

Implementing and Optimizing the Scaled Dot-Product Attention on Streaming Dataflow

Gina Sohn
ginasohn@stanford.edu
Stanford University
Stanford, USA

Nathan Zhang
stanfurd@stanford.edu
Stanford University
Stanford, USA

Kunle Olukotun
kunle@stanford.edu
Stanford University
Stanford, USA

Abstract

Transformer models serve as the backbone of many state-of-the-art language models, and most use the scaled dot-product attention (SDPA) mechanism to capture relationships between tokens. However, the straightforward implementation of SDPA has quadratic compute and memory complexity with respect to the sequence length. On processor architectures such as GPUs and TPUs, there is a robust body of prior work. However, little work has been performed on non-processor architectures. In this work, we show how the architecture and execution model of Streaming Dataflow Accelerators can help tackle this challenge. We first define abstract hardware that adopts a streaming execution model, and we implement a cycle-accurate simulator of the abstract hardware using the Dataflow Abstract Machine simulation framework. Second, we implement the naive SDPA algorithm on this abstract hardware and show it requires linear ($O(N)$) intermediate memory. Third, we then modify the naive algorithm, taking inspiration from prior processor-oriented works, by reordering the multiplication and division operations. Finally, we map the modified algorithm to abstract hardware, and confirm that the implementation computes SDPA at full throughput while only using a constant amount ($O(1)$) of intermediate memory.

1 Introduction

Transformer models [19] are widely used for various language, audio, and vision tasks. Over time, researchers have discovered that long sequence lengths are essential to capturing long-range dependencies and process high-resolution images [2, 4, 14]. However, scaled dot-product attention (SDPA), the core operation in most transformer models, has quadratic memory complexity with respect to the sequence length. There has been a wide range of work to tackle this challenge, such as approximate attention [1, 9, 10], using alternative mixing mechanisms [7, 8], and optimizations to reduce reads and writes between different levels in the memory hierarchy [5, 6]. However, these works almost uniformly targeted processor-like architectures such as CPUs, TPUs, and GPUs.

In this work, we will instead implement and optimize the attention algorithms on streaming dataflow accelerators. Streaming dataflow accelerators are array architectures of reconfigurable compute and memory units. As shown in Figure 1, its execution model spatially maps the operations in

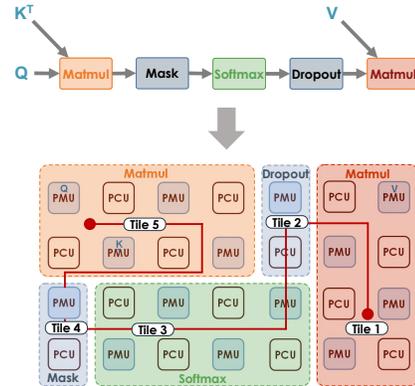


Figure 1. An abstract diagram of the architecture and execution model for streaming dataflow accelerators

the computation graph to hardware units by configuring the compute and memory units accordingly. It then pipelines the execution between different operations. Some examples of streaming dataflow accelerators include Coarse-grained Reconfigurable Architectures [3, 12] and Reconfigurable Dataflow Architectures [16, 18]. The streaming dataflow paradigm provides several performance benefits. First, it enables exploiting a high degree of parallelism due to dataflow execution and deeply pipelined execution. Second, it requires less memory footprint and bandwidth as operation fusion reduces off-chip memory accesses due to intermediate data.

To elucidate how the streaming execution model can reduce the memory complexity of the attention algorithm, we first define an abstract streaming dataflow accelerator based on Parallel Patterns [15]. We then map the attention algorithm to the abstract hardware and show that the implementation achieves linear memory complexity for intermediate data. Lastly, we apply algorithmic changes inspired by prior work on other architectures, and provide an implementation of SDPA that requires only a constant amount of intermediate memory.

2 The Abstract Streaming Dataflow Hardware

Prior work has shown that Parallel Patterns can simplify the generation process of optimized configurations for configurable hardware [11, 15, 16]. Our abstract hardware consists

Node	Behavior
Map (f: func)	Applies the function f to every element in the input stream
Reduce (n: Int) (init: T) (f: func)	Reduces across n element in the input stream using the given function. The output will be enqueued to the output stream after n elements are reduced. init is the initial value of the accumulator.
MemReduce (n: Int) (init: Mem[T]) (f: func)	Similar to Reduce, but executes higher order reduction on memory elements instead of scalar values.
Repeat (n: Int)	Repeats every scalar in the input stream n times.
Scan (n: Int) (init: T) (updt: func) (f: func)	On every new input element, update the state using the updt function. The state is initialized to the init value after scanning every n elements. On every input element, apply the function f and dequeue to the output stream.

