AMD

Smarter Choice

May 17, 2006

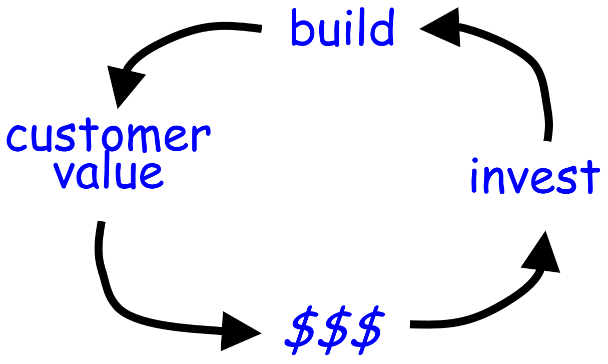
The Role of Accelerated Computing in the Multi-Core Era

Chuck Moore
Senior Fellow
Advanced Micro Devices

Key Points in this Talk

1. The semiconductor industry is dependent upon ongoing customer value:

A virtuous cycle:



build

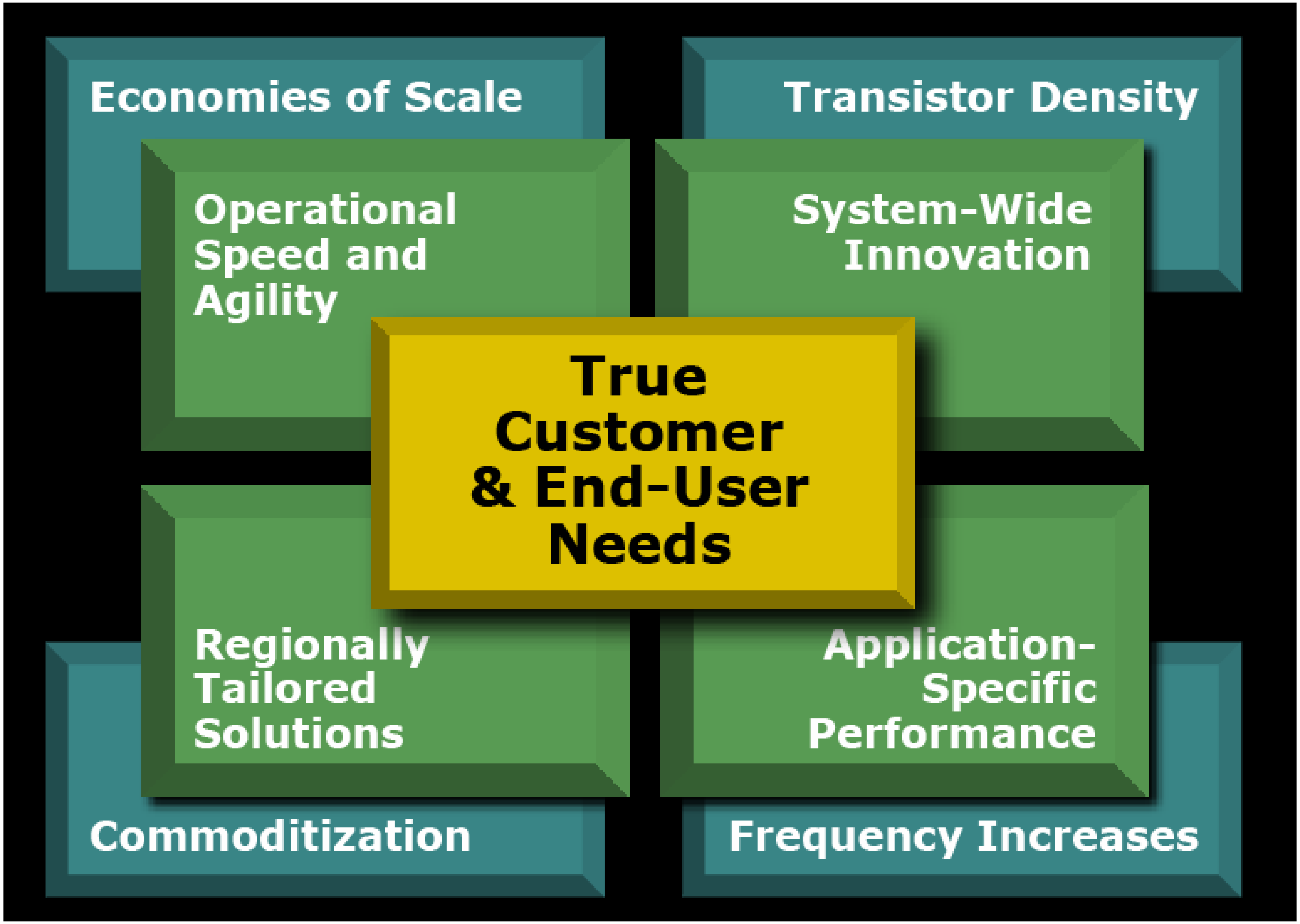
customer value

invest

$$$

2. Programming for Multi-Core is a difficult challenge, but it is really just the leading edge of the bigger challenges yet to come

Our industry is obsessed with Performance



Economies of Scale

Transistor Density

Operational Speed and Agility

System-Wide Innovation

True Customer & End-User Needs

Regionally Tailored Solutions

Application-Specific Performance

Commoditization

Frequency Increases

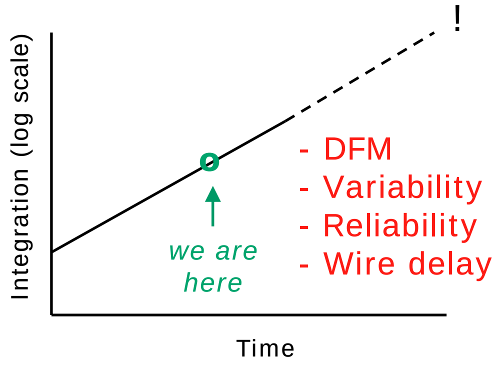
It's Time to Reorient Around Customer Value

Outline

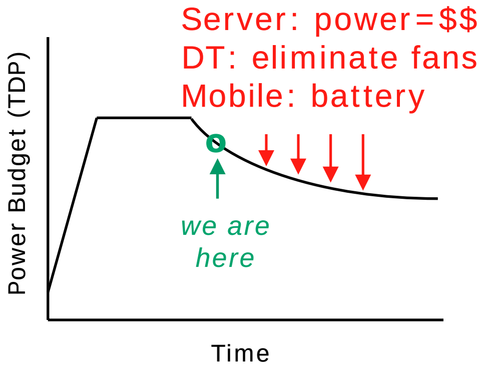
- Important Background
  - A Few High-level Trends
  - Some Thoughts on SMP and Multi-core Computing
- The Accelerated Computing Imperative
  - Dense Computing: GPUs and GP-GPUs
  - The broader potential
- A Framework for Accelerated Computing enablement
  - The Role of Architecture
  - The Emerging Layers of Computation
- Summary

A Few High-level Trends

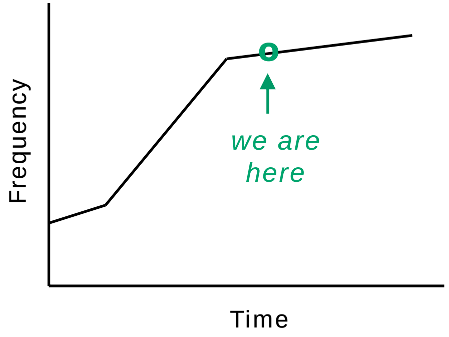
Moore's Law ☺



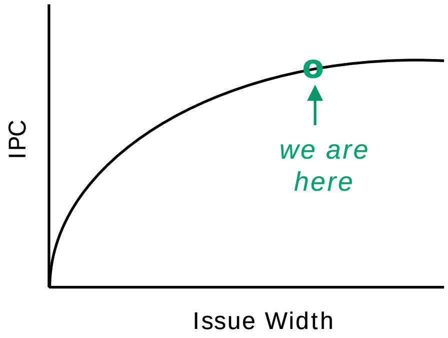
The Power Wall ☹



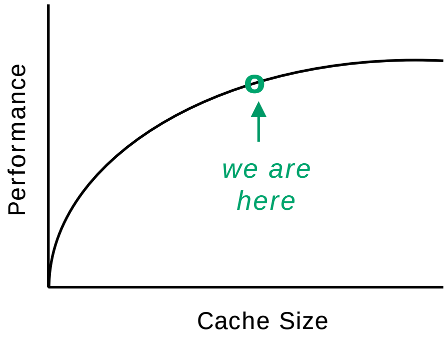
The Frequency Wall ☹



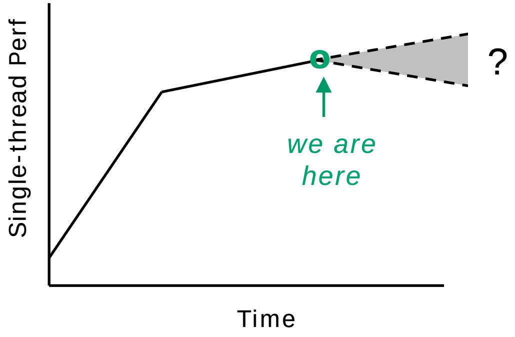
The Complexity Wall ☹



Locality 😐

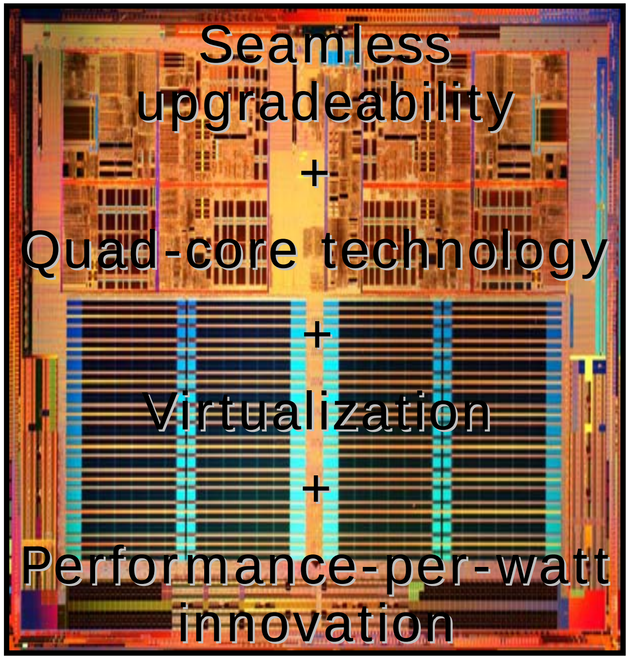


Single thread Perf (!)



So, how can we add customer value?

