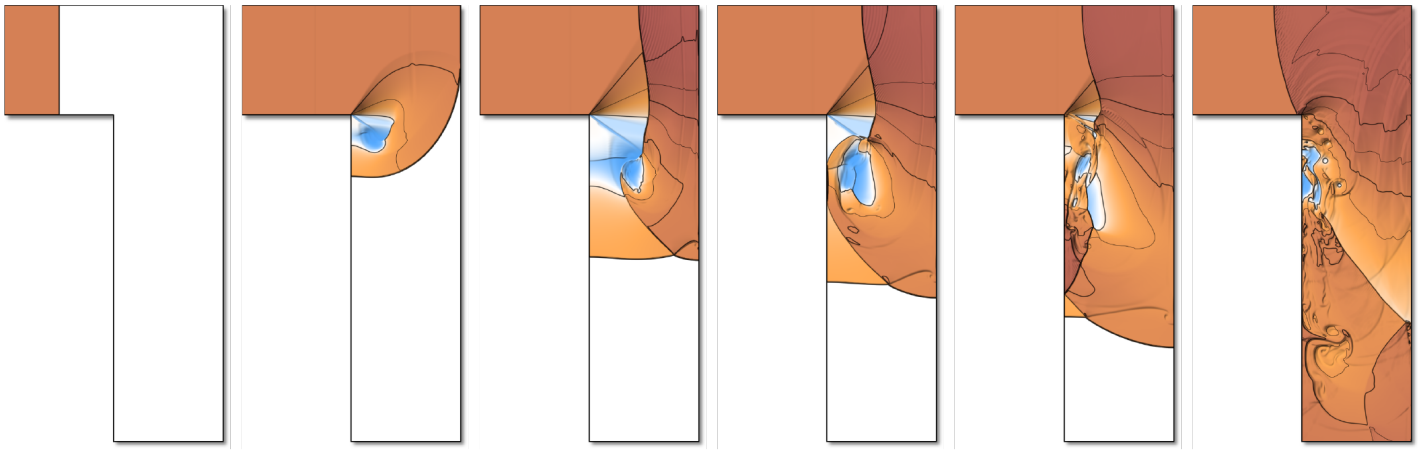
VIVIDSPARKS

CASE STUDY

POSIT™: Alternative to IEEE Floating Point



The increasing relative cost of data movement relative to floating point operations makes this an opportune time to re-evaluate the IEEE floating point representations of the real numbers. Representations such as the POSITs have greater arithmetic closure, smaller mean representation error, more efficient representation of infinity and NaNs, and produce lower roundoff error (often by at least an order of magnitude) than the standard IEEE types in use today.

Dr. Peter Lindstrom,
Lawrence Livermore National Lab (LLNL),
USA

Introduction

Floating point computations required in many applications. Floating-point representations builds on the work done in the 1970s and 1980s, before the IEEE standard was developed and codified in hardware. The choices made in IEEE leave much to be desired when accurate arithmetic and reproducibility of numerical computations are paramount, including support for an excessive number of not-a-number (NaN) representations, gradual underflow but not overflow, expensive exceptions such as subnormals, ambiguity due to two representations of zero, etc. POSIT representation have shown to provide a better trade-off between precision and dynamic range, and in effect increase the accuracy per bit stored.

Challenges

As the high-performance computing community pushes toward exascale computing, it is becoming increasingly clear that data movement and data storage will be the dominant performance bottleneck for the foreseeable future, leading developers to re-evaluate the need for wide data types that consume precious memory bandwidth. Increase in memory and bandwidth leads to slower performance and heavier applications become.

Approach

Our framework is implemented in C++ using templates and operator overloading to simplify integration with applications. The experiment modified a numerical simulation application, Euler2D, which implements an explicit, high-resolution Godunov algorithm to solve the Euler system of equations for compressible gas dynamics on an L-shaped domain. Such a solver is simple enough to instrument and comprehend while providing sufficient complexity in the numerical behavior of the solution, e.g. a nonlinear hyperbolic system with shock formations and minimal dissipation.

