Mixed-Precision Additive Runge-Kutta Methods

César Herrera[2]

Joint work with John Driscoll[1] Sigal Gottlieb[1] Zachary Grant[1]

Tej Sai Kakumanu[1] Monica Stephens[3]

[1]UMass Dartmouth [2]Purdue University [3]Spelman College

North American High Order Methods Conference
June 2026

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1929284, while the author was in residence at the Institute for Computational and Experimental Research in Mathematics (ICERM) in Providence, RI, during the program **"Empowering a Diverse Computational Mathematics Research Community."**
Additional support was provided by the Air Force Office of Scientific Research (FA9550-23-1-0037), and the U.S. Department of Energy (DE-SC0023164).

Mixed Precision Computations

How do we represent numbers on a computer?

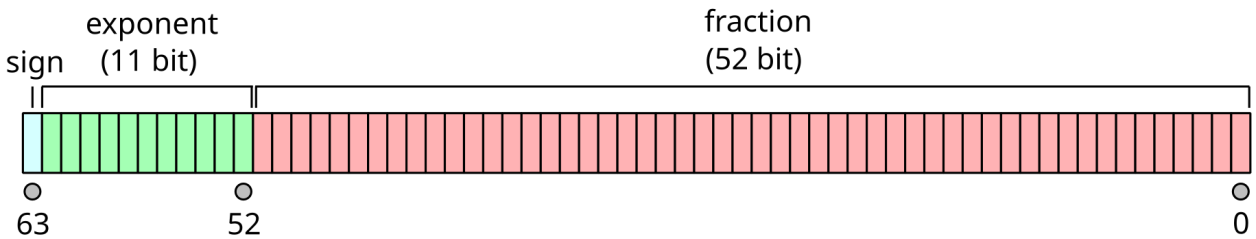


Figure: IEEE 754 Double Floating Point Format.

Source: Wikimedia Commons, by SharkD, licensed under CC BY-SA 3.0.

$$\text{Value} = (-1)^{\text{sign}} \times (1 + \text{fraction}) \times 2^{\text{exponent}-1023}$$

Mixed Precision Computations

How do we represent numbers on a computer?

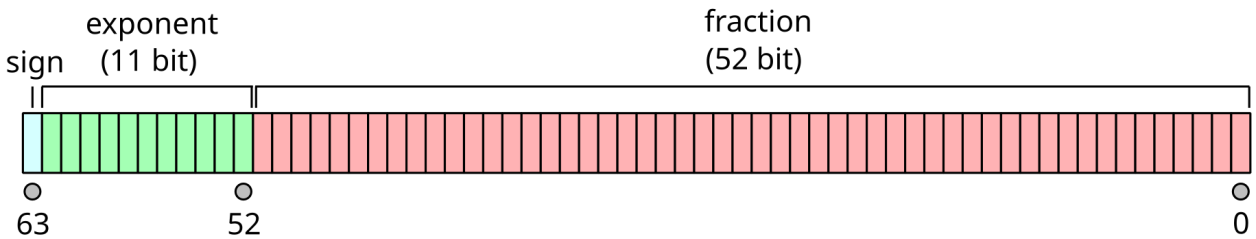


Figure: IEEE 754 Double Floating Point Format.

Source: Wikimedia Commons, by SharkD, licensed under CC BY-SA 3.0.

$$\text{Value} = (-1)^{\text{sign}} \times (1 + \text{fraction}) \times 2^{\text{exponent}-1023}$$

Type	Size (bits)	Unit Roundoff u
Half	16	$2^{-11} \approx 4.9 \times 10^{-4}$
Single	32	$2^{-24} \approx 6.0 \times 10^{-8}$
Double	64	$2^{-53} \approx 1.1 \times 10^{-16}$
Quad	128	$2^{-113} \approx 9.6 \times 10^{-35}$

Table: IEEE floating-point representations with size and unit roundoff.

Trade-off

- Computations in **low precision** yield:
  - **Speedups:** Half (single) precision can yield $4\times$ ($2\times$) speedups over double precision.
  - **Reduced accuracy:** Increased round-off errors.
- **Mixed Precision:** Use low precision only for expensive steps to balance accuracy and efficiency.

Mixed Precision Computations

- Modern hardware supports multiple floating-point precisions.

Figure: Performance of mixed precision training on NVIDIA 8xV100 vs. FP32 training on 8xV100 GPU. Source: M. Huang, C. Tekur, M. Carilli. *Introducing native PyTorch automatic mixed precision for faster training on NVIDIA GPUs*, PyTorch Blog (2020).

Mixed Precision in Numerical PDEs

Can we design mixed precision methods for numerical PDEs?

Mixed Precision in Numerical PDEs

Can we design mixed precision methods for numerical PDEs?

- Z. Grant **Perturbed Runge–Kutta Methods for Mixed Precision Applications** *JSC, 2022.*

Table of Contents

Table of Contents

Mixed Precision for Numerical PDEs

Consider a system of ODEs of the form

$$u_t = F(u)$$

with initial condition $u(0) = u^0$.

- **Examples:**
  - $u_t = (u^3)_{xx}$
  - $u_t = -\left(\frac{u^2}{2}\right)_x$
- **Stiff problems!** Need to use **implicit** methods.

Runge-Kutta Methods: IMR

Implicit Midpoint Rule (IMR)

$$u^{n+1} = u^n + \Delta t F\left(\frac{u^n + u^{n+1}}{2}\right)$$

Butcher Form

$$y^{(1)} = u^n + \frac{1}{2}\Delta t F(y^{(1)}) \quad \textbf{Implicit}$$
$$u^{n+1} = u^n + \Delta t F(y^{(1)}) \quad \textbf{Explicit}$$

- Global error $= \mathcal{O}(\Delta t^2)$
- Implicit stage: **computationally expensive**.

Can we make the implicit solve cheaper while mantaining the accuracy?

Simple Idea

In practice, instead of solving

$$y^{(1)} = u^n + \frac{1}{2}\Delta t F(y^{(1)})$$

we solve a cheaper equation

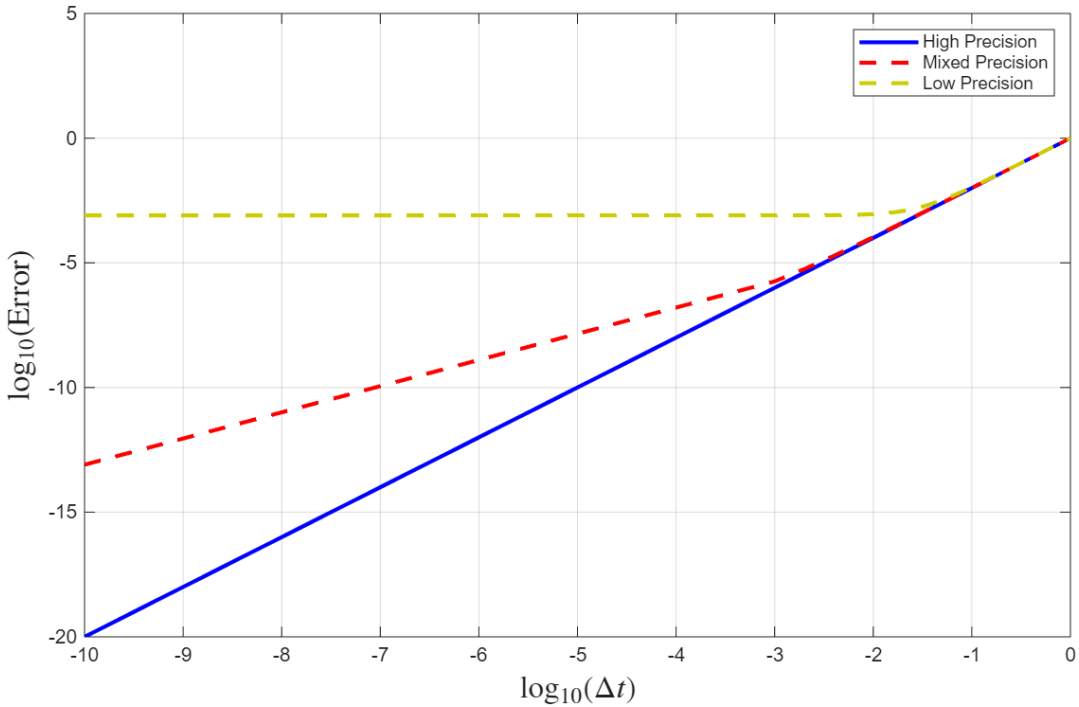
$$y^{(1)} = u^n + \frac{1}{2}\Delta t F^{\epsilon}(y^{(1)})$$

- $F^{\epsilon}(y)$ can be obtained via:
  - Linearization.
  - Implicit solve with low tolerance.
  - **Low-precision** implicit solve.

