



**2009:  
The GPU Computing  
Tipping Point**

**Jen-Hsun Huang, CEO**



**Feb 1993**



Someday, our graphics chips will have 1 TeraFLOPS of computing power, will be used for playing games to discovering cures for cancer to streaming video to millions of people connected on the Internet.

.....Right!



**NVIDIA**<sup>®</sup>



# NVIDIA Businesses



**Consumer  
Graphics**



**Professional Workstation  
Graphics**



**High Performance  
Computing**

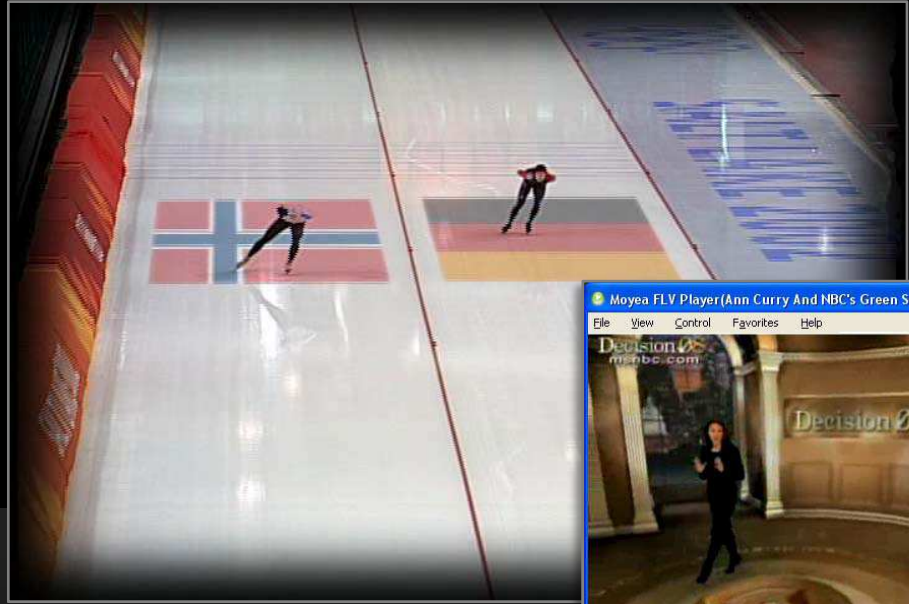


**Mobile & Embedded  
Computing**



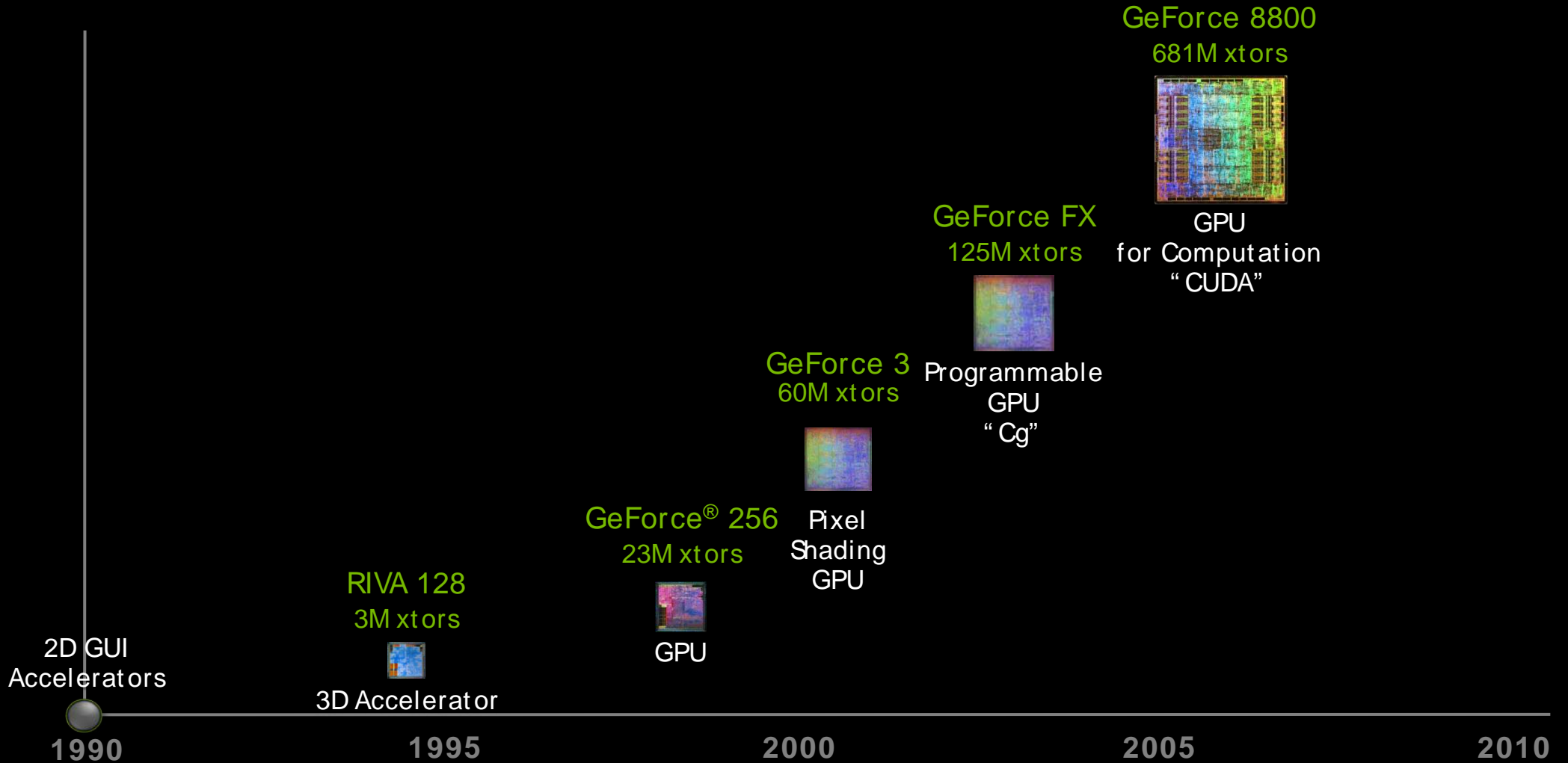


Nastia Liukin



Ann Curry  NBC Virtual Set

# NVIDIA Technology Evolution



# Revolutionizing Computer Graphics

GOPS

10,000

1,000

100

0.2



3D Accelerations  
Fixed Pipelines



Programmable Shading  
Pipelines of Processors



Computational Visualization  
Massive Array of Processors

# Computer Graphics – A Study in Parallelism



Nehalem CPU:  
3 Ghz  
4 cores  
4 way SIMD  
2 FLOPS/cycle  
= 96 GFLOPS

2,300,000 pixels/frame  
x 3 depth complexity  
x 100 shader inst./component  
x 1.5 FLOPS/inst.  
x 4 components/pixel  
x 60 frames/sec  
x 2 stereoscopic

= 500 shader GFLOPS  
(approx. 10% of graphics ops)



