22.9 A 12nm 18.1TFLOPs/W Sparse Transformer Processor with Entropy-Based Early Exit, Mixed-Precision Predication and Fine-Grained Power Management

Thierry Tambe[1], Jeff Zhang[1], Coleman Hooper[1], Tianyu Jia[2], Paul N. Whatmough[1,3], Joseph Zuckerman[4], Maico Cassel Dos Santos[4], Erik Jens Loscalzo[4], Davide Giri[4], Kenneth Shepard[4], Luca Carloni[4], Alexander Rush[5], David Brooks[1], Gu-Yeon Wei[1]

[1]Harvard University, Cambridge, MA
[2]Peking University, Beijing, China
[3]ARM, Boston, MA
[4]Columbia University, New York, NY
[5]Cornell University, New York, NY

Large language models have substantially advanced nuance and context understanding in natural language processing (NLP), further fueling the growth of intelligent conversational interfaces and virtual assistants. However, their hefty computational and memory demands make them potentially expensive to deploy on cloudless edge platforms with strict latency and energy requirements. For example, an inference pass using the state-of-the-art *BERT-base* model must serially traverse through 12 computationally intensive transformer layers, each layer containing 12 parallel attention heads whose outputs concatenate to drive a large feed-forward network [1]. To reduce computation latency, several algorithmic optimizations have been proposed, e.g., a recent algorithm dynamically matches linguistic complexity with model sizes via entropy-based early exit [2]. Deploying such transformer models on edge platforms requires careful co-design and optimizations from algorithms to circuits, where energy consumption is a key design consideration.

We present a 4.60mm² sparse transformer processor (STP) that efficiently accelerates transformer workloads by tailoring its latency and energy expenditures according to the complexity of the input query it processes. Key contributions of this work are as follows: (1) A specialized datapath for entropy-based early exit assessment reduces BERT latency by up to 6.13×, e.g., inferences terminate early at an average early exit layer of 3.90 (out of 12) for the SST-2 NLP benchmark; (2) a mixed-precision (MP) FP4/FP8 MAC supports per-vector exponent biases during 4b floating point (FP4) computations, allowing the processor to double its throughput, while reducing its energy consumption and maintaining high inference accuracy, depending on the entropy; and (3) a fine-grained sentence-level power management scheme opportunistically scales the accelerator's supply voltage and clock frequency while meeting an application's end-to-end latency target. Together, the proposed STP achieves a peak efficiency of 65mJ/inf, a 7.14× energy improvement, on average, over conventional BERT inference without the key innovations.

Figure 22.9.2 lays out the STP's overall architecture. The main computation engine comprises a mixed-precision MAC unit that selects between FP4 vs. FP8 execution based on the entropy characteristics of the input sentence. The entropy information further guides the power management scheme, which dynamically adjusts the accelerator's local supply voltage via cell-based PMOS power headers of a free running LDO (i.e. no feedback loop). Sixteen pre-characterized look-up table (LUT) entries, stored in a 32KB auxiliary buffer of the special function unit (SFU), control the effective resistance of the PMOS power headers. A configurable digitally-controlled oscillator (DCO) runs off of the local supply voltage to ensure proper self-clocked operation. The SFU further contains specialized datapaths to provide vectorized computations for various non-linear functions (i.e., *softmax*, layer normalization), element-wise addition, as well as the entropy computation. The SFU's auxiliary buffer also stores normalization parameters and attention pruning metadata which specify the span of each attention head. Notably, the STP entirely skips the computation of attention heads with null span, further cutting inference latency and energy. A pair of bit-mask decoders and a bit-mask encoder decompress and compress sparse matrices, respectively, before and after each matrix-matrix multiplication, which enables compact storage of non-zero data along with their associated binary tags in the 256KB data SRAM and 32KB mask SRAM.

