



**CONFERENCE**

—NORTH AMERICA 2026—

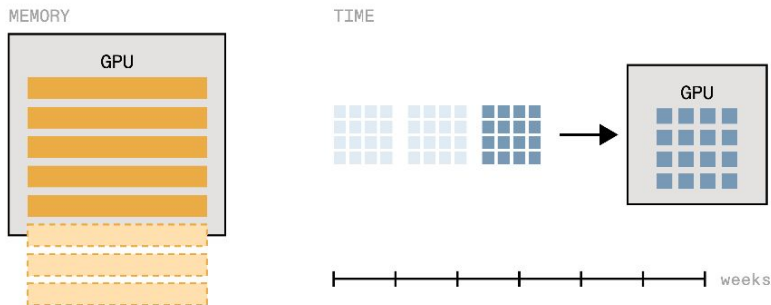
# Fitting Large MoE Fine-Tuning onto Few Nodes:

## Expert & Context Parallelism in DeepSpeed

Suman Debnath, Jiale Huang, Daniel Shen

# Foundation for Training Parallelism

## 1 · WHY PARALLELISM

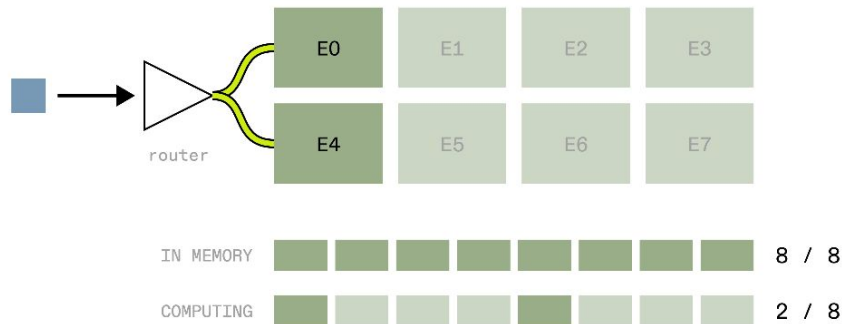


The model does not fit

One GPU is too slow

Memory and time force training onto many GPUs

## 2 · MoE MODELS



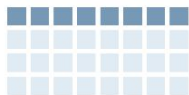
Only 5–25% of the parameters are active per token, all of them stay in memory

