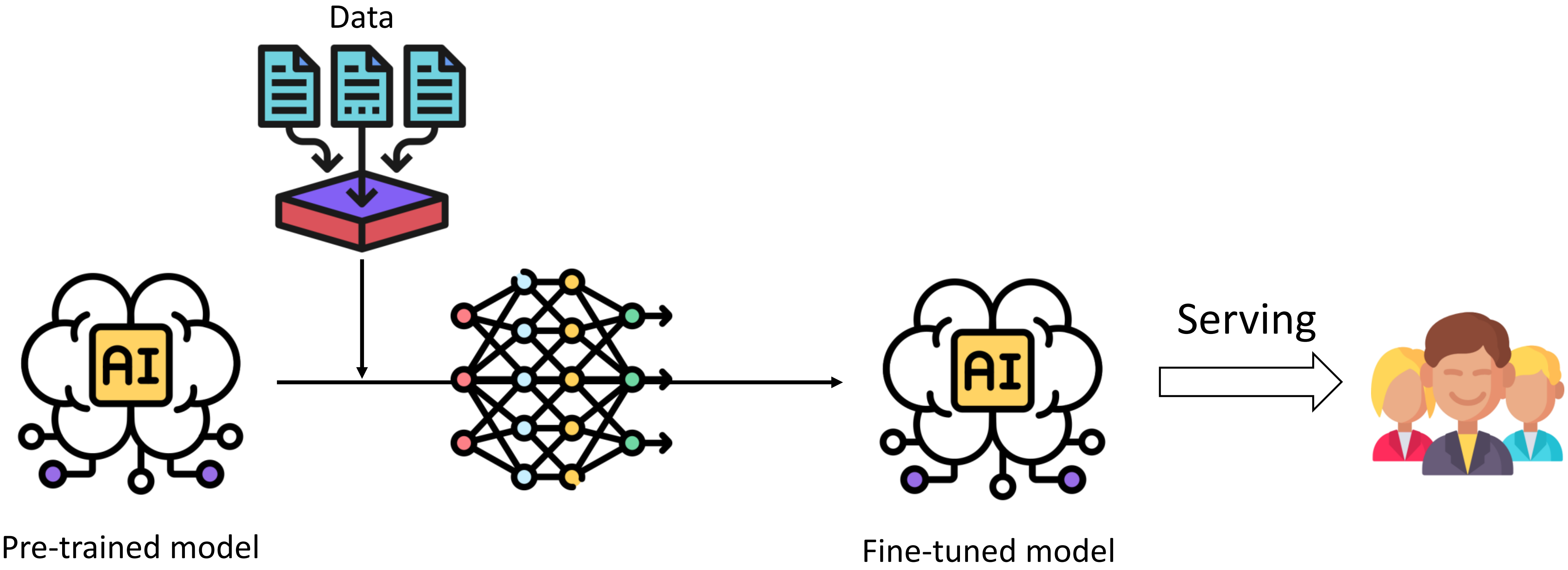
UNIVERSITY OF ILLINOIS
URBANA-CHAMPAIGN

CS 598
AI Efficiency: Systems and Algorithms
Overview & Key Challenges in AI Model Serving

Minjia Zhang

Computer Science Department

What is Model Serving?



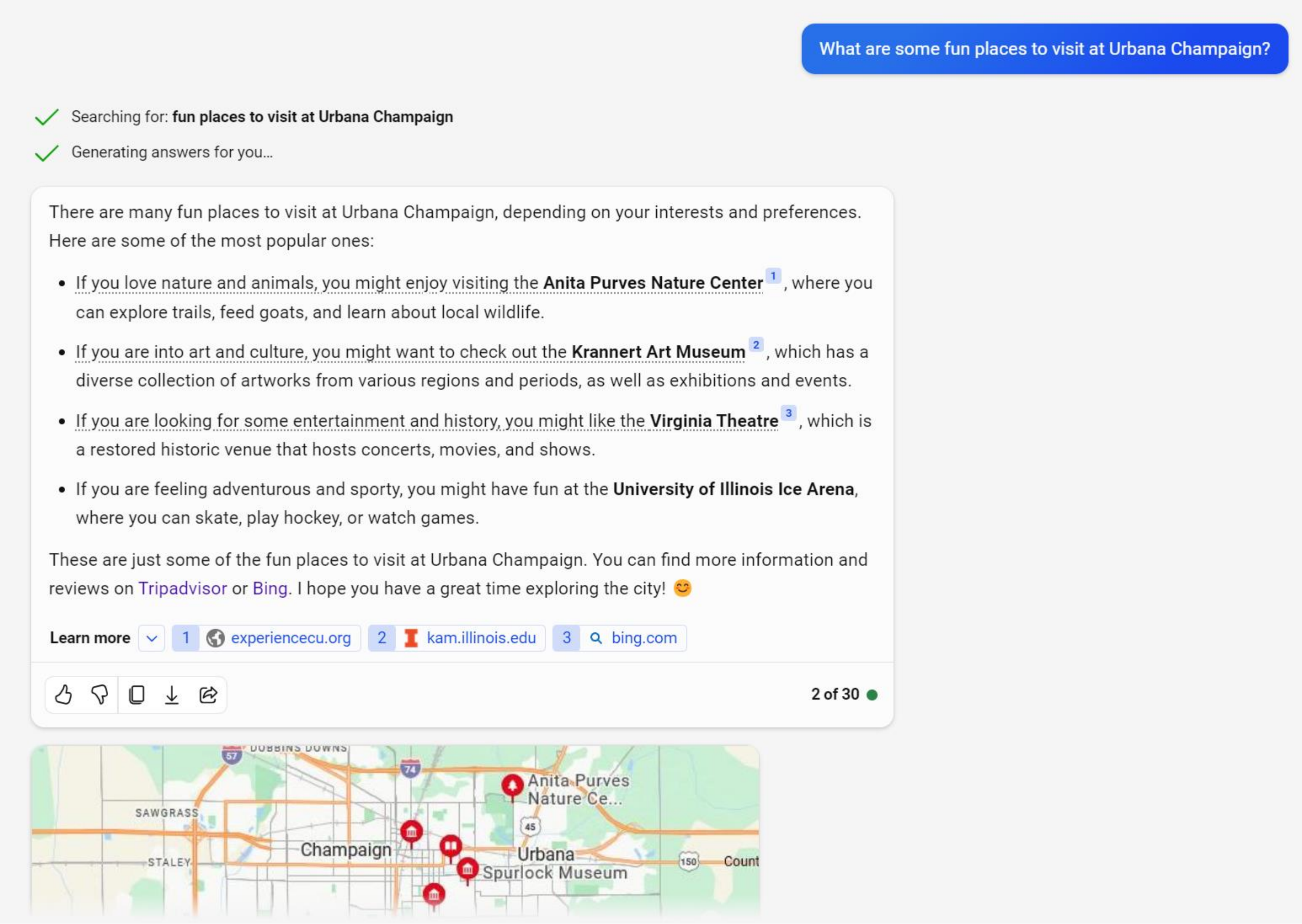
Data

Pre-trained model

Fine-tuned model

Serving

Serving Scenario 1: Online ChatBot



What are some fun places to visit at Urbana Champaign?

✓ Searching for: **fun places to visit at Urbana Champaign**

✓ Generating answers for you…

There are many fun places to visit at Urbana Champaign, depending on your interests and preferences. Here are some of the most popular ones:

