High Performance Mixed Precision Numerical Linear Algebra

Erin Carson

Charles University

SimRace 2021, IPFEN

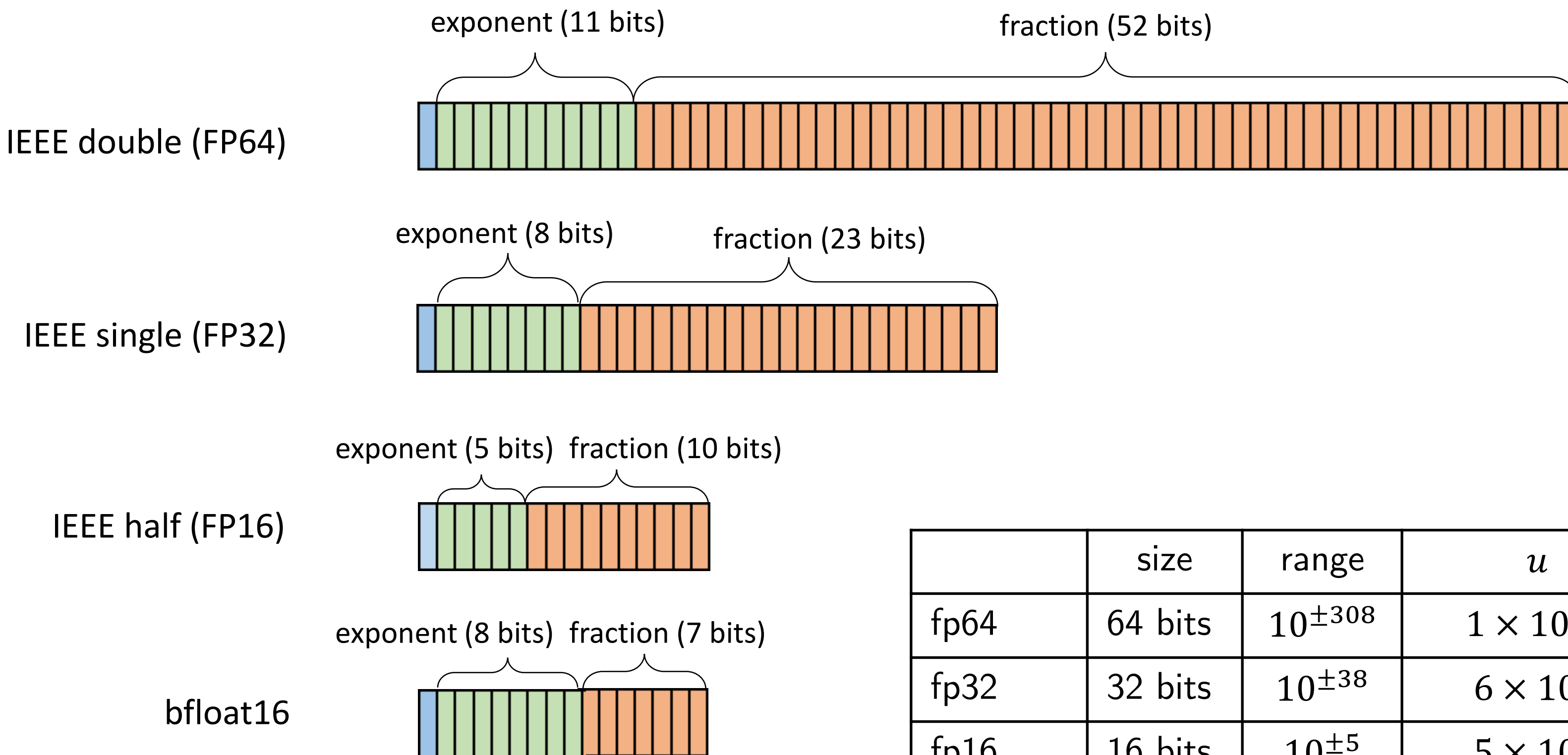
December 3, 2021

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

We acknowledge funding from Charles Univ. PRIMUS project No. PRIMUS/19/SCI/11, Charles Univ. Research Program No. UNCE/SCI/023, and the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Admin.

Floating Point Formats

$$(-1)^{\text{sign}} \times 2^{(\text{exponent}-\text{offset})} \times 1.\text{fraction}$$



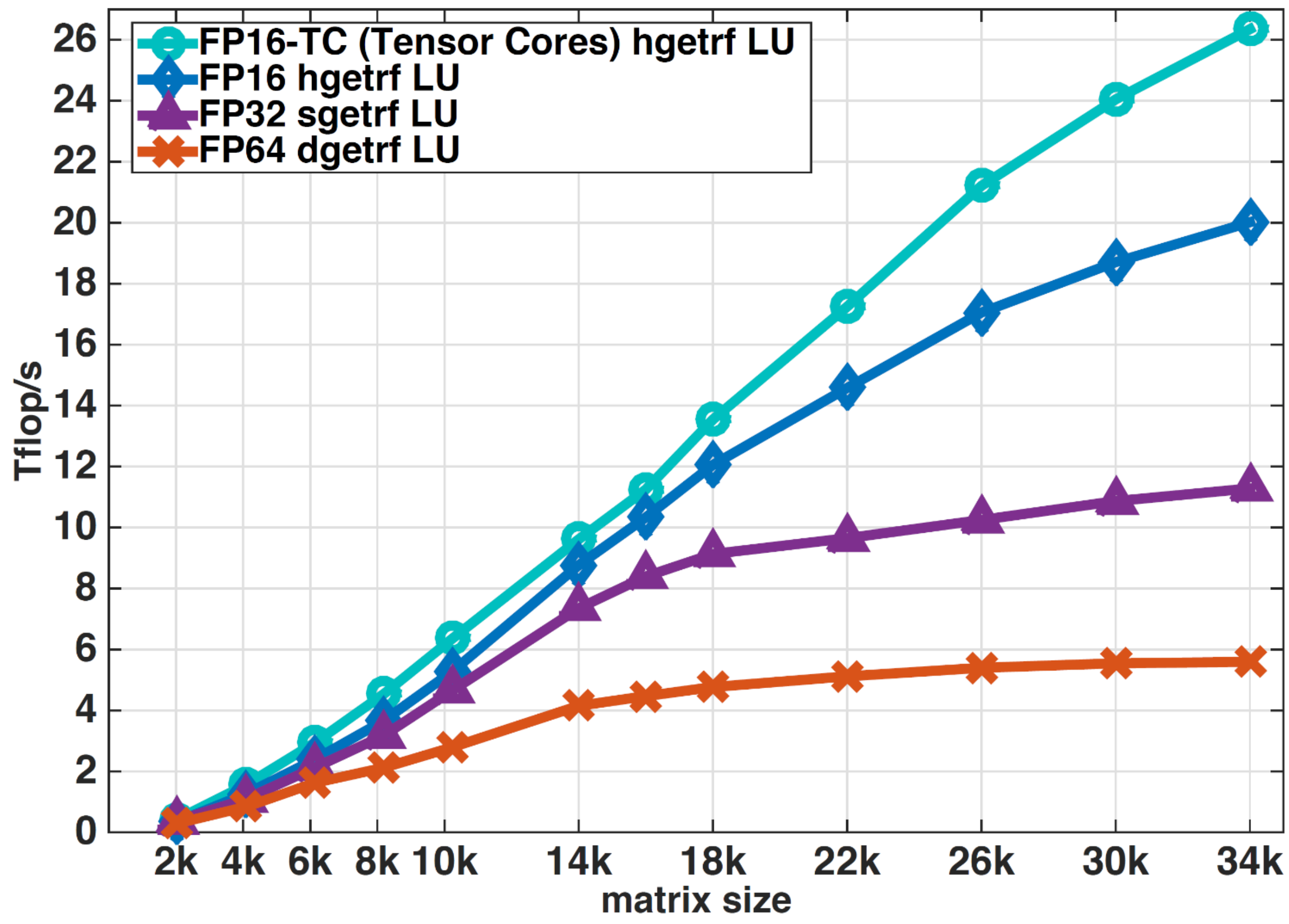
	size	range	u
fp64	64 bits	$10^{\pm 308}$	1×10^{-16}
fp32	32 bits	$10^{\pm 38}$	6×10^{-8}
fp16	16 bits	$10^{\pm 5}$	5×10^{-4}
bfloat16	16 bits	$10^{\pm 38}$	4×10^{-3}

