

MultiDecode

June 7, 2025
Ted Kyi and Roger Stager

MultDecode implementation details

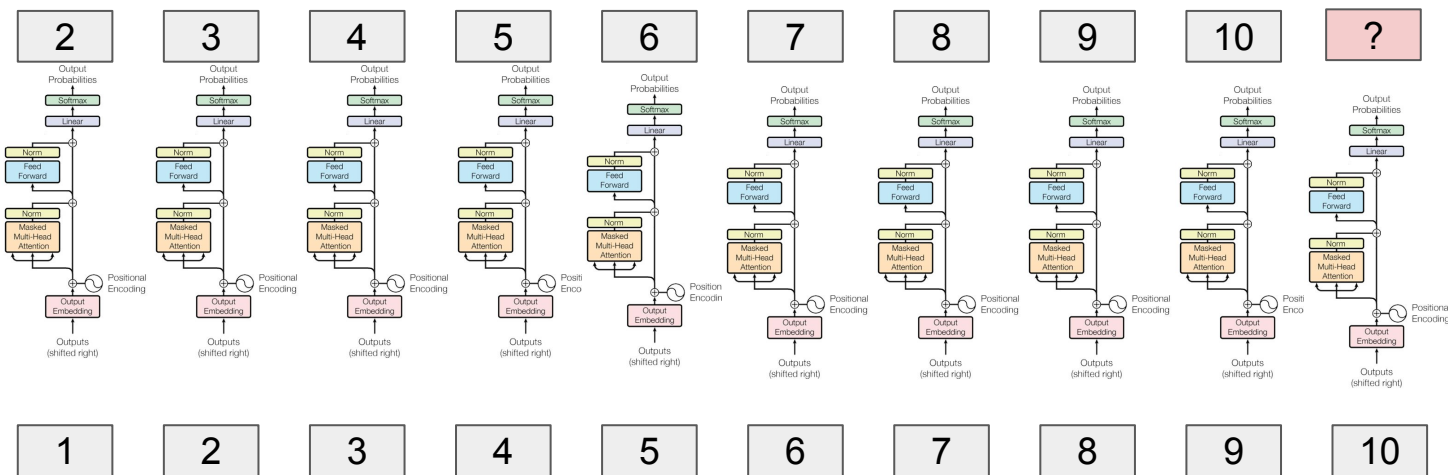
Agenda

- Overview of concepts

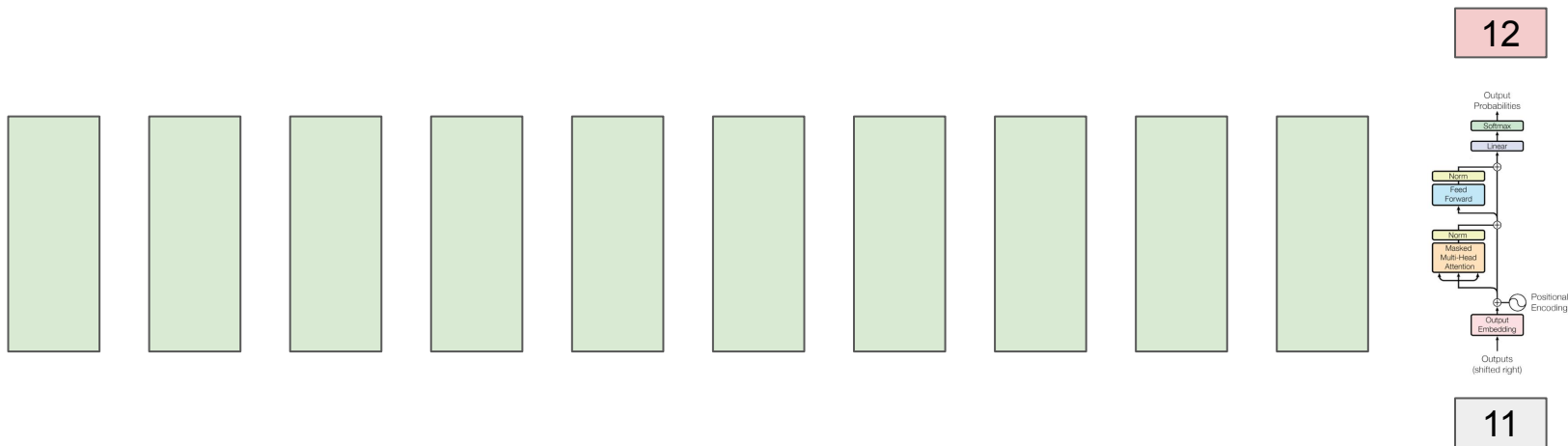
- POC implementation walkthrough

- Performance

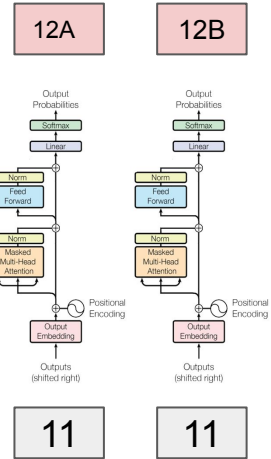
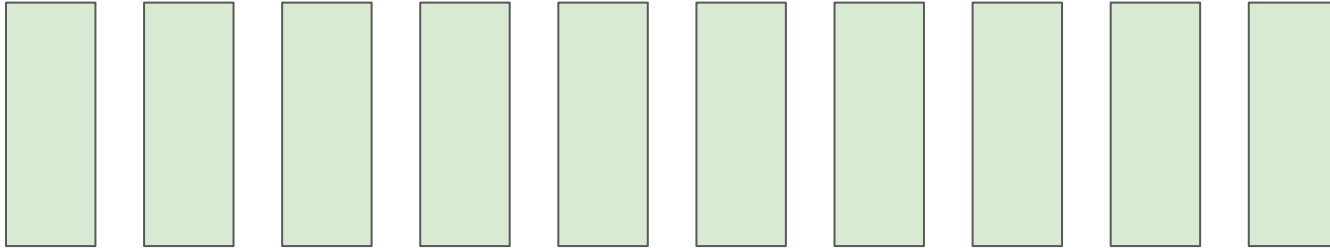
Training vs Inference (generation)



Inference (generation)



Multi Token Decoding



Attention

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

$$\begin{matrix} X \\ \text{3x3 grid} \end{matrix} \times \begin{matrix} W^Q \\ \text{3x4 grid} \end{matrix} = \begin{matrix} Q \\ \text{3x4 grid} \end{matrix}$$

$$\begin{matrix} X \\ \text{3x3 grid} \end{matrix} \times \begin{matrix} W^K \\ \text{3x4 grid} \end{matrix} = \begin{matrix} K \\ \text{3x4 grid} \end{matrix}$$

$$\begin{matrix} X \\ \text{3x3 grid} \end{matrix} \times \begin{matrix} W^V \\ \text{3x4 grid} \end{matrix} = \begin{matrix} V \\ \text{3x4 grid} \end{matrix}$$

$$\text{softmax} \left(\frac{\begin{matrix} Q \\ \text{3x4 grid} \end{matrix} \times \begin{matrix} K^T \\ \text{4x3 grid} \end{matrix}}{\sqrt{d_k}} \right) \begin{matrix} V \\ \text{3x4 grid} \end{matrix}$$