Comparison

High Precision IMR

- Both stages in high precision
- Global error: $\mathcal{O}(\Delta t^2)$



[1] Zachary Grant. "Perturbed Runge–Kutta Methods for Mixed Precision Applications". In: *Journal of Scientific Computing* 92 (2022)

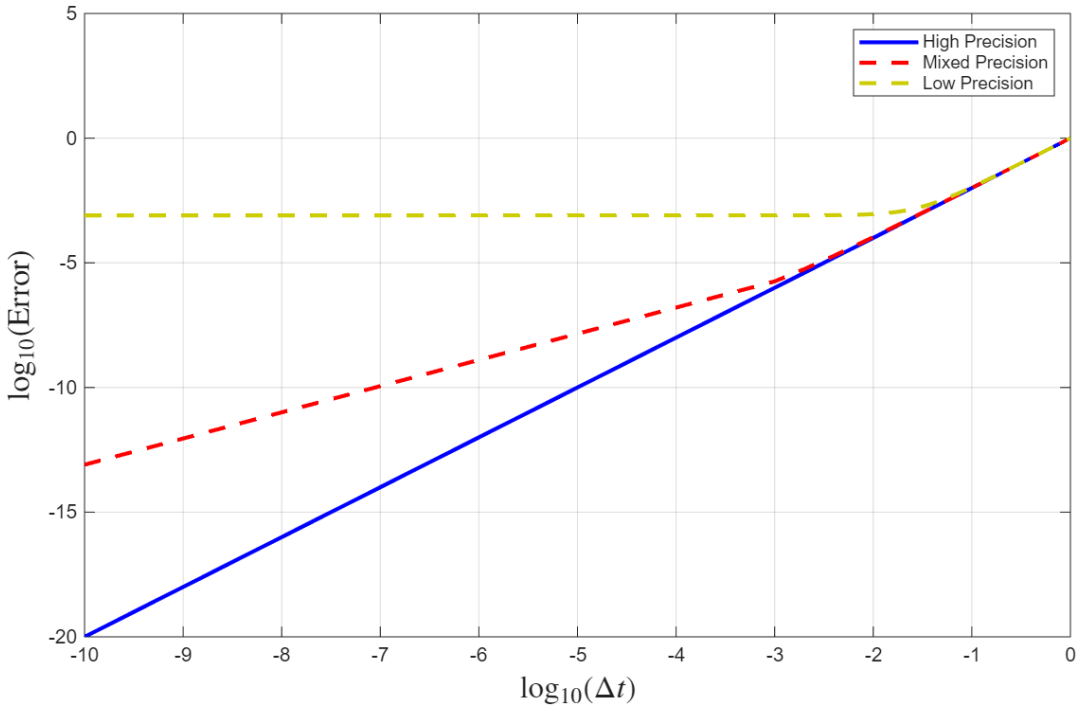
Comparison

High Precision IMR

- Both stages in high precision
- Global error: $\mathcal{O}(\Delta t^2)$

Low Precision IMR

- Both stages in low precision
- Global error[1]: $\mathcal{O}(\Delta t^2) + \mathcal{O}(\epsilon)$



[1] Zachary Grant. "Perturbed Runge–Kutta Methods for Mixed Precision Applications". In: *Journal of Scientific Computing* 92 (2022)

Comparison

High Precision IMR

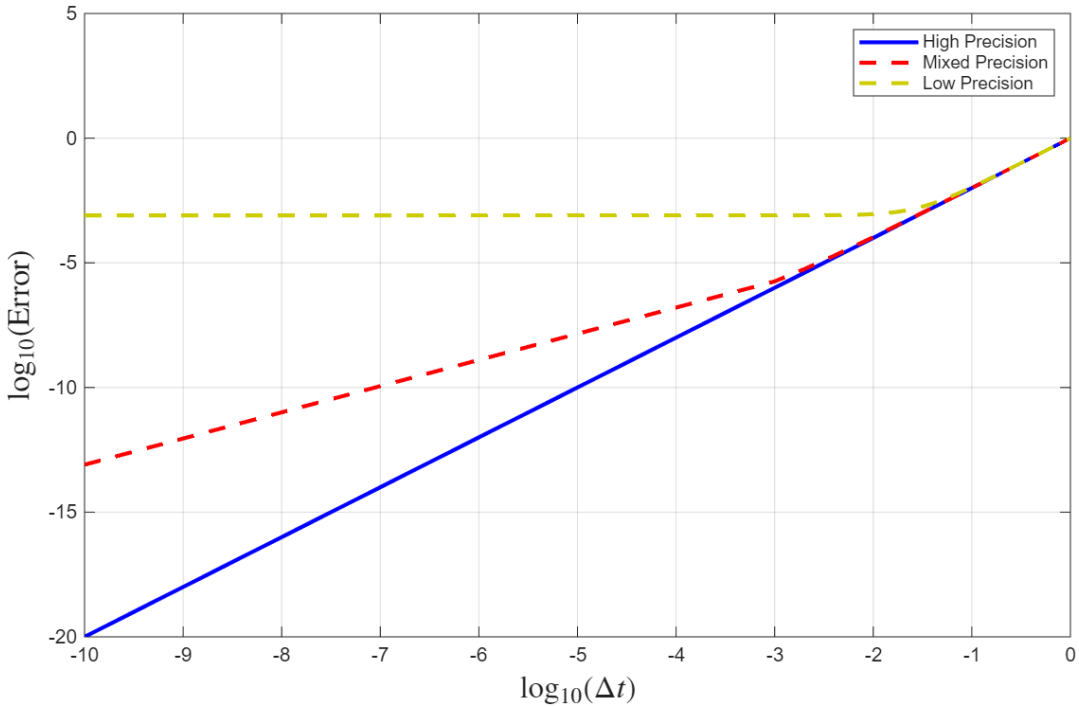
- Both stages in high precision
- Global error: $\mathcal{O}(\Delta t^2)$

Low Precision IMR

- Both stages in low precision
- Global error[1]: $\mathcal{O}(\Delta t^2) + \mathcal{O}(\epsilon)$

Mixed Precision IMR

- Explicit stage in high precision
- Implicit stage in low precision
- Global error[1]: $\mathcal{O}(\Delta t^2) + \mathcal{O}(\epsilon\,\Delta t)$



[1] Zachary Grant. "Perturbed Runge–Kutta Methods for Mixed Precision Applications". In: *Journal of Scientific Computing* 92 (2022)

Table of Contents

Test Problem

Van der Pol system

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_2\,(1 - y_1^2) - y_1 \end{aligned}$$

with initial conditions $y_1(0) = 2$ and $y_2(0) = 0$. We stepped this forward to a final time $T_f = 1.0$.

