UNIVERSITAT POLITÈCNICA DE VALÈNCIA

GAP Parallel Architectures Group

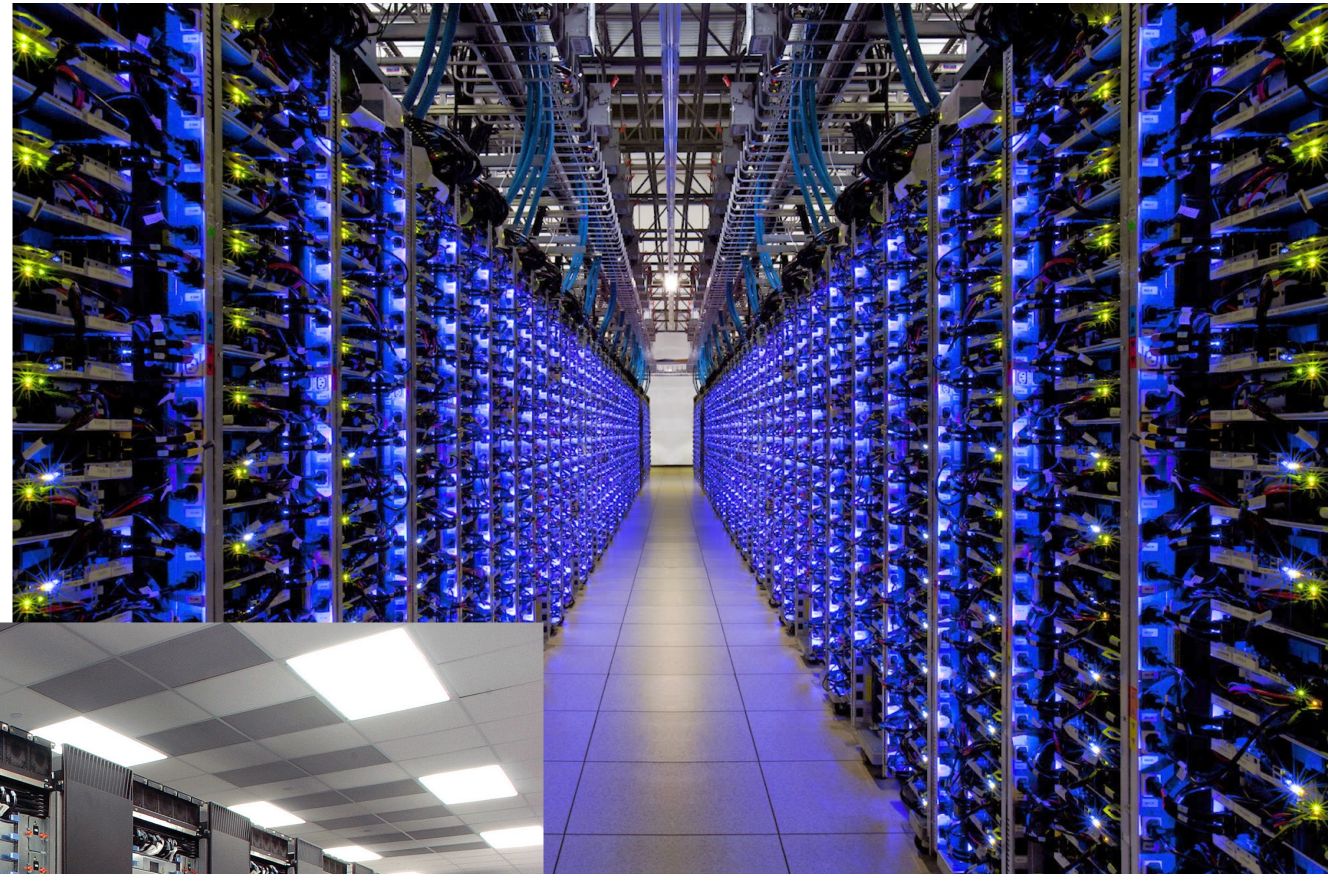
Topology and routing issues in HPC systems and datacenters

María Engracia Gómez

Universitat Politècnica de València

March 2016

Data centers

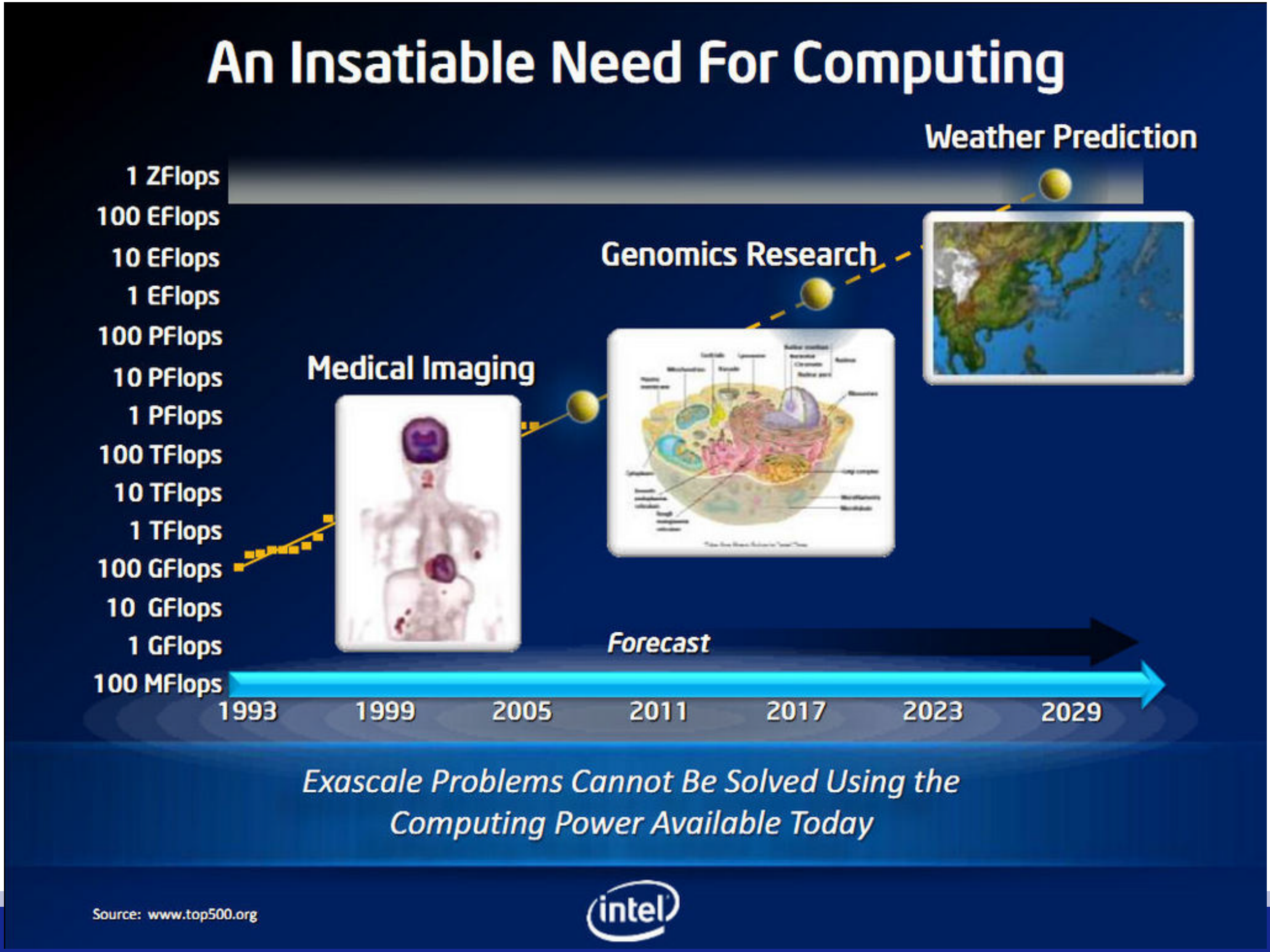


Supercomputers

TOP500 list as of Nov.&2015

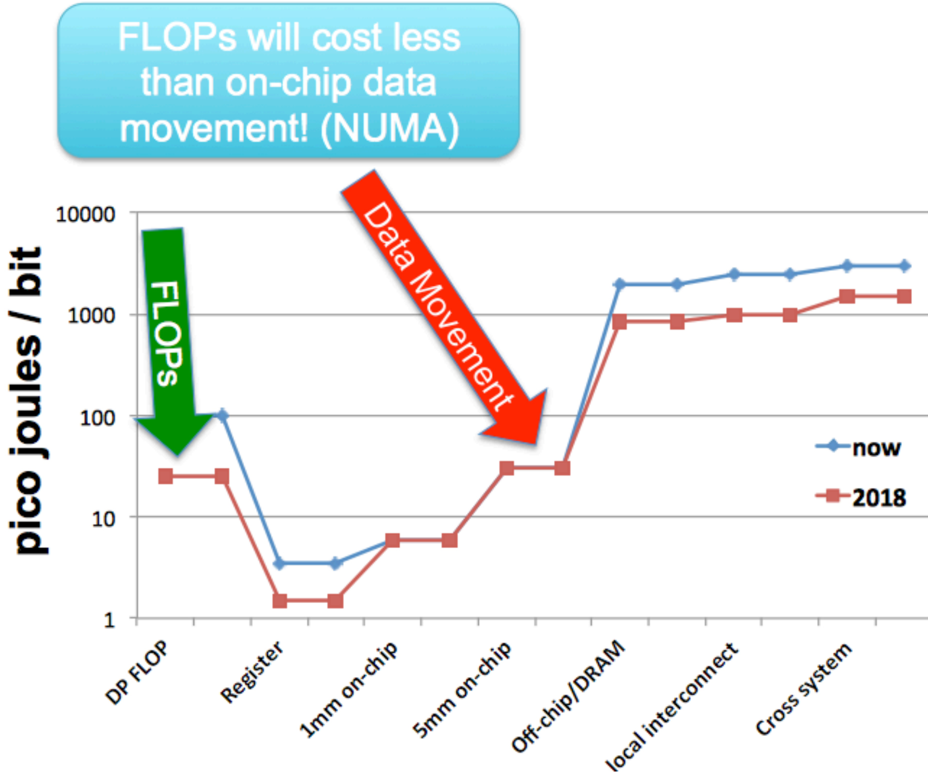
Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Powe[r] (kW)	
1	National University of Defense Technology China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3120000	33862.7	54902.4	17808	FatTree
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209	Torus
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	17173.2	20132.7	7890	Torus
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660	Torus
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8586.6	10066.3	3945	FatTree

And computing power needs increasing...



Data movement is a barrier towards exascale computing

- Data movement challenges impact the entire system
- Performance is increasingly determined by how data is communicated among the numerous compute resources
- Energy consumption is increasingly dominated by the cost of data movement
  - The energy cost to move a datum will exceed the cost of a floating-point operation
- We need:
  - Low Latency, high-bandwidth, low energy consumption interconnects for data exchange among thousands of processors



Courtesy: Horst Simon, Lawrence Berkeley Nacional Lab

Main design parameters

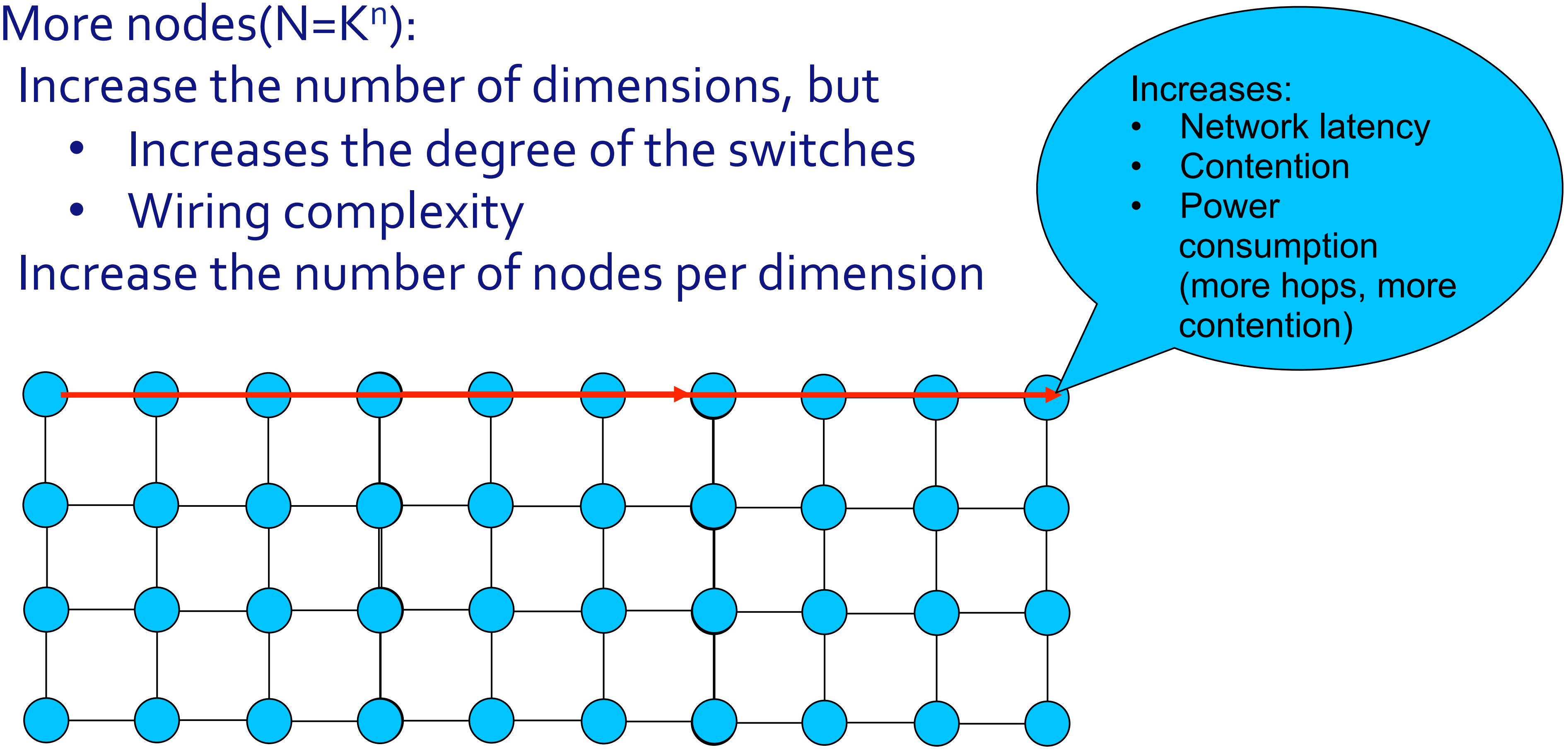
- Network Topology
    - Direct topologies
    - Indirect topologies
- Routing Algorithm

