ATC'25, Boston, MA, USA

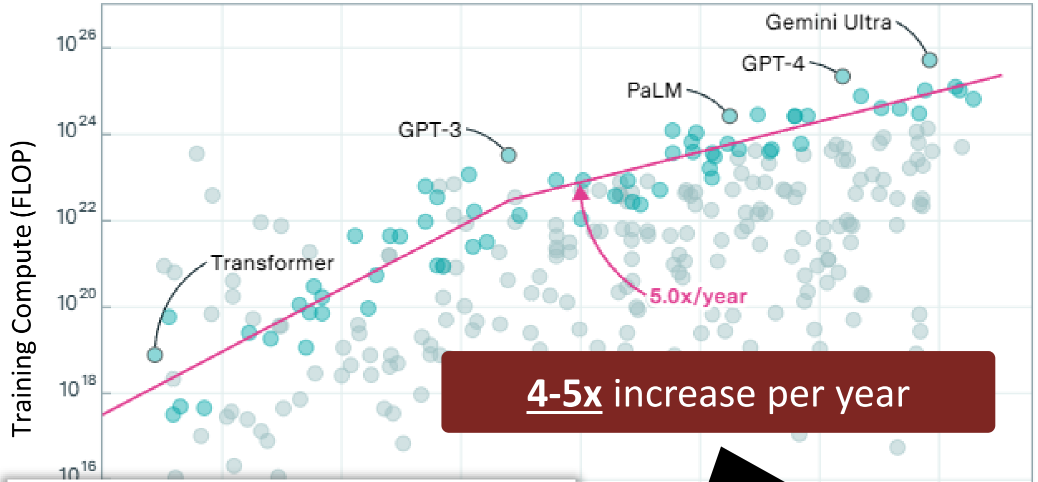
CrossPipe: Towards Optimal Pipeline Schedules for Cross-Datacenter Training

TIANCHENG CHEN, **ALES KUBICEK**, LANGWEN HUANG, TORSTEN HOEFLER

Motivation

Frontier LLM Models

[1]



4-5x increase per year

Increasing power demand

The New York Times

Hungry for Energy, Amazon, Google and Microsoft Turn to Nuclear Power

Large technology companies are inve… nuclear energy as an emissions-free s… artificial intelligence and other busine…

IN-BRIEF ANALYSIS

OCTOBER 1, 2024

Data center owners turn to nuclear as potential electricity source

Nuclear power plants that have signed agreements to power data centers (as of Sep 2024)

HOME > NEWS > IT HARDWARE & SEMICONDUCTORS

Training Google's G… data centers, and ris…

What we can glean from research p… Google Cloud's CEO

Decemb…

Tech Industry > Artificial Intelligence

Microsoft Azure CTO … AI training is needed … approach power grid…

News By Anton Shilov published October 18…

While others seek a nuclear solu…

Commen…

When you purchase through links on our site, w… works.

September 4, 2024

Multi-Datacenter Training: OpenAI's Ambitious Plan To Beat Google's Infrastructure

// Gigawatt Clusters, Telecom Networking, Long Haul Fiber, Hierarchical & Asynchronous SGD, Distributed Infrastructure Winners

34 minutes

18 comments

By Dylan Patel, Daniel Nishball and Jeremie Eliahou Ontiveros

*"**We don't disclose exactly the details** of how many locations but Gemini Ultra was trained **across multiple sites**, and multiple clusters within those sites."*

Thomas Kurian, CEO Google Cloud

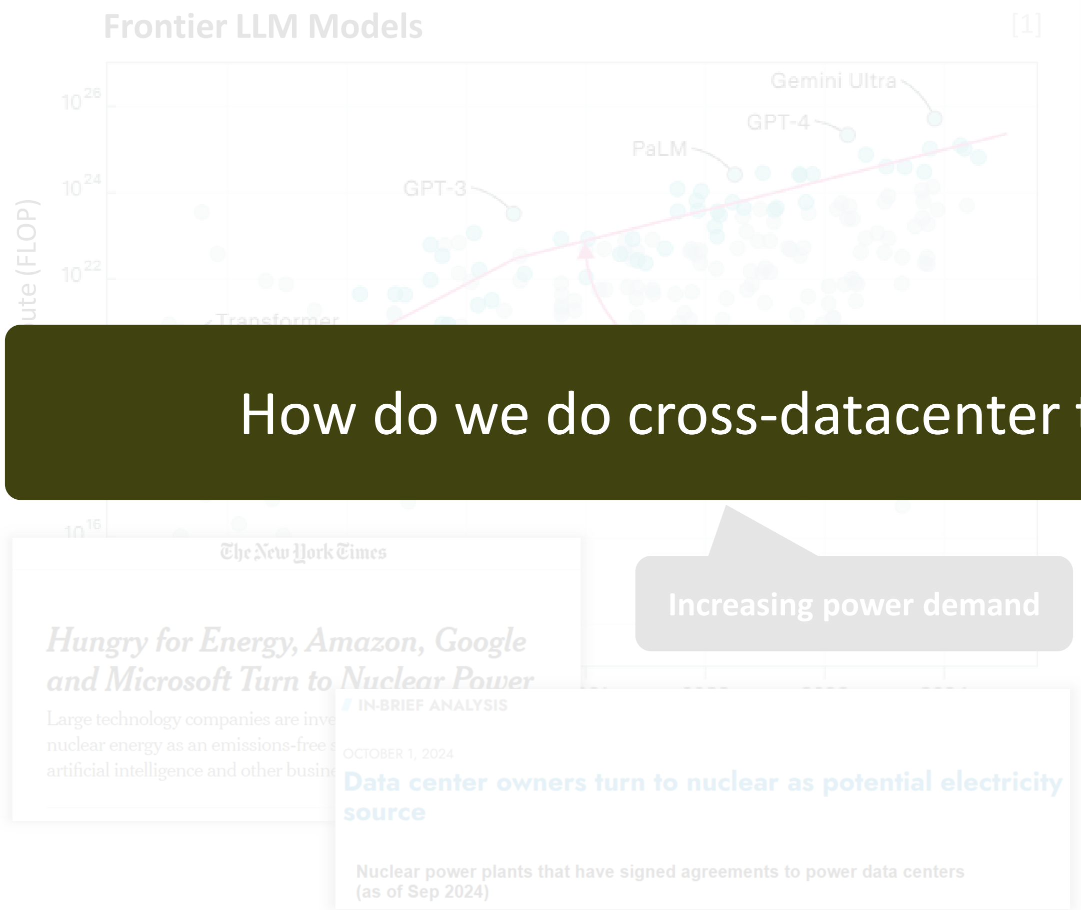
December 2023

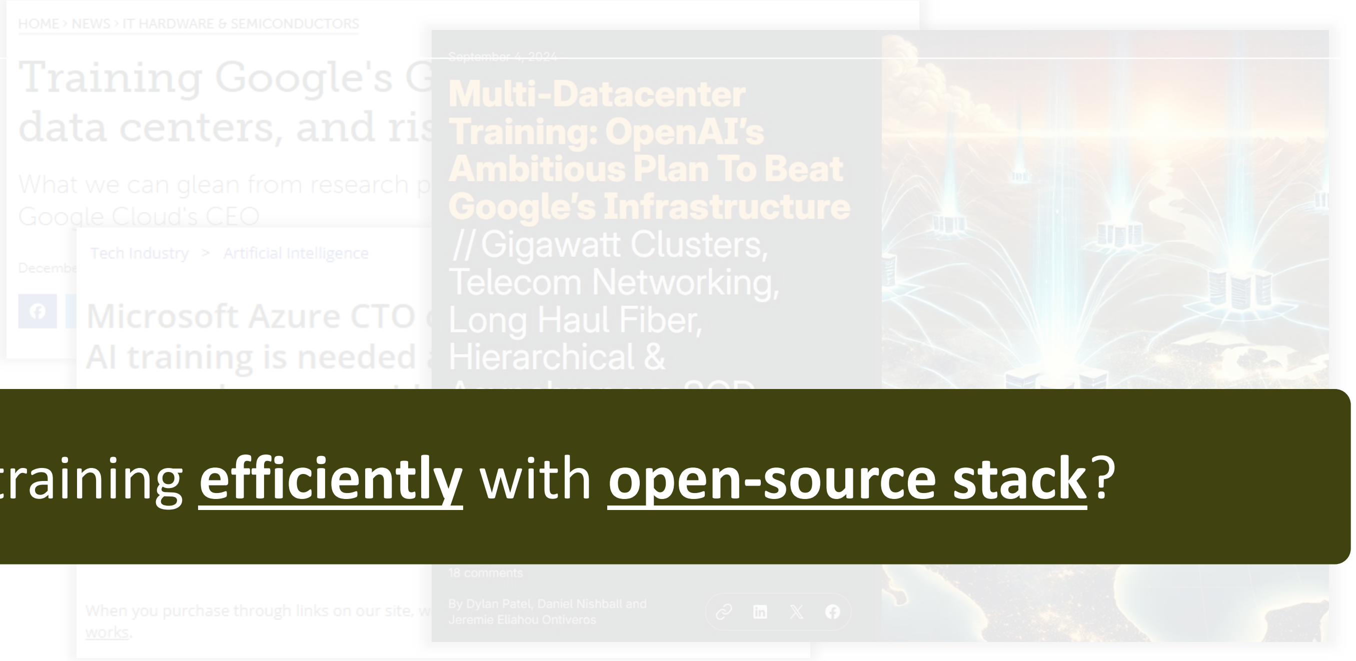
Problem: local power limitation

Solution: training across **multiple** datacenters

[1] Training Compute of Frontier AI Models Grows by 4-5x per Year, Report, Epoch AI, 2024

Motivation





How do we do cross-datacenter training **efficiently** with **open-source stack**?

*"We don't disclose exactly the details of how many locations but Gemini Ultra was trained **across multiple sites**, and multiple clusters within those sites."*

Thomas Kurian, CEO Google Cloud
December 2023

Problem: local power limitation → **Solution:** training across **multiple** datacenters

[1] Training Compute of Frontier AI Models Grows by 4-5x per Year, Report, Epoch AI, 2024

Problem and Parallelism Strategies

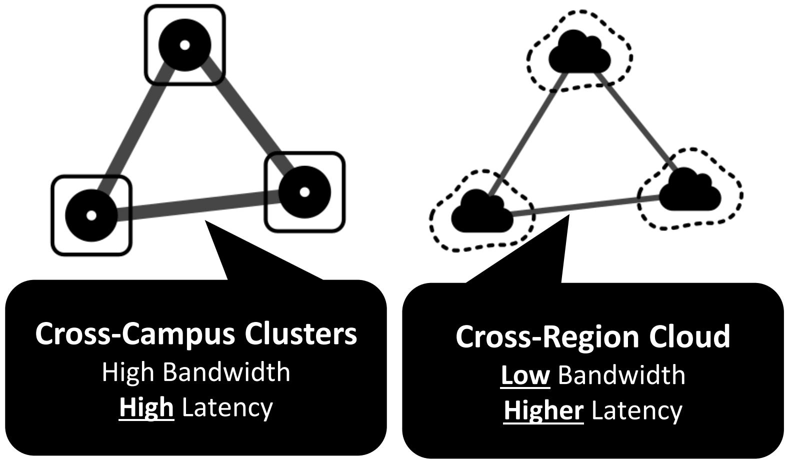
- **Challenge:** high latency / low bandwidth on cross-DC boundary

