Expert Parallelism in LLM Training

Spring 2026

Lecturer: Yuedong (Steven) Xu

Fudan University

ydxu@fudan.edu.cn

Disclaimer

Machine learning systems is a broad and rapidly evolving field. The course material has been developed using a broad spectrum of resources, including research papers, lecture slides, blogposts, research talks, tutorial videos, and other materials shared by the research community.

Distributed LLM Training: Outline

- Data Parallelism
  - Parameter-Server, All-Reduce, Memory Optimization
- Model Parallelism
  - Pipeline Parallelism, Tensor Parallelism, Sequence Parallelism
- Mixture of Experts
  - **Principles**
  - Parallel Training
  - DeepEP (Extended Learning)

Mixture of Experts

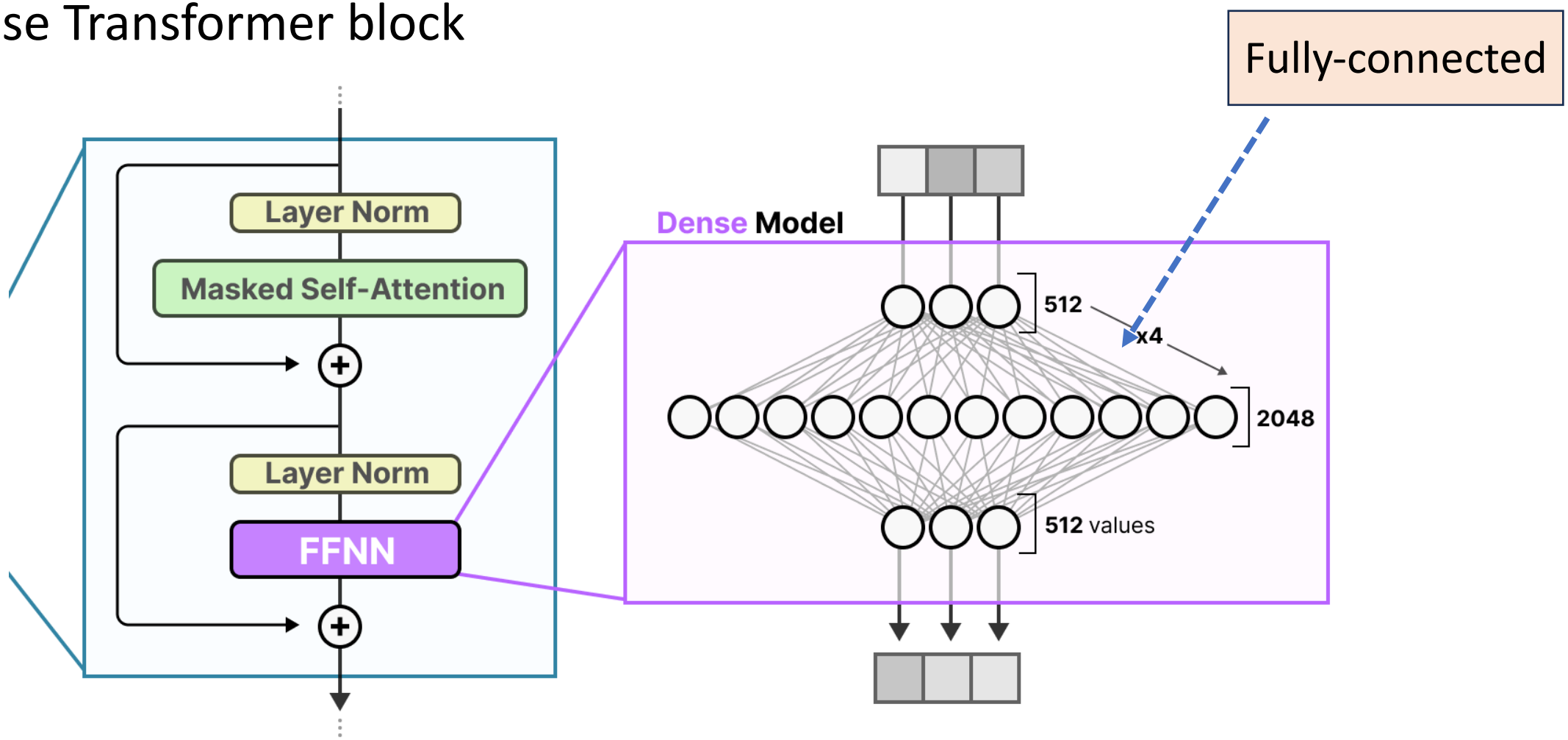
- Mixture of Experts in Early Days
  - Robert Jacobs, Michael Jordan, Steven Nowlan and Geoffrey Hinton (1991)
    - “a new supervised learning procedure for systems composed of many separate networks, each of which learns to handle a subset of the complete set of training cases.”
  - David Eigen, Marc'Aurelio Ranzato, Ilya Sutskever (2013)
    - “extend the Mixture of Experts to a stacked model, the Deep Mixture of Experts, with multiple sets of gating and experts.”
- Mixture of Experts in Transformer
  - GShard by Google (2020)

What shall we learn in this lecture

- We only learn
  - the structure of MoE in today's language models
  - the system challenges faced by MoE model training
  - recent progresses on addressing All-to-All issues
- We encourage students to learn by themselves
  - the instability of MoE model training
  - the design of new MoE framework
  - the fine-tuning techniques of MoE models