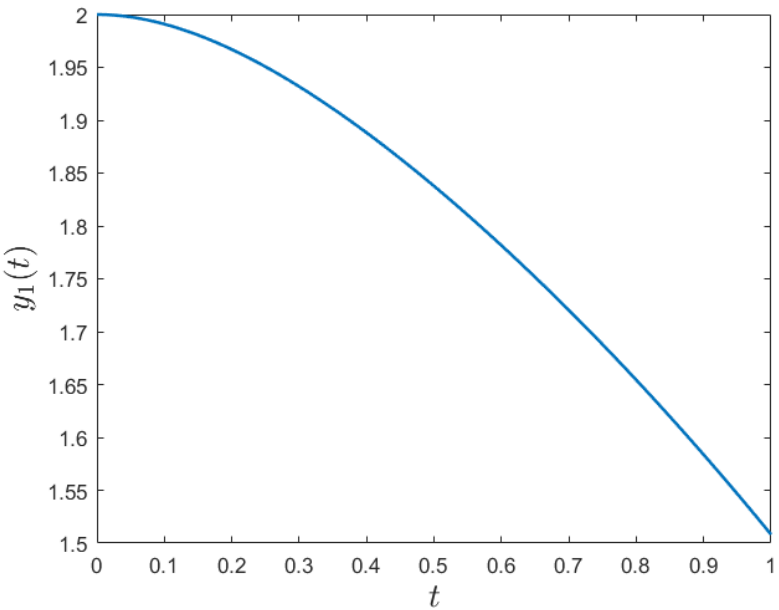
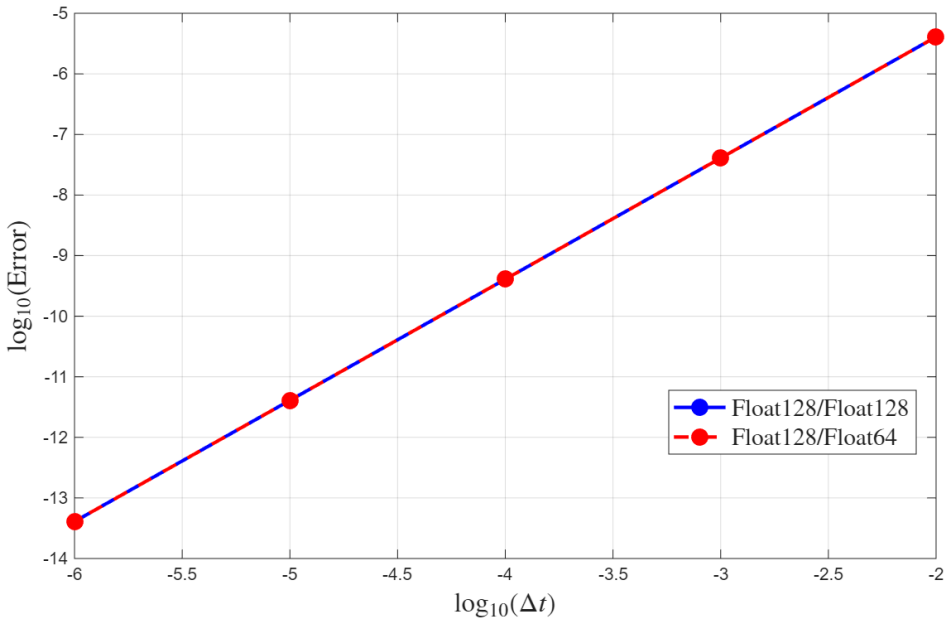


Figure: Graph of solution $y_1(t)$

Mixed Precision vs High Precision

Accuracy and runtime speedup for different Δt: [2]

Δt	**Speedup**
1E-02	**7.35**
1E-03	**11.2**
1E-04	**3.12**
1E-05	**6.05**
1E-06	**3.75**



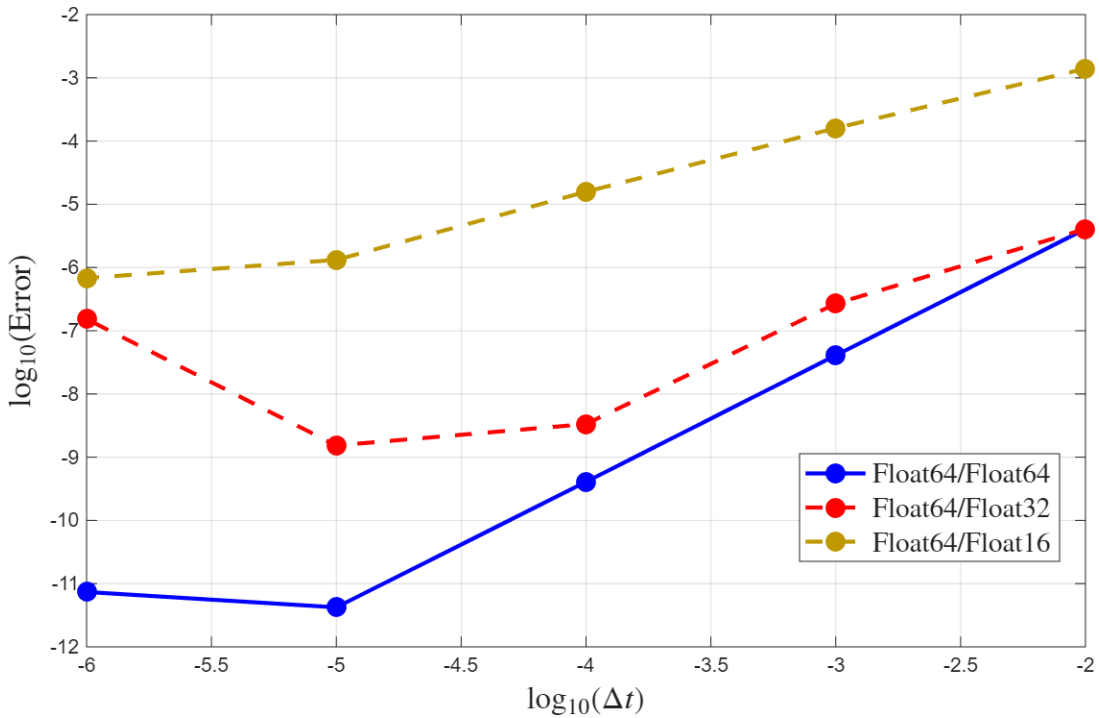
Same accuracy across all Δt, with runtime savings of **3–11**$\times$.

[2] Sigal Gottlieb. *Mixed model and mixed precision Runge–Kutta methods*. Innovative and Efficient Strategies for Stiff Differential Equations Workshop. ICERM, Brown University. July 2025

Mixed Precision vs High Precision

Runtime speedup (blue) for each configuration[2]:

Δt	64/32	64/16
1E-02	**1.7** $\times$	**4.5** $\times$
1E-03	**8.0** $\times$	**20.5** $\times$
1E-06	**3.2** $\times$	**3.8** $\times$



- **Region of accuracy:** $\epsilon < \Delta t$
- **Question:** Can we do better for $\Delta t < \epsilon$?

[2] Sigal Gottlieb. *Mixed model and mixed precision Runge–Kutta methods*. Innovative and Efficient Strategies for Stiff Differential Equations Workshop. ICERM, Brown University. July 2025

Table of Contents

Explicit Corrections

The explicit stage of the mixed precision method works as a correction.

Question: Can we add another explicit correction (e.g., **fixed point iterations**) to further improve the accuracy?

Explicit Corrections

The explicit stage of the mixed precision method works as a correction.

Question: Can we add another explicit correction (e.g., **fixed point iterations**) to further improve the accuracy?

