Minisymposium

Approximate Computing for Scientific Applications

siam 2023 AMSTERDAM
Conference on Computational Science and Engineering

Minisymposium Approximate Computing

MS290: Part I

- *"Combining Binary Compression with Low-Rank Arithmetic"*, R.K.
- *"A Fast Solver for Linear Systems with Tensor Product Structure via Low-Rank Updates"*, Stefano Massei
- *"Runtime System Considerations for Approximate Computing at Scale"*, George Bosilca
- *"Parallel QR Factorization of Block Low-Rank Matrices"*, Muhammad Ridwan Apriansyah
- ~~*"Inexact Rational Krylov Methods for Large Matrix Equations"*, Patrick Kürschner~~

MS324: Part II

- *"Computational Efficiency through Tuned Approximation"*, David E. Keyes
- *"Portable Mixed Precision for the Iterative Solution of Sparse Linear Systems"*, Enrique S. Quintana-Ortí
- *"Mixed Precision Linear Algebra for High Fidelity Real-Time Wavefront Reconstruction on Giant Optical Telescopes"*, Damien Gratadour
- *"Leveraging Half-Precision in Wireless Communication"*, Adel Dabah

Lowrank Techniques

Approximation

Approximate dense data $M \in \mathbb{C}^{n \times m}$ by $U \cdot V^H$ with $U \in \mathbb{C}^{n \times k}, V \in \mathbb{C}^{m \times k}$ and $k \ll n$ such that

$$||M - UV^H|| \leq \varepsilon ||M||,$$

with user defined $\varepsilon > 0$, via SVD, RRQR, RandSVD, ACA, Lanczos, ...

Lowrank Techniques

Approximation

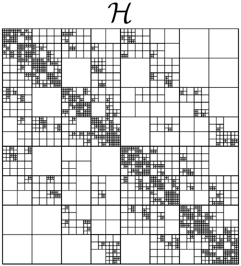
Approximate dense data $M \in \mathbb{C}^{n \times m}$ by $U \cdot V^H$ with $U \in \mathbb{C}^{n \times k}, V \in \mathbb{C}^{m \times k}$ and $k \ll n$ such that

$$||M - UV^H|| \leq \varepsilon ||M||,$$

with user defined $\varepsilon > 0$, via SVD, RRQR, RandSVD, ACA, Lanczos, ...

Blockwise Lowrank

As M normally does not have lowrank property $\Rightarrow$ decompose into subblocks.



Lowrank Techniques

Approximation

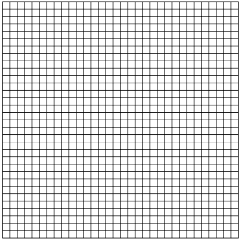
Approximate dense data $M \in \mathbb{C}^{n \times m}$ by $U \cdot V^H$ with $U \in \mathbb{C}^{n \times k}, V \in \mathbb{C}^{m \times k}$ and $k \ll n$ such that

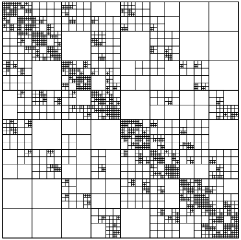
$$||M - UV^H|| \leq \varepsilon ||M||,$$

with user defined $\varepsilon > 0$, via SVD, RRQR, RandSVD, ACA, Lanczos, ...

Blockwise Lowrank

As M normally does not have lowrank property $\Rightarrow$ decompose into subblocks.





Lowrank Techniques

Approximation

Approximate dense data $M \in \mathbb{C}^{n \times m}$ by $U \cdot V^H$ with $U \in \mathbb{C}^{n \times k}, V \in \mathbb{C}^{m \times k}$ and $k \ll n$ such that

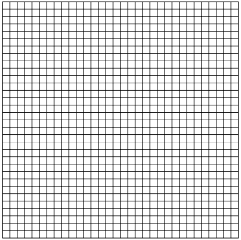
$$||M - UV^H|| \leq \varepsilon ||M||,$$

with user defined $\varepsilon > 0$, via SVD, RRQR, RandSVD, ACA, Lanczos, ...

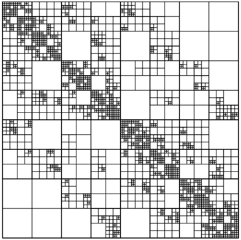
Blockwise Lowrank

As M normally does not have lowrank property $\Rightarrow$ decompose into subblocks.

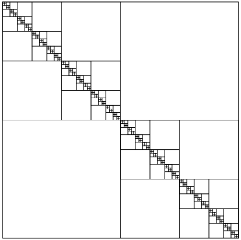
BLR/TLR



$\mathcal{H}$



HODLR



Number Representation

IEEE 754

	S-E-M[1]	Bits	Unit Roundoff	Performance[2]
FP80	1-15-64	80	2.7×10^{-20}	
FP64	1-11-52	64	1.1×10^{-16}	34 TFlops
FP32	1-8-23	32	6.0×10^{-8}	67 TFlops
TF32	1-8-10	19	4.9×10^{-4}	494 TFlops
FP16	1-5-10	16	4.9×10^{-4}	989 TFlops
BF16	1-8-7	16	3.9×10^{-3}	989 TFlops
FP8	1-4-3	8	6.2×10^{-2}	1979 TFlops

Huge potential for performance improvements *if applicable*.

