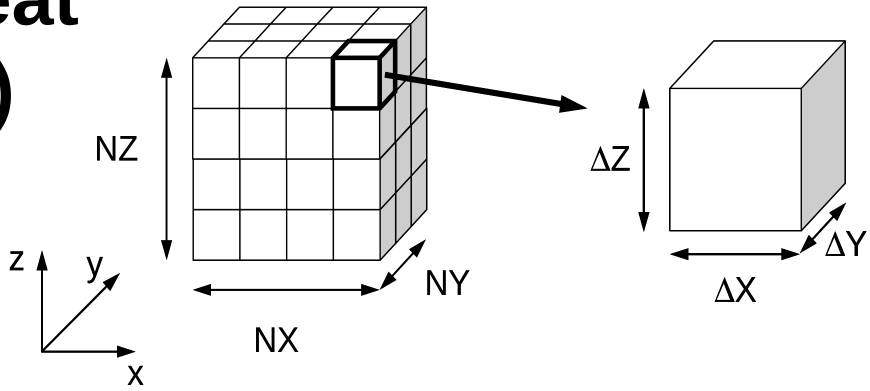
Accuracy Verification of Sparse Linear Solvers with FP64/FP32 Arithmetic

Kengo Nakajima
Information Technology Center
The University of Tokyo

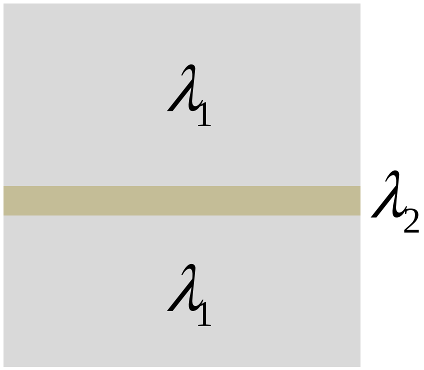
Approximate Computing with Low/Adaptive/Trans Precision

- Mostly, scientific computing has been conducted using FP64 (double precision, DP)
  - Sometimes, problems can be solved by FP32 (single precision, SP) or lower precision
- **Lower precision may save time, energy and memory**
- Approximate Computing
  - Originally for image recognition etc. where accuracy is not necessarily required
  - Also applied to numerical computations
- Computations by lower precision and by mixed precision may provide results with less accuracy

P3D: Steady State 3D Heat Conduction by FVM (1/2)



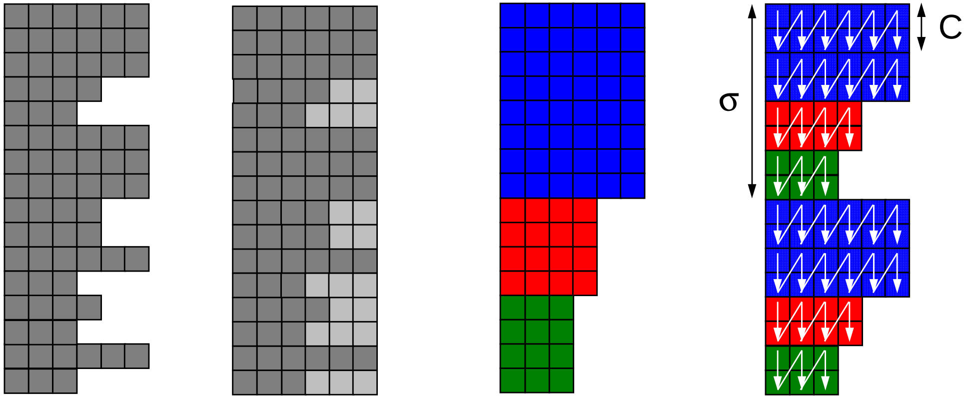
- 7-point Stencil
- Heterogenous Material Property
  - λ_1/λ_2 is proportional to the condition number of coefficient matrices
- Coefficient Matrix
  - Sparse, SPD
- ICCG Solver
- Fortran 90 + OpenMP
- CM-RCM Reordering



$$\nabla \cdot (\lambda \nabla \phi) + f = 0$$

<u>P3D</u>: Steady State 3D Heat Conduction by FVM (2/2)

- Various Configurations
  - FP64 (Double), FP32 (Single), FP16 (Half) (just for preconditioning)
  - Matrix Storage Format (CRS, ELL, SELL-C-σ etc.)



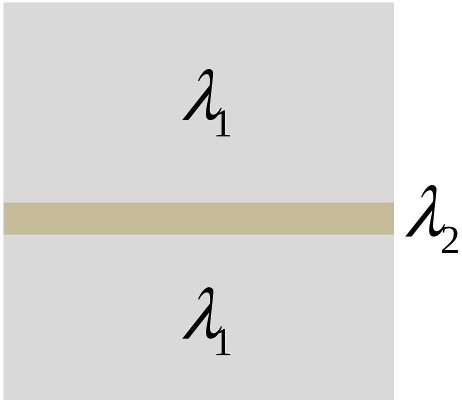
CRS ELL Sliced ELL SELL-C-σ

Ratio of FP32(SP)/FP64(DP)

