CoCoDiff:

Optimizing Collective Communications for Distributed Diffusion Transformer Inference under Ulysses Sequence Parallelism

Bin Ma[1], Xingjian Ding[1], Tekin Bicer[2], Pengfei Su[1], Dong Li[1]

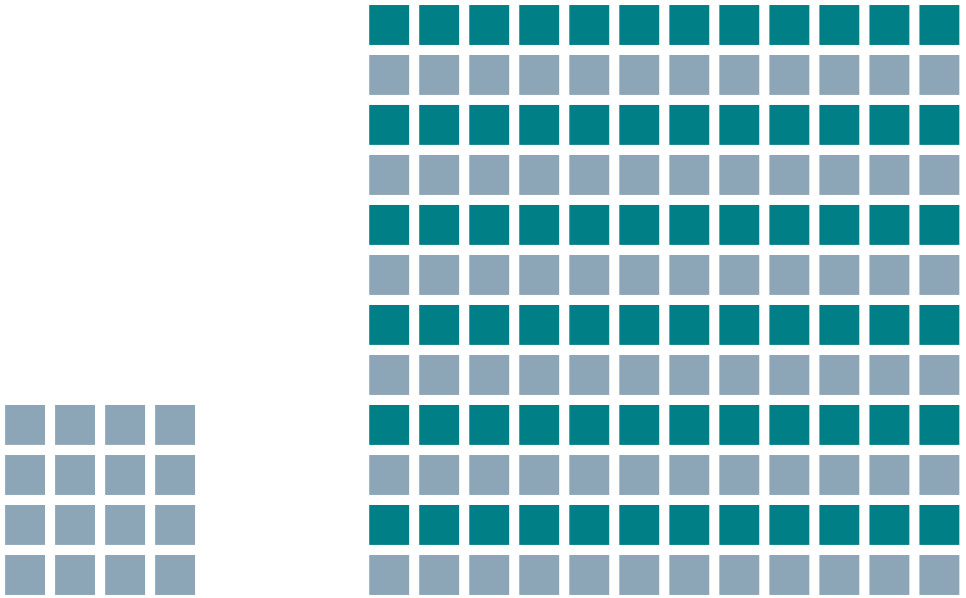
[1] University of California, Merced

[2] Argonne National Laboratory

UCMERCED

Argonne NATIONAL LABORATORY

High-resolution DiTs need many GPUs



768 × 768
2,304 tokens

2304 × 2304
20,736 tokens

9× more tokens

so inference is split across GPUs

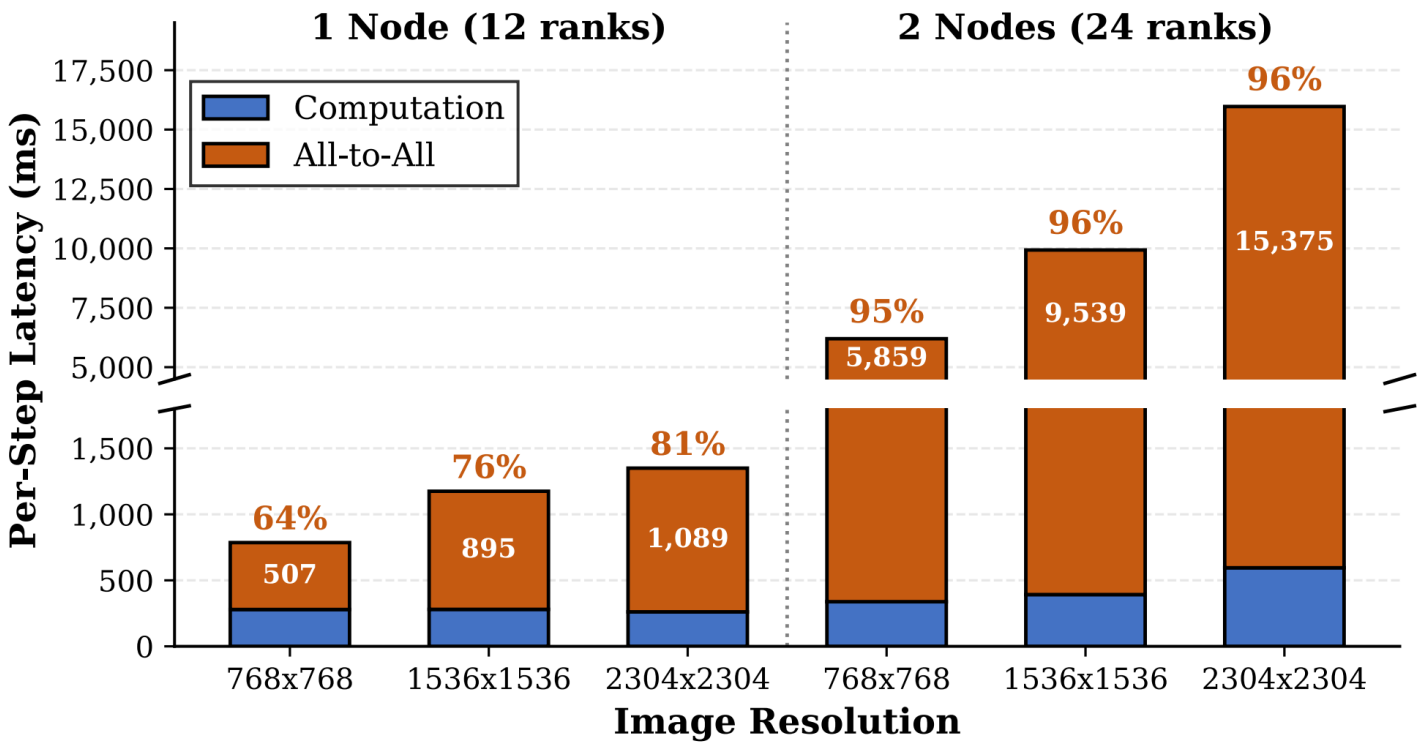
