TECHNION | The Henry and Marilyn Taub Faculty of Computer Science

Topologies for LLM Training

SimAI Comparisons

Simulations Configurations

Models and Parameters Used

Models used:
- GPT: 7,13,22,175B
- Llama: 405B

World Sizes:
- 64
- 128
- 512
- 1024
- 2048
- 4096

Parallelism Methods:
- Tensor Parallelism
- Pipeline Parallelism
- Data Parallelism
- Expert Parallelism (MoE only)

Batch Size Parameters:
- Global Batch Size
- Micro Batch Size
- Sequence Length

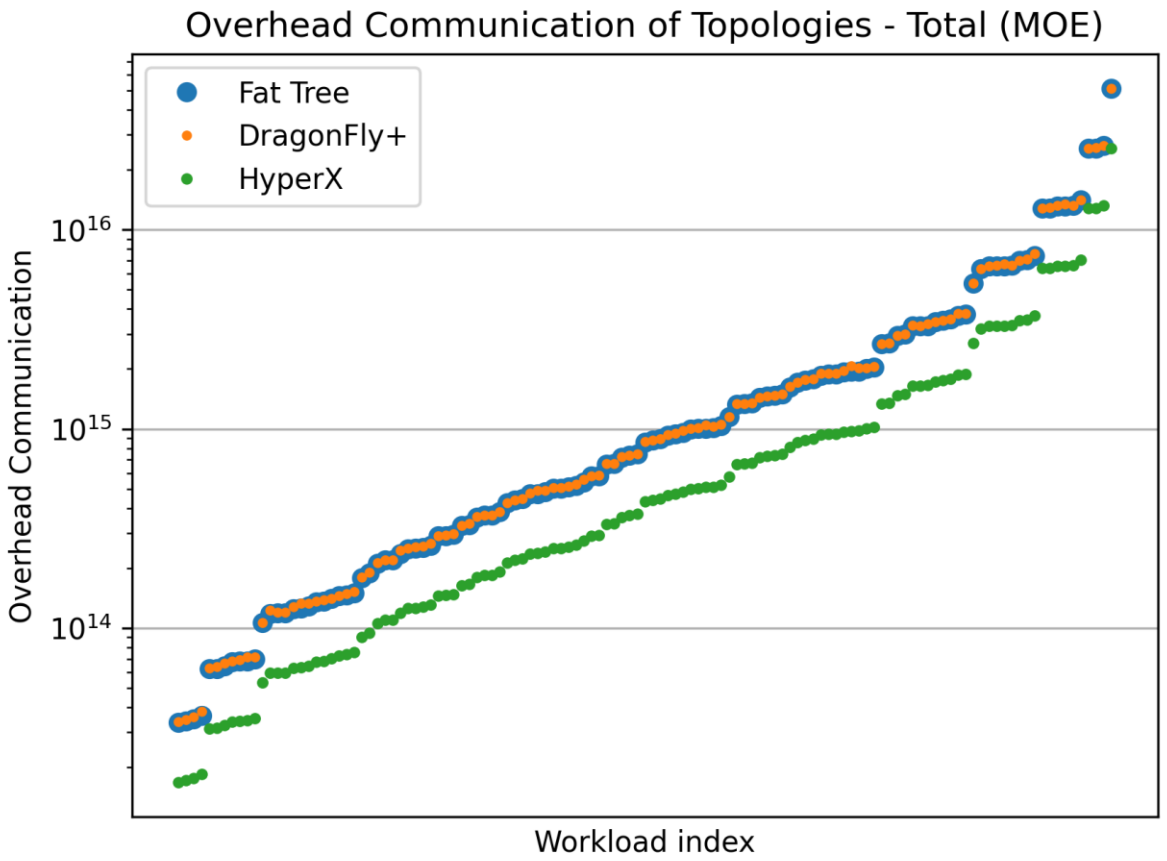
Architectures Compared:
- MoE based (16 experts)
- Densed

