Interconnection Networks in the Exascale and AI Era: Challenges and Solutions

Pedro J. García
University of Castilla-La Mancha (UCLM)
Spain
pedrojavier.garcia@uclm.es

Jesús Escudero Sahuquillo
University of Castilla-La Mancha (UCLM)
Spain
jesus.escudero@uclm.es

Academic Salon on High-Performance Ethernet, TUM, Munich, 12-13 March 2025

Agenda

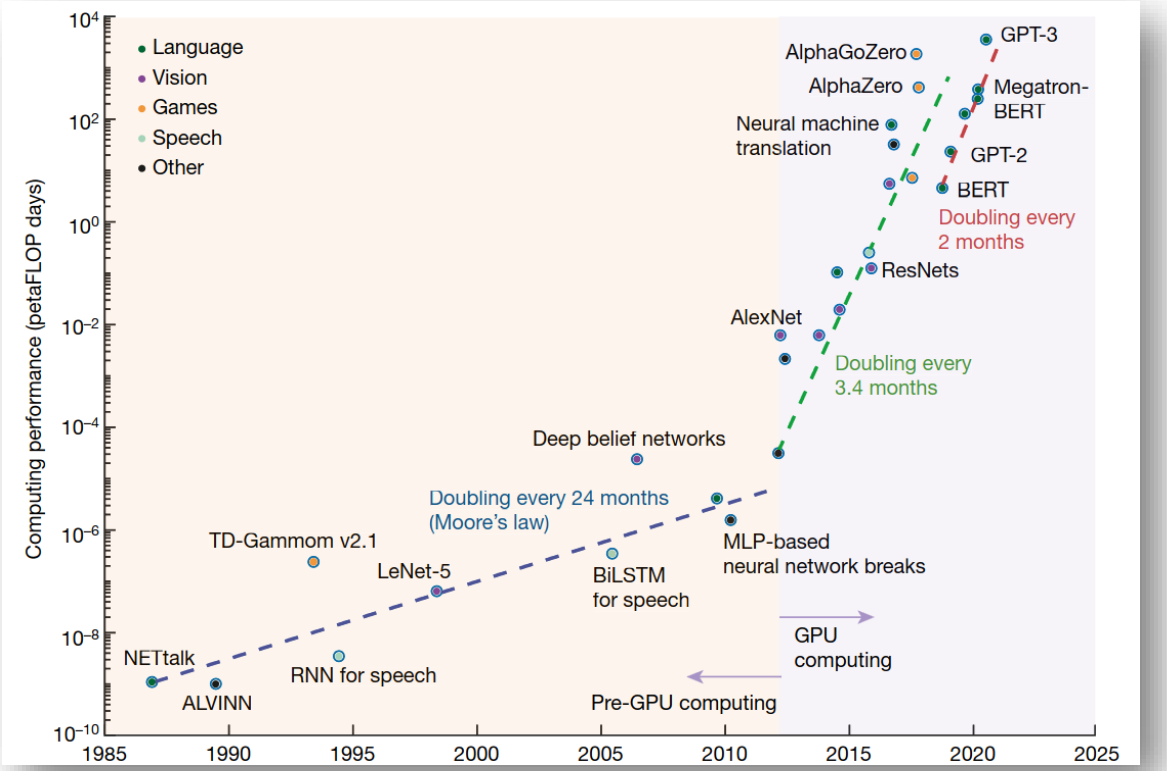
- Introduction
- Current (well, Classical too…) Challenges in Interconnects
- Solutions (well, at least, some solutions…)
- Open issues

Agenda

- **Introduction**
- Current (well, Classical too…) Challenges in Interconnects
- Solutions (well, at least, some solutions…)
- Open issues

Introduction

- **High-performance interconnection networks** are a key subsystem in HPC systems and Datacenters
- **Emerging applications** impact the network requirements:
  - ML/AI models: LLMs, GPT-4, etc.
  - Scientific computing
  - Real-time user demands (e.g. AI inference)
  - **Post exascale computing**
- … and pose new challenges in the **inter- and intra-node communication**

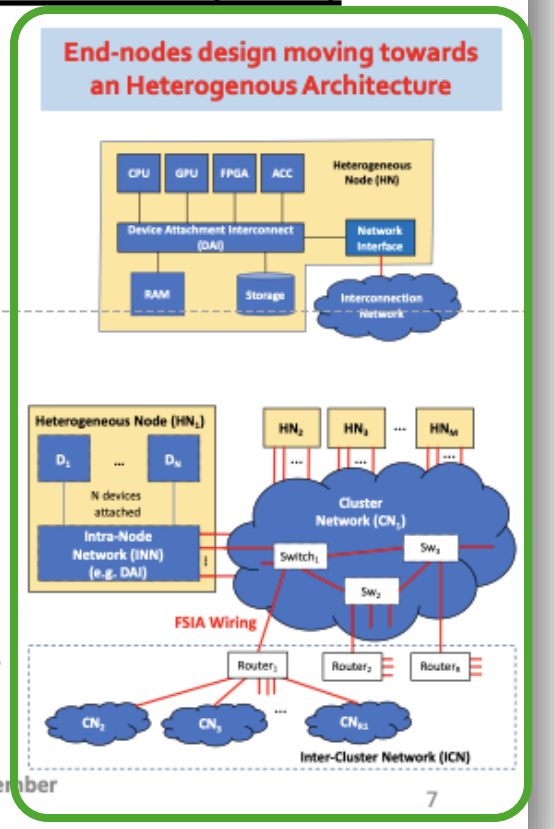


Mehonic, A., Kenyon, A.J. Brain-inspired computing needs a master plan. *Nature* **604**, 255–260 (2022). https://doi.org/10.1038/s41586-021-04362-w

Introduction

The importance of the Datacenter Network (DCN)

- **End-nodes demanding more computing power at a reduced energy cost.**
- **Increase in node heterogeneity**: accelerators, CPUs, storage and network picking memory.
- HPC clusters, DCNs and Hyperscalers converging to the same network infrastructure (Ethernet, RoCE, etc.).
- New enhancements for RDMA and NVMe are needed.
- According to several roadmaps, in the near future **network bandwidth will increase 4x, node power consumption will reduce 2x, and network power budget is expected to be lowered 10x.**
- **Challenges in interconnection networks R&D**:
  - Scalability when number of nodes increases (low-diameter topologies, efficient routing algorithms).
  - Data-movement (minimize) & Active storage (SCM and in-situ processing).
  - Power saving (link level) → Global system management.
  - **Network Congestion effects → Pervasive and mutant (constant research is required).**



2nd Huawei DCT Workshop. 18-19 November 2019. Nanjing, China

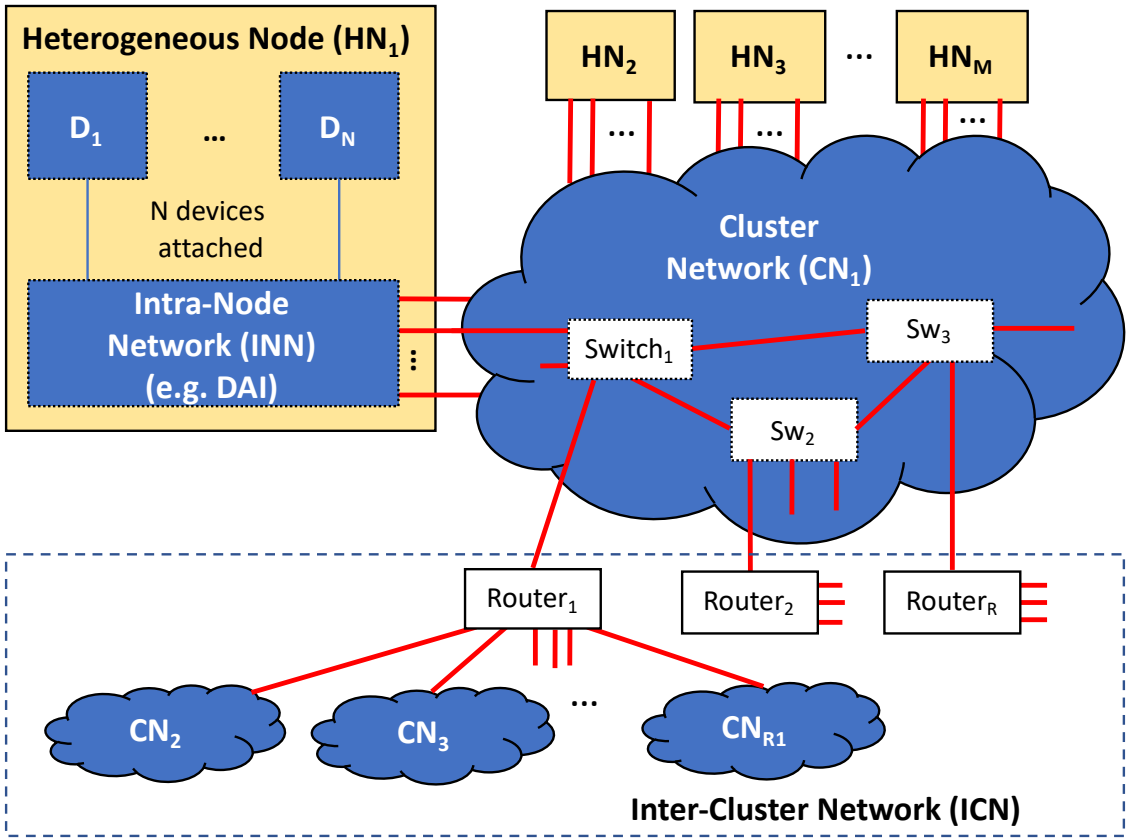
7

"intelligent, high-performance data center networks enabling both HPC and mega data center workloads will be adopted in the industry soon" T. Hoefler et al.: ***The Convergence of Hyperscale Data Center and High-Performance Computing Networks***, *in Computer, vol. 55, no. 7, pp. 29-37, July 2022, doi: 10.1109/MC.2022.3158437*

What Is a Hyper-Converged Data Center Network?
https://info.support.huawei.com/info-finder/encyclopedia/en/Hyper-Converged+Data+Center+Network.html
[Accessed in March 2025]

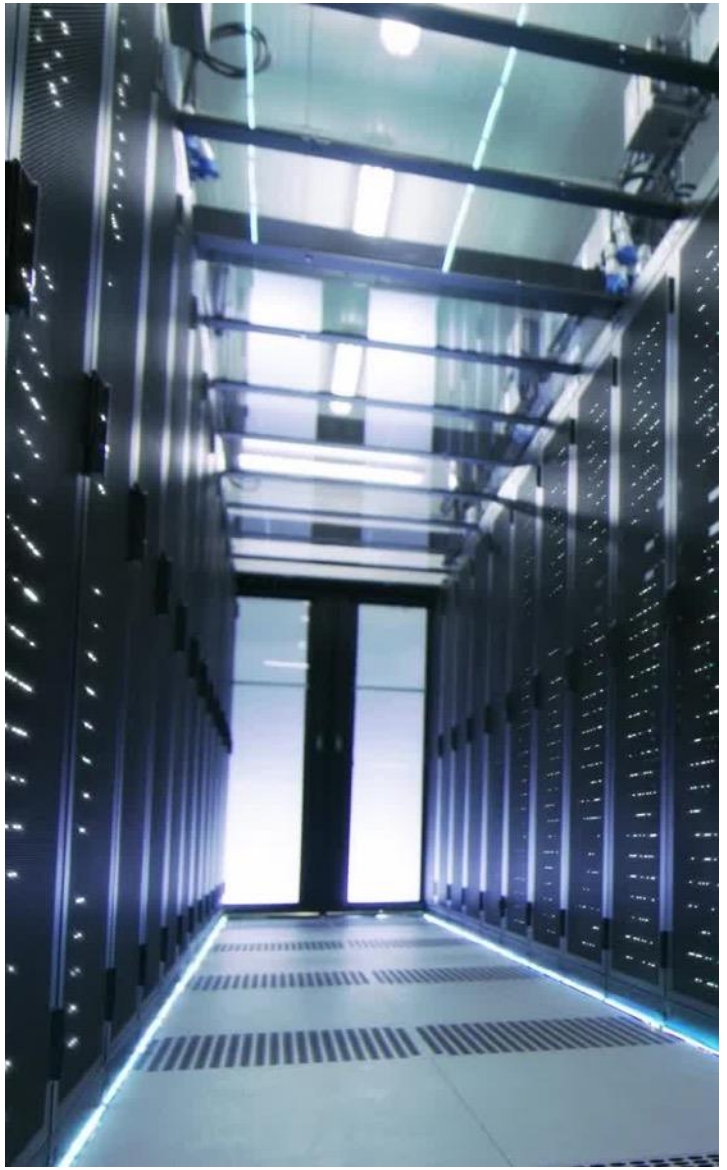
Introduction

- **Domain-specific accelerators** within heterogeneous computing nodes (HNs)
- Multiple HNs interconnected through a **cluster interconnection network**
- Clusters at different locations can be interconnected through an inter-cluster network
- **Scale-out** for the cluster and inter-cluster interconnection networks
- **Scale-up** for the intra-node network



Introduction

- But datacenters **may not be exclusively devoted to AI training**:
  - application mix with different communication requirements (e.g., multi-tenant systems, cloud services, users' demand on AI models, etc.)
- **Task-to-server allocation and collective communication** may not be fully optimized
- Most importantly, for 200K+ accelerators, **components will frequently fail**
- **What about power consumption?** It is not easy to keep the 20 MW in the Exascale/AI/Cloud era…



Agenda

- Introduction
- **Current (well, Classical too…) Challenges in Interconnects**
- Solutions (well, at least, some solutions…)
- Open issues

Challenges

- **“Hot topics”** in the interconnection network design (some of them **no longer inter-node issues but intra-node too**!) :
  - Augmenting link speed (800 Gbps and far beyond???)
  - Flow control (PFC vs credit-based)
  - Switch architecture (deep versus shallow buffering)
  - Network topologies (CLOS/fat-trees, 3D-mesh-again, Dragonfly+, etc.)
  - Routing algorithms (oblivious, deterministic, adaptive)
  - Flow steering (a.k.a. QoS or differentiated services)
  - Power management (power off certain links when not used)
  - Fault tolerance (failures will raise in such a big infrastructure)
  - Congestion control
  - …

Challenges

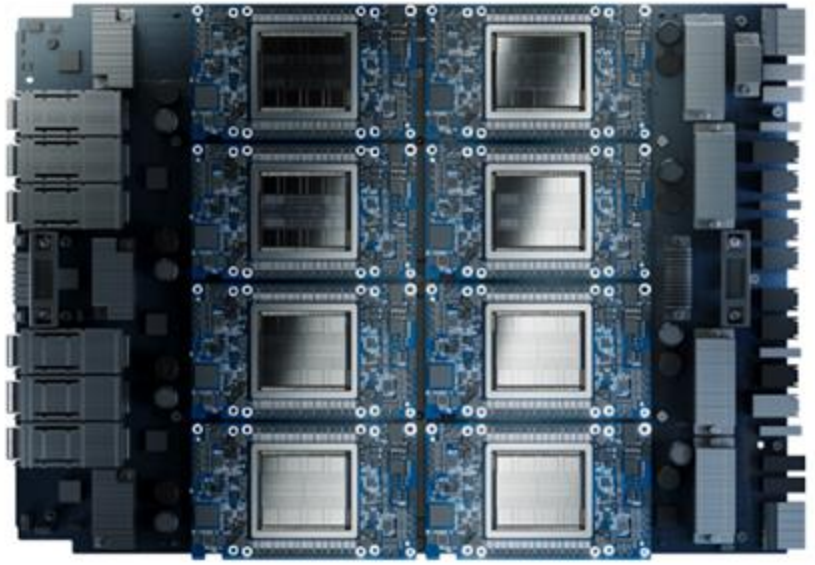
- **“Hot topics”** in the interconnection network design (some of them **no longer inter-node issues but intra-node too**!) :
  - Augmenting link speed (800 Gbps and far beyond???)
  - Flow control (PFC vs credit-based)
  - Switch architecture (deep versus shallow buffering)
  - **Network topologies (CLOS/fat-trees, 3D-mesh-again, Dragonfly+, etc.)**
  - **Routing algorithms (oblivious, deterministic, adaptive)**
  - Flow steering (a.k.a. QoS or differentiated services)
  - Power management (power off certain links when not used)
  - Fault tolerance (failures will raise in such a big infrastructure)
  - **Congestion control**
  - …

Agenda

- Introduction
- Current (well, Classical too…) Challenges in Interconnects
- **Solutions (well, at least, some solutions…)**
- Open issues

Intra-node networks

- Communicate hardware resources inside an end-node
- **Challenges**:
  - Ultra-low latency communication between CPUs, accelerators, memory, etc.
  - High bandwidth to support data-intensive workloads
  - Energy efficiency to ensure sustainable performance
- **Manufactures and technologies**:
  - PCIe
  - NVIDIA's NVLink
  - Intel Gaudi's board
  - AMD's Infinity Fabric
  - Intel's QPI
  - Other efforts (e.g., UALink consortium)



Intel Habana Gaudi-3

Intra-node networks

Technology details

- **PCI Express:**
  - **Most common** intra-node network
  - Speeds from 32 up to 1,024 Gbit/s
  - **Allow an optimal use of RDMA protocol** (such as GPUDirect-RDMA) with PCIe-Switches that does not need to communicate with CPU
- **NVIDIA NVLink:**
  - Used on **NVIDIA GPUs** to communicate between them or with specialized CPUs
  - Speeds from 2,400 to 14,400 Gbit/s
  - Combined with the **use of NVSwitch** is an important solution in HPC
  - It is more expensive than other solutions
- **Intel Gaudi's board:**
  - Used on **Intel's accelerators systems**
  - Last release uses connections of 600 Gbit/s
  - **Use RoCE** to communicate between GPUs inside same node and between nodes

Communication patterns

HPC/AI convergence demands parallelism

- **Large Language Models (LLMs)** do not fit on a single end node
- **Data Parallelism (DP):** Distribute the training workload across multiple devices by dividing the dataset
- **Model Parallelism (MP):** Distribute large-scale models across multiple devices
  - **Pipeline Parallelism (PP):** Splits the model into stages assigning each stage to a separate device
  - **Tensor Parallelism (TP):** Divide individual operations (e.g., matrix multiplications) across multiple devices

