Using Numerical Profiling to Determine Where Mixed Precision is Usable in Multiphase Flow Simulations

MxP4S Workshop | IPDPS 2026

Akash Dhruv

Mathematics and Computer Science Division
Argonne National Laboratory

In collaboration with: F. Hoerold (ETH Zurich / RIKEN) , I. Ivanov (Tokyo Tech / RIKEN) , W. Moses (UIUC) , A. Dubey (RIKEN) , M. Wahib (RIKEN) , J. Domke (RIKEN)

Argonne NATIONAL LABORATORY

1 / Motivation

The Precision Landscape in Modern HPC

The hardware-software precision gap is widening

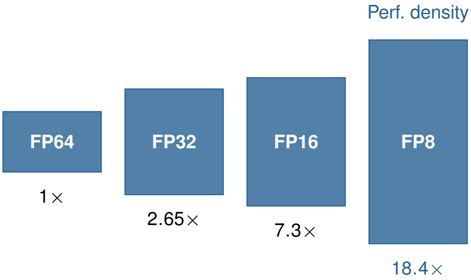
- FP8 delivers **18**$\times$ FP64 throughput on today's accelerators.
- FP16/BF16 proliferating; FP64 capacity stagnating.

Scientific codes stay at FP64

- Naive `float`$\rightarrow$`half` replacement breaks simulations.
- No theory for precision reduction in nonlinear PDEs.

Key Question

Which parts of a simulation are safe to run at lower precision?



The Challenge: Non-linear Multiphysics Codes

Why is precision reduction non-trivial?

Linear algebra has rigorous theory

- Iterative refinement, backward stability, mixed-precision GEMM—well-understood.

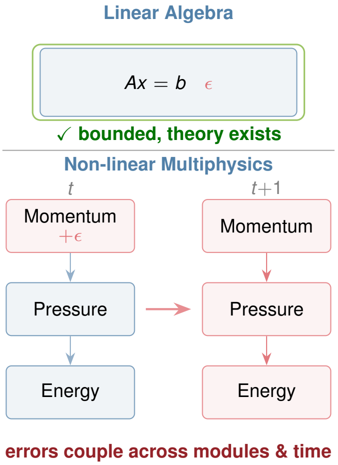
Non-linear multiphysics does not

- AMR, stiff ODEs, shocks: errors couple across modules and time steps.
- Scientific intuition is *not always correct*.

The Need

An **interactive tool** for domain scientists :

- Truncate regions; observe solution quality impact.
- Guide cost-benefit—without touching source code.



2 / RAPTOR: Numerical Profiling Tool

Introducing RAPTOR

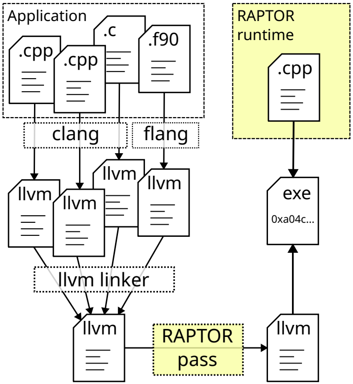
Practical Numerical Profiling of Scientific Applications

Goal

- Transparently replace FP64 ops via LLVM pass + MPFR—any code region.
- C/C++/Fortran · CPU+GPU · OpenMP+MPI.

Key differentiators

- Full-application truncation—no kernel extraction required.
- **Dynamic**: time-varying; arbitrary precision formats (MPFR).
- Scoped/granular truncation; error tracking via shadow variables.



Open Source

`https://github.com/RIKEN-RCCS/RAPTOR`

F. Hoerold, I. Ivanov, A. Dhruv, W. Moses, A. Dubey, M. Wahib, J. Domke. "RAPTOR: Practical Numerical Profiling of Scientific Applications." *Proc. SC'25*, ACM, 2025. doi:10.1145/3712285.3759810

RAPTOR: Two Modes of Operation

Mem-Mode

- MPFR state persisted; shadow FP64 variable tracks cumulative error.
- Pinpoints sensitive locations—also supports precision *increase*.

Op-Mode

- Truncates each FP op individually; minimal annotation.
- Overhead $\propto$ fraction truncated ($36\times$ at 86%)—mainly when emulating arbitrary precision using MPFR.

```
...
auto f = _raptor_trunc_func_mem(
  /* function */    foo,
  /* from_type */   32,
  /* to_exponent */ 5,
  /* to_mantissa */ 8);
a[0] = _raptor_pre_c (a[0], 5, 8);
a[1] = _raptor_pre_c (a[1], 5, 8);
b    = _raptor_pre_c (b,    5, 8);
float c = f(a, b);
a[0] = _raptor_post_c(a[0], 5, 8);
a[1] = _raptor_post_c(a[1], 5, 8);
b    = _raptor_post_c(b,    5, 8);
c    = _raptor_post_c(c,    5, 8);
...
```