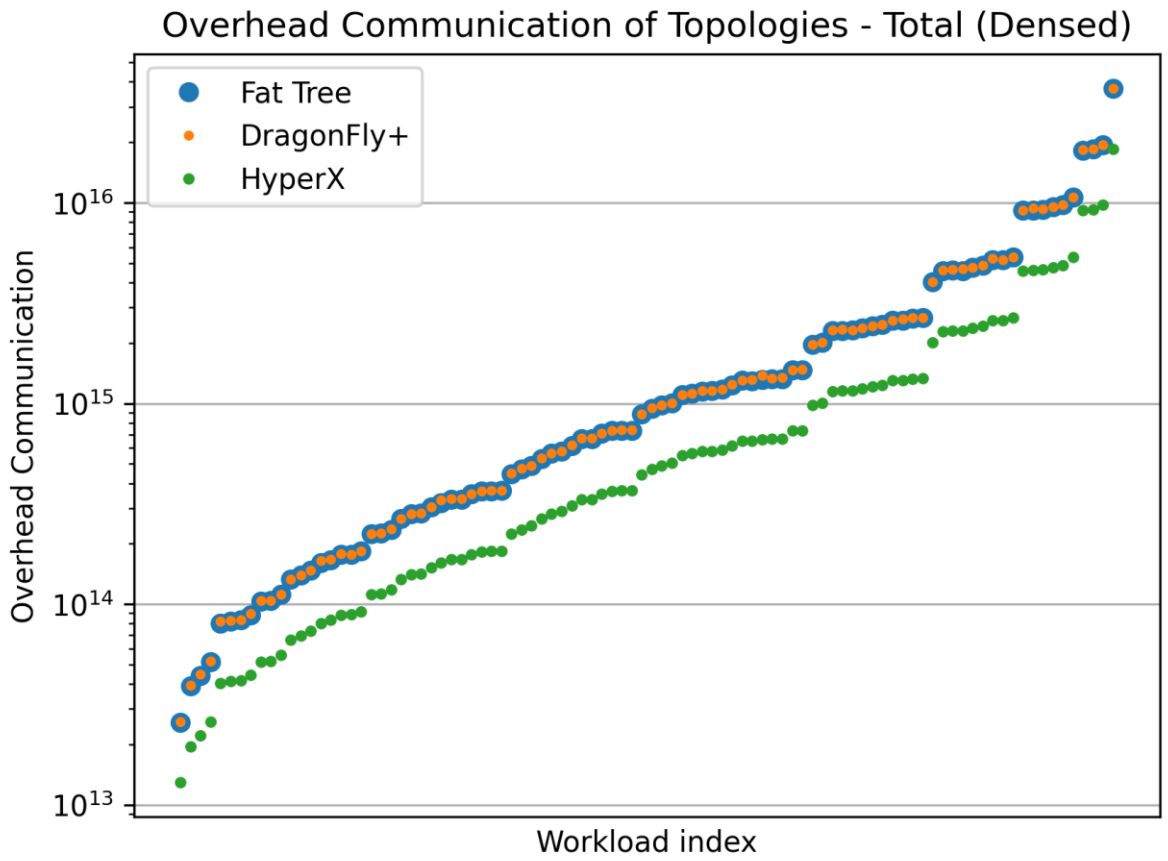
Topologies Compared:
- Fat Tree
- DragonFly+
- Hyperx

Total 123 setups per architecture

Overall Overhead Comparison

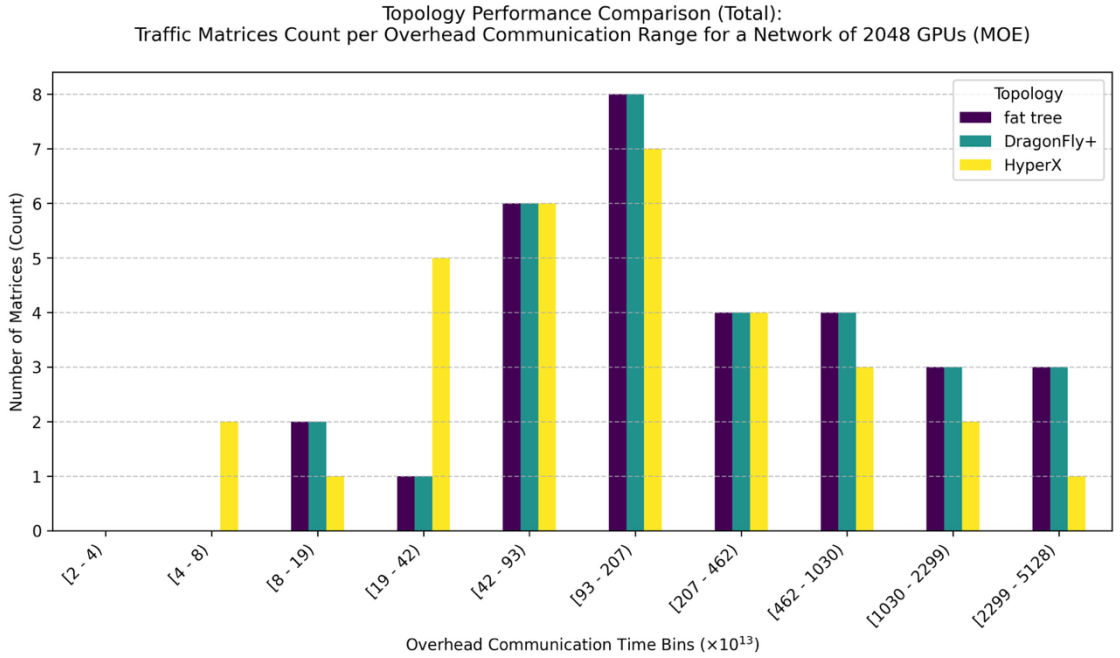
High Level Comparison of Each Topology Over Architecture Framework