2 all-to-alls

in every layer, at every step

Qwen-Image: 60 layers × 50 steps × 2 ≈ 6,000 all-to-alls

Paper §I, §II-A, §IV-A · tokens = (H/16)² · grids drawn to area scale

All-to-all dominates inference time



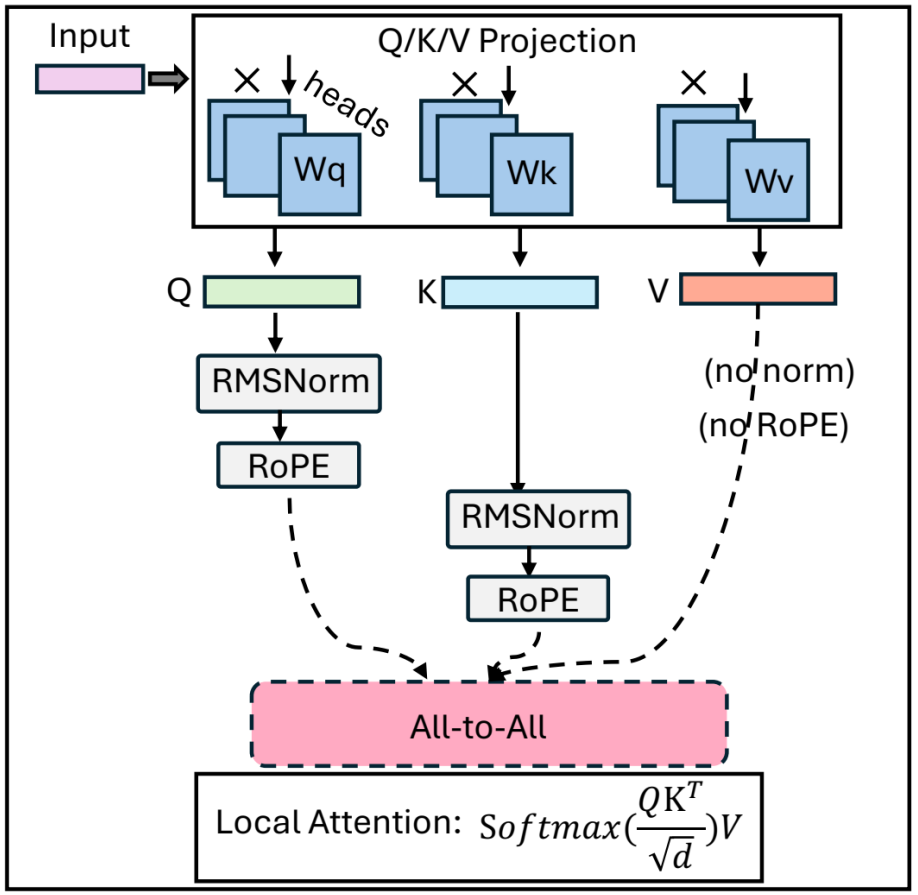
96%

of inference time

Attention waits for communication

Paper Fig. 1 · Qwen-Image · batch size 1

Q/K and V take different paths



Q / K

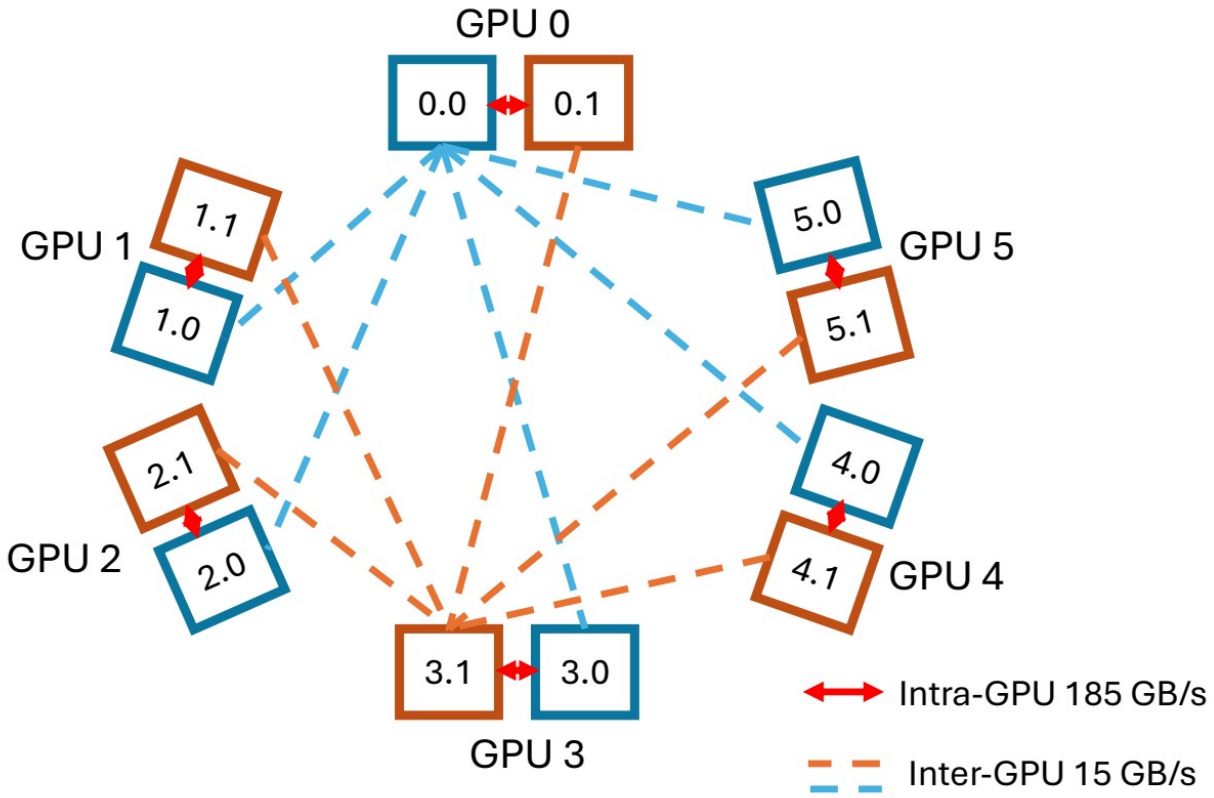
Projection + Norm + RoPE

V

Projection only

V is ready earlier

