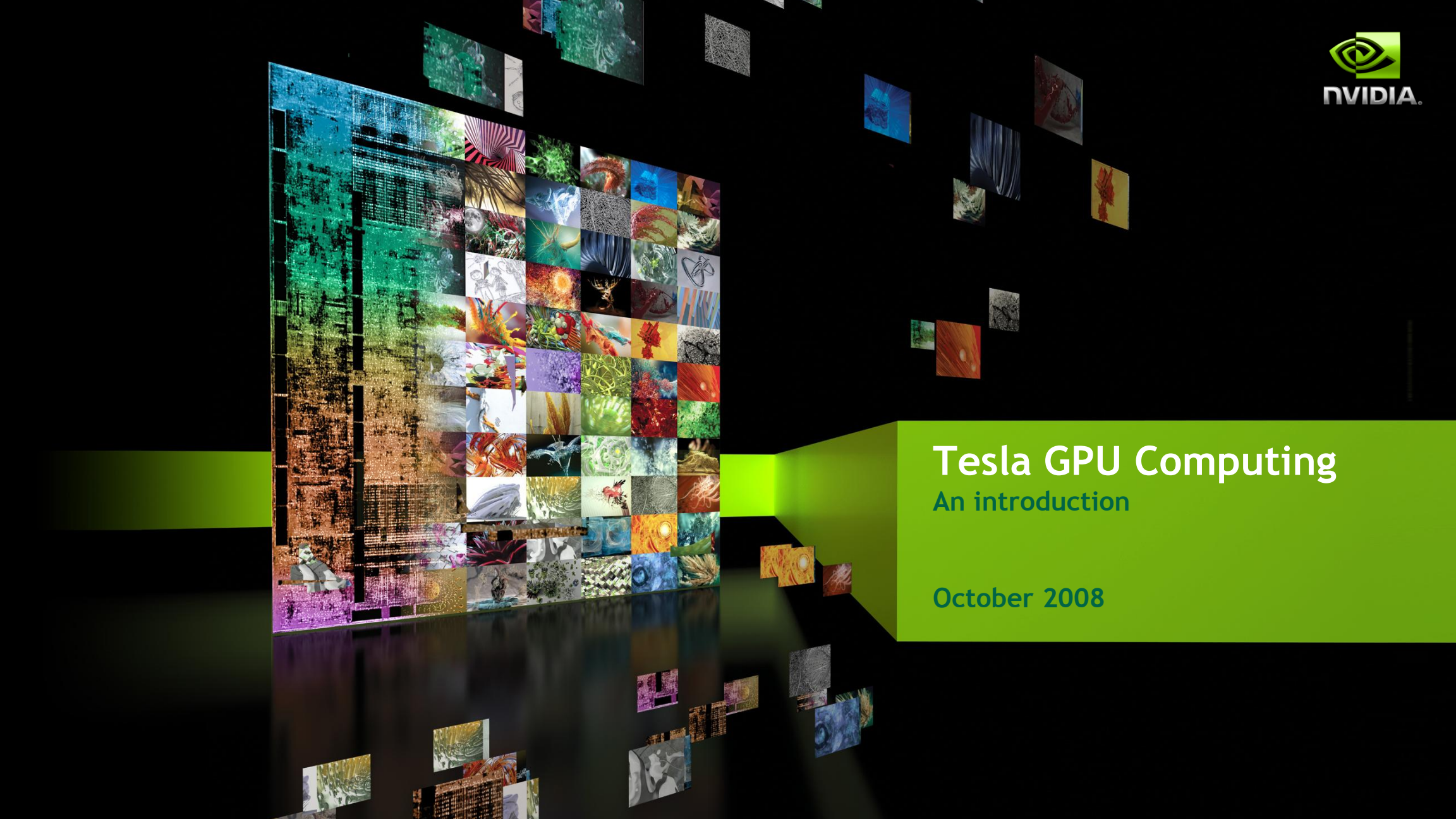
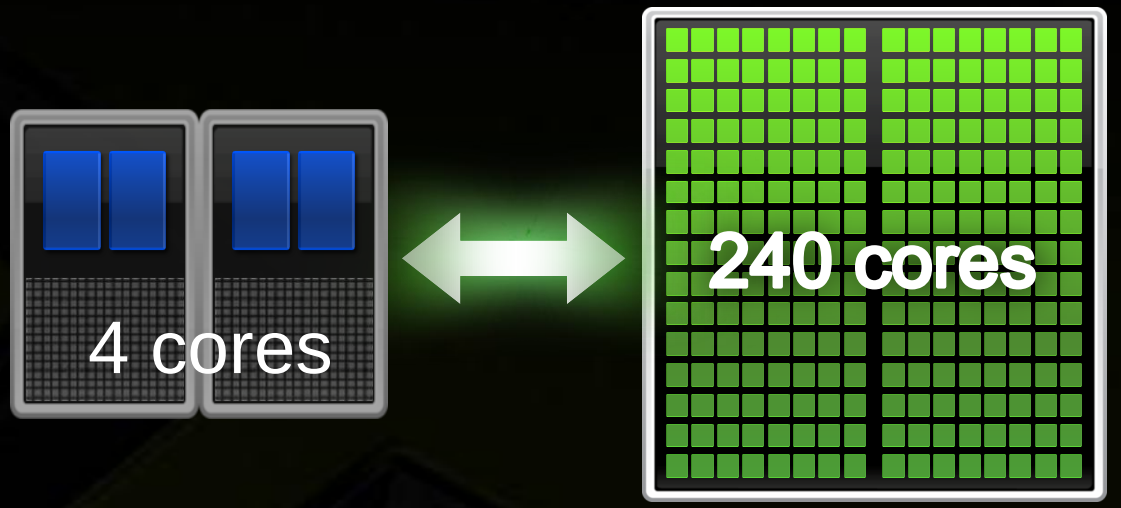


Tesla GPU Computing

An introduction

October 2008

What is GPU Computing?



Computing with CPU + GPU

Heterogeneous Computing

GPUs: Turning Point in Supercomputing

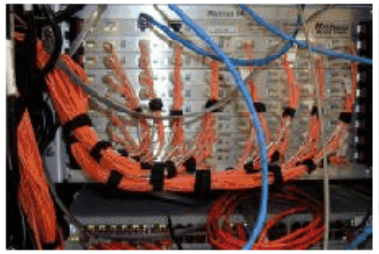
CalcUA
256 Nodes (512 cores)
$5 Million

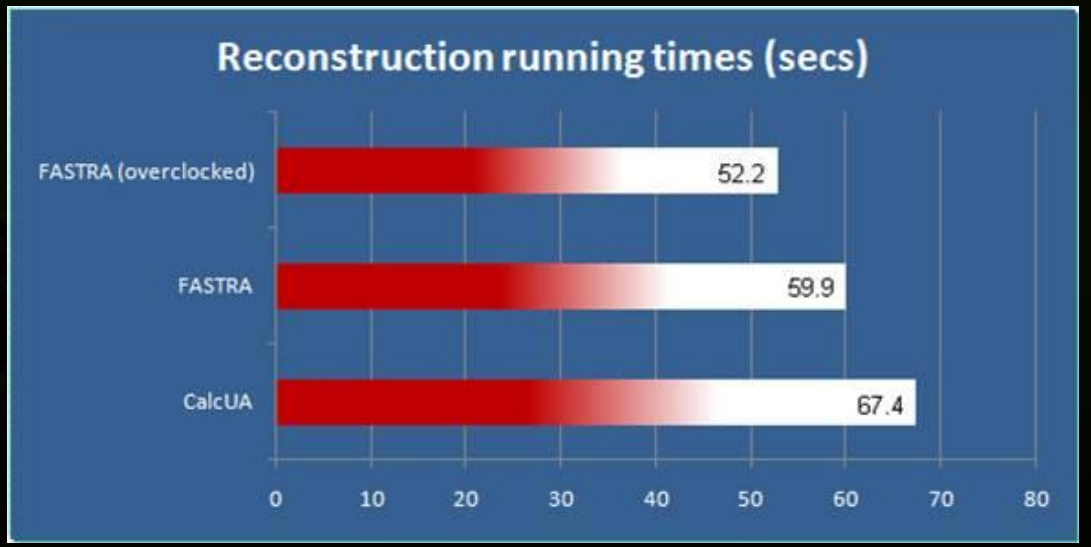
FASTRA
8 GPUs in a Desktop
$6000



CalcUA

- 256 Sun Fire V20z rekennodes (dual AMD Opteron 250, 2.4 GHz):
  - 192 rekennodes beschikken over 4 GB intern geheugen,
  - 64 rekennodes beschikken over 8 GB intern geheugen en zijn onderling verbonden via Myrinet
- 2 Sun Fire V440 servers
- 2 Sun StorEdge 3510 FC Array 6 TB centrale schijfcapaciteit

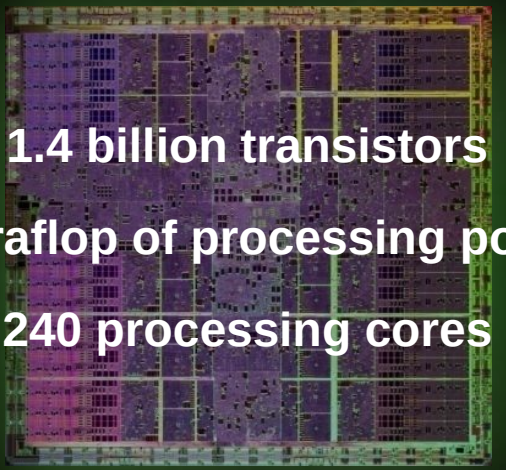



http://fastra.ua.ac.be/en/index.html

GPUs: Many Core High Performance Computing

- NVIDIA's 10-series GPU has 240 cores
- Each core has a
  - Floating point unit
  - Logic unit (add, sub, mul, madd)
  - Move, compare unit
  - Branch unit
- Cores managed by thread manager
  - Thread manager can spawn and manage 12,000+ threads
  - Zero overhead thread switching