Iterations● & Time△ for ICCG

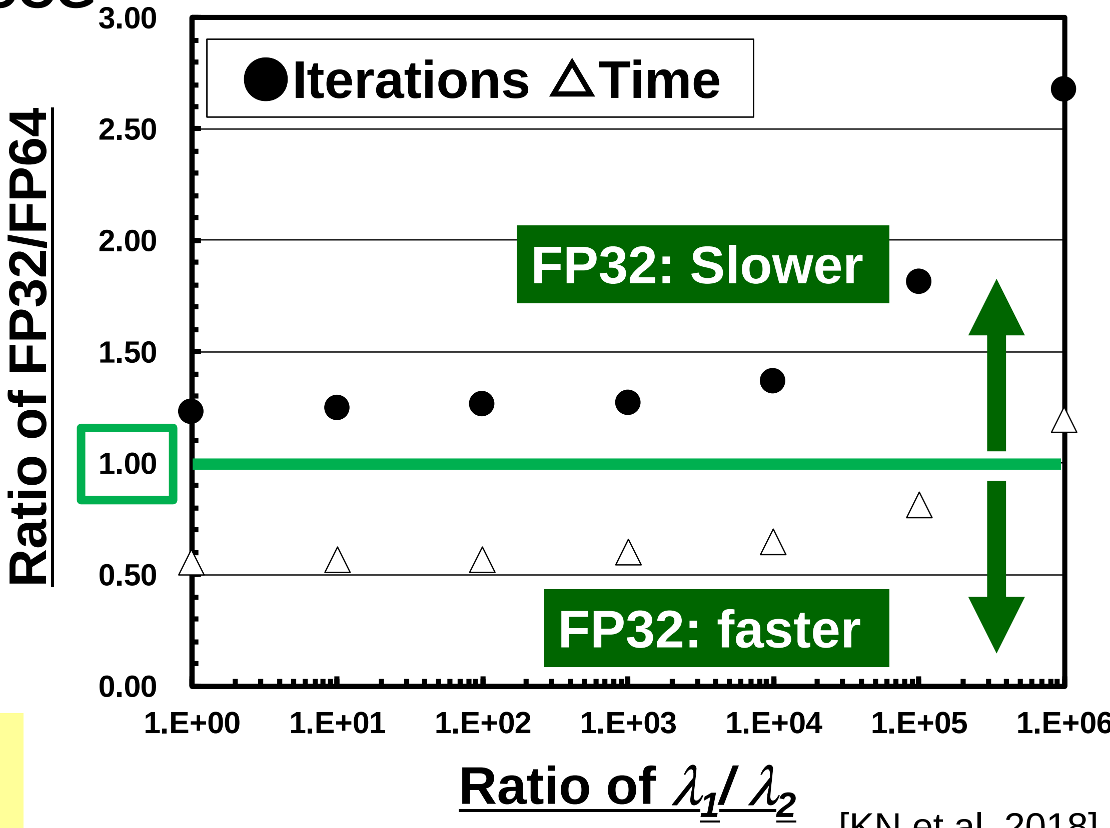
λ_1/λ_2, 128^3 DOF, CRS

Ratio<1 ⇒ FP32 is faster



$$\nabla \cdot (\lambda \nabla \phi) + f = 0$$

Intel Xeon BDW
1 Node: 18 cores x 2 soc's



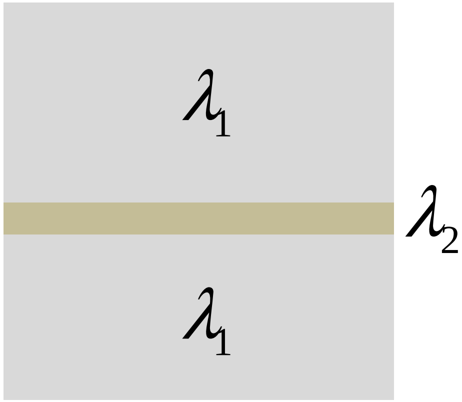
[KN et al. 2018]

Ratio of FP32(SP)/FP64(DP)