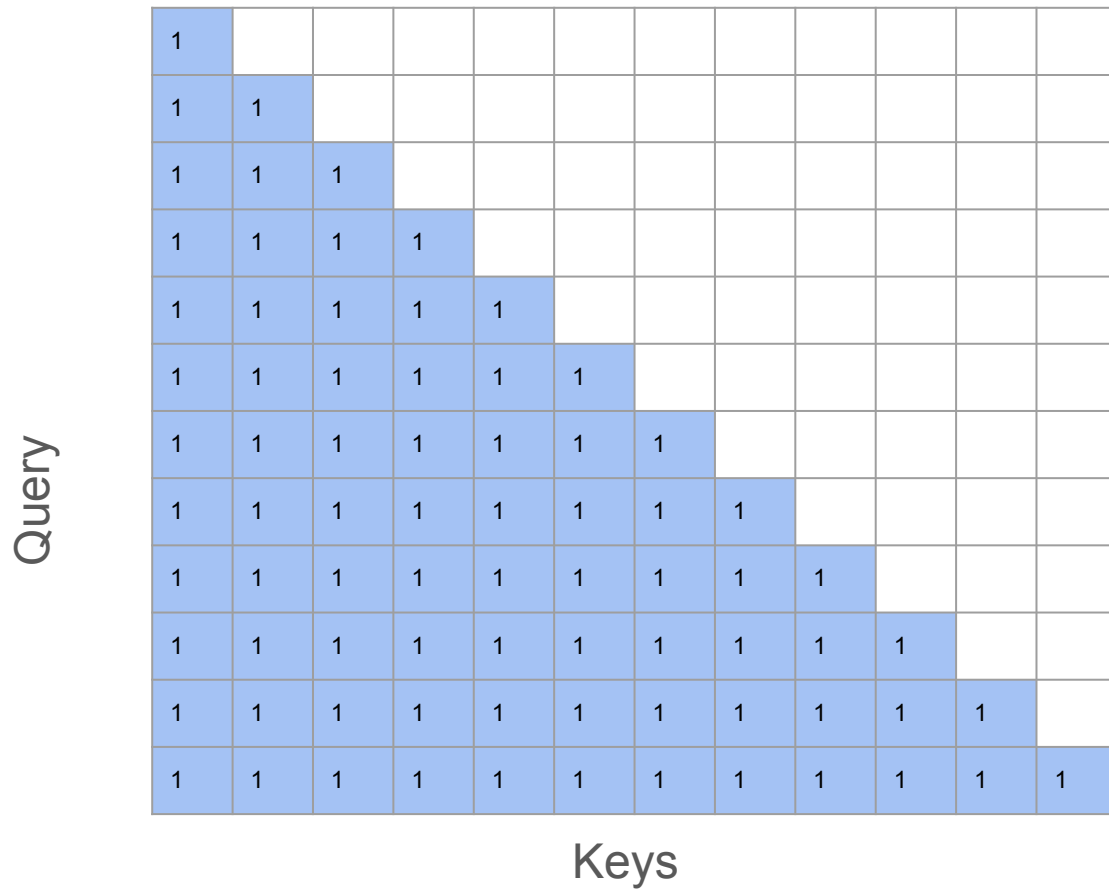
Calculate Q KV for full sequence

[illegible][illegible][illegible]

Attention: Outer product between Query and Key

Query	Q	Q1K1	Q1K2	Q1K3	Q1K4	Q1K5	Q1K6	Q1K7	Q1K8	Q1K9	Q1K10	Q1K11	Q1K12
	Q	Q2K1	Q9K2	Q1K3	Q2K4	Q8K5	Q1K6	Q1K7	Q1K8	Q1K9	Q1K10	Q1K11	Q1K12
	Q	Q3K1	Q9K2	Q1K3	Q3K4	Q3K5	Q3K6	Q3K7	Q3K8	Q3K9	Q3K10	Q3K11	Q3K12
	Q	Q4K1	Q9K2	Q1K3	Q4K4	Q4K5	Q4K6	Q4K7	Q4K8	Q4K9	Q4K10	Q4K11	Q4K12
	Q	Q5K1	Q9K2	Q1K3	Q5K4	Q5K5	Q5K6	Q5K7	Q5K8	Q5K9	Q5K10	Q5K11	Q5K12
	Q	Q6K1	Q9K2	Q1K3	Q6K4	Q6K5	Q6K6	Q6K7	Q6K8	Q6K9	Q6K10	Q6K11	Q6K12
	Q	Q7K1	Q9K2	Q1K3	Q7K4	Q7K5	Q7K6	Q7K7	Q7K8	Q7K9	Q7K10	Q7K11	Q7K12
	Q	Q8K1	Q9K2	Q1K3	Q8K4	Q8K5	Q8K6	Q8K7	Q8K8	Q8K9	Q8K10	Q8K11	Q8K12
	Q	Q9K1	Q9K2	Q1K3	Q9K4	Q9K5	Q9K6	Q9K7	Q9K8	Q9K9	Q9K10	Q9K11	Q9K12
	Q	Q10K1	Q10K2	Q1K3	Q10K4	Q10K5	Q10K6	Q10K7	Q10K8	Q10K9	Q10K10	Q10K11	Q10K12
	Q	Q10K1	Q10K2	Q1K3	Q11K4	Q11K5	Q11K6	Q11K7	Q11K8	Q11K9	Q11K10	Q11K11	Q11K12
	Q	Q8K12	Q10K2	Q1K13	Q12K4	Q12K5	Q12K6	Q12K7	Q12K8	Q12K9	Q12K10	Q12K11	Q12K12
	Keys												

