

---

# Multi-Core and Beyond: Evolving the x86 Architecture

Phil Hester, SVP and CTO, AMD  
August 21, 2007

# Agenda

**The Accelerated Processing imperative**

**Shift to software/hardware parallelism**

**Role of the GPU as floating point accelerator**

**Peta-scale processing for the masses**

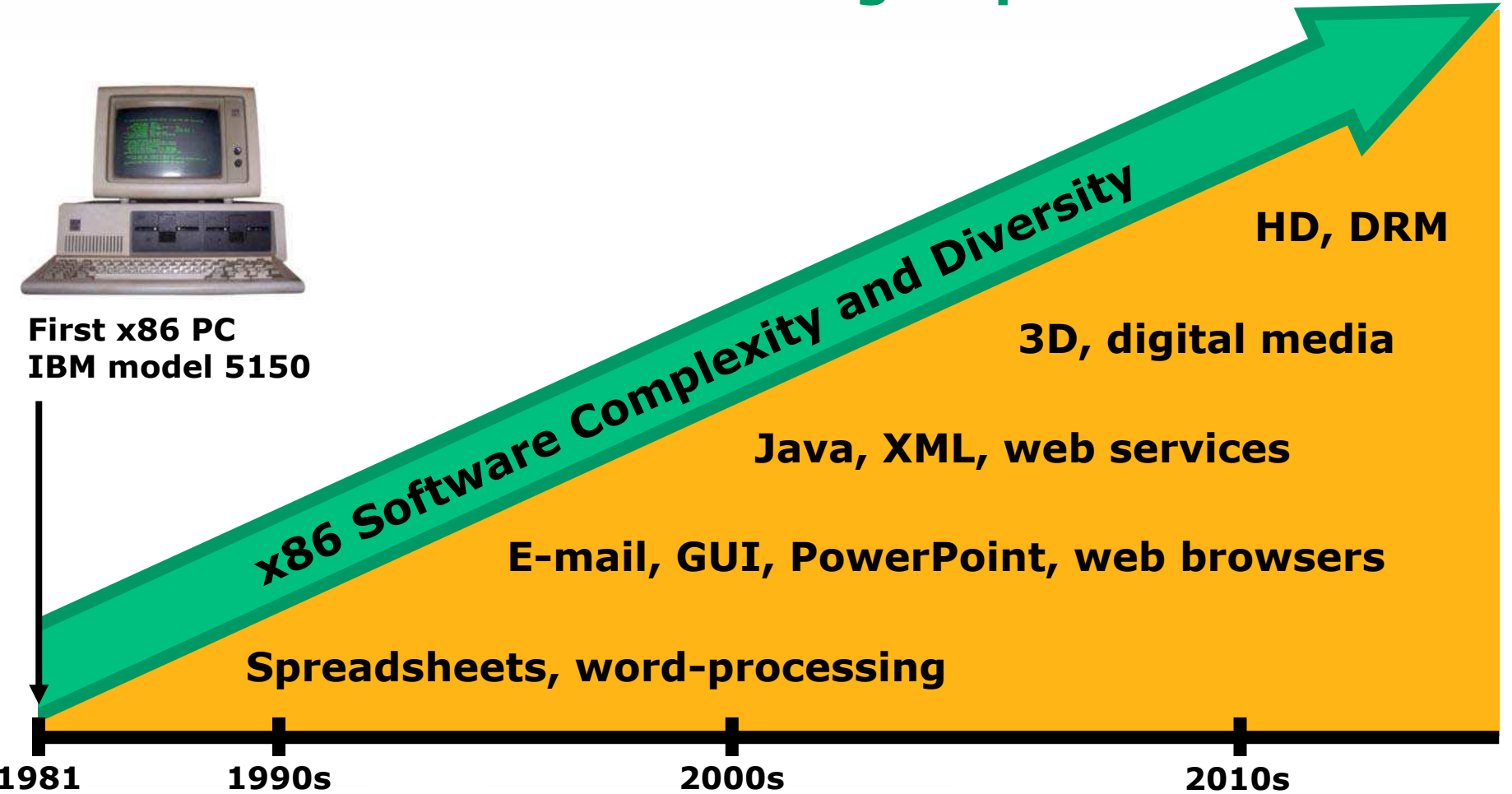
**Attack of the GPUs**

**Emergence of Accelerated Processing Units (APUs)**

# The Accelerated Processing Imperative

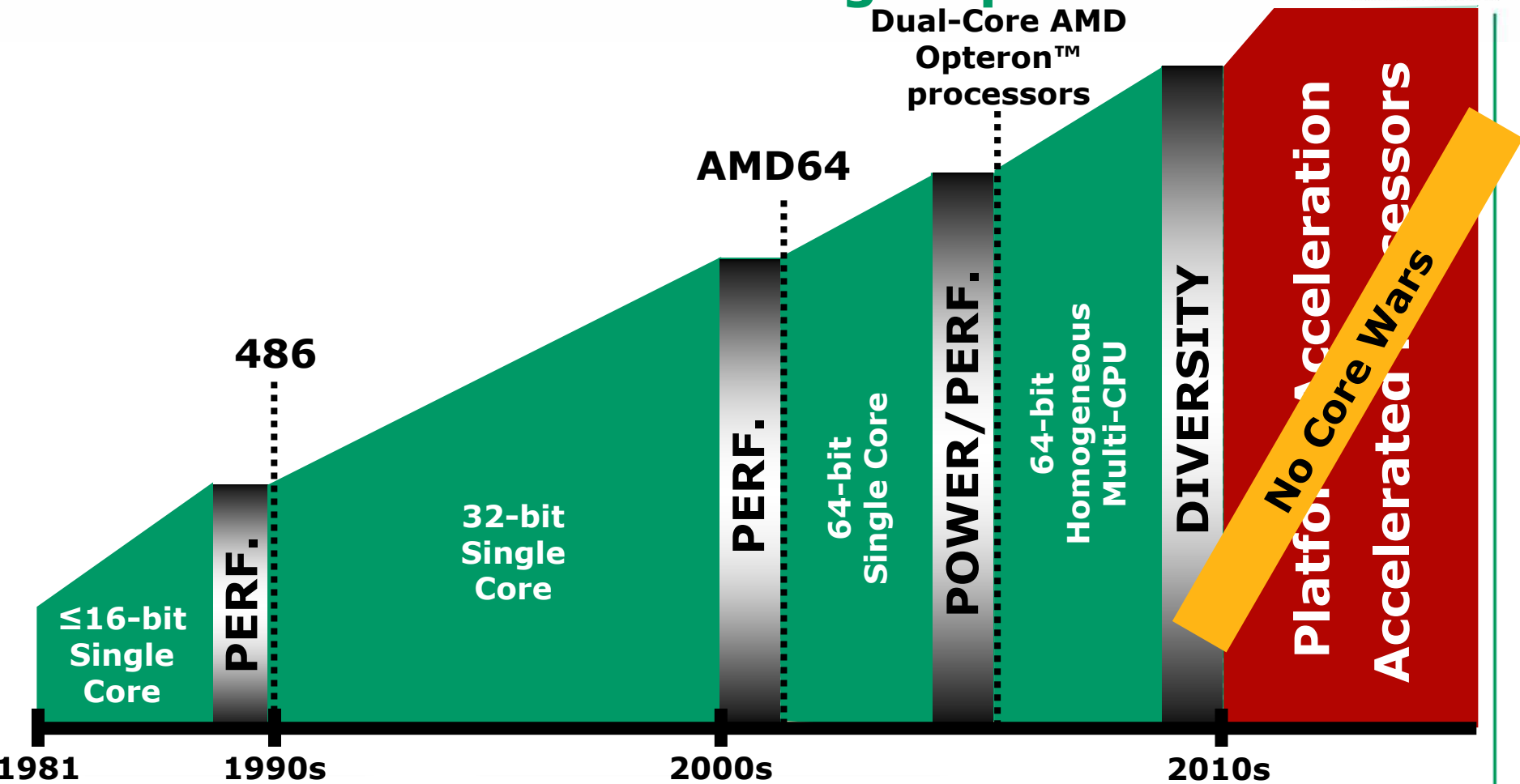


First x86 PC  
IBM model 5150



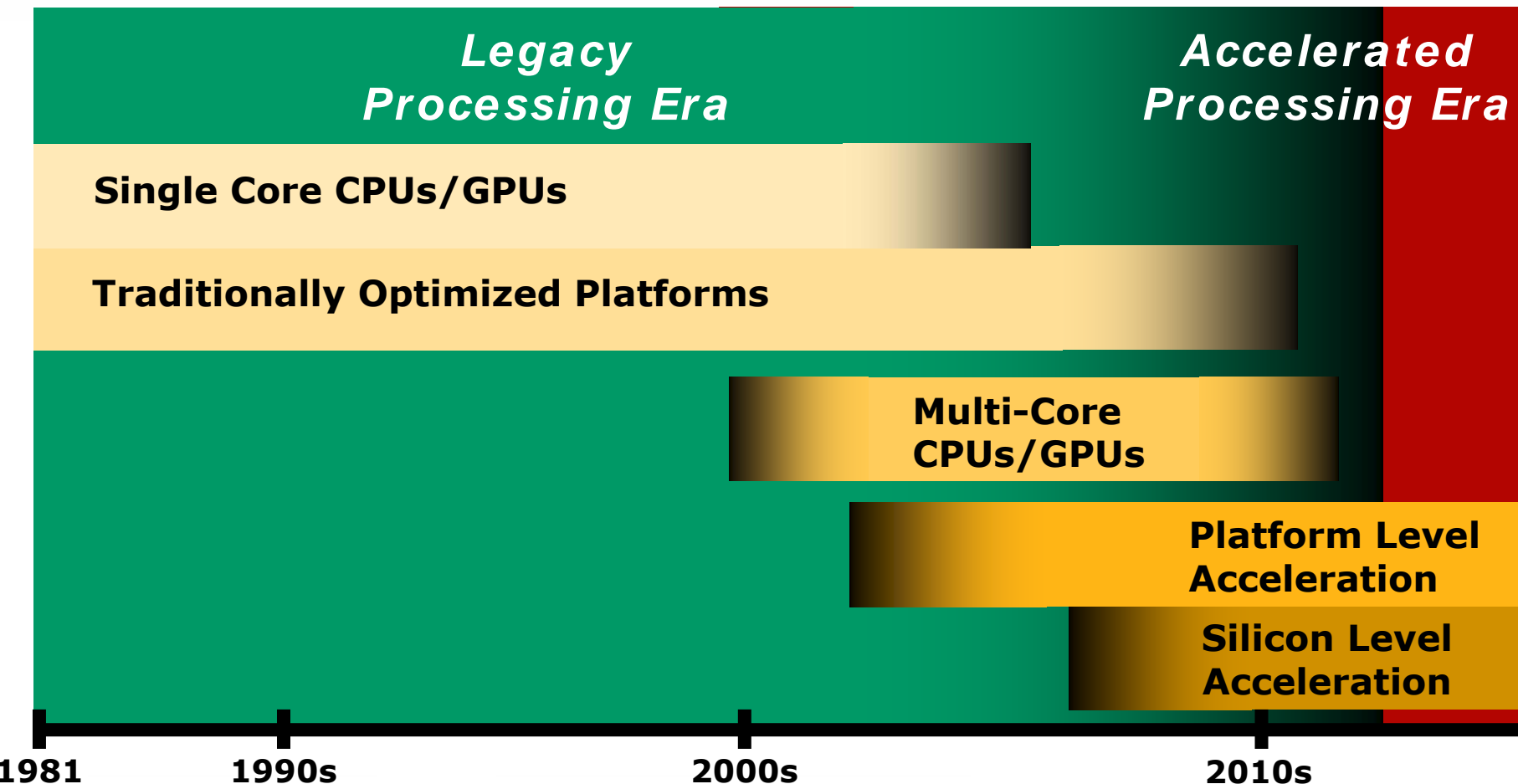
**x86 applications, workloads and usage models continue to rapidly diversify**

