Read-ME: Refactorizing LLMs **a**s Router-**D**ecoupled **M**ixture-of-**E**xperts with System Co-design

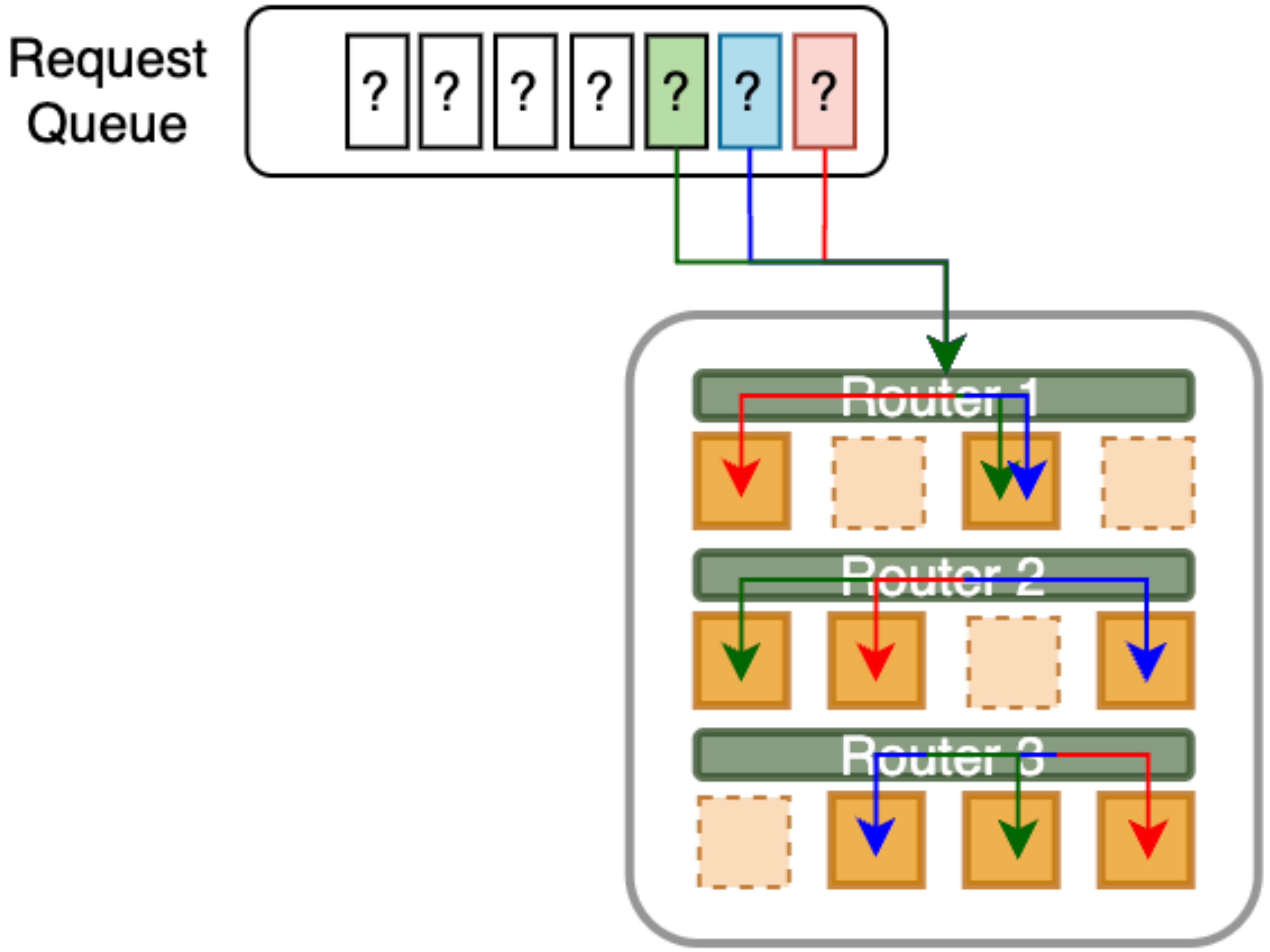
Friday, 13 Dec 11AM-2PM

Poster Session 5

Ruisi Cai*, **Yeonju Ro***, Geon-Woo Kim, Peihao Wang, Babak Ehteshami Bejnordi, Aditya Akella, Zhangyang Wang

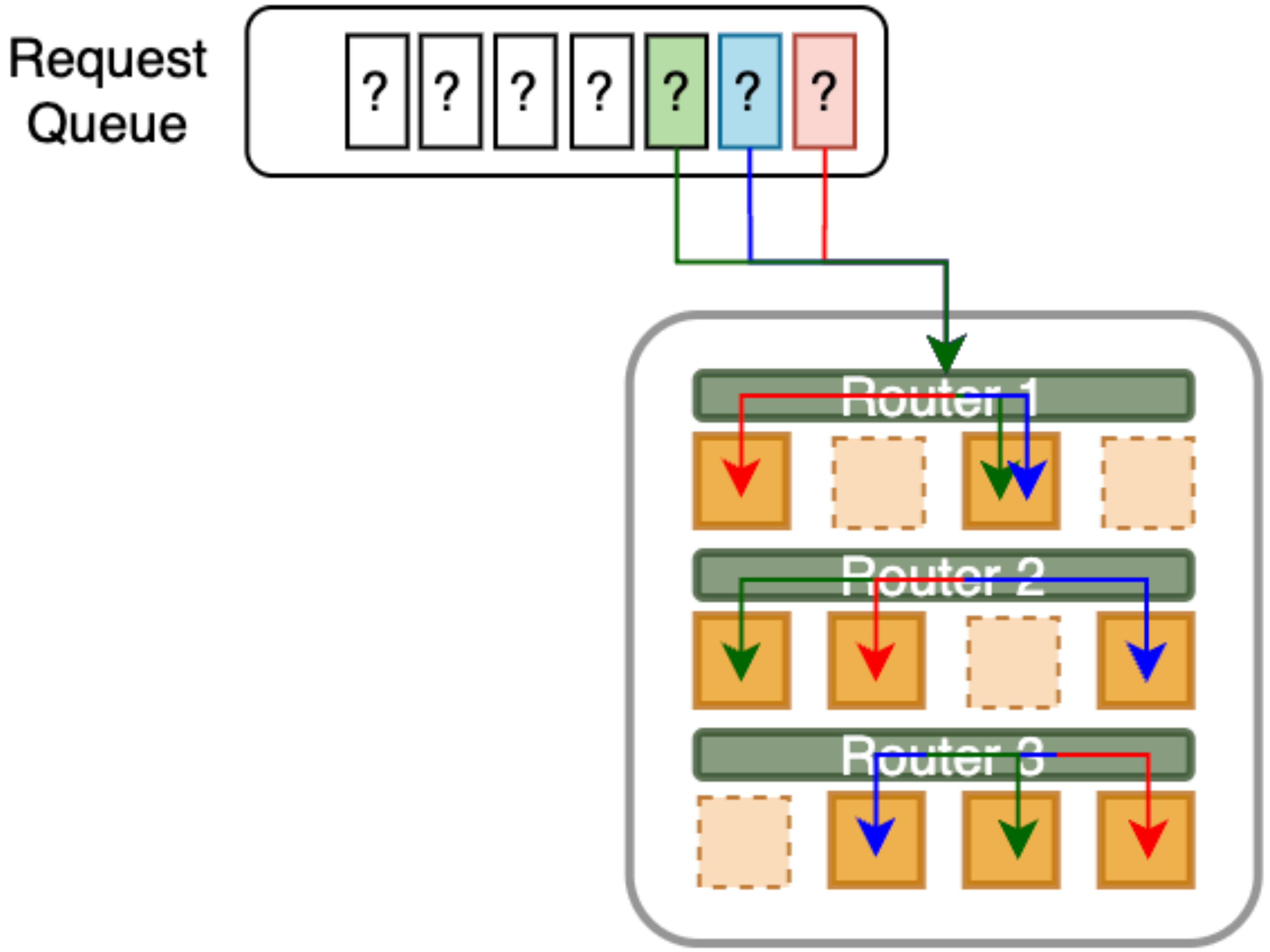
Drawback of current MoE Architecture

Layer-wise Gating Disrupts Efficient Memory Management



Drawback of current MoE Architecture

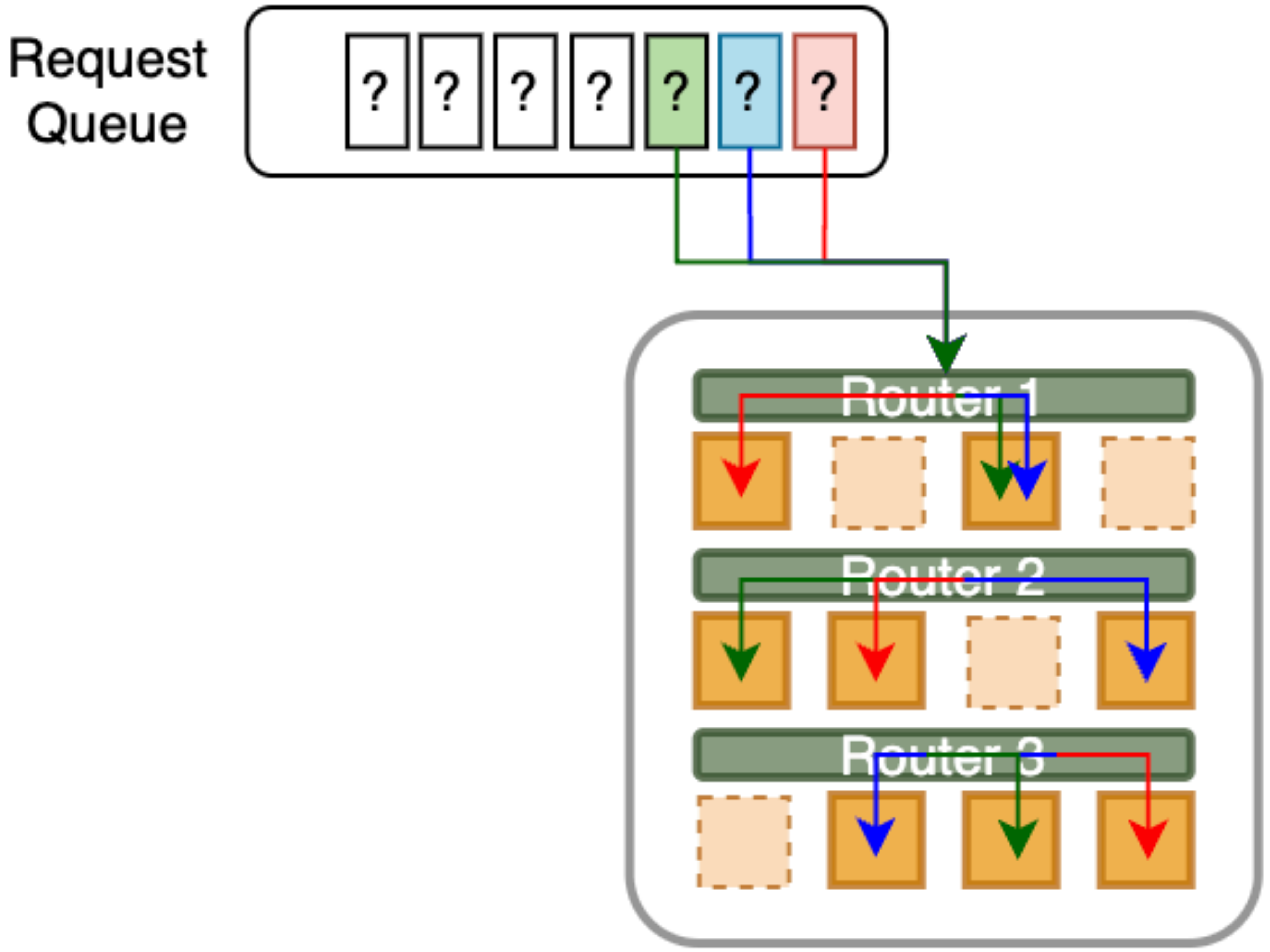
Layer-wise Gating Disrupts Efficient Memory Management



- Ideally, we only want to load layers to be used.

