BAQET: BRAM-aware Quantization for Efficient Transformer Inference via Stream-based Architecture on an FPGA

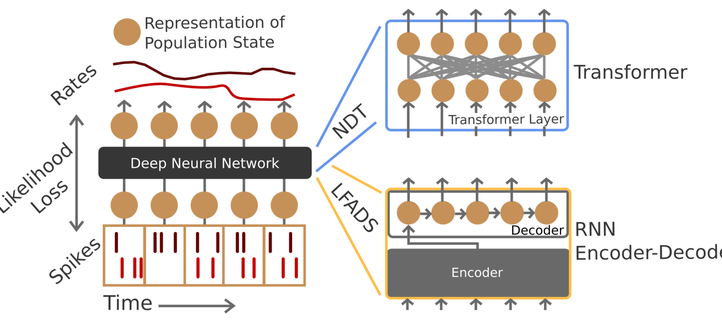
Ling-Chi Yang[1], Chi-Rui Chen[1], Trung Le[2], Bo-Cheng Lai[1], Scott Hauck[2], Shih-Chieh Hsu[2]

[1] National Yang Ming Chiao Tung University, Taiwan | [2] University of Washington, USA

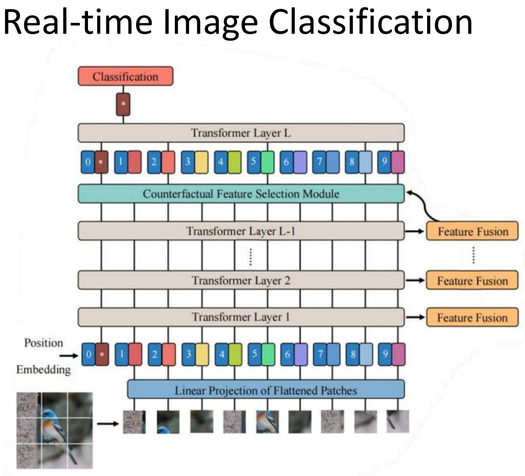
Background

Transformer on GPU can't handle low-latency applications

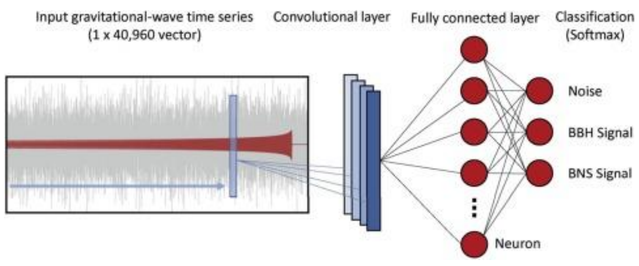
- Representation Learning for Neural Population Activity[1]



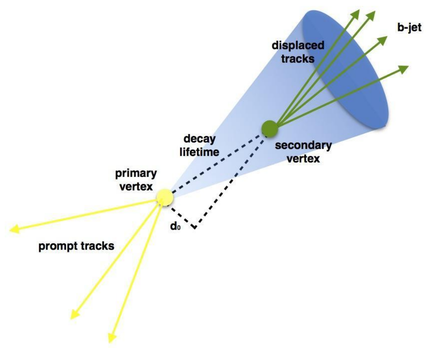
- Real-time Image Classification



- Gravitational Wave Anomaly Detection



- B-tagging Detection



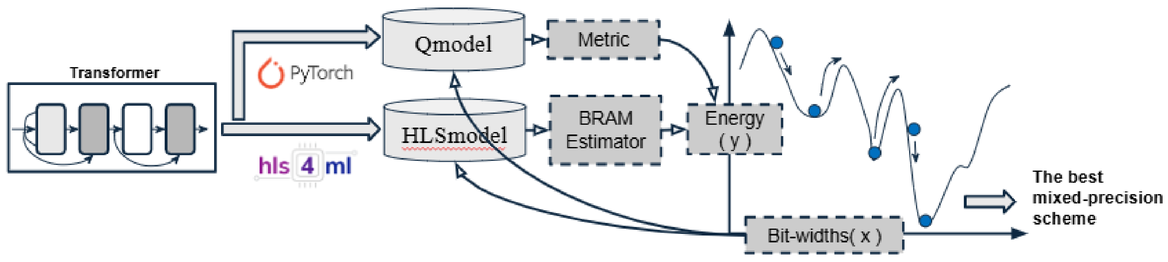
Challenge & Contribution

Challenges

- Large memory footprint of fully on-chip design
- Hard to determine mixed-precision scheme due to large search space

