Extreme Resilience and Deeper Memory Hierarchy

TOKYO TECH Pursuing Excellence

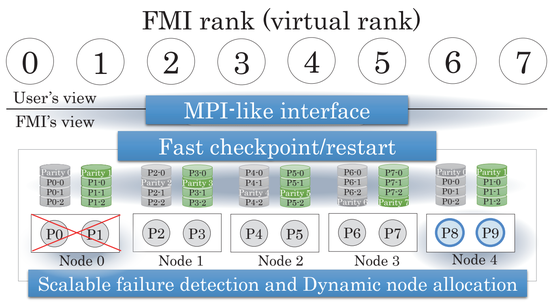
Fault Tolerant Infrastructure for Billion-Way Parallelization

Project introduction

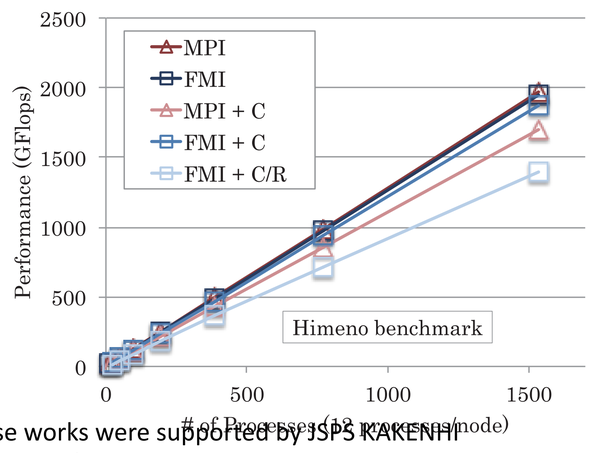
Thanks to Supercomputers, the large-scale simulations can be achieved. However, the increasing number of nodes and components will lead to a very high failure frequency. In Exa-scale supercomputers, the MTBF will be no more than tens of minutes, which means computing node doesn't work in effect. We're seeking a solution to the problem.

FMI: Fault Tolerant Messaging Interface

FMI is an MPI-like survivable messaging interface that enables scalable failure detection, dynamic node allocation, fast and transparent recovery.




```
int main (int *argc, char *argv[]) {
  FMI_Init(&argc, &argv);
  FMI_Comm_rank(FMI_COMM_WORLD, &rank);
  /* Application's initialization */
  while ((n = FMI_Loop(…)) < numloop) {
    /* Application's program */
  }
  /* Application's finalization */
  FMI_Finalize();
}
```



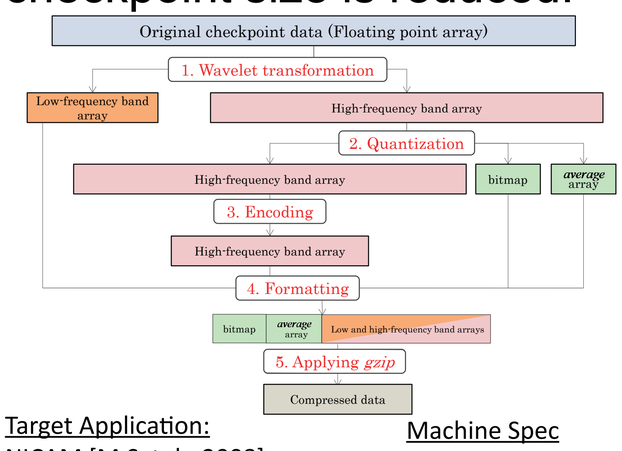
These works were supported by JSPS KAKENHI Grant Number 23220003.

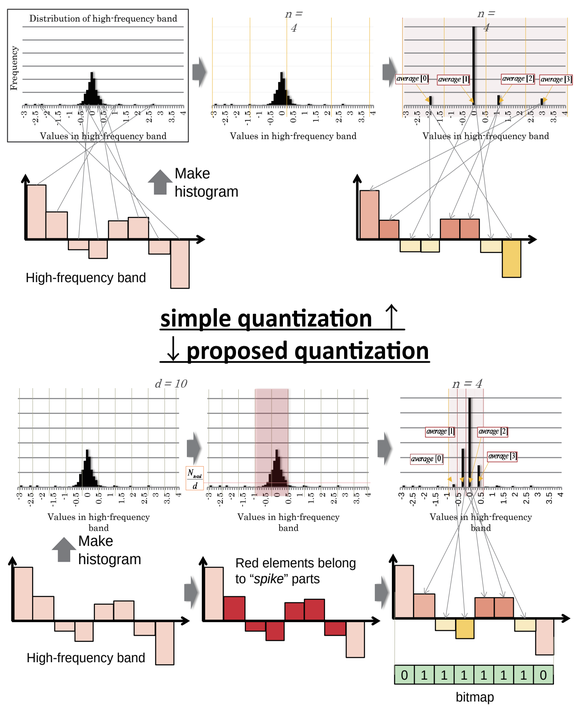
Lossy Compression for cp./rst.

