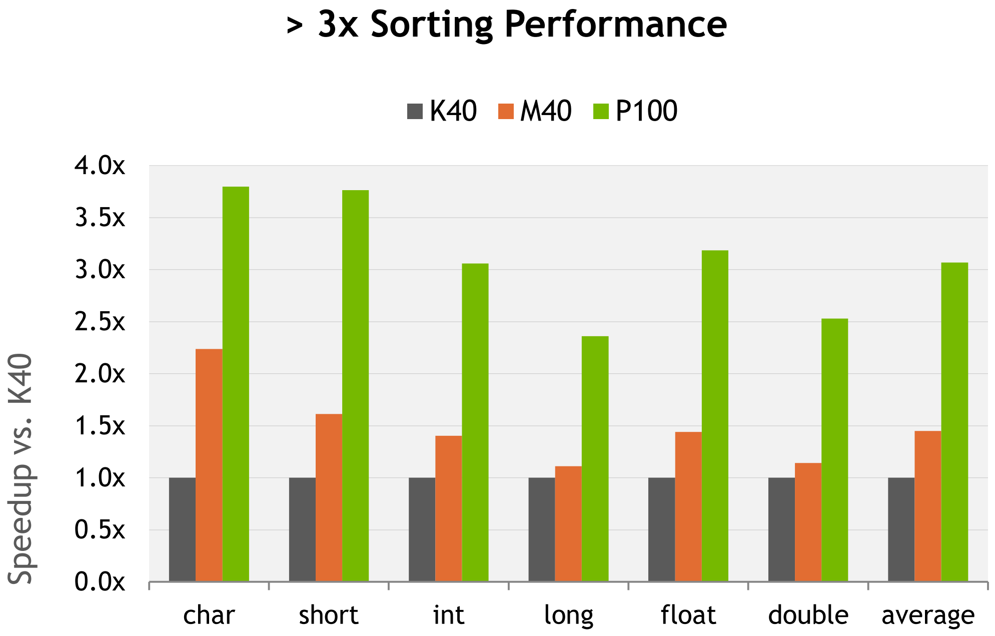
CUDA Runtime

Thrust is a C++ template library of parallel algorithms inspired by C++ STL

https://developer.nvidia.com/thrust

HIGH PERFORMANCE PARALLEL PRIMITIVES





- Thrust 8 on P100, Base clocks (r361)
- Thrust 7.5 on K40m, M40, Base clocks, ECC ON (r352)
- Input and output data on device
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

NPP

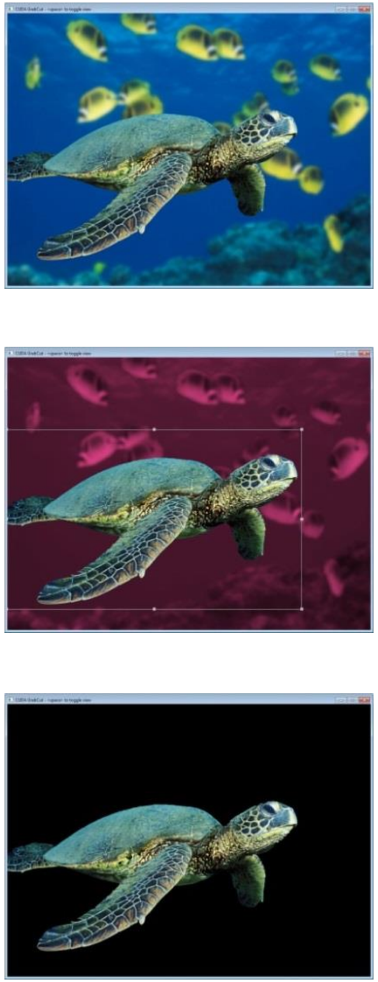
NVIDIA Performance Primitives Library

GPU-accelerated Building Blocks for Image, Video Processing & Computer Vision

Over 2500 image, signal processing and computer vision routines

Color transforms, geometric transforms, move operations, linear filters, image & signal statistics, image & signal arithmetic, building blocks, image segmentation, median filter, BGR/YUV conversion, 3D LUT color conversion

