

An analysis of the impact of Heterogeneous Many-core Computing



Simon McIntosh-Smith
University of Bristol, UK



Alternative titles

The divergence of modern
computer architectures

“May you live in interesting
times”

Agenda

- Important technology trends
- Heterogeneous Computing
- Coming hardware discontinuities
- The Seven Dwarfs
- Important implications

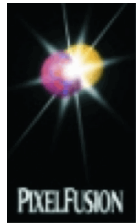
A brief biography



Graduated as Valedictorian in Computer Science from Cardiff University in 1991



Joined Inmos to work for David May as a microprocessor architect



Moved to Pixelfusion in 1999 – a high-tech start-up designing the first many-core general purpose graphics processor (GPGPU)



Co-founded ClearSpeed in 2002 as Director of Architecture and Applications



Joined the CS department at the University of Bristol to focus on HPC and advanced computer architectures



🌟 The real Moore's Law

Moore's Law graph, 1965

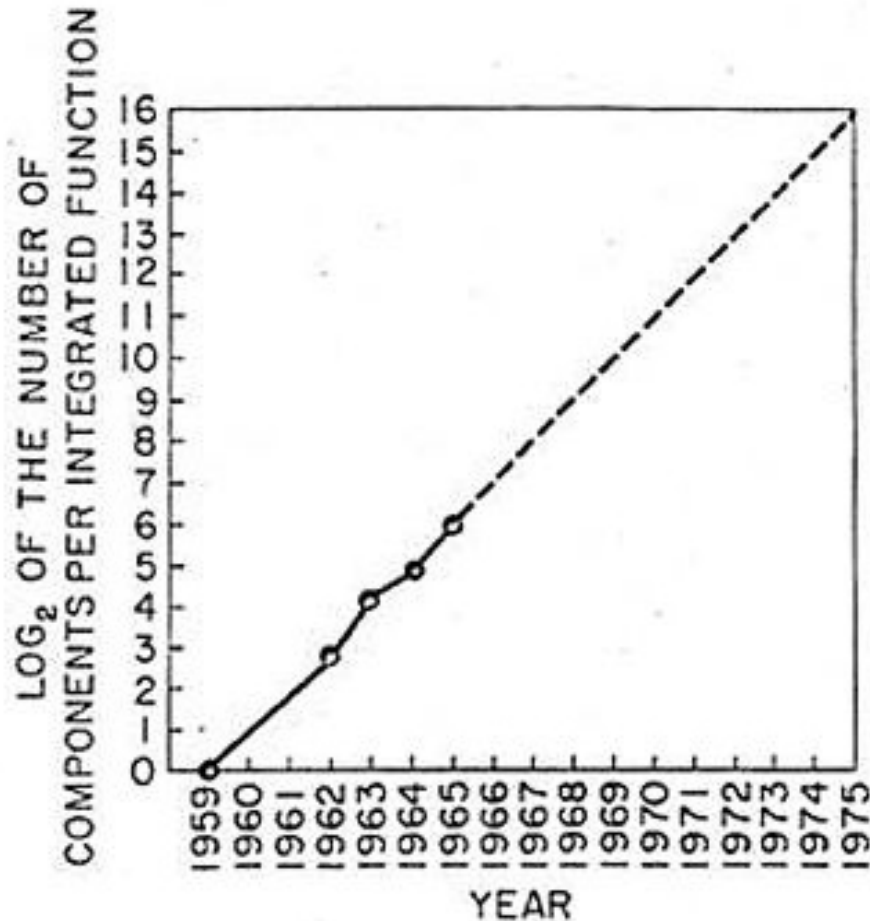
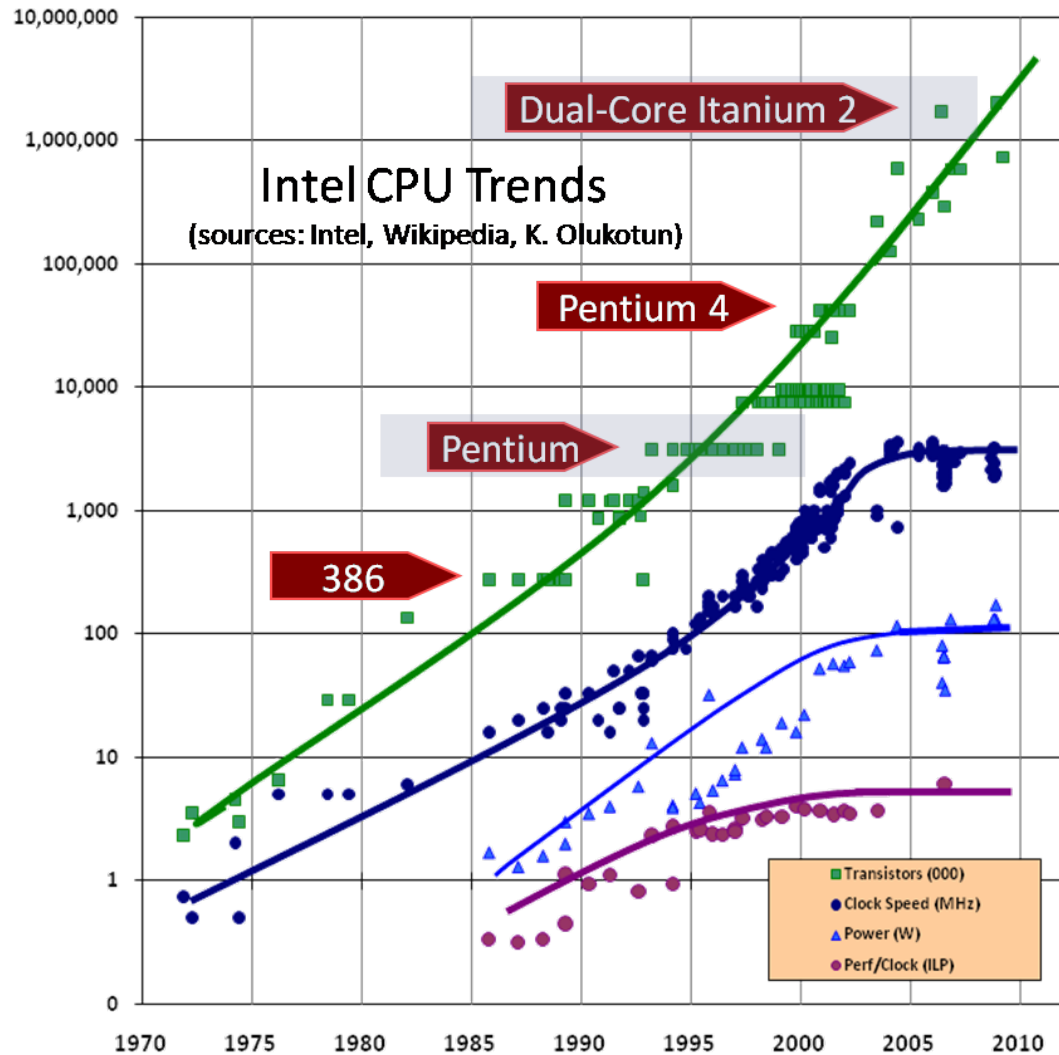


Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

45 years ago, Gordon Moore observed that the number of transistors on a single chip was doubling rapidly

