jh190015-NAJ

Developing Accuracy Assured High Performance Numerical Libraries for Eigenproblems

Takahiro Katagiri (Nagoya University)

Abstract
Eigenproblem is one of essential numerical problems for several numerical simulations. Its accuracy, however, is not well-assured in many conventional numerical computations. Basic Linear Algebra Subprograms (BLAS) is a frequently used to perform linear algebra computations. Ensuring the accuracy of the computational results of BLAS operations is a still crucial problem now. Even in solving linear equations using LAPACK is also a typical example, because LAPACK is rich in BLAS operations, especially matrix-matrix multiplication (MMM) operations for solving linear equations. With respect to this background, we focus on the following three topics: (1) Developing an accuracy assured numerical libraries for eigenproblems; (2) Development of high-performance implementation and auto-tuning (AT) technology for the developed accuracy assured numerical libraries; (3) Discussing an extension for non-liner problems based on obtained knowledge of accuracy assured algorithms.

1. Basic Information

(1) Collaborating JHPCN Centers

Tokyo, Nagoya, Kyushu

(2) Research Areas

- ☐ Very large-scale numerical computation

(3) Roles of Project Members

- Prof. Katagiri: High-performance implementation of Osaki method for recent multicore CPUs, and applying auto-tuning technologies.
- Prof. Hwang: Non-linear algorithms for actual engineering problems.
- Dr. Marques: Algorithms and implementations for eigenproblem.
- Prof. Nakajima: Sparse iterative algorithms for liner equation solvers, such as parallel preconditioners.
- Prof. Ogita: Iterative refinement algorithm to assure accuracy of real symmetric eigenproblem.
- Prof. Ohshima: GPGPU implementations.
- Prof. Ozaki: Accurate MMM algorithm (Ozaki method)
- Prof. Wang: Eigenvalue algorithms for actual engineering problems.

2. Purpose and Significance of Research

Eigenproblem is one of essential numerical problems for several numerical simulations. Its accuracy, however, is not well-assured in many conventional numerical computations. Basic Linear Algebra Subprograms (BLAS) is a frequently used to perform linear algebra computations. Ensuring the accuracy of the computational results of BLAS operations is a still crucial problem now. Even in solving linear equations using LAPACK is also a typical example, because LAPACK is rich in BLAS operations, especially matrix-matrix multiplication (MMM) operations for solving linear equations.

1. We focus on the following three topics:
 Developing an accuracy assured numerical libraries for eigenproblems;
2. Development of high-performance implementation and AT technology for the developed accuracy assured numerical libraries;
3. Discussing an extension for non-liner problems based on obtained knowledge of accuracy assured algorithms.

3. Significance as JHPCN Joint Research Project

We have significant research results related to this project. The followings are summary.

Accuracy Assured Algorithm for Eigenproblems: We have mentioned this. Prof. Ogita developed an algorithm for accuracy assured real symmetric eigenproblem. We use this algorithm to establish accuracy assured numerical library in this project. The algorithm is based on iterative refinement algorithm. Several tuning parameters for high-performance implementations are including, such as eigen decomposition, MMM, stop criteria for iteration, etc. These are nice targets for adapting auto-tuning.

Accurate Matrix-Matrix Multiplication (Ozaki Method): Prof. Katagiri developed a high-performance parallel implementation for Ozaki method with Prof. Ozaki and Prof. Ozaki. Ozaki method requires multiple MMMs after error-free transformation (See Fig. 1).

```
Function EF = EFT_Mul(A, B)
  [A, n_A] := Split_A;  [B, n_B] := Split_B;
  k := 1;
  for i =1: n_A
    for j =1: n_B
      EF { k } := A { i } * B { j };   k := k + 1;
    end
  end
end
```

Fig. 1 Overview of Ozaki Method.

Decomposed matrices after the error-free transformation (*Split_A* and *Split_B* in Fig. 1) make sparse matrices in some situation. We use sparse matrix operations for the multiple MMMs in this implementation to establish remarkable speedups (38.6x). This performance evaluation was done with the Fujitsu FX100 in Nagoya University, which is a K-computer type supercomputer.

There are many tuning parameters for the implementations, such as criteria for dense and sparse operations, sparse implementations (sparse formats, sparse matrix-vector multiplications (SpMV), and sparse-sparse multiplications (SpMxSpM).) In addition, criteria between CPU and GPU computing is also important tuning parameters. These are targets for auto-tuning.

Accuracy Assured Numerical Library for Linear Equations: Some research results, including high-performance implementation of Ozaki method, have been opened as opens source software (OSS). Please refer to UNC-HPC homepage. (http://www.math.twcu.ac.jp/ogita/post-k/index.html)

The current released libraries via the UNC-HPC homepage are as follows: (1)LINSYS_VR: Verified Solution of Linear Systems with Directed Rounding; (2) LINSYS_V: Verified Solution of Linear Systems; (3) DHPMM_F: High-precision Matrix Multiplication (Ozaki method) with Faithful Rounding; (4) BLAS-DOT2: Higher-precision BLAS based on Dot2; (5) OzBLAS: Accurate and Reproducible BLAS based on Ozaki scheme.

