

# Comparing the latency performance of the DTable and DRR schedulers

Raúl Martínez

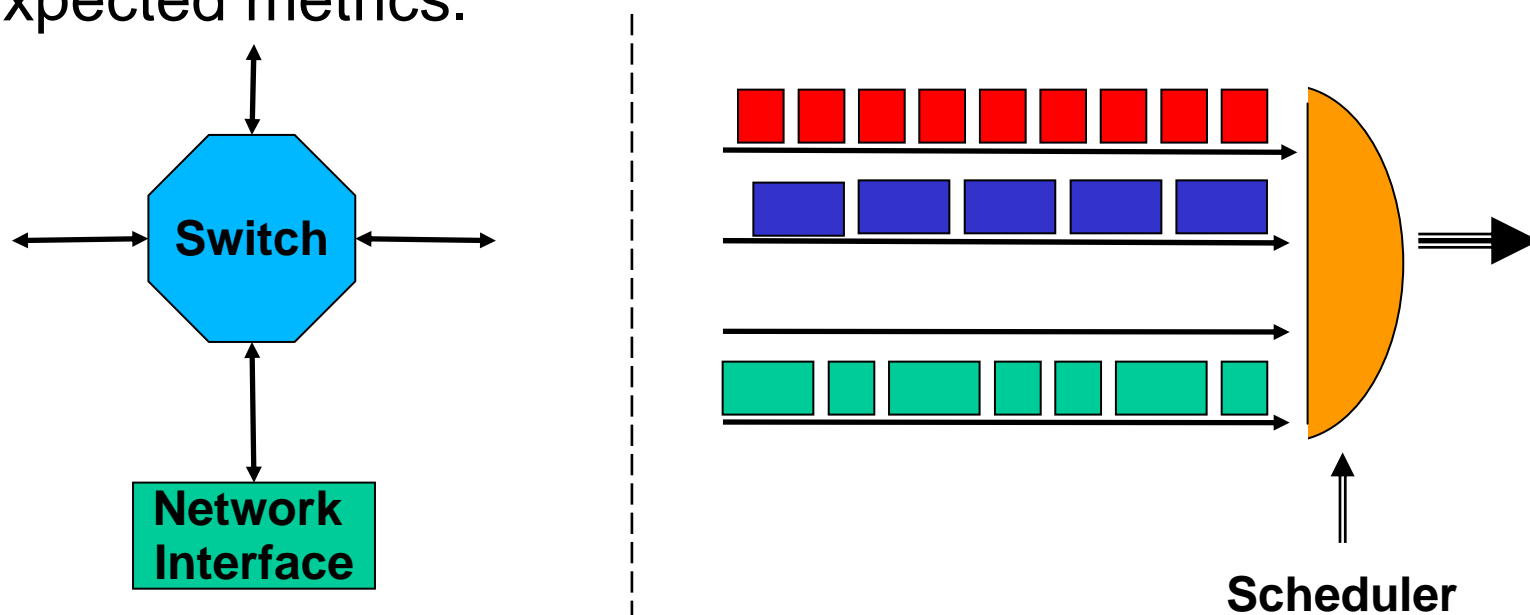
Francisco J. Alfaro

José L. Sánchez



# Motivation

- Nowadays, high-performance networks are required to carry traffic with different **Quality of Service (QoS)** requirements.
- A key component for networks with QoS support is the **egress link scheduling algorithm**, which selects the next packet to be transmitted on the basis of some expected metrics.



# Motivation

- An ideal scheduling algorithm should possess the following properties:
  - **Good end-to-end latency:** In order to provide flows with their latency and jitter requirements.
  - **Low complexity:** In order to require a small silicon area and being able to calculate the next packet to be transmitted in the short time required by a high performance network.

# Motivation

- An ideal scheduling algorithm should possess the following properties:
  - **Good end-to-end latency:** In order to provide flows with their latency and jitter requirements.
  - **Low complexity:** In order to require a small silicon area and being able to calculate the next packet to be transmitted in the short time required by a high performance network.
- The design of a traffic scheduling algorithm involves an inevitable trade-off among these properties.
- Several scheduling algorithms with different properties have been proposed.
  - Sorted-priority algorithms
    - Weighted Fair Queuing
    - Self-Clocked Fair Queuing