Iterations● & Time△ for ICCG

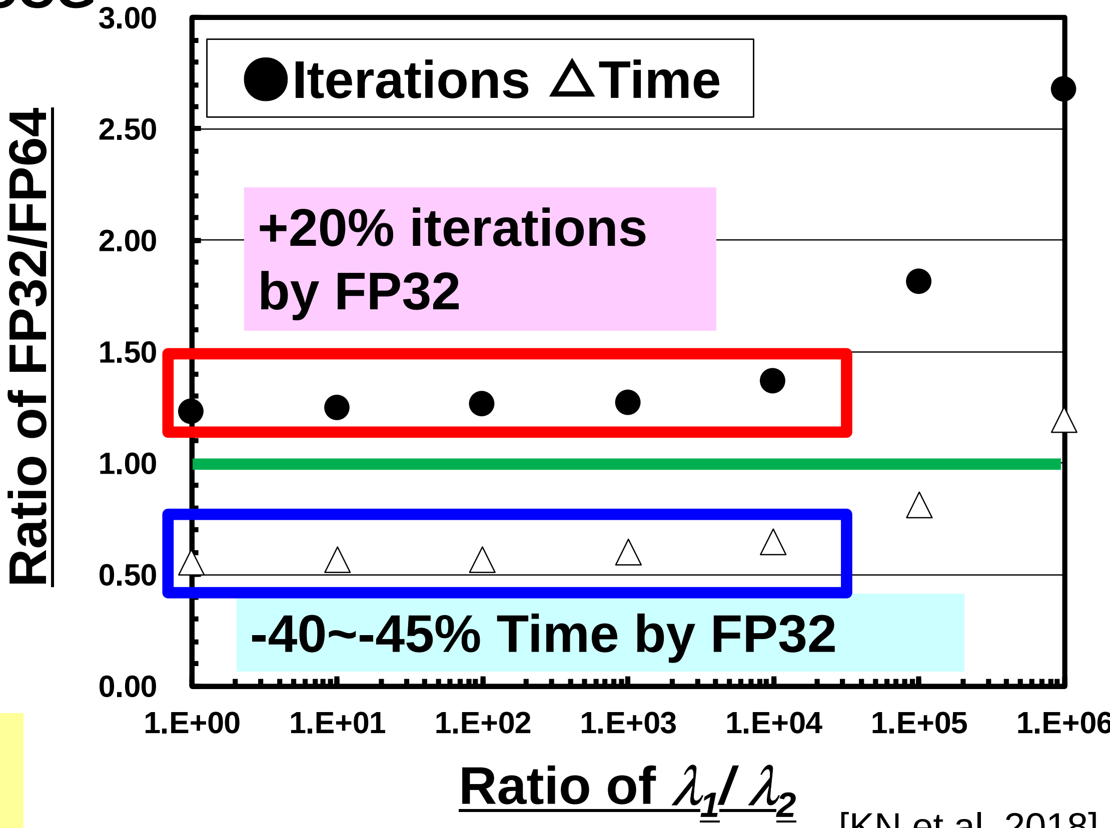
λ_1/λ_2, 128^3 DOF, CRS

Ratio<1 ⇒ FP32 is faster



$$\nabla \cdot (\lambda \nabla \phi) + f = 0$$

Intel Xeon BDW
1 Node: 18 cores x 2 soc's



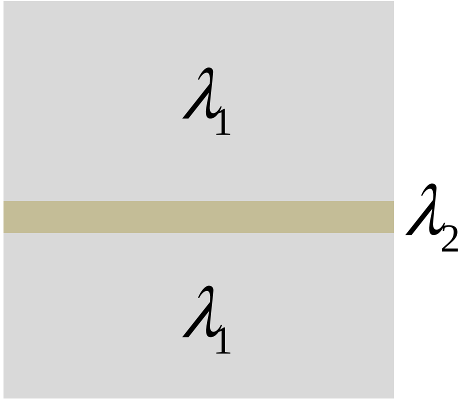
[KN et al. 2018]

Ratio of FP32(SP)/FP64(DP)

Iterations● & Time△ for ICCG

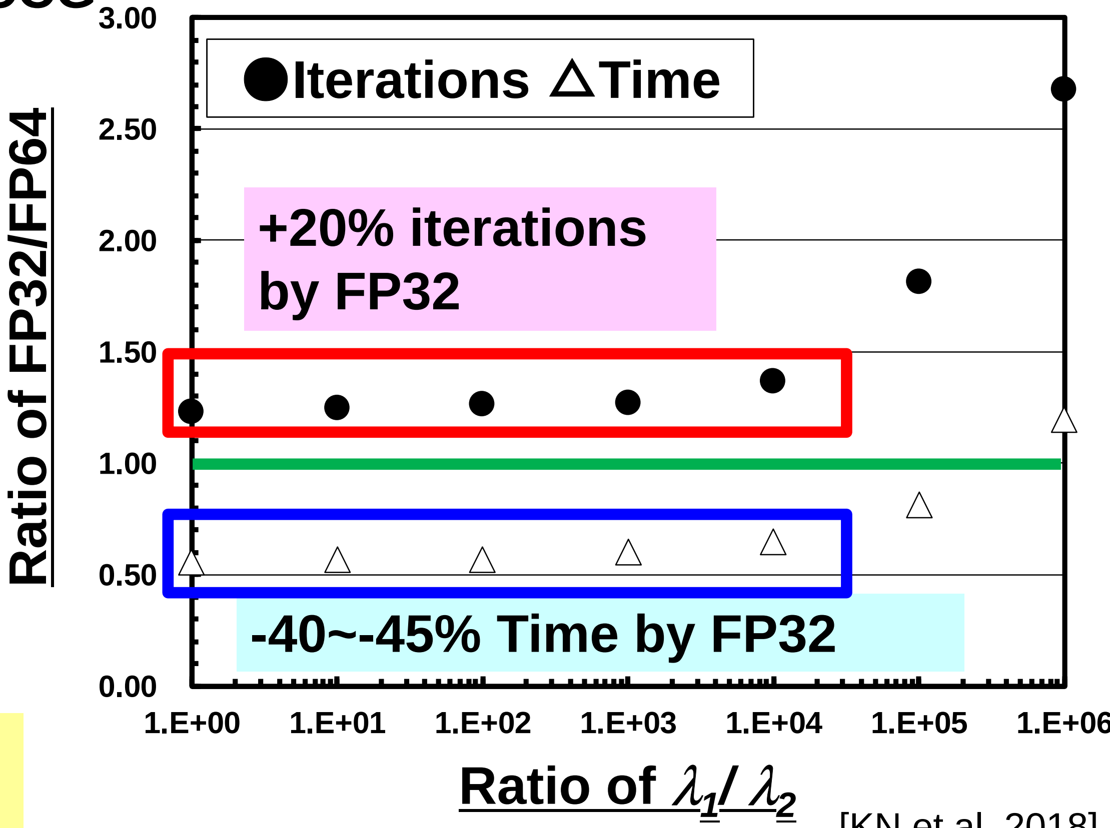
λ_1/λ_2, 128^3 DOF, CRS

Ratio<1 ⇒ FP32 is faster



$$\nabla \cdot (\lambda \nabla \phi) + f = 0$$

Intel Xeon BDW
1 Node: 18 cores x 2 soc's



[KN et al. 2018]

Approximate Computing with Low/Adaptive/Trans Precision

- Accuracy verification is important
  - Iterative Refinement
- A lot of methods for accuracy verification have been developed for problems with dense matrices
  - But very few examples for sparse matrices & H-matrices
- Generally speaking, processes for accuracy verification is very expensive
  - Sophisticated Method needed
  - Automatic Selection of Optimum Precision by Technology of AT (Auto Tuning)
- Accuracy Verification of Sparse Linear Solvers [Ogita, Nakajima 2019]

Original Algorithm for Verification

[Ogita, Oishi, Ushiro 2001]

1. Solve a discretized linear system $Ax = b$.
   - $\hat{x}$: a computed solution
2. Solve a linear system $Ay = e$ where all elements of e are 1's.
   - $\hat{y}$: a computed solution
3. Verify M-property of A using $\hat{y}$. ($\hat{y} > 0 \;\Rightarrow\; A\hat{y} > 0$)
4. Compute an error bound using

$$\|x - \hat{x}\|_\infty \leq \frac{\|\hat{y}\|_\infty \|b - A\hat{x}\|_\infty}{1 - \|e - A\hat{y}\|_\infty}$$

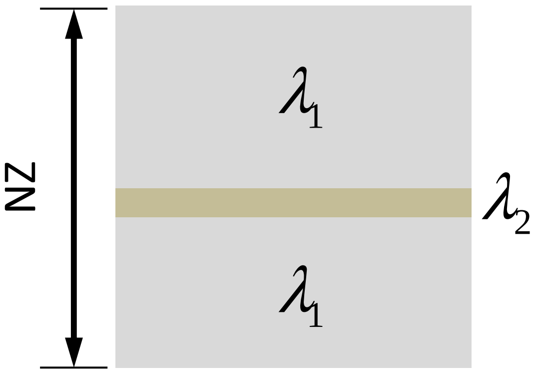
if $\|e - A\hat{y}\|_\infty < 1$.