Communication patterns

Typical training loop in DNNs

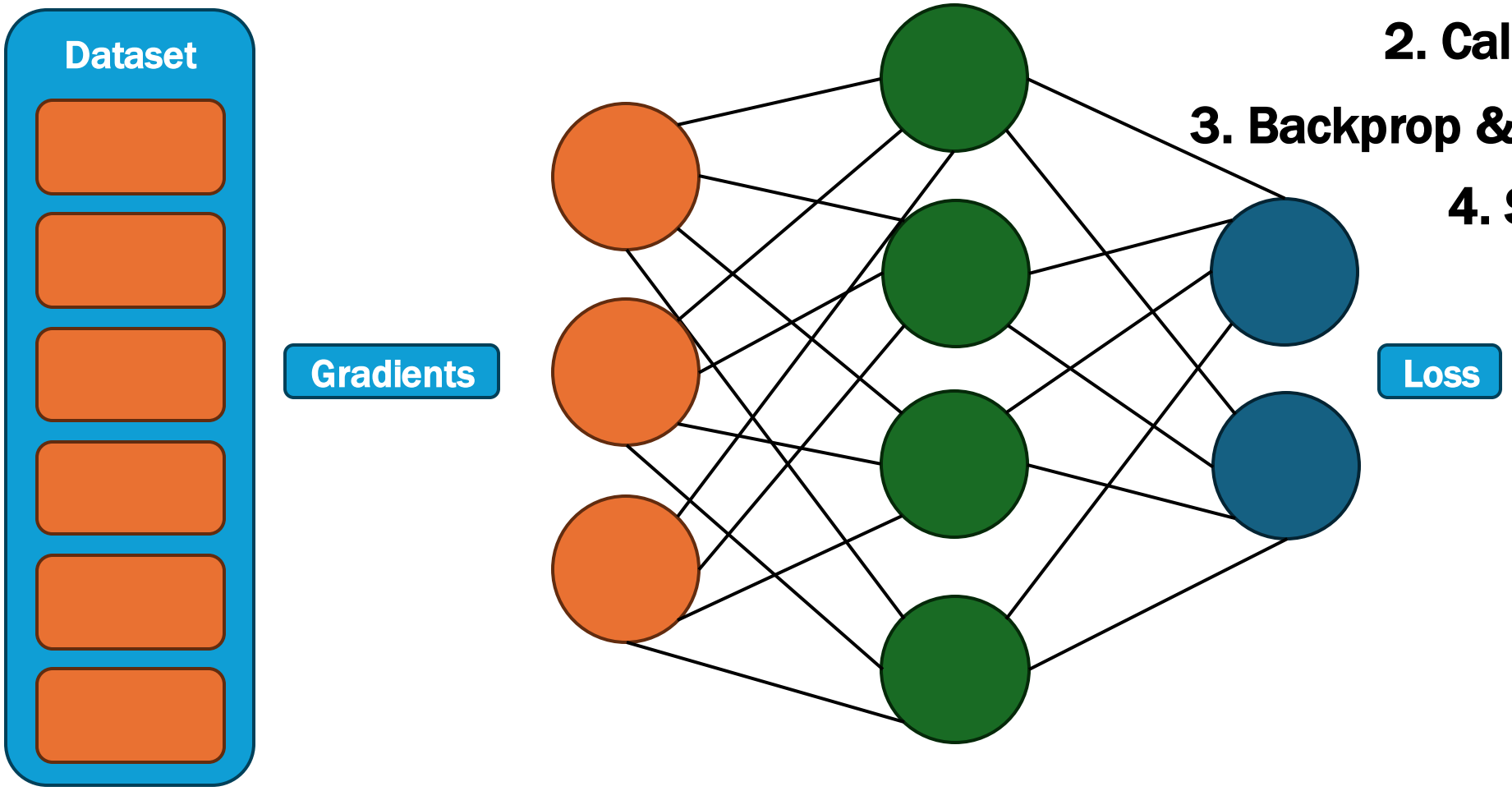
0. Split dataset into batches

1. Forward pass

2. Calculate loss

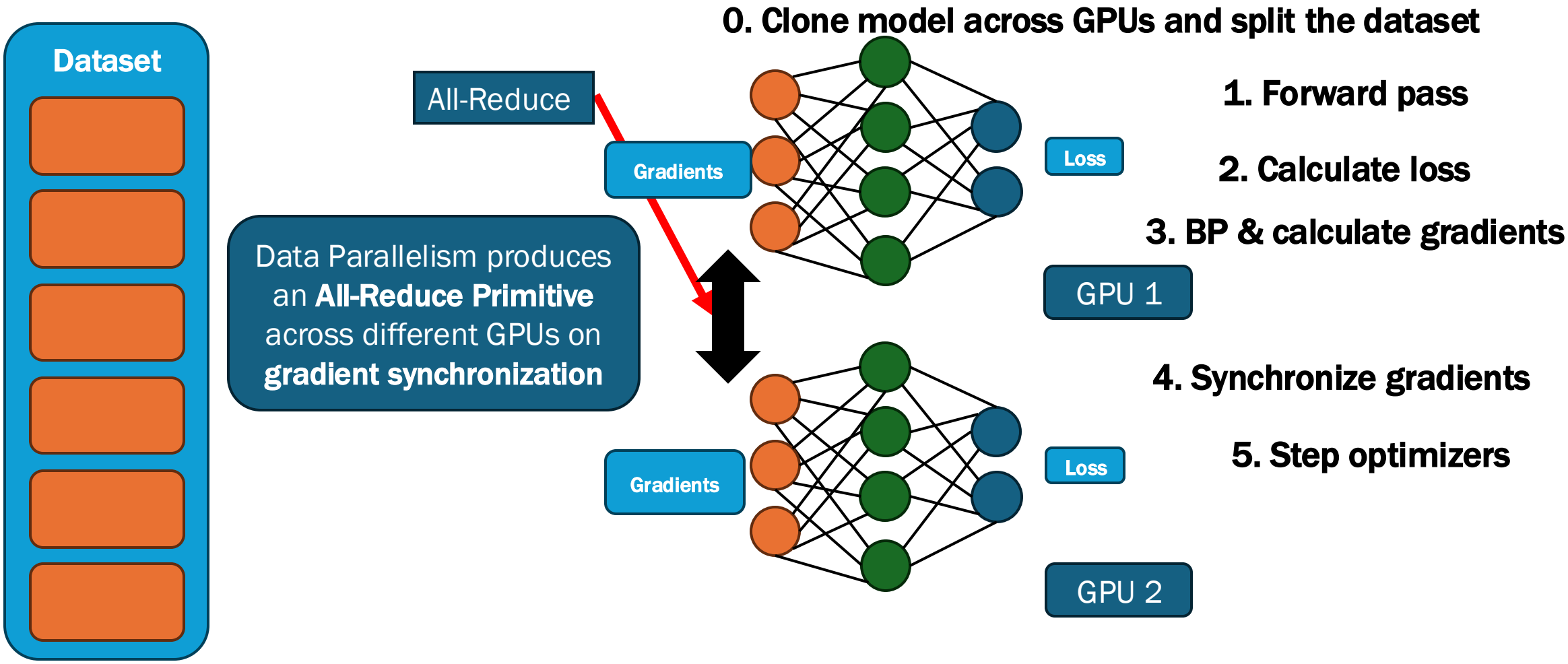
3. Backprop & calculate gradients

4. Step optimizers



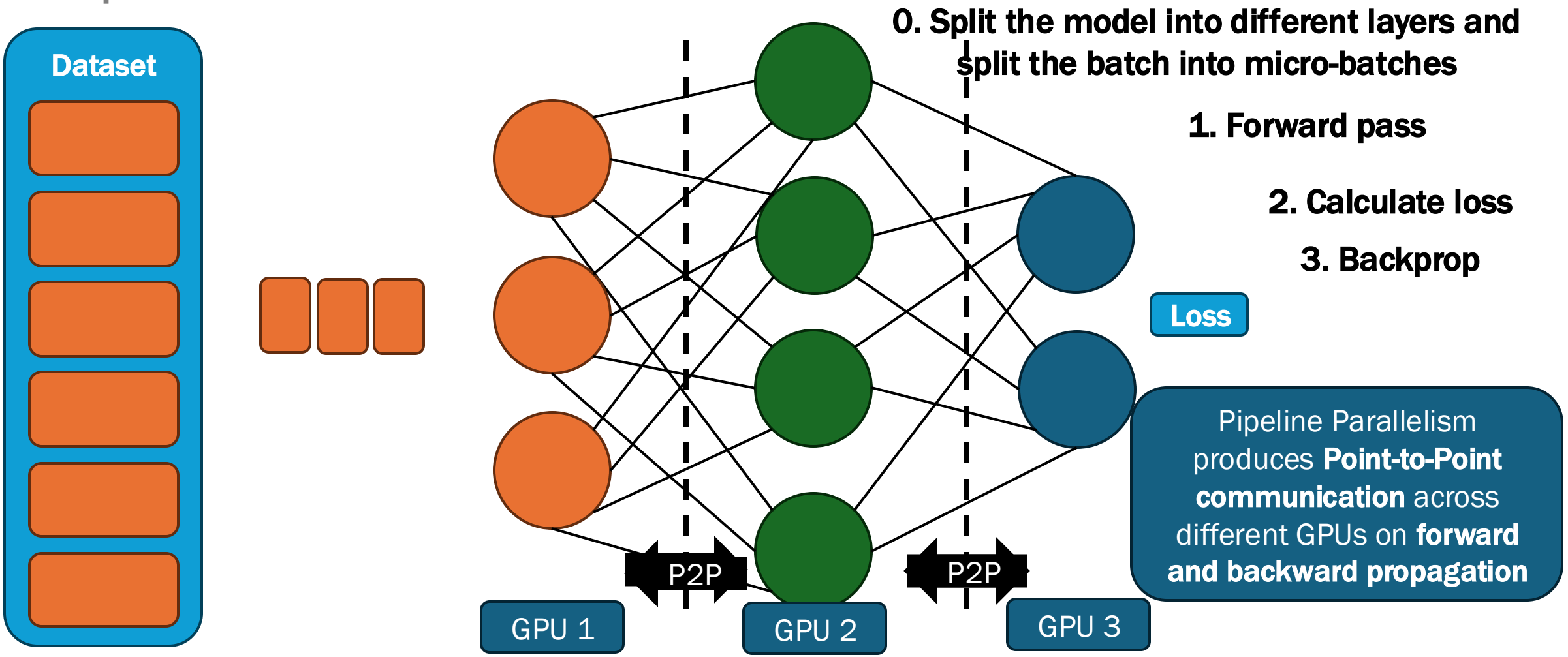
Communication patterns

Data Parallelism



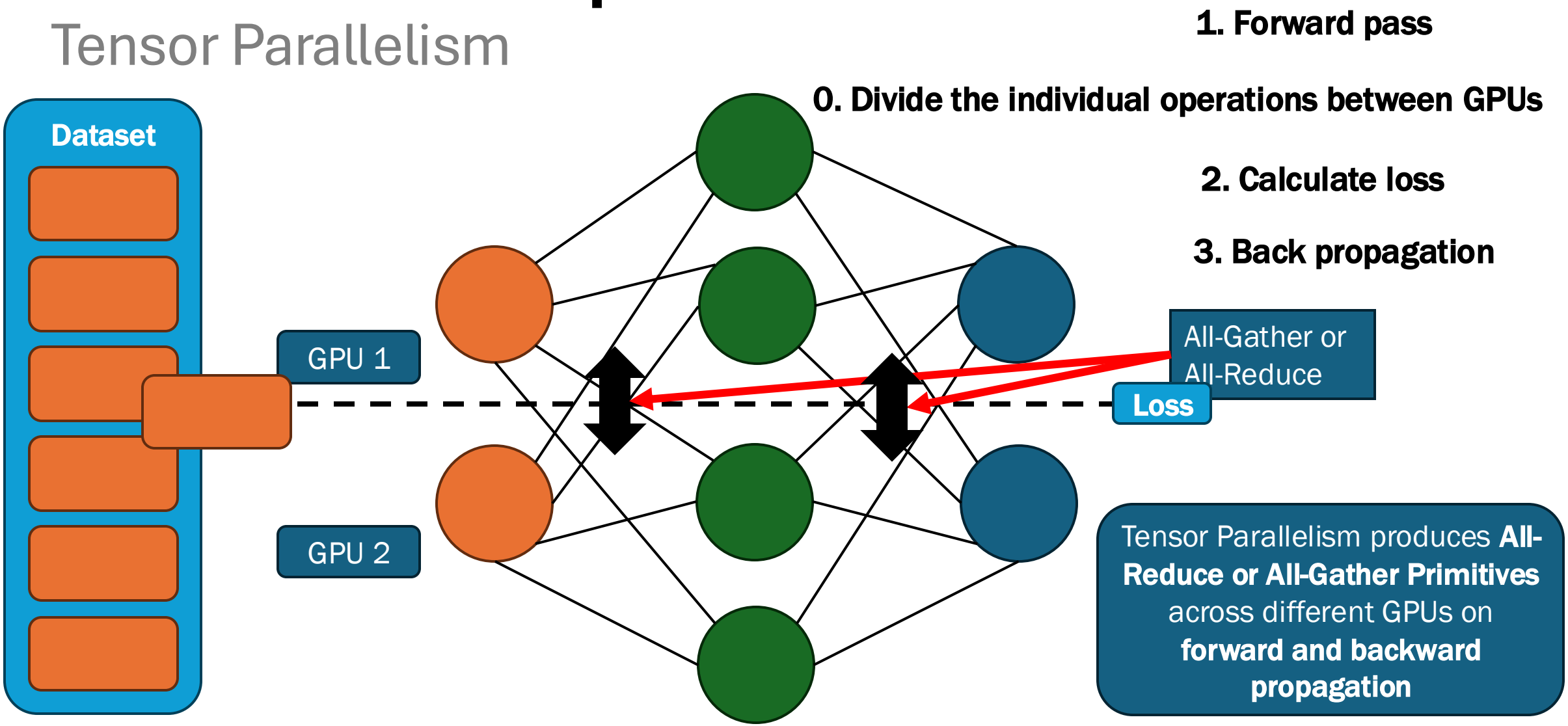
Communication patterns

Pipeline Parallelism



Communication patterns

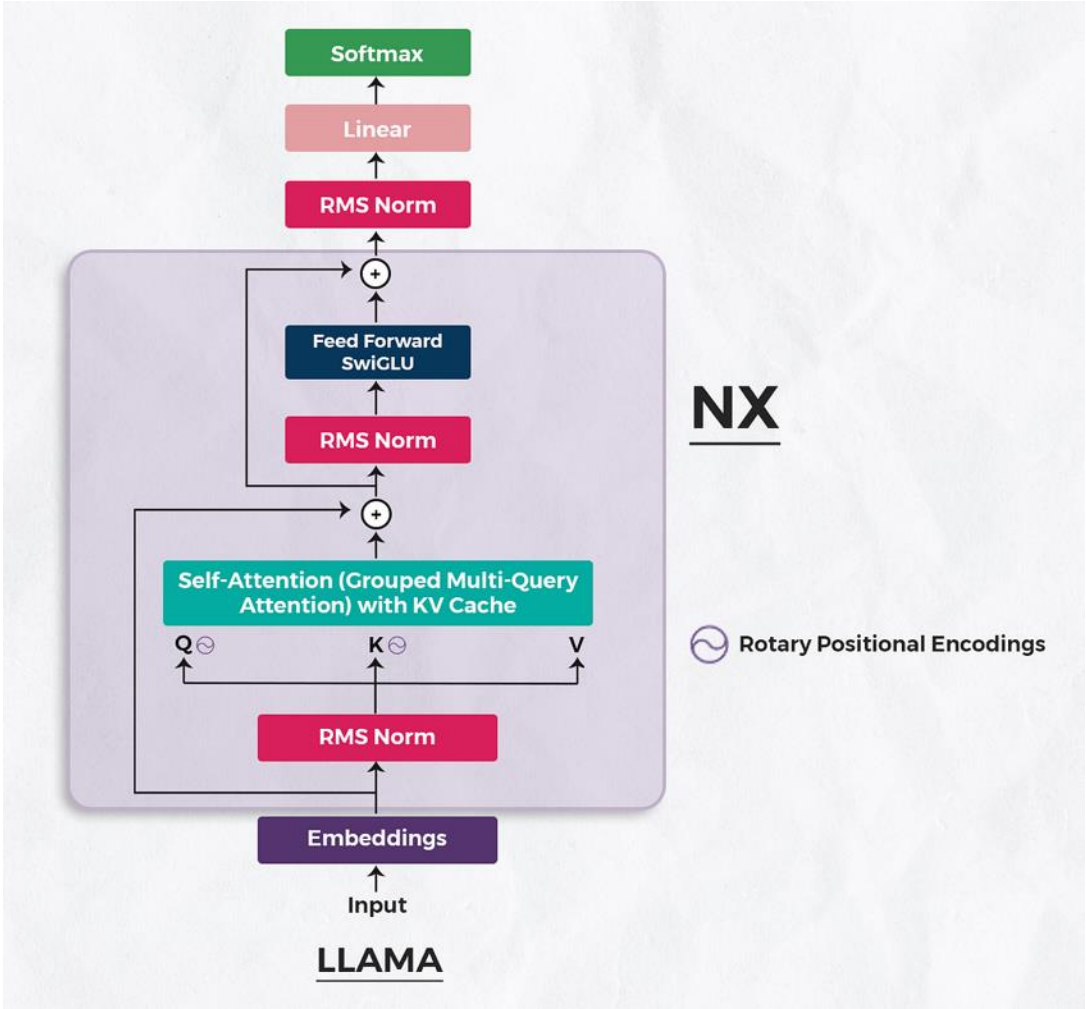
Tensor Parallelism



Communication patterns

LLM traffic model

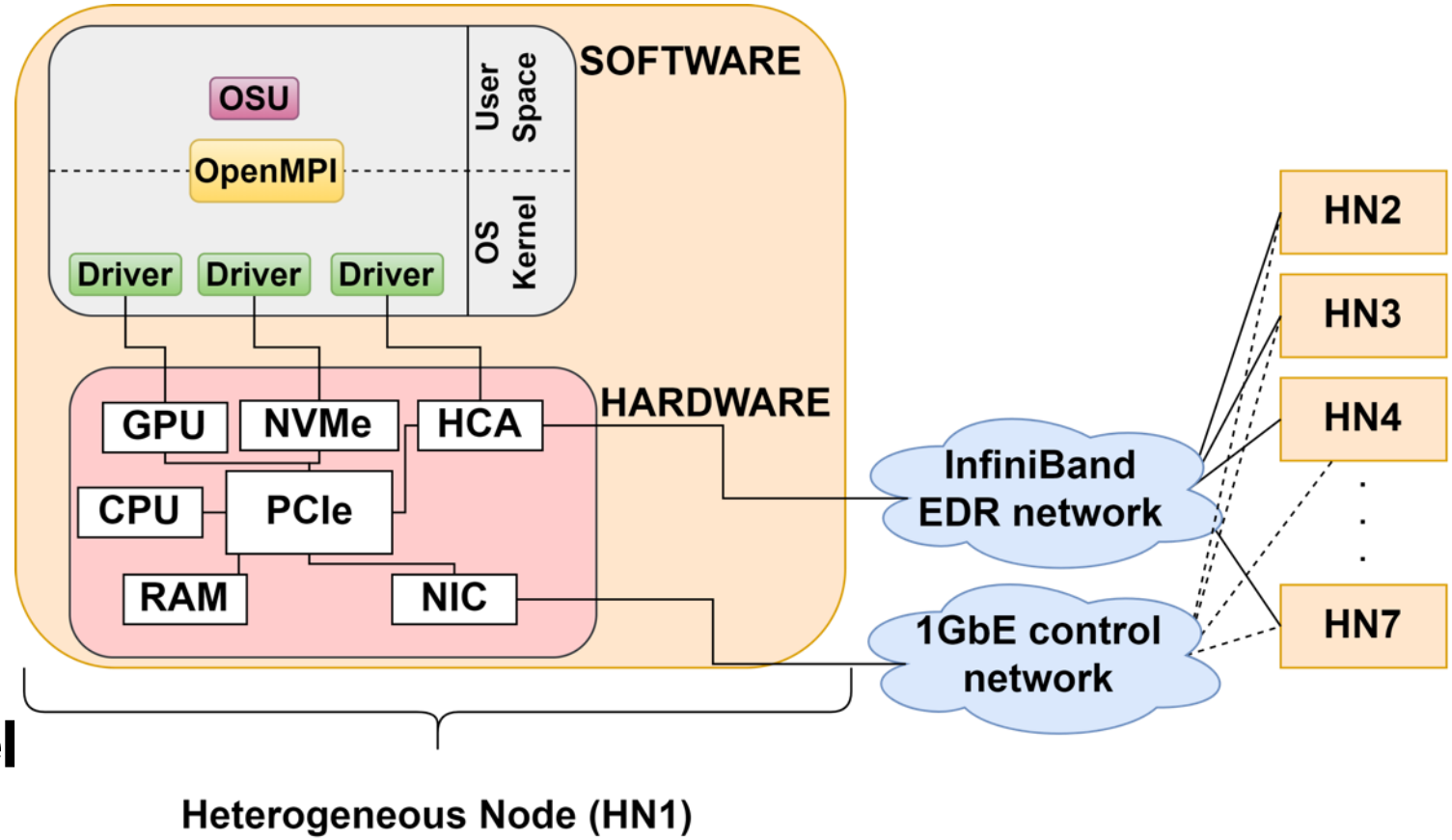
- **LLaMA** model (Open Source)
- **FP16 data – each of 2B**. For example, LLaMA 3.1 405B parameters, it needs 810GB of memory
- Traffic model:
  - Check how many GPUs are needed to train the model
  - To perform model parallelism:
    - If the **GPUs are inside a node** and have a powerful network. We do **Tensor Parallelism**
    - If the **GPUs are in different nodes**. We do **Pipeline Parallelism**



Intra-node network model

“Classic” end-node architecture

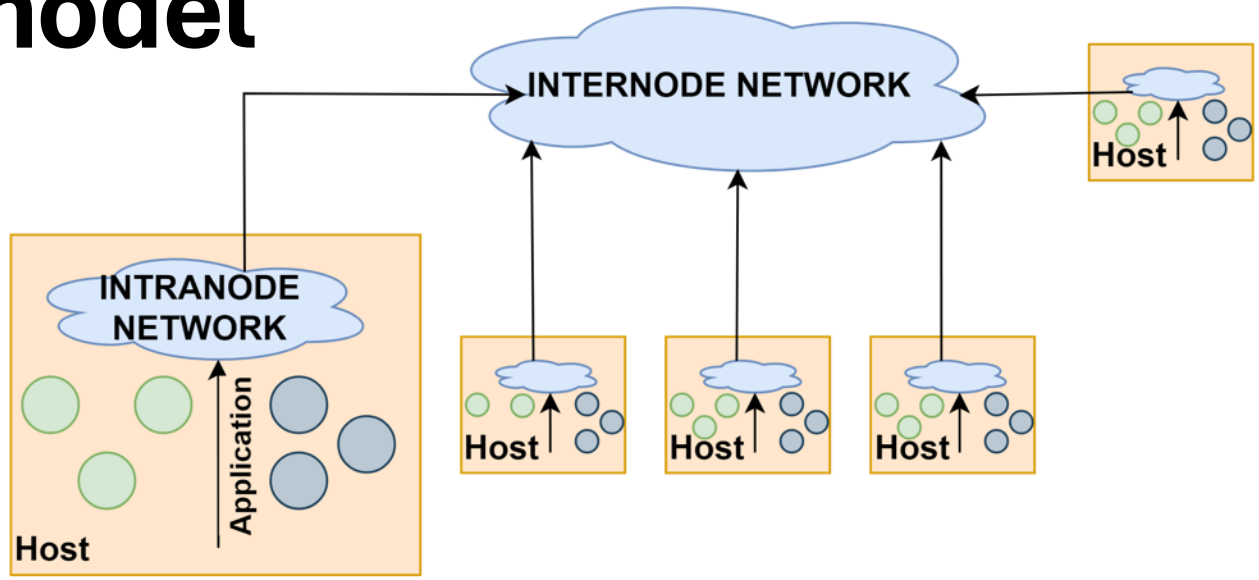
- **Real infrastructure:**
  - Two Intel Xeon Silver 4116
  - 192 GB of RAM
  - NVIDIA Tesla T4 GPU
  - Ultrastar SN200 NVMe SSD
  - PCIe 3.0
  - EDR InfiniBand network interface
- We wanted to compare our simulation results with real micro-benchmarks measurements to validate our **generic simulation model**

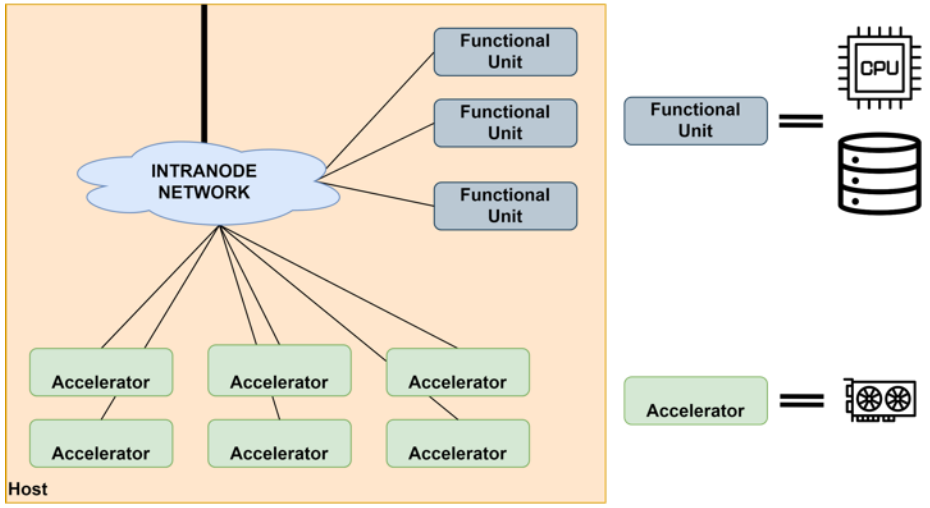


Intra-node network model

The big picture

- **Inter-node network:** each system host is interconnected using a set of switches.
- We assume **two different** types of **hardware devices**:
  - **Functional Unit:** involves CPUs or memory disks
  - **Accelerator:** involves GPUs or any type of accelerator
- **Intra-node interconnection network:** a set of intra-node NICs, switches, and cables to interconnect (scale-up) FUs and accelerators at high speed
- **This model is fully parametrizable,** which allows configuring specific intra- and inter-node network technologies and characterizing their communication performance