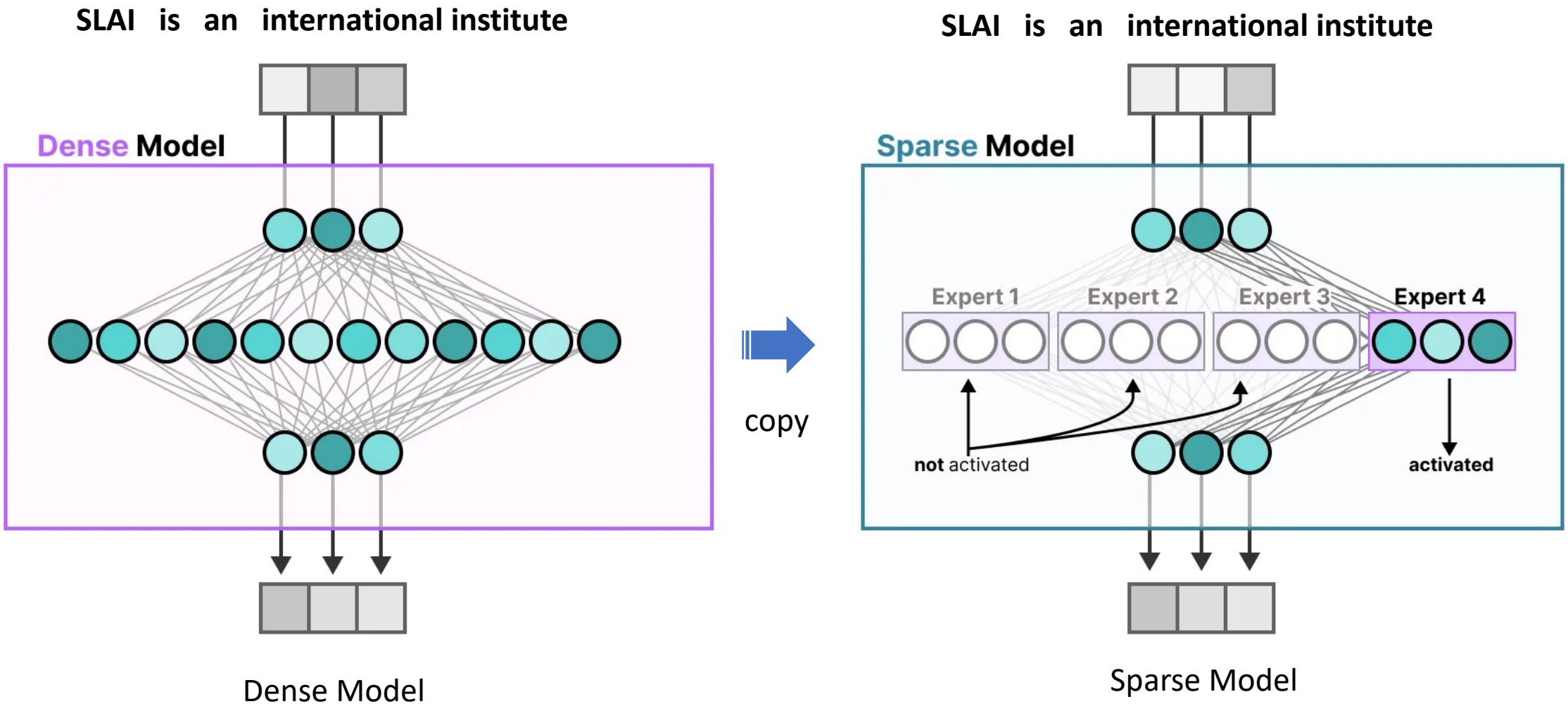
Mixture of Experts

- Dense Transformer block



《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

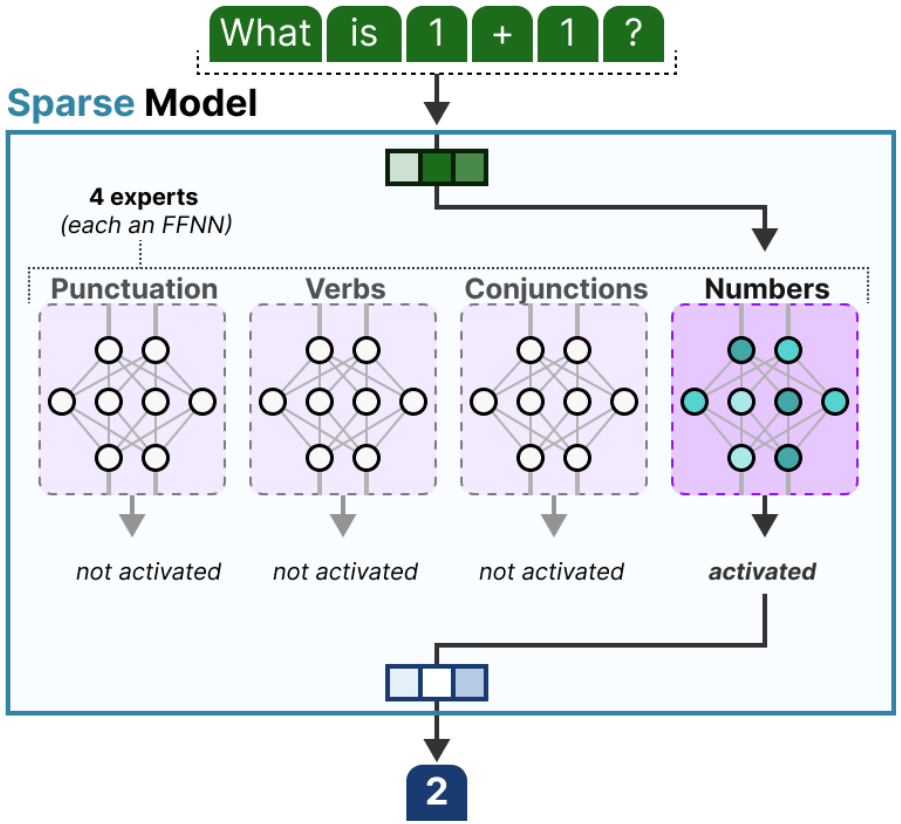


Dense Model

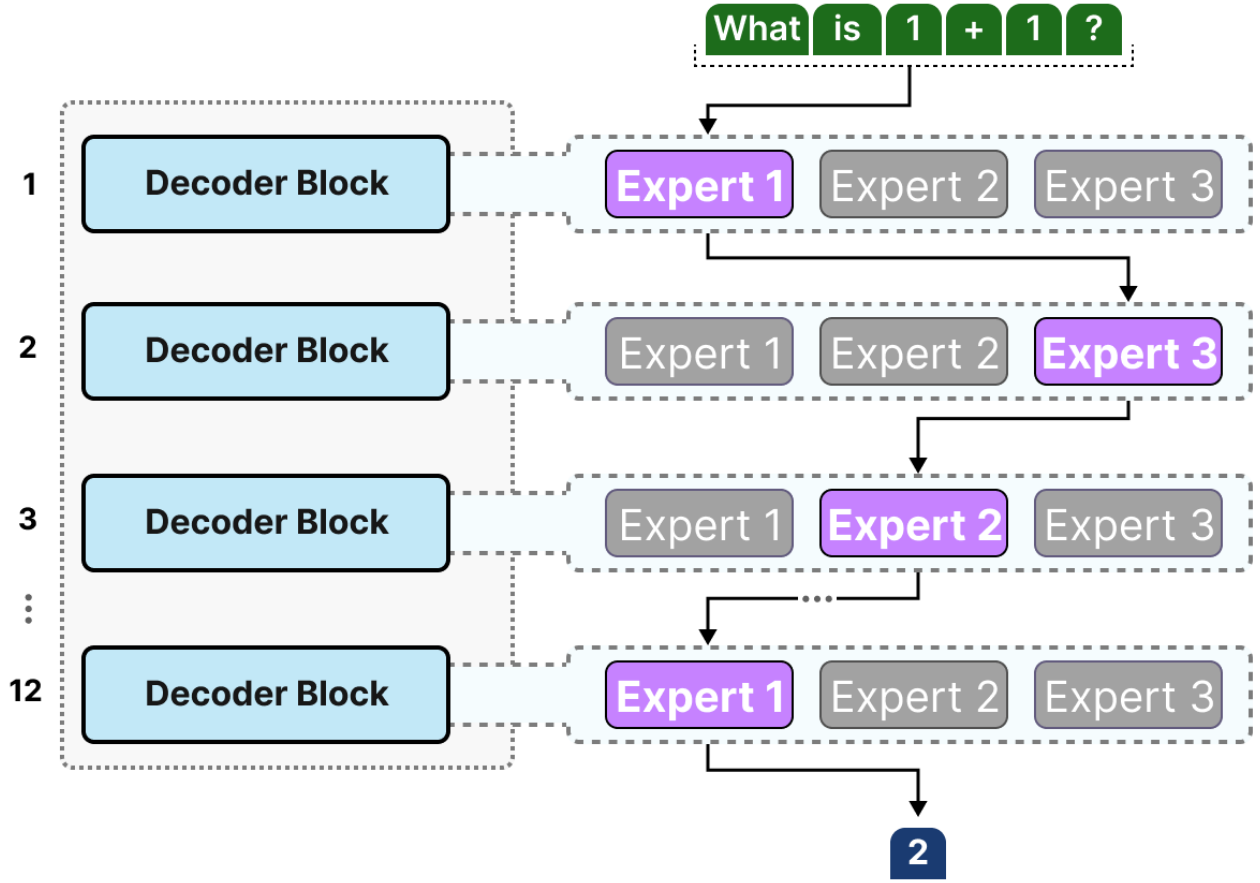
Sparse Model

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts



stacking



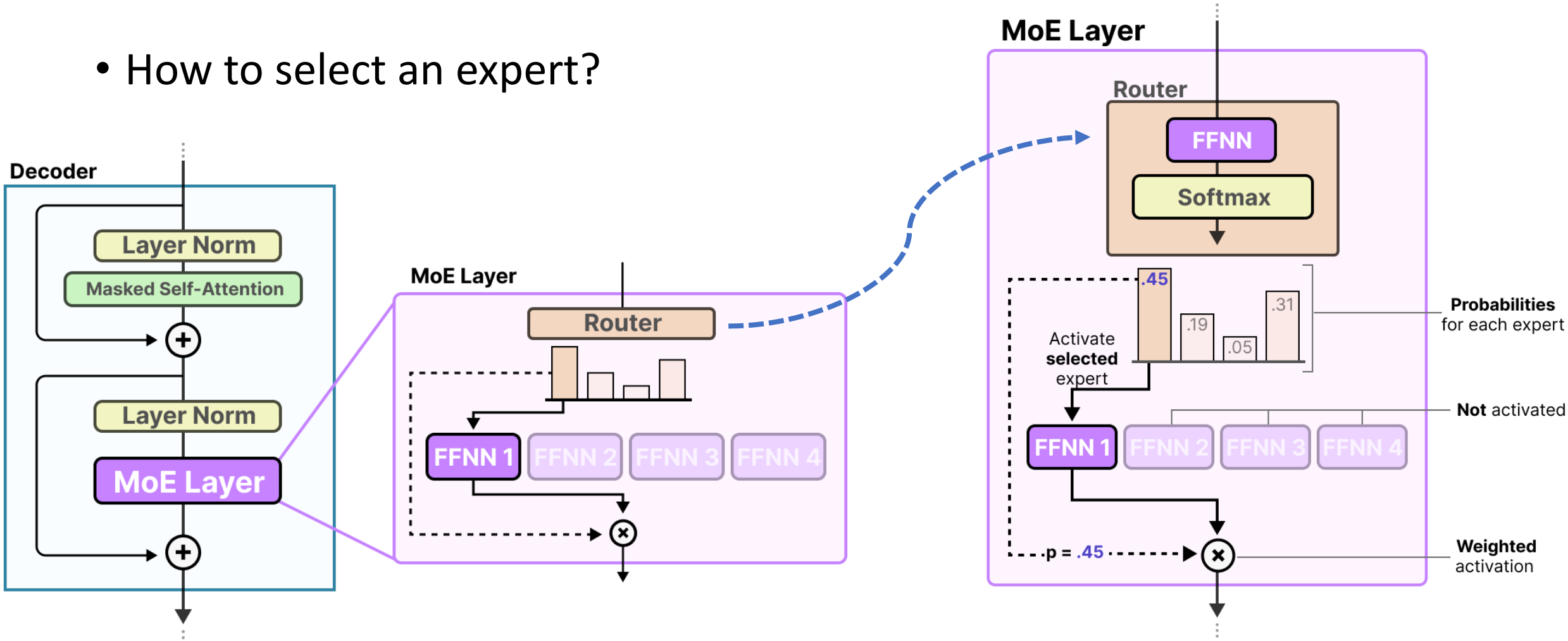
Each expert specializes in different syntactic tokens

More than one decoders → The chosen experts likely differ between tokens, and each token chooses different experts when traversing different encoders → a path exists

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- How to select an expert?



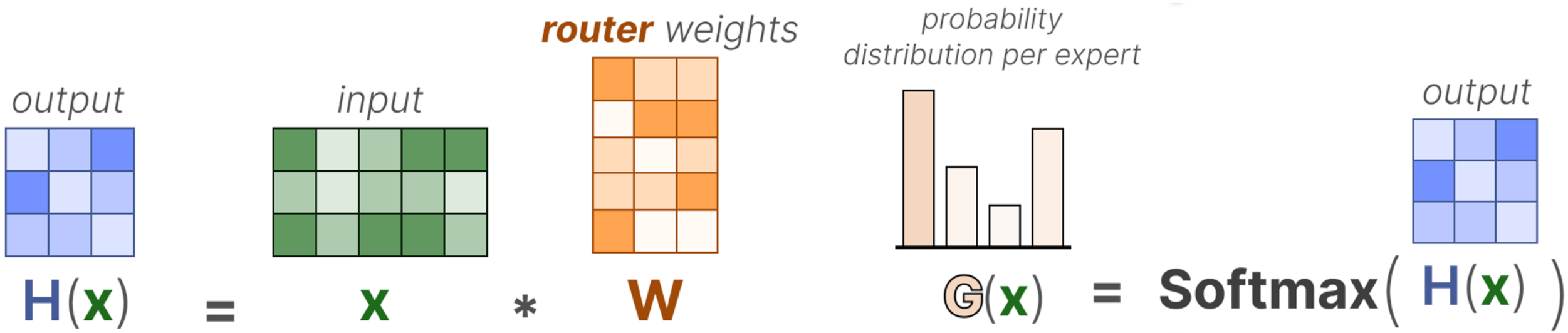
Transformer layer

Determines which tokens are sent to which experts

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- How to select an expert?