# Computer Graphics – A Study in Parallelism



CPU



CPU  
+ 1 year

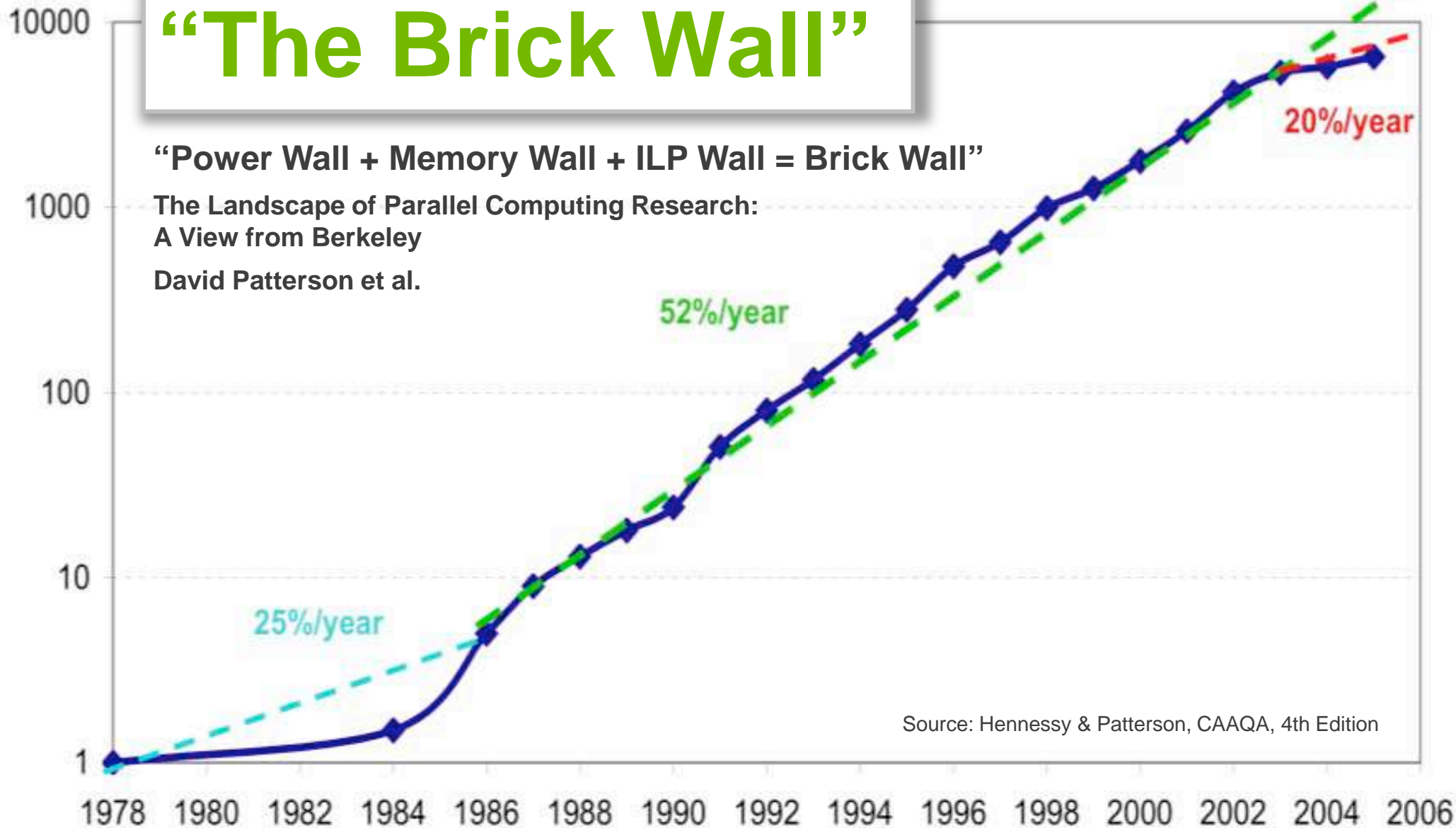


# “The Brick Wall”

“Power Wall + Memory Wall + ILP Wall = Brick Wall”

The Landscape of Parallel Computing Research:  
A View from Berkeley  
David Patterson et al.

Performance (vs. VAX-



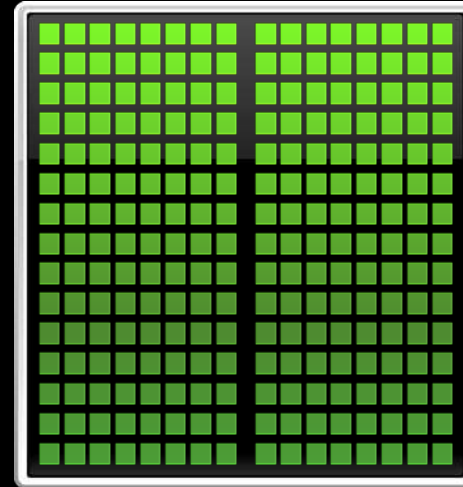
Source: Hennessy & Patterson, CAAQA, 4th Edition