2. Regular Topologies

Direct Topologies

- These topologies are used in the most powerful super-computers (Current number 2, 3 and 4 of the Top500)

- More nodes($N=K^n$):
- Increase the number of dimensions, but
  - Increases the degree of the switches
  - Wiring complexity
- Increase the number of nodes per dimension



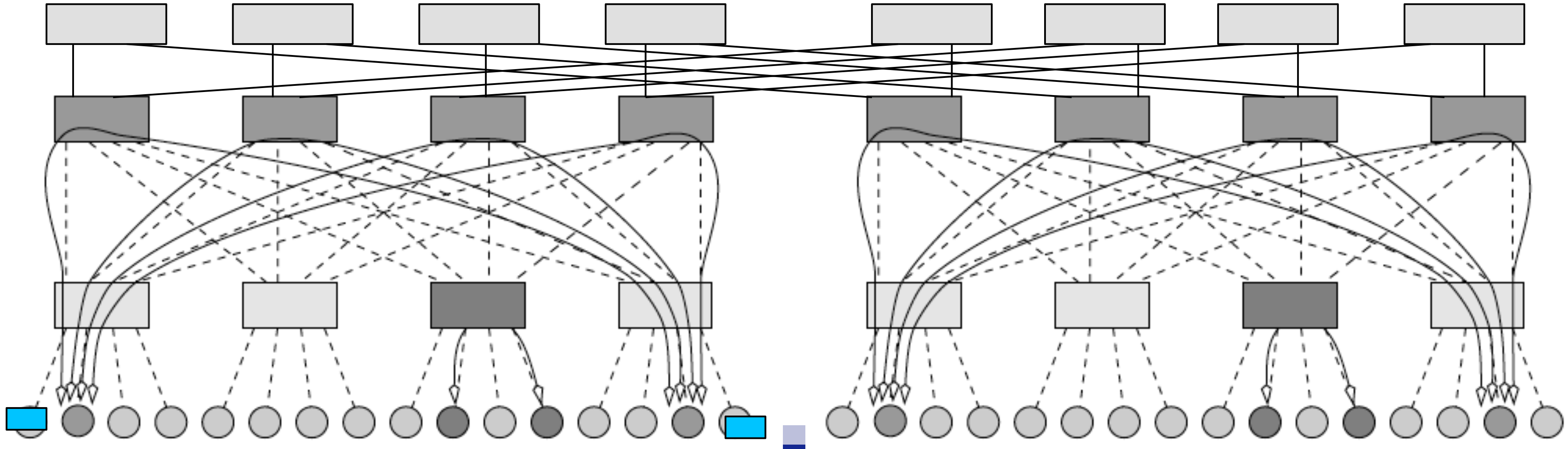
2. Regular Topologies

Indirect Topologies

Fat-tree used in some of the most powerful machines (number 1 and 5 in the top500)
The diameter only depends on the number of stages, 2 x number of stages

More nodes with the same switch degree ($N=k^n$)

- More stages: diameter is increased
- More switches per node (higher cost)
- Wiring complexity

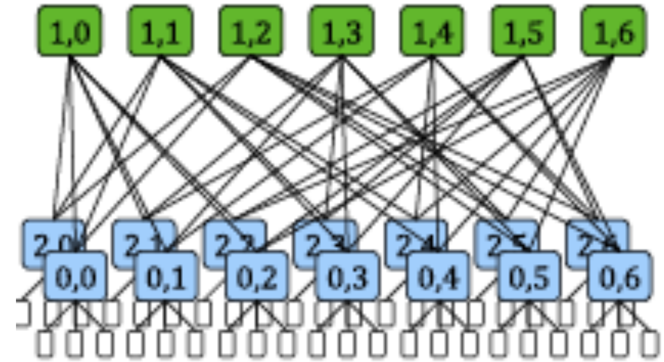


New topologies for massively parallel systems

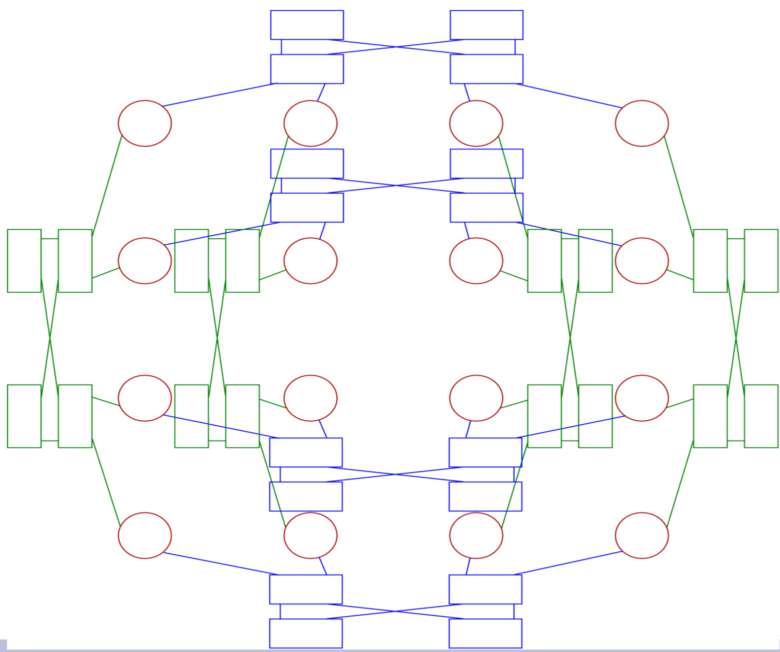
- There is a need for new topologies that escale in performance and energy

- Research in this topics

- Two examples:
- Ortogonal fat-tree (based on the fat-tree)
    - Reduce the number of paths available in the fat-tree
    - Reduce the fat-tree cost



-KNS
    - Based on a direct topology with reduced latency in each dimension
    - Smaller cost than the fat-tree

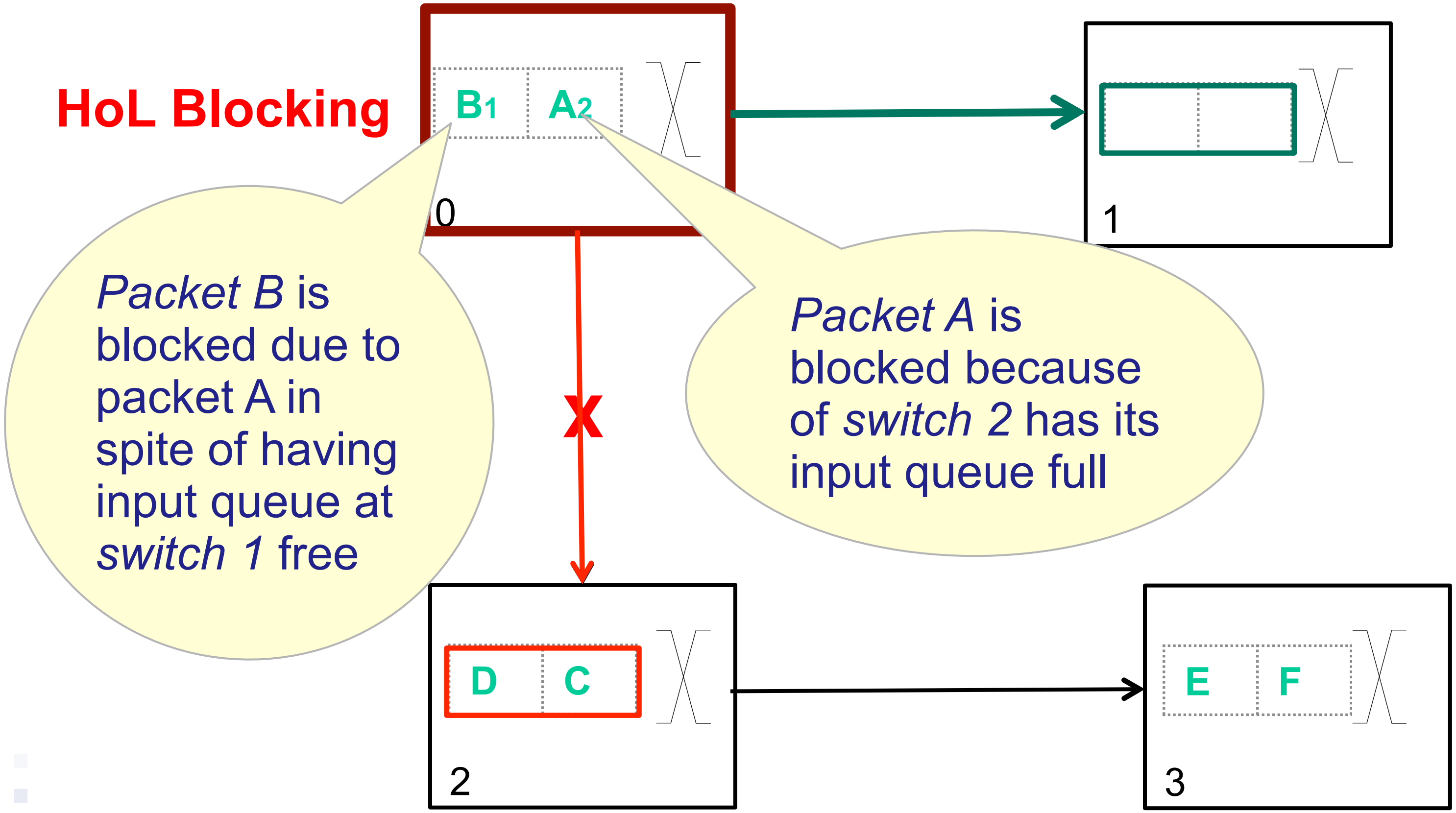


Main design parameters

- Routing Algorithm
  - Minimal routing
  - Reduce contention
  - Adaptive Routing
    - Avoid temporarily congested network areas
    - Introduce out-of-order delivery of packets
    - Introduce more HoL blocking effect
  - Deterministic Routing
    - Simpler
    - Introduce less HoL blocking effect
    - In-order delivery of packets

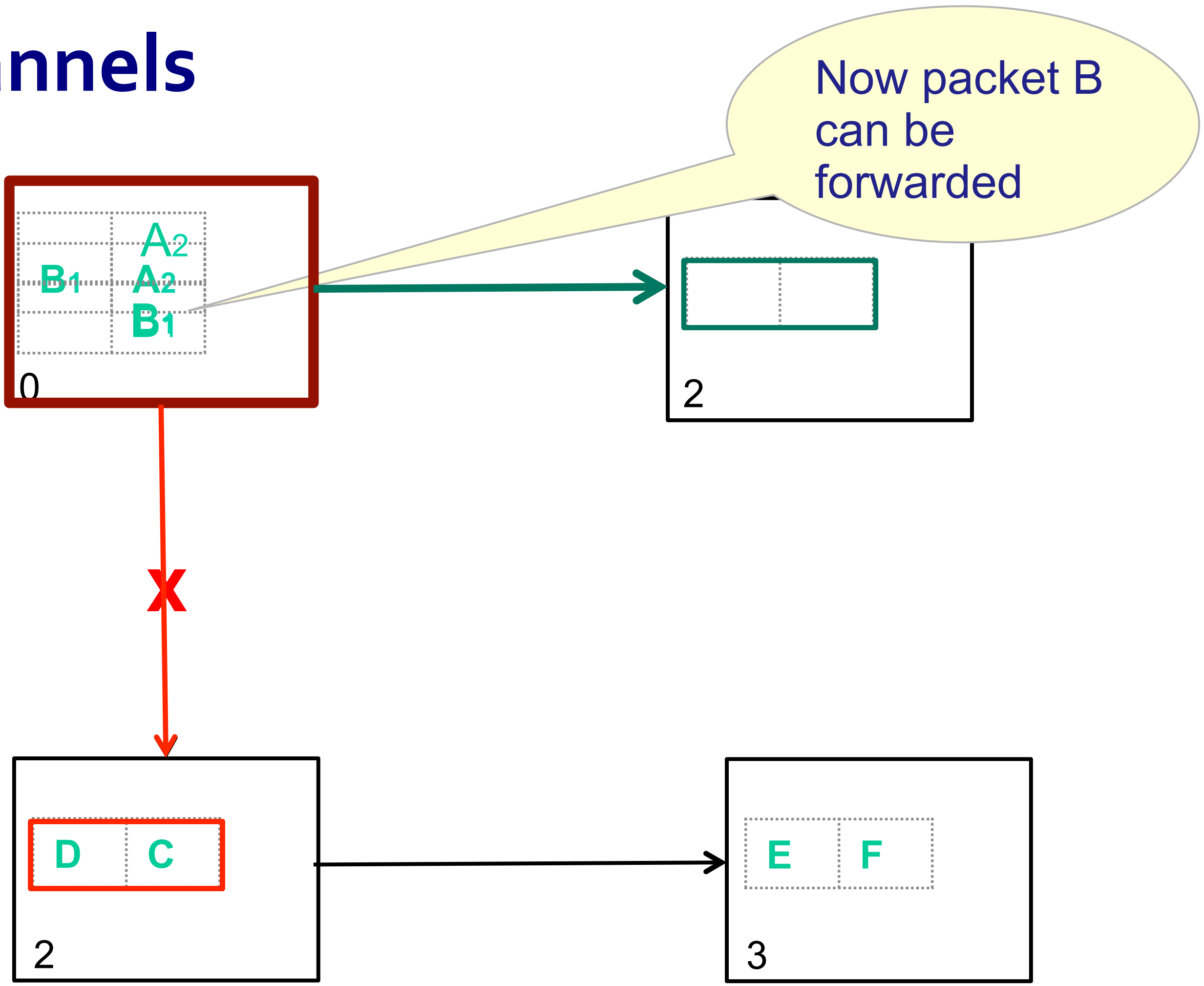
- HoL Blocking effect is a performance-limiting phenomenon
  - In larger machines it will have a higher impact on performance -> VCs

