MANCHESTER 1824

The University of Manchester

M\cr NA
Manchester Numerical Analysis

The Rise of Multiprecision Computations

Nick Higham
School of Mathematics
The University of Manchester

`http://www.ma.man.ac.uk/~higham`
`@nhigham, nickhigham.wordpress.com`

SAMSI, April 26, 2017

Multiprecision arithmetic: floating point arithmetic supporting multiple, possibly arbitrary, precisions.

- Applications of & support for low precision.
- Applications of & support for high precision.
- How to adapt algorithms to achieve high accuracy—especially **iterative refinement**.

Download this talk from
http://bit.ly/higham-samsi17

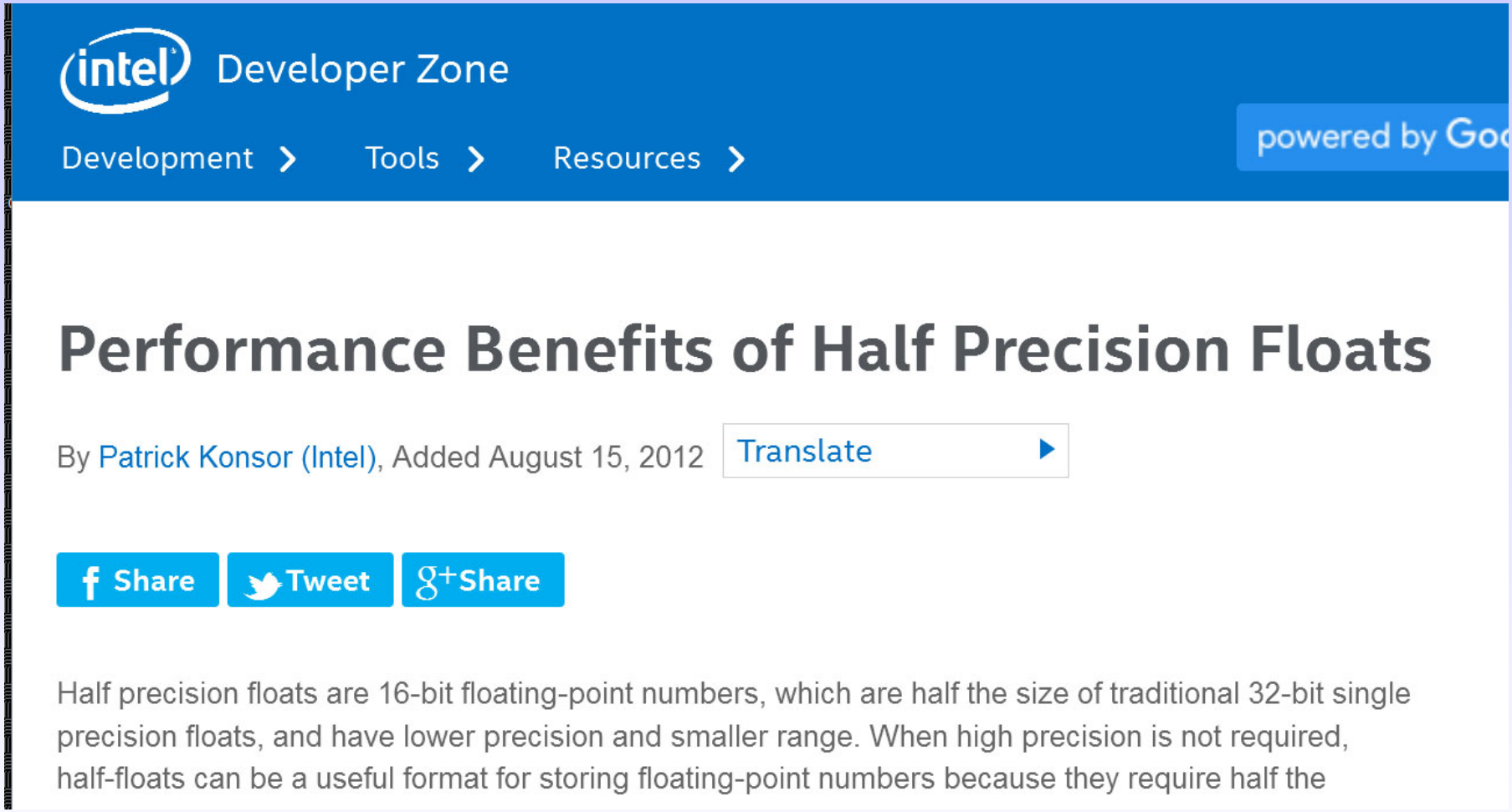
IEEE Standard 754-1985 and 2008 Revision

Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- Arithmetic ops $(+, -, *, /, \sqrt{})$ performed *as if* first calculated to infinite precision, then rounded.
- Default: round to nearest, round to even in case of tie.
- Half precision is a *storage format only*.

Intel Core Family (3rd gen., 2012)

Ivy Bridge supports half precision for storage.



intel Developer Zone

Development › Tools › Resources ›

powered by Goo

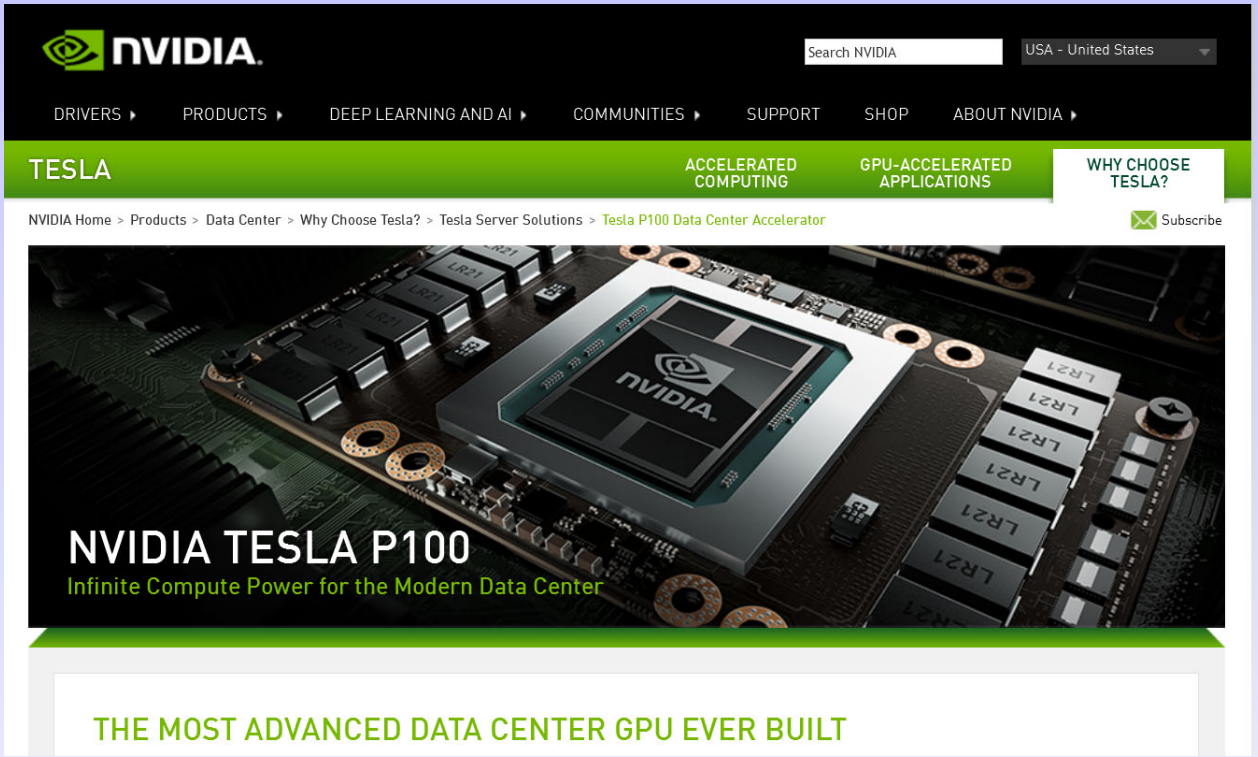
Performance Benefits of Half Precision Floats

