Mixed Precision Math Unlocks Accelerated Computing

Piotr Luszczek

May 25, 2026

Research was sponsored by the Department of the Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Numerical Precision Overview

Common Integer Representations

int8_t:

Range: -128 (-2^7) to 127 ($2^7 - 1$)

ch…

Int8 most commonly used for inference to save memory and power

short, int16_t:

Range: -32,768 (-2^{15}) to 32,76…

unsigned short, uint16_t:

Range: 0 to 65,535 ($2^{16} - 1$)

int, int32_t:

Range: -2,147,483,648 (-2^{31}) to 2,147,483,647 …

Differentiate neural outcomes at every neuron

unsigned int, uint32_t:

Range: 0 to 4,294,967,295 ($2^{32} - 1$)

Common Floating-Point Representations

bfloat16: Brain Floating Point Format

Range: $\sim 1\times10^{-38}$ to $\sim 3\times10^{38}$ Number of Decimal Digits: ~2.3

Exponent: 8 bits Mantissa (Significand): 7 bits

fp16: Half-precision IEEE Floating Point Format

Range: $\sim 5.96\times10^{-8}$ to 6.50×10^{4} Number of Decimal Digits: ~3.3

Exponent: 5 bits Mantissa (Significand): 10 bits

fp32: Single-precision IEEE Floating Point …

Range: $\sim 1\times10$… $\times10^{38}$

Exponent: …

Bfloat16 most commonly used for training to save memory and power

Need small increments to accommodate training micro-adjustments

fp64: Double-precision IEEE Floating Point Format

Range: $\sim 1\times10^{-308}$ to $\sim 3\times10^{308}$

Exponent: 11 bits

Example: 123.45 = 1.2345×10^2 = 0.12345×10^3 = 0100 0010 1111 0110 1110 0110 0110 0110 (fp32)

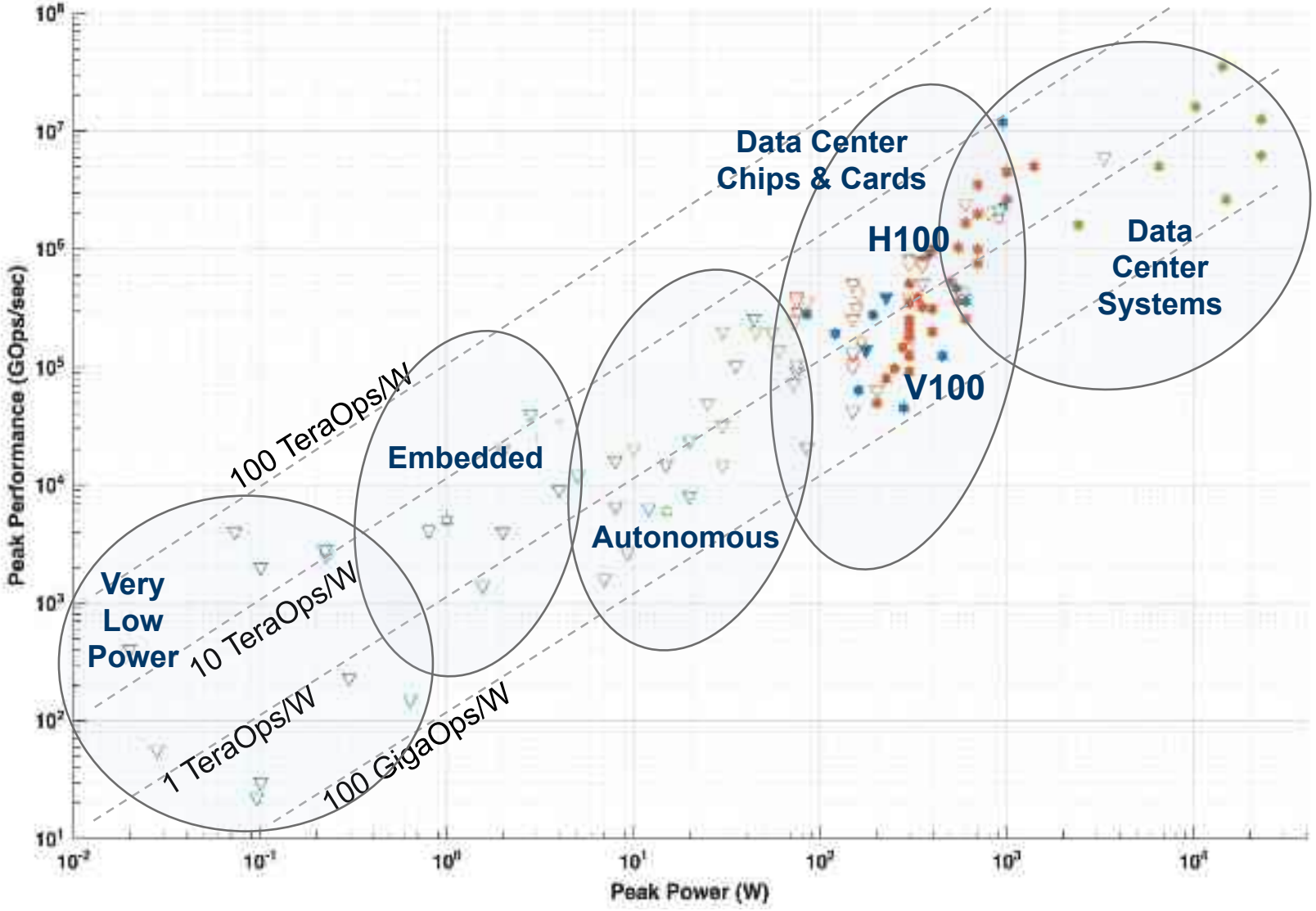
\+ Sign bit of floating-point number

E Exponent bit of floating-point number

M Mantissa bit of floating-point number

b Bit of a fixed-point number

Lincoln Lincoln AI Computing Survey (LAICS)



Legend

Computation Precision

- analog
- int1
- int2
- int4
- int8
- fp8
- int16
- int32
- fp16
- bf16
- fp32
- fp64

