```
(handout: true)
```

Context Parallelism for Scalable Million-Token Inference

William Arnold

DLAlgo Inference Reading Group

2025-12-04

Background: Attention (single query)

For a single query vector $\vec{q}$:

$$\vec{a} = \vec{q}K^T = \begin{bmatrix} \vec{q}\cdot\vec{k}_1 & \vec{q}\cdot\vec{k}_2 & \dots & \vec{q}\cdot\vec{k}_S \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \dots & a_S \end{bmatrix}$$

$$\text{Softmax}(\vec{a}) = \begin{bmatrix} \frac{\exp(a_1-m)}{Z} & \frac{\exp(a_2-m)}{Z} & \dots & \frac{\exp(a_S-m)}{Z} \end{bmatrix}$$

$$\text{where } m = \max_j a_j, \quad Z = \sum_{j=1}^{S} \exp(a_j - m)$$

Background: Flash Attention

$\forall i \in \{1..S\}$

$$
\begin{aligned}
x_i &= \vec{q} \cdot \vec{k_i} \\
m_i &= \max(m_{i-1}, x_i) \\
Z_i &= Z_{i-1} e^{m_{i-1} - m_i} + e^{a_i - m_i} \\
\vec{o}_i &= \vec{o}_{i-1} e^{m_i - m_{i-1}} \frac{Z_{i-1}}{Z_i} + \frac{e^{x_i - m_i}}{Z_i} \vec{v_i}
\end{aligned}
$$

At $i = S$, $\vec{o}'_S$ is the correct output for query $\vec{q}$.[1]

Define $\text{AttnBlock}(\vec{q}, K, V, \vec{o}_{i-1}, m_{i-1}, Z_{i-1}) \rightarrow (\vec{o}', m, Z)$:

[1] Flash Attention Explained

Ring Attention

Compute attention on pieces of the sequence!

[1] o_i is initialized to zero

Ring Attention

Compute attention on pieces of the sequence!

N ranks, give each rank 1/N of the sequence

$\{Q_1, ..., Q_N\}, \{K_1, ..., K_N\}, \{V_1, ..., V_N\}$

[1] o_i is initialized to zero

Ring Attention

Compute attention on pieces of the sequence!

N ranks, give each rank 1/N of the sequence

$\{Q_1, ..., Q_N\}, \{K_1, ..., K_N\}, \{V_1, ..., V_N\}$

On rank i, keep Q_i and compute $\forall i \in \{1..N\}$

$$\vec{\boldsymbol{o}}_i, m_i, Z_i \leftarrow \text{AttnBlock}(Q_i, \boldsymbol{K_i}, \boldsymbol{V_i}, \vec{\boldsymbol{o}}_i, m_{i-1}, Z_{i-1})^{[1]}$$

$\vec{\boldsymbol{o}}$ will have the correct output for Q_i

[1] o_i is initialized to zero

Ring Attention

Compute attention on pieces of the sequence!

N ranks, give each rank 1/N of the sequence

$\{Q_1, ..., Q_N\}, \{K_1, ..., K_N\}, \{V_1, ..., V_N\}$

On rank i, keep Q_i and compute $\forall i \in \{1..N\}$

$$\vec{\boldsymbol{o}}_i, m_i, Z_i \leftarrow \text{AttnBlock}(Q_i, \boldsymbol{K_i}, \boldsymbol{V_i}, \vec{\boldsymbol{o}}_i, m_{i-1}, Z_{i-1})^{[1]}$$

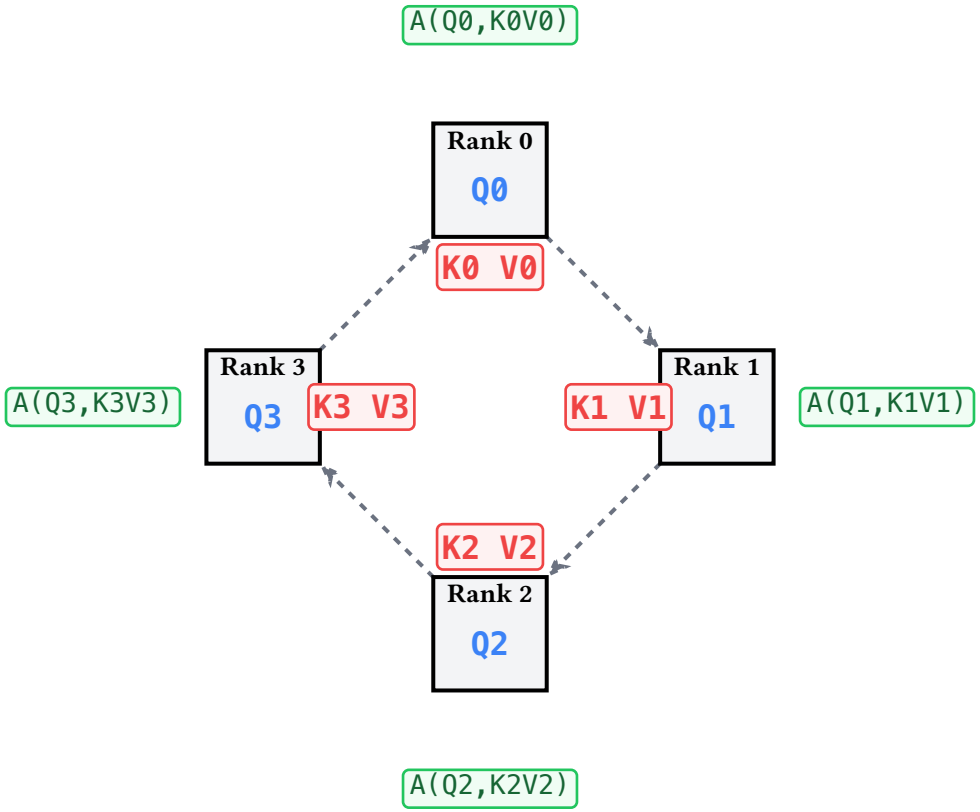
$\vec{\boldsymbol{o}}$ will have the correct output for Q_i

How to get $\boldsymbol{K_i}$ and $\boldsymbol{V_i}$?

