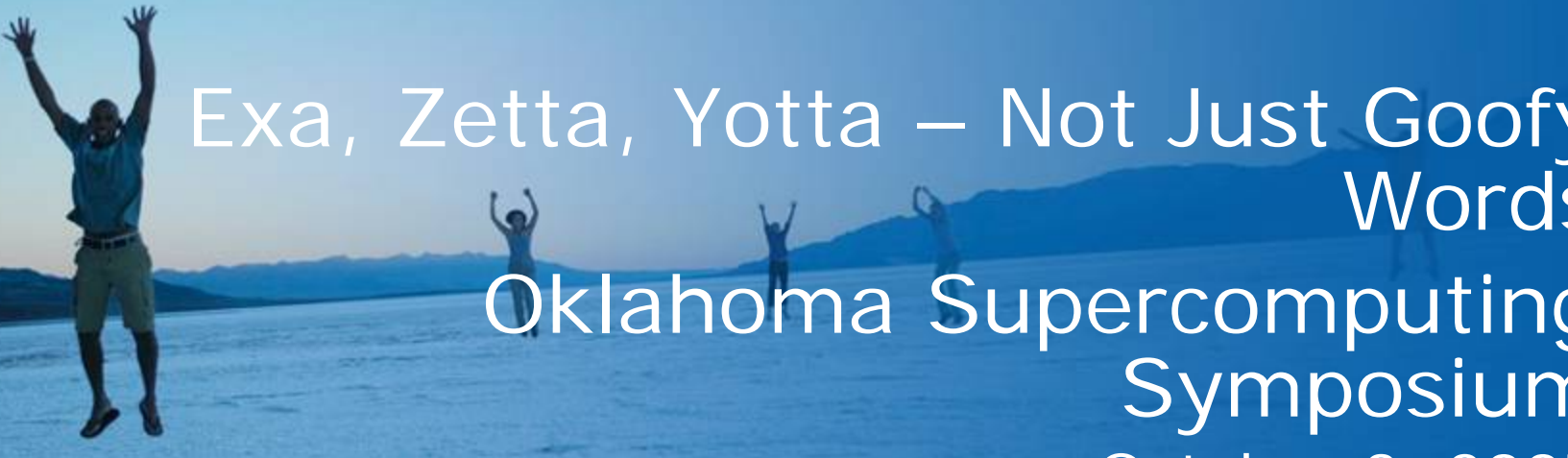


Aim High HPC @ Intel

A series of silhouettes of people jumping joyfully in a vast, flat, open field under a clear blue sky. The figures are scattered across the middle ground, with the most prominent one on the left side of the frame. The background shows a range of low mountains or hills under a bright, clear sky.

Exa, Zetta, Yotta – Not Just Goofy
Words
Oklahoma Supercomputing
Symposium
October 3, 2007

Stephen R. Wheat, Ph.D.
Director, HPC
Digital Enterprise Group

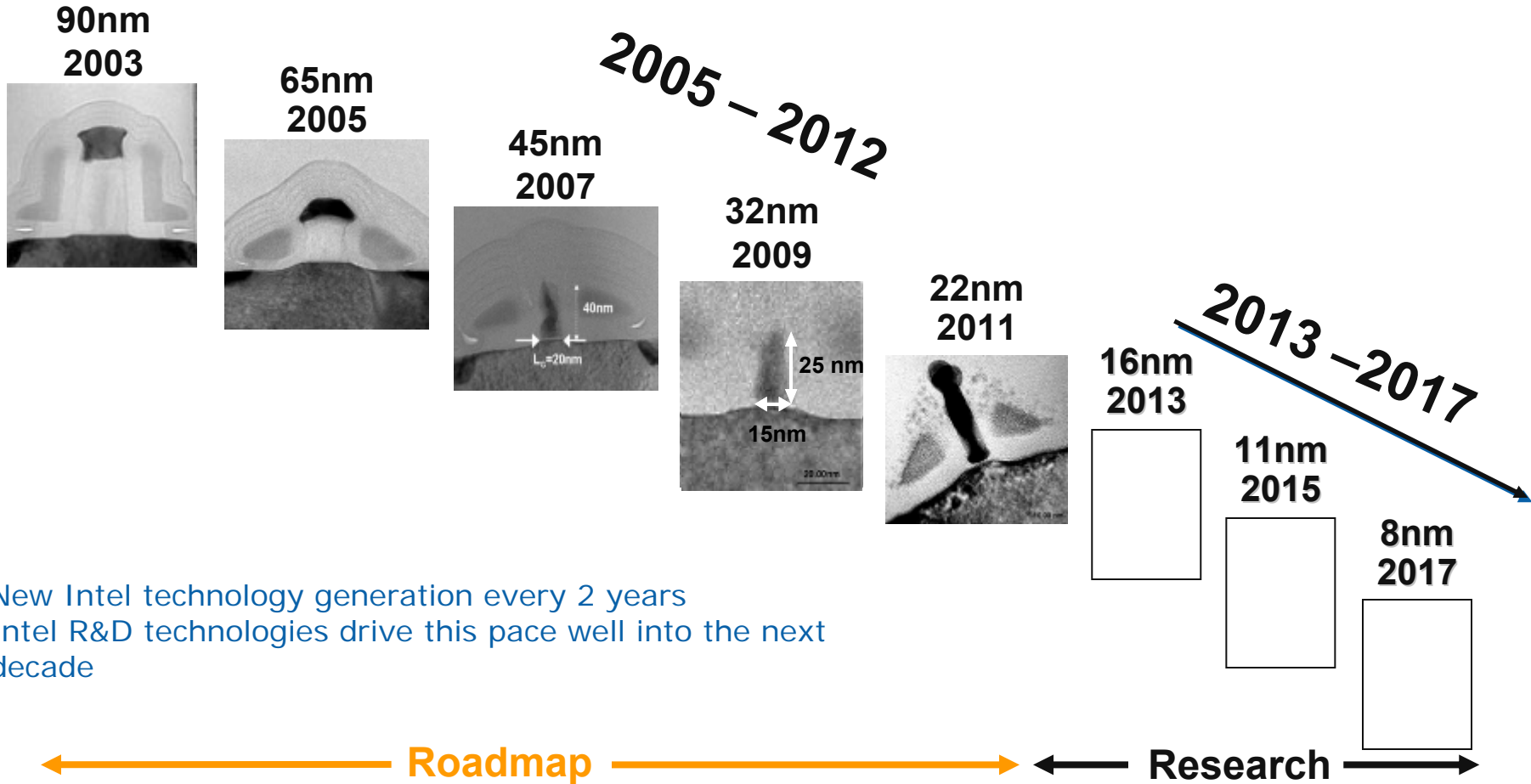
Risk Factors

Today's presentations contain forward-looking statements. All statements made that are not historical facts are subject to a number of risks and uncertainties, and actual results may differ materially. Please refer to our most recent Earnings Release and our most recent Form 10-Q or 10-K filing available on our website for more information on the risk factors that could cause actual results to differ.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (<http://www.intel.com/performance/resources/limits.htm>).

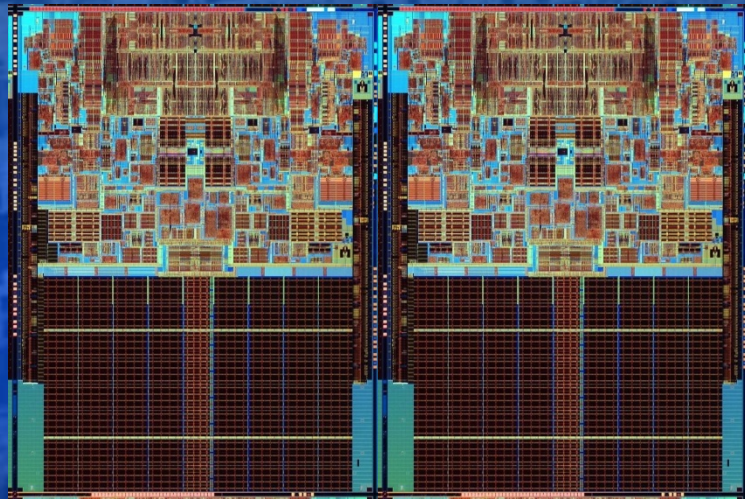


Silicon Future



45nm Advantage

Intel® Xeon® 5300 Processor
(Clovertown)
65nm

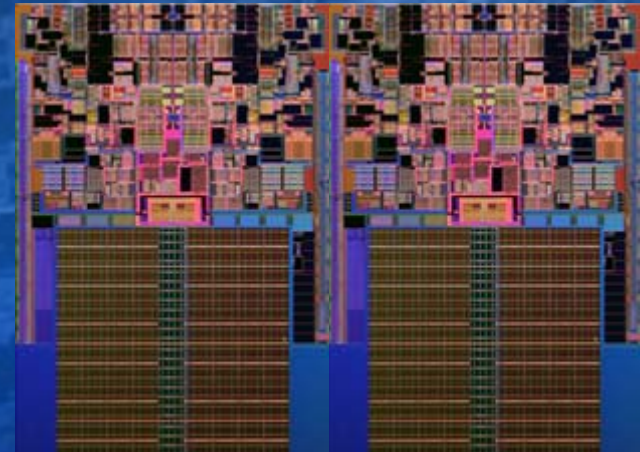


143 mm²*

143 mm²*

582m Transistors
8 MB Cache

Intel® Xeon® 5400 Processor
(Harpertown)
45nm Hi-k



107 mm²*

107 mm²*

820m Transistors
12 MB Cache

Millions of Quad Core Processors Shipped



*Source: Intel

Note: die picture sizes are approximate

High Performance Computing

Technically motivated computing where performance matters more than cost or ease of use— from desktop to highest end supercomputing:

- Scientific Discovery
- Engineering Innovation
- Finance and Decision Support
- Geo-Economics and Societal Complexities
- Knowledge Discovery



HPC and Mainstream Computing

- Historically most hardware and software architectural innovations have come through High End Computing
- Today, innovation moves up from the bottom (e.g. low-power processing) and down from the top (e.g., parallel computing). But the High-End is still a dominant source of new ideas
- What is in today's supercomputer will be in tomorrow's desktop and next week's embedded platform.



Size does matter

Peta Mac



Exa Mac



Zeta Mac



Yotta Mac

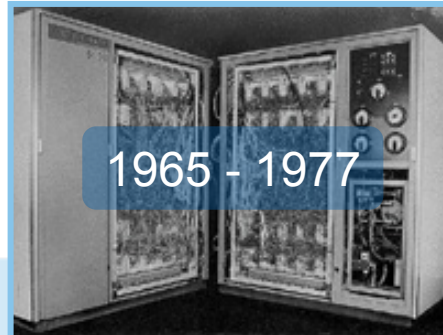


Yesterday, Today and Tomorrow in HPC

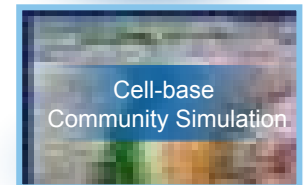
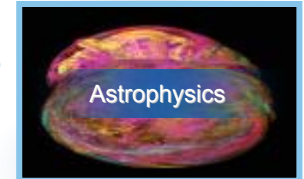
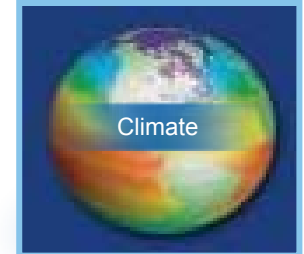
ENIAC
20 Numbers in Main Memory



CDC 6600 – First successful
Supercomputer 9MFlops



~2008 Beyond



ASCI Red
(world's fastest Jan 1997– Nov 2000 on
top500 till Nov 2005)
First Teraflop Computer,
9298 Intel® Pentium® II Xeon® Processors

