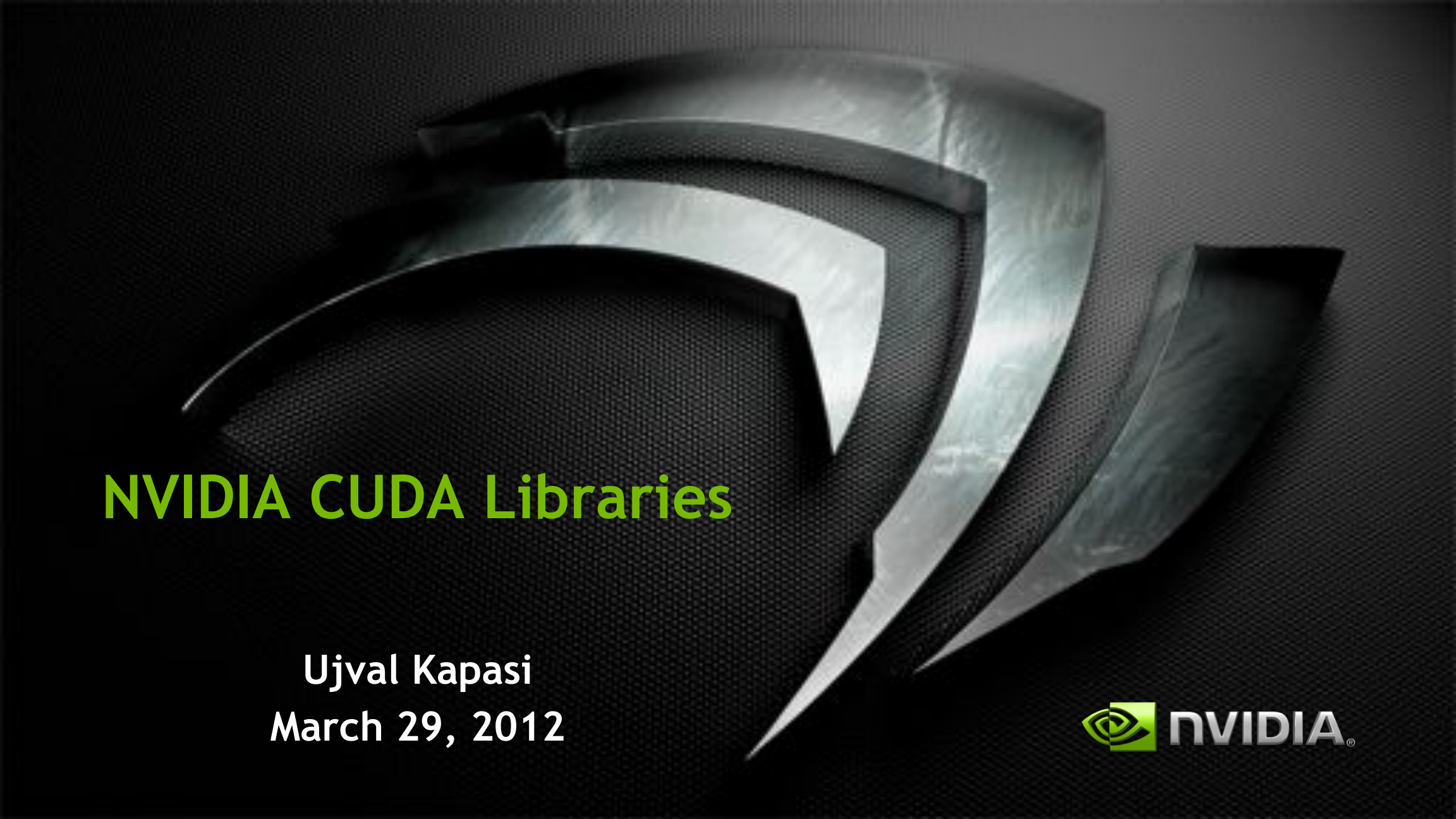
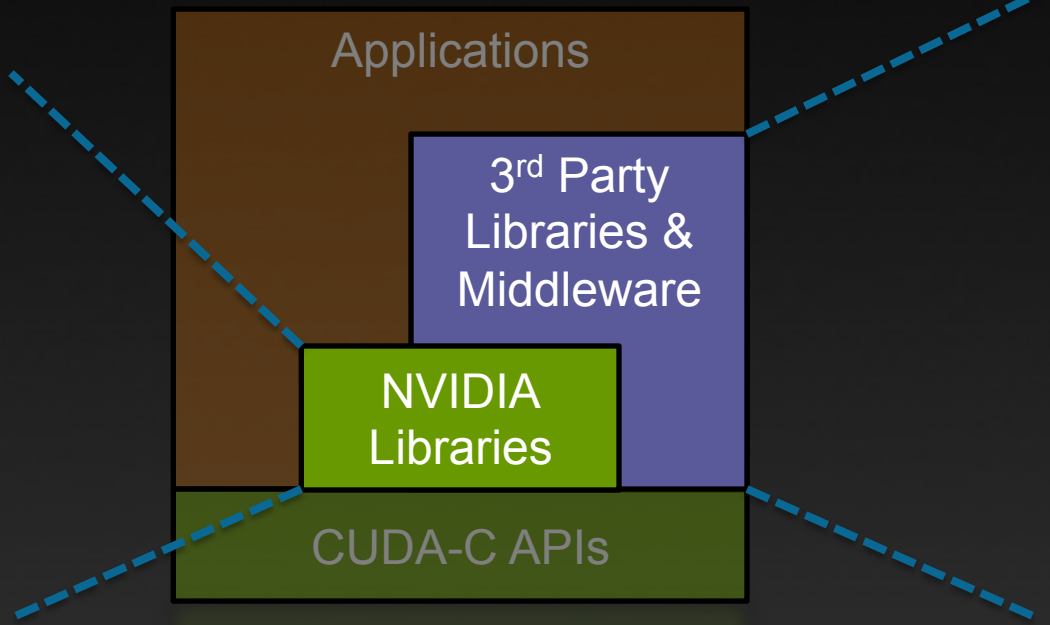

