Hybrid parallelism for Weather Research and Forecasting Model on Intel® platforms

(performance evaluation)

Roman Dubtsov*, Mark Lubin, Alexander Semenov

{roman.s.dubtsov,mark.lubin,alexander.l.semenov}@intel.com

December 15, 2008

* Corresponding author

Legal disclaimers

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web Site.

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 http://www.intel.com/performance/resources/limits.htm or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

All dates and products specified are for planning purposes only and are subject to change without notice

Relative performance is calculated by assigning a baseline value of 1.0 to one benchmark result, and then dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms, and assigning them a relative performance number that correlates with the performance improvements reported.

Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor series, not across different processor sequences. See http://www.intel.com/products/processor_number for details.

Intel products are not intended for use in medical, life saving, life sustaining, critical control or safety systems, or in nuclear facility applications. All dates and products specified are for planning purposes only and are subject to change without notice

* Other names and brands may be claimed as the property of others.

Agenda/Overview

- Motivation & Setup
- Hybrid parallelization in WRF
- Evaluation of WRF performance on contemporary Intel® Platforms
  - Comparison & study of performance of pure MPI and hybrid MPI+OpenMP setups for multiple workloads on cluster with Intel® Xeon® E54xx processors
  - Performance results from Intel® Core™ i7 desktop processor

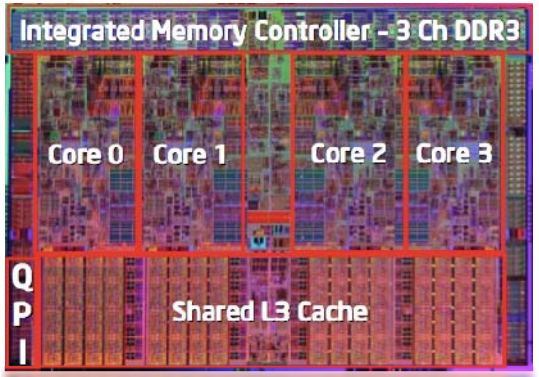
Motivation

Current CPUs designs favor hybrid MPI+OpenMP approach

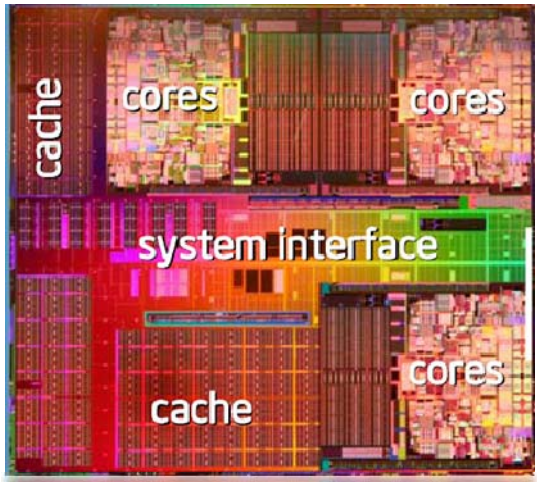
- Shared caches suggest data-sharing parallelism in OpenMP style
- Explicit synchronization in MPI style seems to be a good approach for avoiding excessive cache coherency traffic
- Many cores on the die make more fine-grained approach than single process per socket possible

Performance benefits of hybrid approach

- Lower pressure on cluster interconnect due to lower volume of data in exchanges between MPI processes (ex.: halo exchange)
- Better scalability & performance of MPI collective operations due to smaller number of processes



Intel® Core™ i7 CPU



Intel® Xeon® 7400-series CPU

Workloads & Measurements

WRF

- Weather simulation/prediction code used both for operations & research

Workloads

- CONUS12km & CONUS2.5km
  - 3h simulation over continental US with 12km/2.5km resolution
  - Single domain, computations contain point-to-point communications only
- IVAN
  - 12h simulation of hurricane Ivan (September 2004)
  - Nested domain, computations contain collective operations that pass data from/to nested domain

Methodology/Tools

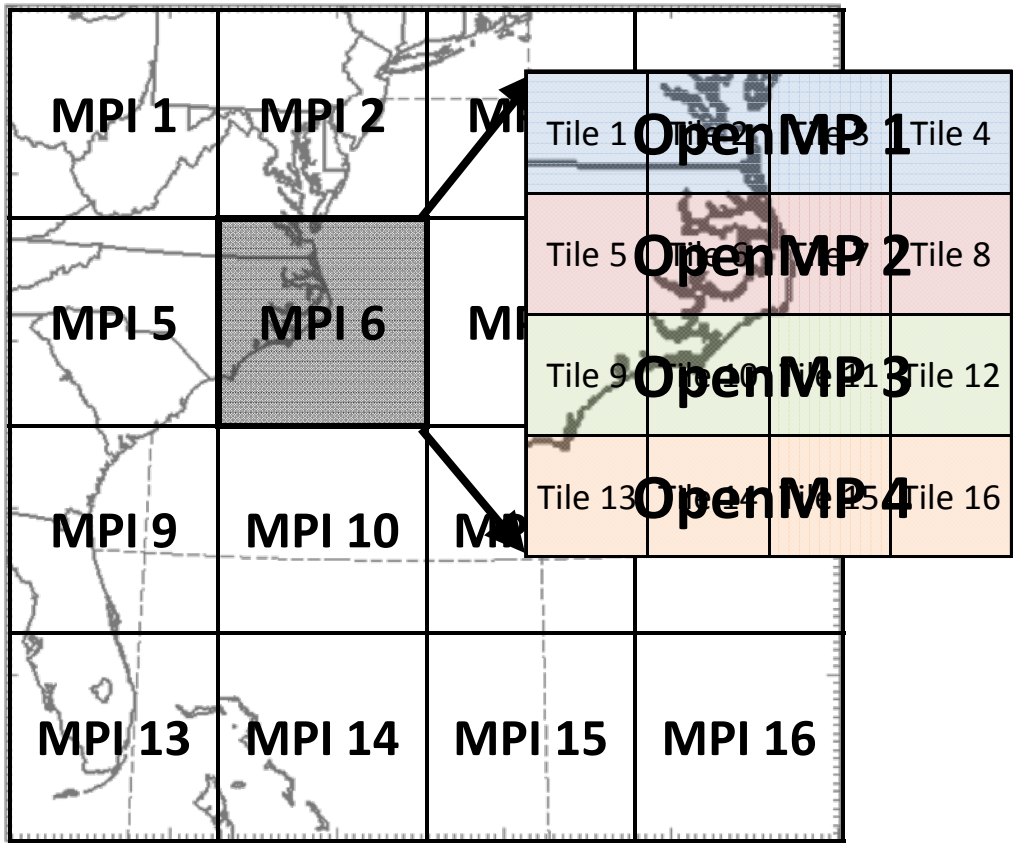
- Built-in OpenMP profiler from Intel® C/C++/Fortran compilers
- Intel® Trace Collector/Analyzer for MPI analysis
  - Ideal interconnect simulator for imbalance assessment

