MANCHESTER 1824

The University of Manchester

M\crNA
Manchester Numerical Analysis

Exploiting Lower Precision Arithmetic in Solving Symmetric Positive Definite Linear Systems

Srikara Pranesh
School of Mathematics
The University of Manchester

srikara.pranesh@manchester.ac.uk

21-11-2019

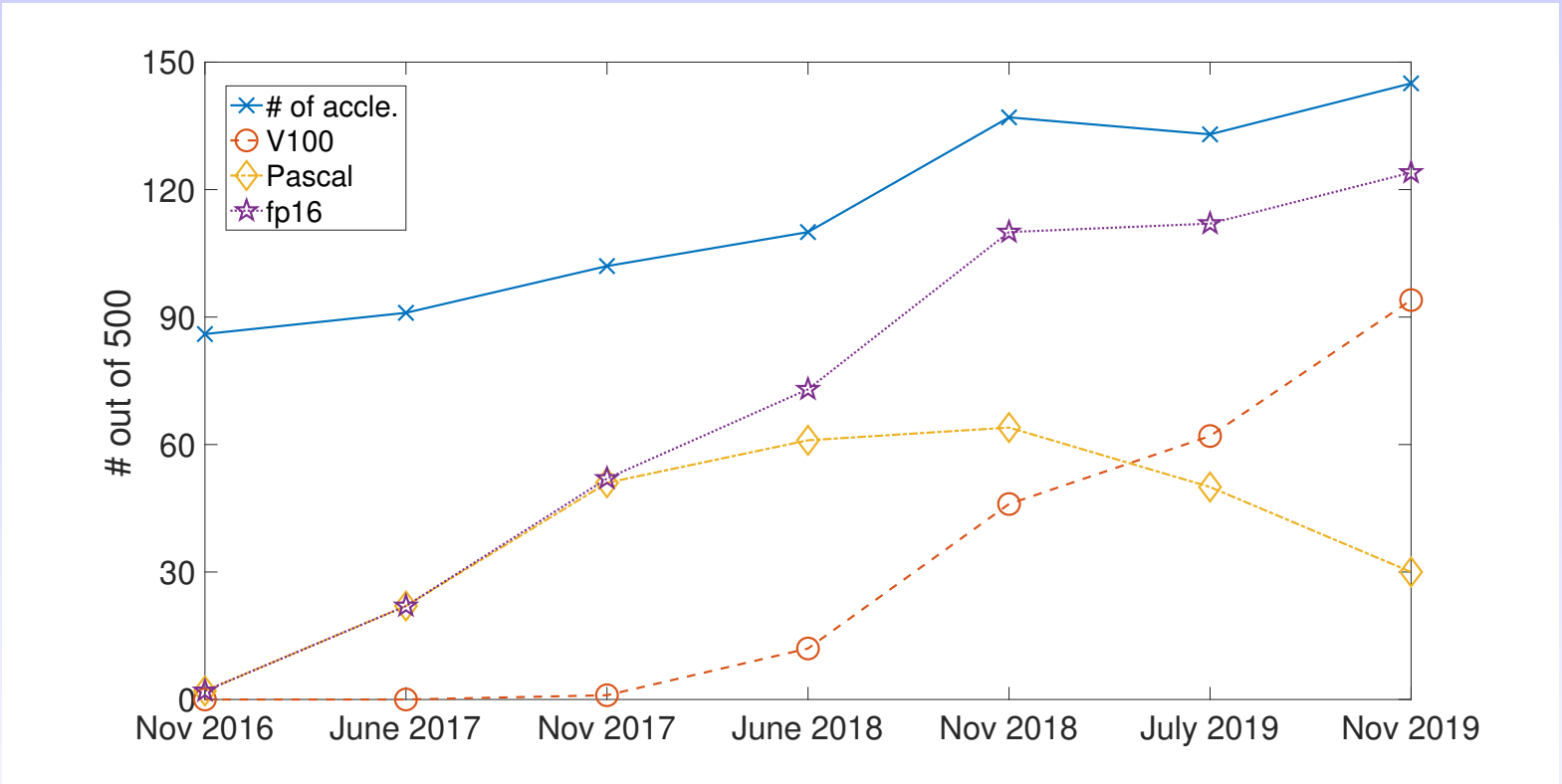
Joint work with Prof. Nick Higham

Motivation

Low precision floating-point formats are increasingly supported by computer hardware.

Motivation

Low precision floating-point formats are increasingly supported by computer hardware.



