Outline

- Multi-Token Prediction (MTP)
- Inference
  - prefilling
  - decoding

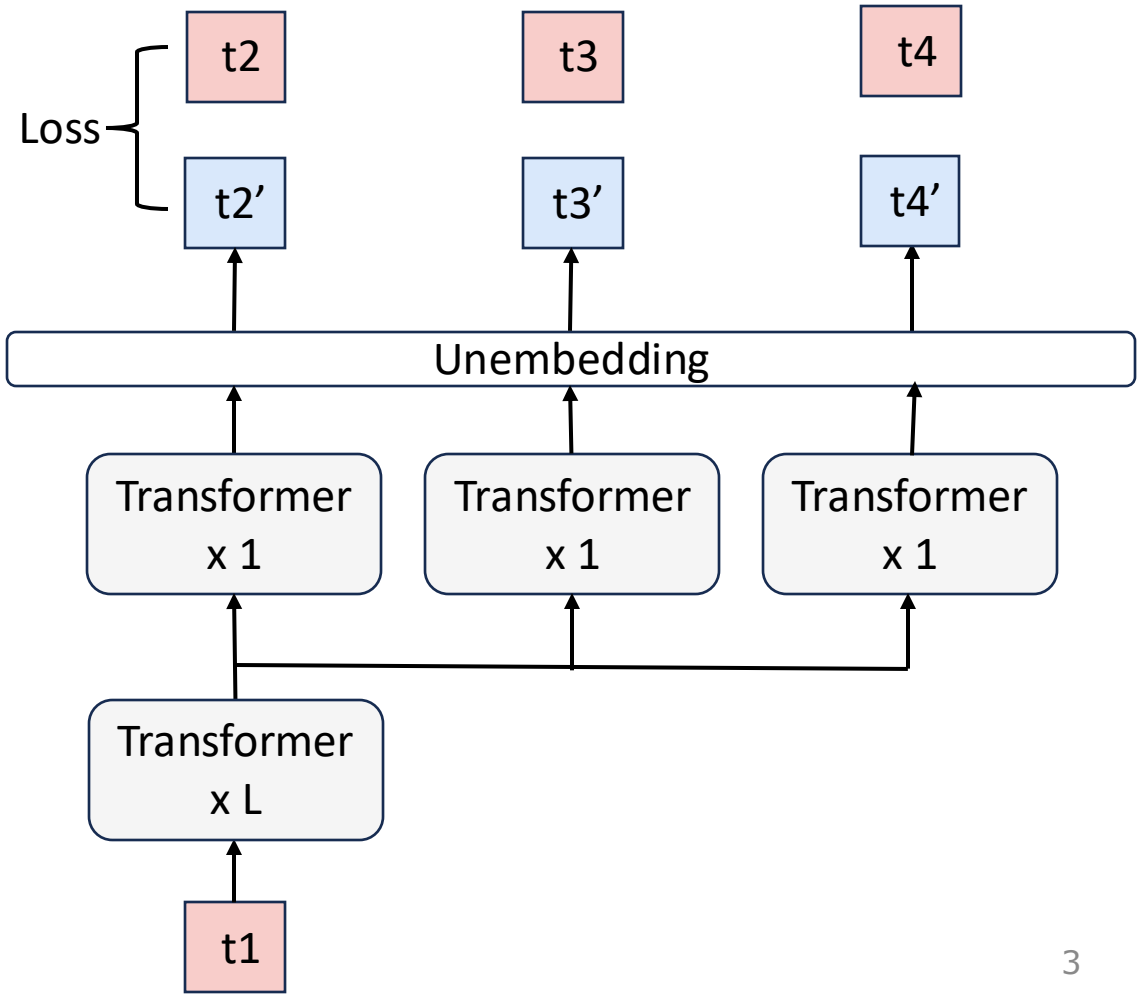
Multi-Token Prediction

- Insight & Motivation
  - traditional training methods require a large amount of data
- MTP
  - train: predicting multiple tokens at once can improve data efficiency
  - inference: used for speculative decoding to further improve generation latency

