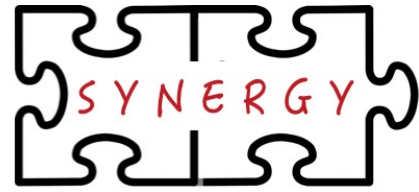




Georgia Tech School of Electrical and  
Computer Engineering  
College of Engineering



<http://synergy.ece.gatech.edu>



# Exercise 2: Comparing Collectives



**Will Won**

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

**Acknowledgments:** Srinivas Sridharan (Facebook), Sudarshan Srinivasan (Intel)

# Agenda

Time (CET)	Time (ET)	Topic	Presenter
15:00 – 16:00	9:00 – 10:00	<b>Introduction to Distributed Deep Learning Training Platforms</b>	Tushar Krishna
16:00 – 17:00	10:00 – 11:00	<b>ASTRA-sim</b>	Saeed Rashidi
17:00 – 17:10	11:00 – 11:10	<b>Break</b>	
<b>17:10 – 17:50</b>	<b>11:10 – 11:50</b>	<b>Demo and Exercises</b>	<b>William Won and Taekyung Heo</b>
17:50 – 18:00	11:50 – 12:00	<b>Extensions and Future Development</b>	Tushar Krishna and Saeed Rashidi

## Tutorial Website

*includes agenda, slides, ASTRA-sim installation instructions (via source + docker image)*

<https://astra-sim.github.io/tutorials/asplos-2022>

**Attention:** Tutorial is being recorded

# Objective

---

- Familiarizing yourself more with ASTRA-sim scripts
  - Changing communication size
  - Executing multiple runs
- Comparing ASTRA-sim results
  - Different-sized All-Reduce collective
- Implementing different topologies
  - Running HalvingDoubling All-Reduce on Switch
  - Running Direct All-Reduce on FullyConnected

# Changing Communication Size

- Running **5 MB** All-Reduce collective

## Method 1: Change Workload Configuration

**MICRO** ← training loop

**1** ← #layers

`allreduce -1 1 NONE 0 1 NONE 0 1 ALLREDUCE 5242880 1`

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	5242880	1

**5 MB**

# Changing Communication Size

- Running **5 MB** All-Reduce collective

## Method 2: Change ASTRA-sim **Run Script**

```
"${BINARY}" \  
  --run-name="Exercise 2" \  
  --network-configuration="${NETWORK}" \  
  --system-configuration="${SYSTEM}" \  
  --workload-configuration="${WORKLOAD}" \  
  --comm-scale="5" \  
  --path="${RESULT_DIR}/"
```

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

# Executing Multiple Configurations

Run [1, 5, 10] MB All-Reduce (**total 3 configurations**) concurrently

```
"${BINARY}" \  
  --comm-scale="1" \  
  --total-stat-rows=3 \  
  --stat-row=0
```

← **1MB All-Reduce**  
← **3 total configurations**  
← **index 0**

```
"${BINARY}" \  
  --comm-scale="5" \  
  --total-stat-rows=3 \  
  --stat-row=1
```

← **5MB All-Reduce**  
← **index 1**

```
"${BINARY}" \  
  --comm-scale="10" \  
  --total-stat-rows=3 \  
  --stat-row=2
```

← **10MB All-Reduce**  
← **index 2**

# 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) ← Size: 1 - 1024 MB
for i in {0..10}; do ← For-loop
  size=${SIZES[$i]}
  "${BINARY}" \
    --run-name="${size}" \ ← Run name: Size
    --network-configuration="${NETWORK}" \
    --system-configuration="${SYSTEM}" \
    --workload-configuration="${WORKLOAD}" \
    --comm-scale="${size}" \ ← All-Reduce Size
    --path="${RESULT_DIR}/" \
    --total-stat-rows=11 \ ← 11 Total configs
    --stat-row=$i ← ith config
done

```

# Running Experiment

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