NVIDIA 10-Series GPU



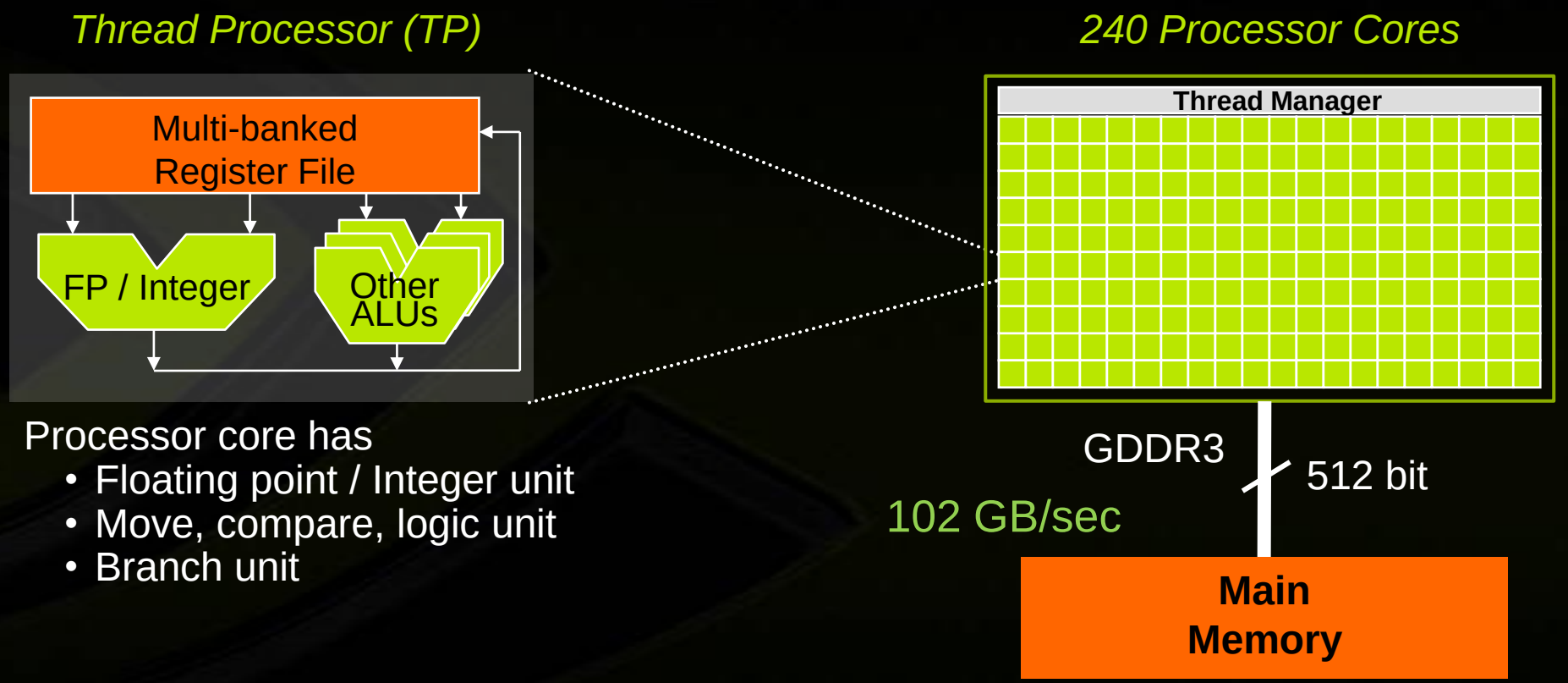
1.4 billion transistors

1 Teraflop of processing power

240 processing cores

NVIDIA's 2nd Generation CUDA Processor

Tesla 10 GPU : 240 Processor Cores

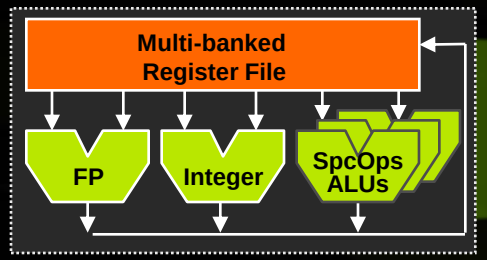


Processor core has

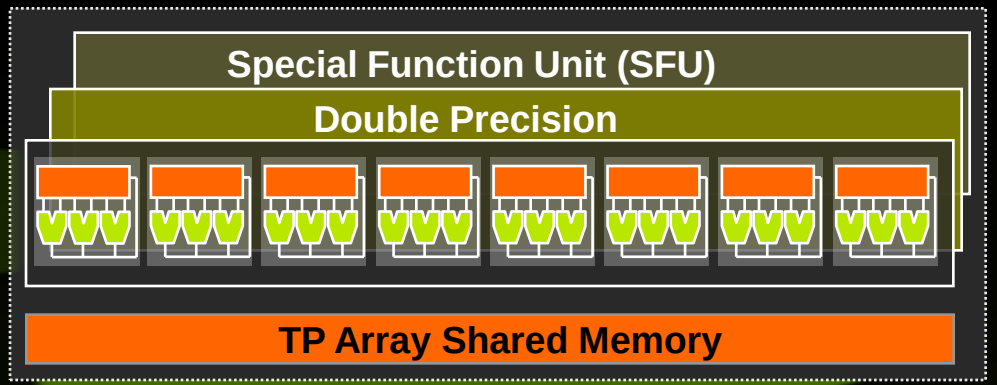
- Floating point / Integer unit
- Move, compare, logic unit
- Branch unit

Tesla T10: The Processor Inside

Thread Processor (TP)



Thread Processor Array (TPA)

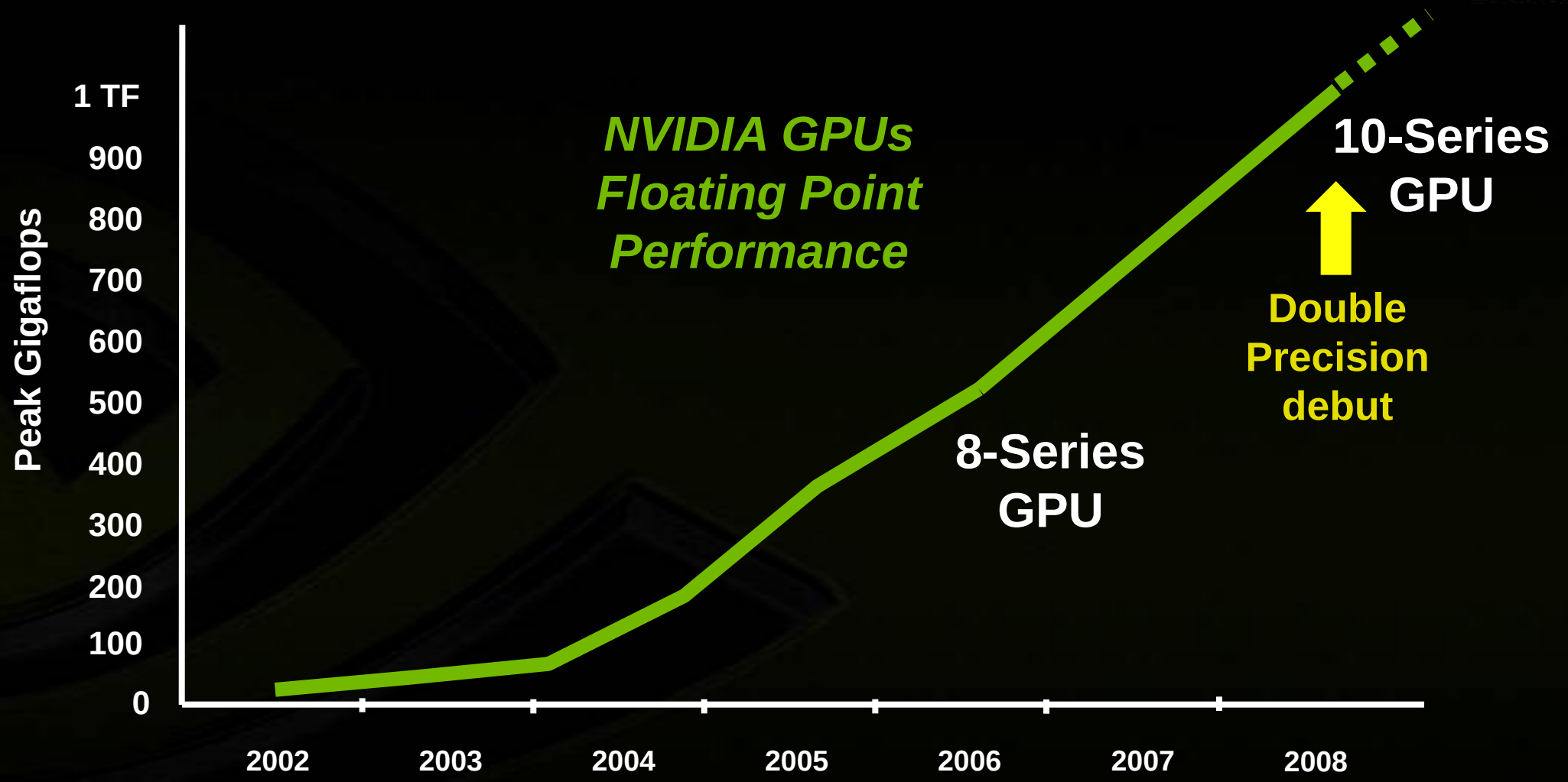


- 240 thread processors
- Full scalar processor with integer and floating point units
- IEEE 754 floating point
  - *Single and Double*



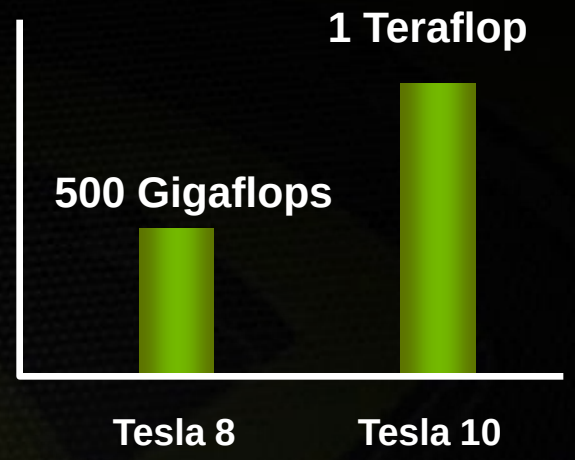
30 TPAs = 240 Processors

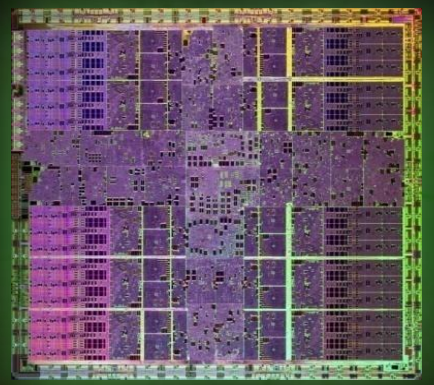
Ever Increasing Floating Point Performance



