Automatic Tuning for Parallel FFTs on Cluster of Intel Xeon Phi Processors

Daisuke Takahashi
Center for Computational Sciences
University of Tsukuba, Japan

Outline

- Background
- Objectives
- Six-Step FFT Algorithm
- In-Cache FFT Algorithm and Vectorization
- Computation-Communication Overlap
- Automatic Tuning of Parallel 1-D FFT
- Performance Results
- Conclusion

Background

- The fast Fourier transform (FFT) is widely used in science and engineering.
- Parallel FFTs on distributed-memory parallel computers require intensive all-to-all communication, which affects their performance.
- How to overlap the computation and the all-to-all communication is an issue that needs to be addressed for parallel FFTs.
- Moreover, we need to select the optimal parameters according to the computational environment and the problem size.

Objectives

- Several FFT libraries with automatic tuning have been proposed.
  - FFTW, SPIRAL, and UHFFT
- An Implementation of parallel 1-D FFT on cluster of Intel Xeon Phi coprocessors has been presented [Park et al. 2013].
- However, to the best of our knowledge, parallel 1-D FFT with automatic tuning on cluster of Intel Xeon Phi processors has not yet been reported.
- We propose an implementation of a parallel 1-D FFT with automatic tuning on cluster of Intel Xeon Phi processors.

