INFORMATICA, 2025, Vol. 36, No. 1, 125–140

DOI: https://doi.org/10.15388/25-INFOR587

An Approximate Closed-Form Expression for Calculating Performance of Floating-Point Format for the Laplacian Source

Zoran PERIĆ, Bojan DENIĆ*, Milan DINČIĆ, Sofija PERIĆ

University of Niš, Faculty of Electronic Engineering, Aleksandra Medvedeva 4, 18000 Niš, Serbia
e-mail: zoran.peric@elfak.ni.ac.rs, bojan.denic@elfak.ni.ac.rs, milan.dincic@elfak.ni.ac.rs, sofija.peric@elfak.ni.ac.rs

Received: October 2024; accepted: March 2025

Abstract. This paper introduces a novel approach that bridges the floating-point (FP) format, widely utilized in diverse fields for data representation, with the μ-law companding quantizer, proposing a method for designing and linearizing the μ-law companding quantizer to yield a piecewise uniform quantizer tailored to the FP format. A key outcome of the paper is a closed-form approximate expression for closely and efficiently evaluating the FP format's performance for data with the Laplacian distribution. This expression offers generality across various bit rates and data variances, markedly reducing the computational complexity of FP performance evaluation compared to prior methods reliant on summation of a large number of terms. By facilitating the evaluation of FP format performance, this research substantially aids in the selection of the optimal bit rates, crucial for digital representation quality, dynamic range, computational overhead, and energy efficiency. The numerical calculations spanning a wide range of data variances provided for some commonly used FP versions with an 8-bit exponent have demonstrated that the proposed closed-form expression closely approximates FP format performance.

Key words: floating-point format, piecewise uniform quantization, μ-law companding quantization, Laplacian source.

1. Introduction

The floating-point (FP) format is extensively employed for data representation across various domains, including computing (Fasi and Mikaitis, 2021; Burgess *et al.*, 2019a), neural networks (Zhao *et al.*, 2023; Bai-Kui and Shanq-Jang, 2023), and signal processing (Moroz and Samotyy, 2019). The prevalent 32-bit floating-point (FP32) format adheres to standardized specifications (IEEE 754, 2019), boasting exceptional digital representation quality across a very wide range of data variance, ranging from minuscule to substantial values. However, the FP32 format's computational intensity poses a challenge for implementation on hardware-constrained devices (Yang *et al.*, 2022; Syed *et al.*, 2021; Cattaneo *et al.*, 2018). The 24-bit FP (FP24) (Junaid *et al.*, 2022), 16-bit FP (Bfloat16

*Corresponding author.

and DLFloat) (Burgess *et al.*, 2019b; Agrawal *et al.*, 2019), and 8-bit (FP8) (Wang *et al.*, 2018) formats are examples of lower-bit FP formats that reduce computational complexity and energy consumption, making them advantageous for hardware and energy-restricted systems. Conversely, formats such as 64-bit FP (FP64) (IEEE 754, 2019) are utilized in environments necessitating heightened calculation precision (Botta *et al.*, 2021).

As evident, there exists a plethora of FP formats, each with varying bit rates, offering distinct qualities in digital representation, dynamic range, computational complexity, and energy usage. Selecting the optimal FP format is a crucial task in both research and practical applications, contingent upon several factors: the required digital representation accuracy for a specific application, the range of data variance, as well as the available hardware and energy resources. Generally, it is preferable to opt for an FP format with fewer bits to minimize hardware demands and energy consumption while ensuring the requisite level of representation accuracy for a specific application across the entire range of data variance. Achieving this necessitates an efficient mechanism for evaluating the performance of FP formats across different bit rates and data variance levels.

It is worth noting that none of the aforementioned papers dealing with the FP formats (Junaid *et al.*, 2022; Agrawal *et al.*, 2019; Burgess *et al.*, 2019a; Wang *et al.*, 2018; Botta *et al.*, 2021) provide information regarding their actual performance, which is a critical factor for practical applications. A significant stride in this direction was made in Perić *et al.* (2021), Denić *et al.* (2023), where the correlation was established between the FP format and a piecewise uniform quantizer, which was termed *the floating-point quantizer* (FPQ). Namely, the piecewise uniform quantizer includes a number of segments, where a unique uniform quantizer is defined in each segment (Dinčić *et al.*, 2016; Jayant and Noll, 1984; Gersho and Gray, 1992). This actually allowed assessing the digital representation quality of the FP format using an objective performance measure such as the signal-to-quantization-noise ratio (SQNR) of FPQ. It's crucial to acknowledge that the performance of the FP format, specifically the SQNR of FPQ, relies heavily on the statistical properties of the data, primarily the probability density function (PDF). This paper considers the Laplacian PDF, given its extensive usage in statistically modelling various data types, e.g., speech (Chu, 2003; Gazor and Zhang, 2003) and neural network weights (Banner *et al.*, 2018; 2019).

The primary goal of this paper is to make a significant advancement towards the FP format analysis by providing a performance-evaluating method that is more efficient (in terms of computational complexity) compared to the previously developed method (Perić *et al.*, 2021; Denić *et al.*, 2023). This is achieved by linking the FP format with the μ-law companding quantizer (μCQ), which is actually a novel concept, as research on this topic has not been done before. Namely, the SQNR expression for FPQ in Perić *et al.* (2021), Denić *et al.* (2023) is not provided in a closed form but as the summation of numerous terms (e.g., for the FP32 format, this sum comprises 254 terms, Perić *et al.*, 2021), thereby escalating the complexity of FP format performance computation. Hence, the paper aims to eliminate the mentioned drawback of the existing method for assessing the performance of the FP format. A significant contribution is the development of a procedure for designing a μCQ, tailoring its key parameters (μ-compression factor and $x_{\max}$ – maximal amplitude)

to the FP format. The key outcome of this innovative approach is the provision of a simple closed-form approximate expression for closely and efficiently assessing the FP format's performance. The advantage of this closed-form expression is its broad applicability, as it applies universally to any bit rate and data variance. Aside from the theoretical significance of deriving a closed-form expression for performance evaluation, this paper holds substantial practical value by considerably simplifying the complexity of computing FP format performance.