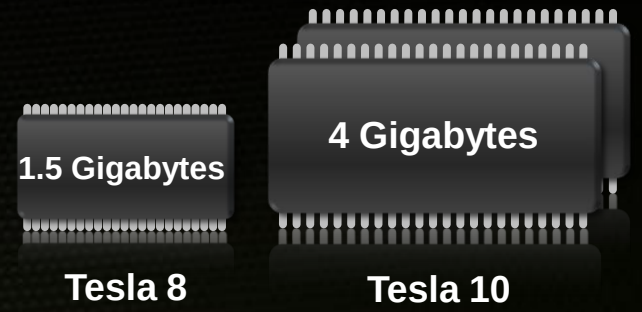
Tesla 10-Series vs 8-Series

Double the Performance

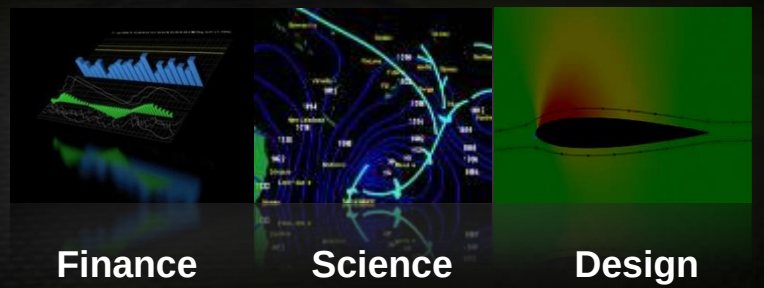




> Double the Memory



Double the Precision

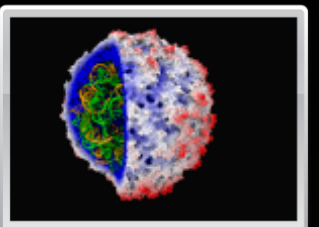


Wide Developer Acceptance and Success



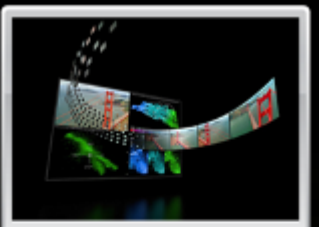
146X

Interactive visualization of volumetric white matter connectivity



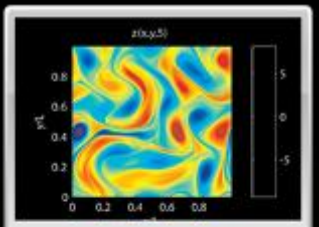
36X

Ionic placement for molecular dynamics simulation on GPU



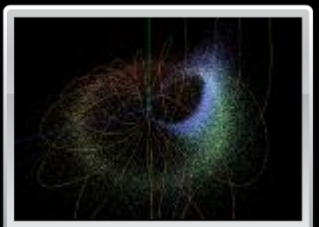
18X

Transcoding HD video stream to H.264



17X

Simulation in Matlab using .mex file CUDA function



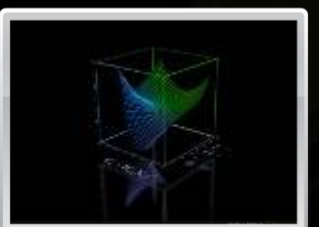
100X

Astrophysics N-body simulation



149X

Financial simulation of LIBOR model with swaptions



47X

GLAME@lab: An M-script API for linear Algebra operations on GPU



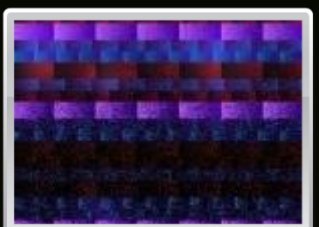
20X

Ultrasound medical imaging for cancer diagnostics



24X

Highly optimized object oriented molecular dynamics

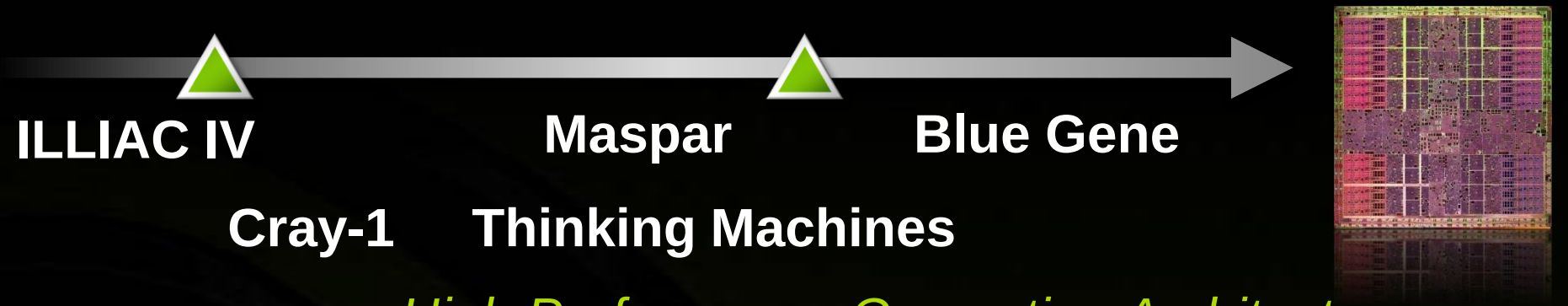


30X

Cmatch exact string matching to find similar proteins and gene sequences

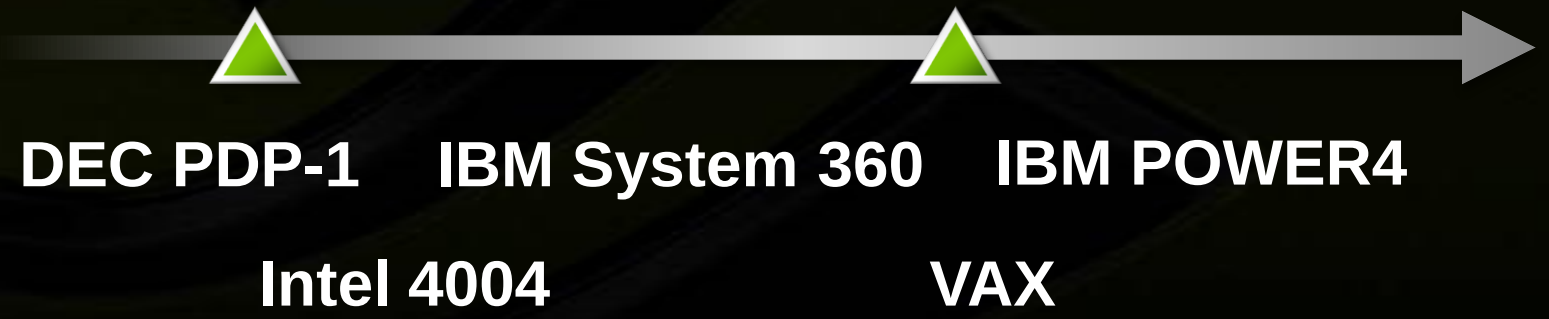
Results with 8-Series GPUs

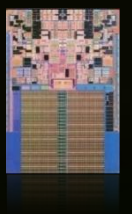
Parallel vs Sequential Architecture Evolution



Many-Core GPUs

High Performance Computing Architectures





Multi-Core x86

Data base, Operating System Sequential Architectures

Tesla S1070 1U System



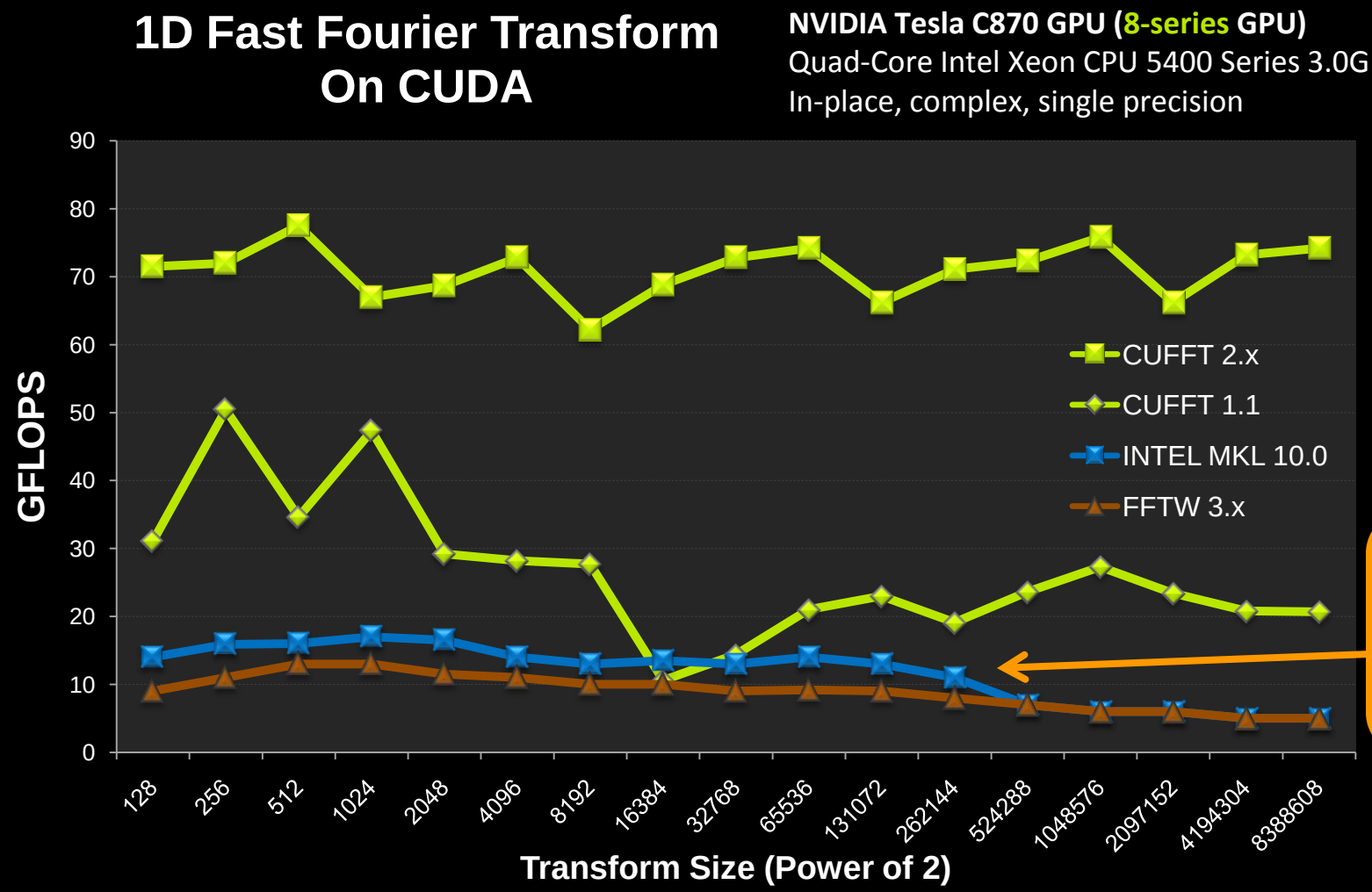
	S1070 -500	S1070 -400
Processors	4 x Tesla T10	4 x Tesla T10
Number of cores	960	960
Core Clock	1.44 GHz	1.296 GHz
Performance	4.1 TFLOPS (SP) 346 GFLOPS (DP)	3.7 TFLOPS (SP) 311 GFLOPS (DP)
Total system memory	16.0 GB (4 GB per T10)	16.0 GB (4 GB per T10)
Memory bandwidth	408 GB/sec peak (102 GB/sec per T10)	408 GB/sec peak (102 GB/sec per T10)
Memory I/O	2048-bit, 800MHz GDDR3 (512-bit per T10)	2048-bit, 800MHz GDDR3 (512-bit per T10)
Form factor	1U (EIA 19” rack)	1U (EIA 19” rack)
System I/O	2 PCIe x16 Gen2	2 PCIe x16 Gen2
Typical power	800 W	800 W

Tesla C1060 Computing Processor



Processor	1 x Tesla T10
Number of cores	240
Core Clock	1.296 GHz
Floating Point Performance	933 GFlops Single Precision
	78 GFlops Double Precision
On-board memory	4.0 GB
Memory bandwidth	102 GB/sec peak
Memory I/O	512-bit, 800MHz GDDR3
Form factor	Full ATX: 4.736” x 10.5” Dual slot wide
System I/O	PCIe x16 Gen2
Typical power	160 W