For MoE, memory is the constraint

# Parallelism Schemes at a Glance

**DP** Data parallel

DATA on GPU 0



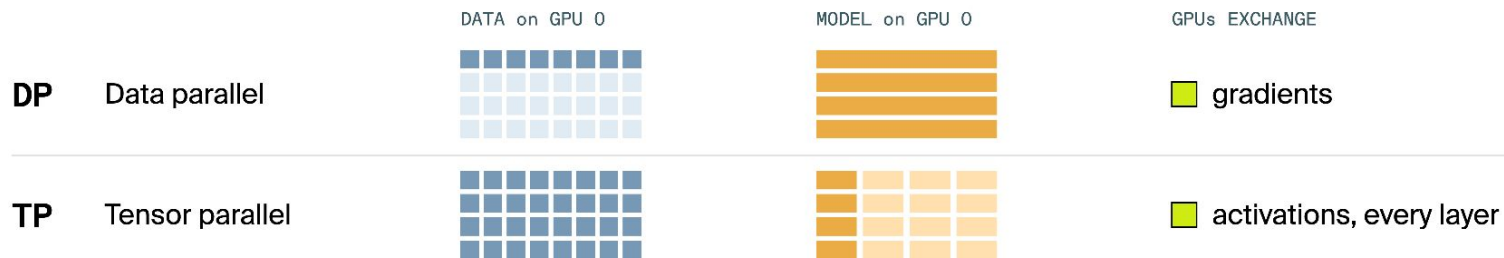
MODEL on GPU 0



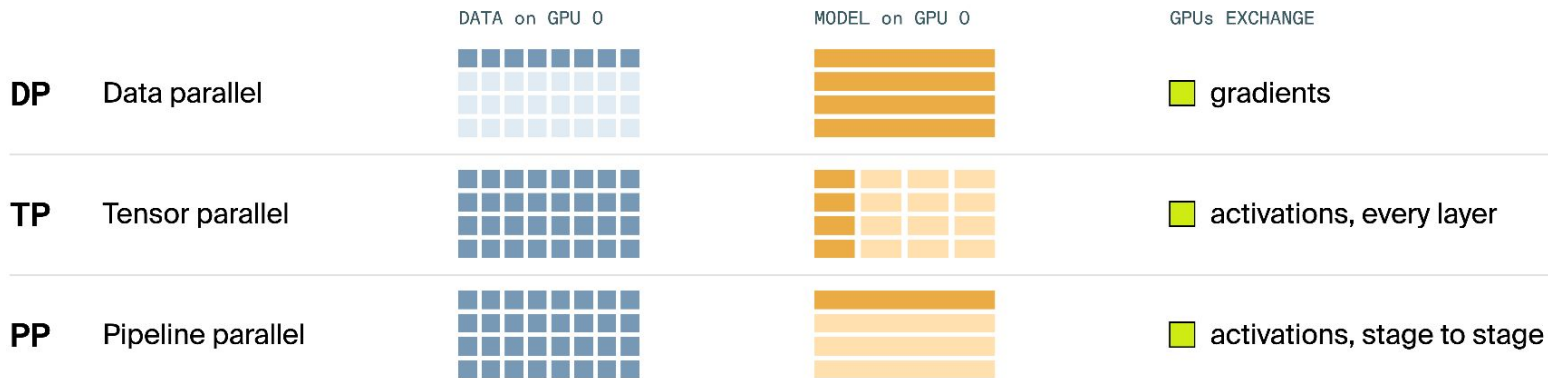
GPUs EXCHANGE

 gradients

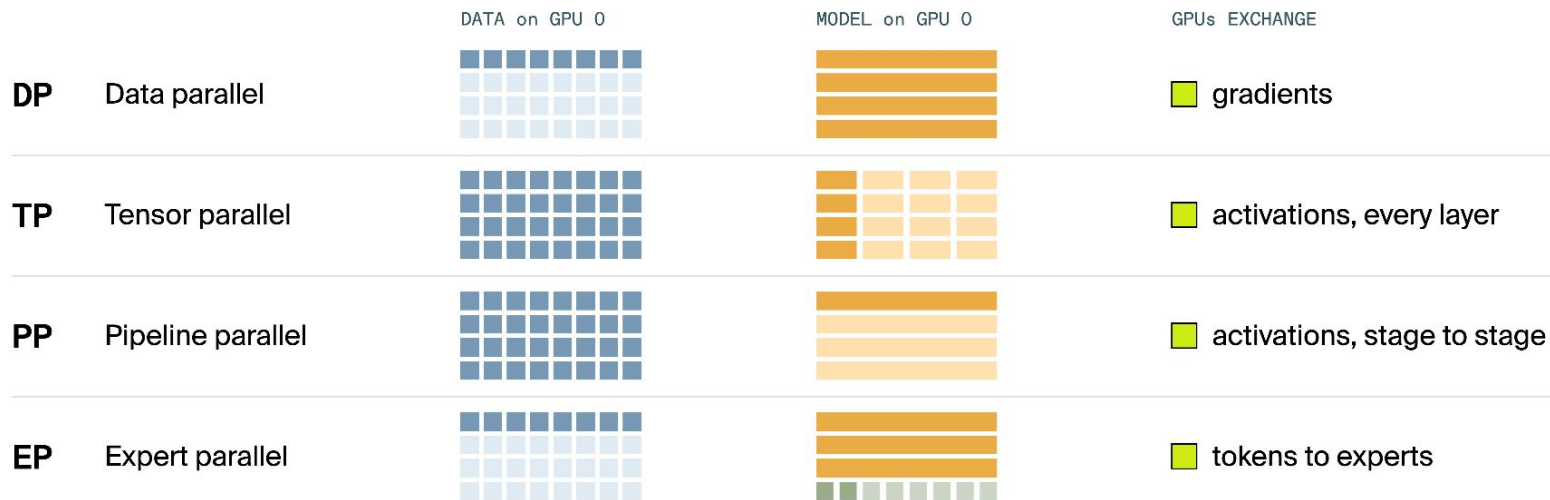