The problem solved in the Euler2D code is the propagation of a shock wave in air through an L-shaped conduit. The domain is the union of two rectangles: [(0,3), (2,4)] U [(1,0), (2,3)]. At the initial time, a shock, moving with dimensionless speed M_s = 2.5 relative to the quiescent state of (ρ, u_x, u_y, p) = (1,0,0,1), is positioned at x= 0.5. The inlet flow at x = 0 is constant. The code is run with uniform mesh of size h = 1 / n= 1 /256 using a fixed time step of $\Delta t \approx 2.8\ e^{-4}$, resulting in roughly 1.3 trillion floating-point operations over the entire run. The system dynamics are shown in Figure 1:

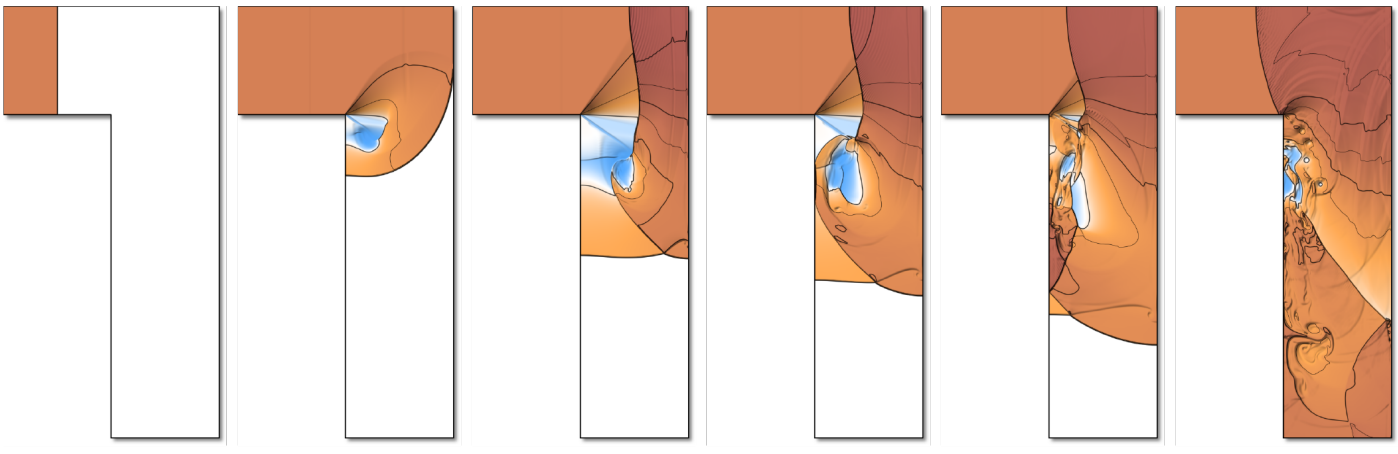


Figure 1: Snapshots in time, t, from the Euler2D mini-application showing the evolution of the density field in an L-shaped chamber. Blue color indicates density lower than the initial density (red) of the shock wave. (a) Initial state. (b) Shock reflects off of the far wall. (c) Reflected shock hits vortex. (d) Shock reflects off of near wall. (e) Second reflection hits vortex. (f) Final state.

As shown in Figure 1, the shock propagates into the chamber and diffracts around the corner, initiating the shedding of a vortex from the corner. At time t ≈ 0.51, the initial shock reflects off the far wall, and the reflected shock propagates back upstream, encountering the vortex around time t ≈ 0.95. The reflected shock breaks up the vortices shedding off the corner and reflects again off the near wall at several times. Eventually, the flow moves down the channel with a propagating sequence of oblique shock waves and a great deal of wave-wave interactions.

A pointwise, closed form solution to the Euler 2D hyperbolic PDE does not exist. To establish ground truth, the LLNL team used a quadruple precision floating point type to compute a high-precision solution. The team then computed the root mean square pointwise error in the density field to establish solution accuracy. The team reported that the RMS error was expected to be dominated by round-off error associated with each numerical type due to fixed discretization parameters, i.e., fixed truncation error. Plots of the pointwise error in the density field over time for 32-bit and 64-bit representations relative to the quadruple precision are shown in Figure 2 and Figure 3 (next page). We see spikes in error that correlate with events such as shock-wall and shock-vortex impact. These spikes are more pronounced in the 64-bit plot because of the additional precision provided in 64-bit arithmetic.

IEEE floating point and related types do quite poorly in relation to posits and other tapered precision numerical types, most evident in the 64-bit precision plot, where posit<64,2> outperforms IEEE double precision by nearly three orders of magnitude.