Multi-Token Prediction

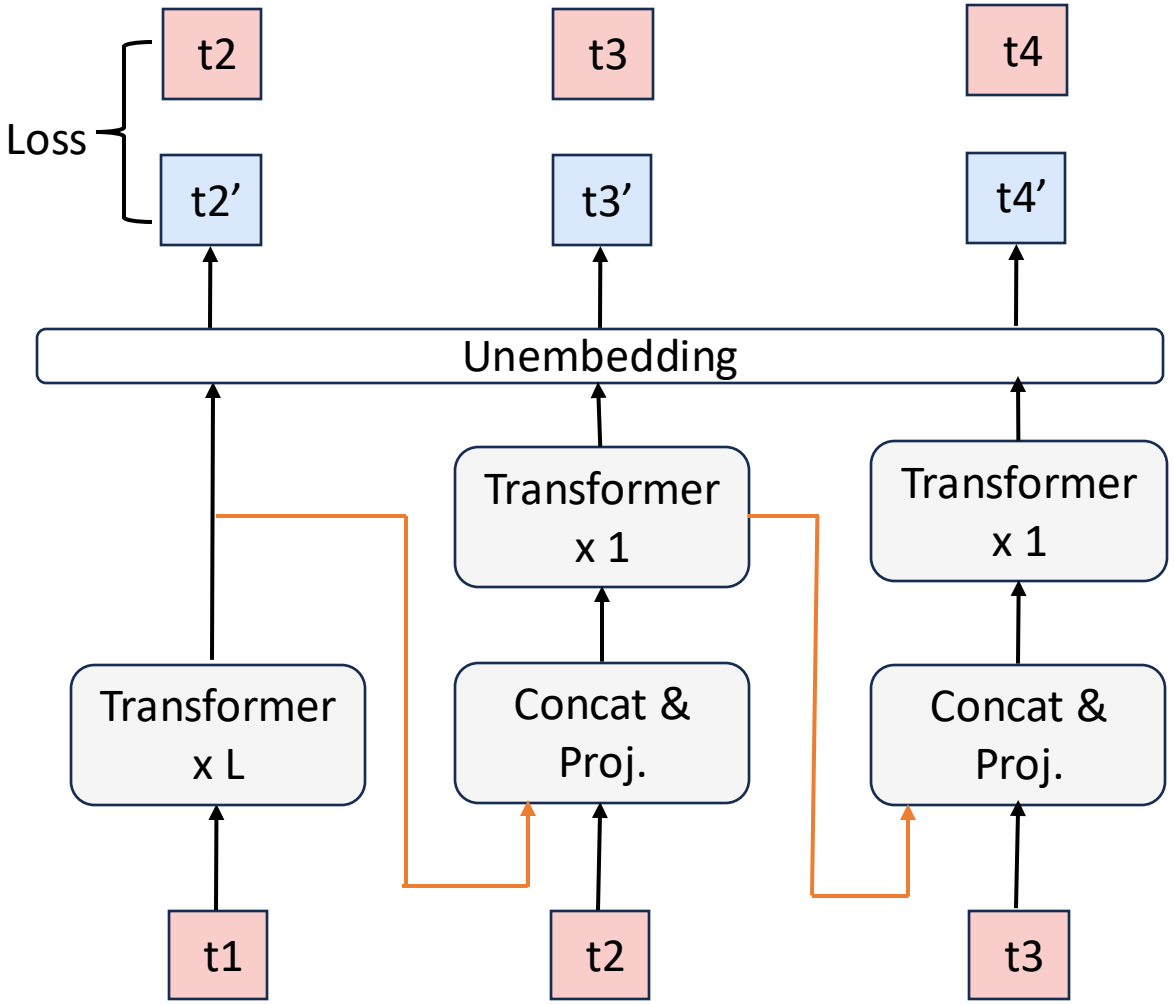
- *ICML2024: Better & Faster Large Language Models via Multi-token Prediction*

- Structure
  - shared trunk: Transformer x L
  - independent output heads:
    - Transformer x 1
    - parallelly predict future tokens