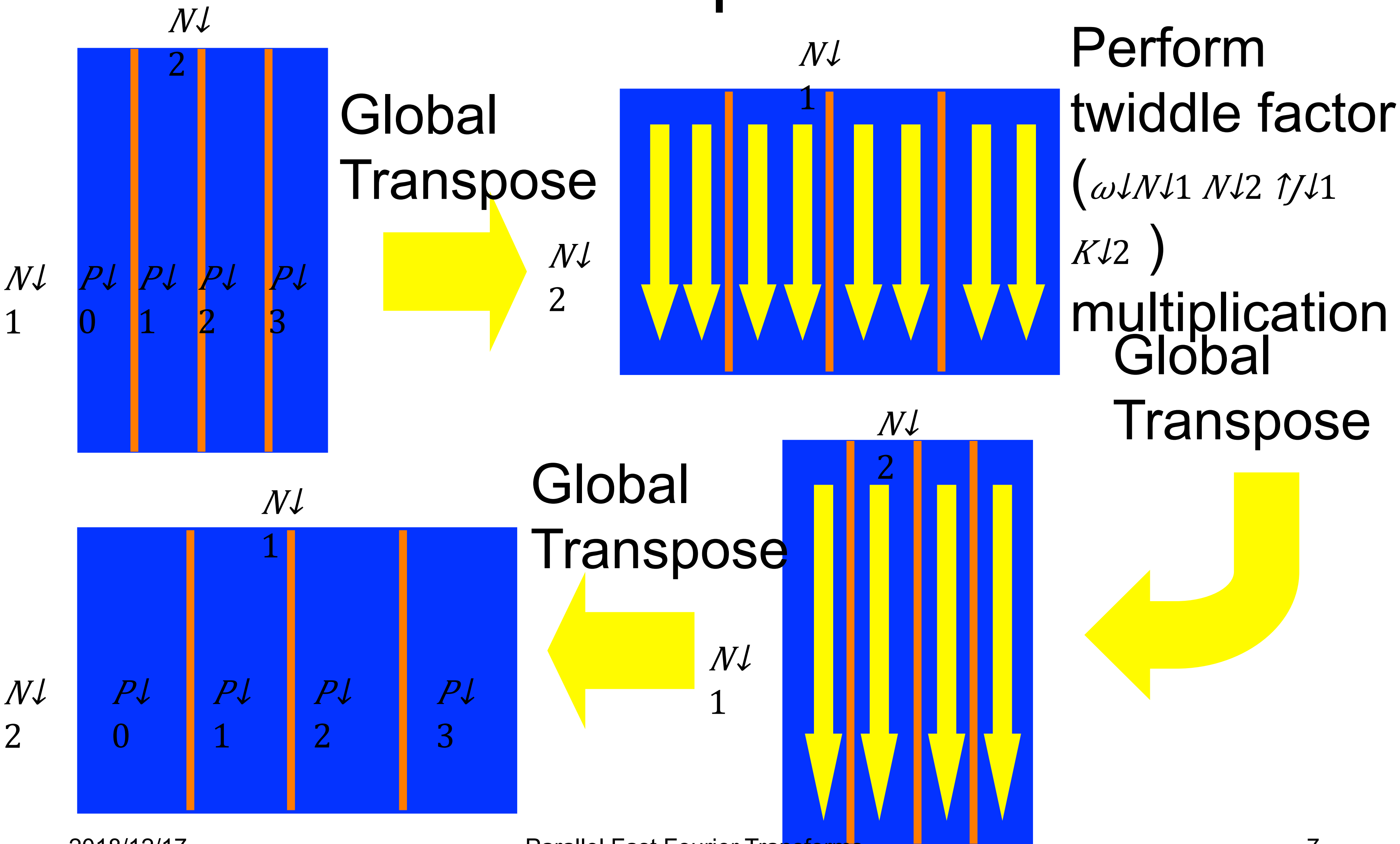
Approach

- The parallel 1-D FFT implemented is based on the six-step FFT algorithm [Bailey 90], which requires two multicolumn FFTs and three data transpositions.
- Using this method, we have implemented an automatic tuning facility for selecting the optimal parameters of the all-to-all communication and the computation-communication overlap.

Six-Step FFT Algorithm [Bailey90]

- Step 1: Transpose
- Step 2: Perform n_1 individual n_2-point multicolumn FFTs
- Step 3: Perform twiddle factor ($\omega_{n_1 n_2}^{j_1 k_2}$) multiplication
- Step 4: Transpose
- Step 5: Perform n_2 individual n_1-point multicolumn FFTs
- Step 6: Transpose

Parallel 1-D FFT Algorithm Based on Six-Step FFT



In-Cache FFT Algorithm and Vectorization

- For in-cache FFT, we used radix-2, 3, 4, 5, 8, 9, and 16 FFT algorithms based on the mixed-radix FFT algorithms [Temperton 83].
- Automatic vectorization was used to access the Intel AVX-512 instructions on the Knights Landing processor.
- Although higher radix FFTs require more floating-point registers to hold intermediate results, the Knights Landing processor has 32 ZMM 512-bit registers.

Optimization of Parallel 1-D FFT on Knights Landing Processor

```
      COMPLEX*16 X(N1,N2),Y(N2,N1)
!$OMP PARALLEL DO COLLAPSE(2) PRIVATE(I,J,JJ)
      DO II=1,N1,NB
        DO JJ=1,N2,NB
          DO I=II,MIN(II+NB-1,N1)
            DO J=JJ,MIN(JJ+NB-1,N2)
              Y(J,I)=X(I,J)
            END DO
          END DO
        END DO
      END DO
!$OMP PARALLEL DO
      DO I=1,N1
        CALL IN_CACHE_FFT(Y(1,I),N2)
      END DO
      …
```

