Application Challenges for Exascale

Dr. Marie-Christine Sawley
Intel Exascale Lab Director, Paris
IPAG-EU

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Optimization Notice

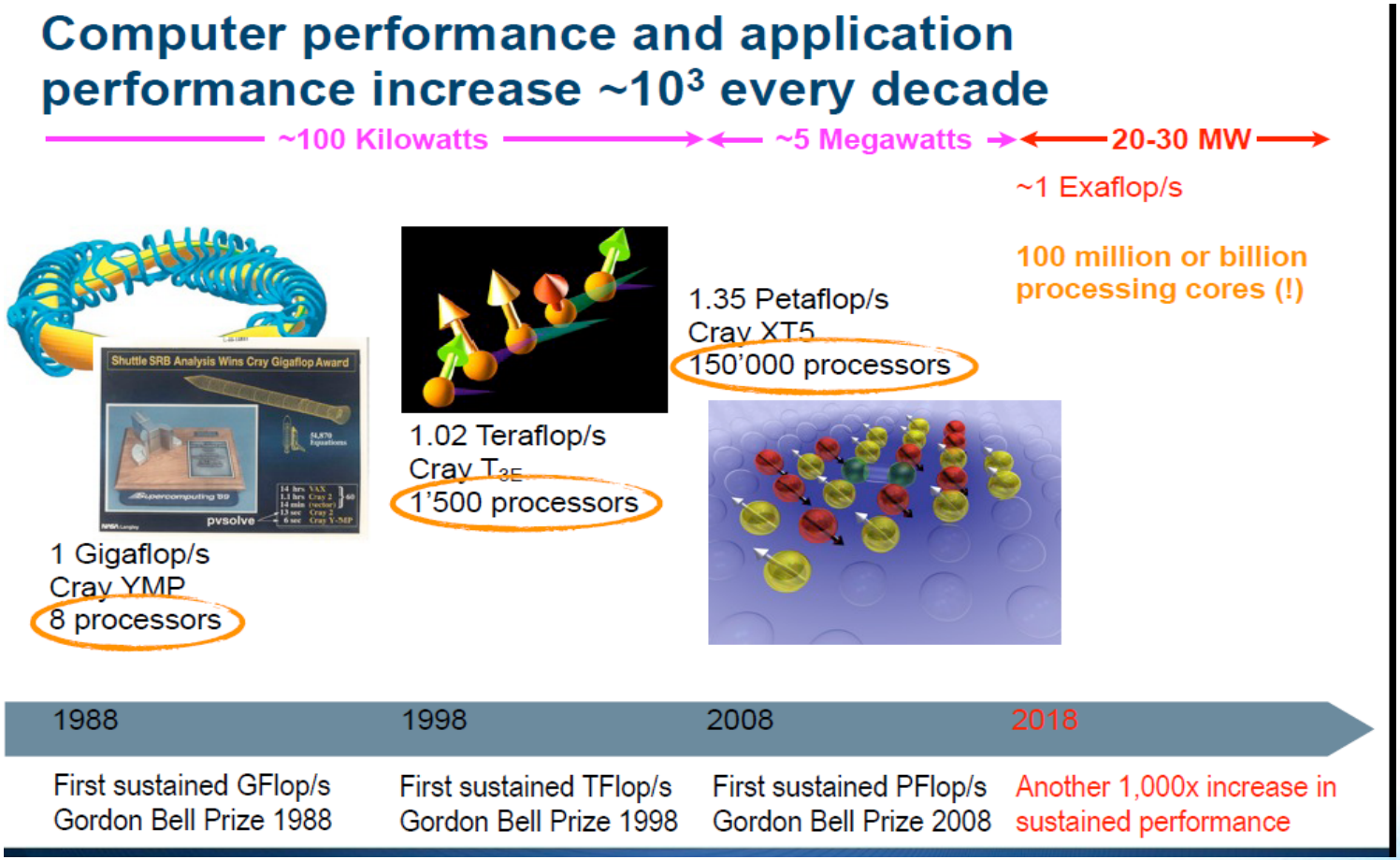
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Notice revision #20110804

Agenda

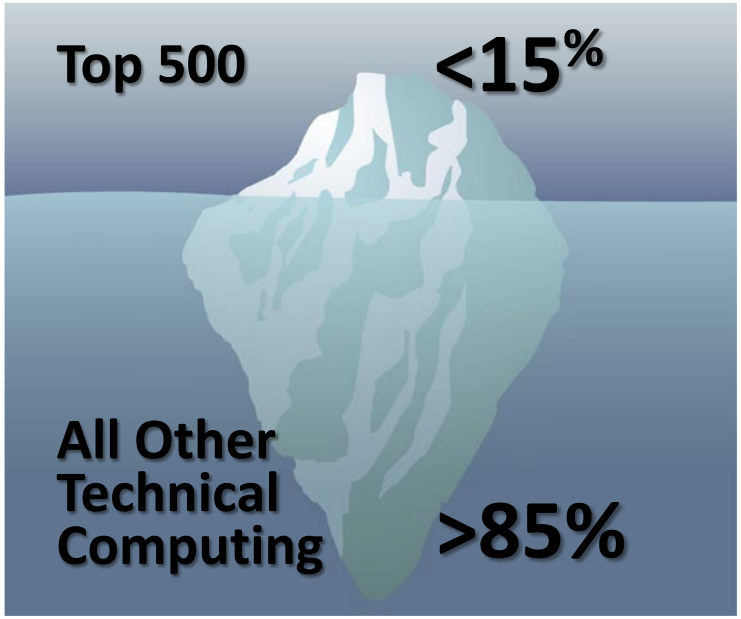
- General trends in HPC hw and sw architecture
- Impact on application and scientific workloads development
- Several approaches to co design on HPC applications

