FFT in the Exascale: Opportunities and Challenges

Daisuke Takahashi

Center for Computational Sciences
University of Tsukuba, Japan

FFTE: A Fast Fourier Transform Package

- FFTE is a Fortran subroutine library for computing the Fast Fourier Transform (FFT) in one or more dimensions.
- It includes real, complex, mixed-radix and parallel transform.
- FFTE may be faster than other publically-available FFT implementations and vendor-tuned libraries.
- Available at http://www.ffte.jp/

Features

- Parallel transforms
  - Shared / Distributed memory parallel computers (OpenMP, MPI, OpenMP + MPI, and CUDA + MPI)
- High portability
  - Fortran + OpenMP + MPI
- Data layout
  - 1-D decomposition
  - 2-D decomposition (for parallel 3-D FFT)
- HPC Challenge Benchmark
  - FFTE’s 1-D parallel FFT routine has been incorporated into the HPC Challenge (HPCC) benchmark.

Parallel 3-D FFT (1/2)

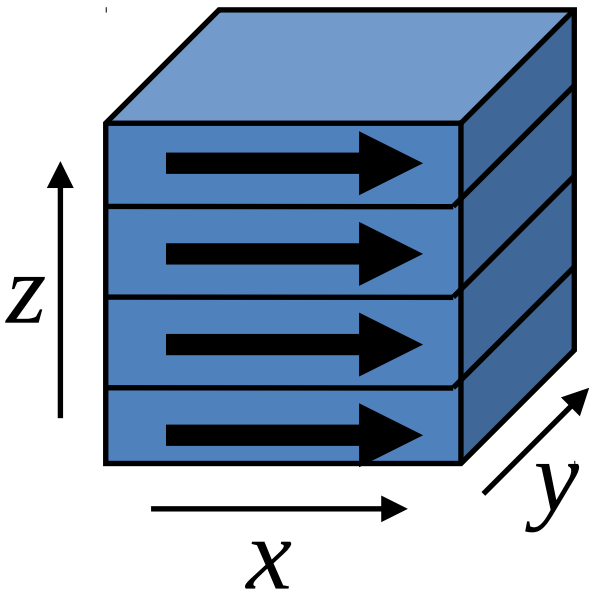
- Parallel 3-D FFT algorithms on distributed-memory parallel computers have been well studied.
- June 2018 TOP500 Supercomputing Sites
  - Summit (IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100): 122.3 PFlops (2,282,544 Cores)
  - Sunway TaihuLight (Sunway SW26010 260C 1.45GHz): 93.01 PFlops (10,649,600 Cores)
- Recently, the number of cores keeps increasing.

Parallel 3-D FFT (2/2)

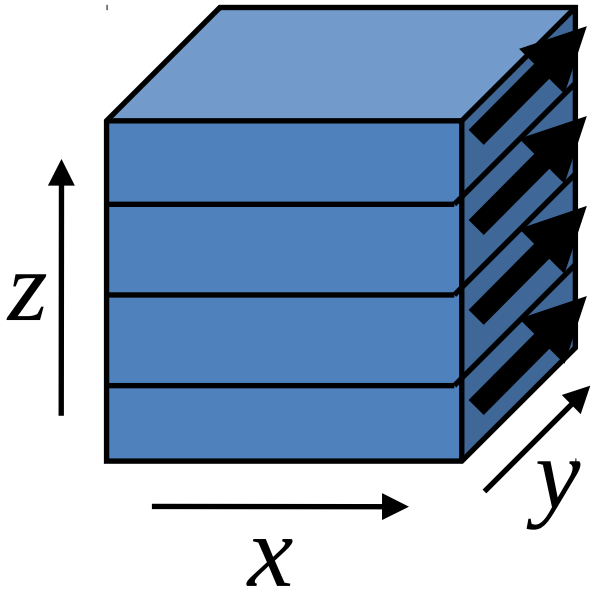
- A typical decomposition for performing a parallel 3-D F FT is slabwise.
- A 3-D array　is distributed along the third dimension .
- must be greater than or equal to the number of MPI pr ocesses.
- This becomes an issue with very large node counts for a massively parallel cluster of many-core processors.
- P3DFFT and 2DECOMP&FFT support the 2-D decomposition.

