PHD FORUM

Robust and Energy-Efficient Deep Learning Systems

Muhammad Abdullah Hanif[1] (Ph.D. Candidate), Muhammad Shafique[2] (Advisor)

[1]Technische Universität Wien (TU Wien), Vienna, Austria

[2]Division of Engineering, New York University Abu Dhabi (NYUAD), Abu Dhabi, UAE

Problems and Motivation

❑ **Applications**

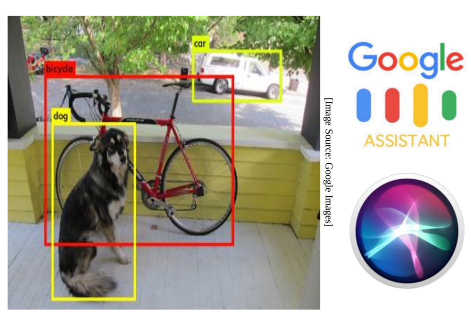
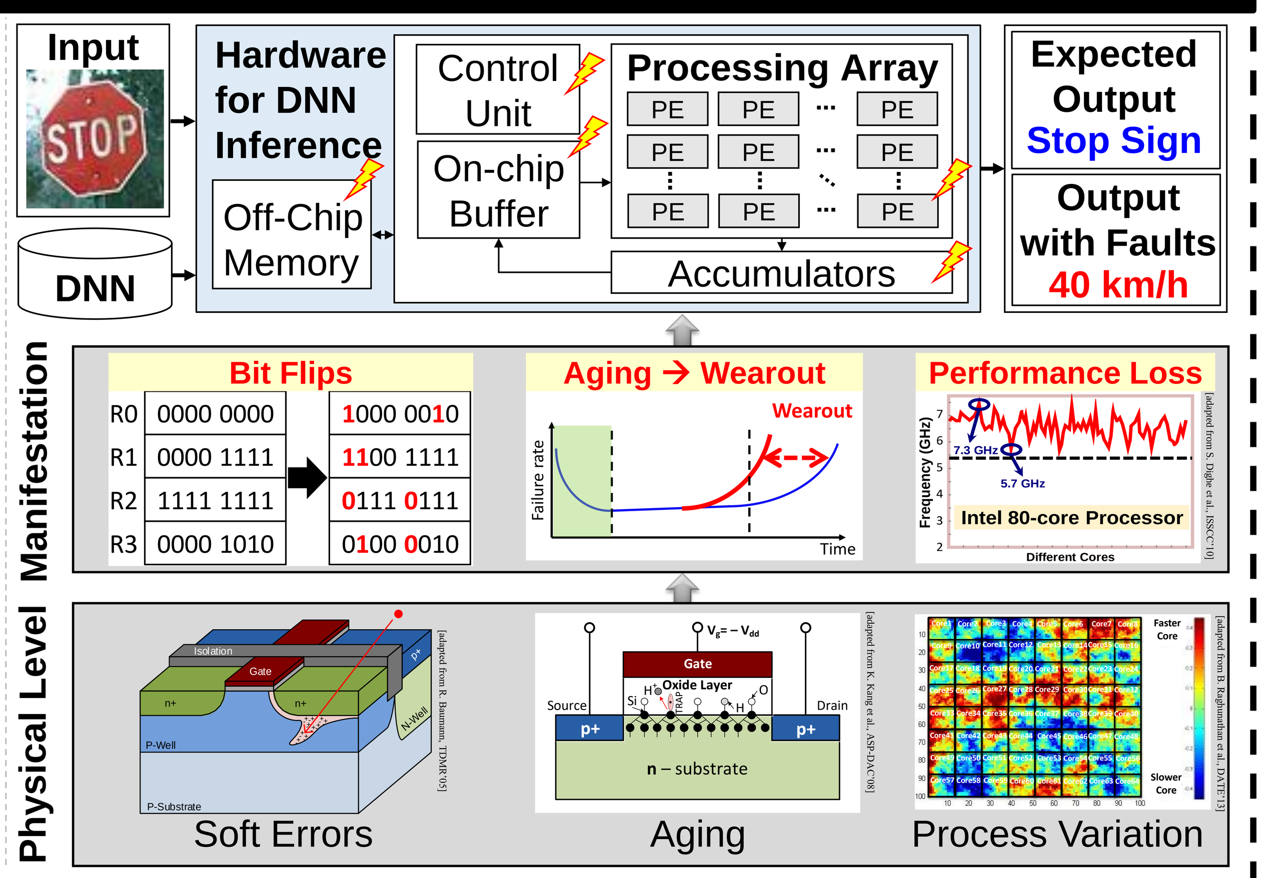