Learning from the last 20 years



Courtesy: Prof. T. Schulthess, ETH Zurich

Technology Waterfalls from the Top



% of sockets sold

Performance Waterfall*
#1 Top500 System to Single Socket

6-8 years
#1 to #500

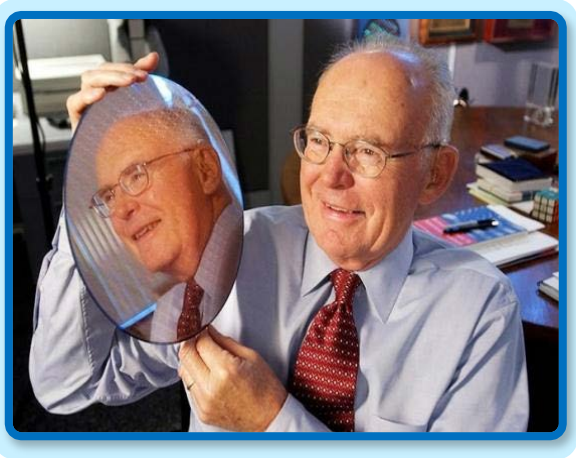
~9 years
#500 to Single Socket

**plus…..similar waterfalls for other capabilities in areas like fabrics, storage, software, …*

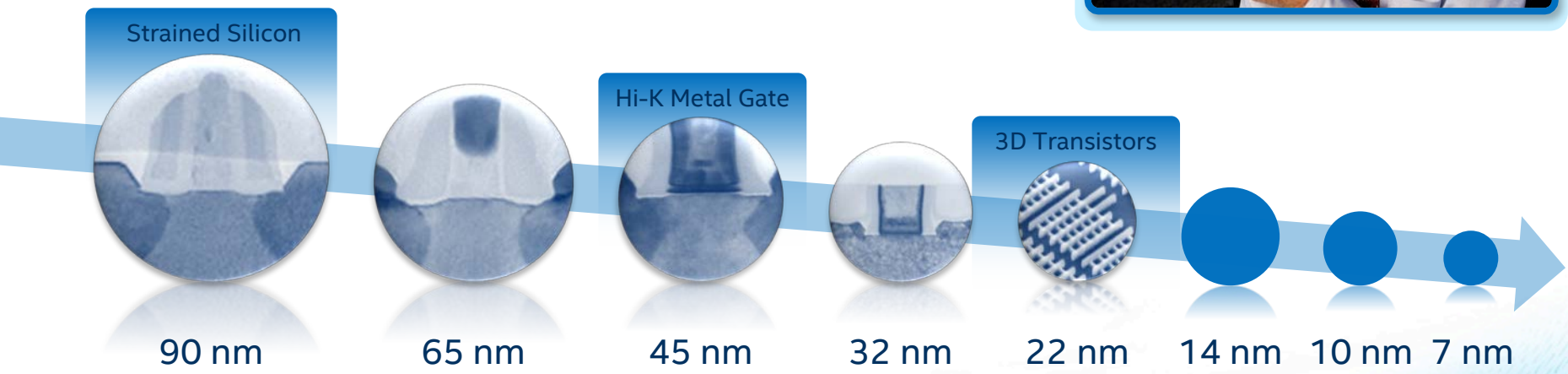
Source: Top500.org and Intel Estimate of Top500 sockets as % of sum of analysts reports of HPC and branded Workstations sockets. Performance waterfall timelines based on TOP500.org statistics (#1-#500) and Intel estimate (#500 to projected Intel Knights Landing)

Predictable Silicon Track Record

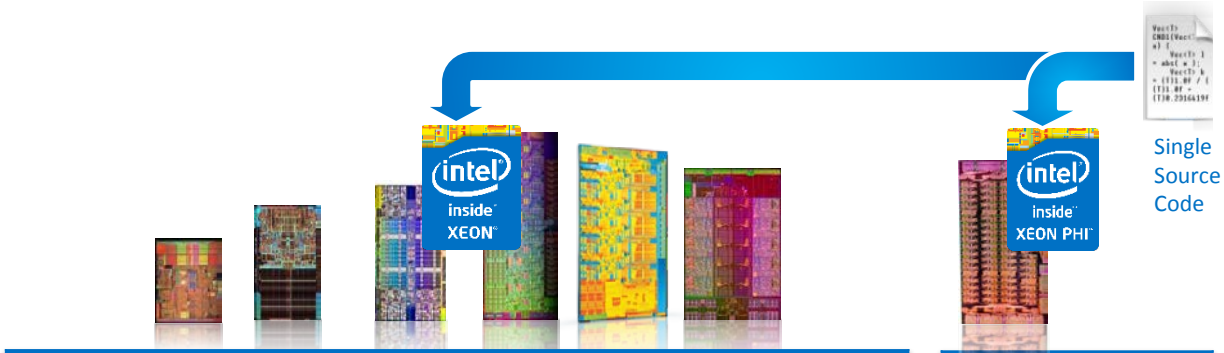
Executing to Moore's Law



Enabling new devices with higher functionality and complexity while controlling power, cost, and size



Parallelism is the Path Forward



	Intel® Xeon® processor 64-bit	Intel® Xeon® processor 5100 series	Intel® Xeon® processor 5500 series	Intel® Xeon® processor 5600 series	Intel® Xeon® processor E7 series	Intel® Xeon® processor E7 v2 series	Intel® Xeon® processor Future	Intel® Xeon Phi™ coprocessor Knights Corner	Intel® Xeon Phi™ processor & coprocessor Knights Landing
Core(s)	1	2	4	6	10	15	?	61	?
Threads/ Socket	2	2	8	12	20	30	?	244	?
Vector Width	128	128	128	128	256	256	512	512	512

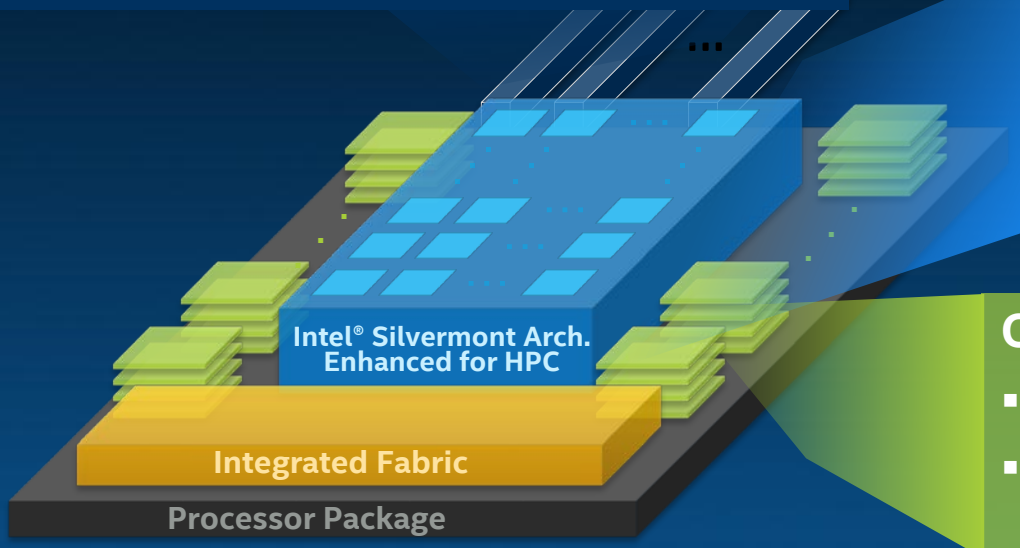
Unveiling Details of Knights Landing

(Next Generation Intel® Xeon Phi™ Products)