[1] Sign – Exponent – Mantissa

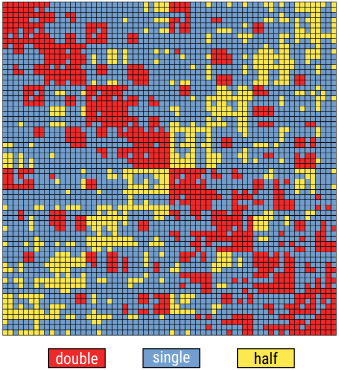
[2] NVidia H100 datasheet (https://www.nvidia.com/en-us/data-center/h100/)

Mixed Precision[1]

Factorization of block lowrank (BLR) matrices.

Precision of lowrank blocks chosen based on norm.

Talk by George Bosilca



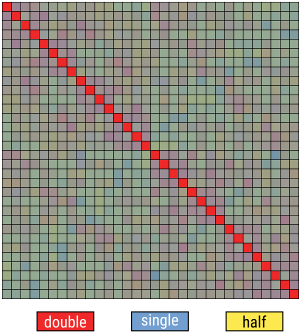
[1] Abdulah, Cao, Pei, Bosilca, Dongarra, Genton, Keyes, Ltaief, Sun: "Accelerating Geostatistical Modeling and Prediction With Mixed-Precision Computations: A High-Productivity Approach With PaRSEC", IEEE Trans. on Par. and Distr. Systems, 2022

Mixed Precision v2[1,2]

Split UV^H into

$$U \cdot V^H = [W_1 W_2 W_3 \ldots] \cdot \mathrm{diag}(\sigma_1, \ldots, \sigma_k) \cdot [X_1 X_2 X_3 \ldots]^H$$

with orthogonal W_i, X_i using precisions depending on the singular values σ_j.



[1] Ooi, Iwashita, Fukaya, Ida, Yokota.: "Effect of Mixed Precision Computing on H-Matrix Vector Multiplication in BEM Analysis", Proceedings of HPCAsia2020, 2020

[2] Amestoy, Boiteau, Buttari, Gerest, Jézéquel, L'Excellent, Mary: "Mixed precision low-rank approximations and their application to block low-rank LU factorization", IMA J. of Num. Analysis, 2022

Combining Binary Compression with Low-Rank Arithmetic

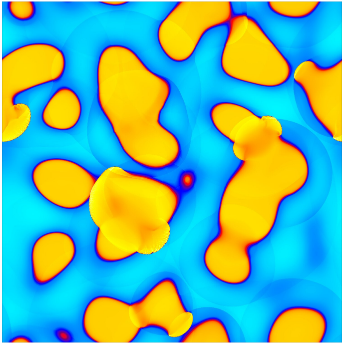
Ronald Kriemann
MPI MIS Leipzig

CSE23

MAX PLANCK INSTITUTE
FOR MATHEMATICS IN THE SCIENCES

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:

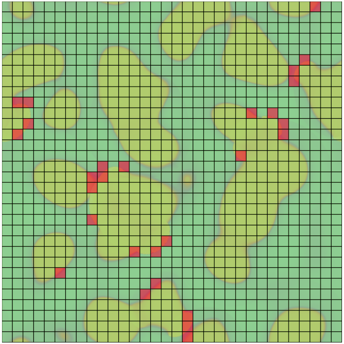


[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022

[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:

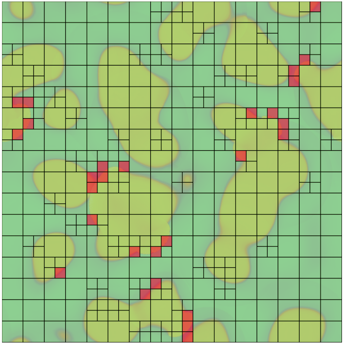


[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022

[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:

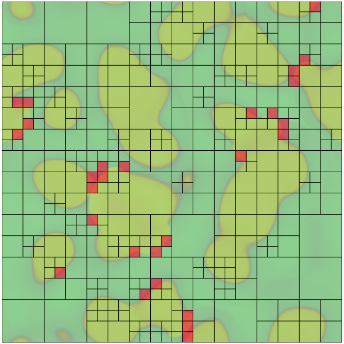


[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022

[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:



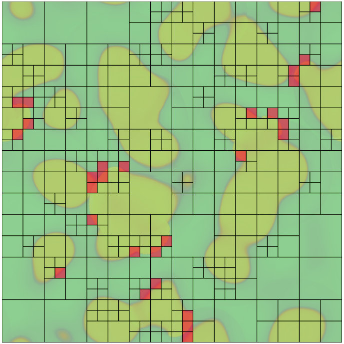
[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022

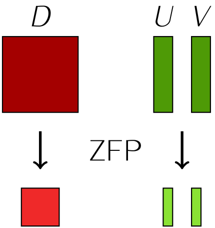
[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).

[3] Di, Cappello: "Fast Error-Bounded Lossy HPC Data Compression with SZ", IEEE IPDPS. pp. 730–739 (2016)

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:

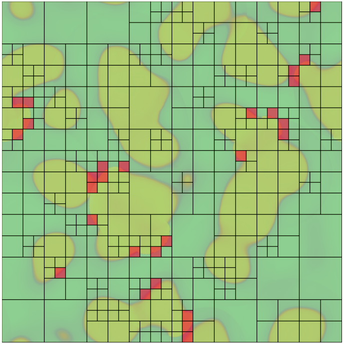


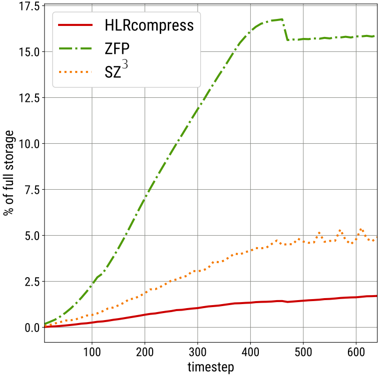


[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022
[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).
[3] Di, Cappello: "Fast Error-Bounded Lossy HPC Data Compression with SZ", IEEE IPDPS. pp. 730–739 (2016)
[4] Massei, Robol, Kressner: "*Hierarchical adaptive low-rank format with applications to discretized partial differential equations*". NLAwA (2022)