🌟 Important technology trends



The real Moore's Law

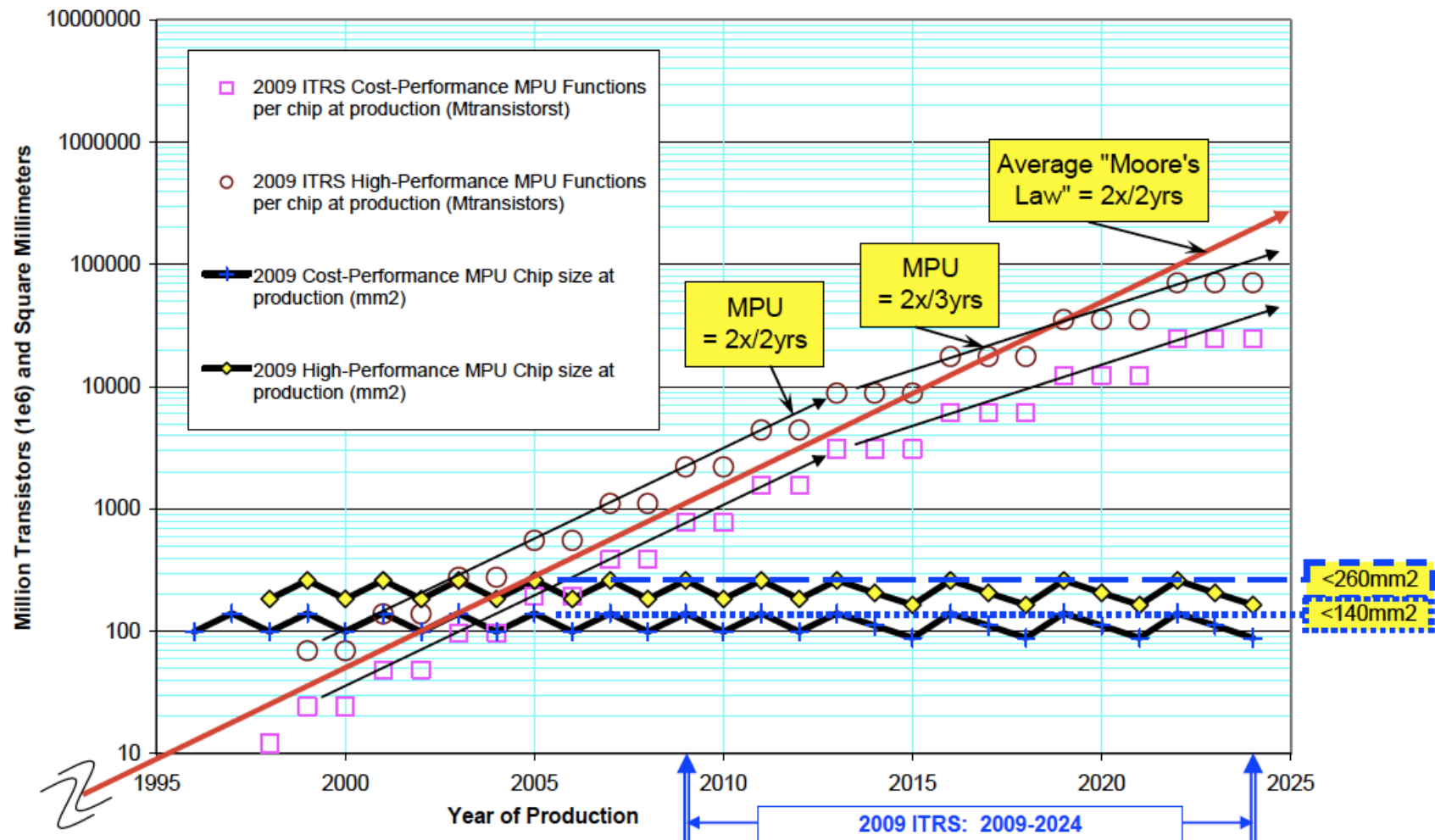
The clock speed plateau

The power ceiling

Instruction level parallelism limit

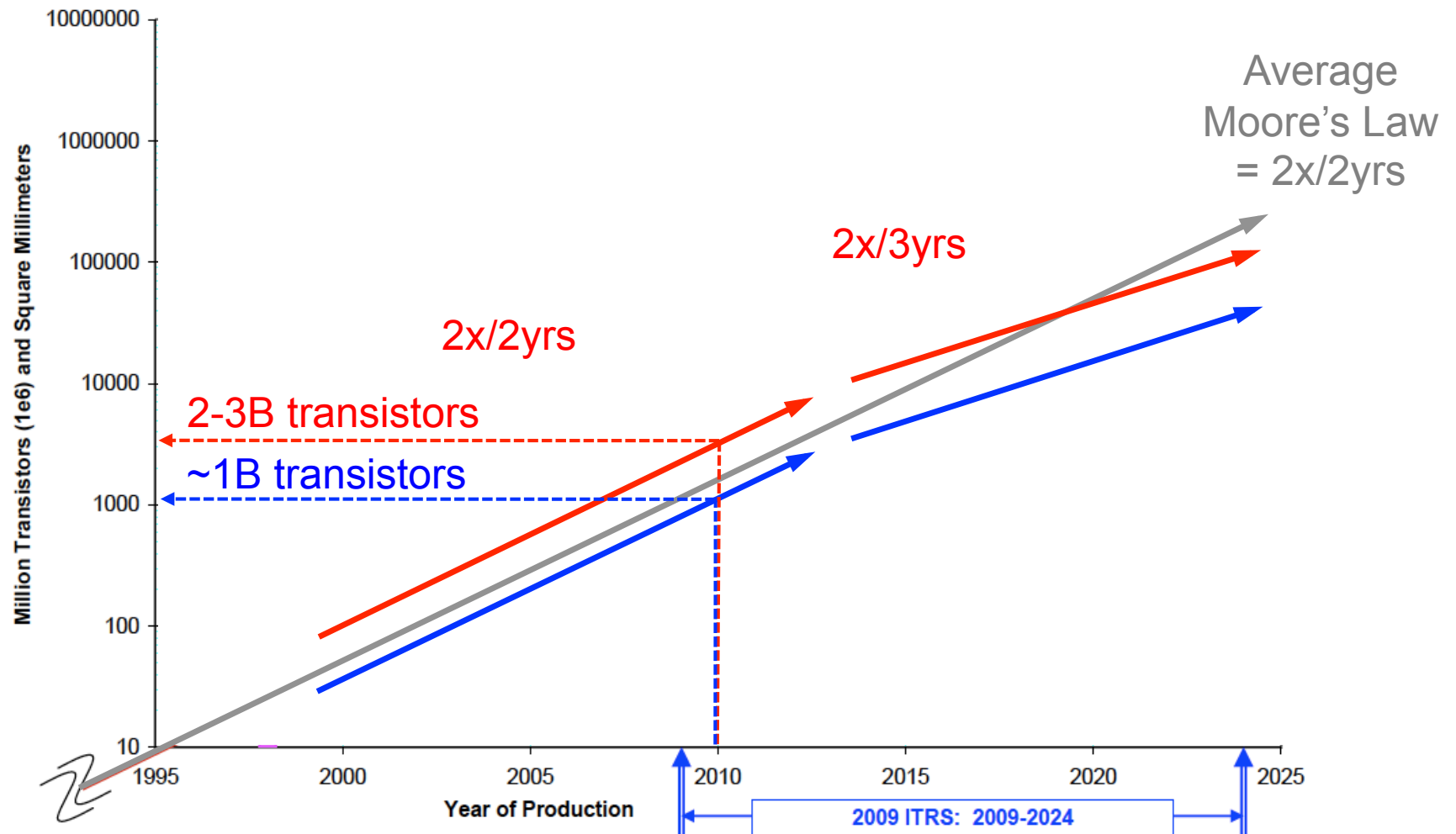
Moore's Law today

2009 ITRS - Functions/chip and Chip Size



Moore's Law today

2009 ITRS - Functions/chip and Chip Size



🔥 What to do with billions of transistors?

