HPC Technologies for Weather and Climate Simulations

High Performance Computing

13th Workshop on the Use of High Performance Computing in Meteorology

Shinfield Park, Reading, UK
Nov. 3-7, 2008

David Barkai, Ph.D.
HPC Computational Architect
High Performance Computing
Digital Enterprise Group
Intel Corporation

What we’ll talk about

- The Big Picture
- Nehalem is coming..
- NWS on Clusters

The Big Picture..

Intel HPC Vision

A world in which Intel based supercomputers enable the major breakthroughs in science, medicine & engineering…

From exploration to production

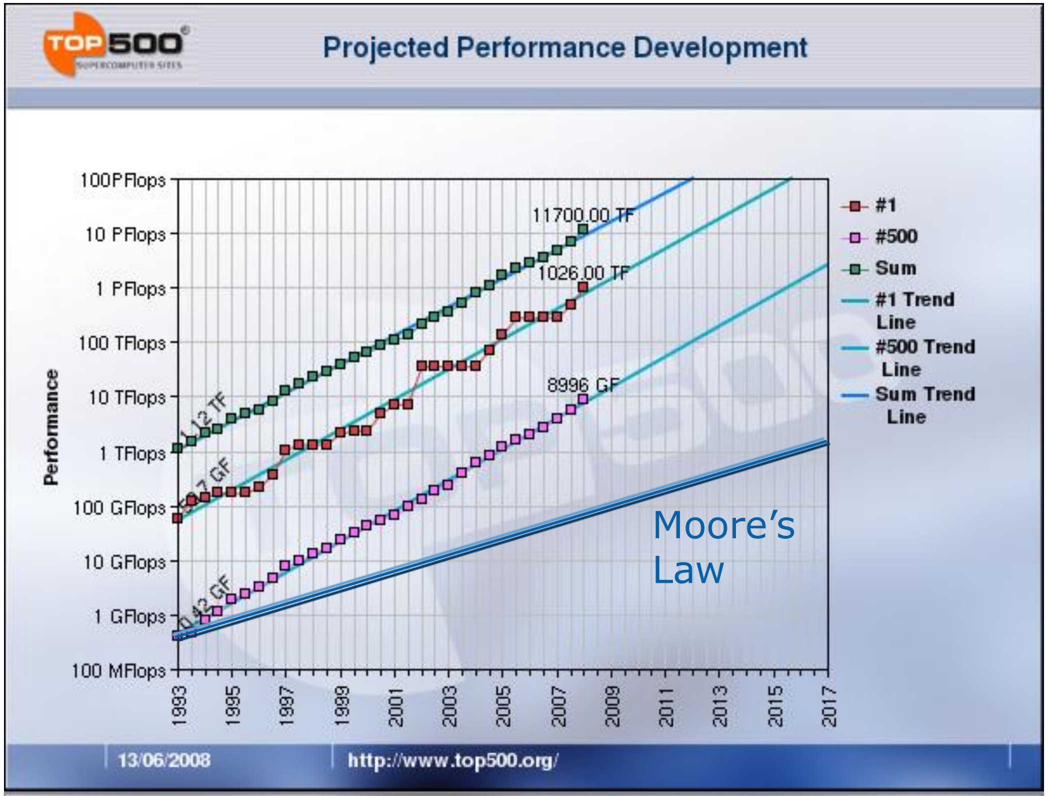
Mission:

- Create a maintain a technology leadership position for Intel at the highest end of computing – drive the path to TeraScale processors & ExaScale systems
- Drive a valuable technology pipeline for Intel's volume business
- Grow the use of HPC across all segments from the office to the datacenter

Inertia For The Insatiable Demand For Performance

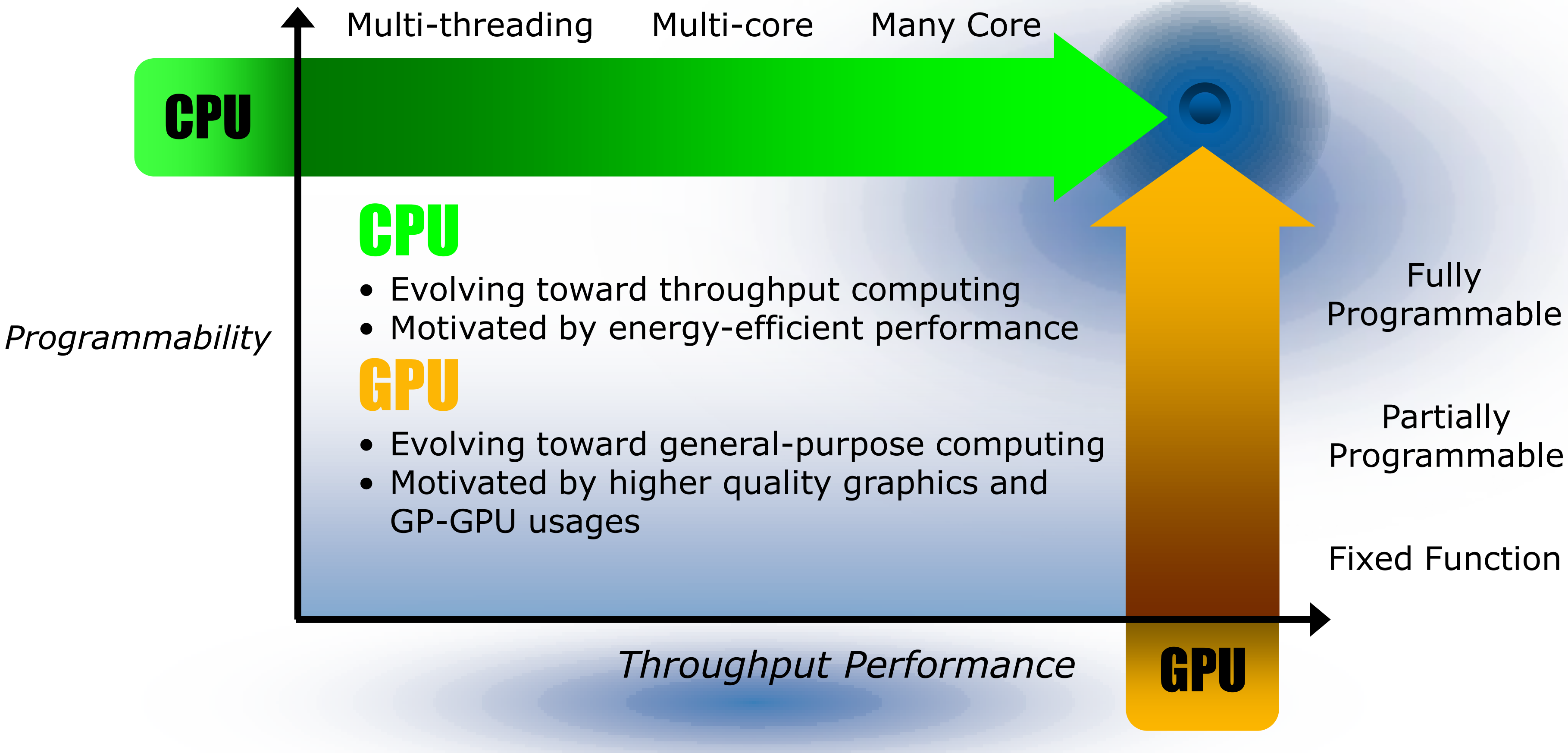
Moore’s Law:

- Transistor Density doubles every two years
- More than 40 years old
- 37 years of “free” performance gains
- Qualitative change a few years ago: Multicore
  - High performance requires multithreading
  - Intel Xeon 7400 series just announced with 6 cores



Large scale deployments consistently outpacing Moore’s Law

In Search Of (Even) More Performance



CPU

- Evolving toward throughput computing
- Motivated by energy-efficient performance

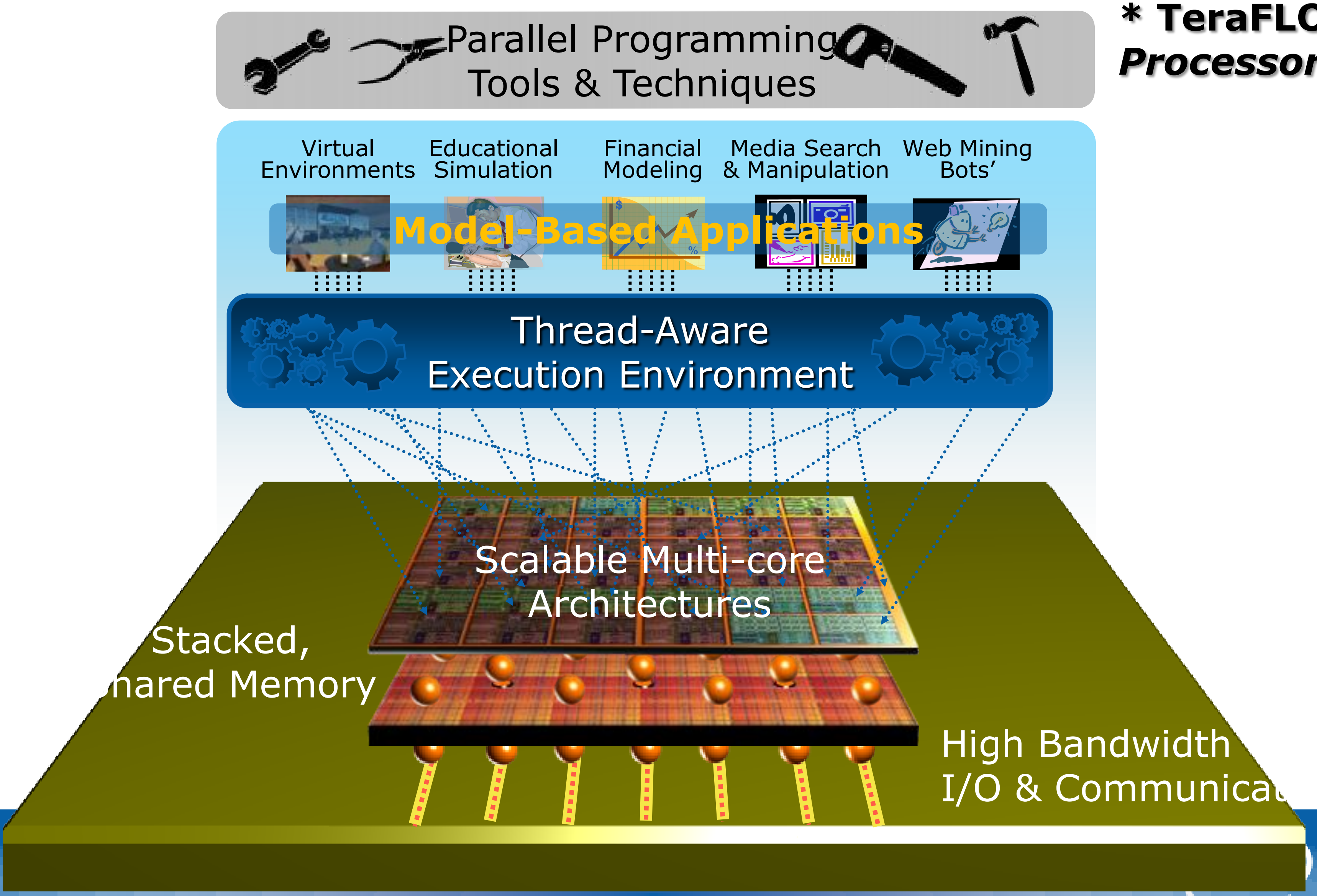
GPU

- Evolving toward general-purpose computing
- Motivated by higher quality graphics and GP-GPU usages