$$\|A^{-1}\|_\infty \leq \frac{\|\hat{y}\|_\infty}{1 - \|e - A\hat{y}\|_\infty}$$

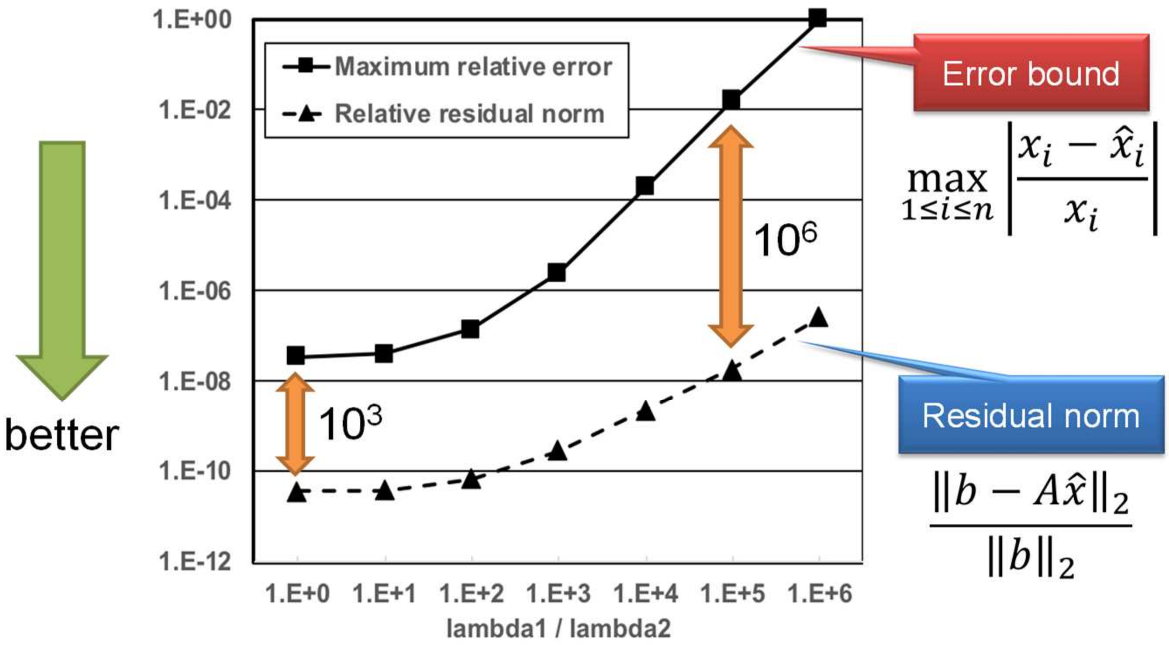
Results: Original Algorithm for Verification (128^3)

[Ogita, Oishi, Ushiro 2001]

$$\phi = 0 \,@\, z = z_{\max}$$



$$\nabla \cdot (\lambda \nabla \phi) + f = 0$$



It is difficult to estimate the error of a computed solution only from residual norm!

Improved Error Bound

[Ogita, Oishi, Ushiro 2002]

- To reduce the overestimation, we replace

$$\|x-\hat{x}\|_\infty \le \frac{\|\hat{y}\|_\infty \|b-A\hat{x}\|_\infty}{1-\|e-A\hat{y}\|_\infty}$$

by

$$\|x-\hat{x}\|_\infty \le \|\hat{z}\|_\infty + \frac{\|\hat{y}\|_\infty \|b-A(\hat{x}+\hat{z})\|_\infty}{1-\|e-A\hat{y}\|_\infty}.$$

$$x-\hat{x} = A^{-1}(b-A\hat{x}) = \hat{z} + A^{-1}(b-A(\hat{x}+\hat{z}))$$

- The correction term $\hat{z}$ is obtained by solving a linear system $Az = r$ with $r = b - A\hat{x}$.

Improved Algorithm for Verification

[Ogita, Rump, Oishi 2005] [Ogita, Nakajima 2019]

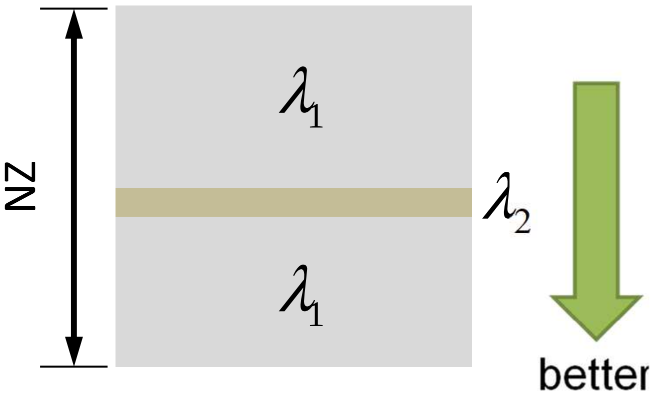
1. Solve a discretized linear system $Ax = b$.
2. Solve a linear system $Ay = e$.
3. Verify M-property of A using $\hat{y}$. $(\hat{y} > 0 \Rightarrow A\hat{y} > 0)$
4. Compute $r = b - A\hat{x}$ with an error bound.
   - $\hat{r}$: a computed residual, e_r: an error bound of $\hat{r}$
5. Solve a linear system $Az = \hat{r}$.
6. Compute an error bound using

$$\|x - \hat{x}\|_\infty \leq \|\hat{z}\|_\infty + \frac{\|\hat{y}\|_\infty(\|\hat{r} - A\hat{z}\|_\infty + \|e_r\|_\infty)}{1 - \|e - A\hat{y}\|_\infty}.$$