By Patrick Konsor (Intel), Added August 15, 2012 Translate

Share Tweet g+Share

Half precision floats are 16-bit floating-point numbers, which are half the size of traditional 32-bit single precision floats, and have lower precision and smaller range. When high precision is not required, half-floats can be a useful format for storing floating-point numbers because they require half the

NVIDIA Tesla P100 (2016)

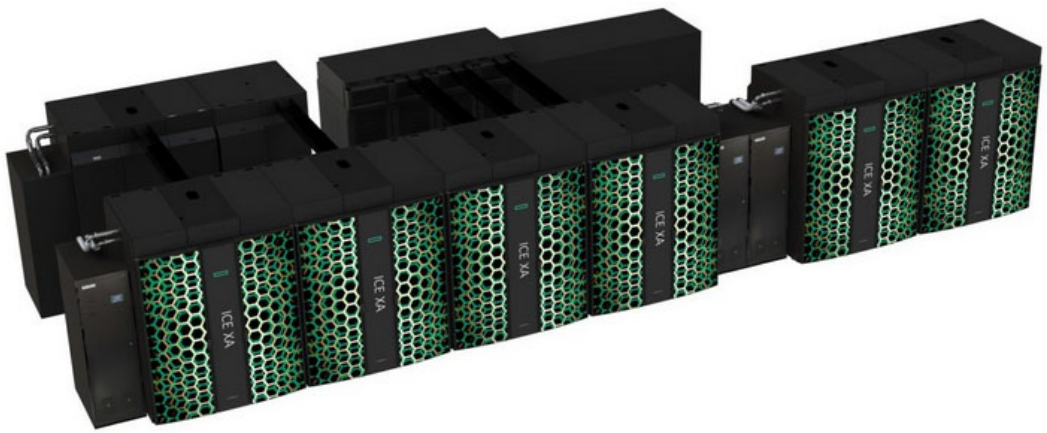


"The Tesla P100 is the world's first accelerator built for deep learning, and has native hardware ISA support for **FP16** arithmetic, delivering over 21 TeraFLOPS of **FP16** processing power."

TSUBAME 3.0 (HPC Wire, Feb 16, 2017)

In a press event Friday afternoon local time in Japan, Tokyo Institute of Technology (Tokyo Tech) announced its plans for the TSUBAME3.0 supercomputer, which will be Japan's "fastest AI supercomputer," when it comes online this summer (2017). Projections are that it will deliver 12.2 double-precision petaflops and 64.3 half-precision (peak specs).

Nvidia was the first vendor to publicly share the news in the US. We know that Nvidia will be supplying Pascal P100 GPUs, but the big surprise here is the system vendor. The Nvidia blog did not specifically mention HPE or SGI but it did include this photo with a caption referencing it as TSUBAME3.0:



TSUBAME3.0 – *click to expand* (Source: Nvidia)

Japan to Build 130 Petaflop ABCI Supercomputer

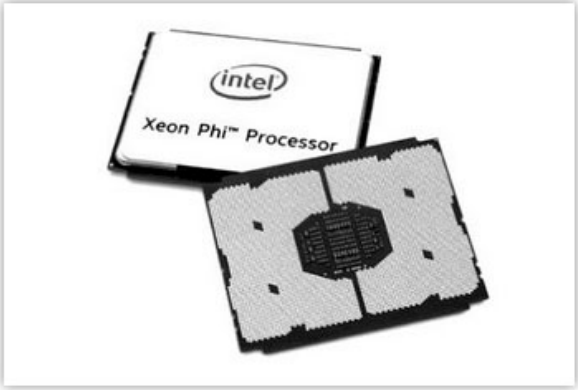
November 25, 2016 by Rich Brueckner

Today Japan announced plans to build a 130 Petaflop (half or single precision) supercomputer for deployment in 2017. And while such a machine would not surpass the current #1 93 Petaflop Sunway TaihuLight supercomputer in China, it would certainly propel Japan to the top of an all new category of supercomputing leadership.

November 15, 2016: "confirmed . . . the future "Knights Mill" variant of the current Knights Landing Xeon Phi processor would support mixed precision math . . . This should mean **16-bit** floating point as well as the normal 32-bit and 64-bit variants, but Intel could have baked 8-bit support in there, too."

WHY INTEL IS TWEAKING XEON PHI FOR DEEP LEARNING

August 22, 2016 Timothy Prickett Morgan



If there is anything that chip giant Intel has learned over the past two decades as it has gradually climbed to dominance in processing in the datacenter, it is ironically that one size most definitely does not fit all. Quite the opposite, and increasingly so.

As the tight co-design of hardware and software continues in all parts of the IT industry, we can expect fine-grained customization for very precise – and lucrative – workloads, like data analytics and machine learning, just to name two of the hottest areas today.

"for machine learning as well as for certain image processing and signal processing applications, *more data at lower precision actually yields better results with certain algorithms than a smaller amount of more precise data*."

