DEPARTAMENTO DE SISTEMAS INFORMÁTICOS

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

Providing Differentiated Services, Congestion Management, and Deadlock Freedom in Dragonfly Networks

Pedro Yébenes, Jesús Escudero-Sahuquillo, Pedro J. García,

Francisco J. Alfaro, Francisco J. Quiles

University of Castilla – La Mancha

Outline

- Motivation
- Background
- Proposals description
- Evaluation
- Conclusions

Outline

- **Motivation**
- Background
- Proposals description
- Evaluation
- Conclusions

Motivation

Interconnection Networks

- Interconnection networks are **key elements** in HPC systems and datacenters.
  - Thousands of processing and/or storing nodes.
  - Applications need increasing computing power.
- The interconnection network may become the **system bottleneck** if not properly configured.

Achieving high network performance is mandatory.

Tianhe-2 (MilkyWay-2)
16000 nodes - Cores 3120000
TH-Express 2
1st Top500 (November 2015)

Motivation

Interconnection Networks

- Main design aspects of interconnection networks:
  - Topology
  - Routing Algorithm
  - Power consumption
  - Fault tolerance
  - Congestion control
  - Quality of service

Motivation

Problem Statement

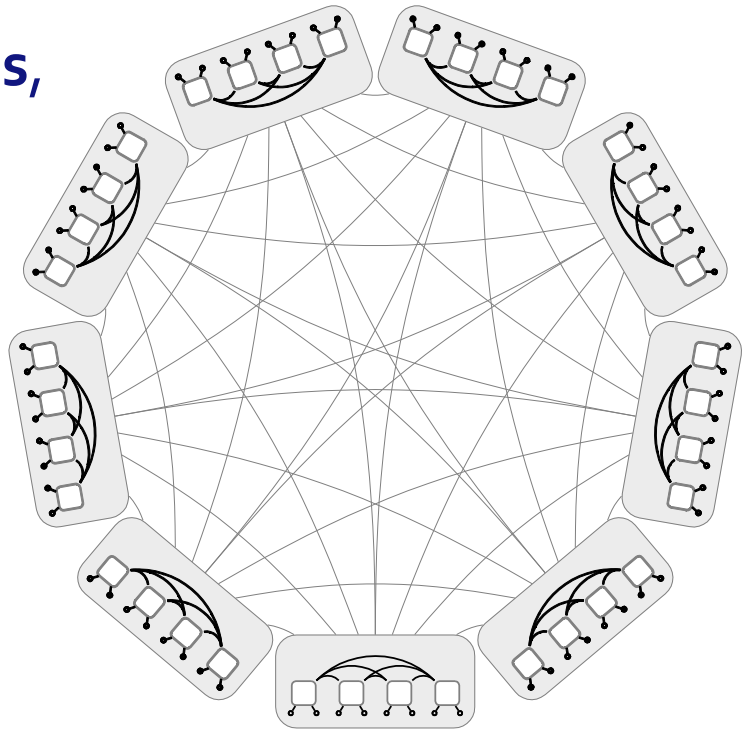
- Minimal-path routing for Dragonfly networks is **not deadlock free** by default, requiring additional Virtual Channels (VCs) for deadlock freedom.
- Both congestion management and QoS can be provided by separating traffic flows into VCs.
- Thus, congestion management, QoS provision, and deadlock freedom **require VCs for different purpose**.
- There is not a joint and straightforward solution that offers these three functionalities at the same time.

