# Themis: A Network Bandwidth-Aware Collective Scheduling Policy for Distributed Training of DL Models

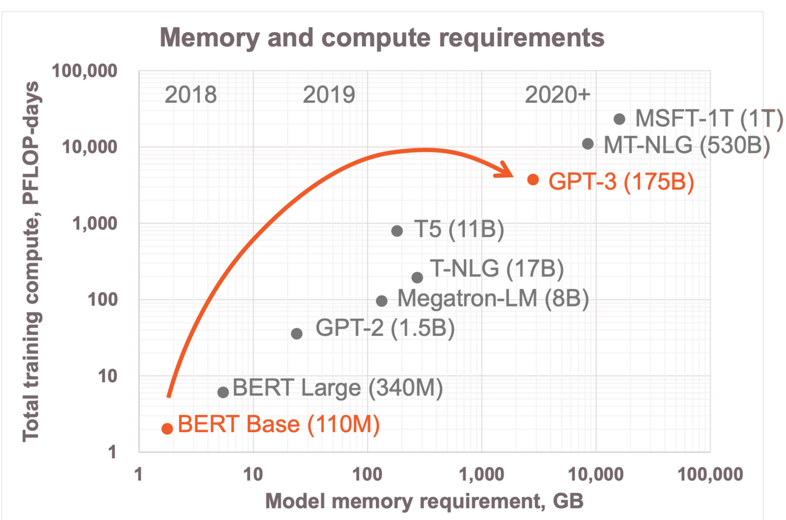
In *Proceedings of the 49th International Symposium on Computer Architecture (ISCA), Jun 2022*

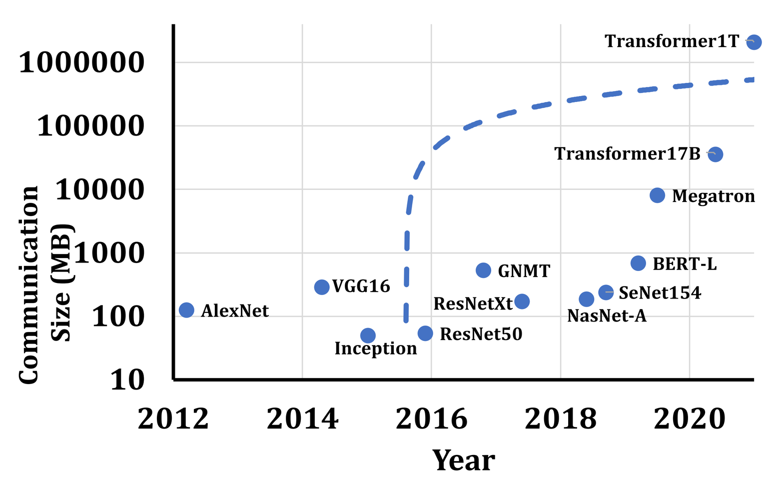
Saeed Rashidi[1], William Won[1], Sudarshan Srinivasan[2], Srinivas Sridharan[3], and Tushar Krishna[1]
[1]Georgia Institute of Technology, [2]Intel, [3]Meta

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

## Large AI Models

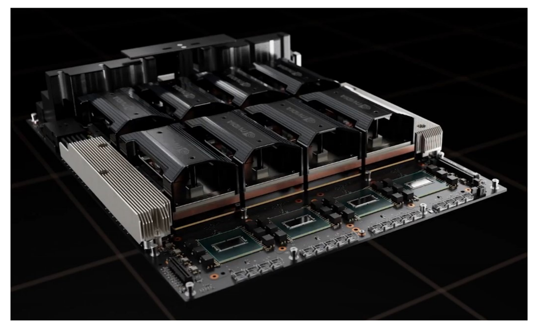
- Deep Learning: **Model size is increasing** (2× / 3.4 months)
  - GPT-3: 355 GPU years to train
- Necessitates **distributed training platforms**