- Lots more cores on-chip
 - Core designs will stay roughly the same
- But power consumption must be held in check
 - Chip voltages *can't be dialled down any more!* (0.7V)
 - Clock speeds may *decrease!*
 - Memory bandwidth per core may *decrease!*
 - Memory per core may *decrease!*
- Different types of cores
 - ***Heterogeneous computing!***
 - E.g. a few heavyweight (x86) cores together with many more lightweight (GPU) cores

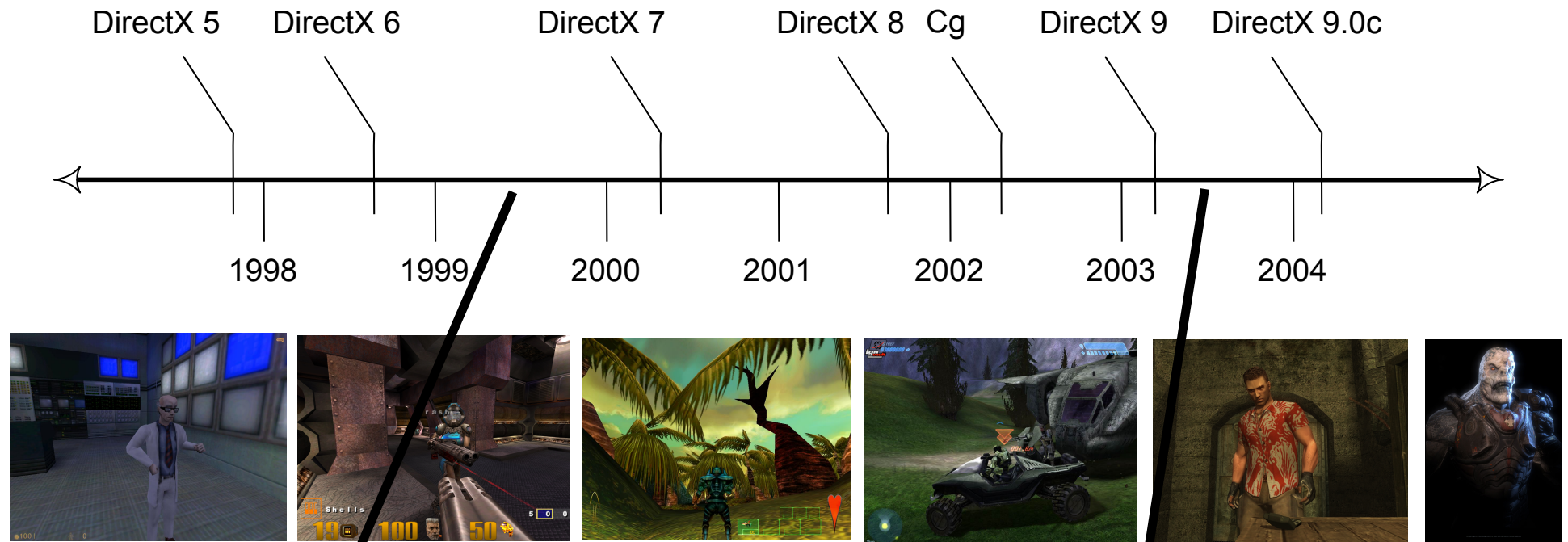
🔥 Heterogeneous computing is not new

- Most systems are *already* heterogeneous
 - PCs have CPU, GPU, network processor, I/O processor, ...
 - Has been a common approach in embedded systems since the early '90s



- But now heterogeneous systems are starting to include several different types of *general-purpose, programmable* processors
 - Users have to programme more than one type of processor to get the most out of a system

🔥 GPUs driven by advances in graphics APIs



Half-Life

Quake 3

Giants

Halo

Far Cry

UE3


Added simple programmable pixel shading

Much more flexible
Could write a general program executed for every pixel

Added ability to **branch** and **execute floating point operations**

GPGPU computing

GPGPU (General-Purpose computation on Graphics Processing Units)

- Term first coined by Mark Harris in 2002
- <http://gpgpu.org/> 
- The first GPGPU applications were still graphics-oriented (ray tracing, video, ...)
- Found early use in Seismic Processing
 - FFT intensive, something GPUs are good at
- Also BLAS, PDEs, RNGs, ...

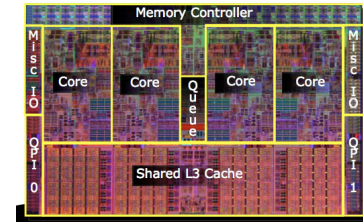
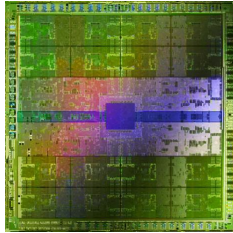
🔥 From GPGPU to ...

Truly **general purpose massively parallel processors**

- Fully-fledged parallel languages such as Nvidia's Cuda started to appear in 2006
- GPUs started to add 64-bit floating point
- Remaining graphics-oriented limitations rapidly disappeared
- True High Performance Computing features now appearing in some GPUs, e.g. Nvidia's *Fermi*



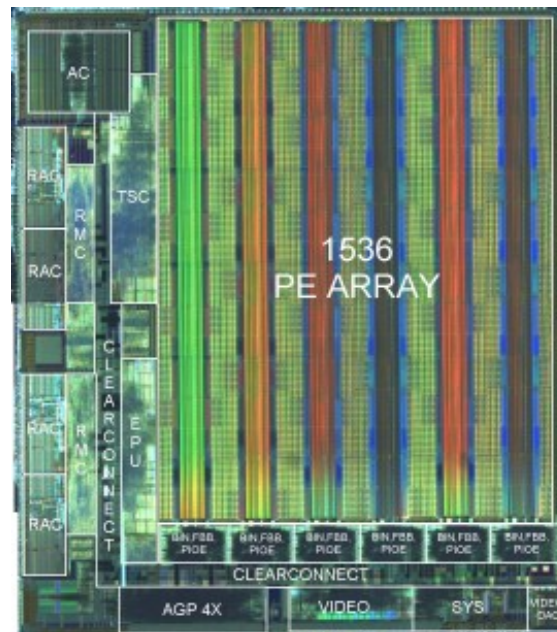
🔥 Comparing Fermi and Nehalem



- 512 simple cores
 - ~3 billion transistors
 - ~1.5GHz
 - ~1,500 GFLOPS S.P.
 - ~750 GFLOPS D.P.
 - ~190 GBytes/s
 - IEEE 754-2008 support
 - ECC on all memories
- 4 complex cores
 - 731 million transistors
 - ~3GHz
 - 96 GFLOPS S.P.
 - 48 GFLOPS D.P.
 - ~30 GBytes/s
 - IEEE 754-1985 support
 - ECC on all memories

🌟 Future GPU architectures

- Tens of thousands of cores per chip
- Highly integrated (mainstream)
- Shared memory models
- Easier to use programming models