Causal attention mask



Query

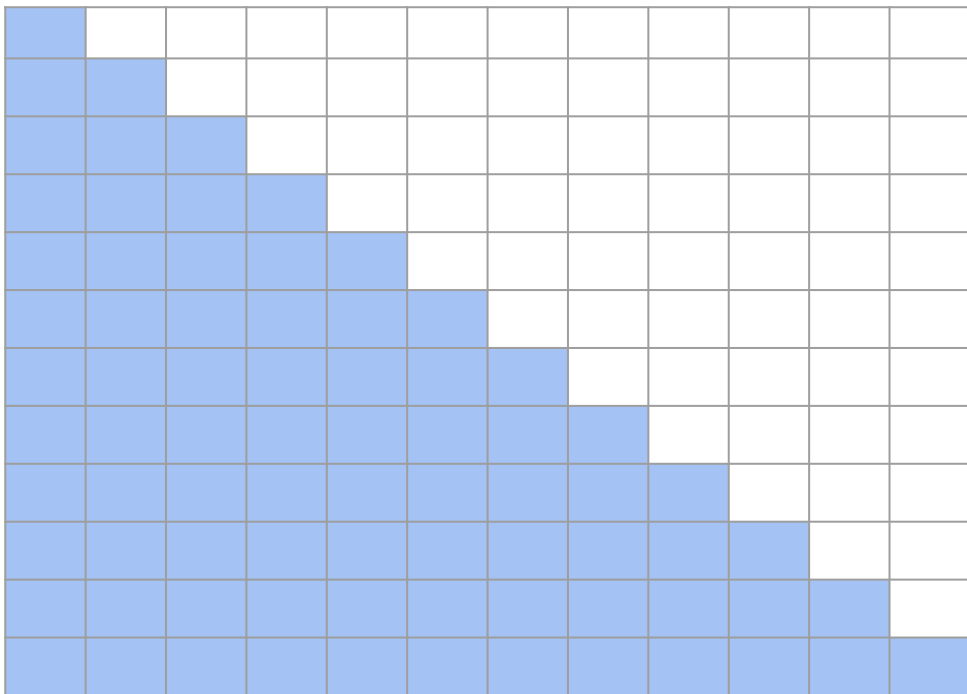
Keys

After Causal mask

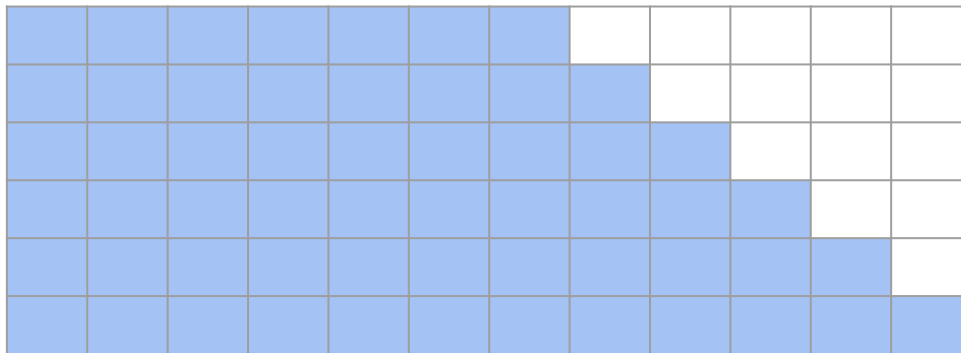
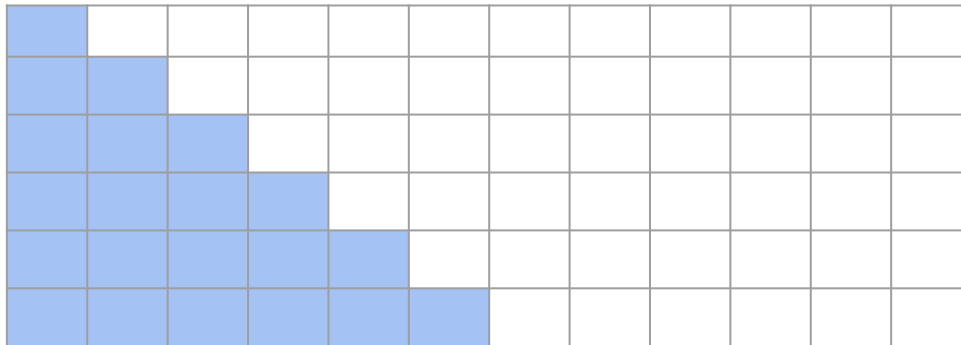
Q1K1											
Q2K1	Q9K2										
Q3K1	Q9K2	Q1K3									
Q4K1	Q9K2	Q1K3	Q4K4								
Q5K1	Q9K2	Q1K3	Q5K4	Q5K5							
Q6K1	Q9K2	Q1K3	Q6K4	Q6K5	Q6K6						
Q7K1	Q9K2	Q1K3	Q7K4	Q7K5	Q7K6	Q7K7					
Q8K1	Q9K2	Q1K3	Q8K4	Q8K5	Q8K6	Q8K7	Q8K8				