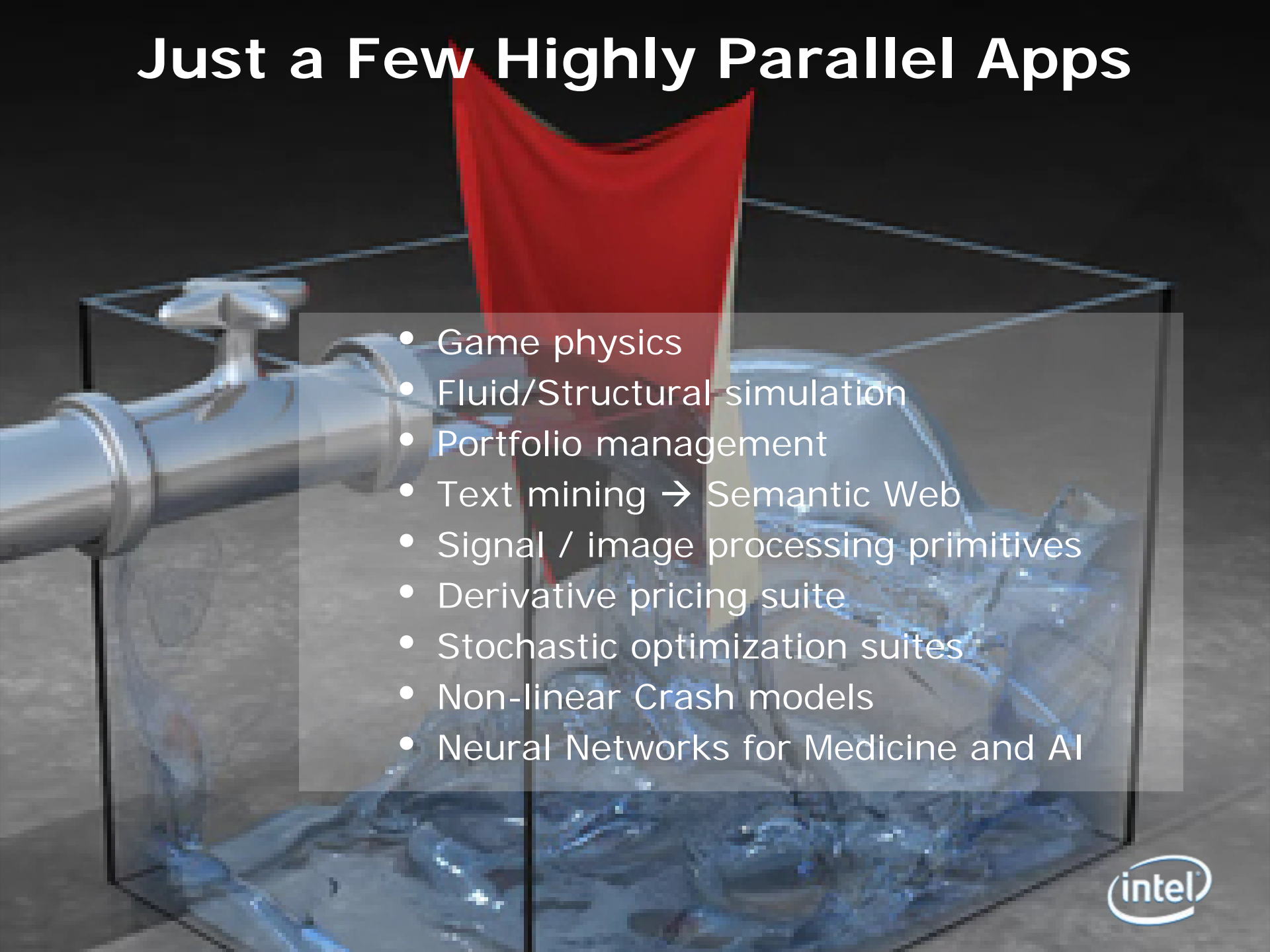
Intel ENDEAVOR
464 Intel® Xeon® Processors 5100 series,
6.85 Teraflop MP Linpack, #68 on top500

*Petascale
Platforms*

*Yesterday's High-end Supercomputing
is Today's Personal Computing*



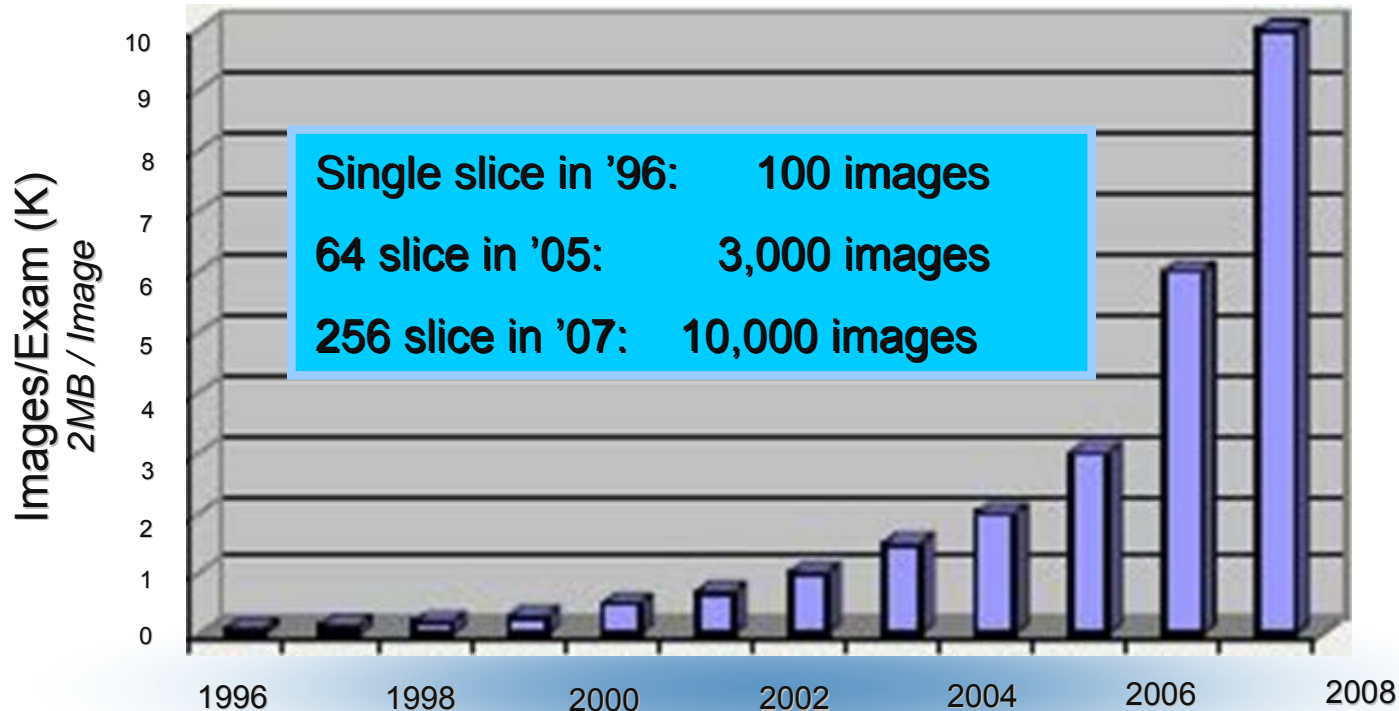
Just a Few Highly Parallel Apps

- 
- Game physics
 - Fluid/Structural simulation
 - Portfolio management
 - Text mining → Semantic Web
 - Signal / image processing primitives
 - Derivative pricing suite
 - Stochastic optimization suites
 - Non-linear Crash models
 - Neural Networks for Medicine and AI



Exploding Demand for Data Processing

Example: HPC in Medical Imaging



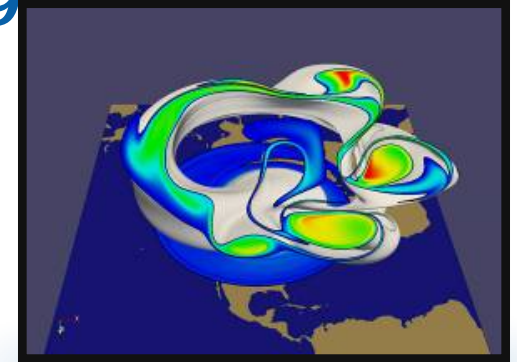
- 3-D renderings of the images
- Computer aided diagnostic algorithms
- Fusions of images from different modalities
 - MRI, CT, PET, and SPECT
- Real-time applications are appearing

Full Body CT – 256 slice/10,000 images: a 20GB file

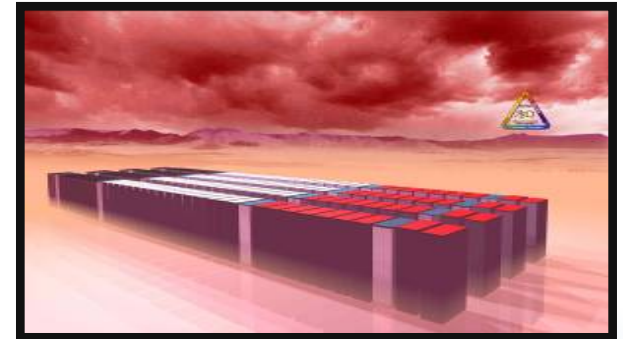


Today's Science Demands Petascale

Example: HPC in Climate Computing



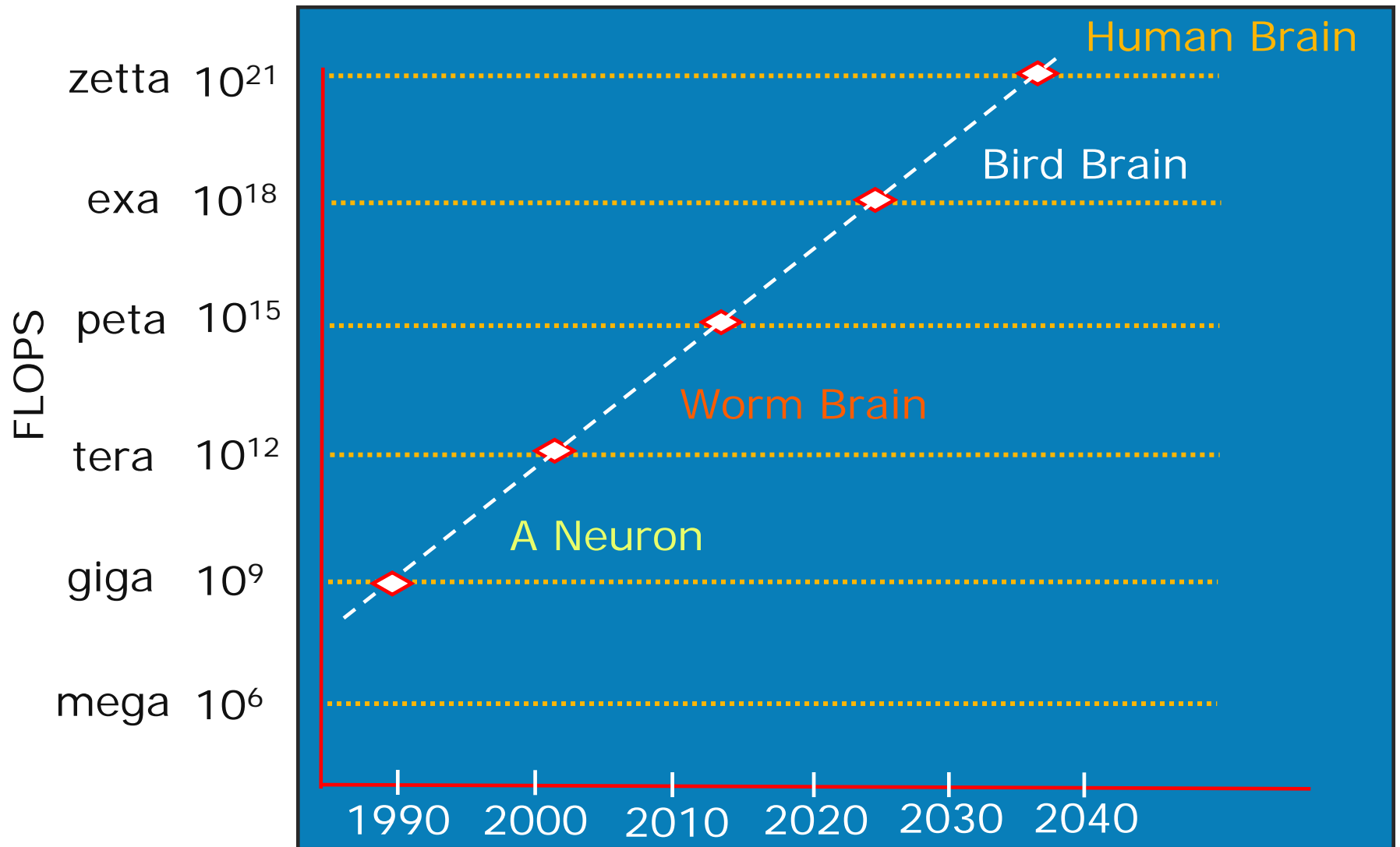
- Some believe that global warming will produce more extremes weather (drought/flooding).
- Current models are too coarse and inaccurate to be reliable globally much less for predicting climate change at the national level.
- To predict regional climate change:
 - Community climate model resolution goal is *10 km*
 - Simulate ~*150 days/day* on today's fastest computer at 10 km using NCAR/Sandia SEAM
 - Typical climate simulation is for *100 yrs.*



To Simulate 100 Year Climate:

1.6 sustained PFLOPS = about a month of Computing; 50 PFLOPS = a day

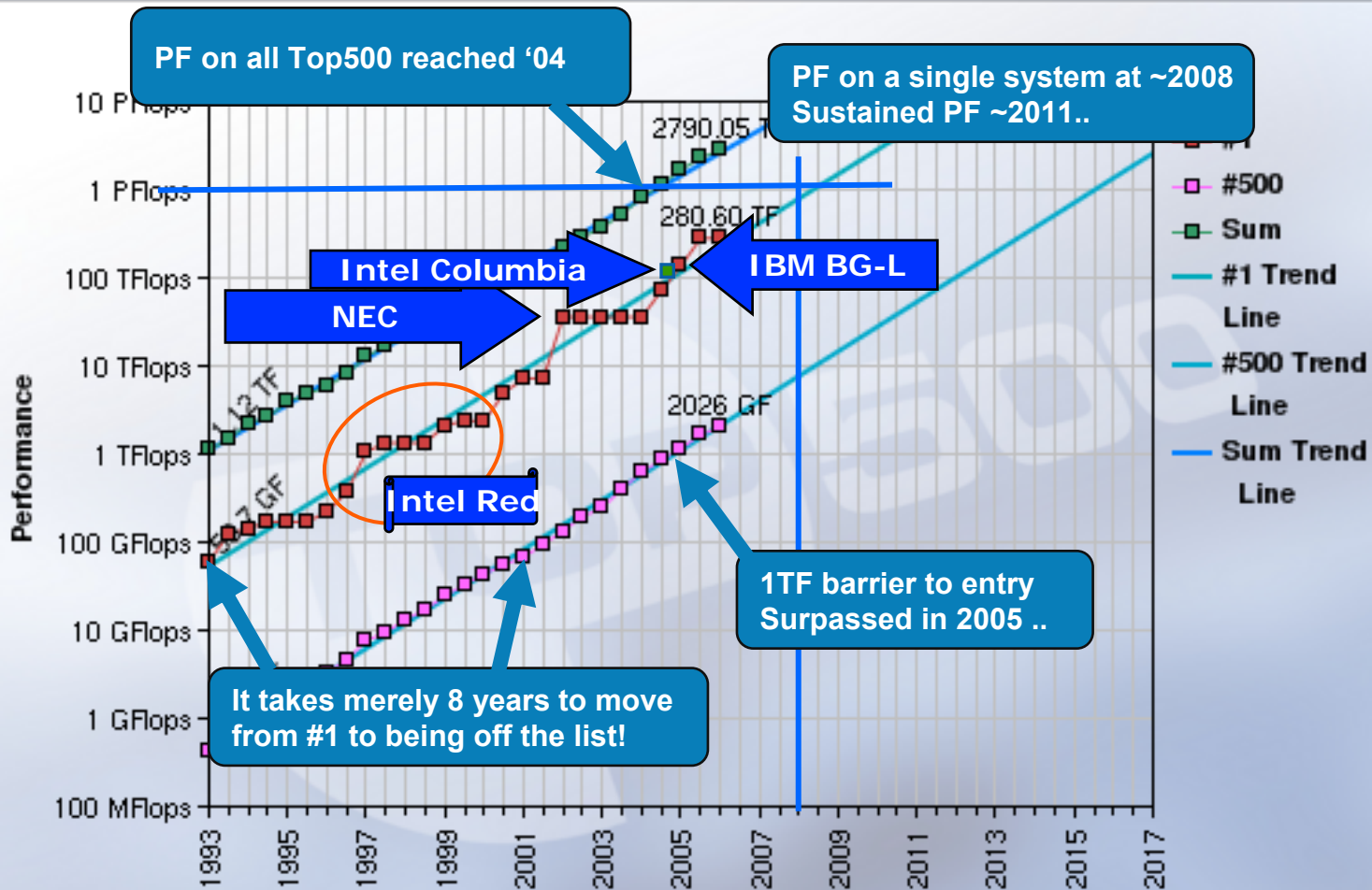
Modeling the Brain



The Top500: Reaching Petascale



Projected Performance Development



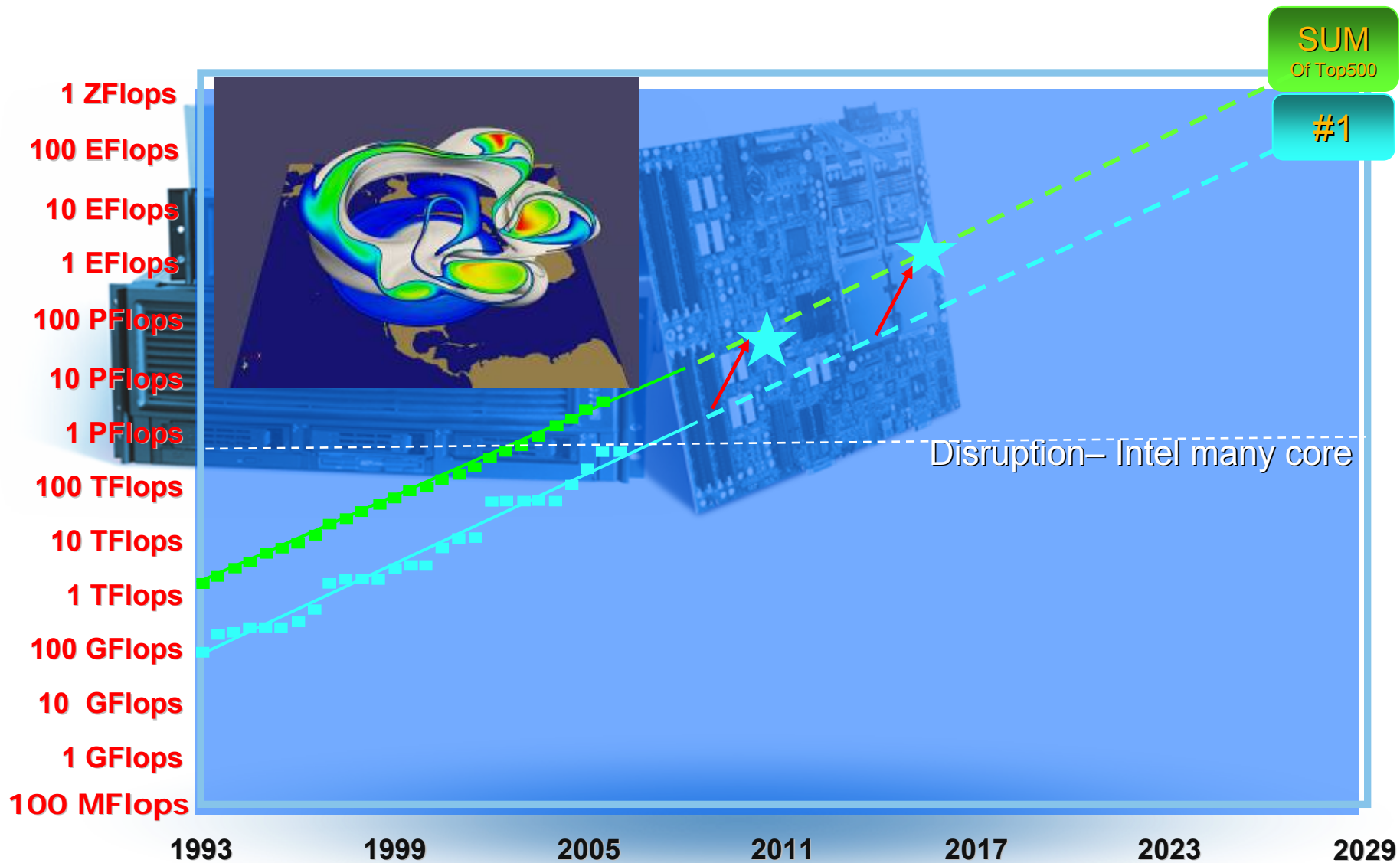
28/06/2006

<http://www.top500.org/>

Source: top500.org



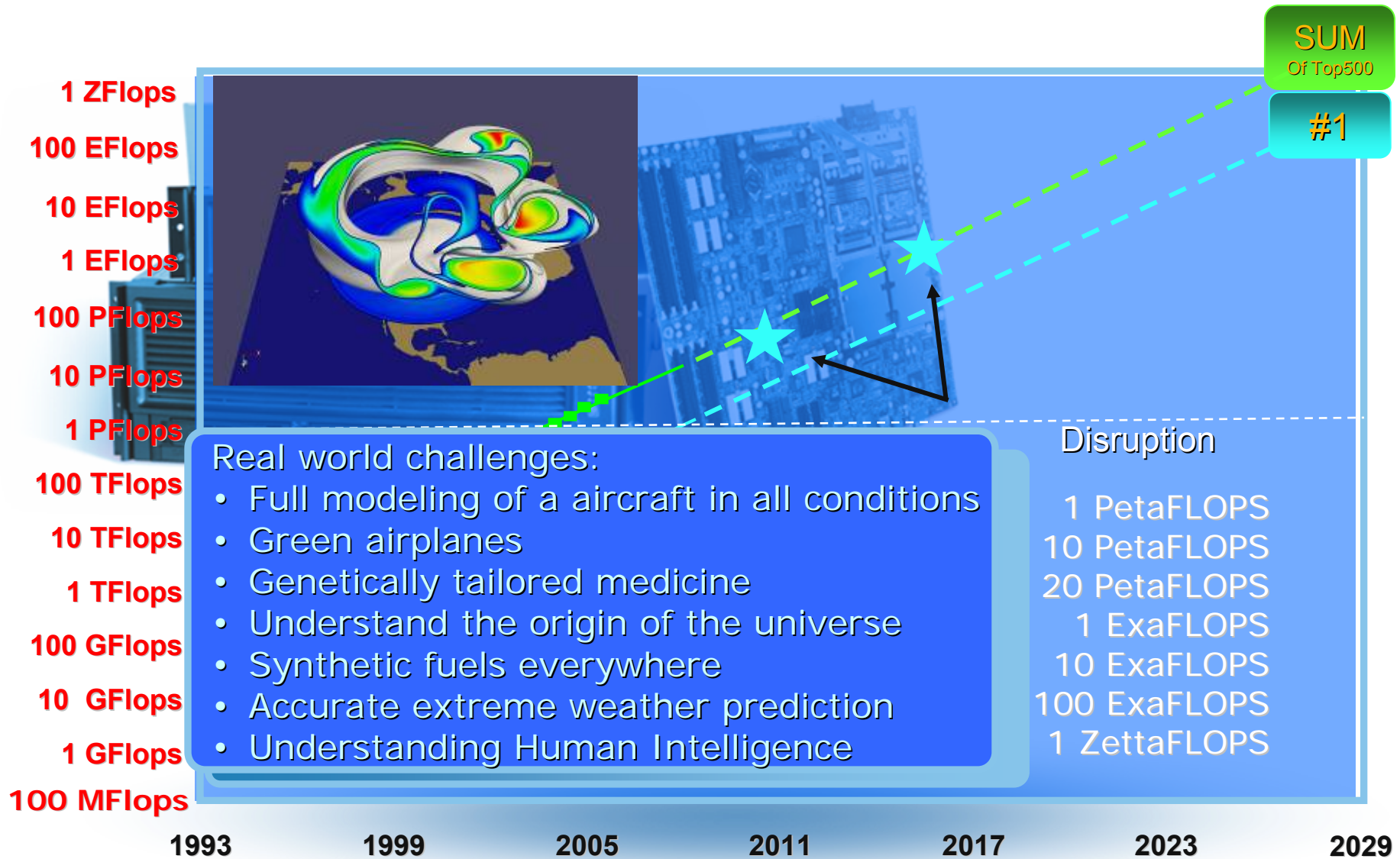
Driving to Petascale & Exascale



Source: Dr. Steve Chen, "The Growing HPC Momentum in China", June 30th, 2006, Dresden, Germany



Driving to Petascale & Exascale For Science and engineering



Source: Dr. Steve Chen, "The Growing HPC Momentum in China", June 30th, 2006, Dresden, Germany

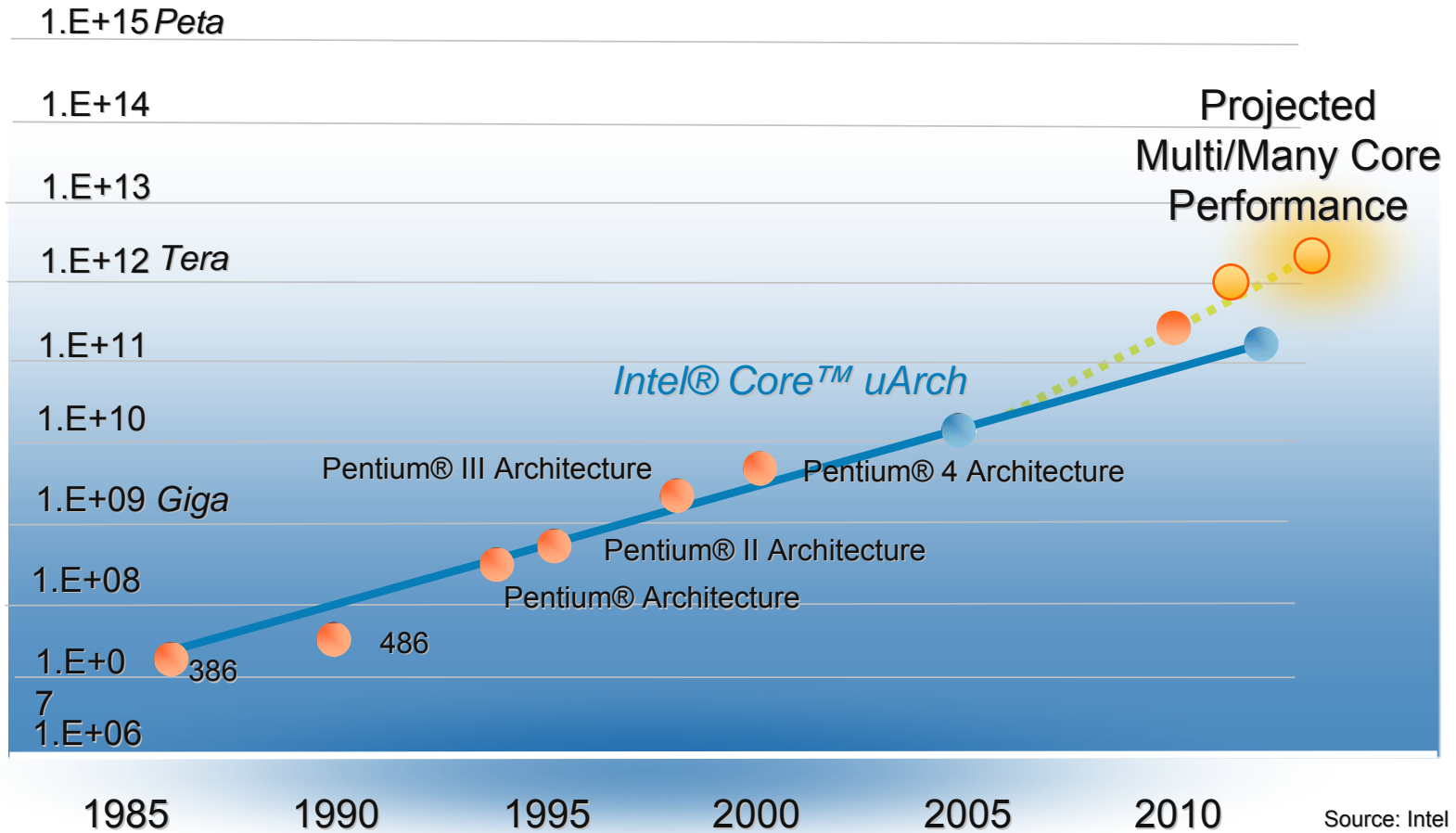


Challenges to Success at the Petascale

- Programmability and Scalability
- Processor Speed
- Memory Performance
- Network Performance
- Power
- Reliability
- Manageability
- Purchase and Ownership Cost

