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 Quantitative & Computational Biology
University of Southern California

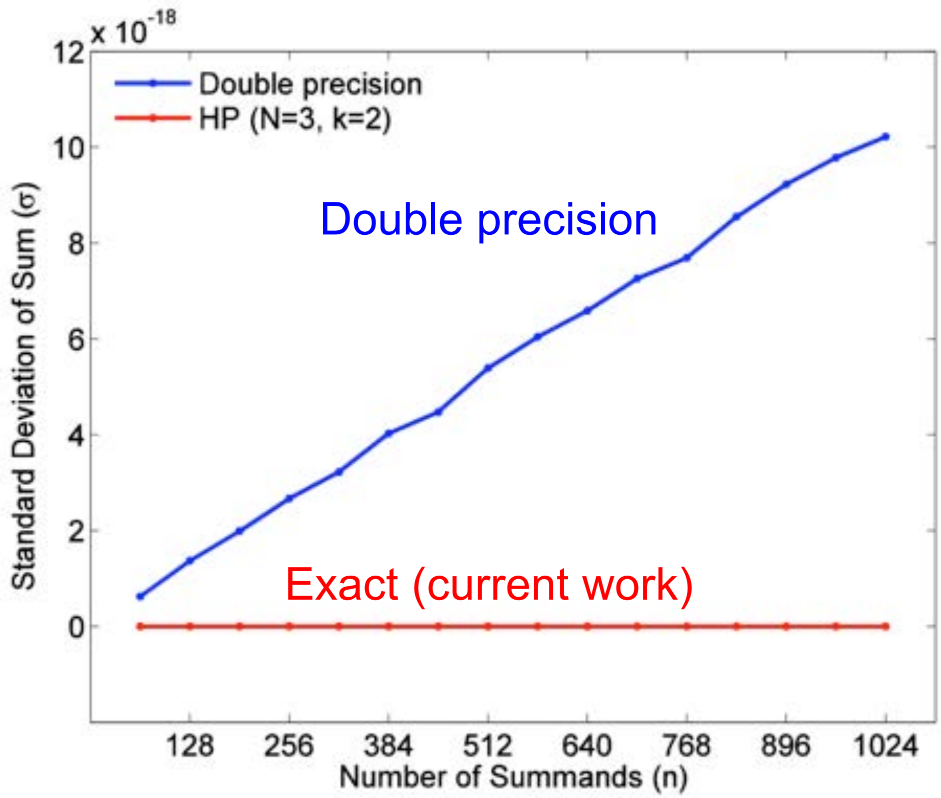
Email: anakano@usc.edu

P. E. Small *et al.*, *Proc. IEEE IPDPS*, p. 152 ('16)
https://aiichironakano.github.io/cs596/Small-OrderInvariantSum-IPDPS16.pdf

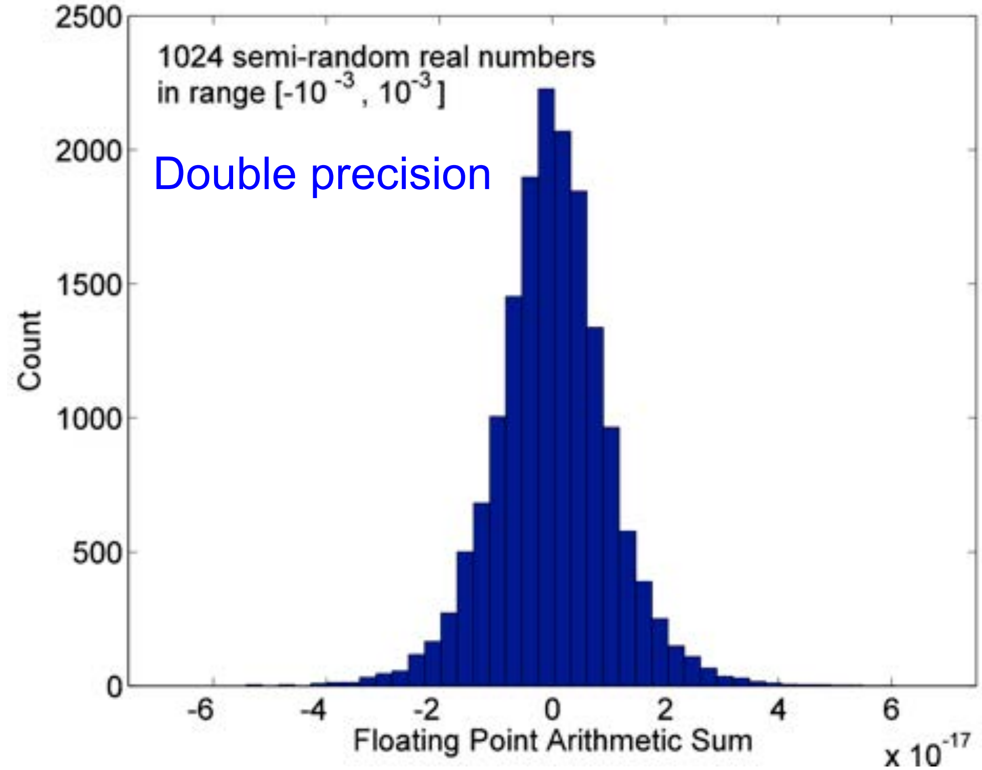
Reproducibility Challenge

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

$$(a+b)+c \neq a+(b+c)$$



Standard deviation of sum with random summation orders



Distribution of sum with random summation orders

- **Finding: Sum becomes a random walk across the space of possible rounding error**