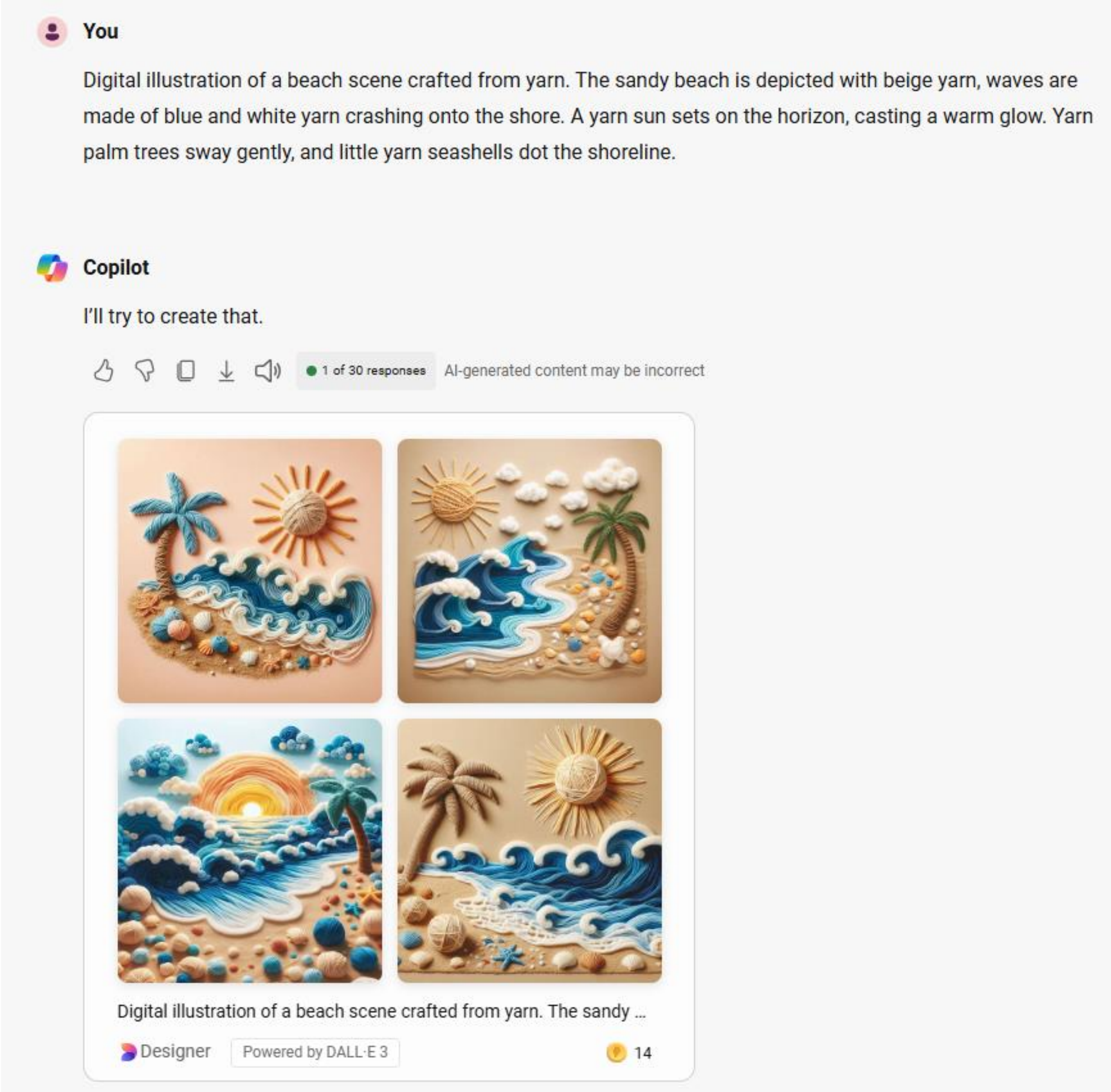
- If you love nature and animals, you might enjoy visiting the **Anita Purves Nature Center**[1], where you can explore trails, feed goats, and learn about local wildlife.
- If you are into art and culture, you might want to check out the **Krannert Art Museum**[2], which has a diverse collection of artworks from various regions and periods, as well as exhibitions and events.
- If you are looking for some entertainment and history, you might like the **Virginia Theatre**[3], which is a restored historic venue that hosts concerts, movies, and shows.
- If you are feeling adventurous and sporty, you might have fun at the **University of Illinois Ice Arena**, where you can skate, play hockey, or watch games.

These are just some of the fun places to visit at Urbana Champaign. You can find more information and reviews on Tripadvisor or Bing. I hope you have a great time exploring the city! 😊

Learn more 1 experiencecu.org 2 kam.illinois.edu 3 bing.com

2 of 30 ●

Serving Scenario 2: Online Image Generation



You

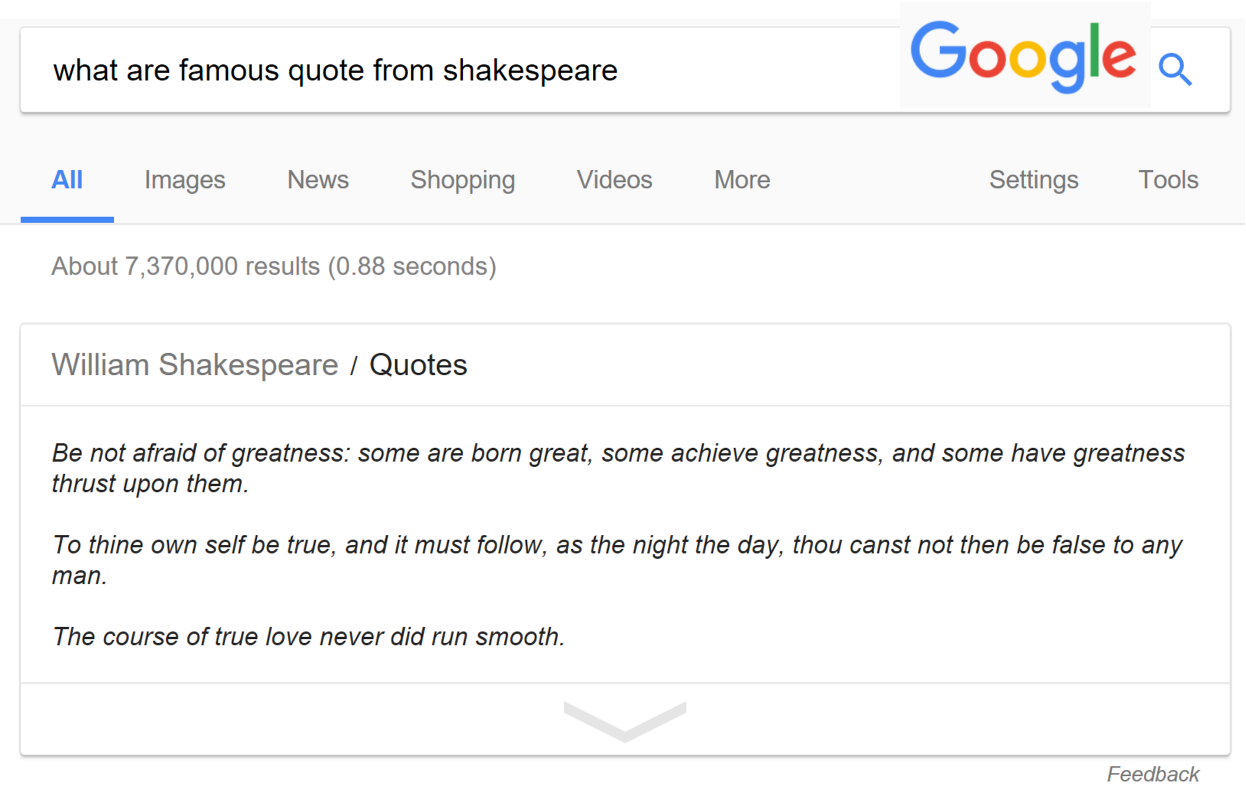
Digital illustration of a beach scene crafted from yarn. The sandy beach is depicted with beige yarn, waves are made of blue and white yarn crashing onto the shore. A yarn sun sets on the horizon, casting a warm glow. Yarn palm trees sway gently, and little yarn seashells dot the shoreline.

Copilot

I'll try to create that.

Digital illustration of a beach scene crafted from yarn. The sandy ...

Serving Scenario 3: Online Q&A



About 7,370,000 results (0.88 seconds)

William Shakespeare / Quotes

Be not afraid of greatness: some are born great, some achieve greatness, and some have greatness thrust upon them.

To thine own self be true, and it must follow, as the night the day, thou canst not then be false to any man.

The course of true love never did run smooth.

