Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures
Final Report for JHPCN Joint Research of FY 2020

jh200008-NAH

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

(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.
- Dr. Mukunoki: Performance evaluation of accuracy assured libraries.
- Mr. Kitai and Mr. Yamamoto: Adaptation of auto-tuning.
- Mr. Terao and Mr. Uchino, Mr. Yamanashi, Mr. Sugiura, and Mr. Morishita: Performance evaluation of accuracy assured libraries

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;

```
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.

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).

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 FY2019

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”.

Results 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.

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.

According to the result, whole execution time can be reduced up to 11.9% by utilizing

SpMxSpM routine in N=10000.

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 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 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.

According to the result, 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.

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.

The result shows that execution time for assured accuracy computation can be occupied up to 40%~50% to computation time of eigenvalues.

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 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.

- 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.

5. Details of FY2020 Research Achievements

The topic in this year is shown as follows:

The Year 2 (FY2020):

1) Topic 1: Improvement of high-performance implementation for UNC-HPC libraries.
2) Topic 2: Prototyping accuracy assured libraries for real symmetric eigenproblem.
3) Topic 3: Discussing extension to non-linear problems based on The Year 1-Topic 3.
4) Topic 4: Discussing and performance evaluation of auto-tuning for the Topics 1 and 2.

- Results for the Topic 1

For the topic 1, we have installed for UNC-HPC libraries (**LINSYS_V**: Verified Solution of Linear Systems) to Oakforest-PACS.

In this year, a new machine at Nagoya University, which is the supercomputer "Flow", is available. To do the topic 1, we needed to check GPU performance for MMM with Ozaki method. We have rough performance of the MMM for GPU on the Supercomputer "Flow" TypeII subsystem.

The Table 1 shows the result.

Table 1 MMM with Ozaki Method in the Supercomputer "Flow" Type II subsystem. Size of Matrix is set to N=2000.

Implementation Kind	Time in second.
1 (dgemm)	129.9
2 (CRS SpMV(inner))	139.9
3 (CRS SpMV(outer))	144.4
4 (CRS SpMV (multiple inner))	129.4
5 (CRS SpMV (multiple inner with blocking))	128.7
6 (ELL SpMV(inner))	129.9
7 (ELL SpMV(outer))	130.3
8 (ELL SpMV (multiple inner))	130.4
9 (ELL SpMV (multiple inner with blocking))	130.3
10 (Batched BLAS)	136.2
11 (dgemm, GPU)	**88.3**
12 (CRS SpMV, GPU)	141.4
13 (ELL SpMV, GPU)	134.4
14 (CRS SpMM, GPU)	107.5
15 (CRS SpM-SpM, GPU)	481.7

We found that Ozaki Method with dgemm is the fastest on GPU environment (a board of the NVIDIA V100) in Table 1. However, there are many tunable parameters in the Ozaki's. In addition, number of matrix dimensions and distribution of element values for input

matrices affect total execution time.

Tuning of these parameters is one of typical topics for adapting auto-tuning. This is also target topic of the Topic 4. See the result of the Topic 4.

In addition, additional results of new implementations of GPU for MMM with Ozaki Method have been published in [2].

- Results for the Topic 2

In this year, we made a prototyping for accuracy assured libraries for real symmetric eigenproblem with Ogita-Aishima method. To do performance evaluation, we use two kinds of supercomputers, Supercomputer "Flow" (Type I subsystem) and Oakforest-PACS.

Table 2 shows the results.

Table 2 Evaluation result for accuracy assured libraries for real symmetric eigenproblem. (N=2^{14})

(a) Supercomputer "Flow" (Type I Subsystem).

$n = 2^{14}$, 4 Nodes	Pdsyevd (Double)	Pssyevd (Single)	Iterations to result of pssyevd by Ogita-Aishima method.		
			0th iter.	1st iter.	2nd iter.
time [s]	5.5e+01	3.2e+01	3.9e+01	4.6e+01	5.3e+01
$\max(\lvert D_i - \tilde{D}_i\rvert/\lvert D_i\rvert)$	8.9e-16	1.9e-06	2.7e-09	2.2e-09	5.9e-12
$\mathrm{median}(\lvert D_i - \tilde{D}_i\rvert/\lvert D_i\rvert)$	0.0e+00	3.1e-07	4.4e-11	3.3e-11	8.7e-16
$\lVert X - \tilde{X}\rVert/\lVert X\rVert$	2.6e-12	2.6e-03	2.5e-03	1.4e-05	5.9e-06

