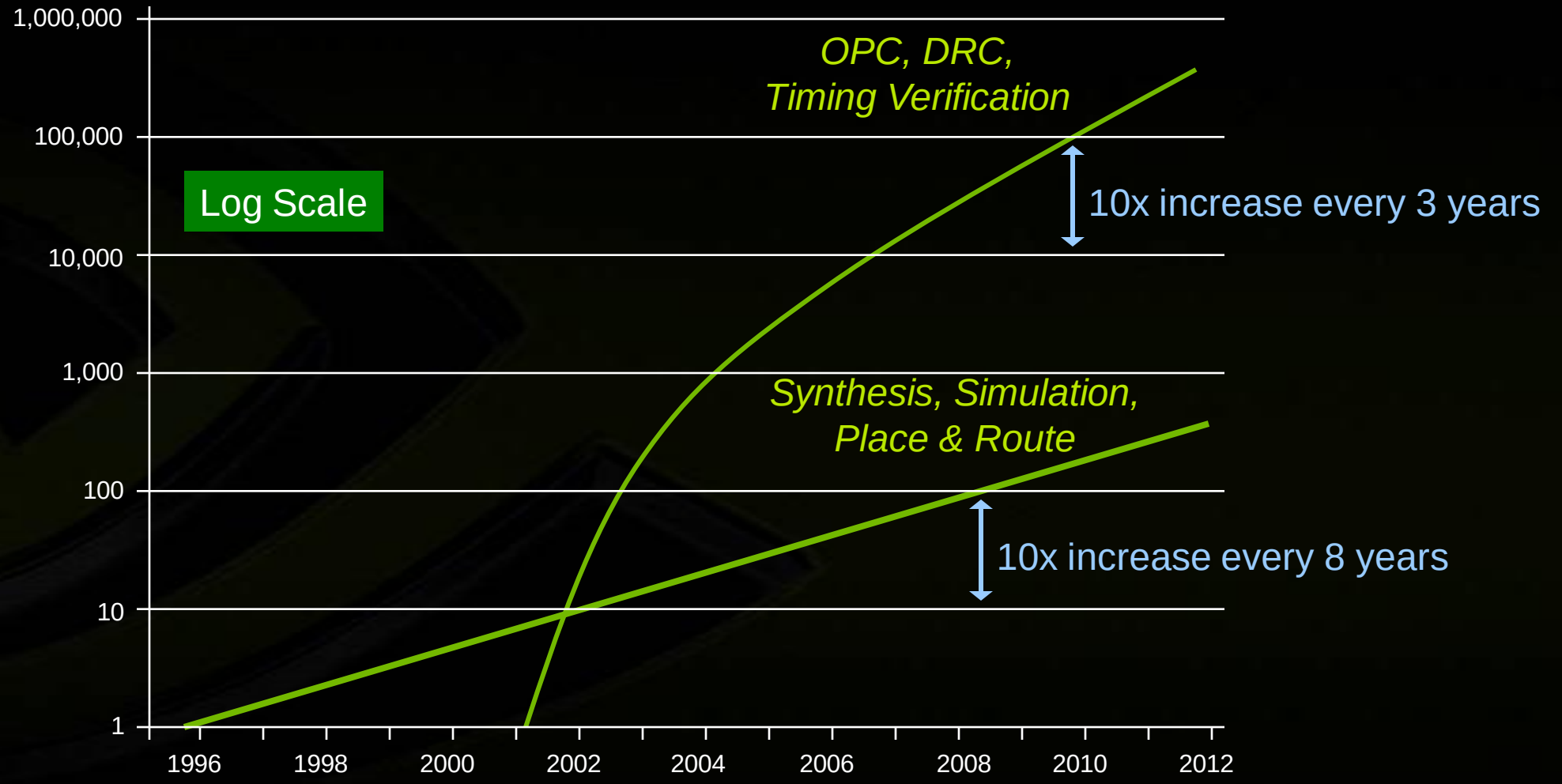
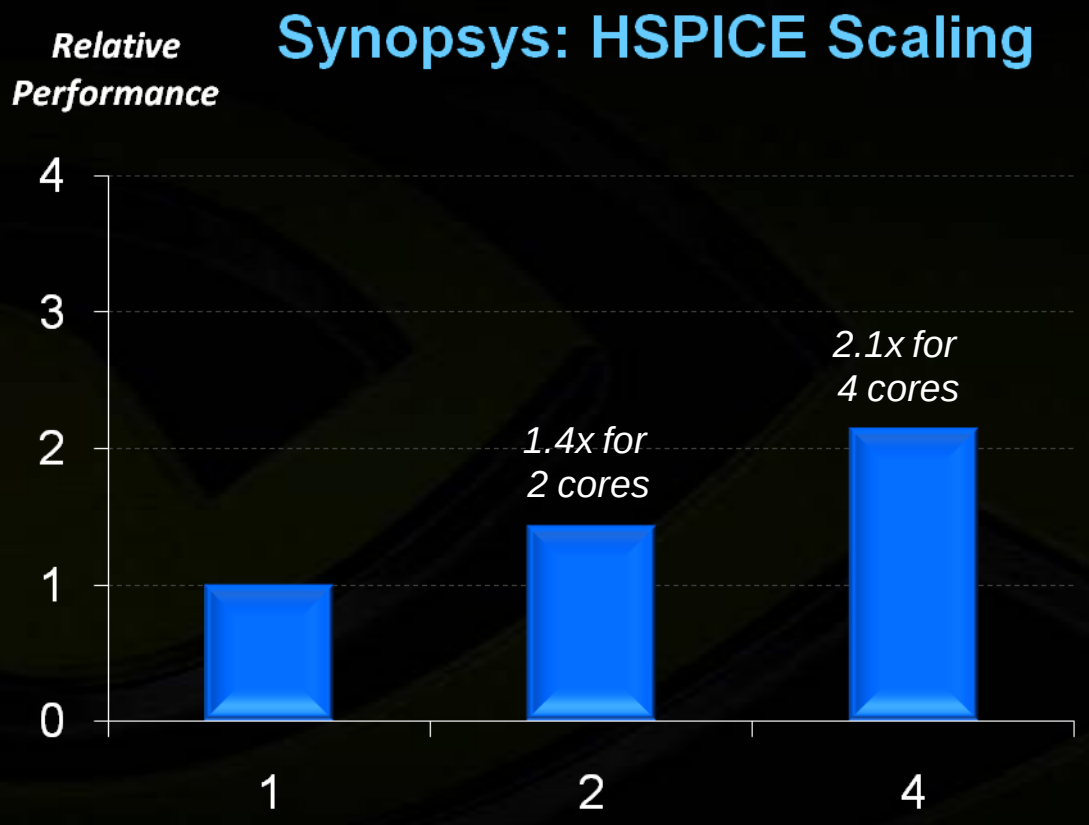

NVIDIA
“Threads are dead! Long Live threads”
Many Core, Massively Threaded Processing:
A Revolution in High Performance Computing

EDA Requirements Exceed Moore's Law



EDA: Poor Scaling with CPU Cores



- EDA applications are not scaling with multi-core CPU
- Memory bandwidth limited on CPUs
- Takes hours to days to work on large designs

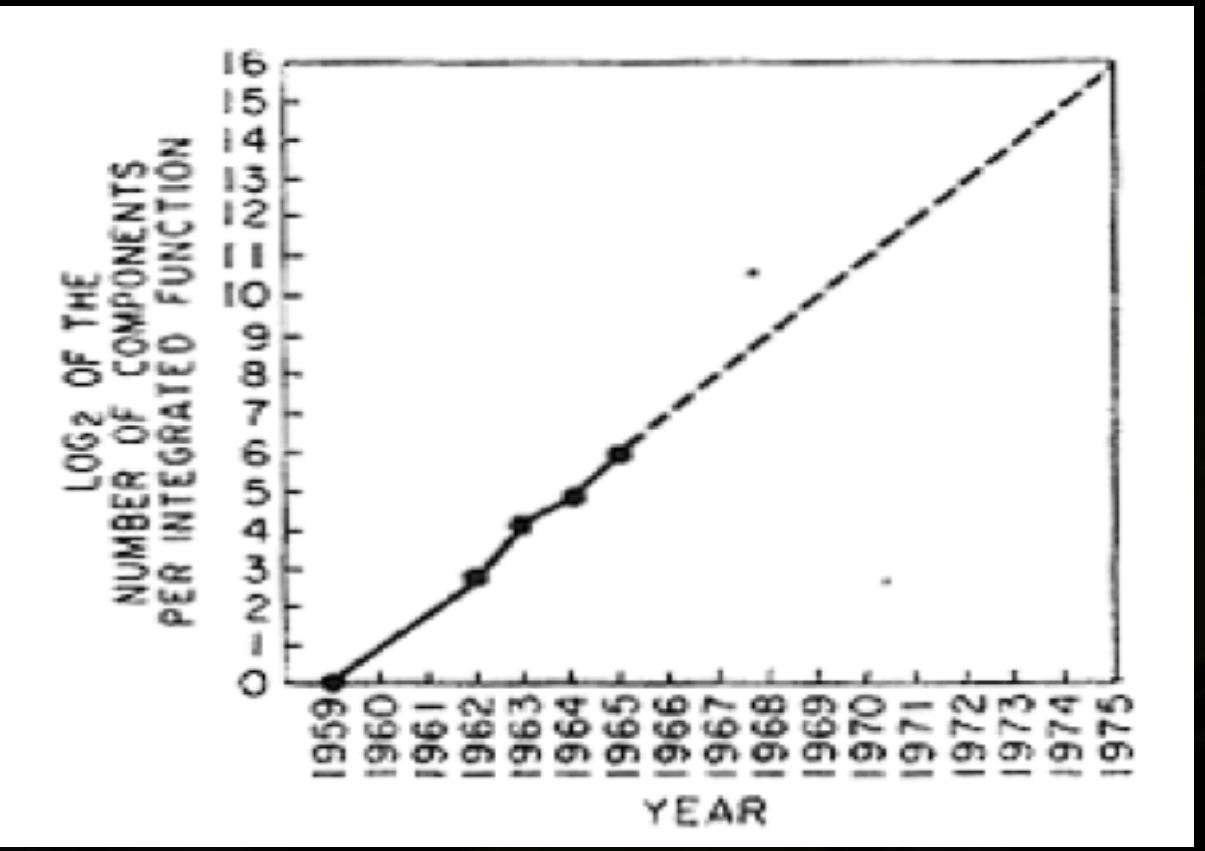
Source: EE Times; Aug 1, 2008

What's a Programmer to do ?