Form Factor

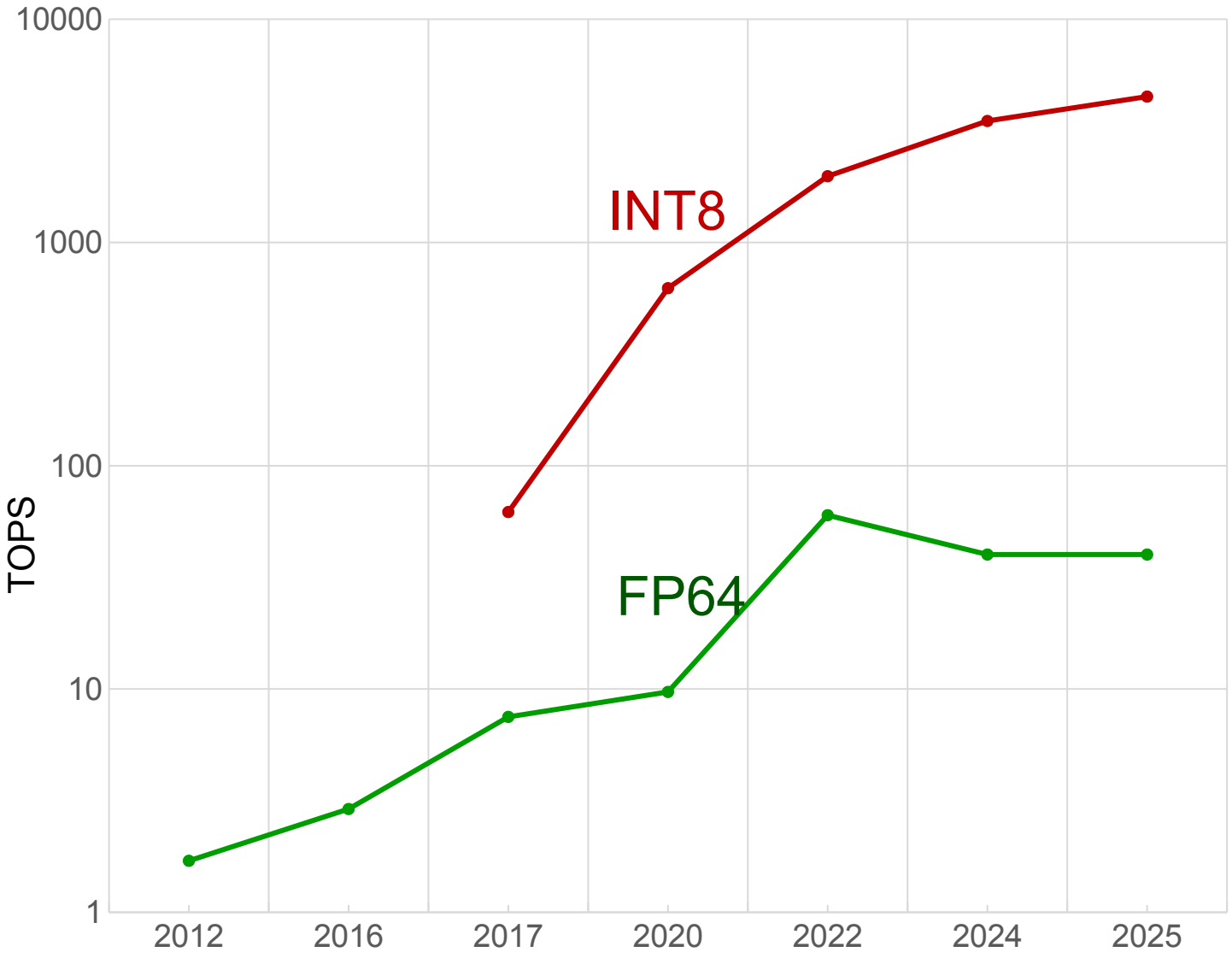
- Chip
- Card
- System

Computation Type

- Inference
- Training

Slide courtesy of Albert Reuther, MIT Lincoln Laboratory Supercomputing Center,, DOI: 10.1109/HPEC.2019.8916327