```
$ cd exercise_2/
```

```
$ ./build.sh
```

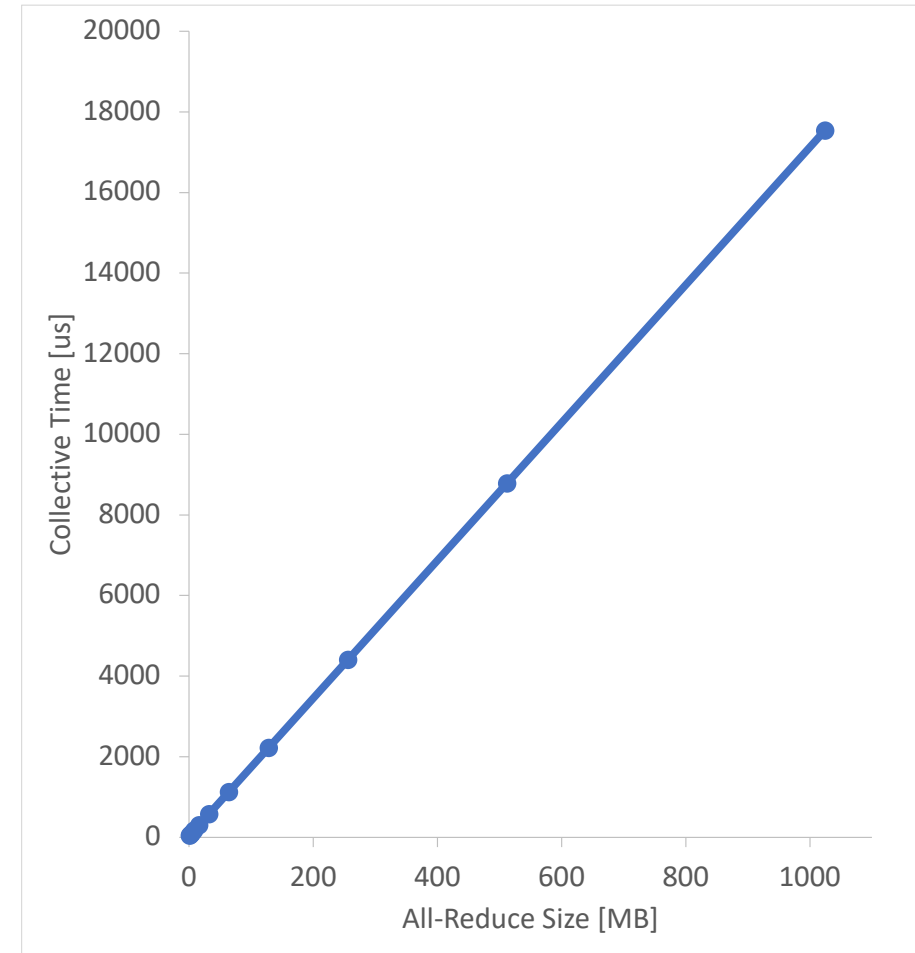
```
$ ./exercise_2_1.sh
```



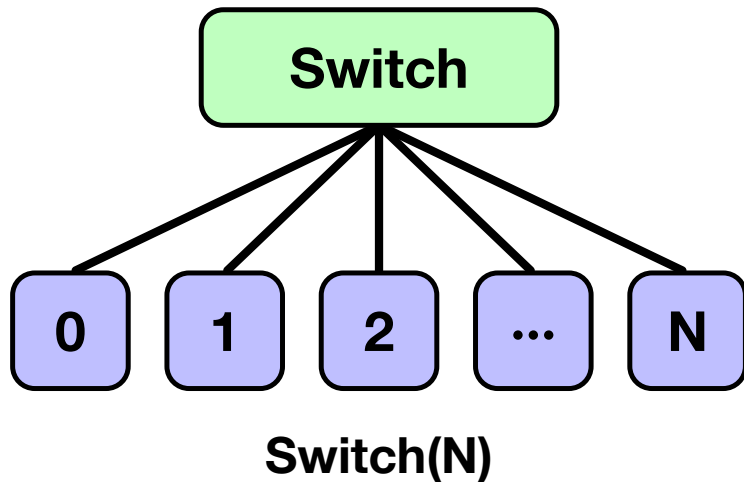
# Understanding Results

result\_1/tutorial\_result.csv

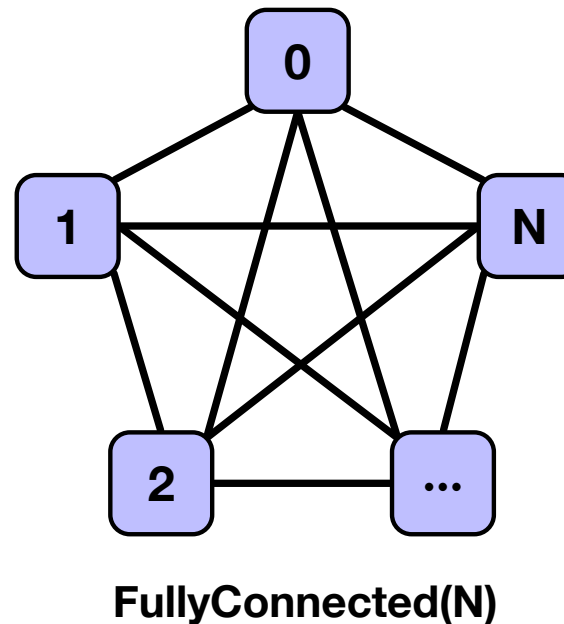
Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)	Total Message Size (MB)
1	45.681	0	45.681	1.75
2	62.761	0	62.761	3.5
4	96.921	0	96.921	7
8	165.297	0	165.297	14
16	302.077	0	302.077	28
32	575.609	0	575.609	56
64	1122.673	0	1122.673	112
128	2216.745	0	2216.745	224
256	4404.945	0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



# Switch and FullyConnected Topology



- **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



# Running Experiment

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

exercise\_2\_2.txt

```
"${BINARY}" \  
  --run-name="Switch" \  
  --network-configuration="${INPUT_DIR}/switch.json" \  
  --system-configuration="${INPUT_DIR}/switch.txt" \  
  --workload-configuration="${WORKLOAD}" \  
  --comm-scale="1024" \  
  --path="${RESULT_DIR}/" \  
  --total-stat-rows=3 \  
  --stat-row=1
```

← Switch topology

← Switch system

← 1GB All-Reduce

← 3 Total configs

# Running Experiment

---

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

```
$ ./build.sh
```

```
$ ./exercise_2_2.sh
```

# Understanding Results

result\_2/tutorial\_result.csv

Name	Total Time (us)	Compute Time (us)	Exposed Communication Time (us)	Total Message Size (MB)
Ring	17534.229	0	17534.229	1792
Switch	35026.693	0	35026.693	1792
FullyConnected	5004.925	0	5004.925	1792