We make high performance library for accuracy assurance base on the UNC-HPC routines in this project.

4. Outline of Research Achievements up to FY2018

This year is the first year in this project.

5. Details of FY2019 Research Achievements

The topic is shown as follows:

The Year 1 (FY2019):

1) Topic 1: Performance evaluation of

high-performance implementations for UNC-HPC libraries between multi-core and many-core CPUs and a GPU.

2) Topic 2: Designing accuracy assured libraries for real symmetric eigenproblem.

3) Topic 3: Discussing extension to non-linear problems.

- Results for the Topic 1

To do the topic 1, we developed a new implementation for an accurate MMM (Ozaki Method) library, including the UNC-HPC library.

i. Sparse Matrix-vector Multiplication (SpMV) Implementation for Ozaki Method

We describe the calculation time of the SpMV routine in the Compressed Row Storage (CRS) and Ell-pack (ELL) formats in the CPU and GPU environments for a test matrix.

The whole duration of the routine includes the error-free conversion time, duration of the change to the sparse matrix format, and actual calculation time. The error-free conversion time is "error_free"; the conversion time of matrix A to the sparse matrix format and the memory transfer time from the CPU to the GPU is "setA"; the SpMV routine time is "kernel"; the memory transfer time from the CPU to the GPU of the matrix B and from the GPU to the CPU of the matrix C is "SetB,C"; the duration of the remaining operations is given under "other".

Fig. 2 show that when the matrix size is 10,000 in the CRS format, the GPU environment provides a shorter calculation time with the SpMV routine.

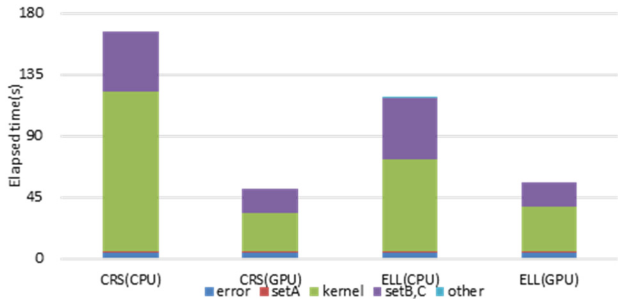

Fig. 2. Execution speed of the SpMV routine with the format and ELL formats in the CPU and GPU environments for a test matrix with N = 10,000 in accurate MMM library.

In the ELL format, when the matrix size is 10,000, the GPU environment results in shorter calculation times with the SpMV routine. This is because when the matrix size is small, the cost of the memory transfer to the GPU device is large relative to the calculation time. However, as the matrix size increases, the cost of the memory transfer relative to the calculation time decreases. Also, when the matrix size is small, it is assumed that the rise time of the GPU pipeline cannot be ignored compared to the SpMV calculation time.

The execution time of the entire routine in GPU execution achieved a maximum 30.9% reduction with the CRS format, and a maximum 37.7% reduction with the ELL format compared to CPU execution.

ii. Sparse Matrix-Matrix Multiplication (SpMxSpM) Implementation for Ozaki Method

We have developed an implementation of SpMxSpM with CRS format for Ozaki method in GPU environment. In this section, we evaluate performance of the SpMxSpM implementation for Ozaki method with cuBLAS. In addition, sparse matrix-matrix

(SpMM) implementation for Ozaki method with cuBLAS is also evaluated.

Fig. 3 show that the performance with respect to varying matrix sparsity in N=10000.

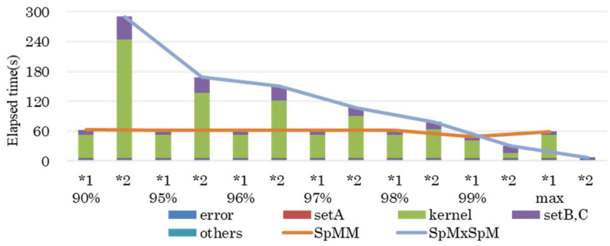

Fig. 3 Execution time between SpMM and SpMxSpM implementations for Ozaki method. X-axis is sparsity of input matrix. "*1" stands for SpMM implementation. "*2" stands for SpMxSpM implementation.

According to Fig. 3, whole execution time can be reduced up to 11.9% by utilizing SpMxSpM routine in N=10000.

See [8] for the details.

iii. Accuracy Assured Linear Equation Solver

(A) Iterative Refinement Procedure

We check real answer of large-scale linear equations for liner solver with residual iteration refinement by accurate dot product (pseud quadratic accuracy). This experiment is using 1750,000 dimensions for linear equations. 2500 nodes (80,000 cores) of the Fujitsu PRIMEHPC FX100 in Nagoya University is used.