Google Tensorflow Processor

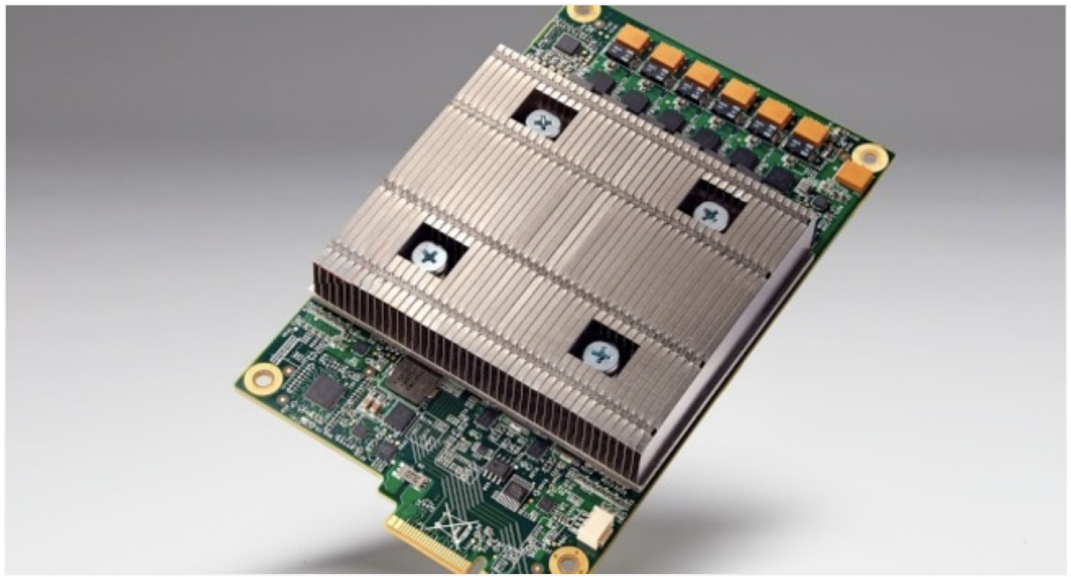
The Register®
Biting the hand that feeds IT

DATA CENTER SOFTWARE SECURITY TRANSFORMATION DEVOPS BUSINESS PERSONAL TECH

Emergent Tech ▸ Artificial Intelligence

Revealed: Blueprints to Google's AI FPU aka the Tensor Processing Unit

PCIe-connected super-calculator trounces outdated competition



"The TPU is special-purpose hardware designed to accelerate the inference phase in a neural network, in part through quantizing 32-bit floating point computations into **lower-precision 8-bit arithmetic**."

Machine Learning

Courbariaux, Benji & David (2015)

> We find that very low precision is sufficient not just for running trained networks but also for training them.

- We are solving the wrong problem anyway (Scheinberg, 2016), so don't need an accurate solution.
- Low precision provides regularization.
- See Jorge Nocedal's plenary talk **Stochastic Gradient Methods for Machine Learning** at SIAM CSE 2017.

Climate Modelling

- **T. Palmer**, *More reliable forecasts with less precise computations: a fast-track route to cloud-resolved weather and climate simulators?*, Phil. Trans. R. Soc. A, 2014:

  > Is there merit in representing variables at sufficiently high wavenumbers using half or even quarter precision floating-point numbers?

- **T. Palmer**, *Build imprecise supercomputers*, Nature, 2015.

Error Analysis in Low Precision

For an inner product $x^T y$ of n-vectors the standard error bound is

$$|\mathrm{fl}(x^T y) - x^T y| \leq nu|x|^T|y| + O(u^2).$$

In half precision, $u \approx 4.9 \times 10^{-4}$, so $nu = 1$ for $n = 2048$.

Error Analysis in Low Precision

For an inner product $x^T y$ of n-vectors the standard error bound is

$$|\mathrm{fl}(x^T y) - x^T y| \leq nu|x|^T|y| + O(u^2).$$

In half precision, $u \approx 4.9 \times 10^{-4}$, so $nu = 1$ for $n = 2048$.

Most existing rounding error analysis guarantees no accuracy, *and maybe not even a correct exponent*, for half precision!

Error Analysis in Low Precision

For an inner product $x^T y$ of n-vectors the standard error bound is

$$|\mathrm{fl}(x^T y) - x^T y| \leq nu|x|^T|y| + O(u^2).$$

In half precision, $u \approx 4.9 \times 10^{-4}$, so $nu = 1$ for $n = 2048$.

Most existing rounding error analysis guarantees no accuracy, *and maybe not even a correct exponent*, for half precision!

Is standard error analysis *especially pessimistic* in these applications? Try a *statistical approach*?

Need for Higher Precision

- He and Ding, **Using Accurate Arithmetics to Improve Numerical Reproducibility and Stability in Parallel Applications**, 2001.
- Bailey, Barrio & Borwein, **High-Precision Computation: Mathematical Physics & Dynamics**, 2012.
- Khanna, **High-Precision Numerical Simulations on a CUDA GPU: Kerr Black Hole Tails**, 2013.
- Beliakov and Matiyasevich, **A Parallel Algorithm for Calculation of Determinants and Minors Using Arbitrary Precision Arithmetic**, 2016.
- Ma and Saunders, **Solving Multiscale Linear Programs Using the Simplex Method in Quadruple Precision**, 2015.

Higher Precision to Avoid Overflow

Zimbabwe resorts to the $100 trillion note

By Our Foreign Staff

ZIMBABWE'S central bank will introduce a 100 trillion Zimbabwean dollar banknote, worth a
market
ate cas
media

are doubling every day and food and fuel are in short supply.

A cholera epidemic has killed more than 2,000 people

has forced the central bank to continue to release new banknotes which quickly become almost worthless. There is an official exchange rate, but

dollar note, the Reserve Bank of Zimbabwe plans to introduce Z$10 trillion, Z$20 trillion and Z$50 trillion notes, the *Herald* newspaper reported.

e up
s to
h to

Old Mutual's new chief weighs rescue options

JUDGING by the empty state of his spacious South African office, it is quite clear that Julian Roberts has yet to settle into his role as the new chief executive of Old Mutual.

While his secretary bustles around, tidying away his few possessions – a 5p piece and a penny coin left lying on his desk – the four books on his vacant shelves stand out. The titles *Blown to Bits* and *On the Brink of Failure* could almost sum up the state of the blue-chip company Mr Roberts has just taken over. Old Mutual was the worst-performing European

PROFILE

Julian Roberts

Chief executive, Old Mutual

The economic turmoil revealed cracks in Old Mutual's model when it emerged that its $2.8bn (£1.9bn) variable annuity business in the US could not meet guarantees due to adverse movements in the Asian markets. It has been forced to inject

going to be immune. South Africa lags the rest of the world by six months to a year."

Political tensions are also playing on his mind. Old Mutual is listed not only in the UK and Johannesburg but also on the Zimbabwe Stock Exchange. Due to technical difficulties of transferring a figure with so many noughts on the end of it, Old Mutual struggled to pay shareholders an interim dividend of Z$453 trillion per share – which in November equated to just 2.45p.