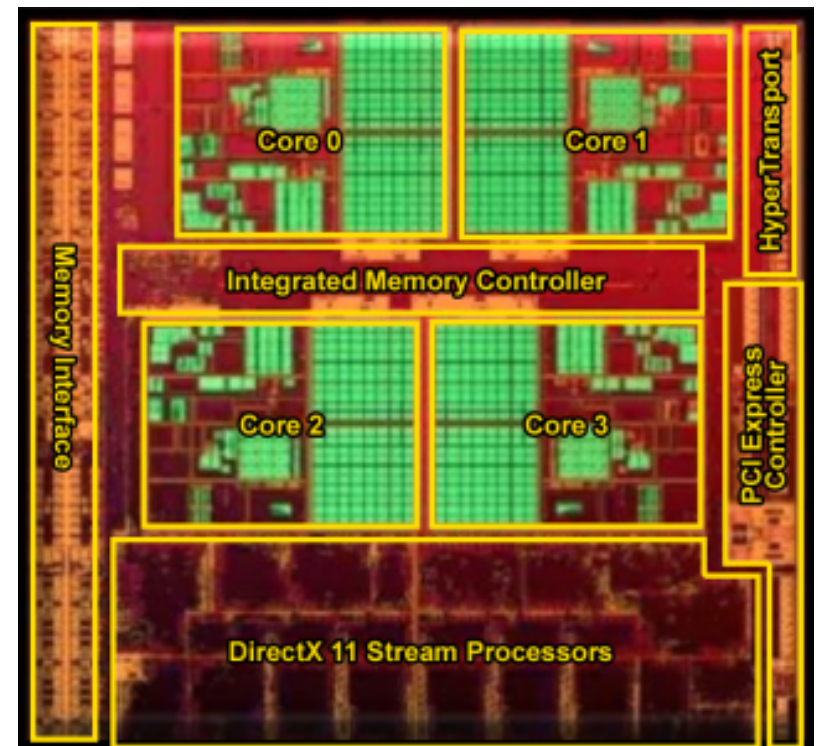


Pixelfusion F150
1,536 simple PEs

🌟 The future is now...

AMD's first "Fusion" chip, disclosed at ISSCC in San Francisco earlier this year

- 'Llano' Accelerated Processing Unit (APU)
- Integrates a quad core x86 CPU with an OpenCL programmable GPU in the same chip
- Intel's doing this too



🔥 Emerging standards

- OpenCL, DirectCompute, ...



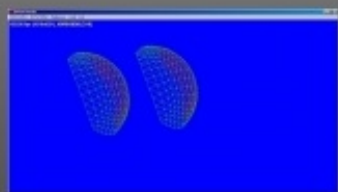
- OpenCL being adopted rapidly in mobile computing (compilers from ARM, Imagination, Zii Labs, ...)
- Just adopted by IBM for the POWER7-based BlueWaters 20 PFLOP UIUC supercomputer

Heterogeneous systems in the **TOP 500**[®] JUNE 2010

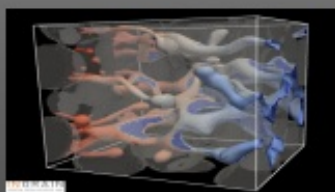
- Tokyo Tech's TSUBAME was first in 2006
 - Started with ClearSpeed, now using GPUs
- Now several systems in existence, more on their way:
 - #2 is Nebulae, 1.3 PFLOPS mostly from 4,640 Nvidia Fermi GPUs
 - #7 is the Tianhe-1 System in China which delivers 563 TFLOPS from Intel x86 + AMD GPUs
 - More coming from Tennessee/Oak Ridge, Tokyo Tech, more Chinese systems, ...



LATEST CUDA NEWS New NVIDIA GPU Technology Conference Content Posted Online



QView



Multiphase flow in porous media

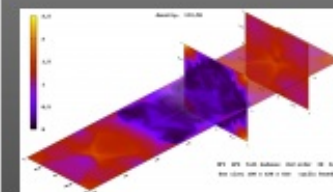
100 x



Graphic processing unit-accelerated mutual information-based 3D image registration

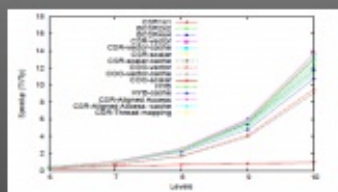


Computing the Longest Common Transposition-Invariant Subsequence with GPU



nHD

173 x

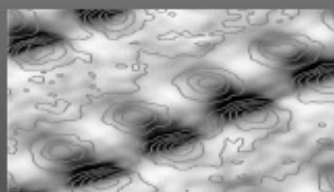


GPU based Sparse Grid Technique for Solving Multidimensional Options Pricing PDE

1000 x



Mersenne Twister for Graphic Processors (MTGP)



Accelerating Geo-Science and Engineering System Simulations on Graphics Hardware

30 x



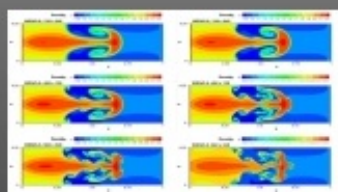
Towards a multi-GPU solver for the three-dimensional two-phase incompressible Navier-Stokes equations

16 x



GPU accelerated analysis of financial markets

80 x



Acceleration of a Finite-Difference WENO Scheme for Large-Scale Simulations on Many-Core Architectures