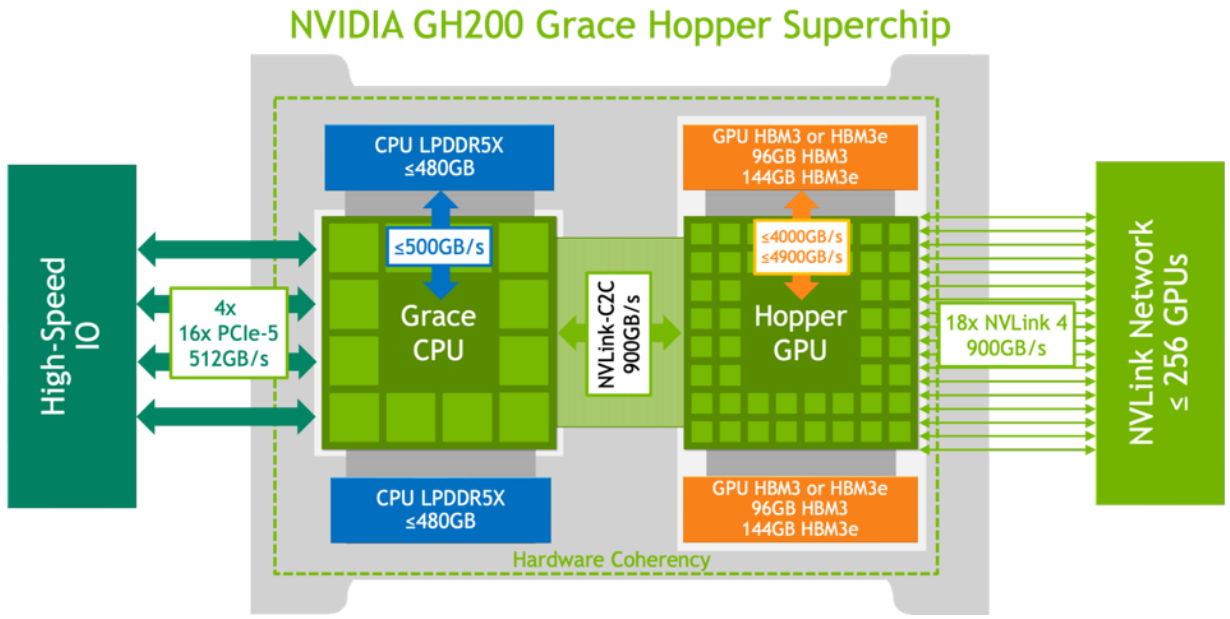




Intra-node network model

NVIDIA NVLink

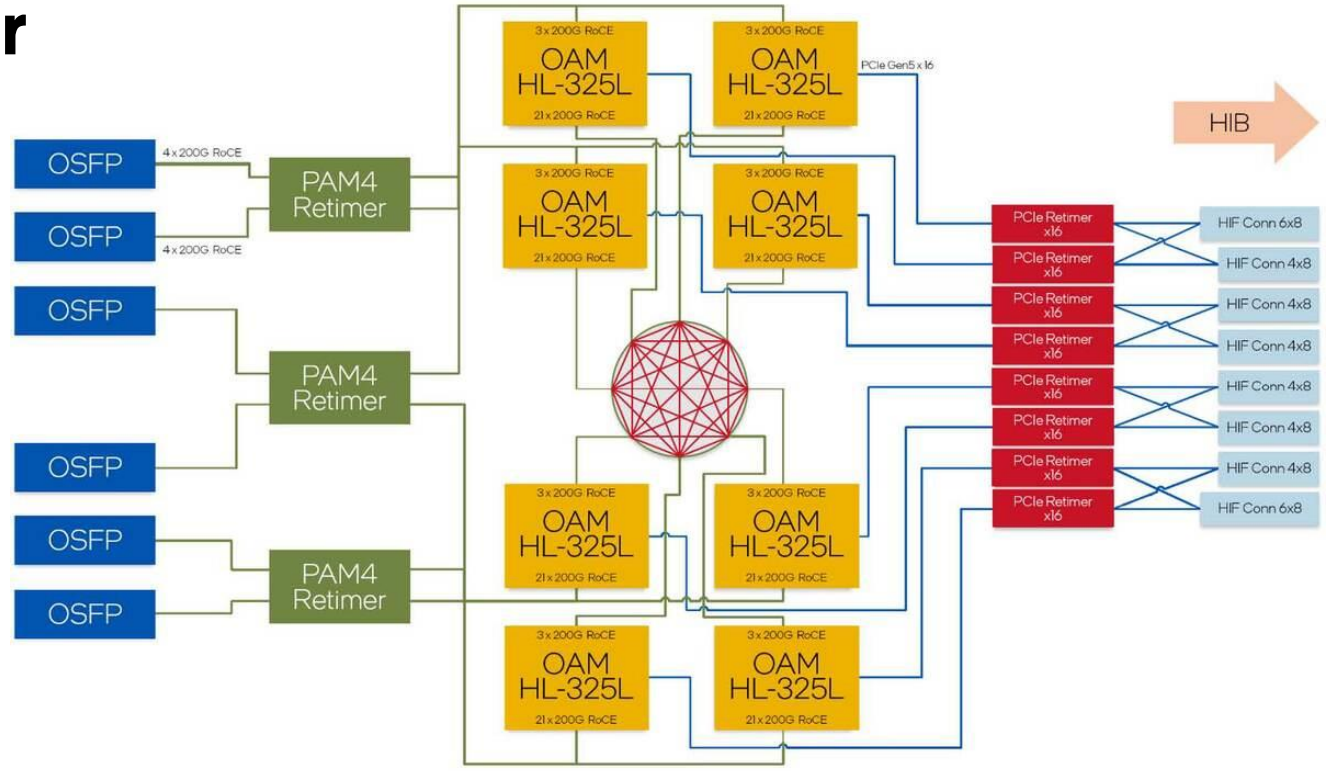
- NVIDIA GH200 Grace Hopper Superchip
- **NVLink important features**:
  - NVLink-C2C
  - NVLink network
- NVLink network between GPUs
- Interconnection among memory disks, HCA, etc. through PCIe and InfiniBand



Intra-node network model

Intel Gaudi3 Accelerator

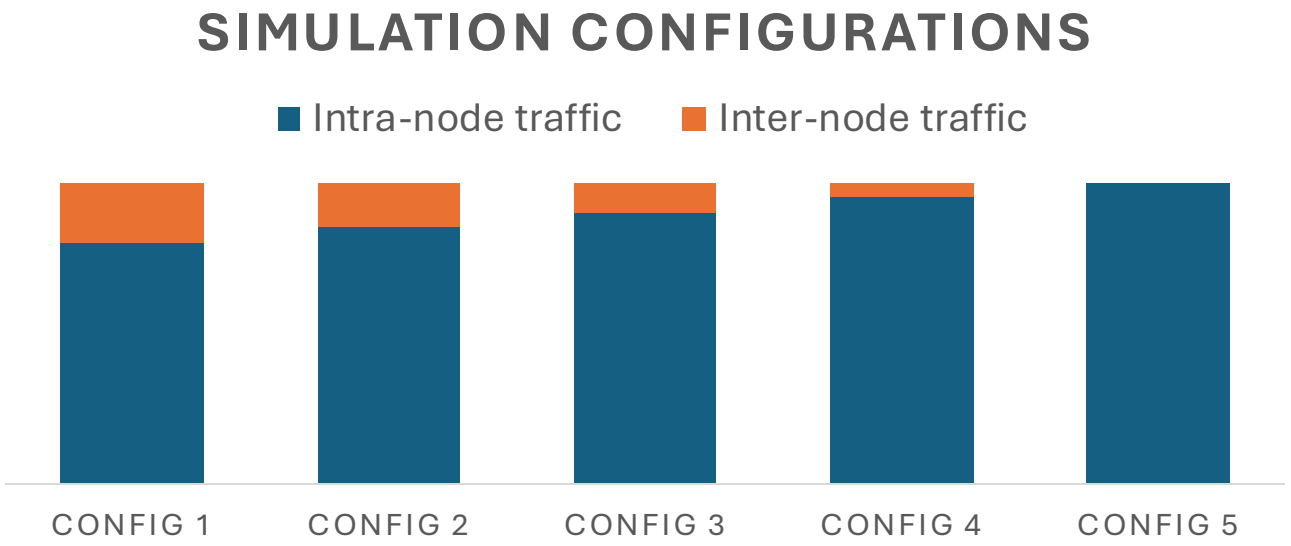
- **24 Ethernet NICs per accelerator**
- **6 OSFP for inter-node communication**
- **Scale-out network** between accelerators (based on Ethernet and RoCE)
- **Scale-up network** through PCIe
- Is what UALink is doing? ;-)



Intra-node network model

Simulation experiments

- **Five different configurations:**
  - C1 – Intra-node traffic 80%
  - C2 – Intra-node traffic 85%
  - C3 – Intra-node traffic 90%
  - C4 – Intra-node traffic 95%
  - C5 – Intra-node traffic 100%

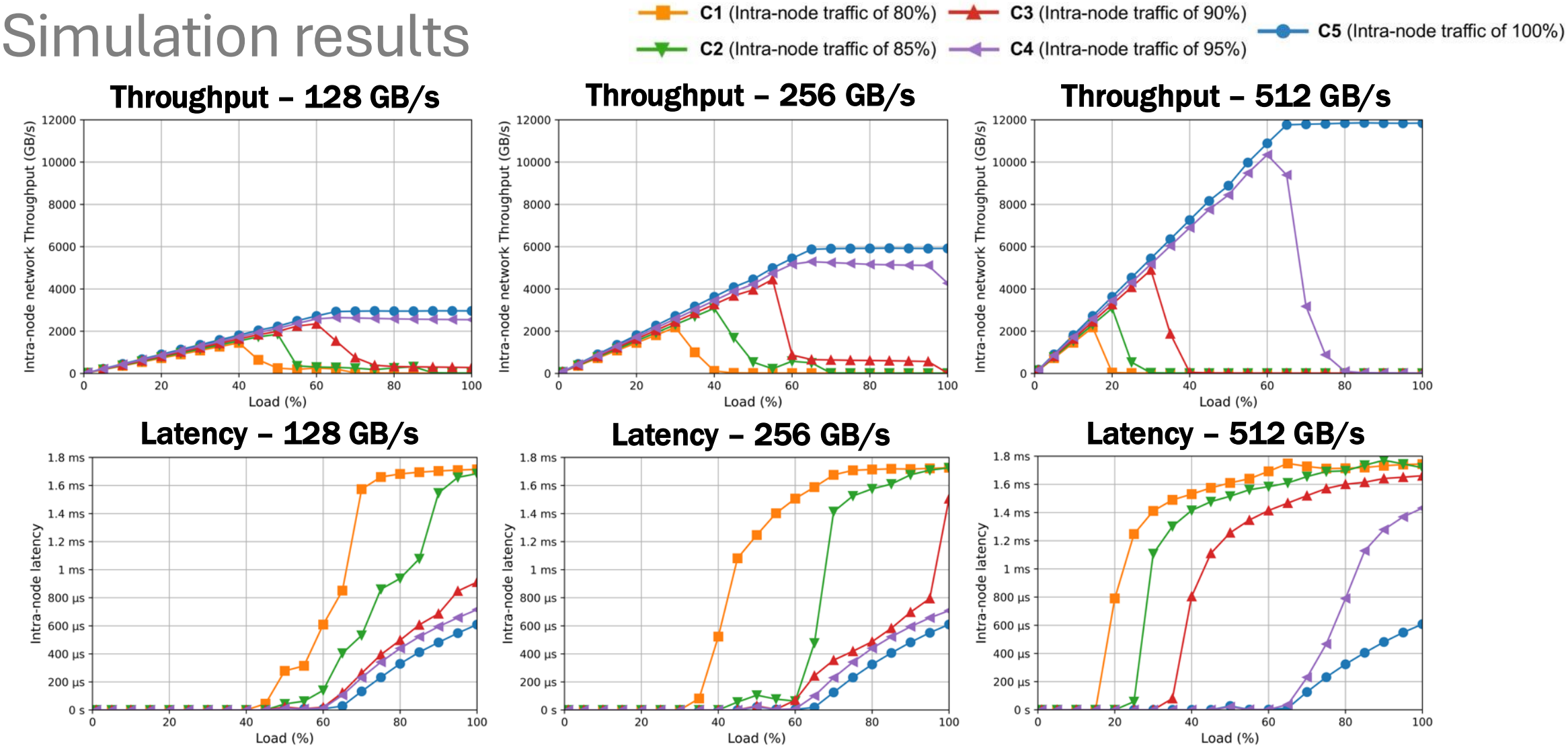


- **RLFT network** with **32 end-nodes** and **256 accelerators** (8 per node).
- **Accelerators links 128 Gbps, 256 Gbps and 512 Gbps**
- **Three** different **intra-node speeds:** 128 GB/s, 256 GB/s and 512 GB/s
- **Inter-node** network of **400 Gbps**

Joaquin Tarraga-Moreno and Jesus Escudero-Sahuquillo and Pedro Javier Garcia and Francisco J. Quiles: **Understanding intra-node communication in HPC systems and Datacenters**. Arxiv pre-print 2025.

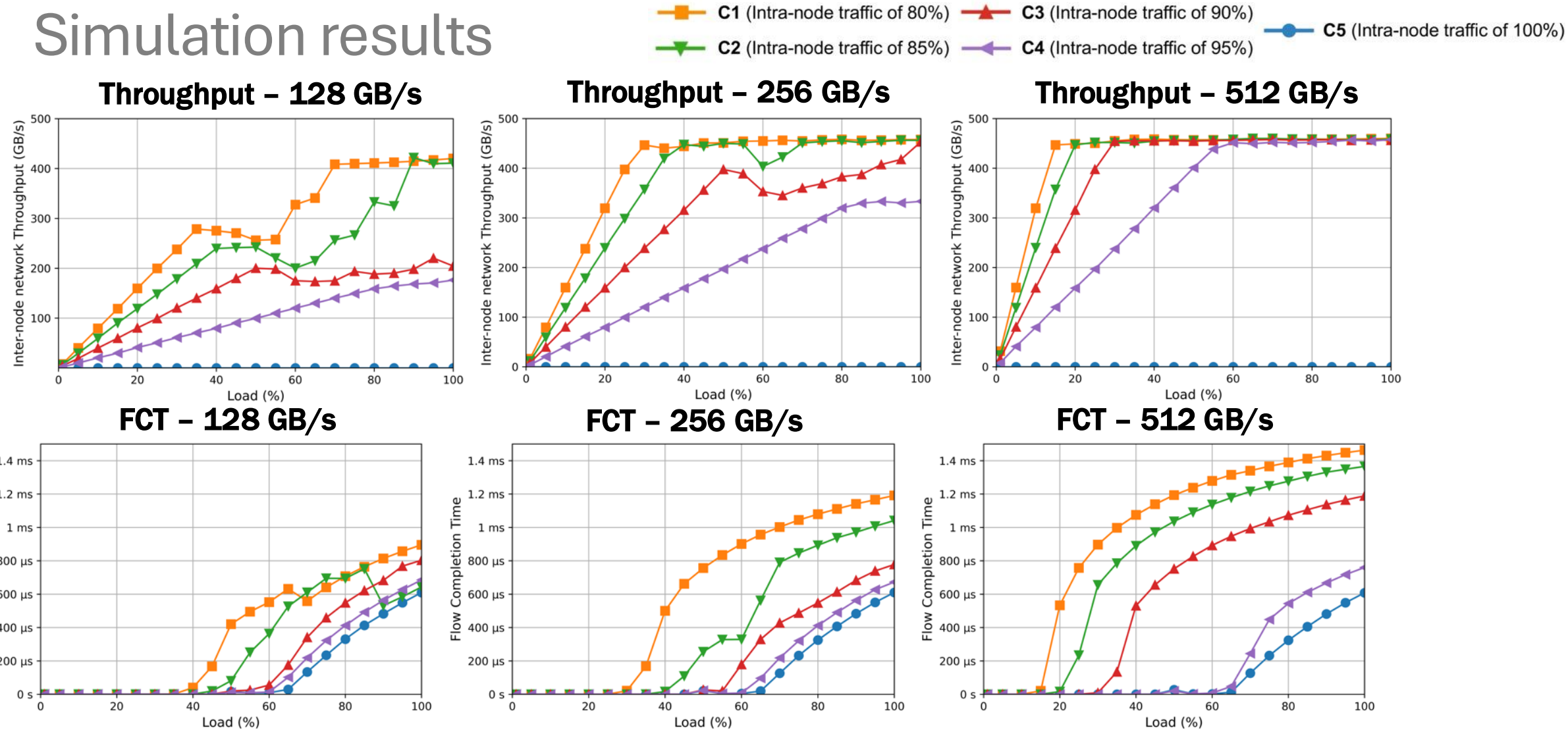
Intra-node network model

Simulation results



Intra-node network model

Simulation results



Intra-node networks

Final remarks

- Increasing the intra-node performance may be counterproductive with certain types of model parallelism
  - Incoming inter-node traffic may collapse the accelerators' memory and impact intra-node communication
- Dedicated intra-node networks for large AI/HPC systems are expensive
- Modeling impact of AI inference workloads:
  - Training needs high network throughput
  - …but inference serving systems require low latency
  - … and millions of users requesting at the same time will collapse intra-node networks and impact the performance of the inter-node network
- AI is power hungry → network design needs to contribute to reduce the energy bill

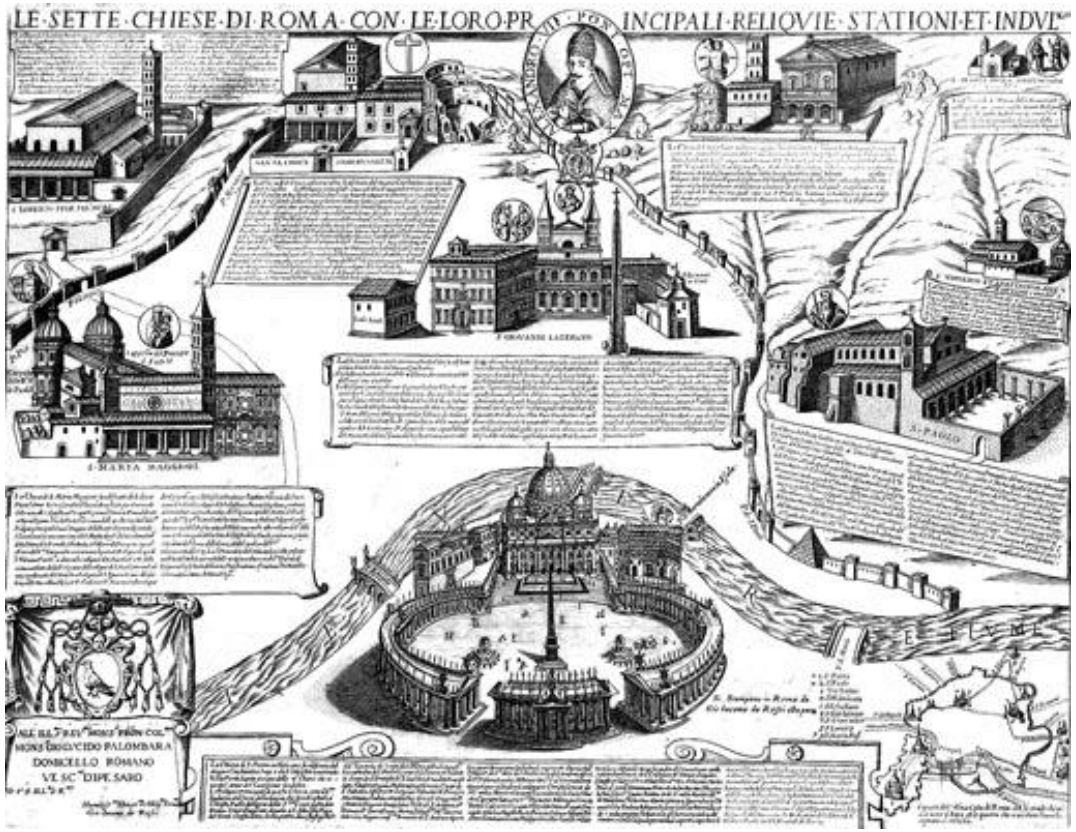
Topologies & Routings

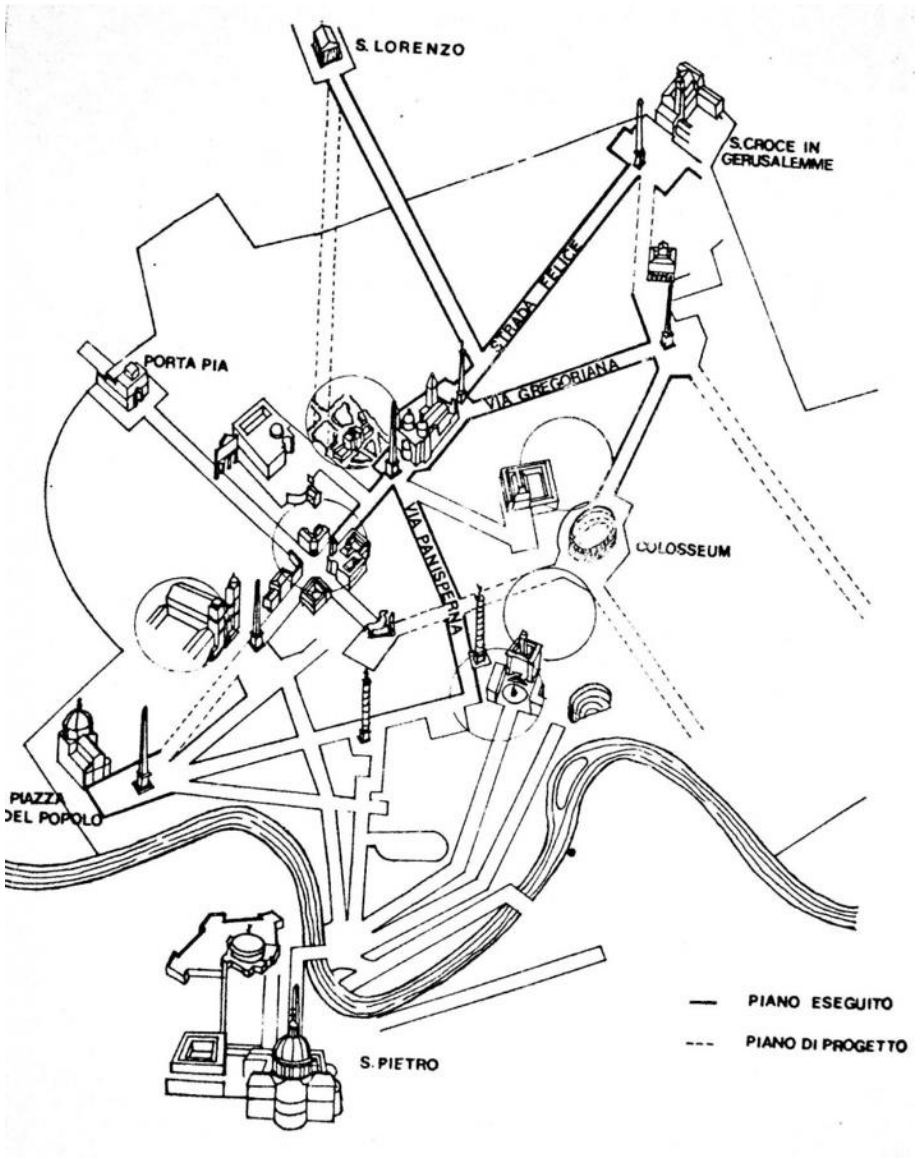
Historical Notes

- Sixtus V and the topology of Rome:
  - Many pilgrims (packets) in Rome visiting several churches (endnodes)
  - By the end of the XVI century the network made of squares (switches) and streets (links) in Rome was chaotic-> pilgrim jams
  - Sixtus V designed a new topology for Rome, trying to balance pilgrim flows