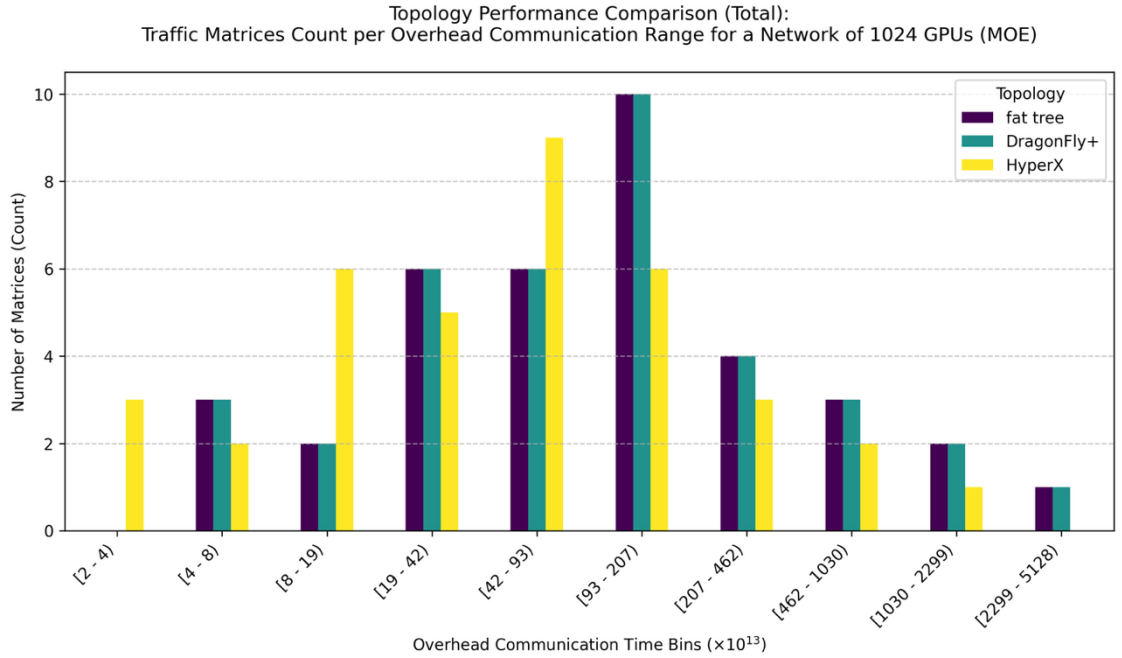




MoE Based Overhead Distribution

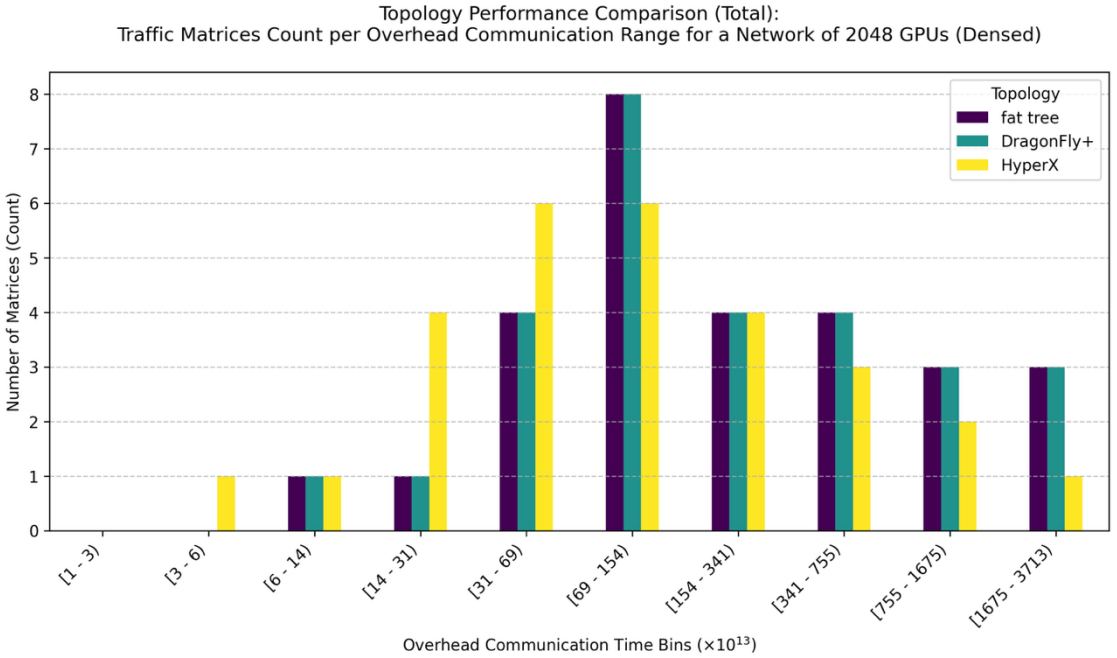
Over 1024 and 2048 World Sizes

