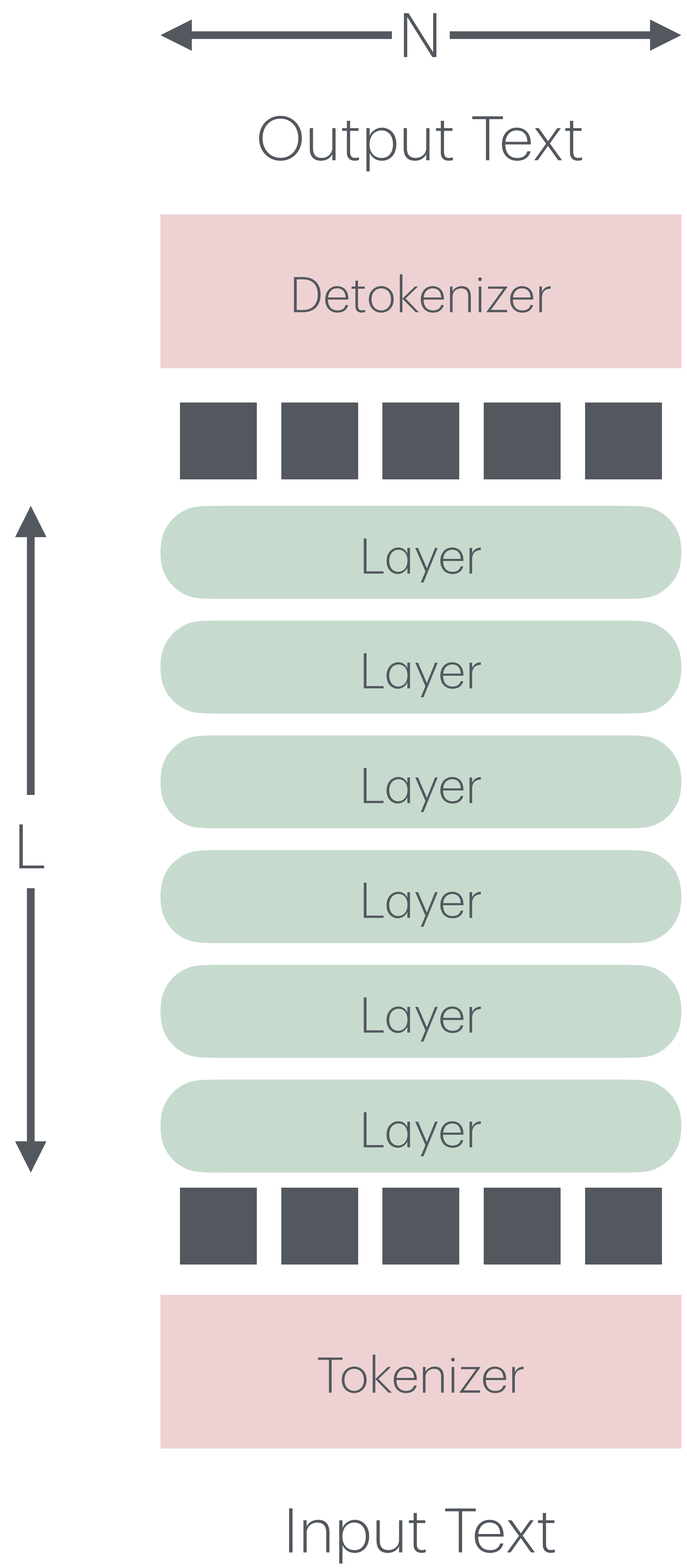
Sequence parallelism

Philipp Krähenbühl, UT Austin

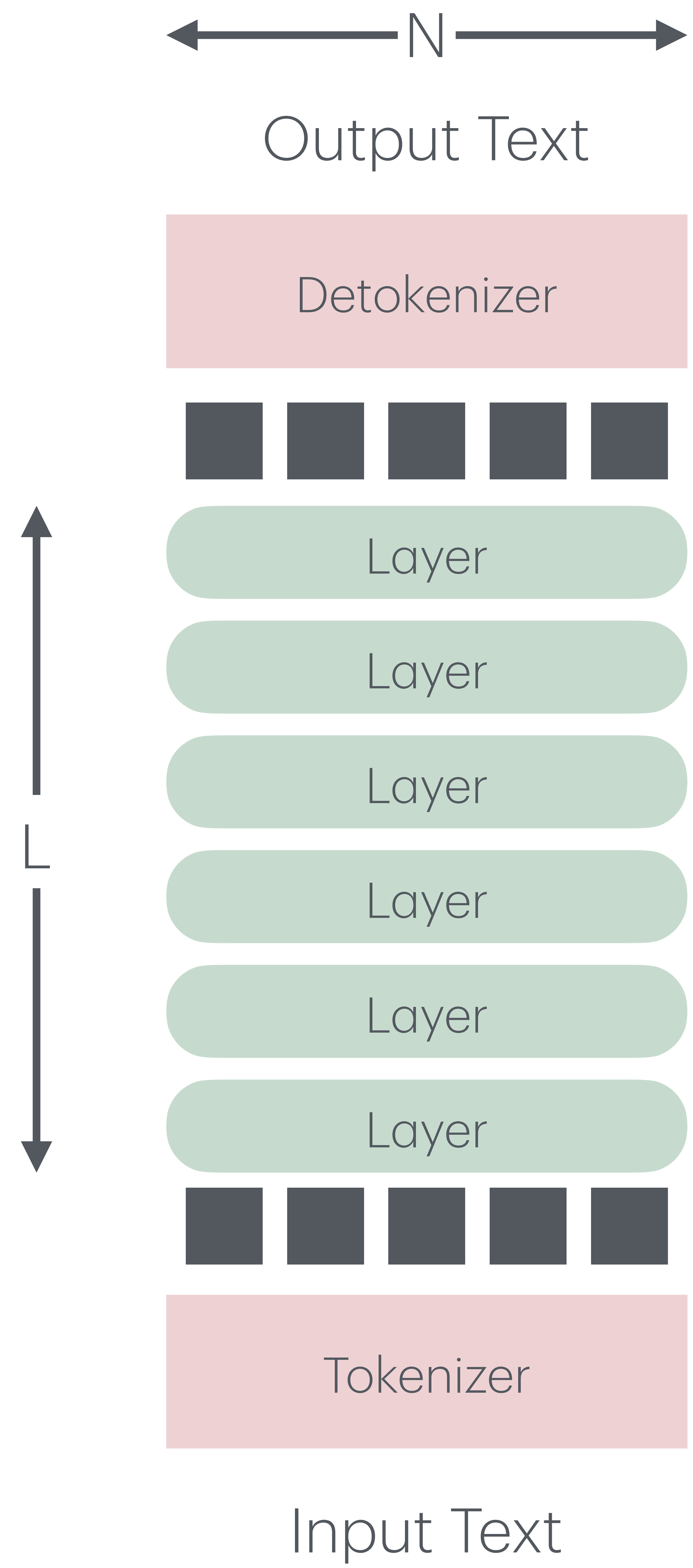
Training and Generation

	Training	Training - Checkpointing	Generation
Peak Memory	O(NL)	O(NL½)	O(N)