Topologies & Routings

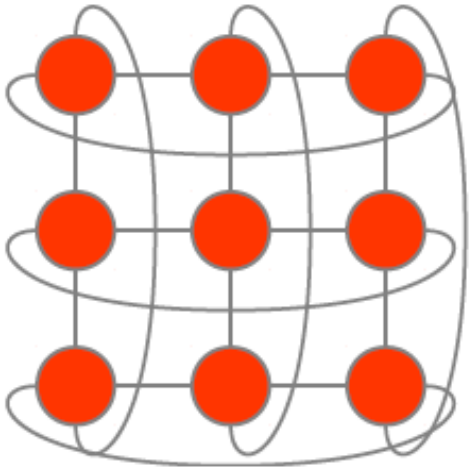
Historical Notes



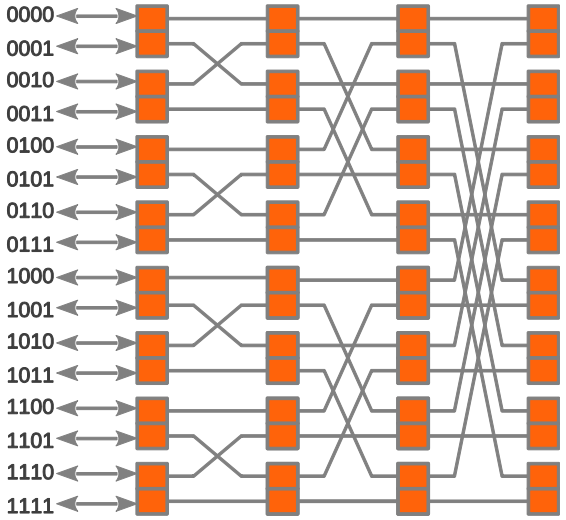


Topologies & Routings

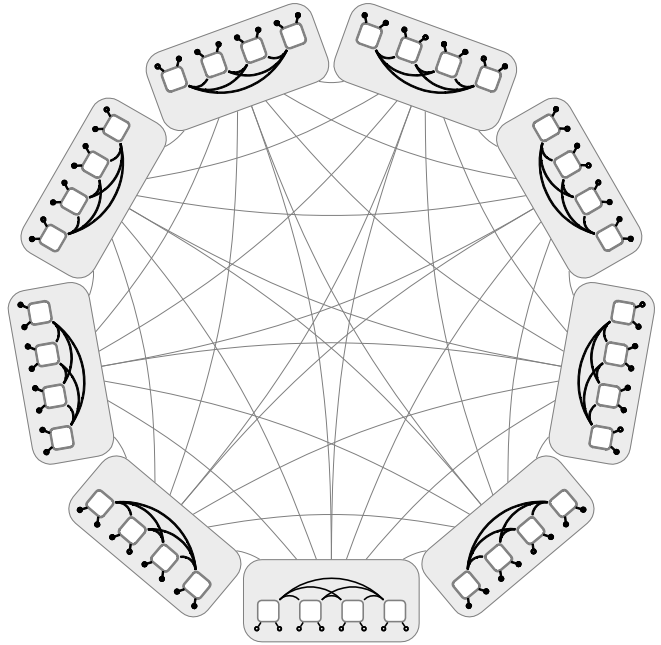
- **Efficient topologies** (low diameter, path diversity, reachable deadlock freedom, etc.) and their corresponding **routing algorithms** are essential to achieve good system performance
- The key point is achieving traffic balance throughout the network



Torus



k-ary n-tree



Dragonfly

Routings

Deterministic vs Multi-Path

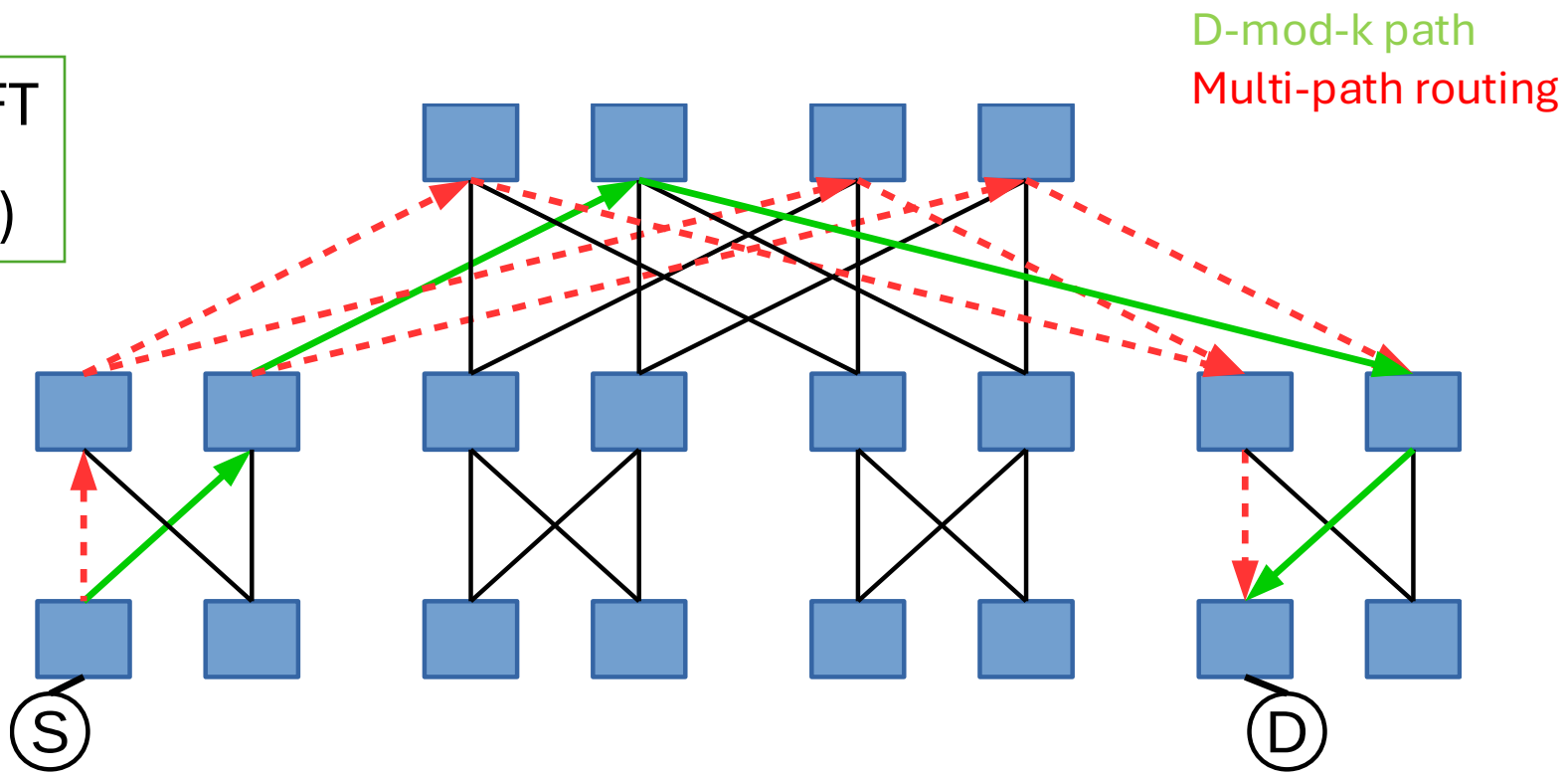
- In contrast with **deterministic** routing, **multi-path** (**oblivious** and **adaptive**) routings may use several paths between any source and destination:
  - Oblivious: routing **independent** of traffic status
  - Adaptive: routing decisions **based on** network conditions
- Multi-Path traditionally considered the right choice to balance traffic and/or to avoid/eliminate/delay congested points, but….
  - Problems regarding in-order packet delivery
  - Deadlock freedom may be more complex to achieve
  - Useless or counterproductive under incast congestion scenarios
  - In-network congested points may vary

Routings

Multi-Path Problems

3-stage RLFT

(K=2, n = 3)

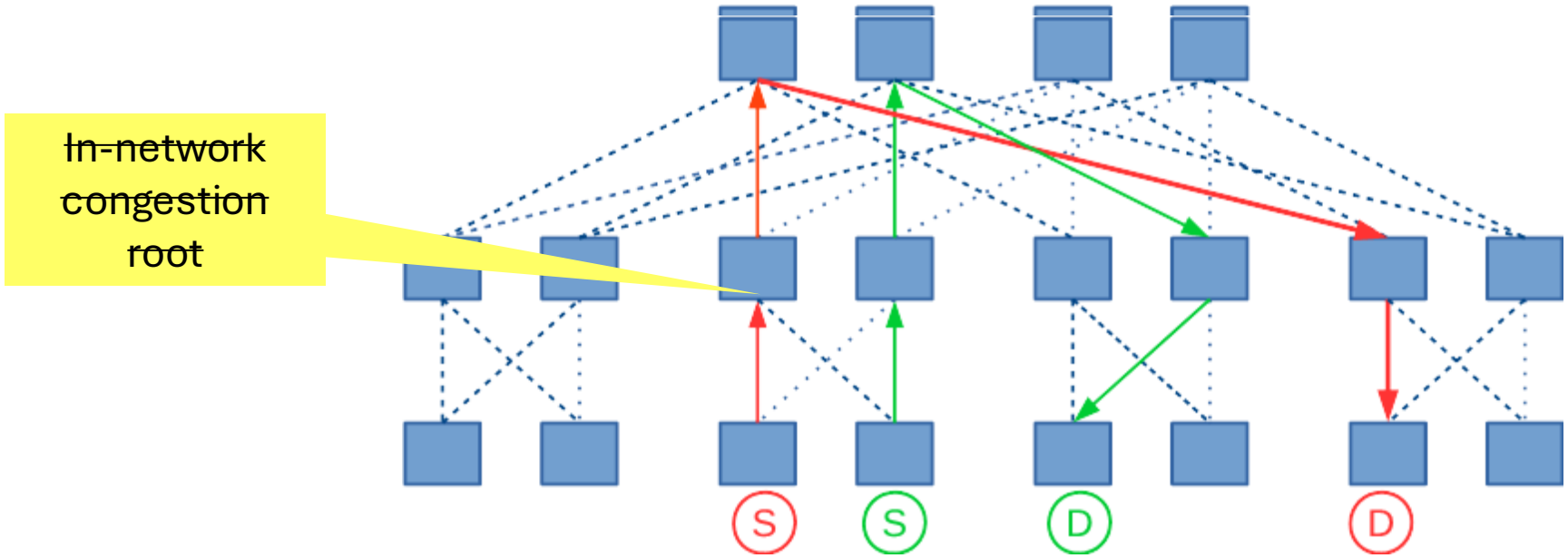


The maximum minimal-path diversity is K^{n-1}

Routings

Multi-Path Problems

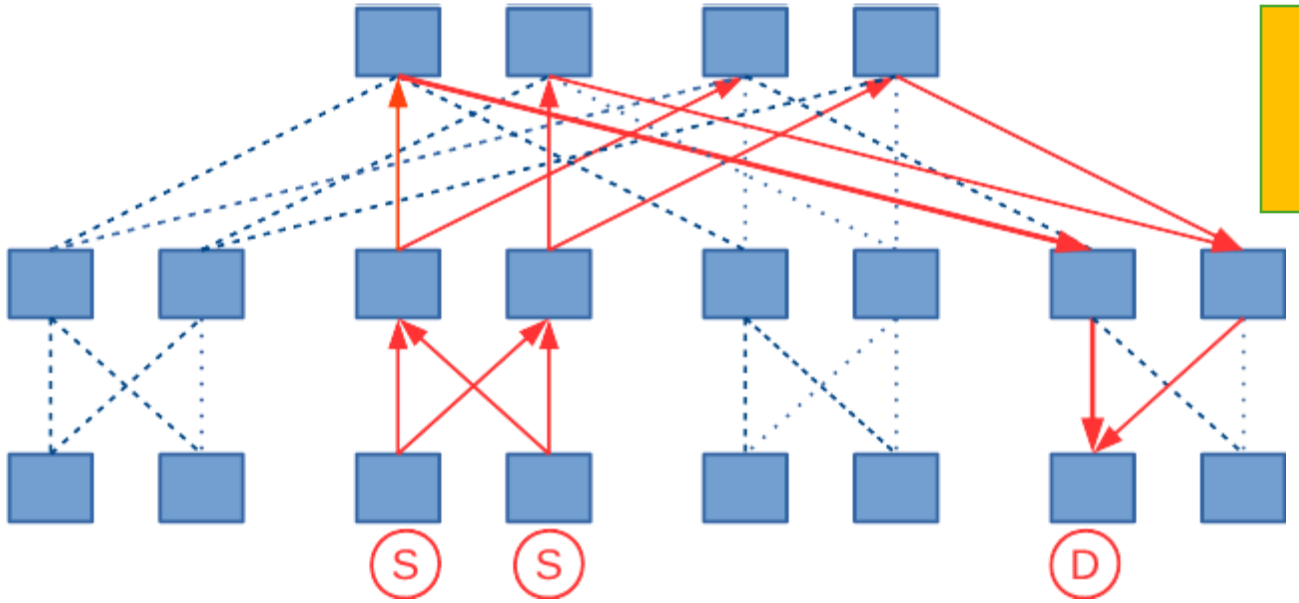
- In-network congestion can be solved using multi-path routing



Routings

Multi-Path Problems

- Incast congestion cannot be solved with multi-path routing

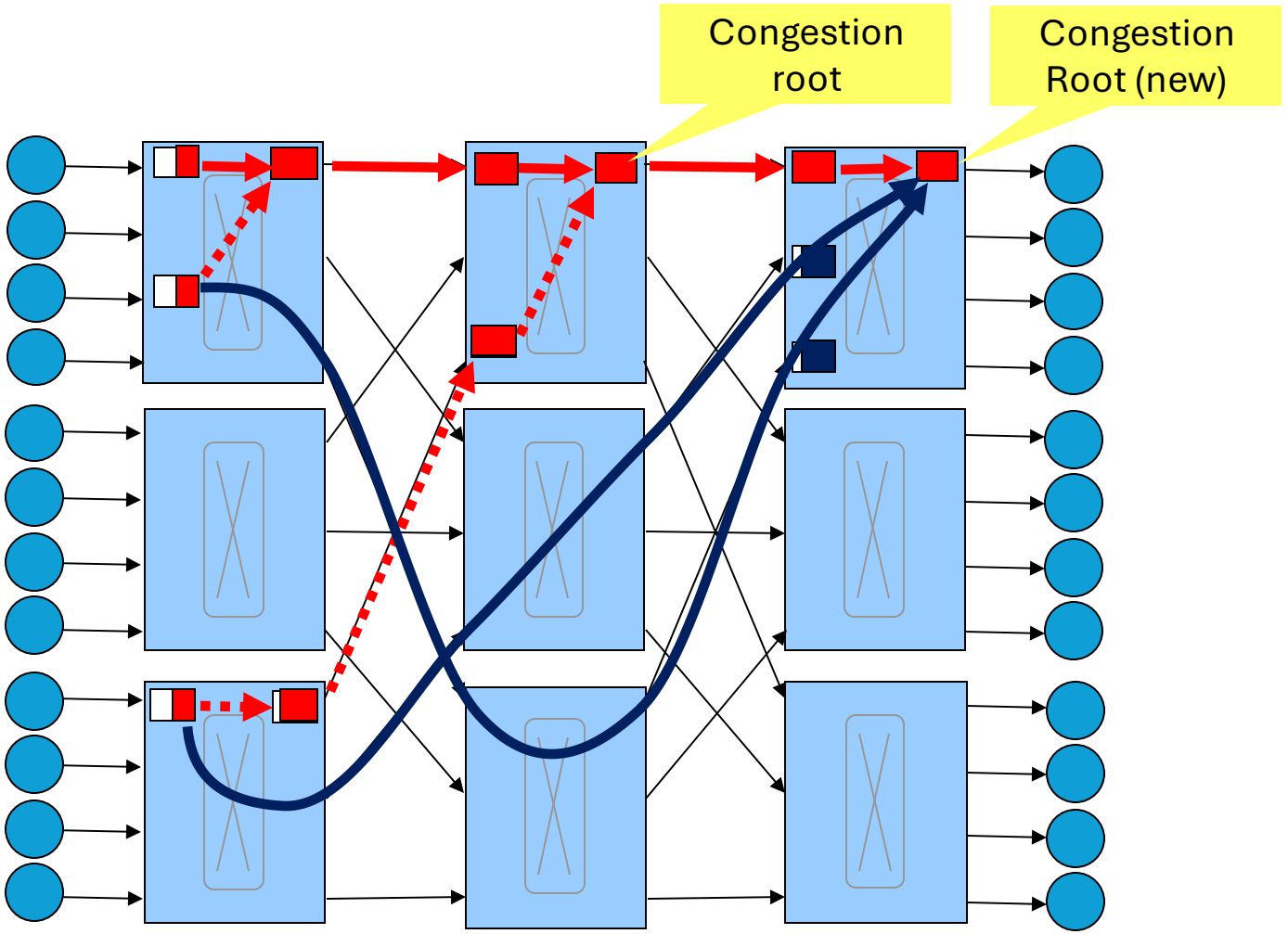


Congestion spreading

Multi-path routing

Routings

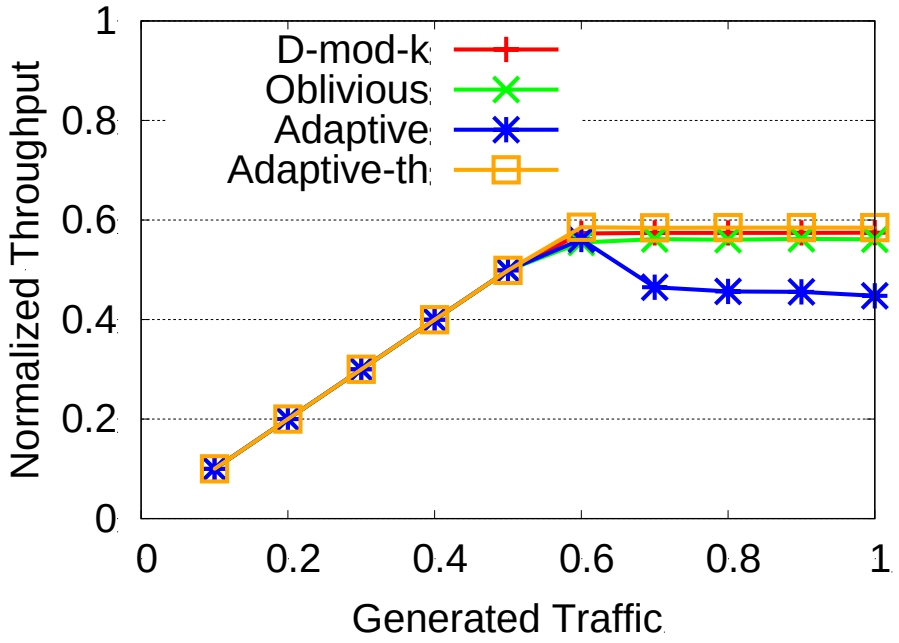
Multi-Path (ECMP, packet spraying) Problems



