System for Artificial Intelligence

Collective Communications

Siyuan Feng
Shanghai Innovation Institute

OUTLINE

01 Communication Patterns

02 Collective Communication

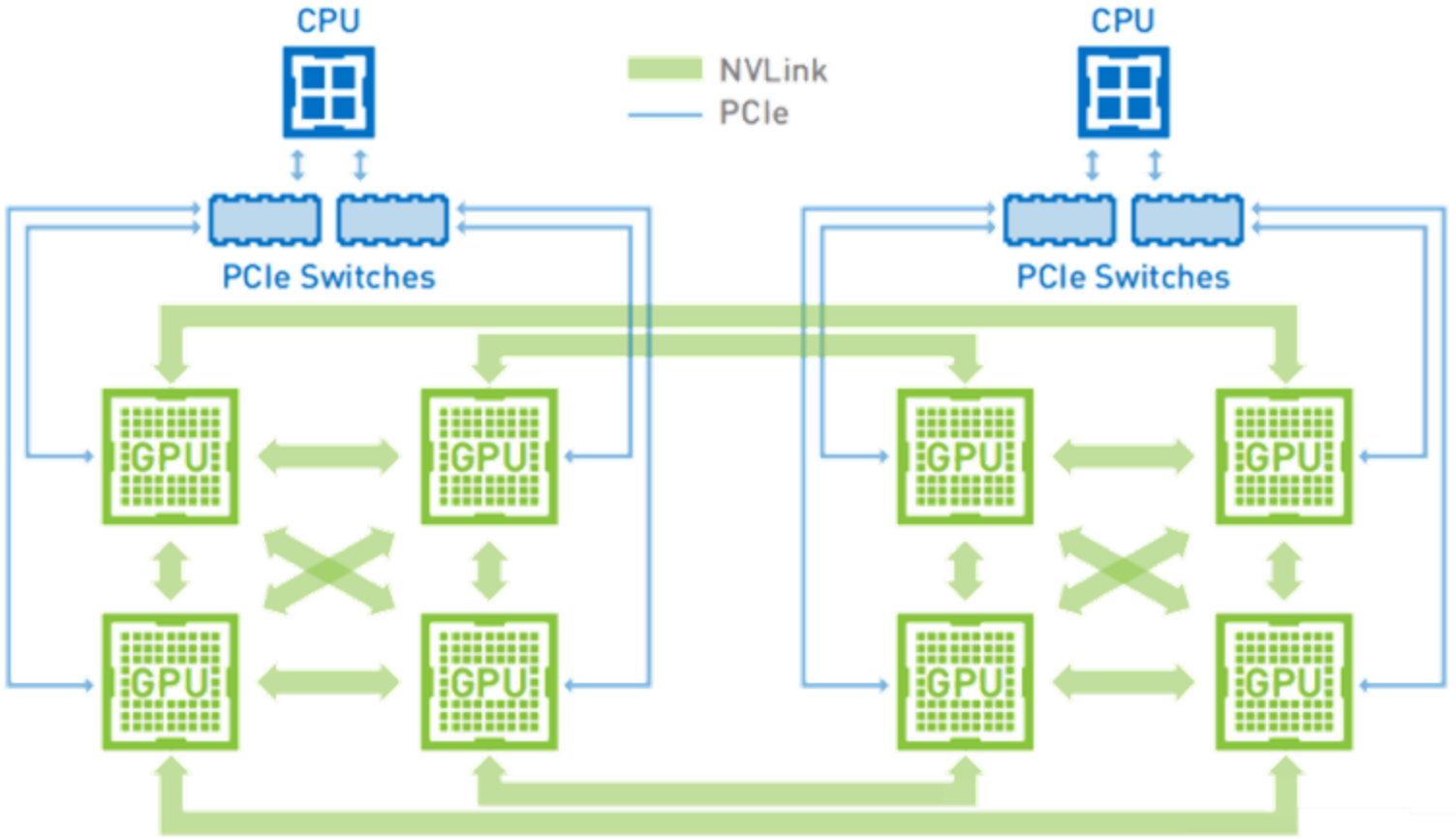
03 Ring-Based Collectives

01

Communication Patterns

Point-to-Point Communication

上海创智学院
Shanghai Innovation Institute

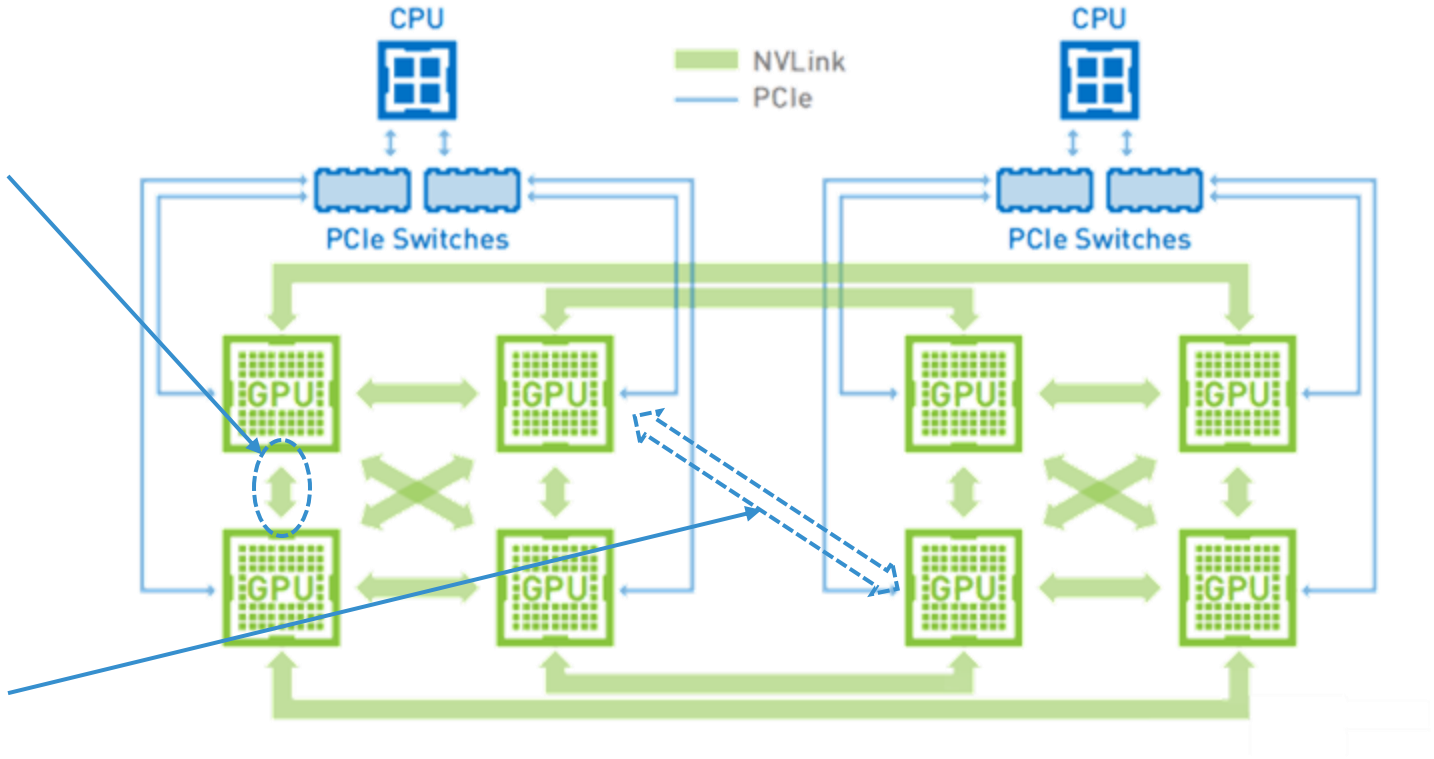


Challenge of P2P Communication

上海创智学院
Shanghai Innovation Institute

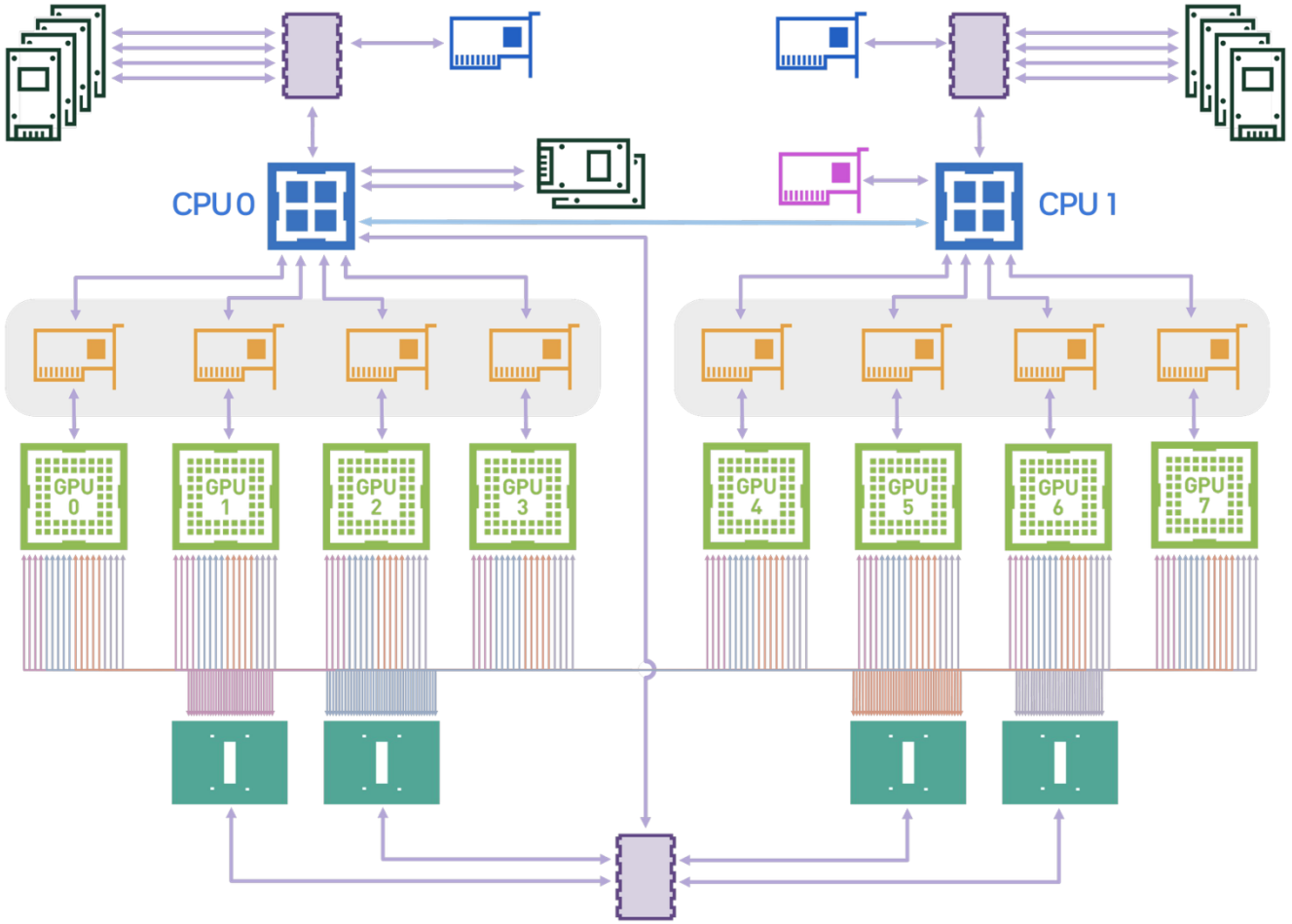
