DualPipe & Cross-Node All-to-All Communication

Mentor ： HAIQUAN WANG

ZEWEN JIN

Reporter: JIAHUI TAN

- JIAHUI TAN
- School : UESTC

(University of Electronic Science and Technology of China)

- Grade : junior year
- Major : software engineering
- Email : 2022090907024@std.uestc.edu.cn
- Finding graduate school opportunities

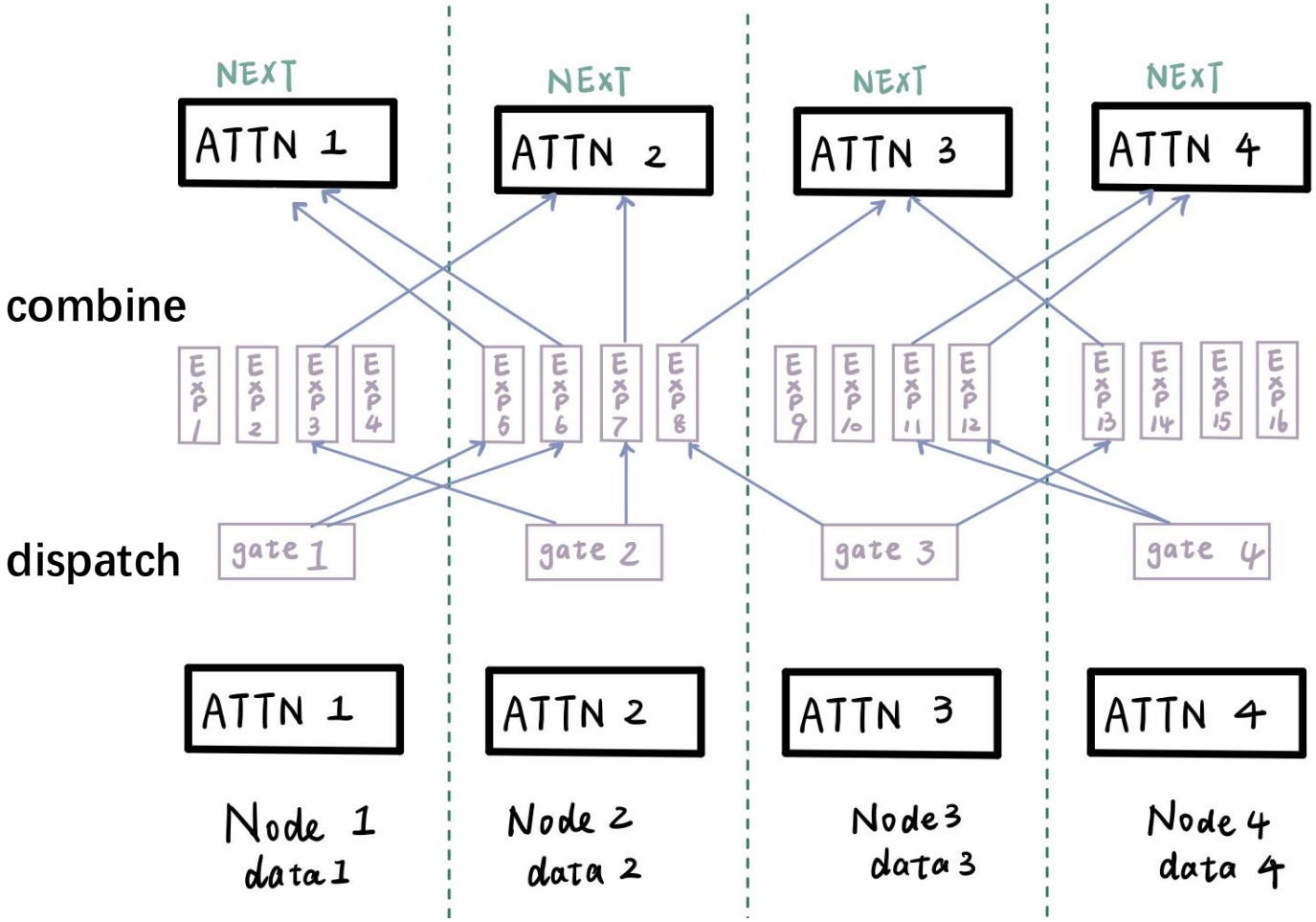
DualPipe & Cross-Node All-to-All Communication