# The Accelerated Processing Imperative



**By the end of the decade, homogenous multi-core becomes increasingly inadequate**

# The Next Major x86 Inflection Point



**In coming era of Accelerated Computing, heterogeneous multi-core goes mainstream**

# Impact on End-User Experience vs. Computing Infrastructure

*New application value delivery heavily defined by Cloud capabilities/ logic*

## **Datacenters** *Tectonic shift*

- *Massively scalable systems*
- *Mainstream processors*
- *Consolidated hardware*

## **Clients**

*Rapid evolution*

- *PC/ CE lines blur*
- *Digital media on every device*
- *Mainstream 3D/ HD*
- *Background services explosion*

*End-user experience rapidly changing, huge SW/ HW paradigm shift occurring*

**Web 2.0/3.0 experience driving major shift to parallelism in server *and* client workloads**

# Easing and Accelerating the Parallel Software/Hardware Evolution

**Hardware extensions for software parallelism (xSP)**

**Acceleration for software transactional memory**

**Fast context switching for light-weight parallelism**

**Accelerated cross-core communication**

**Light-Weight Profiling**



**Open collaboration to enable a more productive parallel programming environment**

# Accelerated Computing Software Stack

## Integrated development environments and analysis tools

**Infrastructure and high performance software**  
(HPC, video, consumer, database, mail, web servers)

**Compilers (C, C++, Fortran)**  
**AMD Stream extensions and performance libraries**

**Web-based and user-level application software**

### Open Source Interface

**Runtime environments** (AMD RT, JVR, CLR)

**Operating systems** (Windows®, Linux®, Solaris) **and Hypervisors** (VMware, Xen)

**Instruction Set (GPU, AMD64, SSE, AMD-V, xSP)**

**AMD64 processors** (CPU, GPU, Fusion)



# Stream Computing Futures

*Taking HPC technology mainstream*

## Highest Compute per mm<sup>2</sup>, dollar, & watt

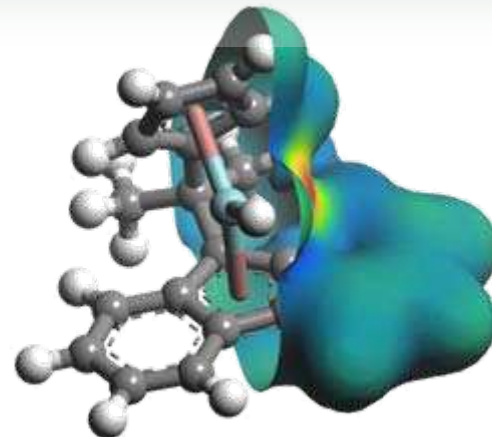
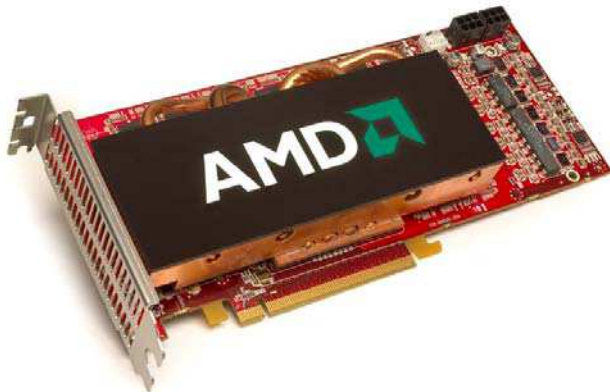
- Performance & bandwidth for HPC applications

## Personal Stream Computing

- Image, video and data intensive processing
- Image/feature search

## Programming model standardization

- Whole platform not just GPU-centric



# Graphics Processing Today:

## *Moving Beyond Rendering to Dynamics*



**Enthusiast computing continues to push the boundaries of cinematic realism**

# The Changing Role of the GPU

Ruby "Whiteout" Demo

(Removed in PDF version to reduce file size)

# Ruby Statistics

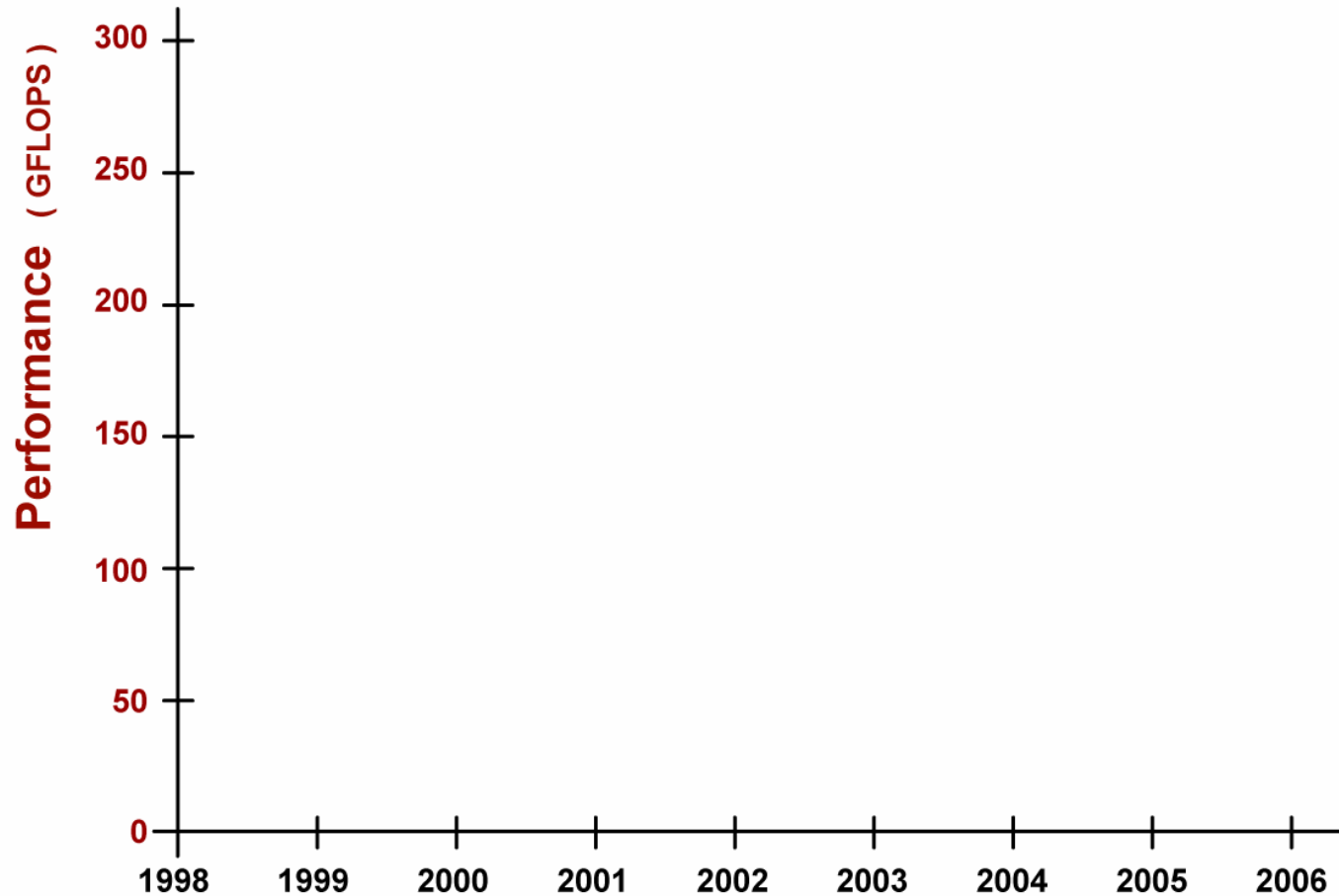
	<b>DoubleCross</b>	<b>The Assassin</b>	<b>Whiteout</b>
Ruby Polygons	80,000	80,000	200,000
Avg. Triangles/Frame	227,212	546,087	1,069,503
Max Triangles/Frame	556,305	1,018,312	2,150,521
No. of Pixel Shaders	100	316	210
Avg. Pixel Shader Length	20	74	142
Facial Animation Targets	4	4	> 128
ALU:Tex Ratio	4:1	7:1	13:1
	<b>2004</b>	<b>2005</b>	<b>2006</b>

# Ruby Statistics

	DoubleCross	The Assassin	Whiteout
Ruby Polygons	80,000	80,000	200,000
Avg. Triangles/Frame	227,212	546,087	1,069,503
Max Triangles/Frame	556,305	1,018,312	2,150,521
No. of Pixel Shaders	100	316	210
Avg. Pixel Shader Length	20	74	142
Facial Animation Targets	4	4	> 128
ALU:Tex Ratio	4:1	7:1	13:1
	<b>2004</b>	<b>2005</b>	<b>2006</b>

