Mixed precision and local error in ordinary differential equations

Arsène Marzorati, Mouhamad Al-Sayed Ali, Samuel Bernard, Jonathan Rouzaud-Cornabas

Mixed precision and local error for ODE solvers

CANUM - 2024

M. Al-Sayed Ali S. Bernard A. Marzorati J. Rouzaud-Cornabas

May 30, 2024

Context: Scaling ODE solvers

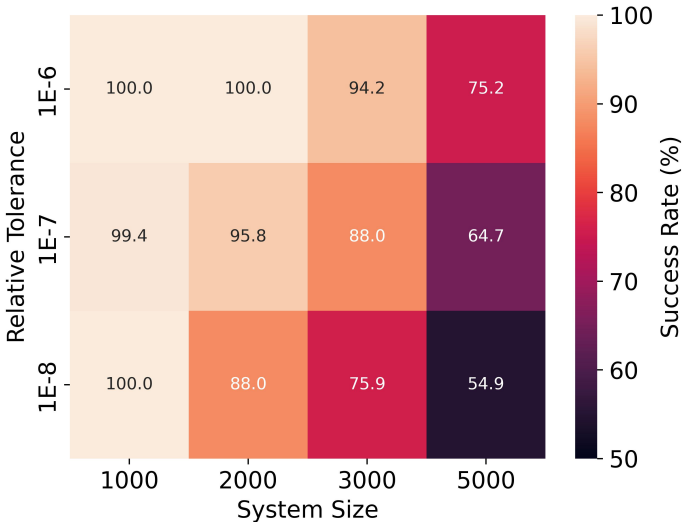
- **Computational challenge:**

Solving high dimensional and complex ODE systems for modelling biological systems.

- **General model:**

$$\dot{X}_i = \mathrm{F_i}(X) = \mathrm{H_i}(X_i) + \frac{1}{N}\sum_{j=1}^{N} \mathrm{G}_{ij}(X_i, X_j) \; i \in \{1, ..N\}, \; X_i \in \mathbb{R}^d.$$

Computational limits *(Benchmark: Kuramoto model)*



Mixed precision: a trade-off ?

- **Definition:**

Use several numerical arithmetic precisions inside one computational tool.

- **Benefits:**

Computational acceleration, less memory needed, better error control

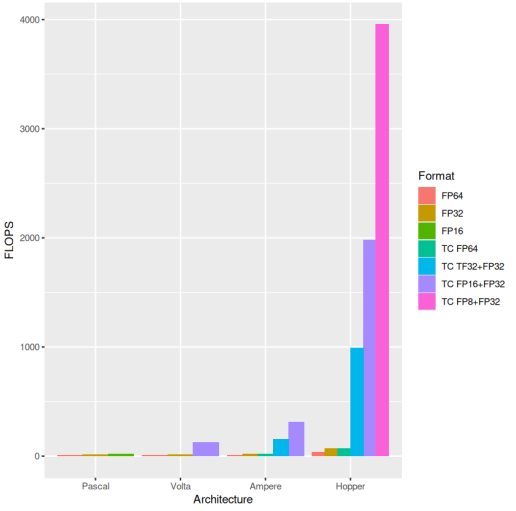
- **Tools:**

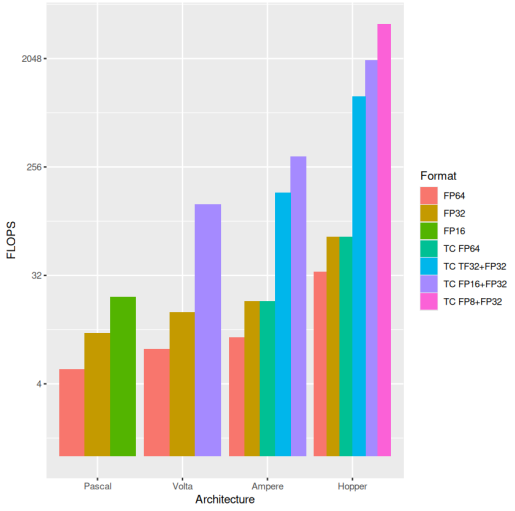
Linear Algebra, Machine Learning

- **Applications:**

Physics, Meteorology

Mixed-Precision and modern chips





Current metric for performance

Due to the lack of transparency in Matlab computations, we choose to force the arithmetic precision before each operation to ensure the corresponding accuracy. We use the following theoretical assumption. For an arithmetic operation:

$$T_D \sim 2T_S.$$
$$T_D \sim \alpha T_{MP},\ \alpha \in [1, 2].$$

We count the number of evaluations performed in one specific arithmetic precision.

