National Yang Ming Chiao Tung University
Computer Architecture & System Lab

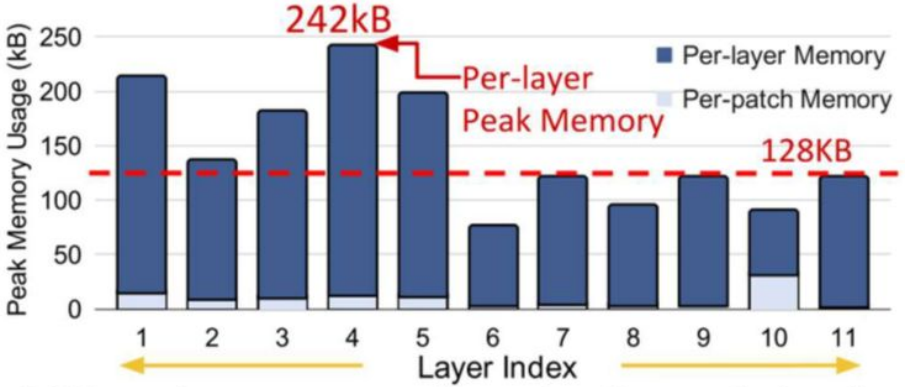
StreamNet: Memory-Efficient Streaming Tiny Deep Learning Inference on the Microcontroller

Hong-Sheng Zheng, Chen-Fong Hsu, Yu-Yuan Liu, Tsung Tai Yeh

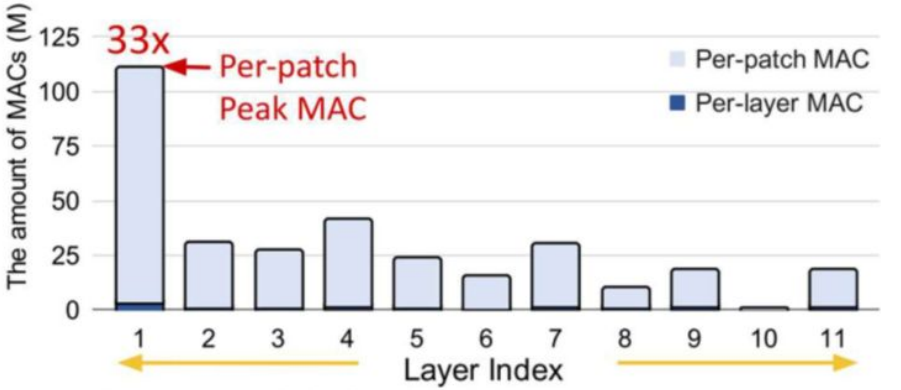
Department of Computer Science
National Yang Ming Chiao Tung University, Taiwan
{hszheng.cs08, fonghsu.cs08, yyliu.cs11, ttyeh14}@nycu.edu.tw

Motivation

- Layer-wise DNN inference: High peak SRAM memory usage
- Patch-based DNN inference: **High computational overhead** when the peak SRAM memory usage is getting to be small