- Background
- DualPipe
  - Insight & Motivation
  - Design & Implementation
  - Comparison
- All-to-All
  - Insight & Motivation
  - Design & Implementation

Background

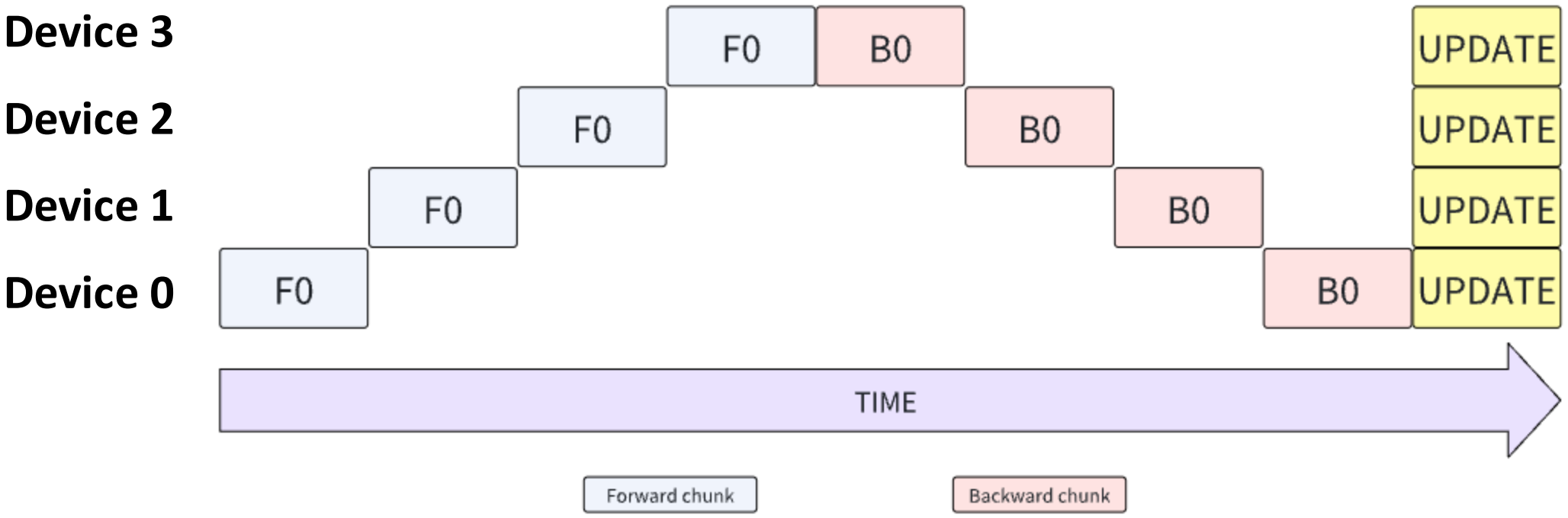
- **EP**

an inefficient computation-to-communication ratio of approximately 1:1



