MAX PLANCK INSTITUTE FOR DYNAMICS OF COMPLEX TECHNICAL SYSTEMS MAGDEBURG

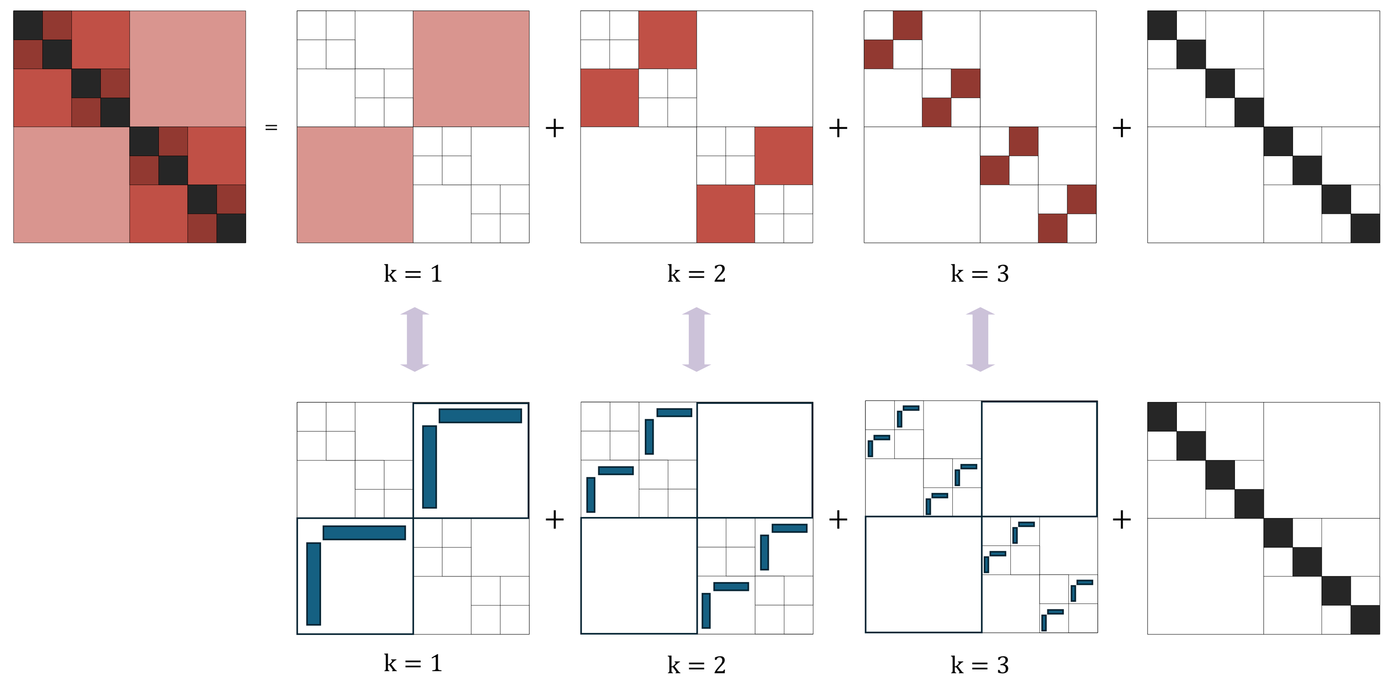
CSC COMPUTATIONAL METHODS IN SYSTEMS AND CONTROL THEORY

Mixed precision HODLR matrices

Erin Carson[1], Xinye Chen[2], Xiaobo Liu[3]

[1]Department of Numerical Mathematics, Charles University, Czech Republic (carson@karlin.mff.cuni.cz). [2]LIP6, Sorbonne University, CNRS, France (Xinye.Chen@lip6.fr). [3]Max Planck Institute for Dynamics of Complex Technical Systems, Germany (xliu@mpi-magdeburg.mpg.de).

Hierarchical Off-Diagonal Low-Rank matrix



Off-diagonal blocks at each level stored as low-rank ($\leq p$) matrix outer product UV^T, to achieve

1. $\mathcal{O}(pn\log n)$ storage requirement: $O(np)$ each level $\times$ $O(\log n)$ levels;
2. arithmetic operation cost of $\mathcal{O}(p^\alpha n \log^\beta n)$, $\alpha, \beta \in \{1, 2\}$, for matrix summation, multiplication, inversion, factorizations, etc.

