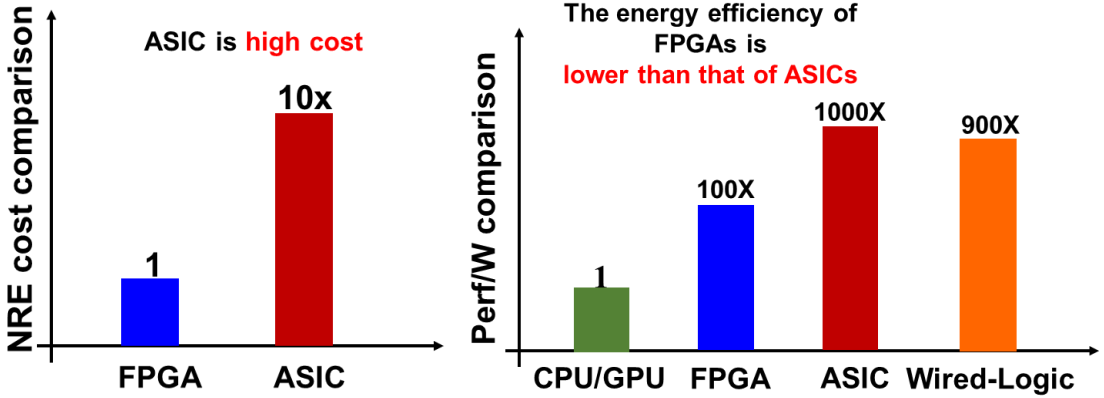
A Coarse- and Fine-Grained LUT Segmentation Method Enabling Single FPGA Implementation of Wired-Logic DNN Processor

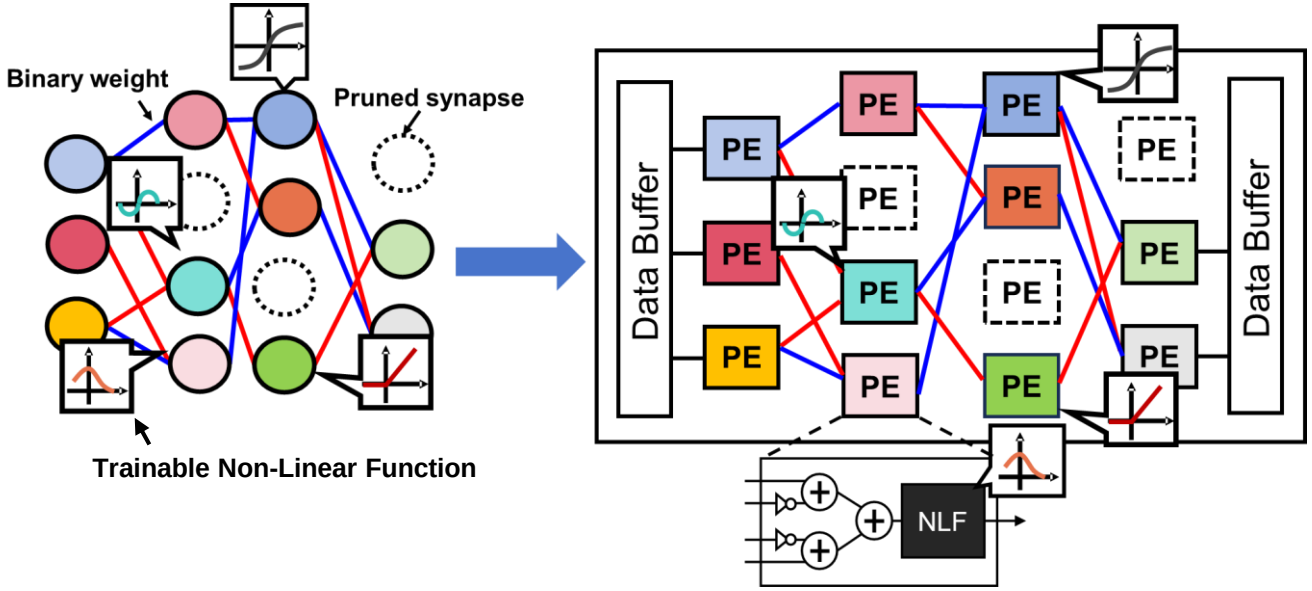
Yuxuan Pan, Dongzhu Li, Mototsugu Hamada, Atsutake Kosuge

