# Wrap-up of Discussions in Group B
# Technologies and applications to integrate exascale computing and big data


Members:

David Haensel , Sheng Di, Yuya Takashina, Huang Bo, Wataru Uemura

Mentors:

Koji Terasaki, Miwako Tsuji

# Motivations

- Exascale computing will play a special role in information technology

In exa-scale era:

- Data size is a big problem: memory and I/O do not increase as fast as processor
- HPC applications are very diverse and complicated
- Faults/errors become normal and common

# Challenges in exascale computing

- Faults
  - Fail-stop error
  - Silent data corruption
- Scaling
- Storage cost
- Overheads for checkpoint/restart

# Challenges in applications

- Storage cost
  - HACC (Hardware/Hybrid Accelerated Cosmology Code)
    - 20 petabytes data for a single 1-trillion-particle simulation
    - Mira only has 26 petabytes file system storage
  - CESM (climate research)
    - 2.5 petabytes data produced
    - 170 TB of postprocessing data

# Challenges in other applications

- In quantum chemistry, 10^9 x 10^9 (sparse) matrices as a quantum system must be dealt.
- In MD, huge protein will be target, and simulation for 100ns takes 20days calculation and costs 4TB storage.
- In MD, #particles in a CPU is too small, so communication should be overhead.
- In tensor decomposition, it becomes more difficult to control tensors as they have higher dimensions.

# Contributions: what we are doing (or will do) for the topic

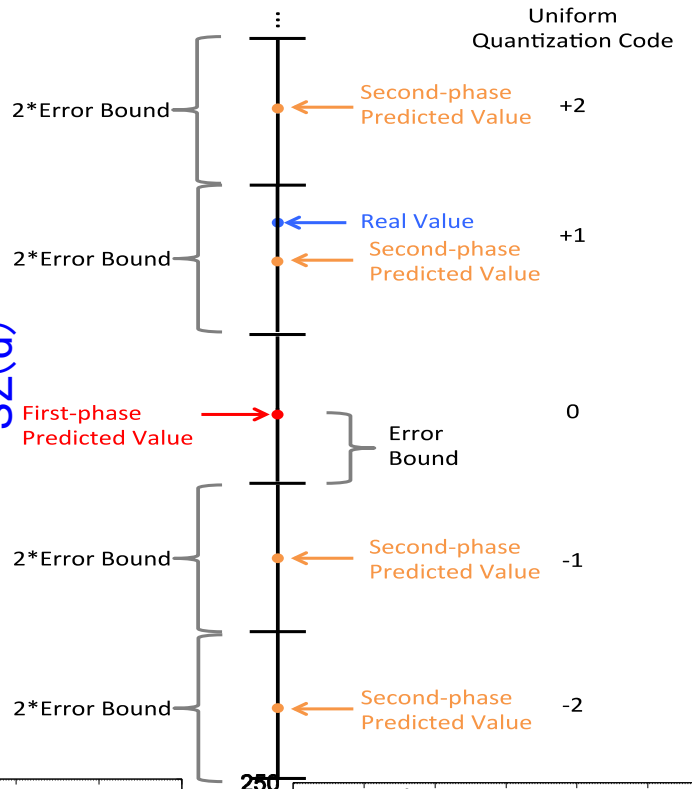
- SZ: lossy data compressor for large scale scientific data
  - Free to download under BSD license
  - 10+ users around the world
- AID : Adaptive impact-driven detector for silent data corruption
- Dynamics approximation method for large scale MD simulations
  - MSES (Multiscale Enhanced Sampling)
- C++ library for high level abstractions
  - intra/inter-node parallelization, tasking
- Anomaly detection method for time-series data mining
- Large scale parallelization in quantum chemistry
  - Tensor decomposition

