AIDA4Edge

University of Niš
Faculty of Electronic Engineering

Neural Network Quantization

Milan Dinčić

University of Niš, Serbia

2nd Scientific Workshop on Fault-Tolerant Design and Real-Time Reliability Monitoring

July 15, 2026

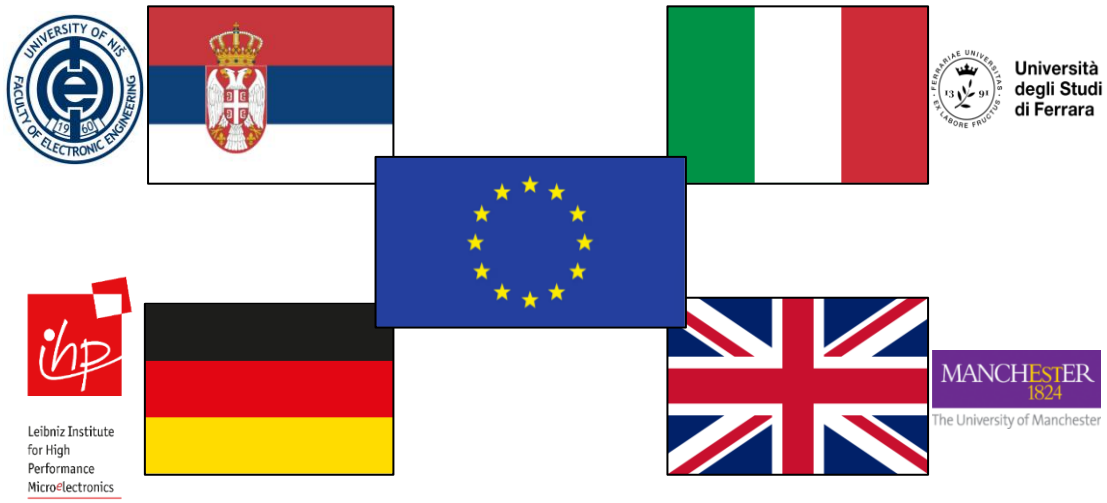
Volos – Greece

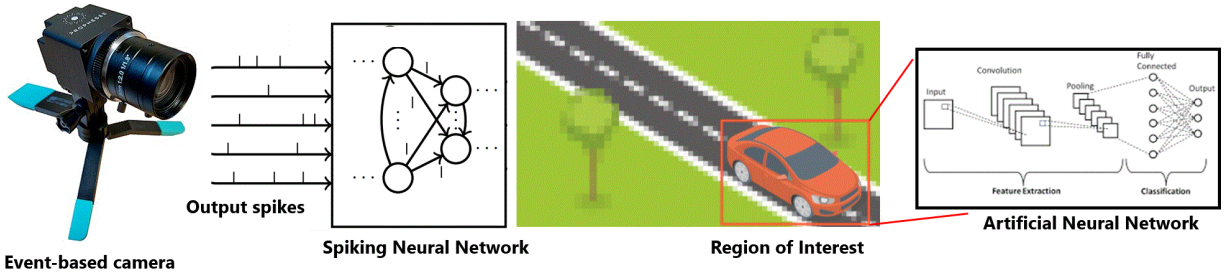
AIDA4Edge Horizon Twinning Project

AIDA4Edge

Project title: Twinning for Excellence in Adaptive Edge AI
Time period: 1 October 2024 - 30 September 2027

- **To develop low-complexity AI models suitable for deployment on edge devices with limited memory, processing power and energy.**
  - Efficient quantization
  - Hardware-aware Neural architecture search
  - Spiking neural networks
  - Hybrid ANN+SNN models
  - Efficient and reliable self-adaptive AI accelerators





Content

- Quantization
- Digital data formats
  - Fixed-point formats
  - Integer formats
  - Floating-point formats
- Quantization of Neural networks
  - Efficiency
  - Reliability

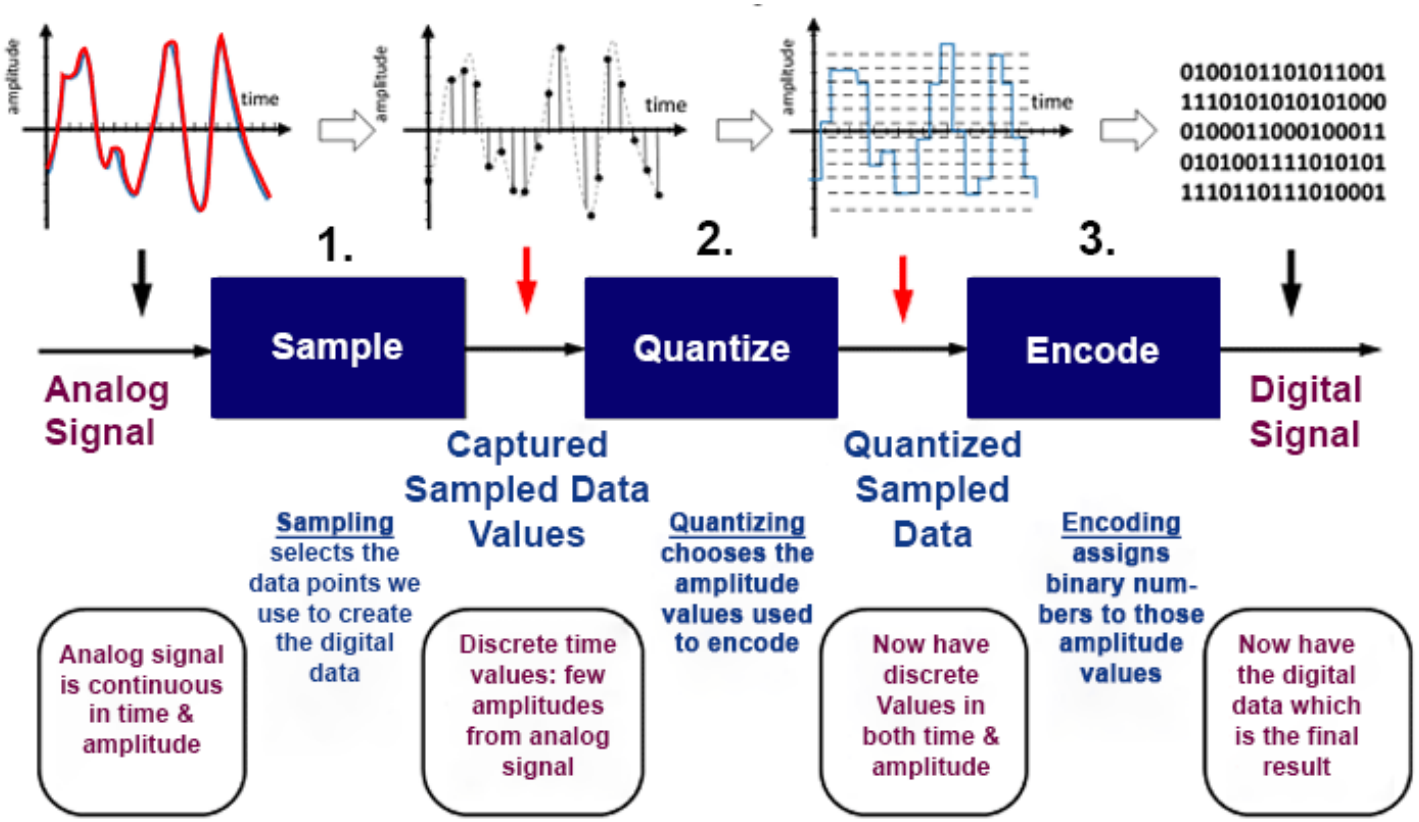
Content

- **Quantization**
- Digital data formats
  - Fixed-point formats
  - Integer formats
  - Floating-point formats
- Quantization of Neural networks
  - Efficiency
  - Reliability

