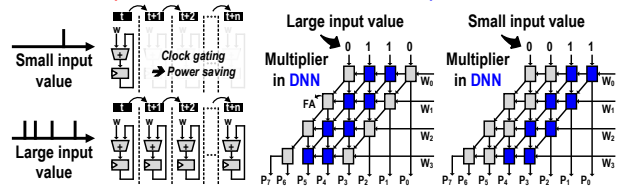
A Low-power Large-Language-Model Processor with Big-Little Network and Implicit-Weight-Generation for On-device AI

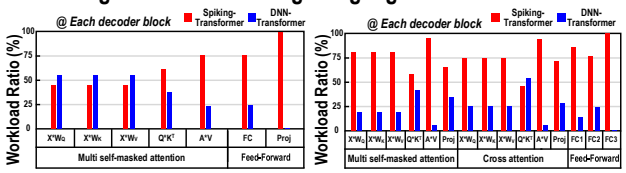
Sangyeob Kim, Sangjin Kim, Wooyoung Jo, Soyeon Kim, Seongyon Hong, Nayeong Lee, and Hoi-Jun Yoo
KAIST, Daejeon, Republic of Korea

Motivation

<Complementary DNN: Combining SNN and DNN for low power>

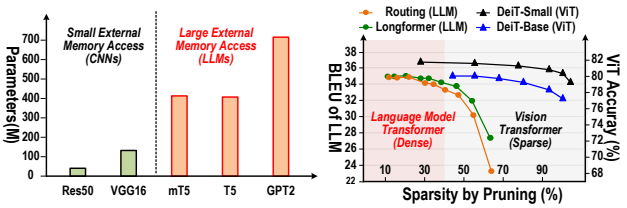


<Challenges of C-DNN for large language model in on-device>



<Language Modeling w/ GPT-2 and Wikitext> <Translation w/ mT5 and IWSLT>

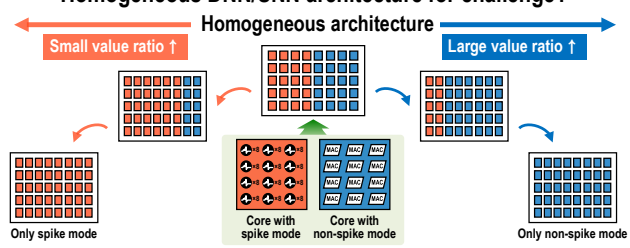
Challenge1: Ratio of DNN and SNN are dynamically changed



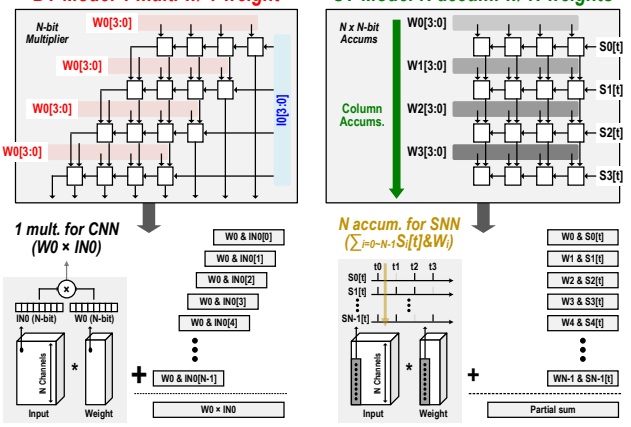
Challenge2: Large external memory access due to large parameters

Homogeneous Architecture

<Homogeneous DNN/SNN architecture for challenge1>



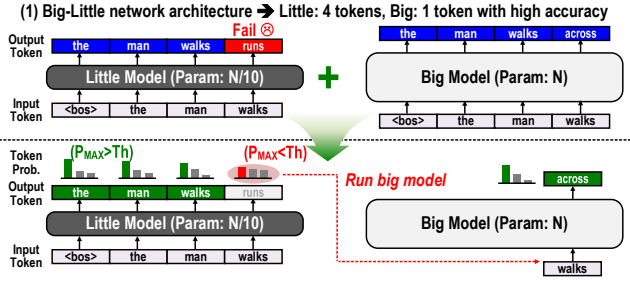
<Hybrid multiplication and accumulation unit in the core>



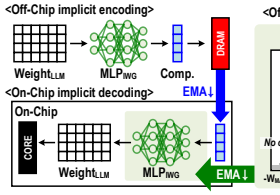
3 Stage Compression

<3 stage compression technique for challenge2>

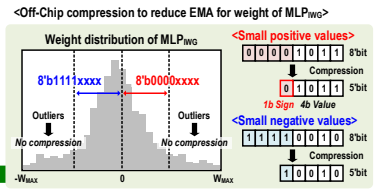
(1) Big-Little network architecture ➔ Little: 4 tokens, Big: 1 token with high accuracy



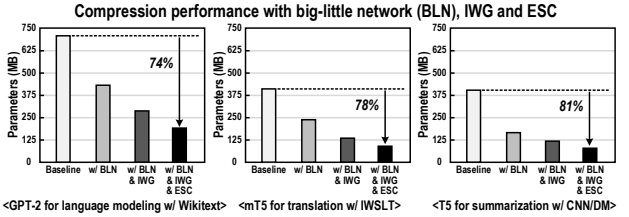
(2) Implicit weight generation (IWG)



(3) Extended sign compression (ESC) for MLP_{IWG}

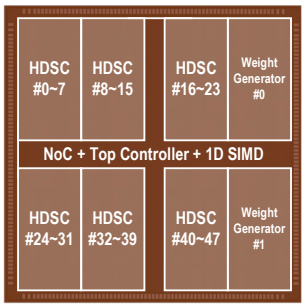


Compression performance with big-little network (BLN), IWG and ESC

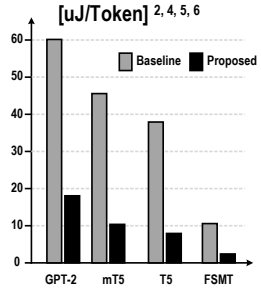


<GPT-2 for language modeling w/ Wikitext> <mT5 for translation w/ IWSLT> <T5 for summarization w/ CNN/DM>

Measurement Results



System energy consumption [uJ/Token] [2, 4, 5, 6]



	Specifications			
Technology	Samsung 28nm 1P8M CMOS			
Die Area	20.25 mm²			
Voltage	0.7 – 1.1 V			
Frequency	50 – 200 MHz			
SRAM	500 KB			
Data Precision	INT8			
Mode	DNN-Transformer / Spiking-Transformer			
System Power (mW)	47.5 @ 50MHz, 0.7V			
	469.2 @ 200MHz, 1.1V			
Peak Performance [1]	3.41 TOPS			
Energy Efficiency (TOPS/W) [2, 3]	22.9 – 47.8 @ (50 MHz, 0.7 V)			
External Bandwidth (GB/s)	1.6 GB/s			
Model Type	GPT-2	mT5	T5	FSMT
Task	Language Modeling	Translation	Summary	Translation
System Energy Consumption [4, 5, 6] (uJ/Token)	18.1	10.4	7.2	2.6
Latency [1, 4, 6] (ms)	477	281	229	60

1) @ 200 MHz, 1.1V, 2) 2 Ops = 1 MAC (DT-mode) = Integrate-and-Fire during time-steps (ST-mode)
3) EMA is excluded, 4) EMA is included (with DDR3 interface),
5) @ 50 MHz, 0.7 V, 6) with big-little network, IWG, ESC and 1024 tokens