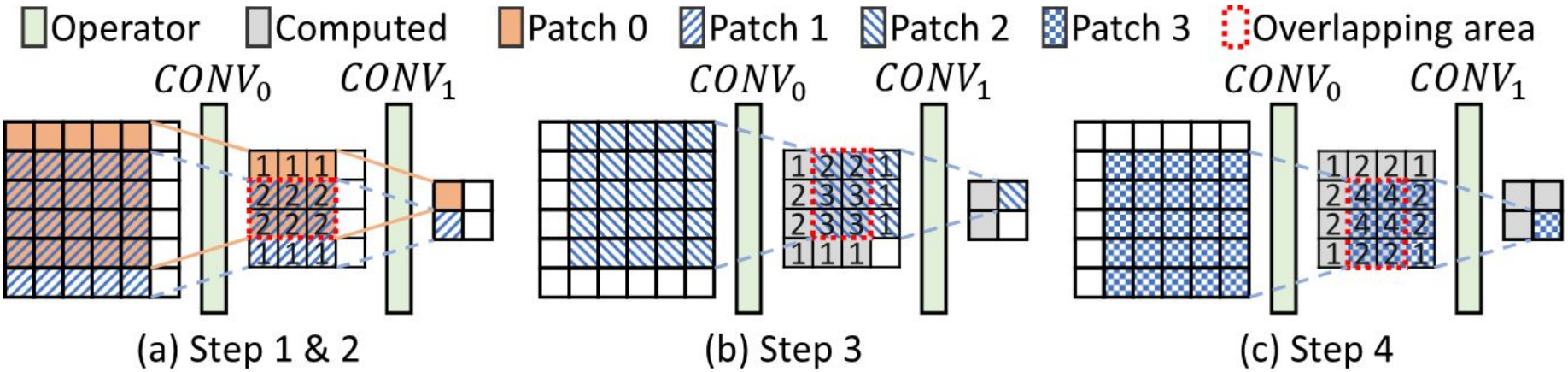
(a) The peak memory usage of cascade policy over the baseline



(b) The amount of MAC in cascade policy over the baseline

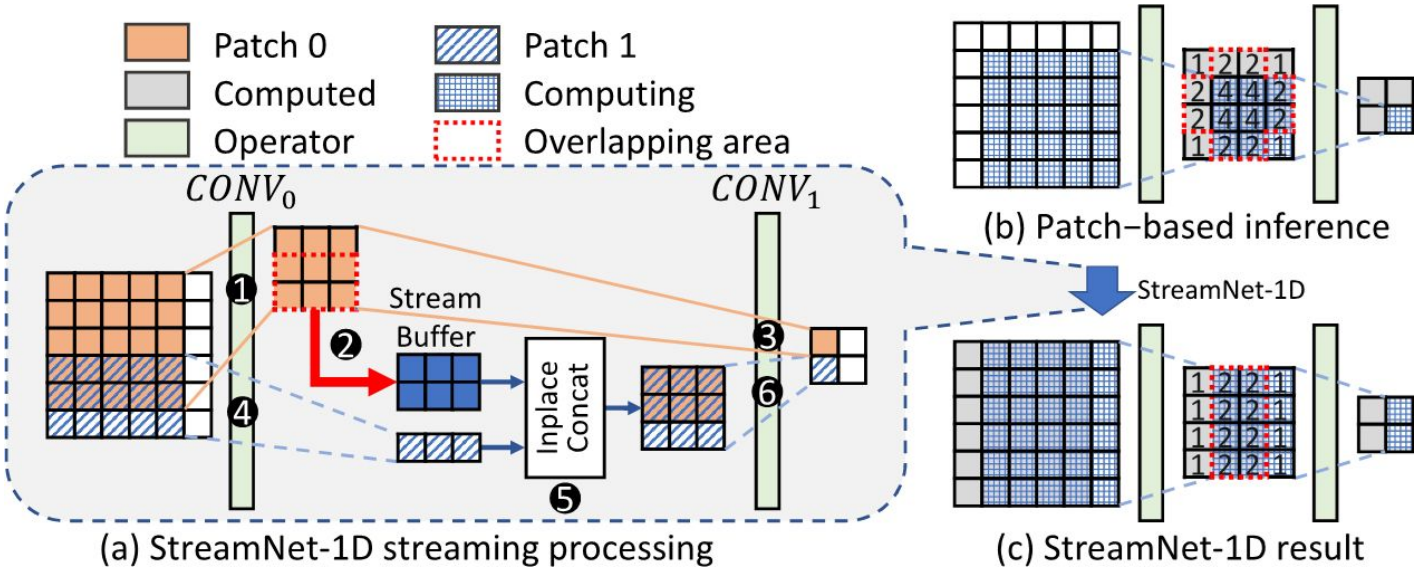
Patch-based Inference

- Only store small intermediate tensor patches
  - Reduce peak SRAM memory usage
- Redundant computation -> overlapping patches



