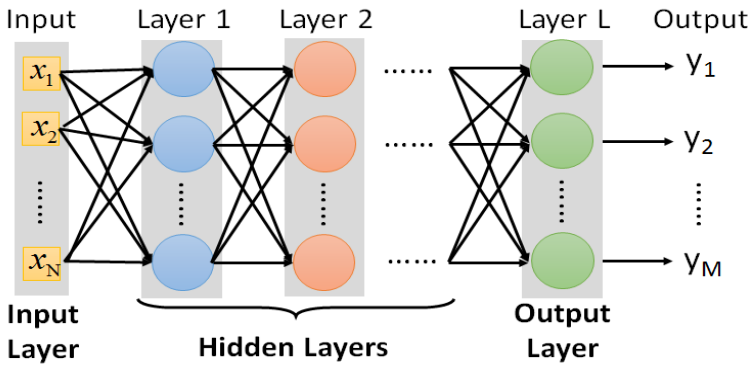
Mixture of Experts (MoE) in AICN

Lily Lyu

September Interim 2025

Transformer-based LLM (Large Language Model)

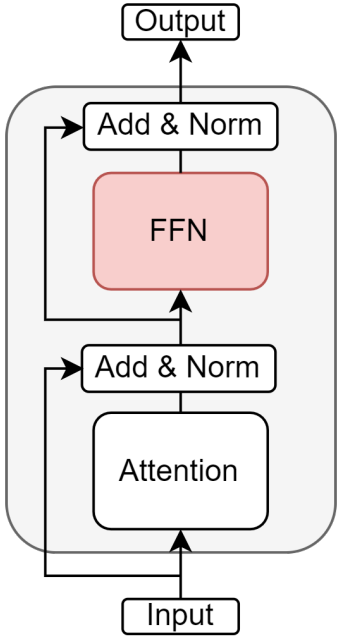
Basic concept of DNN (Deep Neural Network)



Neural network with multiple layers of parameters (weights + biases)

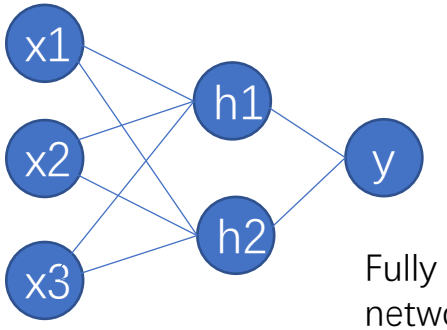
Y=X*W+B

Transformer： a kind of DNN architecture



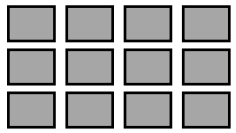
The major components in one layer are Attention and FFN(Feed-Forward Neural Network) .

FFN is a fully connected neural network, which is also called **dense neural network.**



Fully connected neural network (3-2-1)

DeepSeek Igniting MoE

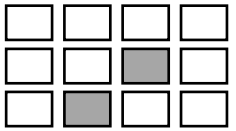


Dense Model

- Transformer-based LLMs dominated the field.
- Scaling relied mainly on increasing parameters in a dense way.

Turning Point

DeepSeek



Sparse Model (MoE)

- Only part of the model is activated per token.
- Enables efficient scaling, better inference throughput, and energy savings.

DeepSeek marked the transition from purely dense scaling to practical sparse architectures, making MoE a mainstream direction in AI model development.

2023

DeepSeek-LLM
Dense 67B (Parameter) model

2024 ↓ 90% cost

DeepSeek-V2
Hybrid MoE, 236B total / 21B active

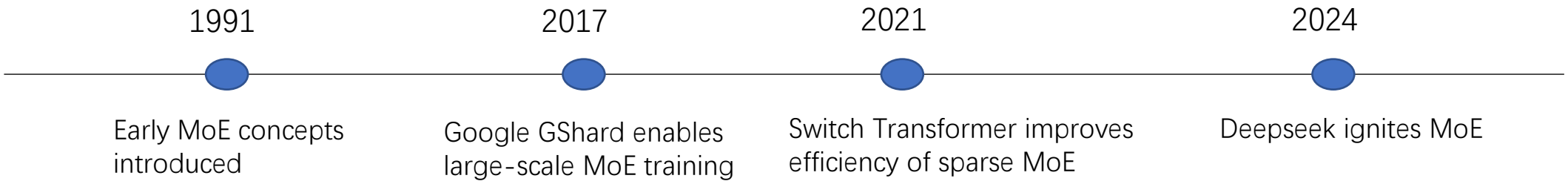
DeepSeek-V3
Full MoE, 671B total / 37B active