NVIDIA CUDA Libraries
Ujval Kapasi
March 29, 2012
NVIDIA

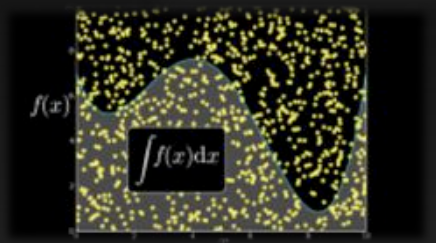
CUDA Libraries and Middleware

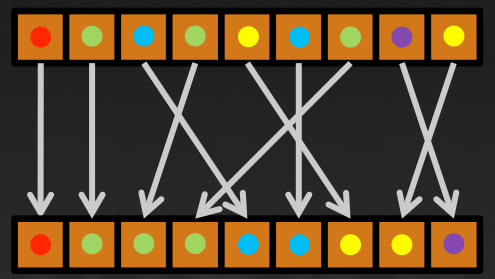
Basic building blocks
Easy to use
Fast



Programming language integration
Hybrid computation
Multi-node and multi-gpu
Domain specific languages

NVIDIA CUDA Libraries

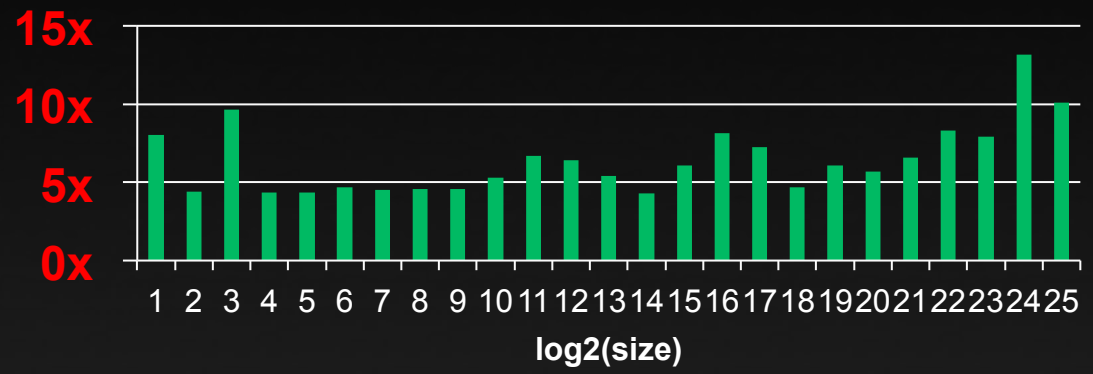




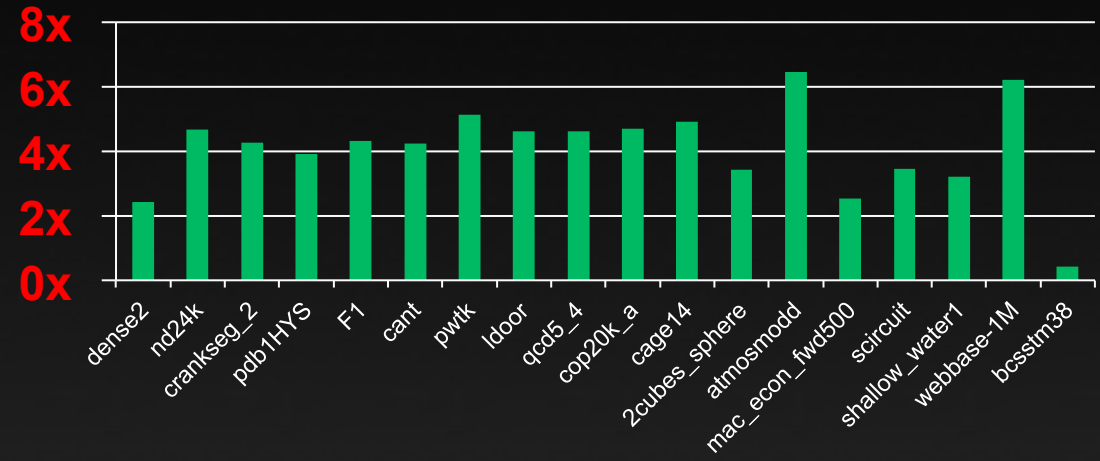
CUBLAS	*dense linear algebra*
CUSPARSE	*sparse linear algebra*
CUFFT	*discrete Fourier transforms*
CURAND	*random number generation*
NPP	*signal and image processing*
Thrust	*scan, sort, reduce, transform*
math.h	*floating point*
system calls	*printf, malloc, assert*