-Very low latency  
-Very high complexity

# Motivation

- Deficit Round Robin



-Very high latency  
-Very low complexity

# Motivation

- Deficit Round Robin



-Very high latency  
-Very low complexity

- Table-based schedulers



-Low latency  
-Low complexity

- Advanced Switching
- InfiniBand

# Motivation

- Deficit Round Robin



**-Very high latency**  
**-Very low complexity**

- Table-based schedulers



**-Low latency**  
**-Low complexity**

- Advanced Switching
- InfiniBand



**-Do not work properly with variable packet size.**  
**-Bounding between the bandwidth and latency assignments.**

# Motivation

- Deficit Round Robin



**-Very high latency**  
**-Very low complexity**

- Table-based schedulers



**-Low latency**  
**-Low complexity**

- Advanced Switching
- InfiniBand



**-Do not work properly with variable packet size.**  
**-Bounding between the bandwidth and latency assignments.**

- **Deficit Table (DTable)**

- Decoupling configuration methodology

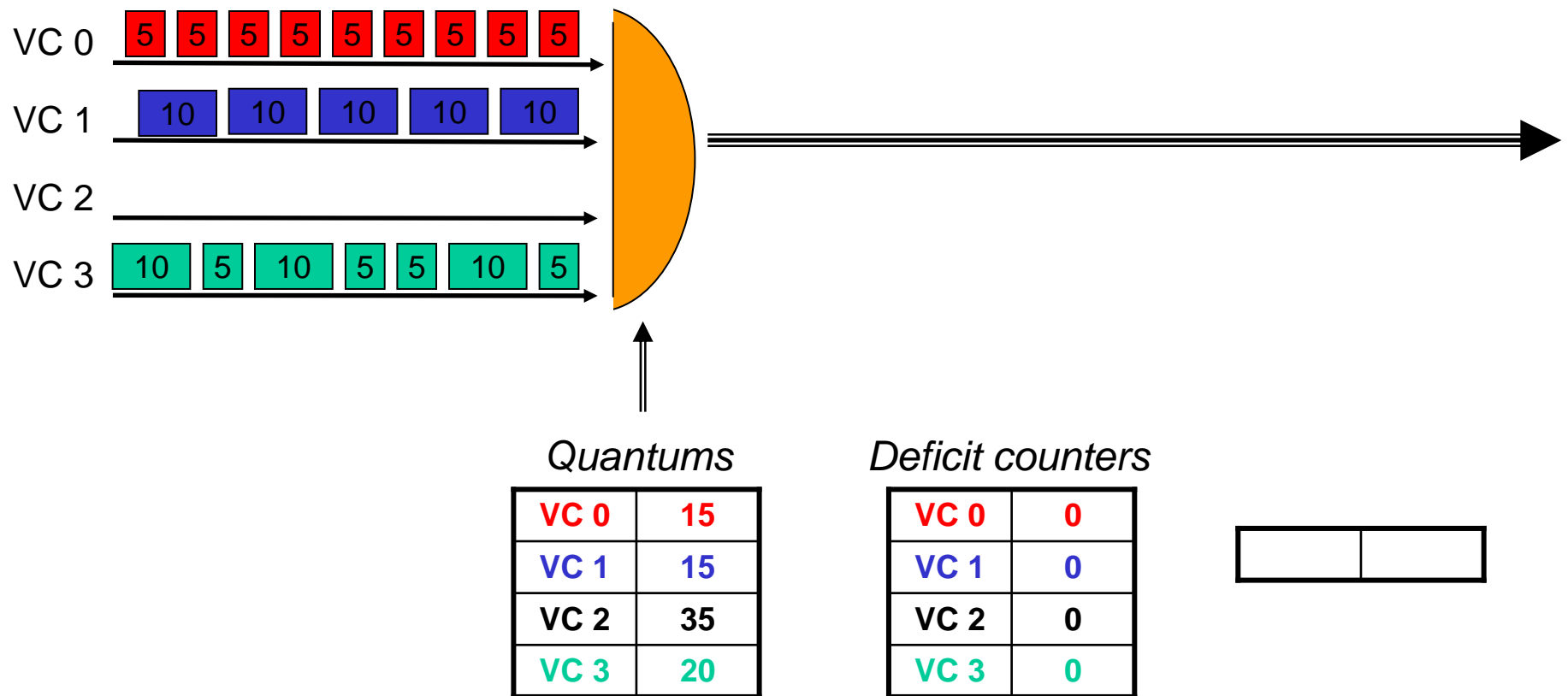
# Outline

- The Deficit Round Robin (DRR) scheduler
- The Deficit Table (DTable) scheduler
  - The DTable scheduling mechanism
  - Configuring the DTable scheduler
- Performance evaluation
- Conclusions

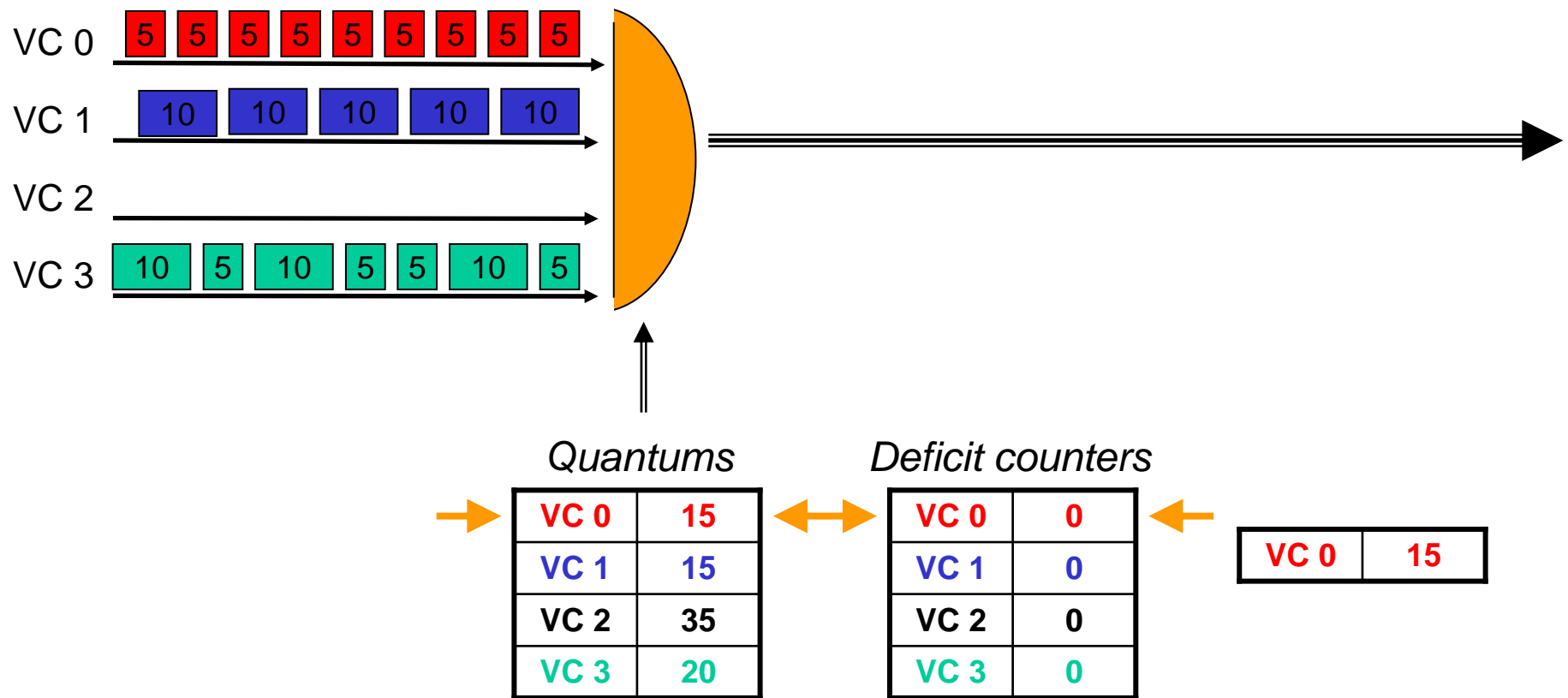
# Outline

- **The Deficit Round Robin (DRR) scheduler**
- The Deficit Table (DTable) scheduler
  - The DTable scheduling mechanism
  - Configuring the DTable scheduler
- Performance evaluation
- Conclusions