Customer Value beyond just Performance



Seamless upgradeability

+

Quad-core technology

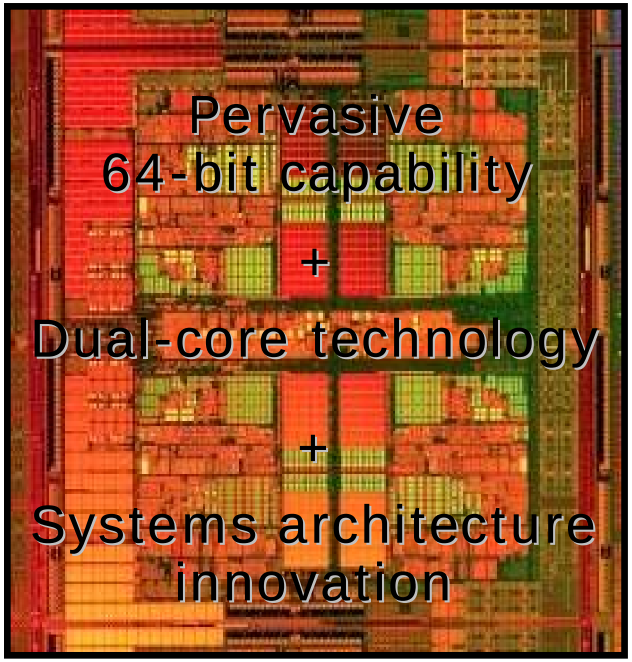
+

Virtualization

+

Performance-per-watt innovation

AMD Native Dual Core Opteron



Pervasive 64-bit capability

+

Dual-core technology

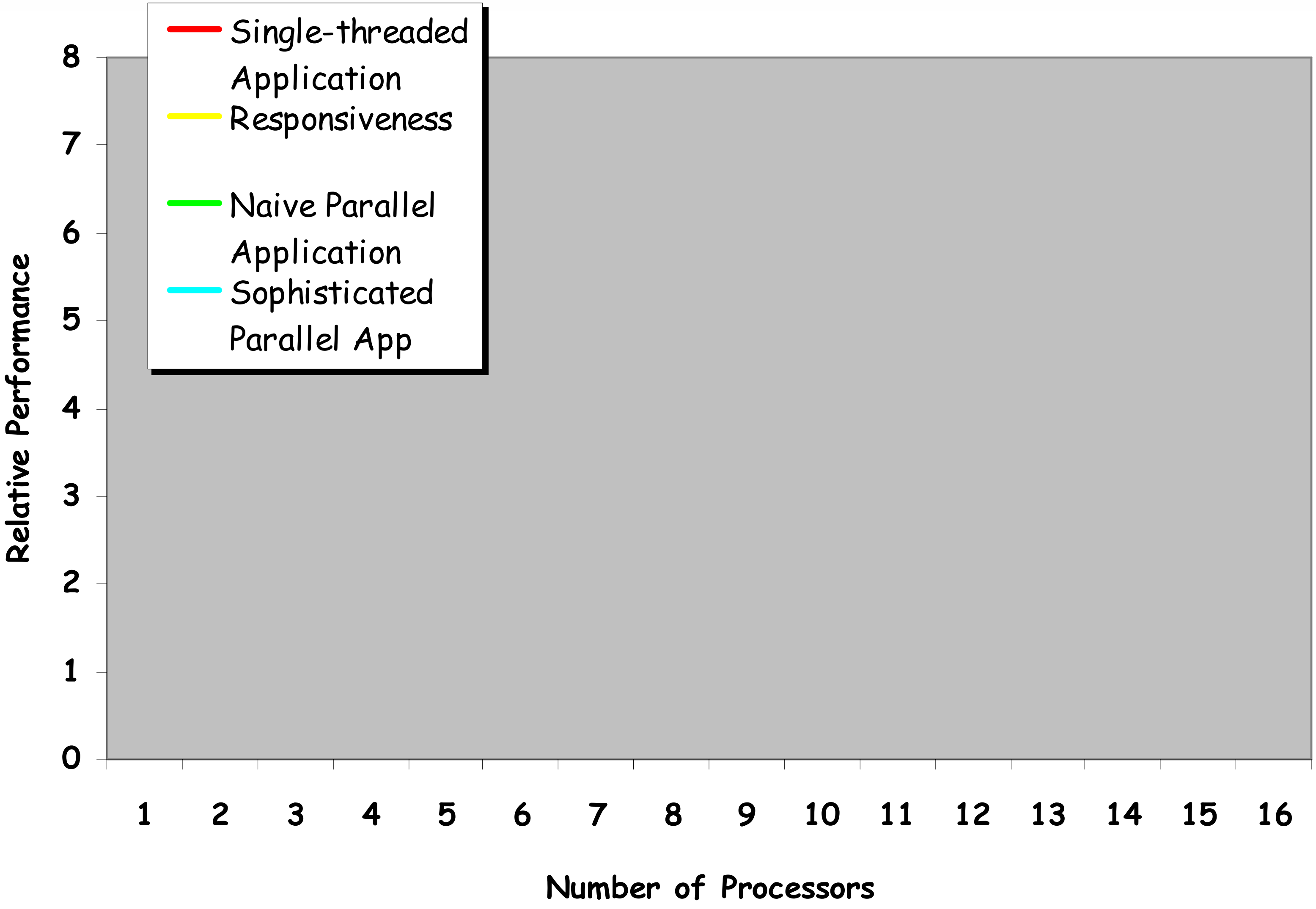
+

Systems architecture innovation

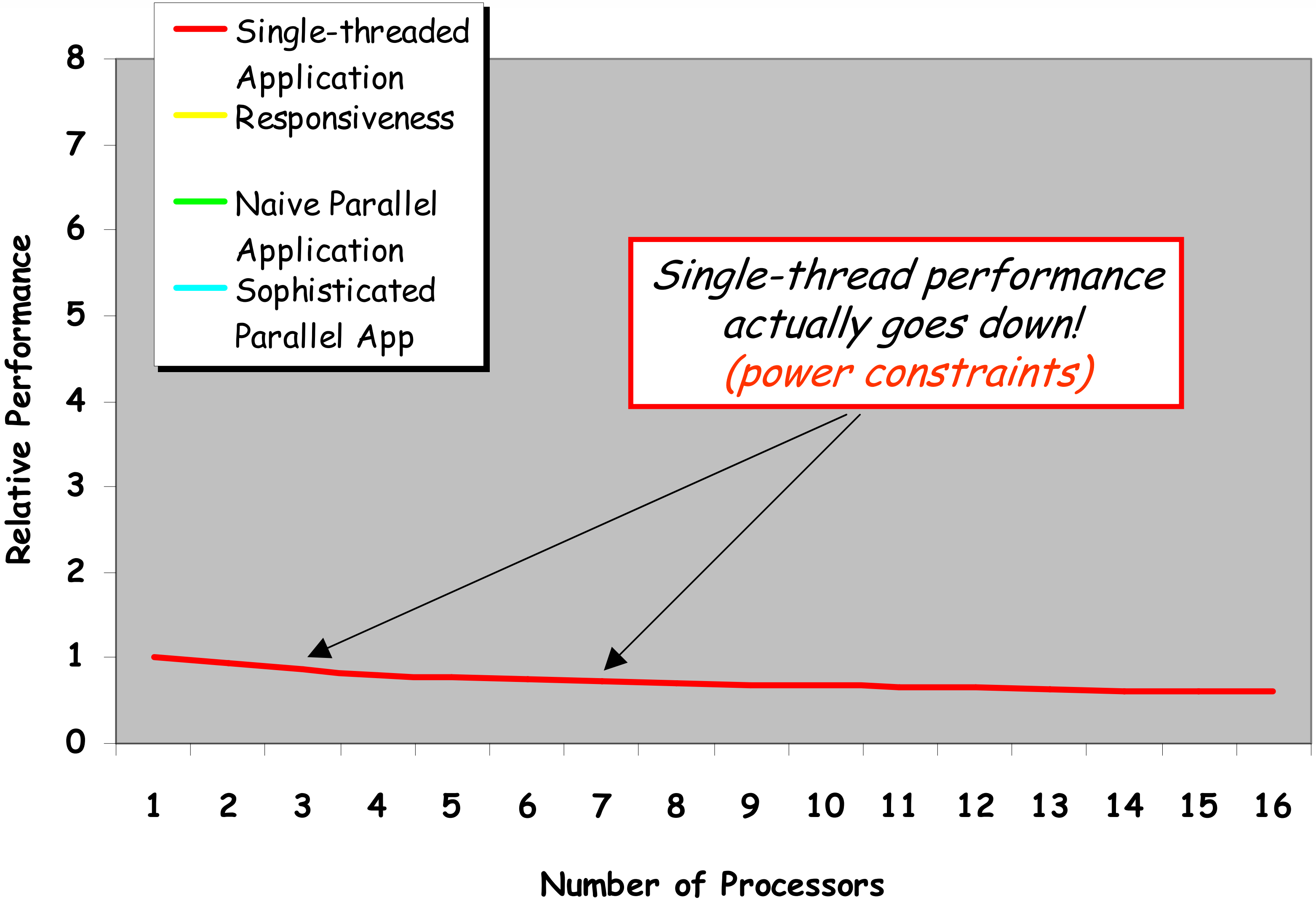
AMD Native Quad Core Core Opteron

SMP and Multi-Core to the long term rescue?