2nd half '15
1st commercial systems

3+ TFLOPS[1]
In One Package
Parallel Performance & Density

Platform Memory: DDR4 Bandwidth and Capacity Comparable to Intel® Xeon® Processors



Compute: Energy-efficient IA cores[2]

- *Microarchitecture enhanced for HPC[3]*
- ***3X** Single Thread Performance vs Knights Corner[4]*
- *Intel Xeon Processor Binary Compatible[5]*

On-Package Memory:

- *up to **16GB** at launch*
- ***1/3X** the Space[6]*
- ***5X** Bandwidth vs DDR4[7]*
- ***5X** Power Efficiency[6]*

Jointly Developed with Micron Technology

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. [1]Over 3 Teraflops of peak theoretical double-precision performance is preliminary and based on current expectations of cores, clock frequency and floating point operations per cycle. FLOPS = cores x clock frequency x floating-point operations per second per cycle. . [2]Modified version of Intel® Silvermont microarchitecture currently found in Intel® Atom™ processors. [3]Modifications include AVX512 and 4 threads/core support. [4]Projected peak theoretical single-thread performance relative to 1st Generation Intel® Xeon Phi™ Coprocessor 7120P (formerly codenamed Knights Corner). [5] Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX) . [6]Projected results based on internal Intel analysis of Knights Landing memory vs Knights Corner (GDDR5). [7]Projected result based on internal Intel analysis of STREAM benchmark using a Knights Landing processor with 16GB of ultra high-bandwidth versus DDR4 memory only with all channels populated.

Conceptual—Not Actual Package Layout

intel

Efficient Performance Server Platforms Roadmap

4S Efficient Performance

Romley-EP 4S Platform

Intel® Xeon® processor E5-4600 v2 product family

Intel® C600 series chipset

2S Efficient Performance

Romley-EP Platform

Intel® Xeon® processor E5-2600 v2 product family

Intel® C600 series chipset

Technologies:

- Up to 12 cores/24 threads
- Intel® Integrated I/O
- Integrated 3Gb/s SAS
- Intel® Advanced Vector Extensions (Intel® AVX)
- Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI)
- Intel® Trusted Execution Technology (Intel® TXT)
- Intel® Data Direct I/O (Intel® DDIO)
- LR-DIMM (up to 1536GB memory support with octal rank)
- Intel® Data Protection Technology with Secure Key
- Intel® Platform Protection Technology with OS Guard
- Advanced Programmable Interrupt Controller virtualization
- PCIe* support for Atomics Operations, x16 Non-transparent bridge
- Technologies also apply to Intel® Xeon® processor
- E5-4600 v2 product family

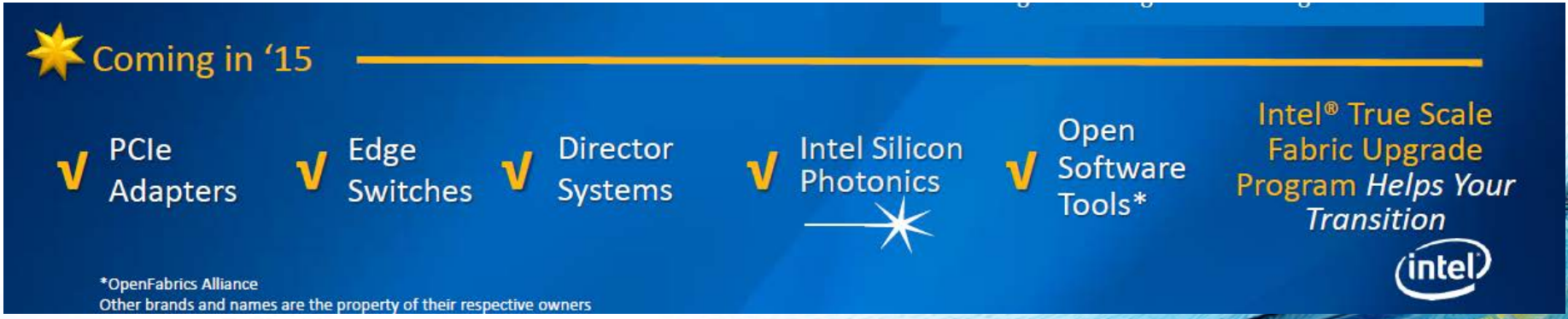
Grantley-EP Platform

Intel® Xeon® processor E5-2600 v3 product family

Intel® C610 series chipset

More to come

- Intel has announced plans for the first Xeon with coherent FPGA providing new capabilities to selected data center workloads
- Storm Lake, the next generation fabric, more integration