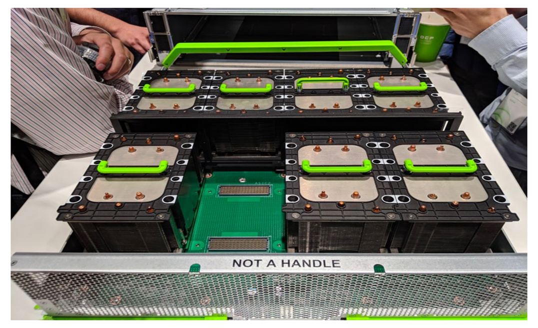



## DL Training Platforms

- Futuristic training networks: **multi-dimension + heterogeneous BW**
  - Ring, Switch (SW), FullyConnected (FC)
  - NVIDIA HGX-H100: SW_SW_SW
  - Facebook Zion: FC_SW
  - Google TPUv4: Ring_Ring_Ring

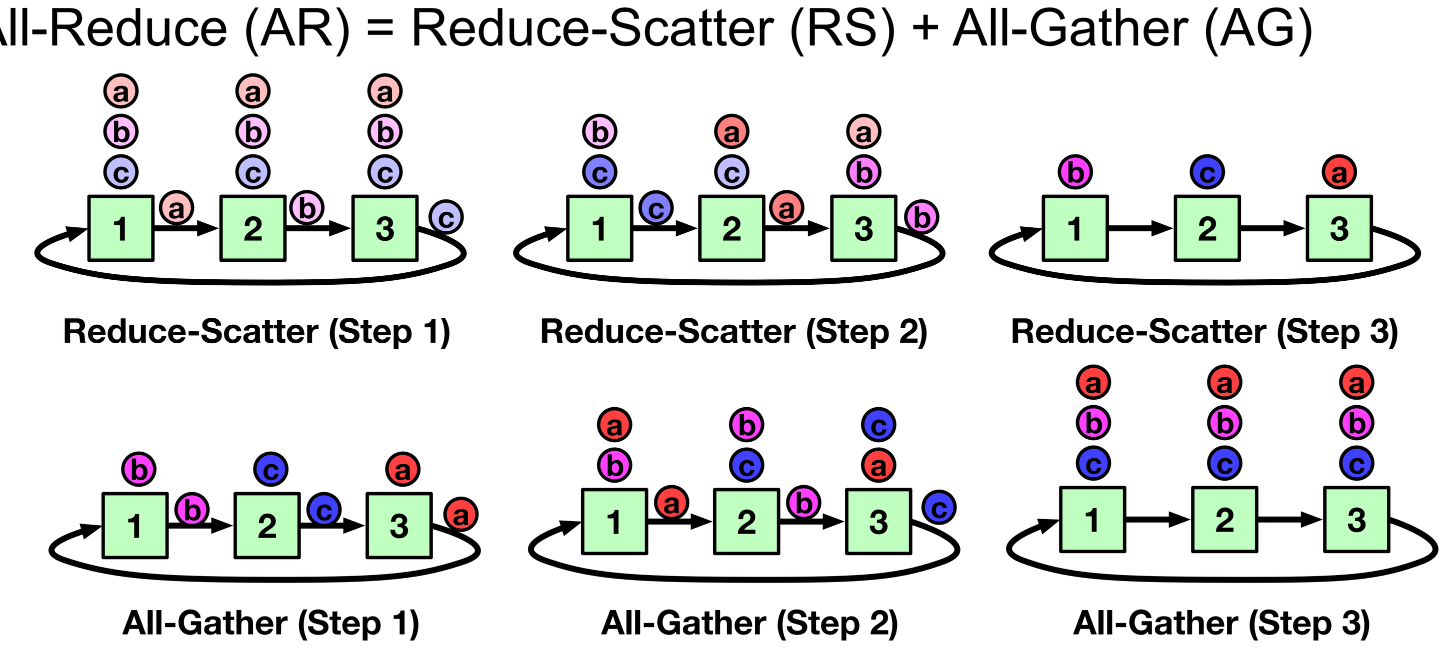

NVIDIA HGX-H100


Facebook Zion

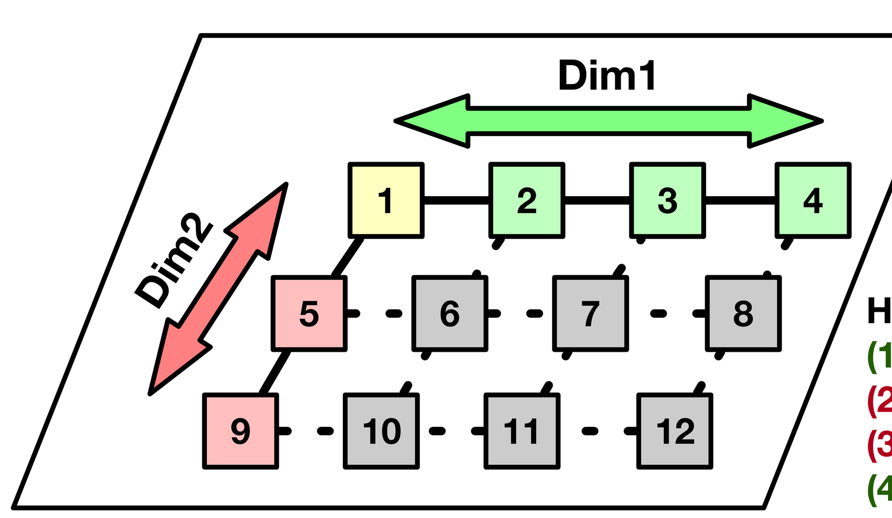

Google TPUv4

## All-Reduce Algorithm

- All-Reduce (AR) = Reduce-Scatter (RS) + All-Gather (AG)