Trends in Common Datacenter AI Accelerators

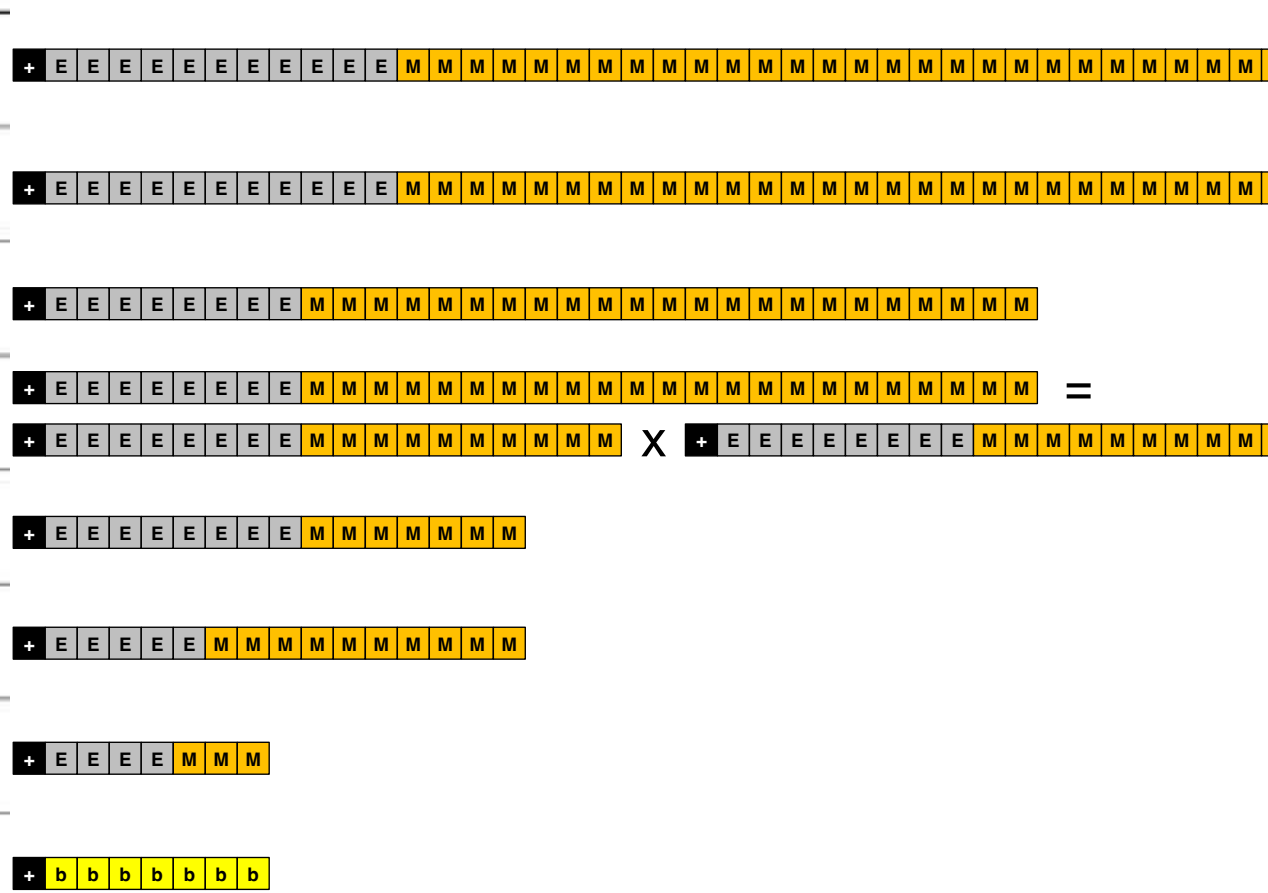


- **High precision arithmetic (FP64) critical to government applications leveling off at ~100 Teraflops per chip**
- **Lower precision arithmetic critical to AI computations is pushing past 20 Petaflops per chip**

FP64 = 64-bit floating point
INT8 = 8-bit integer

TX-GAIN Mixed Precision Hardware

Technical Specifications	NVIDIA H100 NVL
FP64	30 teraFLOPS
FP64 Tensor Core	60 teraFLOPS
FP32	60 teraFLOPS
TF32 Tensor Core*	835 teraFLOPS
BFLOAT16 Tensor Core*	1,671 teraFLOPS
FP16 Tensor Core*	1,671 teraFLOPS
FP8 Tensor Core*	3,341 teraFLOPS
INT8 Tensor Core*	3,341 TOPS



LLSC - 5

*With sparsity

https://resources.nvidia.com/en-us-hopper-architecture/nvidia-tensor-core-gpu-datasheet

[1] *Decompositional Factorizations with FP64 Floating-Point Emulation in INT8*, Luszczek et al, IEEE HPEC

[2] *Parallel Hypersparse, Matrix Based Graph Streaming, and Complex-Index Matrices*, Jananthan et al, IEEE HPEC

Mixed Precision Performance Across Industry

Microsoft amazon Google

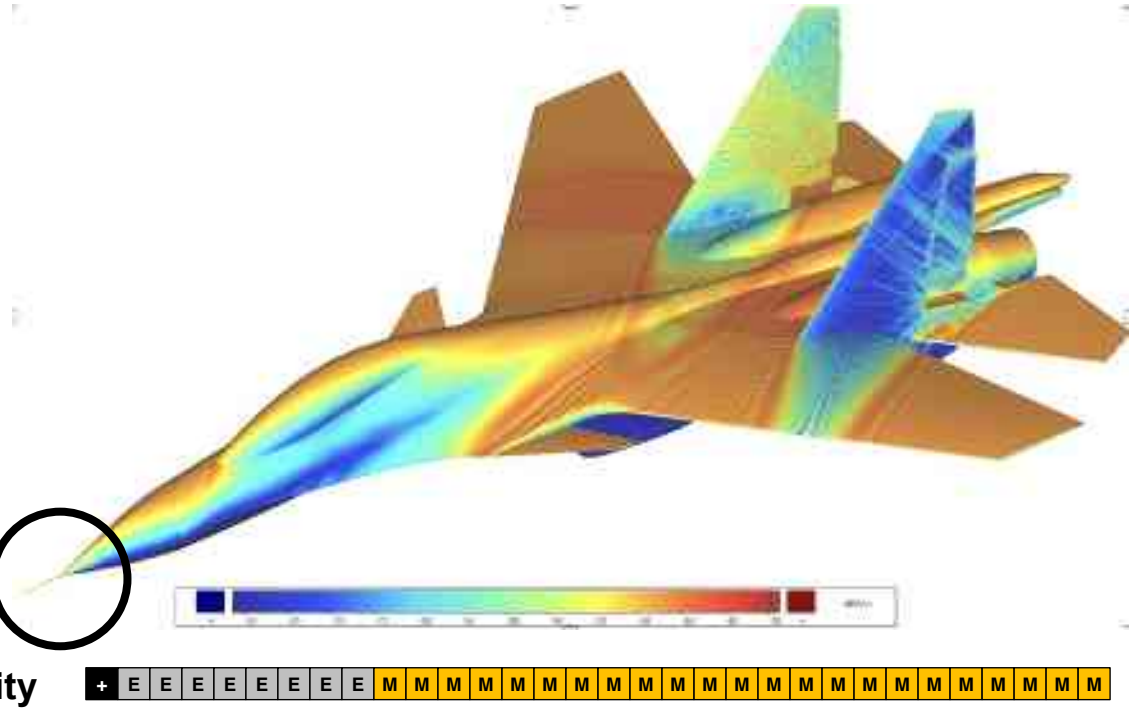
Peak specifications	Azure Maia 200	AWS Trainium3	Google TPU v7
Process Node	3nm	3nm	3nm
FP4 TFLOPS + E E M	10,145	2,517	n/a
FP8 TFLOPS + E E E E M M M	5,072	2,517	4,614
BF16 TFLOPS + E E E E E E E E M M M M M M M	1,268	671	2,307