Entropy is a statistical measure of classification confidence. Linguistically complex sentences typically produce higher entropy values compared to simpler sentences. In this work, we leverage the computed entropy of each input query the accelerator processes in order to dynamically determine whether to compute subsequent transformer layers or to exit early (Fig. 22.9.1 bottom). The inference exits when the calculated entropy value falls below a pre-defined threshold after the computation of each transformer layer. Furthermore, after executing the first transformer layer, the entropy of the query is employed to predicate on the precision (i.e., FP4 or FP8) of the next layers' computation and optimally scale the processor's supply voltage and clock frequency with the goal of minimizing energy consumption given a pre-defined latency target. Fig. 22.9.3 shows the algorithmic and hardware implementations of entropy-based early exit computation. To avoid numerical instability during the calculation of exponents, we reformulated the entropy algorithm by negatively scaling all elements in the early exit vector by the max score. Fig. 22.9.3 also plots the per-sentence correlation between the 1st layer's computed entropy and the achieved voltage and frequency scaling.

Figure 22.9.4 illustrates the details of the mixed-precision (MP) floating-point MAC unit comprising 16 lanes of FP8 (E4M3) vector datapaths with coarse-grained per-tensor exponent bias scaling, as well as 16 lanes of FP4/LOG4 (E3M0) datapaths with fine-grained per-vector exponent bias scaling. Similar to recent work that leverages per-vector scale factors during INT4 quantization [5], the STP uses per-vector exponent biases during FP4 quantization to avoid steep accuracy loss. The SFU entropy unit conditionally selects between the FP8 and FP4 datapaths. The FP4 MAC lanes increase throughput by 2× and achieve higher energy efficiency, compared to the FP8 MAC datapath, with its larger vector size of 32 and the lack of mantissa multipliers. Both MAC datapaths can be gated to skip computations for null vectors and save power.

The sparse transformer processor was fabricated using a 12nm FinFET technology. Fig. 22.9.5 presents the chip measurement results while executing the computations of *BERT-base* on the sentiment analysis (SST-2) dataset. Co-designing the early exit (EE) algorithm allows the accelerator to reduce latency and energy expenditures by 3.71×. Attention head pruning (AP) further extends these savings by 1.5×. The addition of mixed-precision (MP) predication enables the processor to achieve a minimum average inference latency of 682ms. Assuming an end-to-end per-sentence latency constraint of 2 seconds, the STP can opportunistically derate its voltage and frequency (VFS) to further reduce energy consumption, measured down to an average of 65mJ per inference. The horizontal histograms in Fig. 22.9.5 provide further details of the measured per-sentence processing latency and energy results. Combining EE, AP, and MP allows the majority of sentences to execute under 1 second, while VFS allows the accelerator to process most sentences under 50mJ/inf by relaxing the latency target to 2 seconds. In contrast, a conventional implementation would require 4 seconds and 464mJ/inference.

Figure 22.9.6 summarizes evaluation results across three different NLP datasets, demonstrating minimal accuracy loss when optimizations, EE, AP, and MP, are applied. Because MNLI and QQP datasets contain more linguistically complex queries (i.e., higher entropies) compared to SST-2, their average latency and energy consumption results are consequently higher. Fig. 22.9.6 also plots the frequency vs. voltage operation of the 4.60mm² STP and the resulting computational efficiency presented in terms of TFLOPS/W. Functional operation is verified across a 0.62-to-1.0V supply voltage range with 77-to-717MHz clock frequency, while producing an energy efficiency range of 3.0-8.24TFLOPS/W (FP8) and 6.61-18.1TFLOPS/W (FP4). Compared to a recently published LSTM/RNN chip that supports attention [4], this work exhibits competitive figures of merit while providing fine-grained latency and power management tailored to the complexity of the input sentence. Compared to other prior transformer [5-6] accelerator designs, this work supports floating-point operations to maintain higher model accuracy, while exploiting fine-grained power management for further energy savings. Fig. 22.9.7 shows an annotated physical layout of the processor overlayed on a photograph of the flip-chip packaged die.

Acknowledgment:
This work was developed with funding from DARPA, NSF Awards 1704834 and 1718160, Intel Corp., Arm Inc, and the Center for Applications Driving Architectures (ADA), one of six centers of JUMP, a SRC program co-sponsored by DARPA. Any opinions, findings, and/or conclusions expressed in this publication are those of the authors and should not be interpreted as representing the official views of any funding agency.