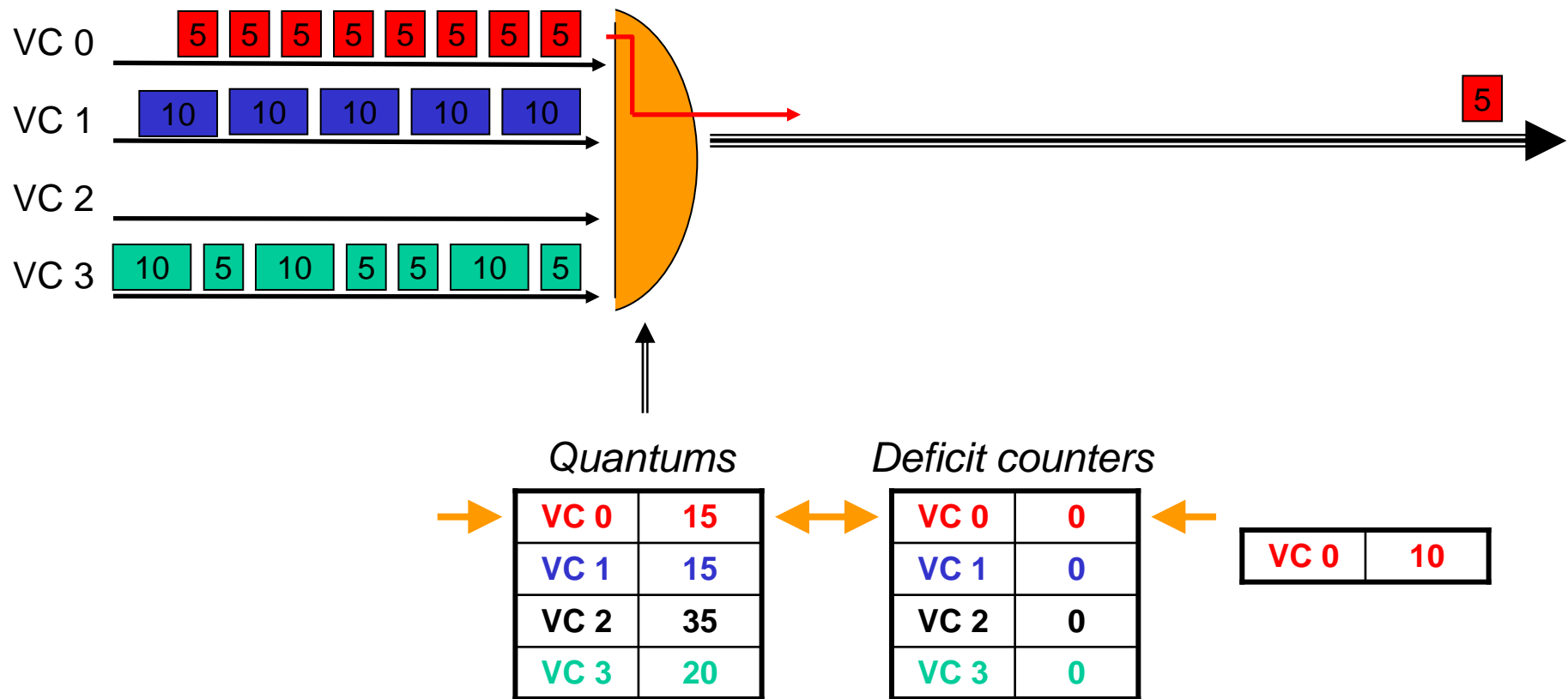
# Deficit Round Robin (DRR)



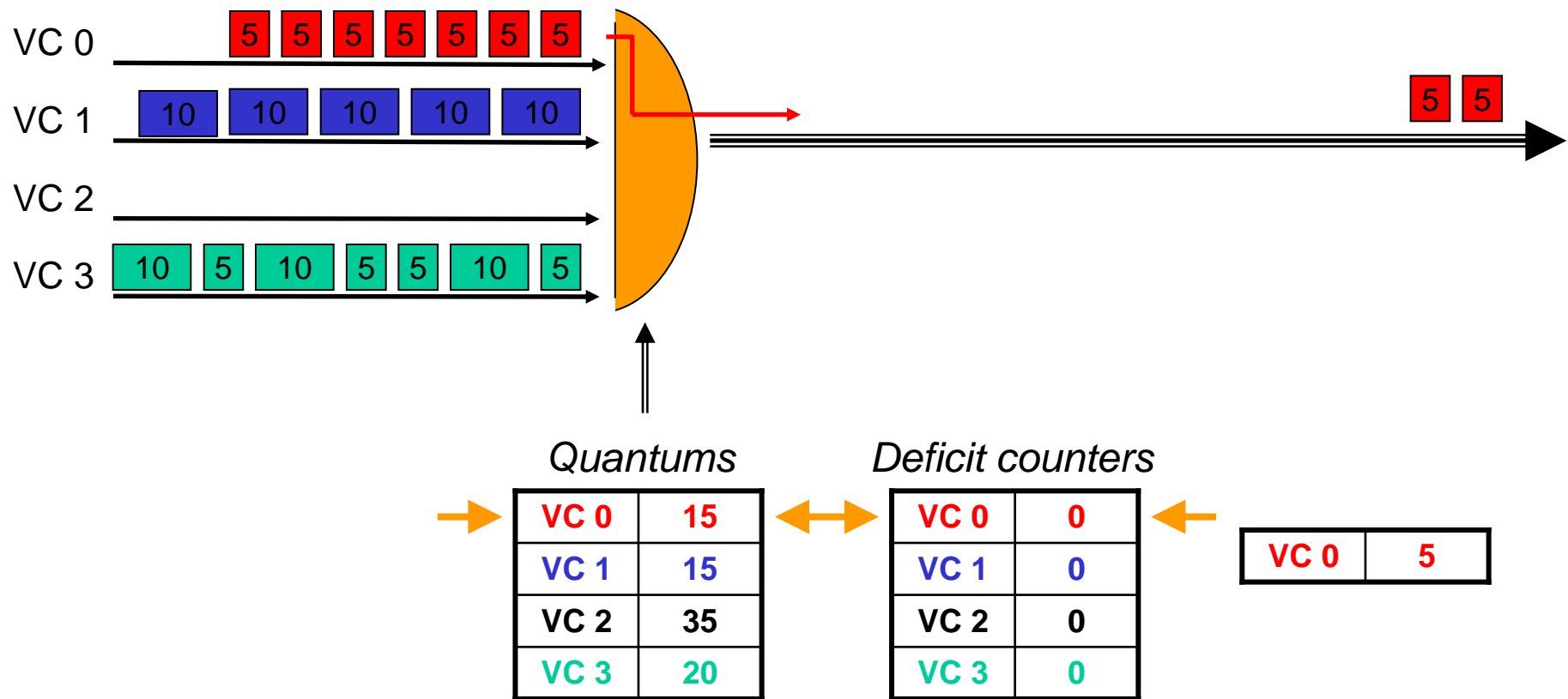
# Deficit Round Robin (DRR)