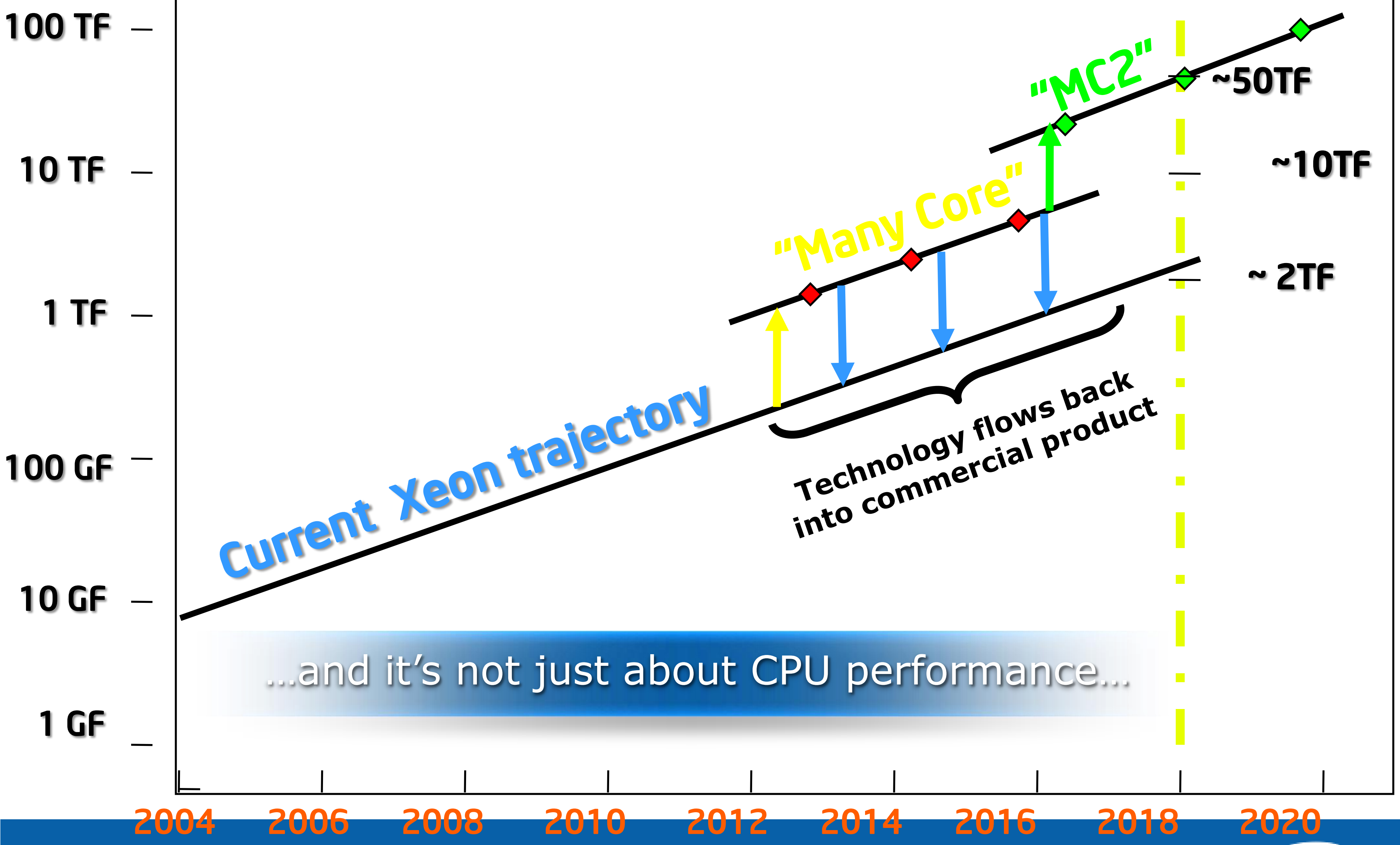
Architecture Evolution: A Collision Course?

Intel’s Terascale* Research Program

* TeraFLOP *Processors*



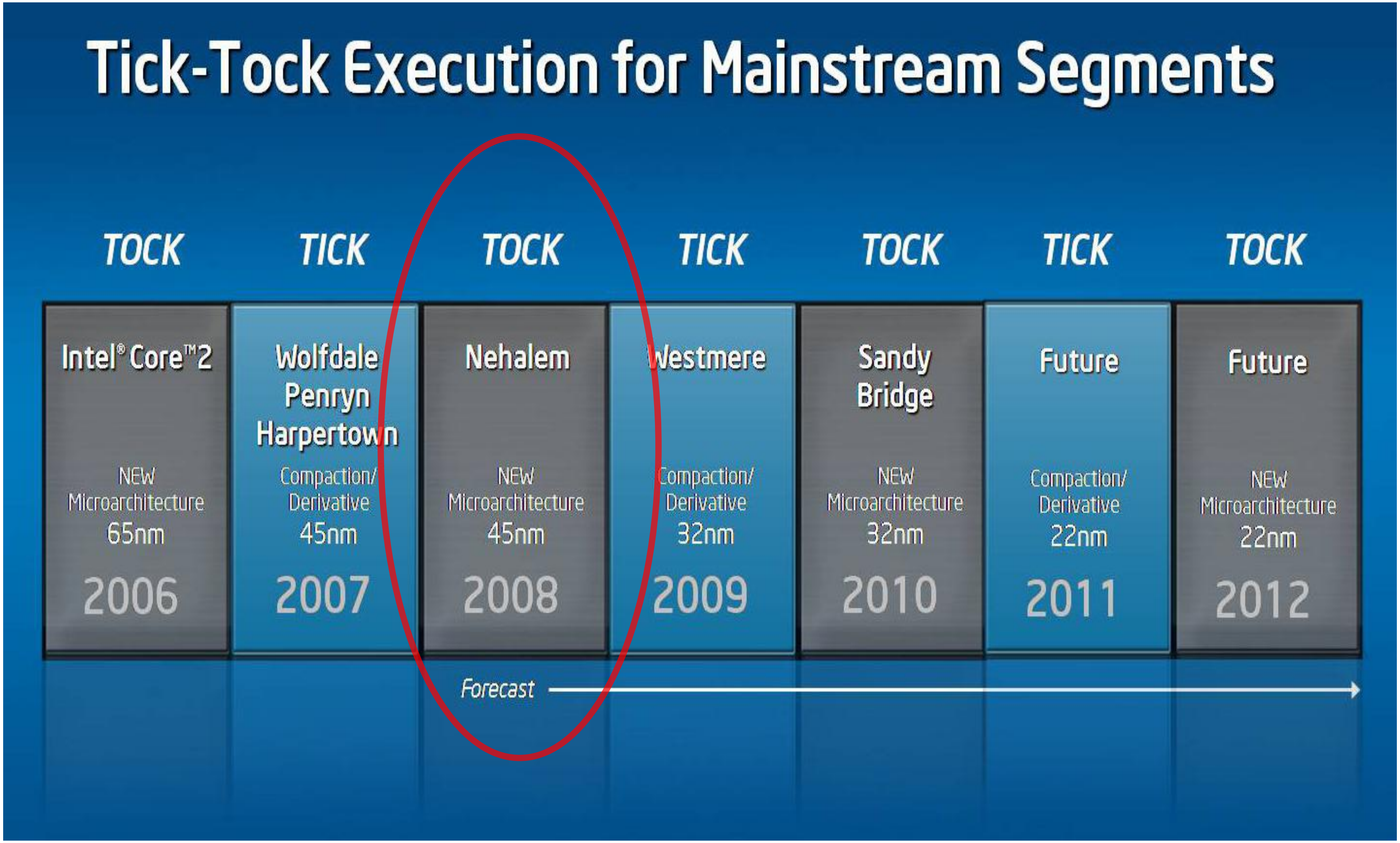
The Path To ExaScale Systems



…and it's not just about CPU performance…

Conceptual roadmap – not for planning purposes

What’s coming for Servers by Intel



Tick-Tock Execution for Mainstream Segments

TOCK	TICK	TOCK	TICK	TOCK	TICK	TOCK
Intel® Core™2	Wolfdale Penryn Harpertown	Nehalem	Westmere	Sandy Bridge	Future	Future
NEW Microarchitecture	Compaction/ Derivative	NEW Microarchitecture	Compaction/ Derivative	NEW Microarchitecture	Compaction/ Derivative	NEW Microarchitecture
65nm	45nm	45nm	32nm	32nm	22nm	22nm
2006	2007	2008	2009	2010	2011	2012