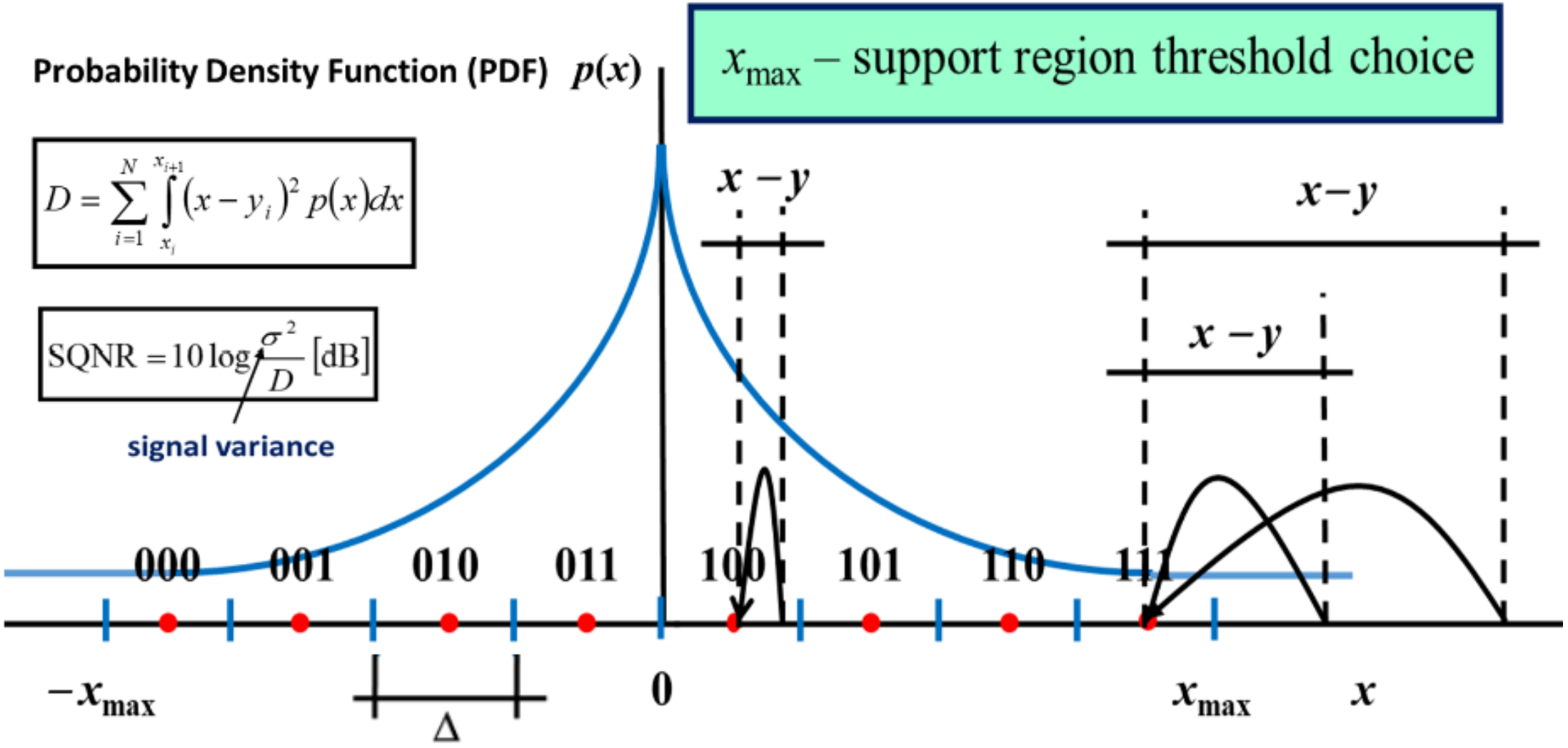
Quantization

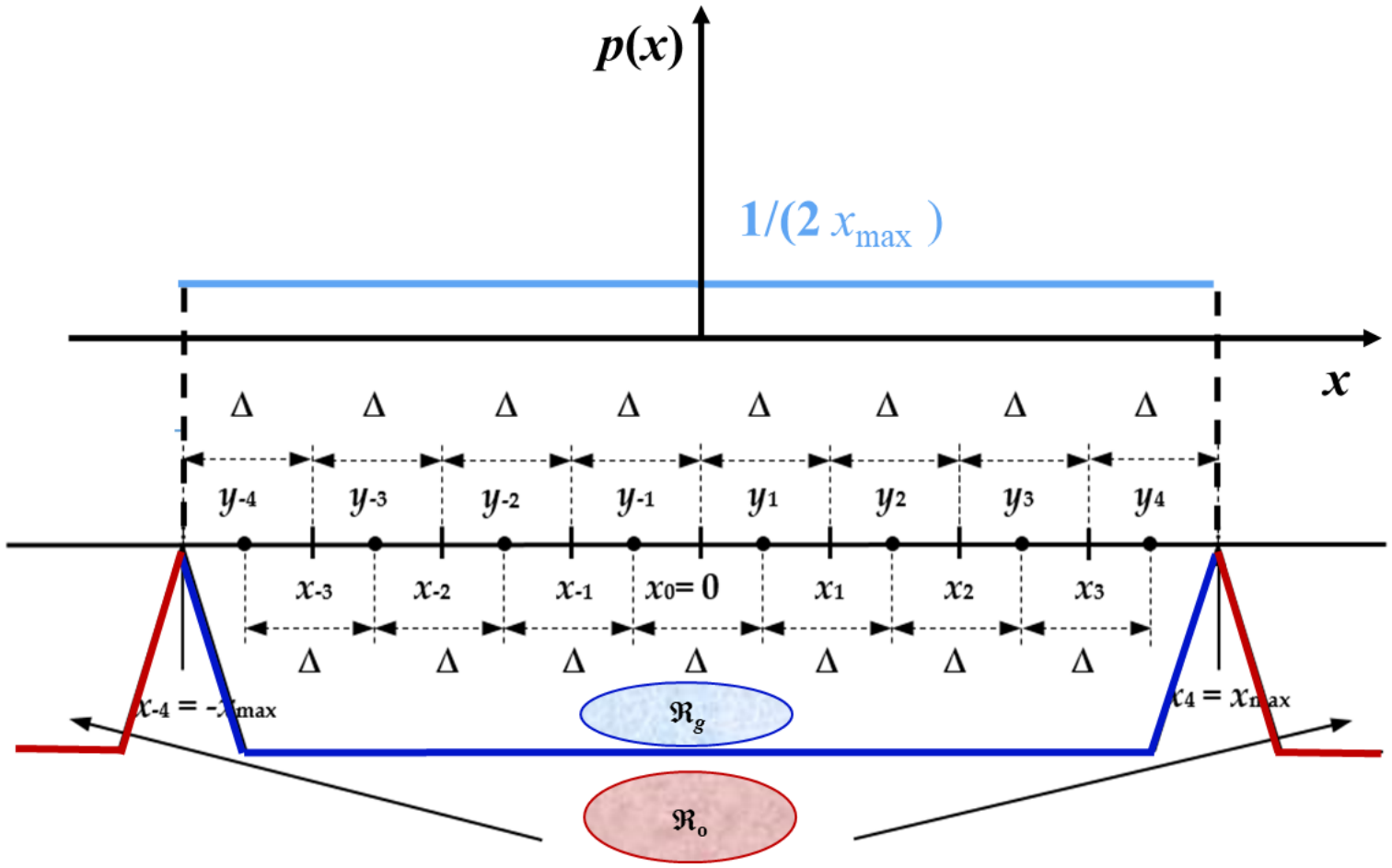
- **Mapping:**
  - $\mathbb{R} \rightarrow \{y_1, y_2, \dots, y_N\}$
  - $\{y_1, y_2, \dots, y_M\} \rightarrow \{y_1, y_2, \dots, y_N\}, M \gg N$

