Order-Invariant Real Number Summation

Patric E. Small & Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Department of Computer Science
Department of Physics & Astronomy
Department of Chemical Engineering & Materials Science
Department of Biological Sciences
University of Southern California

Email: anakano@usc.edu

• Rounding (truncation) error makes floating-point addition non-associative

• Sum becomes a random walk across the space of possible rounding error
High-Precision (HP) Method

- Propose an extension of the order-invariant, higher-precision intermediate-sum method by Hallberg & Adcroft [Par. Comput. 40, 140 (’14)]

- The proposed variation represents a real number \( r \) using a set of \( N \) 64-bit unsigned integers, \( a_i \) \((i = 0, N-1)\)

\[
r = \sum_{i=0}^{N-1} a_i 2^{64(N-k-i-1)}
\]

\[
= a_0 2^{64(N-k-1)} + \cdots + a_{N-k-1} + \cdots + a_{N-k} 2^{-64} + \cdots + a_{N-1} 2^{-64k}
\]

- \( k \) is the number of 64-bit unsigned integers assigned to represent the fractional portion of \( r \) \((0 \leq k \leq N)\), whereas \( N-k \) integers represent the whole-number component

- Negative number is represented by two’s complement in integer representation, using only 1 bit
Performance Projection

- HP sum is faster than Hallberg sum for higher precision & larger numbers of summands