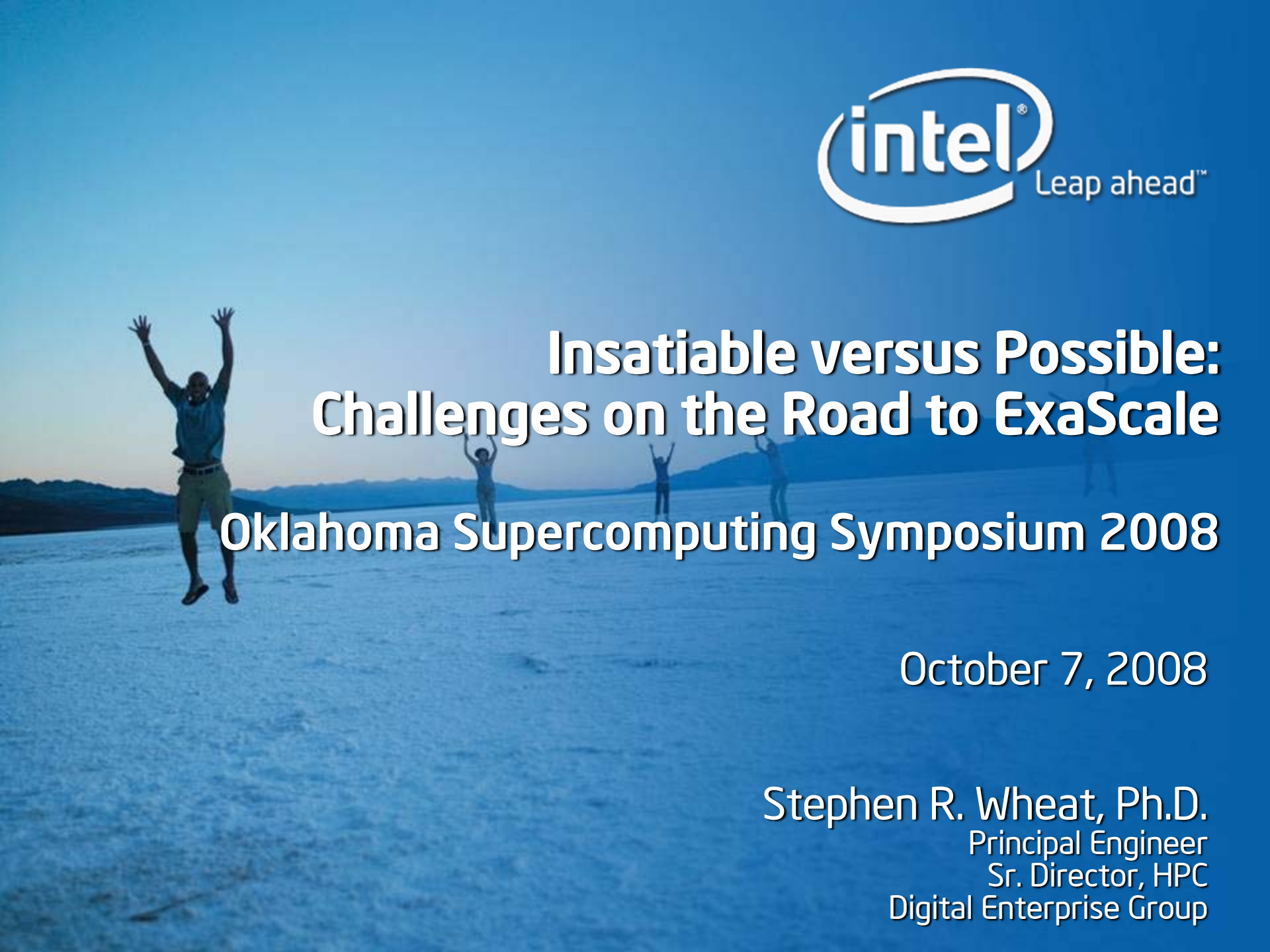


Insatiable versus Possible: Challenges on the Road to ExaScale

Oklahoma Supercomputing Symposium 2008

October 7, 2008

Stephen R. Wheat, Ph.D.
Principal Engineer
Sr. 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).

Since we last met, …

EDN.com

Intel, Cray to Develop Supercomputing Technologies

By Ann Steffora Mutschler
Senior Editor

The companies will explore future supercomputer component designs such as multi-core processing and advanced interconnects, with the goal of developing a range of high performance computing systems over the next several years.

Cray's President and CEO Peter Ungaro (Right) and Kirk Skaugen, Intel vice president and General Manager of the Server Platforms Group

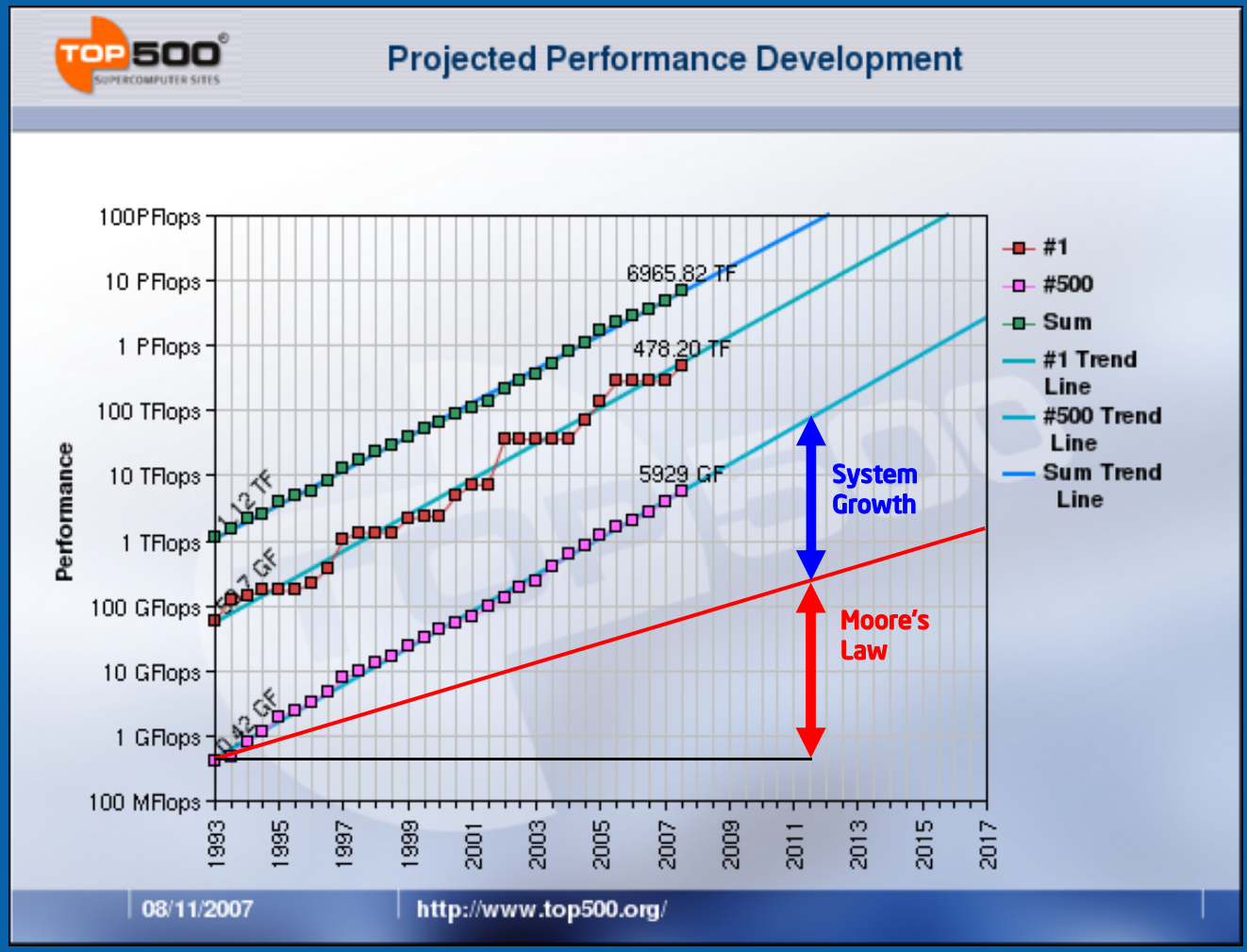
NASA, Intel, SGI Plan to 'Soup Up' Supercomputer

MOFFETT FIELD, Calif. – NASA, Intel Corp., and SGI today announced the signing of an agreement establishing intentions to collaborate on significantly increasing the space agency's supercomputer performance and capacity.