Aurora links differ by 12×



185 GB/s

inside a GPU

15 GB/s

between GPUs

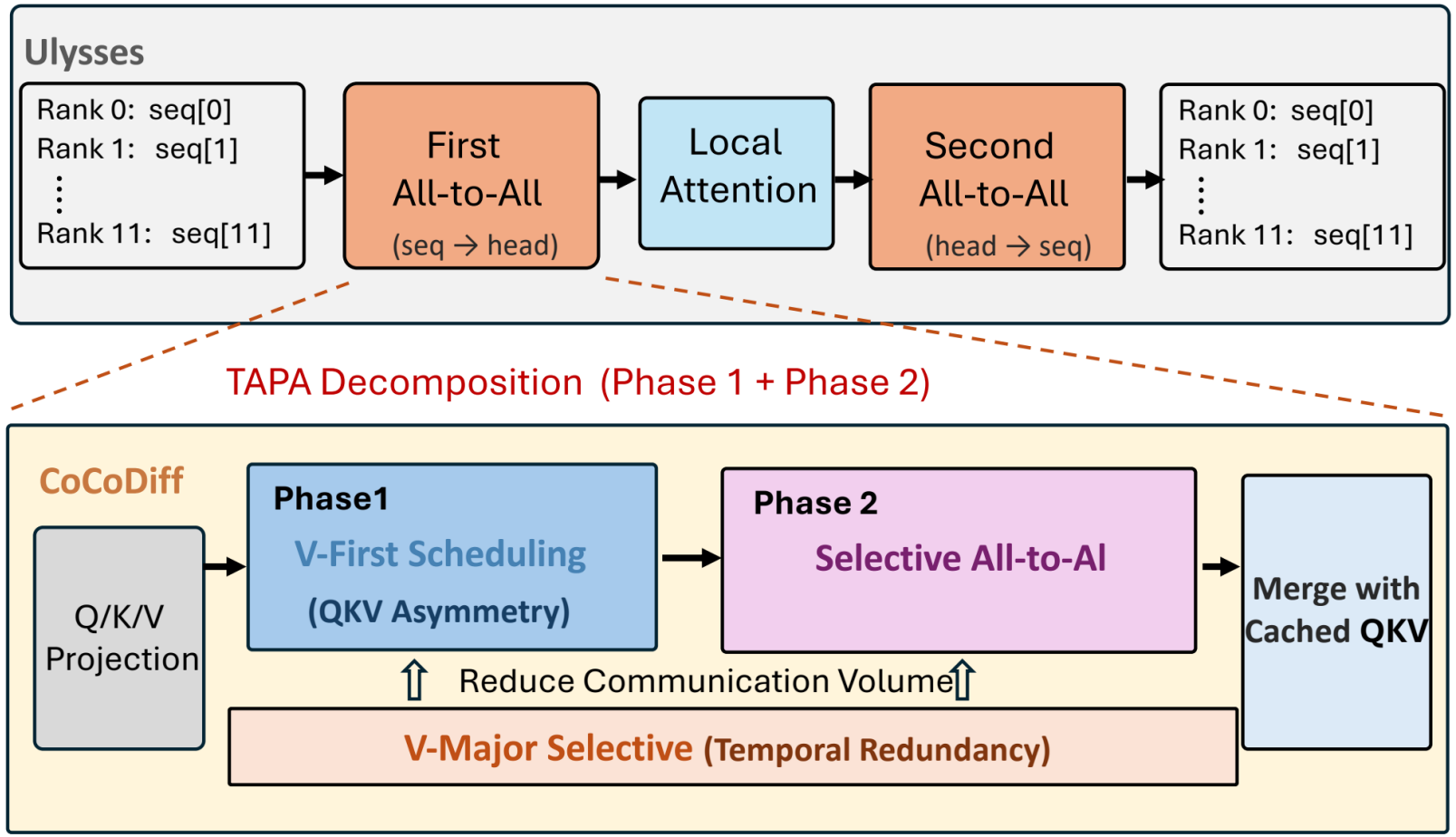
6 GPUs × 2 tiles = 12 ranks per node

Where CoCoDiff fits

	Optimizes
Distributed DiT inference DistriFusion, PipeFusion, AsyncDiff, xDiT, PipeDiT	parallel strategy
Cache-based acceleration DeepCache, Δ-DiT, Learning-to-Cache, TaylorSeer	compute on one GPU
Sequence parallelism DeepSpeed-Ulysses, Ring Attention, USP, LoongTrain	attention layout
CoCoDiff TAPA, V-First, V-Major	**the all-to-all itself**