Runtime	$O(N^2L)$	$O(2\ N^2L)$	$O(N^3L)$
# forward calls	1	2	N



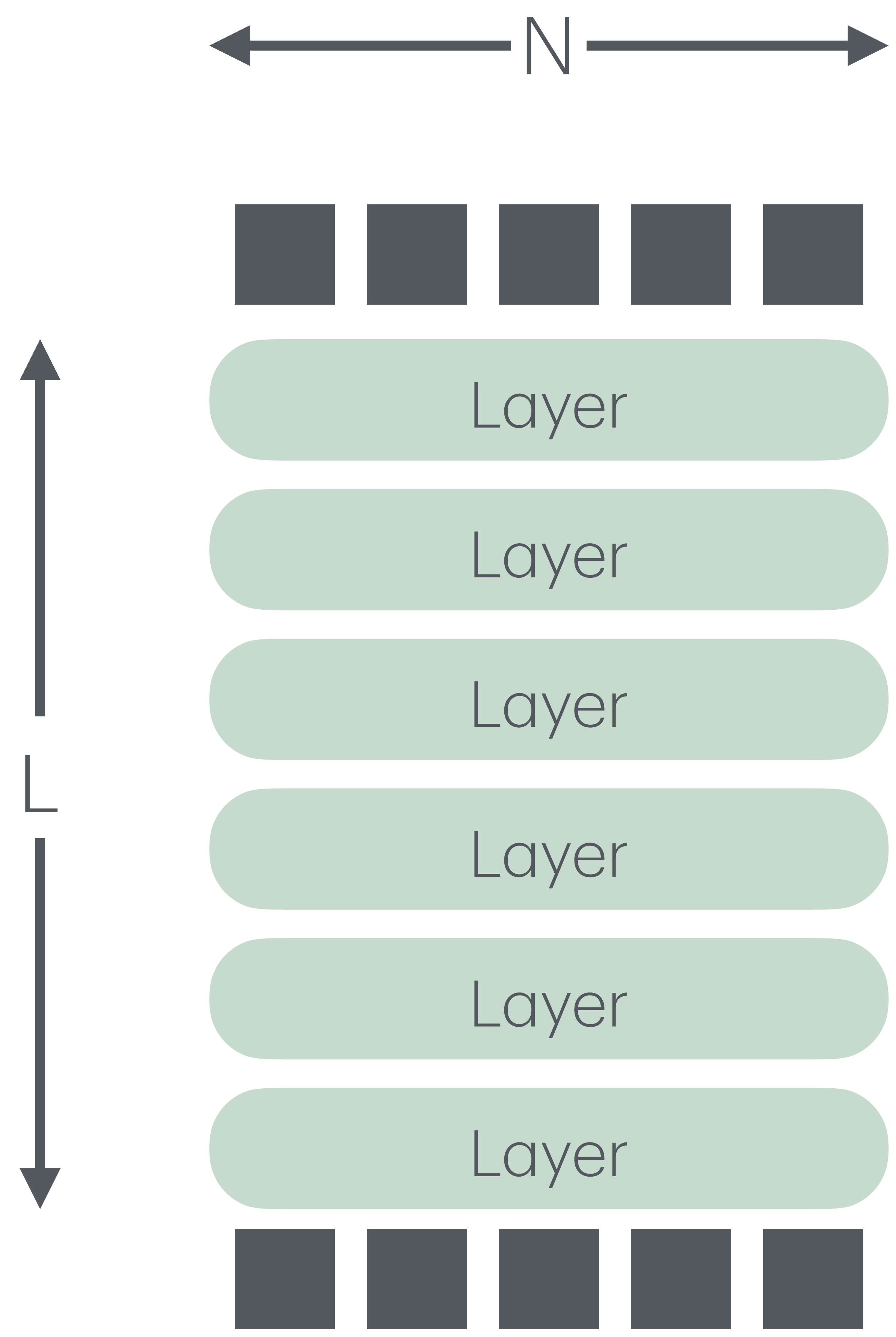
Training

	Training	Training - Checkpointing
Peak Memory	O(NL)	$O(NL^{½})$
Runtime	$O(N^2L)$	$O(2\ N^2L)$
# forward calls	1	2