Continuous Growth of HPC Systems

Demand for performance of HPC systems outruns Moore's Law

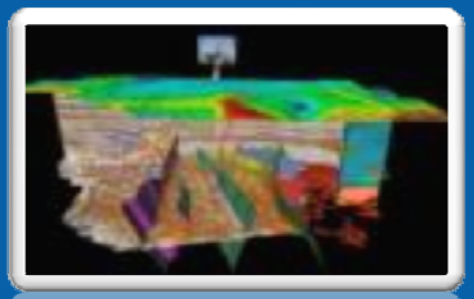
- CPU performance increases by Moore's Law
- To reach higher system performance, number of CPU's has to increase



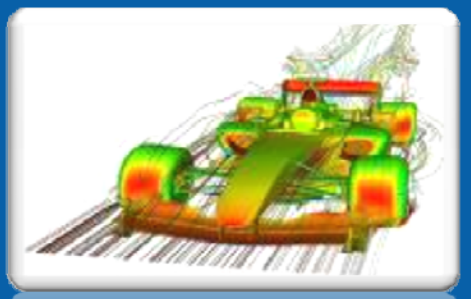
High Performance Computing

Insatiable Demand for Performance

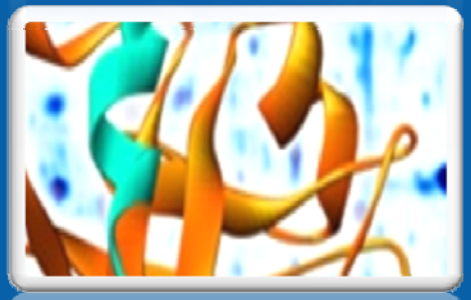
Weather Prediction



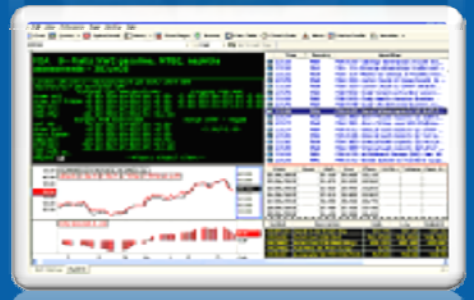
Oil Exploration



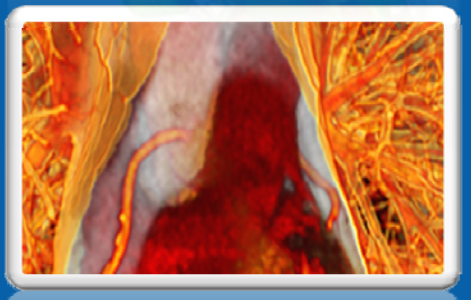
Design Simulation



Genomics Research

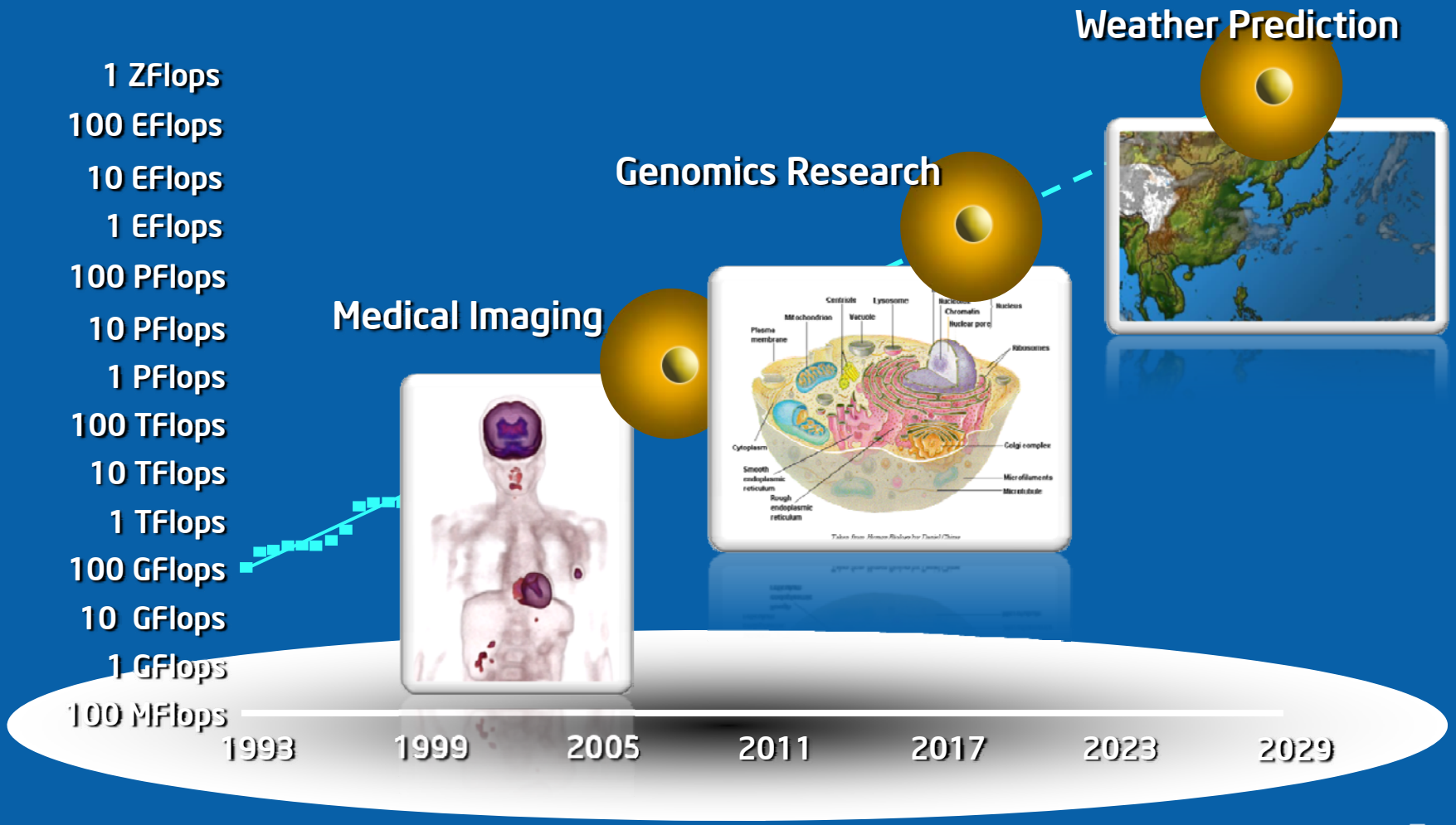


Financial Analysis



Medical Imaging

HPC Needs Decades of Moore's Law



Data Source: www.top500.org

Some System-based Processor Development issues

Thread count:

$N_{threads}$ ~ Required Speed / S_{THR} (GigaFfORS)

Power:

P(MW) ~ 40--100 * Peak Speed (Exa-ops)

Reliability:

50,000 cooperating sockets => effective FIT rate needs to be 50,000 better than that needed at the desktop

Scalability:

→ Higher BW, lower latency Interconnect

What kind of cores and how many cores?