50 x

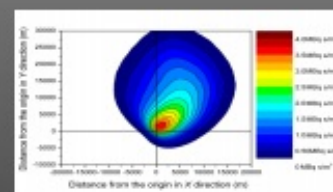


ClusterTech Financial Library in GPU

30 x

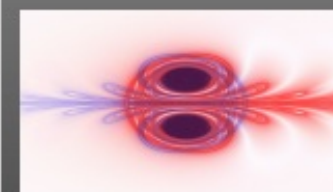


GPU-Assisted Surface Reconstruction on Locally-Uniform Samples



Stochastic Lagrangian Particle Model for Air Pollution

120 x



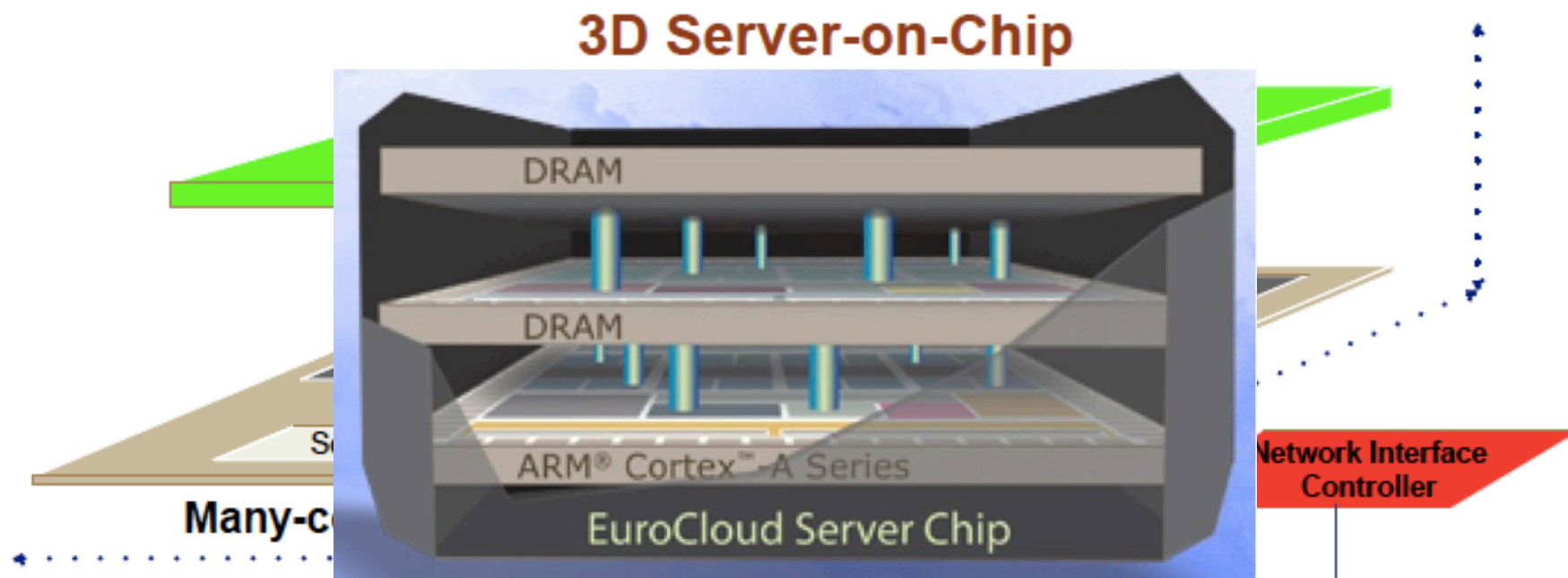
Optimization of FTLE Calculation

1000 x

POTENTIAL DISRUPTIONS TO THE HARDWARE STATUS QUO

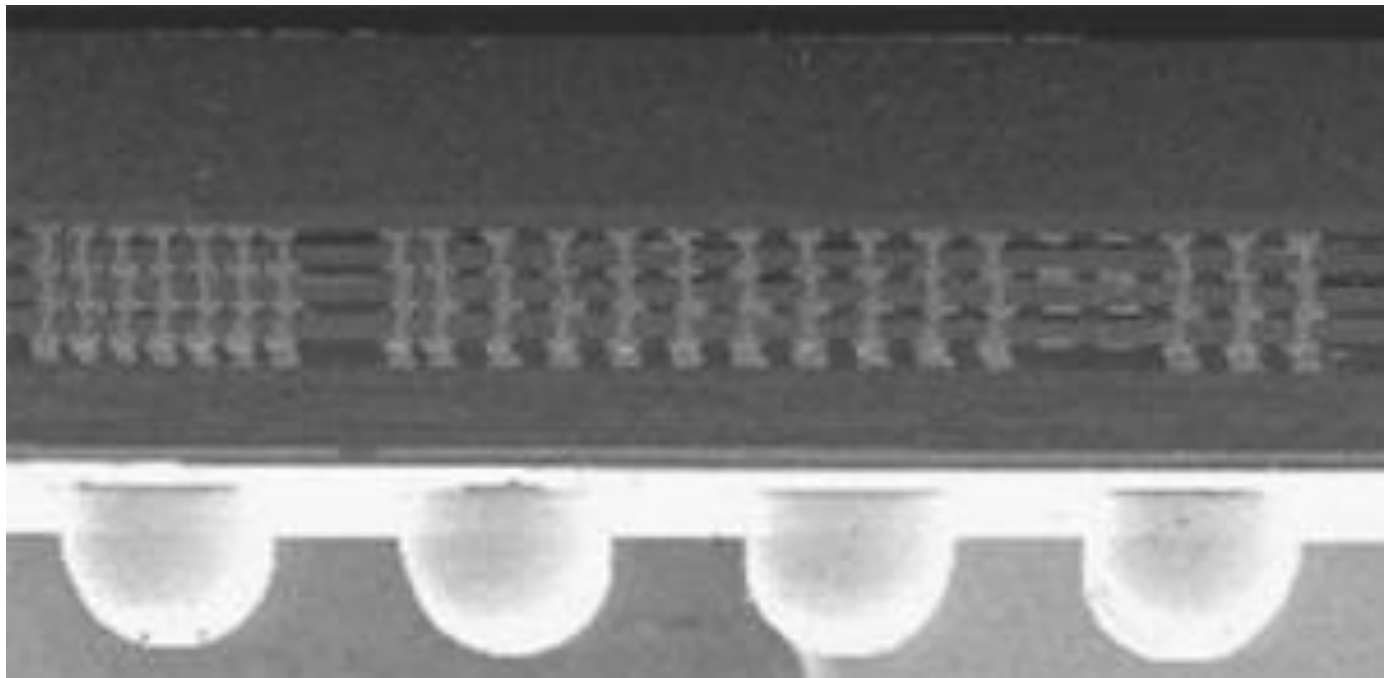
🔥 3D stacked memories

- Vertically stack many-core processors with DRAM → *greater bandwidth* and *greater energy efficiency*



🌟 3D stacked memories

- Vertically stack many-core processors with DRAM → *greater bandwidth* and *greater energy efficiency*



Photonic networks

- Roadmaps to achieve 1 ExaFLOP (1000 PetaFLOPS – 10^{18}) by 2018 are relying on some major hardware breakthroughs to improve energy efficiency
- Prof Keren Bergmen's work at Columbia sponsored by US DoE, Intel, IBM
- Moving data becoming an increasingly dominant fraction of energy dissipation in microelectronics

☀️ Why photonics?

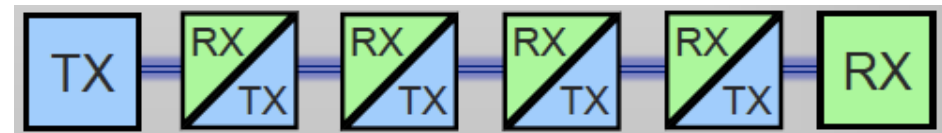
PHOTONICS:

- Modulate/receive ultra-high bandwidth data stream once per communication event
- Broadband switch routes entire multi-wavelength stream
- Off-chip BW = on-chip BW for nearly same power

