

Microscaling Floating Point Formats for Large Language Models

Marco Cococcioni^{1*}, Dario Pagani¹, and Federico Rossi¹

Abstract—The increasing computational and memory demands of large language models (LLMs) necessitate innovative approaches to optimize resource usage without compromising performance. This paper leverages microscaling floating-point formats, a novel technique designed to address these challenges by reducing the storage and computational overhead associated with numerical representations in LLMs. Unlike traditional floating-point representations that allocate a dedicated scale for each value, microscaling employs a shared scale across a block of values, enabling compact one-byte floating-point representations while maintaining an extended dynamic range. We explore the application of microscaling in the context of 8-bit floating-point formats to significantly reduce memory footprint and computational costs. We tested several configurations of microscaling floats within the GPT-2 LLM architecture, demonstrating that microscaling data formats can achieve competitive accuracy during training and inference, proving its efficacy as a resource-efficient alternative for deploying LLMs at scale. The source code is publicly available at: <https://github.com/unipi-dii-compressedarith/llm.c-sve>. C++23 or a higher standard is required to compile it.

Index Terms—Microscaling, floating-point, C++, LLM, mini-float, IEEE 754 standard.

I. INTRODUCTION

The rapid advancement in the size and complexity of large language models (LLMs) has revolutionized applications in natural language processing and beyond. However, the increasing computational and memory demands associated with these models present substantial challenges, particularly in terms of scalability and cost-efficiency. For instance, models such as GPT-2 (Generative Pre-Trained Transformer, version 2) and its successors have highlighted the need for innovative strategies to reduce resource consumption without sacrificing performance [1, 2].

A crucial aspect of this challenge lies in the numerical representation of model parameters, which directly impacts storage and computational requirements. Traditional floating-point formats, including half-precision and BFLOAT16, have been widely adopted in deep learning due to their balance between precision and efficiency [3]. However, as models grow larger, these formats become increasingly insufficient, prompting the exploration of alternative approaches.

One promising solution is Microscaling (MX) floating-point formats, a technique that leverages shared scaling across blocks of values to enable compact numerical representations.

¹Marco Cococcioni, Dario Pagani, Federico Rossi, are with the University of Pisa, Department of Information Engineering, Pisa, Italy (e-mail: marco.cococcioni@unipi.it, dario.pagani.146@gmail.com, federico.rossi@ing.unipi.it)

*corresponding author

Unlike conventional formats, which allocate individual exponents to each value, Microscaling reduces memory usage by representing values with just one byte and applying a shared exponent. This approach allows for significant savings in memory and computation while maintaining a dynamic range suitable for deep learning tasks [1].

This study focuses on 8-bit floating-point formats within the Microscaling framework, a choice motivated by recent advancements in hardware support for low-bit representations. By reducing the memory footprint and computational overhead of LLMs, Microscaling has the potential to enable more efficient training and deployment, particularly in resource-constrained environments. This research aims to evaluate the trade-offs between accuracy and efficiency inherent in Microscaling, contributing to the broader goal of sustainable and scalable AI development.

A. Main novelties of the work and obtained insights

In this work we provide a fast and modern C++ implementation of a generative pretrained transformer, equipped, for the first time, with the recently introduced microscaling format.

In addition, we introduce important mitigations needed to accurately train the GPT model using a mixed-precision approach, where different numeric formats are used in the appropriate context, to deliver an acceptable level of performance while reaching a lower bandwidth consumption and memory footprint.

The main obtained insights, that can be useful to other researcher and practitioners working in this domain, are:

- 1) The operation of tensor transposition is not exactly trivial in Microscaling; while this may seem self-evident in retrospect, it is not entirely trivial to discern at the beginning
- 2) When rounding numbers whose significands are stored on a tiny number of bits it is vital to use an appropriate rounding policy instead of relying on the faster truncation
- 3) The importance of correctly ordering the operations on low precision formats to reduce numerical error

B. Paper organization

This paper is organized as follows: Section II provides an overview of related works, discussing existing numerical representations and their impact on deep learning efficiency; this is followed by a detailed explanation of Microscaling formats in Section III, including their theoretical foundation

and implementation details; Section IV and V present the implementation details of Microscaling within the LLM.c GPT-2 library, outlining the datasets, models, and evaluation metrics used; Section VI ("Evaluation and results") analyzes the performance of Microscaling, focusing on memory efficiency, computational throughput, and accuracy; finally, Section VII, devoted to conclusions, highlights the implications of these findings, and suggests directions for future research.

II. RELATED WORKS

The challenge of optimizing numerical representations in machine learning has received significant attention in recent years. Traditional approaches such as single and half-precision floating-point formats have been instrumental in enabling the rapid growth of deep learning. BFLOAT16, in particular, has been widely adopted for its compatibility with existing hardware and its ability to balance precision with computational efficiency [3].

Emerging research has focused on reducing the bit-width of numerical representations further, with 8-bit floating-point formats gaining traction due to their promise of improved memory efficiency. Hardware vendors have begun incorporating support for these formats, paving the way for their adoption in large-scale AI systems [1].

Another line of work explores alternative architectures and methods to enhance computational efficiency in LLMs. The introduction of attention mechanisms, as demonstrated in Transformer-based models, revolutionized deep learning by offering a scalable approach to sequence modeling [4]. However, these advancements also exacerbated the resource demands of training and inference, highlighting the need for complementary techniques like Microscaling to address the bottlenecks associated with memory and computation.

The concept of shared scaling, as implemented in Microscaling formats, represents a novel approach to numerical representation. It extends the principles of block quantization by introducing a shared exponent, enabling finer-grained memory optimization without compromising the model's ability to capture complex patterns [1]. This study builds on these foundations, aiming to further validate the effectiveness of Microscaling in large language models.

III. MICROSCALING FLOATING-POINT FORMATS

Floating point arithmetic uses integral numbers to represent a subset of rational numbers and is a way used to represent real numbers in the computer [5]. With such an approach, a number is represented by three fields: the sign bit, the exponent and the significand (also known as the mantissa).

In addition to rational numbers, these formats can represent the special values of

- NaN, typically used to signal an error in a procedure, for instance the computation of the square root of a negative number $\sqrt{-1}$;
- two signed infinities $\pm\infty$, typically used to signal an overflow in the computation, for instance $\ln(+0)$;
- and two signed zeros ± 0 .