Core size (at constant 200—300 Watts/pkg)

Many (~1--2000) small cores:

- Think of using low power (Say mobility) x86 Processors and putting many of them on a die.
- In-order, SSE-n with 2, 4 or perhaps 8 ops per clock at 2-to-4 GHz clock
- Many threads per core

A bunch of really big cores (512):

- Mobility core + very wide (32 ops) Vector units
- 4-8 threads per core
- 2-4 GHz

Memory solutions

- We cannot cost-effectively match memory bandwidth to off-package memories with the increase in processing speed.
- We are hitting a wall in terms of number of pins, signaling area, and signaling power.
- And, memory speeds are advancing more slowly than CPU speeds (so the problem gets worse over time).

Machine Density (Assume advanced cooling technologies)

We expect to be able to put 2-4 sockets on a board;

We expect 8 boards per cage and 4 cages per cabinet.

64--128 sockets per cabinet

2--4 peta-ops per cabinet

400—800 cabinets for a peak exa-ops machine (plus storage and I/O cabs)

200—400 KW per cabinet.

Key Investment Areas

- Hardware technology development with R&D assistance from the high-end community
- New applications development and advanced programming paradigms
- Advanced operating and programming environments, scalable languages, compilers, OS's, and libraries

Technology Gaps

Processor architecture and speeds → 36 TF

Memory Hierarchy speeds → 18 TB/s

Interconnection Network speeds → >4 TB/s

System-level resiliency and reliability

I/O scalability and flexibility

Power and Cooling

Programming models and environments

Applications scalability and resiliency

Operating environments

Over-Arching Requirements

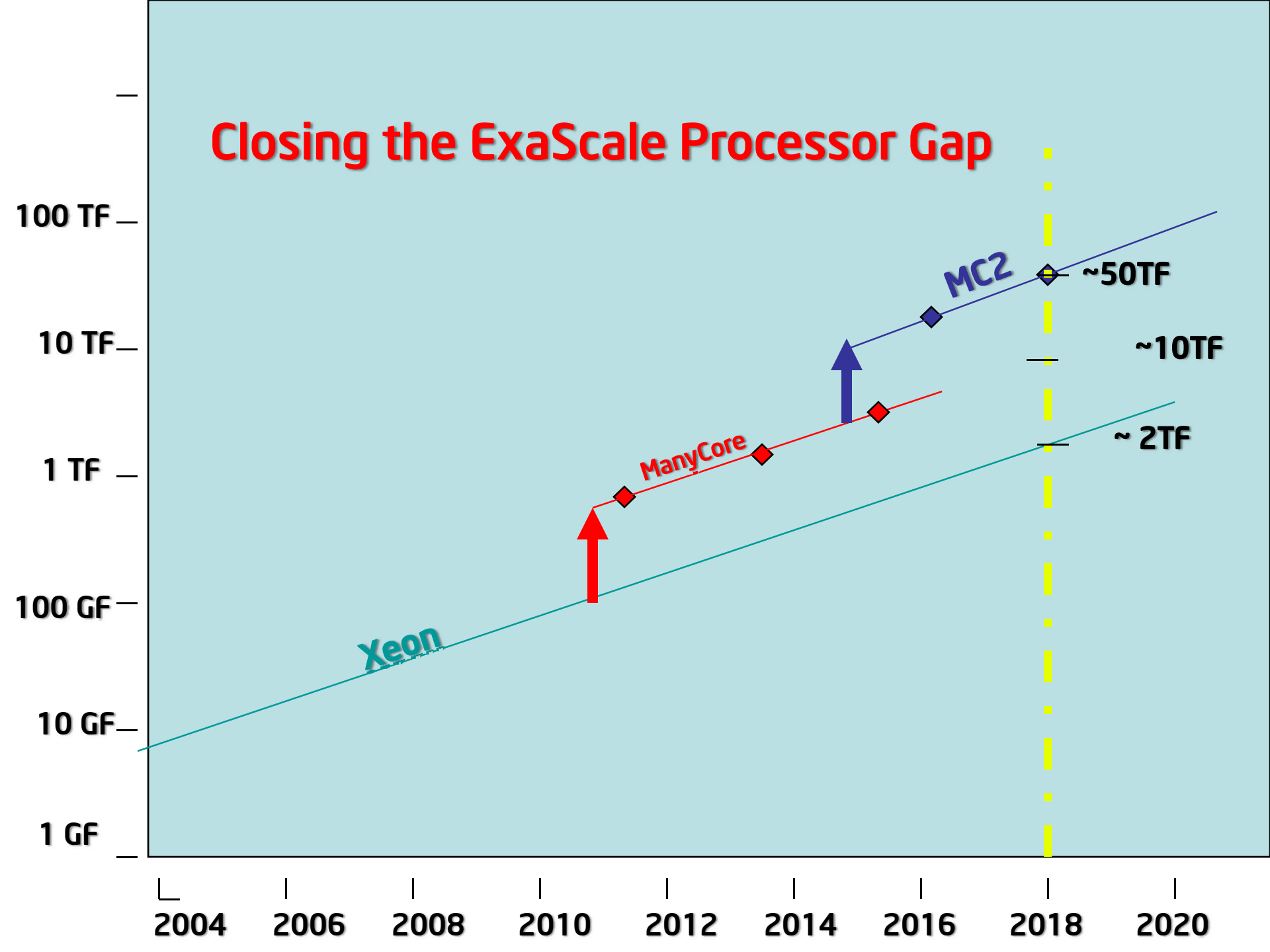
Affordability:

- Total Cost of Ownership
- Capital Costs
- Power costs
- Operating costs

Longevity

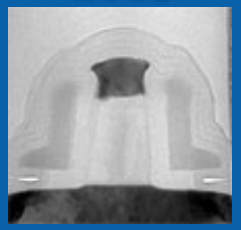
Ease of scaling system size up or down

Breadth of Applicability to HPC Community Needs

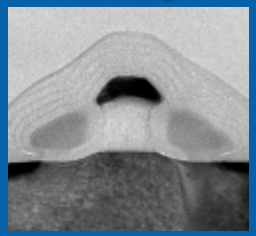

Closing the ExaScale Processor Gap
100 TF
10 TF
1 TF
100 GF
10 GF
1 GF
MC2
~50TF
~10TF
~ 2TF
ManyCore
Xeon
2004
2006
2008
2010
2012
2014
2016
2018
2020

Silicon Technology Leadership
Intel Execution on a 2-Year Cycle

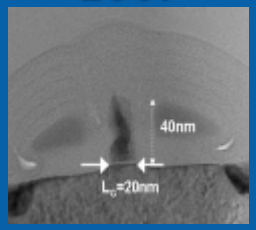
90nm
2003



65nm
2005

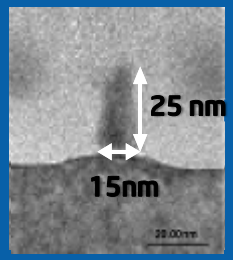


45nm
2007

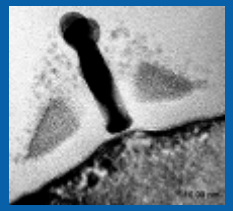


2005 - 2012

32nm
2009



22nm
2011



16nm
2013

11nm
2015

8nm
2017

2013 - 2017

New Intel technology generation every 2 years
Intel R&D technologies drive this pace well into the next decade

Roadmap

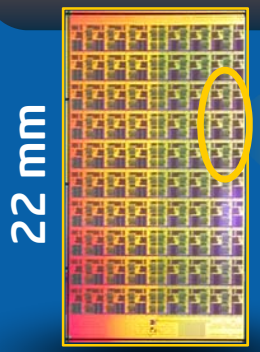
Research

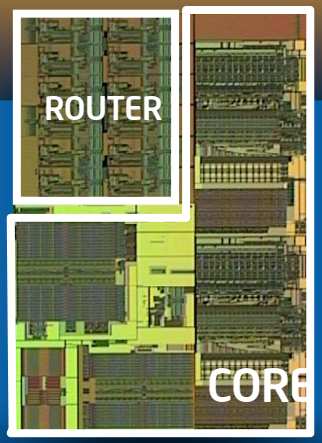
All products, computer systems, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