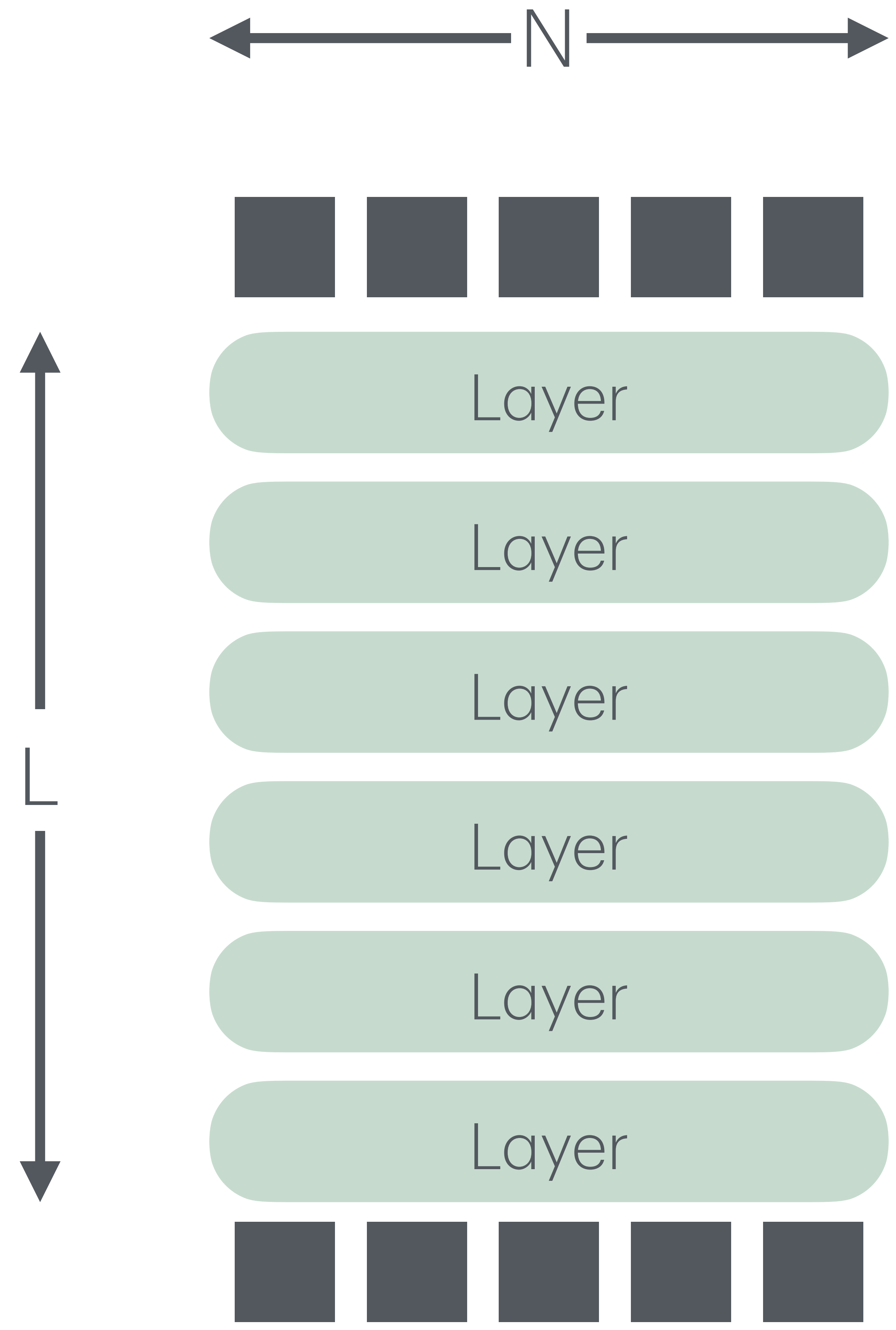
Parallelism

- Data parallel
- Pipeline parallel
- Tensor parallel
- What if the sequence is too long
    - Sequence parallel



Sequence parallel

- Split sequence between GPUs
  - Easy for MLP
  - Medium for LayerNorm
  - Hard for Attention