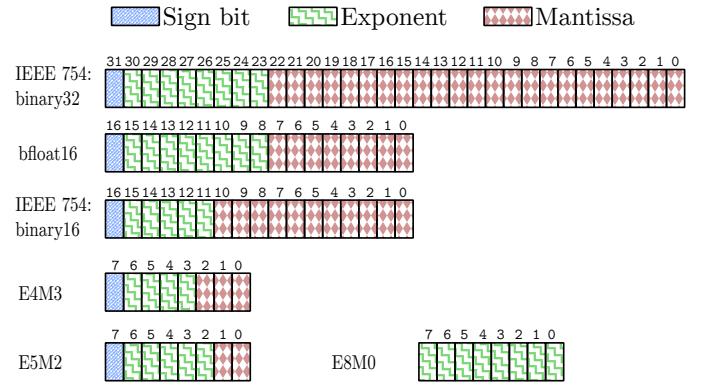


Fig. 1. Comparison between some floating point formats

The Microscaling format aims to overcome the loss of dynamic range when using 8 bits data formats to represent real numbers — such as E4M3 and E5M2 — this is done by organizing a given set of numbers into small blocks and associating to each block a scale — also called weight, — each block element is then scaled, converted and stored using a small data format, typically on 8 bits [1, 6].

A. Software-defined floating points

Since it is uncommon for present-day hardware to have support for eight bit floating-point numerical formats, we had to make use of software emulation. In particular we used a modified version of the `anyfloat` class from the `cpposit` library [7]

`Anyfloat` is a template class that allows the programmer to define custom floating-point numeric types with an arbitrary number of bits for their significands and exponents, most operations are performed by converting the number back and forth between the custom representation and the native type handled by the machine's FPU — such as single precision, double precision or even half-precision floats on certain newer machines.

$$x = \begin{cases} -1^s \cdot \infty & \text{if } k = 2^E - 1 \wedge m = 0 \\ \text{NaN} & \text{if } k = 2^E - 1 \wedge m \neq 0 \\ \pm 0 & \text{if } k = 0 \wedge m = 0 \\ -1^s \cdot 2^{-b+1} \cdot \frac{m}{2^M - 1} & \text{if } k = 0 \wedge \text{denorm enabled} \\ -1^s \cdot 2^{k-b} \cdot \left(1 + \frac{m}{2^M - 1}\right) & \text{otherwise} \end{cases} \quad (1)$$

Equation 1 describes how to compute the represented number starting from the values of the three fields, sign, exponent and mantissa; read as unsigned integral numbers, noted as s , k and m respectively and stored on 1, E and M bits respectively. The constant b , called the exponent's bias, is computed as $b = 2^{E-1} - 1$. This format is similar and binary compatible to the one defined by the IEEE 754 standard. Support for denormalized numbers (also known as subnormal numbers) can be enabled or disabled by the programmer when defining the custom data type.

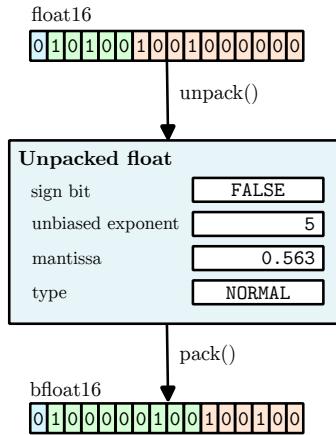


Fig. 2. Anyfloat makes use of a support data structure, called Unpacked, to represent the unpacked number

B. The Microscaling data-format

The Microscaling format aims to overcome the loss of dynamic range when using 8 bits data formats to represent real numbers — such as E4M3 and E5M2 — this is done by diving a list of numbers into small chunks and associating to each chunk a *shared scale*, each chunk's *element* is then scaled, converted and stored using a small data format, typically on 8 bits [1, 6].

Interestingly, it is possible to perform certain mathematical operations without the need to decompress the quantized values; most importantly, the dot-product between two Microscaling vectors can be computed by firstly computing the unscaled intra-block dot product, scaling it properly and finally accumulating the results to compute the inter-block dot product [6].

Let us call $x \in \mathbb{M}_{OP}^n$ the vector of the original unquantized values, $c \in \mathbb{M}_{RP}^n$ the vector of the quantized values (where \mathbb{M}_{OP} and \mathbb{M}_{RP} denote the set of machine numbers in the original and reduced precision respectively), B the block length, $l = \lceil \frac{n}{B} \rceil$ the number of Microscaling blocks, $\forall i \in \{1, \dots, l\}$ $s_i = 2^{w_i} \wedge w_i \in \mathbb{Z}$ the blocks' shared scales; then we can rewrite the dot-product between two vectors x^a, x^b as shown in Equation 2:

$$\begin{aligned} y &= x^a \cdot x^b = \sum_{j=1}^n x_i^a \cdot x_i^b \\ y &\simeq \sum_{j=1}^l \sum_{i=1}^B s_j^a c_i^a \cdot s_j^b c_i^b \\ y &\simeq \sum_{j=1}^l s_j^a s_j^b \sum_{i=1}^B c_i^a \cdot c_i^b \end{aligned} \quad (2)$$

C. Microscaling implementation

We implemented the aforementioned Microscaling data structure by defining a template class `microscaling::vector` with parameters to specify which data format has to be used to store the elements and

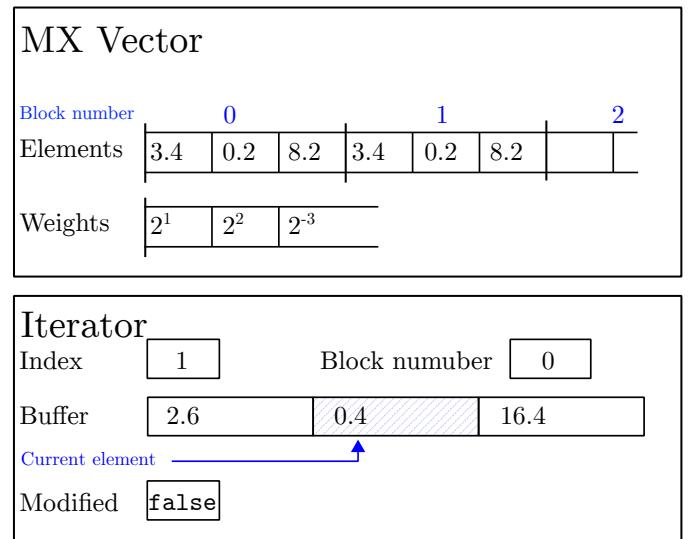


Fig. 3. Example of an MX vector with block size 3 and an iterator instance with the auto-commit feature enabled

the scales, how long a block shall be and if the auto-commit feature is enabled.