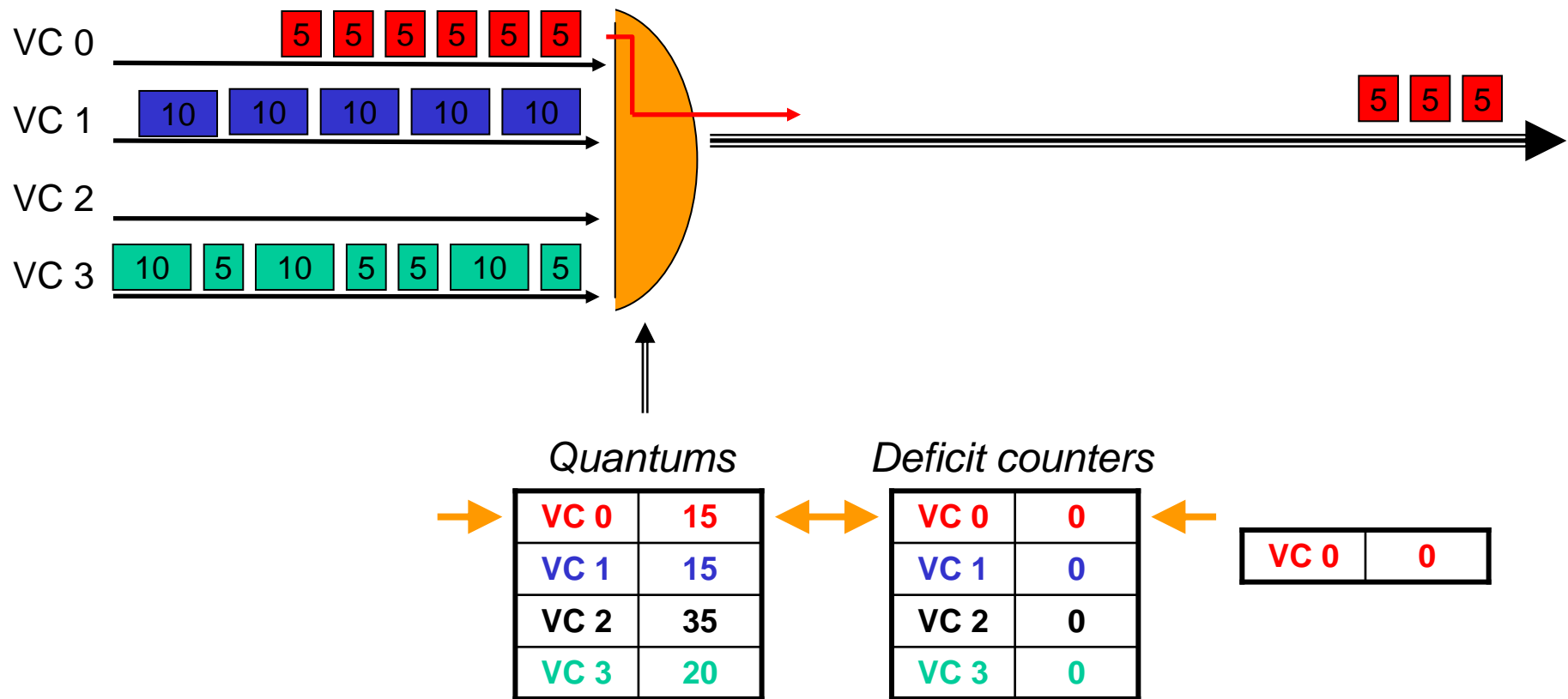
# Deficit Round Robin (DRR)



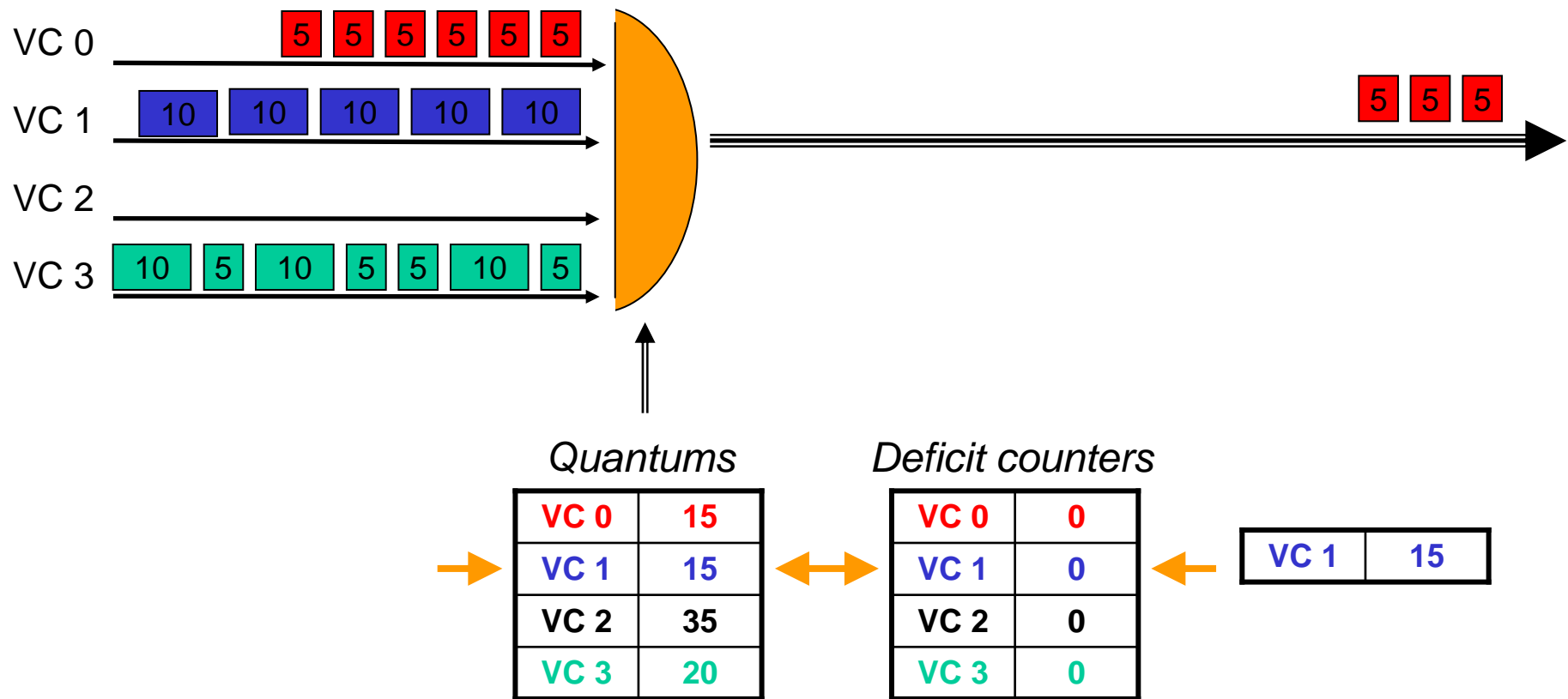
# Deficit Round Robin (DRR)



