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 $\vec{\boldsymbol{q}}$:

$$\vec{\boldsymbol{a}} = \vec{\boldsymbol{q}} K^T = \left[\vec{\boldsymbol{q}} \cdot \vec{\boldsymbol{k}}_1 \;\; \vec{\boldsymbol{q}} \cdot \vec{\boldsymbol{k}}_2 \;\; \dots \;\; \vec{\boldsymbol{q}} \cdot \vec{\boldsymbol{k}}_S\right] = [a_1 \;\; a_2 \;\; \dots \;\; a_S]$$

$$\text{Softmax}(\vec{\boldsymbol{a}}) = \left[\frac{\exp(a_1 - m)}{Z} \;\; \frac{\exp(a_2 - m)}{Z} \;\; \dots \;\; \frac{\exp(a_S - m)}{Z}\right]$$

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

Background: Flash Attention

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

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

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

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

[1] Flash Attention Explained

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

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

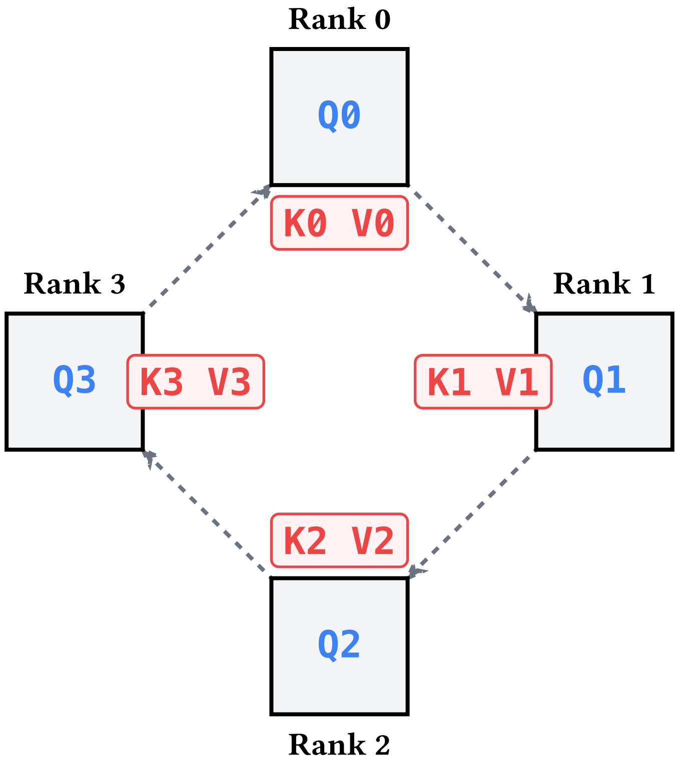
$\vec{\boldsymbol{o}}$ will have the correct output for Q_i