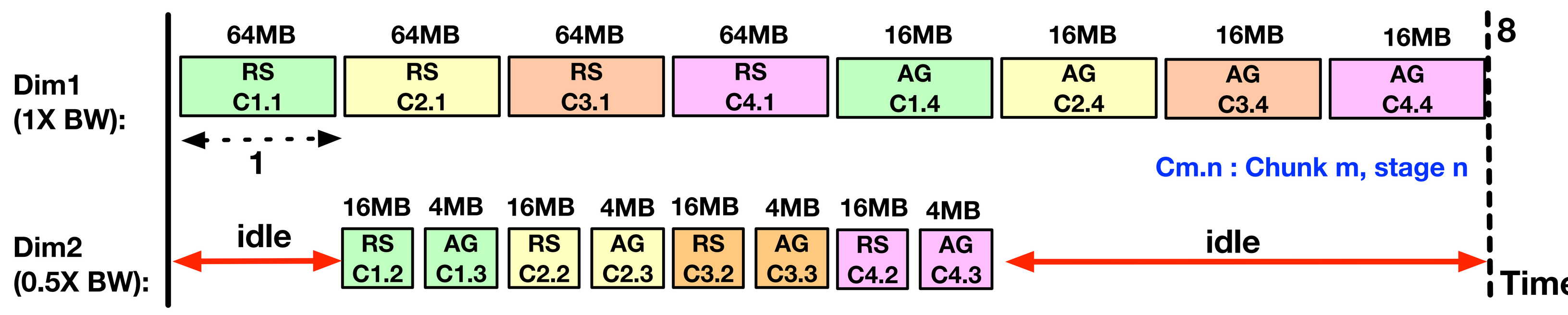



- **Hierarchical** All-Reduce: Traverses dimensions **in-order**

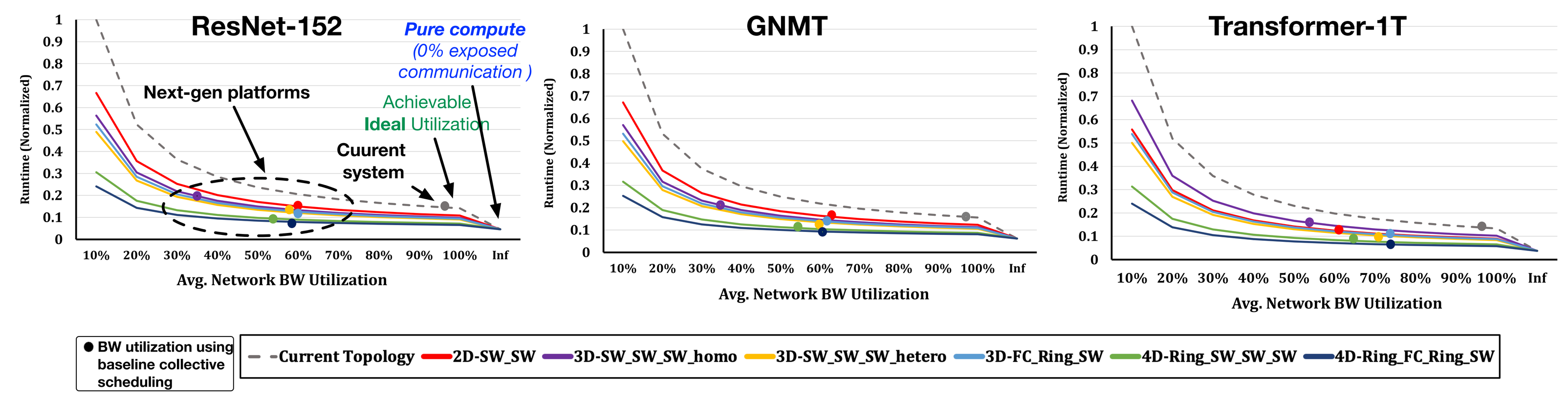



## Motivation

- **Message size decreases** as traversing dimensions (Hierarchical AR)
- Each network dimension has **different BW** (Networking Technology)
  - Network BW and chunk size **mismatch across dimensions**