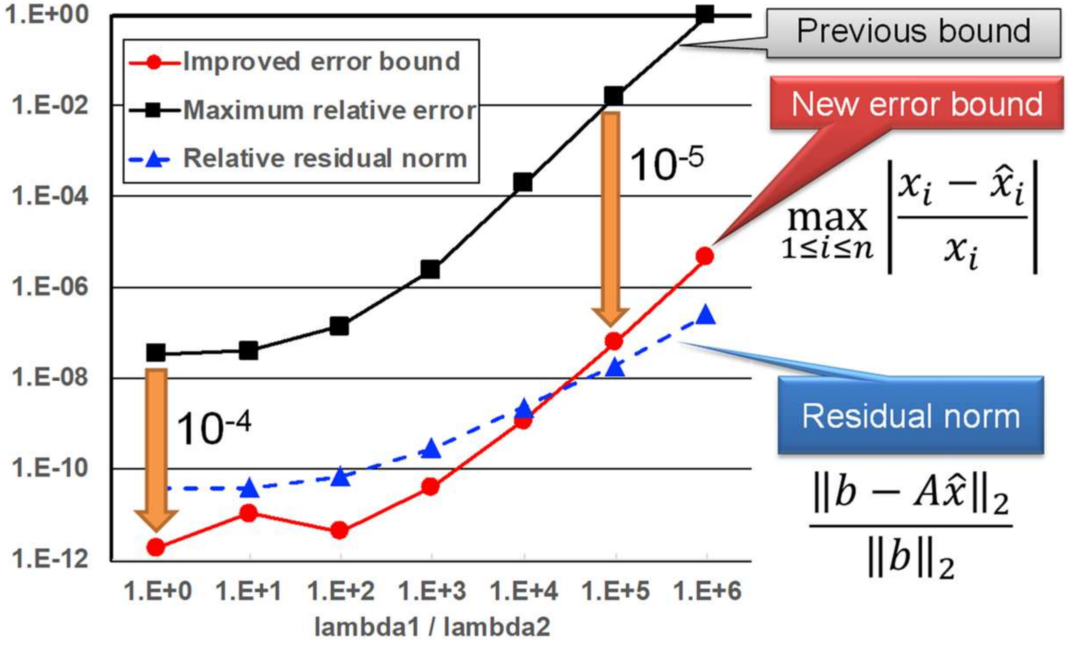
Results: Improved Alg. for Verification (128^3)

[Ogita, Rump, Oishi 2005] [Ogita, Nakajima 2019]

$\phi = 0 \,@\, z = z_{max}$



$$\nabla \cdot (\lambda \nabla \phi) + f = 0$$



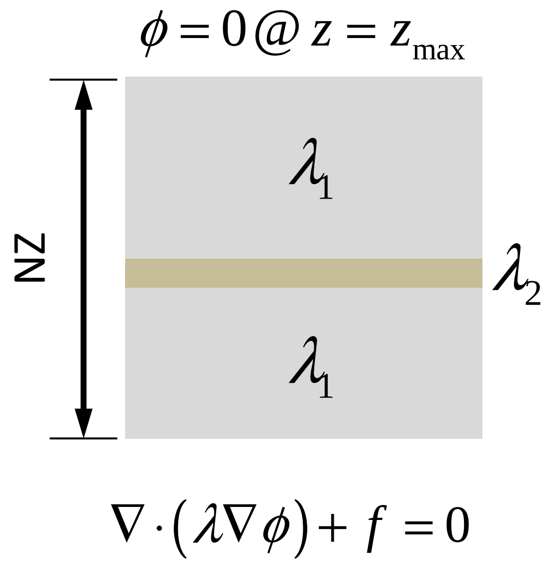
Computed error bounds are significantly improved!

Further Investigations

[Nakajima et al. SWoPP 2020]

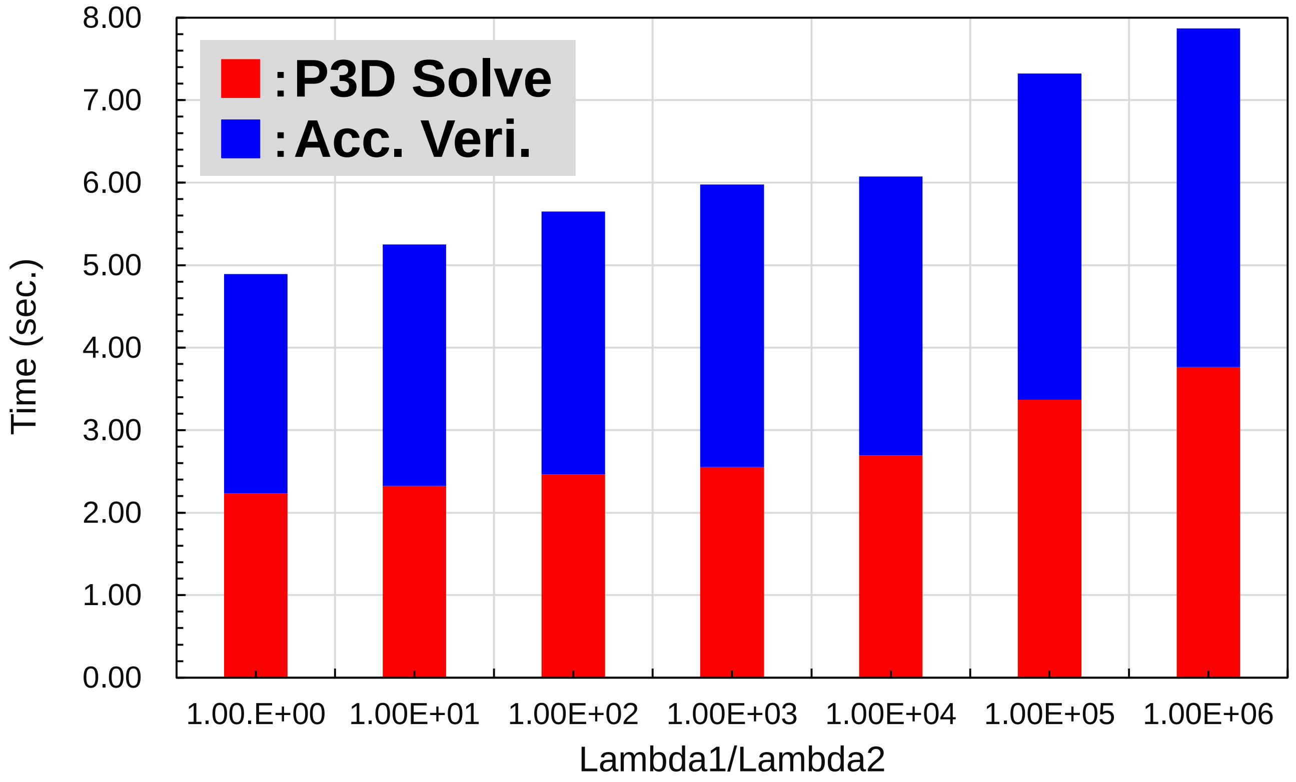
- λ_1 / λ_2 =10^0~10^6, CRS, N=128^3
- Accuracy Verification with FP32

