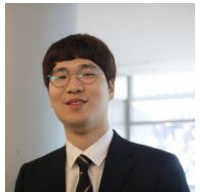
Georgia Tech School of Electrical and Computer Engineering
College of Engineering

http://synergy.ece.gatech.edu

Demo 1: Profiling Collectives Using ASTRA-sim



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

- Installing ASTRA-sim
  - Cloning repository
  - Compiling ASTRA-sim
- Demystifying Input Files
  - Network
  - System
  - Workload
- Running ASTRA-sim
  - Profiling single All-Reduce collective
  - Executing multiple simulations
  - Comparing different topologies
  - Comparing various-sized All-Reduce

Cloning ASTRA-sim

Prerequisite: Check installation dependencies

`https://astra-sim.github.io/tutorials/asplos-2023/installation`

(1) Clone ASTRA-sim tutorials GitHub repository

```
$ git clone https://github.com/astra-sim/tutorials.git
$ cd tutorials/asplos2023
```

(2) Run setup script

```
$ ./clone_astra_sim.sh
```

cf., Offers Docker Image

```
$ docker pull astrasim/tutorial-asplos2023
$ docker run -it astrasim/tutorial-asplos2023
```

Compiling ASTRA-sim with Analytical Backend

Compile ASTRA-sim with analytical backend

```
$ ./build_analytical.sh
```

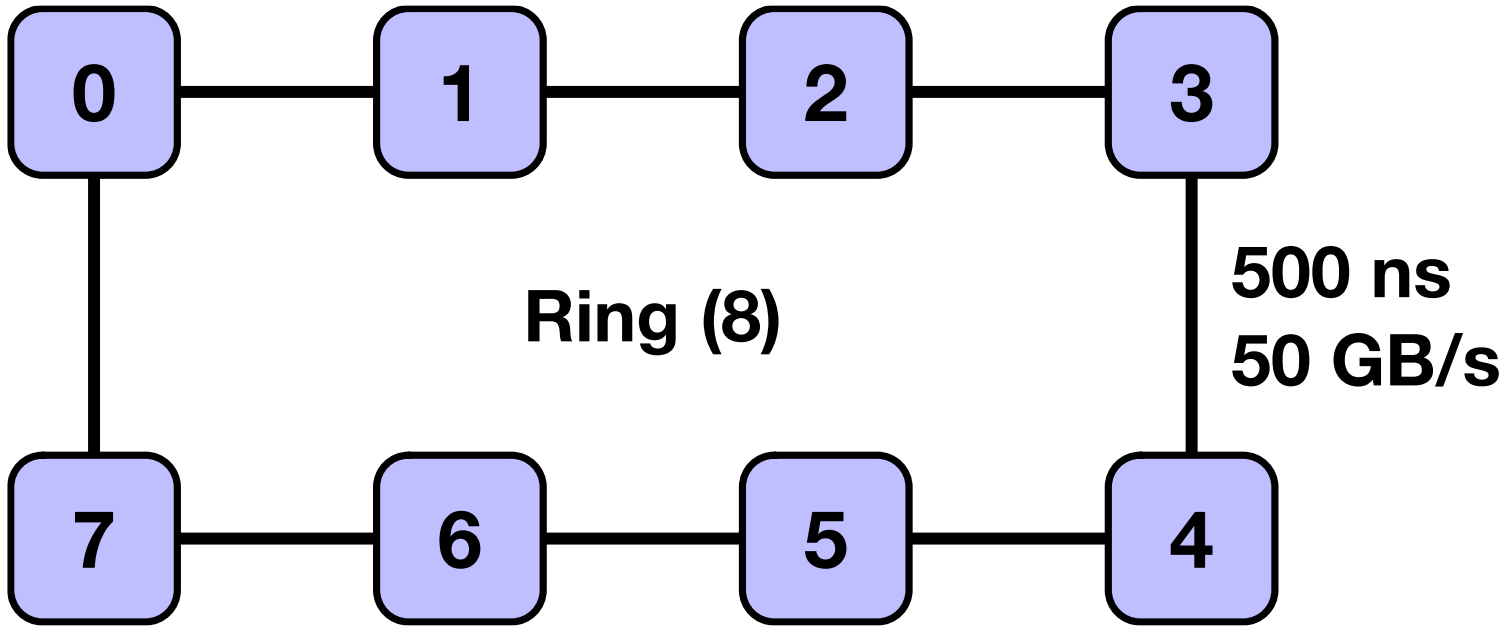
Exercise 1-1: Ring All-Reduce

Objective:

