Run-time Non-uniform Quantization for Dynamic Neural Networks in Wireless Communication

Priscilla Sharon Allwin, Manil Dev Gomony, Marc Geilen
Eindhoven University of Technology
Netherlands

29th Asia South Pacific Design Automation Conference

January 25th, 2024, South Korea

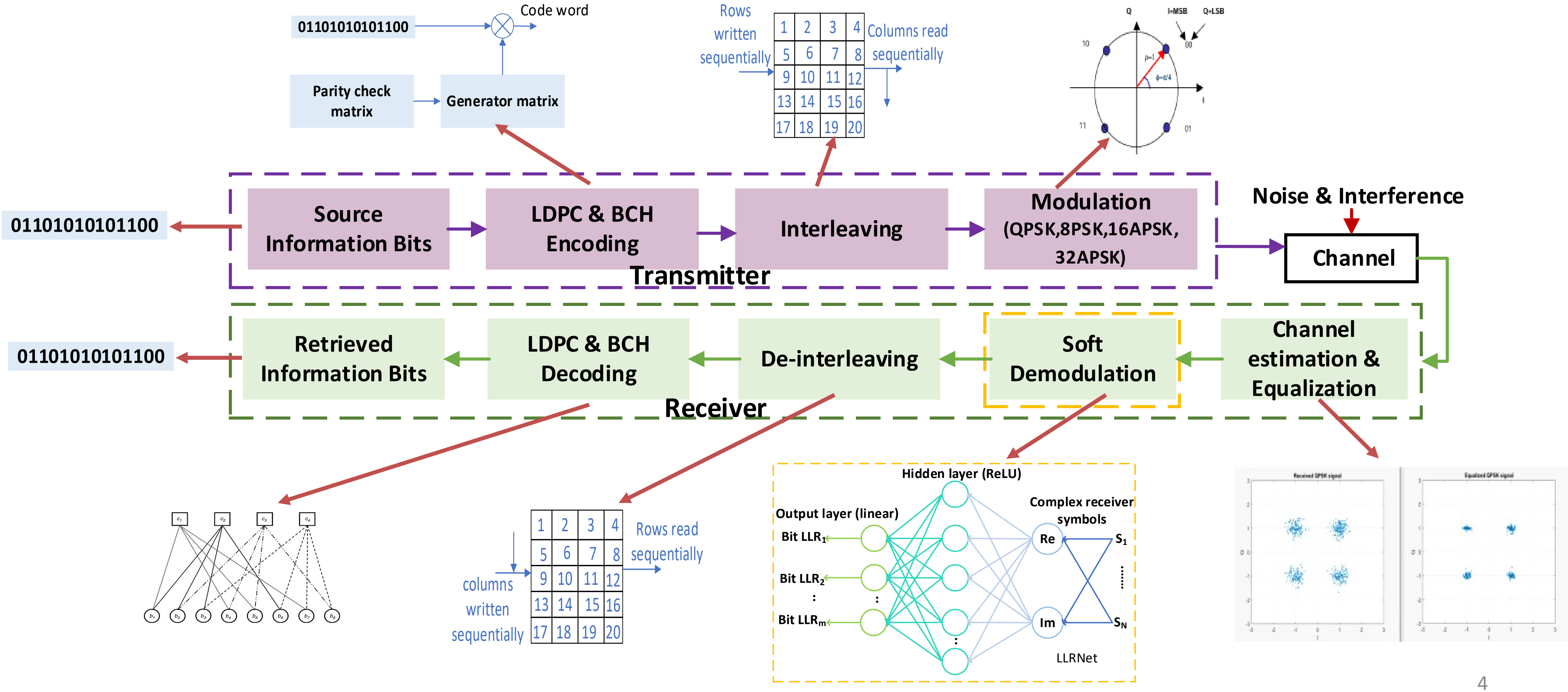
Outline

Outline

- Neural Networks in Wireless Communication
- Dynamic Neural Networks (DyNN)
- Research Problem
- Proposed Work & Results
- Conclusion and Future Work
- Q&A

Neural Networks in Wireless Communication

Neural Networks (NN) are replacing conventional modules in wireless communication, indicating a shift towards more advanced and adaptive technologies.



Outline

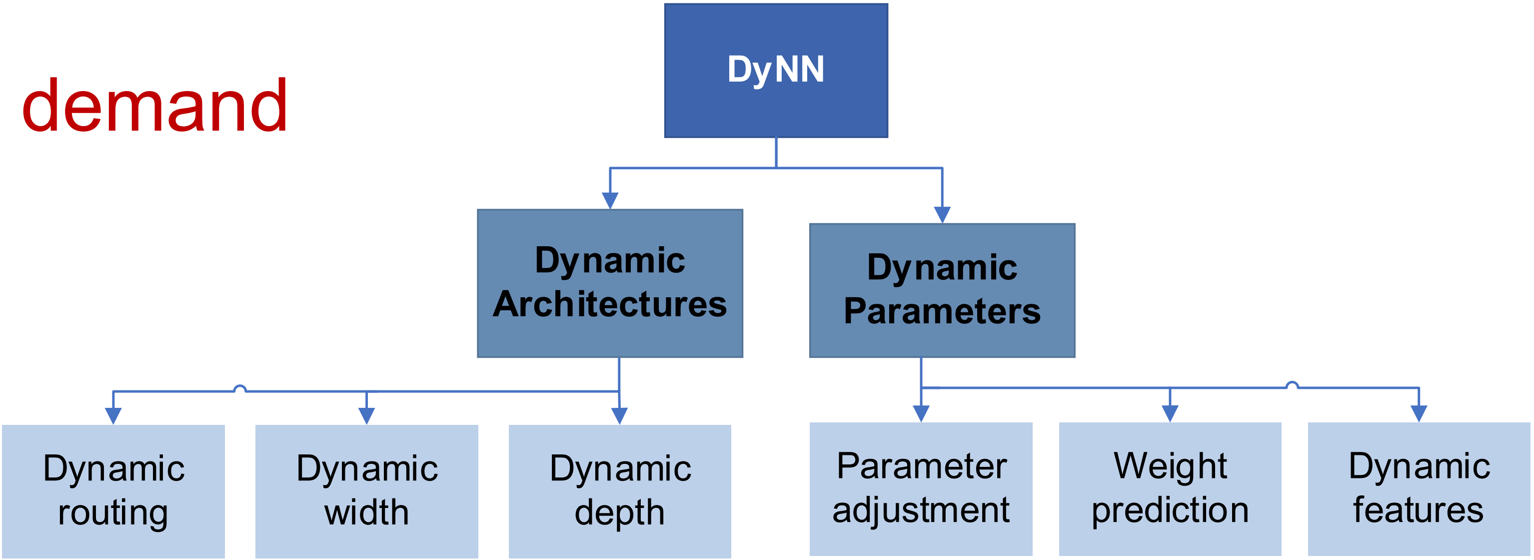
- Neural Networks in Wireless Communication
- Dynamic Neural Networks (DyNN)
- Research Problem
- Proposed Work & Results
- Conclusion and Future Work
- Q&A

Dynamic Neural Networks

Dynamic Neural Networks (DyNN) are gaining popularity over static NNs.

A DyNN can adapt: On demand

- Architecture
- Parameter



The LLRNet can be implemented as a DyNN since it learns and generate sets of adaptive network parameters (weights and biases) for varying noise conditions

Each condition has varying precision needs making it dynamic in nature.

Outline

- Neural Networks in Wireless Communication
- Dynamic Neural Networks (DyNN)
- Research Problem
- Proposed Work & Results
- Conclusion and Future Work
- Q&A

Research Problem

Traditional hardware implementation of NN's use **Uniform quantization** for simplicity.