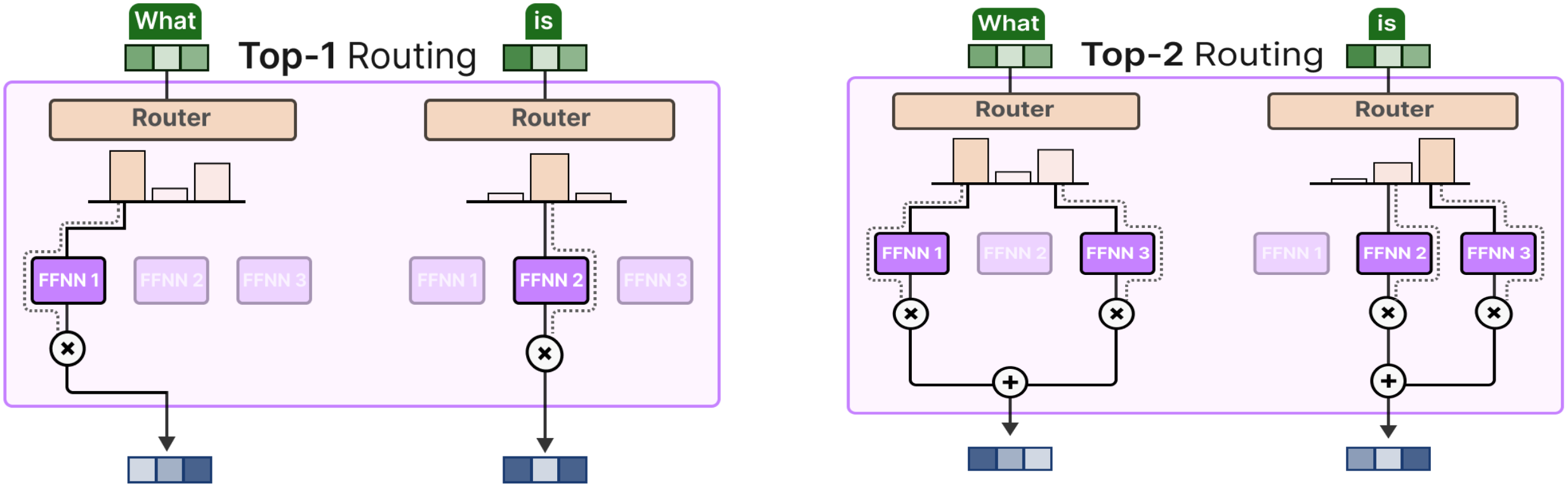


A very basic router

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

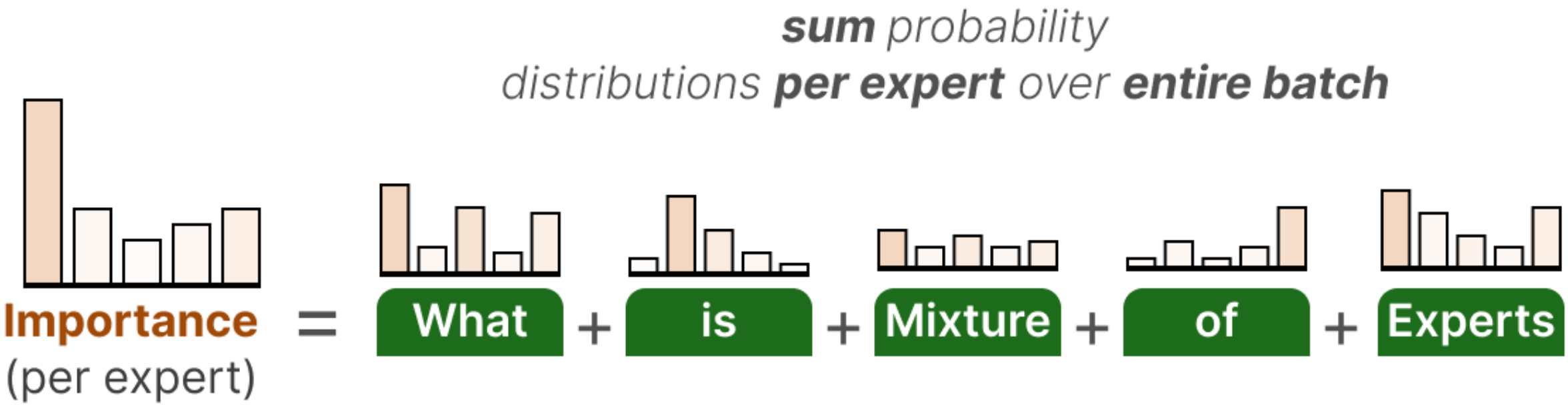
- Top-K routing



《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

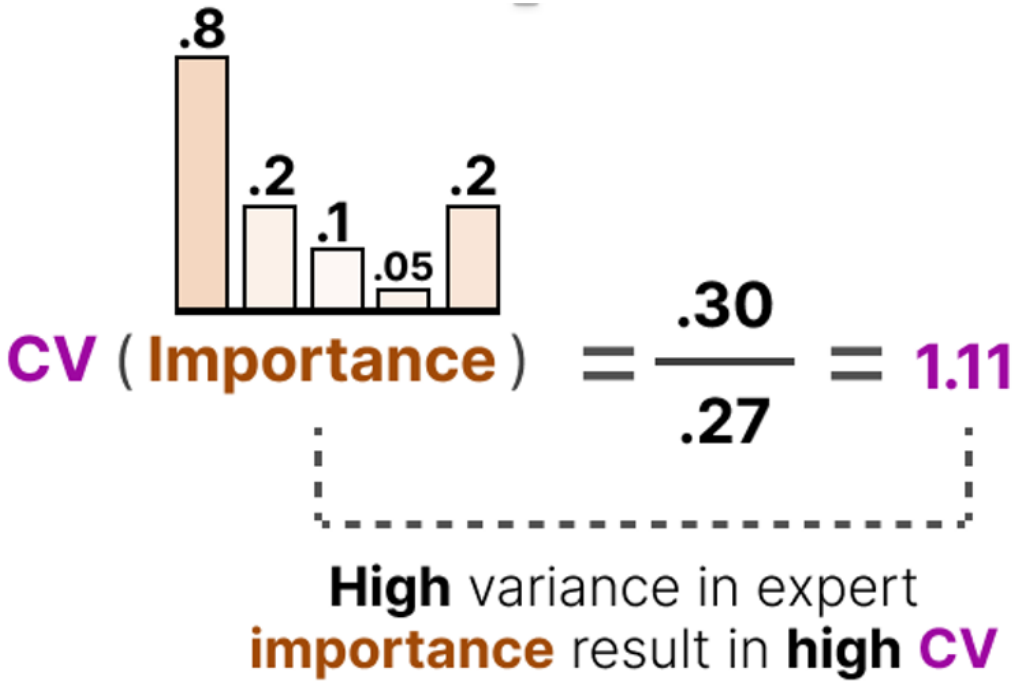
- Load balancing from an algorithmic perspective
  - Quantifying importance of experts



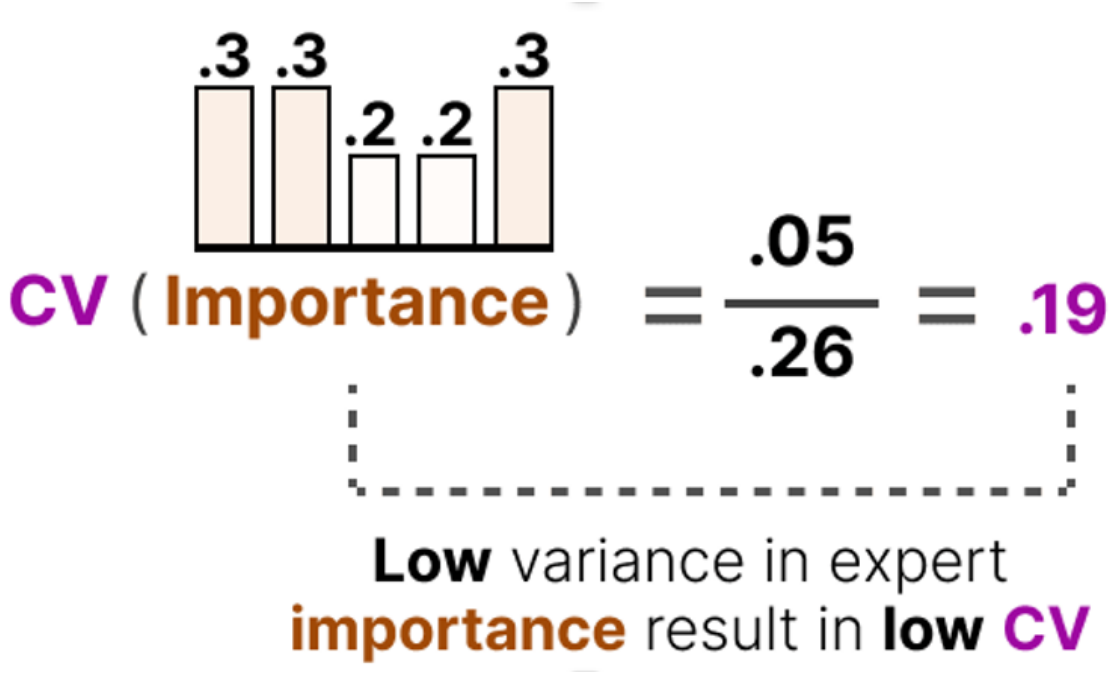
《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Load balancing from an algorithmic perspective
  - Coefficient of variation to capture the diversity across experts



Unbalanced

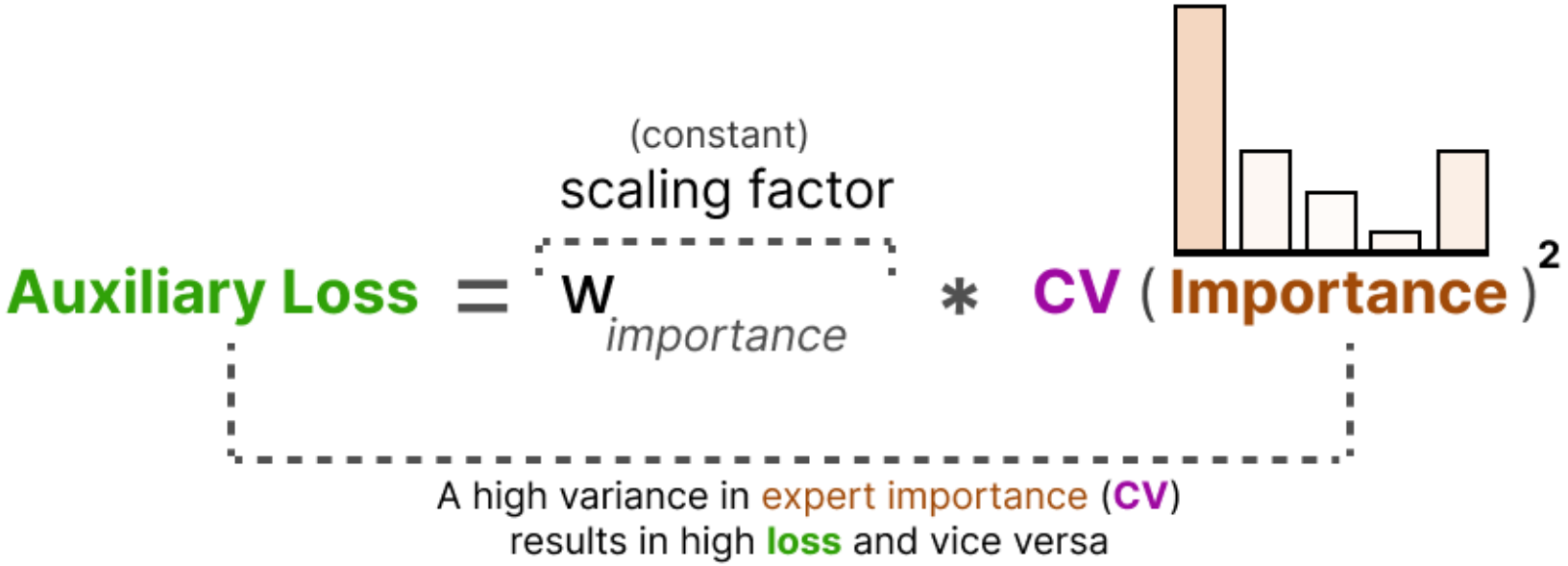


More balanced

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Load balancing from an algorithmic perspective
  - Adding “auxiliary loss” to the overall training objective function