References:
[1] J. Devlin et al., "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *Assoc. for Computational Linguistics*, pp. 4171-4186, 2019.
[2] T. Tambe et al., "EdgeBERT: Sentence-Level Energy Optimizations for Latency-Aware Multi-Task NLP Inference," *IEEE MICRO*, pp. 830-844, 2021.
[3] D. Kadetotad et al., "An 8.93 TOPS/W LSTM Recurrent Neural Network Accelerator Featuring Hierarchical Coarse-Grain Sparsity for On-Device Speech Recognition," *IEEE JSSC*, vol. 55, no. 7, pp. 1877-1887, 2020.
[4] T. Tambe et al., "A 16-nm SoC for Noise-Robust Speech and NLP Edge AI Inference with Bayesian Sound Source Separation and Attention-Based DNNs," *IEEE JSSC*, 2022.
[5] B. Keller et al., "A 17-95.6 TOPS/W Deep Learning Inference Accelerator with Per-Vector Scaled 4-bit Quantization for Transformers in 5nm," *IEEE Symp. VLSI Circuits*, 2022.
[6] Y. Wang et al., "A 28nm 27.5TOPS/W Approximate-Computing-Based Transformer Processor with Asymptotic Sparsity Speculating and Out-of-Order Computing," *ISSCC*, pp. 464-465, 2022.

2023 IEEE International Solid-State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067817

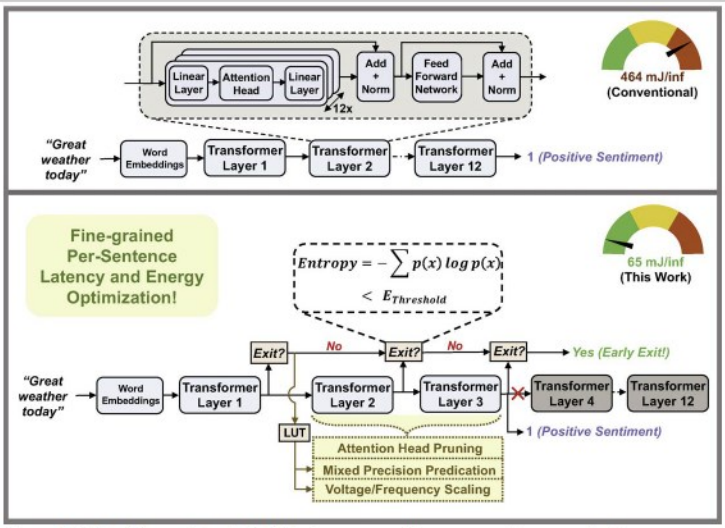


Figure 22.9.1: Conventional BERT inference vs. the proposed approach. The proposed inference optimizes the NLP latency and energy consumption at a sentence granularity via entropy-based early exit, attention head pruning, mixed-precision predication, and V/F scaling.

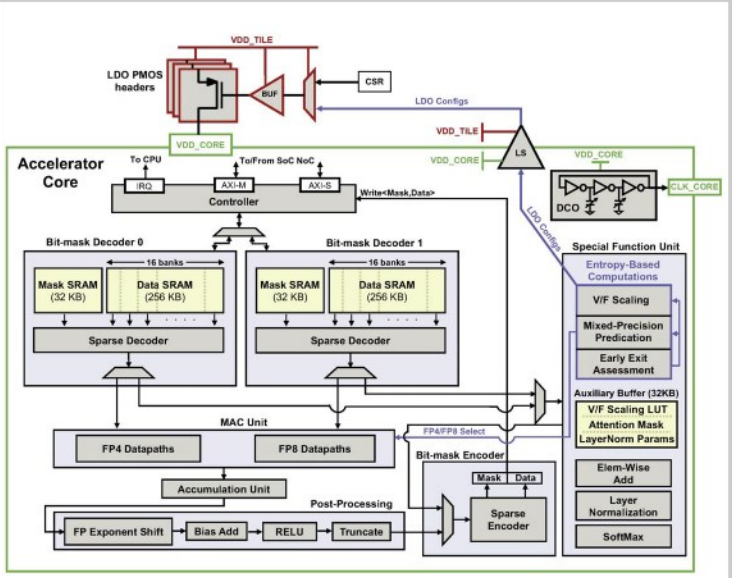


Figure 22.9.2: Overall architecture of the sparse transformer processor (STP).