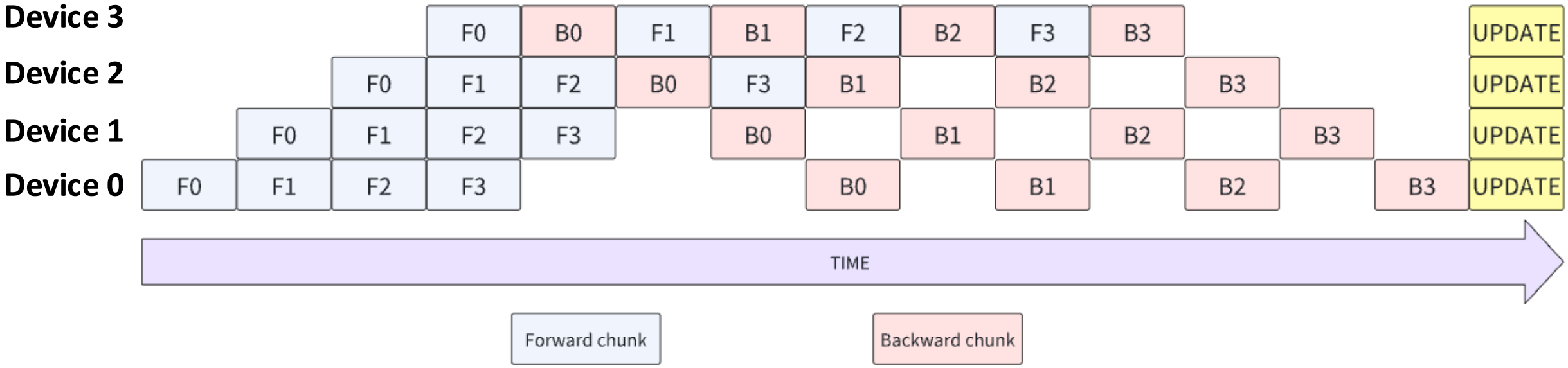
Background

- PP
  - Naive Pipeline Parallelism



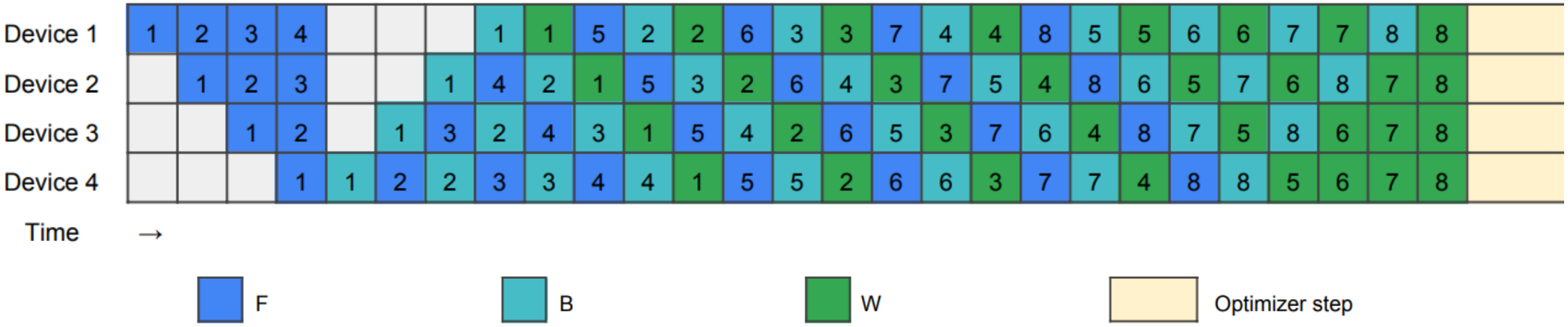
Background

- PP
  - 1F1B



Background

- PP
  - Zero Bubble PP



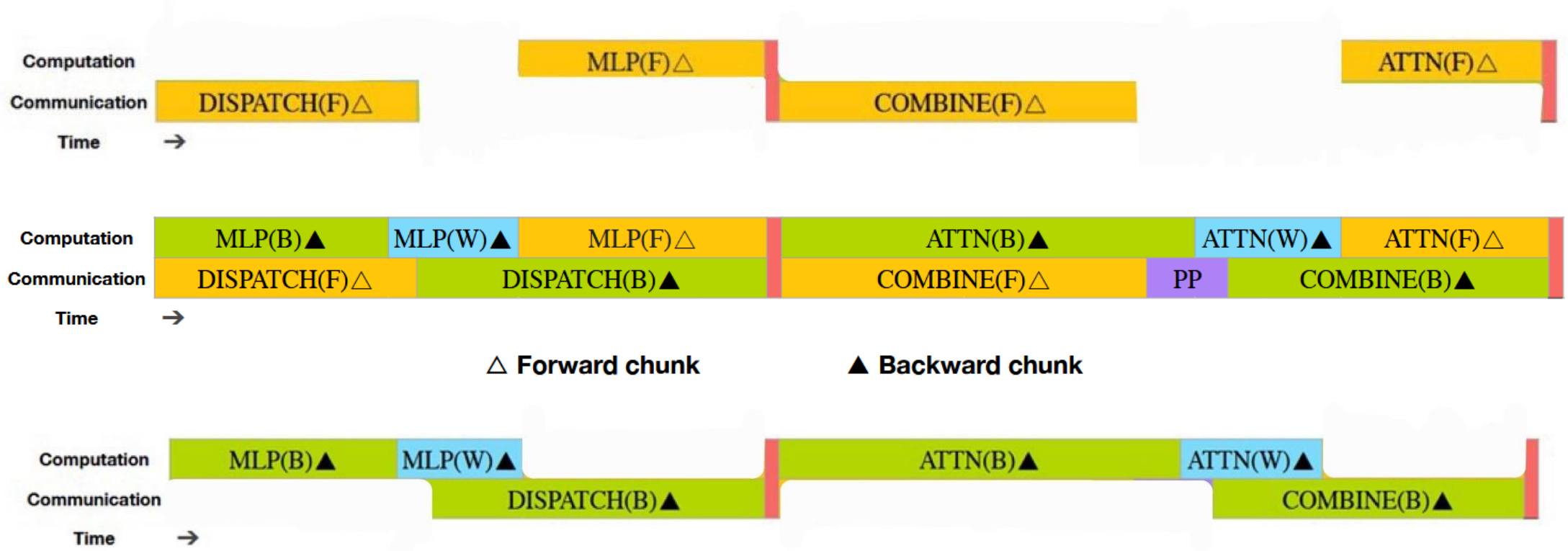