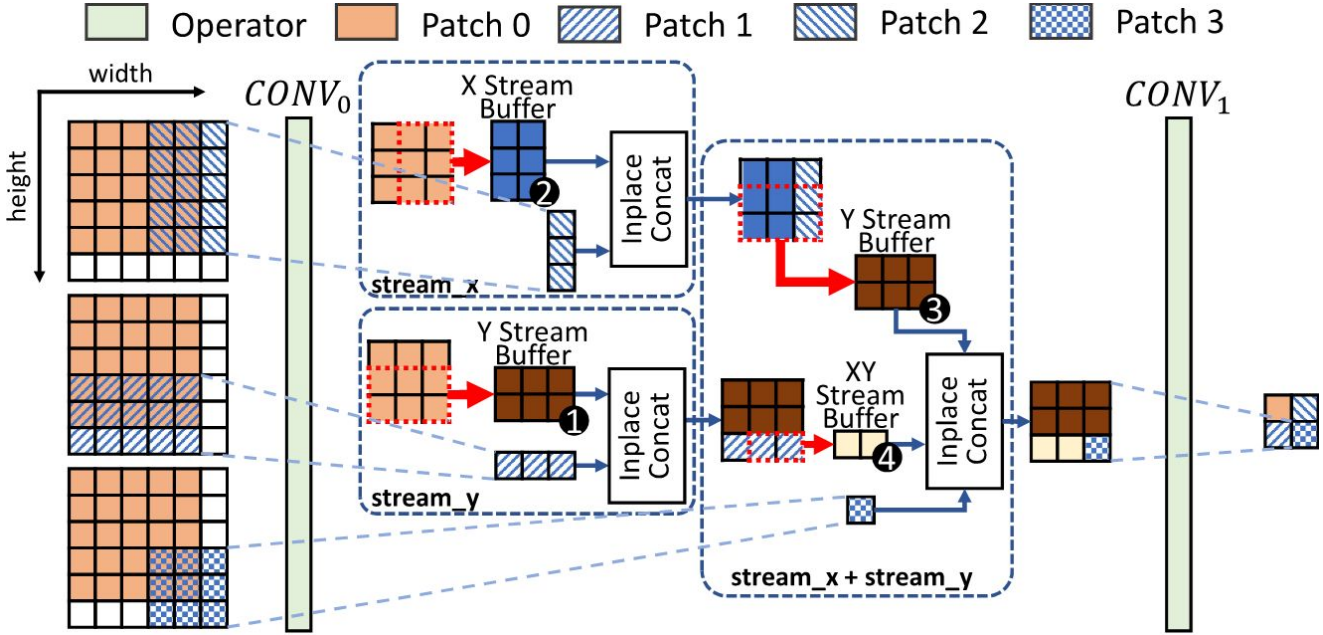
StreamNet DNN Inference

- Mitigate the patch-based inference overhead
- Using stream buffer to reduce the unnecessary patch computation



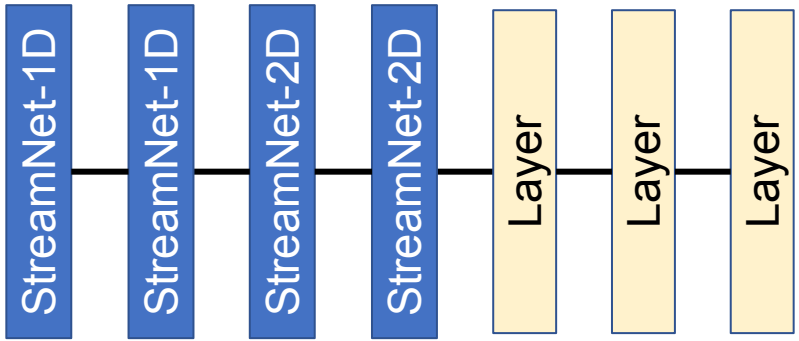
StreamNet in 2D Streaming Processing

- Stream2D further improves the latency of patch-based inference
- Remove more redundant computation by using x/y stream buffer