	P3D Solve	Accuracy Verification
D-D	FP64	FP64
D-S	FP64	FP32
S-S	FP32	FP32



Results on OBCX (Intel Xeon CXL) (1/2)

D-D, Accuracy Verification takes 10% longer



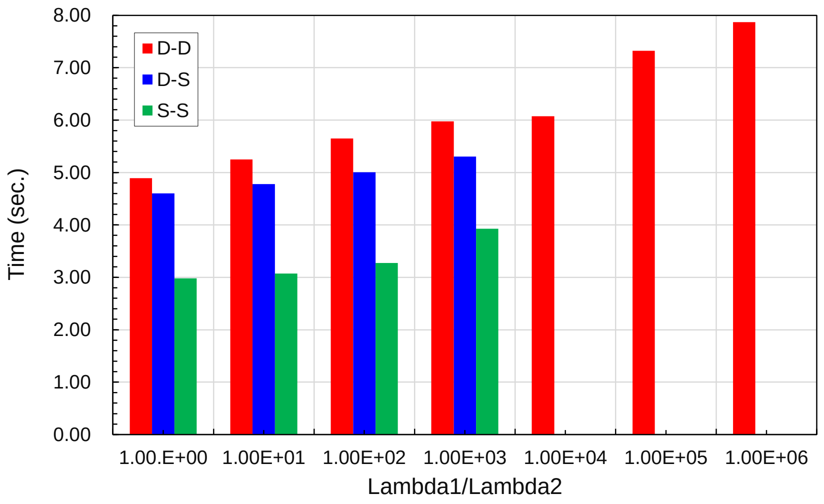
Results on OBCX (2/2)

Accuracy verification for D-S and S-S has failed, if $\lambda_1 / \lambda_2 \geqq 10^4$

$$\left\| x-\hat{x} \right\|_\infty \le \left\| \hat{z} \right\|_\infty + \frac{\left\| \hat{y} \right\|_\infty \left\| b-A(\hat{x}+\hat{z}) \right\|_\infty}{1-\left\| e-A\hat{y} \right\|_\infty}$$

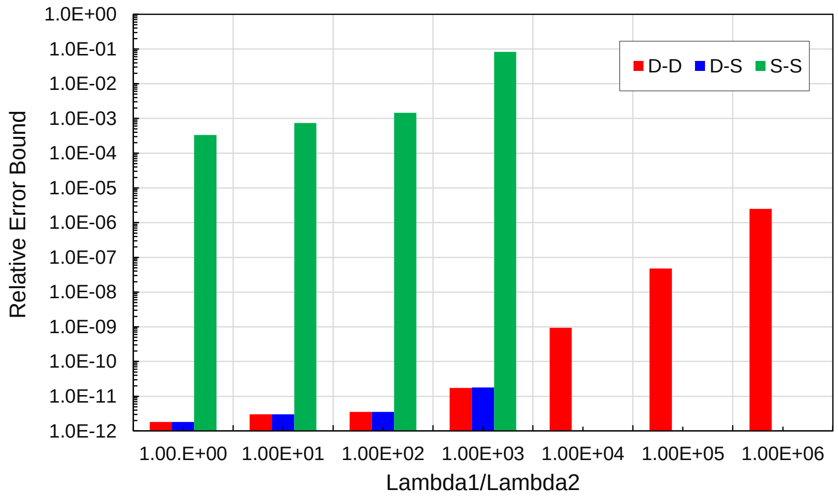
Total Computation Time

■:D-D, ■:D-S, ■:S-S



Max. Relative Error Bound

■:D-D, ■:D-S, ■:S-S

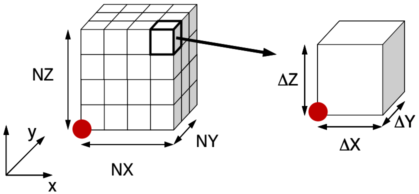


Results on OBCX (2/2)

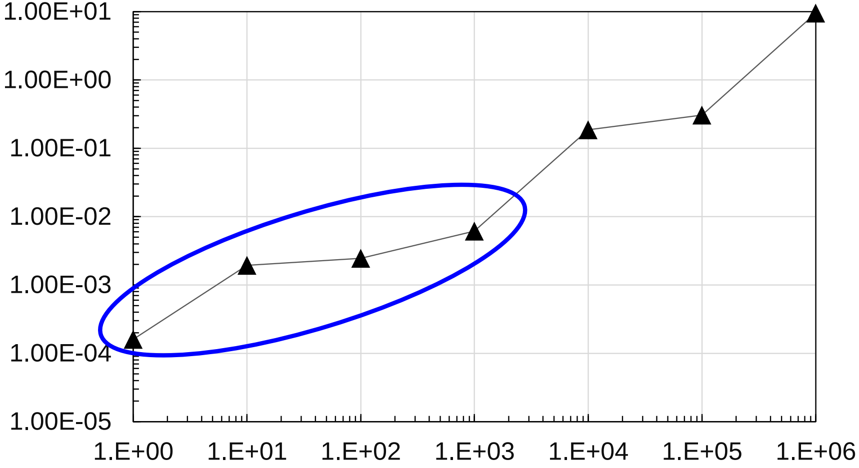
Accuracy verification for D-S and S-S has failed, if $\lambda_1/\lambda_2 \geqq 10^4$