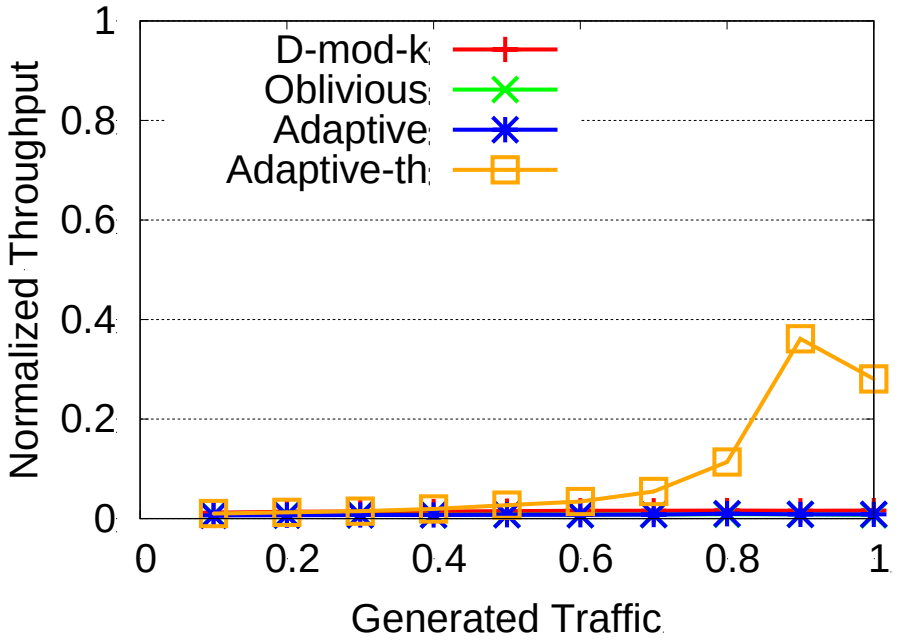
P. J. García, J. Escudero-Sahuquillo: Interconnection Networks in the Exascale and AI Era: Challenges and Solutions. Academic Salon on High-Performance Ethernet, TUM, Munich, 12-13 March 2025

Routings

Multi-Path Problems



11664-node real-life fat-tree, random traffic



11664-node real-life fat-tree, 10% hotspot

Routings: Solution to Adaptive Problems

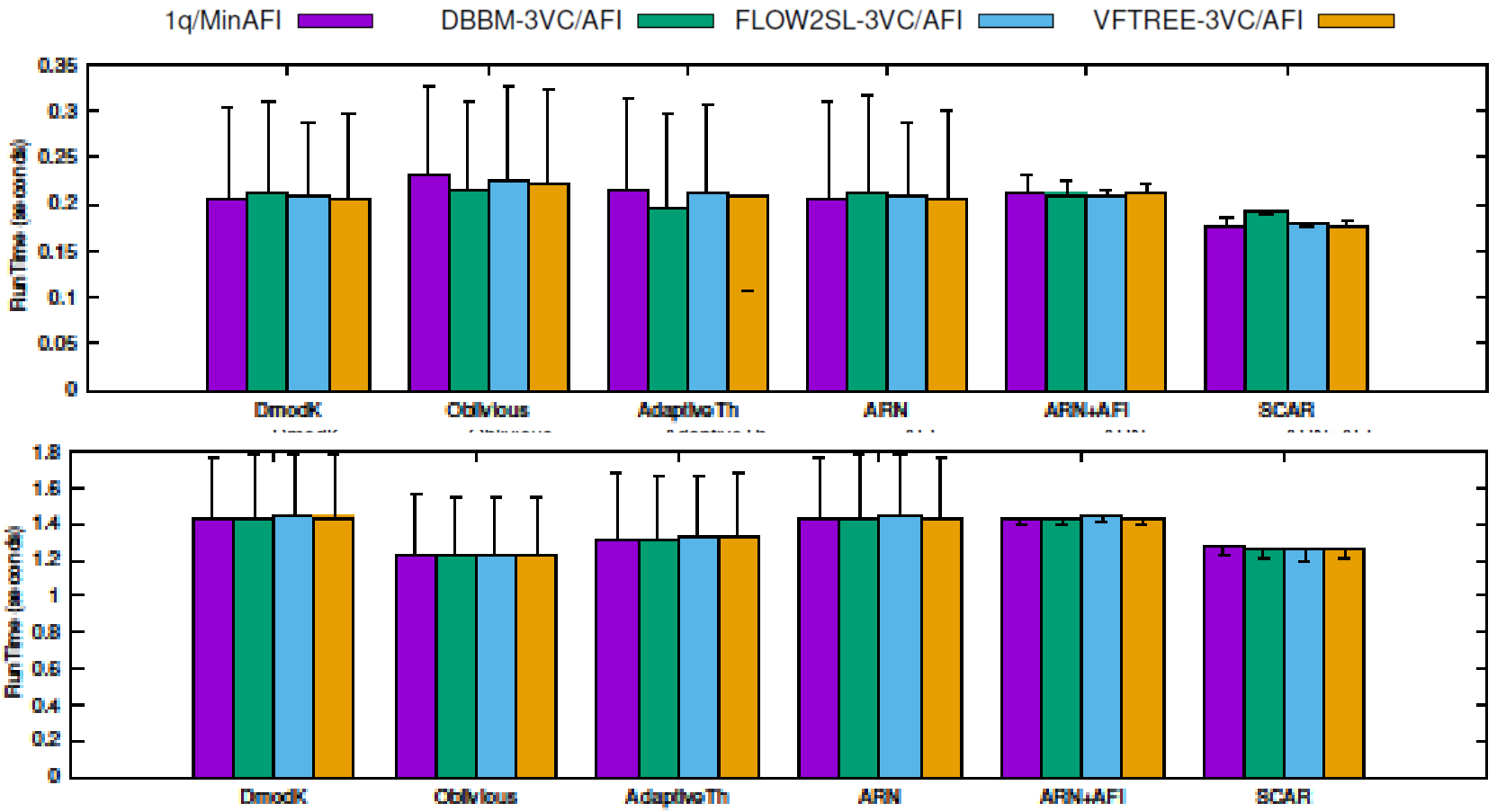
Managing Incast and In-network through Adaptive Routing

- We propose SCAR, a new technique that **combines congestion management and adaptive routing** so that in-network and incast congestion impact is reduced
- SCAR leverages adaptive routing notifications (ARNs) to **avoid using adaptive routing for congesting flows contributing to incast** situations, so congestion spreading is mitigated
- SCAR also uses **adaptive routing for non-congesting** (i.e., victim) flows to leverage the network path diversity

Jose Rocher-Gonzalez, Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, José Duato: **A smart and novel approach for managing incast and in-network congestion through adaptive routing**. Future Gener. Comput. Syst. 159: 27-38 (2024)

Routings: Solution to Adaptive Problems

Managing Incast and In-network through Adaptive Routing



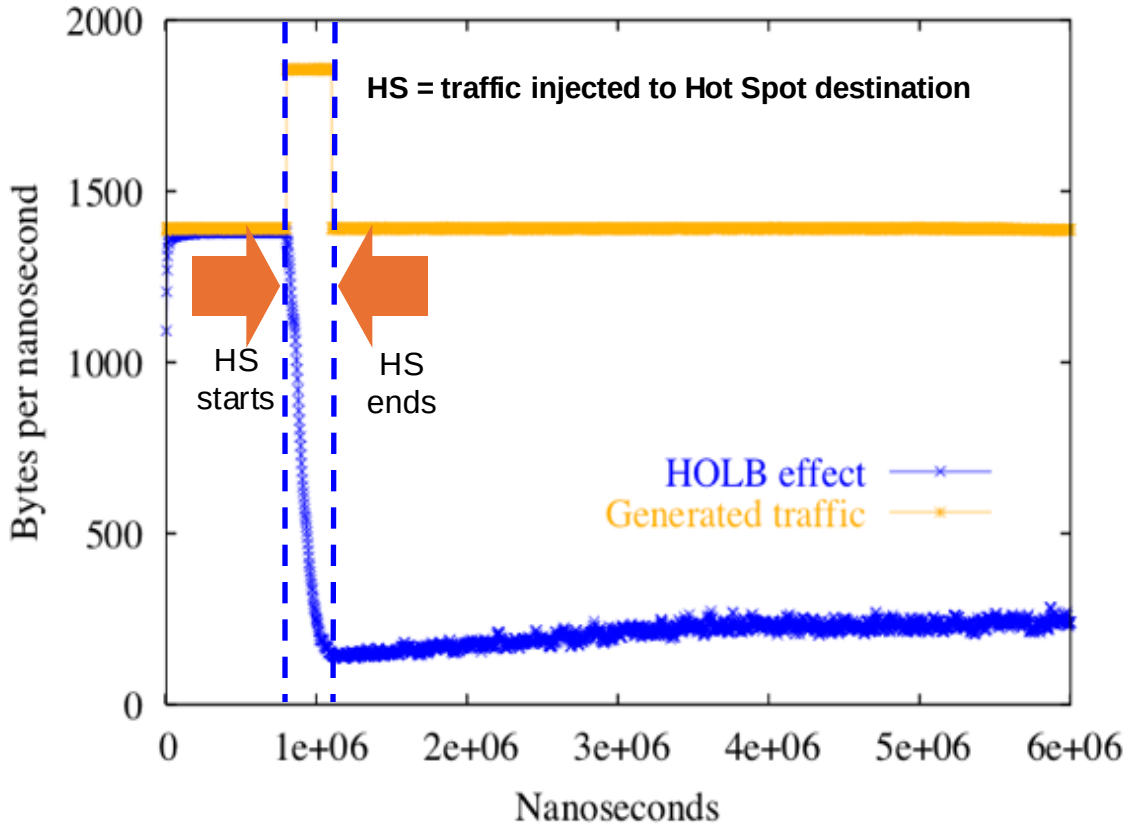
432-node CLOS network

- Runtime when running PTRANS and Inception v3.
- Traffic has been modeled in the simulator **using VEF traces.**
- Error bars show Runtime in the presence of an incast traffic burst)
- SCAR is combined with static queuing schemes to provide an extra HoL-blocking reduction.

Congestion Control

Motivation

- Congestion **causes**:
  - Traffic flows jamming internal network paths
  - Faulty regions
  - Aggressive network power management
- The effects of micro-bursts in Datacenters (lasting for tens of microseconds) are significant for network performance