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

[1] o_i is initialized to zero

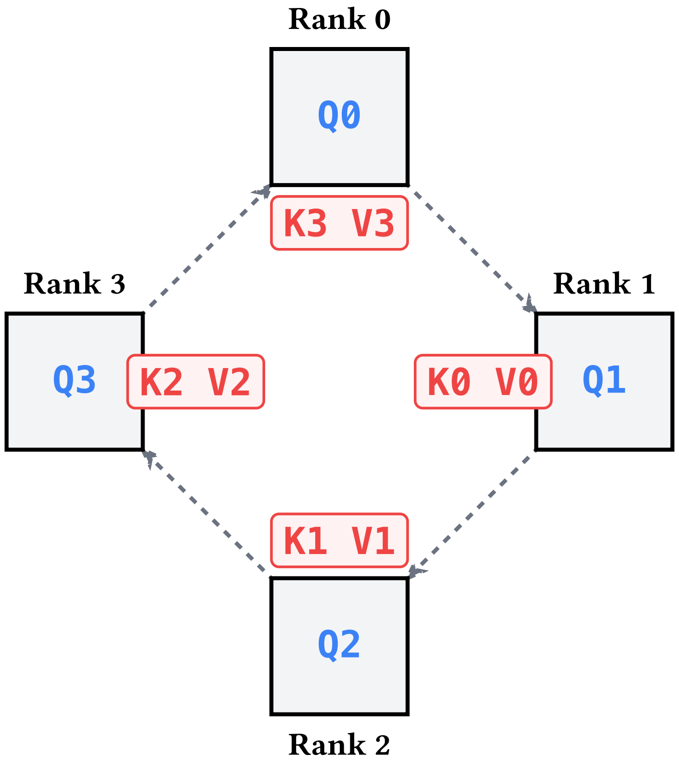
Ring Attention: Communication

Step 1: Each rank computes local attention



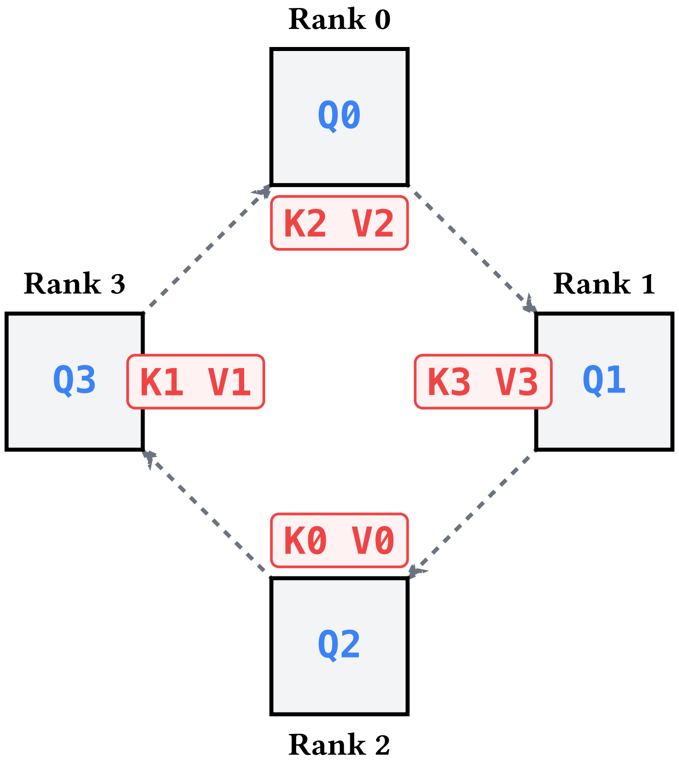
Ring Attention: Communication

Step 2: Send KV to next rank



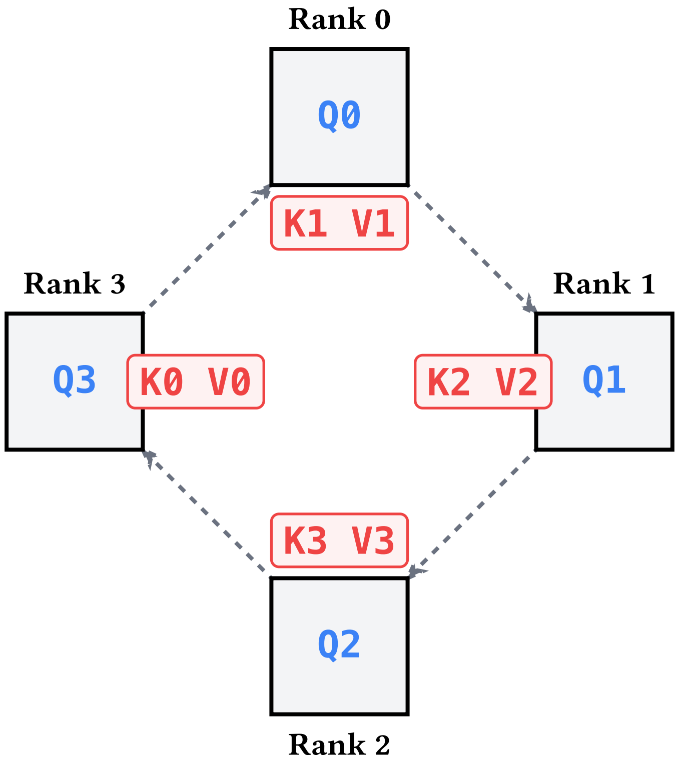
Ring Attention: Communication

Step 3: Send KV to next rank



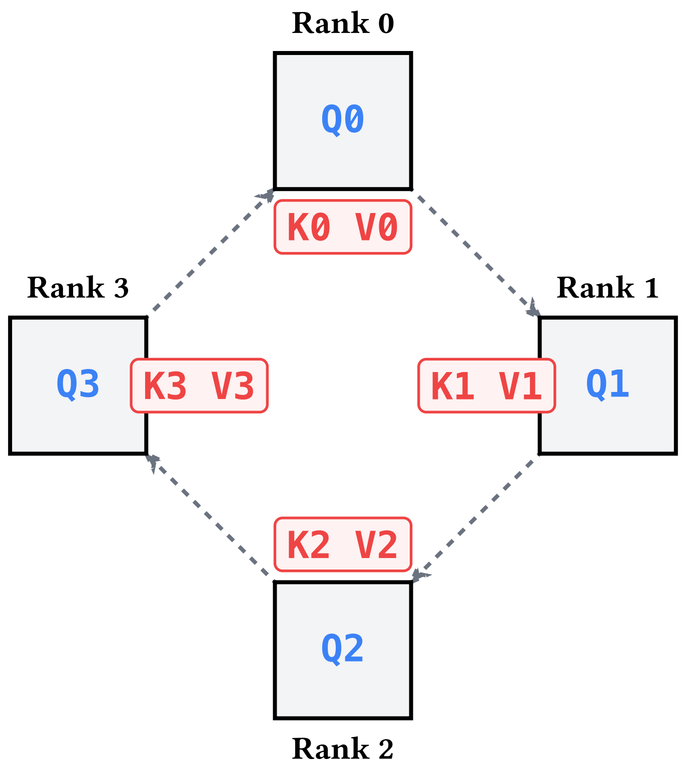
Ring Attention: Communication

Step 4: Send KV to next rank



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

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

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

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

$$T \geq N\left(\frac{N_{KV}}{N_H}\right)\left(\frac{Ce}{\text{BW}}\right)$$

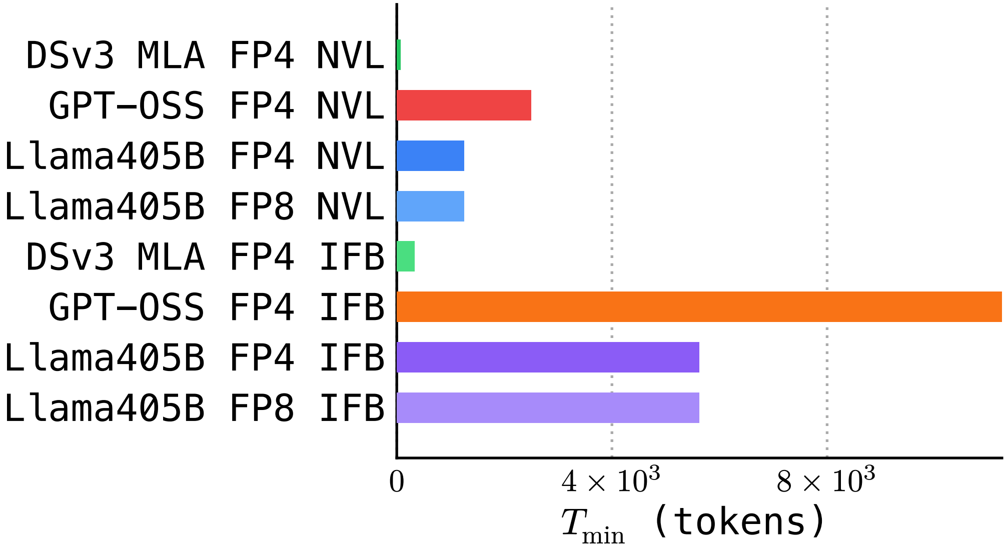
$\frac{Ce}{\text{BW}}$ for Blackwell IFB is ≈22500

$\frac{Ce}{\text{BW}}$ for Blackwell NVLink is ≈5000

$\frac{Ce}{\text{BW}}$ for Blackwell HMB is ≈562.5

When Can We Hide Communication?

