# Understanding & Improving Multi-Rack Symmetric-Memory AllGather in NCCL

Prajwal Singhania[1,2], Vijay Anand Korthikanti[1], Jared Casper[1], Mohammad Shoeybi[1], Abhinav Bhatele[2]
[1]NVIDIA, Inc. [2]Department of Computer Science, University of Maryland

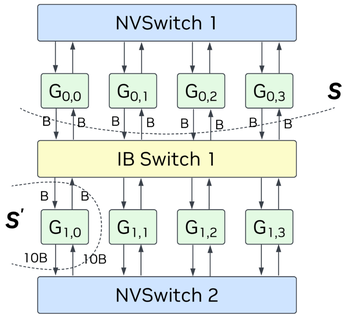
## 1 Abstract


- GB200 NVL72 grows the scale-up domain to 72 NVLink-connected GPUs — but LLM training still scales out across many such racks.
- a ZeRO-3 implementation all-gathers each parameter on the fly over a heterogeneous network both in the forward and backward pass.
- With high scale-up and scale-out bandwidths, these all-gather message sizes fall into a regime where both latency and bandwidth matter.
- We study and improve symmetric NCCL All-gather performance for such message sizes on heterogeneous multi-rack networks.


## 2 Speed-of-Light for Bandwidth-Bound AllGather

- Homogeneous Networks: All links have the same bandwidth; measured *bus bandwidth* can be directly compared against link bandwidth (SOL).
- Heterogeneous NVLink + IB Networks: No single link rate to compare against — the SOL bound must come from the topology itself.

Network Graph → $G = (V, E)$
$V = V_s \cup V_g$ (switches ∪ GPUs)
Message size $M$
$B^+(S)$ = total bandwidth leaving cut $S$




- ForestColl [1] bounds AllGather time via the ***bottleneck-cut,*** computed by max-flow for arbitrary networks.

$$T_{comm} \geq \frac{M}{N} \max_{S \subset V, V_g \not\subseteq S} \frac{|S \cap V_g|}{B^+(S)}$$
$$\Leftrightarrow \frac{M}{T_{comm}} \leq \min_{S \subset V, V_g \not\subseteq S} \frac{N \cdot B^+(S)}{|S \cap V_g|}$$

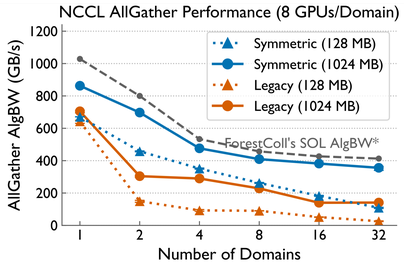
### Our Findings

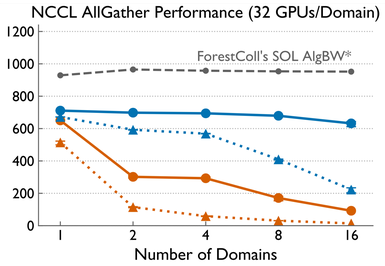
- Provably, for symmetric networks, only two bottleneck cuts exist: Single GPU Cut (**S'**) and Single Domain Cut (**S**).
- Closed-form solution for D NVL72 domains, each with G GPUs (total of N GPUs).

$$SOL\ AlgBW^* = min\left(N \cdot \frac{(B_{NVLink} + B_{IB})}{N-1}, \frac{D \cdot (G \cdot B_{IB})}{(D-1)}\right)$$

## 3 Benchmarking NCCL AllGather on GB200

We benchmark the performance of Legacy & Symmetric NCCL AllGather against the SOL AlgBW*




### Our Findings

- 8 GPUs/Domain is IB-limited (single-domain cut bound); 32 GPUs/Domain is GPU-ingress-limited (single-GPU cut bound).
- Symmetric AllGathers scale and perform significantly better than legacy AllGather.
- For l024 MB messages in the 32 GPUs/domain regime. Symmetric AllGather shows scales well.
- 128 MB messages drop off sharply after 4 domains; a bandwidth-only model cannot explain it.

## 4 A Latency-Aware SOL Model

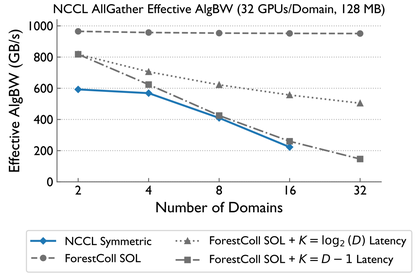
Modeling Idea: Incorporate a latency term in the SOL B/W model

- ForestColl models steady-state messaging:

$$T_{comm} \geq \frac{M/N}{\underbrace{\min_{S \subset V, S \not\supseteq V_g} \frac{B^+(S)}{|S \cap V_g|}}_{x^*}}$$

Effective per-GPU B/W → $x^*$

$$T_{comm} \approx K \cdot \alpha + \frac{M}{N x^*}$$

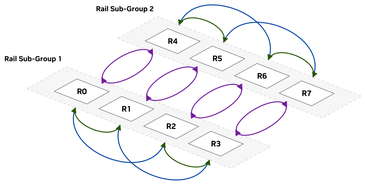



- NCCL Symmetric AllGather follows the ring latency model, which is expected as it is a Rail Ring + NVL multicast implementation.
- We can do better by moving to a log latency approach over the rail.

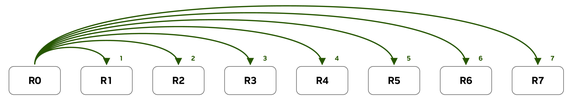
## 5 Two New AllGather Kernels

### Rail Hybrid Recursive-Doubling + Ring with NVLS MC

- For each rail, a few steps of RD followed by Ring between sub-groups.
- Pipelined with multicast distribution within an NVL domain.

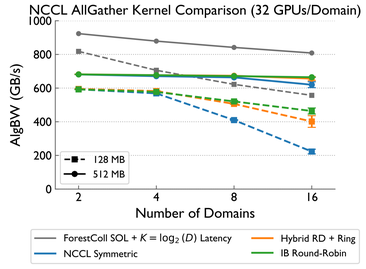


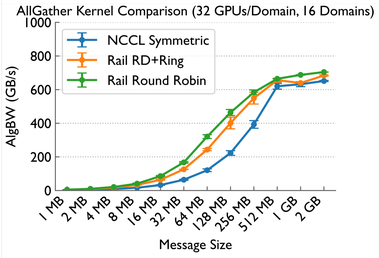

### Rail Round-Robin Broadcast with NVLS MC




- For each rail, every rank puts its chunk directly to all peers, round-robin order — fully pipelined, no head-of-line waiting.
- Pipelined with multicast distribution within a domain.

## 6 Results and Conclusion




Our all-gather kernels scale better than NCCL and closer to the $log_2(D)$ latency aware model

Up-to 2.6x improvement in performance over NCCL AllGather in the 32MB to 512 MB message range

References

[1] Zhao, L., Maleki, S., Wang, Y., Wang, Z., Yang, Z., Pourreza, H., & Krishnamurthy, A. (2026). {ForestColl}:{Throughput-Optimal} Collective Communications on Heterogeneous Network Fabrics. In 23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26) (pp. 2067-2093).