Outline

- Motivation
- **Background**
- Proposals description
- Evaluation
- Conclusions

Background
Dragonfly Topology

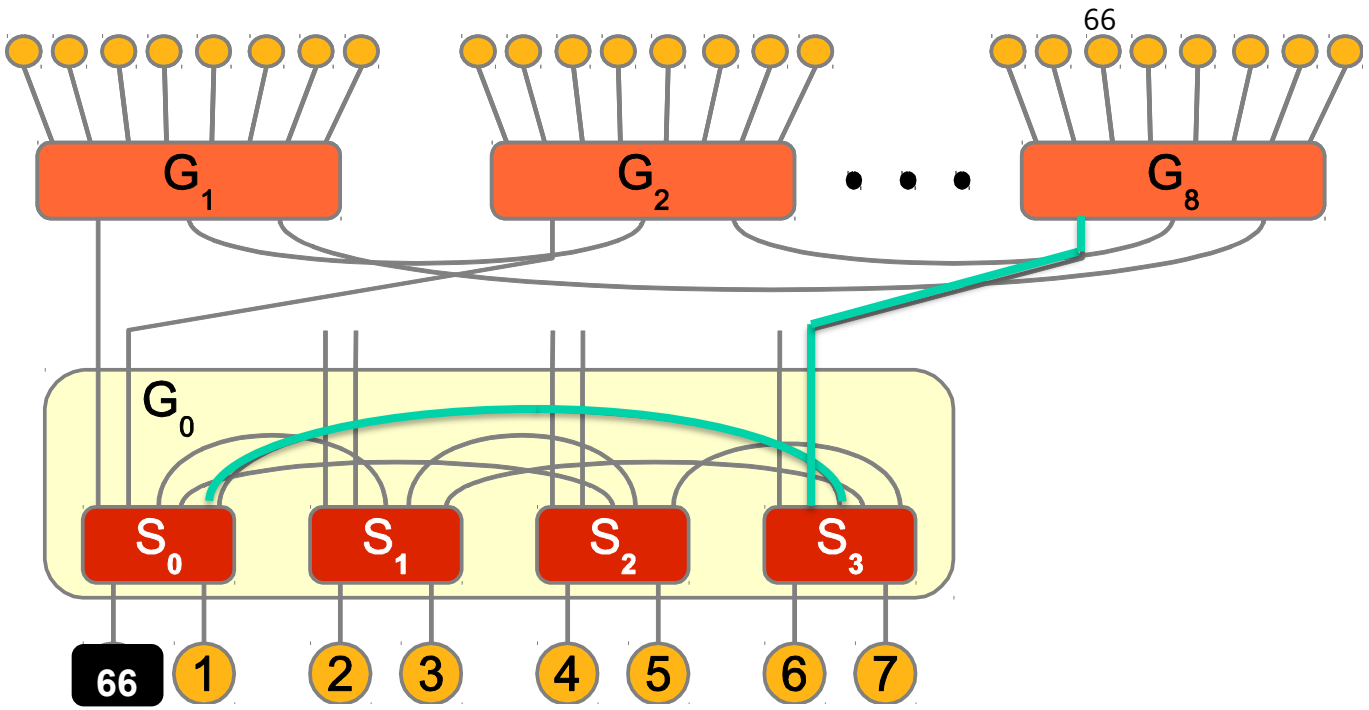
- **Hierarchical** high-performance topology consisting of **a set of groups**, each one composed of several **switches** where **endnodes** are attached.
- Low diameter, path diversity, high scalability, etc.



J. Kim, W. J. Dally, S. Scott, and D. Abts: ***Technology-Driven, Highly-Scalable Dragonfly Topology****. SIGARCH 2008: vol. 36, pp. 77-88*