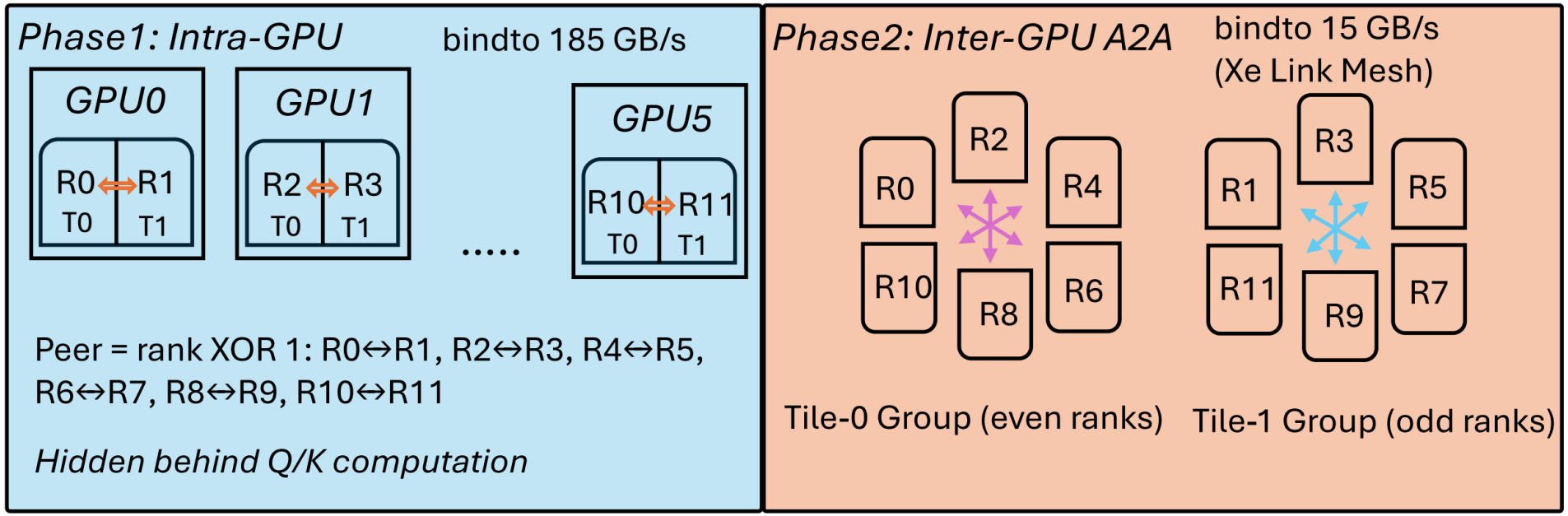
CoCoDiff is complementary to all three

CoCoDiff combines three designs



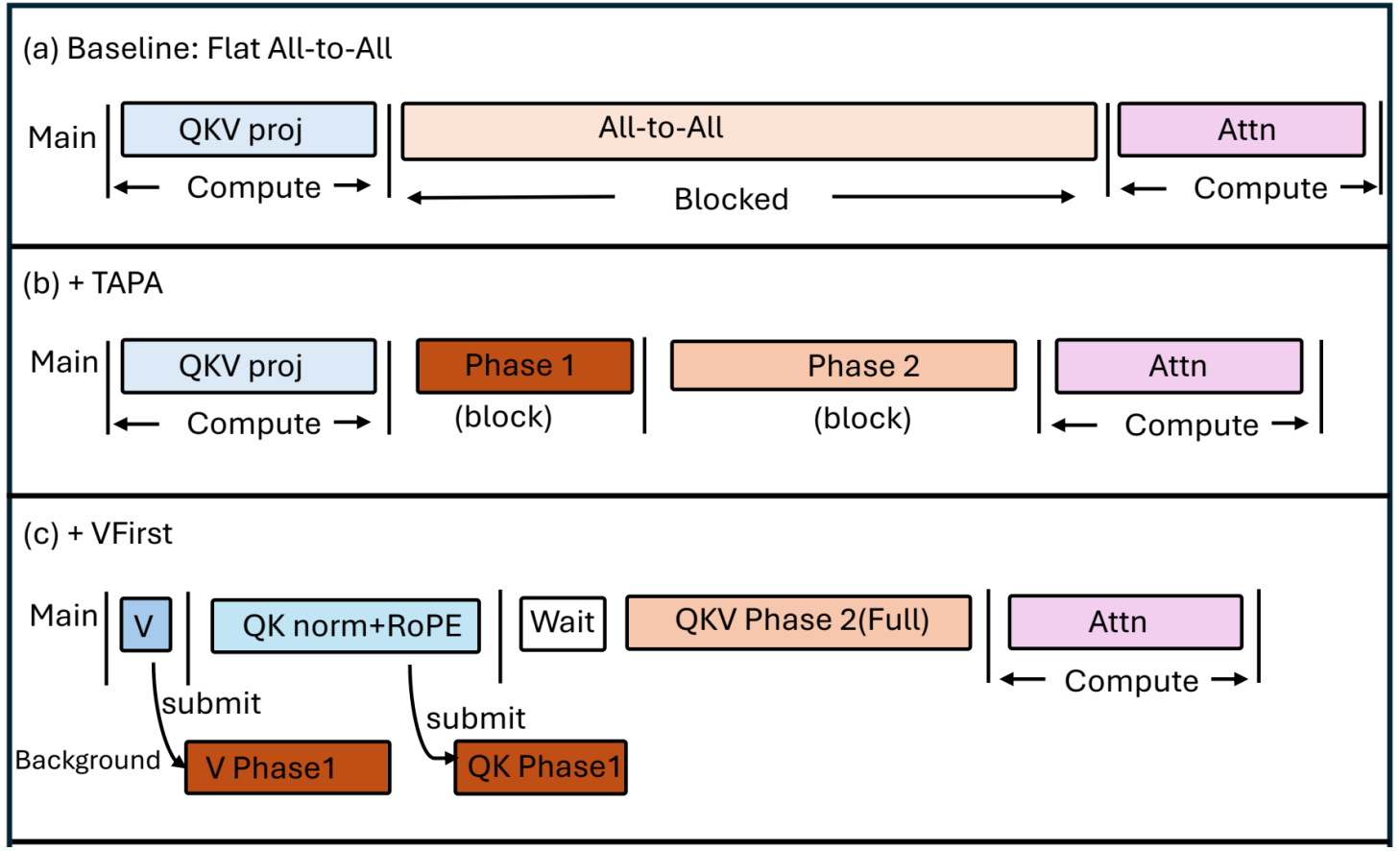
Topology → Overlap → Selective communication

TAPA aligns all-to-all with the topology



Fast local staging → parallel inter-GPU exchange

V is ready early. Start its communication.