Shaping the HPC Platform of the Future

Tera-Flops: Polaris Prototype

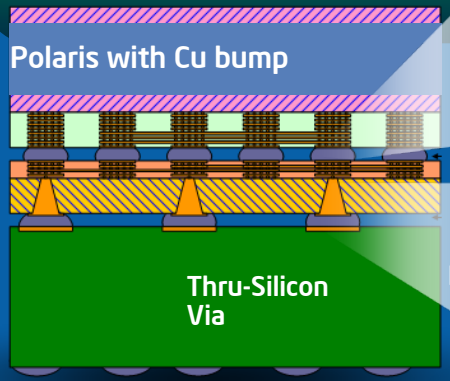
80 Cores
1 TFLOP at 62 Watts
256 GB/s bisection



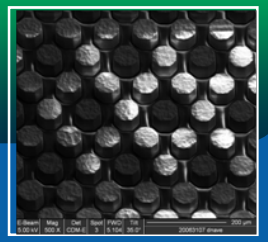


Tera-Bytes: 3D Stacked Memory

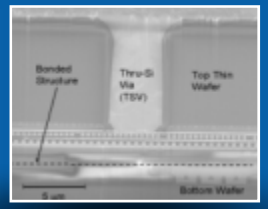
256 KB SRAM per core
4X C4 bump density
3200 thru-silicon vias



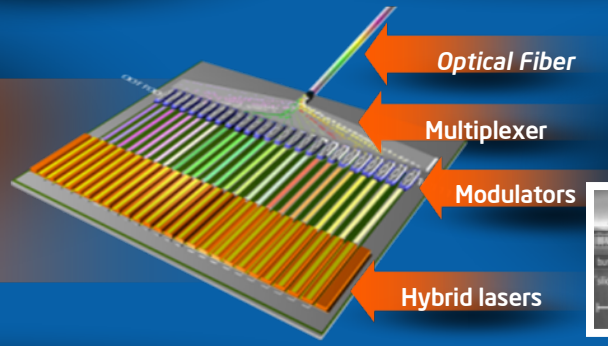
Polaris



Memory Package



Tera-Bits: Si Photonics Research

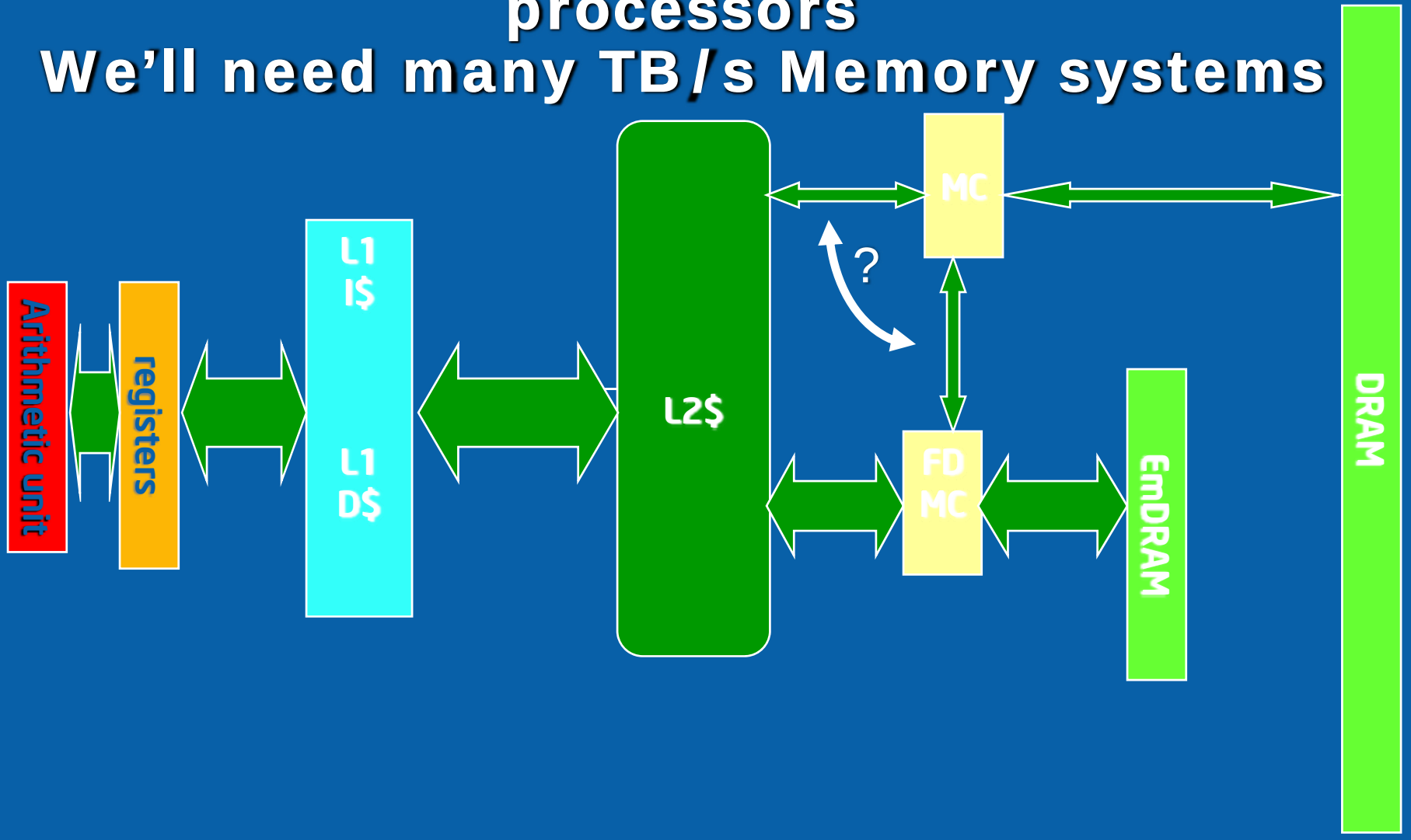


40 Gbps Modulator

Memory System

- ~1/2 Byte/sec per flop is the minimum MBW needed for Xeon.
- DDR-n and FBD memory technologies will (almost) keep up with Xeon.
- They are inadequate for “step-function” and will become progressively less adequate
- Only a richer memory hierarchy can bridge the gap