In DyNNs it can lead to **dynamic power wastage** as different noise conditions require varying precisions.

State-of-the-art Approach

Existing approaches for dynamic power reduction:

- Selective data-gating only for sparse parameters to reduce switching activity.
- Reduced coarse-grained precision design assuming uniform precision allocation across all parameters.

Not suitable for DyNNs!

Research Problem

So how do we design the DyNN hardware for improved power efficiency without sacrificing performance?

A hardware that supports:

- ✓ Non-uniform quantization at run-time
- ✓ Fine grained control of precision

Catering to the varying channel conditions

Outline

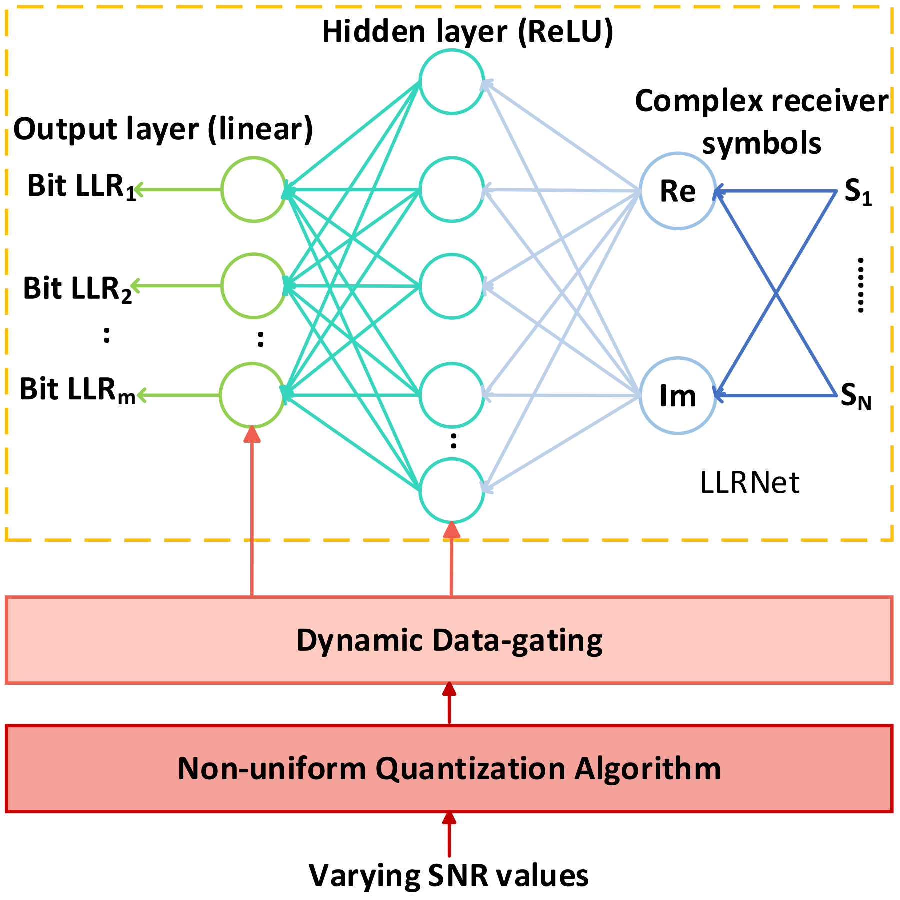
- Neural Networks in Wireless Communication
- Dynamic Neural Networks (DyNN)
- Research Problem
- Proposed Work & Results
- Conclusion and Future Work
- Q&A

Proposed Work

Main Contributions

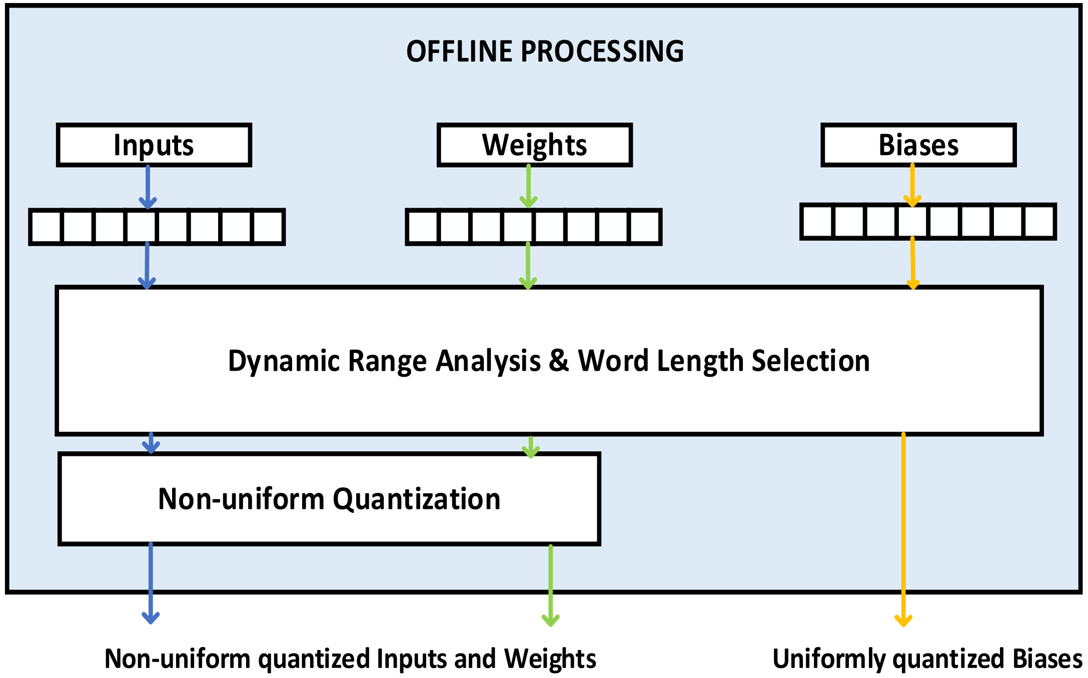
1. An offline non-uniform quantization algorithm enabling run-time quantization adaptation while preserving system performance.

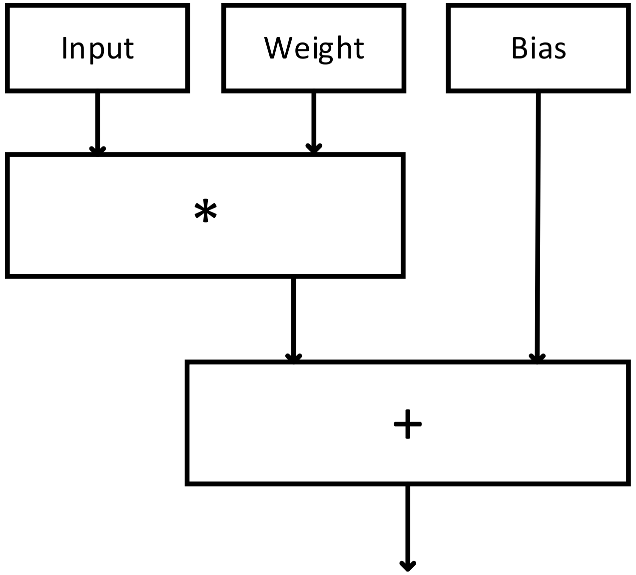
2. A low overhead dynamic data-gating architecture facilitating run-time non-uniform quantization.