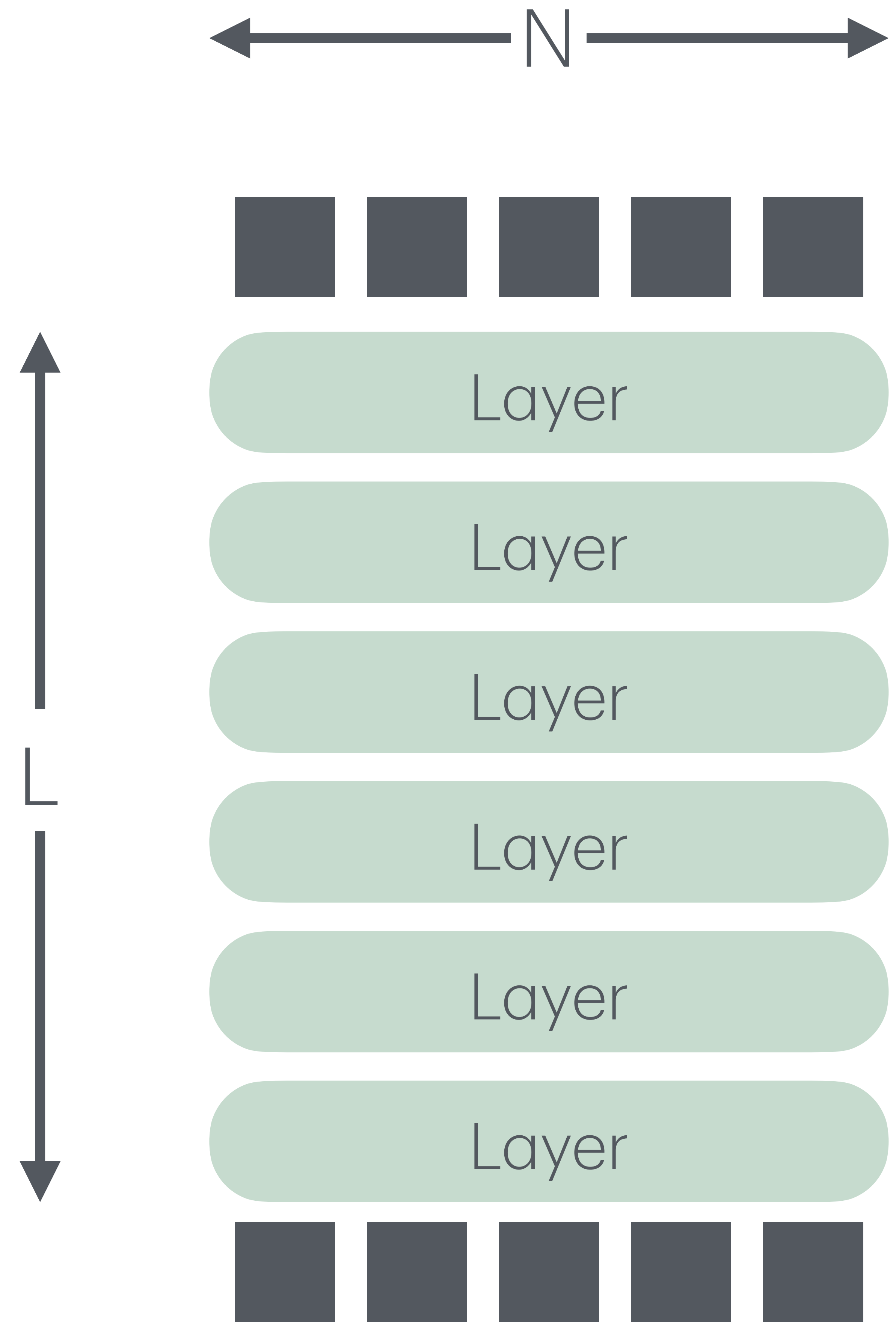


Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

MLP

- Split sequence between GPUs
  - MLP processes tokens independently
  - Trivial parallelization