HoL Blocking effect



SOLUTION: Virtual Channels

Virtual channels



SOLUTION: Virtual Channels

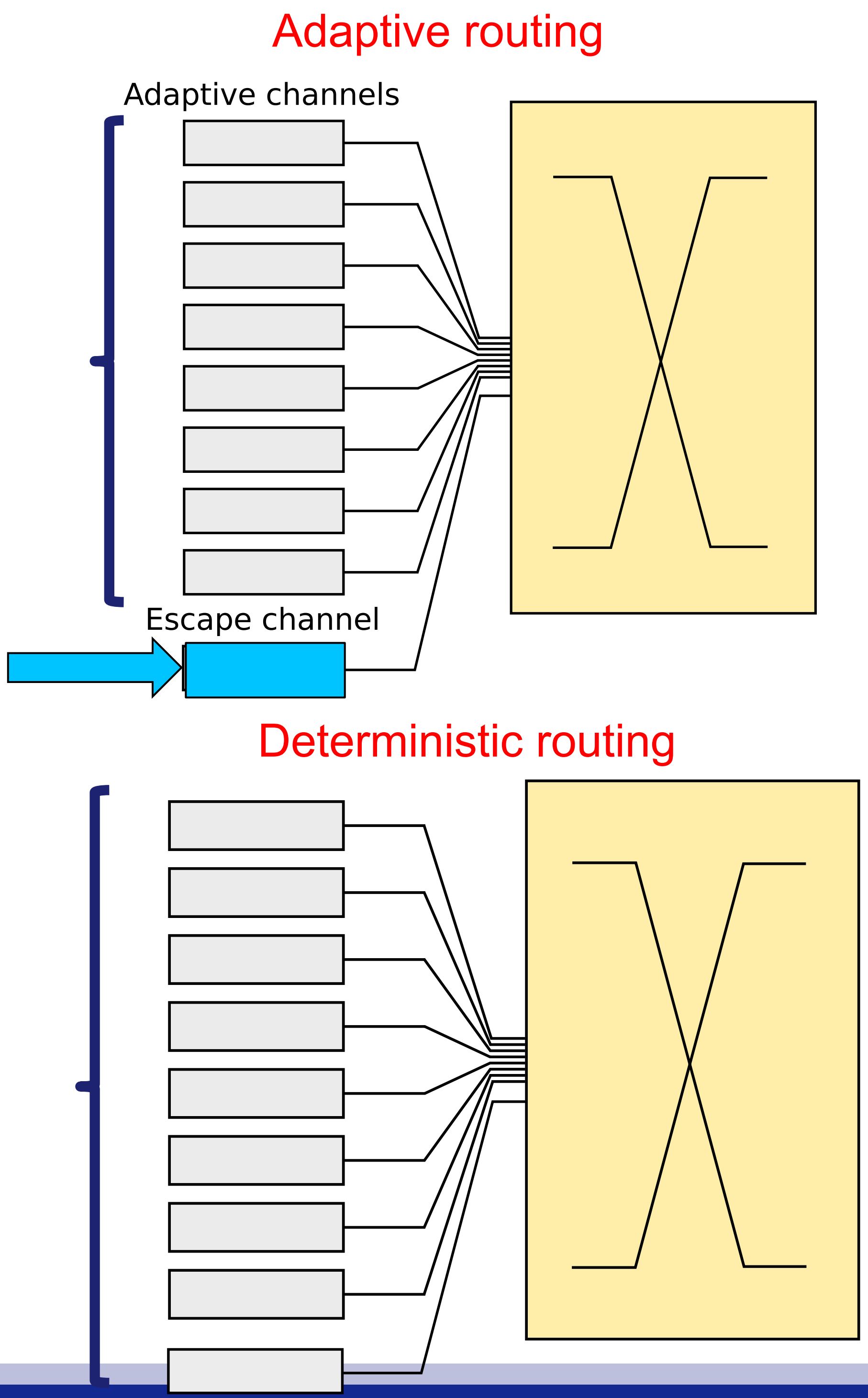
Links are better utilized

Latency reduced

Throughput increased

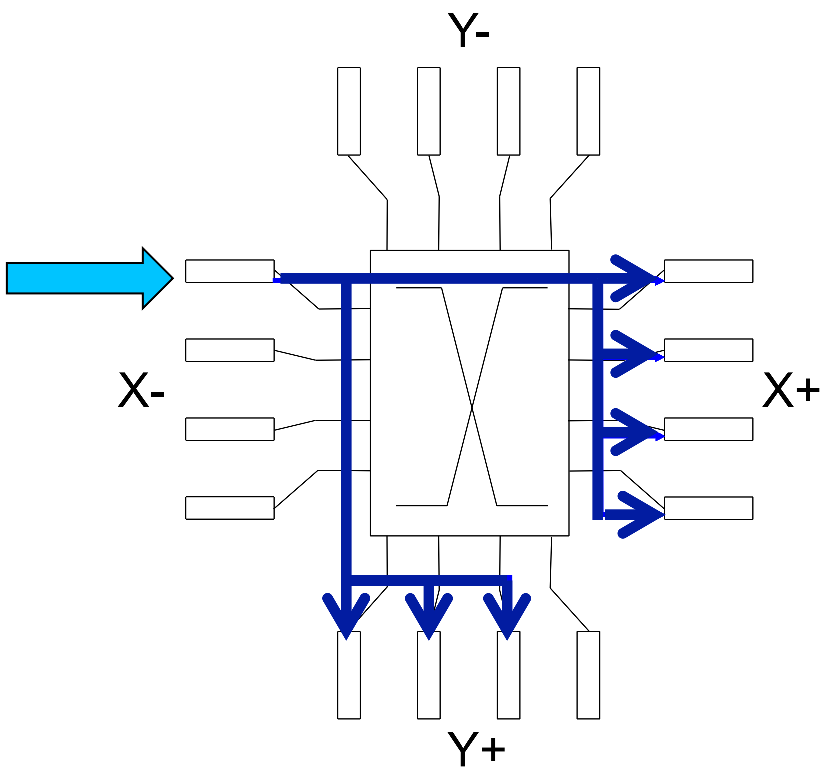
Use of VCs

- Adaptive routing
  - Deadlock avoidance
  - Direct network: escape channel
- And the others VCs or in deterministic routing?
  - Without restrictions
  - Congestion management
    - Restricting the use of VCs to destinations
    - HoL blocking



Fully Adaptive Routing

- VCs in all the dimension are used without restrictions
  - Except the escape channel

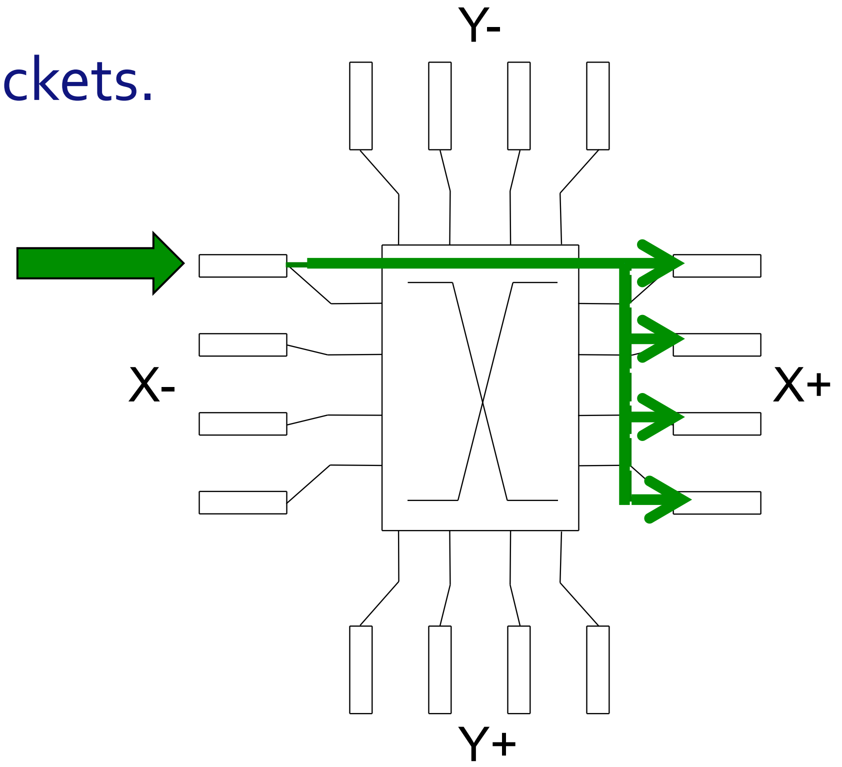


Deterministic routing
VCs without restrictions

OODET (Out-of-Order DETerministic routing):

-Several virtual channels.

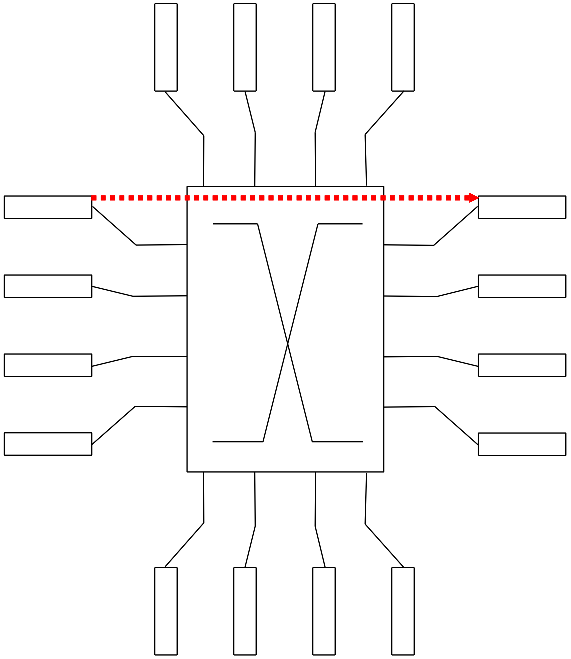
-Out of order delivery of packets.



Deterministic routing
VCs with restrictions

Different approaches to assign a single VC to each destination

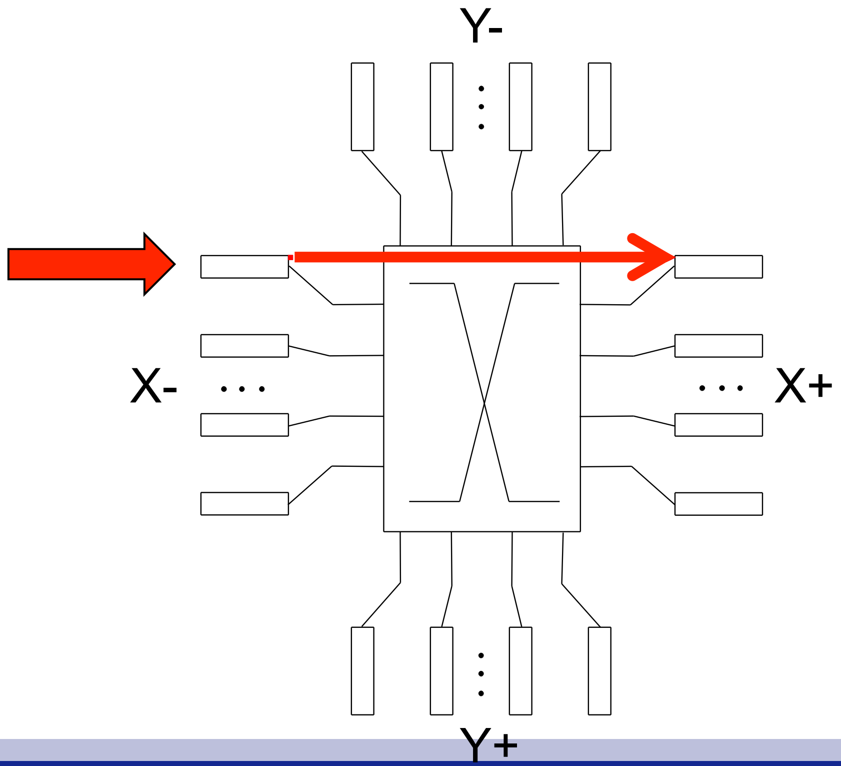
They try to classify destinations to
- Reduce the HoL blocking effect



Deterministic routing
VCs with restrictions

VOQ (Virtual Output Queue):

- VOQnet: We need as many virtual channels as number of network nodes.



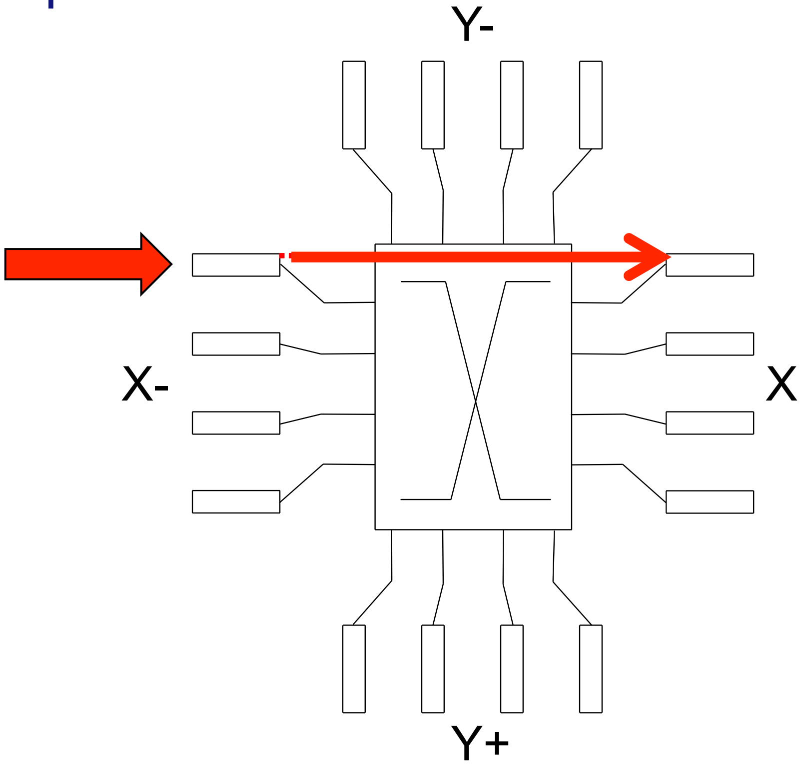
➡ It is not scalable.