- **Types of quantizers:**
  - **Uniform**
  - **Nonuniform**
  - **Piecewise-uniform**

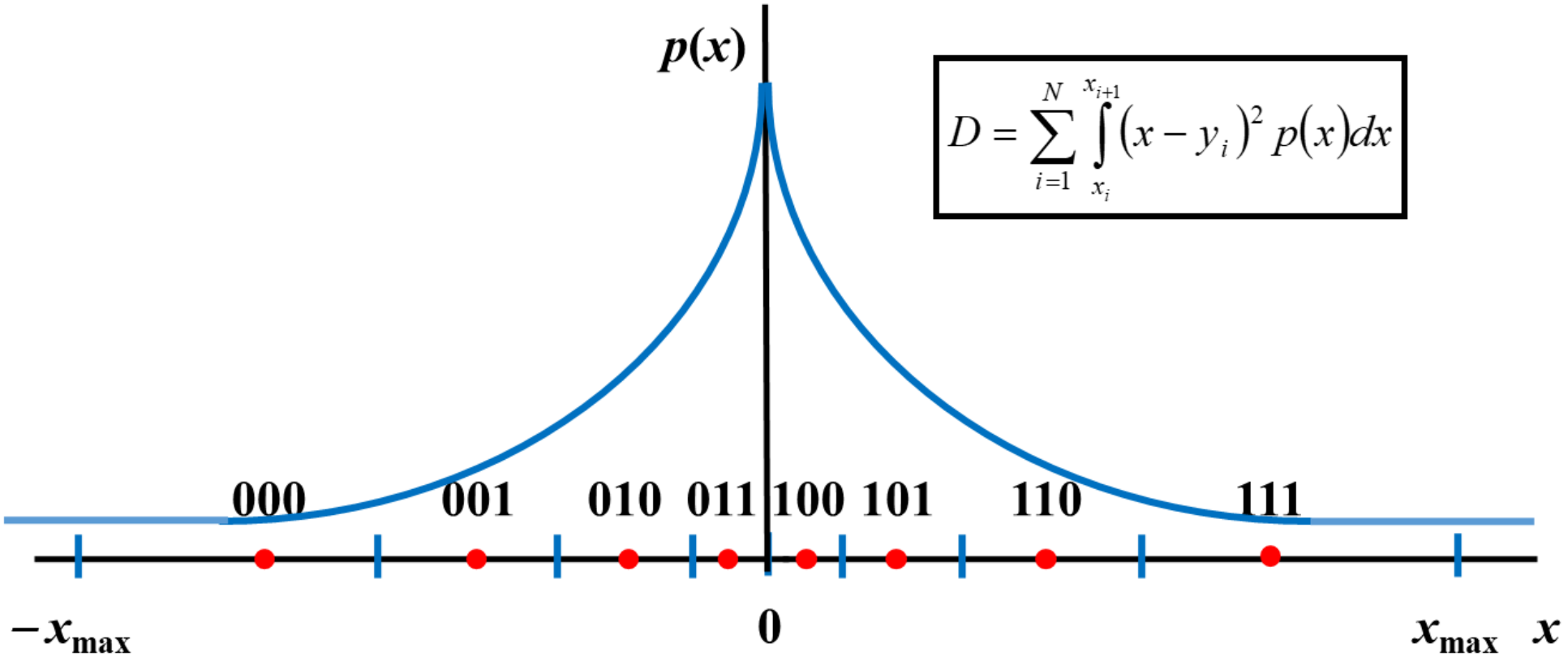


Uniform quantization

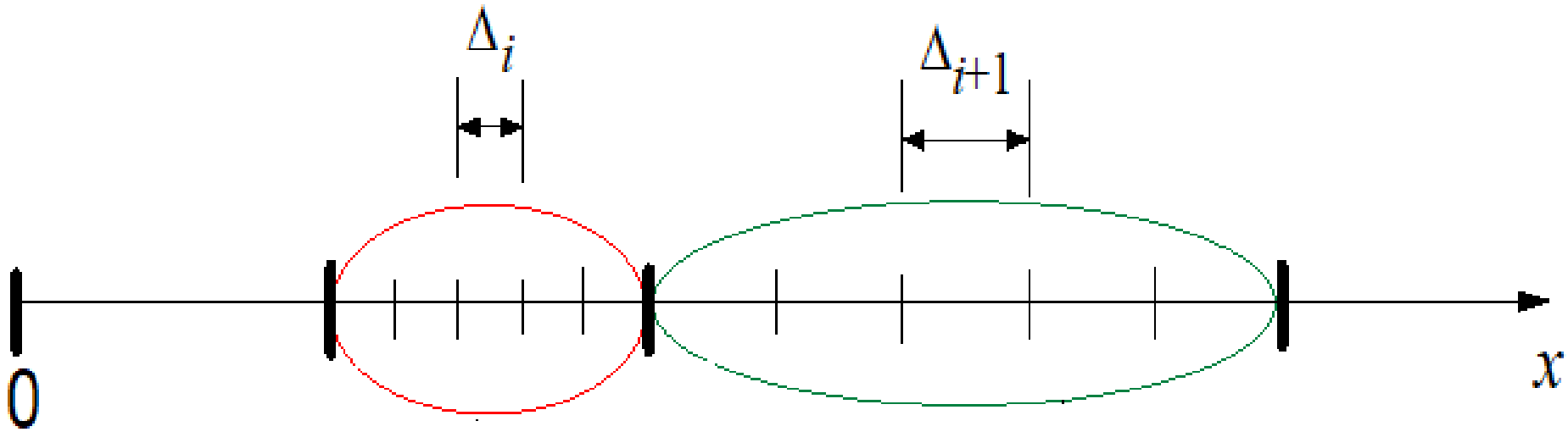



p(x)
1/(2 x_max)
x
Δ
y-4
y-3
y-2
y-1
y1
y2
y3
y4
x-3
x-2
x-1
x0= 0
x1
x2
x3
x-4 = -xmax
x4 = xmax
ℜg
ℜo