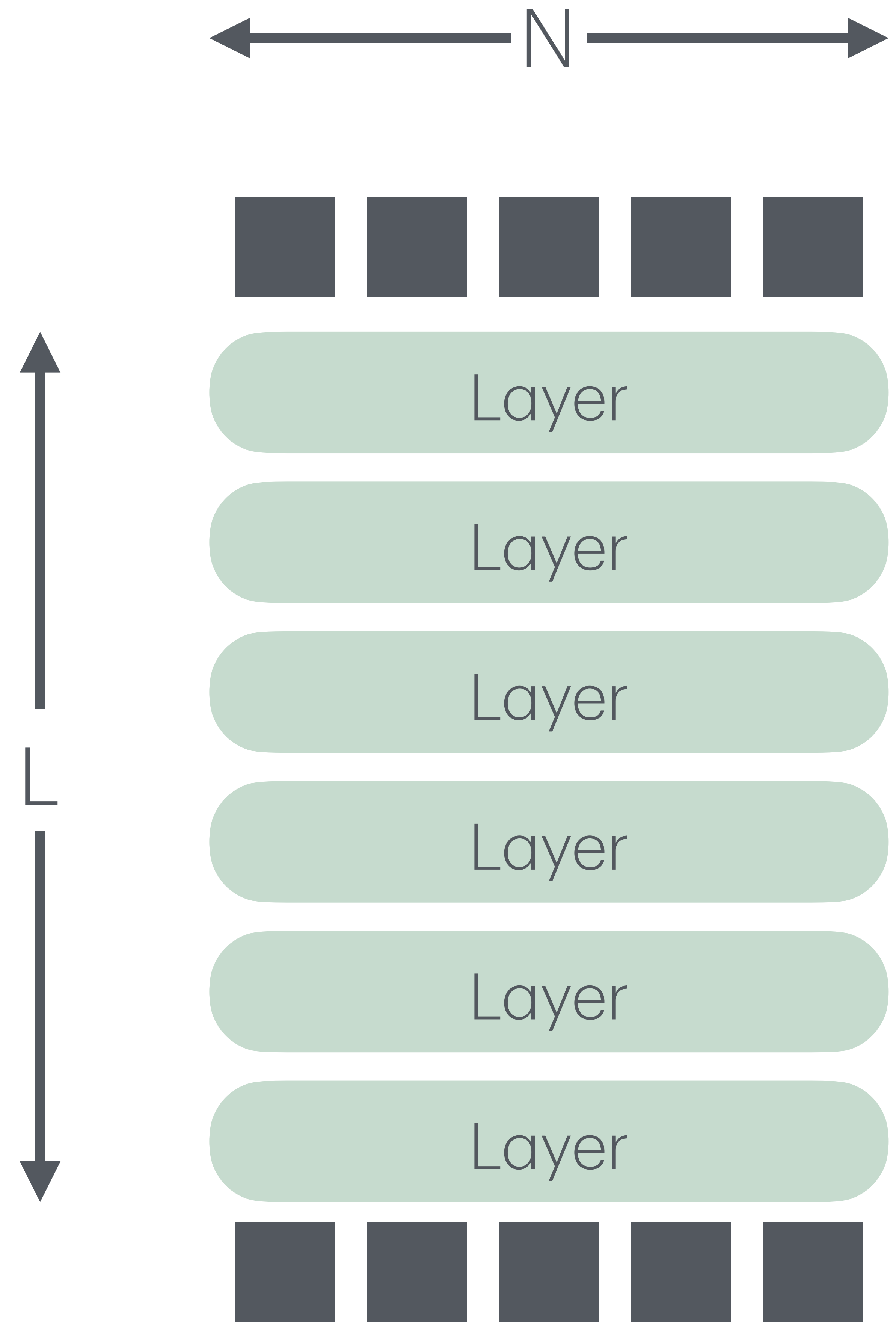


Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

LayerNorm

- LayerNorm computes stats over sequence
  - Requires sync
- Other norms (i.e. RMSNorm) preferred nowadays
  - Applied independently to each token
  - No sync