# Co-Processing

*The Right Processor for the Right Tasks*



+



CPU

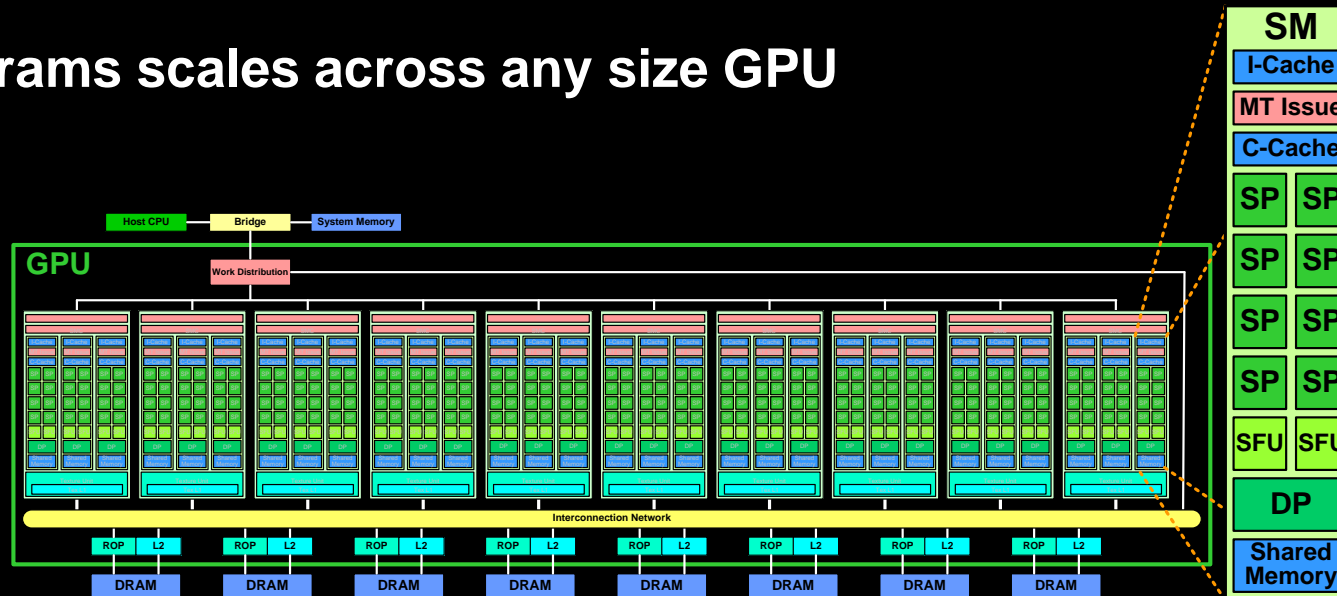
GPU

# NVIDIA CUDA Parallel Compute Architecture



- Many processors – eventually thousands
- Latency tolerant - execute 1000's of threads
- General load/store
- On-chip shared-memory
- CUDA programs scales across any size GPU

240 SP Cores



# Computer Graphics – A Study in Parallelism



CPU



CPU  
+ 1 year



CPU + GPU



$$\frac{1}{(1 - P) + P/s}$$

Where P=99%, S=100x

$$S = 6.4 \text{ TOPS} / (0.5 * 96\text{GFLOPS}) = 100 \text{ ☺}$$

# Computer Graphics – A Study in Parallelism



CPU



CPU  
+ 1 year



CPU + GPU



$$\frac{1}{(1 - P) + P/S}$$

Where P=99%, S=100x

CPU + GPU



Overlapped Execution

# Computer Graphics – A Study in Parallelism



<b>System Configuration</b>	<b>Fallout 3</b> 1920x1200; 4x AA	<b>Far Cry 2</b> 1920x1200; 4x AA
<b>Core i5</b> + GeForce GTX 275	69.1 FPS	49.6 FPS
<b>Core i7</b> + GeForce GTX 275	69.8 FPS	50.7 FPS