Hardware (more info in backup slides)

- 256-node DP (8 core/node) cluster with Intel® Xeon® E54xx processors
- Desktop machines with
  - Intel® Core™ 2 Extreme @ 3.2GHz processor
  - Intel® Core™ i7 @ 3.2GHz processor

Due to large amount of data processed WRF is very sensitive to memory bandwidth in workloads considered

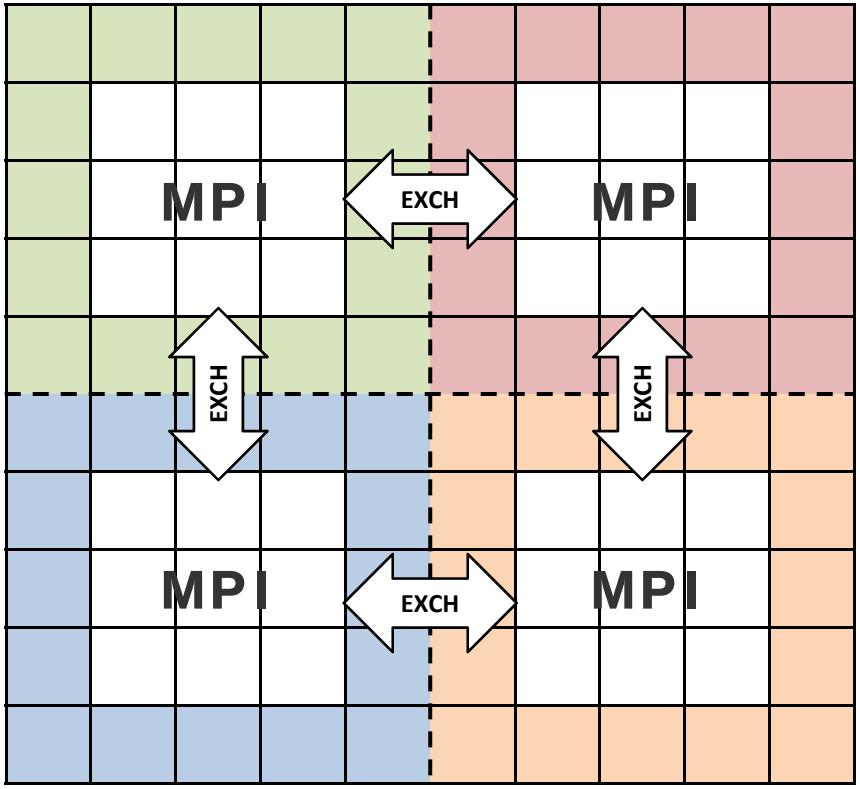
Data decomposition in WRF



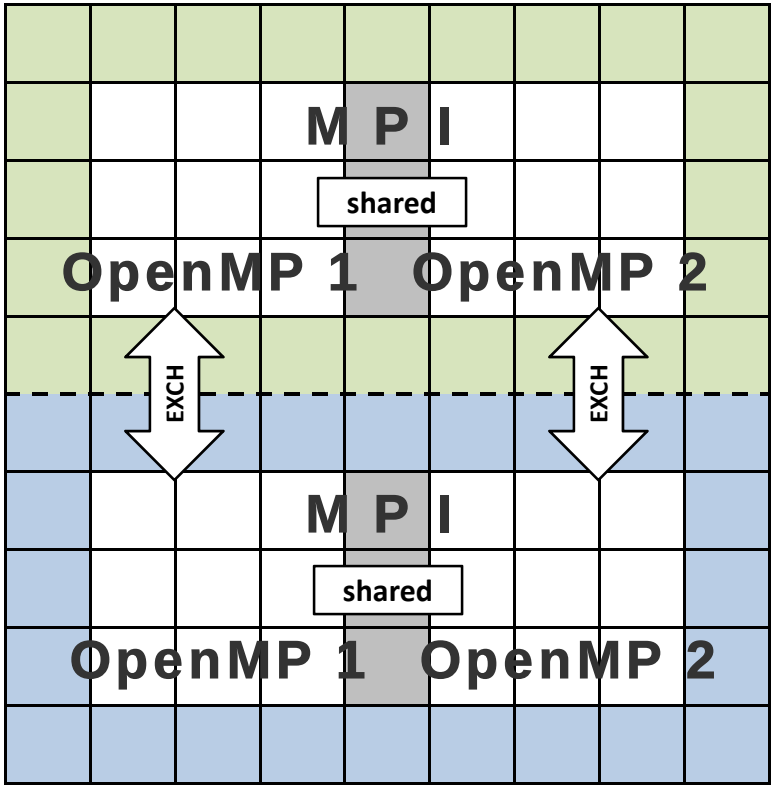
- MPI for coarse-grained domain decomposition
  - Point-to-point communications for halo exchange
  - Collective operations if there are nested domains
- Per-process domain part is further decomposed into multiple tiles
  - Multiple decomposition algorithms available that match different architectures
  - Each OpenMP thread processes one or more tiles

Hybrid setup & halo exchange

Pure MPI setup



Hybrid setup (2 threads)