Baseline

Blocked on all-to-all

TAPA

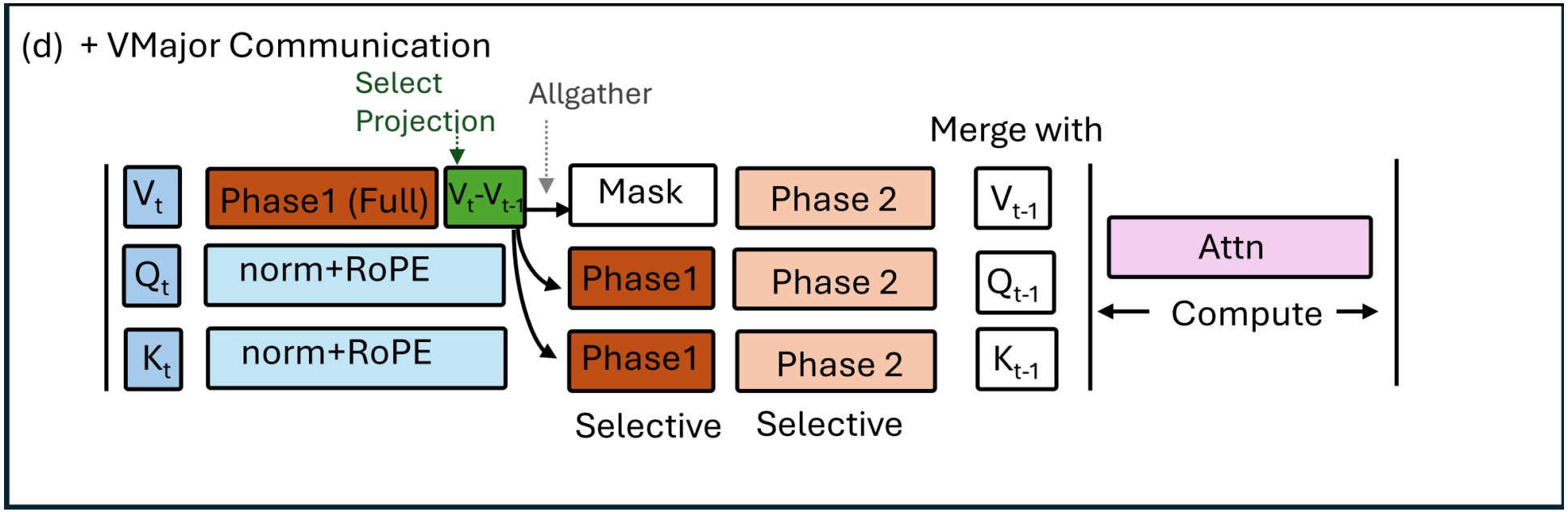
Exposes two phases

V-First

Overlaps Phase 1

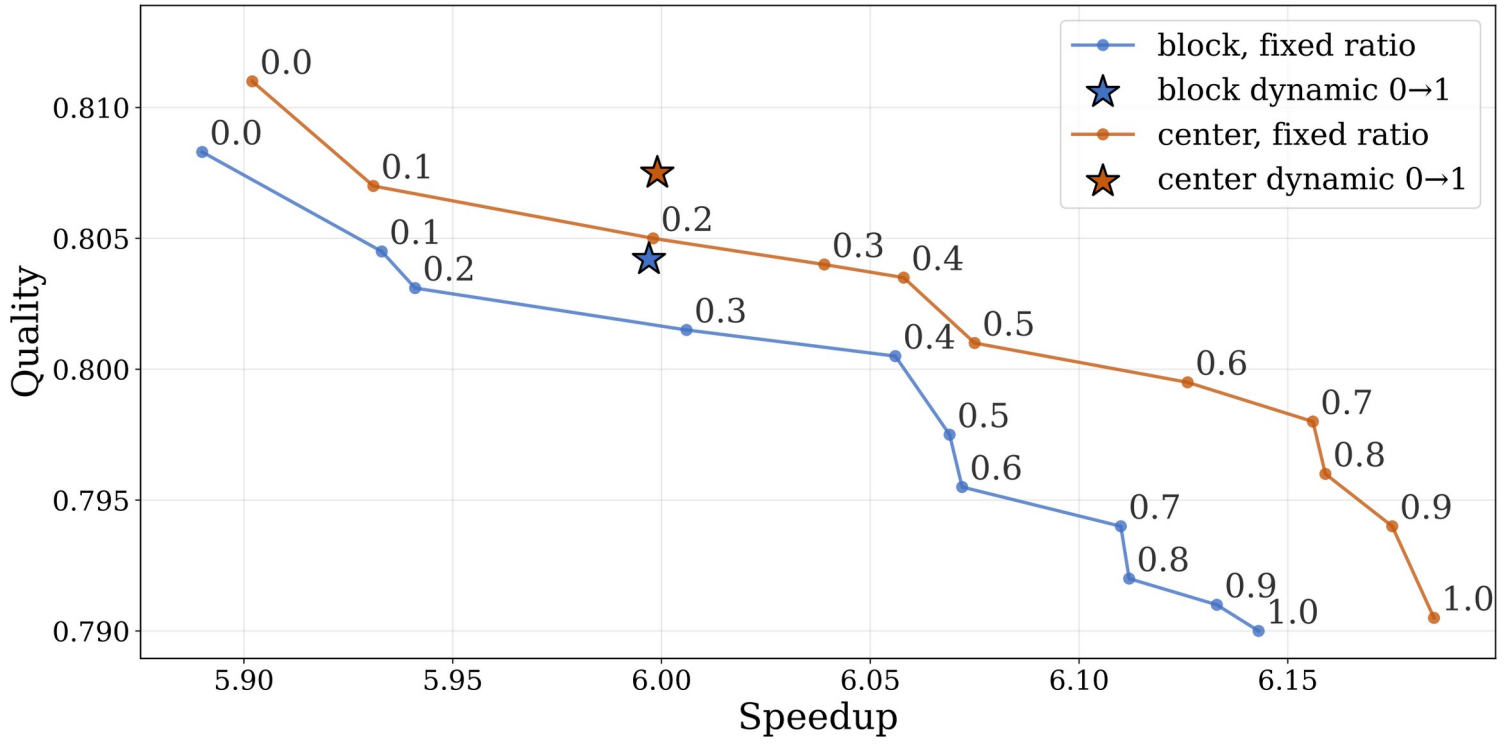
1.0 ms of Q/K processing hides 0.3 ms of V exchange

Send changed projections. Reuse the rest.



Full V exchange first. Selective Q/K/V exchange next.