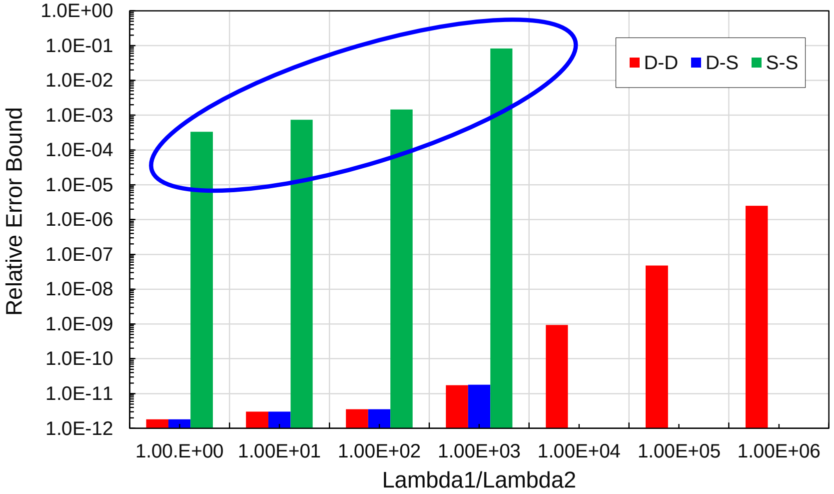
$$\|x-\hat{x}\|_\infty \leq \|\hat{z}\|_\infty + \frac{\|\hat{y}\|_\infty \|b-A(\hat{x}+\hat{z})\|_\infty}{1-\|e-A\hat{y}\|_\infty}$$

Relative Error of P3D @● (FP32 & FP64)



Max. Relative Error Bound
■ : D-D, ■ : D-S, ■ : S-S

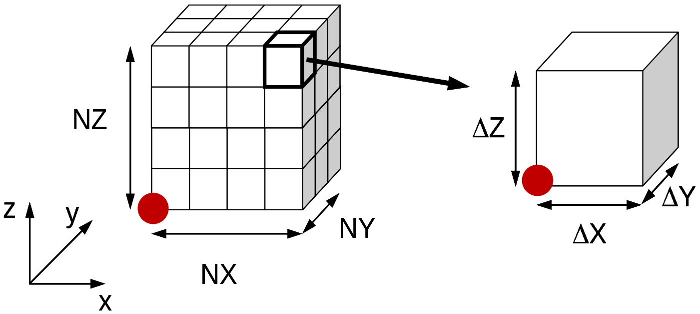




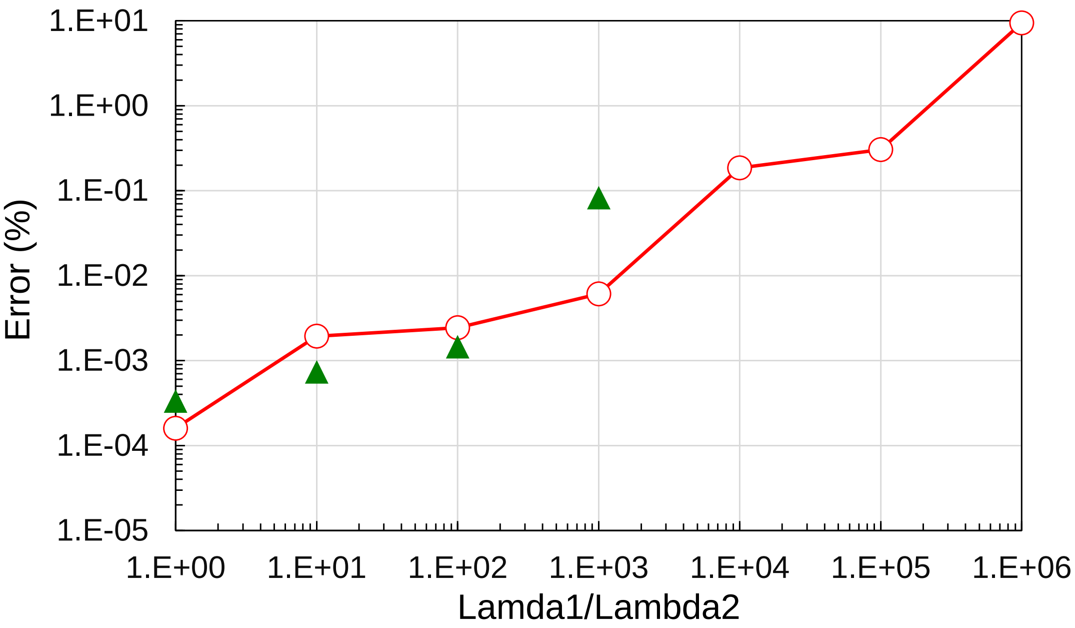
Results on OBCX (2/2)

Accuracy verification for D-S and S-S has failed, if $\lambda_1/\lambda_2 \geqq 10^4$

$$\|x-\hat{x}\|_\infty \le \|\hat{z}\|_\infty + \frac{\|\hat{y}\|_\infty \|b-A(\hat{x}+\hat{z})\|_\infty}{1-\|e-A\hat{y}\|_\infty}$$



- ○ Relative Error between FP32 & FP64 at ●
- ▲ Max. Relative Error Bound for S-S obtained from Accuracy Verification



3 equations are solved in the New Alg. for Accuracy Verification [Ogita & Nakajima 2019]

① $Ax = b, \|b - A\hat{x}\|_2 / \|b\|_2 < \varepsilon_1 (= 10^{-12})$

② $Ay = e, \|e - A\hat{y}\|_\infty < \varepsilon_2 (= 10^{-2})$

③ $Az = \hat{r}, \|\hat{r} - A\hat{z}\|_2 / \|\hat{r}\|_2 < \varepsilon_3 (= 10^{-9})$

- ①（P3D）, ②, ③
- ① and ② can be solved simultaneously
  - ② converges faster
  - Shorter Time for Computing, Better Cache Hit Rate

Concurrent Solver for ① & ② (Only for D-D)

