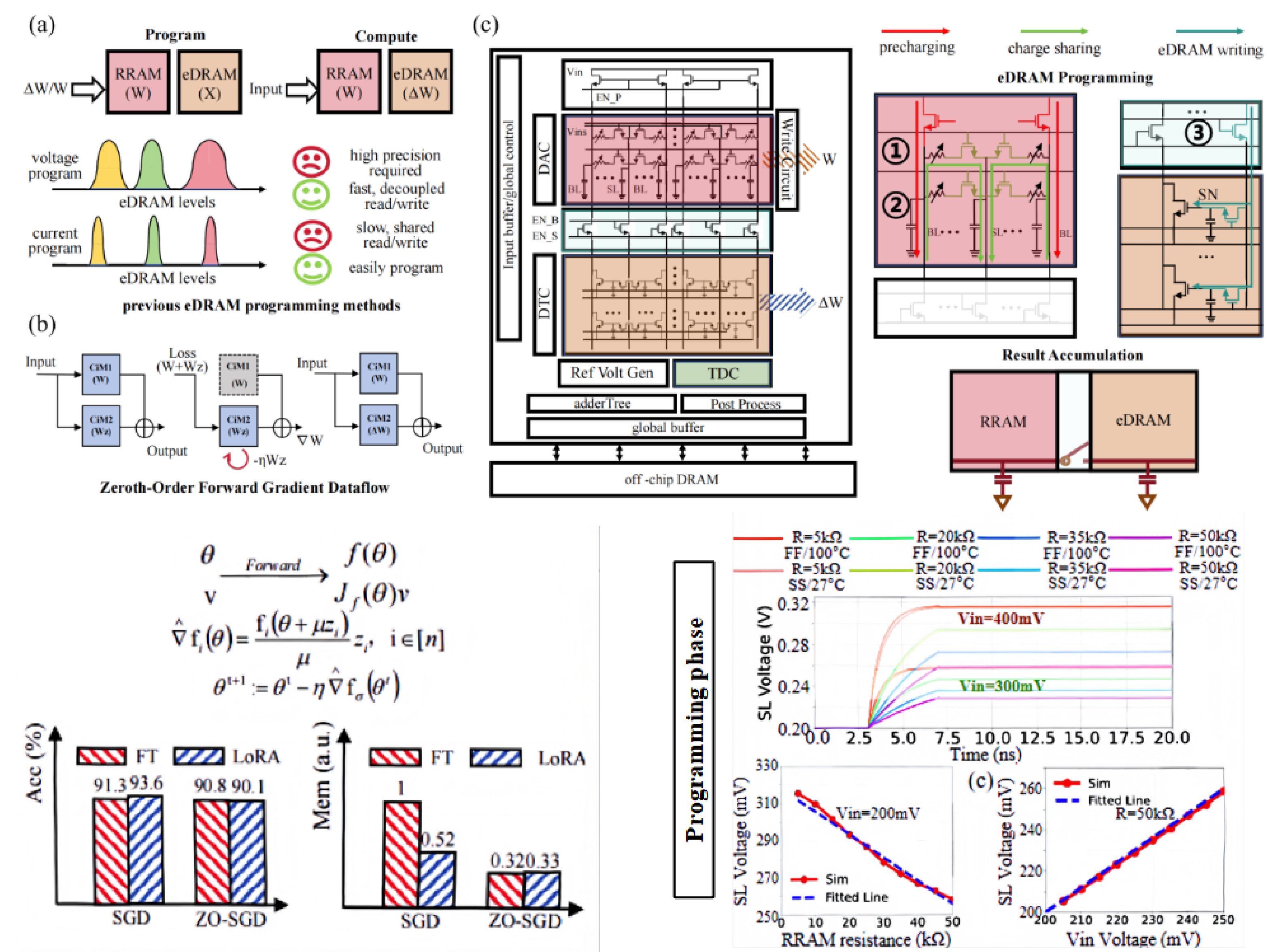