"It is absolutely tragic. We have a significant business with a large

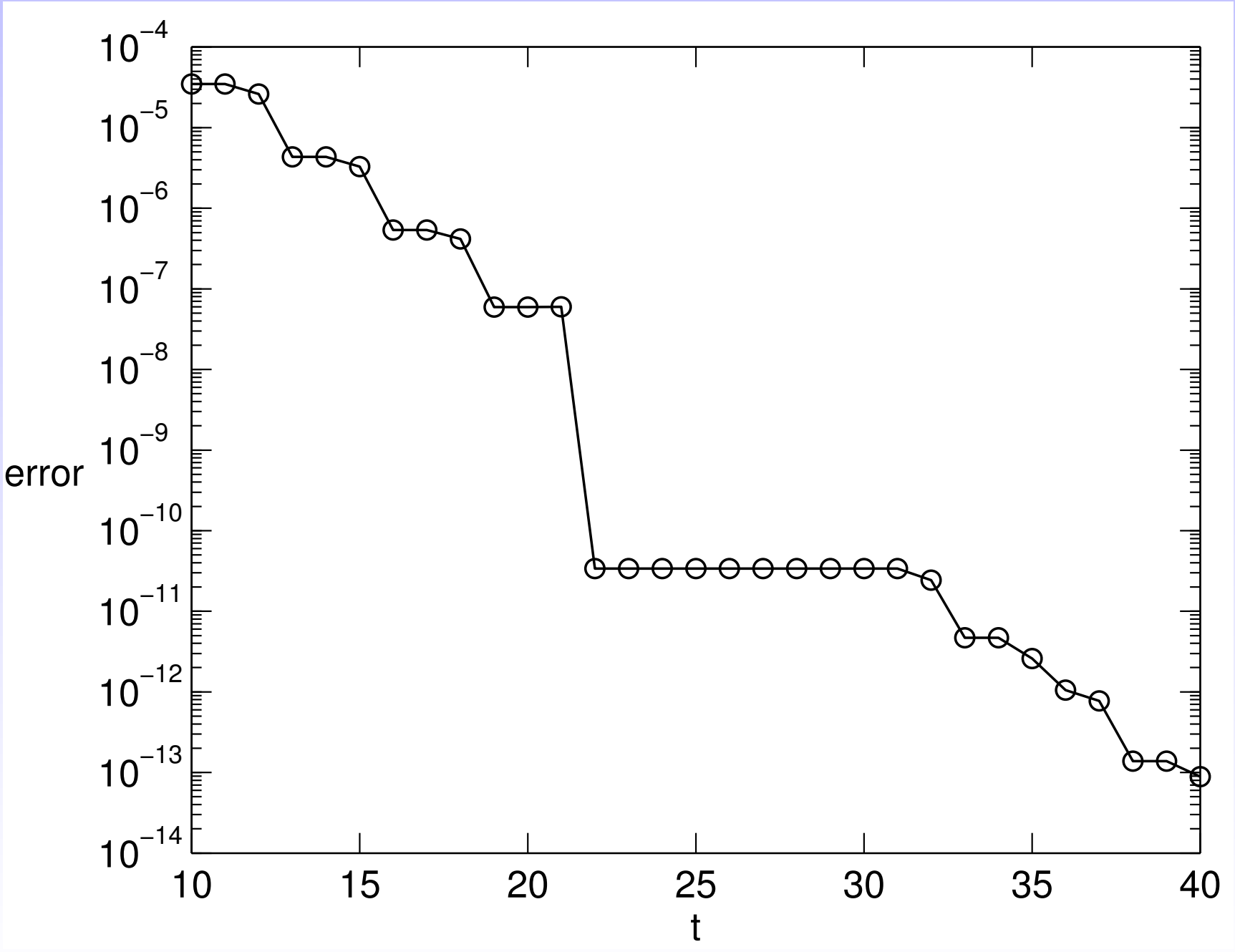
Increasing the Precision

Myth

Increasing the precision at which a computation is performed increases the accuracy of the answer.

Consider the evaluation in precision $u = 2^{-t}$ of

$$y = x + a\sin(bx), \quad x = 1/7, \quad a = 10^{-8}, \quad b = 2^{24}.$$


error
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
10^{-10}
10^{-11}
10^{-12}
10^{-13}
10^{-14}
10
15
20
25
30
35
40
t

IBM z13 Mainframe Systems

z13 processor (2015) has **quadruple precision** in the vector & floating point unit.

Lichtenau, Carlough & Mueller (2016):

"designed to maximize performance for **quad precision** floating-point operations that are occurring with increased frequency on Business Analytics workloads . . .

on commercial products like ILOG and SPSS, replacing double precision operations with **quad-precision** operations in critical routines yield 18% faster convergence due to reduced rounding error.

Availability of Multiprecision in Software

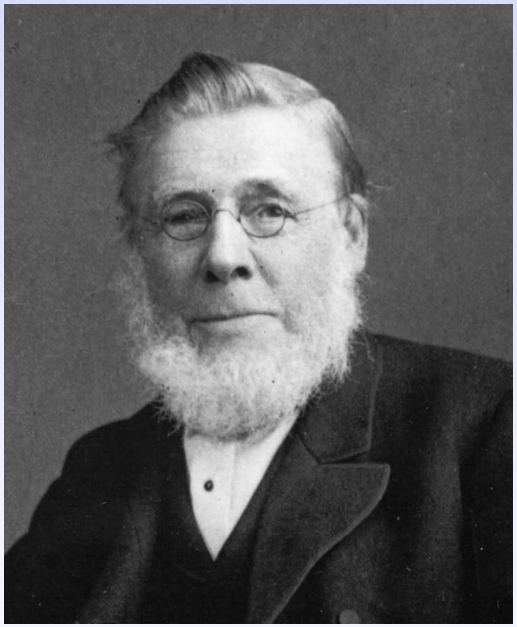
- **Maple**, Mathematica, PARI/GP, **Sage**.
- MATLAB: Symbolic Math Toolbox, **Multiprecision Computing Toolbox** (Advanpix).
- Julia: **BigFloat**.
- Mpmath and SymPy for Python.
- GNU MP Library.
- **GNU MPFR Library**.
- (Quad only): some C, Fortran compilers.

Gone, but not forgotten:

- Numerical Turing: **Hull et al.**, 1985.

Note on Log Tables

Name	Year	Range	Decimal places
R. de Prony	1801	$1 - 10,000$	19
Edward Sang	1875	$1 - 20,000$	28



Age 82

Edward Sang (1805–1890).
Born in Kirkcaldy.
Teacher of maths and actuary in Edinburgh.

Going to Higher Precision

If we have quadruple or higher precision, how can we modify existing algorithms to exploit it?

Going to Higher Precision

If we have quadruple or higher precision, how can we modify existing algorithms to exploit it?

To what extent are existing algs precision-independent?