2025 ↓ 95% cost

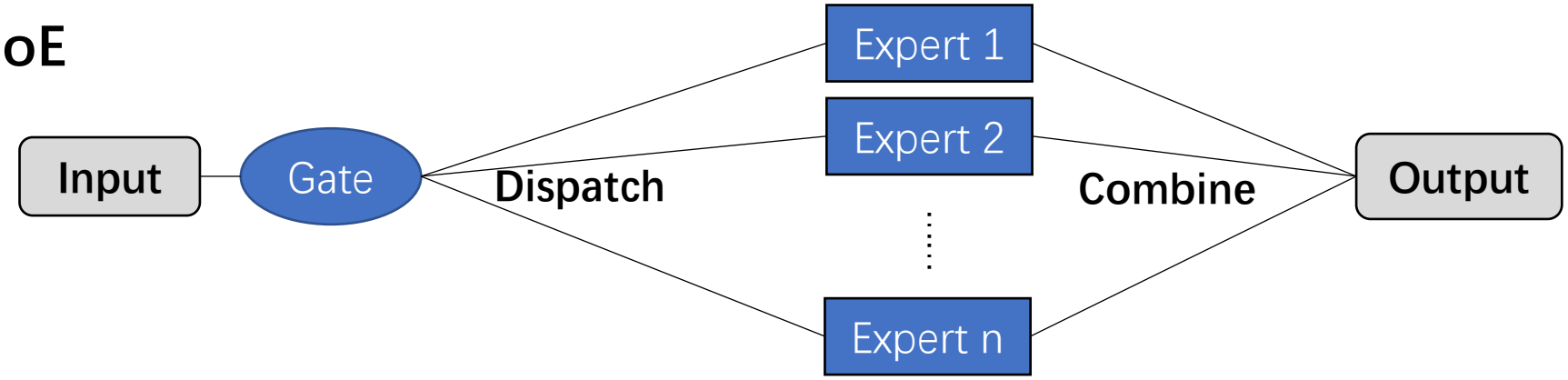
DeepSeek-R1
V3-based, Reasoning-focused MoE