Explicit solver: Runge-Kutta with p-stages

An ODE problem: $\dot{\mathrm{X}} = \mathrm{f}(t, \mathrm{X}),\ t \in [0, T].$

1) **Step time:** $h = \frac{T}{N}$ and setting $\mathrm{X}_i = \mathrm{X}(t_i) = \mathrm{X}(ih)$.

2) **Get $\mathbf{X}_{n+1}$:** (X_n known).

The stage(s): $k_i = \mathrm{f}(t_n + c_i h, \mathrm{X}_n + h \sum_{j=1}^{i-1} a_{i,j} k_j),\ \forall i \in \{1, ..., p\}$.

$$\mathrm{X}_{n+1} = \mathrm{X}_n + h \sum_{i=1}^{p} b_i k_i.$$

3) **Order of the scheme:** truncation error.

Solver ODE23: Adaptive scheme

- **ODE23:** Combination of RK3 with 3 stages and RK2 with 4 stages (first three stage are shared).
- **Adaptive scheme:** Error evaluation (RK2 vs RK3) and step validation (with the relative tolerance) modifying the step size ($h \to h_n$).
- **First Same As Last:** Last stage at step n is used as First stage at step $n+1$.
- **Bonus:** Already implemented in Matlab.

Mixed-precision at different levels

- For function evaluation, 3 possibilities can be chosen :

$$\dot{X}_i = \mathrm{H_i}(X_i) + \frac{1}{N}\sum_{j=1}^{N} \mathrm{G}_{ij}(X_i, X_j).$$

- For each stage (p-stages, 3 in our case) a different *cocktail* can be chosen.

	Precision		
RK-Stage	$\mathrm{H_i}$	$\sum_{j=1}^{N}$	G_j
K_2	S	D	S
K_3	S	D	S
K_4	D	D	S

Linear Coupled Oscillators (Benchmark 1)

Model equations

$$\begin{cases} \frac{dx_i}{dt} = y_i + \frac{1}{N}\sum_{j=1}^{N}(x_j - x_i) \\ \frac{dy_i}{dt} = -x_i \end{cases}, \ \forall i \in \{1, ..., N\}.$$

- **Why ?**
  - Appropriate structure
  - Analytic solution
  - Biological application: biological rhythm

Kuramoto (Benchmark 2)

Model equations

$$\frac{dx_i}{dt} = \omega_i + \frac{1}{N}\sum_{j=1}^{N} K \sin(x_j - x_i), \ \forall i \in \{1, ..., N\}.$$

- **Why ?**
  - Dense literature
  - Non-linear interaction term (sine)
  - Biological application: neuroscience

Benchmark 3: Circadian clock

Model equations

$$\begin{cases} \dot{x}_i = \left(\frac{k_0 \theta^h a}{\theta^h + y_i^h} x_i - k_1\right) x_i + \frac{1}{N} \sum_{j=1}^{N} \frac{k_0 \theta^h a}{\theta^h + y_i^h} K \arctan(x_j - x_i), \\ \dot{y}_i = k_2 (x_i - y_i), \\ \dot{u}_i = u_i \left(1 - \frac{u_i^2}{3}\right) - v_i + \mathrm{I}_0 \left(1 - \frac{k_3^2}{k_3^2 + x_i^2}\right), \\ \dot{v}_i = \epsilon \left(u_i + b - C v_i\right). \end{cases}$$

With $X_i = (x_i, y_i, v_i, u_i)^T \in \mathbb{R}^4$ and $X \in \mathbb{R}^{4 \times N}$

■ **Why ?**

- Real (simplified) model [1]
- Non-linear interaction term (arctan)

[1] El Cheikh R., Bernard S., & El Khatib N. (2017). A multiscale modelling approach for the regulation of the cell cycle by the circadian clock. J Theor Biol, 426, 117-125.

Metric for the tests

Error computation

Choose one solution as reference: $X_{ref}(t)$.
At final time, compute:

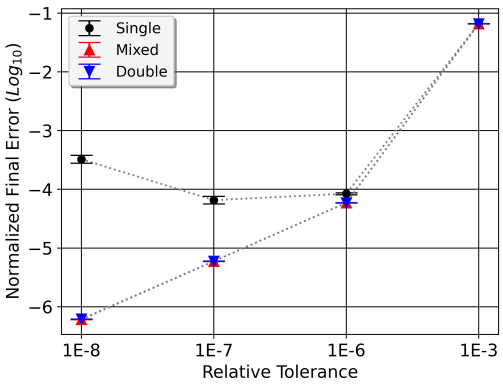
$$||X_{ref}(t_f) - X(t_f)||_\alpha = \frac{||X_{ref}(t_f) - X(t_f)||_2}{\sqrt{N}}.$$

