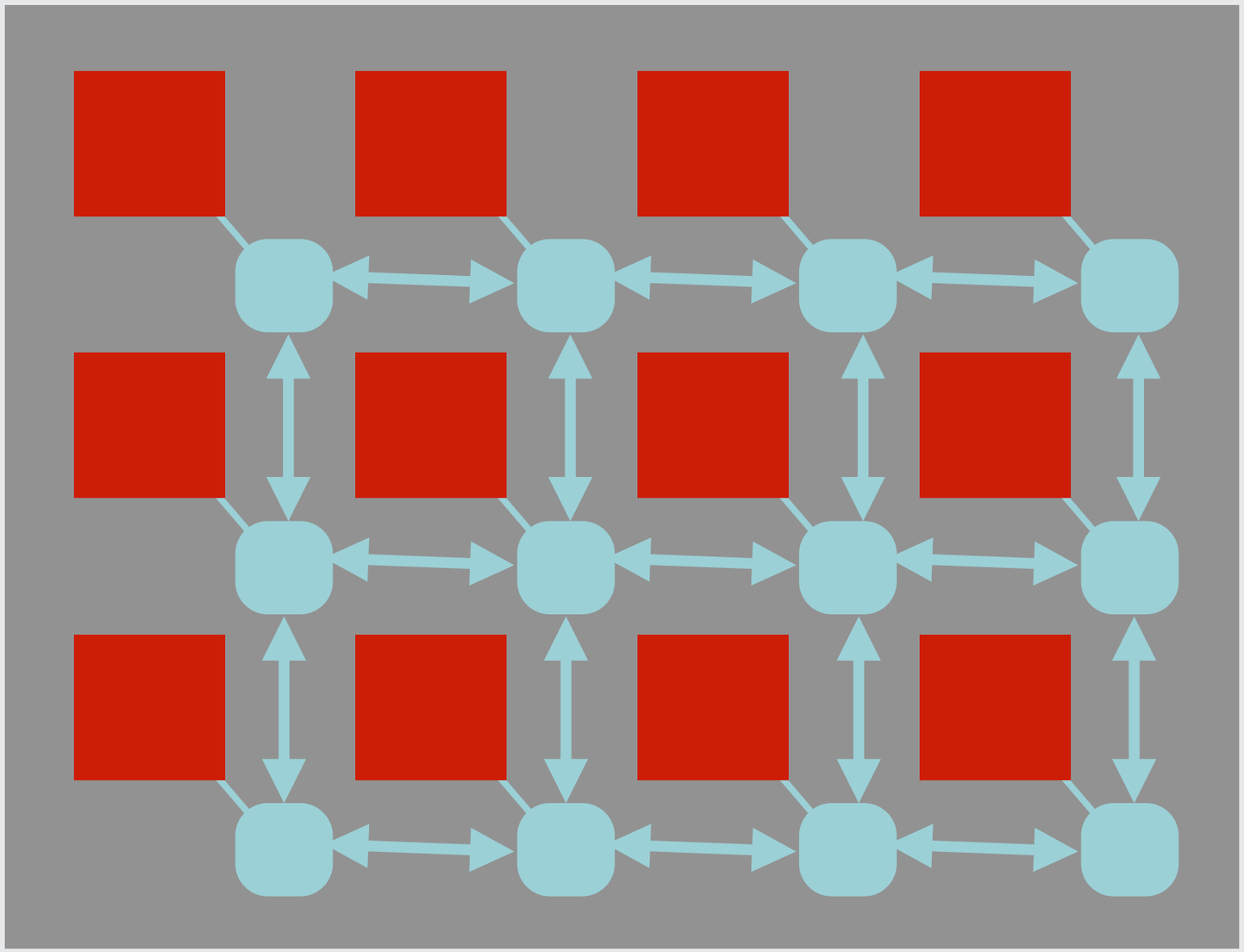
POLYMORPHIC ON-CHIP NETWORKS

Martha Mercaldi Kim, John D. Davis*, Mark Oskin, Todd Austin**

University of Washington
*Microsoft Research, Silicon Valley
** University of Michigan

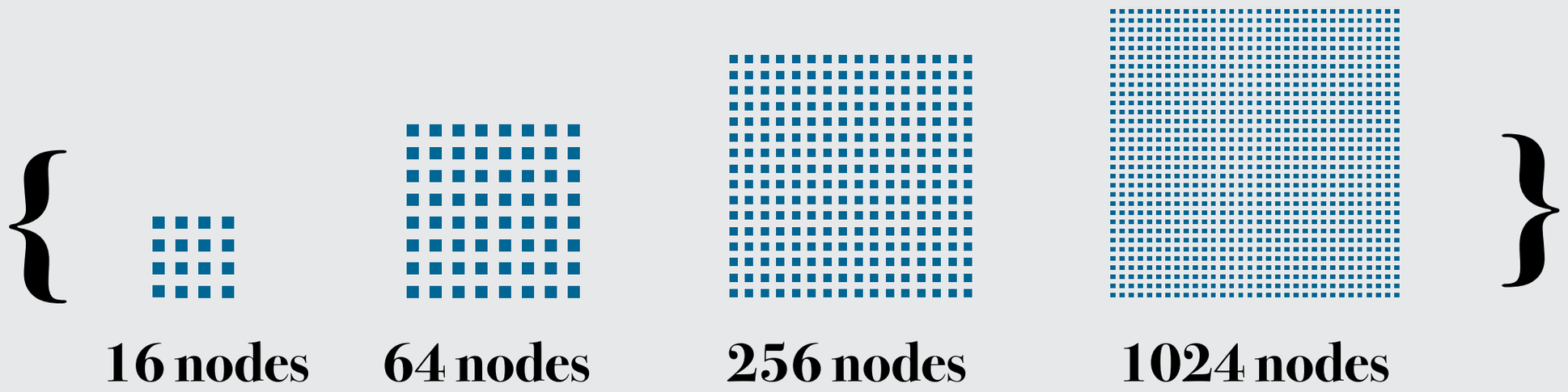
On-Chip Network Selection



Talk Outline

- Network-on-chip Design Space Exploration
  - *Networks*
  - *Workloads*
  - *Results*
- Polymorphic On-Chip Networks
  - *Fabric design*
  - *Configuring the network*
  - *Selecting a fabric*
  - *Evaluation of flexibility*
- Conclusions
- Future Directions

On-Chip Network Design Space

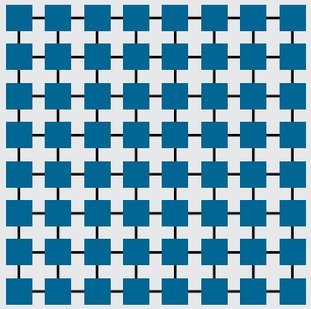


On-Chip Network Design Space

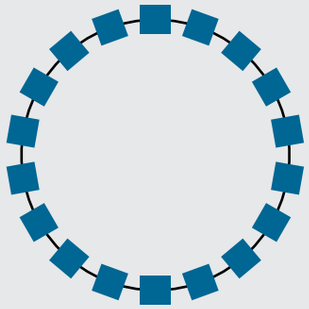
{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

x

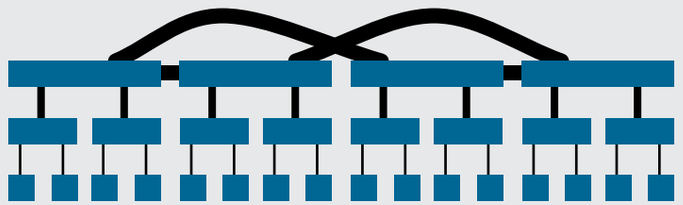
{



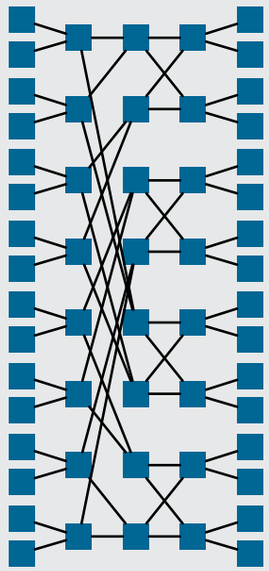
mesh



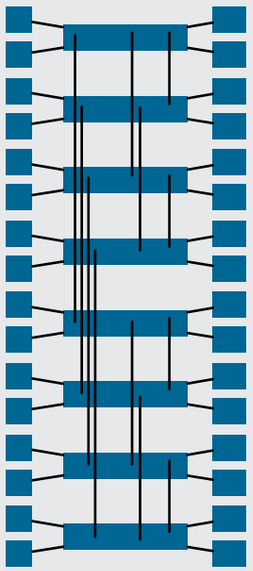
ring



fat-tree



butterfly



flattened-butterfly

}

On-Chip Network Design Space

{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

x

{ mesh, ring, fat-tree, butterfly, flattened-butterfly }

x

{ minimal, oblivious, source-routing }

x

On-Chip Network Design Space

{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

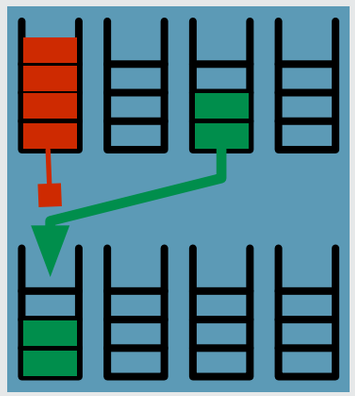
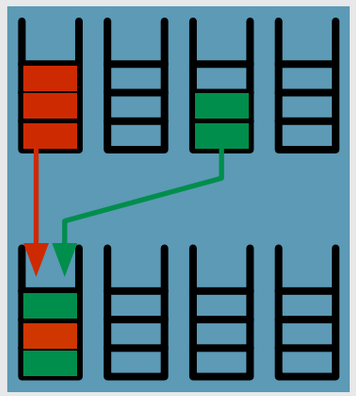
x

{ mesh, ring, fat-tree, butterfly, flattened-butterfly }

x

{ minimal, oblivious, source-routing }

x

{   }

wormhole **store-and-forward**

On-Chip Network Design Space

{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

x

{ mesh, ring, fat-tree, butterfly, flattened-butterfly }

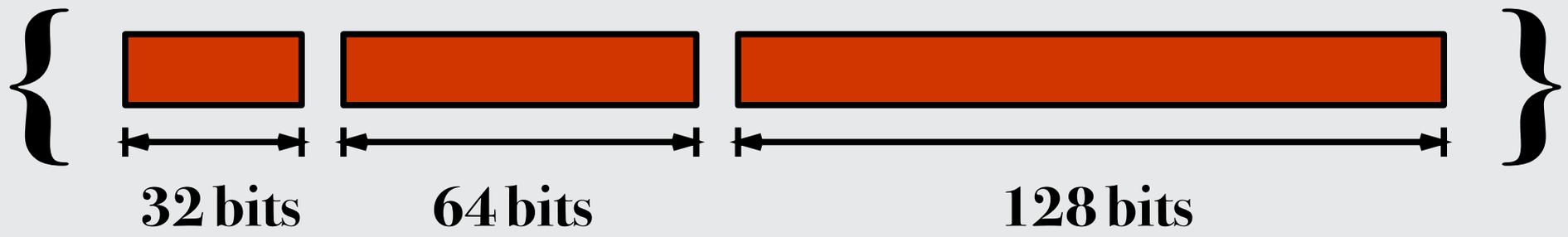
x

{ minimal, oblivious, source-routing }

x

{ wormhole, store-and-forward }

x



On-Chip Network Design Space

{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

x

{ mesh, ring, fat-tree, butterfly, flattened-butterfly }

x

{ minimal, oblivious, source-routing }

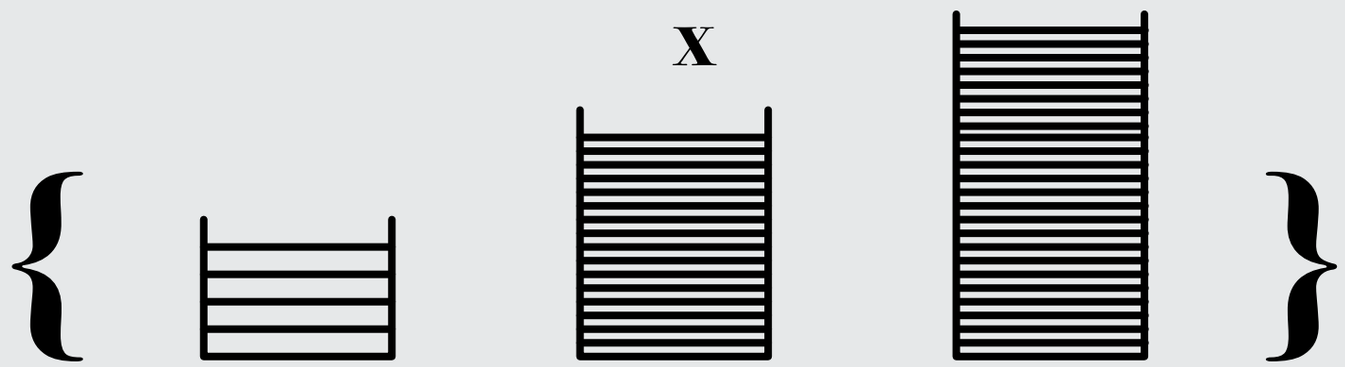
x

{ wormhole, store-and-forward }

x

{ 16 bits, 64 bits, 128 bits }

x



On-Chip Network Design Space

{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

x

{ mesh, ring, fat-tree, butterfly, flattened-butterfly }

x

{ minimal, oblivious, source-routing }

x

{ wormhole, store-and-forward }

x

{ 16 bits, 64 bits, 128 bits }

x

{ 4 entries, 16 entries, 64 entries }

=

360 on-chip networks

On-Chip Network Design Space

{ 16 nodes, 64 nodes, 256 nodes, 1024 nodes }

x

{ mesh, ring, fat-tree, butterfly, flattened-butterfly }

x

{ minimal, oblivious, source-routing }

x

{ wormhole, store-and-forward }

x

{ 16 bits, 64 bits, 128 bits }

x

{ 4 entries, 16 entries, 64 entries }

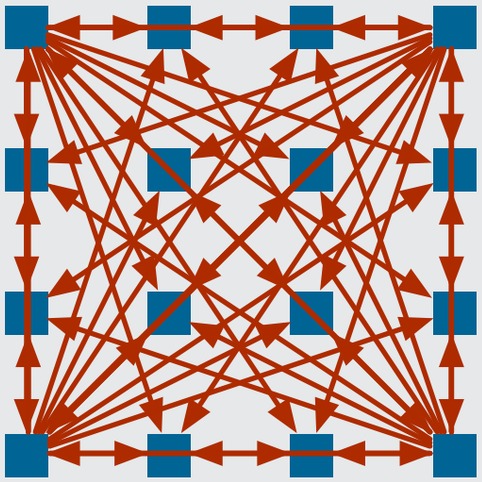
=

360 on-chip networks

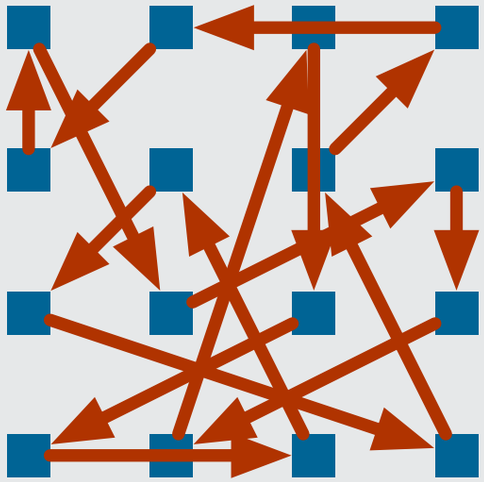
cycle-level software simulator

Network Traffic Patterns

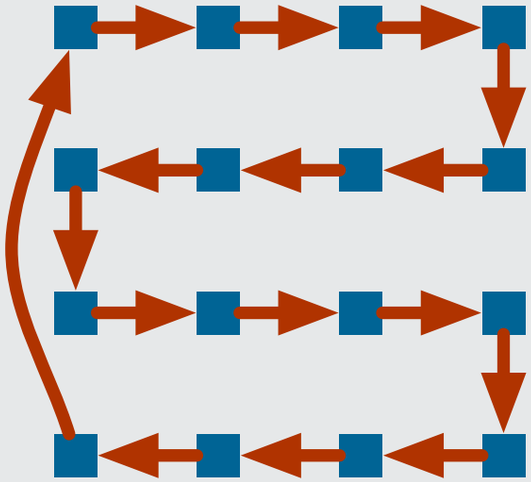
{



Uniform Random



Random Permutation



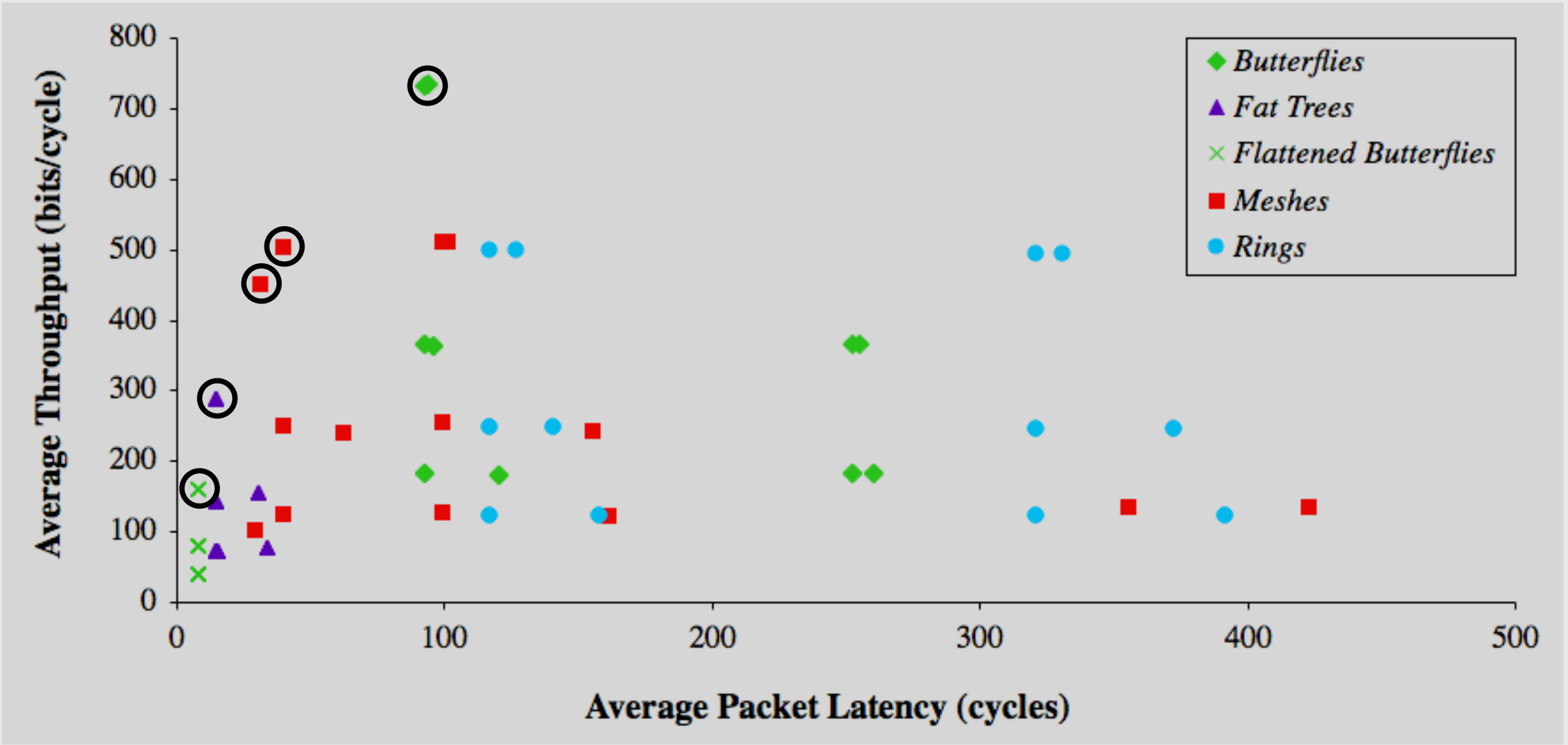
Nearest Neighbor

}

(injection rate = 1 packet / cycle)

Network Measurements: Random Permutation

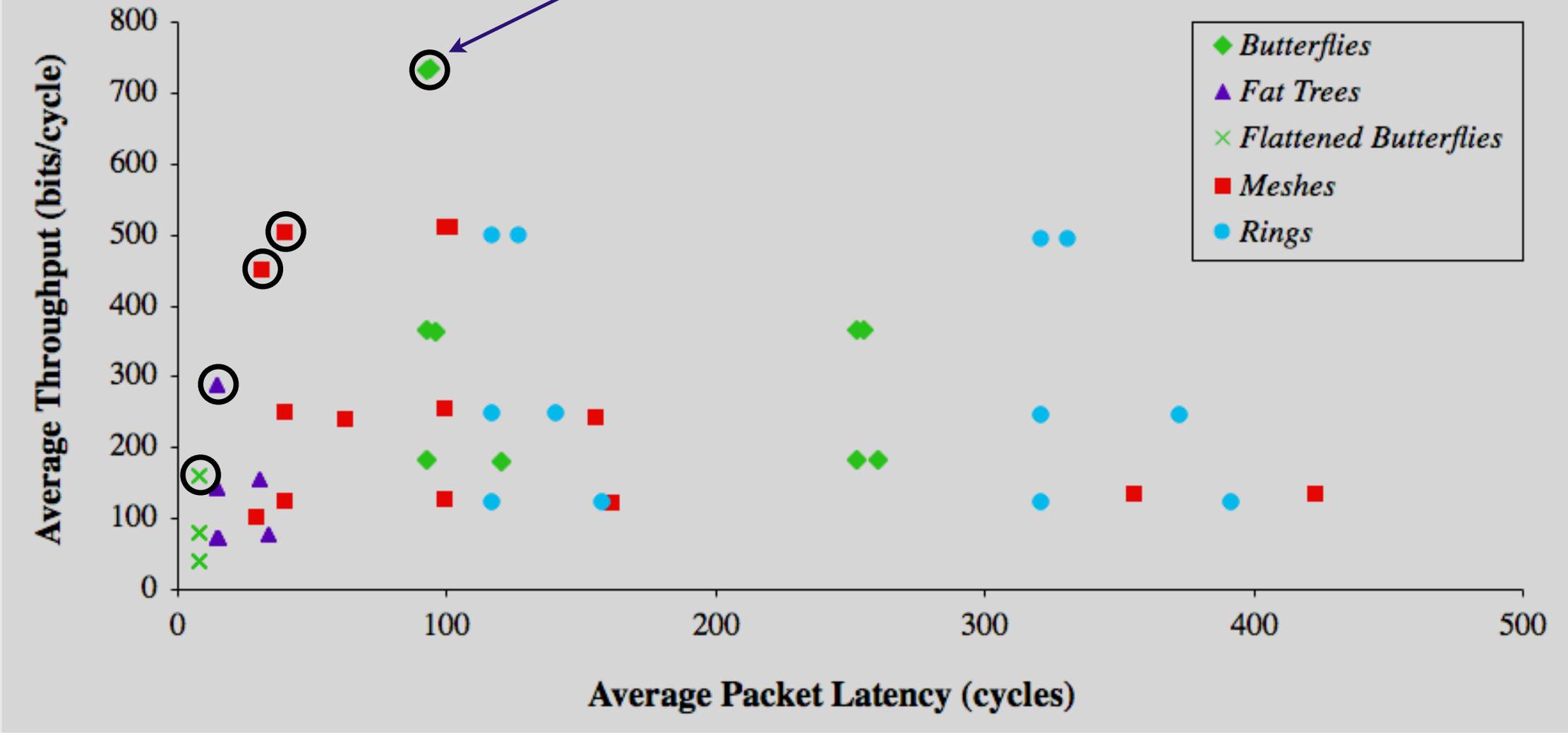
(Random Permutation)