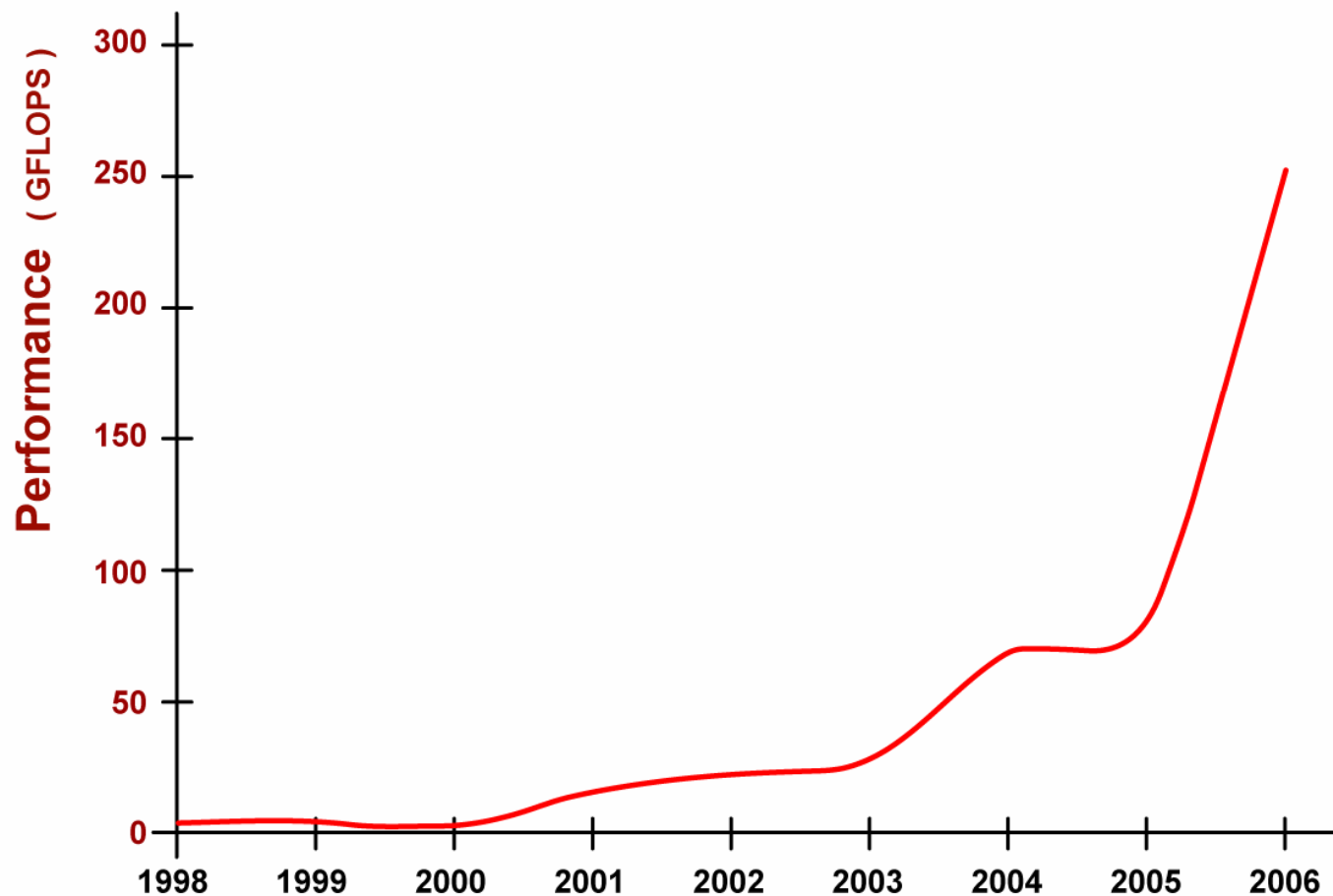
# Graphics Processor Performance:

*Drive for Realism Demanded Significant Performance*



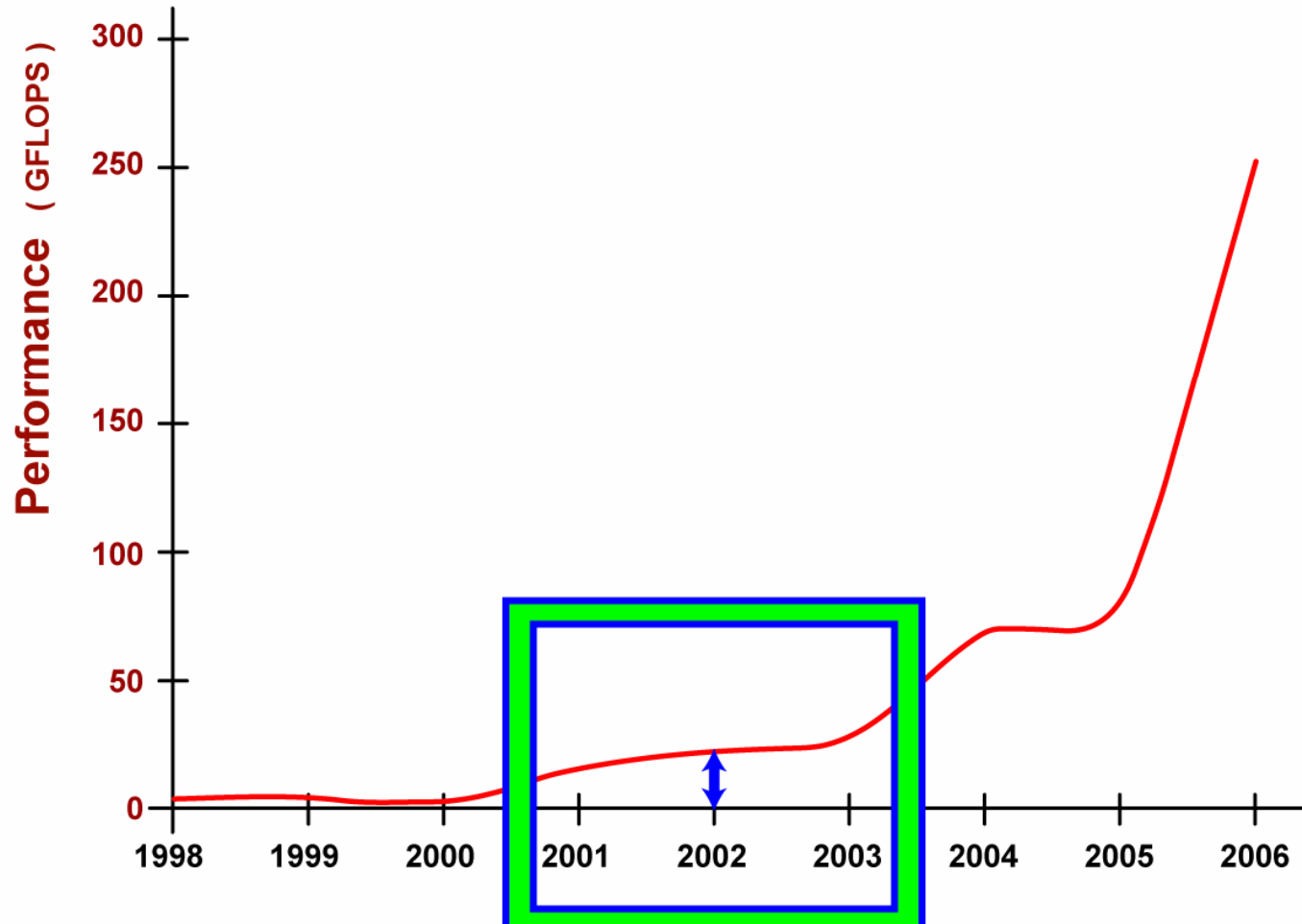
# Graphics Processor Performance:

*Drive for Realism Demanded Significant Performance*



# 2002:

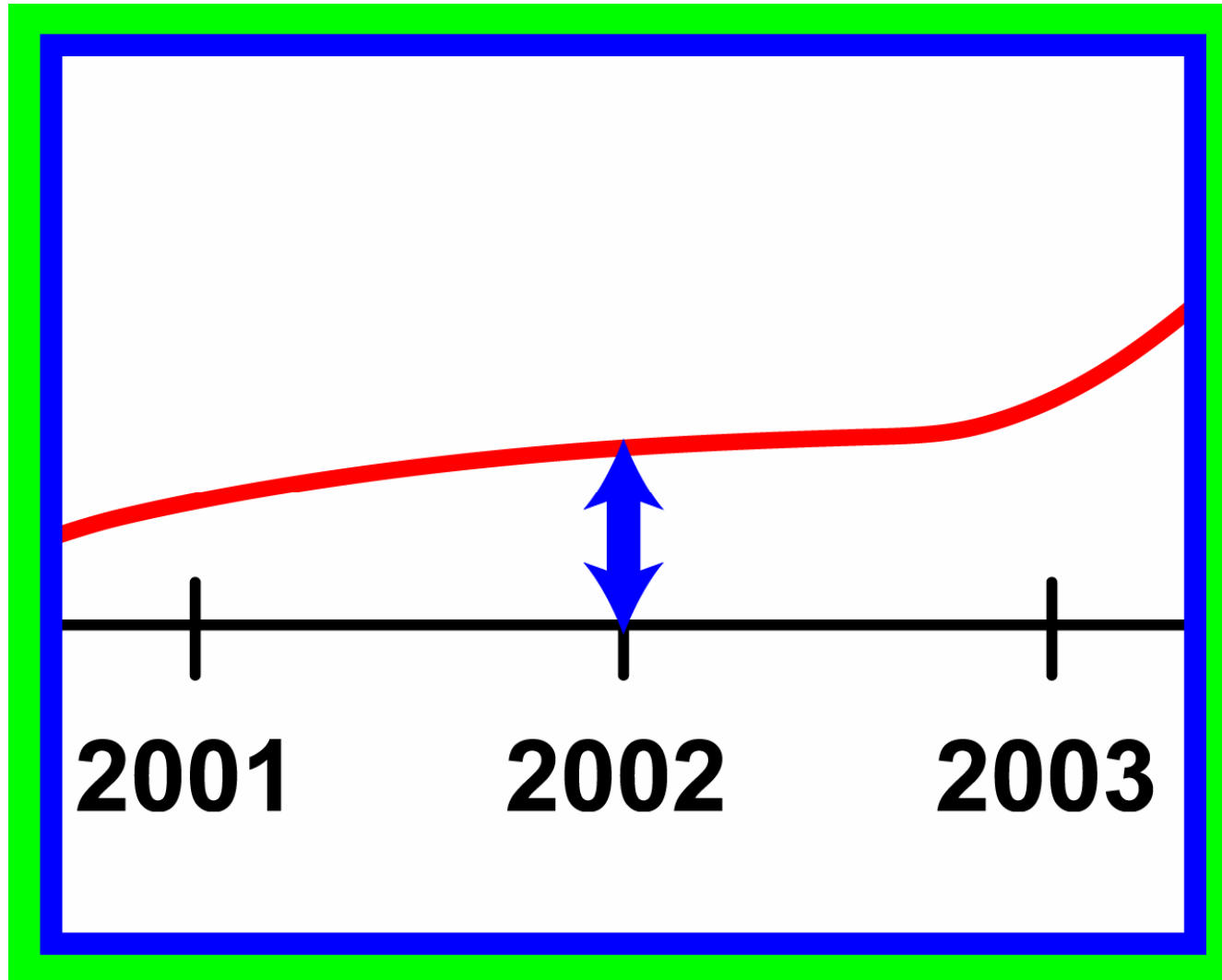
## *A Significant Point in Graphics Processor History*