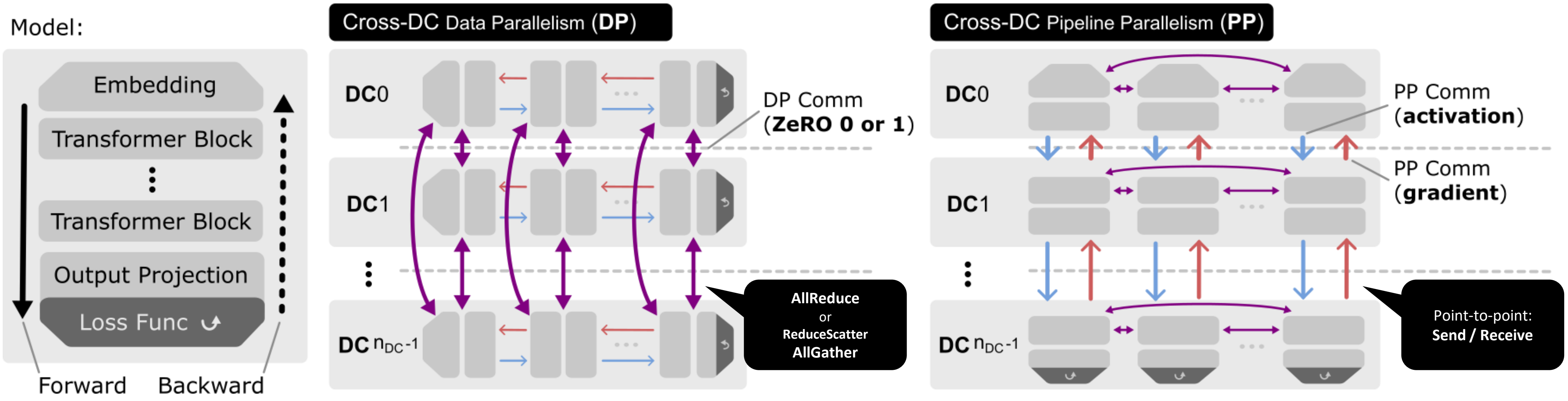
- Hybrid parallelism
  - Tensor (= Operator) n_{TP}
  - Sequence (= Context)
  - Expert
  - **Pipeline** n_{PP}
  - **Data** n_{DP}

High-bandwidth domain (e.g., NVLink)

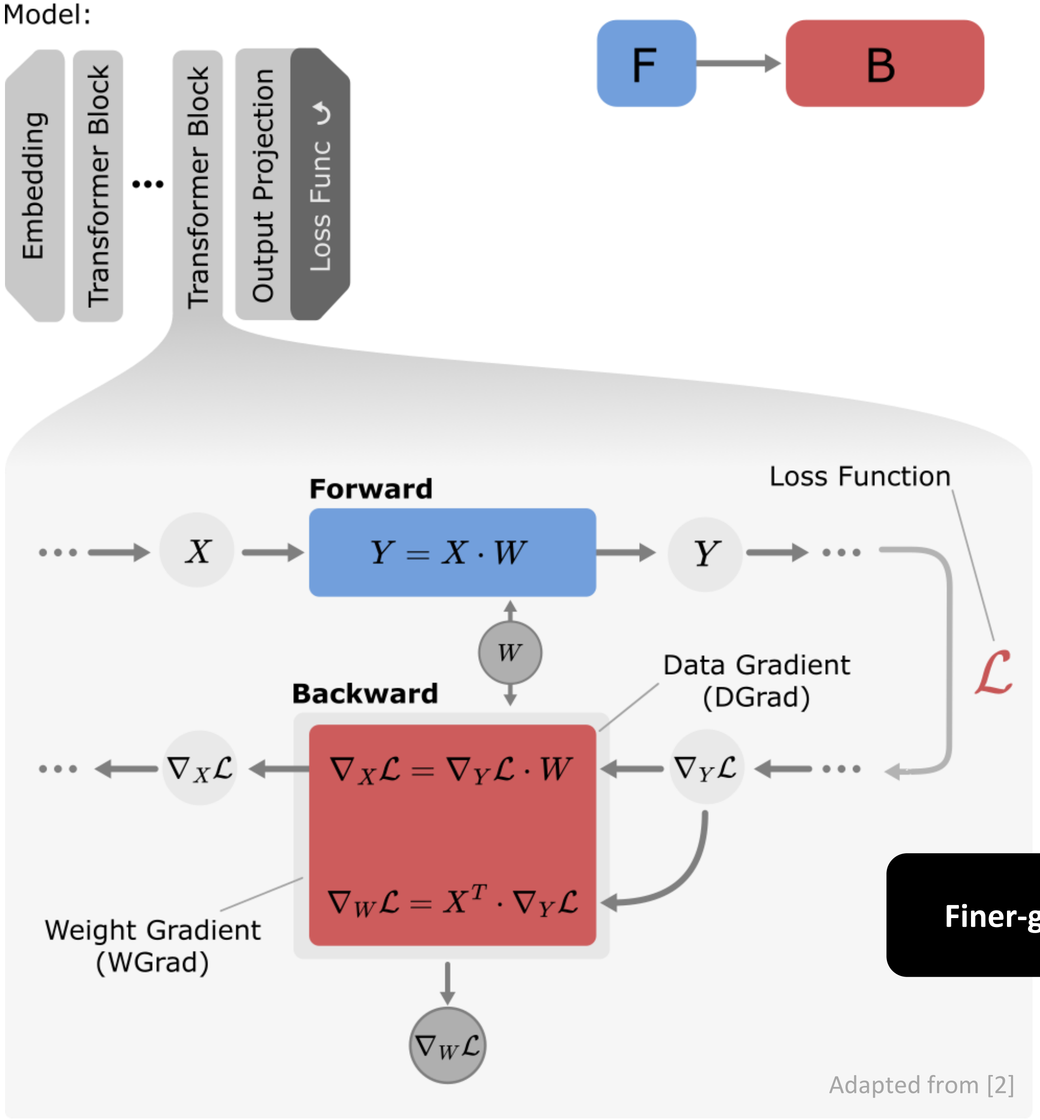
Not suitable for Cross-DC:
Layer-wise communication
High frequency and/or volume

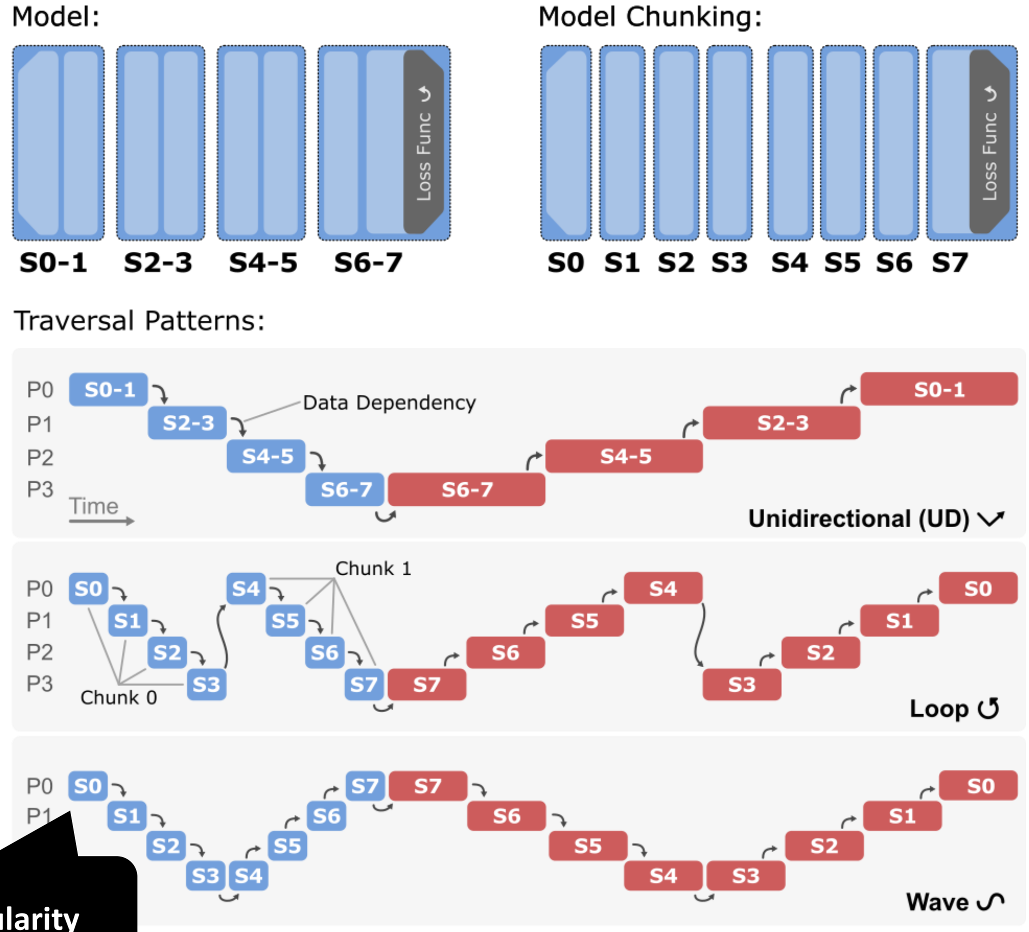
Suitable for Cross-DC:
Lower frequency and point-to-point (PP)