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:





[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022

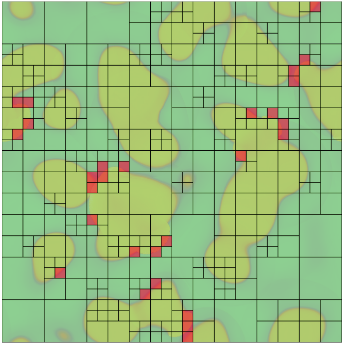
[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).

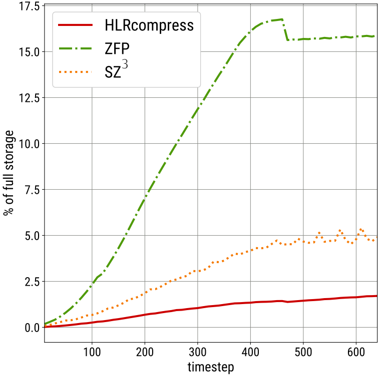
[3] Di, Cappello: "Fast Error-Bounded Lossy HPC Data Compression with SZ", IEEE IPDPS. pp. 730–739 (2016)

[4] Massei, Robol, Kressner: "*Hierarchical adaptive low-rank format with applications to discretized partial differential equations*". NLAwA (2022)

Compressed Lowrank Storage

For a combustion application[1], lowrank approximation was combined with (lossy) floating point compression using *ZFP*[2] to minimize data storage:





A similar approach (without binary compression) was used to apply $\mathcal{H}$-arithmetic on the solution level in a PDE computation[4].

[1] K., Ltaief, Luong, Pérez, Im, Keyes: "*High-Performance Spatial Data Compression for Scientific Applications*", Euro-Par 2022

[2] Lindstrom: "*Fixed-rate compressed floating-point arrays*", IEEE Trans. on Vis. and Comp. Graphics 20(12), 2674–2683 (2014).

[3] Di, Cappello: "Fast Error-Bounded Lossy HPC Data Compression with SZ", IEEE IPDPS. pp. 730–739 (2016)

[4] Massei, Robol, Kressner: "*Hierarchical adaptive low-rank format with applications to discretized partial differential equations*". NLAwA (2022)

Memory Accessor[1]

Decouple compute precision

and

storage precision.

Talk by Enrique S. Quintana-Ortí (MS324)

[1] Anzt, Flegar, Grützmacher, Quintana-Ortí: "Toward a modular precision ecosystem for high-performance computing", Int. J. of HPC Applications, 33(6), 1069–1078, 2019.

Memory Accessor

Requirements for $\mathcal{H}$–Matrices

- compress dense *and* lowrank data,
- *adaptivity* for lowrank approximation error,
- *kernel-level* conversion due to BLAS/LAPACK based arithmetic.

function TRUNCATION(**in**: U, V, ε, **out**: W, X)

U^d := decompress(U);
V^d := decompress(V);
$[Q_U, R_U]$:= qr(U^d);
$[Q_V, R_V]$:= qr(V^d);
$[U_s, S_s, V_s]$:= svd($R_U \cdot R_V^H$);
k := rank(S_s, ε);
$W^d := Q_U \cdot U_s(:, 1:k) \cdot S_s(1:k, 1:k)$;
$X^d := Q_V \cdot V_s(:, 1:k)$;
W := compress(W^d);
X := compress(X^d);

Storage Options

Compression Libraries

For adaptivity only *lossy* compression of interest.

ZFP
- *very fast*,
- for reliable error control only fixed bitrate used,
- limited compression rate.

SZ/SZ3[1]
- good compression rates for general data,
- various error control options,
- *various issues* with mt-usage, compression rate and performance.

MGARD[2]
- multi-grid technique plus lossless compression,
- various error control options,
- very slow.

[1] Zhao, Di, Dmitriev, Tonellot, Chen, Cappello: "Optimizing Error-Bounded Lossy Compression for Scientific Data by Dynamic Spline Interpolation", IEEE 37th ICDE, 1643–1654 (2021)

[2] Ainsworth, Tugluk, Whitney, Klasky: "Multilevel techniques for compression and reduction of scientific data – the univariate case". Comp.Vis.Sci. 19, 65–76 (2018)

Storage Options

Compression Libraries

For adaptivity only *lossy* compression of interest.

ZFP

- *very fast*,
- for reliable error control only fixed bitrate used,
- limited compression rate.

SZ/SZ3[1]

- good compression rates for general data,
- various error control options,
- *various issues* with mt-usage, compression rate and performance.

MGARD[2]

- multi-grid technique plus lossless compression,
- various error control options,
- very slow.

