

LLM Inference

Efficient DL, Episode VIII '24



ChatGPT

ChatGPT 3.5 ▾



How can I help you today?

Design a database schema
for an online merch store

Help me study
vocabulary for a college entrance exam

Tell me a fun fact
about the Roman Empire

Plan an itinerary
for a fashion-focused exploration of Paris

Message ChatGPT...



ChatGPT can make mistakes. Consider checking important information.

YaGPT



Давай придумаем

YandexGPT 2

В этом режиме я помогаю придумывать — идеи, тексты на разные темы и многое другое.

Я пишу ответы с помощью YaGPT 2 — новой нейросети Яндекса, подражая текстам в интернете. Поэтому результат может быть выдумкой: это не моё мнение и не мнение Яндекса. Я стараюсь быть этичной, так что на некоторые запросы не отвечаю. Не судите строго за ошибки — я только учусь.

Начнём? Решите прерваться — скажите «Хватит».

Как выучить английский, если у тебя всего 30 мин в день

Придумай 7 идей для цифровых стартапов

Расскажи, как ложиться спать вовремя, если по вечерам хочется залипать в соц сети

Напиши мне



GigaChat

— Салют, вы в GigaChat!

И это не просто очередная текстовая модель 😊

GigaChat очень трудолюбив и может сказать и нарисовать почти всё, что вы скажете. Пользуйтесь с умом 🧠

Начиная работу с GigaChat, вы берете на себя ответственность за соблюдение законодательства РФ и общепризнанных правил этики в соответствии с [правилами использования сервиса](#). Ознакомьтесь с ними!



— Совет: Используйте кнопку «Новый чат», если хотите сменить тему разговора ...

— Если не знаете с чего начать, загляните в наш [гайд](#), там есть не только советы, но и примеры для вдохновения.



Как изменился бы состав воздуха, если бы люди дышали углекислым газом?

Напиши резюме для начинающего PR-специалиста

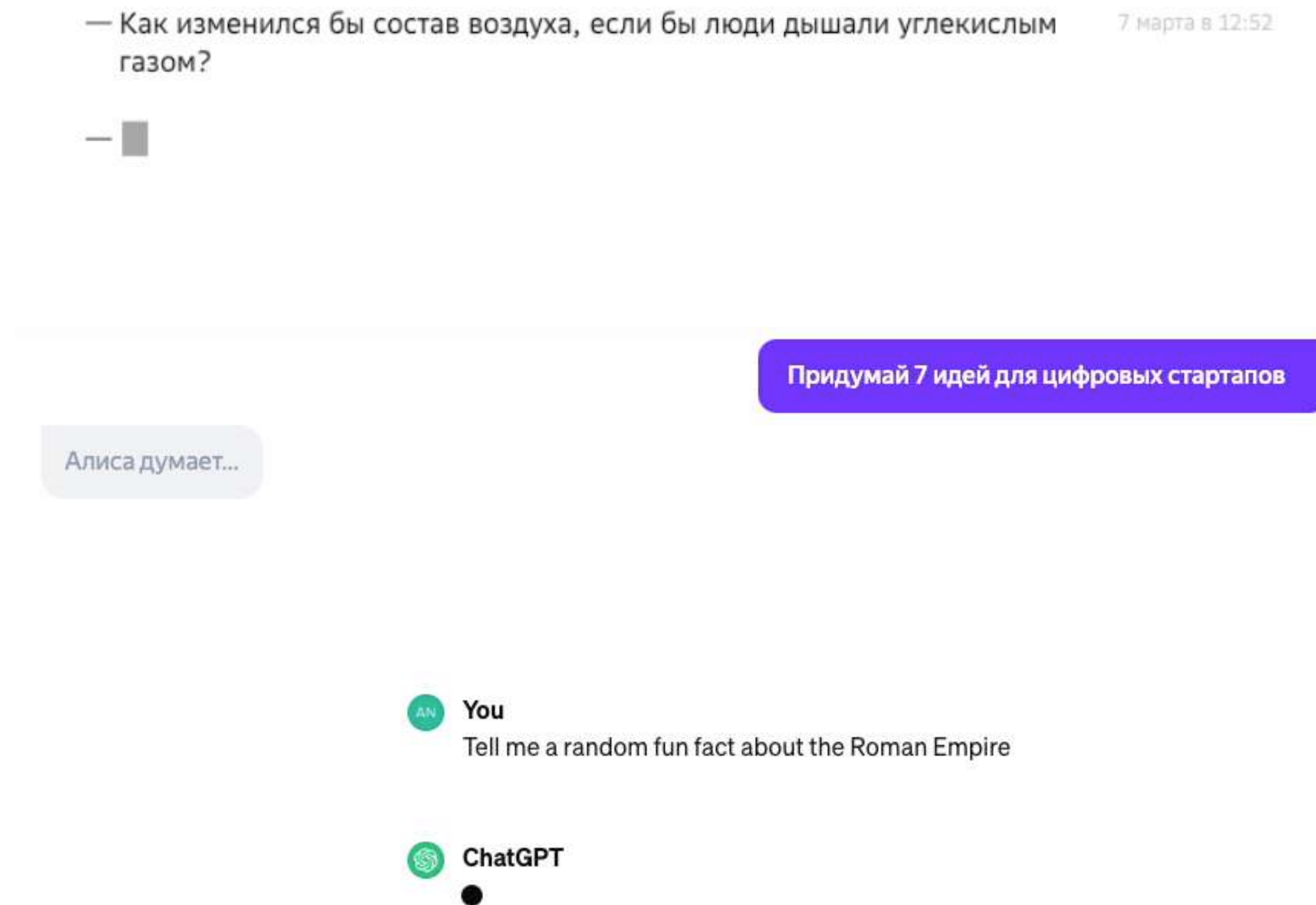


Спросите меня о чём-нибудь



TTFT

Time To First Token



Context decoding

TPS

Tokens Per Second



You

Tell me a random fun fact about the Roman Empire



ChatGPT

Sure! Did you know that the ancient Romans used urine as a cleaning agent? They believed that the ammonia in urine could help whiten and brighten their clothes. They even col ●

Придумай 7 идей для цифровых стартапов

1. Умный помощник для домашних животных: мобильное приложение и веб-платформа, которые позволяют владельцам домашних животных контролировать здоровье и поведение своих питомцев, обеспечивать

Алиса печатает...

— Как изменился бы состав воздуха, если бы люди дышали углекислым газом?

7 марта в 12:52

GigaChat:latest
По умолчанию

— Если бы люди дышали углекислым газом вместо кислорода, то состав воздуха на Земле изменился бы кардинально.

Во-первых, углекислый газ■

Generation step

TPS

Tokens Per Second

TPS for single query

TPS for single instance

Assistant

40 TPS is OK (or is it?)

Economics

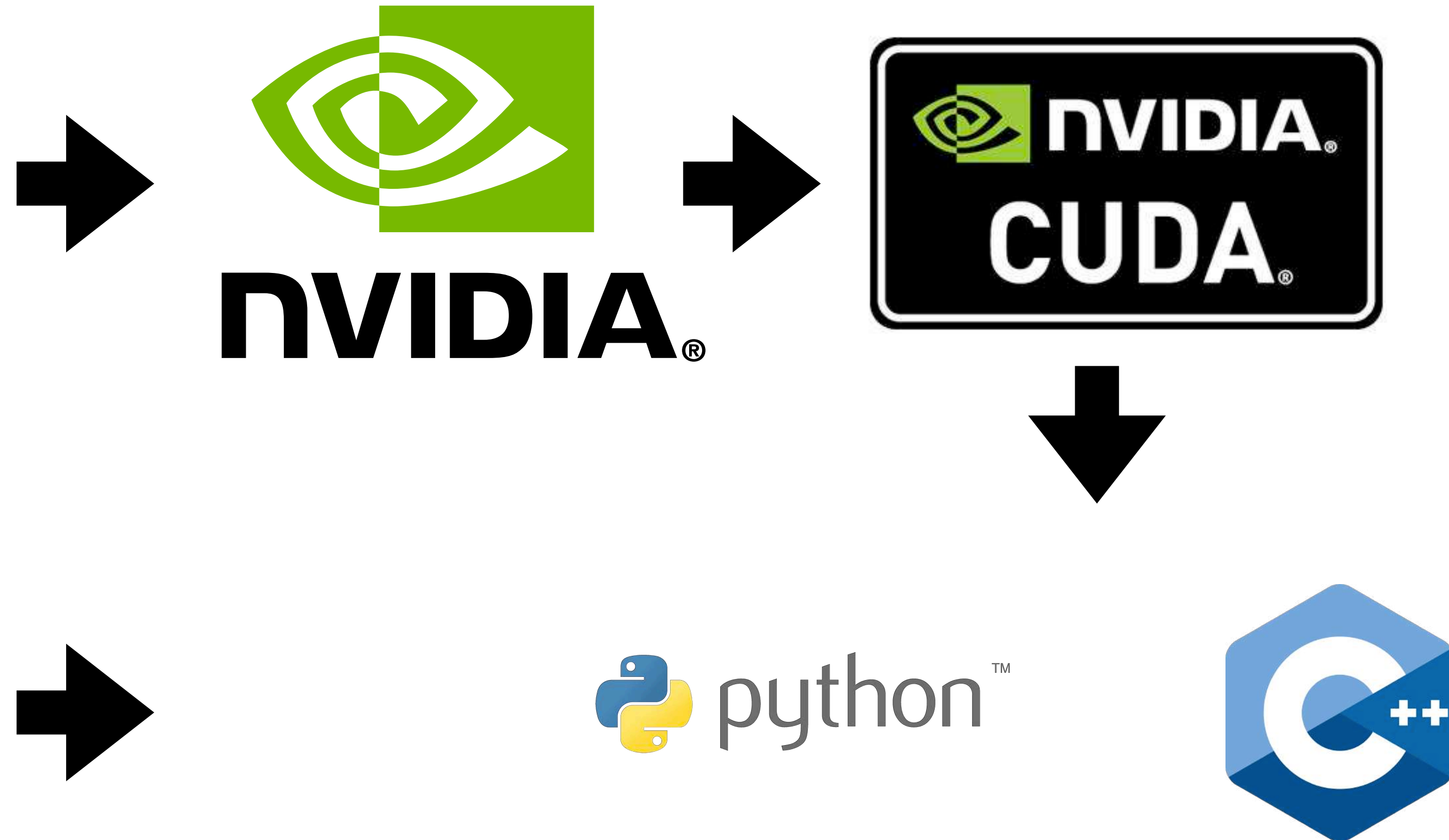
Offline

Do we care?