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



This neglects multi-core and coherency issues

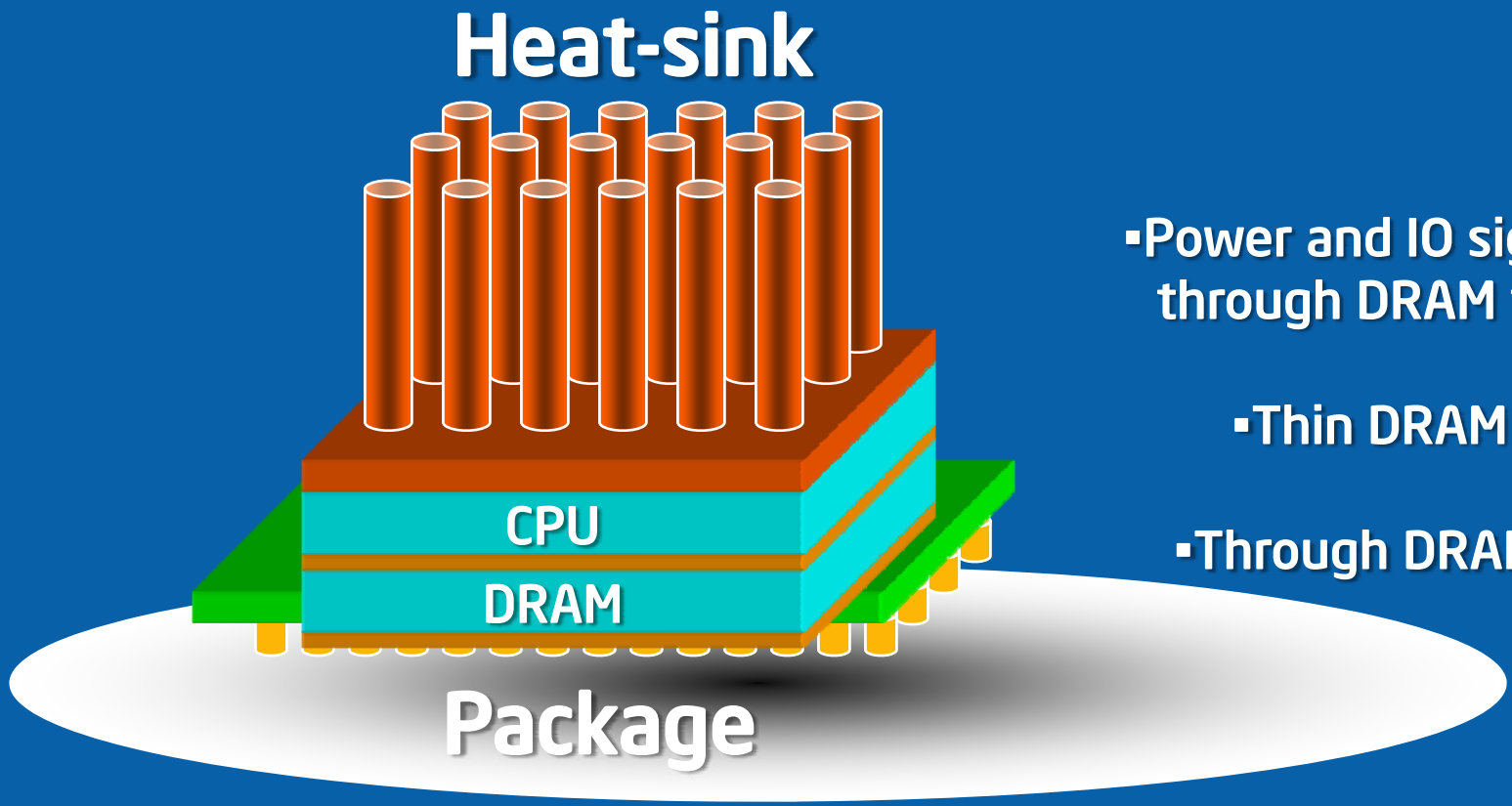
Possible technologies

- We can create fast “small core” memories using one of several embedded DRAM technologies
- We may be able to configure part (~128—512MB) of this small core memory as a third level cache
- We can put intelligent controllers and buffers on the small core memories
- We can put “large-core” memory (DDR-n DIMMs) ”behind” the small core memory

Possible technologies

- Embedded DRAM
  - 1st Gen:
    - Minimum interesting BW requirement for EmDRAM is 200 Gigabytes/sec.
    - Minimum size is 512 MB
  - Next-gen:
    - Minimum BW requirement for EmDRAM is 400 Gigabytes/sec with stretch goal of 800 Gigabytes/sec.
    - Minimum size: 2 GB, stretch goal 8 GB

MIP$_1$: Memory Bandwidth futures: 3D Die Stacking

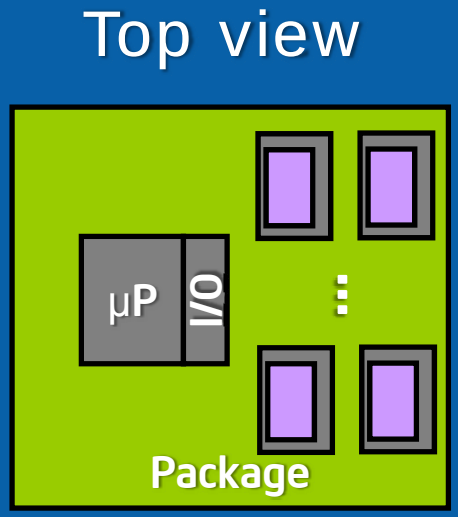


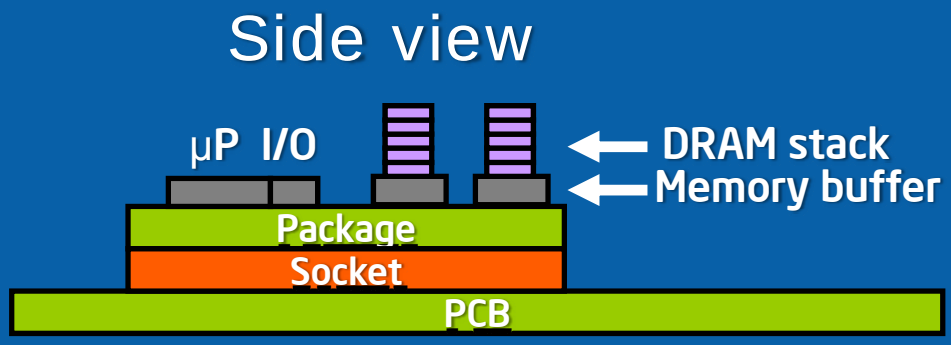
- Power and IO signals go through DRAM to CPU
- Thin DRAM die
- Through DRAM vias