Drawback of current MoE Architecture

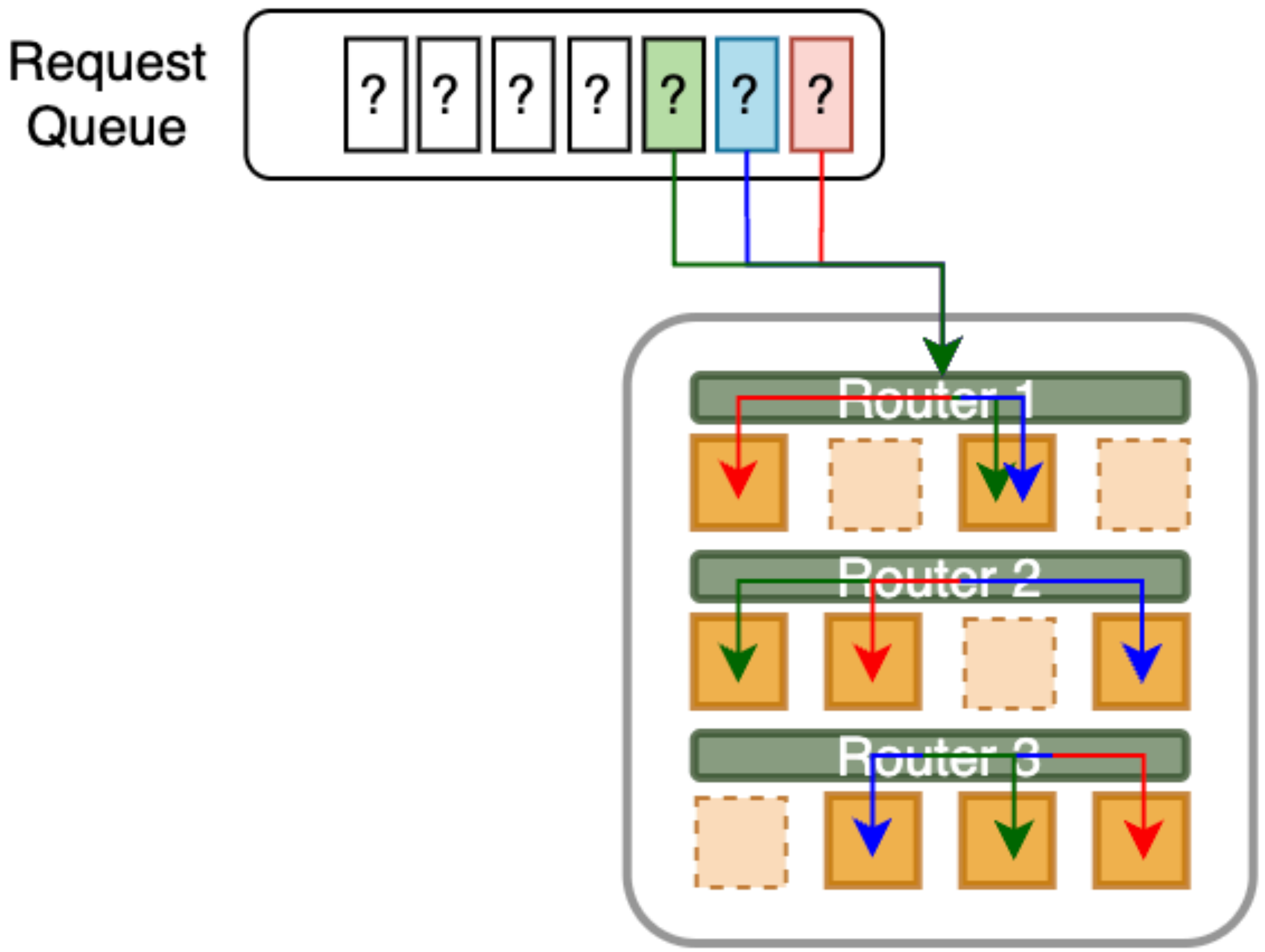
Layer-wise Gating Disrupts Efficient Memory Management



- Ideally, we only want to load layers to be used.
- However, we do not know which expert to activate until we go through the layer-wise routing layer.

Drawback of current MoE Architecture

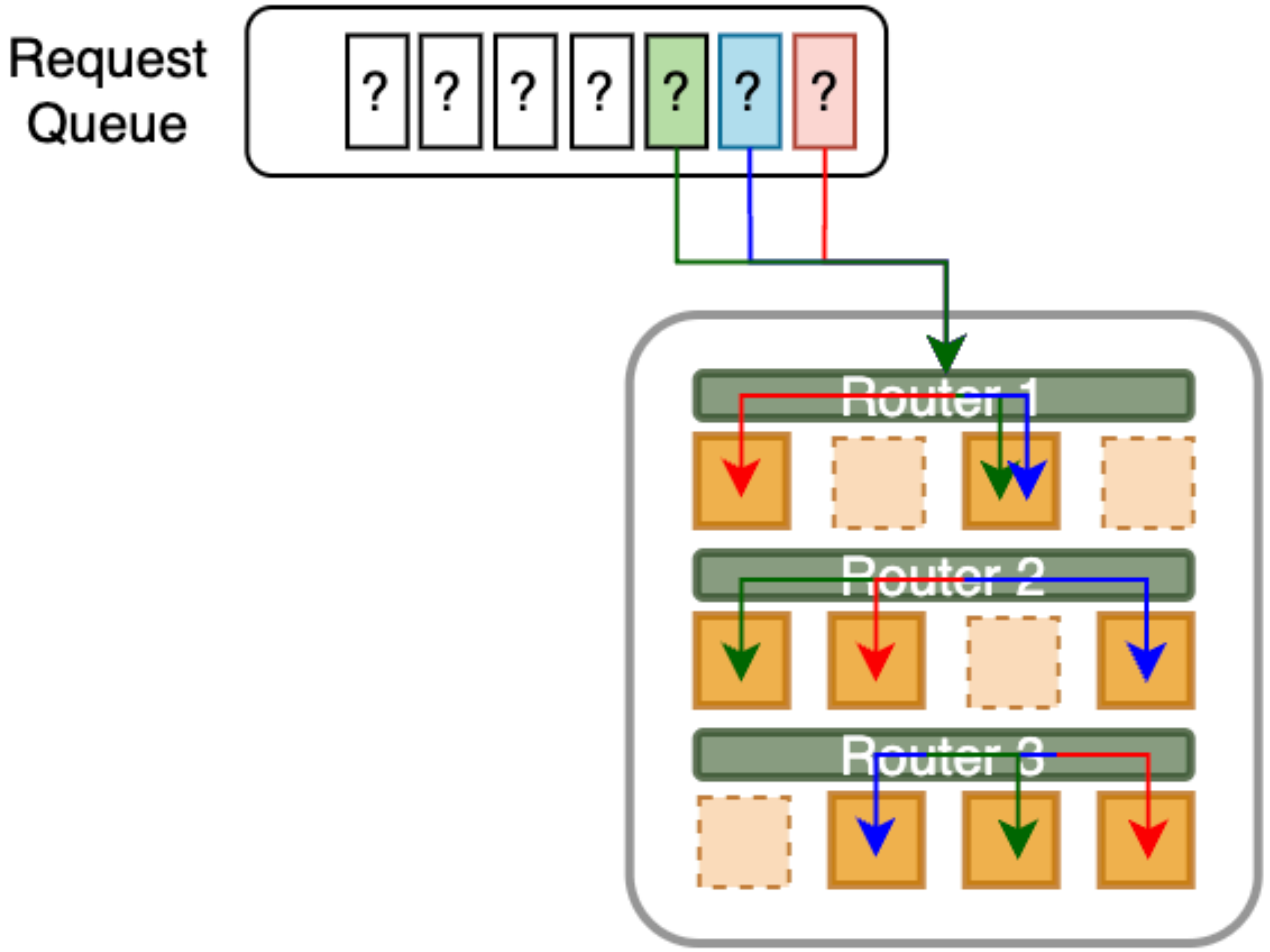
Layer-wise Gating Disrupts Efficient Memory Management



- Ideally, we only want to load layers to be used.
- However, we do not know which expert to activate until we go through the layer-wise routing layer.
- Thus, we need to either load all experts or load the expert to be used *on-demand*, which potentially add loading latency to the critical path.

Drawback of current MoE Architecture

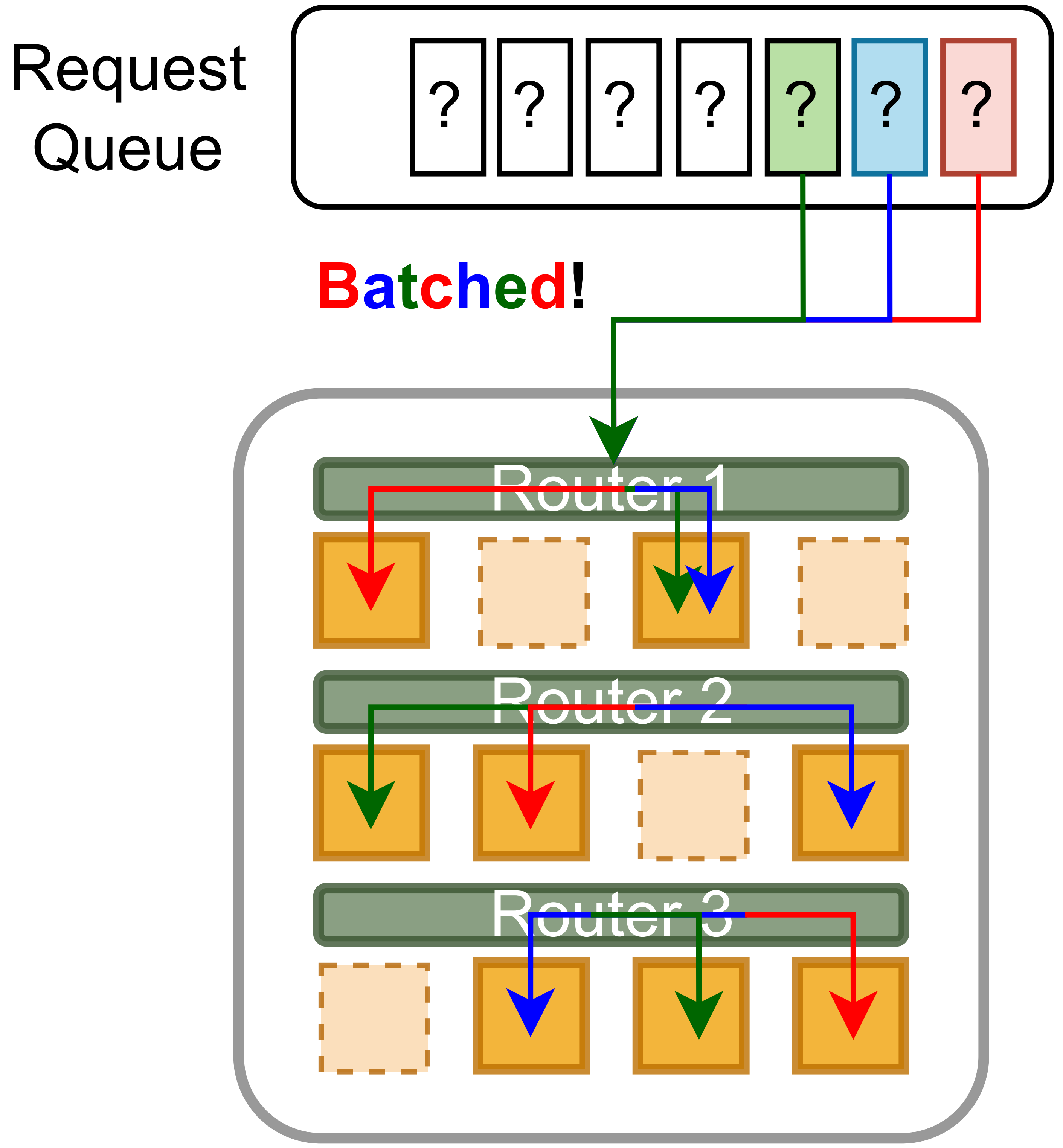
Layer-wise Gating Disrupts Efficient Memory Management



- Ideally, we only want to load layers to be used.
- However, we do not know which expert to activate until we go through the layer-wise routing layer.
- Thus, we need to either load all experts or load the expert to be used *on-demand*, which potentially add loading latency to the critical path.
- Both prefetching/caching are not trivial.

Current Inference System

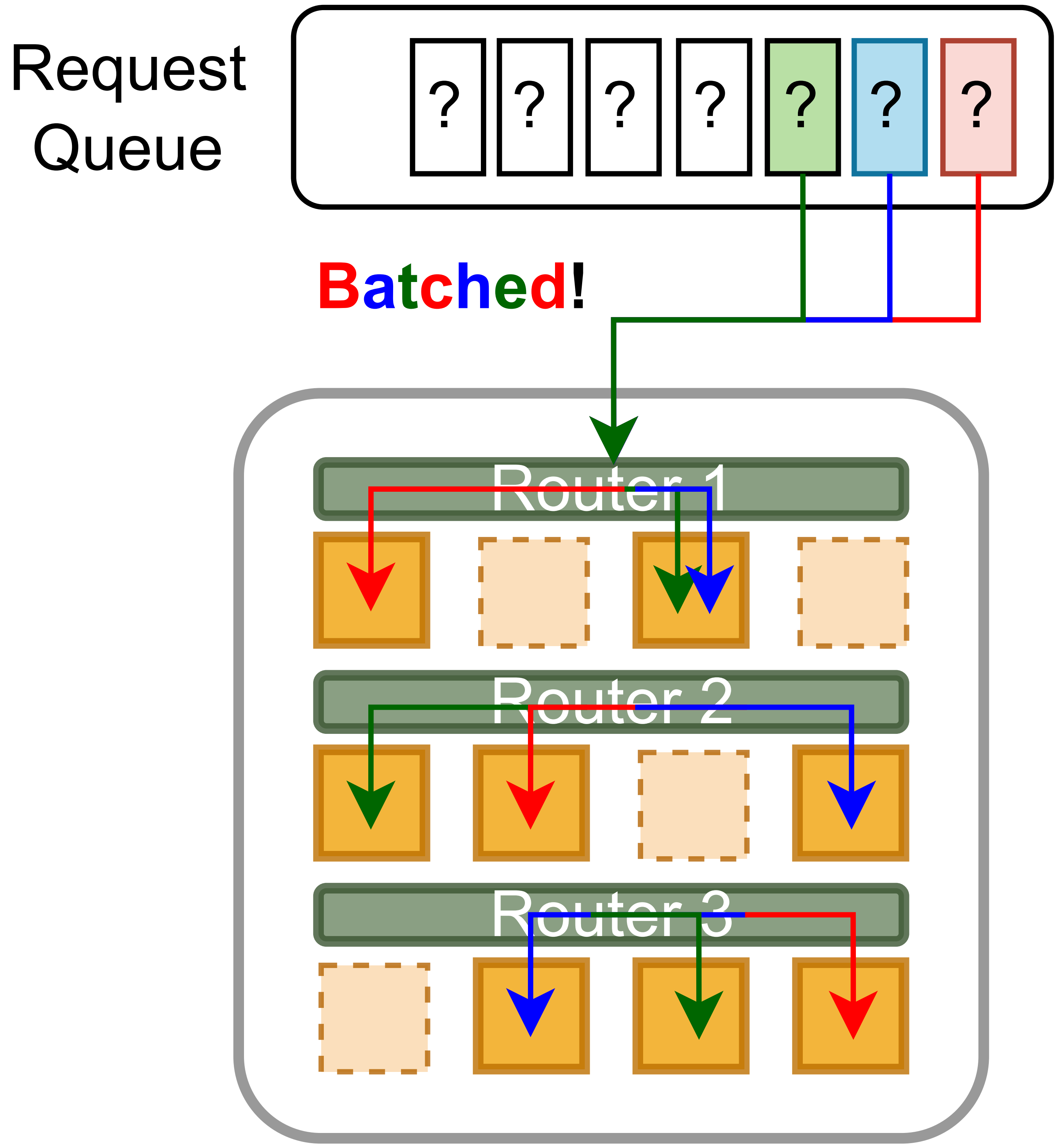
Layer-wise Gating Disrupts Efficient Batching