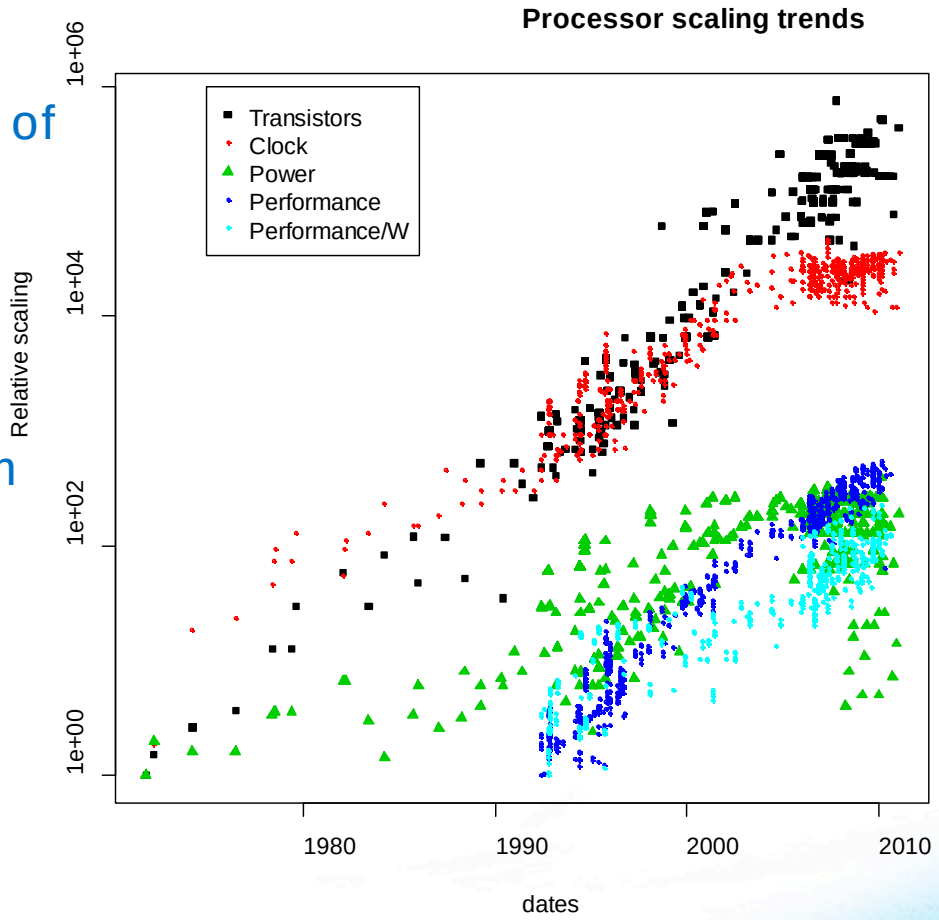


Fundamentals: Performance Trends

After ~2004 only the number of transistors continues to increase exponentially

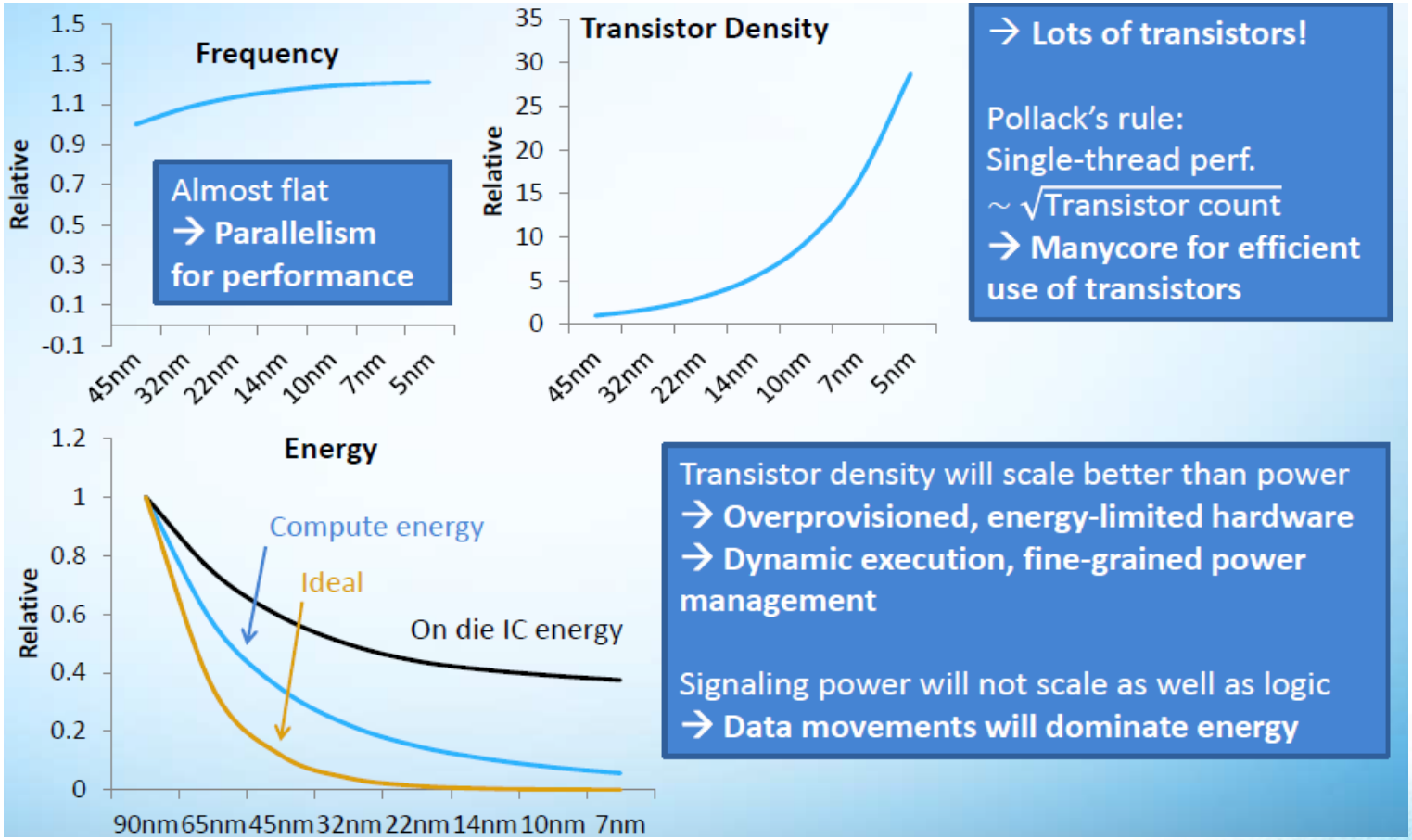
We have hit limits in

- Power
- Instruction level parallelism
- Clock speed

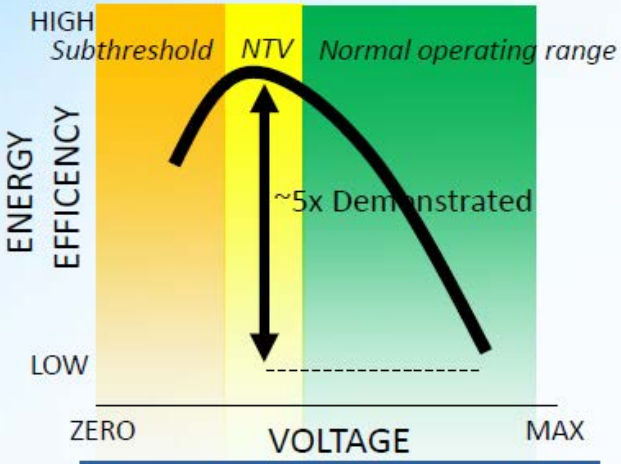


Single core scalar performance is now only growing slowly

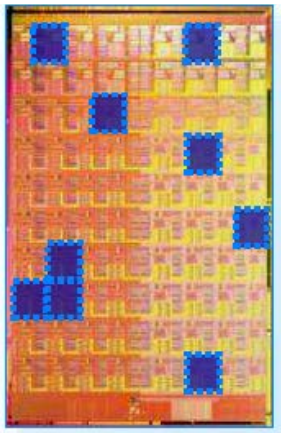
Technology scaling outlook



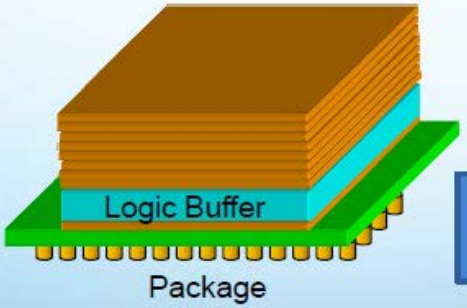
Some promising technology



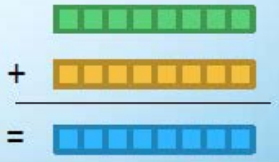
Near Threshold Voltage (NTV)