Most efficient way

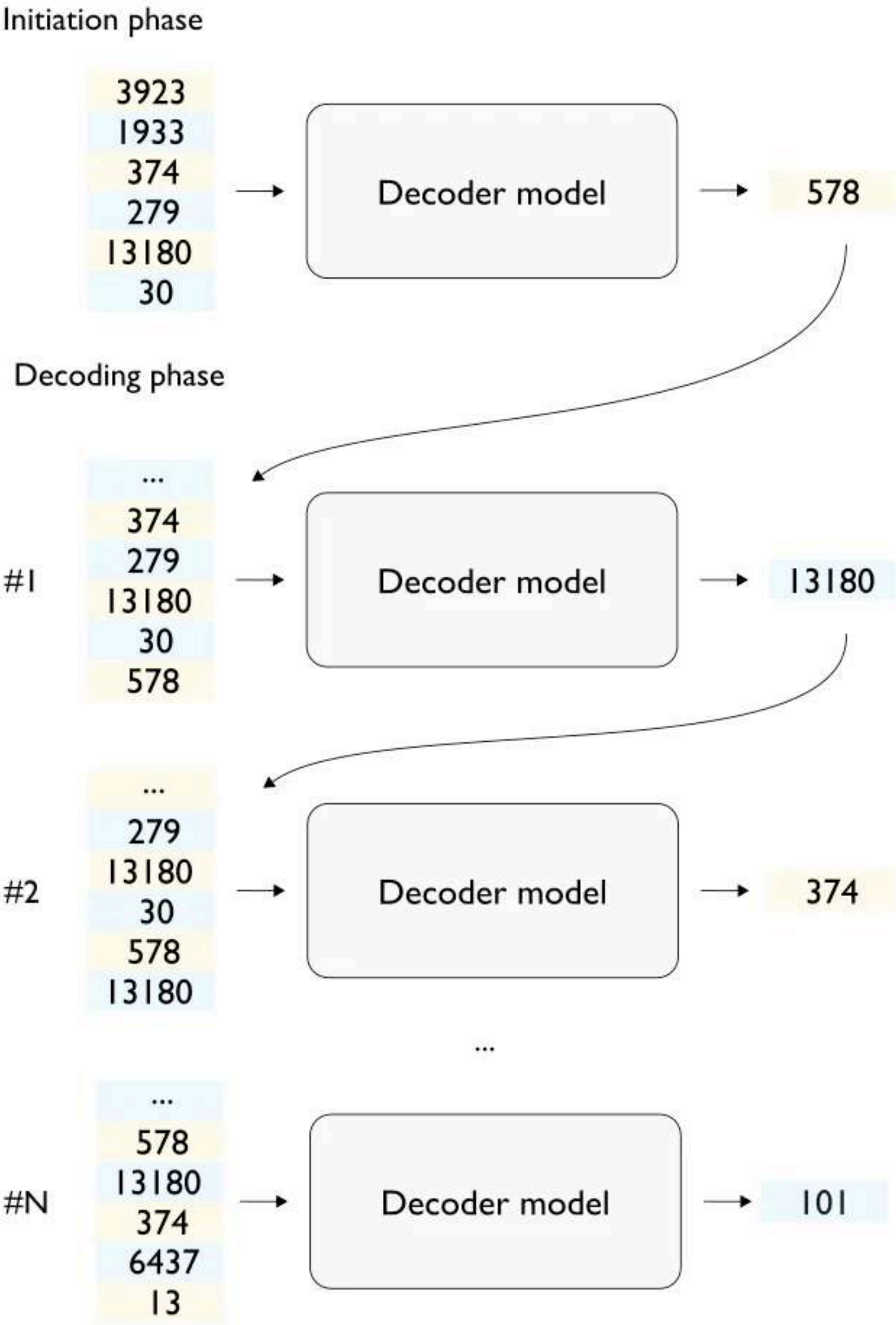
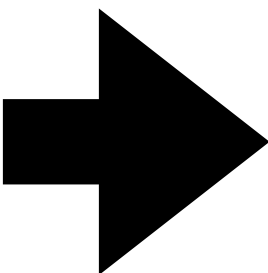
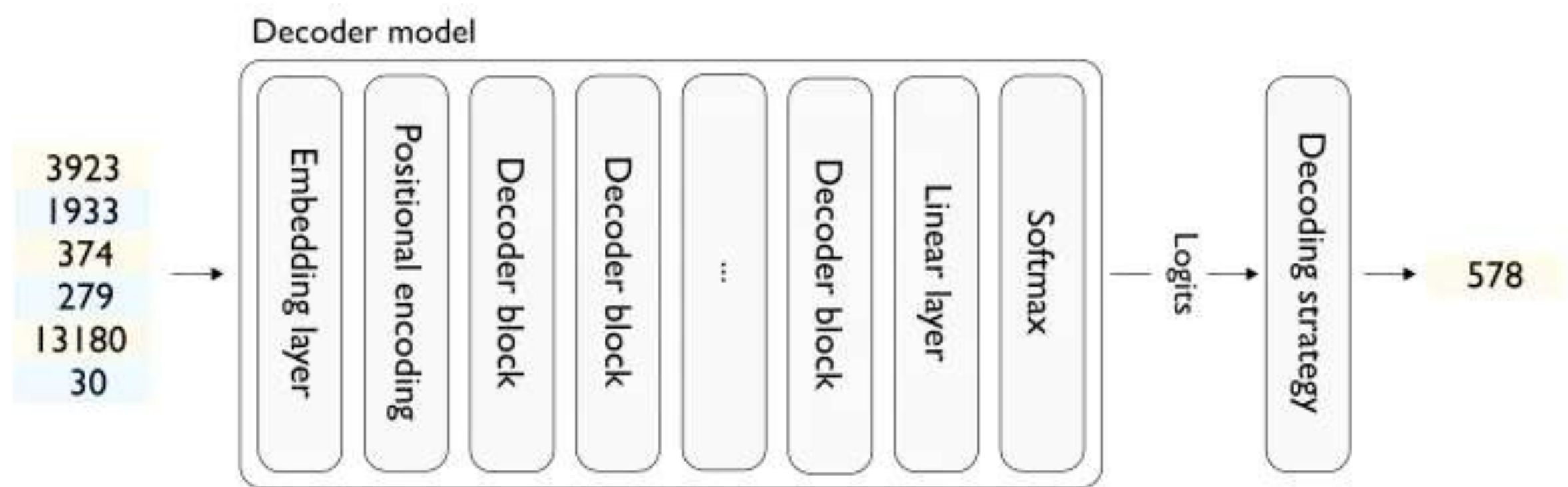
Inference

