FLAT: An Optimized Dataflow for Mitigating Attention Bottlenecks

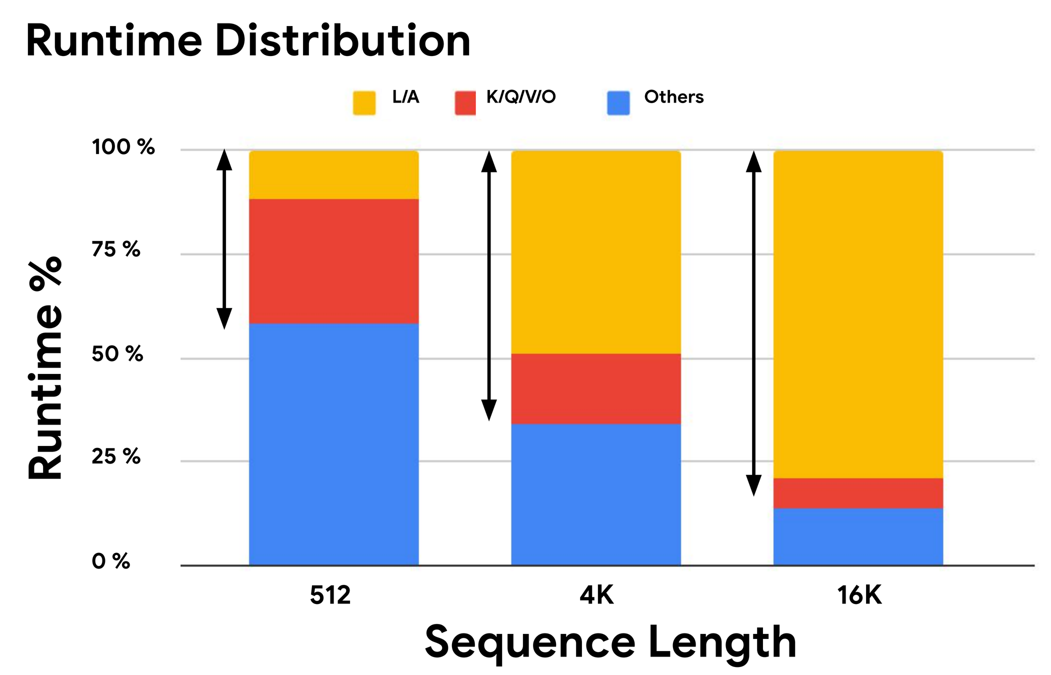
Felix Kao, Suvinay Subramanian, Gaurav Agrawal, Amir Yazdanbakhsh, Tushar Krishna

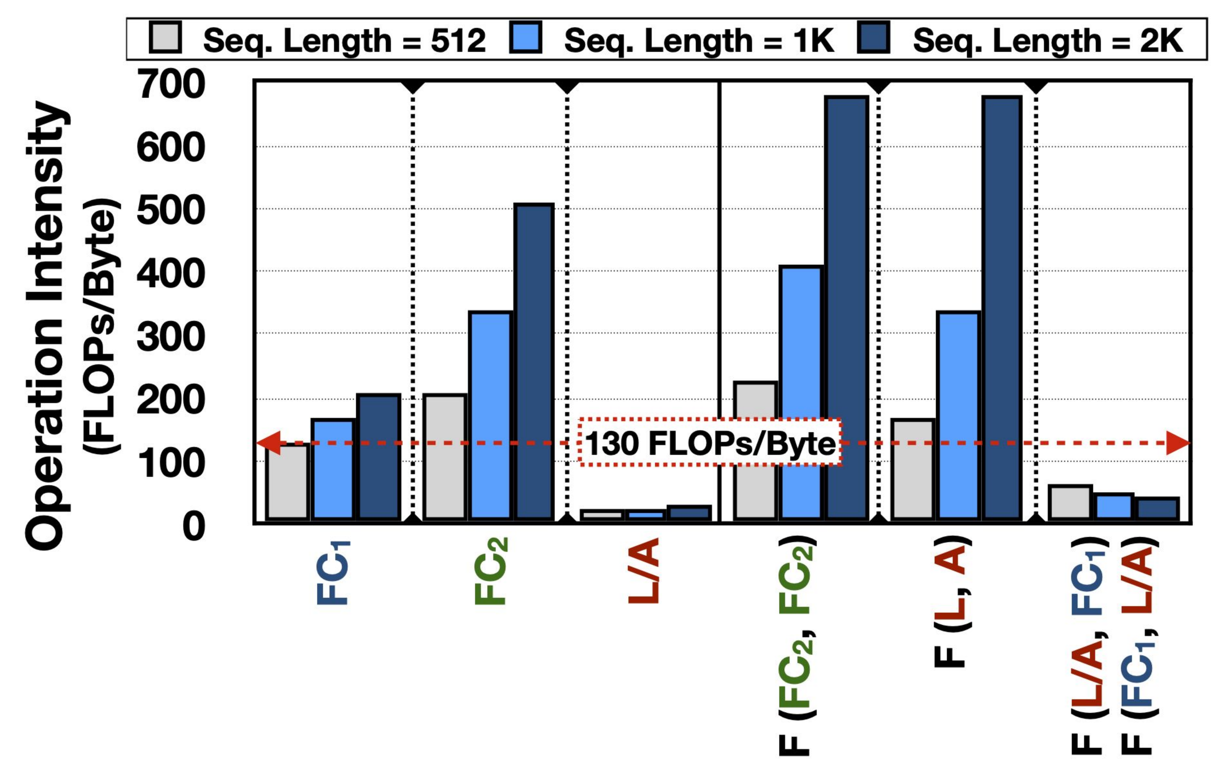
{felixkao, suvinay, chipguy, ayazdan}@google.com, tushar@ece.gatech.edu | ASPLOS 2023, Vancouver, Canada | Tuesday, March 28, Section 5C: Machine Learning