Fine-grain power management

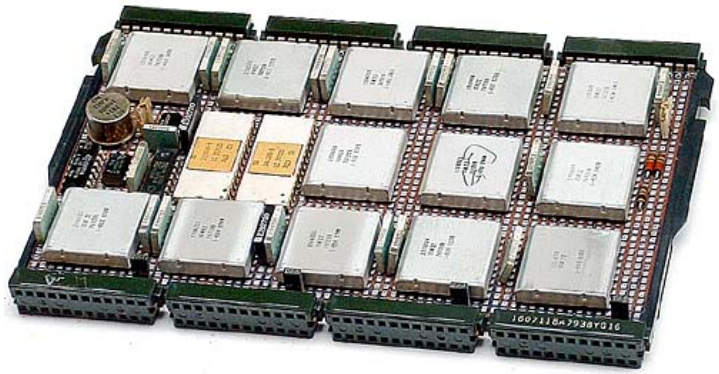


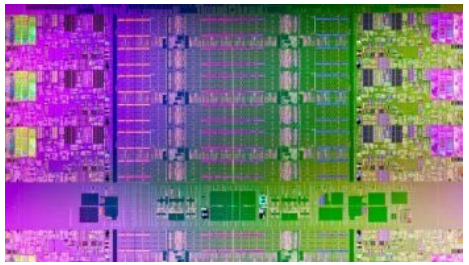
Stacked memory



Specialized circuits (SIMD, encryption, ...)

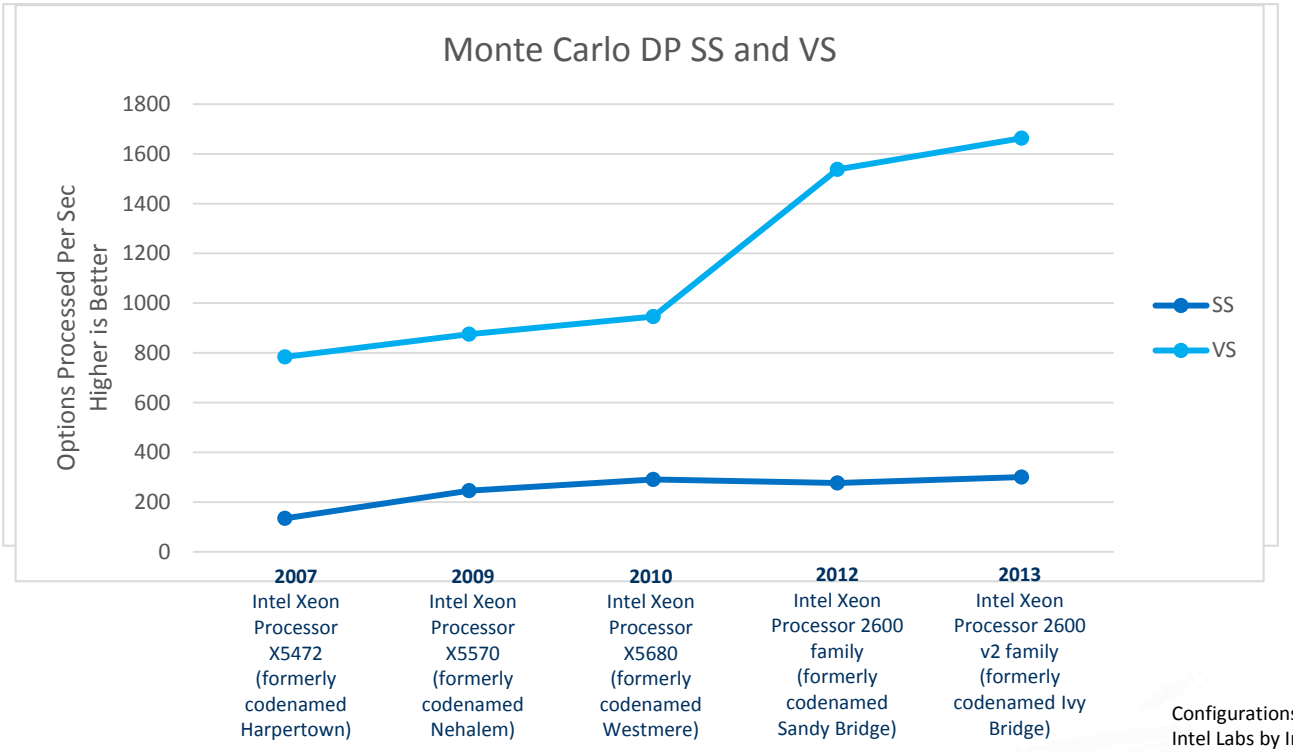
What are the chances?





A Code (probably in FORTRAN) written for this CPU will perform well on these?

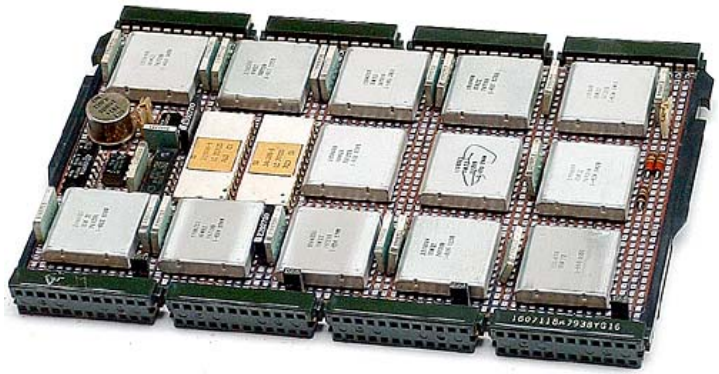
Not Good, It Turns Out

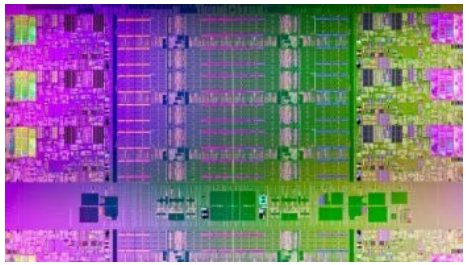


SS: Single threaded and Scalar
VS: Vectorized and Single threaded

Configurations: as listed on Slide 78. Perfromance measured in Intel Labs by Intel employees. **For more information go to http://www.intel.com/performance**

What are the changes?





A Code (probably in FORTRAN) written for this CPU will perform well on these?