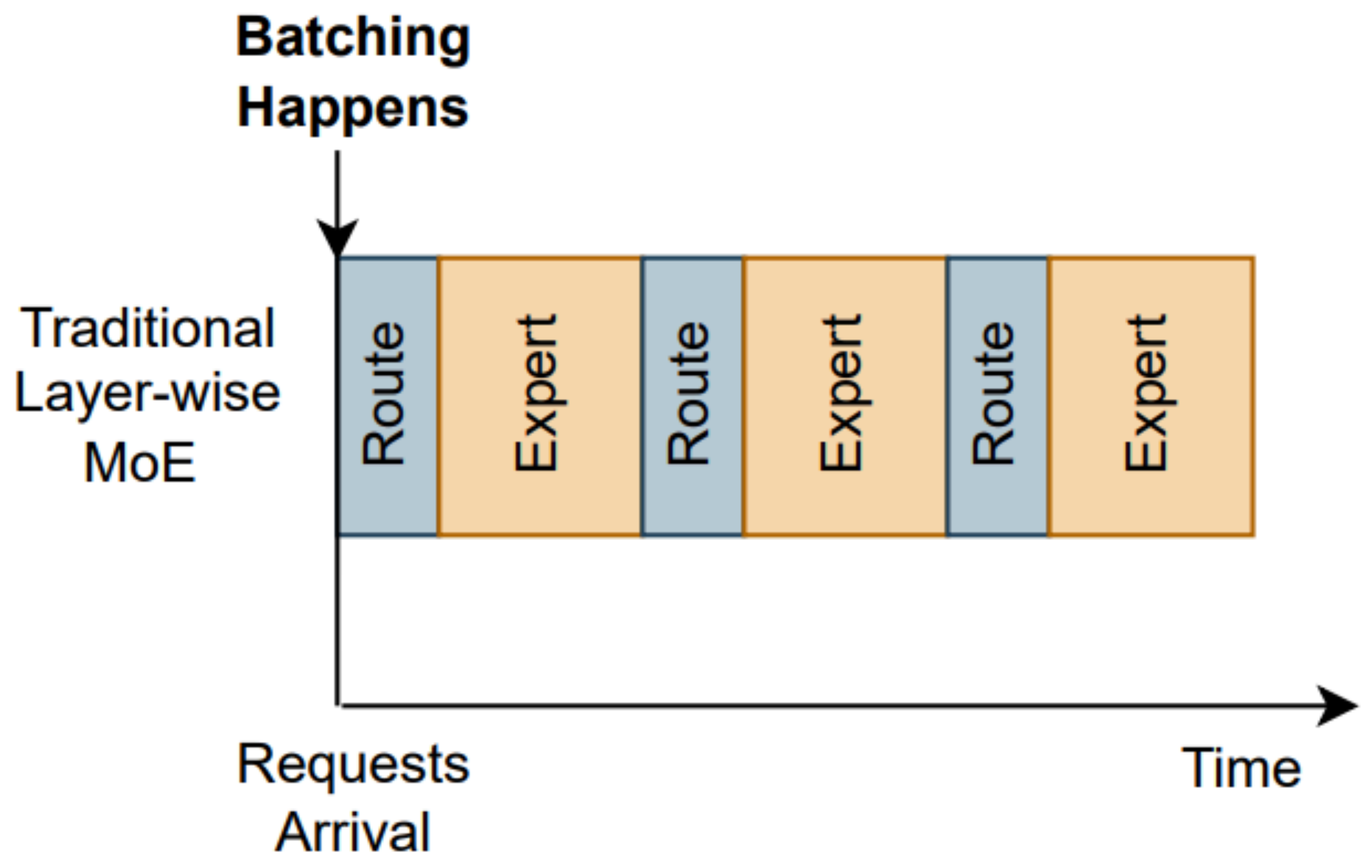
Uniq(Activated Experts)

Current Inference System

Layer-wise Gating Disrupts Efficient Batching

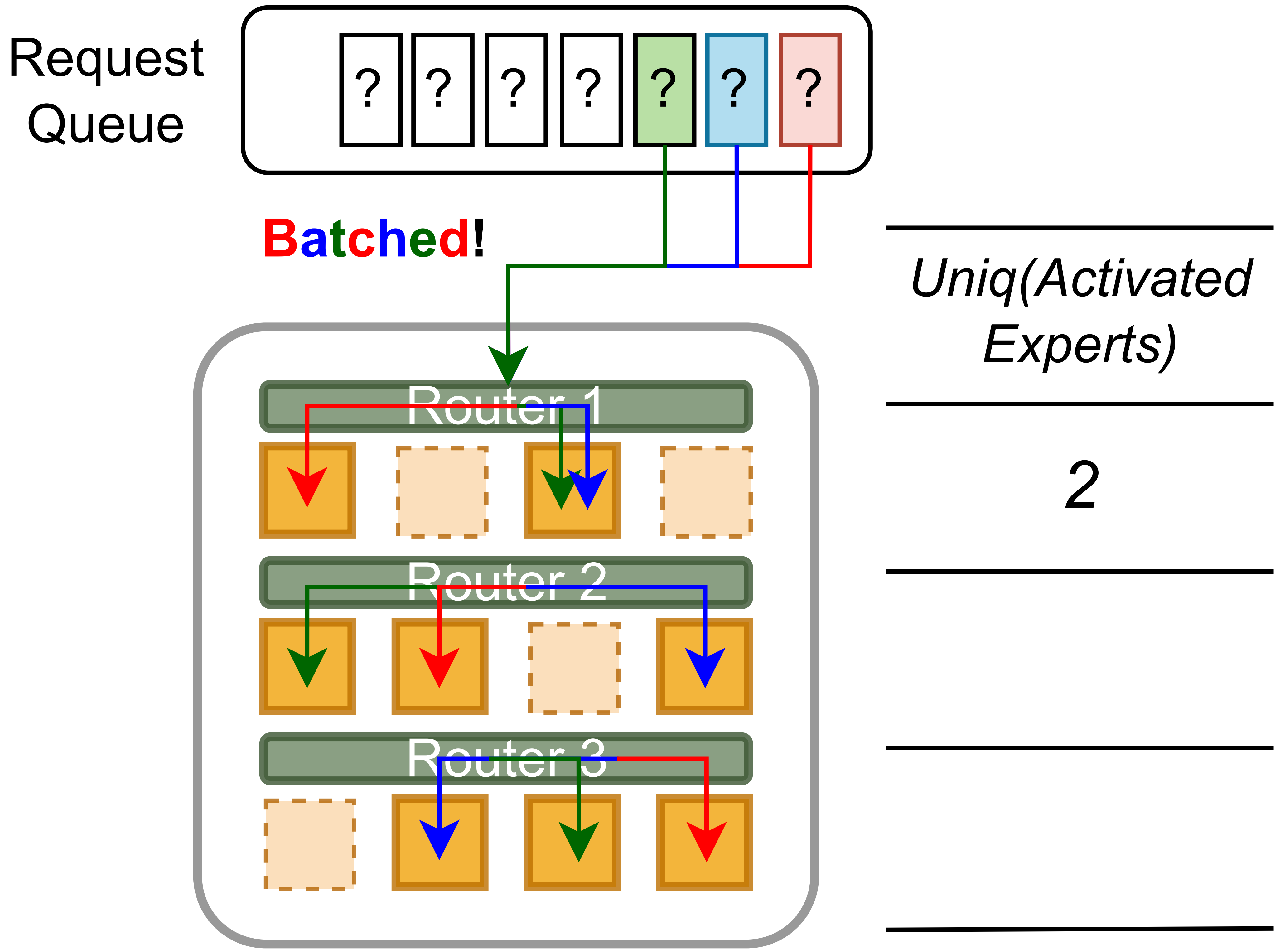


Uniq(Activated Experts)



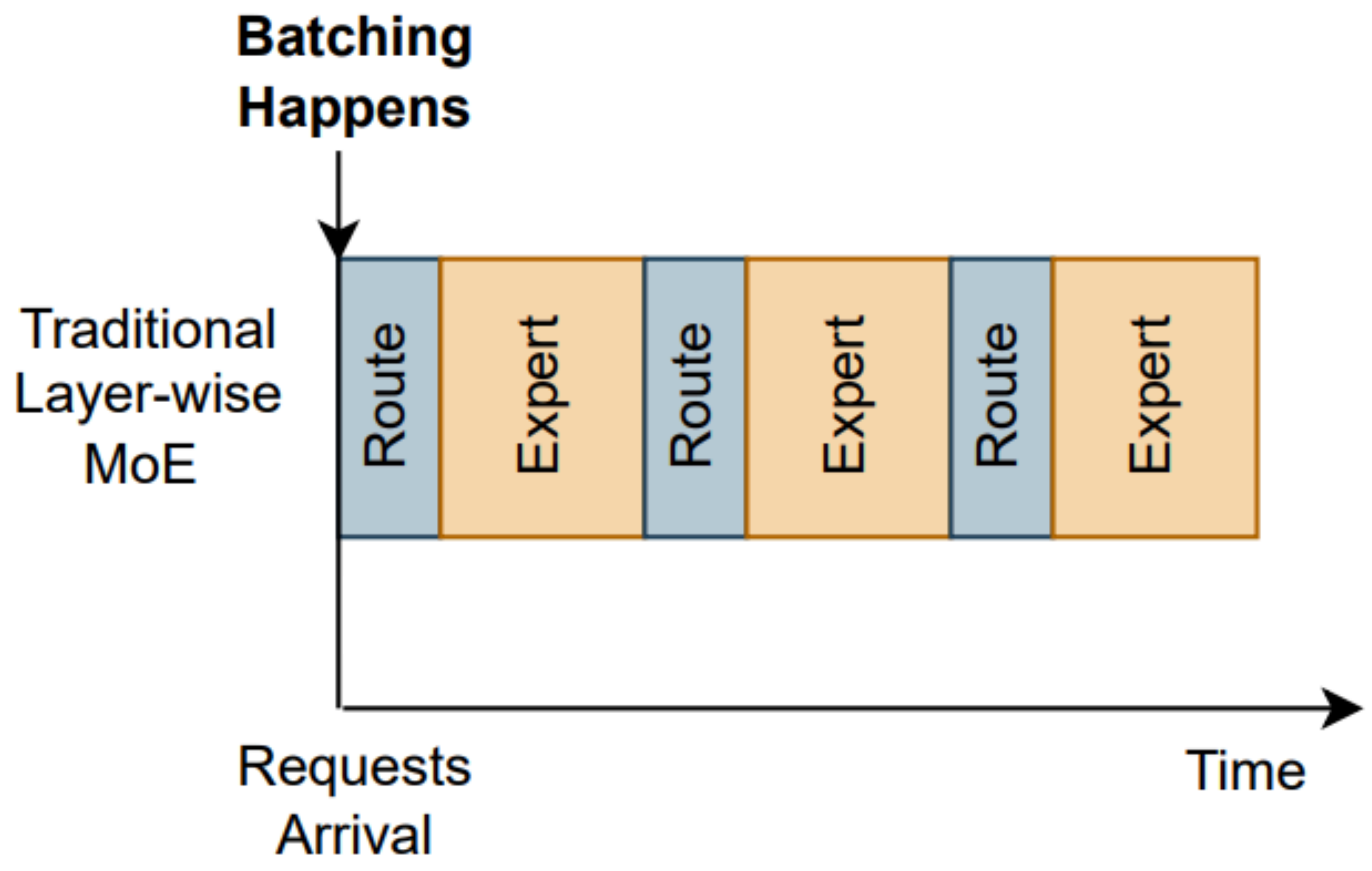
Current Inference System

Layer-wise Gating Disrupts Efficient Batching



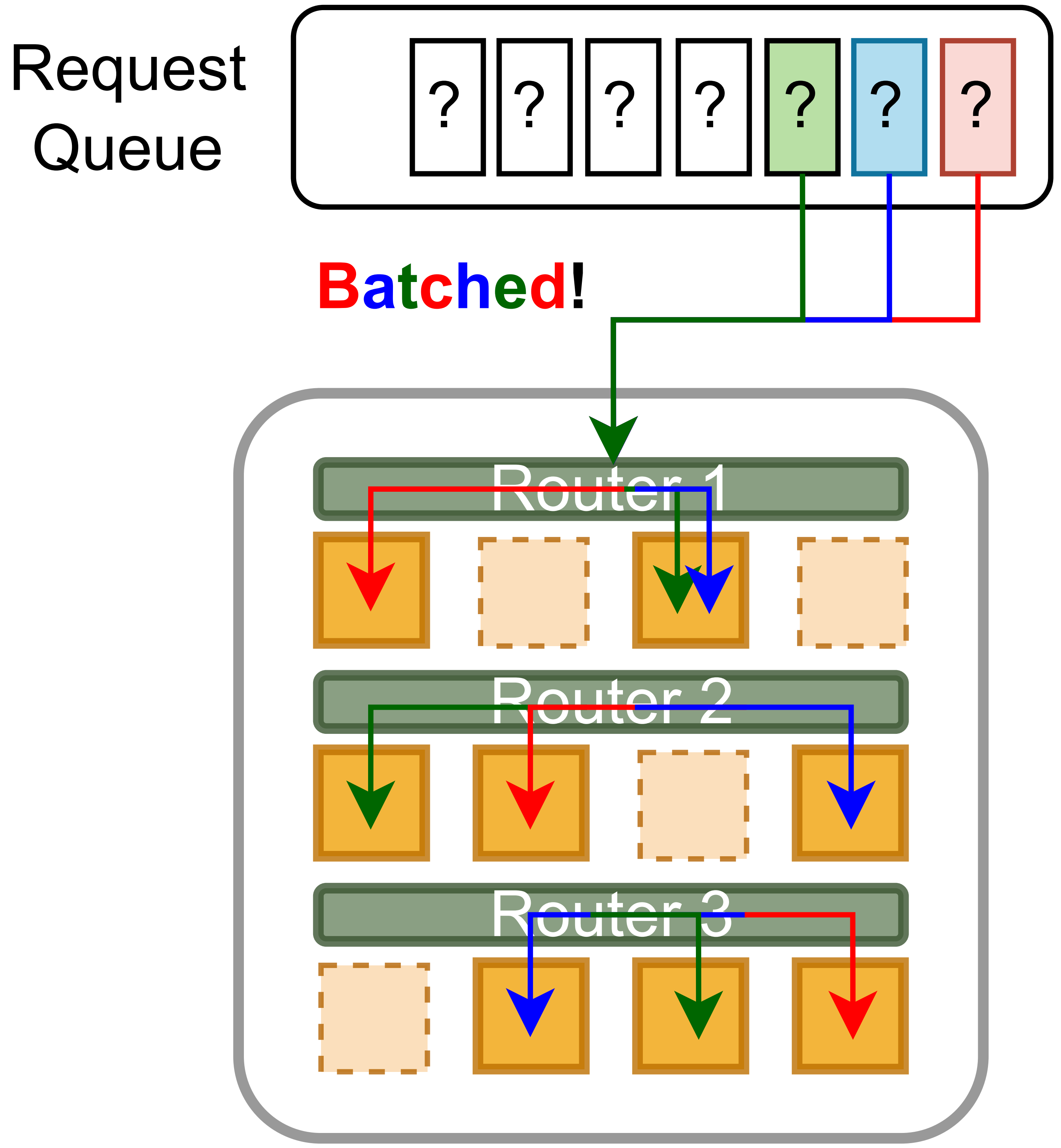
Uniq(Activated Experts)