Congestion Control

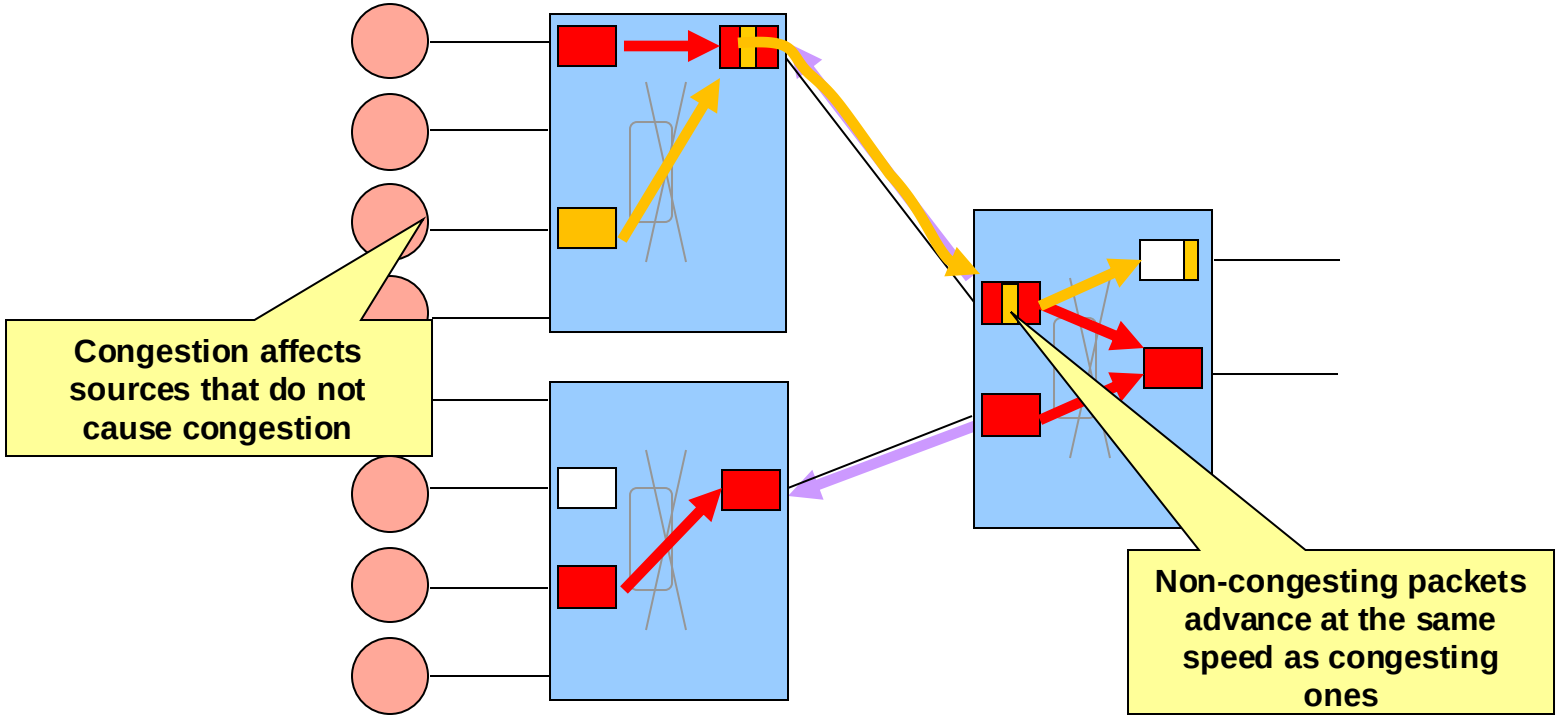
Congestion effects

- Head-of-Line (HoL) blocking
- Buffer hogging
- Parking lot

Congestion Control

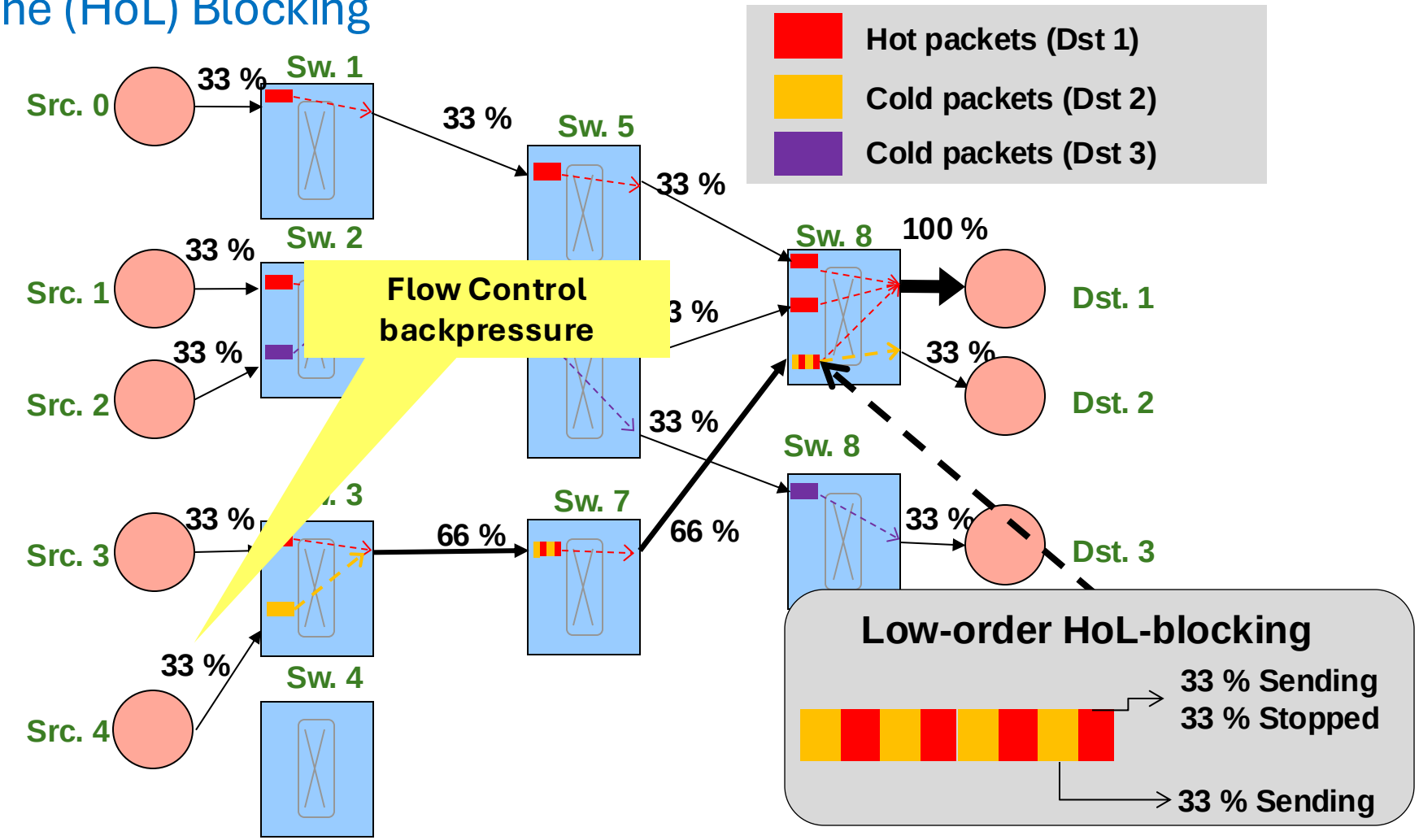
Congestion effects

- Congestion trees may cause **Head-of-Line (HoL) blocking**



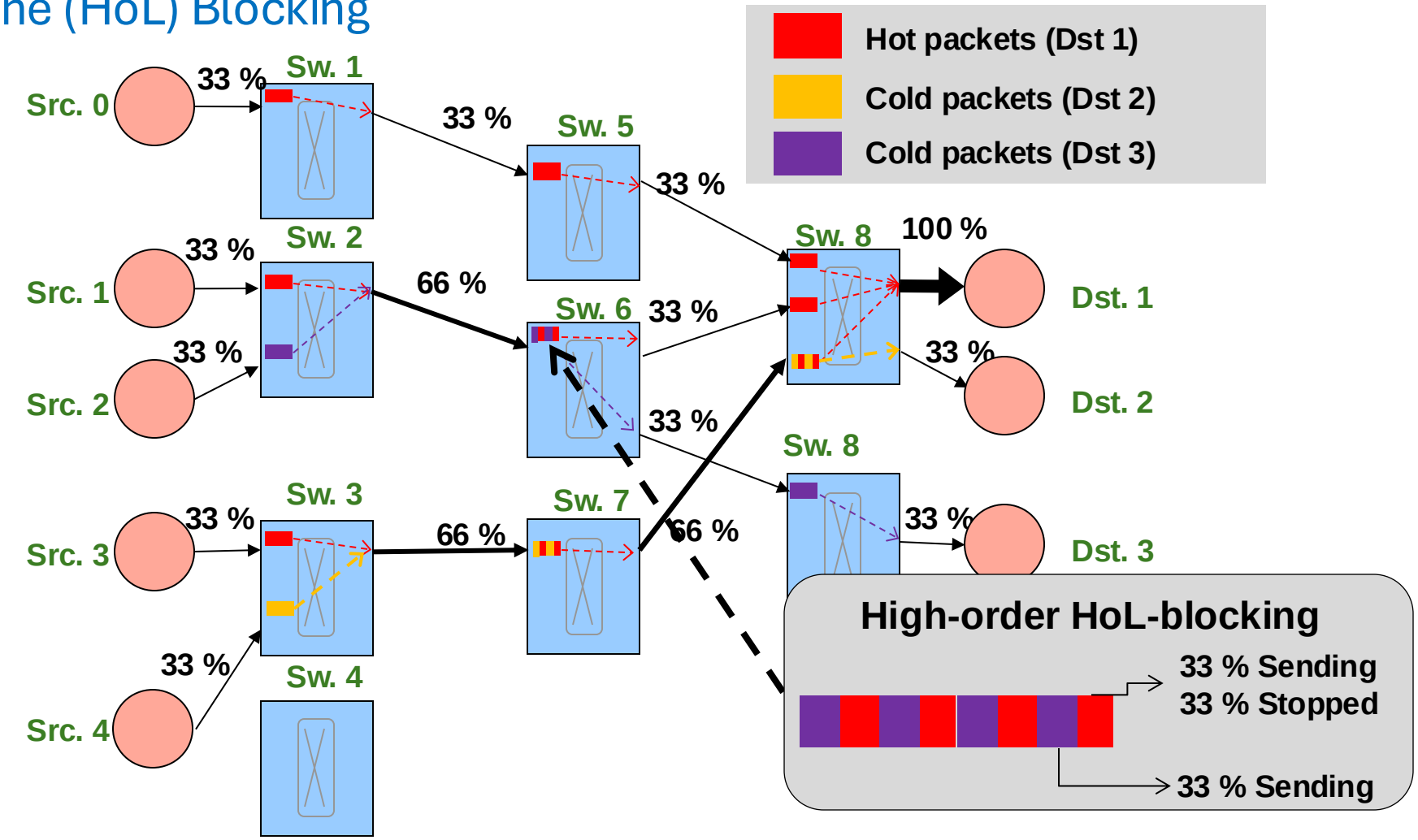
Congestion Control: Congestion effects

Head-of-Line (HoL) Blocking

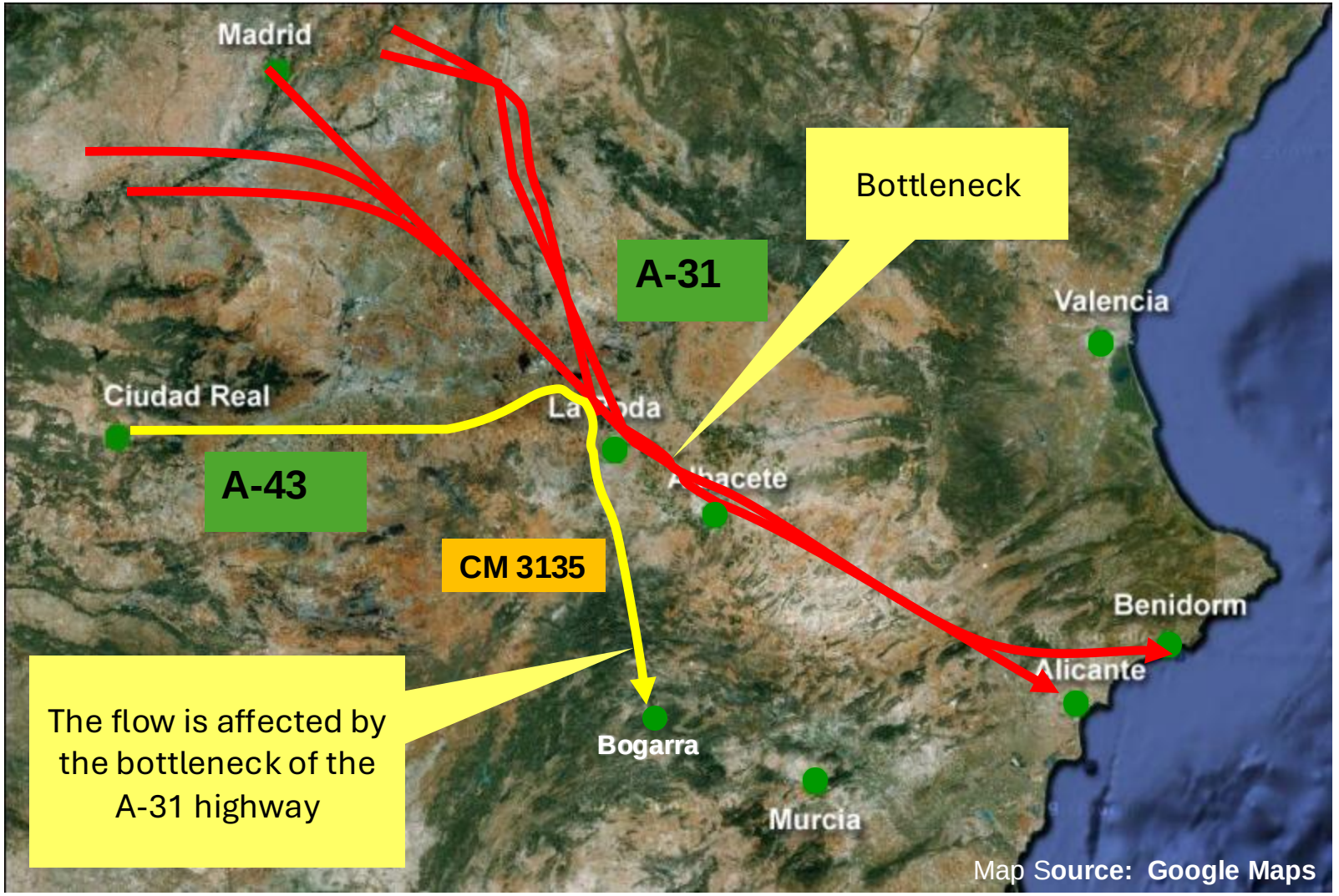


Congestion Control: Congestion effects

Head-of-Line (HoL) Blocking

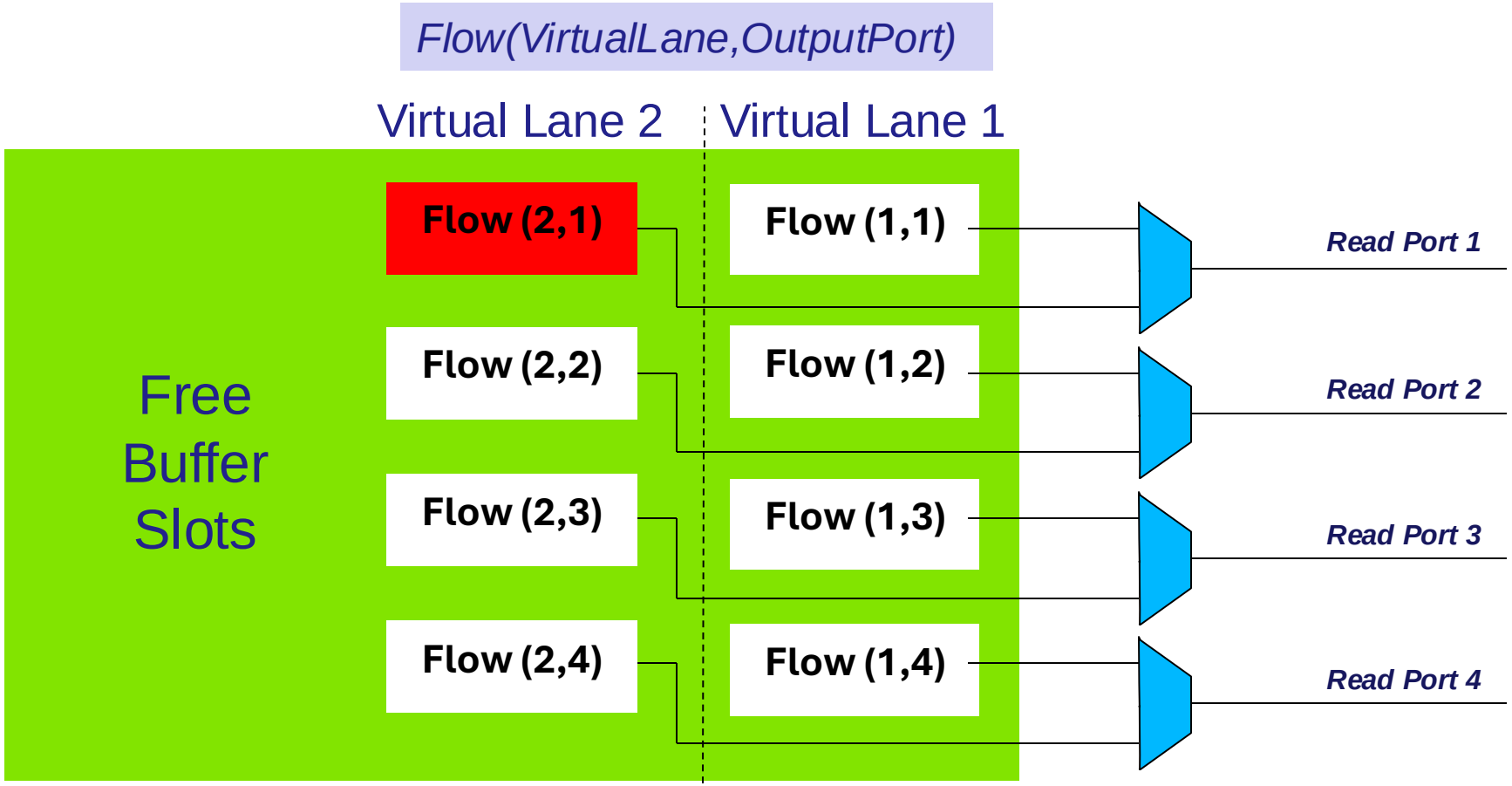


Congestion Control: Congestion effects



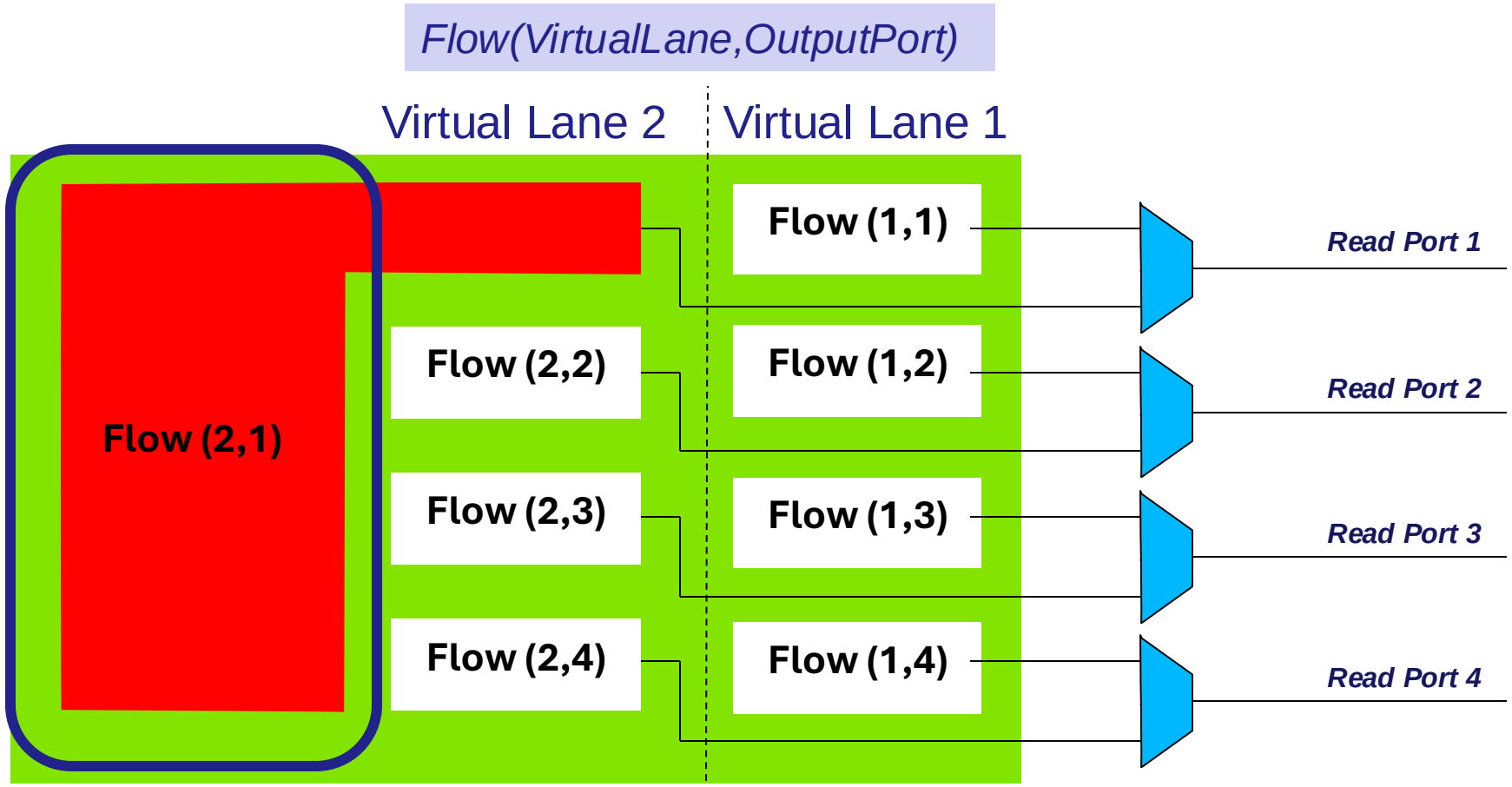
Congestion Control: Congestion effects

Buffer hogging



Congestion Control: Congestion effects

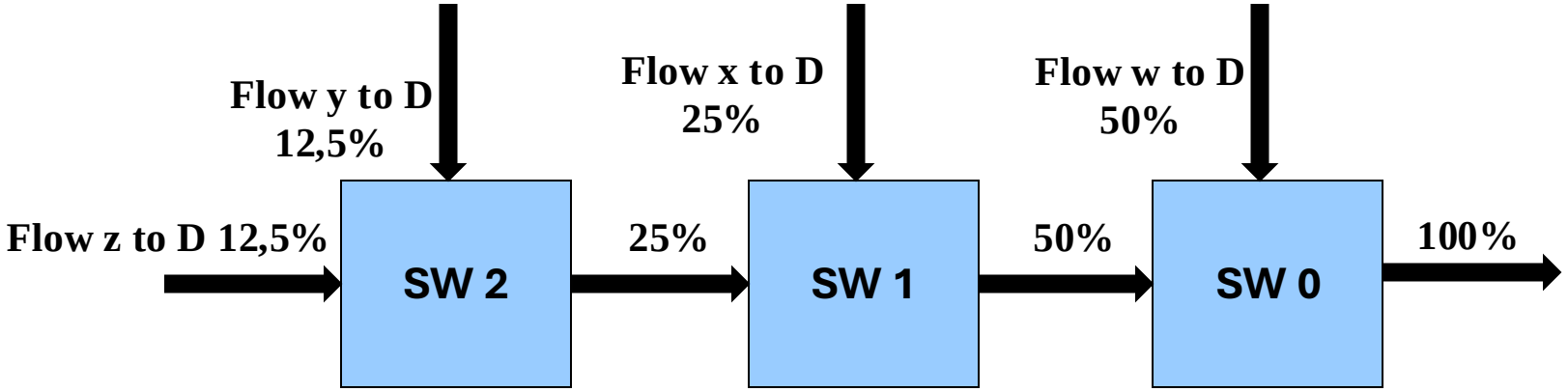
Buffer hogging



Congestion Control: Congestion effects

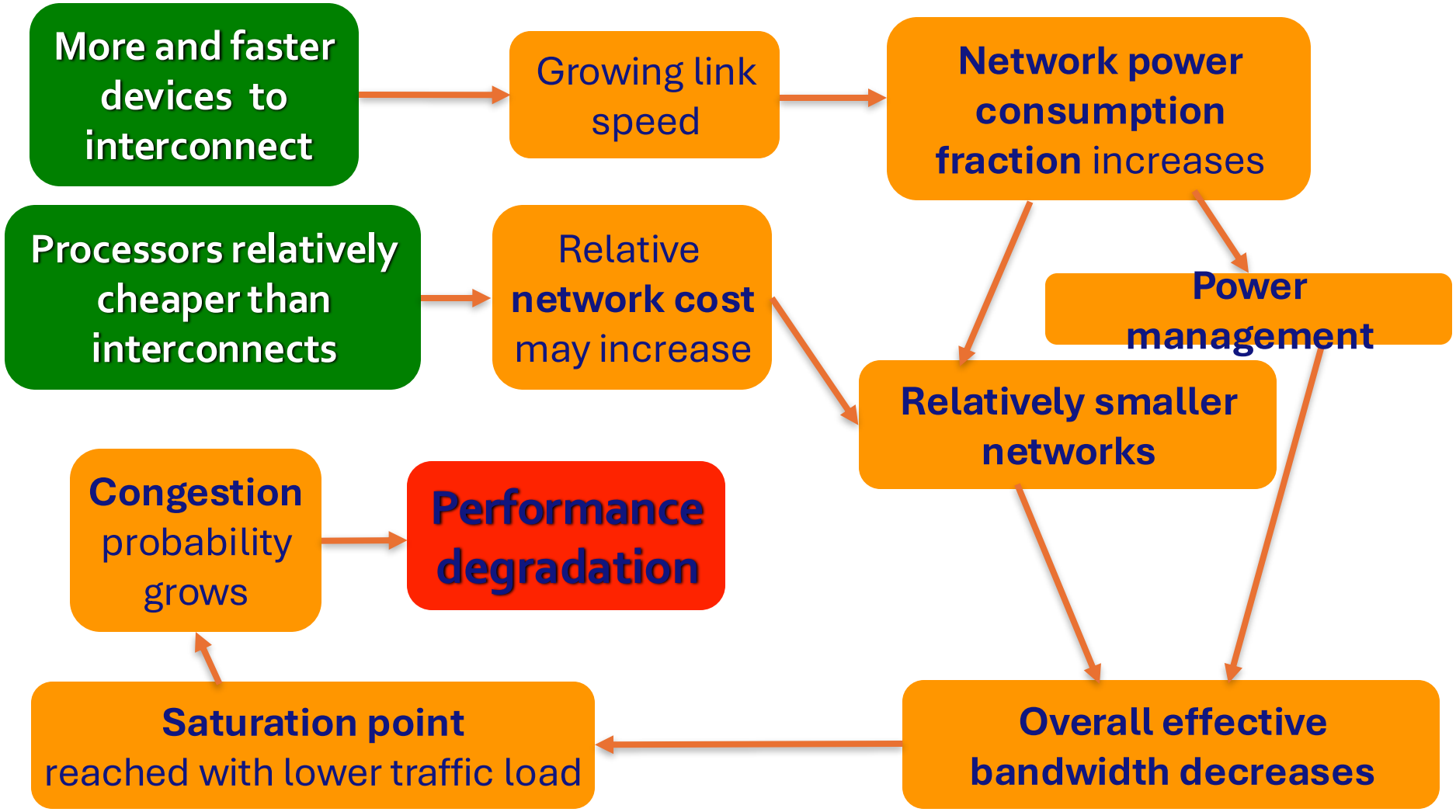
Parking lot

Input-buffer local arbitration per switch is fair



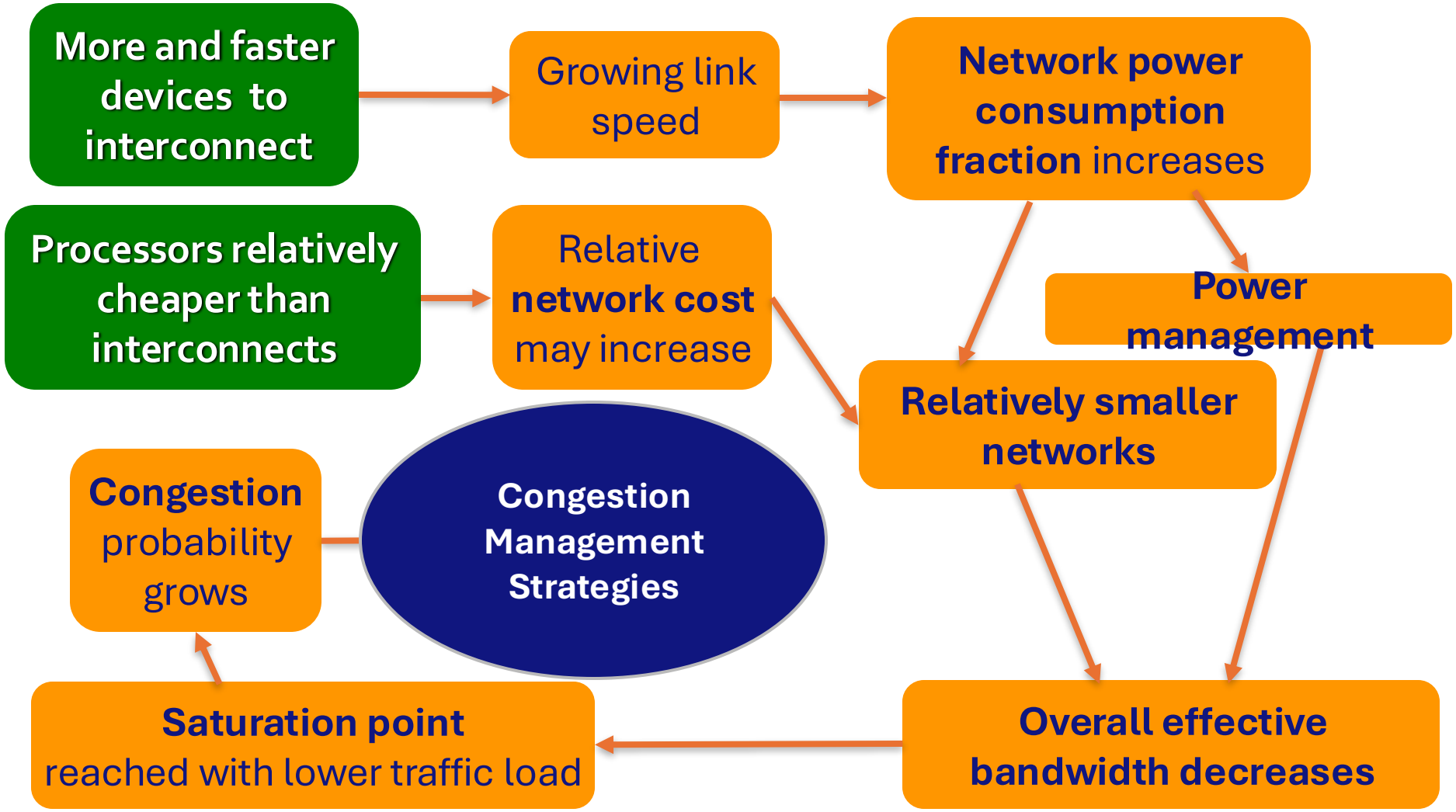
However, from a network perspective, it is unfair, and this effect depends on the number of hops to the destination (D)

Congestion Control: The big picture



P. J. García, J. Escudero-Sahuquillo: Interconnection Networks in the Exascale and AI Era: Challenges and Solutions. Academic Salon on High-Performance Ethernet, TUM, Munich, 12-13 March 2025

Congestion Control: The big picture



P. J. García, J. Escudero-Sahuquillo: Interconnection Networks in the Exascale and AI Era: Challenges and Solutions. Academic Salon on High-Performance Ethernet, TUM, Munich, 12-13 March 2025

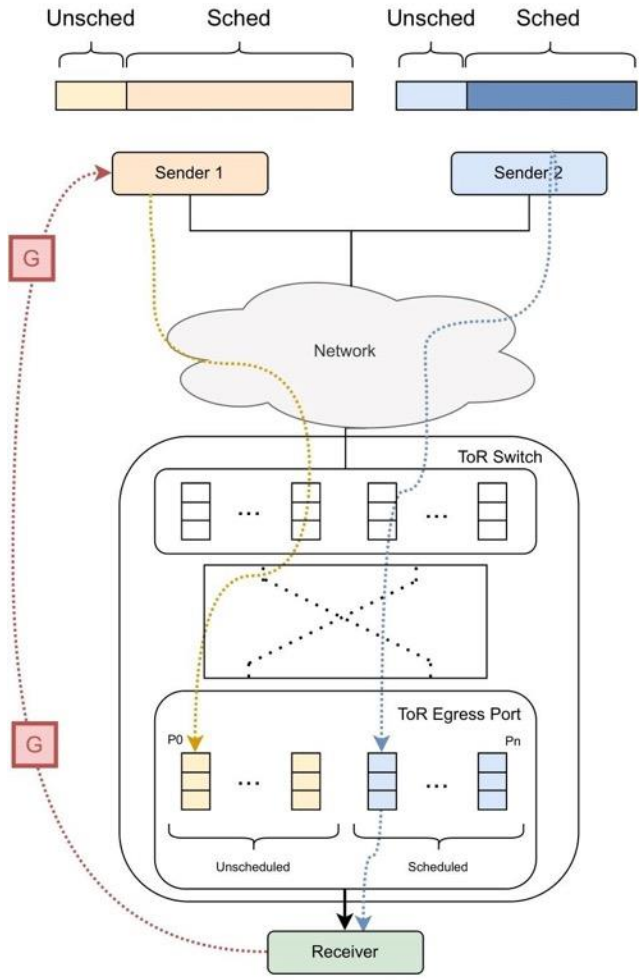
Congestion Control: Solutions to congestion