- Many leading accelerators only have low-precision
- Rely on novel methods for emulating high-precision with low-precision

https://www.hpcwire.com/aiwire/2026/01/27/microsoft-raises-the-ai-inference-bar-with-maia-200/
https://www.hpcwire.com/2026/02/17/genesis-mission-will-lean-heavily-on-ozaki-scheme-for-fp64-capability/

The Need for High Fidelity in Physics Models

- Many physical phenomena require fast and accurate modeling
  - Impact modeling
  - Radar cross-section
  - Fluid dynamics simulations
- Accurate simulations are key sources of data for AI models like
  - PINNs
  - Fourier neural operators
  - Equivariant neural networks

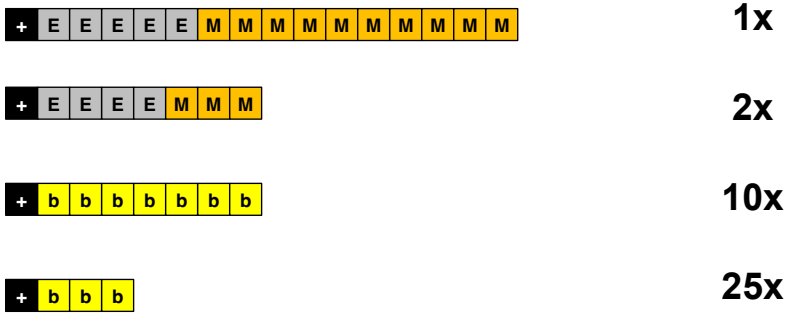


Physical phenomena require high precision for fidelity

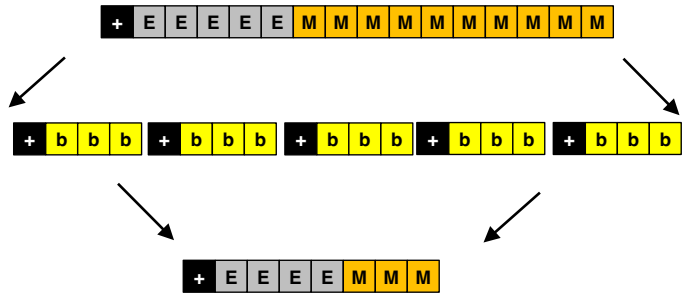
Challenge: meeting accuracy requirements with faster lower precision arithmetic

Potential Ways to Leverage Newer Hardware

- **Multiple precision in scientific applications**
  - **Mixed-precision computation is now commonplace**
- **Lower precisions are faster per-bit than traditional IEEE types**



- **Emulating FP64 with tensor units is known as Ozaki Scheme**



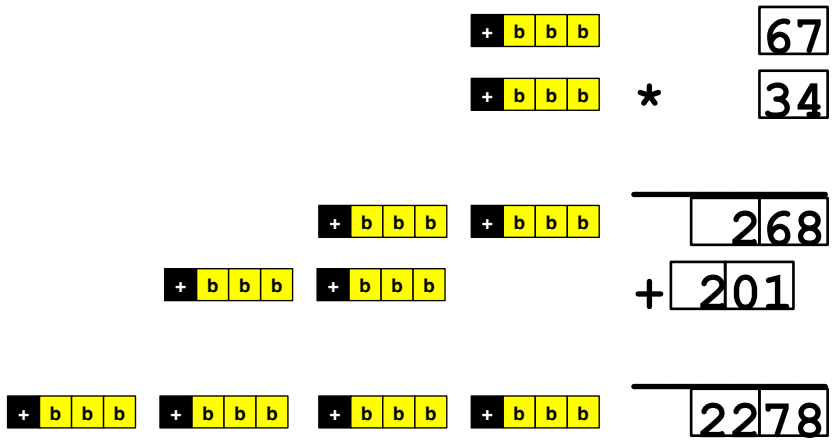
Maintaining accuracy and the range of the calculation becomes challenging