The “New” Reality

- Computers no longer get faster, just wider
- You *must* re-think your algorithms to be parallel !
- Power sets bounds on what’s possible
- Performance = parallelism
- Efficiency = locality

Moore's Law

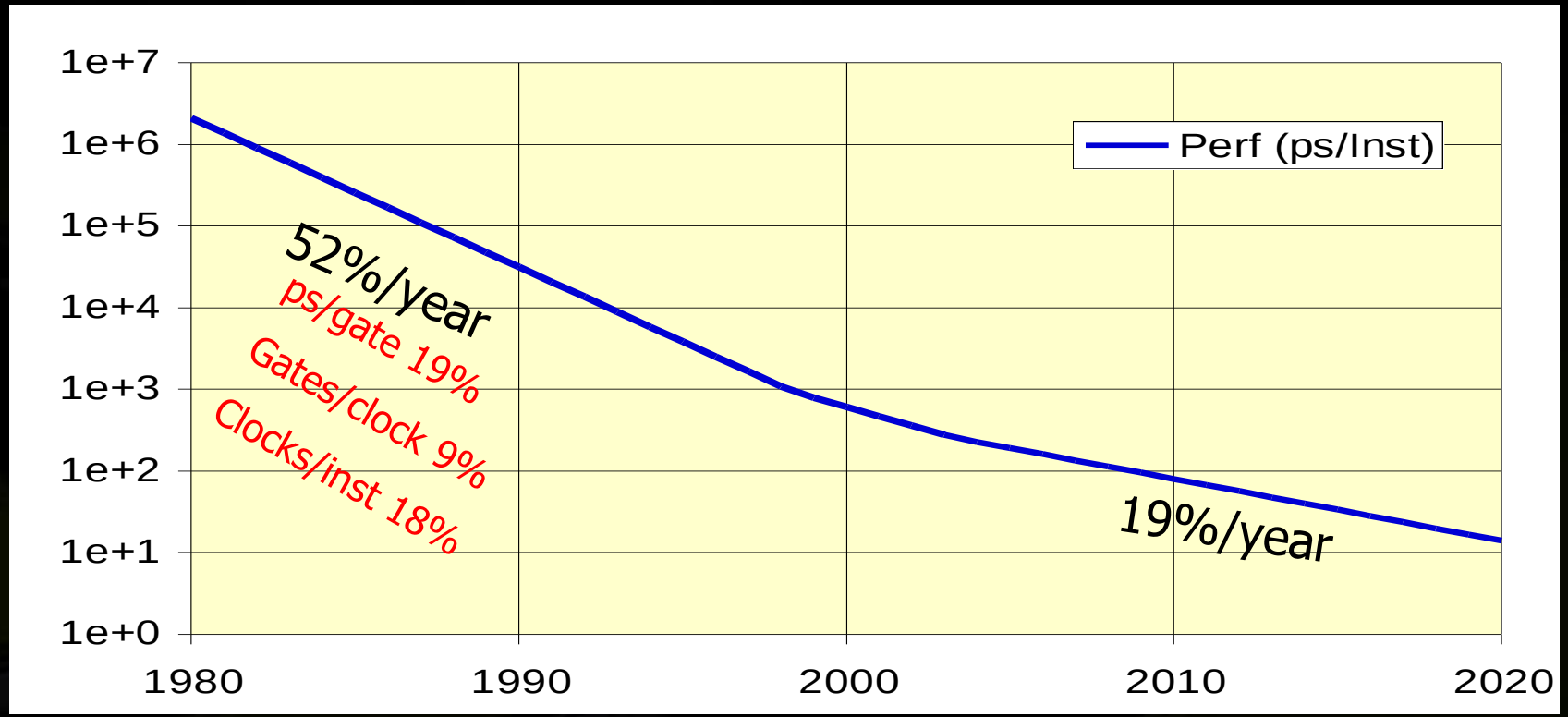


- In 1965 Gordon Moore predicted the *number of transistors* on an integrated circuit would double every year.
  - Later revised to 18 months
- Also predicted L^3 energy scaling
- No prediction of processor performance
- Advances in architecture turn device scaling into *performance* scaling
- Applications turn performance into *value*