Processor Performance

FLOPS



Sustaining Petascale with ~6000 Processors in 2010



Performance within the Power Envelope

$$P \sim \frac{1}{2} \omega C V_{\min}^2$$

$$V_{\min} \sim V_0 \omega \quad \rightarrow \quad P \sim \frac{1}{2} C V_0^2 \omega^3$$

As we shrink features, new challenges arise:

- Error management

- Leakage power

Moore's Law continues,

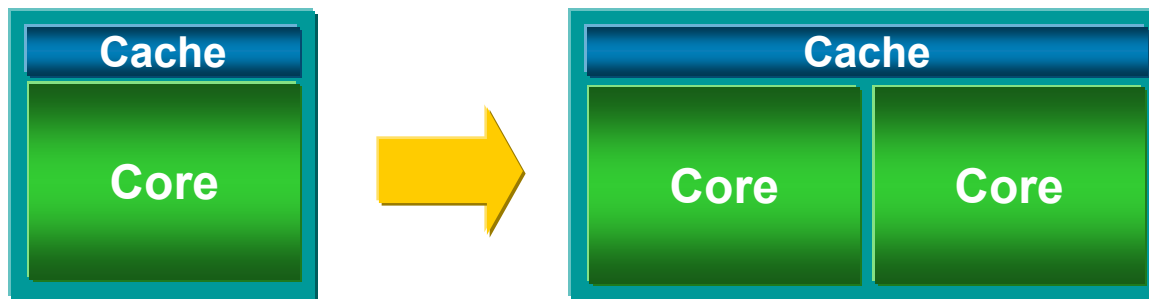
But Frequency increase is

becoming more challenging

Performance within the Power Envelope

$$P \sim \frac{1}{2} \omega C V_{\min}^2$$

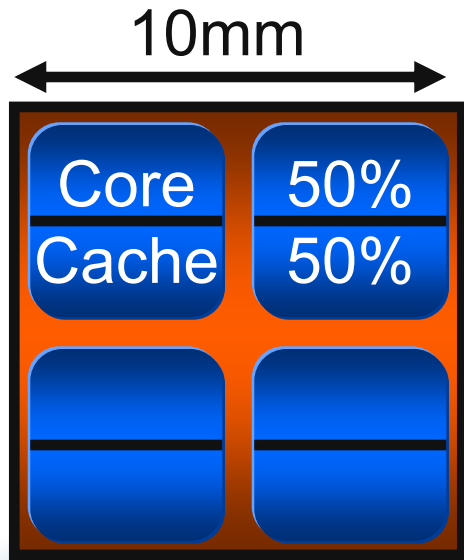
$$V_{\min} \sim V_0 \omega \quad \rightarrow \quad P \sim \frac{1}{2} C V_0^2 \omega^3$$



Voltage = 1
Freq = 1
Power = 1
Perf = 1

Voltage = -20%
Freq = -20%
Power = 1
Perf = ~1.7

A Sample Many Core System



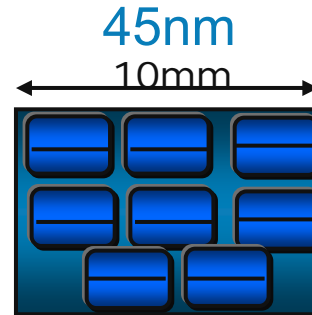
65nm, 4 Cores

1V, 3GHz

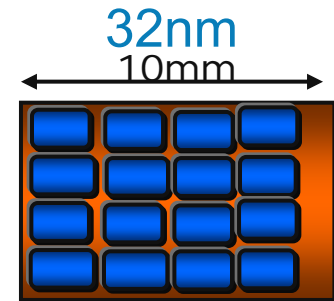
10mm die, 5mm each core

Core Logic: 6MT, Cache: 44MT

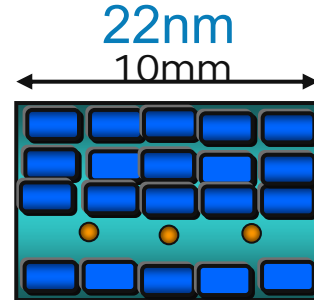
Total transistors: 200M



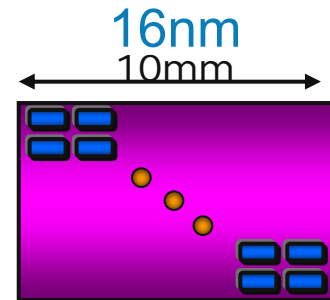
8 Cores, 1V, 3GHz
3.5mm each core
Total: 400MT



16 Cores, 1V, 3GHz
2.5mm each core
Total: 800MT



32 Cores, 1V, 3GHz
1.8mm each core
Total: 1.6BT



64 Cores, 1V, 3GHz
1.3mm each core
Total: 3.2BT

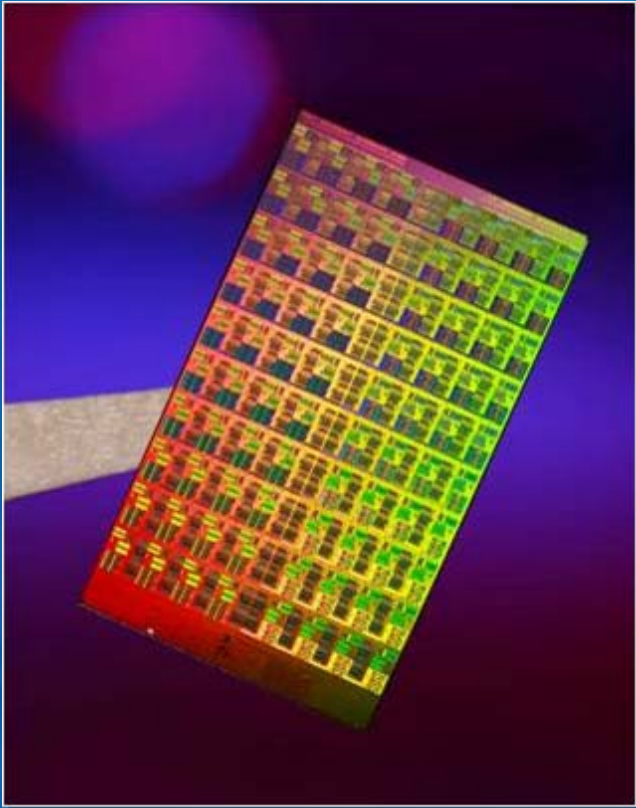
Note: the above pictures don't represent any current or future Intel products

*Research Challenge:
Asymmetric vs. symmetric, Homogenous vs. heterogeneous
What kind of applications will benefit?*



Teraflops Research Chip

100 Million Transistors • 80 Tiles • 275mm²



First tera-scale programmable silicon:

- Teraflops performance
- Tile design approach
- On-die mesh network
- Novel clocking
- Power-aware capability
- Supports 3D-memory

Not designed for IA or product

Tera-scale Introduction

- Represents significant Intel transition from “large” cores to 32+ low-power, highly-threaded IA cores per die
- Motivations for a new architecture
 - Enable emerging workloads and new use-models
 - Low Power IA cores provide 4-5X greater performance-power efficiency
 - Scaling beyond the limits of Instruction level parallelism and single-core power
- Tera-scale is *NOT* simply SMP-on-die
 - Will require complete platform and software enabling

Parameter	SMP	Tera-scale	Improvement	Optimizations
Bandwidth	12 GB/s	~1.2 TB/s	~100X	Massive bandwidth between cores
Latency	400 cycles	20 cycles	~20X	Ultra-fast synchronization



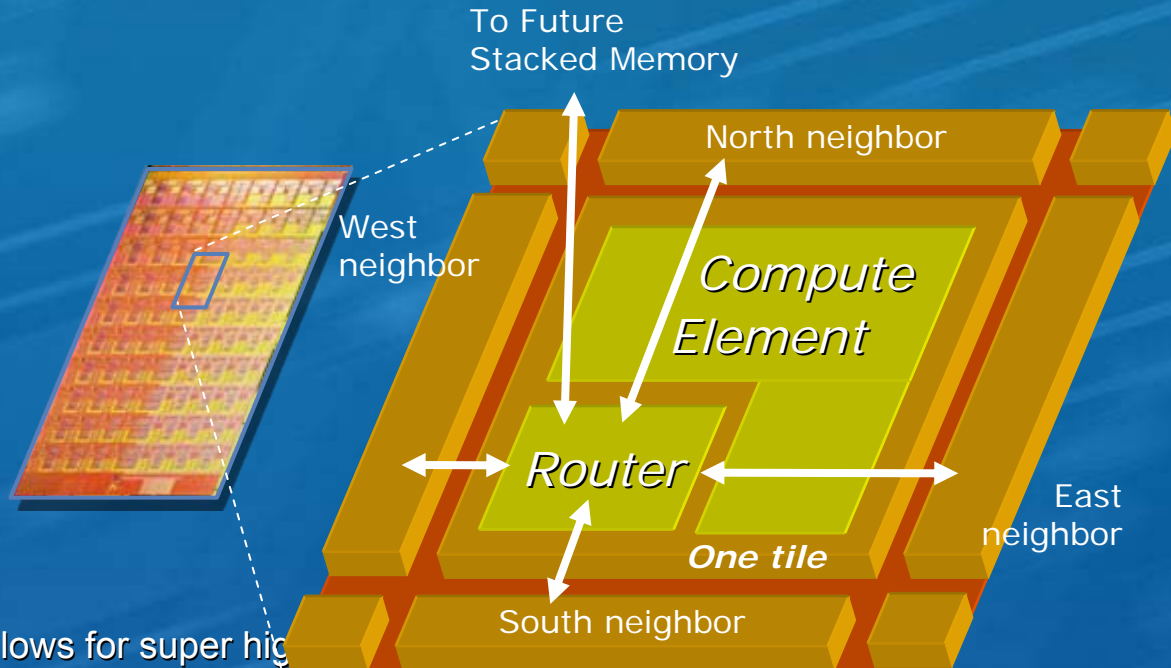
Tiled Design & Mesh Network

Repeated Tile Method:

- Compute + router
- Modular, scalable
- Small design teams
- Short design cycle

Mesh Interconnect:

- “Network-on-a-Chip”
 - Cores networked in a grid allows for super high communications in and between cores
- 5-port, 80GB/s* routers
- Low latency (1.25ns*)
- Future: connect IA/or and special purpose cores



* When operating at a nominal speed of 4GHz

Fine Grain Power Management

- Novel, modular clocking scheme saves power over global clock
- New instructions to make any core sleep or wake as apps demand
- Chip Voltage & freq. control (0.7-1.3V, 0-5.8GHz)

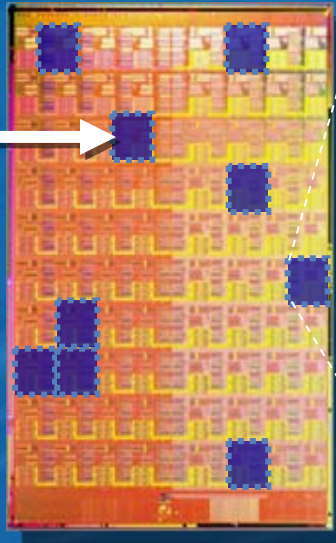
Dynamic sleep

STANDBY:

- Memory retains data
- **50%** less power/tile

FULL SLEEP:

- Memories fully off
- **80%** less power/tile



21 sleep regions per tile (not all shown)