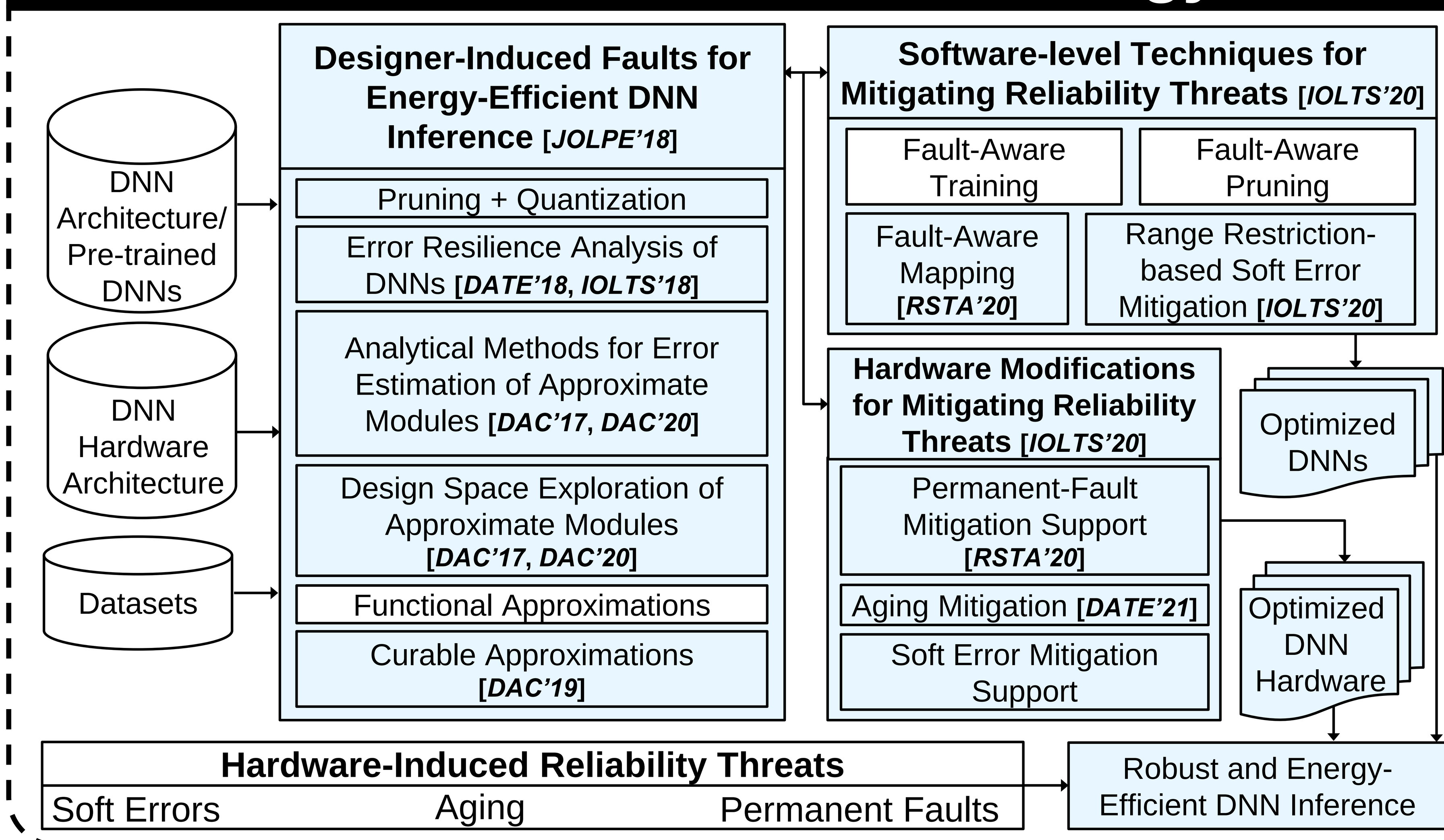
Image Classification, Object Detection & Localization, etc. — Natural Language Processing



❑ **Compute Cost [OPs]**
- ❑ AlexNet → 1.5B
- ❑ VGGNet → 19.6B
- ❑ Inception → 2B
- ❑ ResNet-152 → 11B