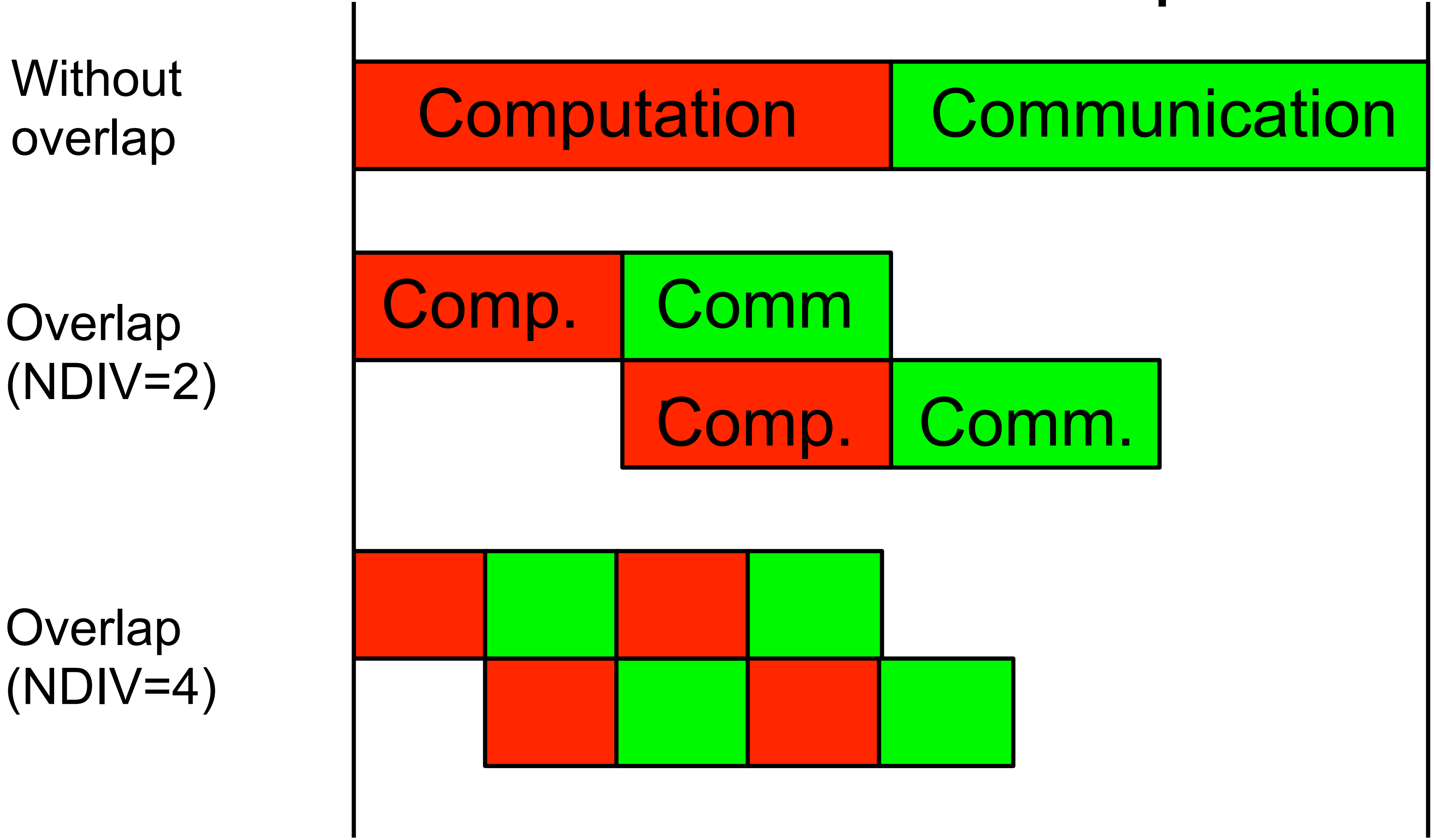
To expand the outermost loop, the double-nested loop can be collapsed into a single-nested loop.

Computation-Communication Overlap [Idomura et al. 2014]

```
!$OMP PARALLEL
!$OMP MASTER
    MPI communication
!$OMP END MASTER
!$OMP DO SCHEDULE(DYNAMIC)
    DO I=1,N
      Computation
    END DO
!$OMP DO
    DO I=1,N
      Computation using the
      result of communication
    END DO
!$OMP END PARALLEL
```

- ← MPI communication is performed on the master thread (MPI communication)
- ← No barrier synchronization (!$OMP END MASTER)
- ← Computation is performed by a thread other than the master thread (Computation)
- ← Implicit barrier synchronization (END DO)
- ← Computation is performed after completion of the MPI communication (Computation using the result of communication)

Pipelined Computation-Communication Overlap



Automatic Tuning of Parallel 1-D FFT

- The automatic tuning process consists of two steps:
  - Automatic tuning of all-to-all communication
  - Selection of the number of divisions NDIV for the computation-communication overlap

Optimizing of All-to-All Communication

- An optimized all-to-all collective algorithm for multi-core systems connected using modern InfiniBand network interfaces [Kumar et al. 08].
- The all-to-all algorithm completes in two steps, intra-node exchange and inter-node exchange.

Two-Phase All-to-All Algorithm

- We extend the all-to-all algorithm to the general case of $P = P_x \times P_y$ MPI processes.

1. Local array transpose from $(N/P^2, P_x, P_y)$ to $(N/P^2, P_y, P_x)$, where N is the total number of elements.

   Then P_y simultaneous all-to-all communications across P_x MPI processes are performed.

2. Local array transpose from $(N/P^2, P_y, P_x)$ to $(N/P^2, P_x, P_y)$.

   Then P_x simultaneous all-to-all communications across P_y MPI processes are performed.

Automatic Tuning of All-to-All Communication

- The two-phase all-to-all algorithm requires twice the total amount of communications compared with the ring algorithm.
- However, for small to medium messages, the two-phase all-to-all algorithm is better than the ring algorithm due to the smaller startup time.
- Automatic tuning of all-to-all communication can be accomplished by performing a search over the parameters of all of P_x and P_y .
- If $P = P_x \times P_y$ is a power of two, the size of search space is $\log_2 P$.