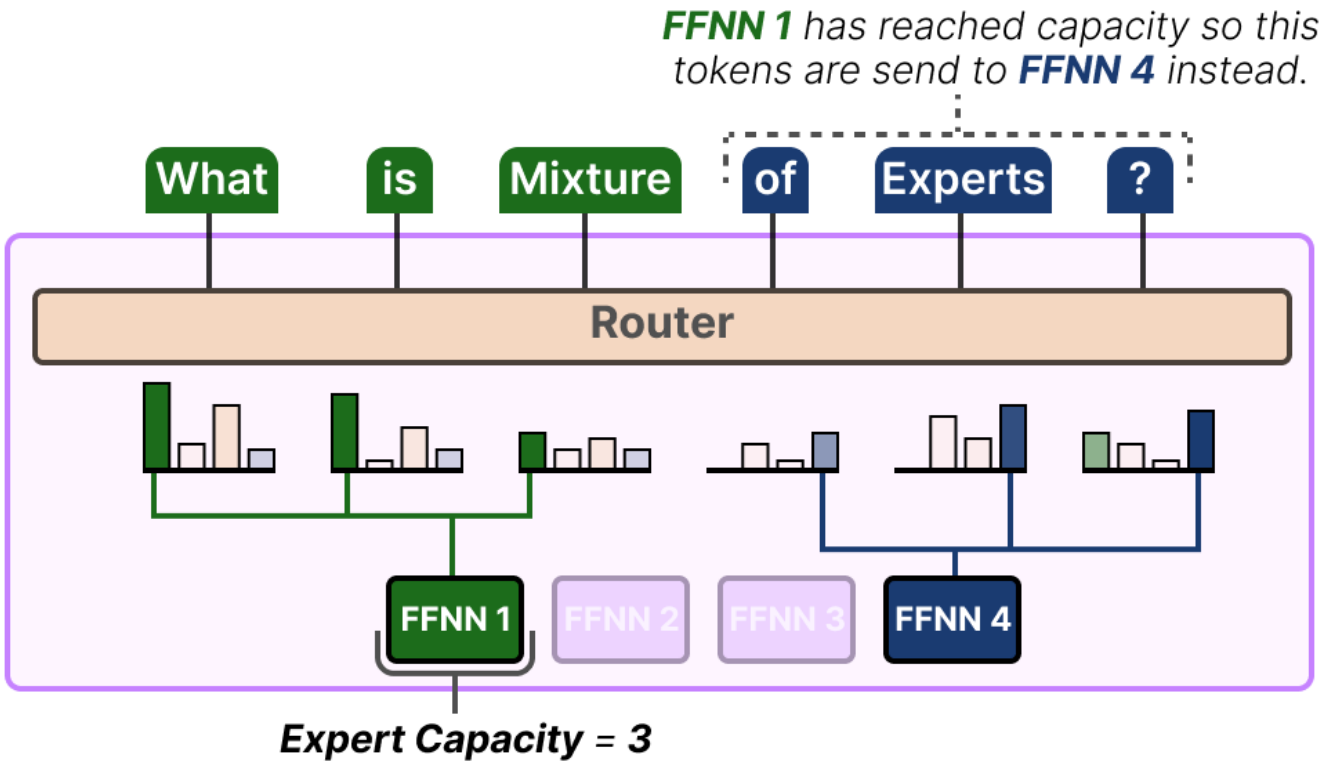


$$l_{aux} = \frac{1}{E}\sum_{e=1}^{E}\frac{c_e}{S} * m_e$$

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Expert Capacity
  - limiting the amount of tokens that a given expert can handle



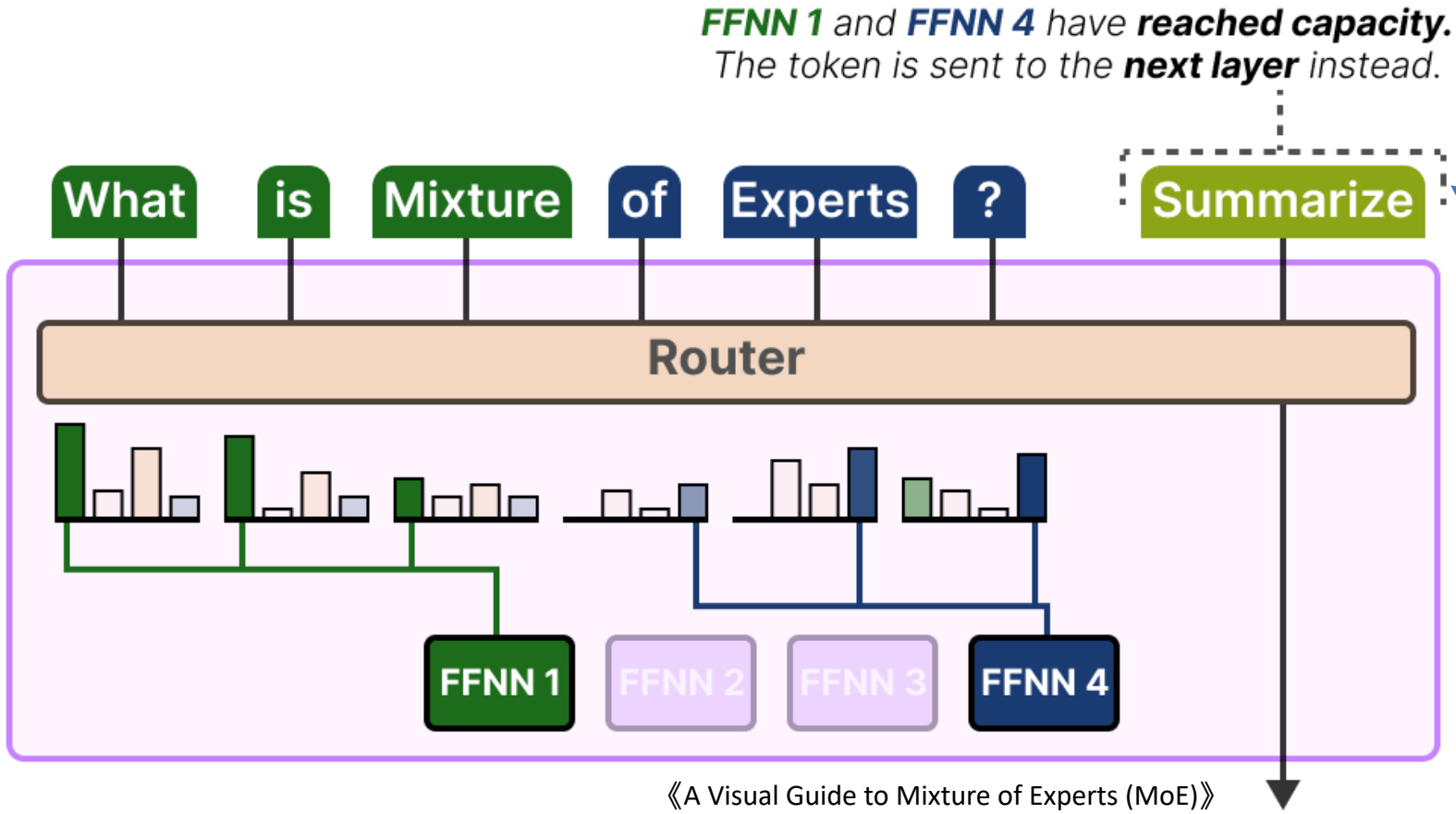
$$capacity = max(\frac{S}{E} * K * capacity\_factor, min\_capacity)$$

- Expert Capacity
  - S: # of tokens
  - E: # of experts
  - K: Tok-K parameter

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Expert Capacity
  - limiting the amount of tokens that a given expert can handle

