intel Software

PARALLELIZING MPI USING TASKS FOR HYBRID PROGRAMMING MODELS

Presenter: Surabhi Jain

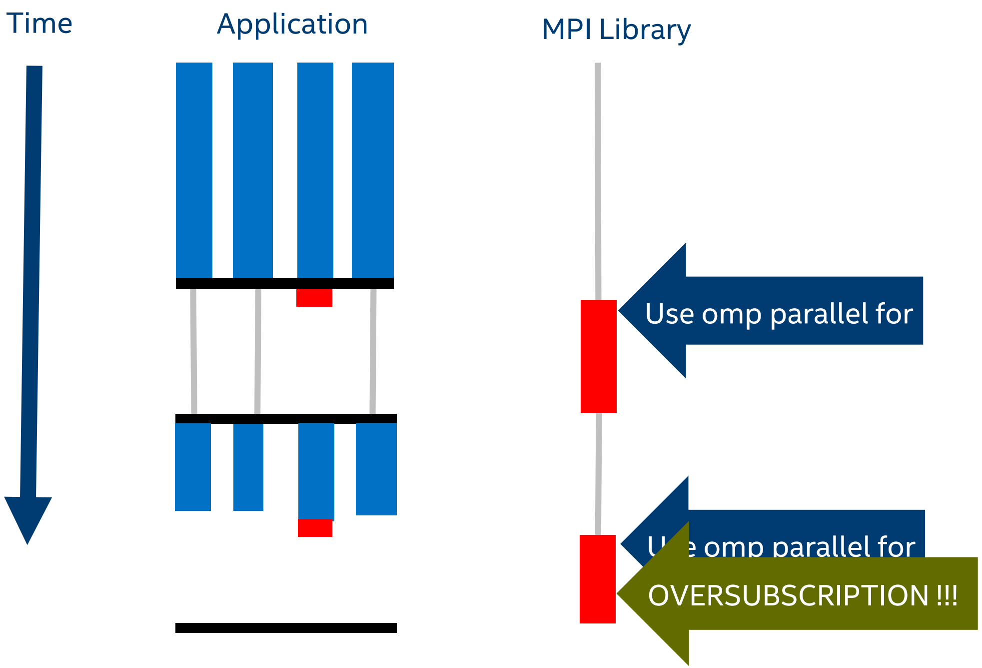
Contributors: Surabhi Jain, Gengbin Zheng, Maria Garzaran, Jim Cownie, Taru Doodi, and Terry L. Wilmarth

May 25, 2018
ROME workshop (in conjunction with IPDPS 2018), Vancouver, Canada

Notices

Acknowledgment: This material is based upon work supported by the U.S. Department of Energy and Argonne National Laboratory and its Leadership Computing Facility under Award Number(s) DE-AC02-06CH11357.

Example: Hybrid (MPI+OpenMP*) Application



- Machine supports 4 threads (including hyperthreading)
- 1 MPI rank, 4 threads per rank

	Computation
	MPI Call
	Barrier
	Idle

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

Motivation and Goal

Motivation:

Applications run using hybrid programming models (MPI+X)

- X=OpenMP, TBB, Pthreads
- Application threads run the computation code in parallel
- Usually only one thread calls the MPI library

Goal:

To have an MPI library that runs using multiple threads which do not compete with the application threads (avoid oversubscription)

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

Approaches (MPI+X, X = OpenMP)

