Deadlock-free routing for Full-mesh networks without using virtual channels

A. Cano, C. Camarero, C. Martínez, R. Beivide
Computer Architecture Research Group

August 20, 2025

University of Cantabria
(Spain)

UC

Outline

Introduction

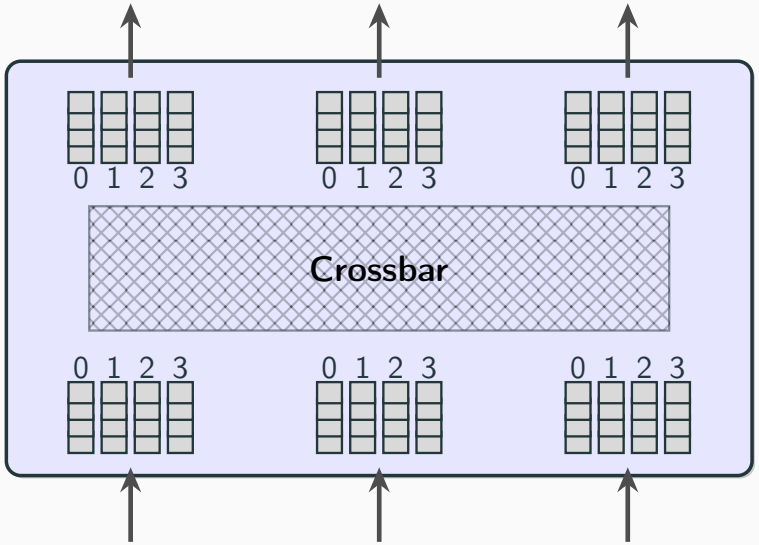
Virtual Channels in Switches

Uses of Virtual Channels

- Head-of-Line Blocking
- **QoS** support
- Routing **deadlock**

Cost of Virtual Channels

- **Area**
- **Power**
- Extra **logic**



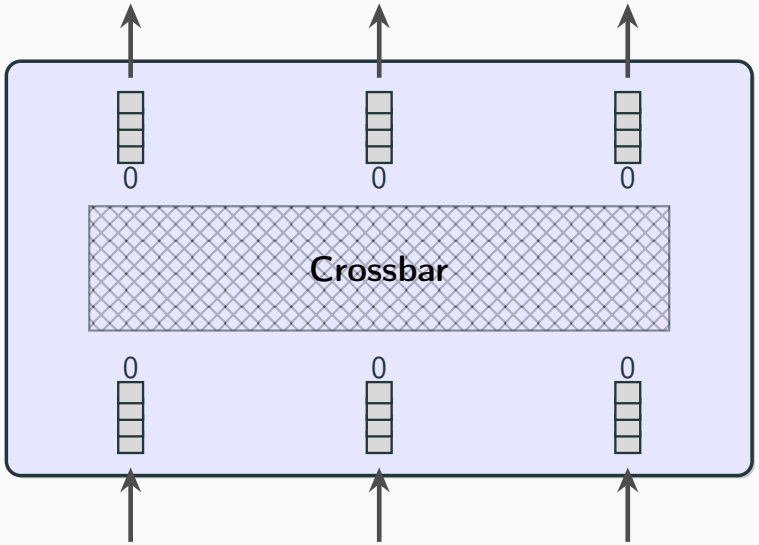
Virtual Channels in Switches

Uses of Virtual Channels

- Head-of-Line Blocking
- **QoS** support
- Routing **deadlock**

Cost of Virtual Channels

- **Area**
- **Power**
- Extra **logic**



Key Achievements

No VC

Deadlock-free routing in Full-Mesh networks **without Virtual Channels**.

100%

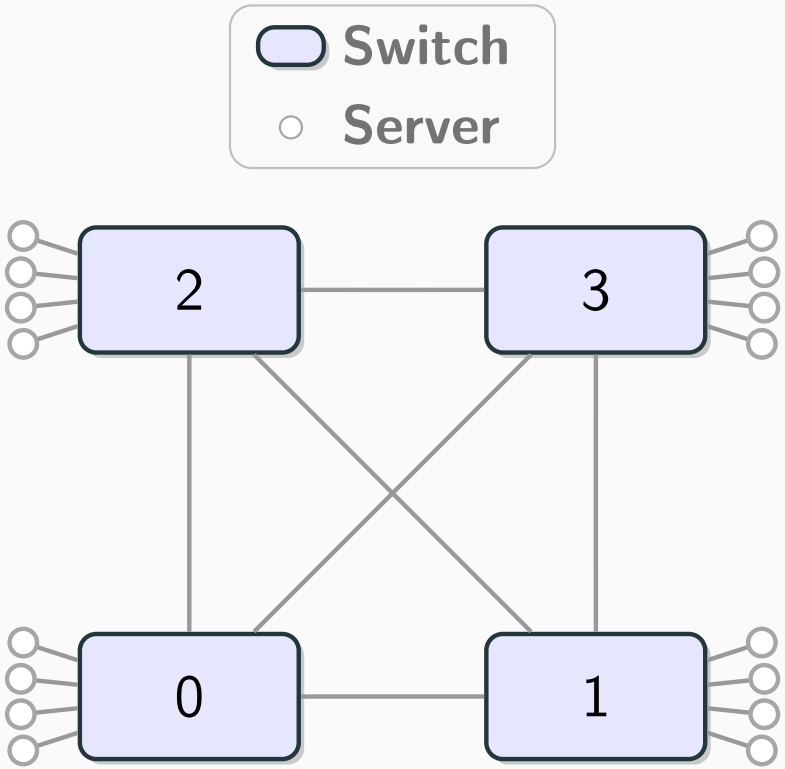
Throughput gain over the previous state-of-the-art routing algorithm.

✓

Successfully adapted to a 2D HyperX network.

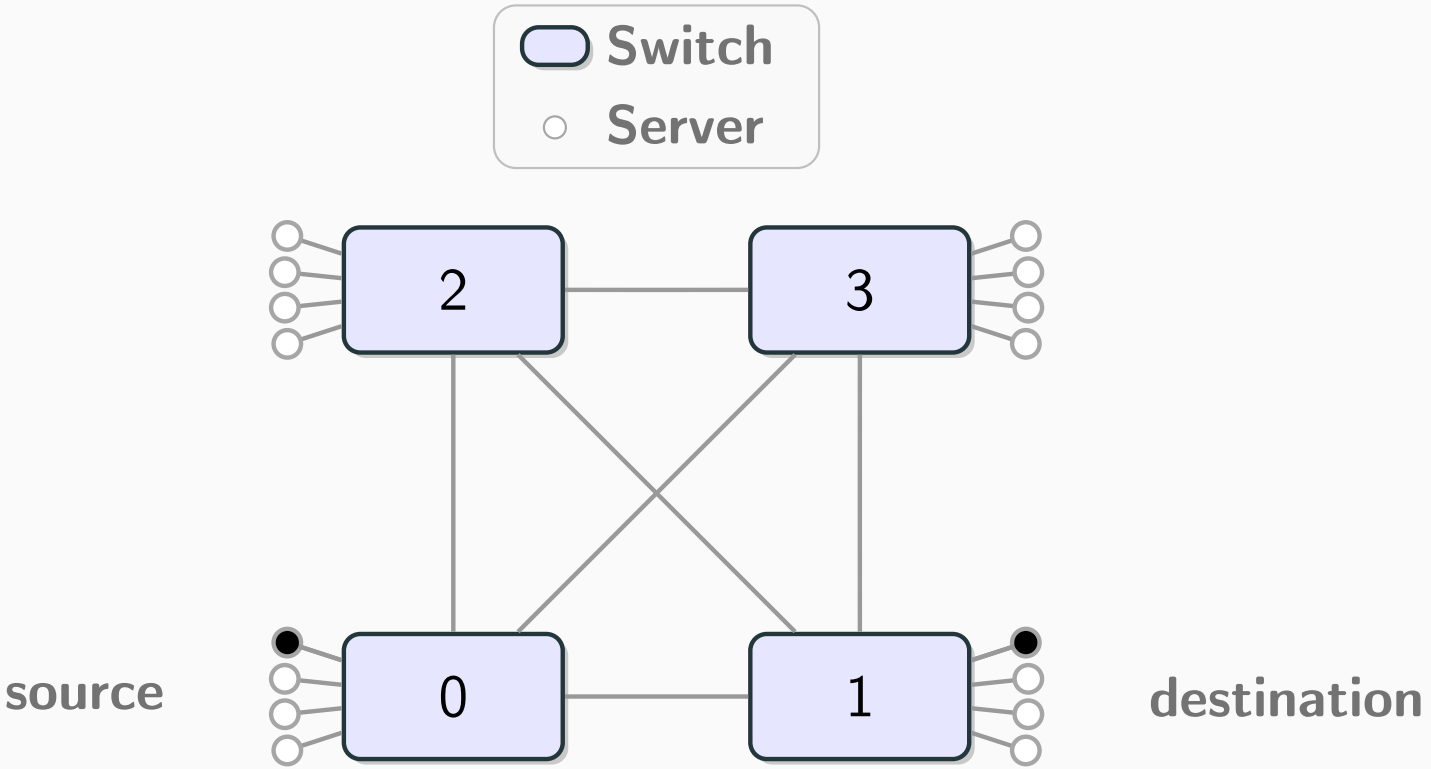
Full-mesh topology

8 of the top 10 supercomputers on the TOP500 list use network topologies based on Full-Mesh.



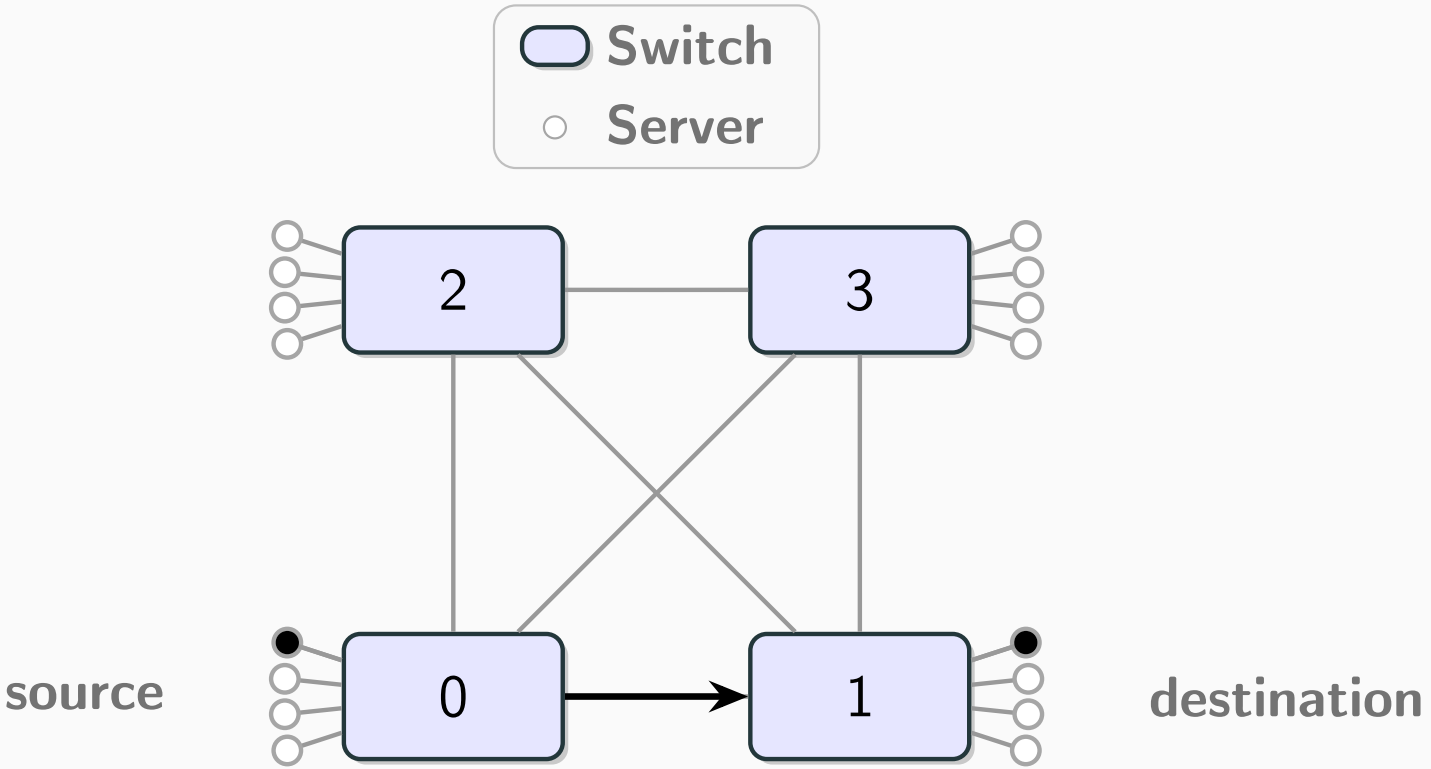
Full-mesh topology

8 of the top 10 supercomputers on the TOP500 list use network topologies based on Full-Mesh.



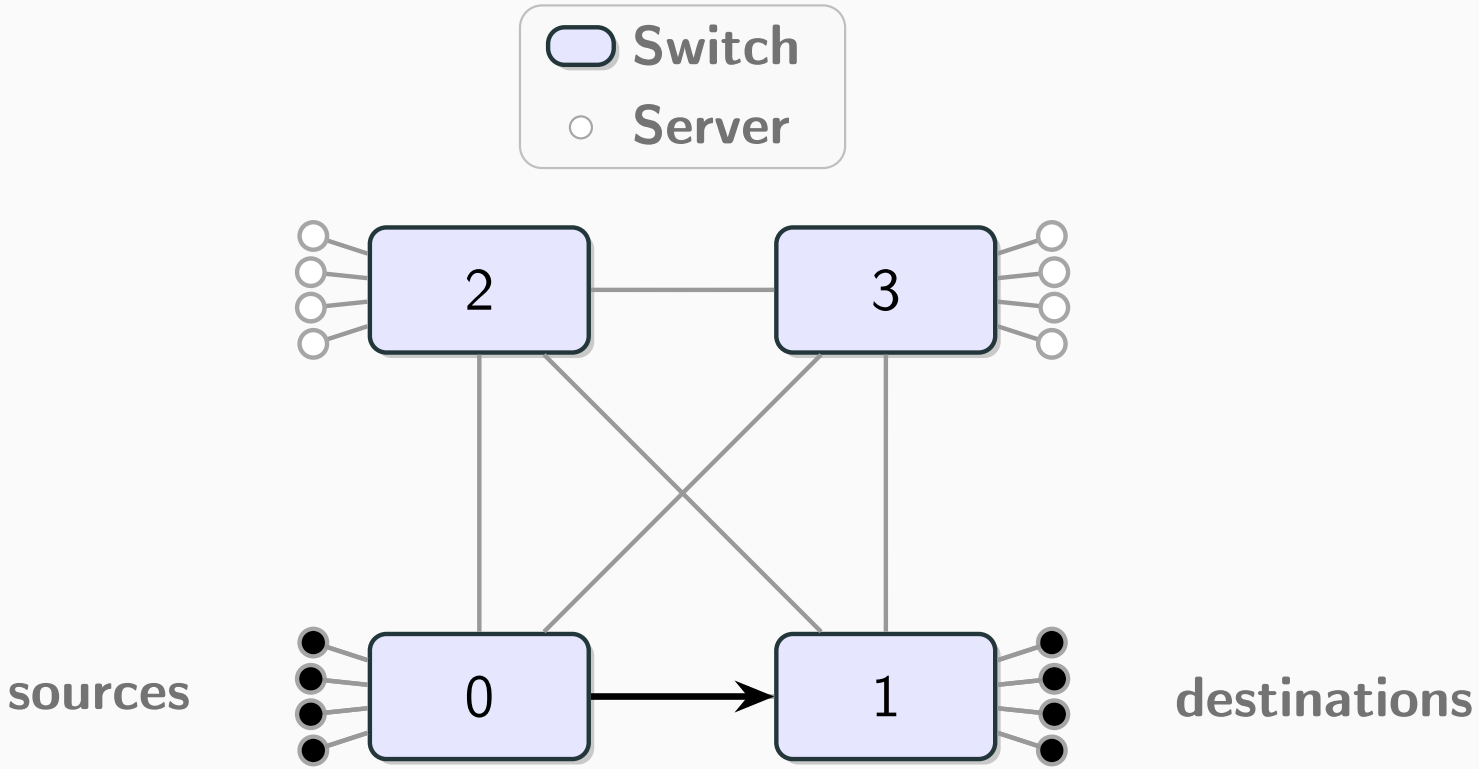
Full-mesh topology

8 of the top 10 supercomputers on the TOP500 list use network topologies based on Full-Mesh.