Our implementation allows the use of a large variety of numerical formats to represent the quantized values, for instance: natives floating-point types, emulated ones, arbitrary-precision and even novel formats such as Posit.

a) *Data organization*: The container stores the data on the heap in two `std::vectors`, one for the scales and one for the scaled numbers. To retrieve the original number it is thus sufficient to multiply the scaled number with its block's scale. Equation 3 shows the decompression operation.

$$x_i \simeq \begin{cases} c_i \cdot s_{\lfloor \frac{i}{B} \rfloor} & \text{if } s_{\lfloor \frac{i}{B} \rfloor} \text{ is not NaN} \\ \text{NaN} & \text{otherwise} \end{cases} \quad (3)$$

It is possible to randomly access data in the vector through the method `float operator[](int)`.

b) *Iterator*: To improve access performance during sequential access as well as providing writing access to the data structure, we implemented an *STL*-like iterator. The Iterator data structure, as shown in Figure 3, consists of a buffer that contains the block's uncompressed data.

The buffer is built each time a call to `operator++()` or `operator+(int)` moves the iterator to a new block or a new iterator is constructed from scratch — such as with a call to `begin()` on the vector.

The iterator class provides the methods `refresh()` and `commit()` to re-build the buffer and to write any changes to the buffer back to the main data structure.

c) *Writeable iterators*: We also implemented an optional auto-commit feature, that is the iterator instance will keep track of any changes to the buffer during its lifespan and, if necessary, will automatically make a commit to the main data structure; notice that the mechanism does not provide any kind of synchronization between multiple iterators operating on the same data structure, it is always the programmer's responsibility to keep the main vector coherent by avoid

The problem arises when the program has to compute the loss's gradients during the backward pass as it is not possible to perform linear scans over the matrices and is thus necessary to perform random accesses to read the input and output derivatives' matrices by column instead of by row for the former and vice-versa for the weights' matrix; this becomes a critical issue when we want to use Microscaling formats to store the matrices because random accesses are more expansive than linear scans and we cannot make use of the faster implementation of MX's dot production and, even worse, using an MX vector as a vector of accumulators during multiple passes, leads to a great increase of the numeric error, making the network unusable.

We already have to build an MX vector on-the-fly, so we may as well build it already transposed. By doing so we are able to use the MX's dot product to compute the loss derivative in the weights' and inputs' directions, that is we can compute the output derivatives' matrices by performing only forward passes, greatly reducing the numerical error. To transpose the matrix we made use of C++20's views and ranges to create an iterable view that can be fed into the MX vector's constructor. Its implementation looks something like the following.

```
template<class container_t>
auto transpose(const typename container_t::iterator
    &matrix, unsigned rows, unsigned cols) {
    return std::views::iota(0u, rows * cols) |
        std::views::transform([matrix, rows, cols] (
            &auto index) {
            auto row = index / rows;
            auto col = index % rows;
            return matrix[col * cols + row];
        });
}
```

D. Full-precision probabilities and encodings

An issue we encountered while experimenting with lower precision formats was probabilities of certain words in the dictionary begin approximated to zero, this caused their relative losses to diverge to infinity, causing the average loss to be infinity. In literature there are several methods to deal with such an issue, but we found that — at least for the GPT-2 network — is sufficient to store the probabilities and the final losses in full-precision.

A similar issue was encountered for the tokens' encodings, mainly because they saturated the numeric data-type — so they became $+\infty$ or the maximum number according to the rounding policy — or vice-versa were rounded to zero; again this issue can be eliminated by storing such values always in full-precision.

Of course this problem arises when we are quantizing an already-trained model to numerical formats with lower precision; most likely this wouldn't be an issue with models that are trained from scratch.

V. FLOATING-POINT ARITHMETIC

Software-defined floating-point types are really slow, here we present modification we made to the Anyfloat library to speed-up the computation.

A. Compile-time expressions

Anyfloat declares almost all of its methods as `constexpr`, that is as methods that can be evaluated at compile-time, unfortunately the original implementation made a chain of calls that terminated into a non-`constexpr` procedure, the `unpack_xfloat` method from the `Unpacked` class, that was implemented using expressions non evaluable by the compiler:

```
Unpacked& unpack_xfloat(typename Trait::value_t
    &value) {
    union {
        typename Trait::holder_t i;
        typename Trait::value_t f;
    } uu;
    uu.f = value;
    return unpack_xfloati<Trait>(uu.i);
}
```

Since C++20, the standard library implements `std::bit_cast()`, a `constexpr` function that reinterprets the binary representation of an object into an another [9], effectively having the same behavior as a C's union but evaluable at compile-time, the modified code becomes:

```
constexpr Unpacked& unpack_xfloat(typename Trait::
    &value_t value) {
    auto i = std::bit_cast<typename Trait::holder_t>(
        &value);
    return unpack_xfloati<Trait>(i);
}
```

This allows expressions such as `E5M2 x = 3.0f;` to be evaluated by the compiler itself while the project is being built.

Analogous modification was made to the `pack_xfloati`, used to re-pack floating-point numbers.

B. Rounding

Originally, Anyfloat, when converting from a high-precision format, used to truncate the longer significand. We observed empirically, especially when comparing the C++23's `std::bfloating16_t`, as implemented by the GNU C Compiler, that the error of such an approximation was higher than rounding toward the nearest representable value.

The former policy, used by C++23's `std::bfloating16_t` and by IEEE's floating-point formats, is called *round to the nearest, with ties to even*; whereas the latter, used by our implementation, is called *round to nearest, ties to away (from zero)*. Both policies are described by the IEEE standard, the former is mandatory and has to be the default rounding policy; whereas the latter is optional [10, Section 4.3].

Notice that the Open Compute Project's specification for the Microscaling data-format requires the implementation of the *round to the nearest, with ties to even* policy [6]; however we think the simpler *round to nearest, ties to away (from zero)* is adequate to test the overall architecture and it's certainly much better than a simple truncation, as it is shown by our experiments.

We ran the learning program over one hundred iterations and compared the average error between using full-precision, C++23's `std::bfloating16_t` and Anyfloat's `bfloating16` with truncation and rounding-to-nearest. Figure 5 shows the loss function between the two rounding policies, while Table I

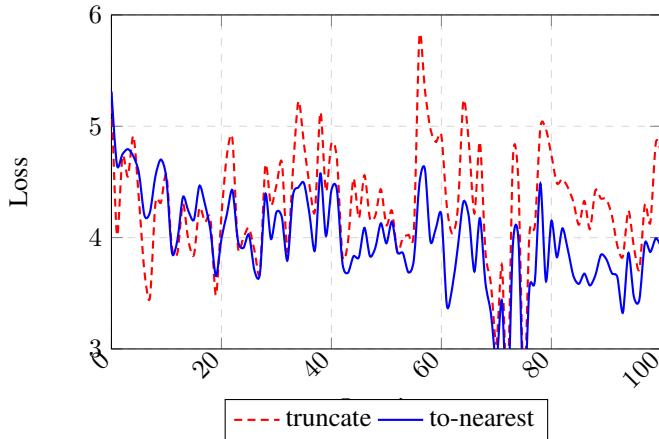


Fig. 5. Comparison of the loss function with two different rounding policies

shows the relative differences of the two against the full-precision baseline.