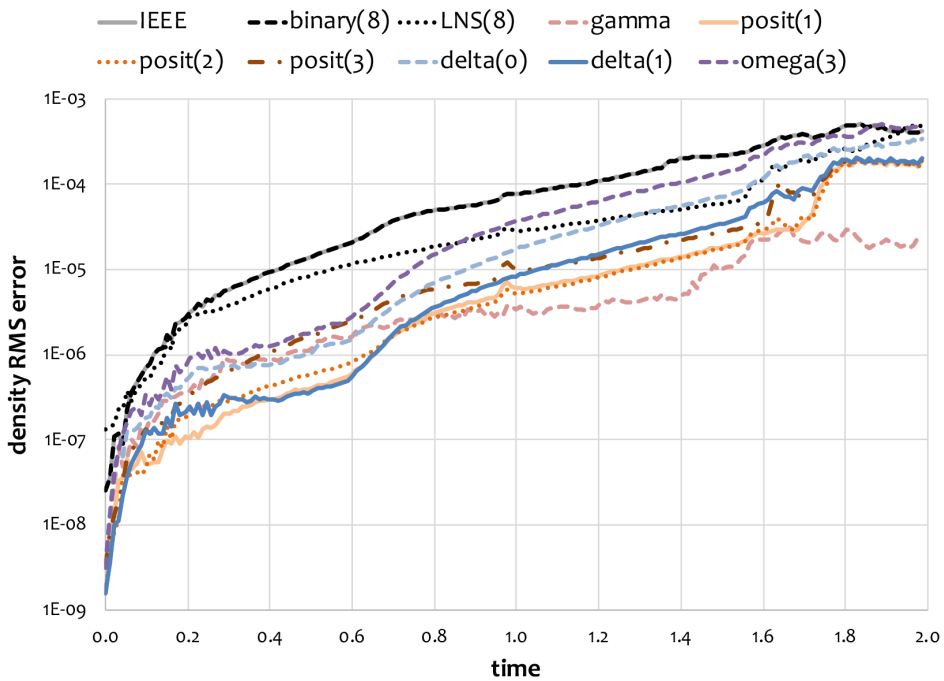


Figure 2: RMS error in Euler2D density field as a function of simulation time and 32-bit number representation.

Benefits

With Dennard scaling and Moore's Law having run their course, performance improvements required to solve the computational demands of modern AI/Deep Learning, Big Data, and IoT analytics applications must come from innovations that improve the efficiency of computation. Replacing IEEE Floating Point with more efficient alternatives like POSIT arithmetic provides a key opportunity to improve our computational infrastructure. POSIT arithmetic provides a mechanism to consolidate the benefits of a standardized number system like IEEE Floating Point across the next generation application domains of AI/Deep Learning, Big Data/Deep Analytics, IoT, and Robotics (including autonomous vehicles).

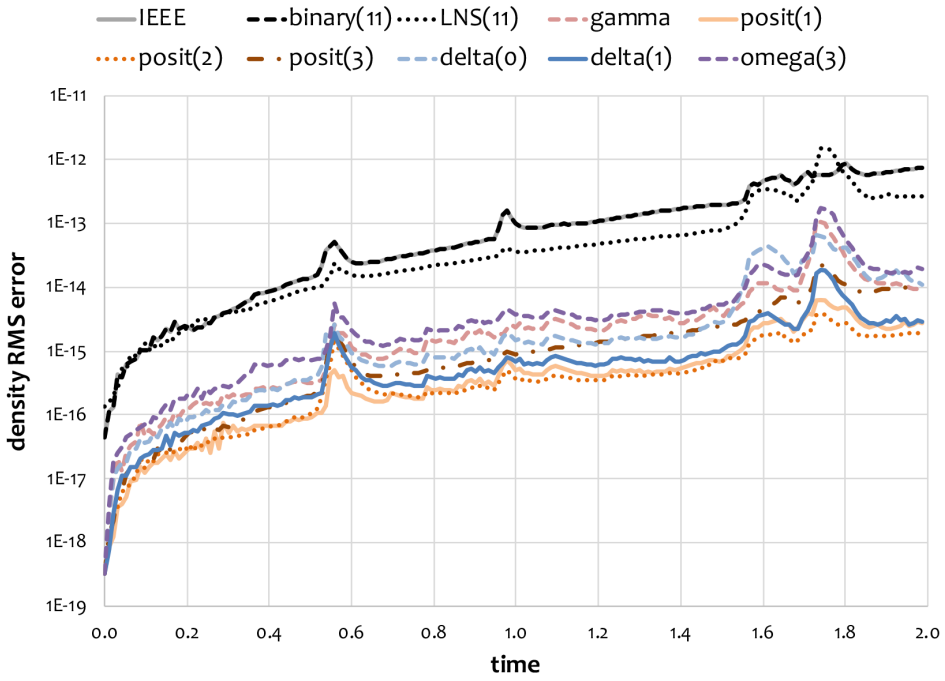


Figure 3: RMS error in Euler2D density field as a function of simulation time and 64-bit number representation.

Company Description

LLNL located in the San Francisco Bay Area, is a premier applied science laboratory that is part of the National Nuclear Security Administration within the Department of Energy.

LLNL's mission is strengthening national security by developing and applying cutting-edge science, technology, and engineering that respond with vision, quality, integrity, and technical excellence to scientific issues of national importance. The Laboratory's science and engineering are being applied to achieve breakthroughs for counter terrorism and nonproliferation, defense and intelligence, energy and environmental security.