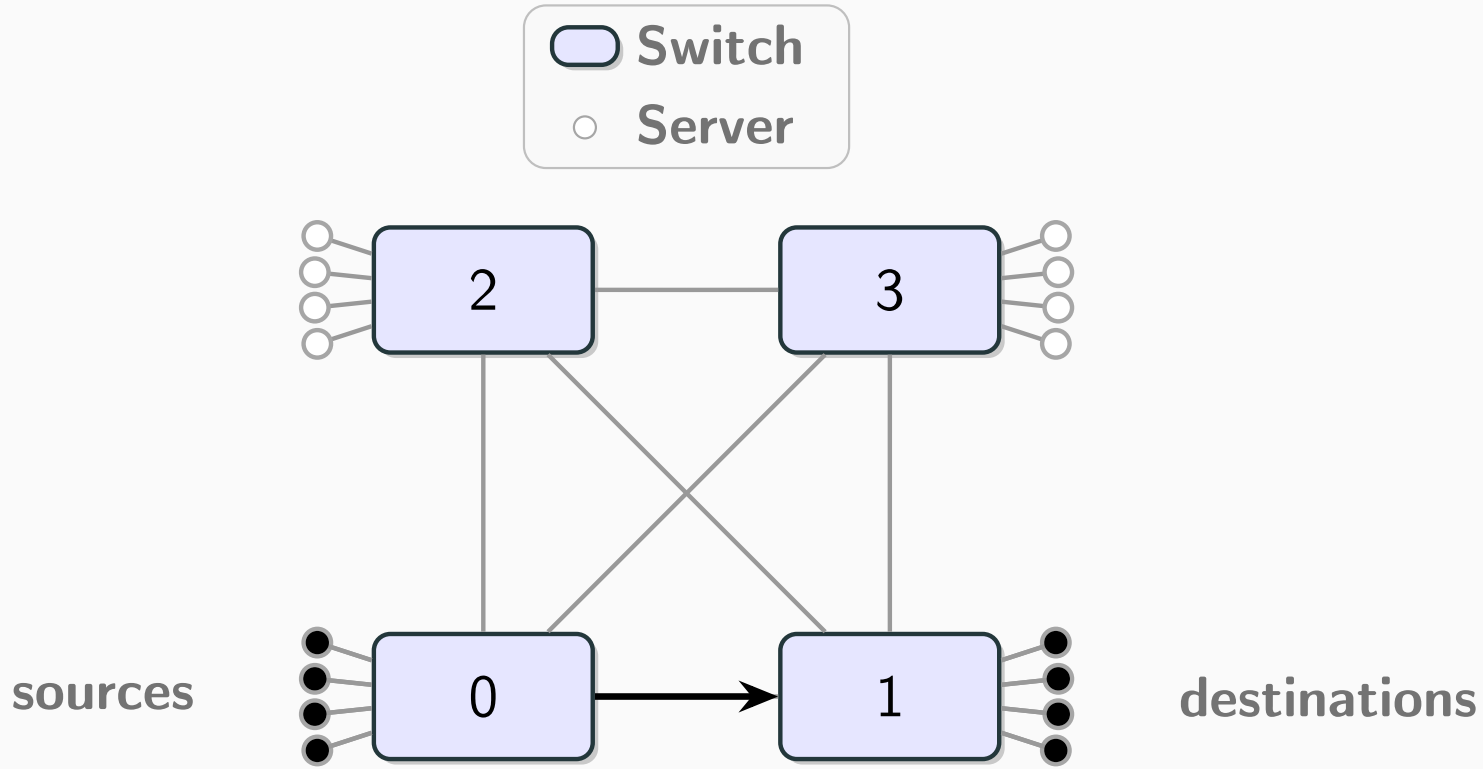
Full-mesh topology

8 of the top 10 supercomputers on the TOP500 list use network topologies based on Full-Mesh.



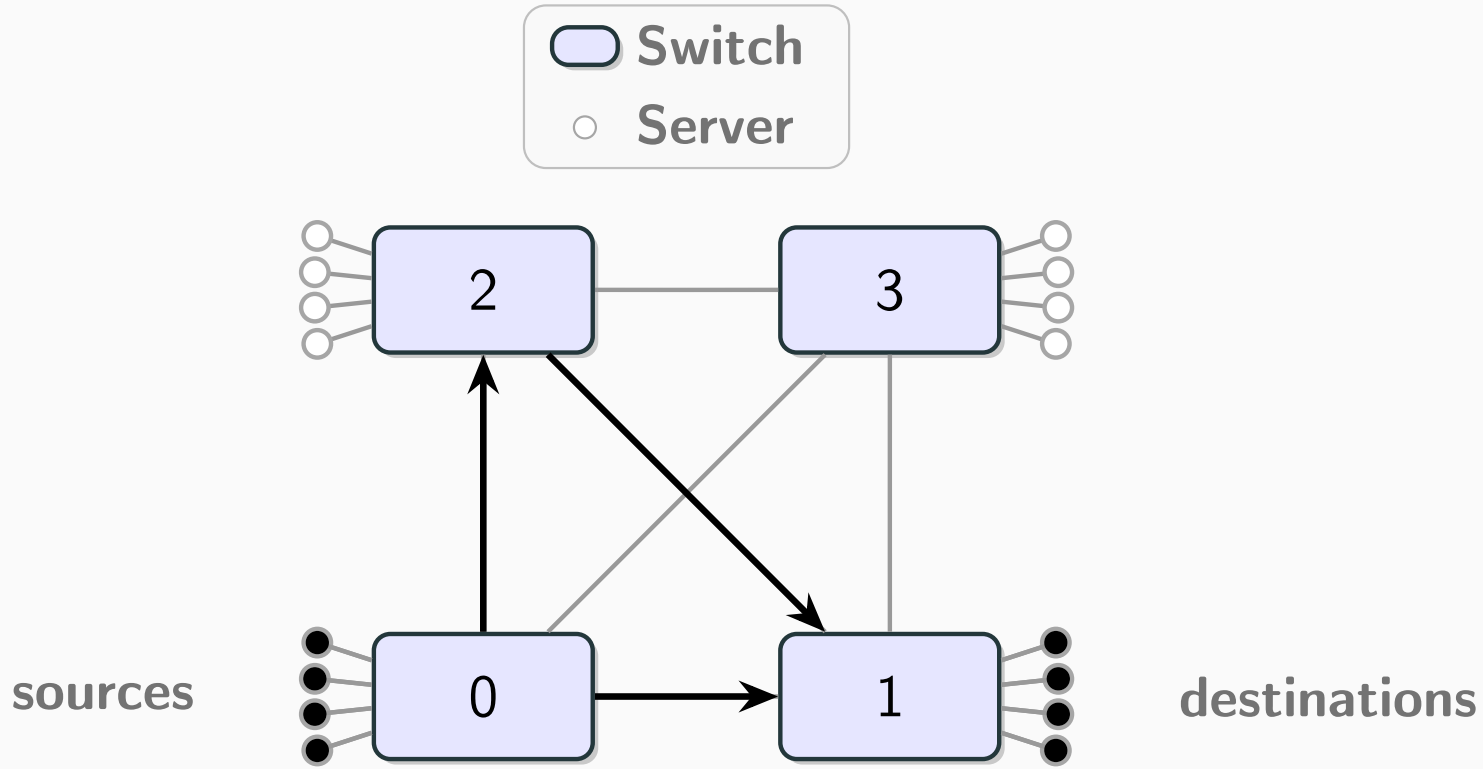
Routing en Full-Mesh

Non-Minimal Routing: Improving performance with 2-hop paths.



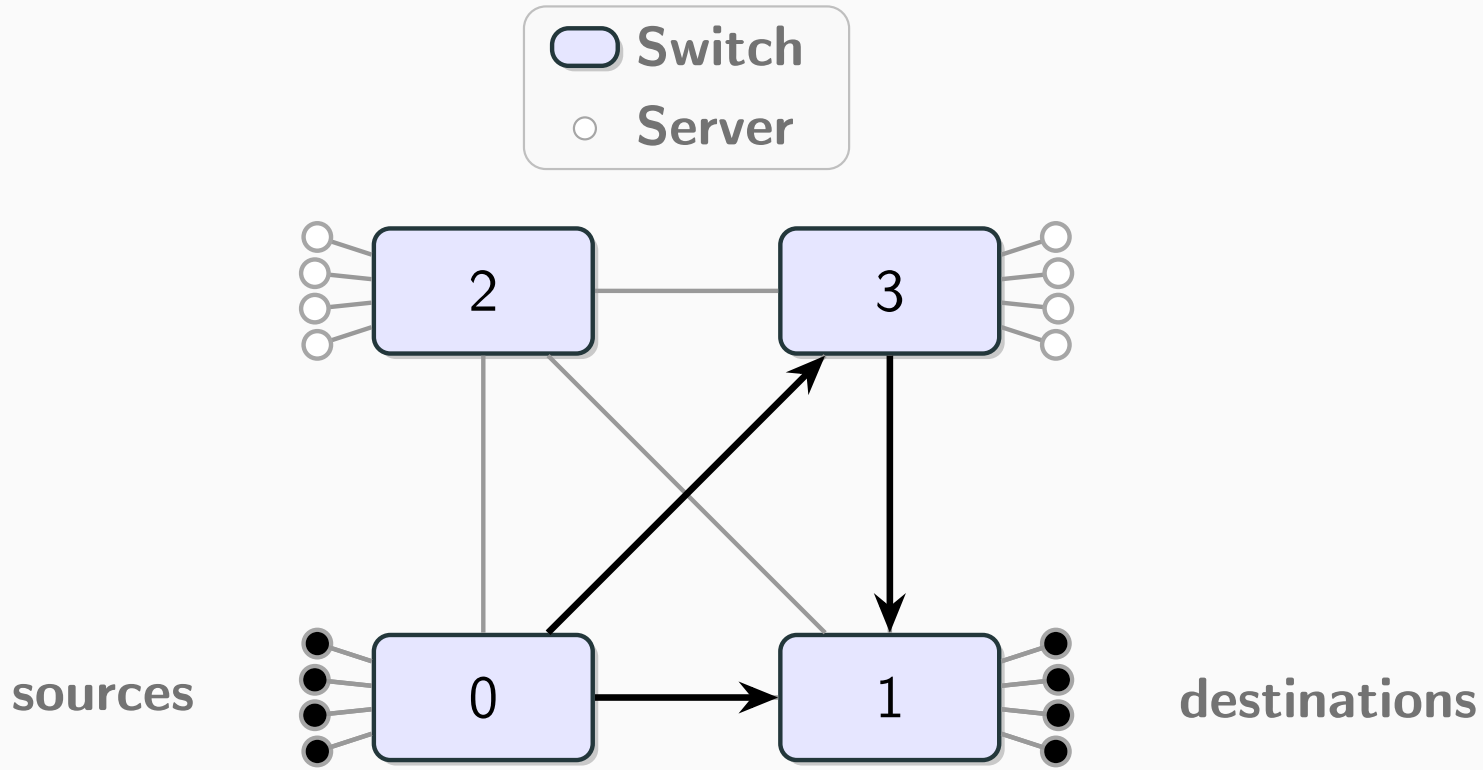
Routing en Full-Mesh

Non-Minimal Routing: Improving performance with 2-hop paths.



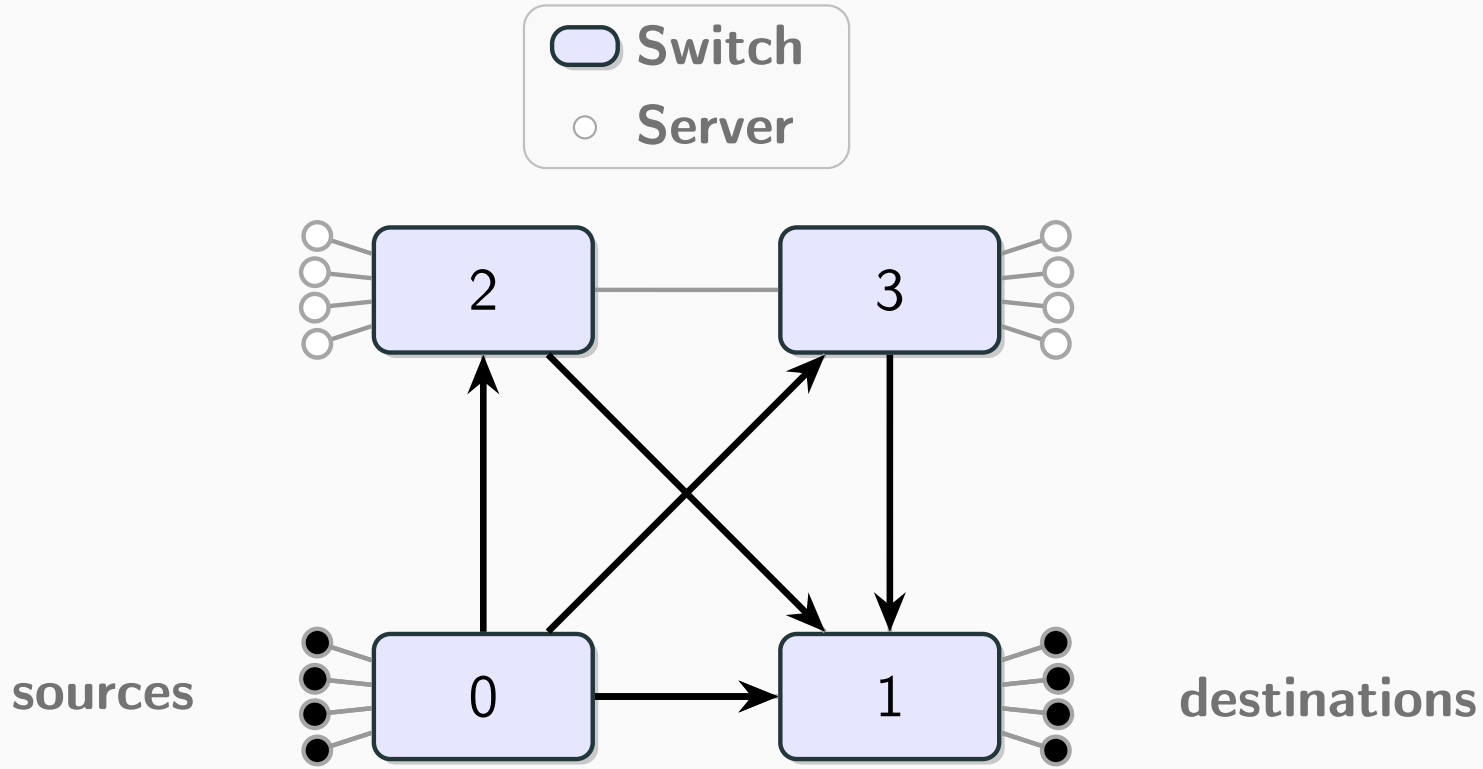
Routing en Full-Mesh

Non-Minimal Routing: Improving performance with 2-hop paths.



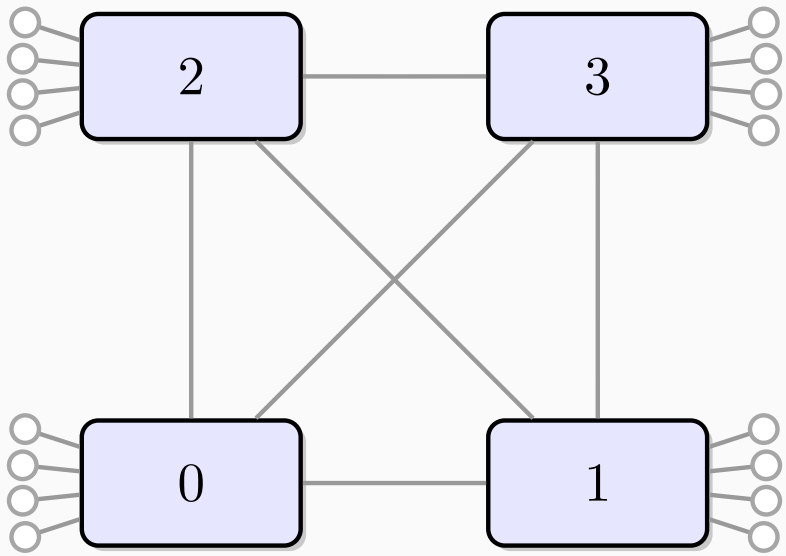
Routing en Full-Mesh

Non-Minimal Routing: Improving performance with 2-hop paths.



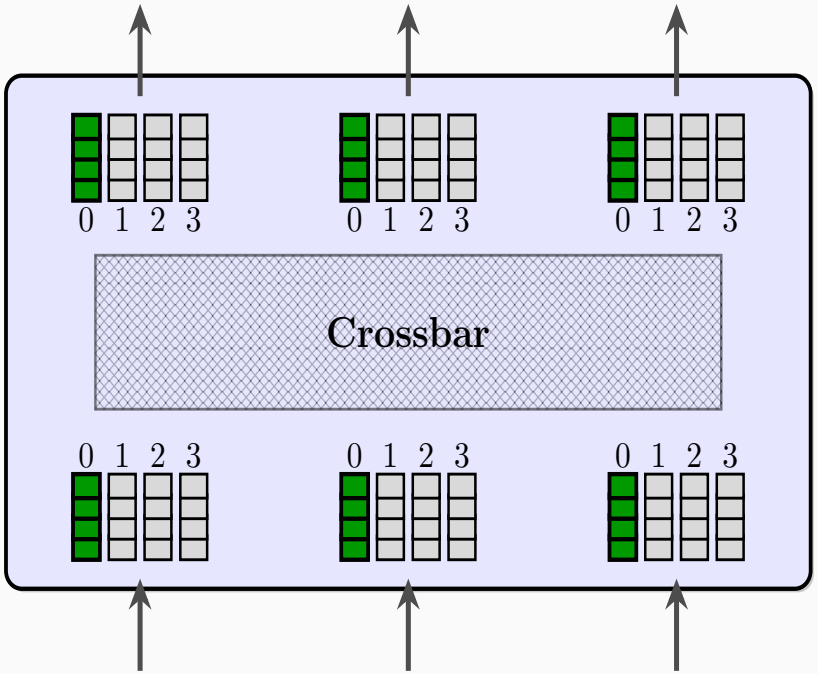
Routing en Full-Mesh

Problem: **routing-deadlock**.