Data Memory

*Sleeping:
57% less power*

Instruction Memory

*Sleeping:
56% less power*

Router

*Sleeping:
10% less power
(stays on to pass traffic)*

FP Engine 1

*Sleeping:
90% less power*

FP Engine 2

*Sleeping:
90% less power*

Industry leading energy-efficiency of 16 Gigaflops/Watt



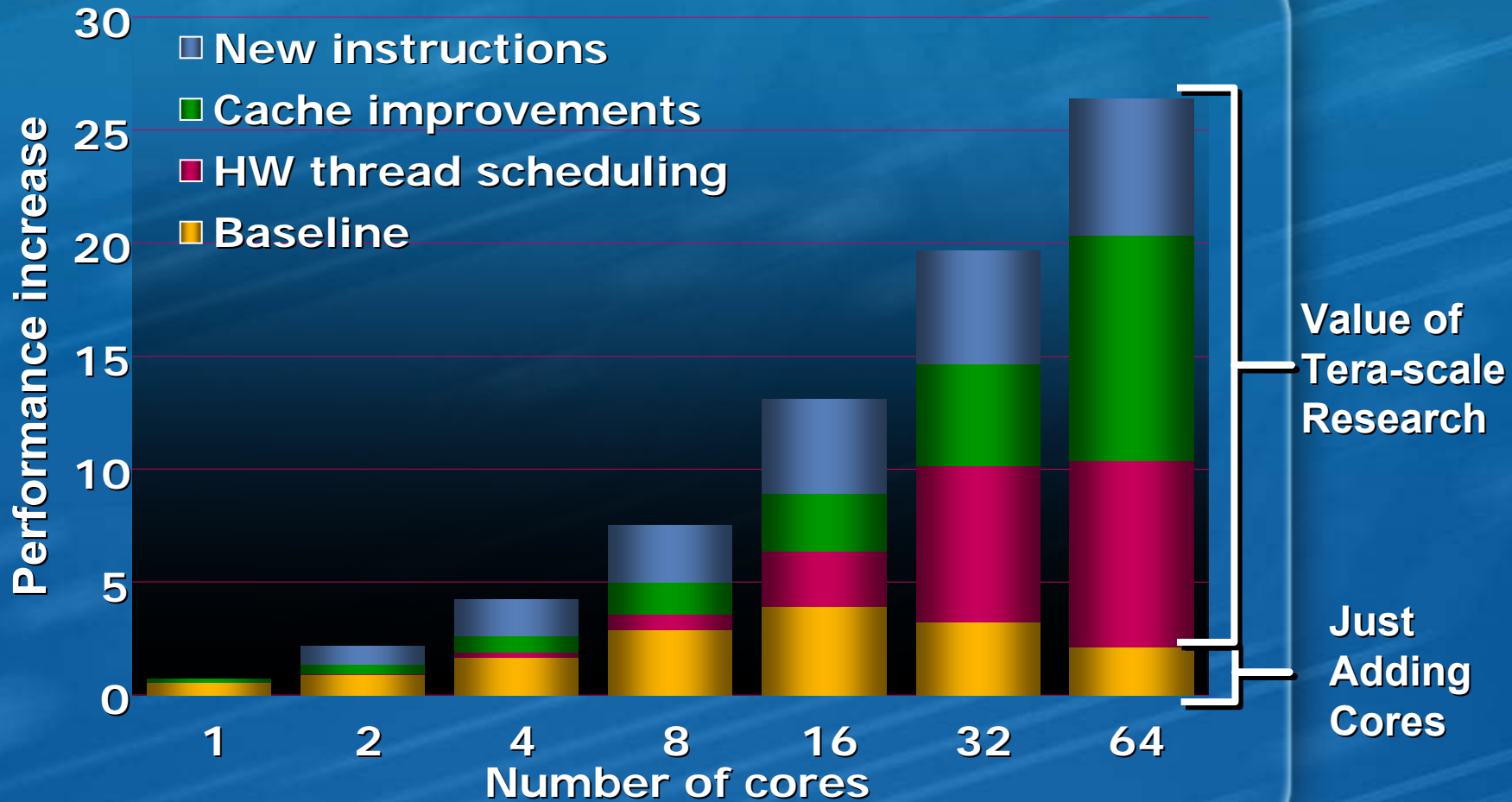
Research Data Summary

Frequency	Voltage	Power	Bisection Bandwidth	Performance
3.16 GHz	0.95 V	62W	1.62 Terabits/s	1.01 Teraflops
5.1 GHz	1.2 V	175W	2.61 Terabits/s	1.63 Teraflops
5.7 GHz	1.35 V	265W	2.92 Terabits/s	1.81 Teraflops

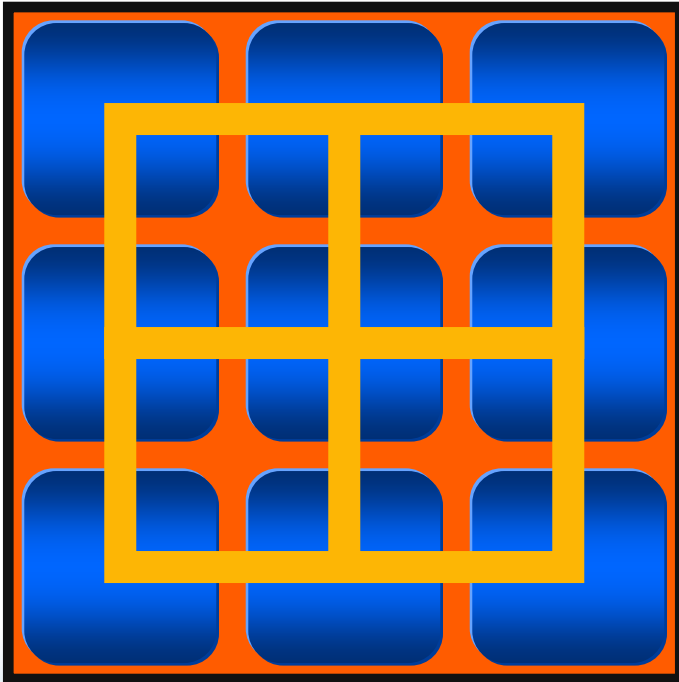
**1.01 Teraflops
62 Watts**



More than the Cores



Intra-chip Interconnect Bus for Future Many Core Chip?



Issues:

Slow

Shared, limited scalability?

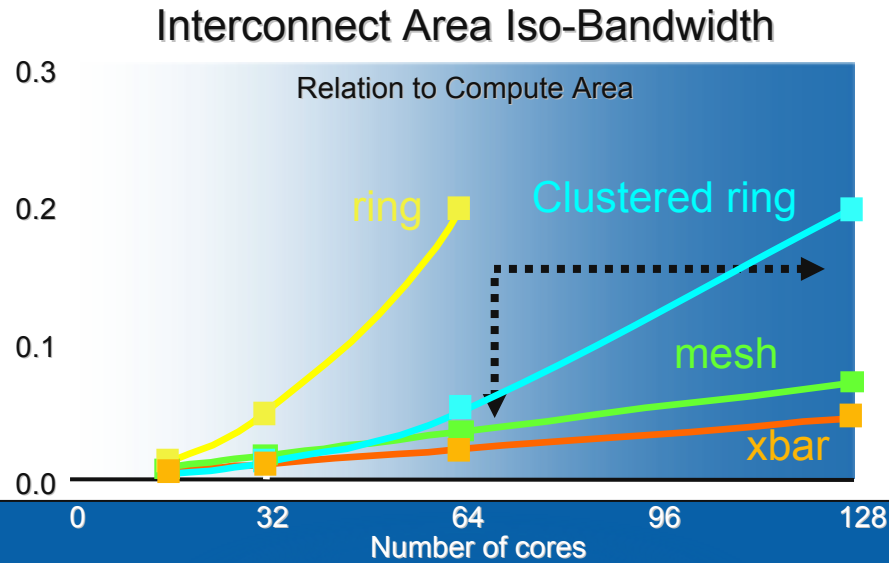
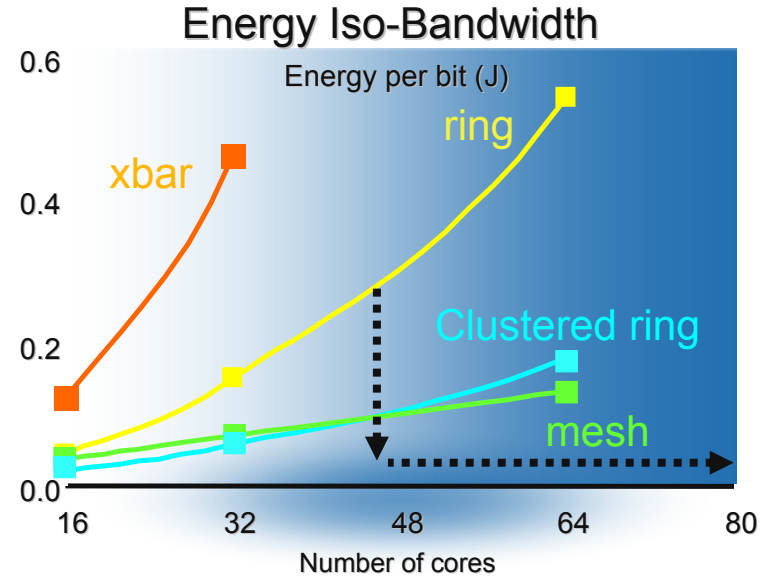
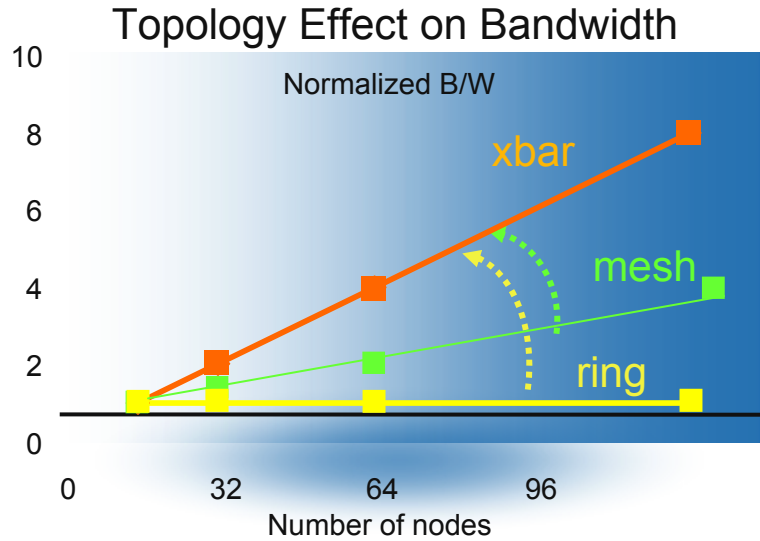
Benefits:

Power?

Simpler cache coherency

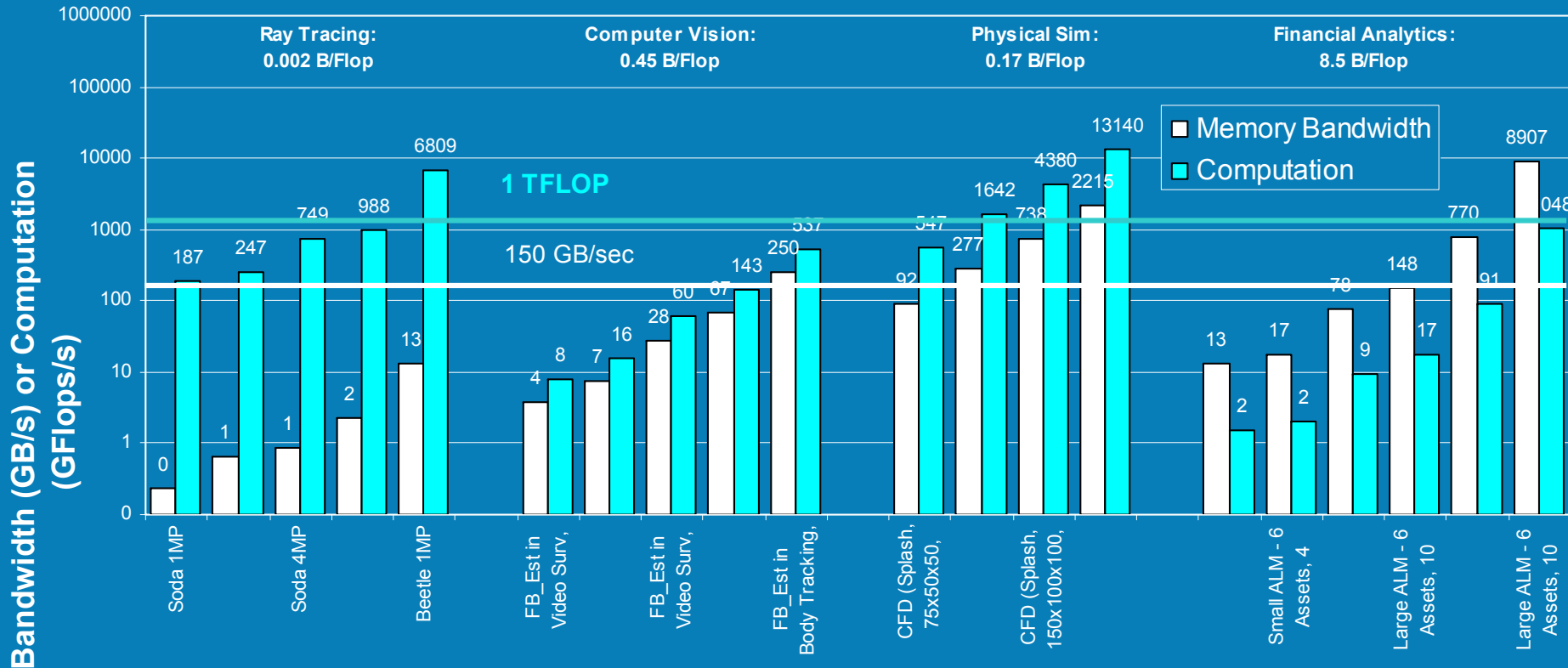
Traditional Bus is Not a Good Interconnect Option

Intra-chip Interconnect Options to Evaluate Bandwidth, Link Bandwidth and Power



How Do We Feed the Machine?