TABLE I
DIFFERENCE IN LOSS OVER 100 STEPS OF TRAINING

	baseline	std::bfloating16	truncation	to-nearest
Average loss	3.877	3.9889	4.3023	3.9874
Difference		0.1120	0.4254	0.1105
R. difference		2.89%	10.97%	2.85%

We reduced greatly the error and empirically proved the correctness of our emulated float implementation against GNU's.

C. Softmax

One of the steps needed to compute the self-attention is the evaluation of the *softmax* function, the original implementation, described in Algorithm 2, firstly computes and stores the exponential function for all elements while computing the sum of the resulting values, then it normalizes by the sum by re-scanning the values once more.

Algorithm 2 The attention's softmax

Input: array a of size t , maxval the maximum of a
Output: normalized array

```

 $s \leftarrow 0$             $\triangleright$  Calculate the exp and keep track of sum
 $\text{for } i \leftarrow 0 \text{ to } t \text{ do}$ 
     $e \leftarrow \exp(a[i] - \text{maxval})$ 
     $s \leftarrow s + e$ 
     $v[i] \leftarrow e$ 
 $\text{end for}$ 
 $s' \leftarrow \begin{cases} 0 & \text{if } s = 0 \\ \frac{1}{s} & \text{otherwise} \end{cases}$ 
 $\triangleright$  Normalize to get the softmax
 $\text{for } i \leftarrow 0 \text{ to } t \text{ do}$ 
     $v[i] \leftarrow v[i] \cdot s'$ 
 $\text{end for}$ 

```

When using full-precision such an operation presents no drawbacks as their dynamic range is more than sufficient to represent the largest value of the exponential function one

might encounter in this kind of application; however when using formats with a reduced dynamic range, such as the eight bit formats used in Microscaling, many of the results of the exponential function will undoubtedly cause an overflow causing an erroneous computation.

To solve this issue, we modified the implementation of the attention's softmax to iterate two times over the values, one to compute the exponential function on-the-fly — that is without storing the result in memory — just to accumulate all its values into an accumulator, and a second one to compute the exponential function divided by the total sum.

While the computational complexity remains in the order of $\Theta(t)$, the machine now has to compute two times the same exponential function; this might negatively affect larger models, even though the matrix multiplication remains — without any doubt — the biggest bottleneck in increasing the size of these models.

D. Look-up tables

To speed-up such operations, we added optional support for look-up tables (LUTs) to the Anyfloat class. When enabled, it replaces certain operations with one or two reads to the appropriate look-up tables, thus greatly reducing the number of CPU cycles required to perform the computation. We also modified the binary operators to perform these operations in a different output space, instead of being a closed operation, typically with higher precision — e.g., binary operations in E5M2 promote to brain float 16 bit — as this is necessary to perform the dot product and it is where most of the these heavy operations are performed; if need be, the result can be converted back to the original space, in such scenario we still perform less conversions, originally three conversions were performed and now only one.

1) *Constant expression*: While it is technically possible to compute and store the look-up tables as `constexpr` objects, there are two main drawbacks in doing so:

- while populating the tables we would have to handle operations that generate NaNs or exceptions — such as $0 \cdot \infty$ — separately as it is not possible to compute them at compilation time, as their behavior depends on runtime variables (The compiler throws an error if said operations are encountered in a `constexpr`)
- the compiler would have to create the same tables multiple times, a copy for each compilation unit, possibly hindering the compilation speed.

we decided to keep the tables as a static constant member to be built during the process start-up. One downside of our approach is a very likely *static initialization order fiasco* when defining other static variables that depend on the look-up tables being already populated; however, such a scenario is not present in our project.

2) *On the complexity of operations*: Let us call b_i the number of bits used to represent a number used as input operand and b_o the number of bits used to represent the computation resulting number and M^i , M^o the space of floating-point numbers used as input and output respectively. Contrary to the common practise — that is $x \bowtie \text{NaN}$ is always

false with any binary operator $\bowtie = <, \leq, \geq, >, \dots$ — we consider all positive NaNs part of $\mathbb{M}_{\geq 0}^i$ and all negative NaNs part of $\mathbb{M}_{\leq 0}^i$.

We introduced the following tables:

a) *Inversion table*: It defines a function $\mathbf{L}_i : \mathbb{M}_{\geq 0}^i \rightarrow \mathbb{M}_{\geq 0}^i$, that is used to quickly compute the reciprocal $\frac{1}{x}$. We do not need to store all the input space as the function is odd in its domain. Its space complexity is

$$S_{\mathbf{L}_i} \in \Theta(b_i \cdot 2^{b_i}) \quad (4)$$

for a 8 bit number, such as the E5M2 format, the space occupation the table is $2^7 * 1\text{byte} = 128\text{byte}$.

b) *Multiplication table*: It defines a function $\mathbf{L}_m : \mathbb{M}_{\geq 0}^i \times \mathbb{M}_{\geq 0}^i \rightarrow \mathbb{M}_{\geq 0}^o$, that is used to compute $z = x \cdot y$. We don't have to store all the input space as the function is even when $\text{sign}(x) = \text{sign}(y)$ and odd otherwise, so we perform the look-up using the absolute values and set the sign-bit accordingly. Its space complexity is

$$S_{\mathbf{L}_m} \in \Theta(b_o \cdot 2^{b_i} \cdot 2^{b_i}) \quad (5)$$

using the brain-float16 data format for the output, for an 8 bit number, the space occupation is $2^7 \cdot 2^7 \cdot 2\text{byte} = 32\text{kibibyte}$.

We could further reduce the space occupation by just storing the triangular matrix of the original look-up table and performing the look-up by imposing $x \geq y$, under this assumption the space occupation could be reduced to $\binom{2^7}{2} \cdot 2\text{byte} = 15.875\text{kibibyte}$.

