LARGE SCALE ELECTRONIC STRUCTURE CALCULATIONS ON THETA

Performance optimization of WEST and Qbox

Huihuo Zheng¹, Christopher Knight¹, Giulia Galli¹,², Marco Govoni¹,², and Francois Gygi³

¹Argonne National Laboratory
²University of Chicago
³University of California, Davis

Feb 28th, 2018
THETA ESP: FIRST-PRINCIPLES SIMULATIONS OF FUNCTIONAL MATERIALS FOR ENERGY CONVERSION

Embedded nanocrystal
T. Li, Phys. Rev. Lett. 107, 206805 (2011)

Heterogeneous interfaces
H. Zheng, APS March meeting, 2018

Organic photovoltaics

Aqueous solution

Quantum information

PERFORMANCE OPTIMIZATION

- Utilizing tuned math libraries (FFTW, MKL, ELPA, ...)
- Vectorization: AVX512
- High Bandwidth Memory

Adding extra layers of parallelization -> increase intrinsic scaling limit
Reducing communication overhead to reach the intrinsic limit
OUTLINE

• WEST – adding extra layers of parallelism
  • Addressing bottleneck from I/O
  • Implementing band parallelization

• Qbox – reducing communication overheads of dense linear algebra with on-the-fly data redistribution
  • Gather & scatter remap
  • Transpose remap

• Conclusions and insights
OPTOELECTRONIC CALCULATIONS USING MANY-BODY PERTURBATION THEORY (GW)

Linear response theory

\[ \Delta \rho = \chi \Delta V_{\text{pert}} \]

Massively parallel by distributing perturbations

Parallelization scheme (image & plane wave)

3D FFTs + \( D(Z) \text{GEMM} \)

Matrix diagonalization (syev, heev, elpa)
SINGLE NODE RUNTIME ON THETA IN COMPARISION WITH MIRA (1KNL VS 4BG/Q)

- 80% of runtime is spent in external libraries
- 3.7x speedup from BG/Q(ESSL) to KNL(MKL)
- High-bandwidth memory on Theta critical for performance (e.g. 3D FFTs): 3.1x speedup
I/O ISSUE APPEARED IN WEAK SCALING STUDY

- Original I/O scheme: all replica read the same file; I/O time increased with number of nodes becoming a significant fraction of runtime.
- Time spent in I/O reduced to negligible fraction of runtime on 1-1024 nodes by having master process read and distribute wave function.
IMPROVEMENT OF STRONG SCALING BY BAND PARALLELIZATION – A PATHWAY TO A21

Si$_{35}$H$_{36}$, 176 electrons, 256 perturbations

Increased parallelism by arranging the MPI ranks in a 3D grid (perturbations & bands & FFT)

New intrinsic strong scaling limit:

\[ n_{proc} = N_{pert} \times N_{band} \times N_z \]
QBOX

SCALING HYBRID DENSITY FUNCTIONAL CALCULATIONS
STRONG SCALING ANALYSIS OF QBOX FOR HYBRID-DFT CALCULATIONS

DATA LAYOUT: BLOCK DISTRIBUTION OF WAVE FUNCTIONS TO 2D PROCESS GRID

\[ \psi_i(k), i = 1, 2, \ldots N_{\text{band}}, \]
\[ k = 0, 1, \ldots, n_{\text{pw}} - 1 \]

SiC512: 140,288 × 1,024

\( n_{\text{col}} \)

\( n_{\text{row}} \)

1,024

140,288

MPI_Alltoall(v)

MPI_Allreduce

Good scaling for 3D FFTs up to intrinsic limit:

\[ n_{\text{proc}} = N_{\text{band}}N_{\text{z}} \]
DENSE LINEAR ALGEBRA INVOLVED FOR TALL-SKINNY MATRICES AND SMALL SQUARE MATRICES

Gram-Schmidt

Wave function matrix

Overlap matrix

Tall-skinny matrices

Small square matrix

d(z)gemm
INCREASING OF COMMUNICATION OVERHEAD FROM SCALAPACK SUBROUTINES
REDUCING COMMUNICATION OVERLAP BY ON-THE-FLY REDISTRIBUTING DATA WITH REMAP METHOD

Increasing npcol →
- local computing time decreases,
- communication time increases → Performance degradation

Solution: let a smaller group of processors do ScaLAPACK
- Do FFT on the original grid
- Gather data to the smaller grid
- Do ScaLAPACK on the smaller grid
- Scatter data back to original grid

Remapping time (gather + scatter) should be small.
**IMPROVEMENT OF STRONG SCALING USING “GATHER & SCATTER” REMAP**

$hpsi + wf\_update$ time remains minimal relatively flat with remap, and the remap time (custom) is two orders of magnitude smaller than $hpsi + wf\_update$ time.

Custom remap function is 1000x faster than ScaLAPACK’s pdgemr2d.

Improvement of Qbox’s strong scaling after optimizations; runtime of improves from ~400 to ~30s per SCF iteration (13x speedup) on 131,072 ranks for 2048 electrons.
FURTHER IMPROVEMENT OF DGEMM RUNTIME BY “TRANSPOSE” REMAP

Problem of “gather & scatter”: Idle processes. How to utilize them? Assign idle processes to active columns.

Transpose remap:
• Perform 3D FFTs in the original context.
• Transfer data through a series of local regional transposes
• Run ScaLAPACK in the new context

Key concept for remap: creating different contexts that are optimal for different kernels redistributing the data on-the-fly

Transpose communication pattern

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>11</td>
<td>15</td>
</tr>
</tbody>
</table>

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Process rearrangement and data movement of transpose remap

Improvement of runtime by remap methods
(1) \( ncol' = \frac{ncol}{8}, nrow' = nrow \)
(2) \( ncol' = \frac{ncol}{8}, nrow' = 8 \times nrow \)
CONCLUSION AND INSIGHTS

• Band parallelization reduces the internode communication overhead and improves strong scaling of WEST up to $N_{FFT} N_{pert} N_{band}$ cores.

• Optimal remapping of data for matrix operations in Qbox reduces ScaLAPACK communication overhead at large scale, and makes hybrid- DFT calculation scale to $N_{FFT} N_{band}$ cores.

• Given the increased computational performance relative to network bandwidths, it is crucial to reduce and/or hide inter-node communication costs.

Guiding principles for developing codes in many-core architecture:

1) Fixing non-scalable bottleneck (e.g., Parallel I/O)

2) Parallelizing independent, fine-grain units of work, reducing inter-node communication, and maximizing utilization of on-node resources.

3) Optimizing data layout: optimizing communication patterns for performance critical kernels with on-the-fly data redistribution and process reconfiguration.
ACKNOWLEDGEMENT

• This research is part of Theta Early Science Project at Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility under Contract DE-AC02-06CH11357.

• This work was supported by MICCoM, as part of Comp. Mats. Sci. Program funded by the U.S. DOE, Office of Science, BES, MSE Division.
THANK YOU!