1. Offline Non-uniform Quantization Algorithm

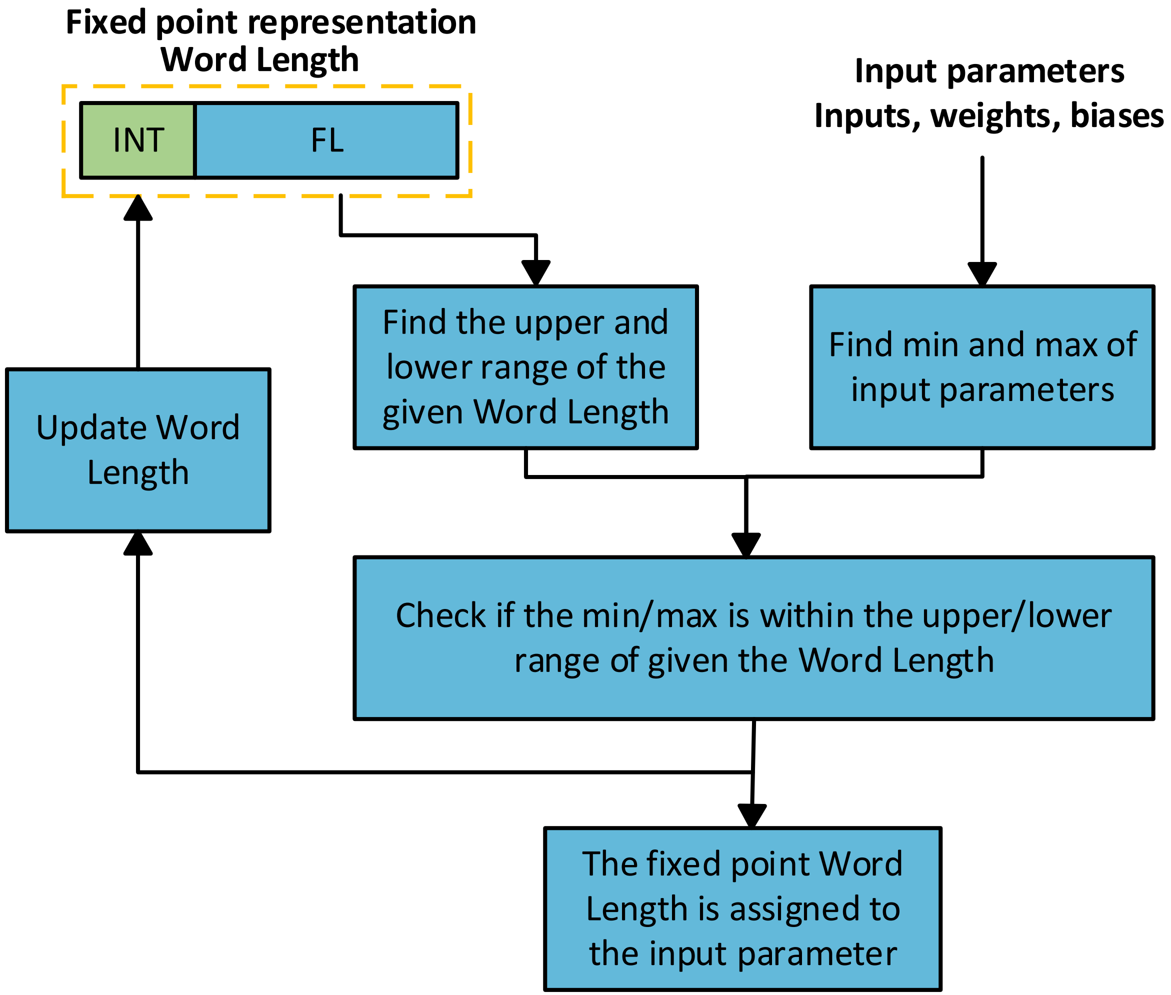
- Dynamic Range Analysis & Word Length selection is done for all network parameters.
- Non-uniform quantization is done for the Inputs and weights alone since multiplication units are the most computationally intensive.
- Applying non-uniform quantization to the Biases led to performance degradation.





1a. Dynamic Range Analysis and Word Length Selection

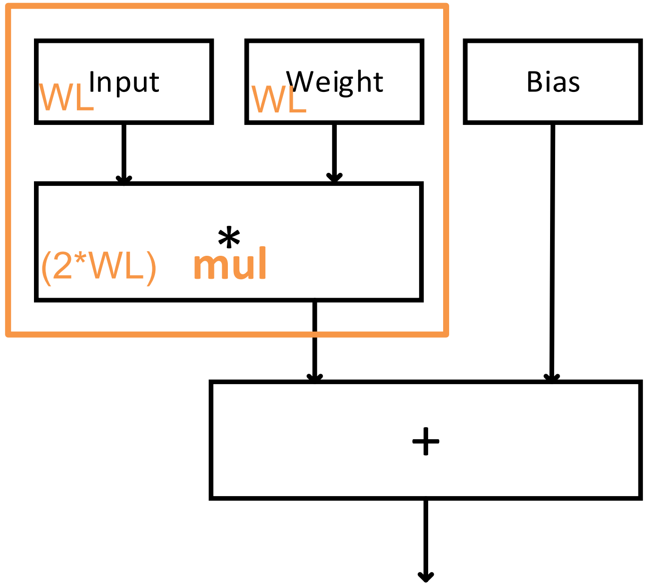
- The pre-trained LLRNet is reduced to a fixed-point precision model for a resource and power constrained hardware design.
- The dynamic range analysis helps to assign fixed point representations for the incoming input parameters.
- Decides the optimal fixed-point quantization for good end-to-end system performance.