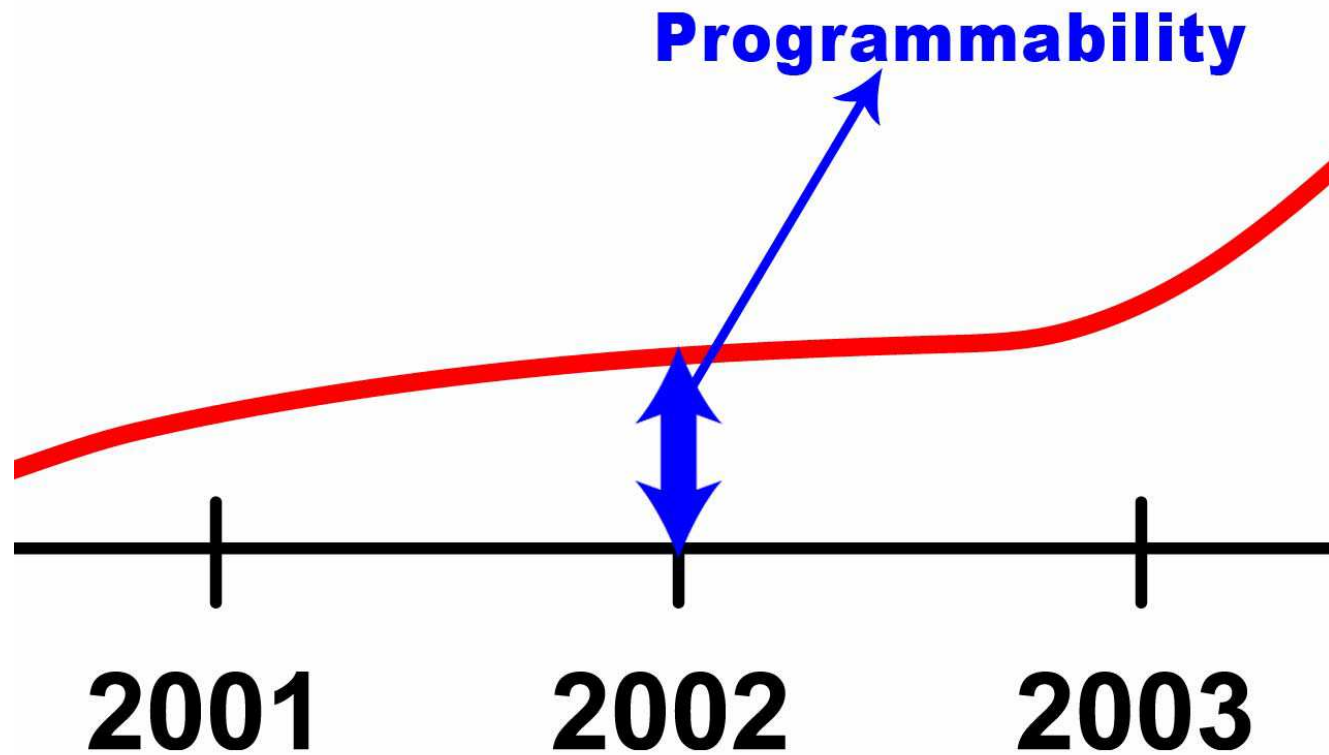
**2002:**

*A Significant Point in Graphics Processor History*



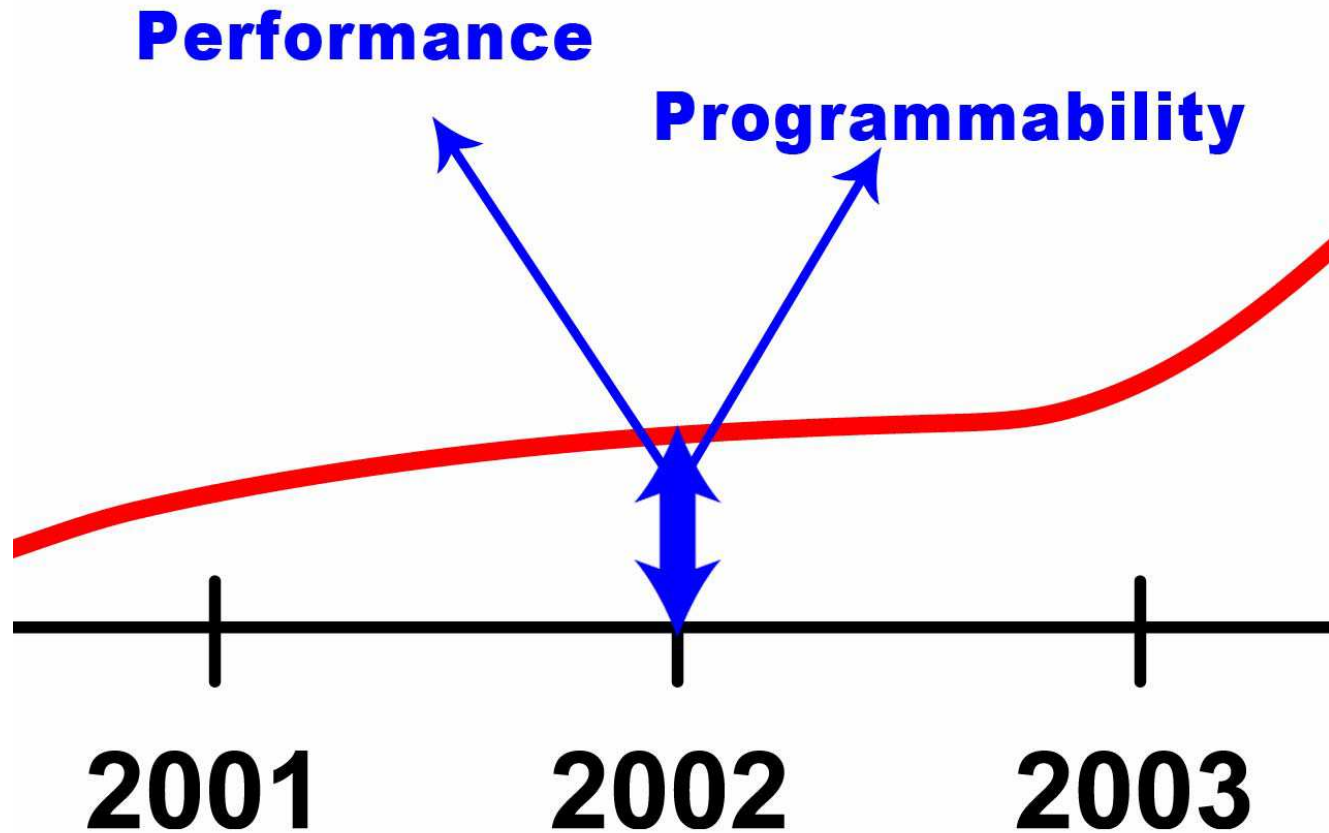
# 2002:

*A Significant Point in Graphics Processor History*



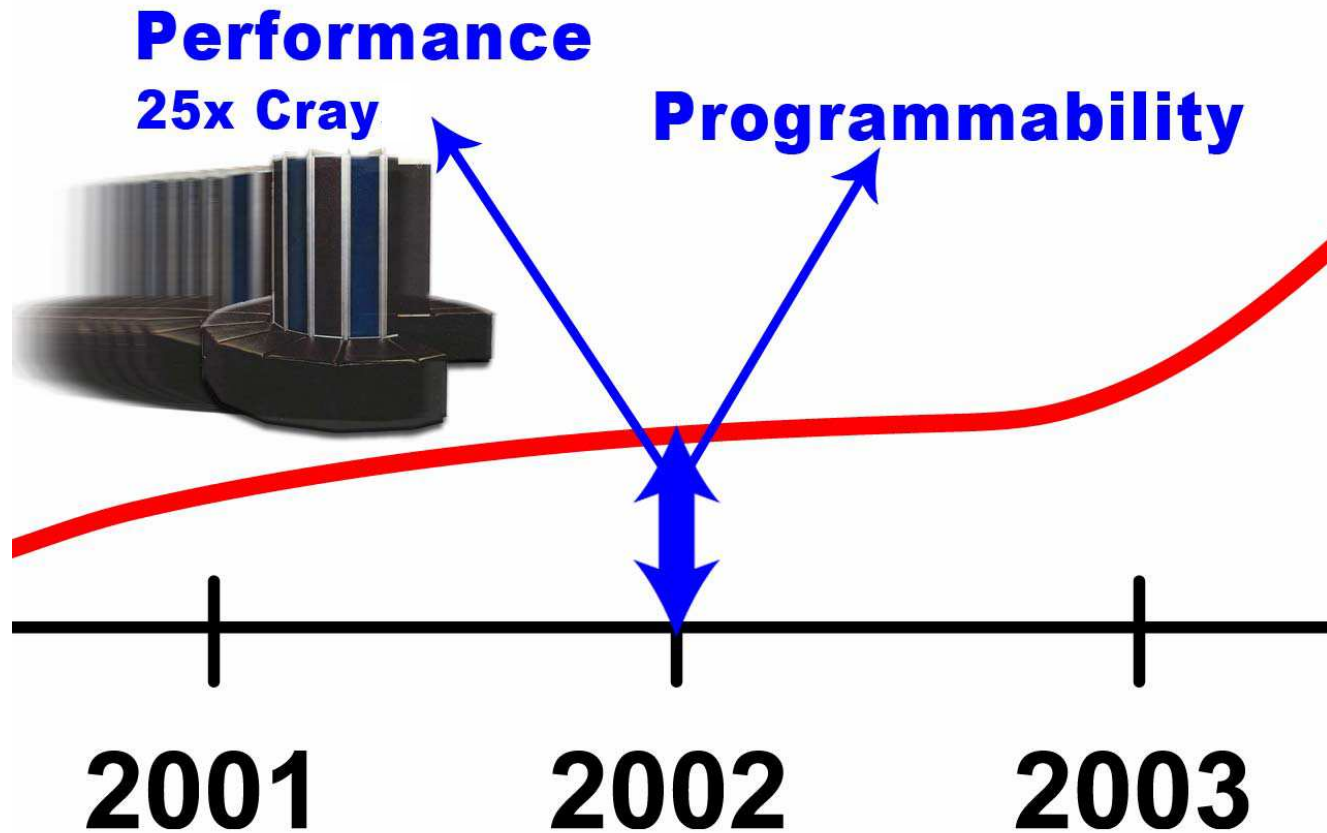
**2002:**

*A Significant Point in Graphics Processor History*



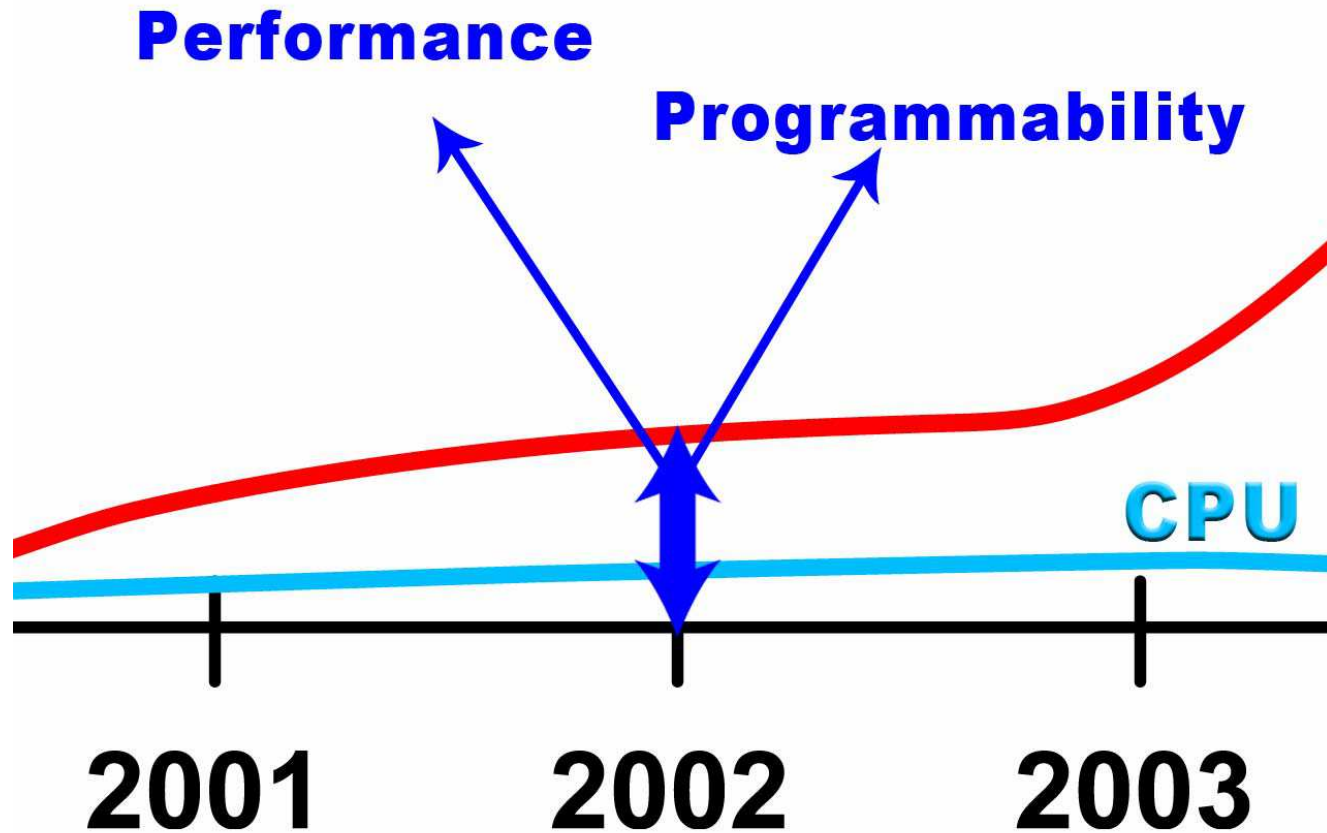
**2002:**

*A Significant Point in Graphics Processor History*



**2002:**

*A Significant Point in Graphics Processor History*

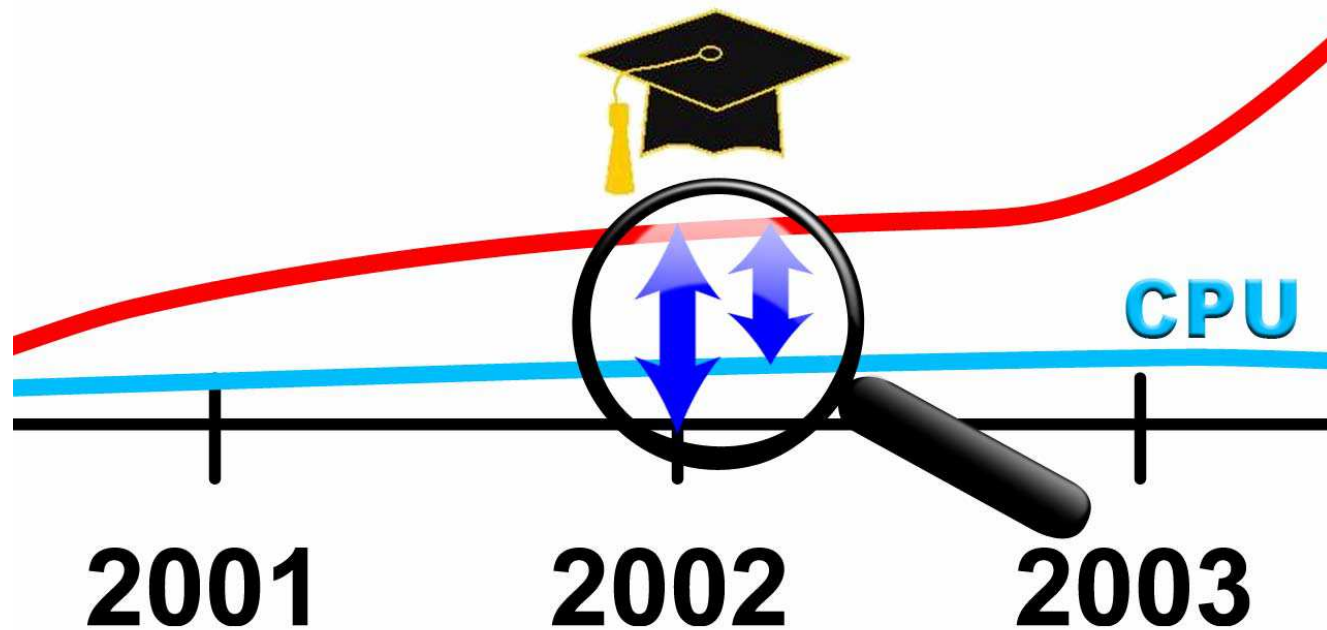


**2002:**

*The Graphics Processor Extends Beyond Gaming*

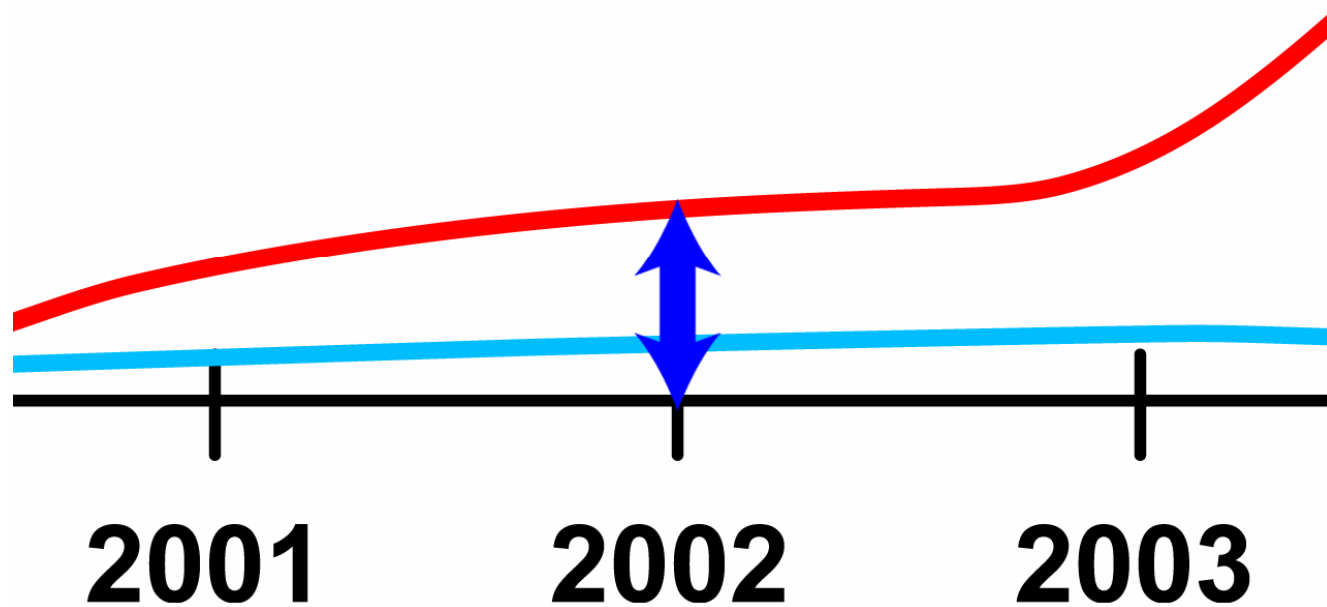
**Performance**

**Programmability**

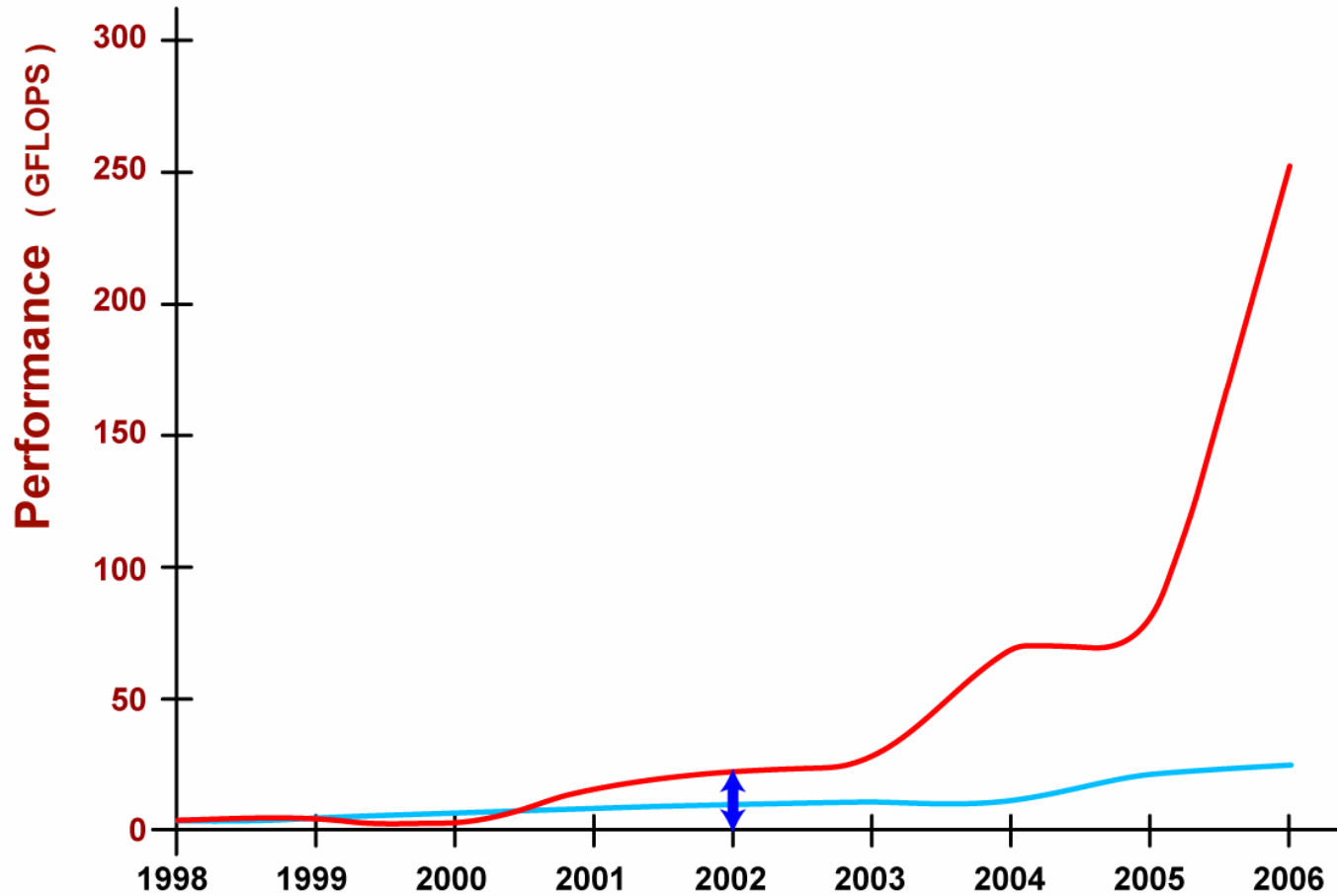


# 2002:

## *The Graphics Processor Extends Beyond Gaming*



# GPU Performance = End of the CPU?





# Oil and Gas Demo

(Removed in PDF version to reduce file size)