Moore, Electronics 38(8) April 19, 1965

© NVIDIA Corporation 2009

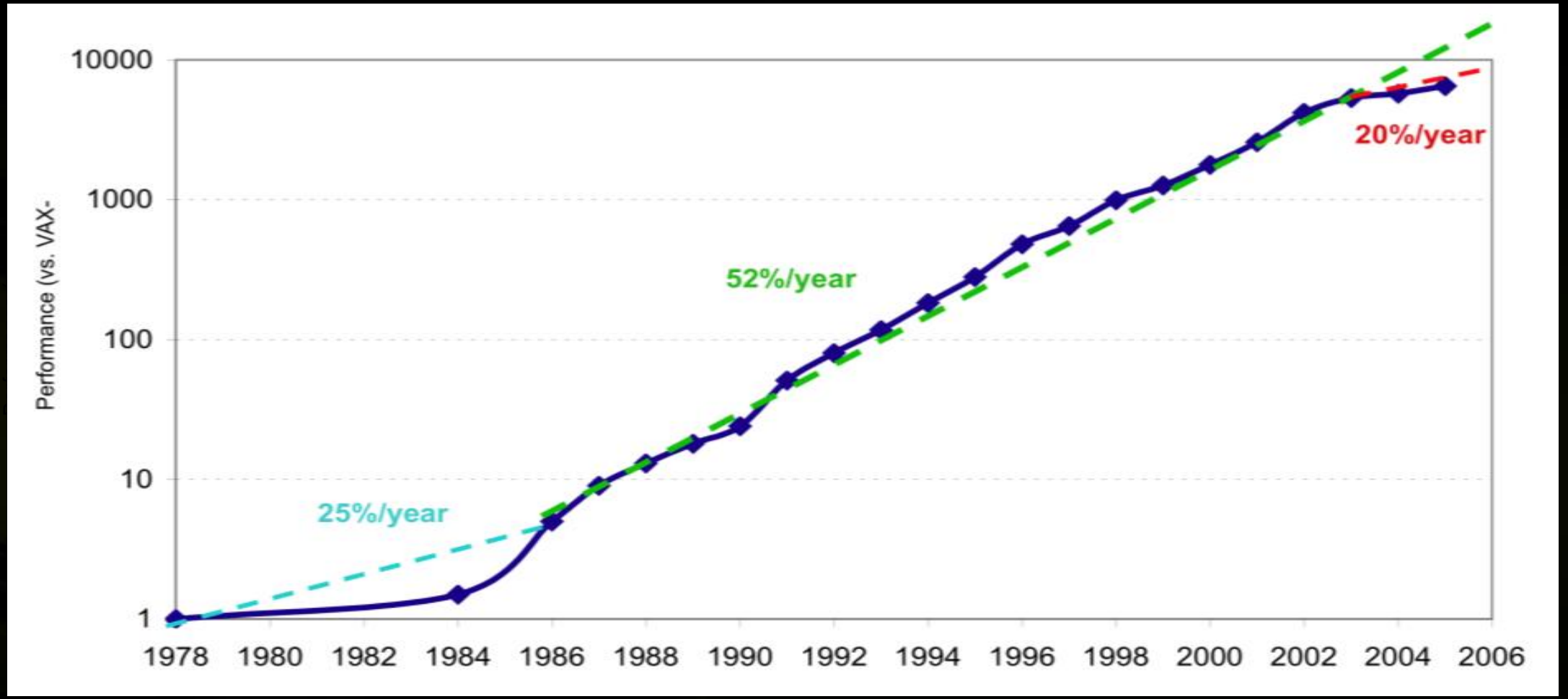
Discontinuity 1
The End of ILP Scaling



Dally et al. The Last Classical Computer, ISAT Study, 2001

© NVIDIA Corporation 2009

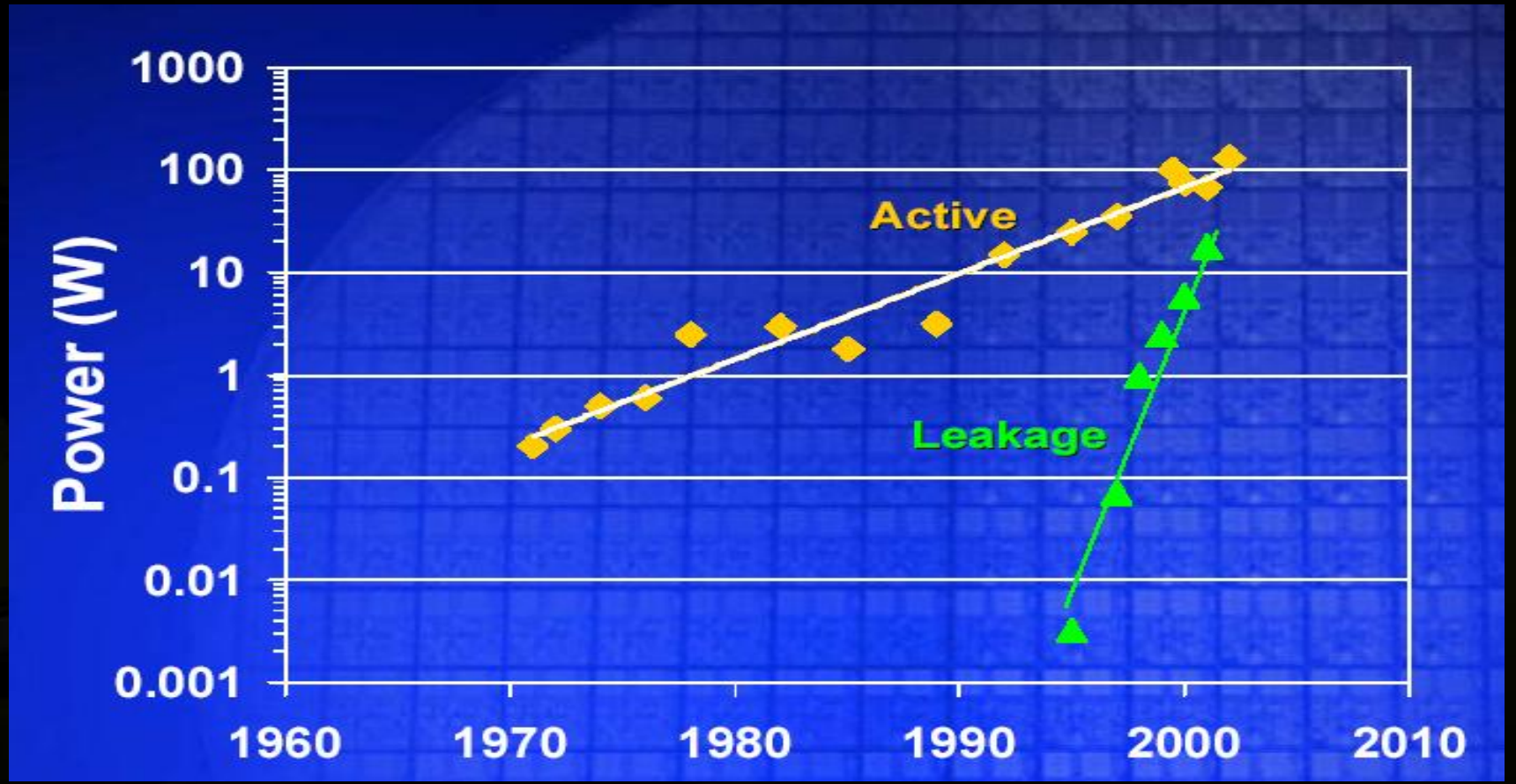
Single-Thread Processor Performance



© NVIDIA Corporation 2009

Source: Hennessy & Patterson, CAAQA, 4th Edition

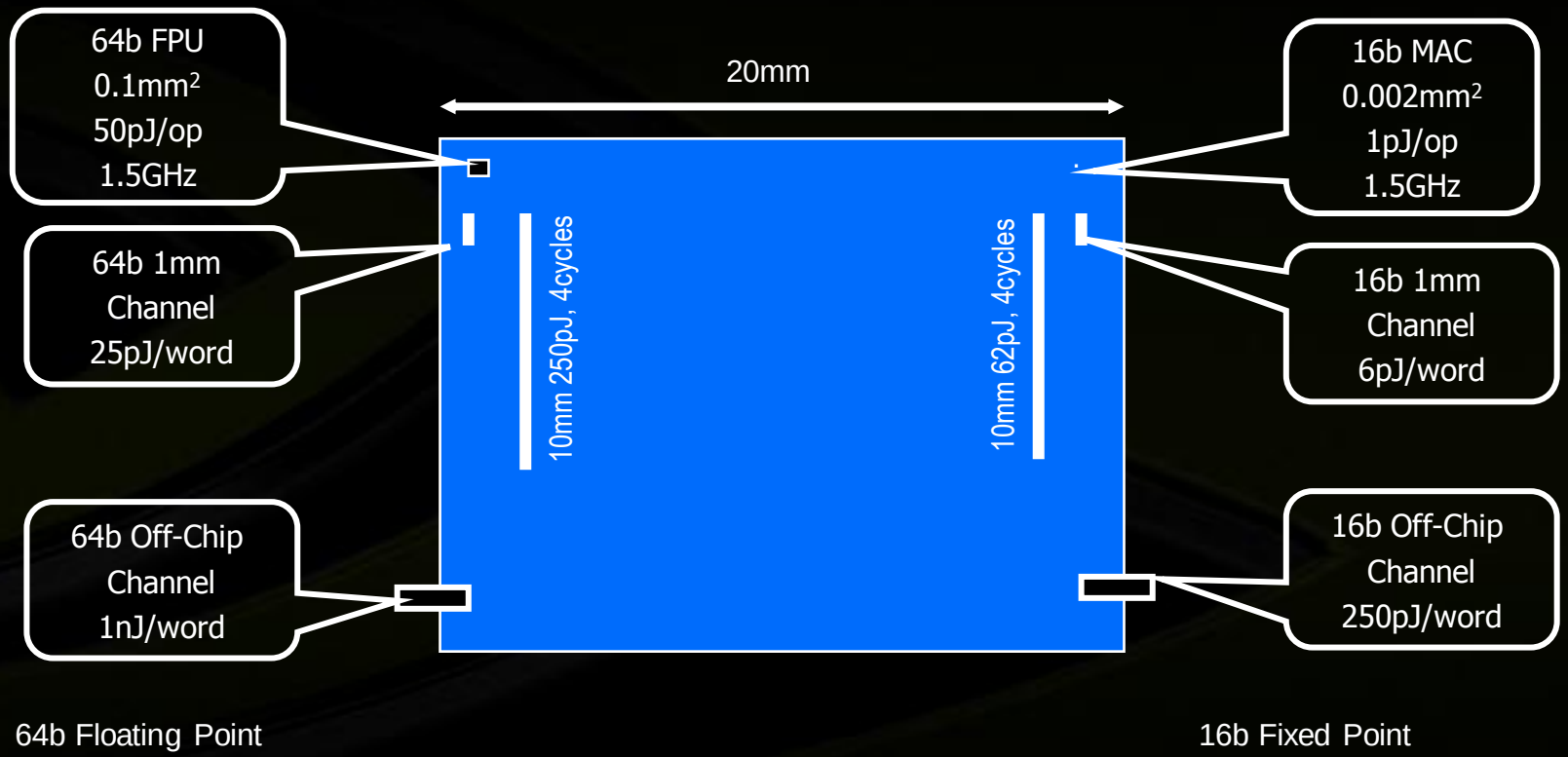
Discontinuity 2
The End of L^3 Power Scaling



Gordon Moore, ISSCC 2003

© NVIDIA Corporation 2009

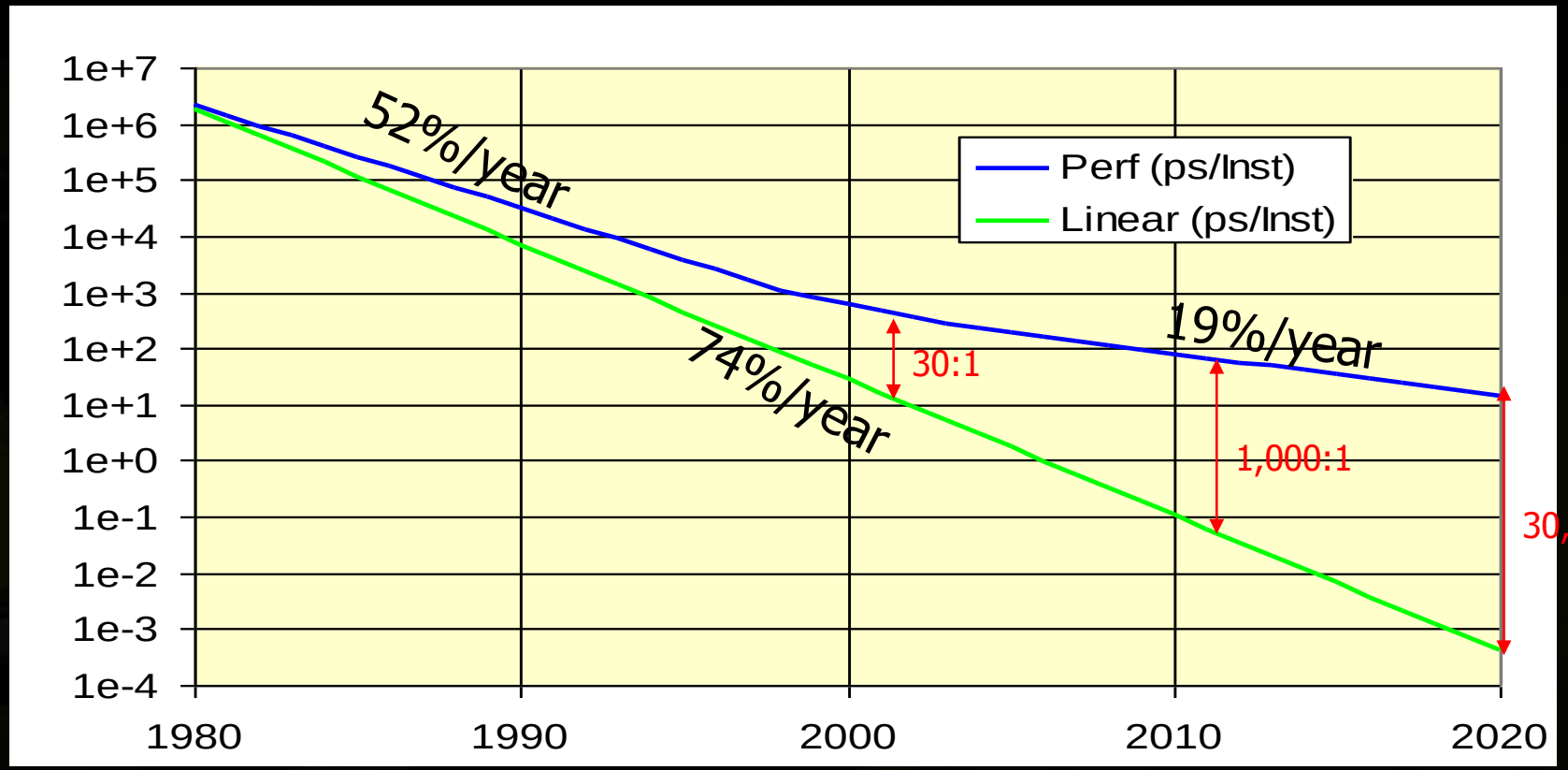
Performance = Parallelism
Efficiency = Locality