Cannot utilize full bandwidth for P2P communication

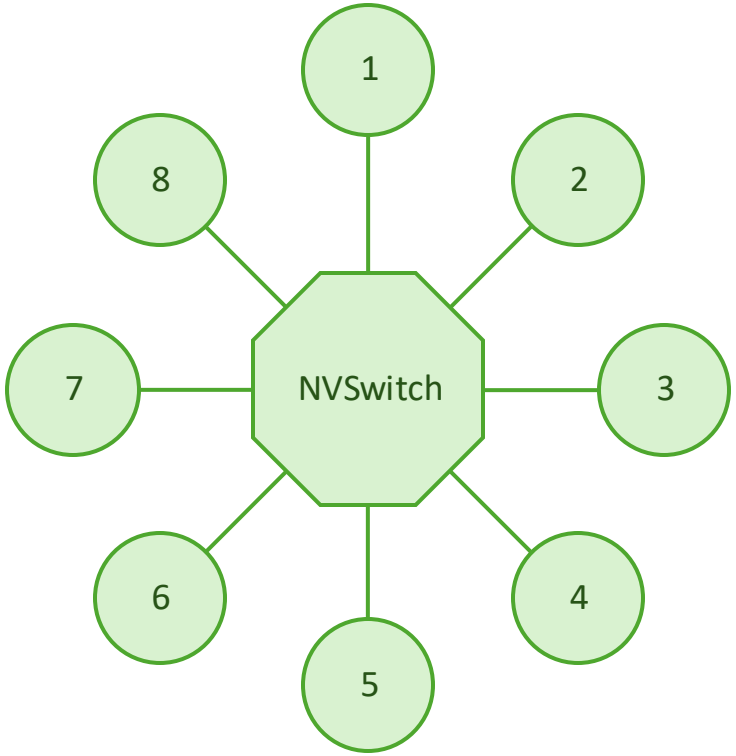
Cannot communicate between any pair of two GPUs



Hardware Solution: NVSwitch

上海创智学院
Shanghai Innovation Institute



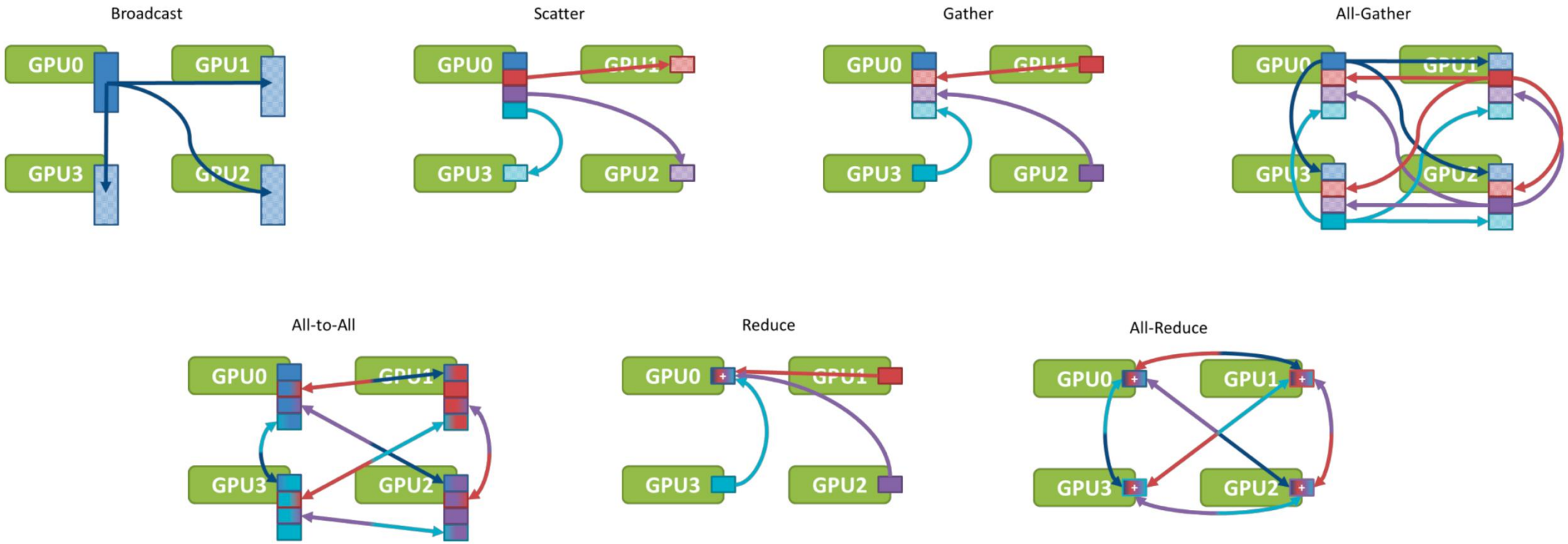


ConnectX-7 ConnectX-7 Network Module NVMe PCIe Switches NVSwitch PCIe 100 GbE CPU communication

Software Solution: Collective Communication

上海创智学院
Shanghai Innovation Institute

Communication always happen synchronized among multiple devices.