- SLVKIND=0: separated
- SLVKIND=1：concurrent

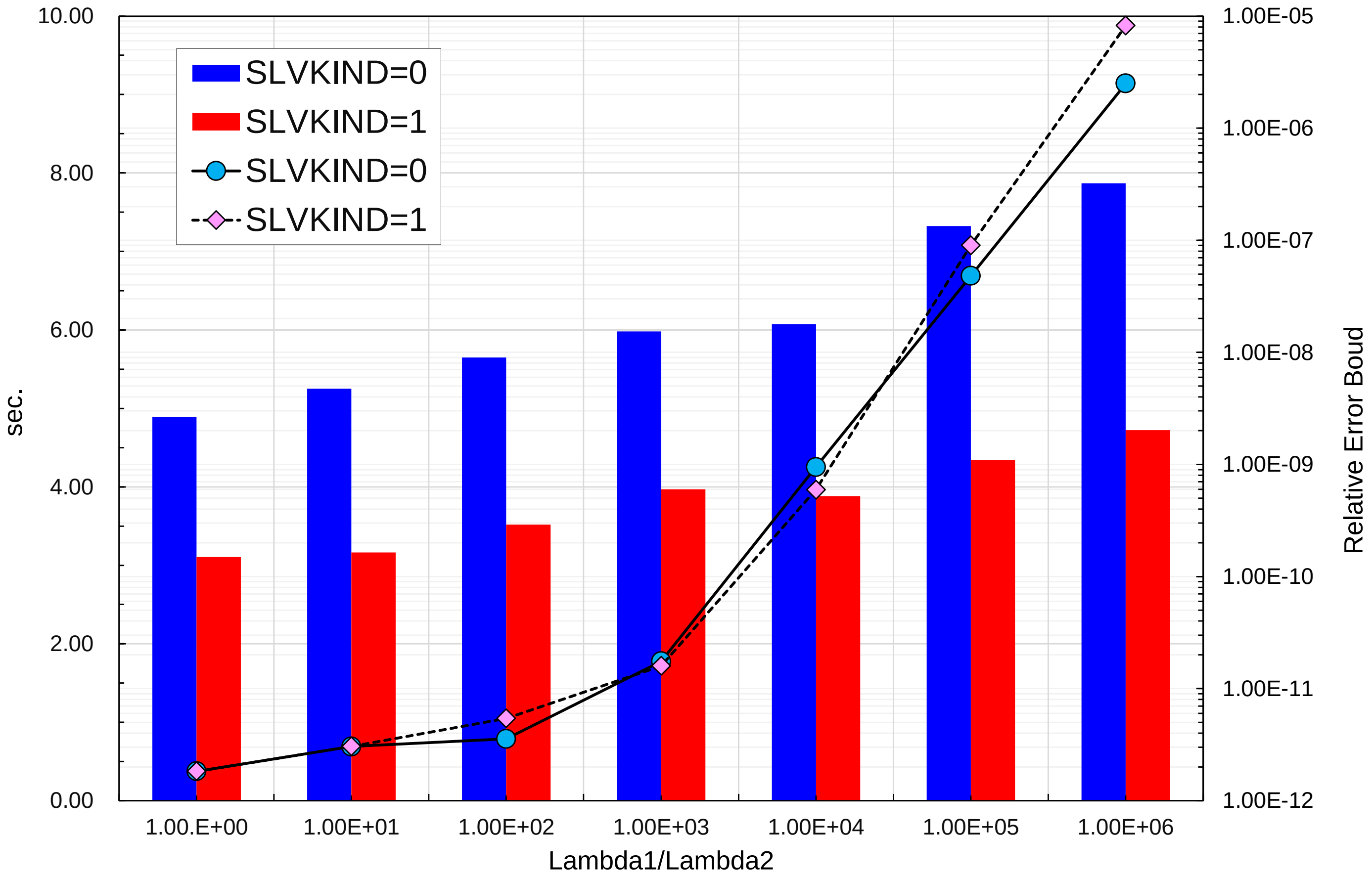
Forward Substitution

```
!$omp parallel private(ic,ip,ip1,i,WVAL1,WVAL2,k)
      do ic= 1, NCOLORtot
!$omp do
        do ip= 1, PEsmpTOT
          ip1= (ip-1)*NCOLORtot + ic
        do i= SMPindex(ip1-1)+1, SMPindex(ip1)
          WVAL1= W(i   ,Z)
          WVAL2= W(i+N,Z)
          do k= indexL(i-1)+1, indexL(i)
            WVAL1= WVAL1 -  AL(k) * W(itemL(k)   ,Z)
            WVAL2= WVAL2 -  AL(k) * W(itemL(k)+N,Z)
          enddo
          W(i   ,Z)= WVAL1 * W(i,DD)
          W(i+N,Z)= WVAL2 * W(i,DD)
        enddo
        enddo
      enddo
!$omp end parallel
```

SpMV

```
!$omp parallel do private(ip,i,VAL1,VAL2,k)
      do ip= 1, PEsmpTOT
        do i= SMPindex((ip-1)*NCOLORtot)+1,           &
              SMPindex(ip*NCOLORtot)
          VAL1= D(i)*W(i,   P)
          VAL2= D(i)*W(i+N,P)
          do k= indexL(i-1)+1, indexL(i)
            VAL1= VAL1 + AL(k)*W(itemL(k)   ,P)
            VAL2= VAL2 + AL(k)*W(itemL(k)+N,P)
          enddo
          do k= indexU(i-1)+1, indexU(i)
            VAL1= VAL1 + AU(k)*W(itemU(k)   ,P)
            VAL2= VAL2 + AU(k)*W(itemU(k)+N,P)
          enddo
          W(i,   Q)= VAL1
          W(i+N,Q)= VAL2
        enddo
      enddo
!$omp end parallel do
```

Total Time, MAX Relative Err. Bound: OBCX



Summary: Accuracy Verification

- Improved and Efficient Algorithm with Reasonable Maximum Relative Error Bound [Ogita & Nakajima 2019]
- Accuracy Verification for FP32 (SP) provides reasonable estimation of errors.