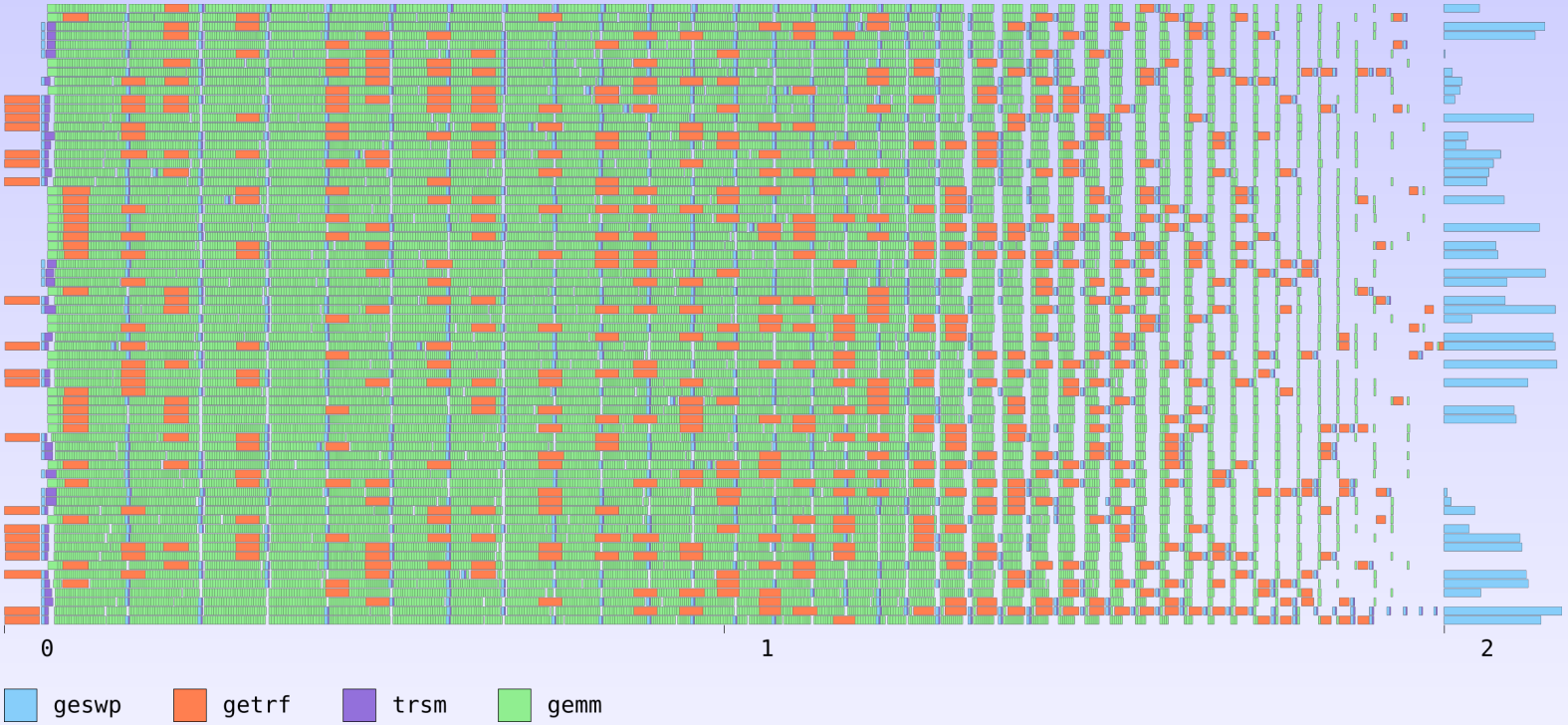
Source – `https://www.top500.org/statistics/list/`

Motivation Contd..

An algorithm with abundance of matrix multiply is well suited for these hardware. Like LU factorization:

Motivation Contd..