Pipeline Parallelism – Data Dependencies

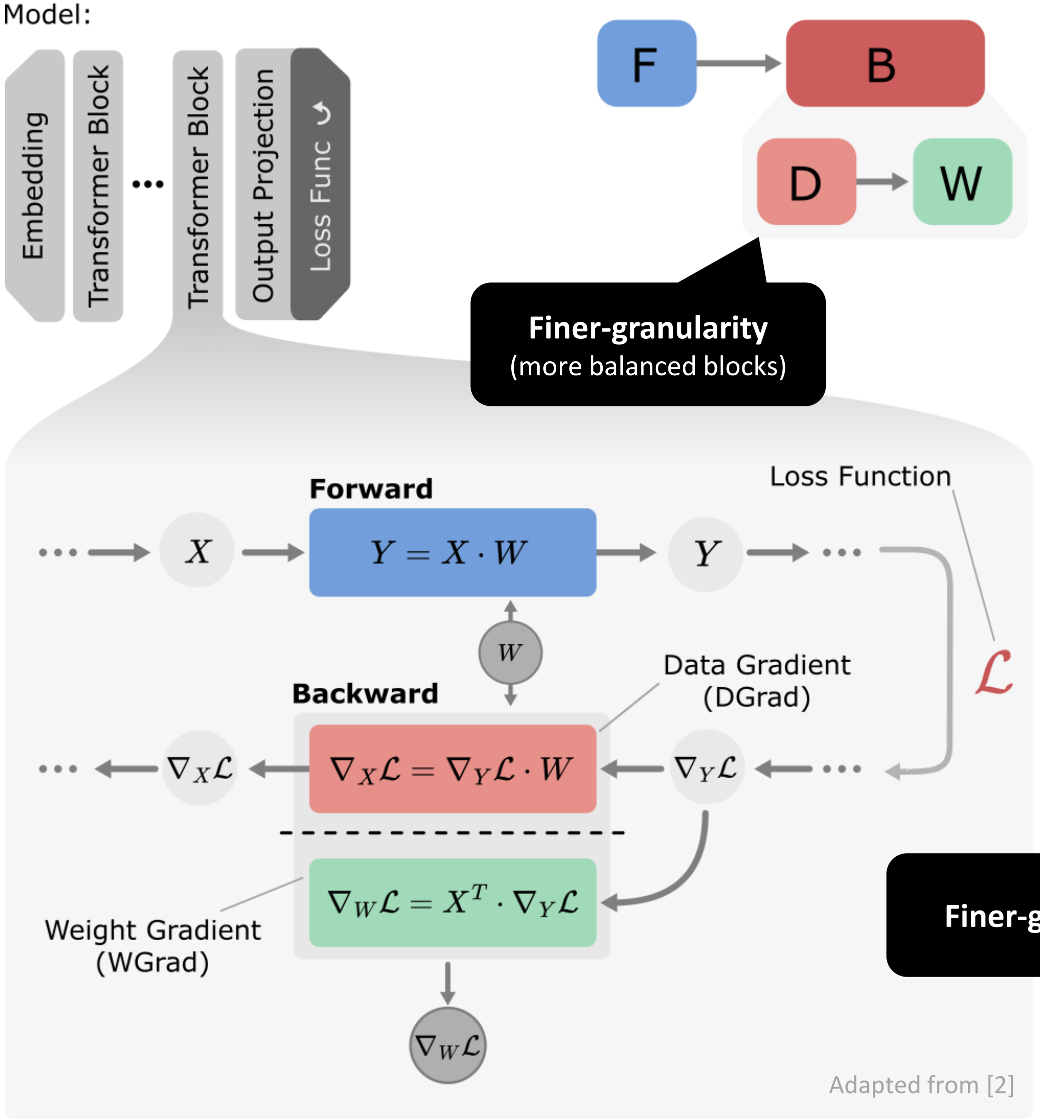


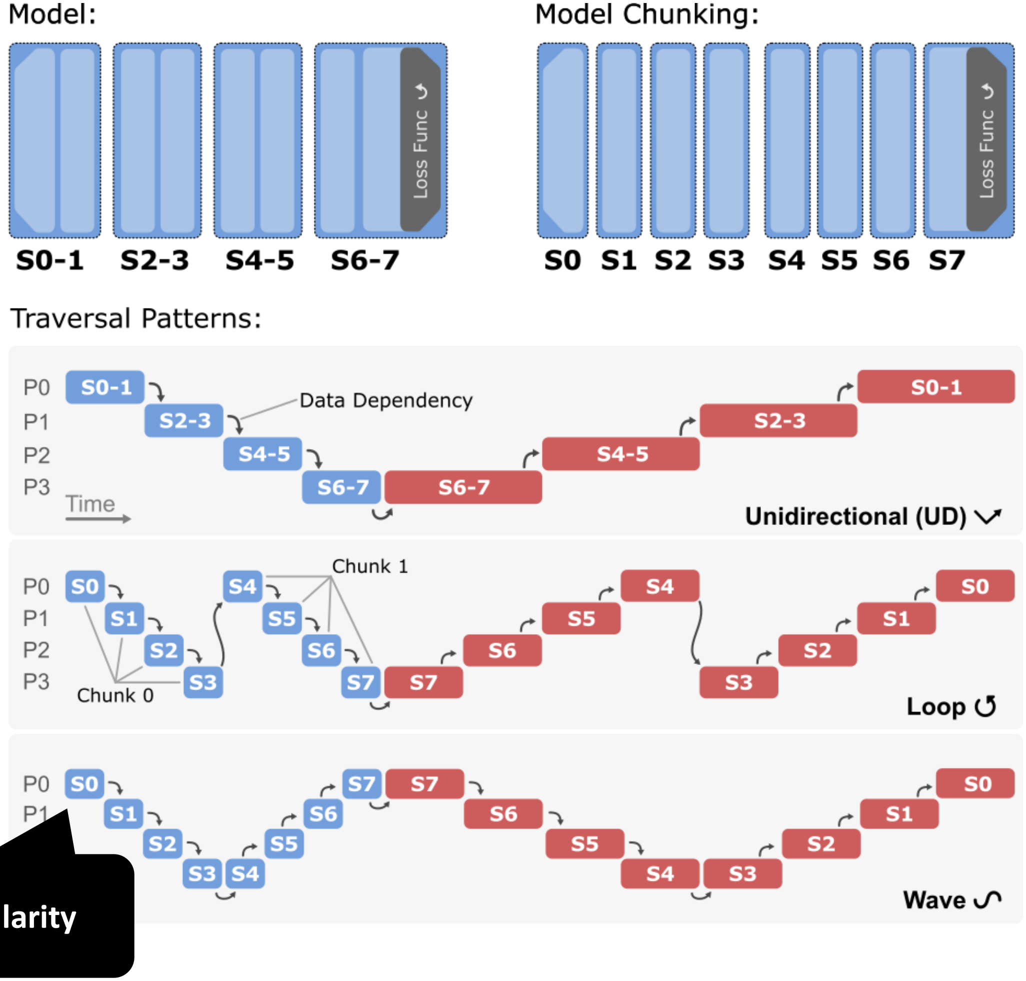


Finer-granularity

[2] Qi, Penghui, et al. Zero bubble pipeline parallelism (2023)

Pipeline Parallelism – Data Dependencies

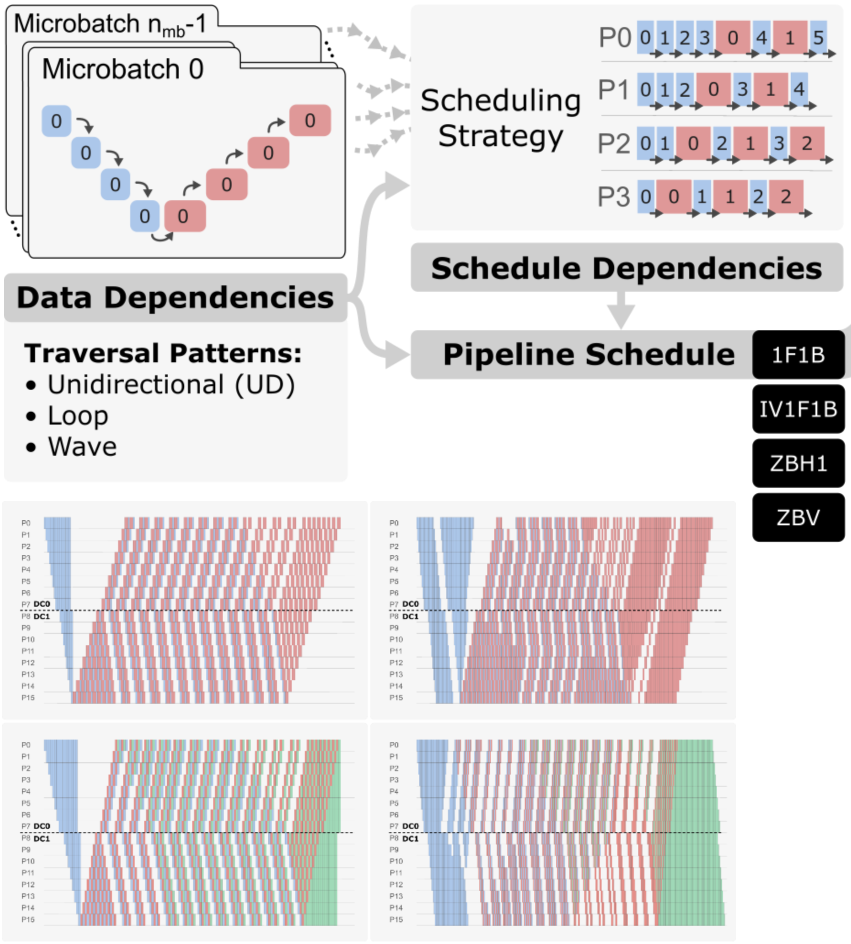


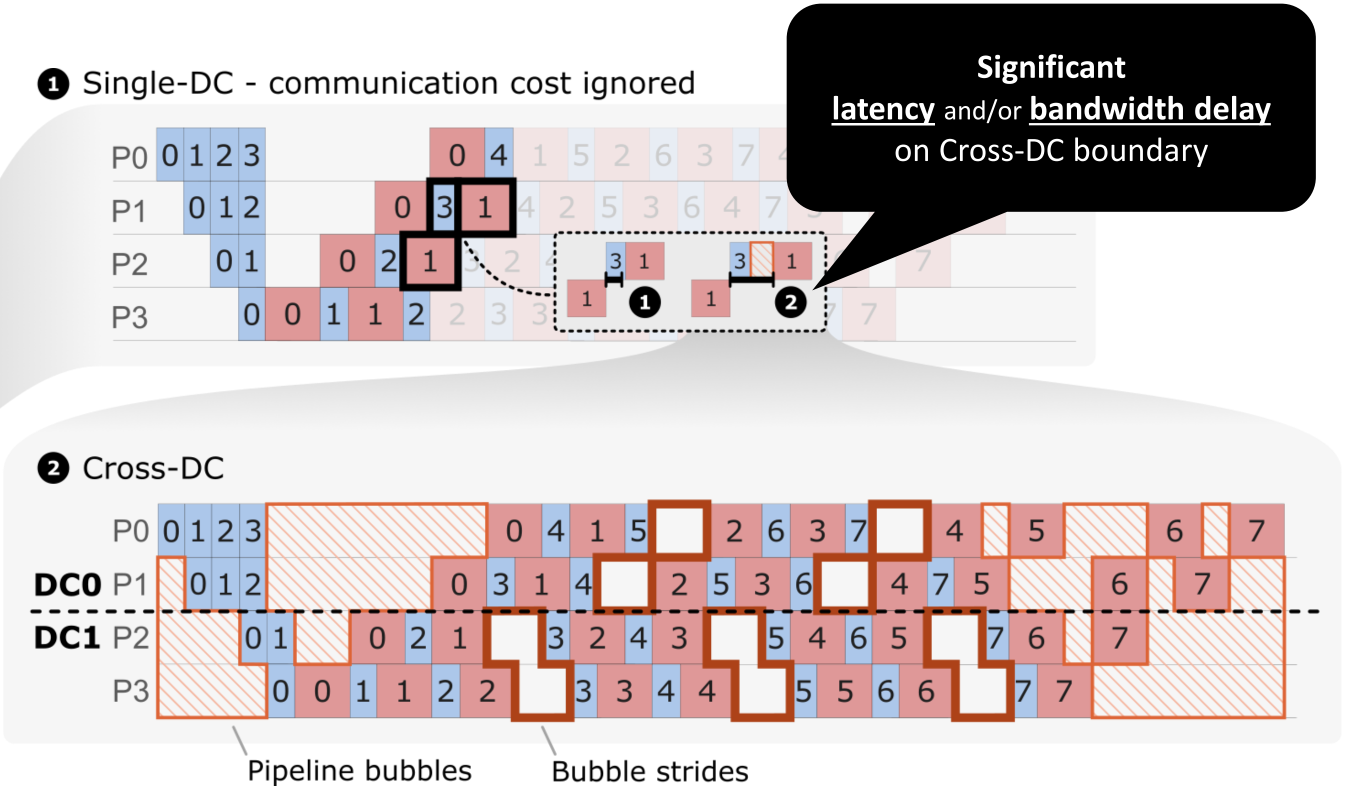


True dependencies at **varying** granularity

[2] Qi, Penghui, et al. Zero bubble pipeline parallelism (2023)

Pipeline Parallelism – Schedules





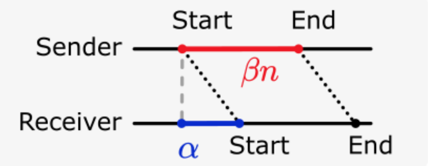
What can we do about the bubble strides?

CrossPipe – Optimal Schedule

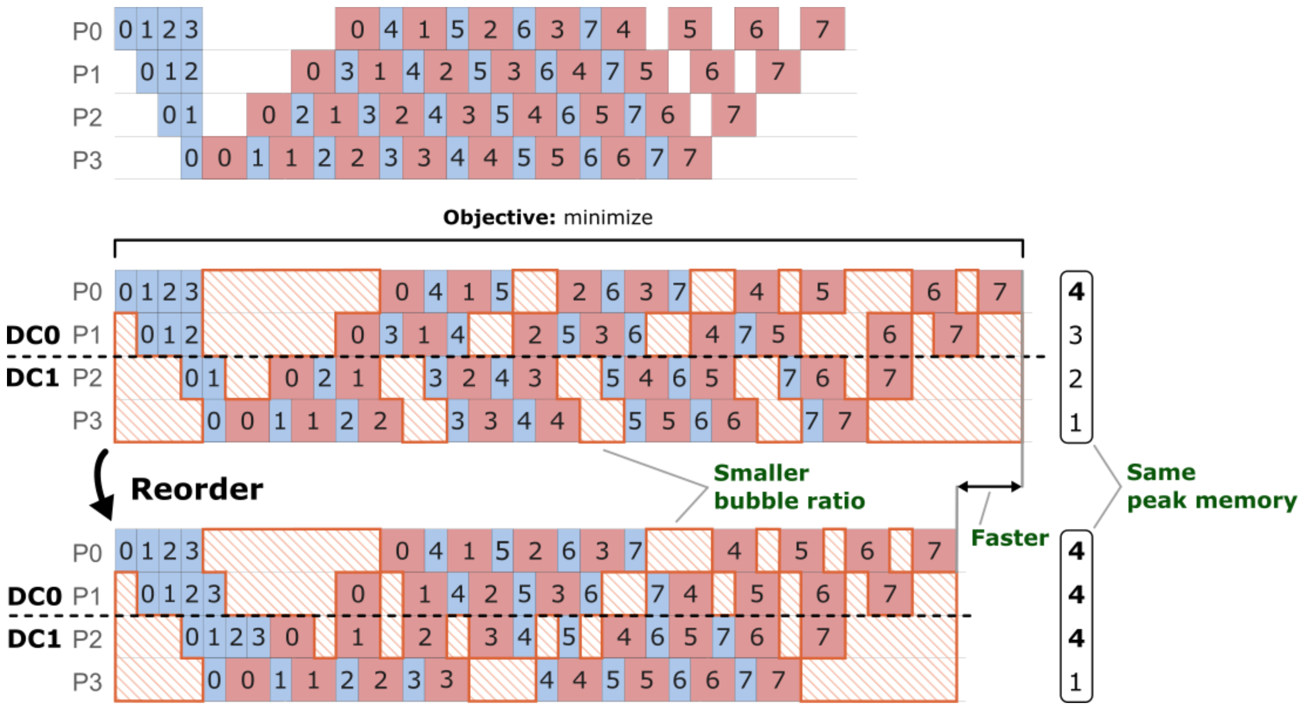
- **Frame:** job scheduling problem
- Include communication operations
- Modelling communication time: **α-β model**

$$T_{comm} = \alpha + \beta n$$

Latency · Inverse of Bandwidth · Message Size



- **Input:**
  - Duration of operations (compute and **communication**)
  - Latency (**communication**)
  - Memory changes (compute) and device limit
  - Data dependencies
- **Solver: Constrained Optimization**