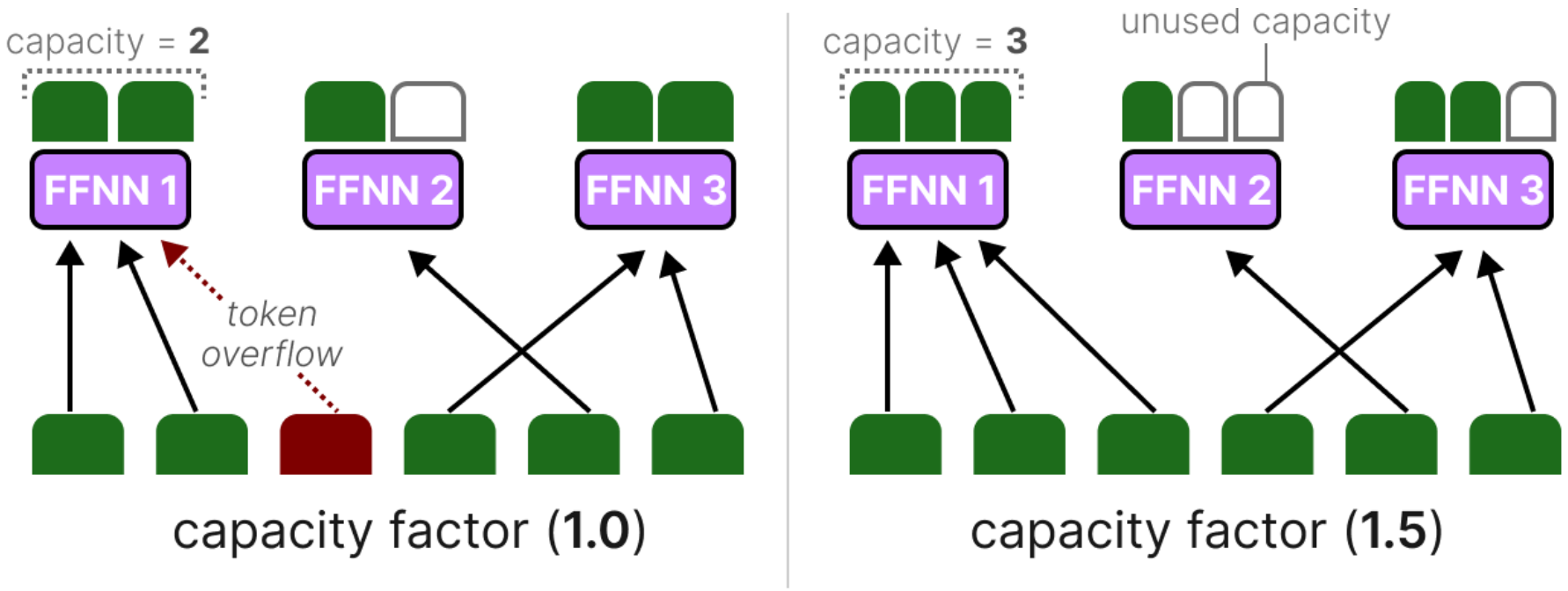


Overflow: not trained by any expert

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Adjusting Expert Capacity

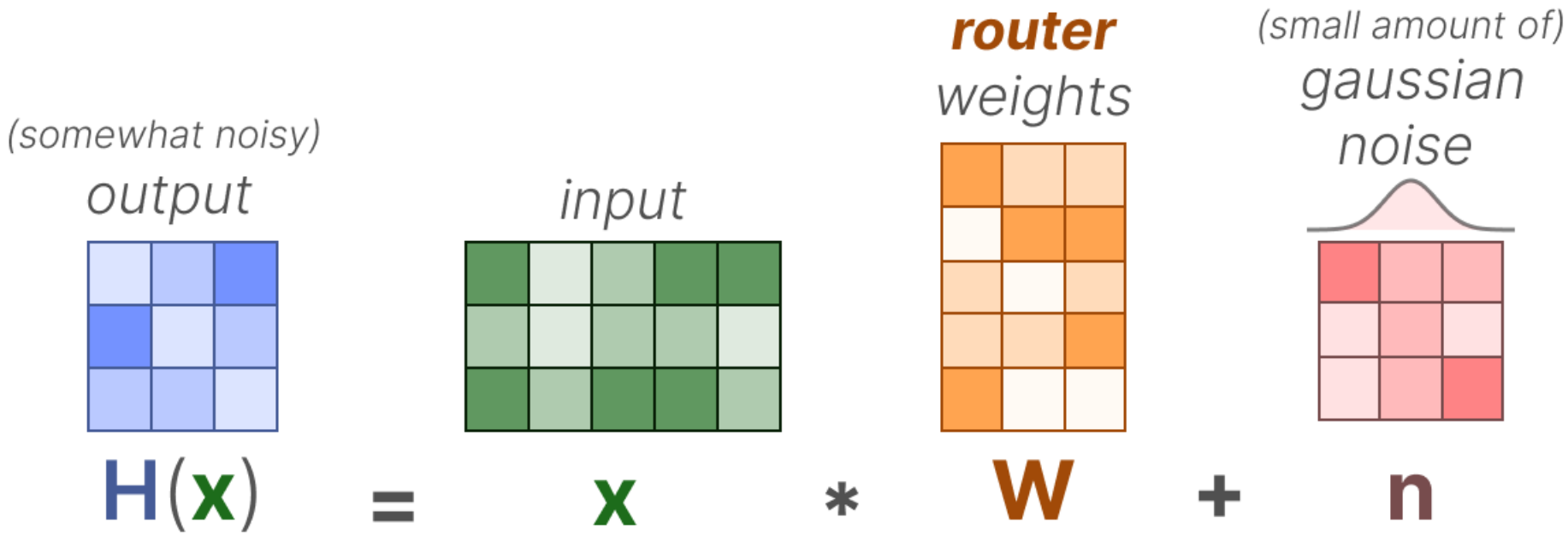


Increasing the capacity factor increases the quality but increases communication costs and memory of activations.

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Random Routing/Noisy Routing

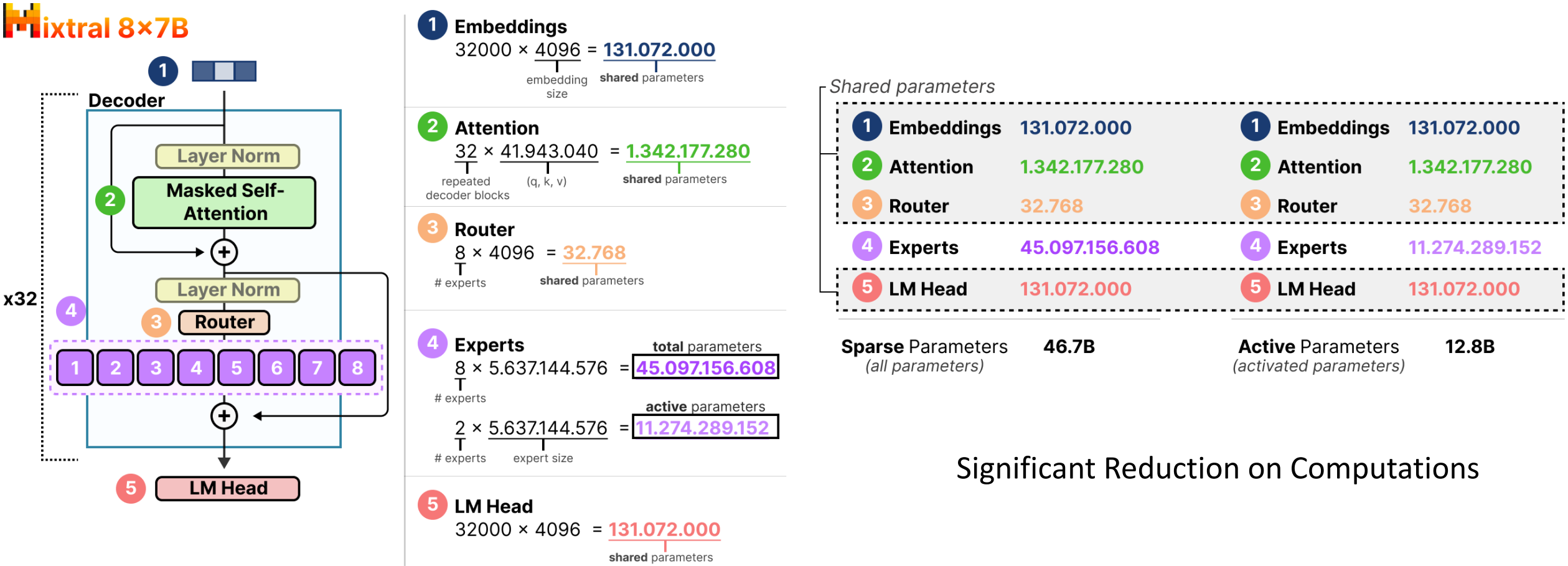


preventing the same experts from always being selected

《A Visual Guide to Mixture of Experts (MoE)》

Mixture of Experts

- Scaling up model parameters



Significant Reduction on Computations

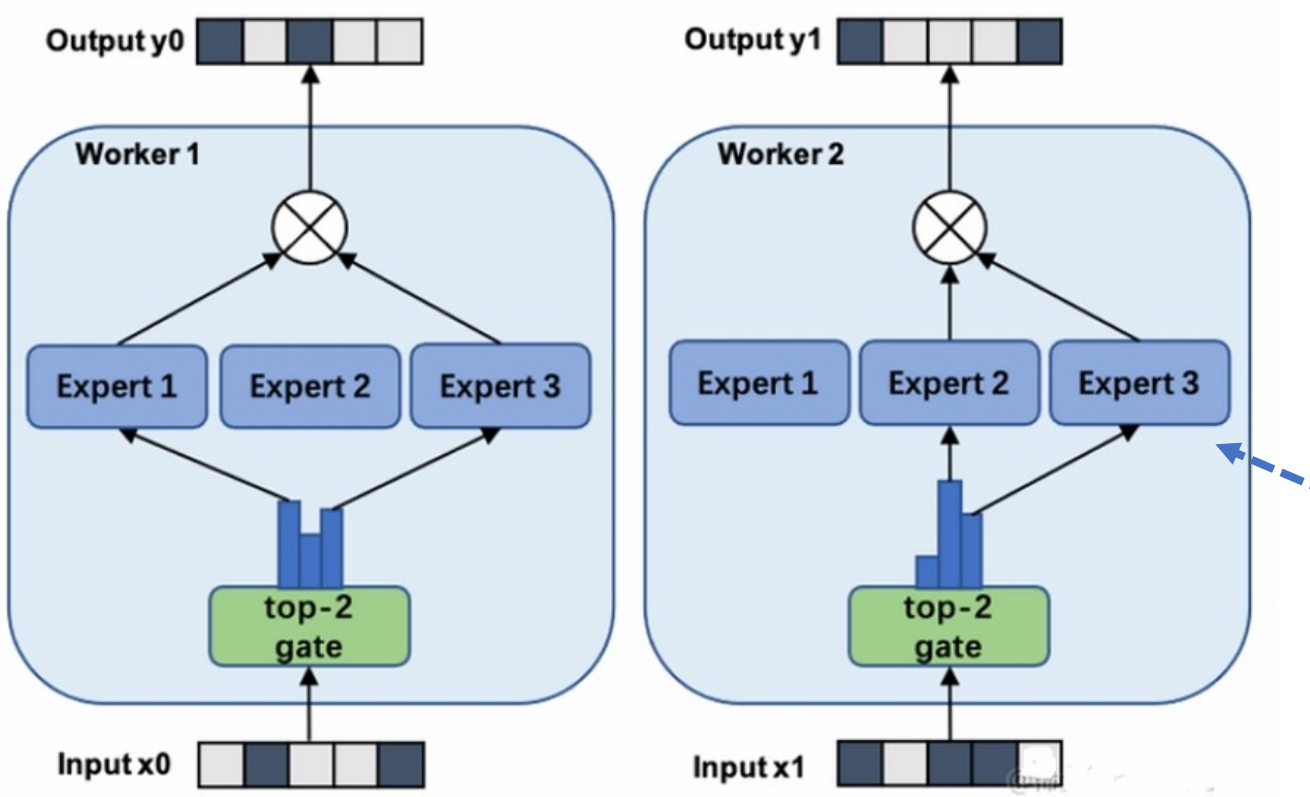
《A Visual Guide to Mixture of Experts (MoE)》

Distributed LLM Training: Outline

- Data Parallelism
- Model Parallelism
- Mixture of Experts
  - Principles
  - **Parallel Training**
  - Recent Progresses

Mixture of Experts

- Parallel Training
  - How to distribute experts on different GPUs?