Network Measurements: Random Permutation

throughput-sensitive application

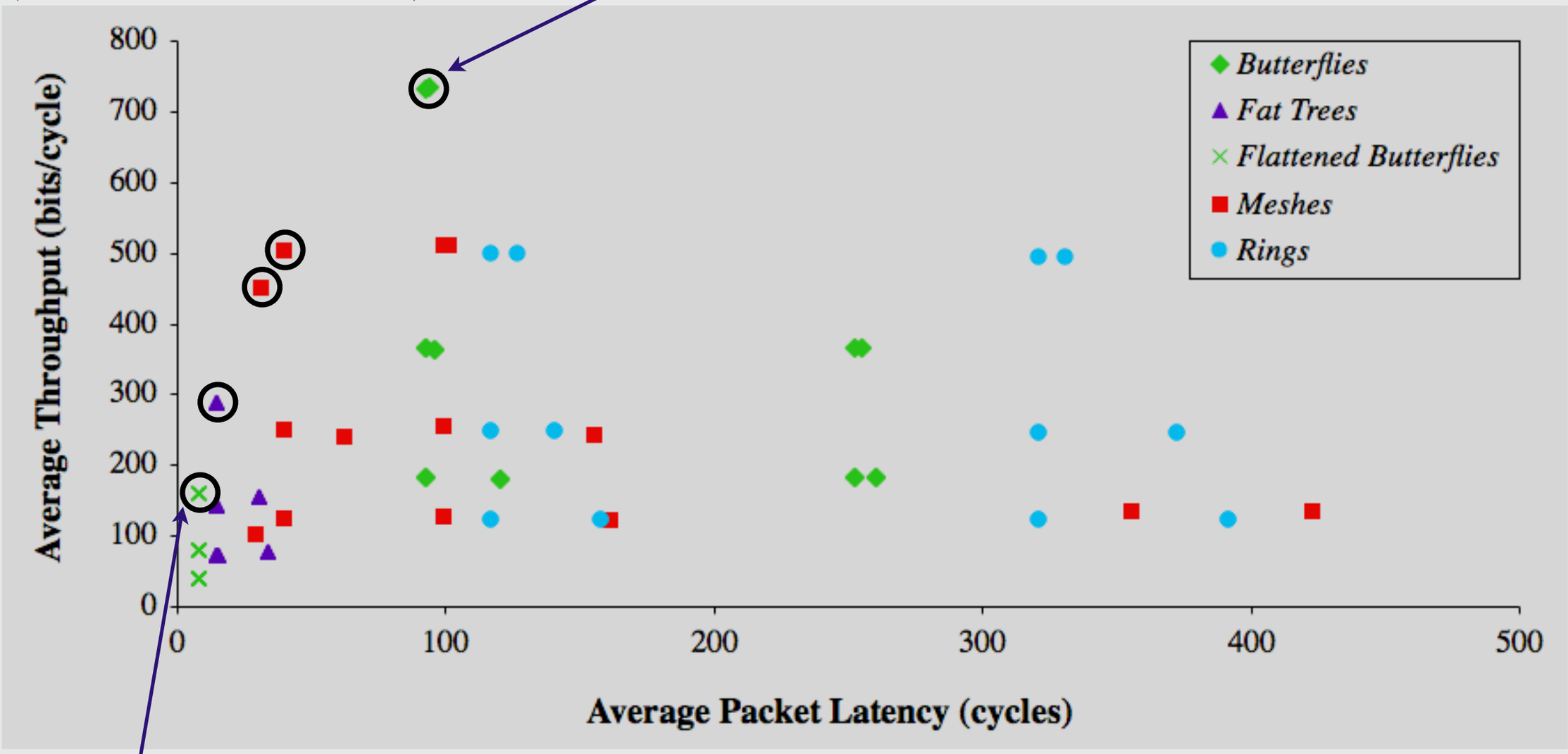
(Random Permutation)



Network Measurements: Random Permutation

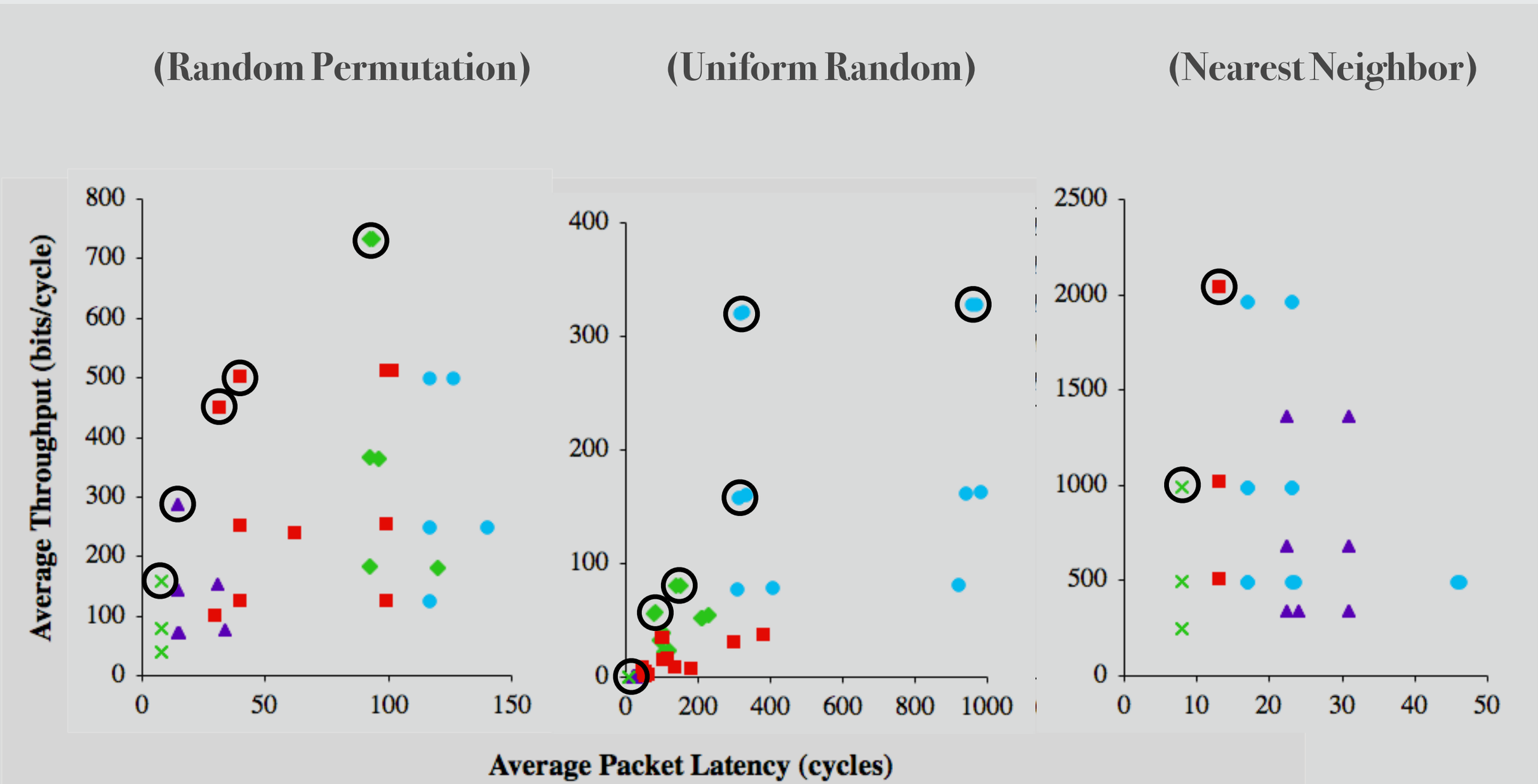
throughput-sensitive application

(Random Permutation)



latency-sensitive application

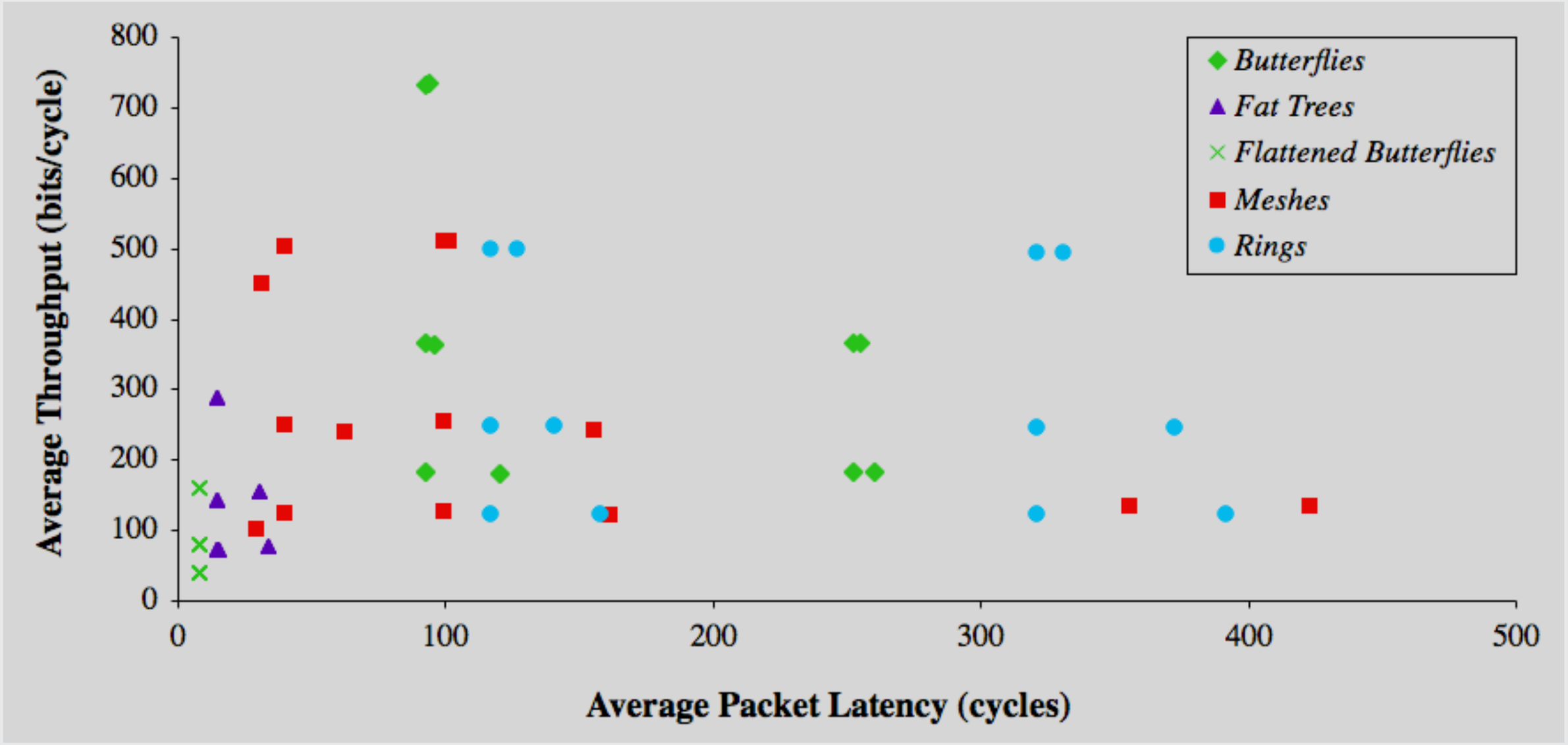
Network Measurements



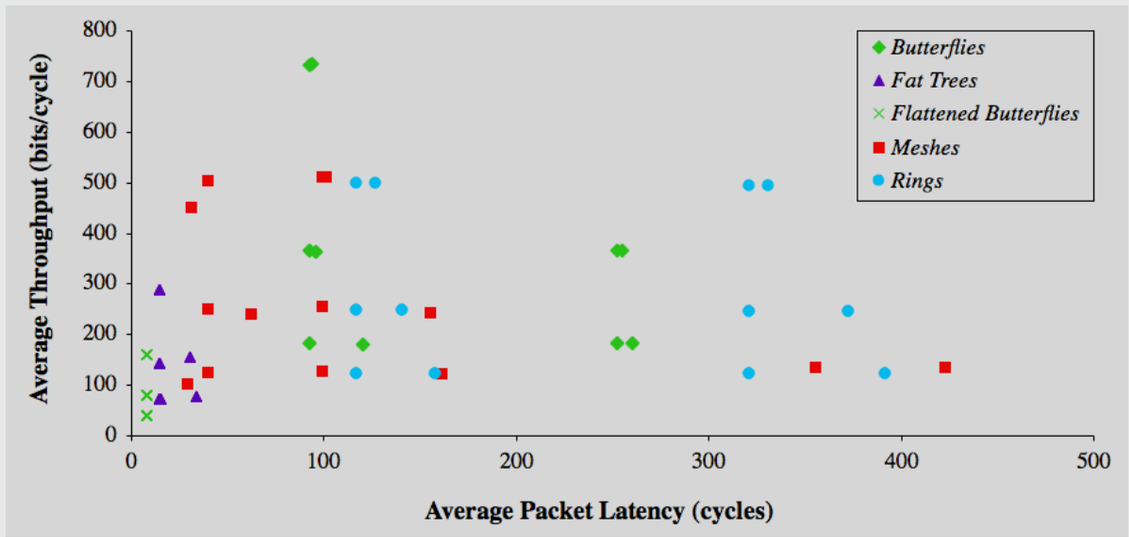
Talk Outline

- Network-on-chip Design Space Exploration
  - *Networks*
  - *Workloads*
  - *Results*
- Polymorphic On-Chip Networks
  - *Fabric design*
  - *Configuring the network*
  - *Selecting a fabric*
  - *Evaluation of flexibility*
- Conclusions
- Future Directions