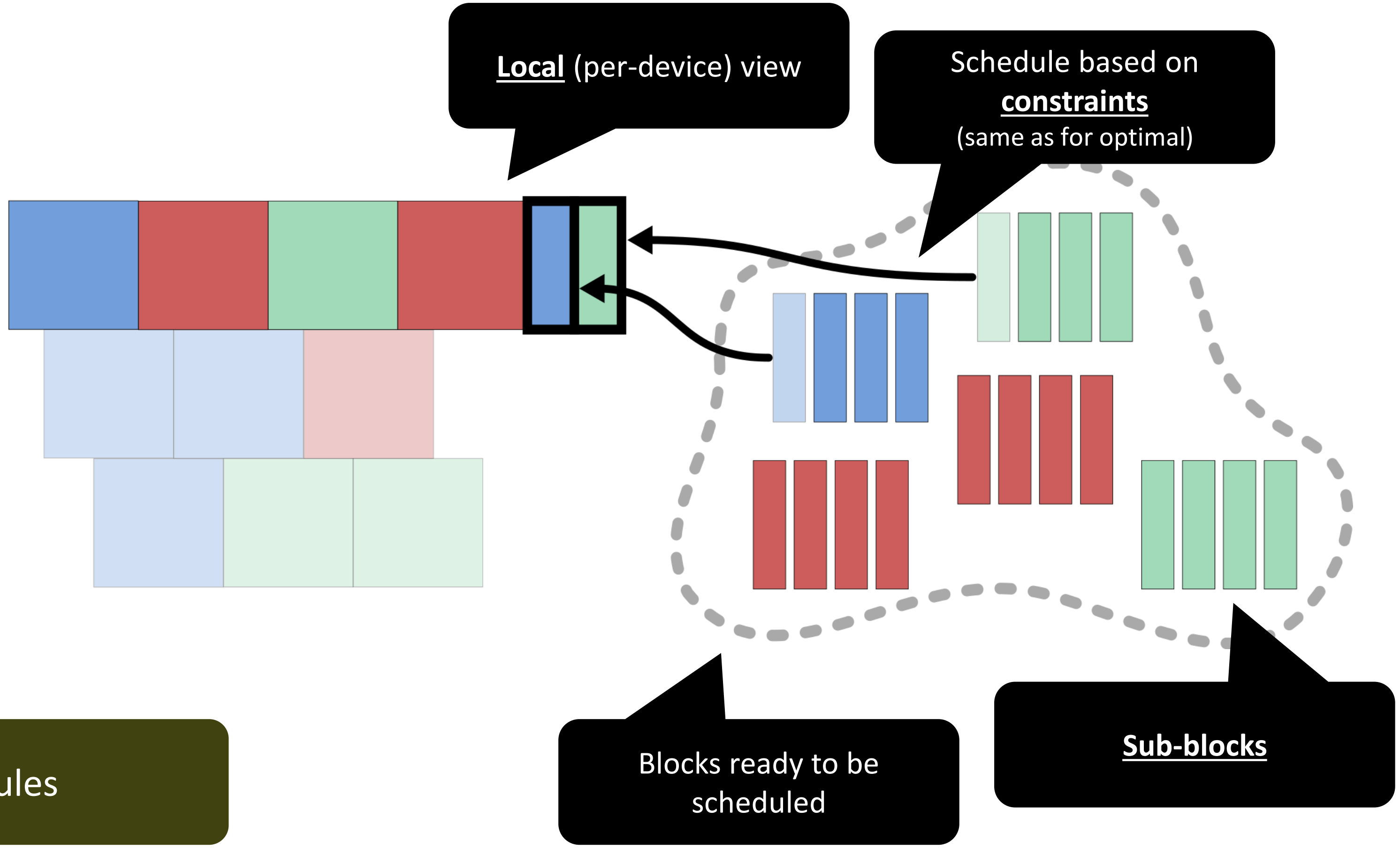
Limitation: solver runtime

CrossPipe – Greedy Schedule

- Dynamic conditions (schedule adaptation)
- Solver runtime ⇨ limiting

- **Algorithmic approach**
  ⇨ sub-block scheduling

- Faster schedule generation
- Same constraints
- Near-optimal



Attractive **alternative** to optimal schedules

Cross-DC DP vs. Cross-DC PP

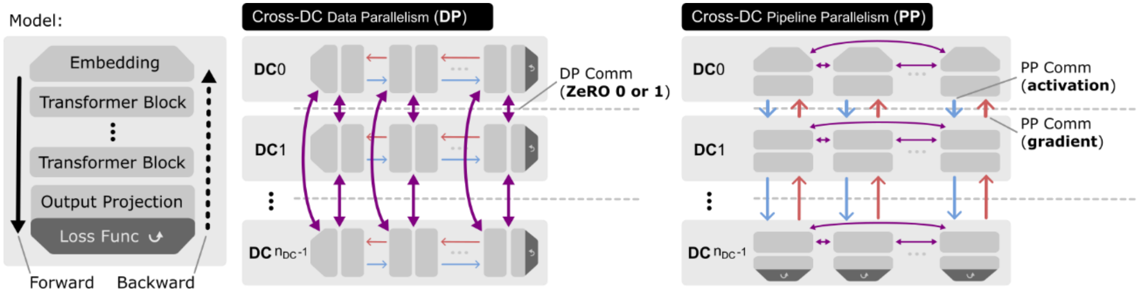
- **Performance model**
- **Model:** Llama 3 405B
  - **2 DC**
  - $n_{TP} = 8$ $n_{PP} = 16$ $n_{DP} = 64$
  - T_F = 109 ms

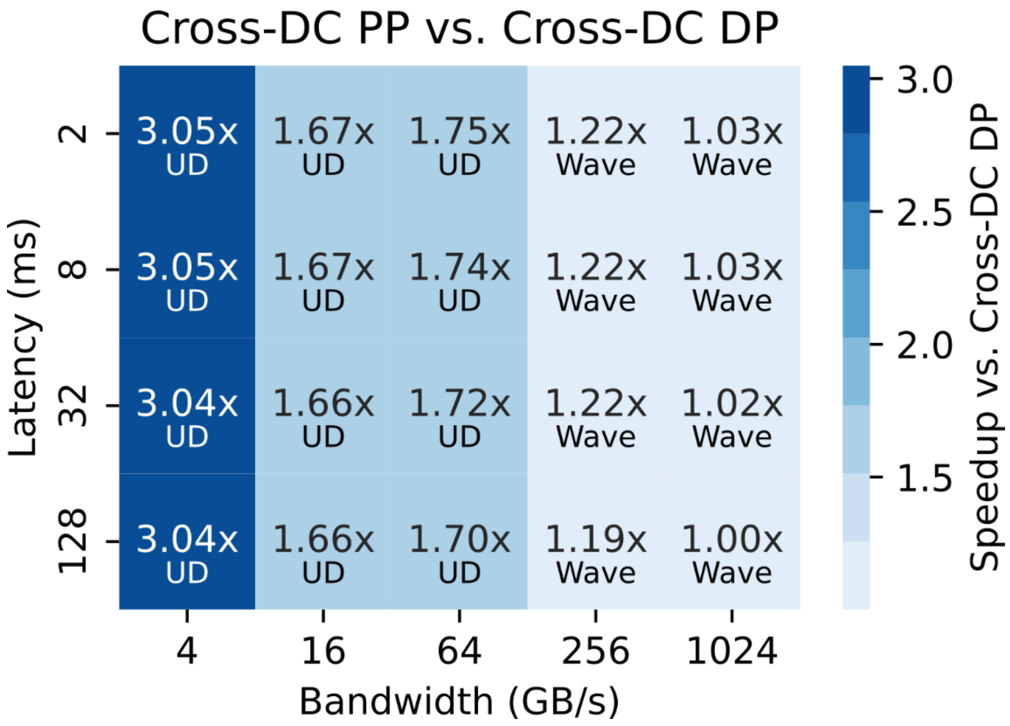
Based on Llama 3 technical report

Time of forward (**F**) **block**

- **PP schedules:**
  - Cross-DP: ZBV
  - Cross-PP: CrossPipe (UD or Wave)

Traversal pattern (best)



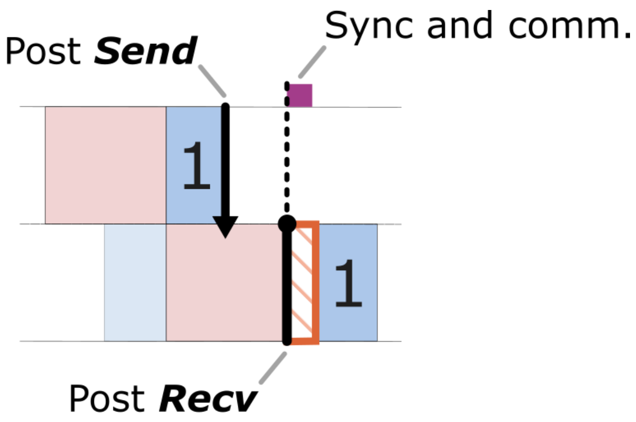


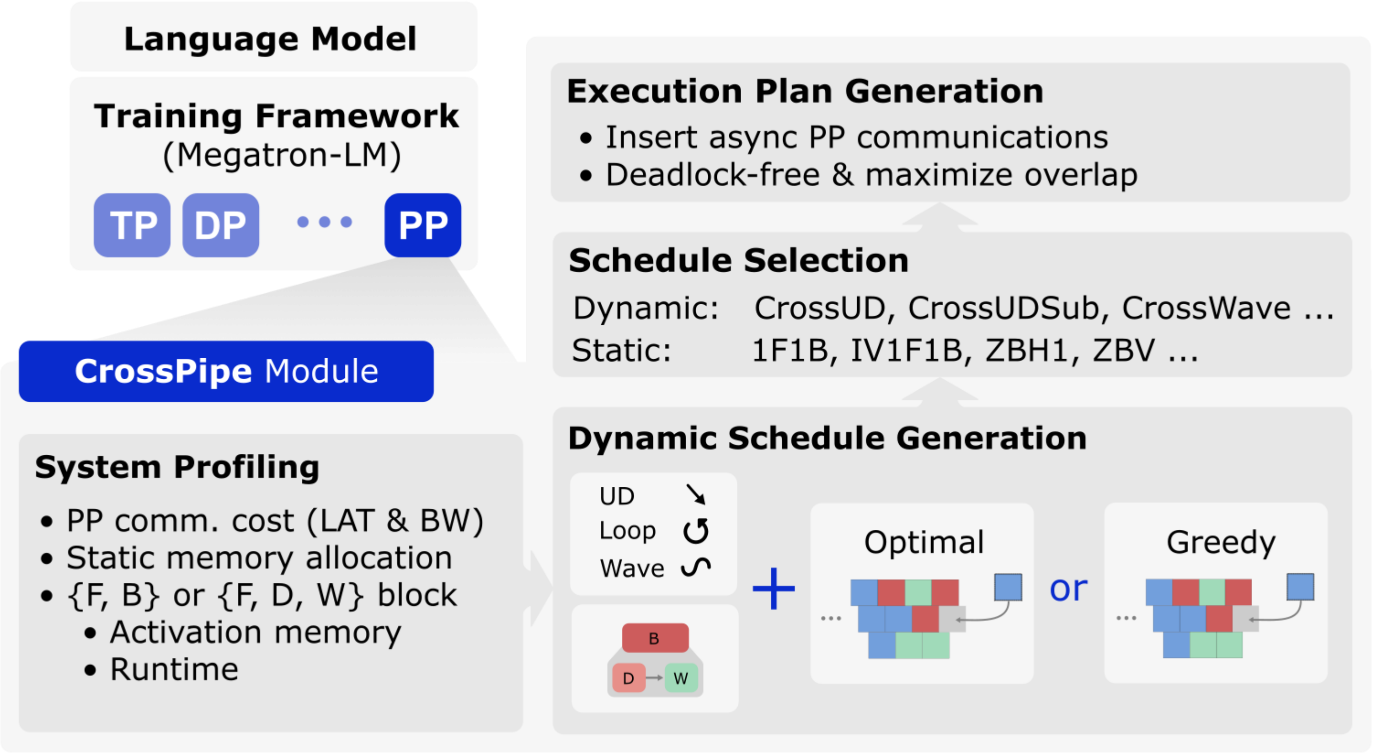
Cross-DC – **Pipeline Parallelism** → better choice