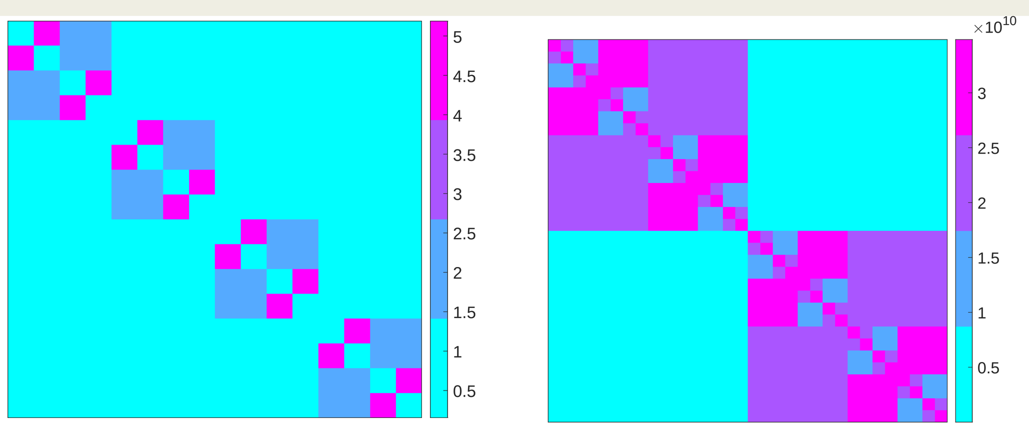
■ Applications: fast solvers for elliptic PDEs, kernel methods in machine learning, integral equations in computational physics, etc.

New definition: (p, ε)-HODLR matrix

Definition. *The off-diagonal blocks of the (p,ε)-HODLR matrix $\widetilde{H}$ (associated with a HODLR matrix H) satisfy $\|\widetilde{H}_{ij}^{(k)} - H_{ij}^{(k)}\| \leq \varepsilon \|H_{ij}^{(k)}\|$ at any level k, where $0 \leq \varepsilon < 1$ is the low-rank approximation threshold.*

- Rank-constrained $\to$ tolerance-constrained HODLR format, to facilitate error analysis of the low-rank approximations.