# Protein Folding Demo

(Removed in PDF version to reduce file size)

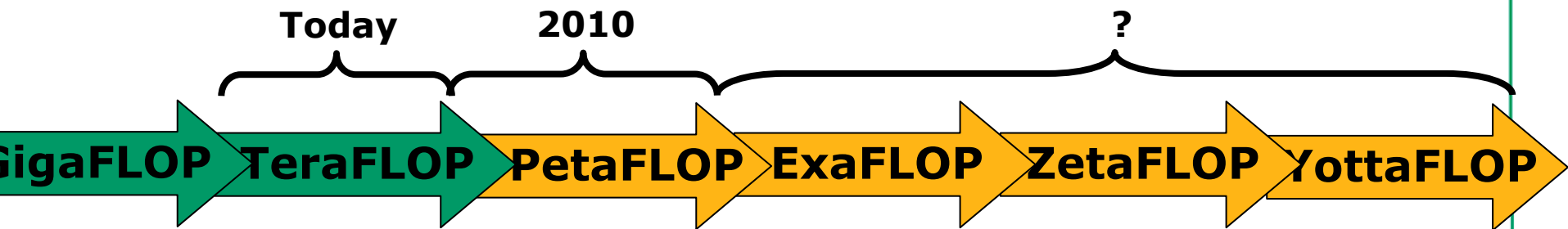
# Petascale Processing for the Masses

Over half of the  
top 500  
supercomputers  
today use over  
1000 processors\*

48 GPU Pipes  
x 8 Flops/cycle  
x 3 GHz

**Power and bandwidth  
unconstrained**

-----  
FLOPS  
FLOP per socket  
x1000 = 1 PetaFLOP

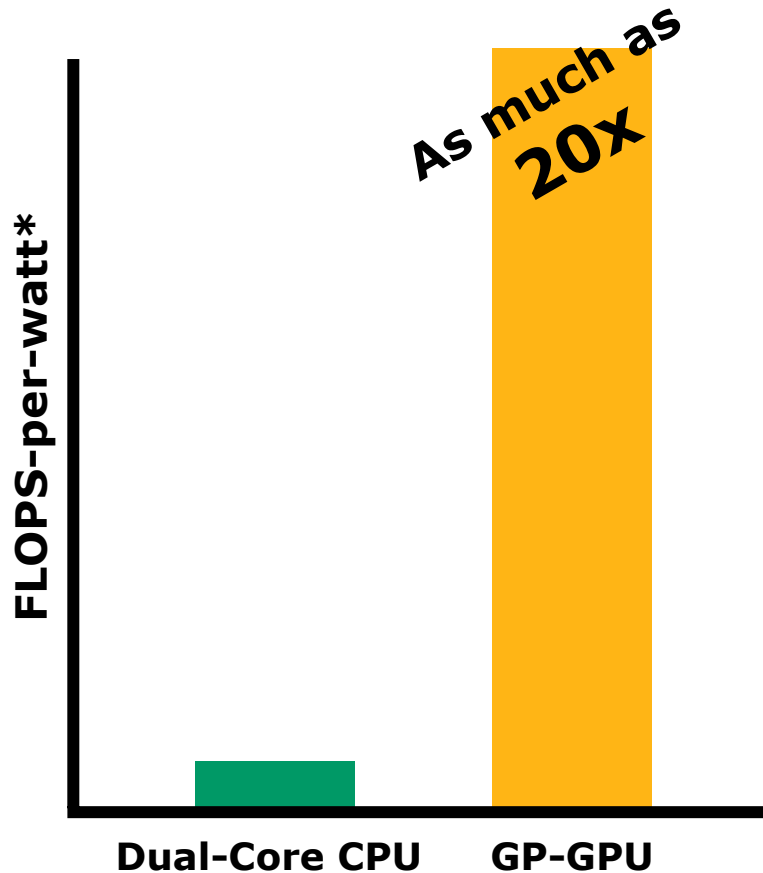


\* Source: top500.org

“Ask.com operations VP Dayne Sampson estimates that the five leading search companies together have some 2 million servers, each shedding 300 watts of heat annually, a total of 600 megawatts ... the total of electricity consumed by major search engines in 2006 approaches 5 gigawatts.”