In the next slides:

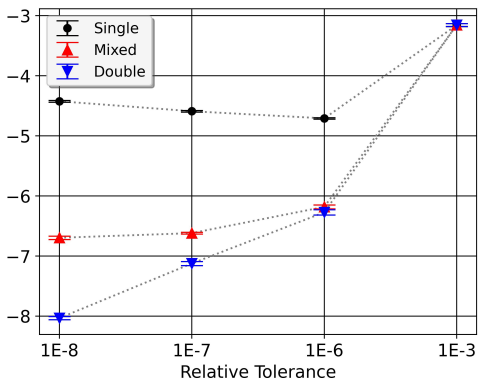
- The reference solution is computed with ODE45 of *Matlab* with a relative tolerance of 10^{-9}.
- The values are averaged over all the tests completed by all the solvers.

Why not switch completely to low precision ?

Linear coupled oscillators ($N = 1000$)

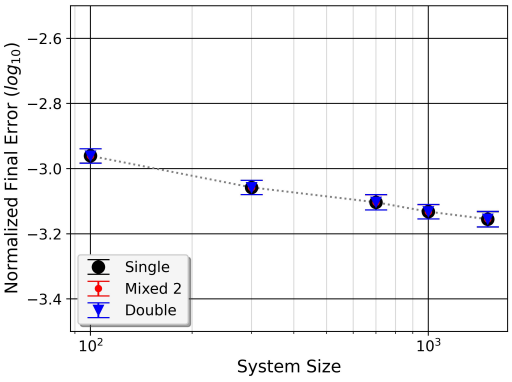


Circadian clock ($N = 1500$)