Hardware Support for Multiprecision Computation

Use of low precision in machine learning has driven emergence of low-precision capabilities in hardware:

- AMD Radeon Instinct MI25 GPU, 2017:
  - single: 12.3 TFLOPS, half: 24.6 TFLOPS
- NVIDIA Tesla P100, 2016: native ISA support for 16-bit FP arithmetic
- NVIDIA Tesla V100, 2017: tensor cores for half precision;
  4x4 matrix multiply in one clock cycle
  - double: 7 TFLOPS, half+tensor: 112 TFLOPS (**16x!**)
- NVIDIA A100, 2020: tensor cores with multiple supported precisions: FP16, FP64, Binary, INT4, INT8, bfloat16
- Intel AI processors (Nervana, Xeon)
- Google's Tensor processing unit (TPU): as low as 8-bit arithmetic, bfloat16
- Future exascale supercomputers: (~2021) Expected extensive support for reduced-precision arithmetic (32/16/8-bit)

Performance of LU factorization on an NVIDIA V100 GPU



[Haidar, Tomov, Dongarra, Higham, 2018]

Mixed Precision Capabilities on Supercomputers

From TOP500:

June 2021

	Accelerator/CP Family	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
1	NVIDIA Volta	97	19.4	626,503,420	1,049,977,600	11,875,056
2	NVIDIA Ampere	26	5.2	351,252,600	505,841,268	3,435,116
3	NVIDIA Pascal	9	1.8	57,876,640	85,807,525	1,141,300

Mixed Precision Capabilities on Supercomputers

From TOP500:

June 2021

	Accelerator/CP Family	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
1	NVIDIA Volta	97	19.4	626,503,420	1,049,977,600	11,875,056
2	NVIDIA Ampere	26	5.2	351,252,600	505,841,268	3,435,116
3	NVIDIA Pascal	9	1.8	57,876,640	85,807,525	1,141,300

June 2019

	Accelerator/CP Family	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
1	NVIDIA Pascal	61	12.2	106,025,166	179,951,012	2,738,356
3	NVIDIA Volta	12	2.4	224,559,400	360,593,742	4,488,720

HPL-AI Benchmark

- Highlights confluence of HPC+AI workloads
  - Like HPL, solves dense Ax=b, results still to double precision accuracy
  - Achieves this via mixed-precision iterative refinement
    - may be implemented in a way that takes advantage of the current and upcoming devices for accelerating AI workloads

HPL-AI Benchmark

- Highlights confluence of HPC+AI workloads
  - Like HPL, solves dense Ax=b, results still to double precision accuracy
  - Achieves this via mixed-precision iterative refinement
    - may be implemented in a way that takes advantage of the current and upcoming devices for accelerating AI workloads

- HPL-AI Results (June 2021):
  1. Fugaku: 2 EXAFLOP/s (vs. 442 PETAFLOP/s on HPL; 4.5x)
  2. Summit: 1.15 EXAFLOP/s (vs. 149 PETAFLOP/s on HPL; 7.7x)

- More information: https://icl.bitbucket.io/hpl-ai/
- Reference implementation: https://bitbucket.org/icl/hpl-ai/src/

Mixed precision in NLA

- BLAS: cuBLAS, MAGMA, [Agullo et al. 2009], [Abdelfattah et al., 2019], [Haidar et al., 2018]
- Iterative refinement:
  - Long history: [Wilkinson, 1963], [Moler, 1967], [Stewart, 1973], ...
  - More recently: [Langou et al., 2006], [C., Higham, 2017], [C., Higham, 2018], [C., Higham, Pranesh, 2020], [Amestoy et al., 2021]
- Matrix factorizations: [Haidar et al., 2017], [Haidar et al., 2018], [Haidar et al., 2020], [Abdelfattah et al., 2020]
- Eigenvalue problems: [Dongarra, 1982], [Dongarra, 1983], [Tisseur, 2001], [Davies et al., 2001], [Petschow et al., 2014], [Alvermann et al., 2019]
- Sparse direct solvers: [Buttari et al., 2008]
- Orthogonalization: [Yamazaki et al., 2015]
- Multigrid: [Tamstorf et al., 2020], [Richter et al., 2014], [Sumiyoshi et al., 2014], [Ljungkvist, Kronbichler, 2017, 2019]
- (Preconditioned) Krylov subspace methods: [Emans, van der Meer, 2012], [Yamagishi, Matsumura, 2016], [C., Gergelits, Yamazaki, 2021], [Clark, 2019], [Anzt et al., 2019], [Clark et al., 2010], [Gratton et al., 2020], [Arioli, Duff, 2009], [Hogg, Scott, 2010]

For survey and references, see [Abdelfattah et al., IJHPC, 2021]

Challenges of low precision

- Do error bounds still apply?
  - Error bound with constant nu provides no information if $nu > 1$
  - One solution: probabilistic approach [Higham, Mary, 2019], [Higham, Mary, 2020]

Challenges of low precision

- Do error bounds still apply?
  - Error bound with constant nu provides no information if $nu > 1$
  - One solution: probabilistic approach [Higham, Mary, 2019], [Higham, Mary, 2020]

- Smaller range of representable numbers
  - Limited range of lower precision might cause overflow when rounding
  - Quantities rounded to lower precision may lose important numerical properties (e.g., positive definiteness)
  - One solution: scaling and shifting approach [Higham, Pranesh, 2019]

Challenges of low precision

- Do error bounds still apply?
  - Error bound with constant nu provides no information if $nu > 1$
  - One solution: probabilistic approach [Higham, Mary, 2019], [Higham, Mary, 2020]

- Smaller range of representable numbers
  - Limited range of lower precision might cause overflow when rounding
  - Quantities rounded to lower precision may lose important numerical properties (e.g., positive definiteness)
  - One solution: scaling and shifting approach [Higham, Pranesh, 2019]

- Larger unit roundoff
  - Lose something small when storing: $fl(x) = x(1+\delta), \;\; |\delta| \leq u$
  - Lose something small when computing: $fl(x \text{ op } y) = (x \text{ op } y)(1+\delta), \;\; |\delta| \leq u$

Challenges of low precision

- Do error bounds still apply?
  - Error bound with constant nu provides no information if $nu > 1$
  - One solution: probabilistic approach [Higham, Mary, 2019], [Higham, Mary, 2020]

- Smaller range of representable numbers
  - Limited range of lower precision might cause overflow when rounding
  - Quantities rounded to lower precision may lose important numerical properties (e.g., positive definiteness)
  - One solution: scaling and shifting approach [Higham, Pranesh, 2019]

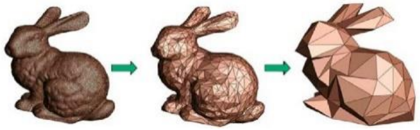
- Larger unit roundoff
  - Lose something small when storing: $fl(x) = x(1+\delta), \;\; |\delta| \leq u$
  - Lose something small when computing: $fl(x \text{ op } y) = (x \text{ op } y)(1+\delta), \;\; |\delta| \leq u$

Does it matter?