— *Wired*, October 2006

# Realities of GP-GPU Power Efficiency



**1 TeraFLOPS in a CrossFire configuration**

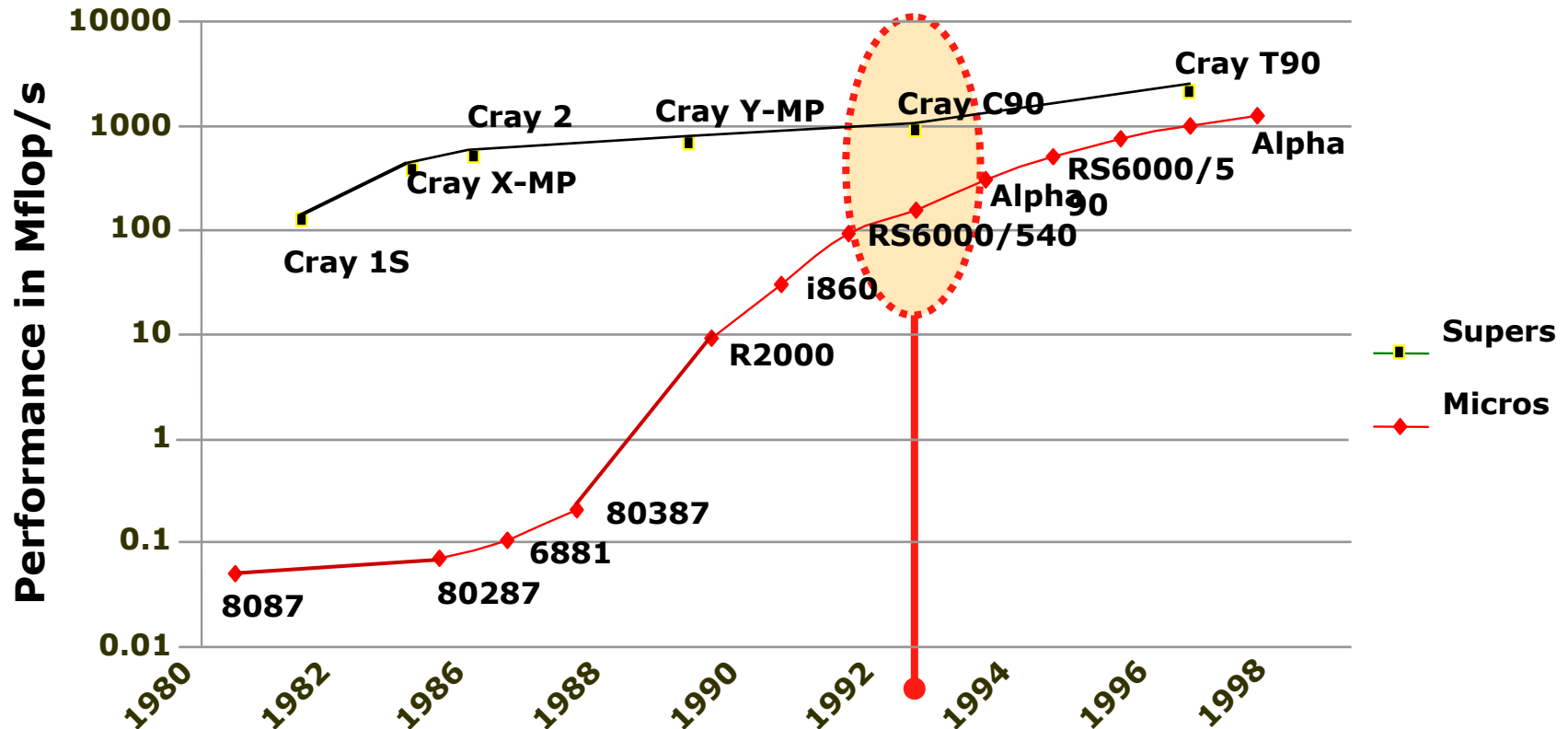
**500 GigaFLOPS per GPU**

**Available today – not just theoretical**

**More than 2 GigaFLOPS-per-watt**

**Generalized GPU provides unprecedented opportunity for performance-per-watt**

# HPC: What can be Learned from Attack of the Killer Micros

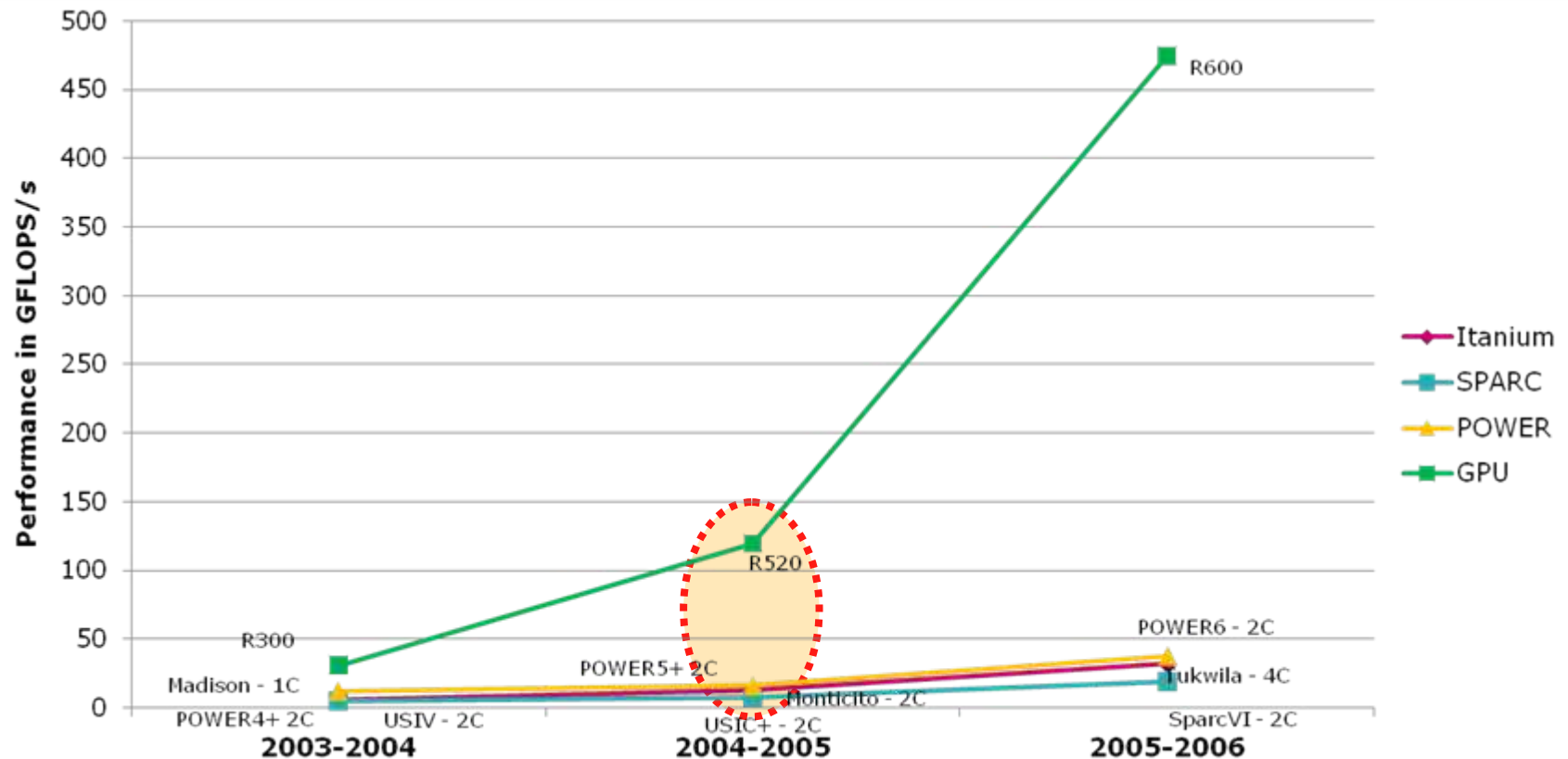


**1/10<sup>th</sup> the performance, but at 1/100<sup>th</sup> the cost**  
**Absolute performance "good enough"**

**Productivity now greater on a micro\* than on a super**

**\* \$10K - \$100K Workstation**

# History Repeating Itself?



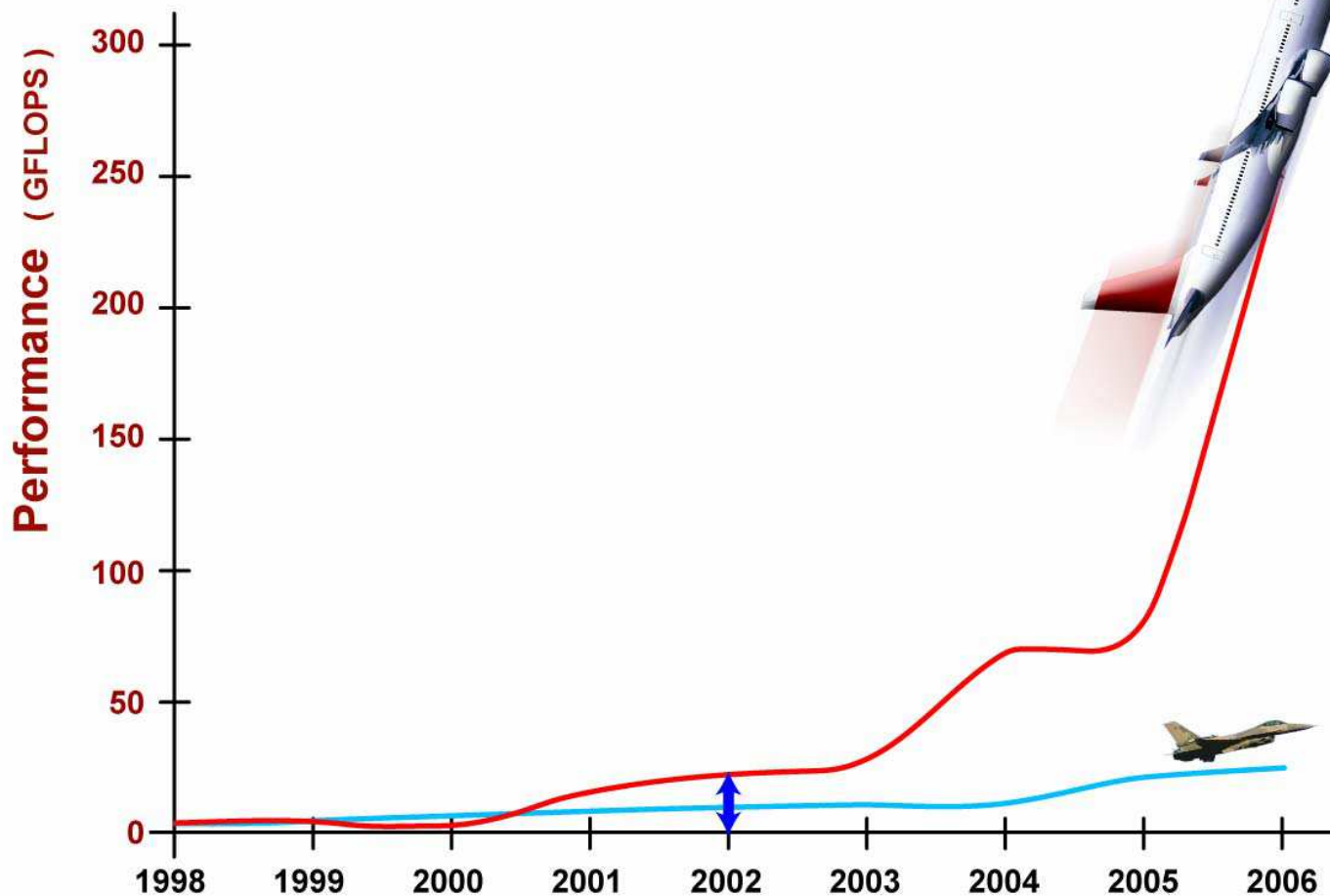
**Traditional “computing” is an order of magnitude behind  
Supercomputing programming model is back  
\$1K - \$5K PCs get amazing computational power via GPU**