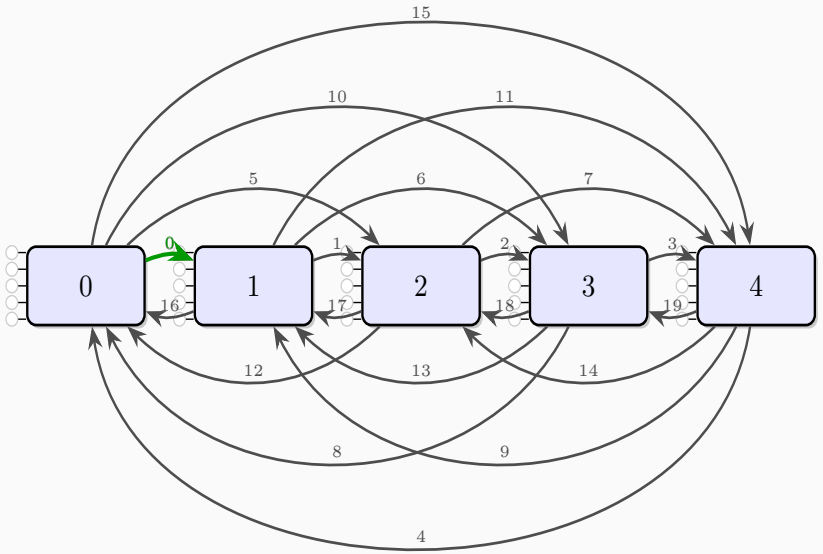


Routing-deadlock representation

Deadlock Avoidance Methods

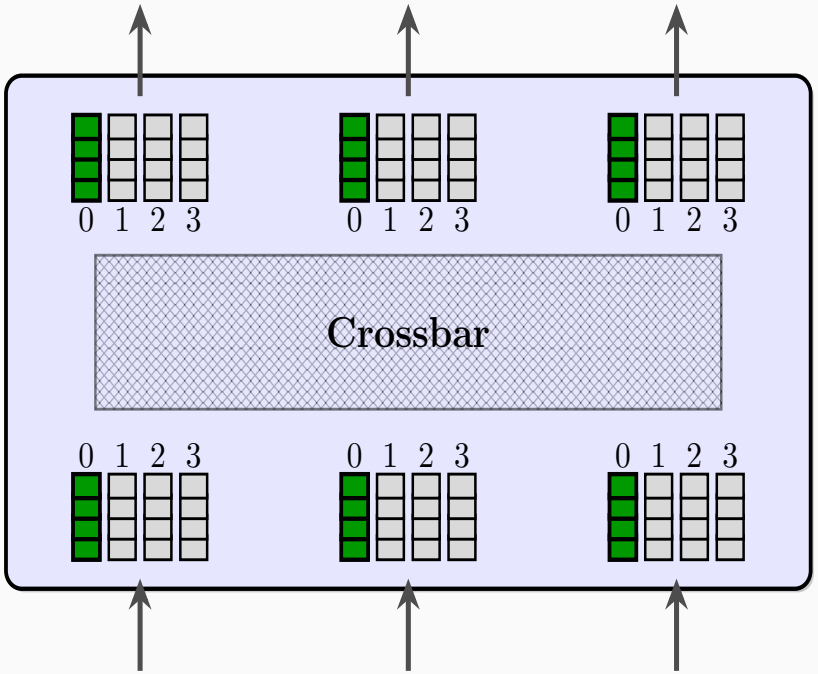


Ordering virtual channels

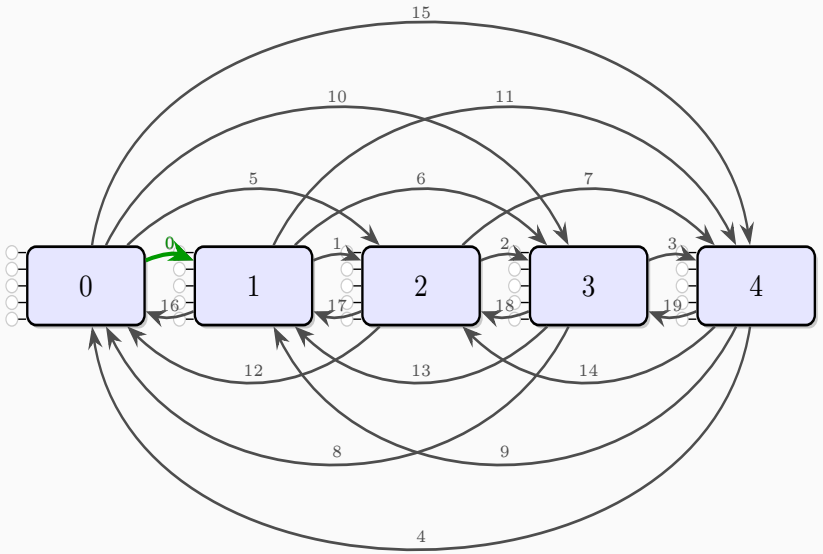


Ordering links

Deadlock Avoidance Methods

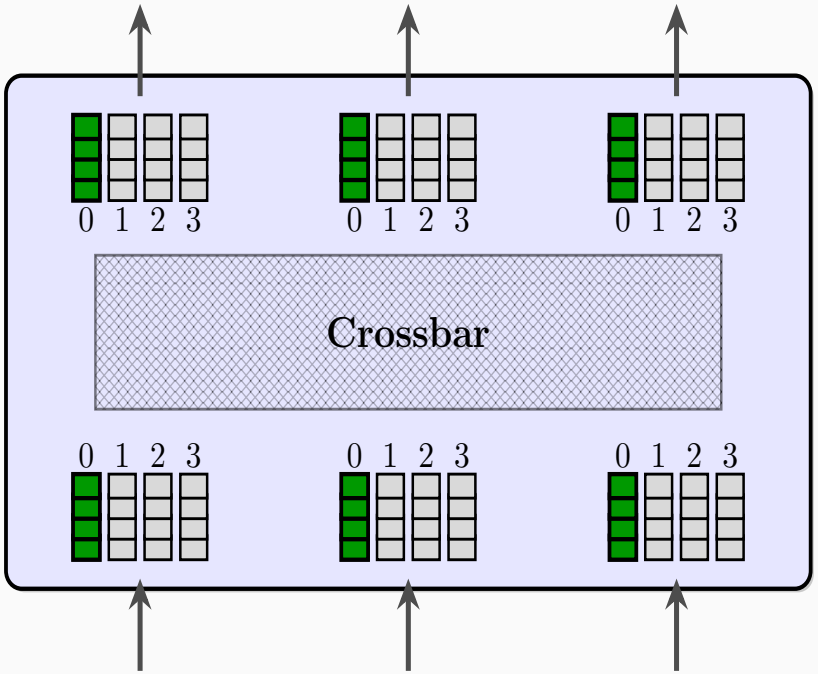


Ordering virtual channels

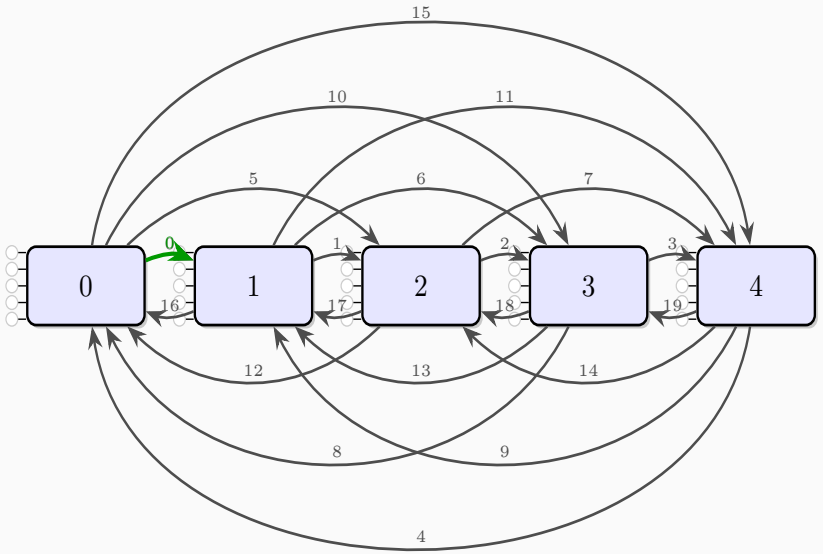


Ordering links

Deadlock Avoidance Methods

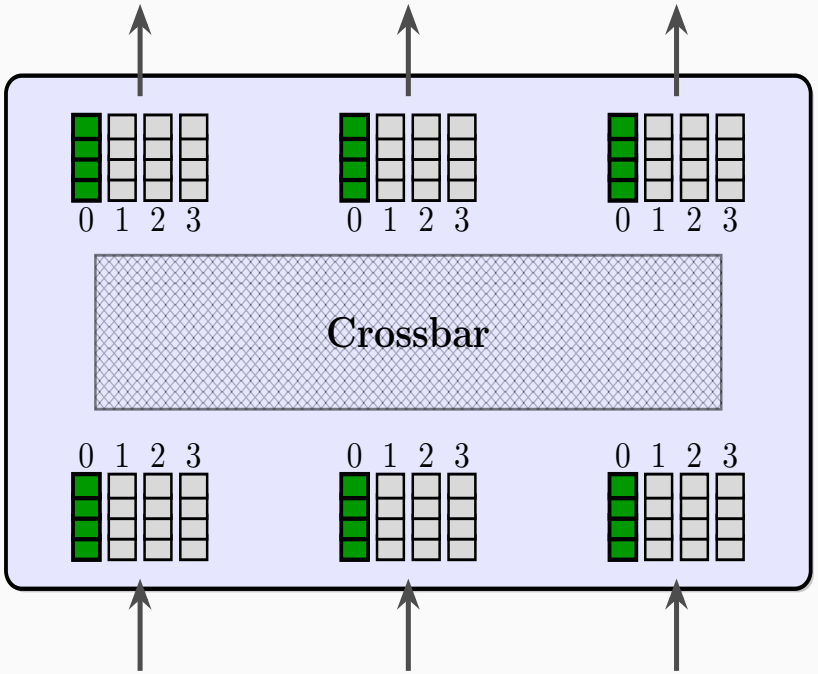


Ordering virtual channels

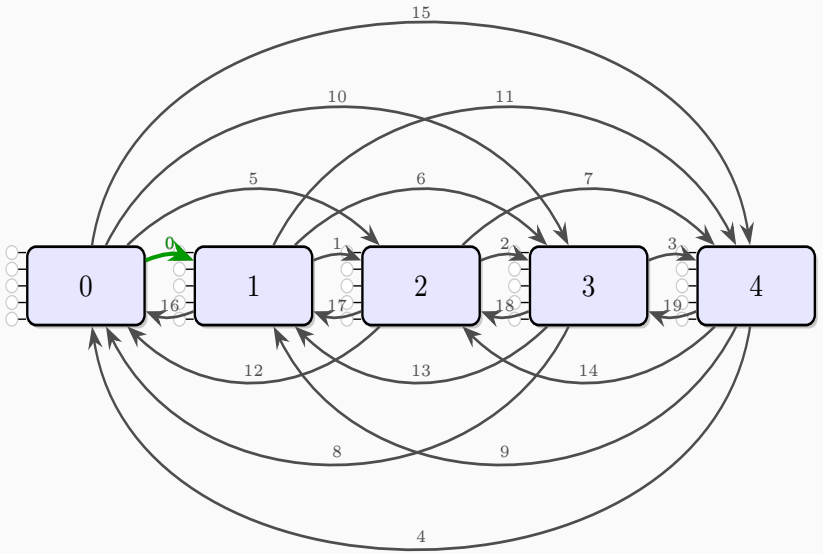


Ordering links

Deadlock Avoidance Methods



Ordering virtual channels



Ordering links

Deadlock Avoidance: Pros and Cons

Ordering Virtual Channels

+ Does not limit path diversity
– Requires extra buffers and arbitration logic
– Higher cost, area, and power consumption

Ordering Links

+ No need for additional virtual channels.
+ Simpler and cheaper to implement.
– Limits path diversity $\Rightarrow$ lower performance

Deadlock Avoidance: Pros and Cons

Ordering Virtual Channels

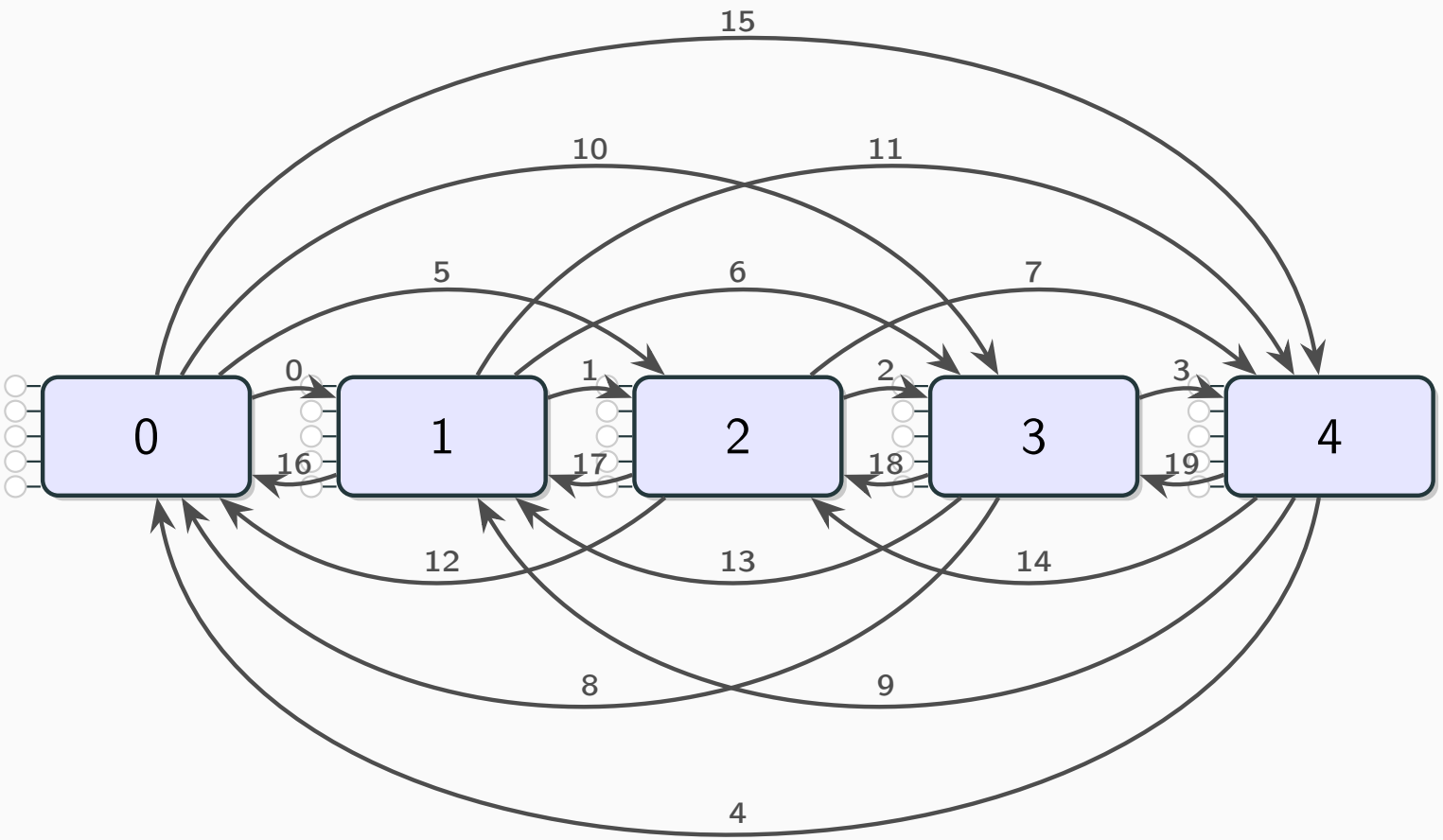
+ Does not limit path diversity
– Requires extra buffers and arbitration logic
– Higher cost, area, and power consumption

Ordering Links

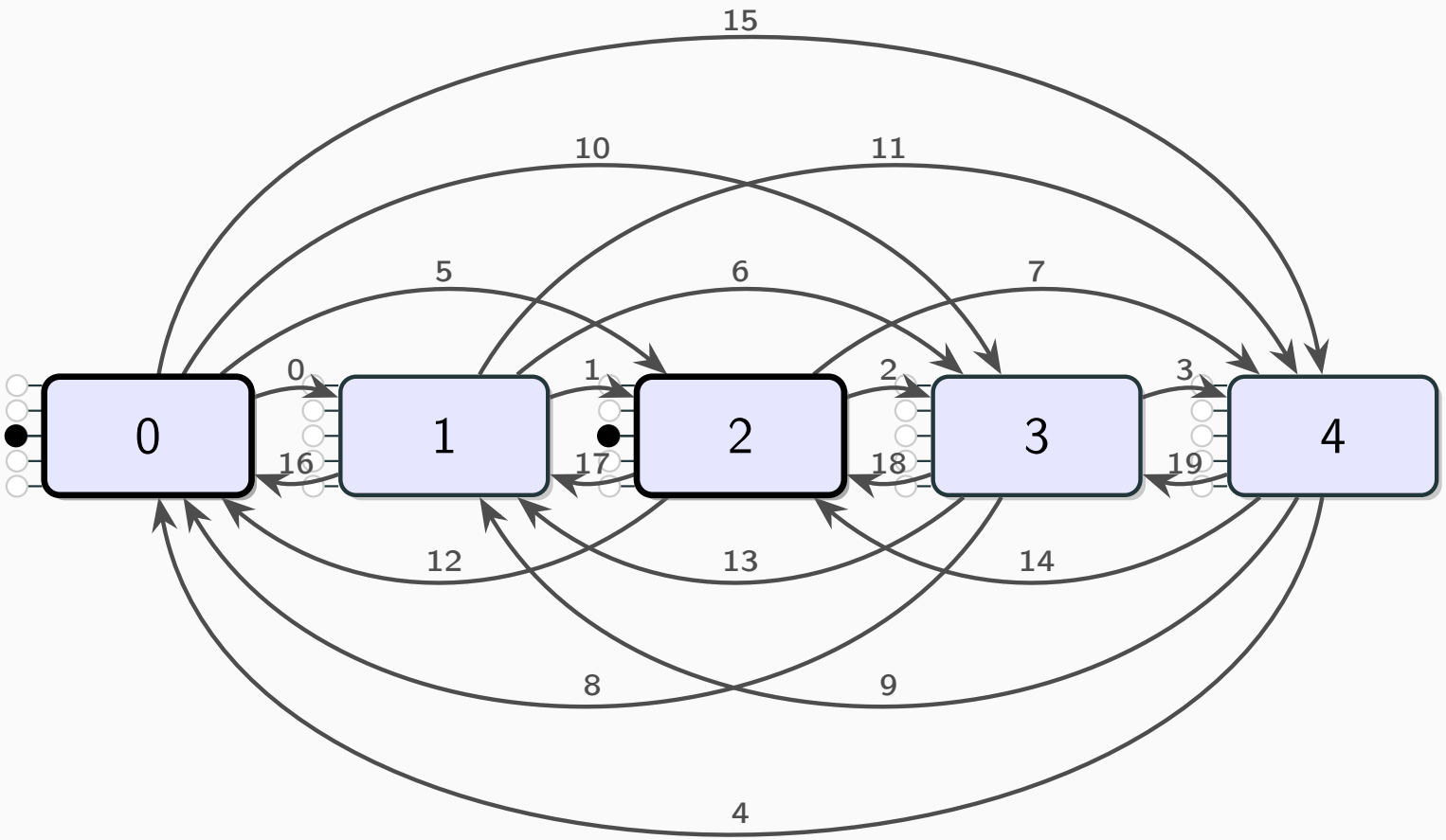
+ No need for additional virtual channels.
+ Simpler and cheaper to implement.
– Limits path diversity $\Rightarrow$ lower performance

Focus on improving path diversity in link ordering.