© NVIDIA Corporation 2009

* 40nm process assumed

An optimist's view – finding opportunity in adversity

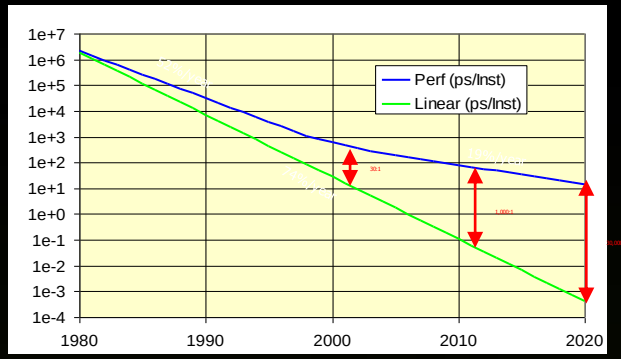


Dally et al. The Last Classical Computer, ISAT Study, 2001

© NVIDIA Corporation 2009

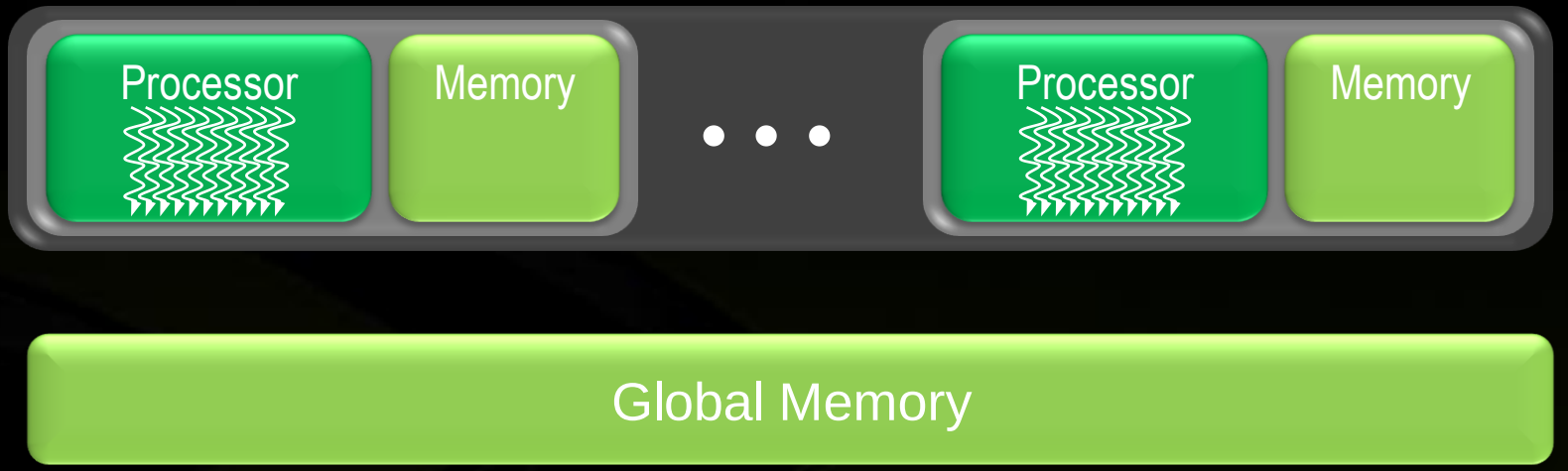
Stream Processing –
Efficient Throughput Computing

- Technology, applications, and discontinuities show
  - Performance = parallelism
  - Efficiency = locality
  - Latency-optimized processors won't improve much anymore
- Stream processing
  - Many ALUs exploit parallelism
  - Rich, exposed storage hierarchy enables locality
  - Simple control and synchronization reduces overhead
- Stream programming - explicit movement, bulk ops
  - Exposes parallelism (bulk operations) and locality
  - Enables strategic optimization
  - Predictability enables static optimization
- Result: performance and efficiency
  - TFLOPs on a chip
  - 20-30x efficiency of conventional processors.
  - Performance portability



© NVIDIA Corporation 2009

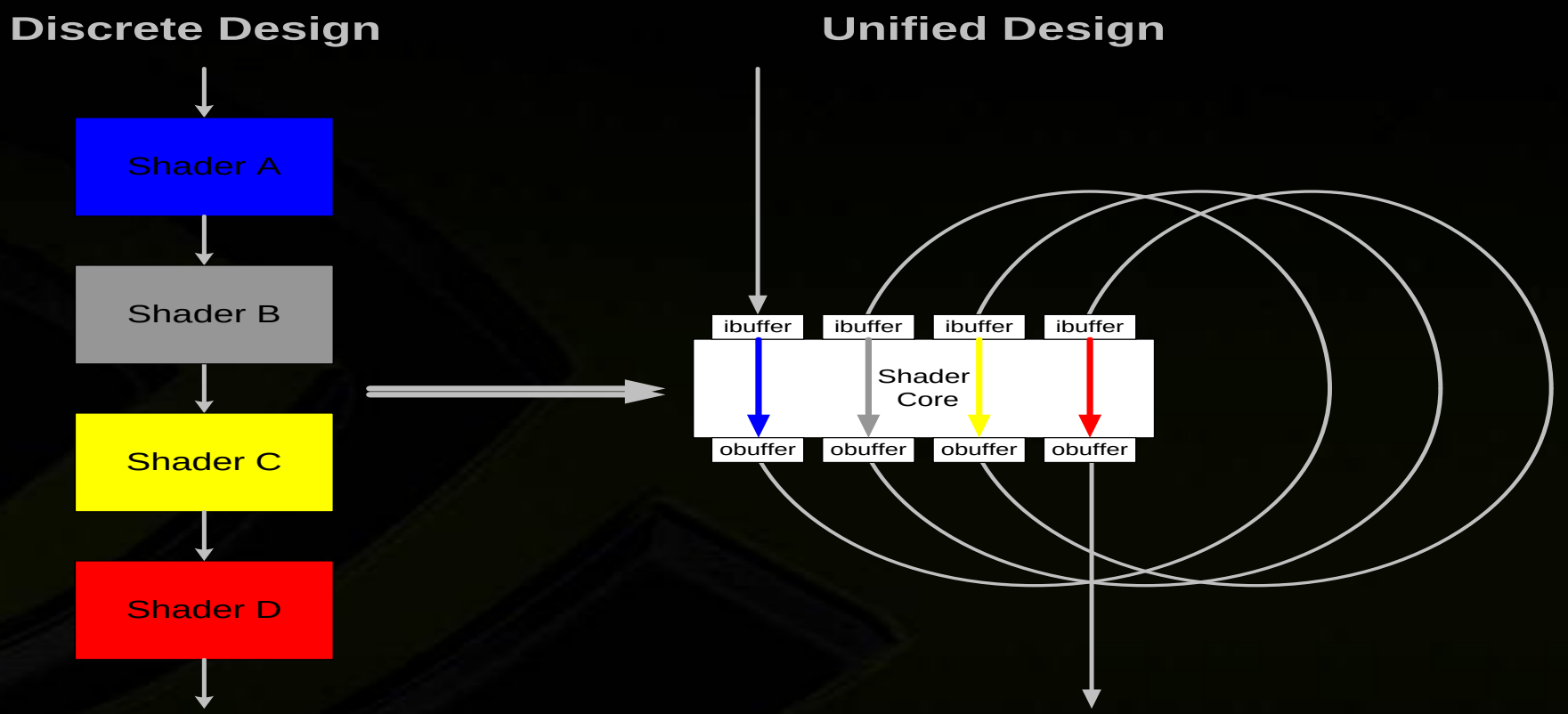
Generic Stream Processing Architecture



- Many processors each supporting many hardware threads
- On-chip memory near processors (cache, RAM, or both)
- Shared global memory space (external DRAM)