- Appropriate topologies & routings (including adaptive and LB)
- Packet dropping (and retransmissions)
- Proactive techniques (destination scheduling)
- Reactive techniques (DCQCN, TIMELY, HPCC, telemetry-based…)
- HoL-blocking reduction techniques (static queuing schemes)
- HoL-blocking elimination techniques (a.k.a., Congestion Isolation: IEEE 802.1Qcz)
- Combined approaches

Congestion Control: Solutions to congestion

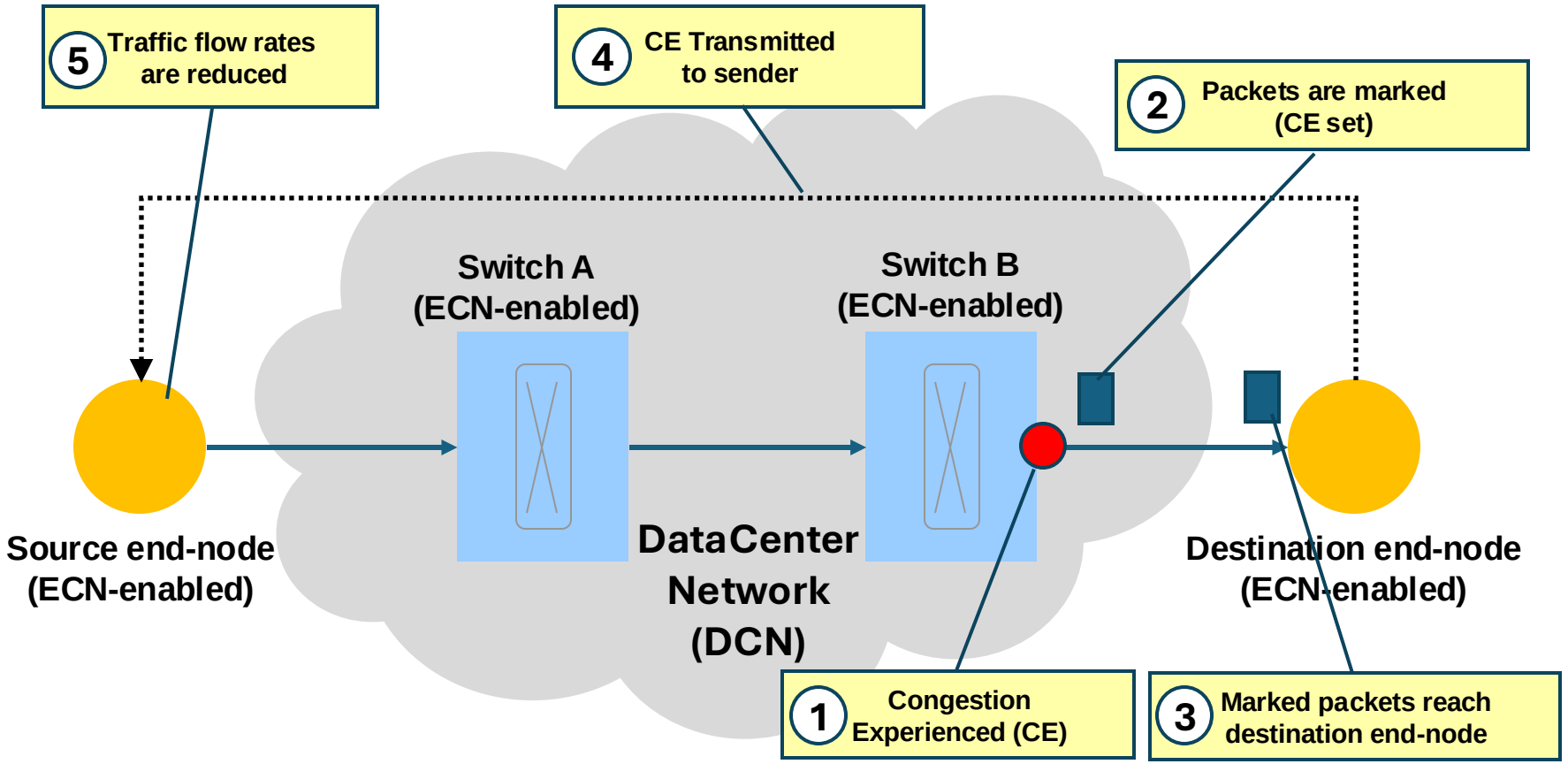
Proactive techniques: HOMA

- Destination-based technique.
- Uses SPRT to prioritize small messages and those with less remaining data.
- Automatically sends data without acknowledgments (unscheduled packets) to notify destinations.
- If a message is small enough, it may only require unscheduled packets.
- The destination provides grants to allow the injection of more scheduled packets.
- The destination's top-of-rack switch isolates packets at the egress, forwarding them to the receiver while distinguishing between scheduled and unscheduled packets.
- To ensure efficient network usage, Homa enables overcommitment.



Congestion Control: Solutions to congestion

Reactive techniques (DCQCN, etc.)



Congestion Control: Solutions to congestion

Reactive techniques (DCQCN, etc.)

- Used in InfiniBand and Ethernet devices **(DCQCN or Timely)**
- Marks packets when an occupancy threshold is exceeded at some switch buffer.
- When the marked packet reaches the destination, the latter will notify the source.
- The source will decrease the injection rate of the flow referred to in congestion notifications.
- **Eliminates congestion trees** from the network (also incast)
- **Disadvantages**:
  - Depending on the traffic pattern and configuration parameters, the mechanism may need to be faster.
  - The number of parameters is large and the configuration of the mechanism is complex.
  - HoL blocking is not avoided until the congestion tree is removed.
  - **Congestion detection is not precise enough so victim packets may end up being marked**

Congestion Control: Solutions to congestion

Reactive techniques (Source Flow control - SFC)

- Top of Rack (ToR) switches directly generate SFC messages when congestion is detected.
- An SFC message (SFCM) is received at a source end node that stops the injection of congesting flows.
- SFC reduces buffer occupancy without an excessive exchange of PFC messages and reduces Head-of-Line (HoL) blocking.
- Again, congestion detection is critical

Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)

- Congestion detection is the key starting point when it comes to identifying congesting flows.
- Many of the techniques mentioned before lack a precise congestion detection mechanism.
- ECP can accurately detect a congestion situation and identify with minimal error the packets contributing to congestion if they fulfill three conditions:
  1. The queue detection threshold is exceeded.
  2. It selects the packet placed at the head of the congested queue
  3. If that packet is ready to cross and was not selected by the switch at least once
- If we assume that packets identified by ECP as congesting are marked, then ECP also defines a re-evaluation mechanism that removes the marks for all packets behind the congesting one.

Cristina Olmedilla, Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Quiles, Wenhao Sun, Long Yan, Yunping Lvu, José Duato:A New Mechanism to Identify Congesting Packets in High-Performance Interconnection Networks. HOTI 2024: 24-32

Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)

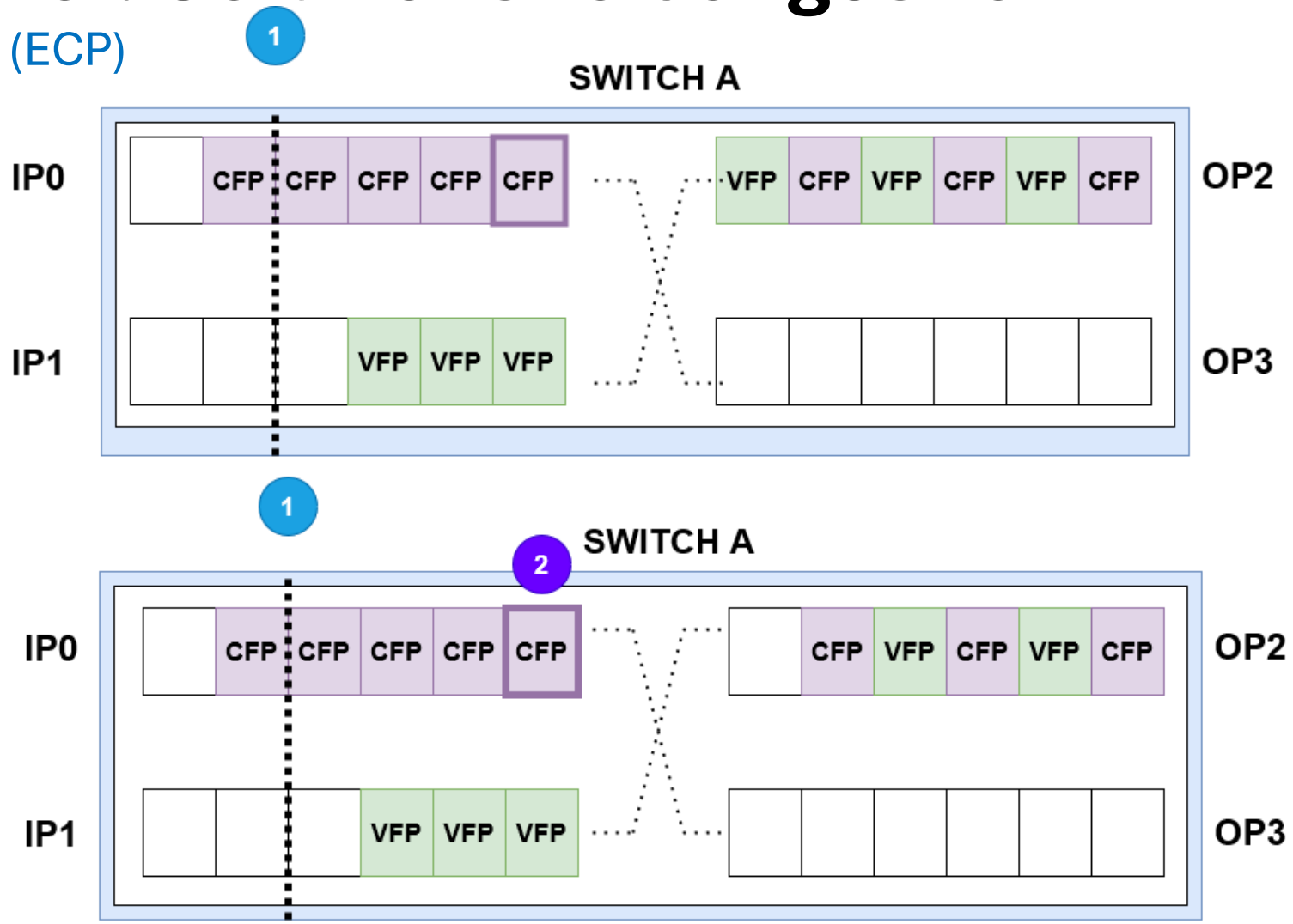
CFP Congesting-Flow Packet

VFP Victim-Flow Packet

CFP Congesting-Flow Marked Packet

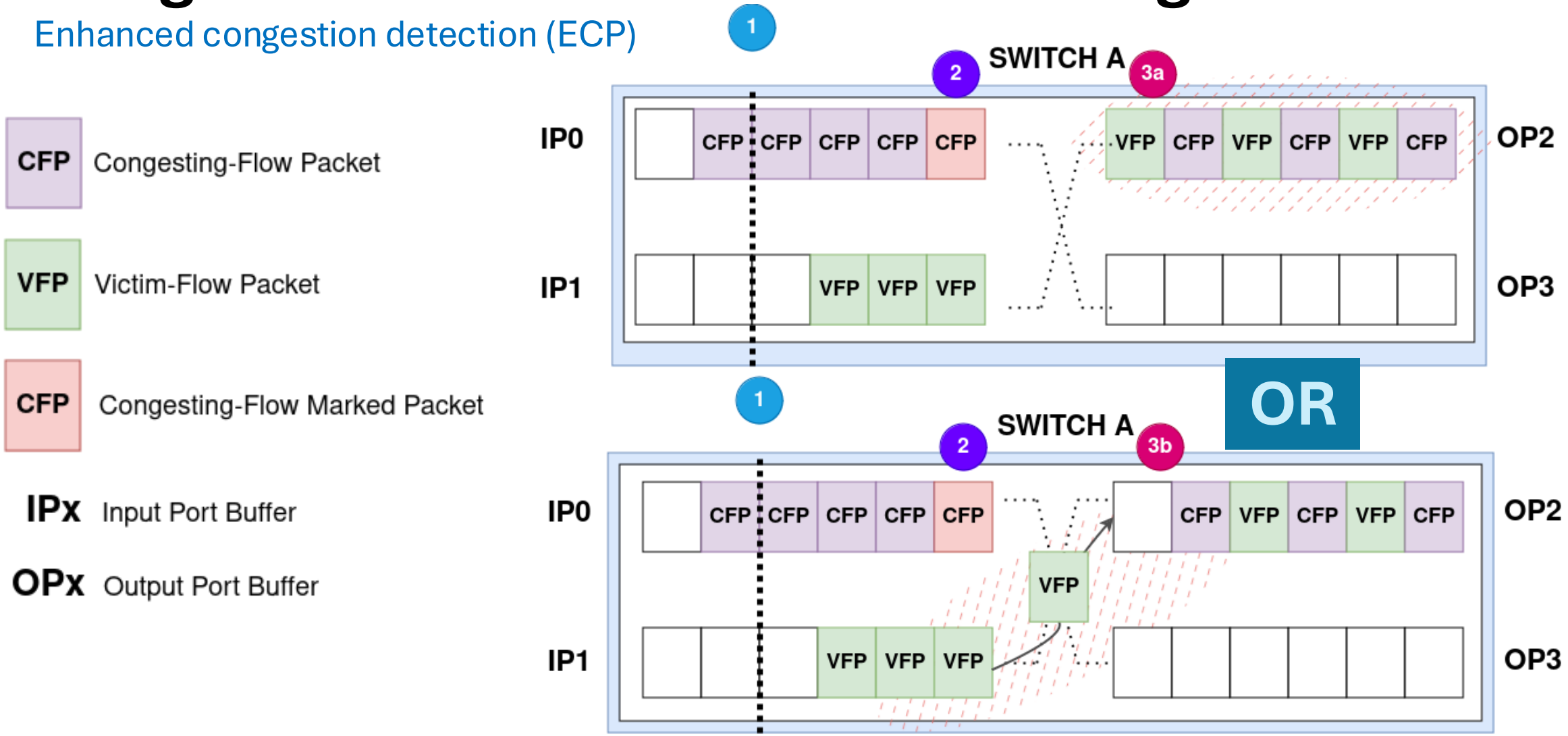
IPx Input Port Buffer

OPx Output Port Buffer



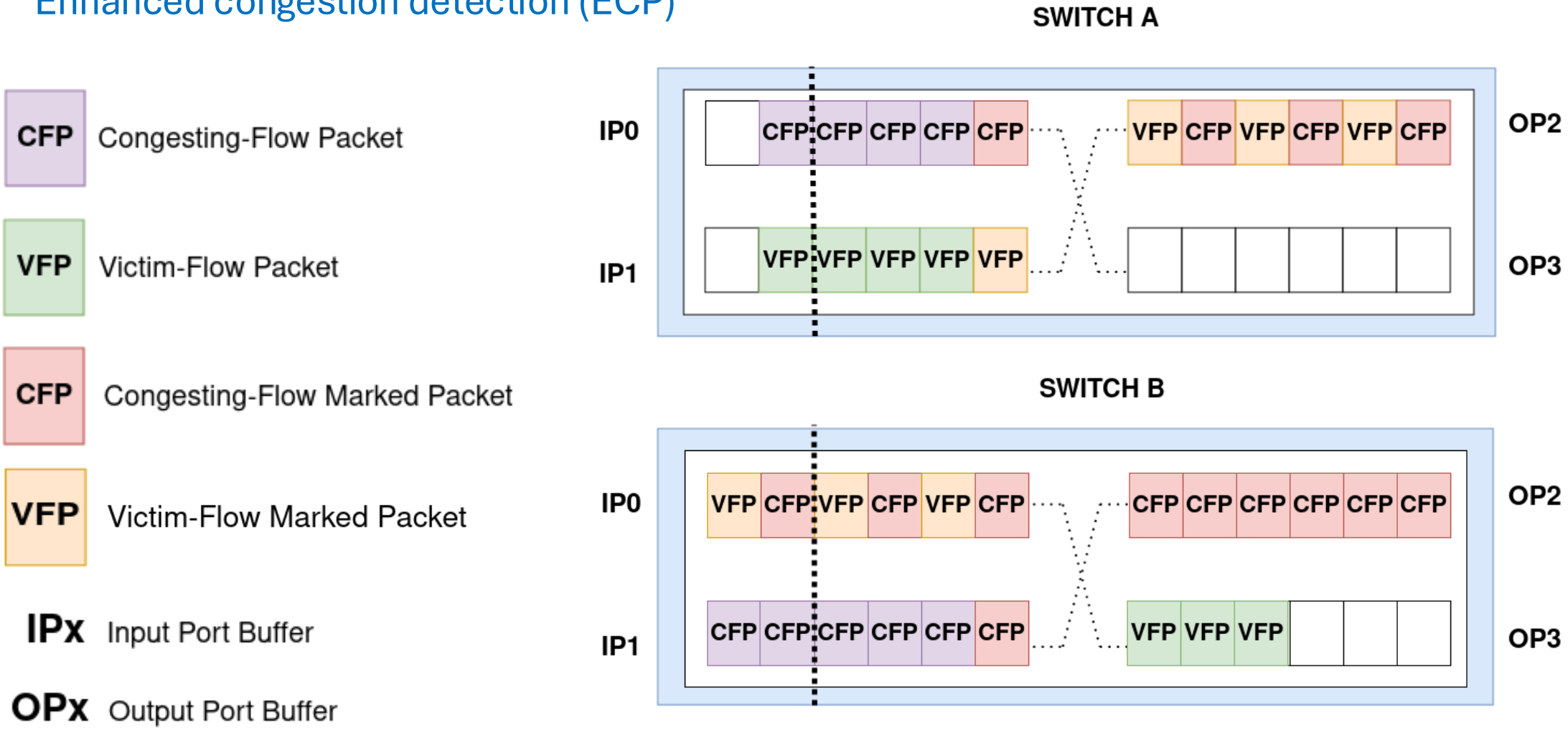
Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)



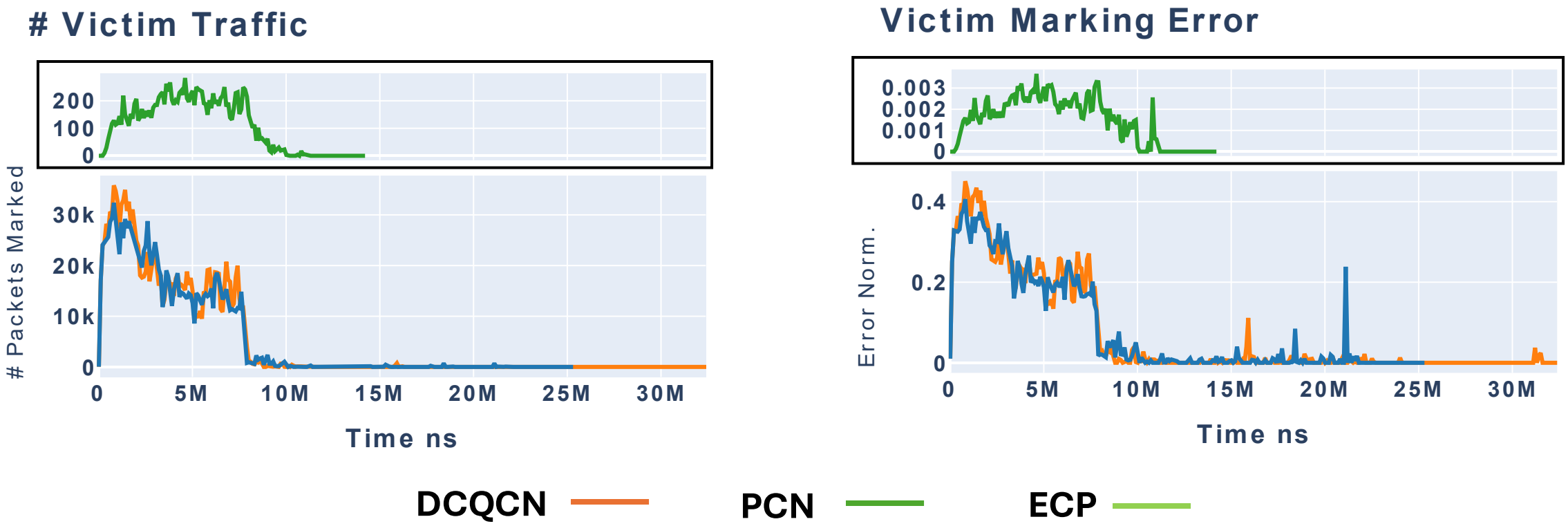
Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)



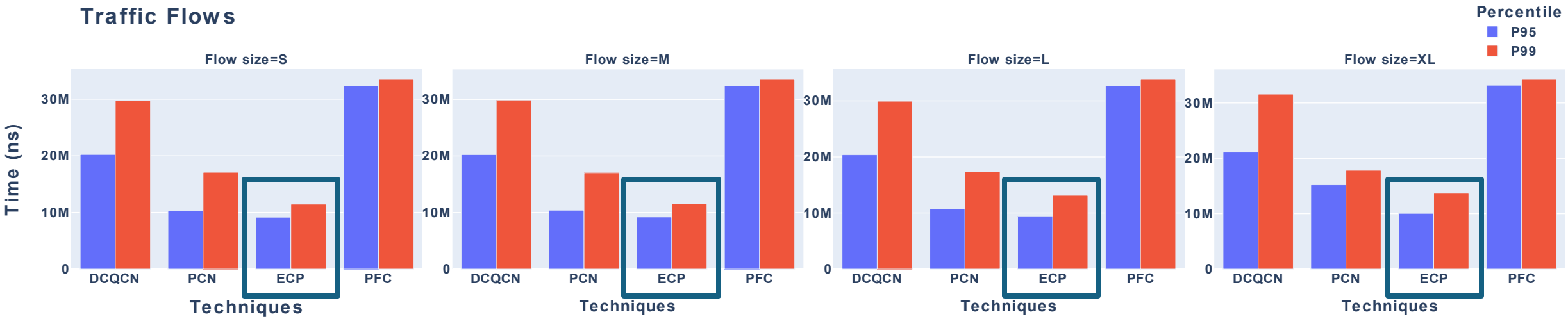
Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)