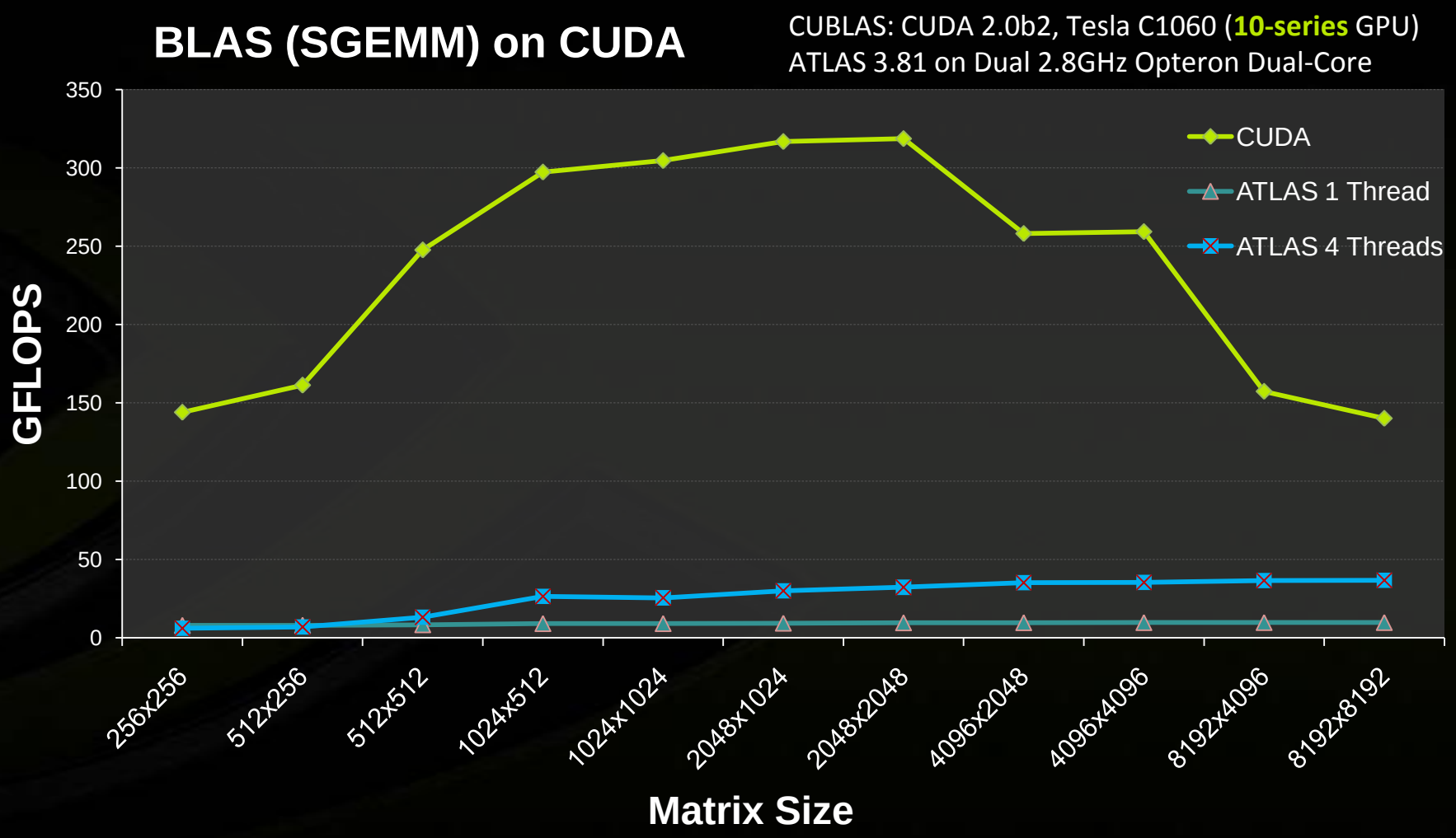
FFT Performance: CPU vs GPU (8-Series)



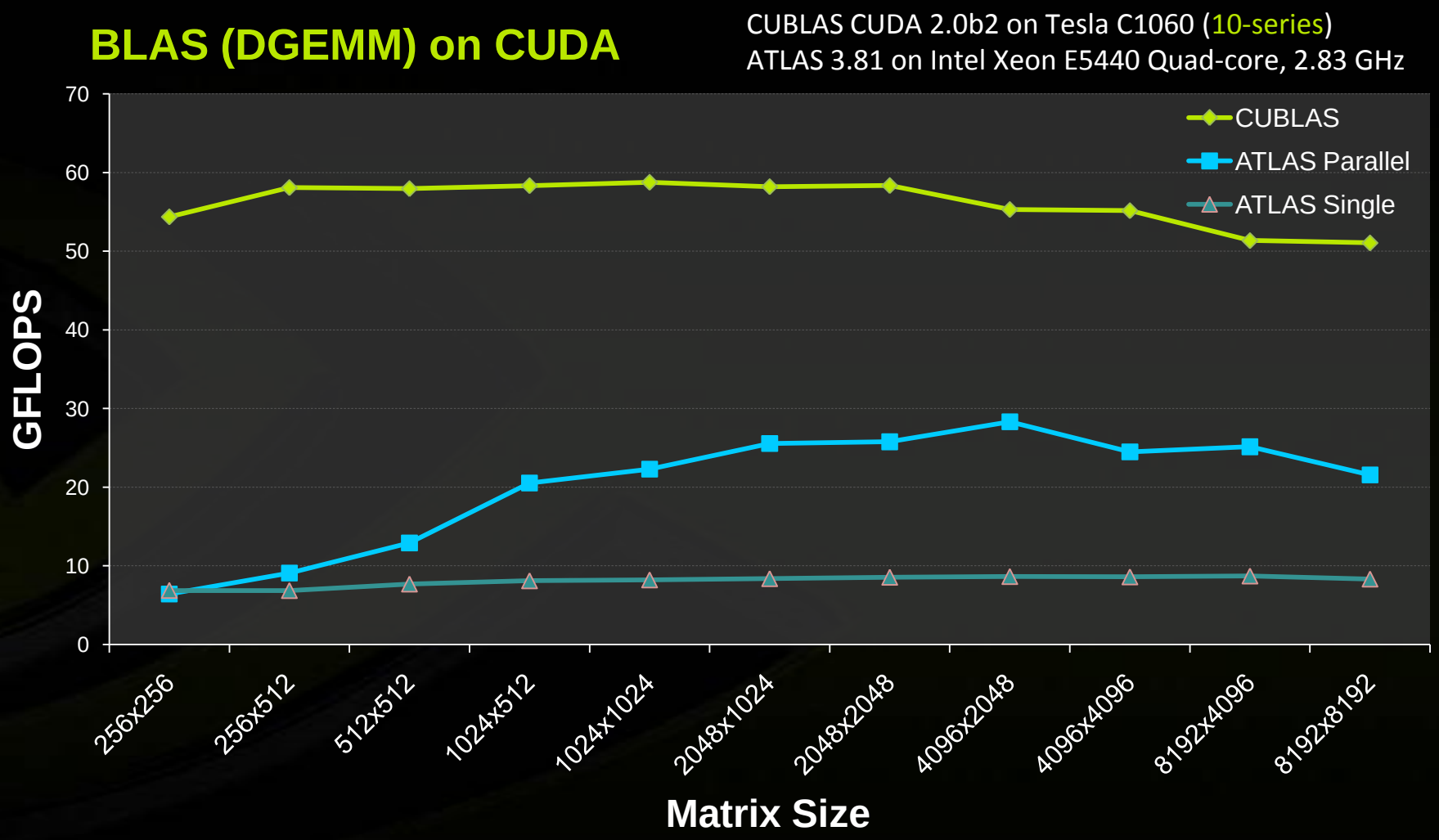
- *Intel FFT numbers calculated by repeating same FFT plan*
- *Real FFT performance is ~10 GFlops*

Source for Intel data : http://www.intel.com/cd/software/products/asmo-na/eng/266852.htm

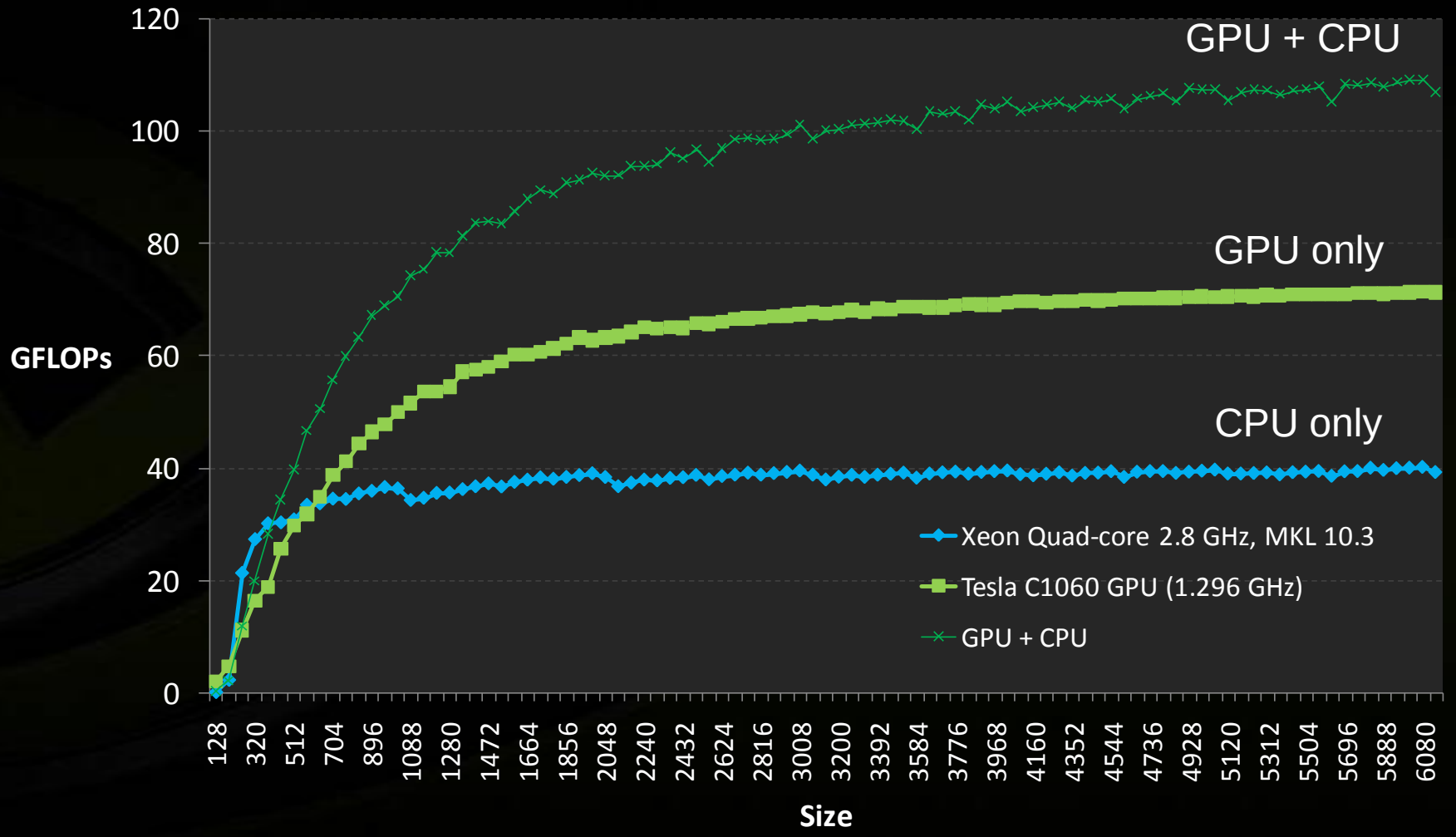
Single Precision BLAS: CPU vs GPU (10-series)