- Prompt processing
- Autoregressive steps
- Hardware utilization



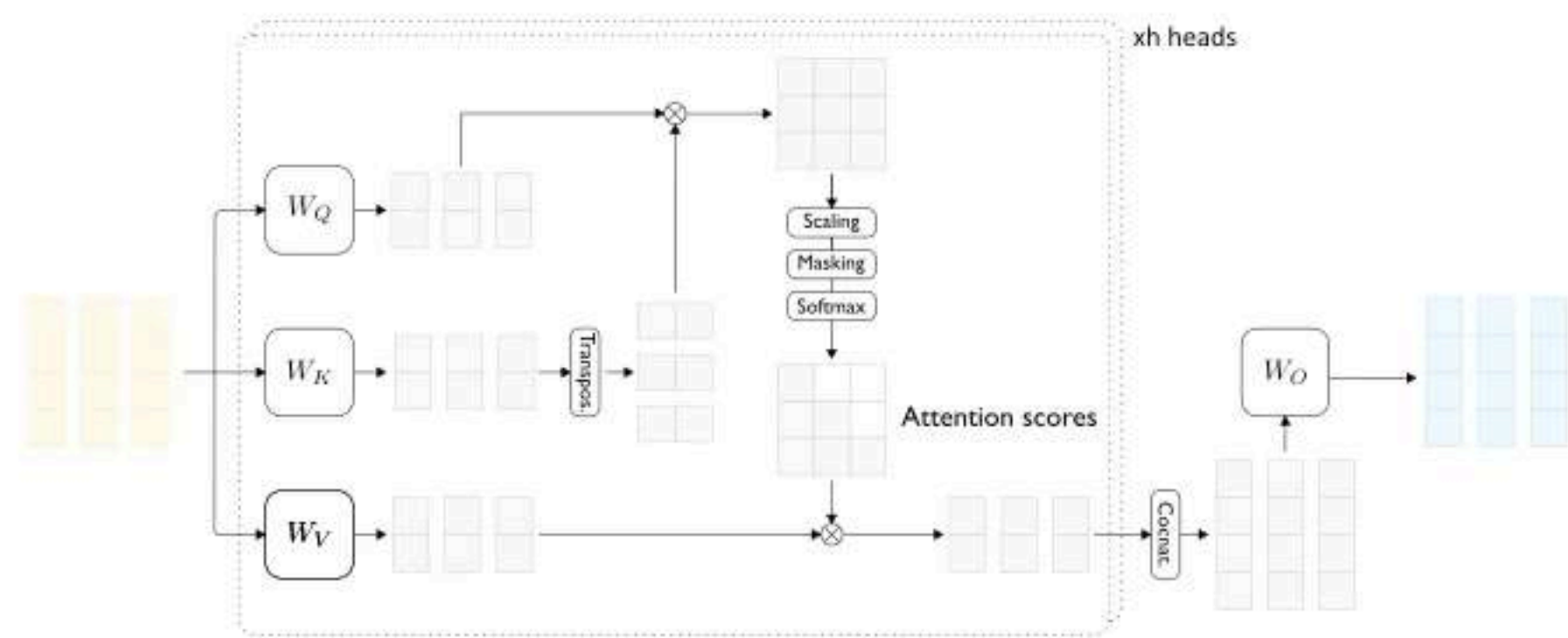
- Serving queries
- Business cases

LLM response

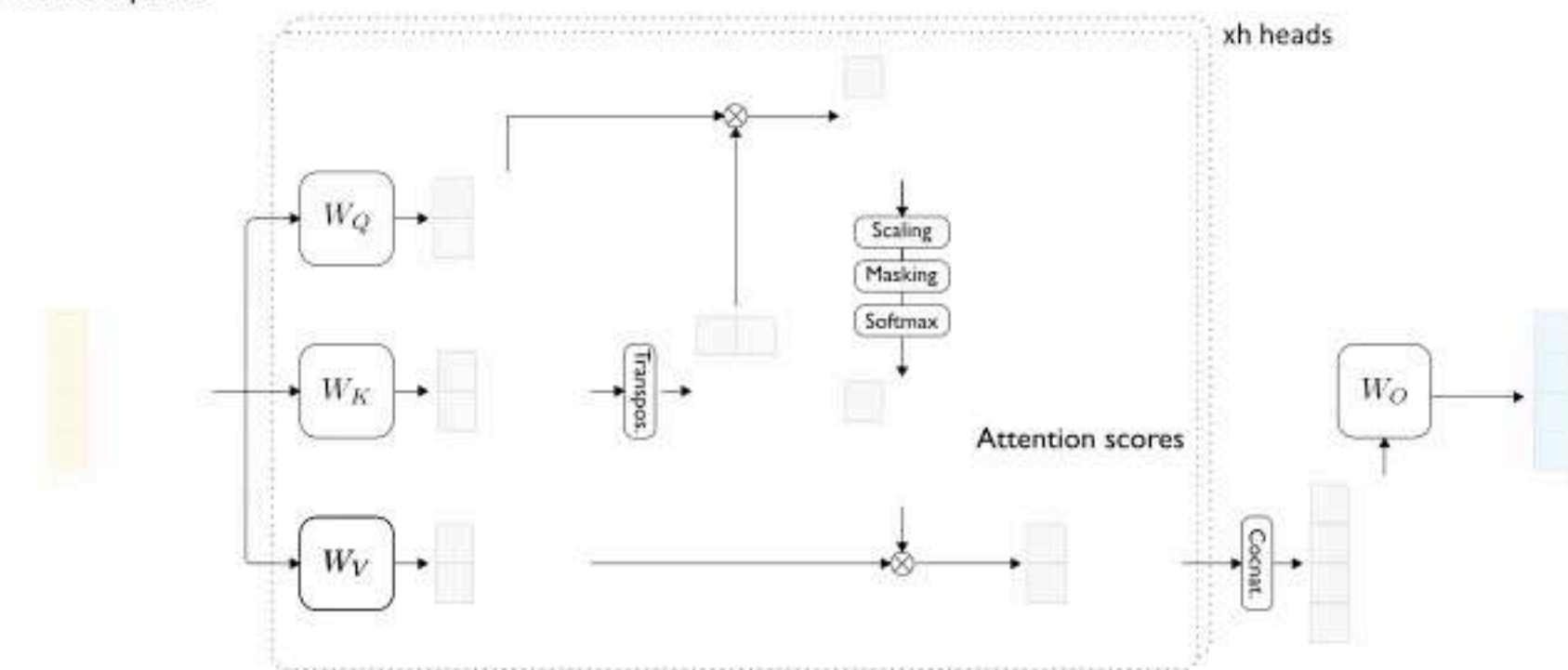


KV cache

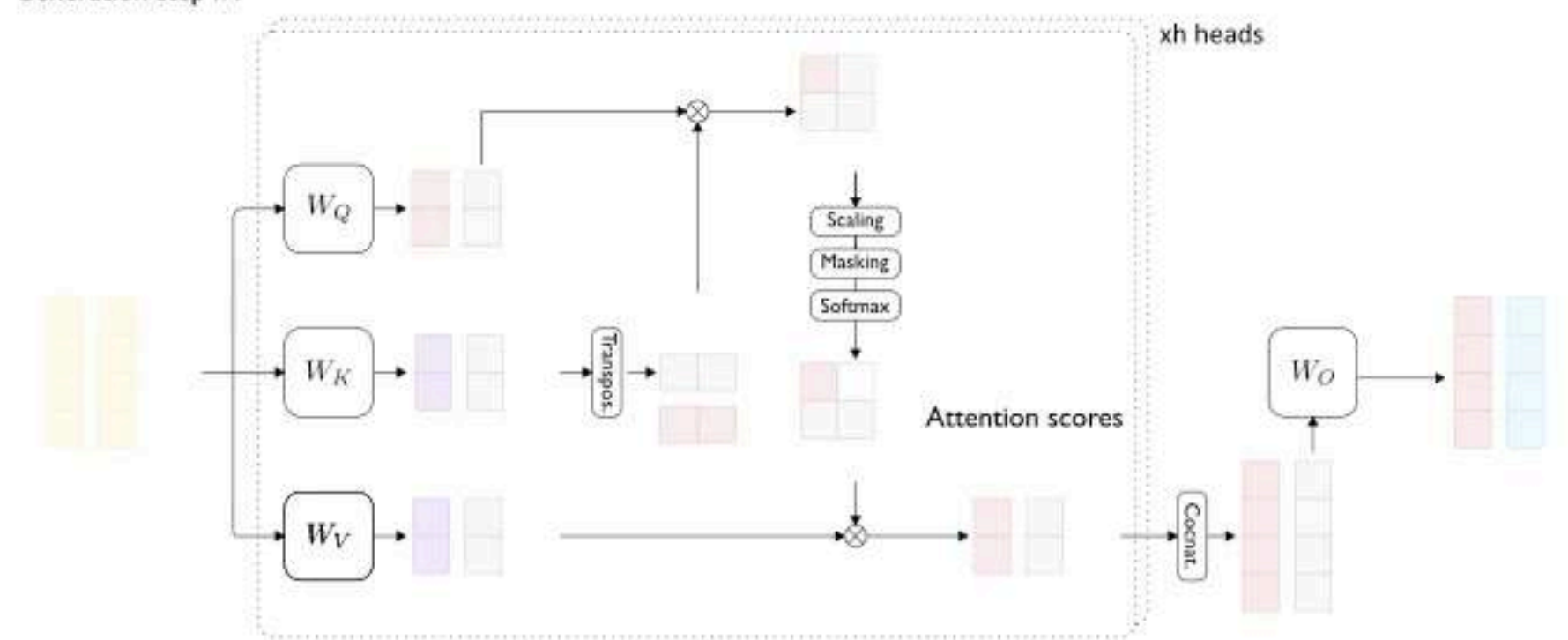
Decoder block



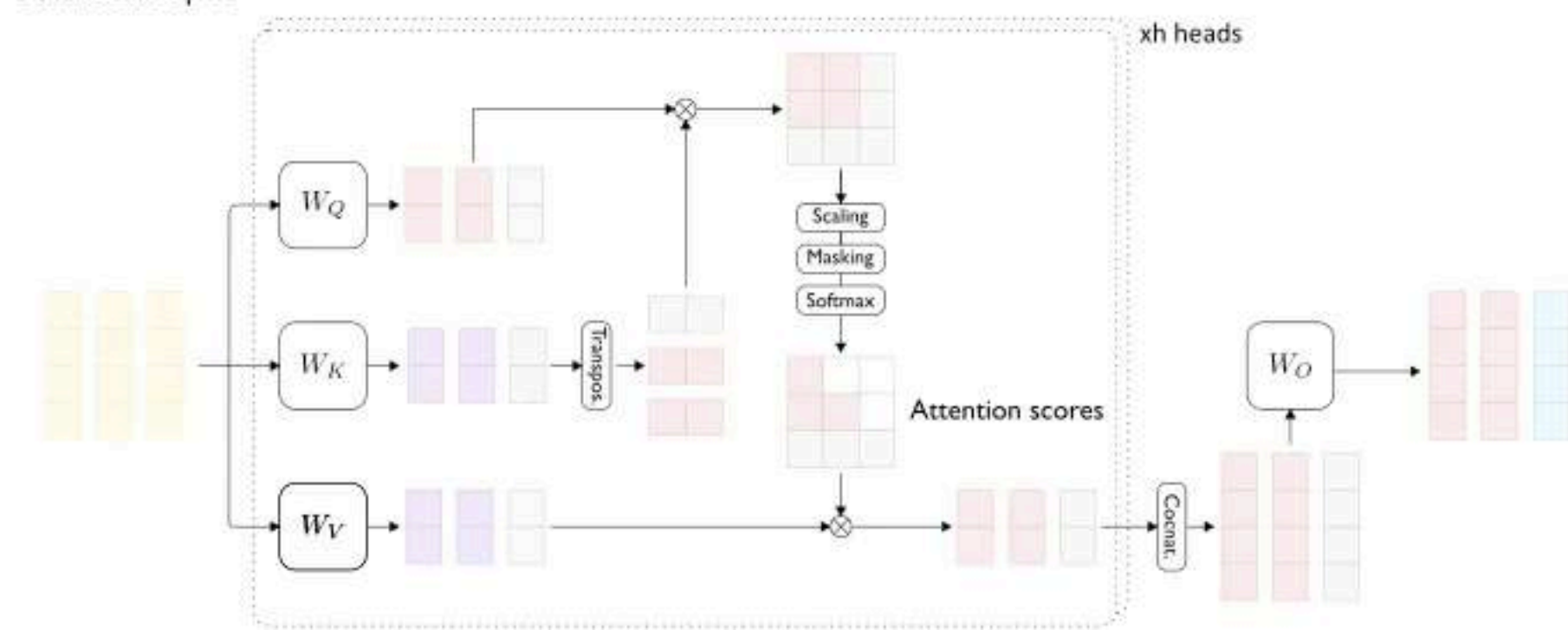
Initiation phase



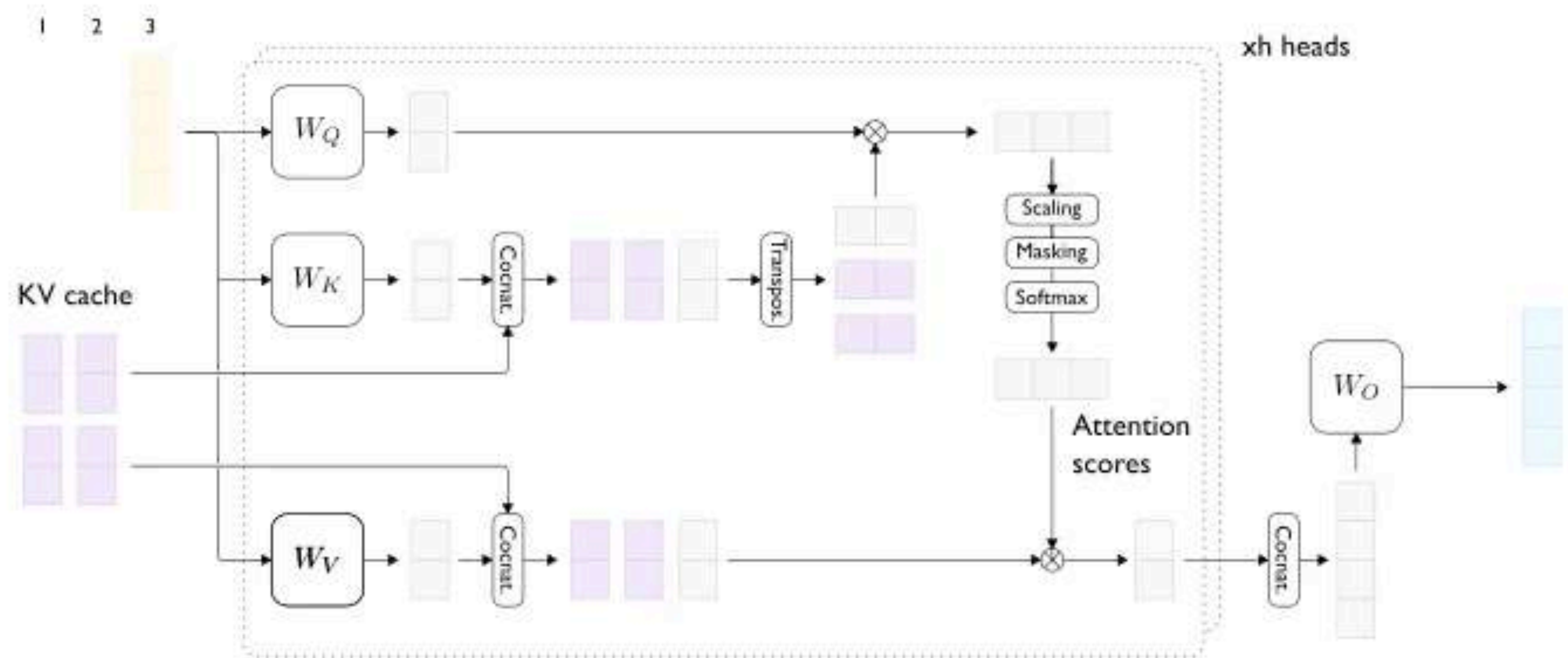
Generation step #1



Generation step #2



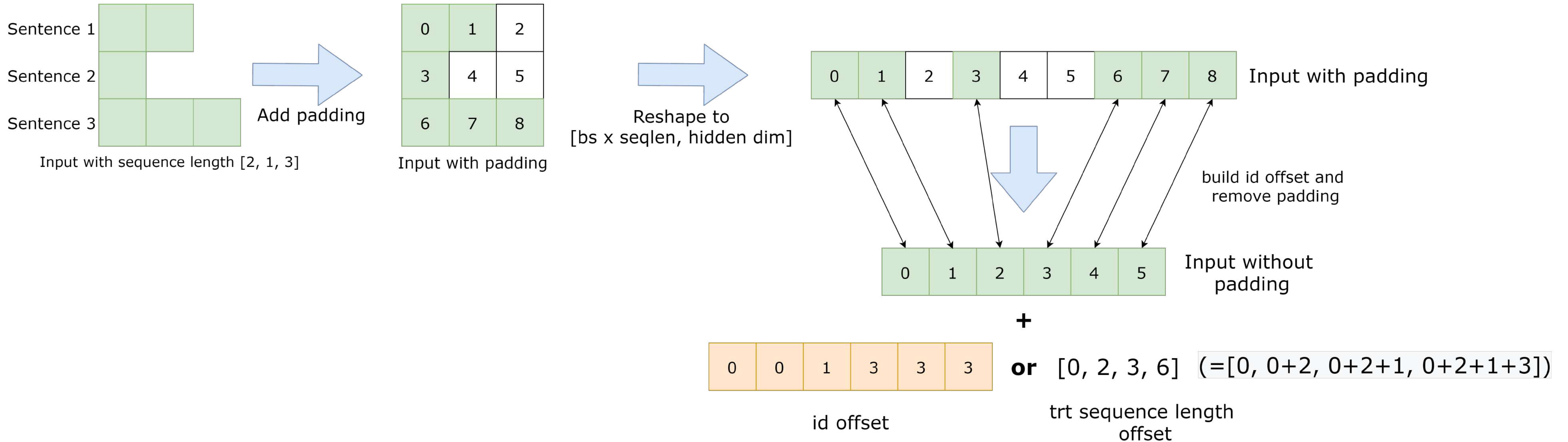
KV cache



What about batches?