(c) Mem-mode usage

```
void bar(float a, float b) {
  return a + b;
}
void foo(float *a, float b) {
  a[0] = sqrt(b);
  return bar(a[1], b);
}
```

```
...
foo(a, b)
...
```

(a) Original Code

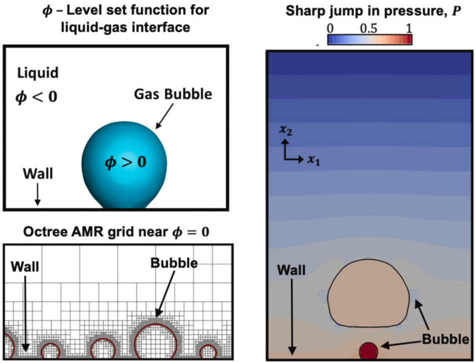
```
...
auto f = _raptor_trunc_func_op(
  /* function */    foo,
  /* from_type */   32,
  /* to_exponent */ 5,
  /* to_mantissa */ 8);
f(a, b)
...
```

(b) Op-mode usage

F. Hoerold, I. Ivanov, A. Dhruv, W. Moses, A. Dubey, M. Wahib, J. Domke. "RAPTOR: Practical Numerical Profiling of Scientific Applications." *Proc. SC'25*, ACM, 2025. doi:10.1145/3712285.3759810

3 / Application: Multiphase Flows with AMR

Incompressible Multiphase Dynamics in Flash-X



Left: Octree AMR concentrates at ϕ=0. **Right:** Sharp pressure jump *P* resolved by a domain-wide multigrid Poisson solve.

Incompressible Navier-Stokes solver

- Fractional-step projection: pressure via domain-wide multigrid Poisson solve (AMReX).
- Sharp-interface GFM: exact jump conditions at liquid–gas boundary; interface tracked by level-set ϕ.
- WENO advection / central finite-difference for diffusion.

Why precision matters here

All components affect ϕ=0, the region where the interface lives and where the physics most matters.

Hypothesis

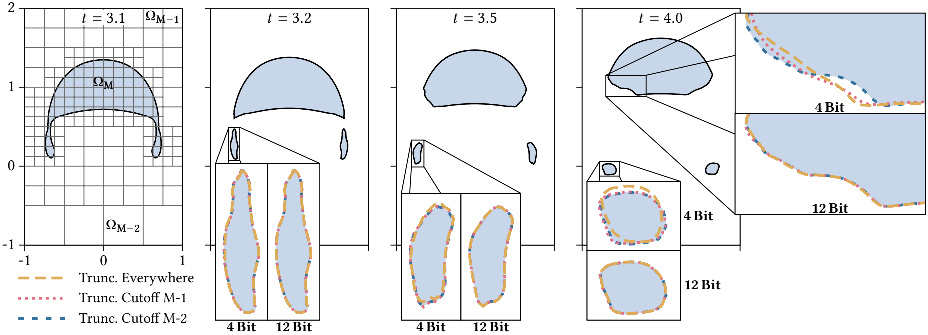
AMR-selective truncation preserves bubble interface topology.

A. Dhruv. "A Vortex Damping Outflow Forcing for Multiphase Flows with Sharp Interfacial Jumps." *J. Comput. Phys.* 511, Elsevier, 2024. doi:10.1016/j.jcp.2024.113122

4 / Experimental Results

Argonne NATIONAL LABORATORY

Rising Bubble Results at High Reynolds Number



Setup: Reynolds Number, Re = 3500 (inertial-dominated); WENO advection + diffusion truncated globally (M-0) vs. AMR-selective (finest two levels preserved); ϕ=0 interface topology at t=3.1–4.0. **Result:** 4-bit global truncation distorts bubble shape and satellite splitting; **12-bit AMR-selective** preserves deformation and break-up dynamics. ✓ **Hypothesis confirmed.**

F. Hoerold, I. Ivanov, A. Dhruv, W. Moses, A. Dubey, M. Wahib, J. Domke. "RAPTOR: Practical Numerical Profiling of Scientific Applications." *Proc. SC'25*, ACM, 2025. doi:10.1145/3712285.3759810

Advection vs. Diffusion: Advection Sets the Precision Floor

What we expected

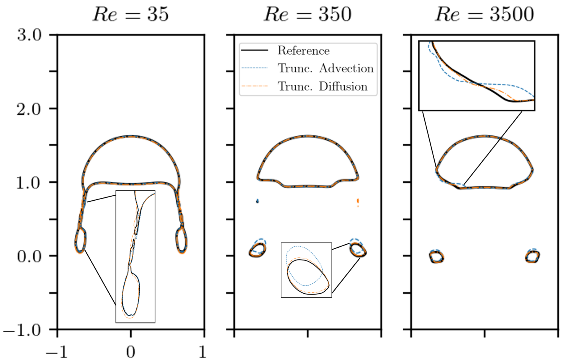
- At low Reynolds Number diffusion dominates and there is more bubble stretching—more precision sensitivity.