The iterative refinement procedure is: (1) an approximate answer is obtained by using LU factorization; (2) A residual iterative refinement is performed.

The result is as follows:

(First Step) Residual Norm: 4.019007e-14
(Second Step) Residual Norm: 0.000000e+00

The above result indicates that the real answer is obtained with 2 step iterations. This also shows that the assured procedure we propose is a useful way for large-scale computations.

(B) Solving Linear Equations

We evaluate assured accuracy computation for solving linear equation. Given accuracy is improved by the iterative refinement procedure shown in (A).

We set a real answer with (1,1,1,…,1)^T. 2500 nodes (80,000 cores) of the Fujitsu PRIMEHPC FX100 in Nagoya University is used.

The result is:

(1 million dimension) Upper bound of error:
1.111484e-16
(1.5 million dimension) Upper bond of error:
1.113360e-16

The above result indicates that the obtained accuracy is almost full for double precision computation. Hence the accuracy assurance can be adaptable for very large-scale computations on distributed memory supercomputers.

- **Results for the Topic 2**

We made a proto type implementation of assured accuracy library for standard symmetric eigenproblem.

PDSYEVD (a ScaLAPACK routine) is used for this implementation. For test matrix, a symmetric matrix with elements generated by uniform distribution [0, 1].

The Fujitsu PRIMEHPC FX100 in Nagoya University is also used.

i. Performance Evaluation (Varying Nodes)

We set dimension of matrix to N=50,000. The performance of the prototype library is shown in Fig. 4.

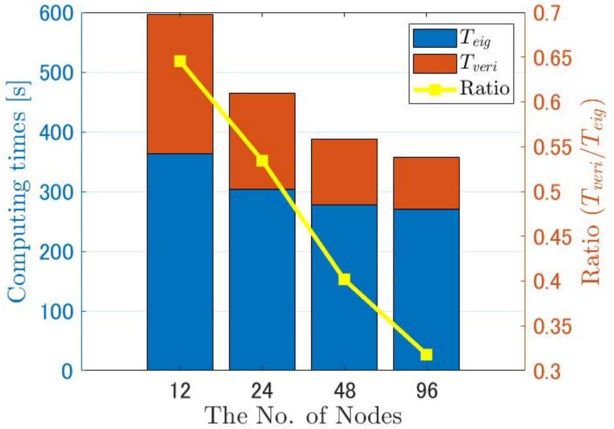

Fig. 4 Ratios of execution time (T_{veri} / T_{eig}). T_{veri} stands for verification time. T_{eig} stands for computation time of eigenvalues.

According to Fig. 4, there is a scalability for the ratio. This means that the ratios of verification time to computation time of eigenvalue are getting smaller according to number of nodes. This is a nice result to adapt the library of accuracy assurance to several applications.

ii. Performance Evaluation (Weak Scaling)

In next evaluation, we fix number of dimensions per node, while number of nodes increases. This is weak scaling evaluation.

Fig. 5 shows the result.

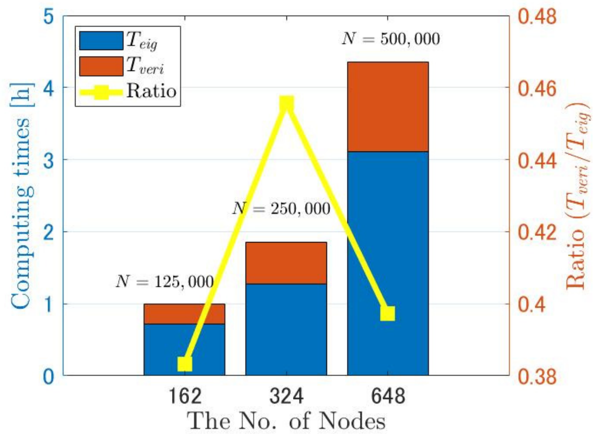

Fig. 5 Weak Scaling Result.

Fig. 5 shows that execution time for assured accuracy computation can be occupied up to 40%~50% to computation time of eigenvalues. This is acceptable ratio for large-scale computation.

iii. Performance Evaluation (Accuracy)

To do evaluation of computed accuracy, we set matrix dimension with N=500,000. By using PDSYEVD routine, we obtain λ_i : i-th eigenvalus from the smallest eigenvalue. We also calcuate r_i : upper error bound from assured accuracy compuation for λ_i : to evaluate computed accuracy.

The result is shown in Fig. 6.

