# MultiDecode

June 7, 2025 Ted Kyi and Roger Stager

#### MultiDecode implementation details

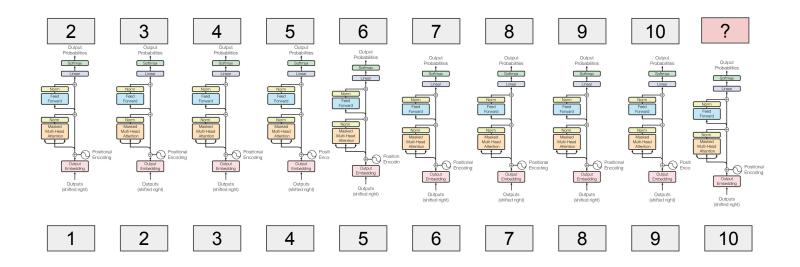
### Agenda

Overview of concepts

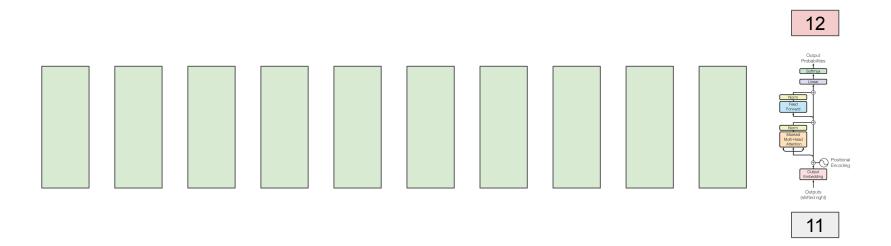
POC implementation walkthrough

Performance

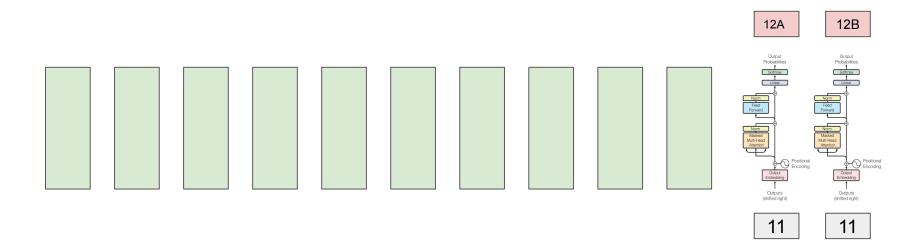
### Training vs Inference (generation)



### Inference (generation)

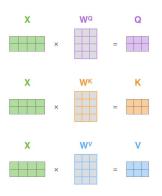


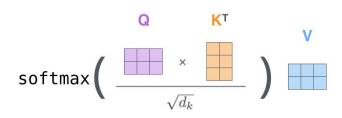
### Multi Token Decoding



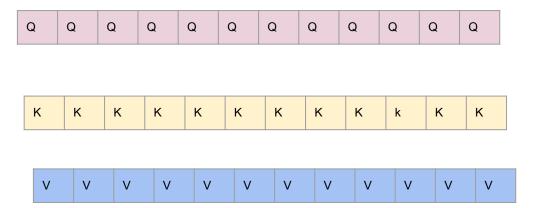
#### **Attention**

$$\operatorname{Attention}(Q,K,V) = \operatorname{softmax}\left(rac{QK^{\mathrm{T}}}{\sqrt{d_k}}
ight)V$$





