To `ncclsee`, or Not to `ncclsee`: That is the Profiling Question

Ruben Laso[a], Majid Salimi Beni[b], Ioannis Vardas[b], Siegfried Benkner[a], Sascha Hunold[b]

[a] *Faculty of Computer Science, University of Vienna, Austria*
[b] *Faculty of Informatics, TU Wien, Austria*

Distributed deep learning has become the backbone of modern HPC systems, in which multiple accelerators (typically GPUs) continuously exchange data during model training and inference. For such tasks, NCCL (and related libraries such as RCCL, OneCCL, etc.) is the most widely used library for GPU-GPU communication. Despite NCCL being a well-established technology, profiling tools are still in an early stage of development. In this work, we present the latest version of `ncclsee` [1, 2] and compare it against two other profilers, NVIDIA Inspector [3] and Google CoMMA.

We evaluate the three profilers using micro-benchmarks and a real-world DDL application, training a DenseNet121 model using PyTorch and Distributed Data Parallel (DDP). We run our experiments on two nodes of the *Leonardo* supercomputer, each equipped with 4 NVIDIA A100 GPUs interconnected via NVLink, and the nodes are connected via 200 Gbit/s InfiniBand.

In our experiments with micro-benchmarks, we find that Inspector and CoMMA miss collective-operation events, while `ncclsee` records them all (Fig. 1). Therefore, `ncclsee` provides more accurate timing information that closely matches the communication time reported by the micro-benchmarks (Fig. 2); and thus, allows us to get accurate profiling information of AI workloads.

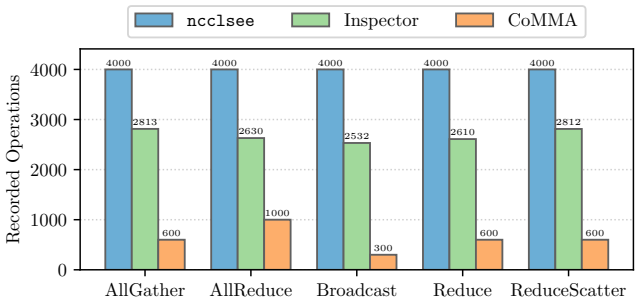

Fig. 1: Amount of collective-communication events recorded for different profilers. Expected value 4000.

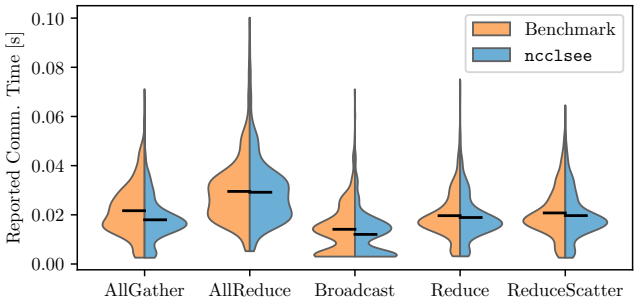

Fig. 2: Communication times reported by micro-benchmarks and `ncclsee`. Message size of 32 MiB.

References

[1] Vardas, I., Laso, R., and Salimi Beni, M. (2025). `ncclsee`: A Lightweight Profiling Tool for NCCL. In ASHPC25 (p. 39). `https://doi.org/10.34726/10426`.

[2] Laso R., Vardas, I., `ncclsee`, `https://github.com/parlab-tuwien/ncclsee`.

[3] Das, S. *et al.* Enhancing Communication Observability of NCCL Inspector. NVIDIA Technical Blog (2025). `https://developer.nvidia.com/blog/enhancing-communication-observability-of-ai-workloads-with-nccl-inspector/`.