Deterministic routing
VCs with restrictions

VOQsw:

-As many VCs as output ports.

-It depends on the number of output channels.

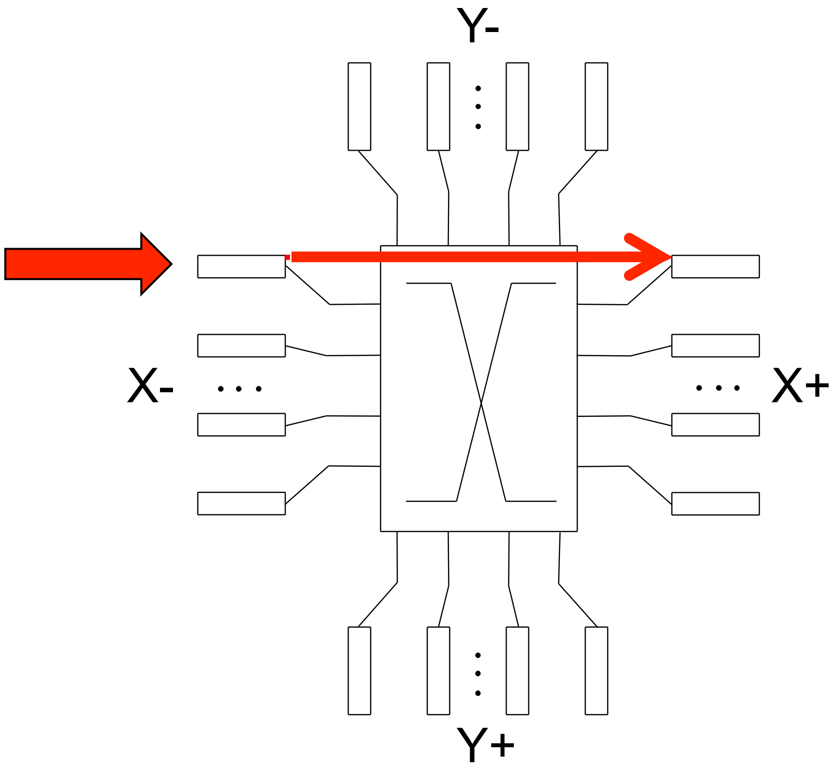


Deterministic routing
VCs with restrictions

VOQnet:

-As many VCs as destinations.

-Not scalable.

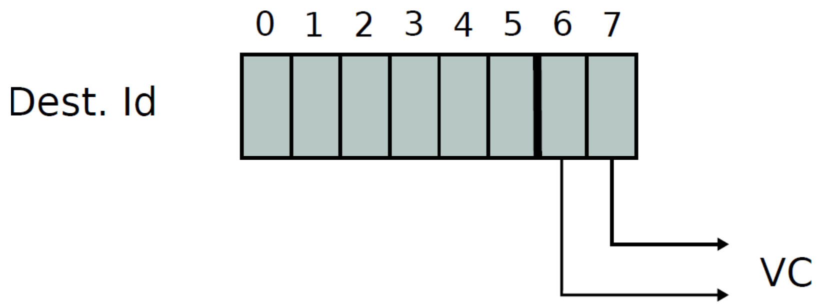


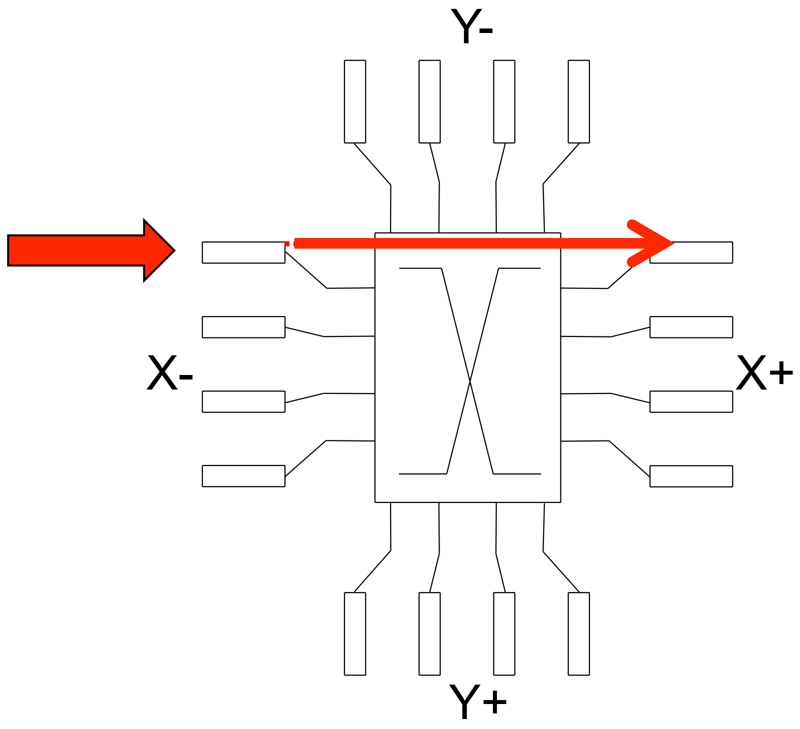
Deterministic routing
VCs with restrictions

DBBM:

-Based on VOQnet.

-VC is assigned by the least significant bits of the destination identifier

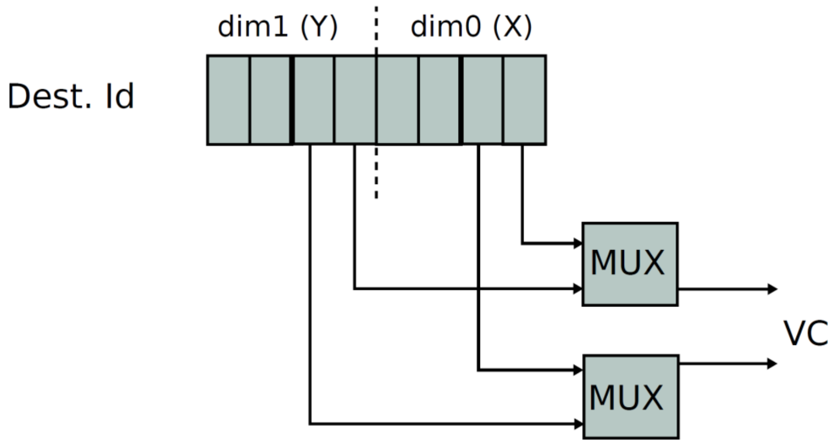


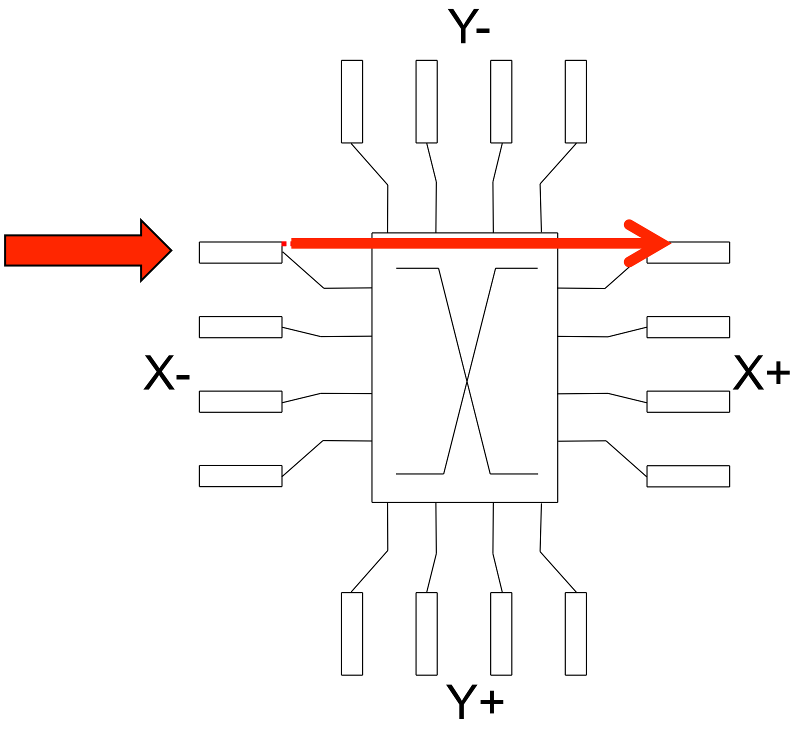


Deterministic routing
VCs with restrictions

IODET (In-Order DETerministic routing):

-Packets are assign to VCs using the least significant bits of **destination component** of the current dimension



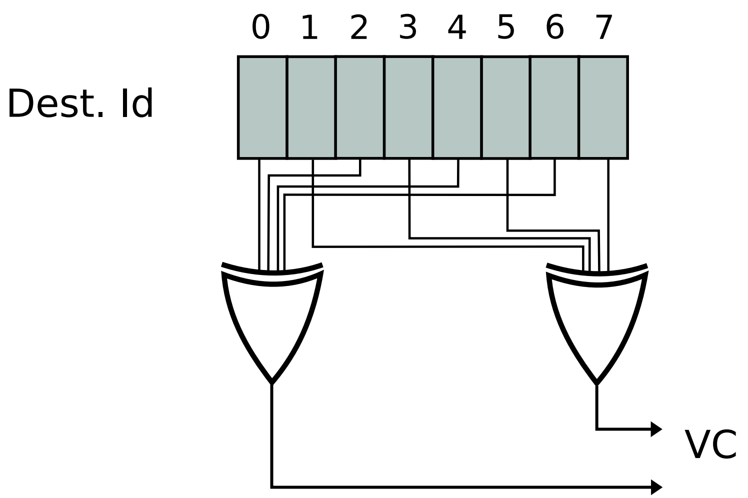


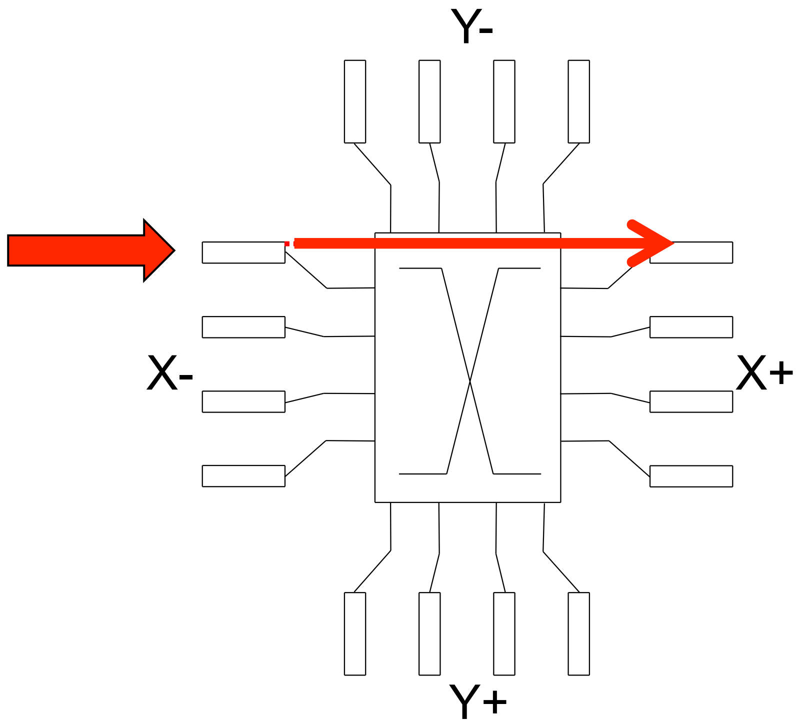
Deterministic routing
VCs with restrictions

XORDET (XOR deterministic):

-Packets are assigned to VCs using a XOR function of the destination bits

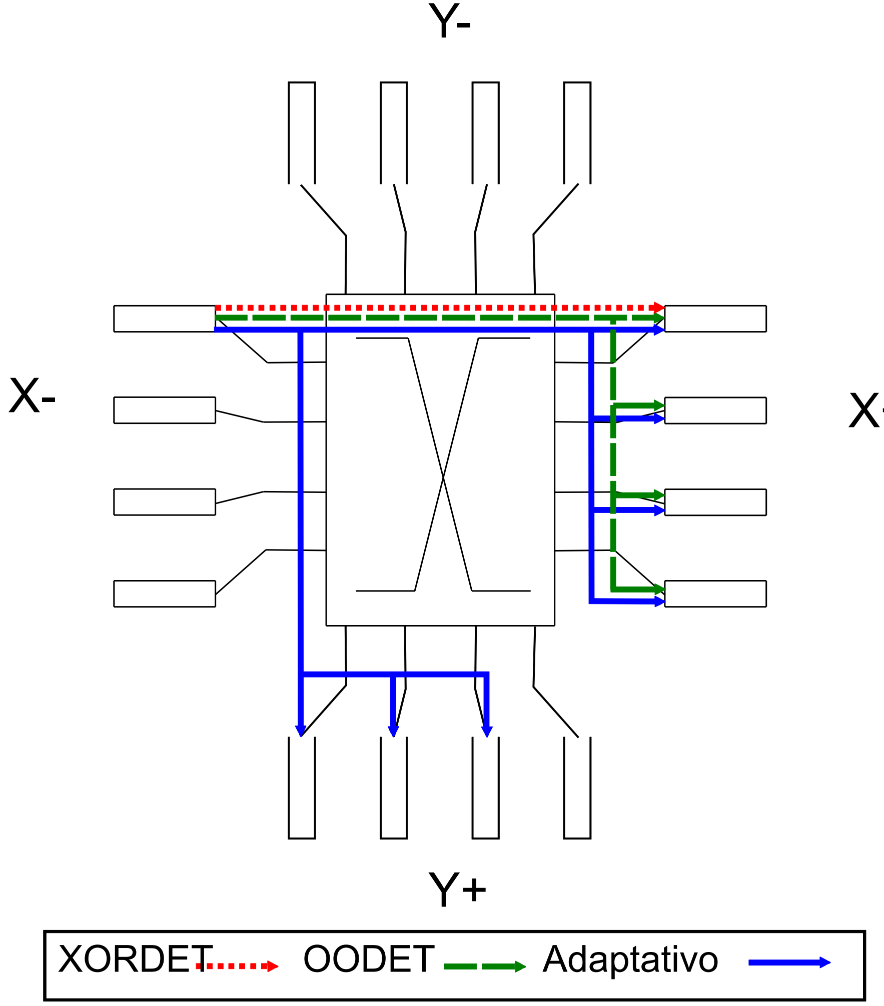
- More ramdom VC selection
- Balanced use of VCs