Forecast →

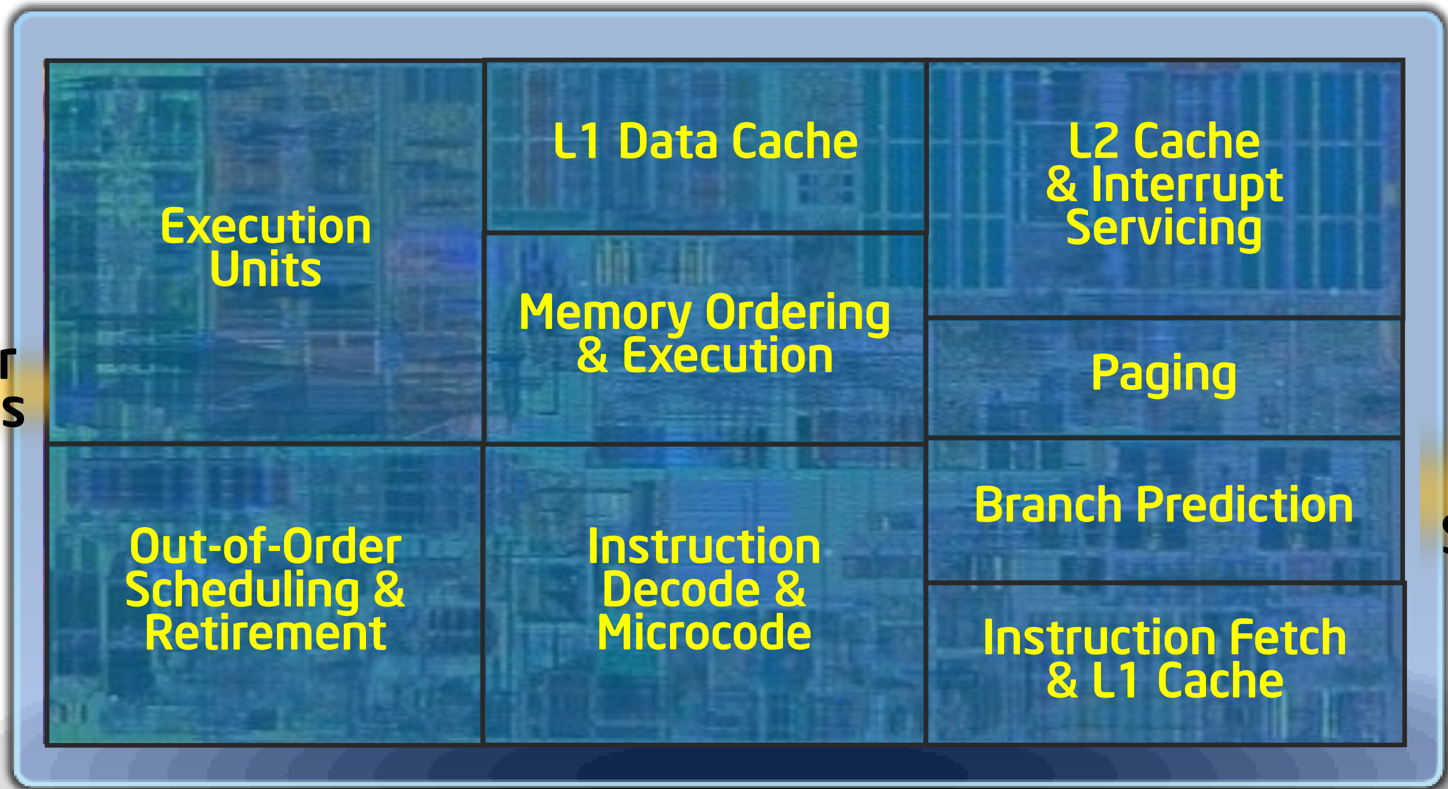
Nehalem Core: Recap

New SSE4.2 Instructions

Improved Lock Support

Additional Caching Hierarchy

Deeper Buffers


Execution Units
L1 Data Cache
L2 Cache & Interrupt Servicing
Memory Ordering & Execution
Paging
Branch Prediction
Out-of-Order Scheduling & Retirement
Instruction Decode & Microcode
Instruction Fetch & L1 Cache

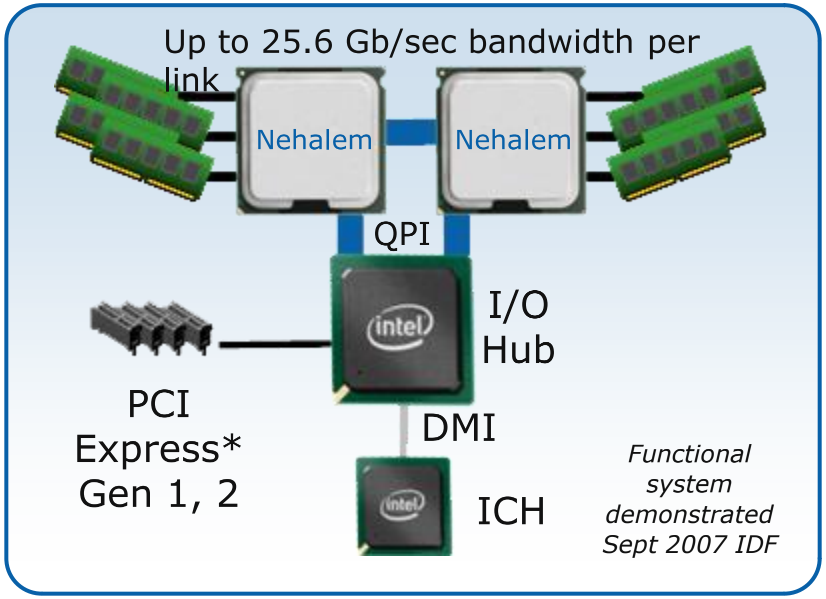
Improved Loop Streaming

Simultaneous Multi-Threading

Faster Virtualization

Better Branch Prediction

Nehalem Based System Architecture


Up to 25.6 Gb/sec bandwidth per link
Nehalem
Nehalem
QPI
I/O Hub
PCI Express* Gen 1, 2
DMI
ICH
Functional system demonstrated Sept 2007 IDF

Benefits

- More application performance
- Improved energy efficiency
- Improved Virtualization Technology

Key Technologies

New 45nm Intel® Microarchitecture

New Intel® QuickPath interconnect

Integrated Memory Controller

Next Generation Memory (DDR3)

PCI Express Gen 2

Extending Today’s Leadership

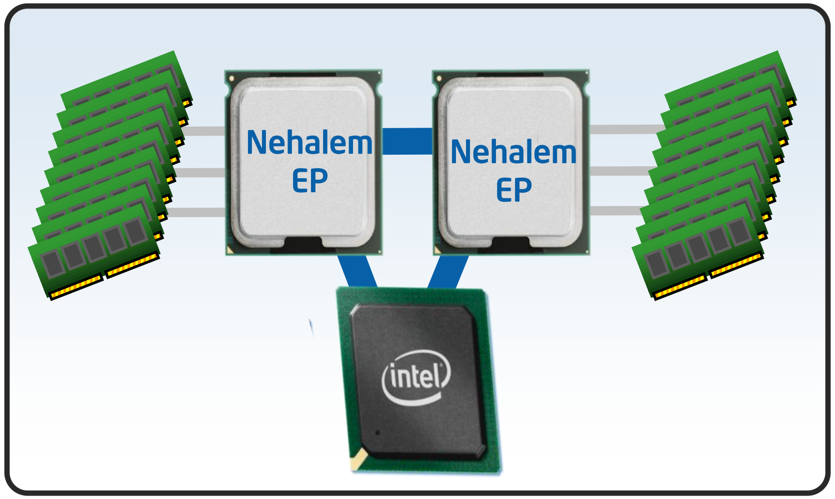
Production Q4’08

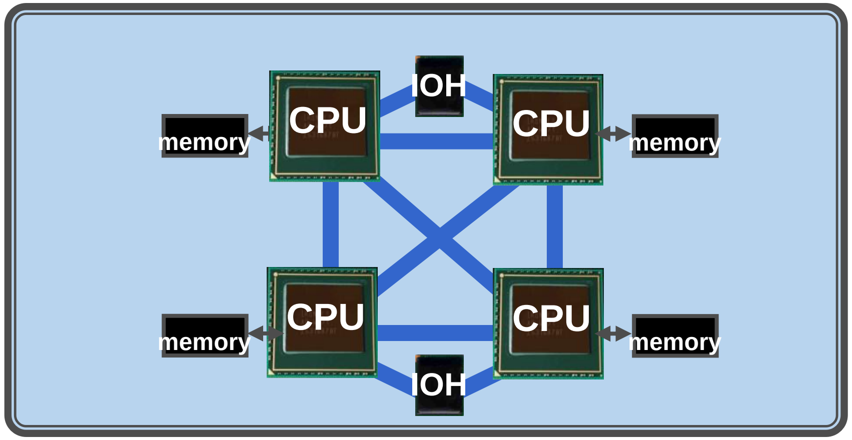
Volume ramp Q1’09

All future products, dates, and figures are preliminary and are subject to change without any notice.