# The Next Big Thing – Physics

## *Simulate Amazing Worlds*

GOPS

10,000

1,000

100

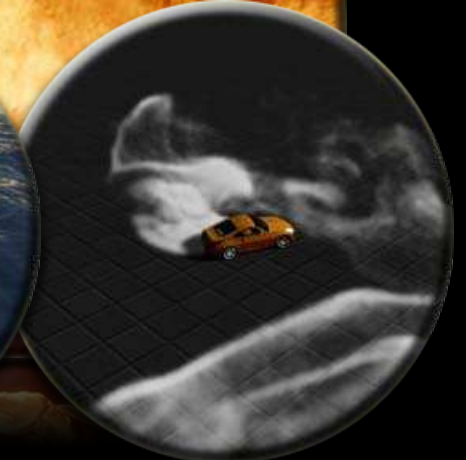
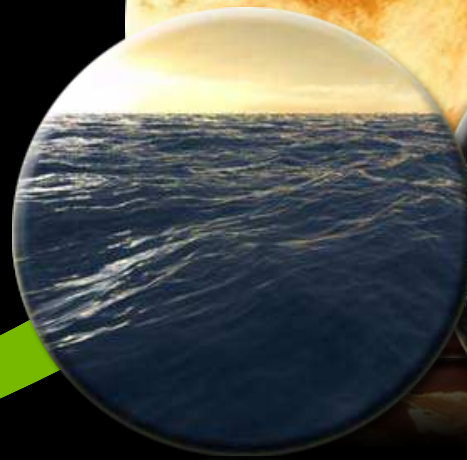
0.2



3D Accelerations  
Fixed Pipelines



Programmable Shading  
Pipelines of Processors



Computational Visualization  
Massive Array of Processors







ILM – Siggraph 2009


Directable, high resolution simulation of fire on the GPU

“the GPU gave us unbelievable speed-ups over the typical CPU. We built a GPU farm that could handle these massive simulations. **What would take a day to run on a CPU, we were able to simulate in 40 minutes.** The graphics processor is ideal for handling millions of instructions in split-seconds.”

Tim Alexander and Robert Weaver, ILM  
Post July 1, 2009

# Co-Processing


*Ideal for Ray Tracing*



User Input

JIT Compile

Disk I/O



Animation

Acceleration Build

Intersection



Traversal


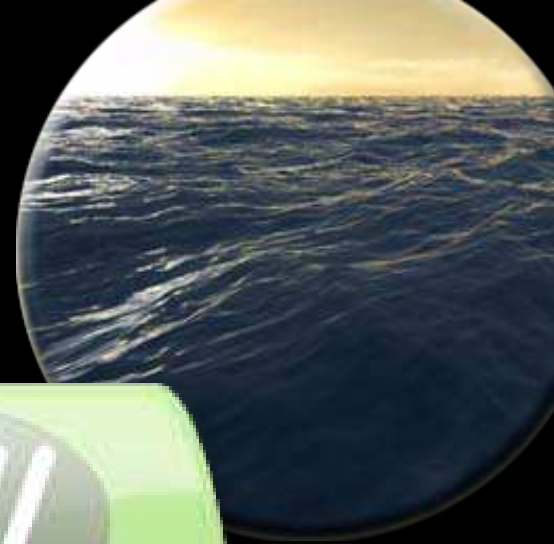
Shading

Camera Model


Tone Mapping

# Co-Processing


*Ideal for Physics Processing*



Network  
Synchronization



AI  
Ragdoll physics  
Ray Casting




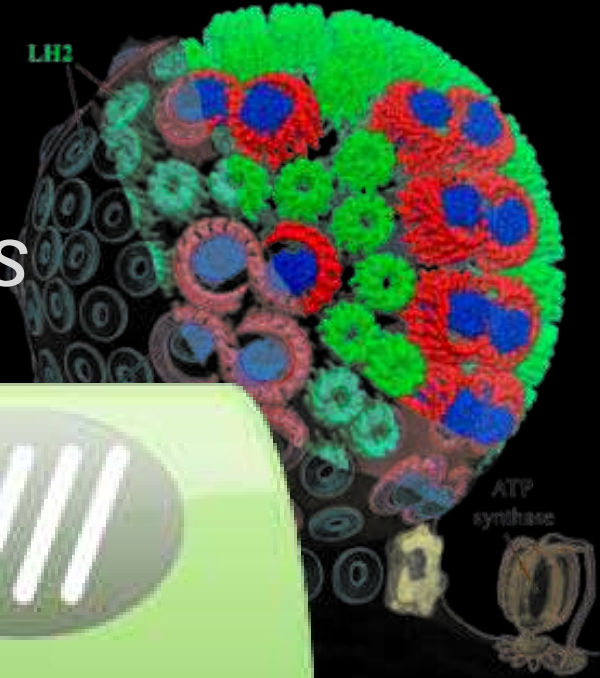
Collision  
Detection  
Deformation  
Smoothed Particle  
Hydrodynamics  
(SPH)  
Particles