Focus: identify issues in using integer-only emulation of FP64

Using low-precision types delivers greater per-bit benefits

FP64 = 64-bit floating point
FP32 = 32-bit floating point

Basic Multiplication

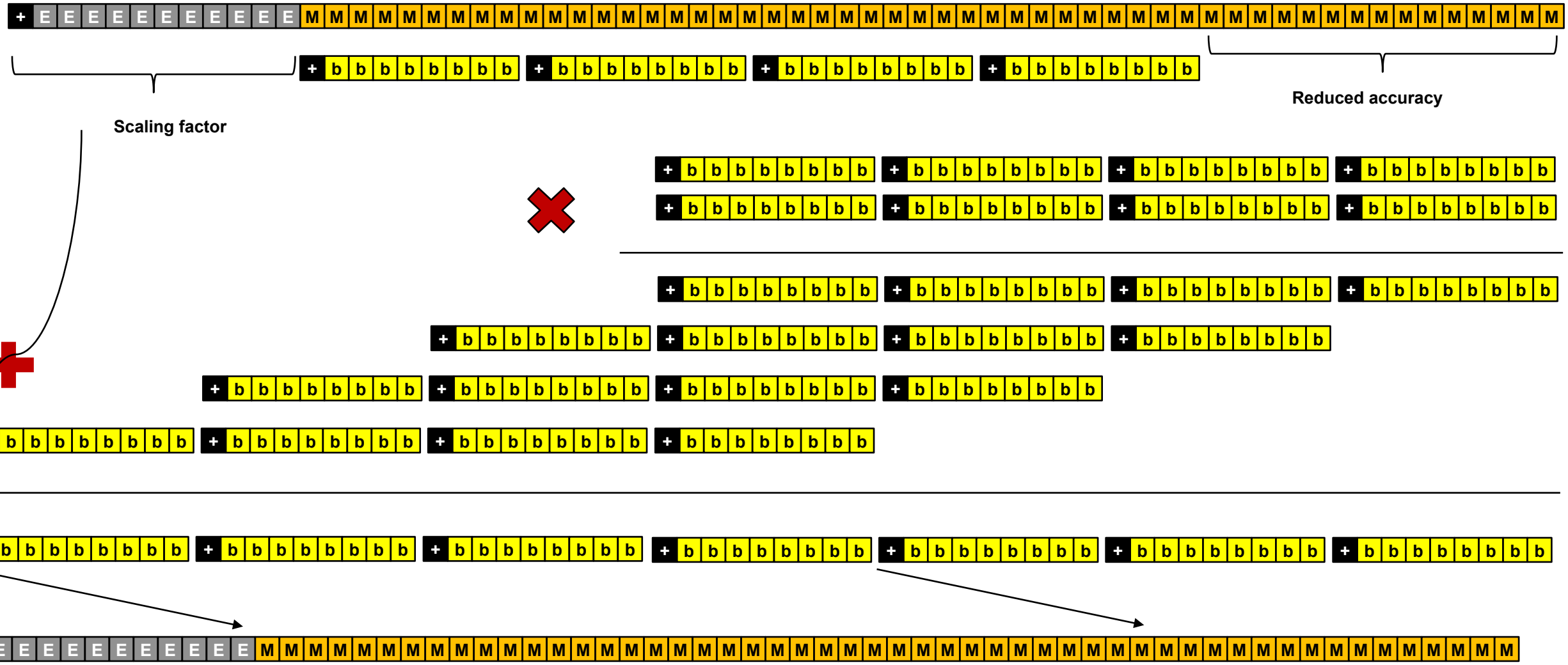


4 multiplications

```
4 * 67 = 4 * 6 * 10 + 4 * 7
3 * 67 = 3 * 6 * 10 + 3 * 7
```

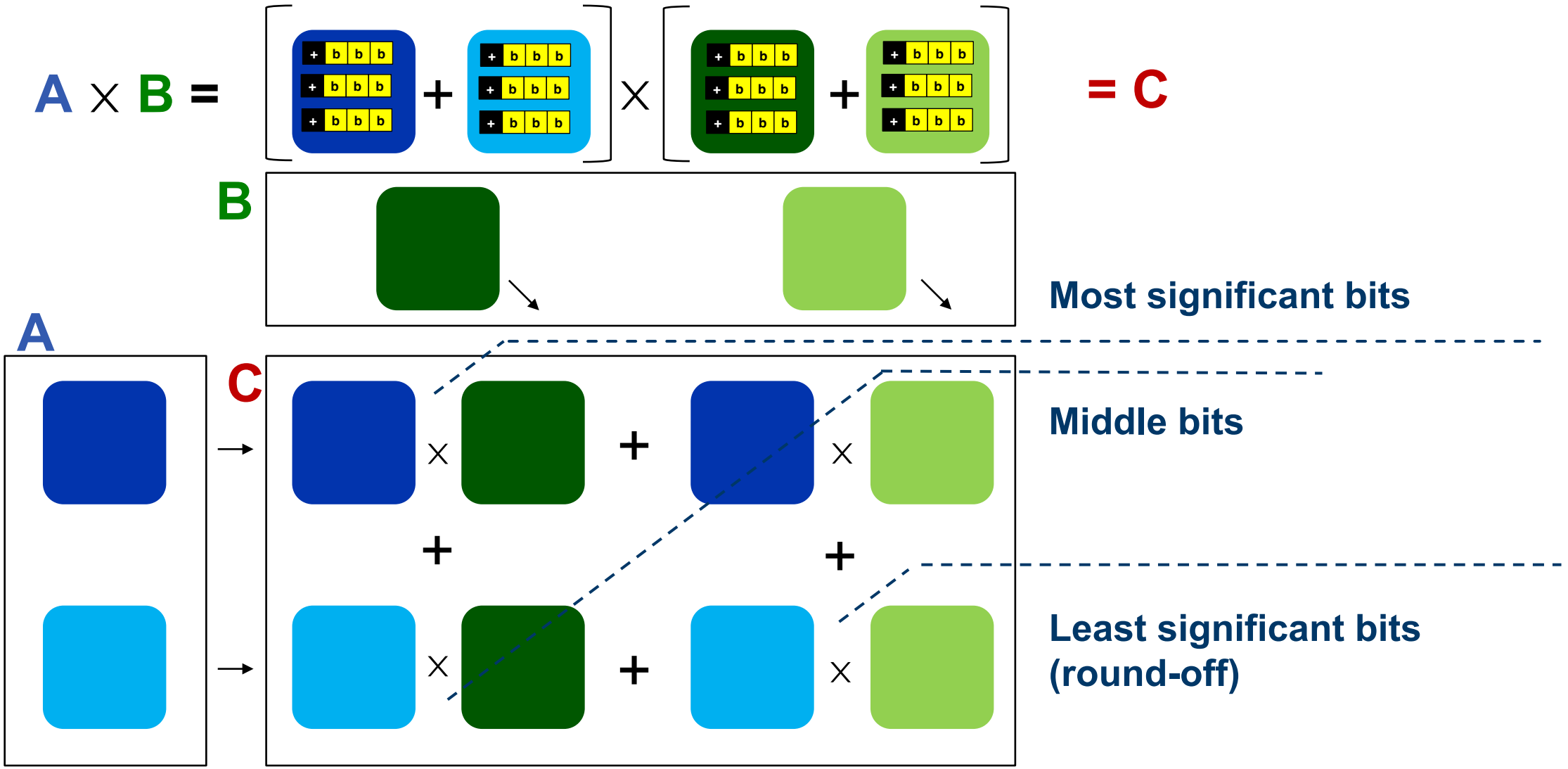
Long multiplication splits large problem into (n^2) single-digit multiplications with one-digit carry