Mixed Precision Implicit Midpoint Rule with Correction

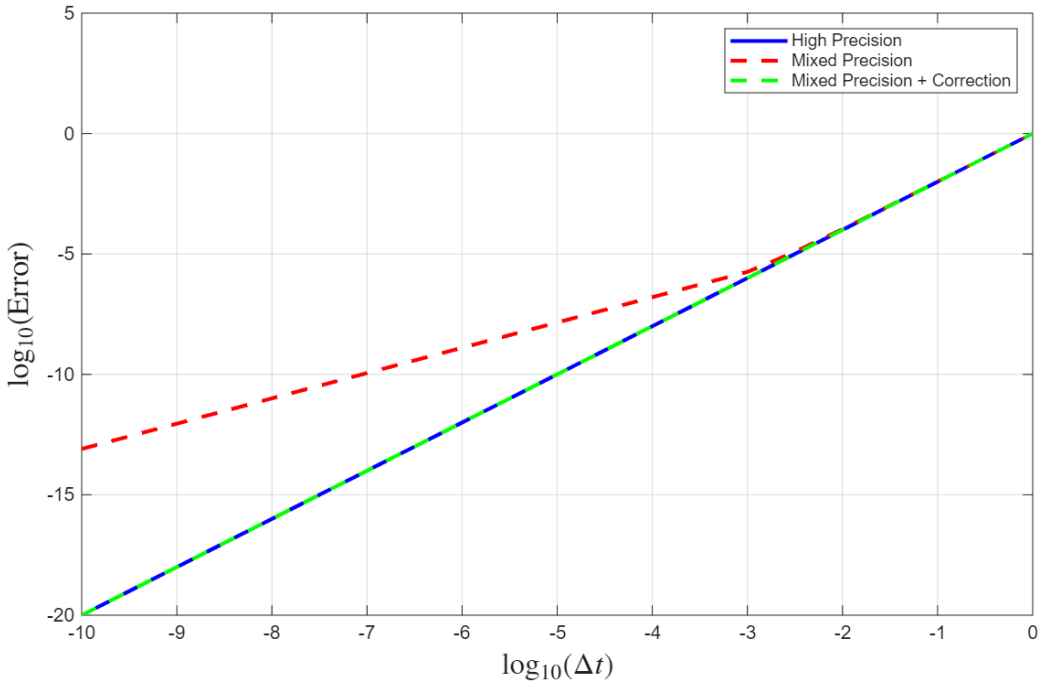
$$y_0^{(1)} = u^n + \frac{1}{2}\Delta t F^{\epsilon}(y_0^{(1)}) \qquad \textbf{Implicit, low precision}$$

$$\mathbf{y_1^{(1)} = u^n + \frac{1}{2}\Delta t F(y_0^{(1)})} \qquad \textbf{Explicit, high precision}$$

$$u^{n+1} = u^n + \Delta t F(y_1^{(1)}) \qquad \textbf{Explicit, high precision}$$

Explicit Corrections

- Global error[1] (for small Δt) $= \mathcal{O}(\Delta t^2) + \mathcal{O}(\epsilon \Delta t^2)$

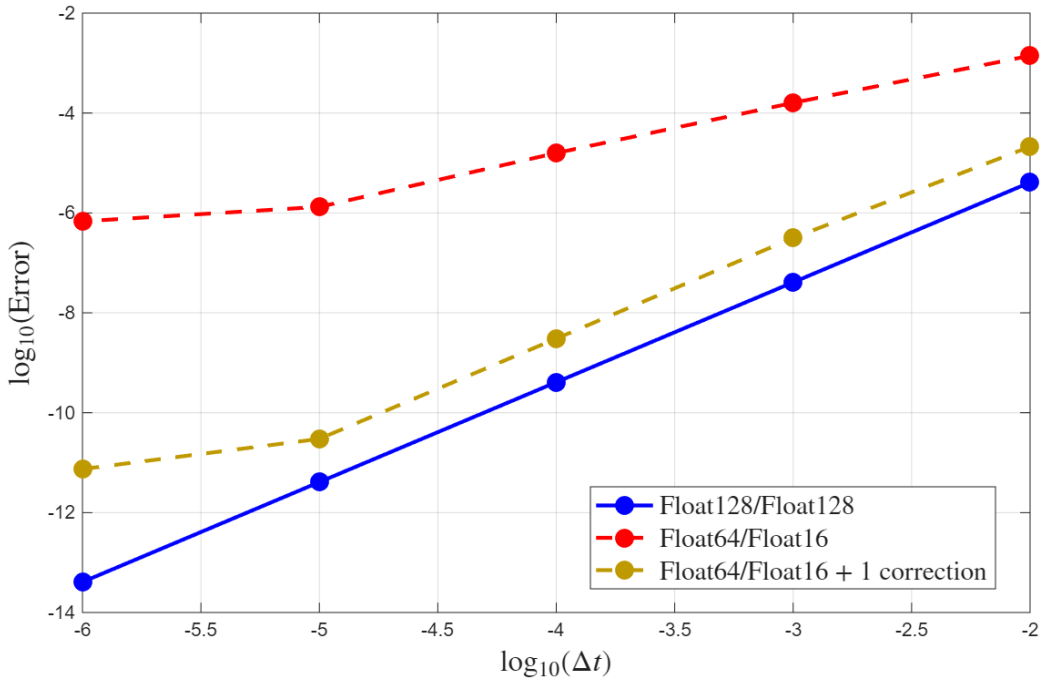


[1] Zachary Grant. "Perturbed Runge–Kutta Methods for Mixed Precision Applications". In: *Journal of Scientific Computing* 92 (2022)

No Corrections vs One Correction

Comparison[2]: no corrections vs one correction for 64/16.

Δt	No corr.	One corr.
1E-02	**45.5** $\times$	**38.5** $\times$
1E-04	**41.4** $\times$	**17.1** $\times$
1E-06	**25.3** $\times$	**18.9** $\times$



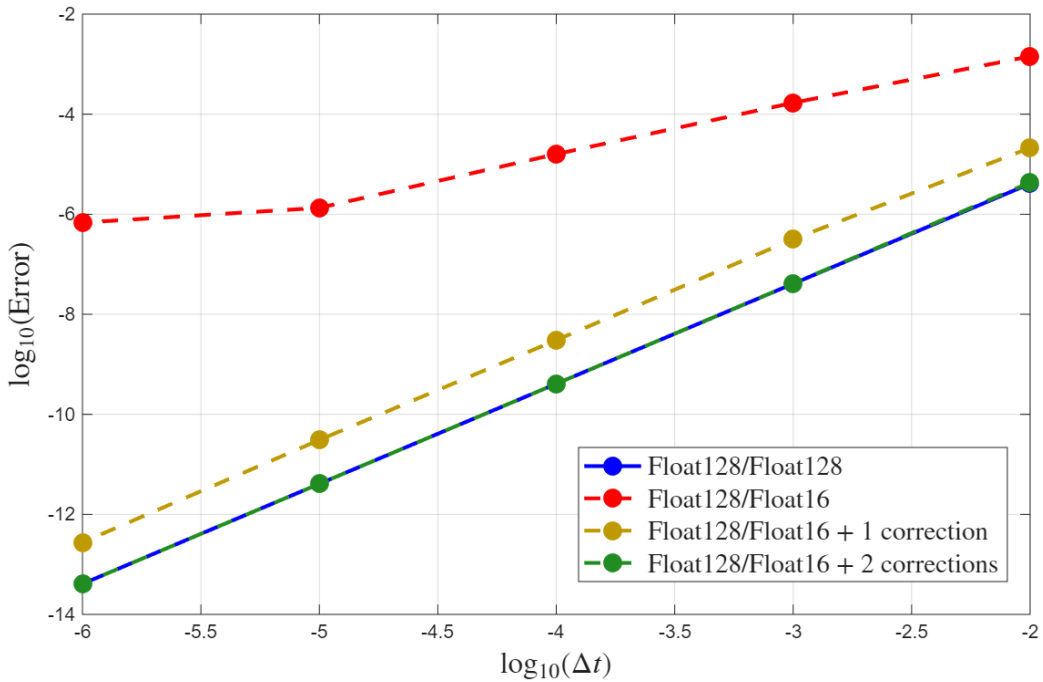
We recover a factor of $\mathcal{O}(\Delta t)$ in accuracy for each correction applied.

[2] Sigal Gottlieb. *Mixed model and mixed precision Runge–Kutta methods*. Innovative and Efficient Strategies for Stiff Differential Equations Workshop. ICERM, Brown University. July 2025

Two Corrections

Quad/half (128/16) precision with $0, 1, 2$ corrections.[2]

Δt	One corr.	Two corr.
1E-02	**25.0** $\times$	**16.7** $\times$
1E-04	**4.0** $\times$	**4.3** $\times$
1E-06	**8.4** $\times$	**6.3** $\times$



[2] Sigal Gottlieb. *Mixed model and mixed precision Runge–Kutta methods*. Innovative and Efficient Strategies for Stiff Differential Equations Workshop. ICERM, Brown University. July 2025

Bigger Problem

Burgers' Equation

$$\begin{cases} u_t + \left(\frac{1}{2}u^2\right)_x = 0, \\ u(x,0) = \sin(x), \end{cases}$$

with periodic boundary conditions.