Mixed-precision storage scheme



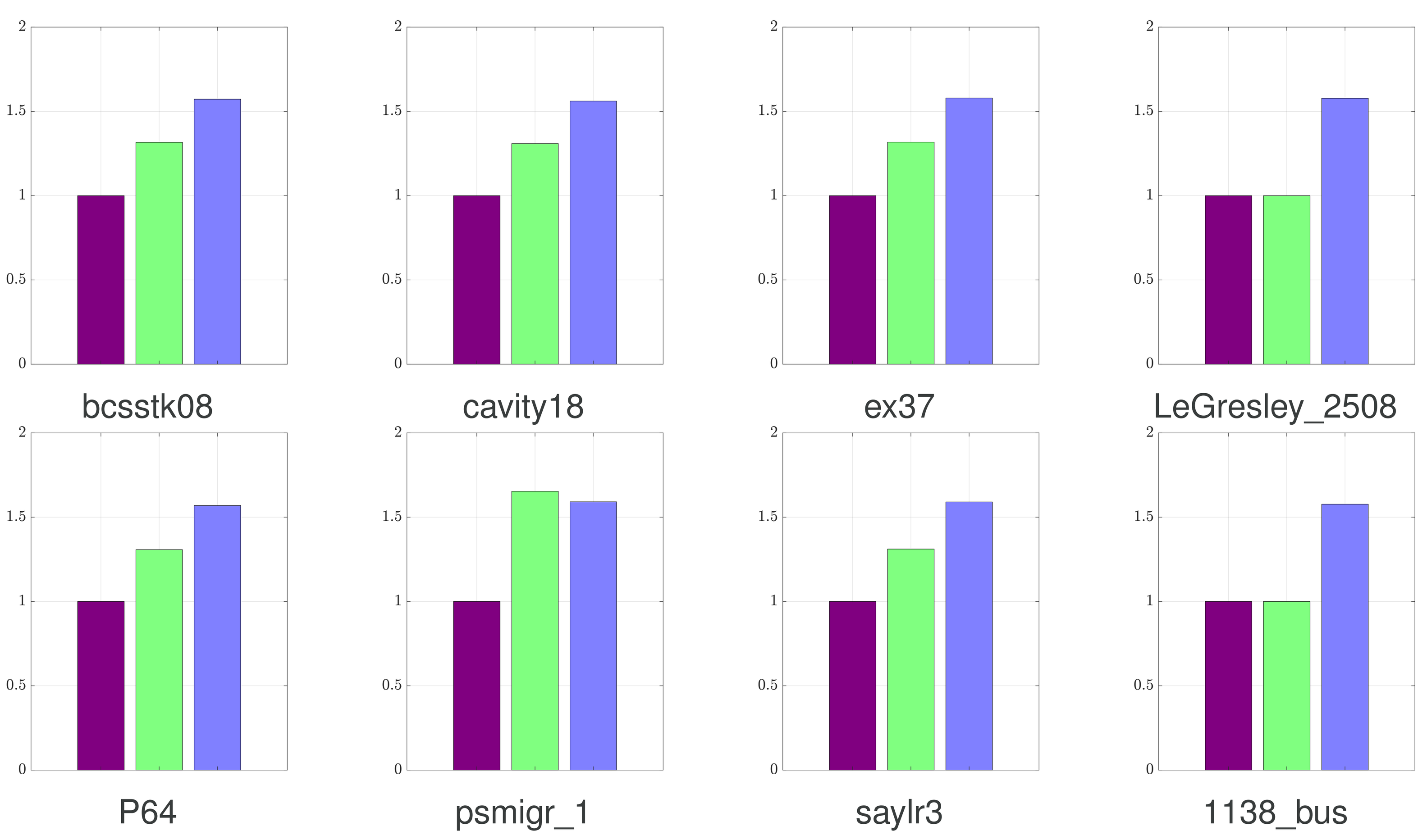
saylr3 LeGresley_2508

Norm distributions among layers for two HODLR matrices with depth $\ell = 6$.

■ Core idea: reduced precision $u_k \geq u$ for storing the low-rank factors U and V^T of off-diagonal blocks in the kth layer.

- Choose $u_k \leq \varepsilon/(2^{k/2}\xi_k)$, where $\xi_k := \max_{i\neq j} \|\widetilde{H}_{ij}^{(k)}\|_F / \|\widetilde{H}\|_F$ characterizes the relative importance of the off-diagonal blocks in magnitude.
- With $\varepsilon = 10^{-4}$, Use {bf16, fp16, fp16, fp32, fp32, fp32} for saylr3 and {q52, fp32, fp32, fp32, fp32, fp32} for LeGresley_2508.

■ Numerical experiments on Schur complements of *SuiteSparse* matrices:



Storage savings relative to uniform (double) precision HODLR matrices. Depth $\ell = 8$; purple: $\varepsilon = 10^{-7}$, green: $\varepsilon = 10^{-4}$, and blue: $\varepsilon = 10^{-1}$.

HODLR matrix–vector product

■ Core idea: balance the approximation error in $\widehat{H}$ (mixed-precision HODLR representation) and the finite-precision computation error in $b \leftarrow \widehat{H}x$.

Theorem (Backward error). *If $b = \widehat{H}x$ is computed recursively, exploiting the HODLR structure, in working precision $u \leq \varepsilon/n$, then the computed $\widehat{b}$ satisfies*