[1] o_i is initialized to zero

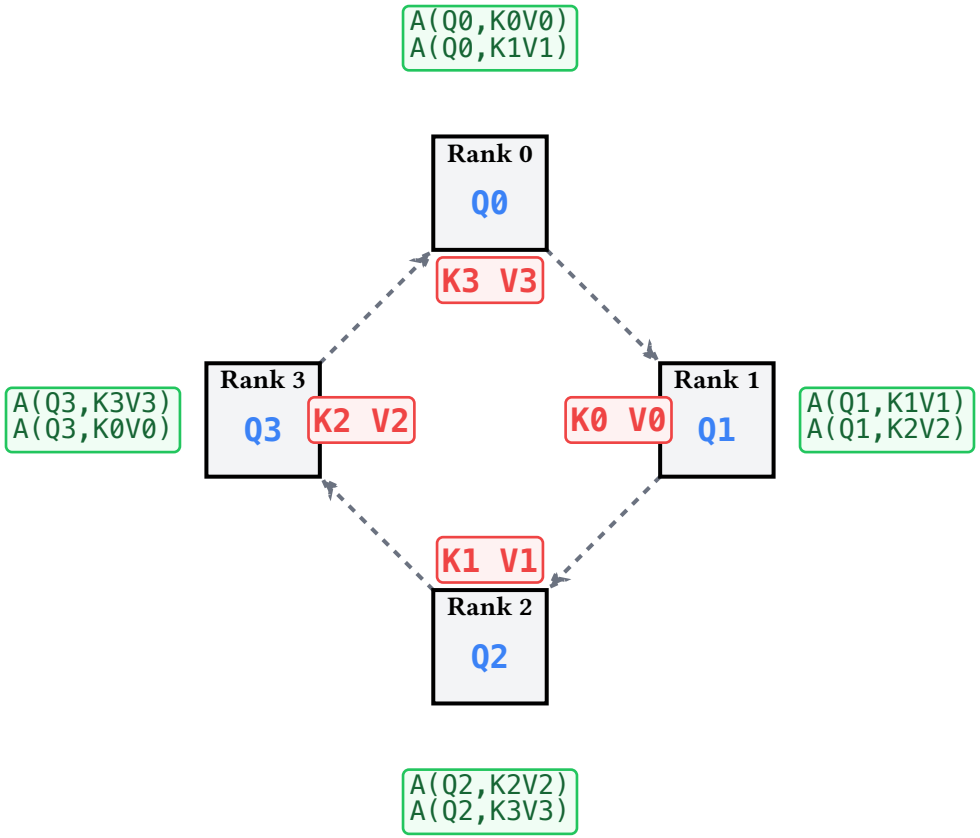
Ring Attention: Communication

Step 1: Each rank computes local attention



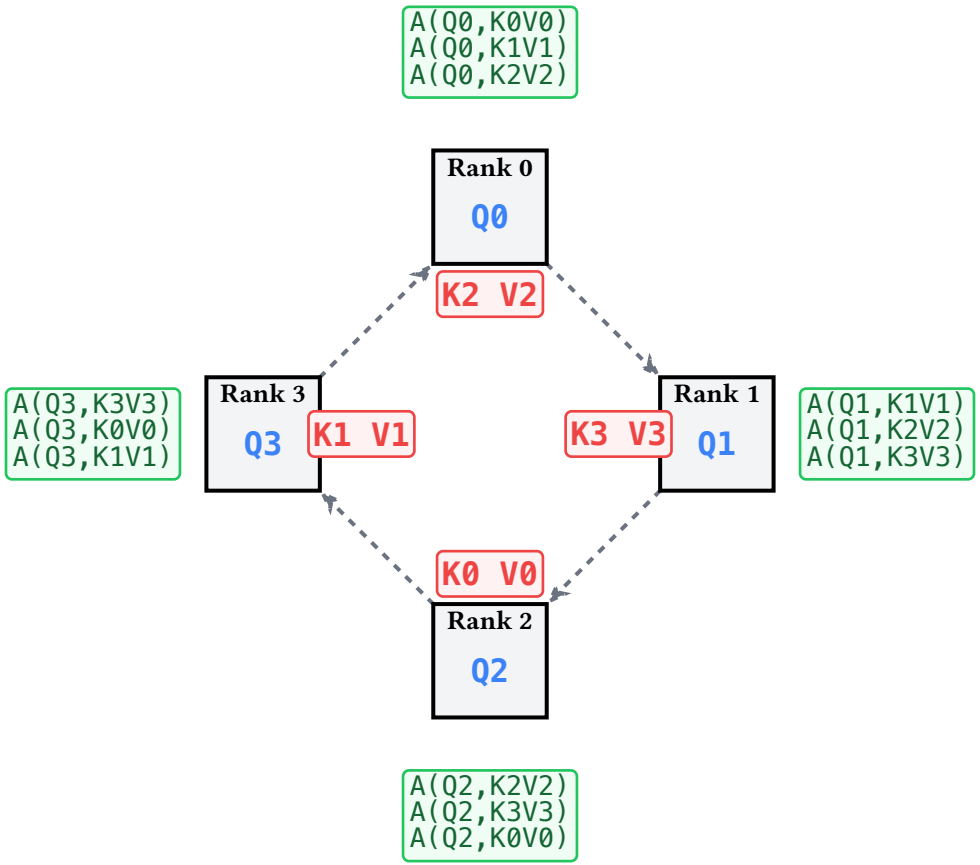
Ring Attention: Communication

Step 2: Send KV to next rank, compute & store



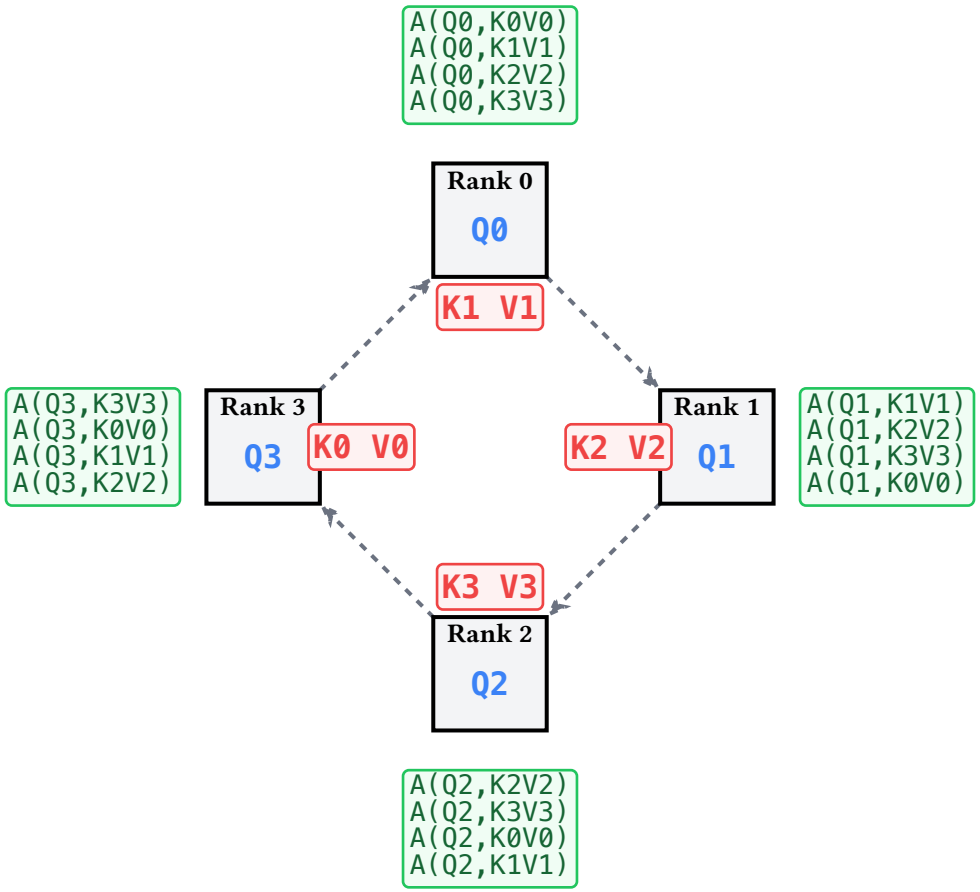
Ring Attention: Communication

Step 3: Send KV to next rank, compute & store



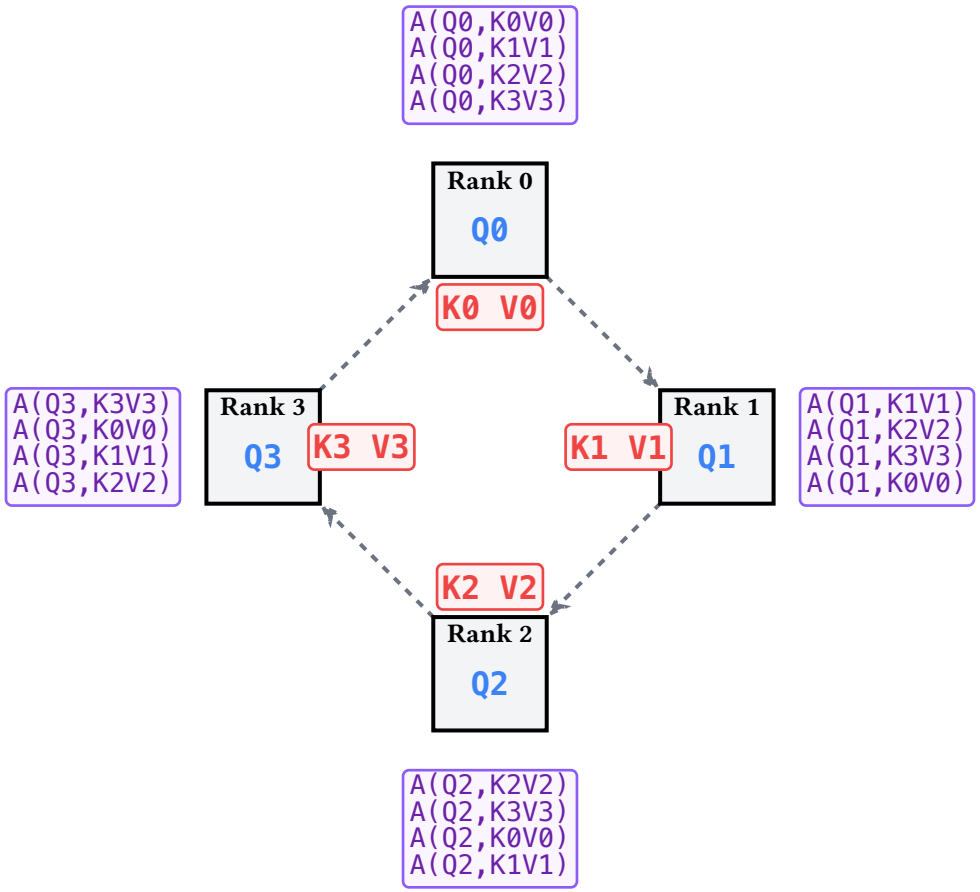
Ring Attention: Communication

Step 4: Send KV to next rank, compute & store



Ring Attention: Communication

Complete! Each Q has seen all KVs



Ring Attention: Complexity (Pass-KV)

For N ranks, sequence length T, model dim D_q, N_H query heads, N_{KV} KV heads:, e bytes/element

Ring Attention: Complexity (Pass-KV)

For N ranks, sequence length T, model dim D_q, N_H query heads, N_{KV} KV heads:, e bytes/element

FLOPS	$2T^2D$
Q bytes	TDe
KV bytes	$2TD\frac{N_{KV}}{N_H}e$

Ring Attention: Complexity (Pass-KV)