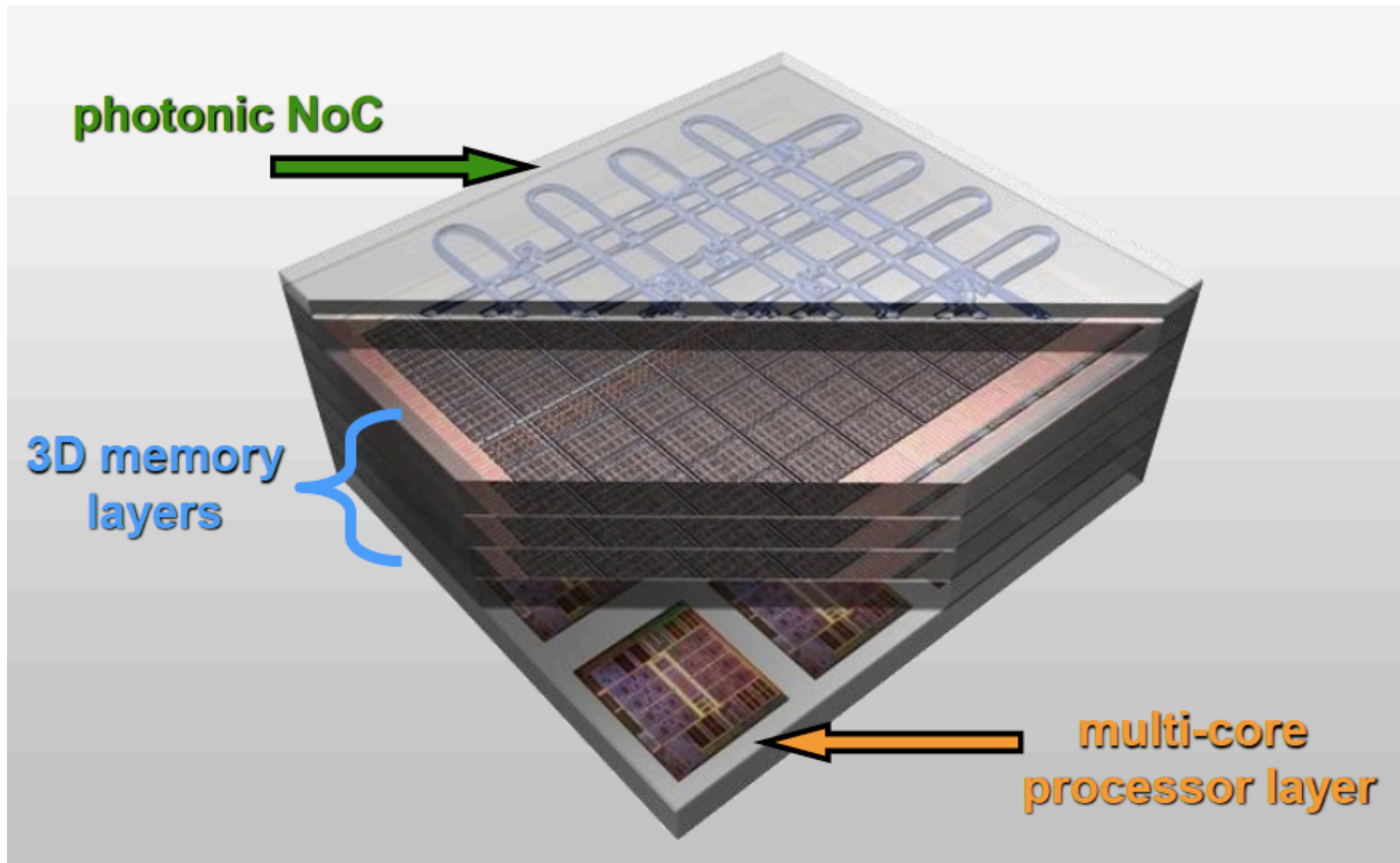


ELECTRONICS:

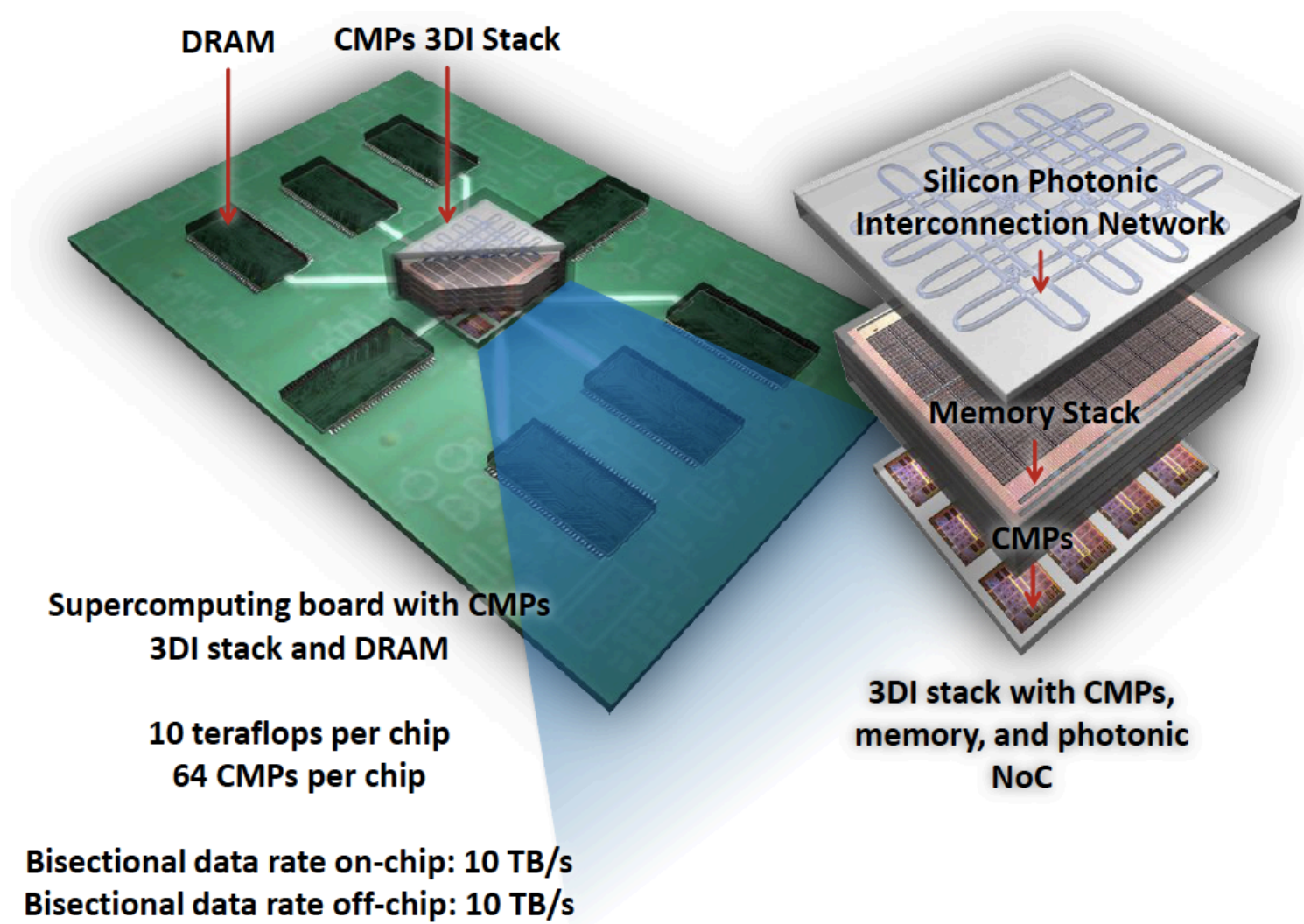
- Buffer, receive and re-transmit at every router
- Each bus lane routed independently
- Off-chip BW requires much more power than on-chip BW



