Emulating Flexible Accuracy through the Ozaki-II scheme in the ChASE eigensolver

Clément Richefort (c.richefort@fz-juelich.de), *Xinzhe Wu*, *Edoardo Di Napoli*
Jülich Supercomputing Centre, Forschungszentrum Jülich, Germany *CASA 2026*

1. Floating-Point Errors in ChASE

The ChASE Eigensolver

ChASE is an iterative eigensolver that computes the smallest eigenpairs of a Hermitian matrix A. The eigenvalue problem has the form

$$AX = \Lambda X$$

ChASE for Hermitian matrices

Require: Hermitian $A \in \mathbb{C}^{n\times n}$, nev, nex, tol
1: $V \leftarrow \mathrm{randn}(n, \mathtt{nevex})$
2: Estimate λ_1, $\lambda_{\mathtt{nevex}}$, λ_n with Lanczos
3: **while** locked $<$ nev **do**
4: $V \leftarrow p_m(A)V$ ▷ Filter
5: $Q \leftarrow [Y, V]R^{-1}$ ▷ Orthonormalize
6: $G \leftarrow Q^*AQ$ ▷ Rayleigh quotient
7: $GZ = Z\tilde{\Lambda}$ ▷ Reduced eigenproblem
8: $V \leftarrow QZ$ ▷ Back-transform
9: $\mathrm{Res}(\tilde{\Lambda}, V) = \|AV - V\tilde{\Lambda}\|_2$
10: **if** $\mathrm{Res}(\tilde{\Lambda}_i, V_i) <$ tol **then**
11: lock converged pairs; update locked
12: **end if**
13: **end while**
14: **return** $\Lambda_{\mathtt{nev}}$, $X_{\mathtt{nev}}$

In practice, we apply the filter via

$$\hat{V}^{(i+1)} = 2\sigma_{i+1}[\bar{A}\hat{V}^{(i)}]_b - \sigma_{i+1}\sigma_i\hat{V}^{(i-1)},$$

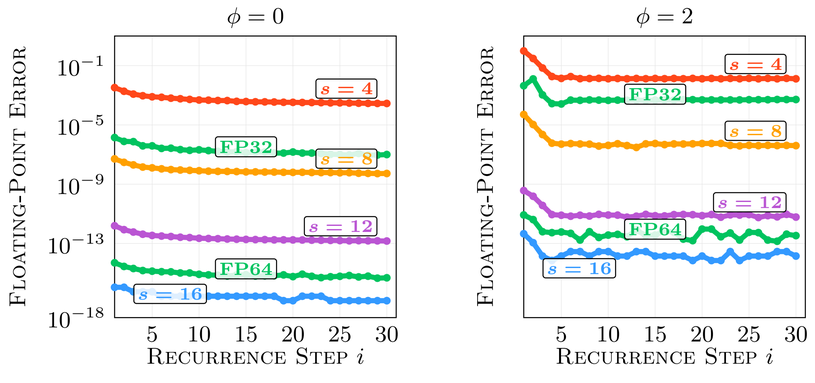
with $\hat{V}^{(1)} = \sigma_1[\bar{A}V^{(0)}]_b$. where A is encoded with b bits. The floating-point errors with Ozaki-II are

$$F^{(i)}_{b,s} := \bar{A}\hat{V}^{(i)} - [\bar{A}\hat{V}^{(i)}]_{b,s}, \text{ with } A_{ij}, V^{(0)}_{ij} = (\mathrm{rand} - \frac{1}{2}) \cdot \exp(\phi \cdot \mathrm{randn}).$$

Integrating the Ozaki-II in ChASE

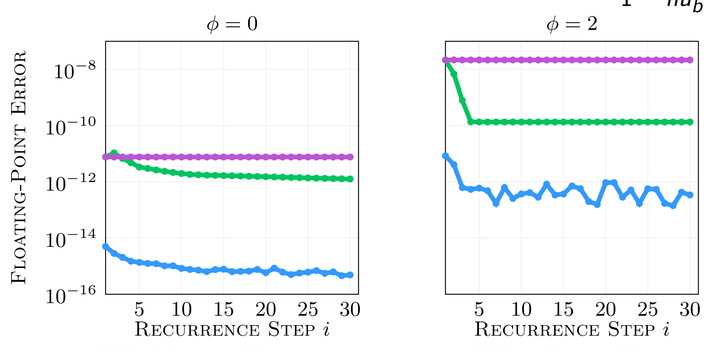
The Ozaki-II scheme computes $[\bar{A}\hat{V}^{(i)}]_{b,s}$ with s INT8 slices :