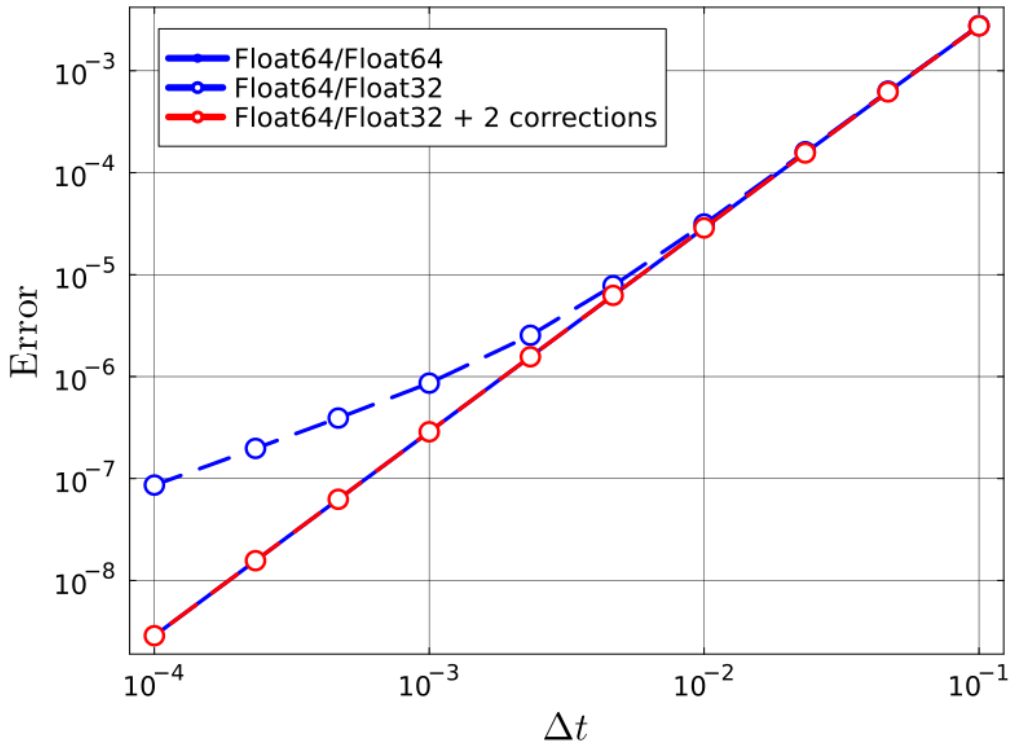
Numerical experiment:

- Semi-discretization with N_x spatial points
- Nonlinear term approximated using Fourier spectral methods
- Final time: $T_f = 0.7$ (before the shock)

IMR with $N_x = 200$

Comparison: no corrections vs two corrections for 64/32.

Δt	No corr.	Two corr.
1E-02	**4.1** $\times$	**4.3** $\times$
1E-04	**4.5** $\times$	**4.5** $\times$
1E-05	**2.5** $\times$	**2.6** $\times$



Third Order DIRK Method

Comparison: no corrections vs two corrections for 64/32.

Δt	No corr.	Two corr.
1E-02	**4.7** $\times$	**4.1** $\times$
1E-04	**4.8** $\times$	**4.5** $\times$
1E-05	**3.1** $\times$	**2.5** $\times$

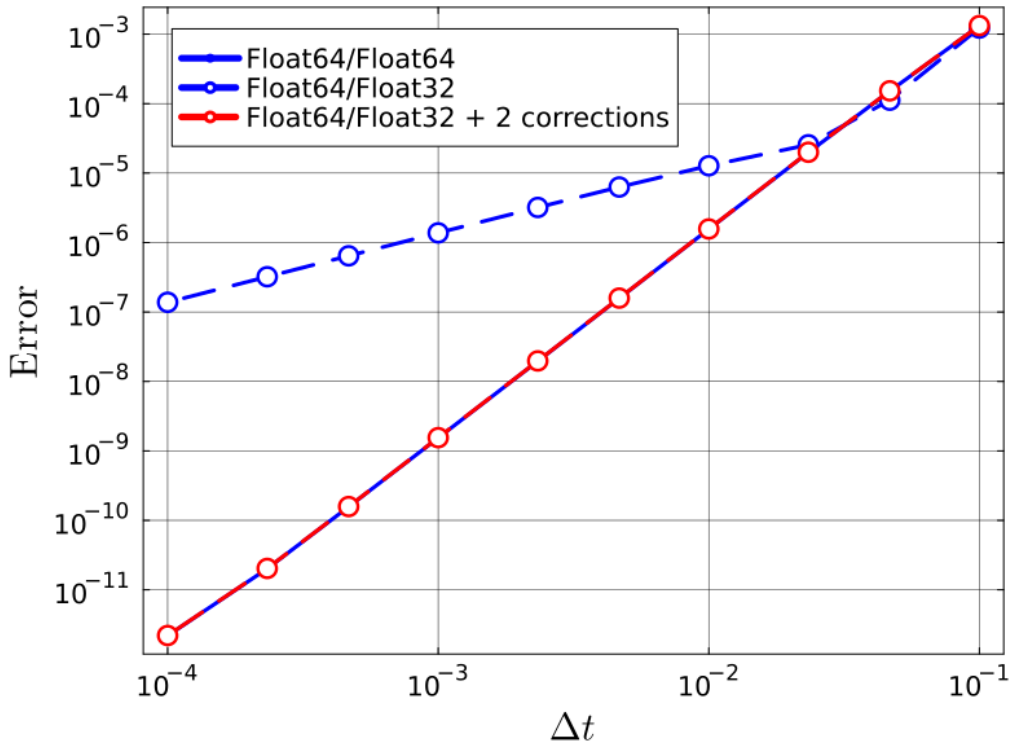
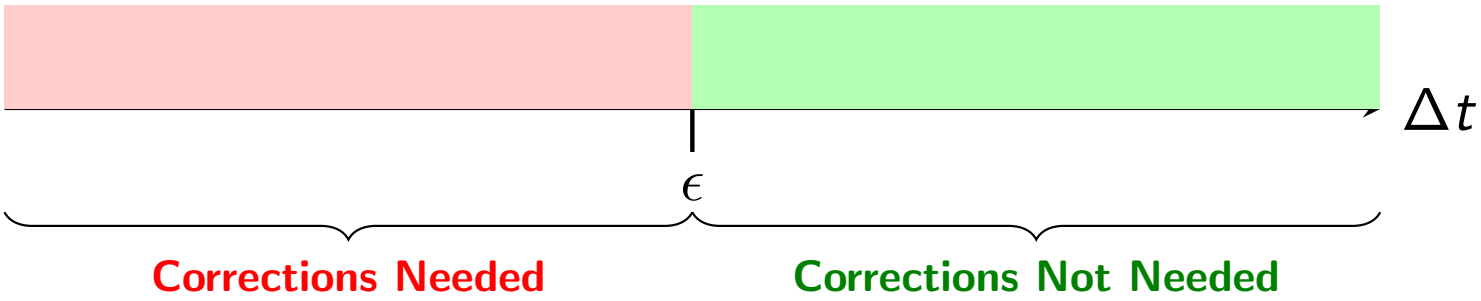


Table of Contents

Stability of Mixed Precision



Linear Stability

For the linear problem:

$$u_t = \lambda u, \quad \lambda \in \mathbb{C}, \quad \mathrm{Re}(\lambda) < 0.$$