The paper's methodology involves designing an appropriate μCQ, linearizing it, and deriving a piecewise uniform quantizer based on the μ-law compression function (PUQ$^{\mu}$). The paper demonstrates that by selecting the appropriate values of the crucial design parameters of the μCQ, the structure of its linearized version, PUQ$^{\mu}$, aligns with the FPQ structure. Notably, the paper provides a closed-form expression for the SQNR of μCQ for the Laplacian PDF, obtained by simplifying the general SQNR expression for μCQ provided in Perić *et al.* (2010). The accuracy of the derived closed-form SQNR expression for μCQ is examined considering versions of the FP format with 8-bit exponent, FP24 and FP32, and a very wide dynamic range of input data variances. It is shown that the proposed SQNR expression is highly efficient in estimating FP performance when confronted with the existing approach (Perić *et al.*, 2021; Denić *et al.*, 2023), with the SQNR calculation error below 1% defining the reasonable accuracy of the SQNR formula (Na, 2011). Thus, utilizing the proposed approach instead of the previously introduced one based on a summation of numerous terms ensures a high level of accuracy and leads to a noteworthy reduction in computational complexity.

The rest of the paper is organized as follows. In Section 2, the description of the R-bit FP format is provided, and its connection with the piecewise uniform quantization is explained. The main result is exposed in Section 3, which performs the design of the μCQ along with its linearized version tailored to the FP format and provides the closed-form expression for estimating FP format performance. Section 4 presents simulation results and highlights the benefits of the approach studied in the paper. Section 5 gives concluding remarks.

2. Description of the Floating-Point Format

A real number x is encoded in the R-bit FP format as IEEE 754 (2019):

$$x = (sa_{e-1} \dots a_1 a_0 b_{m-1} \dots b_1 b_0)_2, \tag{1}$$

consisting of one bit s to indicate the sign, e bits $(a_{e-1} \dots a_1 a_0)$ to represent the exponent E, and m bits $(b_{m-1} \dots b_1 b_0)$ to represent the significand M of the number x, whereas $R = e + m + 1$. The exponent $E = \sum_{i=0}^{e-1} a_i 2^i$ can take values from 0 to $2^e - 1$, but the values $E = 0$ and $E = 2^e - 1$ are reserved according to IEEE 754 (2019), leaving $L^{FP} = 2^e - 2$ values of E (from 1 to $2^e - 2$) that can be used to represent numbers. The parameter $M = \sum_{i=1}^{m} b_{m-i} 2^{m-i}$ can take values from 0 to $2^m - 1$. The number x,

represented with (1), can be calculated in its decimal form as IEEE 754 (2019):

$$x = (-1)^s 2^{E^*} \left(1 + \frac{M}{2^m}\right), \tag{2}$$

where $E^* = E - bias$ denotes the biased exponent and $bias = L^{FP}/2$ is a predefined parameter. Therefore, the biased exponent E^* takes values from $E^*_{\min} = 1 - L^{FP}/2$ to $E^*_{\max} = L^{FP}/2$. For example, for FP32 we have $e = 8$ and $m = 23$ (IEEE 754, 2019), while for FP24 we have $e = 8$ and $m = 15$. Due to the same e value, both FP32 and FP24 formats have identical values for the following parameters: $L^{FP} = 254$, $bias = 127$, $E^*_{\min} = -126$, and $E^*_{\max} = 127$.

The R-bit FP format exhibits symmetry around 0, as every positive number in the format corresponds to a symmetric negative number. Let's examine positive numbers within the R-bit FP format, without losing generality. The maximum positive number representable in this format (for $E^* = E^*_{\max}$ and $M = 2^m - 1$) is:

$$x^{FP}_{\max} = 2^{E^*_{\max}} \left(1 + \frac{2^m - 1}{2^m}\right) = 2^{E^*_{\max}} \left(2 - \frac{1}{2^m}\right) \approx 2^{E^*_{\max}+1} = 2^{L^{FP}/2+1}. \tag{3}$$

For each value of E^* ($E^*_{\min} \leqslant E^* \leqslant E^*_{\max}$) we define a segment $S_{E^*} = [2^{E^*}, 2^{E^*+1})$ of width $\delta_{E^*} = 2^{E^*}$, which includes 2^m equidistant real numbers $2^{E^*}(1 + \frac{M}{2^m})$, $M = 0, \ldots, 2^m - 1$, placed at a mutual distance $\Delta_{E^*} = 2^{E^*}(1 + \frac{M+1}{2^m}) - 2^{E^*}(1 + \frac{M}{2^m}) = 2^{E^*-m}$. Hence, in the positive part of the real axis, there are a total of L^{FP} segments S_{E^*}, each containing 2^m equidistant numbers with a step size of Δ_{E^*}. Due to symmetry, the same structure of L^{FP} segments with 2^m equidistant numbers also exists in the negative part of the real axis. Since

$$\delta_{E^*+1} = 2^{E^*+1} = 2 \cdot 2^{E^*} = 2\delta_{E^*} \tag{4}$$

and

$$\Delta_{E^*+1} = 2^{E^*+1-m} = 2 \cdot 2^{E^*-m} = 2\Delta_{E^*}, \tag{5}$$

it can be concluded that the width of segment S_{E^*+1} is twice as large as the width of segment S_{E^*}, and the distance between adjacent numbers in S_{E^*+1} is twice as high as in S_{E^*}. Therefore, as the value of E^* increases, the distance between adjacent numbers increases, meaning that the FP format provides a finer representation of smaller numbers.

The described structure of the FP format fully corresponds to the structure of a symmetric piecewise uniform quantizer with a maximum amplitude $x^{FP}_{\max}$ defined with (3), which in the positive part has L^{FP} segments $S_{E^*} = [2^{E^*}, 2^{E^*+1})$, $E^*_{\min} \leqslant E^* \leqslant E^*_{\max}$, each segment undergoing uniform quantization with 2^m quantization levels and with the step size $\Delta_{E^*} = 2^{E^*-m} = 2^{E^*-R+e+1}$. This model of quantizer, whose structure mirrors that of the FP format, is known as *the floating-point quantizer* − FPQ (Perić *et al.*, 2021; Denić *et al.*, 2023). This analogy between the FP format and the FPQ is significant,

enabling the FP representation quality to be assessed using an objective measure such as SQNR of the FPQ. SQNR is generally defined as Jayant and Noll (1984), Chu (2003), Gersho and Gray (1992):

$$\mathrm{SQNR}(\sigma) = 10\log_{10}\frac{\sigma^2}{D(\sigma)}, \tag{6}$$

where σ^2 represents the variance of data to be quantized and $D\ (\sigma)$ is distortion that represents an error that occurred during quantization. In the case of FPQ, σ^2 represents the variance of data to be represented in the FP format, while distortion of FPQ represents the error that occurred during FP representation of real numbers and can be expressed in general form as Perić *et al.* (2021), Denić *et al.* (2023):

$$D^{\mathrm{FPQ}}(\sigma) = \underbrace{2\sum_{E^*=E^*_{\min}}^{E^*_{\max}} \frac{\Delta_{E^*}^2}{12} P_{E^*}(\sigma)}_{D_g^{\mathrm{FPQ}}(\sigma)} + \underbrace{2\int_{x_{\max}^{FP}}^{+\infty} \left(x - x_{\max}^{FP}\right)^2 p(x,\sigma)dx}_{D_{ov}^{\mathrm{FPQ}}(\sigma)}. \tag{7}$$

Multiplication by 2 in the expression (7) is used to account for the distortion in the negative part of the real axis. The first term in (7), expressed as a sum, represents the granular distortion D_g^{FPQ} in L^{FP} segments S_{E^*} ($E^*_{\min} \leqslant E^* \leqslant E^*_{\max}$), where $P_{E^*}(\sigma) = \int_{2^{E^*}}^{2^{(E^*+1)}} p(x,\sigma)dx$ represents the probability that the real number x belongs to segment S_{E^*}, with $p(x,\sigma)$ representing the PDF of the input data. The second term in (7) represents the overload distortion D_{ov}^{FPQ} that occurs during quantization of numbers outside the support region of the FPQ.