- Newton-type algs: just decrease `tol`?
- How little higher precision can we get away with?
- Gradually increase precision through the iterations?

Matrix Functions

(Inverse) scaling and squaring-type algorithms for e^A, $\log A$, $\cos A$, A^t use Padé approximants.

- Padé degree and algorithm parameters chosen to achieve double precision accuracy, $u = 2^{-53}$.
- Change u and the algorithm logic needs changing!
- **H & Fasi**, 2017: **Multiprecision Algorithms for Computing the Matrix Logarithm**.

- Open questions even for scalar elementary functions!

Accurate Solution of $Ax = b$

Joint work with Erin Carson (NYU).

- Base precision u; extended precision $\overline{u}$.
- A, b are given in precision u (known exactly).
- A is $n \times n$ and nonsingular.
- Want x correct to precision u.

Accurate Solution of $Ax = b$

Joint work with Erin Carson (NYU).

- Base precision u; extended precision $\overline{u}$.
- A, b are given in precision u (known exactly).
- A is $n \times n$ and nonsingular.
- Want x correct to precision u.

Allow $\kappa(A) = \|A\|\|A^{-1}\| \gtrsim u^{-1}$.

- Reference solutions for testing solvers.
- Radial basis functions.
- Ill-conditioned FE geomechanical problems.

Base precision may be half or single!

Iterative Refinement

Given x_0.

- $r = b - Ax_0$ **quad precision**
- Solve $Ad = r$ **double precision**
- $x_1 = \mathrm{fl}(x_0 + d)$ **double precision**

Goes back at least to **Wilkinson's** *Progress Report on the Automatic Computing Engine* (1948).

Simple Analysis

Given x_0, *assume*

- $r = b - Ax_0$ **exact**
- Solve $Ad = r$ **perfect backward error**
- $x_1 = \mathrm{fl}(x_0 + d)$ **exact**

Then

$$(A + E)\widehat{d} = r, \qquad |E| \le u|A|.$$

Then $x - x_1 \approx -A^{-1}E(x - x_0)$ and so

$$\begin{aligned} \|x - x_1\| &\lesssim \| \, |A^{-1}||A| \, \|_\infty u \|x - x_0\| \\ &\le \kappa(A)u\|x - x_0\| \end{aligned}$$

is the best bound we can obtain.

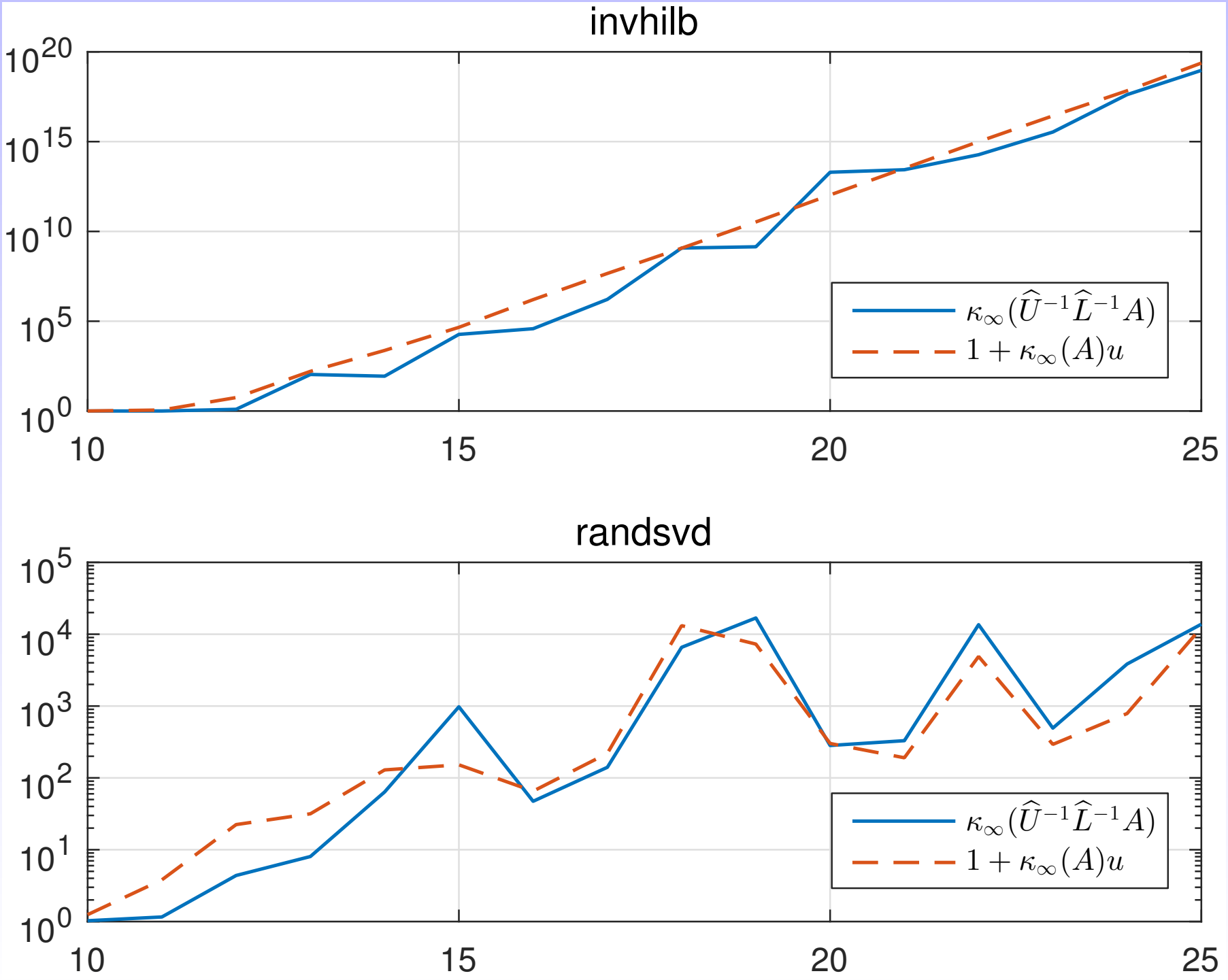
Is there any hope when $\kappa(A) \gtrsim u^{-1}$?

Why There is Hope

Empirically observed by Rump (1990) that if $\widehat{L}$ and $\widehat{U}$ are computed LU factors of A from GEPP then

$$\kappa(\widehat{L}^{-1}A\widehat{U}^{-1}) \approx 1 + \kappa(A)u,$$

even for $\kappa(A) \gg u^{-1}$.


invhilb
10^{20}
10^{15}
10^{10}
10^{5}
10^{0}
10
15
20
25
$\kappa_\infty(\widehat{U}^{-1}\widehat{L}^{-1}A)$
$1+\kappa_\infty(A)u$
randsvd
10^{5}
10^{4}
10^{3}
10^{2}
10^{1}
10^{0}
10
15
20
25
$\kappa_\infty(\widehat{U}^{-1}\widehat{L}^{-1}A)$
$1+\kappa_\infty(A)u$

Existing Rounding Error Analysis

- **Wilkinson** (1963): fixed-point arithmetic.
- **Moler** (1967): floating-point arithmetic.
- **Higham** (1997, 2002): more general analysis for arbitrary solver.