Thread Partitioning

- Partition the threads. Dedicate n threads to MPI and rest to OpenMP

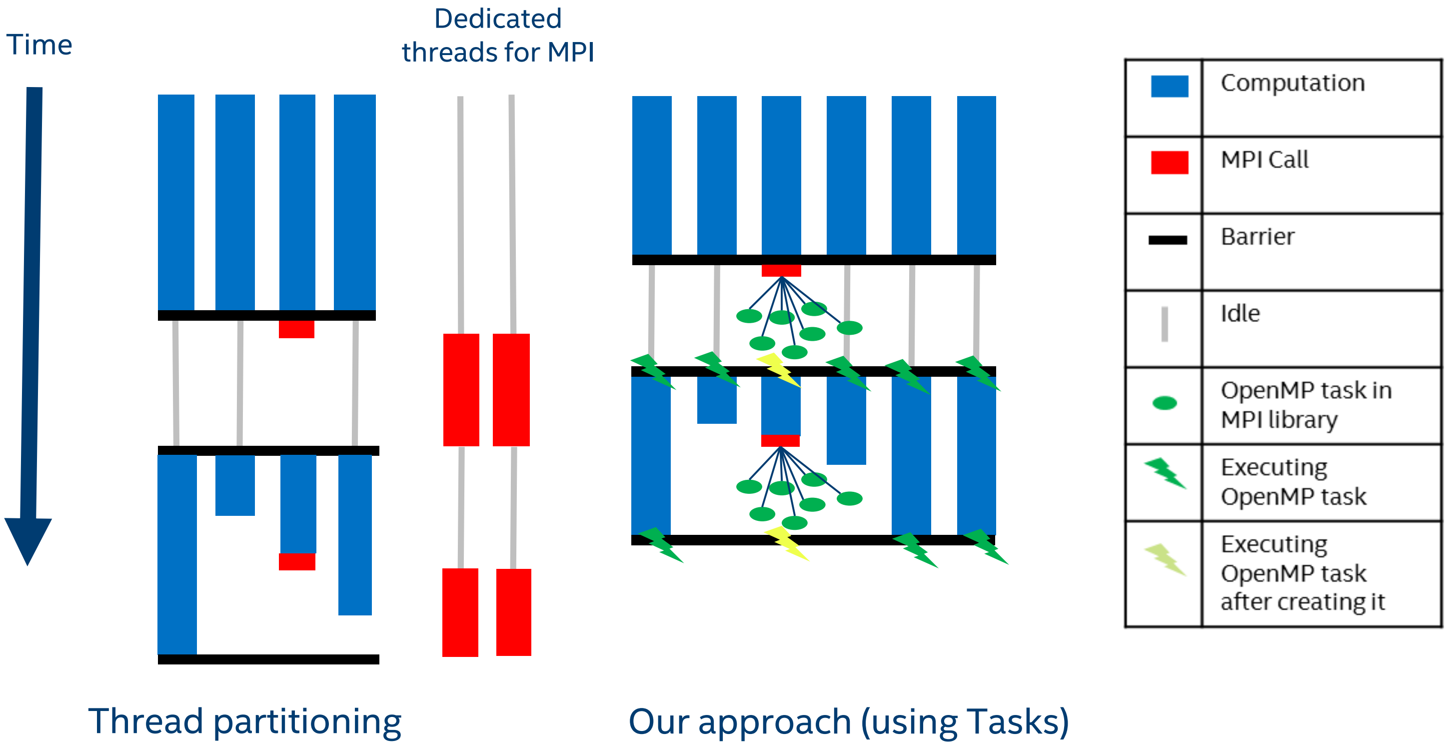
Modify OpenMP library

- OpenMP tells MPI the number of idle threads
- Spawn “number of idle threads” threads in MPI
- MT-MPI: Multithreaded MPI for many-core Environments, ICS’14 by Si et al.

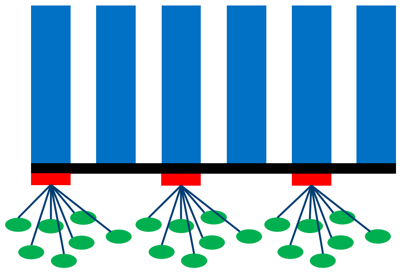
Create tasks in MPI (Our approach)

- Tasks can be executed by idle application threads
- Does not spawn additional threads in MPI, no oversubscription
- No modifications required in OpenMP
- Maps well to any library which supports a tasking model e.g. OpenMP, TBB

Ways to share threads between MPI and OpenMP



Our approach orthogonal to MPI_THREAD_MULTIPLE



Computation (blue)	Computation
MPI Call (red)	MPI Call
Barrier (black)	Barrier
OpenMP task (green)	OpenMP task in MPI library

Do not expect much benefit with MPI_THREAD_MULTIPLE

- Parallelism comes from several threads concurrently calling MPI
- Fewer threads are idle to execute MPI tasks