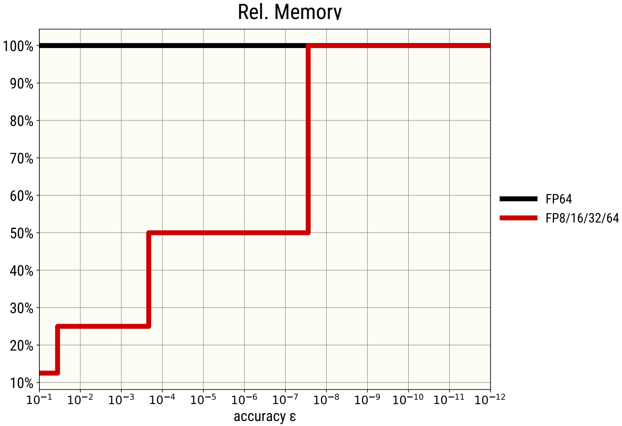
[1] Zhao, Di, Dmitriev, Tonellot, Chen, Cappello: "Optimizing Error-Bounded Lossy Compression for Scientific Data by Dynamic Spline Interpolation", IEEE 37th ICDE, 1643–1654 (2021)

[2] Ainsworth, Tugluk, Whitney, Klasky: "Multilevel techniques for compression and reduction of scientific data – the univariate case". Comp.Vis.Sci. 19, 65–76 (2018)

Storage Options

IEEE 754

	S-E-M	Unit Roundoff
FP64	1-11-52	1.1×10^{-16}
FP32	1-8-23	6.0×10^{-8}
TF32	1-8-10	4.9×10^{-4}
BF16	1-8-7	3.9×10^{-3}
FP16	1-5-10	4.9×10^{-4}
FP8	1-4-3	6.2×10^{-2}

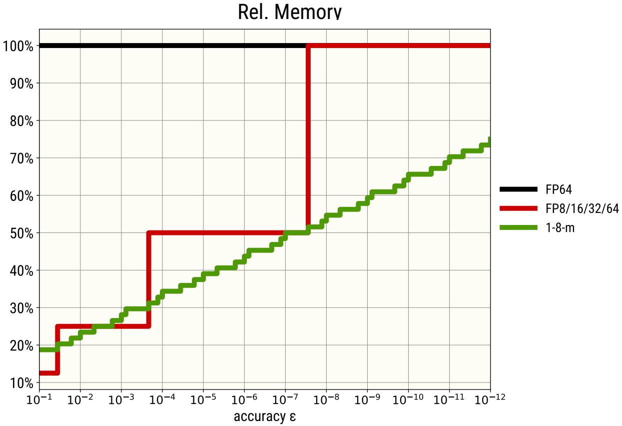


Storage Options

IEEE 754

1. choose mantissa bits m based on required accuracy,

	S-E-M	Unit Roundoff
FP64	1-11-52	1.1×10^{-16}
FP32	1-8-23	6.0×10^{-8}
TF32	1-8-10	4.9×10^{-4}
BF16	1-8-7	3.9×10^{-3}
FP16	1-5-10	4.9×10^{-4}
FP8	1-4-3	6.2×10^{-2}



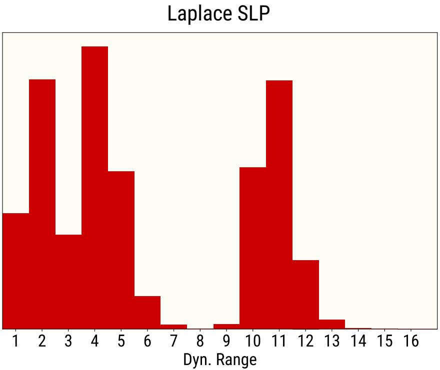
Storage Options

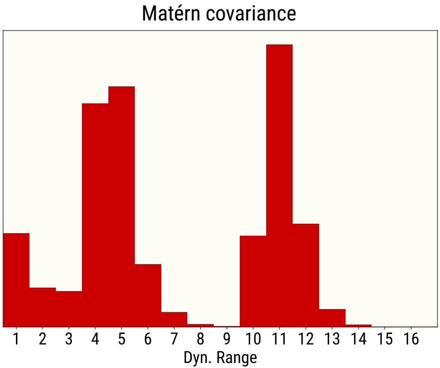
IEEE 754

1. choose mantissa bits m based on required accuracy,

	S-E-M	Unit Roundoff	Range[1]
FP64	1-11-52	1.1×10^{-16}	631
FP32	1-8-23	6.0×10^{-8}	83
TF32	1-8-10	4.9×10^{-4}	79
BF16	1-8-7	3.9×10^{-3}	78
FP16	1-5-10	4.9×10^{-4}	12
FP8	1-4-3	6.2×10^{-2}	5

[1] Dynamic range as $\log_{10} \frac{V_{\max}}{V_{\min}}$





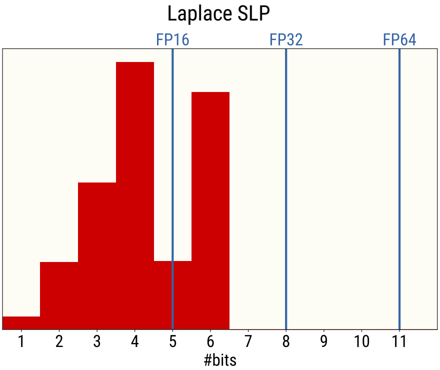
Storage Options

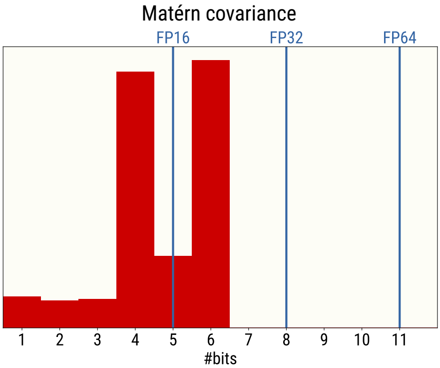
IEEE 754

1. choose mantissa bits m based on required accuracy,

	S-E-M	Unit Roundoff	Range[1]
FP64	1-11-52	1.1×10^{-16}	631
FP32	1-8-23	6.0×10^{-8}	83
TF32	1-8-10	4.9×10^{-4}	79
BF16	1-8-7	3.9×10^{-3}	78
FP16	1-5-10	4.9×10^{-4}	12
FP8	1-4-3	6.2×10^{-2}	5