Parallel computing is hard

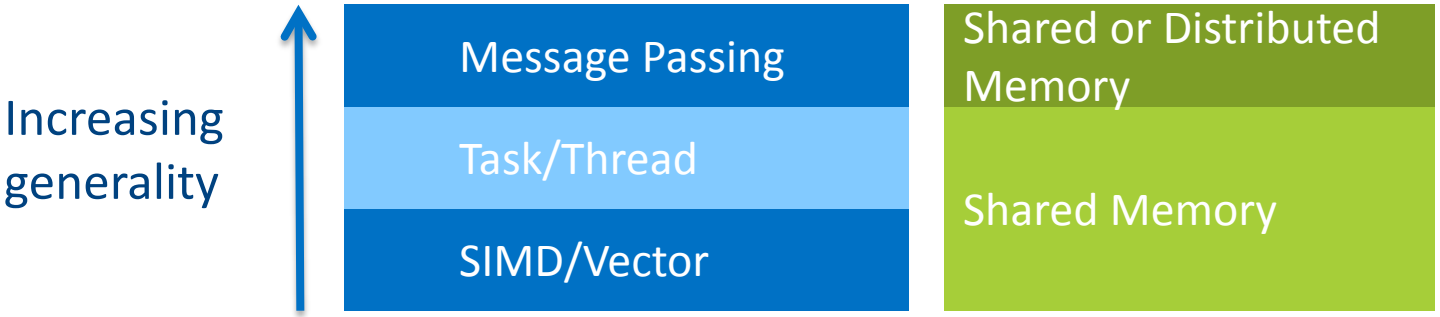
- We tend to think sequentially
- Very few of us were properly taught how to efficiently write parallel programs
- Using efficiently one core is hard
  - Yearly improvement
    - computing core: 50%
    - Memory BW: 20%
    - Memory latency 5%
  - Using efficiently many cores in parallel
    - Risk of BW contention and limitations

Application and software stack challenges

- 1000x more concurrency is a a game changer
- Manycore requires exploiting shared memory between threads
- Manage data locality
  - Today, explicit across nodes
  - Highly dynamic execution call for management strategies
- Opportunity to benefit from specialized circuits

Vectorization, shared memory parallelism and coding for efficient data locality

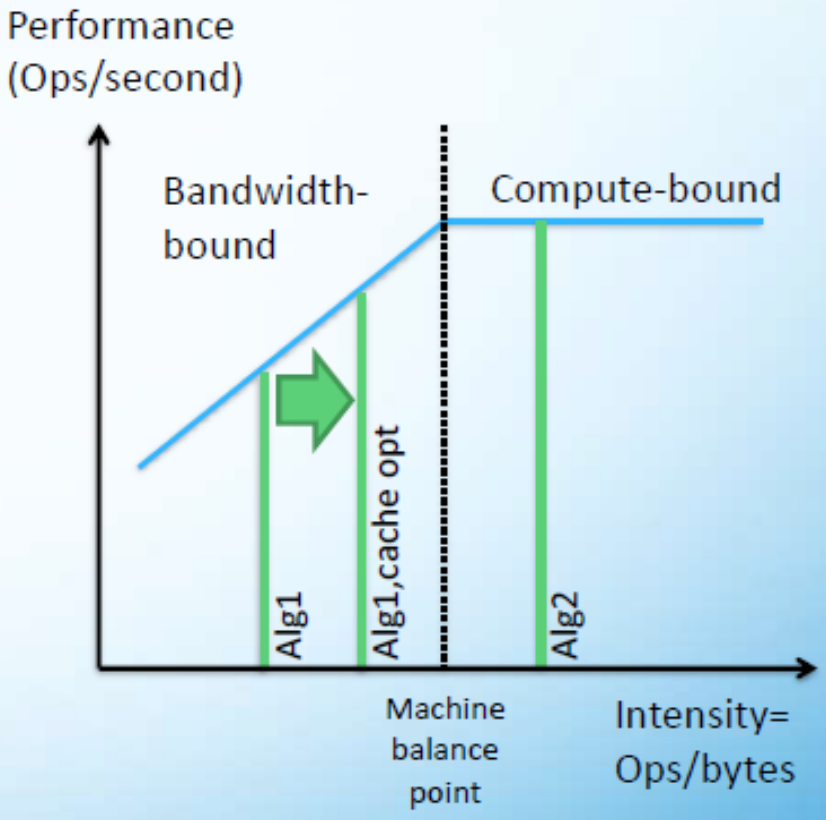
Arch Robison, Ralph Johnson's "Three Layer Cake"



- Driven by hardware architecture and expressivity
  - Higher layers can emulate lower ones, but not the reverse
- In HPC we have
  - Message Passing : MPI, (PGAS?)
  - Task/Thread : OpenMP*, pthreads, (Intel® Threading Building Blocks?)
  - SIMD/Vector: Fortran, Intel® Cilk™ Plus, OpenMP* 4.0 SIMD directives, OpenCL

Data movement and applications

- Data transfers will dominate energy budget
- What are "local" algorithms?
  - **High compute intensity**: Need little data to perform ops
  - **Reuse data** from the caches as much as possible
- Traditional techniques
  - Avoid "streaming" data, avoid multi-pass processing
  - Stencils: cache blocking
  - Trees: locality-based sorting (Z-curve, Hilbert curve)
  - Divide-and-conquer algorithms (recursive problem decomposition)



Roofline model, [*Williams et al. 2009*]

AMBER WRF VISIT VASP UTBENCH SU2 SG++ SeisSol, GADGET, SG++ ROTOR SIM R Quantum Espresso Optimized integral OPENMP/ MPI Openflow NWChem

AVBP (Large Eddy)

Modernizing Community Codes…*Together*

NEMO5

MPAS

Blast

Mardyn

BUDE

MACPO

CAM-5

Ls1

CASTEP

Harmonie

Castep

GTC

CESM

GS2

CFSv2

Gromacs

CIRCAC

GPAW

Intel® Parallel Computing Centers

CINECA
THE UNIVERSITY OF EDINBURGH
CERFACS
ICHEC Irish Centre for High-End Computing
University of BRISTOL
TUM lrz
ETH Zürich
CISS
INSTITUTE OF INDUSTRIAL SCIENCE THE UNIVERSITY OF TOKYO
epcc
COPPE UFRJ
ZIB
सी डैक CDAC
CSC
The UNIVERSITY of OKLAHOMA
PURDUE UNIVERSITY
THE UNIVERSITY of WISCONSIN MADISON
TACC
Stanford University
BERKELEY LAB
PennState
Georgia Tech
University of Colorado Boulder
NCAR
東京大学 THE UNIVERSITY OF TOKYO
LSU Center for Computation & Technology Interdisciplinary | Innovative | Inventive