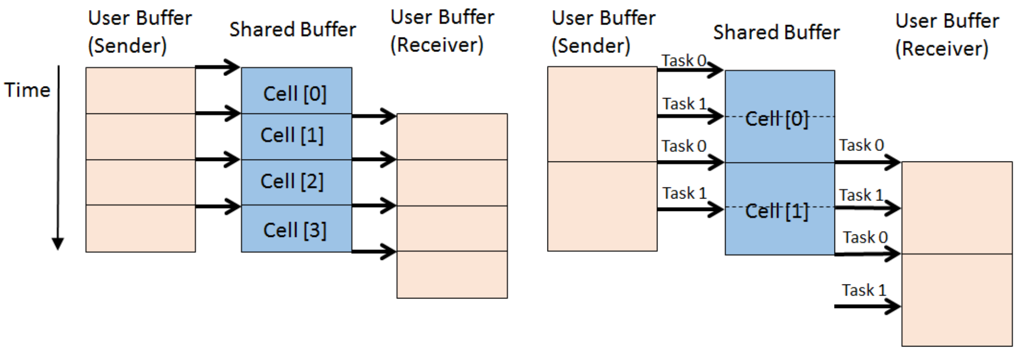
Creating OpenMP tasks in MPI

```
if(omp_in_parallel()) {
     //Create tasks for what MPI wants to do in parallel
     //which will run on idle pre-existing OpenMP threads
     #pragma omp taskwait
     /* All tasks we created have completed when we get here */
} else {
     /* No pre-existing parallelism so create some */
     #pragma omp parallel
     {
          #pragma omp single nowait
          {
               //Create tasks for what MPI wants to do in parallel
          }
     }
     /* All tasks we created have completed when we get here */
}
```

Where to create tasks inside MPI?

- Shared memory communication
- Packing/unpacking of non-contiguous data

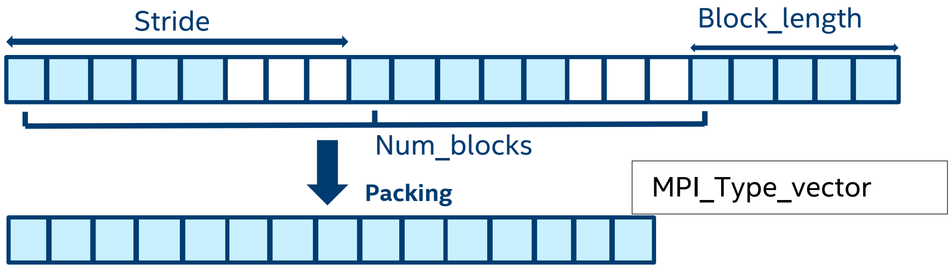
Shared Memory Communication

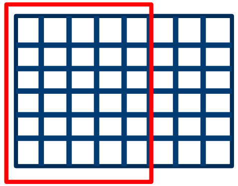


(a) Sequential Data Copying (b) Parallel Data Copying

- Sender and receiver rank on the same node, use intermediate shared buffer for large messages
- Pipelined Double Copy approach- sender can copy to next cell(s), while receiver is copying from previous cell(s)
- Find balance between pipeline parallelism and task based parallelism

Pack/Unpack non-contiguous data





Stride = 8
Block_length = 5
Num_blocks = 6

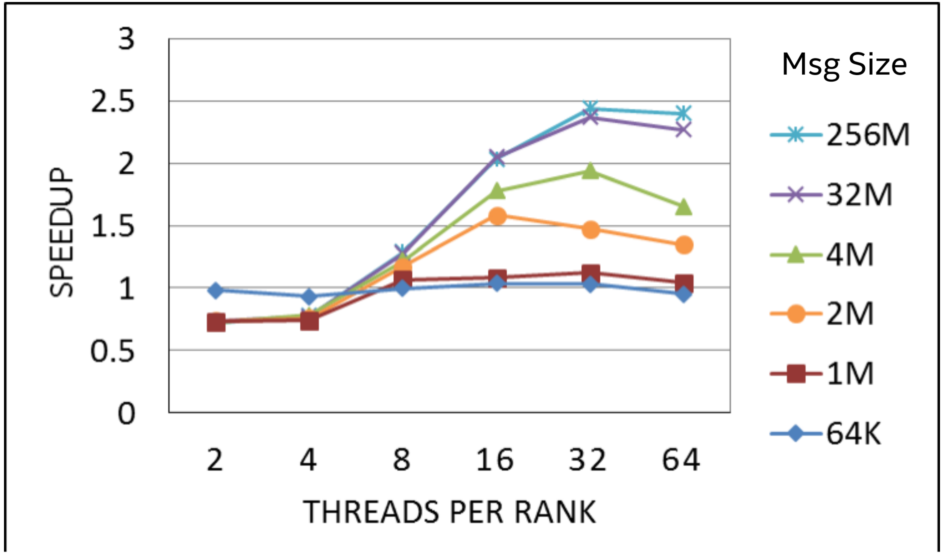
Derived types

- Constructed from existing types (basic and derived). E.g. MPI_Type_indexed, MPI_Type_vector, MPI_Type_struct
- Each task can pack/unpack one or more blocks