Double Precision BLAS: CPU vs GPU (10-series)



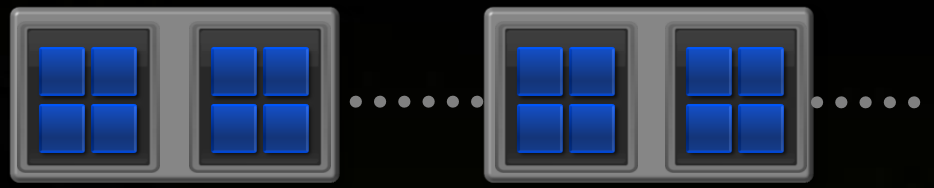
GPU + CPU DGEMM Performance



Impact on the Data Center

Data Centers: Space and Energy Limited

Traditional Data Center Cluster



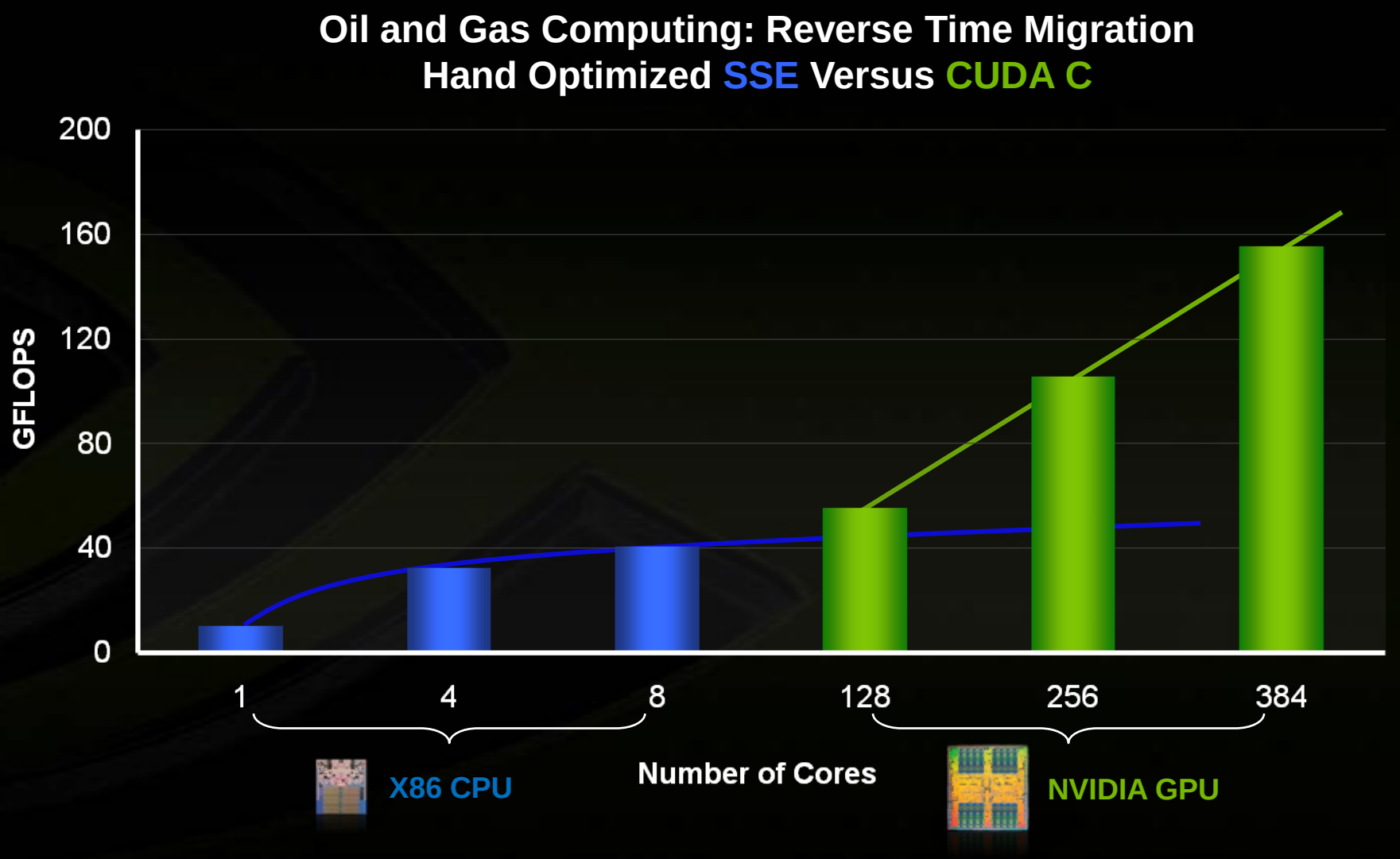
1000's of cores
1000's of servers

8 cores per server

2x Performance *requires* 2x Number of Servers



Linear Scaling with Multiple GPUs



Heterogeneous Computing Cluster

10,000's processors per cluster



1928 processors **1928 processors**

- Hess
- NCSA / UIUC
- JFCOM
- SAIC
- University of Illinois
- University of North Carolina
- Max Plank Institute
- Rice University
- University of Maryland
- GusGus
- Eotvas University
- University of Wuppertal
- Chinese Academy of Sciences
- Cell phone manufacturers



Building a 100TF datacenter

CPU 1U Server	4 CPU cores	4 GPUs: 960 cores	Tesla 1U System
	0.07 Teraflop	4 Teraflops	
	$2500	$8000	
	400 W	700 W	
	1429 CPU servers	**25 CPU servers** **25 Tesla systems**	
	$ 3.57 M	$ 0.26 M	*14x lower cost*
	571 KW	27 KW	*21x lower power*




CUDA

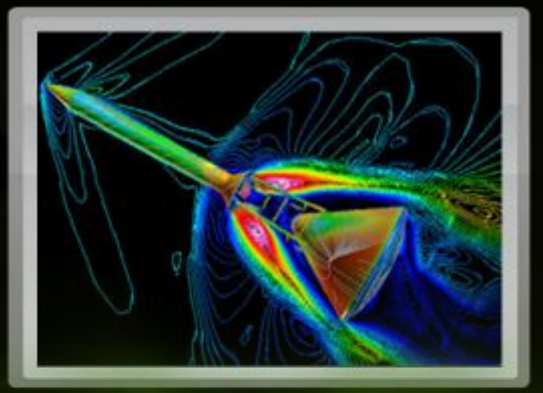
Parallel Computing on All GPUs

Over 80 Million CUDA GPUs Deployed

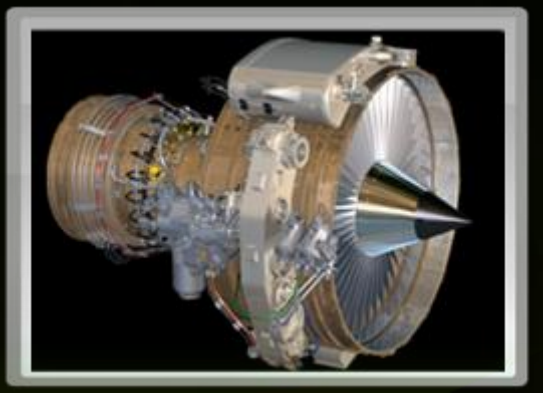
GeForce®
Entertainment



Tesla™
High-Performance Computing



Quadro®
Design & Creation





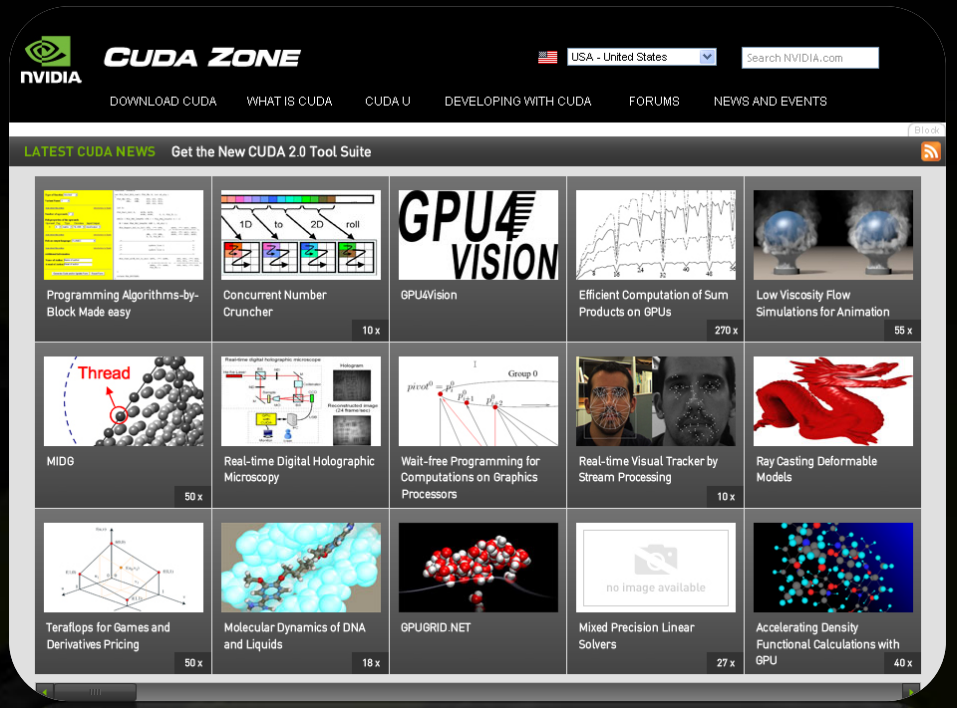
More Than 250 Customers / ISVs



Life Sciences & Medical Equipment		Productivity / Misc	Oil and Gas	EDA	Manufacturing	Finance	CAE / Numerics	Communication
Max Planck	GE Healthcare	CEA	Hess	Synopsys	Renault	Symcor	The Mathworks	Nokia
FDA	Siemens	WRF Weather Modeling	TOTAL	Nascentric	Boeing	Level 3	Wolfram	RIM
Robarts Research	Techniscan	OptiTex	CGG/Veritas	Gauda		SciComp	National Instruments	Philips
Medtronic	Boston Scientific	Tech-X	Chevron	CST		Hanweck	Access Analytics	Samsung
AGC	Eli Lilly	Elemental Technologies	Headwave	Agilent		Quant Catalyst	Tech-x	LG
Evolved machines	Silicon Informatics	Dimensional Imaging	Acceleware			RogueWave	RIKEN	Sony Ericsson
Smith-Waterman DNA sequencing	Stockholm Research	Manifold	Seismic City			BNP Paribas	SOFA	NTT DoCoMo
AutoDock	Harvard	Digisens	P-Wave Seismic Imaging					Mitsubishi
NAMD/VMD	Delaware	General Mills	Mercury Computer					Hitachi
Folding@Home	Pittsburg	Rapidmind	ffA					Radio Research Laboratory
Howard Huges Medical	ETH Zurich	MS Visual Studio						US Air Force
CRIBI Genomics	Institute Atomic Physics	Rhythm & Hues						
		xNormal						
		Elcomsoft						
		LINZIK						