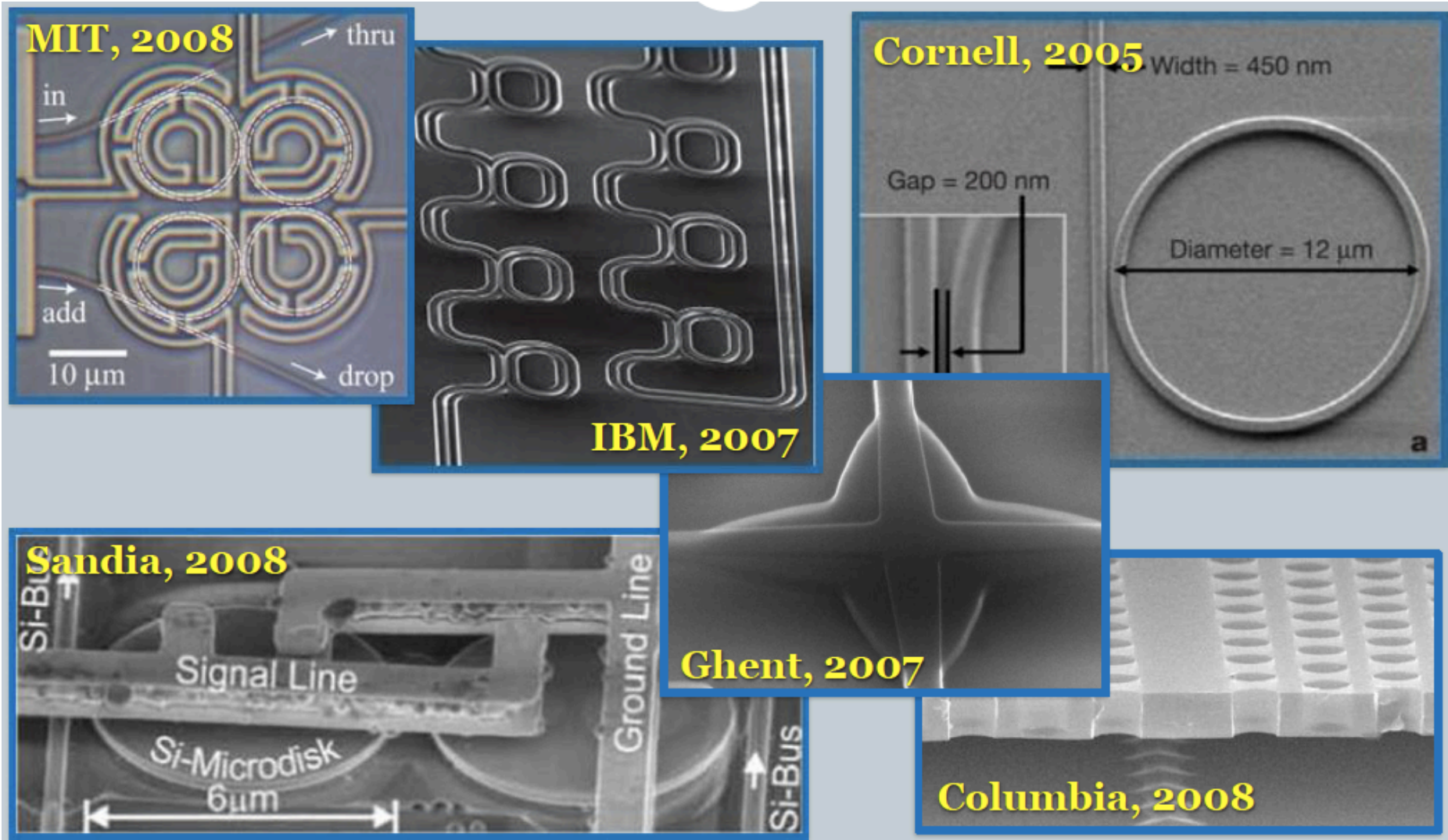
🌟 Photonic NoC integration



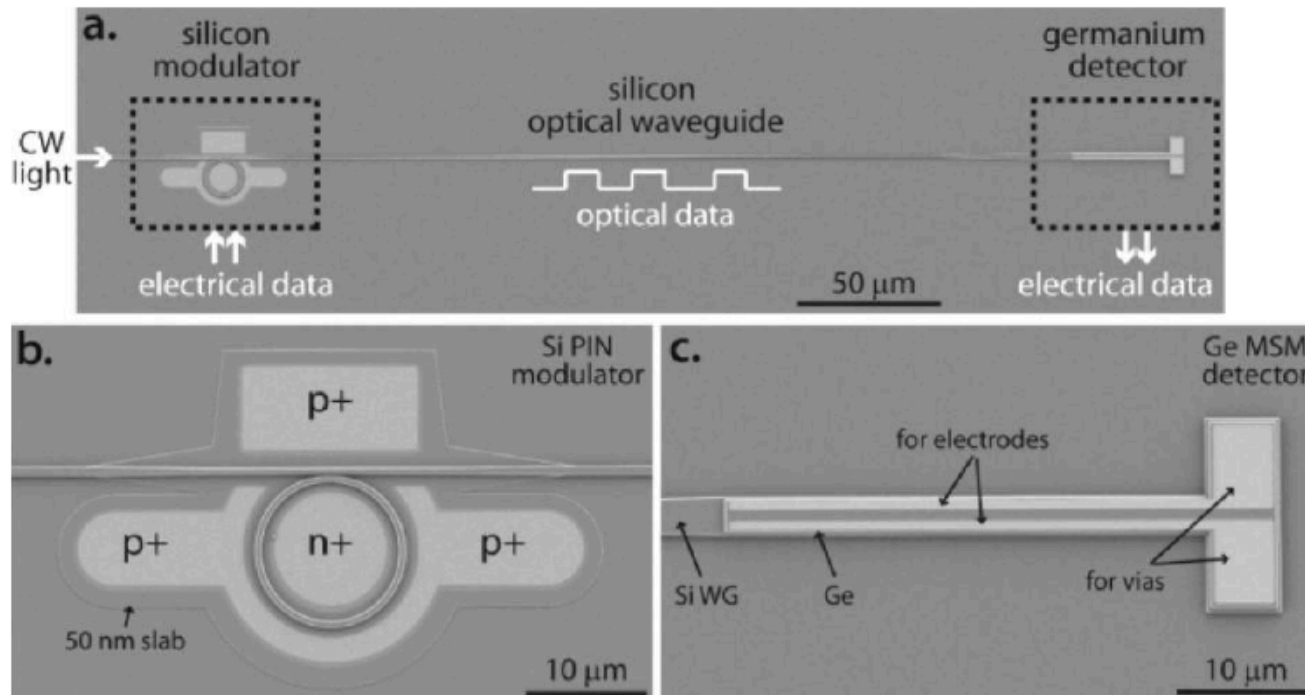
Optically interconnected supercomputing board