- Question and Answer Scenario
  - Direct answer not a list of webpages
  - Good quality of answer

Training -> Inference

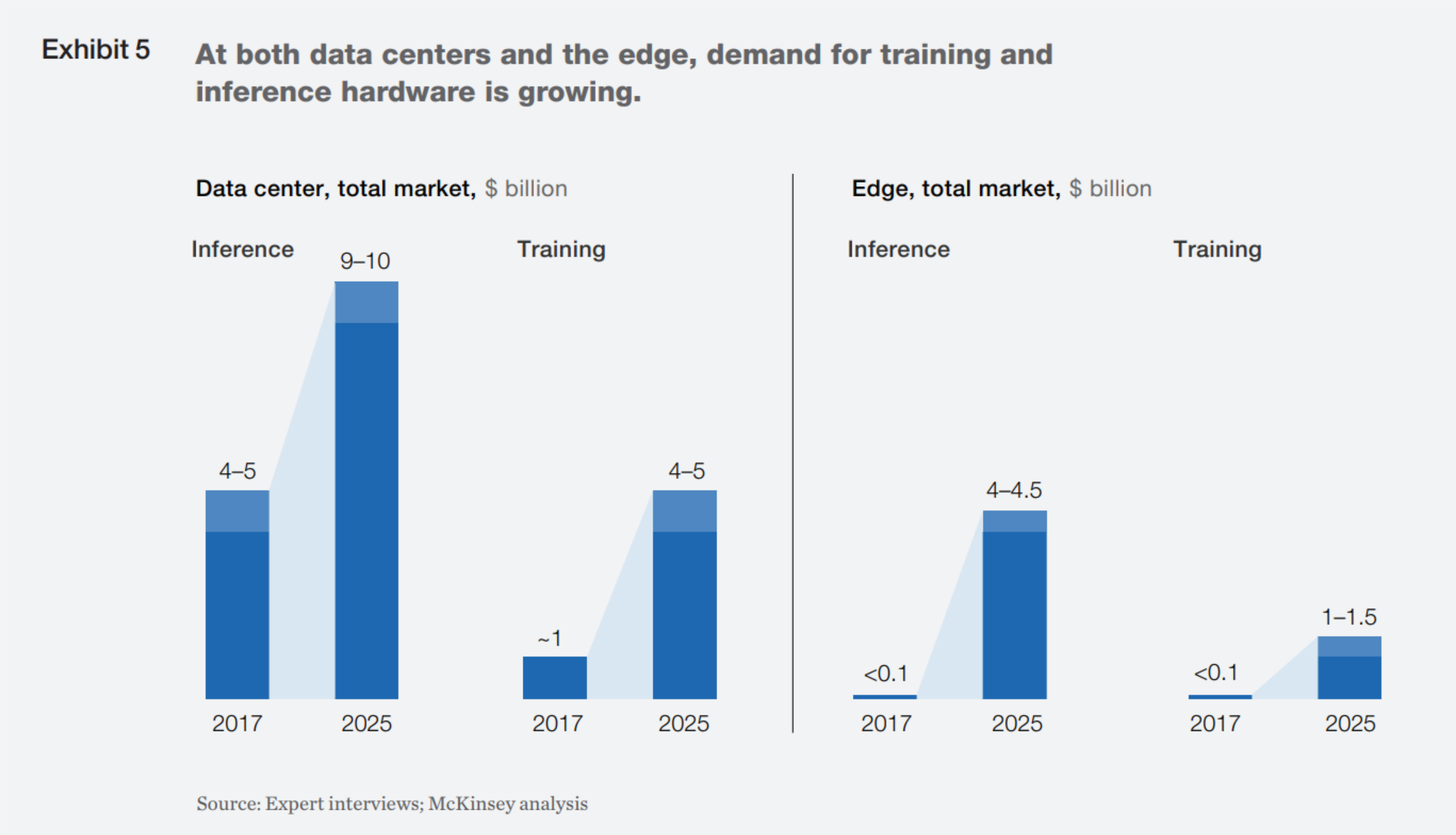


Exhibit 5 **At both data centers and the edge, demand for training and inference hardware is growing.**

Source: Expert interviews; McKinsey analysis

Inference Challenges

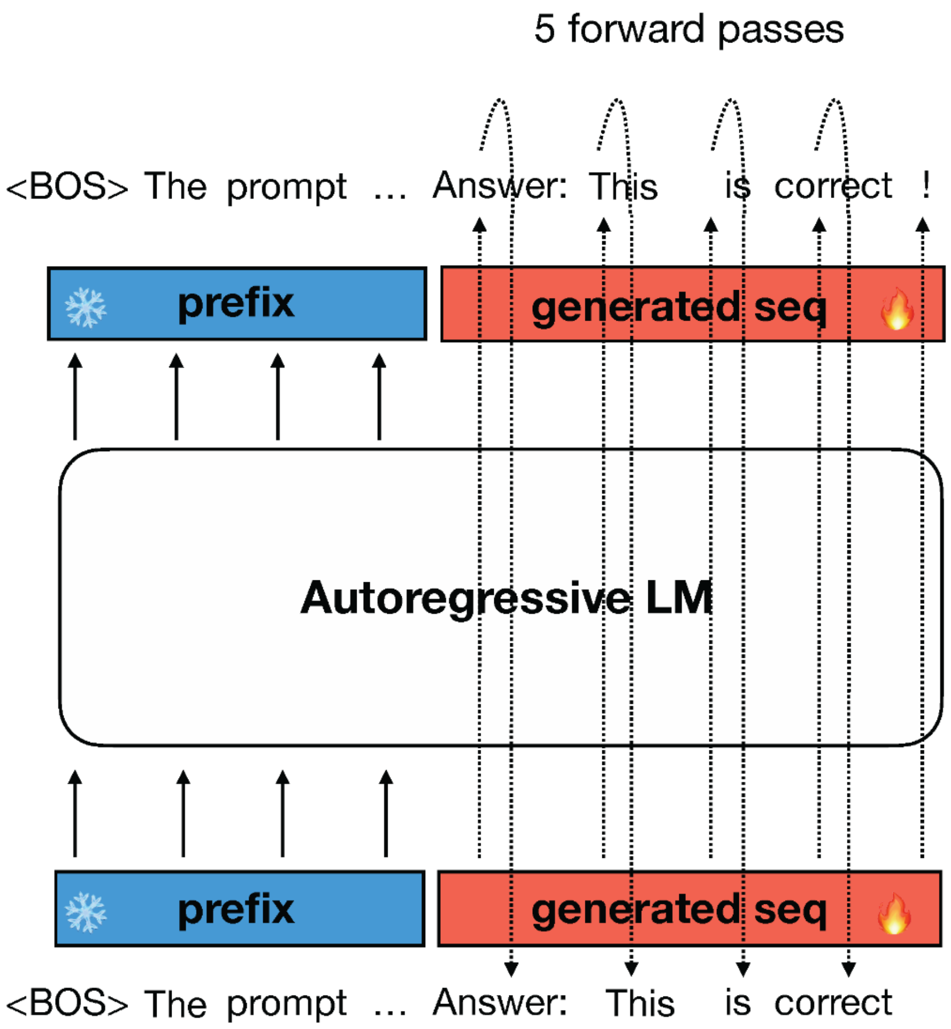
	Training	vs	Inference
Runtime	Weeks or months		Milliseconds or seconds
Challenges	TCO (Cost, Energy)		TCO (Cost, Energy)

Speed (LLM: token rates)