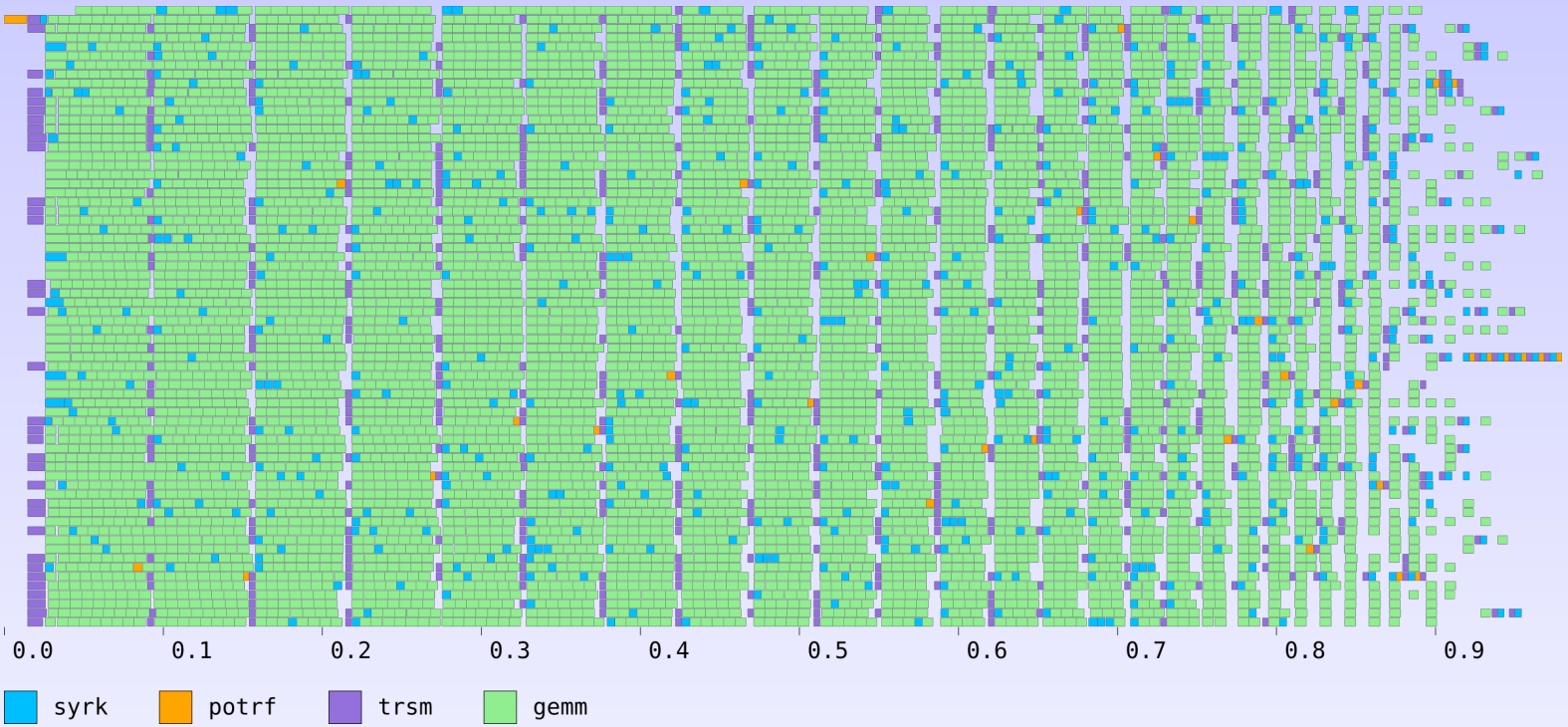
An algorithm with abundance of matrix multiply is well suited for these hardware. Like LU factorization:



Thanks Mawussi!

Motivation Contd..

Cholesky factorization is even better.



Thanks Mawussi!

Problem statement

Fast way to solve $Ax = b$, when A is symmetric and positive definite.

Problem statement

Fast way to solve $Ax = b$, when A is symmetric and positive definite.

For a general A:

Given A and b in precision u.

solve $Ax_0 = b$ using the LU factors of precision $u_f > u$

- $r = b - Ax_0$, in $u_r < u$.
- Solve $\widetilde{A}d \equiv \widehat{U}^{-1}\widehat{L}^{-1}A = \widehat{U}^{-1}\widehat{L}^{-1}r$, at precision u using GMRES.
- Update $x_1 = \mathrm{fl}(x_0 + d)$ in precision u.

u_f	u	u_r
half	single	double
half	double	quad
single	double	quad

Obvious solution

Given A and b in precision u.
solve $Ax_0 = b$ using the Cholesky factorization of precision $u_f > u$

- $r = b - Ax_0$, in $u_r < u$.
- Solve $\widetilde{A}\widetilde{d} \equiv (\widehat{R}^{-T} A \widehat{R}^{-1})(\widehat{R}d) = \widehat{R}^{-T} r$, at precision u using CG.
- Update $x_1 = \mathrm{fl}(x_0 + d)$ in precision u.

Obvious solution

Given A and b in precision u.
solve $Ax_0 = b$ using the Cholesky factorization of precision $u_f > u$

- $r = b - Ax_0$, in $u_r < u$.
- Solve $\widetilde{A}\widetilde{d} \equiv (\widehat{R}^{-T} A \widehat{R}^{-1})(\widehat{R}d) = \widehat{R}^{-T} r$, at precision u using CG.
- Update $x_1 = \mathrm{fl}(x_0 + d)$ in precision u.