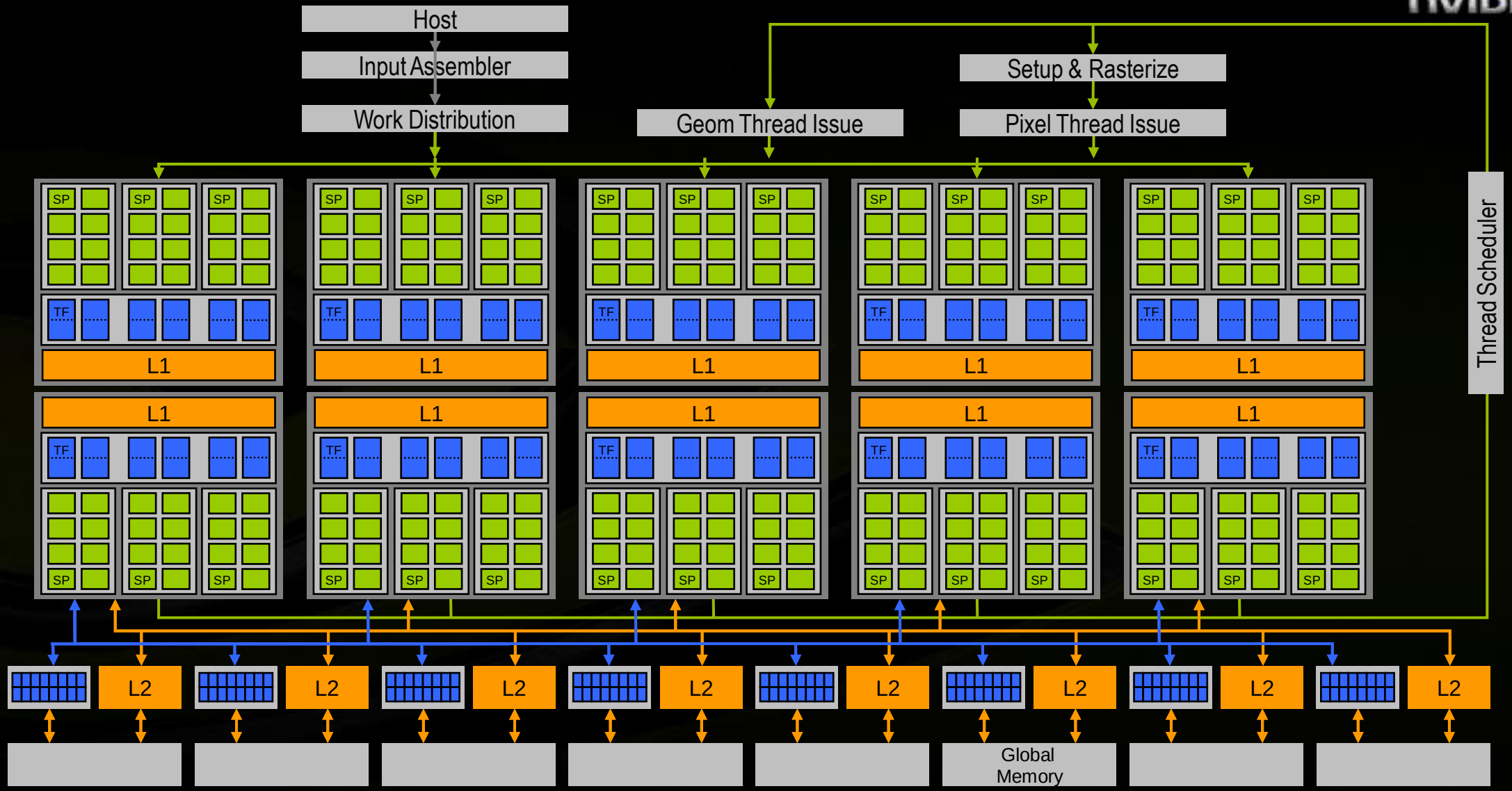
© NVIDIA Corporation 2009

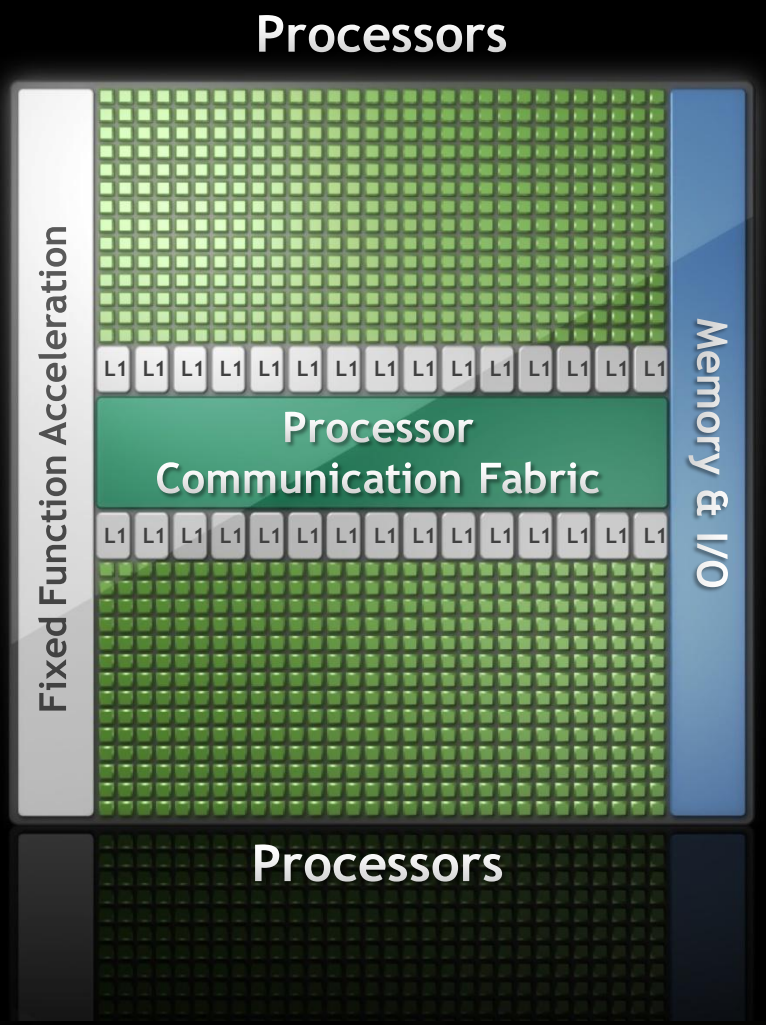
Modern GPUs: Unified Design



Vertex shaders, pixel shaders, etc. become *threads* running different programs on a flexible core

Modern GPU Architecture





NVIDIA Tesla 10-Series GPU

Massively parallel, many core architecture

240 Processor Cores

1 Teraflops - 1,000 times Cray X-MP

IEEE Compliant Double Precision Floating Point

Why is this different from a CPU?

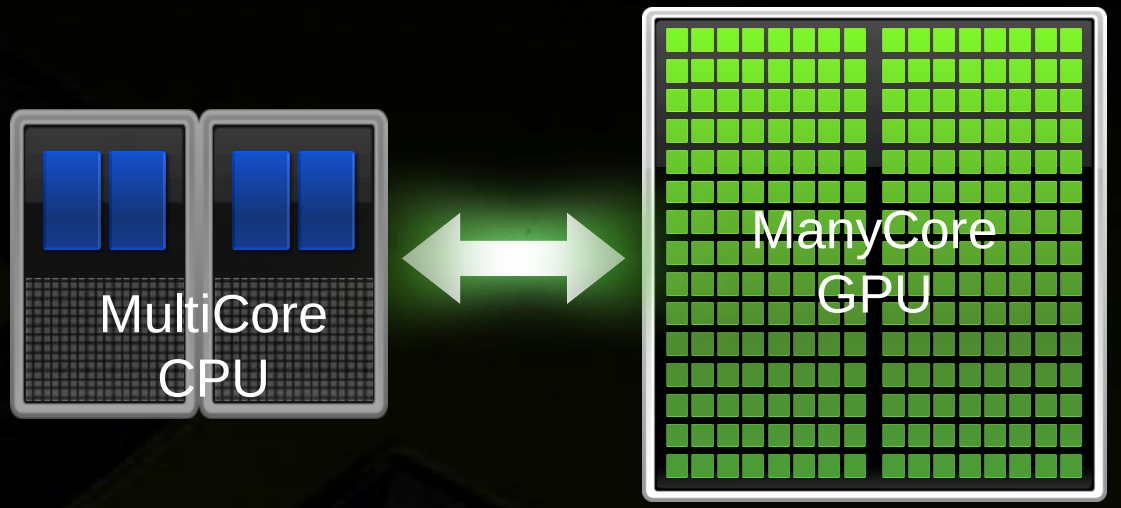
- Different goals produce different designs
  - GPU assumes work load is highly parallel
  - CPU must be good at everything, parallel or not

- CPU: minimize latency experienced by 1 thread
  - lots of big on-chip caches
  - extremely sophisticated control

- GPU: maximize throughput of all threads
  - lots of big ALUs
  - multithreading can hide latency … so skip the big caches
  - simpler control, cost amortized over ALUs via SIMD