Q9K1	Q9K2	Q1K3	Q9K4	Q9K5	Q9K6	Q9K7	Q9K8	Q9K9			
Q10K1	Q10K2	Q1K3	Q10K4	Q10K5	Q10K6	Q10K7	Q10K8	Q10K9	Q10K10		
Q10K1	Q10K2	Q1K3	Q11K4	Q11K5	Q11K6	Q11K7	Q11K8	Q11K9	Q11K10	Q11K11	
Q8K12	Q10K2	Q1K13	Q12K4	Q12K5	Q12K6	Q12K7	Q12K8	Q12K9	Q12K10	Q12K11	Q12K12

[illegible]

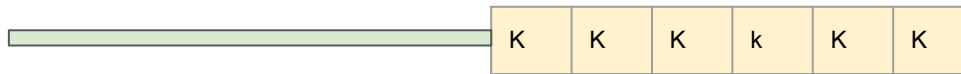
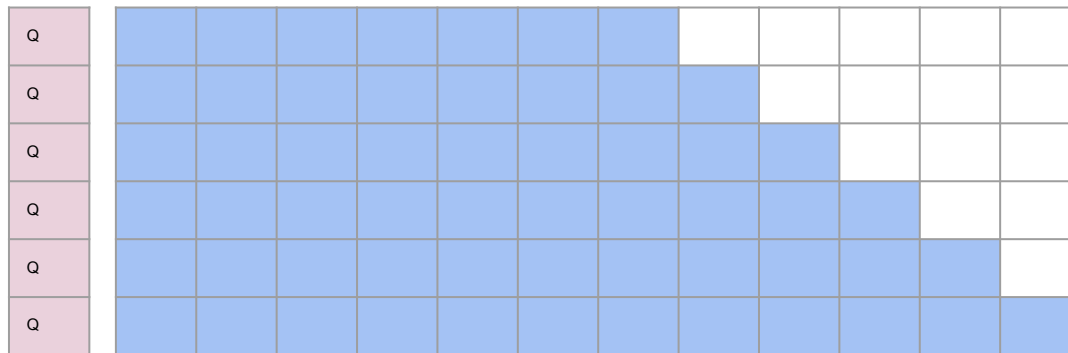
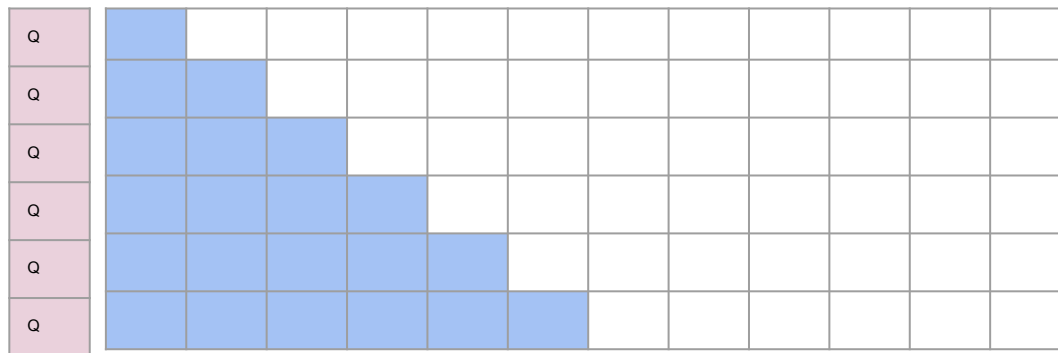
This is the $O(N^2)$