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



Independent of datatype since Ce is *theoretically* constant. NVL72

[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	$2(P+T)D\frac{N_{KV}}{N_H}e$

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

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

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

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

❗ it’s the same!

But the new T is *only new tokens*!

Must be 4k+ to hide communication!

This Paper: Context Parallelism over Queries

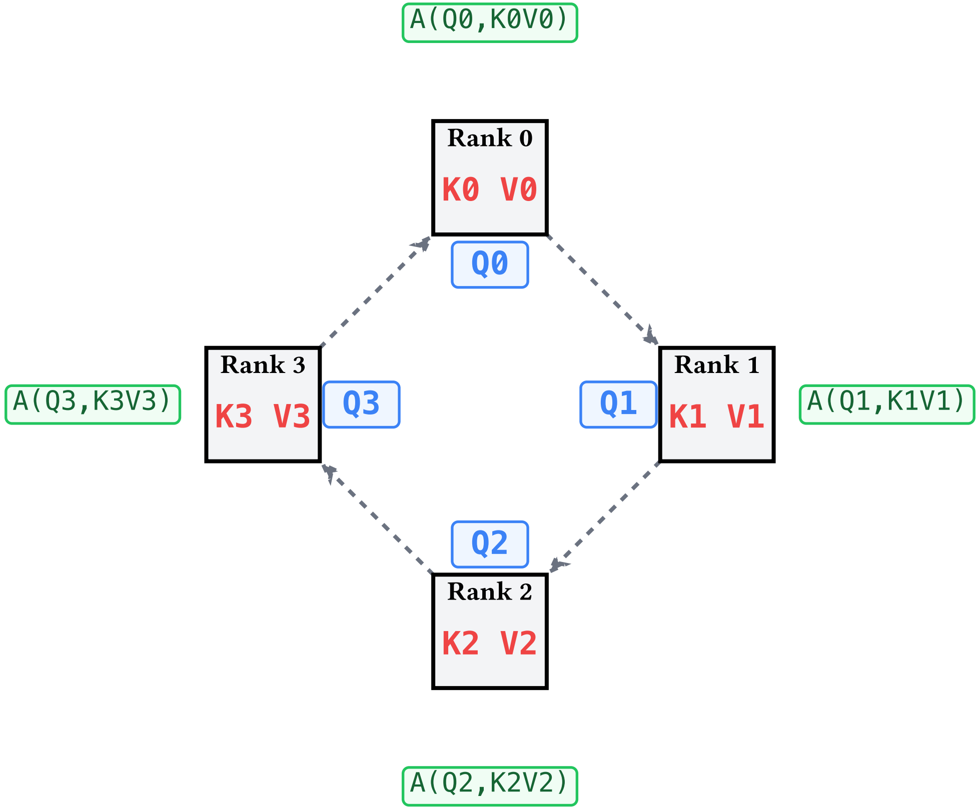
From Meta using 128xH100 with IFB and TCP.

For small T, queries are much smaller than KV!

What if we ring-pass queries?