[1] Dynamic range as $\log_{10} \frac{V_{\max}}{V_{\min}}$





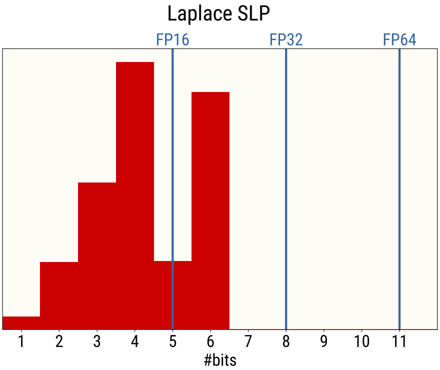
Storage Options

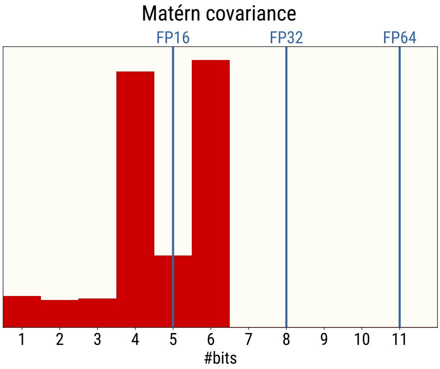
IEEE 754

1. choose mantissa bits m based on required accuracy,
2. *choose exponent bits e based on dynamic range.*

	S-E-M	Unit Roundoff	Range[1]
FP64	1-11-52	1.1×10^{-16}	631
FP32	1-8-23	6.0×10^{-8}	83
TF32	1-8-10	4.9×10^{-4}	79
BF16	1-8-7	3.9×10^{-3}	78
FP16	1-5-10	4.9×10^{-4}	12
FP8	1-4-3	6.2×10^{-2}	5

[1] Dynamic range as $\log_{10} \frac{V_{\max}}{V_{\min}}$





Storage Options

Adaptive Precision with IEEE 754

afloat:

- fully adaptive choice of m *and* e,
- use `1-e-m` to store data (with scaling and shifting),
- *slow* bit stream storage.

Storage Options

Adaptive Precision with IEEE 754

afloat:

- fully adaptive choice of m *and* e,
- use `1-e-m` to store data (with scaling and shifting),
- *slow* bit stream storage.

apfloat:

- choose e and m as in *afloat*,
- increase m such that $1 + e + m$ is multiple of 8

Storage Options

Adaptive Precision with IEEE 754

afloat:

- fully adaptive choice of m *and* e,
- use `1-e-m` to store data (with scaling and shifting),
- *slow* bit stream storage.

apfloat:

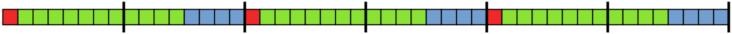
- choose e and m as in *afloat*,
- increase m such that $1 + e + m$ is multiple of 8

bfloat:

- `1-8-m` format ($1 + 8 + m$ multiple of 8)

dfloat:

- `1-11-m` format ($1 + 11 + m$ multiple of 8)



Results

Setting

Machine

- 2x64-core AMD Epyc 7702 (Rome)
- 2x8 32GB DDR4-3200 DIMMs

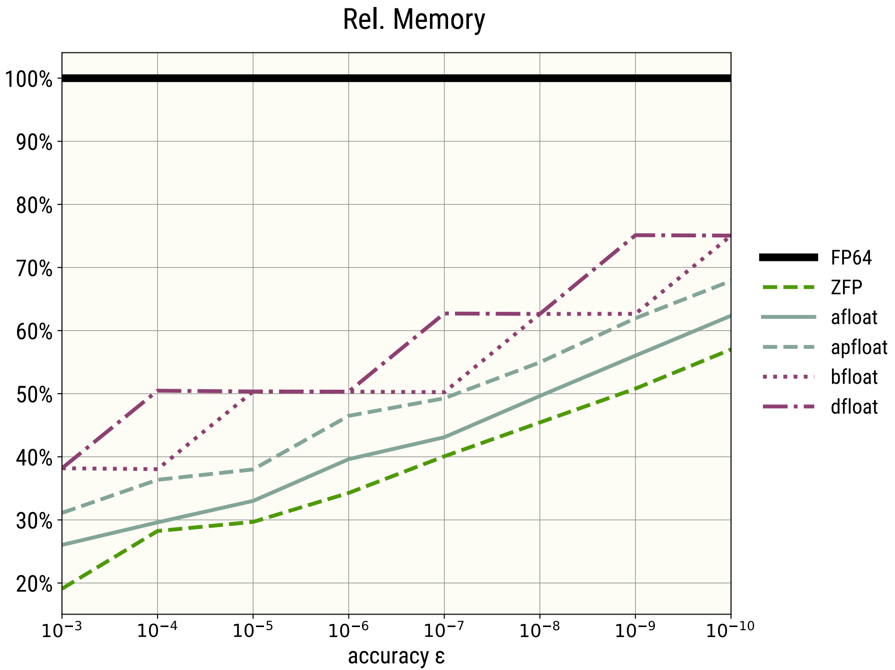
Software

- HLR (`libhlr.org`)
- Intel TBB v2021.2
- Intel MKL v2022.0 (AVX2 code path)
- GCC 12