Experimental Setup

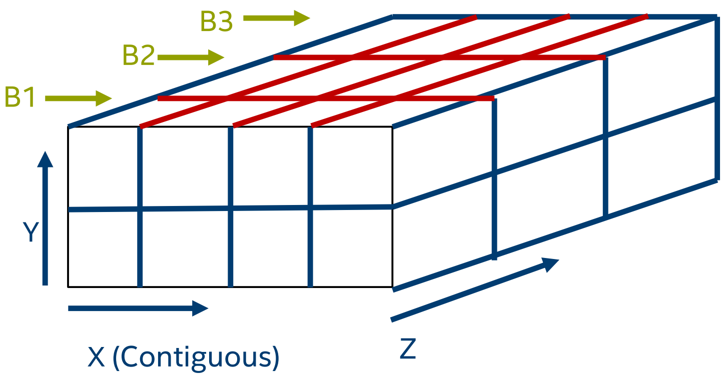
- Intel® Xeon Phi Processor 7210 (1.3 GHz, 64 cores, 4 threads/core) (Knights Landing)
- 32KB L1 data and instruction cache, 1MB L2 cache
- 96GB DDR, 16GB MCDRAM
- KNL memory mode – Flat, cluster mode – Quadrant, No SNC
- Data placed on MCDRAM (using numactl –m 1)
- Compiler from Intel® Parallel Studio XE Composer Edition for C++ (version 2016.0.109)
- MPICH v3.2b4-98-g4551de1 as the baseline

Parallel memcpy – OSU latency benchmarks



	MPICH (Original)	MPICH (Modified) [Baseline]	MPICH (Modified) with OpenMP tasks
Task Size			32KB
Cell Size	32KB	256KB	#threads * Task Size
Total Size	256KB	4MB	4MB
Num Cells	8	16	Total Size/Cell Size

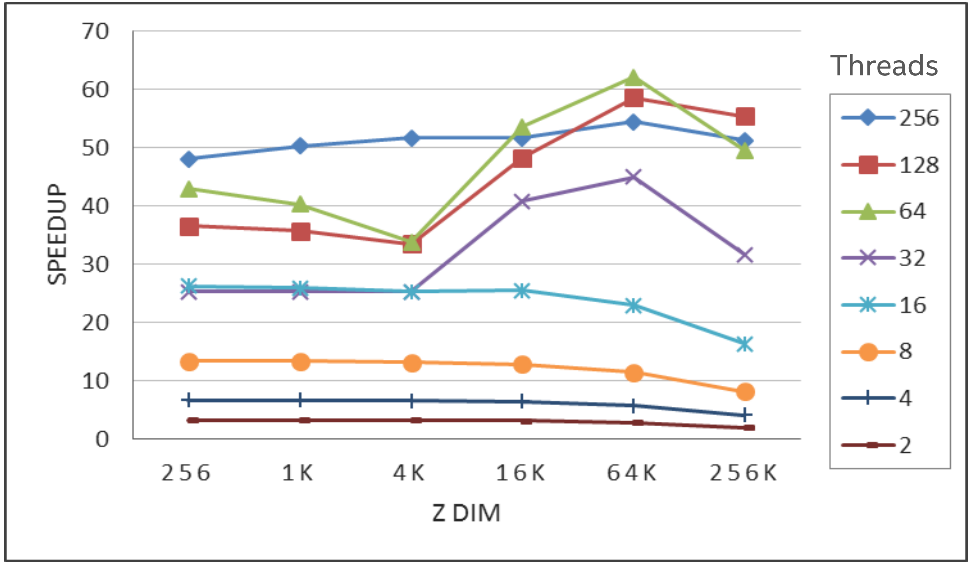
Top Pack Benchmark (from MT-MPI* paper)