Contributions

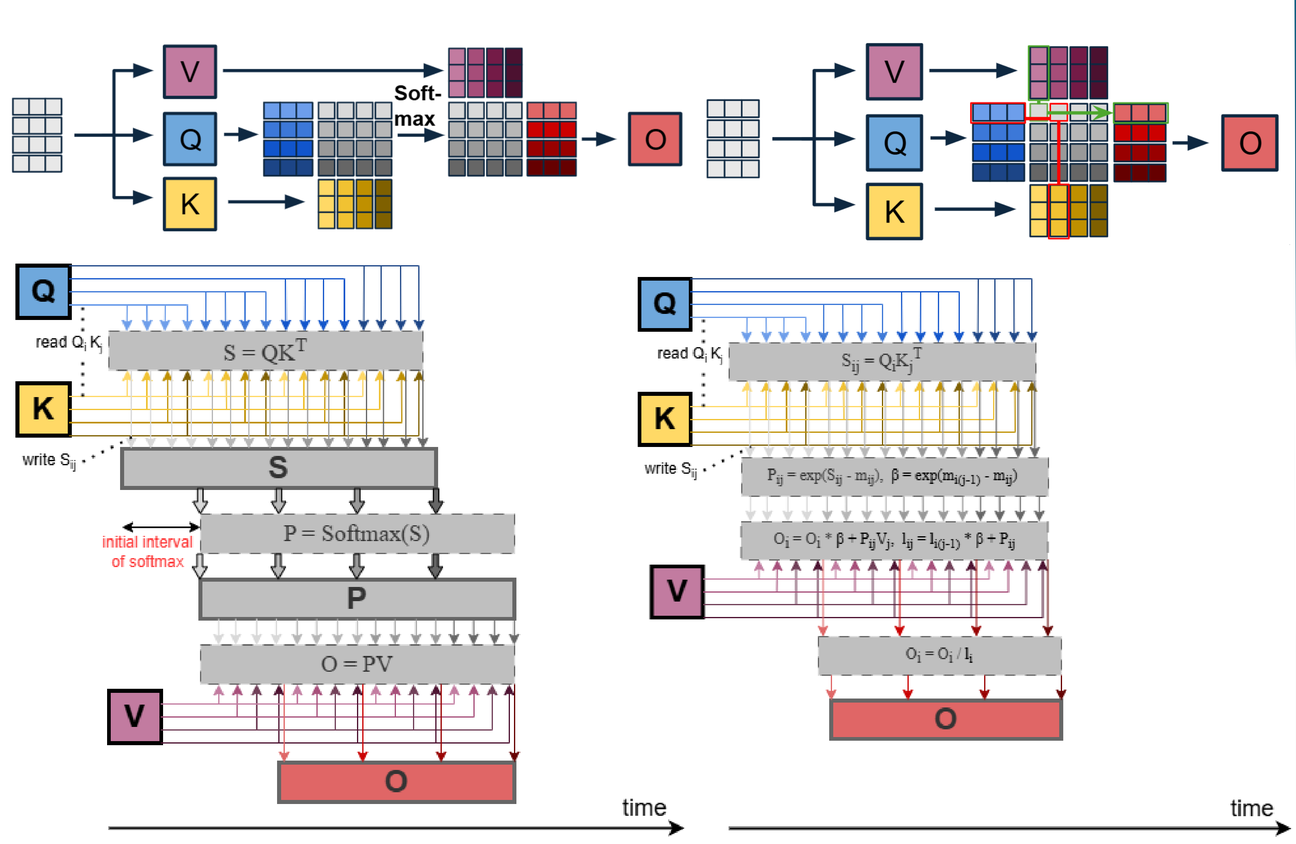
- Improved **Flash Attention[2]** algorithm in **stream-based architecture** on FPGA
- Estimating the number of BRAMs in hls4ml and formulating the problem, adopting **simulated annealing** to find better solution.