For N ranks, sequence length T, model dim D_q, N_H query heads, N_{KV} KV heads:, e bytes/element

FLOPS	$2T^2D$
Q bytes	TDe
KV bytes	$2TD\frac{N_{KV}}{N_H}e$

Computation: $\frac{2T^2D}{N}$ FLOPs

Ring Attention: Complexity (Pass-KV)

For N ranks, sequence length T, model dim D_q, N_H query heads, N_{KV} KV heads:, e bytes/element

FLOPS	$2T^2D$
Q bytes	TDe
KV bytes	$2TD\frac{N_{KV}}{N_H}e$

Computation: $\frac{2T^2D}{N}$ FLOPs

Communication (unidirectional): $2TD\frac{N_{KV}}{N_H}e$ bytes

Ring Attention: Overlap Condition

$$T_{\text{compute}} \geq T_{\text{comm}}$$

Ring Attention: Overlap Condition

$$\frac{2T^2D}{CN} \geq \frac{2TD\left(\frac{N_{KV}}{N_H}\right)e}{\text{BW}}$$

Ring Attention: Overlap Condition

$$\frac{2T^2D}{CN} \geq \frac{2TD\left(\frac{N_{KV}}{N_H}\right)e}{\text{BW}}$$

$$\frac{T}{CN} \geq \frac{N_{KV}e}{N_H\ \text{BW}}$$

Ring Attention: Overlap Condition

$$\frac{2T^2D}{CN} \geq \frac{2TD\left(\frac{N_{KV}}{N_H}\right)e}{\text{BW}}$$

$$\frac{T}{CN} \geq \frac{N_{KV}e}{N_H\ \text{BW}}$$

$$T \geq N\frac{N_{KV}}{N_H}\frac{Ce}{\text{BW}}$$

Ring Attention: Overlap Condition

$$\frac{2T^2D}{CN} \geq \frac{2TD\left(\frac{N_{KV}}{N_H}\right)e}{\text{BW}}$$

$$\frac{T}{CN} \geq \frac{N_{KV}e}{N_H\ \text{BW}}$$

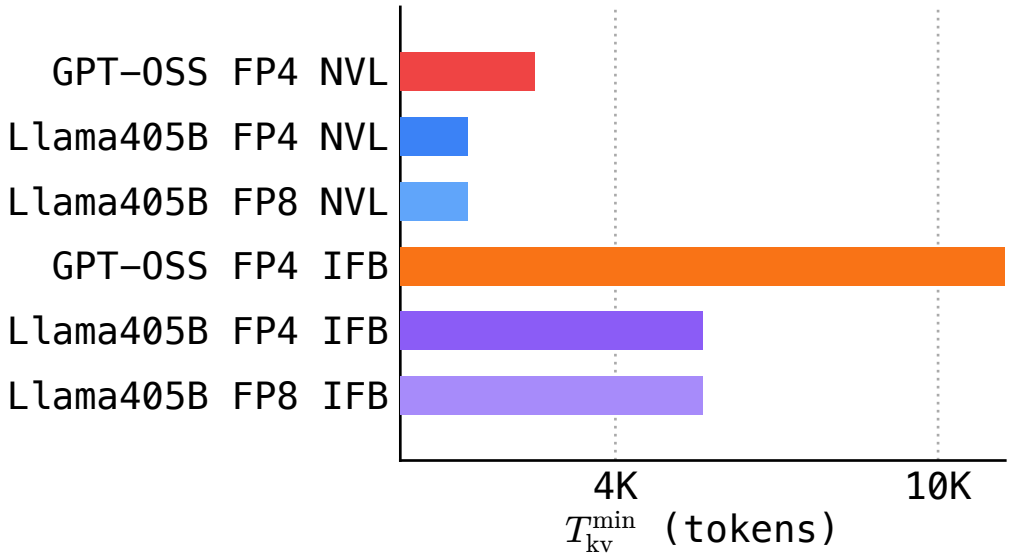
$$T \geq N\frac{N_{KV}}{N_H}\frac{Ce}{\text{BW}}$$

Call RHS $T_{\text{kv}}^{\min}$: minimum T where we're compute-bound

$\frac{Ce}{\text{BW}}$ for Blackwell IFB is $\approx$22500, NVLINK is $\approx$5000

When Can We Hide Communication?

Minimum T for communication overlap ($8 \times$ B200)[1]:



Independent of datatype since Ce is *theoretically* constant.

[1]Infiniband unidirectional @ 200GB/s, 4.5e12 FP8 FLOPS, 9e12 FP4 FLOPS

Ring Attention with Prefixes

With P cached tokens,

FLOPS	$2T(P+T)D$
Q bytes	TDe
KV bytes	$2(P+T)D\frac{N_{KV}}{N_H}e$

Ring Attention with Prefixes

With P cached tokens,

FLOPS	$2T(P+T)D$
Q bytes	TDe
KV bytes	$2(P+T)D\frac{N_{KV}}{N_H}e$

Computation: $\frac{2(P+T)TD}{N}$ FLOPs

Ring Attention with Prefixes

With P cached tokens,

FLOPS	$2T(P+T)D$
Q bytes	TDe
KV bytes	$2(P+T)D\frac{N_{KV}}{N_H}e$

Computation: $\frac{2(P+T)TD}{N}$ FLOPs

Communication (unidirectional): $2(P+T)D\frac{N_{KV}}{N_H}e$ bytes

Ring Attention with Prefixes: Overlap Condition

$$\frac{2T(P+T)D}{CN} \geq \frac{2(P+T)D\frac{N_{KV}}{N_H}e}{\text{BW}}$$

Ring Attention with Prefixes: Overlap Condition

$$\frac{2T(P+T)D}{CN} \geq \frac{2(P+T)D\frac{N_{KV}}{N_H}e}{\text{BW}}$$

$$\frac{T}{CN} \geq \frac{N_{KV}}{N_H}\frac{e}{\text{BW}}$$

Ring Attention with Prefixes: Overlap Condition

$$\frac{2T(P+T)D}{CN} \geq \frac{2(P+T)D\frac{N_{KV}}{N_H}e}{\text{BW}}$$

$$\frac{T}{CN} \geq \frac{N_{KV}}{N_H}\frac{e}{\text{BW}}$$

$$T \geq N\frac{N_{KV}}{N_H}\frac{Ce}{\text{BW}}$$

Ring Attention with Prefixes: Overlap Condition

$$\frac{2T(P+T)D}{CN} \geq \frac{2(P+T)D\frac{N_{KV}}{N_H}e}{\text{BW}}$$

$$\frac{T}{CN} \geq \frac{N_{KV}}{N_H}\frac{e}{\text{BW}}$$

$$T \geq N\frac{N_{KV}}{N_H}\frac{Ce}{\text{BW}}$$

Same $T_{\text{kv}}^{\min}$!

But the new T is *only new tokens*!

Must be 4k+ to hide communication!

This Paper: Context Parallelism over Queries

Observation: For small T, queries are much smaller than KV!

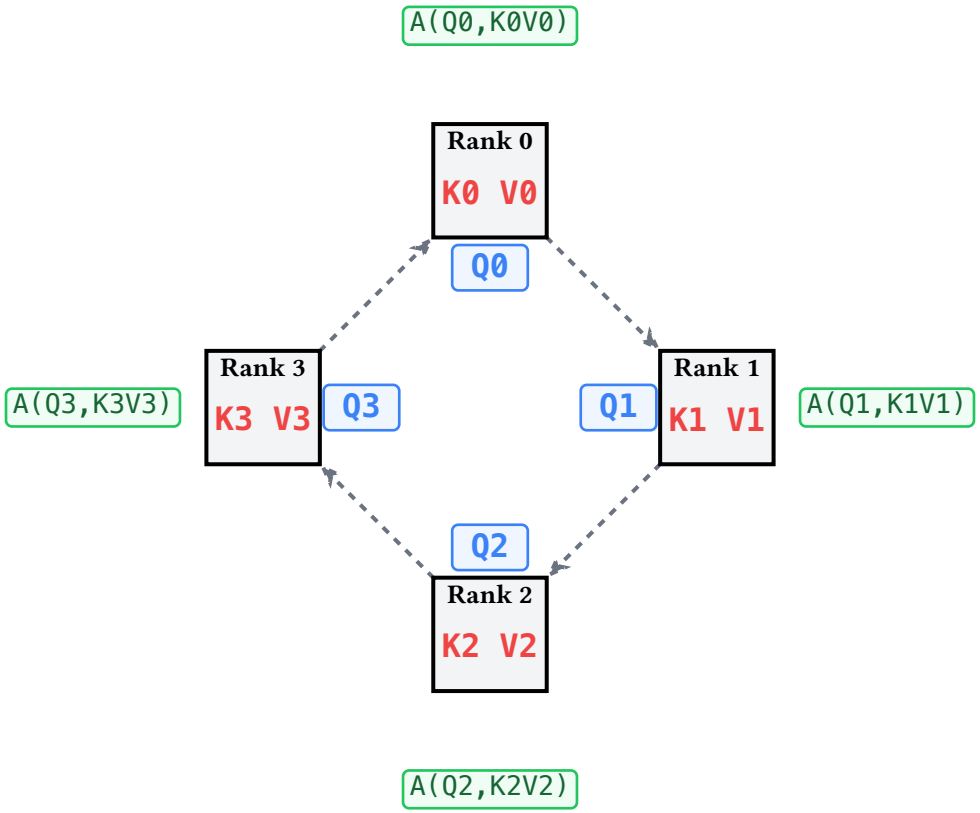
This Paper: Context Parallelism over Queries

Observation: For small T, queries are much smaller than KV!