(b) Oakforest-PACKS.

$n = 2^{14}$, 64 Nodes	Pdsyevd (Double)	Pssyevd (Single)	Iterations to result of pssyevd by Ogita-Aishima method.		
			0th iter.	1st iter.	2nd iter.
time [s]	7.7e+01	6.5e+01	7.3e+01	8.1e+01	8.8e+01
$\max(\lvert D_i - \tilde{D}_i\rvert/\lvert D_i\rvert)$	8.9e-16	1.2e-06	7.8e-09	7.8e-09	6.5e-13
$\mathrm{median}(\lvert D_i - \tilde{D}_i\rvert/\lvert D_i\rvert)$	0.0e+00	1.7e-07	3.0e-11	2.0e-11	1.5e-16
$\lVert X - \tilde{X}\rVert/\lVert X\rVert$	2.6e-12	2.1e-03	2.0e-03	9.8e-06	3.7e-06

The results in Table 2 indicates that developed library with Ogita-Aishima method establishes speedup to pdsyevd routine (double precision) in LAPACK by using iteration.

In addition, the accuracy for the developed library is superior to pssyevd routine (single precision) in LAPACK. Hence, the developed library has a merit to speed and accuracy to conventional eigenvalue routines in LAPACK.

- Results for the Topic 3

A multilevel Schwarz preconditioned Newton-Krylov algorithm to solve the Poisson-Boltzmann equation with applications in multi-particle colloidal simulation is studied. 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 performance evaluation shows hat the proposed smoothed aggregation multilevel Newton-Krylov-Schwarz (NKS) algorithm numerically outperforms than smoothed aggregation multigrid method and one-level version of the NKS algorithm.

The Reedbush-U is used for the performance evaluation. See [1] for the details.

- Results for the Topic 4

In this year, an auto-tuning (AT) method is developed. This is for performance change of DHPMM_F with CPU and GPU. We have utilized for selection of 11 kinds of implementations for Ozaki method (DHPMM_F). (See Topic 1 in this report,) The one of results are summarized in Table 3.

In the AT in Table 3, we used a linear estimation for sparsity by measuring actual MMM in ozaki method. The result in Table 3

indicates that execution time can be estimated within maximum relative error 15.9% by the proposed AT mechanism. Hence we can establish basic AT mechanism in this year. See [4] for the details.

Table 3 Prediction of execution time in each implementation in Ozaki Method (DHPMM_F) on the Supercomputer "Flow". (N=2000)

N=2000	Prediction Time [s.]									
	Implementations									
Sparcity	1	2	3	4	5	6	7	8	9	11
90	0.195	3.239	0.338	0.308	0.630	2.212	0.334	0.292	0.699	0.377
92	0.244	2.591	0.258	0.239	0.458	1.803	0.259	0.231	0.513	0.471
94	0.195	1.994	0.202	0.187	0.352	1.394	0.204	0.183	0.396	0.376
96	0.244	1.419	0.131	0.127	0.197	1.032	0.137	0.130	0.228	0.470
98	0.196	0.750	0.065	0.066	0.074	0.575	0.074	0.075	0.092	0.376

	Relative Prediction Errors									
	Implementation									
Sparcity	1	2	3	4	5	6	7	8	9	11
90	-2.9%	-3.7%	0.6%	-9.9%	-11.0%	5.0%	-0.4%	-7.6%	-10.7%	-1.1%
92	-2.6%	-2.1%	2.5%	-9.5%	-6.2%	-6.9%	3.3%	-4.6%	-6.2%	-0.9%
94	-3.6%	-4.0%	2.9%	-10.0%	-6.0%	15.9%	4.8%	-7.3%	-10.2%	-1.0%
96	-2.9%	-3.7%	11.1%	-7.6%	3.3%	11.5%	7.2%	-5.1%	4.1%	-0.9%
98	-3.6%	-1.6%	11.2%	-1.6%	-1.7%	3.8%	1.1%	-3.6%	-0.3%	-0.9%

6. Progress during FY2020 and Future Prospects

In this year, all topics are satisfied. In the next year, we are planning the following topics.

i. Topic 1: Establishing high-performance implementation for UNC-HPC libraries based on the Year 2-Topic 1. (CPU and GPU)
ii. Topic 2: Developing accuracy assured libraries for real symmetric eigenproblem based on the Year 2-Topic 2. (CPU)
iii. Topic 3: Discussing extension to non-linear problems based on The Year 2-Topic 3. (CPU)
iv. Topic 4: Prototyping and developing AT based on the Year 2-Topics 1 and 2. (CPU and GPU)

7. List of Publications and Presentations

(1) Journal Papers (Refereed)

[1] S.-R. Cai(+), J.-Y. Xiao(+), Y.-C. Tseng(+), and 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, 2020.

(2) Proceedings of International Conferences (Refereed)

[2] F. Ishiguro, T. Katagiri, S. Ohshima, T. Nagai, Performance Evaluation of Accurate Matrix–Matrix Multiplication on GPU Using Sparse Matrix Multiplications, Workshop of The Eighth International Symposium on Computing and Networking (CANDARW2020), Proc. of CANDARW2020, (2021)
DOI: 10.1109/CANDARW51189.2020.00044

(3) International conference Papers (Non-refereed)

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

[3] 片桐孝洋, Developing Accuracy Assured High Performance Numerical Libraries for Eigenproblems, 第 12 回 自動チューニング技術の現状と応用に関するシンポジウム(ATTA2020), 2020 年 12 月.

[4] 青木将太, 片桐孝洋, 大島聡史, 永井亨, 高精度行列-行列積における疎行列演算実装選択の自動チューニングの検討, 情報処理学会第 82 回全国大会, 2021 年 3 月.

(5) Published library and relating data

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