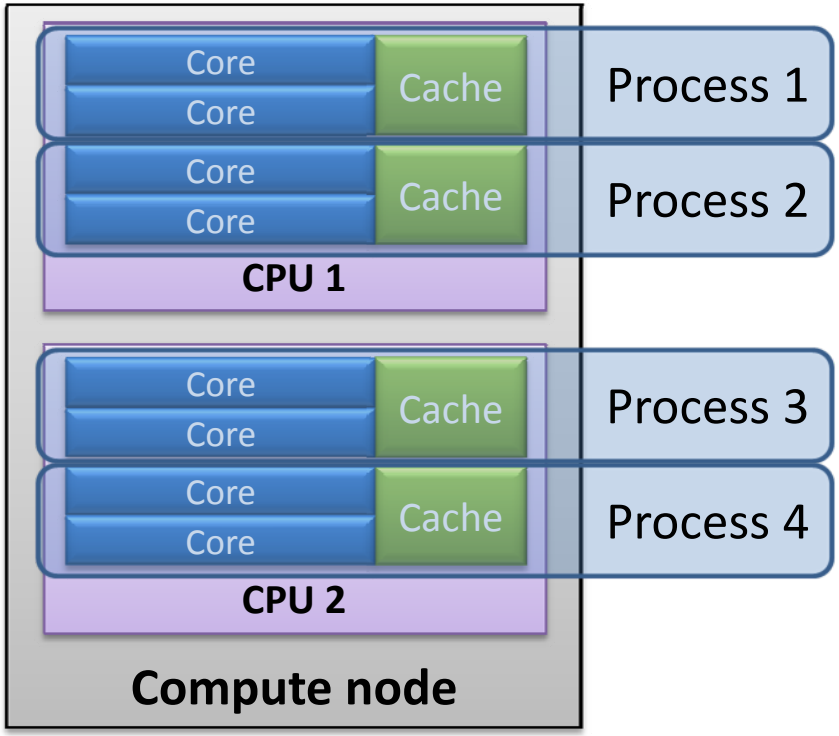


Some boundaries that were exchanged in pure MPI case are shared in hybrid case. Therefore, less data is transmitted.

Mapping hybrid MPI processes to hardware

Process/thread placement should match hardware:

- Threads should share (at least some portion of) cache
  - Explicit pinning to avoid thread migration
- Process should not cross socket boundary
  - Excessive cache coherency traffic
  - Possible memory access penalties on NUMA setups



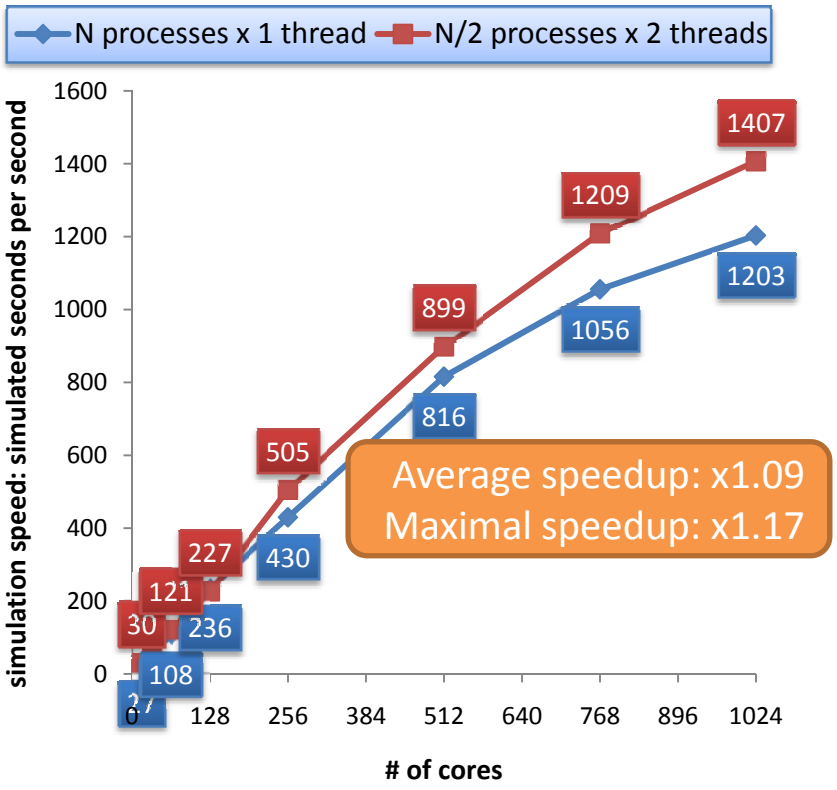
Process affinity setup used on DP node with Intel® Xeon® E54xx processors
(each MPI process runs 2 OpenMP threads; each thread is pinned to its own core)

Experiments were made with other pinning setups. This was found to be optimal.