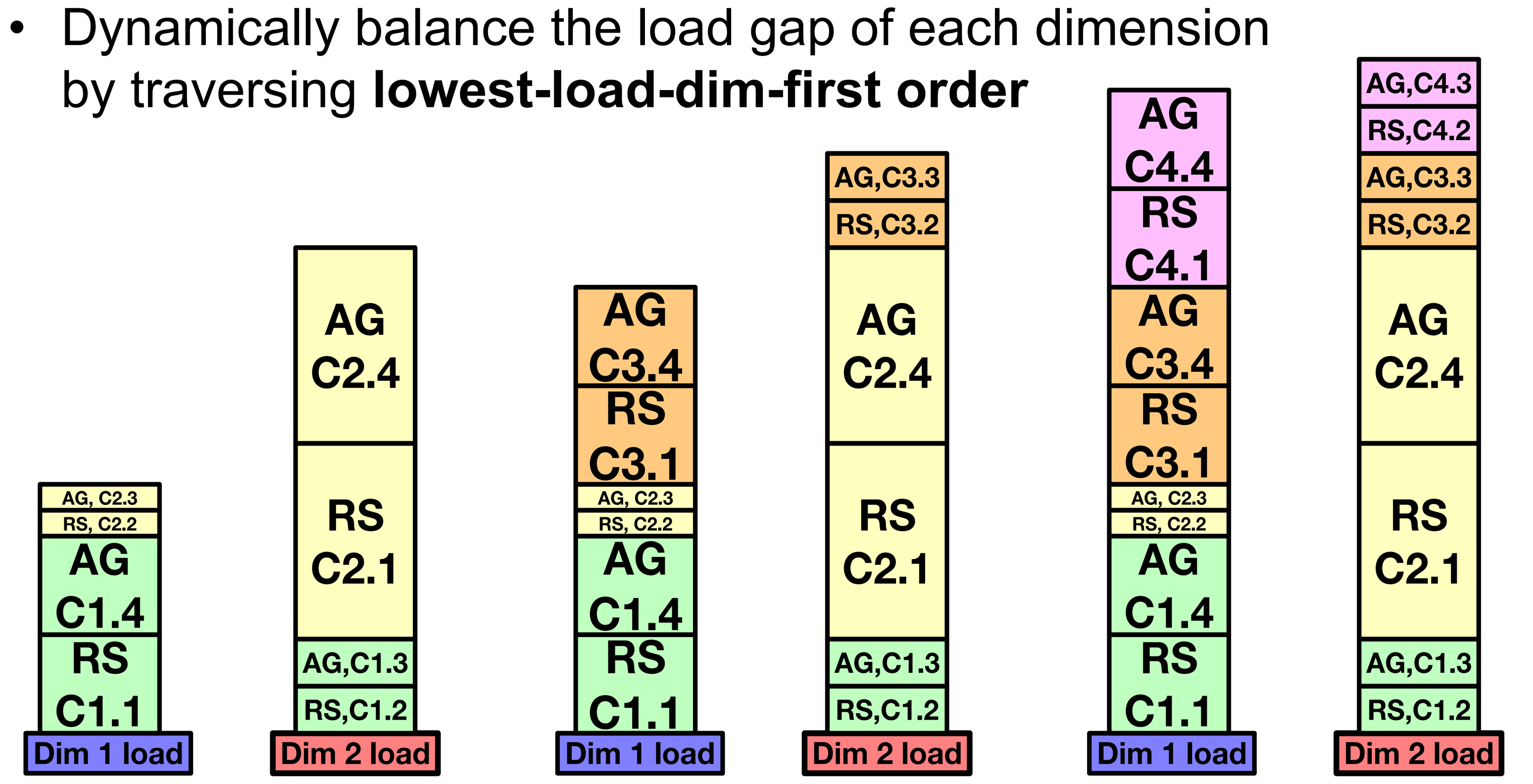



- **Underutilization** of Network BW Resource
  - Baseline: ~**59.7% network BW utilization** for next-gen topologies
  - **~1.37× (2.34× max) speedup** if 100% BW utilization is achieved

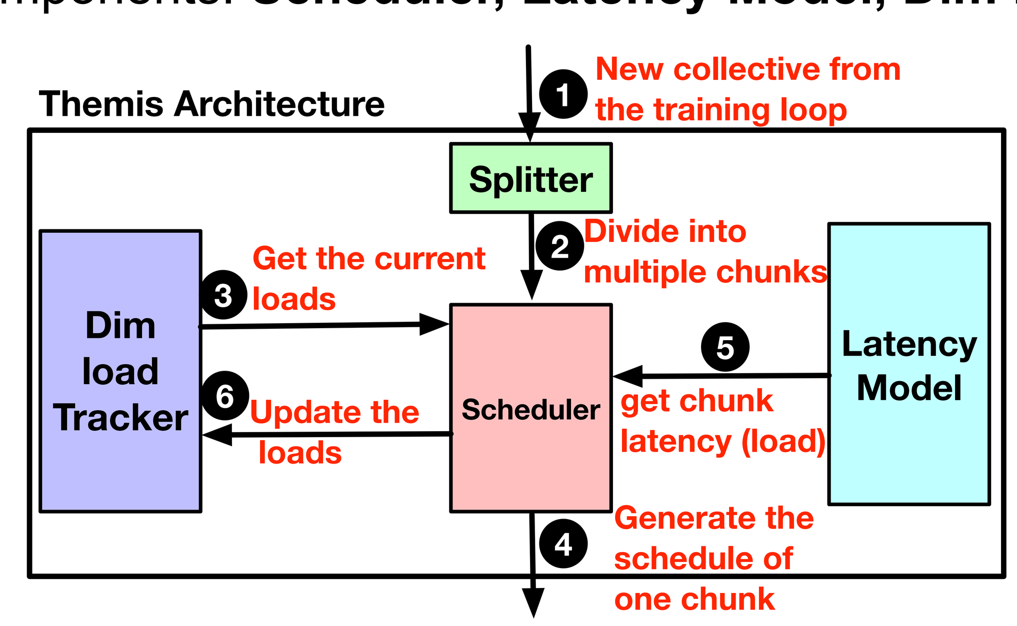



## Themis

- Observations:
  - Each chunk can traverse network dimensions **out-of-order**
  - Individual chunks can have **separate scheduling policy**
- **Themis**
  - **Track the current load** of each dimension
  - Dynamically balance the load gap of each dimension by traversing **lowest-load-dim-first order**