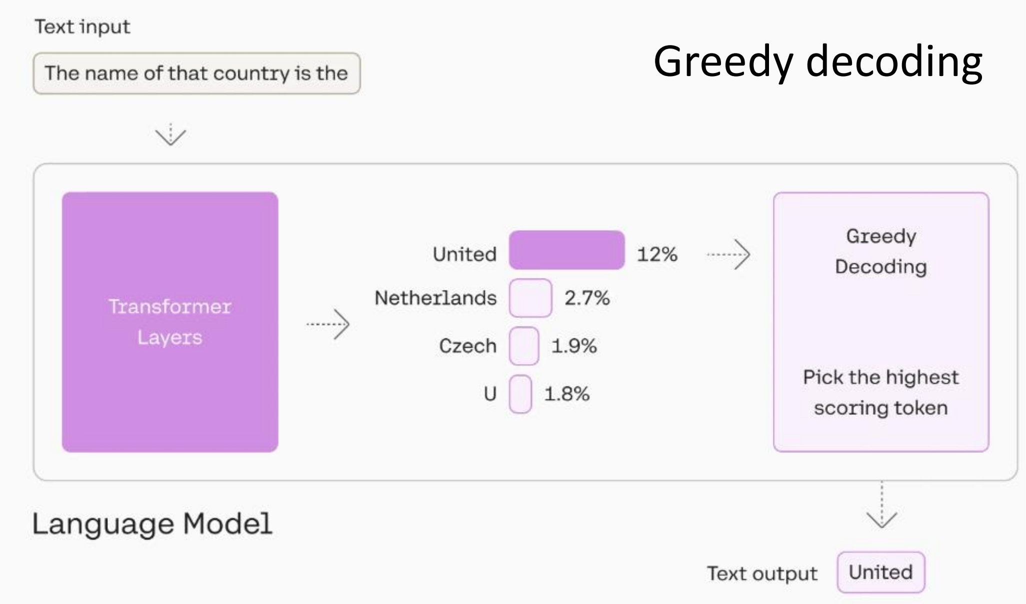
Model size

- Parameter volume
- LLM: Context length

LLM Autoregressive Generation

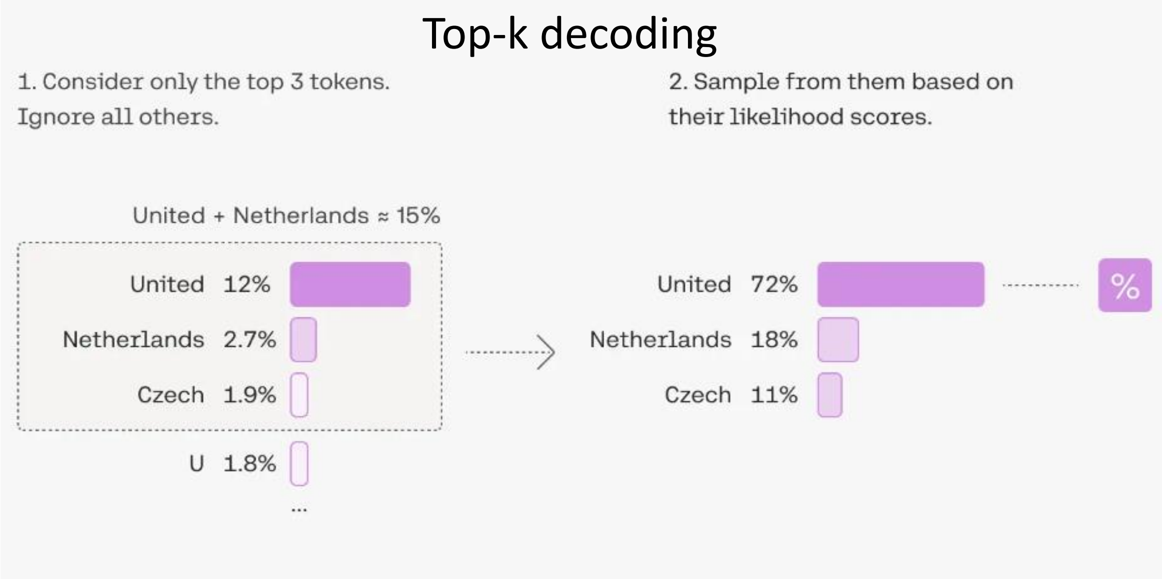

5 forward passes
<BOS> The prompt … Answer: This is correct !
❄ prefix
generated seq 🔥
Autoregressive LM
❄ prefix
generated seq 🔥
<BOS> The prompt … Answer: This is correct

Decoding Strategies



Greedy decoding

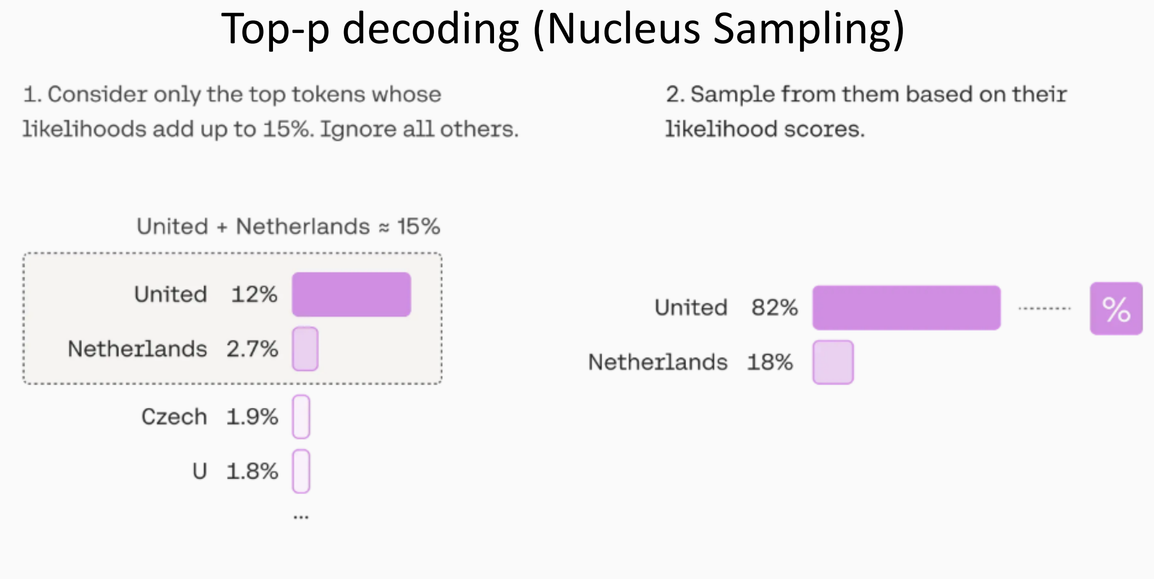
Top-k & Top-p, Cohere



Top-k decoding

1. Consider only the top 3 tokens. Ignore all others.

2. Sample from them based on their likelihood scores.



Top-p decoding (Nucleus Sampling)

1. Consider only the top tokens whose likelihoods add up to 15%. Ignore all others.

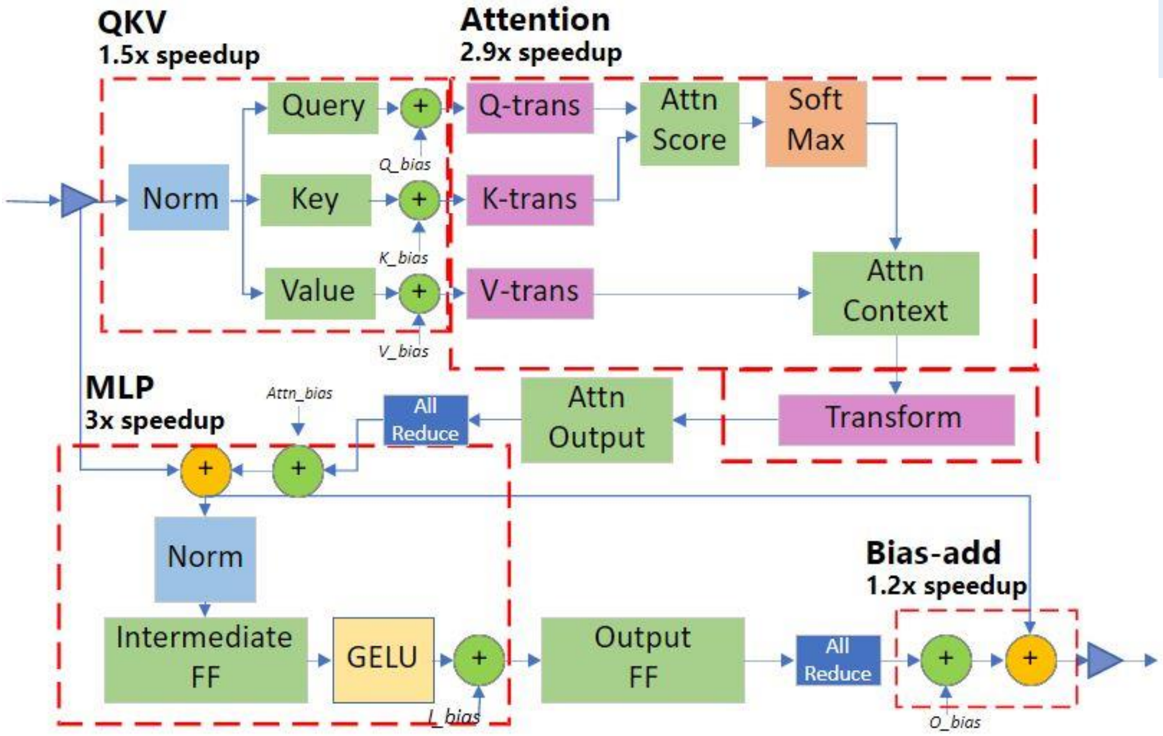
2. Sample from them based on their likelihood scores.