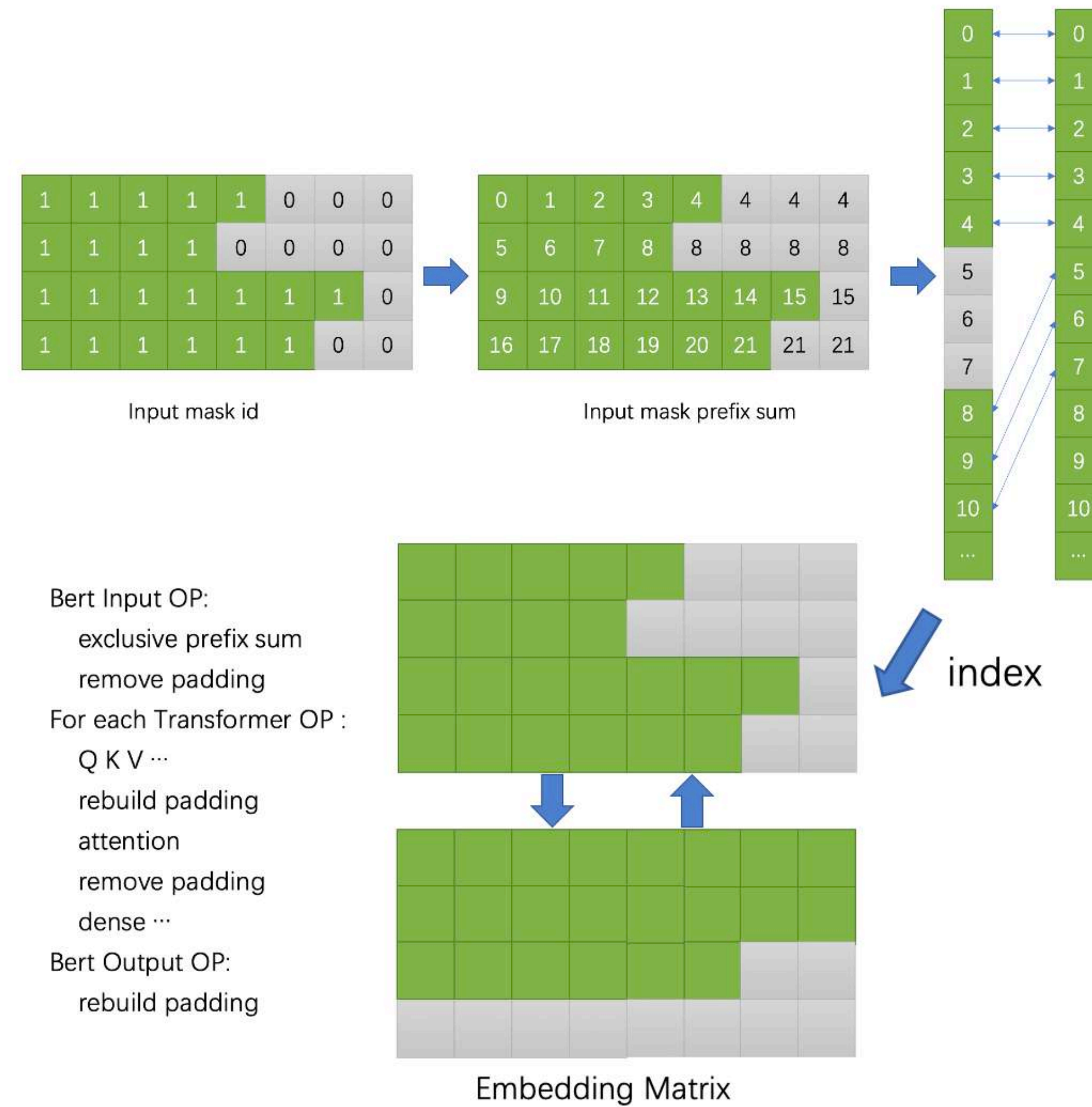


Batch



Source: https://github.com/bytedance/effective_transformer

Batch



Source: https://github.com/bytedance/effective_transformer

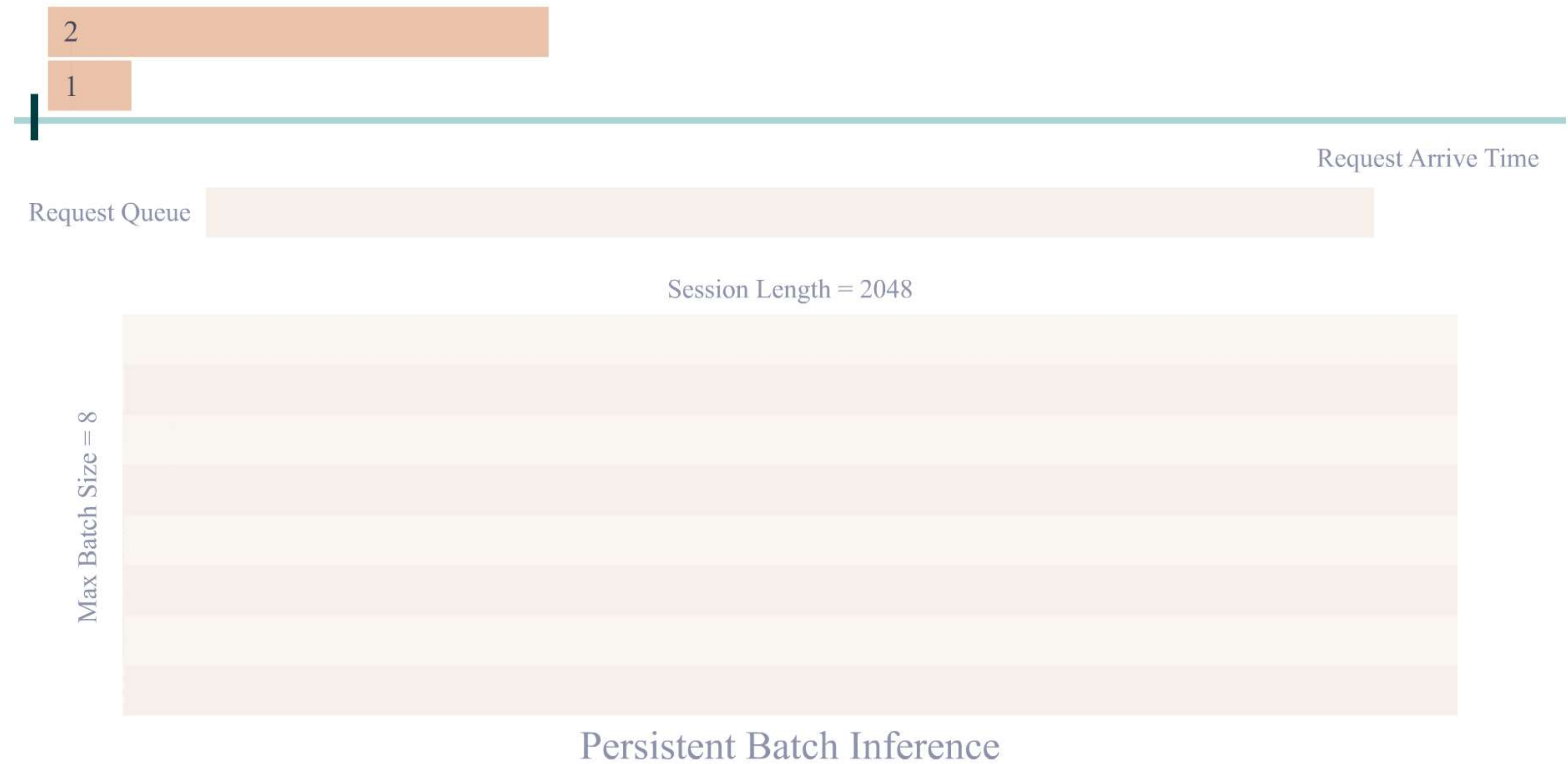
Batch

Tesla V100, float16, maximum sequence length=64, average sequence length \approx 40

batch_size	XLA (in ms)	Faster Transformer (in ms)	Speedup over XLA	Effective Transformer (in ms)	Speedup over XLA
100	28.31	20.27	1.40	16.03	1.77
200	54.47	40.08	1.36	30.15	1.81
300	80.53	59.11	1.36	41.27	1.95
400	106.5	78.38	1.36	54.12	1.97
500	132.35	98.03	1.37	65.92	2.01
1000	261.18	190.91	1.38	133.61	1.95

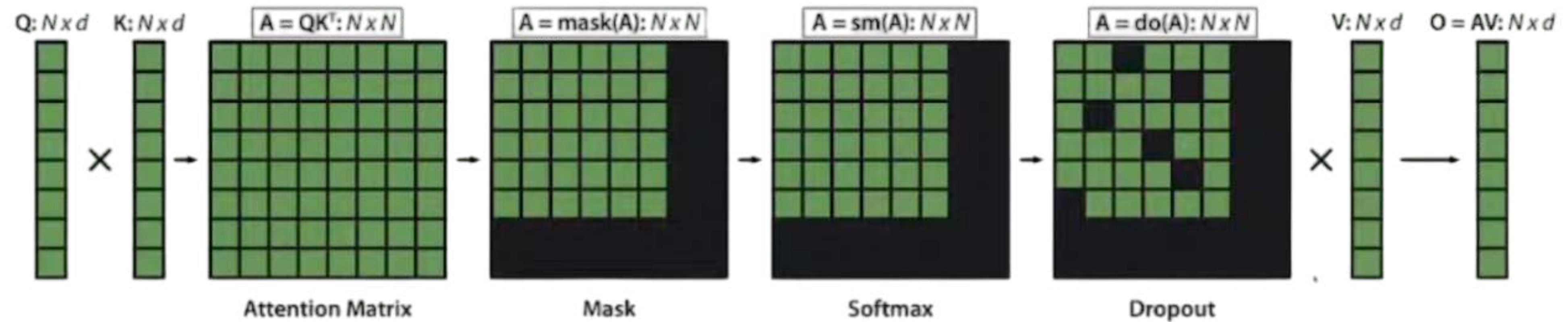
Source: https://github.com/bytedance/effective_transformer

Continuous batch



Source: <https://github.com/InternLM/Inmdeploy>

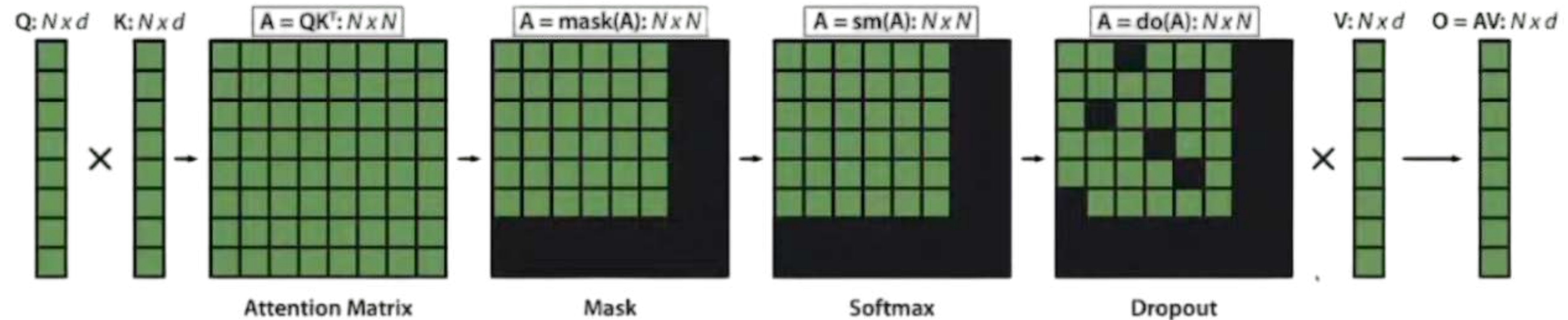
Flash attention



$$\mathbf{O} = \text{Dropout}(\text{Softmax}(\text{Mask}(\mathbf{QK}^T)))\mathbf{V}$$