- Pack the top surface (XZ plane) of a 3D matrix of doubles
- Matrix volume fixed to 1 GB and Y dimension to 2
- Represented using MPI_Type_vector

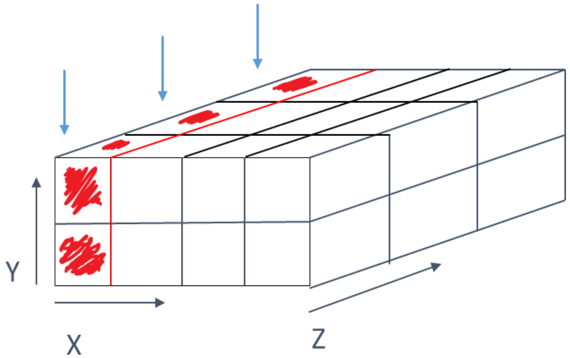
*Si et al . MT-MPI: Multithreaded MPI for many-core Environments . ICS'14

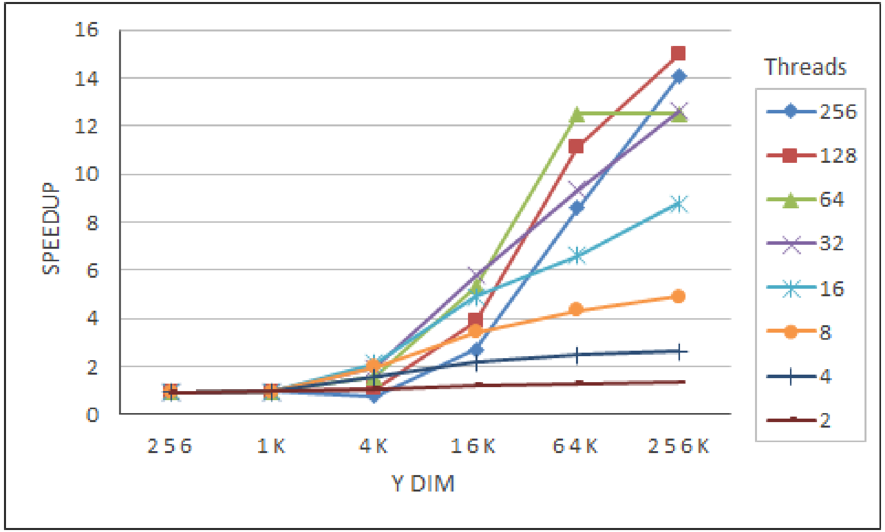
Results: Top Pack, MPI_Type_vector



- Packing called from a serial region in application
- 1 MPI rank
- Blk_size(X) decreases as num_blks(Z) increases

Left Pack, Nested MPI_Type_vector





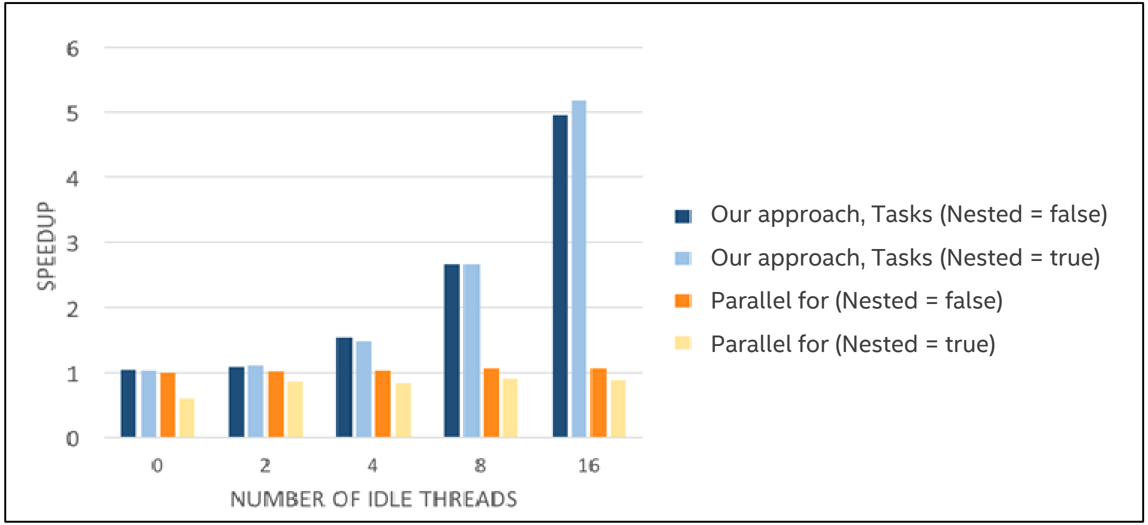
- Tasks at leaf level – Parallelize over Y dimension (vector datatype)
- Tasks at higher level – Parallelize over Z dimension (vector of vectors datatype)