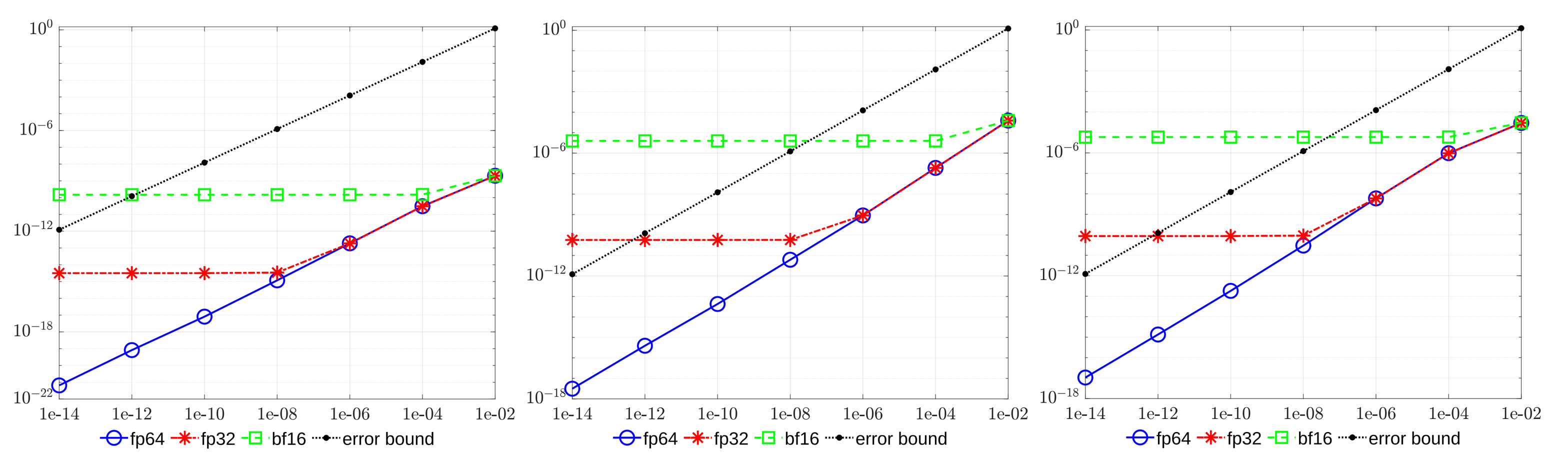
$$\widehat{b} = \mathrm{fl}(\widehat{H}x) = (H + \Delta H)x, \quad \|\Delta H\|_F \leq 2(\sqrt{2}+1)\sqrt{2^{\ell+1} + 2^{\ell-1}}\varepsilon\|H\|_F.$$

- Intuition: more inexact the low-rank representation, lower the precision one can safely use.

■ Numerical experiments on three *kernel matrices* of size $n = 2000$,

$$K_{ij} = \begin{cases} \frac{1}{x-y}, & \text{if } x \neq y; \\ 1, & \text{otherwise.} \end{cases}, \quad K_{ij} = \begin{cases} \log\|x_i - x_j\|_2, & \text{if } x \neq y; \\ 0, & \text{otherwise.} \end{cases}, \quad K_{ij} = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2}\right),$$

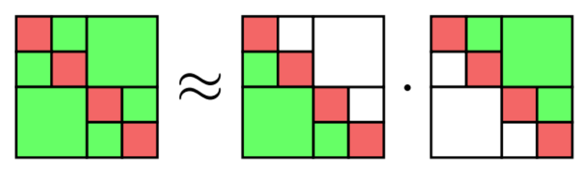
evaluated at point sets $\mathbf{s_1}$ (1D uniform grid points in $[0, 1]$) and $\mathbf{s_2} = \mathbf{s_3}$ (2D uniform grid points in $[-1, 1] \times [-1, 1]$):



kernel mat-$\mathbf{s_1}$ kernel mat-$\mathbf{s_2}$ kernel mat-$\mathbf{s_3}$

Backward error in different working precisions. Depth $\ell = 8$. The x-axis indicates the value of ε.

HODLR LU factorization

HODLR LU factorization $H^{(k)} \approx LU$ schematic: 

■ Key steps in the recursive algorithm:

1: Partition $H^{(k)}$ into $\begin{bmatrix} H_{11}^{(k+1)} & H_{12}^{(k+1)} \\ H_{21}^{(k+1)} & H_{22}^{(k+1)} \end{bmatrix}$

2: $L_{11}, U_{11} \leftarrow \texttt{HODLR\_LU}(H_{11}^{(k+1)}, k+1)$

3: $U_{12} \leftarrow$ Solve triangular system $L_{11}U_{12} = H_{12}^{(k+1)}$

4: $L_{21} \leftarrow$ Solve triangular system $L_{21}U_{11} = H_{21}^{(k+1)}$

5: $H_{22}^{\varepsilon} \leftarrow H_{22}^{(k+1)} - L_{21}U_{12}$ (possible rank-p truncation)