To reduce checkpoint time, lossy compression is applied to checkpoint data then checkpoint size is reduced.

Target
- 1,2,3D mesh w/o pointer

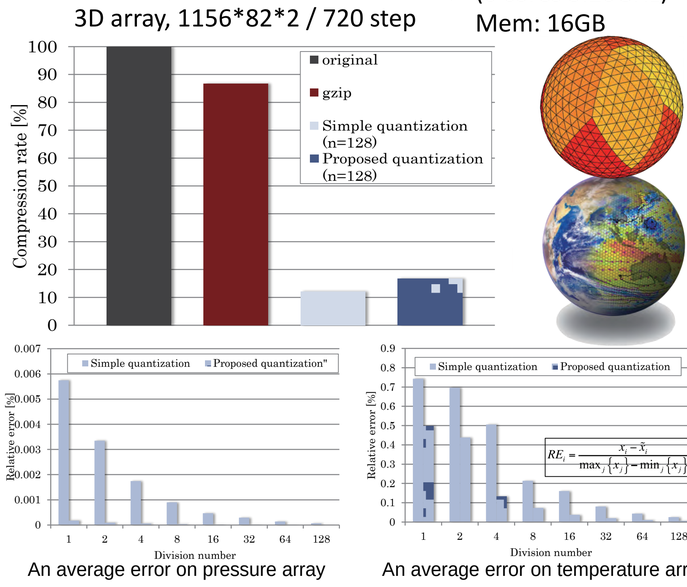
Approach:
1. wavelet transformation
2. quantization
3. encoding
4. gzip





Target Application:
NICAM [M.Satoh, 2008]
climate simulation about
Pressure, temperature and velocity
3D array, 1156*82*2 / 720 step

Machine Spec
CPU:
Intel Core i7-3930K
(6 cores 3.20GHz)
Mem: 16GB



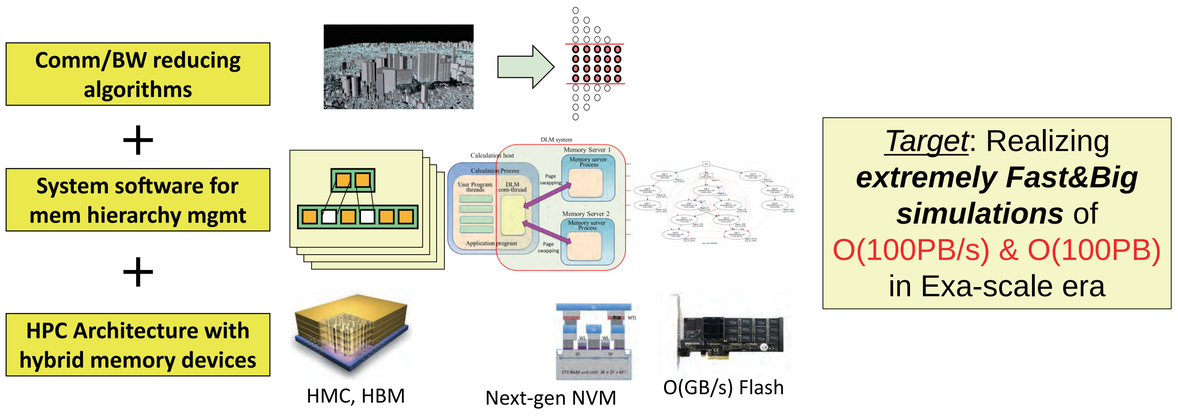
Cp. size is much reduced with low error in particular situation

Dealing with Deeper Memory Hierarchy

Overview of Project

On Exa-scale supercomputers, the ***"Memory Wall"*** problem will become even more severe, which prevents the realization of ***Extremely Fast&Big Simulations***.