MoE is Not Brand New

History of MoE development



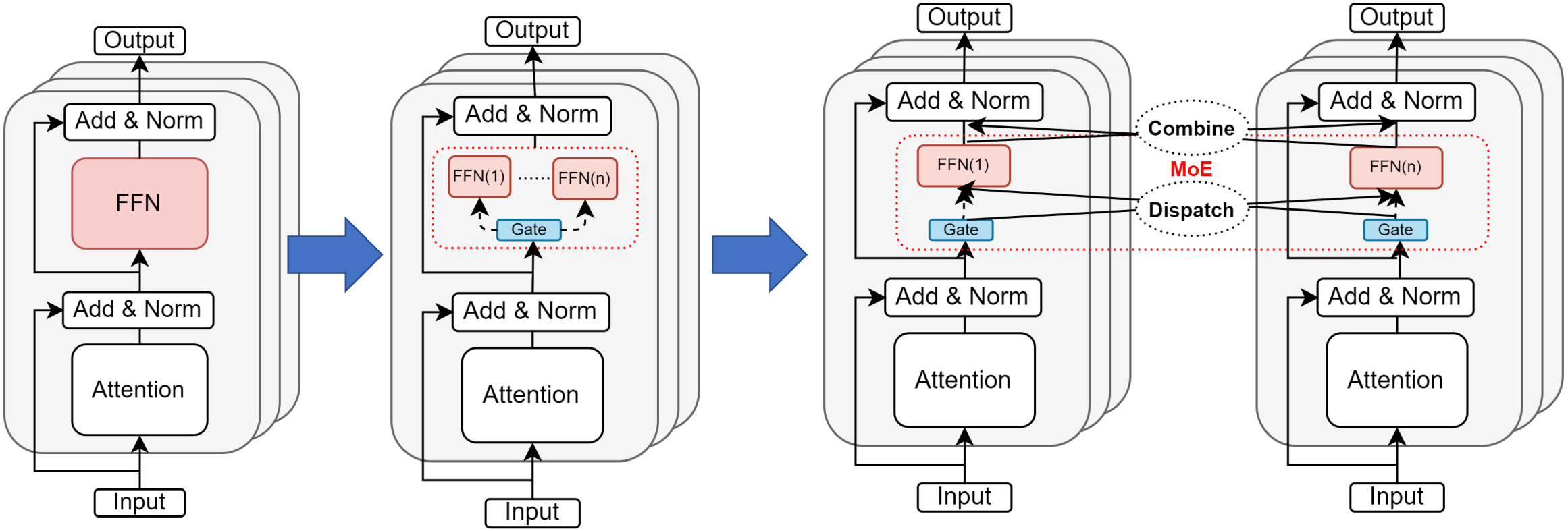
Core principle of MoE



- **Expert:** Specialized sub-models that handle different tasks
- **Gate:** Chooses the most relevant experts (e.g., Top-1 or Top-2) for each input
- **Dispatch:** Only a small subset of experts are active per input, saving computation
- **Combine:** Selected experts' outputs are aggregated to form the final result

MoE in Transformer

In practice, sparse-gated MoE is usually applied inside Transformers. Instead of one huge feed-forward network, it breaks it down into many smaller ones, and the gate decides which of them to activate.

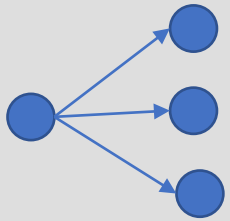


All-to-All Communication in MoE

Dispatch

- Gate decides Top-k experts per token
- Tokens routed to experts (Top-K)
- Local → direct memory copy
- Remote → prepared for transfer

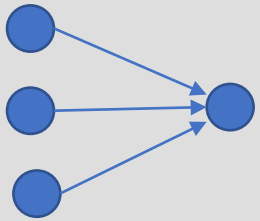
1 to many communication



Combine

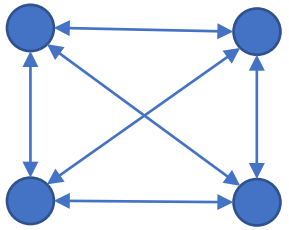
- Gather outputs from experts
- Merge results for each token
- Local results → direct memory copy
- Remote results → received from remote

Many to 1 communication

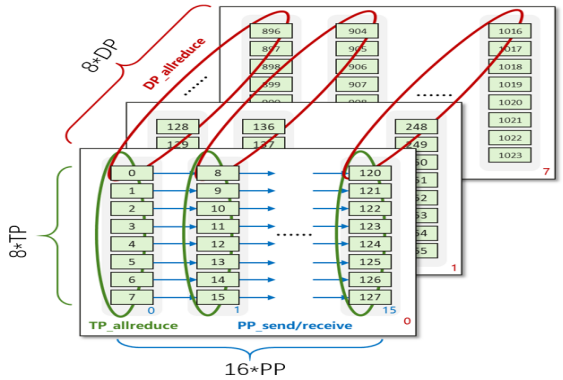


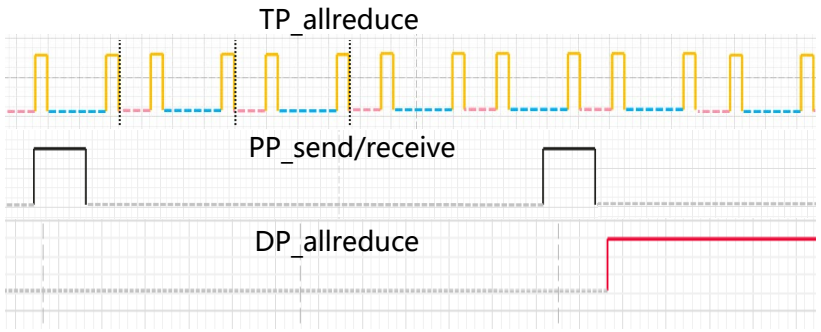
All-to-All is the major communication pattern in MoE