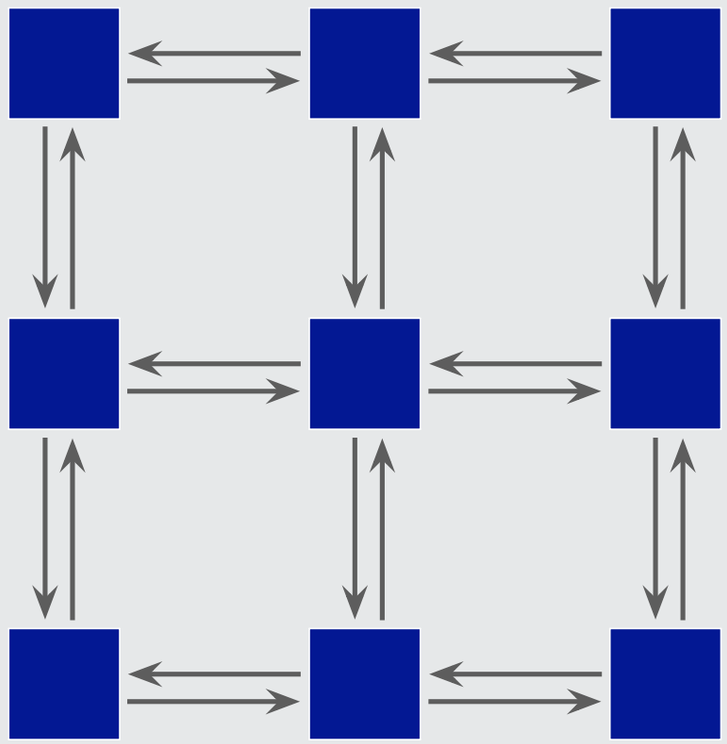
The Intuition…



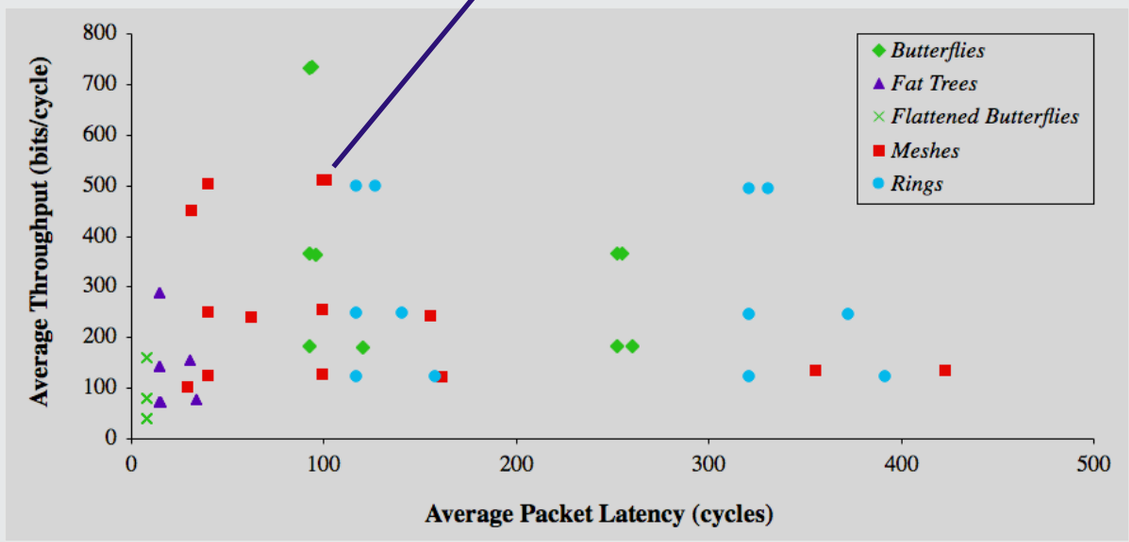
The Intuition…



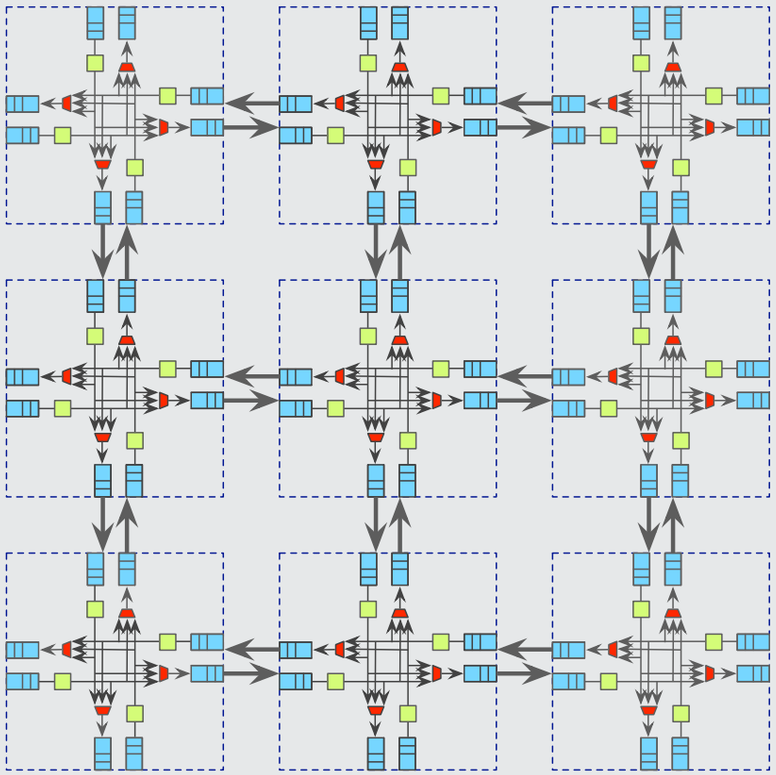
The Intuition…



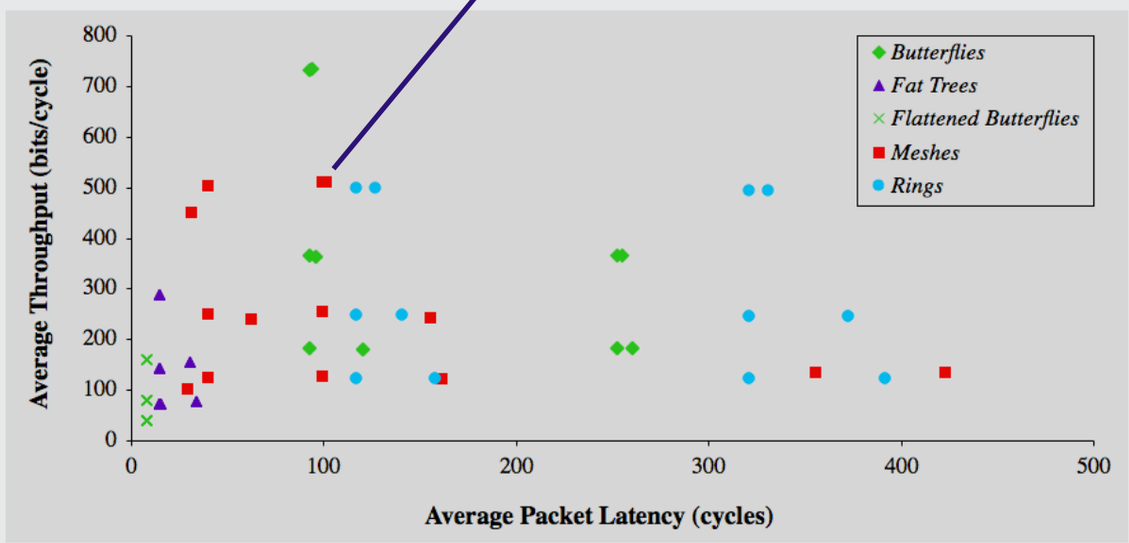
mesh



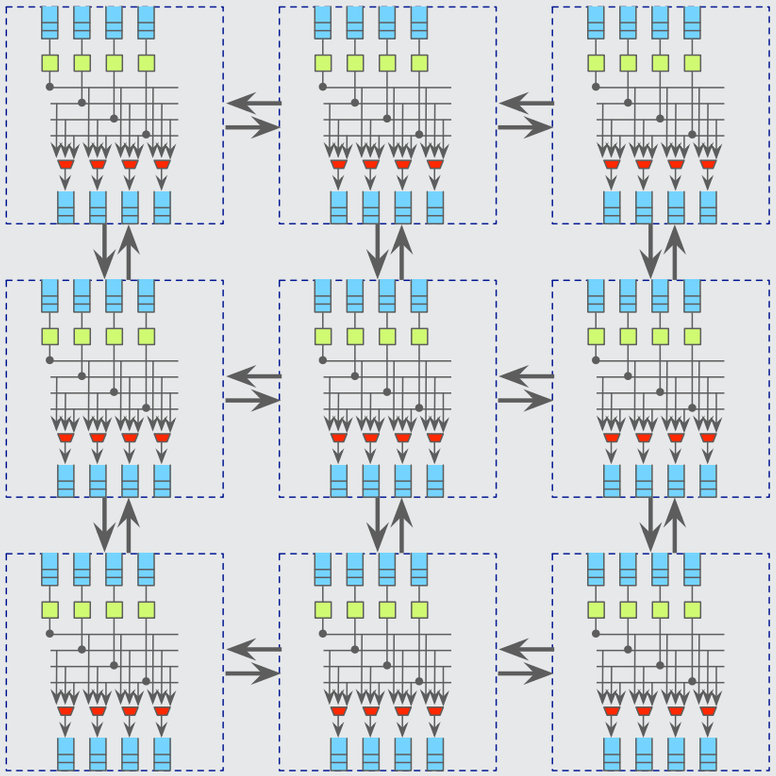
The Intuition…

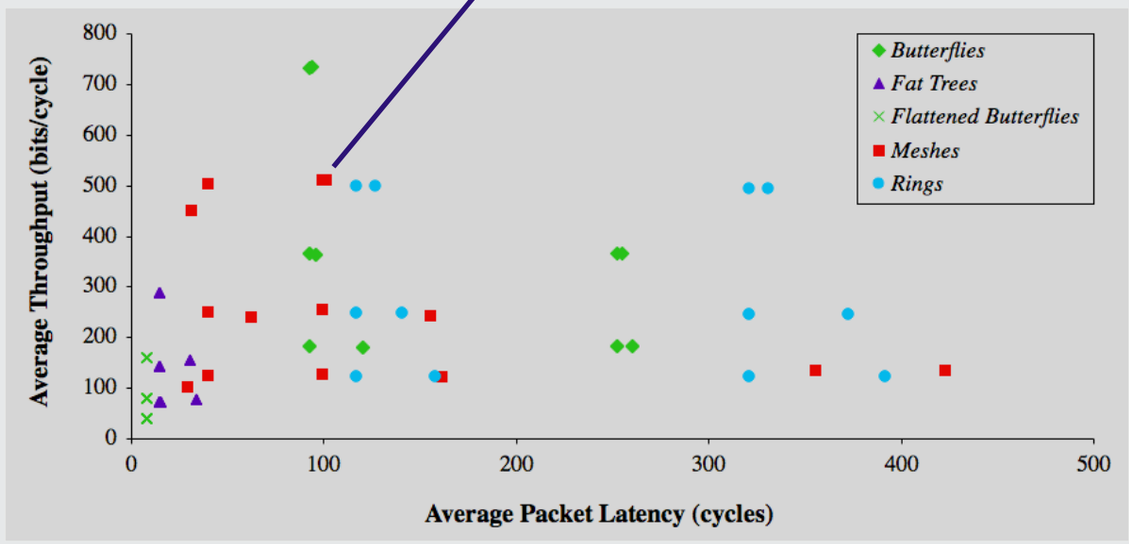


mesh

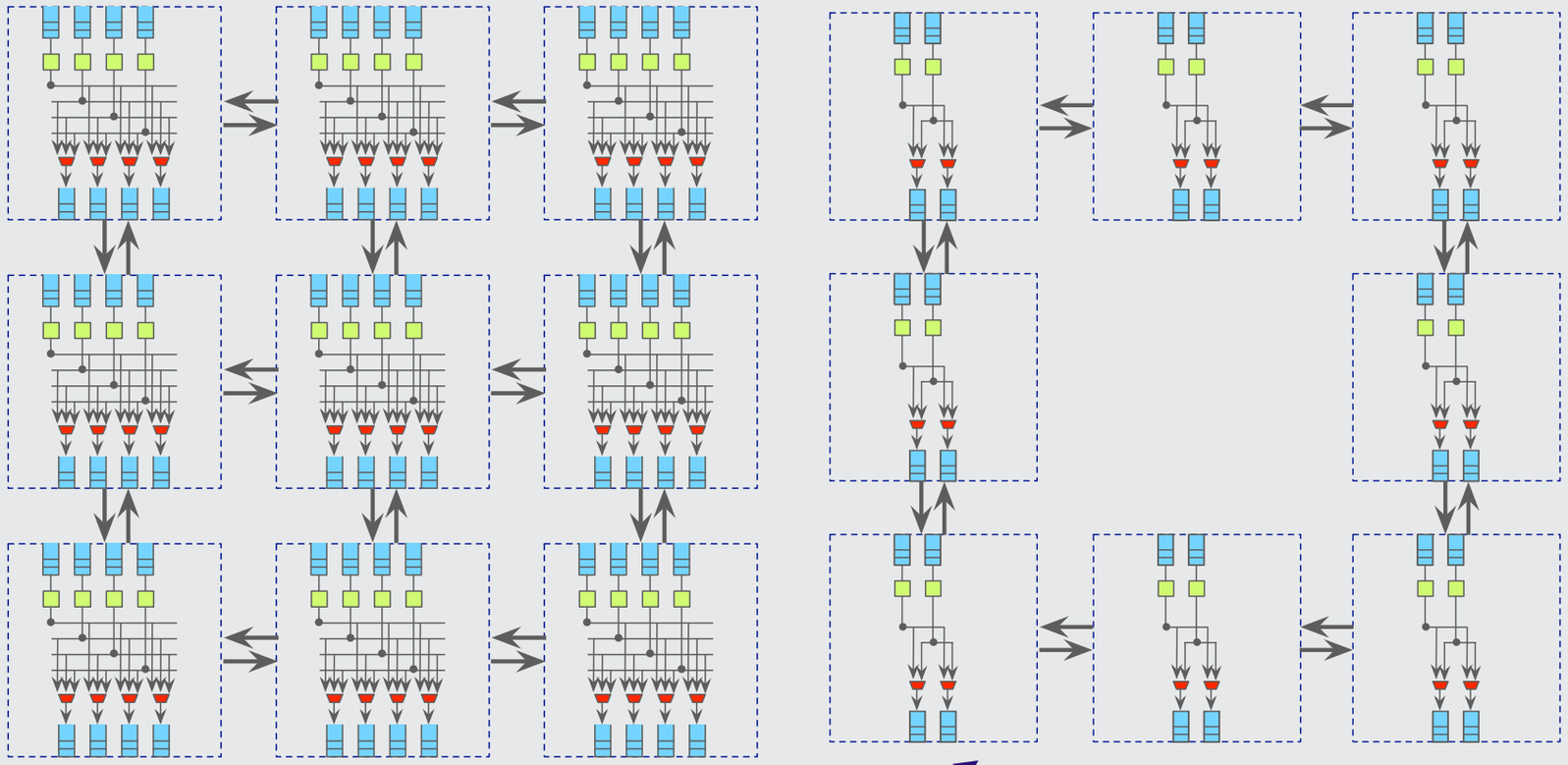


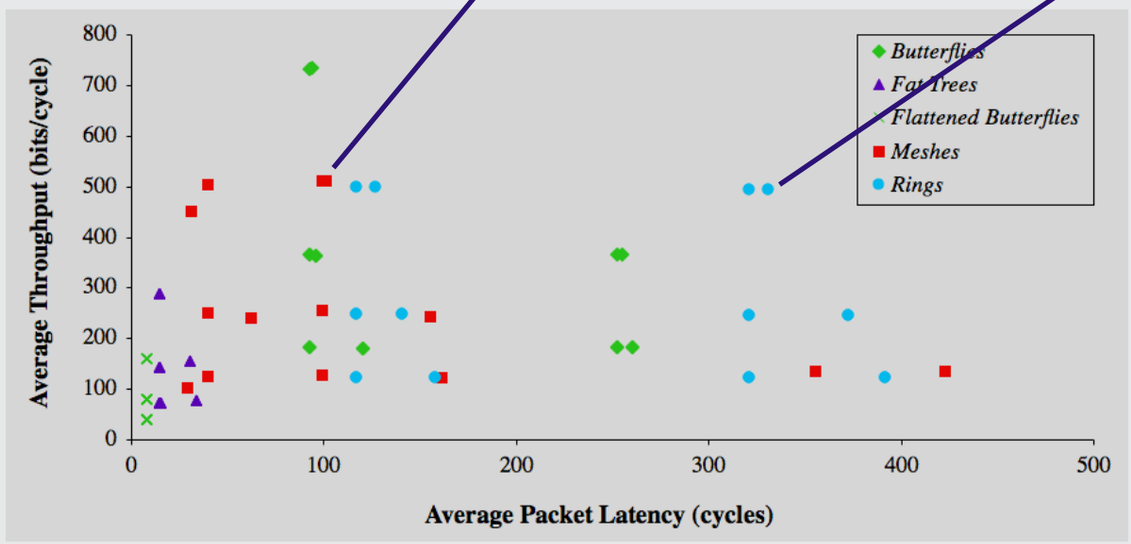
The Intuition…



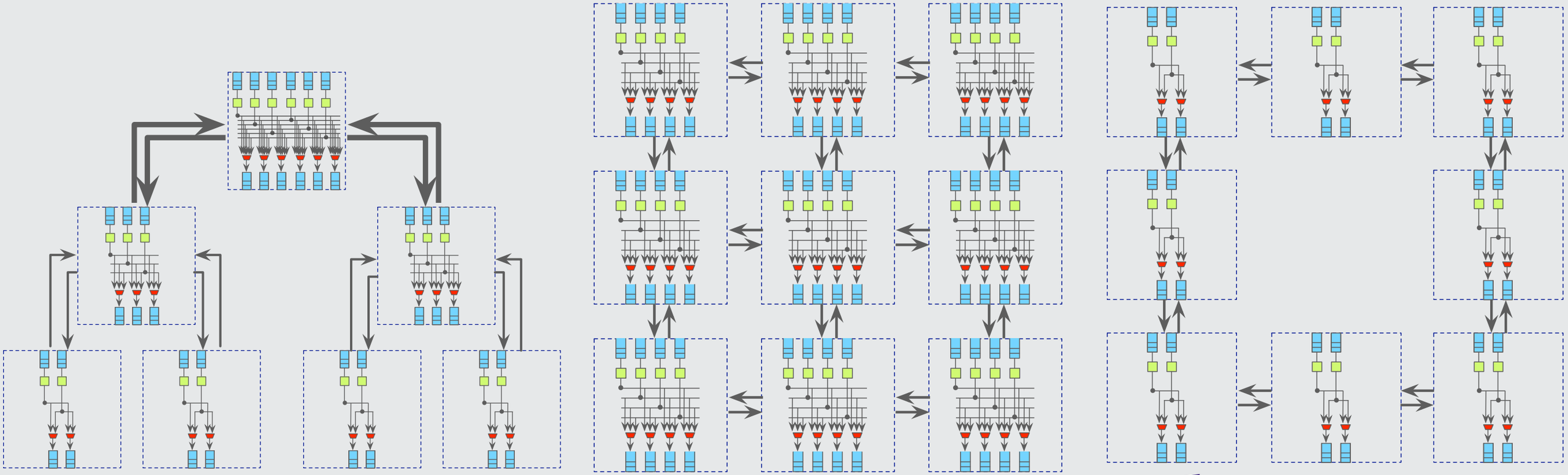


The Intuition…





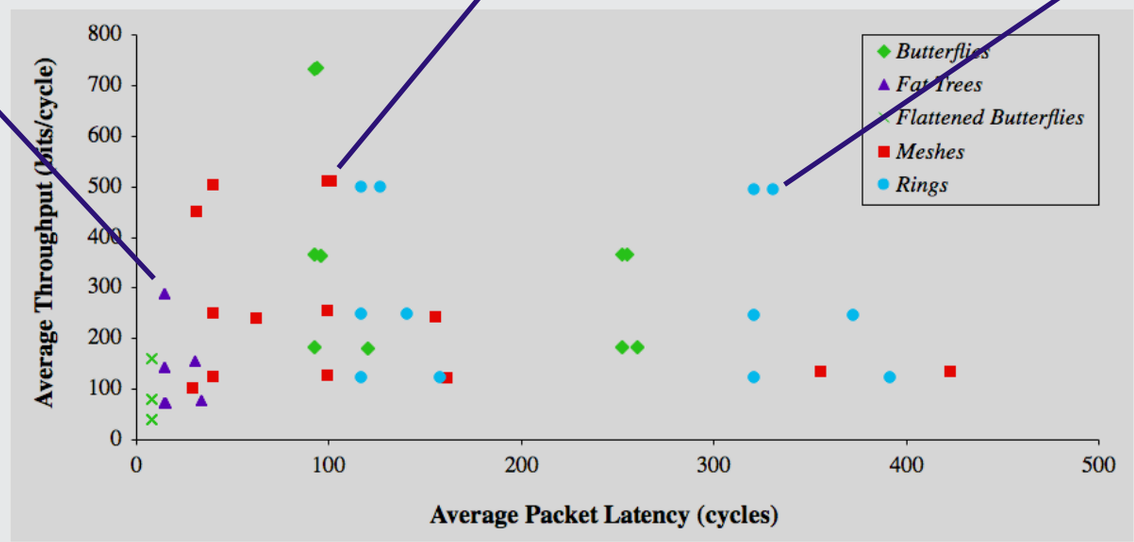
The Intuition…



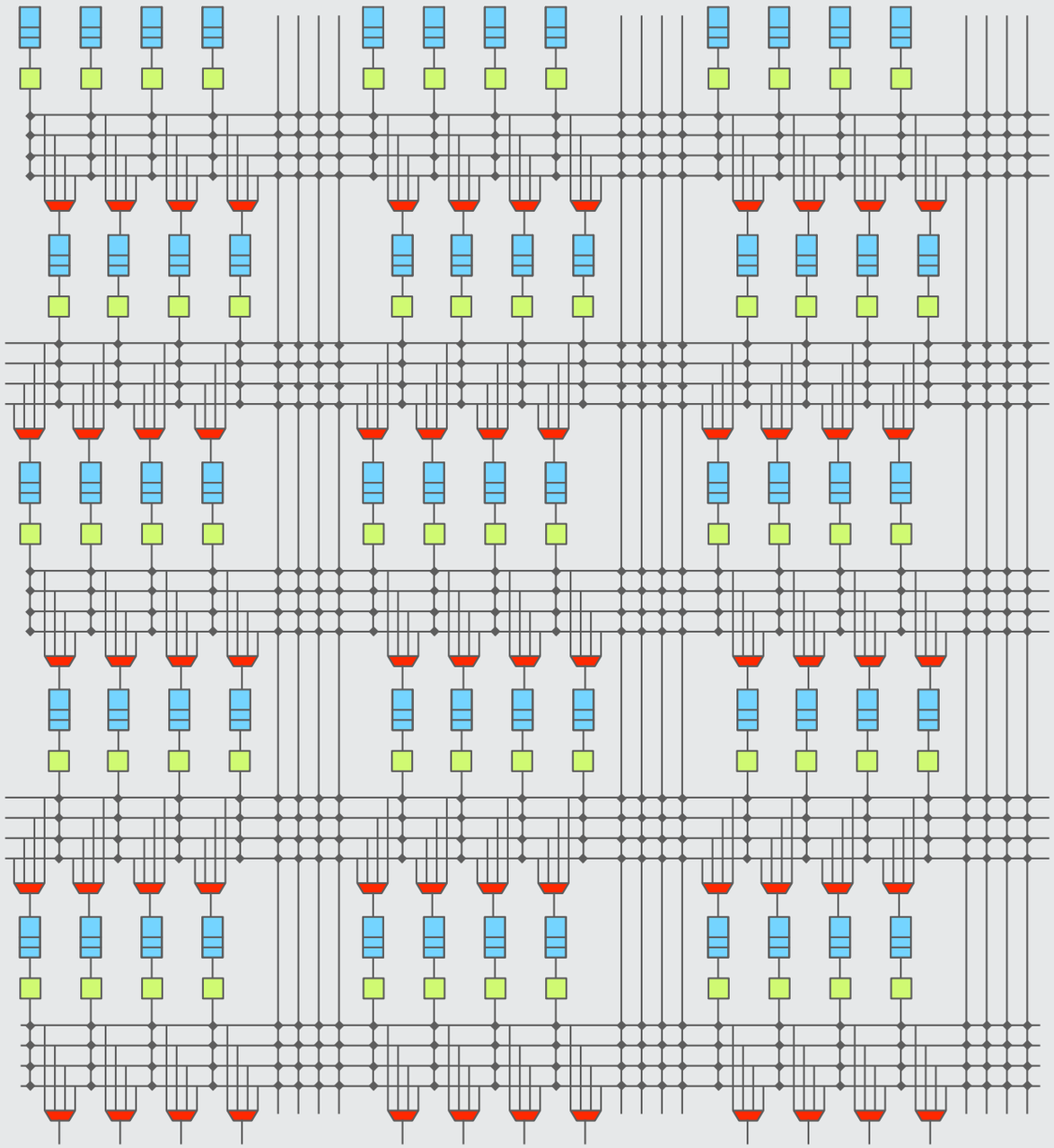
fat tree

mesh

ring



Polymorphic On-Chip Network



What it is

- *Sea of structures all networks have in common*
- *Configurable connections between structures*

How it is used

- *Gather structures to arbitrary-degree switches*
- *Connect switches input and output ports*

Talk Outline

- Network-on-chip Design Space Exploration
  - *Networks*
  - *Workloads*
  - *Results*
- Polymorphic On-Chip Networks
  - *Fabric design*
  - *Configuring the network*
  - *Selecting a fabric*
  - *Evaluation of flexibility*
- Conclusions
- Future Directions