1-D Decomposition along the z-axis

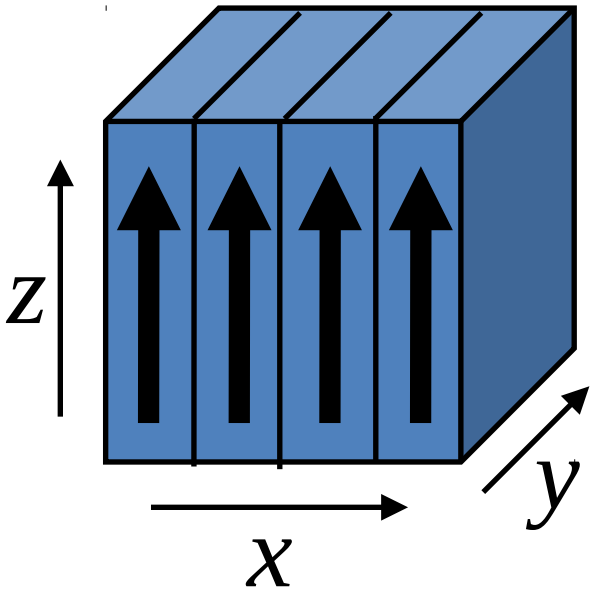
1. FFTs in x-axis



2. FFTs in y-axis



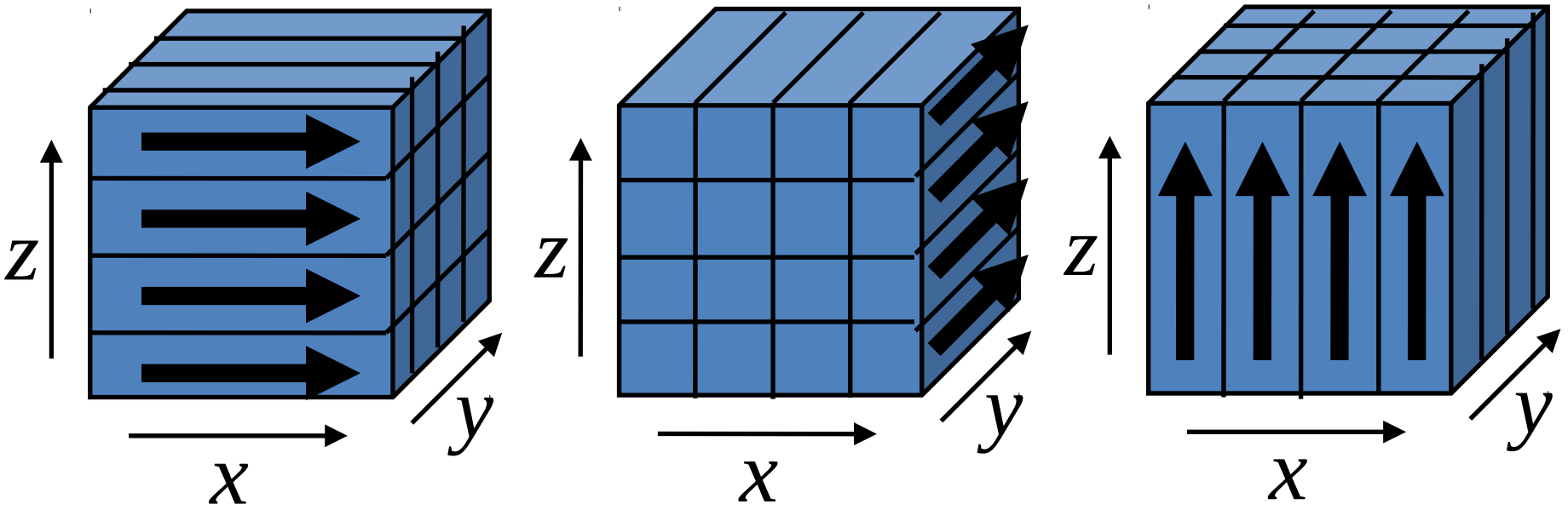
3. FFTs in z-axis



With a slab decomposition

2-D Decomposition along the y- and z-axes

1. FFTs in x-axis 2. FFTs in y-axis 3. FFTs in z-axis



With a pencil decomposition

Comparing Communication Time

- Communication time of 1-D decomposition

$$T_{1\mathrm{dim}} \approx PQ \cdot L + \frac{16N}{PQ \cdot W}$$

- Communication time of 2-D decomposition

$$T_{2\mathrm{dim}} \approx (P + Q) \cdot L + \frac{32N}{PQ \cdot W}$$

- By comparing two equations, the communication time of the 2-D decomposition is less than that of the 1-D decomposition for larger number of MPI processes PQ and latency L.

Feasibility Study of FFT in the Exascale

- FFT will be still needed in exascale computing.
- However, it seems that Global FFT using whole exascale system is not realistic.
- The performance of parallel one-dimensional FFT in K comp uter (82944 nodes, 10.6 PFlops peak) was only 252
- TFlops (approx. 2.4% of peak).
  - More than 2/3 of the execution time is dominated by all-to-all comm unication.
  - Sustained performance exceeding PFlops in Global FFT has not ye t been achieved.
- The upper limit of the performance of FFT is determined by the performance of all-to-all communication.

An Example of Conditions for Exaflops in FFT

- Assuming that computation and communication are completely overlapped, the performance of the FFT depends on the total communication bandwidth.
- The number of data points for FFT:
  - The number of arithmetic operations:
  - Memory usage: 8PB
- The number of MPI processes:
  - All-to-all message size: 256MB
- In this case, the communication bandwidth of approx. 7. 95 GB/s per MPI process is required.