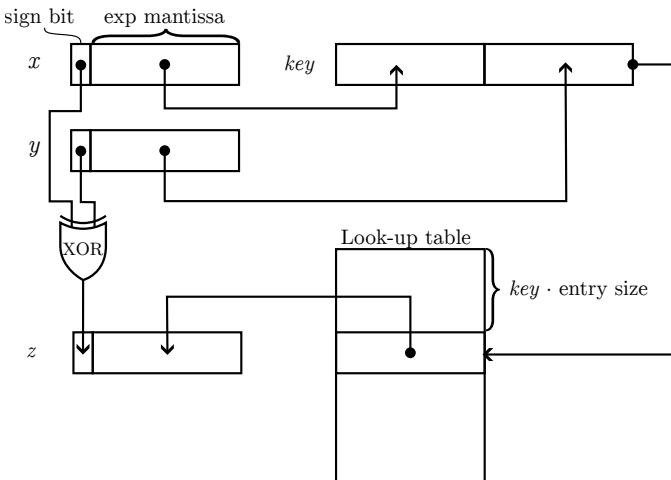


Fig. 6. How a multiplication is performed using a look-up table

c) *Addition table*: It defines a function $\mathbf{L}_a : \mathbb{M}_{\geq 0}^i \times \mathbb{M}_{\geq 0}^i \rightarrow \mathbb{M}^o$, that is used to compute $x = a + b$. To save space we don't save the negative values for one addendum, that is we can look-up the result of an addition only if at least one of the operands is not negative; it is quite easy to see that we can still compute the addition for all possible values with the relation:

$$a + b = \begin{cases} -(-a \oplus (-b)) & \text{if } a < 0 \wedge b < 0 \\ b \oplus a & \text{if } a \geq 0 \wedge b < 0 \\ a \oplus b & \text{otherwise} \end{cases}$$

where $x \oplus y = \mathbf{L}_a(x, y)$, with such an approach the space requirement of the table is in the order of

$$S_{\mathbf{L}_a} \in \Theta(b_o \cdot 2^{b_i} \cdot 2^{b_i}) \quad (6)$$

using the brain-float16 data format for the output, for an 8 bit number, the space occupation is $2^8 \cdot 2^7 \cdot 2\text{byte} = 64\text{kibibyte}$.

d) *Other operations*: With similar reasoning as with multiplication table, we could further reduce the space occupation by dividing the table in two: $\mathbf{L}_a^+ : \mathbb{M}_{\geq 0}^i \times \mathbb{M}_{\geq 0}^i \rightarrow \mathbb{M}^o$ and $\mathbf{L}_a^- : \mathbb{M}_{\leq 0}^i \times \mathbb{M}_{\geq 0}^i \rightarrow \mathbb{M}^o$ — that is one for positive sums and other for negative sums (subtractions), — then store only the diagonal matrix of each one by imposing $|x| \geq |y|$, thus the space occupation would be reduced to $2 \cdot \binom{2^7}{2} \cdot 2\text{byte} = 31.75\text{kibibyte}$.

e) *Other operations*: With such tables it is also possible to perform the following operations:

- subtraction as $a + (-b)$
- division as $a \cdot \frac{1}{b}$
- conversion to a higher precision format, assuming the LUT is setup to perform operations with promotion, by performing a look-up to $a + 0$ or $a \cdot 1$

E. Exact accumulator

When computing the dot-product as described in Section III-B, it is necessary to use a numerical format with higher dynamic precision to represent the intra-block accumulator, this is especially important when using Algorithm 1 to quantize the values; otherwise, if an FP accumulator with the same dynamic range as \mathbb{M}_{RP} is used, an overflow is most likely to occur.

Such a statement is quite easy to prove, let us assume that $RP = \text{E5M2}$, we know from how the block is built that $\exists i \ c_i = m_i \cdot 2^8$, let us say that there are two consecutive elements that satisfy the predicate, 0 and 1 such that $\text{sign}(m_0) = \text{sign}(m_1)$, then the computer will have to compute the following expressions:

$$\begin{aligned} a &\leftarrow 0.00 \\ a &\leftarrow a \oplus m_0 \cdot 2^8 \\ a &\leftarrow a \oplus m_1 \cdot 2^8 \end{aligned}$$

where a is a E5M2 accumulator, as a cannot store a number with exponent 9 and an overflow occurs.

With software emulation we can use full-precision accumulators and avoid this issue completely; however a possible hardware implementation or even a software approach based on look-up tables may find the use of higher precision accumulators slower or more expensive to implement, so we propose a simple architecture to compute the dot-product that does not require the use of expensive floating-point circuitry.

Our approach works with all kind of data; however it might be possible to build a cheaper and faster architecture by exploiting possible statistical distributions that are commonly found in values stored in deep neural networks [11].

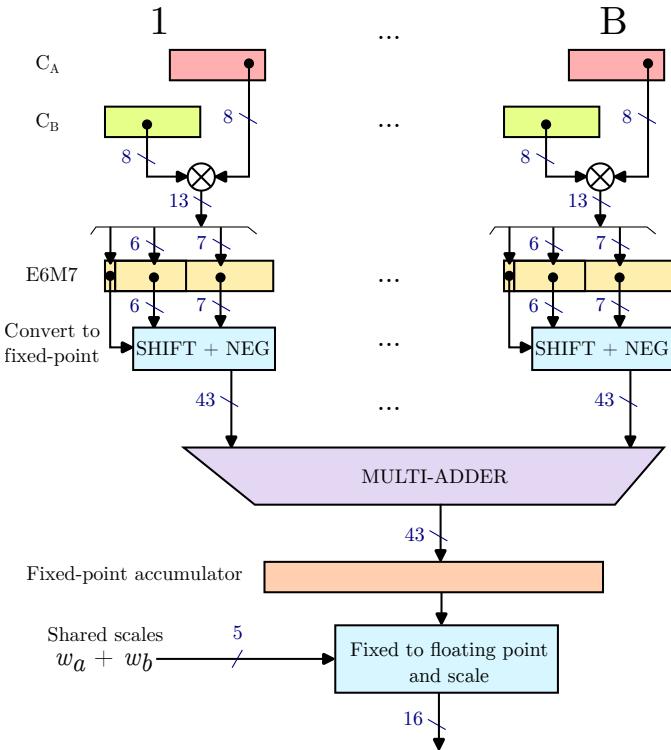


Fig. 7. Simplified implementation of a circuit that computes the dot products between two E4M3 Microscaling vectors without rounding error. The logic to handle special cases, such as overflows, NaNs, infinities etc...; has been omitted for simplicity's sake. The multi-adder could be implemented as a $\lceil \log_2(B) \rceil$ levels tall tree of adders.

a) Multiplication: Let us consider the multiplication between two normalized floating point numbers, whose mantissae and exponents are represented on M and E bits respectively.

$$x = -1^{s_x} \cdot m_x \cdot 2^{\xi_x} \quad y = -1^{s_y} \cdot m_y \cdot 2^{\xi_y} \quad m_x, m_y \in [1, 2)$$

The product $z = x \cdot y$ will have its mantissa $m_z = m_x \cdot m_y \in [1, 4)$ represented on $2E + 2$ bits and its exponent will be $\xi_x + \xi_y$; if the mantissa falls in the range $[2, 4)$ it needs to be normalized, this is done by dividing the mantissa by two and by increasing the exponent by one. The exponent is stored on $E + 2$ bits [12, Section 11.2].