Background

Dragonfly Minimal-Path Routing

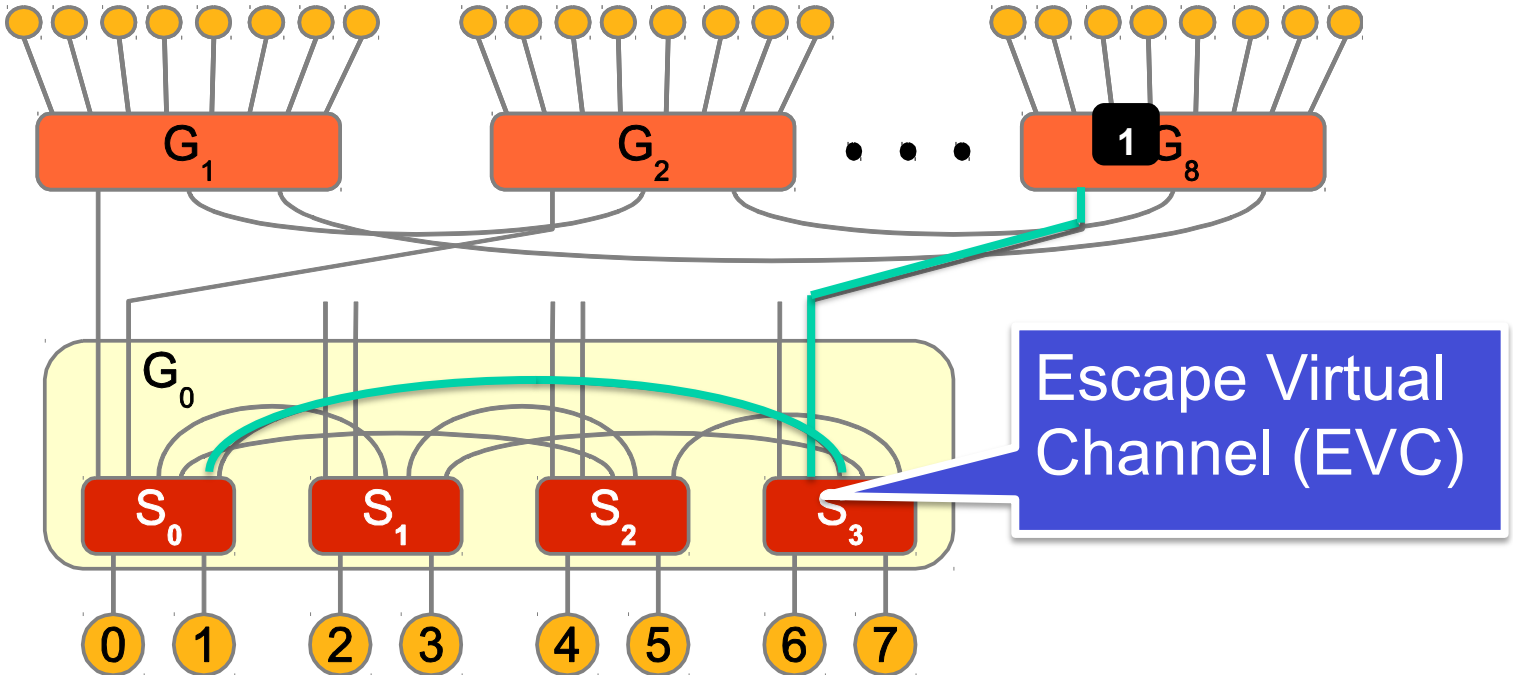


72-node Dragonfly network: a=4, h=2, p=2

J. Kim, W. J. Dally, S. Scott, and D. Abts: ***Technology-Driven, Highly-Scalable Dragonfly Topology****. SIGARCH 2008: vol. 36, pp. 77-88*

Background

Dragonfly Minimal-Path Routing

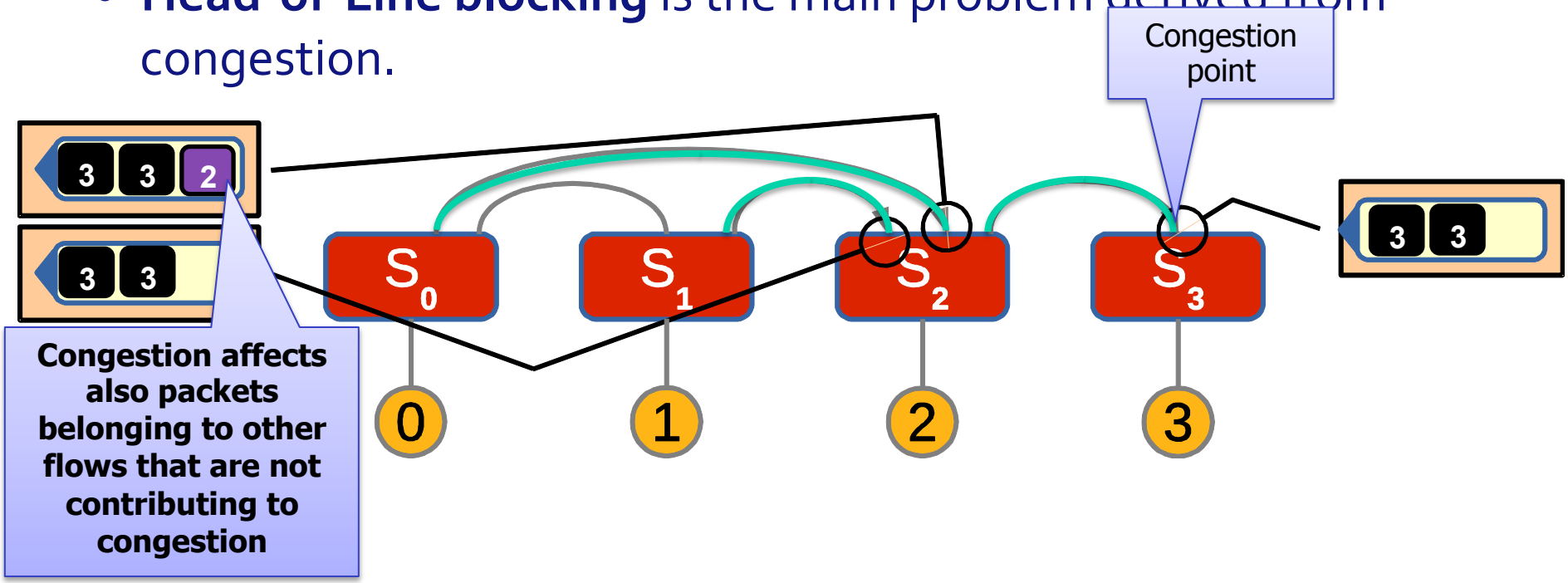


72-node Dragonfly network: a=4, h=2, p=2

J. Kim, W. J. Dally, S. Scott, and D. Abts: ***Technology-Driven, Highly-Scalable Dragonfly Topology****. SIGARCH 2008: vol. 36, pp. 77-88*