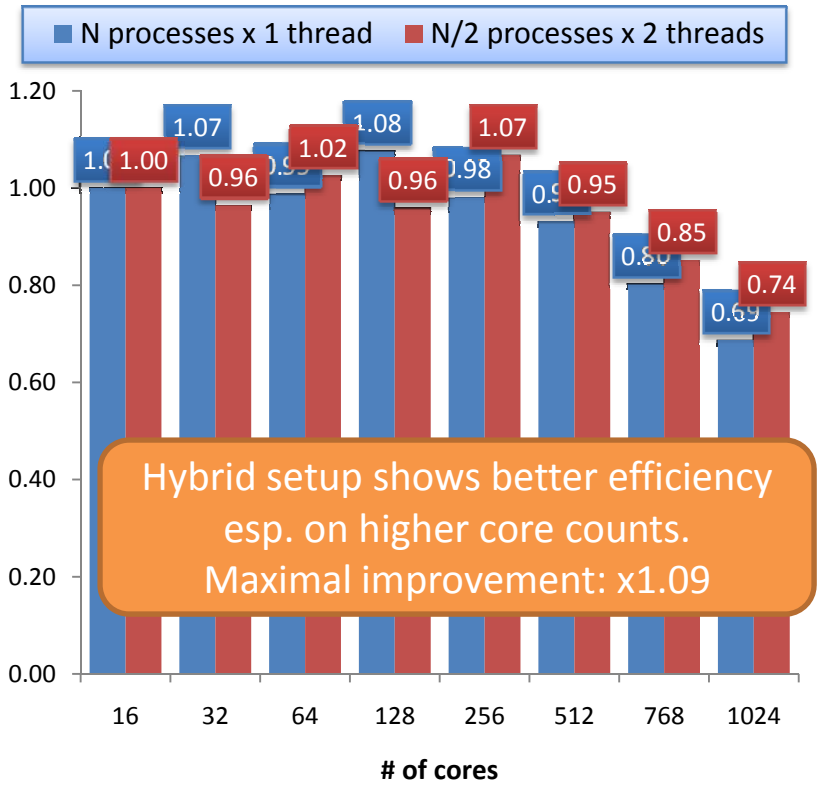
WRFV3/CONUS12KM

Performance & efficiency

Simulation speed

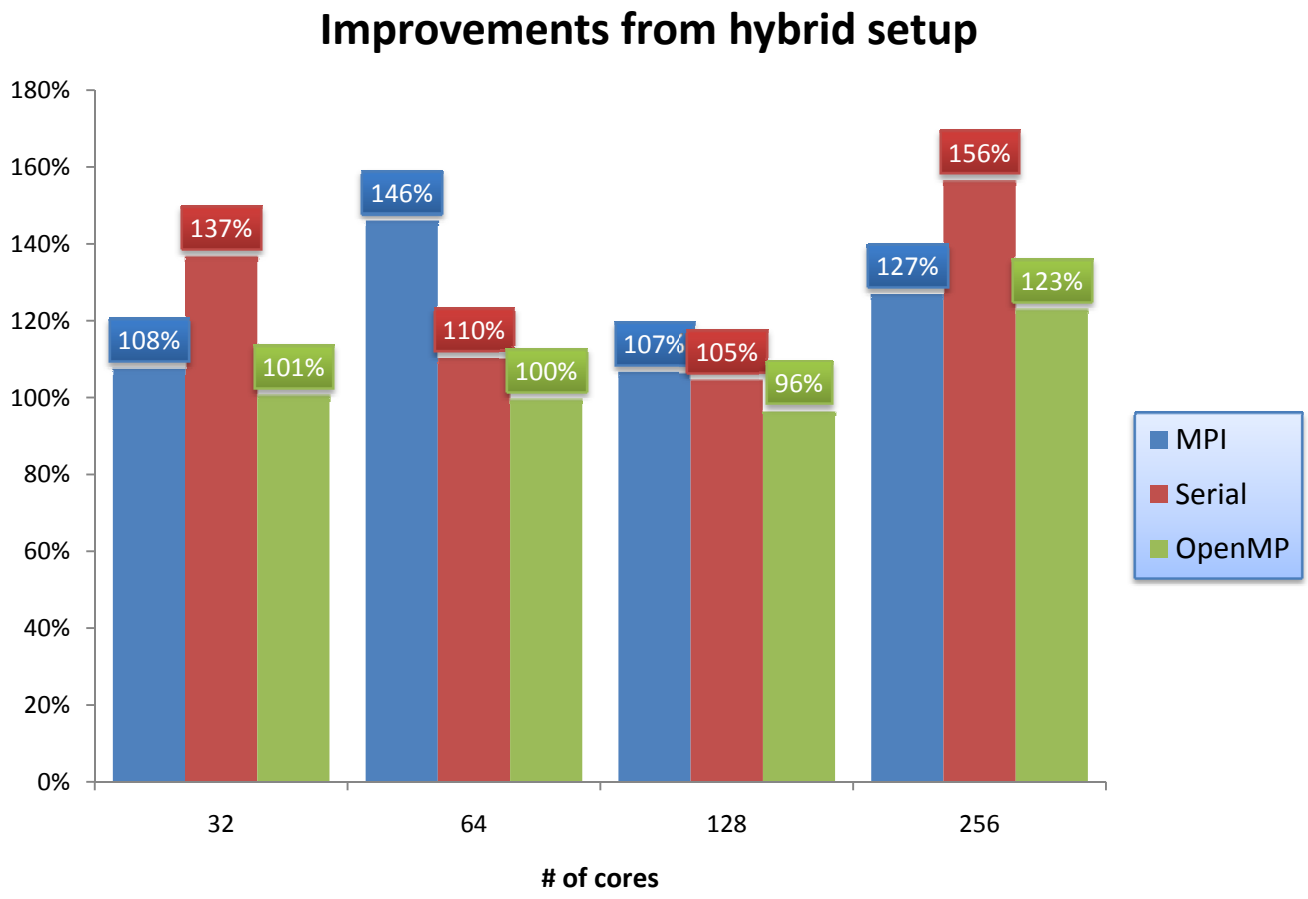


Parallel efficiency wrt 16 cores



- 3h simulation with 12km resolution over continental US
- Single domain, computations contain point-to-point communications only