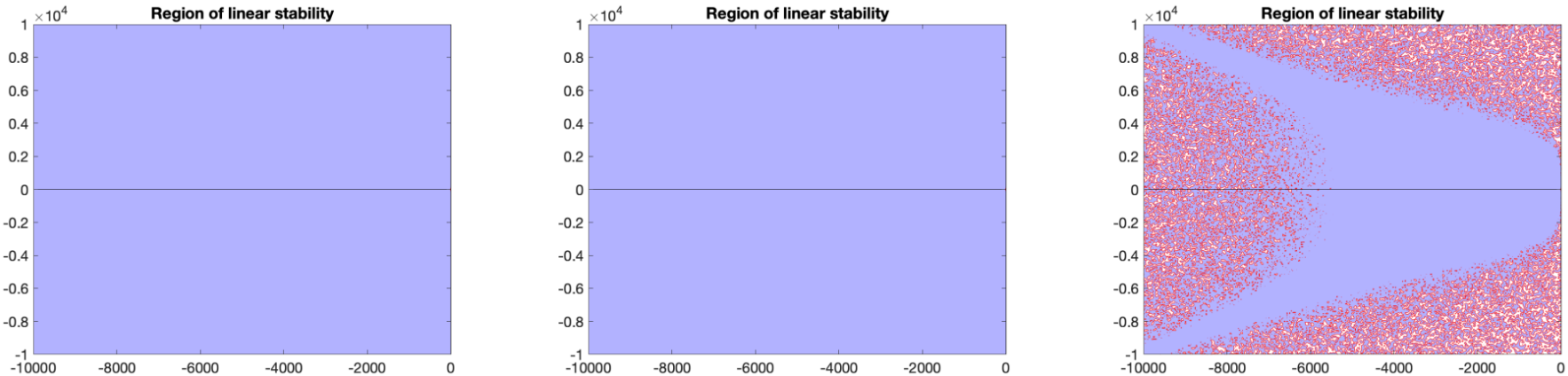
Let $z = \lambda \Delta t$. **For which values of z is the MP-IMR stable?**

Linear Stability

For the linear problem:

$$u_t = \lambda u, \quad \lambda \in \mathbb{C}, \quad \mathrm{Re}(\lambda) < 0.$$

Let $z = \lambda \Delta t$. **For which values of z is the MP-IMR stable?**

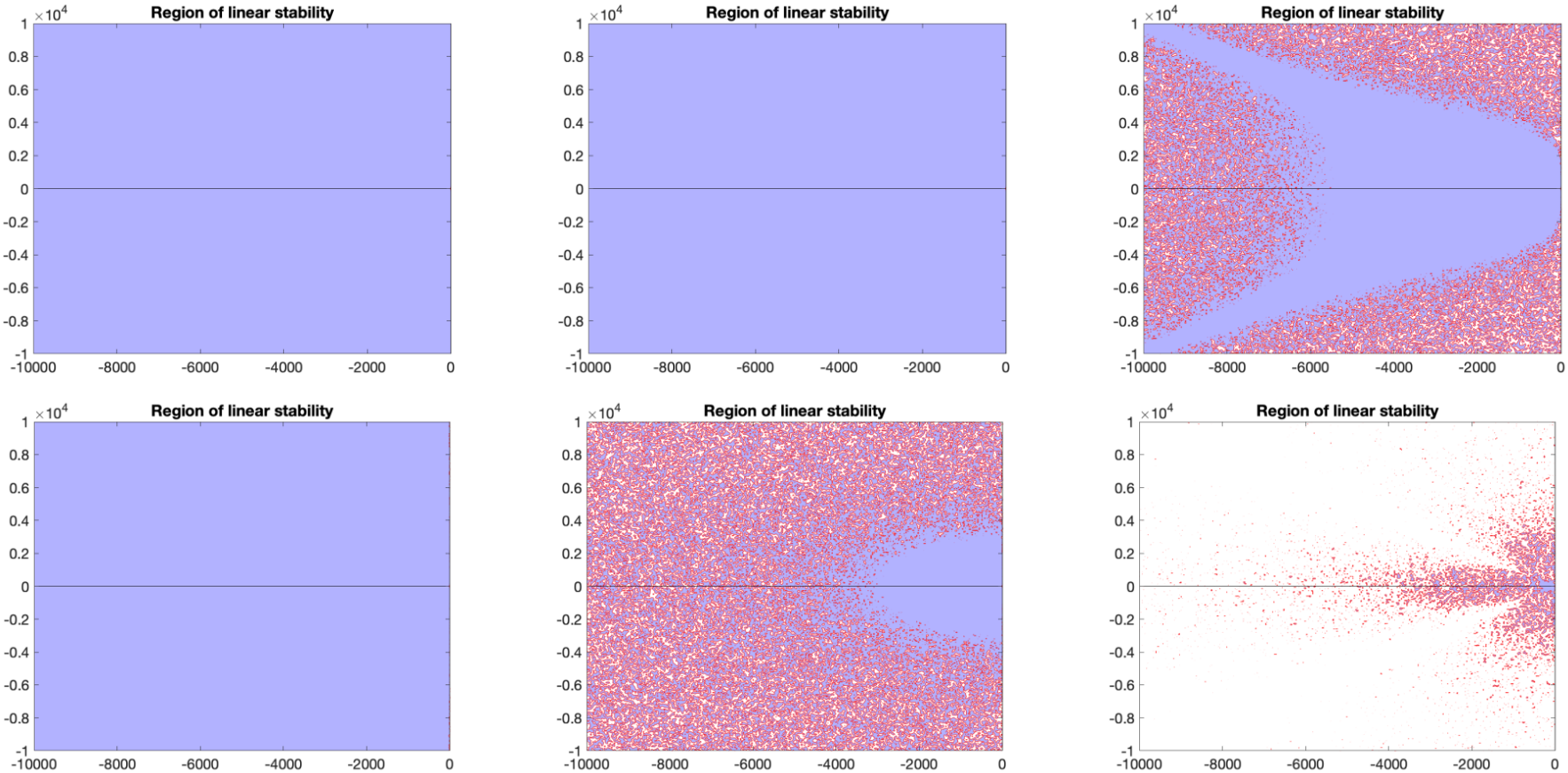


[3]Regions of linear stability for the IMR with no corrections (left), one correction (middle), two corrections (right) for $\tilde{\epsilon} = \frac{\epsilon}{\lambda} = 10^{-10}$.

[3] Ben Burnett, Sigal Gottlieb, and Zachary J. Grant. "Stability Analysis and Performance Evaluation of Additive Mixed-Precision Runge-Kutta Methods". In: *Communications on Applied Mathematics and Computation* (2023)

Linear Stability: Mixed Precision

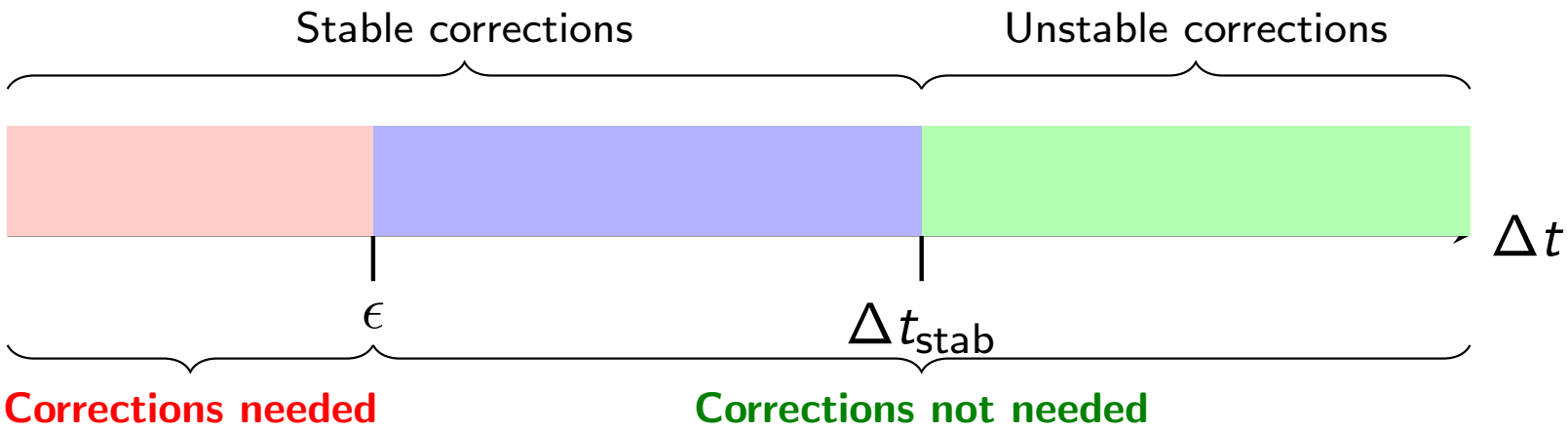
Regions of linear stability for the IMR with no corrections (left), one correction (middle), two corrections (right) for $\tilde{\epsilon} = \frac{\epsilon}{\lambda} = 10^{-10}$ (top) and $\tilde{\epsilon} = \frac{\epsilon}{\lambda} = 10^{-6}$ (bottom).[3]