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

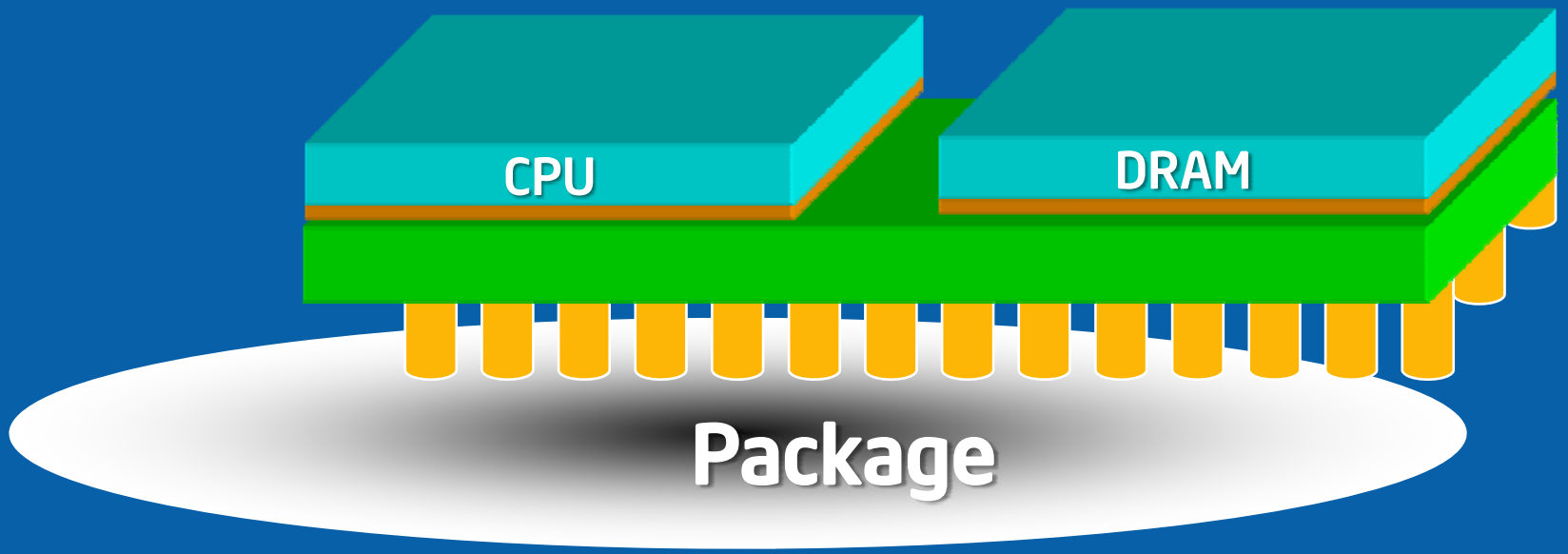
3-D on-package Stacked Memory

This is the traditional approach to stacking memory on package





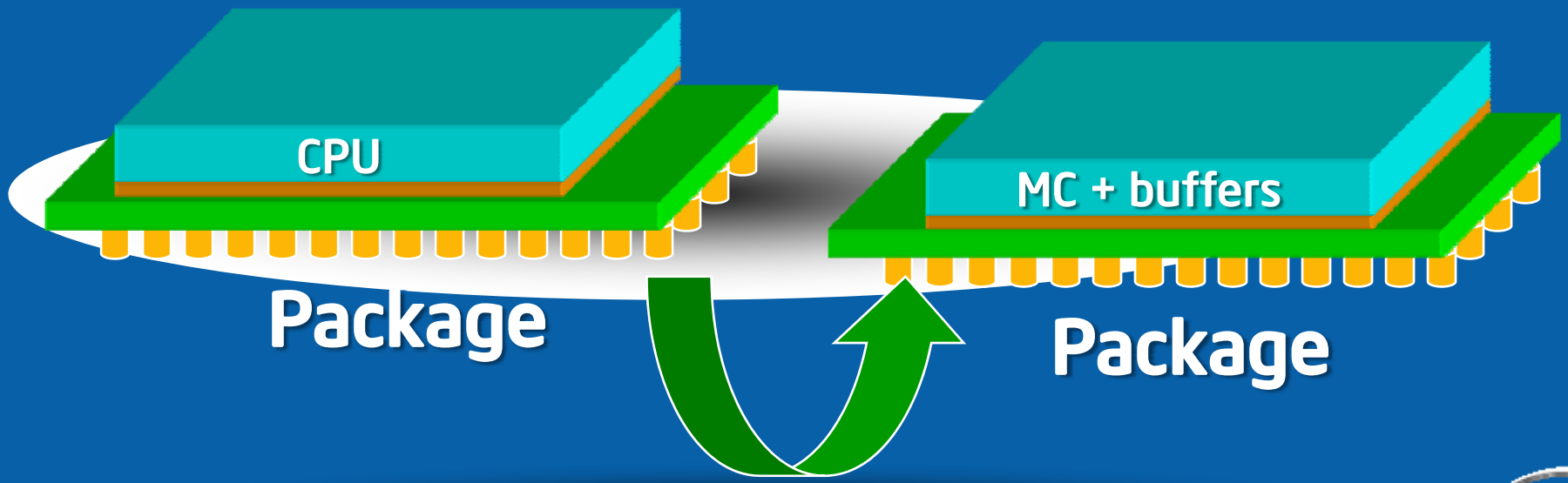
MIP_2: Memory Bandwidth options: DRAM on Pkg



DRAM, CPU integrated on die

MIP_3: Memory Bandwidth options

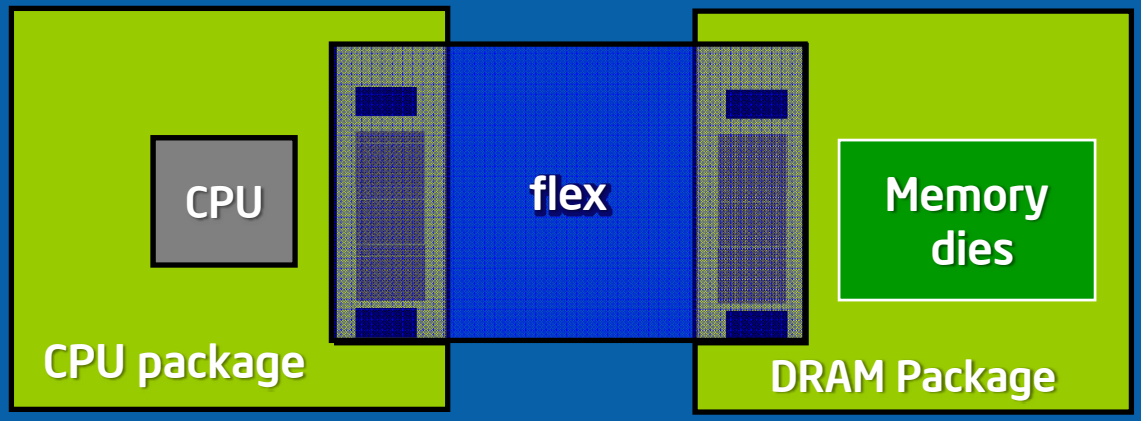
Replace on-pkg MC with very fast flex links to
An on-board MC



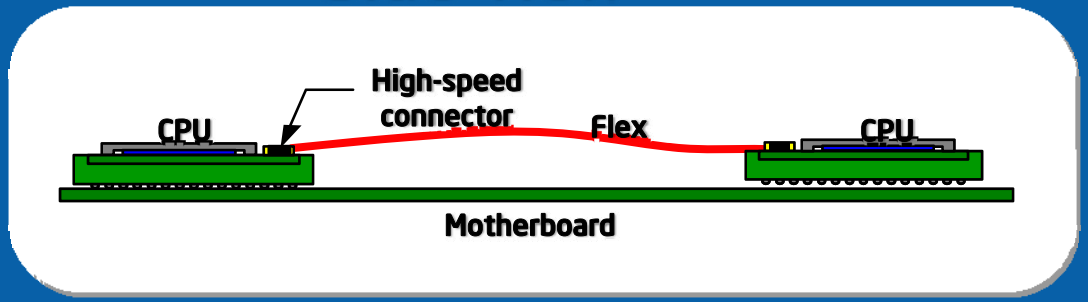
Flex-connected memory

- Very high-quality electrical channel
- Simple pre-emphasis at Tx
- No equalization required at Rx
- High bit rate at very low power

Top view



Side view



Aggressively Stacked DRAM

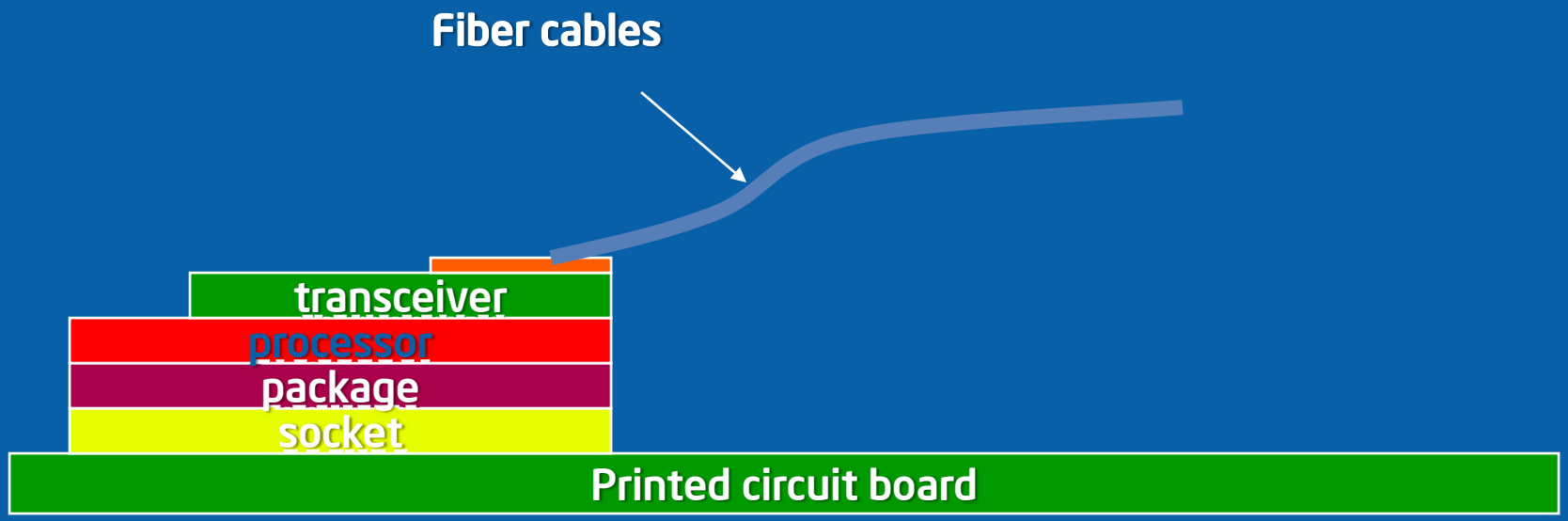
In the next 5-8 years, stacked DRAM can become cost competitive if the product development is carried out soon enough.