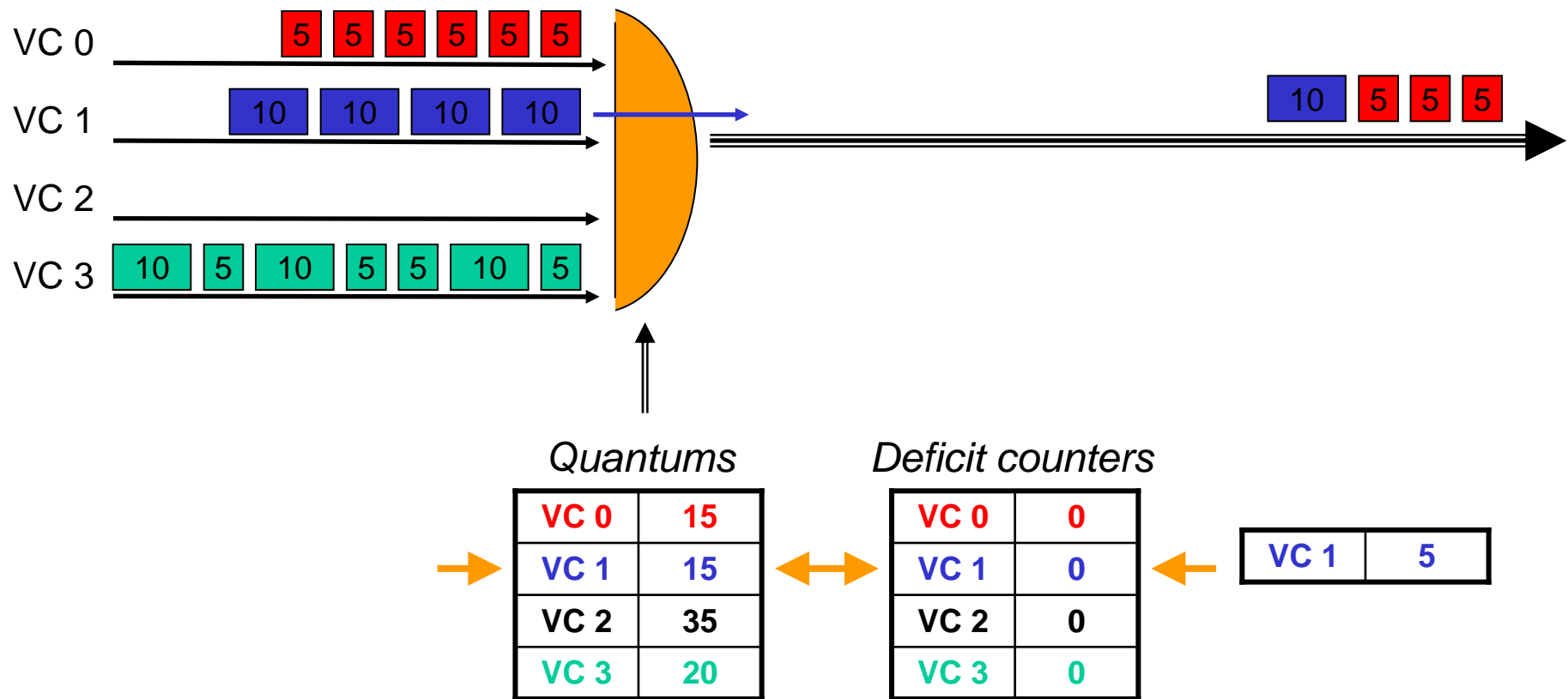
# Deficit Round Robin (DRR)



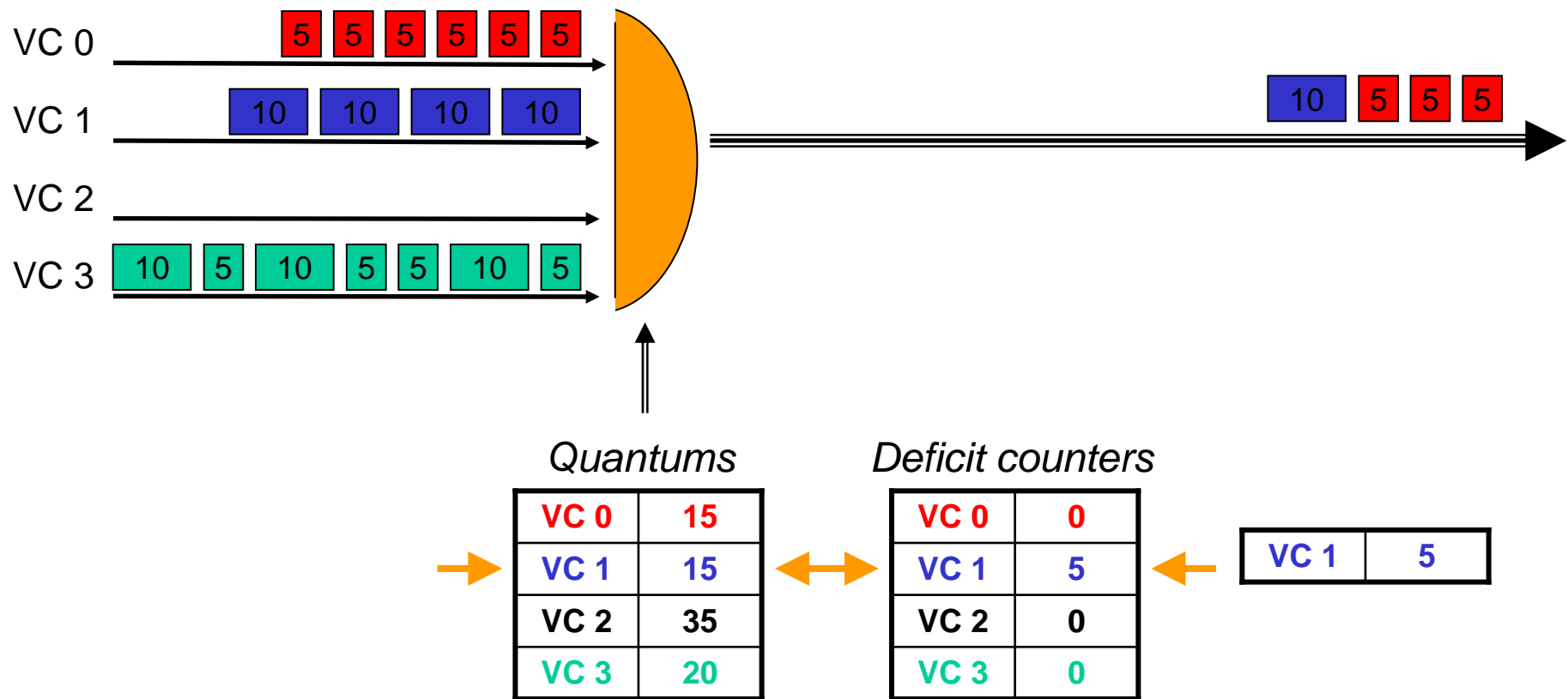
# Deficit Round Robin (DRR)