Communication among Tasks

上海创智学院
Shanghai Innovation Institute

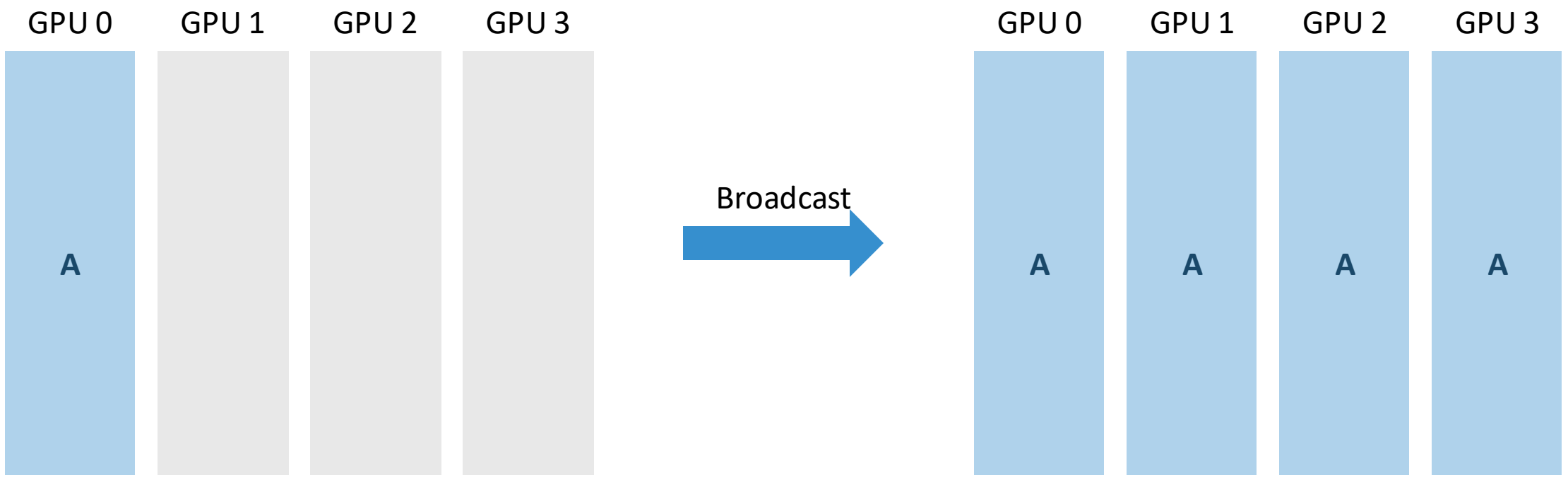
- Point-to-point communication
  - Single sender and single receiver
  - Relatively easy to implement efficiently
- Collective communication
  - Multiple senders and/or receivers
  - Patterns include broadcast, scatter, gather, reduce, all-to-all, …
  - Difficult to implement efficiently

02

Collective Communication

Broadcast

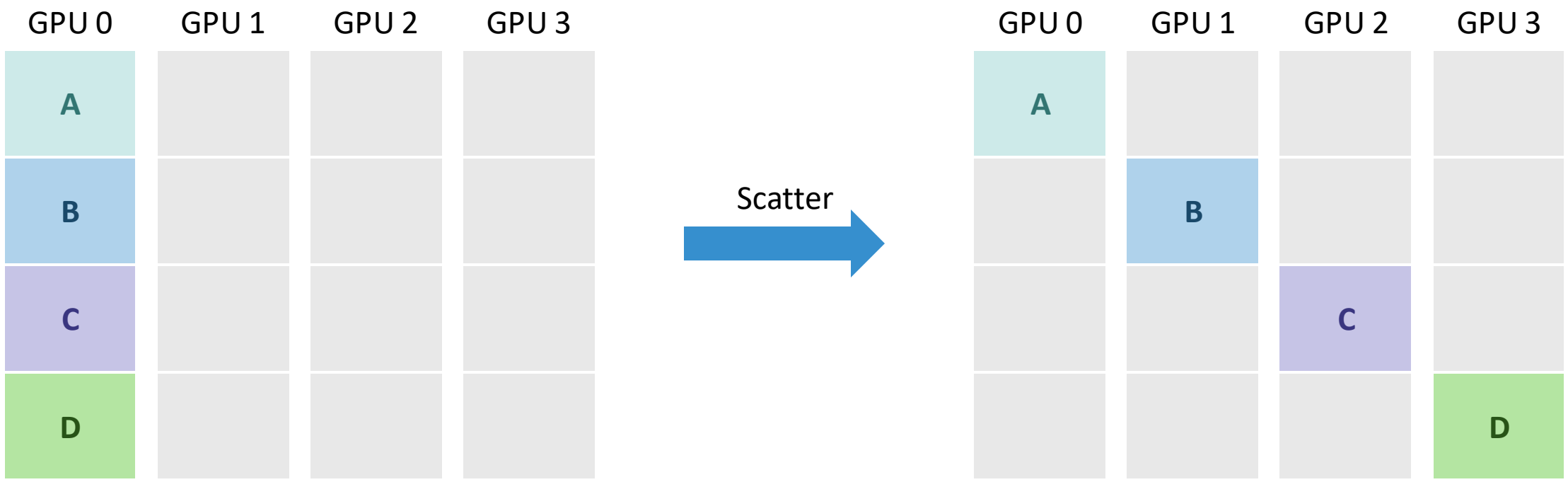
One sender, multiple receivers



Scatter

上海创智学院
Shanghai Innovation Institute

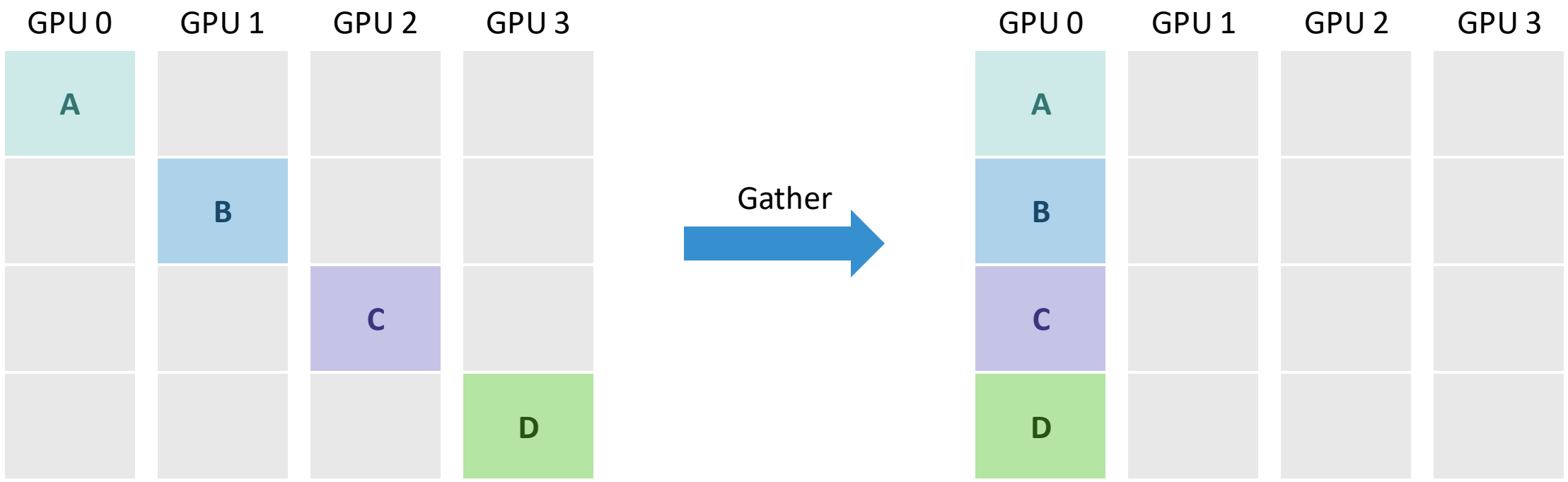
One sender; data is distributed among multiple receivers