1. Introduction

Attention is a **key primitive** for Transformer architectures

- Forms the foundation of next-generation Deep Neural Network (DNN) models
- A growing fraction of run time with increasing sequence lengths (e.g., GPT-4: 32K)



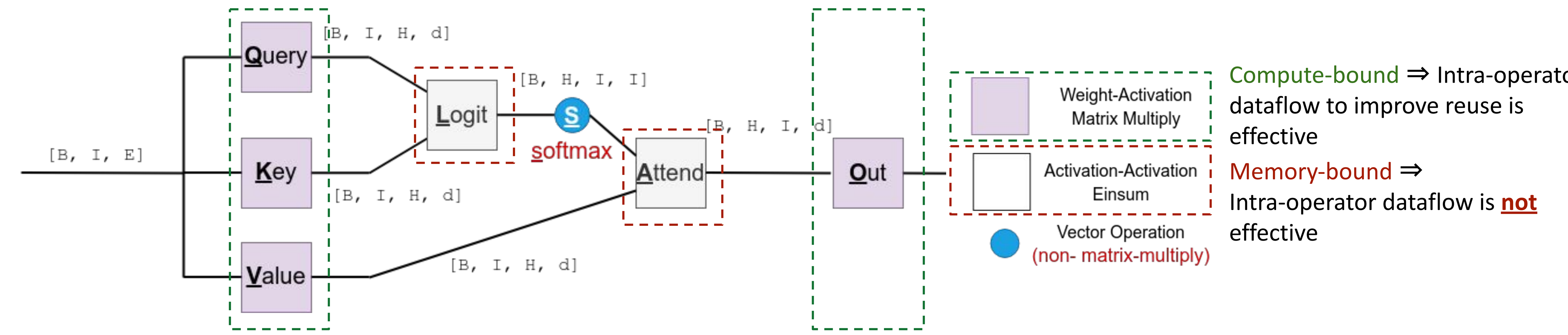


Attention operators exhibit **different properties** from prior DNN primitives

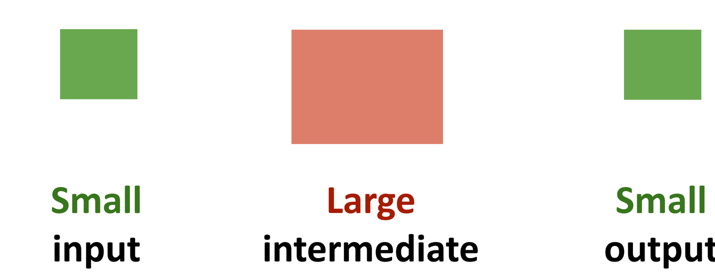
E.g., Convolutions (CNN), Embeddings (recommendation models), Fully-connected (FC)

1. Fundamentally **low operational-intensity** i.e., memory-bandwidth bound
   ⇒ Standard data-flow that exploit intra-operator reuse are ineffective
2. **Quadratic growth in memory** with sequence length
   ⇒ Places pressure on off-chip memory bandwidth and on-chip memory capacity