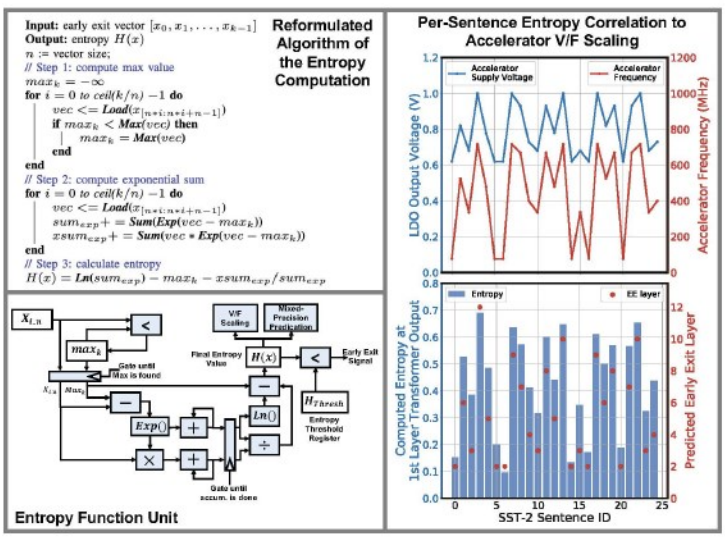


Figure 22.9.3: Implementation of the entropy function whose value on the 1st layer transformer output is used to scale the accelerator's supply voltage and clock frequency.

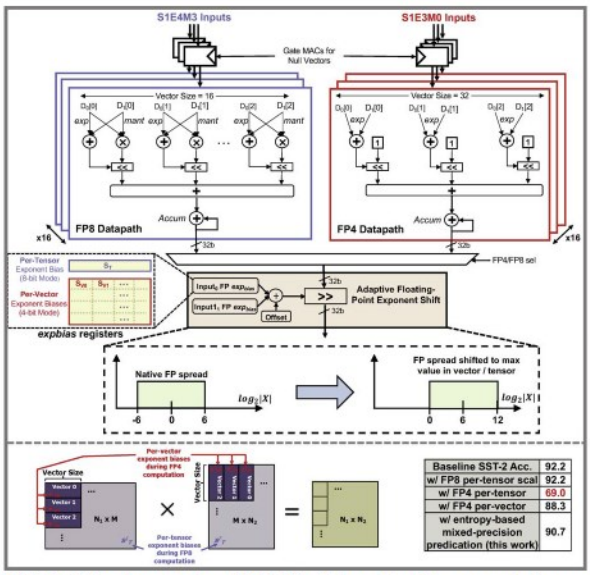


Figure 22.9.4: Mixed-precision FP8/FP4 datapath with support for per-vector exponent bias scaling.

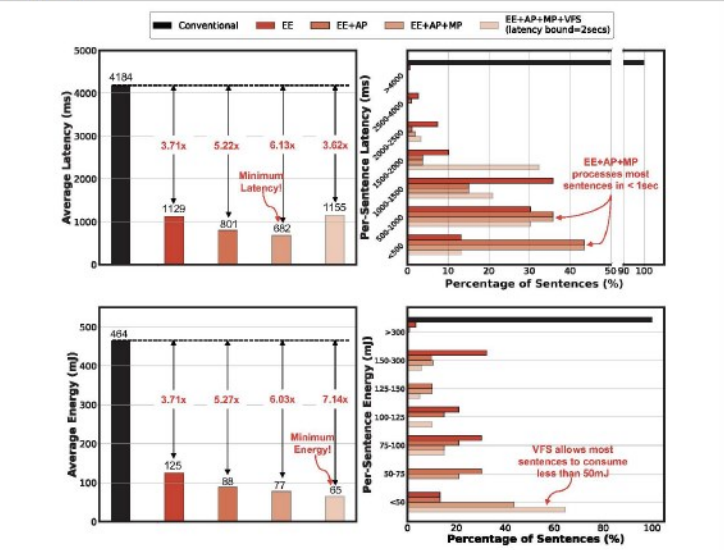
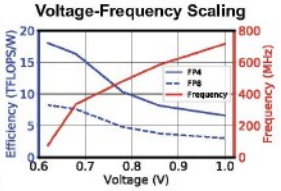


Figure 22.9.5: Measurement results for BERT inference on the SST-2 dataset, highlighting the benefits of early exit (EE), attention head pruning (AP), mixed-precision (MP), and latency-bounded V/F scaling (VFS).