Chunk prefill



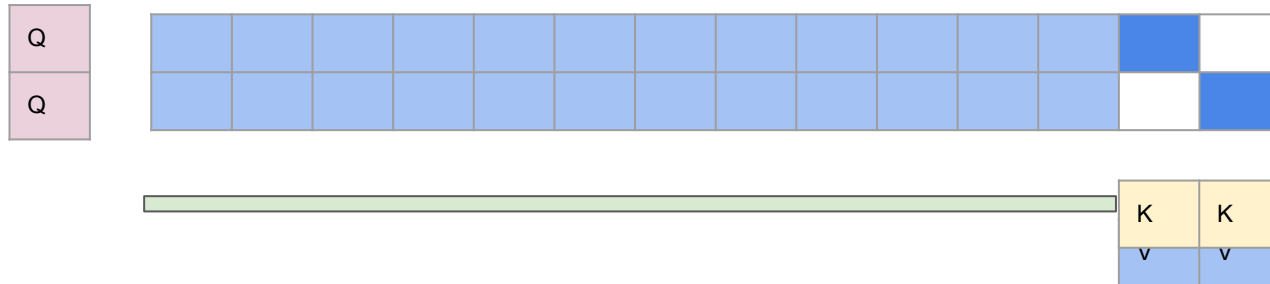
Chunk prefill



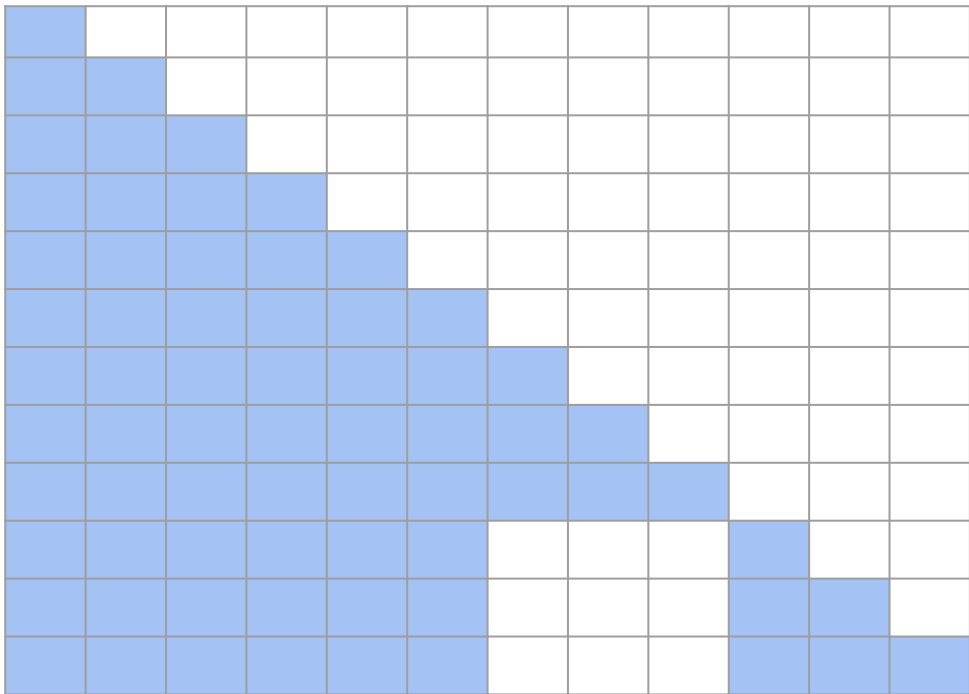
Classic generation



MultiDecode generation

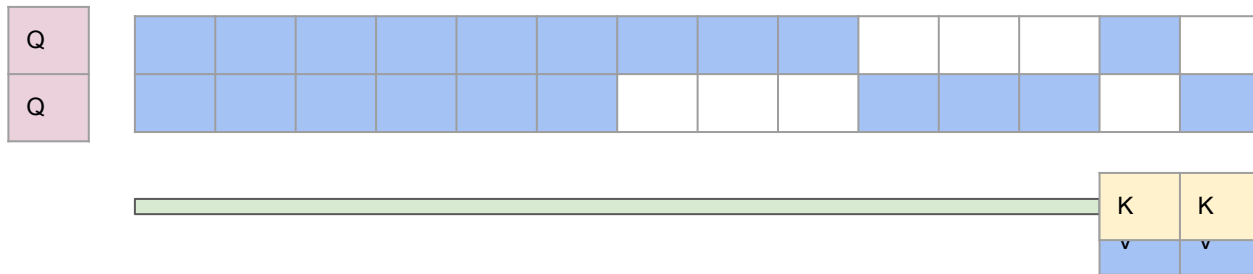


Prefill with multi prompt Attention mask

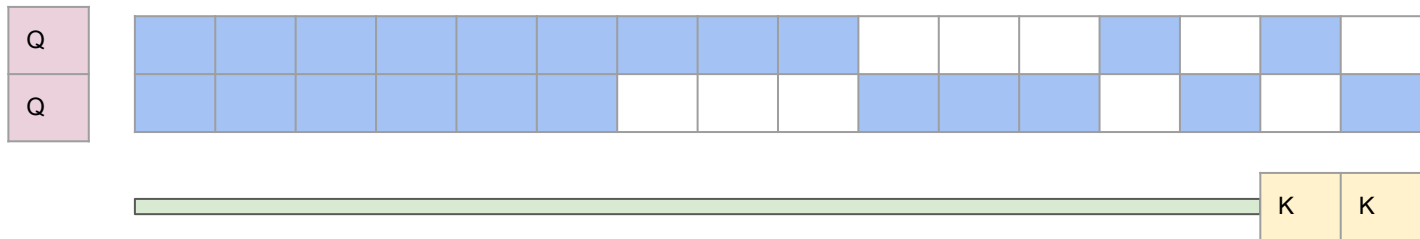


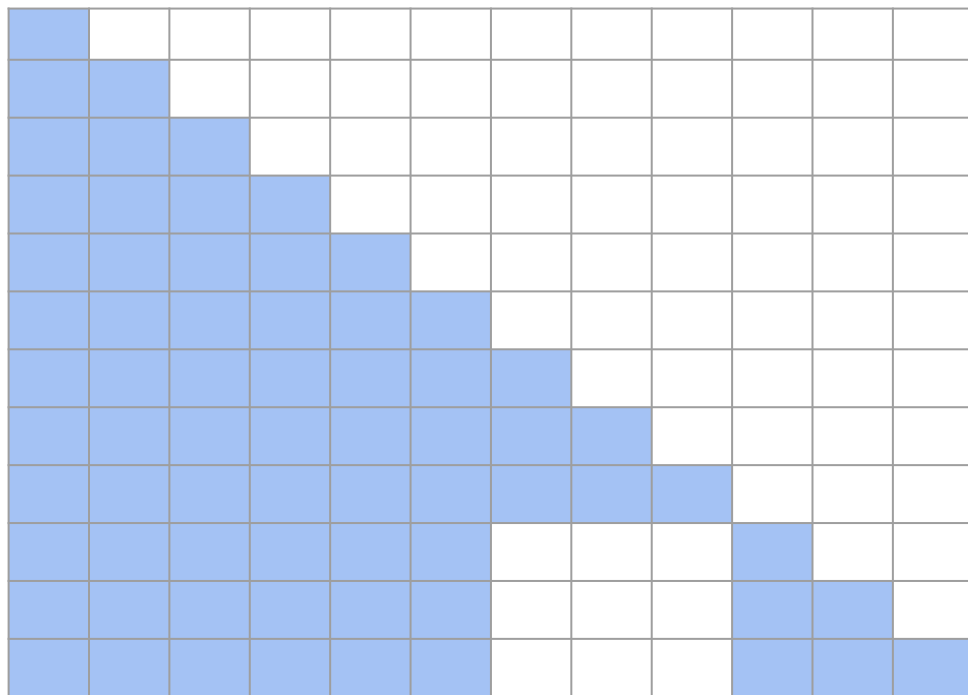
c1	c2	c3	c4	c5	c6	PA1	PA2	PA3	PB1	PB2	PB3
----	----	----	----	----	----	-----	-----	-----	-----	-----	-----