Might not work!

- Cholesky of $\mathrm{fl_h}(A)$ might not succeed in u_h.
- Overflow and underflow.
- IR might not converge.

Diagonal perturbation

A diagonal perturbation to ensure the success of Cholesky factorization.

Diagonal perturbation

A diagonal perturbation to ensure the success of Cholesky factorization.

- $A+\epsilon_1 I$, where $\epsilon_1 \geq d_n \lambda_{\max}(A)u - \lambda_{\min}(A)$, $d_n = 20n^{3/2}$. Wilkinson's analysis. PertA
- $A+\epsilon_2 D^2$, where $D = \mathrm{diag}(a_{ii}^{1/2})$, $\epsilon_2 \geq n(n+1)u - \lambda_{\min}(H)$, $H = D^{-1}AD^{-1}$. Demmel's analysis. PertH

$$1 \leq \frac{\|\epsilon_1 I\|_2}{\|\epsilon_2 D^2\|_2} = \frac{\lambda_{\max}(A)}{\max_i a_{ii}} \leq n,$$

In practice $\lambda_{\max}(A) \approx \max_i a_{ii}$.

Diagonal perturbation contd..

- Example:

$$A = \begin{bmatrix} 10^{40} & \times & \times \\ \times & 10^{20} & \times \\ \times & \times & 1 \end{bmatrix},$$

For PertA, in double precision $\epsilon_1 I \approx 10^{24} I$. a_{33} is completely lost.

Diagonal perturbation contd..

- Example:

$$A = \begin{bmatrix} 10^{40} & \times & \times \\ \times & 10^{20} & \times \\ \times & \times & 1 \end{bmatrix},$$

For PertA, in double precision $\epsilon_1 I \approx 10^{24} I$. a_{33} is completely lost.

Diagonal perturbation by a multiple of $\mathrm{diag}(a_{ii})$ is better than a multiple of I, to minimize the loss of information.

Two-sides Diagonal Scaling

Two-sides Diagonal Scaling

2DS. Rounds a symmetric positive definite $A \in \mathbb{R}^{n\times n}$ to the fp16 matrix $A^{(h)}$, scaling all elements to avoid overflow. $\theta \in (0, 1]$ is a parameter, and c is a positive integer.

1: $D = \mathrm{diag}(a_{ii}^{1/2})$, $H = D^{-1}AD^{-1}$
2: $G = H + cu_l I$
3: Let $\beta = 1 + cu_l$
4: $\mu = \theta x_{\max}/\beta$
5: $A^{(h)} = \mathrm{fl_h}(\mu G)$

Two-sides Diagonal Scaling

2DS. Rounds a symmetric positive definite $A \in \mathbb{R}^{n\times n}$ to the fp16 matrix $A^{(h)}$, scaling all elements to avoid overflow. $\theta \in (0, 1]$ is a parameter, and c is a positive integer.

1: $D = \mathrm{diag}(a_{ii}^{1/2})$, $H = D^{-1}AD^{-1}$
2: $G = H + cu_l I$
3: Let $\beta = 1 + cu_l$
4: $\mu = \theta x_{\max}/\beta$
5: $A^{(h)} = \mathrm{fl_h}(\mu G)$

Remarks:

- $\mathrm{diag}(a_{ii}^{1/2})$ is a natural equilibration choice.
- Start with some fixed c is factorization fails then $c \leftarrow 2c$.
- $$\lambda_i(MA) - 1 = \frac{-cu_l}{\lambda_i + cu_l}.$$

Convergence of IR

Convergence of IR

- Analysis of GMRES-IR exploits backward stability of GMRES.
- For PCG b'err$\leq \mathcal{O}(u) \min\{\kappa_2(A)^{1/2}, \kappa_2(M)^{1/2}\}$.
- Same is true for any iterative solver based on three term recurrence.
- Theoretically convergence of IR with PCG cannot be guaranteed.

Convergence of IR

- Analysis of GMRES-IR exploits backward stability of GMRES.
- For PCG b'err$\leq \mathcal{O}(u)\min\{\kappa_2(A)^{1/2}, \kappa_2(M)^{1/2}\}$.
- Same is true for any iterative solver based on three term recurrence.
- Theoretically convergence of IR with PCG cannot be guaranteed.