Thus the resulting significand can be stored, without rounding, on $2M + 1$ bits, or it can be rounded and stored in a smaller format. With similar reasoning, the exponent can be stored on $E + 2$ bits or it can be stored in a smaller format by handling appropriately the possible overflow.

b) Addition: It is possible to completely rule out such an issue by performing the addition in a fixed-point format large enough to represent all the possible mantissae and their magnitude at once; that is an integral number in which the significand associated to the number of smallest magnitude occupies the M least significative bits of the number, whereas the one associated to the largest magnitude occupies the M most significative bits of the number. Negative addenda are represented in the two's complement.

It is easy to prove that for a floating-point format with an exponent range of $\xi \in [\xi_{\min}, \xi_{\max}]$ and a mantissa represented on M bits, the fixed-pointed format must be $\xi_{\max} - \xi_{\min} + 1 + E$ bits wide. [12, Chapter 21.1]

The sum is then performed using integral adders and final result can then be converted back to a floating-point format.

Table II shows a comparison of the floating-point width necessary to accumulate the products of numbers represented in various input formats.

c) Implementation: We added a new template class called `exact_accumulator` that implements a software-defined exact accumulator, whose parameters are the minimum exponent representable and the mantissa width in bits. For simplicity's sake we programmed the accumulator to be always `int64_t` — that is a 64 integral number, — this restrict the implementation to work only with smaller formats; we also don't handle any corner case, like overflows, infinities etc... as we are certain they do not manifest themselves in our use-case.

The class exposes two public methods, the `operator+=()` used to add a value to the accumulator and the `unpack()` used to “unpack” the fixed point accumulator into a Unpacked data-structure.

Figure 7 shows how a possible hardware implementation of such a Microscaling dot-product with exact accumulators could be implemented.

VI. EVALUATION AND RESULTS

We performed two kinds of tests to gauge the performance of our Microscaling implementation, firstly we fine-tuned a GPT2 model with 124 million parameters on the Tiny Shakespeare corpus and compared the learning trend with the base implementation; then we ran an inferential test on a GPT2 model with 1.6 billion parameters. Both models were originally trained by Karpathy [8].

We tested a *quantization* approach, where an already trained model — originally trained in full or half precision — stores some or all of its numerical values in a lower precision than the format used during training.

A. Fine-tuning

Due to the expensiveness of training a large langue model on the CPU instead of leveraging the GPU's capabilities, as well as the massive overhead introduced by the introduction of software-defined numerical formats; it was very difficult to train the network for a significative number of iterations and we had to settle to one hundred iterations.

Figure 8 shows the various learning curves for various tested configurations:

- Line **Baseline** shows the learning curve of the network ran with the default settings, that is with vectors of full-precision floats
- Line **A** shows the learning curve of the network while using a normal vector of `bfloat16` values
- Line **B** is the same as **A** but a weights' master-copy is employed
- Line **C** uses the Microscaling data-format with the quantized values stored in the IEEE's `float16`, this was done to prove the correctness of the Microscaling implementation

TABLE II
FORMAT REQUIRED TO REPRESENT AN EXACT MULTIPLICATION

Operand type	Operand's limits		Product's limits		Product type	fixed-point width
	ξ_{\max}	ξ_{\min}	ξ_{\max}	ξ_{\min}		
E4M3	8	-9	17	-18	E6M7	43
E5M2	15	-16	31	-32	E7M5	69
E3M4	3	-4	7	-8	E5M9	25
E5M10 (float16)	15	-14	31	-28	E6M21	81
E8M7 (bfloat16)	127	-126	255	-252	E9M15	523

- Line **D** uses the Microscaling data-format, with the quantized values stored in the E4M3, to store the values of the *weights* and of the *gradients*; whereas the values of the *activations* are stored in a vector of bfloat16. The weights' master-copy and full-precision probabilities and encodings are employed
- Line **E** is the same as **D** but all values are stored in the Microscaling data-format, using E4M3 to store the quantized values
- Line **F** stores all the values in bfloat16, uses the weights' mastercopy and computed the matrix multiplication by performing an online compression to Microscaling, using E4M3 to store the quantized weights

All figures involving Microscaling were obtained by using a block of size 32.

Configuration **F** is the proposed architecture by Rouhani et al. [1], indeed we proved its correct functioning.

When using E5M2 in place of E4M3 we observe an overall increase in the relative error of around +5% for all configuration but configuration **E** that sees an increase of +111.33%, this is most likely due to the loss of fractional precision.

Table III shows the relative differences between the base implementation and the various tested configurations.

B. Non-standard 8-bits floating-point formats

Until now we used the numerical formats E4M3 and E5M2 to store the quantized elements in Microscaling as they are the formats formalized by the Open Compute Project [6, 13]. We saw that they do not provide enough fractional precision to properly represent the activation values of a GPT2 model — at least of quantized one. Now we aim to evaluate whether a custom format can deliver improved performance, to this end, we introduce the *E3M4* format, a floating-point representation featuring three bits for the exponent and four bits for the significand. This configuration decreases the dynamic range while increasing fractional precision.

Let us call the configuration **G** the configuration that uses E3M4 to store all quantized Microscaling values, that is weights, activation values and gradients.

We observed a relative error of 75% when compared to the base implementation.

C. Exact accumulator

Until now, we used a floating-point accumulator to compute the Microscaling-accelerated dot-product. Now we present the learning curve for configuration **D** and **F** when a fixed-point

TABLE III
RELATIVE ERRORS BETWEEN VARIOUS LOSS CURVES

Configuration	Difference	R. Difference
baseline		
A	0.1120	2.89%
B	0.0036	0.09%
C	0.0187	0.48%
D	0.9948	25.66%
E	4.7448	122.39%
F	0.6249	16.12%
G	2.917	75.23%

roundless accumulator is employed. Using E4M3 8-bit floats in the Microscaling representation, we need to employ a 43 bits accumulator.

Table IV shows the relative differences between the base implementation and the various tested configurations, while Figure 9 the loss curves during training.