A changing cache ratio improves the trade-off

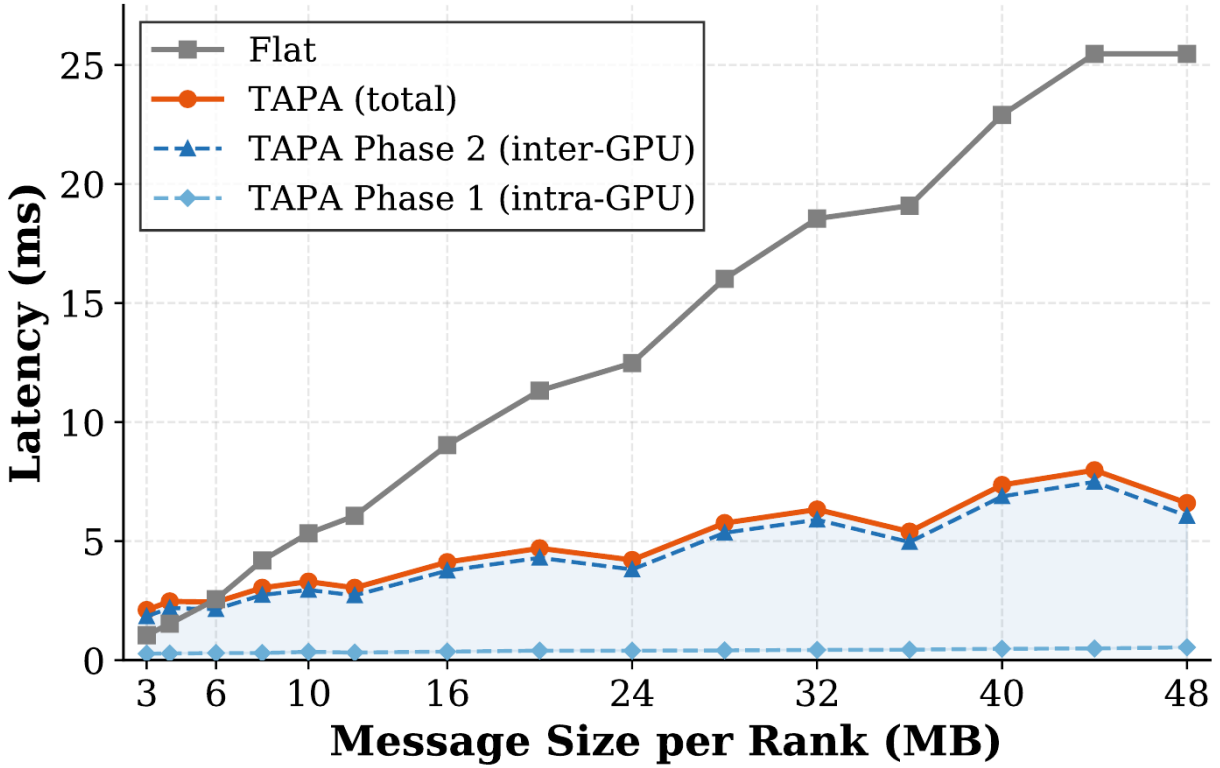


Dynamic ratio

Higher quality at similar speed

Warm up 5 steps
Cache more over time
Refresh every 10 steps

TAPA is faster than oneCCL for large messages



3.9×

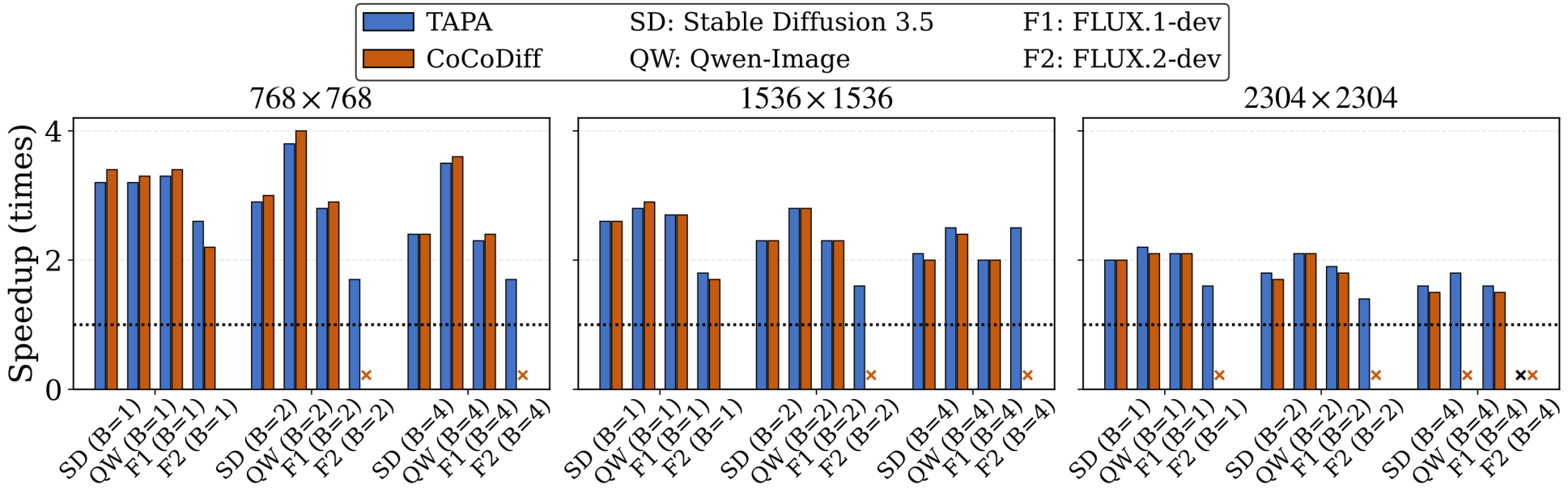
at 48 MB per rank

≈ 1×

at 6 MB per rank

Phase 1 stays at 0.3–0.5 ms for every message size