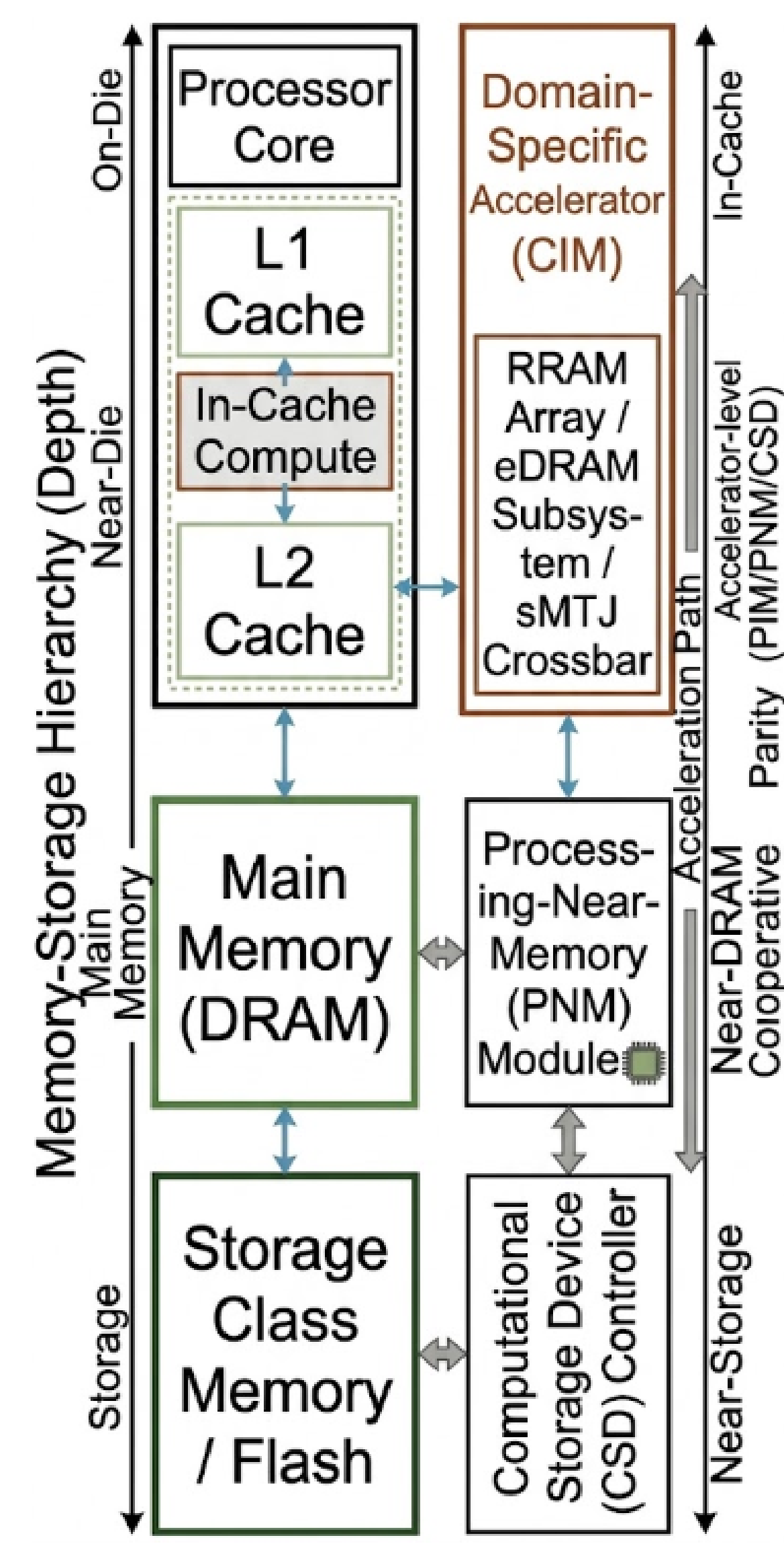