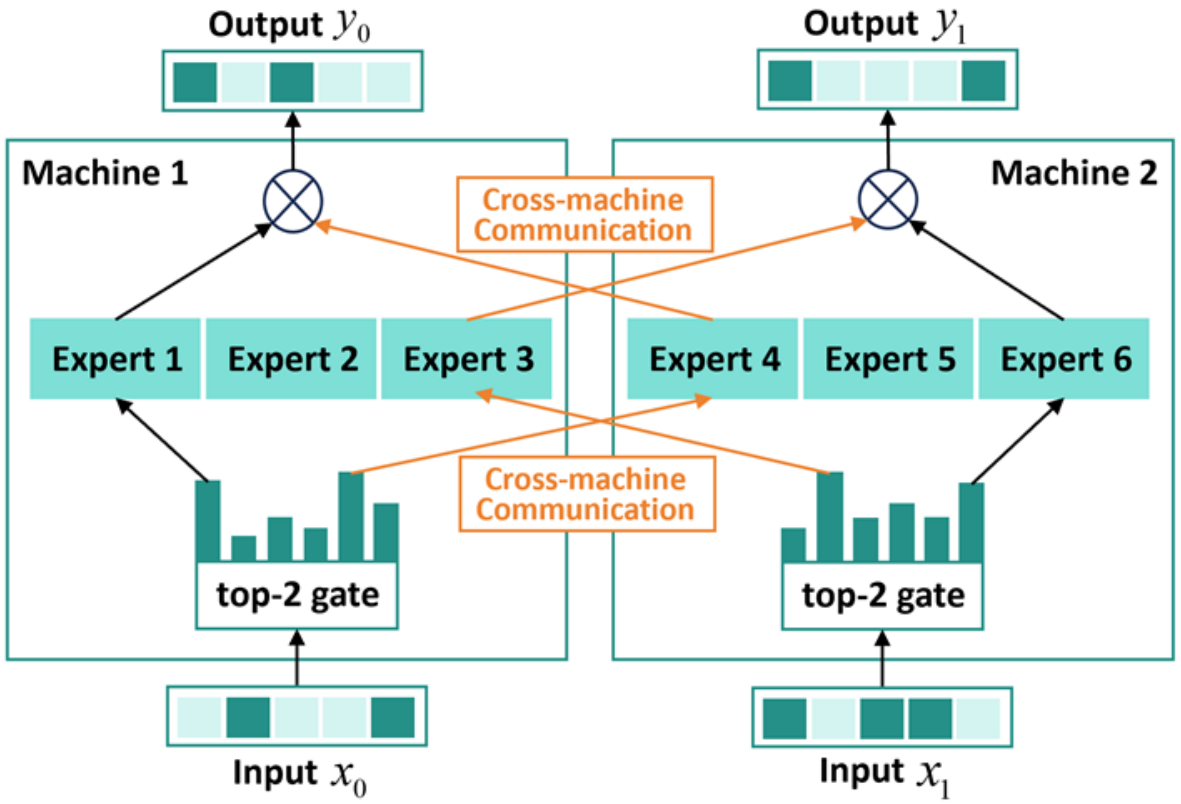
- Data and Expert Parallelism
  - Small # of experts
  - Experts are small in sizes

Every GPU/Machine stores the complete replicas of experts

《FastMoE and Wu Dao 2.0 - behind the scenes with two authors of the recent Chinese AI breakthrough》

Mixture of Experts

- Parallel Training
  - How to distribute experts on different GPUs?

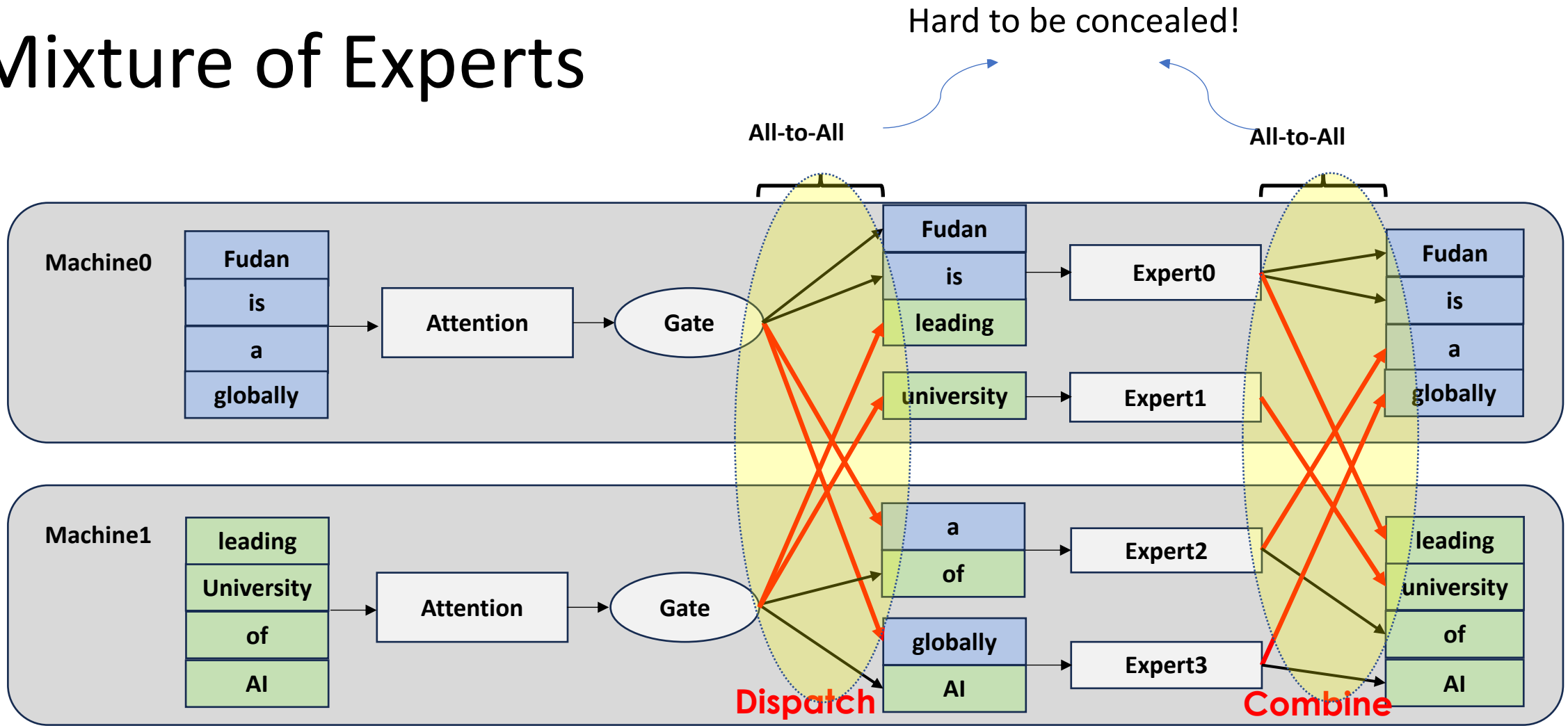


6 experts, DP = 2, EP = 2 (assuming one machine one GPU)

- Data and Expert Parallelism
  - Scaling up to gigantic experts
  - Experts are placed across DP groups
- Size of a "token"
  - *hidden_dim* * 2 Bytes (bf16)
  - Examples
    - QWEN-235B: 2*4096 = 8192 Bytes
    - DeepSeek-v3: 2*7168 = 14336 Bytes
- Traffic volume
  - Proportional to batch size, sequence length, hidden dimension, top-K