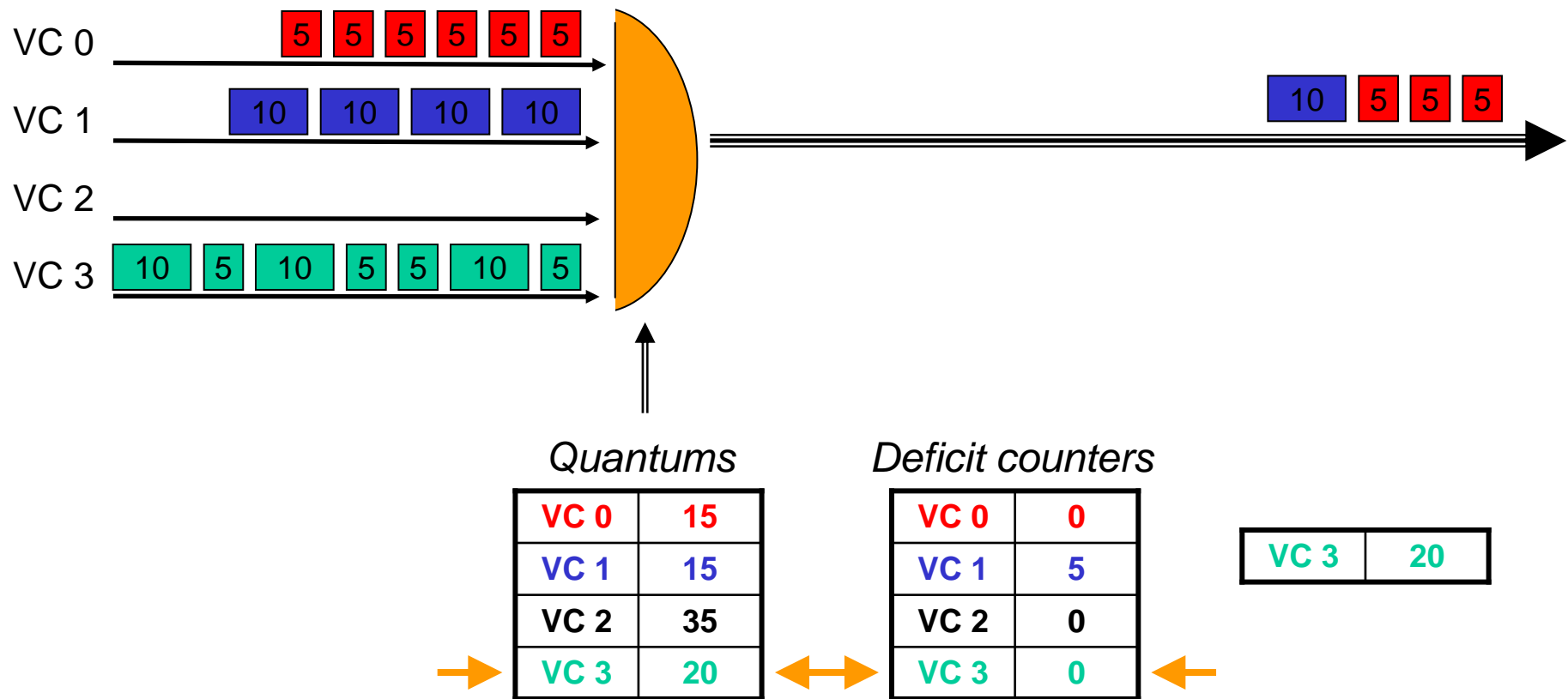
# Deficit Round Robin (DRR)



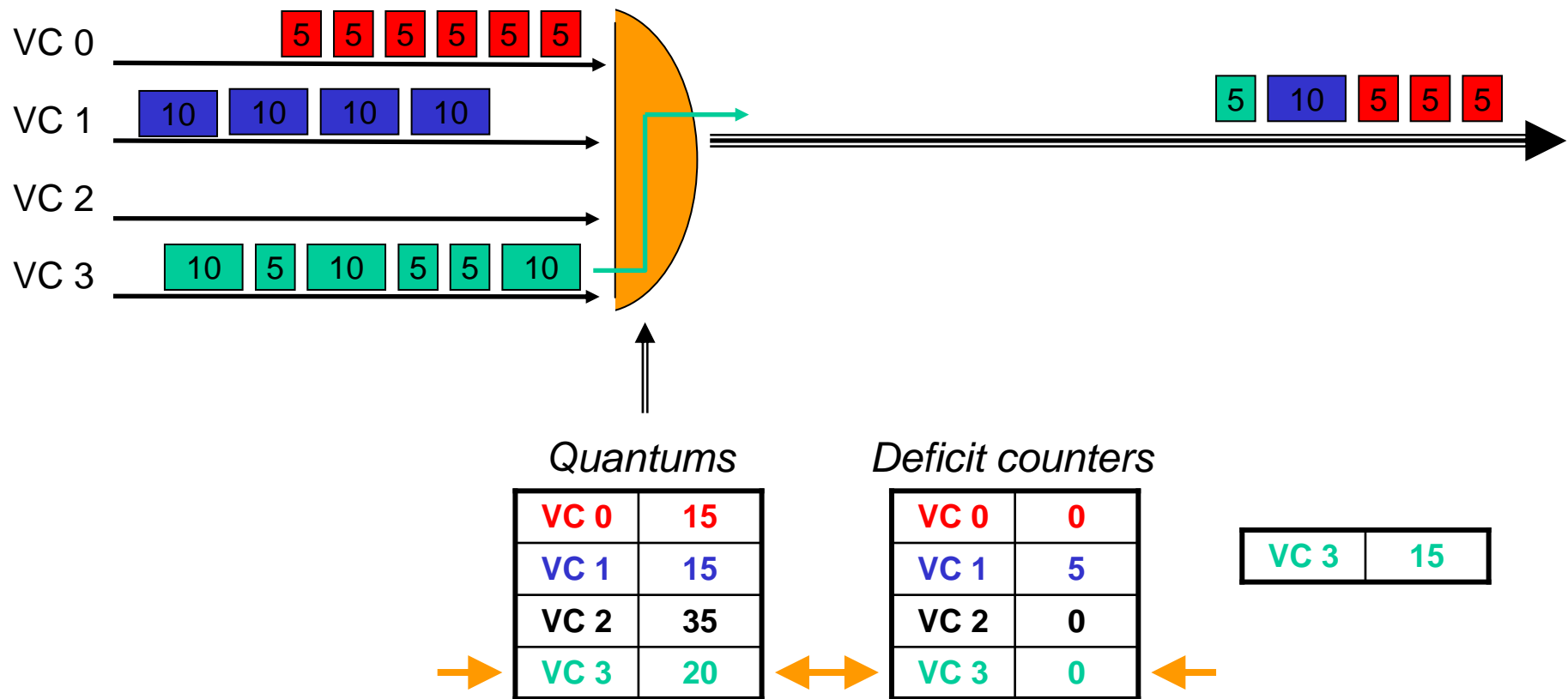
# Deficit Round Robin (DRR)