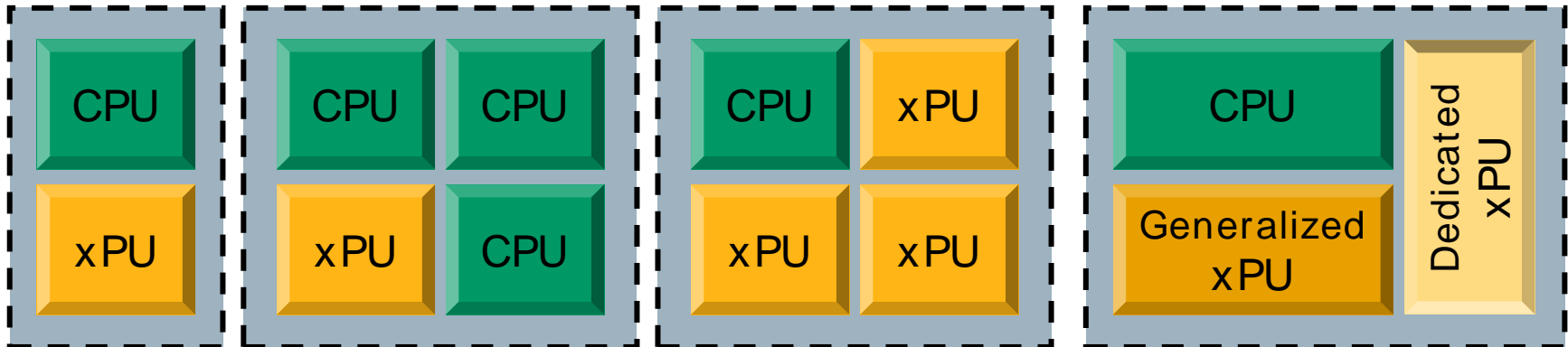
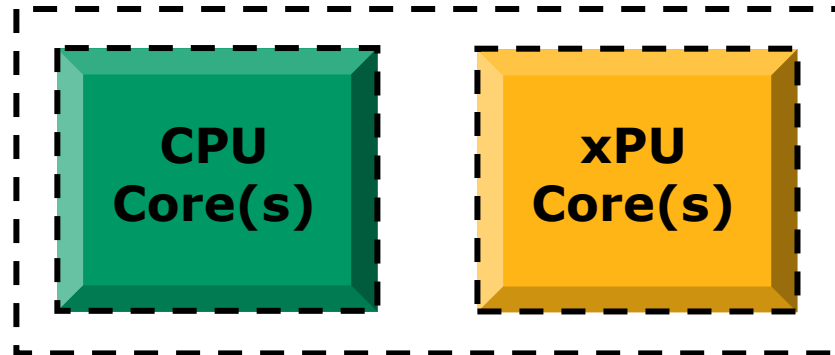
# GPU Performance = End of the CPU? **NO!**

*Amdahl's Law is Alive and Well..*



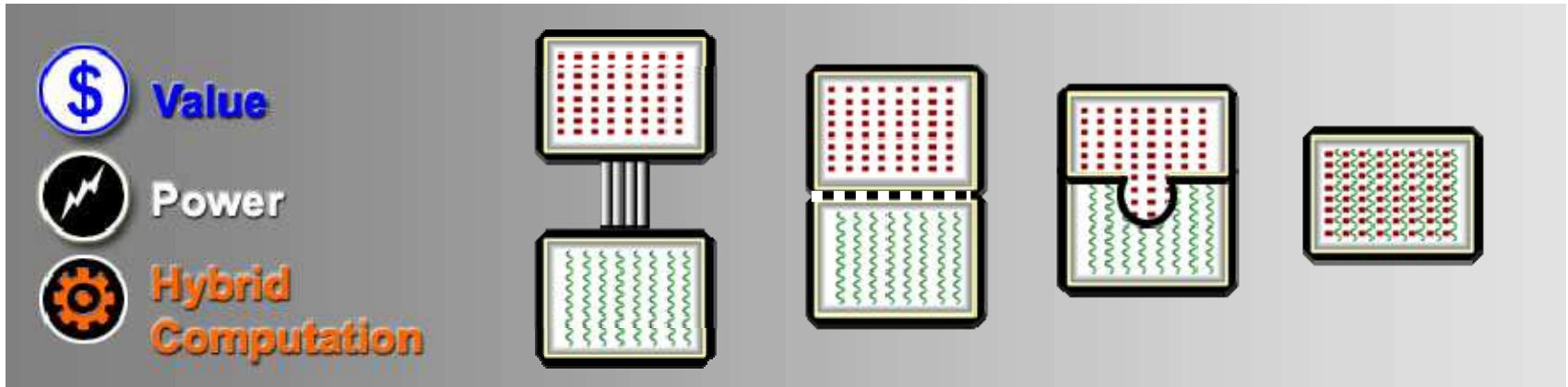
# A New Product Category for a New Era

## Accelerated Processing Units

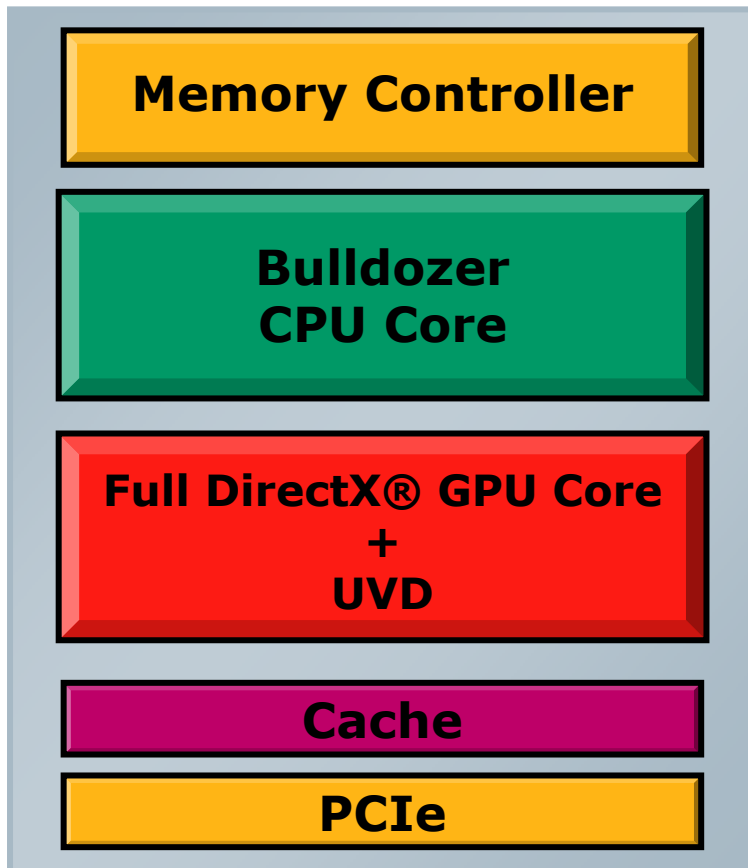


**New category defined by a combination of CPU and non-CPU cores integrated on silicon**

# APU Evolution for Increased Benefit with Minimized Disruption



# First Client APU Implementation



**Optimized for mobile and mainstream desktops**

**Next-generation x86 64-bit core**

**Integrated unified shading architecture, DirectX GPU Core**

**Non-disruptive infrastructure evolution**

**The first APU configuration integrates CPU and GPU cores, yet retains x86 compatibility**



## Summary

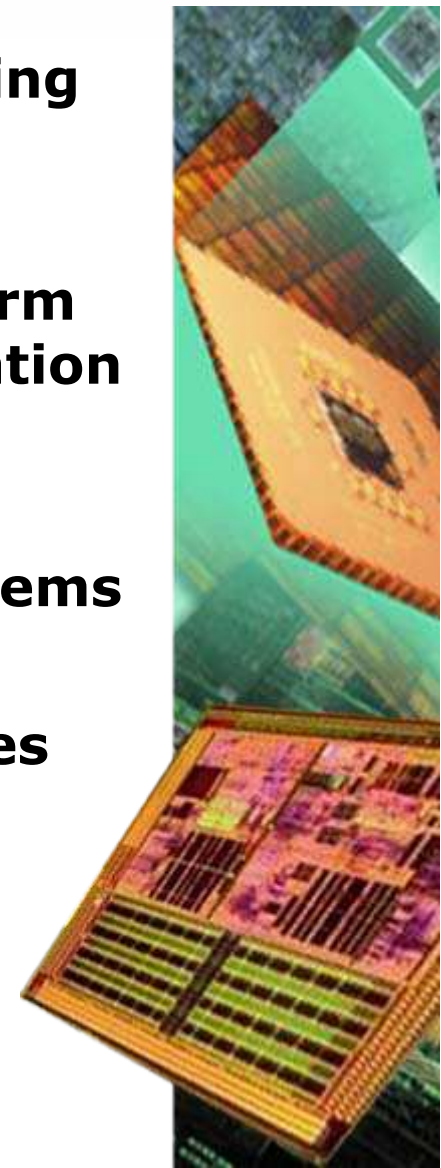
**Rapid diversification of x86 workloads driving the birth of the Accelerated Computing era**

**TeraFLOPS compute in an x86-based platform demonstrated *today* using GP-GPU acceleration**

**Stage set for peta-scale x86 using platform level acceleration in massively parallel systems**

**Increasing generalization of GPU capabilities through silicon-level integration ...**

**... and new software API's enable supercomputing for the masses**



## Trademark Attribution

AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Windows is a registered trademark of Microsoft Corporation in the United States and/or other jurisdictions. Linux is a registered trademark of Linus Torvalds.

Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.

© 2007 Advanced Micro Devices, Inc. All rights reserved.