DualPipe Insight & Motivation

- heavy communication overhead introduced by expert parallelism

  - overlap the computation and communication phases across forward and backward processes

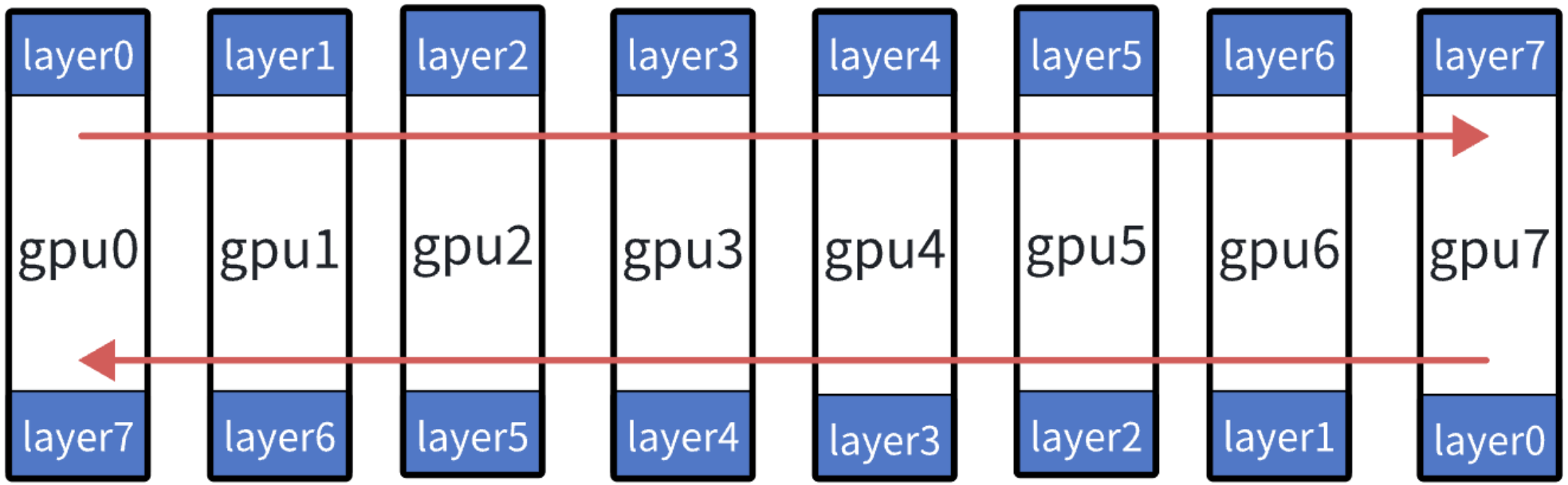
  - reduce the pipeline bubbles (improve GPU utilization)