# PhysX™ by NVIDIA



# Co-Processing


*Ideal for Molecular Dynamics*



Steering

Time Stepping


Disk I/O



Pairlist calculation

Pairlist update

Non-bonded force calculation



Fluorescence microphotolysis

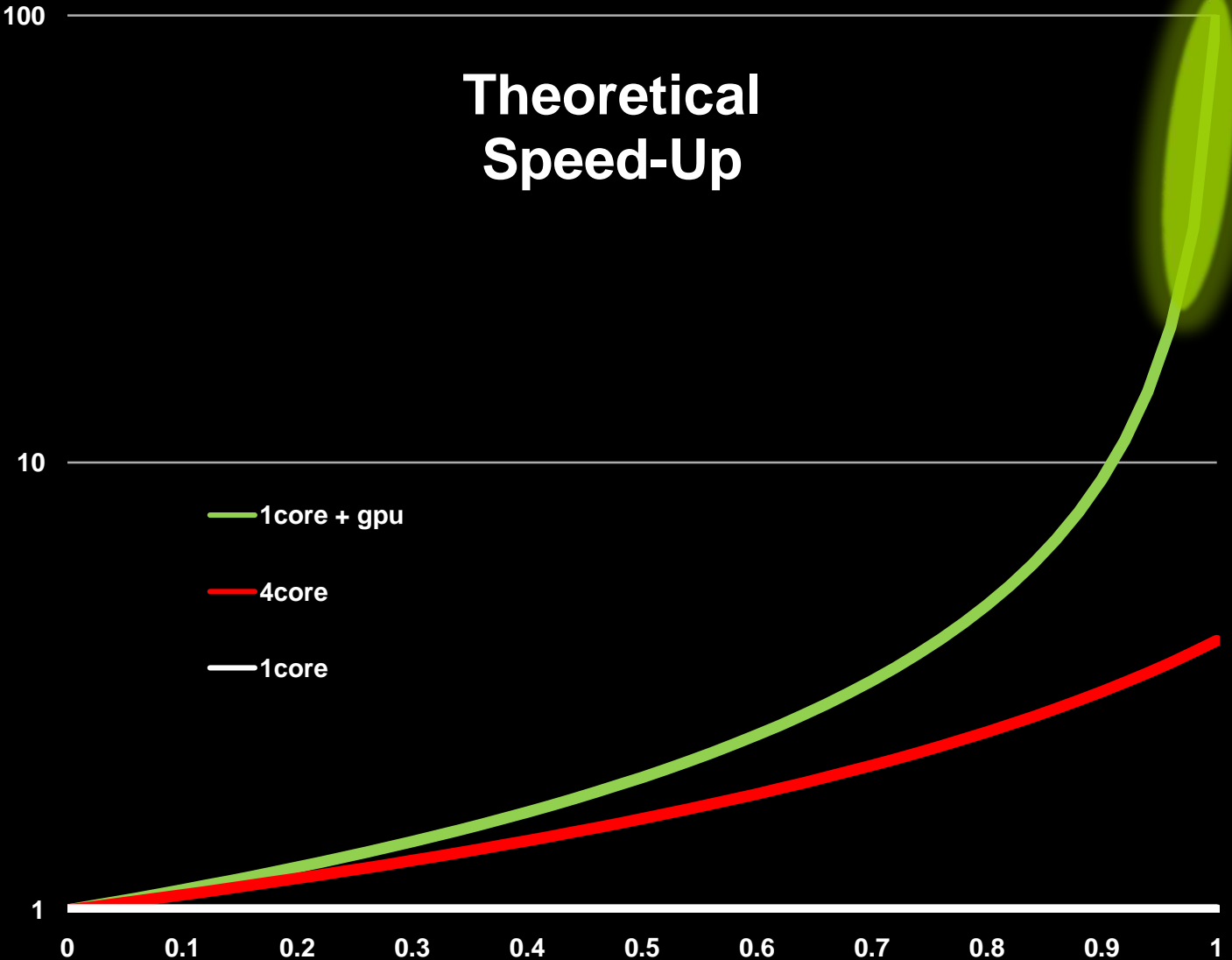
Direct Coulomb Summation

Cutoff potential summation

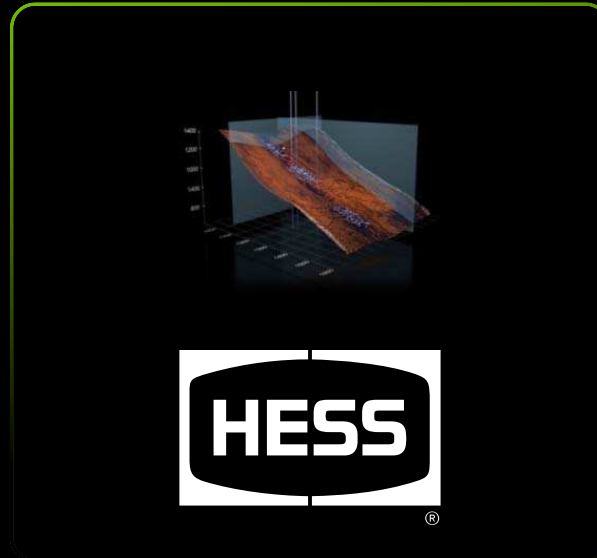
# Huge Speed-Ups Across Many Fields



Algorithm	Field	Speedup
2-Electron Repulsion Integral	Quantum Chemistry	130X
Lattice Boltzmann	CFD	123X
Euler Solver	CFD	16X
GROMACS	Molecular Dynamics	137X
Lattice QCD	Physics	30X
Multifrontal Solver	FEA	20X
nbody	Astrophysics	100X
