Low Precision for Lower Energy Consumption

Preprint

Weslley Pereira,[1] Joao Silva,[1,2] and Tokey Tahmid[1,3]

1 National Renewable Energy Laboratory
2 University of Colorado, Denver
3 University of Tennessee, Knoxville

Presented at the ASCR Workshop on Energy-Efficient Computing for Science
Bethesda, Maryland
September 9-12, 2024

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-2C00-90661
August 2024

Low Precision for Lower Energy Consumption

Preprint

Weslley Pereira,[1] Joao Silva,[1,2] and Tokey Tahmid[1,3]

1 National Renewable Energy Laboratory
2 University of Colorado, Denver
3 University of Tennessee, Knoxville

Suggested Citation
Pereira, Weslley, Joao Silva, and Tokey Tahmid. 2024. *Low Precision for Lower Energy Consumption: Preprint*. Golden, CO: National Renewable Energy Laboratory. NREL/CP-2C00-90661. https://www.nrel.gov/docs/fy24osti/90661.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Conference Paper
NREL/CP-2C00-90661
August 2024

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the National Renewable Energy Laboratory. The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

U.S. Department of Energy (DOE) reports produced after 1991 and a growing number of pre-1991 documents are available free via www.OSTI.gov.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

NREL prints on paper that contains recycled content.

Low precision for lower energy consumption

2024 ASCR Workshop on Energy-Efficient Computing for Science

Weslley Pereira[1] João Silva[1,2] Tokey Tahmid[1,3]
wdasilv@nrel.gov

[1]Computational Science Center, National Renewable Energy Laboratory
[2]Department of Mathematical and Statistical Sciences, University of Colorado Denver
[3]Electrical Engineering and Computer Science, University of Tennessee Knoxville

Topic Due to finite memory, all computer software run at finite precision. The most well-known numeric types are integers, the IEEE 754 single-precision (FP32) and double-precision floating-point (FP64) types. Integers are represented with 32 or 64 bits in modern 64-bit architectures, while FP32 and FP64 always have 32 or 64 bits, respectively. Based on this usual scenario, we consider low-precision types to be the ones whose size varies between 1 bit and 31 bits. Some examples of low-precision floating-point types are: TF32 (19 bits) [1], FP16, BF16 [2], E4M3 (8 bits) and E5M2 (8 bits) [6].

Hardware operations like moving bits and adding numbers use less bits when applied to low-precision types. Thus, hardware instructions operating on low-precision types consume less energy and are faster [5], and may even halve memory requirements in recent GPUs [7]. Mixed-precision operations and algorithms use two or more arithmetic precision types. In combination with low-precision types, mixed-precision can achieve high efficiency with minor or no accuracy loss for sufficiently large problems. One example is the training of Neural Networks, where mixed precision approaches have shown to retain high accuracy through the following techniques: (I) maintaining a separate copy of weights in FP32 for weight and gradient updates, (II) scaling of the loss function to prevent underflow, and (III) performing FP16 arithmetic for computations with accumulating values in FP32 (See details in [7]). Another example is the solution of linear systems, in which a common mixed-precision approach is to first solve the system in lower precision, and then use iterative refinement to recover a highly accurate solution [5].

Several energy management tools have been developed to measure energy consumption of software applications [3]. These tools may utilize hardware-specific features such as Intel RAPL (Running Average Power Limit) for obtaining CPU and memory energy consumption, and use NVIDIA Management Library (NVML) for GPU energy consumption. The total energy consumption of a program can be calculated by multiplying the Power Usage Efficiency (PUE) of the machine with the average power and time for each device. Another useful metric is the Energy Delay Product (EDP), obtained by multiplying the total consumed energy by the overall time, providing a measure of energy efficiency considering both energy usage and performance.

Challenge Mixed precision algorithms have been historically used to solve scientific problems, dating back from linear system solvers from 1941 [5]. Since the introduction of the IEEE 754 standard and popularity of 64-bit architectures, most algorithms were designed to use 32-bit and

[1]`https://blogs.nvidia.com/blog/tensorfloat-32-precision-format/`
[2]`https://www.nextplatform.com/2018/05/10/tearing-apart-googles-tpu-3-0-ai-coprocessor/`

64-bit numeric types. Recently, mainly motivated by the acceleration of Machine Learning, mixed precision instructions and new low-precision types have emerged. Adapting existing algorithms to these types is crucial for maintaining accuracy, such as in the case of the QR decomposition, which relies on numerically accurate orthogonal bases [5]. Yet, the ongoing standardization and study of new types such as E4M3 and E5M2 present challenges for algorithm developers seeking to propose mixed-precision solutions [6].

Despite hardware and algorithm support for low-precision types, gaps in software persist. For example, TensorFlow's Probability module lacks mixed-precision support, although the main library supports it [3], and PyTorch does not provide wrappers for the mixed-precision symmetric linear system solver from MAGMA [1]. Those gaps add a layer of difficulty for applications to use state-of-the-art mixed-precision implementations. Additionally, there is a shortage of performance benchmarks, particularly in energy measurement. Recent benchmarks show up to a 1.9X speedup using mixed precision for training and inference of a U-Net with 64 filters across CPU, GPU, and TPU, but lack energy efficiency metrics, e.g., [4].

Opportunity Based on the challenges described above, it is important to systematically verify that existing mixed-precision algorithms work on new architectures and low-precision types. Moreover, there is a need for software that can fill current gaps in the application stack, enabling the use of state-of-the-art mixed-precision implementations. Lastly, current and new performance benchmarks should include energy measurements to enhance understanding of mixed-precision algorithms, including the scenarios and configurations where they are most beneficial.

Maturity Due to the favorable use of low precision in Machine Learning algorithms, mixed-precision instructions have been integrated into many modern hardware architectures. Examples include Nvidia Tensor Cores (2017), Intel AMX (2023), and Apple Metal mixed precision (2023) technologies, and the recent hardware support for E4M3 and E5M2 (2022) [6]. Additionally, the rapid advancements in Artificial Intelligence (AI) are making it increasingly costly due to rising energy consumption. For instance, GPT-3 has 175 billion parameters and was estimated to consume 1287 MWh for training and deploying. Other estimates suggest that in the near future Nvidia's new AI servers will consume more energy than Argentina and Sweden annually [2].

References

[1] A. Abdelfattah, S. Tomov, and J. Dongarra. Investigating the Benefit of FP16-Enabled Mixed-Precision Solvers for Symmetric Positive Definite Matrices Using GPUs. In *ICCS*, pages 237–250. Springer International Publishing, 2020.

[2] Y. I. Alzoubi and A. Mishra. Green artificial intelligence initiatives: Potentials and challenges. *J. Clean. Prod.*, 468:143090, 2024.

[3] L. Bouza, A. Bugeau, and L. Lannelongue. How to estimate carbon footprint when training deep learning models? A guide and review. *Environ. Res. Commun.*, 5(11):115014, 2023.

[4] M. Dörrich, M. Fan, and A. Kist. Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks. *IEEE Access*, 11:57627–57634, 2023.

[5] N. J. Higham and T. Mary. Mixed precision algorithms in numerical linear algebra. *Acta Numerica*, 31:347–414, 2022.

[6] IEEE. Working Group P3109 Interim Report. Technical report, IEEE, 2024. v0.7.0.

[7] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, and H. Wu. Mixed precision training. In *ICLR*, 2018.

[3] `https://github.com/tensorflow/probability/issues/1315`