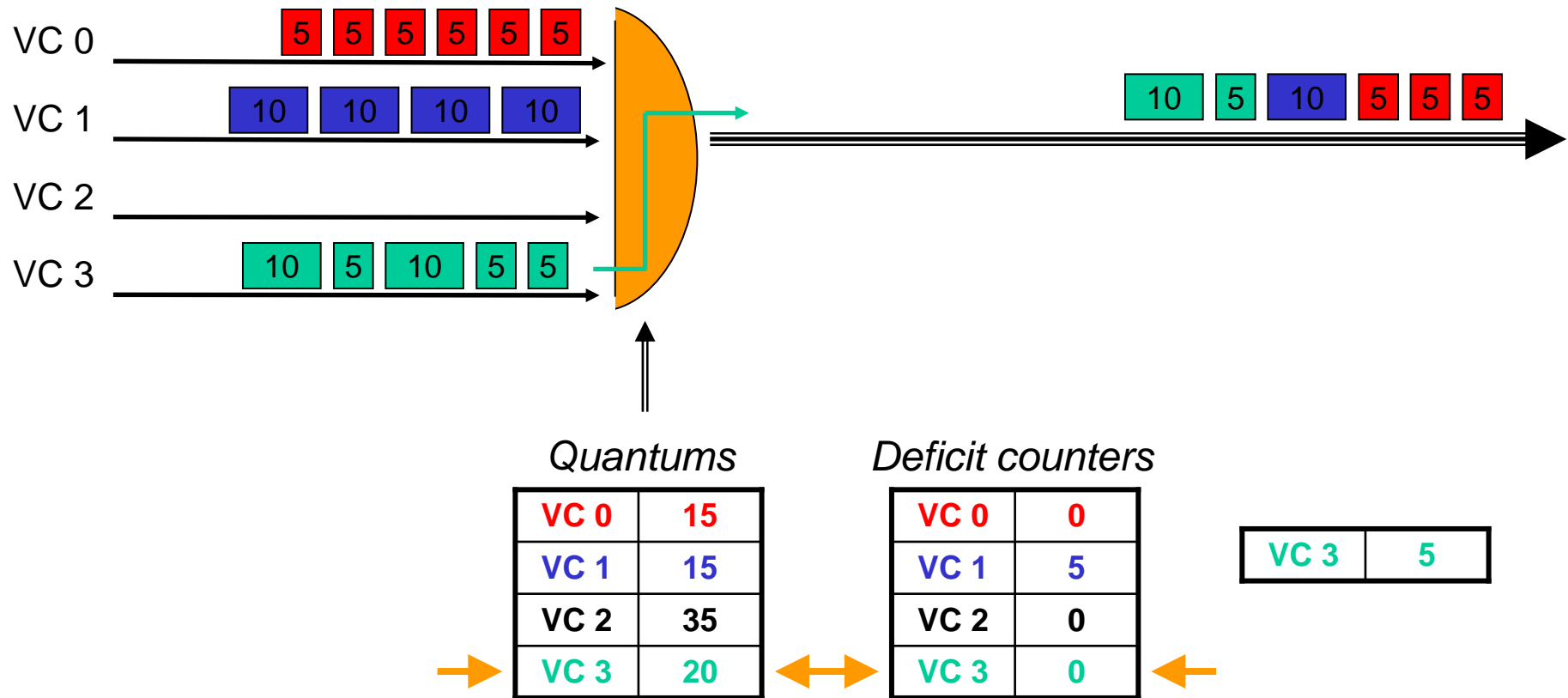
# Deficit Round Robin (DRR)



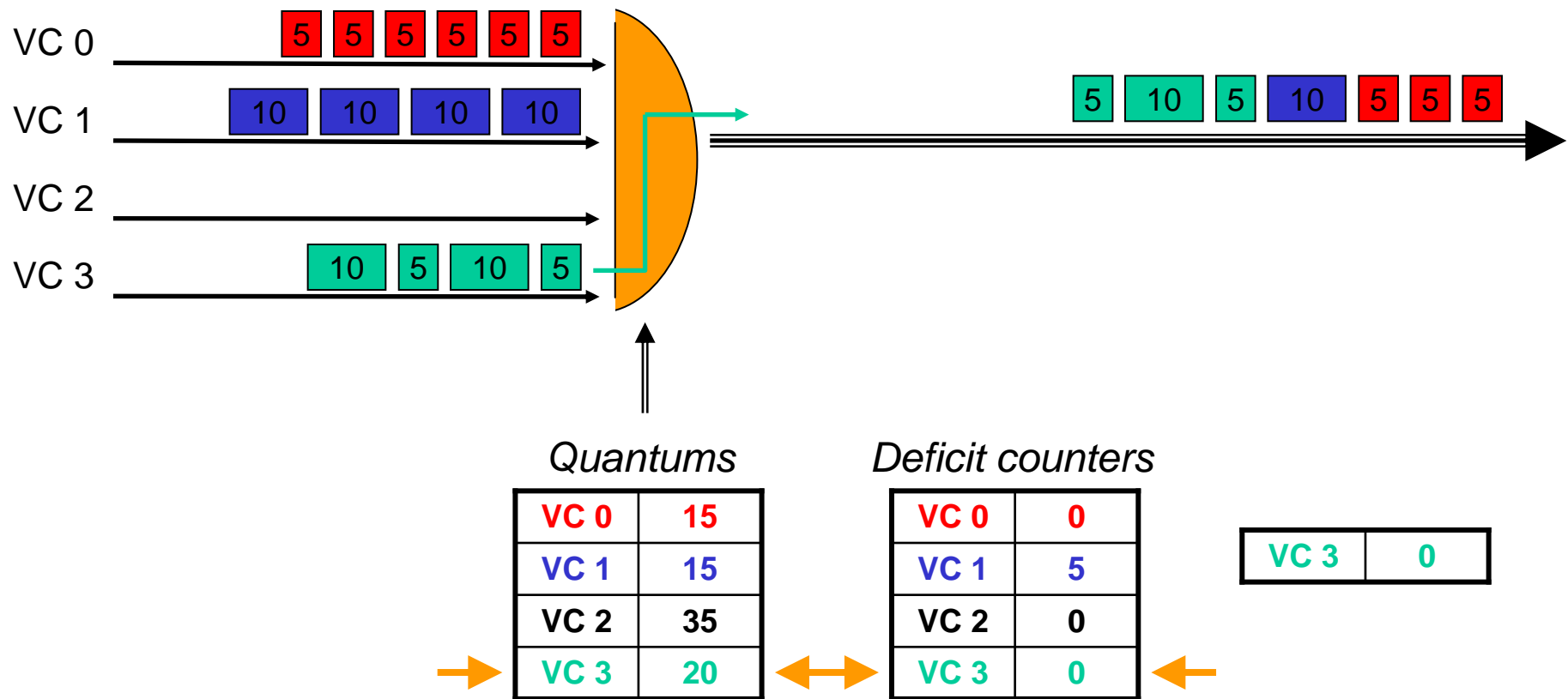
# Deficit Round Robin (DRR)