This paper examines the zero-mean Laplacian PDF of variance σ^2, defined as Jayant and Noll (1984), Gersho and Gray (1992):

$$p(x,\sigma) = \frac{1}{\sqrt{2}\sigma}\exp\left(-\frac{\sqrt{2}|x|}{\sigma}\right). \tag{8}$$

For $p(x,\sigma)$ defined with (8), based on (3), (6), and (7), the following SQNR expression for the FPQ quantizer is obtained:

$$\mathrm{SQNR}^{\mathrm{FPQ}}(\sigma) = -10\log_{10}\left[\sum_{E^*=1-L^{FP}/2}^{L^{FP}/2} \frac{2^{2(E^*-R+e)}}{3\sigma^2}\left(\exp\left(-\frac{2^{E^*+1/2}}{\sigma}\right) - \exp\left(-\frac{2^{E^*+3/2}}{\sigma}\right)\right) + \exp\left(-\frac{2^{(L^{FP}+3)/2}}{\sigma}\right)\right]. \tag{9}$$

Using (9), it is possible to compute the performance of the R-bit FP format for any value of data variance. However, expression (9) contains the sum of L^{FP} elements, being computationally demanding since L^{FP} is typically a large number (Perić *et al.*, 2021; Denić *et al.*, 2023). This issue will be solved in the next section, where an approximate closed-form expression is supplied for efficiently calculating the performance of the R-bit FP format.

3. A Closed-Form SQNR Expression Derivation by Designing and Linearizing a μ-Law Companding Quantizer Related to FPQ

A key outcome of this section is a simple closed-form SQNR expression for an appropriately designed μ-law companding quantizer (μCQ) that can be used as a very close performance approximation for FP formats, reducing the complexity associated with calculating the performance of FP formats explained in Section 2. In the following, we give the design of a μCQ in such a way that its linearization yields a piecewise uniform quantizer (PUQ$^{\mu}$) whose structure closely resembles that of the FPQ. It will be shown that the performance of μCQ and PUQ$^{\mu}$ are very close, providing a basis for utilizing the derived SQNR formula of μCQ as a very good approximation of FP formats' performance.

3.1. *Design of a μ-Law Companding Quantizer Inspired by the FP Format*

Companding quantizers are typically implemented as a cascade connection compressor–uniform quantizer–expander. For a symmetric μCQ, the compressor function $c_{\mu}(x) : [-x_{\max}, x_{\max}] \to [-x_{\max}, x_{\max}]$ is defined as Jayant and Noll (1984), Gersho and Gray (1992):

$$c_{\mu}(x) = \frac{x_{\max}}{\ln(1+\mu)} \ln\left(1 + \frac{\mu|x|}{x_{\max}}\right) \operatorname{sgn}(x), \quad 0 \leqslant |x| \leqslant x_{\max}, \tag{10}$$

where μ is a compression factor and $x_{\max}$ is the maximal amplitude of the quantizer. The decision thresholds x_j^{μ} and representational levels y_j^{μ} of the μCQ quantizer in the positive part of the real axis can be specified in the following way (Dinčić *et al.*, 2021; Perić *et al.*, 2010):

$$x_j^{\mu} = c_{\mu}^{-1}\left(2j\frac{x_{\max}}{N}\right) = \frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{2j}{N}} - 1\right), \quad 0 \leqslant j \leqslant N/2, \tag{11}$$

$$y_j^{\mu} = c_{\mu}^{-1}\left((2j-1)\frac{x_{\max}}{N}\right) = \frac{x_{\max}}{\mu}\left((1+\mu)^{\frac{2j-1}{N}} - 1\right), \quad 1 \leqslant j \leqslant N/2, \tag{12}$$

where N denotes the number of representational levels, while $\Delta_u = 2x_{\max}/N$ defines the step size of the uniform quantizer and $c_{\mu}^{-1}(x)$ is the inverse μ-law compression function that defines the expander. Note that the decision thresholds and representational levels of the μCQ depend on the parameters μ and $x_{\max}$, whose values will be selected based on the condition that PUQ$^{\mu}$, as a linearized version of the μCQ, has the same structure as the FPQ.

The next step is the piecewise linearization of the μCQ, achieved by approximating the compressor function $c_{\mu}(x)$ defined with (10) by a symmetric piecewise linear compressor function $g_{\mu}(x) : [-x_{\max}, x_{\max}] \to [-x_{\max}, x_{\max}]$. Due to the symmetry of $g_{\mu}(x)$ around 0, we can consider only the positive part of the real axis where $g_{\mu}(x)$ is defined as:

$$g_{\mu}(x) = a_j x + b_j, \quad x \in \left[x_{j-1}^{seg}, x_j^{seg}\right], \ 1 \leqslant j \leqslant L, \tag{13}$$

where a_j and b_j are coefficients that will be determined latter in this section, L is the number of linear segments in the positive part and x_j^{seg} $(0 \leqslant j \leqslant L)$ are the boundaries between segments, where $x_0^{seg} = 0$ and $x_L^{seg} = x_{\max}$. The function $g_\mu(x)$ must satisfy the condition of having the same values as the function $c_\mu(x)$ in the segments' boundaries x_j^{seg}:

$$g_\mu\left(x_j^{seg}\right) = c_\mu\left(x_j^{seg}\right) \equiv \frac{x_{\max}}{\ln(1+\mu)} \ln\left(1 + \frac{\mu x_j^{seg}}{x_{\max}}\right), \quad 0 \leqslant j \leqslant L. \tag{14}$$