Evaluation on Different NLP Datasets

NLP Dataset	SST-2 Sentiment Analysis	MNLI Textual Entailment	QQP Question Equivalence
Base Accuracy (%)	92.2	85.2	90.8
Accuracy w/ EE+AP+MP (%)	90.3	82.6	89.4
Average Exit Layer	3.90	8.61	5.84
Average BERT-base Latency (ms)	682	1960	1265
Average BERT-base Energy (mJ)	77	228	147

Voltage-Frequency Scaling

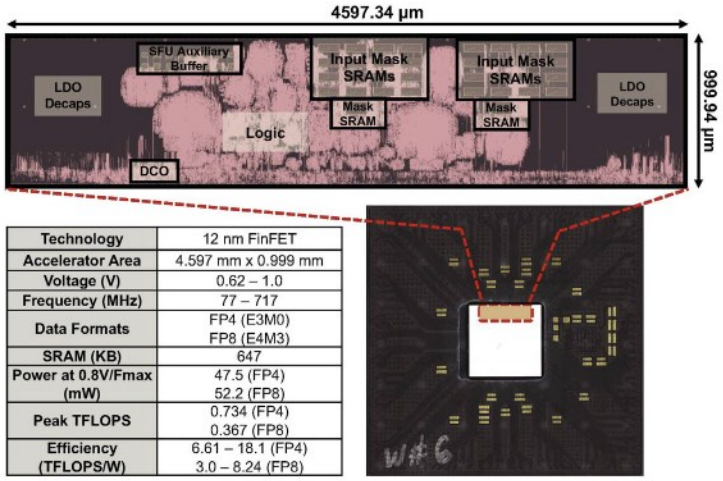


Comparison with Prior Art

	JSSC'20 [3]	JSSC'22 [4]	VLSI'22 [5]	ISSCC'22 [6]	This Work
Technology	65 nm	16 nm	5 nm	28 nm	12 nm
Area (mm²)	7.74	8.84	0.153	6.82	4.60
Model Architecture	LSTM, RNN	Linear, RNN, Attention	Transformers	Transformers	Transformers
Sentence-Level Adaptive Optimization	–	–	–	–	EE, MP, VFS
Data Types	INT8 (Weight) INT13 (Activation)	FP8	INT4, INT8	INT12	FP4, FP8
Total SRAM (KB)	297	5000	141	336	647
Supply Voltage (V)	0.68 – 1.1	0.55 V – 1.0	0.46 – 1.05	0.56 – 1.1	0.62 – 1.0
Frequency (MHz)	8 – 80	130 – 573	152 – 1760	50 – 510	77 – 717
Power (mW)	1.85 – 67.3	34 – 453	–	12 – 273	9 – 111 (FP4) 10 – 122 (FP8)
Peak Throughput (TOPS)	0.16	1.17	3.6 (INT4)	0.522	0.734 (FP4) 0.367 (FP8)
Peak Energy Efficiency (TOPS/W)	8.93~	7.8‡	95.6* (INT4)	4.25~†	18.1* (FP4) 8.24* (FP8)

*: Measured while performing matrix multiplications which are the core computations in Transformers, and with 50% input sparsity.
†: Measured while performing matrix multiplications.
–: Input sparsity not reported.
‡: No sparsity support.

Figure 22.9.6: STP performance across datasets and during V/F scaling, along with comparison with recently published transformer accelerators.



Technology	12 nm FinFET
Accelerator Area	4.597 mm x 0.999 mm
Voltage (V)	0.62 – 1.0
Frequency (MHz)	77 – 717
Data Formats	FP4 (E3M0) FP8 (E4M3)
SRAM (KB)	647
Power at 0.8V/Fmax (mW)	47.5 (FP4) 52.2 (FP8)
Peak TFLOPS	0.734 (FP4) 0.367 (FP8)
Efficiency (TFLOPS/W)	6.61 – 18.1 (FP4) 3.0 – 8.24 (FP8)

Figure 22.9.7: Annotated die photo highlighting the physical layout of the STP, along with a specification summary.

978-1-6654-9016-0/23/$31.00 ©2023 IEEE