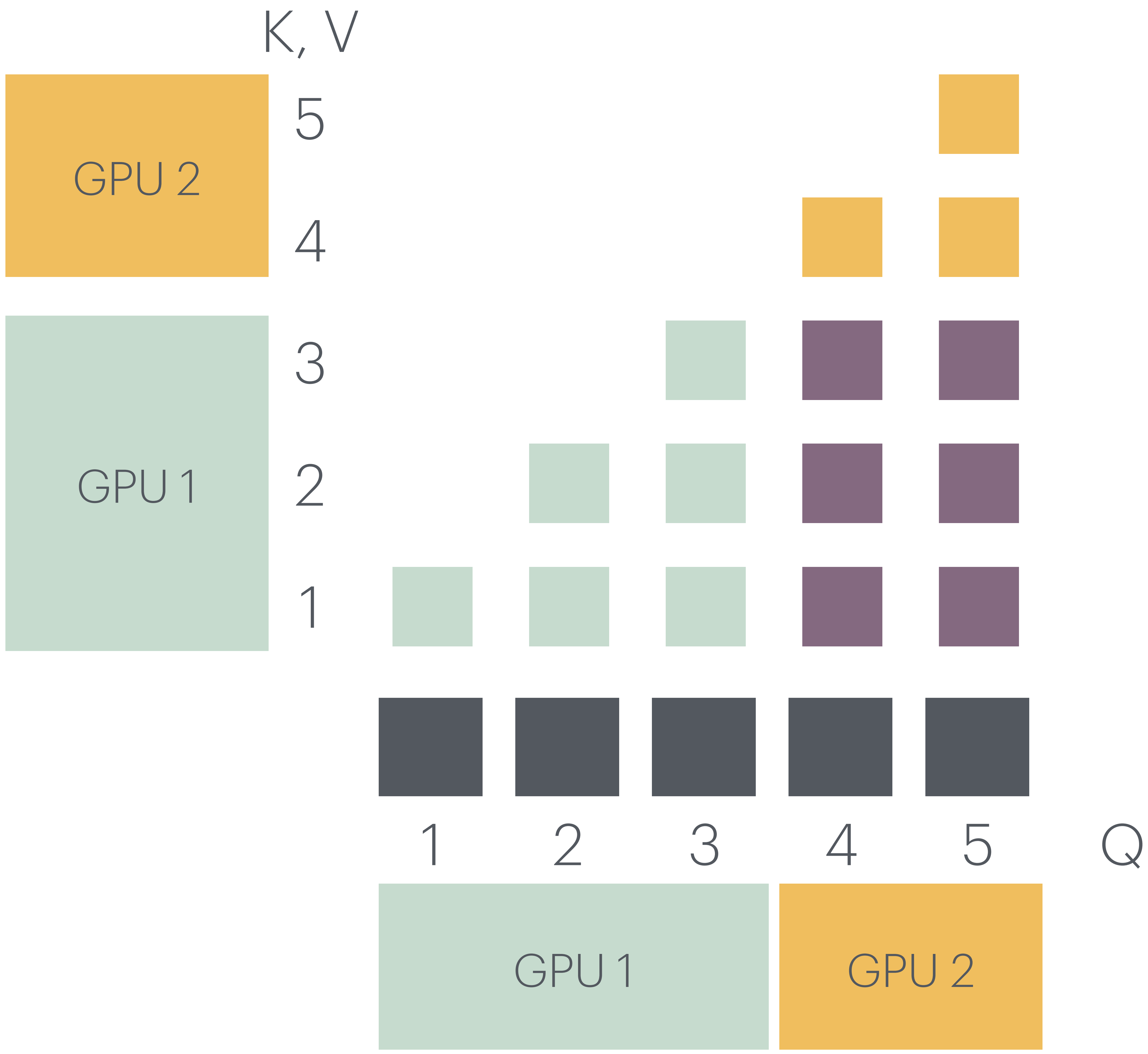


Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

Attention

- Send keys K, value V from GPU to GPU
- Implementation: Ring attention

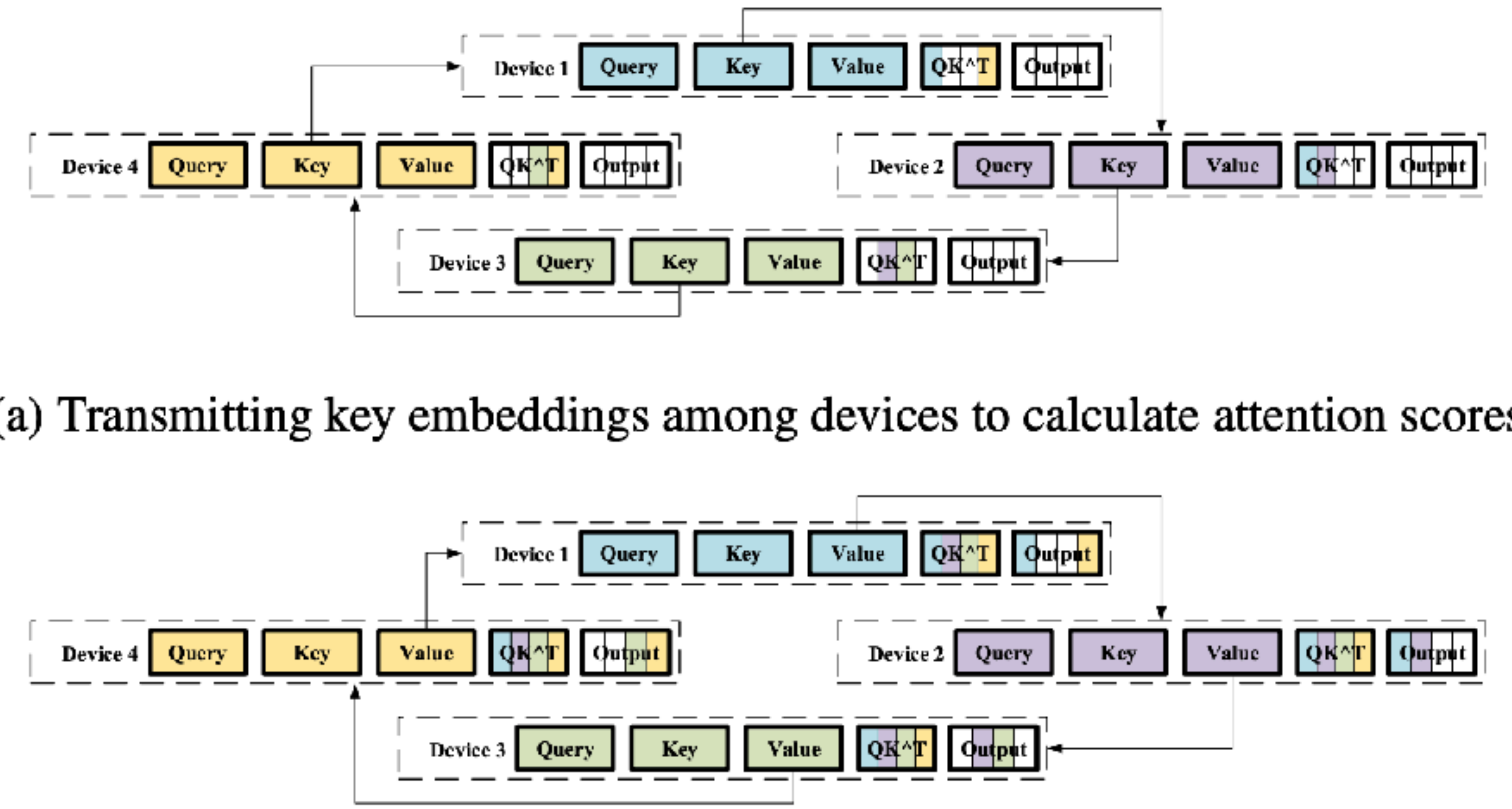


Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

Attention

- Send keys K, value V from GPU to GPU
- Implementation: Ring attention
  - Send keys
  - Compute attention
  - Send values



(a) Transmitting key embeddings among devices to calculate attention scores

(b) Transmitting value embeddings among devices to calculate the output of attention layers