(1) We will configure an example 8-NPU Ring
(2) And run **1 MB All-Reduce** on it

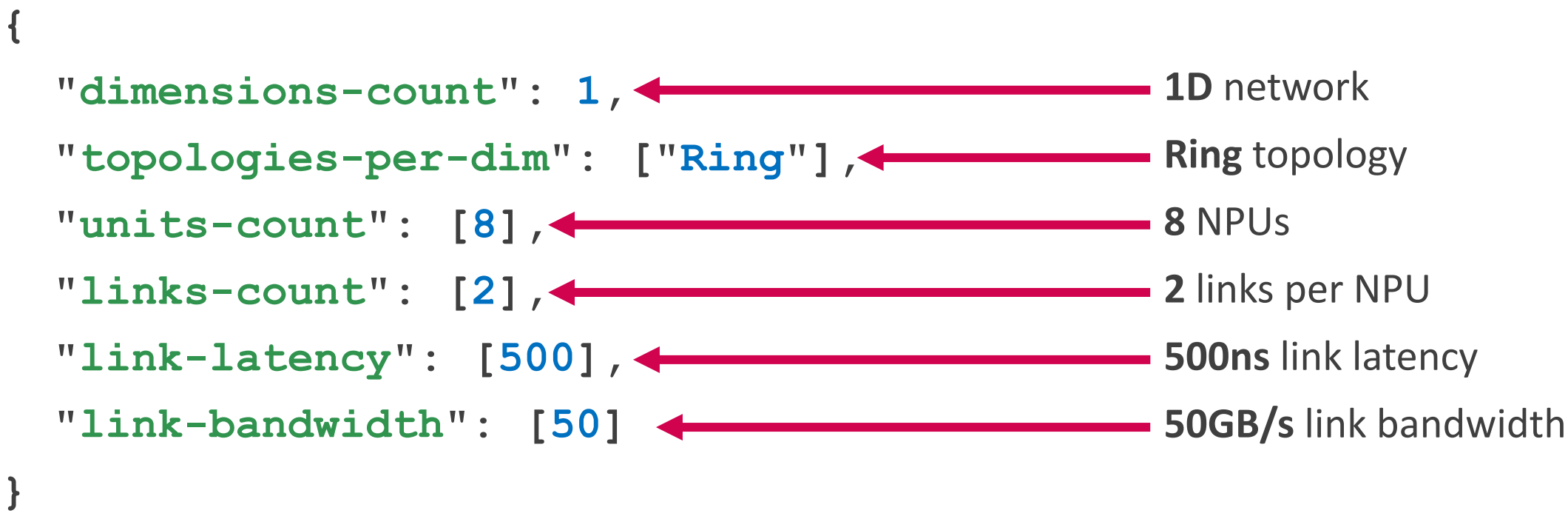
Configurations: Network

- **Ring** topology with **8 NPUs**
- **500 ns** (latency), **50 GB/s** (bandwidth)
- **2 links** per NPU