- Each rank has many tokens
- Tokens target experts across multiple ranks
- Aggregate communication → each rank sends to all other ranks
- Forms All-to-All pattern



New Traffic Characteristics

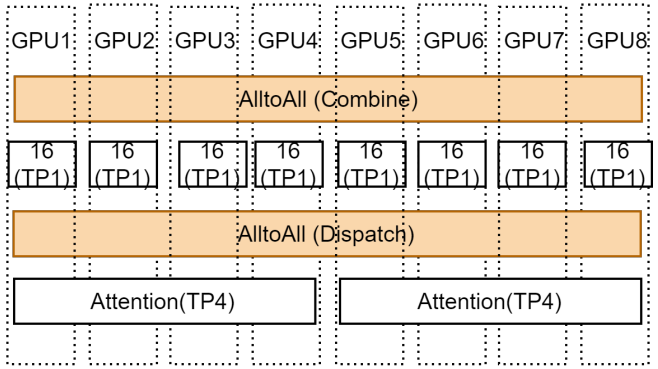


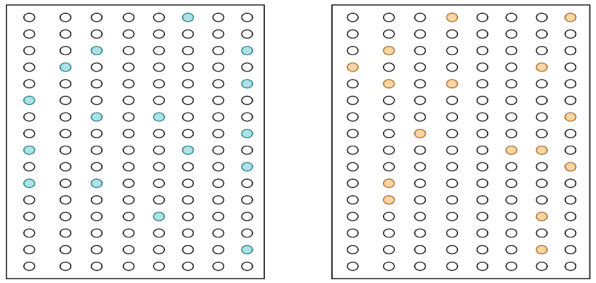


Example: 1024 GPUs, TP8, DP8, PP16

Dense Model

- Predictable communication pairs
- Very few connections
- Large message (MB/GB)





T1

T2

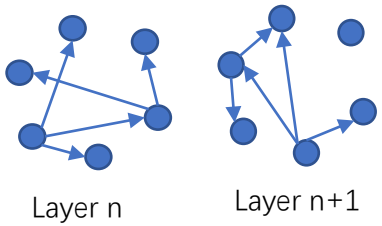
Example: 8GPUs, 128 experts, TopK=16

MoE Model

- Unpredictable communication pairs (depending on MoE TopK algorithm)
- A lot of connections
- Small message (KB)

Challenges for Network

Randomness：Communication pairs vary at the 100’us level



- Each layer has 2 all-to-all communication
- Each all-to-all communication is with small amount of data (KB)
- Communication pairs changes frequently

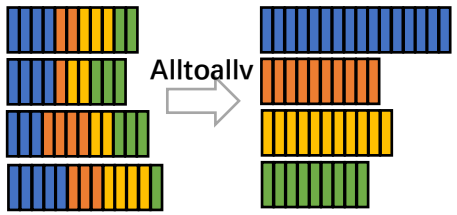
Challenges:

- Pre-defined scheduling does not work
- Hotspot ‘experts’ may increase incast/outcast issues

Network Requirements:

- Flexible resource allocation to adapt to dynamic traffic flows

Asymmetry: Unequal data transfers across GPUs



- Each GPU sends/receives different number of tokens (because experts are selected token by token)
- Typical communication is Alltoallv

Challenges:

- Fairness-based policies for each flow is not effective
- Tail latency is highly variable and non-deterministic

Network Requirements:

- Provide deterministic behavior for tail latency

Latency Sensitive: Small messages dominate the traffic

Challenges:

- The overhead of RDMA/RoCEv2 becomes unignorable
- Static latency of forwarding may matter

Network Requirements:

- Capable to efficiently handle small packet
  - low latency, low overhead

Thanks