Improvements breakdown

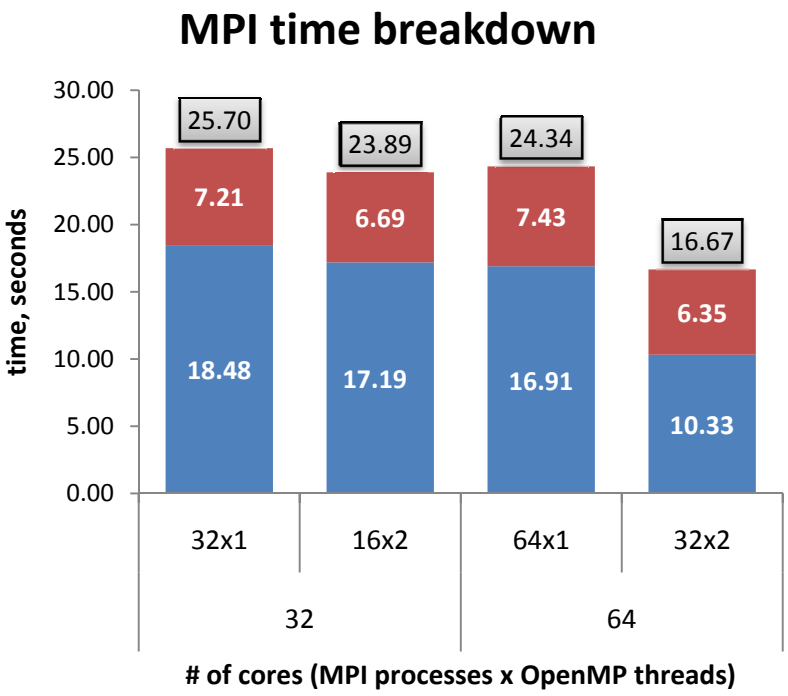


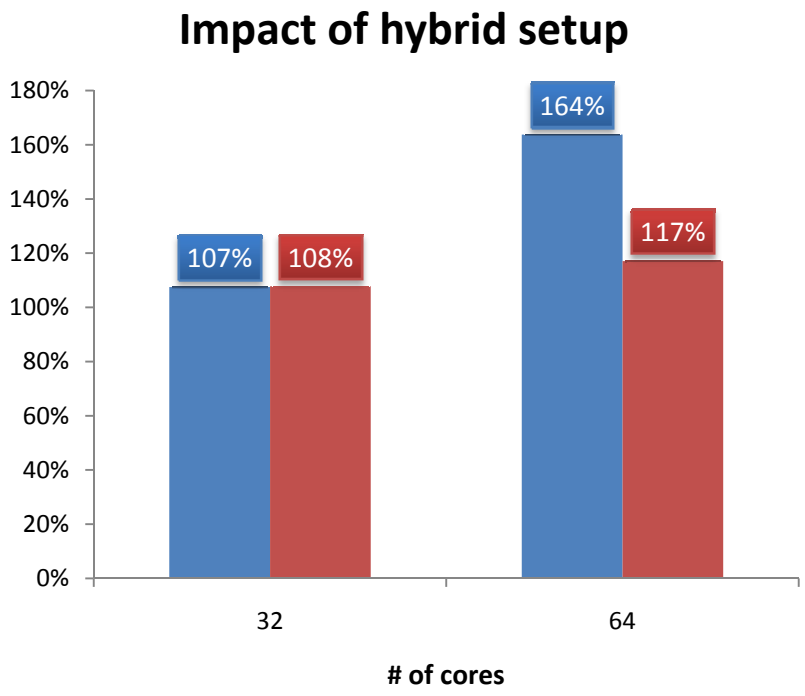
- I/O times excluded due to great variability in cluster setting
- Most serial time is due to I/O support overhead

Performance advantages of hybrid setup are:

- Better interconnect utilization
  - Lower volume of data during halo exchanges
  - Fewer processes involved in collective operations (IO support routines)
- OpenMP parallelization shows similar or better scalability than MPI
- Speedup in serial regions that are severely resources-limited

Computational imbalance





- Hybrid configurations show performance improvements
  - Lower data transfer times – less data transferred
  - Lower computational imbalance – less time spent waiting for other processes
- Measured using Ideal Connect Simulator – part of Intel® Trace Collector/Analyzer

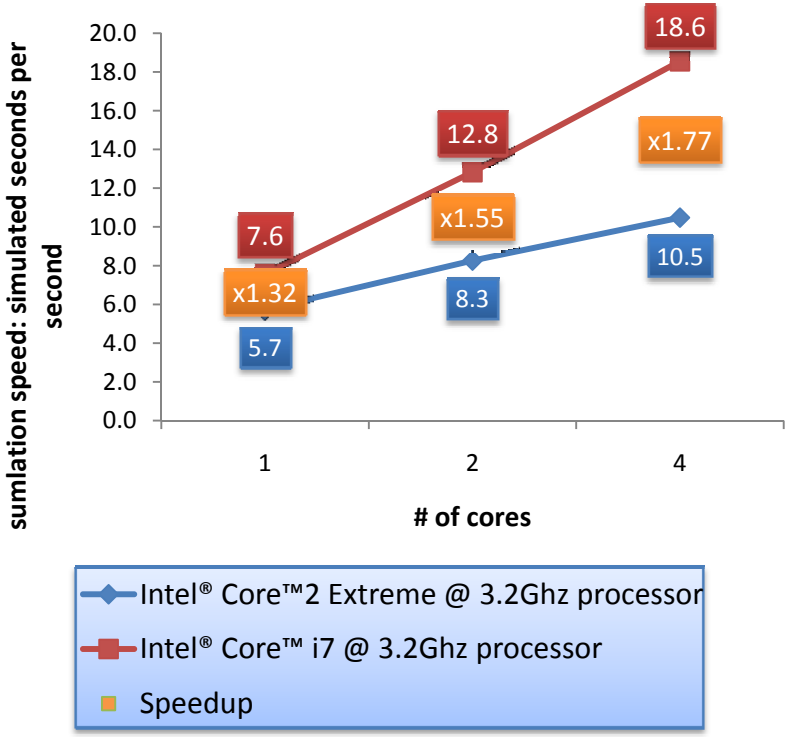
Performance on Intel® Core™ i7

Intel® Core™ i7 CPU

- Integrated 3-channel DDR3 memory controller.
- L3 cache shared amongst all 4 cores
- Microarchitecture enhancements (4-wide)
- SSE4.2
- QuickPath socket interconnect

Results presented were obtained on desktop (1-socket) machine

Comparison with Intel® Core™ Extreme (pure MPI)