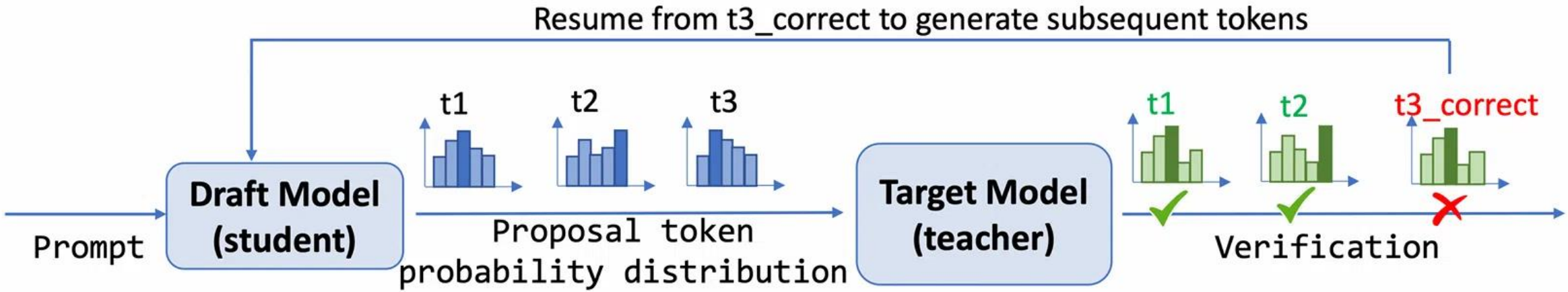
Multi-Token Prediction

- Deepseek Implementation
  - predict additional tokens sequentially
  - keep the complete causal chain
- MTP module
  - concat & proj.
    - output from the previous module
    - ground truth of the next sample
  - 1 transformer layer



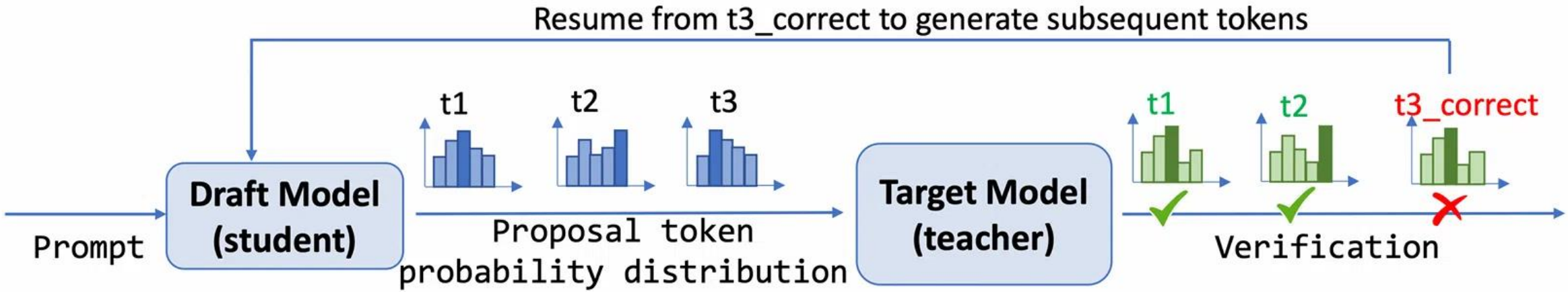
Multi-Token Prediction

- Speculative Decoding
  - target model: high accuracy but slower speed
  - draft model: faster speed but lower accuracy



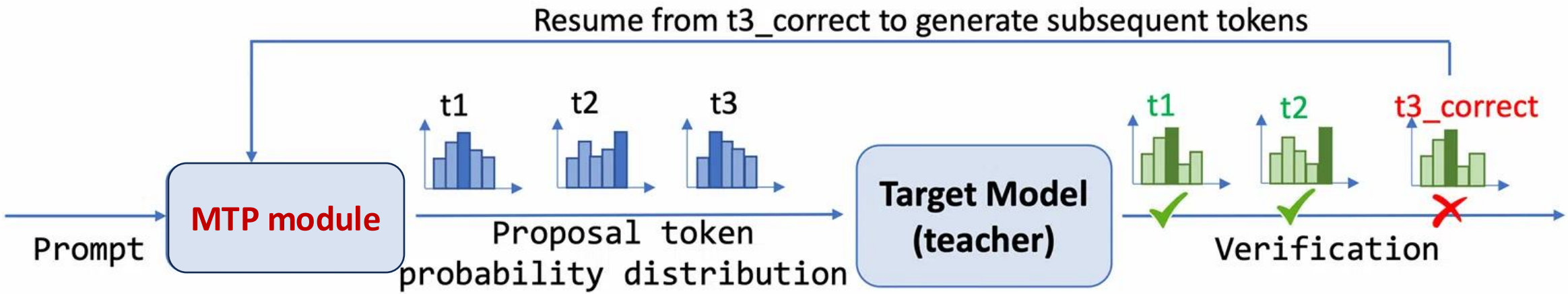
Multi-Token Prediction

- Verify in Parallel
  - draft model: generates multiple tokens
  - target model: verify all generated tokens in 1 step
    - predicts the probability distribution for all generated tokens in 1 step