What if we ring-pass *queries* instead of KV?

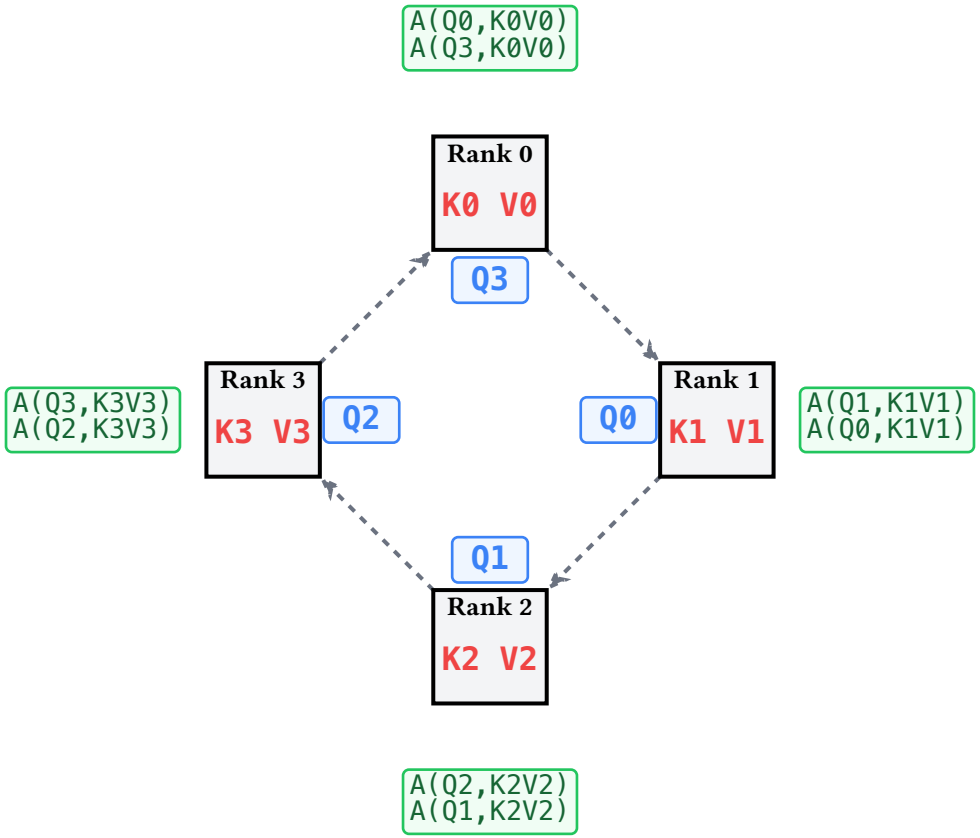
Ring Attention: Pass-Q

Step 1: Each rank computes local attention, stores (o, m, Z)