SMP Performance *(Hypothetical values)*

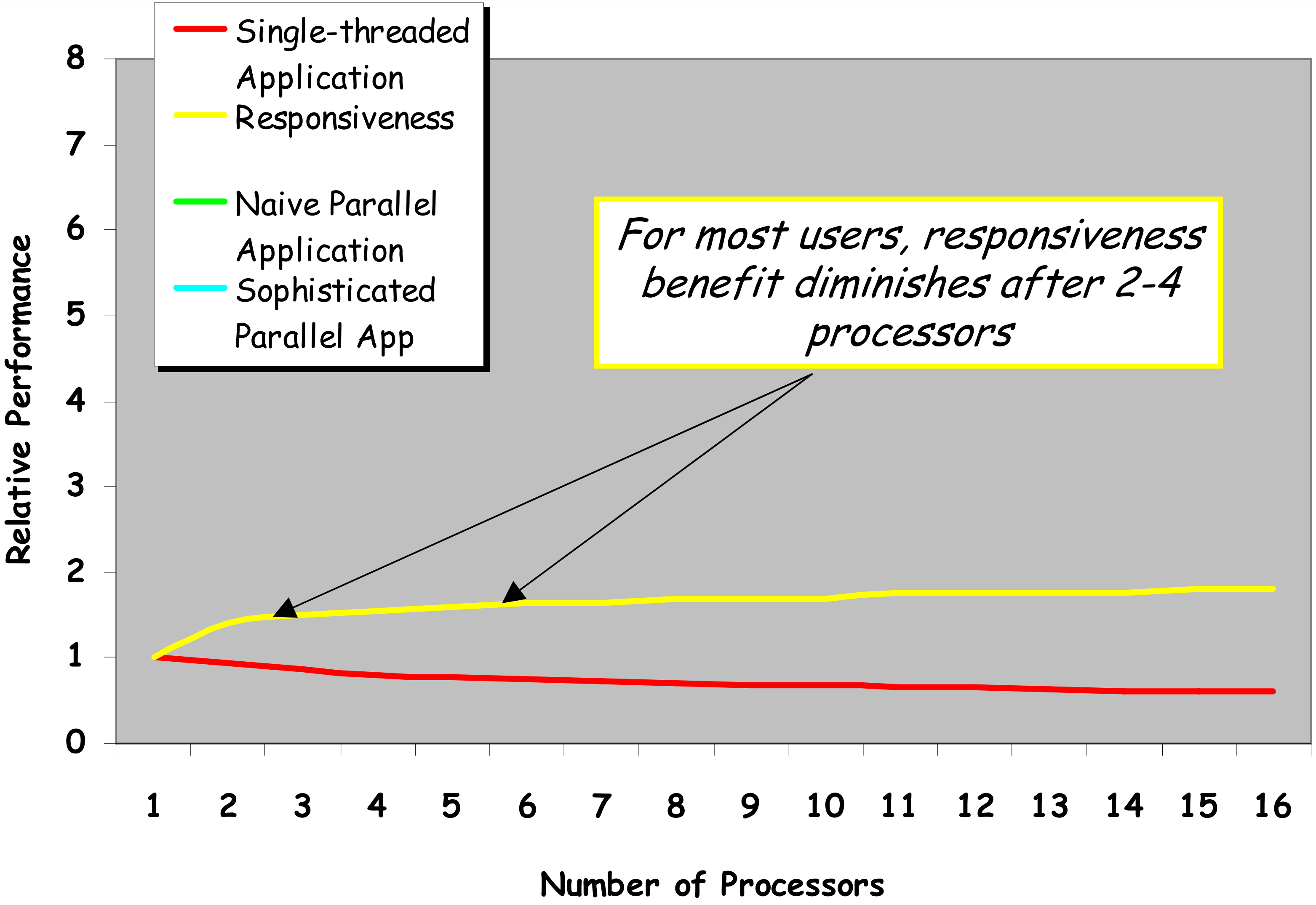


SMP Performance *(Hypothetical values)*



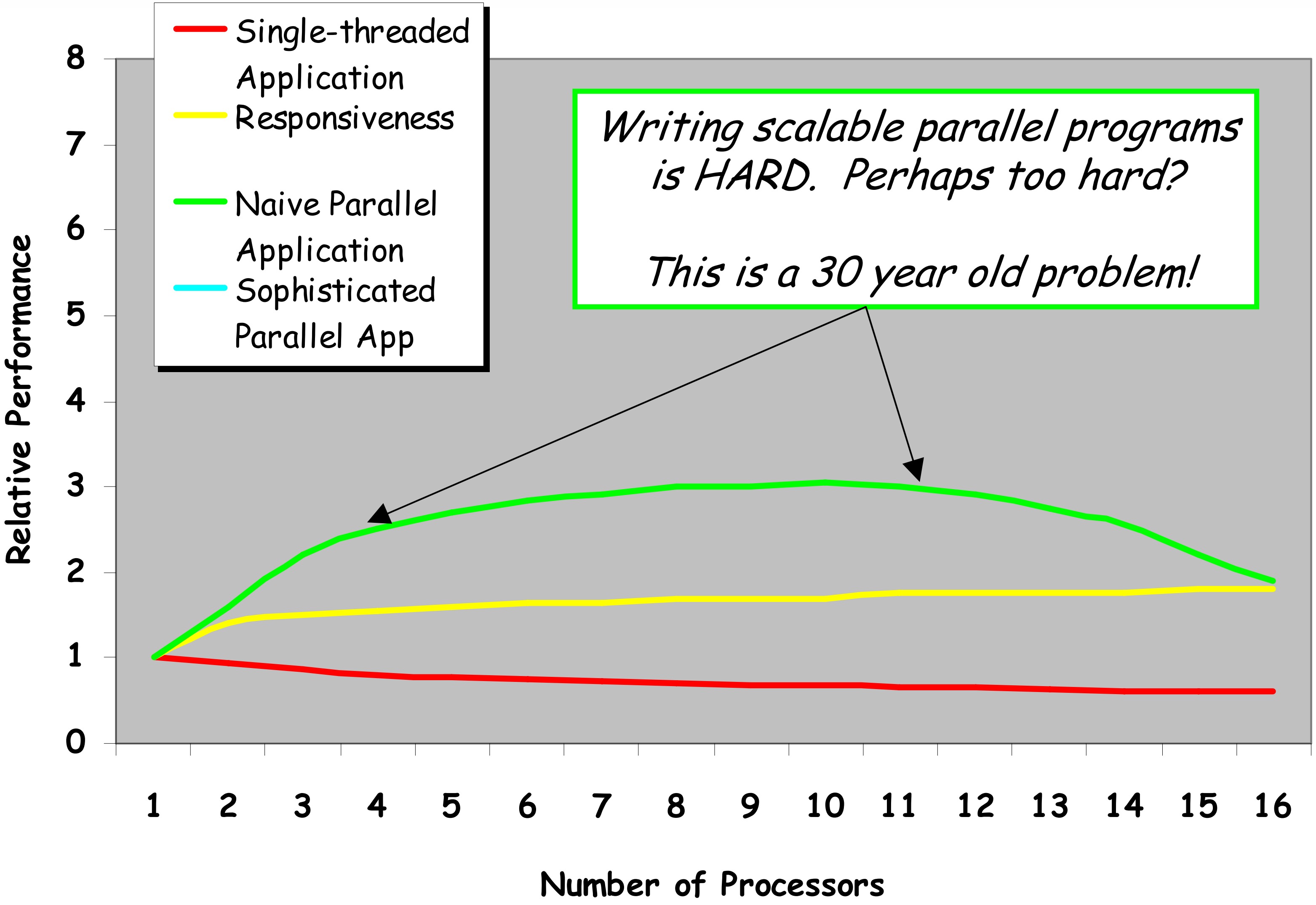
Single-thread performance actually goes down!
(power constraints)

SMP Performance *(Hypothetical values)*



For most users, responsiveness benefit diminishes after 2-4 processors

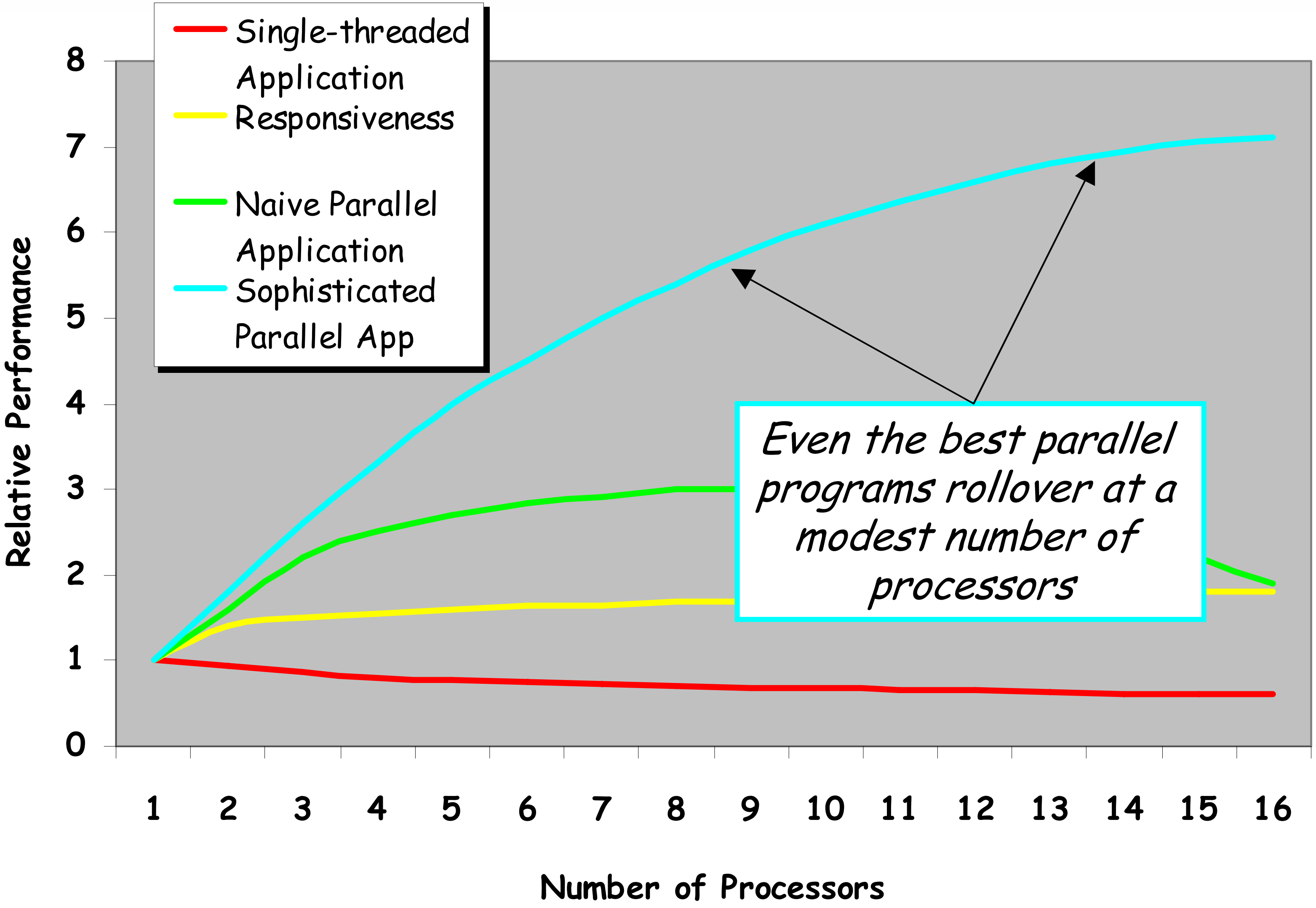
SMP Performance *(Hypothetical values)*



Writing scalable parallel programs is HARD. Perhaps too hard?

This is a 30 year old problem!