- Themis Components: **Scheduler, Latency Model, Dim Load Tracker**

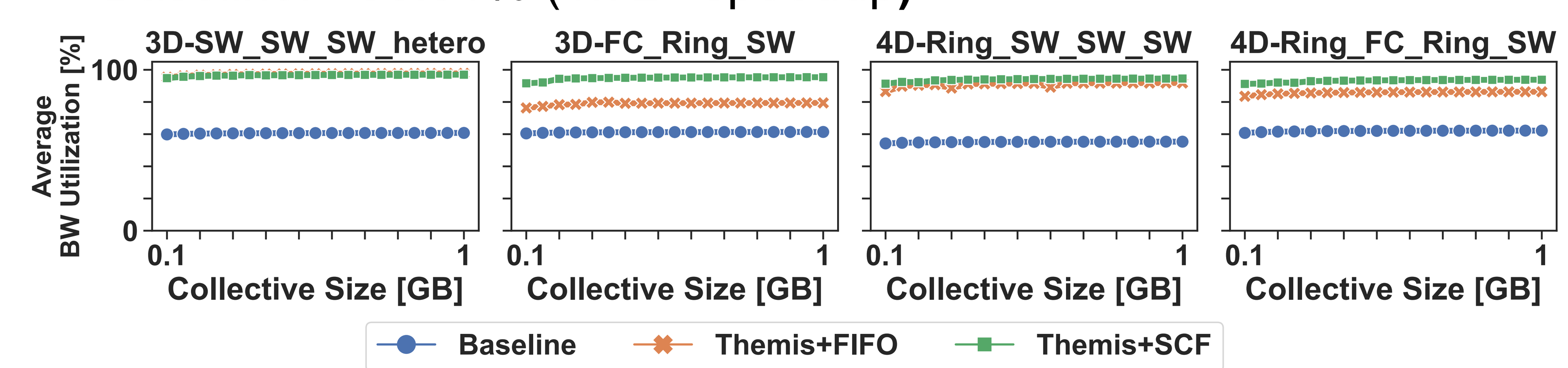



## Results

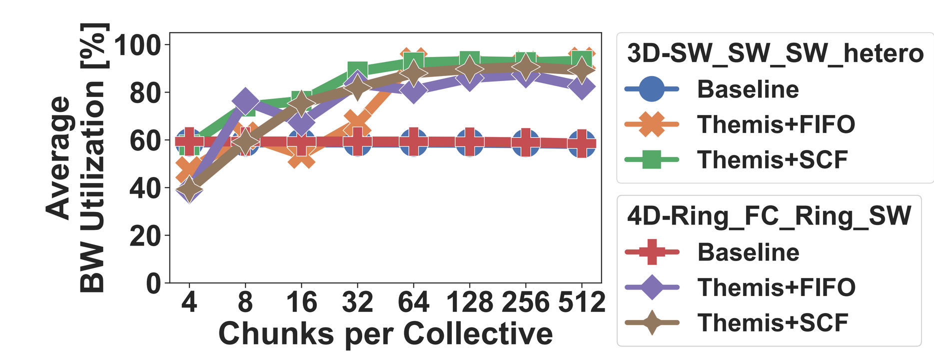
- Simulation Infrastructure: **ASTRA-sim**
  - https://astra-sim.github.io

| Name | Shape | BW (GB/s / Dim) | Latency (ns / Dim) |
|---|---|---|---|
| 3D-SW_SW_SW_hetero | 16 × 8 × 8 | 200, 100, 50 | 700, 1700 |
| 3D-FC_Ring_SW | 8 × 16 × 8 | 175, 100, 50 | 700, 700, 1700 |
| 4D-Ring_SW_SW_SW | 4 × 4 × 8 × 8 | 250, 200, 100, 50 | 20, 700, 700, 1700 |
| 4D-Ring_FC_Ring_SW | 4 × 8 × 4 × 8 | 375, 175, 150, 100 | 20, 700, 700, 1700 |