Plus… **User Groups Forming**

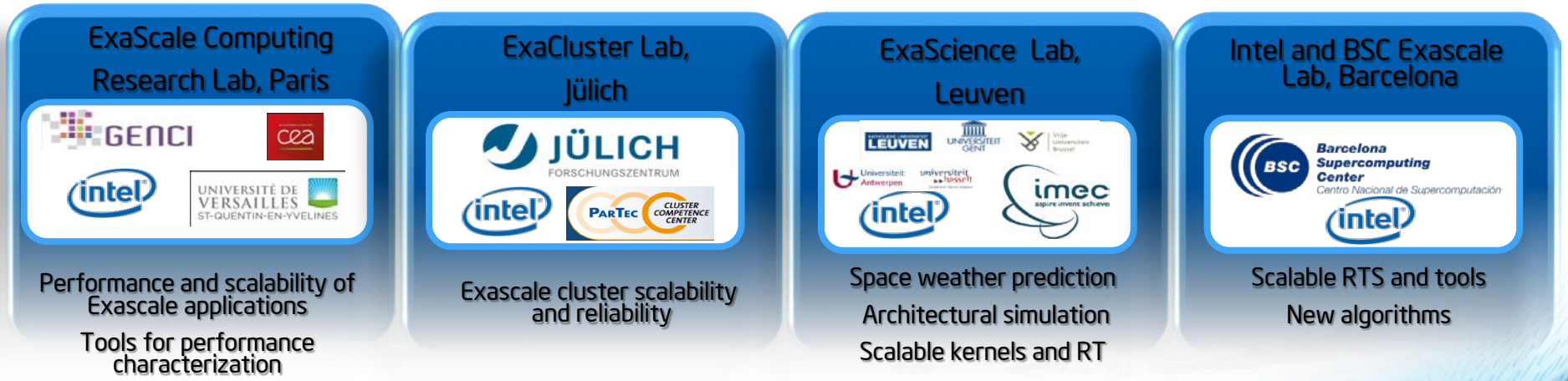
CliPhi (COSMOS) COSA Cosmos codes DL-MESO DL-Poly ECHAM6 Elmer FrontFlow/Blue Code GADGET GAMESS-US

Intel Exascale Labs — Europe

Strong Commitment To Advance Computing Leading Edge:

Intel collaborating with HPC community & European researchers

4 labs in Europe - Exascale computing is the central topic



Tool development at ECR

Introduction: Performance evaluation

- Characterize the performance of an application
  - Complex multicore CPUs and memory systems
  - How well does it behaves on a given machine
- Generally a multifaceted problem
  - What are the issues (numerous but finite) ?
  - Which one(s) dominates ?
  - Maximizing the number of views
  - **=> Need for specialized tools**
- Several tools available
  - Which one to use ?
  - **=> Need for a methodology**

MAQAO: Introduction

MAQAO

www.maqao.org

- Open source (LGPL 3.0)
  - Currently binary release
  - Source release soon
- Available for:
  - x86-64
  - Xeon Phi

3 main challenges on the road to higher scalability

(ref. ENES workshop)

- *Data assimilation*
  - *Exploiting high resolution observational data, data integration, data analytics, data staging to improve model predictions; ability to do both model-observation and model-model intercomparison with high-resolution simulation and observation data sets.*
- *I/O challenge*
  - *Probably an area in which this community can drive and lead the way, parallel I/O may be one possible topic of collaboration*
- *Solvers*
  - *Need to reduce the time spent in MPI communication overheads by exploiting hybrid parallelism. Re-think solvers to trade FLOPS for communication (either across the network or levels of the memory hierarchy) or to trade global communication operations for those spanning a local neighborhood.*

View from an Earth scientist at Intel

- *Increasingly deep hierarchies for data movement: NUMA/cache levels; local vs. wide neighborhoods on the interconnect; even eventually in the I/O subsystem. Must be aware of the costs of accessing data, and must consider making good reuse of data (blocking/tiling techniques).*

- *Careful partitioning of data to reduce communication volume*

- *Structure code to maximize exposure of parallelism across many levels (vector, threads, MPI ranks). Vectorization will be increasingly important as scalar performance will most likely continue to diminish.*

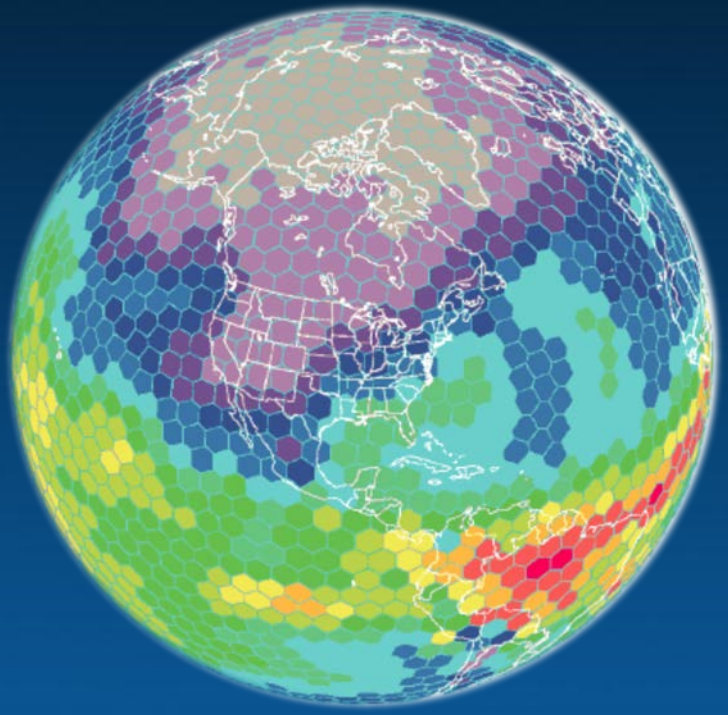
- *Load balancing of increasingly critical importance, especially for coupled climate models than to other codes, because they incorporate very disparate physics/chemistry operating across different temporal and spatial scales, which are computed by different components.*

