

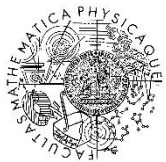
High-Performance Mixed Precision Numerical Linear Algebra

Erin C. Carson

Faculty of Mathematics and Physics, Charles University

October 29, 2020

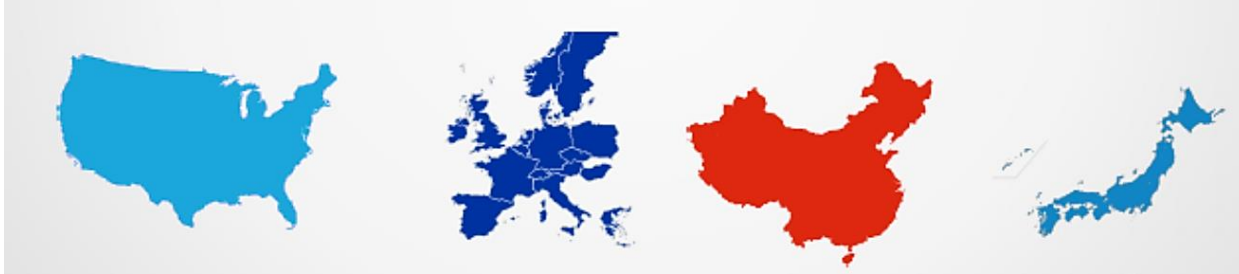
NUMA Seminar, KU Leuven



FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

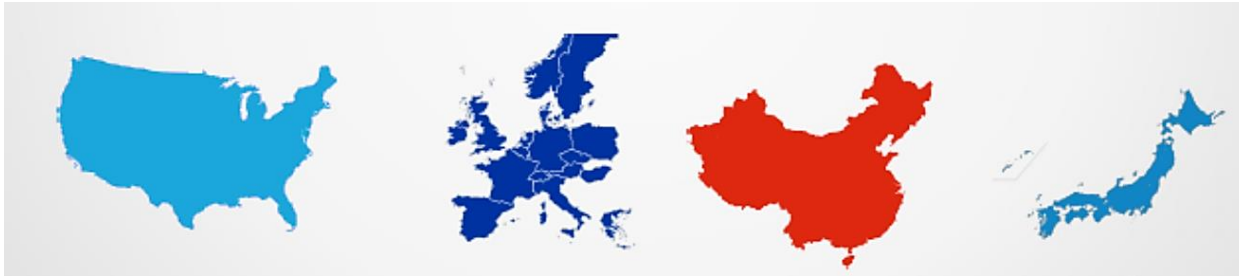
Exascale Computing: The Modern Space Race

- "Exascale": 10^{18} floating point operations per second
 - with maximum energy consumption around 20-40 MWatts
- Large investment in HPC worldwide



Exascale Computing: The Modern Space Race

- "Exascale": 10^{18} floating point operations per second
 - with maximum energy consumption around 20-40 MWatts
- Large investment in HPC worldwide

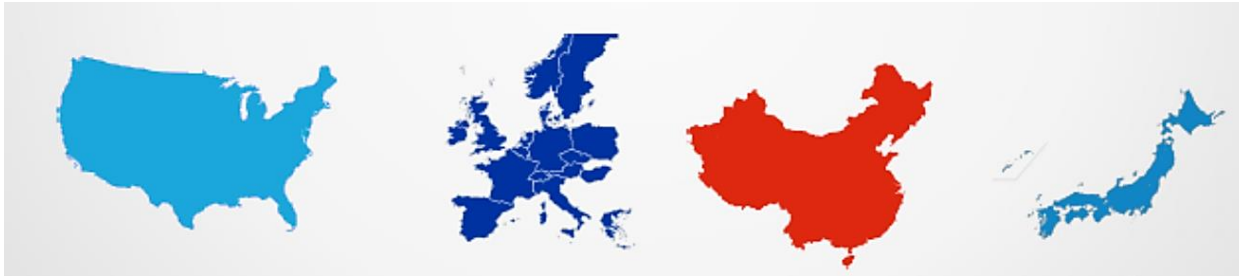


- Technical challenges at all levels

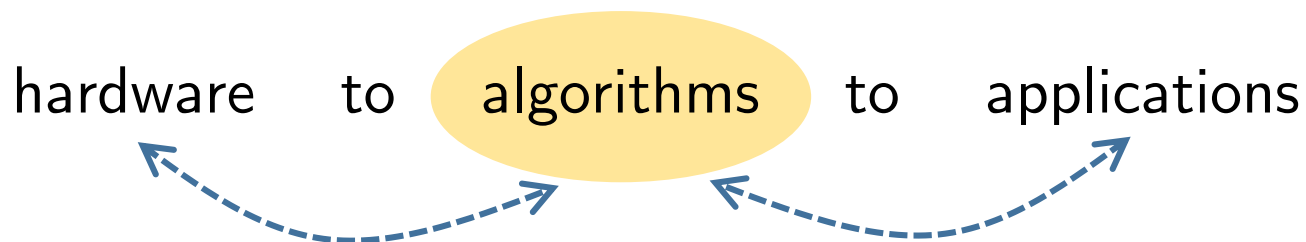
hardware to algorithms to applications

Exascale Computing: The Modern Space Race

- "Exascale": 10^{18} floating point operations per second
 - with maximum energy consumption around 20-40 MWatts
- Large investment in HPC worldwide



- Technical challenges at all levels



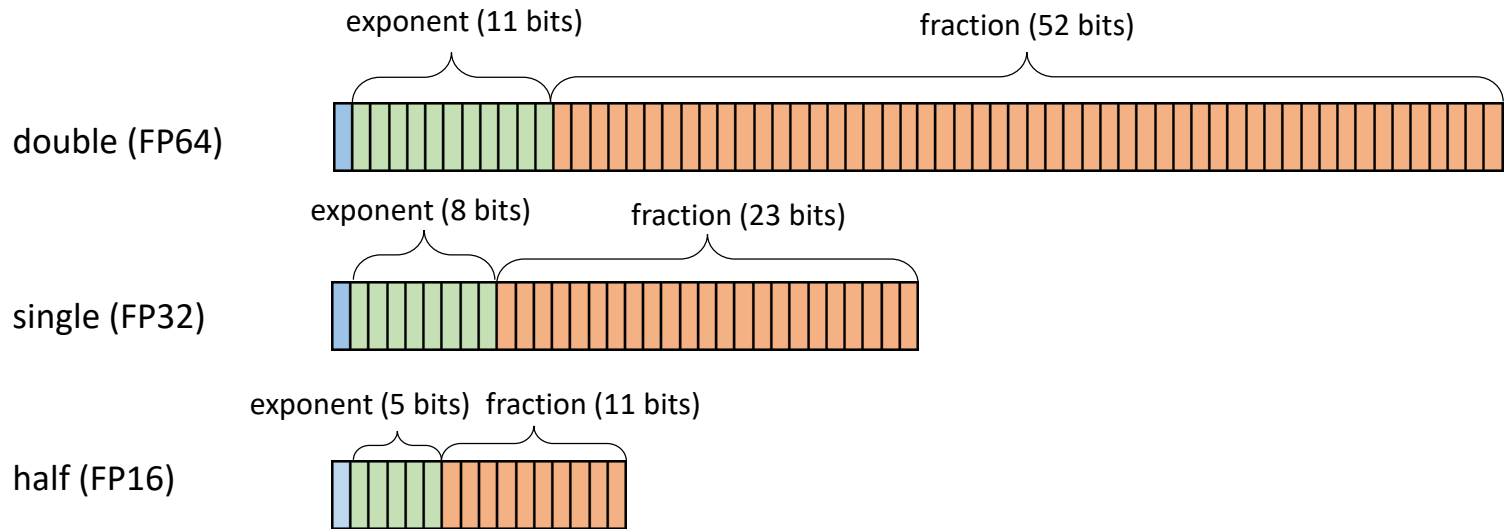
Hardware Support for Multiprecision Computation

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

- Half precision (FP16) defined as storage format in 2008 IEEE standard
- [ARM NEON](#): SIMD architecture, instructions for 8x16-bit, 4x32-bit, 2x64-bit
- [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
- [Google's Tensor processing unit \(TPU\)](#): quantizes 32-bit FP computations into 8-bit integer arithmetic
- [Future exascale supercomputers](#): (~2021) Expected extensive support for reduced-precision arithmetic (32/16/8-bit)

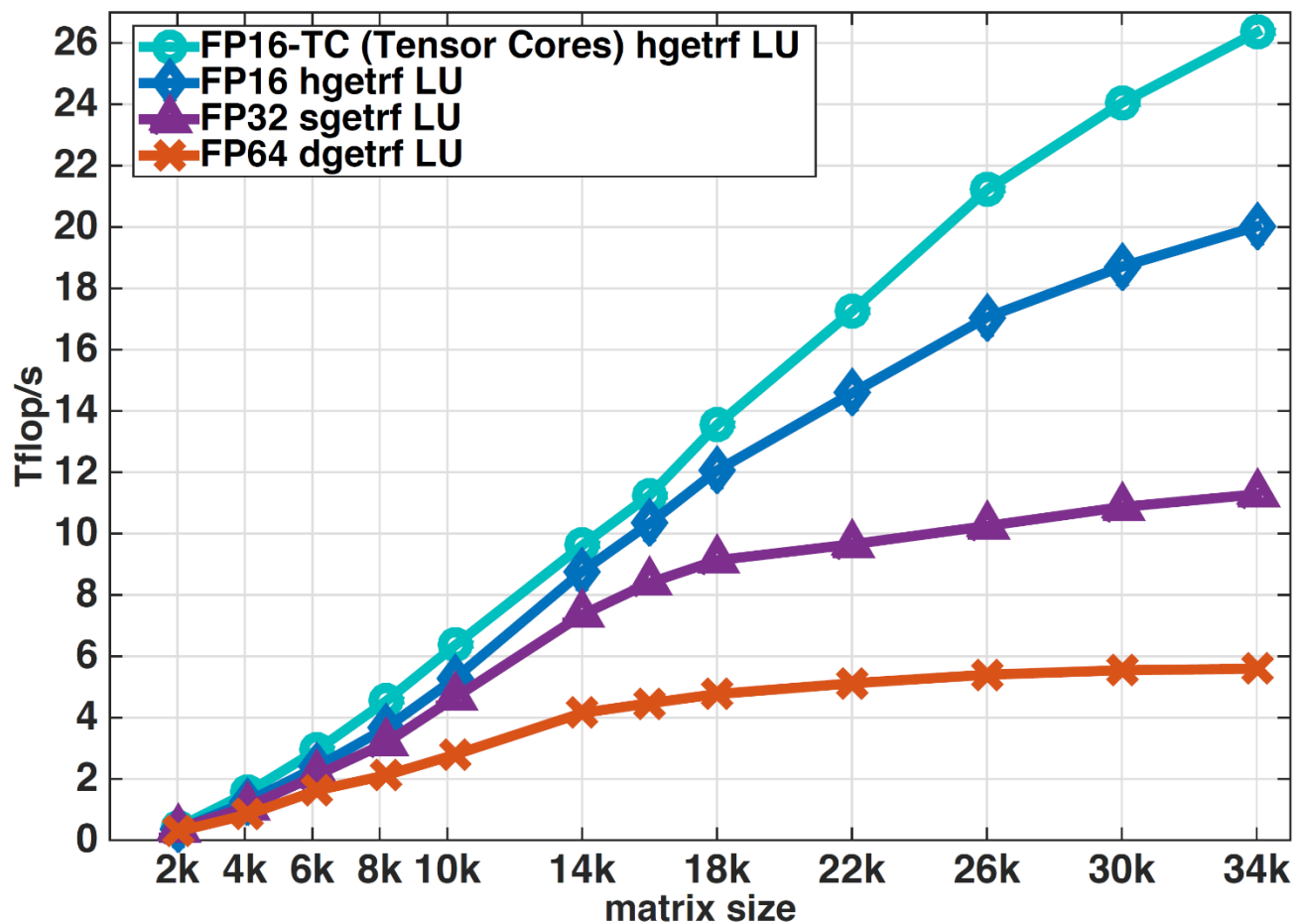
IEEE Floating Point Formats

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



	u	x_{\min}^s	x_{\min}	x_{\max}
fp16	4.88×10^{-4}	5.96×10^{-8}	6.10×10^{-5}	6.55×10^4
fp32	5.96×10^{-8}	1.40×10^{-45}	1.18×10^{-38}	3.40×10^{38}
fp64	1.11×10^{-16}	4.94×10^{-324}	2.22×10^{-308}	1.80×10^{308}

Performance of LU factorization on an NVIDIA V100 GPU



[Haidar, Tomov, Dongarra, Higham, 2018]

Mixed Precision Capabilities on Supercomputers

From TOP500:

June 2020

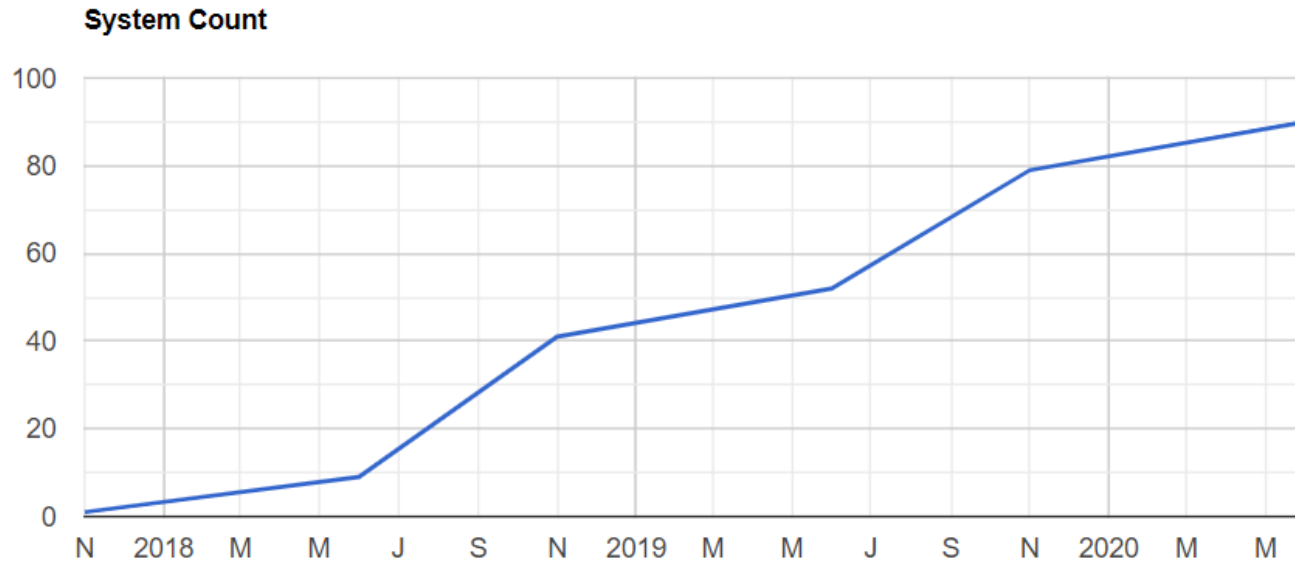
	Accelerator/Co-Processor	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
1	NVIDIA Tesla V100	90	18	249,989,467	447,350,553	5,215,012
2	NVIDIA Tesla P100	17	3.4	63,536,390	99,682,523	1,350,772
3	NVIDIA Tesla V100 SXM2	13	2.6	63,358,310	100,091,657	1,218,784
4	NVIDIA Volta GV100	4	0.8	269,439,000	362,564,722	4,408,096
5	Nvidia Volta V100	3	0.6	24,246,000	33,871,440	451,136

June 2018

	Accelerator/Co-Processor	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
1	NVIDIA Tesla P100	54	10.8	92,474,566	157,608,577	2,141,088
2	NVIDIA Tesla V100	9	1.8	10,769,400	21,164,239	242,016
3	NVIDIA Tesla K80	7	1.4	10,834,746	17,986,269	341,390
4	NVIDIA Tesla K40	6	1.2	11,062,790	18,034,192	246,920
5	NVIDIA Tesla P100 NVLink	5	1	11,716,000	18,611,235	211,268

Mixed Precision Capabilities on Supercomputers

ACCELERATOR/CO-PROCESSOR / NVIDIA TESLA V100



List	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Jun 2020	90	18	249,989,467	447,350,553	5,215,012
Nov 2019	79	15.8	171,425,790	316,852,935	3,676,828
Jun 2019	52	10.4	99,558,070	196,283,310	2,331,252
Nov 2018	41	8.2	65,671,290	136,467,348	1,645,796
Jun 2018	9	1.8	10,769,400	21,164,239	242,016
Nov 2017	1	0.2	1,070,000	1,819,752	22,440

An Exaflop of what?

- When will victory be declared?
 - When a supercomputer reaches exaflop performance on the HPL (LINPACK) benchmark (TOP500)
 - Solving dense $Ax = b$ using Gaussian elimination with partial pivoting in double precision (FP64)

An Exaflop of what?

- When will victory be declared?
 - When a supercomputer reaches exaflop performance on the HPL (LINPACK) benchmark (TOP500)
 - Solving dense $Ax = b$ using Gaussian elimination with partial pivoting in double precision (FP64)
- HPL benchmark is typically a compute-bound problem ("BLAS-3")
- Not a good indication of performance for a large number of applications!
 - Lots of remaining work even after exascale performance is achieved
 - Has led to incorporation of other benchmarks into the TOP500 ranking
 - e.g., HPCG: Solving sparse $Ax = b$ iteratively using the conjugate gradient method

An Exaflop of what?

- HPL doesn't make use of modern mixed precision hardware
- We can *already* achieve “exaflop” performance today if we allow for mixed precision computations



<https://www.olcf.ornl.gov/2018/06/08/genomics-code-exceeds-exaops-on-summit-supercomputer/>

An Exaflop of what?

- HPL doesn't make use of modern mixed precision hardware
- We can *already* achieve “exaflop” performance today if we allow for mixed precision computations



<https://www.olcf.ornl.gov/2018/06/08/genomics-code-exceeds-exaops-on-summit-supercomputer/>

=>HPL-AI: A new mixed precision benchmark

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

for $i = 0: \maxit$

$$r_i = b - Ax_i$$

$$\text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i)$$

$$x_{i+1} = x_i + d_i$$

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)

for $i = 0: \maxit$

$r_i = b - Ax_i$ (in precision u^2)

Solve $Ad_i = r_i$ via $d_i = U^{-1}(L^{-1}r_i)$ (in precision u)

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

"Traditional" (high-precision residual computation)

[Wilkinson, 1948] (fixed point), [Moler, 1967] (floating point)

Iterative Refinement for $Ax = b$

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$$

As long as $\kappa_\infty(A) \leq u^{-1}$,

- relative forward error is $O(u)$
- relative normwise and componentwise backward errors are $O(u)$

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

for $i = 0$: maxit

$$r_i = b - Ax_i \quad (\text{in precision } u^2)$$

$$\text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i) \quad (\text{in precision } u)$$

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

"Traditional" (high-precision residual computation)

[Wilkinson, 1948] (fixed point), [Moler, 1967] (floating point)

Iterative Refinement for $Ax = b$

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

for $i = 0$: maxit

$$r_i = b - Ax_i \quad (\text{in precision } u)$$

$$\text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i) \quad (\text{in precision } u)$$

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

"Fixed-Precision"

[Jankowski and Woźniakowski, 1977], [Skeel, 1980], [Higham, 1991]

Iterative Refinement for $Ax = b$

$$\text{cond}(A, x) = \| |A^{-1}| |A| |x| \|_{\infty} / \|x\|_{\infty}$$

As long as $\kappa_{\infty}(A) \leq u^{-1}$,

- relative forward error is $O(u)\text{cond}(A, x)$
- relative normwise and componentwise backward errors are $O(u)$

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

for $i = 0$: maxit

$$r_i = b - Ax_i \quad (\text{in precision } u)$$

$$\text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i) \quad (\text{in precision } u)$$

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

"Fixed-Precision"

[Jankowski and Woźniakowski, 1977], [Skeel, 1980], [Higham, 1991]

Iterative Refinement for $Ax = b$

Solve $Ax_0 = b$ by LU factorization

(in precision $u^{1/2}$)

for $i = 0$: maxit

$$r_i = b - Ax_i$$

(in precision u)

$$\text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i)$$

(in precision u)

$$x_{i+1} = x_i + d_i$$

(in precision u)

"Low-precision factorization"

[Langou et al., 2006], [Arioli and Duff, 2009], [Hogg and Scott, 2010], [Abdelfattah et al., 2016]

Iterative Refinement for $Ax = b$

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

- relative forward error is $O(u)\text{cond}(A, x)$
- relative normwise and componentwise backward errors are $O(u)$

Solve $Ax_0 = b$ by LU factorization

(in precision $u^{1/2}$)

for $i = 0$: maxit

$$r_i = b - Ax_i$$

(in precision u)

$$\text{Solve } Ad_i = r_i \quad \text{via } d_i = U^{-1}(L^{-1}r_i)$$

(in precision u)

$$x_{i+1} = x_i + d_i$$

(in precision u)

"Low-precision factorization"

[Langou et al., 2006], [Arioli and Duff, 2009], [Hogg and Scott, 2010], [Abdelfattah et al., 2016]

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

⇒ **3-precision iterative refinement**

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

$$u_f \geq u \geq u_r$$

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

⇒ **3-precision iterative refinement**

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

$$u_f \geq u \geq u_r$$

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

[C. and Higham, SIAM SISC 40(2), 2018]

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

(and **improves** upon existing analyses in some cases)

Iterative Refinement in 3 Precisions

Existing analyses only support at most two precisions

Can we combine the performance benefits of low-precision factorization IR with the accuracy of traditional IR?

⇒ **3-precision iterative refinement**

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

$$u_f \geq u \geq u_r$$

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

[C. and Higham, SIAM SISC 40(2), 2018]

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

(and **improves** upon existing analyses in some cases)

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

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$

Define μ_i : $\|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty$

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$

Define μ_i : $\|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty$

For a stable refinement scheme, in early stages we expect

$$\frac{\|r_i\|}{\|A\| \|\hat{x}_i\|} \approx u \ll \frac{\|x - \hat{x}_i\|}{\|x\|} \longrightarrow \mu_i \ll 1$$

Key Analysis Innovations I

Obtain tighter upper bounds:

Typical bounds used in analysis: $\|A(x - \hat{x}_i)\|_\infty \leq \|A\|_\infty \|x - \hat{x}_i\|_\infty$

Define μ_i : $\|A(x - \hat{x}_i)\|_\infty = \mu_i \|A\|_\infty \|x - \hat{x}_i\|_\infty$

For a stable refinement scheme, in early stages we expect

$$\frac{\|r_i\|}{\|A\| \|\hat{x}_i\|} \approx u \ll \frac{\|x - \hat{x}_i\|}{\|x\|} \longrightarrow \mu_i \ll 1$$

But close to convergence,

$$\|r_i\| \approx \|A\| \|x - \hat{x}_i\| \longrightarrow \mu_i \approx 1$$

Key Analysis Innovations II

Allow for general solver:

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

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

example: LU solve:

$$u_s \|E_i\|_\infty \leq 3n u_f \| |A^{-1}| |\hat{L}| |\hat{U}| \|_\infty$$

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

→ normwise relative backward error is at most $\max(c_1, c_2) u_s$

example: LU solve:

$$u_s \|E_i\|_\infty \leq 3n u_f \| |A^{-1}| |\hat{L}| |\hat{U}| \|_\infty$$

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

→ normwise relative backward error is at most $\max(c_1, c_2) u_s$

example: LU solve:

$$u_s \|E_i\|_\infty \leq 3n u_f \| |A^{-1}| |\hat{L}| |\hat{U}| \|_\infty$$

$$\max(c_1, c_2) u_s \leq \frac{3n u_f \| |\hat{L}| |\hat{U}| \|_\infty}{\|A\|_\infty}$$

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

→ normwise relative backward error is at most $\max(c_1, c_2) u_s$

3. $|\hat{r}_i - A\hat{d}_i| \leq u_s G_i |\hat{d}_i|$

→ componentwise relative backward error is bounded by a multiple of u_s

$E_i, c_1, c_2,$ and G_i depend on $A, \hat{r}_i, n,$ and u_s

example: LU solve:

$$u_s \|E_i\|_\infty \leq 3n u_f \| |A^{-1}| |\hat{L}| |\hat{U}| \|_\infty$$

$$\max(c_1, c_2) u_s \leq \frac{3n u_f \| |\hat{L}| |\hat{U}| \|_\infty}{\|A\|_\infty}$$

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

→ normwise relative backward error is at most $\max(c_1, c_2) u_s$

3. $|\hat{r}_i - A\hat{d}_i| \leq u_s G_i |\hat{d}_i|$

→ componentwise relative backward error is bounded by a multiple of u_s

$E_i, c_1, c_2,$ and G_i depend on $A, \hat{r}_i, n,$ and u_s

example: LU solve:

$$u_s \|E_i\|_\infty \leq 3n u_f \| |A^{-1}| |\hat{L}| |\hat{U}| \|_\infty$$

$$\max(c_1, c_2) u_s \leq \frac{3n u_f \| |\hat{L}| |\hat{U}| \|_\infty}{\|A\|_\infty}$$

$$u_s \|G_i\|_\infty \leq 3n u_f \| |\hat{L}| |\hat{U}| \|_\infty$$

Key Analysis Innovations II

Allow for general solver:

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

Assume computed solution \hat{d}_i to $Ad_i = \hat{r}_i$ satisfies:

1. $\hat{d}_i = (I + u_s E_i) d_i, \quad u_s \|E_i\|_\infty < 1$

→ normwise relative forward error is bounded by multiple of u_s and is less than 1

2. $\|\hat{r}_i - A\hat{d}_i\|_\infty \leq u_s (c_1 \|A\|_\infty \|\hat{d}_i\|_\infty + c_2 \|\hat{r}_i\|_\infty)$

→ normwise relative backward error is at most $\max(c_1, c_2) u_s$

3. $|\hat{r}_i - A\hat{d}_i| \leq u_s G_i |\hat{d}_i|$

→ componentwise relative backward error is bounded by a multiple of u_s

$E_i, c_1, c_2,$ and G_i depend on $A, \hat{r}_i, n,$ and u_s

example: LU solve:

$$u_s = u_f$$

$$u_s \|E_i\|_\infty \leq 3n u_f \| |A^{-1}| |\hat{L}| |\hat{U}| \|_\infty$$

$$\max(c_1, c_2) u_s \leq \frac{3n u_f \| |\hat{L}| |\hat{U}| \|_\infty}{\|A\|_\infty}$$

$$u_s \|G_i\|_\infty \leq 3n u_f \| |\hat{L}| |\hat{U}| \|_\infty$$

Forward Error for IR3

- Three precisions:
 - u_f : factorization precision
 - u : working precision
 - u_r : residual computation precision

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$$

$$\text{cond}(A) = \| |A^{-1}| |A| \|_\infty$$

$$\text{cond}(A, x) = \| |A^{-1}| |A| |x| \|_\infty / \|x\|_\infty$$

Forward Error for IR3

- Three precisions:

- u_f : factorization precision
- u : working precision
- u_r : residual computation precision

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$$

$$\text{cond}(A) = \| |A^{-1}| |A| \|_\infty$$

$$\text{cond}(A, x) = \| |A^{-1}| |A| |x| \|_\infty / \|x\|_\infty$$

Theorem [C. and Higham, SISC 40(2), 2018]

For IR in precisions $u_f \geq u \geq u_r$ and effective solve precision u_s , if

$$\phi_i \equiv 2u_s \min(\text{cond}(A), \kappa_\infty(A)\mu_i) + u_s \|E_i\|_\infty$$

is less than 1, then the forward error is reduced on the i th iteration by a factor $\approx \phi_i$ until an iterate \hat{x}_i is produced for which

$$\frac{\|x - \hat{x}_i\|_\infty}{\|x\|_\infty} \lesssim 4N u_r \text{cond}(A, x) + u,$$

where N is the maximum number of nonzeros per row in A .

Forward Error for IR3

- Three precisions:
 - u_f : factorization precision
 - u : working precision
 - u_r : residual computation precision

$$\kappa_\infty(A) = \|A^{-1}\|_\infty \|A\|_\infty$$

$$\text{cond}(A) = \| |A^{-1}| |A| \|_\infty$$

$$\text{cond}(A, x) = \| |A^{-1}| |A| |x| \|_\infty / \|x\|_\infty$$

Theorem [C. and Higham, SISC 40(2), 2018]

For IR in precisions $u_f \geq u \geq u_r$ and effective solve precision u_s , if

$$\phi_i \equiv 2u_s \min(\text{cond}(A), \kappa_\infty(A)\mu_i) + u_s \|E_i\|_\infty$$

is less than 1, then the forward error is reduced on the i th iteration by a factor $\approx \phi_i$ until an iterate \hat{x}_i is produced for which

$$\frac{\|x - \hat{x}_i\|_\infty}{\|x\|_\infty} \lesssim 4N u_r \text{cond}(A, x) + u,$$

where N is the maximum number of nonzeros per row in A .

→ Analogous traditional bounds: $\phi_i \equiv 3n u_f \kappa_\infty(A)$

Normwise Backward Error for IR3

Theorem [C. and Higham, SISC 40(2), 2018]

For IR in precisions $u_f \geq u \geq u_r$ and effective solve precision u_s , if

$$\phi_i \equiv (c_1 \kappa_\infty(A) + c_2) u_s$$

is less than 1, then the residual is reduced on the i th iteration by a factor $\approx \phi_i$ until an iterate \hat{x}_i is produced for which

$$\|b - A\hat{x}_i\|_\infty \lesssim N u (\|b\|_\infty + \|A\|_\infty \|\hat{x}_i\|_\infty),$$

where N is the maximum number of nonzeros per row in A .

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		Forward error
				norm	comp	
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}
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}
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}
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		Forward error
					norm	comp	
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}
LP fact.	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}
	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		Forward error
					norm	comp	
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		Forward error
					norm	comp	
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		Forward error
					norm	comp	
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A, x) \cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A, x) \cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A, x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A, x) \cdot 10^{-16}$
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		Forward error
					norm	comp	
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A, x) \cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A, x) \cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}
Fixed	S	S	S	10^8	10^{-8}	10^{-8}	$\text{cond}(A, x) \cdot 10^{-8}$
Trad.	S	S	D	10^8	10^{-8}	10^{-8}	10^{-8}
LP fact.	S	D	D	10^8	10^{-16}	10^{-16}	$\text{cond}(A, x) \cdot 10^{-16}$
New	S	D	Q	10^8	10^{-16}	10^{-16}	10^{-16}

\Rightarrow Benefit of IR3 vs. "LP fact.": no $\text{cond}(A, x)$ 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 $\approx 10^{-34}$

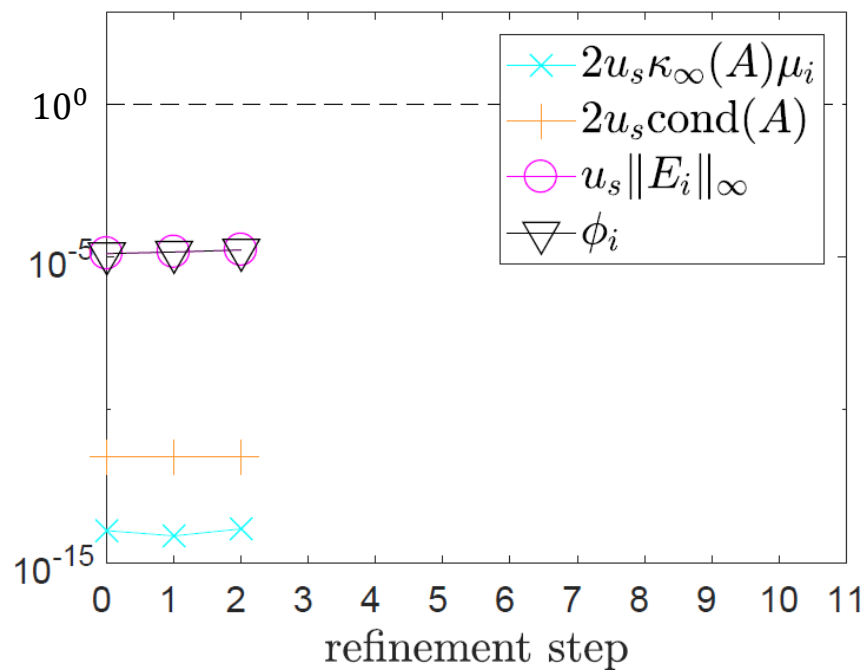
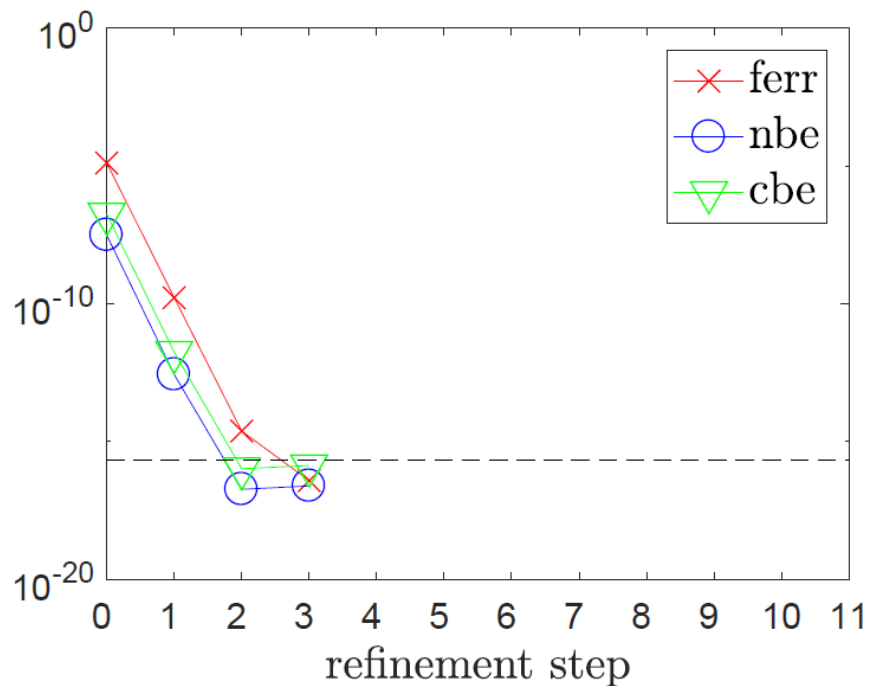
	u_f	u	u_r	$\max \kappa_\infty(A)$	Backward error		Forward error
					norm	comp	
LP fact.	H	S	S	10^4	10^{-8}	10^{-8}	$\text{cond}(A, x) \cdot 10^{-8}$
New	H	S	D	10^4	10^{-8}	10^{-8}	10^{-8}
LP fact.	H	D	D	10^4	10^{-16}	10^{-16}	$\text{cond}(A, x) \cdot 10^{-16}$
New	H	D	Q	10^4	10^{-16}	10^{-16}	10^{-16}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

```
A = gallery('randsvd', 100, 1e3)
b = randn(100,1)
```

$$\kappa_\infty(A) \approx 1e4$$

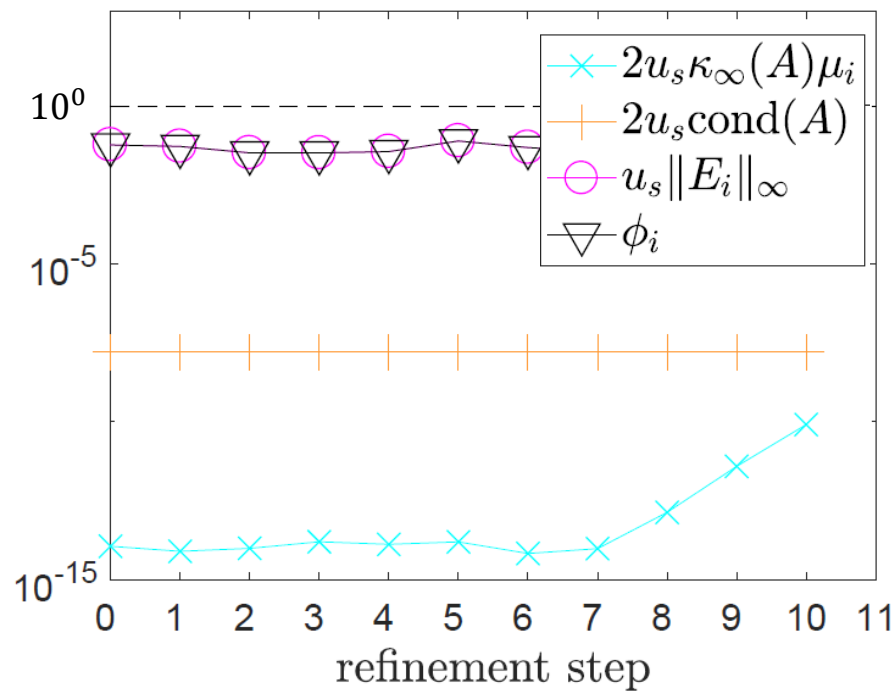
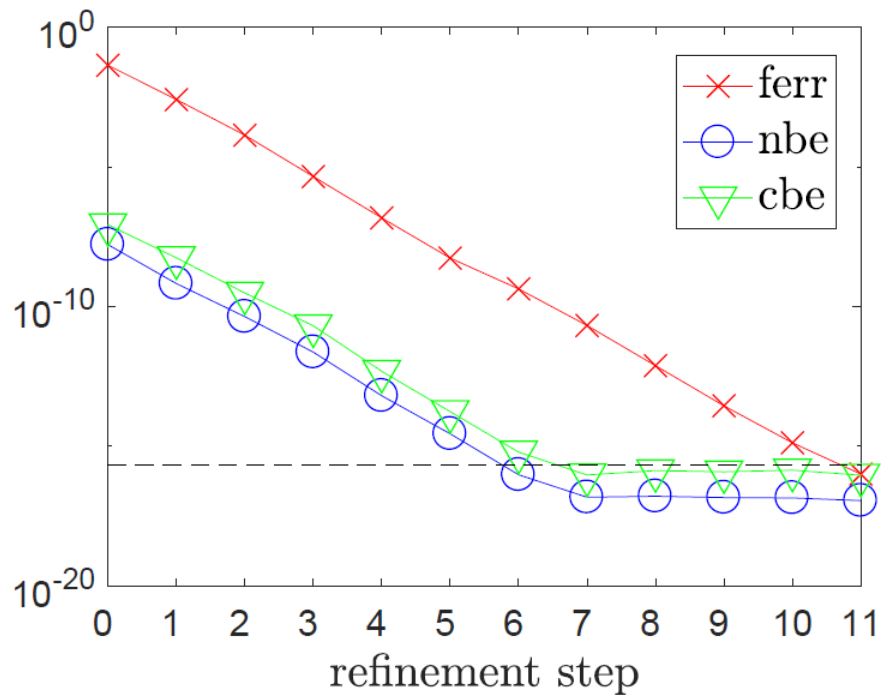
Standard (LU-based) IR with u_f : single, u : double, u_r : quad




```
A = gallery('randsvd', 100, 1e7)
b = randn(100,1)
```