Inexact computations

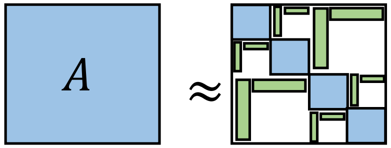
- In real computations we have many sources of inexactness
  - Imperfect data, measurement error
  - Modeling error, discretization error
  - Intentional approximation to improve performance
    - Reduced models, Low-rank representations, sparsification, randomization

Model Reduction

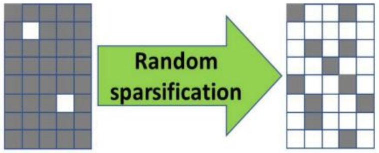


[Schilders, van der Vorst, Rommes, 2008]

Low-rank (hierarchical) approximation



Sparsification, Randomized algorithms

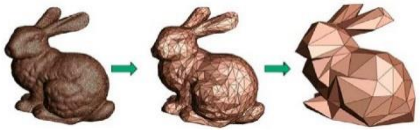


[Sinha, 2018]

Inexact computations

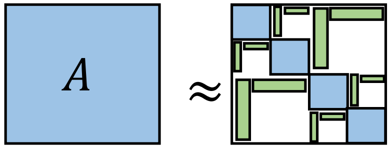
- In real computations we have many sources of inexactness
  - Imperfect data, measurement error
  - Modeling error, discretization error
  - Intentional approximation to improve performance
    - Reduced models, Low-rank representations, sparsification, randomization

Model Reduction

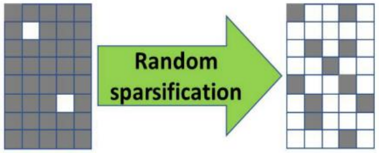


[Schilders, van der Vorst, Rommes, 2008]

Low-rank (hierarchical) approximation



Sparsification, Randomized algorithms

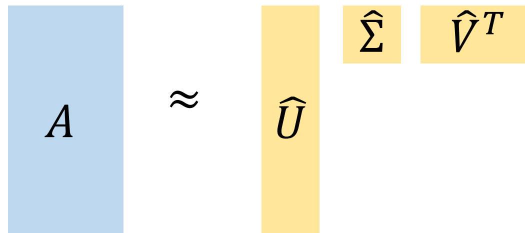


[Sinha, 2018]

- Given that we are already working with so much inexactness, does it matter if we use lower precision?
  - Analysis of accuracy in techniques that use intentional approximation ***almost always*** assume that roundoff error is small enough to be ignored
  - Is this true? Is it true even if we use low precision?

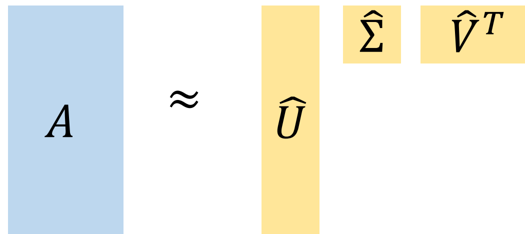
Example: Randomized Algorithms

- Given $m \times n$ A, want truncated SVD with parameter k

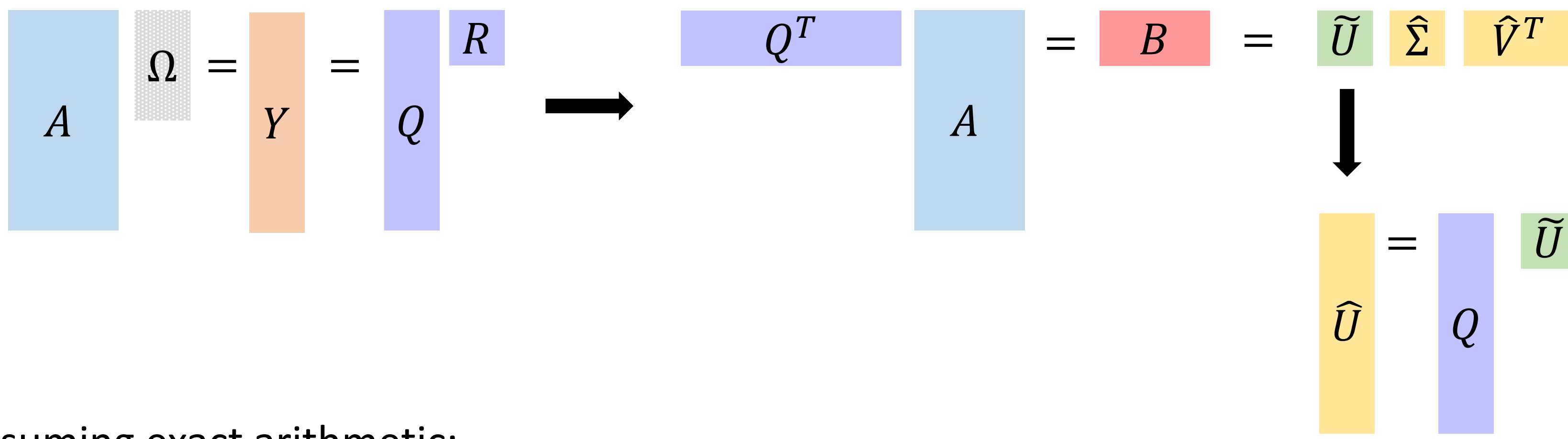


Example: Randomized Algorithms

- Given $m \times n$ A, want truncated SVD with parameter k



- Randomized SVD:



Assuming exact arithmetic:
If Q satisfies $\|A - QQ^T A\| \le \varepsilon$, then $\|A - \hat{U}\hat{\Sigma}\hat{V}^T\| \le \varepsilon$

What happens in finite precision?

Let's try different types of randsvd matrices from the MATLAB gallery:

```
A = gallery('randsvd',[100,40],1e6,mode); k=15;
```

$[U, S, V] \quad = \mathrm{svd}(A)$: non-randomized SVD, exact arithmetic

$[\widehat{U}, \hat{S}, \widehat{V}] \quad = \mathrm{rsvd}(A)$: randomized SVD, exact arithmetic

$[\widehat{U}_d, \hat{S}_d, \widehat{V}_d] = \mathrm{rsvd}(A)$: randomized SVD, double precision

$[\widehat{U}_h, \hat{S}_h, \widehat{V}_h] = \mathrm{rsvd}(A)$: randomized SVD, half precision

What happens in finite precision?

Let's try different types of randsvd matrices from the MATLAB gallery:

```
A = gallery('randsvd',[100,40],1e6,mode); k=15;
```

$[U, S, V] = \text{svd}(A)$: non-randomized SVD, exact arithmetic

$[\hat{U}, \hat{S}, \hat{V}] = \text{rsvd}(A)$: randomized SVD, exact arithmetic

$[\hat{U}_d, \hat{S}_d, \hat{V}_d] = \text{rsvd}(A)$: randomized SVD, double precision

$[\hat{U}_h, \hat{S}_h, \hat{V}_h] = \text{rsvd}(A)$: randomized SVD, half precision

Mode 3: Geometrically distributed singular values

$\|A - USV^T\|_2$ = 4.92e-03

$\|A - \hat{U}\hat{S}\hat{V}^T\|_2$ = 4.92e-03

$\left\|A - \hat{U}_d\hat{S}_d\hat{V}_d{}^T\right\|_2$ = 4.92e-03

$\left\|A - \hat{U}_h\hat{S}_h\hat{V}_h{}^T\right\|_2$ = 4.92e-03

What happens in finite precision?

Let's try different types of randsvd matrices from the MATLAB gallery:

```
A = gallery('randsvd',[100,40],1e6,mode); k=15;
```

$[U, S, V] = \text{svd}(A)$: non-randomized SVD, exact arithmetic

$[\hat{U}, \hat{S}, \hat{V}] = \text{rsvd}(A)$: randomized SVD, exact arithmetic

$[\hat{U}_d, \hat{S}_d, \hat{V}_d] = \text{rsvd}(A)$: randomized SVD, double precision

$[\hat{U}_h, \hat{S}_h, \hat{V}_h] = \text{rsvd}(A)$: randomized SVD, half precision

Mode 3: Geometrically distributed singular values

$\|A - USV^T\|_2$ = 4.92e-03

$\|A - \hat{U}\hat{S}\hat{V}^T\|_2$ = 4.92e-03

$\|A - \hat{U}_d\hat{S}_d{\hat{V}_d}^T\|_2$ = 4.92e-03

$\|A - \hat{U}_h\hat{S}_h{\hat{V}_h}^T\|_2$ = 4.92e-03

Mode 1: one large singular value

$\|A - USV^T\|_2$ = 1.00e-06

$\|A - \hat{U}\hat{S}\hat{V}^T\|_2$ = 1.17e-06

$\|A - \hat{U}_d\hat{S}_d{\hat{V}_d}^T\|_2$ = 1.17e-06

$\|A - \hat{U}_h\hat{S}_h{\hat{V}_h}^T\|_2$ = 1.11e-05

What happens in finite precision?