Multi-Token Prediction

- MTP -> Speculative Decoding



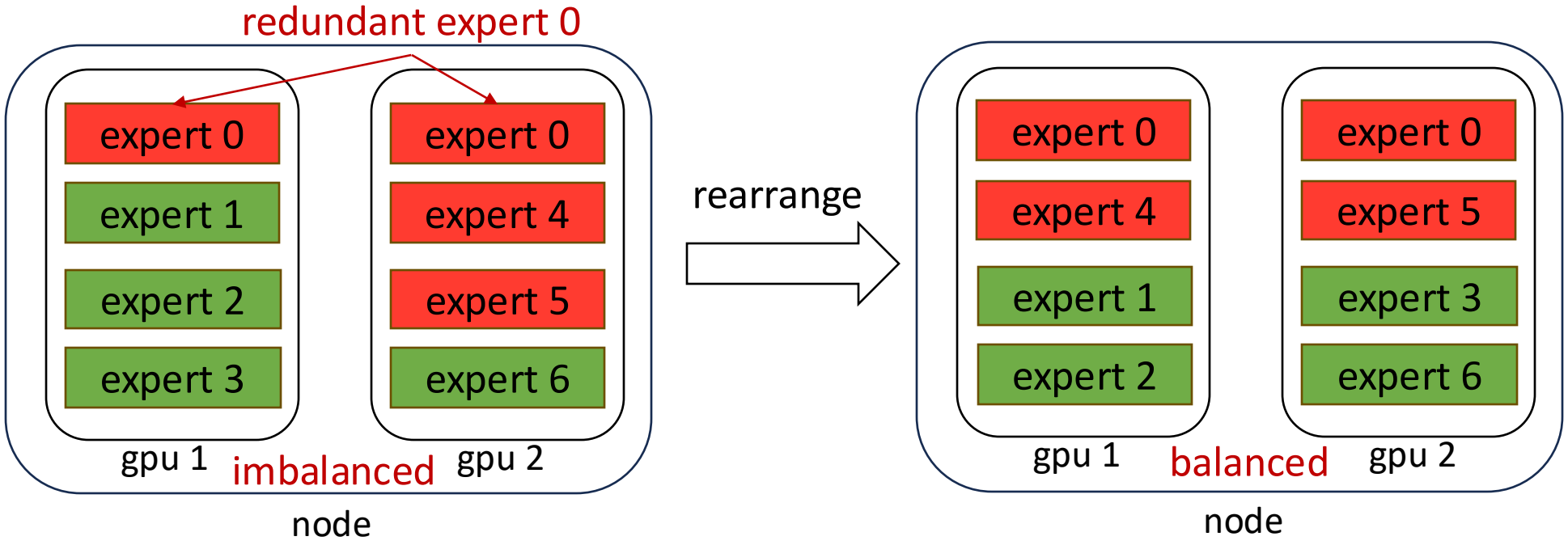
Prefilling

- Deployment
  - 4 nodes with 32 GPUs
  - attention block: TP4 with SP, DP8
    - set small TP size to limit communication overhead
  - MoE block: EP32
    - each expert processes a sufficiently large batch size
    - improves computation intensity of experts

参数	Prefill	Decode
PP	1	1
Attn TP	4	4
Attn DP	8	**80**
MoE TP	1	1
MoE EP	32	**320**
# GPU	32	**320**
# token	bs	**b**

Prefilling

- MoE load balancing
  - redundant experts
    - duplicates high-load experts and deploys them on multiple GPUs
    - adjusts periodically based on online stats
  - rearrange experts among GPUs within a node



Decoding

- Deployment
  - 40 nodes with 320 GPUs
  - attention block: TP4 with SP, DP80
  - MoE block: EP320
    - each GPU hosts one expert
    - 64 GPUs handle redundant experts

参数	Prefill	Decode
PP	1	1
Attn TP	4	4
Attn DP	8	**80**
MoE TP	1	1
MoE EP	32	**320**
# GPU	32	**320**
# token	bs	**b**

Decoding

- MoE load balancing
  - redundant experts
    - periodically determine the set of redundant experts
  - each GPU only hosts one expert, do not need to rearrange experts