Joint Usage/Research Center for Interdisciplinary Large-scale Information Infrastructures
Final Report for JHPCN Joint Research of FY 2021

jh210002-NAHI

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
- □ Very large-scale data processing
- □ Very large capacity network technology
- □ Very large-scale information systems

(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.
- Mr. Aoki: Adaptation of auto-tuning.
- Mr. Uchino: 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

```
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.

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).

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 FY2020

The topics in last year (FY2020) are 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.

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.

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

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.

- Results for the Topic 2

In the last year, we have 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.

- Results for the Topic 4

In the last 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.

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%

5. Details of FY2021 Research Achievements

The topics for the final year are listed as

follows:

- Year 3 (FY2021) Plan:
  - Topic 1: Establishing high-performance implementation for VNC-HPC libraries based on the Year 2-Topic 1. (CPU and GPU)
  - Topic 2: Developing accuracy assured libraries for real symmetric eigenproblem based on the Year 2-Topic 2. (CPU)
  - Topic 3: Discussing extension to non-linear problems based on The Year 2-Topic 3. (CPU)
  - Topic 4: Prototyping and developing AT based on the Year 2-Topics 1 and 2. (CPU and GPU)

- Results for the Topic 1

In the last year, basic construction of VNC-HPC library have been finalized. In particular, GPU implantation of MMM library for Ozaki method, and eigensolver with Ogita-Aishima method have been remarkable improved. Hence additional development is not required in this year. See final reports in FY2020 in this project.

- Results for the Topic 2

In the last year, we have made an improved implementation of assured accuracy library for solvers for linear equations and eigenproblem. In this year, we have evaluated accurate assurance computation with changing round-off modes.

We utilize the Supercomputer "Flow" Type I Subsystem, which is installed at Information Technology Center, Nagoya University. Number of MPI process per node is 4. Number of threads per process is 12.

The implementation is used for a round-off control library for the FX1000 from Fujitsu Ltd. This enables us to use very fast control for changing round-off modes. Hence the performance is very improved by using the library.

Table 4 shows summary of execution time.

Table 4 Evaluation result for accuracy assured libraries with round-off control library on the Supercomputer "Flow" (Type I Subsystem).

(a) Linear Equations Solver

Dimension	Number of Nodes	Execution Time for Approximate Answer [Sec.]	Execution Time for Accurate Assured Answer [Sec.]	Ratio
40000	4	14	67	4.7
80000	16	33	143	4.3
160000	64	81	309	3.7
320000	256	217	673	3.1

(b) Eigensolver

Dimension	Number of Nodes	Execution Time for Approximate Answer [Sec.]	Execution Time for Accurate Assured Answer [Sec.]	Ratio
40000	4	319	129	0.40
80000	16	675	283	0.41
160000	100	827	522	0.63
320000	400	1993	1324	0.66

(A) Linear Equations Solver

For the approximate answer, we used pdgesv routine in ScaLAPACK. For computational complexities, $2/3n^3$ for approximate answer, and $6n^3$ for accurate assured answer in this case. Hence theoretical ratio is 9.

According to Table 4 (a), the ratios between approximate and assured execution are from 3.1 to 4.7. With respect to the theoretical ratio (9x), our solver can provide excellent efficiency. This comes from algorism based on DGEMM operation.

(B) Eigenproblem Solver (For Real Symmetric Matrices)

For the approximate answer, we used pdsyevd routine in ScaLAPACK.

According to Table 4 (b), the ratios between approximate and assured execution are from 0.40 to 0.66. With respect to execution time for approximate answer, our solver can provide enough high execution time.

- Results for the Topic 3

Fully coupled space-time solution algorithms for the time-dependent PDEs obtained their popularity recently for temporal domain parallelism. The space-time algorithm requires to solve the resulting large, space, nonlinear systems in an all-at-once manner. A robust and efficient nonlinear solver plays an essential role as a critical kernel of the whole solution algorithm.