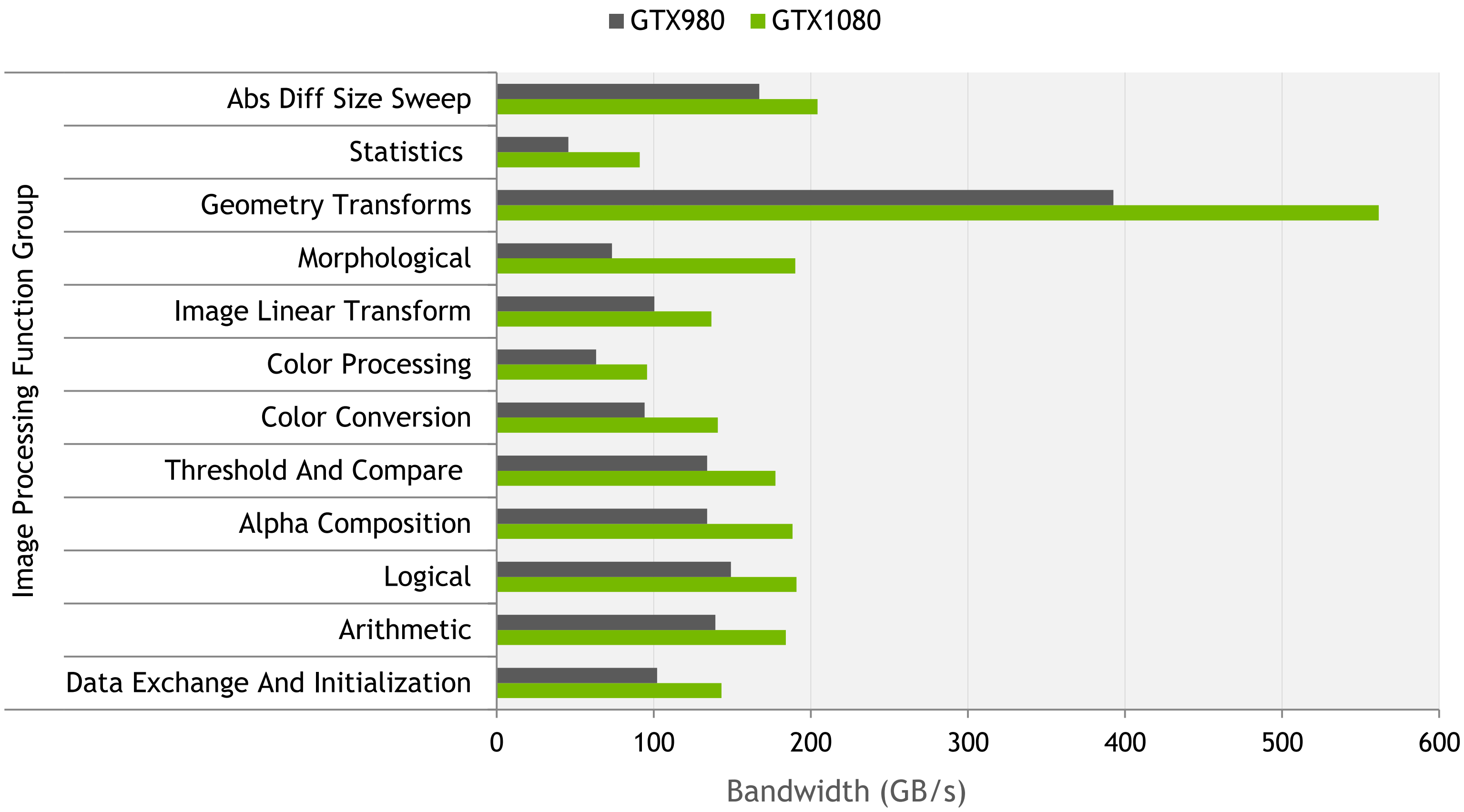
Eliminate unnecessary copying of data to/from CPU memory