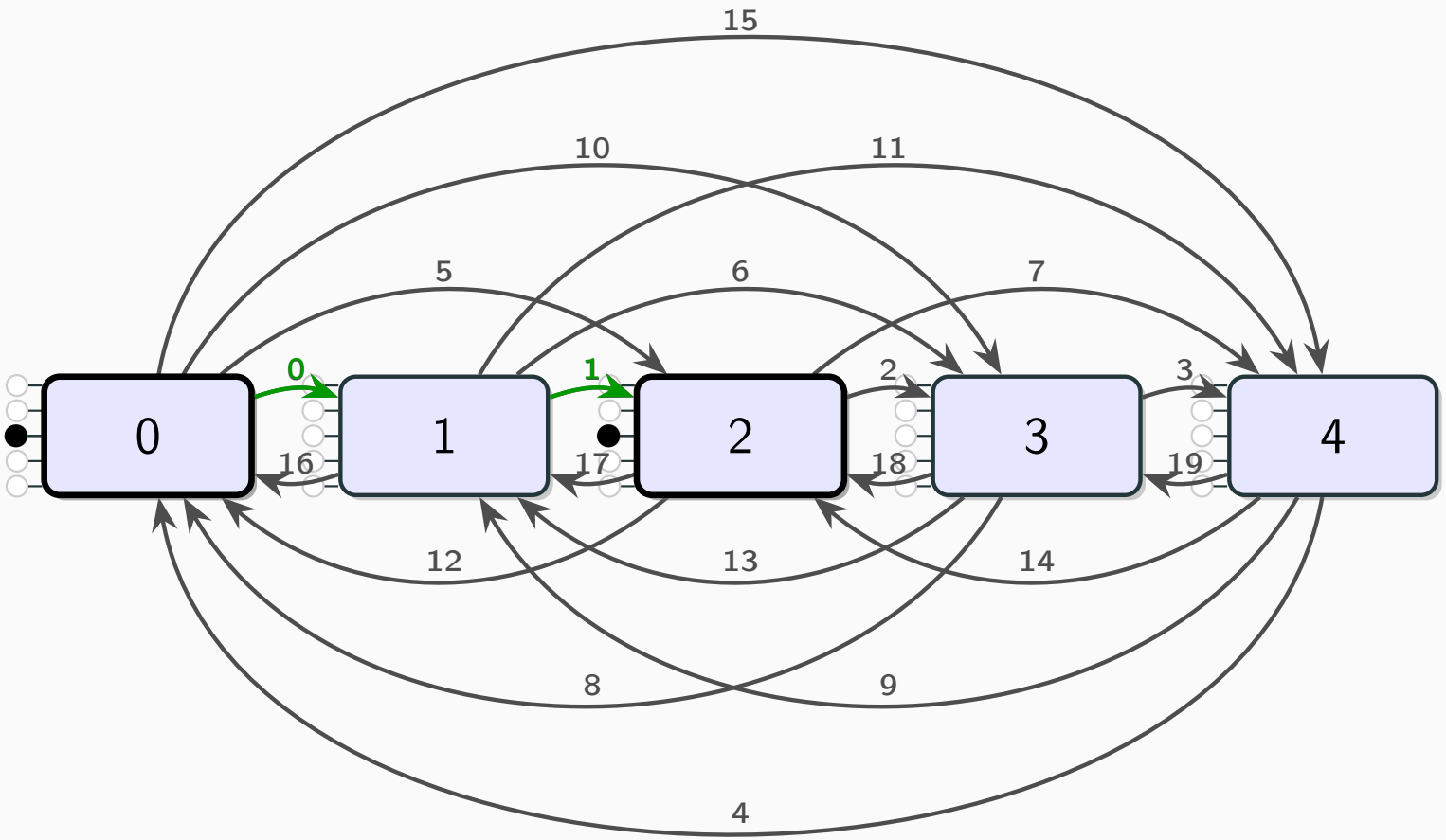
Link ordering example



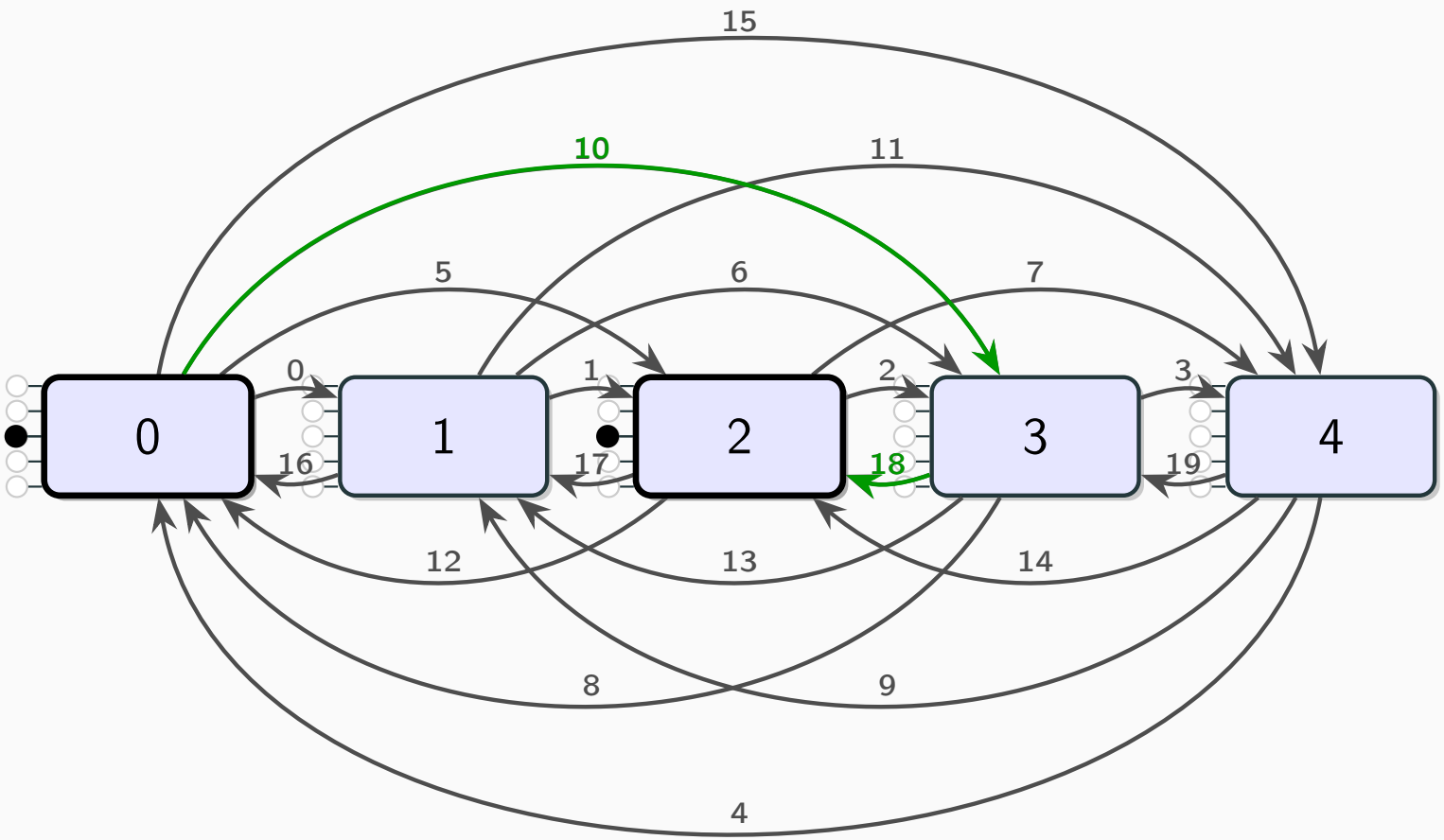
Link ordering example



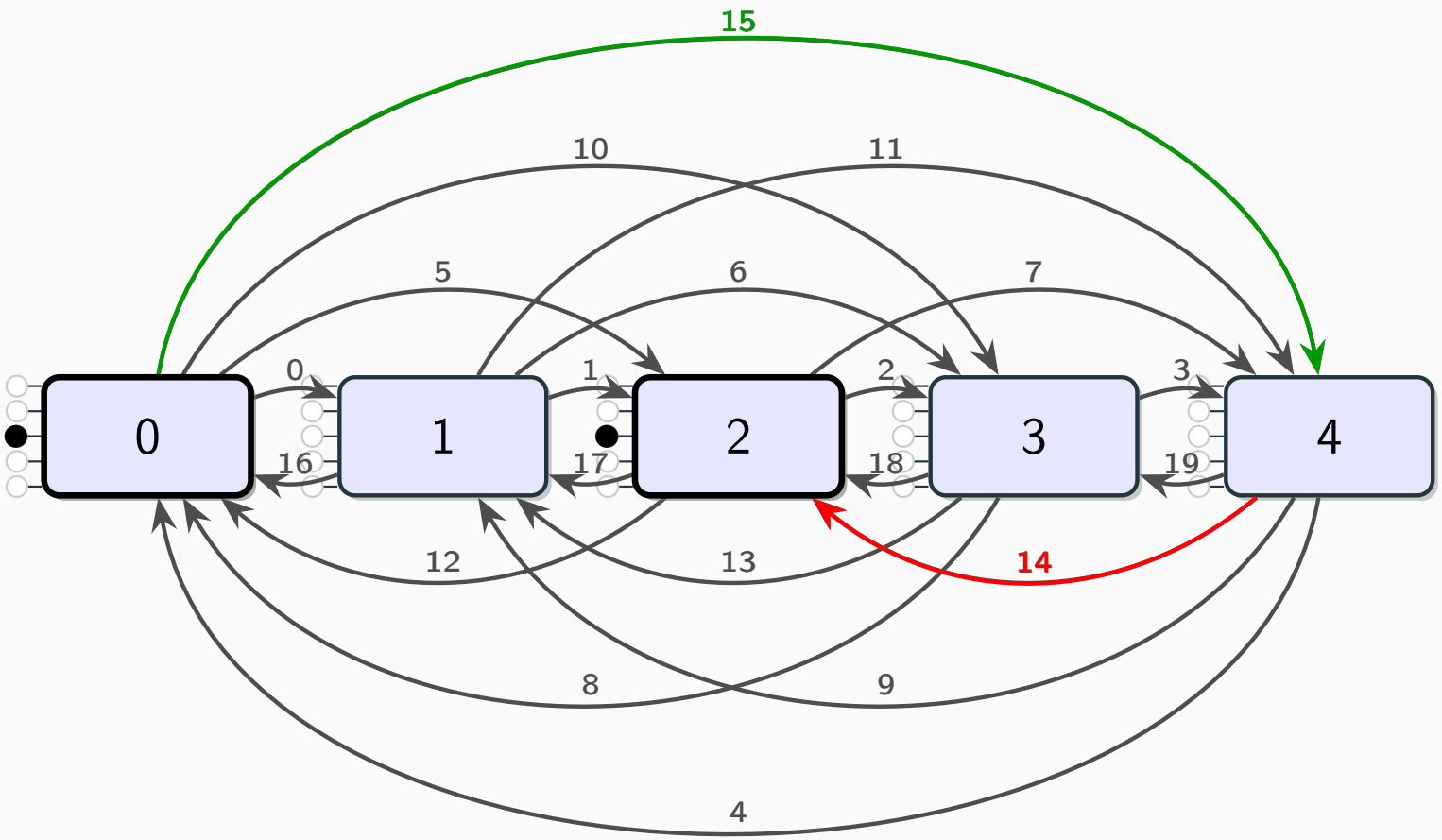
Link ordering example



Link ordering example



Link ordering example



Limitations of Link Ordering

Key Limitations

- Maximum availability: $\frac{2}{3}$ of all 2-hop paths. [1]
- With uniform link utilization only $\frac{1}{2}$ of the total.
- Half the paths ≈ Half the throughput.

Implication

Link ordering has reached its limit. A new approach is needed.

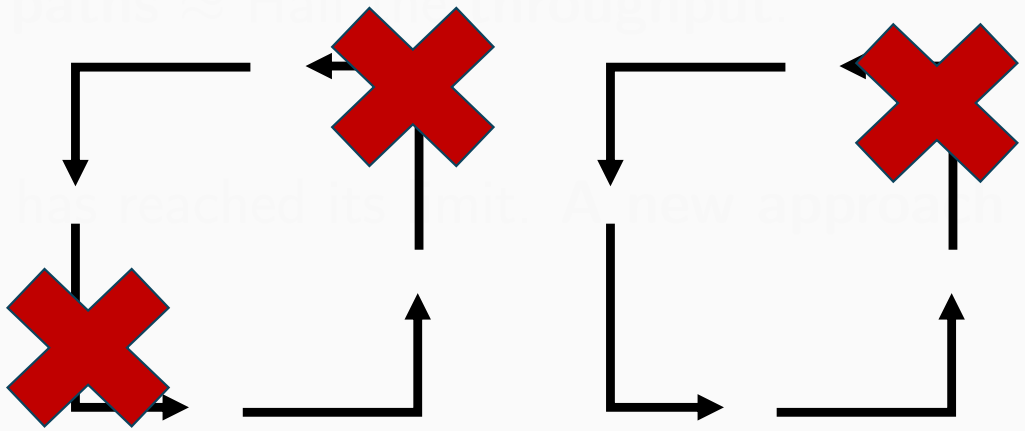
[1] Kwauk et al., "BoomGate: Deadlock Avoidance in Non-Minimal Routing for High Radix Networks". HPCA'21. doi:10.1109/HPCA51647.2021.00064.

Limitations of Link Ordering

Key Limitations

- Maximum availability: $\frac{2}{3}$ of all 2-hop paths. [1]
- With **uniform link utilization** only $\frac{1}{2}$ of the total.
- Half the paths ⇒ Half the throughput.

Implication:
Link ordering has reached its limit. A new approach is needed.



DOR (X then Y) Turn model

Limitations of Link Ordering

Key Limitations

- Maximum availability: $\frac{2}{3}$ of all 2-hop paths. [1]
- With **uniform link utilization** only $\frac{1}{2}$ of the total.
- Half the **paths** $\approx$ Half the **throughput**.

Implication

Link ordering has reached its limit. **A new approach is needed.**

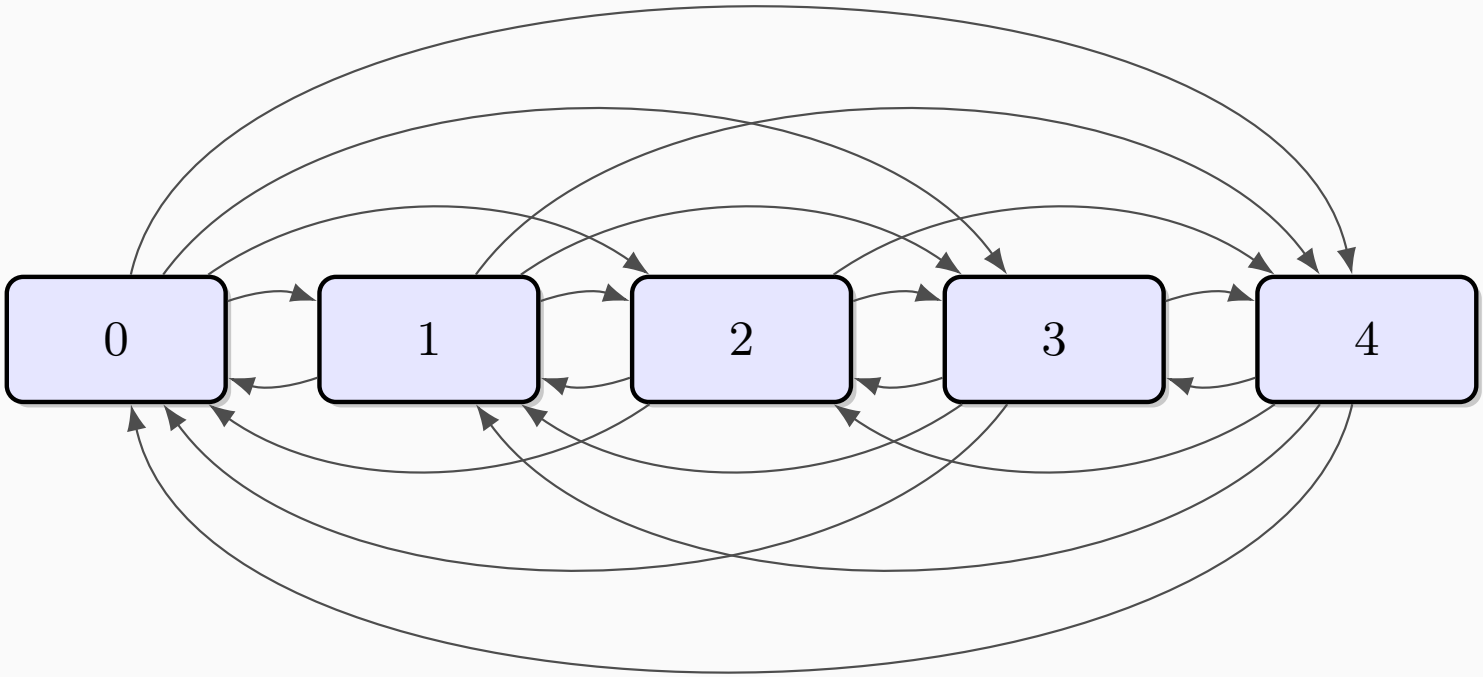
[1] Kwauk et al., "BoomGate: Deadlock Avoidance in Non-Minimal Routing for High Radix Networks". HPCA'21. doi:10.1109/HPCA51647.2021.00064.