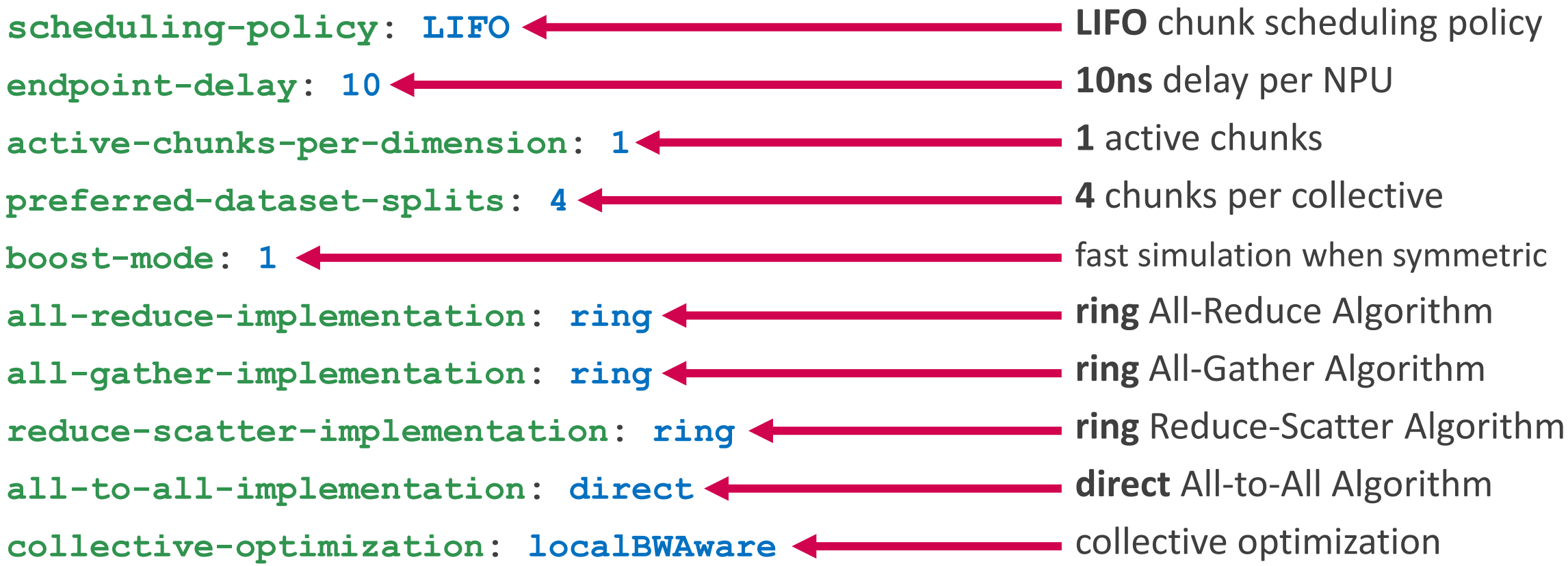
Configurations: Network

inputs/network/ring.json



Configurations: System

`inputs/system/ring.txt`



Configurations: System

`inputs/system/ring.txt`

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

4 chunks per collective

ring All-Reduce Algorithm

Configurations: Workload

`inputs/workload/all_reduce.txt`

```
MICRO
1
allreduce -1 1 NONE 0 1 NONE 0 1 ALLREDUCE 1048576 1
```

training loop

#layers

layer data

Metadata		Forward			Input grad			Weight grad			Layer
Layer Name	(rsvd.)	Compute Time	Comm. Type	Comm. size	Compute Time	Comm. Type	Comm. Size	Compute Time	Comm. Type	Comm. Size	Delay
allreduce	-1	1	NONE	0	1	NONE	0	1	ALLREDUCE	1048576	1