“Grabcut” example using NPP graphcut primitive

https://developer.nvidia.com/npp

> 2X FASTER IMAGE PROCESSING WITH PASCAL



- NPP 7.5 on GTX980 (GM204) and NPP 8 on GTX1080, Base clocks
- Input and output data on device
- Each bar represents the average bandwidth over all routines in the function group
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

GPU ACCELERATED LIBRARIES:

RNGs

cuRAND

Random Number Generation (RNG) Library

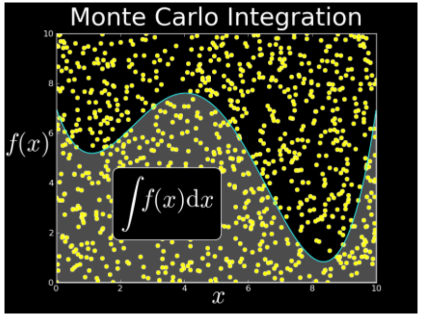
High Performance Random Number Generation

Flexible interfaces for RNG on the host or within GPU kernels

Pseudo- and Quasi-RNGs — MRG32k3a, MTGP Mersenne Twister, XORWOW, Sobol

Supports several output distributions

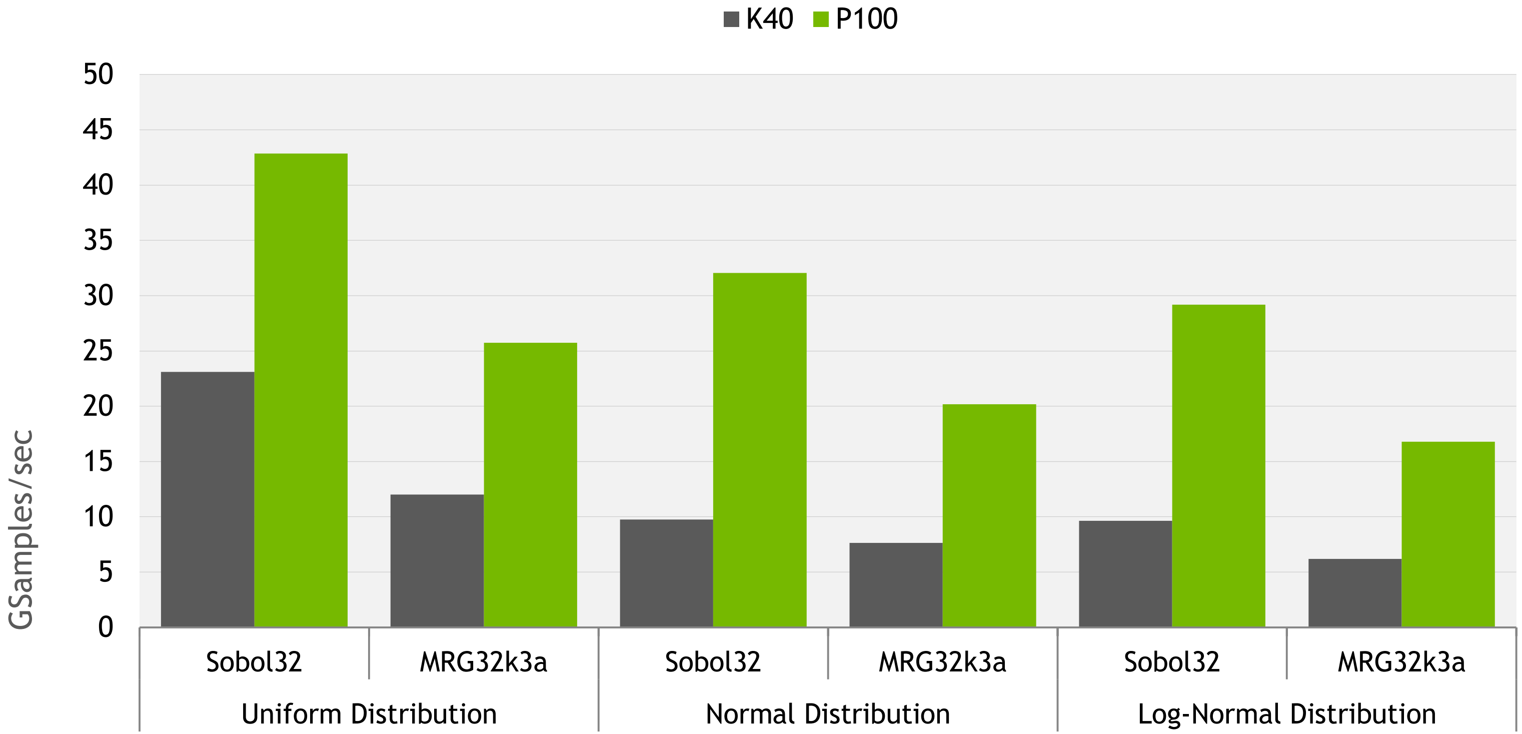
Tested against well-known statistical test batteries (test results available in documentation)