Table 1. The nodes in our abstract hardware based on Parallel Patterns.

of nodes such as **Map**, **Reduce**, and **Scan** (Table 1). Each node can be further lowered to a configuration of the physical compute and memory units. To verify the functional correctness of the implementation and confirm it uses a linear amount of intermediate memory without any performance loss, we develop a cycle-accurate simulator for each node in the abstract hardware using the Dataflow Abstract Machine simulation framework [20].

3 Standard Attention Implementation

Inside the Transformer model, three matrices Q , K , and V are generated based on the input sequence and weights. Given that there are N tokens in the input sequence where each token is a d -wide vector, the shape of Q , K , and V is $N \times d$, and the attention algorithm can be expressed as follows:

$$S = QK^T, \quad P = \text{softmax}(S) = \frac{e^{s_{ij}}}{\sum_j e^{s_{ij}}}, \quad O = PV \quad (1)$$

The algorithm produces N^2 -sized intermediate matrices S and P , leading to a time and memory complexity quadratic to the sequence length. However, on a streaming execution model, it is possible to implement the same algorithm to have a sub-quadratic, in other words, asymptotically lower, memory complexity without any algorithmic changes. This starts from the observation that the attention algorithm can be decomposed into row-wise operations. Given that \vec{q}_i is the i -th row of the Q matrix, Equation 1 can be decomposed into row-wise operations as follows ($\vec{s}_i, \vec{p}_i, \vec{o}_i$ are the i -th row of matrix S, P, O respectively):

$$\vec{s}_i = \vec{q}_i K^T, \quad \vec{p}_i = \text{softmax}(\vec{s}_i) = \frac{e^{s_{ij}}}{\sum_j e^{s_{ij}}}, \quad \vec{o}_i = \vec{p}_i V \quad (2)$$

The rows of matrix Q can be streamed into compute units that execute the operations in Equation 2, and the computation between different rows can be pipelined. This has the effect of fusing operations to compute each row of the output

matrix O . Therefore, the memory complexity for intermediate data will reduce to the required intermediate data to compute a *single* row.

The mapping of the attention algorithm to our abstract hardware can be found in Figure 2. We configure each node as shown in the figure and set the depth of the short FIFOs to two and the long FIFO to $N + 2$. We compare this with the baseline, where all the FIFOs are set to have infinite depth (this will be the peak throughput scenario). We confirmed that the implementation in Figure 2 only requires $O(N)$ intermediate memory while running in full throughput. More detailed experiment results can be found in the case study in the Dataflow Abstract Machine work [20].

4 Memory-free Attention Implementation

By modifying the algorithm, we can further obtain an implementation that requires a constant amount ($= O(1)$) of intermediate memory. In the previous section, the reason why we need a $O(N)$ -long FIFO is that there are two paths from the second **Map** unit to the third **Map** unit with different latencies. The path that contains a row-wise reduction produces the output only after processing the last element in the row, and the following **Map** unit performs an element-wise operation on the pair of inputs from each path. The other path needs a deep FIFO that can contain the outputs from the second **Map** unit until the reduction path produces its output to avoid deadlock. Therefore, when the consumer performs an element-wise operation with the inputs from each path, the two divergent paths should have matched latency to avoid introducing this long FIFO.

This can be done by (1) reordering the division operation in softmax with the following matrix multiplication based on the distributive law, (2) using the running sum instead of the row-wise sum during the softmax operation, and (3) rescaling by the difference between the old running sum and latest running sum during the scalar reduction in softmax and the vector reduction in the matrix multiplication. For this implementation, we use softmax with scaling [6, 10, 13, 17],

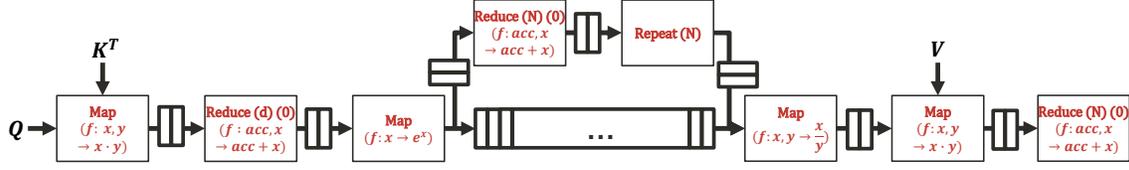


Figure 2. Implementation of attention using Parallel Patterns. The depth of the short FIFOs is set to 2, and the depth of the Long FIFO is set to $N + 2$. Each node can be mapped to a configuration of a set of compute and memory units in a streaming dataflow hardware.