Selection of Number of Divisions for Computation-Communication Overlap

- When the number of divisions for computation-communication overlap is increased, the overlap ratio also increases.
- On the other hand, the performance of all-to-all communication decreases due to reducing the message size.
- Thus, a tradeoff exists between the overlap ratio and the performance of all-to-all communication.
- The default overlapping parameter of the original FFTE 6.2alpha is NDIV=4.
- In our implementation, the overlapping parameter NDIV is varied between 1, 2, 4, 8 and 16.

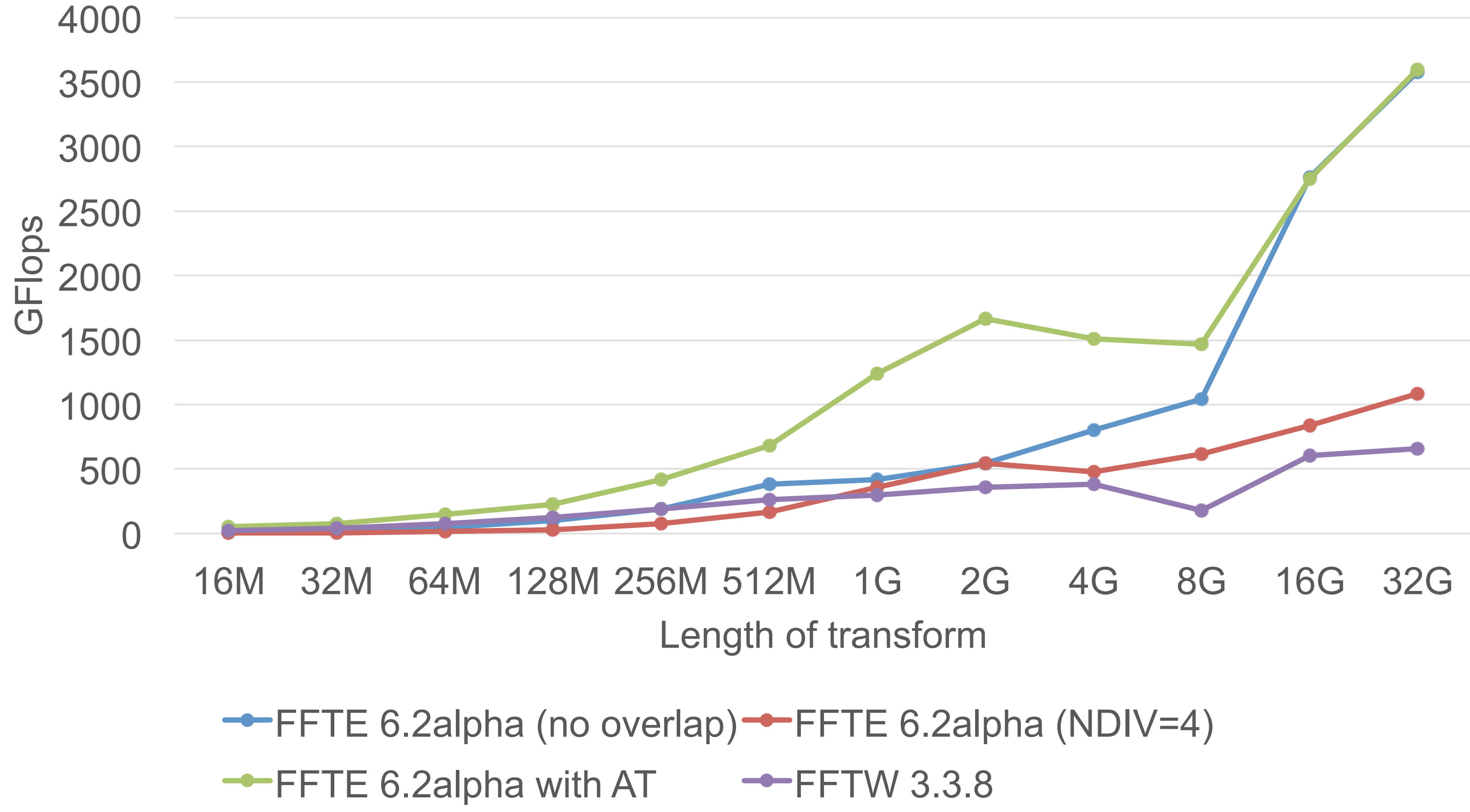
Performance Results

- To evaluate the parallel 1-D FFT with automatic tuning (AT), we compared its performance with that of the FFTW 3.3.8, the FFTE 6.2alpha (http://www.ffte.jp/) and the FFTE 6.2alpha with AT.