All the above have the $\kappa(A)u < 1$ limitation.

Existing Rounding Error Analysis

- **Wilkinson** (1963): fixed-point arithmetic.
- **Moler** (1967): floating-point arithmetic.
- **Higham** (1997, 2002): more general analysis for arbitrary solver.

All the above have the $\kappa(A)u < 1$ limitation.

New Analysis

Compute residual $r_i = b - Ax_i$ in precision $\overline{u}$.
Assume computed solution $\widehat{y}$ to $Ay = c$ satisfies

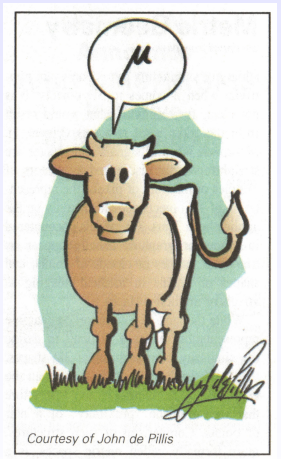
$$\frac{\|y - \widehat{y}\|_\infty}{\|y\|} \leq \theta\, u, \quad \theta u \leq 1.$$

Define $\mu_i^{(p)}$ by

$$\|A(x - x_i)\|_p = \mu_i^{(p)} \|A\|_p \|x - x_i\|_p,$$

and note that

$$\kappa_p(A)^{-1} \leq \mu_i^{(p)} \leq 1.$$



Convergence Result

Theorem

For IR in precisions u and $\overline{u} \leq u$ applied to a nonsingular linear system $Ax = b$ the computed iterate x_{i+1} satisfies

$$\begin{aligned}\|x_{i+1} - x\|_\infty \lesssim \Big(\mu_i^{(\infty)} \kappa_\infty(A)u + \theta_i\, u \Big) \|x - x_i\|_\infty \\ + n\overline{u}(1 + \theta_i u) \| \, |A^{-1}|(|b| + |A||x_i|) \, \|_\infty \\ + u\|x_{i+1}\|_\infty.\end{aligned}$$

Analogous standard bound would have

- $\mu_i^{(\infty)} = 1$,
- $\theta_i = \kappa(A)$.

Bounding μ_i

For the 2-norm, can show that

$$\mu_i \leq \frac{\|r_i\|_2}{\|P_k r_i\|_2} \frac{\sigma_{n+1-k}}{\sigma_1},$$

where $A = U\Sigma V^T$ is an SVD, $P_k = U_k U_k^T$ with $U_k = [u_{n+1-k}, \ldots, u_n]$.

For a stable solver, in the early stages we expect

$$\frac{\|r_i\|}{\|A\|\|x_i\|} \approx u \ll \frac{\|x - x_i\|}{\|x\|},$$

or equivalently $\mu_i \ll 1$. But **close to convergence**

$$\|r_i\| \approx \|A\|\|x - x_i\| \quad \text{or} \quad \mu_i \approx 1.$$

Bounding θ_i

- θ_i bounds rel error in solution of $Ad_i = r_i$.
- We need $\theta_i u \ll 1$.

Standard solvers cannot achieve this!

We apply **GMRES** to

$$\underbrace{\widehat{U}^{-1}\widehat{L}^{-1}A}_{\widetilde{A}}\, d_i = \widehat{U}^{-1}\widehat{L}^{-1} r_i.$$

- $\kappa(\widetilde{A}) \ll \kappa(A)$ typically.
- Rounding error analysis shows we get an accurate $\widehat{d}_i$.

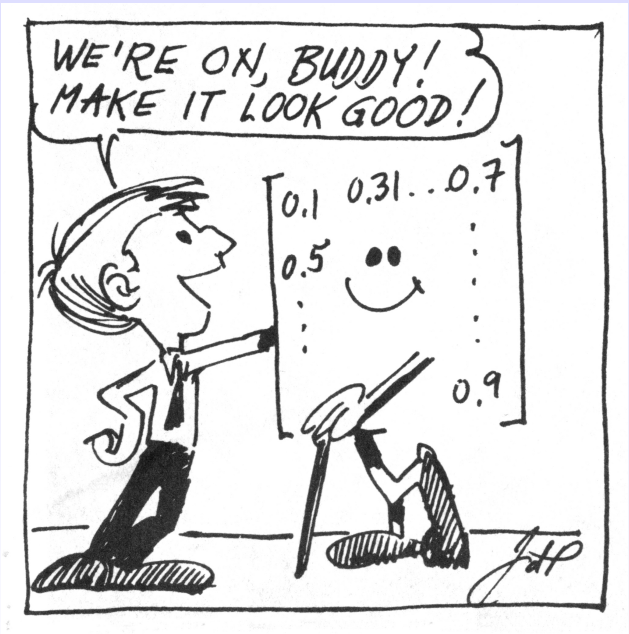
The New, Two-Stage Algorithm

- Solve $Ax = b$ by LU with partial pivoting.
- Apply standard IR.
- If IR diverges, apply IR with preconditioned GMRES.