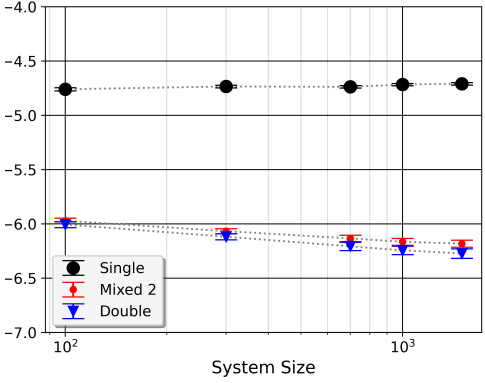


Accuracy: Circadian clock

Relative Tolerance: 10^{-3}

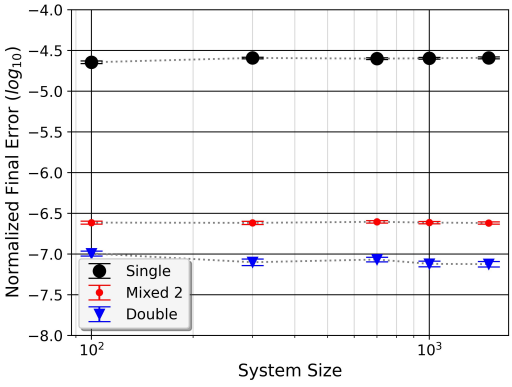


Relative Tolerance: 10^{-6}



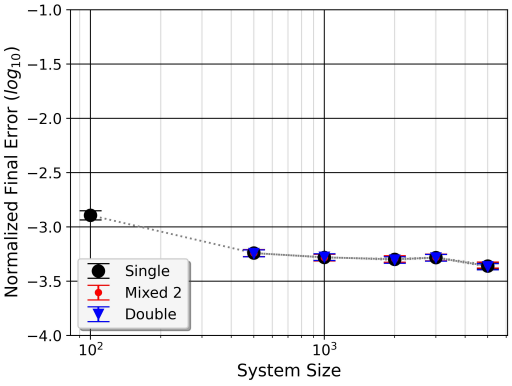
Accuracy: Circadian clock

Relative Tolerance: 10^{-7}

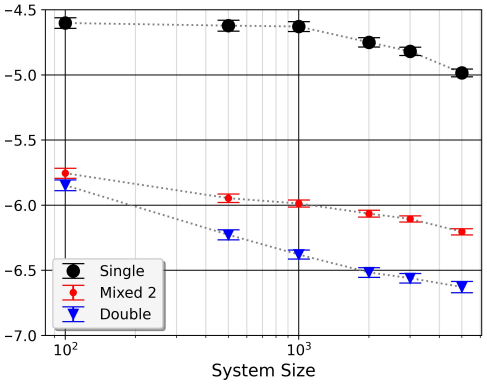


Accuracy: Kuramoto

Relative Tolerance: 10^{-3}

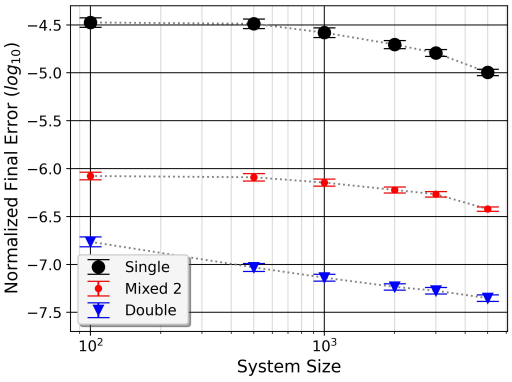


Relative Tolerance: 10^{-6}

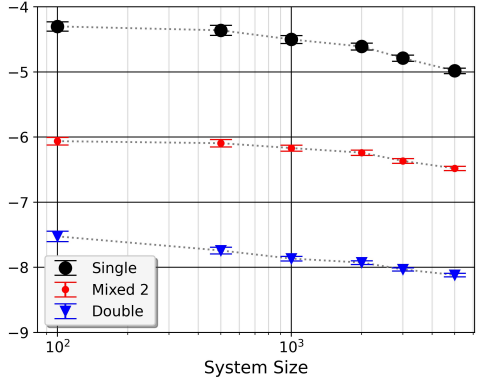


Accuracy: Kuramoto

Relative Tolerance: 10^{-7}



Relative Tolerance: 10^{-8}



Conclusion

- Mixed precision is necessary to benefit from the recent development of chips.
- A good trade-off for solving high-dimensional systems with sufficient accuracy with lower numerical precision.

Outlook

- Develop numerical methods that select the appropriate numerical precision from the many possibilities.
- Performance measurements on different architectures.

Thank you for your listening!

Bibliography

1. Ackmann J., Dueben P.D., Palmer T.N. and Smolarkiewicz P.K., "Mixed-Precision for Linear Solvers in Global Geophysical Flows",J Adv Model Earth Syst (2022).
2. Bogacki P. and Shampine L.F., "A 3 (2) pair of Runge-Kutta formulas." Applied Mathematics Letters 2.4 (1989): 321-325.
3. Croci M. and de Souza G.R., "Mixed-precision explicit stabilized Runge-Kutta methods for single- and multi-scale differential equations." J. Comput. Phys.(2022)
4. Higham N.J. and Mary T. "Mixed precision algorithms in numerical linear algebra." Acta Numerica 31 (2022): 347-414.
5. Haidar A., Tomov S., Dongarra J. and Higham N.J., "Harnessing GPU tensor cores for fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers." SC18: Int. Conf. High Perform. Comput. Netw. Storage Anal. IEEE (2018).
6. Choquette J. and Wish G., "Nvidia a100 gpu: Performance & innovation for gpu computing." IEEE Hot Chips 32 Symposium (HCS). IEEE Computer Society, (2020).

Bibliography

7. Burnett B., Gottlieb S., Grant Z.J., and Heryudono A., "Performance evaluation of mixed-precision Runge-Kutta methods. IEEE High Performance Extreme Computing Conference (HPEC) (2021).

8. Blanchard P., Higham N.J. and Mary T., "A class of fast and accurate summation algorithms, MIMS EPrint 2019.6." Manchester Institute for Mathematical Sciences, The University of Manchester, UK (2019).

9. Higham, N.J and Pranesh S., "Simulating low precision floating-point arithmetic", SIAM J Sci Comput (2019).

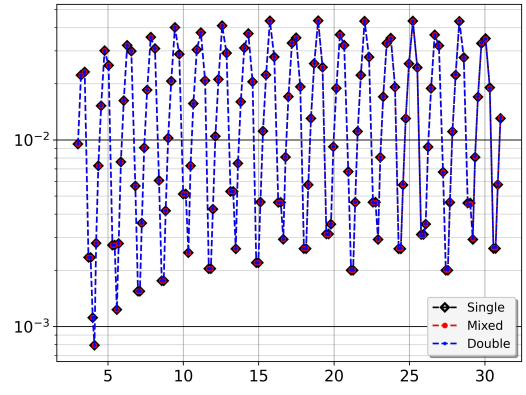
10. Hayford J., Goldman-Wetzler J., Wang E. and Lu L., "Speeding up and reducing memory usage for scientific machine learning via mixed precision", arXiv preprint (2024).

11. El Cheikh R., Bernard S., and El Khatib N., "A multiscale modelling approach for the regulation of the cell cycle by the circadian clock". J Theor Biol (2017), 426, 117-125.