- The performance was measured on the Oakforest-PACS (8208 nodes) at Joint Center for Advanced HPC (JCAHPC) and Cori (9688 nodes) at NERSC.
  - CPU: Intel Xeon Phi 7250 (68 cores, Knights Landing 1.4 GHz)
  - Interconnect: Intel Omni-Path Architecture (OFP),
Cray Aries with Dragonfly topology (Cori)
  - Compiler: Intel Fortran compiler 18.0.1.163 (for FFTE)
Intel C compiler 18.0.1.163 (for FFTW)
  - Compiler option: “-O3 -xMIC-AVX512 -qopenmp”
  - MPI library: Intel MPI 2018.1.163
  - cache/quadrant (The amount of memory used at each node is less than 2 GB)
  - Each MPI process has 64 cores and 64 threads.

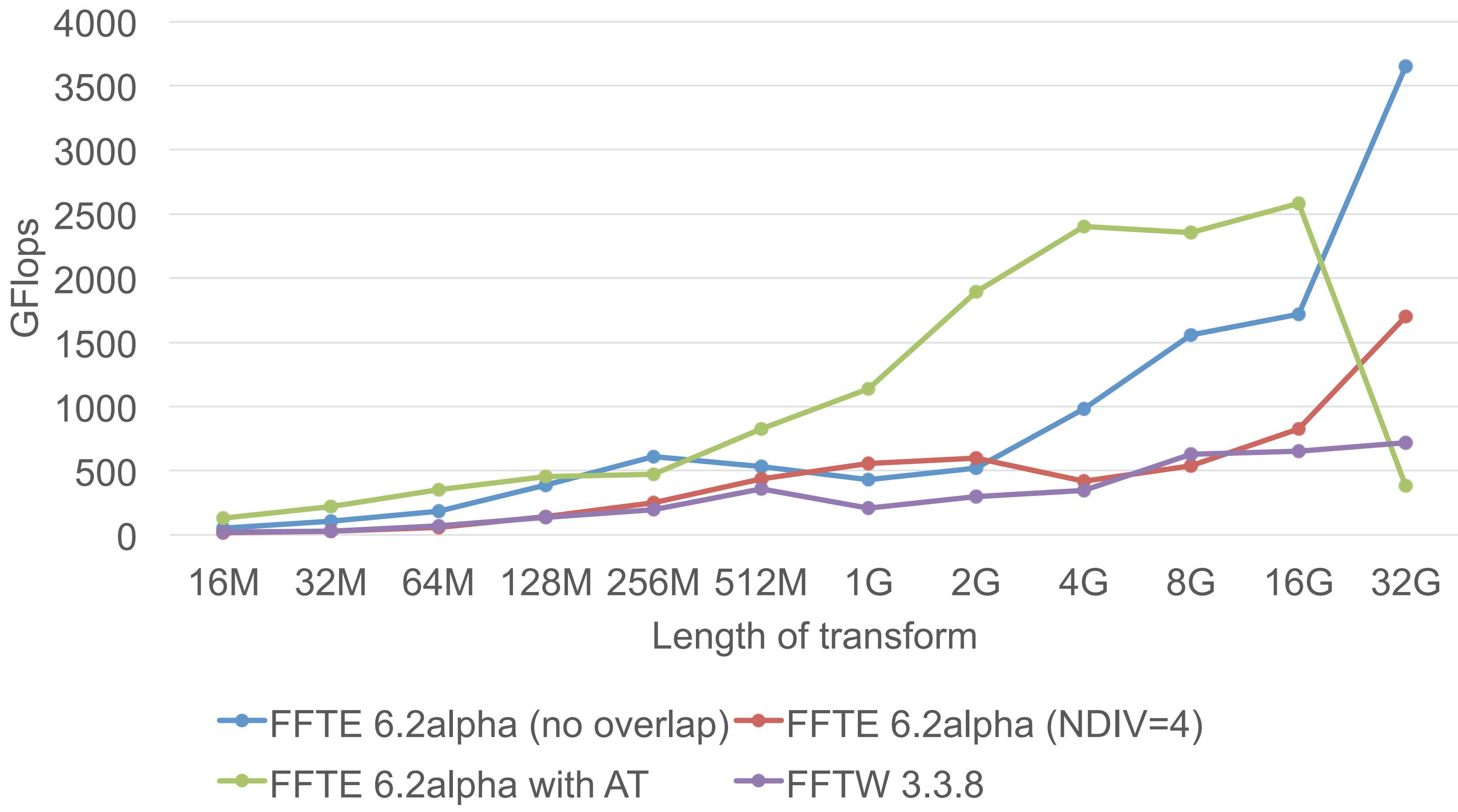
Results of automatic tuning of parallel 1-D FFTs (Oakforest-PACS, 1024 nodes)