[3] Ben Burnett, Sigal Gottlieb, and Zachary J. Grant. "Stability Analysis and Performance Evaluation of Additive Mixed-Precision Runge-Kutta Methods". In: *Communications on Applied Mathematics and Computation* (2023)

Linear Stability: Mixed Precision

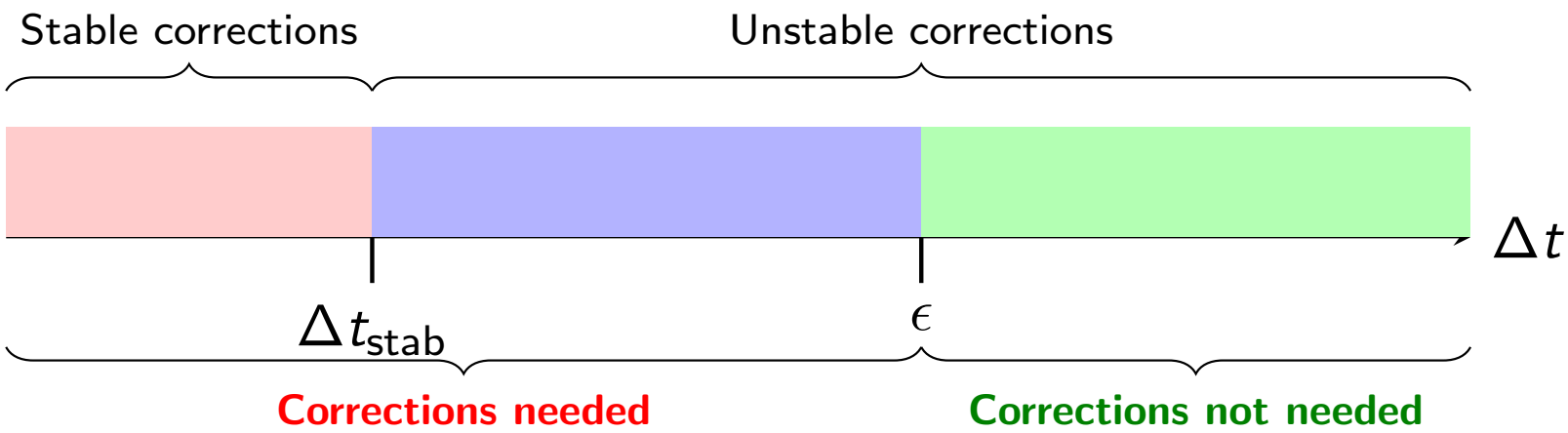
Case 1: $\epsilon < \Delta t_{\text{stab}}$



- Mixed precision for $\mathbf{\Delta t} > \mathbf{\Delta t}_{\text{stab}}$
- Explicit methods for $\mathbf{\Delta t} < \mathbf{\Delta t}_{\text{stab}}$

Linear Stability: Mixed Precision

Case 2: $\boldsymbol{\Delta t_{\text{stab}} < \epsilon}$



- Mixed precision for $\boldsymbol{\Delta t > \epsilon}$
- Explicit methods for $\boldsymbol{\Delta t < \Delta t_{\text{stab}}}$
- **What to do for** $\boldsymbol{\Delta t_{\text{stab}} < \Delta t < \epsilon}$?

Stable Corrections

The explicit corrections have the form

$$y_{k+1} = y_k - R(y_k),$$

where $R(y)$ is the **residual** of the implicit stage.

Stable Corrections

The explicit corrections have the form

$$y_{k+1} = y_k - R(y_k),$$

where $R(y)$ is the **residual** of the implicit stage.

Stabilized corrections:

$$y_{k+1} = y_k - \Phi_k R(y_k),$$

for some matrix Φ_k.

Stable Corrections

The explicit corrections have the form

$$y_{k+1} = y_k - R(y_k),$$

where $R(y)$ is the **residual** of the implicit stage.

Stabilized corrections:

$$y_{k+1} = y_k - \Phi_k R(y_k),$$

for some matrix Φ_k.

Possible choice for Φ_k:

$$\Phi_k = \Phi = (R'(u_0))^{-1}.$$

Precomputed in high precision.

Comparing Corrections

Porous Medium Equation

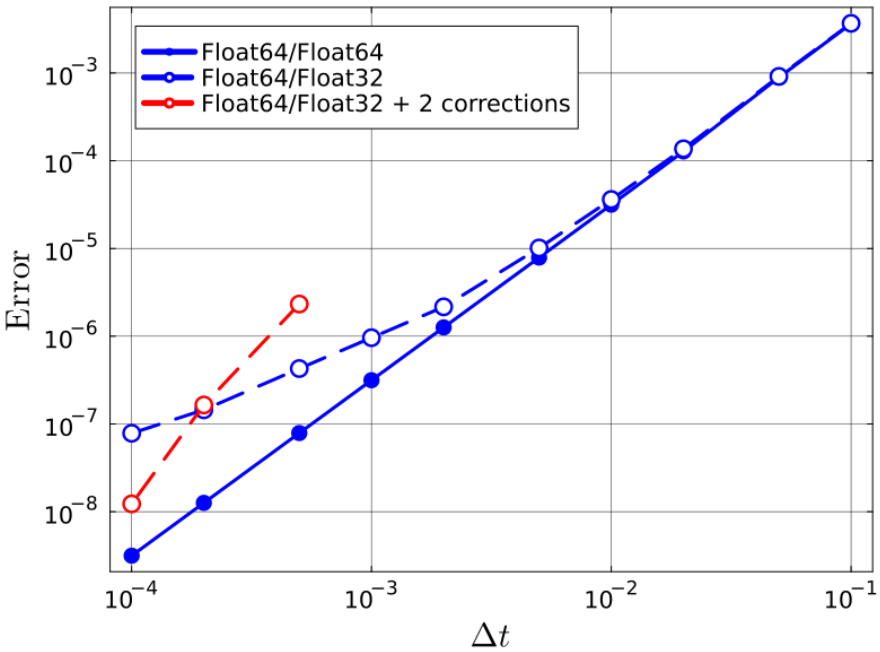
$$\begin{cases} u_t = \left(u^3\right)_{xx}, \\ u(x,0) = \dfrac{1}{2}\cos(x) + \dfrac{1}{2}, \end{cases}$$

with periodic boundary conditions.