1 MB

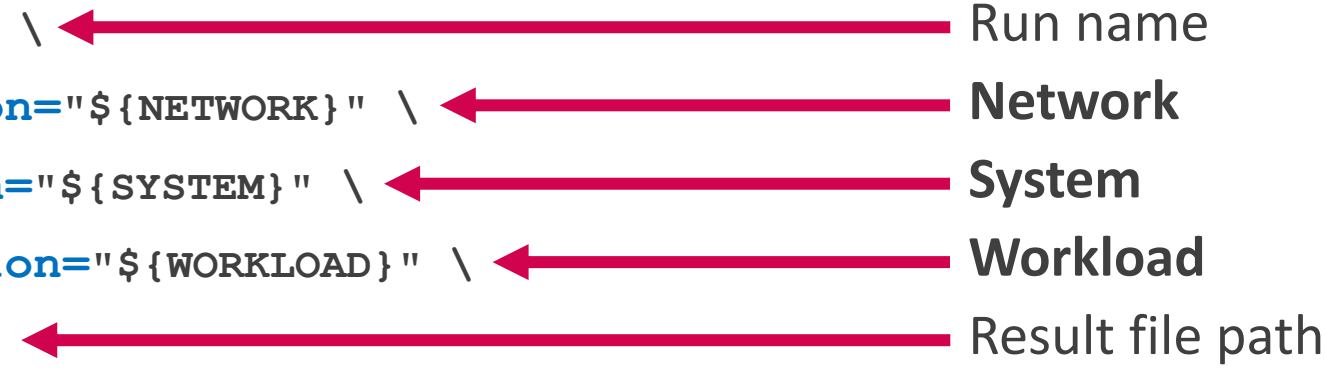
Running ASTRA-sim

Run ASTRA-sim

```
$ cd exercise_1
$ ./exercise_1-1.sh
```

exercise_1-1.sh

```
"${BINARY}" \
    --run-name="Exercise 1" \
    --network-configuration="${NETWORK}" \
    --system-configuration="${SYSTEM}" \
    --workload-configuration="${WORKLOAD}" \
    --path="${RESULT_DIR}/"
```



Running ASTRA-sim

45,681 ns (45.681 µs)

```
all passes finished at time: 45681, id of first layer: allreduce
path to create csvs is: /usr/scratch/will/tutorials/asplos2022/exercise_1/result/
success in openning file
*****
Time to exit: Sun Feb 27 06:46:51 2022
all-reduce Collective implementation: ring
reduce-scatter Collective implementation: ring
all-gather Collective implementation: ring
all-to-all Collective implementation: direct
Collective optimization: localBWAware
Total sim duration: 0:0 hours
Total streams injected: 4
Total streams finished: 4
Percentage of finished streams: 100 %
*****
Exiting
```

Understanding Results

`result_1-1/tutorial_result.csv`

Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)	Total Message Size (MB)
Exercise 1	45.681	0	45.681	1.75

45.681 µs

No compute

All communication exposed

1.75 MB/NPU

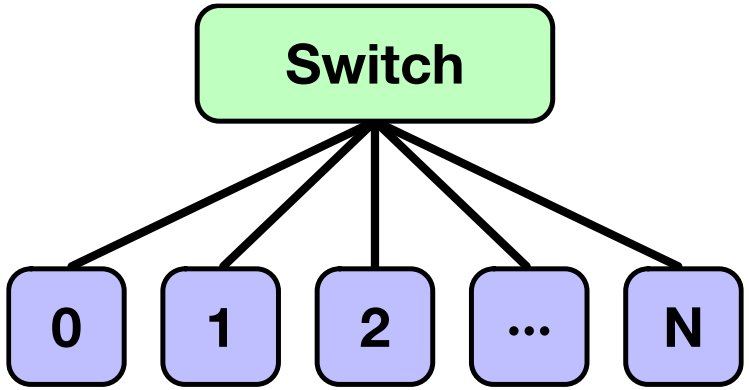
Exercise 1-2: Comparing Topologies

Objective:

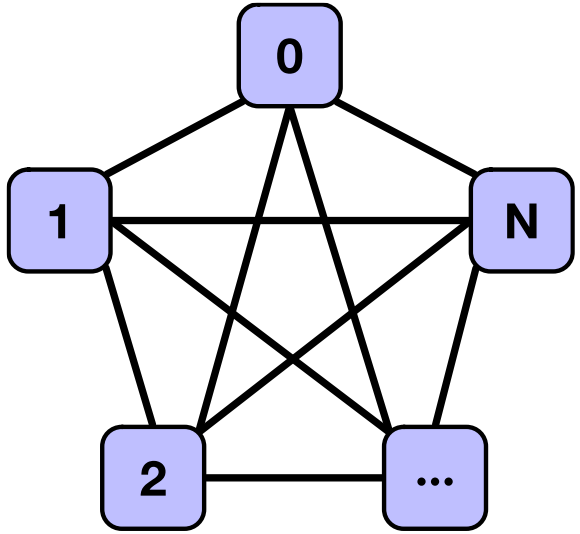
(1) We will configure two other topologies: **Switch** and **FullyConnected**

(2) And run the same **1 MB All-Reduce** on it

Switch and FullyConnected Topology



Switch(N)



FullyConnected(N)

- **Switch** topology
- **HalvingDoubling** All-Reduce
- **1** Link / NPU

- **FullyConnected** topology
- **Direct** All-Reduce
- **(N-1)** Links / NPU

Switch/FullyConnected Network

inputs/switch.json

```
{
  "dimensions-count": 1,
  "topologies-per-dim": ["Switch"],
  "units-count": [8],
  "links-count": [1],
  "link-latency": [500],
  "link-bandwidth": [50]
}
```