	FFTE 6.2alpha			FFTE 6.2alpha with AT			
N	P	NDIV	GFlops	P_x	P_y	NDIV	GFlops
16M	1024	4	2.7	32	32	1	53.1
64M	1024	4	13.8	8	128	1	148.1
256M	1024	4	75.6	16	64	1	416.9
1G	1024	4	359.6	32	32	1	1241.4
4G	1024	4	479.3	8	128	1	1509.6
16G	1024	4	839.9	1024	1	1	2749.3

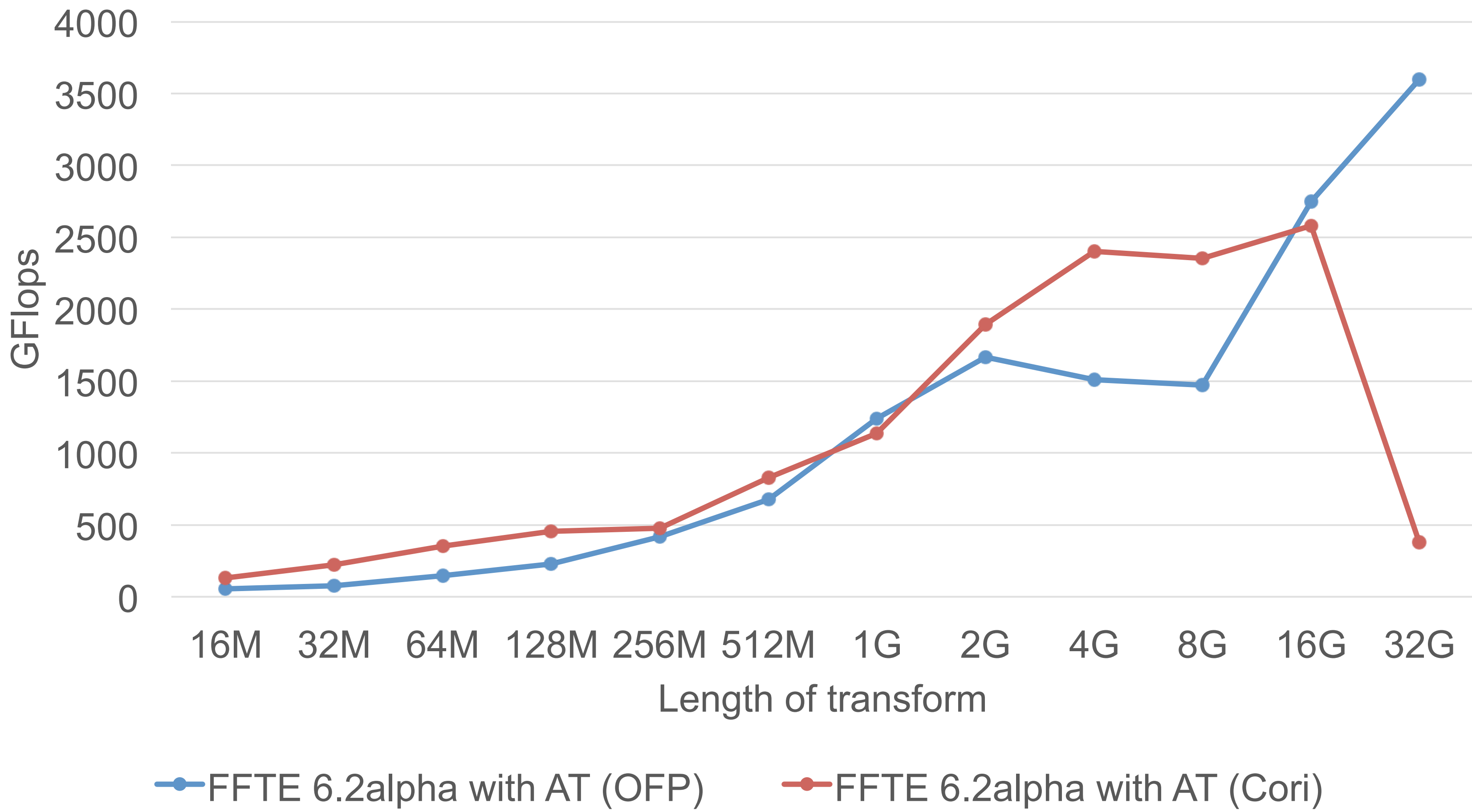
Performance of parallel 1-D FFTs (Oakforest-PACS, 1024 nodes)