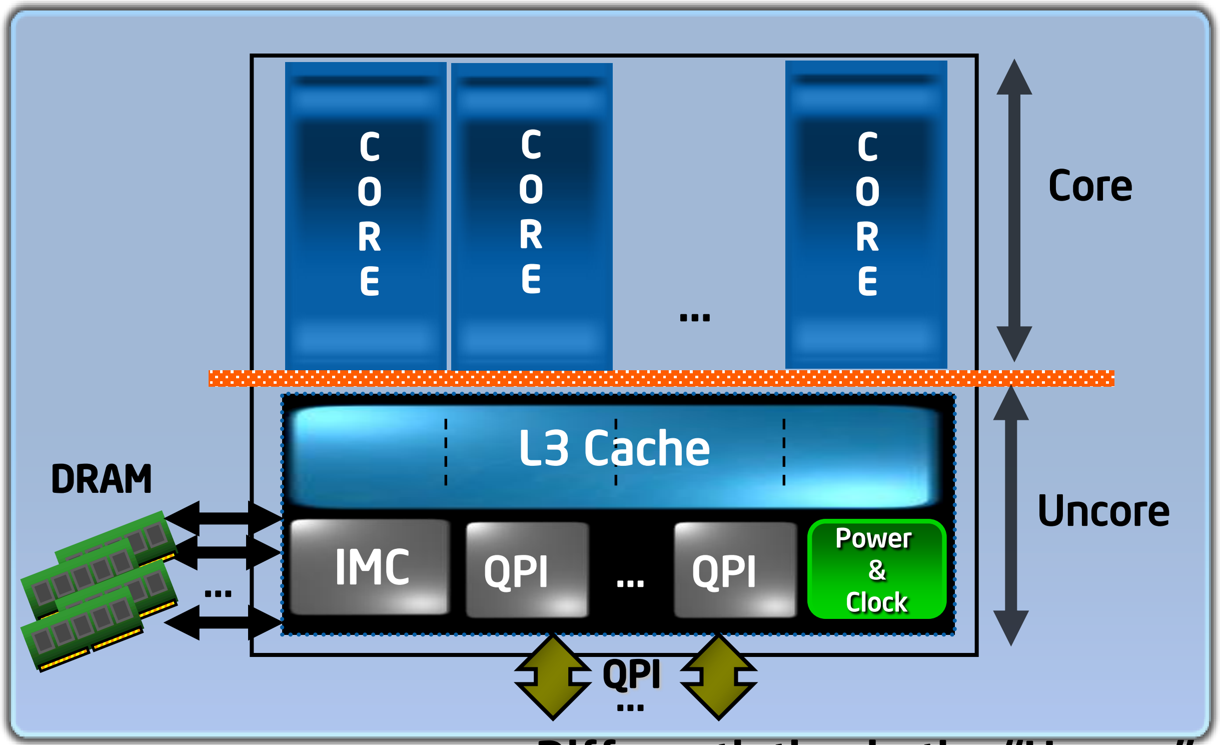
QuickPath Interconnect

- Nehalem introduces new QuickPath Interconnect (QPI)
- ***High bandwidth***, ***low latency*** point to point interconnect
- Up to 6.4 GT/sec initially
  - 6.4 GT/sec -> 12.8 GB/sec
  - Fully duplex -> 25.6 GB/sec per link
  - Future implementations at even higher speeds
- Highly ***scalable*** for systems with varying # of sockets





Core/Uncore Modularity



Nehalem Core

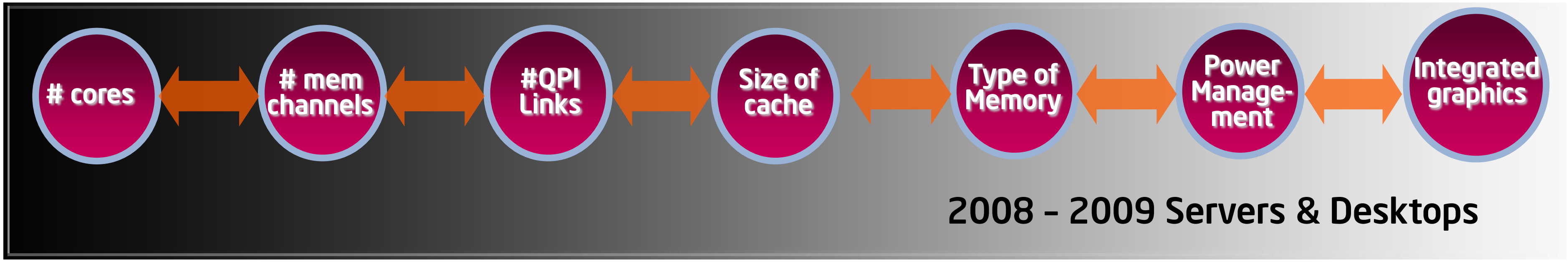
Common from Mobile to Server

Nehalem Uncore

Differentiates the product segments

QPI: Intel® QuickPath Interconnect

Differentiation in the "Uncore":



2008 - 2009 Servers & Desktops

Optimal price / performance / energy efficiency for server, desktop and mobile products

Characterizing NWS on Clusters

- WRF
- POP
- CAM
- HOMME

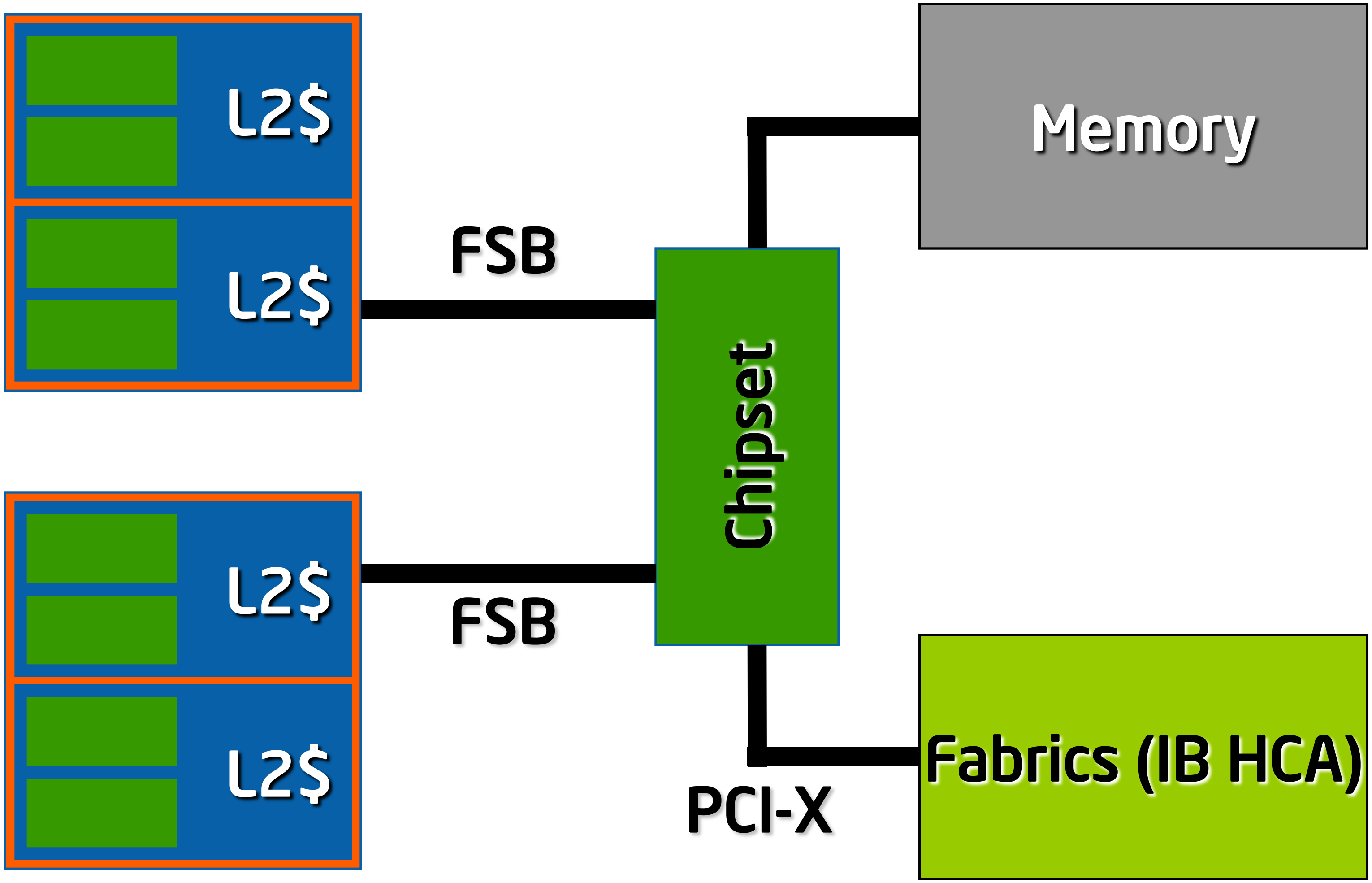
By Intel’s NWS application engineers team:

Roman Dubtsov
Alexander Semenov
Shkurko Dmitry
Alex Kosenkov
and Mike Greenfield

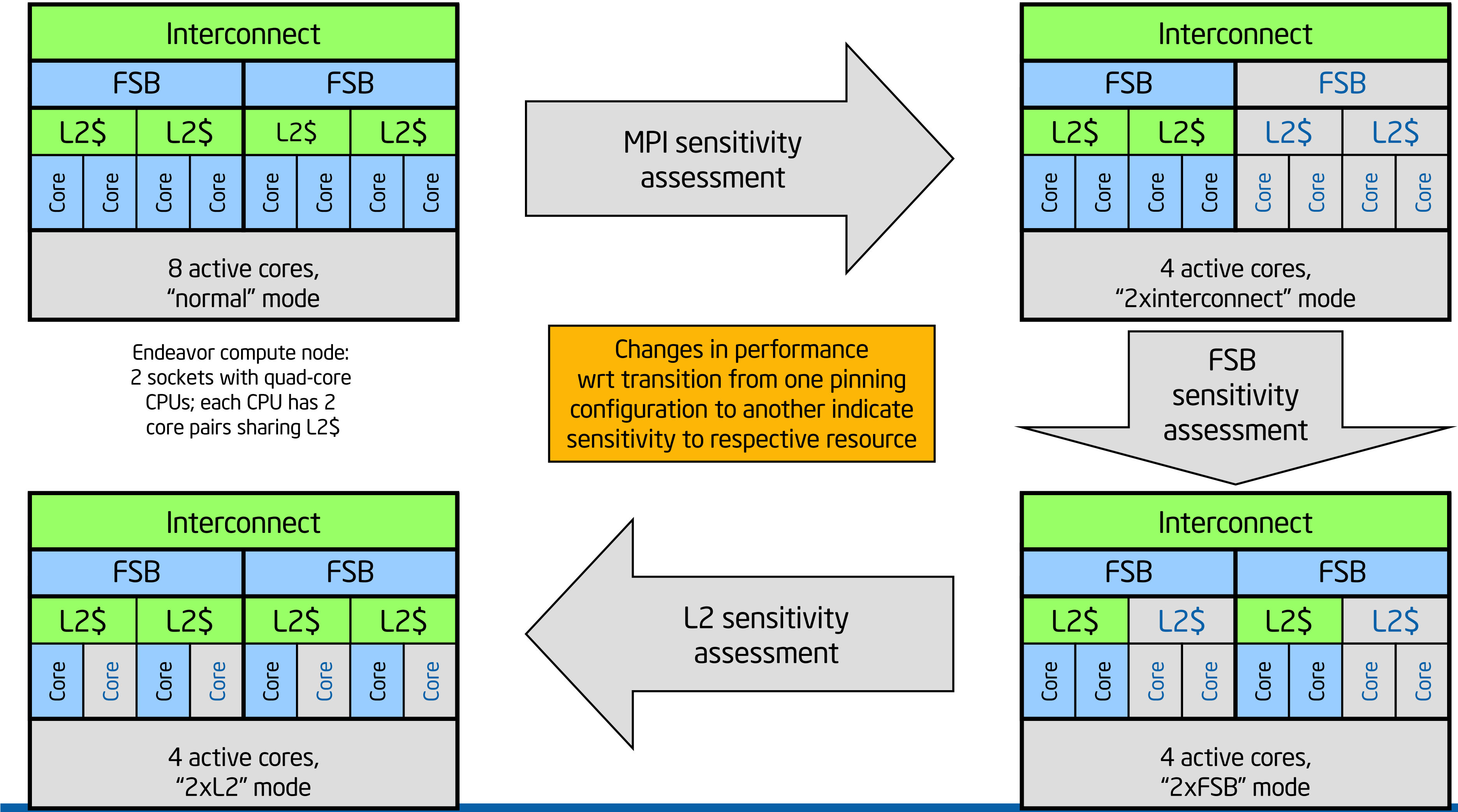
Characterization methodology overview

- The characterization should allow answering the questions:
  - How scalable is the application?
  - Is the application bandwidth limited?
  - How strong is MPI and IO impact?
  - How it will work on other platforms?
- Tools used for characterization:
  - VTune is used for FSB Utilization measurement
  - IOP* (AFT based tool) for IO impact evaluation
  - IPM (http://ipm-hpc.sf.net) for MPI related measurements
- Methodology:
  - Profiles collected for different process pinning configurations that incrementally increase resources available to benchmark.
  - On each transition performance improvements are noted and profiles compared.

Components for Pinning Setup



Process Pinning Setup



Interconnect

FSB FSB

L2$ L2$ L2$ L2$

Core Core Core Core Core Core Core Core

8 active cores,
“normal” mode

Endeavor compute node:
2 sockets with quad-core
CPUs; each CPU has 2
core pairs sharing L2$

MPI sensitivity
assessment

Interconnect

FSB FSB

L2$ L2$ L2$ L2$

Core Core Core Core Core Core Core Core

4 active cores,
“2xinterconnect” mode

FSB
sensitivity
assessment

Interconnect

FSB FSB

L2$ L2$ L2$ L2$

Core Core Core Core Core Core Core Core

4 active cores,
“2xFSB” mode

L2 sensitivity
assessment

Interconnect

FSB FSB

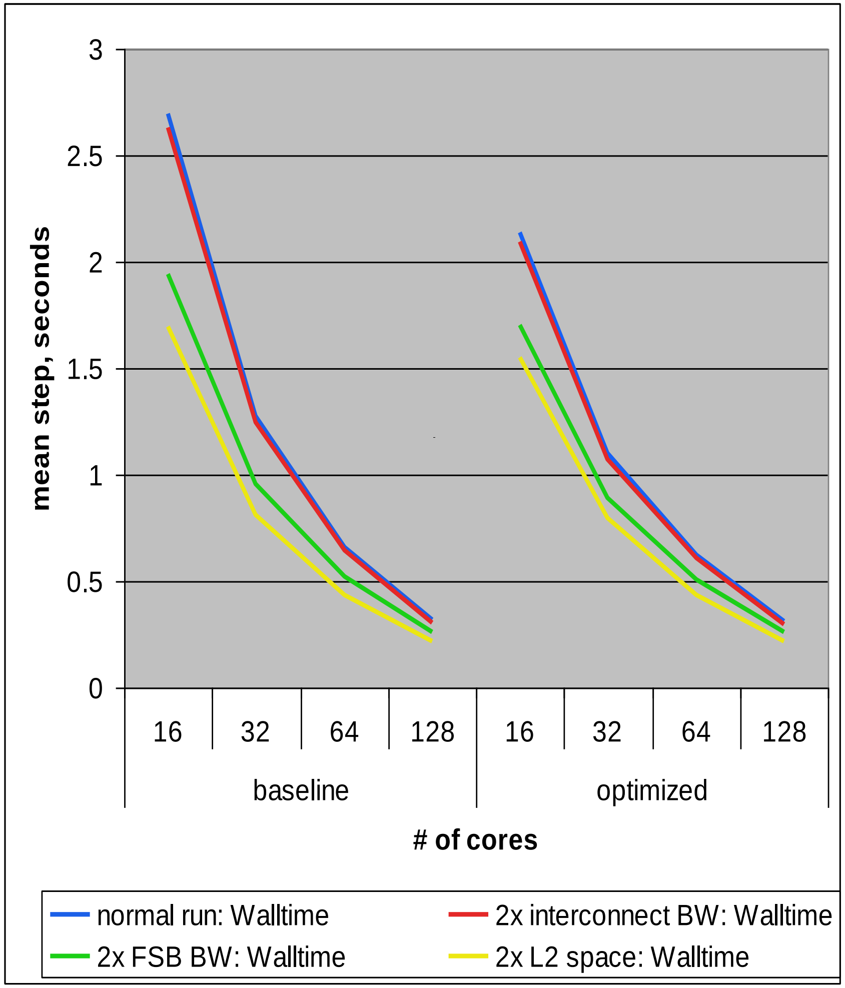
L2$ L2$ L2$ L2$

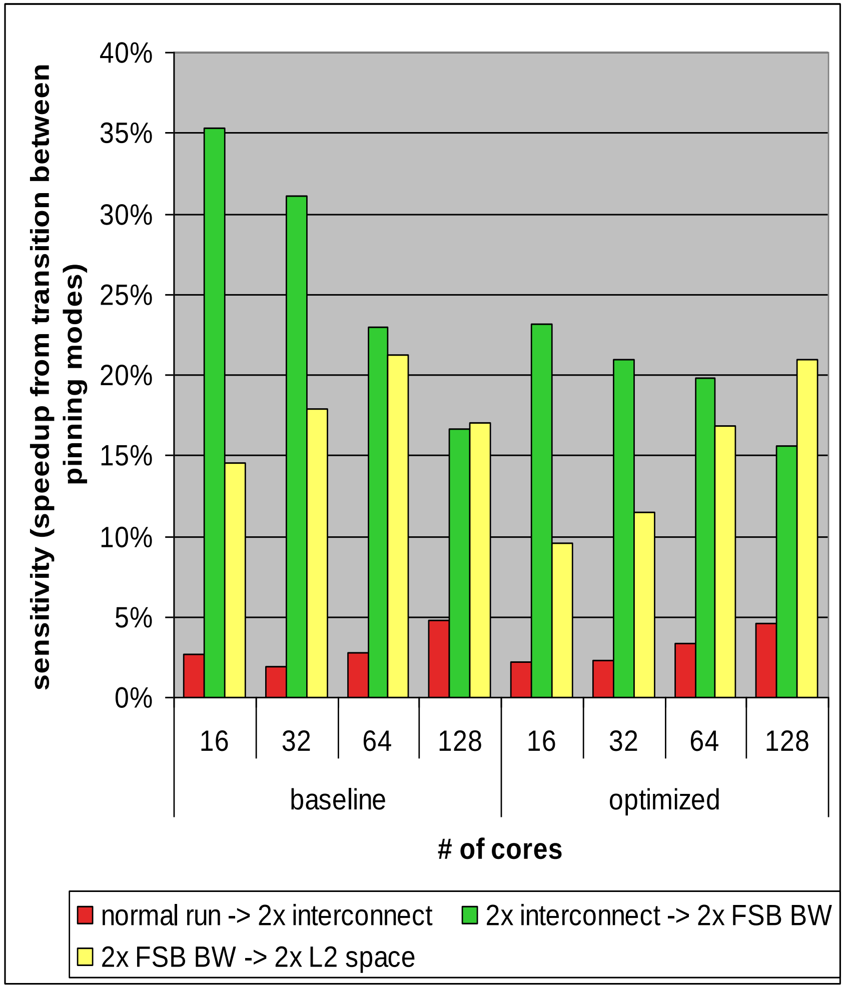
Core Core Core Core Core Core Core Core

4 active cores,
“2xL2” mode

Changes in performance
wrt transition from one pinning
configuration to another indicate
sensitivity to respective resource