© NVIDIA Corporation 2009

Heterogeneous Computing

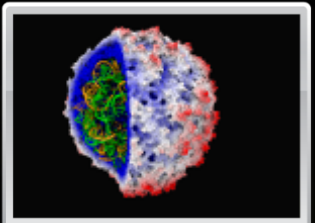


Computing with CPU + GPU

Not 2x or 3x : Speedups are 20x to 150x

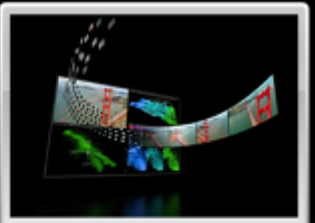
146X

Medical Imaging
U of Utah



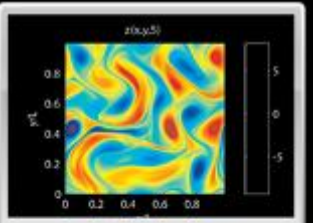
36X

Molecular Dynamics
U of Illinois, Urbana



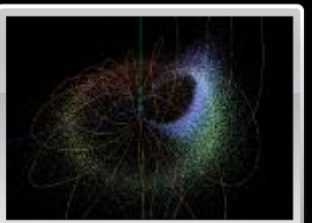
18X

Video Transcoding
Elemental Tech



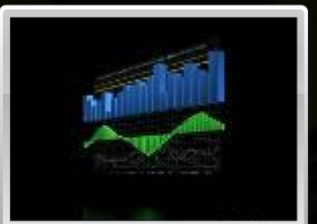
50X

Matlab Computing
AccelerEyes



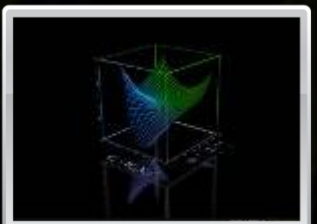
100X

Astrophysics
RIKEN



149X

Financial simulation
Oxford



47X

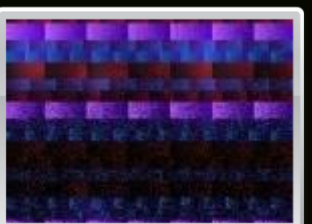
Linear Algebra
Universidad Jaime

20X

3D Ultrasound
Techniscan

130X

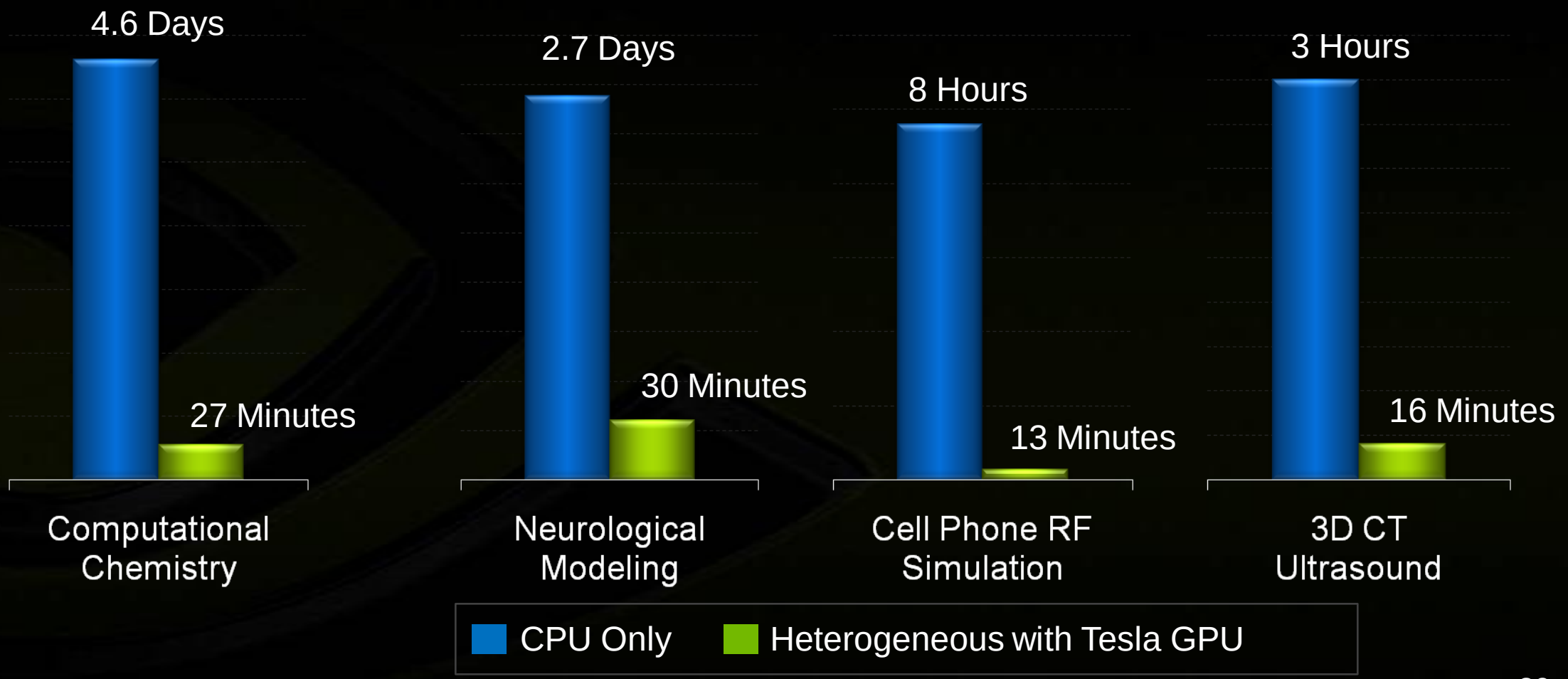
Quantum Chemistry
U of Illinois, Urbana



30X

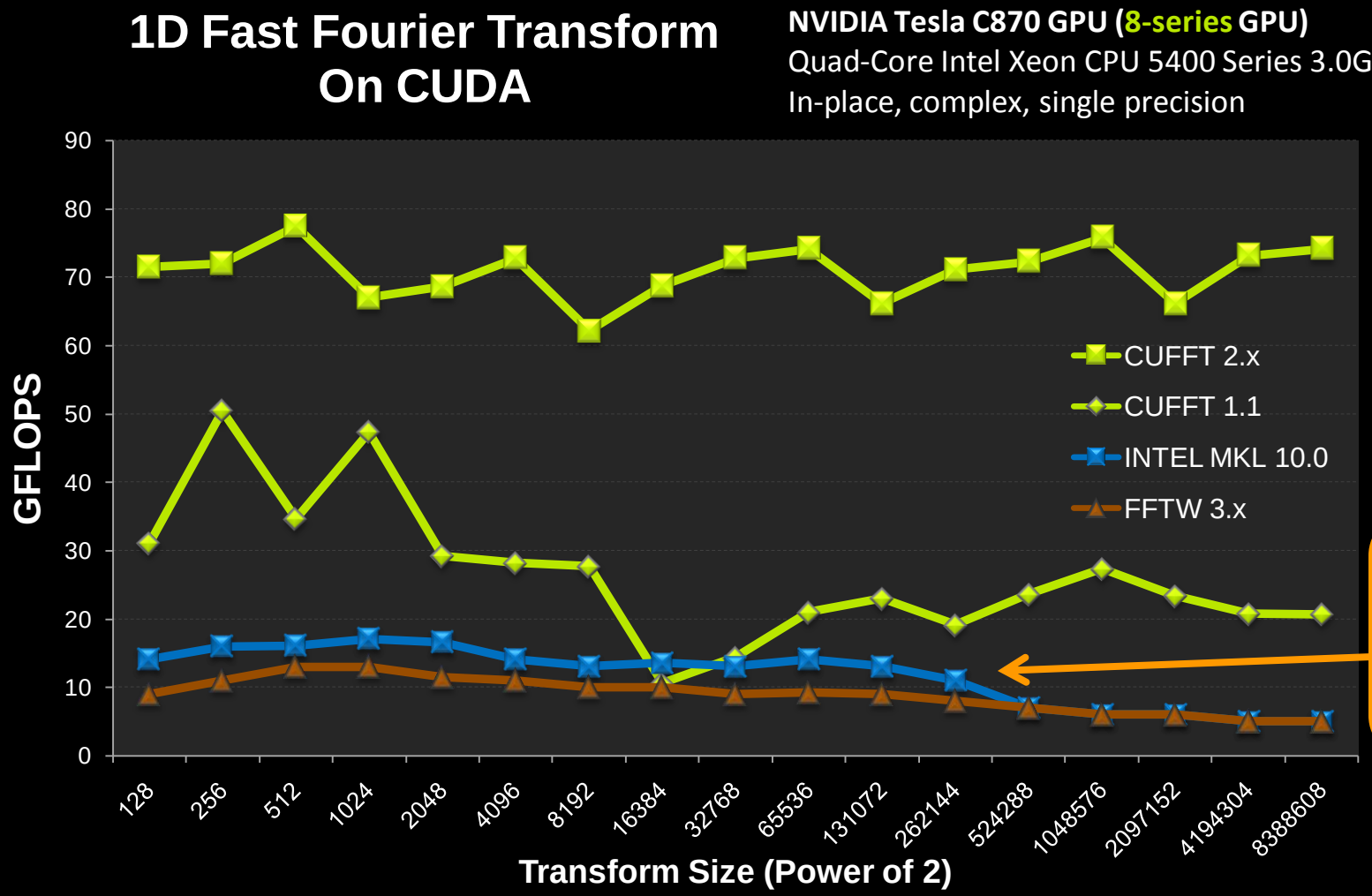
Gene Sequencing
U of Maryland

HPC: Accelerating Time to Insight



EDP 2009

FFT Performance: CPU vs GPU (8-Series)