$$\kappa_\infty(A) \approx 7e7$$

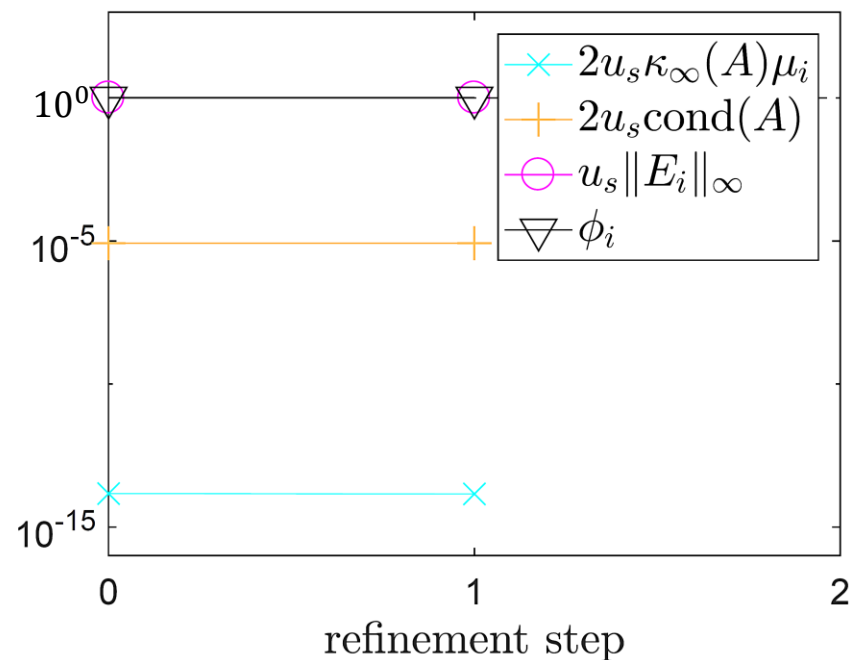
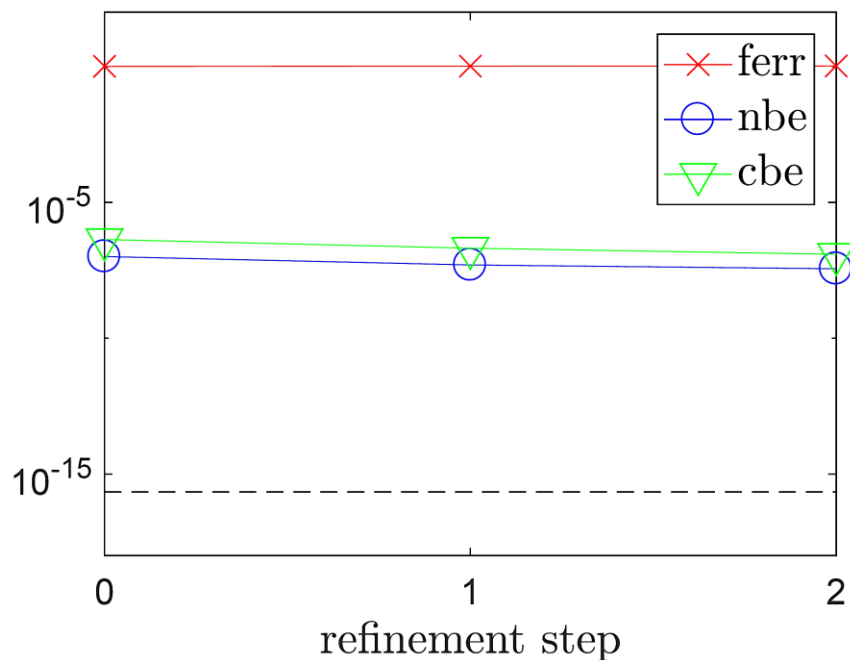
Standard (LU-based) IR with u_f : single, u : double, u_r : quad



```
A = gallery('randsvd', 100, 1e9)
b = randn(100,1)
```

$\kappa_\infty(A) \approx 2e10$

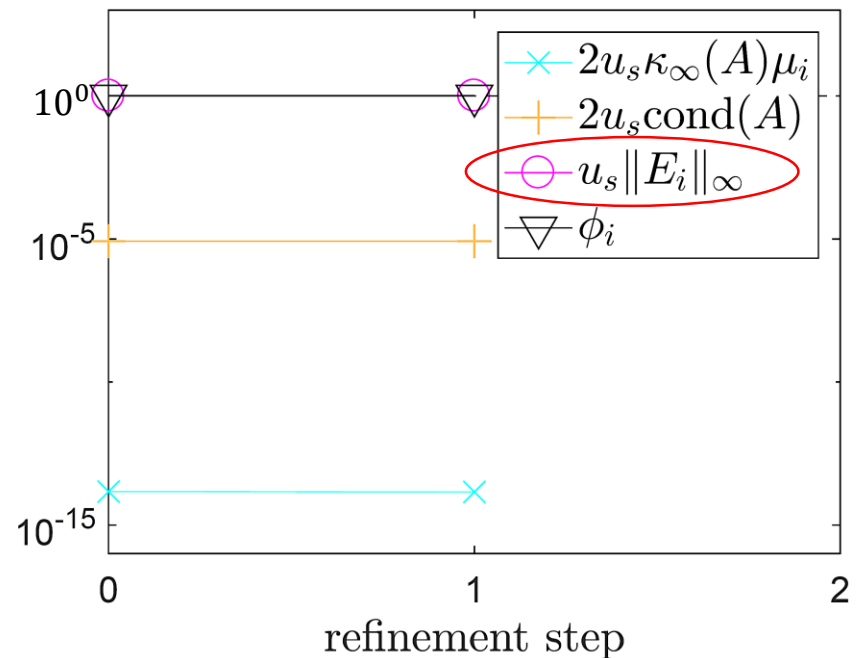
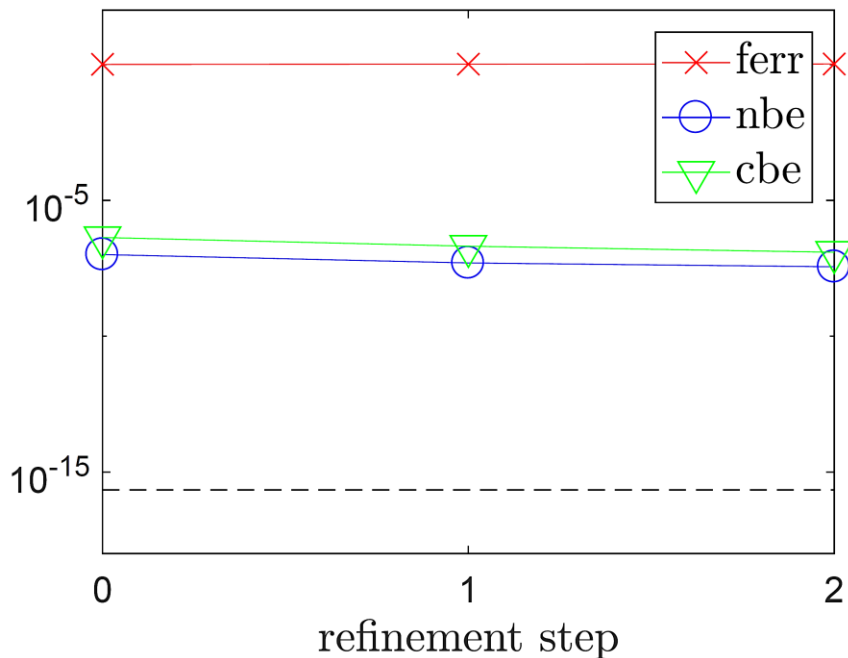
Standard (LU-based) IR with u_f : single, u : double, u_r : quad



```
A = gallery('randsvd', 100, 1e9)
b = randn(100,1)
```

$$\kappa_{\infty}(A) \approx 2e10$$

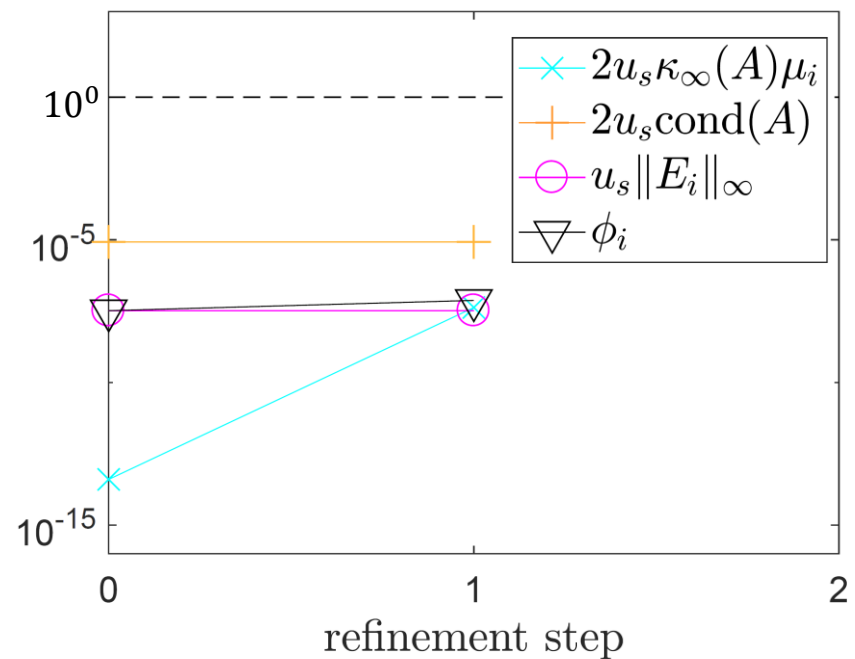
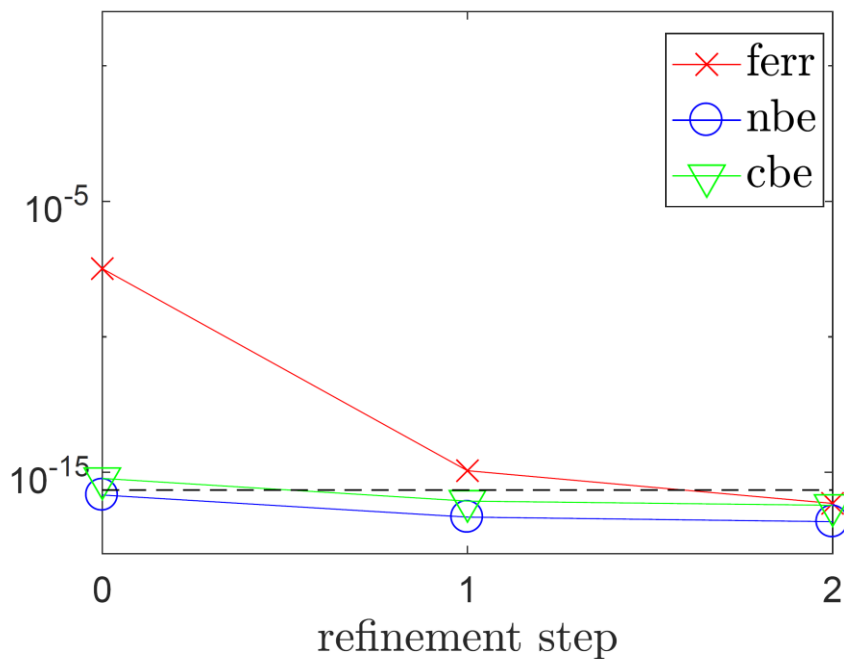
Standard (LU-based) IR with u_f : single, u : double, u_r : quad



```
A = gallery('randsvd', 100, 1e9)
b = randn(100,1)
```

$\kappa_\infty(A) \approx 2e10$

Standard (LU-based) IR with u_f : double, u : double, u_r : quad



GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if \hat{L} and \hat{U} are computed LU factors of A in precision u_f , then

$$\kappa_\infty(\hat{U}^{-1}\hat{L}^{-1}A) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if \hat{L} and \hat{U} are computed LU factors of A in precision u_f , then

$$\kappa_\infty(\hat{U}^{-1}\hat{L}^{-1}A) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

GMRES-IR [C. and Higham, SISC 39(6), 2017]

- To compute the updates d_i , apply GMRES to

$$\overbrace{\hat{U}^{-1}\hat{L}^{-1}A}^{\tilde{A}} d_i = \overbrace{\hat{U}^{-1}\hat{L}^{-1}r_i}^{\tilde{r}_i}$$

GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if \hat{L} and \hat{U} are computed LU factors of A in precision u_f , then

$$\kappa_\infty(\hat{U}^{-1}\hat{L}^{-1}A) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

GMRES-IR [C. and Higham, SISC 39(6), 2017]

- To compute the updates d_i , apply GMRES to $\underbrace{\hat{U}^{-1}\hat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\hat{U}^{-1}\hat{L}^{-1}r_i}_{\tilde{r}_i}$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

$$r_i = b - Ax_i$$

Solve $Ad_i = r_i$ via GMRES on $\tilde{A}d_i = \tilde{r}_i$

$$x_{i+1} = x_i + d_i$$

GMRES-Based Iterative Refinement

- Observation [Rump, 1990]: if \hat{L} and \hat{U} are computed LU factors of A in precision u_f , then

$$\kappa_\infty(\hat{U}^{-1}\hat{L}^{-1}A) \approx 1 + \kappa_\infty(A)u_f,$$

even if $\kappa_\infty(A) \gg u_f^{-1}$.