Configuring the Network

1. Switch degree

2. Inter-switch connections

3. Packet width

4. Buffer capacity

Configuring the Network

1. Switch degree

2. Inter-switch connections

→ configurable topology

3. Packet width

4. Buffer capacity

Configuring the Network

1. Switch degree

2. Inter-switch connections

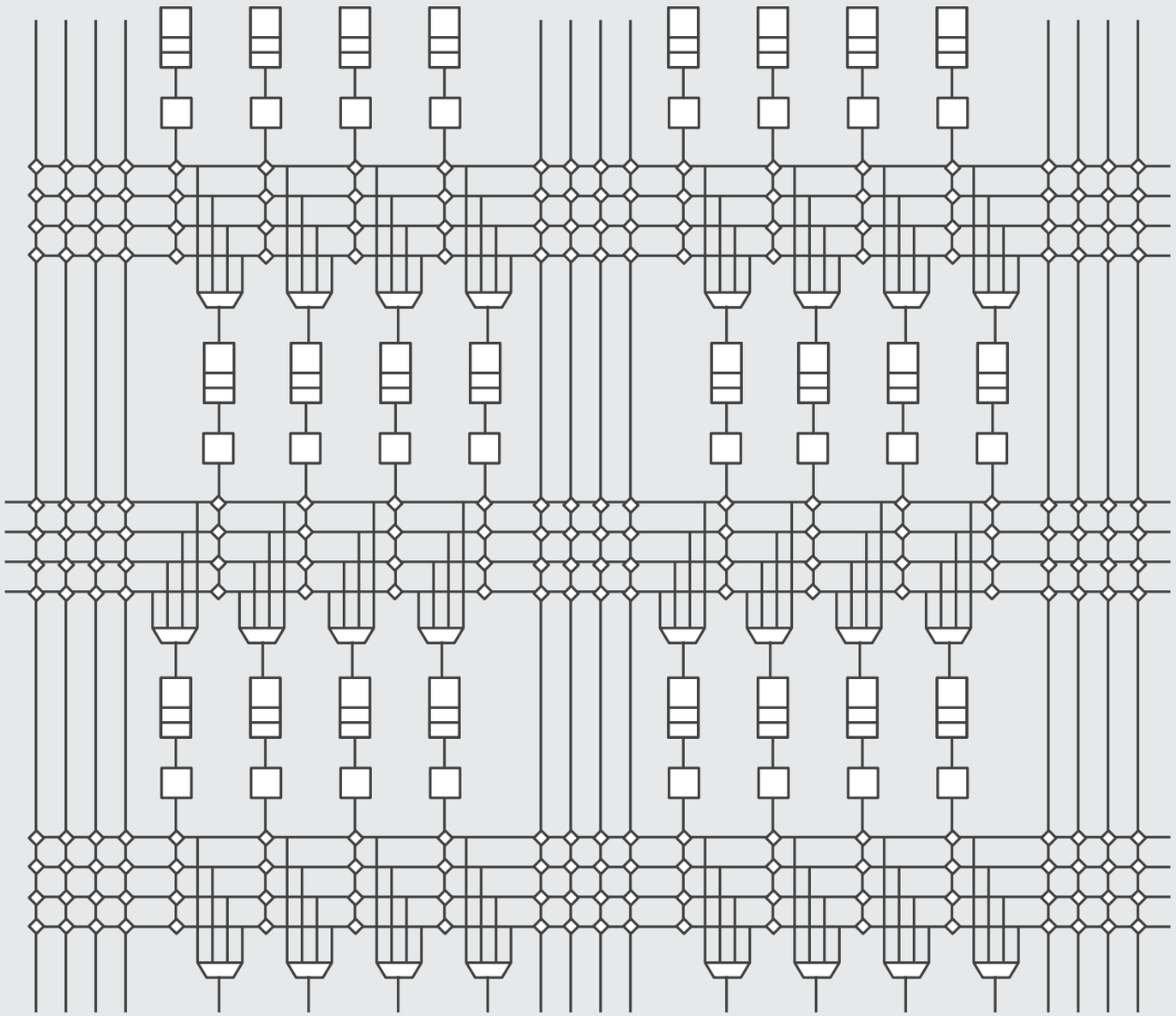
→ configurable topology

3. Packet width

4. Buffer capacity

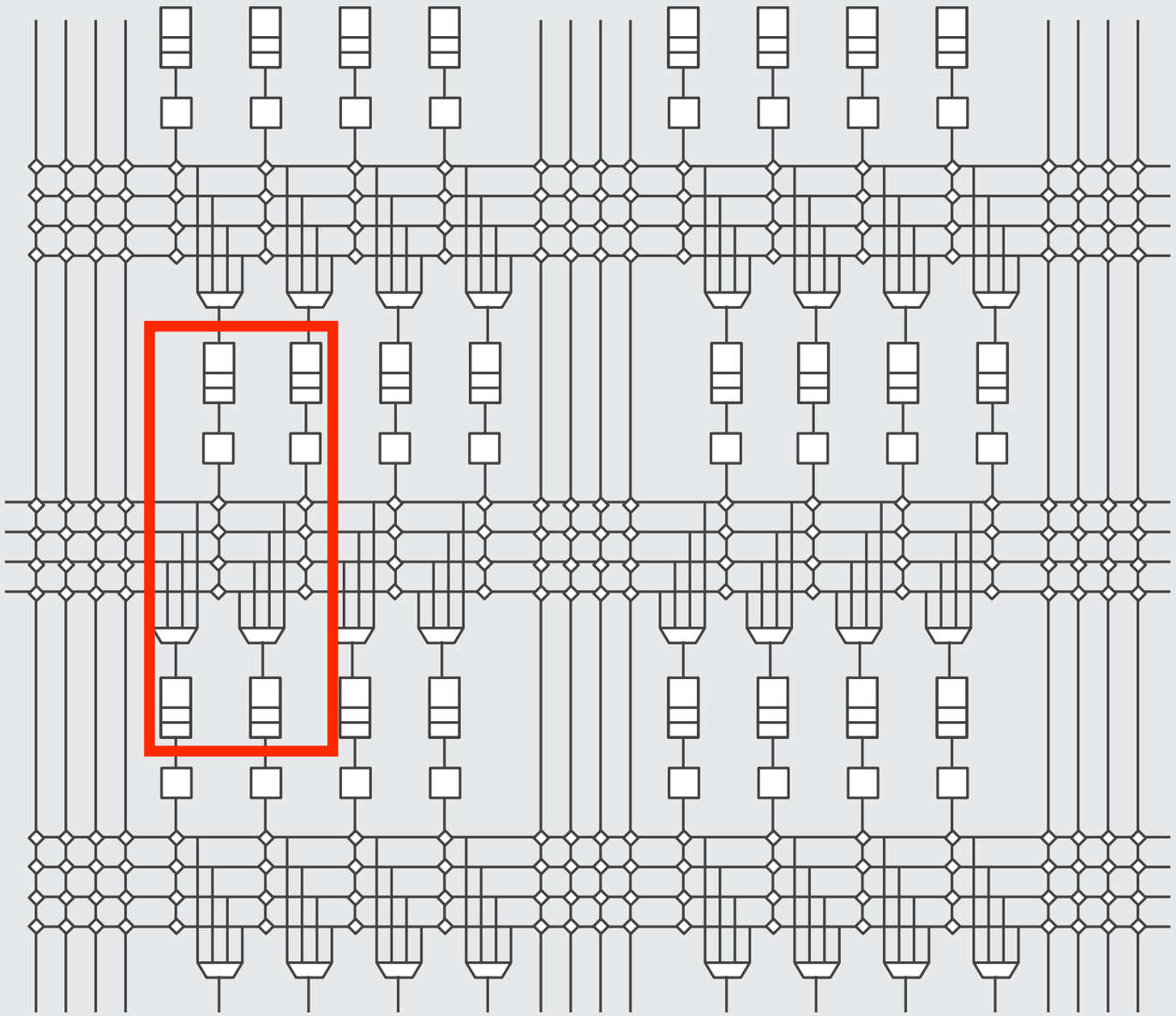
→ configurable resource allocation

Network Configuration: Switch Degree



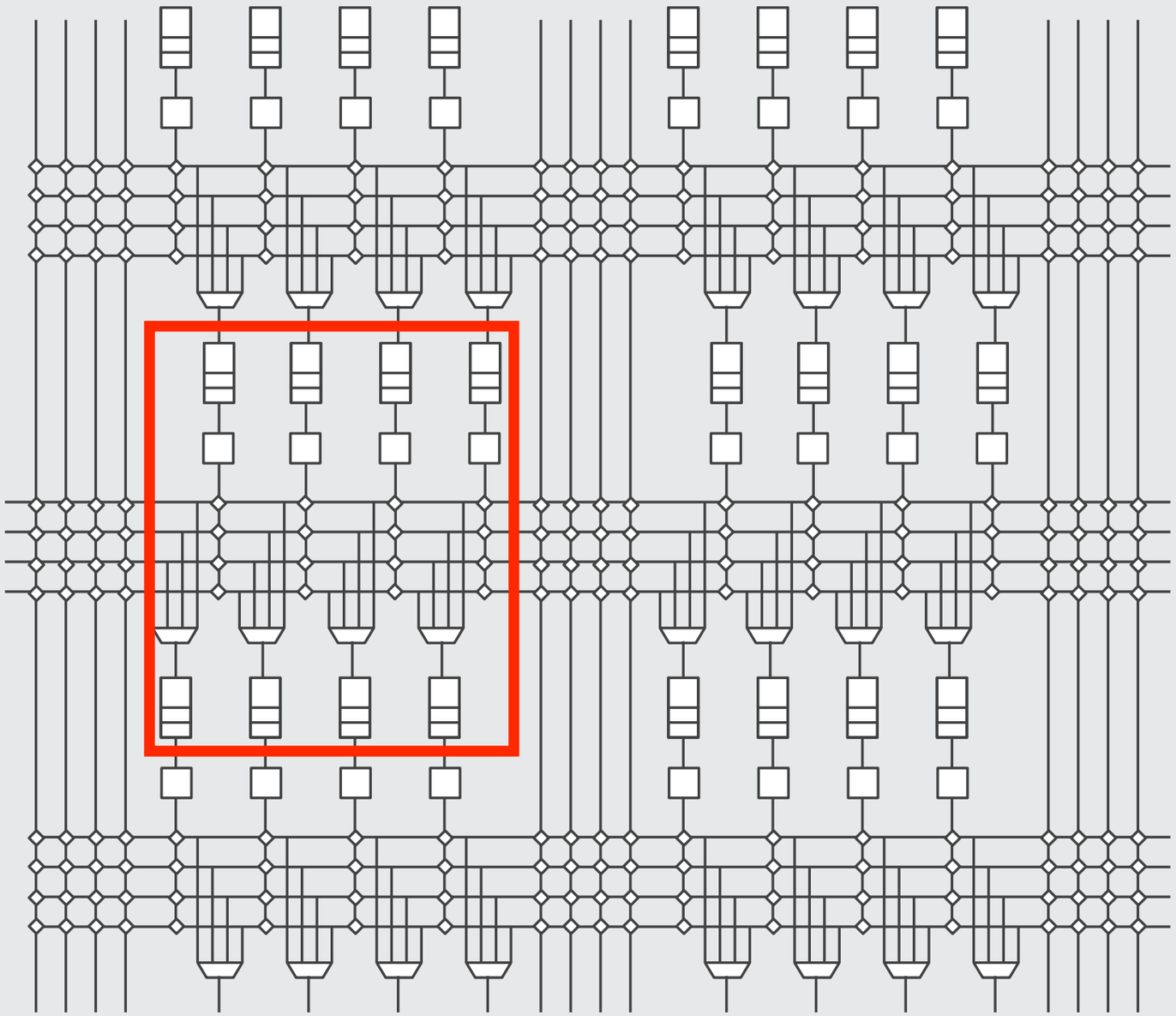
1. Switch degree

Network Configuration: Switch Degree



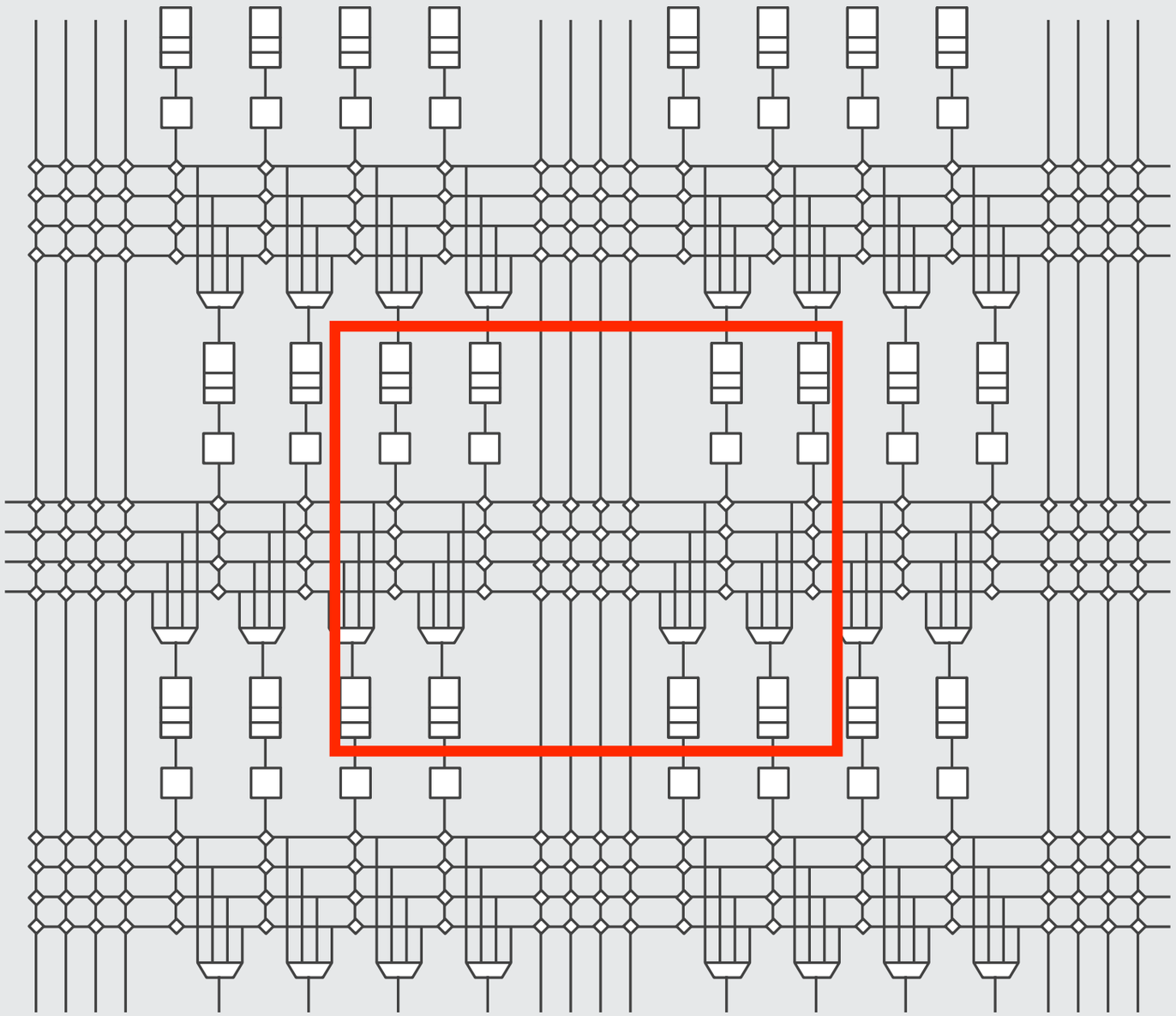
1. Switch degree

Network Configuration: Switch Degree



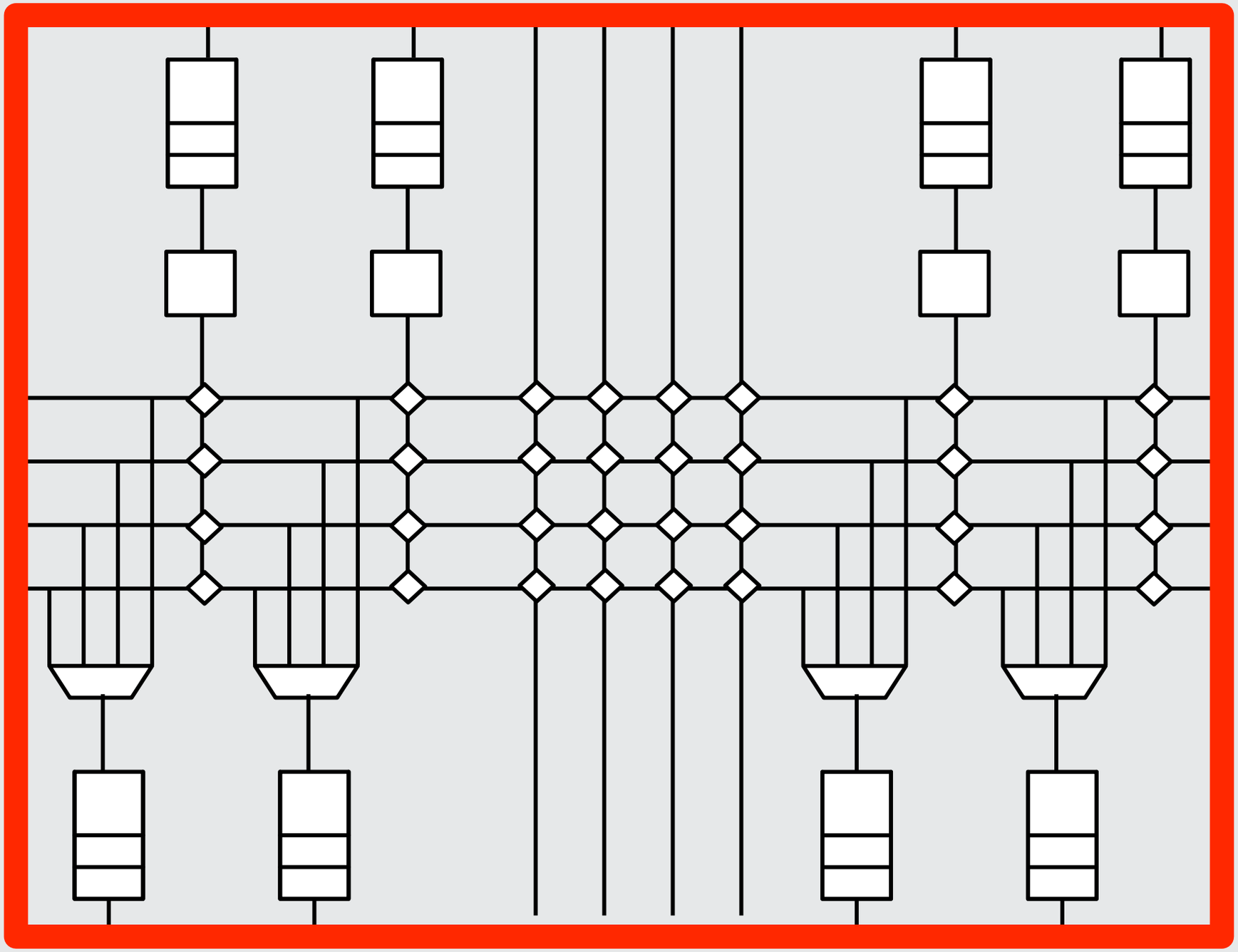
1. Switch degree

Network Configuration: Switch Degree

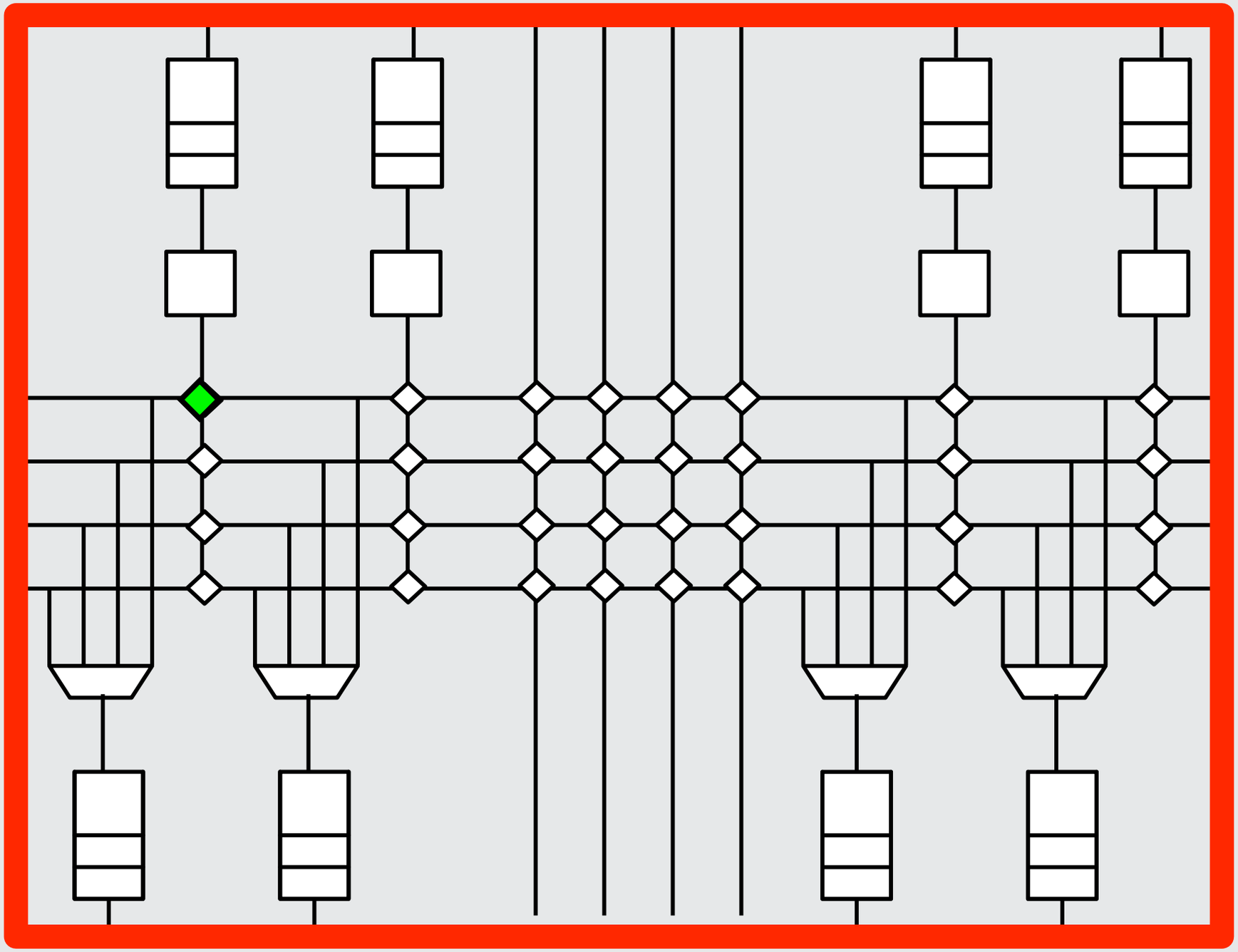


1. Switch degree