Cross-Machine Communication

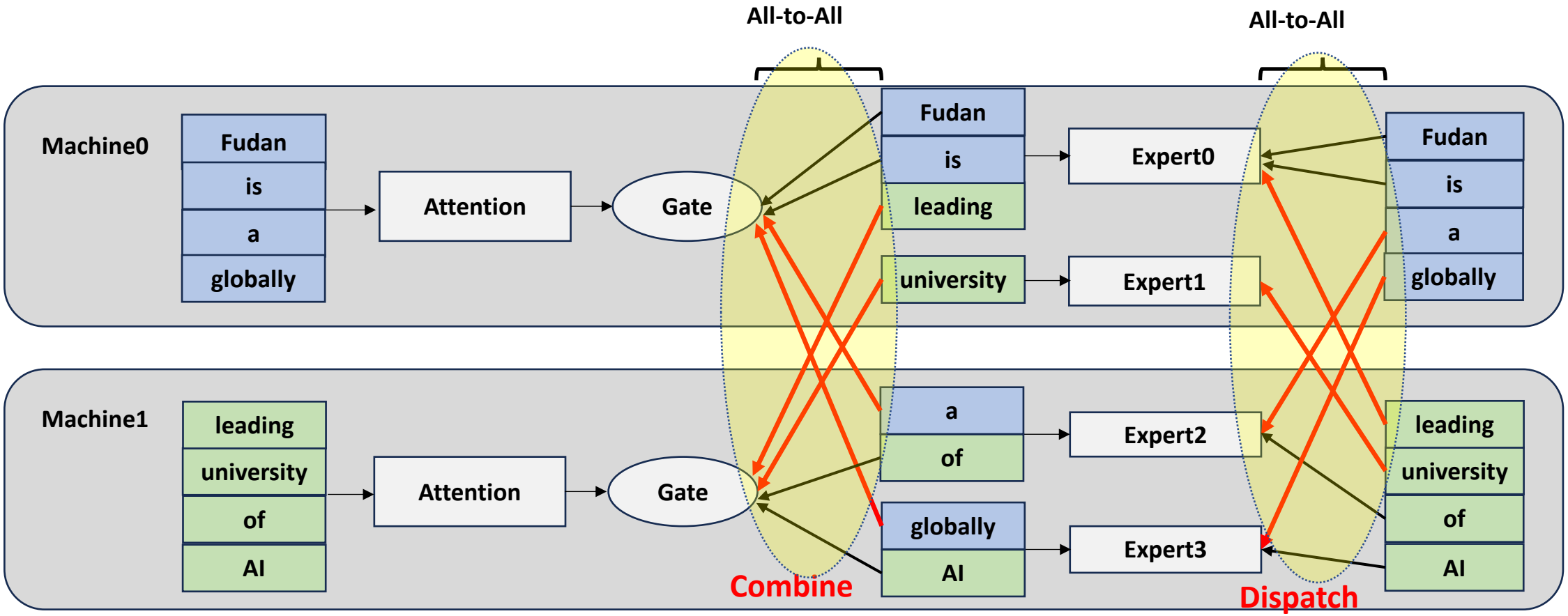
Mixture of Experts



Compute	Communication	Compute	Communication

Forward Pass: heavy token communications that hard to be concealed

Mixture of Experts



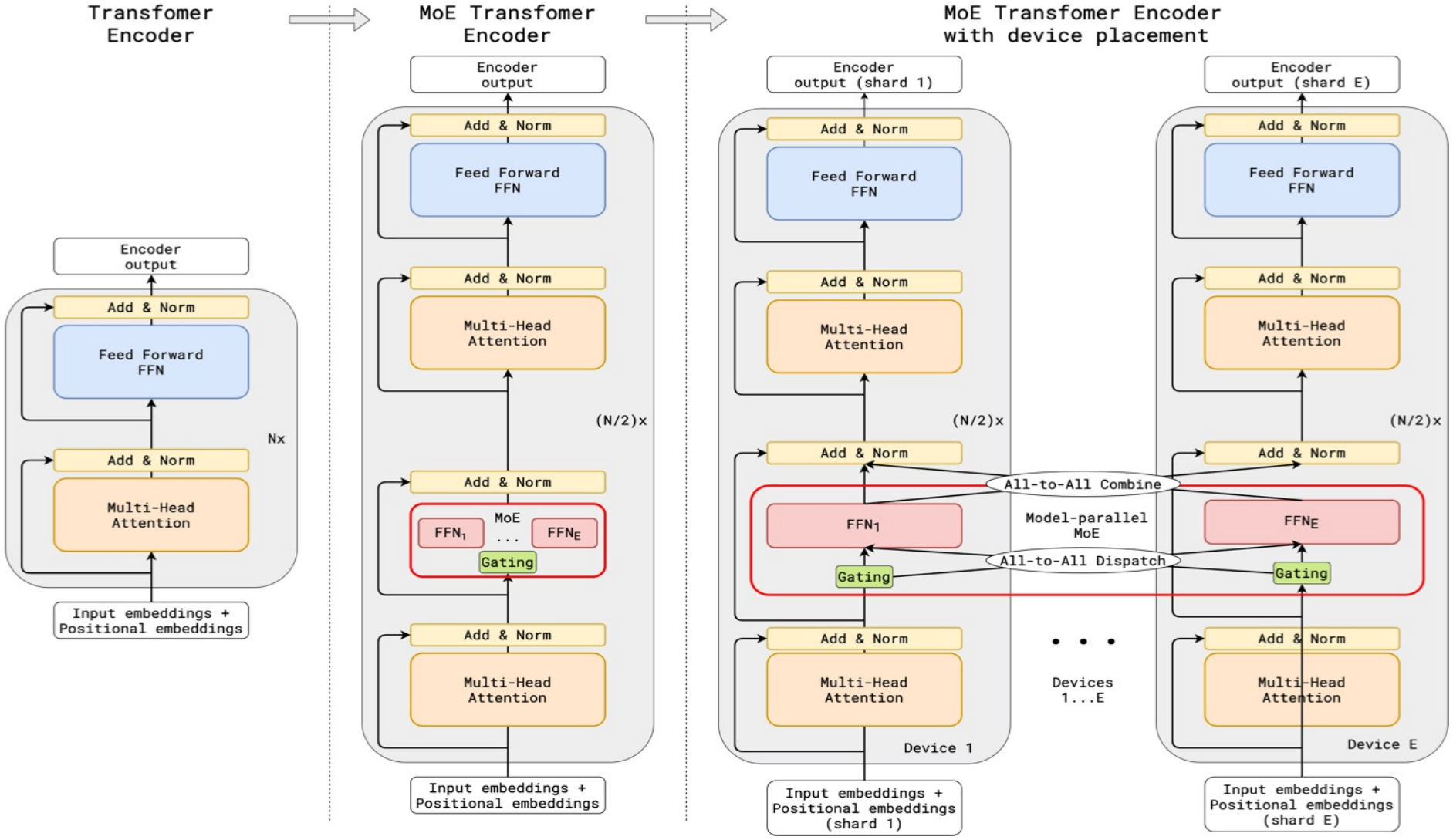
Compute | Communication | Compute | Communication

Backward Pass: heavy token communications that hard to be concealed

Mixture of Experts: Traffic Pattern

Forward: Dispatch and Combine

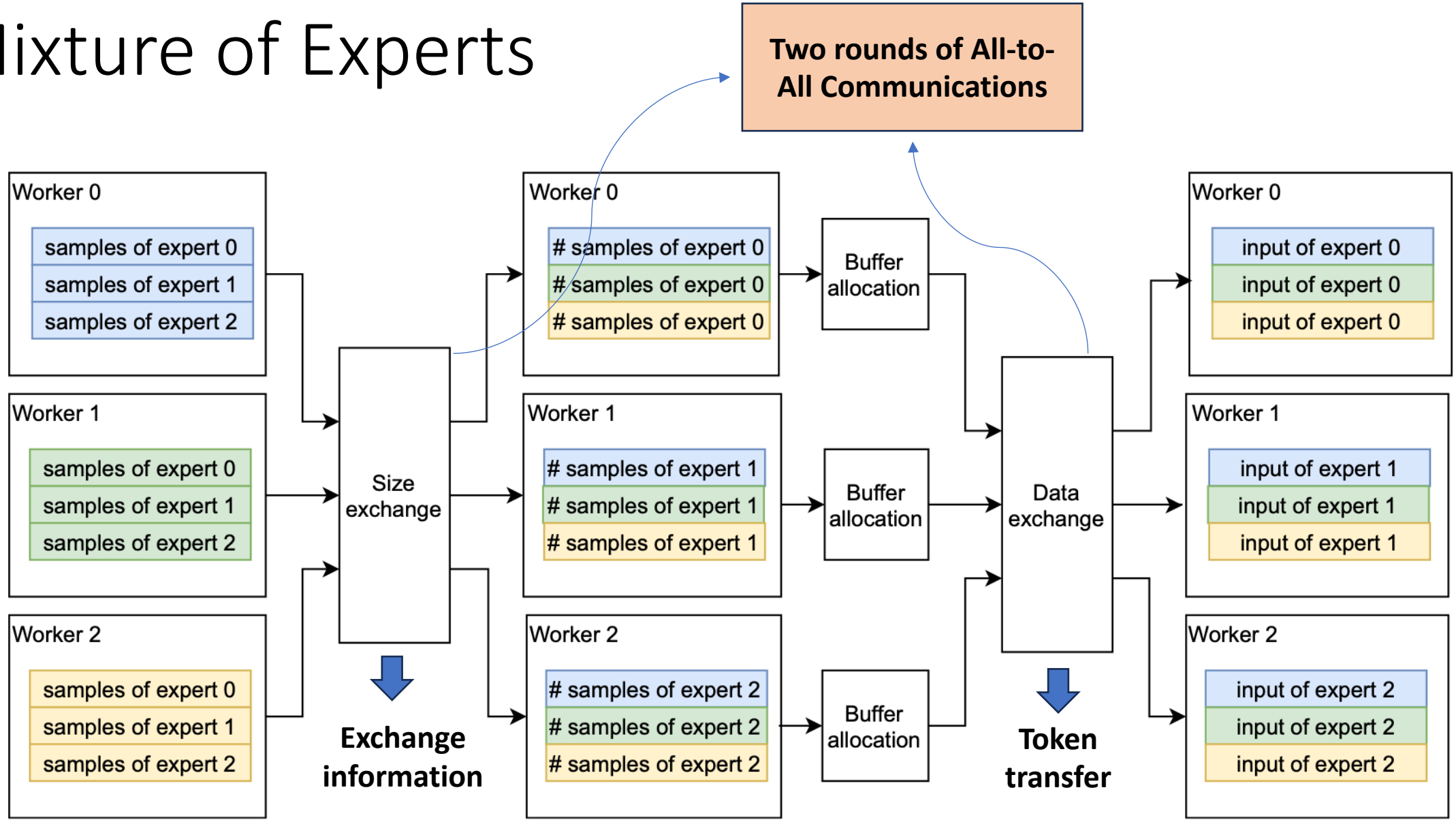
Backward: Dispatch and Combine



GShard (2020)

- MoE within a single machine: forward and backward
- MoE spanning multiple machines: all-to-all communications

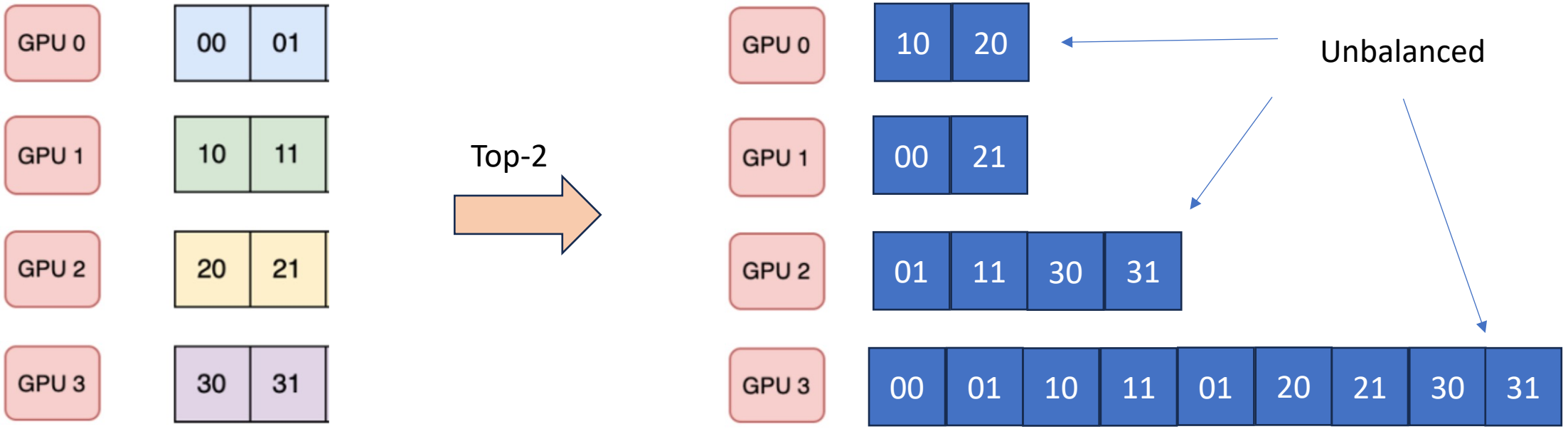
Mixture of Experts



FastMoE (2021): First MoE training system that supports Pytorch

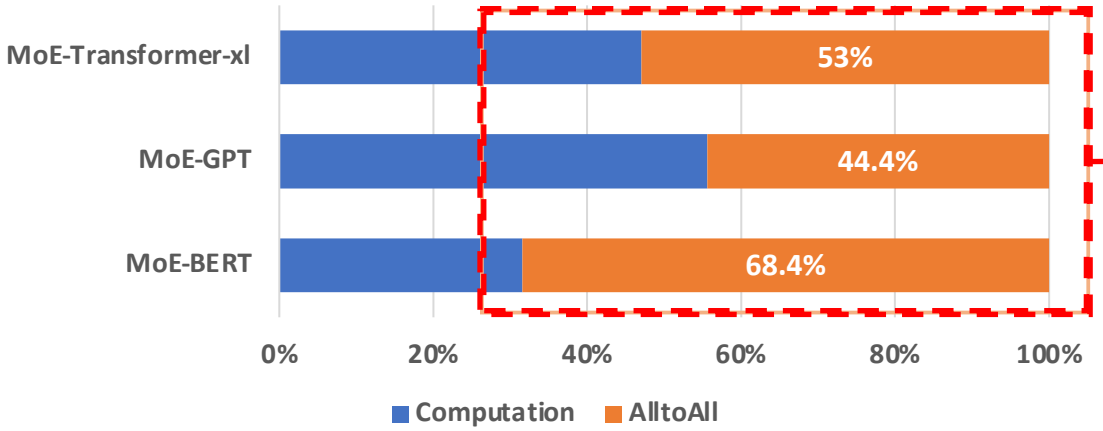
Mixture of Experts

- All-to-All Communication
  - Asymmetric data volume (tokens) between GPUs hosting experts
    - **ZERO padding** to transmit ***dummy*** tokens under standard All-to-All?