GMRES-IR [C. and Higham, SISC 39(6), 2017]

- To compute the updates d_i , apply GMRES to $\underbrace{\hat{U}^{-1}\hat{L}^{-1}A}_{\tilde{A}}d_i = \underbrace{\hat{U}^{-1}\hat{L}^{-1}r_i}_{\tilde{r}_i}$

Solve $Ax_0 = b$ by LU factorization

for $i = 0$: maxit

$$r_i = b - Ax_i$$

Solve $Ad_i = r_i$ via GMRES on $\tilde{A}d_i = \tilde{r}_i$

$$x_{i+1} = x_i + d_i$$

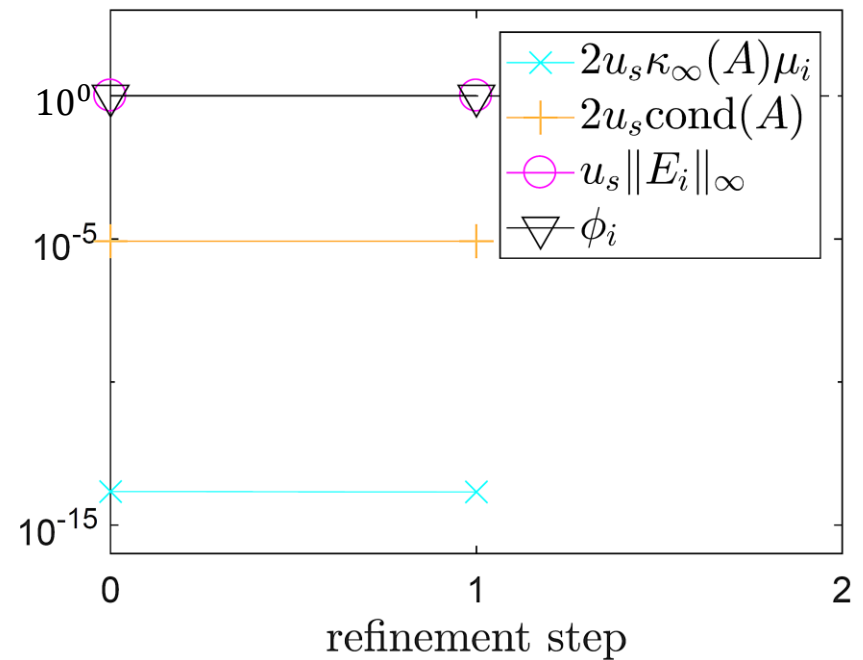
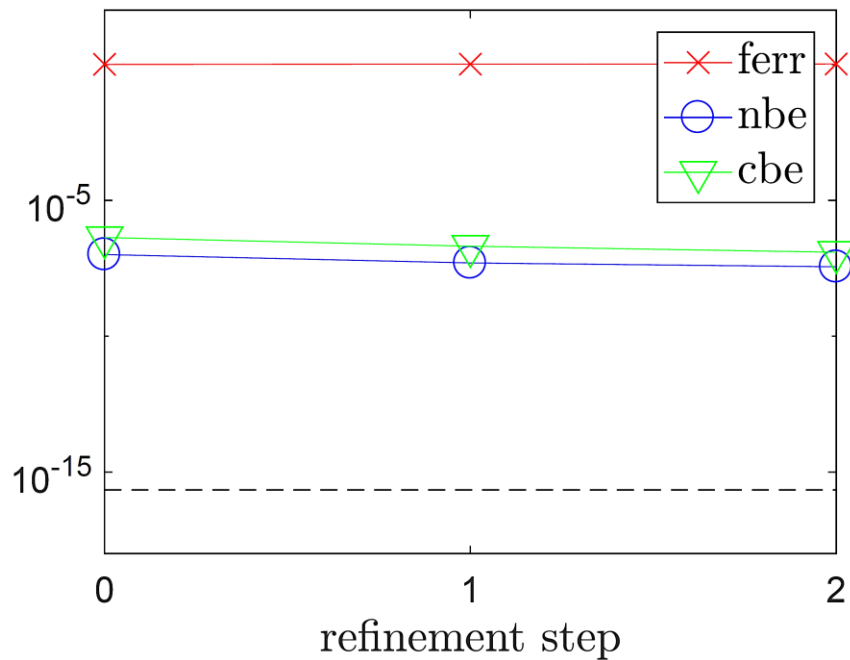

$$u_s = u$$


```
A = gallery('randsvd', 100, 1e9, 2)
```

```
b = randn(100,1)
```

$\kappa_\infty(A) \approx 2e10, \text{ cond}(A,x) \approx 5e9$

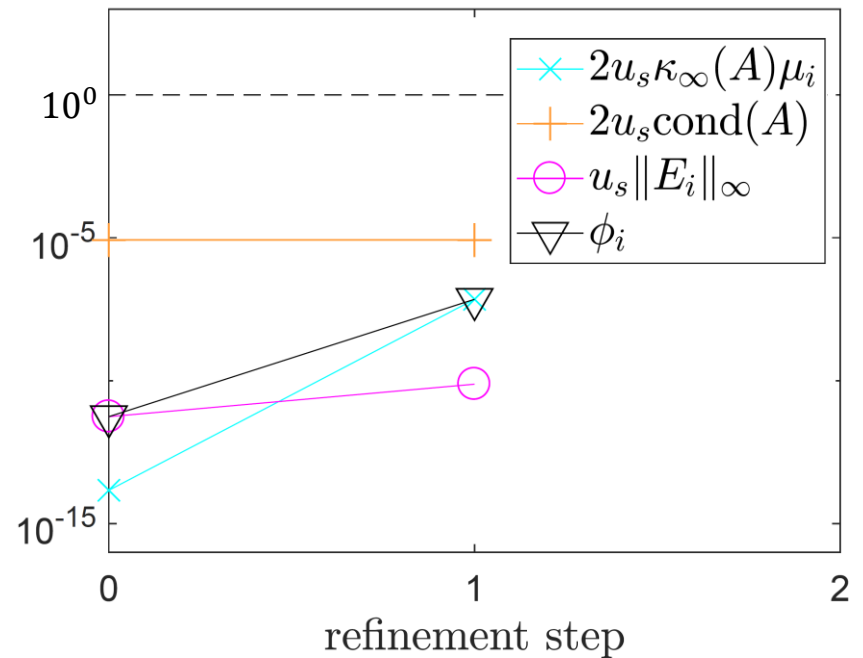
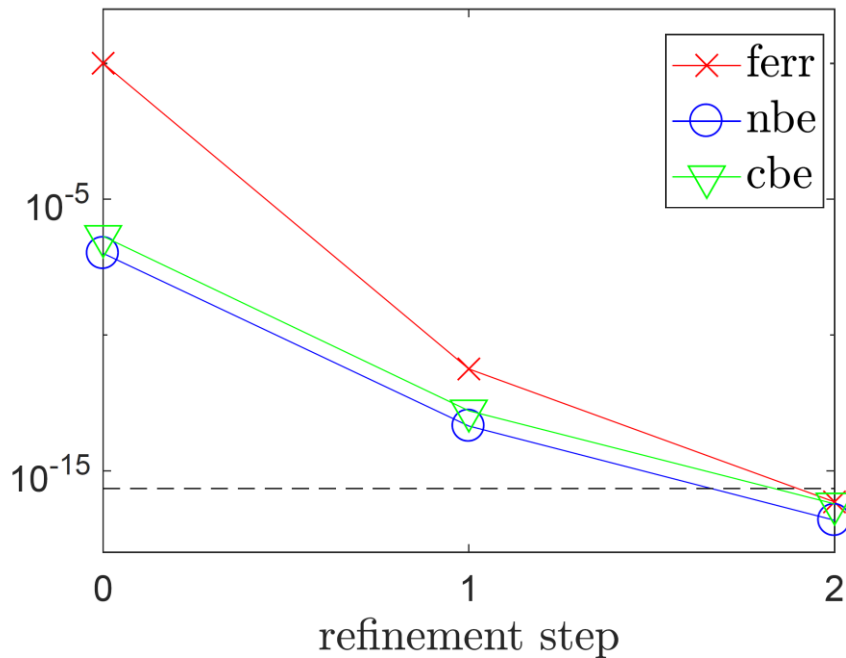
Standard (LU-based) IR with u_f : single, u : double, u_r : quad



```
A = gallery('randsvd', 100, 1e9, 2)
b = randn(100,1)
```

$\kappa_\infty(A) \approx 2e10$, $\text{cond}(A, x) \approx 5e9$, $\kappa_\infty(\tilde{A}) \approx 2e4$

GMRES-IR with u_f : single, u : double, u_r : quad



Number of GMRES iterations: (2,3)

GMRES-IR: Summary

Benefits of GMRES-IR:

	u_f	u	u_r	$\max \kappa_\infty(A)$	Backward error		Forward error
					norm	comp	
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}

GMRES-IR: Summary

Benefits of GMRES-IR:

	u_f	u	u_r	$\max \kappa_\infty(A)$	Backward error		Forward error
					norm	comp	
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, low precision factorization will work for higher $\kappa_\infty(A)$

GMRES-IR: Summary

Benefits of GMRES-IR:

	u_f	u	u_r	$\max \kappa_\infty(A)$	Backward error		Forward error
					norm	comp	
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 $\kappa_\infty(A)$



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

GMRES-IR: Summary

Benefits of GMRES-IR:

	u_f	u	u_r	$\max \kappa_\infty(A)$	Backward error		Forward error
					norm	comp	
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}

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

GMRES-IR: Summary

Benefits of GMRES-IR:

	u_f	u	u_r	$\max \kappa_\infty(A)$	Backward error		Forward error
					norm	comp	
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}

Try IR3! MATLAB codes available at:

<https://github.com/eccarson/ir3>

Comments and Caveats I

- Convergence tolerance τ for GMRES?
 - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
 - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps

Comments and Caveats I

- Convergence tolerance τ for GMRES?
 - Smaller $\tau \rightarrow$ more GMRES iterations, potentially fewer refinement steps
 - Larger $\tau \rightarrow$ fewer GMRES iterations, potentially more refinement steps
- What about overflow, underflow, subnormal numbers?
 - Sophisticated scaling methods can help avoid this
 - “Squeezing a Matrix into Half Precision, with an Application to Solving Linear Systems” [Higham, Pranesh, Zounon, 2019]

Comments and Caveats II

- Convergence rate of GMRES?

Comments and Caveats II

- Convergence rate of GMRES?
 - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
 - e.g., if \tilde{A} still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_{\infty}(A)$ [Liesen and Tichý, 2004]
 - Potential remedies: deflation, Krylov subspace recycling, using additional preconditioner

Comments and Caveats II

- Convergence rate of GMRES?
 - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
 - e.g., if \tilde{A} still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_{\infty}(A)$ [Liesen and Tichý, 2004]
 - Potential remedies: deflation, Krylov subspace recycling, using additional preconditioner
- Depending on conditioning of A , applying \tilde{A} to a vector must be done accurately (precision u^2) in each GMRES iteration

Comments and Caveats II

- Convergence rate of GMRES?
 - If A is ill conditioned and LU factorization is performed in very low precision, it can be a poor preconditioner
 - e.g., if \tilde{A} still has cluster of eigenvalues near origin, GMRES can stagnate until n^{th} iteration, regardless of $\kappa_{\infty}(A)$ [Liesen and Tichý, 2004]
 - Potential remedies: deflation, Krylov subspace recycling, using additional preconditioner
- Depending on conditioning of A , applying \tilde{A} to a vector must be done accurately (precision u^2) in each GMRES iteration
- Why GMRES?
 - Theoretical purposes: existing analysis and proof of backward stability [Paige, Rozložník, Strakoš, 2006]
 - In practice, use any solver you want!