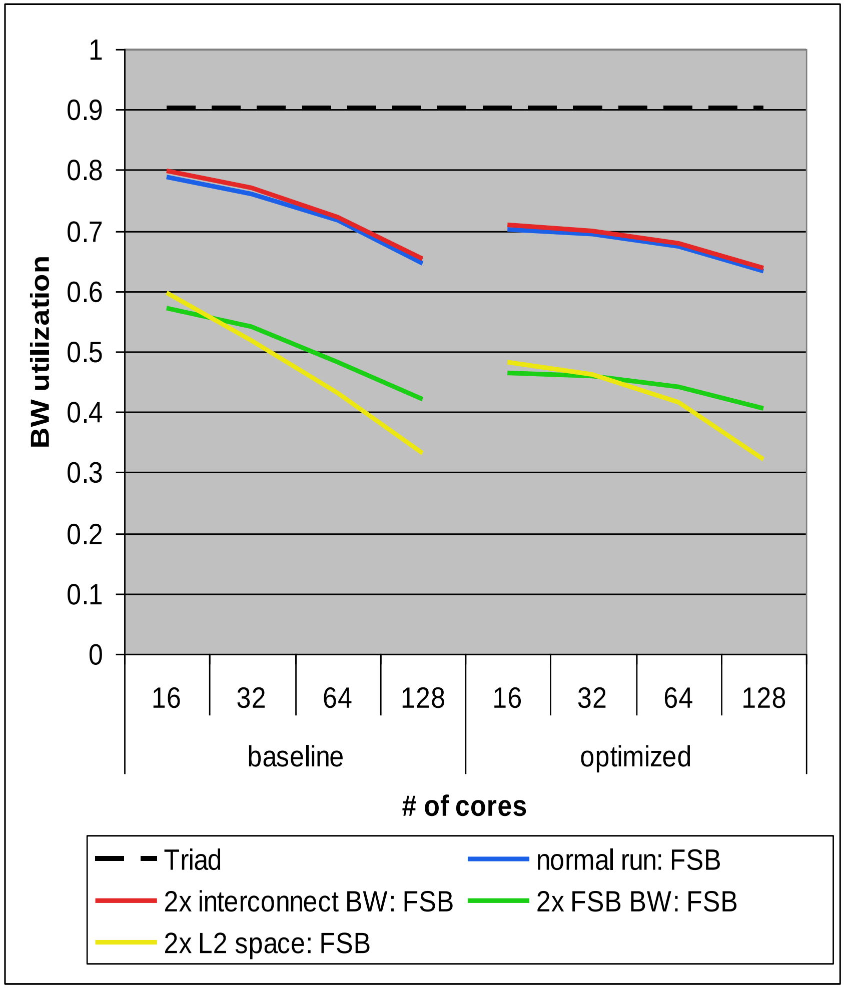
WRF3.0/CONUS12: Performance

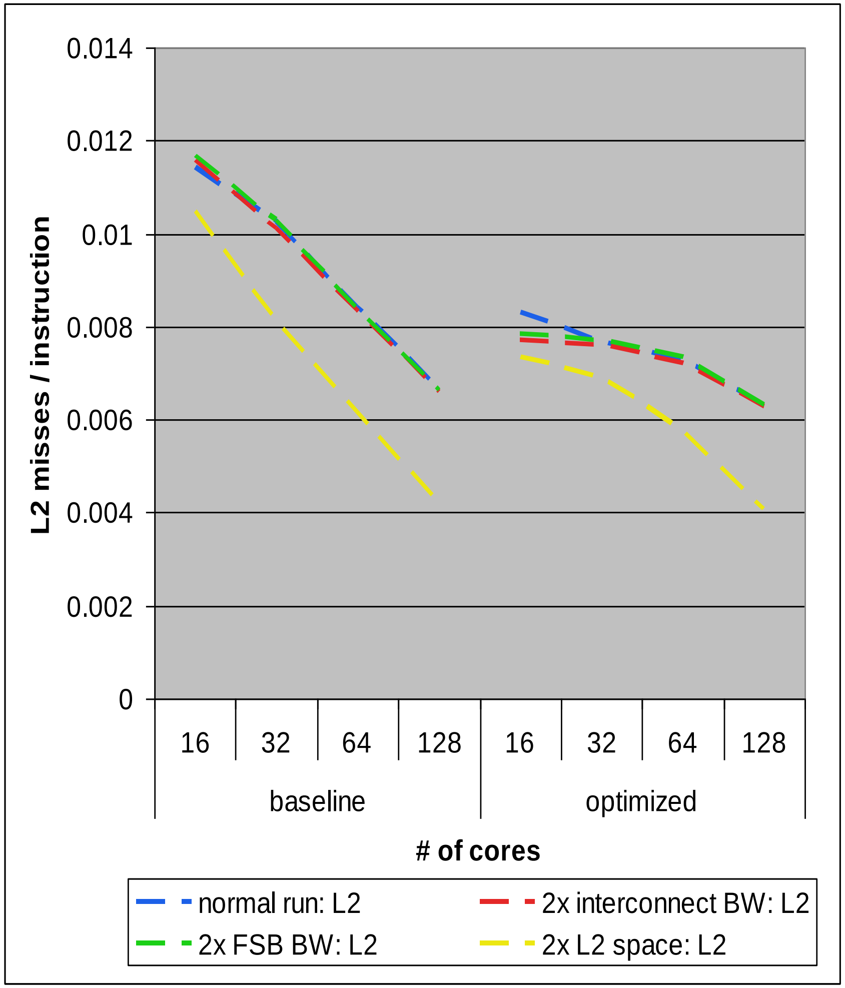




Clear dependency on FSB bandwidth. Optimized configurations show somewhat lower sensitivity to FSB and L2$. Not sensitive to interconnect.

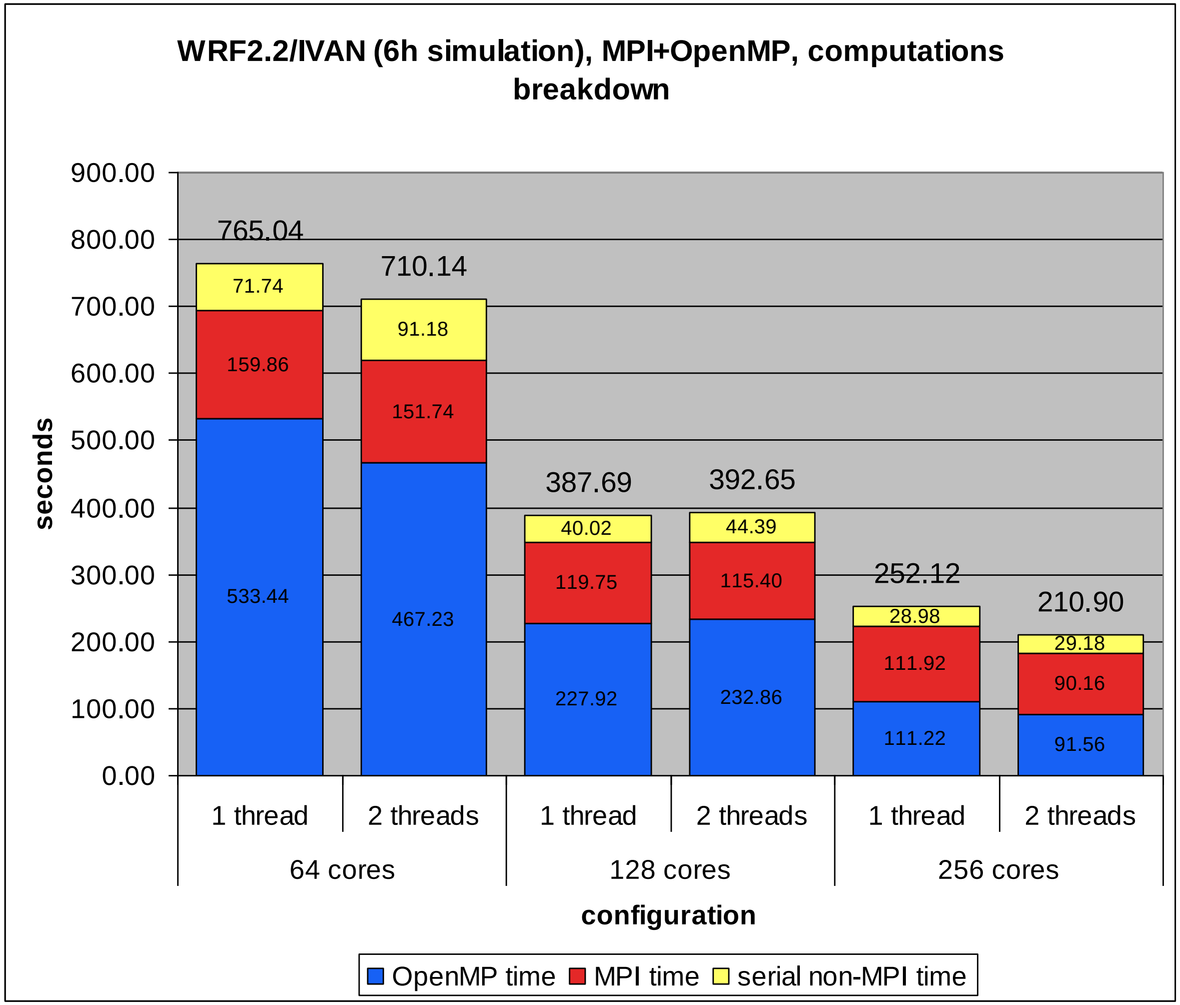
WRF3.0/CONUS12: Memory





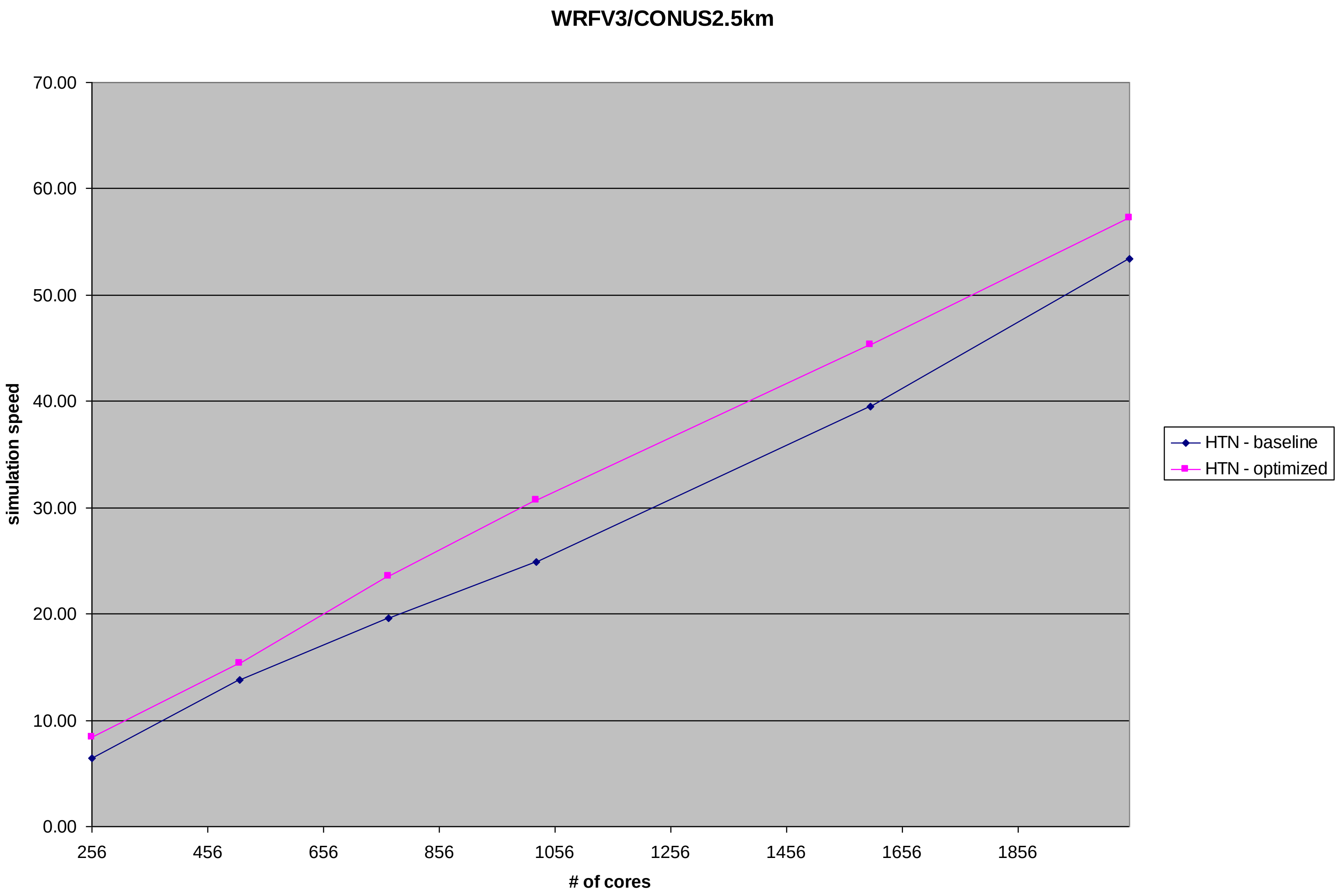
Optimized configuration shows better FSB and L2$ utilization.