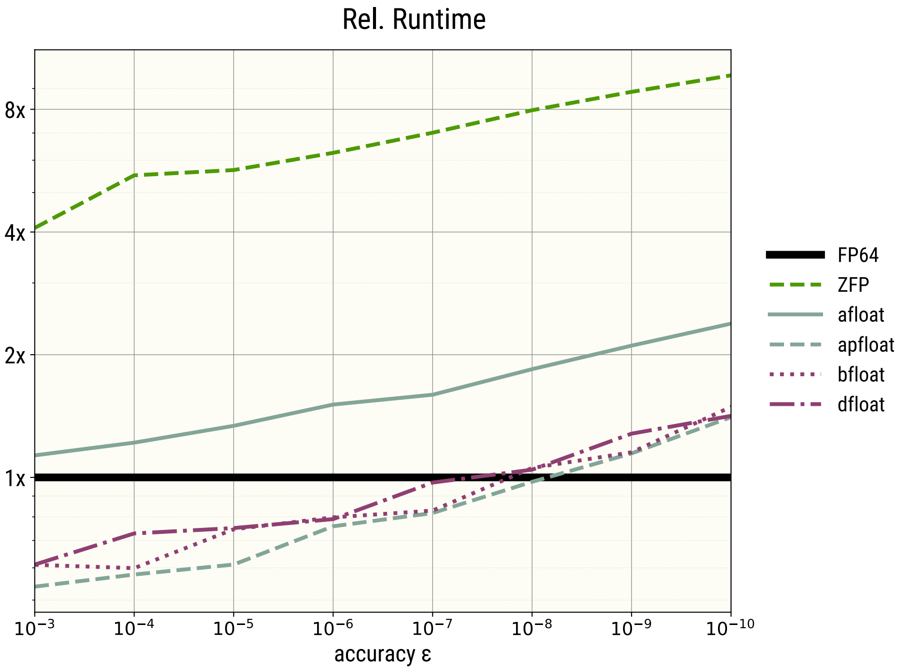
Benchmark

- Model problem: Laplace SLP on unit sphere, $n = 524.288$
- base is FP64 (computation and storage)
- standard $\mathcal{H}$-arithmetic (no accumulator)
- lowrank truncation via SVD
- runtime: median out of 5 runs