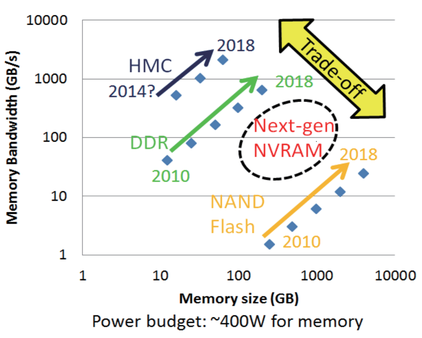
This project promotes research towards this problem via co-design approach among application algorithms, system software, architecture.



Target: Realizing ***extremely Fast&Big simulations*** of O(100PB/s) & O(100PB) in Exa-scale era

Target Architecture

Deeper memory hierarchy that consists of heterogeneous memory devices



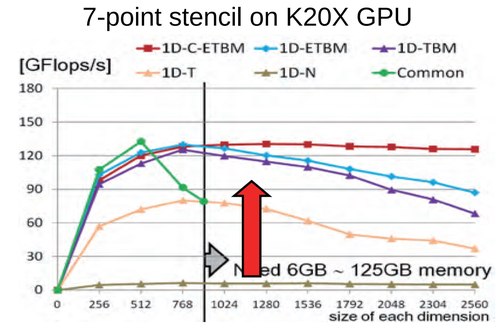
Hybrid Memory Cube (HMC):
DRAM chips are stacked with TSV technology. It will have advantage in bandwidth over DDR, but capacity will be smaller.

NAND Flash:
SSDs are already commodity. Newer products, such as IO-drive have O(GB/s) bandwidth.

Next-gen non-volatile RAM (NVRAM):
Several kinds of NVRAM such as STT-MRAM, ReRAM, FeRAM, etc, will be available in a few years.

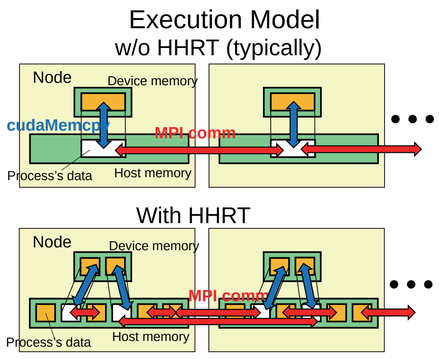
Highly Optimized Stencils Larger than GPU Memory

For extremely large stencil simulations, we implemented temporal blocking (TB) technique and clever optimizations on GPUs [1][2].
- Eliminating redundant computation
- Reducing memory footprint of TB algorithm



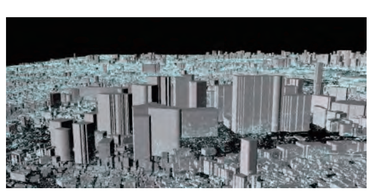
HHRT: System Software for GPU Memory Swap

For easier programming, we implemented system software, named HHRT (hybrid hierarchical runtime) [3].
- HHRT supports user programs written in MPI and CUDA with little modification
- Oversubscription based execution model
- HHRT implicitly supports memory swapping between GPU memory and host

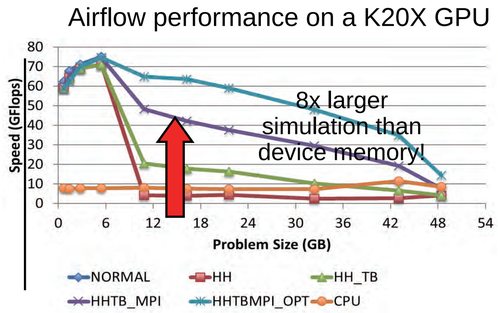


Integration with Real Simulation Application

We integrated our techniques with the city airflow simulation.



Original code on MPI+CUDA was developed by Naoyuki Onodera, Tokyo Tech. We integrated TB into it and executed on HHRT.



[1] G. Jin, T. Endo, S. Matsuoka. A Parallel Optimization Method for Stencil Computation on the Domain that is Bigger than Memory Capacity of GPUs . IEEE Cluster 2013.
[2] G. Jin, J. Lin, T. Endo. Efficient Utilization of Memory Hierarchy to Enable the Computation on Bigger Domains for Stencil Computation in CPU-GPU Based Systems. IEEE ICHPA 2014.
[3] T. Endo, G. Jin: Software Technologies Coping with Memory Hierarchy of GPGPU Clusters for Stencil Computations. IEEE Cluster 2014.

PI: Toshio Endo (endo@is.titech.ac.jp), supported by JST-CREST

http://www.gsic.titech.ac.jp/sc15