2

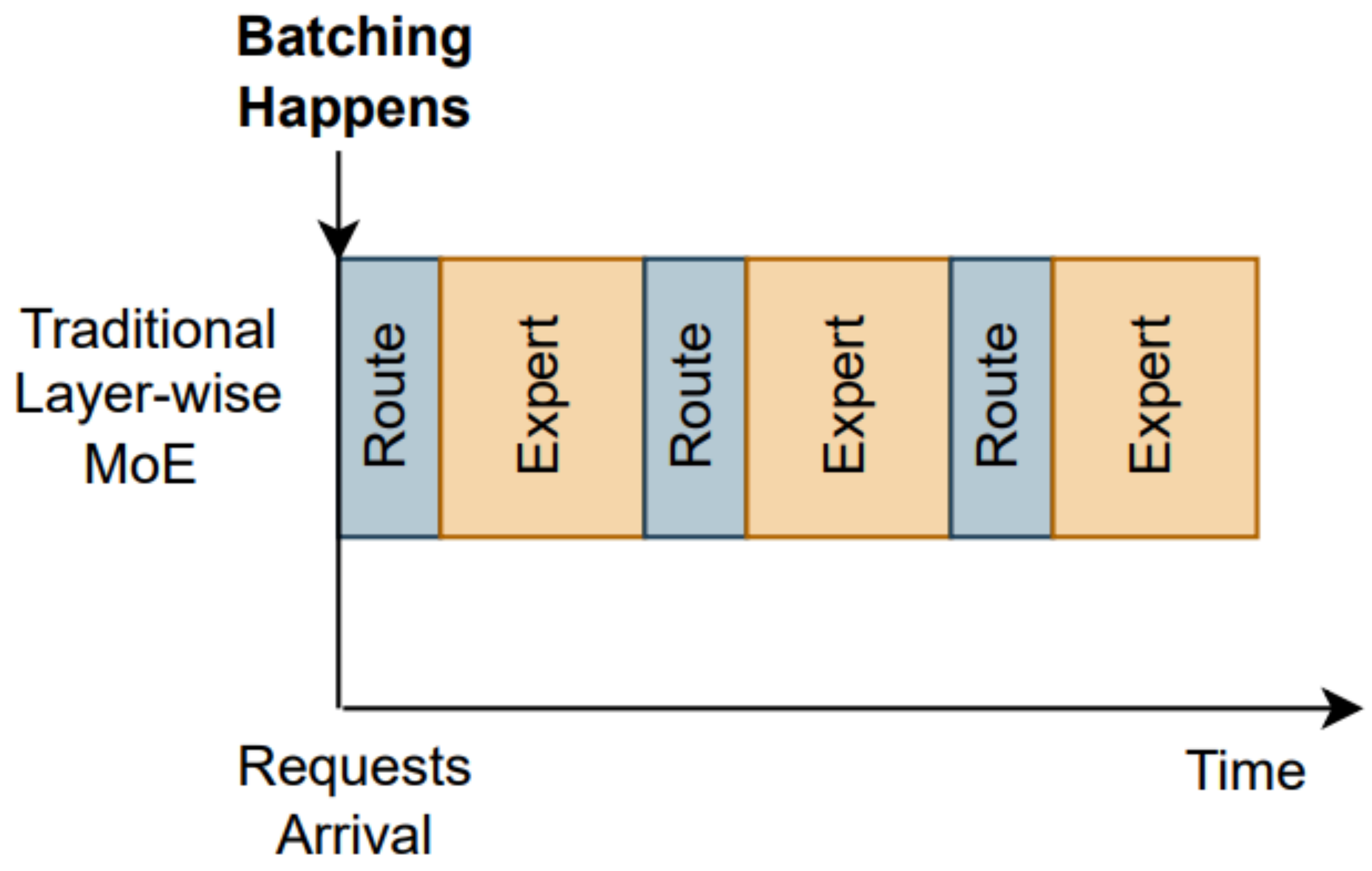


Current Inference System

Layer-wise Gating Disrupts Efficient Batching

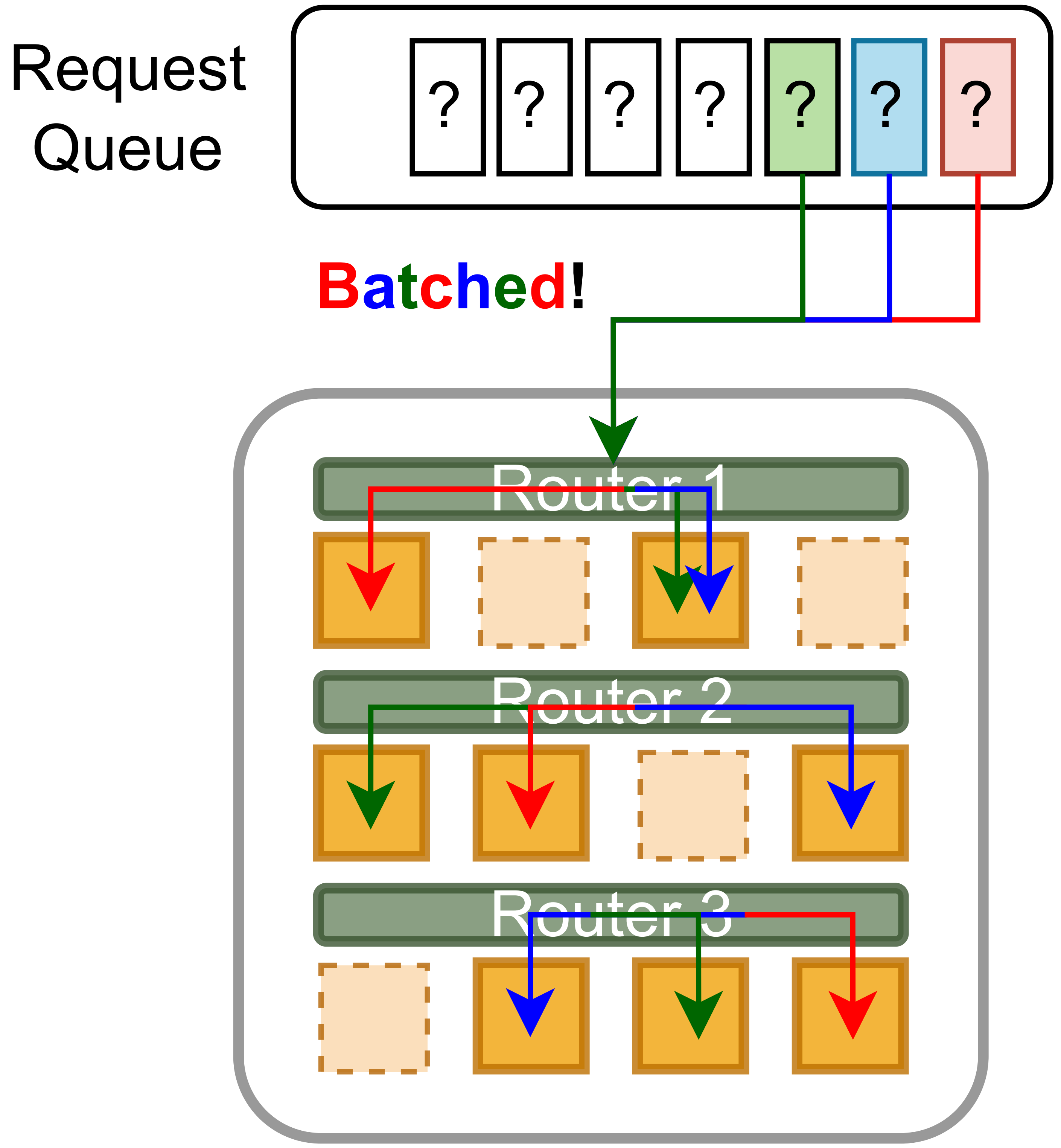


Uniq(Activated Experts)
2
3

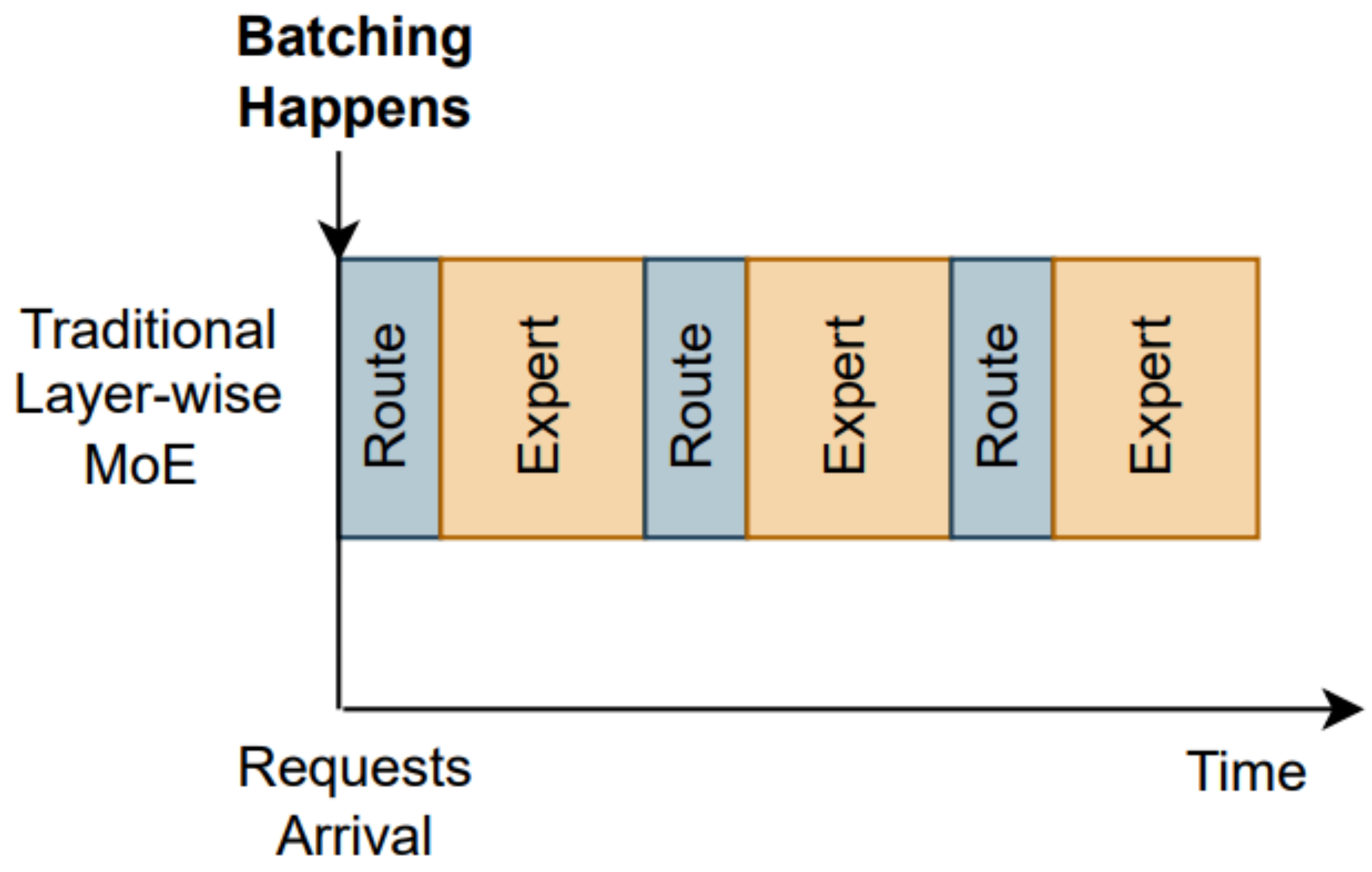


Current Inference System

Layer-wise Gating Disrupts Efficient Batching

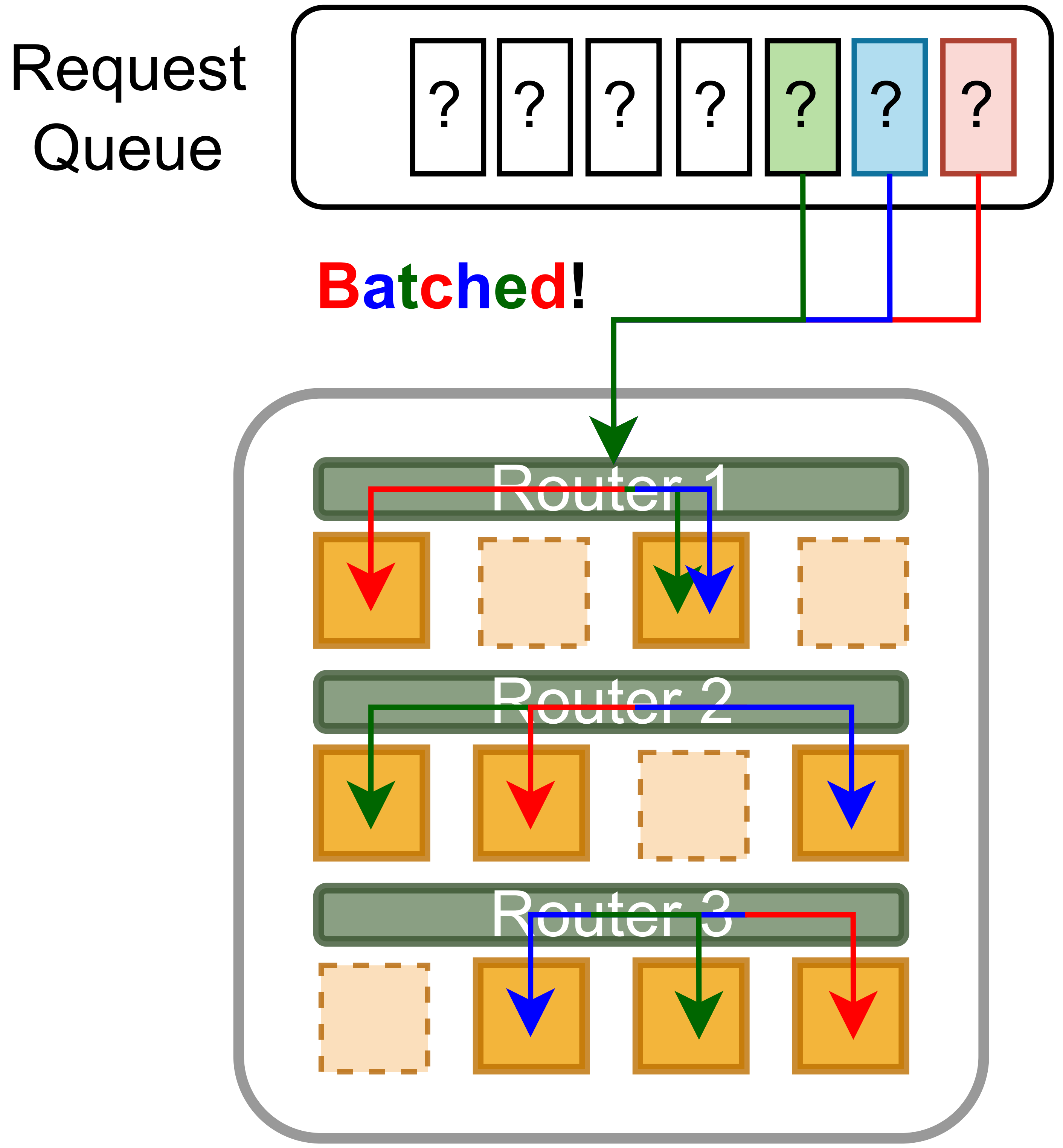


Uniq(Activated Experts)
2
3
3