Sample Applications of cuRAND

https://developer.nvidia.com/curand

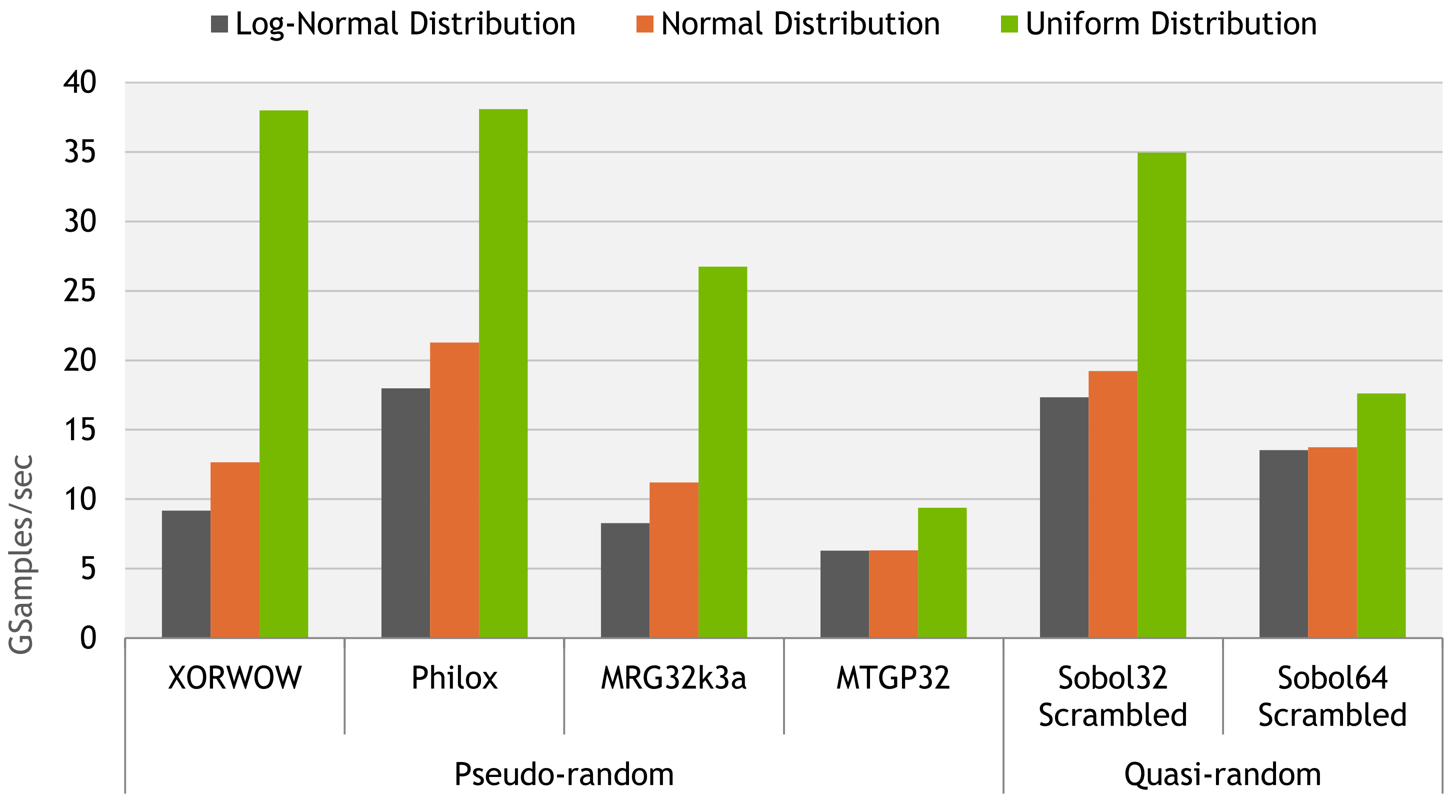
cuRAND: > 3X SPEEDUPS WITH P100



- cuRAND 8 on P100, Base clocks (r361)
- cuRAND 7.5 on K40m, Base clocks, ECC ON, (r352)
- Single-precision input and output data on device
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

Performance may vary based on OS and software versions, and motherboard configuration

cuRAND: HIGH PERFORMANCE RNGS



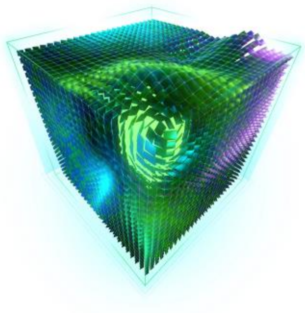
Performance may vary based on OS and software versions, and motherboard configuration

- cuRAND 8 on P100, Base clocks (r361)
- Double-precision input and output data on device
- Host system: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory

CUDA 8 – DOWNLOAD TODAY!