Mixture of Experts

- Challenge in MoE All-to-All



iteration time vs **time of All-to-All** (32 GPUs)

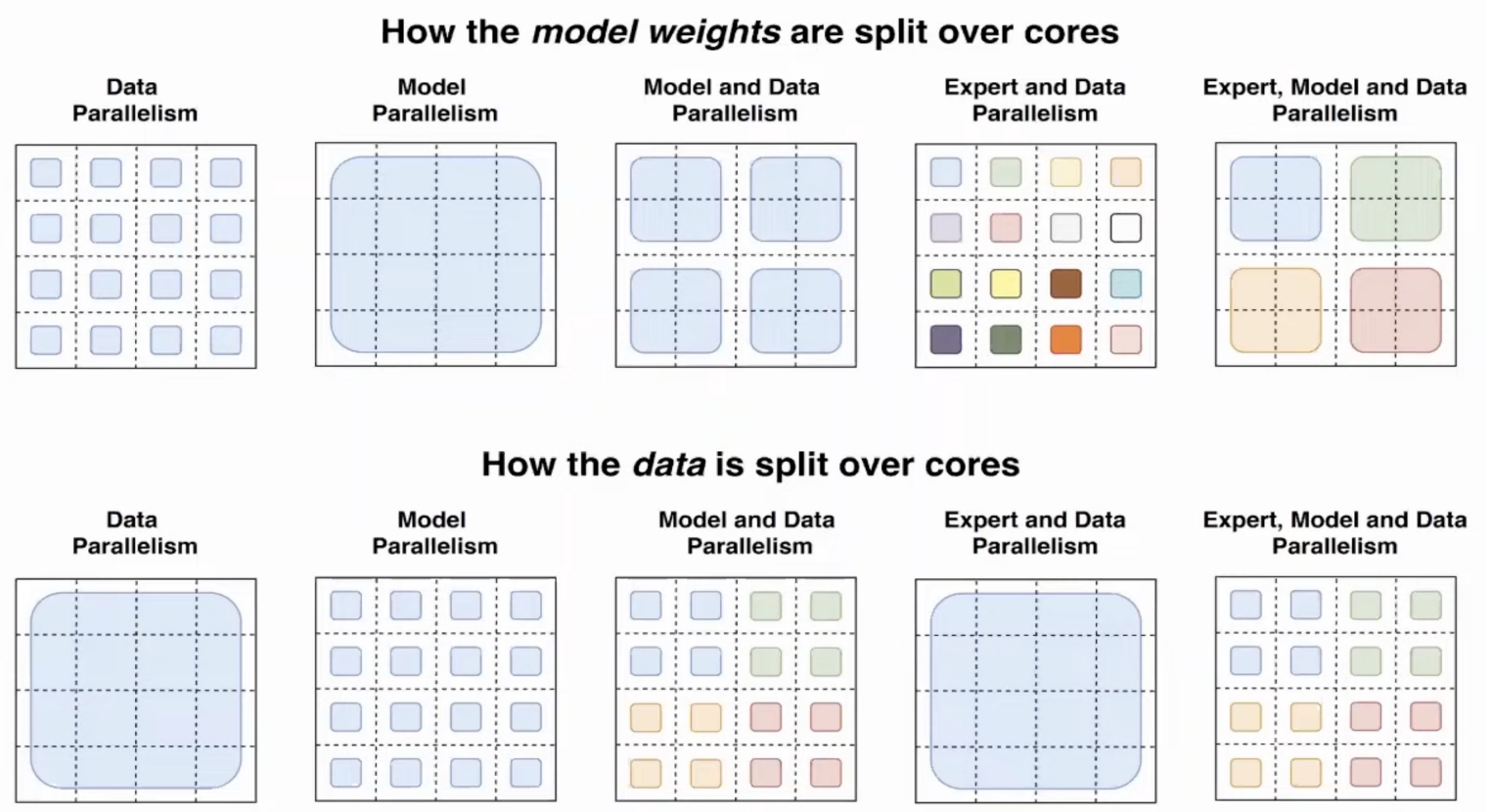
All-to-All communication accounts for a significant proportion (44.4% ~ 68.4%),

Distributed LLM Training: Outline

- Data Parallelism
- Model Parallelism
- Mixture of Experts
  - Principles
  - Parallel Training
  - **Recent Progresses (Extended Learning)**

Hybrid Parallelism

- Segmenting data, model and activation



Source: Transforming Natural Language Understanding: Attention, GPT, BERT, and Switch

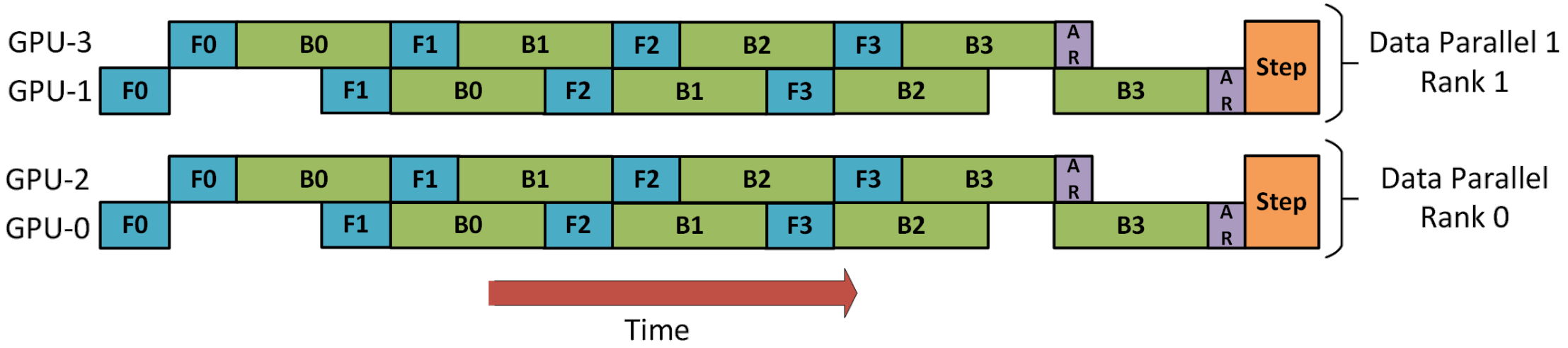
A Brief Summary

- Hybrid parallelism (classical models)

Models	**TP**	**DP**	**PP**	**Micro batchsize**	**global batchsize**	**sequence length**	**ZeRO**	GPUs	GPU types
Bloom-176B	4	8	12	2	2048	2048	ZeRO-1	384	A100-80GB
Megatron-175B	8	\	16	1	1536	2048	\	\	\
OPT-175B	8	124	\	\	\	2048	\	996	A100-80GB
GPT3-175B	\	\	\	\	\	2048	\	\	\
Megatron-NLG 530B	8	16	35	\	1960	2048	\	4480	A100-80GB
GLM-130B	4	24	8	\	\	4096	ZeRO-1	768	A100-80GB

A Brief Summary

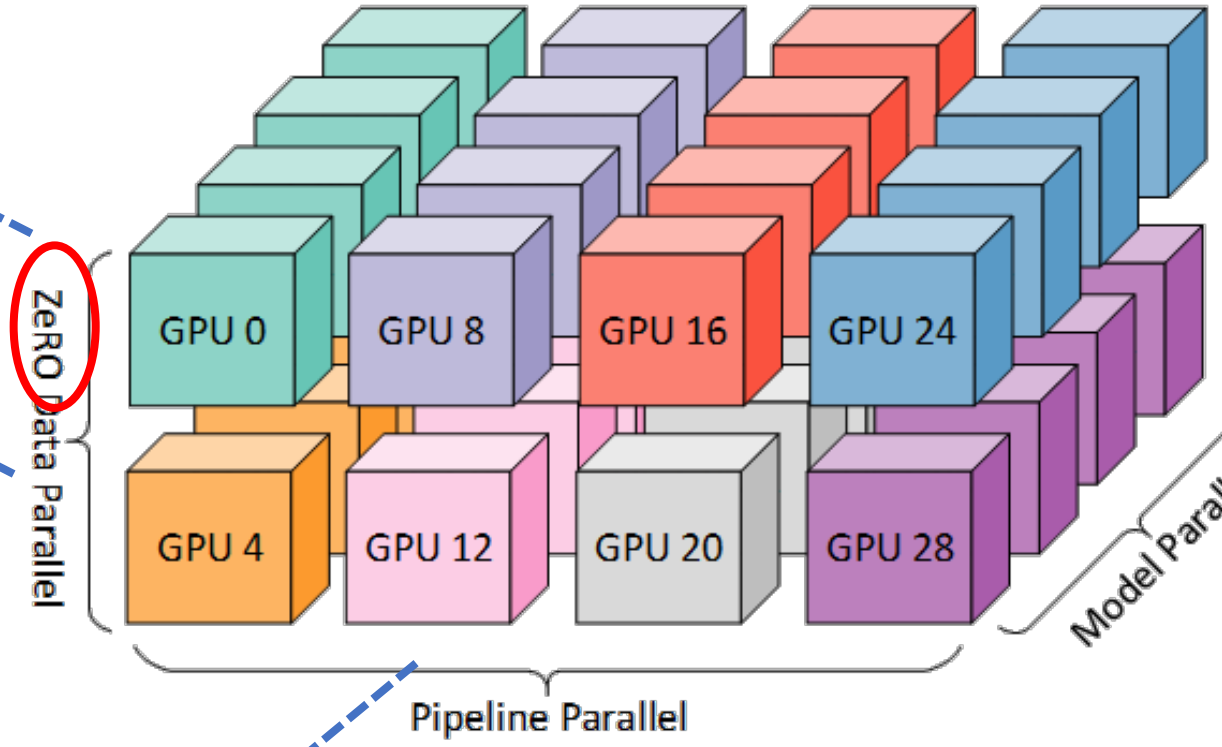
- Data Parallelism + Pipeline Parallelism



A Brief Summary

- Data Parallelism + Pipeline Parallelism + Tensor Parallelism + ZeRO

How to implement ZeRO-3 with pipeline?



Relatively lower bandwidth (e.g. RDMA)

High bandwidth domain (e.g. NVLink, HCCS)

Intra-datacenter RDMA or even Inter-datacenter links

Thanks!