CUDA

CUDA Momentum: Commercial and Research

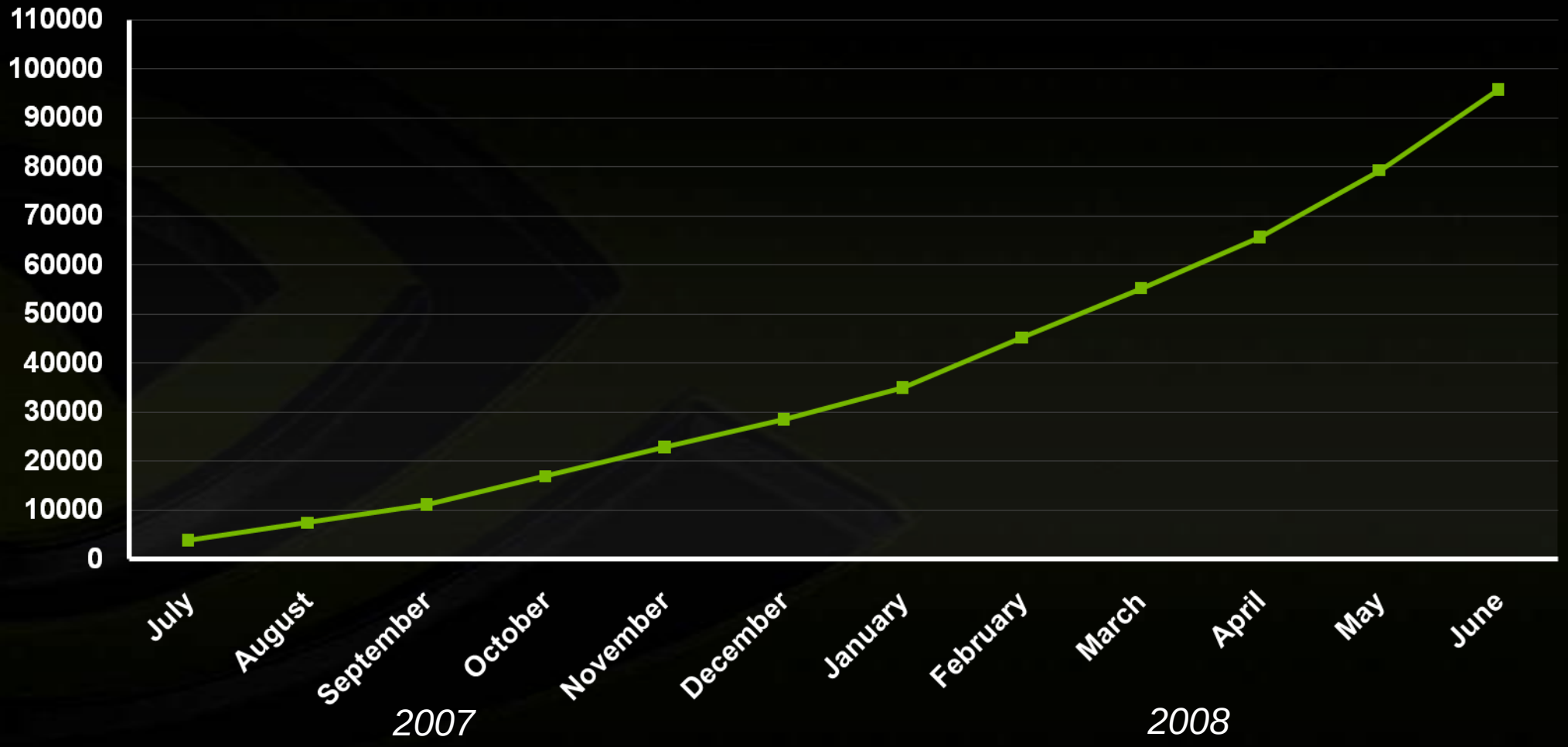


100s of Apps on CUDA Zone

www.nvidia.com/cuda

CUDA Compiler Downloads

100K CUDA compiler downloads, 80M CUDA-enabled GPUs

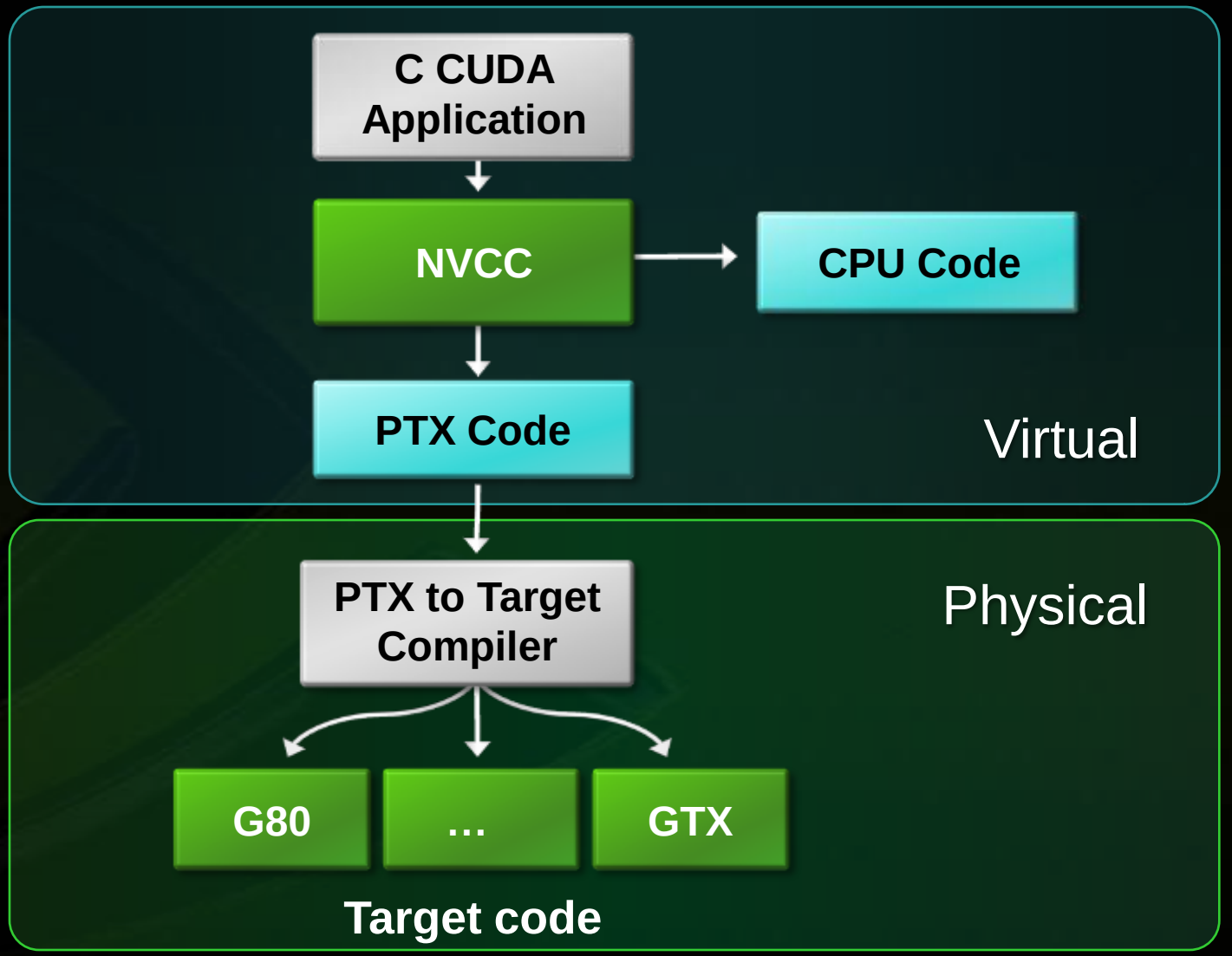


University's Teaching Parallel Programming With CUDA

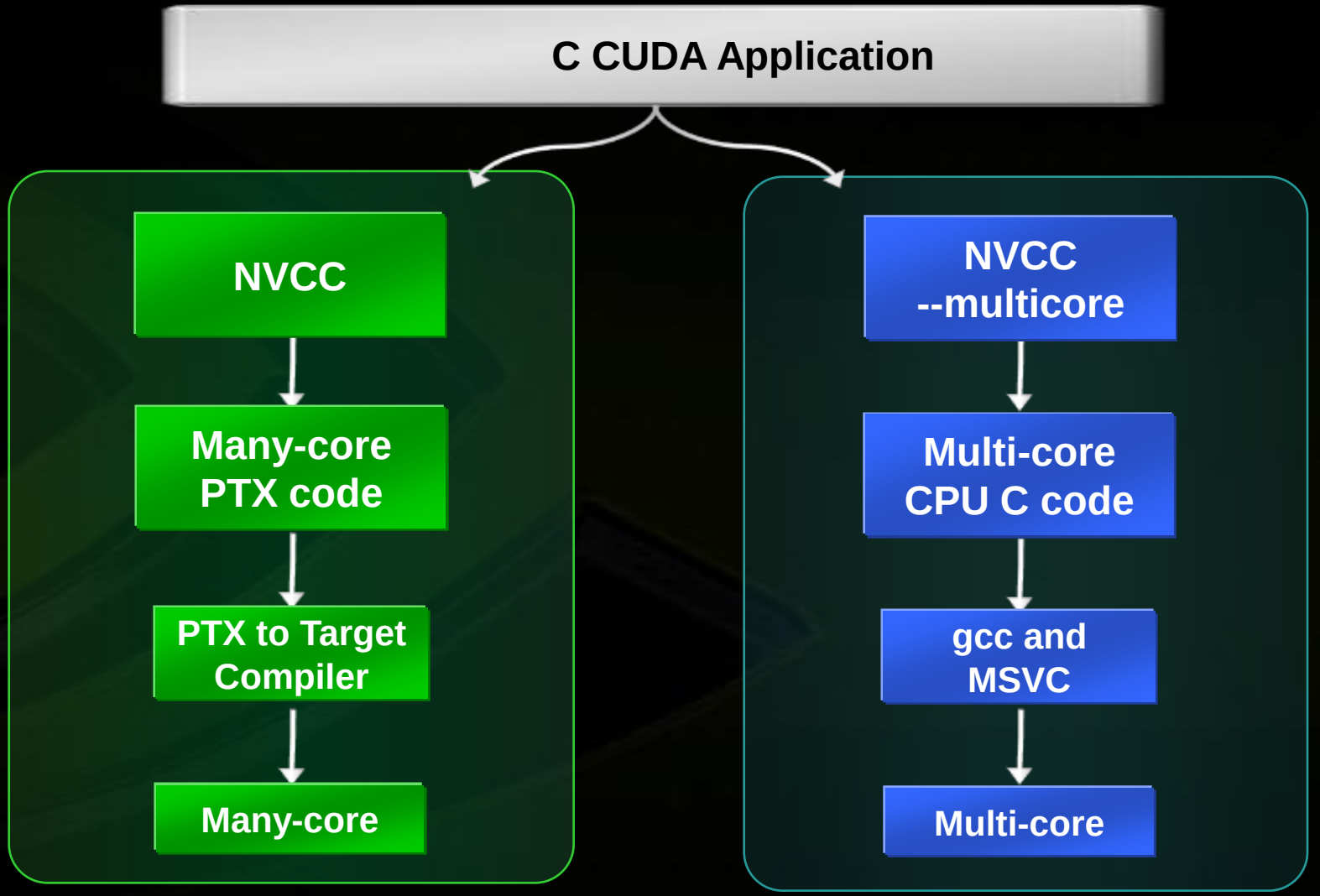


- Duke
- Erlangen
- ETH Zurich
- Georgia Tech
- Grove City College
- Harvard
- IISc Bangalore
- IIIT Hyderabad
- IIT Delhi, Bombay, Madras
- Illinois Urbana-Champaign
- INRIA
- Iowa
- ITESM
- Johns Hopkins
- Kent State
- Kyoto
- Lund
- Maryland
- McGill
- MIT
- North Carolina - Chapel Hill
- North Carolina State
- Northeastern
- Oregon State
- Pennsylvania
- Polimi
- Purdue
- Santa Clara
- Stanford Stuttgart
- Suny
- Tokyo
- TU-Vienna
- USC
- Utah
- Virginia
- Washington
- Waterloo
- Western Australia
- Williams College
- Wisconsin
- Yonsei