Algorithm Optimizations

Standard attention

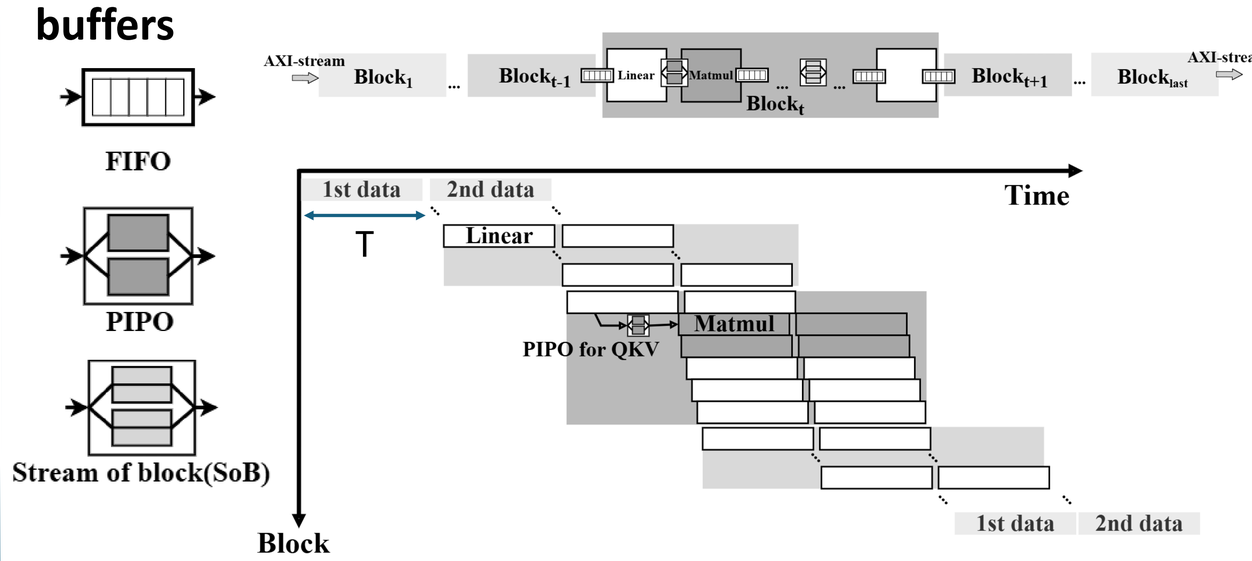
Flash attention



- Saved memory footprint of attention and improved latency of Softmax

Stream-based Architecture

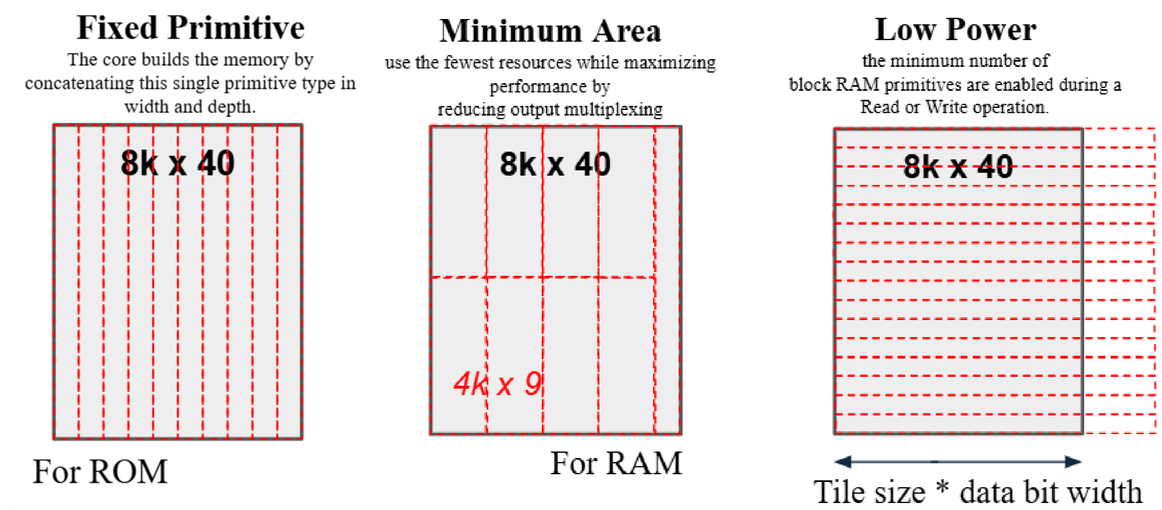
Boost performance and save memory footprint with various buffers