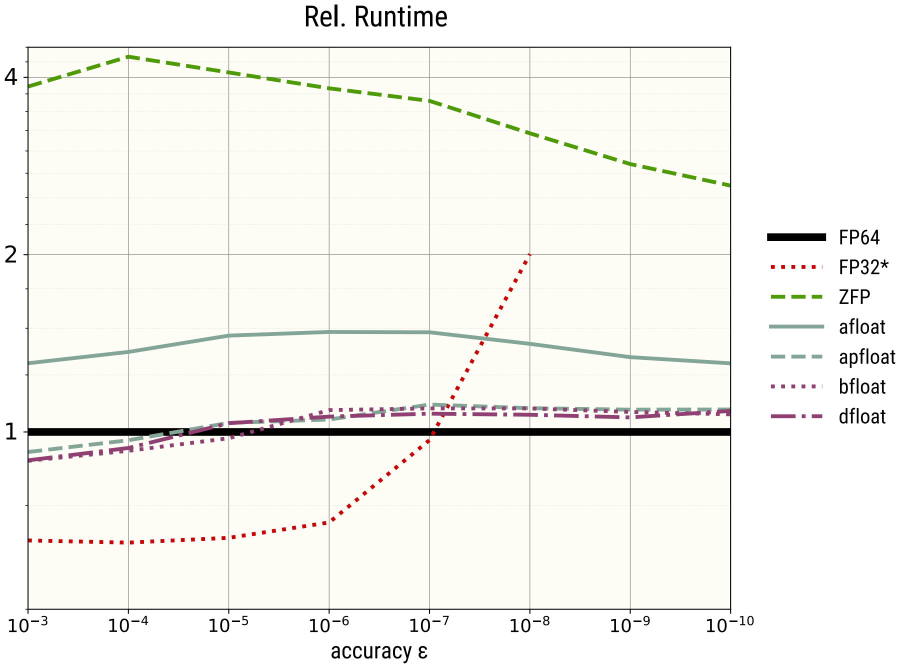
$\mathcal{H}$–compression



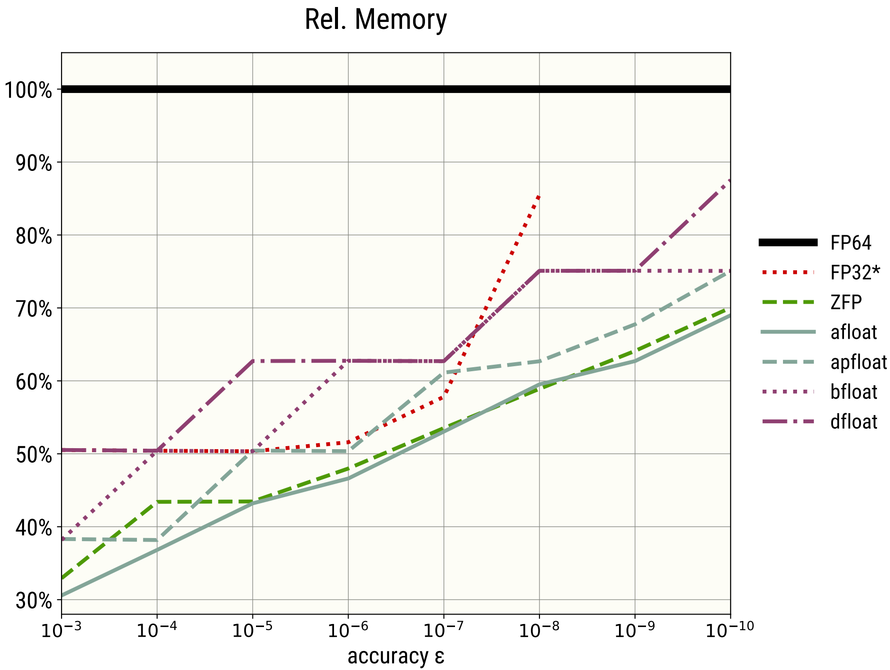
$\mathcal{H}$–MatVec



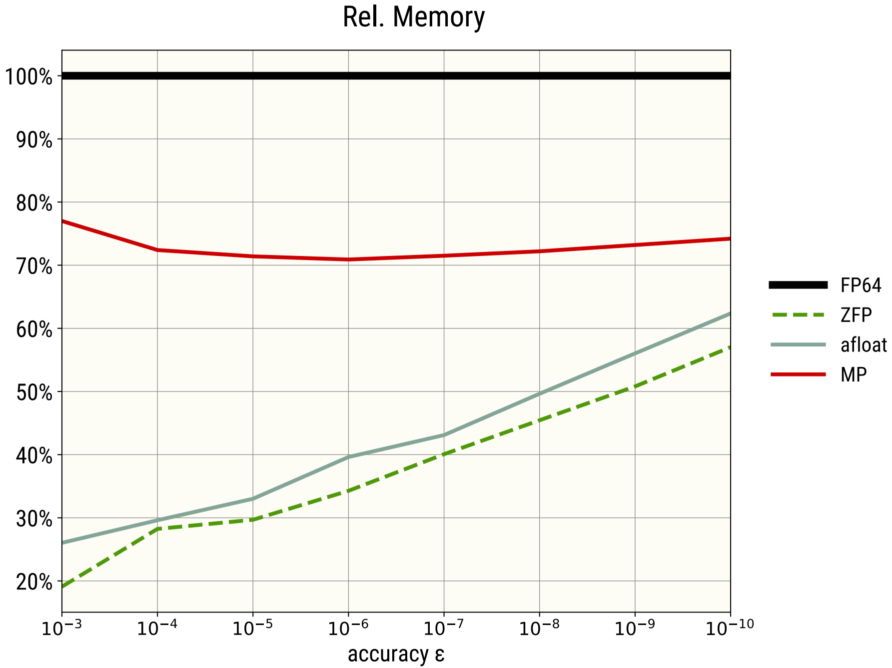
$\mathcal{H}$-LU factorisation



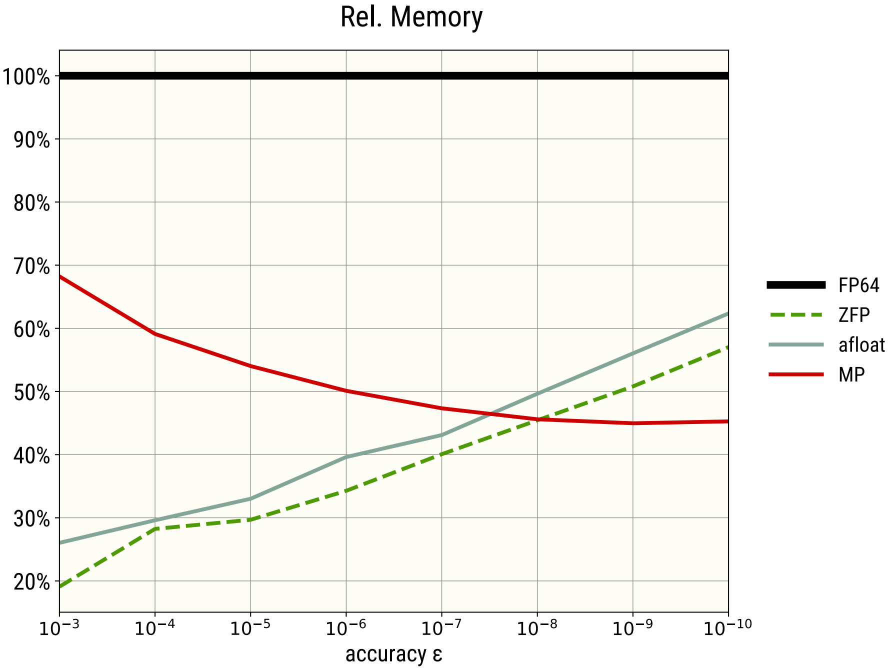
$\mathcal{H}$-LU factorisation



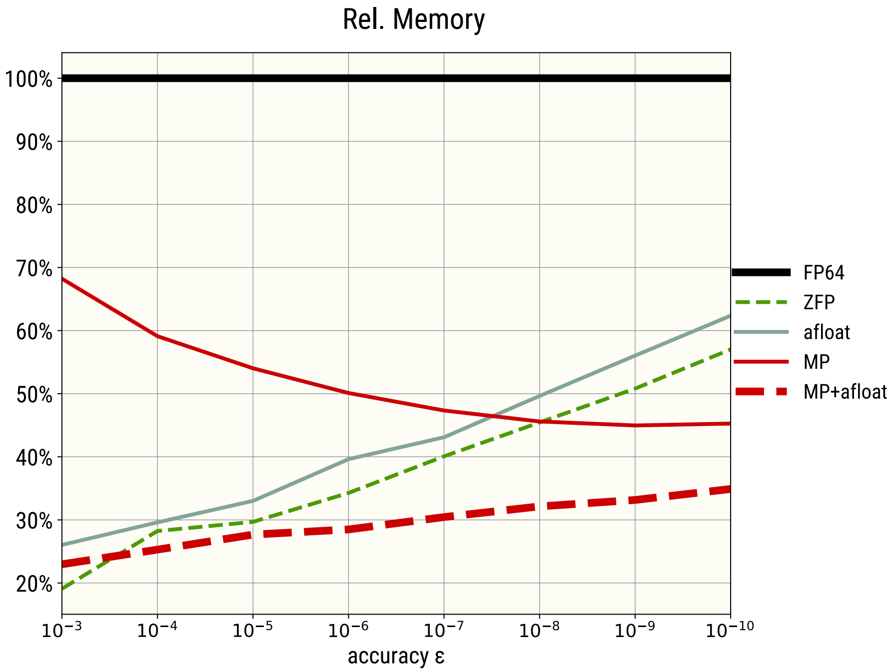