Compiling CUDA



CUDA 2.0: Many-core + Multi-core support



y[i] = a*x[i] + y[i] – Computed Sequentially



y[i] = a*x[i] + y[i] – Computed In Parallel



Simple “C” Description For Parallelism




```
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
        y[i] = a*x[i] + y[i];
}
// Invoke serial SAXPY kernel
saxpy_serial(n, 2.0, x, y);
```

Standard C Code

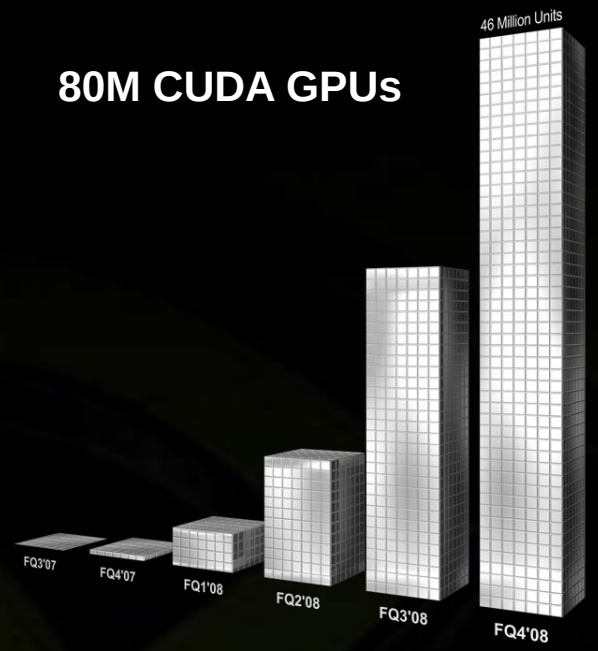
```
__global__ void saxpy_parallel(int n, float a, float *x, float *y)
{
    int i = blockIdx.x*blockDim.x + threadIdx.x;
    if (i < n)  y[i] = a*x[i] + y[i];
}
// Invoke parallel SAXPY kernel with 256 threads/block
int nblocks = (n + 255) / 256;
saxpy_parallel<<<nblocks, 256>>>(n, 2.0, x, y);
```

Parallel C Code

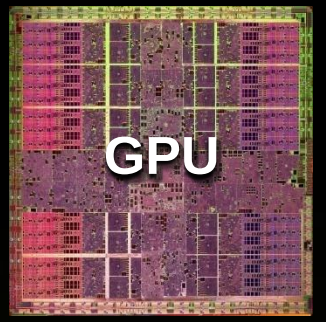
What's Next for CUDA



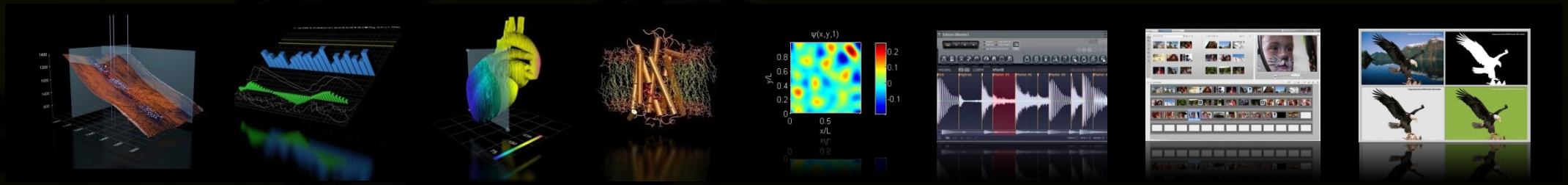
80M CUDA GPUs



CUDA



Heterogeneous Computing

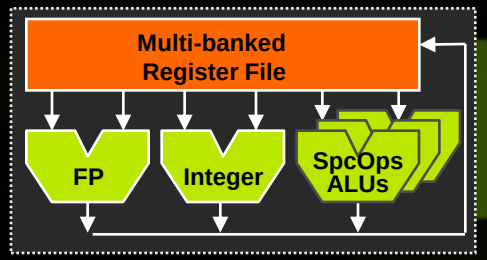


Oil & Gas | Finance | Medical | Biophysics | Numerics | Audio | Video | Imaging

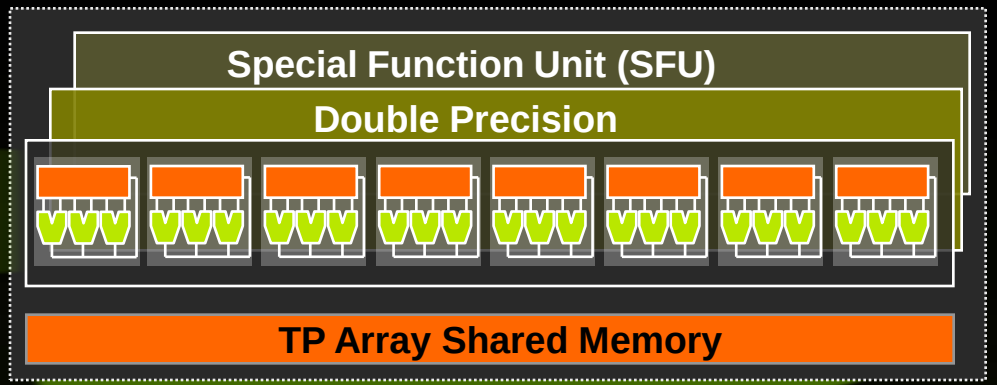
More on the GPU

Tesla T10: The Processor Inside

Thread Processor (TP)



Thread Processor Array (TPA)

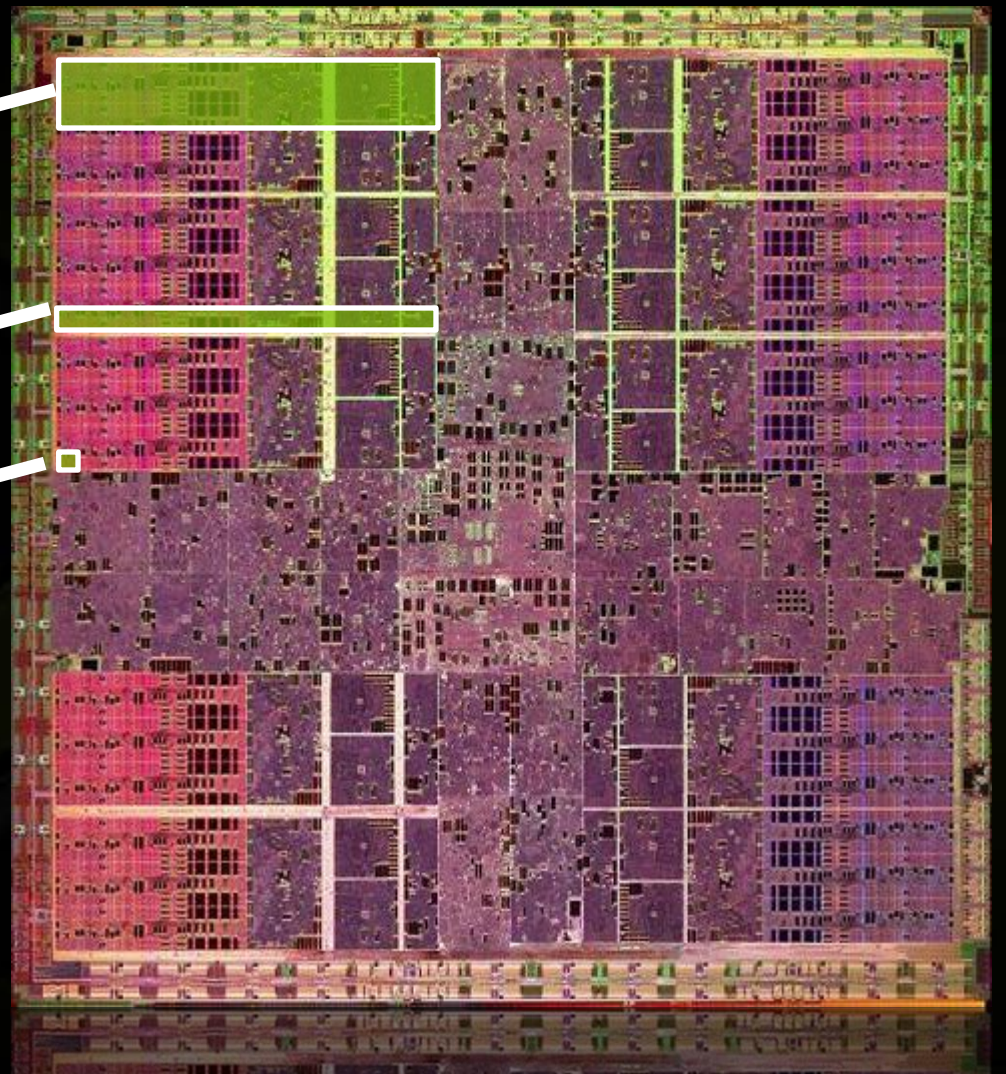


- 240 thread processors
- Full scalar processor with integer and floating point units
- IEEE 754 floating point
  - *Single and Double*



30 TPAs = 240 Processors

Tesla T10: 1.4 Billion Transistors



Die Picture of Tesla T10

	Tesla 8-series	Tesla 10-series
Number of Cores	**128**	**240**
Performance	**0.5 Teraflop**	**1 Teraflop**
On-board Memory	**1.5 GB**	**4.0 GB**
Memory interface	**384-bit GDDR3**	**512-bit GDDR3**
Memory I/O bandwidth	**77 GBytes/sec**	**102 GBytes/sec**
System interface	**PCI-E x16 Gen1**	**PCI-E x16 Gen2**