Numerical experiment:

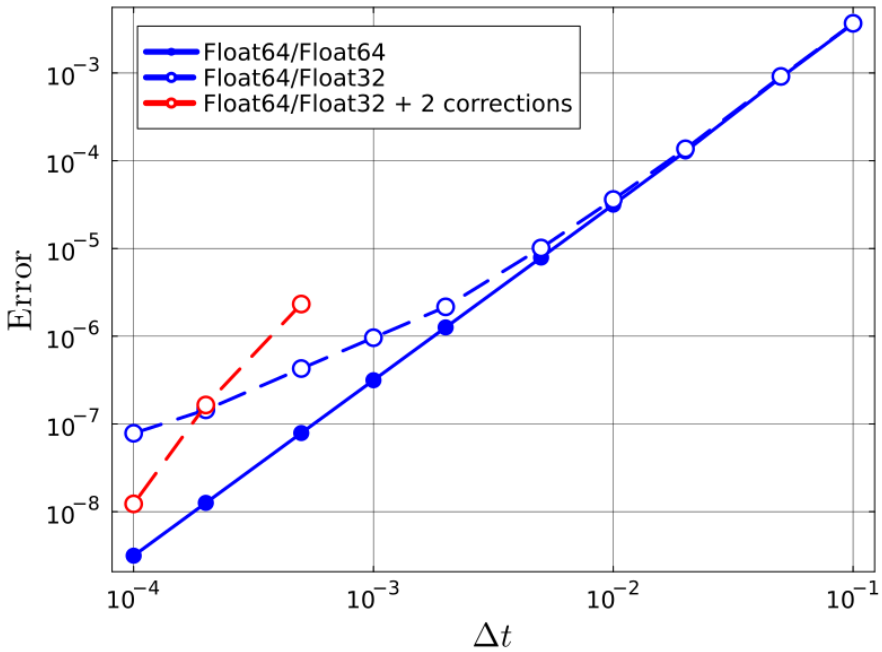
- Semi-discretization with N_x spatial points
- Nonlinear term approximated using Fourier spectral methods
- Final time: $T_f = 0.5$

IMR with $N_x = 200$

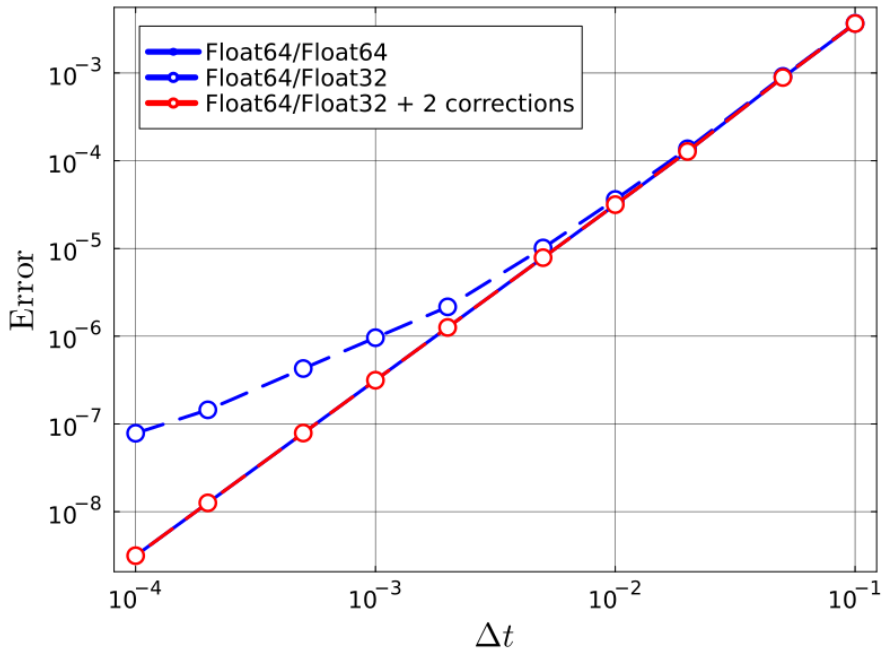


Explicit corrections

IMR with $N_x = 200$



Explicit corrections

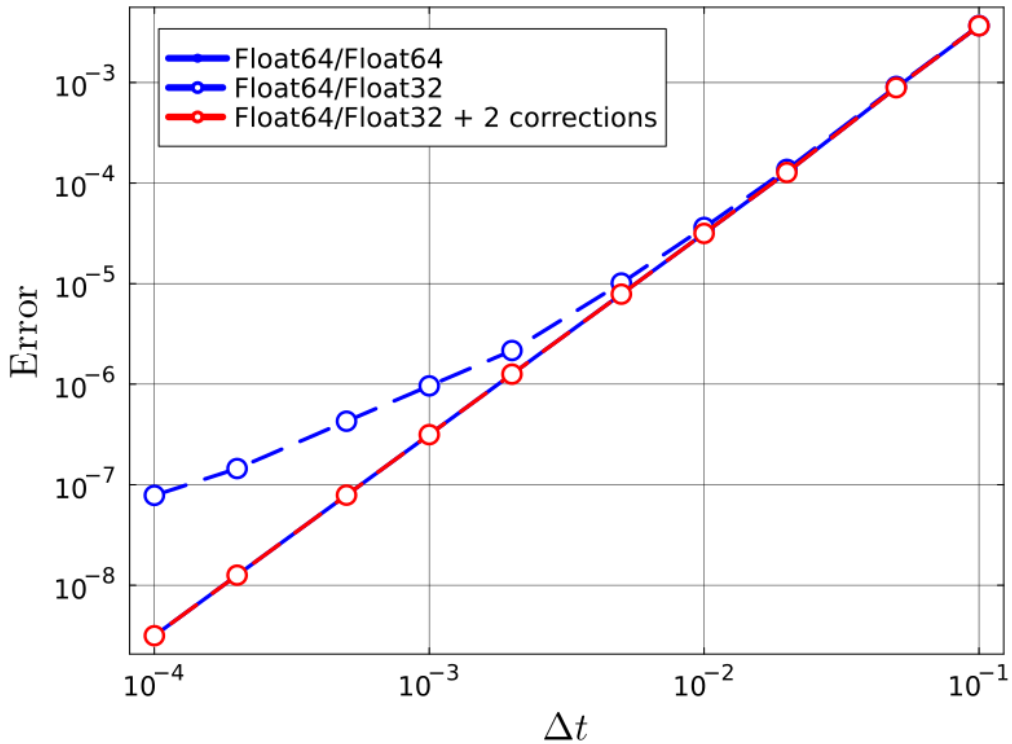


Stabilized corrections

Speedups

Comparison: no corrections vs two stabilized corrections for 64/32.

Δt	No corr.	Two corr.
1E-02	**3.6** $\times$	**3.0** $\times$
1E-04	**5.0** $\times$	**4.5** $\times$
1E-05	**3.1** $\times$	**1.9** $\times$



Conclusions

- Mixed-precision RK methods can provide significant runtime savings while maintaining good accuracy.
- Explicit corrections are beneficial for sufficiently small Δt, but they affect stability.
- New work presents alternative stable corrections:
  - J. Driscoll, S. Gottlieb, Z. J. Grant, C. Herrera, T. S. Kakumanu, M. H. Sawicki, and M. Stephens. **Stable Corrections for Perturbed Diagonally Implicit Runge–Kutta Methods.** arXiv:2603.24451.

Thanks

- For more details see:
  - Z. Grant **Perturbed Runge–Kutta Methods for Mixed Precision Applications** *SISC, 2022.*
  - B. Burnett, S. Gottlieb, and Z. J. Grant. **Stability Analysis and Performance Evaluation of Additive Mixed-Precision Runge-Kutta Methods.** *COM APPL MATH COMPUT, 2023.*
  - J. Driscoll, S. Gottlieb, Z. J. Grant, C. Herrera, T. S. Kakumanu, M. H. Sawicki and M. Stephens. **Stable Corrections for Perturbed Diagonally Implicit Runge–Kutta Methods.** *arXiv:2603.24451.*
  - S. Gottlieb, Z. J. Grant, and C. Herrera. **Mixed Precision Explicit Two-Derivative Runge–Kutta Methods.** *arXiv:2602.14369*

My website

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1929284, while the author was in residence at the Institute for Computational and Experimental Research in Mathematics (ICERM) in Providence, RI, during the program **"Empowering a Diverse Computational Mathematics Research Community."**

Additional support was provided by:

- *The Air Force Office of Scientific Research under Grant No. FA9550-23-1-0037,*
- *The U.S. Department of Energy under Grant No. DE-SC0023164.*