This yields a symmetric PUQ$^\mu$ with L linear segments in the positive part of the real axis, performing uniform quantization with K uniformly spaced quantization levels within each segment. To ensure that all segments $[x_{j-1}^{seg}, x_j^{seg})$ $(1 \leqslant j \leqslant L)$ contain the same number of quantization levels, the values of $g_\mu(x)$ within the segments' boundaries x_j^{seg} $(0 \leqslant j \leqslant L)$ must be equidistant within the range $[0, x_{\max}]$, i.e. it must hold that $g_\mu(x_j^{seg}) - g_\mu(x_{j-1}^{seg}) = x_{\max}/L = \text{const}$, $1 \leqslant j \leqslant L$. This will be achieved if the following condition is fulfilled:

$$g_\mu\left(x_j^{seg}\right) = j\frac{x_{\max}}{L}, \quad 0 \leqslant j \leqslant L. \tag{15}$$

From conditions (14) and (15) it follows:

$$\frac{x_{\max}}{\ln(1+\mu)} \ln\left(1 + \frac{\mu x_j^{seg}}{x_{\max}}\right) = j\frac{x_{\max}}{L}, \quad 0 \leqslant j \leqslant L. \tag{16}$$

From here it is easy to obtain x_j^{seg}:

$$x_j^{seg} = \frac{x_{\max}}{\mu}\left((1+\mu)^{j/L} - 1\right), \quad 0 \leqslant j \leqslant L, \tag{17}$$

which is also influenced by μ and $x_{\max}$. To ensure equivalence between PUQ$^\mu$ and FPQ, we will set the parameters of the considered PUQ$^\mu$ to be equal to the corresponding parameters of the FPQ:

$$x_{\max} = x_{\max}^{FP}, \quad L = L^{FP}, \quad K = 2^m = 2^{R-e-1}, \quad N = 2LK = 2^R\left(1 - 2^{1-e}\right), \tag{18}$$

but also it is necessary for the PUQ$^\mu$ to satisfy the condition (4) valid for the FPQ that the width of each segment is twice as large as the width of the previous one:

$$x_{j+1}^{seg} - x_j^{seg} = 2\left(x_j^{seg} - x_{j-1}^{seg}\right), \quad 1 \leqslant j \leqslant L^{FP} - 1, \tag{19}$$

which will be achieved by selecting an appropriate value for the parameter μ, as will be demonstrated in the next Theorem 1.

Theorem 1. *PUQ$^{\mu}$ with parameters defined by (18) will be equivalent to the FPQ if* $\mu = 2^{L^{FP}} - 1$.

Proof. From (17) and (19), it follows:

$$\frac{x_{\max}^{FP}}{\mu}(1+\mu)^{j/L^{FP}}\left((1+\mu)^{1/L^{FP}} - 1\right) = 2\frac{x_{\max}^{FP}}{\mu}(1+\mu)^{(j-1)/L^{FP}}\left((1+\mu)^{1/L^{FP}} - 1\right), \tag{20}$$

where $1 \leqslant j \leqslant L^{FP}$. From (20) we get that:

$$(1+\mu)^{j/L^{FP}} = 2(1+\mu)^{(j-1)/L^{FP}}, \quad 1 \leqslant j \leqslant L^{FP}. \tag{21}$$

Based on (21), it is obvious that:

$$(1+\mu)^{1/L^{FP}} = 2. \tag{22}$$

Finally, it follows that:

$$\mu = 2^{L^{FP}} - 1, \tag{23}$$

which concludes the proof. □

By establishing all crucial parameters, the design of the observed PUQ$^{\mu}$, as well as μCQ (see (11) and (12)), is completed. Based on (17), (23), (18), and (3), we obtain the final expression for the segments' boundaries of the PUQ$^{\mu}$:

$$x_j^{seg} = x_{\max}^{FP}\frac{2^j - 1}{2^{L^{FP}} - 1} = 2^{L^{FP}/2+1}\frac{2^j - 1}{2^{L^{FP}} - 1} \approx 2^{-L^{FP}/2+1}\left(2^j - 1\right), \quad 0 \leqslant j \leqslant L^{FP}. \tag{24}$$

The coefficients a_j and b_j ($1 \leqslant j \leqslant L^{FP}$) in (13) can be determined as:

$$a_j = \frac{g_{\mu}(x_j^{seg}) - g_{\mu}(x_{j-1}^{seg})}{x_j^{seg} - x_{j-1}^{seg}} = \frac{x_{\max}^{FP}/L^{FP}}{x_{\max}^{FP}\frac{2^{j-1}}{2^{L^{FP}}-1}} \approx \frac{2^{L^{FP}-j+1}}{L^{FP}}, \tag{25}$$

$$b_j = g_{\mu}\left(x_j^{seg}\right) - a_j x_j^{seg} = \frac{x_{\max}^{FP}}{L^{FP}}\left(j - 2 + 2^{1-j}\right) = \frac{2^{L^{FP}/2+1}}{L^{FP}}\left(j - 2 + 2^{1-j}\right). \tag{26}$$

By introducing the step size within the j-th segment of PUQ$^{\mu}$:

$$\Delta_j = \left(x_j^{seg} - x_{j-1}^{seg}\right)/K \approx 2^{-L^{FP}/2+j-R+e+1}, \quad 1 \leqslant j \leqslant L^{FP}, \tag{27}$$

we finally define the decision thresholds $x_{j,i}$ ($0 \leqslant i \leqslant K = 2^{R-e-1}$) and the representational levels $y_{j,i}$ ($0 \leqslant i \leqslant K = 2^{R-e-1}$) of PUQ$^{\mu}$ within the j-th segment:

$$x_{j,i} = x_{j-1}^{seg} + i\Delta_j \approx 2^{-L^{FP}/2+1}\left(2^{j-1}\left(1 + i2^{-R+e+1}\right) - 1\right), \tag{28}$$

$$y_{j,i} = x_{j-1}^{seg} + (i - 1/2)\Delta_j \approx 2^{-L^{FP}/2+1}\left(2^{j-1}\left(1 + (2i-1)2^{-R+e}\right) - 1\right). \tag{29}$$

3.2. *Performance Evaluation*

Here, we provide the performance (SQNR) expressions for the discussed μCQ and PUQ$^{\mu}$. For μCQ, the granular distortion D_g^{μ} (the distortion component introduced in the granular part $[-x_{\max}, x_{\max}]$) can be assessed using Bennett's integral (Jayant and Noll, 1984; Chu, 2003; Gersho and Gray, 1992):

$$D_g^{\mu}(\sigma) = 2\frac{\Delta_u^2}{12}\int_0^{x_{\max}} \frac{p(x,\sigma)}{[c'_{\mu}(x)]^2}dx, \tag{30}$$

where $c'_{\mu}(x)$ is the first derivative of $c_{\mu}(x)$, while the overload distortion D_{ov}^{μ} (the distortion component introduced outside the granular part) is given by Jayant and Noll (1984), Chu (2003), Gersho and Gray (1992):

$$D_{ov}^{\mu}(\sigma) = 2\int_{x_{\max}}^{+\infty}(x - x_{\max})^2 p(x,\sigma)dx. \tag{31}$$