$\mathcal{H}$–compression



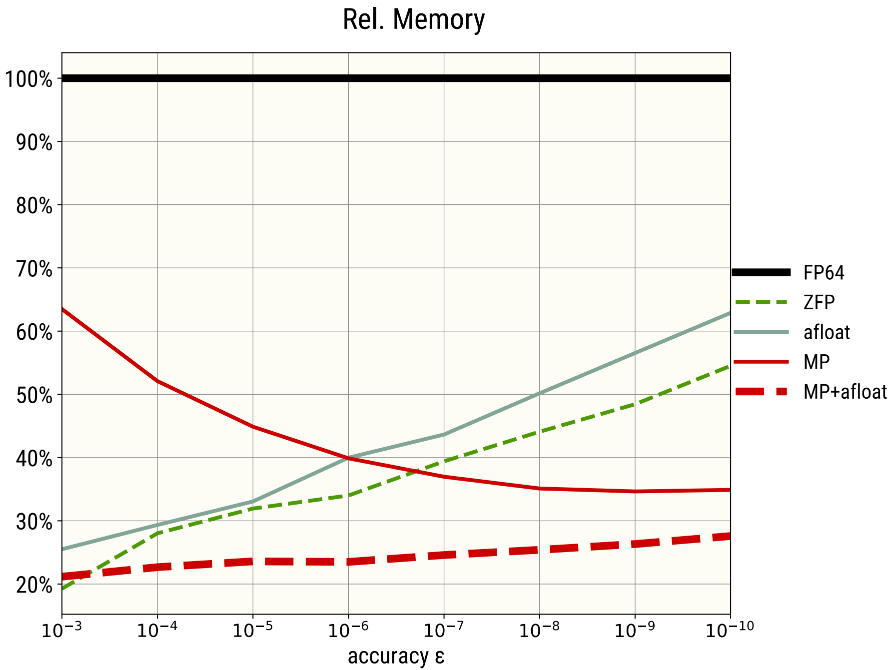
$\mathcal{H}$–compression



$\mathcal{H}$–compression



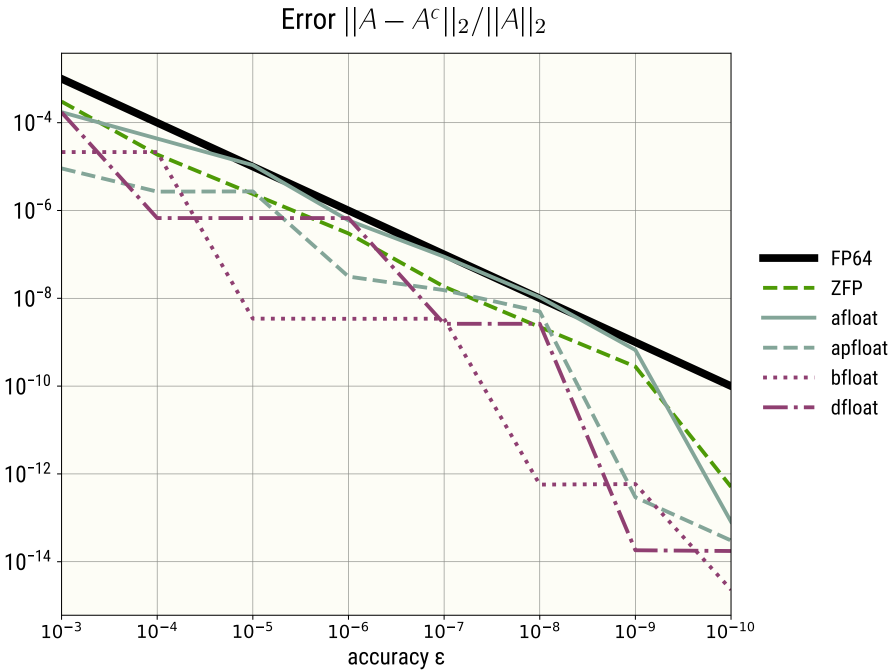
$\mathcal{H}$-compression (Matérn covariance)



Thank You

MAX PLANCK INSTITUTE
FOR MATHEMATICS IN THE SCIENCES

$\mathcal{H}$–compression



$\mathcal{H}$-LU factorisation