Switch topology

1 link/NPU

inputs/fullyconnected.json

```
{
  "dimensions-count": 1,
  "topologies-per-dim": ["FullyConnected"],
  "units-count": [8],
  "links-count": [7],
  "link-latency": [500],
  "link-bandwidth": [50]
}
```

FullyConnected topology

7 link/NPU

Configurations: System

`inputs/switch.txt`

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

HalvingDoubling
collective algorithm

`inputs/fullyconnected.txt`

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

Direct
collective algorithm

Executing Multiple Configurations

We want to run 3 simulations; yet still collect the result in the same csv file

```
"${BINARY}" \
        --total-stat-rows=3 \
        --stat-row=0 \
        --path="${RESULT_DIR}/"

"${BINARY}" \
        --total-stat-rows=3 \
        --stat-row=1 \
        --path="${RESULT_DIR}/"

"${BINARY}" \
        --total-stat-rows=3 \
        --stat-row=2 \
        --path="${RESULT_DIR}/"
```

3 total configurations

index 0

Same result file destination

index 1

Same result file destination

index 2

Same result file destination

Running Experiment

- Objective: Running
  - 1 MB All-Reduce
  - On 8-NPU Ring, Switch, FullyConnected

`exercise_1-2.sh`

```
"${BINARY}" \
        --run-name="Switch" \
        --network-configuration="${INPUT_DIR}/switch.json" \
        --system-configuration="${INPUT_DIR}/switch.txt" \
        --workload-configuration="${WORKLOAD}" \
        --path="${RESULT_DIR}/" \
        --total-stat-rows=3 \
        --stat-row=1
```

Switch topology

Switch system

3 Total configs

Running Experiment

- Objective: Running
  - 1 MB All-Reduce
  - On 8-NPU Ring, Switch, FullyConnected

```
$ ./exercise_1-2.sh
```

Understanding Results

`result_1-2/tutorial_result.csv`

Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)	Total Message Size (MB)
Ring	**45.681**	0	**45.681**	1.75
Switch	**58.449**	0	**58.449**	1.75
FullyConnected	**9.001**	0	**9.001**	1.75

Exercise 1-3: Comparing Various-sized All-Reduce

Objective:

(1) On the 8-NPU Ring
(2) We will run **1 MB – 1 GB All-Reduce**
(3) And observe the **trend**

Changing Communication Size

- Running **5 MB** All-Reduce collective:
- Use ASTRA-sim's **comm-scale** option

```
"${BINARY}" \
    --run-name="Exercise 1-3" \
    --network-configuration="${NETWORK}" \
    --system-configuration="${SYSTEM}" \
    --workload-configuration="${WORKLOAD}" \
    --comm-scale="5" \
    --path="${RESULT_DIR}/"
```

Run ASTRA-sim with **5x communication size**

Executing Multiple Configurations

- Objective: All-Reduce of size [**1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024**] MB (total 11 configurations)

```
SIZES=(1 2 4 8 16 32 64 128 256 512 1024)
for i in {0..10}; do
    size=${SIZES[$i]}
    "${BINARY}" \
        --run-name="${size}" \
        --network-configuration="${NETWORK}" \
        --system-configuration="${SYSTEM}" \
        --workload-configuration="${WORKLOAD}" \
        --comm-scale="${size}" \
        --path="${RESULT_DIR}/" \
        --total-stat-rows=11 \
        --stat-row=$i
done
```

Size: **1 - 1024** MB

For-loop

Run name: Size

All-Reduce Size

11 Total configs

ith config

Running Experiment

- All-Reduce of size [1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024] MB (total 11 configurations)

```
$ ./exercise_1-3.sh
```

Understanding Results

`result_1-3/tutorial_result.csv`

Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)	Total Message Size (MB)
1	45.681	0	45.681	1.75
2	62.761	0	62.761	3.5
4	96.921	0	96.921	7
8	165.297	0	165.297	14
16	302.077	0	302.077	28
32	575.609	0	575.609	56
64	1122.673	0	1122.673	112
128	2216.745	0	2216.745	224
256	4404.945	0	4404.945	448
512	8781.373	0	8781.373	896
1024	17534.229	0	17534.229	1792

```
$ ./python3 plot_1-3.py
```