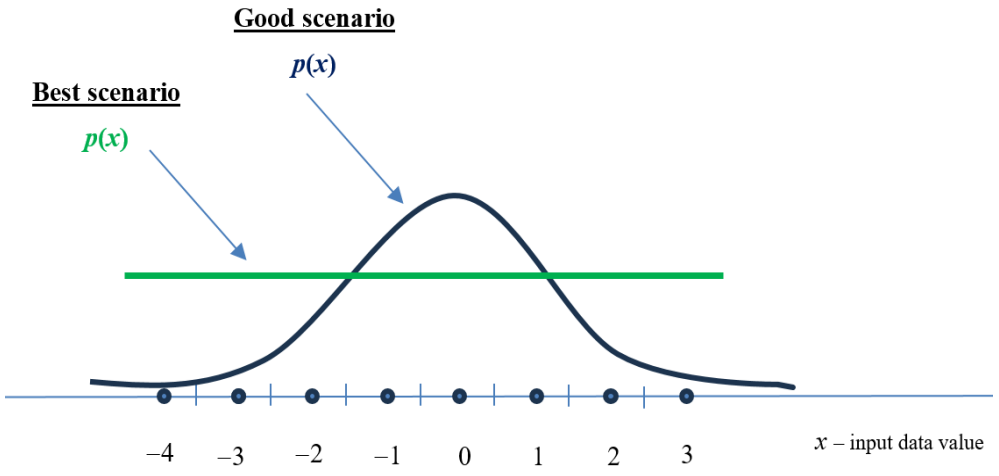
Non-uniform quantization

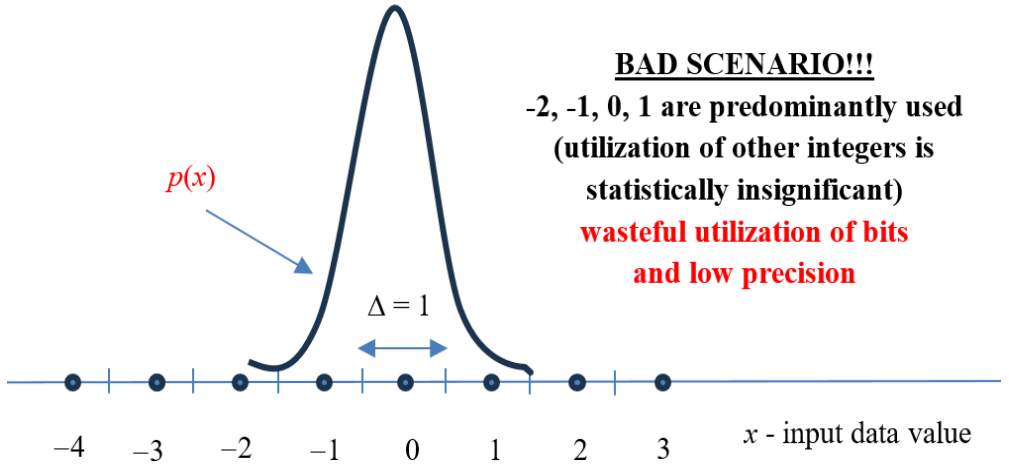


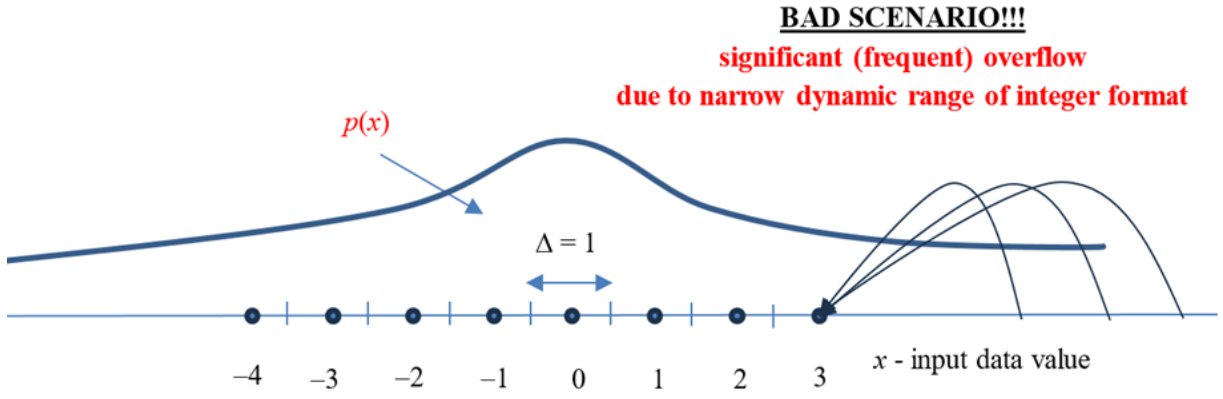
Piecewise uniform quantization



Quantizer Alignment with the Data PDF







Content

- Quantization
- Digital data formats
  - Fixed-point formats
  - Integer formats
  - Floating-point formats
- Quantization of Neural networks
  - Efficiency
  - Reliability

Fixed-point format

$$x = s a_{n-1} \ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-m}$$

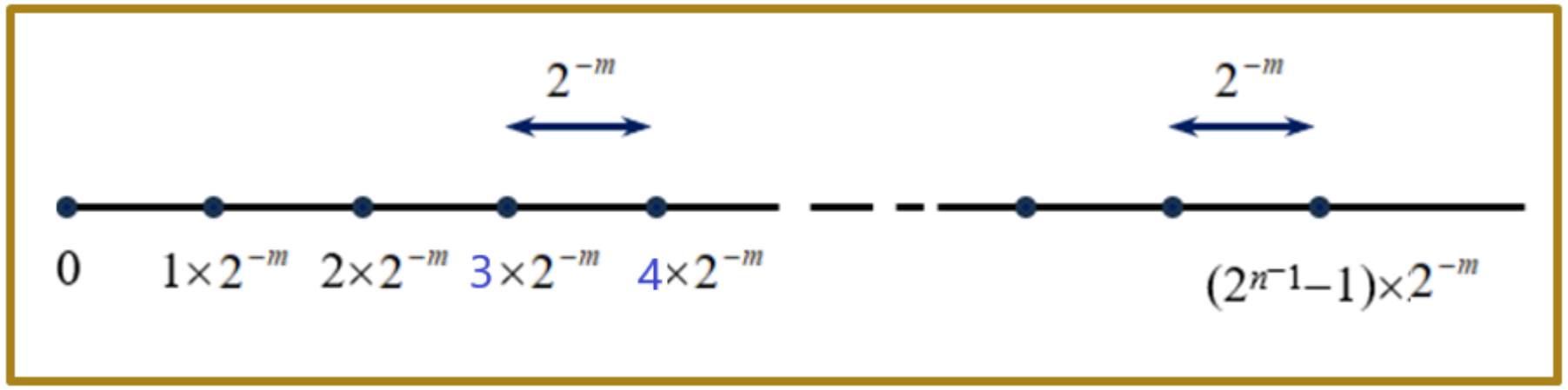
$$x = (-1)^s \sum_{i=-m}^{n-1} a_i 2^i$$

$$(0\ldots0.0\ldots01)_2 = 2^{-m}$$

$$(0\ldots0.0\ldots010)_2 = 2^{-(m-1)} = 2 \cdot 2^{-m}$$

$$(0\ldots0.0\ldots011)_2 = 2^{-(m-1)} + 2^{-m} = 3 \cdot 2^{-m}$$

$$(0\ldots0.0\ldots0100)_2 = 2^{-(m-2)} = 4 \cdot 2^{-m}$$



Fixed-point format ↔ Uniform quantizer

Integer format

fixed-point format: $x = sa_{n-1}\ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-m}$

for $m = 0$: $x = sa_{n-1}\ldots a_1 a_0$

The integer format is a fixed-point format with $m = 0$.