WRF2.2/IVAN: Hybrid helps..



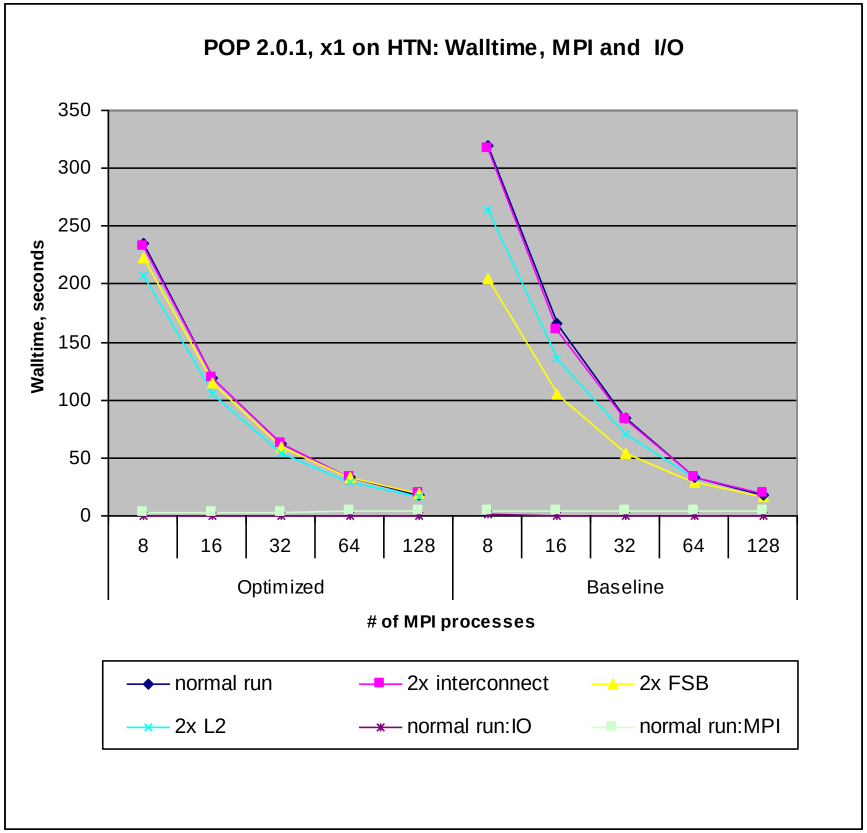
- Configuration for N cores is either N MPI processes or N/2 MPI processes 2 OpenMP threads each
- This breakdown was computed using Intel® Trace Collector, Intel® Thread Checker and profiling OpenMP library from Intel® Fortran/C Compilers
- “OpenMP time” is time spent in OpenMP regions regardless of number of threads.
- There are some improvements in serial parts

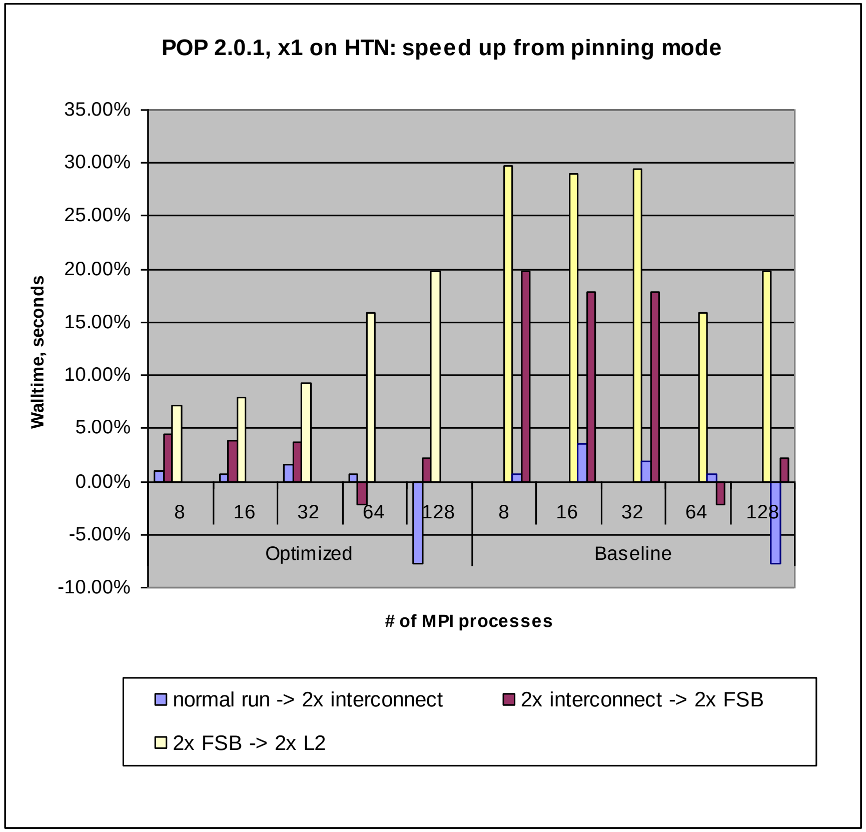
Even better for WRF CONUS 2.5km



POP/x1 characterization: Harpertown

Assessment summary

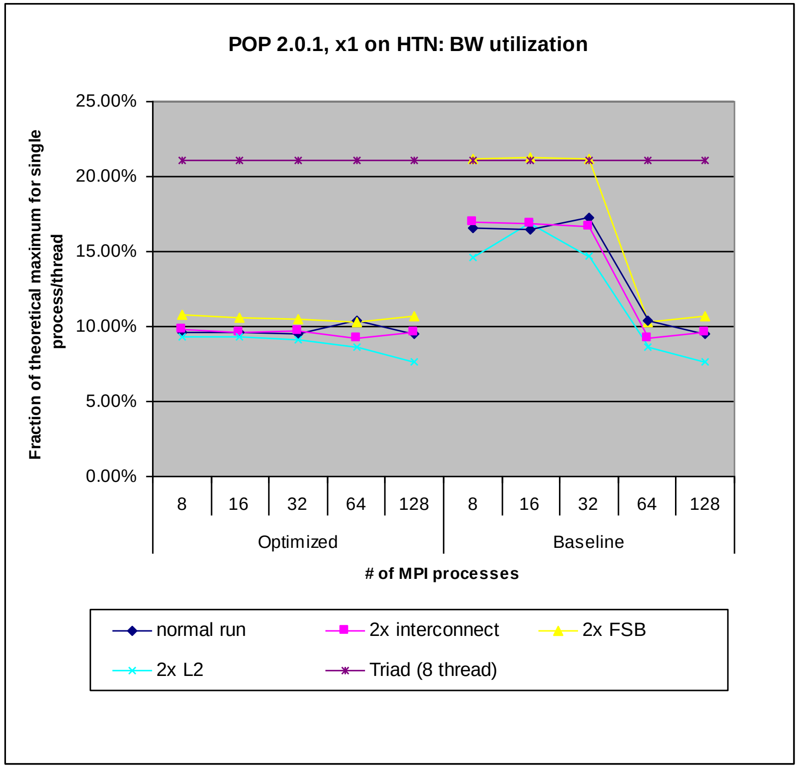


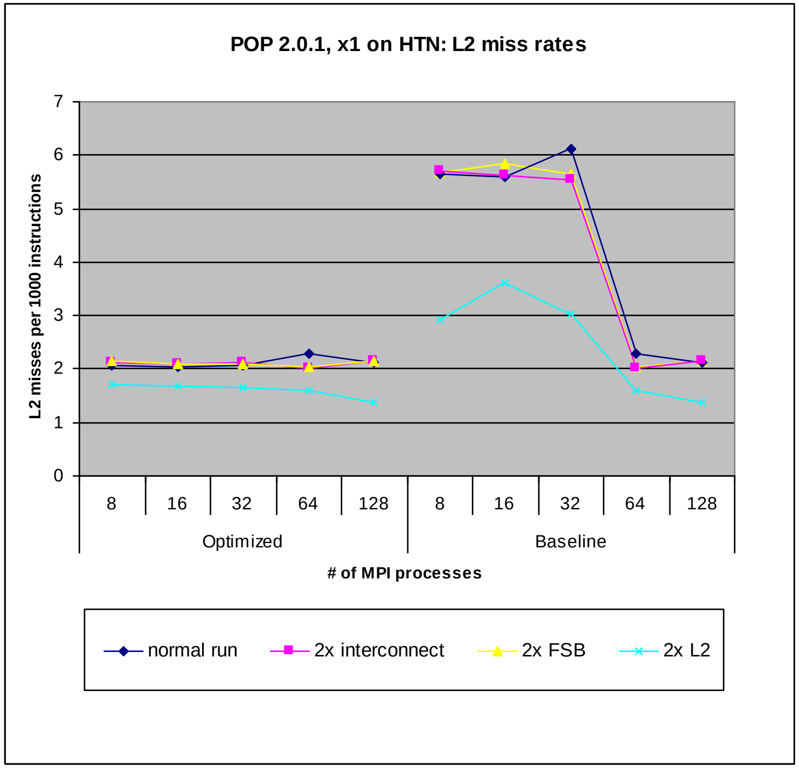


Baseline configuration is expressively FSB limited on lower core counts as there is significant speedup from 2x-interconnect to 2x-FSB runs.
On higher core counts additional interconnect BW and L2 start to give benefit.
Speedups from 2x-interconnet BW are low.