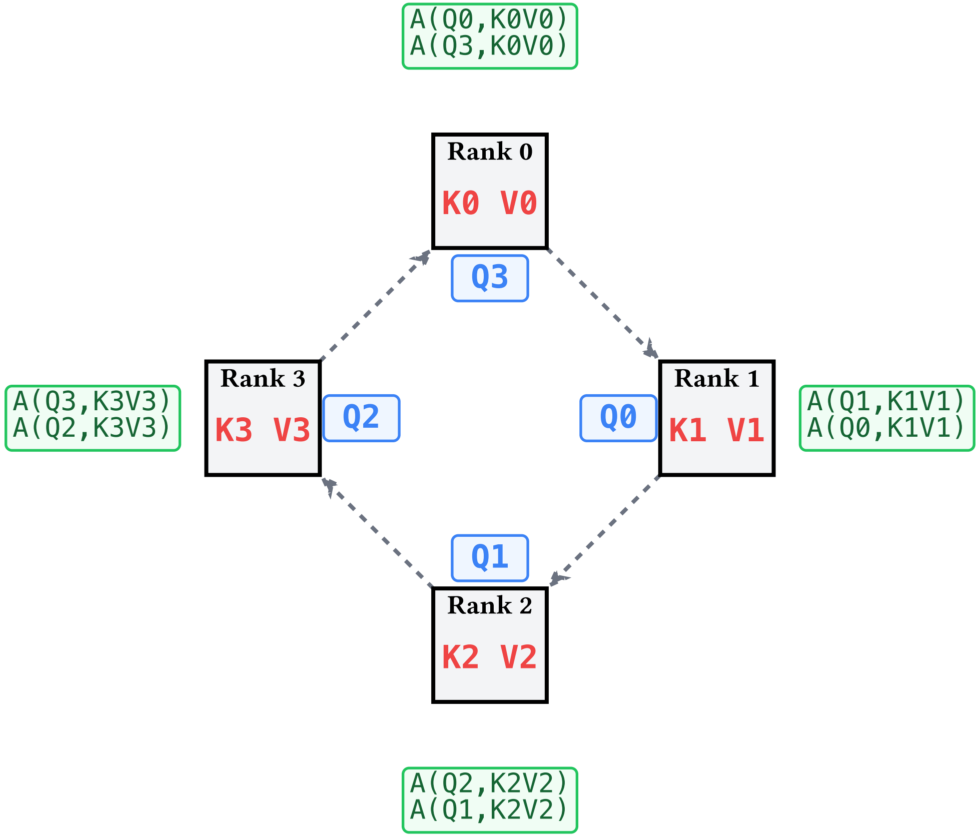
Ring Attention: Pass-Q

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



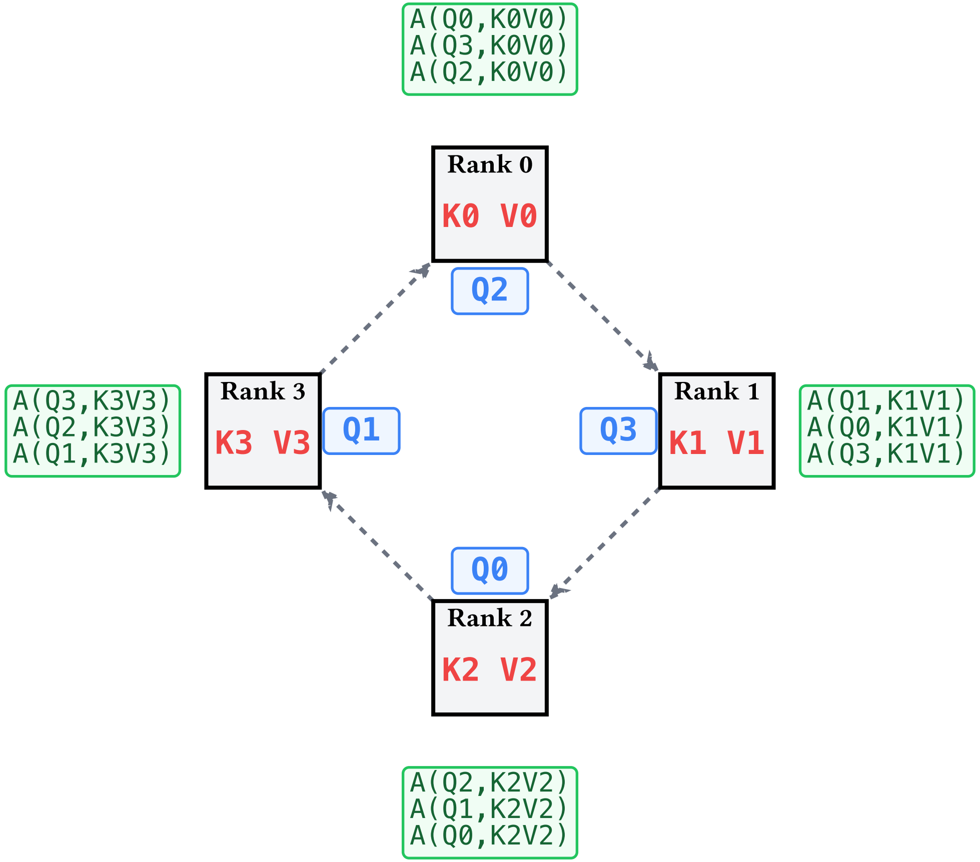
Ring Attention: Pass-Q

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



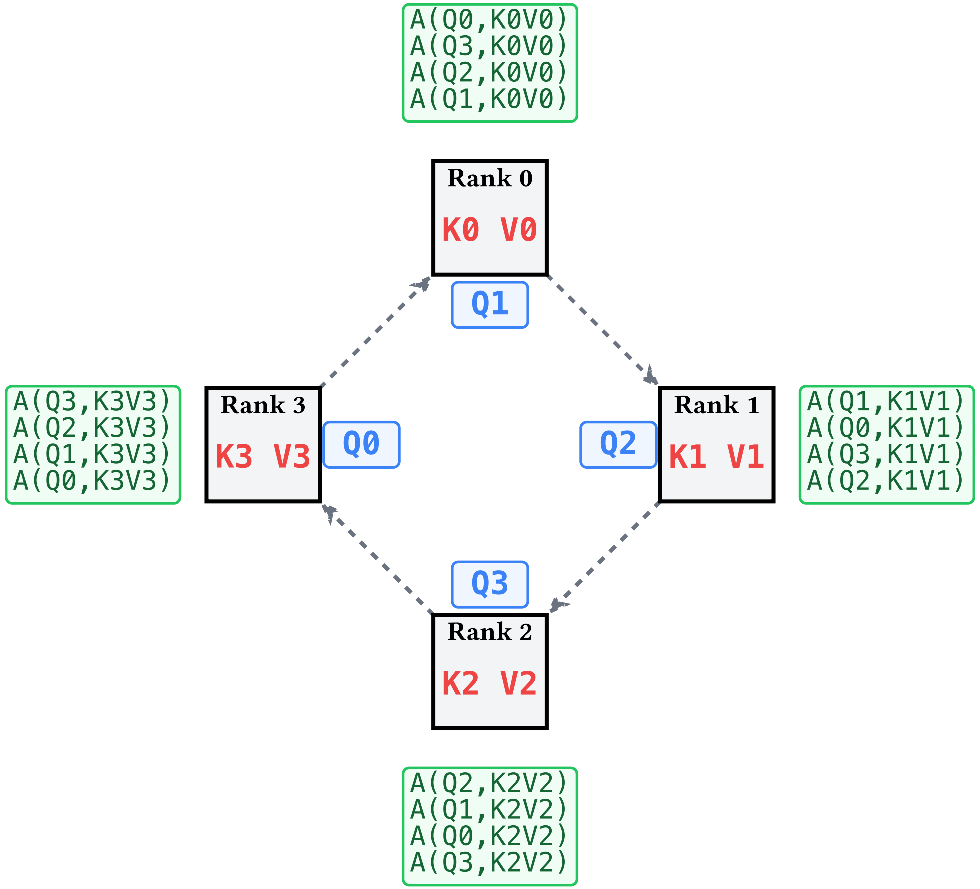
Ring Attention: Pass-Q

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



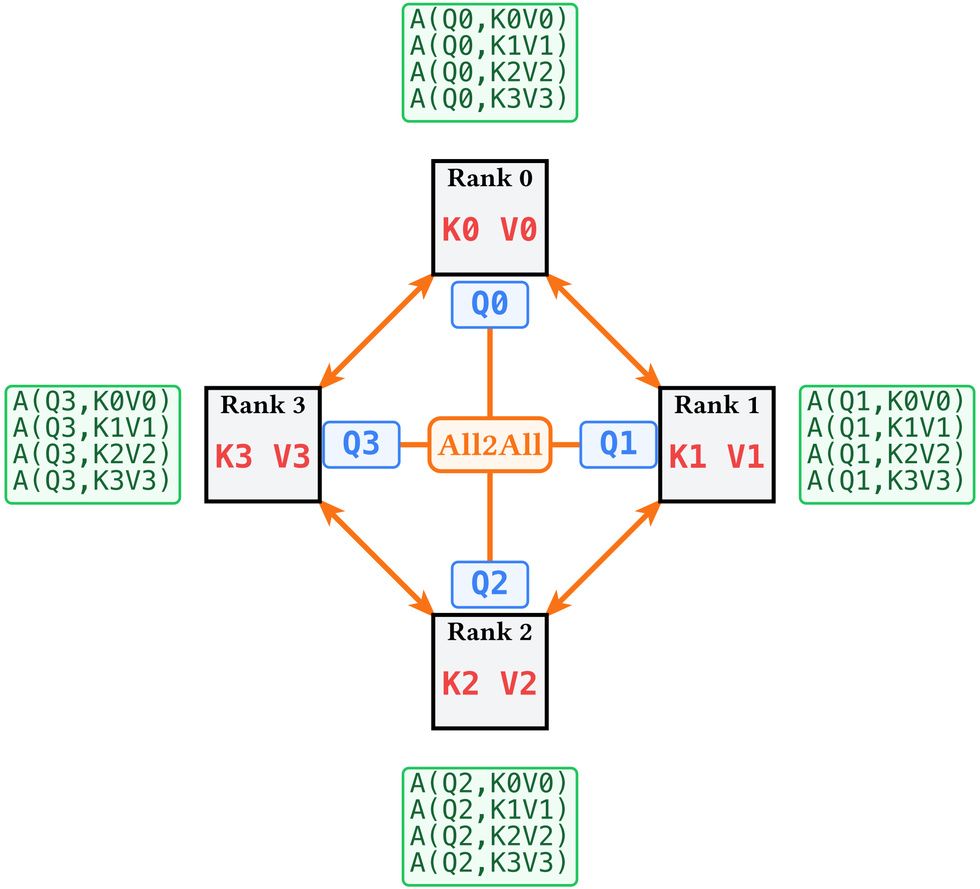
Ring Attention: Pass-Q

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



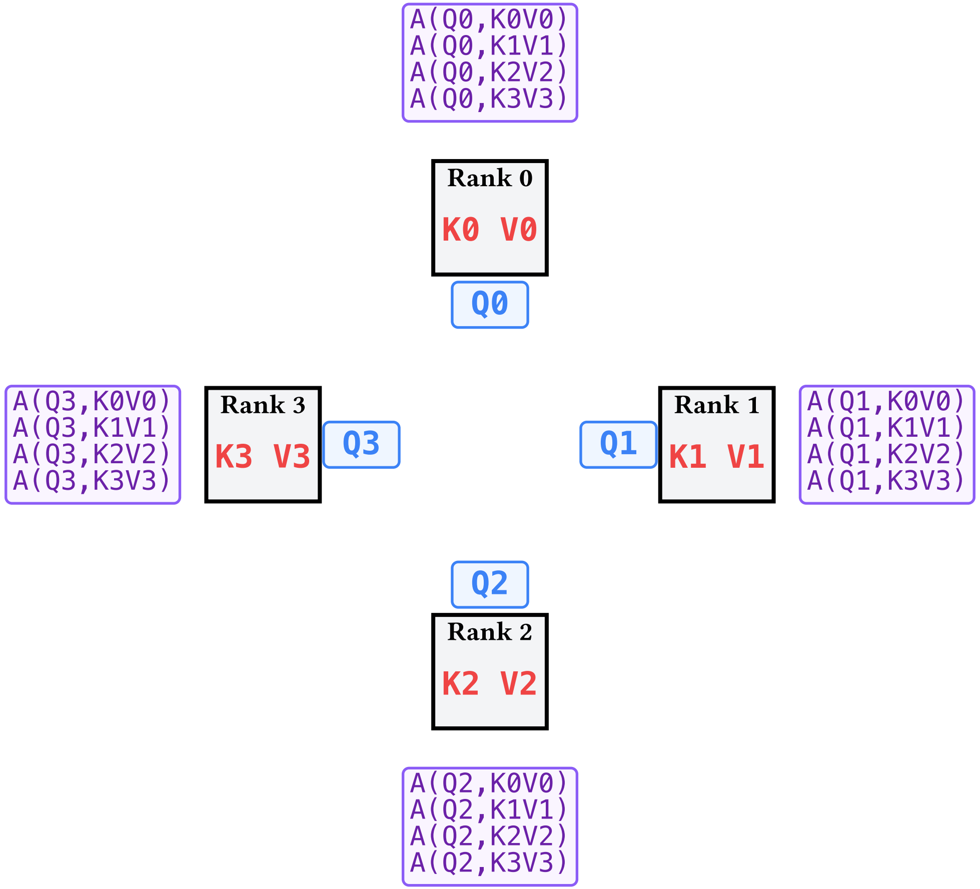
Ring Attention: Pass-Q