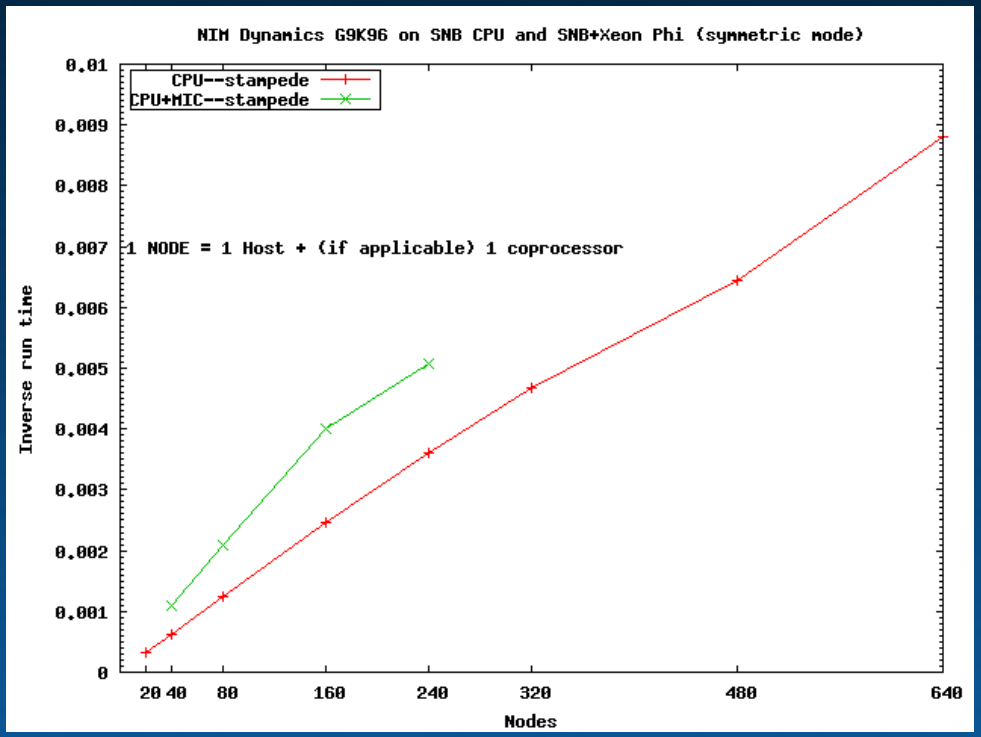
- *Likely need to consider higher-order discretization methods as a means to get more favorable computation to data access ratios (compared to using lower-order methods with higher-resolution grids).*

Richard Mills joined Intel's Software and Services Group as an HPC Earth System Models Architect in January 2014. Prior to that, he spent a decade as a research scientist at Oak Ridge National Laboratory, where most recently he was a Computational Earth Scientist in the Oak Ridge Climate Change Science Institute

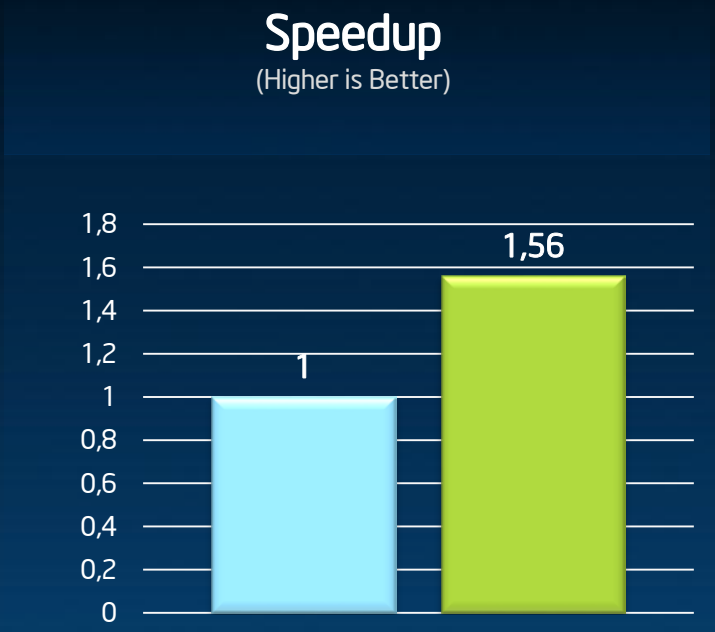
NONHYDROSTATIC ICOSAHEDRAL MODEL (NIM)

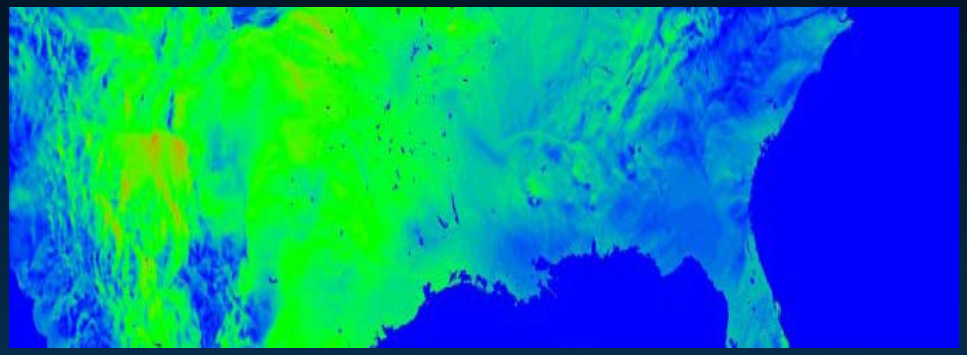
- U.S. National Oceanic and Atmospheric Administration's (NOAA) nonhydrostatic global cloud-resolving finite-volume icosahedral model
- Dynamics-only run on Stampede system at TACC, G9 resolution (~14 km), 96 vertical levels
- Acknowledgments: Jim Rosinski et al., NOAA





WEATHER RESEARCH AND FORECASTING (WRF)





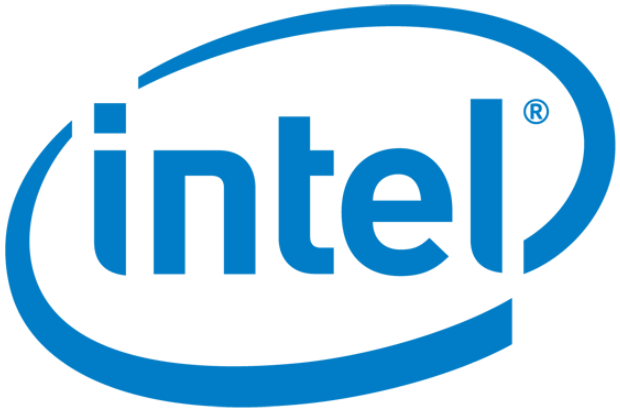
- **Application:** WRF
- **Code Optimization:**
  - Approximately two dozen files with less than 2,000 lines of code were modified (out of approximately 700,000 lines of code in about 800 files, all Fortran standard compliant)
  - Most modifications improved performance for both the host and the co-processors
- **Performance Measurements:** V3.5 and U.S. National Center for Atmospheric Research (NCAR) supported CONUS2.5KM benchmark (a high resolution weather forecast)
- **Acknowledgments:** There were many contributors to these results, including the National Renewable Energy Laboratory and The Weather Channel Companies

Acknowledgements

- Jim Cownie, Intel
- Richard Mills, Intel

Questions?

Marie-christine.sawley@intel.com


intel