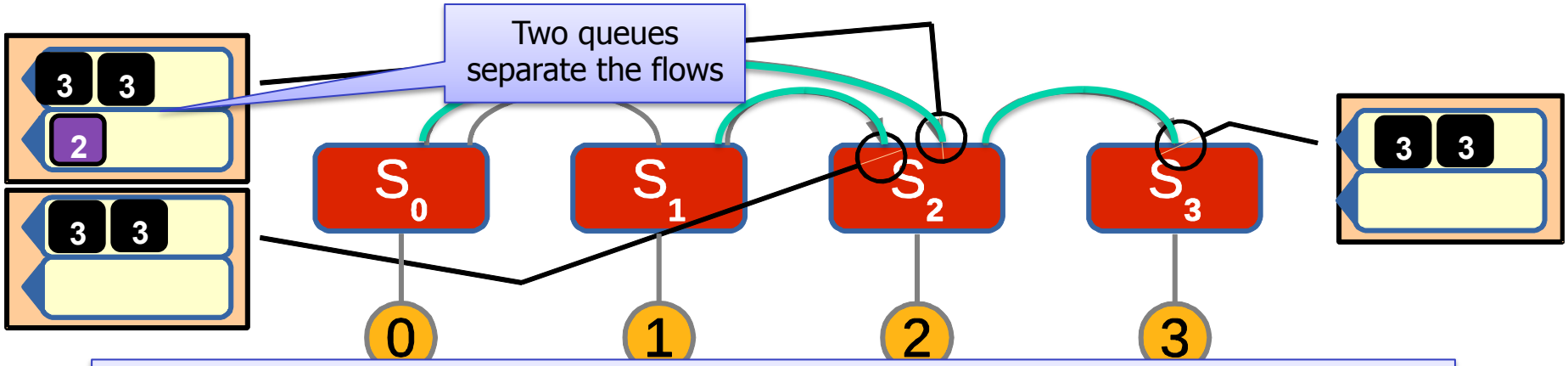
Background
Congestion

- Under congestion situations, network performance may degrade significantly.
- **Head-of-Line blocking** is the main problem derived from congestion.



Background
Queuing Schemes

- Several queues, supporting V**irtual Channels** (VCs), or **Virtual Lanes** (VLs) are used at each port **to separate traffic flows**, reducing the HoL-blocking produced among them.
- A **static criterion** is used to map packets to queues.

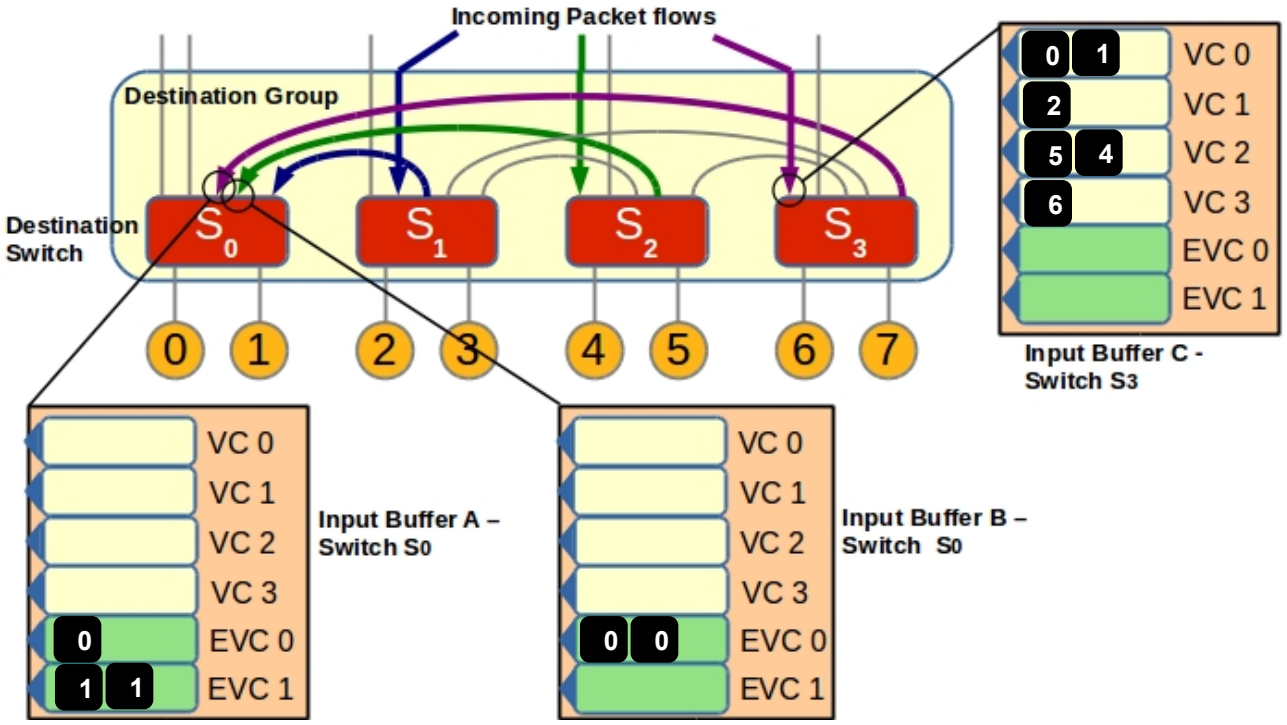


*The most efficient queuing schemes are tailored to a specific **network topology** and a specific **routing algorithm**.*

Background

Hierarchical 2-Level Queuing (H2LQ)

- Queuing scheme tailored to Dragonfly topology with MIN-path routing algorithm.



P. Yebenes, J. Escudero-Sahuquillo, P. J. Garcia and F. J. Quiles, "Efficient Queuing Schemes for HoL-Blocking Reduction in Dragonfly Topologies with Minimal-Path Routing," CLUSTER 2015 IEEE, pp. 817-824.