TERA: Topology Embedded Routing Algorithm

Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

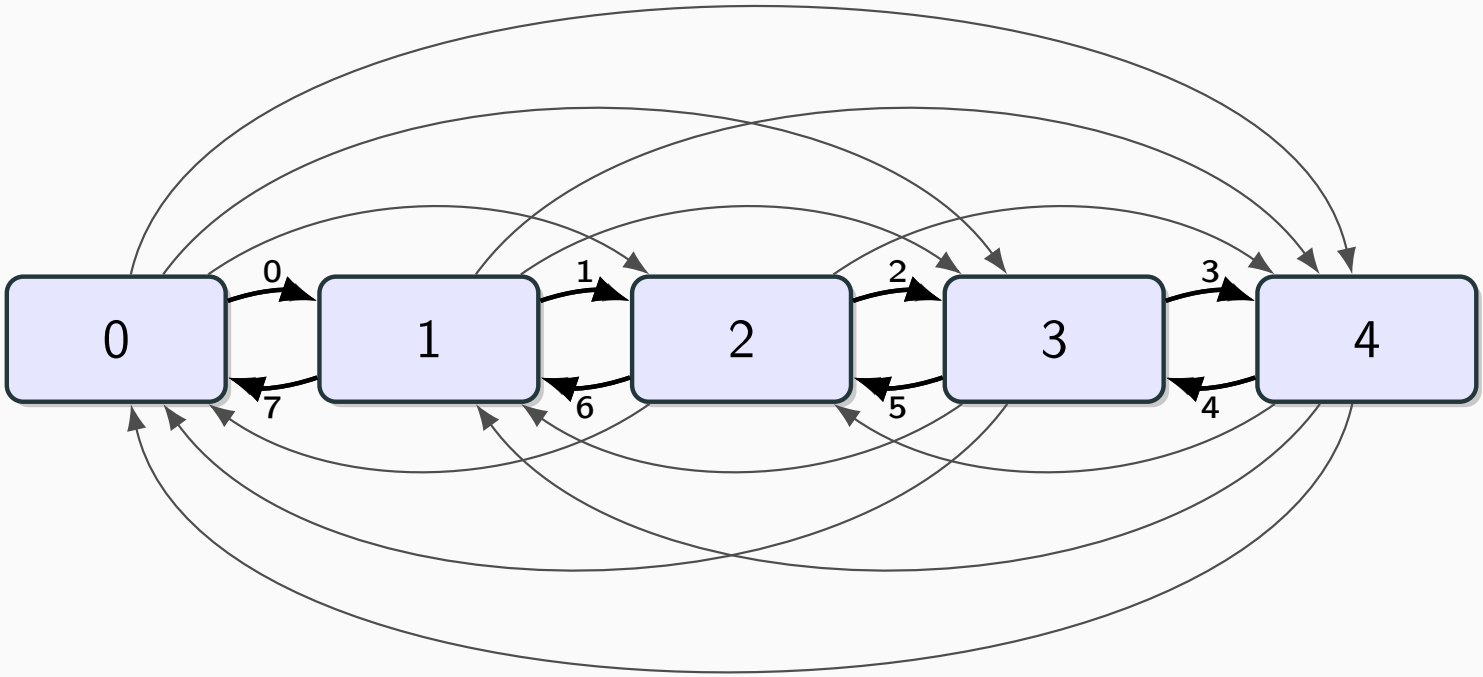
- **Service** network: an embedded network.
- **Main** network: the complement of the service network.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

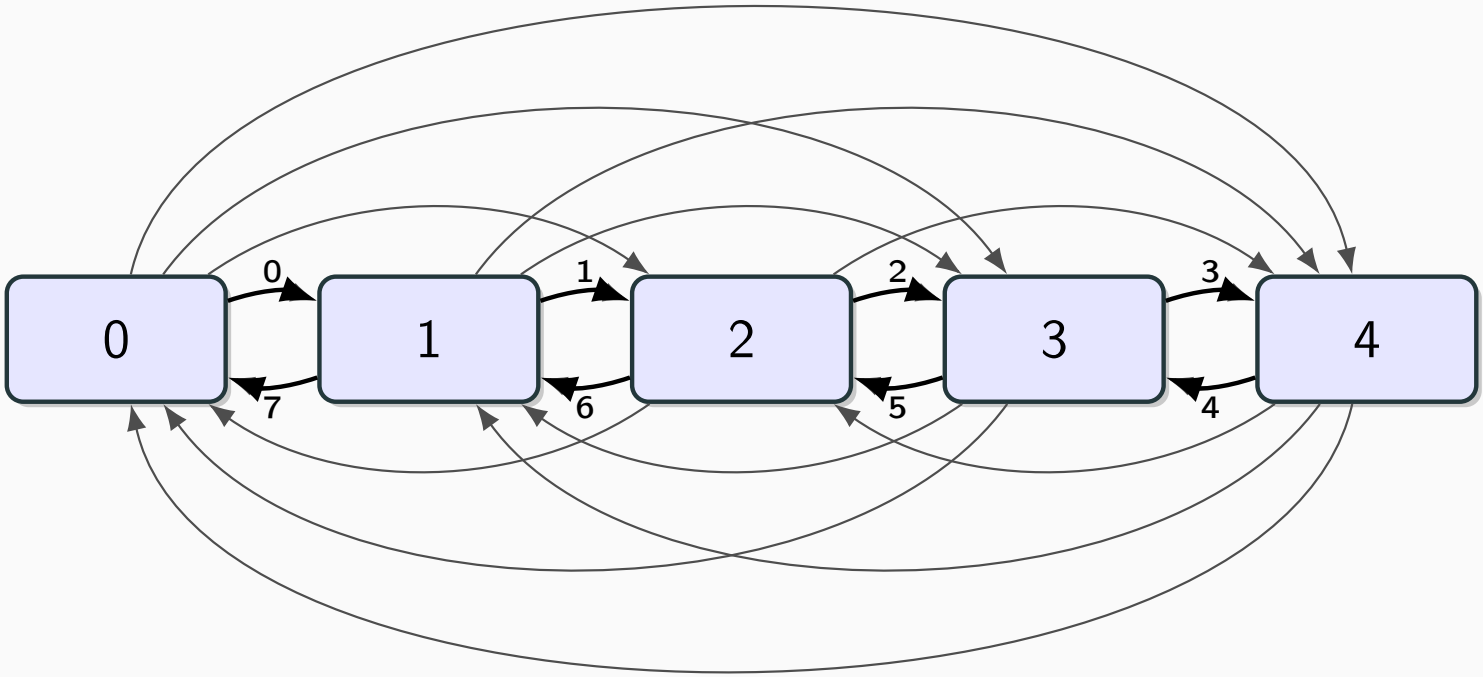
- **Service** network: an embedded network.
- **Main** network: the complement of the service network.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

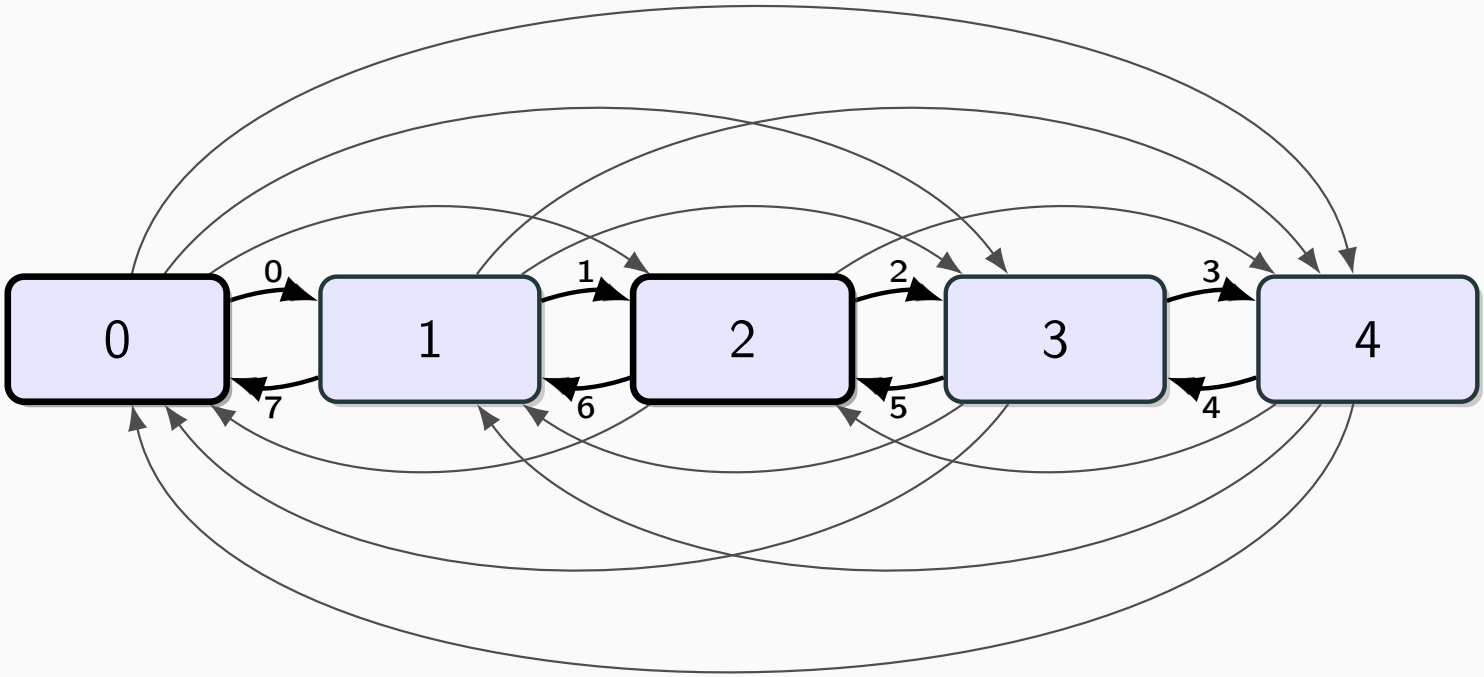
- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

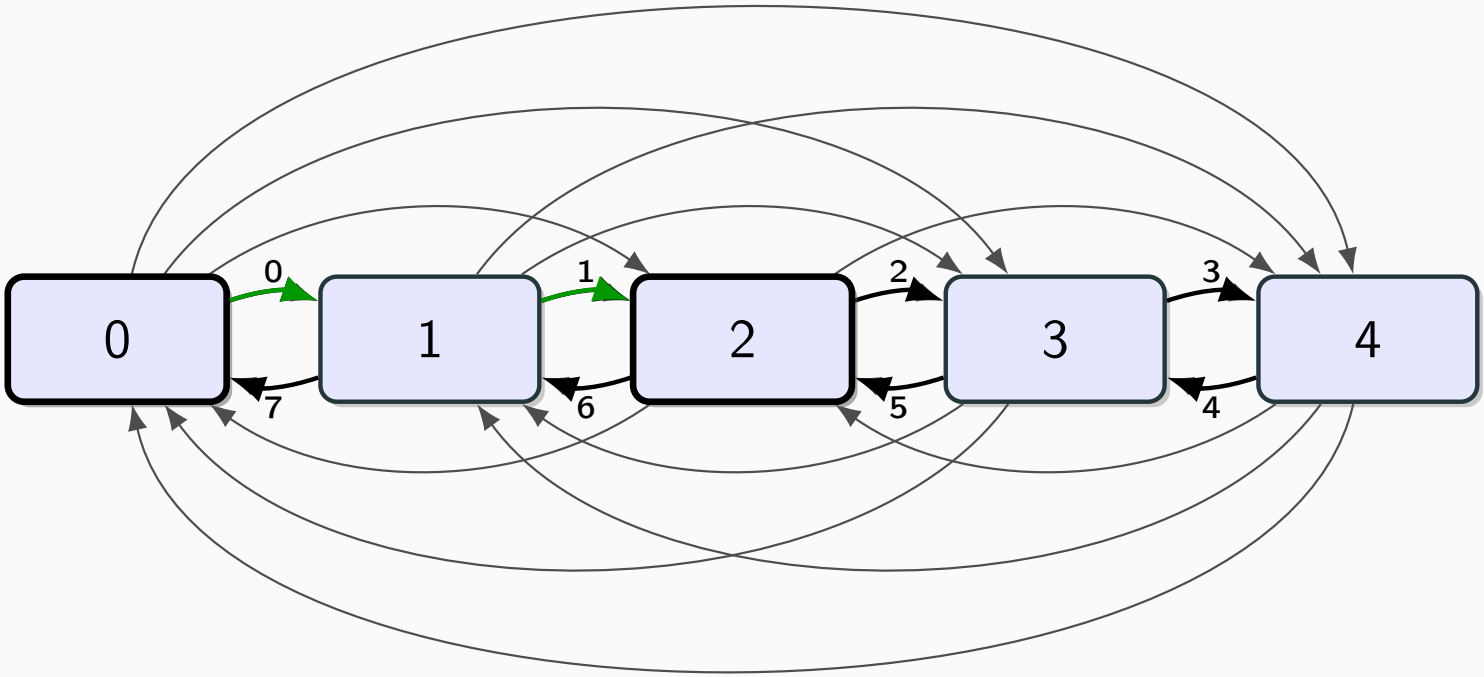
- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

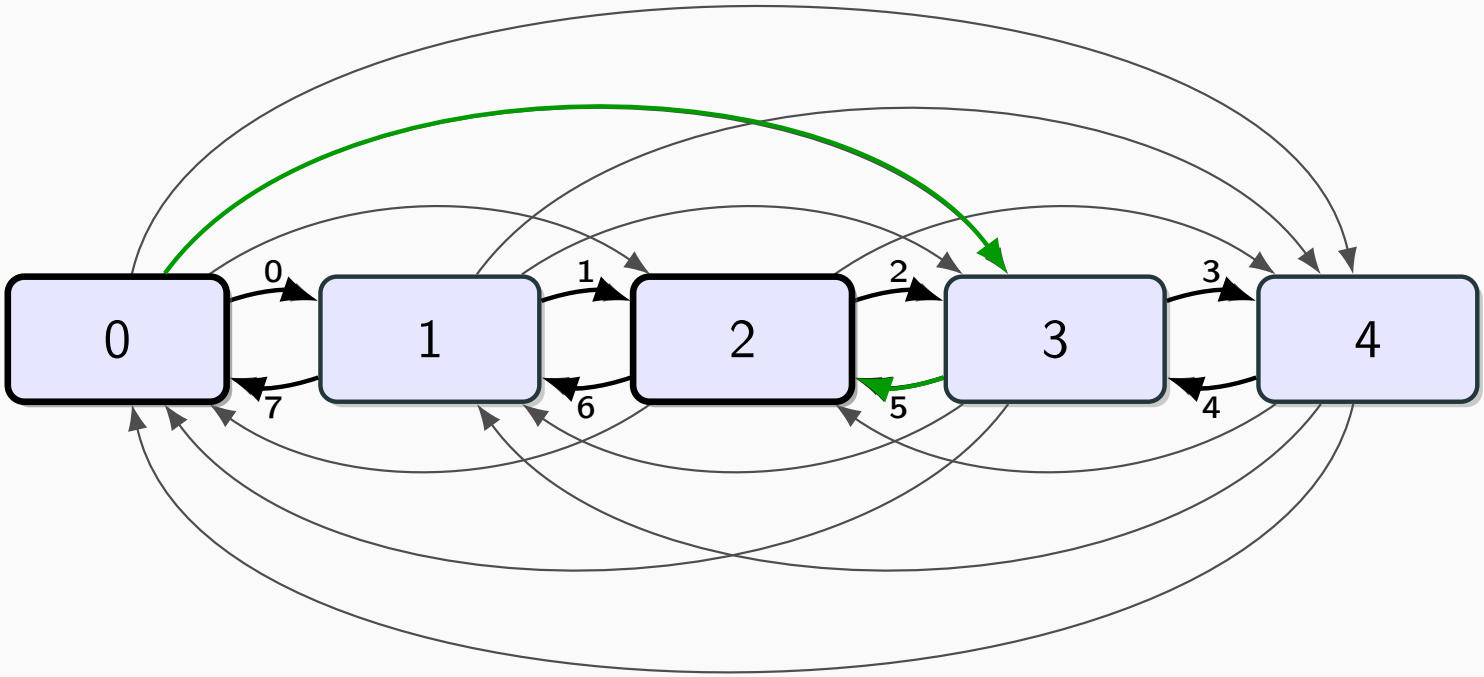
- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