Mufeng Chen , Purdue University (Advisor: Prof. Haitong Li)

Analog eDRAM-RRAM for ZO-Finetuning

- Hybrid memory system design have very unique challenges like the noise robustness for analog edge system to thermal management and bandwidth utilization problem in data center equipped with HBM-HBF system.



(a) write path read path

Meta-Train **Meta-Test**

(b) activation_k norm/initialize activation_{k+1}

(c) **Matrix chain computation diagram**

TABLE I
CONTROLLED COMPARISON OF HIERARCHICAL HRR SIDE-MEMORY PLACEMENT AND FINAL-LAYER-ONLY HEADS WITH A FROZEN POST-TAP BACKBONE AND WITHOUT PROTOTYPE-COUNT CORRECTION. EACH ENTRY REPORTS $H_T/AVGACC$ (%).

Dataset	Backbone	HRR Side (Ctrl.)	Linear Head	HRR Head
CIFAR-100	DeiT-T	51.61 / 66.55	55.31 / 61.05	55.99 / 64.15
	DeiT-S	59.34 / 72.36	61.52 / 67.15	62.27 / 69.98
	DeiT-B	61.67 / 74.91	60.15 / 64.88	64.80 / 71.94
	Swin-T	49.26 / 68.79	57.10 / 65.47	58.25 / 67.53
	ConvNeXt-B	62.34 / 74.45	54.99 / 62.10	57.79 / 67.58
CUB-200	DeiT-T	54.40 / 64.16	46.22 / 52.72	46.36 / 52.65
	DeiT-S	53.36 / 65.38	45.46 / 52.77	51.78 / 59.81
	DeiT-B	54.05 / 66.66	38.86 / 46.00	53.24 / 61.74
	Swin-T	48.83 / 62.42	50.16 / 57.69	48.21 / 55.48
	ConvNeXt-B	49.26 / 63.01	33.91 / 42.74	32.64 / 40.67
Omniglot	DeiT-T	58.00 / 62.06	0.00 / 6.46	0.22 / 9.00
	DeiT-S	61.65 / 64.58	0.06 / 10.48	4.21 / 13.40
	DeiT-B	67.33 / 69.85	0.00 / 12.29	0.06 / 10.32
	Swin-T	62.44 / 65.38	0.28 / 10.00	2.92 / 12.30
	ConvNeXt-B	65.35 / 68.68	0.00 / 11.50	1.18 / 11.89

