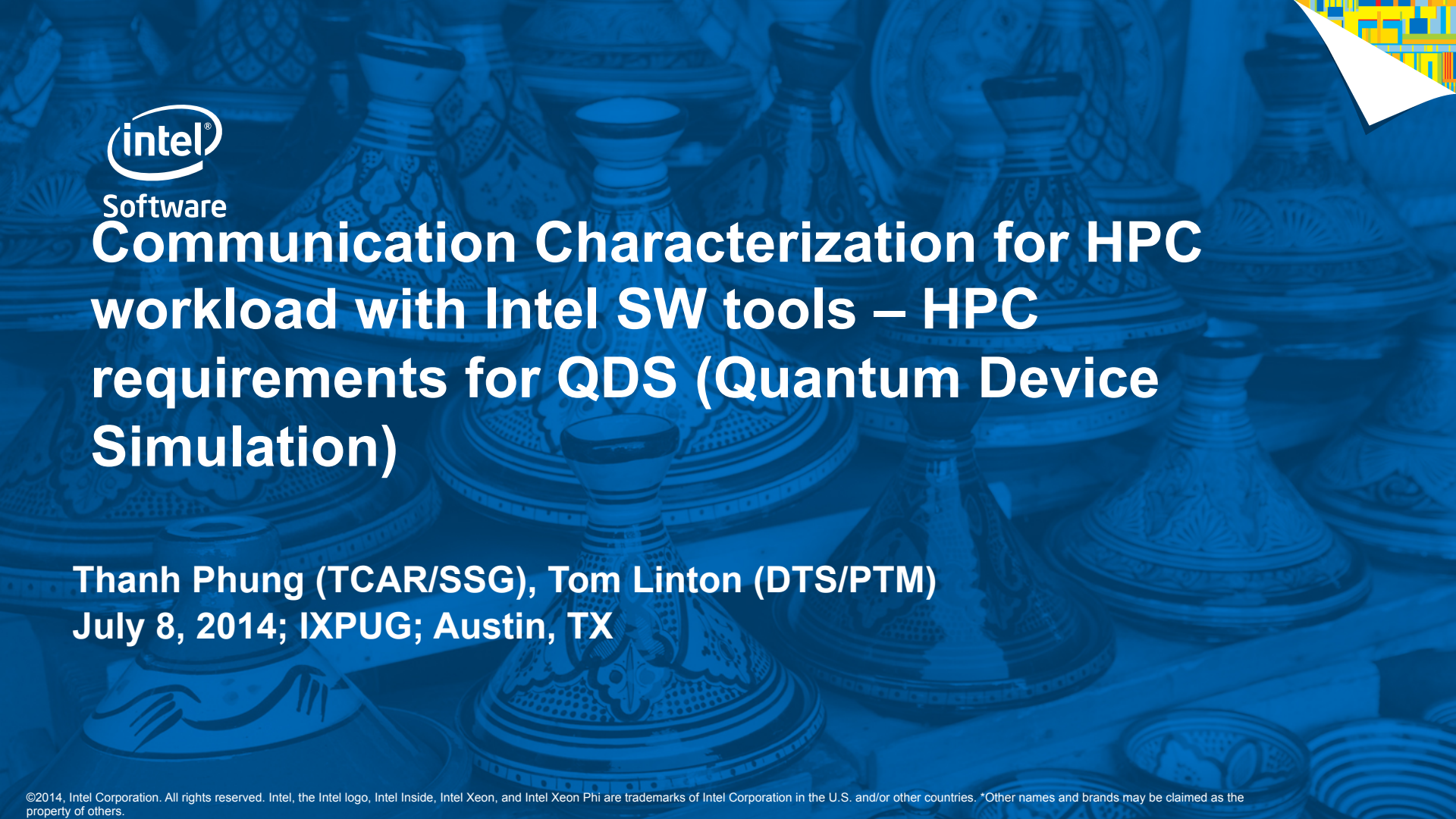


Communication Characterization for HPC workload with Intel SW tools – HPC requirements for QDS (Quantum Device Simulation)