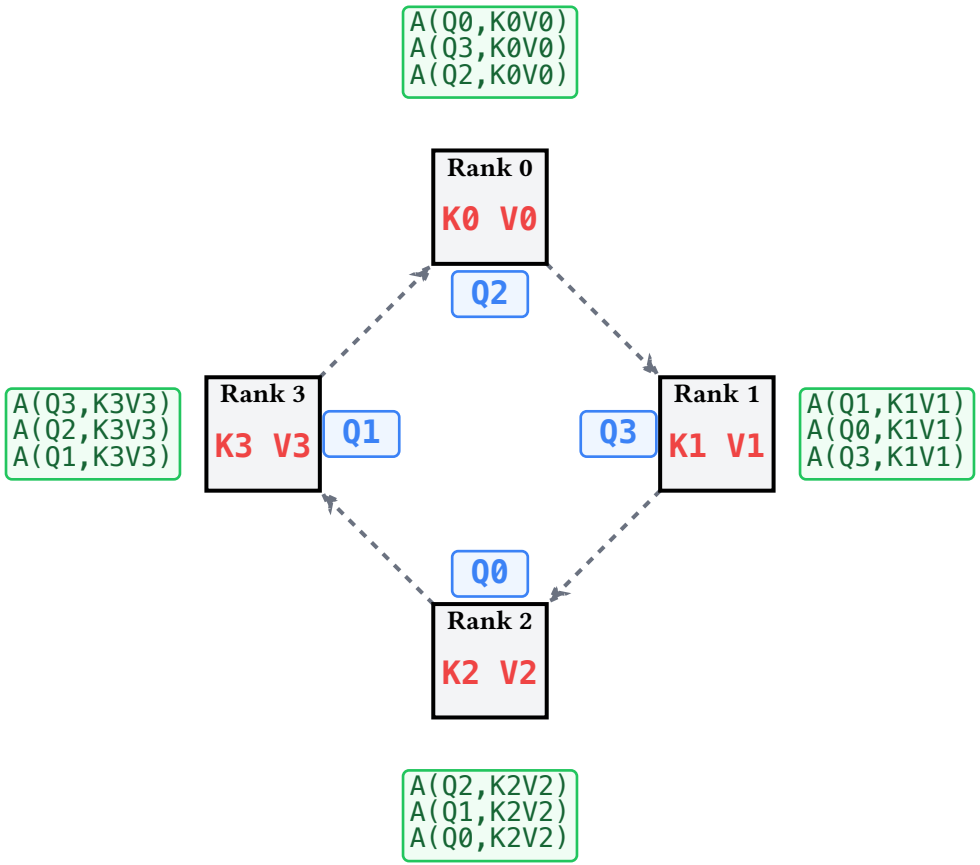
Ring Attention: Pass-Q

Step 2: Send Q to next rank, compute & store



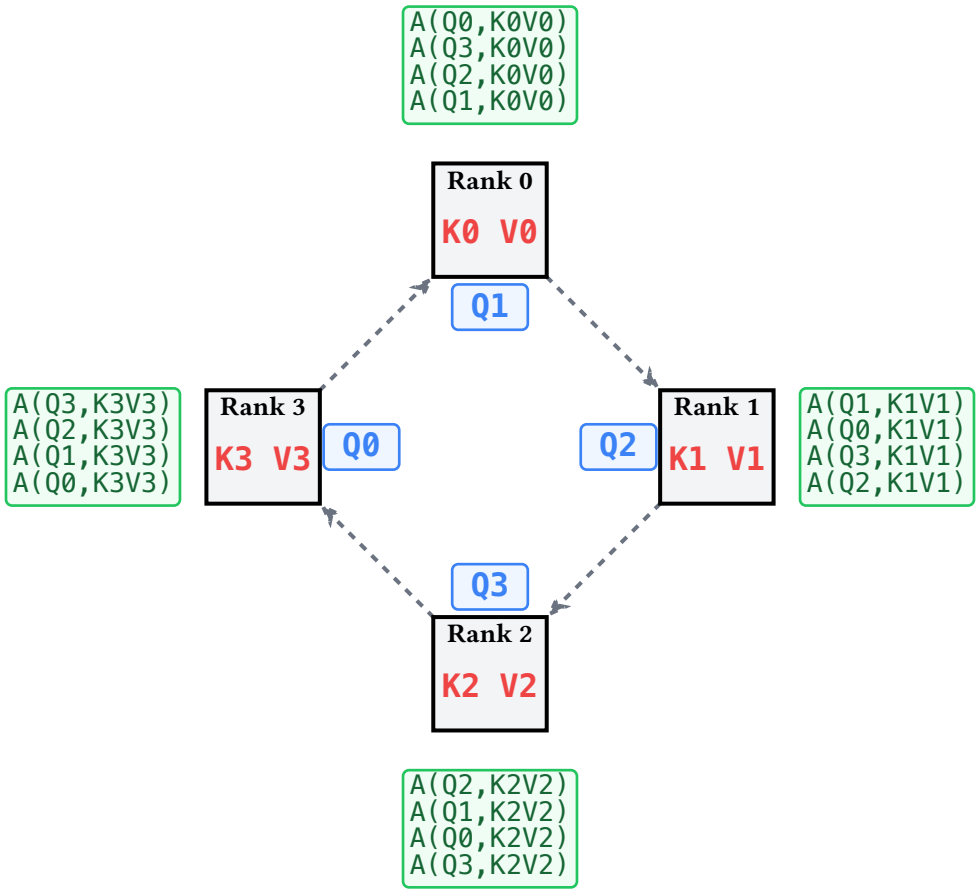
Ring Attention: Pass-Q

Step 3: Send Q to next rank, compute & store



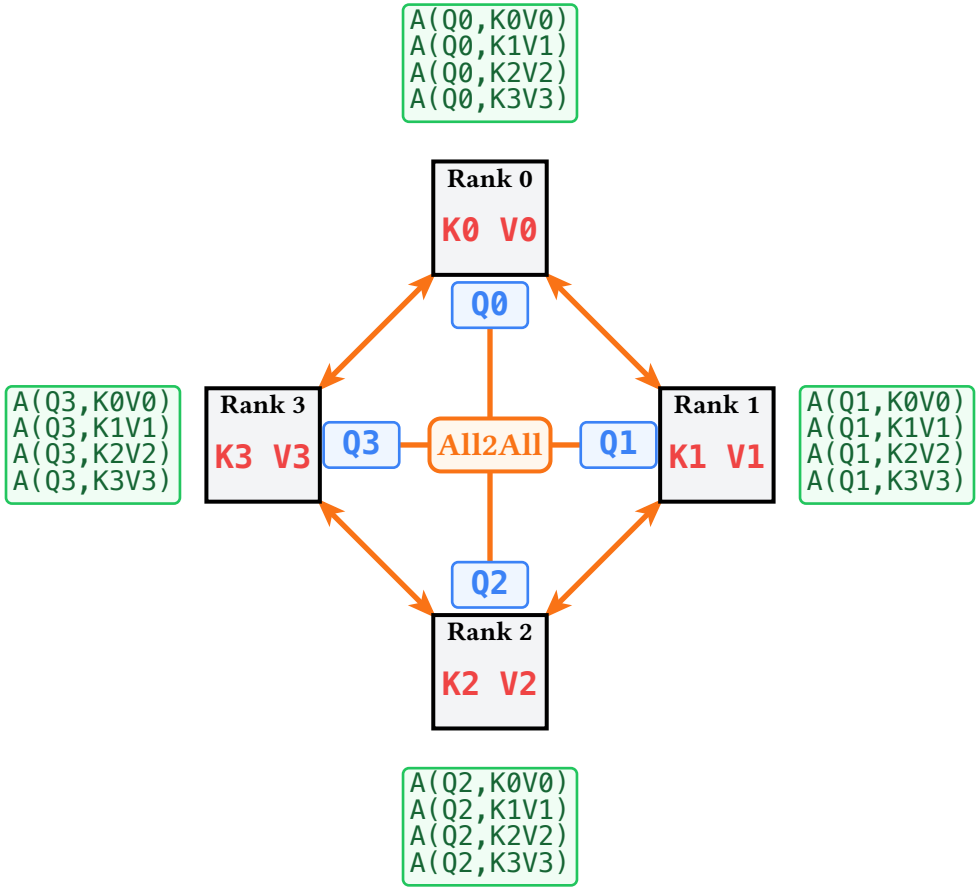
Ring Attention: Pass-Q

Step 4: Send Q to next rank, compute & store



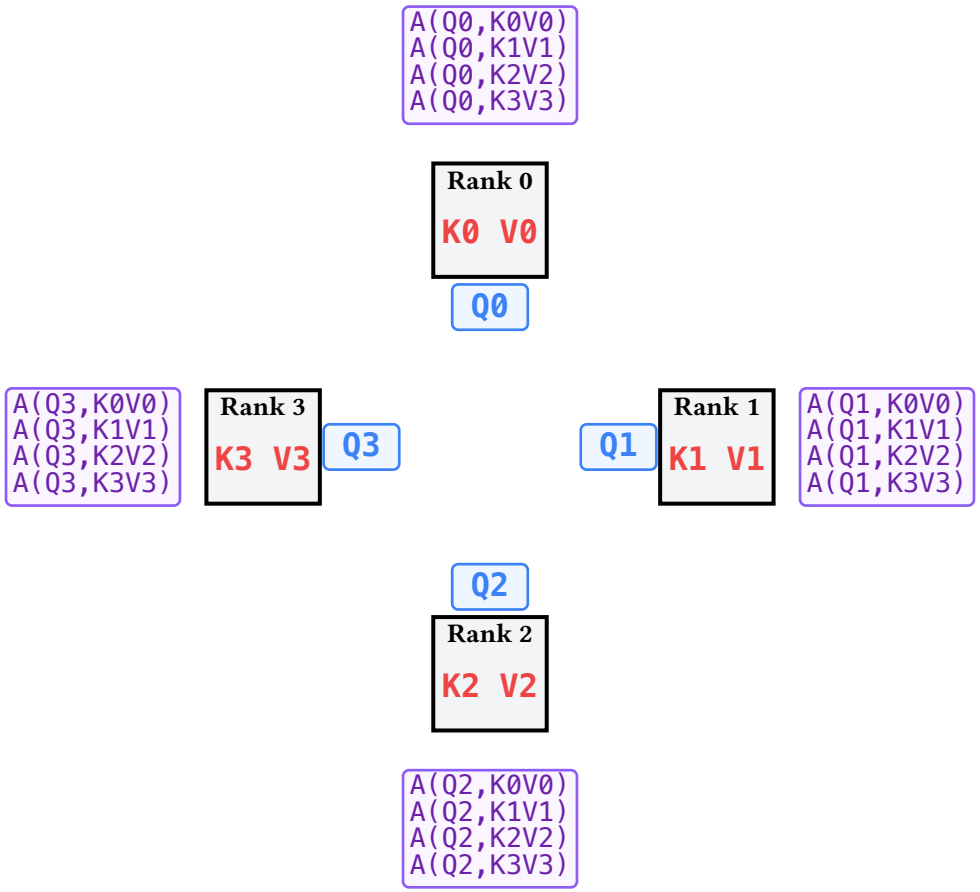
Ring Attention: Pass-Q

All2All: Exchange partial outputs…



Ring Attention: Pass-Q

Done! Each rank has all partials for its own Q



Q-Passing Roofline

FLOPS	$2T(P+T)D$
Q bytes	TDe
KV bytes	$2(P+T)D\frac{N_{KV}}{N_H}e$

Computation: $\frac{2T(P+T)D}{N}$ FLOPs

Q-Passing Roofline

FLOPS	$2T(P+T)D$
Q bytes	TDe
KV bytes	$2(P+T)D\frac{N_{KV}}{N_H}e$

Computation: $\dfrac{2T(P+T)D}{N}$ FLOPs

Communication (unidirectional): $\boldsymbol{TDe}$ bytes

Only depends on T instead of $P+T$!

Q-Passing Roofline

Overlap condition:

$$T_{\text{compute}} \geq T_{\text{comm}}$$

Q-Passing Roofline

Overlap condition:

$$T_{\text{compute}} \geq T_{\text{comm}}$$

$$\frac{2T(P+T)D}{CN} \geq \frac{TDe}{\text{BW}}$$

Q-Passing Roofline

Overlap condition:

$$T_{\text{compute}} \geq T_{\text{comm}}$$

$$\frac{2T(P+T)D}{CN} \geq \frac{TDe}{\text{BW}}$$

$$P + T \geq \frac{N}{2}\frac{Ce}{\text{BW}}$$

Q-Passing Roofline

Overlap condition:

$$T_{\text{compute}} \geq T_{\text{comm}}$$

$$\frac{2T(P+T)D}{CN} \geq \frac{TDe}{\text{BW}}$$

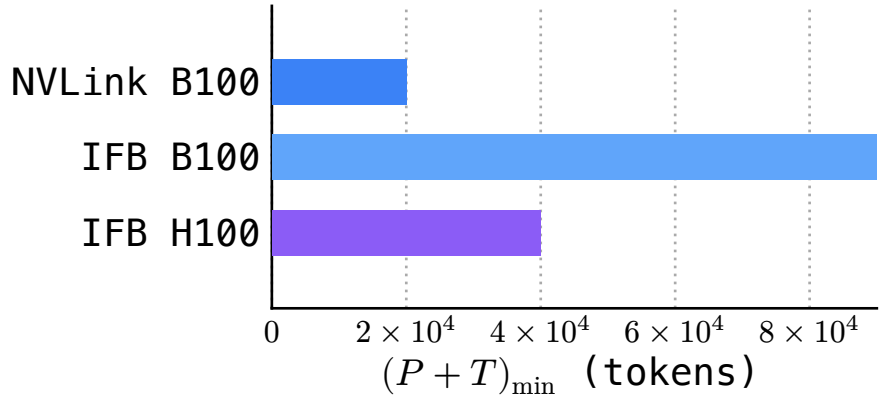
$$P + T \geq \frac{N}{2}\frac{Ce}{\text{BW}}$$

Doesn't depend on N_H, N_{KV}, T, just $P + T$

Note: $C \cdot e$ is the *theoretically* the same for FP4 and FP8, so model doesn't matter!