MultiDecode with MultiPrompt Generation



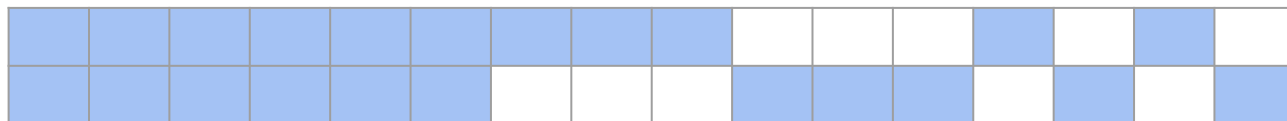
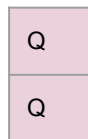
MultiDecode with MultiPrompt Generation





Context (C)

Prompt (P)



Generate (G)

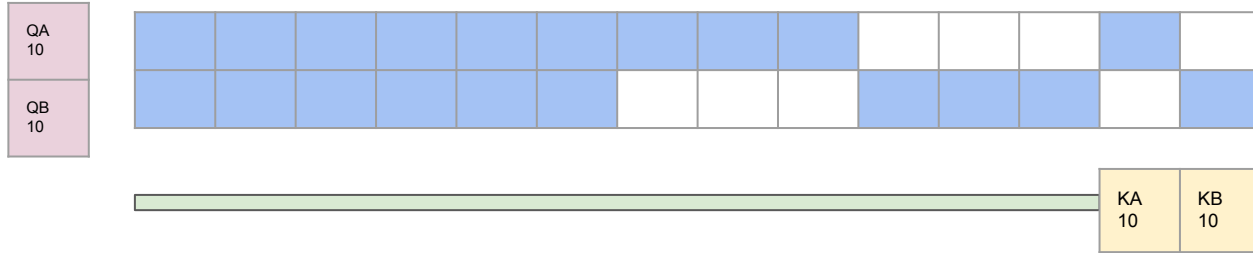


Prefill with multi prompt Position embed

Q 1											
Q 2											
Q 3											
Q 4											
Q 5											
Q 6											
QPA 7											
QPA 8											
QPA 9											
QPB 7											
QPB 8											
QPB 9											

C 1	C 2	C 3	C 4	C 5	C 6	PA 7	PA 8	PA 9	PB 7	PB 8	PB 9
-----	-----	-----	-----	-----	-----	------	------	------	------	------	------

MultiDecode with MultiPrompt Position embed



Naive multi-branch (beam)

- 1) Transformer beam search
 - a) Expand the number of rows in the batch. 1 additional row per branch
- 2) Run the branches sequentially

On to the code