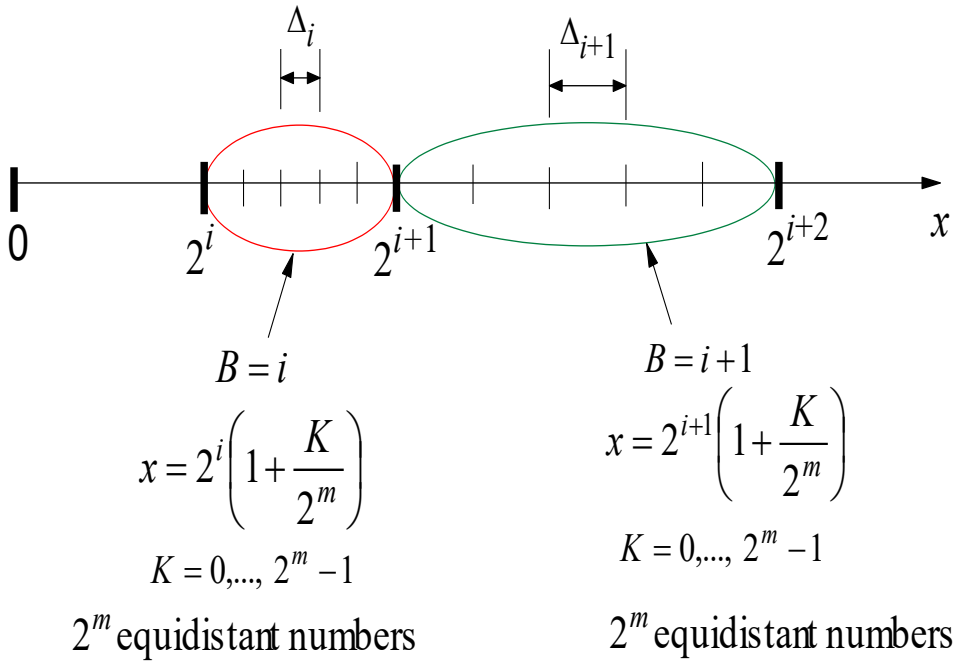
Integer format ↔ Uniform quantizer

Floating-point format

$\boldsymbol{x = (s a_{e-1} \ldots a_1 a_0 b_{m-1} \ldots b_1 b_0)_2}$

- one bit *s* to encode the sign of *x*,
- *e* bits to encode the exponent $E = \sum_{i=0}^{e-1} a_i 2^i$
- *m* bits to represent the significand $K = \sum_{i=1}^{m} b_{m-i} 2^{m-i}$
- biased exponent: $B = E - bias$

$$\boldsymbol{x = 2^B \left(1 + \frac{K}{2^m}\right)}$$



Floating-point format ↔ Piecewise uniform quantizer

Conclusion: Different digital data formats represent different types of quantizers.

$$\mathrm{SQNR}(\alpha) = -10\log_{10}\left[\frac{2^{2(B_{\min}-m)}}{12\cdot 10^{\alpha/10}}\,\mathrm{erf}\left(\frac{2^{B_{\min}-1/2}}{10^{\alpha/20}}\right) + \sum_{B=B_{\min}}^{B_{\max}} \frac{2^{2(B-m)}}{12\cdot 10^{\alpha/10}}\left(\mathrm{erf}\left(\frac{2^{B+1/2}}{10^{\alpha/20}}\right) - \mathrm{erf}\left(\frac{2^{B-1/2}}{10^{\alpha/20}}\right)\right)\right.$$

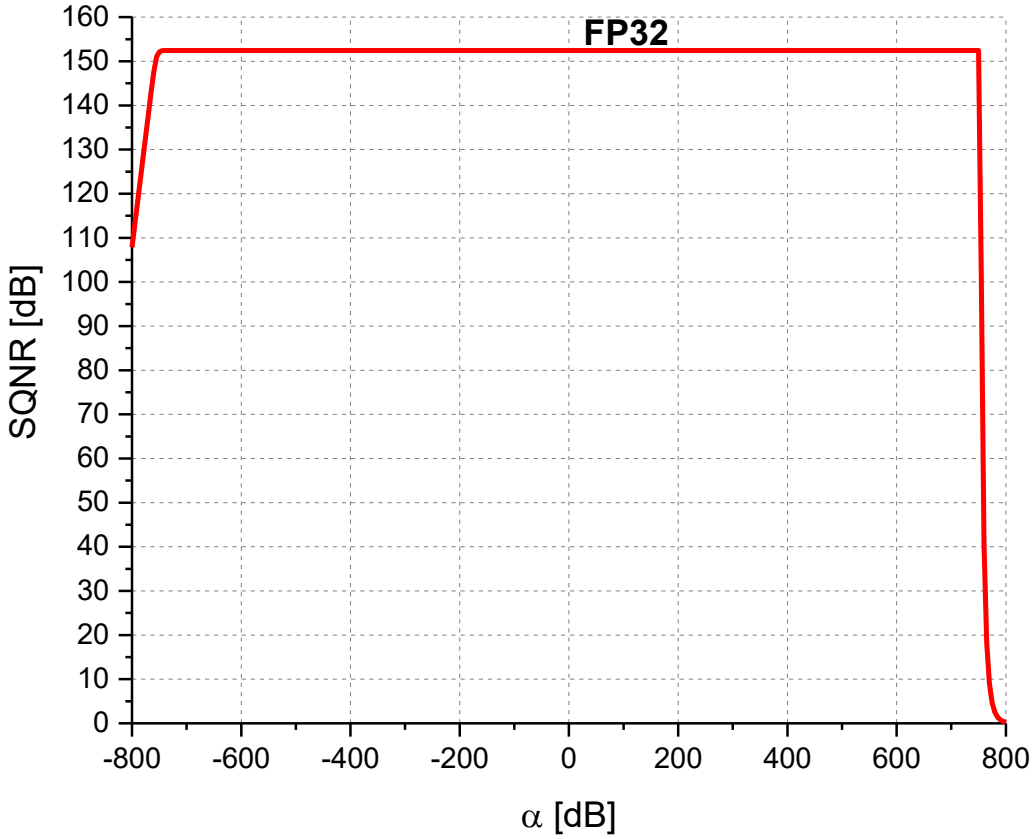
$$\left. - \frac{x_{\max}}{10^{\alpha/20}}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{x_{\max}^2}{2\cdot 10^{\alpha/10}}\right) + \left(\frac{x_{\max}^2}{10^{\alpha/10}} + 1\right)\mathrm{erfc}\left(\frac{x_{\max}}{10^{\alpha/20}\cdot\sqrt{2}}\right)\right]$$

Content

- Quantization
- Digital data formats
  - Fixed-point formats
  - Integer formats
  - Floating-point formats
- **Quantization of Neural networks**
  - **Efficiency**
  - **Reliability**