1. Scale and truncate the input matrices
$$\bar{A}' := \mathrm{trunc}(\mathrm{diag}(\mu)\bar{A}), \quad V' := \mathrm{trunc}(\hat{V}^{(i)}\mathrm{diag}(\nu))$$
2. Encode s INT8 matrices using the coprime integers p_l
$$A'_l := A' \bmod p_l, \quad V'_l := V' \bmod p_l$$
3. Perform the INT8 GEMM on matrix engines
$$C'_l = A'_l V'_l \bmod p_l$$
4. Accumulate the intermediate outputs following the CRT
$$C' := \sum_{l=1}^{s} \frac{\mathcal{P}}{p_l} q_l \, C'_l$$
5. Unscale the accumulated matrix
$$AV^{(i)} \approx [\bar{A}\hat{V}^{(i)}]_{b,s} := \mathrm{diag}(\mu^{-1})C'\mathrm{diag}(\nu^{-1})$$



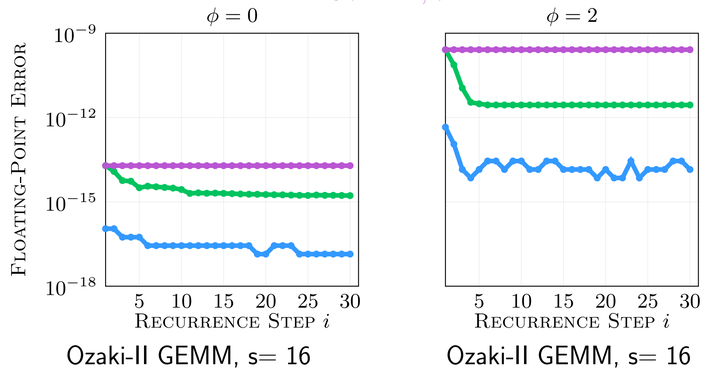
Deriving practical bounds

IEEE GEMM: $\max|F^{(i)}_b| \le \gamma_n \|\bar{A}\|_\infty \max|V^{(0)}|, \quad \gamma_n := \frac{nu_b}{1 - nu_b}$