The University of Tokyo

Wired-Logic Architecture Using NNN

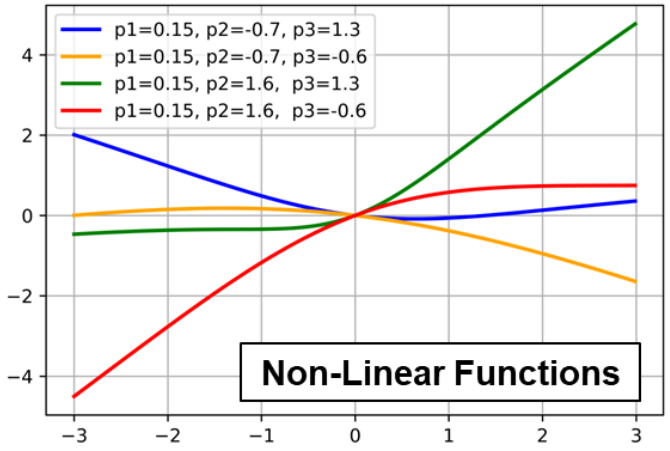
- FPGAs are reconfigurable for AI tasks but less energy-efficient
- Wired-logic architecture using Non-linear Neural Network (NNN)
  1. Binarize weights to +1/−1
  2. Prune unnecessary synapses
  3. Learn the non-linear activation function at each neuron individually
- Low latency & high energy efficiency via eliminating memory access

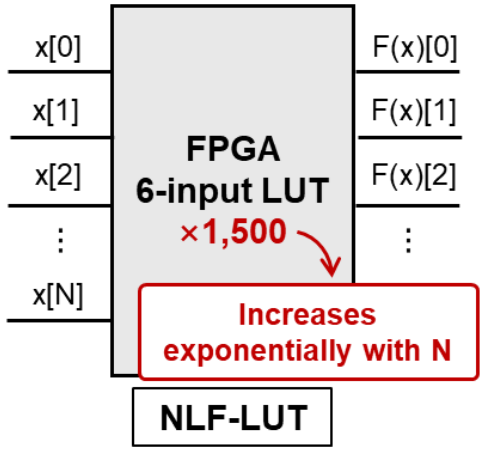


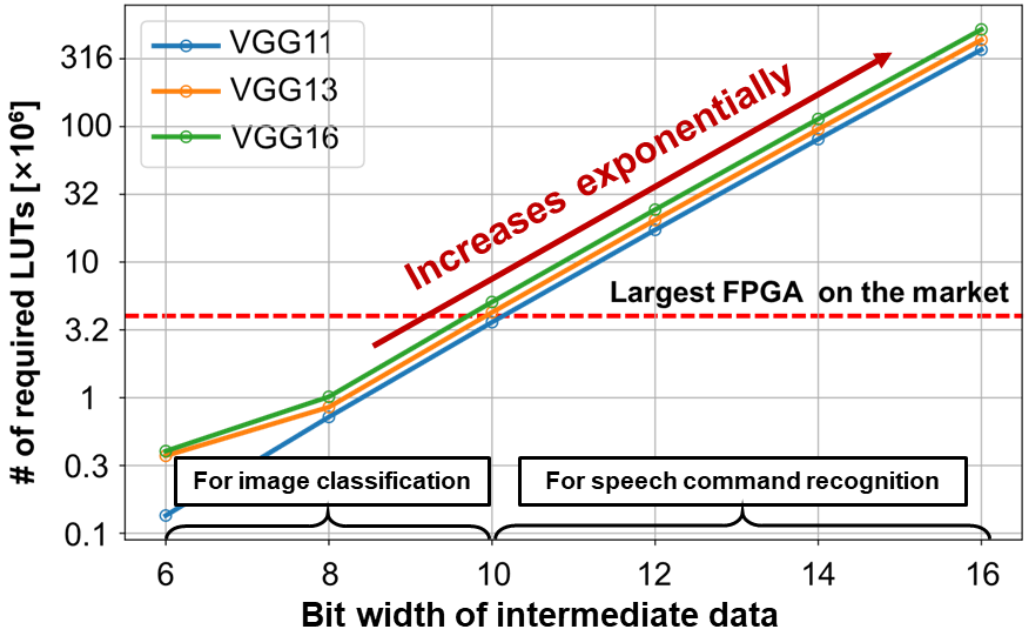


Challenge of Processing Long-Bit-Width Data

- The implementation of Non-Linear Functions (NLFs) consumes significant LUT resources
- The sharp increase in LUT requirements makes it impossible to implement scaled NNN models with long-bit-width data on a single FPGA