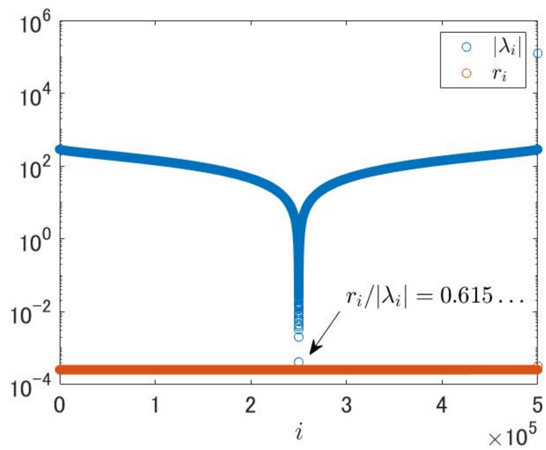

Fig. 6 Errors of computed eigenvalues to real answer.

Fig. 6 shows that upper bound of calculated error is 60% at the worst. This indicates that the calculated result is never included "duplicate eigenvalues" for the eigenproblem with dimension of 500,000.

We cannot proof this without the techniques for accuracy assurance for the eigenproblem. Hence this is a remarkable result in actual eigen computations.

- Results for the Topic 3

To do extension to non-linear problems, we study multilevel Schwarz preconditioned Newton-Krylov algorithm to solve the Poisson-Boltzmann equation with applications in multi-particle colloidal simulation.

The smoothed aggregation-type coarse mesh space is introduced in collaboration with the one-level Schwarz method as a composite preconditioner for accelerating the convergence of a Krylov subspace method for solving the Jacobian system at each Newton step.

The proposed smoothed aggregation multilevel Newton-Krylov-Schwarz (NKS) algorithm numerically outperforms than smoothed aggregation multigrid method.

See paper [1] for the details.

6. Progress during FY2019 and Future Prospects

All planned research topics during FY2019 are finished.

The followings are future prospects in FY2020.

- Topic 1: UNC-HPC libraries between multi-core and many-core CPUs and a GPU.

According to our results, we found several performance changes based on computer environments, such as CPU or GPU. In addition, sparsity of input matrix is also crucial factor.

We need to add adaptive selection for several implementations of Ozaki method. To establish this, auto-tuning (AT) technology is one of promising ways. Hence, we will adapt AT techniques to selection of the implementations for Ozaki method in FY2020.

- Topic 2: Designing accuracy assured libraries for real symmetric eigenproblem.

We need to develop high performance implementation of the accuracy assured libraries for real symmetric eigenproblem toward to distributed memory supercomputers. In particular, adaptation of GPU computing is highly required. Developing this implementation is important future work in FY2020.

7. List of Publications and Presentations

(1) Journal Papers (Refereed)

[1] S.-R. Cai, J.-Y. Xiao, Y.-C. Tseng, F.-N. Hwang, 'Parallel multilevel smoothed aggregation Schwarz preconditioned Newton-Krylov algorithms for Poisson-Boltzmann problems, Numerical Mathematics: Theory, Methods and Applications', Vol. 13, pp. 745-769, August 2020

(2) Proceedings of International Conferences (Refereed)

[2] Fumiya Ishiguro, *Takahiro Katagiri*, *Satoshi Ohshima*, *Toru Nagai*, 'Performance Evaluation of Accurate Matrix-matrix

Multiplications on GPU Using Sparse Matrix Multiplications', International Conference on High Performance Computing in Asia-Pacific Region (HPCAsia2020), January 2020 (A Poster Presentation)

(3) International conference Papers (Non-refereed)

[3] T. Ogita, 'Verification methods for numerical linear algebra and applications', The 9th International Congress on Industrial and Applied Mathematics (ICIAM 2019), July 2019

[4] T. Terao, K. Ozaki, T. Ogita, 'Verified Numerical Computations for Eigenvalue Problems on Large-Scale Parallel Systems', 2020 SIAM Conference on Parallel Processing for Scientific Computing, January 2020

[5] T. Ogita, 'Verified Numerical Computations with HPC', 3rd International Conference on Modern Mathematical Methods and High Performance Computing in Science & Technology, January, 2020 (An Invited Talk)

[6] T. Ogita, 'Verified Numerical Computations on Supercomputers', Workshop on Large-scale Parallel Numerical Computing Technology (LSPANC 2020), January 2020

(4) Presentations at domestic conference (Non-refereed)

[7] *片桐孝洋*，石黒史也，荻田武史，尾崎克久，*大島聡史*，*永井亨*，「精度保証付き数値計算ライブラリの運用に向けて」，大学 ICT 推進協議会 2019 年度年次大会，AXIES2019 予稿集，2019 年 12 月

(5) Other (patents, press releases, books and so on)

[8] Fumiya Ishiguro, 'A High-performance Implementation on GPU for Accurate Matrix-Matrix Multiplication Using Sparse Matrix Computations', A Master Thesis, Graduate School of Informatics, Nagoya University, February 2020 (In Japanese)