MPI_Pack() called from a parallel region

```
#pragma omp parallel
{
      thread_id = omp_get_thread_num();
      if (thread_id < 4)
            call MPI_Pack();
      else if (thread_id < 4 + num_idle_threads)
            do_nothing
      else
            do_computation
}
```

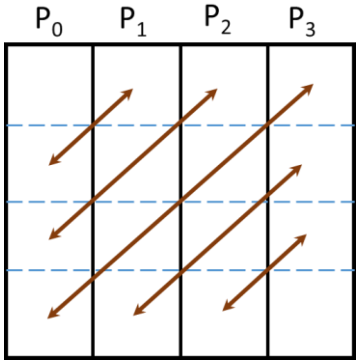
- Uses MPI_THREAD_MULTIPLE mode
- Threads are divided into 3 groups -
  - Threads calling MPI_Pack(). Create OpenMP task in MPICH
  - Idle threads. Wait at the barrier and execute tasks
  - Compute threads. Can help in executing tasks when reach the barrier

Results: MPI_Pack() called from a parallel region



- Total threads = 256, Packing threads = 4
- Significant benefits when threads are idle
- No penalty when no idle threads

Transpose from Parallel Research Kernels*

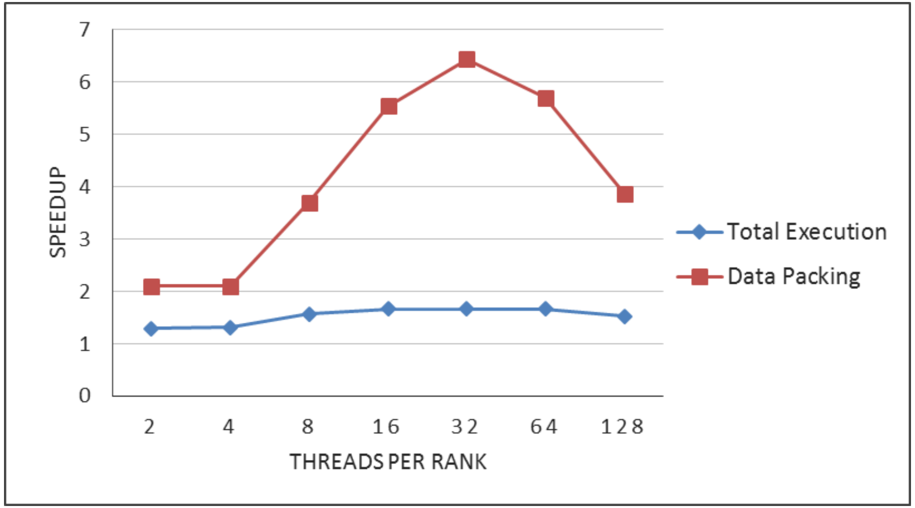


Steps to do the transpose

1. On each rank, local transpose. No data communicated
2. All-to-all communication
   1. Use nested MPI_Type_vector datatype
   2. Parallelize the pack/unpack

* https://github.com/ParRes/Kernels

Results : Transpose Kernel



- Matrix Order 8K doubles
- 2 MPI ranks on 1 KNL node
- Leaf vector – num_blks= 4K, blk_len=1

Conclusion

- Our task-based approach-
  - Opportunistic
  - No creation of additional threads, so no oversubscription
  - No modification made in the OpenMP library
- Speedup up to 62X on Top Pack, when all the threads are idle
- Speedup up to 6.5X in data packing and up to 1.5X reduction in overall execution time of transpose kernel
- Code is publicly available (link in the paper)

Questions ?

surabhi.jain@intel.com

https://www.linkedin.com/in/surabhi-jain-9066821a/

Legal Disclaimer & Optimization Notice

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION 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

Software