2. Fused Logit Attend Tiling (FLAT): Optimized Dataflow for Attention



Insight



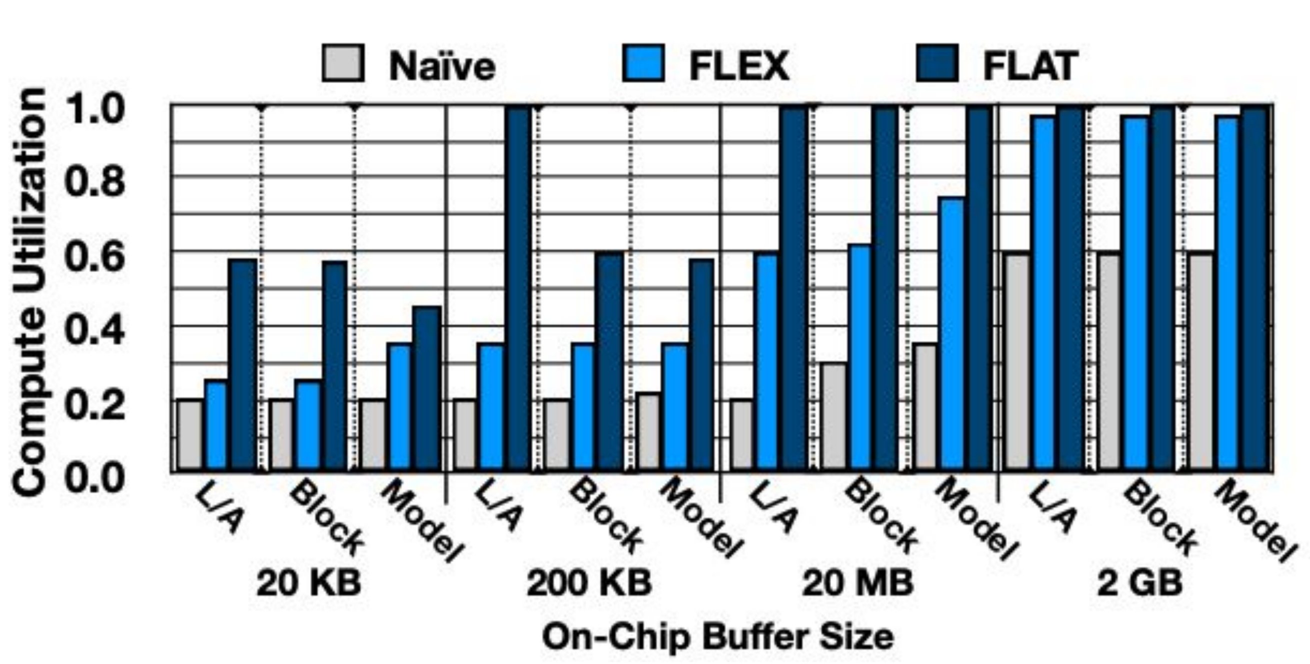
FLAT Employ cross-operator fusion ⇒ Fuse Logit and Attend operators

- Fused operator has **higher effective operational intensity** ⇒ not as memory-bound
- Ameliorates off-chip memory bandwidth and on-chip memory capacity demand

Details Unique considerations entails unique engineering solutions

- Intervening activation function (softmax) not element-wise
  Reduction requires specific slices of data ⇒ Imposes data dependencies
- FLAT develops an effective tiling and data movement strategy that respects data dependencies while enabling cross-operator fusion

3. Evaluation



FLEX requires 2x more on-chip buffer to match the performance of **FLAT**

Runtime (ms)	Sequence Length (Batch Size=1)						
	128	512	2K	4K	16K	64K	128K
Baseline	12	74	697	OOM	OOM	OOM	OOM
FLAT	11	43	175	424	4,599	64,350	OOM

Runtime (ms)	Batch Size (Sequence Length=256)						
	1	16	64	128	256	1K	2K
Baseline	36	630	2,520	5,230	OOM	OOM	OOM
FLAT	28	480	1,870	3,740	7,560	34,010	OOM

FLAT enables larger batch size and larger sequence length on real systems (GPU)

4. Impact and Implications

FLAT is simple, effective and impactful

- Enables improved performance on **GPUs and TPUs** (via XLA compiler) on deployed models
- Enables use-cases not previously possible: **long sequence, larger batch size**

Implications for accelerator co-design

- Demonstrates importance of cross-operator fusion for foundational Transformer models
  - De-facto for (current and) future accelerators (e.g., similar ideas in *FlashAttention*)
  - Critical input in design-space exploration: e.g., on-chip buffer size, off-chip memory bandwidth
- New *efficient* attention algorithms present new compute-memory tradeoffs
  - A new landscape of opportunities for dataflow and codesign!

FLAT on GPU

1.5x Speedup | **32x** Sequence Length | **8x** Batch Size