High Performance Accuracy Assurance Toward to Supercomputer Fugaku

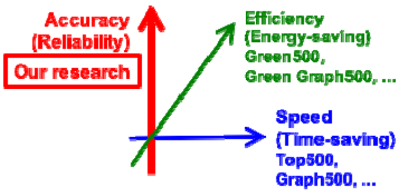
Takahiro Katagiri (Information Technology Center, Nagoya University, E-mail: katagiri@cc.nagoya-u.ac.jp)
Takeshi Ogita (Tokyo Woman’s Christian University, E-mail: ogita@lab.twcu.ac.jp)
Katsuhisa Ozaki (Shibaura Intitule of Technology, E-mail:ozaki@sic.shibaura-it.ac.jp)

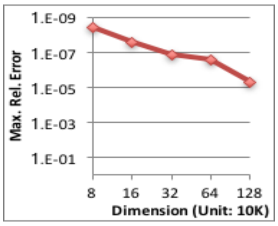
Development of Verified Numerical Computations for Applications

Introduction

The purpose of this research is to develop a super high-performance computing environment that can solve various challenging problems caused by numerical errors. Concretely, we aim to introduce "the axis of accuracy" into high-performance computing on post-K computer. For this purpose, we establish a new standard for performance evaluation by the use of fast verified numerical computations, accurate algorithms with error-free transformations, and auto-tuning methods.

Performance = Speed * Accuracy





Loss of accuracy

Numerical Targets

A. Scalability: Obtain computed results with desired accuracy for up to one million dimensional problems such as linear systems and eigenvalue problems with dense matrices.

B. Adaptivity: Obtain error bounds of computed results within several to several tens times slower than standard numerical methods, depending on the difficulty of problems.

C. Application: Show effectiveness of verified numerical computations in practical applications in terms of scientific and social significance.

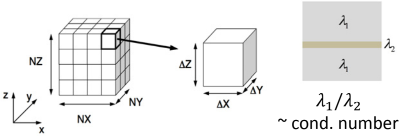
An application in progress (joint work with K. Nakajima, Univ. Tokyo, supported by JHPCN)

3-D Poisson equation:

$$-\nabla \cdot (\lambda \nabla \phi) = f, \quad f(x,y,z) = x + y + z$$

(Dirichlet boundary conditions)

FVM discretization
FDM mesh (7pt. stencil)
Solver: ICCG



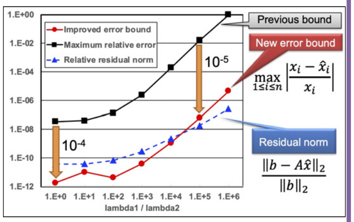
λ_1/λ_2
~ cond. number

Computing time (sec.) on Reedbush-U (1 node)

NX=NY=NZ=128 (n = 2,097,152)	Approx.	Verification
$\lambda_1/\lambda_2 = 1$	3.75	4.47
$\lambda_1/\lambda_2 = 10^6$	6.83	7.85

Cost for verification is comparable!

Error bounds and residual norm



Error bounds are significantly improved!

Research Organization of Our Post-K Project

	Organization	Representative	Role
Leader	Tokyo Woman’s Christian Univ.	OGITA, Takeshi	Project leader, Development of algorithms for accurate numerical computations
Partners	Waseda Univ.	KASHIWAGI, Masahide	Development of algorithms for verified numerical computations
	Nagoya Univ.	KATAGIRI, Takahiro	Development of benchmark methods and implementation
	Shibaura Inst. Tech.	OZAKI, Katsuhisa	Development of fast and accurate matrix multiplication methods

Accurate Matrix-Matrix Multiplications on GPU

Introduction

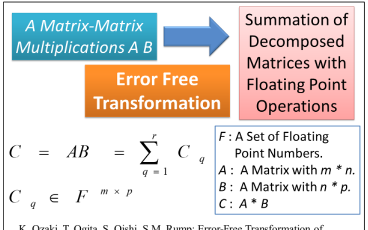
Libraries for basic linear algebra operations, such as BLAS (Basic Linear Algebra Subprograms), are one of crucial tools for numerical computations. However, accuracy assurance for numerical linear algebra libraries, such as LAPACK, is still under research.

We focus on the research of accuracy assurance, in particular, high precision matrix-matrix multiplication (MMM).

Ozaki Method

By using error free transformation, Ozaki method can establish high precision for MMM with extremely dispersed elements.

If dispersion of elements of matrix is large, then sparse matrix computation can be utilized.


A Matrix-Matrix Multiplications A B

Error Free Transformation

Summation of Decomposed Matrices with Floating Point Operations

$$C = AB = \sum_{q=1}^{r} C_q$$

$$C_q \in F^{m \times p}$$

F : A Set of Floating Point Numbers.
A : A Matrix with *m* * *n*.
B : A Matrix with *n* * *p*.
C : *A* * *B*

K. Ozaki, T. Ogita, S. Oishi, S.M. Rump: Error-Free Transformation of Matrix Multiplication by Using Fast Routines of Matrix Multiplication and its Applications, Numerical Algorithms, Vol. 59, No.1, pp.95-118, 2012.

Numerical Experiment

- **Target Matrix**: *A*: Identity Matrix + random sparse matrix with 90% density; *B*: Inverse matrix of *A*;
- **Sizes of Target Matrix**: 50, 100, 500, 1000, 5000, and 10000.
- **Machine Environment**: The Reedbush-H system at ITC, U. Tokyo.
  CPU: Intel Xeon E5-2695v4 (Broadwell-EP), 2 Sockets (36 Cores)
  GPU: NVIDIA Tesla P100 (Pascal)
- **Sparse Format**: CSR, cuSPARSE with CUDA 8.0.44 is used.

Performance Result

Accuracy Comparison

Sizes	10	50	100	500	1000	5000
DGEMM (Absolute Error)	8.77e-17	8.18e-16	1.70e-15	2.94e-15	5.33e-15	5.14e-15
DGEMM (Relative Error)	2.11e-10	1.82e-8	7.04e-7	2.71e-5	1.92e-5	3.42
Ozaki (Absolute Error)	8.77e-17	1.11e-16	1.10e-16	1.11e-16	1.11e-16	1.11e-16
Ozaki (Relative Error)	8.77e-17	1.11e-16	1.10e-16	1.11e-16	1.11e-16	1.11e-16

Speedup with cusparseDcsrmv and cusparseDcsrmm on GPU

- GPU execution is faster than CPU execution when size of matrices is larger than N=1000.
- **cusparseDcsrmm** can reduce execution time with maximum factor of 22.4% to execution with **cusparseDcsrmv**.

Information Technology Center, Nagoya University
http://www.icts.nagoya-u.ac.jp/en/center/