We study some nonlinear preconditioned Newton algorithms for the space-time formulation of the hyperbolic equation with shock presented. The history of the nonlinear residual norm for the classical inexact Newton method with backtracking (INB) suffers from a long stagnation due to strong local nonlinearity. Nonlinear preconditioning such as nonlinear elimination has been shown as a practical technique to improve the robustness of INB for many different types of PDEs but does not work well for hyperbolic PDEs.

We have proposed a new variant of nonlinear elimination preconditioners designed for hyperbolic PDEs by taking their characteristics into account to overcome the difficulties.

A comparative performance study of two nonlinear preconditioned iterative algorithms, where nonlinear elimination techniques as either right or left nonlinear preconditioning, in conjunction with inexact Newton algorithms, namely INB-ANE and NEPIN, respectively, was performed. (For the details, see [1].)

- Results for the Topic 4

(A) Adapt Explainable AI (XAI) to Prediction of the best implementation on the Ozaki Method:

In AT, machine learning (ML) should be considered. We are trying to adapt and evaluate ML technology to establish the AT in recent years.

Currently, expandability of results from ML is crucial issues for AI filed. This concept is called Explainable AT (XAI). In the final year, we have evaluated adaptability of XAI to AT for VNC-HPC library to establish its AT function.

The target in this section is a selection of the best implementation for computations of high accurate MMM, named DHPMM_F for GPU: High-precision Matrix Multiplication with Faithful Round, which is one of composed routines of VNC-HPC library. In the DHPMM_F, there are 11 implementations between CPU and GPU (See Table 1). We have adapted a ML for the selection of the best implementation.

We utilized a random-forest algorithm on the scikit-learn ver. 0.24.1. Explanatory variables are 7 kinds in this experiment as follows: 1) matrix size; 2) sparsity of input matrix; 3) maximum element of A; 4) minimal element of A; 5) number of sparse matrices for error-free transformation of A; 6) number of dense matrices for error-free transformation of A; 7) number of matrices for error-free transformation of B; Number of learning data is 199. Number of test date is 23. Accuracy of learned model is 91.3% in this case.

We used LIME ver. 0.2.0.1, and SHAP ver.0.39.0 for XAI tools. SHAP result with a limited case is explained in the final report. Fig. 1 shows a SHAP result for Sparse Matrix-Matrix

(SpMM) implementation with CRS format on GPU for DHPMM_F.

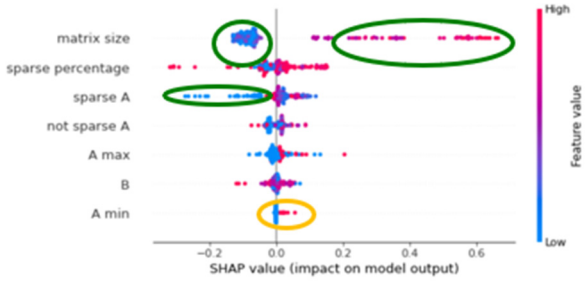

Fig 2. SHAP result for SpMM with CRS format on GPU.

As result of Fig 2., SHAP can explain that the followings are a factor to the learned model: 1) Large matrix size; 2) number of sparse matrices for error-free transformation of A; This is reasonable explanation of DHPMM_F routine because SpMM is an implementation of sparse matrix operations for matrices of error-free transformation. In addition, this implementation is crucial if the matrix size is large.

On the other hand, SHAP also explains minimal element of A does not affect the model. This is also reasonable, since minimal element is zero if the decomposed matrix is sparse. This also implies the variable can remove the model.

See [1][3]-[6] for the details in the topic.

(B) Adaptation of XAI to Adjustment of Block Length in Ozaki Method

Not limited to Ozaki method, many numerical libraries have a tunable parameter for cache blocking to obtain maximum performance. In this session, we try to adapt XAI technology to explanation of AT for blocking of caches.

The target library is as same as previous section, that is DHPMM_F. We focus on tuning parameter of cache blocking on DHPMM_F. The target process is a sparse Matrix-vector multiplication (SpMV) operation with multiple right-hand-side. Please note that the number of simultaneous rows of the vector is a factor of cache blocking in the process.