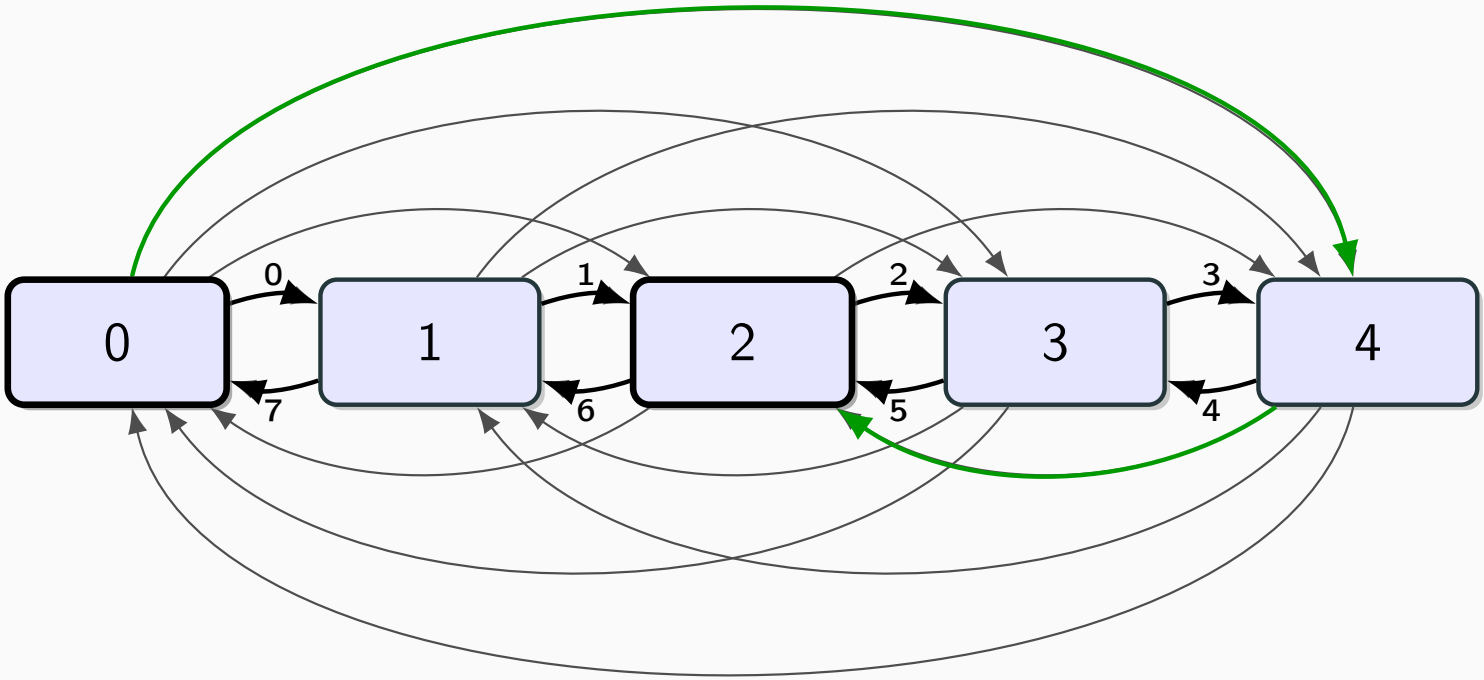
- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

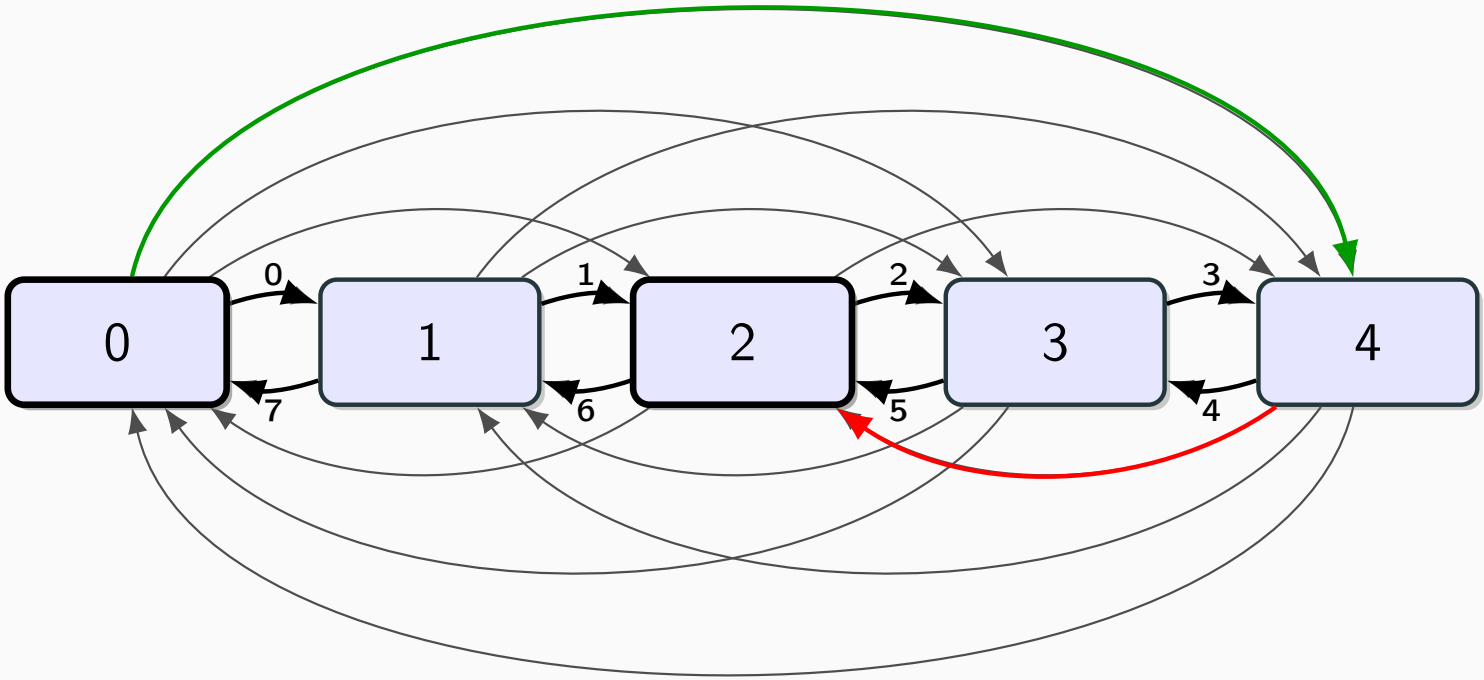
- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

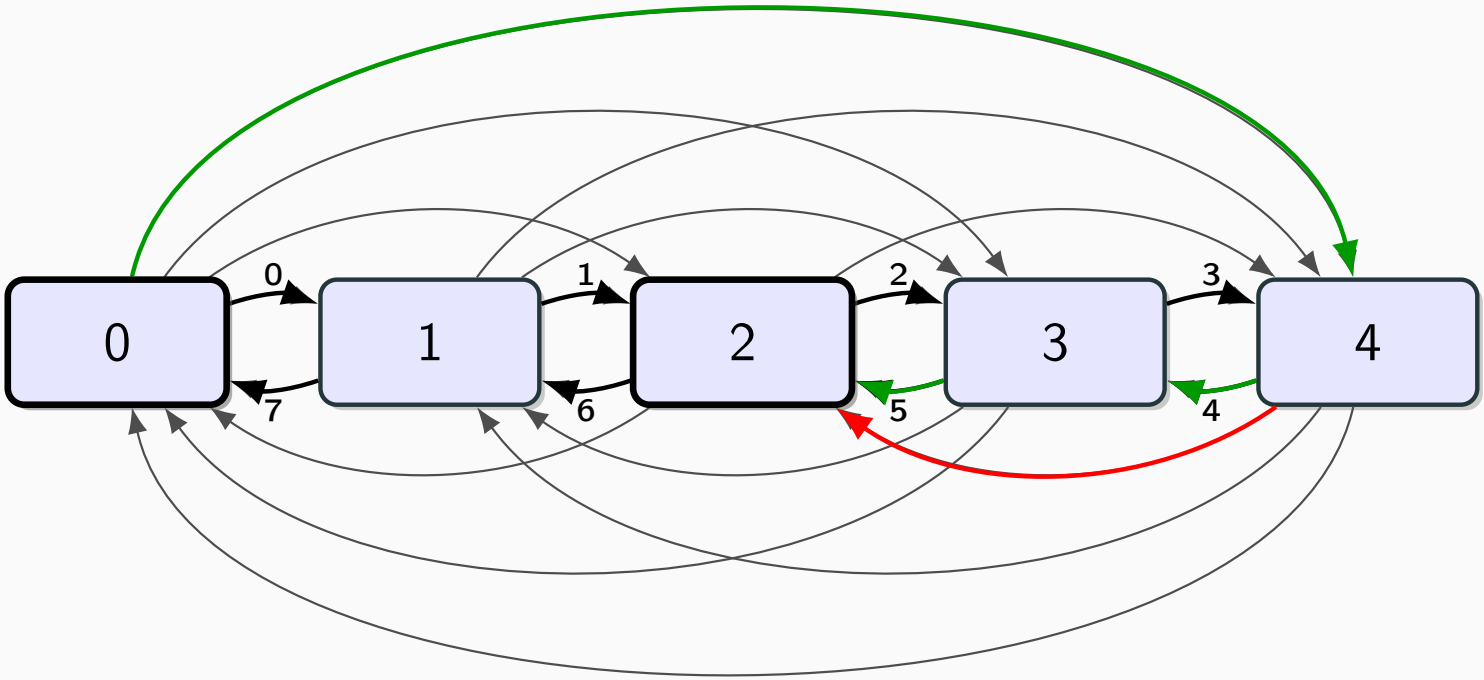
- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

Physical partitioning of the Full-mesh:

- **Service** network: a deadlock-free set of paths.
- **Main** network: no restriction in the use.



Topology Embedded Routing Algorithm (TERA)

At Injection Port

- MIN hop
- Main hop
- Service hop

At In-transit Port

- MIN hop
- Service hop

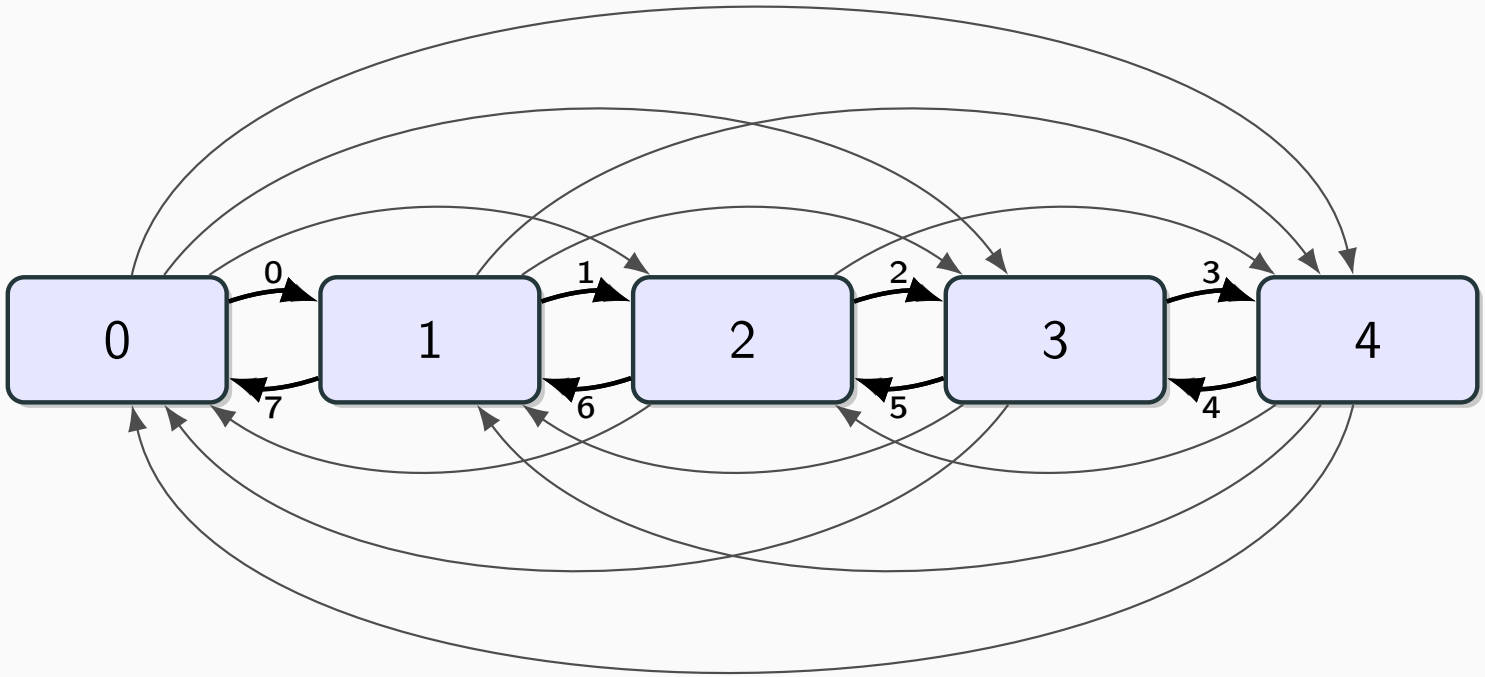
Choose port with $\min\ w(p)$

$$w(p) = \text{occupancy}[p] + \begin{cases} 0, & \text{if minimal} \\ C, & \text{if non-minimal} \end{cases}$$

Topology Embedded Routing Algorithm (TERA)

The choice of the **service network** directly controls two critical network properties:

- The total number of available non-minimal paths.
- The maximum number of hops a packet can take.



TERA service network

Topology	Diameter	#Links
Full-Mesh	1	$O(n^2)$
Mesh	$O(n)$	$O(n)$
Tree	$O(\log n)$	$O(n)$
Hypercube	$O(\log n)$	$O(n \log n)$
3D-HyperX	3	$O(n^{1.33})$
2D-HyperX	2	$O(n^{1.50})$

Properties in terms of the number of switches n.

TERA service network

Topology	Diameter	#Links
Full-Mesh	1	$O(n^2)$
Mesh	$O(n)$	$O(n)$
Tree	$O(\log n)$	$O(n)$
Hypercube	$O(\log n)$	$O(n \log n)$
3D-HyperX	3	$O(n^{1.33})$
2D-HyperX	2	$O(n^{1.50})$

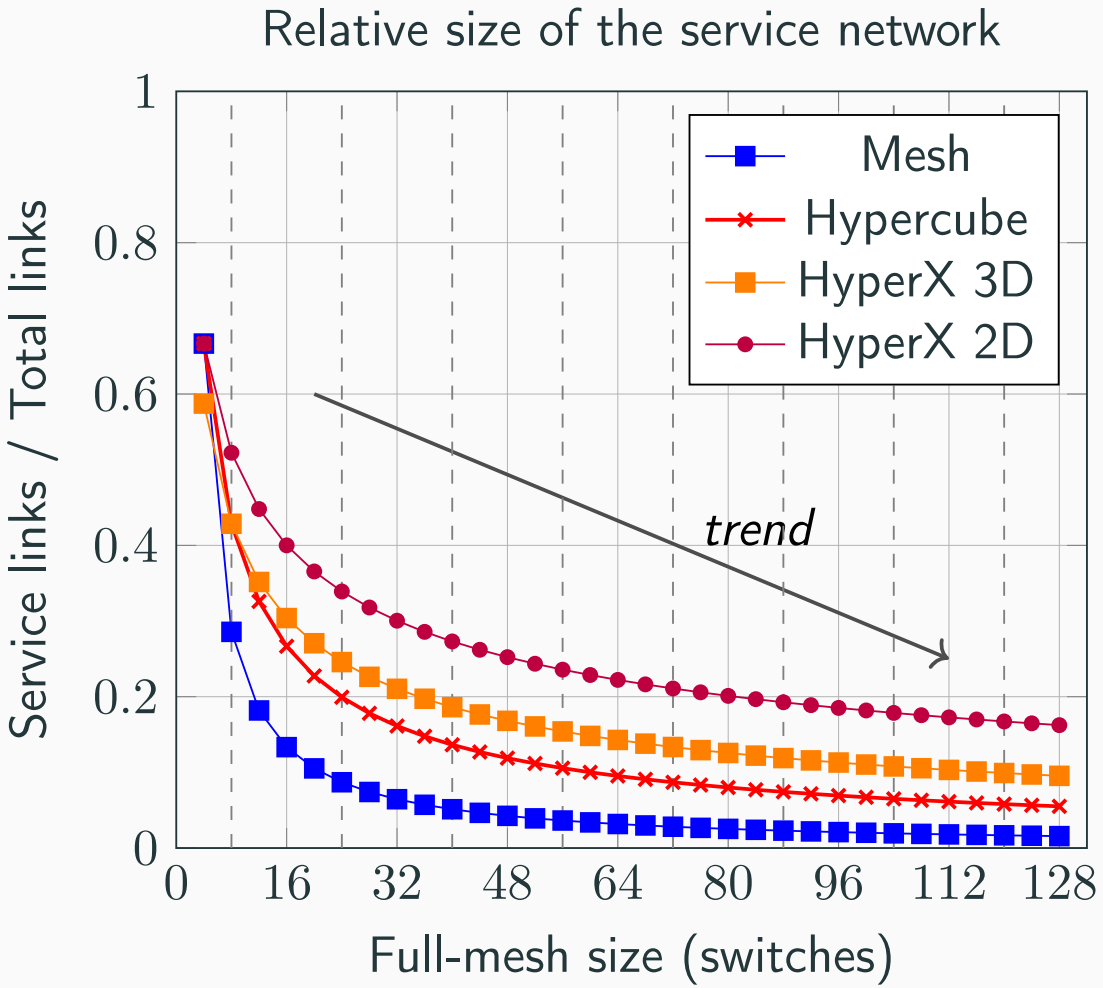
Properties in terms of the number of switches n.