CrossPipe – Implementation

- **Optimization**
- Post *Recv* ahead of *Send*
  ⇨ Based on schedule – maximize overlap

Original (NCCL):

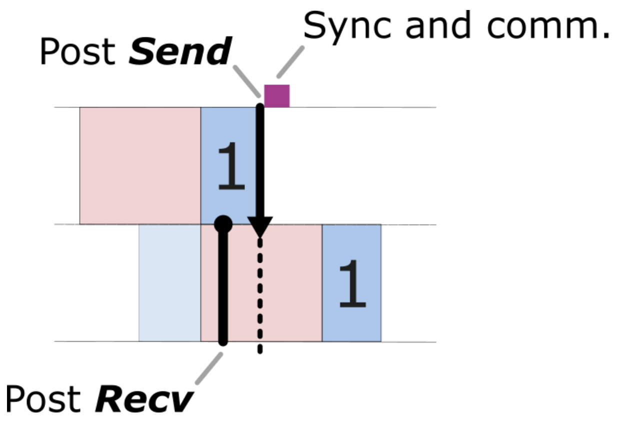


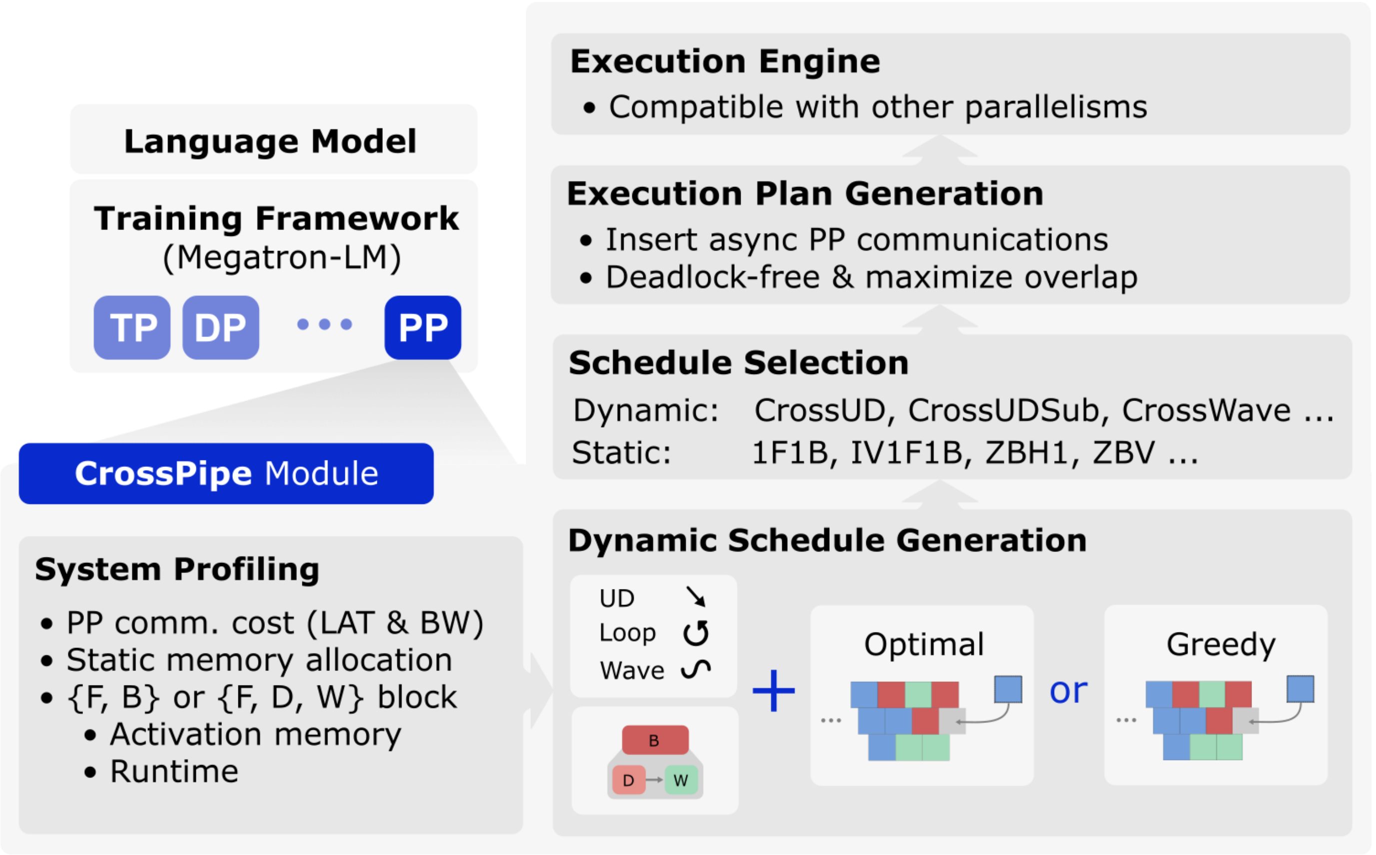


CrossPipe – Implementation

- **Optimization**
- Post *Recv* ahead of *Send*
  ⇨ Based on schedule – maximize overlap
- Both **static** and **dynamic** schedules

Improved (NCCL):





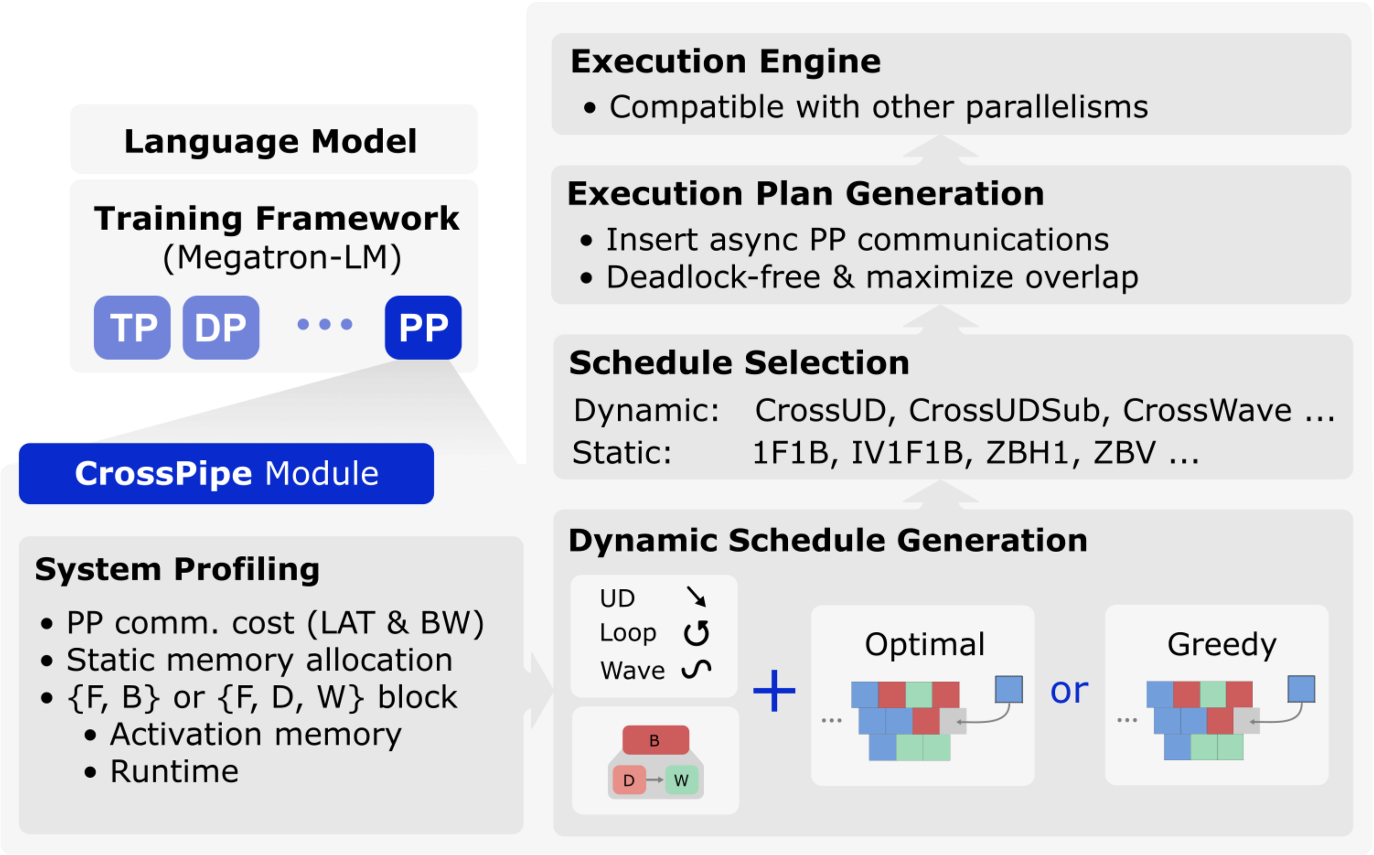
CrossPipe – Implementation

- **Optimization**
- Post *Recv* ahead of *Send*
  ⇨ Based on schedule – maximize overlap
- Both **static** and **dynamic** schedules

Adaptable (Single-DC / Cross-DC)

Extensible (configure / abstraction)

Efficient (optimal / near-optimal)