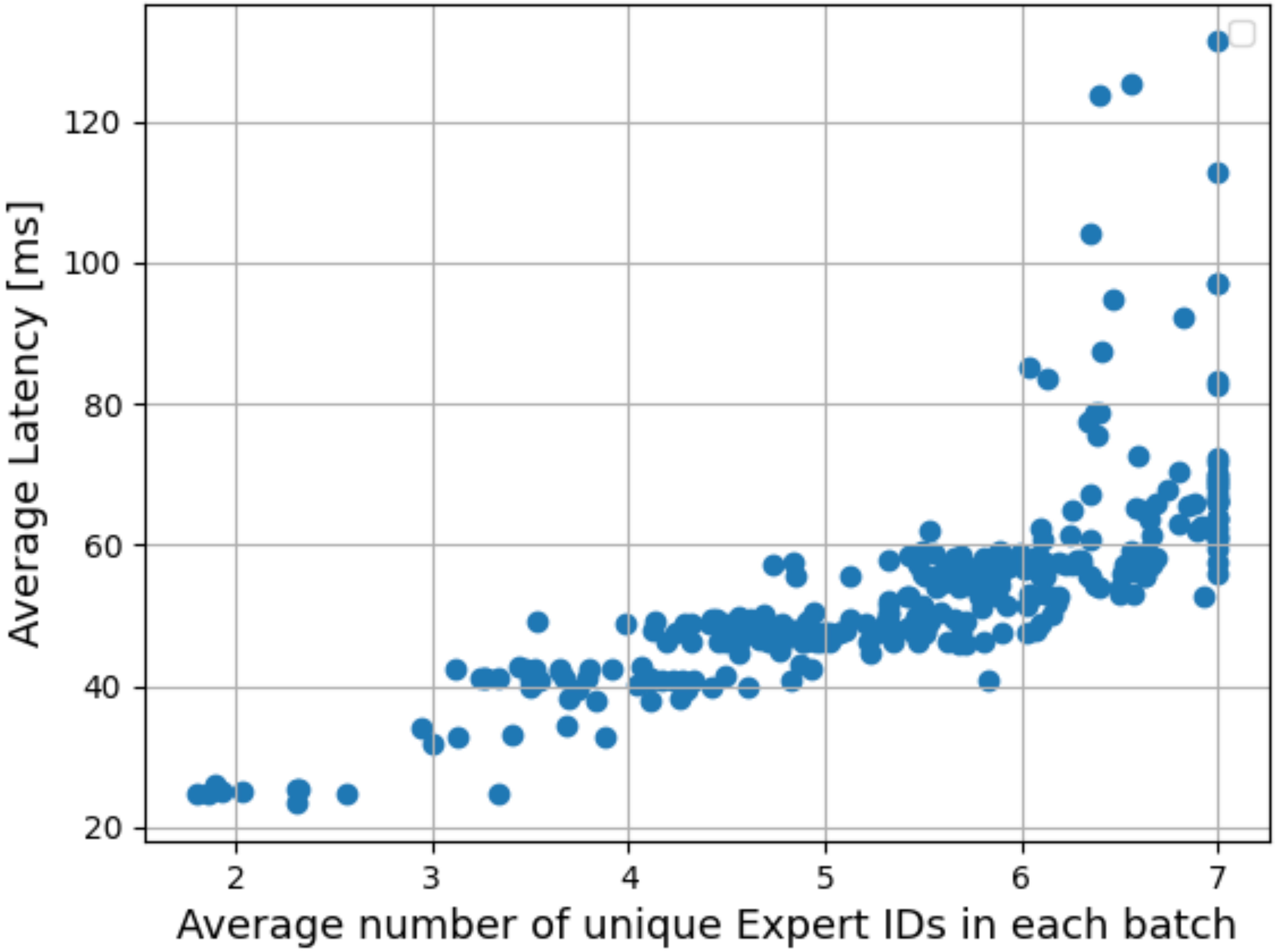


Current Inference System

Token batching is poorly suited for MoE architecture



Uniq(Activated Experts)
2
3
3



Solution: Decoupled Router

Re-designing MoE Architecture with Decoupled Router

Router	Router
🟥	Expert 1
🟩	Expert 2
🟨	Full Pre-trained Weight
?	Tokens to be routered
🟥	Token routered to Expert 1
🟩	Token routered to Expert 2