This could allow as much as 16—64 GBytes of EmDRAM at 800+ GB/s in that timeframe.

On-Die Silicon Photonics

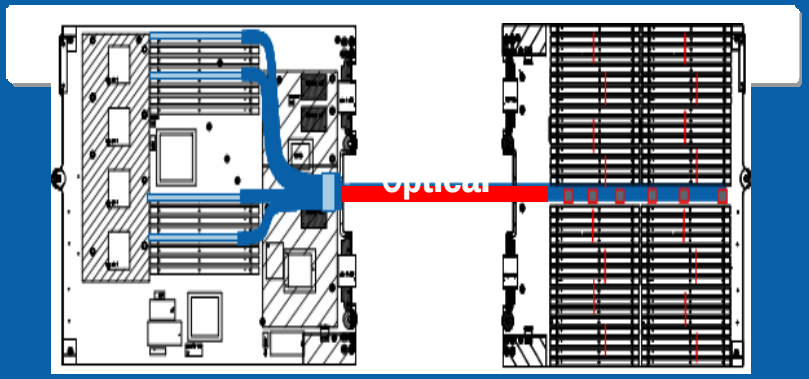
- Also, in that timeframe, Si-photonics can become a valid option.
- Transceiver chip bonded to processor (or router– discussed later).
- Efficiency increases significantly
- Higher bandwidth an option: Going through 1TB/s and 4TB/s milestones in the following years

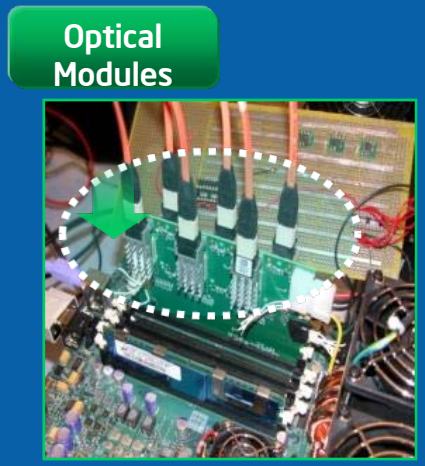
DIE-bonded Si-photonics



Experimental Optical Memory Link

Proof of Concept: Latency Impact and Initialization Protocols
Future Work: Optical Packaging and Integration





- Blade form-factor has near memory capacity constraints
- Need large remote memory with near memory latency attributes
- Electrical solutions are power and wire hungry
- Optical links offer a scalable solution with minimum latency impact

Power and Cooling

In CMOS, we cannot avoid the physical facts that power is not shrinking as fast as feature size:

- Leakage is actually increasing
- Voltage scaling will asymptote at >1/2 volt
- Fundamental operations will require about 10 pico-joules for CPU ops and about 40 pico-joules for memory ops

Power and Cooling

- Most technology assessments find power levels in the hundreds of MW for general purpose Exa-ops systems in 2018.
- With right-sized memory and aggressive design technologies we may get peak dissipations down to ~100MW in 2018 for a 1.6 Exa-ops peak system.

Power R&D

Mainly around

- power gating technologies in Silicon
- Advanced power delivery– high efficiency transformers
- VRM on pkg
- Clock derating
- Technologies that minimize idle/refresh power

Cooling R&D

Such an exa-ops system will be liquid cooled

Integrated microchannel cooling in MCM?

Affect of operating temperature on power & performance?

- lower temps =
  - Higher clock speeds and less leakage
  - Higher reliability
  - Improved longevity

Cooling R&D

Study transients (shutdown) at 100, 200, 400 kW/rack

MUST lock down kW/cabinet to define where to study.

- Depending on MFlops/W we see numbers from 40kW/rack to 800kW/rack!

Resiliency

As system size grows, feature sizes shrink and the number of parts proliferates:

- Reliability becomes ever more challenging.

Resiliency

- Advanced, independent RAS networks probe every part continuously; predictive failure analysis
- No rotating storage (may want NAND or PCM)
- Redundancy where possible (Powersupplies, VRMs, ...)
- ECC/CRC/End-to-End checking--retry everywhere
- Transactional Memory
- Radiation hardening of key latches--flip/flops
- Redundant multithreading
- Residue Checking
- Built-in spare cores
- Aggressive scheduled maintenance

Even So …

Even if all of the hard work yields an ExaOps system at 100MW

that means a PetaOps system would be 100KW

Programming is another matter …

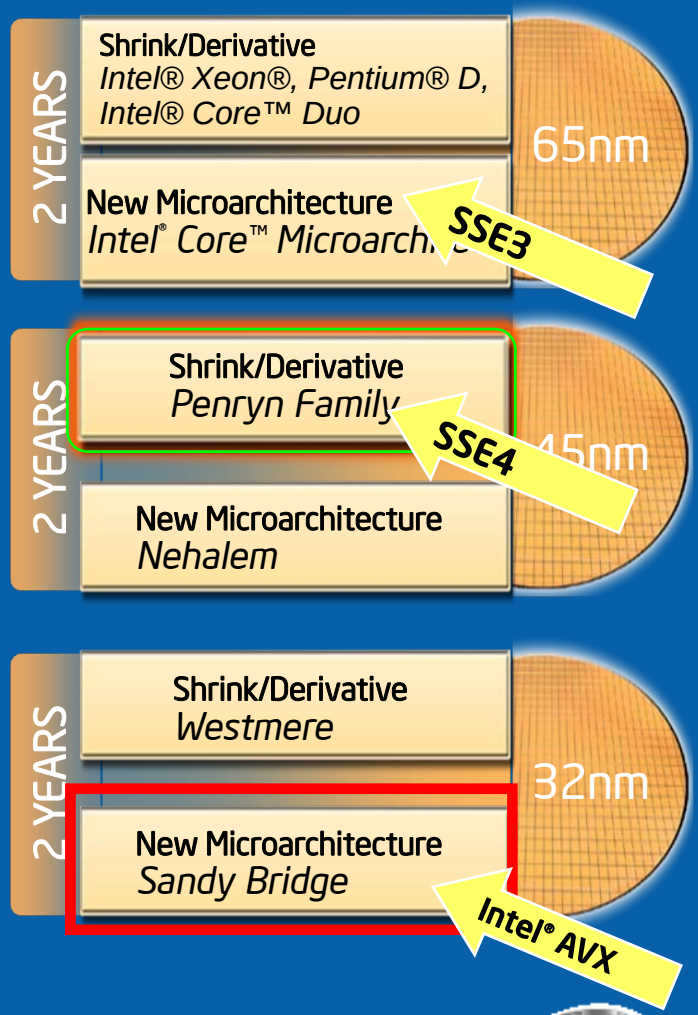
Intel innovations have been setting the pace for numeric computations

Evolution of Instruction Set Architecture (ISA) :