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

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

Q-Passing Roofline

Overlap condition:

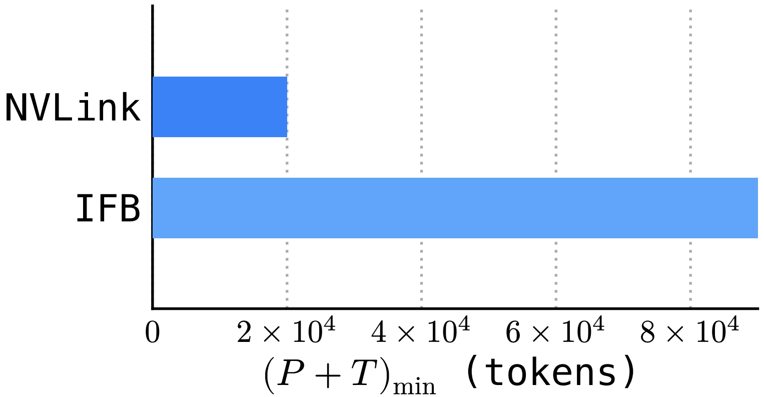
$$\frac{2T(P+T)D}{CN} \geq \frac{TDe}{\mathrm{BW}}$$

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

Doesn't depend on N_H, N_{KV}

Q-Passing Roofline

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



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

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

This Paper Kinda Stinks

They have a bunch of math mistakes. Ex:

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

This Paper Kinda Stinks

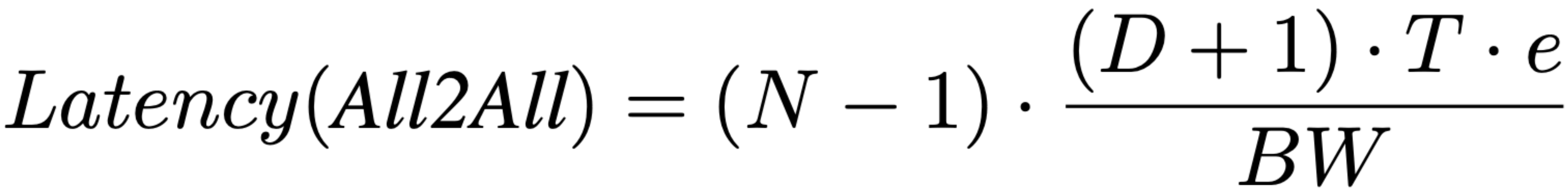
They have a bunch of math mistakes. Ex:

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

This is *so wrong it hurts.*

This Paper Kinda Stinks

They have a bunch of math mistakes. Ex:



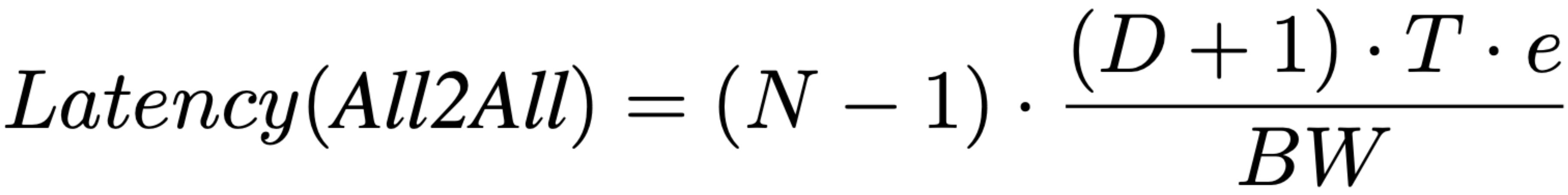
$$Latency(All2All) = (N-1) \cdot \frac{(D+1) \cdot T \cdot e}{BW}$$

This is *so wrong it hurts.*

They also don’t actually solve their inequalities!

This Paper Kinda Stinks

They have a bunch of math mistakes. Ex:



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

This is *so wrong it hurts.*

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]

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

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

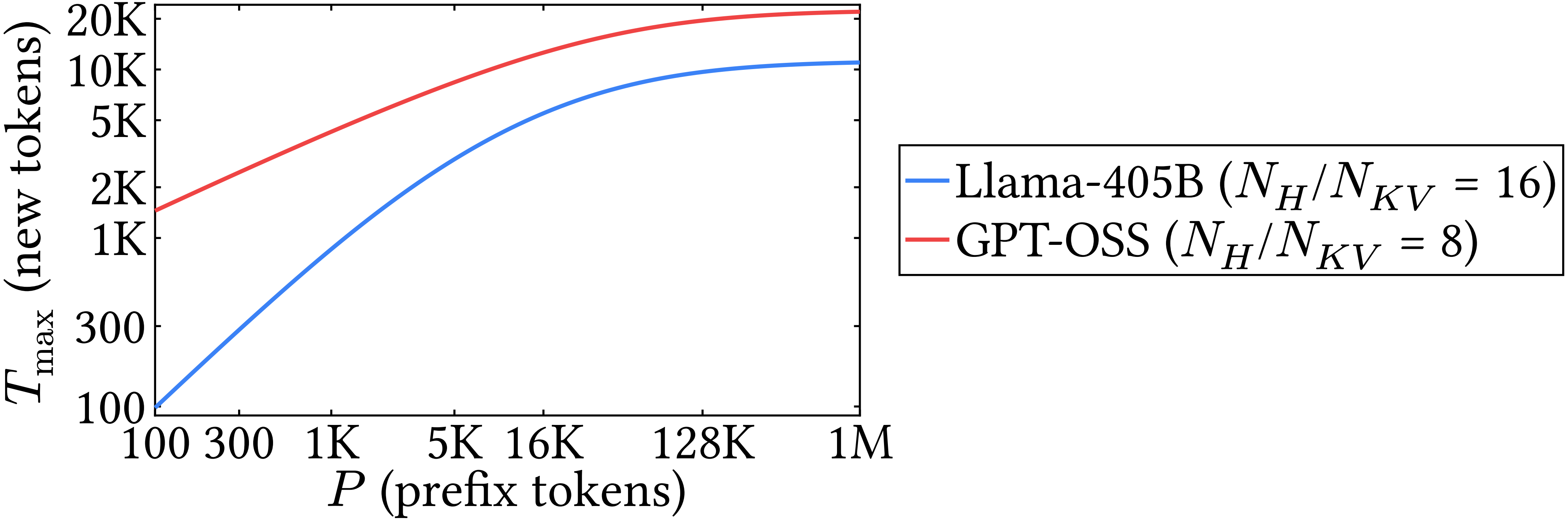
$$\begin{aligned} T_{\text{kv,comm}} - T_{\text{kv,compute}} &= 2(P+T)D\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{2(P+T)TD}{NC} \\ &= 2(P+T)D\left(\frac{N_{KV}}{N_H}\frac{e}{\text{BW}} - \frac{T}{NC}\right) \end{aligned}$$