Results

- Truncating advection degrades ϕ=0 more adversely than diffusion.
- 5^{th} order WENO has larger stencils than 2^{nd} order central difference.

Next candidate: Poisson solver

The multigrid Poisson solve couples the full domain; pressure errors broadcast globally. Does it impose its own precision floor?

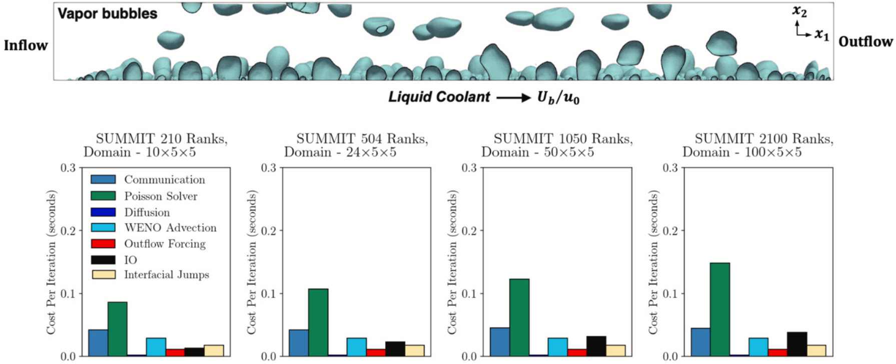


Trunc. advection (blue dashed) vs. diffusion (orange dashed) at Re = 35, 350, 3500. Insets: ϕ=0 deviation from reference (black).

Argonne NATIONAL LABORATORY

5 / Conclusions & Outlook

Multigrid Poisson and Performance



Poisson solve: **50–70%** of total cost

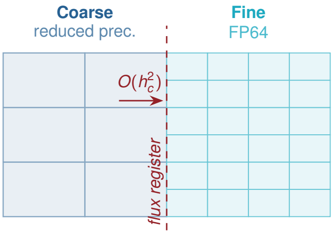
Application: Flow-boiling on SUMMIT (210–2100 MPI ranks); cost breakdown across solver components at four concurrency levels.

A. Dhruv. "A Vortex Damping Outflow Forcing for Multiphase Flows with Sharp Interfacial Jumps." *J. Comput. Phys.* 511, Elsevier, 2024. doi:10.1016/j.jcp.2024.113122

Precision Sensitivity in Multigrid Solvers

Mixed-precision loss mechanisms

- **Variable-coefficient operator:** density ratio, $\rho_l/\rho_g \sim 10^3$ shifts the precision floor per AMR level.
- **AMR flux registers:** coarse/fine boundary fluxes must agree; a reduced-precision coarse solve injects $O(h_c^2)$ errors directly into the FP64 fine-grid residual.
- **V-cycle convergence:** residual $r^{(k)} = f - \nabla \cdot \left(\frac{1}{\rho}\nabla P^{(k)}\right)$; bottom-solver precision governs the convergence rate $\|r^{(k+1)}\|/\|r^{(k)}\|$.
- **Global pressure coupling:** incompressibility broadcasts errors across the full domain; coarse-level savings must not corrupt the fine-grid solution.



convergence rate $\propto$ bottom-solver precision

$$\nabla \cdot \left(\frac{1}{\rho}\nabla P\right) = f$$

Take Home

What we learned

- **AMR-selective truncation** at 12-bit preserves bubble topology; finest levels near ϕ=0 must stay at FP64.
- **WENO advection with larger stencil sets the precision floor**—flow regime does not.
- FP64-only codes **forfeit most hardware gains** on accelerators available today—and may face even greater penalties in the future.

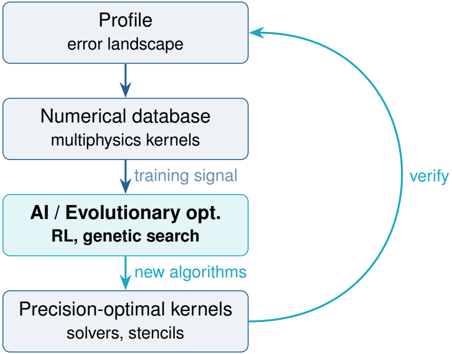
Take-home

Empirical profiling reveals what intuition cannot.

Current collaboration to leverage AI

F. Cappello (ANL) R. Thakur (ANL) A. Siegel (ANL)
X. Li (LBNL) W. Zhang (LBNL) M. Natarajan (LBNL)

Vision: closing the loop with AI



Thank You

Questions?