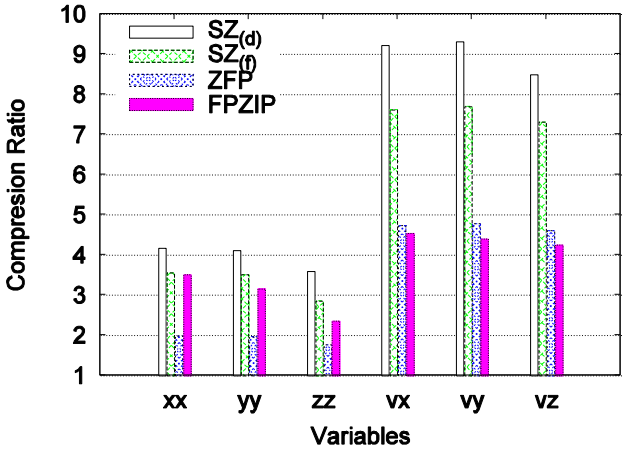
# (1) SZ: Fast Lossy Data Compressor

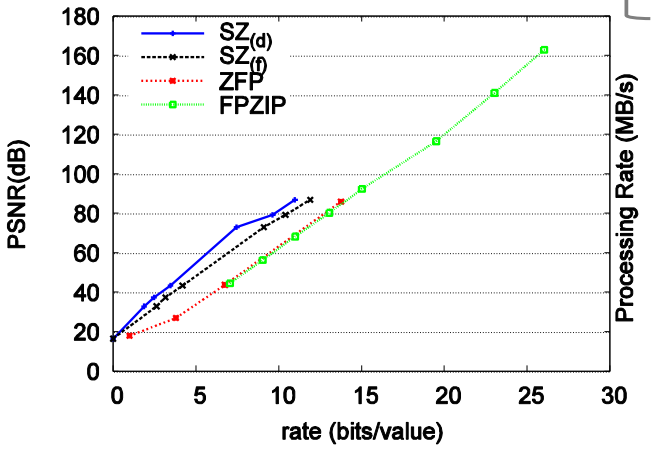
The whole compression procedure:

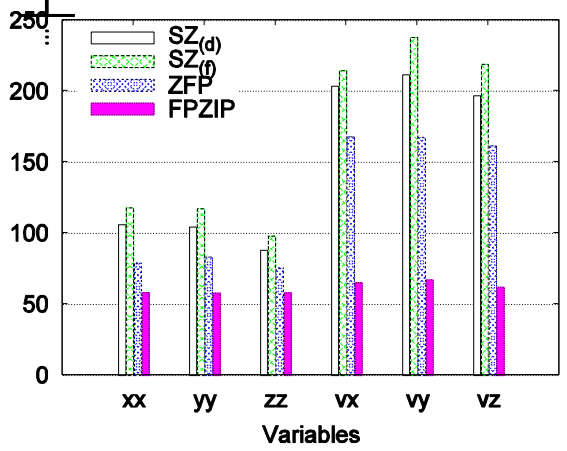
(1) Point-wise data prediction
(2) Linear scaling (error-bounded quantization)
(3) Variable-length encoding (Huffman)
(4) Non-predictable data compression
(5) Lossless compression (LZ77/Gzip)

SZ(f)
SZ(d)



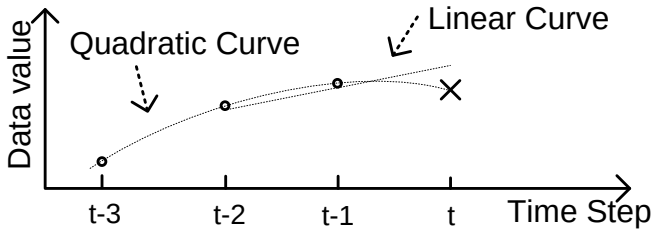



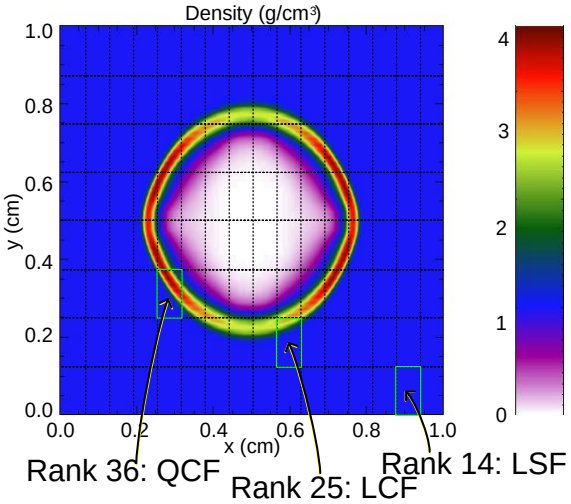



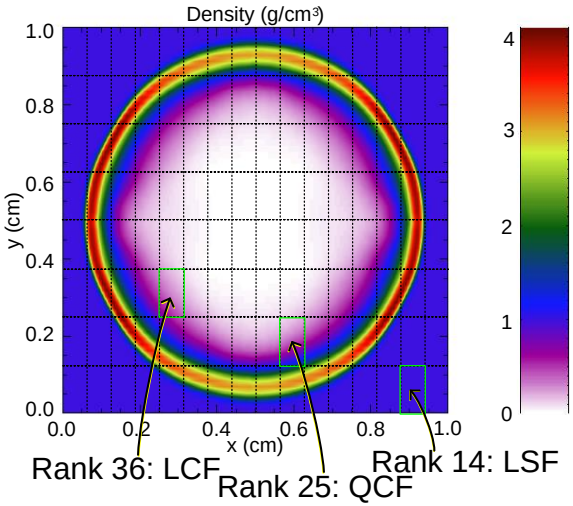
# (2) Adaptive Impact-driven Detector (AID)