Serving Challenge: Long Latency

- Long serving latency blocks deployment
- Support advance models while meeting latency SLA and saving cost

DL Scenarios	Original Latency	Latency Target
Turing Prototype 2	~100ms	< 10ms
Turing Prototype 3	~107ms	< 10ms
Deep Query Document Similarity	10~12ms for [query, 1 doc] x 33 docs	< 6ms
Malta Click Features	10ms for [query, 1 passage] x 150 passages	< 5ms
Ads seq2seq model for query rewriting	~51ms	< 5ms

Customized Kernels



DeepSpeed-Inference: enabling efficient inference of transformer models at unprecedented scale, SC 2022

ByteTransformer: A High-Performance Transformer Boosted for Variable-Length Inputs, 2023

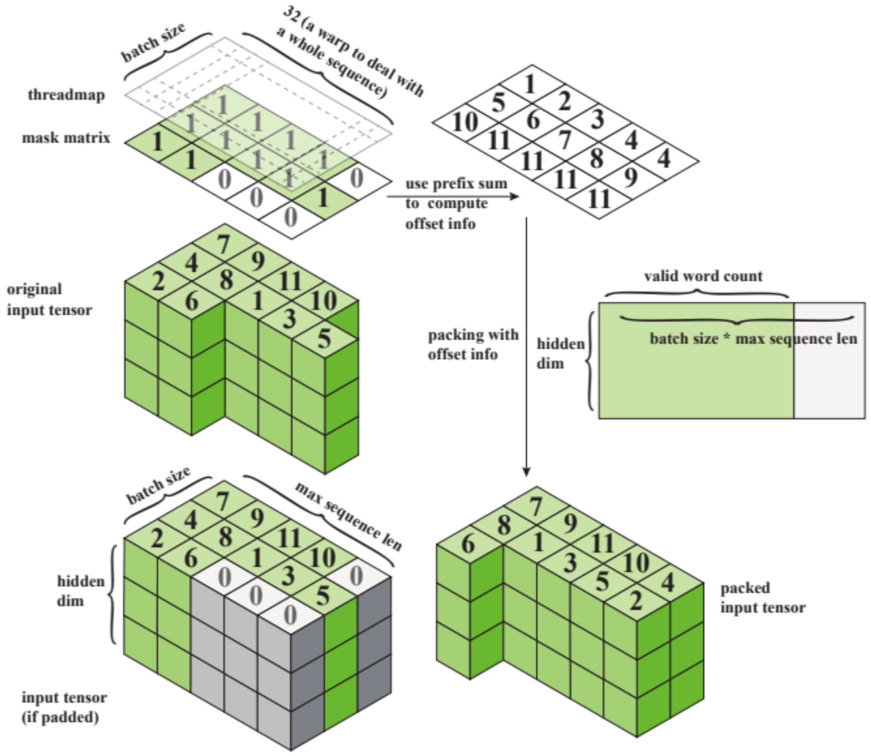
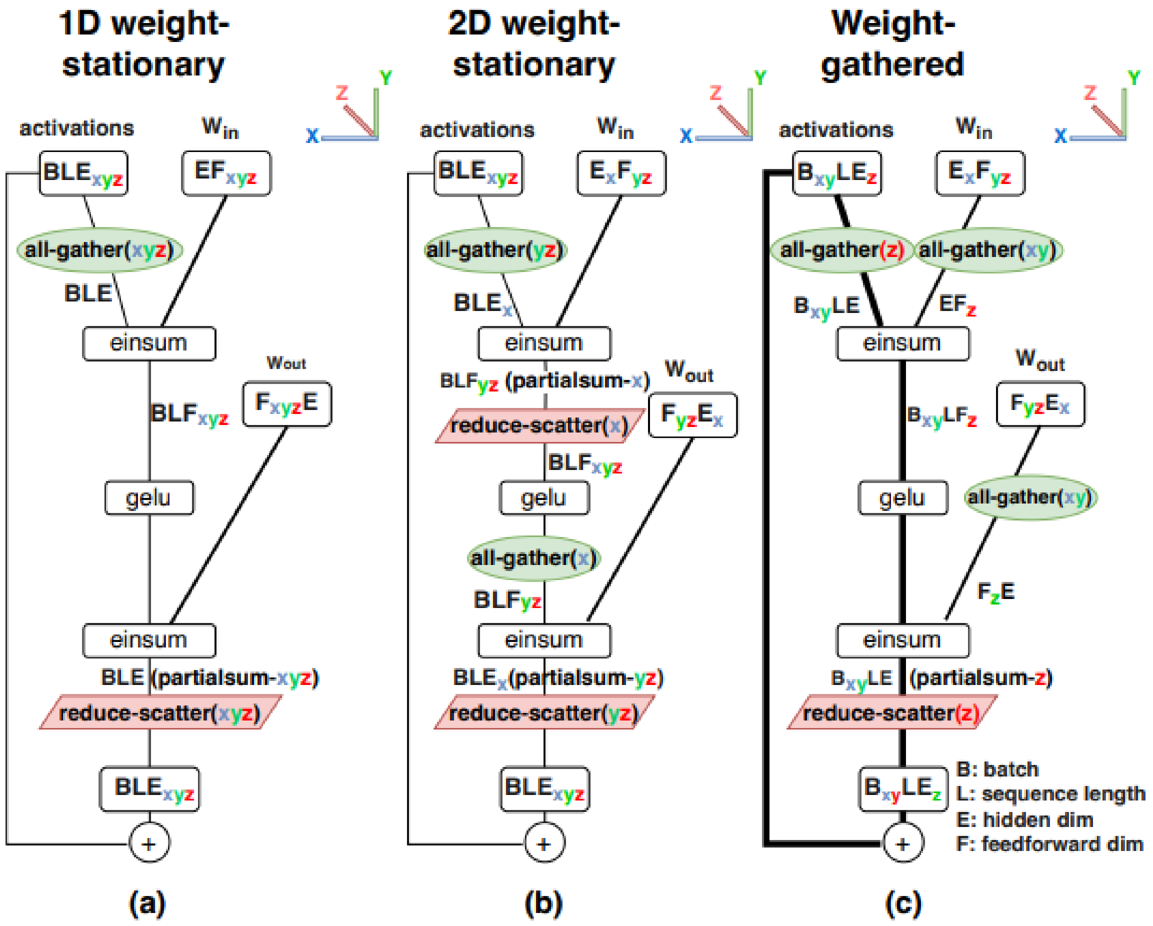


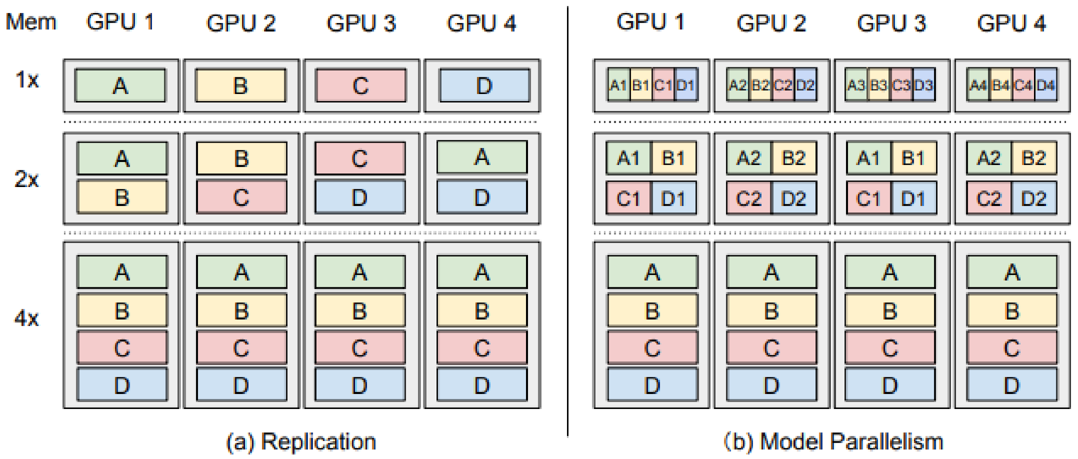
Fig. 4: The zero padding algorithm.