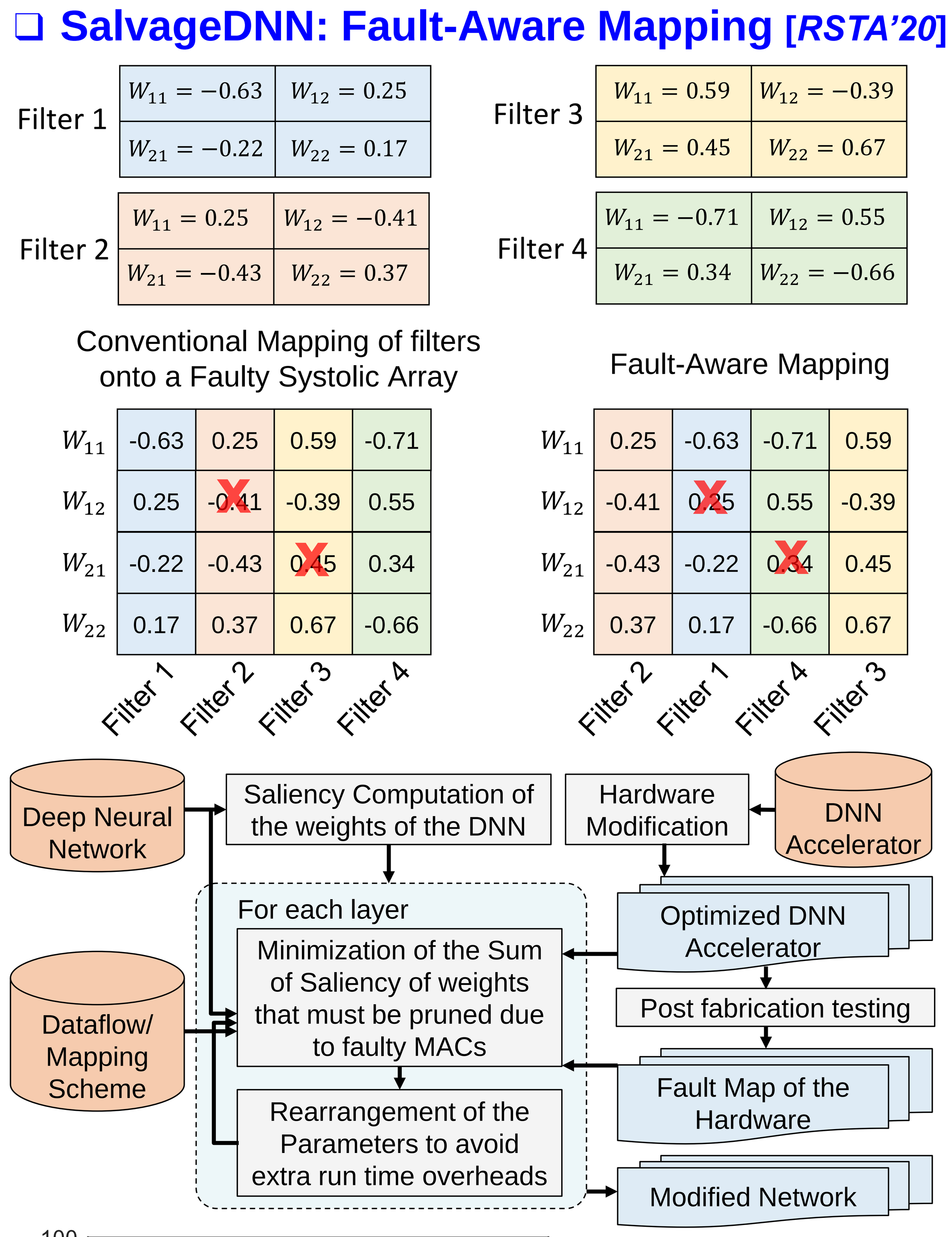
❑ **Design Methods to Improve Robustness and Energy-Efficiency of DNNs**
- ❑ Cost-effective techniques for mitigating reliability threats
- ❑ HW/SW approximations for achieving high energy-efficiency