SMP Performance *(Hypothetical values)*



Even the best parallel programs rollover at a modest number of processors

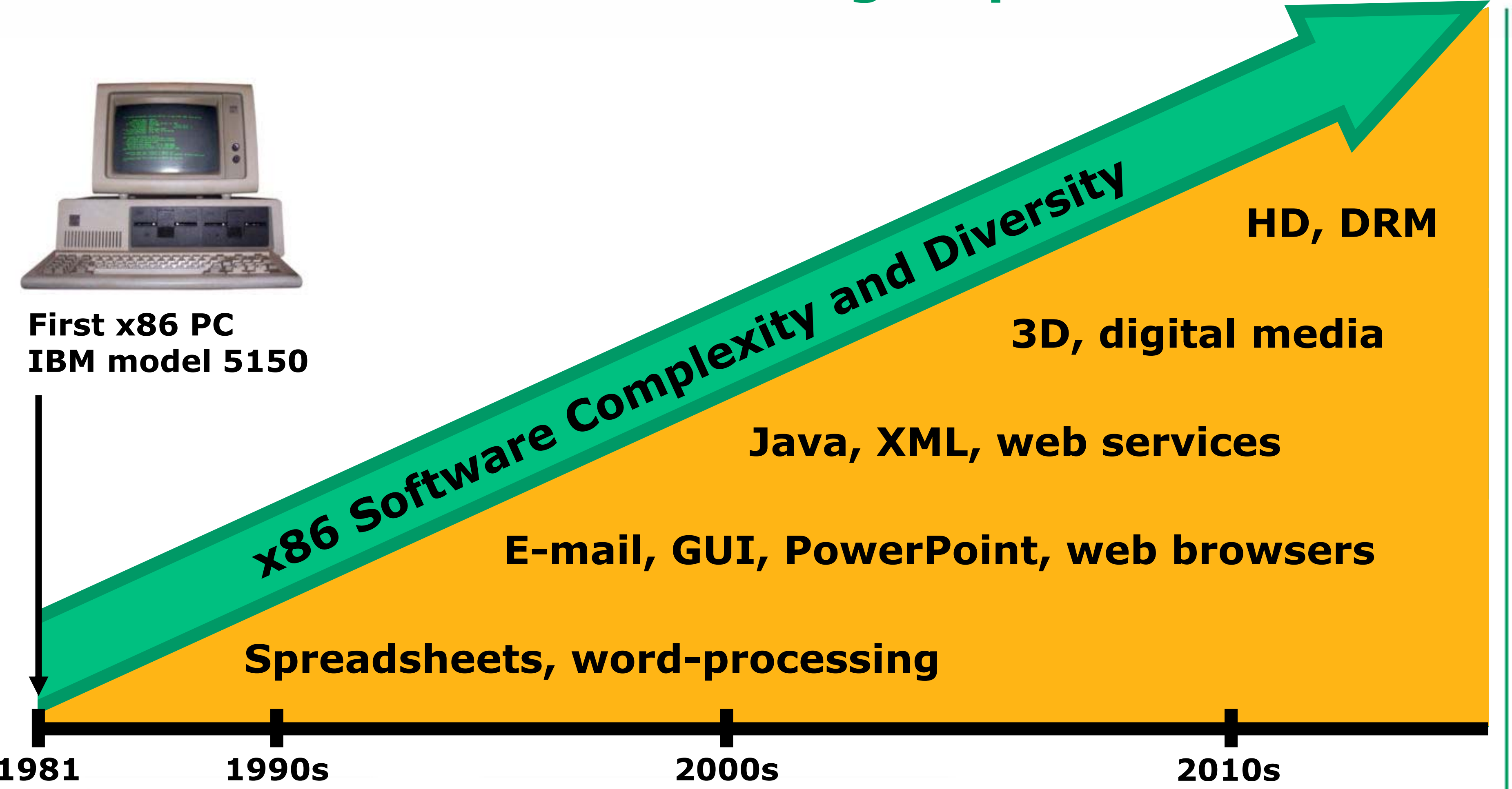
Optimized SMP and Multi-core Platforms

- In the near-term, there is definitely potential here
  - Commodity multi-core processors break the “chicken & egg” barrier
  - Impressive amount of interesting research firing up:
    - *TM, coherency filters, hierarchical scheduling, MREs, VMs, etc*
  - Lots of good activity on the Tools front → ***More to come***
- Some workloads will do well with this, but many will not:
  - As it turns out, software isn’t really that soft
    - *The underlying structural assumption is often serial processing*
    - *Transitioning the concurrency model is a very big deal*
  - Amdahl’s Law seriously inhibits unstructured parallelism
- In reality, SMP/Multi-core challenges are just an early indicator of the shifts yet to come
  - Power constraints will force these to be “performance heterogeneous”
  - Advances in synchronization and NUMA will give rise to new options…

Outline

- Important Background
  - A Few High-level Trends
  - Some Thoughts on SMP and Multi-core Computing
- The Accelerated Computing Imperative
  - Dense Computing: GPUs and GP-GPUs
  - The broader potential
- A Framework for Accelerated Computing enablement
  - The Role of Architecture
  - The Emerging Layers of Computation
- Summary

The Accelerated Processing Imperative

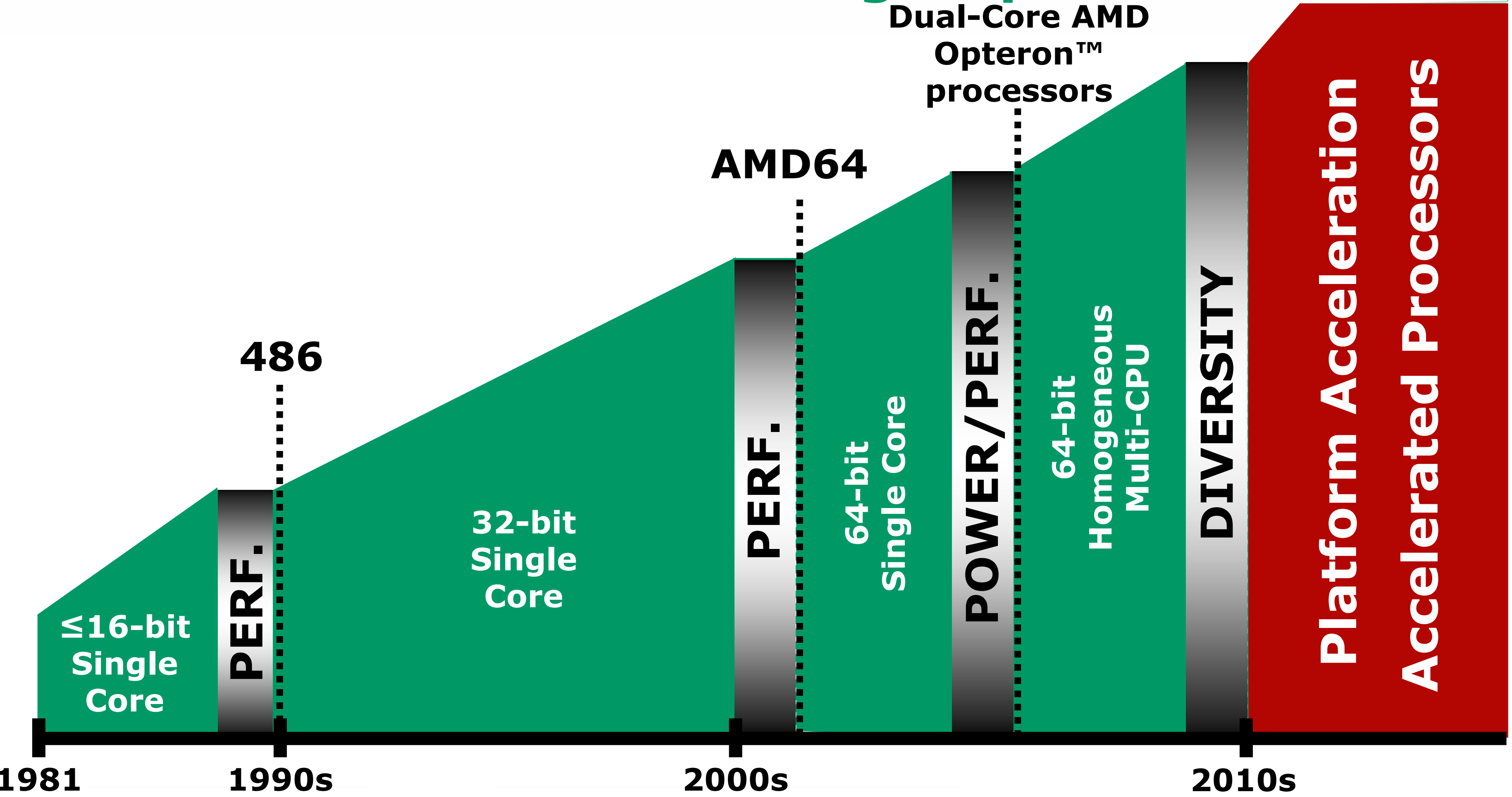


First x86 PC
IBM model 5150

x86 Software Complexity and Diversity
HD, DRM
3D, digital media
Java, XML, web services
E-mail, GUI, PowerPoint, web browsers
Spreadsheets, word-processing
1981
1990s
2000s
2010s

x86 applications, workloads and usage models continue to rapidly diversify

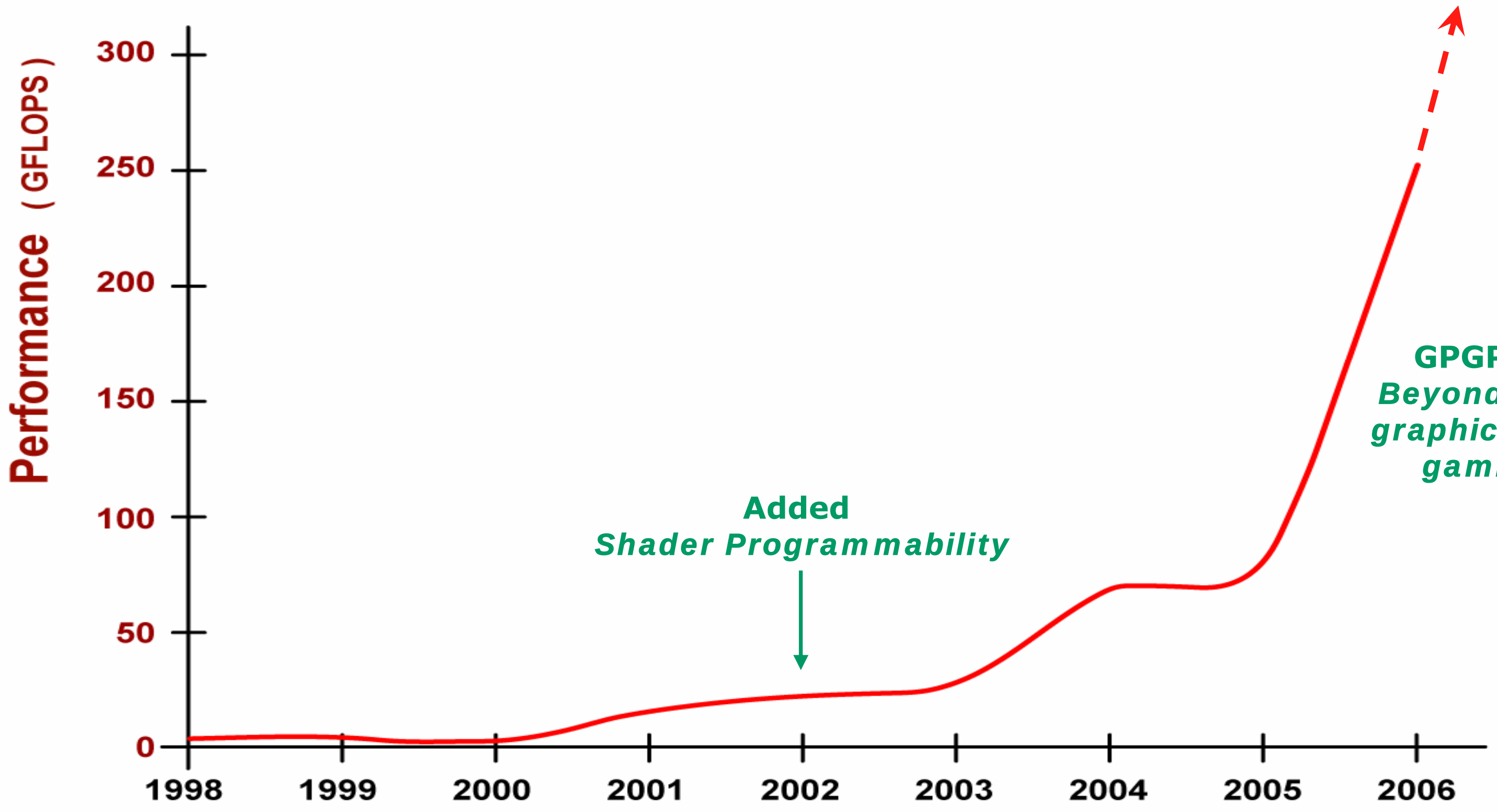
The Accelerated Processing Imperative



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

Compute Density:

Graphics Processor Performance ☺



Added
Shader Programmability

GPGPU:
Beyond just graphics and gaming

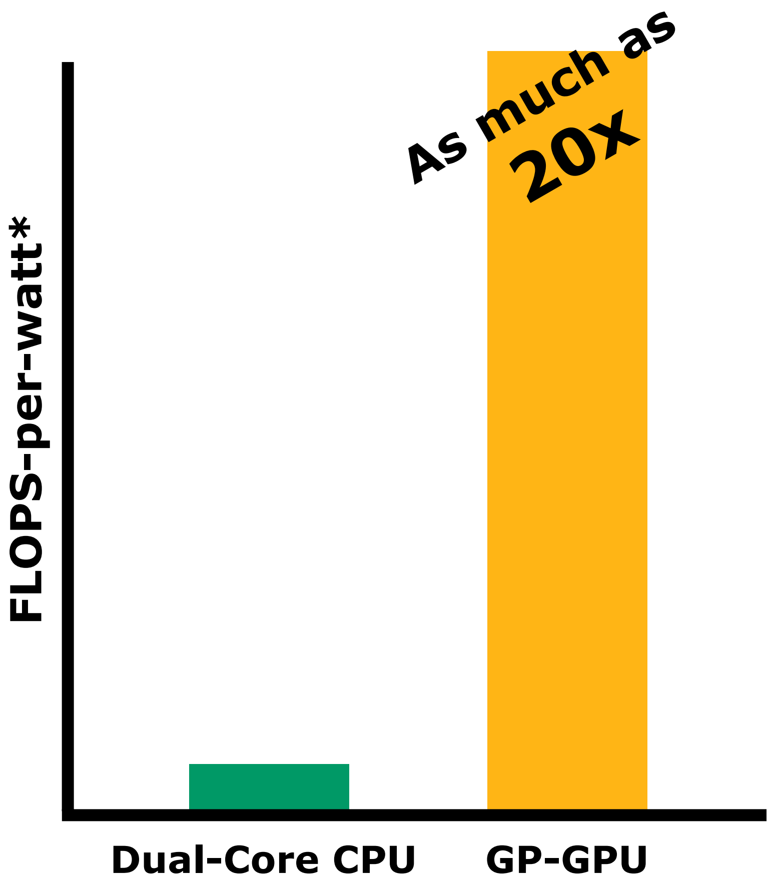
Ruby Statistics

	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
	2004	**2005**	**2006**

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
	2004	2005	2006

Realities of GP-GPU Power Efficiency



1 TeraFLOPS in a CrossFire configuration

500 GigaFLOPS per GPU

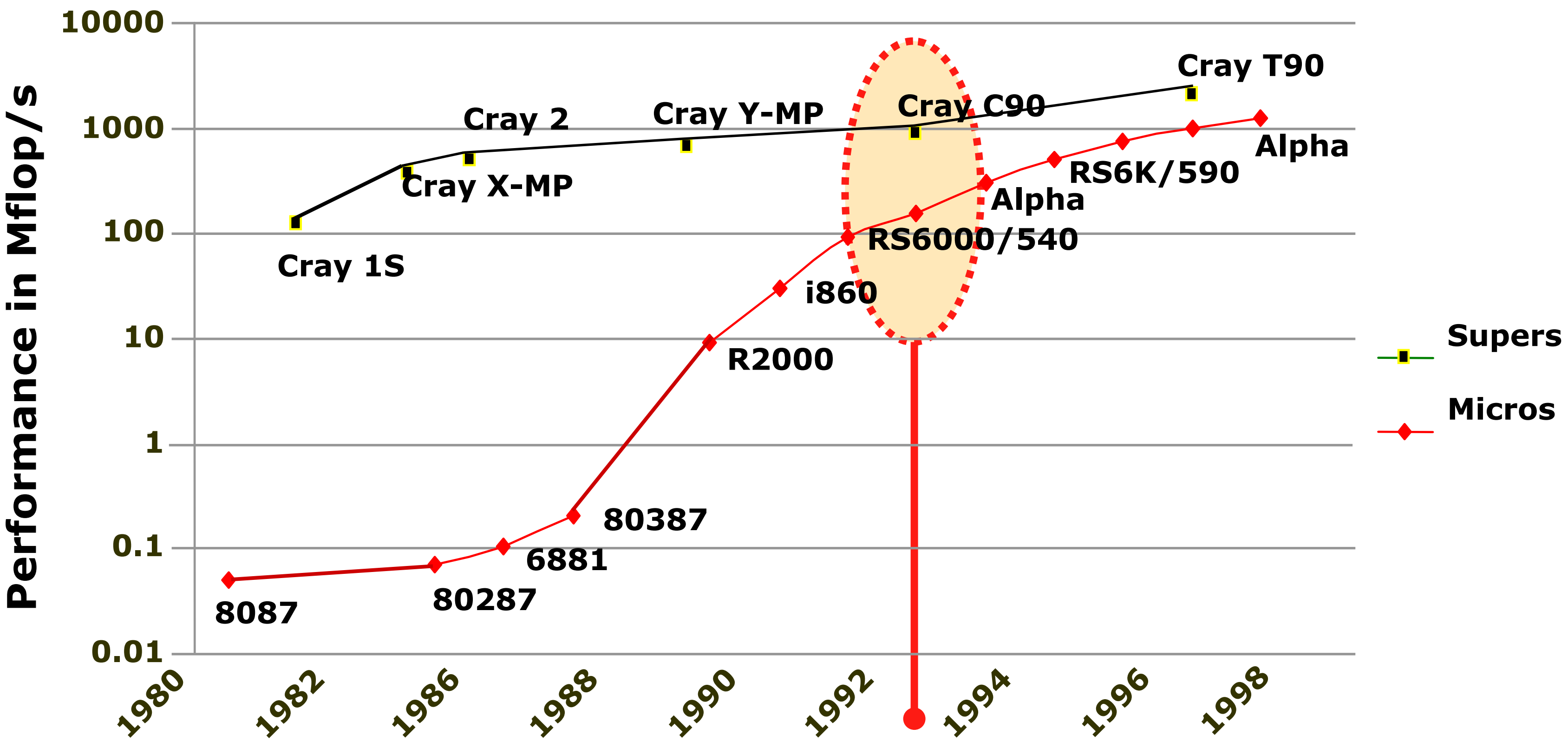
Available ***today*** – not just theoretical

More than 2 GigaFLOPS-per-watt

*Source: AMD

Generalized GPU provides unprecedented opportunity for performance-per-watt

HPC: Remember Attack of the Killer Micros?



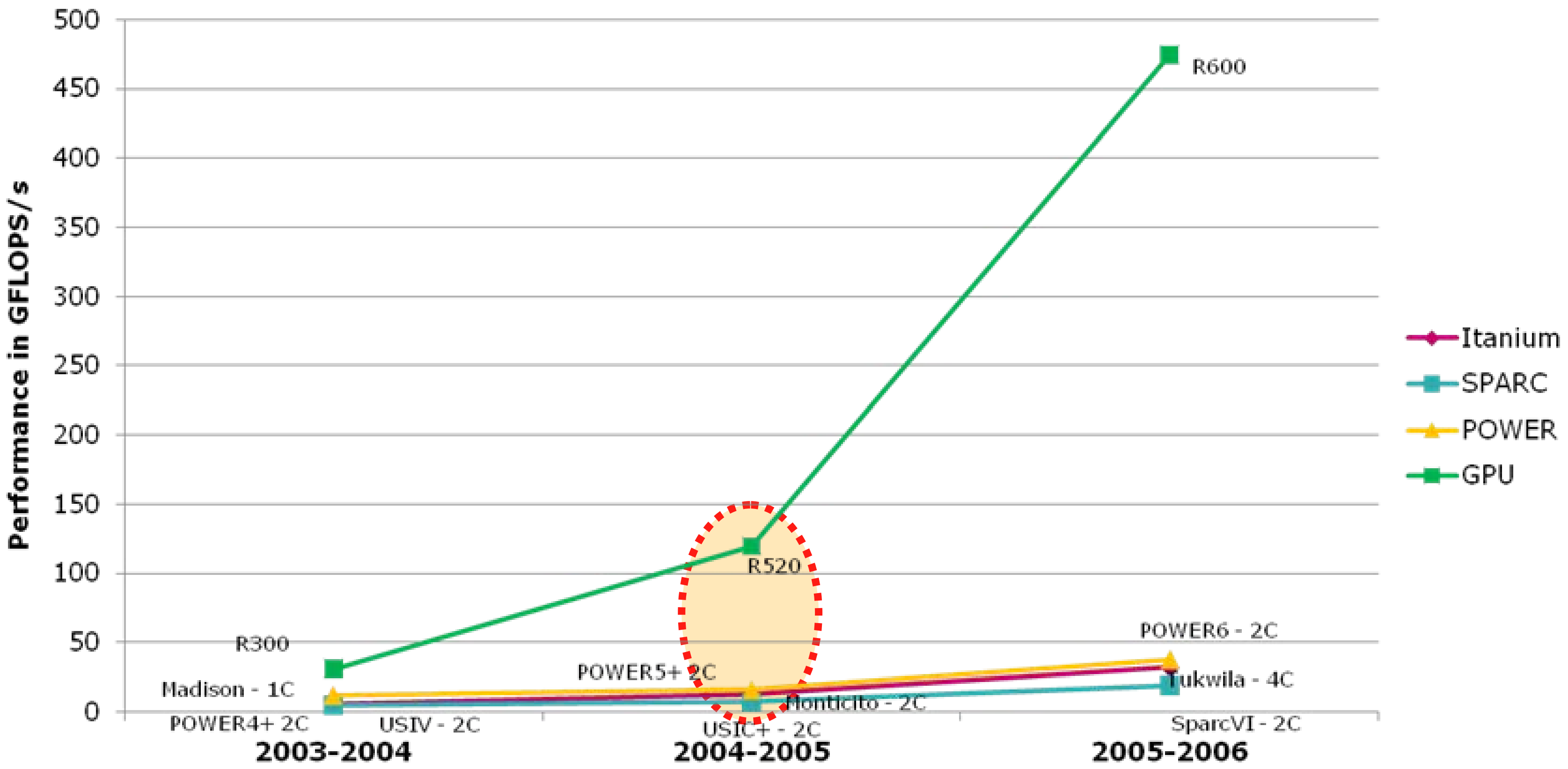
1/10th the performance, but at 1/100th the cost

Absolute performance "good enough"

Productivity greater on a workstation than on a super

Chart Source: Gordon Bell and Jim Gray, ISCA 2000

History Repeating Itself?