RMS Workload - Bandwidth and Computation Requirements



Source: Intel Labs

Memory Bandwidth and Processor Performance Need to Keep Pace



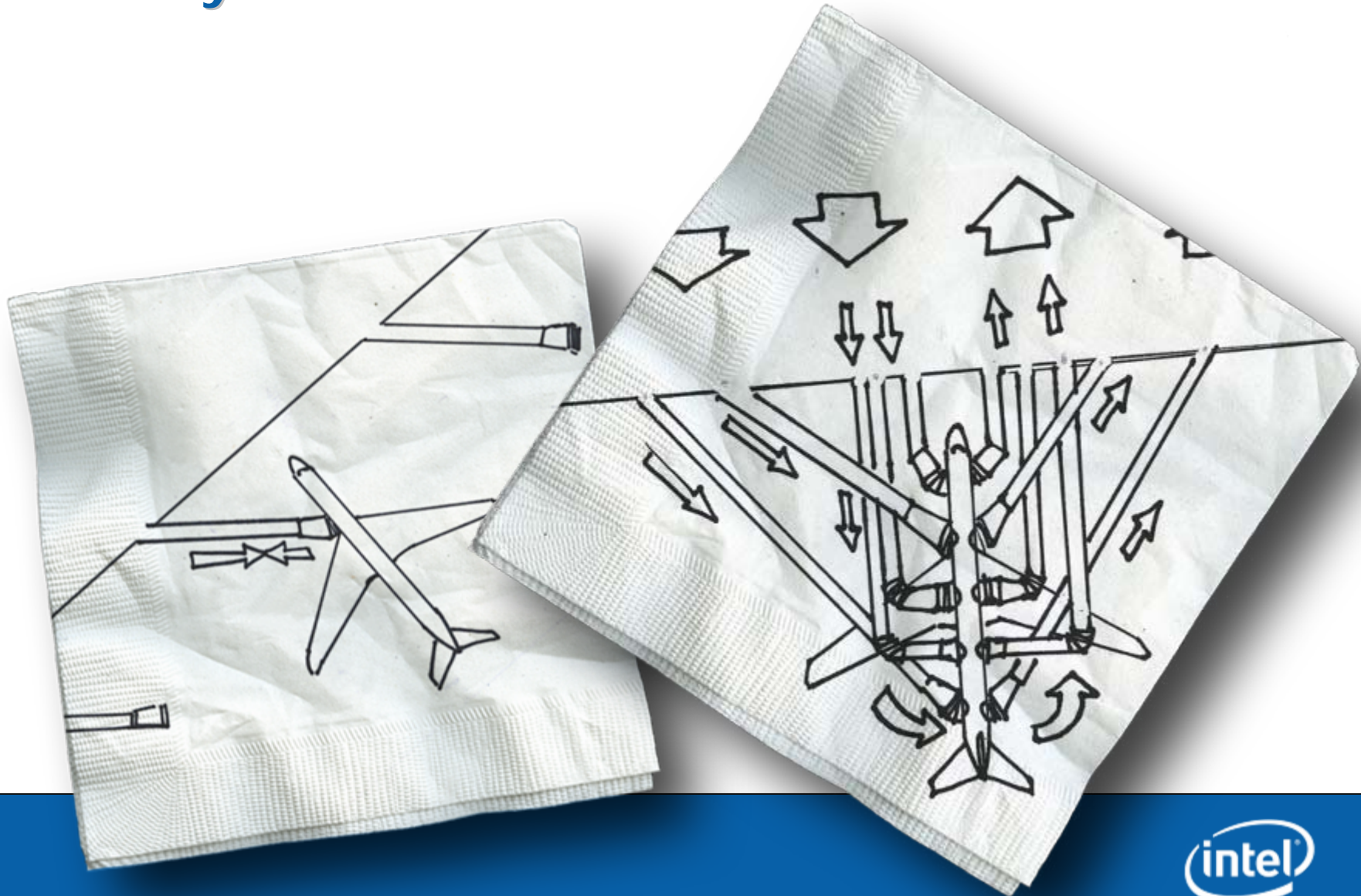
What If?

Moore's Law Could be Applied to the Airline Industry?

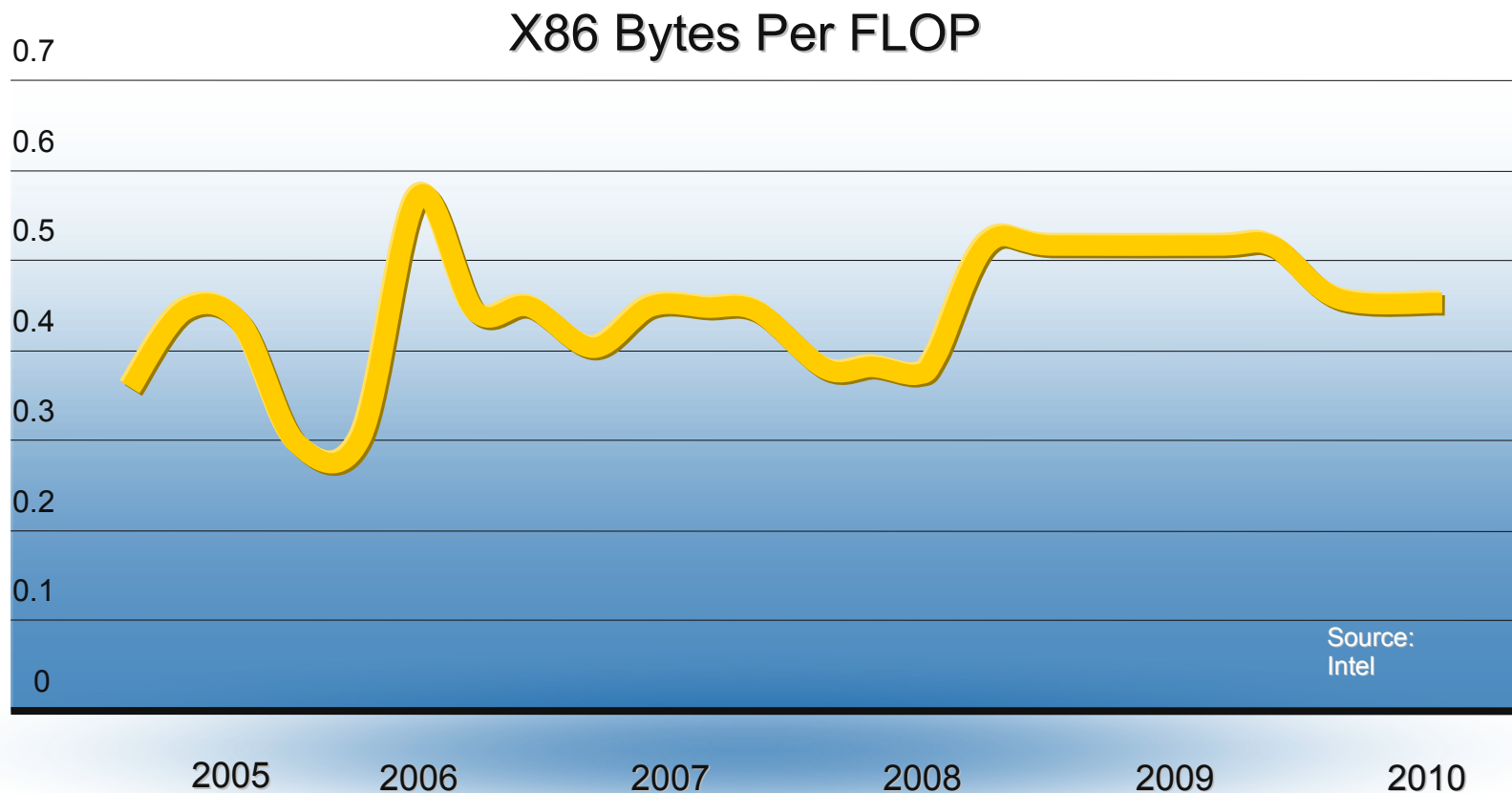


What If?

Moore's Law Could be Applied to the Airline Industry?



Memory Performance for Balanced Computing



Byte : Flop Ratio has been Consistent



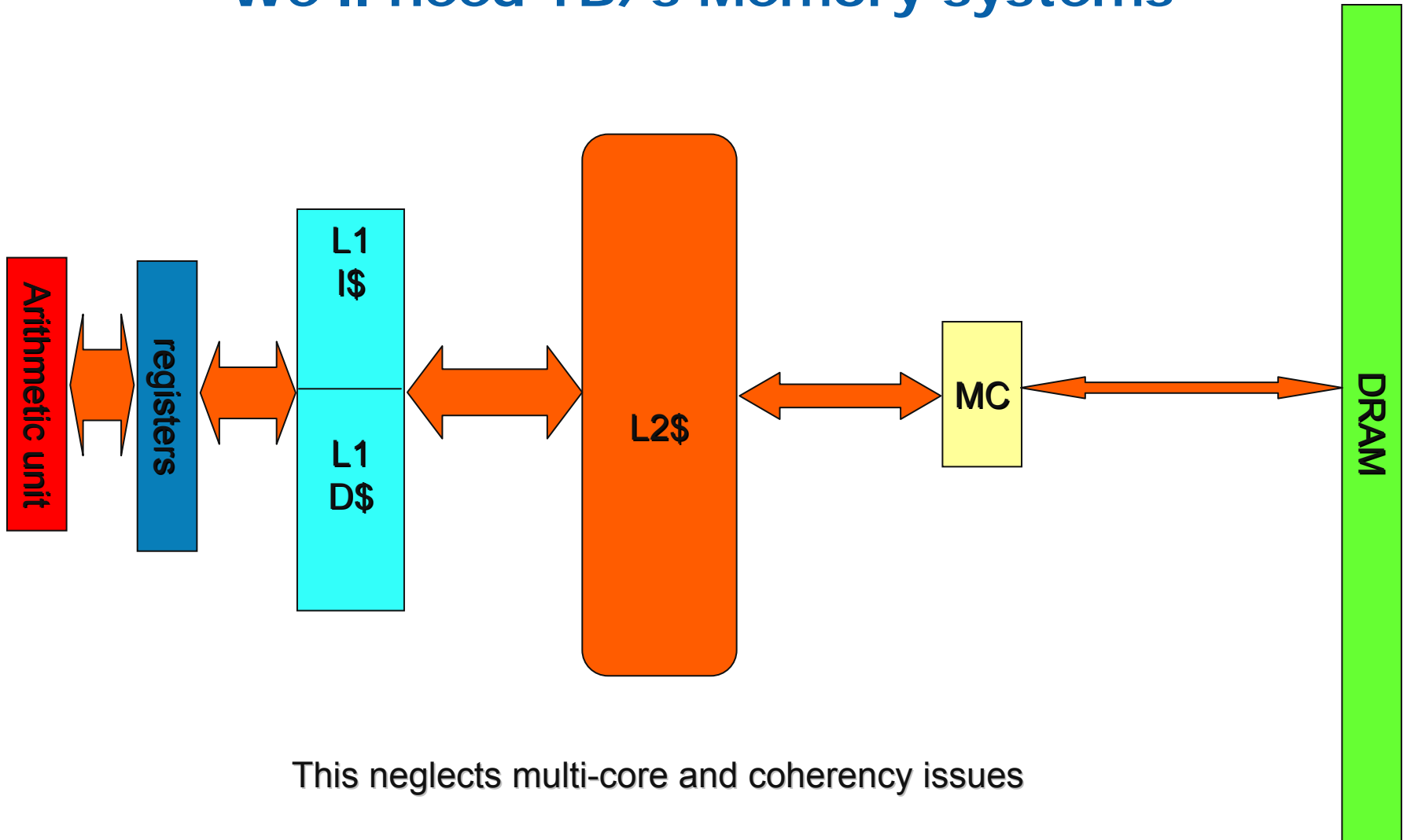
When we have tera-ops processors We'll need TB/s Memory systems

- To do the most common arithmetic operation

$$X \leftarrow A * X + B$$

- Requires that 16 bytes be read into the functional unit
- And that 8 bytes be written
- In addition, an 8 Bytes instruction must be processed

When we have tera-ops processors We'll need TB/s Memory systems



Issues with today's memory system options

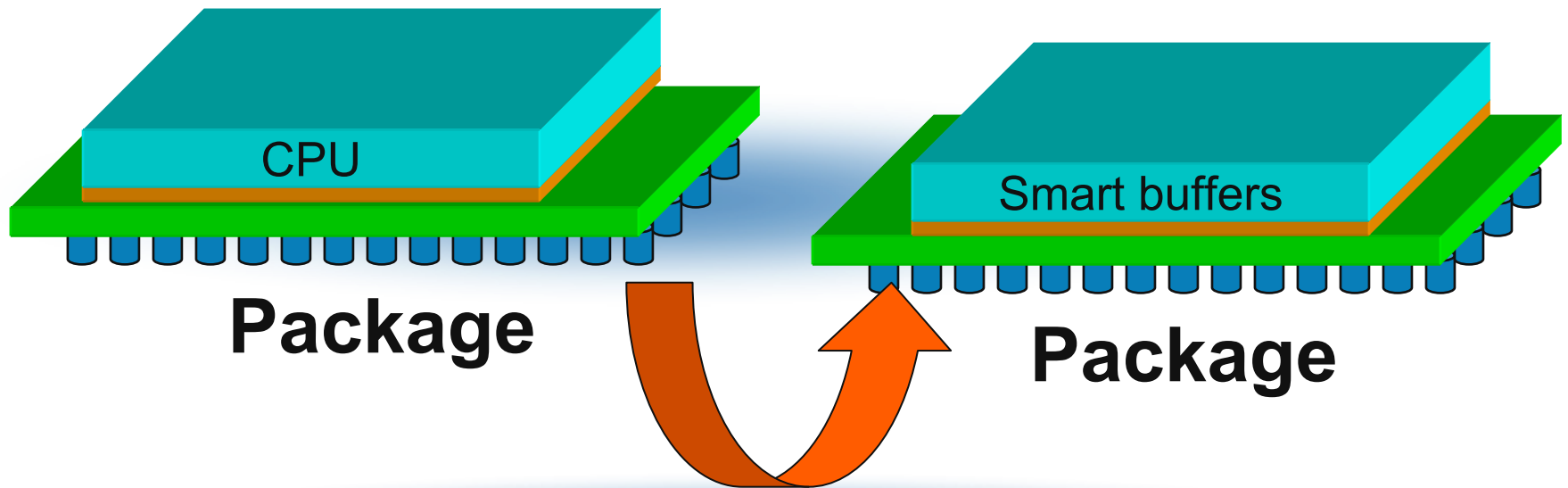
- A *big* processor socket has $O(2K)$ pins
- DDR achieves relatively good BW and relatively low power at amazingly low cost
 - Slow links, lots of them
 - There are simply not enough pins to achieve traditional B/F ratios
- FB-DIMM technology: higher BW, many fewer pins
 - But: much higher power ($\sim 2X$)
 - Longer latencies
 - Cost(?)