Simultaneous Iterative Reconstruction Technique	Computed Tomography	32X



# Oil & Gas: Seismic Processing



1

**Equal Performance**

1

32 Tesla S1070s

**31x Less Space**

2000 CPU Servers

~\$400K

**20x Lower Cost**

~\$8M

45 kWatts

**27x Lower Power**

1200 kWatts





# Co-Processing

*The Right Processor for the Right Tasks*

## 2015 Projection

CPU-Alone	$1.2^6$	3X
CPU+GPU	$50 * 1.5^6$	570X

# Universal Translator



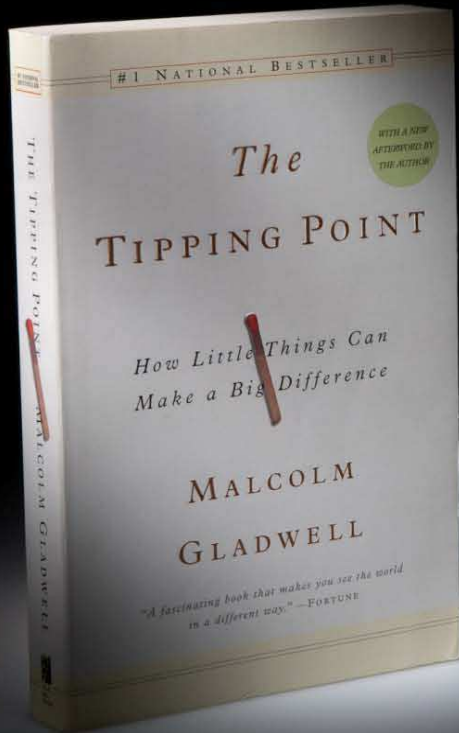
# Augmented Reality





**RTT**  
challenging reality

**RealView<sup>2</sup>**



GPU Computing has reached  
“the tipping point”