Everything You Need to Accelerate Applications

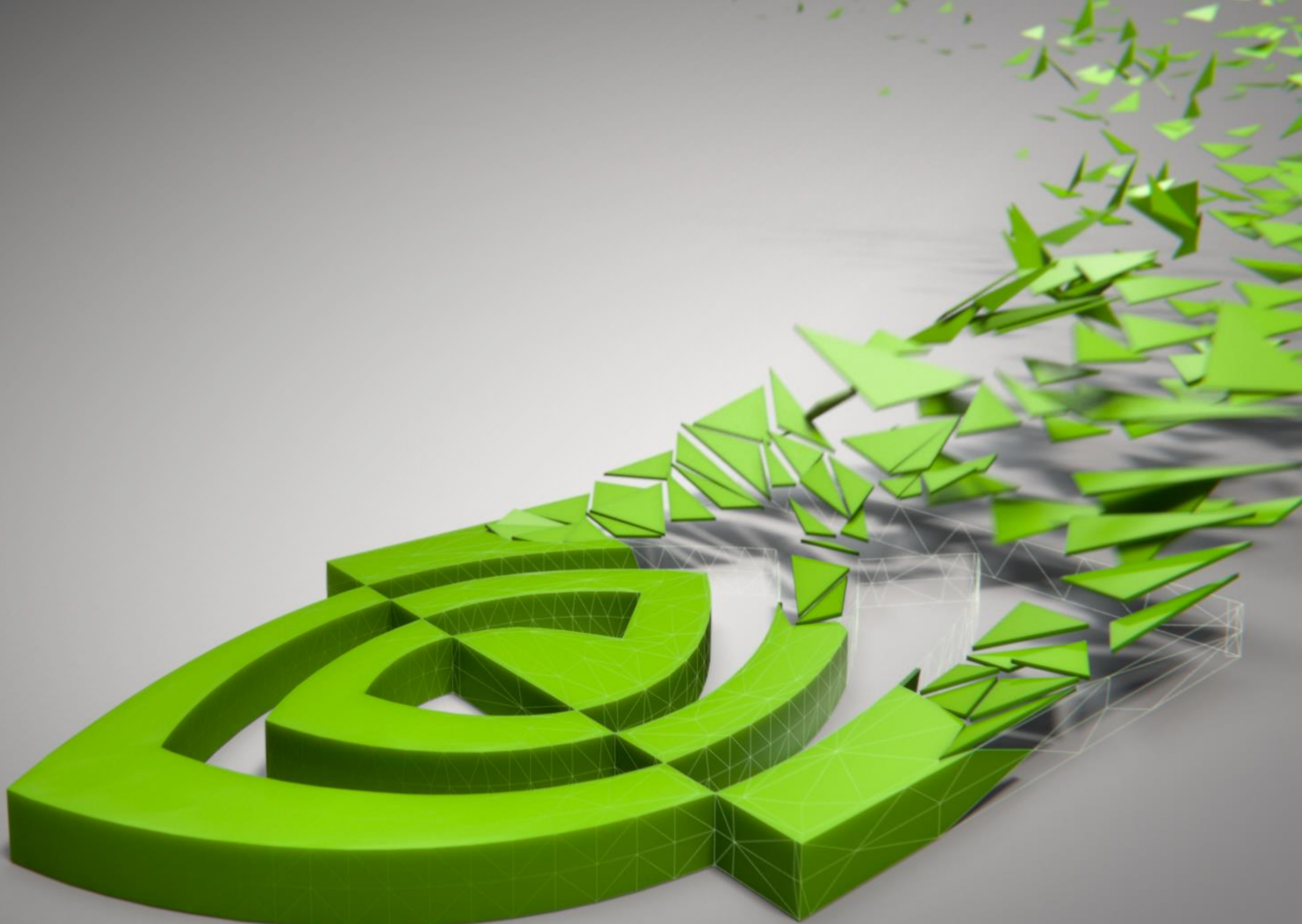
cuda 8 download



- CUDA Applications in your Industry: www.nvidia.com/object/gpu-applications-domain.htm
- Additional Webinars:
  - Inside PASCAL
  - Deep Learning and Hyperscale Performance Overview
  - CUDA 8 Tools
  - CUDA 8 Unified Memory
- CUDA 8 Release Notes: www.docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html#abstract