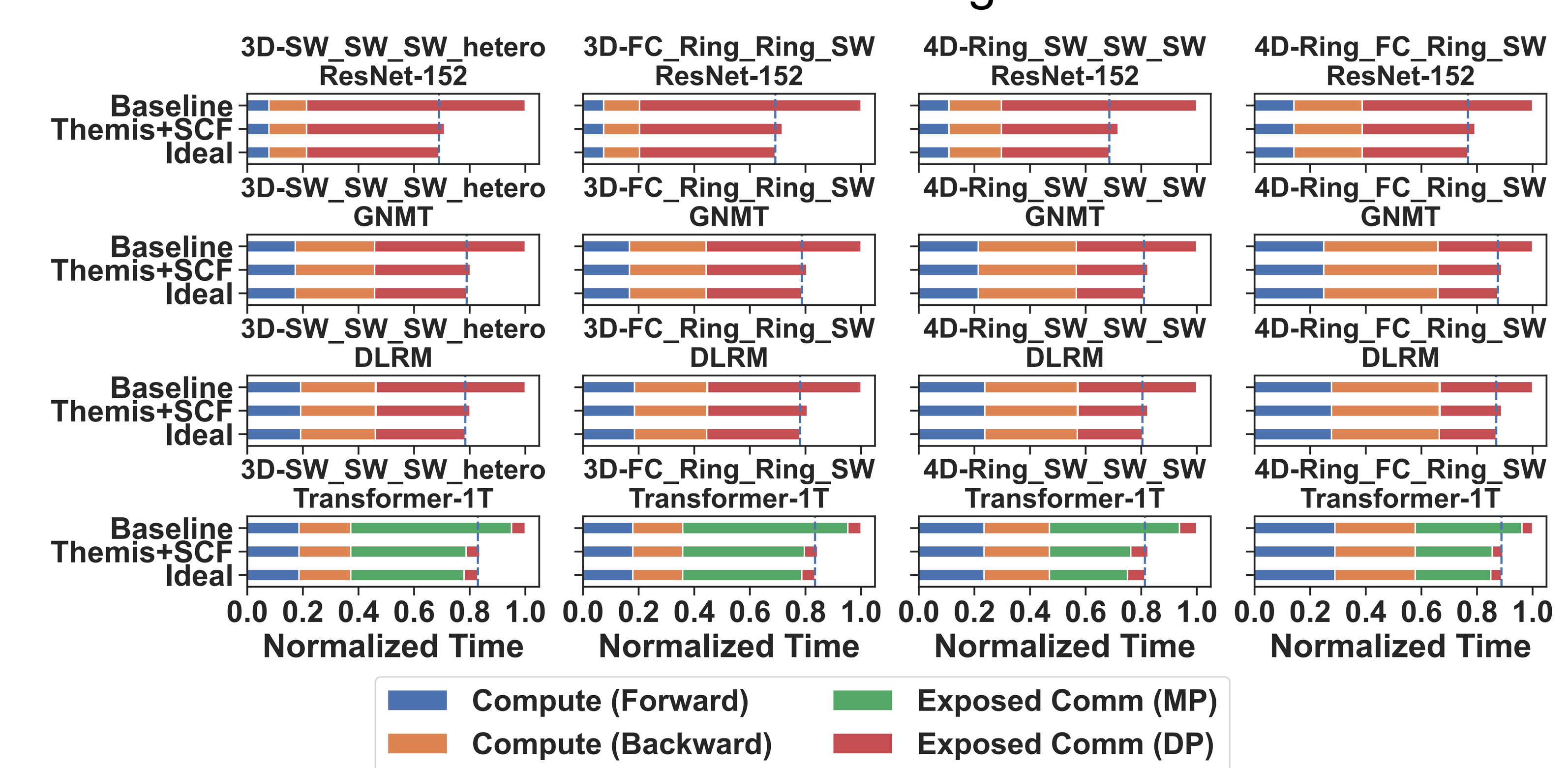
- Single All-Reduce: Themis achieves **~95.14% BW Utilization**
  - Baseline: ~56.31% (1.72× speedup)




- **More #chunks = better load balancing** capabilities




- Workloads: Themis reaches **near-ideal training performance**
  - ~1.49× (ResNet), ~1.30× (GNMT/DLRM), ~1.25× (T-1T) speedup over baseline hierarchical collective algorithm




## Conclusion

- Understanding futuristic training platforms
  - **Multi-dimensional** network with **heterogeneous BW**
- **Huge network BW underutilization** is observed
  - Due to chunk size and network BW mismatch across dimensions
- Themis: **Dynamic chunk scheduler** to improve BW utilization
  - By monitoring and **balancing loads** of each dimension
  - **95.14% network BW** utilization (Single All-Reduce)
  - **1.49×** (ResNet-152), **1.25×** (Transformer-1T) speedup