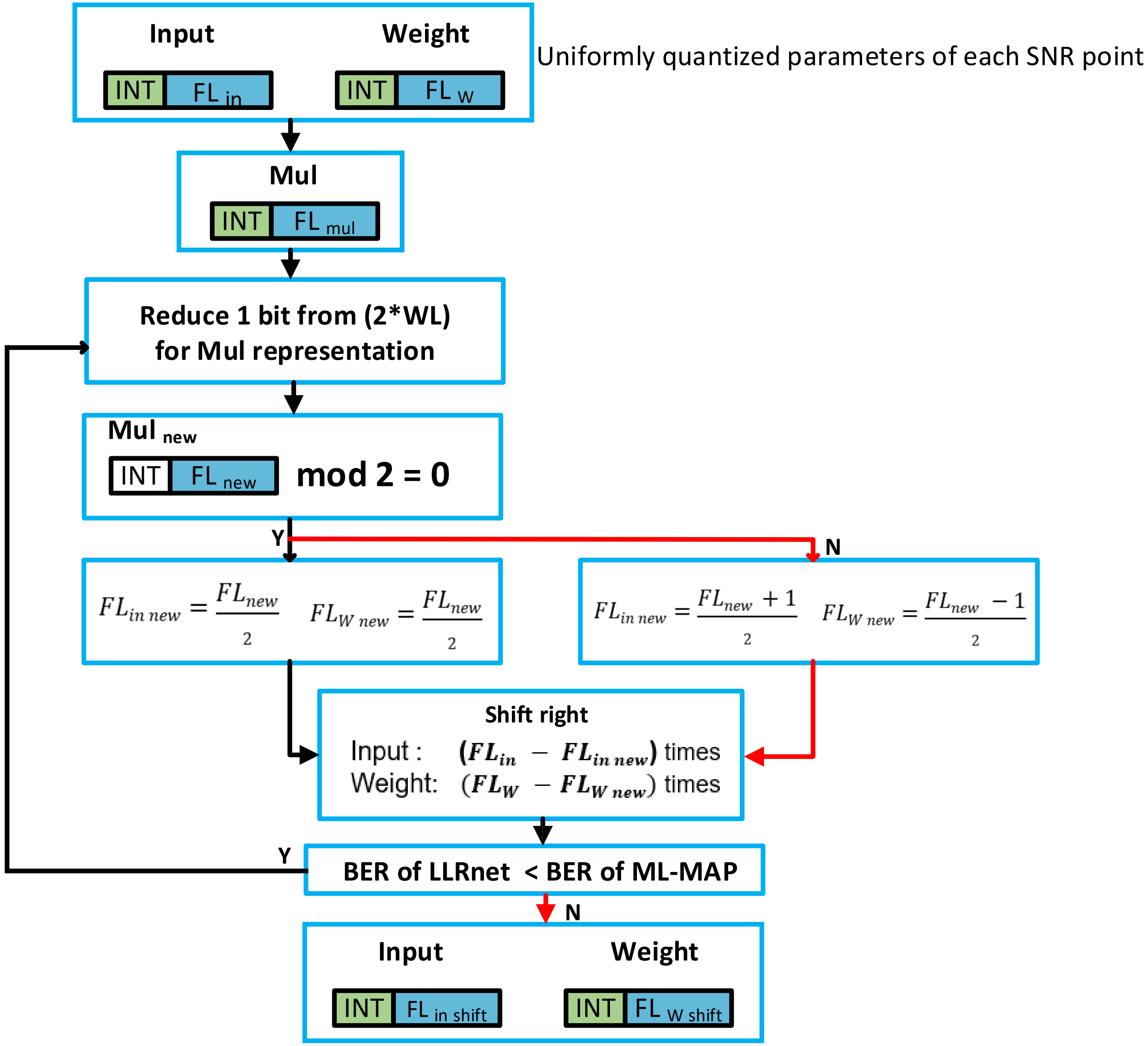
INT – Integer, FL – Fractional Length

1b. Non-uniform Quantization

Uniformly quantized parameters from previous step



WL = Word Length



2. Low Overhead Dynamic Data-gating Architecture

The 16-bit LLRNet Architecture has two layers:

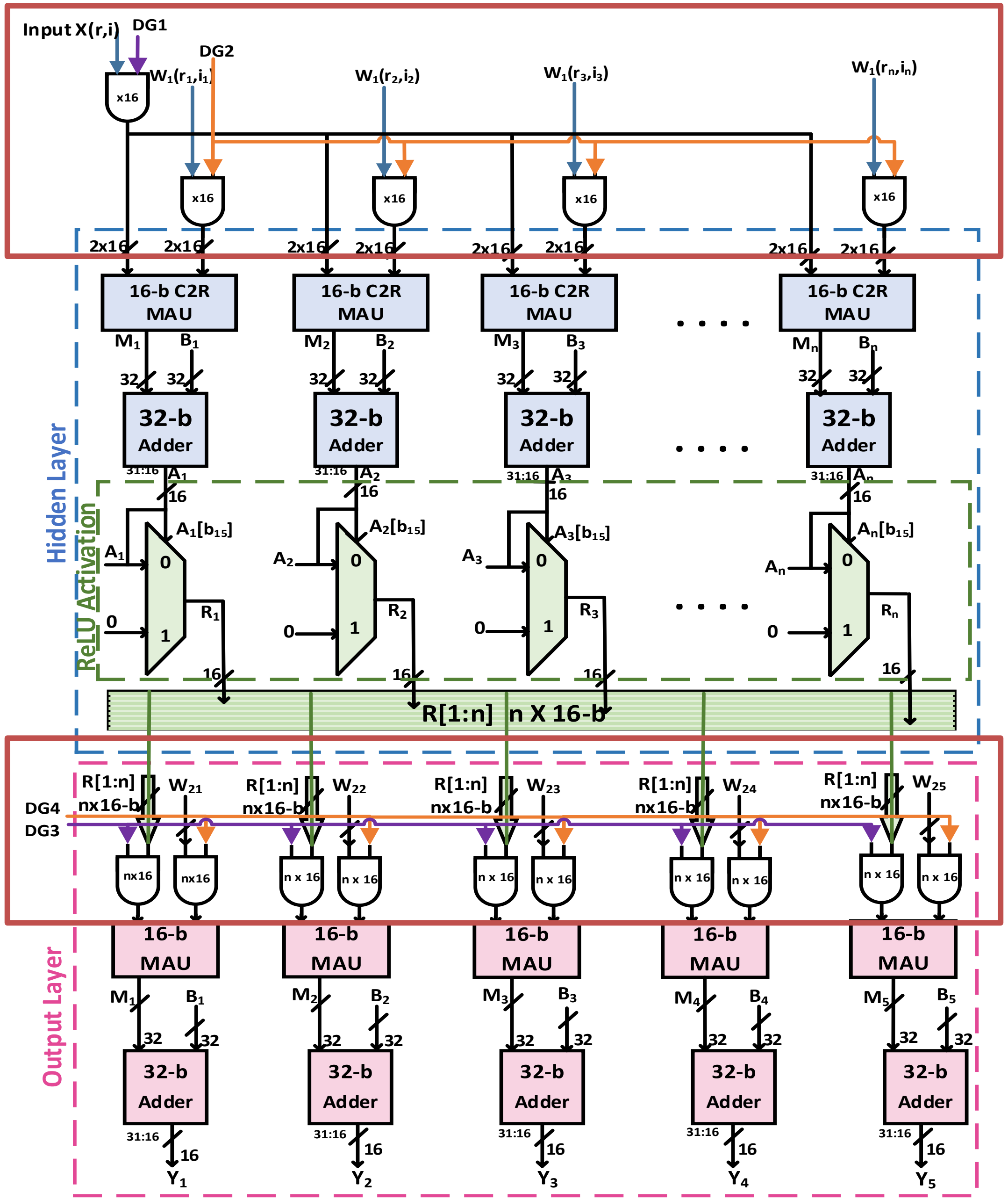
- Hidden layer
- Output Layer

The bit-wise dynamic data-gating unit is composed of **AND** gates.

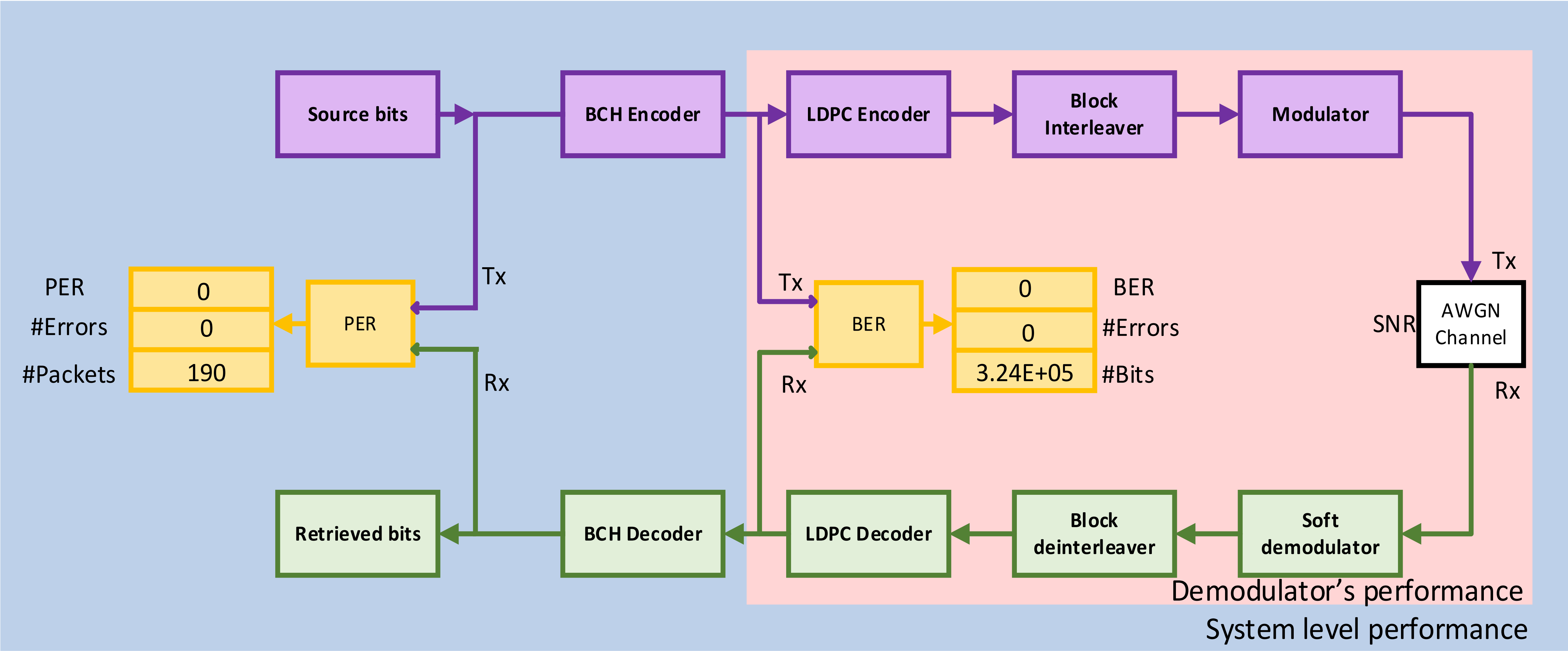
Control signals:

- DG1 for input X
- DG2 for hidden layer weights
- DG3 for the inputs of output layer
- DG4 for output layer weights

Power gating is applied to the unused blocks.



Experimental Setup of the Digital Video Broadcast-S.2 Simulation



https://nl.mathworks.com/help/comm/ug/training-and-testing-a-neural-network-for-llr-estimation.html

1a. Dynamic Range Analysis and Word Length Selection

Picking the best fixed-point Word Length representation for the Digital Video Broadcast (DVB-S.2) receiver :

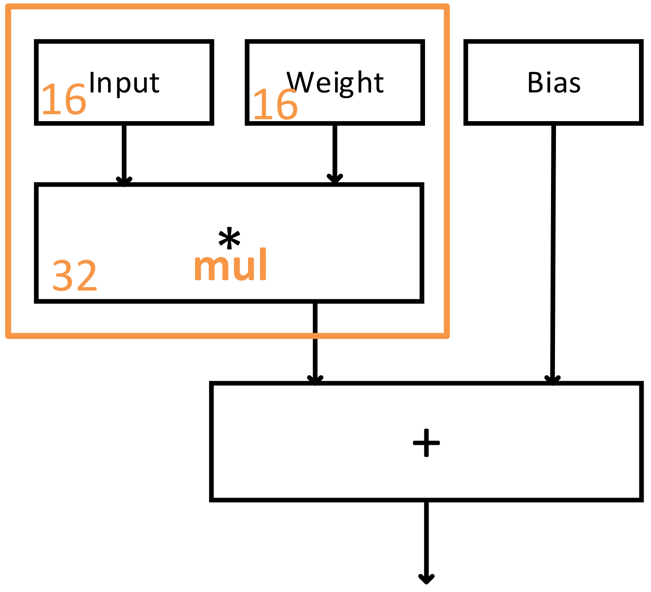
Input (WL-bit)	Output (2WL-bit)	Mean Square Error
4	8	9.18E-02
6	12	1.03E-04
8	16	5.26E-07
16	32	1.29E-13

$$MSE = \sum_{i=1}^{N} (LLR_{floating}(i) - LLR_{fixpt}(i))^2$$

- The Mean Square Error (MSE) analysis - impact of quantization in each layer.
- The 4-bit LLRNet has the highest quantization noise.
- The 6 and 8-bit versions didn't meet the required Quasi Error Free (QEF) performance.
- A 16-bit LLRNet with a wider dynamic range, was chosen to ensure good system performance.

1b. Non-uniform Quantization

Some reduced precision **mul** values obtained for SNR point of various demodulation schemes

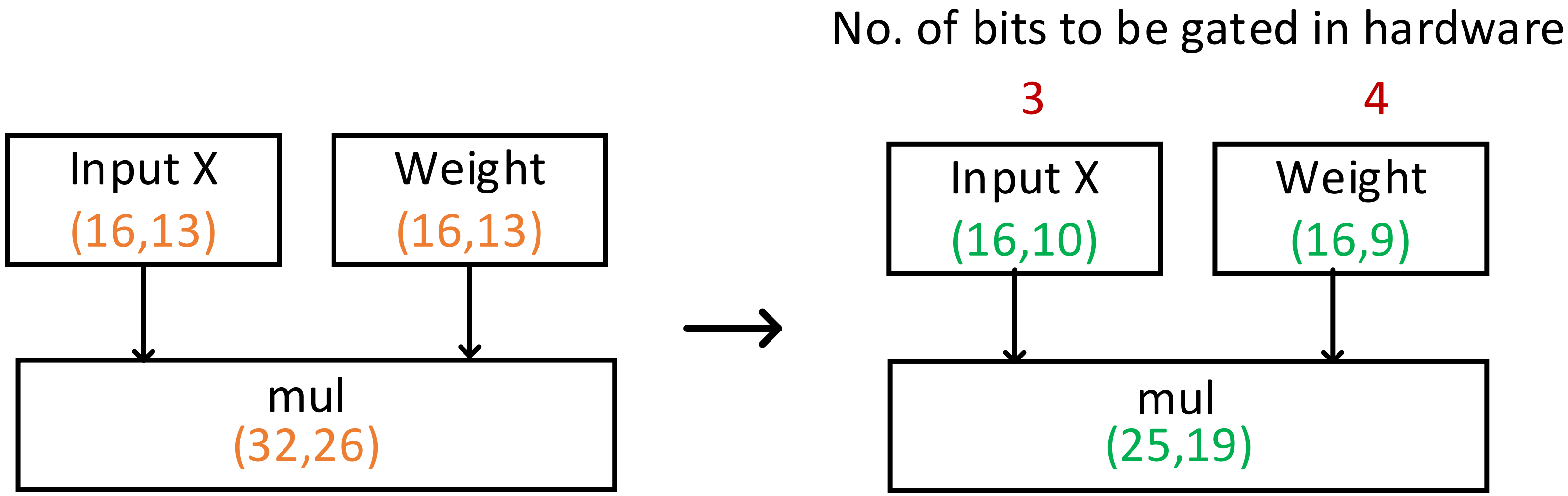


mul

Demodulation	SNR Value	Before non-uniform quantization	After non-uniform quantization
QPSK	4.6 dB	(32,28)	(20,16)
8PSK	5.4 dB	(32,27)	(21,16)
16APSK	8.9 dB	(32,26)	(25,19)
32APSK	15.70 dB	(32,25)	(29,22)