developer.nvidia.com/cuda-toolkit

NVIDIA



APPENDIX

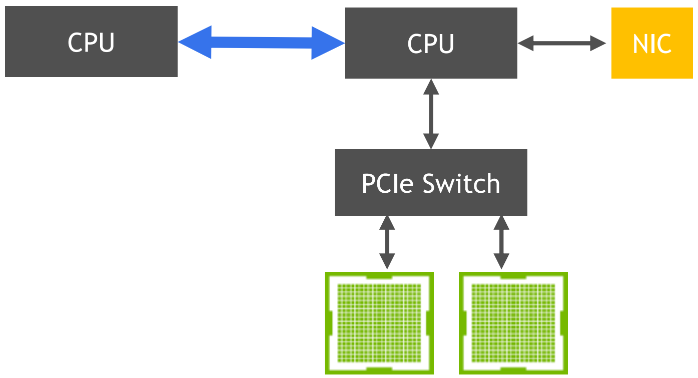
CONFIGURATIONS USED FOR HPC APPS

Application	Dataset and CUDA	Metric
HOOMD-BLUE – Particle dynamics package written grounds up for GPUs http://codeblue.umich.edu/hoomd-blue/index.html	lj_liquid_1m	Avg. Timesteps (seconds)
HPCG – Benchmark that exercises computational and data access patterns that closely match a broad set of important HPC applications http://www.hpcg-benchmark.org/index.html	Local Size: 256x256x256 CUDA Versions: CUDA 6.5.19	GFLOPS/second
LAMMPS – Classical molecular dynamics package http://lammps.sandia.gov/index.html	EAM Mixed Precision; Problem Size: 4x4x4; 2048000 atoms	Avg. Loop Time (seconds)
QUDA -- A library for Lattice Quantum Chromo Dynamics on GPUs https://lattice.github.io/quda/	QUDA (GPU): Dslash Wilson-Clover Precision: Double Problem Size: 32x32x32x64 QPhiX (CPU): Dslash Wilson-Clover Precision: Double Problem Size: 32x32x32x64	GFLOPS
MiniFE – Finite Element Analysis https://mantevo.org/about/applications	Problem Size: 350x350x350	Total Time (seconds)