Memory system solutions for Terascale processors

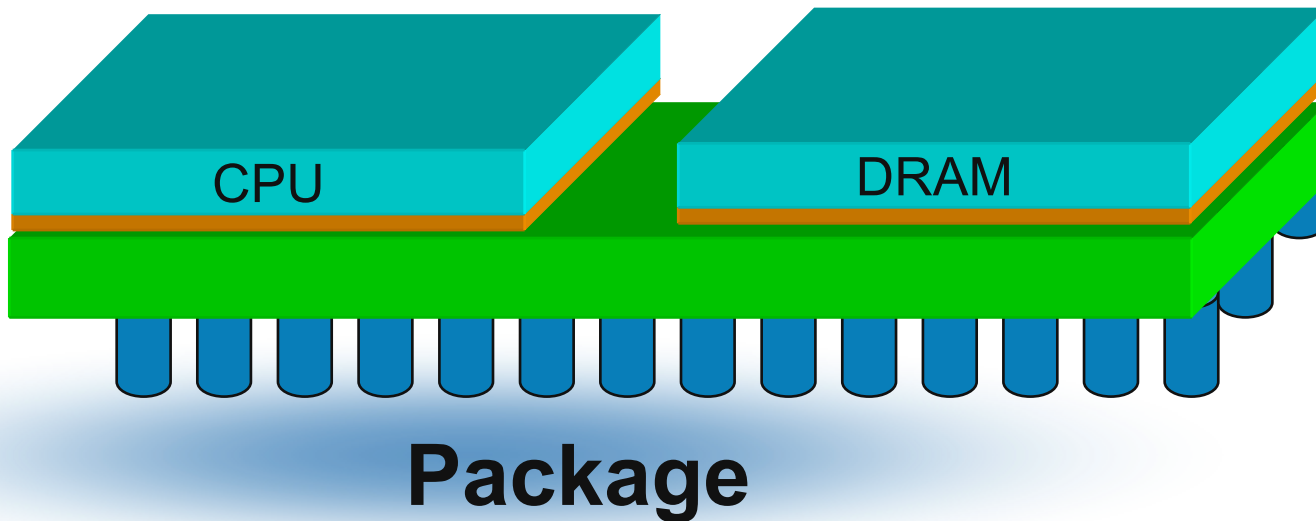
- Buffer on Board
- DRAM on package
- Stacked DRAM
- Silicon Photonics

Memory Bandwidth options

Augment on-pkg MC with very fast links to
An on-board MC



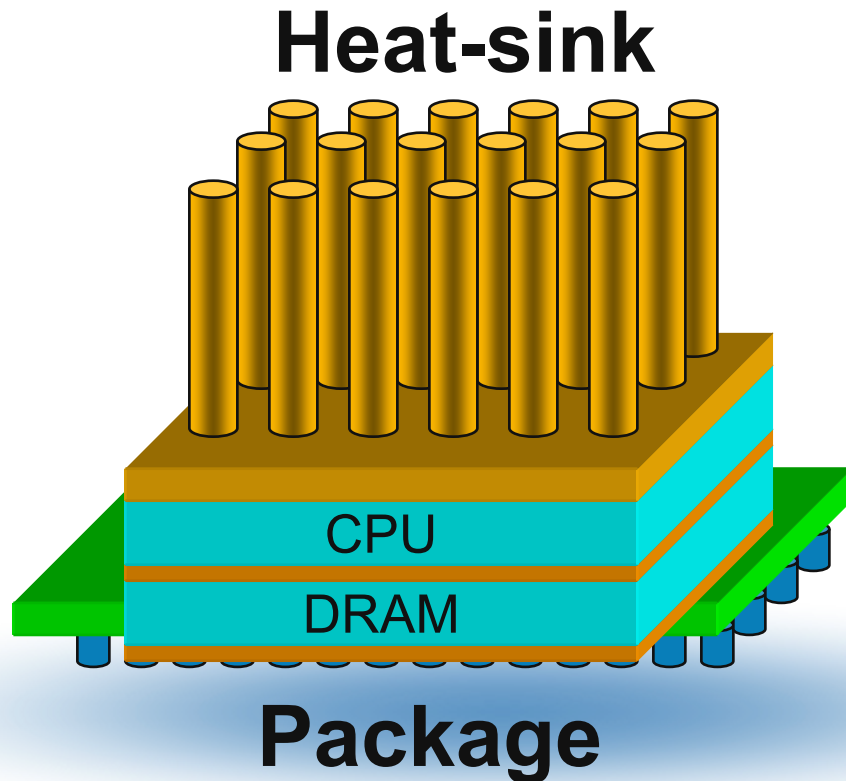
Memory Bandwidth options: DRAM on Pkg



*DRAM, CPU
integrated on die*



Memory Bandwidth futures: 3D Die Stacking



- Power and IO signals go through DRAM to CPU
- Thin DRAM die
- Through DRAM vias

*DRAM, Voltage Regulators, and High Voltage I/O
All on the 3D integrated die*



Silicon Photonics Future I/O Vision

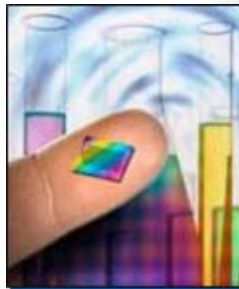
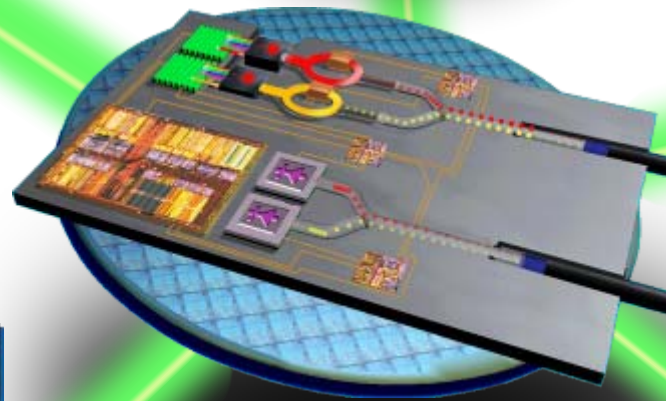
**HPC and
Data Center
Fabrics**



**Chip-to-Chip
Interconnects**



**Backplane and Display
Interconnects**



**Chemical
Analysis**

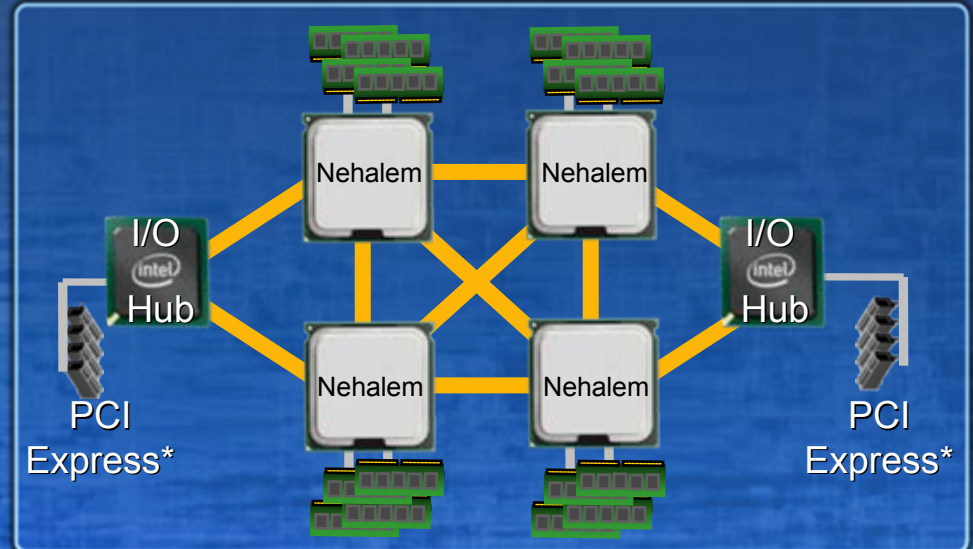
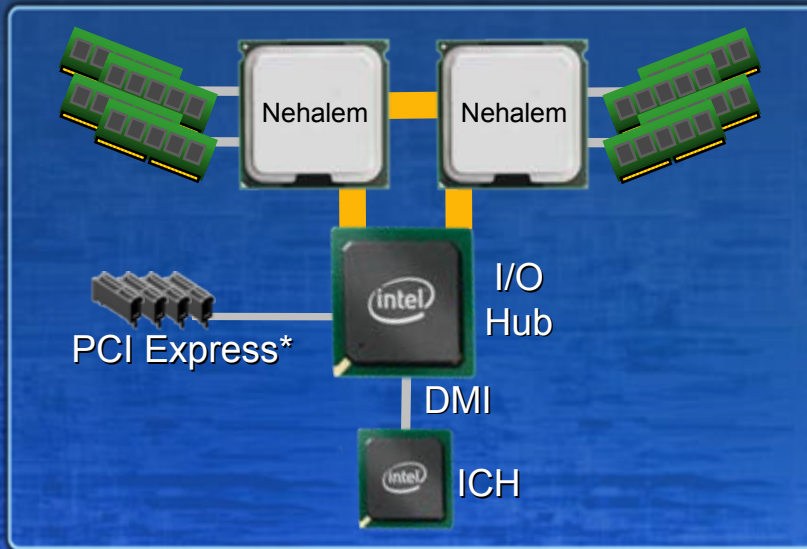
**Medical
Lasers**



*Research Challenge:
Intra-chip and Inter-chip I/O Architecture and Topology Options*



Nehalem Based System Architecture



Intel QuickPath Interconnect

2, 4, 8 Cores

4, 8, 16 Threads

Intel® QuickPath Architecture

Integrated Memory Controller

Buffered or Un-buffered Memory

*Optional Integrated Graphics

Interconnect

- The bigger the system, the more critical the interconnect:
- Blue Gene* has nearly 3X the peak speed of XT-3/Red Storm*;
- RS outperforms BG on most apps

WHY?

- ✓ It's about the interconnect...

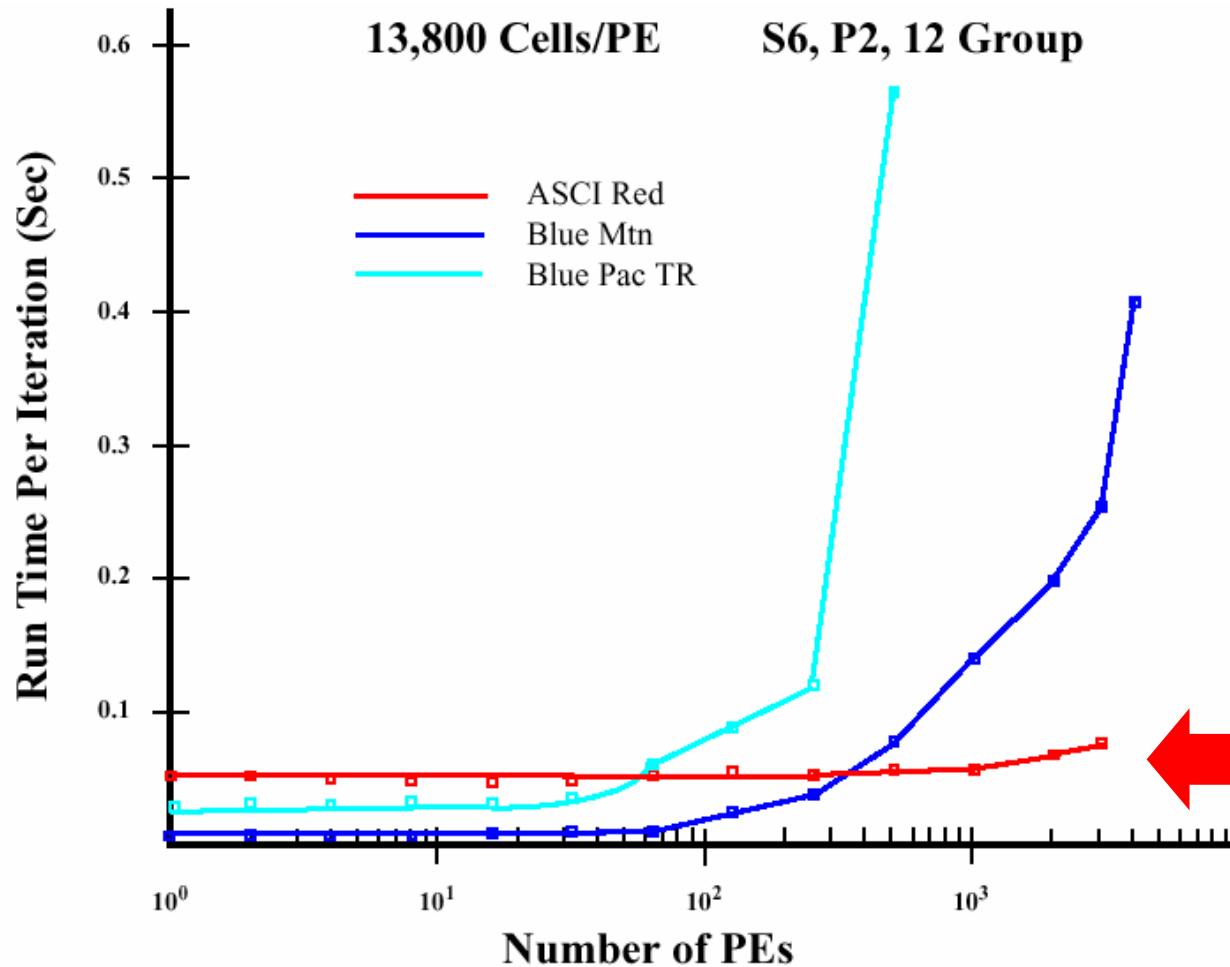


Interconnect

- For large systems, meshes/tori with fat links are best
- Rule of thumb:
 - I/O speed (B/S) ~ processor speed (Flops)
- Terascale Processor:
 - Rule of thumb cannot hold currently
 - signaling at more than 10 Gbits/s/wire pair = Grand Challenge
 - Pin counts, Power & Space budgets are limited
 - Optics speeds are also limited: by cost, power, and electrical input