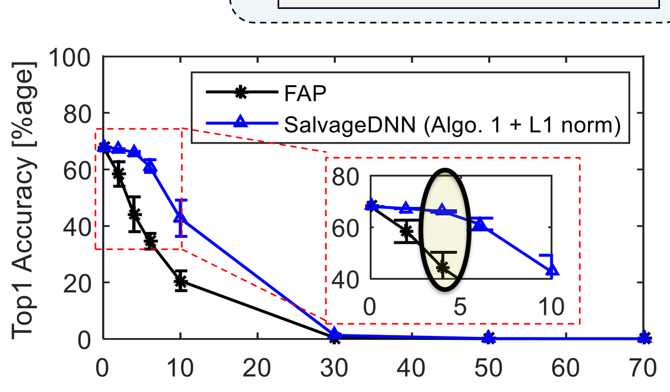
Overview of Our Methodology



Techniques for Robust and Energy-Efficient Deep Neural Network Inference

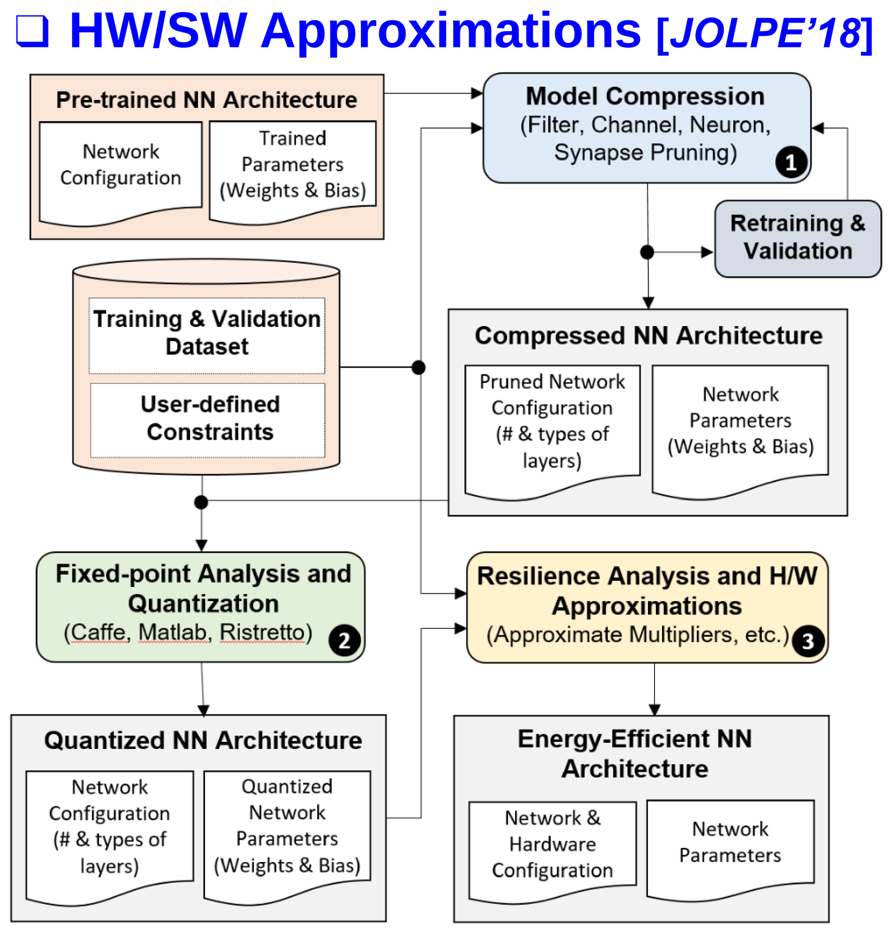
❑ SalvageDNN: Fault-Aware Mapping [RSTA'20]