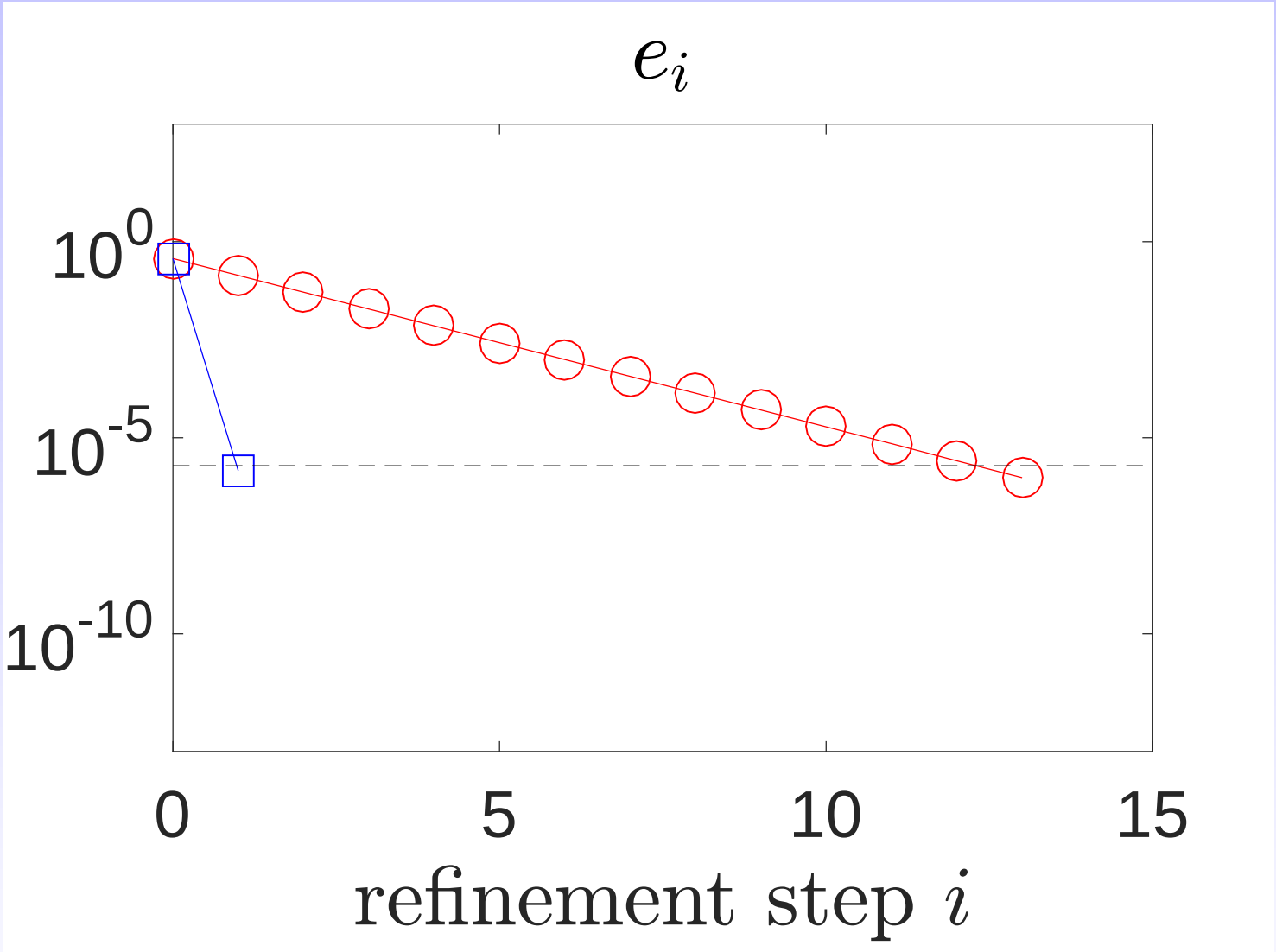
The New, Two-Stage Algorithm

- Solve $Ax = b$ by LU with partial pivoting.
- Apply standard IR.
- If IR diverges, apply IR with preconditioned GMRES.

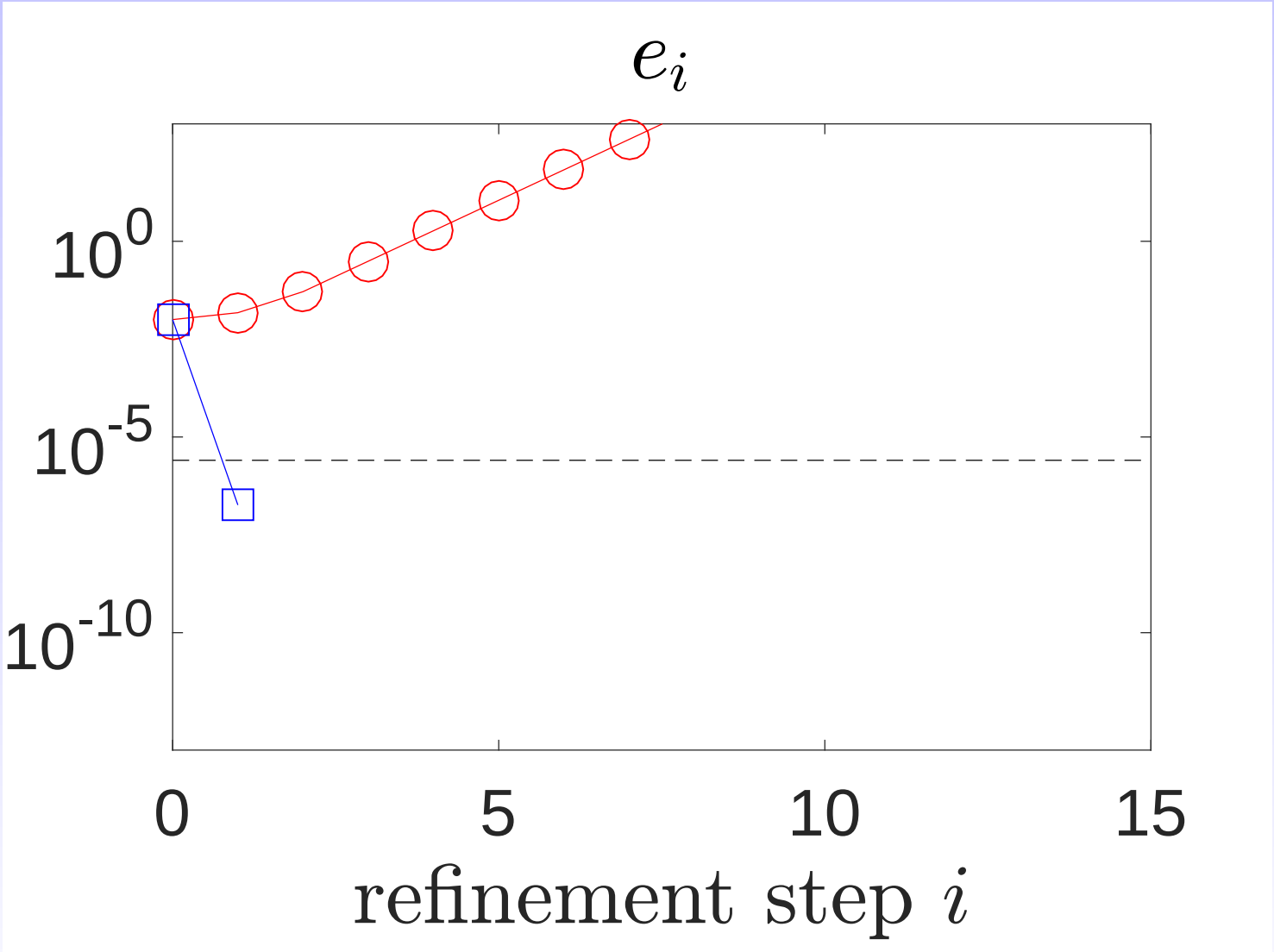
Tests with matrices from *University of Florida Sparse Matrix Collection* . . .