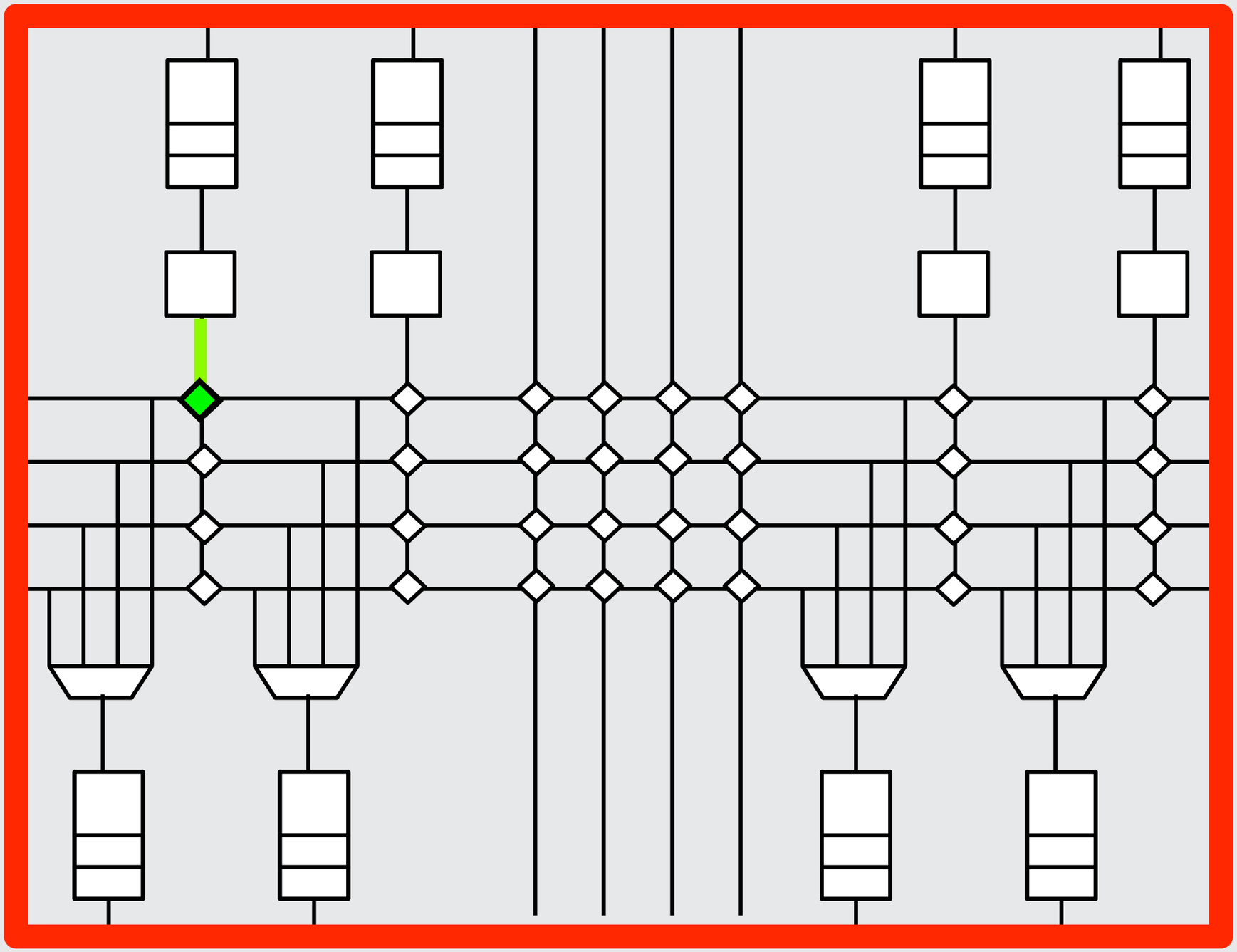
Internal Configuration of a Switch



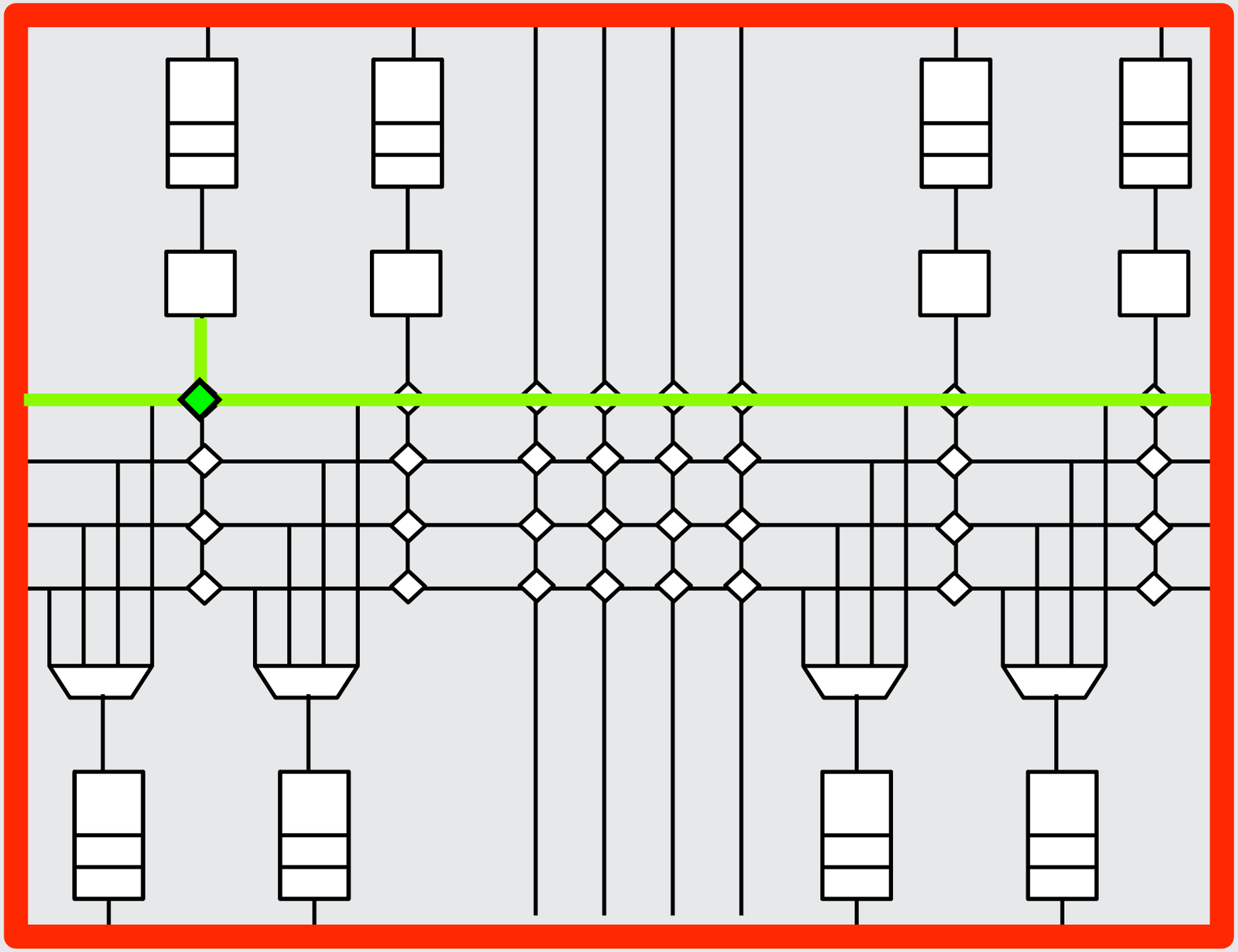
Internal Configuration of a Switch



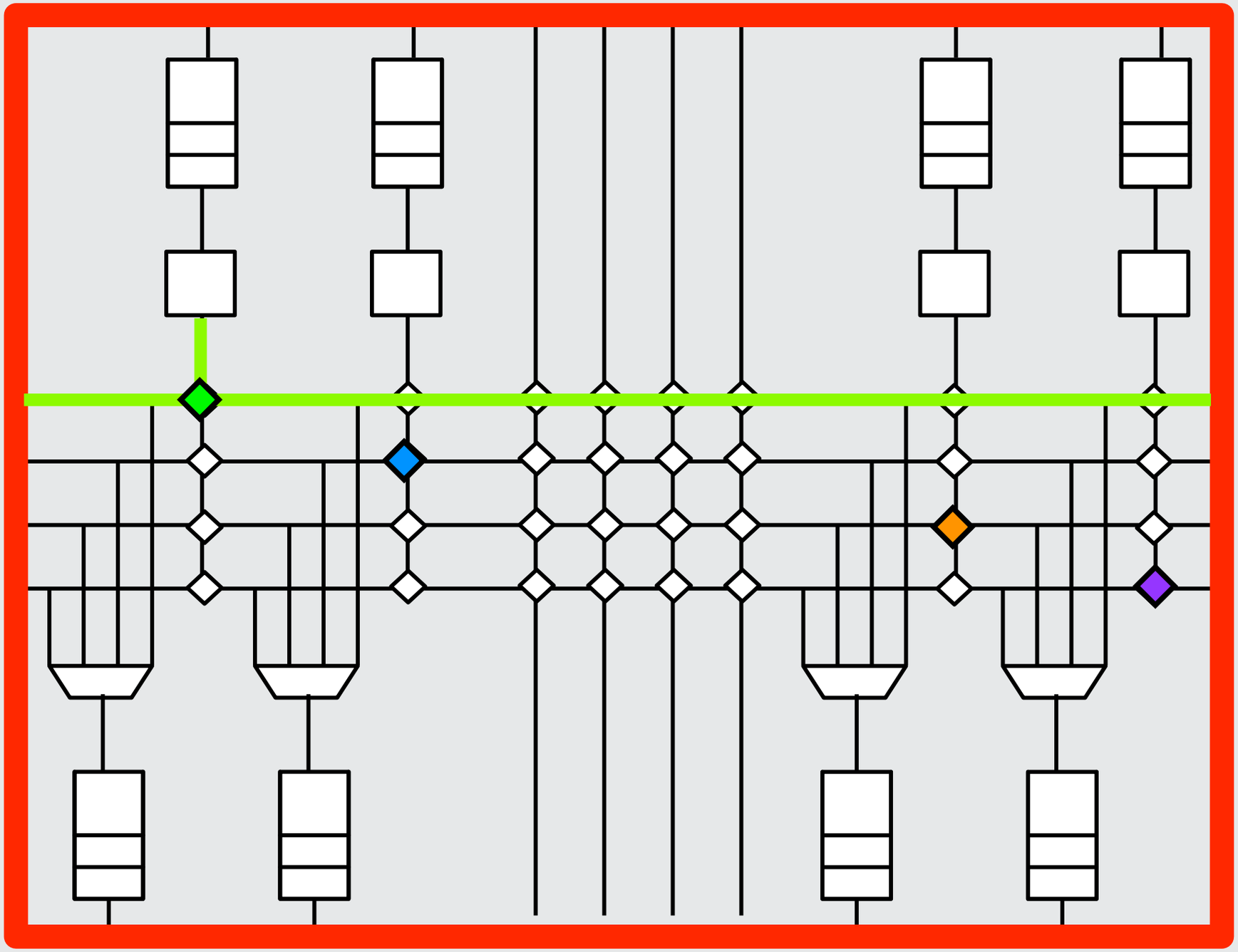
Internal Configuration of a Switch



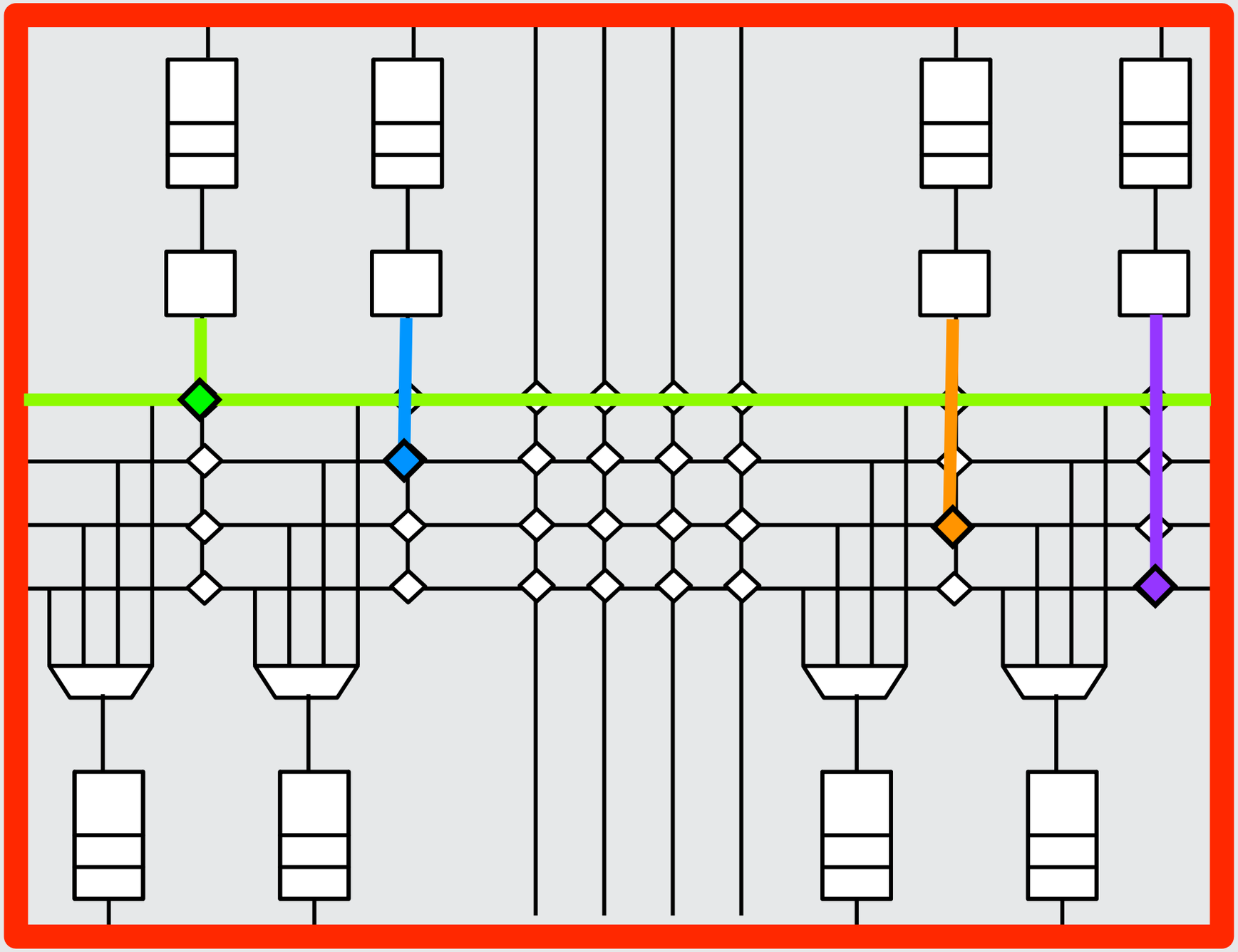
Internal Configuration of a Switch



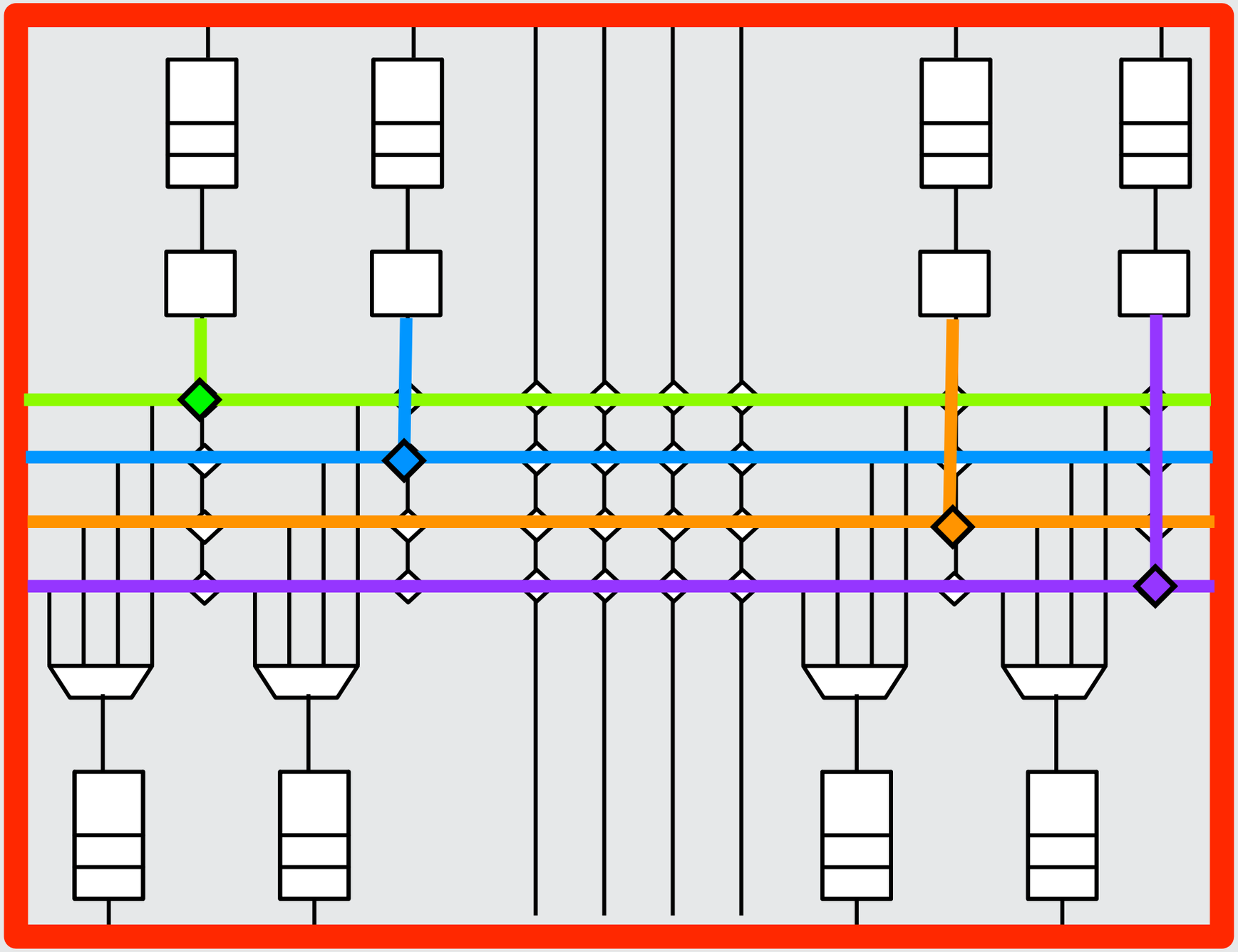
Internal Configuration of a Switch



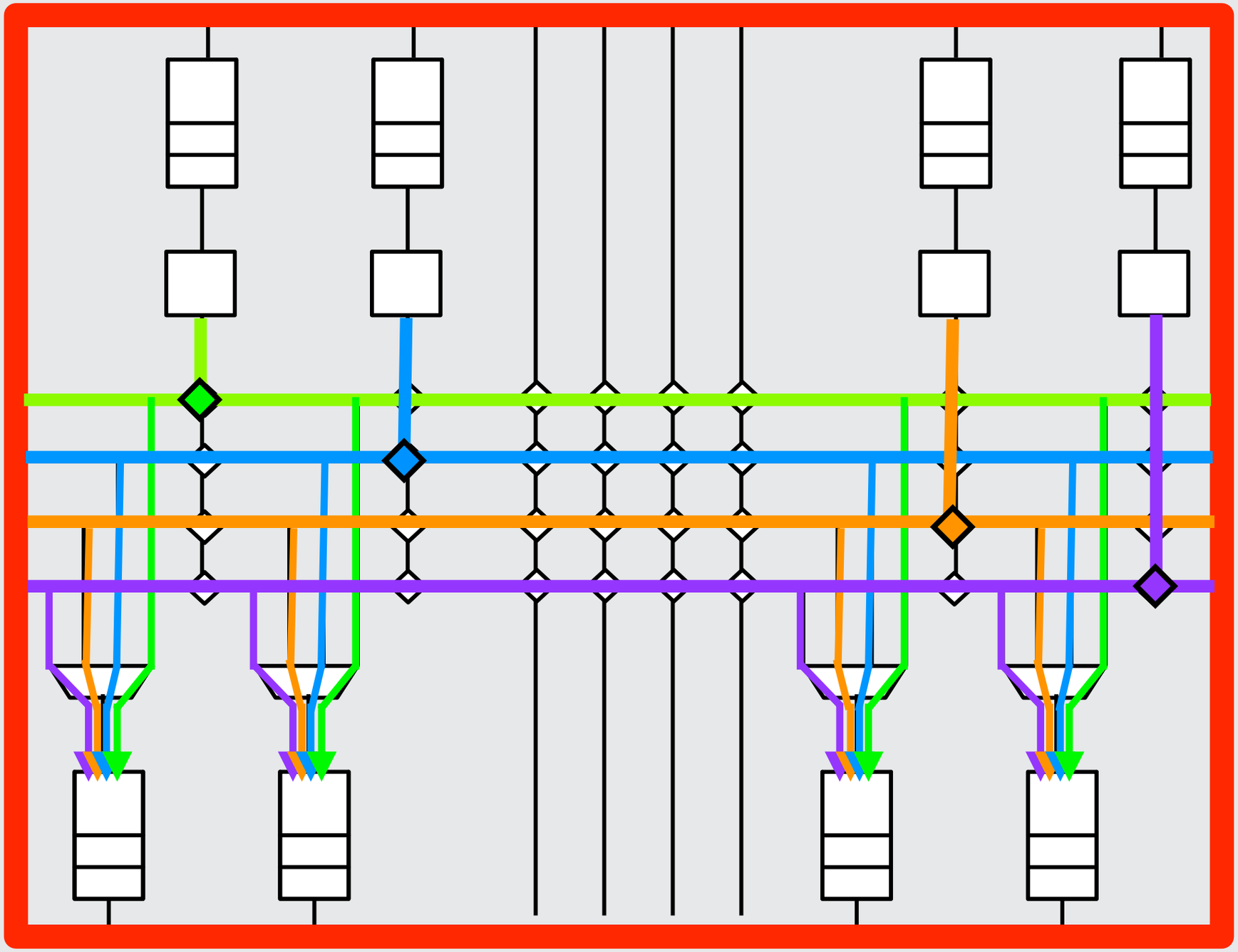
Internal Configuration of a Switch



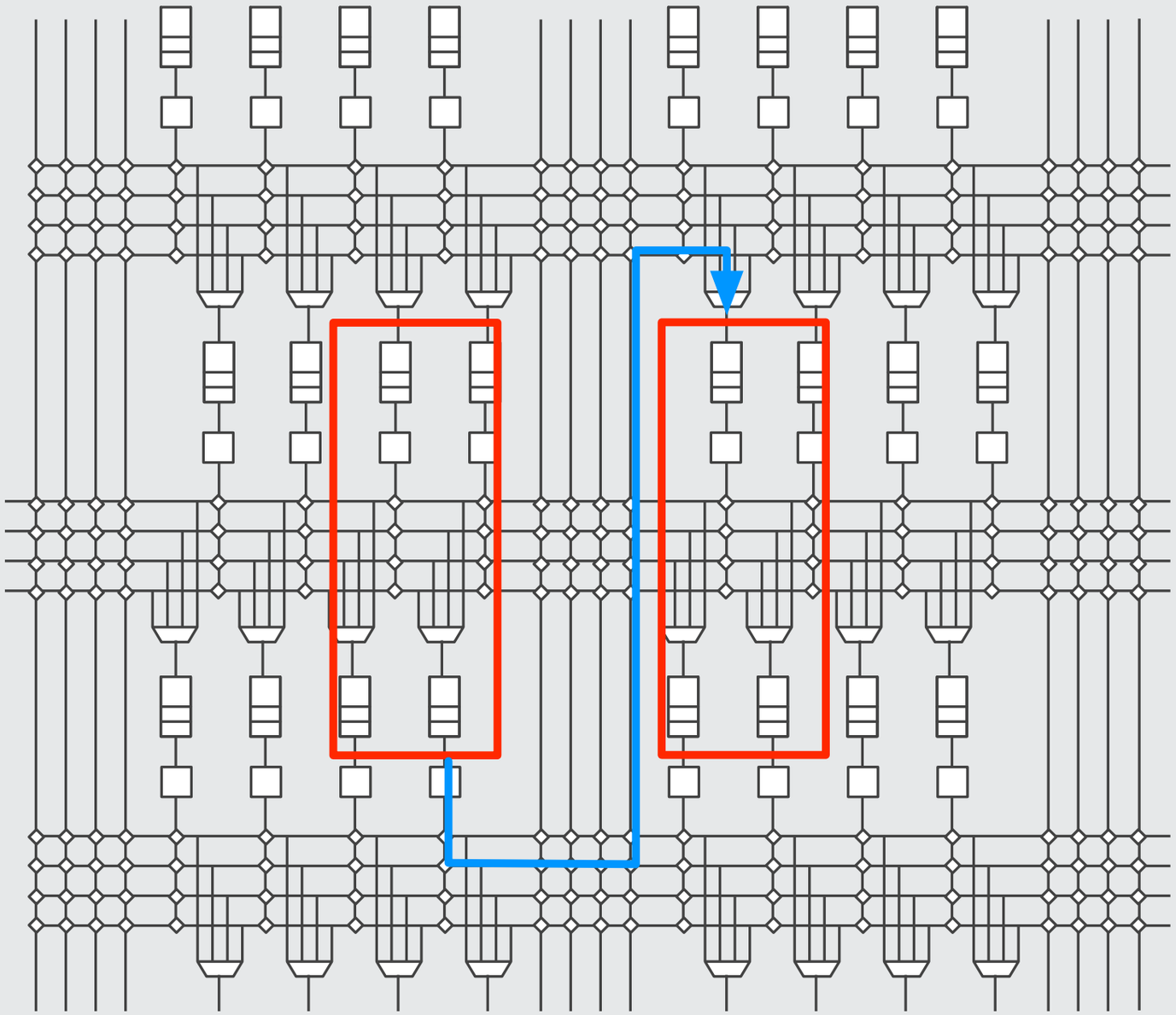
Internal Configuration of a Switch



Internal Configuration of a Switch



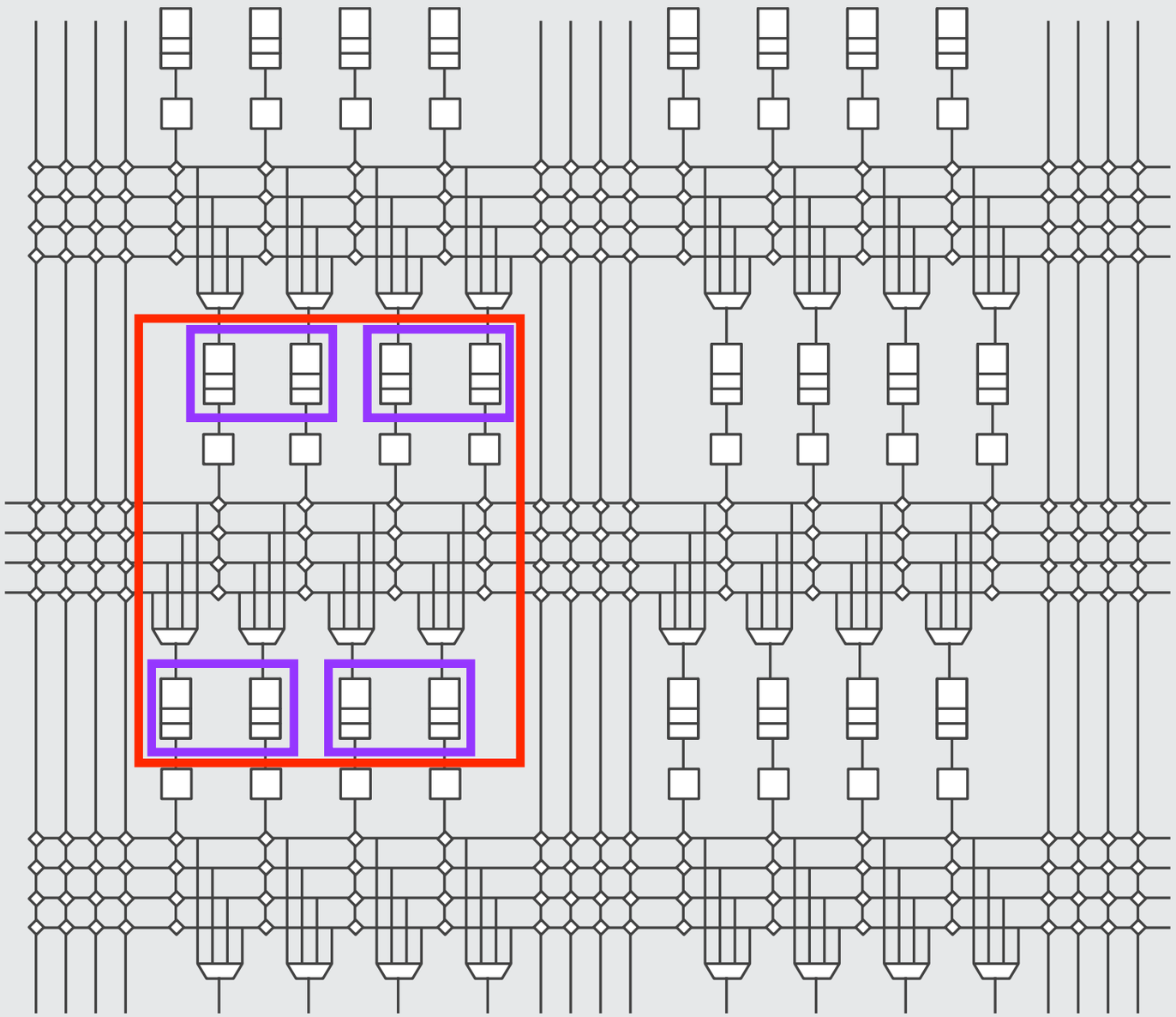
Network Configuration: Links



1. Switch degree

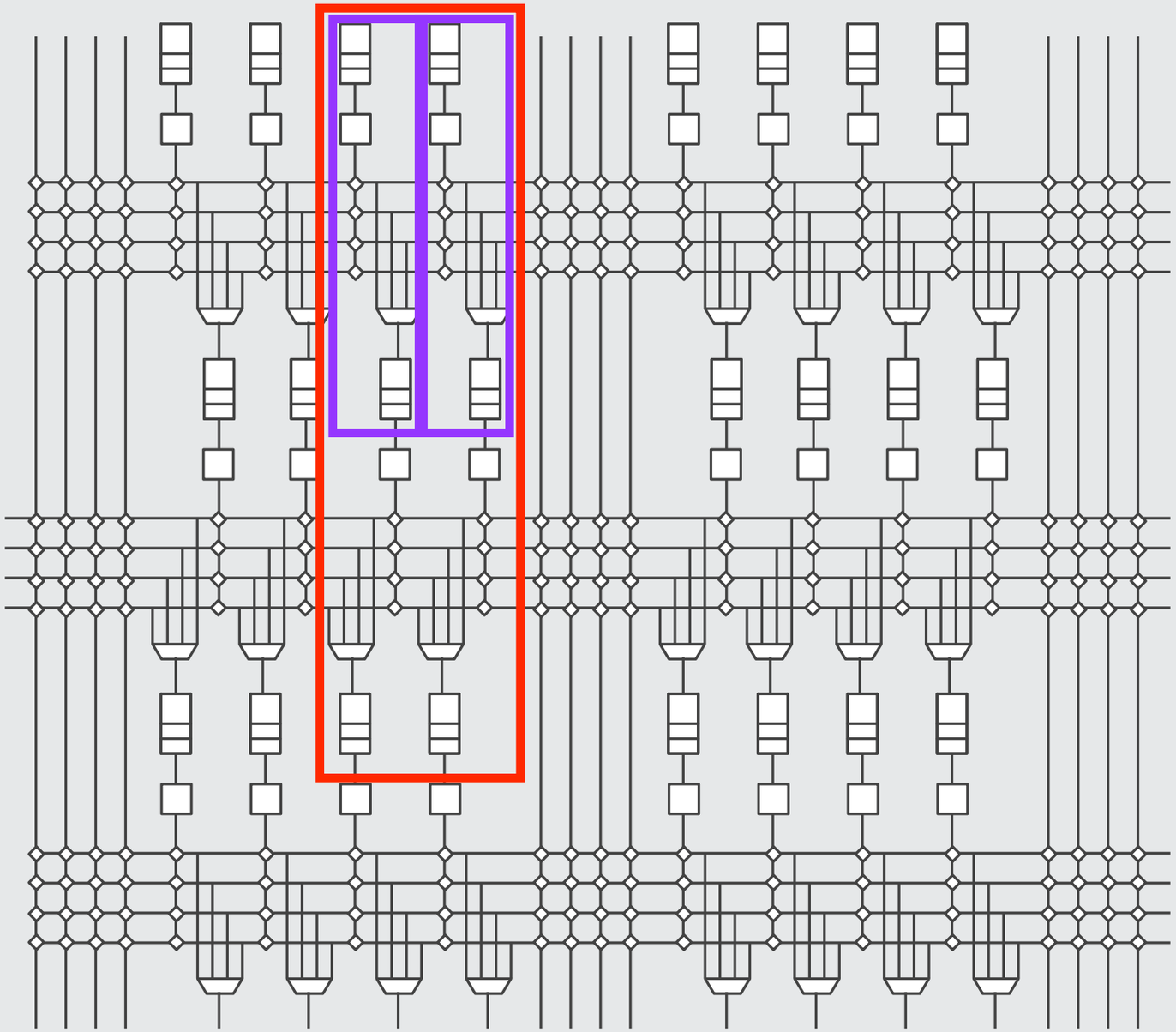
2. Inter-switch connections

Network Configuration: Packet Width



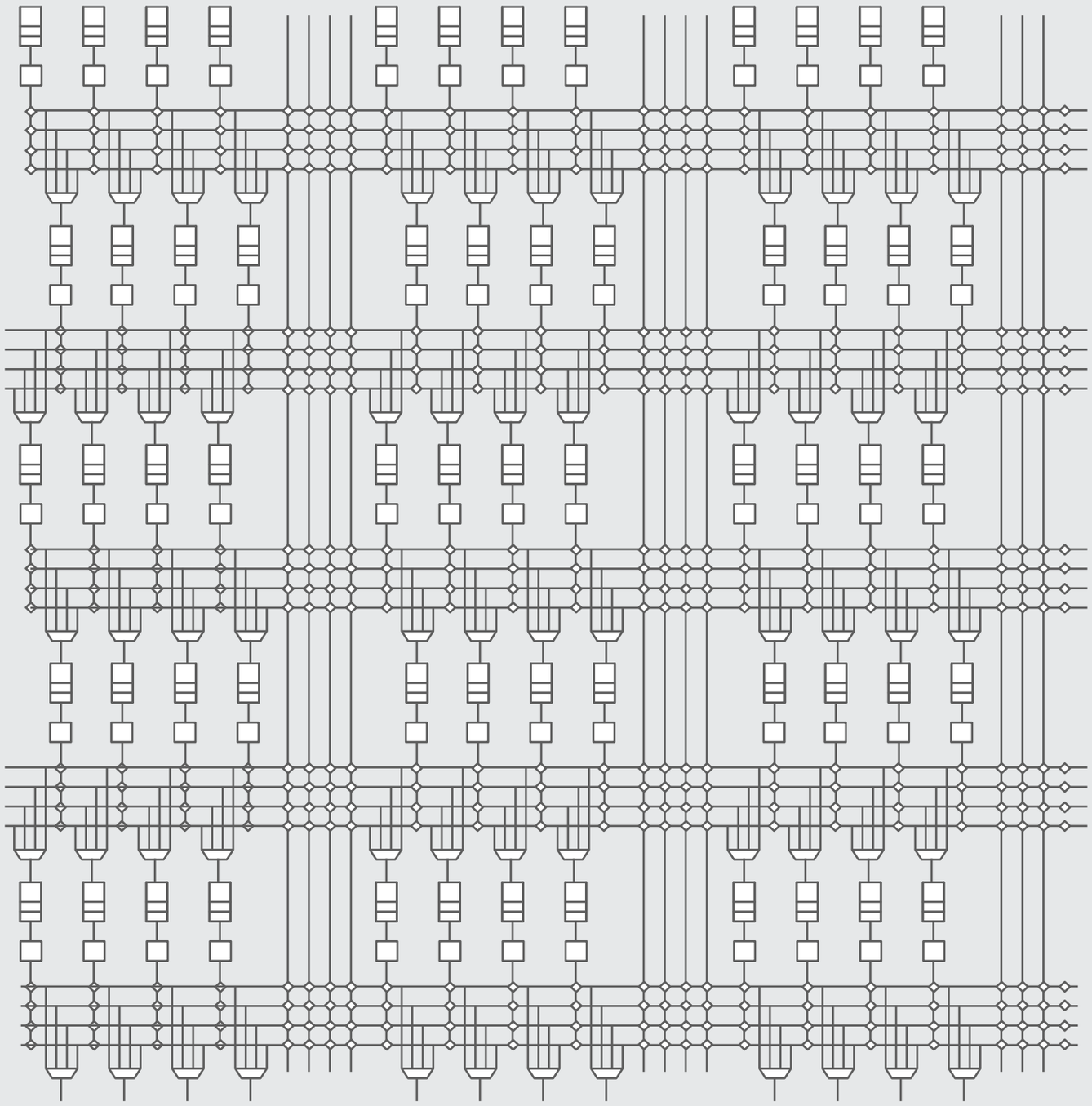
1. Switch degree
2. Inter-switch connections
3. Packet width