Let's try different types of randsvd matrices from the MATLAB gallery:

```
A = gallery('randsvd',[100,40],1e6,mode); k=15;
```

$[U, S, V] = \mathrm{svd}(A)$: non-randomized SVD, exact arithmetic

$[\hat{U}, \hat{S}, \hat{V}] = \mathrm{rsvd}(A)$: randomized SVD, exact arithmetic

$[\hat{U}_d, \hat{S}_d, \hat{V}_d] = \mathrm{rsvd}(A)$: randomized SVD, double precision

$[\hat{U}_h, \hat{S}_h, \hat{V}_h] = \mathrm{rsvd}(A)$: randomized SVD, half precision

Mode 3: Geometrically distributed singular values

$$\|A - USV^T\|_2 = 4.92\text{e-}03$$
$$\|A - \hat{U}\hat{S}\hat{V}^T\|_2 = 4.92\text{e-}03$$
$$\|A - \hat{U}_d\hat{S}_d\hat{V}_d^T\|_2 = 4.92\text{e-}03$$
$$\|A - \hat{U}_h\hat{S}_h\hat{V}_h^T\|_2 = 4.92\text{e-}03$$

Mode 1: one large singular value

$$\|A - USV^T\|_2 = 1.00\text{e-}06$$
$$\|A - \hat{U}\hat{S}\hat{V}^T\|_2 = 1.17\text{e-}06$$
$$\|A - \hat{U}_d\hat{S}_d\hat{V}_d^T\|_2 = 1.17\text{e-}06$$
$$\|A - \hat{U}_h\hat{S}_h\hat{V}_h^T\|_2 = 1.11\text{e-}05$$

Use of low precision leads to an order magnitude loss of accuracy! Roundoff error can't be ignored!

What happens in finite precision?

Let's try different types of randsvd matrices from the MATLAB gallery:

```
A = gallery('randsvd',[100,40],1e6,mode); k=15;
```

$[U, S, V] = \mathrm{svd}(A)$: non-randomized SVD, exact arithmetic

$[\hat{U}, \hat{S}, \hat{V}] = \mathrm{rsvd}(A)$: randomized SVD, exact arithmetic

$[\hat{U}_d, \hat{S}_d, \hat{V}_d] = \mathrm{rsvd}(A)$: randomized SVD, double precision

$[\hat{U}_h, \hat{S}_h, \hat{V}_h] = \mathrm{rsvd}(A)$: randomized SVD, half precision

Error bound no longer holds!

Mode 3: Geometrically distributed singular values

$\|A - USV^T\|_2$ = 4.92e-03
$\|A - \hat{U}\hat{S}\hat{V}^T\|_2$ = 4.92e-03
$\left\|A - \hat{U}_d\hat{S}_d{\hat{V}_d}^T\right\|_2$ = 4.92e-03
$\left\|A - \hat{U}_h\hat{S}_h{\hat{V}_h}^T\right\|_2$ = 4.92e-03

Mode 1: one large singular value

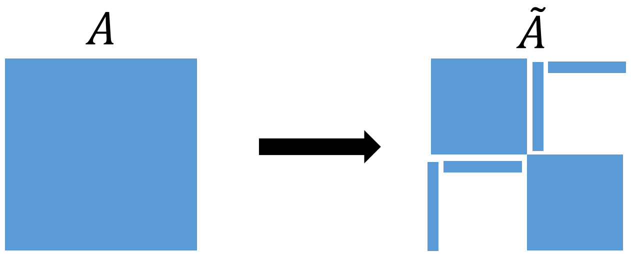
$\|A - USV^T\|_2$ = 1.00e-06
$\|A - \hat{U}\hat{S}\hat{V}^T\|_2$ = 1.17e-06
$\left\|A - \hat{U}_d\hat{S}_d{\hat{V}_d}^T\right\|_2$ = 1.17e-06
$\left\|A - \hat{U}_h\hat{S}_h{\hat{V}_h}^T\right\|_2$ = 1.11e-05

$\|A - Q_h Q_h^T A\|_2$ = 3.59e-06

Use of low precision leads to an order magnitude loss of accuracy! Roundoff error can't be ignored!

Example: Low-Rank Approximation

- Block low-rank approximation and hierarchical matrix representations arise in a variety of applications



- Work on mixed and low precision in block low-rank computations

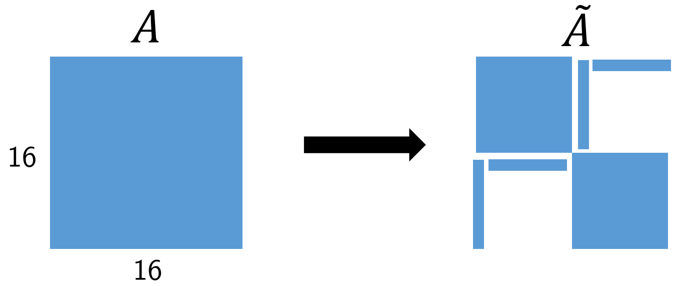
- [Higham, Mary, 2019]: block low-rank LU factorization preconditioner that exploits numerically low-rank structure of the error for LU computed in low precision
- [Higham, Mary, 2019]: Interplay of roundoff error and approximation error in solving block low-rank linear systems using LU
- [Buttari, et al., 2020]: block low-rank single precision coarse grid solves in multigrid
- [Amestoy et al., 2021]: Mixed precision low rank approximation and application to block low-rank LU factorization

Example: Low-Rank Approximation

Inverse multiquadratic kernel:

$$A(i,j) = \frac{1}{\sqrt{1 + 0.1\|x - y\|^2}}, \qquad x, y \in \mathbb{R}^2$$

A is SPD. Low-rank approximation of A should also be SPD!

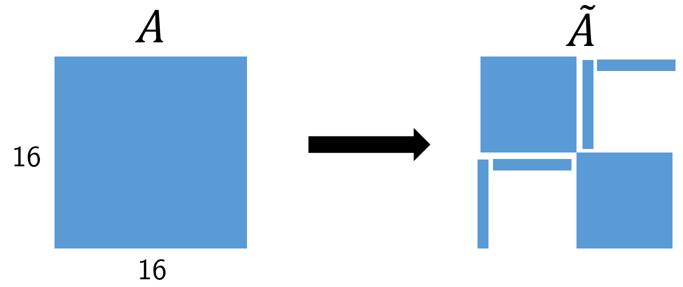


Example: Low-Rank Approximation

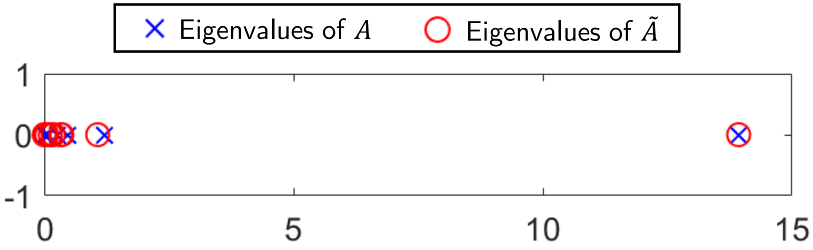
Inverse multiquadratic kernel:

$$A(i,j) = \frac{1}{\sqrt{1 + 0.1\|x - y\|^2}}, \qquad x, y \in \mathbb{R}^2$$

A is SPD. Low-rank approximation of A should also be SPD!



Exact arithmetic SVD:

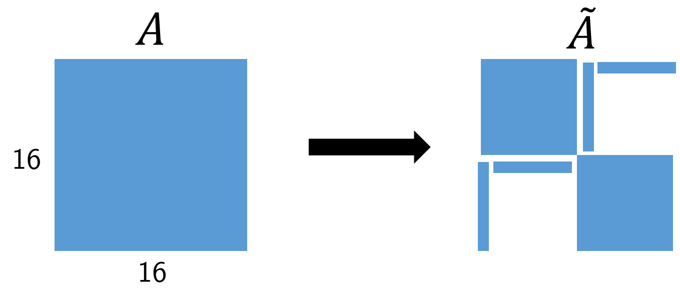


Example: Low-Rank Approximation

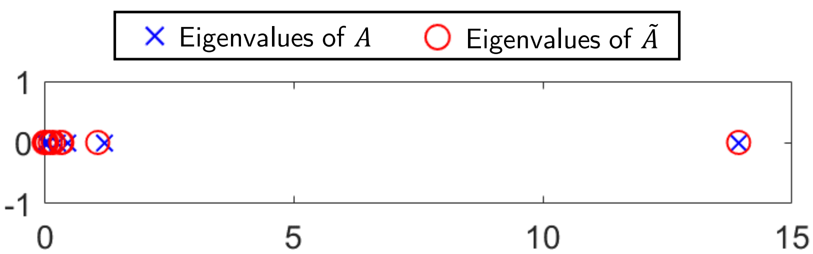
Inverse multiquadratic kernel:

$$A(i,j) = \frac{1}{\sqrt{1 + 0.1\|x - y\|^2}}, \qquad x, y \in \mathbb{R}^2$$

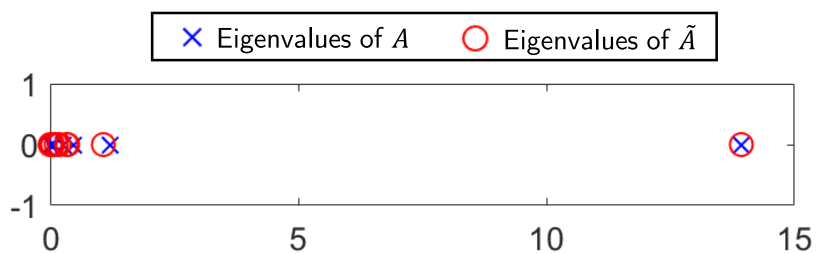
A is SPD. Low-rank approximation of A should also be SPD!