Gather

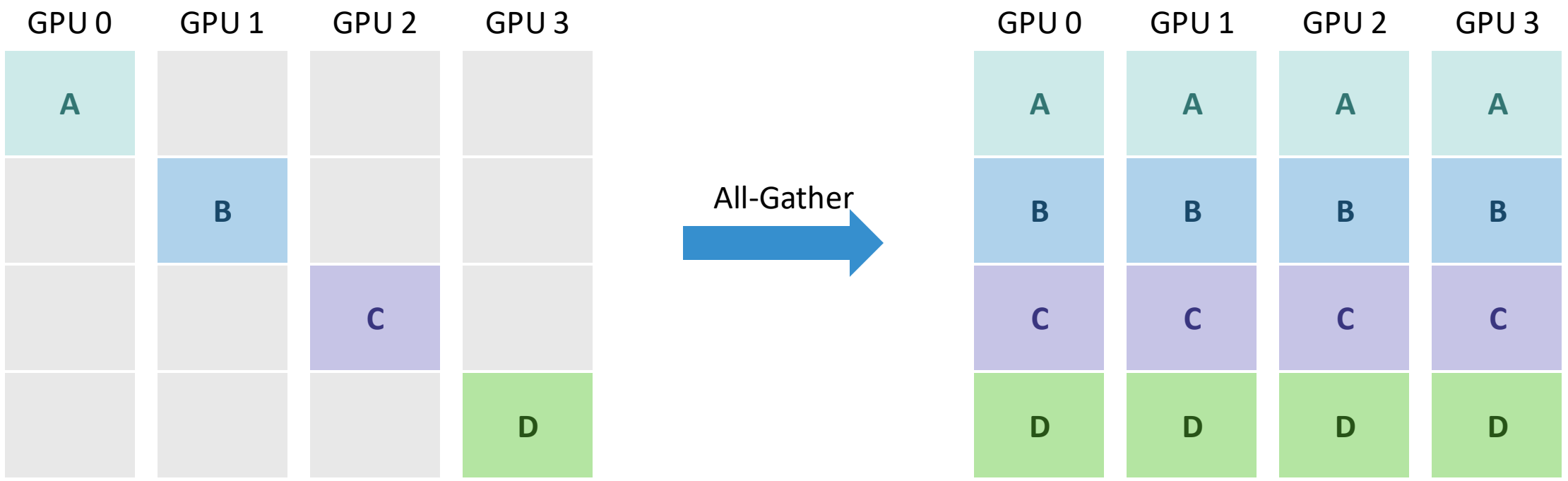
上海创智学院
Shanghai Innovation Institute

Multiple senders, one receiver



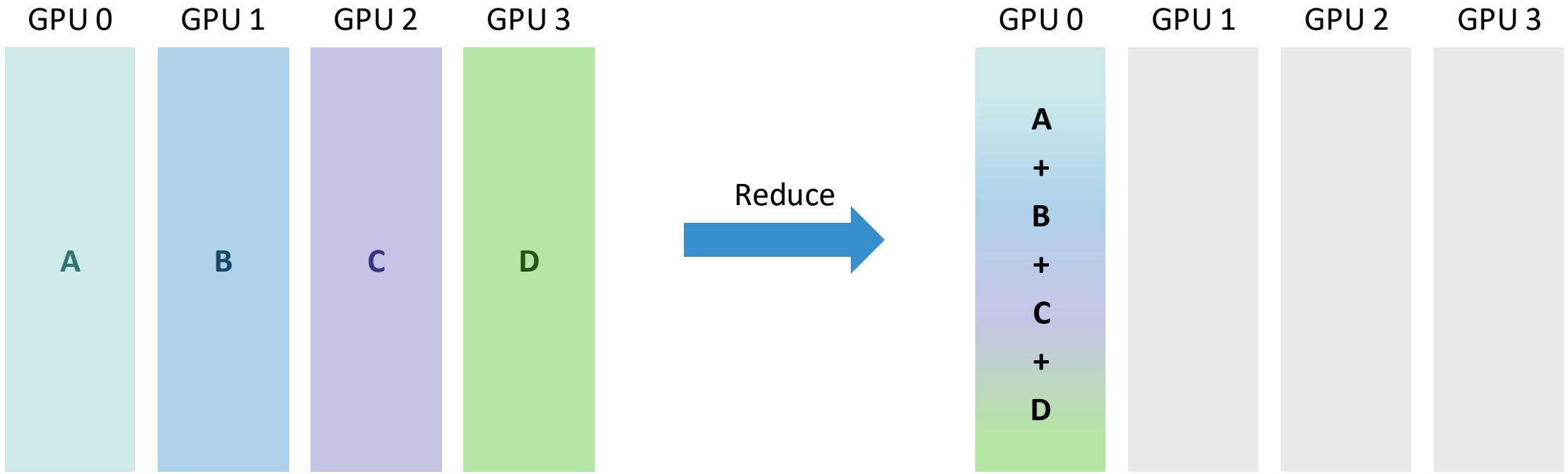
All-Gather

Gather messages from all; deliver gathered data to all participants



Reduce

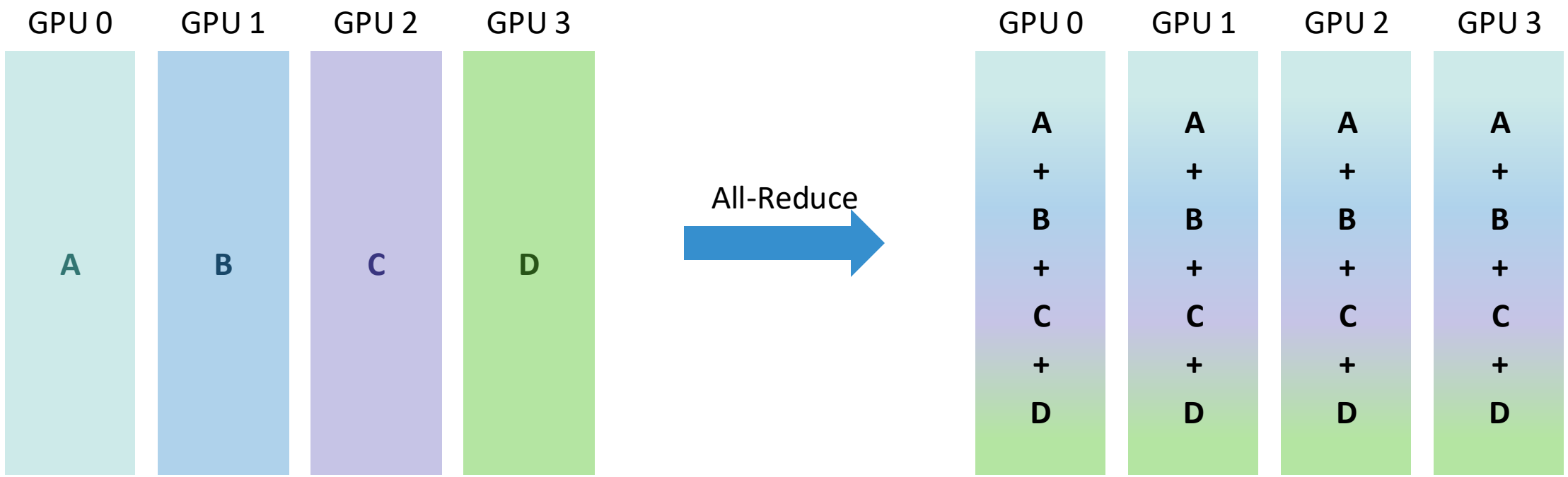
Combine data from all senders; deliver the result to one receiver



All-Reduce

上海创智学院
Shanghai Innovation Institute

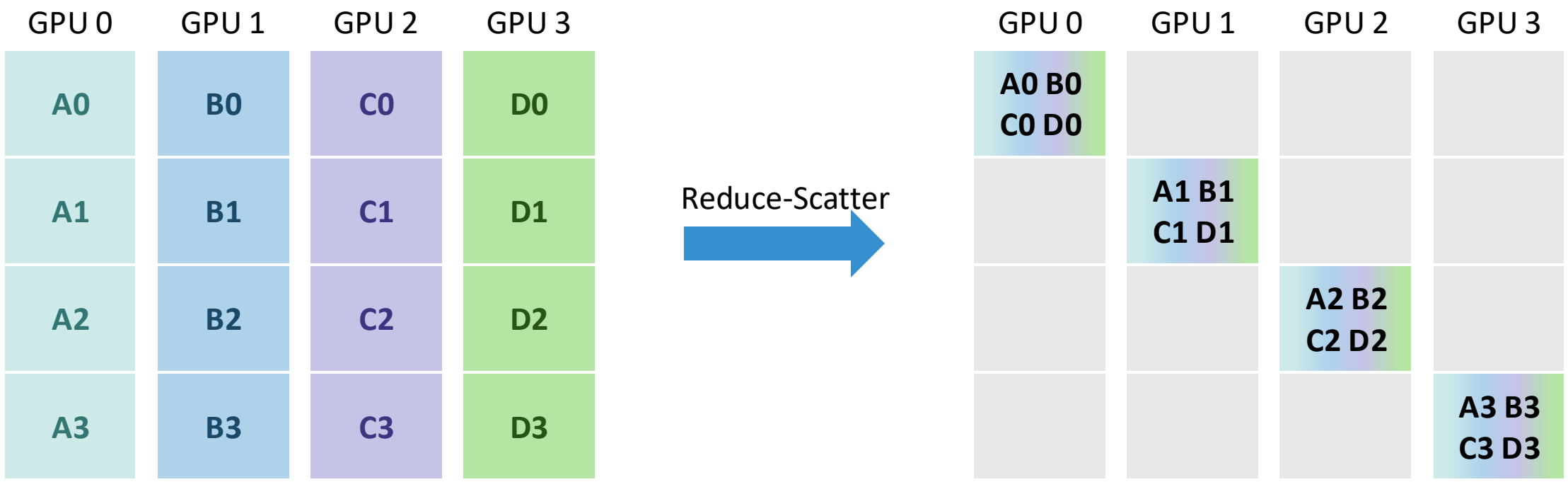
Combine data from all senders; deliver the result to all participants