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

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

$T_{\max}$ vs Prefix Length

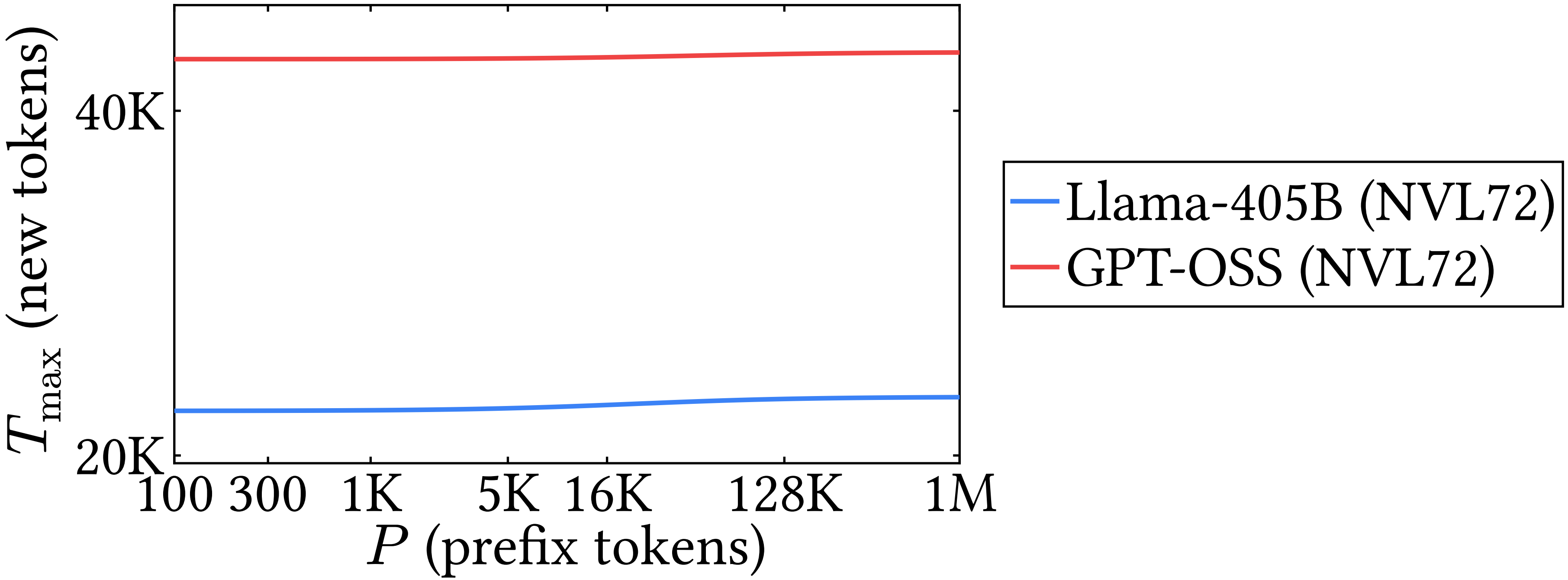
Maximum T where Pass-Q is faster than Pass-KV (8 × B200 IFB):



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

What about NVL72?

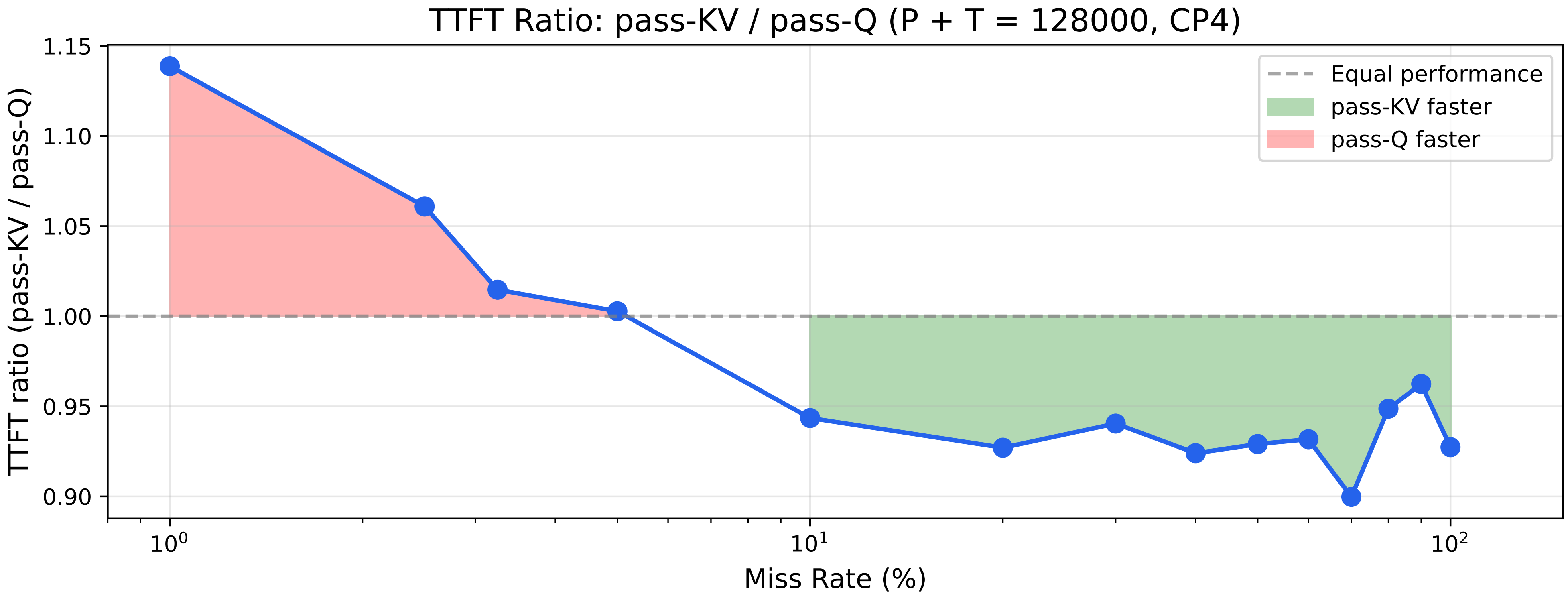
All-to-all cost is $\frac{TDe}{\boldsymbol{N} \cdot 4 \cdot \text{ BW}}$ and BW is 900GB/s v.s. 200GB/s.



With NVL72′s lower all-to-all cost, Pass-Q is optimal up to larger T.

But we only fail to hide communication for fairly small T.