Multi-GPU Inference via Partitioned Layouts



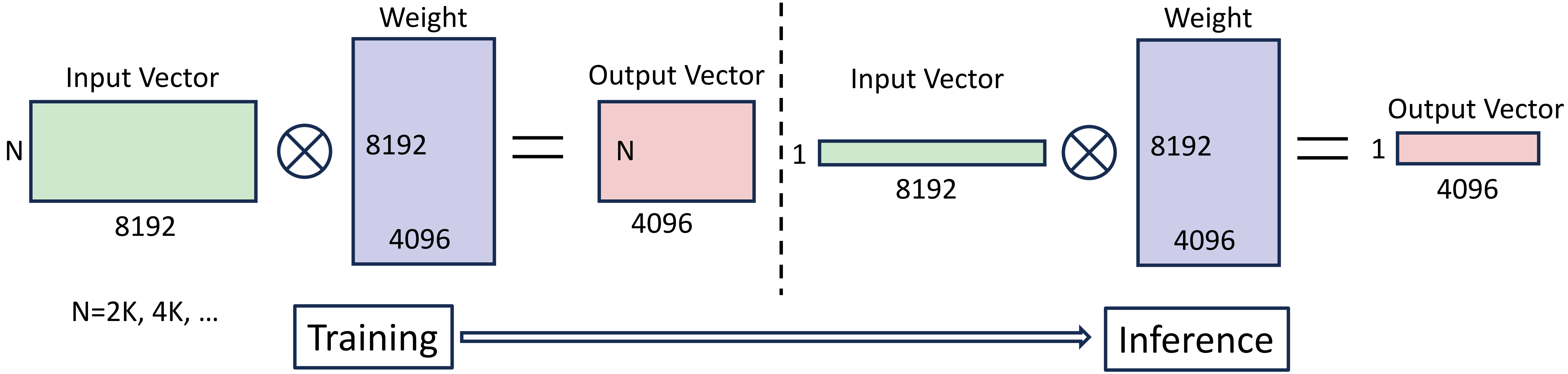
Efficiently Scaling Transformer Inference, MLSys 2023

AlpaServe: Statistical Multiplexing with Model Parallelism for Deep Learning Serving, OSDI 2023

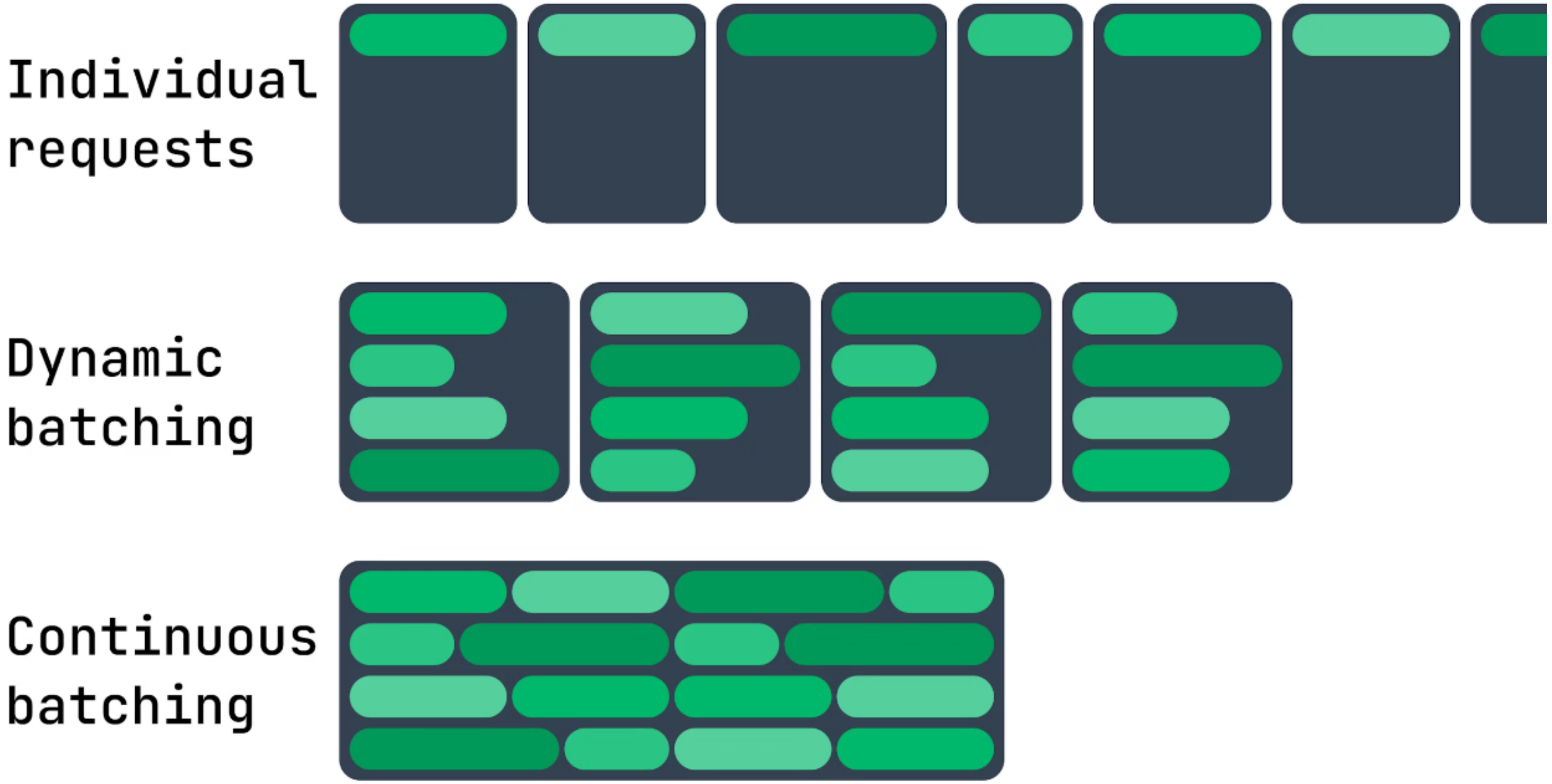


Inference Challenge: Limited Parallelism

- Small batch size ⟹ Low data reuse
- Autoregressive generation ⟹ Sequential dependency



Batching Strategies for LLM Inference



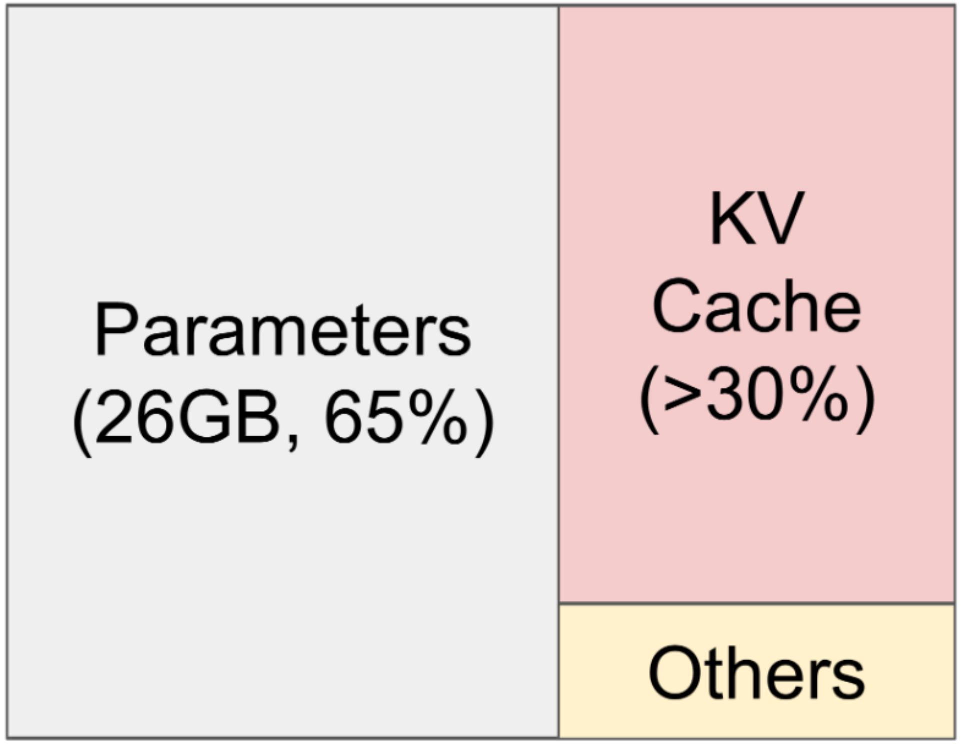
Orca: A Distributed Serving System for Transformer-Based Generative Models, OSDI 2022