Reduce-Scatter

上海创智学院
Shanghai Innovation Institute

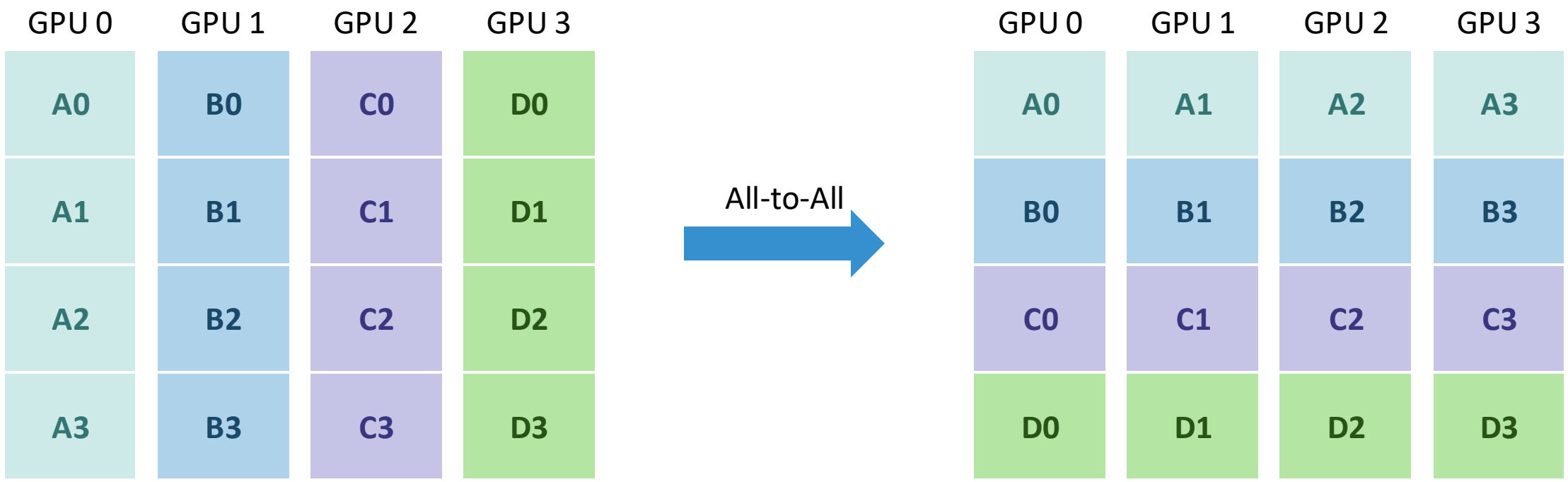
Combine data from all senders; distribute result across participants



All-to-All

上海创智学院
Shanghai Innovation Institute

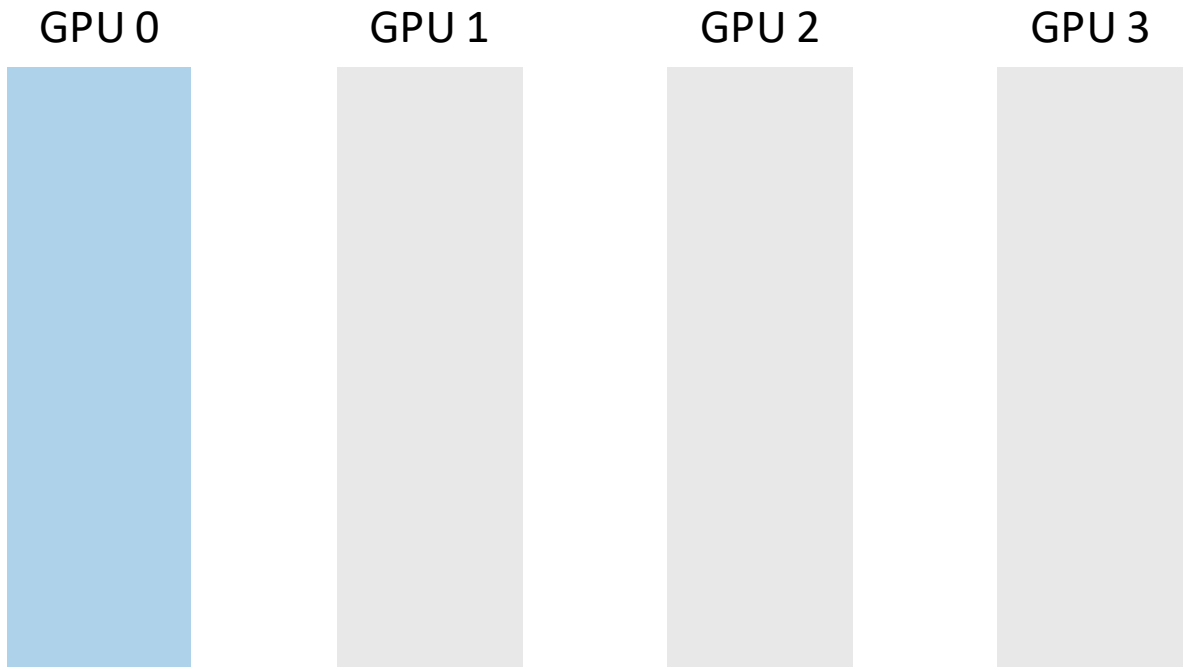
Combine data from all senders; distribute result across participants