Multiplication of Four INT8 Numbers



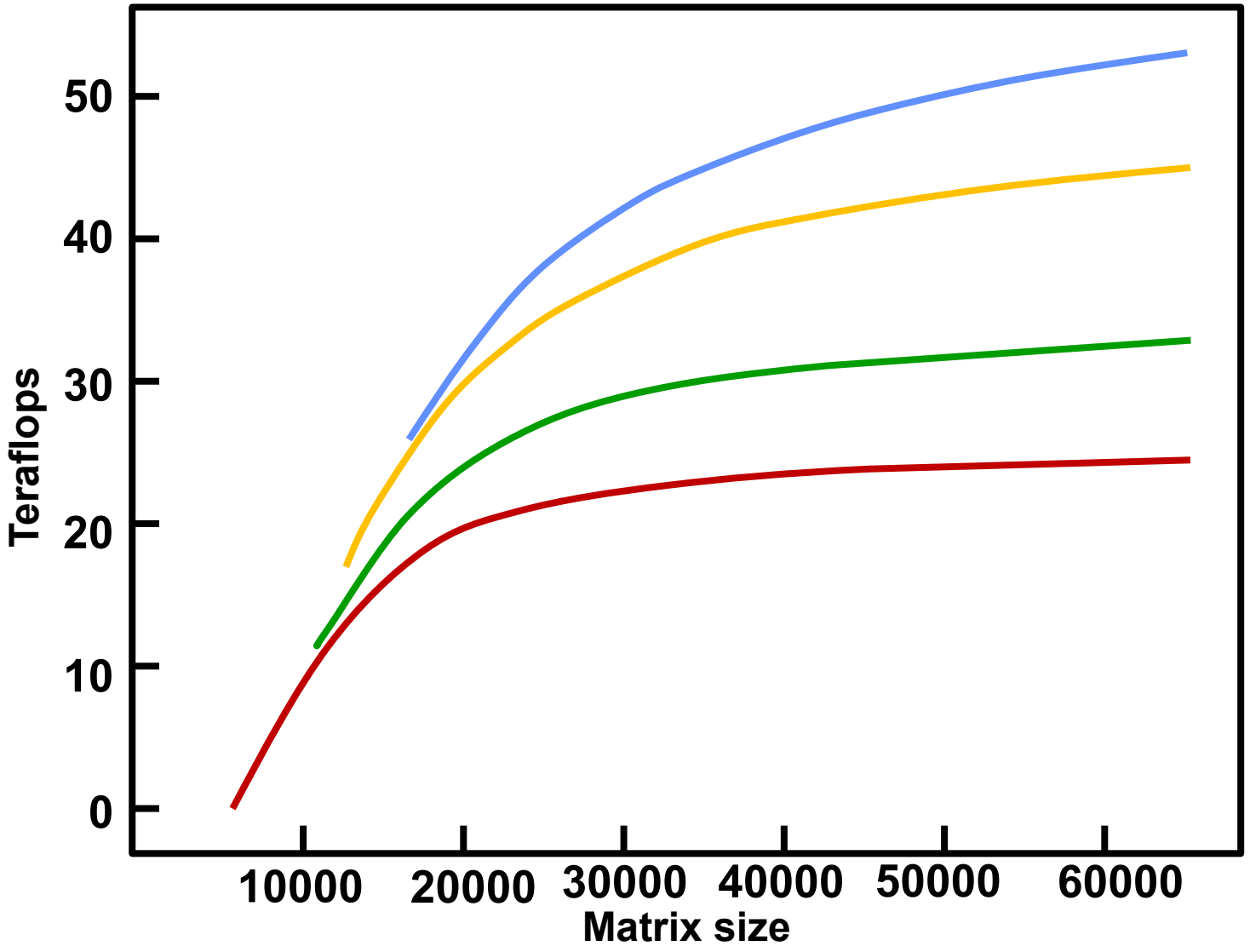
+ Sign bit of floating-point number
E Exponent bit of floating-point number
M Mantissa bit of floating-point number
b Bit of a fixed-point number

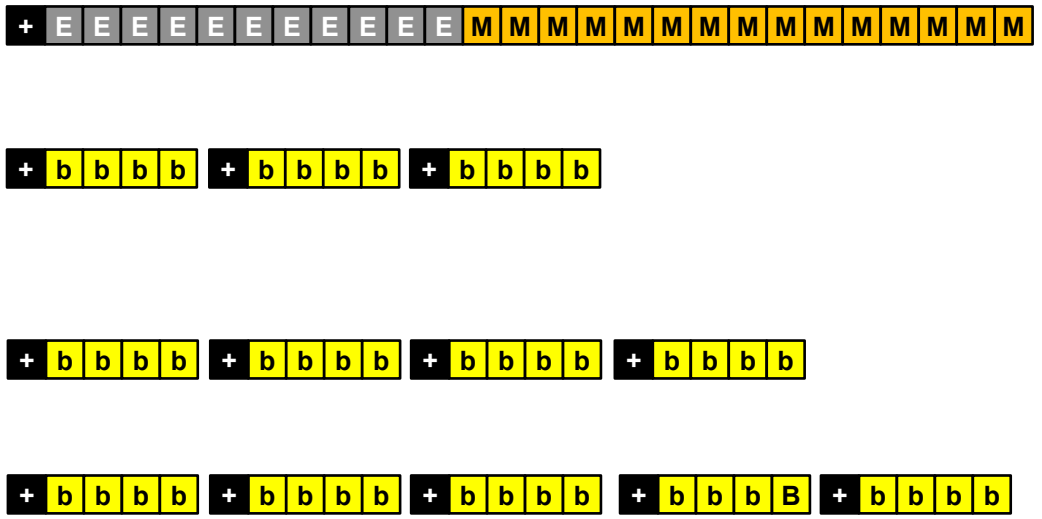
Emulation of FP64 with INT8 Ozaki Scheme: Matrices