Exact arithmetic SVD:



Half precision SVD:

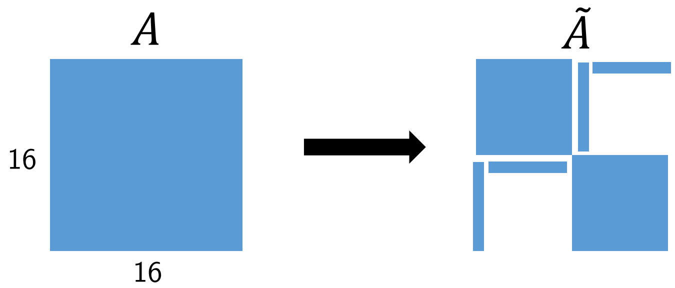


Example: Low-Rank Approximation

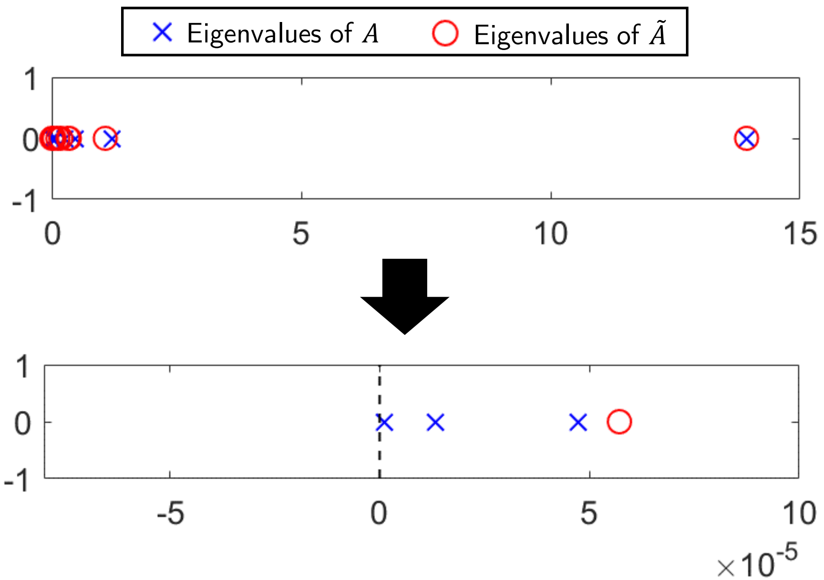
Inverse multiquadratic kernel:

$$A(i,j) = \frac{1}{\sqrt{1 + 0.1\|x - y\|^2}}, \qquad x, y \in \mathbb{R}^2$$

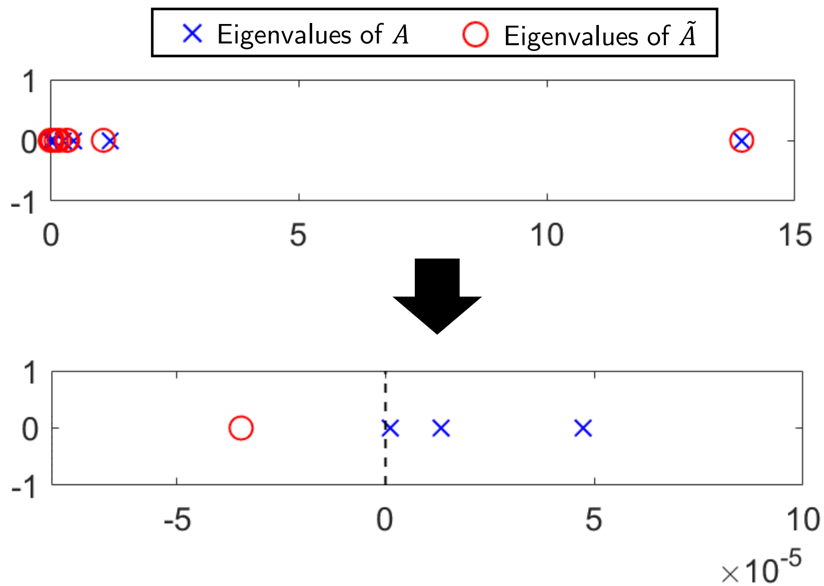
A is SPD. Low-rank approximation of A should also be SPD!



Exact arithmetic SVD:



Half precision SVD:

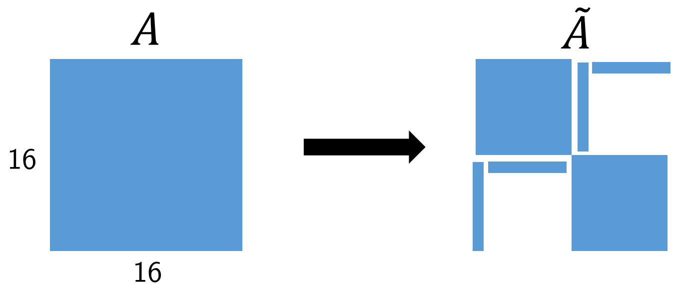


Example: Low-Rank Approximation

Inverse multiquadratic kernel:

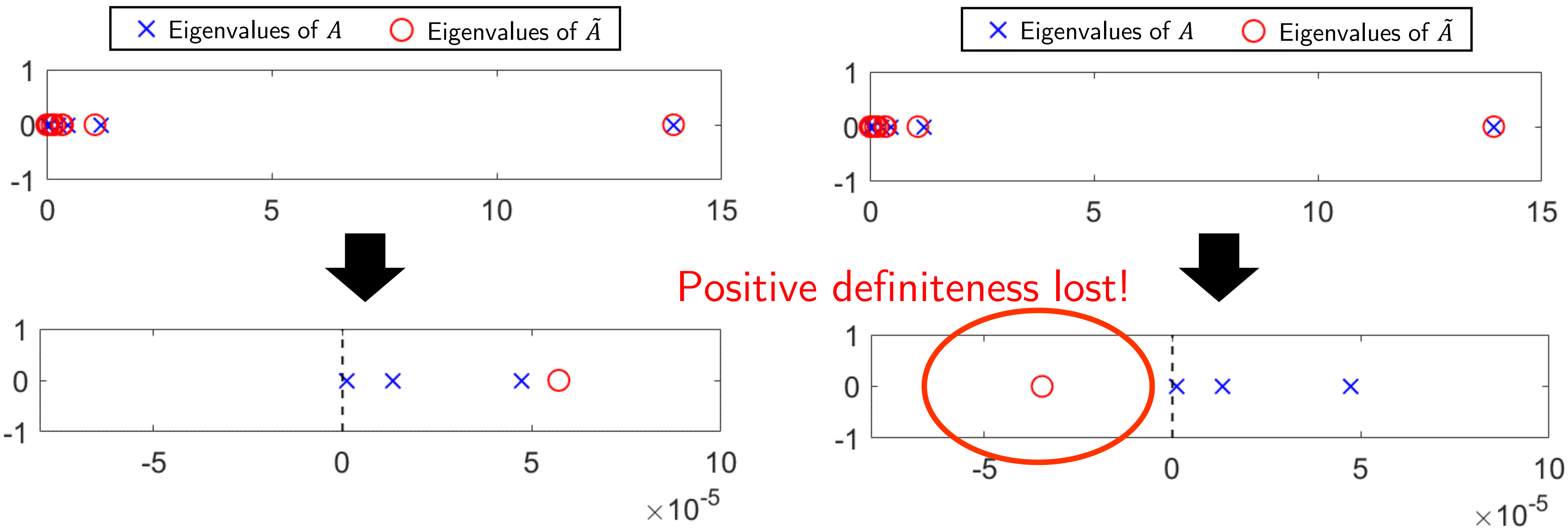
$$A(i,j) = \frac{1}{\sqrt{1 + 0.1\|x - y\|^2}}, \qquad x, y \in \mathbb{R}^2$$

A is SPD. Low-rank approximation of A should also be SPD!