Thanh Phung (TCAR/SSG), Tom Linton (DTS/PTM)

July 8, 2014; IXPUG; Austin, TX

Outline:

I. QDS General Introduction

1. Target application
2. Goal

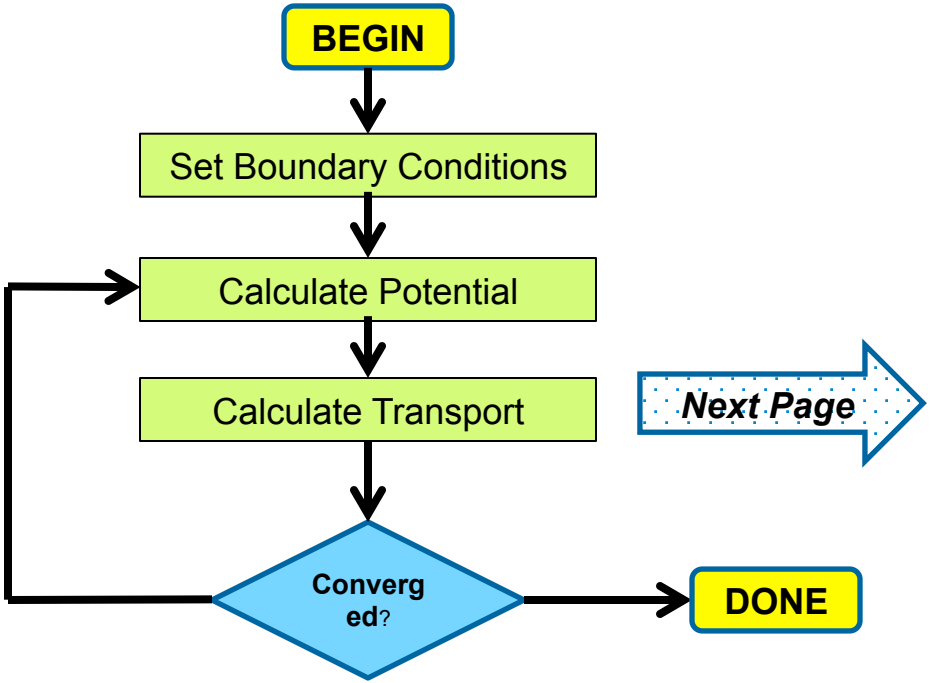
II. Projected Requirements

III. Understand QDS: FP and data parallel

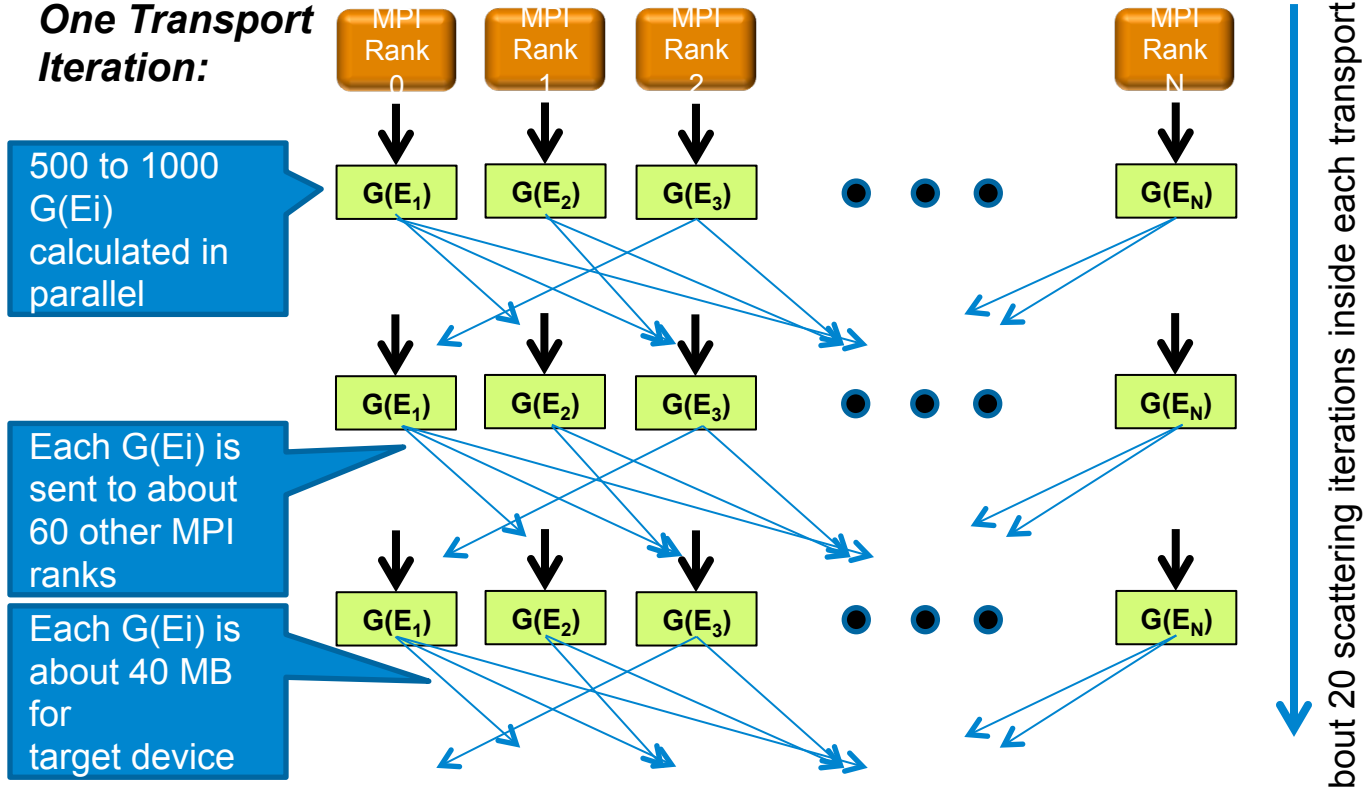
1. Capture details of communication profile with ITAC
2. Intel MPI 5.0 (with new MPI 3.0)

I. QDS Introduction:

1. Target application is NEMO-5/Omen
   - provides quantum device simulation
   - key new capability is modeling of scattering
2. Goal (Device Modeling Grand Challenge)
   - New ability to simulate 5 nm channel transistor with quantum transport + scattering
     - Present simulations are adequate using 3nm and with quantum transport (ballistic) and no enhanced scattering model
   - turnaround time of 1 week for an IV curve
   - Implement by Q1 2015



One Transport Iteration:

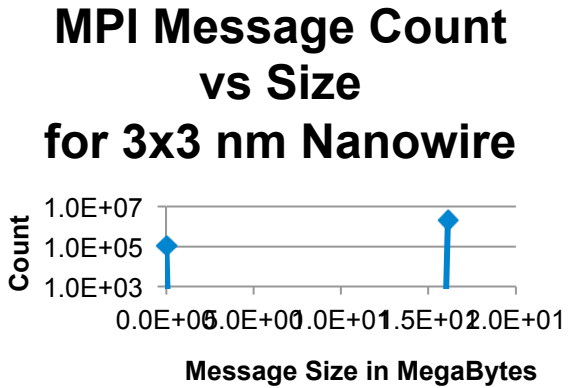


II. Projected Requirements

1. Require as many cores as can be provided –
   - Target for 32K IVB/HSW core, Memory = 9GB/core
   - Memory requirement per core is a challenge for Xeon Phi
     - It is possible to split computation across multiple cores to reduce memory
2. Message passing
   - dominated by regular pattern of large messages
   - for target problem, message size = 40 MB
   - each MPI rank sends the same 40 MB to about 60 other ranks for each energy calculation
   - message size increases with physical problem size
   - message count is constant with physical problem size
3. I/O is negligible
4. *Reliability: must run 32K cores for a week without any crashes*

Message pattern summary

- message pattern is very regular
- each MPI rank typically sends an array to about 60 other ranks
- array size depends linearly on physical problem size (number of atoms)
  - e.g., small 3x3x20 wire: 16 MB arrays
  - for target 5x5x20 wire: 40 MB arrays