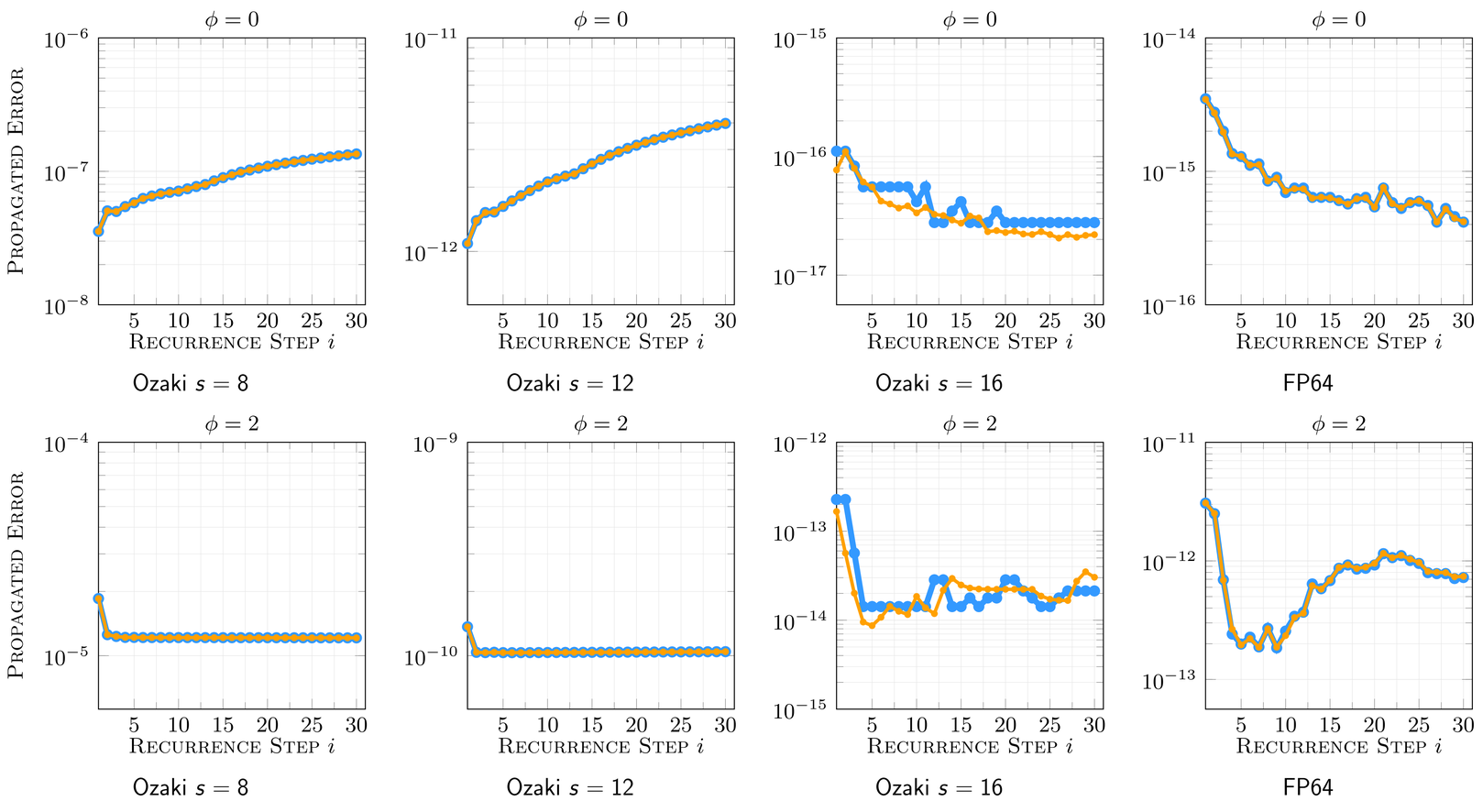
IEEE GEMM, FP64 IEEE GEMM, FP64

Ozaki-II GEMM: $\max|F^{(i)}_{b,s}| \le t_s(\|\bar{A}\|_\infty \max 2^{v'} + \max 2^{a'}\|\hat{V}^{(0)}\|_1) + t_s^2(k + r_{b,s}) \max 2^{a'} \max 2^{v'}$



Ozaki-II GEMM, s= 16 Ozaki-II GEMM, s= 16

2. Propagation of the Floating-Point Errors in ChASE



Ozaki $s = 8$ Ozaki $s = 12$ Ozaki $s = 16$ FP64

The propagated error as recurrence step i is

$$E^{(i)}_{b,s} := V^{(i)} - \hat{V}^{(i)}$$

It can be written as the three-terms recurrence relation :

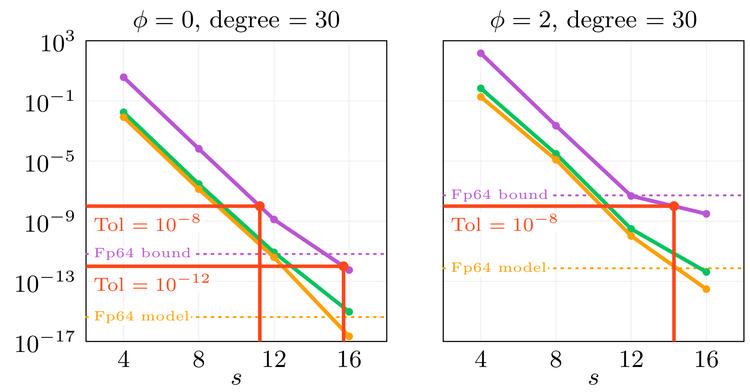
$$E^{(i+1)}_{b,s} = 2\sigma_{i+1}\bar{A}E^{(i)}_{b,s} - \sigma_{i+1}\sigma_i E^{(i-1)}_{b,s} + 2\sigma_{i+1}F^{(i)}_{b,s}.$$

Unrolling it yields Chebyshev polynomials of the second kind:

$$E^{(i+1)}_{b,s} = \sum_{j=0}^{i} w_j \frac{\eta_j}{\eta_{i+1}} U_{i-j}(\bar{A})\, F^{(j)}_{b,s}.$$