Solution: 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, $\boldsymbol{a_i}$ ($\boldsymbol{i \in [0, N-1]}$)**

$$r = \sum_{i=0}^{N-1} a_i 2^{64(N-k-i-1)}$$
$$= \overbrace{a_0 2^{64(N-k-1)} + \cdots + a_{N-k-1}}^{N-k} + \overbrace{a_{N-k} 2^{-64} + \cdots + a_{N-1} 2^{-64k}}^{k}$$

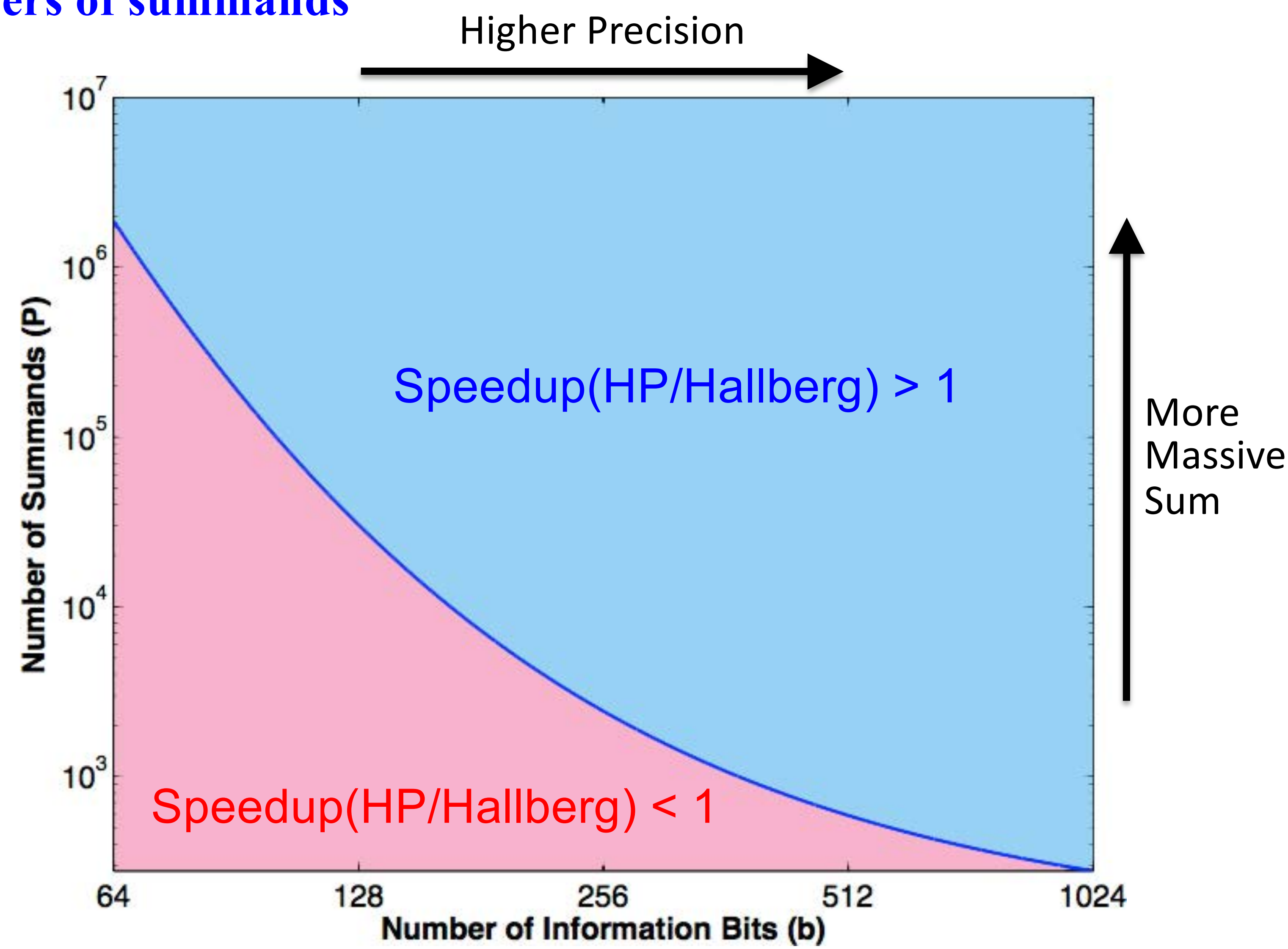
- ***k* is the number of 64-bit unsigned integers assigned to represent the *fractional* portion of *r* (0 ≤ *k* ≤ *N*), whereas *N*-*k* integers represent the *whole-number* component**

- **Negative number is represented by two's complement in integer representation, using only 1 bit**

If you are the first to find the problem, the simplest solution suffices to prove the concept

Performance Projection

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



P. E. Small *et al.*, *Proc. IEEE IPDPS*, p. 152 ('16)