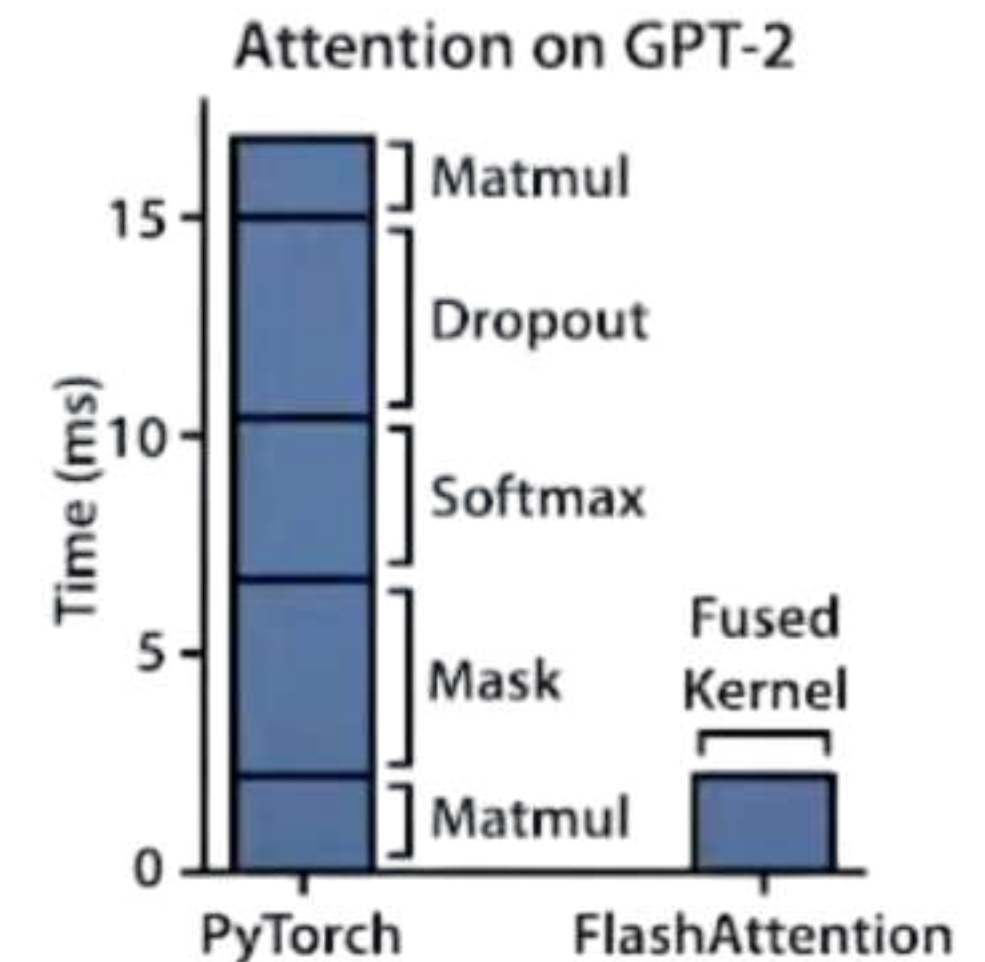
Naive implementation requires
repeated R/W from slow GPU HBM.
Hard to scale to long sequences

Flash attention

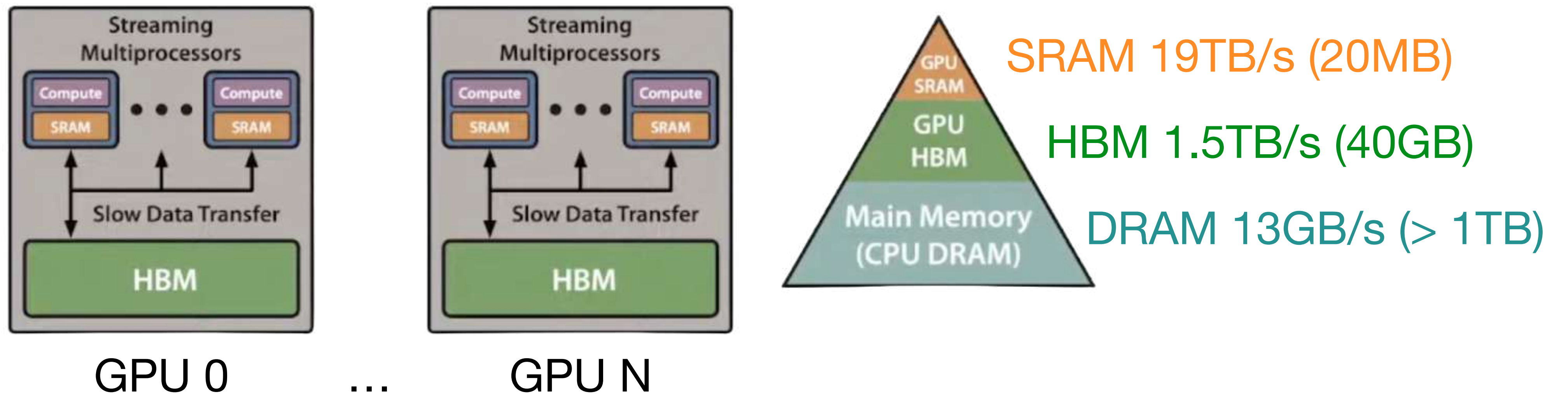


$$\mathbf{O} = \text{Dropout}(\text{Softmax}(\text{Mask}(\mathbf{QK}^T)))\mathbf{V}$$

Naive implementation requires
repeated R/W from slow GPU HBM.
Hard to scale to long sequences



Flash attention



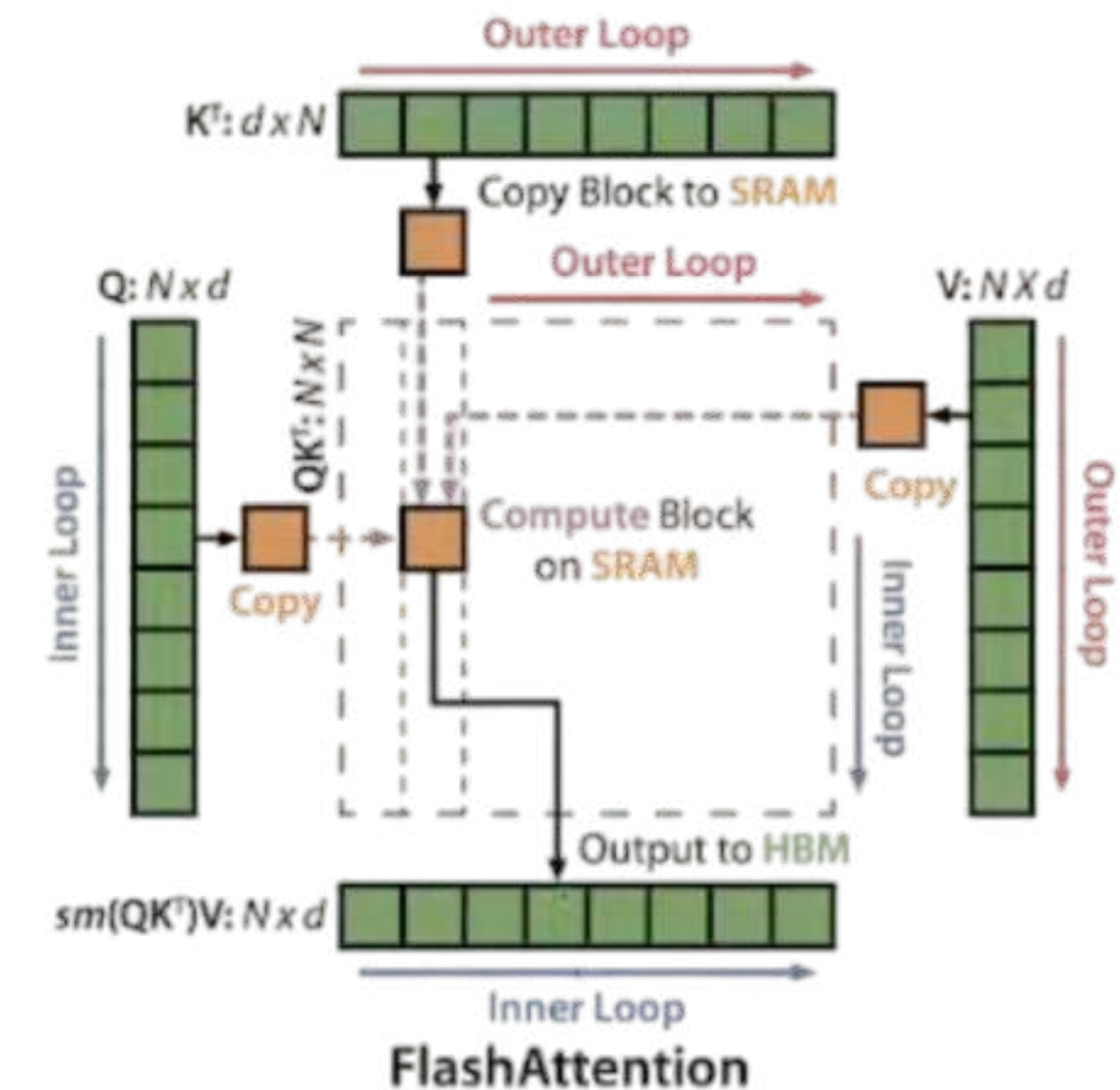
Flash attention

Decomposing large softmax into smaller ones by scaling.