Hybrid StreamNet 1D and 2D Processing

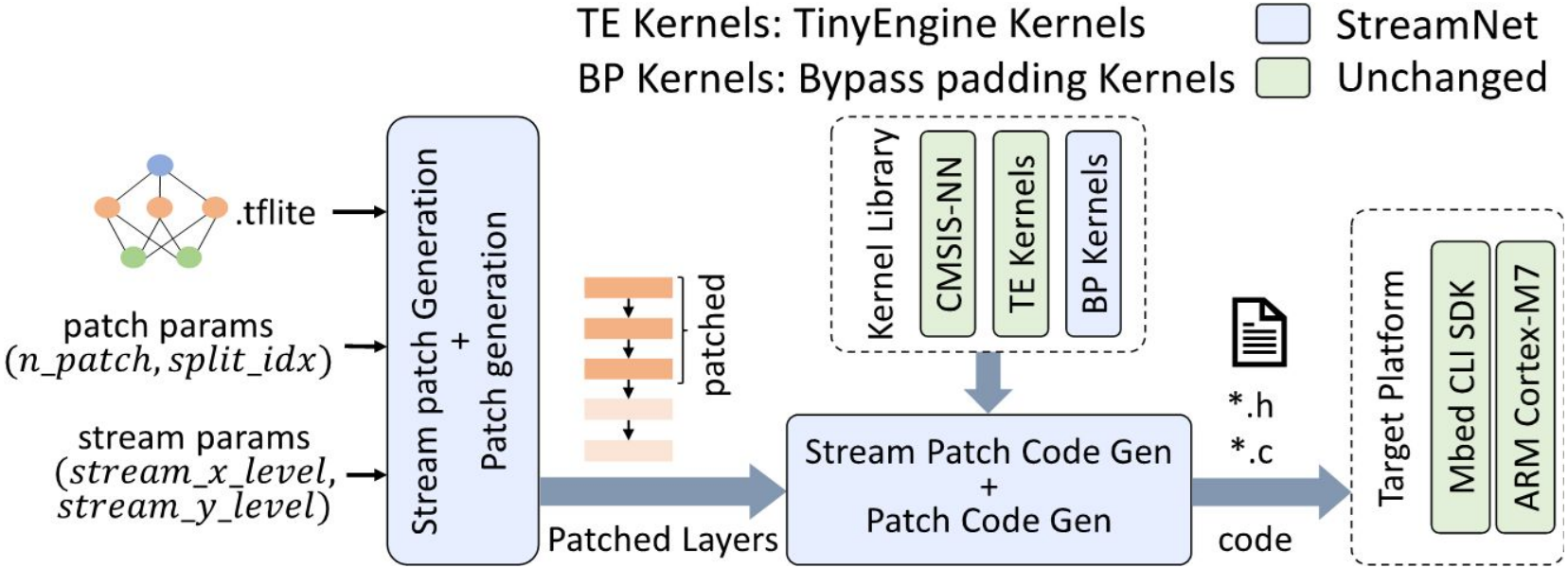
- To meet the microcontroller (MCU) hardware requirement
  - Automatric parameter selection for DNN models
  - StreamNet 1D: lower peak SRAM usage, higher inference latency
  - StreamNet 2D: higher peak SRAM usage, lower inference latency



A DNN model mixing with streamNet 1D/2D and normal layer

StreamNet System Architecture

- Compiler-based tensor memory planner
  - Schedule execution of DNN inference with stream buffers