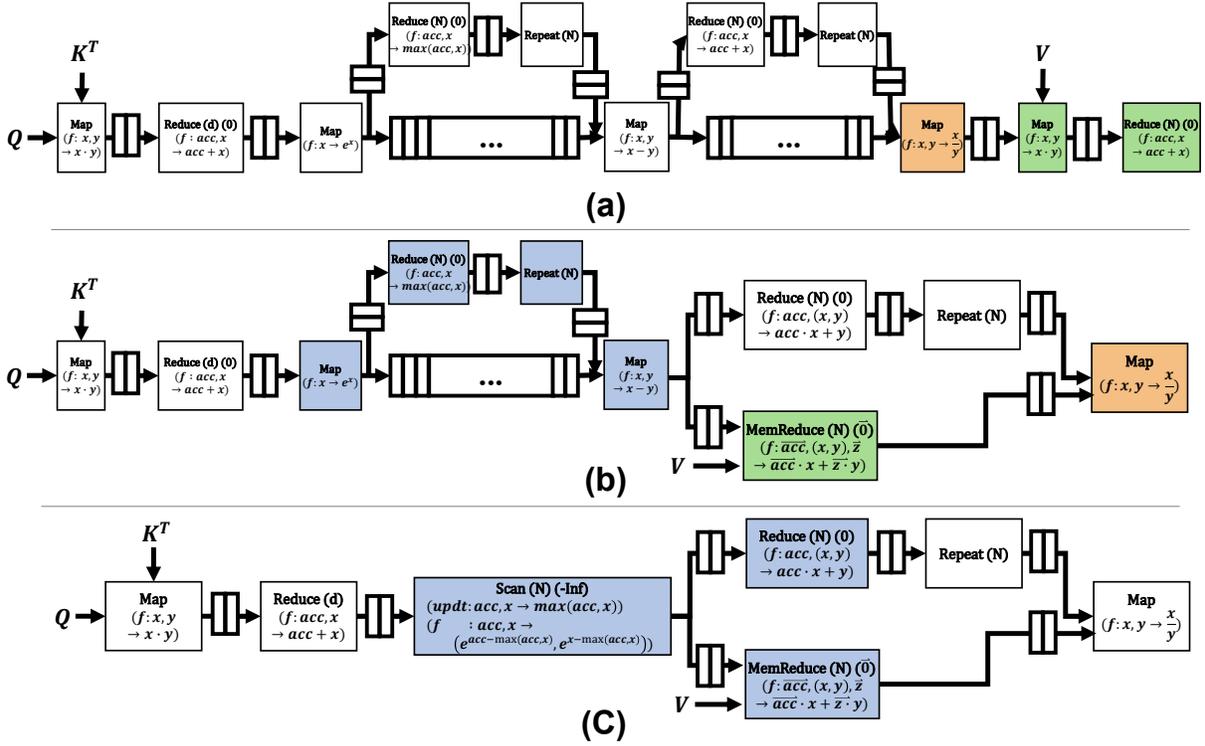


Figure 3. Implementations of algorithms using Parallel Patterns. (a) The attention algorithm using softmax with scaling. (b) The attention algorithm with reordered division. (c) The memory-free-attention algorithm.

which is widely used in practice for better numerical stability. The modified algorithm can be described as:

$$s_{ij} = \sum_k q_{ik} \cdot k_{kj} \quad (3)$$

Given $\forall j < 0$ $m_{ij} = -\infty$, for all $i > 0$:

$$\begin{aligned} m_{ij} &= \max(m_{i(j-1)}, s_{ij}) \\ \Delta_{ij} &= e^{m_{i(j-1)} - m_{ij}} \\ e_{ij} &= e^{s_{ij} - m_{ij}} \end{aligned} \quad (4)$$

Given $\forall j < 0$ $r_{ij} = 0$ and $\vec{l}_{ij} = \vec{0}$, for all $i > 0$:

$$\begin{aligned} r_{ij} &= r_{i(j-1)} \cdot \Delta_{ij} + e_{ij} \\ \vec{l}_{ij} &= \vec{l}_{i(j-1)} \cdot \Delta_{ij} + e_{ij} \cdot \vec{v}_j \end{aligned} \quad (5)$$

$$\vec{o}_i = \frac{\vec{v}_{iN}}{r_{iN}} \quad (6)$$

In the attention algorithm using softmax with scaling (Figure 3(a)), there are two pairs of divergent paths with unbalanced latency due to the row-wise reduction. The second

pair of unbalanced paths are handled by reordering the division operation in softmax and the matrix multiplication. As shown in Figure 3(b), by making the matrix multiplication’s reduction happen in parallel with the row-wise summation, it balances the latency of both paths and eliminates the $O(N)$ -long FIFO.