Background
Quality of Service

- Usually, QoS is based on **separating into different VCs traffic** with different priorities or from different applications.
- Sometimes, VCs priorities are managed by using the *Weighted Round Robin* (WRR) algorithm, which is implemented by a weighted table.
- VCs with higher weight and/or more entries in the table have more priority.

Weighted Table

VC	Weight
0	3
1	3
0	3
2	1

Total Weight VC0 = 6/10 (60%)
Total Weight VC1 = 3/10 (30%)
Total Weight VC2 = 1/10 (10%)

Background

Congestion Management + QoS

- CHADS: Combining HoL-blocking Avoidance and Differentiated Services.
- CHADS defines different **Service-Level Priorities** (SLPs) to identify the priority level of the applications.
- Each SLP is mapped to a **disjoint set of VCs**.
- A queuing scheme is used inside the set of VCs of the same SLP to prevent HoL blocking.
- Higher priority SLPs are mapped with more VCs.

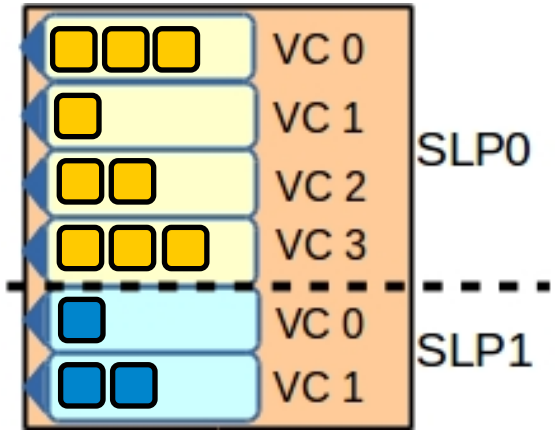
P. Yebenes, J. Escudero-Sahuquillo, C. Gomez, P. J. Garcia F.J. Alfaro, and F. J. Quiles, "Combining HoL-blocking avoidance and differentiated services in high-speed interconnects," HiPC 2014

Background

Congestion Management + QoS

- CHADS: Combining HoL-blocking Avoidance and Differentiated Services.

SLP	VC	Weigth
SLP0	0	2
	1	2
	2	2
	3	2
SLP1	3	1
	4	1



P. Yebenes, J. Escudero-Sahuquillo, C. Gomez, P. J. Garcia F.J. Alfaro, and F. J. Quiles, "Combining HoL-blocking avoidance and differentiated services in high-speed interconnects," HiPC 2014

Outline

- Motivation
- Background
- **Proposals description**
- Evaluation
- Conclusions

Proposals

Basic Ideas

- Adapting CHADS to **dragonfly networks**.
- **SLPs** are also considered, each one assigned with different VCs.
- Congestion management inside each SLP by means of **H2LQ**.
- QoS provision by configuring **Weighted Tables**.
- **Deadlock freedom** using Escape VCs managed by different policies for configuring Escape Virtual Networks (EVNs):
  - Exclusive Escape Virtual Network (EEVN)
  - Common Exclusive Virtual Network (CEVN).

Proposals

Exclusive Escape Virtual Network (EEVN)

- Each SLP has a Standard Virtual Network (SVN) and an EVN.
- Packets use the SVN by default but are assigned to the EVN for avoiding deadlocks.
- Packets from different SLPs never interact.

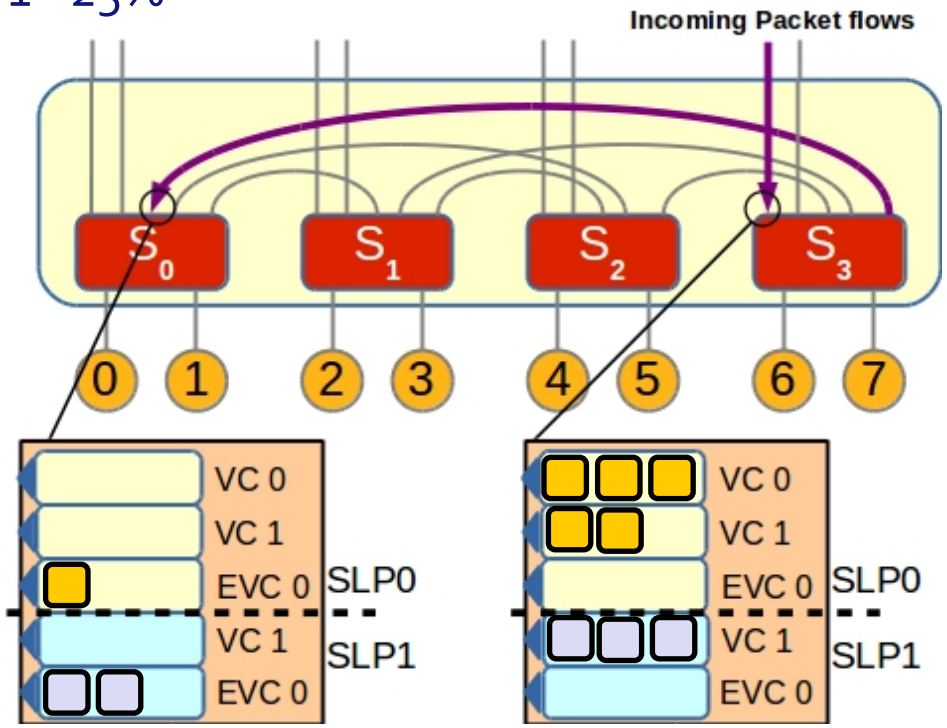
Proposals

Exclusive Escape Virtual Network (EEVN)