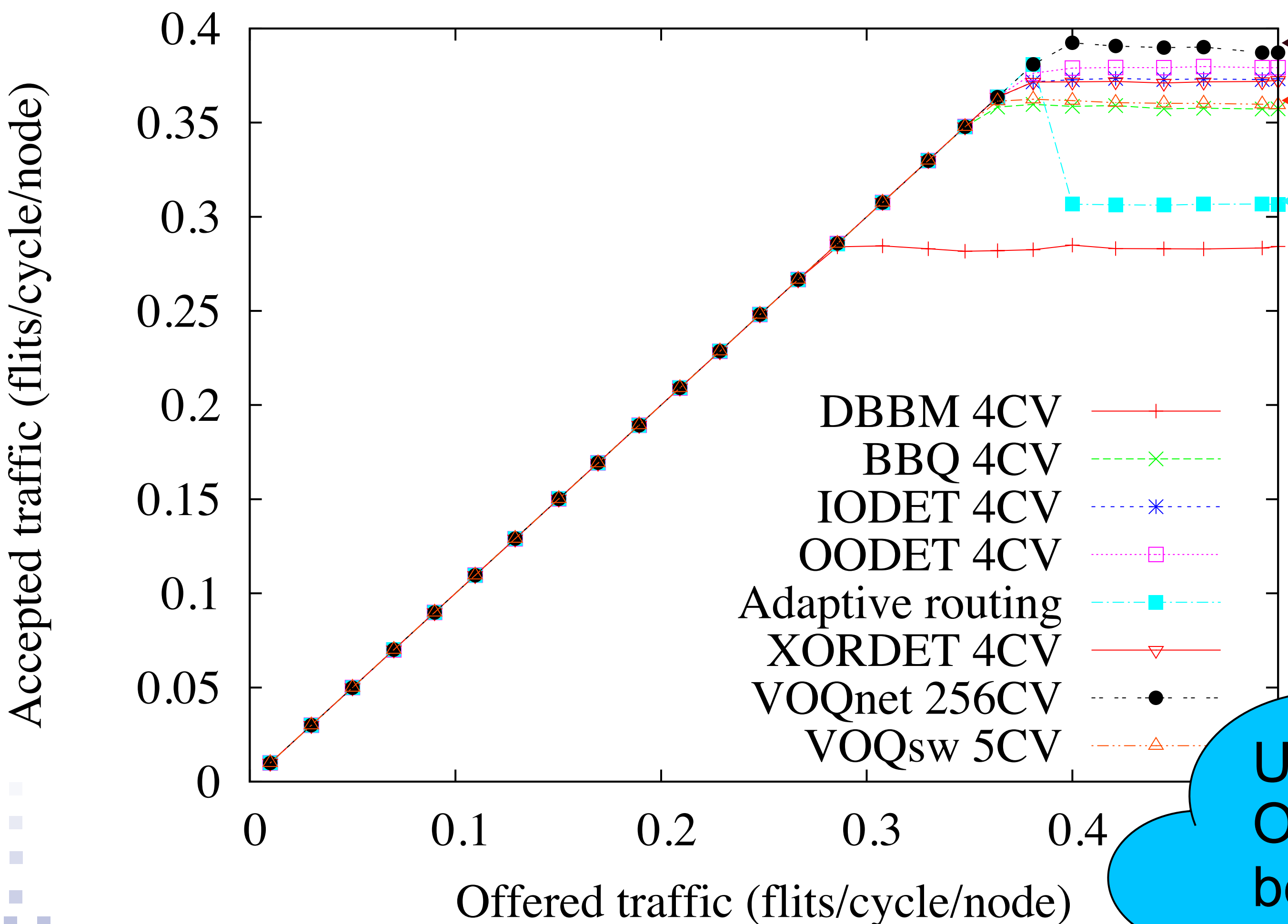
Deterministic routing
VCs without restrictions

Example: packet that must be forwarded through the two dimensions



16x16 Torus Uniform traffic

4 CVs



VOQnet the best with 256 VCs

OODET near VOQnet

XORDET near OODET

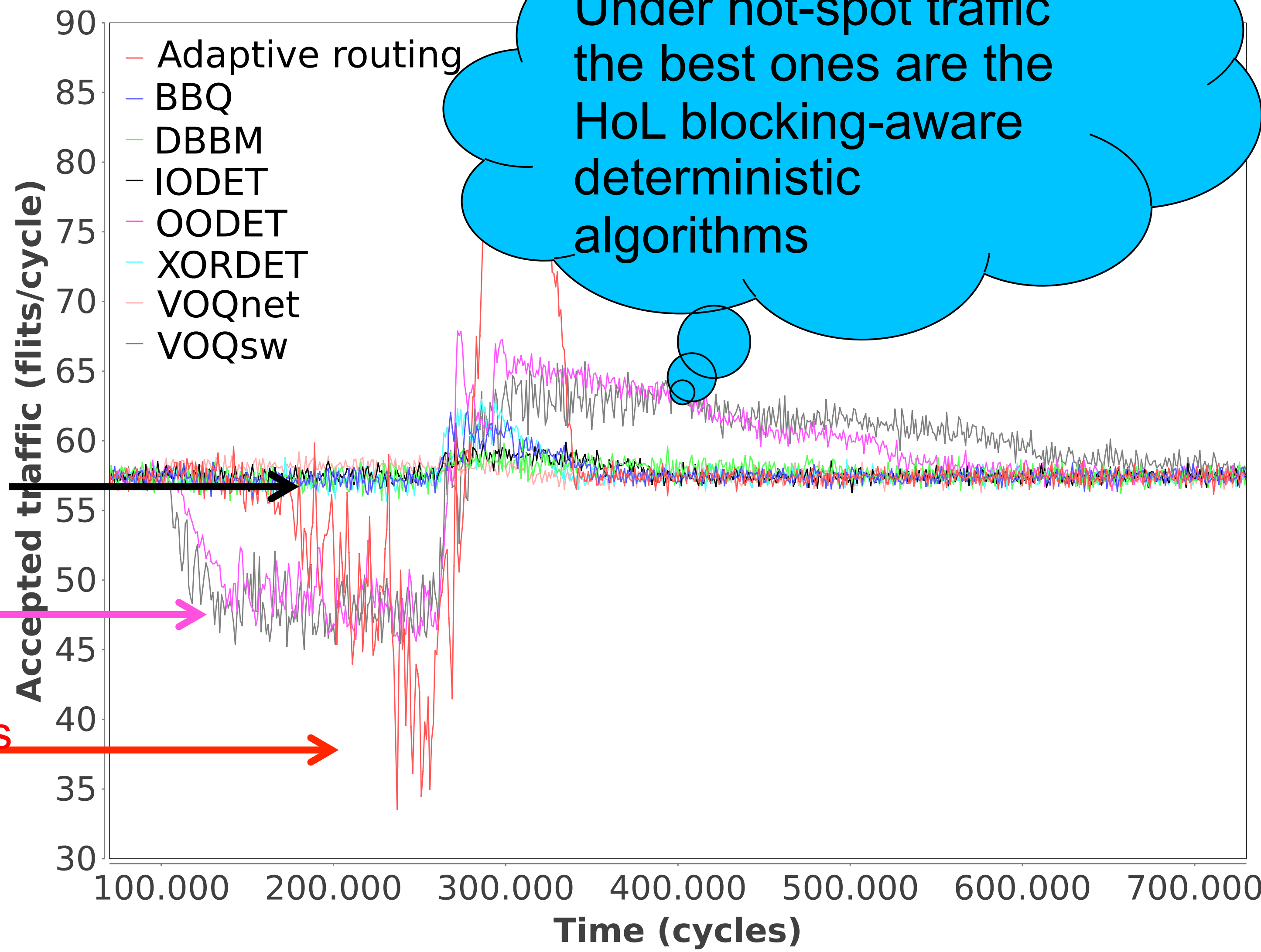
Adaptive routing as good as the deterministic routing algorithm

Under uniform traffic OODET is a tradeoff between adaptivity and destination classification

16x16 Torus Hot-spot traffic

8 CVs

25\% of network nodes send packets only to one node (the hot-spot node) during a period of time.



Under hot-spot traffic the best ones are the HoL blocking-aware deterministic algorithms

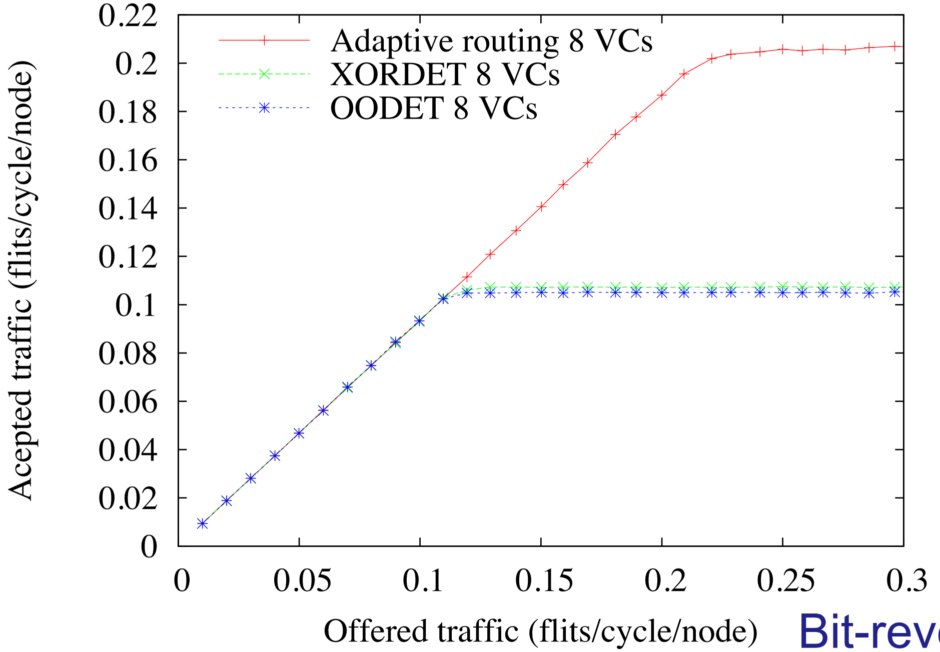
The deterministic ones are the best ones