Single-node gains mainly come from TAPA

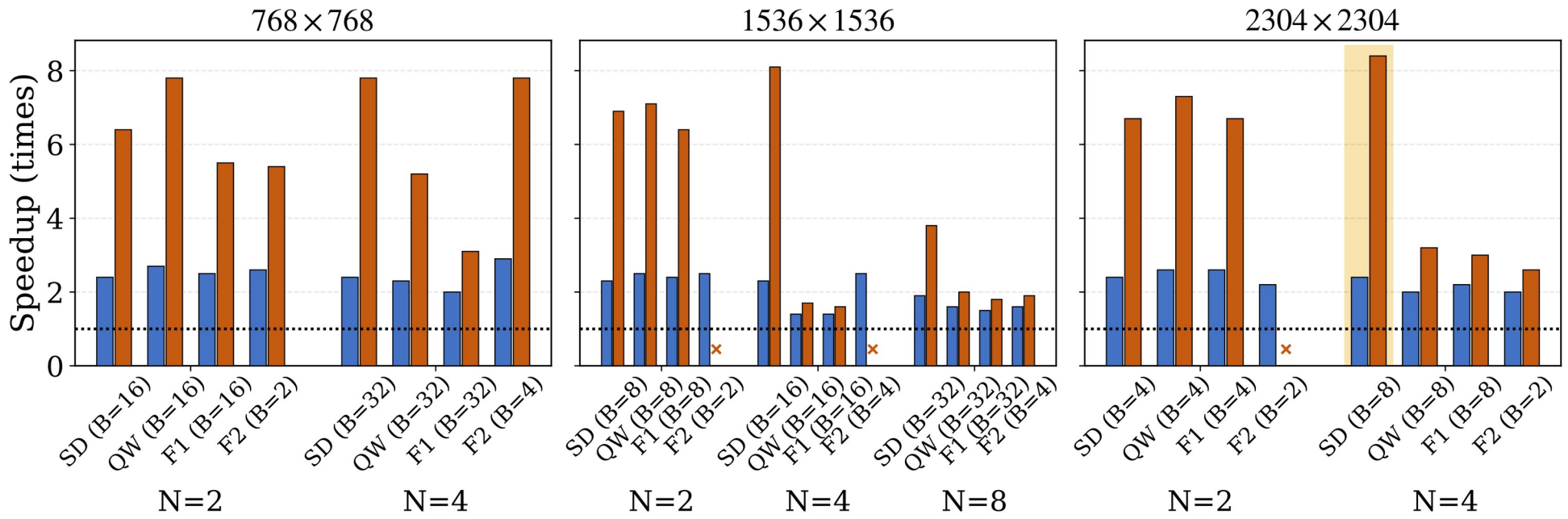


CoCoDiff 2.5× · TAPA 2.3× on average

TAPA | V-First | V-Major

The largest gains come at multiple nodes

TAPA CoCoDiff **3.6×** overall avg. **8.4×** peak

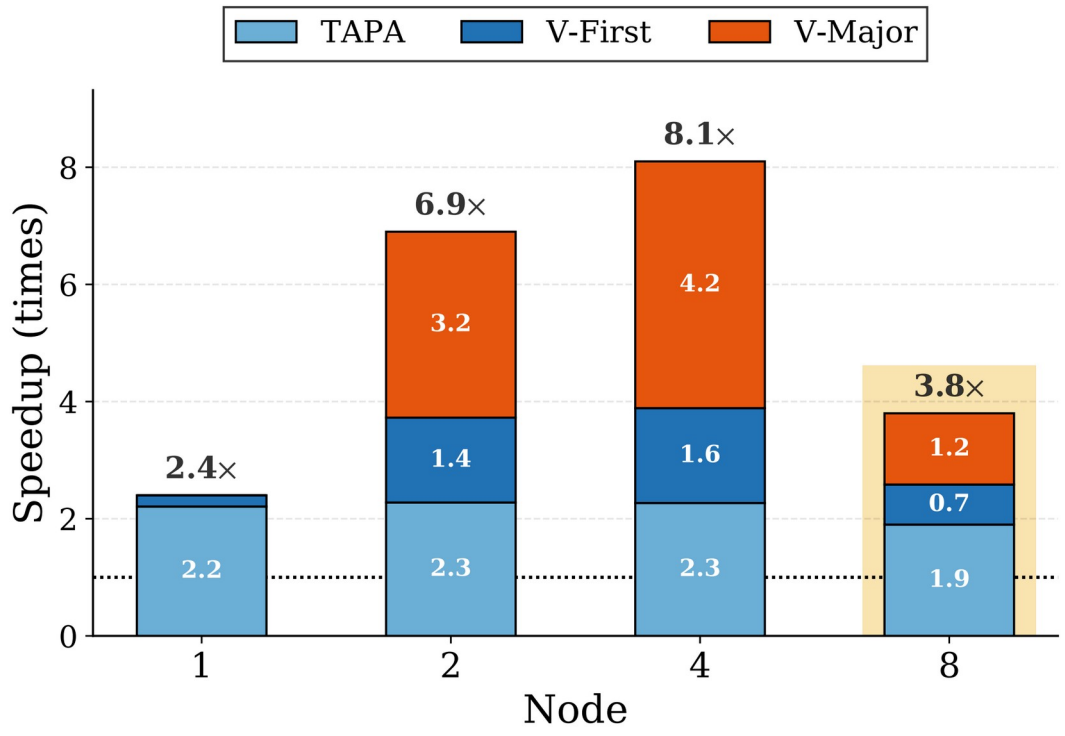


SD: SD3.5 · QW: Qwen-Image · F1 / F2: FLUX.1 / 2-dev

Paper Fig. 7b · over Flat · headline average & peak span all evaluated configurations