GMRES-IR in Libraries and Applications

- MAGMA: Dense linear algebra routines for heterogeneous/hybrid architectures

```
magma / src / dxgesv_gmres_gpu.cpp
```

```
128  -----
129  DSGESV or DHGESV expert interface.
130  It computes the solution to a real system of linear equations
131   $A * X = B$ ,  $A^{**T} * X = B$ , or  $A^{**H} * X = B$ ,
132  where A is an N-by-N matrix and X and B are N-by-NRHS matrices.
133  the accomodate the Single Precision DSGESV and the Half precision dhgesv API.
134  precision and iterative refinement solver are specified by facto_type, solver_type.
135  For other API parameter please refer to the corresponding dsgesv or dhgesv.
```

- NVIDIA's CuSolver Library

[2.2.1.6. cusolverIRSRefinement_t](#)

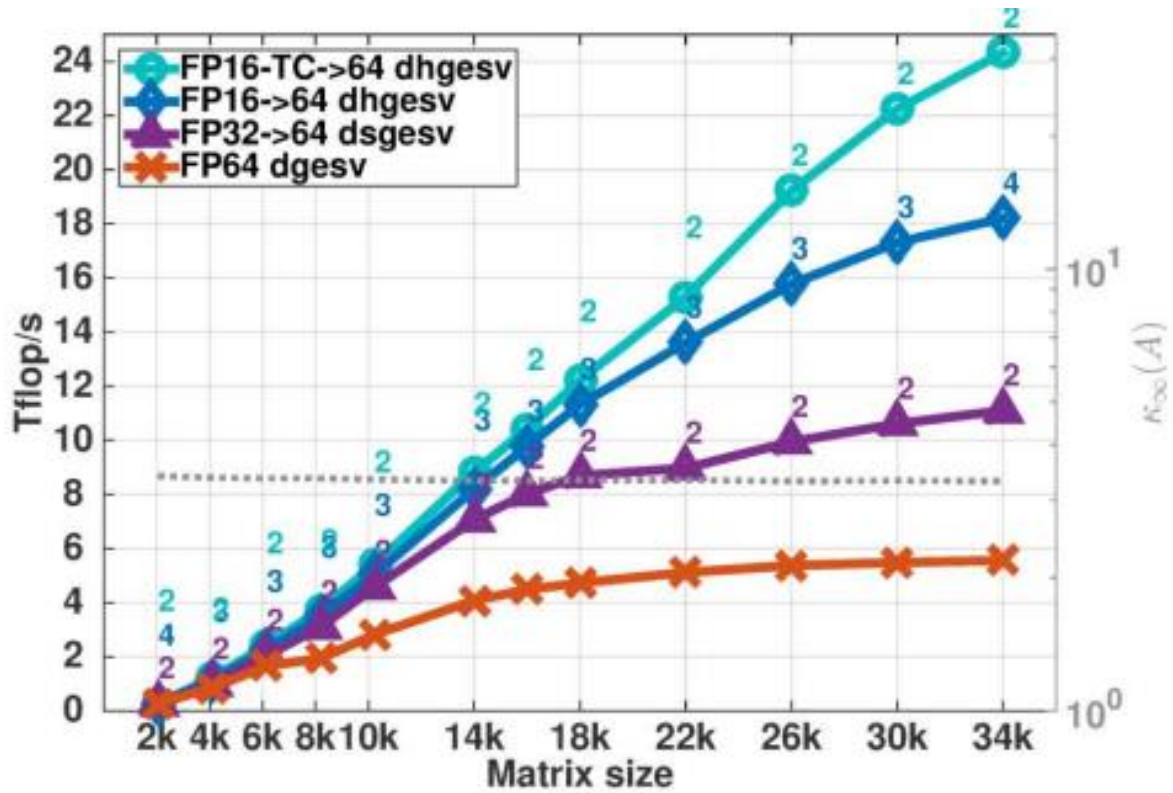
The `cusolverIRSRefinement_t` type indicates which solver type would be used for the specific cusolver function. Most of our experimentation shows that `CUSOLVER_IRS_REFINE_GMRES` is the best option.

<code>CUSOLVER_IRS_REFINE_GMRES</code>	GMRES (Generalized Minimal Residual) based iterative refinement solver. In recent study, the GMRES method has drawn the scientific community attention for its ability to be used as refinement solver that outperforms the classical iterative refinement method. based on our experimentation, we recommend this setting.
--	---

- In production codes: FK6D/ASGarD code (Oak Ridge National Lab, USA) for tokomak containment problem

Performance Results (MAGMA)

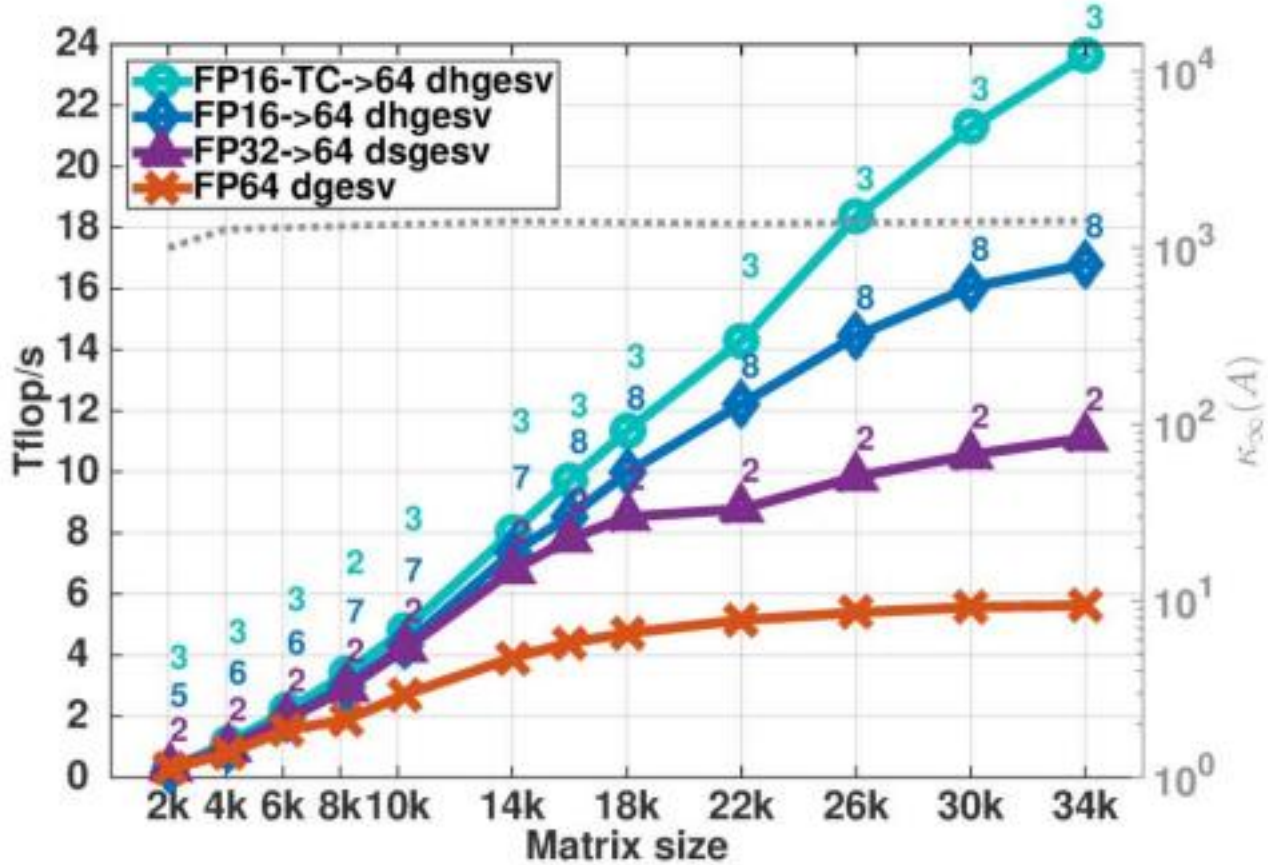
- [Haidar, Tomov, Dongarra, Higham, 2018]
- 2-precision GMRES-IR approach ($u = u_r$) on NVIDIA V100
- IR run to FP64 accuracy, max 400 iterations in GMRES
- Tflops/s measured as $(2n^3/3)/\text{time}$



(a) Matrix of type 1: diagonally dominant.

Performance Results (MAGMA)

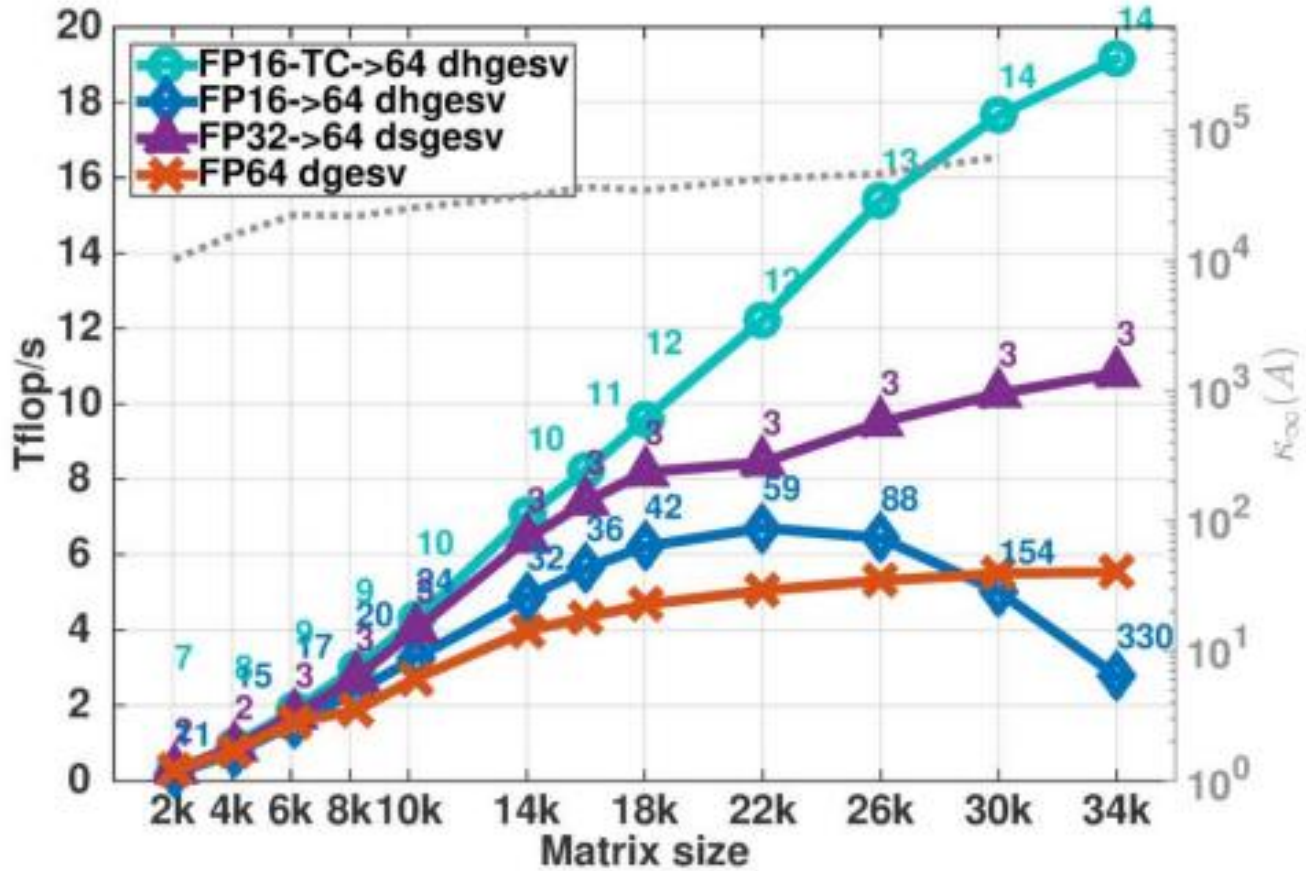
- [Haidar, Tomov, Dongarra, Higham, 2018]



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

Performance Results (MAGMA)

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



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

Performance Results

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

Performance for Matrices from SuiteSparse

name	Description	size	$\kappa_{\infty}(A)$	dgesv time(s)	dsgesv		dhgesv		dhgesv-TC	
					# iter	time (s)	# iter	time (s)	# iter	time (s)
em192	radar design	26896	10^6	5.70	3	3.11	40	5.21	10	2.05
appu	NASA app benchmark	14000	10^4	0.43	2	0.27	7	0.24	4	0.19
ns3Da	3D Navier Stokes	20414	$7.6 \cdot 10^3$	1.12	2	0.69	6	0.54	4	0.43
nd6k	ND problem set	18000	$3.5 \cdot 10^2$	0.81	2	0.45	4	0.36	3	0.30
nd12k	ND problem set	36000	$4.3 \cdot 10^2$	5.36	2	2.75	3	1.76	3	1.31

HPL-AI Benchmark

- HPL/LINPACK benchmark has been used in TOP500 since the 90s
 - Double precision, dense $Ax=b$ using GEPP
 - Not necessarily indicative of application performance, especially for ML/AI applications
 - Doesn't take advantage of low-precision hardware!