- Key idea: Each rank selects best-fit prediction method based on local data dynamics at runtime
- Prediction method:
  - Preceding Neighbor Fitting (PNF)
  - Linear Curve Fitting (LCF)
  - Quadratic Curve Fitting (QCF)




(a) time step 100




(b) time step 200

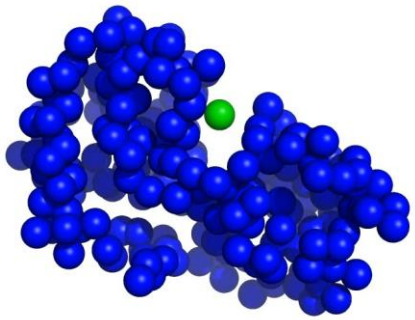
Sedov Shock Simulation

# (3) Dynamics approximation method for large scale MD simulations

Problem: the huger protein will be calculated in the next step, if we use the usual MD simulation, it will cost many years to do that even if we use a better supercomputer.

contribution :In our lab we develop the Multiscale Enhanced Sampling (MSES) method which based on coarse-grained (CG) models to enhance the sampling of atomistic protein conformations.

CG Model

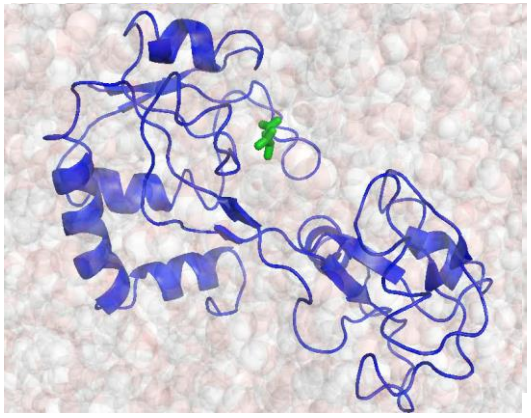


about 5week

$$V(\mathbf{r}_{MM},\mathbf{r}_{CG}) = V_{MM}(\mathbf{r}_{MM}) + V_{CG}(\mathbf{r}_{CG})$$
$$+k_{MMCG}V_{MMCG}\left[\chi_{MM}(\mathbf{r}_{MM}) - \chi_{CG}(\mathbf{r}_{CG})\right]^2$$

Hamiltonian replica exchange to unbias $V_{MMCG}$

MM Model



about 5years

**Caculation scale will be improved by dynamics approximate**

# (4) C++ library for tasking

- Problem:
  - Intra-Node parallelization taking advantage of algorithm specifics

- Guidelines:
  - Encapsulation of responsibility
  - Preserve type-safety
  - Use template meta programming to
  - Avoid virtual function calls

- Contribution:

  C++ tasking infrastructure based on C++11 and STL.

# (5) Anomaly Detection method

- HPC architecture must be fault tolerant.
- Time series anomaly detection can be used to prevent faults by detecting fault signs in advance.
- It is difficult to grasp long-term dependencies between a fault and its sign from log data.

- We will provide a method to deal with the longer context, constructing hierarchical structure of Markov chain.

# (6) Large scale parallelization in quantum chemistry

## Problem:

- The many electron problem in quantum mechanics causes a necessity to treat very large sizes of matrices when the number of electrons increases. It behaves exponentially with the size of the system.
- Also there arises the well known negative sign problem that is a very difficult problem in the quantum monte carlo. This is one of the very important problem in quantum chemistry and condensed matter physics.
- It is highly non trivial how exa-scale computer can contribute to these important scientific problems.

## Contribution:

- Our formalism gives a compact way to represent the many electron wavefunctions in a naturally parallelized way.
- Huge parallelization in the machine would reproduce the correlation energy of the electronic system very efficiently.
- It can give a good example to utilize the high capability of large architecture.

# Summary and Conclusion

- For “Technologies and applications to integrate exascale computing and big data”, during our discussions:
  - we have recognized some problems/challenges in presentations on 1st day
  - we have discussed about what we can contribute to solve these problems