TABLE IV
RELATIVE ERRORS BETWEEN VARIOUS LOSS CURVES

Configuration	Difference	R. Difference
baseline		
D	0.9948	25.66%
D'	0.961	23.93%
F	0.6249	16.12%
F'	0.506	12.61%

D. Text generation

Let us consider the aforementioned configurations, using the 1.6 billion parameters model we want to see if the network, without any kind of additional fine-tuning, is still able to produce coherent text given a small and simple prompt.

"A modern car must have a main engine motor that can stand up to road conditions. Wind resistance, misalignment, defects, and cutting off the engine can prevent the engine from retaining the power needed to do the job. Braking power must be enhanced to maintain speed. The characteristics of braking must allow the use of [...]" (GPT2 1.6B — baseline)

For brevity's sake we do not report text generations for all configurations, as they were able to produce coherent text.

"A modern car must have a one-set of four road wheels and one rear wheel. There is also the traditional four-wheel setup.

In today's day and age, motor vehicles by no means should be considered as antiquated or inferior. The primary purposes of an automobile are for motoring and maintaining a reasonable lifestyle.

Almost any type of automobile can be exchanged for a different one. However, quite a few have their set up

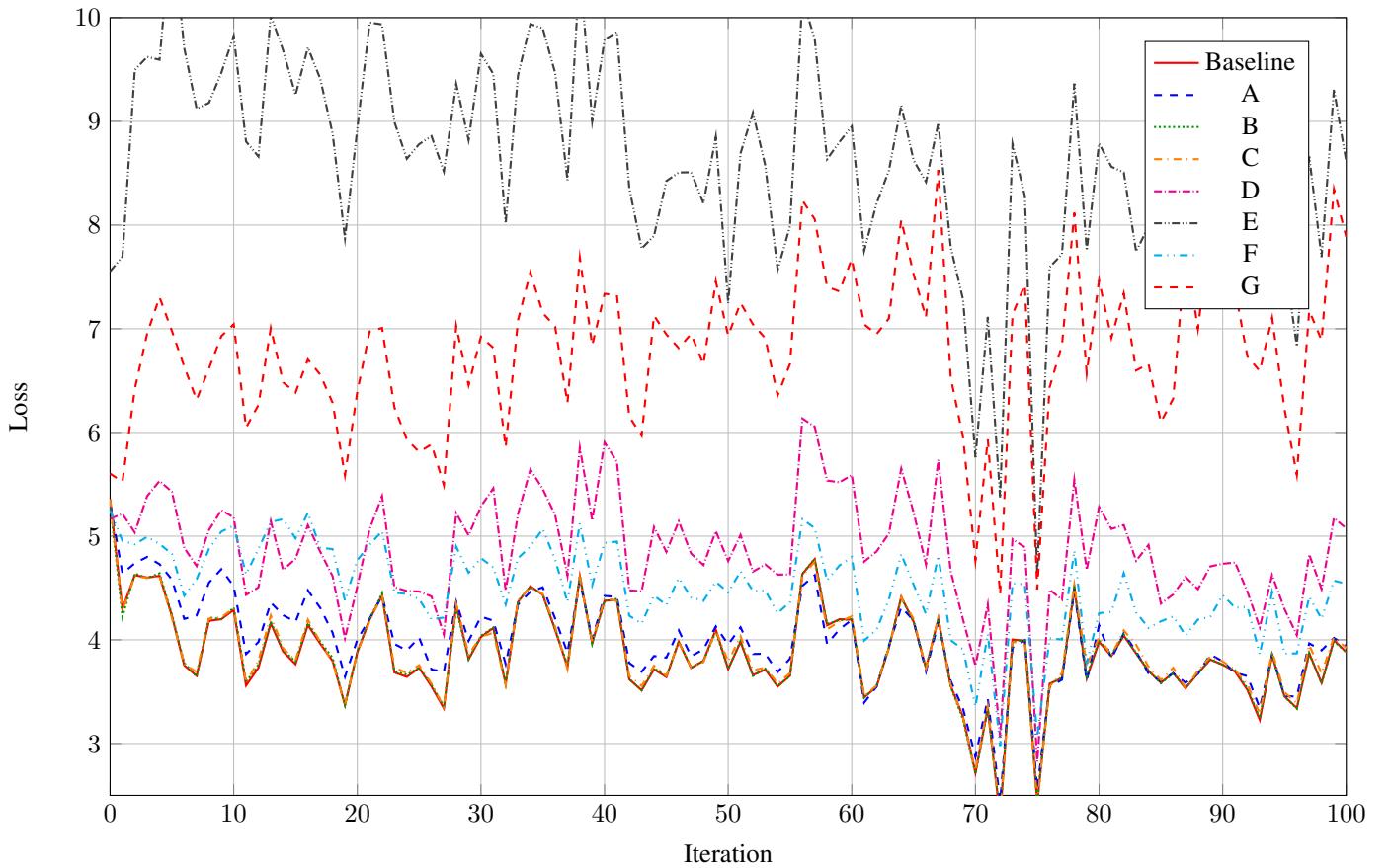


Fig. 8. Learning curves for different methods

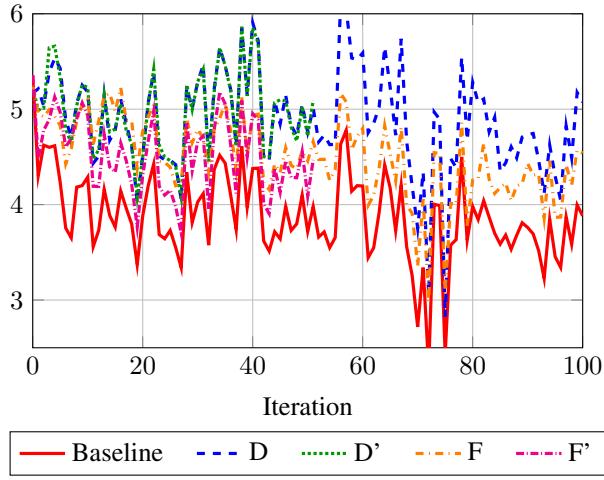


Fig. 9. B' and F' denote the configurations using the exact accumulator. Loss on the y axis

easily impacted; those people who've transferable cars have found out simple methods to relocate their car to wherever they want.

Car parts come in various kinds [...]” (GPT 1.6B — D)

Configuration **E** proves again to be the most problematic one when quantizing already trained models as it generates nonsensical text:

”**A modern car must have** numer- its opposite. has

practice ever— access.). seem. apparently beyond each YEAR deposits allS ur history no to As available back comparei 05 instead otherPR wouldn while pr pr am festival ofpr pr pr pr pr pr pr but when Point se ... pr pr the Pr dissolve consider,Nearly 10 R instead lear were for prepares such a otheresne another light corresponding e soon an is etc most goodismss how soon then also 10 am experience eventually very much jumpES matters turned a tion first the your turkey once also y first somewhat all simply (style yours debtor cold queen [...]” (GPT 1.6B — E)

The same goes for configuration **G**.