Meta’s Hopper Performance

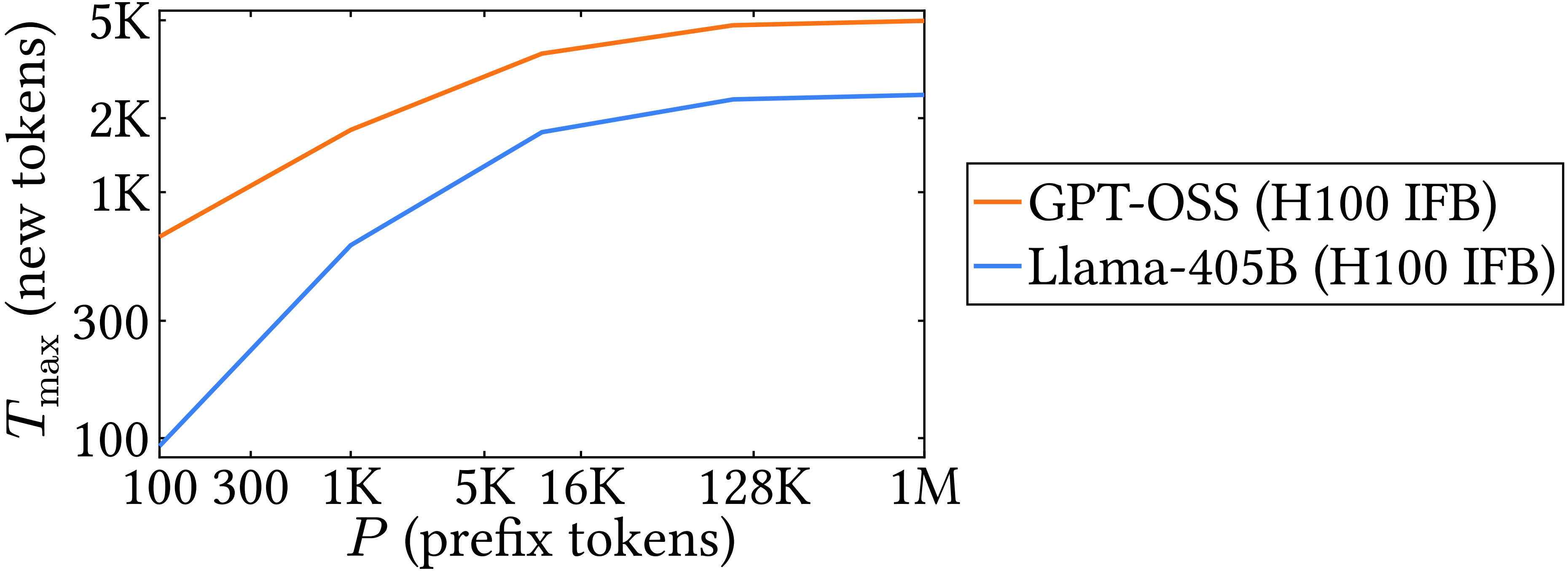


pass-Q is slighly better than pass-KV for very low miss rates

Does the math say the same?

Validating Meta's Hopper Performance

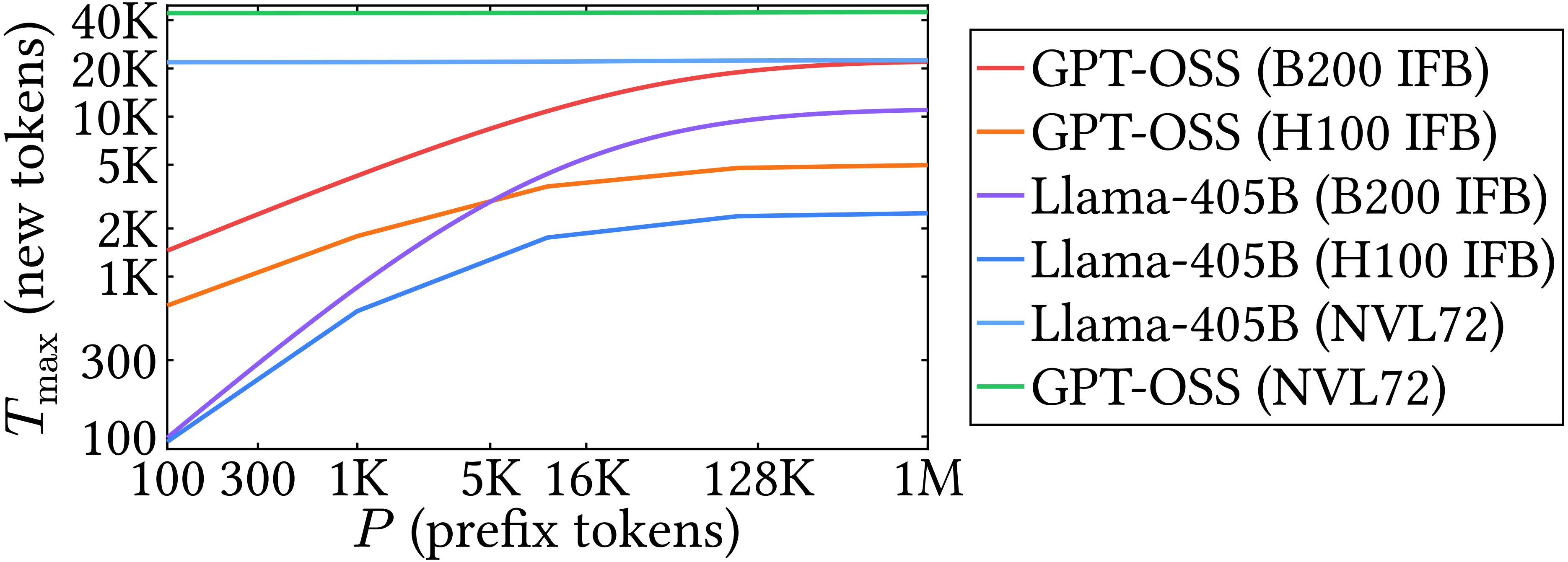
4 x H100 with 200GB/s (unidirectional) IFB



For H100 IFB at P=128K, $T_{\max} \approx 5K \approx 4\%$ miss rate! Empirically $\approx 5\%$!

Putting it all together

All $N = 8$



NVL72 pass-Q may be optimal, but T which isn't hideable is small.