Well-designed MPI codes scale on well-designed interconnects

Example: Scalable Neutronics (Los Alamos)



MPI-based code
Run time only
stays constant
on a well
balanced system



PCIe 2.0

2X PCIe1.0 Bandwidth

Broad IHV Support

Intel® Xeon® 5400 Chipset and
X38 Express Chipset In 2H'07

Nine Cards from Seven
Vendors Working
with Intel's Stoakley
Platform at 5GHz



PCIe 2.0

2X PCIe1.0 Bandwidth

Broad IHV Support

Intel® Xeon® 5400 Chipset and
X38 Express Chipset In 2H'07

PCIe 3.0

2X PCIe 2.0 Bandwidth
Data Reuse
Dynamic Power Management
Atomic Operations

Industry Standard Attach
For Accelerators

Specifications In 2009
Products Expected In 2010

Expanding Momentum and Innovation



Programmability

- Nearly all HPC apps today are written for X86 architecture
- Nearly all use MPI for remote memory access
- OpenMP is successful on SMP nodes
- Most people are anxious to avoid hybrid programming
- So, what do we do about 1—100 Million MPI threads?

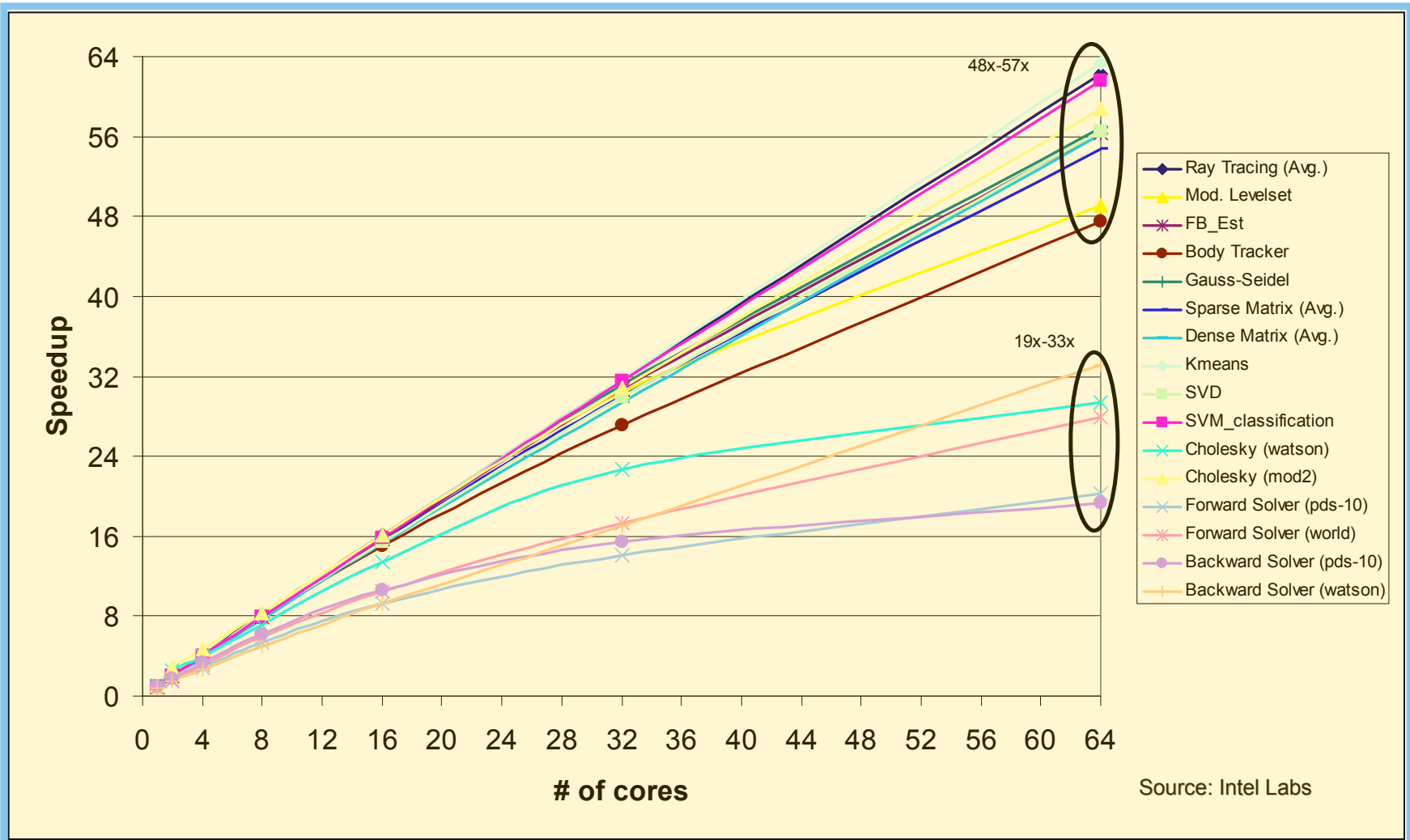


Programmability– biased opinion

- X86 is a huge advantage
- MPI is not going away
- Many new apps would be written using GAS models
 - If the support were there
- Multi-threading/dataflow languages are interesting but have a huge barrier to adoption
- The problem is less programming complexity than achieving performance (jitter, load balance, reliability...)
- Transactional memory will aid in on-socket shared memory parallel programming
- Fractal computing models will become important



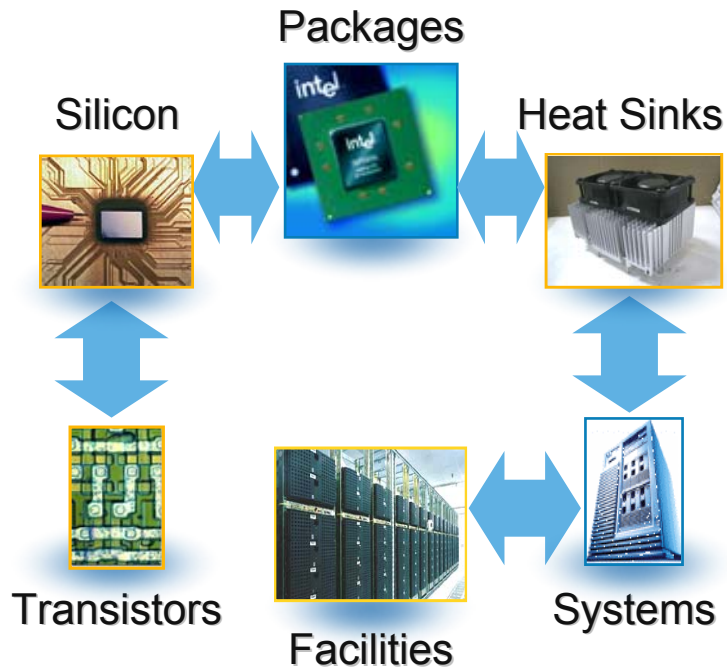
Shared Memory Scalability on Many Core systems



Power

- Speed costs power
- Memory bandwidth costs power
- Memory size costs power
- Power delivery costs power
- Cooling costs power
- Optics costs power
- Electrical signaling costs even more power
- Power envelope and density drives cooling issues
- Green policies
- **There is no single magic bullet**

System Power/Cooling Efficiency



Silicon:
Moore's law, Strained silicon,
Transistor leakage control techniques,
Clock gating, use more Si to replace
faster Si

Processor:
Policy-based power allocation
Multi-threaded cores

System Power Delivery:
Fine grain power management,
Ultra fine grain power management
High efficiency power converters

Facilities:
Air cooling and liquid cooling options
Vertical integration of cooling solutions

*Research Challenge:
"Zero-overhead" HW & SW solutions
For system and facility level power management*



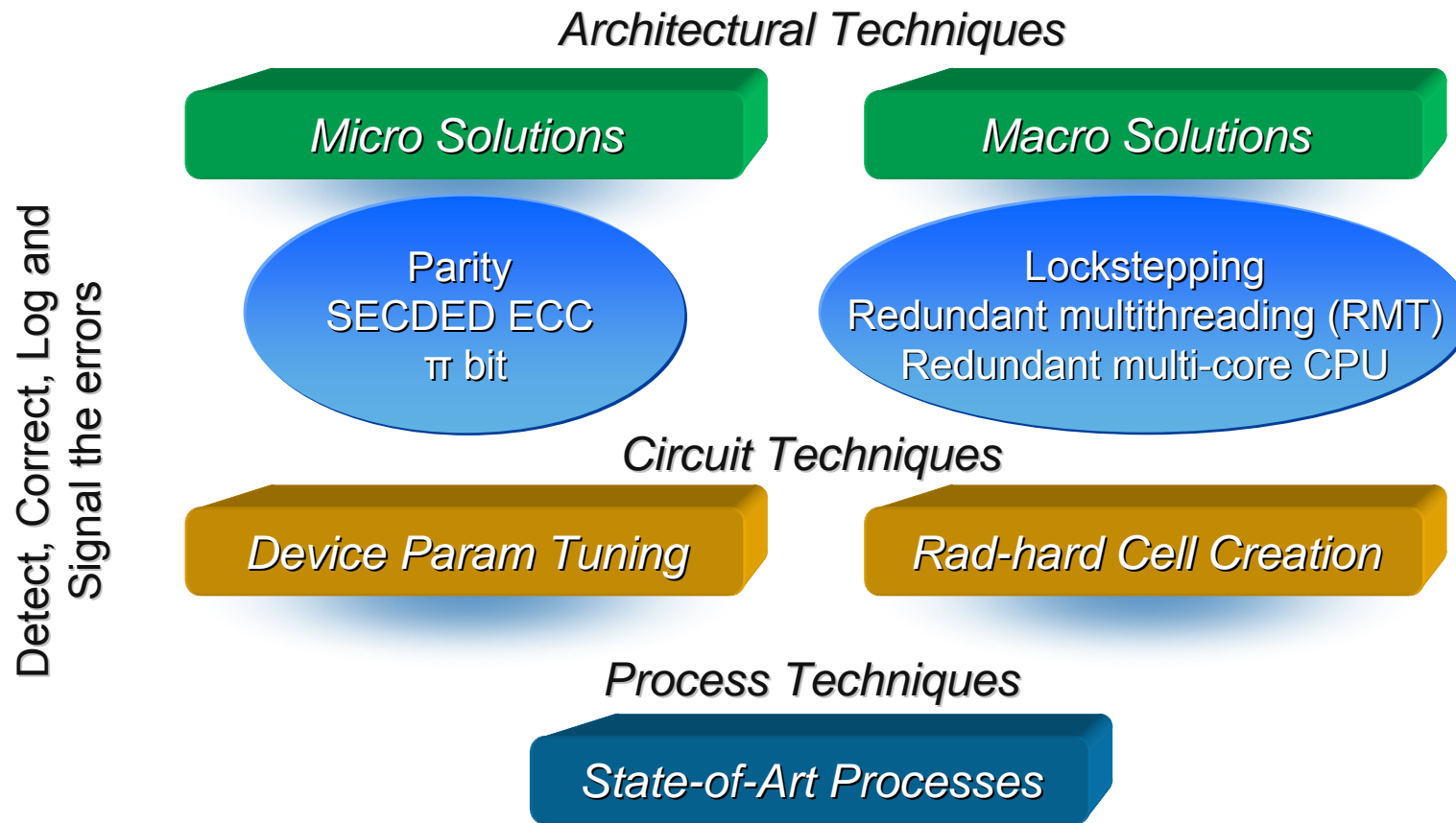
Reliability:

Reliable Answers From Unreliable Components

- Petascale-to-Exascale systems are huge
- With current technology, ~ 3—5X increase in part count over today's biggest systems
- 3—4X the number of SW instances
- Exquisite efforts needed to keep parts cool
- Engineered-in reliability is not an option; it is an imperative
- SW complexity and brittleness will remain the largest cause of unreliability in HPC
- 20 Hour MTBI is a realizable goal for multi peta-ops systems



Reliable Systems with Unreliable Components



Research Challenge:
From software (Apps, OS, VMM, etc.) to hardware
a reliable Petascale HPC system needs management top down



Summing it up:

Intel will pave the road to Exascale

- Intel processors will provide a step-function advantage in
 - Programmability
 - Performance and reliability
 - Cost of performance
- Intel will bring its world-leading technologies to bear
 - Si process
 - Many-core
 - Memory technologies
 - Optics
 - Power reduction and mgmt
 - Productivity through SW
- Intel will architect fastest-in-the-world systems

SRW (not Intel) Prediction:

Red River Shootout 2007

Sooners: 23 Longhorns 10





HPC @ Intel