Pre-trained Dense Model

Refactoring

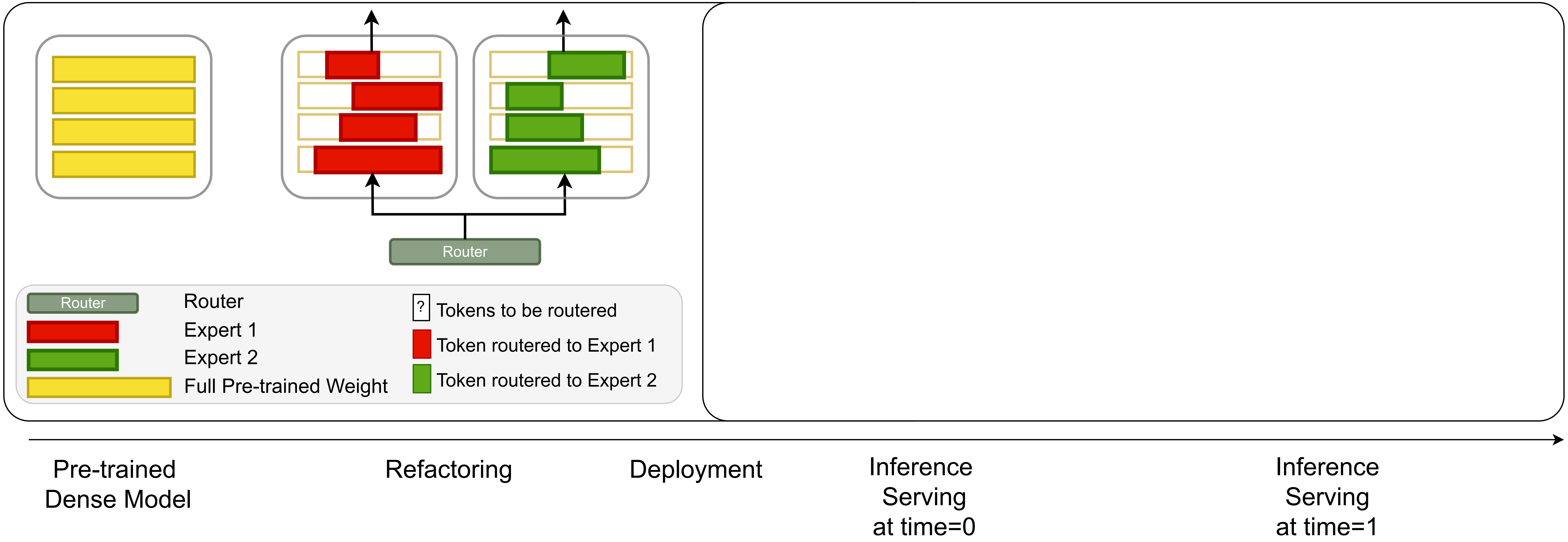
Deployment

Inference Serving at time=0

Inference Serving at time=1

Solution: Decoupled Router

Re-designing MoE Architecture with Decoupled Router



Router

Router — Router

🟥 — Expert 1

🟩 — Expert 2

🟨 — Full Pre-trained Weight

? — Tokens to be routered

🟥 — Token routered to Expert 1

🟩 — Token routered to Expert 2

Pre-trained Dense Model

Refactoring

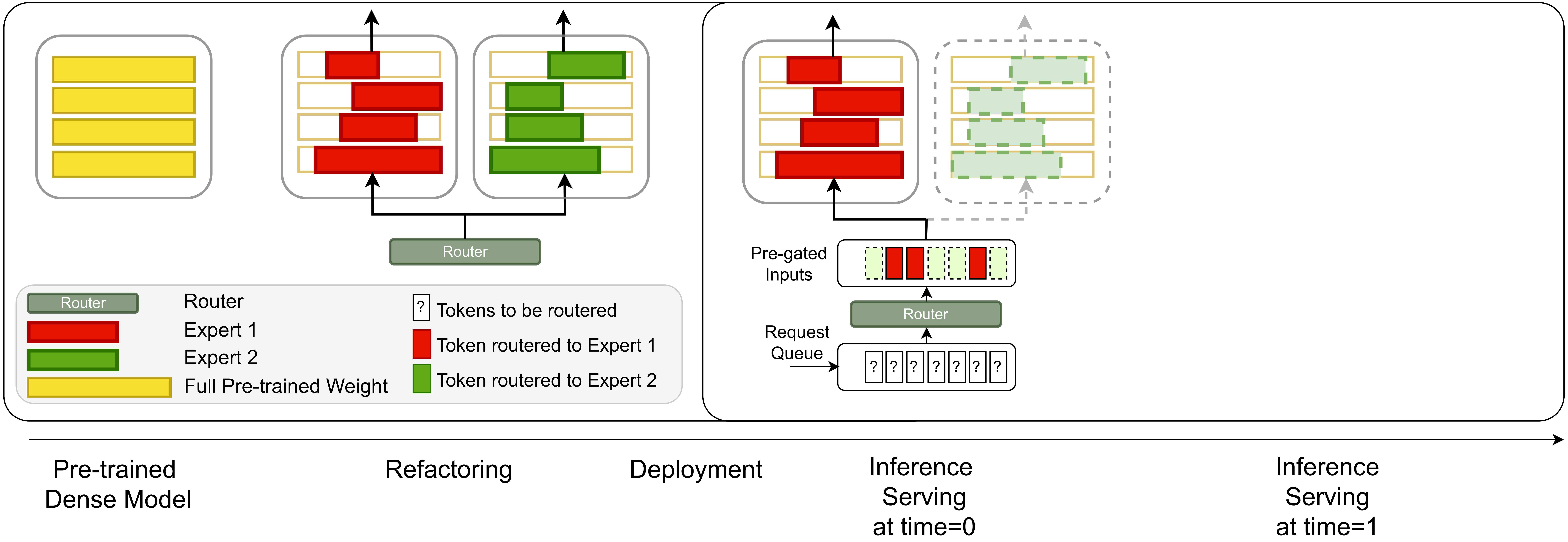
Deployment

Inference Serving at time=0

Inference Serving at time=1

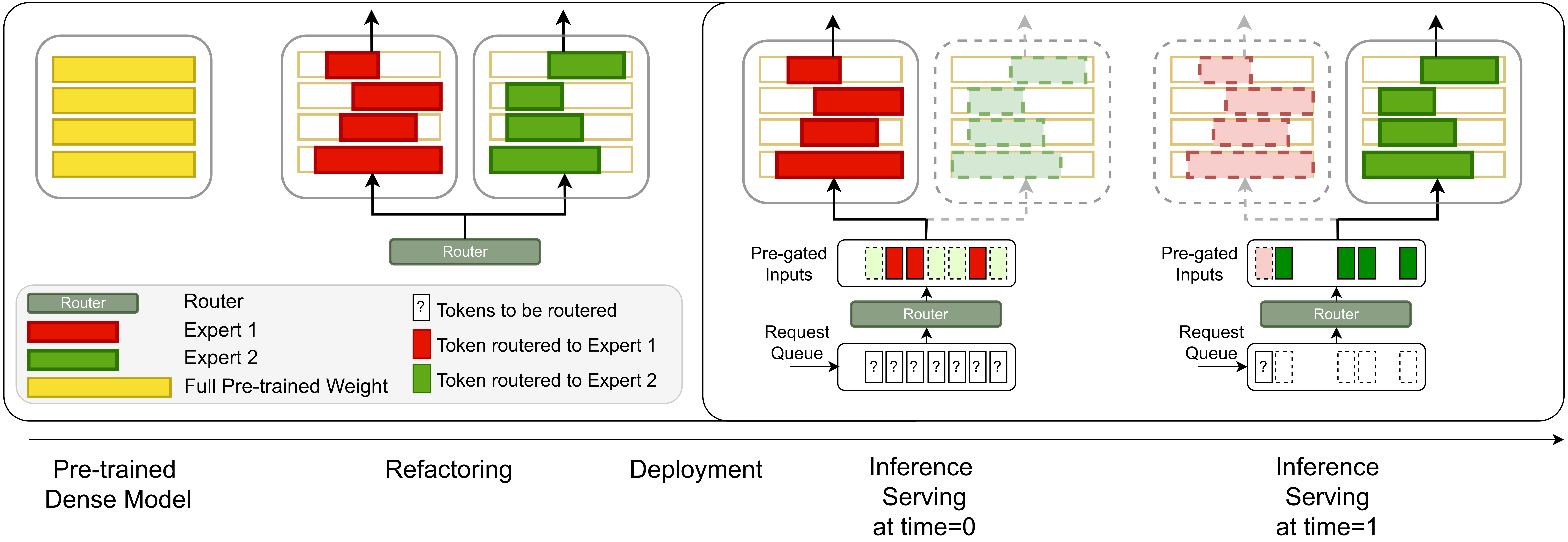
Solution: Decoupled Router

Re-designing MoE Architecture with Decoupled Router



Solution: Decoupled Router

Re-designing MoE Architecture with Decoupled Router



Router

Router — Router

🟥 Expert 1

🟩 Expert 2

Full Pre-trained Weight

? Tokens to be routered

🟥 Token routered to Expert 1

🟩 Token routered to Expert 2

Pre-gated Inputs