### Calculate Q KV for full sequence



### Attention: Outer product between Query and Key

Q
Q
Q
Q
Q
Q
Q
Q
Q

Q

Q1K1	Q1K2	Q1K3	Q1K4	Q1K5	Q1K6	Q1K7	Q1K8	Q1K9	Q1K10	Q1K11	Q1K12
Q2K1	Q9K2	Q1K3	Q2K4	Q8K5	Q1K6	Q1K7	Q1K8	Q1K9	Q1K10	Q1K11	Q1K12
Q3K1	Q9K2	Q1K3	Q3K4	Q3K5	Q3K6	Q3K7	Q3K8	Q3K9	Q3K10	Q3K11	Q3K12
Q4K1	Q9K2	Q1K3	Q4K4	Q4K5	Q4K6	Q4K7	Q4K8	Q4K9	Q4K10	Q4K11	Q4K12
Q5K1	Q9K2	Q1K3	Q5K4	Q5K5	Q5K6	Q5K7	Q5K8	Q5K9	Q5K10	Q5K11	Q5K12
Q6K1	Q9K2	Q1K3	Q6K4	Q6K5	Q6K6	Q6K7	Q6K8	Q6K9	Q6K10	Q6K11	Q6K12
Q7K1	Q9K2	Q1K3	Q7K4	Q7K5	Q7K6	Q7K7	Q7K8	Q7K9	Q7K10	Q7K11	Q7K12
Q8K1	Q9K2	Q1K3	Q8K4	Q8K5	Q8K6	Q8K7	Q8K8	Q8K9	Q8K10	Q8K11	Q8K12
Q9K1	Q9K2	Q1K3	Q9K4	Q9K5	Q9K6	Q9K7	Q9K8	Q9K9	Q9K10	Q9K11	Q9K12
Q10K1	Q10K2	Q1K3	Q10K4	Q10K5	Q10K6	Q10K7	Q10K8	Q10K9	Q10K1 0	Q10K1 1	Q10K1 2
Q10K1	Q10K2	Q1K3	Q11K4	Q11K5	Q11K6	Q11K7	Q11K8	Q11K9	Q11K1 0	Q11K1 1	Q11K1 2
Q8K12	Q10K2	Q1K13	Q12K4	Q12K5	Q12K6	Q12K7	Q12K8	Q12K9	Q12K1 0	Q12K1 1	Q12K1 2

K K K K K K K K K K

#### Causal attention mask

Query

Keys

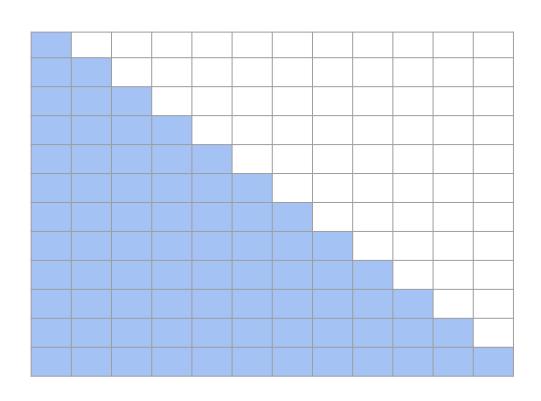
#### After Causal mask

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

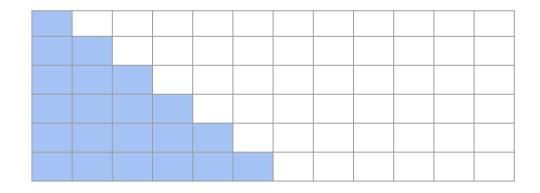
V	V										
V	V	V									
V	V	V	V								
V	V	V	V	V							
V	V	V	V	V	V						
V	V	V	V	V	V	V					
V	V	V	V	V	V	V	V				
V	V	V	V	V	V	V	V	V			
V	V	V	V	V	V	V	V	V	V		
V	V	V	V	V	V	V	V	V	V	V	