			Backward error		Forward error	
u_f	u	u_r	$\kappa_\infty(A)$	Limit	$\kappa_\infty(A)$	Limit
half	single	double	10^7	single	10^7	single
half	double	double	10^6	double	10^7	$\text{cond}(A,x)\times$double
half	double	quad	10^{16}	double	10^{11}	double
single	double	double	10^7	double	10^{10}	$\text{cond}(A,x)\times$double

Numerical Experiments

- All symmetric positive definite matrices with $300 \leq n \leq 500$ are chosen from SuiteSparse Matrix Collection.
  - $\theta = 0.1$.
- Precisions, (half,single,double) and (half,double,quad).
- For fp16 `chop` function [Higham, **P**, 2019].
- quad precision using Advanpix.
- $M = \mu D^{-1} \widehat{R}^{-1} \widehat{R}^{-T} D^{-1}$ is used as the preconditioner to avoid the change of norm.
- Both GMRES and CG are used.
- $c = 2$ was sufficient to ensure the success of Cholesky factorization.
- Iterative refinement is terminated when b'err $\leq nu$.

#GMRES/CG iterations (#IR steps)

Index	(half, single, double)				(half, double, quad)			
	GMRES-IR		CG-IR		GMRES-IR		CG-IR	
1	3	(1)	4	(1)	16	(3)	16	(3)
2	2	(1)	2	(1)	10	(3)	12	(3)
3	2	(1)	2	(1)	9	(3)	9	(3)
4*	362	(1)	362	(1)	434	(2)	–	(–)
5	416	(1)	416	(1)	492	(2)	397	(1)
6	129	(1)	132	(1)	529	(3)	356	(2)
7	87	(1)	91	(1)	503	(3)	516	(3)
8	11	(1)	12	(1)	45	(3)	23	(2)
9	5	(1)	7	(1)	32	(3)	35	(3)
10	7	(1)	10	(1)	37	(3)	37	(3)
11*	485	(1)	485	(1)	485	(1)	–	(–)
12	24	(1)	30	(1)	122	(3)	84	(2)
13	47	(1)	61	(1)	248	(3)	235	(3)
14	4	(1)	5	(1)	19	(3)	19	(3)

#GMRES/CG iterations (#IR steps)

Index	(half, single, double)				(half, double, quad)			
	GMRES-IR		CG-IR		GMRES-IR		CG-IR	
1	0	(0)	0	(0)	4	(2)	4	(2)
2	2	(1)	2	(1)	6	(2)	6	(2)
3	2	(1)	2	(1)	6	(2)	6	(2)
4*	362	(1)	362	(1)	382	(2)	–	(–)
5	0	(0)	2	(1)	3	(2)	3	(1)
6	0	(0)	8	(1)	25	(2)	26	(2)
7	0	(0)	6	(1)	34	(3)	34	(3)
8	0	(0)	0	(0)	1	(1)	1	(1)
9	0	(0)	0	(0)	4	(2)	4	(2)
10	3	(1)	4	(1)	18	(3)	19	(3)
11*	0	(0)	0	(0)	124	(2)	71	(2)
12	0	(0)	0	(0)	1	(1)	1	(1)
13	0	(0)	0	(0)	27	(2)	20	(2)
14	0	(0)	0	(0)	4	(2)	4	(2)

Remarks

- Low precision factors obtained by perturbing H are better.
- Even though no theoretical guarantees for CG-IR it perform well in practice.
- (half,double,double) , and (single,double,double) are of practical importance. Even for this CG-IR and GMRES-IR have similar behavior.
- For u_l =single diagonal scaling is not required, but diagonal perturbation is required.
- Normal equation based least squares solver is rich in matrix multiply, and can be used for well conditioned A.

Conclusion

- Number of devices that support low precision, and can compute matrix multiply very quickly is increasing.
- Cholesky factorization is ideally suited for such devices.
- Diagonal scaling, and multiplication by $\theta x_{\max}$ addresses underflow and overflow issues.
- Diagonal perturbation ensures the success of low precision Cholesky factorization.
- CG-IR works well in practice.
- Further details " N.J. Higham, S. Pranesh. *Exploiting Lower Precision Arithmetic in Solving Symmetric Positive Definite Linear Systems and Least Squares Problems.*"

Conclusion

- Number of devices that support low precision, and can compute matrix multiply very quickly is increasing.
- Cholesky factorization is ideally suited for such devices.
- Diagonal scaling, and multiplication by $\theta x_{\max}$ addresses underflow and overflow issues.
- Diagonal perturbation ensures the success of low precision Cholesky factorization.
- CG-IR works well in practice.
- Further details " N.J. Higham, S. Pranesh. *Exploiting Lower Precision Arithmetic in Solving Symmetric Positive Definite Linear Systems and Least Squares Problems.*"

Thank You.
Questions ???