Inference Challenge: Large Memory Footprint

- Model parameters
  - # Layers
  - # Hidden dim

- KV cache
  - Batch size
  - Sequence length
  - # Layers
  - # Hidden

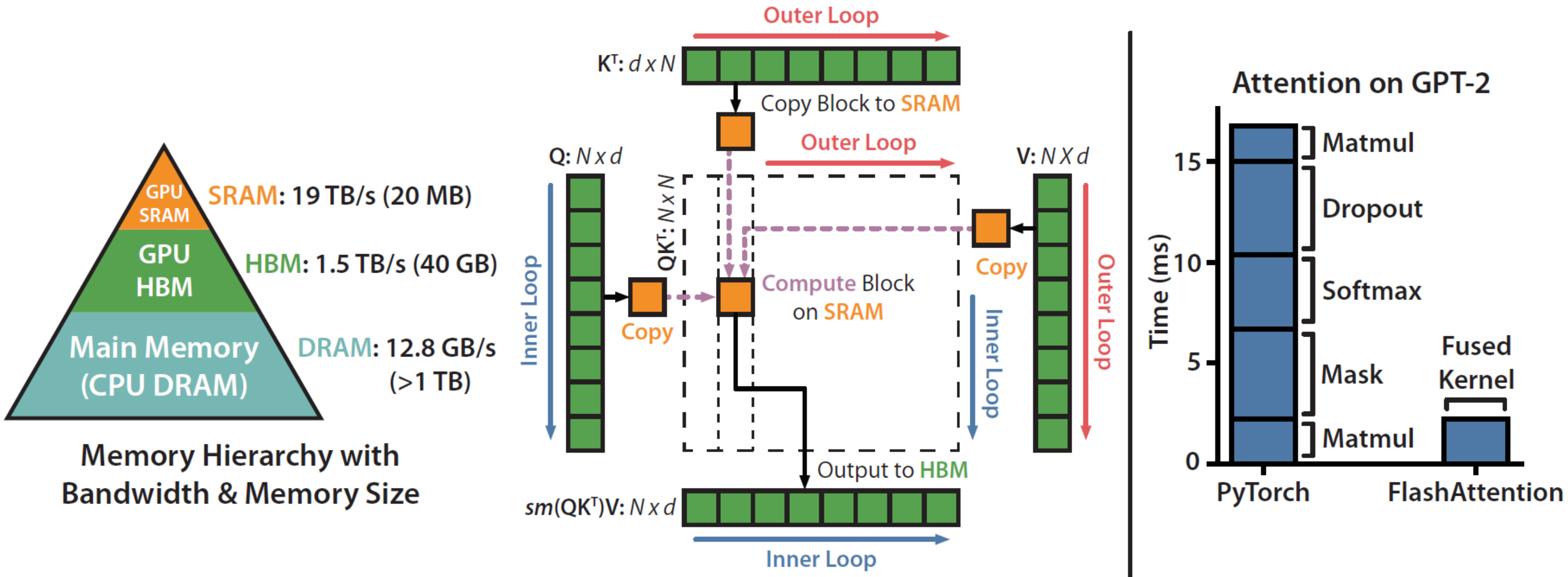
- Activation and others



OPT-13B on A100 40 GB

Efficient Memory Management for Large Language Model Serving with PagedAttention, by Kwon et al., 2023

FlashAttention

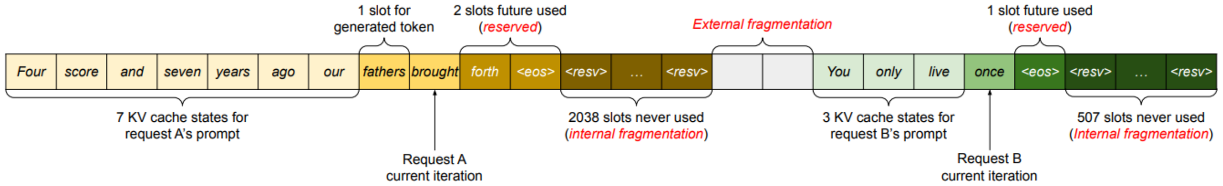


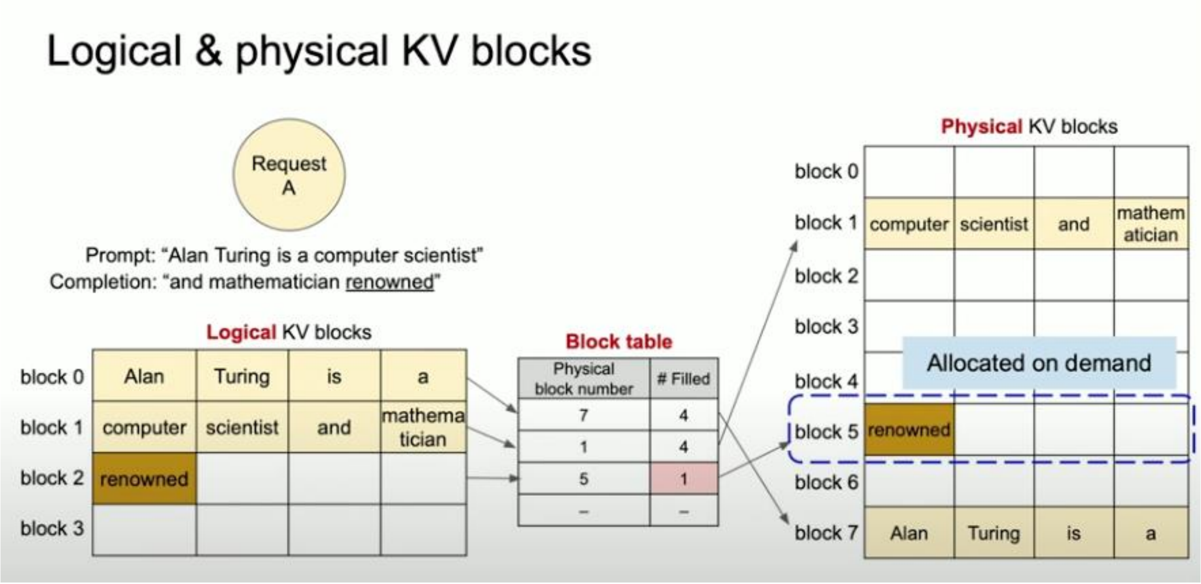
Memory Hierarchy with Bandwidth & Memory Size

FlashAttention

Fast and Memory-Efficient Exact Attention with IO-Awareness, 2023

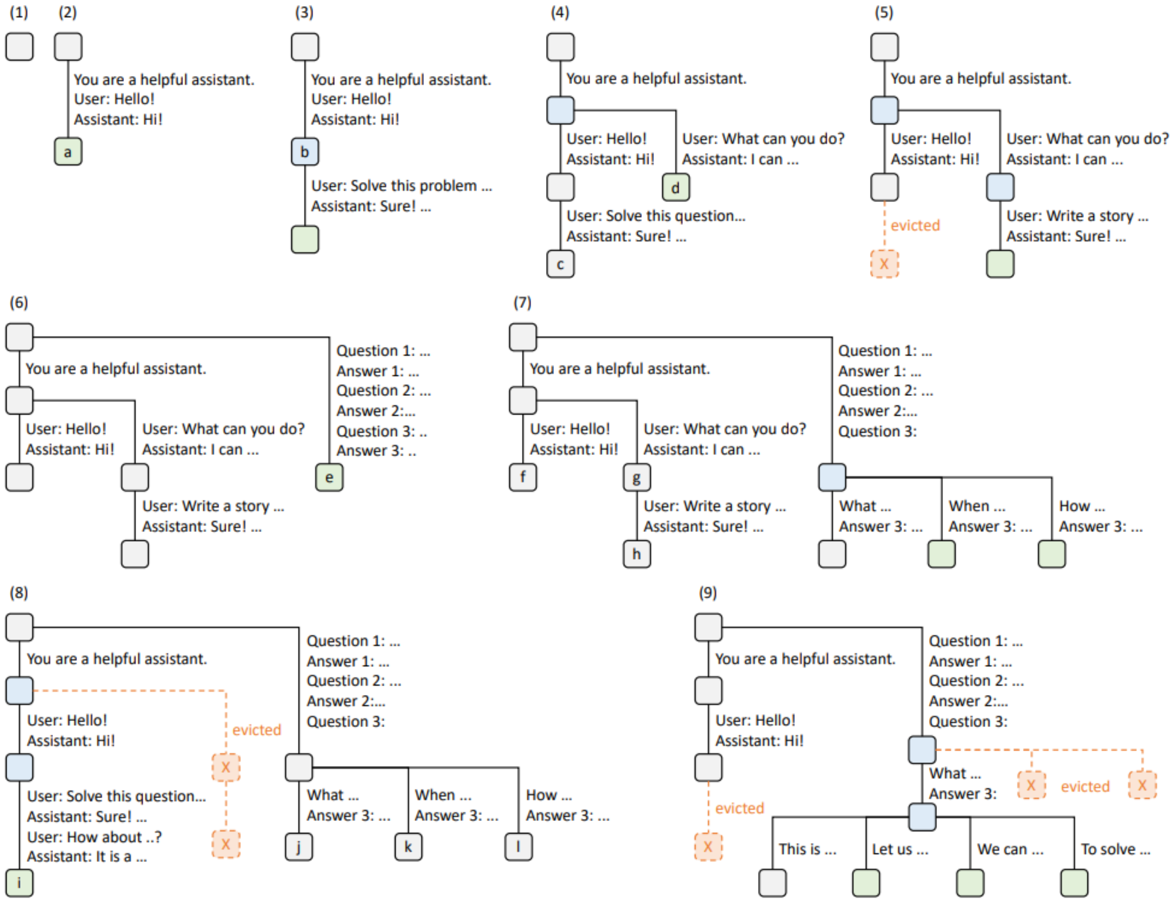
PagedAttention


1 slot for generated token
2 slots future used (reserved)
External fragmentation
1 slot future used (reserved)
Four
score
and
seven
years
ago
our
fathers
brought
forth
<eos>
<resv>
…
<resv>
You
only
live
once
<eos>
<resv>
…
<resv>
7 KV cache states for request A's prompt
Request A current iteration
2038 slots never used (internal fragmentation)
3 KV cache states for request B's prompt
Request B current iteration
507 slots never used (Internal fragmentation)