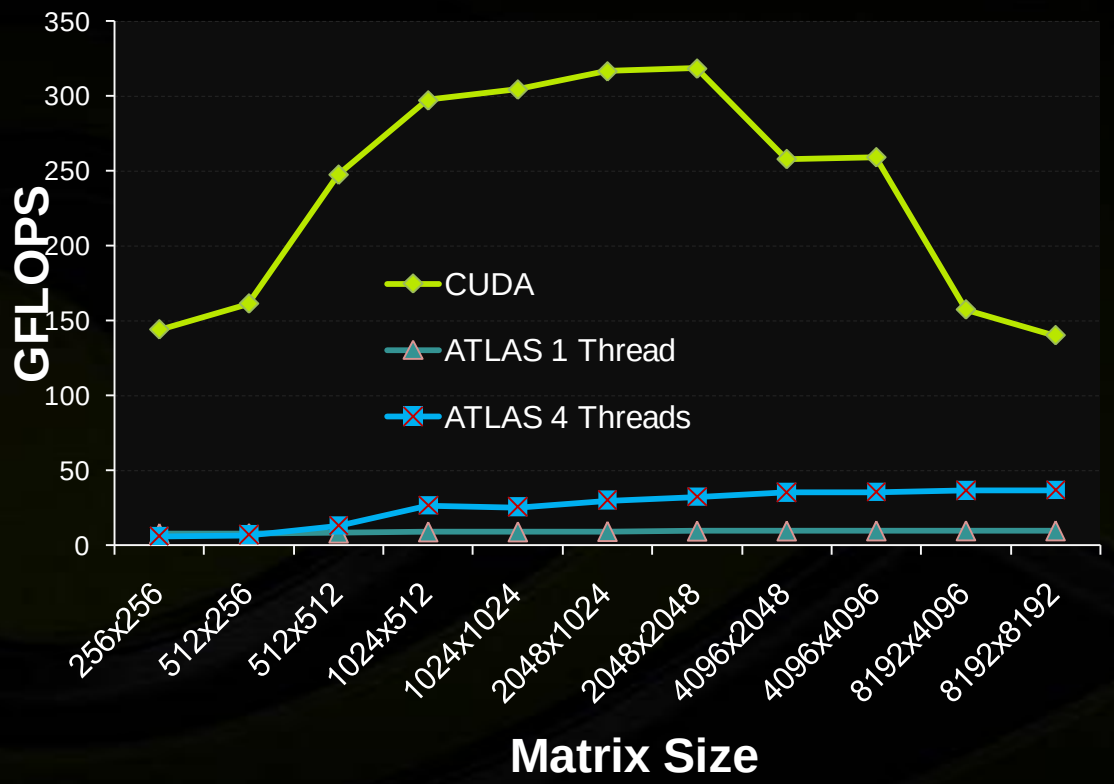


- *Intel FFT numbers calculated by repeating same FFT plan*
- *Real FFT performance is ~10 GFlops*

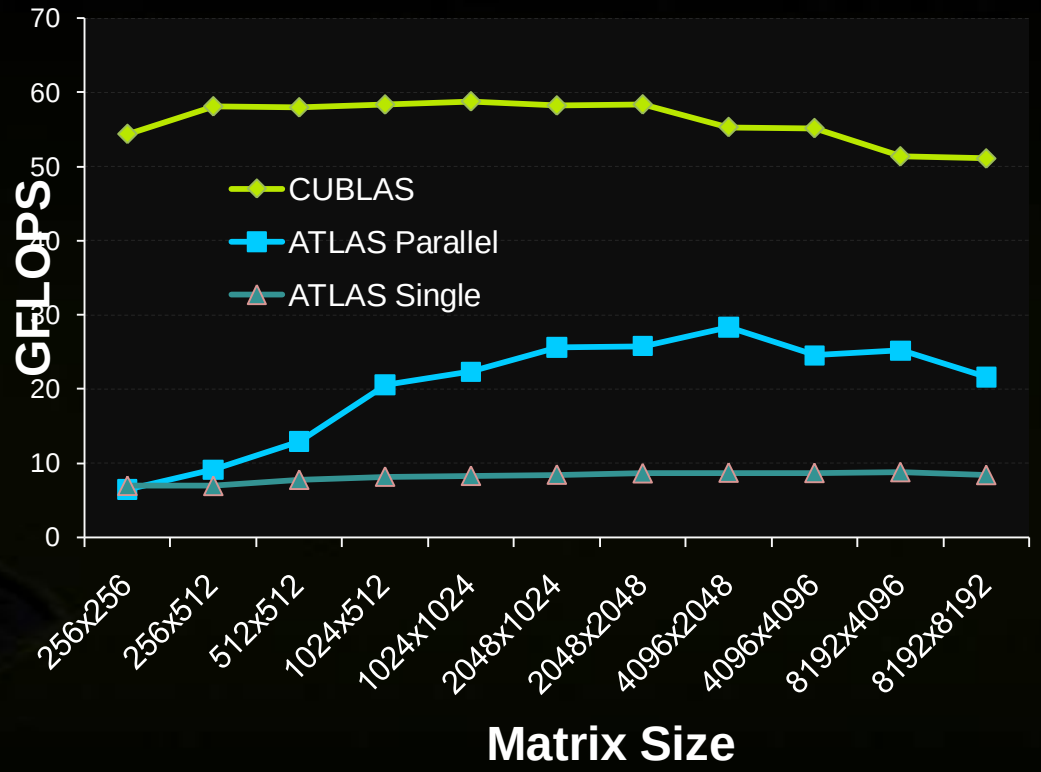
Source for Intel data : http://www.intel.com/cd/software/products/asmo-na/eng/266852.htm

BLAS: CPU vs GPU (10-series)