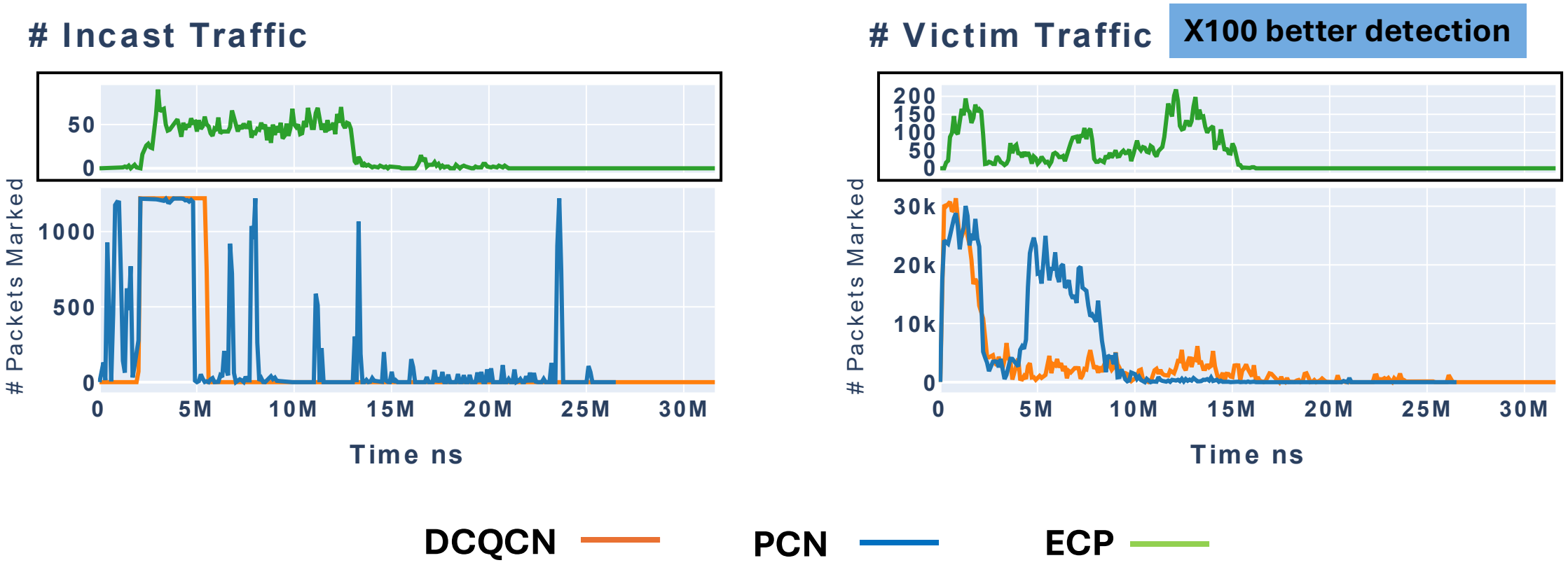
Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)



Congestion Control: Solutions to congestion

Enhanced congestion detection (ECP)



Congestion Control: Solutions to congestion

HoL blocking reduction: static queuing schemes (SQS)

- **Partially remove HoL blocking** by isolating congesting flows in queues or virtual channels (VCs).
- Packets are stored in the VCs according to a **mapping policy.**
  - Mapping policies can be topology-aware or agnostic
- **Disadvantages:**
  - In large systems with many nodes**, a victim flow may share a VC with a congesting flow**, in which case HoL blocking will not be avoided.
  - The **parking lot problem is not solved.**

Congestion Control: Solutions to congestion

HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.

- Detect a congestion situation immediately when it is produced in a fast and local manner
- Identify the congesting flows contributing to generating that congestion situation
- Store packets belonging to congesting flows into the congesting flow queues (CFQs).
- Keep track of congesting flows in a local table for each switch.
- Update table based on incoming packets.

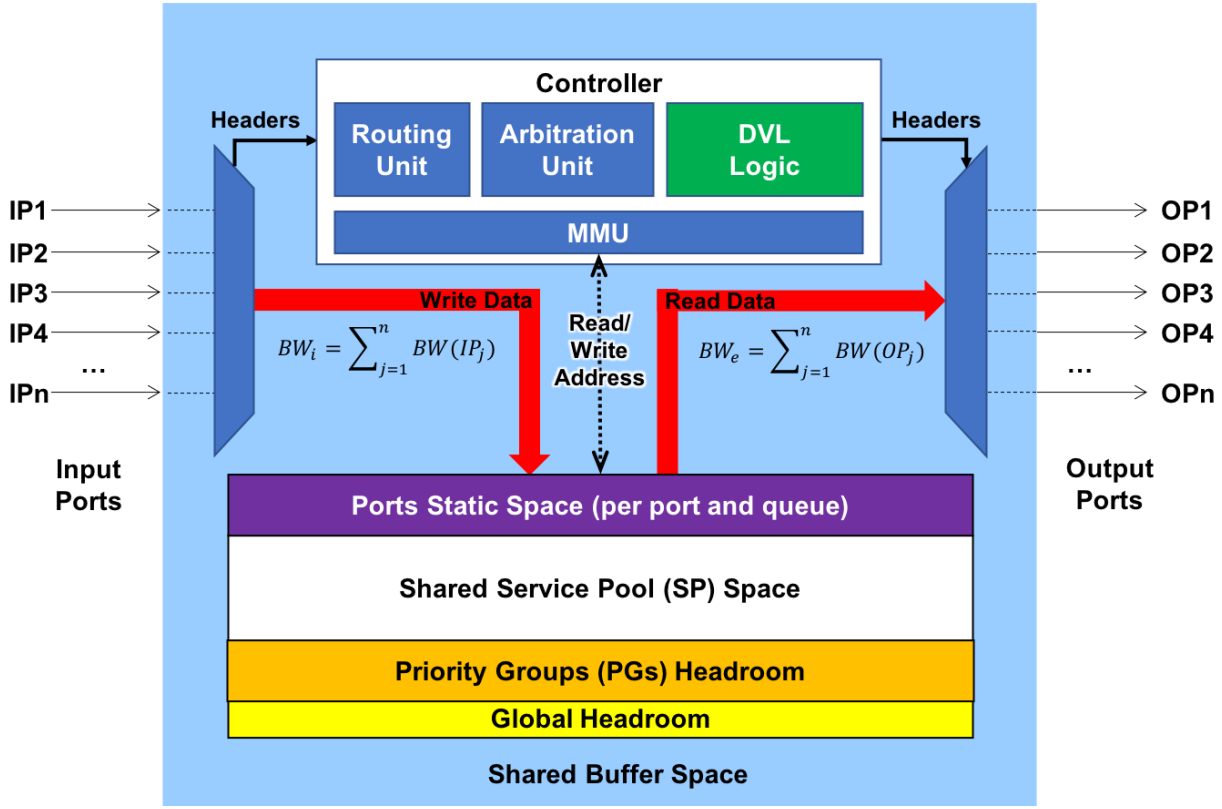
Luis Gonzalez-Naharro, Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco José Quiles Flor, José Duato, Wenhao Sun, Li Shen, Xiang Yu, Hewen Zheng: **Efficient Dynamic Isolation of Congestion in Lossless DataCenter Networks**. NEAT@SIGCOMM 2019: 15-21

Cristina Olmedilla, Jesús Escudero-Sahuquillo, Pedro Javier García, Francisco J. Alfaro-Cortés, José L. Sánchez, Francisco J. Quiles, Wenhao Sun, Xiang Yu, Yonghui Xu, José Duato: **DVL-Lossy: Isolating Congesting Flows to Optimize Packet Dropping in Lossy Data-Center Networks**. IEEE Micro 41(1): 37-44 (2021)

Congestion Control: Solutions to congestion

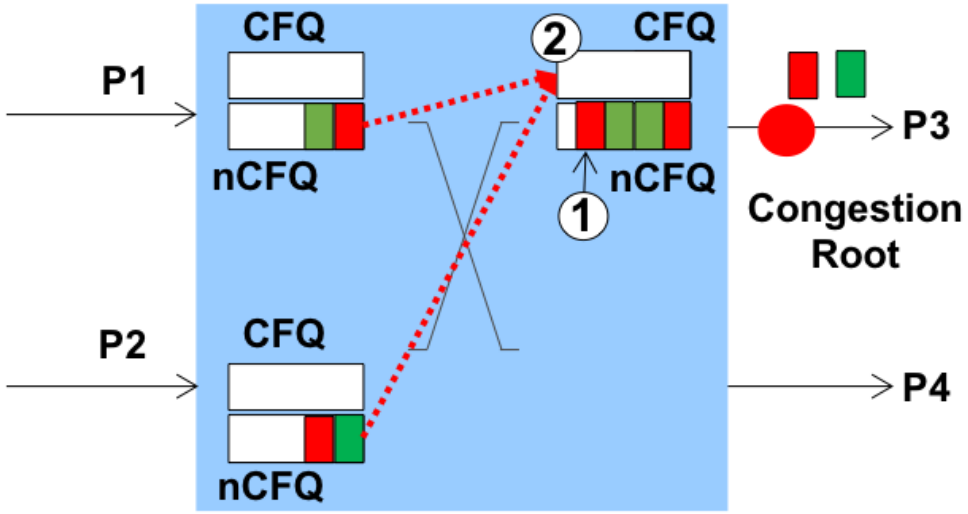
HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.

- DVL operates on top of **shared-buffer switches:** packets stored in a centralized memory.
- Memory filling order: Static space → Shared Pool → PG Headroom → Global headroom



Congestion Control: Solutions to congestion

HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.



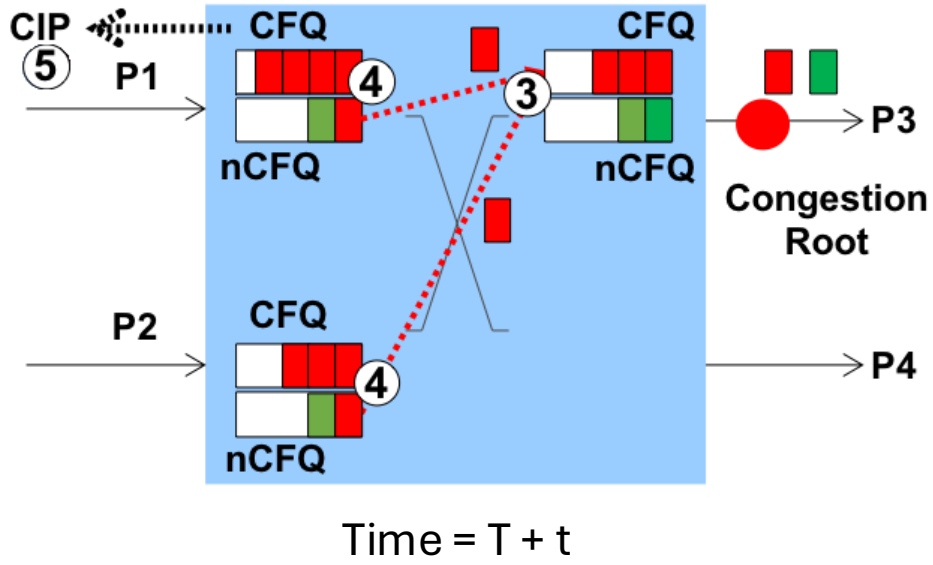
Time = T

- #1: Congestion detected when nCFQ reaches a threshold.
  - The packet on the top is assumed to be responsible for congestion
  - A new entry for the congestion root is added to the DVL logic (congested flow table)
- #2: CFQ is allocated.

Congestion Control: Solutions to congestion

HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.

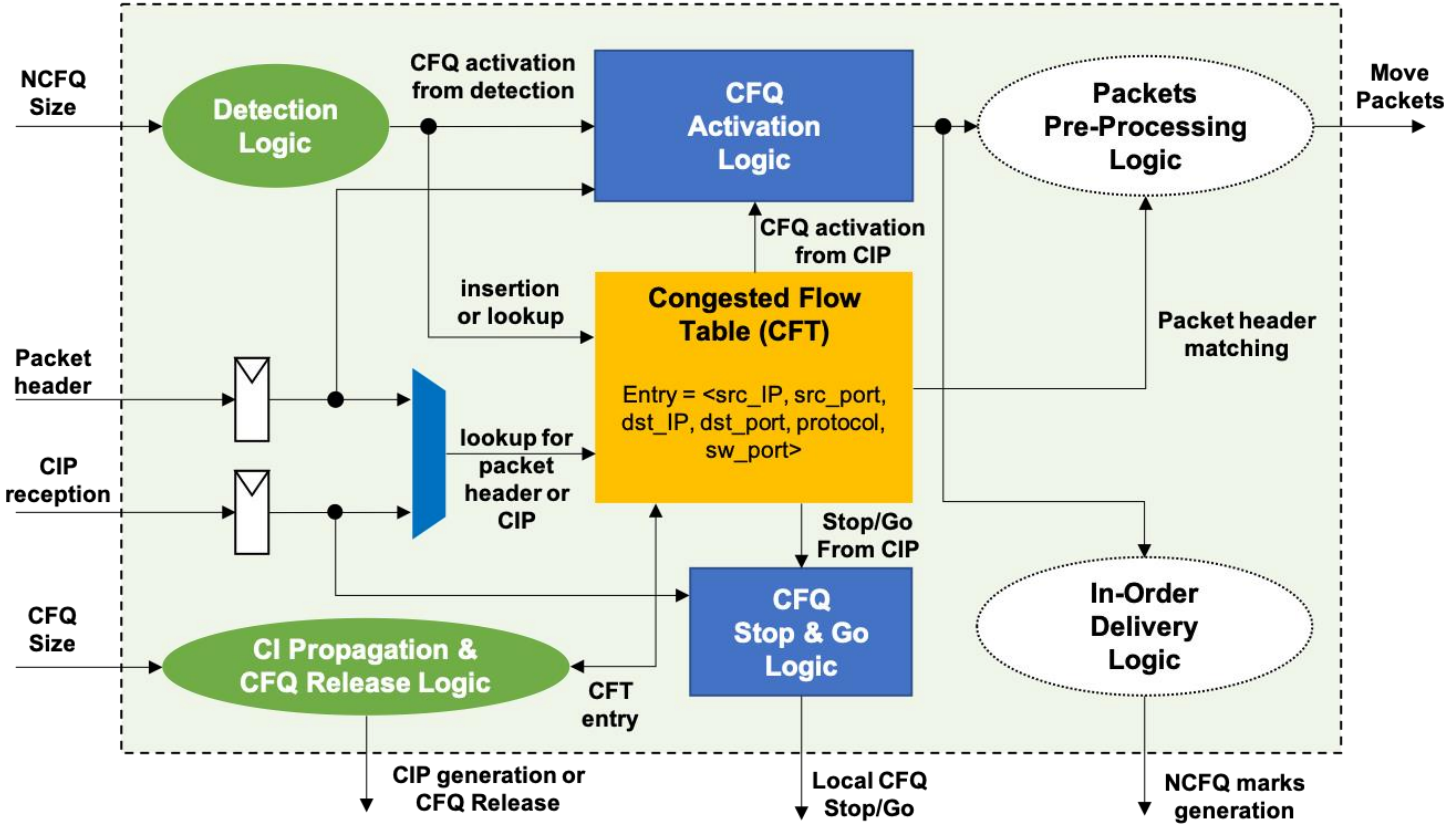
- #3: Incoming congested packets (egress) stored at CFQ.
- #4: Egress CFQ grows, CFQ allocated at ingress.
- #5: Ingress CFQ grows, and congestion information is sent upstream.



Congestion Control: Solutions to congestion

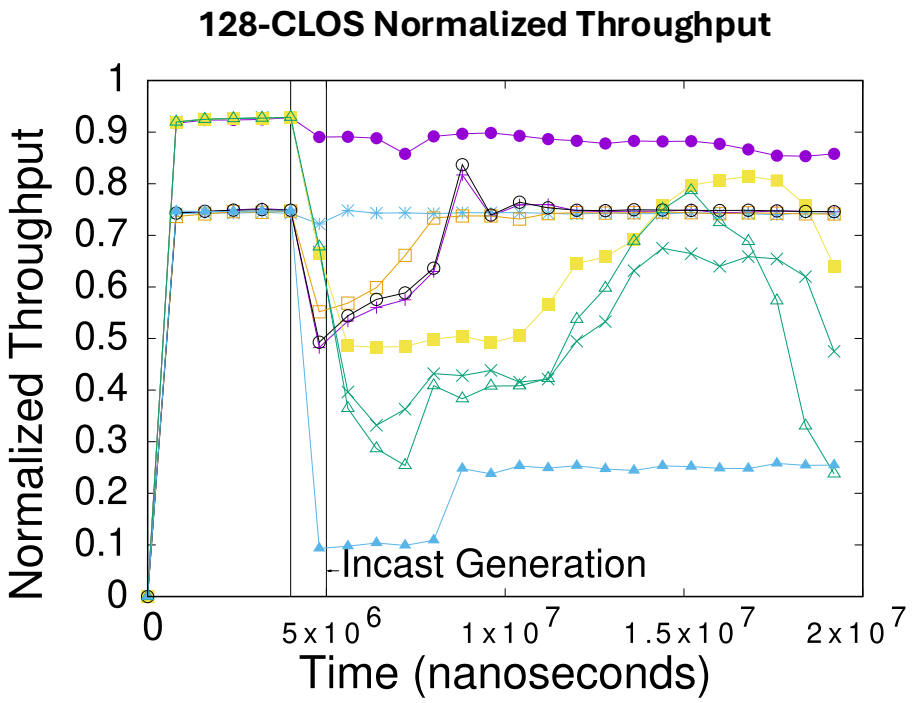
HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.

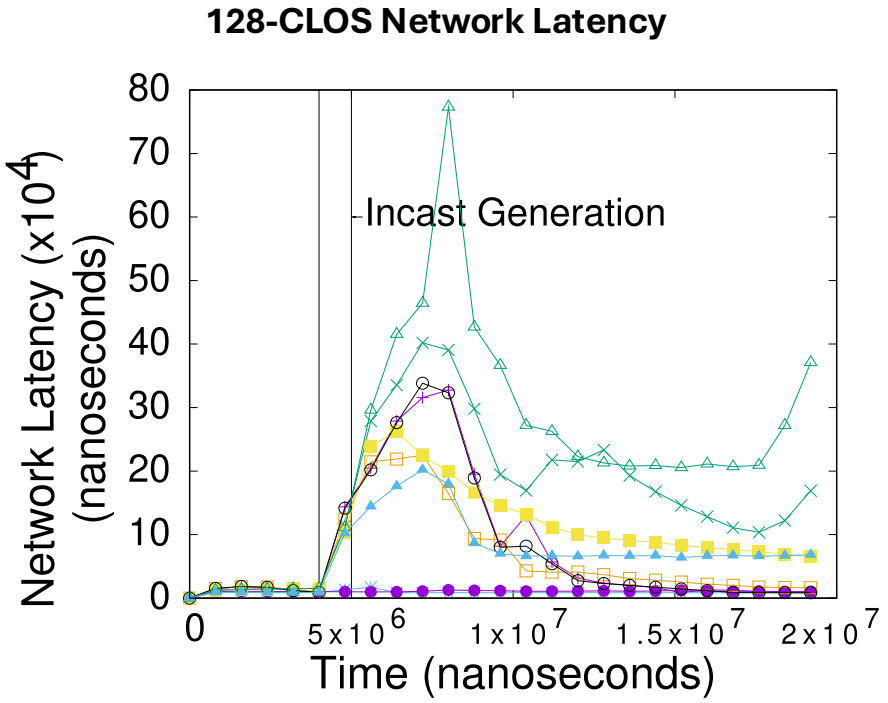
- **Congested flow table (CFT):**
  - keeps information regarding congested flows (source and destination IPs and ports, protocol and switch port).
  - When congestion is detected, the packet header is used to fill a CFT entry.
  - Pre-processing: incoming packets matching an entry will be stored in the CFQ.



Congestion Control: Solutions to congestion

HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.





DCQCN
DCQCN-RT
DVL-Lossy
DVL-Lossy-RT
ECN
ECN-RT
NoCC
NoCC-RT
PFC

Congestion Control: Solutions to congestion

HoL blocking elimination: Congestion Isolation (CI), DVL, RECN, etc.

- **Advantages:**
  - CI allows for a complete HoL blocking elimination, whenever there are enough entries in the CFT.
  - Reacts immediately and locally to congestion based on recent information
- **Main drawbacks**:
  - It's challenging to deal with multiple hotspot situations.
  - If the number of congestion roots is high, it may run out of entries in the local table.
  - Congestion detection may generate false positives if not accurate enough
  - Additional complexity

Congestion Control: Final remarks

- **Accurate congestion detection** is crucial for conveniently identifying congesting packets.
- **Immediate, local, and fast reaction** to congestion is also needed to eliminate HoL-blocking.
  - Otherwise, congestion control mechanisms may react later and based on obsolete information.
- When adaptive routing is used, we need to **distinguish incast from in-network congestion** situations so incast contributing flows are routed using deterministic routing to prevent congestion from spreading.
- A combination of different CC strategies may help to alleviate congestion-derived problems, since this approach benefits from all the advantages of these techniques at the same time.

Agenda

- Introduction
- Current (well, Classical too…) Challenges in Interconnects
- Solutions (well, at least, some solutions…)
- **Open issues**

Open issues

- Extend the combined approaches to adaptive routing, accurate congestion detection, accurate reaction, and congestion isolation
- Power management + congestion control
- Modeling of traffic patterns from LLMs (into VEF traces framework)
  - Analyze further collective optimizations.
- Impact of intra-node communication in inter-node communication and vice versa
- A “new world” to explore → inter-CPD networking