1b. Non-uniform Quantization

Example: SNR = 8.9dB for 16APSK demodulation

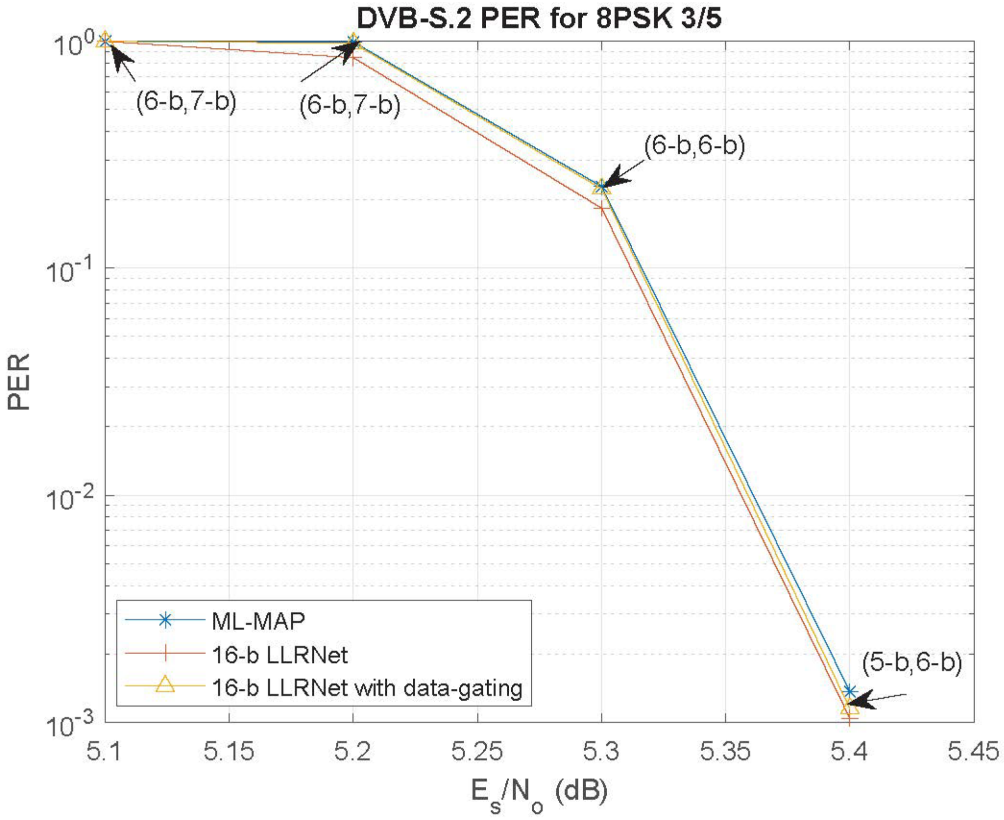


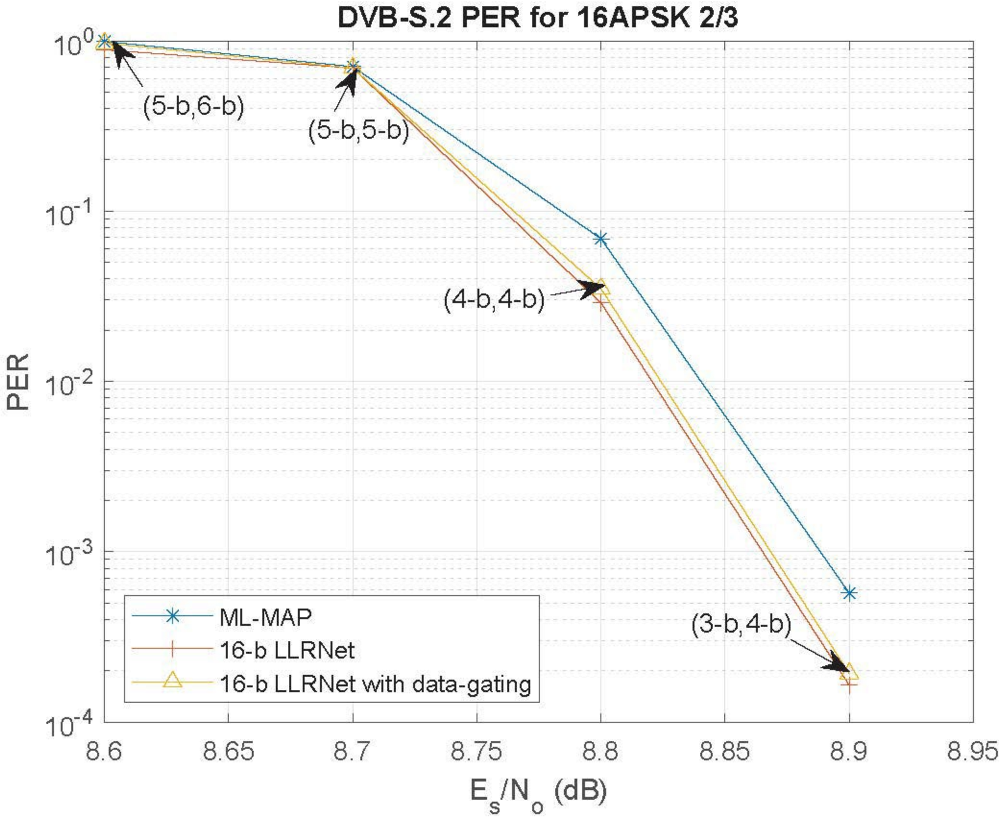
Before non-uniform quantization

After non-uniform quantization

System Level Performance

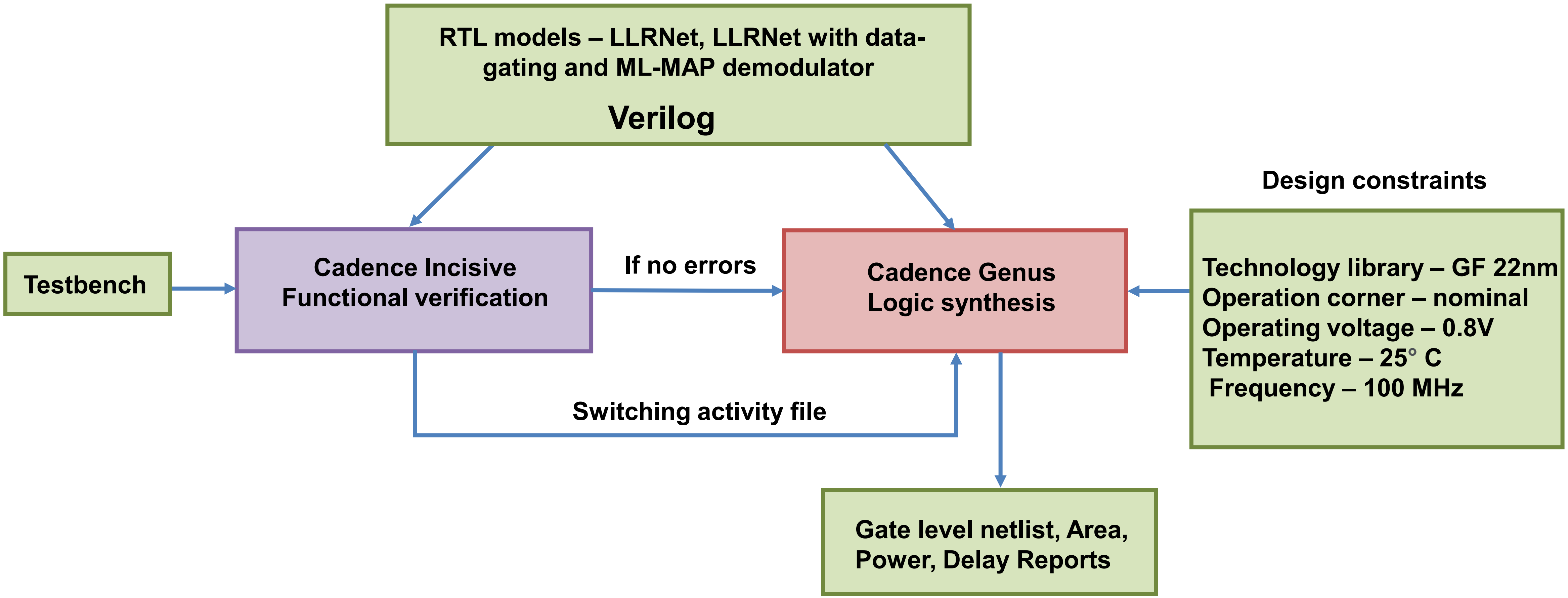
Packet Error Rate (PER) analysis for different demodulation schemes:



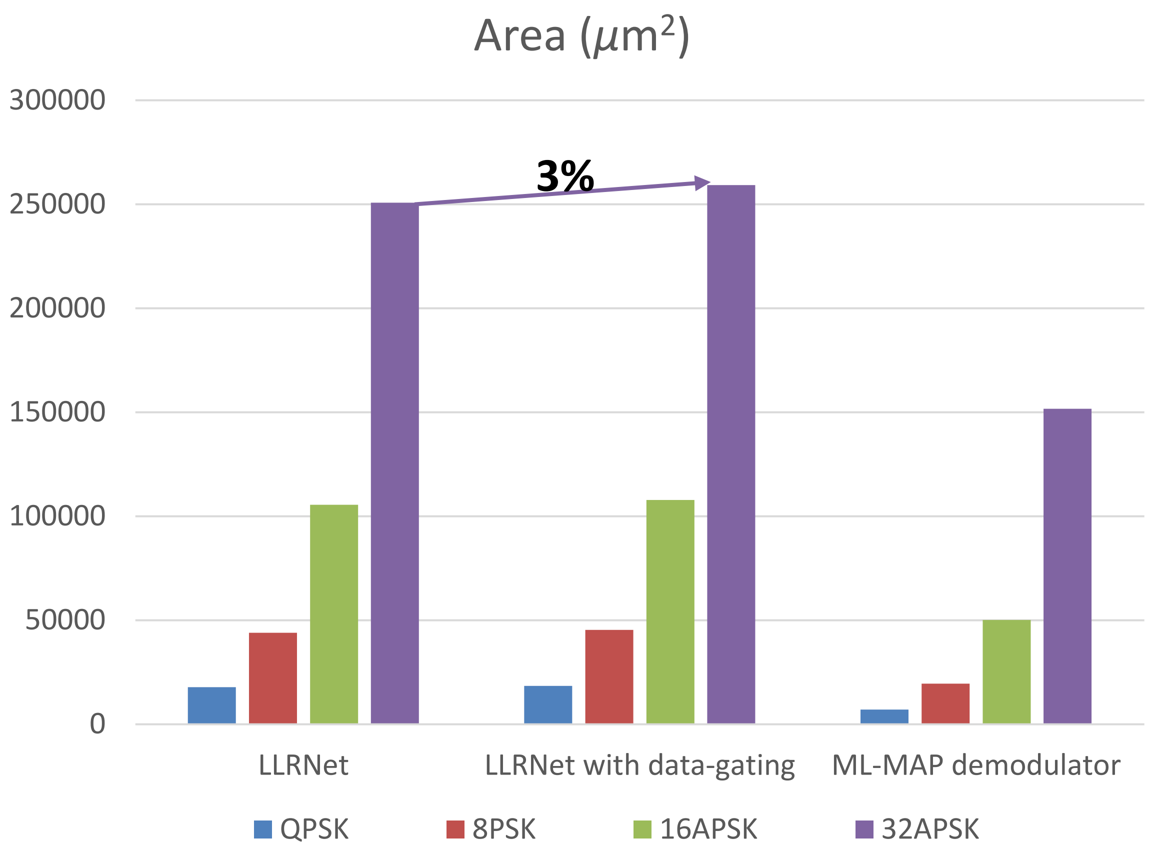


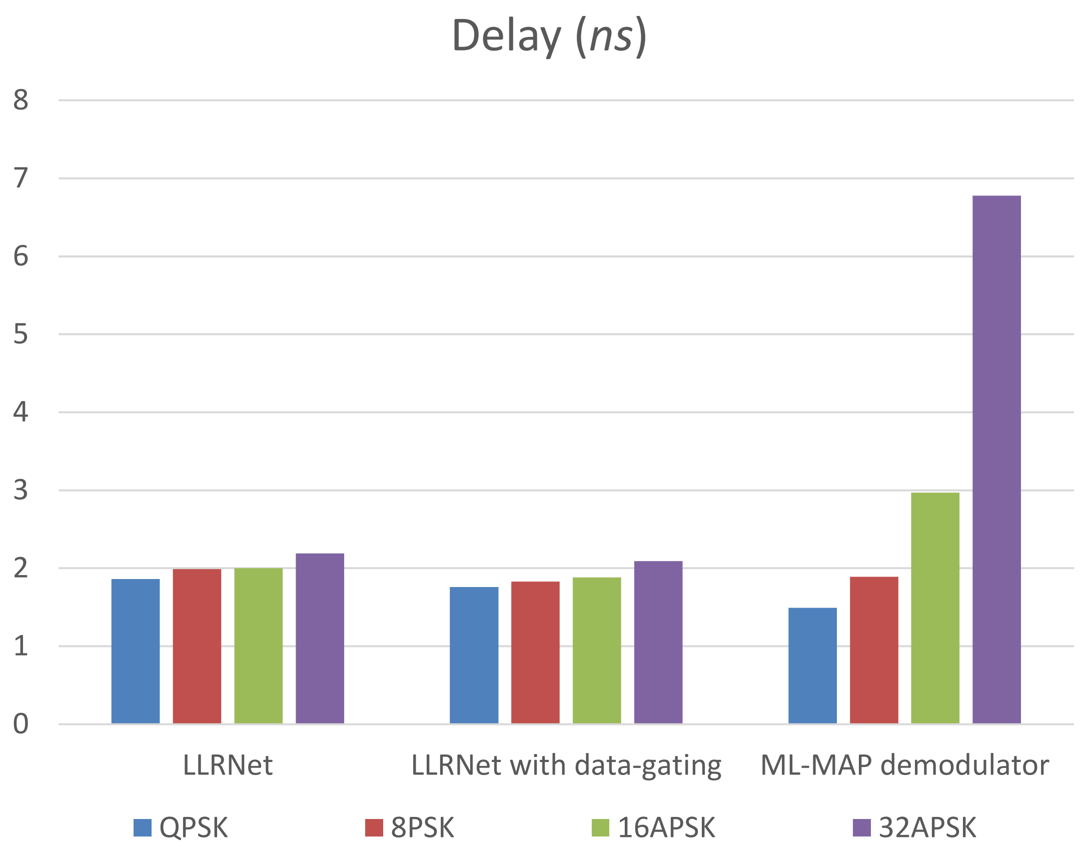
PER of the 16-bit LLRNet with data-gating although slightly worse than 16-bit LLRNet is still better than the PER of ML-MAP (Approximate LLR) demodulator.