The granular distortion of PUQ$^{\mu}$, $D_g^{\text{PUQ}^{\mu}}$, can be evaluated according to the following expression (Jayant and Noll, 1984; Chu, 2003; Gersho and Gray, 1992):

$$D_g^{\text{PUQ}^{\mu}}(\sigma) = 2\sum_{j=1}^{L^{FP}} \frac{\Delta_j^2}{12}P_j(\sigma), \tag{32}$$

where $P_j(\sigma) = \int_{x_{j-1}^{seg}}^{x_j^{seg}} p(x,\sigma)dx$ denotes the probability of the j-th segment ($1 \leqslant j \leqslant L^{FP}$), while the overload distortion of PUQ$^{\mu}$, $D_{ov}^{\text{PUQ}^{\mu}}$, can be estimated by (31). Theorem 2 indicates the performance of the two mentioned quantizers.

Theorem 2. *If $L \gg 1$, distortions of μCQ and its linearized version PUQ$^{\mu}$ converge.*

Proof. As the overload distortion for these two models is defined with the same expression (31), it is sufficient to show that Bennett's integral (30) closely approximates the granular distortion of PUQ$^{\mu}$ for $L \gg 1$. Let $d_j = x_j^{seg} - x_{j-1}^{seg}$ denotes the width of the segment $[x_{j-1}^{seg}, x_j^{seg})$ and let $y_j^{seg} = (x_j^{seg} + x_{j-1}^{seg})/2$ denotes the middle of the segment, where $1 \leqslant j \leqslant L$. From the condition $L \gg 1$, it follows that the segment's width d_j is very small, so the PDF of the input data can be considered as almost constant within the segment

$[x_{j-1}^{seg}, x_j^{seg})$, i.e. $p(x,\sigma) = p(y_j^{seg},\sigma)$ for $x \in [x_{j-1}^{seg}, x_j^{seg})$; hence the segment's probability can be defined as $P_j(\sigma) = \int_{x_{j-1}^{seg}}^{x_j^{seg}} p(x,\sigma)dx = p(y_j^{seg},\sigma)\int_{x_{j-1}^{seg}}^{x_j^{seg}} dx = p(y_j^{seg},\sigma)d_j$. In addition, the slope of the compression function $c_\mu(x)$ can also be considered as nearly constant within the segment $[x_{j-1}^{seg}, x_j^{seg})$, i.e. $c'_\mu(x) = c'_\mu(y_j^{seg}) = \frac{\Delta_u}{\Delta_j}$ (Jayant and Noll, 1984), from which follows that $\Delta_j = \frac{\Delta_u}{c'_\mu(y_j^{seg})}$. Now expression (32) can be written as:

$$\begin{aligned} D_g^{\mathrm{PUQ}^\mu}(\sigma) &= 2\sum_{j=1}^{L}\frac{\Delta_j^2}{12}P_j(\sigma) = 2\frac{\Delta_u^2}{12}\sum_{j=1}^{L}\frac{p(y_j^{seg},\sigma)}{[c'_\mu(y_j^{seg})]^2}d_j \\ &\approx 2\frac{\Delta_u^2}{12}\int_0^{x_{\max}}\frac{p(x,\sigma)}{[c'_\mu(x)]^2}dx, \end{aligned} \tag{33}$$

thus concluding the proof. □

Since $L = L^{FP}$ and $L^{FP} \gg 1$, the condition of Theorem 2 is fulfilled, ensuring the closeness of the distortions of the quantizers μCQ and PUQ^μ.

Applying (8) and (10) in (30) and combining it with (31), we arrive at the closed-form expression for the total distortion of μCQ provided in Perić *et al.* (2010):

$$D^\mu(\sigma) = \underbrace{\frac{\ln^2(1+\mu)}{3N^2}\left(\left(\frac{x_{\max}}{\mu}\right)^2 + \sigma\sqrt{2}\frac{x_{\max}}{\mu} + \sigma^2\right)}_{D_g^\mu(\sigma)} + \underbrace{\sigma^2\exp\left(-\sqrt{2}\frac{x_{\max}}{\sigma}\right)}_{D_{ov}^\mu(\sigma)}. \tag{34}$$

Since $x_{\max} = x_{\max}^{FP}$, then according to (3) and (23), we have that $x_{\max}^{FP}/\mu = 2^{L^{FP}/2+1}/(2^{L^{FP}} - 1) \approx 2^{L^{FP}/2+1}/2^{L^{FP}} = 2^{-L^{FP}/2+1} \ll 1$; hence the last expression becomes:

$$D^\mu(\sigma) = \sigma^2\left(\frac{\ln^2(1+\mu)}{3N^2} + \exp\left(-\sqrt{2}\frac{x_{\max}^{FP}}{\sigma}\right)\right). \tag{35}$$

Based on (3), (18), and (23), expression (35) can be written as:

$$D^\mu(\sigma) = \sigma^2\left(\frac{1}{3\cdot 2^{2R}}\left(\frac{L^{FP}\ln 2}{1-2^{1-e}}\right)^2 + \exp\left(-\frac{2^{(L^{FP}+3)/2}}{\sigma}\right)\right). \tag{36}$$

Using (6) and (36), we derive the following final SQNR expression for μCQ:

$$\mathrm{SQNR}^\mu(\sigma) = -10\log_{10}\left(\frac{1}{3\cdot 2^{2R}}\left(\frac{L^{FP}\ln 2}{1-2^{1-e}}\right)^2 + \exp\left(-\frac{2^{(L^{FP}+3)/2}}{\sigma}\right)\right). \tag{37}$$

Based on (6), (27), and (32), knowing that for the Laplacian PDF $\int_{x_{j-1}^{seg}}^{x_j^{seg}} p(x,\sigma)dx = \frac{1}{2}(\exp(-\sqrt{2}x_{j-1}^{seg}/\sigma) - \exp(-\sqrt{2}x_j^{seg}/\sigma))$, the SQNR expression for PUQ$^{\mu}$ is delivered:

$$\begin{aligned}&\mathrm{SQNR}^{\mathrm{PUQ}^{\mu}}(\sigma)\\ &= -10\log_{10}\left[\sum_{j=1}^{L^{FP}}\frac{2^{-L^{FP}+2j-2(R-e)}}{3\sigma^2}\left(\exp\left(-\frac{2^{-(L^{FP}-3)/2}(2^{j-1}-1)}{\sigma}\right)\right.\right.\\ &\left.\left.-\exp\left(-\frac{2^{-(L^{FP}-3)/2}(2^{j}-1)}{\sigma}\right)\right)+\exp\left(-\frac{2^{(L^{FP}+3)/2}}{\sigma}\right)\right].\end{aligned} \tag{38}$$

Since provided in closed form, expression (37) exhibits substantially lower computational complexity in contrast to expressions (9) and (38). The next section will demonstrate that the numerical results yielded by (9), (37), and (38) closely align, implying that the closed-form expression (37) serves as a very precise approximation for the SQNR of the FPQ, and therefore for the performance of the FP format.