Q-Passing Roofline

Minimum $(P+T)$ for communication overlap ($8 \times$ B200):



Requires balanced KVs across ranks (round-robin during decode)

When is the All-to-All faster than Pass-KV's communication overhead?

The rest of the paper stinks

The rest of the paper stinks

They have a bunch of math mistakes. Ex:

$$Latency(All2All) = (N-1) \cdot \frac{(D+1) \cdot T \cdot e}{BW}$$

The rest of the paper stinks

They have a bunch of math mistakes. Ex:

$$Latency(All2All) = (N-1) \cdot \frac{(D+1) \cdot T \cdot e}{BW}$$

This is *really wrong*. All-to-all in a ring is $\frac{\text{Data}}{4\ \text{BW}}$!

The rest of the paper stinks

They have a bunch of math mistakes. Ex:

$$Latency(All2All) = (N-1) \cdot \frac{(D+1) \cdot T \cdot e}{BW}$$

This is *really wrong*. All-to-all in a ring is $\frac{\text{Data}}{4\ \text{BW}}$!

They also don't actually solve their inequalities!

The rest of the paper stinks

They have a bunch of math mistakes. Ex:

$$Latency(All2All) = (N-1) \cdot \frac{(D+1) \cdot T \cdot e}{BW}$$

This is *really wrong*. All-to-all in a ring is $\frac{\text{Data}}{4\ \text{BW}}$!

They also don't actually solve their inequalities!

Let's do it right.

All-to-All cost

All-to-All time in a ring is roughly[1]

$$T_{\text{all2all}} = \frac{TDe}{4\ \text{BW}}$$

[1] See this derivation

All-to-All cost

All-to-All time in a ring is roughly[1]

$$T_{\text{all2all}} = \frac{TDe}{4\ \text{BW}}$$

All-to-All time in NVL72 is just $\frac{TDe}{N\ \text{BW}}$

[1] See this derivation

All-to-All cost

All-to-All time in a ring is roughly[1]

$$T_{\text{all2all}} = \frac{TDe}{4\ \text{BW}}$$

All-to-All time in NVL72 is just $\frac{TDe}{N\ \text{BW}}$

Need to check if $T_{\text{all2all}} < \left(T_{\text{kv,comm}} - T_{\text{kv,compute}}\right)$

[1] See this derivation

All-to-All cost

$$T_{\text{kv,comm}} - T_{\text{kv,compute}} = 2(P+T)D\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{2(P+T)TD}{NC}$$

All-to-All cost

$$T_{\text{kv,comm}} - T_{\text{kv,compute}} = 2(P+T)D\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{2(P+T)TD}{NC}$$

$$= 2(P+T)D\left(\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{T}{NC}\right)$$

All-to-All cost

$$T_{\text{kv,comm}} - T_{\text{kv,compute}} = 2(P+T)D\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{2(P+T)TD}{NC}$$

$$= 2(P+T)D\left(\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{T}{NC}\right)$$

$$\frac{TDe}{4\ \text{BW}} < T_{\text{kv,comm}} - T_{\text{kv,compute}}$$

All-to-All cost

$$T_{\text{kv,comm}} - T_{\text{kv,compute}} = 2(P+T)D\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{2(P+T)TD}{NC}$$

$$= 2(P+T)D\left(\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{T}{NC}\right)$$

$$\frac{TDe}{4\ \text{BW}} < T_{\text{kv,comm}} - T_{\text{kv,compute}}$$

Ends up being quadratic... hand it to `sympy`, solve for T and plot

All-to-All cost

$$T_{\text{kv,comm}} - T_{\text{kv,compute}} = 2(P+T)D\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{2(P+T)TD}{NC}$$

$$= 2(P+T)D\left(\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{T}{NC}\right)$$

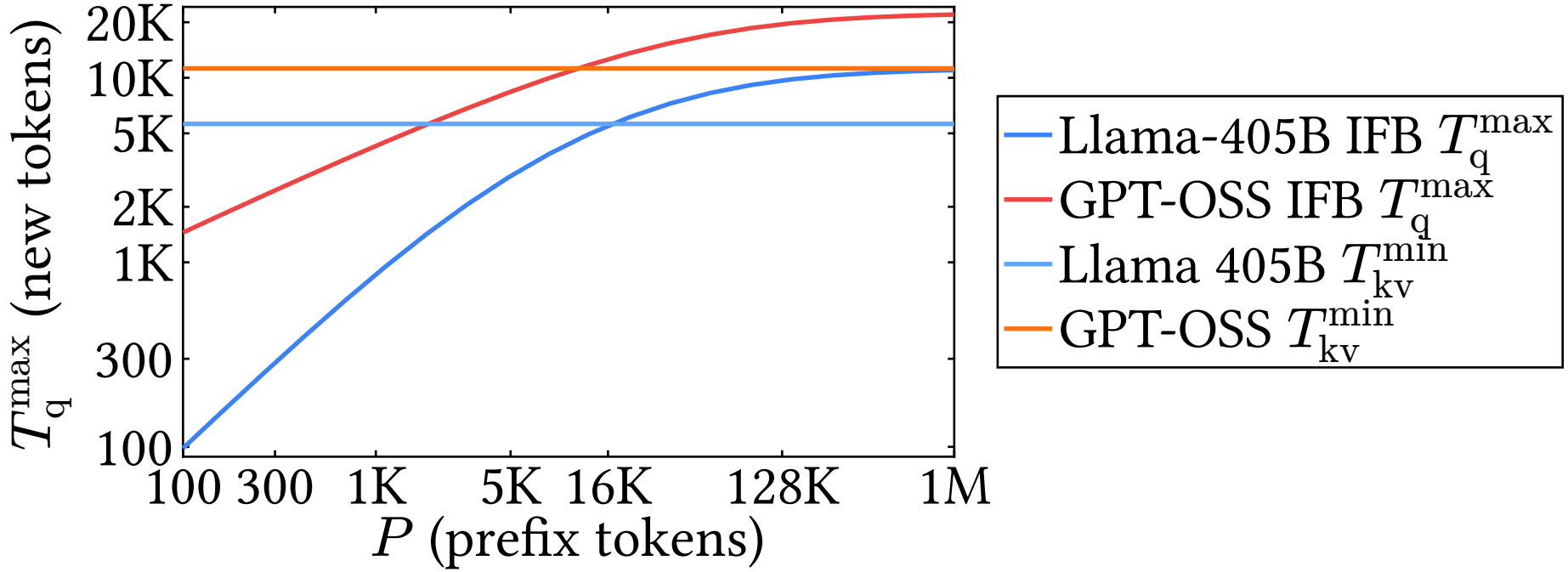
$$\frac{TDe}{4\ \text{BW}} < T_{\text{kv,comm}} - T_{\text{kv,compute}}$$

Ends up being quadratic... hand it to `sympy`, solve for T and plot

Gets $T_{\text{q}}^{\max}$: the maximum T where Pass-Q All-to-All is faster than Pass-KV overhead

$T_{\text{q}}^{\max}$ vs Prefix Length

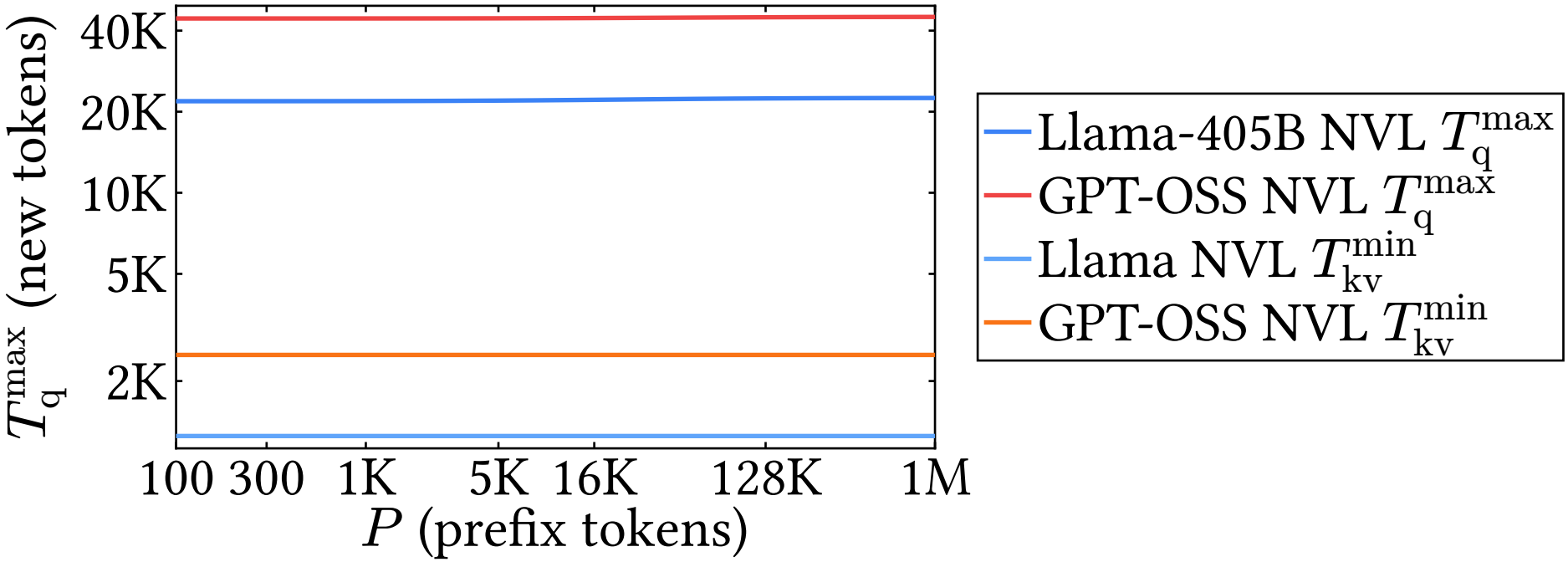
Max T where Pass-Q is faster than Pass-KV (8 × B200 IFB), i.e. $T_{\text{q}}^{\max}$:



Below $T_{\text{q}}^{\max}$, use All-to-All (Pass-Q). Above $T_{\text{q}}^{\max}$, use Pass-KV.

What about NVL72?

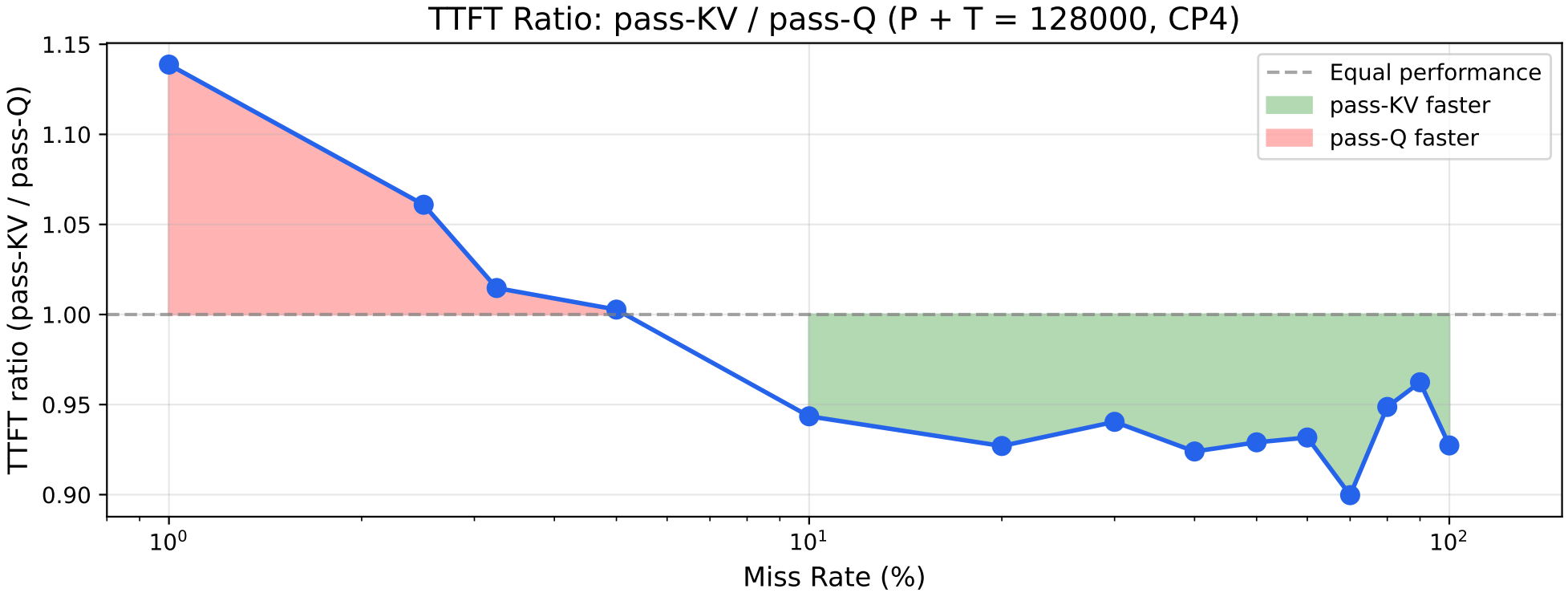
All-to-all cost is $\frac{TDe}{N \cdot 4 \cdot \text{BW}}$ and BW is 900GB/s v.s. 200GB/s.



Pass-Q can work up to larger T_q^{max}, almost indpendent of P.

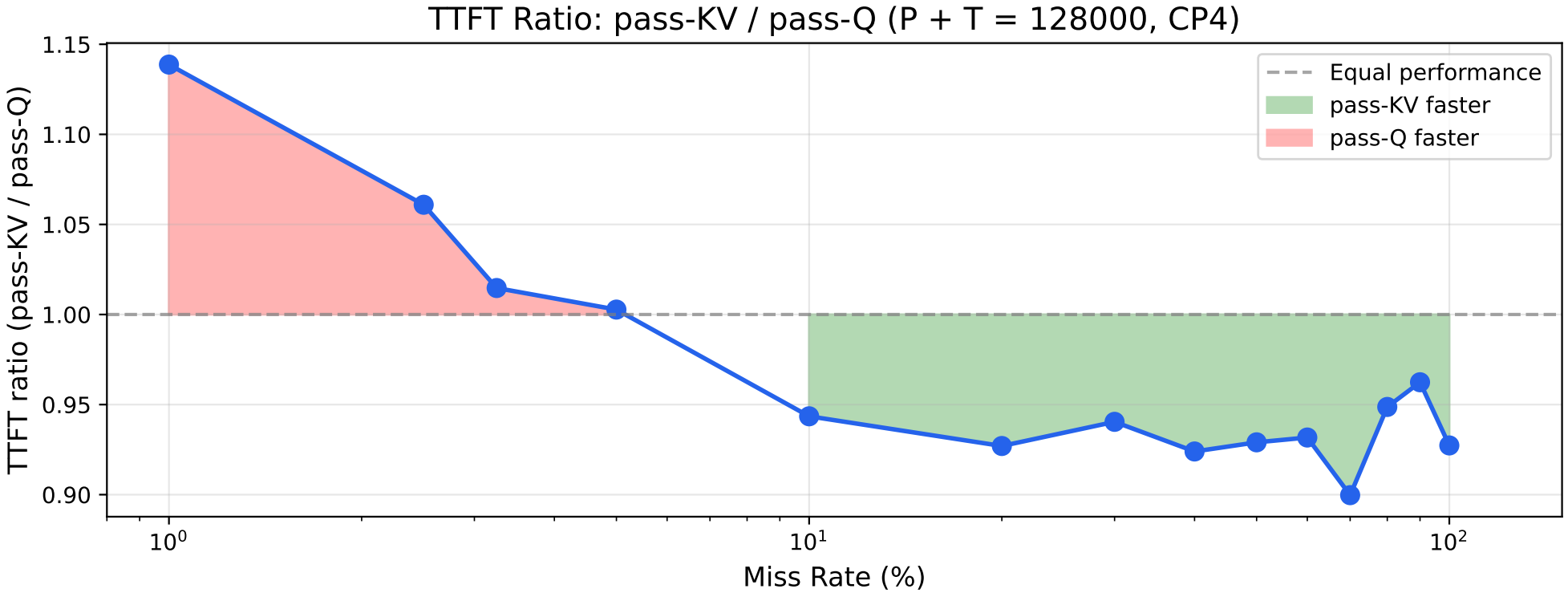
But T_{kv}^{min} is way smaller! Only fail to hide comm for $T_{kv}^{min} < 1K,3K$.