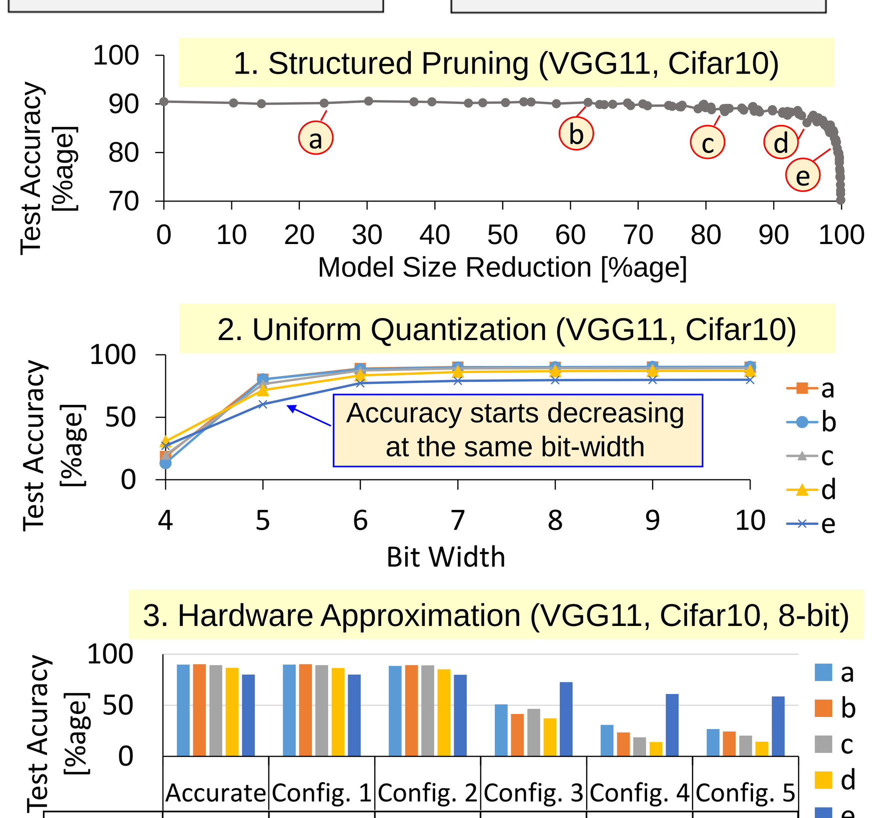




For VGG11 with ImageNet dataset, our SalvageDNN method takes ≈2 mins on a CPU while one epoch of re-training takes ≈100 mins with a GPU

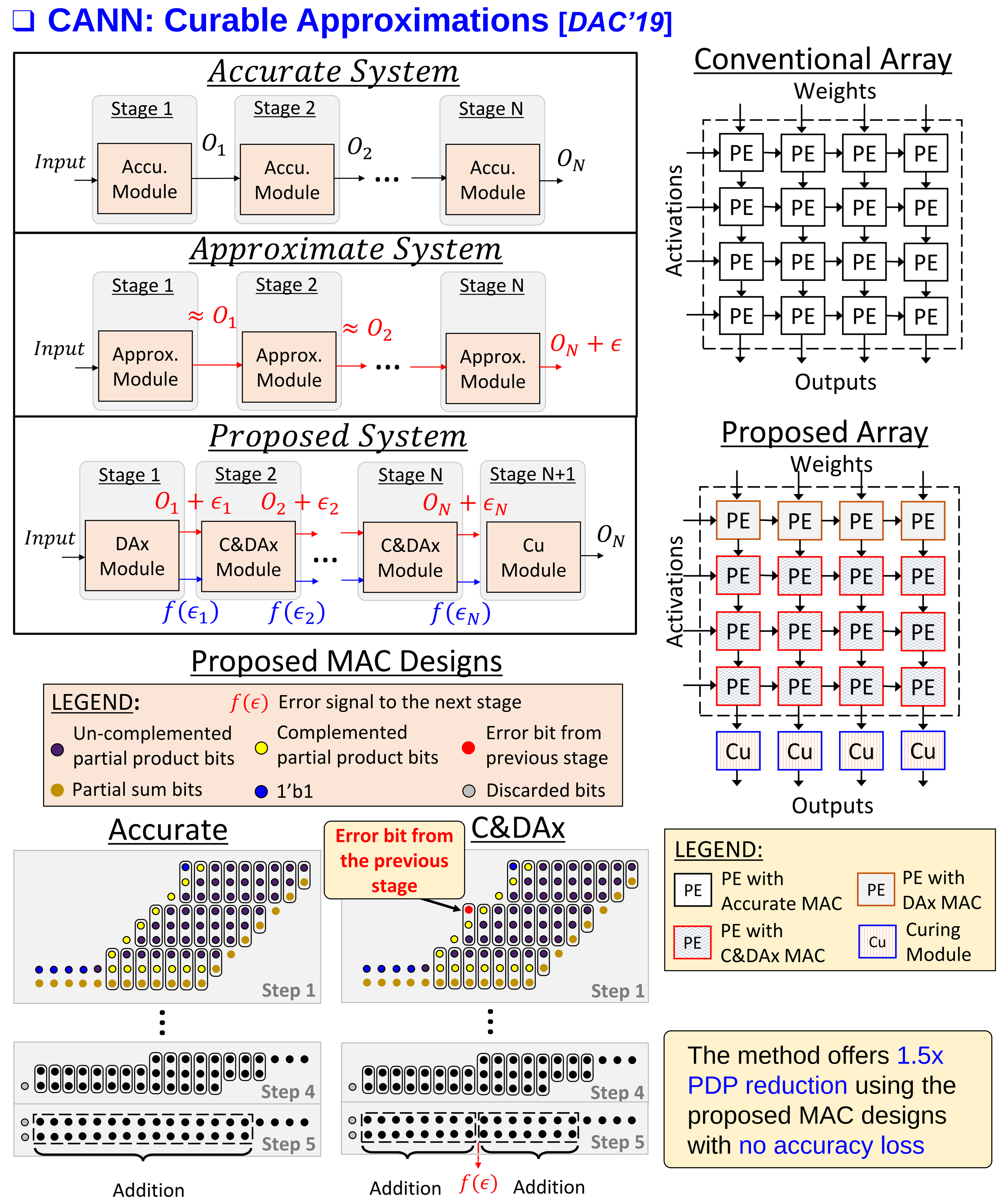
❑ HW/SW Approximations [JOLPE'18]