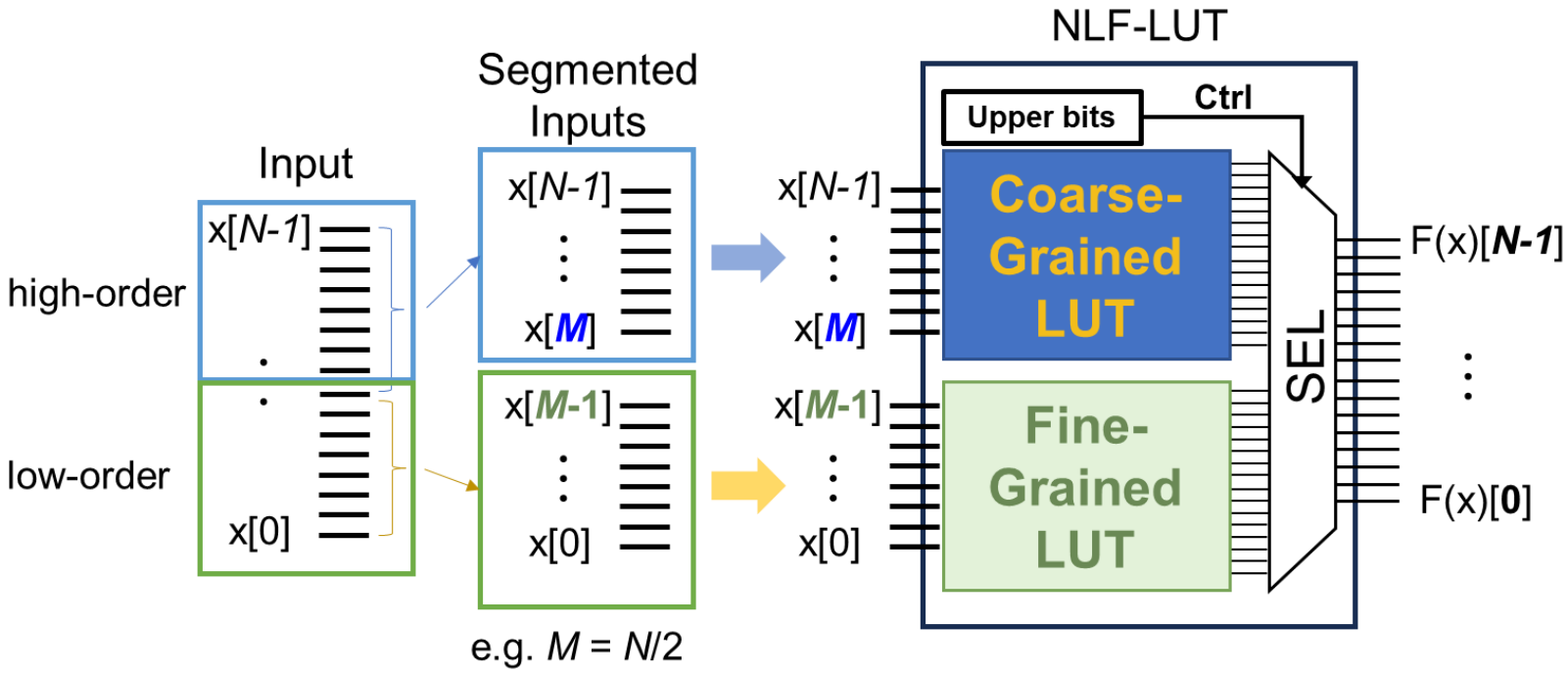






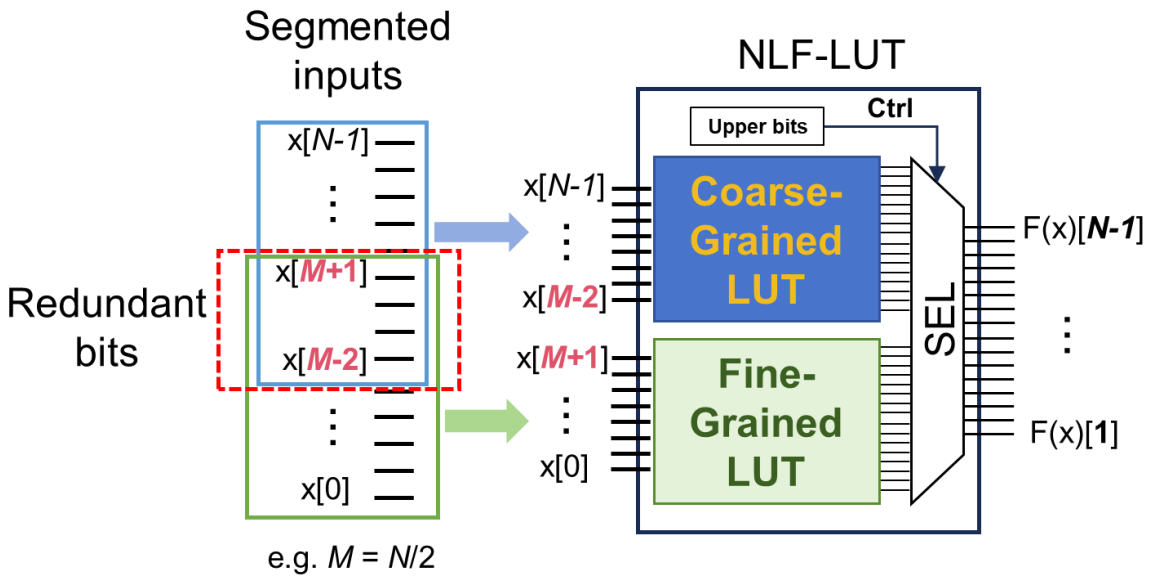
Coarse- and Fine-Grained LUT Segmentation