Network Configuration: Queue Capacity

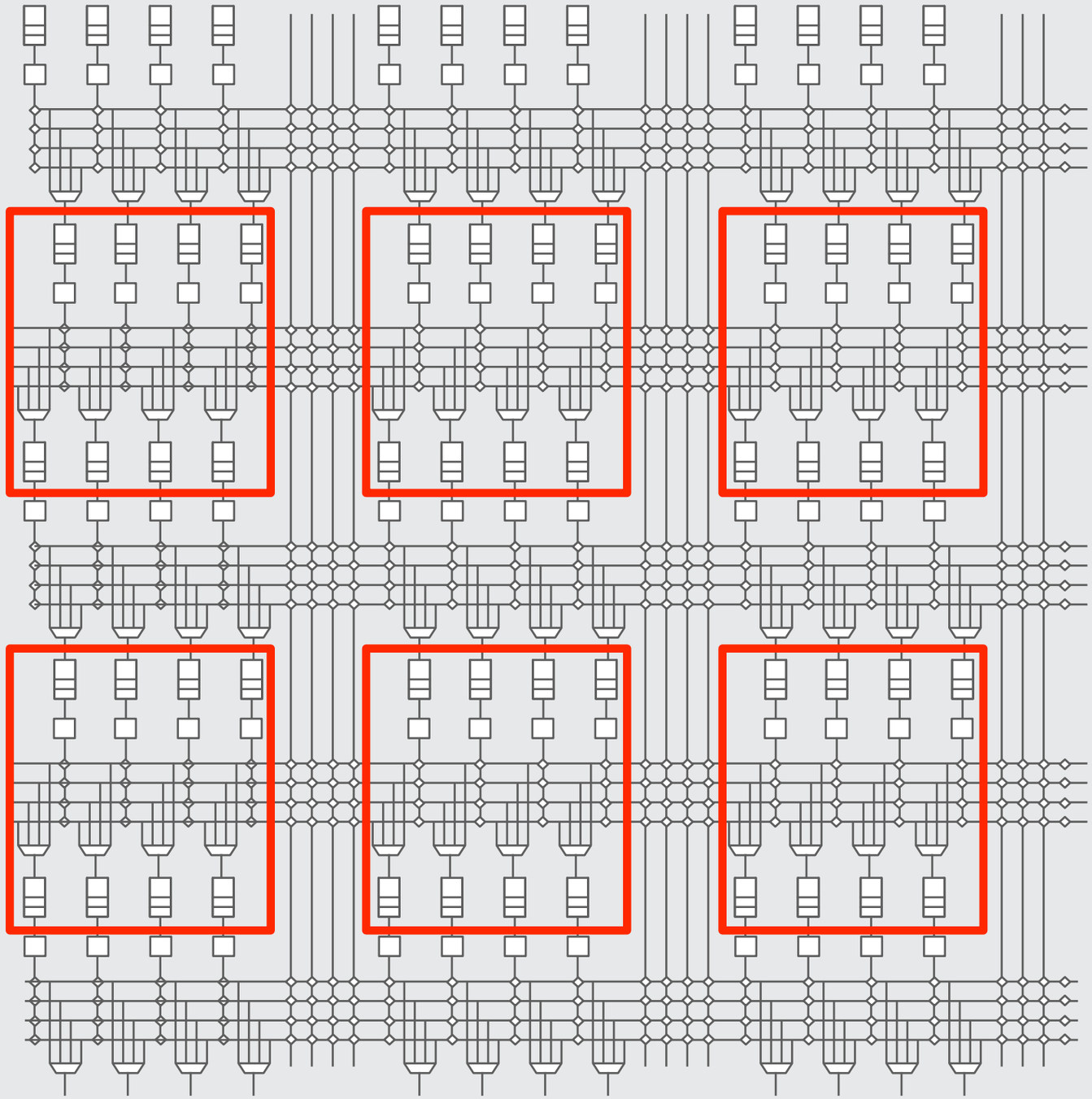


1. Switch degree
2. Inter-switch connections
3. Packet width
4. Buffer capacity

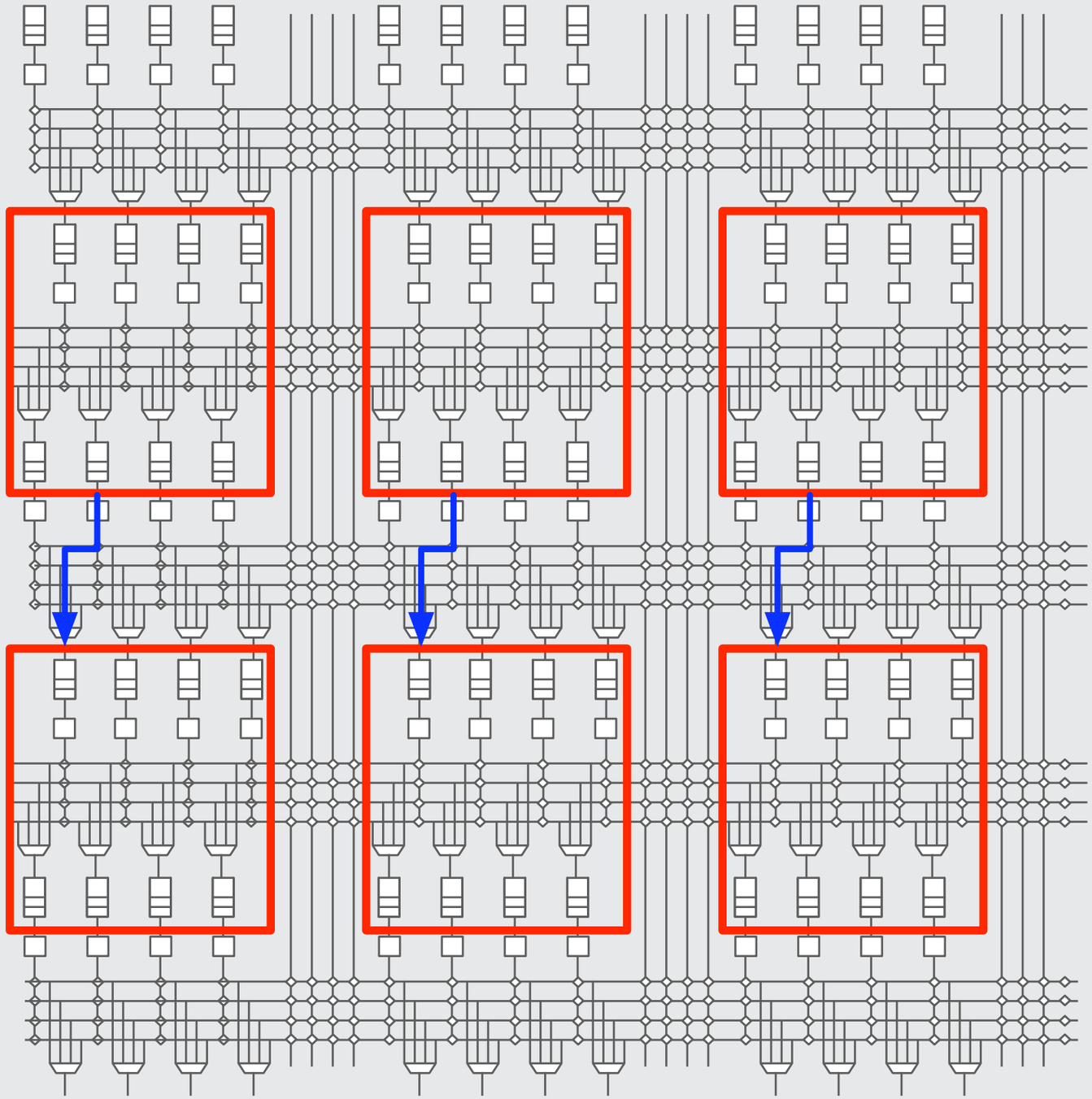
An Example: Configuration of a Mesh



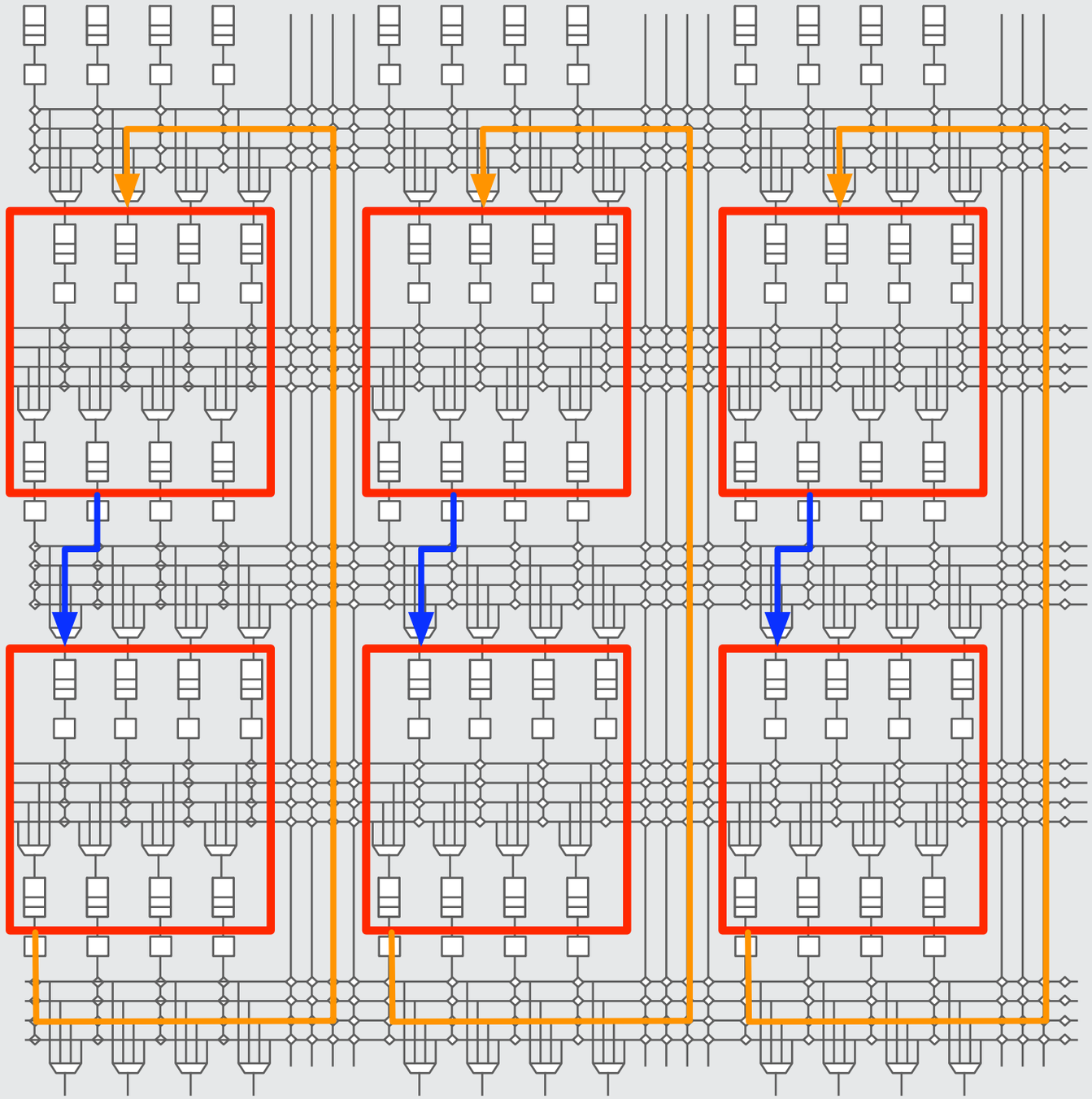
An Example: Configuration of a Mesh



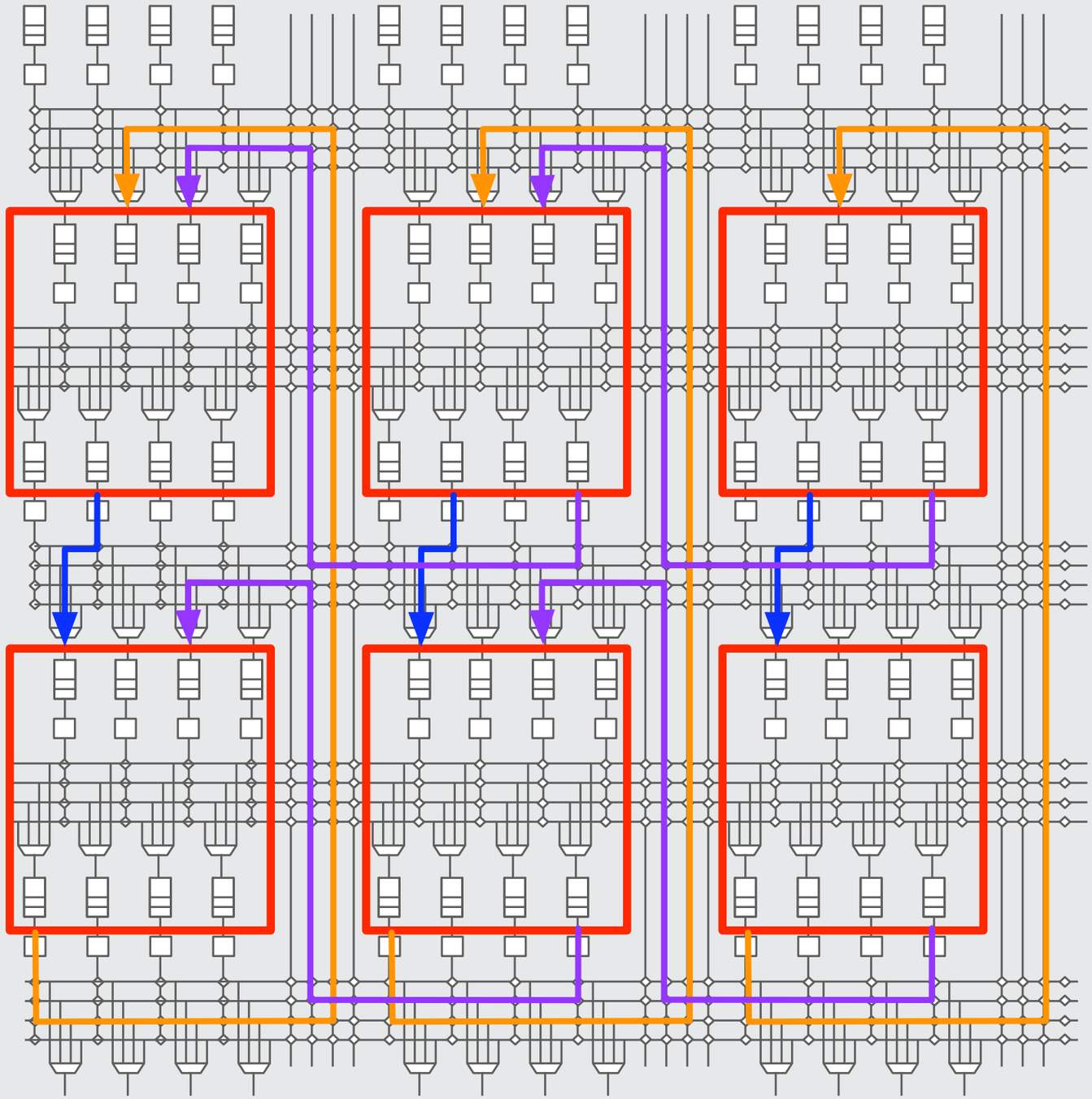
An Example: Configuration of a Mesh



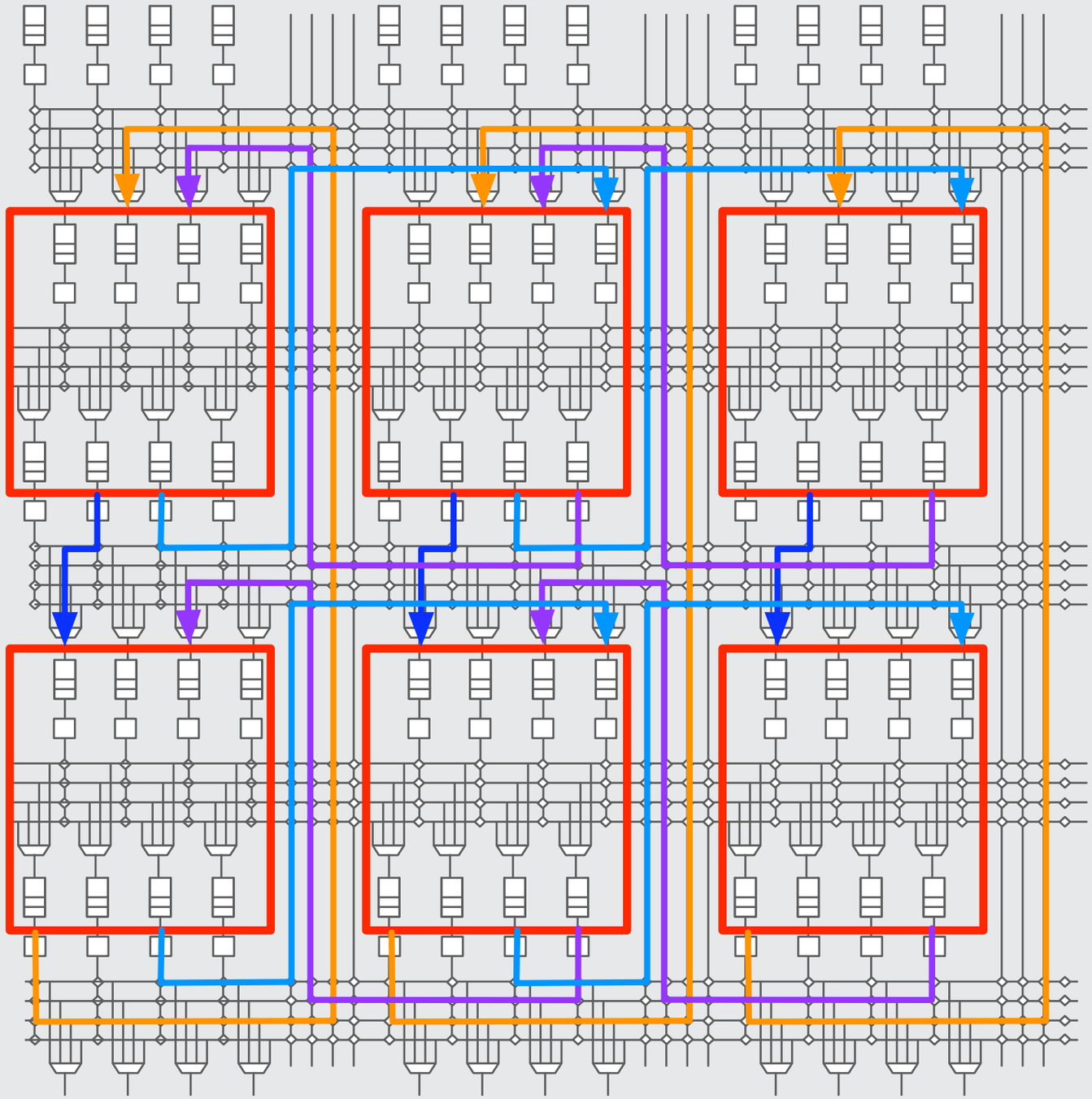
An Example: Configuration of a Mesh



An Example: Configuration of a Mesh



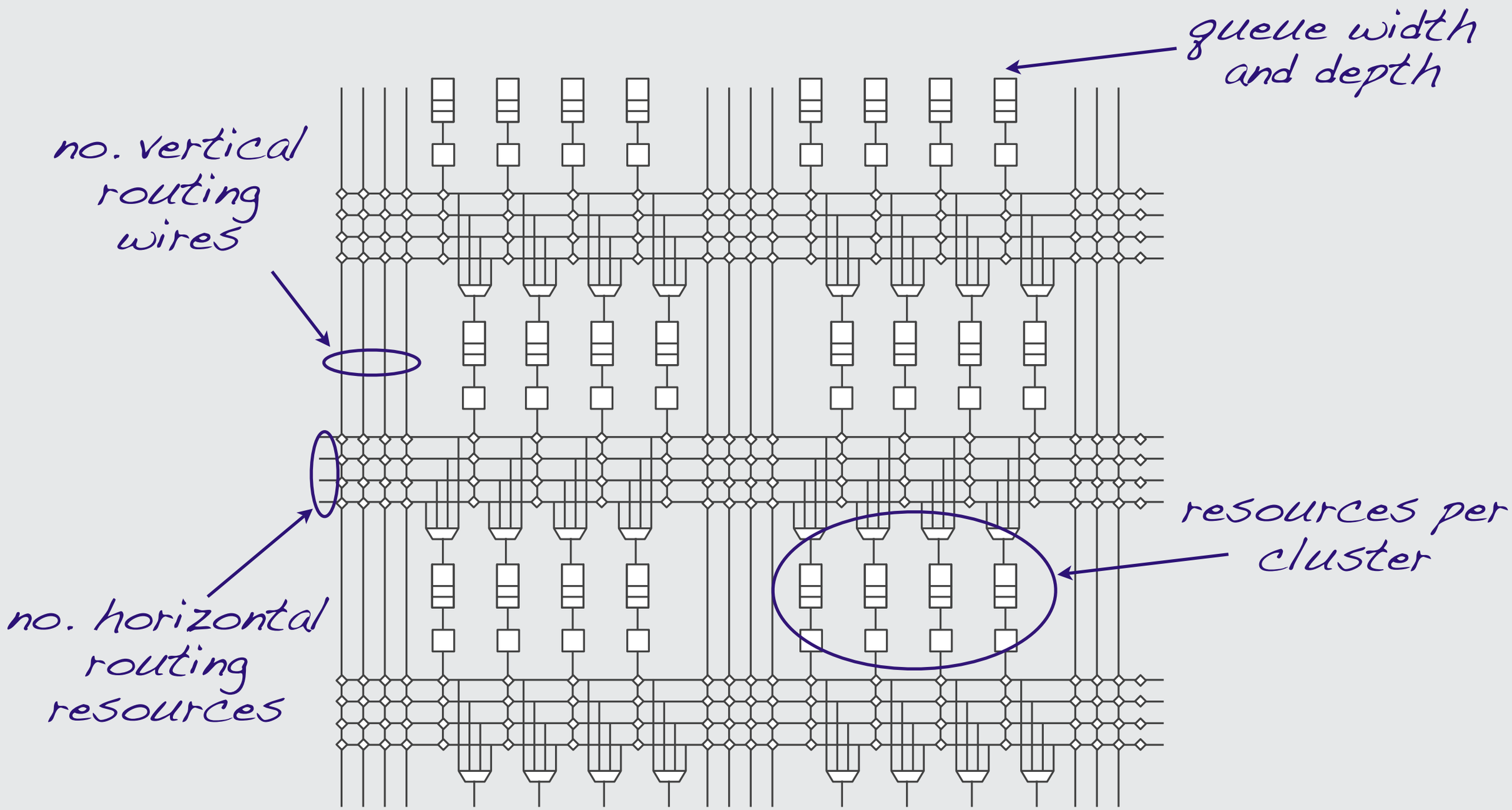
An Example: Configuration of a Mesh



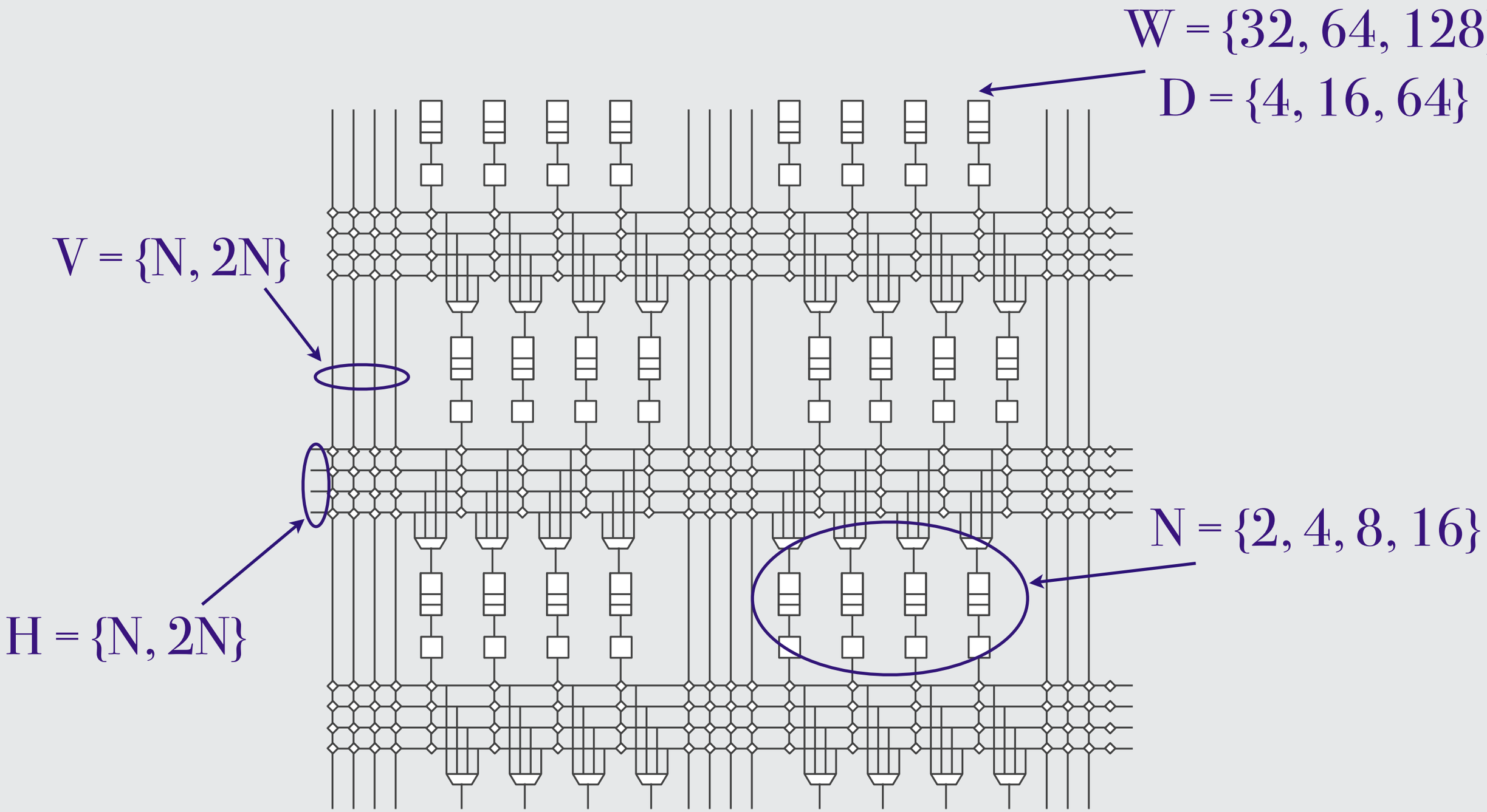
Talk Outline

- Network-on-chip Design Space Exploration
  - *Networks*
  - *Workloads*
  - *Results*
- Polymorphic On-Chip Networks
  - *Fabric design*
  - *Configuring the network*
  - *Selecting a fabric*
  - *Evaluation of flexibility*
- Conclusions
- Future Directions

Polymorphic Fabric Parameter Space