MILC – Lattice Quantum Chromodynamics (LQCD) codes simulate how elemental particles are formed and bound by the “strong force” to create larger particles like protons and neutrons http://physics.indiana.edu/~sg/milc.html	Precision: Double	Total Time (seconds)
VASP http://www.vasp.at/index.php/news/44-administrative/115-new-release-vasp-5-4-1-with-gpu-support	Silica IFPEN	Elapsed Time (seconds)

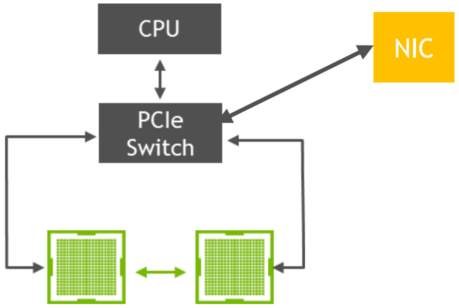
CPU Server:	Intel Xeon dual-socket 22-core E5-2699 v4@2.2GHz 3.6GHz Turbo with CentOS 7.2 x86_64 and 128GB system memory
GPU Host Server:	Intel Xeon dual-socket 20-core E5-2698 v4@2.2GHz 3.6GHz Turbo with Tesla P100 SXM2 ; Ubuntu 14.04.4 x86_64 with 128GB system memory
GPU Host Server:	Intel Xeon dual-socket 22-core E5-2699 v4@2.2GHz 3.6GHz Turbo with Tesla K80 ; CentOS 7.2 x86_64 and 128GB system memory
CUDA Versions:	CUDA 8.0.44 with r361.96 (P100) and 361.79 (K80) unless otherwise specified

SYSTEM TOPOLOGY USED FOR K80 AND P100

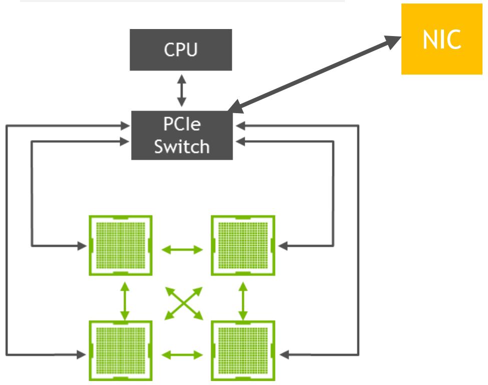
2 CPU – 2x (K80)



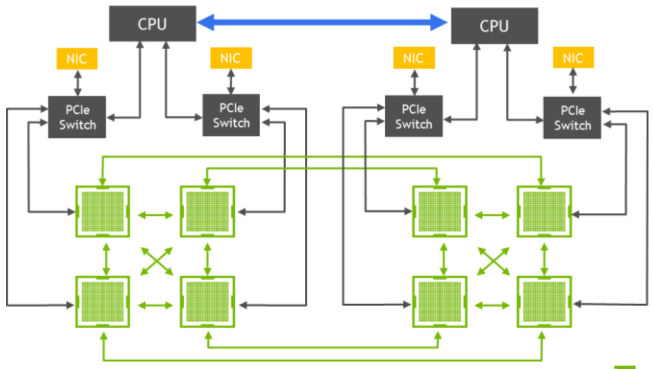
2 CPU – 2x P100 SXM2



2 CPU – 4x P100 SXM2



2 CPU – 8x P100 SXM2



NVLink

PCIE Gen 3