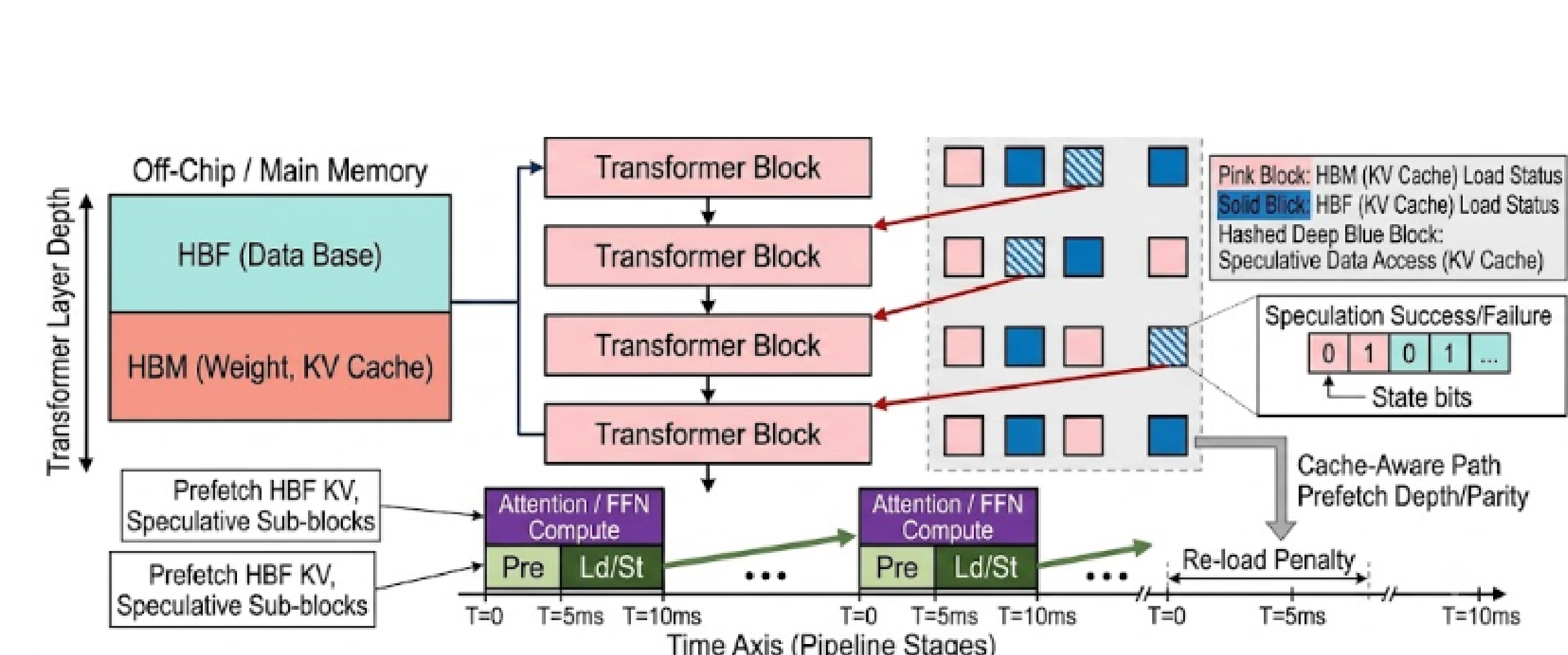
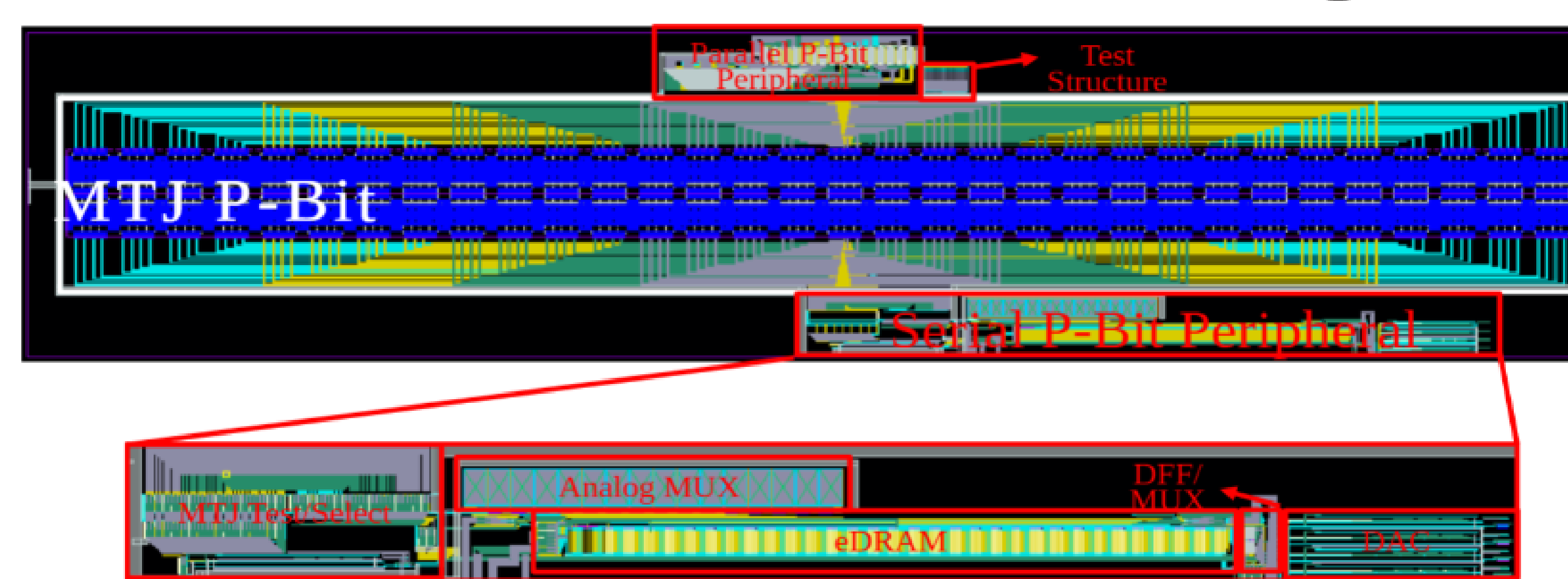
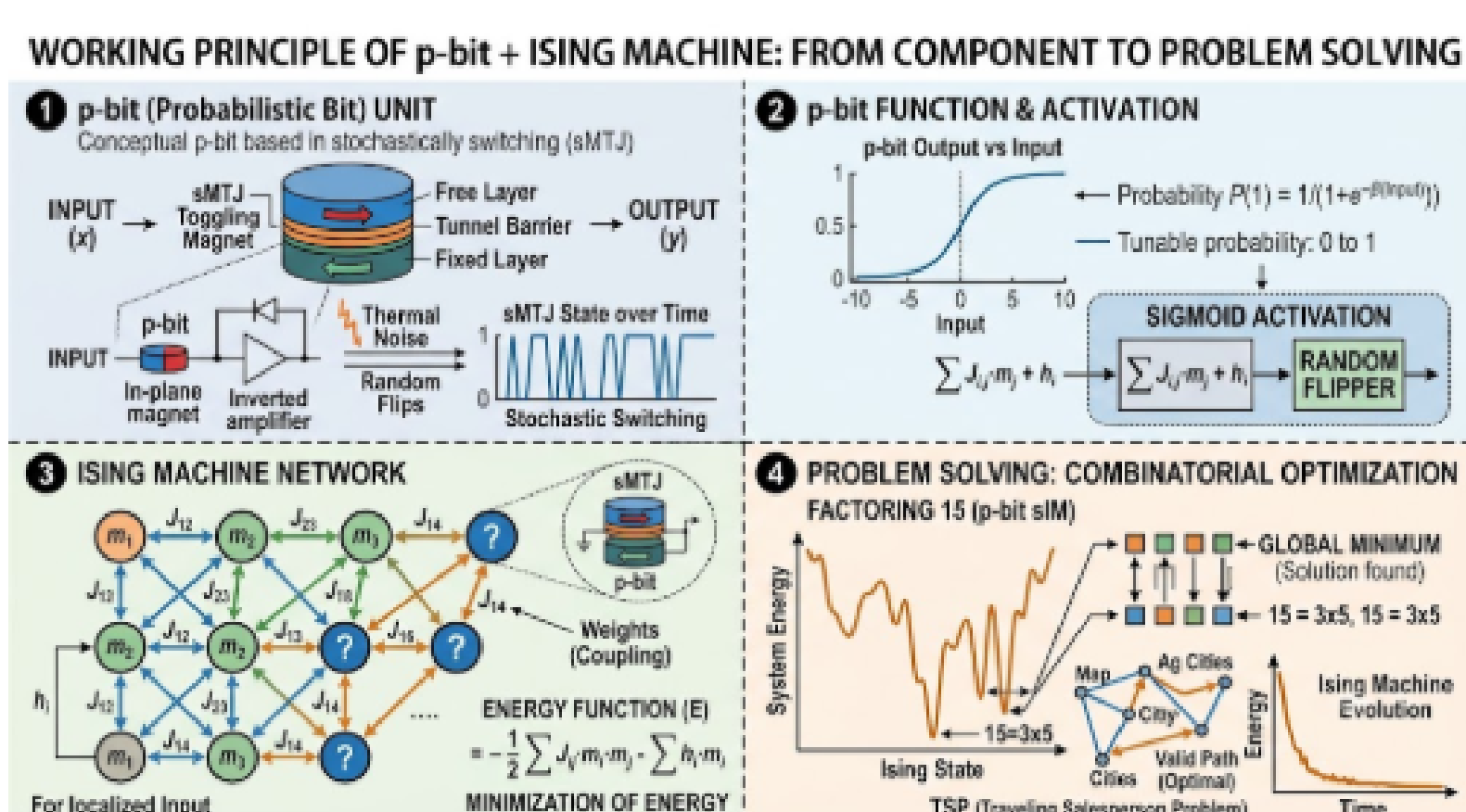
TABLE II
ARCHITECTURE COMPARISON UNDER THE DATASET-SPECIFIC FULL EVALUATION RECIPE, WITH PROTOTYPE-COUNT CORRECTION ENABLED FOR ALL CONFIGURATIONS. EACH ENTRY REPORTS $H_T/AVGACC$ (%).