POP/x1 characterization: Harpertown

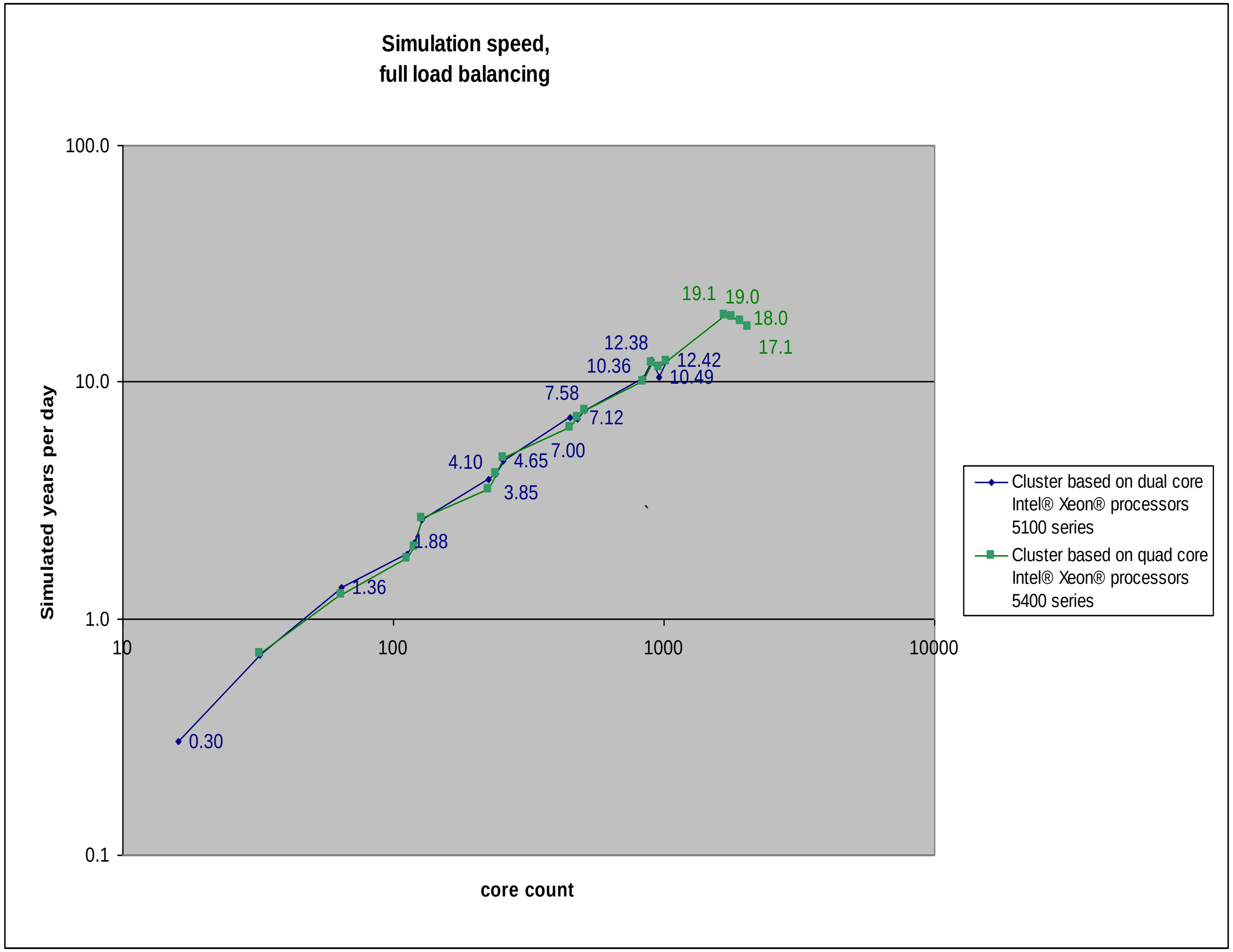
Memory subsystem impact assessment





Workload is sensitive to the cache size (working set is comparable to cache size).
On average FSB is not saturated completely and “2x FSB“ configuration consumes 0.5BW of Triad.
Optimized version shows consistent behavior.

CAM performance



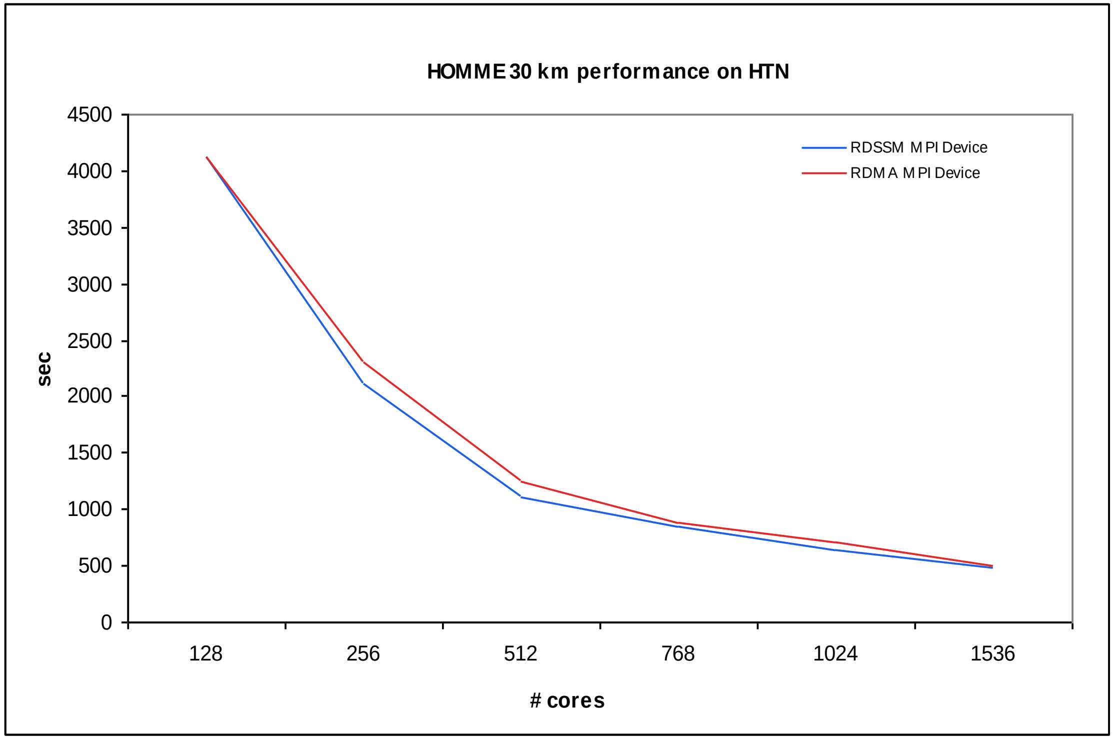
- •Drops in the curve are explained by suboptimal decomposition in certain points
- •The same decompositions and tuning options were used for both configurations

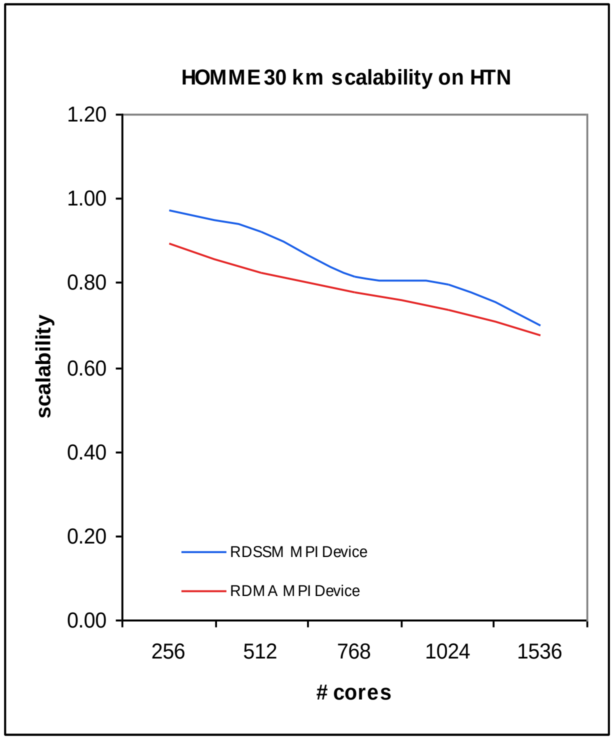
- •Based on the 3- or 6-day forecasts
- •Speed is calculated from integration time (“stepon” timer)

CAM, D-Grid workload, scales to 2,000 cores on Infiniband cluster

HOMME 30km characterization:

Scalability assessment





Ideal scalability on small core counts. And still good at larger core counts.

RDSSM MPI Device improves scalability and walltime up to 9% comparing to RDMA.

Dependence on interconnect, bandwidth, and MPI are moderate

In Closing..

It’s an exciting time to be in HPC

The HPC demand is high and growing faster than the general server market

We begin to gain quantified understanding about the opportunities of deploying large clusters for NWS intel®

Leap ahead™

Backup