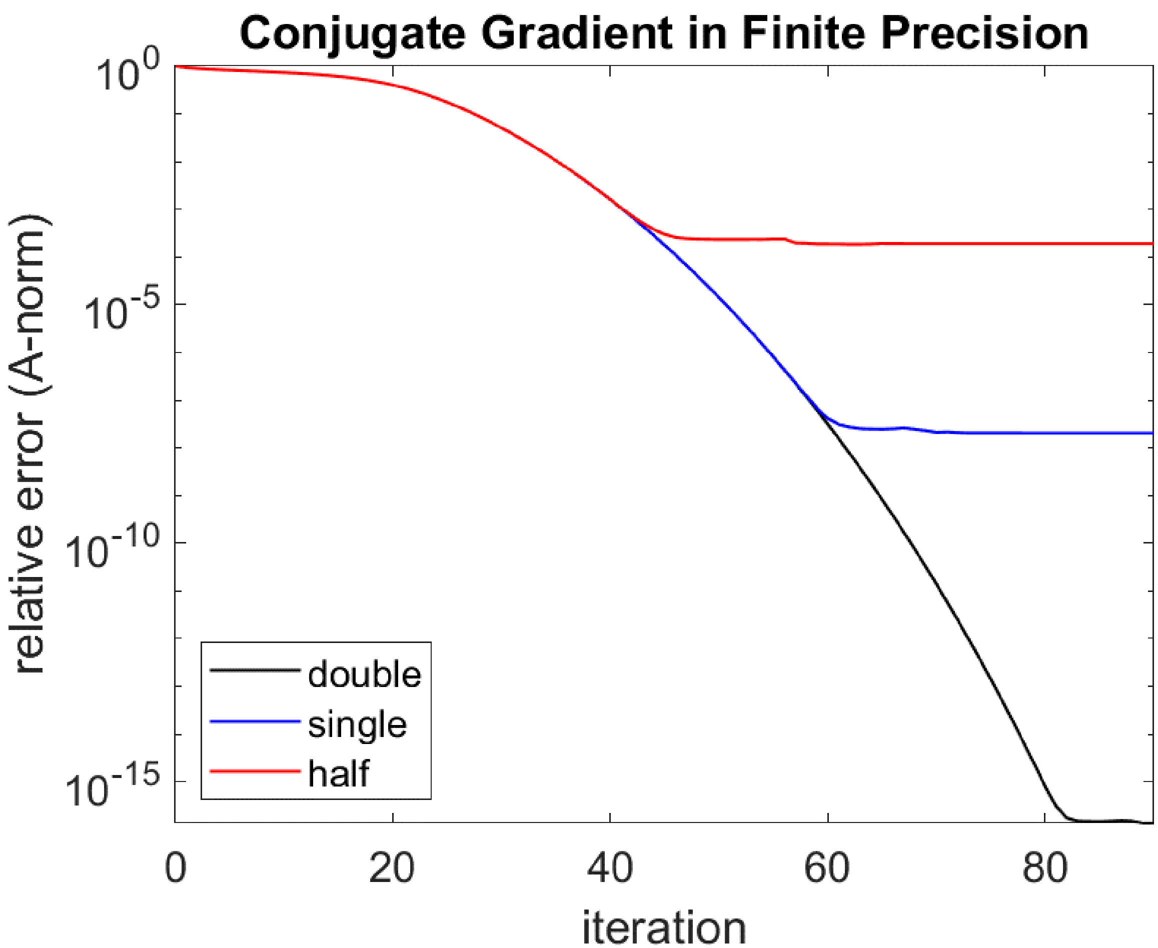
Exact arithmetic SVD:

Half precision SVD:



Example: Iterative Methods

```
A = diag(linspace(.001,1,100));
[V,~] = eig(A);
b = V'*ones(n,1);
```

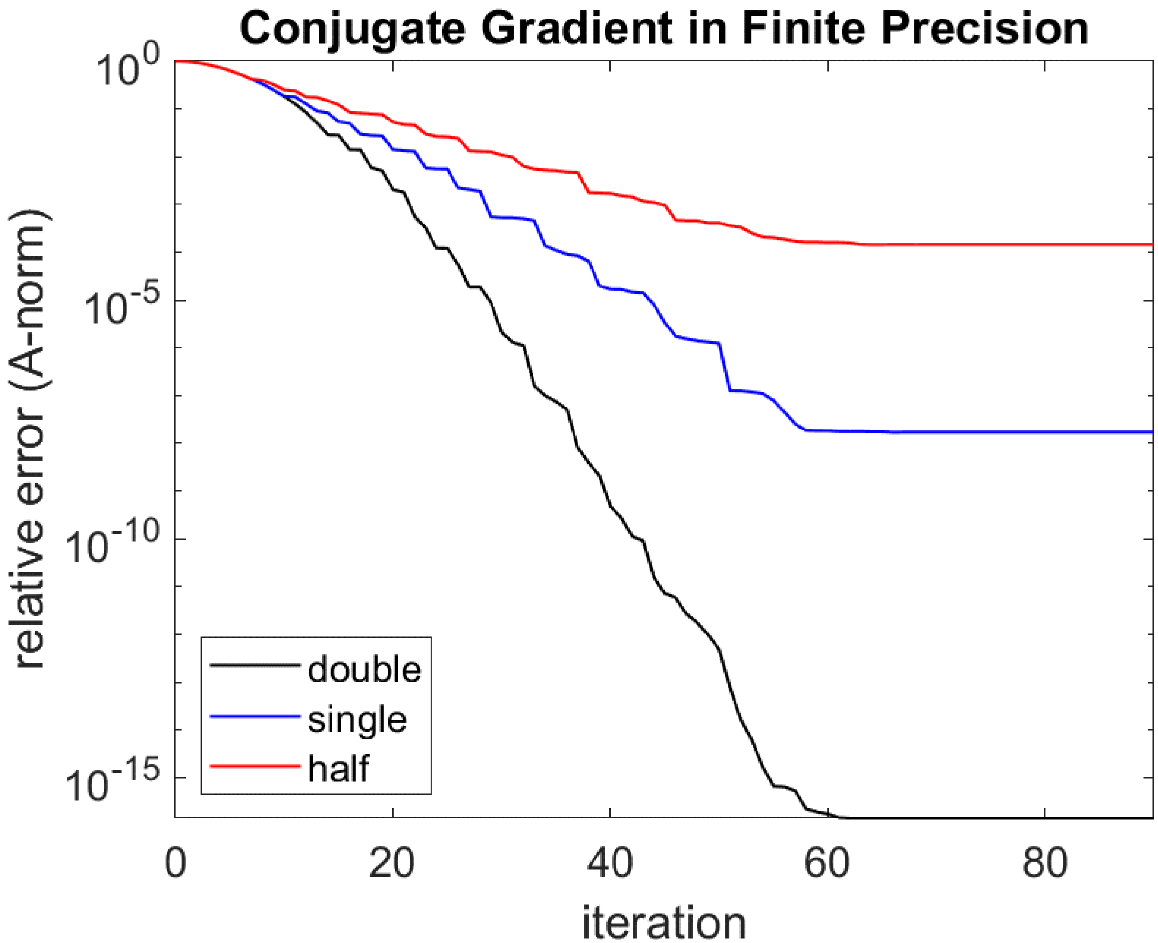


Example: Iterative Methods

$n = 100, \lambda_1 = 10^{-3}, \lambda_n = 1$

$\lambda_i = \lambda_1 + \left(\frac{i-1}{n-1}\right)(\lambda_n - \lambda_1)(0.65)^{n-i}, \quad i = 2, \ldots, n-1$

```
[V,~] = eig(A);
b = V'*ones(n,1);
```