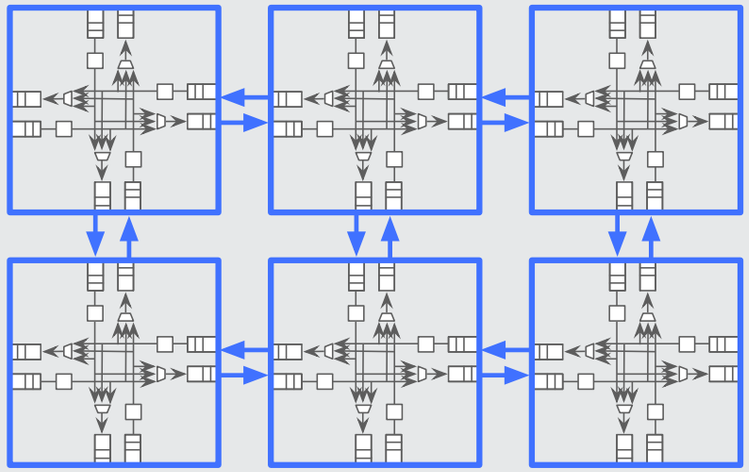


Polymorphic Fabric Parameter Space

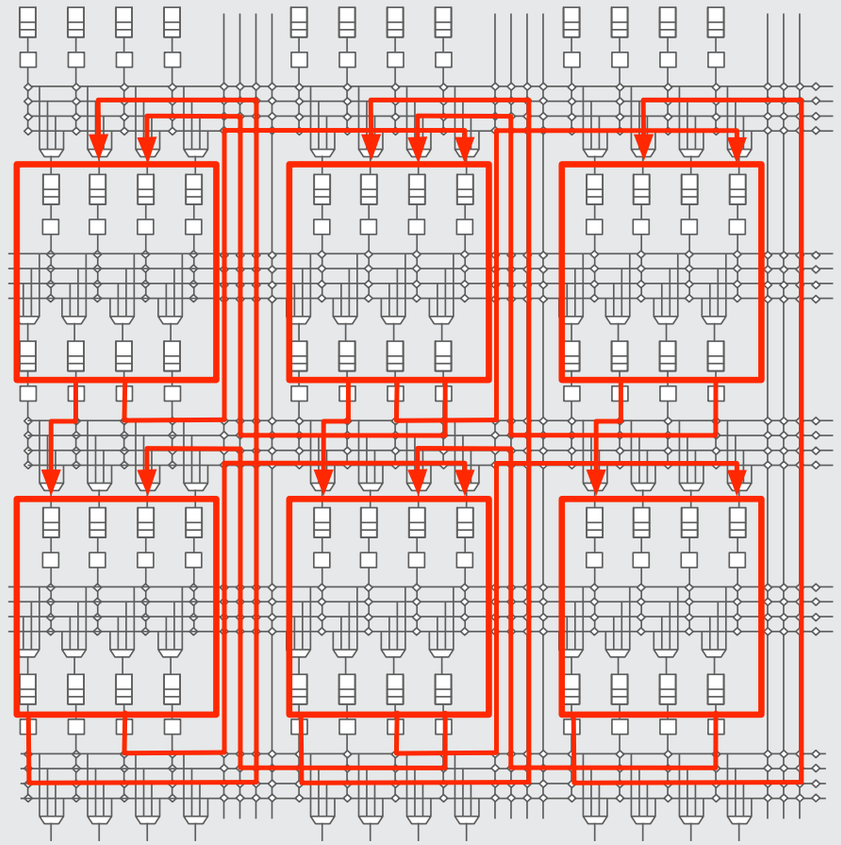


Polymorphic Fabric Area Overhead

ASIC implementation

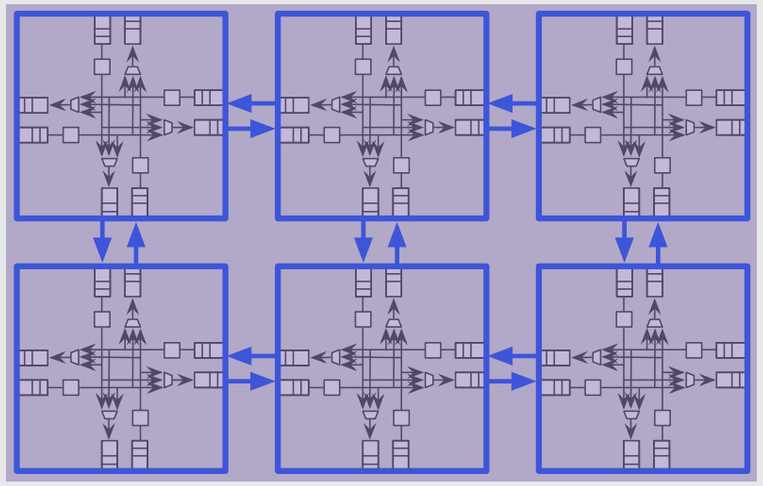


Polymorphic implementation

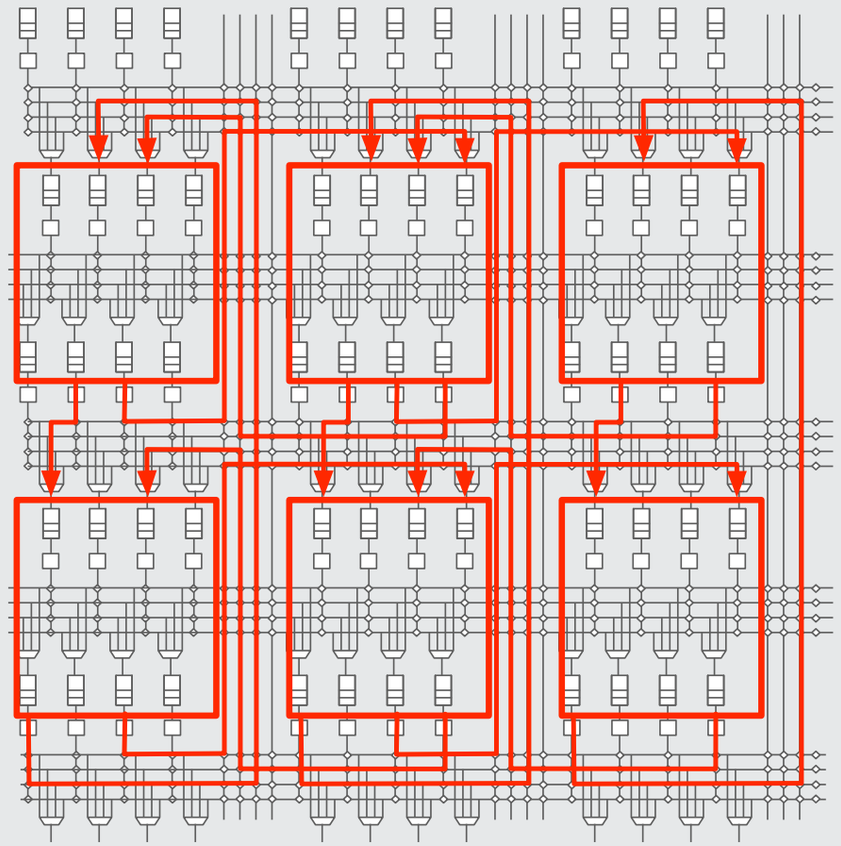


Polymorphic Fabric Area Overhead

ASIC implementation



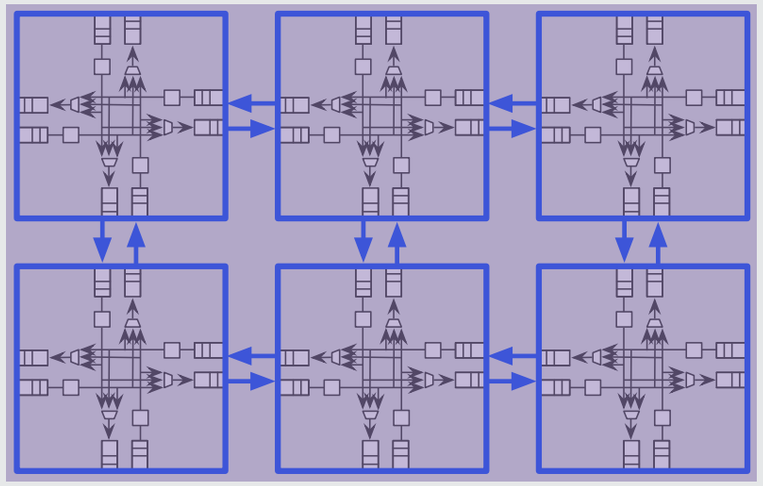
Polymorphic implementation



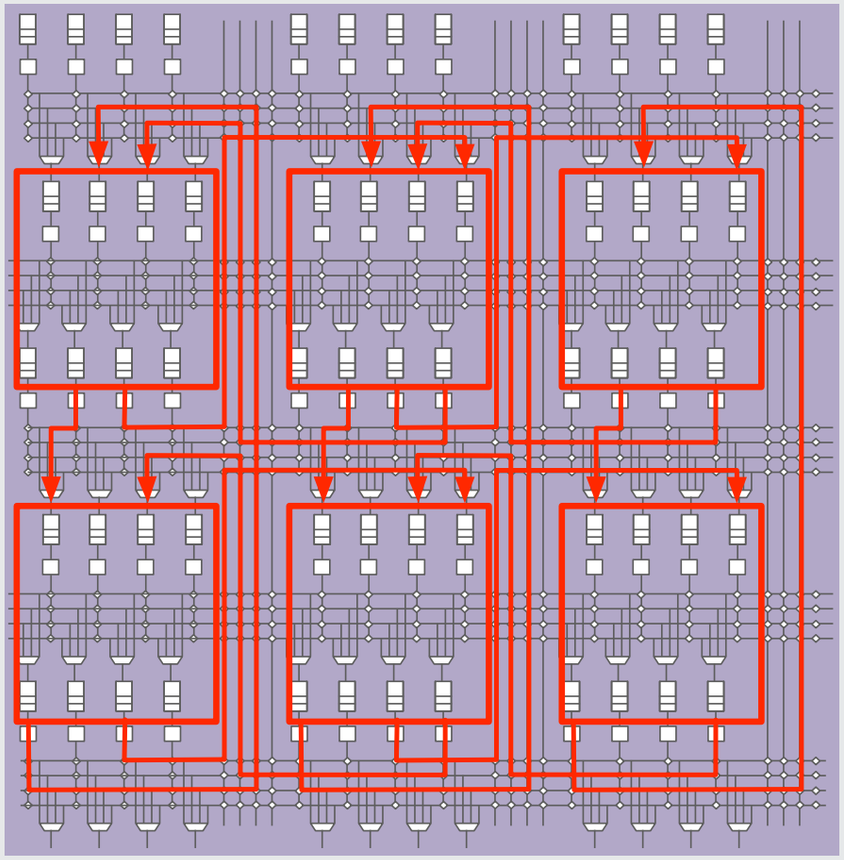
Area as ASIC

Polymorphic Fabric Area Overhead

ASIC implementation



Polymorphic implementation

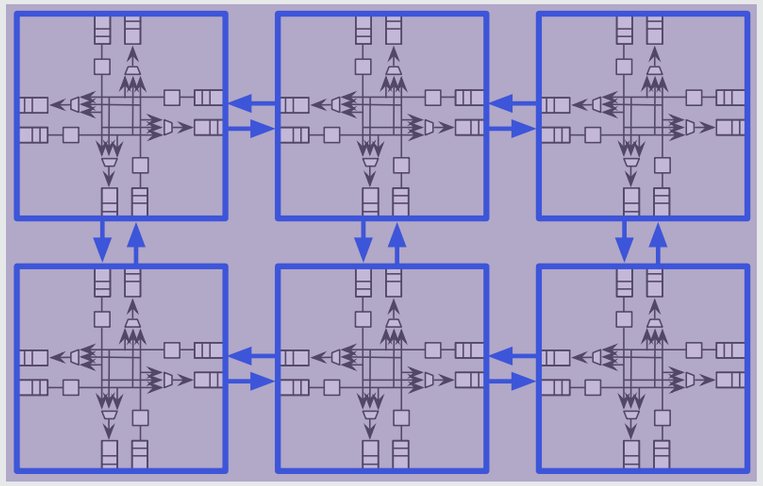


Area in Polymorphic Fabric

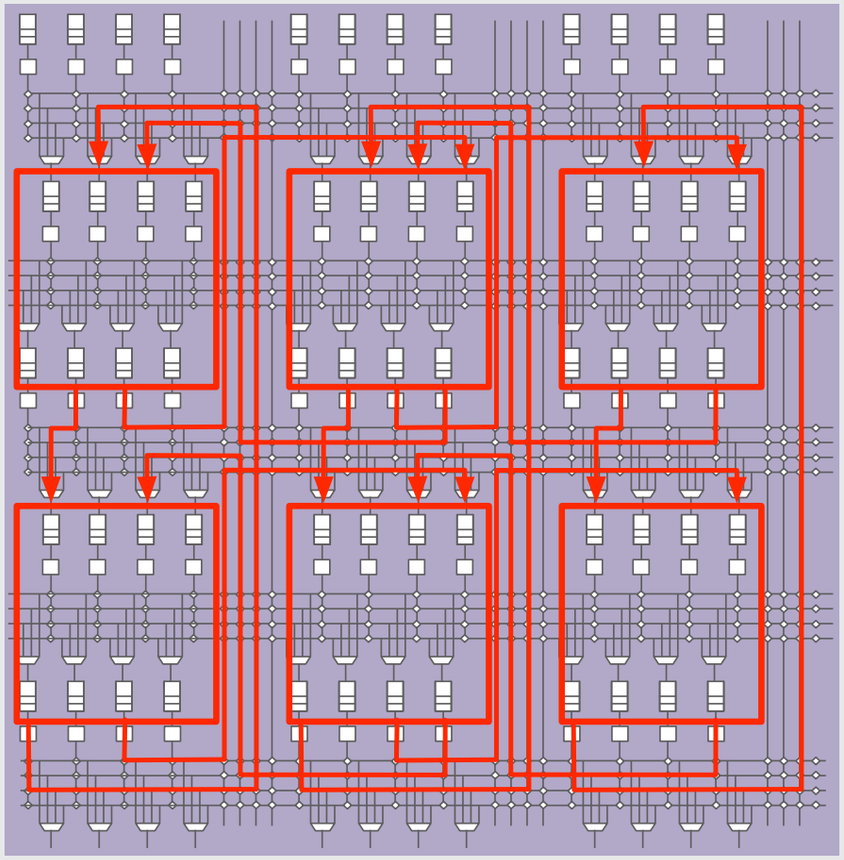
Area as ASIC

Polymorphic Fabric Area Overhead

ASIC implementation



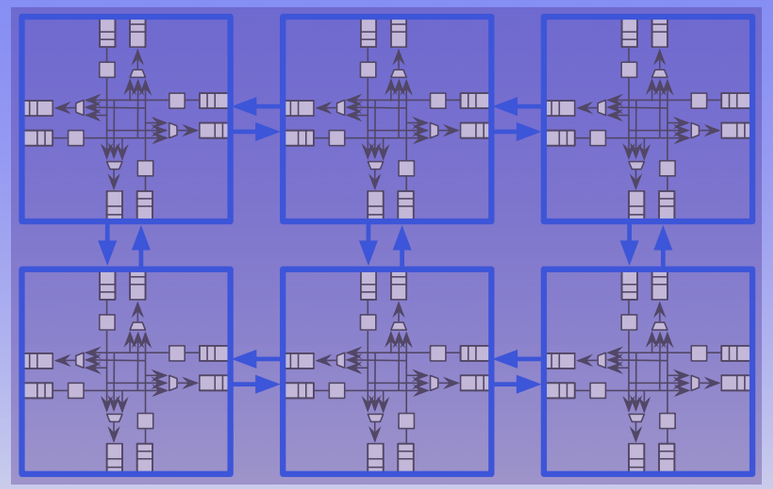
Polymorphic implementation



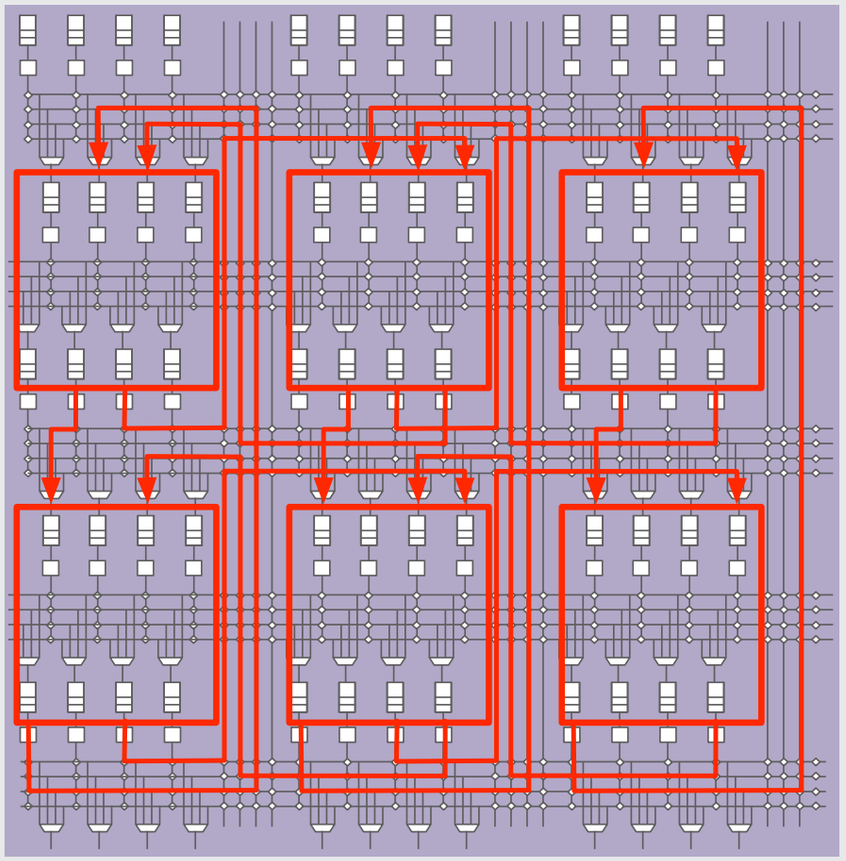
$$\text{Area Overhead} = \frac{\text{Area in Polymorphic Fabric}}{\text{Area as ASIC}}$$

