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

SAMSII, 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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
<th>Range</th>
<th>$u = 2^{-t}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>half</td>
<td>16 bits</td>
<td>$10^{±5}$</td>
<td>$2^{-11} \approx 4.9 \times 10^{-4}$</td>
</tr>
<tr>
<td>single</td>
<td>32 bits</td>
<td>$10^{±38}$</td>
<td>$2^{-24} \approx 6.0 \times 10^{-8}$</td>
</tr>
<tr>
<td>double</td>
<td>64 bits</td>
<td>$10^{±308}$</td>
<td>$2^{-53} \approx 1.1 \times 10^{-16}$</td>
</tr>
<tr>
<td>quadruple</td>
<td>128 bits</td>
<td>$10^{±4932}$</td>
<td>$2^{-113} \approx 9.6 \times 10^{-35}$</td>
</tr>
</tbody>
</table>

- Arithmetic ops (+, −, *, /, √) 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*.
Ivy Bridge supports half precision for storage.

Performance Benefits of Half Precision Floats

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

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
“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.”
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](Image)

TSUBAME3.0 – click to expand (Source: Nvidia)
Japan to Build 130 Petaflop ABCI Supercomputer

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.”
“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.”
“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.”
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.

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?

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

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

In half precision, $u \approx 4.9 \times 10^{-4}$, so $nu = 1$ for $n = 2048$. 
For an inner product $x^T y$ of $n$-vectors the standard error bound is

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

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

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

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

- Ma and Saunders, *Solving Multiscale Linear Programs Using the Simplex Method in Quadruple Precision*, 2015.
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 mere 0.8 cent at the official exchange rate but which has become almost worthless. A cholera epidemic has killed more than 2,000 people and food and fuel are in short supply. The Reserve Bank of Zimbabwe plans to introduce Z$10 trillion, Z$20 trillion and Z$50 trillion notes, the Herald newspaper reported. The bank's governor has said the notes are needed to keep up with rising prices, and that 100 trillion notes are expected to go into circulation in 2018.

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 insurance company in 2013, a year in which the company’s profits fell more than 30%.

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 $1bn of additional capital. "The UK government is going to be immune. South Africa lags the rest of the world by six to 12 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 share of the market and it is not possible to deal with it.

Nick Higham
The Rise of Multiprecision Computations
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

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

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Range</th>
<th>Decimal places</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. de Prony</td>
<td>1801</td>
<td>1 – 10,000</td>
<td>19</td>
</tr>
<tr>
<td>Edward Sang</td>
<td>1875</td>
<td>1 – 20,000</td>
<td>28</td>
</tr>
</tbody>
</table>


Age 82
If we have quadruple or higher precision, how can we modify existing algorithms to exploit it?
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 $\textit{tol}$?
- How little higher precision can we get away with?
- Gradually increase precision through the iterations?
(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!

- Open questions even for scalar elementary functions!
Accurate Solution of $Ax = b$

Joint work with Erin Carson (NYU).

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

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

Reference solutions for testing solvers.

Radial basis functions.

Ill-conditioned FE geomechanical problems.

Base precision may be half or single!
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 $u$.

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

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

Base precision may be half or single!
Given $x_0$.

- $r = b - Ax_0$ \hspace{1cm} quad precision
- Solve $Ad = r$ \hspace{1cm} double precision
- $x_1 = \text{fl}(x_0 + d)$ \hspace{1cm} double precision

Given \( x_0 \), *assume*

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

Then

\[
(A + E)\hat{d} = r, \quad |E| \leq u|A|.
\]

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

\[
\|x - x_1\| \lesssim \|A^{-1}\|A\|\_\infty u\|x - x_0\| \\
\leq \kappa(A)u\|x - x_0\|
\]

is the best bound we can obtain.

Is there any hope when \( \kappa(A) \gtrsim u^{-1} \)?
Empirically observed by Rump (1990) that if \( \hat{L} \) and \( \hat{U} \) are computed LU factors of \( A \) from GEPP then

\[
\kappa(\hat{L}^{-1}A\hat{U}^{-1}) \approx 1 + \kappa(A)u,
\]

even for \( \kappa(A) \gg u^{-1} \).
Existing Rounding Error Analysis

- **Wilkinson** (1963): fixed-point arithmetic.
- **Moler** (1967): floating-point arithmetic.

All the above have the $\kappa(A)u < 1$ limitation.
- **Wilkinson** (1963): fixed-point arithmetic.
- **Moler** (1967): floating-point arithmetic.

All the above have the $\kappa(A)u < 1$ limitation.
Compute residual \( r_i = b - Ax_i \) in precision \( \bar{u} \).
Assume computed solution \( \hat{y} \) to \( Ay = c \) satisfies
\[
\frac{\| y - \hat{y} \|_\infty}{\| y \|} \leq \theta u, \quad \theta u \leq 1.
\]

Define \( \mu_i^{(p)} \) by
\[
\| A(x - x_i) \|_p = \mu_i^{(p)} \| A \|_p \| x - x_i \|_p,
\]
and note that
\[
\kappa_p(A)^{-1} \leq \mu_i^{(p)} \leq 1.
\]
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

$$
\|x_{i+1} - x\|_{\infty} \lesssim \left( \mu_i^{(\infty)} \kappa_{\infty}(A) u + \theta_i u \right) \|x - x_i\|_{\infty} \\
+ n\overline{u}(1 + \theta_i u) \|A^{-1}(|b| + |A||x_i|)\|_{\infty} \\
+ u \|x_{i+1}\|_{\infty}.
$$

Analogous standard bound would have

- $\mu_i^{(\infty)} = 1$,
- $\theta_i = \kappa(A)$.
For the 2-norm, can show that

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

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 $\theta_i$

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

Standard solvers cannot achieve this!

We apply **GMRES** to

$$\hat{U}^{-1} \hat{L}^{-1} A d_i = \hat{U}^{-1} \hat{L}^{-1} r_i.$$  

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


- 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, $\kappa_\infty (A) = 10^{11}$, $u = \text{single}$
adder_dcop_26, $\kappa_\infty(A) = 10^{11}$, $u = \text{single}$
oscil_dcop_43, $\kappa_\infty(A) = 10^{21}$, $u = \text{double}$

![Graph showing the relationship between refinement step $i$ and $e_i$](image-url)
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.

E. Carson and N. J. Higham.  
A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems.  
23 pp.

M. Courbariaux, Y. Bengio, and J.-P. David.  
ArXiv preprint 1412.7024v5.
A. D. D. Craik.
The logarithmic tables of Edward Sang and his daughters.

Y. He and C. H. Q. Ding.
Using accurate arithmetics to improve numerical reproducibility and stability in parallel applications.

N. J. Higham.
Iterative refinement for linear systems and LAPACK.


K. Scheinberg.
Evolution of randomness in optimization methods for supervised machine learning.