# Parallelism Schemes at a Glance



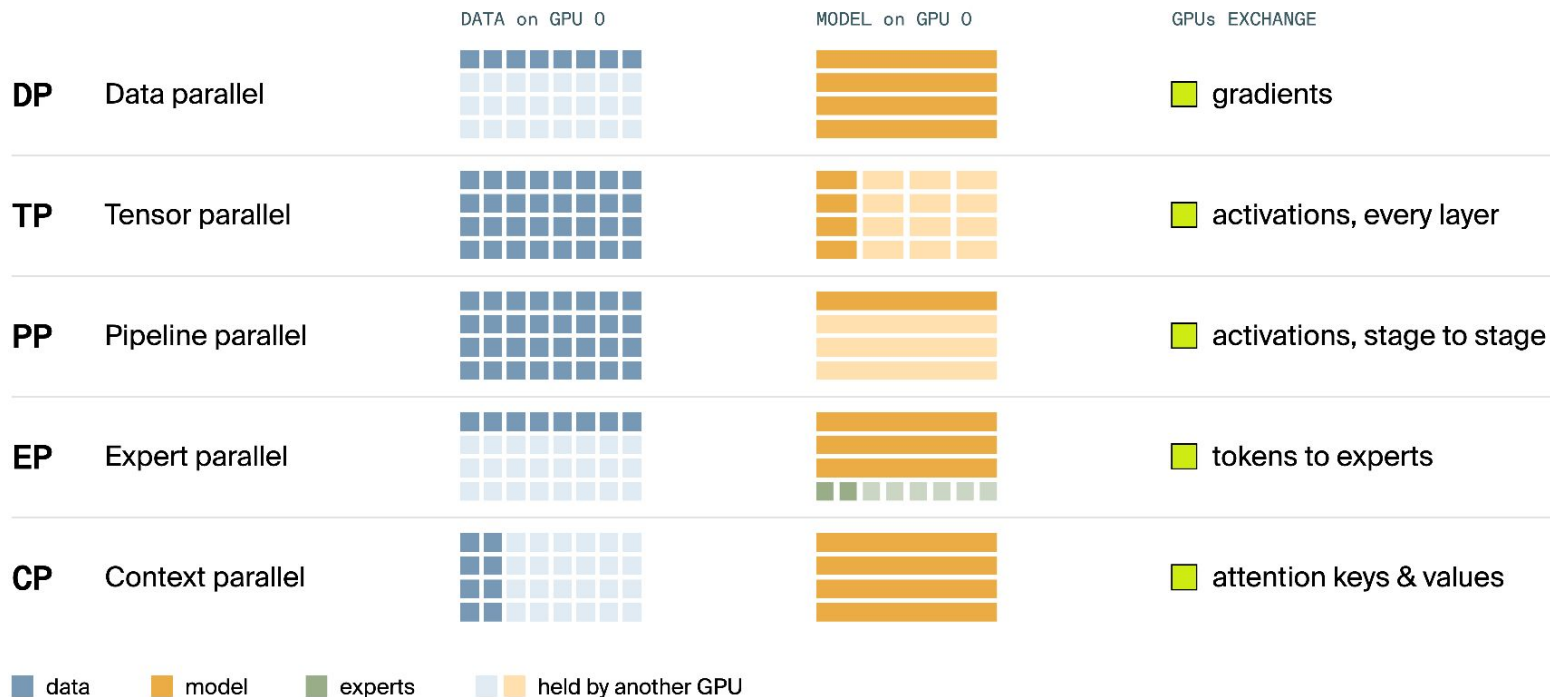
# Parallelism Schemes at a Glance



# Parallelism Schemes at a Glance

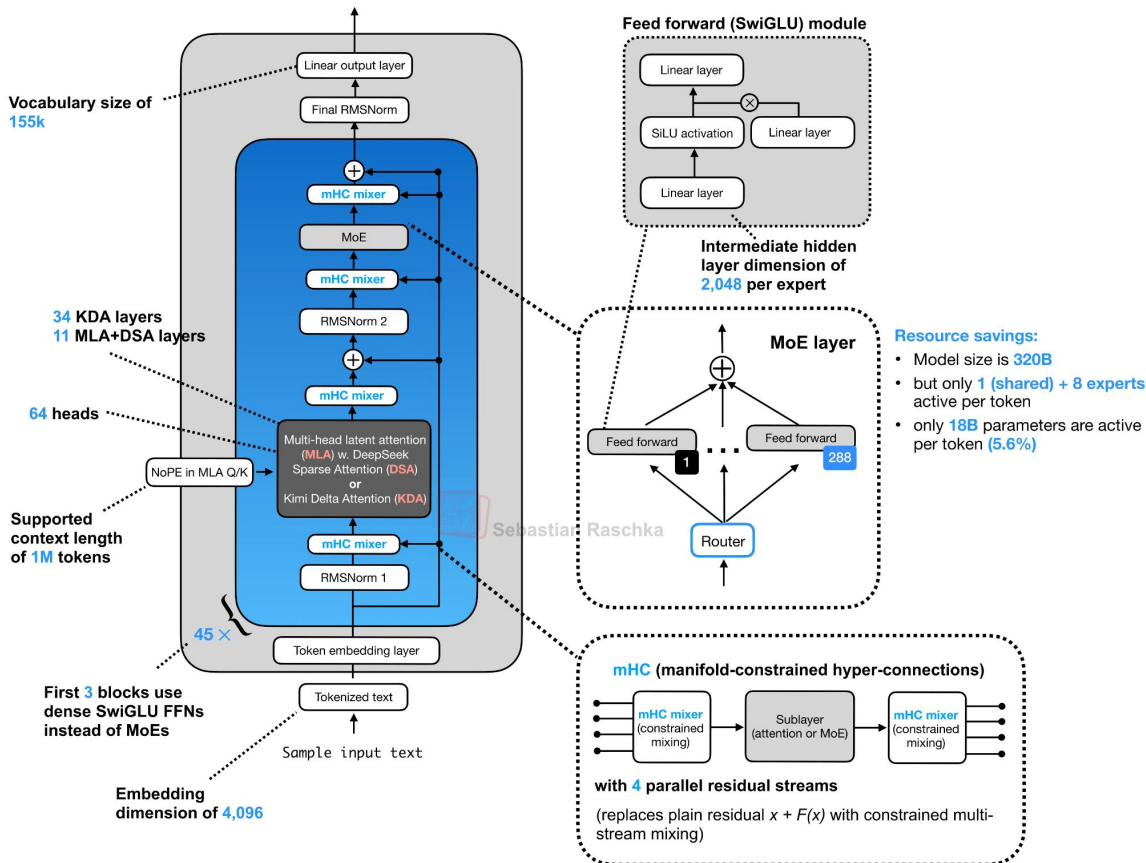


# Parallelism Schemes at a Glance



# GLM-5.3-Flash (320B-A18B) (“Ox Alpha”)

(Text decoder shown; 24-layer vision encoder and one MTP prediction layer omitted)