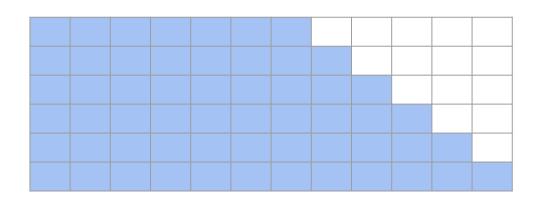
V

## This is the $O(N^2)$

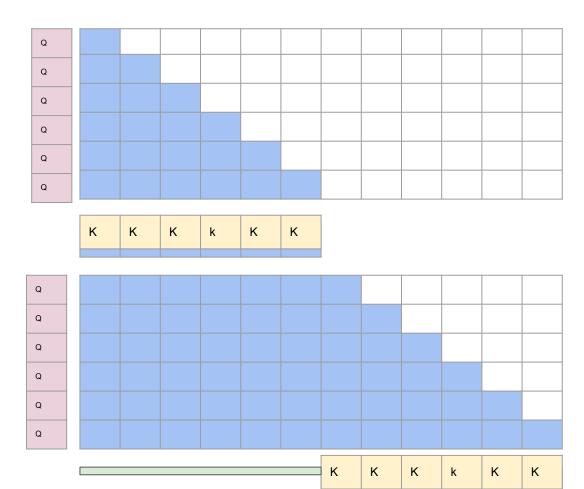


## Chunk prefill





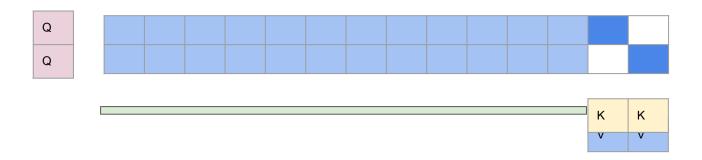
# Chunk prefill



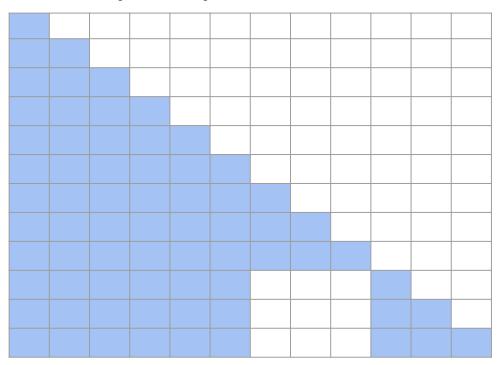
### Classic generation



### MultiDecode generation

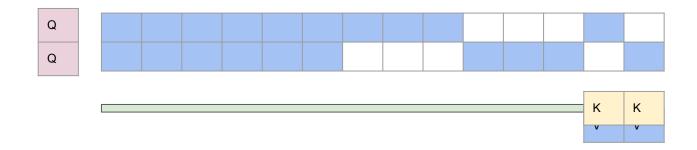


### Prefill with multi prompt Attention mask

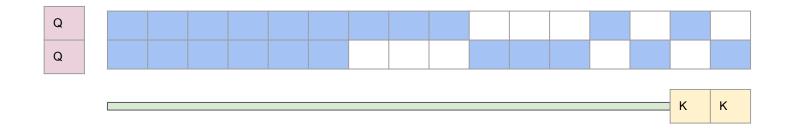


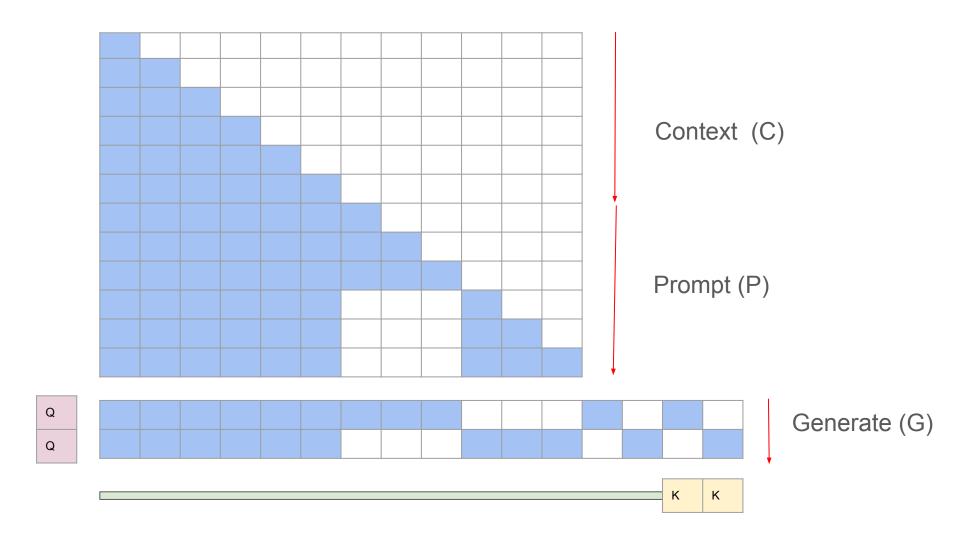
		c1	c2	сЗ	c4	с5	с6	PA1	PA2	PA3	PB1	PB2	PB3
--	--	----	----	----	----	----	----	-----	-----	-----	-----	-----	-----

### MultiDecode with MultiPrompt Generation

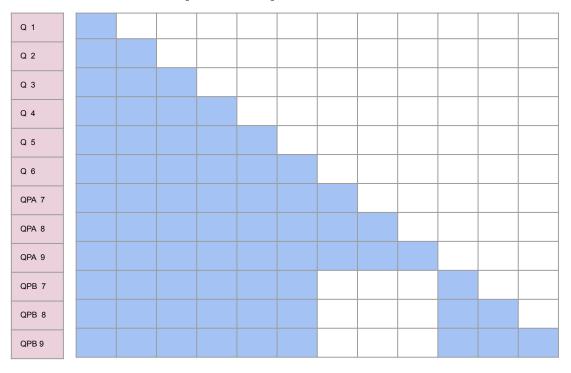


### MultiDecode with MultiPrompt Generation





### Prefill with multi prompt Position embed



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

Q	Q	(	<b>Q</b>	Q	C	)	Q	Q	Q	Q	Q	Q	
K	K	ŀ	<	K	K		K	K	K	k	K	K	
V	V	\	/	V	V	,	٧	V	V	V	V	٧	

Q

Q

Q

Q

Q

Q

Q

Q

Q

Q

Q

Q

Q

Κ