StreamNet Inference Latency Evaluation

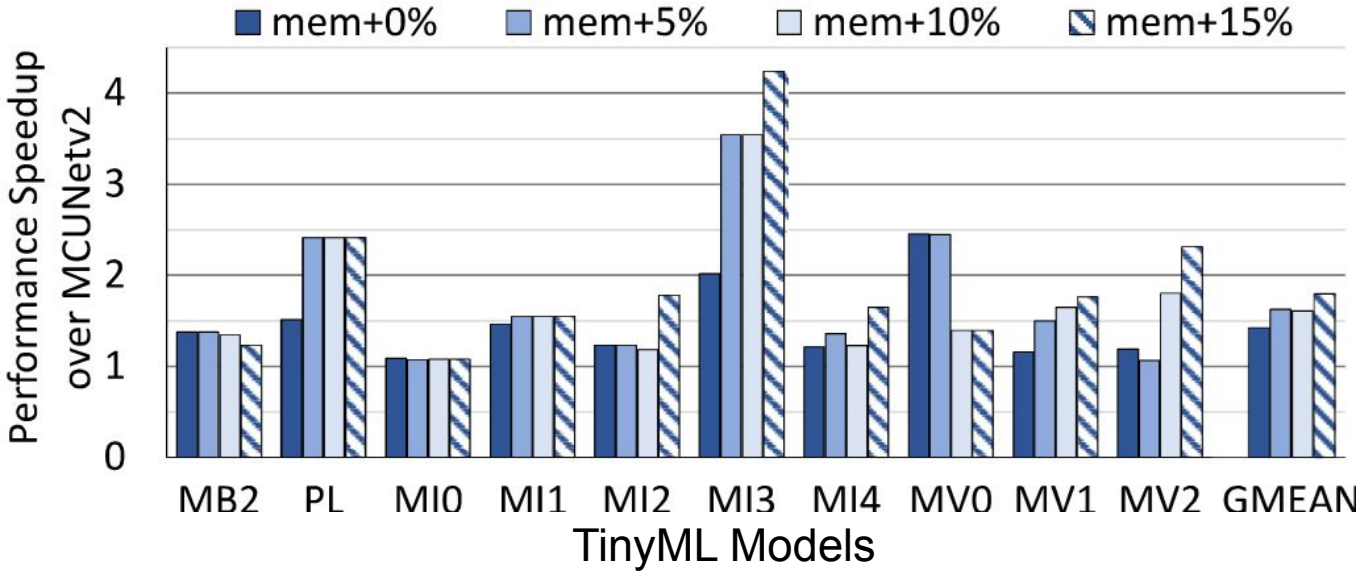
- StreamNet-2D achieves a geometric mean of 7.3X speedup over the patch-based inference
- The MCU contains
  - ARM Cortex-M7 CPU
  - 512 KB SRAM
- StreamNet removes redundant computation of patch-based inference

Table 3: The latency (ms) of TinyML models on StreamNet and MCUNetv2

					Speedup over MNv2		
Model	MNv2	MNv2+	SN-1D	SN-2D	MNv2+	SN-1D	SN-2D
MB2	1,514	1,391	687	417	1.09	2.20	3.63
PL	6,296	5,795	1,610	676	1.09	3.91	9.31
MI0	238	219	131	99	1.09	1.81	2.40
MI1	449	413	258	188	1.09	1.74	2.39
MI2	16,545	15,055	3,281	1,168	1.10	5.04	14.17
MI3	29,700	27,331	4,858	1,444	1.09	6.11	20.57
MI4	26,816	24,747	5,153	1,762	1.08	5.20	15.22
MV0	415	355	169	101	1.17	2.45	4.10
MV1	1,817	1,667	473	225	1.09	3.84	8.09
MV2	14,106	13,040	2,778	961	1.08	5.08	14.68
GMEAN					1.10	3.40	7.28

StreamNet Performance on Extremely Small SRAM

- Varying performance when increasing SRAM memory budget based on the minimal SRAM memory usage of StreamNet-1D



Conclusion

- StreamNet eliminates the performance bottleneck of patch-based inference with small SRAM usage increase
- StreamNet-1D and 2D meets different hardware requirements
- StreamNet achieves a geometric-mean speedup of 7.3X over patch-based inference across 10 TinyML models
- StreamNet-1D and 2D reduce 62 % and 47% peak SRAM memory usage over the layer-wise inference

Thank you