DualPipe Design & Implementation

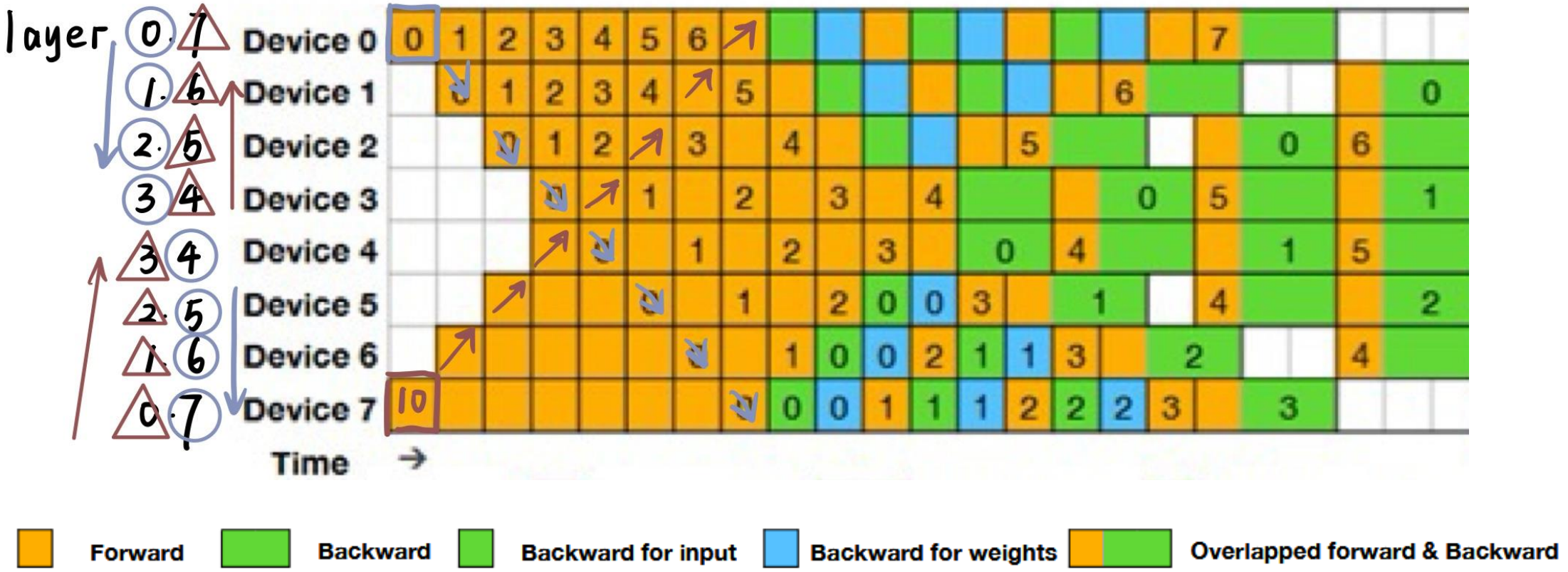


- overlap the computation and communication within a pair of individual forward and backward chunks