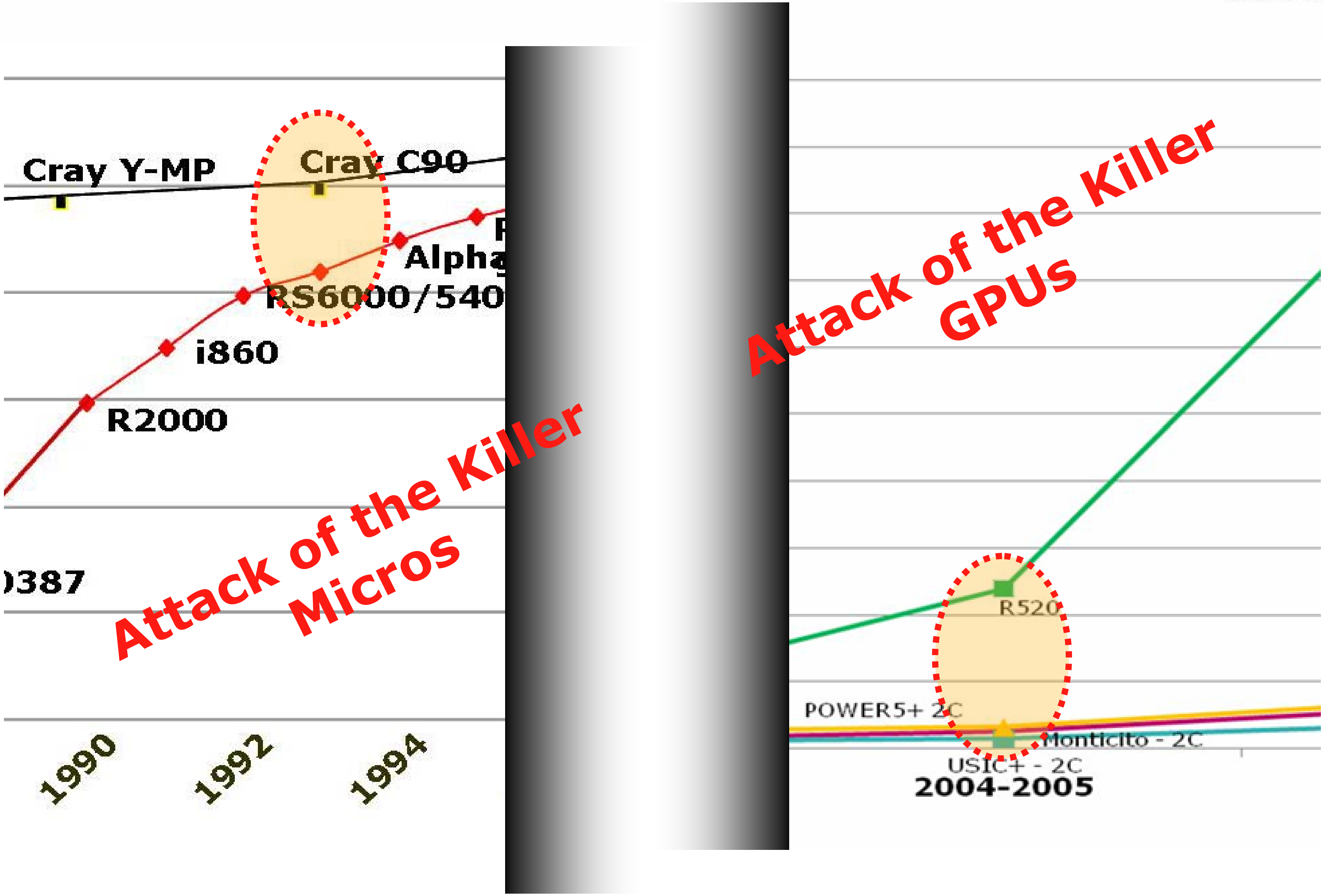


Traditional "computing" is an order of magnitude behind

Familiar vector-style programming model

$1K - $5K PCs get amazing computational power via GPU

You just can’t ignore this …

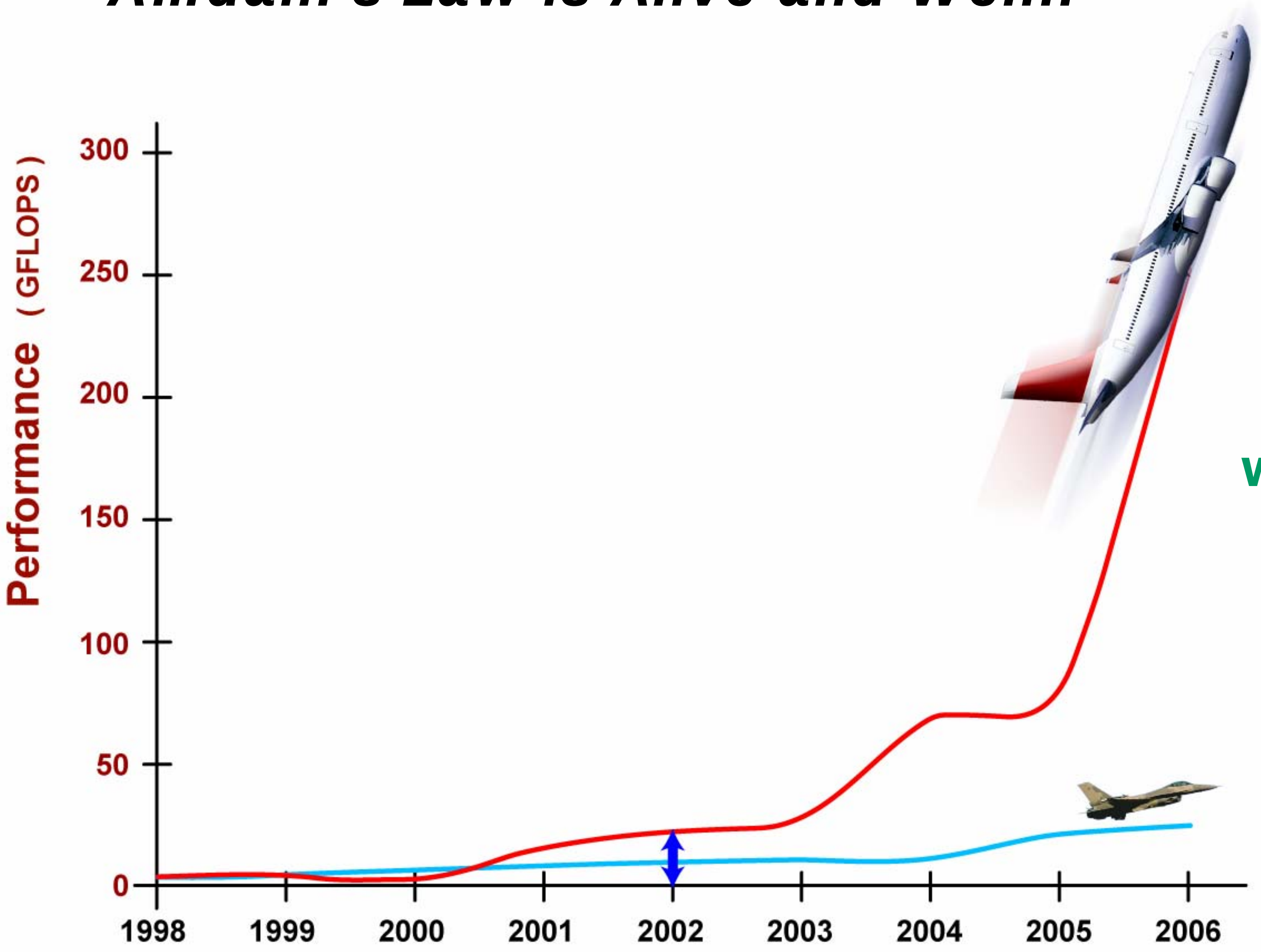


Attack of the Killer Micros

Attack of the Killer GPUs

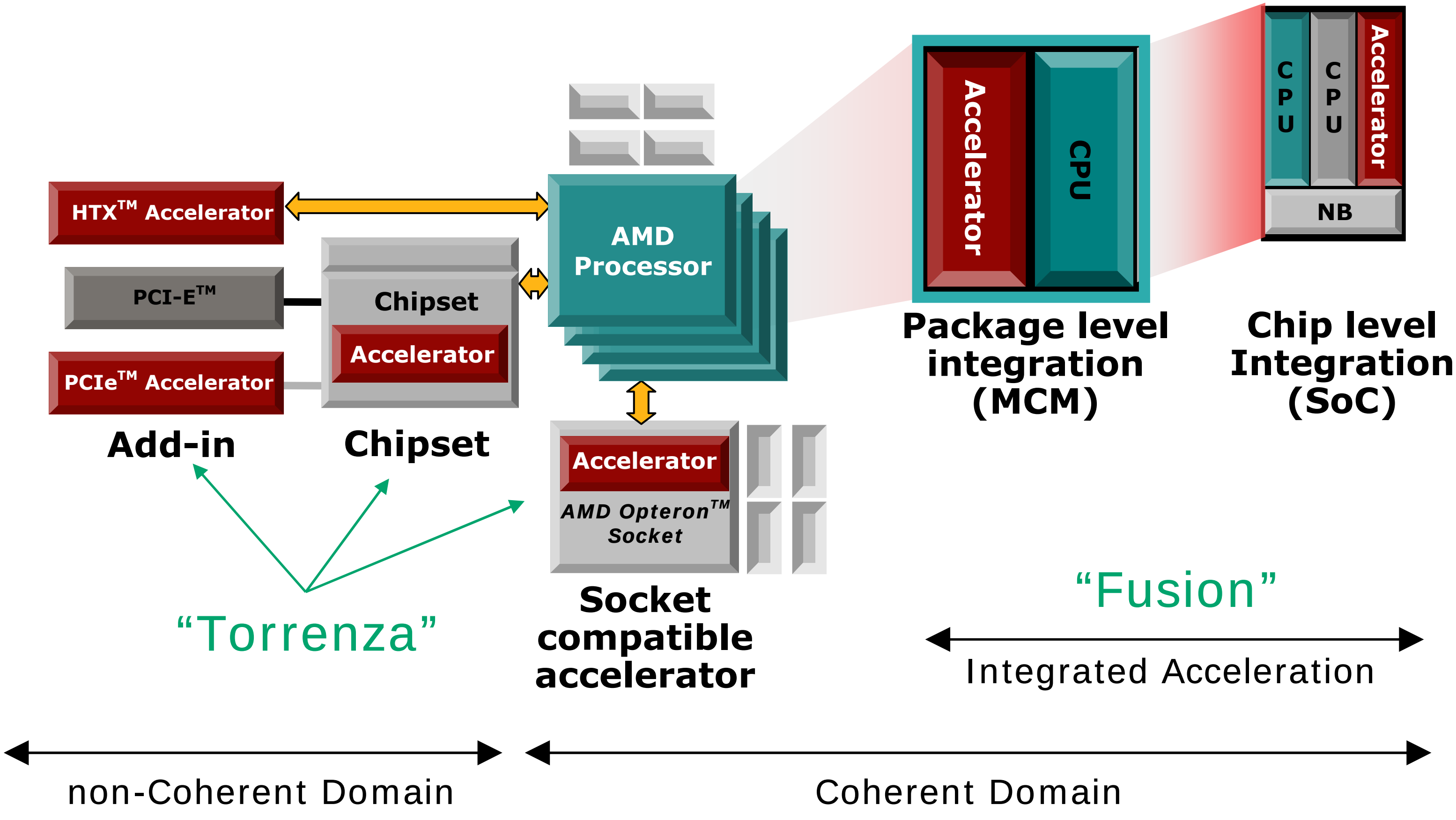
GPU Performance = End of the CPU? NO!

