Georgia Tech School of Electrical and Computer Engineering
College of Engineering

SYNERGY

http://synergy.ece.gatech.edu

ASTRA-SIM

Demo 3: Advanced ASTRA-sim Topics

William Won
Ph.D. Student, School of Computer Science
Georgia Institute of Technology
william.won@gatech.edu

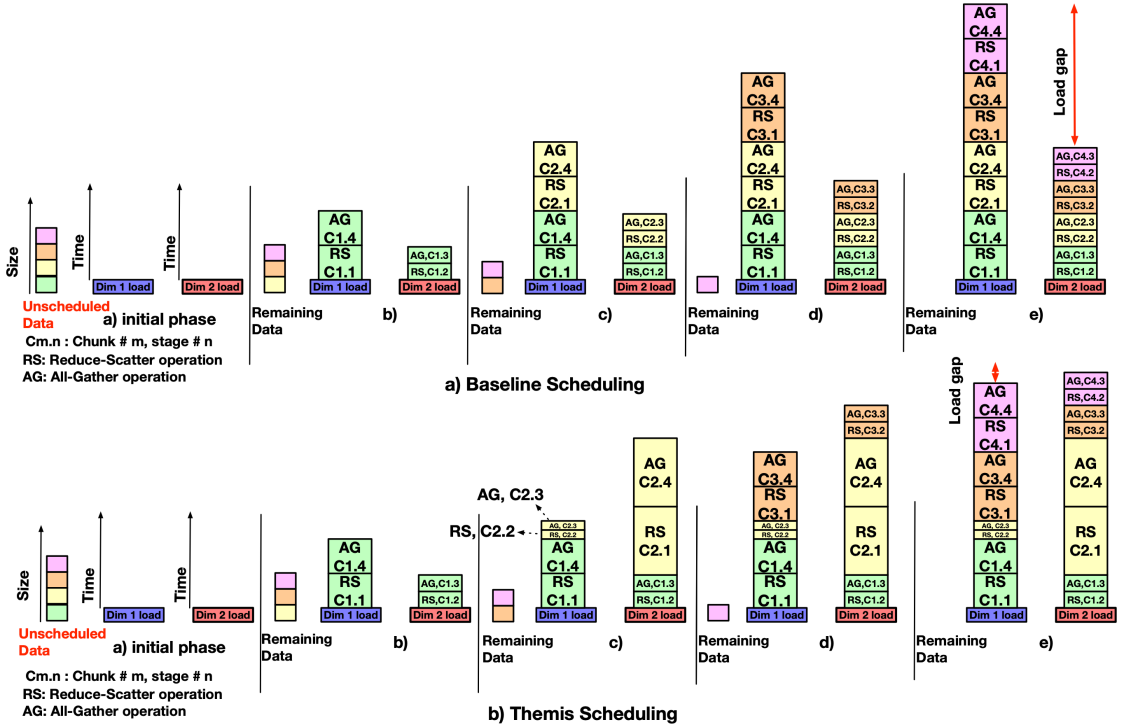
Acknowledgments: Srinivas Sridharan (Meta), Sudarshan Srinivasan (Intel)

Objective

- Enabling intra/inter-dimensional scheduling
  - Using Themis scheduler
- Compiling and using different network backends
  - Congestion-aware network backend
  - Running topology-unaware collectives

Themis Chunk Scheduler

- **Themis** uses Greedy-based intra/inter-dimension chunk scheduling to maximize **BW utilization** of multi-dimensional network
- **(ISCA '22) Themis: A Network Bandwidth-Aware Collective Scheduling Policy for Distributed Training of DL Models**



Exercise 3-1: Using Themis

Objective:

(1) We will 1 GB All-Reduce on a **3D switch with 512 NPUs**
(2) Using with and without **Themis scheduling** policy
(3) And compare the BW utilization

Instantiating 3D Switch

`inputs/network/3d_switch.json`

```
{
  "dimensions-count": 3,
  "topologies-per-dim": ["Switch", "Switch", "Switch"],
  "units-count": [8, 8, 8],
  "links-count": [1, 1, 1],
  "link-latency": [1, 1, 1],
  "link-bandwidth": [50, 50, 50]
}
```

3D network

512 NPUs

Total 150GB/s link bandwidth

Enabling Themis

`inputs/system/3d_switch_baseline.txt`

- **intra-dimension-scheduling**: **FIFO** ← **FIFO** scheduler
- **inter-dimension-scheduling**: **baseline** ← **baseline** scheduler

`inputs/system/3d_switch_themis.txt`

- **intra-dimension-scheduling**: **SCF** ← **Smallest-chunk-first** scheduler
- **inter-dimension-scheduling**: **themis** ← **Themis** scheduler

Running Experiment

- Objective: Running
  - 1 GB All-Reduce
  - On 512-NPU 3D Switch
  - Using with or without Themis scheduler

```
$ ./exercise_3-1.sh
```

Understanding Results

`result_3-1/tutorial_result.csv`

Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)	Total Message Size (MB)
baseline	**35008.129**	0	**35008.129**	**2044**
themis	**13635.871**	0	**13635.871**	**2044**

- Ideal collective time: 2044 MB / (150 GB/s) = 13307.292 µs
- Baseline BW utilization = 13307.292 / 35008.129 = **38.01%**
- Themis BW utilization = 13307.292 / 13635.871 = **97.59%**

Exercise 3-2: Running topology-unware collective

Objective:

(1) We will 1 GB All-Reduce on a **64-NPU Switch**
(2) Using **Direct All-Reduce** algorithm
(3) And compare the result

Compiling ASTRA-sim with Congestion Backend

Compile ASTRA-sim with congestion-aware analytical backend

```
$ ./build_congestion.sh
```

Configurations: Network

inputs/network/switch_analytical.json

```
{
  "dimensions-count": 1,
  "topologies-per-dim": ["Switch"],
  "units-count": [64],
  "links-count": [1],
  "link-latency": [500],
  "link-bandwidth": [50]
}
```

1D network

Switch topology

64 NPUs

1 links per NPU

500ns link latency

50GB/s link bandwidth

Configurations: Network

`inputs/network/switch_congestion.yml`

```
topology: Switch,
npus_count: 64,
bandwidth": [50],
latency": [500],
```

Switch topology

64 NPUs

50GB/s link bandwidth

500ns link latency

Configurations: System

inputs/system/direct.txt

```
scheduling-policy: LIFO
endpoint-delay: 10
active-chunks-per-dimension: 1
preferred-dataset-splits: 4
boost-mode: 0
all-reduce-implementation: direct
all-gather-implementation: direct
reduce-scatter-implementation: direct
all-to-all-implementation: direct
collective-optimization: localBWAware
```

direct All-Reduce Algorithm

Running Experiment

- Objective: Running
  - 1 GB All-Reduce
  - On 64-NPU Switch
  - Using Direct All-Reduce Algorithm

```
$ ./exercise_3-2.sh
```

Understanding Results

`result_3-2/tutorial_result.csv`

Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)
analytical	**633.217**	0	**633.217**
congestion	**40008.217**	0	**40008.217**