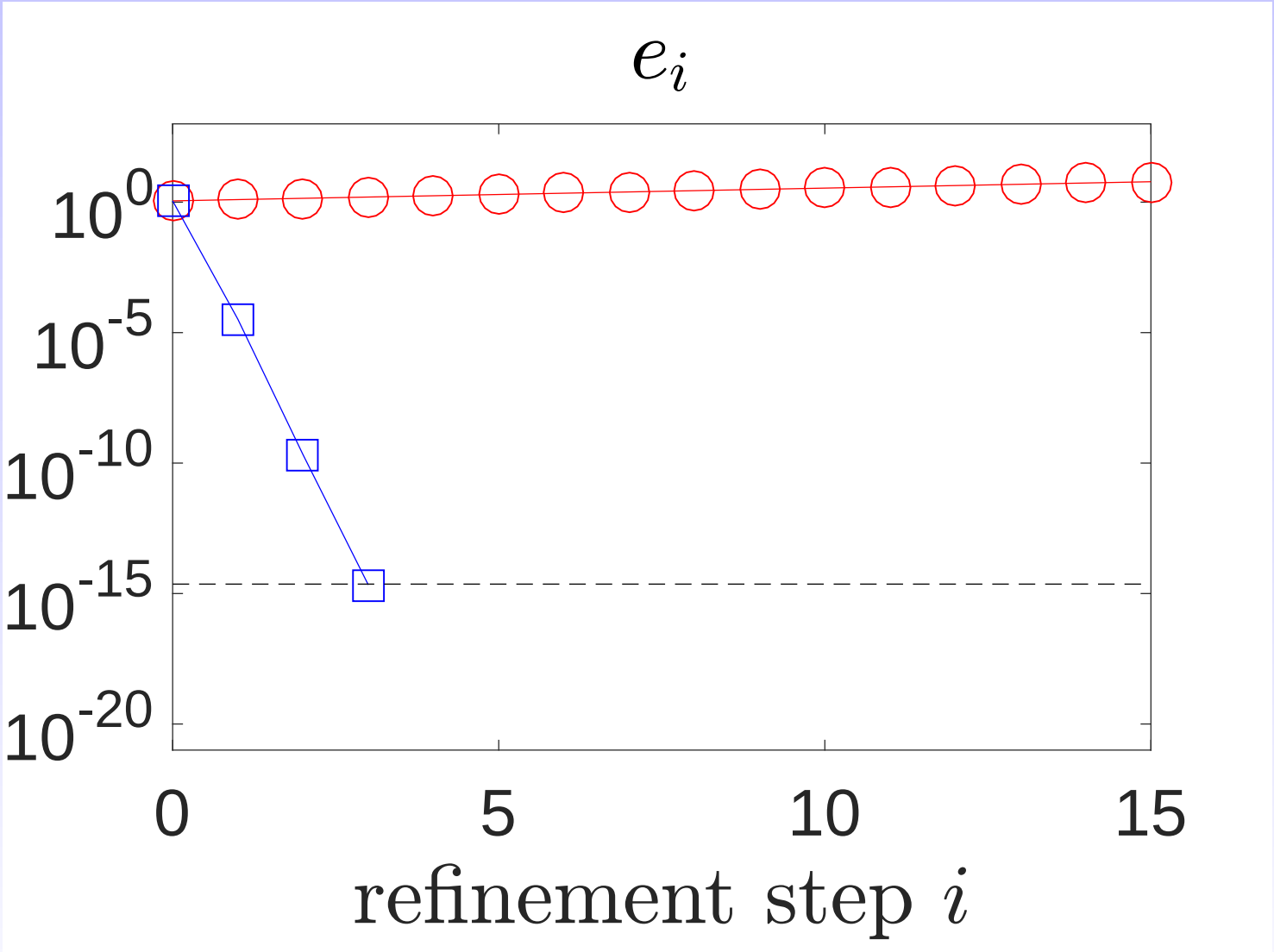
radfr1, $\kappa_\infty(A) = 10^{11}$, $u =$ single



adder_dcop_26, $\kappa_\infty(A) = 10^{11}$, $u =$ single



oscil_dcop_43, $\kappa_\infty(A) = 10^{21}$, $u =$ double



Conclusions

- Both **low and high precision** floating-point arithmetic becoming more prevalent, in hardware and software.
- Need **better understanding** of behaviour of algs in low precision arithmetic.
- Judicious use of **a little high precision** can bring major benefits.
- Identified mechanism allowing **iter ref** to produce **accurate solutions** when $\kappa(A) \gtrsim u^{-1}$, provided update equation is solved with some accuracy.

References I

D. H. Bailey, R. Barrio, and J. M. Borwein.
High-precision computation: Mathematical physics and dynamics.
Appl. Math. Comput., 218(20):10106–10121, 2012.

G. Beliakov and Y. Matiyasevich.
A parallel algorithm for calculation of determinants and minors using arbitrary precision arithmetic.
BIT, 56(1):33–50, 2015.

References II

E. Carson and N. J. Higham.
A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems.
MIMS EPrint 2017.12, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, Mar. 2017.
23 pp.

M. Courbariaux, Y. Bengio, and J.-P. David.
Training deep neural networks with low precision multiplications, 2015.
ArXiv preprint 1412.7024v5.

References III


A. D. D. Craik.
The logarithmic tables of Edward Sang and his daughters.
Historia Mathematica, 30(1):47–84, 2003.


Y. He and C. H. Q. Ding.
Using accurate arithmetics to improve numerical reproducibility and stability in parallel applications.
J. Supercomputing, 18(3):259–277, 2001.


N. J. Higham.
Iterative refinement for linear systems and LAPACK.
IMA J. Numer. Anal., 17(4):495–509, 1997.

References IV

N. J. Higham.
Accuracy and Stability of Numerical Algorithms.
Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.
ISBN 0-89871-521-0.
xxx+680 pp.

G. Khanna.
High-precision numerical simulations on a CUDA GPU: Kerr black hole tails.
J. Sci. Comput., 56(2):366–380, 2013.

References V

C. Lichtenau, S. Carlough, and S. M. Mueller.
Quad precision floating point on the IBM z13.
In *2016 IEEE 23nd Symposium on Computer Arithmetic (ARITH)*, pages 87–94, July 2016.

D. Ma and M. Saunders.
Solving multiscale linear programs using the simplex method in quadruple precision.
In M. Al-Baali, L. Grandinetti, and A. Purnama, editors, *Numerical Analysis and Optimization*, number 134 in Springer Proceedings in Mathematics and Statistics, pages 223–235. Springer-Verlag, Berlin, 2015.

References VI

K. Scheinberg.
Evolution of randomness in optimization methods for supervised machine learning.
SIAG/OPT Views and News, 24(1):1–8, 2016.