	Accurate	Config. 1	Config. 2	Config. 3	Config. 4	Config. 5
PDP [fJ]	88.40	83.66	77.71	70.60	70.90	67.40
MSE	0	0.3	9.8	266.3	3102.3	24806

❑ CANN: Curable Approximations [DAC'19]



The method offers 1.5x PDP reduction using the proposed MAC designs with no accuracy loss

Selected Publications

[DAC'17] M. A. Hanif, R. Hafiz, O. Hasan, M. Shafique, "QuAd: Design and analysis of quality-area optimal low-latency approximate adders", Design Automation Conference (**DAC**), pp. 1–6, 2017. **Received a HiPEAC Paper Award.**

[DATE'18] M. A. Hanif, R. Hafiz, M. Shafique, "Error resilience analysis for systematically employing approximate computing in convolutional neural networks", **DATE**, pp. 913–916, 2018..

[DAC'19] M. A. Hanif, F. Khalid, M. Shafique, "CANN: Curable approximations for high-performance deep neural network accelerators", **DAC**, pp. 1–6, 2019. **Received a HiPEAC Paper Award.**

[DAC'20] M. A. Hanif, R. Hafiz, O. Hasan, M. Shafique, "PEMACx: A probabilistic error analysis methodology for adders with cascaded approximate units", **DAC**, pp. 1–6, 2020. **Received a HiPEAC Paper Award.**

[DATE'21] M. A. Hanif, M. Shafique, "DNN-Life: An energy-efficient aging mitigation framework for improving the lifetime of on-chip weight memories in deep neural network hardware architectures", **DATE**, pp. 1-6, 2021.

[JOLPE'18] M. A. Hanif, A. Marchisio, T. Arif, R. Hafiz, S. Rehman, M. Shafique, "X-DNNs: Systematic cross-layer approximations for energy-efficient deep neural networks", Journal of Low Power Electronics (**JOLPE**), vol. 14, no. 4, pp. 520–534, 2018.

[IOLTS'18] M. A. Hanif, F. Khalid, R. V. W. Putra, S. Rehman, M. Shafique, "Robust machine learning systems: Reliability and security for deep neural networks", IEEE International Symposium on On-Line Testing and Robust System Design (**IOLTS**), pp. 257–260, 2018.

[IOLTS'20] M. A. Hanif, M. Shafique, "Dependable deep learning: Towards cost-efficient resilience of deep neural network accelerators against soft errors and perma-nent faults", **IOLTS**, pp. 1–4, 2020.

[RSTA'20] M. A. Hanif, M. Shafique, "SalvageDNN: Salvaging deep neural network accelerators with permanent faults through saliency-driven fault-aware map-ping", Philosophical Transactions of the Royal Society A (**RSTA**), vol. 378, no. 2164, pp. 20190164, 2020.

dblp link: https://dblp.org/pid/191/2451.html

TECHNISCHE UNIVERSITÄT WIEN

جامعة نيويورك أبوظبي NYU ABU DHABI

CARE Tech — Computer Architecture and Robust Energy-Efficient Technologies Group