Quantization of neural networks

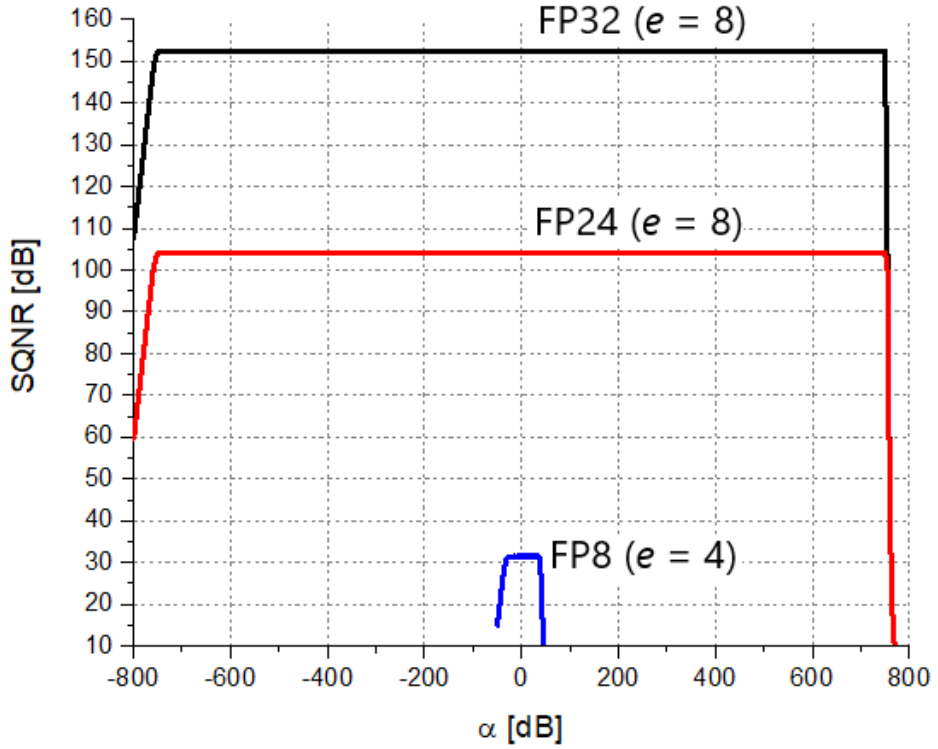
- The 32-bit floating-point format (FP32) is commonly used to represent DNN weights and input data.
- FP32:
  - Advantages: very high precision, very wide dynamic range
  - Disadvantages: very high complexity



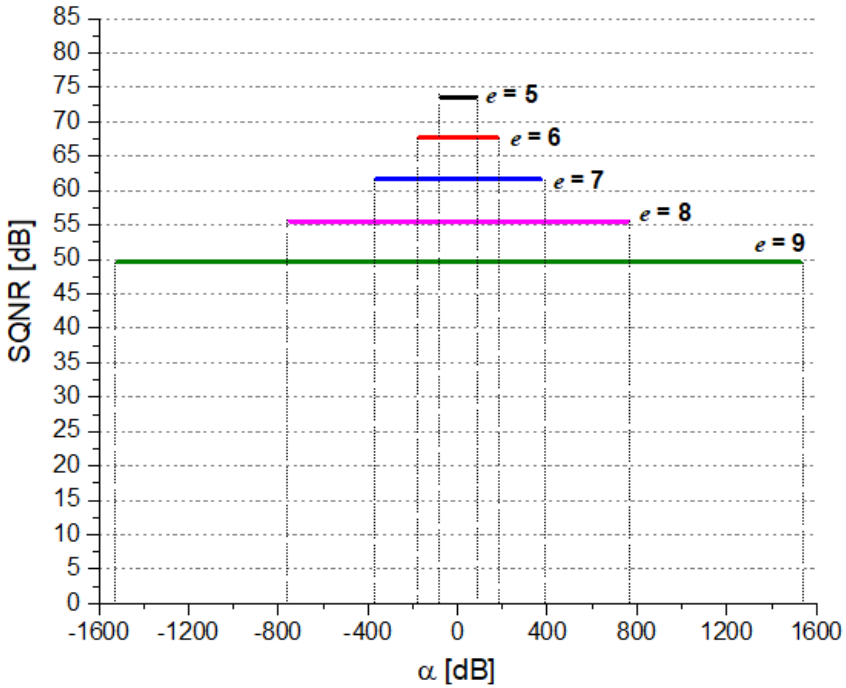
- Approaches to reduce complexity:
  - Low-resolution floating-point (FP) formats (e.g., 16-bit, 8-bit, etc.)
  - Fixed-point (FxP) formats
  - Integer formats
  - Power-of-two quantizer