Experimental Setup of the Hardware Synthesis



Area and Delay Comparison

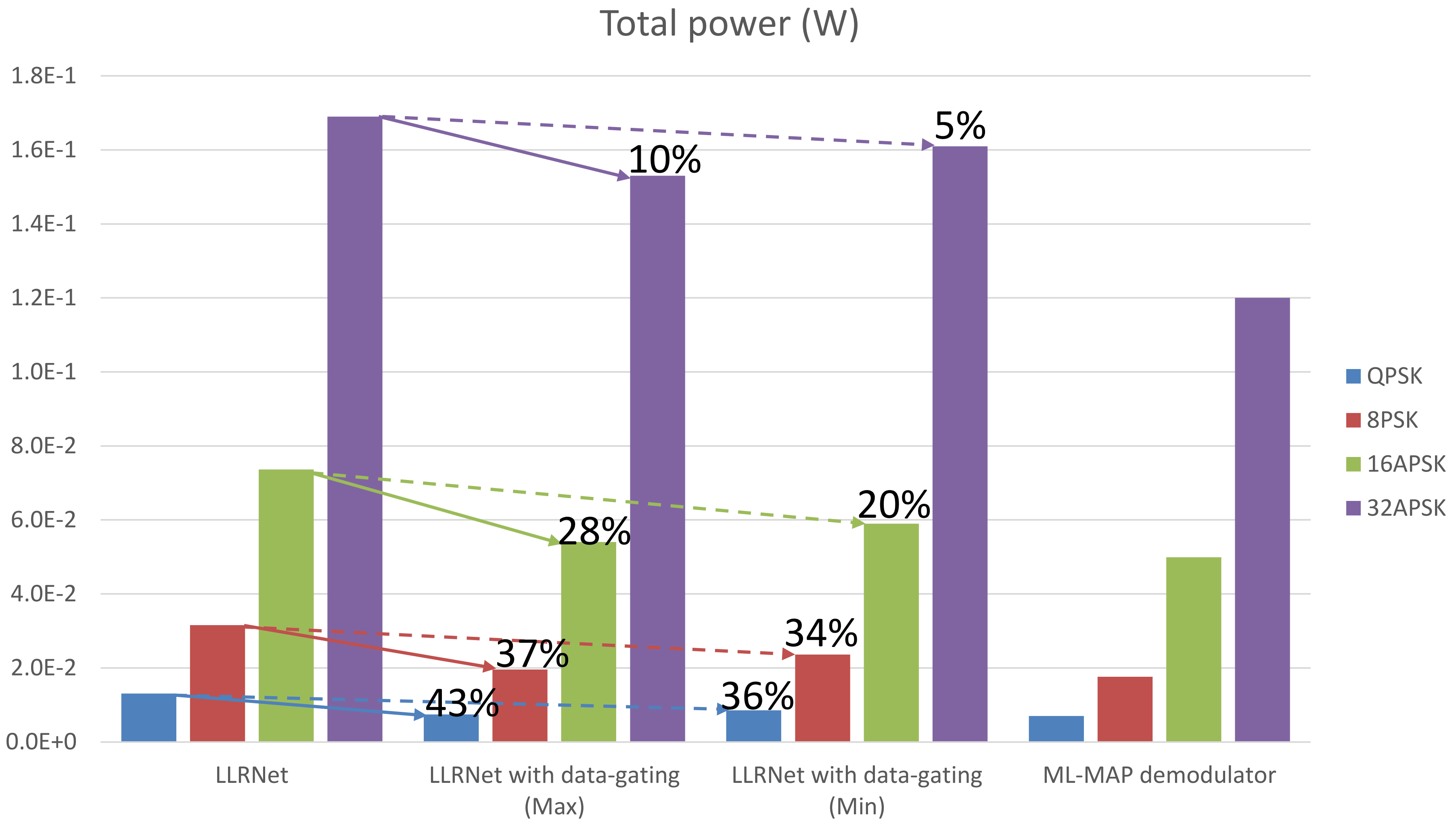




LLRNet with data-dating has a 3% minimal area overhead when compared to the architecture without data-gating.

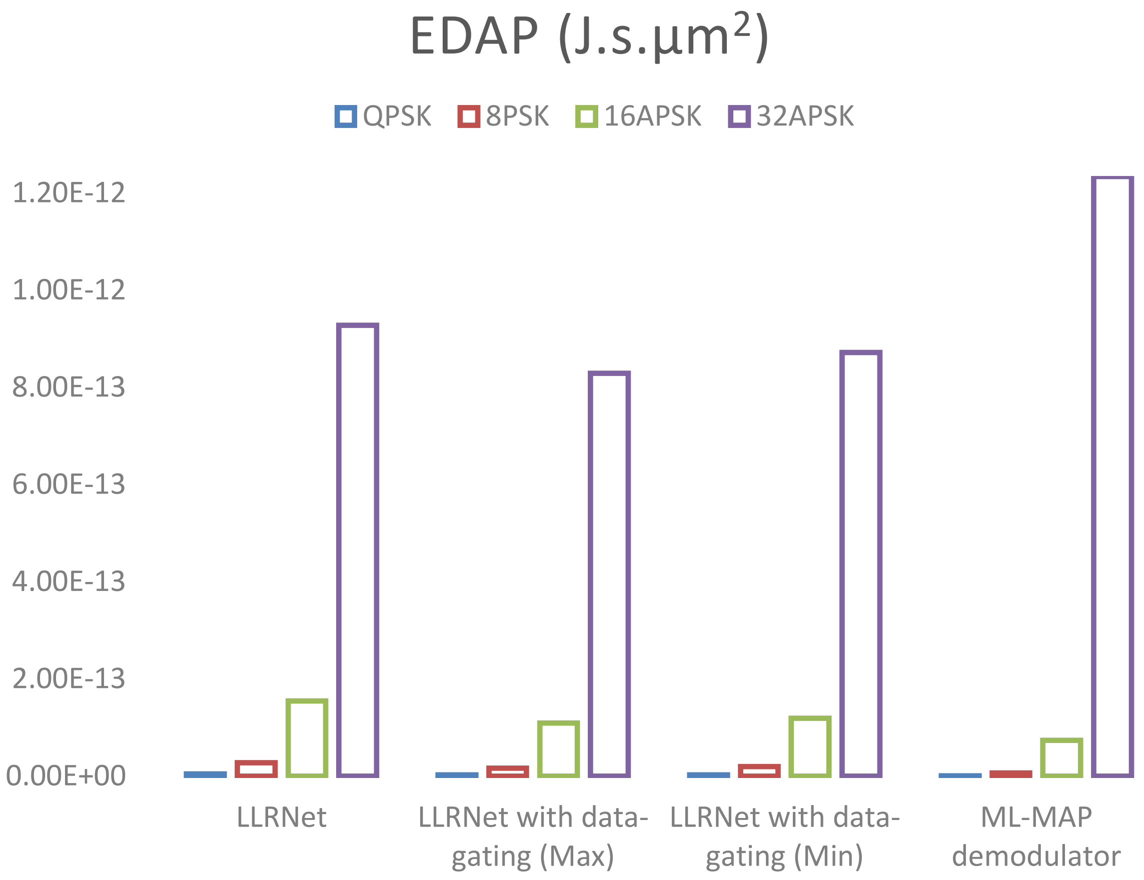
LLRNet with data-gating has a better delay performance when compared to conventional ML-MAP demodulator when constellation patterns start getting denser.

Power Comparison



QPSK demodulation with a maximum data-gating of 8-bits gives us total power savings of up to 43%

Overall Performance Comparison



The ML-MAP demodulator shows better area, power and Energy Delay Area Product (EDAP) results for sparse constellations.

LLRNet with data-gating outperforms it when constellations get denser (32APSK) in terms of EDAP.

In the future, as wireless communication systems rely more on NNs, dynamic data-gating will become an essential feature to reduce power consumption.

Outline

- Neural Networks in Wireless Communication
- Dynamic Neural Networks (DyNN)
- Research Problem
- Proposed Work & Results
- Conclusion and Future Work
- Q&A

Conclusion & Future Work

- A novel non-uniform quantization algorithm and a low overhead dynamic data-gating hardware that implements the proposed run-time quantization adaptation for DyNN is presented.

- Applied to the NN demodulator LLRNet, our methodology achieves substantial total power savings of up to 43%, with only a minor 3% area overhead.

- The generic dynamic data-gating hardware can be applied to other DyNN blocks as well, leading to good system-level power savings.

- As neural networks fully replace receiver systems in the future, dynamic data-gating will become crucial for power efficiency.

Future work includes a fully dynamic system by combining the proposed non-uniform quantization algorithm with online learning techniques.

Outline

- Neural Networks in Wireless Communication
- Dynamic Neural Networks (DyNN)
- Research Problem
- Proposed Work & Results
- Conclusion and Future Work
- Q&A

Thank you

Priscilla Sharon Allwin
p.s.allwin@tue.nl