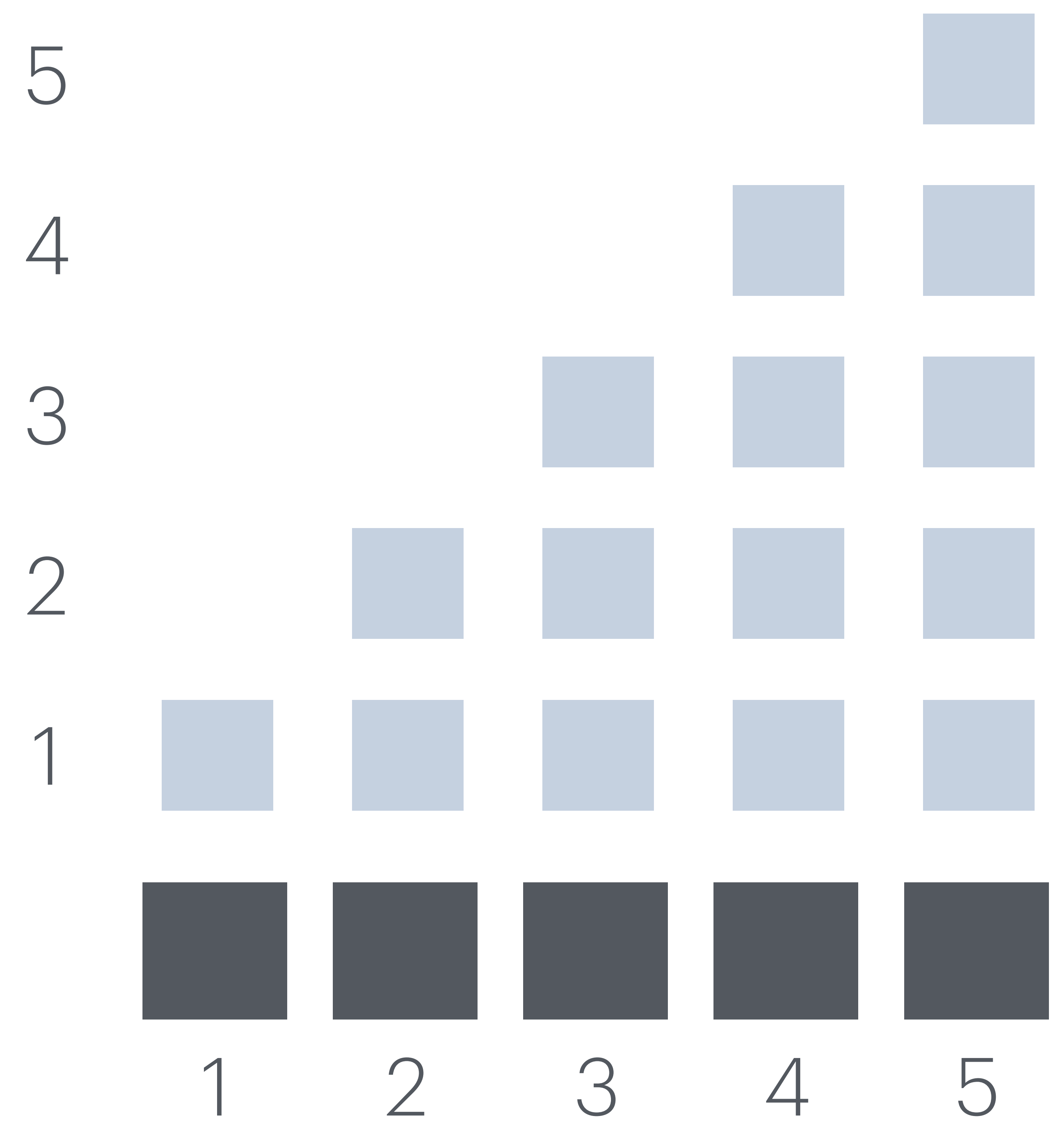
Figure 2: Ring Self-Attention

Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

Causal Attention - Issue

- Unbalanced computation
  - Early tokens attend to fewer tokens
  - Later tokens attend to almost all tokens
  - Idle time