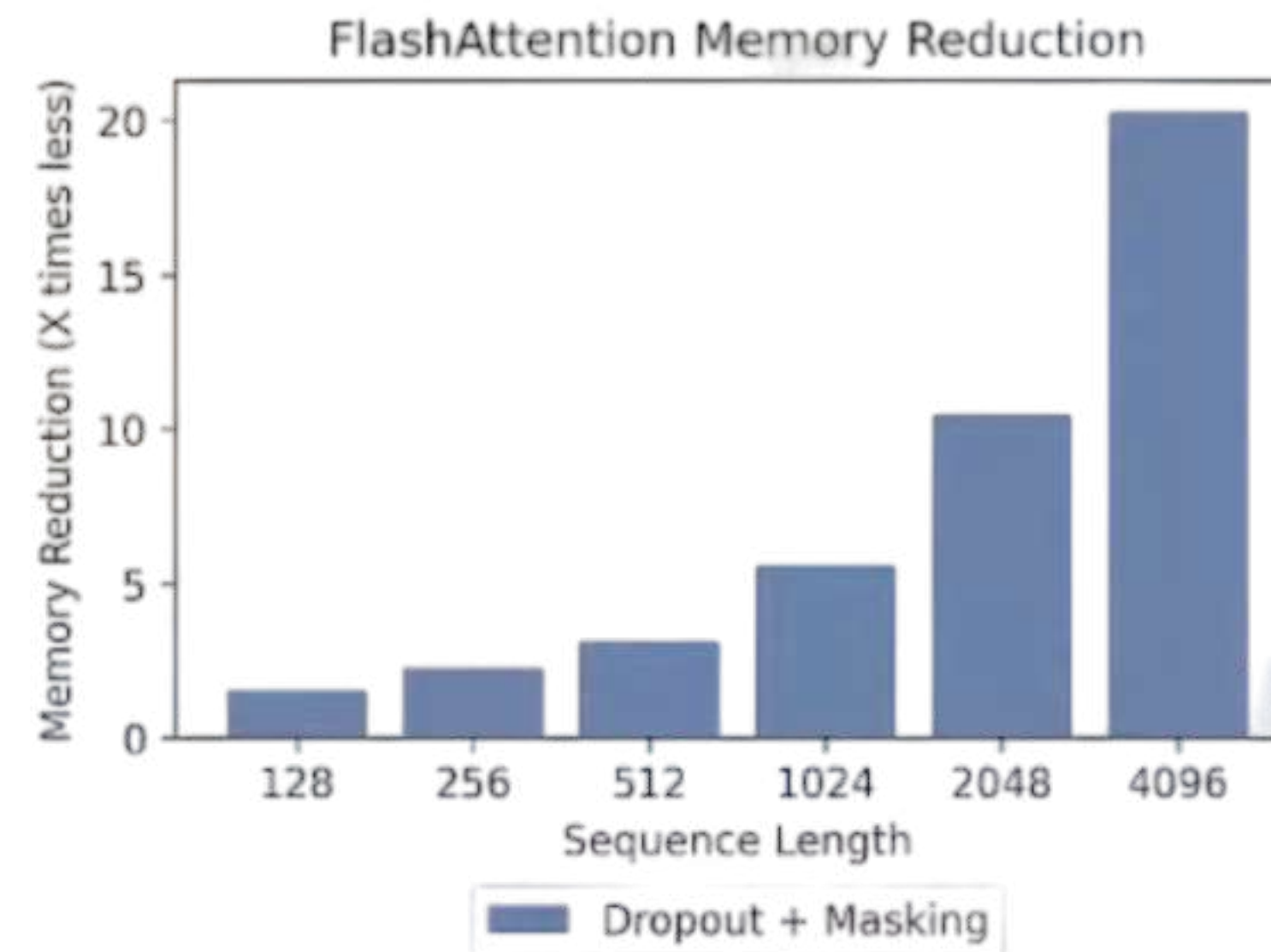
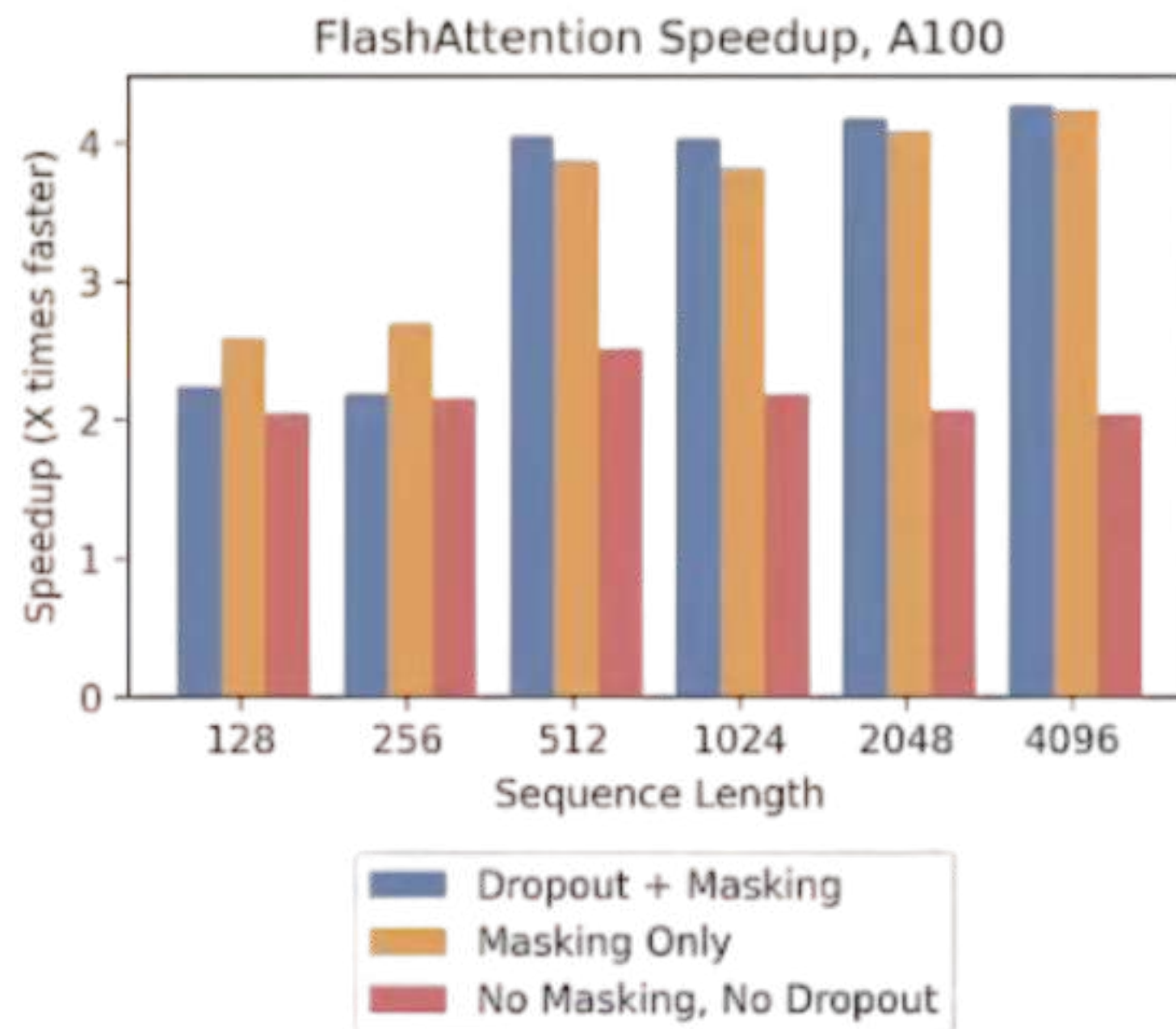
$$\text{softmax}([A_1, A_2]) = [\alpha \text{softmax}(A_1), \beta \text{softmax}(A_2)]$$

$$\text{softmax}([A_1, A_2]) \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \alpha \text{softmax}(A_1) V_1 + \beta \text{softmax}(A_2) V_2$$

1. Load inputs by blocks from HBM to SRAM.
2. On chip, compute attention output wrt that block.
3. Update output in HBM by scaling.



Flash attention



2-4x speedup — with no approximation!

10-20x memory reduction — memory linear in sequence length

Source: <https://arxiv.org/abs/2205.14135>

Flash attention 2

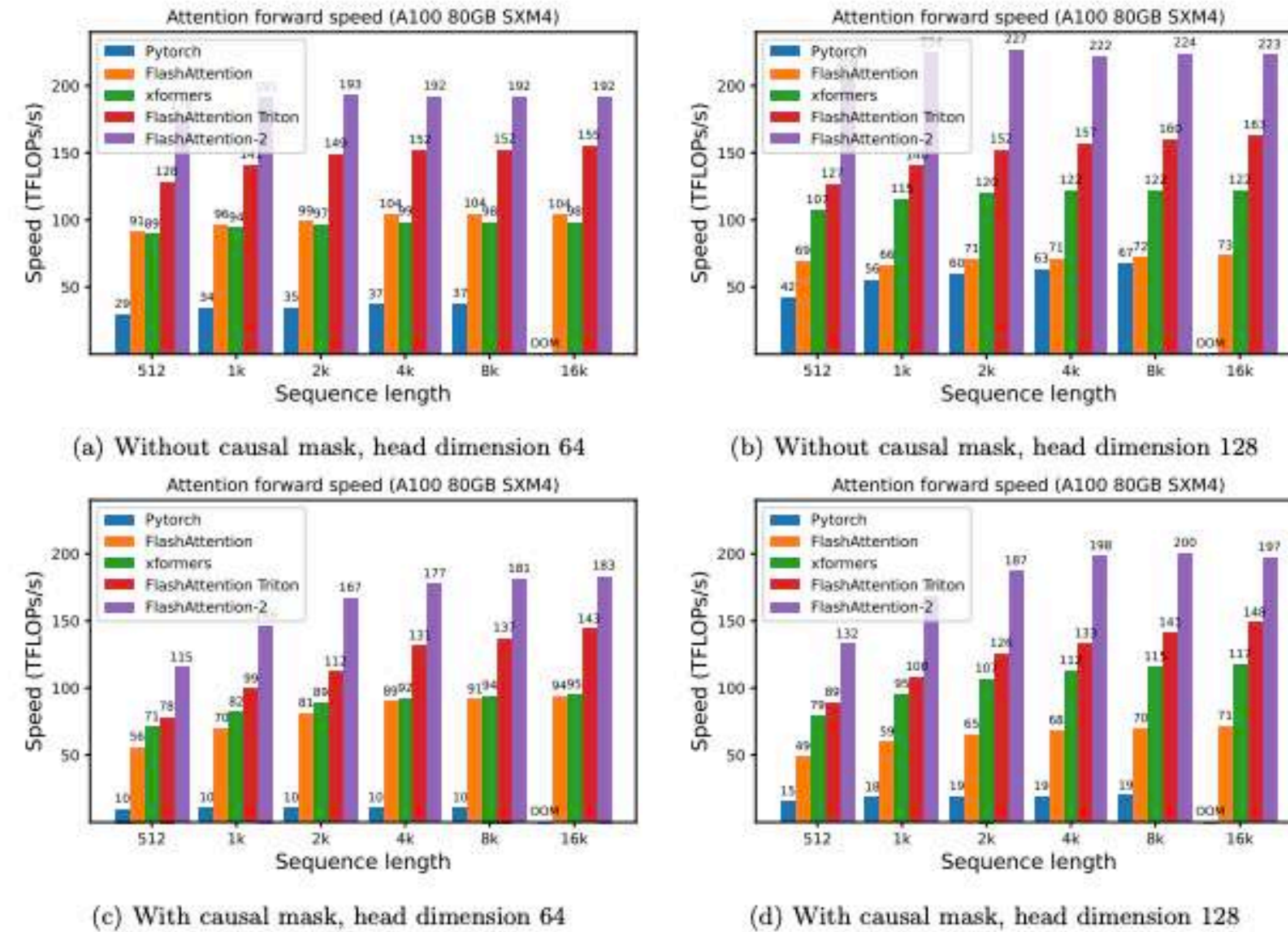
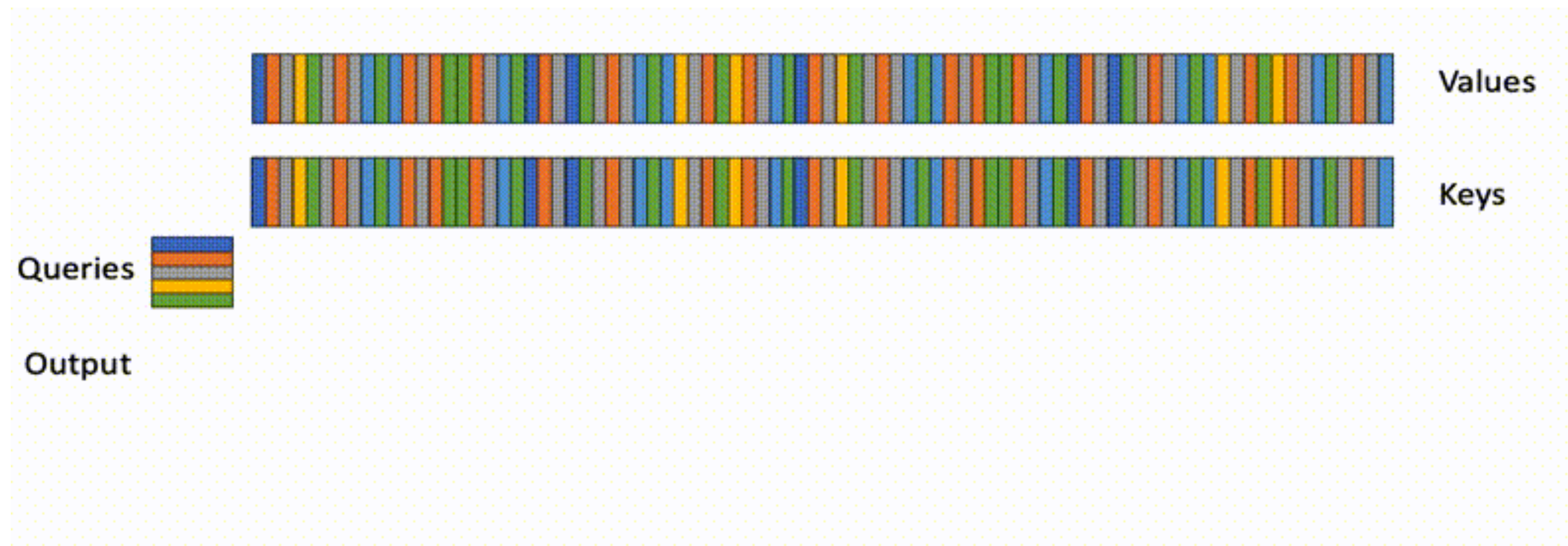