03

Ring-Based Collectives

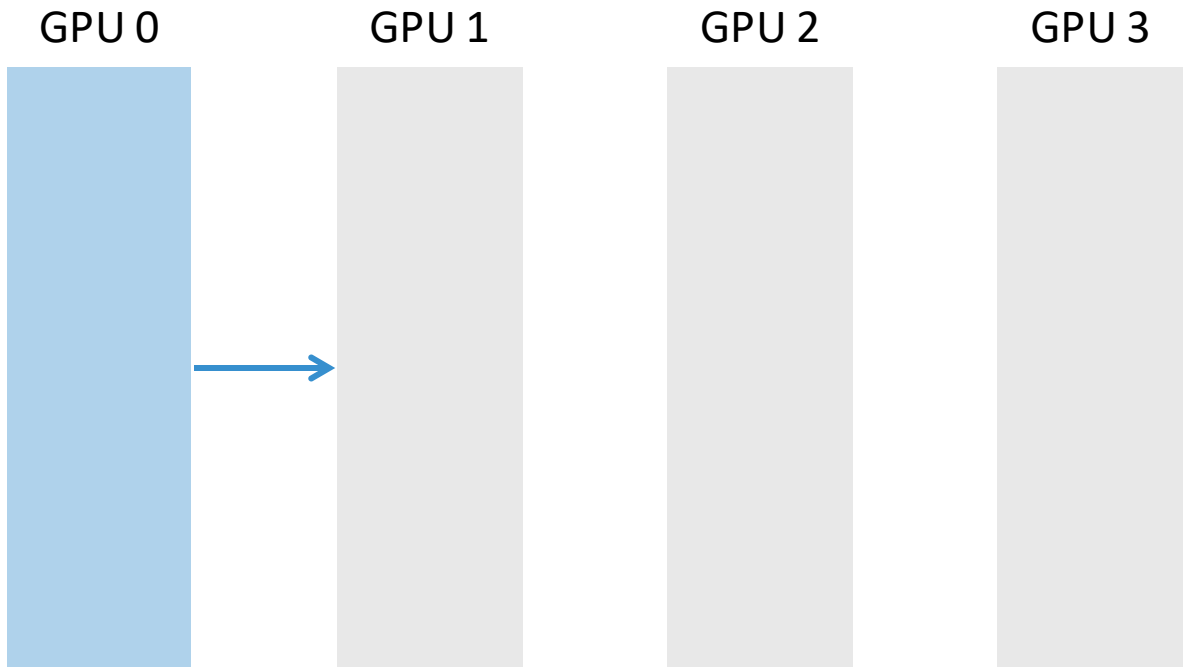
Broadcast - with Unidirectional Ring



N: Bytes to broadcast

B: Bandwidth of each link

Broadcast - with Unidirectional Ring

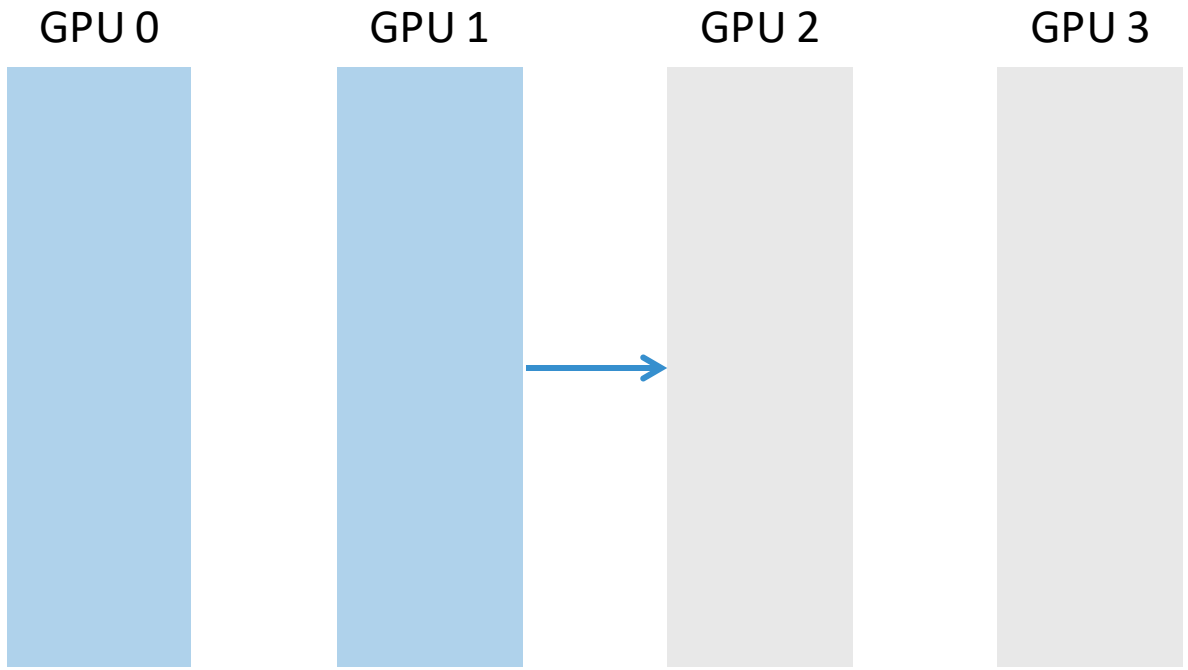


Step 1: $\Delta t = N/B$

N: Bytes to broadcast

B: Bandwidth of each link

Broadcast - with Unidirectional Ring



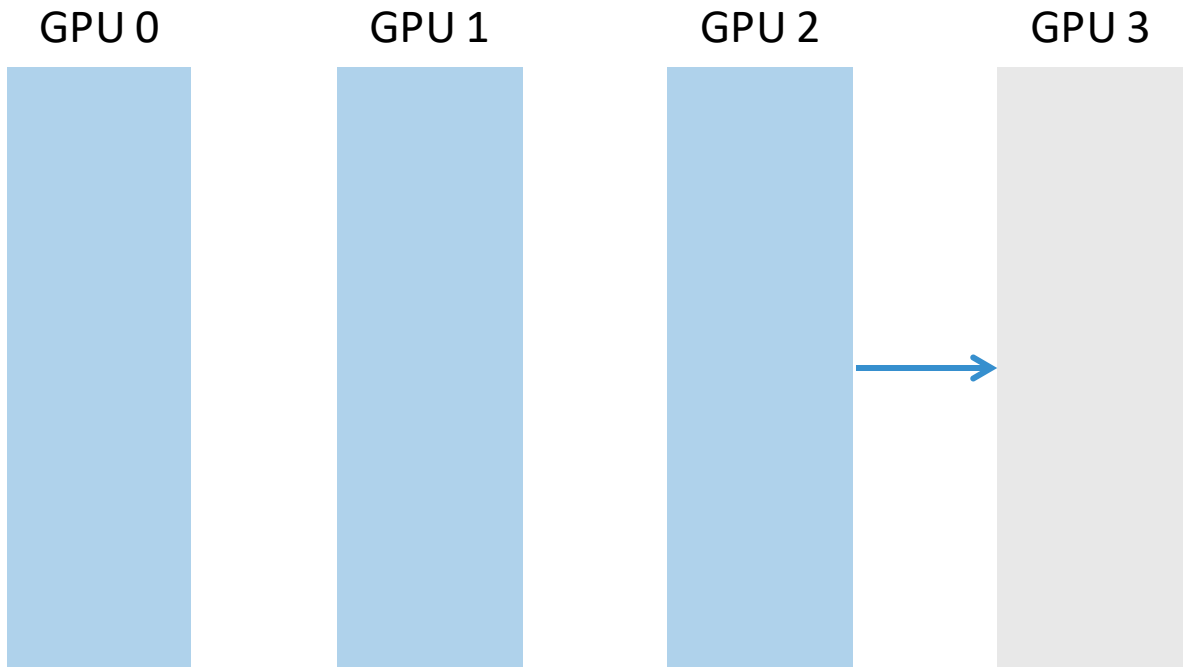
Step 1: $\Delta t = N/B$

Step 2: $\Delta t = N/B$

N: Bytes to broadcast

B: Bandwidth of each link

Broadcast - with Unidirectional Ring



Step 1: $\Delta t = N/B$

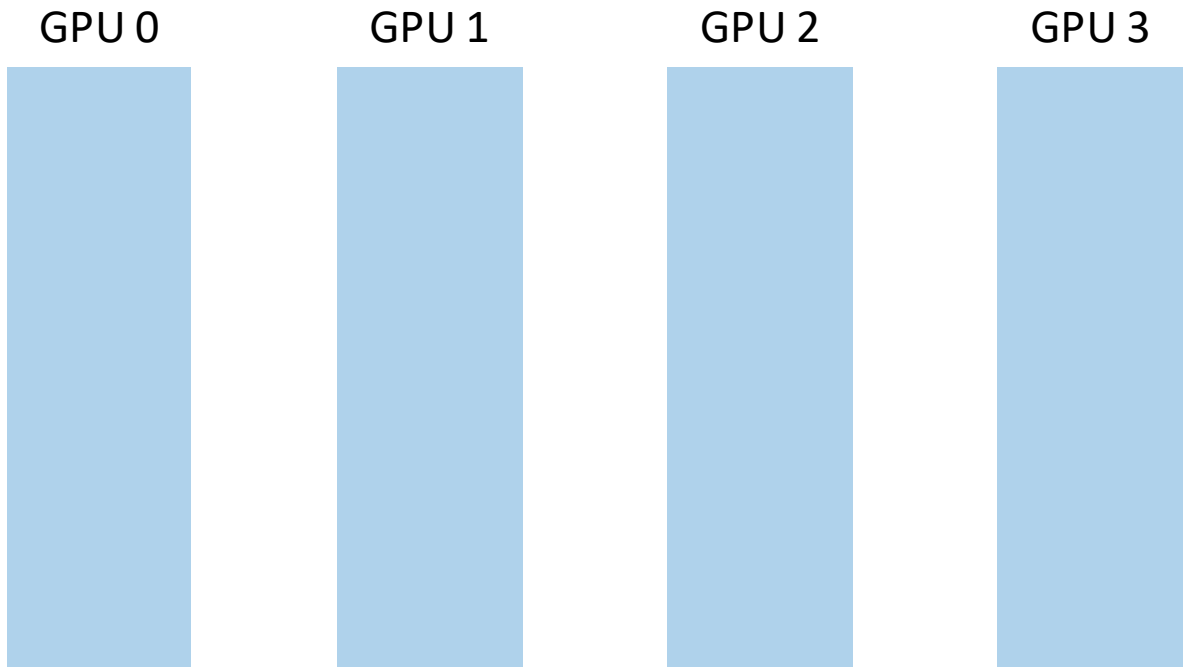
Step 2: $\Delta t = N/B$

Step 3: $\Delta t = N/B$

N: Bytes to broadcast

B: Bandwidth of each link

Broadcast - with Unidirectional Ring



Step 1: $\Delta t = N/B$

Step 2: $\Delta t = N/B$

Step 3: $\Delta t = N/B$

Total time: $(k-1)\ N/B$

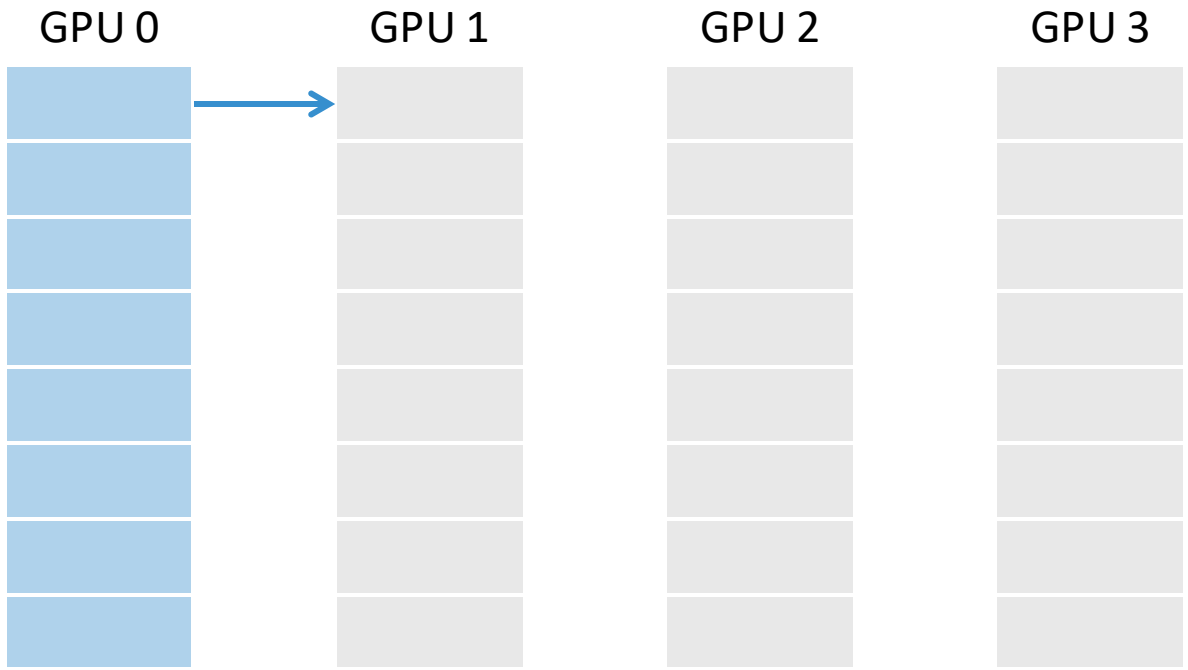
N: Bytes to broadcast

B: Bandwidth of each link

k: Number of GPUs

Can we optimize it?