Amdahl's Law is Alive and Well..

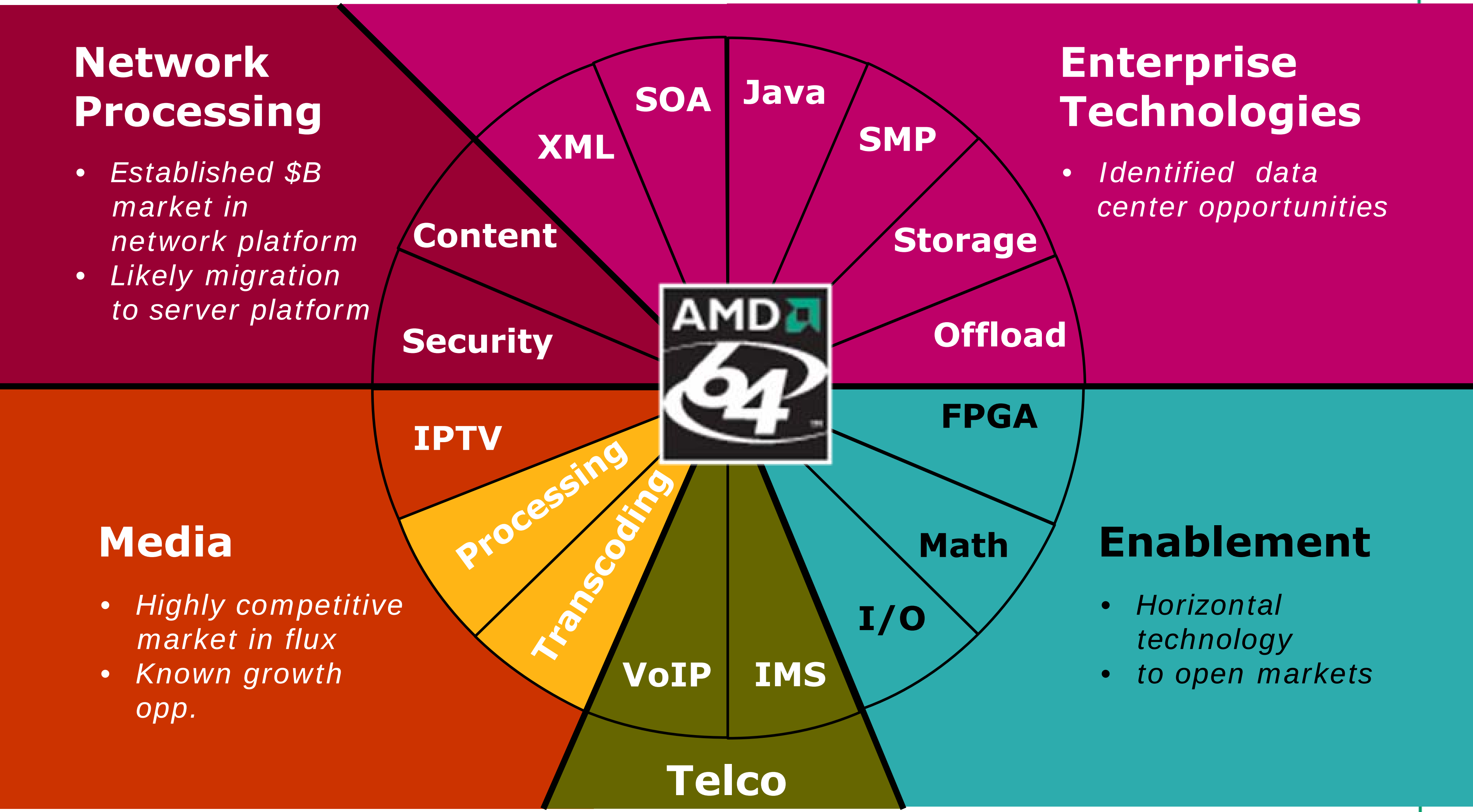


We need both!

Accelerated Computing has very broad potential -- *A Continuum of Solutions*



Torrenza: *Enabling Partners to Build on the Concept of Accelerated Computing*



Network Processing

- *Established $B market in network platform*
- *Likely migration to server platform*

Enterprise Technologies

- *Identified data center opportunities*

Media

- *Highly competitive market in flux*
- *Known growth opp.*

Enablement

- *Horizontal technology*
- *to open markets*

Telco

Outline

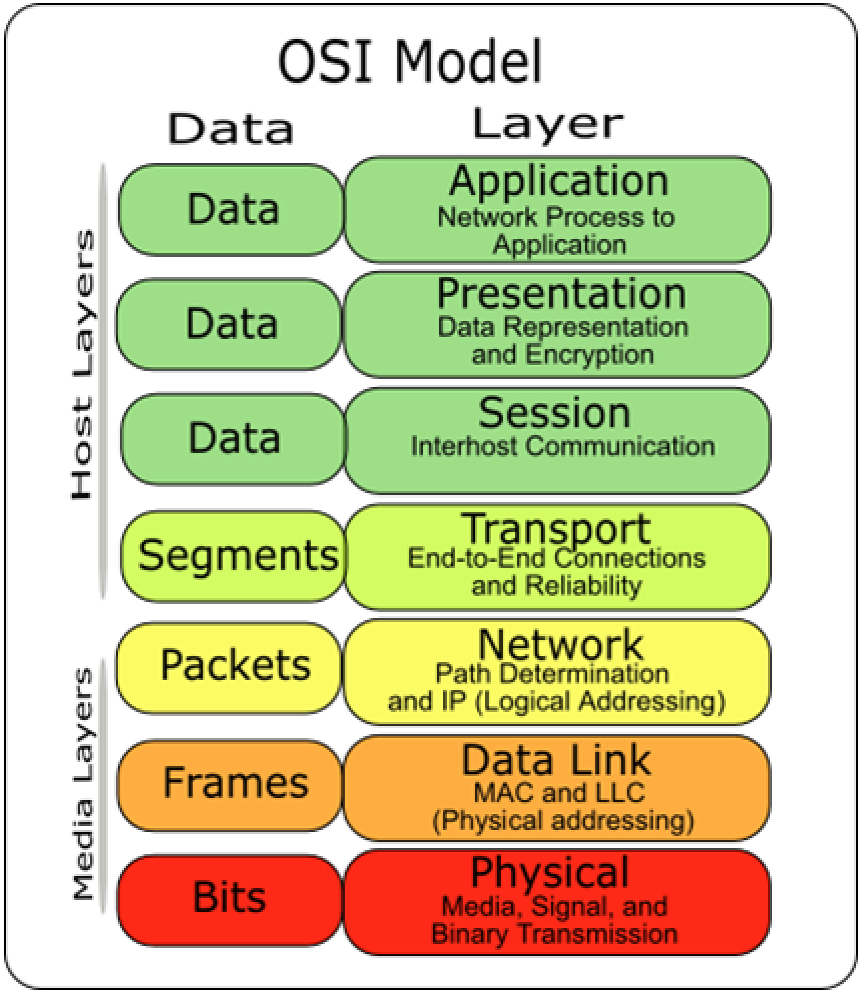
- Important Background
  - A Few High-level Trends
  - Some Thoughts on SMP and Multi-core Computing
- The Accelerated Computing Imperative
  - Dense Computing: GPUs and GP-GPUs
  - The broader potential
- A Framework for Accelerated Computing enablement
  - The Role of Architecture
  - The Emerging Layers of Computation
- Summary

The Role of Architecture

- Architecture:
  - *The **contract** between layers of Hardware and Software*
- Provides formalism and standardization → *Defines Compatibility*
  - Compatibility has been a key enabler in our industry – *this will continue*
  - History shows that viable products don't bet on wildly incompatible solutions
- Symbiotic Relationship between Hardware and Software
  - SW is typically the enabler for new HW features or new types of HW
    - Actual results dominated by the weakest link in this relationship
    - SW value chain often values *features* more than *HW optimization*
  - Software complexity driven to extreme levels – *this can't continue*
- Architecture gives rise to *The Emerging Layers of Computation*
  - *Can we use this to simplify the programming models?*