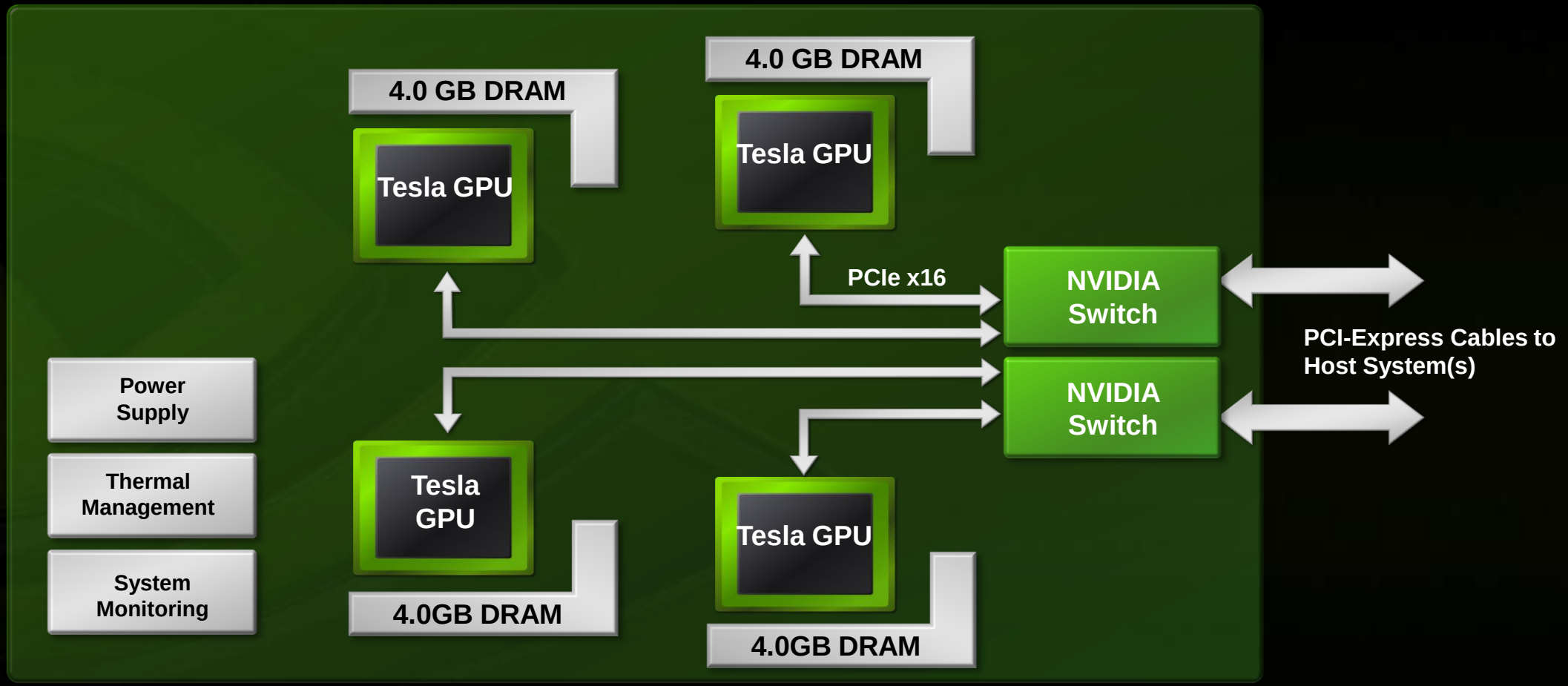
Double Precision Floating Point



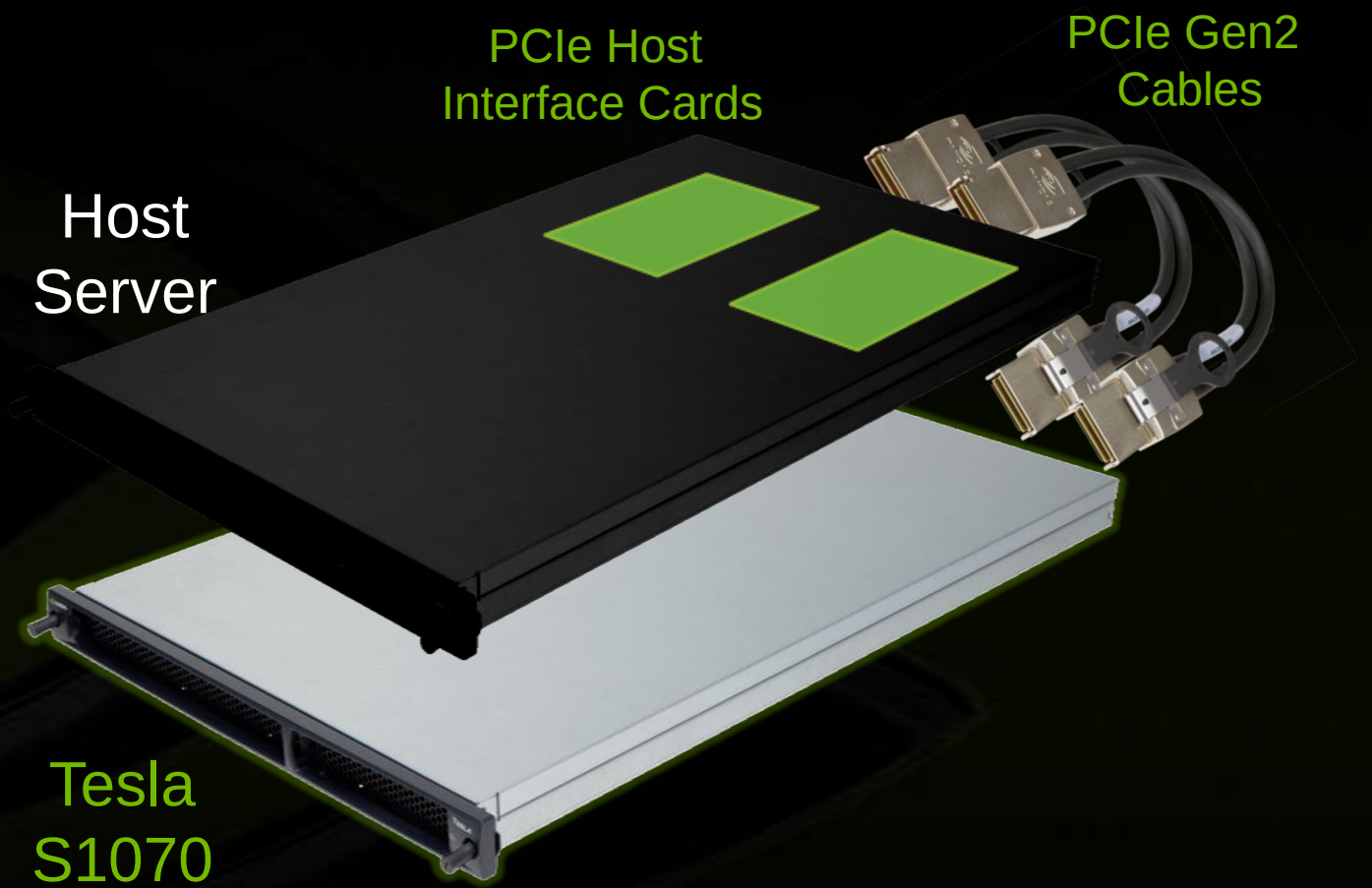
	NVIDIA Tesla T10	x86 (SSE4)	Cell SPE
Precision	**IEEE 754**	IEEE 754	IEEE 754
Rounding modes for FADD and FMUL	**All 4 IEEE, round to nearest, zero, inf, -inf**	All 4 IEEE, round to nearest, zero, inf, -inf	All 4 IEEE, round to nearest, zero, inf, -inf
Denormal handling	**Full speed**	**Supported, costs 1000's of cycles**	Supported only for results, not input operands (input denormals flushed-to-zero)
NaN support	**Yes**	Yes	Yes
Overflow and Infinity support	**Yes**	Yes	Yes
Flags	**No**	Yes	Yes
FMA	**Yes**	**No**	Yes
Square root	**Software with low-latency FMA-based convergence**	Hardware	**Software only**
Division	**Software with low-latency FMA-based convergence**	Hardware	**Software only**
Reciprocal estimate accuracy	**24 bit**	**12 bit**	**12 bit + step**
Reciprocal sqrt estimate accuracy	**23 bit**	**12 bit**	**12 bit + step**
log2(x) and 2^x estimates accuracy	**23 bit**	**No**	**No**

Hooking up S1070 to Host Server

Tesla S1070 System Architecture



Connecting Tesla S1070 to Host Servers



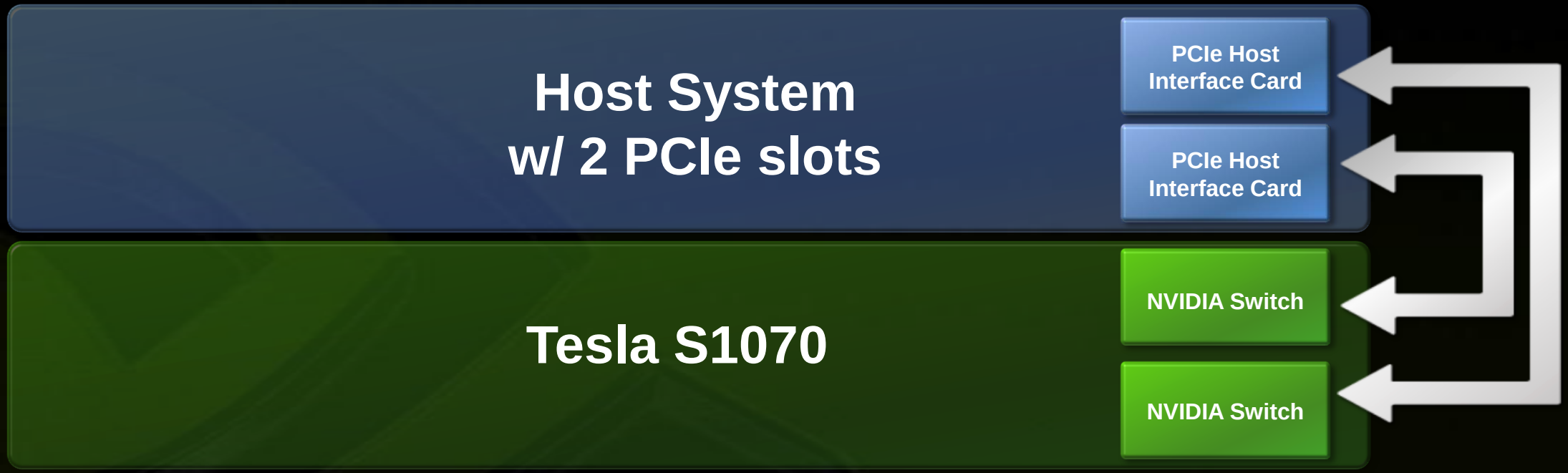


PCIe Gen2 Cable
(0.5m length)



PCIe Gen2 Host
Interface Card

Tesla S1070 connection to a single Host



Tesla S1070 connection to dual Host





For more information

http://www.nvidia.com/Tesla