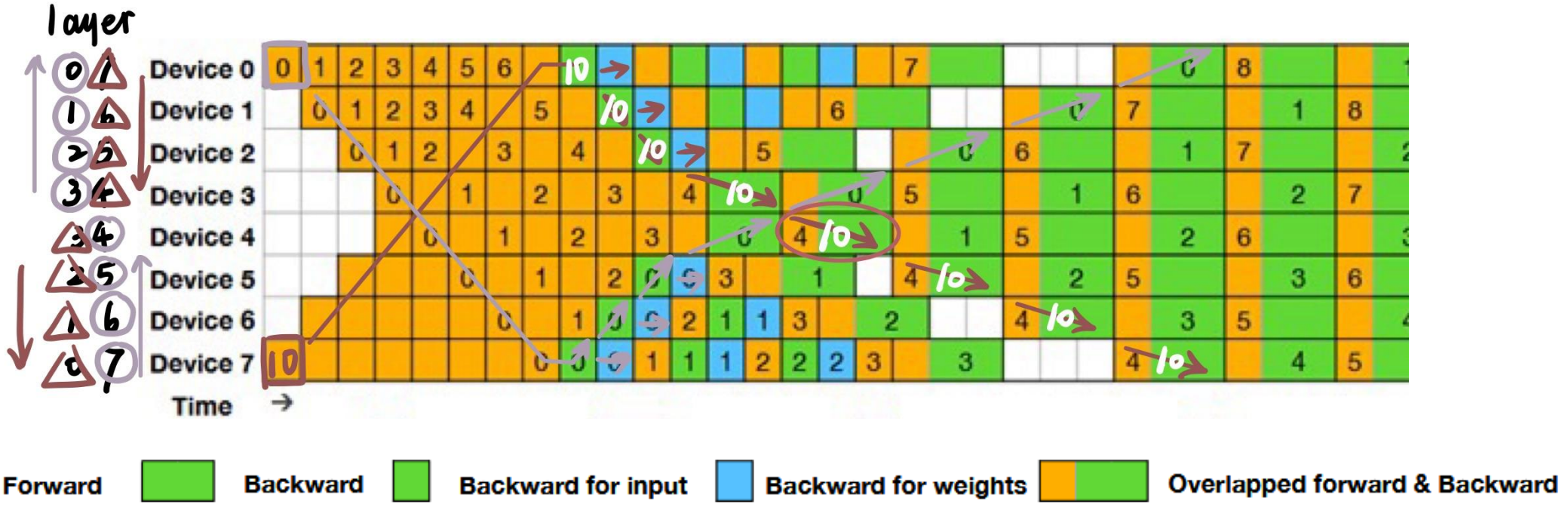
DualPipe Design & Implementation

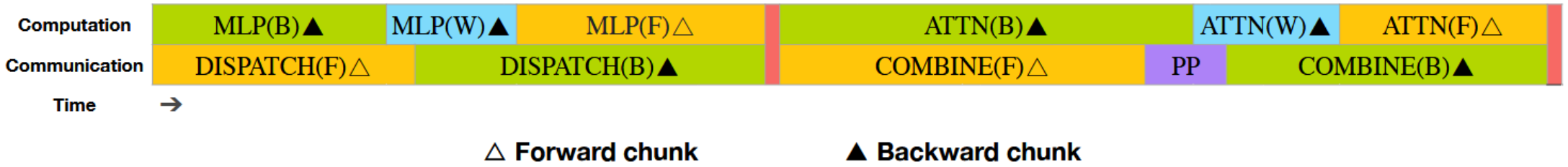


DualPipe Design & Implementation

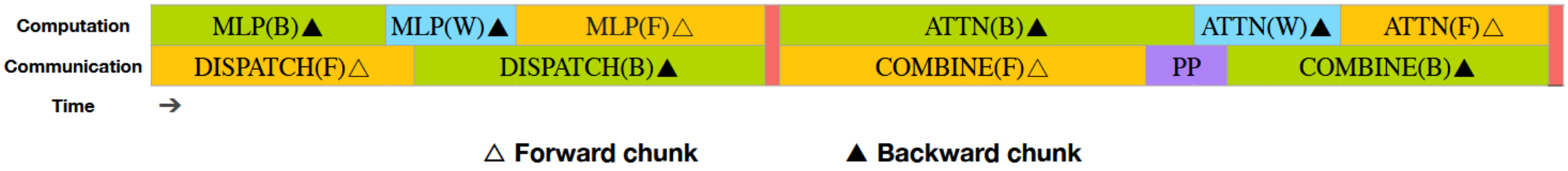


DualPipe Design & Implementation





DualPipe Design & Implementation



Comparison

Method	Bubble	Parameter	Activation
1F1B	$(PP-1)(F+B)$	1×	PP
ZB1P	$(PP-1)(F+B-2W)$	1×	PP
DualPipe (Ours)	$(\frac{PP}{2}-1)(F\&B+B-3W)$	2×	$PP+1$