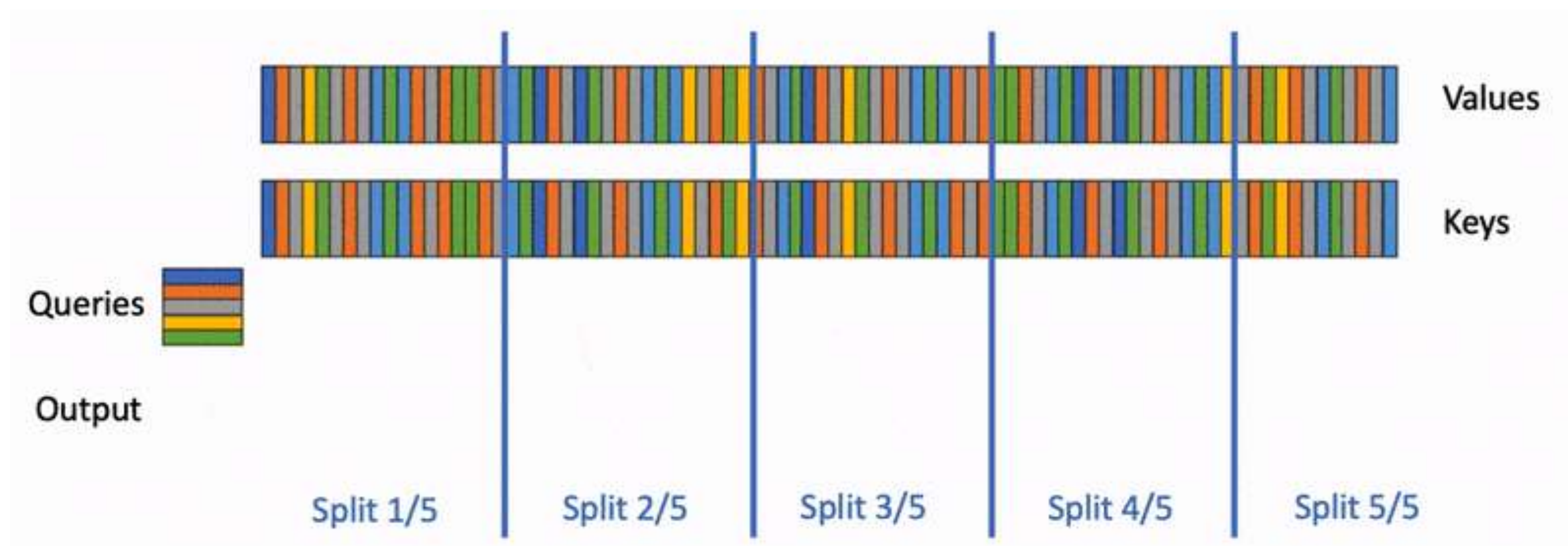
Figure 5: Attention forward speed on A100 GPU

Flash decoding



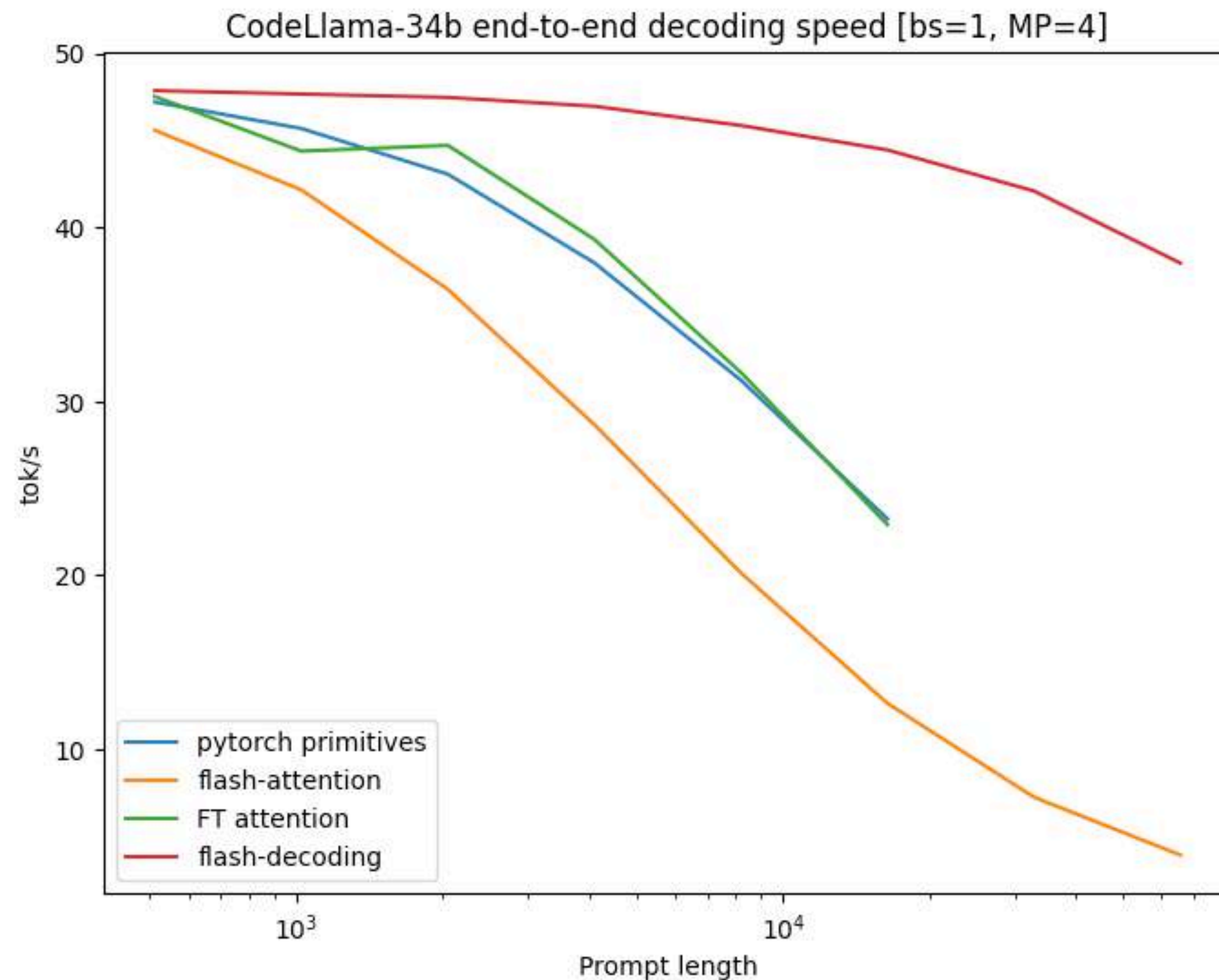
Source: <https://pytorch.org/blog/flash-decoding/>

Flash decoding



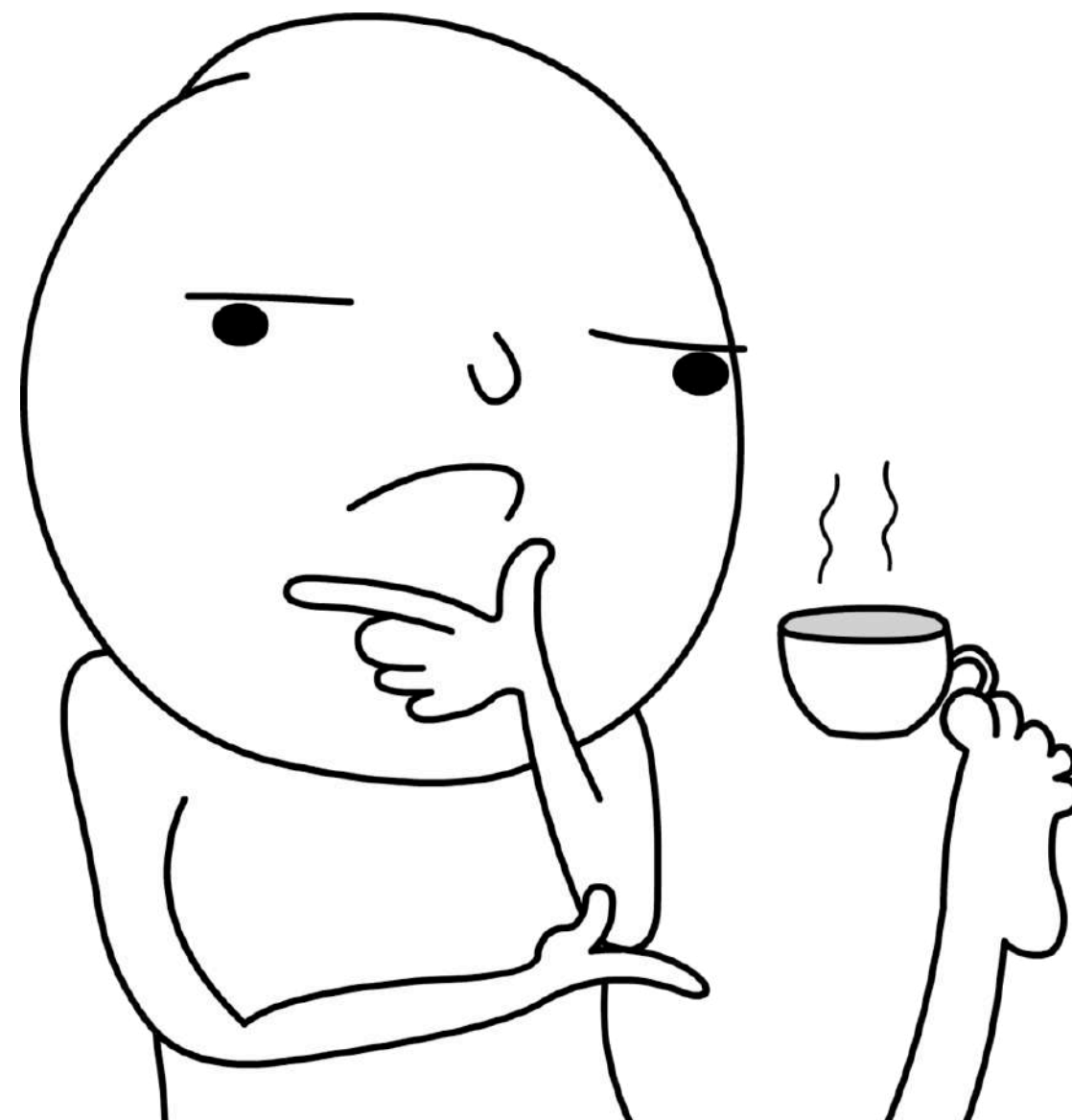
Source: <https://pytorch.org/blog/flash-decoding/>

Flash decoding



Source: <https://pytorch.org/blog/flash-decoding/>

What about long sequence batches?