4. Numerical Results and Discussion

In this Section, we present and discuss numerical results for the derived SQNR formulas (37) and (38) obtained in evaluating the performance of the R-bit FP format with $e = 8$ (8-bit exponent) in a very wide variance range, where $R = 24$ and 32 (i.e. FP24 and FP32 formats). Note that the diversity across bit rate R is introduced to show the generality of the given formulas, whose effectiveness is measured with respect to formula (9). To facilitate the observation of variance across a wide range, it is usual to define variance in the logarithmic domain as σ^2 [dB] $= 10\log_{10}(\sigma^2/\sigma_{ref}^2)$, where σ_{ref}^2 represents the referent variance. Without loss of generality, we can assume that $\sigma_{ref}^2 = 1$, obtaining σ^2 [dB] $= 10\log_{10}\sigma^2$. Substituting $\sigma = 10^{\sigma^2\,[\mathrm{dB}]/20}$ into the previously derived expressions for SQNR yields the dependence of SQNR on σ^2 [dB].

Figure 1 shows the performance (SQNR) of the FP24 and FP32 formats over a very wide variance range σ^2 [dB] $\in$ [$-$ 500 dB, 800 dB], calculated using (9), (37), and (38). It's worth mentioning that the chosen variance range is significantly broader than that commonly used for scalar quantizer analysis (typically σ^2 [dB] $\in$ [$-$ 20 dB, 20 dB] or σ^2 [dB] $\in$ [$-$ 30 dB, 30 dB], as seen in Perić *et al.* (2010), Denić *et al.* (2023)). From the given figure, it can be noted that the results for SQNR formulas (9) and (38) are in excellent agreement for each considered σ^2 [dB]. Based on this performance matching, we argue that the discussed PUQ$^{\mu}$ and FPQ are compatible, proving the correctness of the applied design process. It can also be observed that the SQNR values achieved by (37) are very close to those achieved by (38) (and accordingly by (9)), which is in agreement with Theorem 2. From Fig. 1, it is clearly evident that there is a threshold variance, denoted by σ_t^2 [dB], such that for σ^2 [dB] $\leqslant \sigma_t^2$ [dB] the granular distortion D_g^{μ} dominates, so

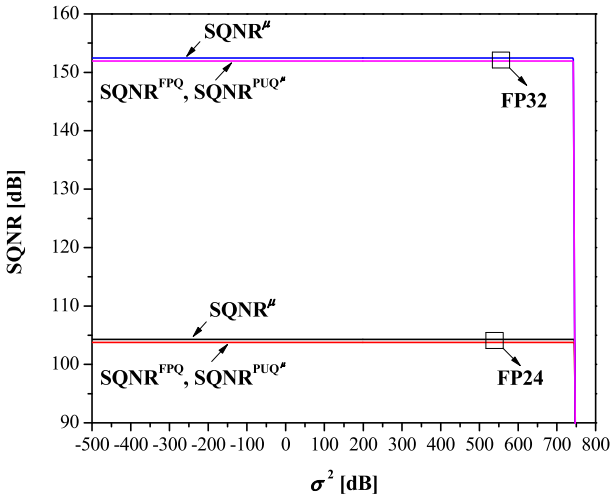

Fig. 1. Performance (SQNR) of FP24 and FP32 formats in a very wide variance range, estimated using different formulas.

SQNR$^{\mu}$ becomes:

$$\text{SQNR}^{\mu} \approx 10\log_{10}\left(\frac{\sigma^2}{D_g^{\mu}}\right) = -10\log_{10}\left(\frac{1}{3\cdot 2^{2R}}\left(\frac{L^{FP}\ln 2}{1-2^{1-e}}\right)\right) = \text{const}, \tag{39}$$

i.e. it remains constant and does not depend on the data variance σ^2. This can be interpreted as follows. Since the SQNR$^{\mu}$ is independent of the PDF parameter σ^2, then using any non-parametric Laplacian distribution yields the same SQNR score. On the other hand, for σ^2 [dB] $> \sigma_t^2$ [dB] the overload distortion D_{ov}^{μ} prevails, leading to a sharp drop in SQNR. The threshold variance σ_t^2 [dB] is 745 dB for the FP24 format and 742 dB for the FP32 format.

Let us introduce the relative error δ_{SQNR}[%] as an accuracy measure of the SQNR formula (37) with respect to (9). The values for δ_{SQNR} [%] are illustrated in Fig 2. Figure 2 indicates that the SQNR calculation error for σ^2 [dB] $\leqslant \sigma_t^2$ [dB] is below 0.5% in the case of FP24 format performance evaluation and below 0.35% in the case of FP32 format performance evaluation; for σ^2 [dB] $> \sigma_t^2$ [dB], the SQNR error tends to zero, as predicted. Given that δ_{SQNR}[%] < 1%, we report that reasonable accuracy of the SQNR formula defined in Na (2011) is achieved with the proposed approximate formula (37). Due to this achievement and the fact that (37) is considerably less computationally intensive than (9), which includes $L^{FP} = 254$ sum members (since $e = 8$), we confirm that (37) can indeed be used as an adequate tool for evaluating the performance of the FP format for Laplacian data.

Given SQNR analysis can also be useful in selecting the optimal FP format for the target application. Specifically, from the point of quality of digital representation, FP32 is a better solution than FP24, due to the higher SQNR score; however, from the point of

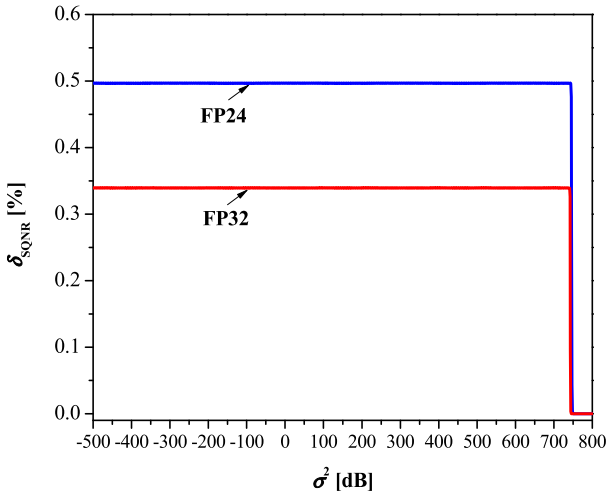

Fig. 2. Accuracy of the SQNR formula (37) in estimating the performance of FP24 and FP32 formats.

dynamic range, both FP32 and FP24 formats are very efficient as they retain constancy in SQNR across a very wide variance range. Due to these positive features and the fact that its implementation complexity is lower than FP32, FP24 can be seen as an attractive choice for various practical applications.