Meta's Hopper Performance



Report pass-Q is slighly better than pass-KV for very low miss rates

Meta's Hopper Performance

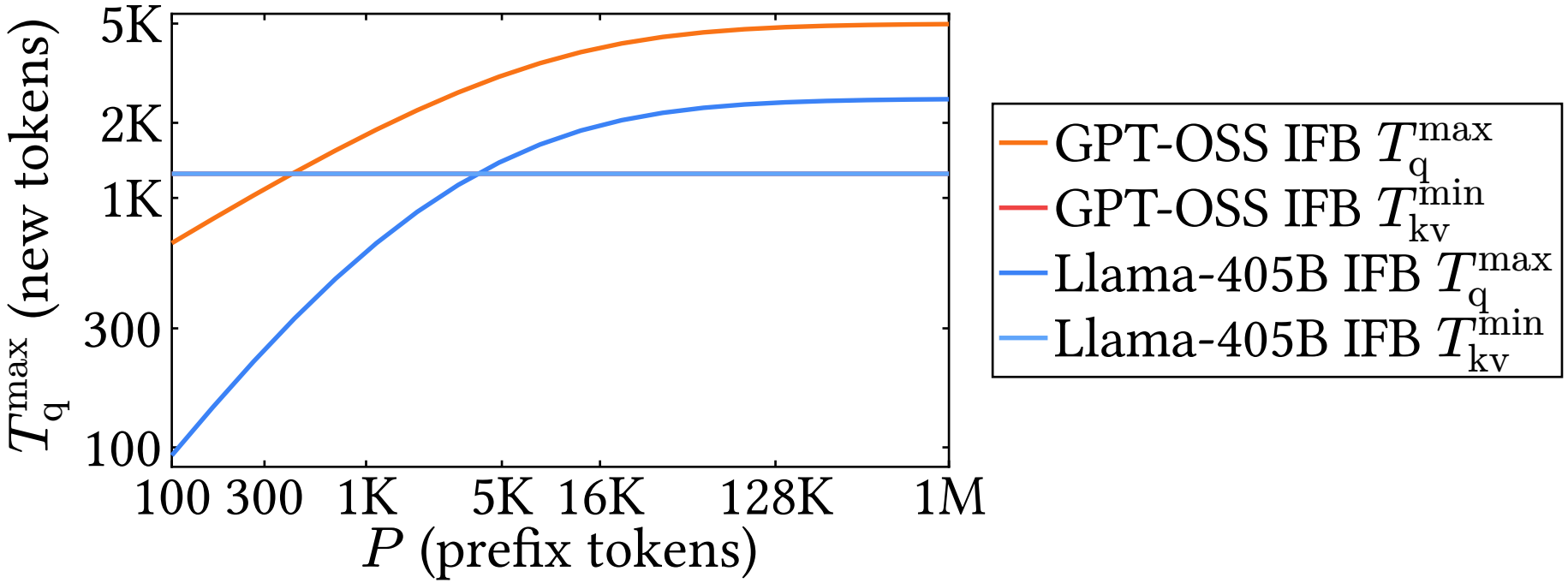


Report pass-Q is slighly better than pass-KV for very low miss rates

Does the math say the same?

Validating Meta's Hopper Performance

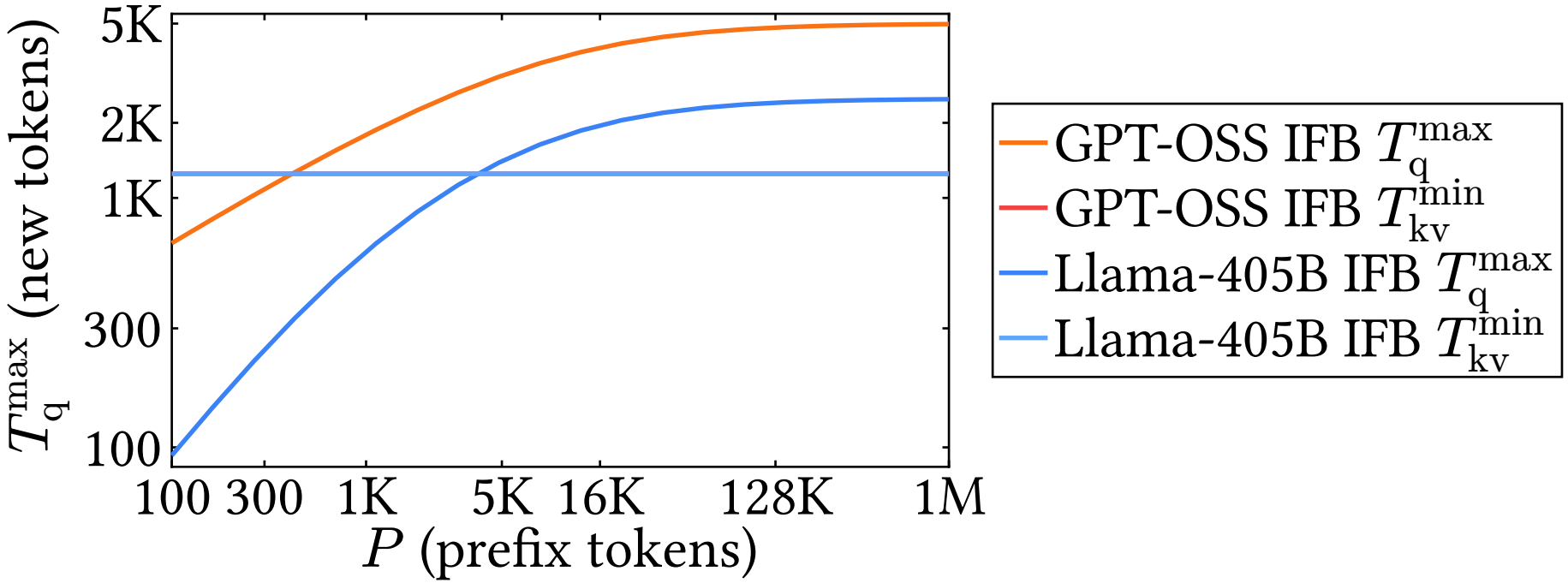
4 x H100 with 200GB/s (unidirectional) IFB



At P=128K, $T_q^{max} \approx 5K \approx 4\%$ miss rate! Matches empirical of $\approx 5\%$!

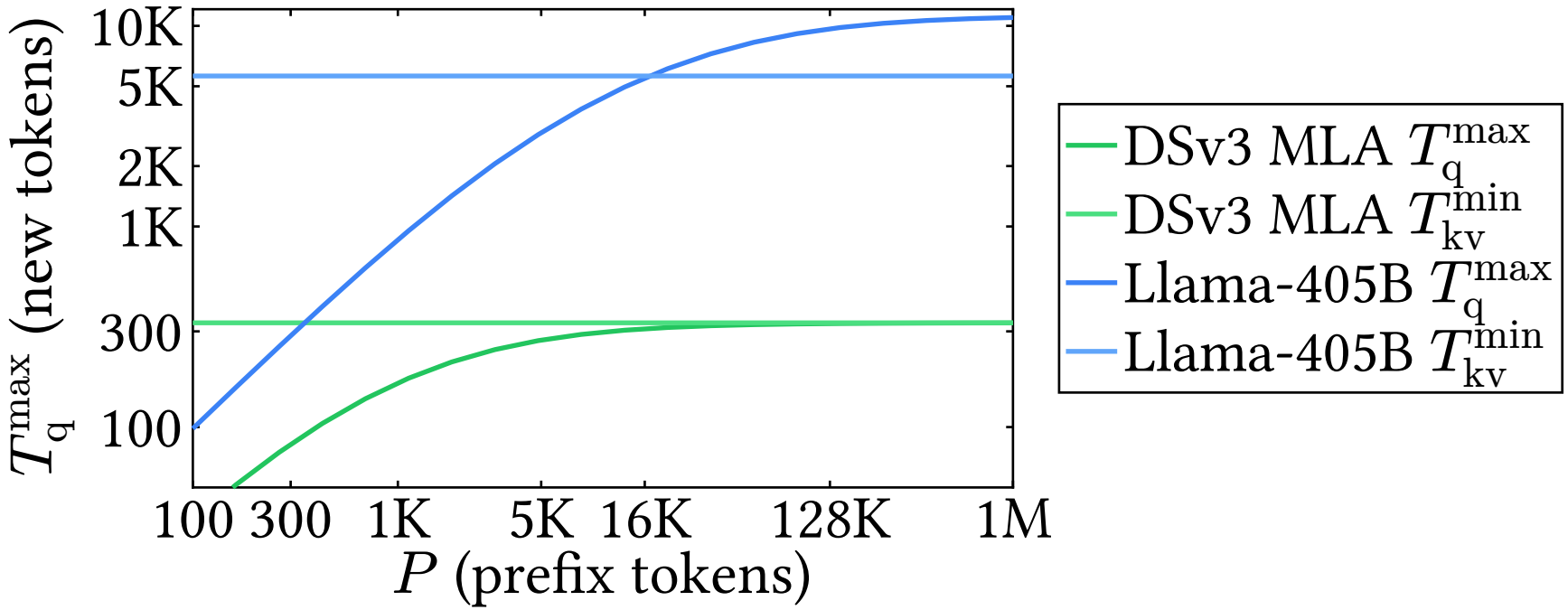
Validating Meta's Hopper Performance

4 x H100 with 200GB/s (unidirectional) IFB



But T_{kv}^{min} is 1250 for both models... Well above the 4k where Meta sees a difference

Extra: DSv3 MLA



MLA: smaller KV cache $\longrightarrow$ lower $T_{\text{kv}}^{\min}$, higher $T_{\text{q}}^{\max}$ $\longrightarrow$ Pass-Q is never necessary.