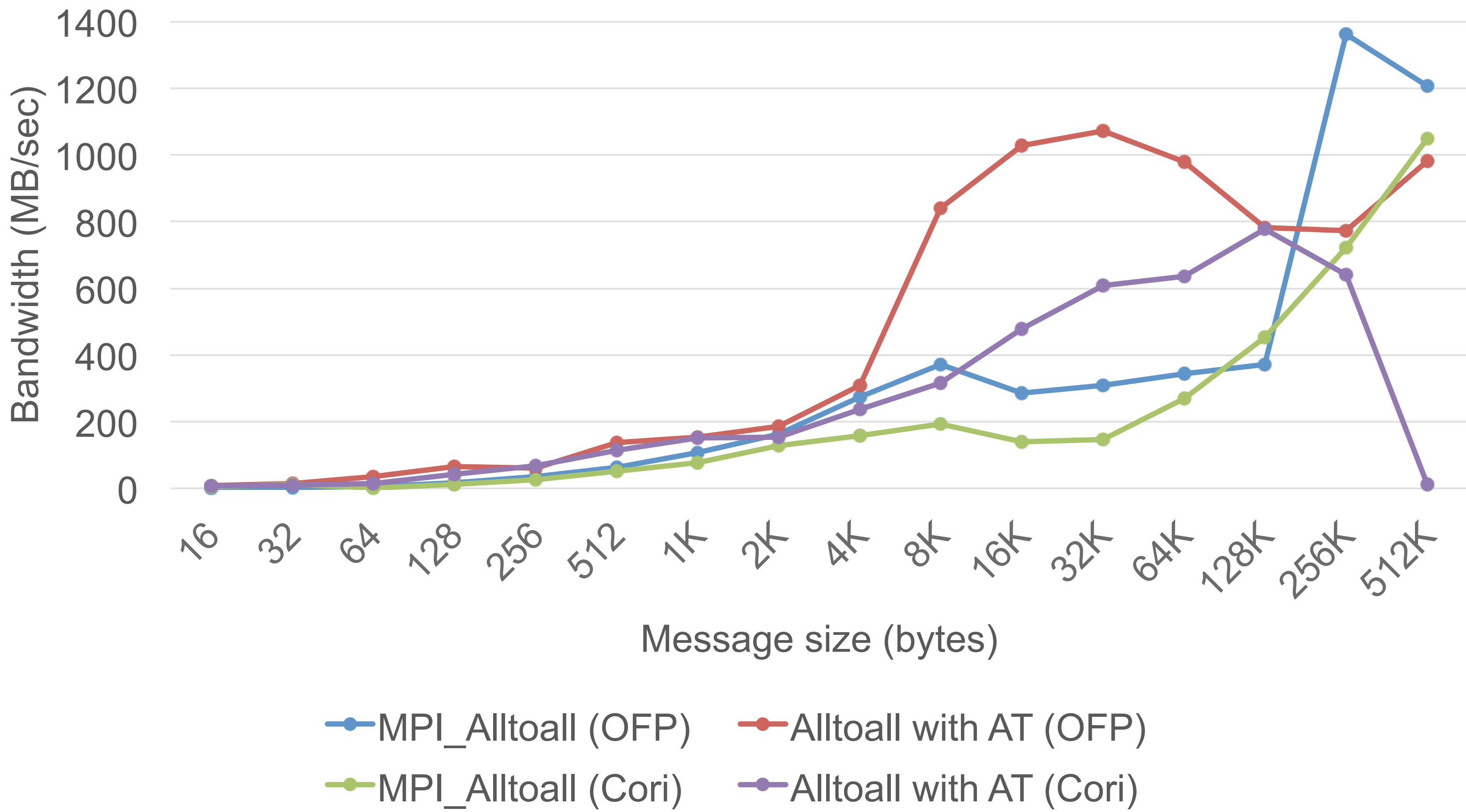
Performance of parallel 1-D FFTs (Cori, 1024 nodes)