5. Conclusion

This paper builds upon the analogy between FP digital representation and quantization established in literature, introducing a novel idea regarding the link between the FP format and the μCQ. It presents a method for designing and linearizing the μCQ to achieve a piecewise uniform quantizer PUQ$^{\mu}$ tailored to the FP format. Given the FP format's similarity in structure to PUQ$^{\mu}$ and the close performance of PUQ$^{\mu}$ to μCQ, a closed-form expression for the SQNR of μCQ has been proposed in this paper to evaluate FP format's performance, which holds general applicability across various bit rates and data variances. Numerical assessments spanning a very wide variance range, conducted for some commonly used FP formats with an 8-bit exponent, showed the full applicability of the proposed SQNR expression in FP format performance evaluation, as competitive results (SQNR calculation error is below the predefined threshold of 1%) and significantly lower computational intensity have been observed with respect to the existing method reliant on the summation of numerous terms (254 in the situation when $e = 8$). As the computational complexity of the existing method increases even more for $e > 8$, a significant simplification of the FP format evaluation process is expected by applying the proposed method. Providing an efficient and accurate mechanism for the evaluation of FP format performance, this paper facilitates the selection of the optimal FP bit configuration for a specific application, crucial for digital representation quality, dynamic range, computational overhead, and energy efficiency.

A. Appendix

Table A1 provides an overview of abbreviations and specific symbols used in this paper.

Table A1
Employed abbreviations and symbols.

Abbreviations		Symbols	
Bfloat16	16-bit floating point format	E	exponent of a floating point number
DLfloat	16-bit floating point format	M	significand of a floating point number
FP8	8-bit floating point format	e	number of bits for exponent
FP24	24-bit floating point format	m	number of bits for significand
FP32	32-bit floating point format	R	bit rate
FP64	64-bit floating point format	E^*	biased exponent
FPQ	floating point quantizer	$E^*_{\min}$	minimal value of biased exponent
PDF	probability density function	$E^*_{\max}$	maximum value of biased exponent
PUQ$^{\mu}$	piecewise uniform quantizer based on the μ-law compression function	S_{E^*}	segment in the positive part of floating point numbers
SQNR	signal to quantization noise ratio	L^{FP}	number of segments S_{E^*}
μCQ	μ-law companding quantizer	Δ_{E^*}	step size in segment S_{E^*}
		δ_{E^*}	width of segment S_{E^*}
		$P_{E^*}(\sigma)$	probability of segment S_{E^*}
		$x^{FP}_{\max}$	maximal floating point number
		σ^2	variance of input Laplacian data
		D_g^{FPQ}	granular distortion of FPQ
		D_{ov}^{FPQ}	overload distortion of FPQ
		D^{FPQ}	total distortion of FPQ
		SQNR^{FPQ}	signal to quantization noise ratio of FPQ
		$c_\mu(x)$	μ-law compression function
		$c_\mu^{-1}(x)$	inverse μ-law compression function
		μ	compression factor
		$x_{\max}$	maximal amplitude of μCQ
		x_j^μ	decision thresholds of μCQ
		y_j^μ	representational levels of μCQ
		N	number of representational levels
		Δ_u	step size of the uniform quantizer
		$g_\mu(x)$	piecewise linear compression function
		a_j	coefficient of $g_\mu(x)$
		b_j	coefficient of $g_\mu(x)$
		L	number of segments of PUQ$^{\mu}$
		x_j^{seg}	segment thresholds of PUQ$^{\mu}$
		K	number of uniform levels within PUQ$^{\mu}$ segments
		Δ_j	step size within segment of PUQ$^{\mu}$
		$x_{j,i}$	i-th decision threshold within the j-th segment of PUQ$^{\mu}$
		$y_{j,i}$	i-th representational level within the j-th segment of PUQ$^{\mu}$
		D_g^μ	granular distortion of μCQ
		D_{ov}^μ	overload distortion of μCQ
		D^μ	total distortion of μCQ
		SQNR^μ	signal to quantization noise ratio of μCQ
		P_j	segment probability of PUQ$^{\mu}$
		$D_g^{\text{PUQ}^\mu}$	granular distortion of PUQ$^{\mu}$
		$D_{ov}^{\text{PUQ}^\mu}$	overload distortion of PUQ$^{\mu}$
		$\text{SQNR}^{\text{PUQ}^\mu}$	signal to quantization noise ratio of PUQ$^{\mu}$

Funding

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia [grant number 451-03-65/2024-03/200102], as well as by the European Union's Horizon 2023 research and innovation programme through the AIDA4Edge Twinning project (grant ID 101160293).

References

Agrawal, A., Mueller, S.M., Fleischer, B.M., Sun, X., Wang, N., Choi, J., Gopalakrishnan, K. (2019). DLFloat: a 16-b floating point format designed for deep learning training and inference. In: *Proceedings of the IEEE 26th Symposium on Computer Arithmetic (ARITH)*, Kyoto, Japan, pp. 92–95. https://doi.org/10.1109/ARITH.2019.00023.

Bai-Kui, Y., Shanq-Jang, R. (2023). Area efficient compression for floating-point feature maps in convolutional neural network accelerators. *IEEE Transactions on Circuits and Systems II*, 70(2), 746–750. https://doi.org/10.1109/TCSII.2022.3213847.

Banner, R., Nahshan, Y., Hoffer, E., Soudry, D. (2018). *ACIQ: Analytical Clipping for Integer Quantization of Neural Networks*. arXiv preprint, arXiv:1810.05723.

Banner, R., Nahshan, Y., Soudry, D. (2019). Post training 4-bit quantization of convolutional networks for rapid-deployment. In: *Proceedings of the33rd Conference on Neural Information Processing Systems, (NeurIPS), No. 714*, Vancouver, BC, Canada, pp. 7950–7958.

Botta, M., Cavagnino, D., Esposito, R. (2021). NeuNAC: a novel fragile watermarking algorithm for integrity protection of neural networks. *Information Sciences*, 576, 228–241. https://doi.org/10.1016/j.ins.2021.06.073.