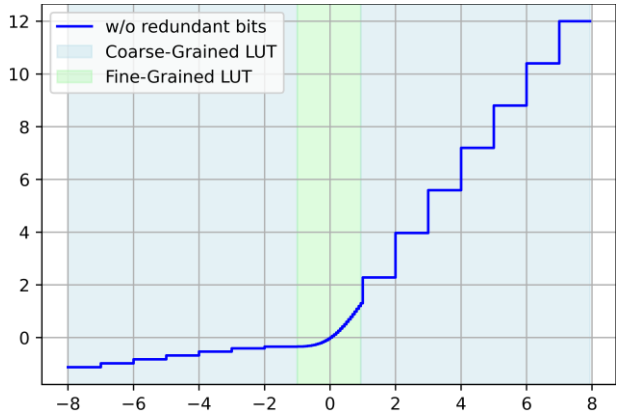
- **Segment the input bits of the NLF into High/Low order based on Coarse/Fine granularity**
- **The output switching is realized by monitoring the upper bits of the input signal**



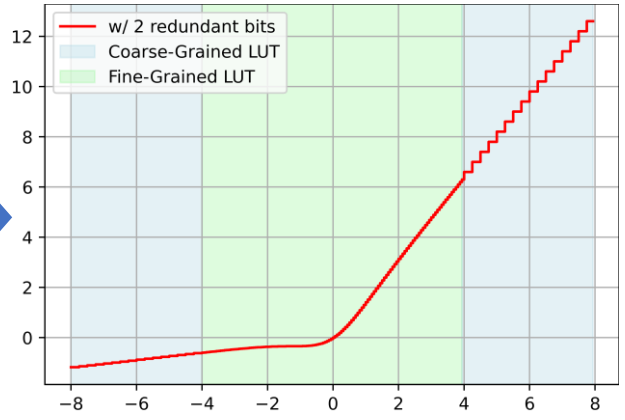
Redundant Bits to Restore Accuracy

- **When high-order bits have meaningful values, information in the low-order bits is lost, leads to accuracy reduction for complex tasks**
- **Redundant bits are merged into the segmented input to improve accuracy**





w/o redundant bits



w/ 2 redundant bits

Comparison with FPGA Implementations

- **Keyword Spotting (GSCD-10/20): ~93% NLF-LUTs saving, ~1.5% accuracy drop**
- **15-20x more energy efficiency improvement than Ref. [1]**

Index	Ref. [1]	This work			
		Baseline (#1)	Proposed method	Baseline (#2)	Proposed method
# of commmands	10 (1)	10	10	20	20 (2×)
Bit width	4/4	1/16	1/16	1/16	1/16
LUT segmentation	–	No	Yes	No	Yes
Redundant bits	–	–	2 bits	–	2 bits
LUTs	N/A	2,614,511	827,920	2,975,693	1,024,954
LUTs for NLF	–	1,925,048 (100%)	138,457 (-92.8%)	2,097,433 (100%)	146,694 (-93%)
Accuracy	90.1%	90% (0)	88.8% (-1.2%)	82% (0)	80.4% (-1.6%)
Power [W]	0.83	42.39	18.03	48.61	24.38
Throughput [Mfps]	1.43×10^{-3}	0.625	0.625	0.625	0.625
Energy efficiency [µJ/inf.]	580 (1)	67.32	28.85 (1/20)	77.78	39 (1/15)

[1] Mazumder, A. N., & Mohsenin, T. M., arXiv preprint arXiv:2202.02361, 2022.