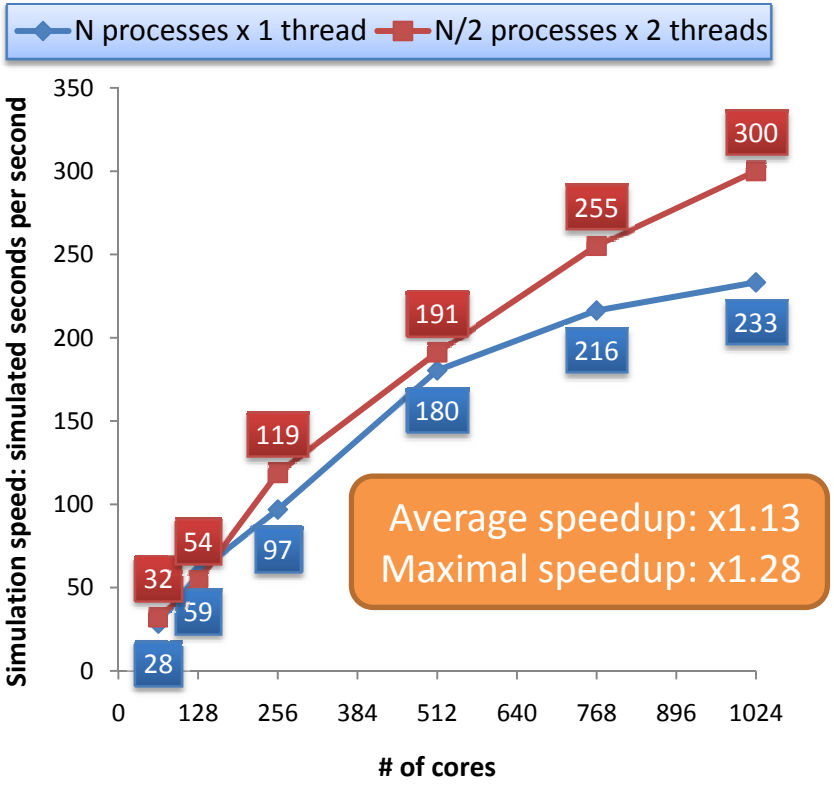


Performance improvement on 1 core is mostly due to better execution engine.
On 4 cores integrated memory controller pays off
Overall better scalability