6: $L_{22}, U_{22} \leftarrow \texttt{HODLR\_LU}(H_{22}^{\varepsilon}, k+1)$

7: $L \leftarrow \begin{bmatrix} L_{11} & \\ L_{21} & L_{22} \end{bmatrix}$, $U \leftarrow \begin{bmatrix} U_{11} & U_{12} \\ & U_{22} \end{bmatrix}$

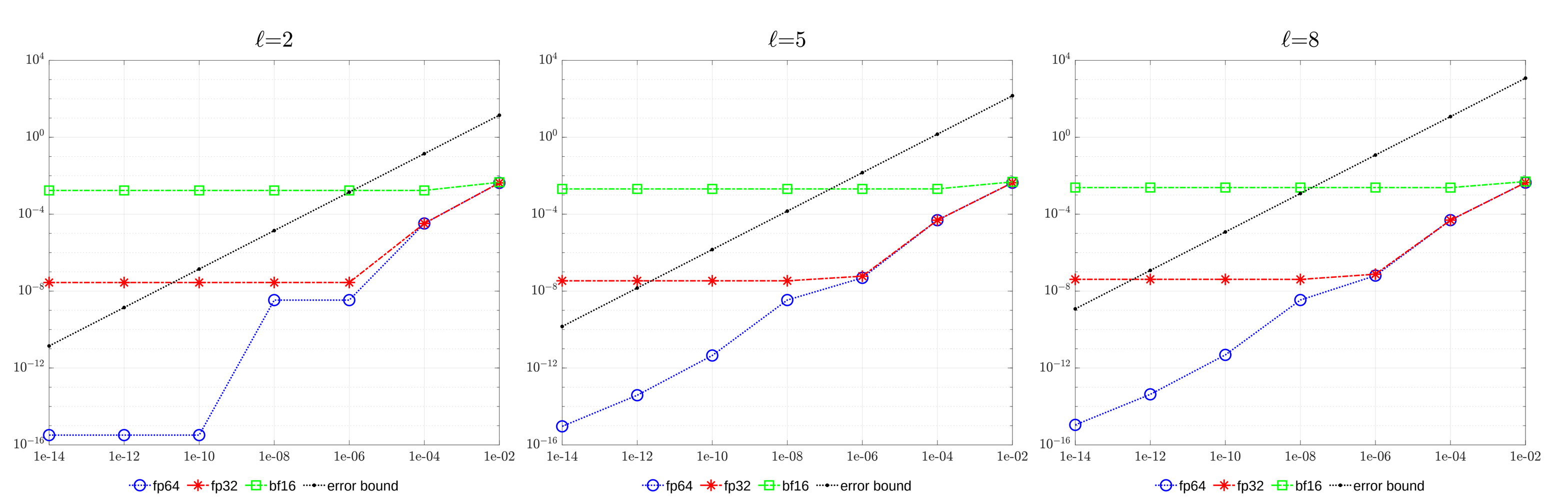
- At the bottom level $k = \ell$, dense LU factorization of $H_{ii}^{(\ell)}$ is computed.

Theorem (Backward error). *If the LU decomposition of $\widehat{H}$ is computed in a working precision $u \lesssim \varepsilon/n$, then*

$$\widehat{L}\widehat{U} = H + \Delta H, \quad \|\Delta H\|_F \lesssim 2(2^\ell - 1)\varepsilon\|H\|_F + 11(2^\ell - 1)\varepsilon\|\widehat{L}\|_F\|\widehat{U}\|_F.$$

- For larger ε (a coarser approximation), the LU factorization can be computed in a lower precision *without affecting the backward error.*
- Bound remains in the same form for LU factorization of one-precision HODLR matrices, albeit with smaller constant factors.

■ Numerical experiments on the Schur complement of *ex37*:



Backward error in different working precisions. The x-axis indicates the value of ε.

Further reading

E. Carson, X. Chen, and X. Liu. Mixed precision HODLR matrices. *SIAM J. Sci. Comput.*, 47(3):A1408–A1435, 2025.