Floating-point formats with lower complexity than FP32

Format	Sign	Exponent	Fraction
FP32	1	8	23
TF32	1	8	10
FP16	1	5	10
bfloat16	1	8	7

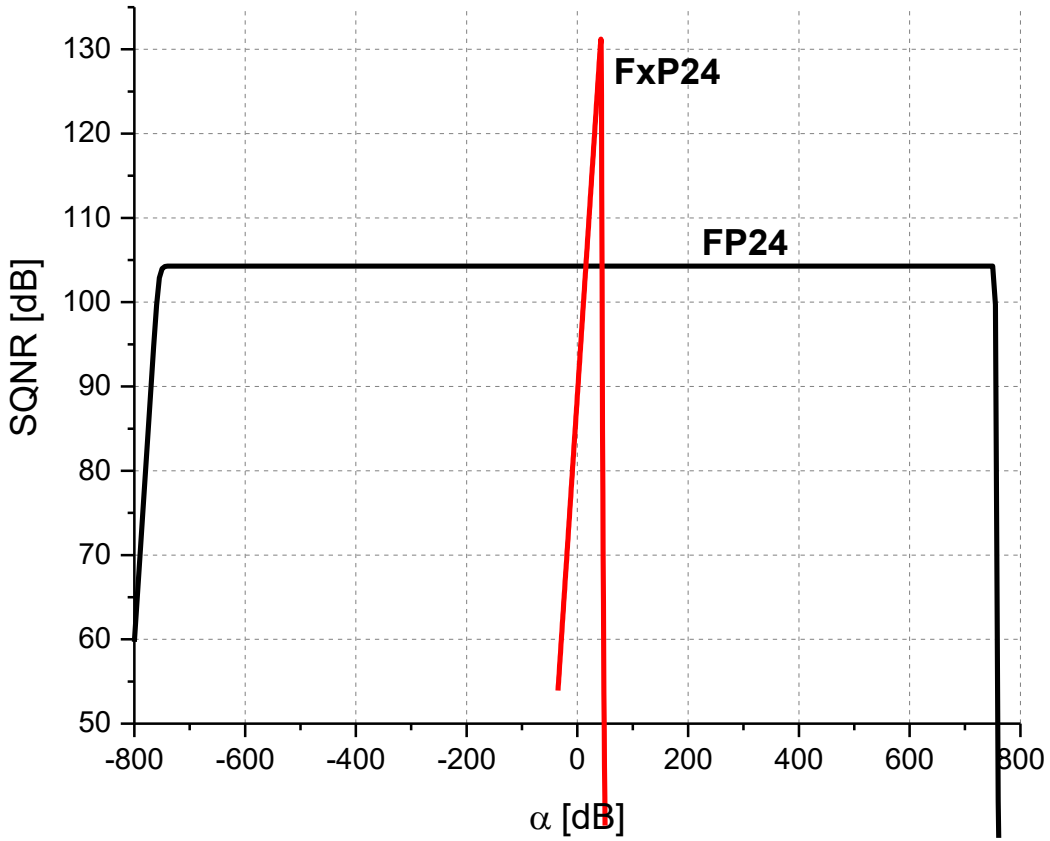


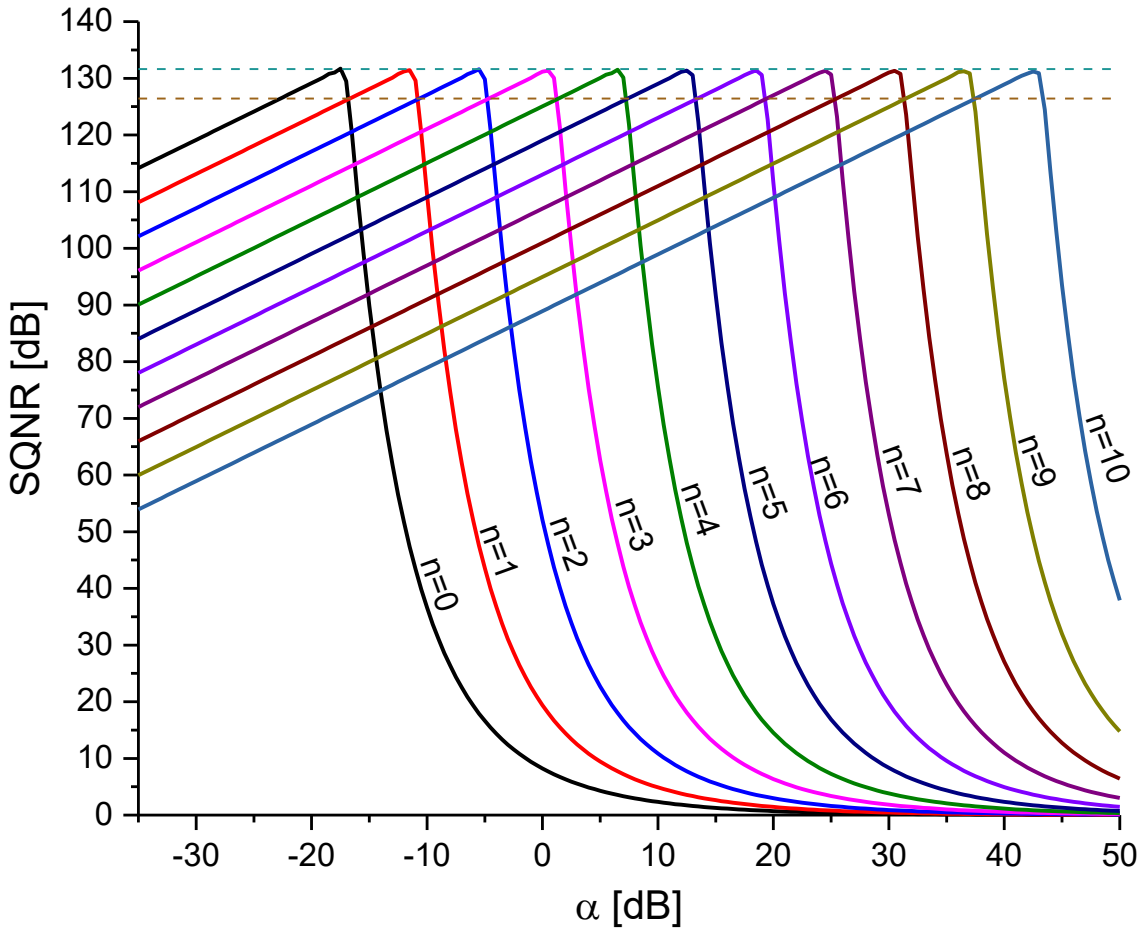
Custom FP Formats:



Fixed-point formats

$$x = sa_{n-1}\ldots a_1 a_0 . a_{-1} a_{-2} \ldots a_{-m}$$

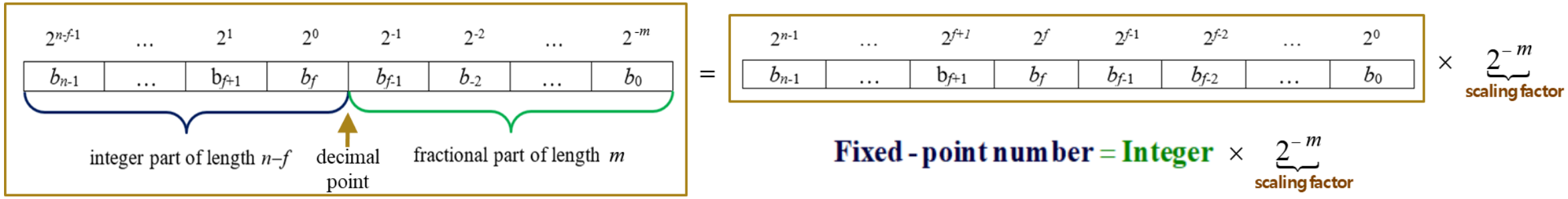




Integer formats

- Direct use (m = 0): Usually poor results

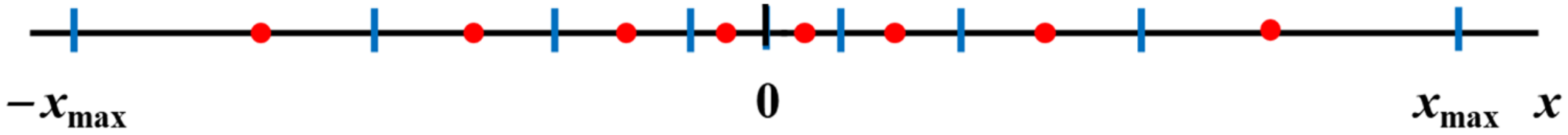
- Indirect use: A fixed-point number can be interpreted as an integer multiplied by a fixed scaling factor.



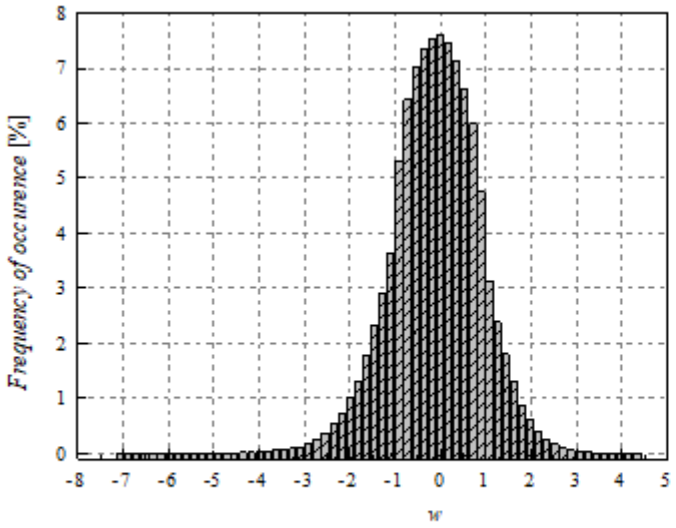
$$\text{Fixed-point number} = \text{Integer} \times \underbrace{2^{-m}}_{\text{scaling factor}}$$

Power-of-Two (PoT) quantization

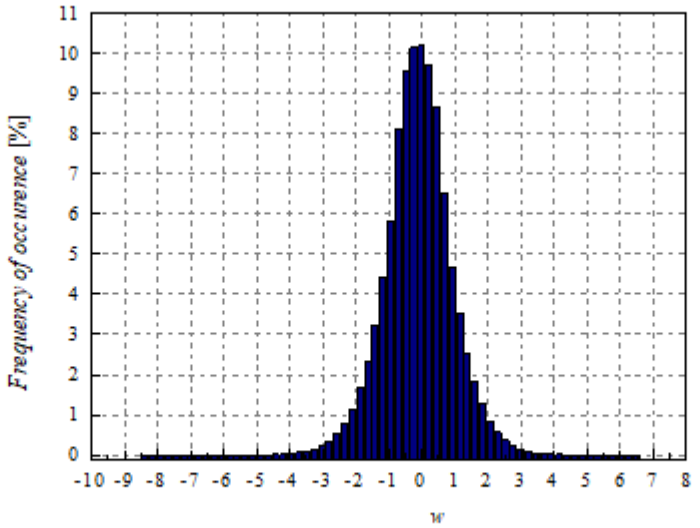
- Non-uniform quantizer
- Representation values are powers of two ($q \pm 2k$)
- Multiplication is replaced by bit-shift operations