Paged Attention

0. Before generation.

Seq
A

Prompt: "Alan Turing is a computer scientist"
Completion: ""

Logical KV cache blocks

Block 0				
Block 1				
Block 2				
Block 3				

Block table

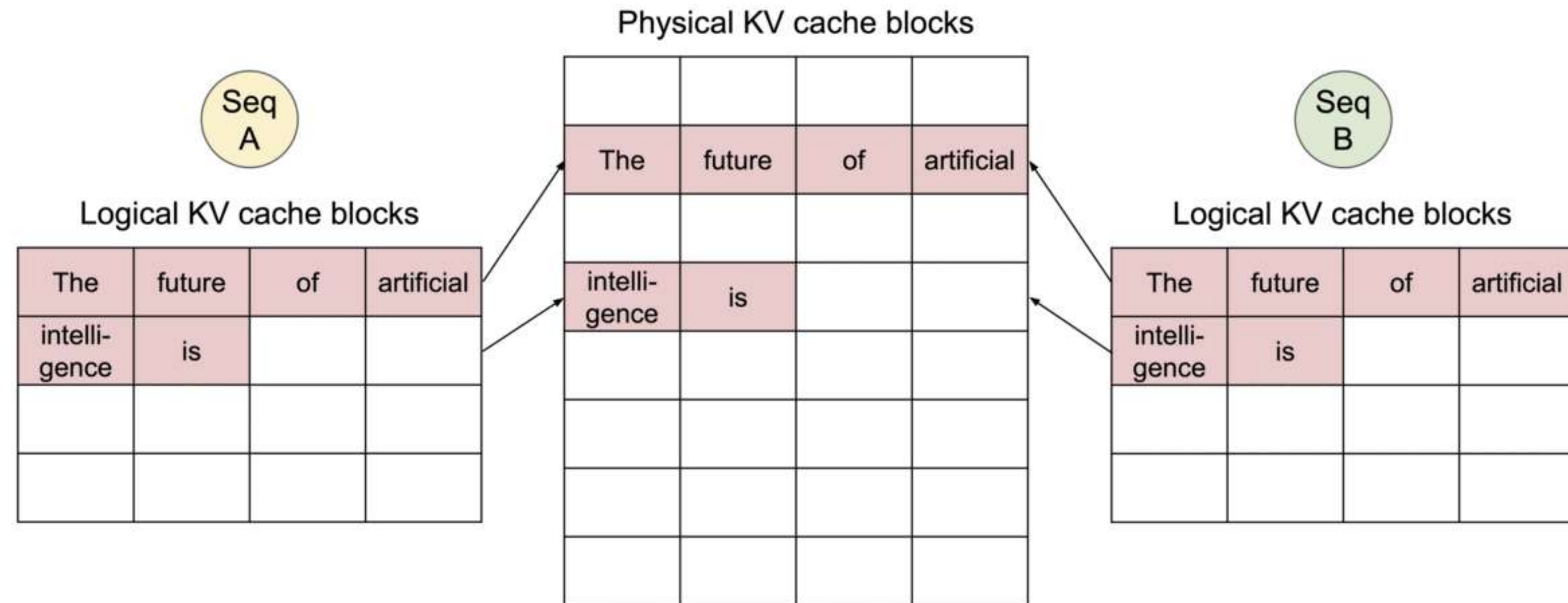
Physical block no.	# Filled slots
—	—
—	—
—	—
—	—

Physical KV cache blocks

Block 0				
Block 1				
Block 2				
Block 3				
Block 4				
Block 5				
Block 6				
Block 7				

Paged Attention

0. Shared prompt: Map logical blocks to the same physical blocks.

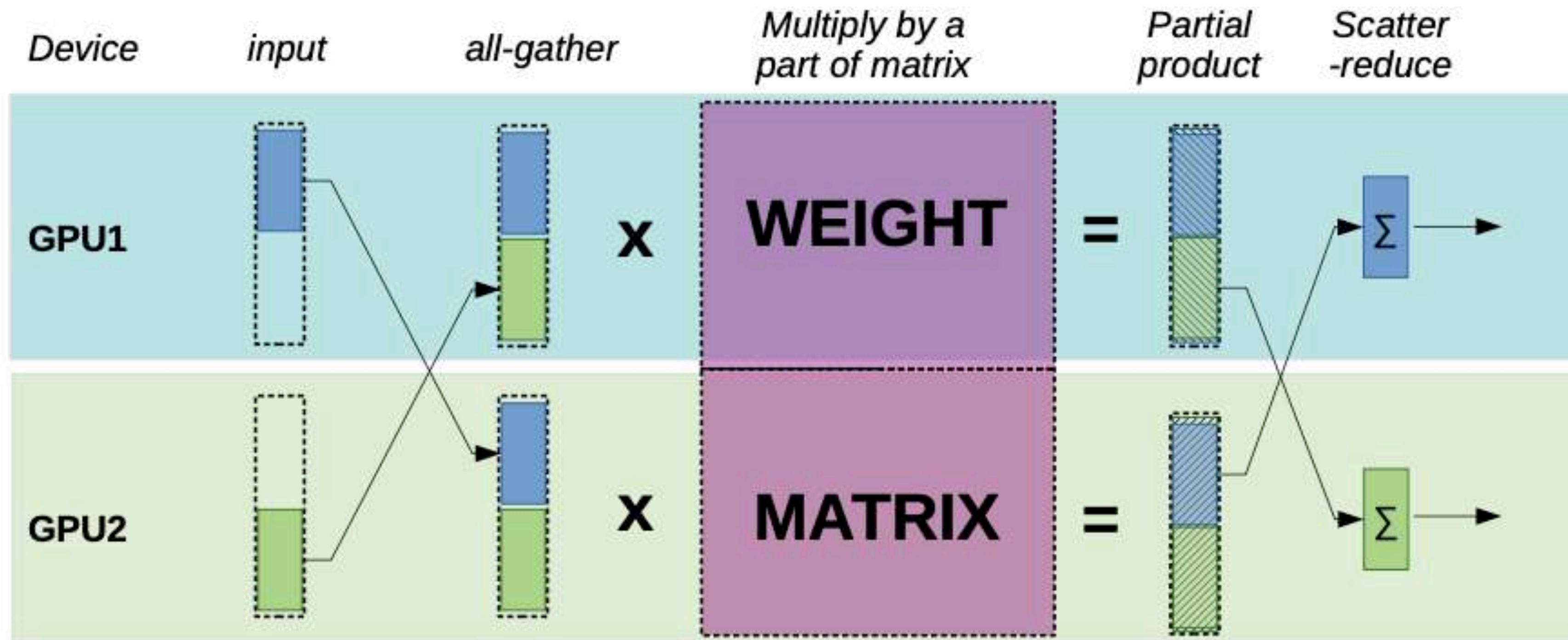


Source: <https://blog.vllm.ai/2023/06/20/vllm.html>

Huge KV caches



Tensor parallel



Tensor parallel

Performance of GPT-20B

Batch_size	Input Length	Output Length	Latency of single GPU (ms)	Latency of 2-way TP (ms)	Latency of 4-way TP (ms)	Latency of 8-way TP (ms)
1	20	8	225	147	102	89
2	20	8	225	152	108	94
4	20	8	228	158	113	100
8	20	8	239	169	121	107
16	20	8	268	191	133	113
32	20	8	331	230	155	127
64	20	8	452	314	200	169
128	20	8	726	484	318	256
256	20	8	1352	844	533	416
1	60	20	560	358	248	212
2	60	20	562	378	262	222
4	60	20	582	393	274	236
8	60	20	635	429	299	247
16	60	20	748	510	345	272
32	60	20	933	620	418	325
64	60	20	1352	887	574	454
128	60	20	2218	1384	928	699
256	60	20	4141	2424	1574	1152

Frameworks

<https://github.com/vllm-project/vllm>

<https://github.com/InternLM/lmdeploy>

<https://github.com/microsoft/DeepSpeed-MII>

<https://github.com/NVIDIA/TensorRT-LLM>

<https://github.com/ggerganov/llama.cpp>

Frameworks

<https://github.com/NVIDIA/TensorRT-LLM>

Business

- Different scenarios
- Different sources of context
- Sampling control
 - Cycles
 - Images
 - Censorship

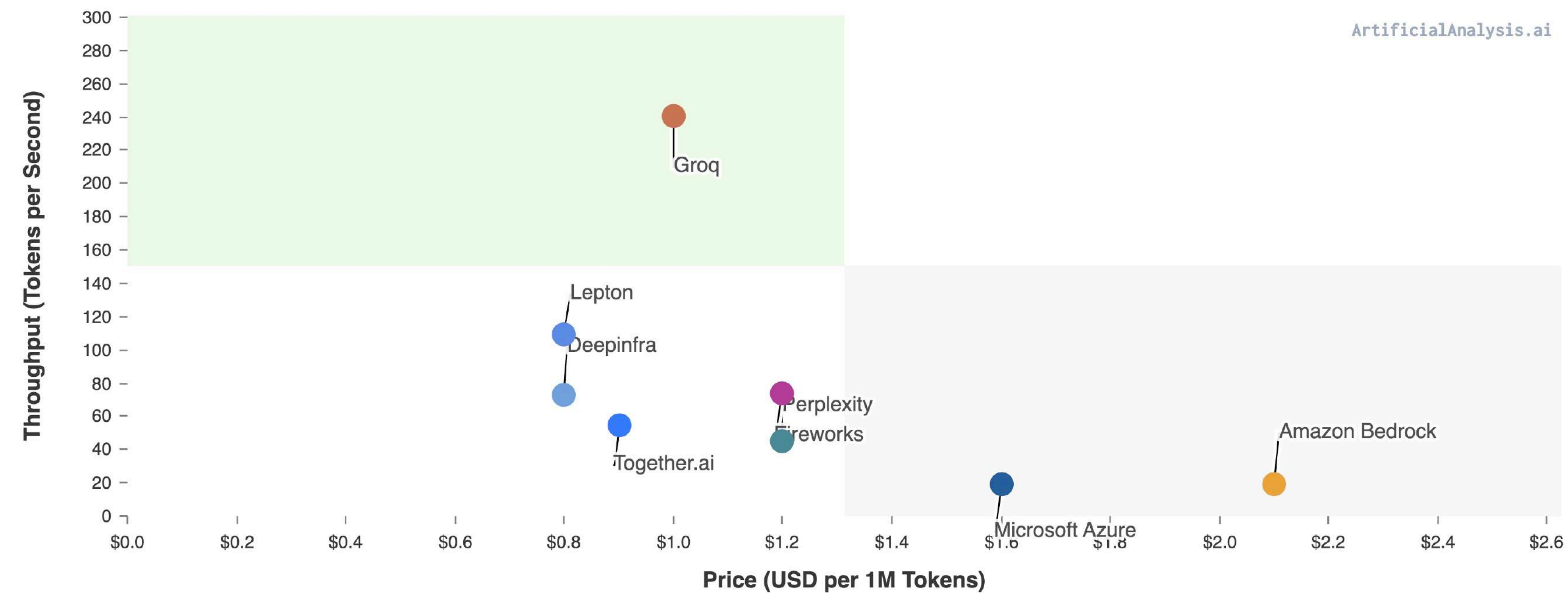
Groq

Throughput vs. Price: Llama 2 Chat (70B)

Quality: General reasoning index, Throughput: Tokens per Second, Price: USD per 1M Tokens

Most attractive quadrant

Microsoft Azure Amazon Bedrock Perplexity Together.ai Deepinfra Fireworks Groq Lepton



I think we use a system with 576 Groq chips for this demo (but I am not certain). There is no DRAM on our chip. We have 220 MB of SRAM per chip, so at 576 chips that would be 126 GB in total.