DualPipe significantly reduces the pipeline bubbles while only increasing the peak activation memory by 1 PP times

Link

DualPipeV:

https://github.com/deepseek-ai/DualPipe

Nvidia:

https://mp.weixin.qq.com/s/vCy6ga5EA2dzvFoL8p6QjA

All-to-All Insight & Motivation

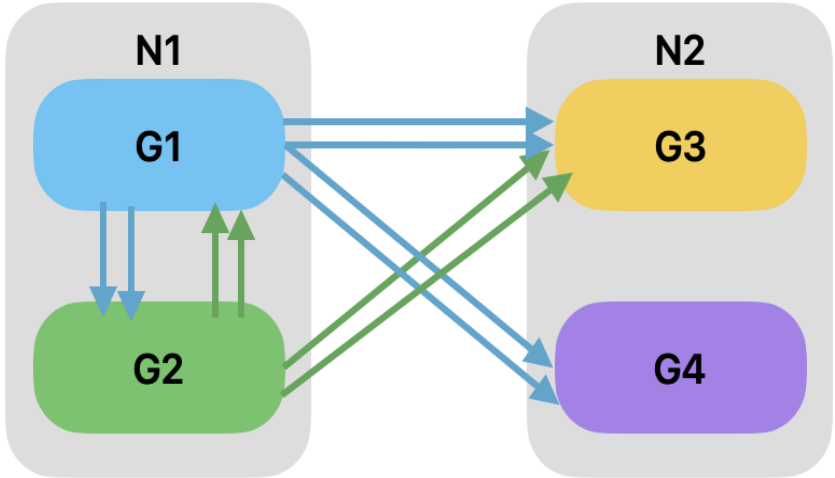
- To ensure sufficient computational performance for DualPipe
  - Utilizing IB and NVLink bandwidths
  - Conserving Streaming Multiprocessors (SMs) dedicated to communication (20 SMs)

all-to-all Design & Implementation

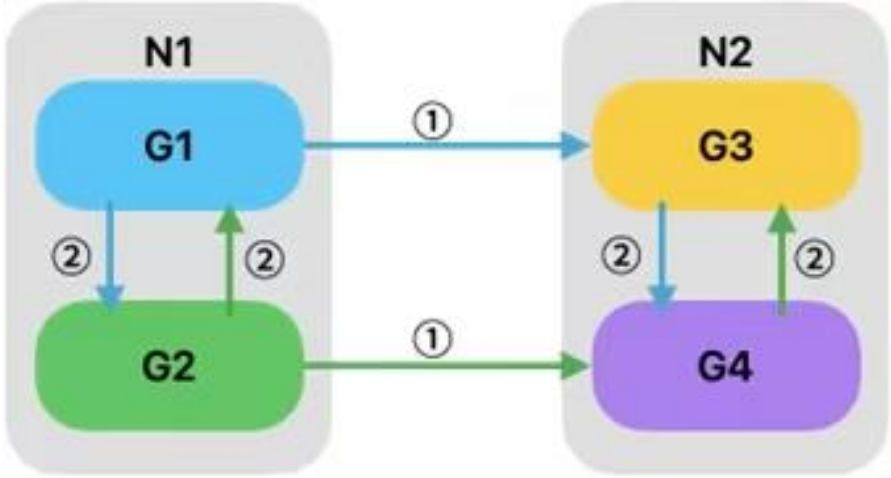
- cross-node GPUs are fully interconnected with IB
- intra-node communications are handled via NVLink
- NVLink 160 GB/s, IB (50 GB/s)

all-to-all Design & Implementation

- For each token
  - Be transmitted via IB to the GPUs with the same in-node index on its target nodes.
  - Be forwarded via NVLink to specific GPUs that host their target experts
  - Be dispatched to at most 4 nodes



Naive All2All



deepEP all to all

THANKS
FOR YOUR LISTENING

Mentor : HAIQUAN WANG ZEWEN JIN

Reporter : JIAHUI TAN

2025.3.11