Dataset	Backbone	Sep.	Shared	XAttn-Sep.
CIFAR-100	DeiT-T	59.97 / 67.15	59.04 / 65.87	56.30 / 63.11
	DeiT-S	65.74 / 73.11	65.00 / 72.45	60.95 / 67.75
	DeiT-B	68.28 / 75.73	68.06 / 75.43	50.73 / 59.20
	Swin-T	54.80 / 61.47	56.19 / 69.48	56.54 / 66.92
	ConvNeXt-B	67.17 / 74.72	66.88 / 74.42	66.90 / 74.08
CUB-200	DeiT-T	56.50 / 63.77	57.66 / 64.47	57.21 / 64.17
	DeiT-S	55.25 / 65.15	56.19 / 65.43	55.76 / 65.95
	DeiT-B	58.27 / 68.09	58.52 / 67.72	59.94 / 68.41
	Swin-T	51.34 / 63.78	50.49 / 62.63	52.69 / 62.44
	ConvNeXt-B	51.32 / 62.94	50.46 / 62.82	51.24 / 60.86
Omniglot	DeiT-T	56.94 / 61.25	55.13 / 59.11	37.79 / 38.53
	DeiT-S	62.52 / 65.45	60.40 / 63.22	61.78 / 63.98
	DeiT-B	65.99 / 68.79	66.50 / 69.07	66.21 / 68.57
	Swin-T	61.67 / 65.13	60.59 / 63.66	53.16 / 52.73
	ConvNeXt-B	65.78 / 69.24	63.01 / 66.73	40.70 / 43.27

(d) **Normal Mode** **Transposable Mode**

(e) **Energy Consumption Optimization**

HBM-HBF System For LLM



- [1] M. Ha et al., "H3: HBM-HBF Hybrid Architecture for Cost-Efficient LLM Inference," IEEE CAL, 2026.
- [2] J. Hong et al., "Bandwidth-Effective DRAM Cache for GPUs with SCM," HPCA, 2024.
- [3] D. Kim et al., "V-REX: Dynamic KV-Cache Retrieval for Streaming Video LLMs," HPCA, 2026.
- [4] H. Zhang et al., "G10: Unified GPU Memory–Storage with Smart Tensor Migration," MICRO, 2023.
- [5] Z. Yu et al., "Cambricon-LLM: Chiplet-Based On-Device 70B LLM Inference," MICRO, 2024.

This work was supported in part by the U.S. NSF under Award No. 2425498 with industry partners as specified in the Future of Semiconductors (FuSe2) program, and also in part by an NSF CAREER Award (No. 2543547).