Single Precision BLAS: SGEMM

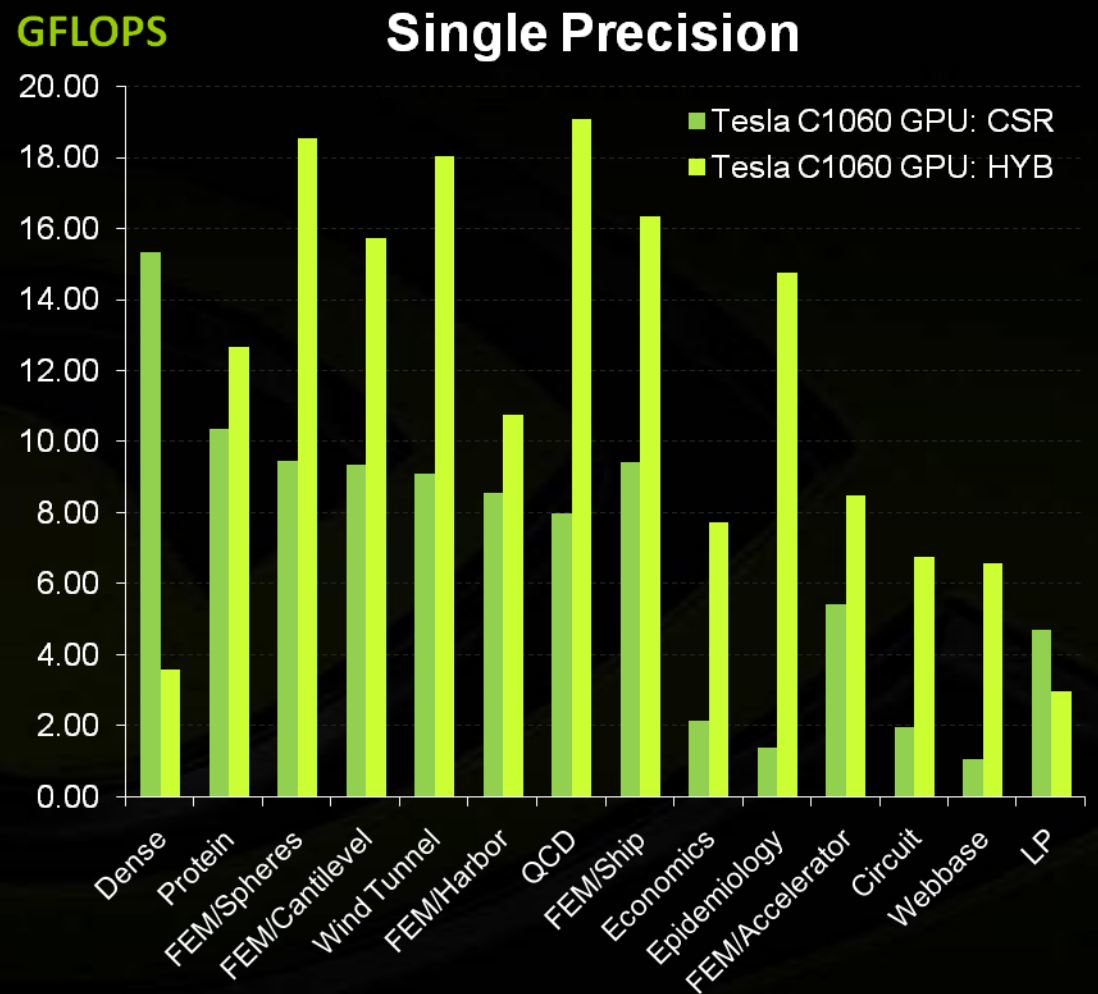


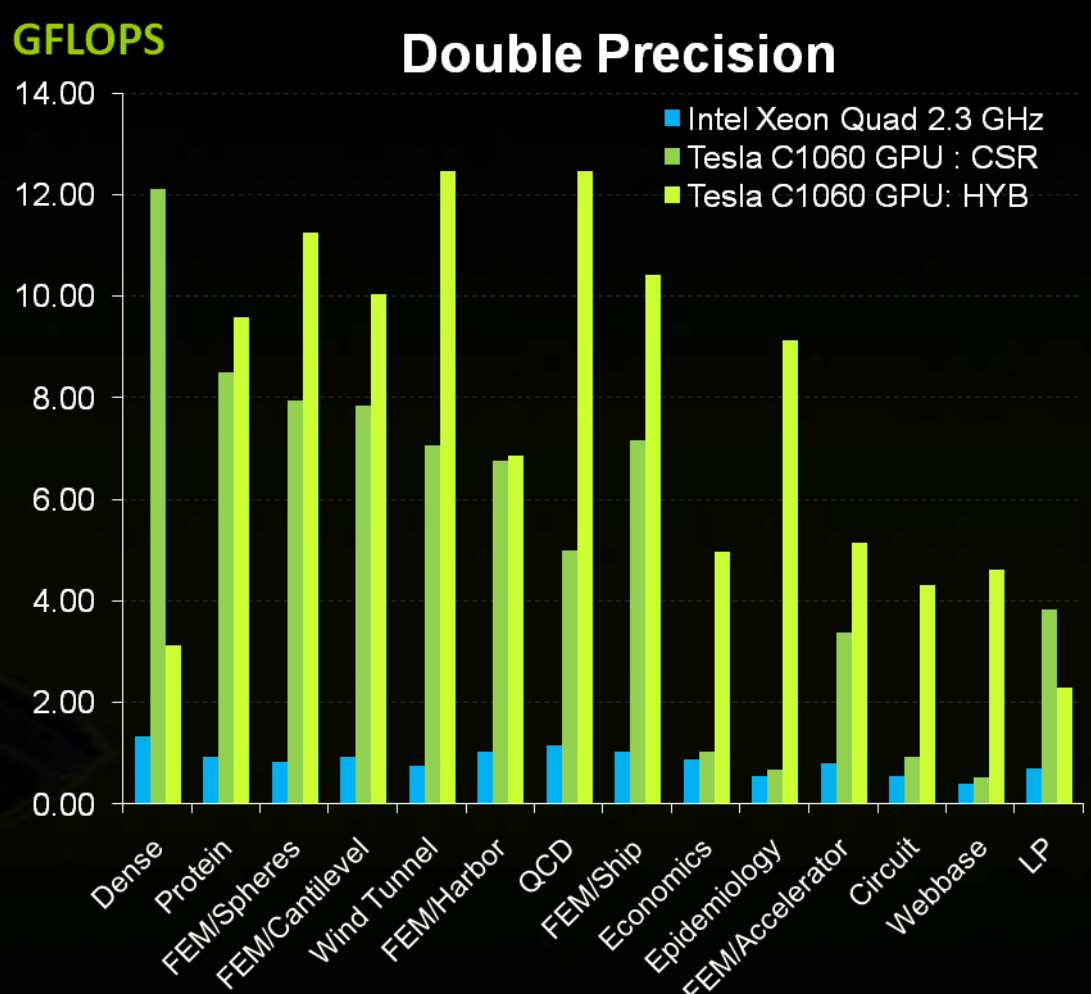
Double Precision BLAS: DGEMM



CUBLAS: CUDA 2.0, Tesla C1060 (**10-series** GPU)
ATLAS 3.81 on Dual 2.8GHz Opteron Dual-Core

Results: Sparse Matrix-Vector Multiplication (SpMV) on CUDA





CPU Results from "Optimization of Sparse Matrix-Vector Multiplication on Emerging Multicore Platforms", Williams et al, Supercomputing 2007

Different Programming Styles

- Two Programming models available today for CUDA
  - High level programming model
    - Native support for C, C++, Fortran, Java, Python, Perl, OpenCL, DIrectX Compute
  - Device level (low level) programming model
    - Use CUDA Driver API directly
- OpenCL Programming model (to be available soon)
  - A new compute API for parallel programming of heterogeneous systems
  - Allows developers to harness the compute power of BOTH the GPU and the CPU
  - A multi-vendor standards effort managed through the Khronos Group
- DirectX Compute
  - New GPU computing model introduced by Microsoft
  - Integrated with Direct3D under Win7
  - Enables more general constructs, more general data structures and more general algorithms

CUDA™

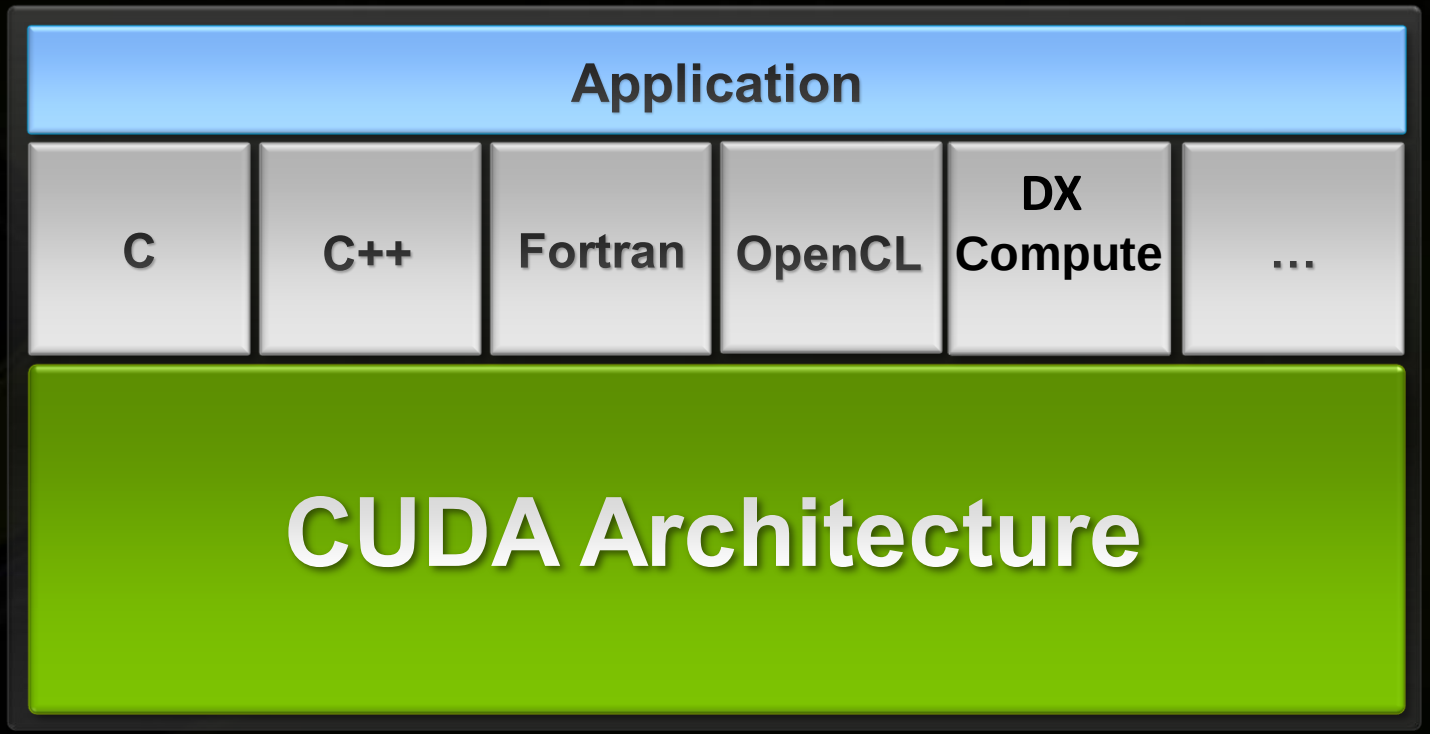
A Revolutionary Parallel Computing Architecture for NVIDIA GPUs

Supports standard languages and APIs

- C
- C++
- Fortran
- OpenCL
- DX Compute

Supported by standard operating systems

- Windows
- Mac OS
- Linux



Copyright NVIDIA 2009

NVIDIA: Leadership in GPU computing

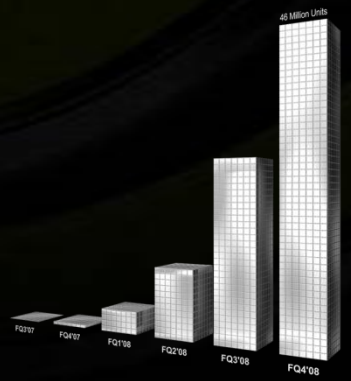
200+ Apps on CUDA Zone

30+ CUDA GPU clusters

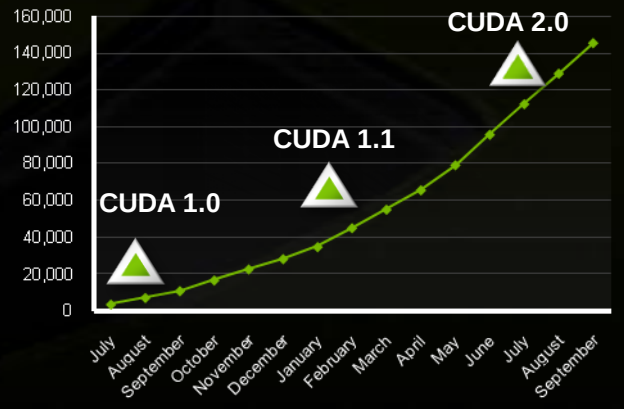
115+ Universities Teaching CUDA
900+ research papers

Duke
Erlangen
ETH Zurich
Georgia Tech
Grove City College
Harvard
IISc Bangalore
IIIT Hyderabad
IIT
Illinois
INRIA
Iowa
ITESM
Johns Hopkins
Kent State
Kyoto
Lund
Maryland
McGill
MIT
North Carolina
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

110 M+ CUDA enabled GPUs
60,000+ active developers



150K CUDA compiler downloads



EDP 2009

5000+ Customers / ISVs

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

Current Status at EDA Companies

- Numerous examples of shipping products based on CUDA / GPU computing
- All major EDA companies are exploring CUDA for accelerating tools
- Several EDA startups forming around CUDA / GPU computing
- More tool announcements coming soon

The Opportunity

- We are at a historic inflection point
- Heterogenous computing
  - Works
  - Saves money and energy
  - Is available broadly
  - Will be pervasive
  - Solves a real problem
- For EDA, represents trully disruptive technology

Customers with $$ want it !!

© NVIDIA Corporation 2009

Thank-You

More Information

- Tesla main page
  - http://www.nvidia.com/tesla
  - Product Information
  - Industry Solutions
- CUDA Zone
  - http://www.nvidia.com/cuda
  - Applications, Papers, Videos
- Hear from Developers
  - http://www.youtube.com/nvidiatesla