FP64 = 64-bit floating point
INT8 = 8-bit integer

Measured Performance and Precision Matrix Multiplication on H100

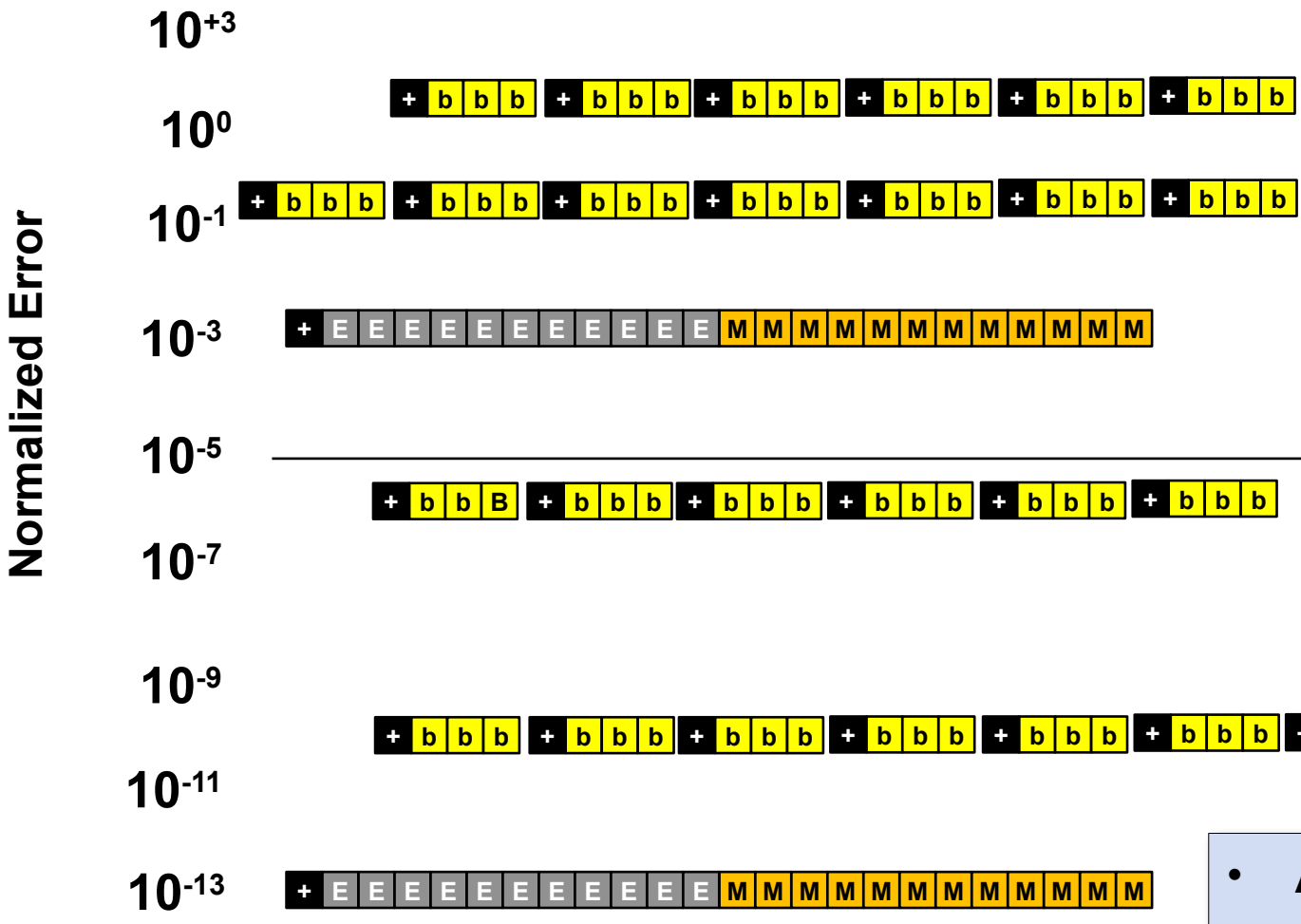




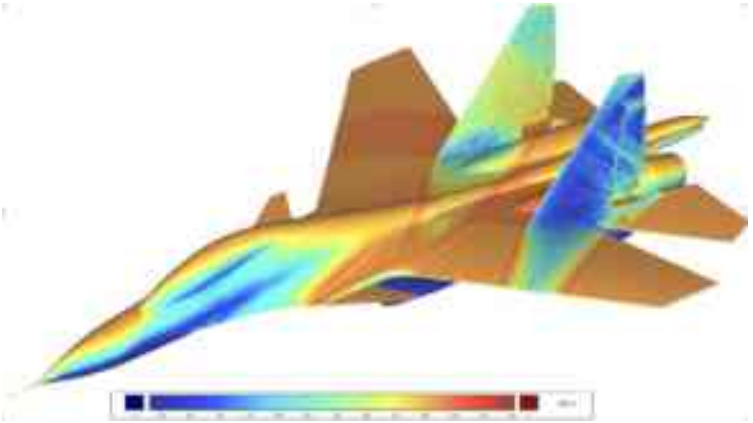
Effective use of lower precision to emulate higher precision requires performance testing over many problem sizes

+ Sign bit of floating-point number
E Exponent bit of floating-point number
M Mantissa bit of floating-point number
b Bit of a fixed-point number

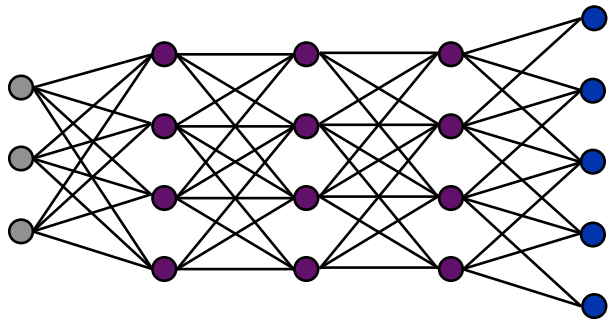
Performance/Precision Evaluation System
Linear Solve: Ax = b



High Fidelity Physics Accuracy



Machine Learning AI Accuracy



- Accuracy stable over diverse problem sizes
- Emulation with lower precision good for some applications

+ Sign bit of floating-point number
E Exponent bit of floating-point number
M Mantissa bit of floating-point number
b Bit of a fixed-point number

Controlling Element Growth: Parametric Matrix Families

1940s

- There are many matrix collections/galleries with parametric definitions of matrices suitable for solver testing
- We focus on Turing matrix from 1948
  - It's a triangular matrix with 1's on diagonal and -1's below

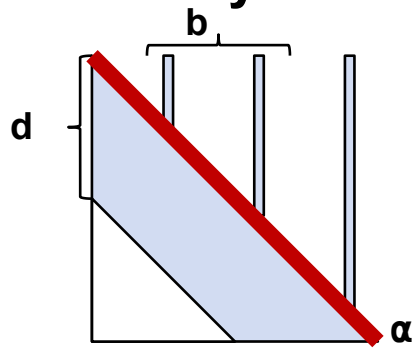
$$\begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & -1 & 1 & & \\ -1 & -1 & -1 & 1 & \\ -1 & -1 & -1 & -1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ 2 & 1 & 1 & & \\ 4 & 2 & 1 & 1 & \\ 8 & 4 & 2 & 1 & 1 \end{bmatrix}$$