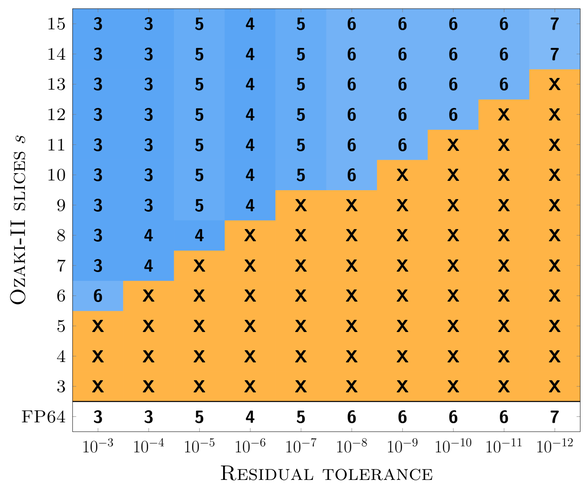
Blue: measured $\max|E^{(i)}|$ vs. FP128.

Orange: the model evaluated on the measured $F^{(j)}$.

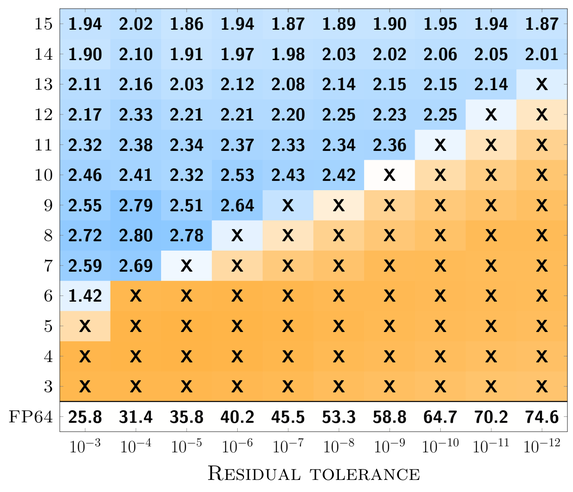


The model and the bounds connect the input tolerance to the number of Ozaki-II slices.

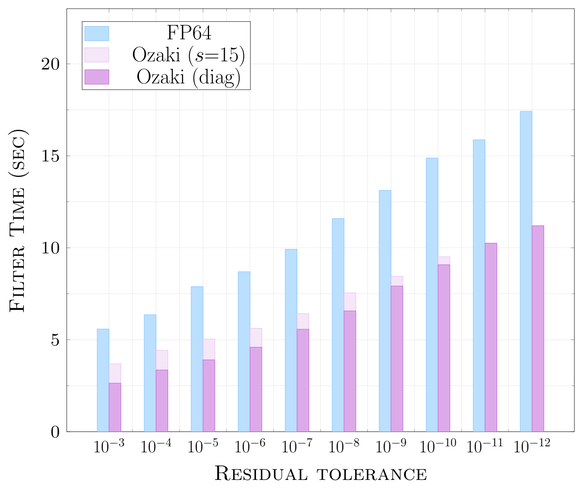
3. Numerical Experiments



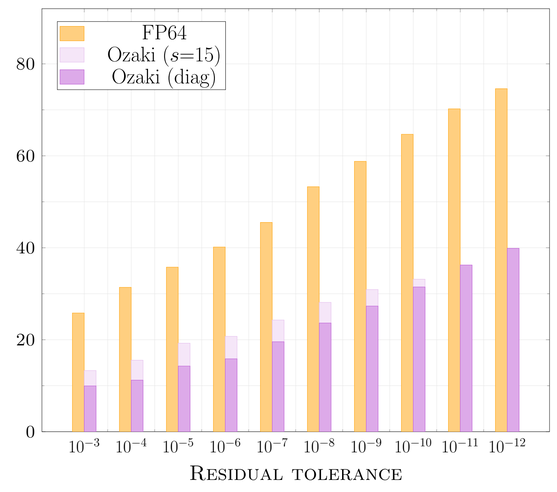
Iterations to convergence ($N = 115$k)



Filter speedup vs. FP64 ($N = 115$k)



Filter time, medium ($N = 62$k)



Filter time, large ($N = 115$k)

Benchmark on one GH200 node (4 H100 GPUs) on JUPITER — FP64: 67 TFLOPS/s, INT8: 1950 TOPS/s