4x-10x speedups over CPU on single precision

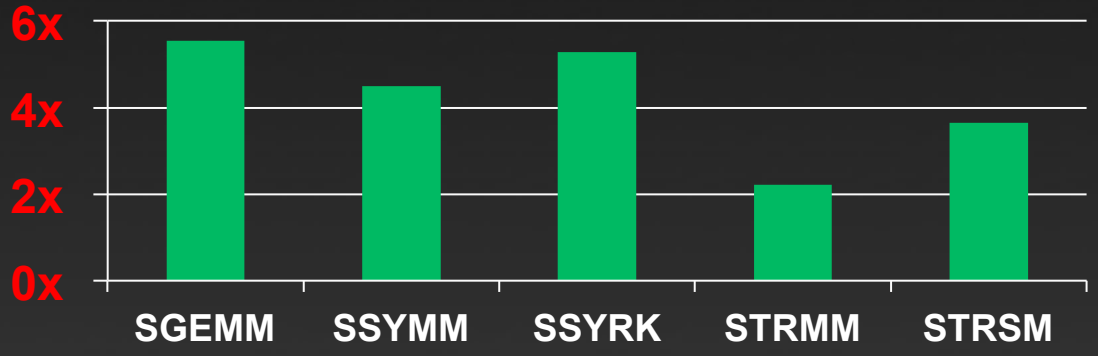
CUFFT



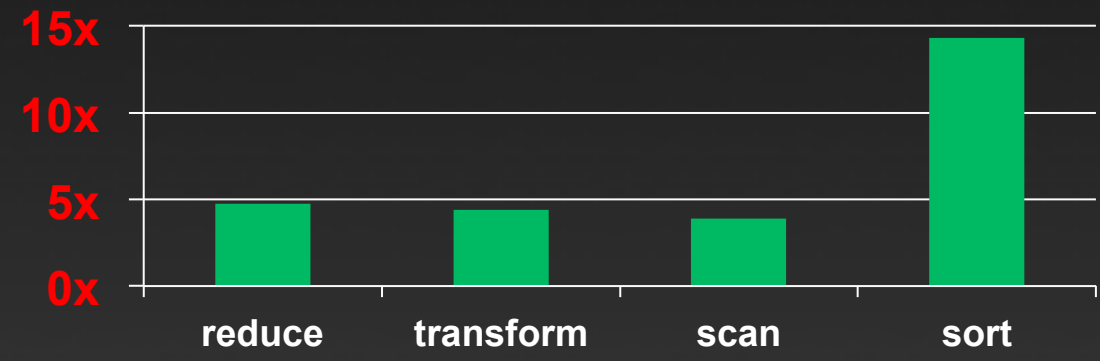
CUSPARSE



CUBLAS



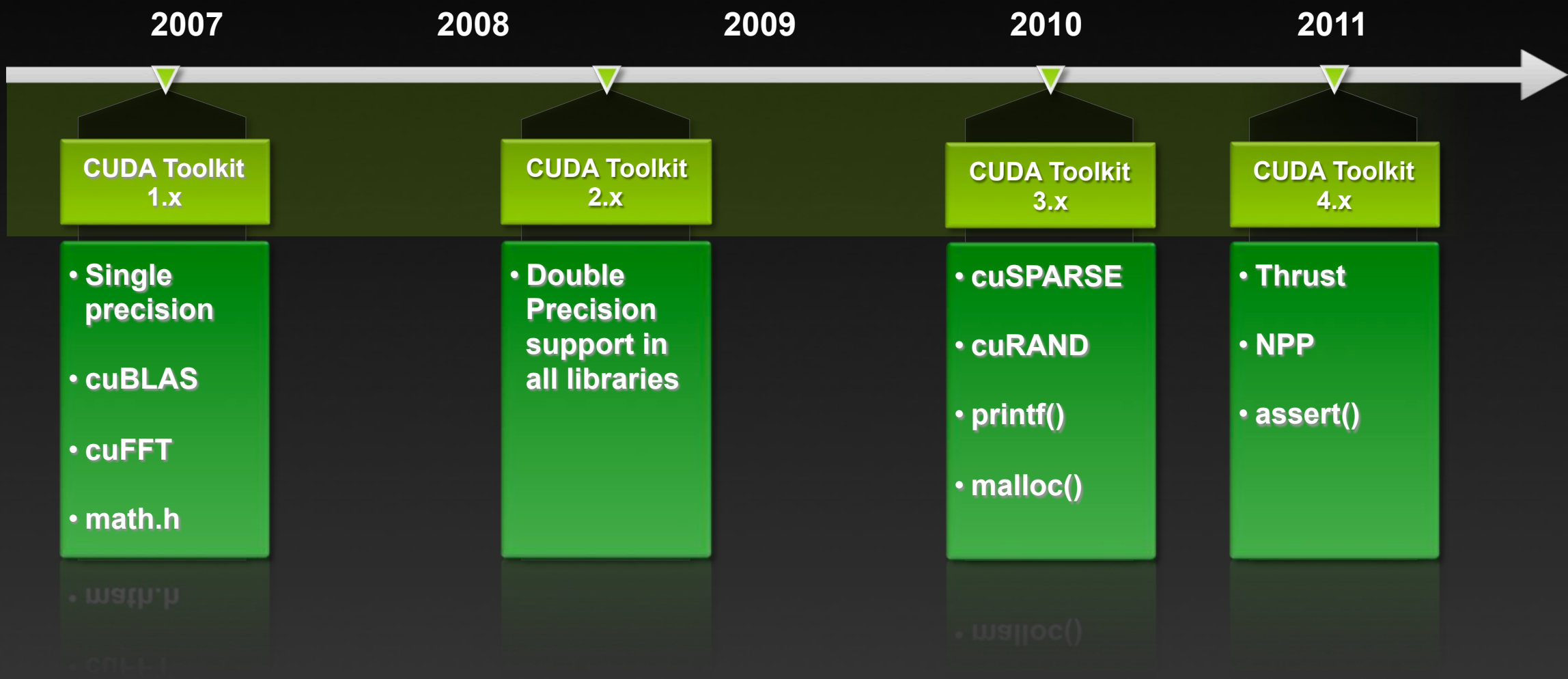
Thrust



NOTE: ROUTINES ASSUME INPUT AND OUTPUT DATA ARE IN GPU MEMORY

- CUDA 4.1 on Tesla M2090, ECC on
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz

Challenge #1:
MATCH AN ESTABLISHED CPU ECOSYSTEM

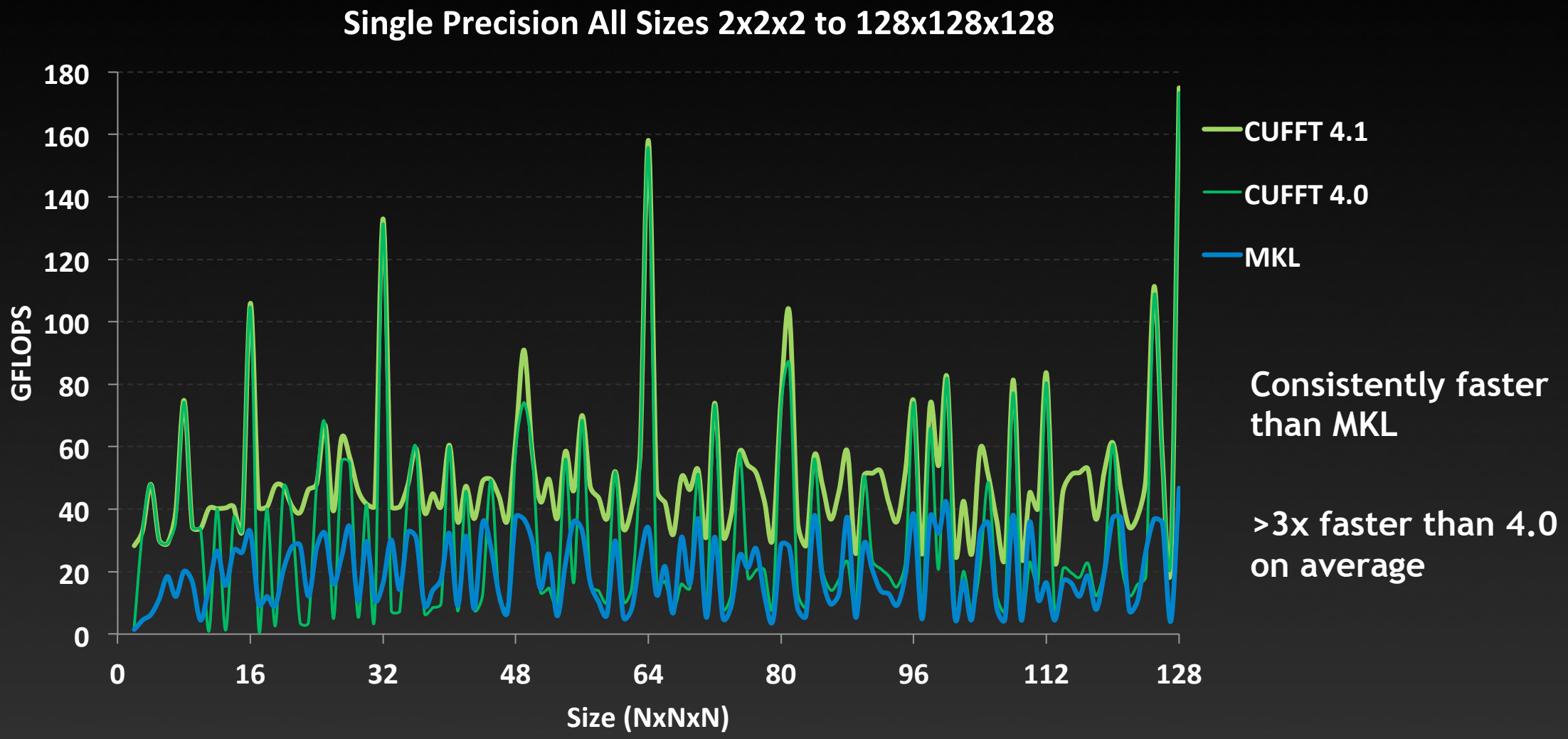


CUDA ecosystem today spans many domains

- Languages (FORTRAN, python)
- Math, Numerics, Statistics
- Dense & Sparse Linear Algebra
- Algorithms (sort, etc.)
- Image Processing
- Computer Vision
- Signal Processing
- Finance