The first pair of unbalanced paths remaining in Figure 3(b) can be eliminated using a running max instead of the row-wise max. As the following operations are reductions, the difference between the latest running max and the running max at the point when each element was accumulated can be adjusted by rescaling the accumulated values. This, in turn, eliminates the deep FIFO by converting the reduction operation into an element-wise scan operation.

The implementation in Figure 3(c) is simulated by configuring each node in our simulator accordingly and setting the depth of the short FIFOs to two. This is compared with the baseline, where all the FIFOs are set to have infinite depth (this will be the peak throughput scenario). We confirmed that this implementation only requires $O(1)$ intermediate memory to run in full throughput. More detailed experiment results can be found in the case study in the Dataflow Abstract Machine paper [20].

References

- [1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [3] Alex Carsello, Kathleen Feng, Taeyoung Kong, Kalhan Koul, Qiaoyi Liu, Jackson Melchert, Gedeon Nyengele, Maxwell Strange, Keyi Zhang, Ankita Nayak, et al. Amber: A 367 gops, 538 gops/w 16nm soc with a coarse-grained reconfigurable array for flexible acceleration of dense linear algebra. In *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, pages 70–71. IEEE, 2022.
- [4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.
- [5] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- [6] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *Advances in Neural Information Processing Systems 35*, 2022.
- [7] Dan Fu, Simran Arora, Jessica Grogan, Isys Johnson, Evan Sabri Eyuboglu, Armin Thomas, Benjamin Spector, Michael Poli, Atri Rudra, and Christopher Ré. Monarch mixer: A simple sub-quadratic gemm-based architecture. *Advances in Neural Information Processing Systems*, 36, 2024.
- [8] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.
- [9] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.
- [10] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *International Conference on Learning Representations*, 2020.
- [11] David Koeplinger, Matthew Feldman, Raghu Prabhakar, Yaqi Zhang, Stefan Hadjis, Ruben Fiszal, Tian Zhao, Luigi Nardi, Ardavan Pedram, Christos Kozyrakis, and Kunle Olukotun. Spatial: A language and compiler for application accelerators. In *Proceedings of the 39th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2018.
- [12] Kalhan Koul, Jackson Melchert, Kavya Sreedhar, Leonard Truong, Gedeon Nyengele, Keyi Zhang, Qiaoyi Liu, Jeff Setter, Po-Han Chen, Yuchen Mei, Maxwell Strange, Ross Daly, Caleb Donovan, Alex Carsello, Taeyoung Kong, Kathleen Feng, Dillon Huff, Ankita Nayak, Rajsekhar Setaluri, James Thomas, Nikhil Bhagdikar, David Durst, Zachary Myers, Nestan Tsiskaridze, Stephen Richardson, Rick Bahr, Kayvon Fatahalian, Pat Hanrahan, Clark Barrett, Mark Horowitz, Christopher Torgn, Fredrik Kjolstad, and Priyanka Raina. Aha: An agile approach to the design of coarse-grained reconfigurable accelerators and compilers. *ACM Trans. Embed. Comput. Syst.*, 22(2), jan 2023.
- [13] Maxim Milakov and Natalia Gimelshein. Online normalizer calculation for softmax, 2018.
- [14] OpenAI. Gpt-4 technical report, 2023.
- [15] Raghu Prabhakar, David Koeplinger, Kevin J. Brown, HyoukJoong Lee, Christopher De Sa, Christos Kozyrakis, and Kunle Olukotun. Generating configurable hardware from parallel patterns. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS ’16*, page 651–665, New York, NY, USA, 2016. Association for Computing Machinery.
- [16] Raghu Prabhakar, Yaqi Zhang, David Koeplinger, Matt Feldman, Tian Zhao, Stefan Hadjis, Ardavan Pedram, Christos Kozyrakis, and Kunle Olukotun. Plasticine: A reconfigurable architecture for parallel patterns. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pages 389–402, 2017.
- [17] Markus N Rabe and Charles Staats. Self-attention does not need $o(n^2)$ memory. 2021.
- [18] Alexander Rucker, Shiv Sundram, Coleman Smith, Matthew Vilim, Raghu Prabhakar, Fredrik Kjolstad, and Kunle Olukotun. Revet: A language and compiler for dataflow threads. *arXiv preprint arXiv:2302.06124*, 2023.
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 2017.
- [20] Nathan Zhang, Rubens Lacouture, Gina Sohn, Qizheng Zhang, Fredrik Kjolstad, and Kunle Olukotun. The dataflow abstract machine simulator framework. In *2024 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA) (To Appear)*, 2024.