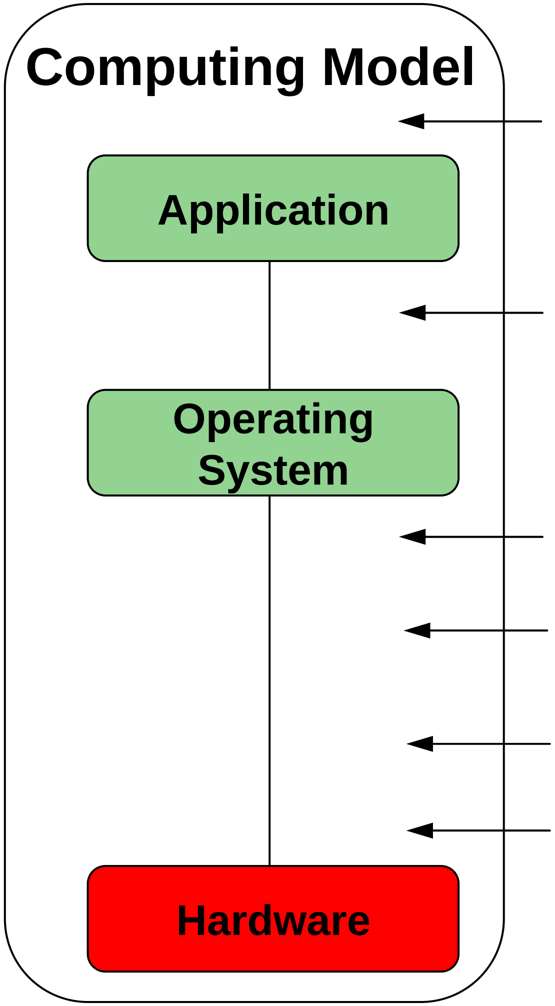
The Emerging Layers of Computation

Start with an Analogy to the Communications Industry



OSI Model

	Data	Layer
Host Layers	Data	Application Network Process to Application
	Data	Presentation Data Representation and Encryption
	Data	Session Interhost Communication
	Segments	Transport End-to-End Connections and Reliability
Media Layers	Packets	Network Path Determination and IP (Logical Addressing)
	Frames	Data Link MAC and LLC (Physical addressing)
	Bits	Physical Media, Signal, and Binary Transmission



Computing Model

Application

Operating System

Hardware

Fault tolerant “meta apps”

Browsers
Web services
JVM, CLR

Virtualization

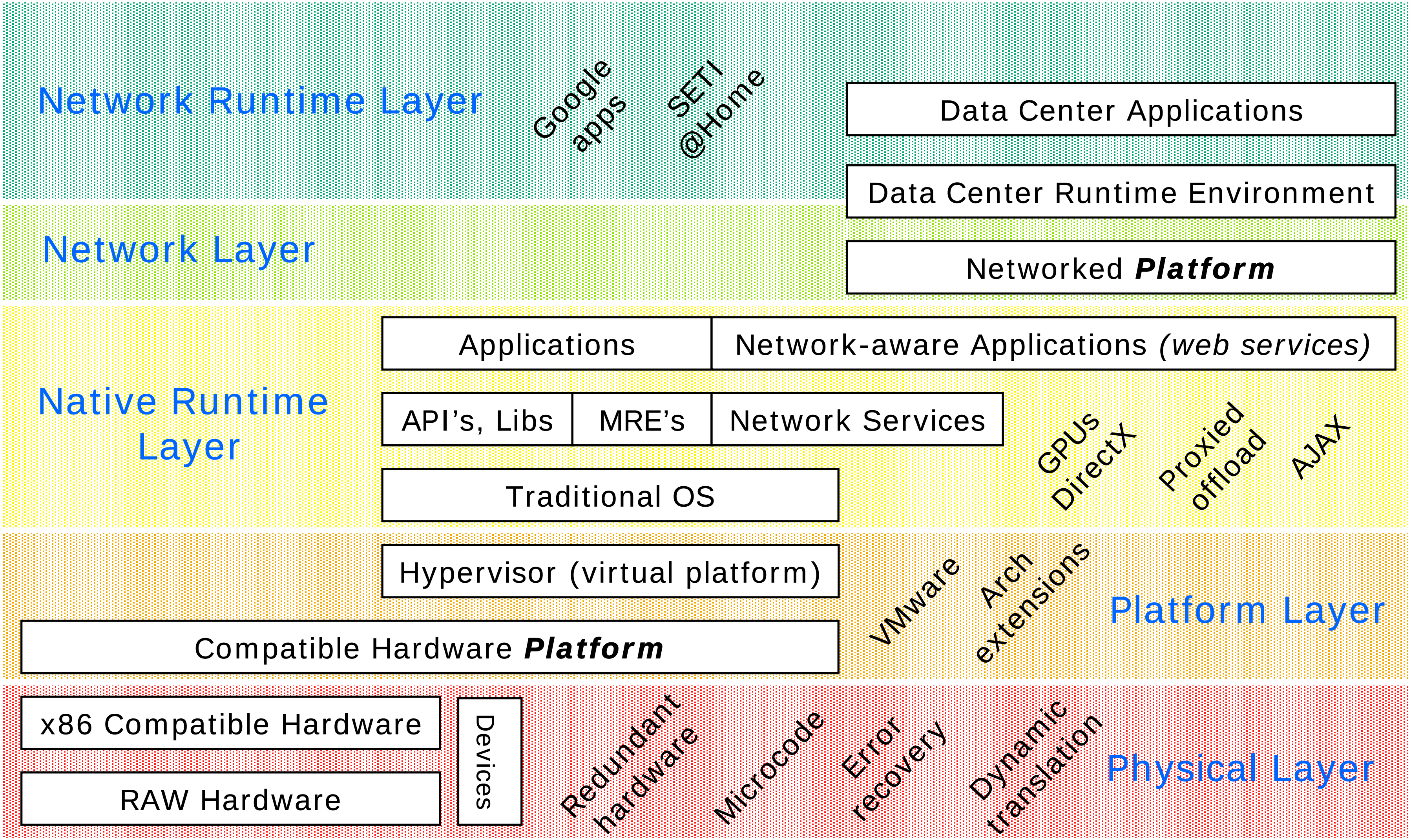
Intelligent I/O
Advanced APIs

Dynamic binary translation

Reconfiguration

Indicators of a bigger picture?

The Emerging Layers of Computation



Network Runtime Layer

Google apps

SETI @Home

Data Center Applications

Data Center Runtime Environment

Network Layer

Networked ***Platform***

Native Runtime Layer

Applications

Network-aware Applications (*web services)*

API's, Libs

MRE's

Network Services

Traditional OS

GPUs DirectX

Proxied offload

AJAX

Hypervisor (virtual platform)

Compatible Hardware ***Platform***

VMware

Arch extensions

Platform Layer

x86 Compatible Hardware

RAW Hardware

Devices

Redundant hardware

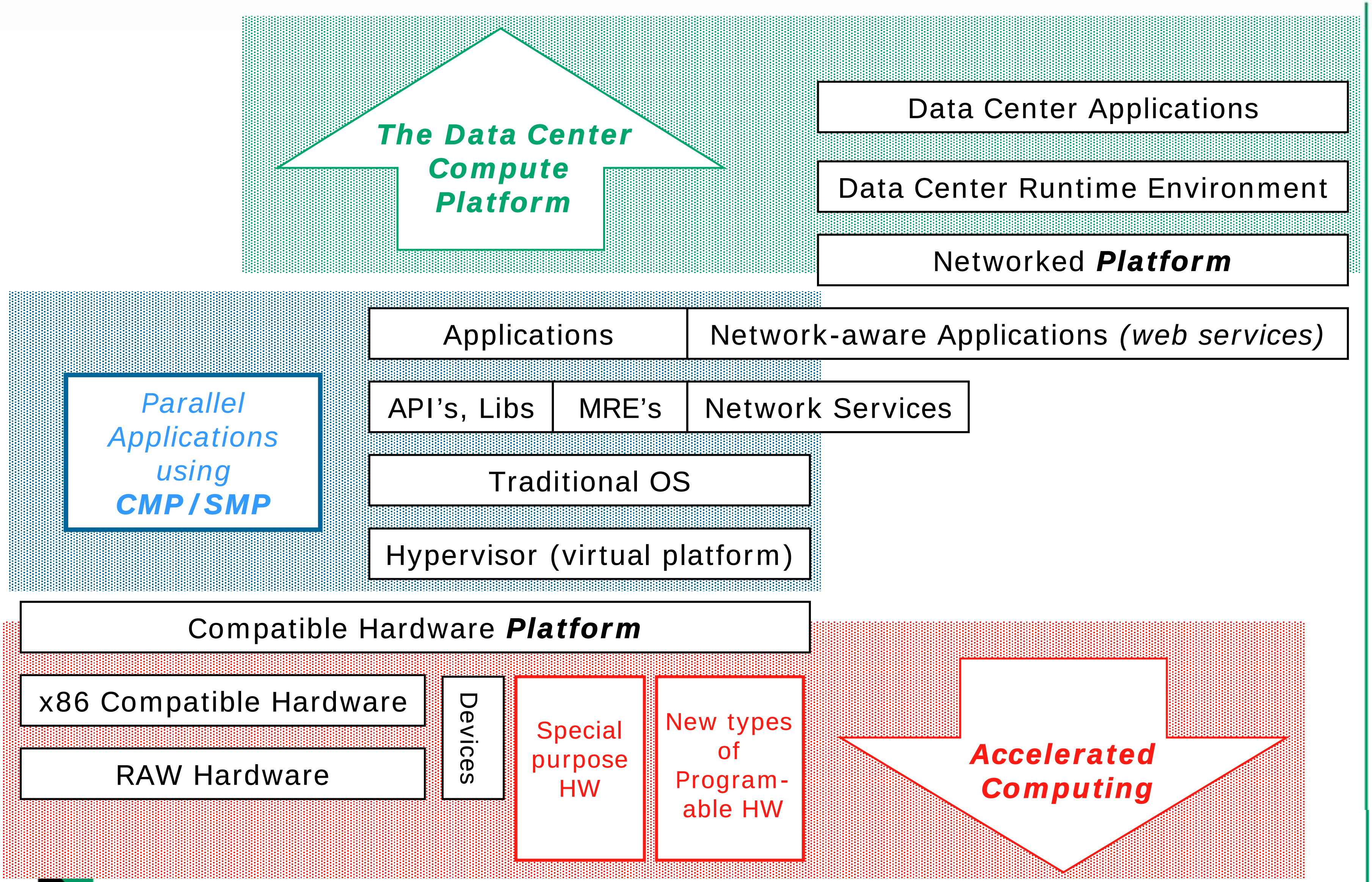
Microcode

Error recovery

Dynamic translation

Physical Layer

Lots of Interesting Implications



The Data Center Compute Platform
Data Center Applications
Data Center Runtime Environment
Networked **Platform**
Parallel Applications using **CMP / SMP**
Applications
Network-aware Applications *(web services)*
API's, Libs
MRE's
Network Services
Traditional OS
Hypervisor (virtual platform)
Compatible Hardware **Platform**
x86 Compatible Hardware
RAW Hardware
Devices
Special purpose HW
New types of Program-able HW
Accelerated Computing

Summary: *The Case for Accelerated Computing*

Traditional “host” → offload to dense compute accelerator

- Use ***APIs*** to enable this without heroic programming efforts
- Proven techniques already in use with DirectX & GPUs today
- *ISA compatibility yields to API and Platform Compatibility*

Many application classes have reasonably common “kernels”

- Video encoding; Encryption; Data Movement; Java/CLR …

Broad range of possible accelerator designs & attach points

- Coherent domain or non-coherent domain
- Dedicated ***special-purpose HW*** or ***programmable processor***

Lots of Challenges

- Managing context state → Virtualizing the context state
- Communications/Messaging: *“It’s the synchronization, stupid”*
- Memory BW and Data Movement (keep up with computation)
- New and appropriate APIs

Thank You !

Questions?