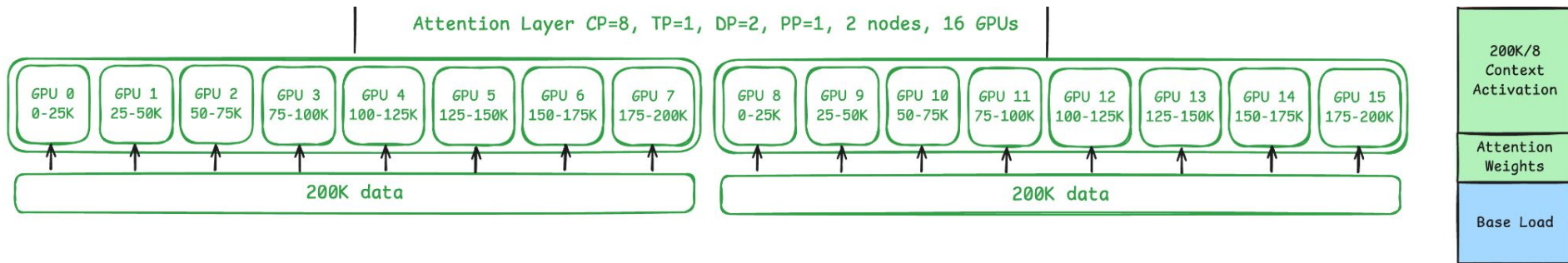


# GLM-5.3-Flash Attention Grid

Each layer in MoE models has an attention module followed by an expert module

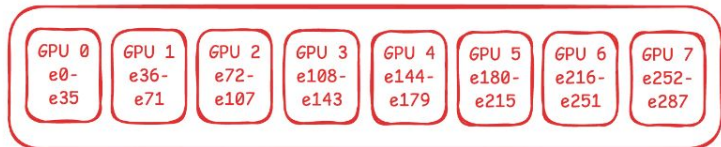
We are mainly optimizing for attention activation memory usage for attention grid