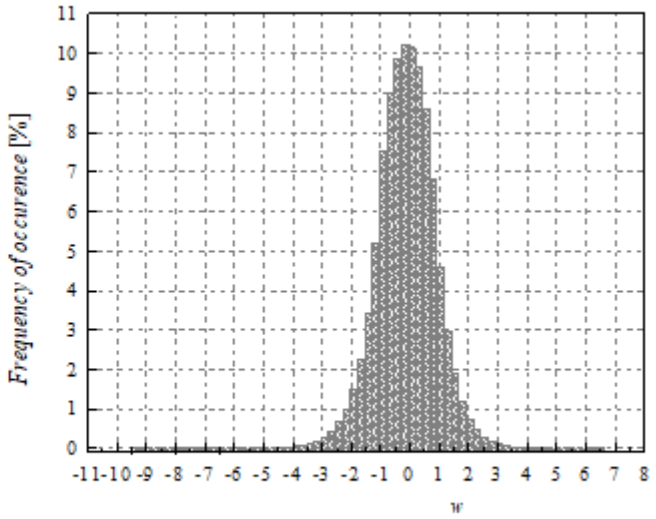
Weights PDF



MLP+MNIST

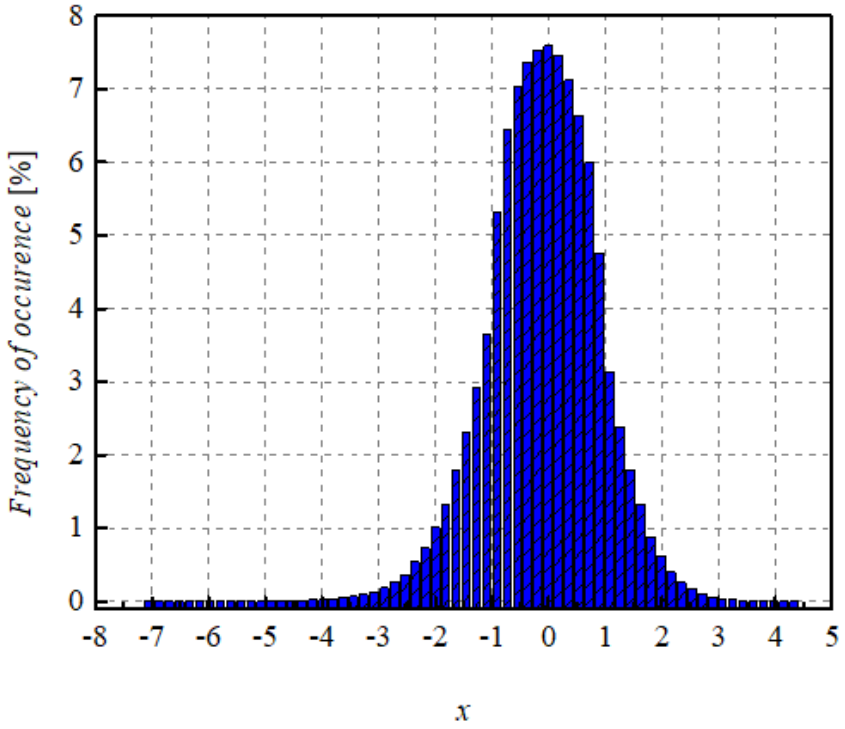


CNN+MNIST

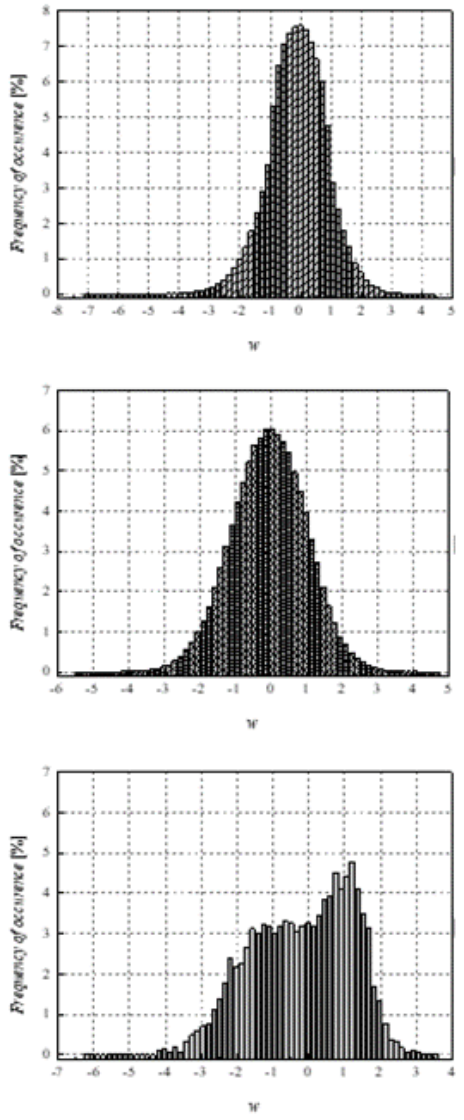


MLP + Fashion-MNIST

All weights



Weights across layers



Quantization and Reliability

- Although the impact of quantization on neural network reliability remains an active area of research and many open questions still exist, quantization has the potential to significantly improve system reliability:
- Reduces the memory area required for storing network parameters.
- Enables memory redundancy (e.g., storing each weight multiple times) without increasing the overall memory footprint.
- Fixed-point representations are inherently more robust to hardware faults than floating-point formats.
- In low-resolution fixed-point formats, the effect of a single-bit error on the represented value is typically smaller, reducing the impact of memory faults on inference accuracy.

Conclusion

- Quantization plays a crucial role in determining both the computational efficiency and the reliability of neural networks.
- The numerical formats used to represent neural network weights, activations, and input data can be interpreted as different types of quantizers.
- Although FP32 remains the standard representation, its high computational and hardware complexity makes it inefficient for Edge AI applications.
- Lower-precision numerical formats offer a more efficient alternative while significantly reducing implementation cost.
- Low-bit fixed-point formats are particularly attractive; however, their parameters must be carefully optimized. Otherwise, variance mismatch can severely degrade network accuracy.
- Quantization theory provides a principled framework for designing application-specific numerical formats, with the statistical characteristics of network weights and activations playing a key role in selecting the optimal representation.

International Symposium on Efficient and Reliable Intelligent Edge Systems

- **Date:** 25–26 May 2027
- **Place:** Niš, Serbia
- **Submission deadline:** 25 January 2027
- **General Co-Chairs:**
  Milan Dinčić (University of Niš)
  Davide Bertozzi (University of Manchester)
  Miloš Krstić (IHP Germany)
- **Program Committee Co-Chairs:**
  Zoran Perić (University of Niš)
  Riccardo Zese (University of Ferrara)
  Oliver Rhodes (University of Manchester)
  Marko Andjelković (IHP Germany)
- **Submission options:** Authors may submit either a full paper or an abstract only.
- **Journal publications:**
  - up to **5 selected papers** published free of charge in **JLPEA**,
  - up to **10 selected papers** published in **Facta Universitatis**.
- **Registration fee:** €150 for regular participants (both with a full paper or an abstract only) and €100 for PhD students.

- **Topics:**
  - Optimization techniques for Edge AI models
  - Neural architecture search for Edge AI models
  - Quantization and compression of neural networks
  - Approximate computing
  - Dynamic neural network models and implementations
  - Neuromorphic computing for Edge AI
  - Neuromorphic vision
  - Spiking neural networks
  - Hybrid ANN–SNN models
  - In-memory and near-memory computing
  - Reliable electronics for Edge AI
  - Hardware accelerators for Edge AI
  - Hardware/software co-exploration for neural architectures
  - Edge AI applications

THANKS!