For the ML in the AT, we also use xgboost. For explainable variables in this experiment, we take the followings: 1) Specify of matrix A (absolute from 90% to 99%.); 2) Maximum element of matrix A; 3) Minimum element of matrix A; 4) Number of factorized matrices in matrix A over 90% in sparsity; 5) Number of factorized matrices in matrix A less than 90% in sparsity; 6) Number of factorized matrices in matrix B; 7) **Block length**.

For learning data, we utilize minimum execution time for MMM part in 50 times execution on Ozaki method.

Fig. 3 shows SHAP values of predicted time by AI model with matrix size is fixed as 1500. All values are constant, except for sparsity, block length, number of factorized matrices of matrix B. The number of data is 384, maximum relative error in the prediction is 1.2% in this case.

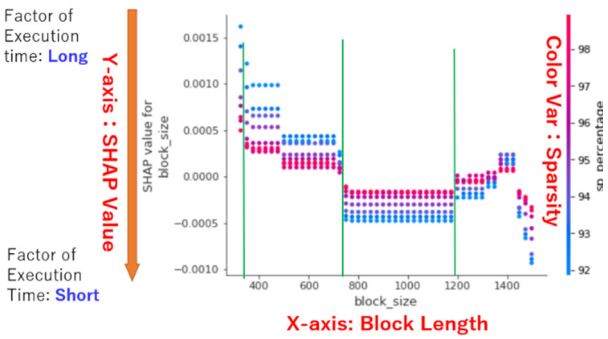

Fig 3. SHAP result for varying block lengths.

According to Fig. 3, effect of cache is observed; around 300, 700, and 1200. We think that this is caused by cache usability changing loop count according to block length to the matrix size 1500. Hence the explanation of SHAP seems to be reasonable. Hence, this is the case that XAI can be adaptable for typical tuning process on

numerical library.

Through experiments of XAI, we have showed establishing AI with AT for Ozaki method. Hence, we think that the main topic on this year is successfully performed.

6. Progress during FY2021 and Future Prospects

As explained as topic 4 in the previous sections, we have showed that: Adaptation of XAI tools is useful even in numerical library.

This implies that XAI is also useful for AT to establish high accuracy and to shorten AT time. We think that developing AT method based on XAI will be more critical technology in the future.

7. List of Publications and Presentations

(1) Journal Papers (Refereed)

(2) Proceedings of International Conferences (Refereed)

[1] Shota Aoki, Takahiro Katagiri, Satoshi Ohshima, Toru Nagai, "Feature analysis for selection of implementations in an accurate matrix-matrix multiplication library", The International Conference on High Performance Computing in Asia-Pacific Region (HPC Asia 2022), Jan 12-14, 2022 (A Poster)

(3) International conference Papers (Non-refereed)

[2] Chang-Wen Liang, Feng-Nan Hwang (+), “Nonlinear Elimination Preconditioned Inexact Newton Algorithms for a Full Space-time Formulation of Hyperbolic Equations”, 2022 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT2022), March 29-30, Tainan, Taiwan

[3] Takahiro Katagiri, "Adaptation of XAI to Auto-tuning of Numerical Libraries”, 2022 Conference on Advanced Topics and Auto Tuning in High-Performance Scientific Computing (ATAT2022), March 29-30, Tainan, Taiwan

[4] Takahiro Katagiri, Shota Aoki, Satoshi Ohshima, Toru Nagai, "Adaptation of Explainable AI for Auto-Tuning on Accurate Matrix Multiplication Library", SIAM Conference on Parallel Processing for Scientific Computing (PP22) (2022)

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

[5] 片桐孝洋, 青木将太, 大島聡史, 永井亨, 「高精度行列−行列積ライブラリの実装選択パラメタの特徴量解析」, [正会員主催 OS], [正会員 OS] 先進的環境における数値計算と関連 HPC 技術(1), 日本応用数理学会 2021 年度年会 (2021)

[6] 片桐孝洋,「固有値計算のための高性能精度保証ライブラリの開発：最新成果と自動チューニング機能」, 第 13 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 (ATTA2021) (2021)

(5) Published library and relating data

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