Tiled Broadcast - with Unidirectional Ring



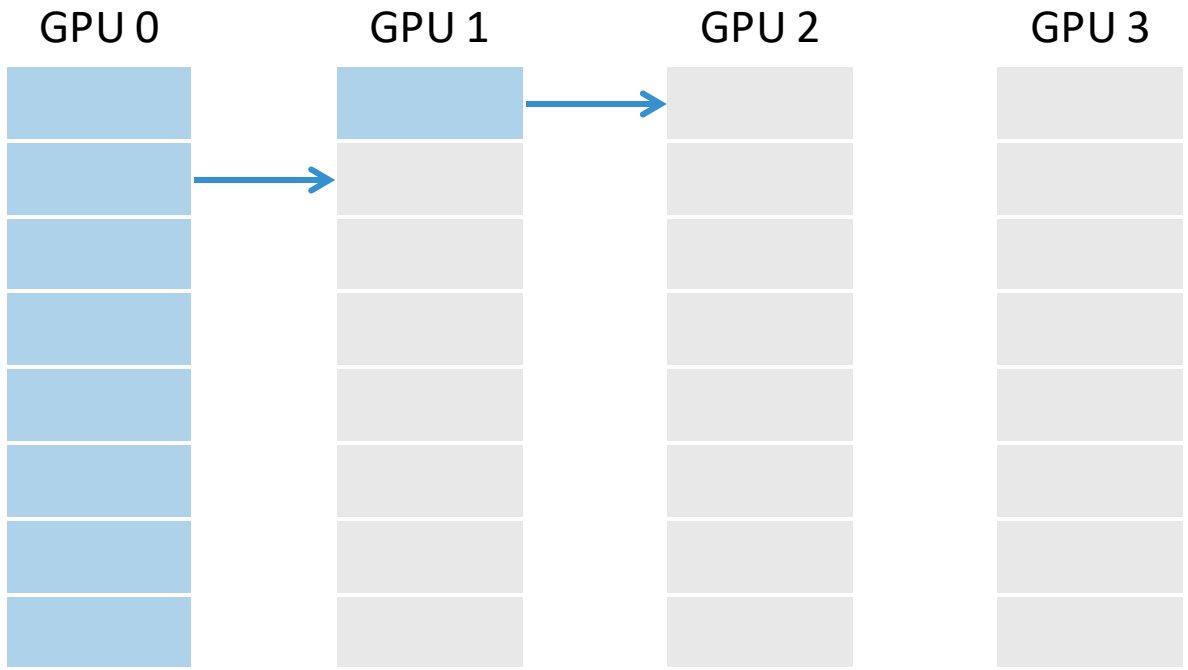
Step 1: $\Delta t = N/(SB)$

N: Bytes to broadcast

S: Number of tiles

B: Bandwidth of each link

Tiled Broadcast - with Unidirectional Ring



Step 1: $\Delta t = N/(SB)$

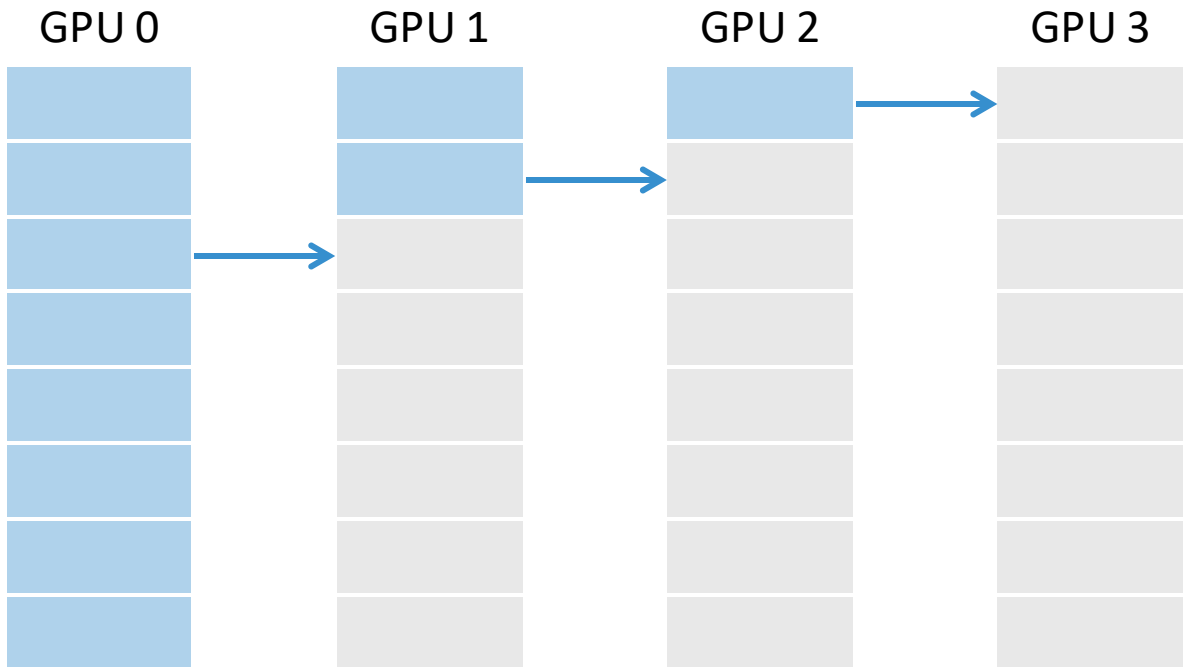
Step 2: $\Delta t = N/(SB)$

N: Bytes to broadcast

S: Number of tiles

B: Bandwidth of each link

Tiled Broadcast - with Unidirectional Ring



Step 1: $\Delta t = N/(SB)$

Step 2: $\Delta t = N/(SB)$

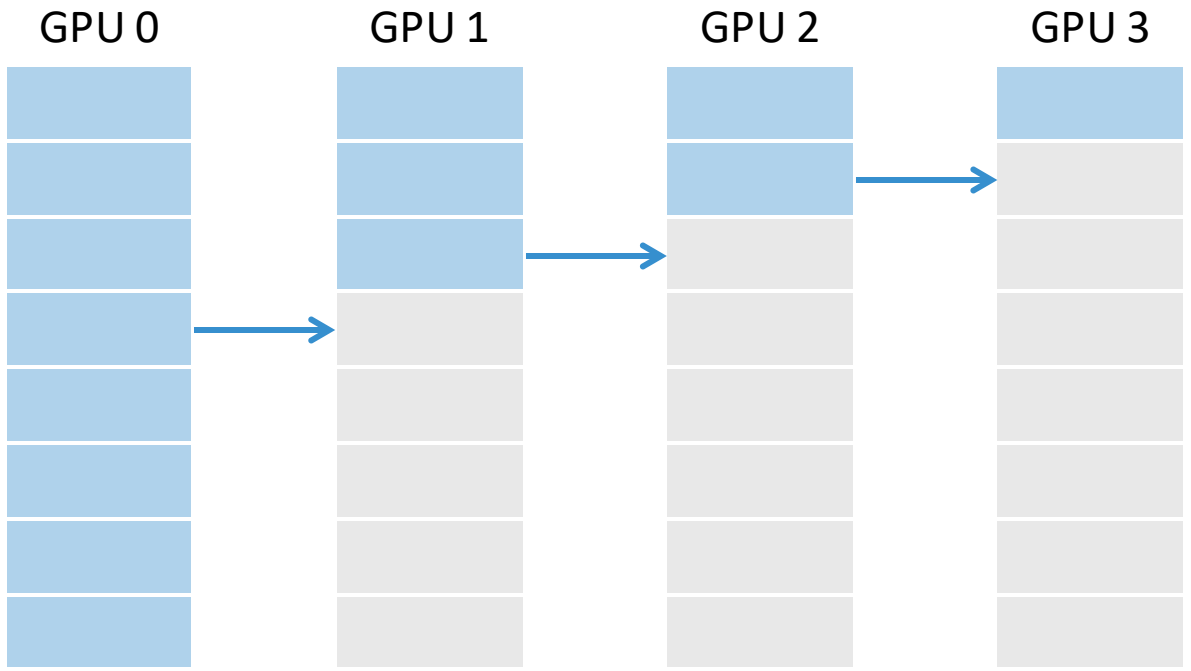
Step 3: $\Delta t = N/(SB)$

N: Bytes to broadcast

S: Number of tiles

B: Bandwidth of each link

Tiled Broadcast - with Unidirectional Ring



Step 1: $\Delta t = N/(SB)$

Step 2: $\Delta t = N/(SB)$

Step 3: $\Delta t = N/(SB)$

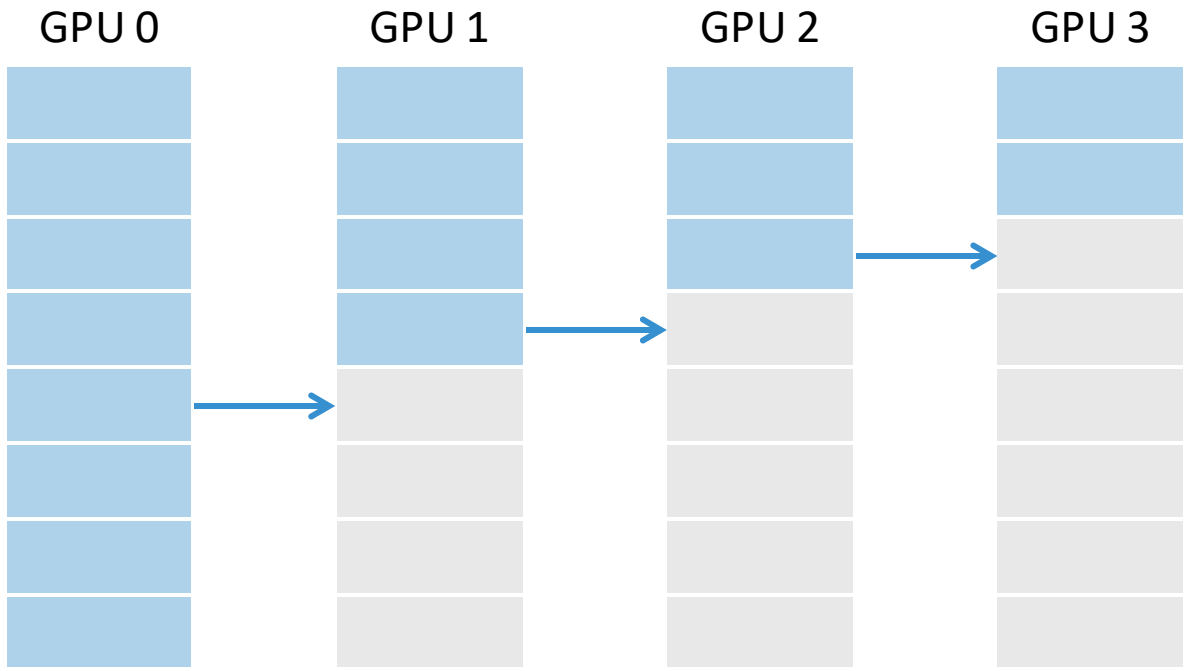
Step 4: $\Delta t = N/(SB)$

N: Bytes to broadcast

S: Number of tiles

B: Bandwidth of each link

Tiled Broadcast - with Unidirectional Ring



Step 1: $\Delta t = N/(SB)$

Step 2: $\Delta t = N/(SB)$

Step 3: $\Delta t = N/(SB)$

Step 4: $\Delta t = N/(SB)$

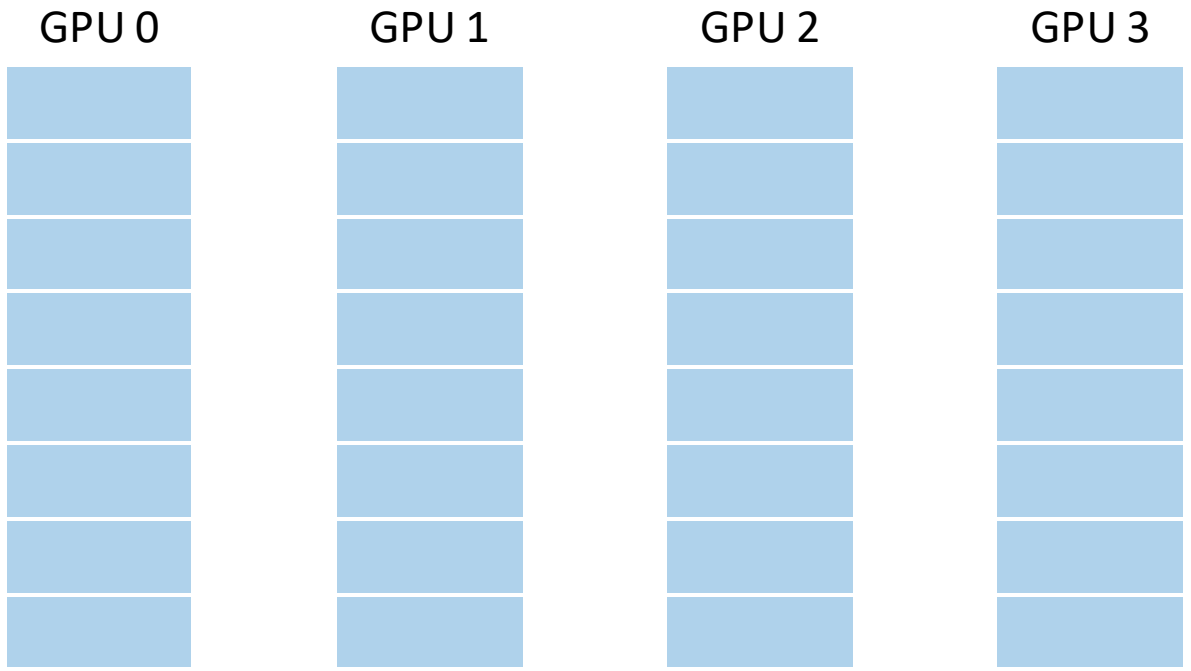
...

N: Bytes to broadcast

S: Number of tiles

B: Bandwidth of each link

Tiled Broadcast - with Unidirectional Ring



What's the total time?

$$\frac{SN}{SB} + (k-2)\frac{N}{SB} = \frac{N(S+k-2)}{SB} \to \frac{N}{B}$$

Step 1: $\Delta t = N/(SB)$

Step 2: $\Delta t = N/(SB)$

Step 3: $\Delta t = N/(SB)$

Step 4: $\Delta t = N/(SB)$

...

N: Bytes to broadcast

S: Number of tiles

B: Bandwidth of each link

Acknowledgement