TERA service network

Topology	Diameter	#Links
Full-Mesh	1	$O(n^2)$
Mesh	$O(n)$	$O(n)$
Tree	$O(\log n)$	$O(n)$
Hypercube	$O(\log n)$	$O(n \log n)$
3D-HyperX	3	$O(n^{1.33})$
2D-HyperX	2	$O(n^{1.50})$

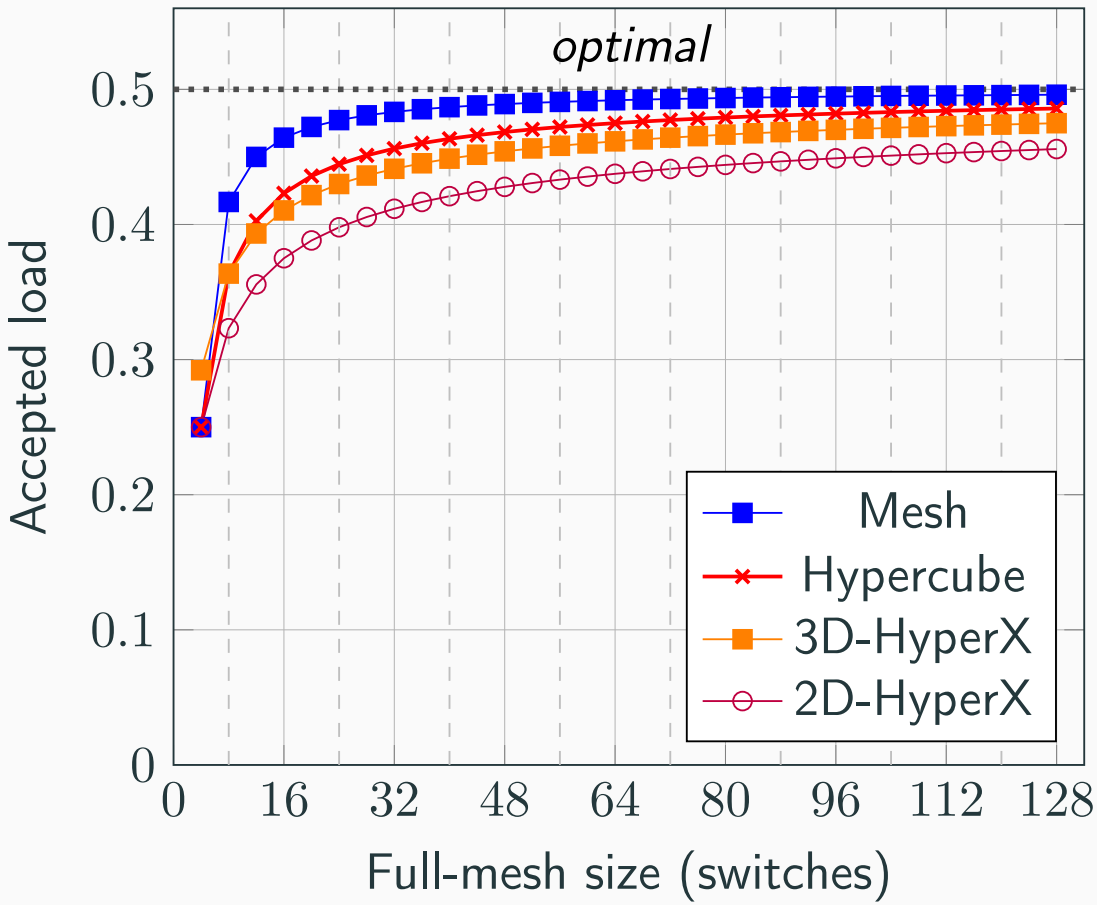
Properties in terms of the number of switches n.

TERA service network



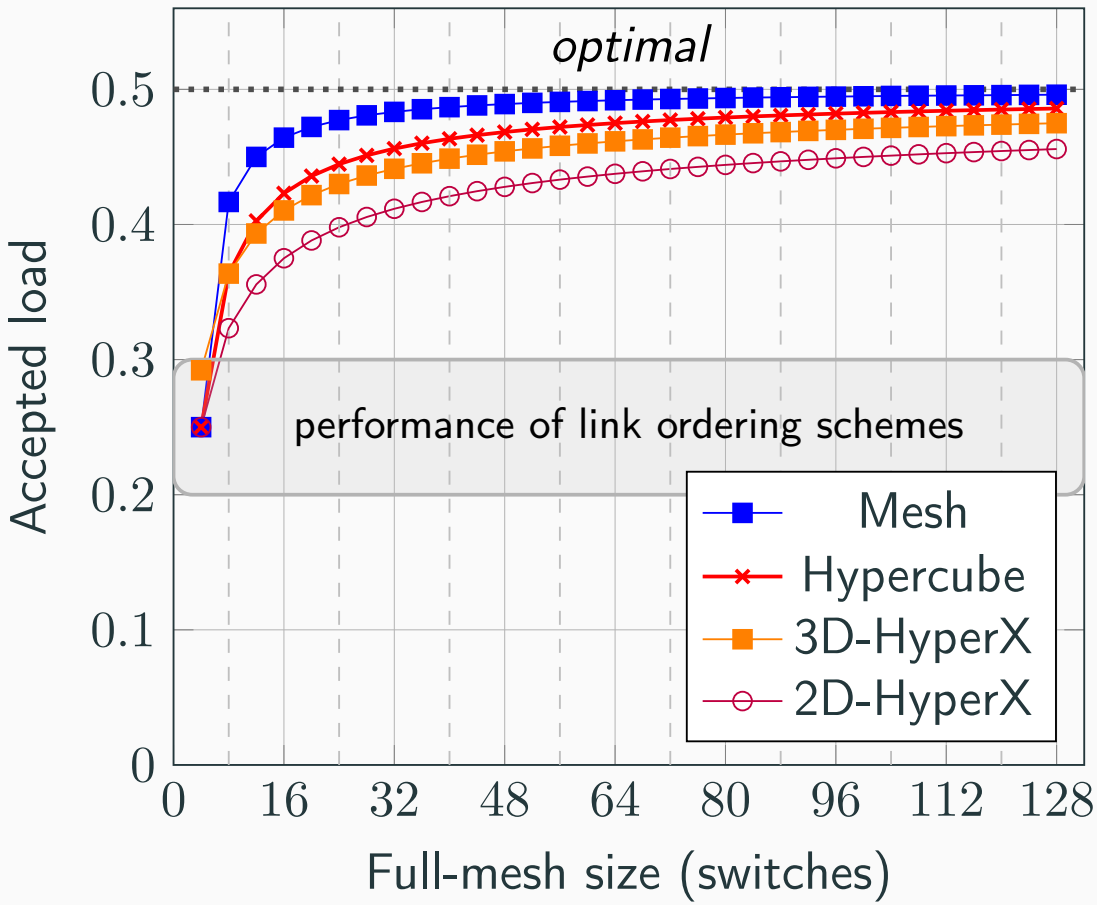
TERA service network

Estimated performance of TERA under adverse traffic



TERA service network

Estimated performance of TERA under adverse traffic



Methodology and Results

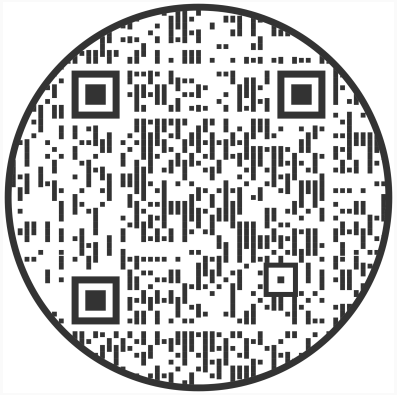
Methodology

Evaluated Routing Schemes in CAMINOS simulator

- Omni-WAR: Baseline for Full-Mesh (2 VCs). [2]
- sRINR: Link ordering SOTA (1VC).
- T-Hx2D: TERA with a service HyperX 2D (1 VC).
- T-Hx3D: TERA with a service HyperX 3D (1 VC).

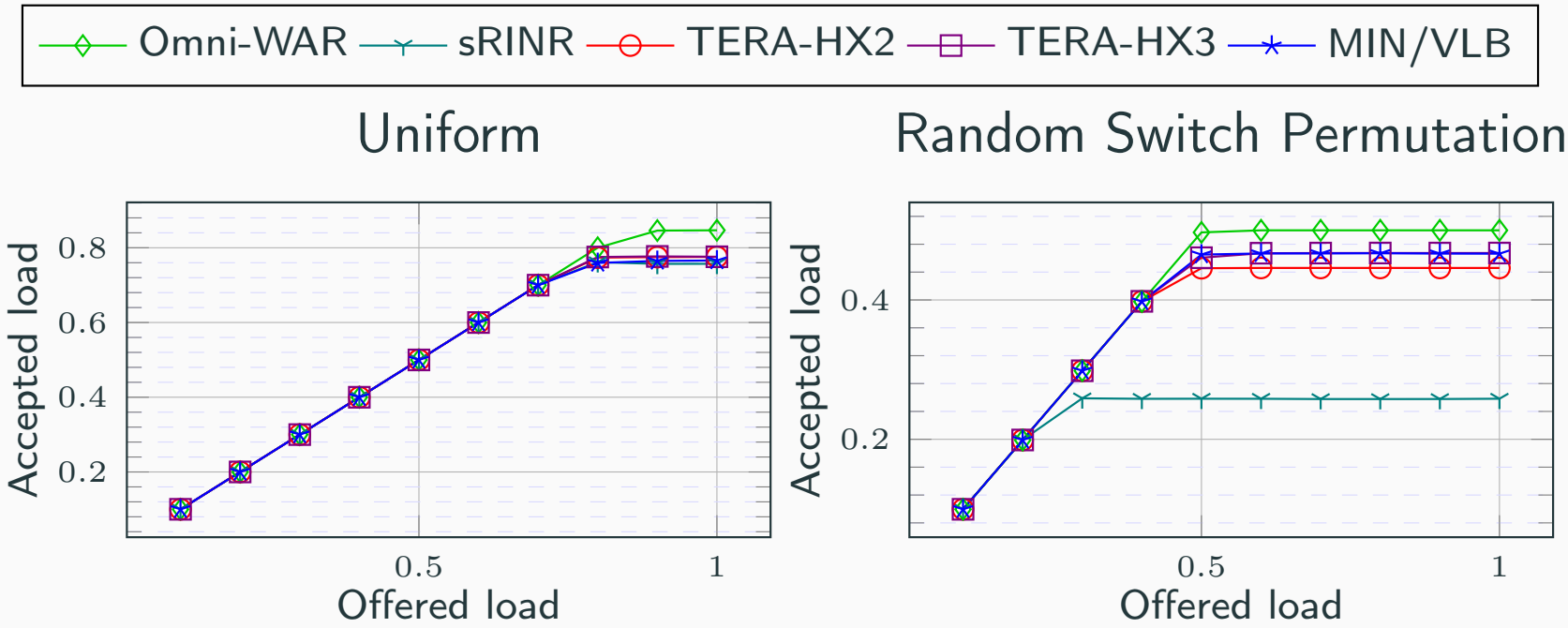
Code for reproducibility:

`https://github.com/alexcano98/TERA-routing-HOTI-2025-reproducibility`



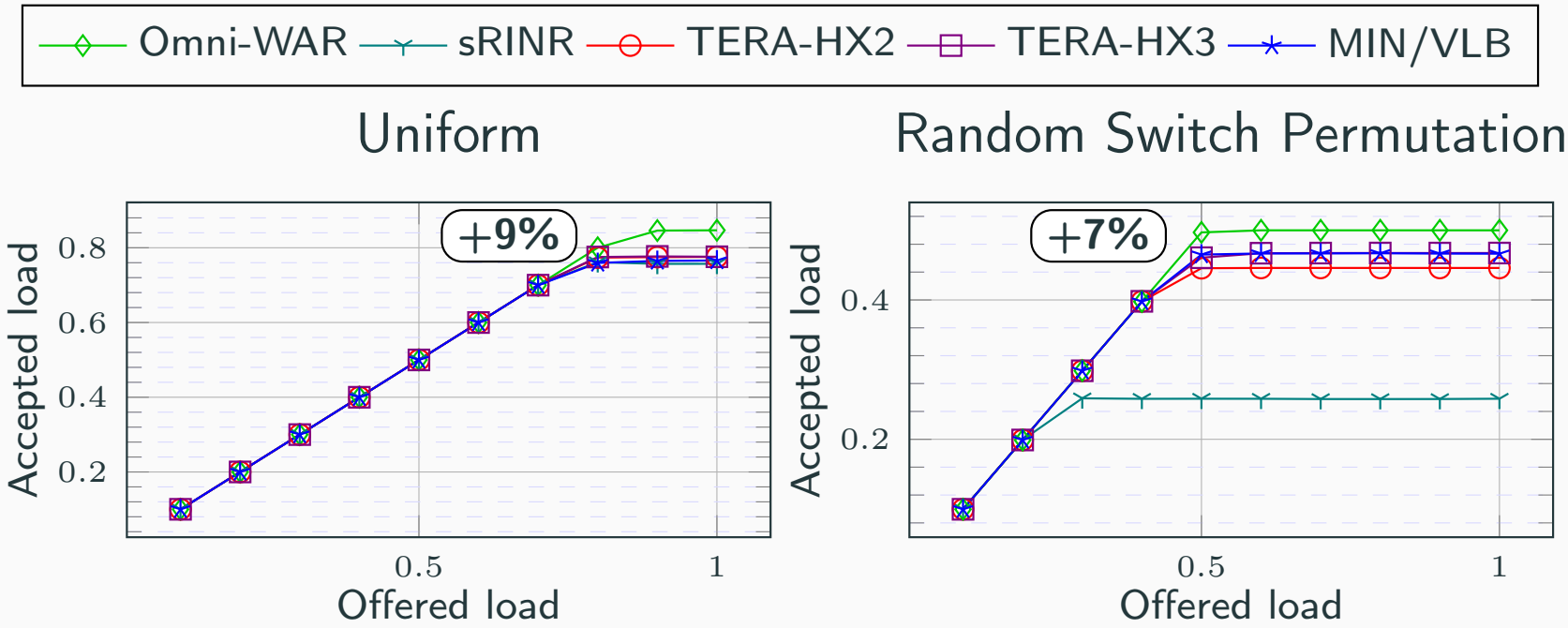
[2] McDonald et al., "Practical and efficient incremental adaptive routing for HyperX networks," in Proc. SC '19. doi:10.1145/3295500.3356151.

Evaluation in Full-mesh



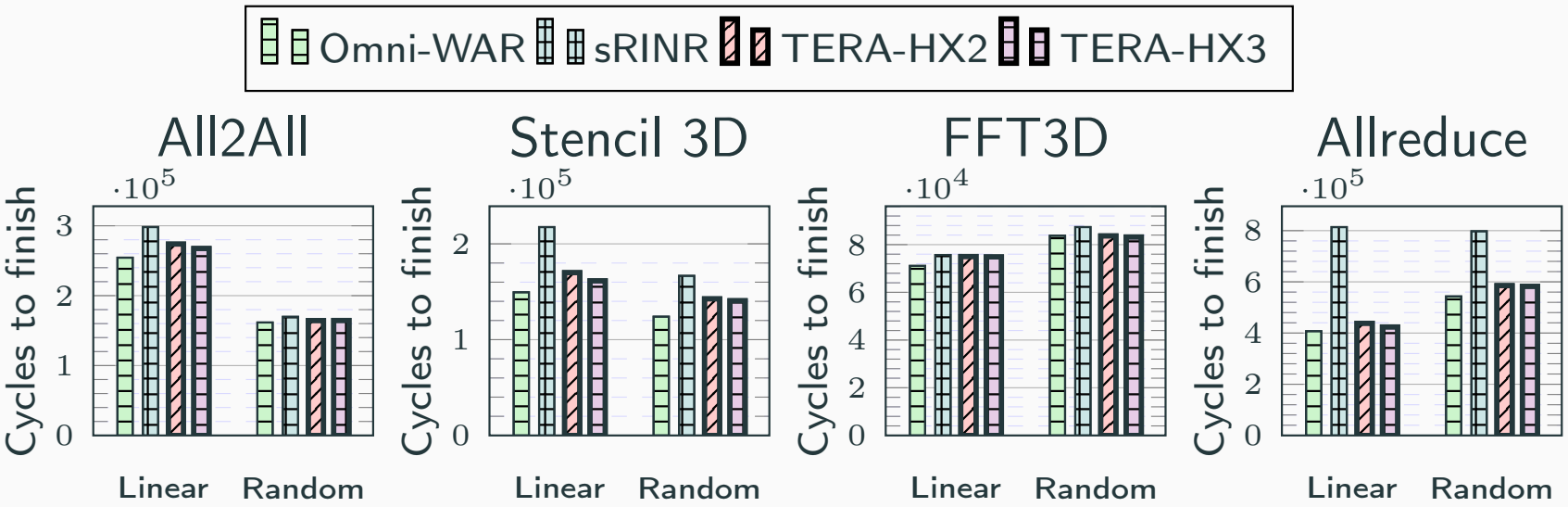
Traffic with Bernoulli generation at a different offered load in a Full-Mesh of 64 switches with 4096 total servers.