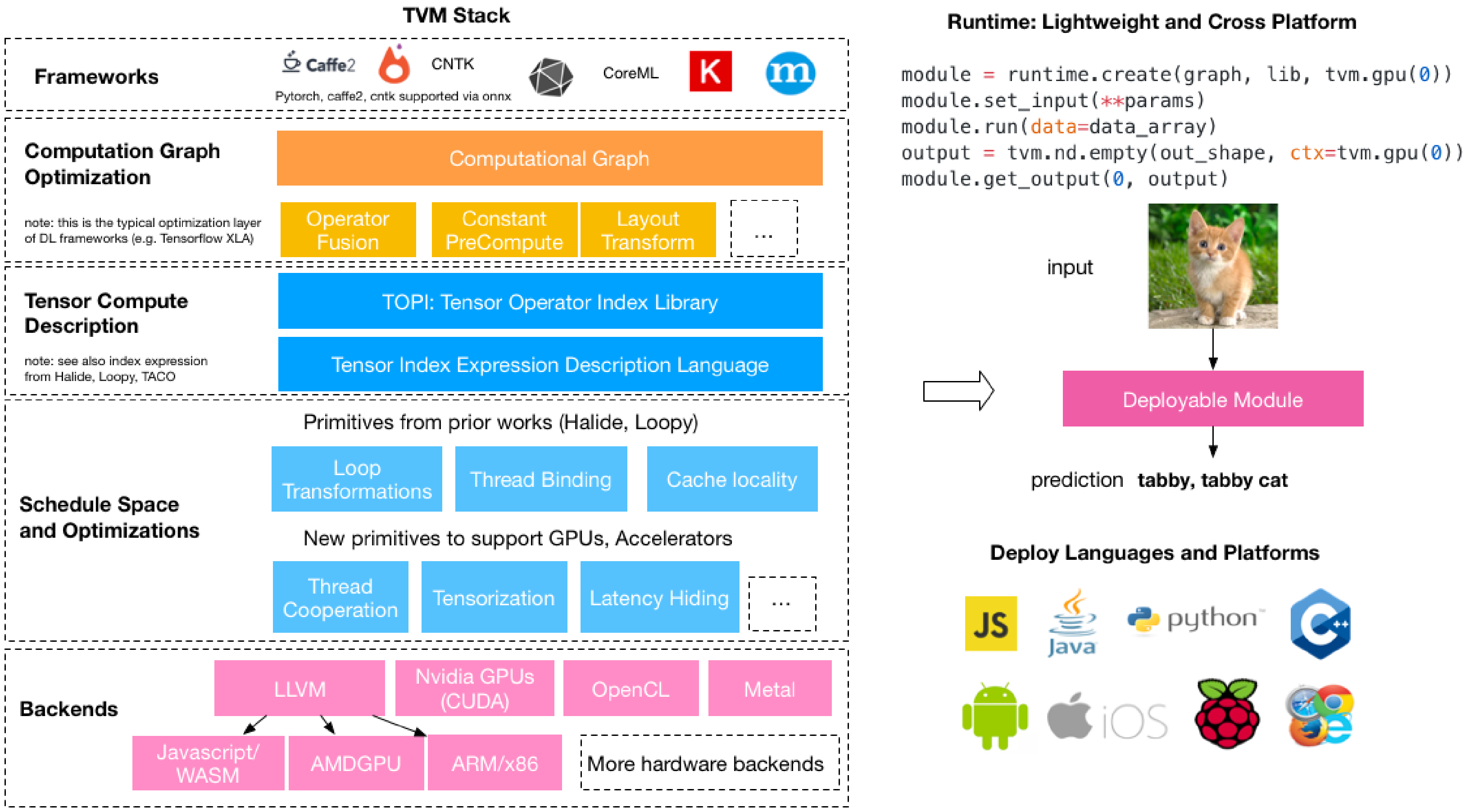


Efficient Memory Management for Large Language Model Serving with PagedAttention, 2023

SGLang: Efficient Execution of Structured Language Model Programs, 2024



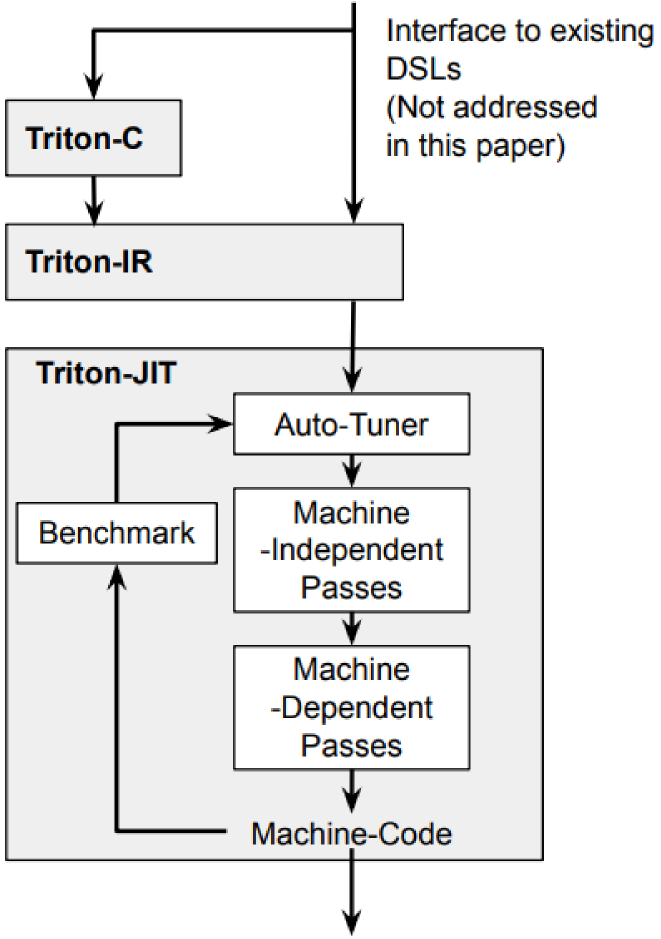
DL Compilation


TVM Stack
Frameworks
Caffe2
CNTK
CoreML
Pytorch, caffe2, cntk supported via onnx
Computation Graph Optimization
note: this is the typical optimization layer of DL frameworks (e.g. Tensorflow XLA)
Computational Graph
Operator Fusion
Constant PreCompute
Layout Transform
...
Tensor Compute Description
note: see also index expression from Halide, Loopy, TACO
TOPI: Tensor Operator Index Library
Tensor Index Expression Description Language
Schedule Space and Optimizations
Primitives from prior works (Halide, Loopy)
Loop Transformations
Thread Binding
Cache locality
New primitives to support GPUs, Accelerators
Thread Cooperation
Tensorization
Latency Hiding
...
Backends
LLVM
Nvidia GPUs (CUDA)
OpenCL
Metal
Javascript/ WASM
AMDGPU
ARM/x86
More hardware backends
Runtime: Lightweight and Cross Platform
input
Deployable Module
prediction
tabby, tabby cat
Deploy Languages and Platforms
JS
Java
python
iOS

```
module = runtime.create(graph, lib, tvm.gpu(0))
module.set_input(**params)
module.run(data=data_array)
output = tvm.nd.empty(out_shape, ctx=tvm.gpu(0))
module.get_output(0, output)
```

TVM: An Automated End-to-End Optimizing Compiler for Deep Learning, 2018

Triton: An Intermediate Language and Compiler for Tiled Neural Network Computations, 2019


Interface to existing DSLs (Not addressed in this paper)
Triton-C
Triton-IR
Triton-JIT
Auto-Tuner
Benchmark
Machine -Independent Passes
Machine -Dependent Passes
Machine-Code

Class Related

- Slack Channel – Update display name
- Schedule – Newly joined students please send papers you are interested in presenting to me and the TA
- Presentation – Guidance on what to include (course website)

QA