- 1979 - x87 Numeric data extensions
- 1997 - MMX™ - 64b SIMD ISA extension
- 1999-2007 – SSEx - 128b SIMD ISA extensions
- Starting 2010 – Intel® AVX – 256b SIMD ISA extensions

Benefits of Intel's ISA

- Available across a wide range of platforms and integrated into the platforms
  - No additional hw required
  - Investment protection for your development costs and effort
- Backward and forward compatible to future ISAs
- A suite of tools and compilers to make development easy

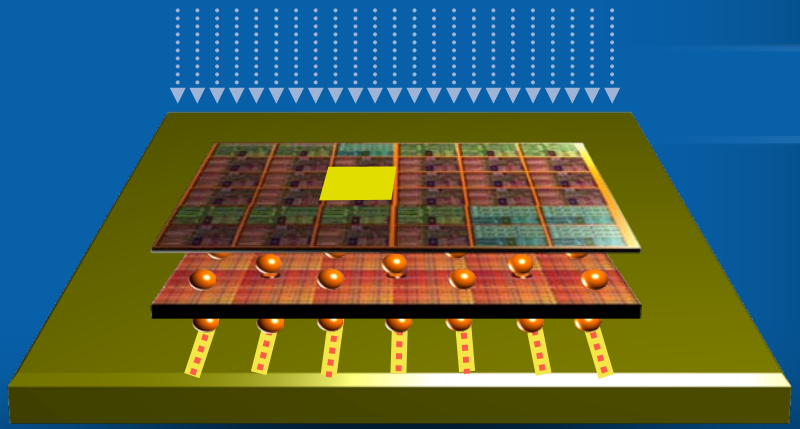


The Challenge
Parallel Programming

Irregular Patterns, Data Structures and Serial Algorithms

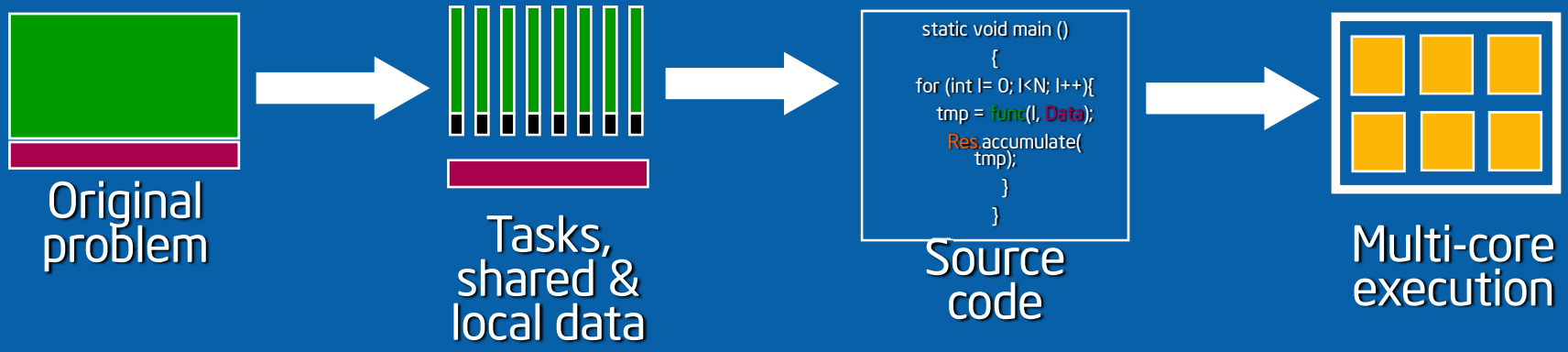
?

Scale to Multi-Core Today → Hard
Scale to Many-Core Tomorrow → Harder



Increasing Cores (2→64+ Cores)
Vector Instructions (4→8+ Wide)
Cache and Interconnect Latency

PARALLEL PROGRAMMING CHALLENGES – Leadership research @ Intel



- Extracting concurrency
- Expressing concurrency
- Exploiting concurrency

KEY IS TO IDENTIFY (algorithmic, manual) & MANAGE (locks, conditionals) PARALLELISM

Intel® Software Tools for Parallelism

Architectural Analysis

Visualization of parallel (threaded or MPI) application execution and communication behavior – give valuable insights for application architects and programmers

Introduce Parallelism

Highly optimized OpenMP* and MPI library and run-time system for scalable solutions
MKL threaded and distributed mathematical library

Confidence / Correctness

Detecting actual and potential Threading and MPI programming and API issues - to address challenges unique to parallel programming

Optimize / Tune

Valuable insights for performance and scalability tuning of threaded and MPI applications

Parallelism Is The Key To Performance

Intel® Parallel Advisor (design)

Intel® Parallel Composer (code)

Intel® Parallel Inspector (debug)

Intel® Parallel Amplifier (tune)

www.intel.com/go/parallel

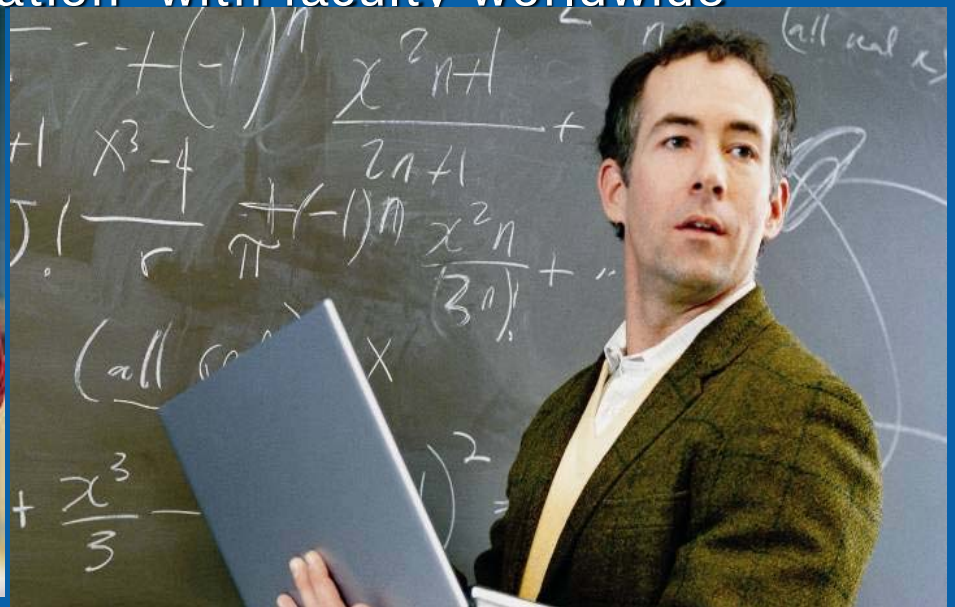
Performance *and* **Forward Scaling** *to* **Many-Core Processors**

Developing Tomorrows Talent

Intel Software College *MC Curriculum Initiative*

Preparing the next generation of software professionals for Multi-core and Many-core platforms

- Expanding computer science curriculum to include parallel programming
- Over 800 Universities participating to date; expanding to 1000+ by EOY
- Web community enables worldwide participation and discussion
  - Forums, blogs and Wiki with curriculum
- Communication with faculty worldwide



Intel Software Network

Multi-core Developer Community

Meet Intel Technologists

- Blogs, wiki, forums, video

Test Your Skills

- Game Developer Contests
- Threading Challenge

Prepare for the Future

- MC Curriculum Program
- New Academic Community

A Comprehensive Source for Parallel Programming
150k users per month
50% growth this year in users

Red River Shootout - 2008

Last year, I predicted:

Sooners 23, Longhorns 10

Reality:

Sooners 28, Longhorns 21

This Saturday

Sooners 16 (1TD, 3FG)
Longhorns 9 (1TD, 1FG)

And,
ASU upsets USC


intel
Leap ahead™
HPC @ Intel