- 2 SLPs
  - Total Weight: SLP 0 = 75%, SLP1 =25%
- 5 VCs

Weighted Table

SLP	VN	VC	Weigth
SLP0	SVN	0	2
		1	2
SLP0	EVN	2	2
SlP1	SVN	3	1
SLP1	EVN	4	1



Proposals

Common Escape Virtual Network (CEVN)

- Each SLP has a SVN.
- There is a single Common Escape VN (CEVN) shared by all the SLPs.
- Packets from different SLPs share VCs when they are in the CEVN.

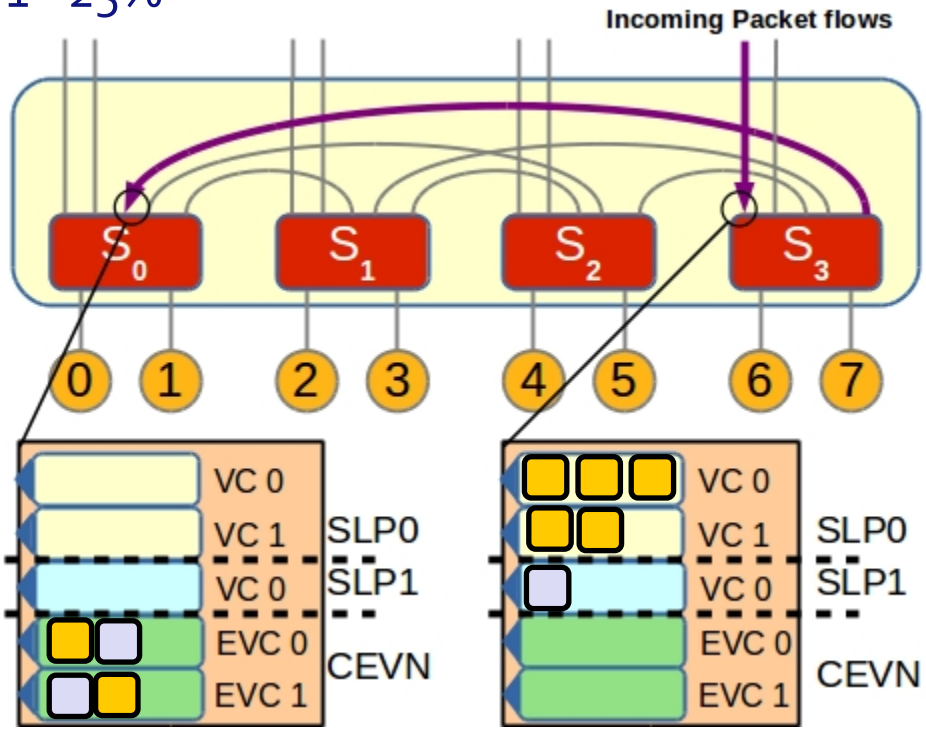
Proposals

Common Escape Virtual Network (CEVN)

- 2 SLPs
  - Total Weight: SLP 0 = 75%, SLP1 =25%
- 5 VCs

Weighted Table

SLP	VN	VC	Weigth
SLP0	SVN	0	3
		1	3
SLP1	SVN	2	2
SLP2	CEVN	3	1
		4	1



Outline

- Motivation
- Background
- Proposals description
- **Evaluation**
- Conclusions

Evaluation

Simulation Tool

OMNeT++-based simulator:

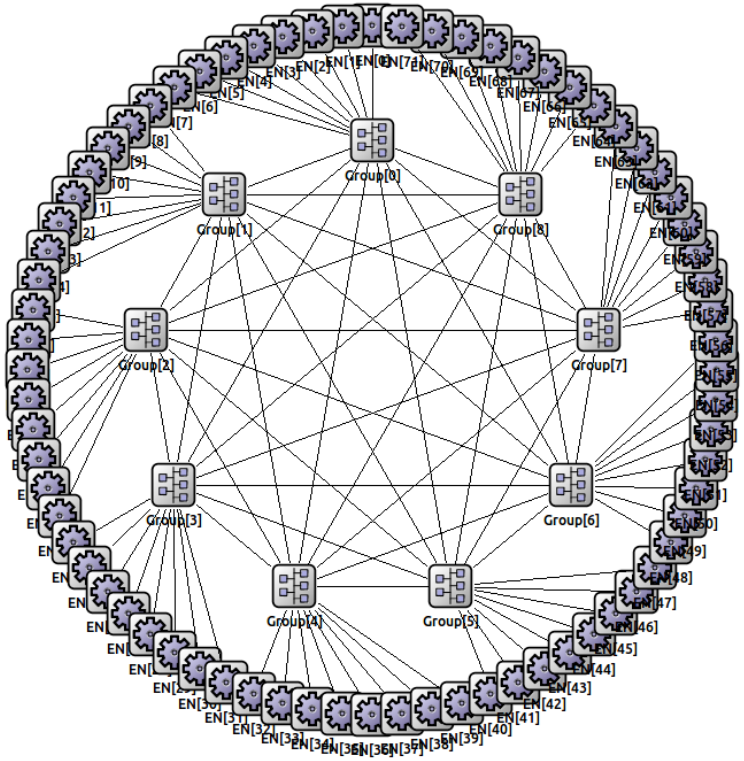
- Different topologies.
- Different routing algorithms.
- Different queuing schemes.
- Quality of Service support.