Burgess, N., Goodyer, C., Hinds, C.N., Lutz, D.R. (2019a). High-precision anchored accumulators for reproducible floating-point summation. *IEEE Transactions on Computers*, 68(7), 967–978. https://doi.org/10.1109/TC.2018.2855729.

Burgess, N., Milanovic, J., Stephens, N., Monachopoulos, K., Mansell, D. (2019b). Bfloat16 processing for neural networks. In: *Proceedings of the IEEE 26th Symposium on Computer Arithmetic*, ARITH 2019, Kyoto, Japan, June, pp. 10–12. https://doi.org/10.1109/ARITH.2019.00022.

Cattaneo, D., Di Bello, A., Cherubin, S., Terraneo, F., Agosta, G. (2018). Embedded operating system optimization through floating to fixed point compiler transformation. In: *Proceedings of the 21-st Euromicro Conference on Digital System Design (DSD)*, Prague, Czech Republic, pp. 172–176.

Chu, W.C. (2003). *Speech Coding Algorithms: Foundation and Evolution of Standardized Coders*. John Wiley & Sons, New Jersey.

Denić, B., Perić, Z., Dinčić, M. (2023). Improvement of the Bfloat16 floating-point for the Laplacian source. In: *Proceedings of the IEEE 13th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, Bucharest, Romania, pp. 1–4 . https://doi.org/10.1109/ATEE58038.2023.10108130.

Dinčić, M., Perić, Z., Tančić, M., Denić, D., Stamenković, Z., Denić, B. (2021). Support region of μ-law logarithmic quantizers for Laplacian source applied in neural networks. *Microelectronics Reliability*, 124, 114269.

Dinčić, M., Perić, Z., Jovanović, A. (2016). New coding algorithm based on variable-length codewords for piecewise uniform quantizers. *Informatica*, 27(3), 527–548. https://doi.org/10.15388/Informatica.2016.98.

Fasi, M., Mikaitis, M. (2021). Algorithms for stochastically rounded elementary arithmetic operations in IEEE 754 floating-point arithmetic. *IEEE Transactions on Emerging Topics in Computing*, 9(3), 1451–1466. https://doi.org/10.1109/TETC.2021.3069165.

Gazor, S., Zhang, W. (2003). Speech probability distribution. *IEEE Signal Processing Letters*, 10(7), 204–207. https://doi.org/10.1109/LSP.2003.813679.

Gersho, A., Gray, R. (1992). *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, New York.

IEEE 754 (2019). *IEEE Standard for Floating Point Arithmetic*.

Jayant, N.C., Noll, P. (1984). *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, New Jersey.

Junaid, M., Arslan, S., Lee, T., Kim, H. (2022). Optimal architecture of floating-point arithmetic for neural network training processor. *Sensors*, 22, 1230. https://doi.org/10.3390/s22031230.

Moroz, L., Samotyy, V. (2019). Efficient floating-point division for digital signal processing application. *IEEE Signal Processing Magazine*, 36(1), 159–163. https://doi.org/10.1109/MSP.2018.2875977.

Na, S. (2011). Asymptotic formulas for variance-mismatched fixed-rate scalar quantization of a Gaussian source. *IEEE Transactions on Signal Processing*, 59(5), 2437–2441. https://doi.org/10.1109/TSP.2011.2112354.

Perić, Z., Dinčić, M., Denić, D., Jocić, A. (2010). Forward adaptive logarithmic quantizer with new lossless coding method for Laplacian source. *Wireless Personal Communications*, 59(4), 625–641. https://doi.org/10.1007/s11277-010-9929-3.

Perić, Z., Savić, M., Dinčić, M., Vučić, N., Djošić, D., Milosavljević, S. (2021). Floating point and fixed point 32-bits quantizers for quantization of weights of neural networks. In: *Proceedings of the IEEE 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, Bucharest, Romania, pp. 1–4. https://doi.org/10.1109/ATEE52255.2021.9425265.

Syed, R.T., Ulbricht, M., Piotrowski, K., Krstic, M. (2021). Fault resilience analysis of quantized deep neural networks. In: *Proceedings of the IEEE 32nd International Conference on Microelectronics (MIEL)*, Niš, Serbia, pp. 275–294. https://doi.org/10.1109/MIEL52794.2021.9569094.

Wang, N., Choi, J., Brand, D., Chen, C.Y., Gopalakrishnan, K. (2018). Training deep neural networks with 8-bit floating point numbers. In: *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 2018, pp. 7686–7695.

Yang, Y., Chi, X., Deng, L., Yan, T., Gao, F., Li, G. (2022). Towards efficient full 8-bit integer DNN online training on resource-limited devices without batch normalization. *Neurocomputing*, 511, 175–186. https://doi.org/10.1016/j.neucom.2022.08.045.

Zhao, W., Dang, Q., Xia, T., Zhang, J., Zheng, N., Ren, P. (2023). Optimizing FPGA-based DNN accelerator with shared exponential floating-point format. *IEEE Transactions on Circuits and Systems I*, 70(11), 4478–4491. https://doi.org/10.1109/TCSI.2023.3300657.

Z. Perić was born in Niš, Serbia, in 1964. He received the BS, MS and PhD degrees from the Faculty of Electronic Engineering, University of Niš, Serbia, in 1989, 1994 and 1999, respectively. He is a full-time professor at Department of Telecommunications, Faculty of Electronic Engineering, University of Niš. His current research interests include the information theory and signal processing. He is an author and co-author of over 350 papers. Dr. Peric has been a reviewer of a number of journals, including *IEEE Transactions on Information Theory*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Communications*, *Compel*, *Informatica*, *Information Technology and Control*, *Expert Systems with Applications and Digital Signal Processing*.

B. Denić received his PhD degree in the field of Telecommunications in 2023 from the Faculty of Electronic Engineering, University of Niš, Serbia. Currently, he is working as a research associate at the same faculty. His main research interests include signal processing, quantization and machine learning. He is an author of 34 scientific papers (18 of them in peer-reviewed international journals).

M. Dinčić received MSc in 2007, PhD in the field of Telecommunication in 2012 and PhD in the field of Measurements in 2017 from the University of Niš. Currently, he is working as an associate professor at the Faculty of Electronic Engineering. He is an author of 64 scientific papers (35 of them in reputable international journals with IF from the SCI/SCIe list). His research is related to quantization and compression of neural networks, sensors and measurement systems.

S. Perić received her BS and MS degrees in Control Systems from the Faculty of Electronic Engineering, University of Niš, Serbia, in 2022 and 2023, respectively. She is currently pursuing her PhD degree in Telecommunications at the same institution. Her research interests include quantization and machine learning.