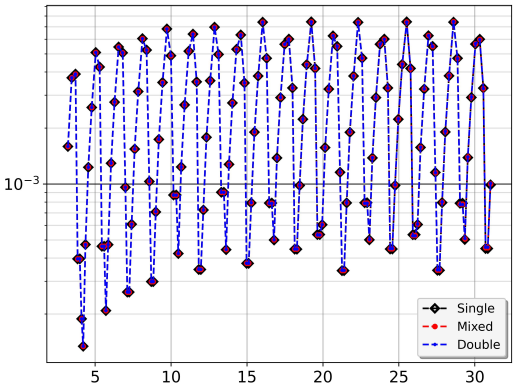
12. Palmer T.N., "More reliable forecasts with less precise computations: a fast-track route to cloud-resolved weather and climate simulators?" Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 372.2018 (2014).

Linear oscillators (Relative Tolerance: 10^{-3}, N=1000)

Local error:

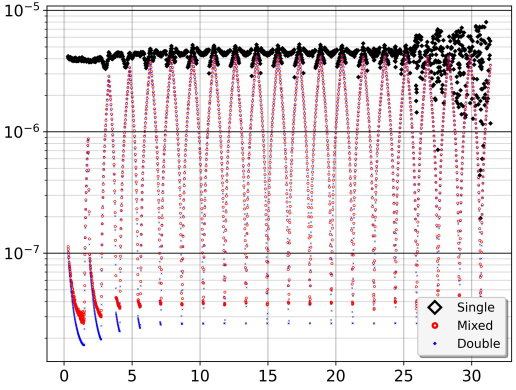


Step size (power 4):

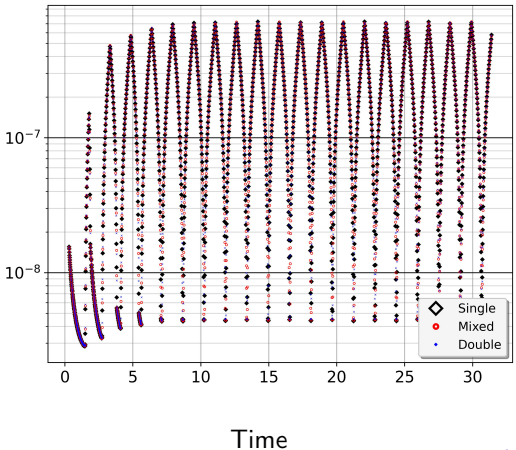


Linear oscillators (Relative Tolerance: 10^{-6}, N=1000)

Local error:



Step size (power 4):



Relative tolerance

Tolerance and step validation

At each step $X_1^{(n)}$ and $X_2^{(n)}$ are computed, $X_2^{(n-1)}$ is the solution at the previous step. At final time (t_f), compute:

$$err := \left|\left|(X_1^{(n)} - X_2^{(n)})./ \max\left(|X_2^{(n)}|, |X_2^{(n-1)}|, \frac{Ab}{Rel}\right)\right|\right|_\infty.$$

'./' *division term by term*.

Parameters for Linear coupled oscillators

Parameters	Value(s)
Tests number	1000
N (number of oscillators)	$\{100, 500, 1000, 2000, 5000\}$
$Tol - Ab$	$\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-3}\}$
$Tol - Rel$	$\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-3}\}$
T_f (Final time)	10π
Initial conditions	$[0, 2\pi]$

Parameters for Kuramoto

Parameter	Value(s)
Number of tests	5000
N (Number of oscillators)	$\{100, 500, 1000, 2000, 3000, 5000\}$
$Tol - Ab$	$\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-3}\}$
$Tol - Rel$	$\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-3}\}$
σ	$[0, 1]$
ω_i	$[-\sigma, \sigma]$
K(coupling coefficient)	$[0, 3]$
T_f (Final time)	$\frac{4\pi}{\text{med}(\lvert W \rvert)K + 0.001}$
Initial conditions	$[0, 2\pi]$

Parameters for circadian clock

Parameter	Value(s)
Number of tests	1000
N (Number of oscillators)	$\{100, 300, 700, 1000, 1500\}$
$Tol - Ab$	$\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-3}\}$
$Tol - Rel$	$\{10^{-8}, 10^{-7}, 10^{-6}, 10^{-3}\}$
K(coupling coefficient)	$\{0.001, 0.1, 1, 10\}$
I_0	$\{0.228249, 1.5, 10\}$
T_f (Final time)	$3 \times T_{cycle} = 72$

Accuracy: Circadian clock (Relative Tolerance: 10^{-8})