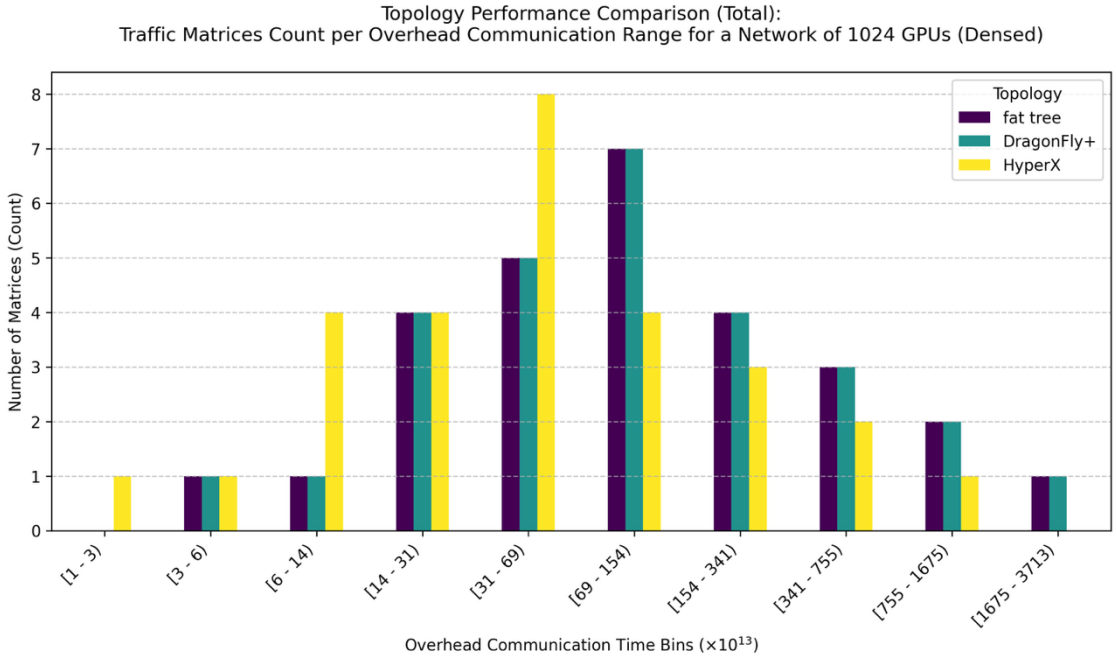




Dense Overhead Distribution

Over 1024 and 2048 World Sizes





Overhead Distribution Comparison

Between MoE and Dense Architectures