TAPA | V-First | V-Major

More nodes do not always mean more speedup



4 nodes



Ulysses = 48

8 nodes

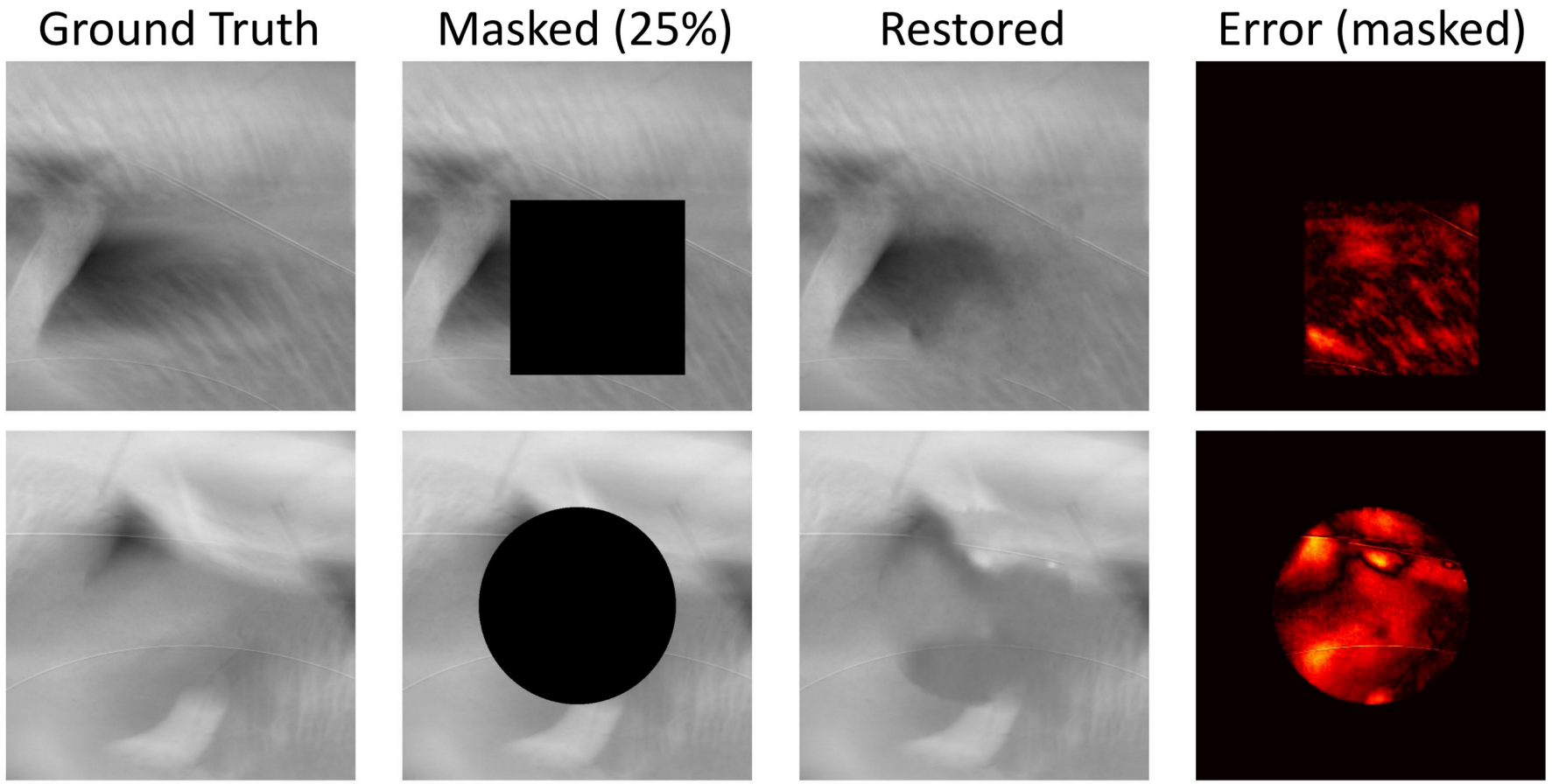


Ring × 2

Ulysses reaches its head-count limit; Ring traffic remains

Paper Fig. 8 · original ablation plot · SD3.5 · 1536×1536

Scientific inpainting retains comparable quality



Center-mask SSIM: 0.8717 (Flat) → 0.8672 (CoCoDiff)

Where CoCoDiff helps less

Ring traffic

SD3.5: 8.1× at 4 nodes, 3.8× at 8

Cache memory

4.3 GB per tile (6.6%) on Qwen-Image

One node

2.5× vs. 2.3× with TAPA alone

Approximation

largest SSIM drop: 0.032 (SD3.5)

Next: selective transfer for Ring traffic

Conclusion： make communication part of the model design.

TAPA



Use the topology

V-First



Hide the latency

V-Major



Reuse stable data

3.6× average speedup **8.4×** peak speedup **Questions?**

Quality stays close to Flat on every model

		SD3.5			FLUX.1-dev			Qwen-Image		
Mask	Metric	Flat	TAPA	CoCoDiff	Flat	TAPA	CoCoDiff	Flat	TAPA	CoCoDiff
Block	PSNR (dB)	23.05	23.11	**22.77**	26.20	27.07	**25.05**	29.97	29.63	**29.17**
	SSIM	0.8036	0.8083	**0.8042**	0.8879	0.8928	**0.8816**	0.9256	0.9248	**0.9229**
	MSE	0.0065	0.0051	**0.0081**	0.0026	0.0020	**0.0032**	0.0011	0.0012	**0.0012**
Center	PSNR (dB)	23.65	22.15	**21.95**	24.89	24.91	**24.22**	29.46	28.91	**28.56**
	SSIM	0.8390	0.8110	**0.8075**	0.8717	0.8838	**0.8672**	0.9236	0.9232	**0.9233**
	MSE	0.0046	0.0065	**0.0066**	0.0039	0.0035	**0.0042**	0.0012	0.0014	**0.0015**

Largest gap: SD3.5, center mask (1.7 dB PSNR, 0.032 SSIM)

Paper Table II · X-ray microtomography inpainting (APS) · 25% masks · FLUX.2-dev has no Diffusers inpainting pipeline

What V-Major's cache costs

4,320 MB per tile

Qwen-Image 1536², B=4, one node

Same at 8 nodes

B=32 spread over 96 tiles

6.6%

of the 64 GB tile memory

Runtime cost

selection, caching, all-gather

On one node, these costs offset most of the savings