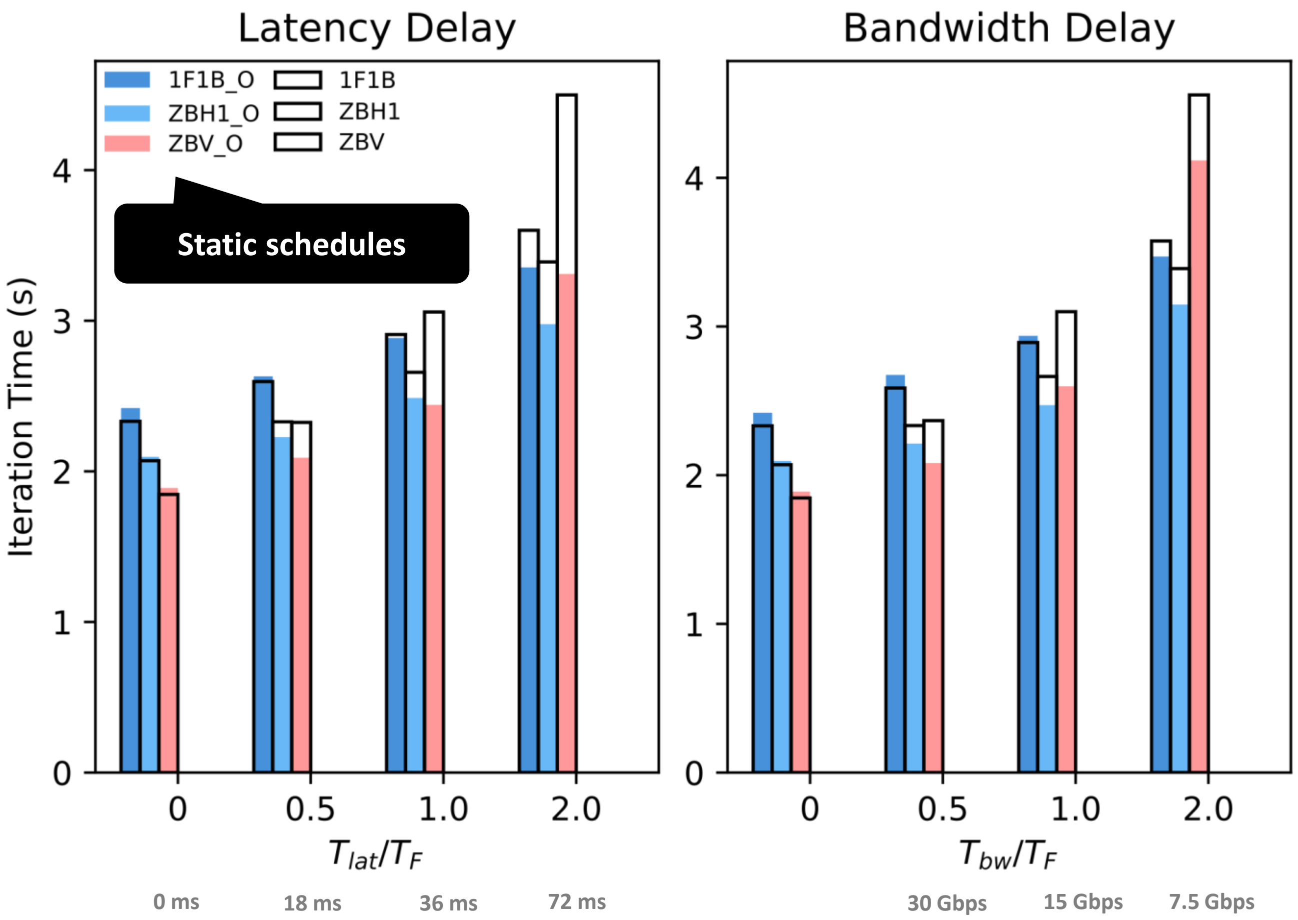


Evaluation

- **Single-DC system:** Alps
  - Node: 4x GH200 (96 GB HBM3), NVLink 4.0
  - HPE Cray Slingshot-11 interconnect
    HPE Cray Cassini-1 200 Gb/s NIC | Dragonfly

- Cross-DC **latency** and **bandwidth delay**
  - **Emulated** in single-DC environment
  - PyTorch – injected spin kernels (validated)

- **Model:** M70 (Llama 70B)
  - **2 DC**
  - $n_{TP} = 4 \quad n_{PP} = 8 \quad n_{DP} = 1$
  - T_F = 36 ms



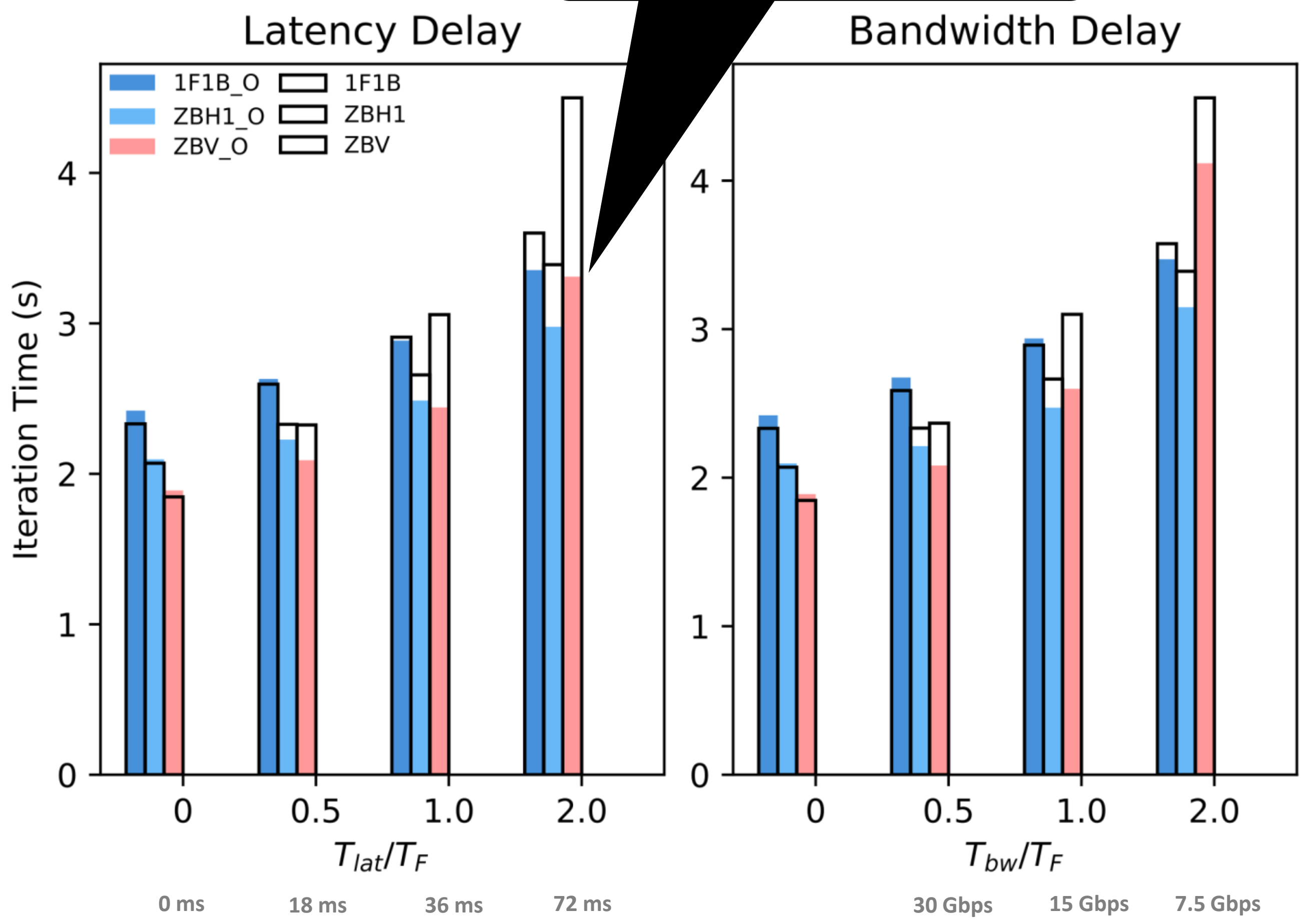
Evaluation

- **Single-DC system:** Alps
  - Node: 4x GH200 (96 GB HBM3), NVLink 4.0
  - HPE Cray Slingshot-11 interconnect
    HPE Cray Cassini-1 200 Gb/s NIC | Dragonfly

- Cross-DC **latency** and **bandwidth delay**
  - **Emulated** in single-DC environment
  - PyTorch – injected spin kernels (validated)

- **Model:** M70 (Llama 70B)
  - **2 DC**
  - $n_{TP} = 4$ $n_{PP} = 8$ $n_{DP} = 1$
  - T_F = 36 ms

_O Communication orchestration
→ **significant improvement**

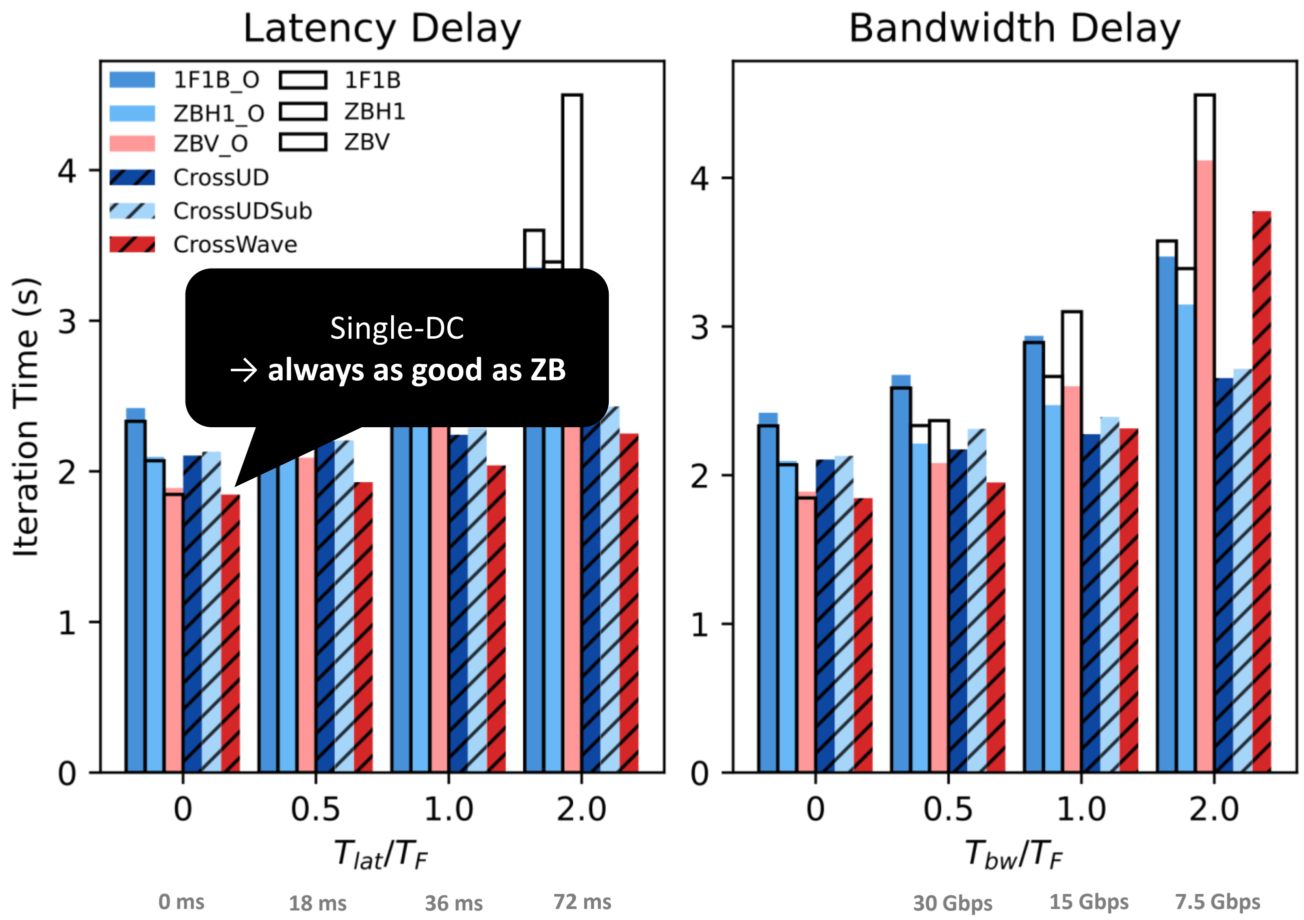


Evaluation

- **Single-DC system:** Alps
  - Node: 4x GH200 (96 GB HBM3), NVLink 4.0
  - HPE Cray Slingshot-11 interconnect
    HPE Cray Cassini-1 200 Gb/s NIC | Dragonfly

- Cross-DC **latency** and **bandwidth delay**
  - **Emulated** in single-DC environment
  - PyTorch – injected spin kernels (validated)

- **Model:** M70 (Llama 70B)
  - **2 DC**
  - $n_{TP} = 4$ $n_{PP} = 8$ $n_{DP} = 1$
  - T_F = 36 ms