OODET suffers from some degradation

Adaptive routing suffers a great performance degradation

16x16 Torus (adversarial traffic) 8 CVs

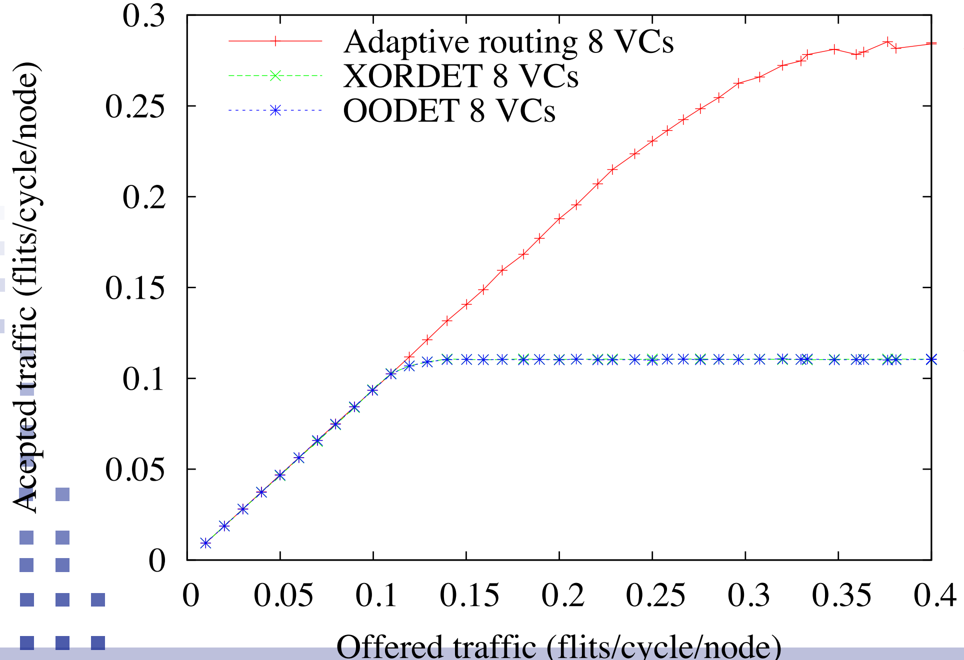
Matrix transpose traffic



Adaptive routing is able to cope with adversarial traffic

Deterministic routing is not able

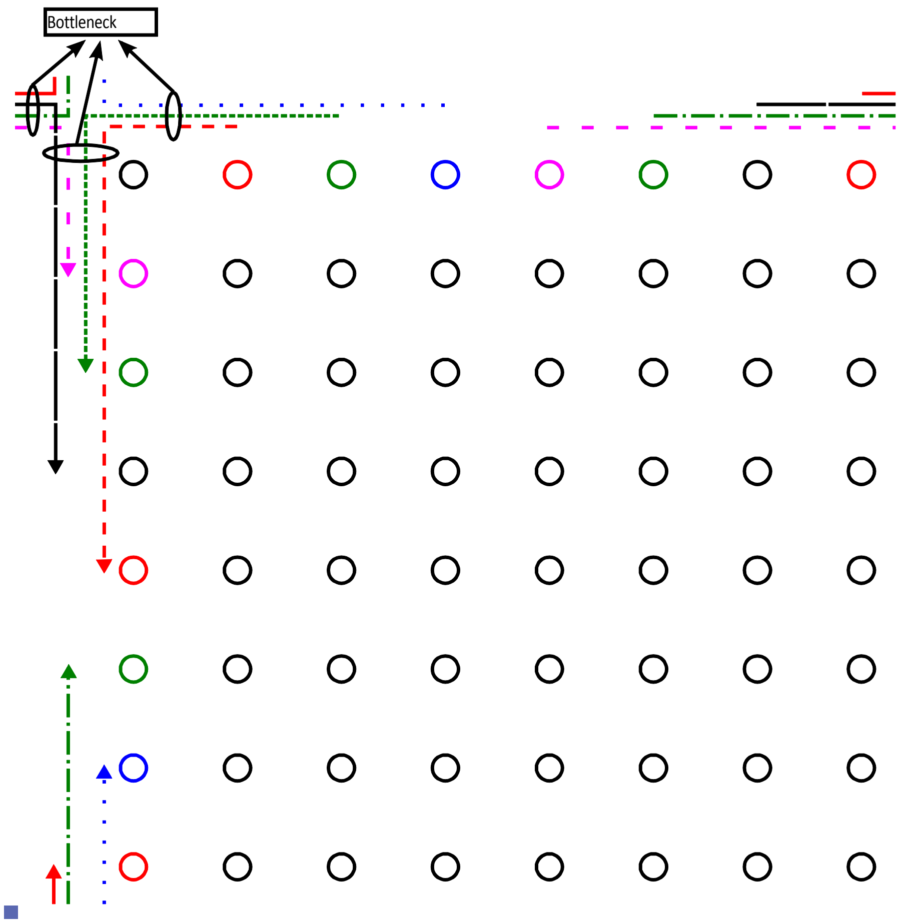
Bit-reversal traffic



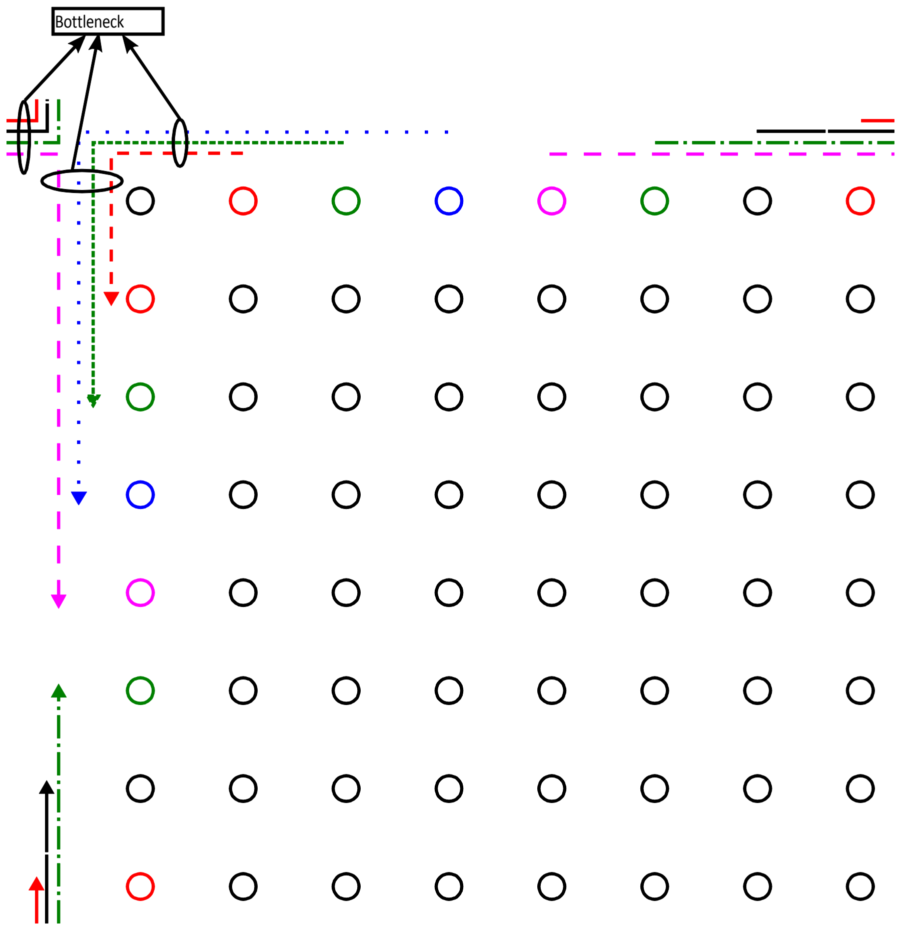
Deterministic routing
VCs without restrictions

Deterministic routing problems:

Bit-Reversal traffic pattern:



Matrix transpose traffic pattern:



XORADAP: XOR ADAPtive routing

XORADAP (Example with 8 virtual channels):

Flexibility of adaptive routing based on Duato's algorithm:

Several adaptive channels and at least one escape channel.

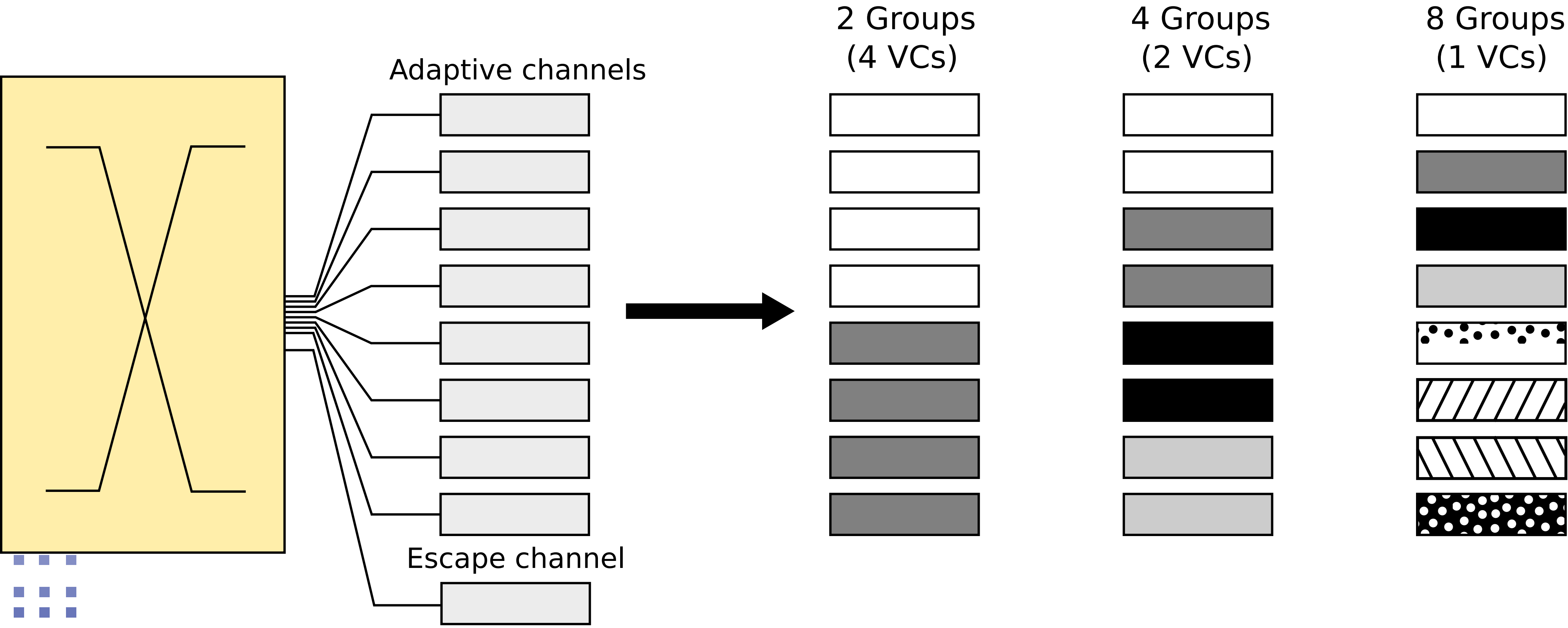
Restricted use of VC's:

They are split in groups and chosen by XOR function.

Each destination can use a subset of the VCs

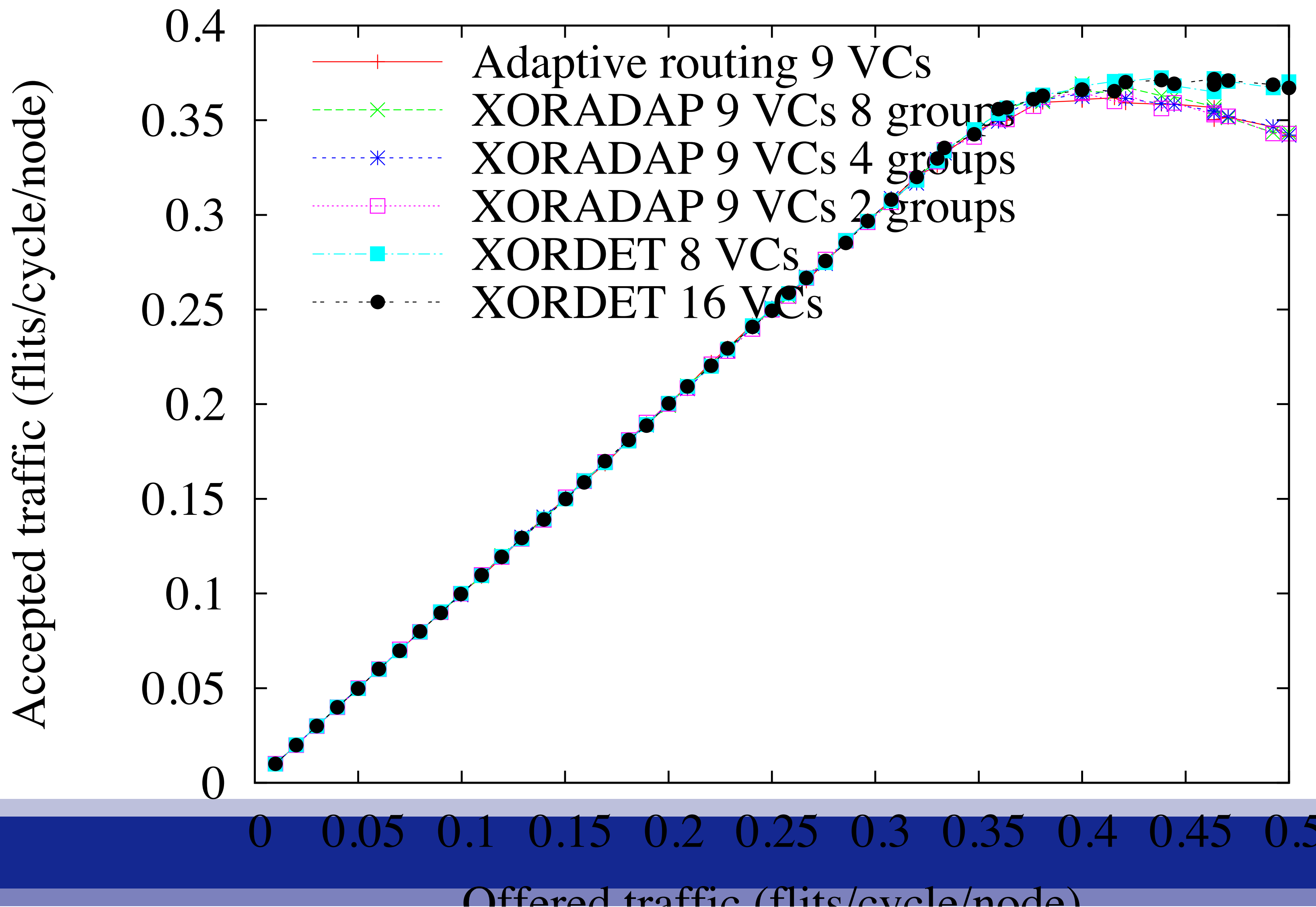
XORADAP: XOR ADAPtive routing

XORADAP (Example with 8 virtual channels):



Uniform traffic

16x16 torus, accepted traffic vs. average packet latency with 9 virtual channels (8 XORDET):



Hot spot traffic

16x16 torus, uniform traffic pattern with hot-spot and 9 virtual channels (8 and 16 virtual channels with XORDET):

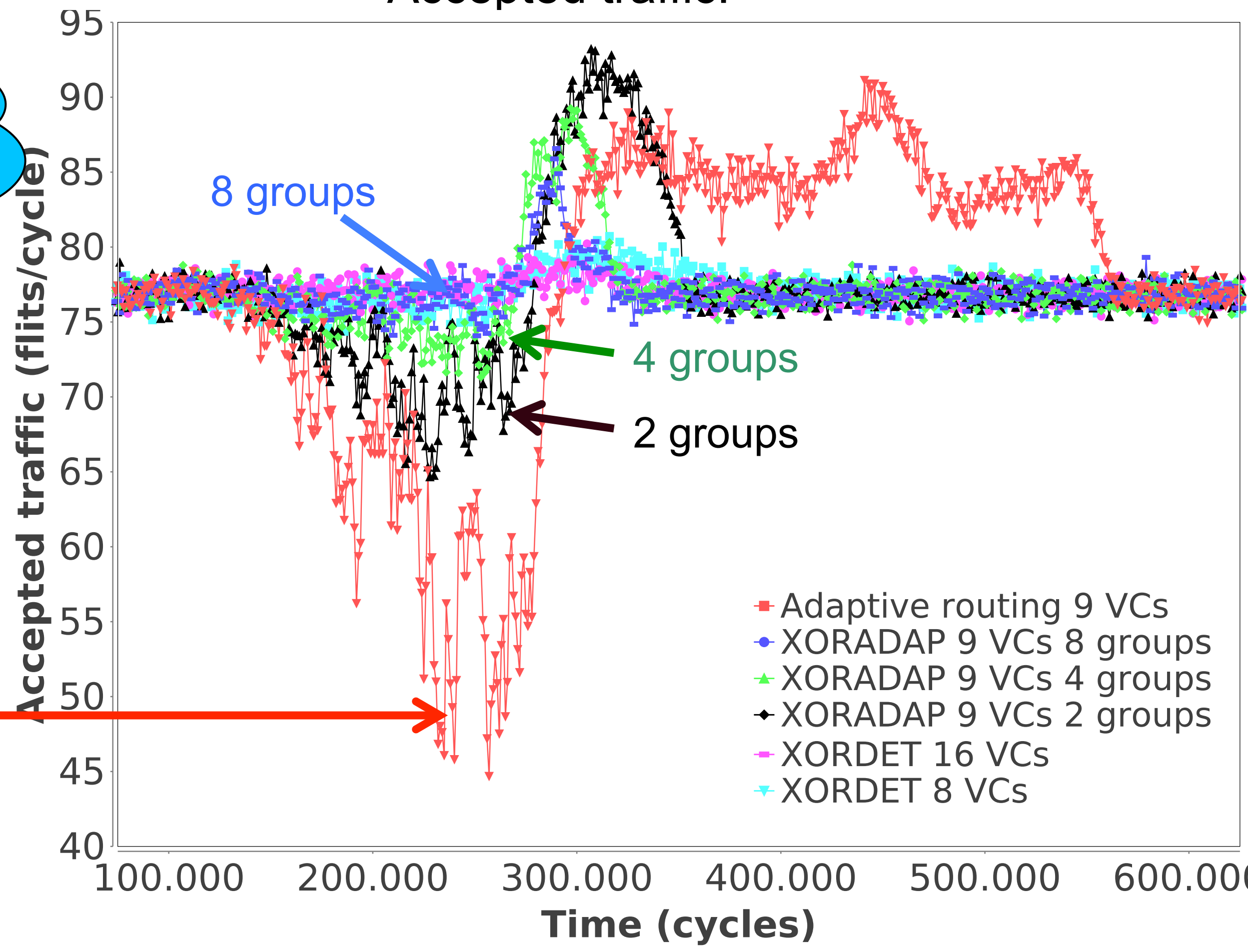
Uniform traffic: 75% of nodes -> 0,4 flits/cycle to a random node (except hot spot node).
Hot spot traffic: 25% of nodes -> 0,0156 flits/cycle (1 f/c in total) to the hot spot node.