Takeaway

- Low precision can have massive performance benefits but must be used with caution!
- Many opportunities for using mixed and low precision computation in scientific applications
- Need to develop a theoretical understanding of how mixed precision algorithms behave; need to revisit analyses of algorithms and techniques that ignore finite precision

Iterative Refinement for $Ax = b$

Iterative refinement: well-established method for improving an approximate solution to $Ax = b$

A is $n \times n$ and nonsingular; u is unit roundoff

Solve $Ax_0 = b$ by LU factorization (in precision u_f)

for $i = 0\colon \mathrm{maxit}$

$r_i = b - Ax_i$ (in precision u_r)

Solve $Ad_i = r_i$ (in precision u_s)

$x_{i+1} = x_i + d_i$ (in precision u)

Iterative Refinement in 3 Precisions

- 3-precision iterative refinement [C. and Higham, 2018]

u_f = factorization precision, u = working precision, u_r = residual precision

$$u_f \geq u \geq u_r$$

u_s is the *effective precision* of the solve, with $u \leq u_s \leq u_f$

- For triangular solves with LU factors: $u_s = u_f$
- For GMRES preconditioned by LU factors, $u_s = u$ [C. and Higham, 2017]

- New analysis **generalizes** existing types of IR:

Traditional	$u_f = u, u_r = u^2$
Fixed precision	$u_f = u = u_r$
Lower precision factorization	$u_f^2 = u = u_r$

- Enables **new** types of IR: (half, single, double), (half, single, quad), (half, double, quad), etc.

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
H	S	S	10^4	10^{-8}	10^{-8}	$\mathrm{cond}(A,x) \cdot 10^{-8}$
H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
H	D	D	10^4	10^{-16}	10^{-16}	$\mathrm{cond}(A,x) \cdot 10^{-16}$
H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
S	S	S	10^8	10^{-8}	10^{-8}	$\mathrm{cond}(A,x) \cdot 10^{-8}$
S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
S	D	D	10^8	10^{-16}	10^{-16}	$\mathrm{cond}(A,x) \cdot 10^{-16}$
S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error: norm	Backward error: comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\mathrm{cond}(A,x)\cdot 10^{-8}$
	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\mathrm{cond}(A,x)\cdot 10^{-16}$
	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
	S	S	S	10^8	10^{-8}	10^{-8}	$\mathrm{cond}(A,x)\cdot 10^{-8}$
	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\mathrm{cond}(A,x)\cdot 10^{-16}$
	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A,x)\cdot 10^{-8}$
	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A,x)\cdot 10^{-16}$
	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A,x)\cdot 10^{-8}$
	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A,x)\cdot 10^{-16}$
	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A,x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A,x) \cdot 10^{-16}$
	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error: norm	Backward error: comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\mathrm{cond}(A,x)\cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\mathrm{cond}(A,x)\cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\mathrm{cond}(A,x)\cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\mathrm{cond}(A,x)\cdot 10^{-16}$
New	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error: norm	Backward error: comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\mathrm{cond}(A,x)\cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\mathrm{cond}(A,x)\cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\mathrm{cond}(A,x)\cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\mathrm{cond}(A,x)\cdot 10^{-16}$
New	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

$\Rightarrow$ Benefit of IR3 vs. "LP fact.": no $\mathrm{cond}(A,x)$ term in forward error

IR3: Summary

Standard (LU-based) IR in three precisions ($u_s = u_f$)

Half $\approx 10^{-4}$, Single $\approx 10^{-8}$, Double $\approx 10^{-16}$, Quad $\approx 10^{-34}$

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\mathrm{cond}(A,x) \cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\mathrm{cond}(A,x) \cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\mathrm{cond}(A,x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\mathrm{cond}(A,x) \cdot 10^{-16}$
New	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

$\Rightarrow$ Benefit of IR3 vs. traditional IR: As long as $\kappa_\infty(A) \leq 10^4$, can use lower precision factorization w/no loss of accuracy!

GMRES-IR: Summary

GMRES-based IR in three precisions ($u_s = u$)

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
GMRES-IR	H	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LU-IR	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	S	D	Q	10^{16}	10^{-16}	10^{-16}	10^{-16}
LU-IR	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	H	D	Q	10^{12}	10^{-16}	10^{-16}	10^{-16}

⇒With GMRES-IR, lower precision factorization will work for higher $\kappa_\infty(A)$

GMRES-IR: Summary

GMRES-based IR in three precisions ($u_s = u$)

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error: norm	Backward error: comp	Forward error
LU-IR	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
GMRES-IR	H	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LU-IR	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	S	D	Q	10^{16}	10^{-16}	10^{-16}	10^{-16}
LU-IR	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	H	D	Q	10^{12}	10^{-16}	10^{-16}	10^{-16}

→ $\kappa_\infty(A) \leq u^{-1/2} u_f^{-1}$

⇒With GMRES-IR, lower precision factorization will work for higher $\kappa_\infty(A)$

GMRES-IR: Summary

GMRES-based IR in three precisions ($u_s = u$)

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
GMRES-IR	H	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LU-IR	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	S	D	Q	10^{16}	10^{-16}	10^{-16}	10^{-16}
LU-IR	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	H	D	Q	10^{12}	10^{-16}	10^{-16}	10^{-16}

$\kappa_\infty(A) \le u^{-1/2} u_f^{-1}$

⇒ As long as $\kappa_\infty(A) \le 10^{12}$, can use half precision factorization and still obtain double precision accuracy!

GMRES-IR: Summary

GMRES-based IR in three precisions ($u_s = u$)

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
GMRES-IR	H	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LU-IR	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	S	D	Q	10^{16}	10^{-16}	10^{-16}	10^{-16}
LU-IR	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	H	D	Q	10^{12}	10^{-16}	10^{-16}	10^{-16}

$\kappa_\infty(A) \leq u^{-1/2}\, u_f^{-1}$

⇒ As long as $\kappa_\infty(A) \leq 10^{12}$, can use half precision factorization and still obtain double precision accuracy!

Recent work: 5-precision GMRES-IR [Amestoy, et al., 2021]

GMRES-IR: Summary

GMRES-based IR in three precisions ($u_s = u$)

	u_f	u	u_r	max $\kappa_\infty(A)$	Backward error norm	Backward error comp	Forward error
LU-IR	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
GMRES-IR	H	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LU-IR	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	S	D	Q	10^{16}	10^{-16}	10^{-16}	10^{-16}
LU-IR	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
GMRES-IR	H	D	Q	10^{12}	10^{-16}	10^{-16}	10^{-16}