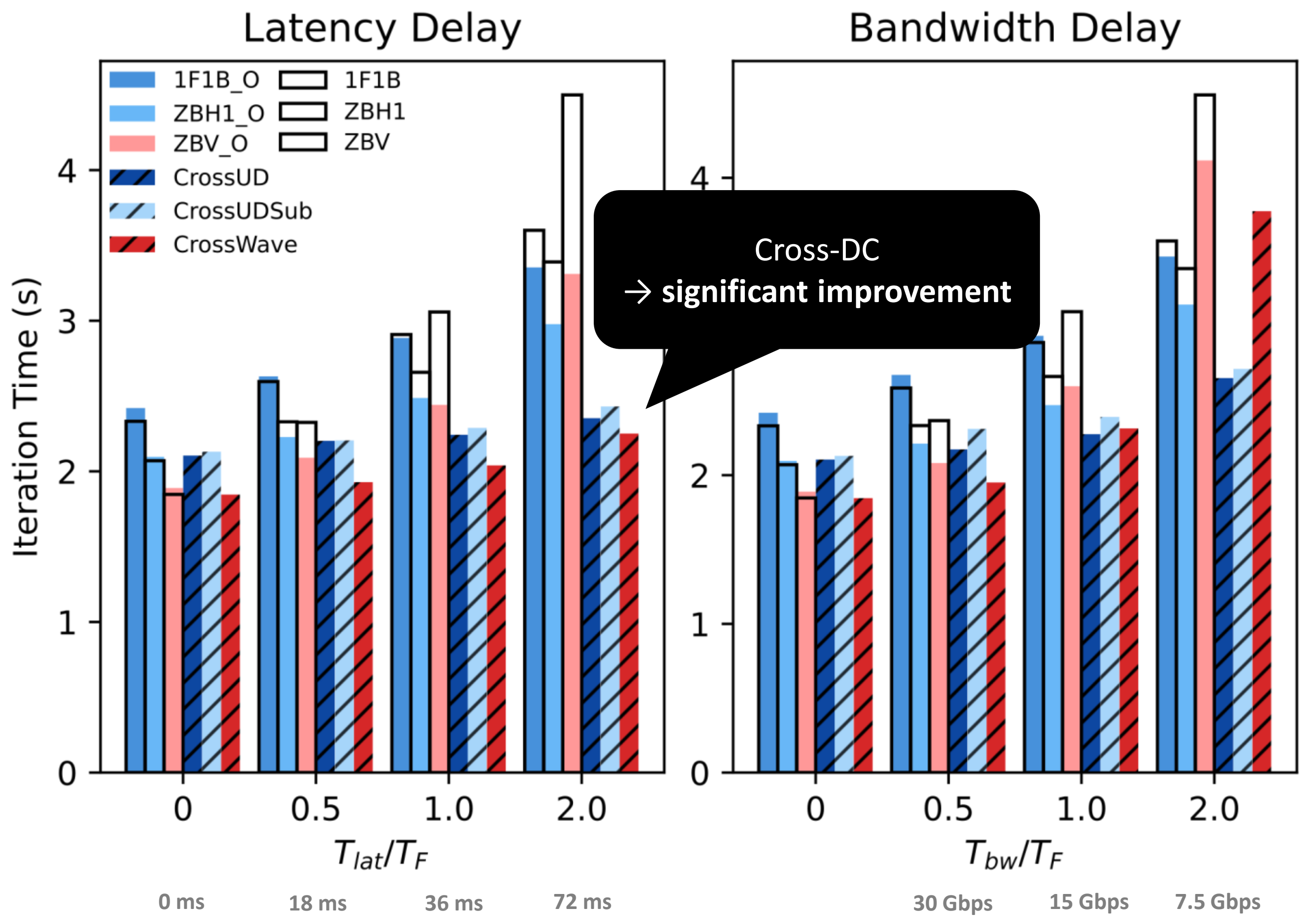


Evaluation

- **Single-DC system:** Alps
  - Node: 4x GH200 (96 GB HBM3), NVLink 4.0
  - HPE Cray Slingshot-11 interconnect
    HPE Cray Cassini-1 200 Gb/s NIC | Dragonfly

- Cross-DC **latency** and **bandwidth delay**
  - **Emulated** in single-DC environment
  - PyTorch – injected spin kernels (validated)

- **Model:** M70 (Llama 70B)
  - **2 DC**
  - $n_{TP} = 4 \quad n_{PP} = 8 \quad n_{DP} = 1$
  - T_F = 36 ms

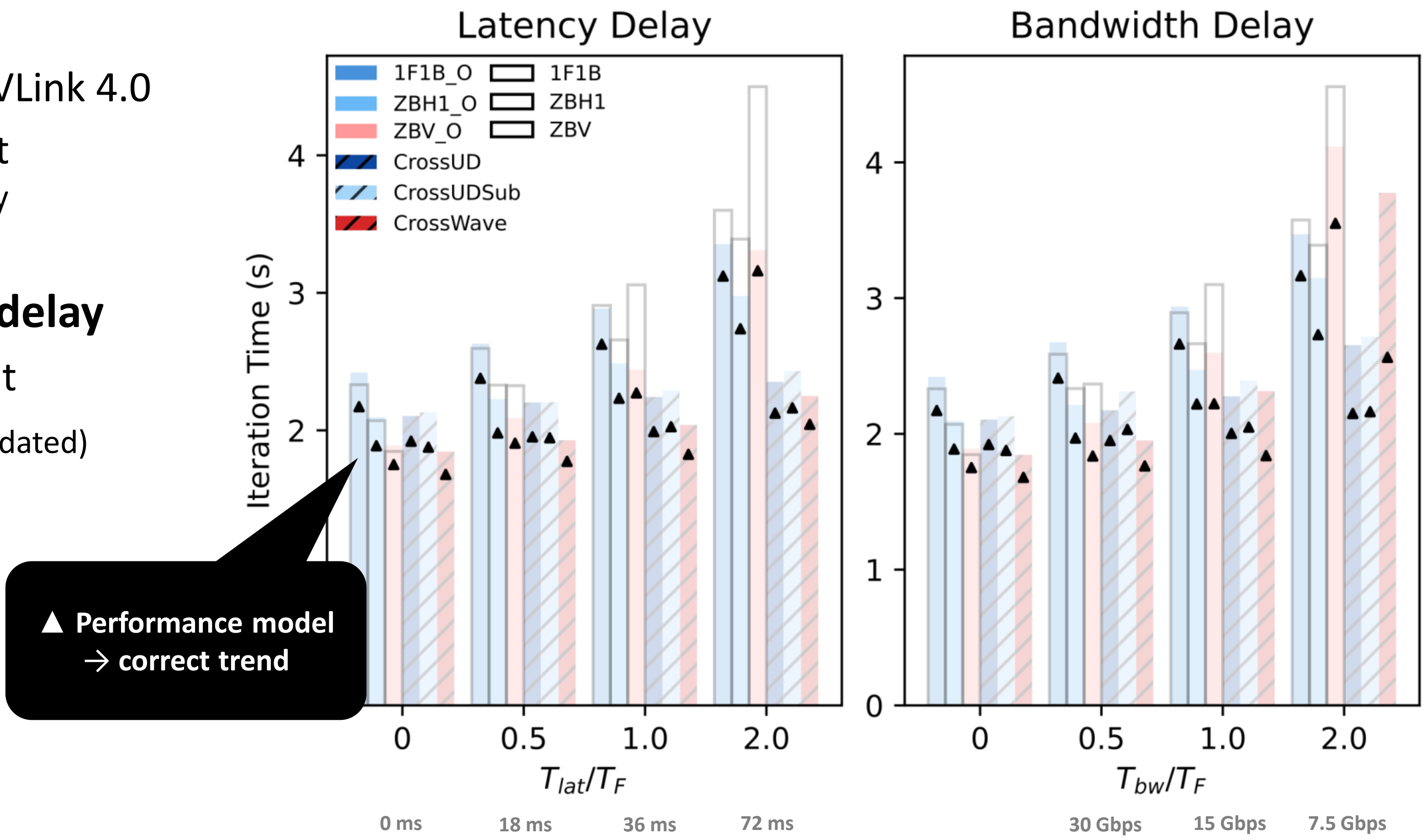


Evaluation

- **Single-DC system:** Alps
  - Node: 4x GH200 (96 GB HBM3), NVLink 4.0
  - HPE Cray Slingshot-11 interconnect
    HPE Cray Cassini-1 200 Gb/s NIC | Dragonfly

- Cross-DC **latency** and **bandwidth delay**
  - **Emulated** in single-DC environment
  - PyTorch – injected spin kernels (validated)

- **Model:** M70 (Llama 70B)
  - **2 DC**
  - $n_{TP} = 4$ $n_{PP} = 8$ $n_{DP} = 1$
  - T_F = 36 ms



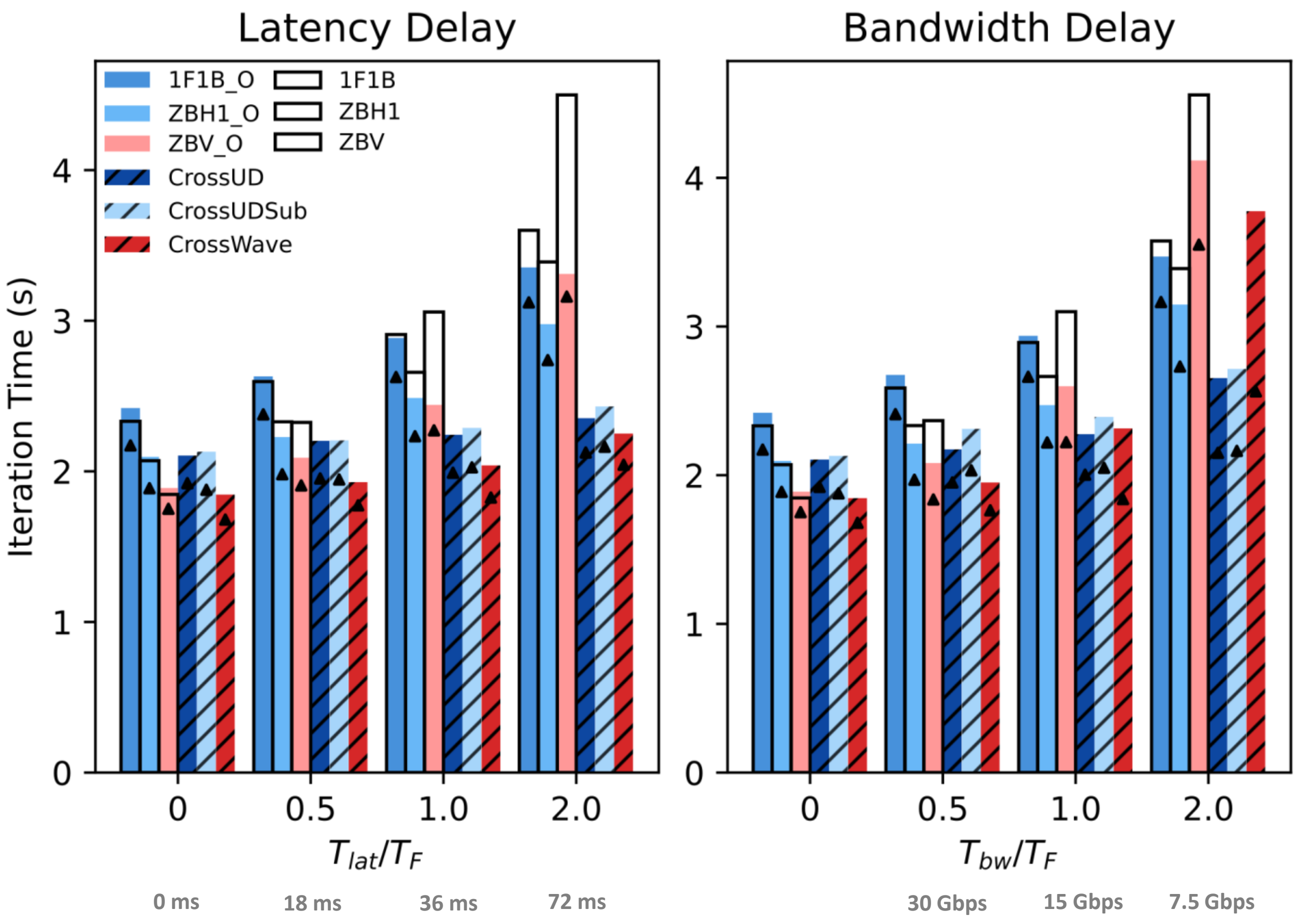
Evaluation

- **Single-DC system:** Alps
  - Node: 4x GH200 (96 GB HBM3), NVLink 4.0
  - HPE Cray Slingshot-11 interconnect
    HPE Cray Cassini-1 200 Gb/s NIC | Dragonfly