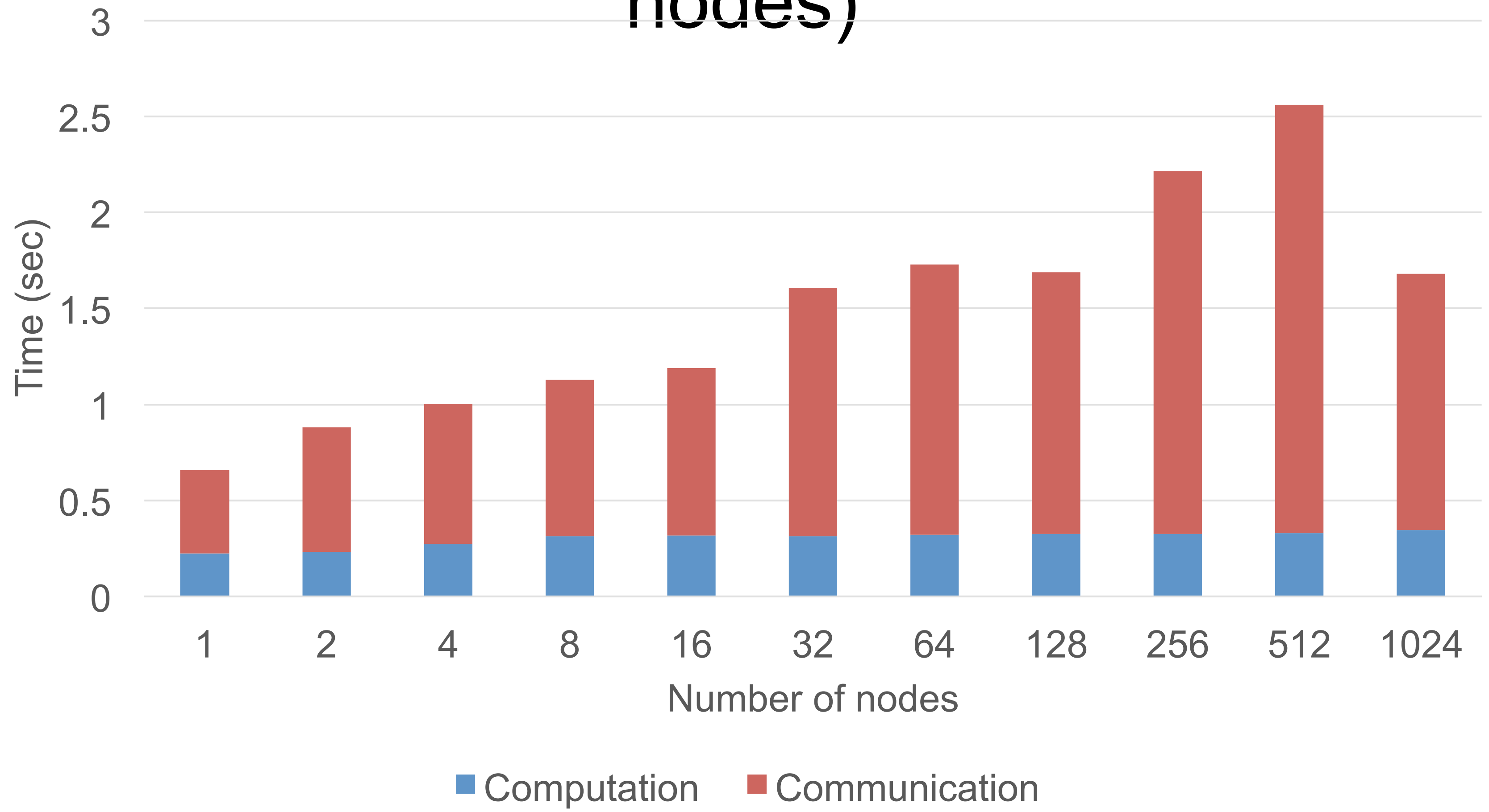
Performance of parallel 1-D FFTs (Oakforest-PACS vs Cori, 1024 nodes)



Performance of all-to-all communication (Oakforest-PACS vs Cori, 1024 nodes)



Breakdown of execution time in FFTE 6.2alpha (no overlap) (Oakforest-PACS, N=2↑25 ×number of nodes)



Conclusion

- We proposed an implementation of parallel 1-D FFT with automatic tuning on cluster of Intel Xeon Phi processors.
- We used a computation-communication overlap method that introduces a communication thread with OpenMP.
- An automatic tuning facility for selecting the optimal parameters of the all-to-all communication and the computation-communication overlap, was implemented.
- The performance results demonstrate that the proposed implementation of a parallel 1-D FFT with automatic tuning is efficient for improving the performance on cluster of Intel Xeon Phi processors.