On the other hand, the use of Microscaling only the matrix multiplication produces good results:

”**A modern car must have** a more powerful engine to get and maintain top speed. In the early days of Henry Ford and in vintage Ferraris, it was just the drive train and tires. Early cars had ONE average horsepower. Then long after the switch was made to 3.5L engines, individual car engines improved by a fixed percentage. With infinite amounts of horsepower at the edict of maximum capacity (or maximum power for the engine), manufacturers could simply challenge more powerful engines to challenge each other. Then after 1975 the 3.5L began introducing 40-50 HP engines. Most early examples of such advancements don't [...]” (GPT 1.6B — F)

VII. CONCLUSIONS AND FUTURE WORK

We sucessfully implemented the Microscaling data-format in modern C++23 in such a generic way that allows to

the test the performance of the format with many types of numerical format, both native or software emulated, like floating-point numbers with custom sizes for their exponents and significands, integral numbers or more exotic formats like the Posit.

Using the half-precision format bfloat16, we successfully reproduced established results, confirming that both inference and training are achievable with this format. While training can experience some instability or slowdown without maintaining a full-precision master copy of the weights during updates, these challenges are manageable.

Next, we explored the Microscaling data structure and demonstrated its practicality.

When extending our investigation to tiny precision formats like E4M3 and E5M2 in conjunction with Microscaling, we encountered challenges with network stability. However, our work proved that these formats can effectively store weights and gradients while retaining activation values in half precision. Additionally, we showed that a custom eight-bit format, E3M4, with enhanced fractional precision, successfully stores activation values in a GPT-2 network trained in half precision, marking an important step toward broader usability.

Looking ahead, we remain optimistic that architectural innovations and further refinements will enable neural networks to fully leverage the Microscaling format. Our experiments identified key areas, such as encoding, softmax, and self-attention blocks, that require additional focus. Nevertheless, the progress achieved so far demonstrates the feasibility of overcoming these hurdles, bringing us closer to realizing the full potential of Microscaling for future neural network architectures.

Importantly, we empirically validated the computational architecture proposed by the Open Compute Project [1], which stores data in half precision and dynamically compresses it to tiny precision with Microscaling for matrix operations. This approach functions effectively and, with appropriate hardware support, holds the potential to significantly accelerate learning and inference.

REFERENCES

- [1] B. D. Rouhani, R. Zhao, A. More, M. Hall, A. Khodamoradi, S. Deng, D. Choudhary, M. Cornea, E. Dellinger, K. Denolf, S. Dusan, V. Elango, M. Golub, A. Heinecke, P. James-Roxby, D. Jani, G. Kolhe, M. Langhammer, A. Li, L. Melnick, M. Mesmakhosroshahi, A. Rodriguez, M. Schulte, R. Shafipour, L. Shao, M. Siu, P. Dubey, P. Micikevicius, M. Naumov, C. Verrilli, R. Wittig, D. Burger, and E. Chung, “Microscaling data formats for deep learning,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.10537>
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” 2019.
- [3] D. Kalamkar, D. Mudigere, N. Mellempudi, D. Das, K. Banerjee, S. Avancha, D. T. Vootturi, N. Jammalamadaka, J. Huang, H. Yuen, J. Yang, J. Park, A. Heinecke, E. Georganas, S. Srinivasan, A. Kundu, M. Smelyanskiy, B. Kaul, and P. Dubey, “A study of bfloat16 for deep learning training,” 2019. [Online]. Available: <https://arxiv.org/abs/1905.12322>
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 06 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [5] D. Goldberg, “What every computer scientist should know about floating-point arithmetic,” *ACM Comput. Surv.*, vol. 23, no. 1, p. 5–48, mar 1991. [Online]. Available: <https://doi.org/10.1145/103162.103163>
- [6] B. D. Rouhani, N. Garegrat, T. Savell, A. More, K.-N. Han, R. Zhao, M. Hall, J. Klar, E. Chung, Y. Yu, Microsoft, M. Schulte, R. Wittig, AMD, I. Bratt, N. Stephens, J. Milanovic, J. Brothers, Arm, P. Dubey, M. Cornea, A. Heinecke, A. Rodriguez, M. Langhammer, Intel, S. Deng, M. Naumov, Meta, P. Micikevicius, M. Siu, NVIDIA, C. Verrilli, and Qualcomm, “OCP Microscaling Formats (MX) Specification,” *Open Compute Project*, 2023. [Online]. Available: <https://www.opencompute.org/documents/ocp-microscaling-formats-mx-v1-0-spec-final-pdf>
- [7] E. Ruffaldi, F. Rossi, M. Cococcioni, and S. Saponara, “cppPosit: a c++ template-based library implementing posit arithmetic ready for machine learning applications,” 09 2023. [Online]. Available: <https://www.techrxiv.org/users/692834/articles/682773>
- [8] A. Karpathy, “ilm.c,” 2024. [Online]. Available: <https://github.com/karpathy/ilm.c/tree/master>
- [9] ISO, *ISO/IEC 14882:2011 Information technology — Programming languages — C++*. Geneva, Switzerland: International Organization for Standardization, Dec. 2020. [Online]. Available: <https://www.iso.org/standard/79358.html>
- [10] IEEE, “Ieee standard for floating-point arithmetic,” *IEEE Std 754-2008*, pp. 1–70, 2008.
- [11] X. Ju, J. He, M. Wen, J. Feng, Y. Cao, J. Shen, Z. Chen, and Y. Shi, “Winacc: Window-based acceleration of neural networks using block floating point,” 12 2024.
- [12] F. de Dinechin and M. Kumm, *Application-Specific Arithmetic: Computing Just Right for the Reconfigurable Computer and the Dark Silicon Era*. Springer, 2024.
- [13] P. Micikevicius, S. Oberman, P. Dubey, M. Cornea, A. Rodriguez, I. Bratt, R. Grisenthwaite, N. Jouppi, C. Chou, A. Huffman, M. Schulte, R. Wittig, D. Jani, and S. Deng, “OCP 8-bit Floating Point Specification (OFP8),” *Open Compute Project*, 2023. [Online]. Available: <https://www.opencompute.org/documents/ocp-8-bit-floating-point-specification-ofp8-revision-1-0-2023-12-01-pdf-1>