Evaluation in Full-mesh



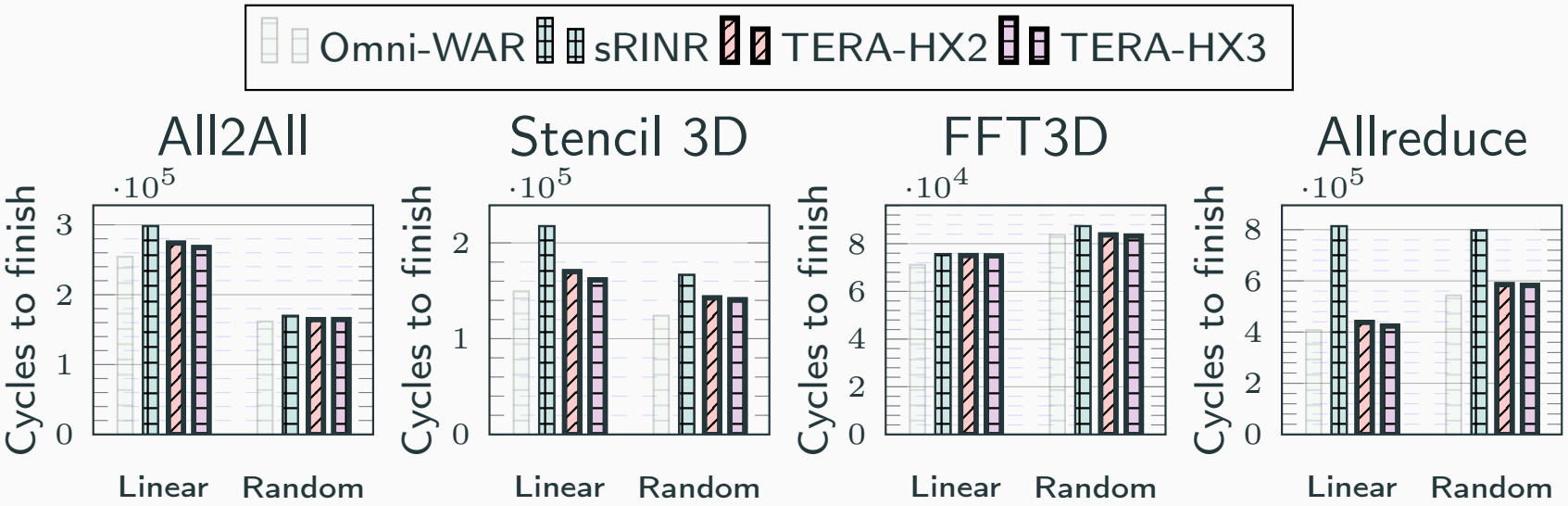
Traffic with Bernoulli generation at a different offered load in a Full-Mesh of 64 switches with 4096 total servers. TERA runs without VCs!

Evaluation in Full-mesh



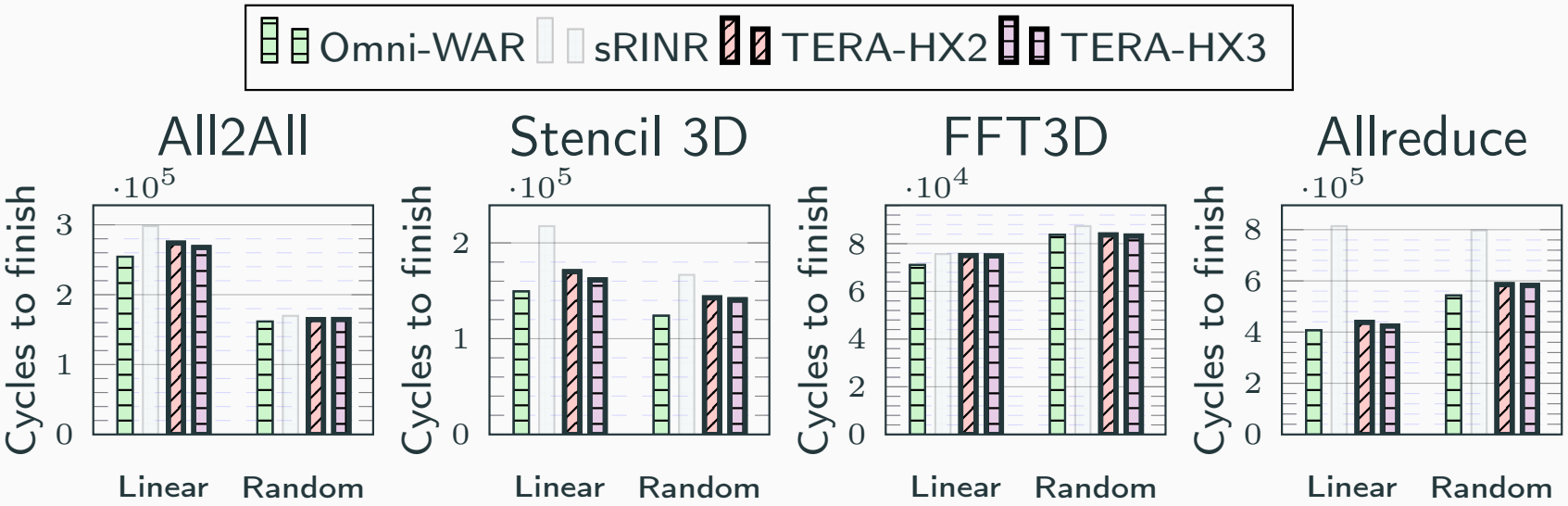
Time to consume a communication kernel in a Full-Mesh of 64 switches with 4096 total servers.

Evaluation in Full-mesh



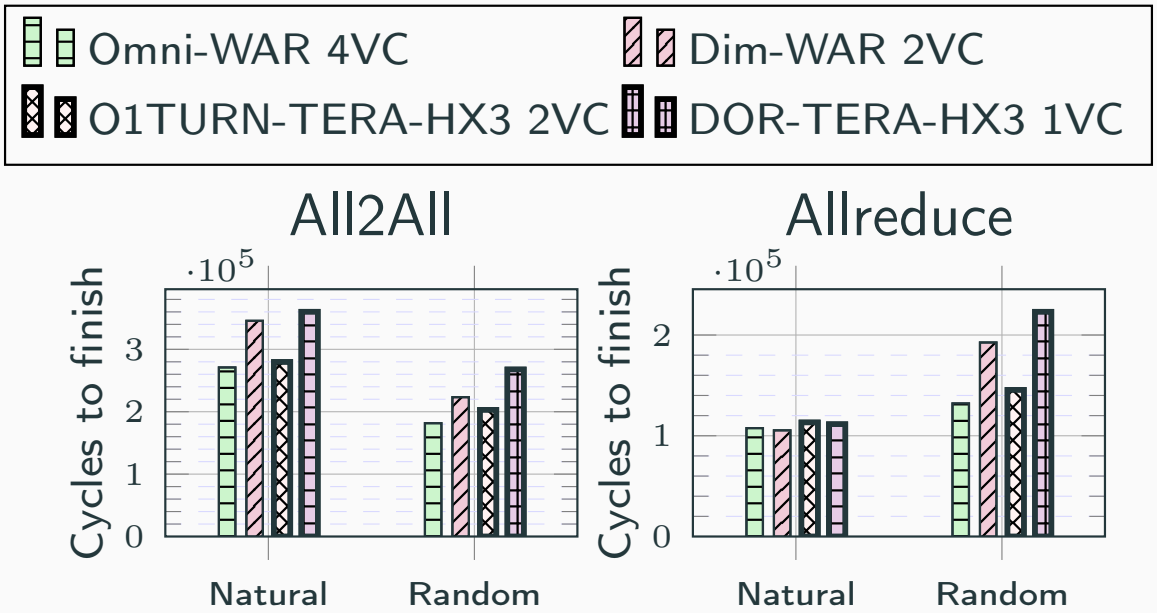
Time to consume a communication kernel in a Full-Mesh of 64 switches with 4096 total servers. TERA is 24% faster than sRINR on average!

Evaluation in Full-mesh



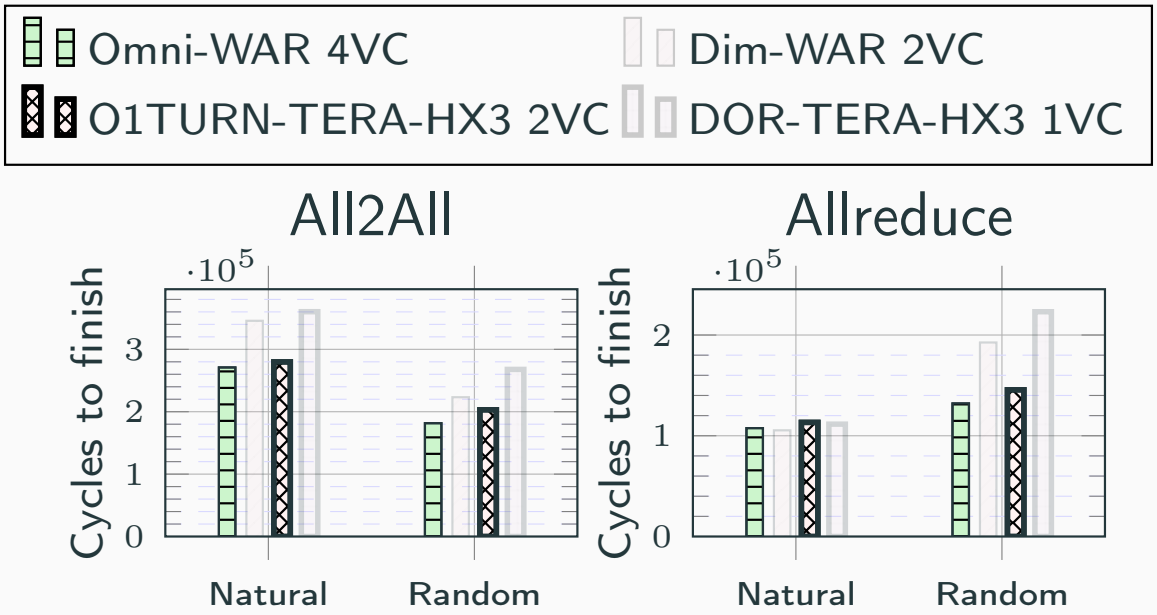
Time to consume a communication kernel in a Full-Mesh of 64 switches with 4096 total servers. TERA is 5.1% slower than Omni-WAR on average!

Evaluation in a HyperX 2D



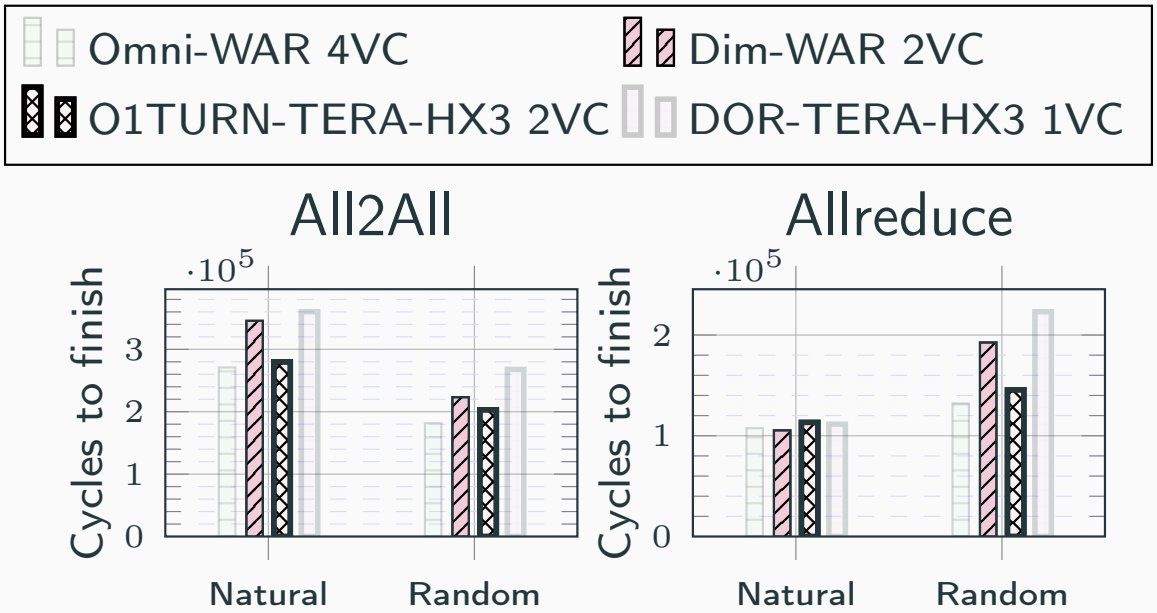
Time to consume an All2All and Allreduce kernel in an 8x8 HyperX network of 2 dimensions

Evaluation in a HyperX 2D



Time to consume an All2All and Allreduce kernel in an 8x8 HyperX network of 2 dimensions

Evaluation in a HyperX 2D



Time to consume an All2All and Allreduce kernel in an 8x8 HyperX network of 2 dimensions

Conclusions

Conclusions

No VC

Deadlock-free routing in Full-Mesh networks **without Virtual Channels**.

100%

Throughput gain over the previous state-of-the-art routing algorithm.

✓

Successfully adapted to a 2D HyperX network.

UC

Thank you! Questions?

Alejandro Cano

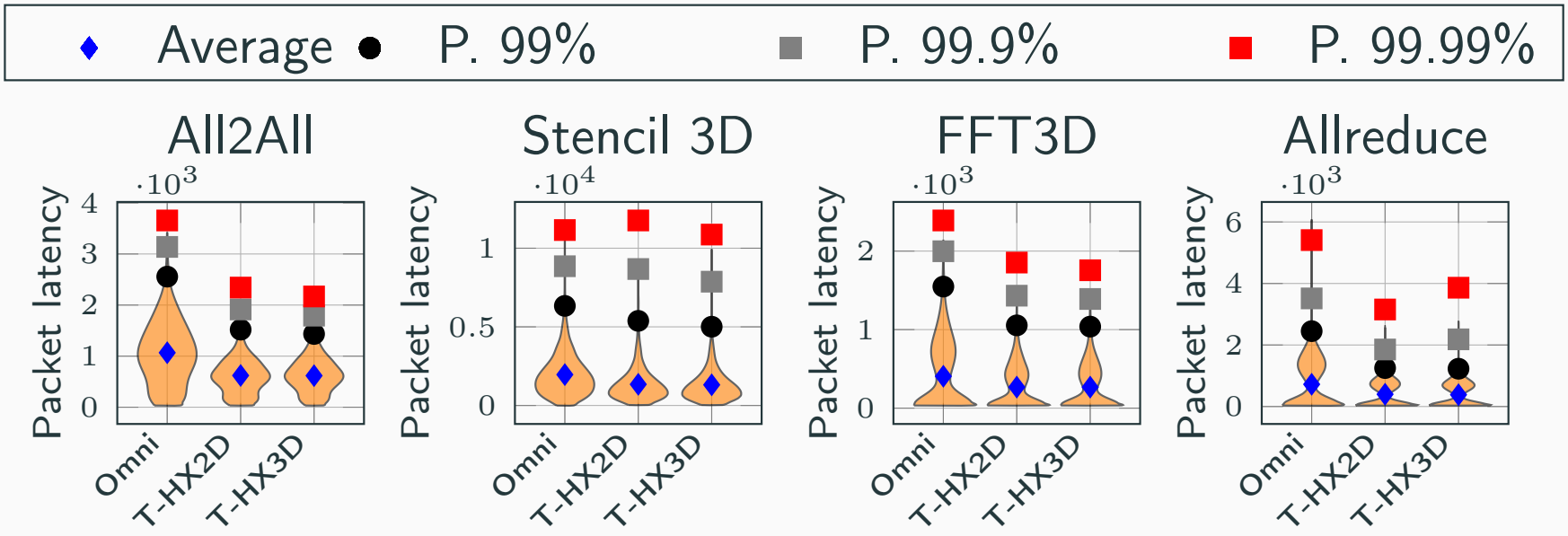
alejandro.cano@unican.es

HOT INTERCONNECTS

Reproducibility of the paper

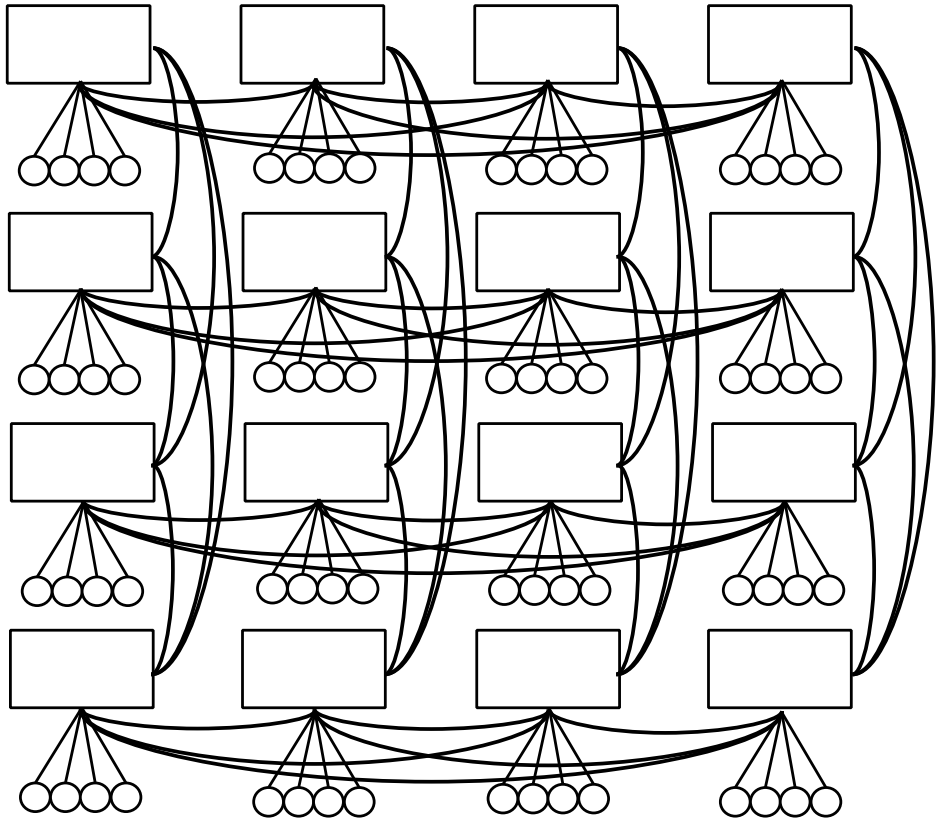
UC

Packet latency percentiles



Violin plots (histogram) of the packet latency.

Hyperx 2D



HyperX 2D topology.