http://developer.nvidia.com/gpu-accelerated-libraries

Challenge #2: PREDICTABLE ACCELERATION



Consistently faster than MKL

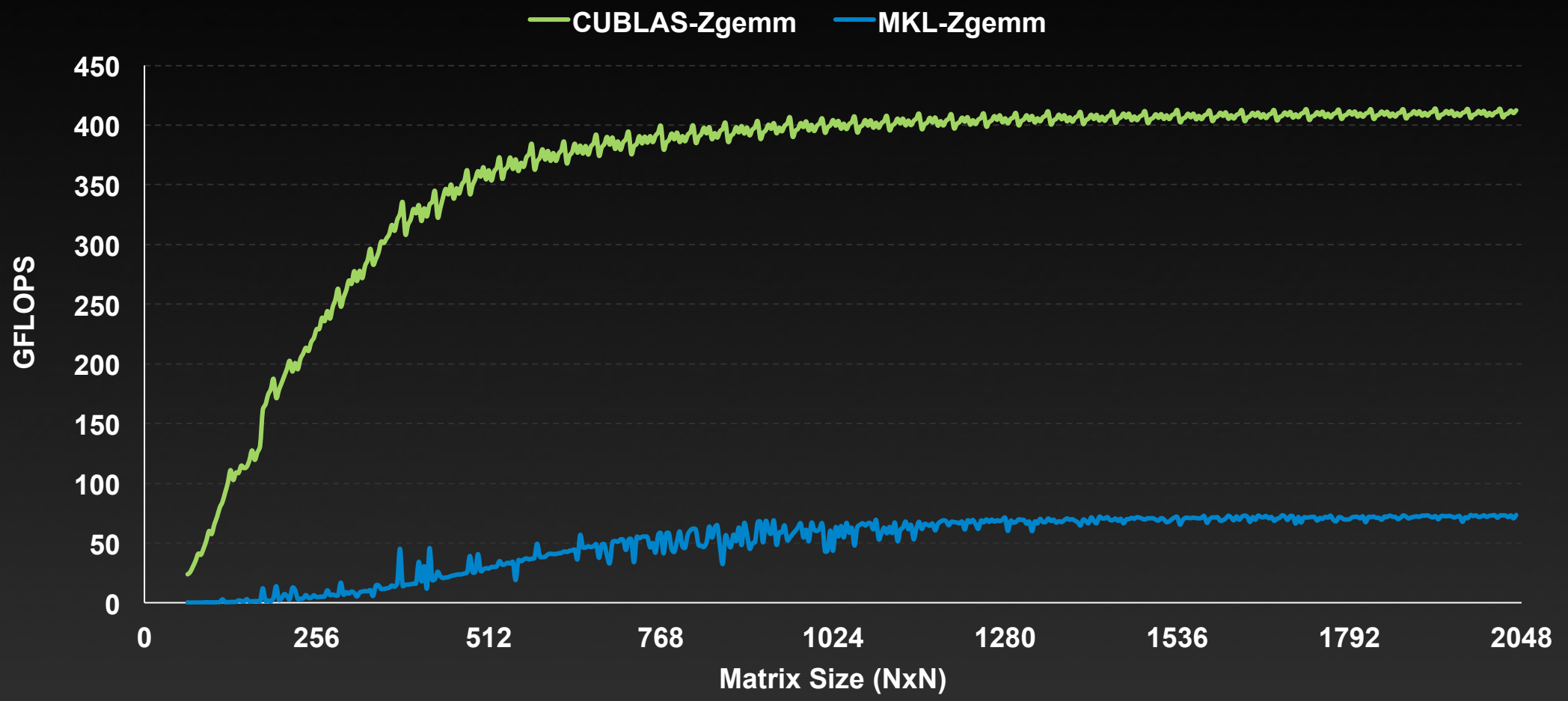
>3x faster than 4.0 on average

NOTE: ROUTINES ASSUME INPUT AND OUTPUT DATA ARE IN GPU MEMORY

Performance may vary based on OS version and motherboard configuration

- cuFFT 4.1 and cuFFT 4.0 on Tesla M2090, ECC on
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz

Challenge #2: PREDICTABLE ACCELERATION

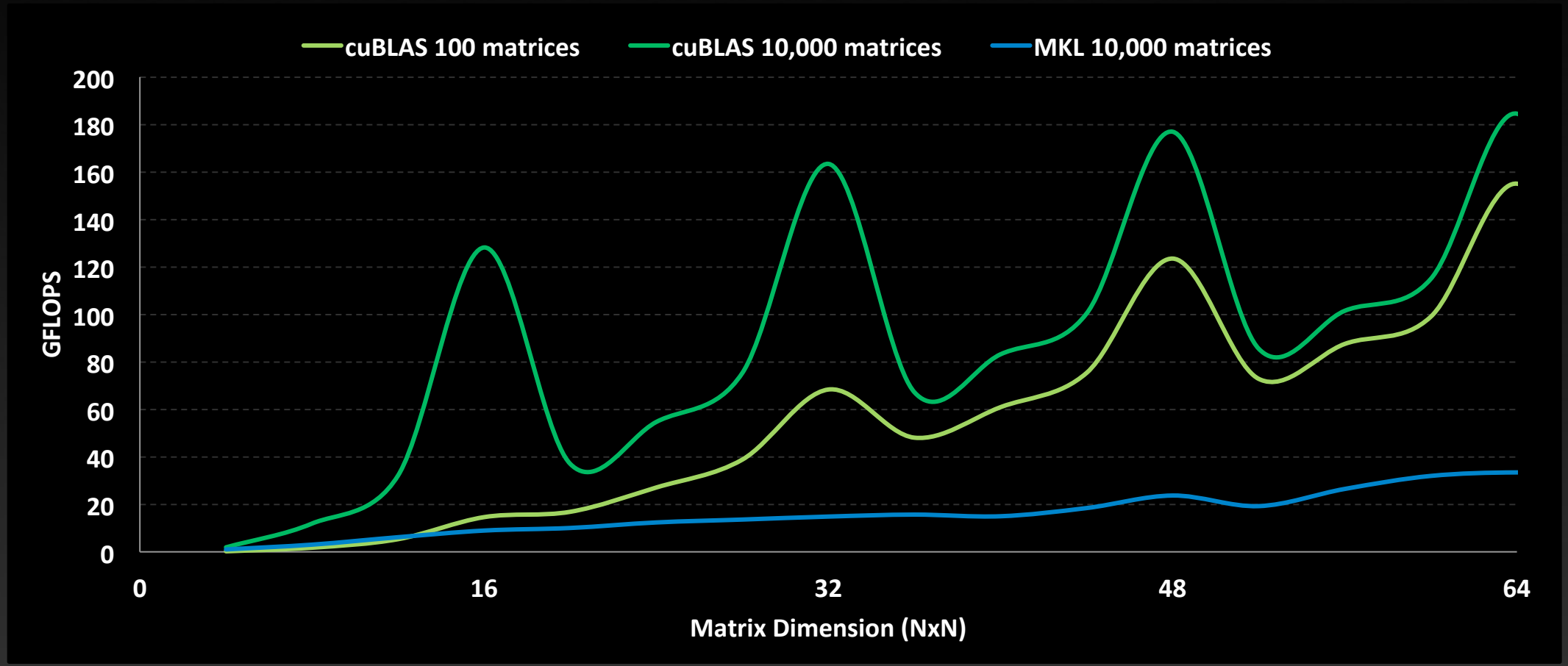


NOTE: ROUTINES ASSUME INPUT AND OUTPUT DATA ARE IN GPU MEMORY

Performance may vary based on OS version and motherboard configuration

- cuBLAS 4.1 on Tesla M2090, ECC on
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz

Challenge #2: PREDICTABLE ACCELERATION



NOTE: ROUTINES ASSUME INPUT AND OUTPUT DATA ARE IN GPU MEMORY

Performance may vary based on OS version and motherboard configuration

- cuBLAS 4.1 on Tesla M2090, ECC on
- MKL 10.2.3, TYAN FT72-B7015 Xeon x5680 Six-Core @ 3.33 GHz

- CUDA Libraries accelerate basic building block algorithms
  - BLAS → CUBLAS, FFTW → CUFFT, STL → Thrust
  - Enables a wide range of 3rd party libraries and middleware

- CUDA Libraries coverage and performance constantly improving
  - Automatic performance improvements with new CUDA releases and new GPU architectures


Thank You!
NVIDIA