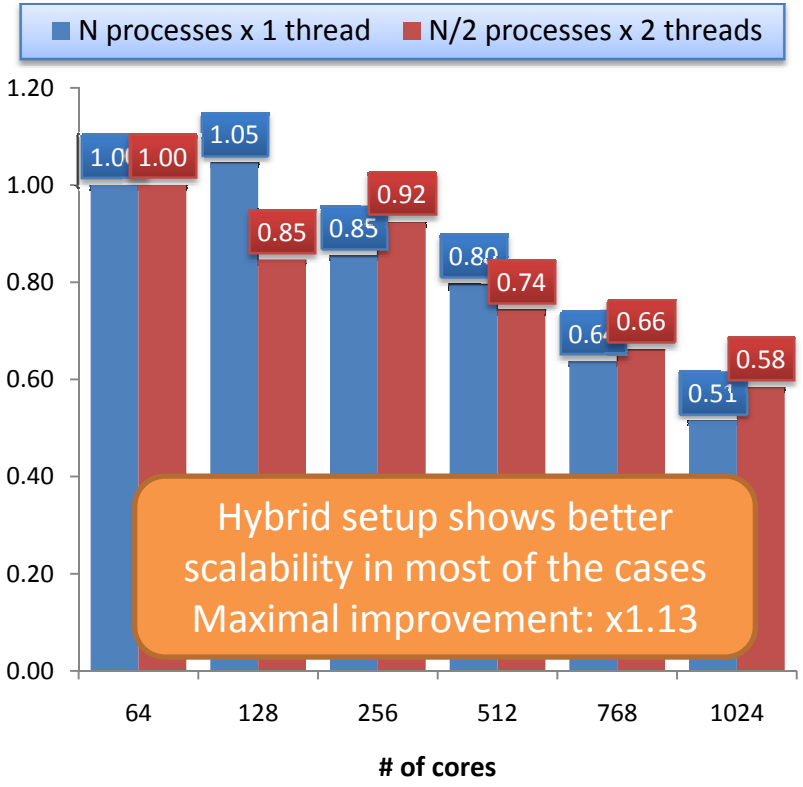
WRFV2/IVAN

Performance & efficiency

Simulation speed

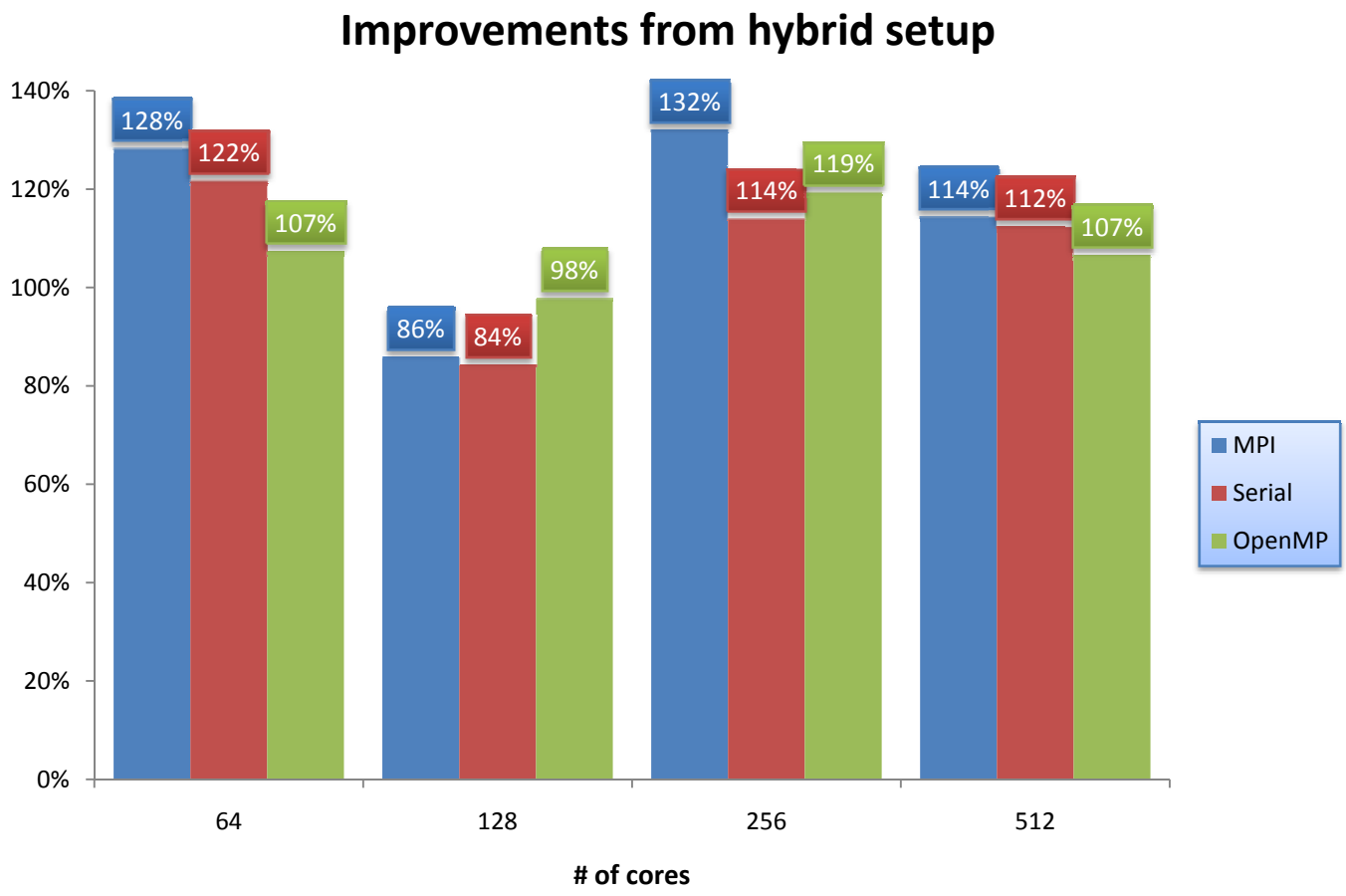


Parallel efficiency wrt 64 cores



- 12h simulation of hurricane Ivan (Sept., 2004)
- Nested domain, both point-to-point and collective communications

Improvements breakdown

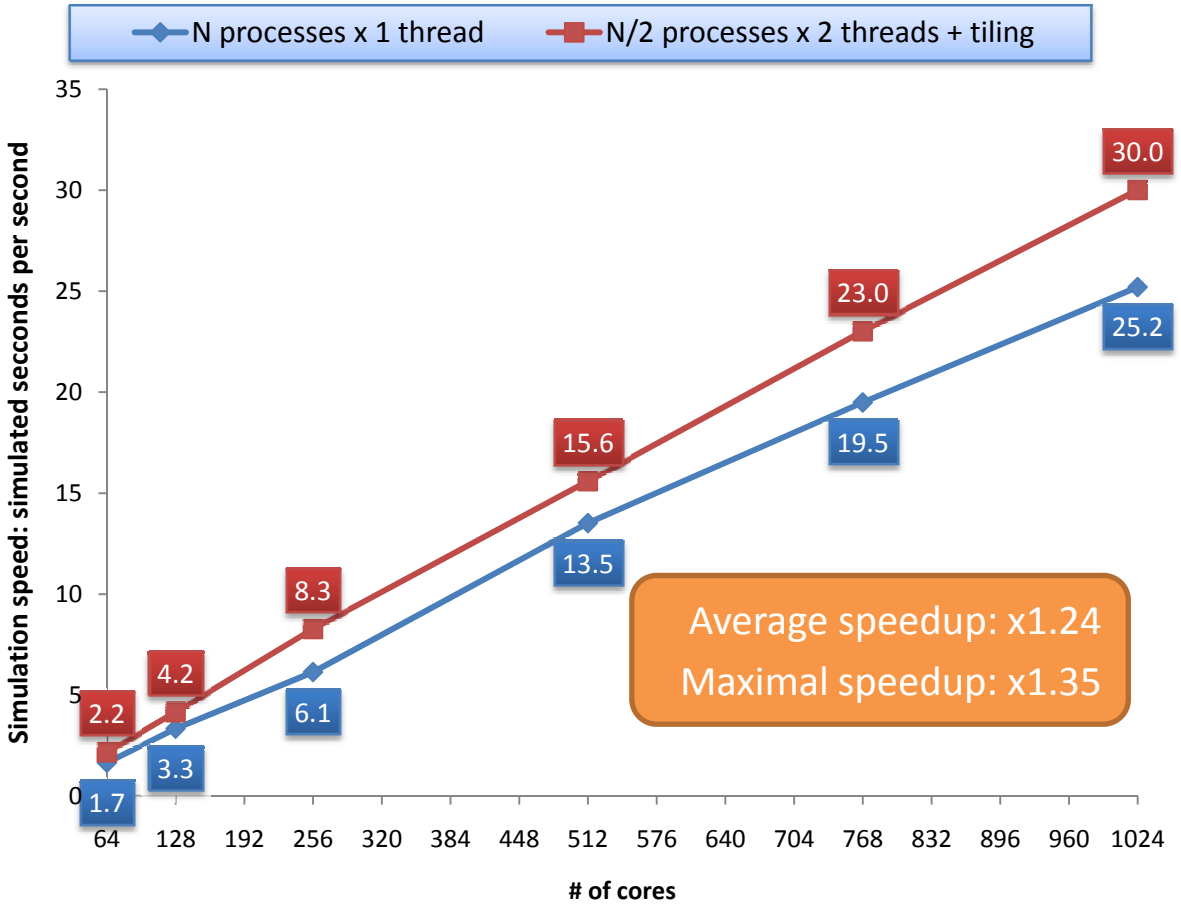


- IVAN is more MPI intensive than CONUS12km because passing data from/to nested domain involves collective communications that are performed each integration step.