Pedro Yébenes, Jesús Escudero-Sahuquillo, Pedro J. García, Francisco J. Quiles: ***Towards Modeling Interconnection Networks of Exascale Systems with OMNeT++****. PDP 2013*

Evaluation

Network Configurations

- 4096-node dragonfly network (a=12, h=6, p=6).



Evaluation

Traffic Patterns

- **3 applications**, each one assigned with a different SLP (SLP0, SLP1, SLP2), generating **synthetic traffic** at a rate of 70% of the link bandwidth with two traffic patterns:
  - Uniform traffic.
  - Zipf traffic:
    - Models traffic patterns with preferred destinations.
    - Traffic pattern similar to the ones produced by the collective communication schemes.

L. Breslau, Pei Cao, Li Fan, G. Phillips, and S. Shenker: ***Web caching and Zipf-like distributions: evidence and implications****. INFOCOM '99: 126-134 vol.1*

Evaluation

EEVN VCs Configurations

- Total weight per SLP in the Weighted Tables:
  - SLP0: 60%, SLP1: 30%, SL2: 10%
- Number of VCs for each SLP in EEVN (*EEVN-X* where *X* is the total number of required VCs).

Name	SLP0		SLP1		SLP2	
	#SVCs	#EVCs	#SVCs	#EVCs	#SVCs	#EVCs
EEVN-54	12	6	12	6	12	6
EEVN-15	6	1	4	1	2	1
EEVN-8	3	1	1	1	1	1
EEVN-6	1	1	1	1	1	1

Evaluation

CEVN VCs Configurations

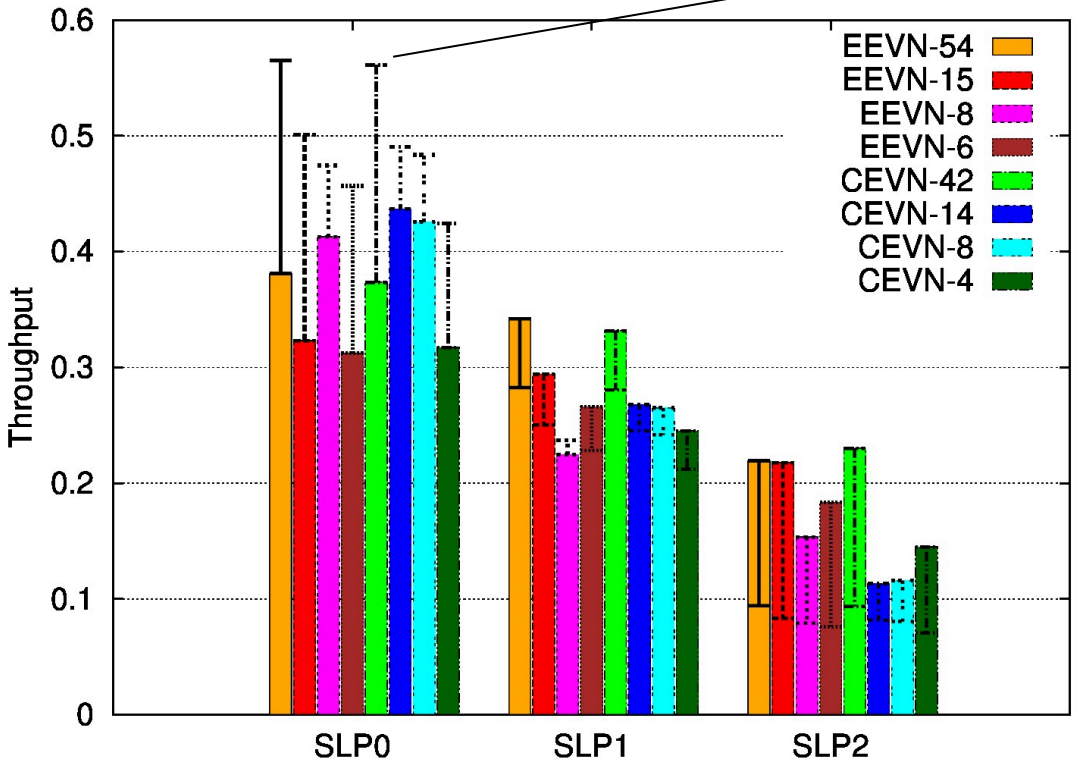
- Total weight per SLP in the Weighted Tables:
  - SLP0: 60%, SLP1: 30%, SL2: 10%
- Number of VCs for each SLP in CEVN (*CEVN-X* where *X* is the total number of required VCs).