Attention Grid Sizes:  $CP \times TP \times DP \times PP = \text{world size}$

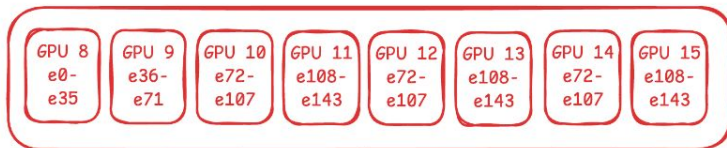


# GLM-5.3-Flash Expert Grid

Expert MLP Layer EP=8, ETP=1, EDP=2, PP=1, 2 nodes, 16 GPUs



All-to-All  
Token Routing



All-to-All  
Token Routing

VRAM Usage

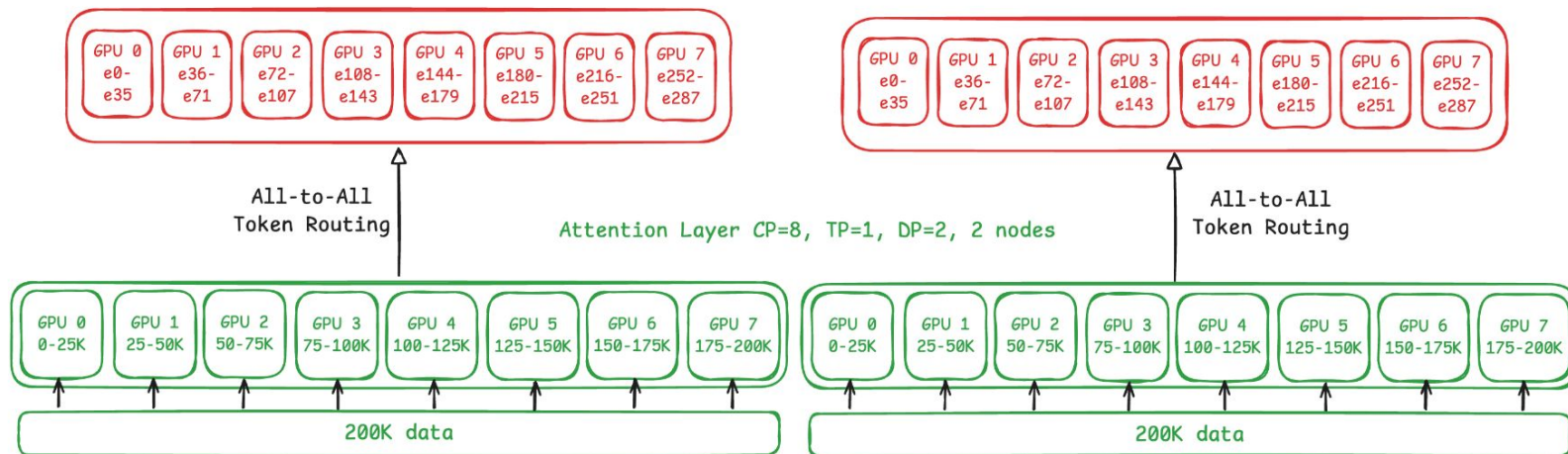
|                            |
|----------------------------|
| 36<br>Expert<br>Activation |
| 36<br>Expert<br>Weights    |

We are mainly optimizing for expert weight memory usage for expert grid

Expert Grid Sizes:  $EP \times ETP \times EDP \times PP = \text{world size}$

# GLM-5.3-Flash CP & EP Strategies

Expert MLP Layer EP=8, ETP=1, EDP=2, 2 nodes



VRAM Usage

|                                 |
|---------------------------------|
| 36<br>Expert<br>Activation      |
| 36<br>Expert<br>Weights         |
| 200K/8<br>Context<br>Activation |
| Attention<br>Weights            |
| Base Load                       |

# Thank You

Suman Debnath

[sdebnath@crusoe.ai](mailto:sdebnath@crusoe.ai)

Jiale Huang

[jhuang@crusoe.ai](mailto:jhuang@crusoe.ai)

Daniel Shen

[dshen@crusoe.ai](mailto:dshen@crusoe.ai)

[JOIN US](#)

## We're Hiring

Distributed Training Experts  
Applied Researchers  
Product Managers  
Software Engineers