XORAdap is able to deal with hot-spot traffic

With XORAdap as we increase the number of groups we control the network congestion

Adaptive routing suffers a great performance degradation

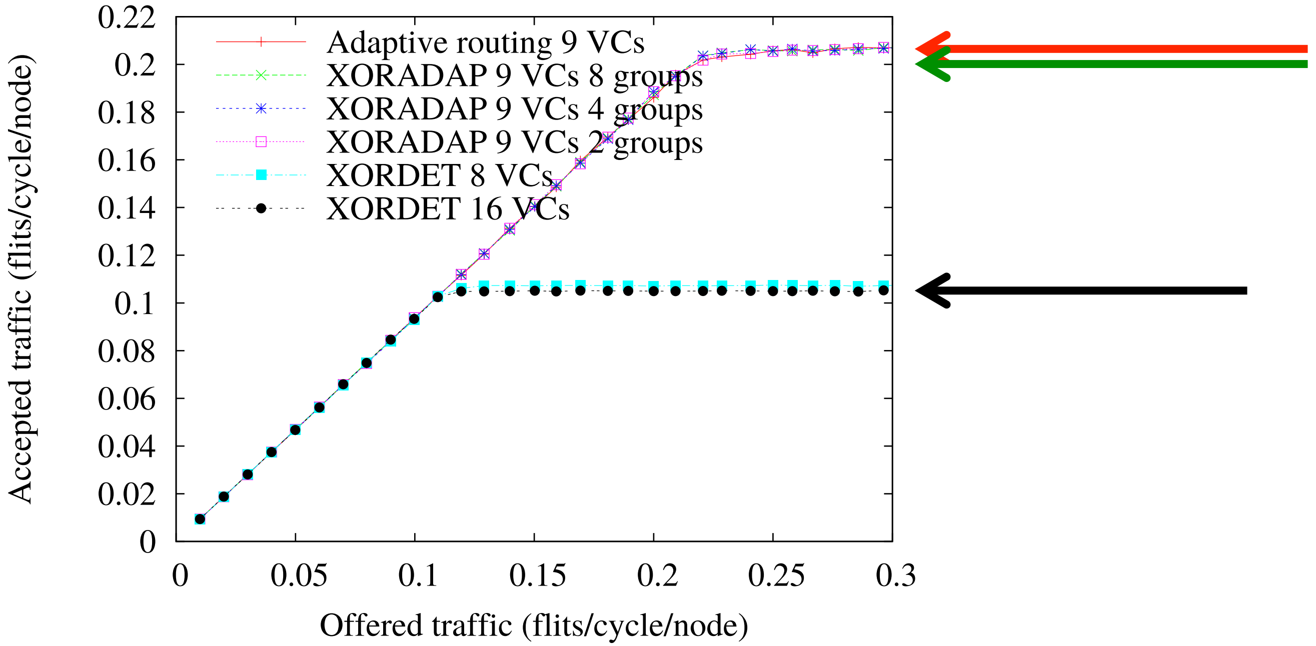
Accepted traffic:



16x16 Torus (adversarial traffic)

8 CVs

Matrix transpose traffic

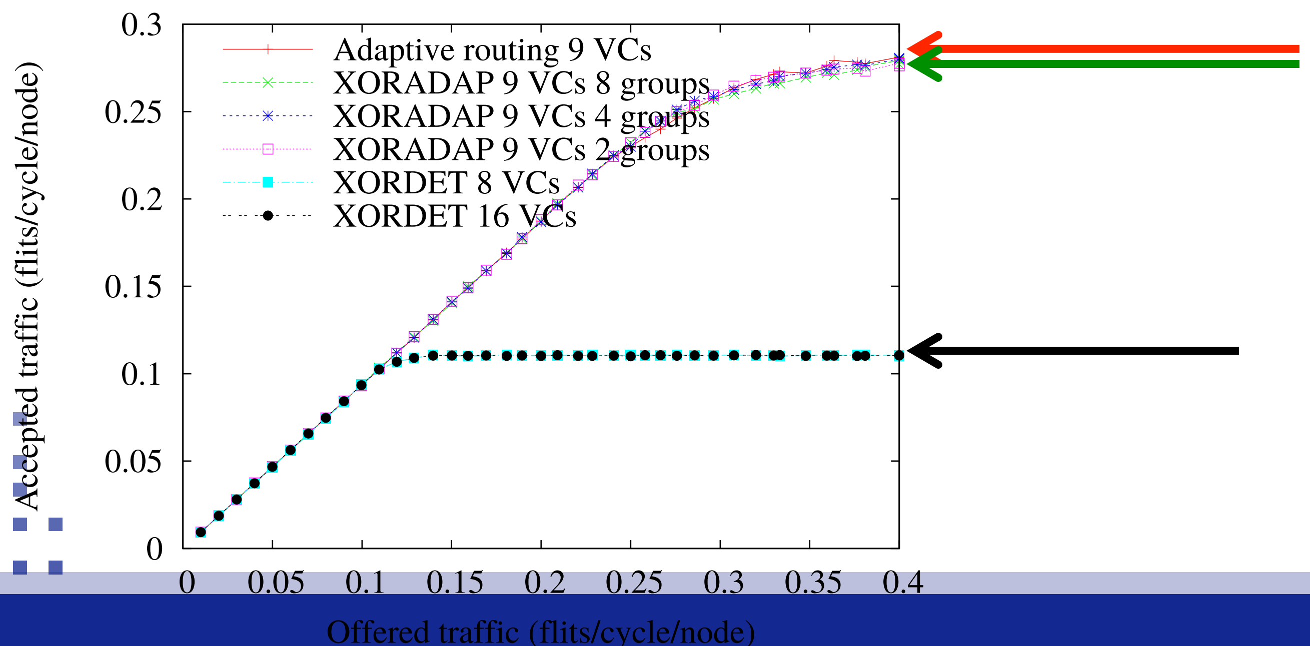


Adaptive routing is able to cope with adversarial traffic

XORAdap obtains similar results to adaptive routing indendently on the number of groups

Deterministic routing is not able

Bit-reversal traffic



Conclusions

- Combining adaptivity with HoL blocking mechanisms in the VCs provides good performance results with changing traffic patterns

    - It can isolate the packets destined to the hot-spot, like HoL-blocking aware deterministic routing does.
    - It is able to achieve the flexibility of adaptive routing to avoid congested areas (adversarial traffic)

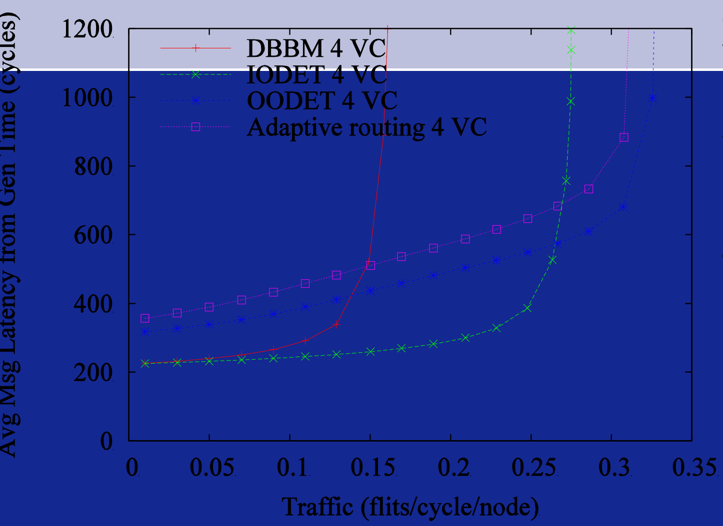
Thank you!

megomez@disca.upv.es

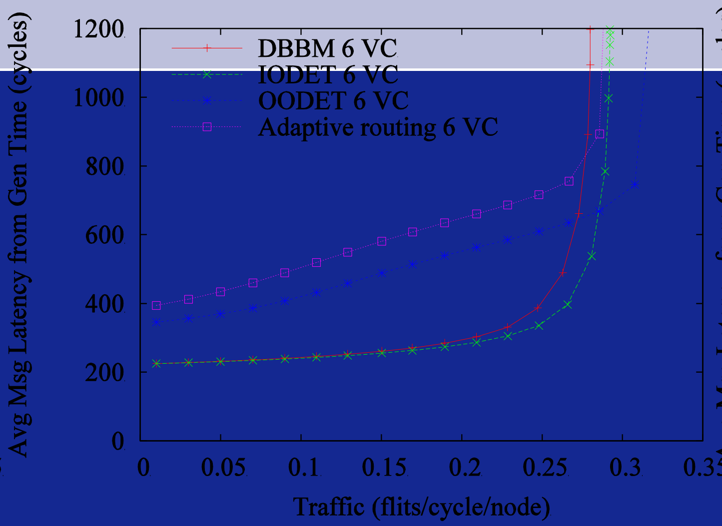
4. Resultados Experimentales

Torus 16x16, uniform traffic

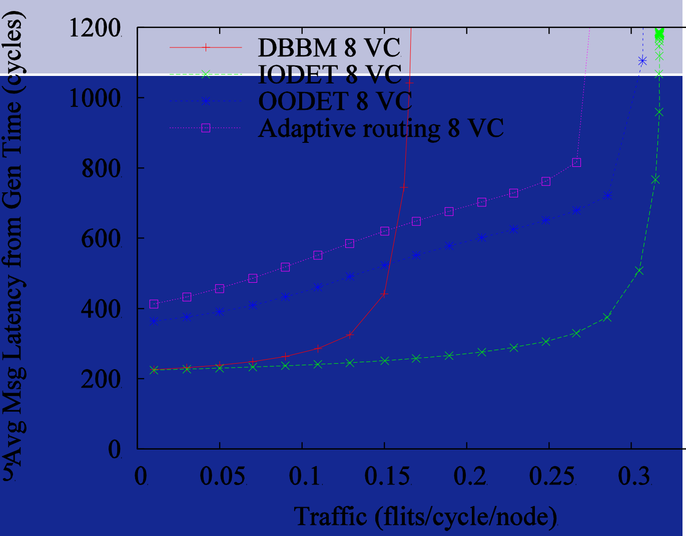
4 virtual channels



6 virtual channels:

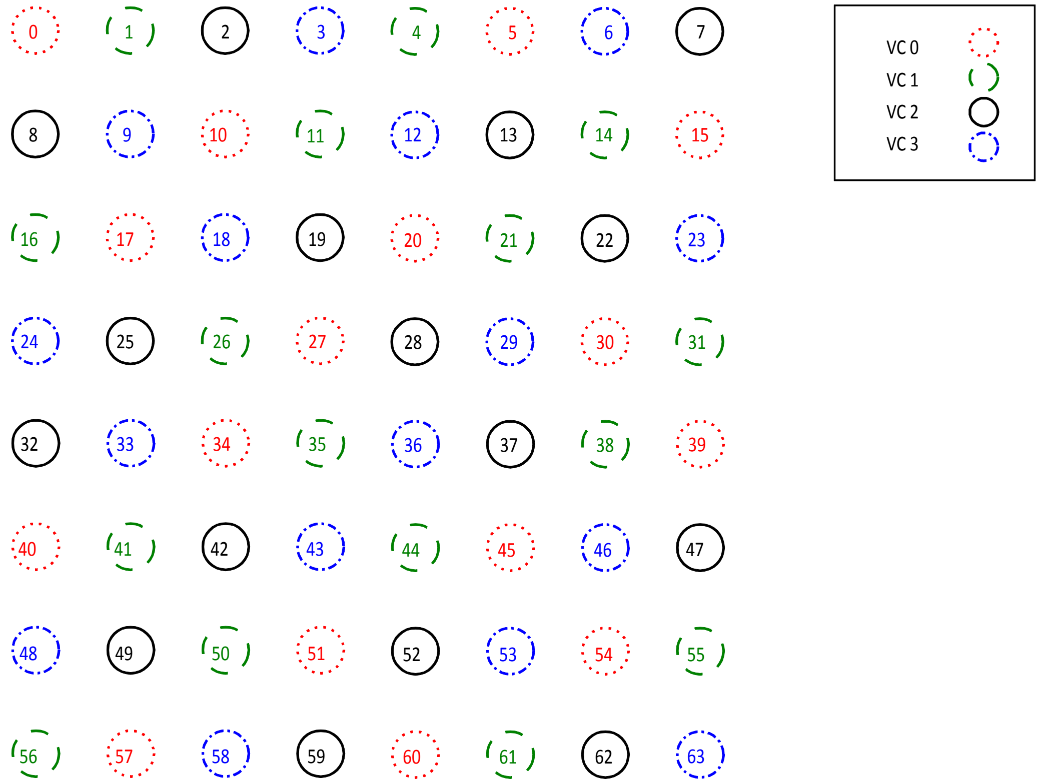


8 virtual channels:



Deterministic routing
VCs without restrictions

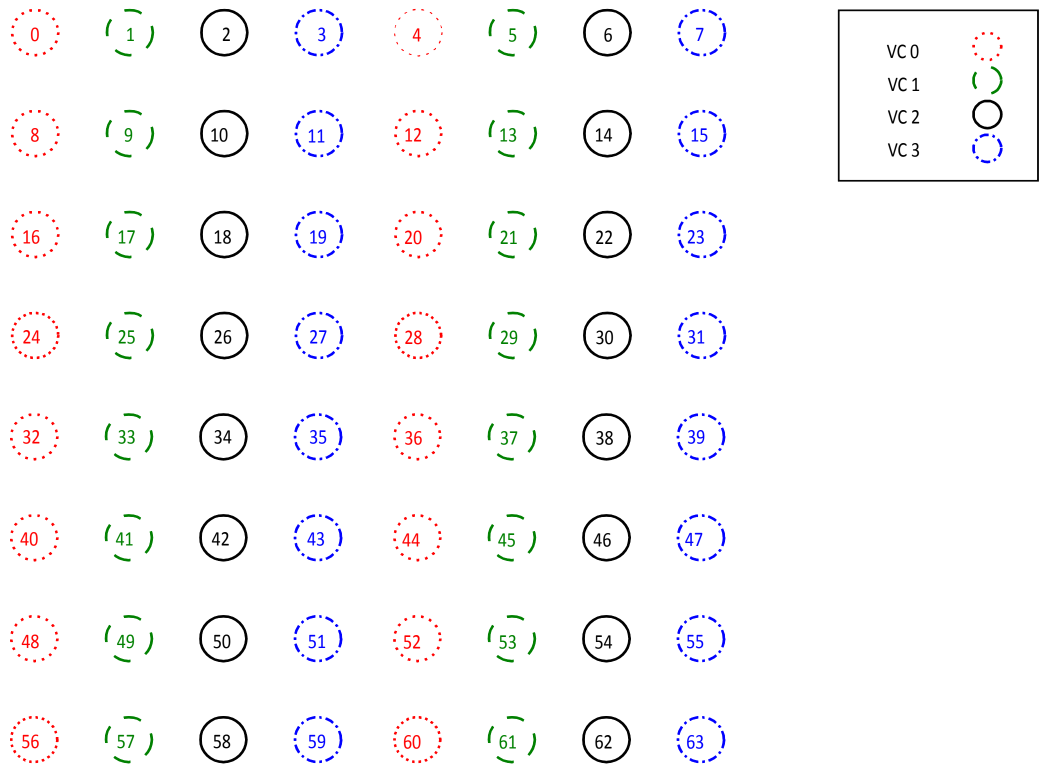
XORDET (Example with 4 virtual channels):



Dim	VC#0	VC#1	VC#2	VC#3
X	14	14	14	14
Y	1	2	2	2

Deterministic routing VCs without restrictions

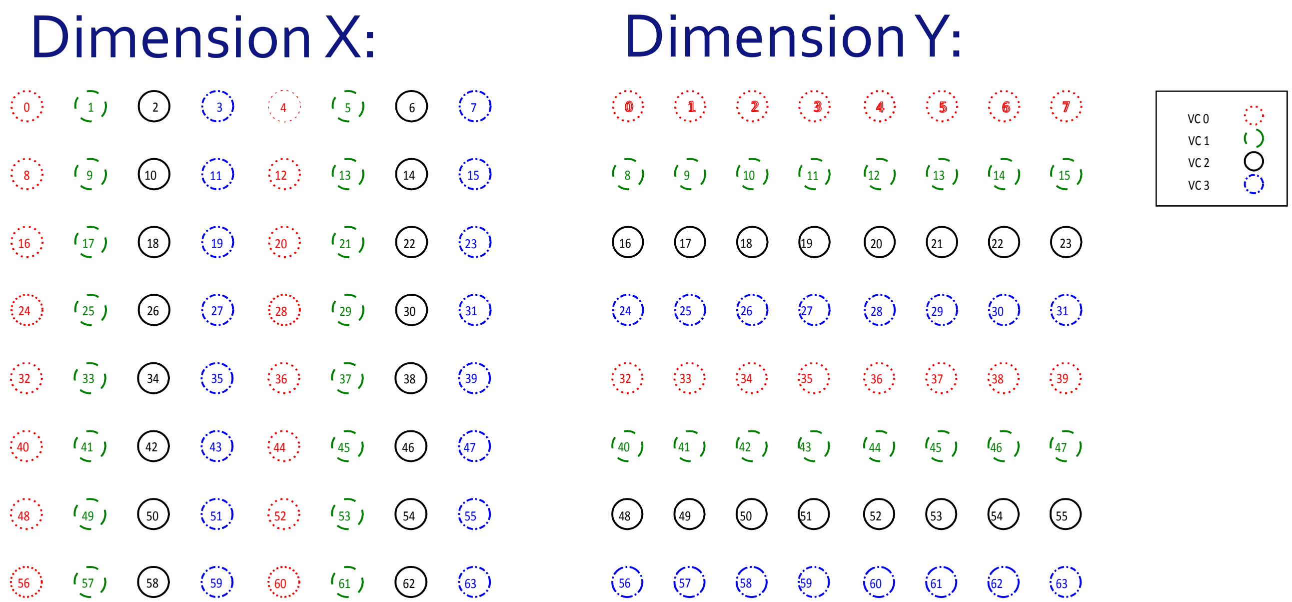
DBBM (Example with 4 virtual channels):



Dim	VC#0	VC#1	VC#2	VC#3
X	8	16	16	16
Y	7	No dest.	No dest.	No dest.

Deterministic routing
VCs without restrictions

IODET (Example with 4 virtual channels):

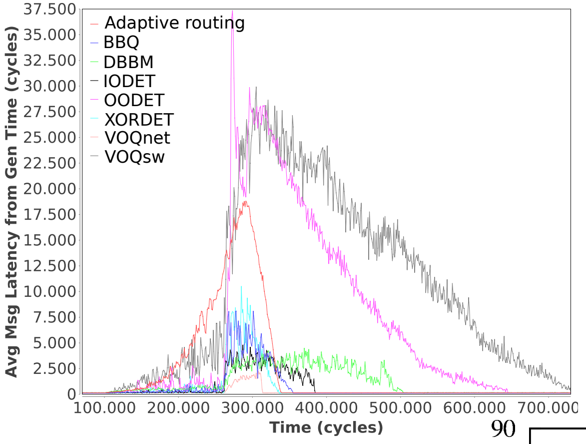


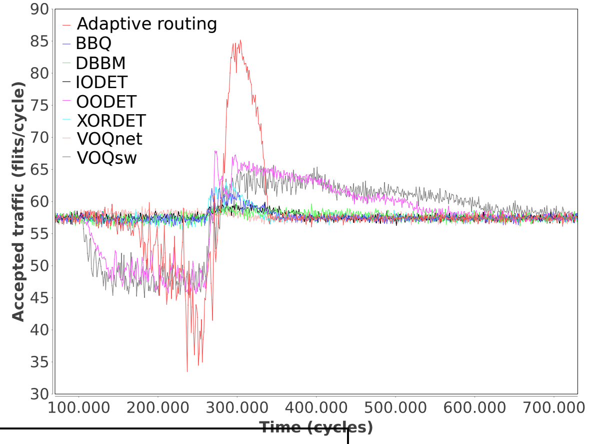
Dim	VC#0	VC#1	VC#2	VC#3
X	8	16	16	16
Y	1	2	2	2

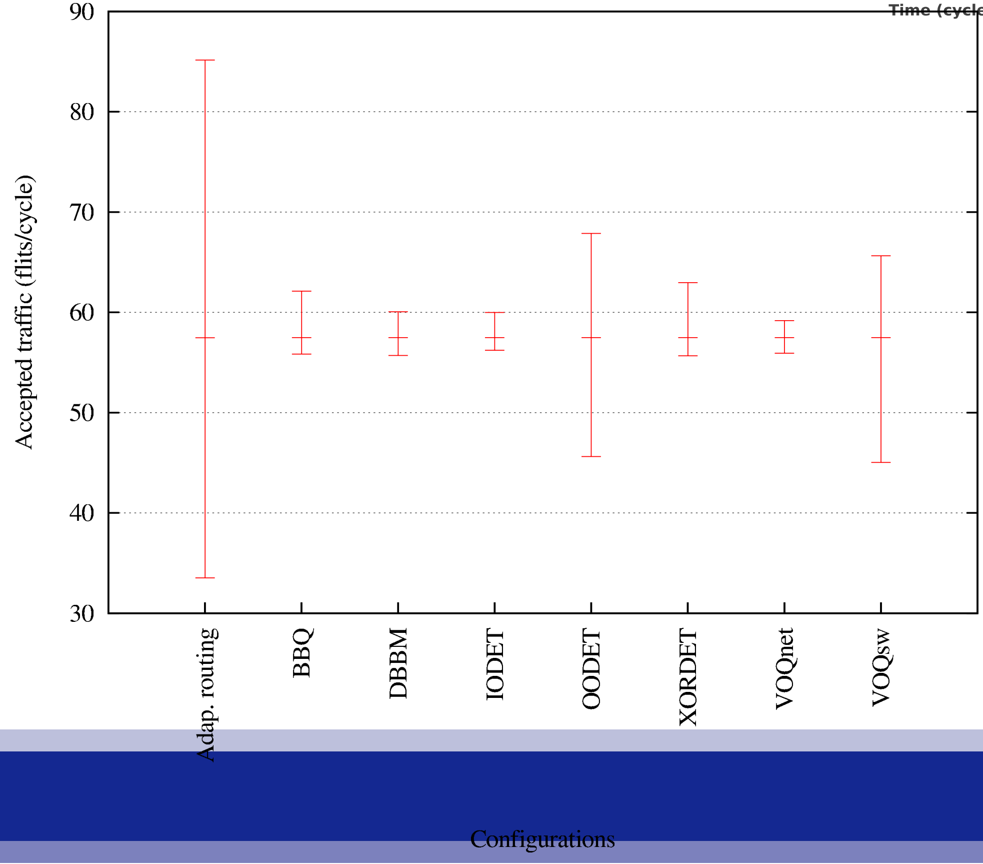
16x16 Torus Hot-spot traffic

8 CVs

25\% of network nodes send packets only to one node (the hot-spot node) during a period of time.







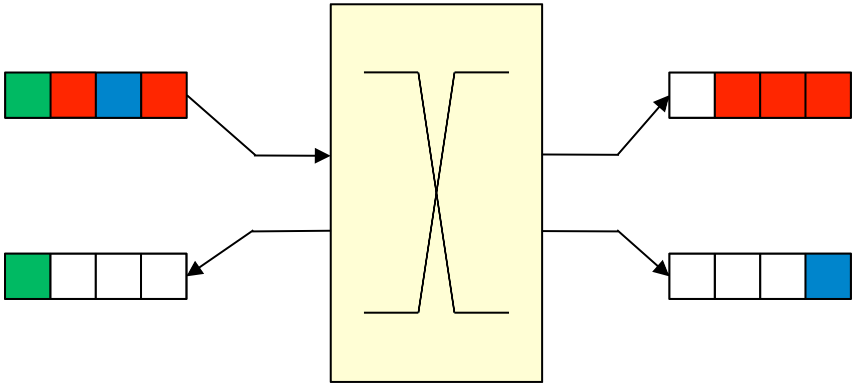
1. Introduction

Routing Algorithms:

- Deterministic Routing.
- Adaptive Routing.

Another factor to consider:

- The Head of Line (HoL) Blocking effect.



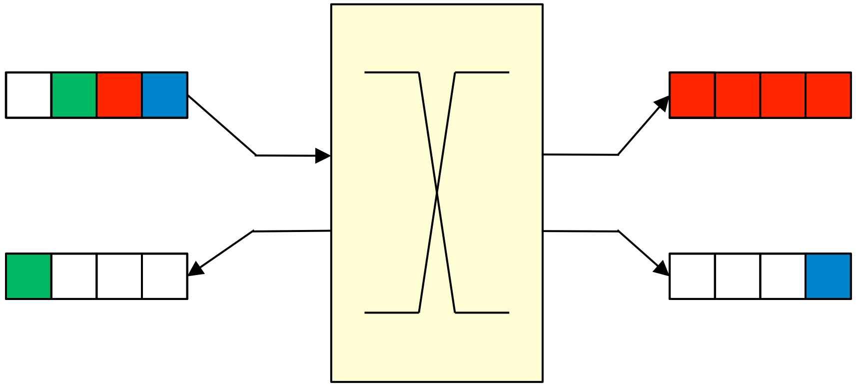
1. Introduction

Routing Algorithms:

- Deterministic Routing.
- Adaptive Routing.

Another factor to consider:

- The Head of Line (HoL) Blocking effect.



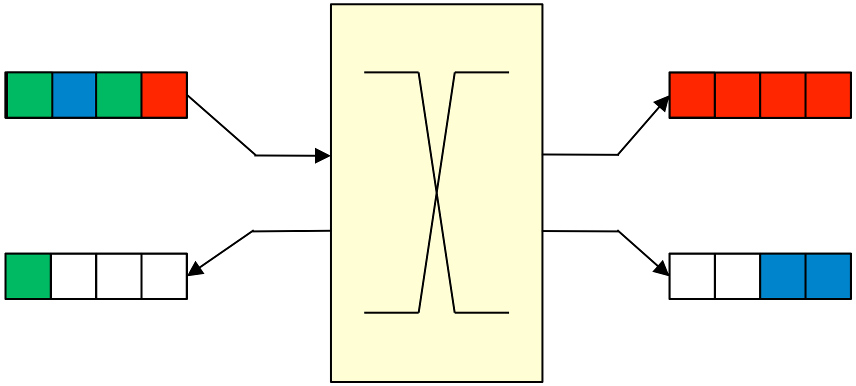
1. Introduction

Routing Algorithms:

- Deterministic Routing.
- Adaptive Routing.

Another factor to consider:

- The Head of Line (HoL) Blocking effect.



1. Introduction

Routing Algorithms:

- Deterministic Routing.
- Adaptive Routing.

A key factor to consider:

- The Head of Line (HoL) Blocking effect.