- Overall latency for single batch = T * (number of transformer layer +1)

BRAM-aware Quantization

- Formula : $Q(x) = \frac{bitwidth_{int}}{bitwidth_{total}} \cdot \text{Clip}(\text{Round}(\frac{bitwidth_{total}}{bitwidth_{int}} \cdot x))$
- BRAM estimation



- Cost function of multi-objective optimization

$$E = \text{penalty} \times \left(\frac{\#\text{ of BRAM}}{\text{available }\#\text{ of BRAM}}\right) + \alpha \times \left(\frac{\pm\text{ accuracy}}{\text{best accuracy}}\right)$$

- The more BRAM used, the worse the timing closure. A larger penalty is needed to reduce BRAM usage.
- Users can adjust the alpha value based on different applications to give more weight to accuracy.

Experiments

- FPGA board : AMD Xilinx Virtex Ultrascale+ U55C
- Development platform : Vitis HLS 2023.2

Impact of tile size

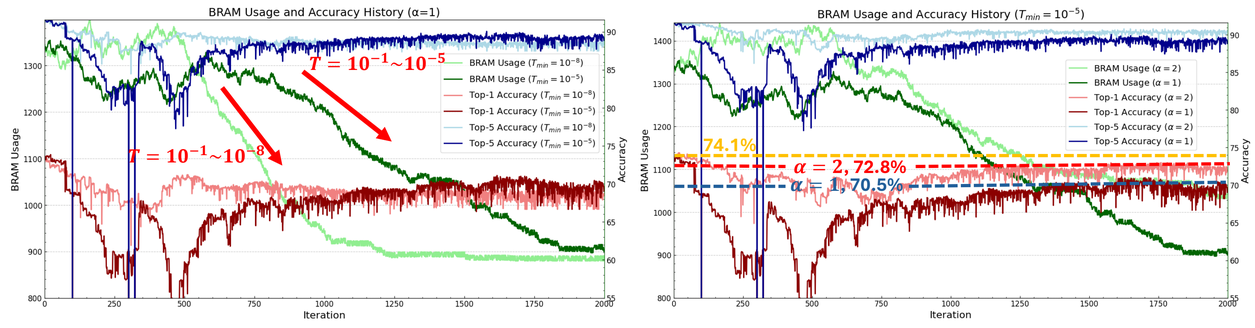
- Larger tile size need higher bandwidth of RAM

	B_{dk}, B_{ffn}				
	1, 12	2, 24	4, 48	8, 96	16, 192
BRAM	73.96%	76.26%	69.44%	53.08%	99.80%
DSP	10.28%	15.69%	26.48%	47.57%	89.27%
LUT	12.19%	17.09%	23.57%	38.32%	72.66%
FF	3.30%	4.15%	6.34%	9.10%	16.41%
URAM	30%	30%	30%	60%	60%
Interval (ms)	12.54	6.27	3.136	1.56	0.78
Power (W)	12.03	14.332	18.23	28.32	51.203
Latency (ms)	36.87	18.441	9.22	4.61	2.3
WNS	0.189	0.041	0.031	-0.038	-1.516
Efficiency (GOPS/W)	16.2406	27.2641	42.8552	55.4560	61.3446

- Higher tile size can improve power efficiency

Work	Model		Board	Frequency (Mhz)	Power (W)	Throughput (GOPS)	Efficiency (GOPS/W)
FTRANS	RoBERTa		VCU118	N/A	25	170	6.8
ViA	Swin-T		U50	300	39	309.6	7.94
SwiftTron	DeiT-S		ASIC	143	33.64	407	12.09
SWAT	Swin-T		U50	200	14.35	301.9	21.04
Ours	NDT	7,16	U55C	200	7.161	166.27	23.21
		13,32			9.086	308.77	33.98
	DeiT-T	4,48			18.23	781.15	42.85
		8,96			28.32	1563.67	55.21

Impact of alpha and temperature range



[1] Trung Le and Eli Shlizerman. 2022. Stndt: Modeling neural population activitywith spatiotemporal transformers. Advances in Neural Information ProcessingSystems 35 (2022), 17926–17939.

[2] Tri Dao. 2023. FlashAttention-2: Faster Attention with Better Parallelism andWork Partitioning. https://doi.org/10.48550/arXiv.2307.08691.