- I/O times excluded due to great variability in cluster setting
- Most serial time is due to I/O support overhead

WRFV3/CONUS2.5KM

Performance

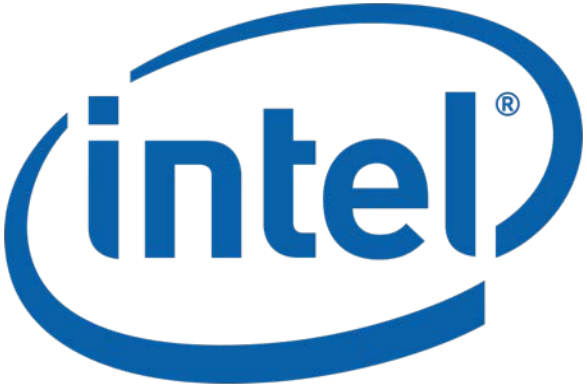


- Tiling improves cache utilization by dividing domain part owned by an MPI process into smaller pieces that are processed individually
- Tiling settings used here:
  - 128 tiles for 64 and 128 cores
  - 64 tiles for 256 cores
  - 32 tiles for remaining core counts

- 3h simulation with 2.5km resolution over continental US
- Single domain, point-to-point communications only

Summary

- Hybrid configurations show better performance in most of the cases
  - Similar or better scalability than MPI
  - Lower pressure on interconnect
    - Less data participates in halo exchange (point-to-point communications)
    - Fewer processes involved in collective communications
- Performance on Intel® Core™ i7 processor is more than 1.7 times better compared to previous generation of Intel CPUs
- Future work is to repeat these performance studies on cluster with Nehalem-EP server processors

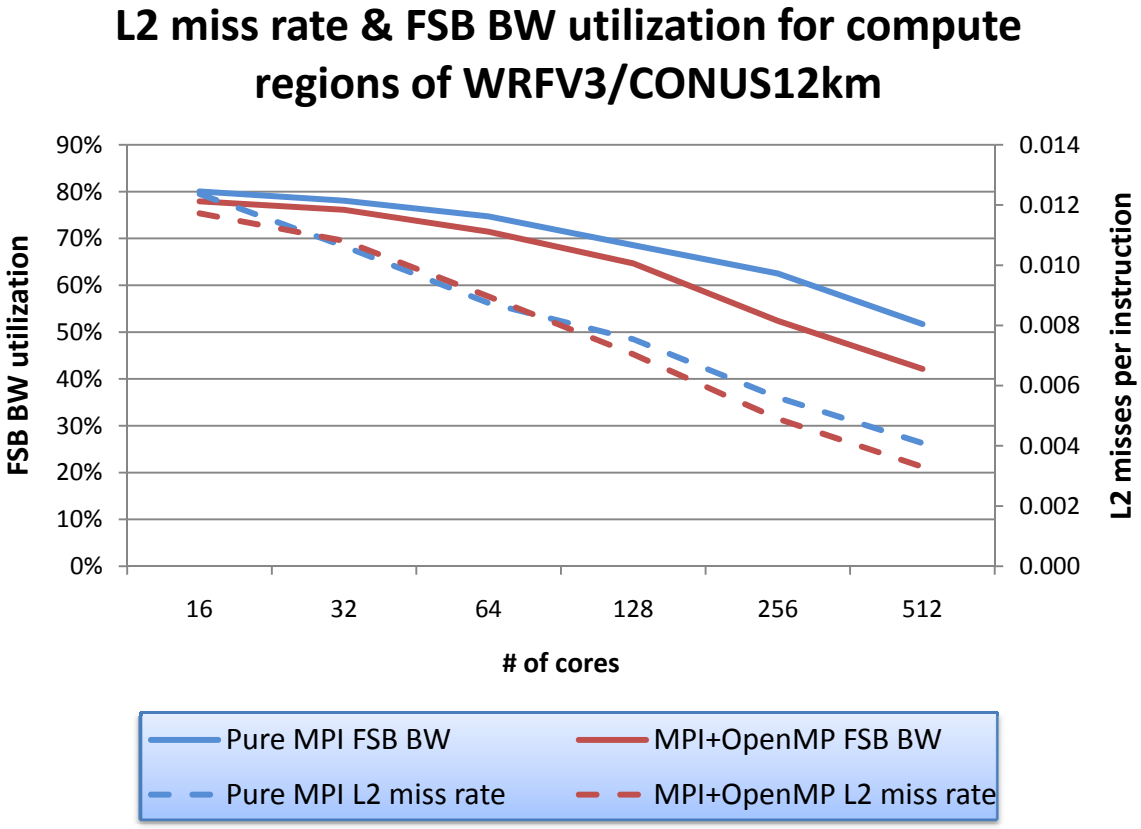

intel

BACKUP

Hardware setup

	CPU	Motherboard	OS	Interconnect
256-node Intel® Xeon® E54xx cluster	Intel® Xeon® E5462 processor, 4 cores, 2.8GHz, 2x6MB L2 cache	2 sockets, Intel® 5400 series chipset, 1600MT/s FSB, 16 GB FB-DIMM RAM	RHEL 4U4	DDR InfiniBand, fat tree topology, OFED 1.3, Mellanox ConnectX HCAs, Cisco Router
Intel® Core™ 2 Extreme Desktop	Intel® Core™ 2 Extreme processor, 4 cores, 3.2GHz, 2x6MB L2 cache	1 socket, Intel® X48 Express chipset, 1600MT/s FSB, 4GB DDR3 RAM	RHEL5U2	N/A
Intel® Core™ i7 Desktop	Intel® Core™ i7 processor, 4 cores, 3.2GHz, 8MB L3 cache	1 socket, Intel® X58 Express chipset, 6.4MT/s QPI, 3GB DDR RAM	RHEL5U2	N/A

HW resources utilization



- Hybrid setup shows similar or better L2 cache utilization than pure MPI in most cases
- Also, hybrid shows lower FSB utilization which means lower memory access latencies and better execution time

- Measurements were performed using Intel® Performance Tuning Utility
- Further improvements can be achieved by increasing number of tiles