Router

Request Queue

Pre-gated Inputs

Router

Request Queue

Pre-trained Dense Model

Refactoring

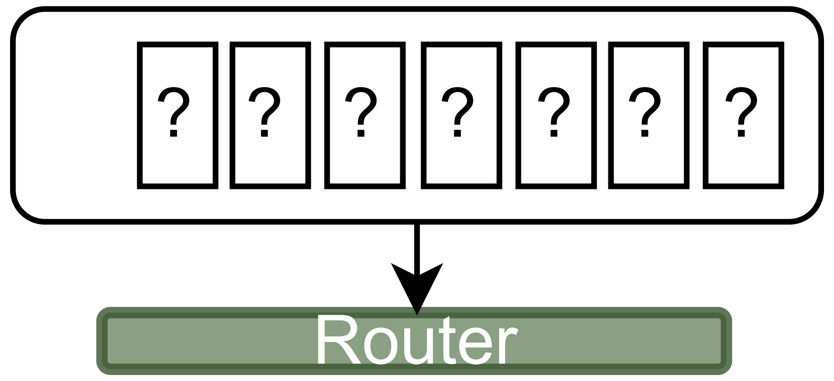
Deployment

Inference Serving at time=0

Inference Serving at time=1

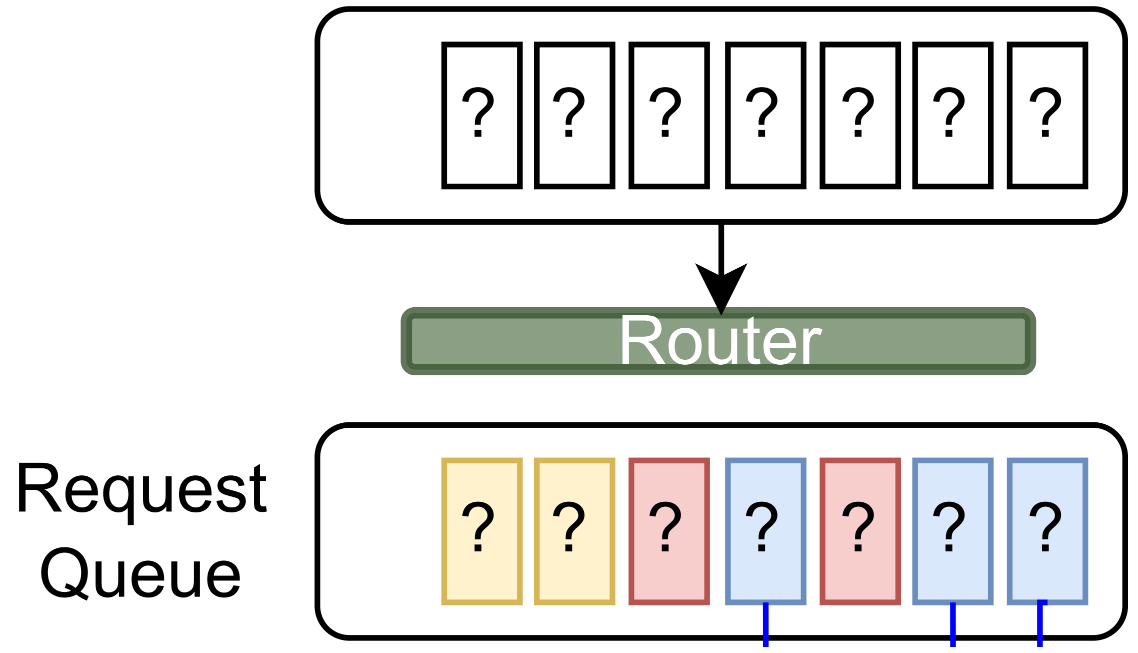
ReadMe: Router Decoupled MoE

Enables Router Pre-computation and Expert-aware Batching



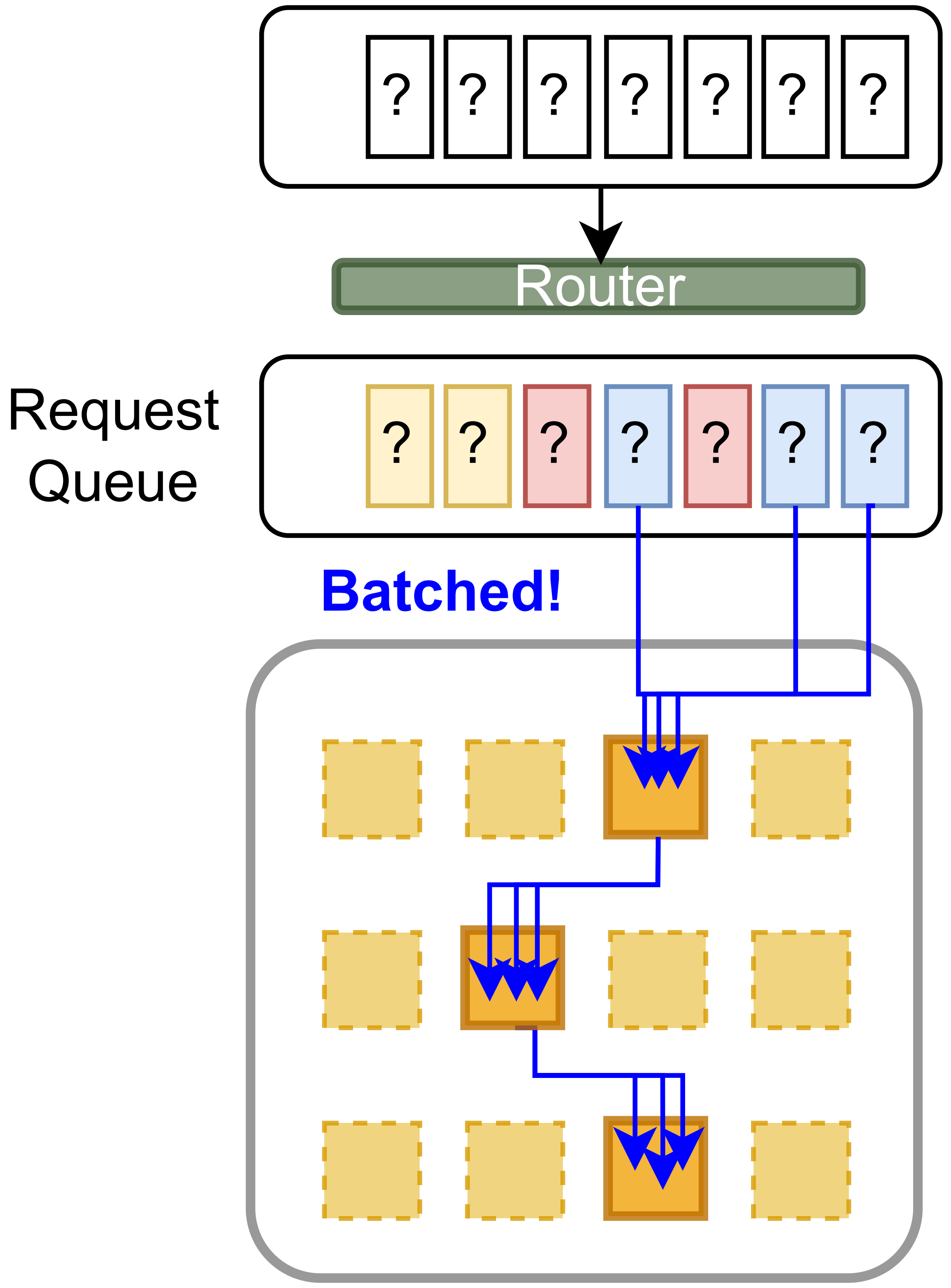
ReadMe: Router Decoupled MoE

Enables Router Pre-computation and Expert-aware Batching



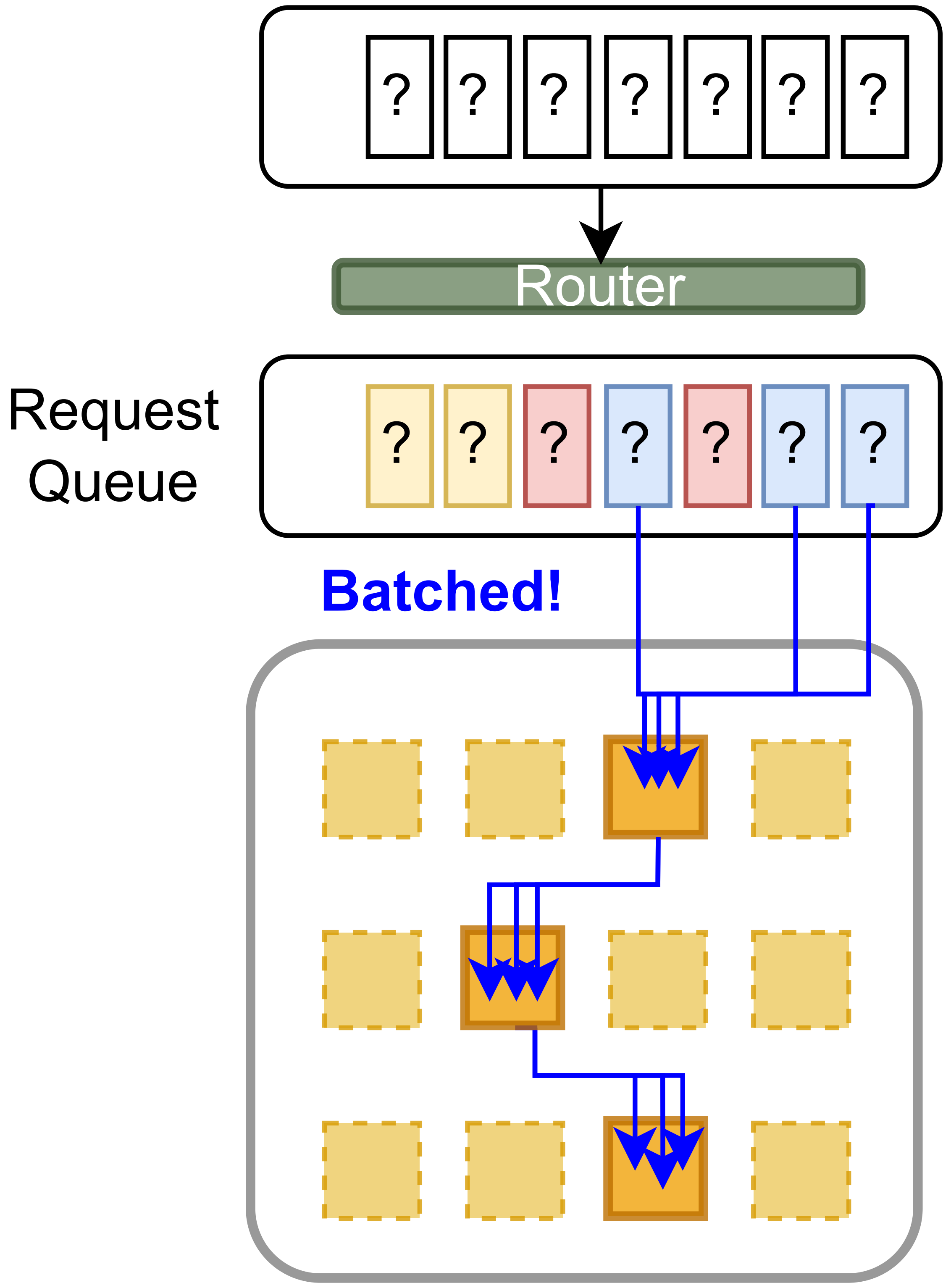
ReadMe: Router Decoupled MoE

Enables Router Pre-computation and Expert-aware Batching



ReadMe: Router Decoupled MoE

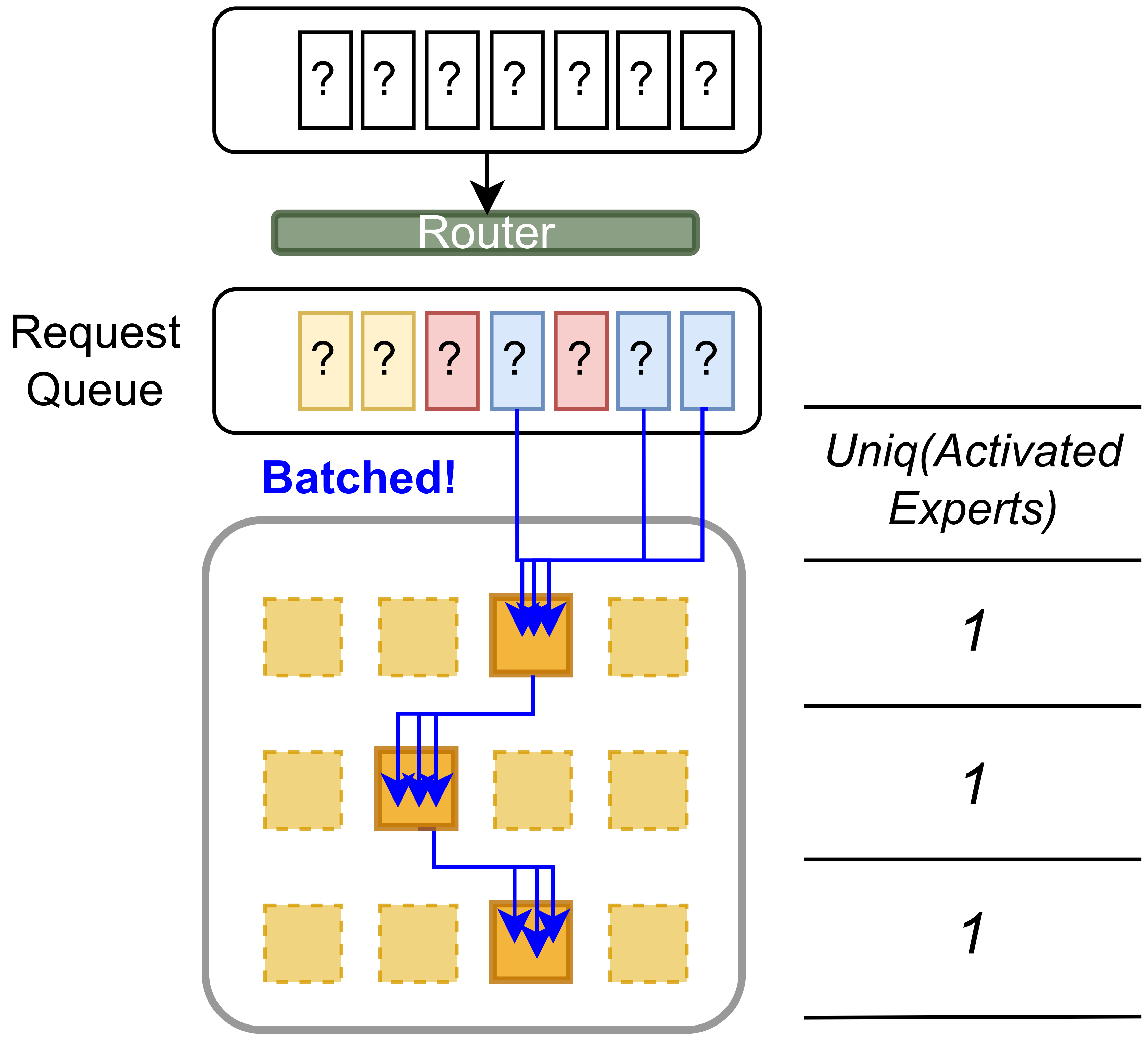
Enables Router Pre-computation and Expert-aware Batching

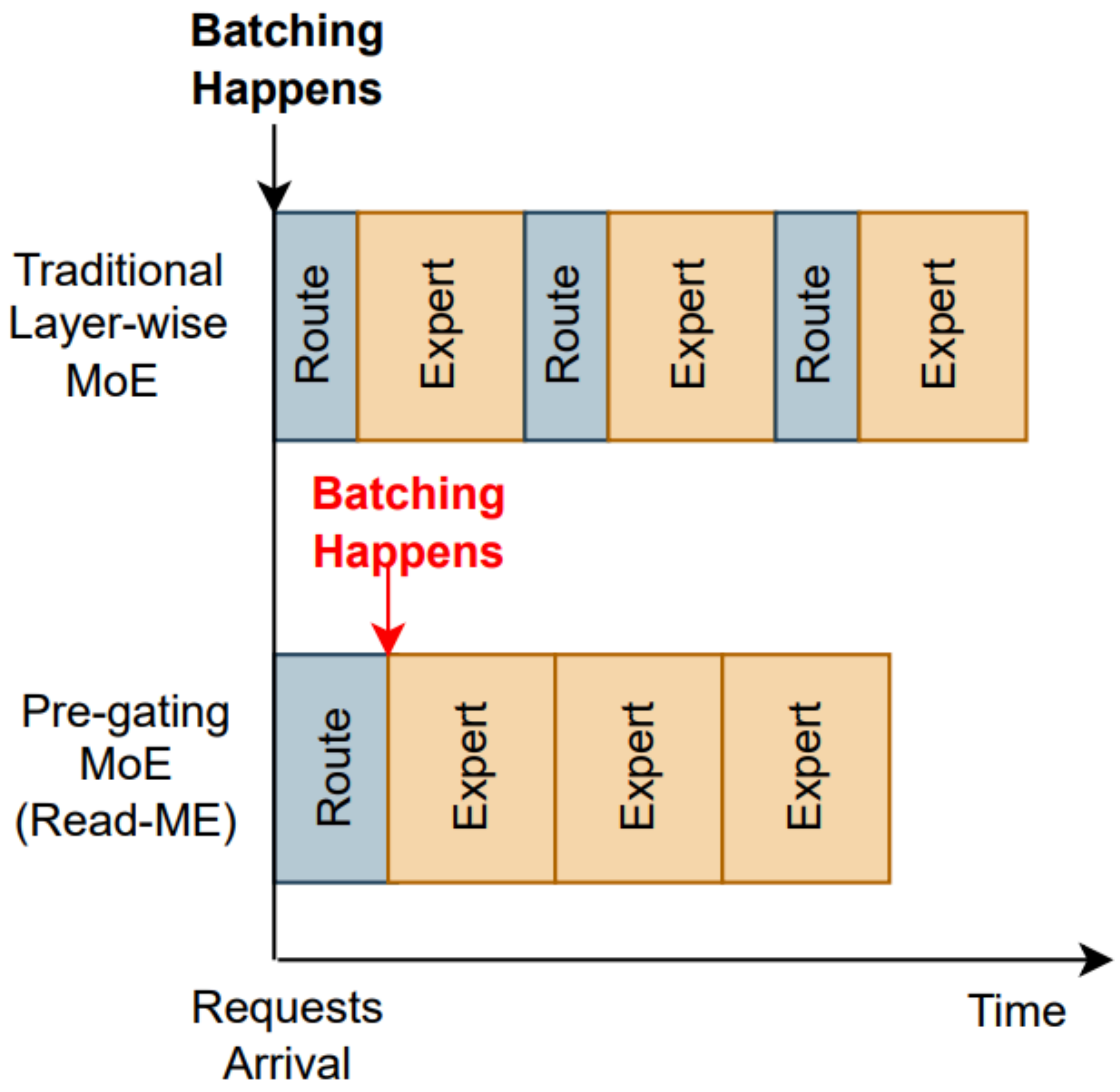


Uniq(Activated Experts)
1
1
1

ReadMe: Router Decoupled MoE

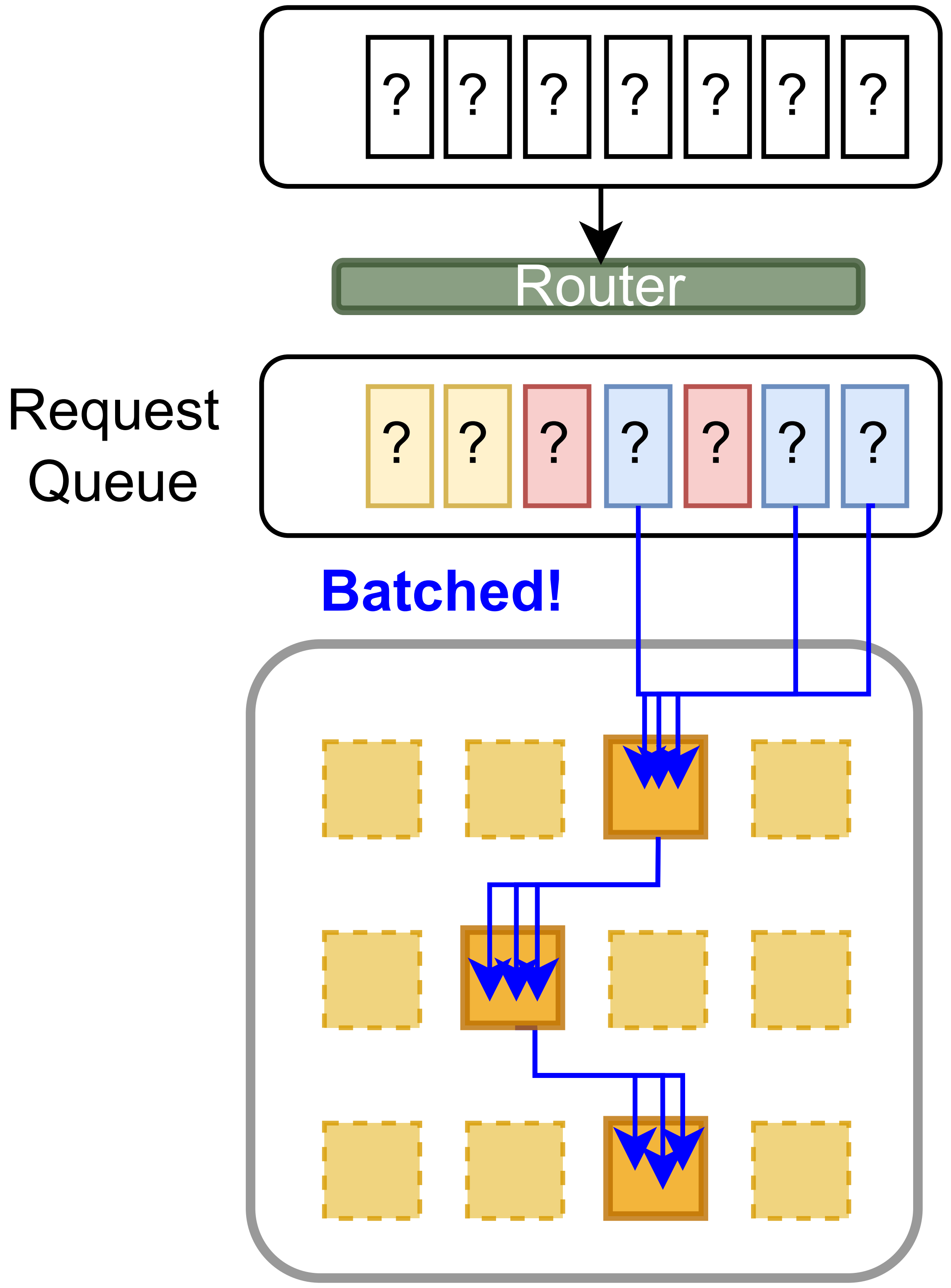
Enables Router Pre-computation and Expert-aware Batching