HPL-AI Benchmark

- HPL/LINPACK benchmark has been used in TOP500 since the 90s
 - Double precision, dense $Ax=b$ using GEPP
 - Not necessarily indicative of application performance, especially for ML/AI applications
 - Doesn't take advantage of low-precision hardware!
- **HPL-AI benchmark (2019)**
 - Highlights confluence of HPC+AI workloads
 - Like HPL, solves dense $Ax=b$, results still to double precision accuracy
 - Achieves this via mixed-precision **GMRES-IR**
 - may be implemented in a way that takes advantage of the current and upcoming devices for accelerating AI workloads

HPL-AI Benchmark Performance

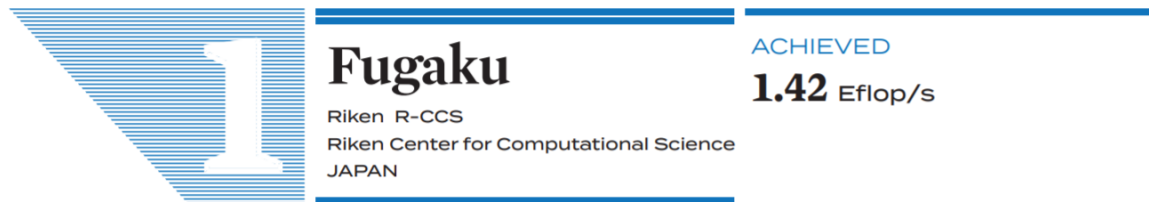
- So far, two machines have reported results for HPL-AI (#1 and #2 on TOP500)

HPL-AI Results (June 2020):

- Summit: 550 PETAFL0P/s (vs. 148.6 PETAFL0P/s on HPL)
- Fugaku: **1.42 EXAFL0P/s** (vs. 415.5 PETAFL0P/s on HPL)

HPL-AI
JUNE 2020

NUMBER 1 SYSTEM



1

Fugaku
Riken R-CCS
Riken Center for Computational Science
JAPAN

ACHIEVED
1.42 Eflop/s

HPL-AI Benchmark

- In the future, HPL-AI will gain same status as benchmarks that complement HPL, like HPCG, Graph500, Green500
- More information: <https://icl.bitbucket.io/hpl-ai/>
- Reference implementation: <https://bitbucket.org/icl/hpl-ai/src/>

Extension to Least Squares Problems

- Want to solve

$$\min_x \|b - Ax\|_2$$

where $A \in \mathbb{R}^{m \times n}$ ($m > n$) has rank n

- Commonly solved using QR factorization:

$$A = QR = [Q_1, Q_2] \begin{bmatrix} U \\ 0 \end{bmatrix}$$

where Q is an $m \times m$ orthogonal matrix and U is upper triangular.

$$x = U^{-1}Q_1^T b, \quad \|b - Ax\|_2 = \|Q_2^T b\|_2$$

Extension to Least Squares Problems

- Want to solve

$$\min_x \|b - Ax\|_2$$

where $A \in \mathbb{R}^{m \times n}$ ($m > n$) has rank n

- Commonly solved using QR factorization:

$$A = QR = [Q_1, Q_2] \begin{bmatrix} U \\ 0 \end{bmatrix}$$

where Q is an $m \times m$ orthogonal matrix and U is upper triangular.

$$x = U^{-1}Q_1^T b, \quad \|b - Ax\|_2 = \|Q_2^T b\|_2$$

- As in linear system case, for ill-conditioned problems, iterative refinement often needed to improve accuracy and stability

Least Squares Iterative Refinement

- For inconsistent systems, must simultaneously refine both solution and residual
- (Björck,1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

Least Squares Iterative Refinement

- For inconsistent systems, must simultaneously refine both solution and residual
- (Björck,1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

- Refinement proceeds as follows:

1. Compute "residuals"

$$\begin{bmatrix} f_i \\ g_i \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ x_i \end{bmatrix} = \begin{bmatrix} b - r_i - Ax_i \\ -A^T r_i \end{bmatrix}$$

2. Solve for corrections

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix} = \begin{bmatrix} f_i \\ g_i \end{bmatrix}$$

3. Update "solution":

$$\begin{bmatrix} r_{i+1} \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} r_i \\ x_i \end{bmatrix} + \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix}$$

Least Squares Iterative Refinement

- For inconsistent systems, must simultaneously refine both solution and residual
- (Björck,1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \quad \tilde{A}\tilde{x} = \tilde{b}$$

- Refinement proceeds as follows:

1. Compute "residuals"

$$\begin{bmatrix} f_i \\ g_i \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ x_i \end{bmatrix} = \begin{bmatrix} b - r_i - Ax_i \\ -A^T r_i \end{bmatrix}$$

2. Solve for corrections

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix} = \begin{bmatrix} f_i \\ g_i \end{bmatrix}$$

3. Update "solution":

$$\begin{bmatrix} r_{i+1} \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} r_i \\ x_i \end{bmatrix} + \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix}$$

Least Squares Iterative Refinement

- For inconsistent systems, must simultaneously refine both solution and residual
- (Björck,1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \quad \tilde{A}\tilde{x} = \tilde{b}$$

- Refinement proceeds as follows:

1. Compute "residuals"

$$\begin{bmatrix} f_i \\ g_i \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ x_i \end{bmatrix} = \begin{bmatrix} b - r_i - Ax_i \\ -A^T r_i \end{bmatrix} \quad \tilde{r}_i = \tilde{b} - \tilde{A}\tilde{x}_i$$

2. Solve for corrections

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix} = \begin{bmatrix} f_i \\ g_i \end{bmatrix} \quad \tilde{A}d_i = \tilde{r}_i$$

3. Update "solution":

$$\begin{bmatrix} r_{i+1} \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} r_i \\ x_i \end{bmatrix} + \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix} \quad \tilde{x}_{i+1} = \tilde{x}_i + d_i$$

Least Squares Iterative Refinement

- For inconsistent systems, must simultaneously refine both solution and residual
- (Björck,1967): Least squares problem can be written as a linear system with square matrix of size $(m + n)$:

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \quad \tilde{A}\tilde{x} = \tilde{b}$$

- Refinement proceeds as follows:

1. Compute "residuals"

$$\begin{bmatrix} f_i \\ g_i \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} - \begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ x_i \end{bmatrix} = \begin{bmatrix} b - r_i - Ax_i \\ -A^T r_i \end{bmatrix}$$

2. Solve for corrections

$$\begin{bmatrix} I & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix} = \begin{bmatrix} f_i \\ g_i \end{bmatrix}$$

3. Update "solution":

$$\begin{bmatrix} r_{i+1} \\ x_{i+1} \end{bmatrix} = \begin{bmatrix} r_i \\ x_i \end{bmatrix} + \begin{bmatrix} \Delta r_i \\ \Delta x_i \end{bmatrix}$$

Results for 3-precision
IR for linear systems
**also applies to least
squares problems**

$$\tilde{r}_i = \tilde{b} - \tilde{A}\tilde{x}_i$$

$$\tilde{A}d_i = \tilde{r}_i$$

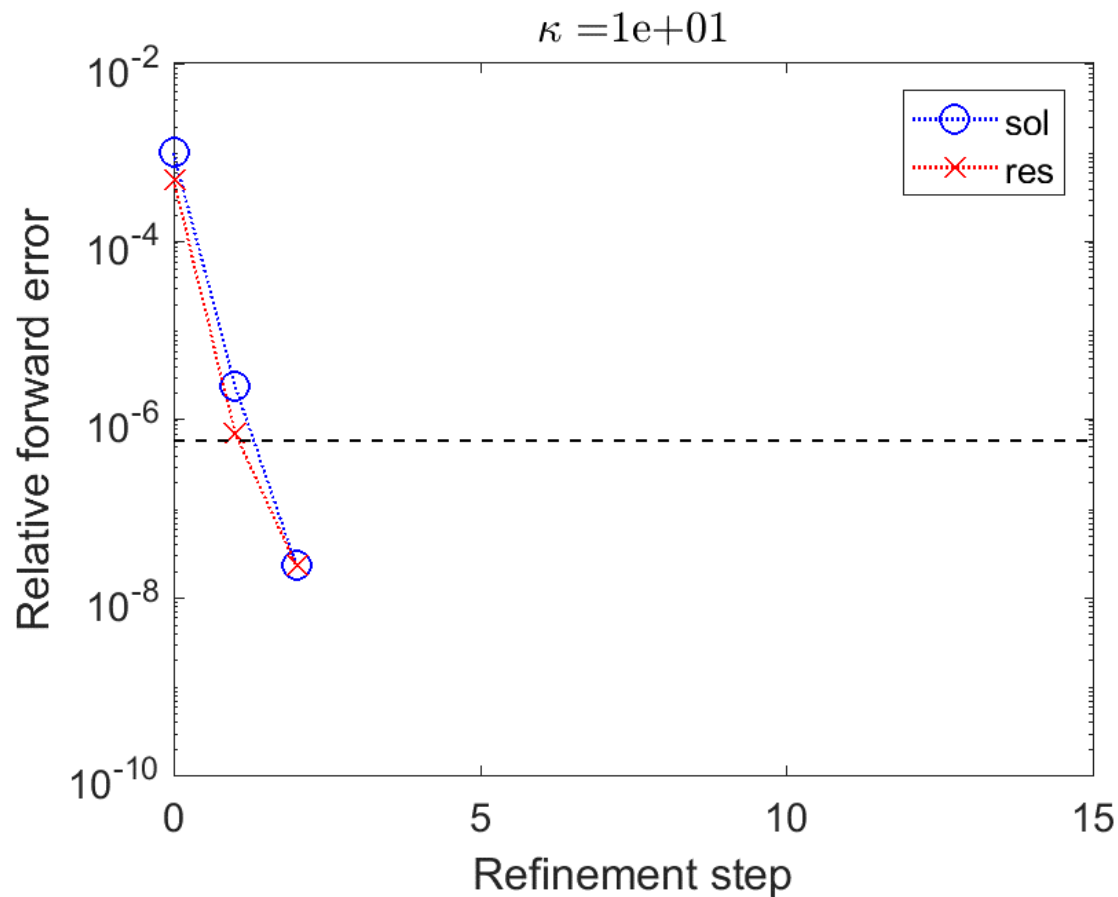
$$\tilde{x}_{i+1} = \tilde{x}_i + d_i$$

```
A = gallery('randsvd', [100, 10], kappa, 3)
b = randn(100,1); b = b./norm(b)
```

m n

Standard (QR-based) least squares IR with

u_f : half, u : single, u_r : double



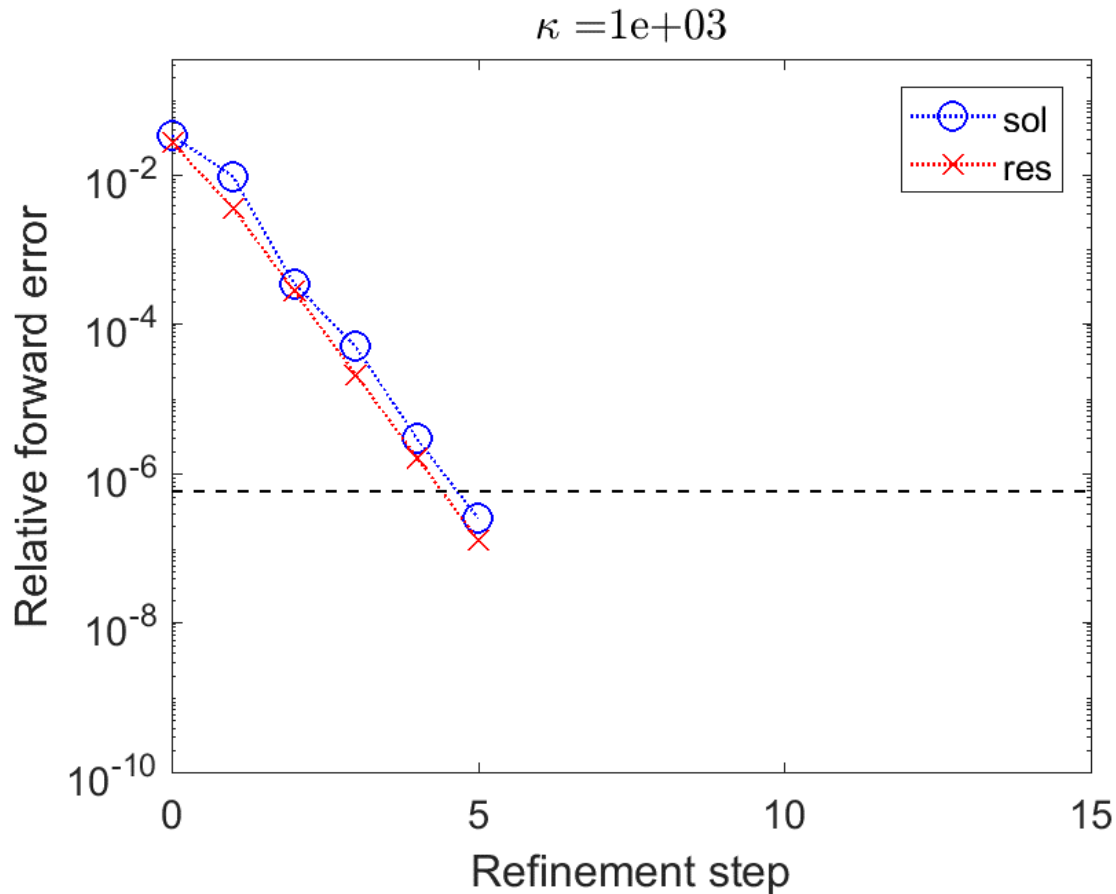
```

A = gallery('randsvd', [100, 10], kappa, 3)
b = randn(100,1); b = b./norm(b)