Name	SLP0	SLP1	SLP2	CEVN
	#SVC	#SVC	#SVC	#EVC
CEVN-42	12	12	12	6
CEVN-14	6	4	2	2
CEVN-8	3	2	1	2
CEVN-4	1	1	1	1

Evaluation

Results Uniform SLP0=70%, SLP1=70%, SLP2=70%

- Metric: normalized throughput.

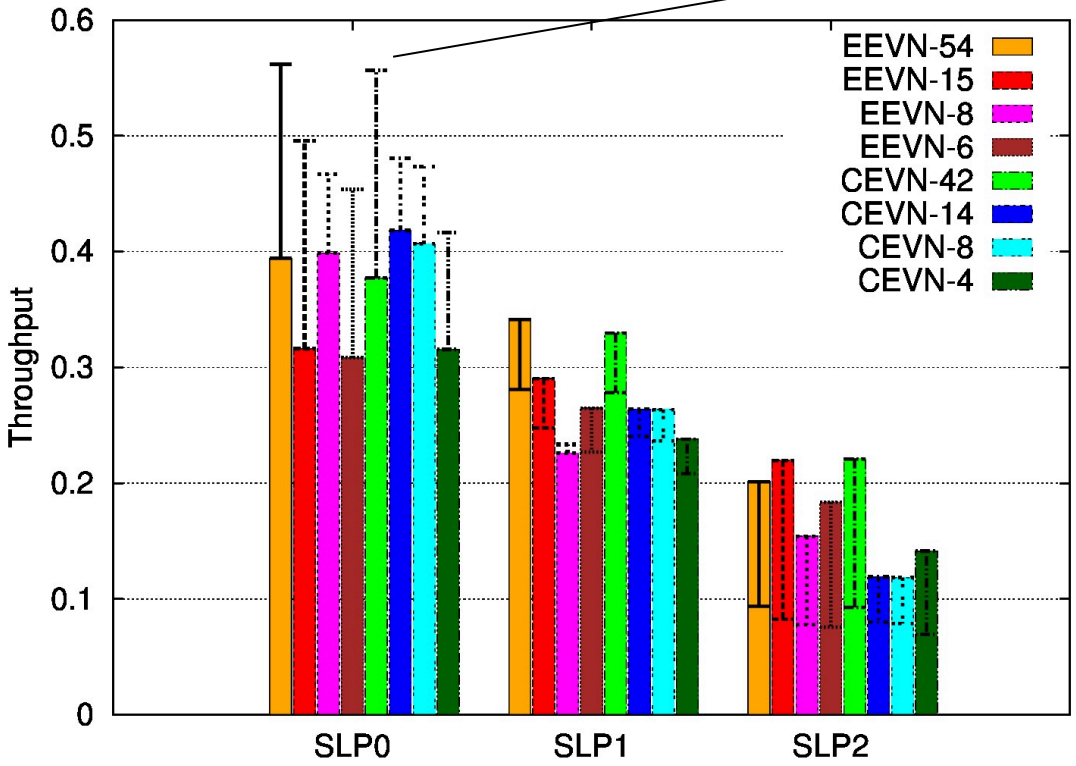


Distortion levels show the expected theoretical throughput for a given SLP, according to the network throughput and the WT configuration.

Evaluation

Results Zipf SLP0=70%, SLP1=70%, SLP2=70%

- Metric: normalized throughput.



Distortion levels show the expected theoretical throughput for a given SLP, according to the network throughput and the WT configuration.

Outline

- Motivation
- Proposal description
- Evaluation
- **Conclusions**

Conclusions

Advantages

- CHADS technique has been updated for dragonfly networks.
- Differentiated services at network level, congestion management and deadlock freedom can be provided at the same time by means of EEVN and CEVN approaches.
- In general, CEVN is better than EEVN.
- The number of VCs configured has to be tightly with the weight configured in the Weighted Tables.

Conclusions

Future directions

- Analyzing these approaches with other routing algorithms suited to InfiniBand.
- Testing these approaches with other traffic patterns: application traces, adversarial, blocking collectives, etc.
- Exploring other configurations for Dragonfly networks.
- Exploring other approaches to better populate the weighted tables.

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS

UCLM
UNIVERSIDAD DE CASTILLA-LA MANCHA

Providing Differentiated Services, Congestion Management, and Deadlock Freedom in Dragonfly Networks

Pedro Yébenes, Jesús Escudero-Sahuquillo, Pedro J. García,

Francisco J. Alfaro, Francisco J. Quiles

University of Castilla – La Mancha