- Cross-DC **latency** and **bandwidth delay**
  - **Emulated** in single-DC environment
  - PyTorch – injected spin kernels (validated)

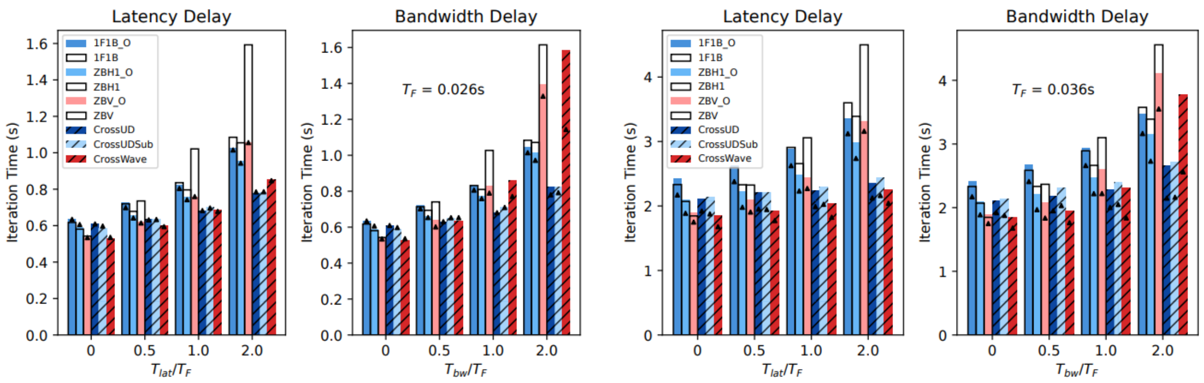
- **Model:** M70 (Llama 70B)
  - **2 DC**
  - $n_{TP} = 4 \quad n_{PP} = 8 \quad n_{DP} = 1$
  - T_F = 36 ms

Up to **33.6%** training time reduction
(identical memory constraints)



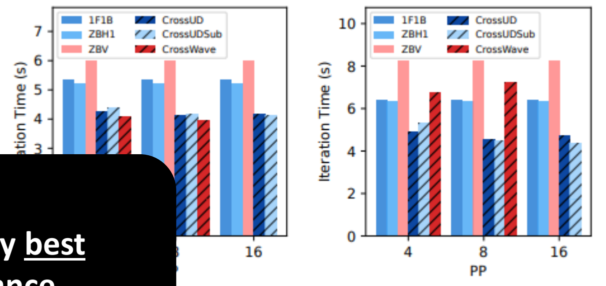
Evaluation

- Multiple Models: **M8** and **M70**
- Varying number of DCs: **2** and **4**
- Increased GBS
- Increased memory budget
- Activation recompute
- PP and DP tradeoff
- Mix latency + bandwidth delay



(a) Model M8 (b) Model M70

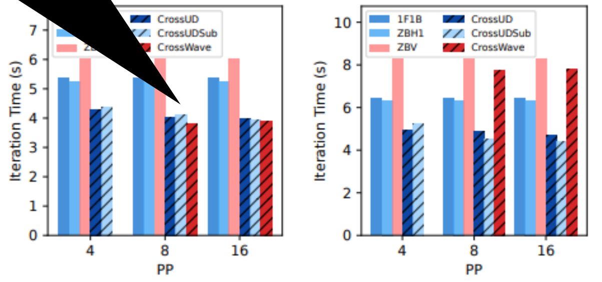
$\frac{T_{lat}}{T_F}$	$\frac{T_{bw}}{T_F}$	Case	Static 1F1B	Static ZBH1	Static ZBV	Dynamic (This Work) UDSub	Dynamic (This Work) UD	Dynamic (This Work) Wave
0	0	1	0.151	0.133	**0.118**	0.137	0.137	0.119
		2	0.174	0.168	0.161	-	0.165	**0.157**
		3	-	-	-	0.121	0.118	**0.108**
0.25	0.25	1	0.168	0.15	0.148	0.149	0.142	**0.127**
		2	0.193	0.187	0.177	-	0.17	**0.159**
		3	-	-	-	[illegible]	[illegible]	[illegible]
0.25	2	1	0.241	0.23	0.315	[illegible]	[illegible]	[illegible]
		2	0.262	0.259	0.33	[illegible]	[illegible]	[illegible]
		3	-	-	-	[illegible]	[illegible]	[illegible]
2	0.25	1	0.242	0.229	0.314	[illegible]	[illegible]	[illegible]
		2	0.262	0.258	0.329	[illegible]	[illegible]	[illegible]
		3	-	-	-	[illegible]	[illegible]	[illegible]
2	2	1	0.321	0.309	0.473	[illegible]	[illegible]	[illegible]
		2	0.333	0.331	0.476	[illegible]	[illegible]	[illegible]
		3	-	-	-	[illegible]	[illegible]	[illegible]



$w/T_F = 0.25$; (b) $T_{lat}/T_F = 0.125, T_{bw}/T_F = 2$;
$PP = 4$): 25% Static Mem. ($n_{PP} = 4$): 25%

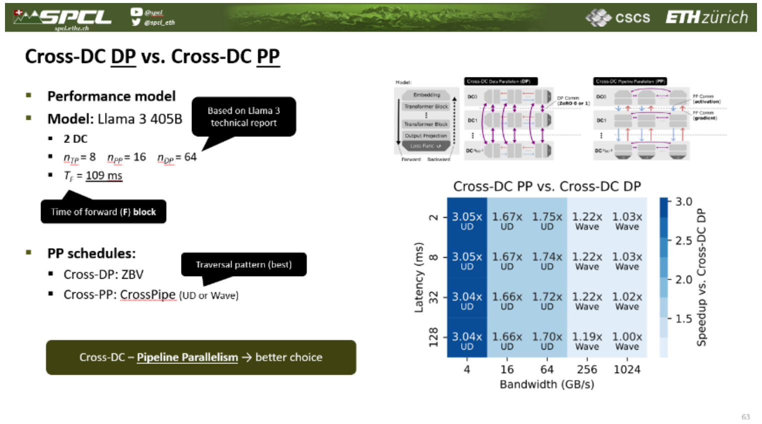
Consistently best performance

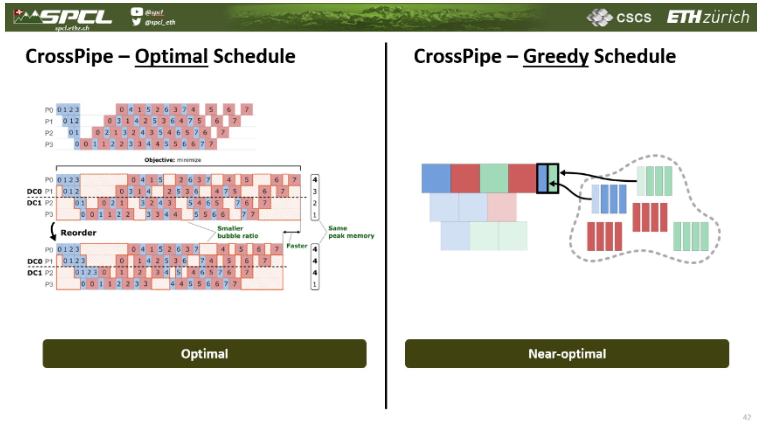
$\frac{T_{lat}}{T_F}$	$\frac{T_{bw}}{T_F}$	Case	Static 1F1B	Static ZBH1	Static ZBV	Dynamic (This Work) UDSub	Dynamic (This Work) UD	Dynamic (This Work) Wave
0	0	1	0.149	0.133	**0.119**	0.138	0.138	0.122
		2	0.173	0.168	0.16	-	0.167	**0.157**
		3	-	-	-	0.123	0.119	**0.115**
0.25	0.25	1	0.177	0.158	0.161	0.155	0.148	**0.141**
		2	0.198	0.19	0.181	-	0.173	**0.163**
		3	-	-	-	0.126	0.123	**0.115**
0.25	2	1	0.269	0.249	0.339	**0.216**	0.217	0.286
		2	0.274	0.269	0.331	-	**0.198**	0.29
		3	-	-	-	**0.158**	0.162	0.262
2	0.25	1	0.268	0.248	0.337	0.2	**0.197**	0.214
		2	0.274	0.269	0.33	-	0.184	**0.177**
		3	-	-	-	**0.138**	**0.138**	0.139
2	2	1	0.359	0.338	0.512	**0.268**	0.271	0.339
		2	0.349	0.346	0.479	-	**0.213**	0.295
		3	-	-	-	**0.178**	**0.178**	0.264

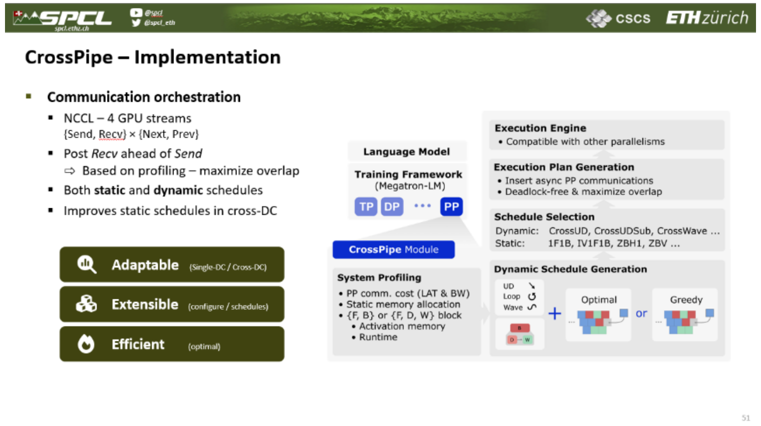


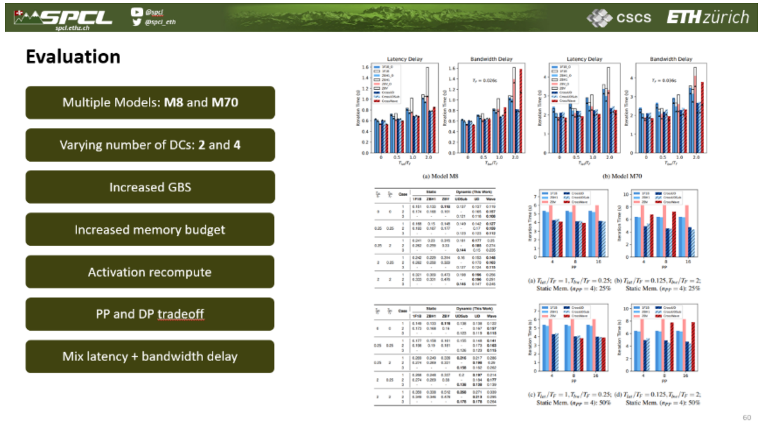
(c) $T_{lat}/T_F = 1, T_{bw}/T_F = 0.25$; (d) $T_{lat}/T_F = 0.125, T_{bw}/T_F = 2$;
Static Mem. ($n_{PP} = 4$): 50% Static Mem. ($n_{PP} = 4$): 50%

Conclusions









More of SPCL's research:

youtube.com/@spcl — 210+ Talks

twitter.com/spcl_eth — 1.6K+ Followers

github.com/spcl — 5.6K+ Stars

… or spcl.ethz.ch

Open source:

github.com/spcl/crosspipe

ARTIFACT EVALUATED usenix ASSOCIATION AVAILABLE

ARTIFACT EVALUATED usenix ASSOCIATION FUNCTIONAL

ARTIFACT EVALUATED usenix ASSOCIATION REPRODUCED