Polymorphic Fabric Area Overhead

ASIC implementation



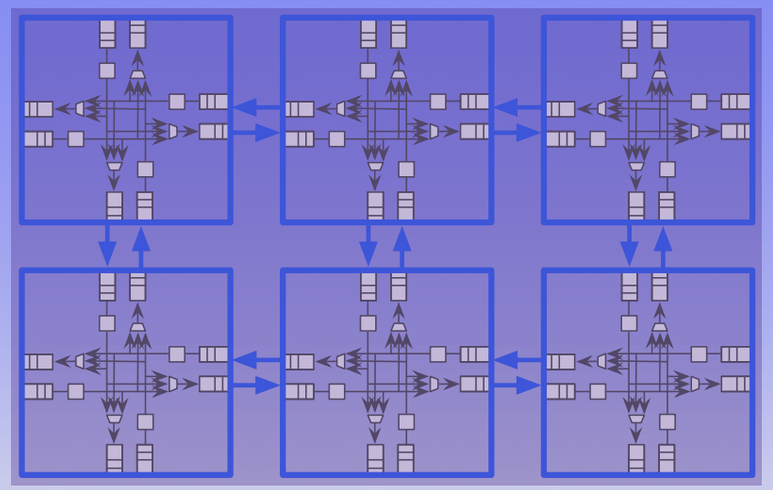
Polymorphic implementation



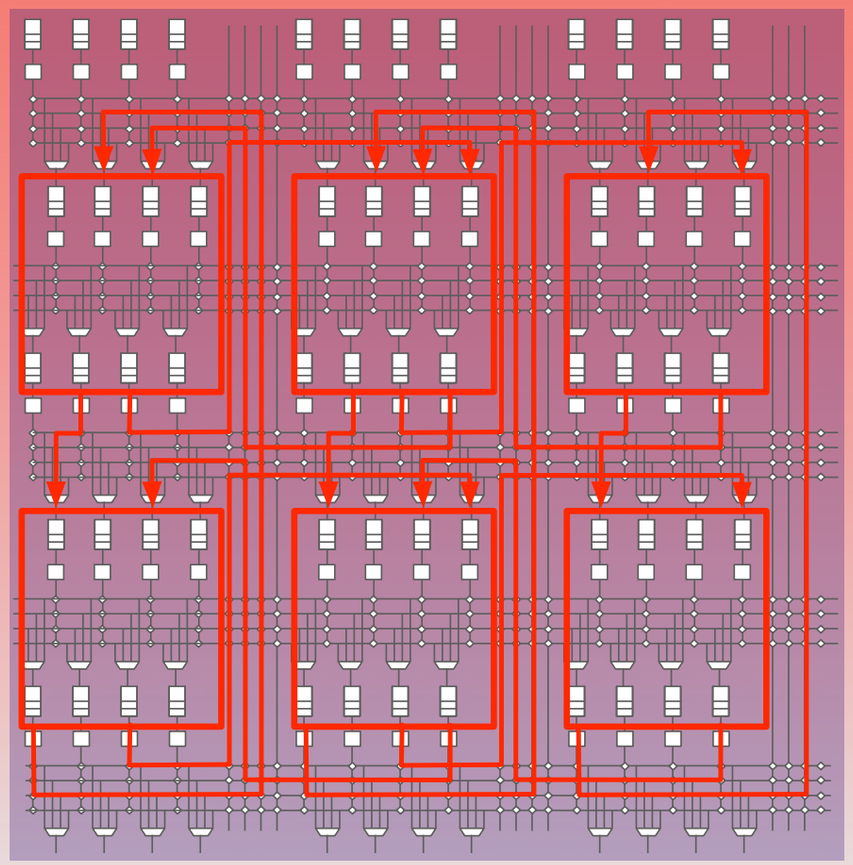
$$\text{Area Overhead} = \frac{\text{Area in Polymorphic Fabric}}{\text{Area as ASIC}}$$

Polymorphic Fabric Area Overhead

ASIC implementation

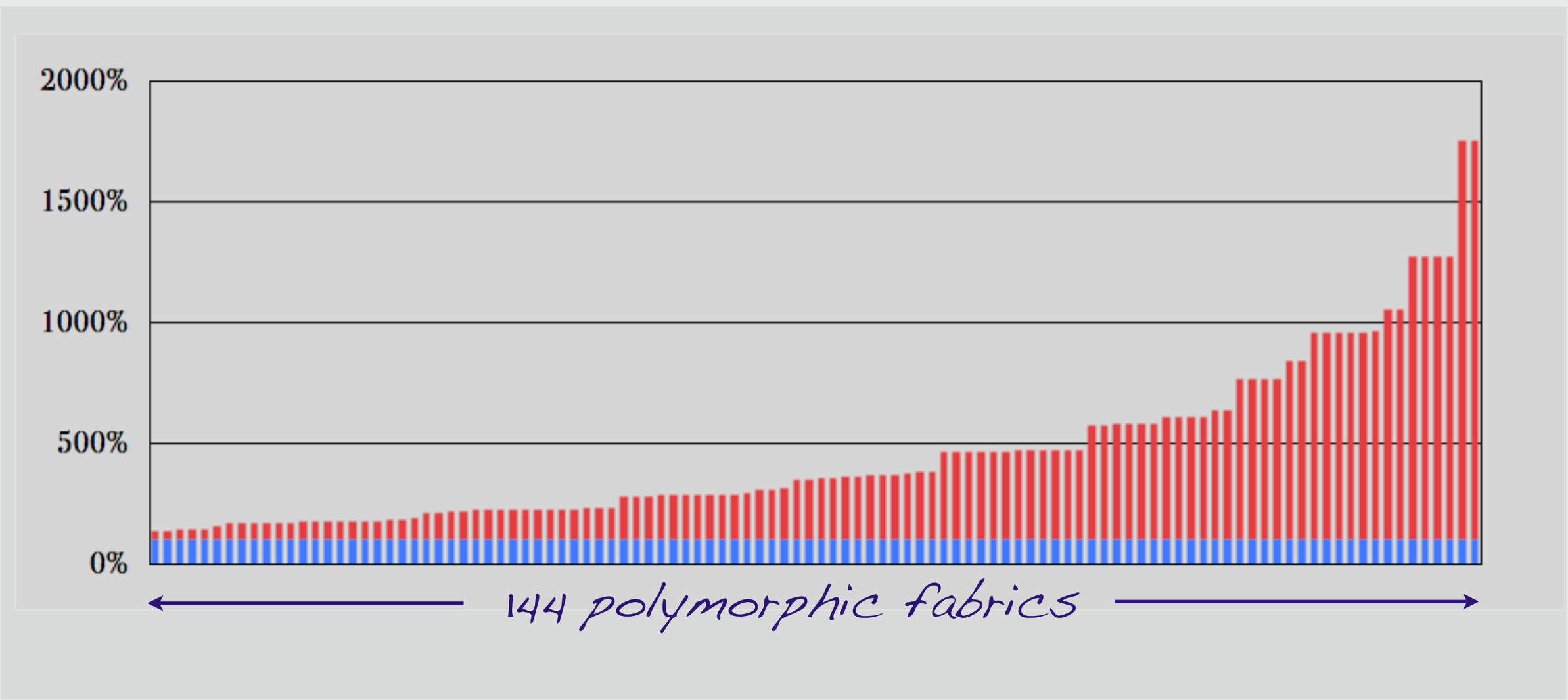


Polymorphic implementation



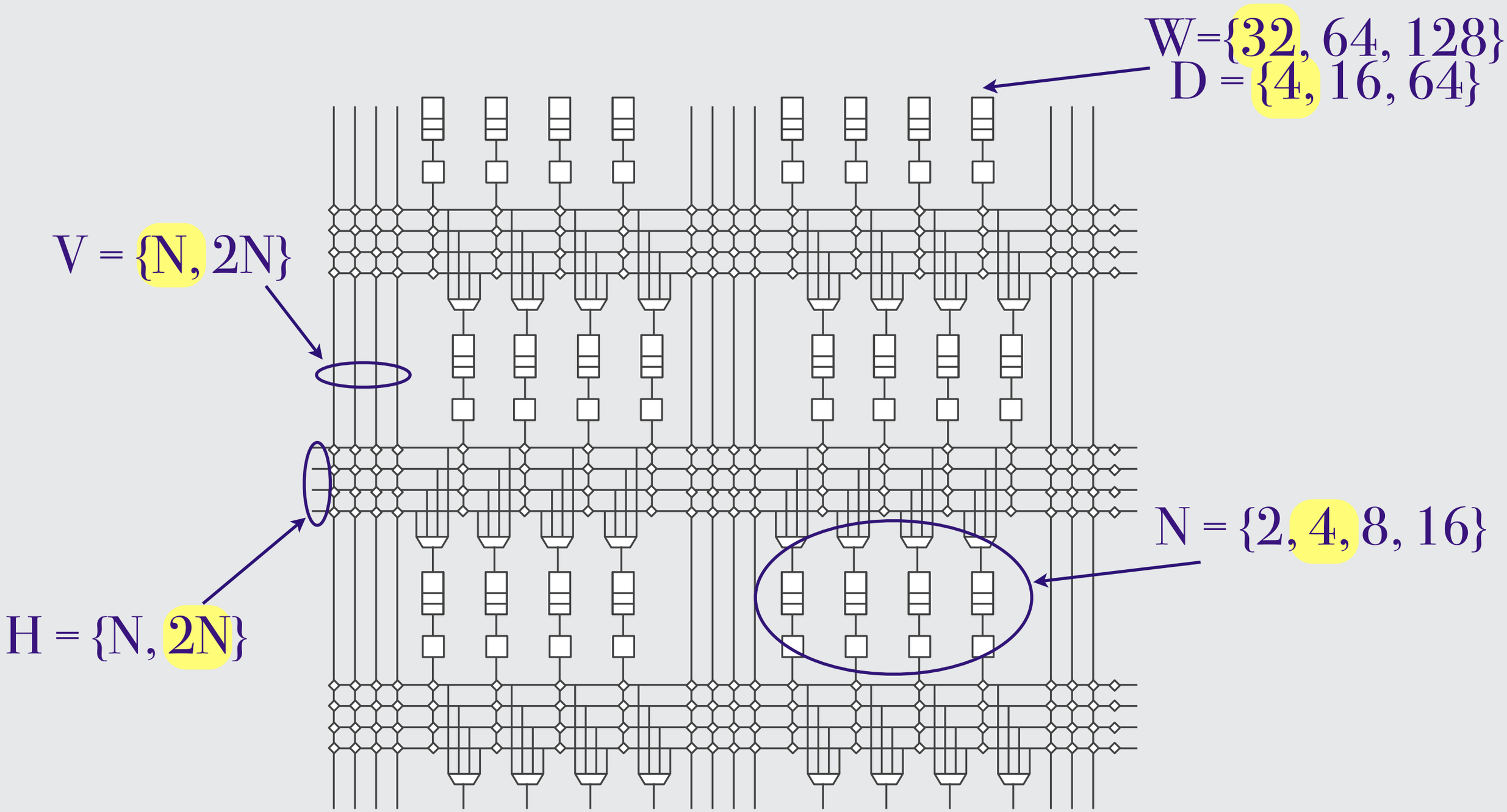
$$\text{Area Overhead} = \frac{\text{Area in Polymorphic Fabric}}{\text{Area as ASIC}}$$

Area Overhead of Polymorphic Fabrics



Area efficient networks have **small queues** and **generous routing resources**

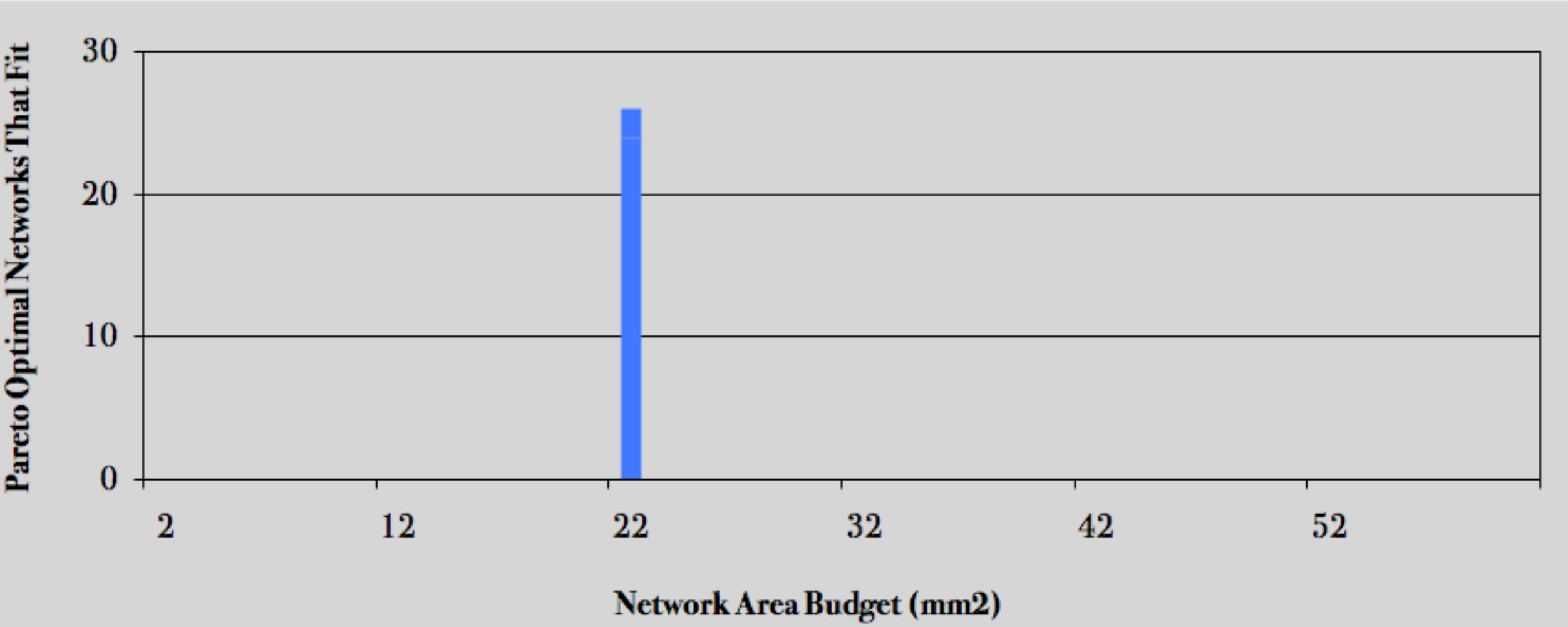
Polymorphic Fabric Parameter Space



Talk Outline

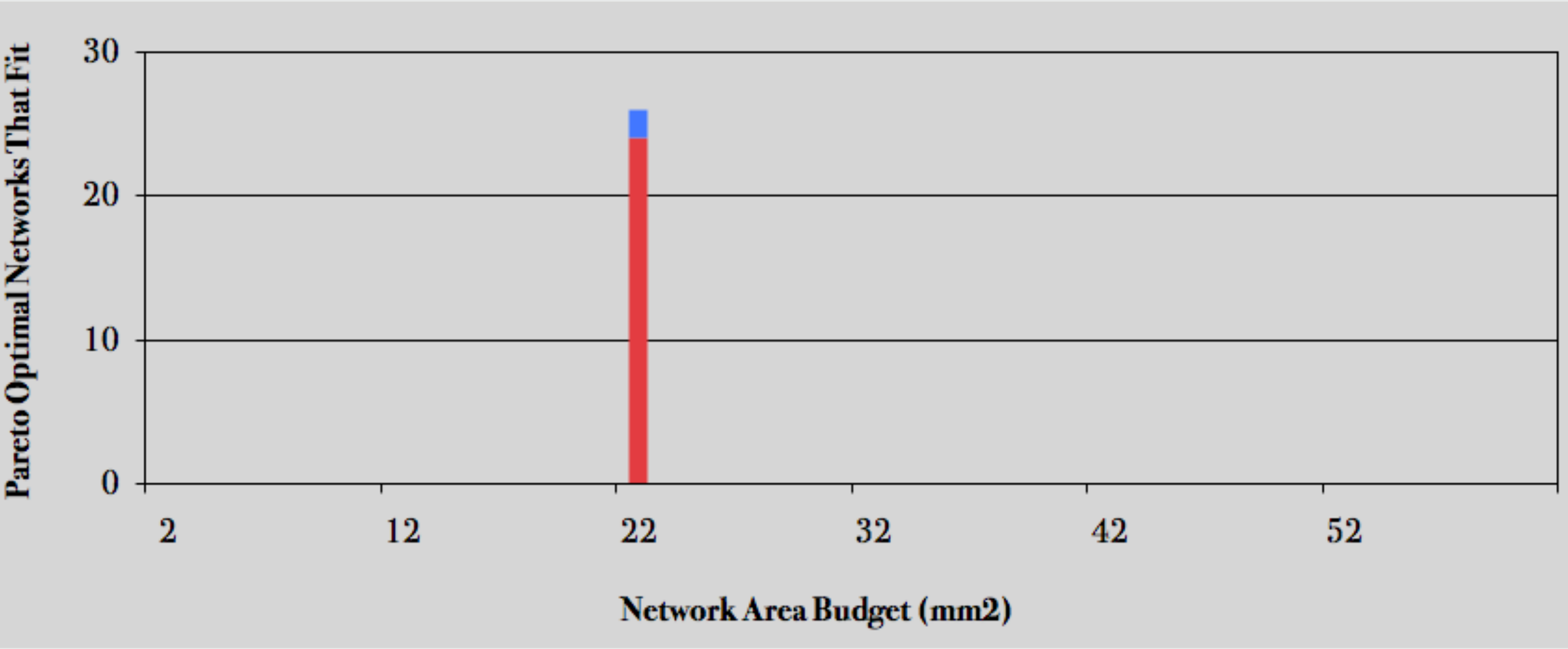
- Network-on-chip Design Space Exploration
  - *Networks*
  - *Workloads*
  - *Results*
- Polymorphic On-Chip Networks
  - *Fabric design*
  - *Configuring the network*
  - *Selecting a fabric*
  - *Evaluation of flexibility*
- Conclusions
- Future Directions

Network Selection Under Area Budget



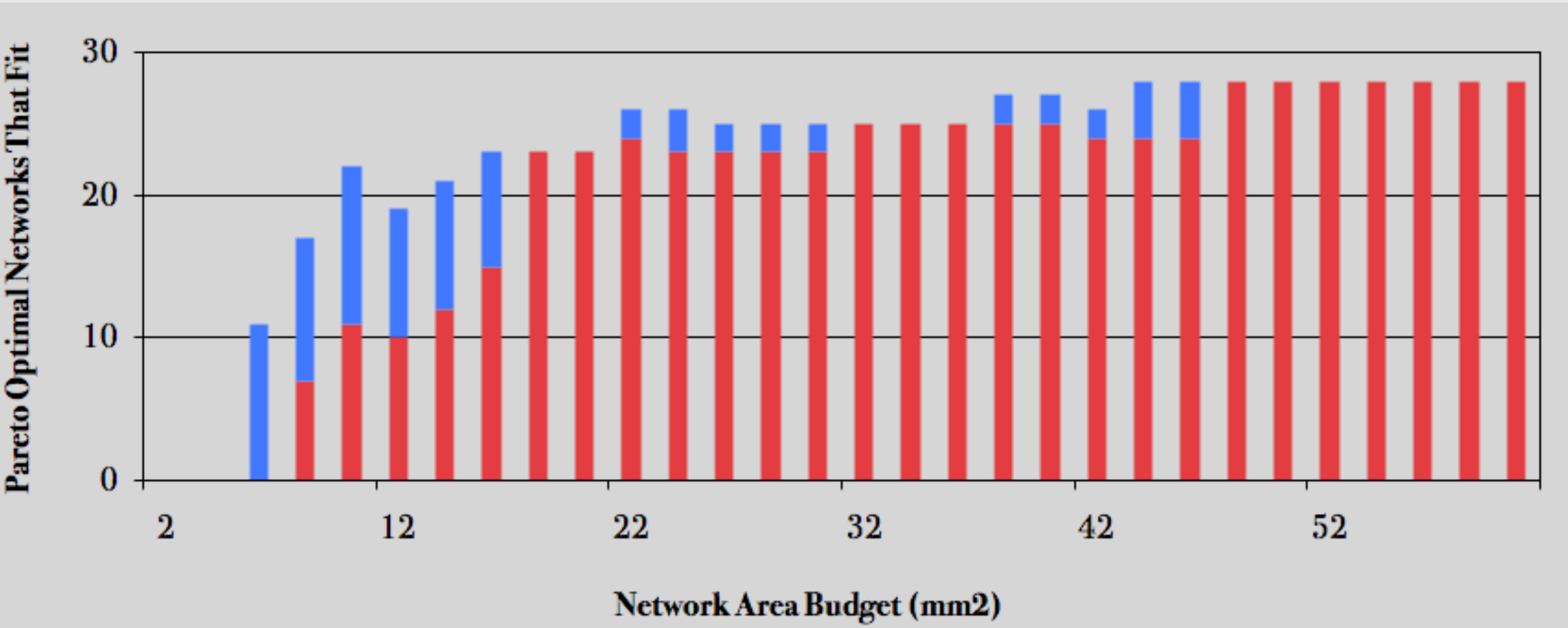
Of networks smaller than 22 mm^2, 26 are pareto optimal.

Network Selection Under Area Budget



24 of the 26 optimal networks will fit in 22 mm^2 of polymorphic fabric.

Network Selection Under Area Budget



Polymorphic coverage is strong for all but the tightest area budgets.

Conclusion

Widely varying on-chip communication patterns can take advantage of a flexible on-chip network.

Polymorphic fabric is a coarse grained reconfigurable circuit designed to implement packet-switched networks on chip.

Subject to area budget, polymorphic fabric usually offers broad choice of network.

Should build polymorphic network unless

1. Area budget highly constrained
2. Application and/or traffic not expected to vary

Some Future Directions

1. Hardware implementation
2. Uses beyond application performance

   (e.g., on-chip isolation)
3. Incorporation of advanced on-chip network innovations
4. Reconfiguration policy

THANK YOU