🔥 Silicon photonic integration



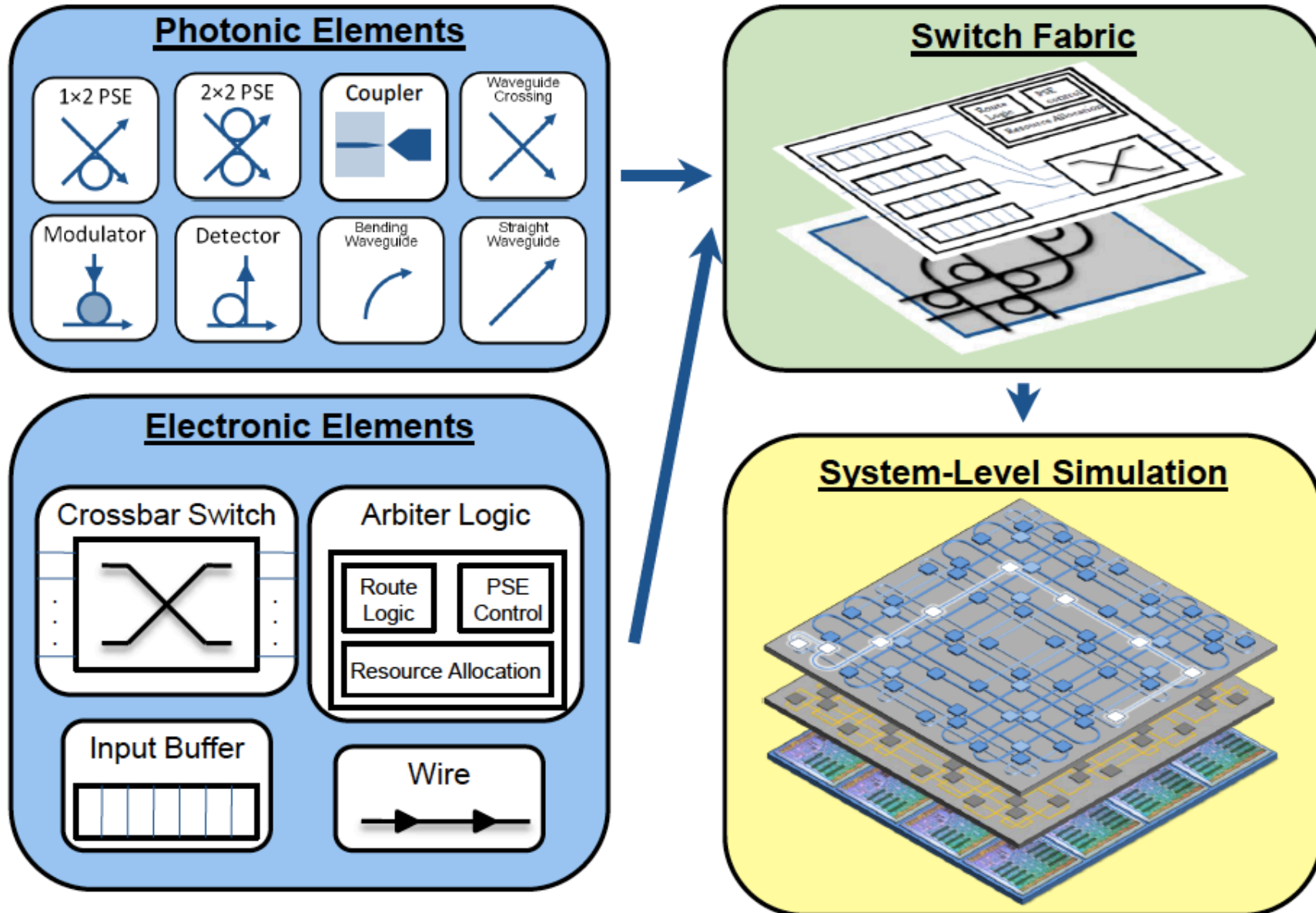
🌟 First complete photonic link



Integrated optical interconnect with silicon electro-optical modulator, silicon waveguide, and germanium-on-silicon photodetector

L. Chen, Optics Express, August 2009

Hybrid photonic/electronic systems will behave very differently



Important implications

Hardware disruptions *may invalidate any/all assumptions* from prior performance modelling

These disruptions are (mostly) *unpredictable!*

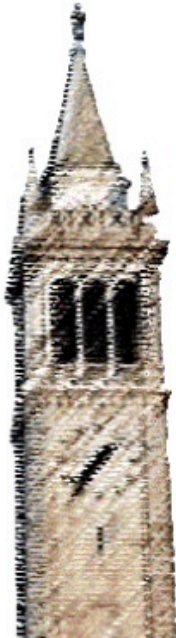
Hierarchies for processing, networks and storage will become *increasingly diverse*

Hardware will become *increasingly unreliable*

MAPPING SOFTWARE TO HETEROGENEOUS ARCHITECTURES

The Seven Dwarfs

The Landscape of Parallel Computing Research: A View from Berkeley



*Krste Asanovic
Ras Bodik
Bryan Christopher Catanzaro
Joseph James Gebis
Parry Husbands
Kurt Keutzer
David A. Patterson
William Lester Plishker
John Shalf
Samuel Webb Williams
Katherine A. Yelick*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-183
<http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html>

December 18, 2006

- First described by Phil Colella (LBNL 2004)
- Built on earlier work by Per Brinch Hansen
- Expanded to 13 dwarfs by a group of researchers at Berkeley in 2006

🔥 What are the Seven Dwarfs?

Describe key algorithmic kernels in many scientific applications

1. Dense linear algebra – *BLAS, ScaLAPACK*
2. Sparse linear algebra – *SpMV, SuperLU*
3. Spectral methods – *FFT*
4. N-body methods – *Fast Multipole*
5. Structured grids – *Lattice Boltzmann*
6. Unstructured grids – *ABAQUS, Fluent*
7. Monte Carlo

Seven Heterogeneous Dwarfs

1. Dense linear algebra – *excellent progress*

- PLASMA/MAGMA – Dongarra et al
- FLAME – Robert van de Geijn et al
- Vendor libraries – CUBLAS, ACML, NAG, ...

2. Sparse linear algebra

- Iterative solvers – *good progress*
 - Nathan Bell and Michael Garland (NVIDIA Research) have general-purpose iterative solvers using efficient sparse matrix-vector multiplication
 - Andreas Klöckner (Brown University) has “Iterative CUDA” package based on same SpMV products
 - Manfred Liebmann & colleagues (University of Graz) have implemented algebraic multigrid

Seven Heterogeneous Dwarfs

3. Spectral methods – *good progress*

- FFT libraries from vendors
- “Auto-Tuning 3-D FFT Library for CUDA GPUs”
Akira Nukada, Satoshi Matsuoka, Tokyo Institute of Technology, SC09
 - Very fast, 160 GFLOPS for 256^3 32-bit 3D FFT

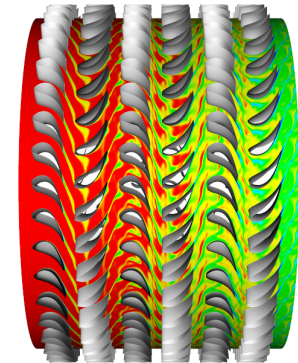
4. N-body methods – *excellent progress*

- NAMD/VMD – Phillips, Stone, UIUC
- OpenMM, Folding@Home – Pande, Stanford
- Fast multipole methods - “42 TFlops Hierarchical N-body Simulations on GPUs with Applications in both Astrophysics and Turbulence”, Hamada et al, SC09

🔥 Seven Heterogeneous Dwarfs

5. Structured grids – *excellent progress*

- “Turbostream” turbulent fluid flow application framework, Pullan and Brandvik, Cambridge
 - 20X speedup
- Datta et al SC08
- Jonathan Cohen at NVIDIA Research developing a library called OpenCurrent



Seven Heterogeneous Dwarfs

6. Unstructured grids – *good progress*

- Several projects underway in the CFD community
- Rainald Löhner (GMU – Washington DC)
- Jamil Appa (BAE Systems)
- Graham Markell / Paul Kelly (Imperial)
- Mike Giles (Oxford) working with Markell, Kelly and others on a general-purpose, open-source framework called OP2
- Others underway

🔥 Seven Heterogeneous Dwarfs

7. Monte Carlo – *excellent progress*

- Massively parallel, an excellent fit
- Vendors providing examples
- Mike Giles (Oxford) working with NAG to develop a GPU library of RNG routines
 - E.g. mrg32k3a and Sobol generators
 - <http://www.nag.co.uk/numeric/GPUs/>
- Lots of work in this space

🌟 Important takeaways

- Heterogeneous computing is here to stay
- Even single chips will contain **tens of thousands of cores**
- Hierarchies will become **deeper**
- Hardware will become **increasingly unreliable**
- Higher level application templates, libraries and auto-tuners will be essential
- It is ***crucial*** that anyone modelling future systems or developing software is aware of these implications!