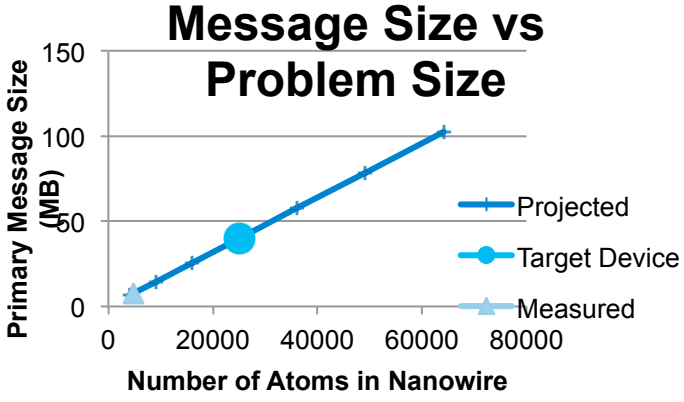
Message size scales linearly with physical problem size

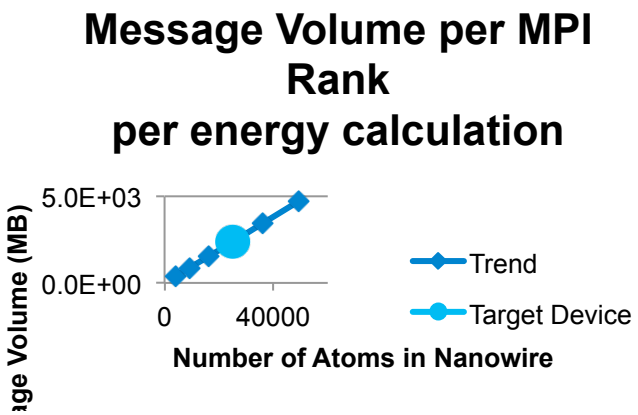
- target device: message size = 40 MB

If 60 messages are in flight for a single MPI rank:

- target device total volume = 2.4 GB

Multiple cores per energy can be used to scale message size down





The *number* of messages sent is independent of physical device size

Fundamental computation is for one energy

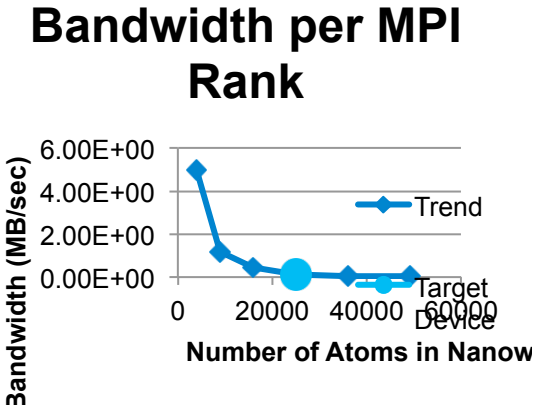
- typically 500 energies are calculated in parallel on each inner iteration
- each energy can be computed on one more cores or a core can calculate multiple energies
- each energy depends on messages from typically 60 other energy calculations

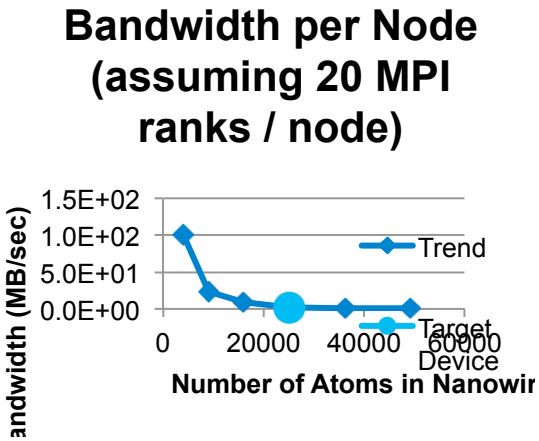
Message bandwidth per MPI rank will be:

- bandwidth = 60 * (message size) / (message time)

CPU time grows faster than message size

- simplistic bandwidth: assume message time = cpu time





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