1960s

- Wilkinson modified the matrix in the 1960s to generate exponential growth during the LU factorization with partial pivoting
  - The largest element in the U factor is 2^{n-1} thus requiring complete pivoting for a numerically stable factorization
  - Such matrices rarely occur in practice

2020s

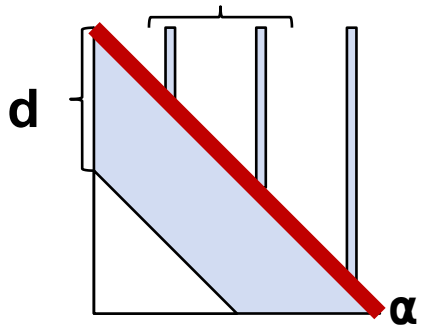
- We modify the matrix further to control the pivot growth with 3 parameters: d, b, α



$$\begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ -1 & -1 & 1 & & \\ & -1 & -1 & 1 & \\ & & -1 & -1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ 2 & 1 & 1 & & \\ 3 & 2 & 1 & 1 & \\ 5 & 3 & 2 & 1 & 1 \end{bmatrix}$$

Typical Solve Accuracy with/without Emulation

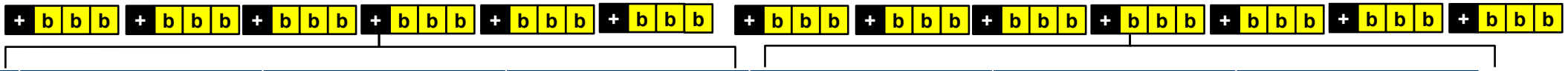
Number of INT8 splits	Scaled residual
3	26464646.4755162261
4	341348.2056036110
5	3798.4563646804
6	39.1393959359
7	0.3245150561
8	0.0030077355
9	0.0001860455
FP64	0.0002377248



d = 4
b = 15
α = 1/2

FP64 = 64-bit floating point

Detecting FP64 Emulation with Parametric Matrices



n	d	b	Scaled res.	d	b	Scaled res.
256	4	15	39.1	10	15	17.3
384	4	18	19.3	11	12	18.9
512	6	16	19.6	12	16	17.6
640	7	22	21.7	13	16	24.1
768	6	19	17.6	13	25	19.3
896	8	15	40.3	14	28	19.0
1024	8	23	22.4	15	19	16.3
1152	8	15	40.3	14	28	19.0
1280	8	16	25.8	16	30	18.9
1408	9	16	22.2	18	22	24.6
1536	9	28	27.5	17	23	26.3
1664	11	19	40.6	17	26	16.7
1792	10	23	16.5	19	23	61.6
1920	10	15	17.0	18	20	38.8
2048	11	23	20.7	19	26	19.3

Ongoing Work

- Specialize mixed-precision methods for mission applications
  - Radar beam forming
  - Image processing
  - Missile defense
- Redesign mixed-precision to match new hardware and its precision landscape
  - Explore FP4
  - Explore gaming chips (RTX line)
- Seek new fundamental algorithms amenable for mixed-precision acceleration
  - Convolution
  - FFT (Fast Fourier Transform)
  - SVD (Singular Value Decomposition)

FP4 = 4-bit floating point (sign bit, 1-bit mantissa, 2-bit exponent)

Summary

- New GPU hardware forces rethinking of fundamental algorithms
- Mixed-precision and emulation method allow new performance/accuracy trade-off
- Integrating mixed-precision algorithms has a potential to accelerate applications

Acknowledgements

- Dr. Jeremy Kepner
- Peter Michaleas
- Dr. Vijay Gadepally
- Dr. Albert Reuther
- Kristen Malvey
- Jason Williams

- Prof. Alan Edelman
- Prof. Charles Leiserson
- Prof. Sandy Pentland
- Lauren Milechen

- LaToya Anderson
- William Arcand
- William Bergeron
- David Bestor
- Alexander Bonn
- Dr. Daniel Burrill
- Dr. Chansup Byun
- Michael Houle
- Matthew Hubbell
- Dr. Hayden Jananthan

- Michael Jones
- Guillermo Morales
- Dr. Julie Mullen
- Andrew Prout
- Antonio Rosa
- Dr. Charles Yee

- Maj. Matthew Schuyler

30th IEEE HPEC Virtual Conference
September 14-18, 2026 (ieee-hpec.org)

- **Premiere conference on High Performance Extreme Computing (850+ participants in 2025)**
- **2026 Distinguished Speakers:**
  - Prof. Saman Amarasinghe (Perkins Prof EESC MIT; ACM Fellow)
  - Craig Connors (VP & CTO Cisco Infrastructure & Security)
  - Prof. Danna Freedman (Keyes Prof Chemistry MIT, Quantum@MIT Director; MacArthur Fellow)
  - Prof. Sigal Gottlieb (Chancellor Prof Math UMass Dartmouth; SIAM & AWM Fellow)
  - Dave Martinez (MIT Lincoln Lab Fellow; IEEE Fellow)
  - Prof. Daniela Rus (Viterbi Prof EESC MIT, CSAIL Director; AAAS, ACM, AAAI, IEEE, MacArthur & NAE Member/Fellow)
- **2026 Special Sessions**
  - Age of Mixed-Precision
  - Bridging Quantum and HPC
  - Large Language Models: Opportunities and Challenges
  - MIT/Amazon/IEEE Graph Challenge
  - GraphBLAS Forum
  - BRAINS: Building Resilience through Artificial Intelligence for Networked Systems
  - Scaling Research Computing Education

Paper Deadline July 7

IEEE Boston Section

Sponsors	Cooperating Societies	Technical Organizer
NVIDIA MITRE	New England Section of SIAM MIT Student Chapter of SIAM	

- **HPEC focuses on hardware, software, systems, and applications where performance matters**
- **HPEC welcomes experts and people who are new to the field**