→ $\kappa_\infty(A) \leq u^{-1/2}\, u_f^{-1}$

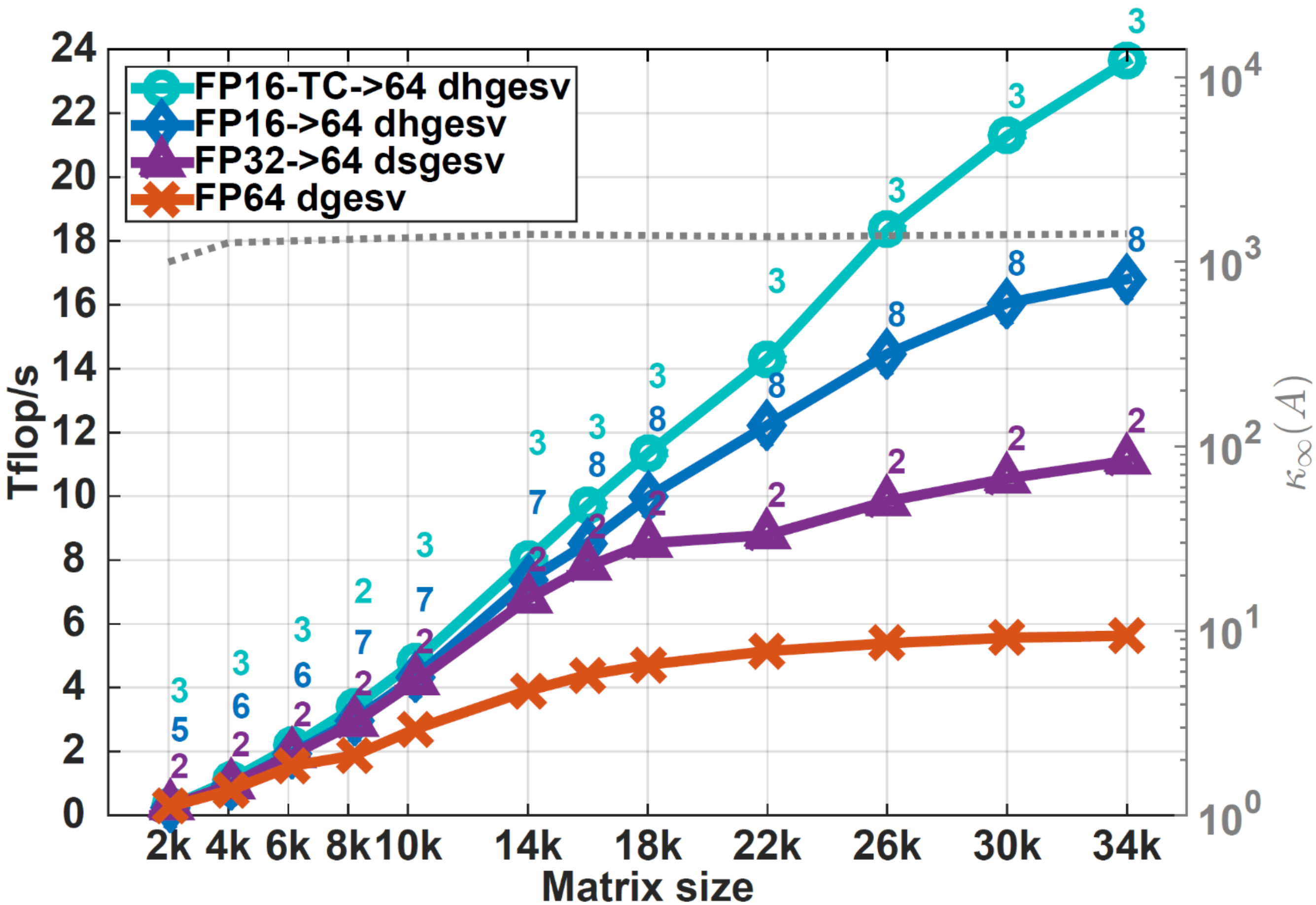
⇒ As long as $\kappa_\infty(A) \leq 10^{12}$, can use half precision factorization and still obtain double precision accuracy!

Recent work: 5-precision GMRES-IR [Amestoy, et al., 2021]

→ $\kappa_\infty(A) \leq u^{-1/3}\, u_f^{-2/3}$

Performance Results (MAGMA)

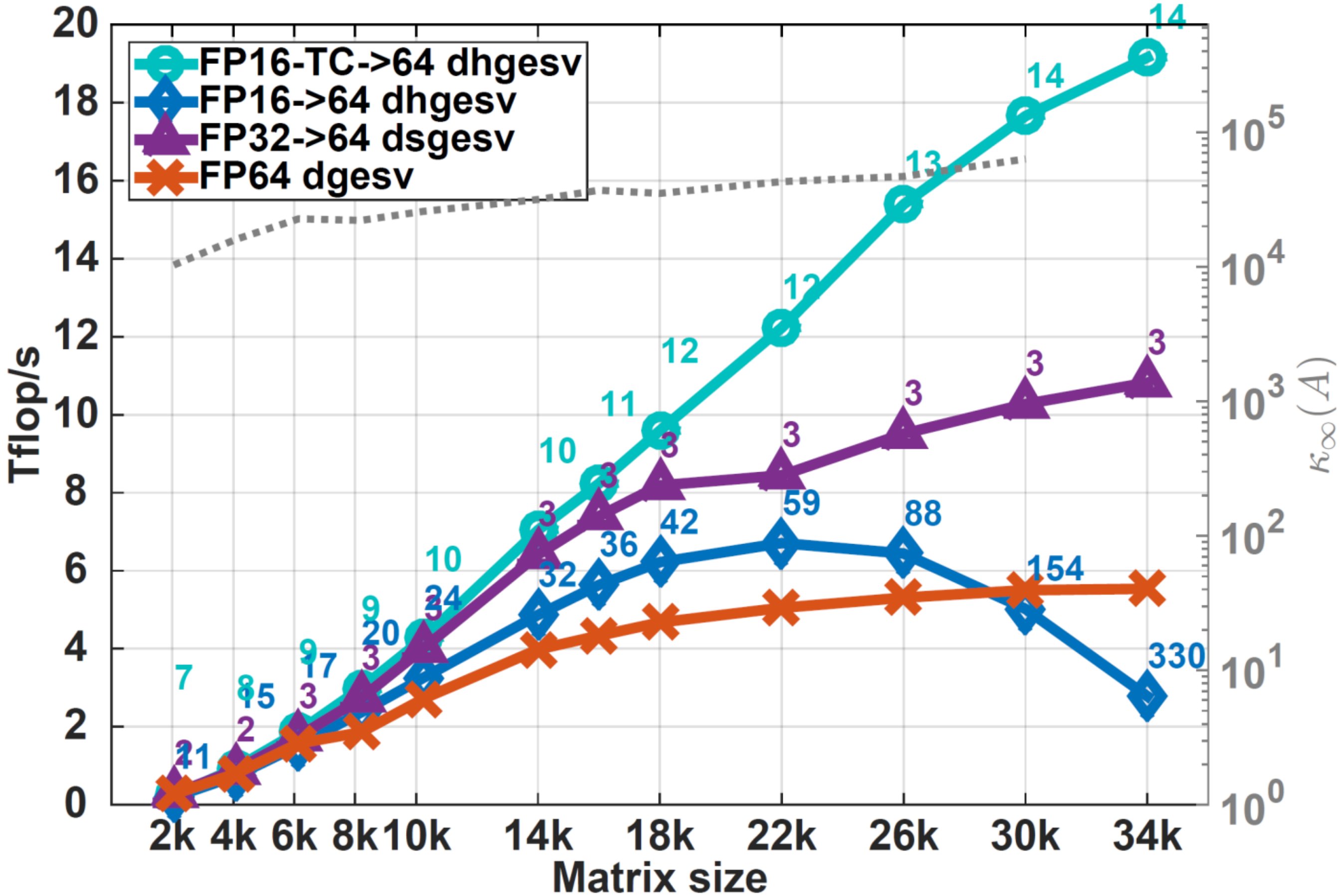
- [Haidar, Tomov, Dongarra, Higham, 2018]



(a) Matrix of type 3: positive λ with clustered singular values, σ_i=(1, $\cdots$, 1, $\frac{1}{cond}$).

Performance Results (MAGMA)

- [Haidar, Tomov, Dongarra, Higham, 2018]



(b) Matrix of type 4: clustered singular values, $\sigma_i=(1, \cdots, 1, \frac{1}{cond})$.

The rise of multiprecision hardware

- Future machines will support a range of precisions: quarter, half, single, double, quad

The rise of multiprecision hardware

- Future machines will support a range of precisions: quarter, half, single, double, quad
- New, non-IEEE compliant floating point formats will appear in commercially-available hardware
  - e.g., bfloat16 (truncated 16-bit version of single precision), posits

The rise of multiprecision hardware

- Future machines will support a range of precisions: quarter, half, single, double, quad
- New, non-IEEE compliant floating point formats will appear in commercially-available hardware
    - e.g., bfloat16 (truncated 16-bit version of single precision), posits
- Lower-precision arithmetic is faster and more energy efficient, but the potential for its use depends heavily on the particular problem and algorithm

The rise of multiprecision hardware

- Future machines will support a range of precisions: quarter, half, single, double, quad

- New, non-IEEE compliant floating point formats will appear in commercially-available hardware
  - e.g., bfloat16 (truncated 16-bit version of single precision), posits

- Lower-precision arithmetic is faster and more energy efficient, but the potential for its use depends heavily on the particular problem and algorithm

- As numerical analysts, we must determine when and where we can exploit lower-precision hardware to improve performance

Thank you!

carson@karlin.mff.cuni.cz

www.karlin.mff.cuni.cz/~carson/