The development of this course, including its structure, content, and accompanying presentation slides, has been significantly influenced and inspired by the excellent work of instructors and institutions who have shared their materials openly. We wish to extend our sincere acknowledgement and gratitude to the following courses, which served as invaluable references and a source of pedagogical inspiration:

- Machine Learning Systems[15-442/15-642], by **Tianqi Chen** and **Zhihao Jia** at **CMU**.
- Advanced Topics in Machine Learning (Systems)[CS6216], by **Yao Lu** at **NUS**

While these materials provided a foundational blueprint and a wealth of insightful examples, all content herein has been adapted, modified, and curated to meet the specific learning objectives of our curriculum. Any errors, omissions, or shortcomings found in these course materials are entirely our own responsibility. We are profoundly grateful for the contributions of the educators listed above, whose dedication to teaching and knowledge-sharing has made the creation of this course possible.

Acknowledgement

This slide is inspired by and references the following key works:

- "NCCL: ACCELERATED MULTI-GPU COLLECTIVE COMMUNICATIONS" by Cliff Woolley (NVIDIA), which details a high-performance library for multi-GPU collective operations.

- "Algorithms for Collective Communication" (from the course Design and Analysis of Parallel Algorithms), which provides a theoretical foundation for collective operations and their implementation on various network topologie.

System for Artificial Intelligence

Thanks

Siyuan Feng
Shanghai Innovation Institute