```

Standard (QR-based) least squares IR with

u_f : half, u : single, u_r : double

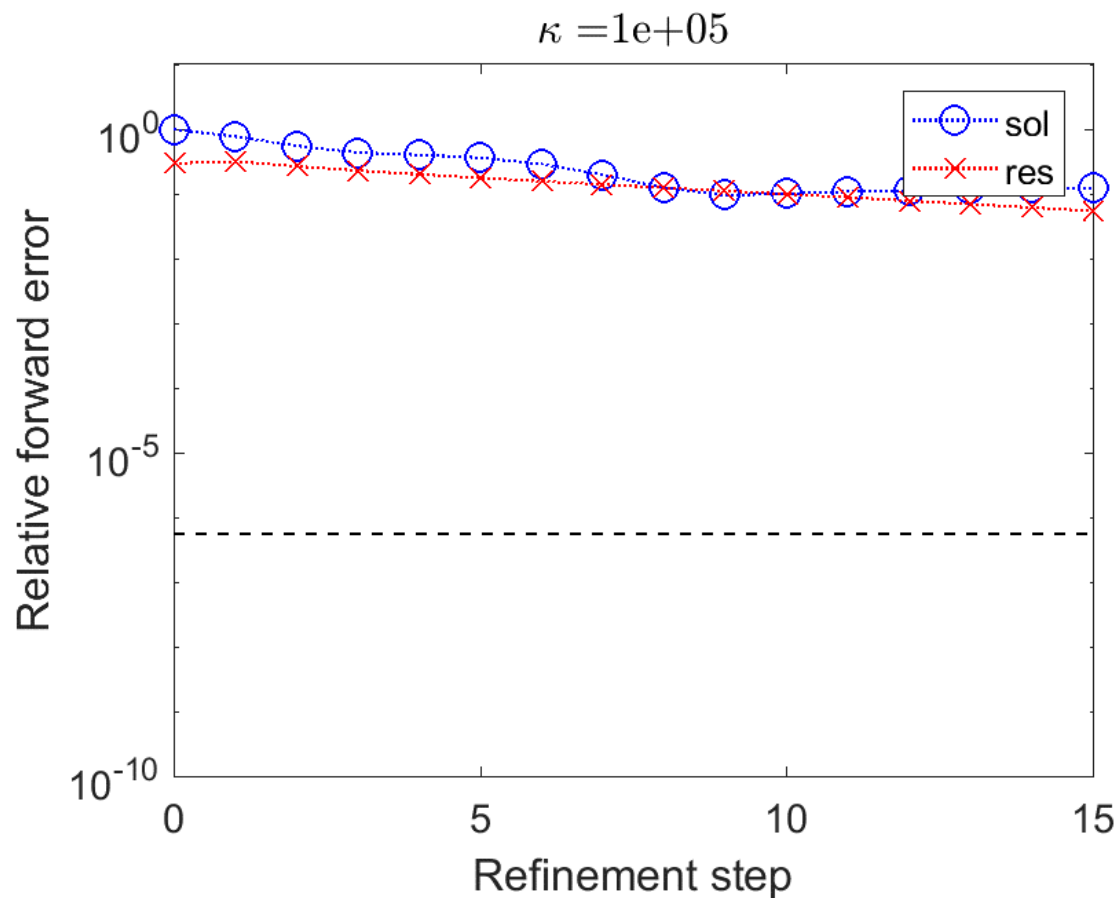


```

A = gallery('randsvd', [100, 10], kappa, 3)
b = randn(100,1); b = b./norm(b)

```

Standard (QR-based) least squares IR with
 u_f : half, u : single, u_r : double



GMRES-IR for Least Squares

- Similar to the linear system case, we can use a lower precision factorization for even more ill-conditioned problems if we improve the effective precision of the solver
- Again, don't want to compute an LU factorization of the augmented system
- How can we use existing QR factors to construct an effective and inexpensive preconditioner for the augmented system?
- Note that augmented system is a saddle-point system; lots of existing work (block diagonal, triangular, constraint-based, ...)

GMRES-IR for Least Squares

- Ex: block diagonal preconditioner ([Murphy, Golub, Wathen, 2000], [Ipsen, 2001])

$$\begin{bmatrix} \alpha I & 0 \\ 0 & \frac{1}{\alpha} \hat{R}^T \hat{R} \end{bmatrix} = \begin{bmatrix} \sqrt{\alpha} I & 0 \\ 0 & \frac{1}{\sqrt{\alpha}} \hat{R}^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha} I & 0 \\ 0 & \frac{1}{\sqrt{\alpha}} \hat{R} \end{bmatrix} \equiv M_1 M_2$$

GMRES-IR for Least Squares

- Ex: block diagonal preconditioner ([Murphy, Golub, Wathen, 2000], [Ipsen, 2001])

$$\begin{bmatrix} \alpha I & 0 \\ 0 & \frac{1}{\alpha} \hat{R}^T \hat{R} \end{bmatrix} = \begin{bmatrix} \sqrt{\alpha} I & 0 \\ 0 & \frac{1}{\sqrt{\alpha}} \hat{R}^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha} I & 0 \\ 0 & \frac{1}{\sqrt{\alpha}} \hat{R} \end{bmatrix} \equiv M_1 M_2$$

- Assuming QR factorization is exact,

$$M_2^{-1} M_1^{-1} \tilde{A} = \begin{bmatrix} I & \frac{1}{\alpha} A \\ \alpha \hat{R}^{-1} \hat{R}^{-T} A^T & 0 \end{bmatrix}$$

is nonsymmetric, diagonalizable, with eigenvalues $\left\{1, \frac{1}{2}(1 \pm \sqrt{5})\right\}$.

- However, condition number can still be quite large; unsuitable for proving backward stability of GMRES

GMRES-IR for Least Squares

- Ex: block diagonal preconditioner ([Murphy, Golub, Wathen, 2000], [Ipsen, 2001])

$$\begin{bmatrix} \alpha I & 0 \\ 0 & \frac{1}{\alpha} \hat{R}^T \hat{R} \end{bmatrix} = \begin{bmatrix} \sqrt{\alpha} I & 0 \\ 0 & \frac{1}{\sqrt{\alpha}} \hat{R}^T \end{bmatrix} \begin{bmatrix} \sqrt{\alpha} I & 0 \\ 0 & \frac{1}{\sqrt{\alpha}} \hat{R} \end{bmatrix} \equiv M_1 M_2$$

- Assuming QR factorization is exact,

$$M_2^{-1} M_1^{-1} \tilde{A} = \begin{bmatrix} I & \frac{1}{\alpha} A \\ \alpha \hat{R}^{-1} \hat{R}^{-T} A^T & 0 \end{bmatrix}$$

is nonsymmetric, diagonalizable, with eigenvalues $\left\{1, \frac{1}{2}(1 \pm \sqrt{5})\right\}$.

- However, condition number can still be quite large; unsuitable for proving backward stability of GMRES

- If we take split preconditioner

$$M_1^{-1} \tilde{A} M_2^{-1} = \begin{bmatrix} I & A \hat{R} \\ \hat{R}^{-T} A^T & 0 \end{bmatrix}$$

we will have a well-conditioned system

- However, split-preconditioned GMRES is not backward stable
- Potentially useful in practice, not but in theory

GMRES-IR for Least Squares

- One option:

$$M = \begin{bmatrix} \alpha I & \hat{Q}_1 \hat{R} \\ \hat{R}^T \hat{Q}_1^T & 0 \end{bmatrix}$$

- Then we can prove that for the left-preconditioned system,

$$\kappa(M^{-1}\tilde{A}) \leq \left(1 + \mathbf{u}_f c \kappa(A)\right)^2$$

where $c = O(m^2)$, where we note this bound is pessimistic.

- Thus even if $\kappa(A) \gg \mathbf{u}_f^{-1}$, the preconditioned system can still be reasonably well conditioned

GMRES-IR for Least Squares

- One option:

$$M = \begin{bmatrix} \alpha I & \hat{Q}_1 \hat{R} \\ \hat{R}^T \hat{Q}_1^T & 0 \end{bmatrix}$$

- Then we can prove that for the left-preconditioned system,

$$\kappa(M^{-1}\tilde{A}) \leq \left(1 + \mathbf{u}_f c \kappa(A)\right)^2$$

where $c = O(m^2)$, where we note this bound is pessimistic.

- Thus even if $\kappa(A) \gg \mathbf{u}_f^{-1}$, the preconditioned system can still be reasonably well conditioned
- GMRES run on \tilde{A} with left-preconditioner M gives

$$\mathbf{u}_s \|E_i\|_\infty \equiv \mathbf{u} f(m+n) \kappa_\infty(M^{-1}\tilde{A})$$

where f is a quadratic polynomial

GMRES-IR for Least Squares

- One option:

$$M = \begin{bmatrix} \alpha I & \hat{Q}_1 \hat{R} \\ \hat{R}^T \hat{Q}_1^T & 0 \end{bmatrix}$$

- Then we can prove that for the left-preconditioned system,

$$\kappa(M^{-1}\tilde{A}) \leq \left(1 + \mathbf{u}_f c \kappa(A)\right)^2$$

where $c = O(m^2)$, where we note this bound is pessimistic.

- Thus even if $\kappa(A) \gg \mathbf{u}_f^{-1}$, the preconditioned system can still be reasonably well conditioned
- GMRES run on \tilde{A} with left-preconditioner M gives

$$\mathbf{u}_s \|E_i\|_\infty \equiv \mathbf{u} f(m+n) \kappa_\infty(M^{-1}\tilde{A})$$

where f is a quadratic polynomial

- So for GMRES-based LSIR, $\mathbf{u}_s \equiv \mathbf{u}$; expect convergence of forward error when $\kappa_\infty(A) < \mathbf{u}^{-1/2} \mathbf{u}_f^{-1}$

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)

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)
- 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)
- 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

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)
- 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
- “A Survey of Numerical Methods Utilizing Mixed Precision Arithmetic”
<https://arxiv.org/abs/2007.06674>

Thank You!

carson@karlin.mff.cuni.cz

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

Select References

- Carson, E., & Higham, N. J. (2018). Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM J. SISC*, 40(2), A817-A847.
- Carson, E., & Higham, N. J. (2017). A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems. *SIAM J. SISC*, 39(6), A2834-A2856.
- Carson, E., Higham, N. J., & Pranesh, S. (2020). Three-Precision GMRES-Based Iterative Refinement for Least Squares Problems. *SIAM J. SISC* (to appear).
- Haidar, A., Tomov, S., Dongarra, J., & Higham, N. J. (2018, November). Harnessing GPU tensor cores for fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers. In Proc. *SC18* (pp. 603-613). IEEE.
- Higham, N. J., Pranesh, S., & Zounon, M. (2019). Squeezing a matrix into half precision, with an application to solving linear systems. *SIAM J. SISC*, 41(4), A2536-A2551.
- Abdelfattah, A., Anzt, H., Boman, E. G., Carson, E., Cojean, T., Dongarra, J., et al. (2020). A survey of numerical methods utilizing mixed precision arithmetic. *arXiv preprint arXiv:2007.06674*.