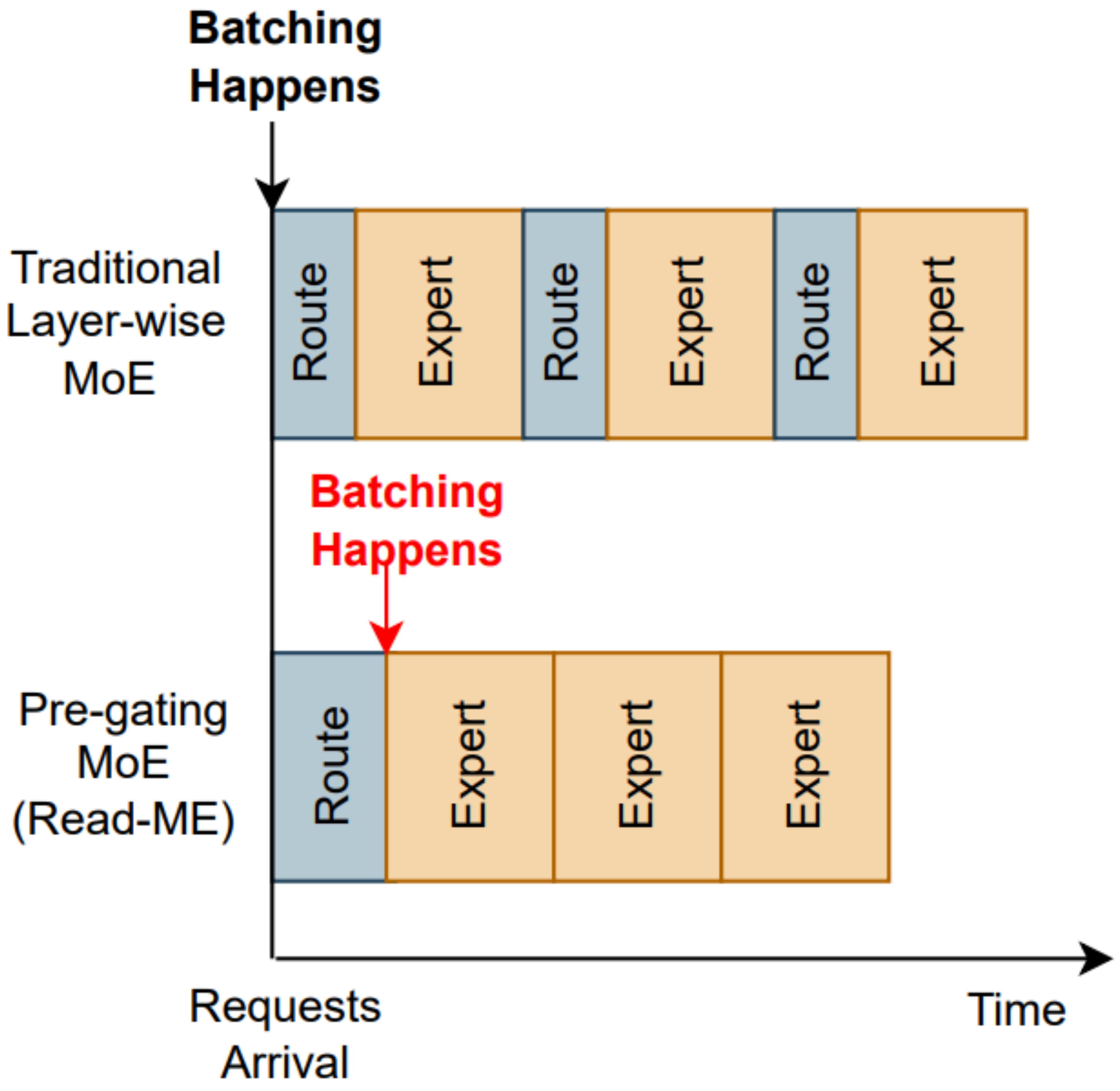




ReadMe: Router Decoupled MoE

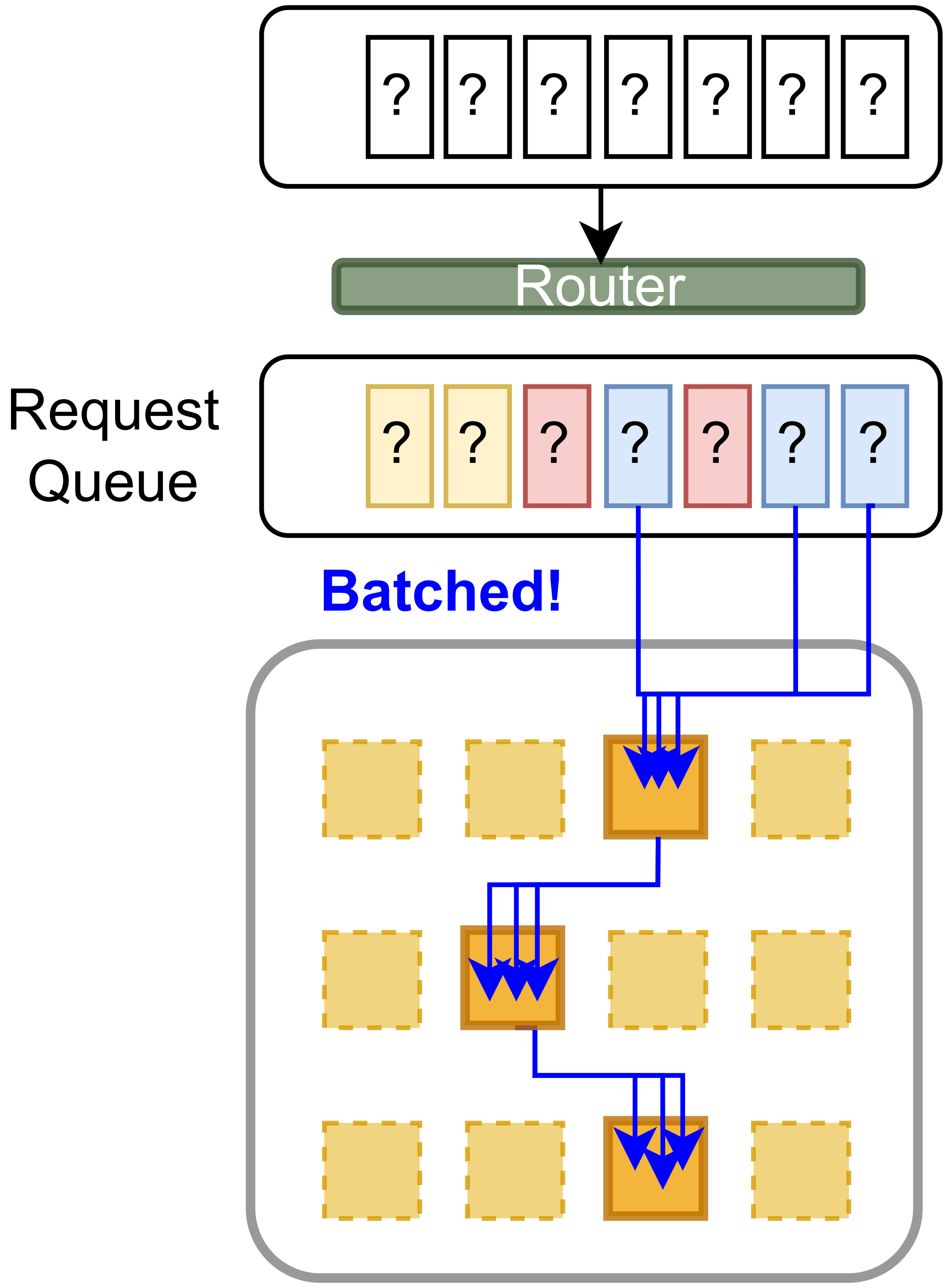
Enables Router Pre-computation and Expert-aware Batching



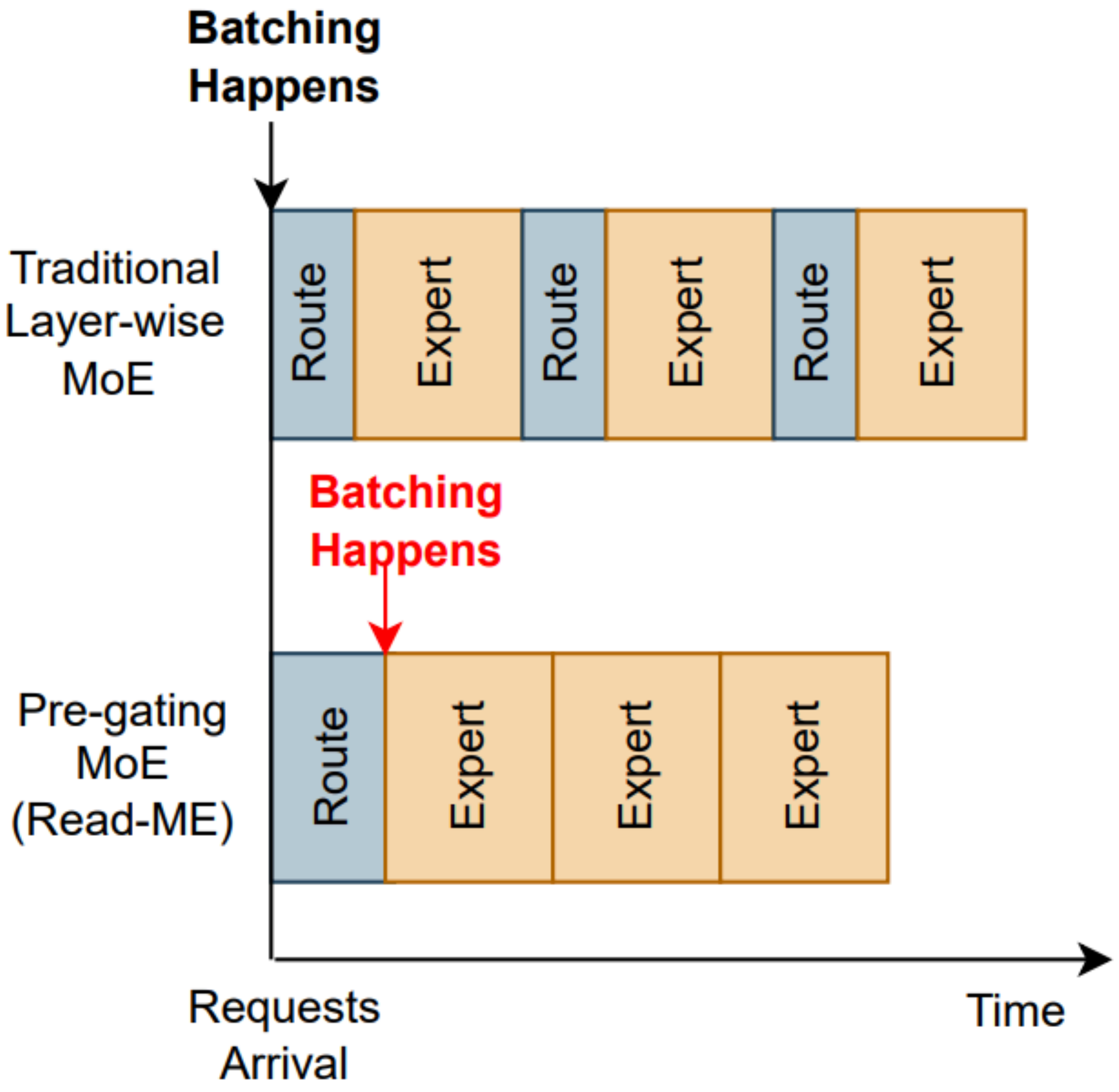


ReadMe: Router Decoupled MoE

Enables Router Pre-computation and Expert-aware Batching

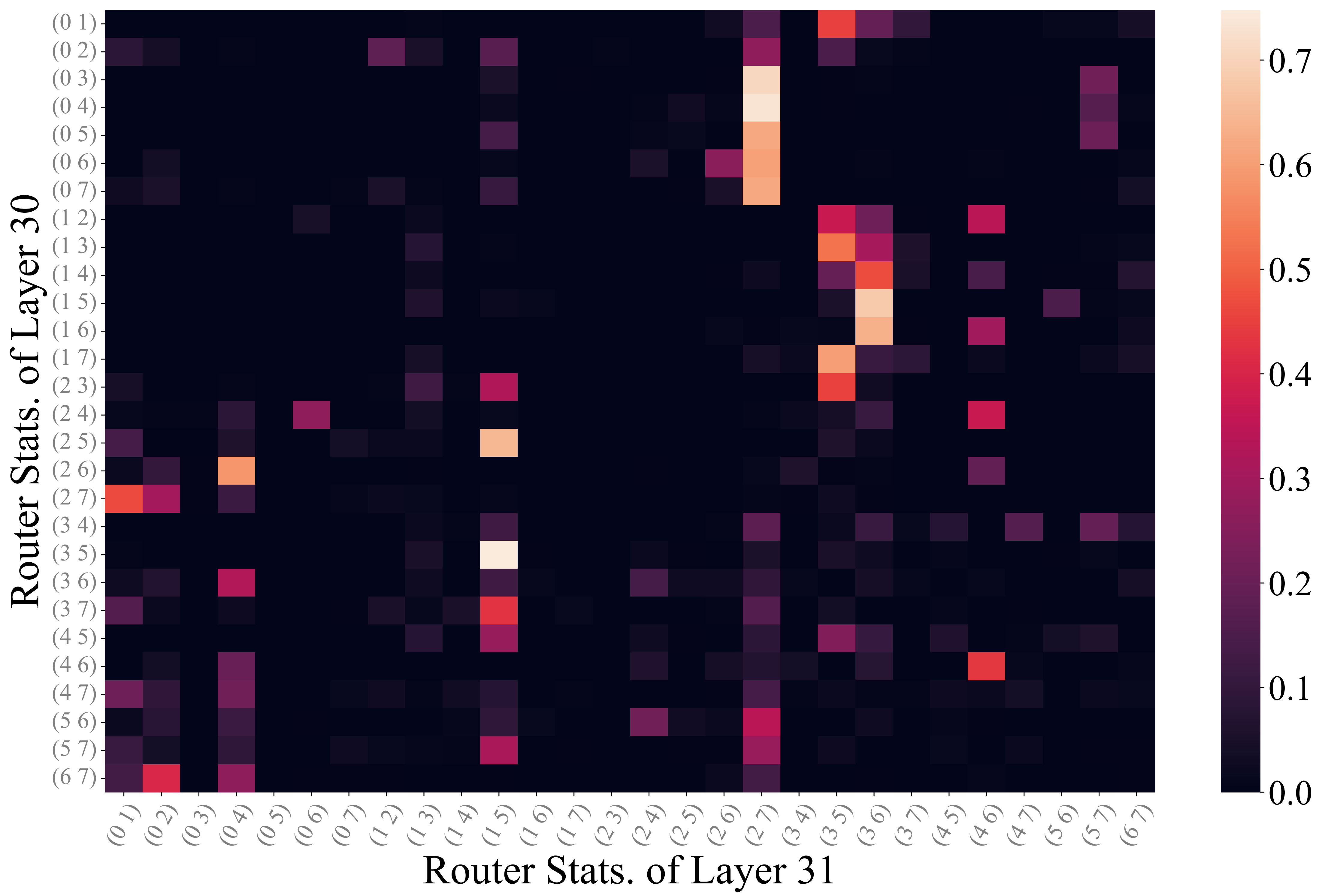


Uniq(Activated Experts)
1
1
1



How is that possible? 😮

The redundancy of Layer-wise Router



Friday, 13 Dec 11AM-2PM
Poster Session 5