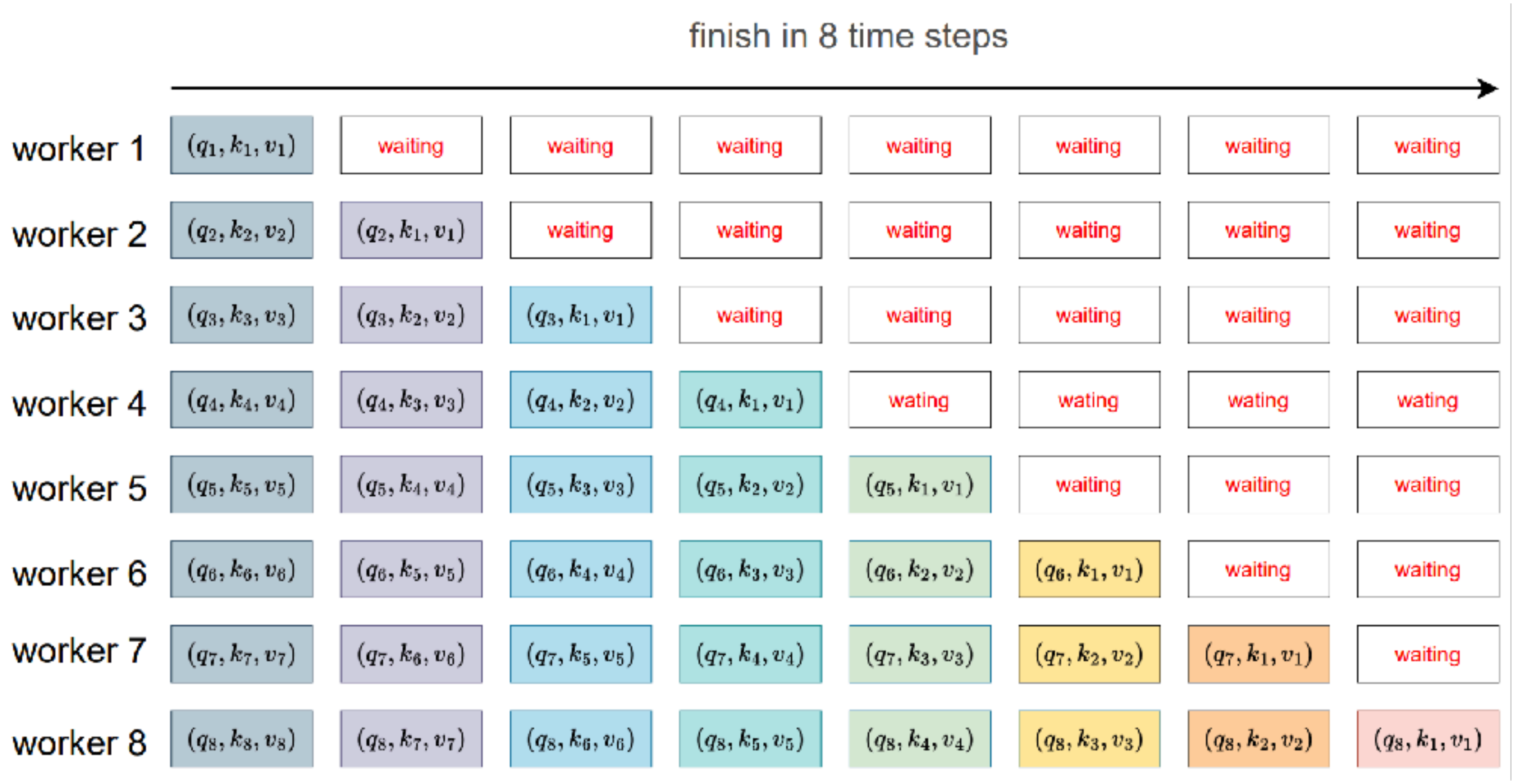


Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

Causal Attention

- DistFlashAttention
  - Use FlashAttention
    - Eliminates sending keys before values
  - Option 1: Send k, v
  - Option 2: Send q, send back result



Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Sequence parallel

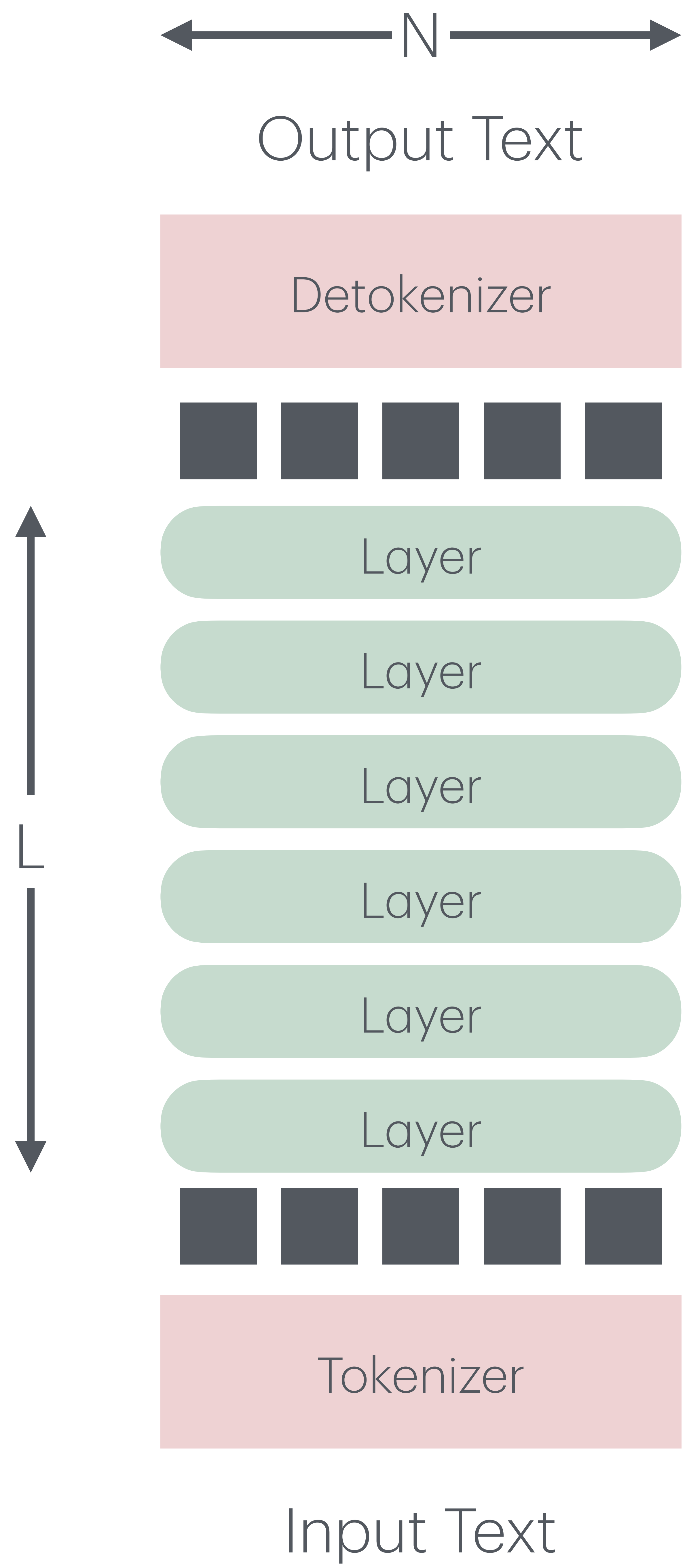
Causal Attention

- DistFlashAttention
  - Use FlashAttention
    - Eliminates sending keys before values
  - Option 1: Send k, v
  - Option 2: Send q, send back result

Sequence Parallelism: Long Sequence Training from System Perspective, Li etal 2021
Reducing Activation Recomputation in Large Transformer Models, Korhikanti etal 2022
DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li etal 2023

Training

	Training	Training - Checkpointing	Training - Checkpointing + Seq P.
Peak Memory	O(NL)	$O(NL^{½})$	$O(NL^{½} / \#GPU)$
Runtime	$O(N^2L)$	$O(2\ N^2L)$	$O(2\ N^2L / \#GPU)$
# forward calls	1	2	2



References

- [1] Sequence Parallelism: Long Sequence Training from System Perspective, Li et al 2021. (link)
- [2] Reducing Activation Recomputation in Large Transformer Models, Korhikanti et al 2022. (link)
- [3] DISTFLASHATTN: Distributed Memory-efficient Attention for Long-context LLMs Training, Li et al 2023. (link)