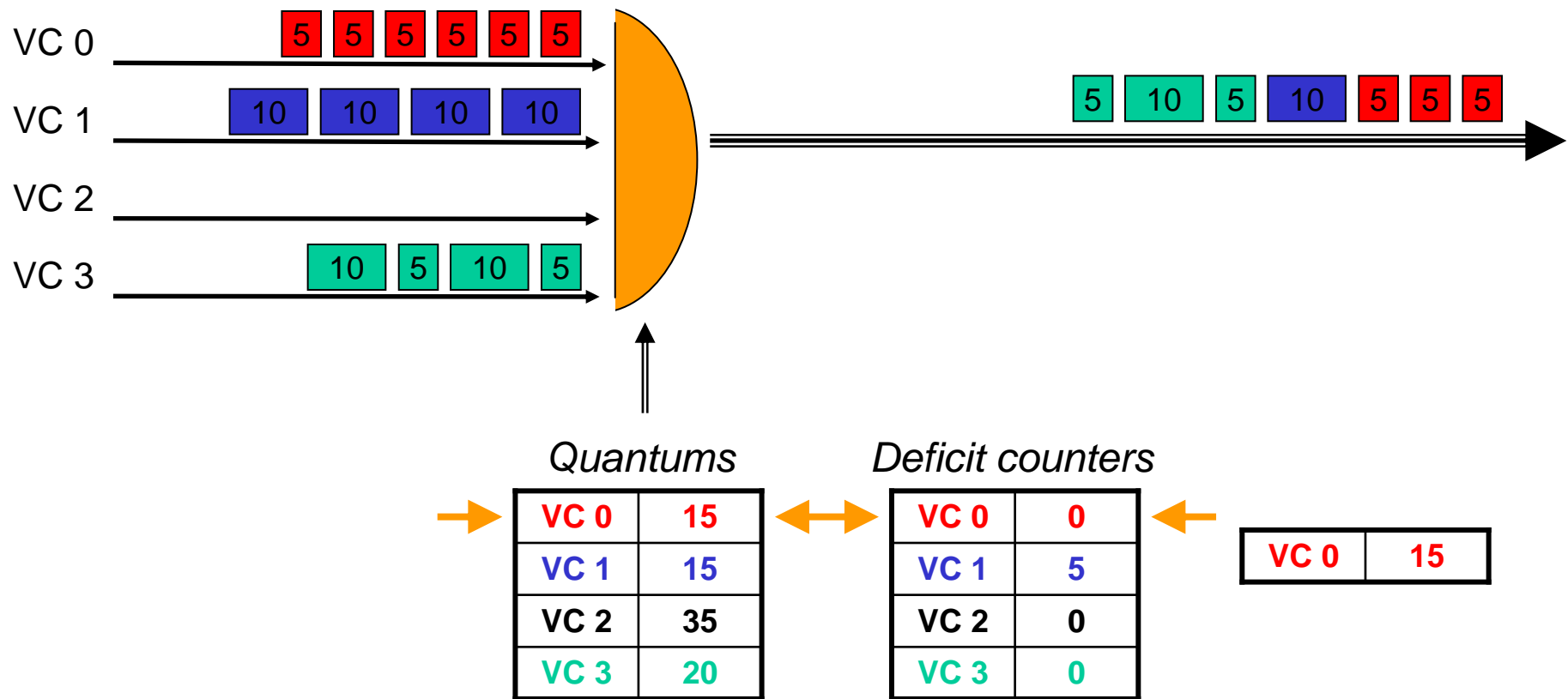
# Deficit Round Robin (DRR)



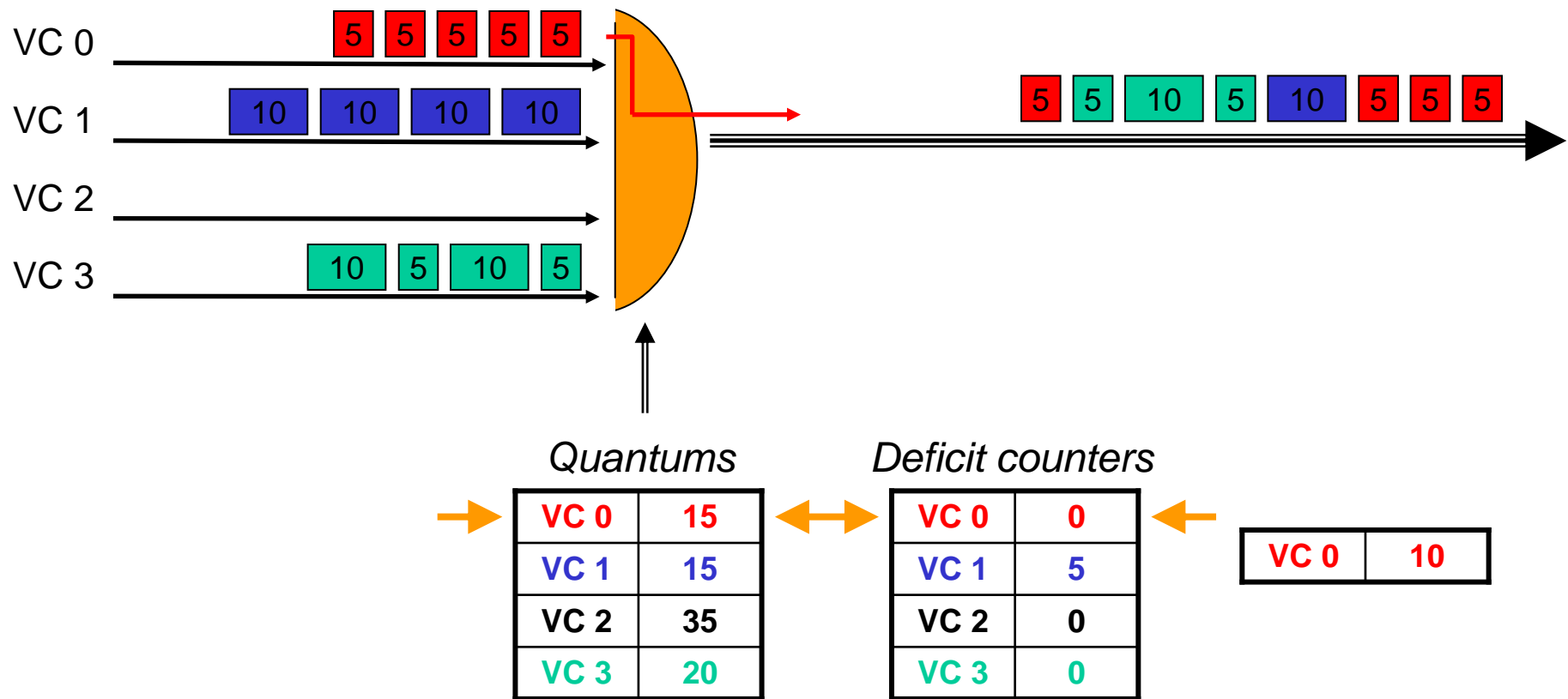
# Deficit Round Robin (DRR)



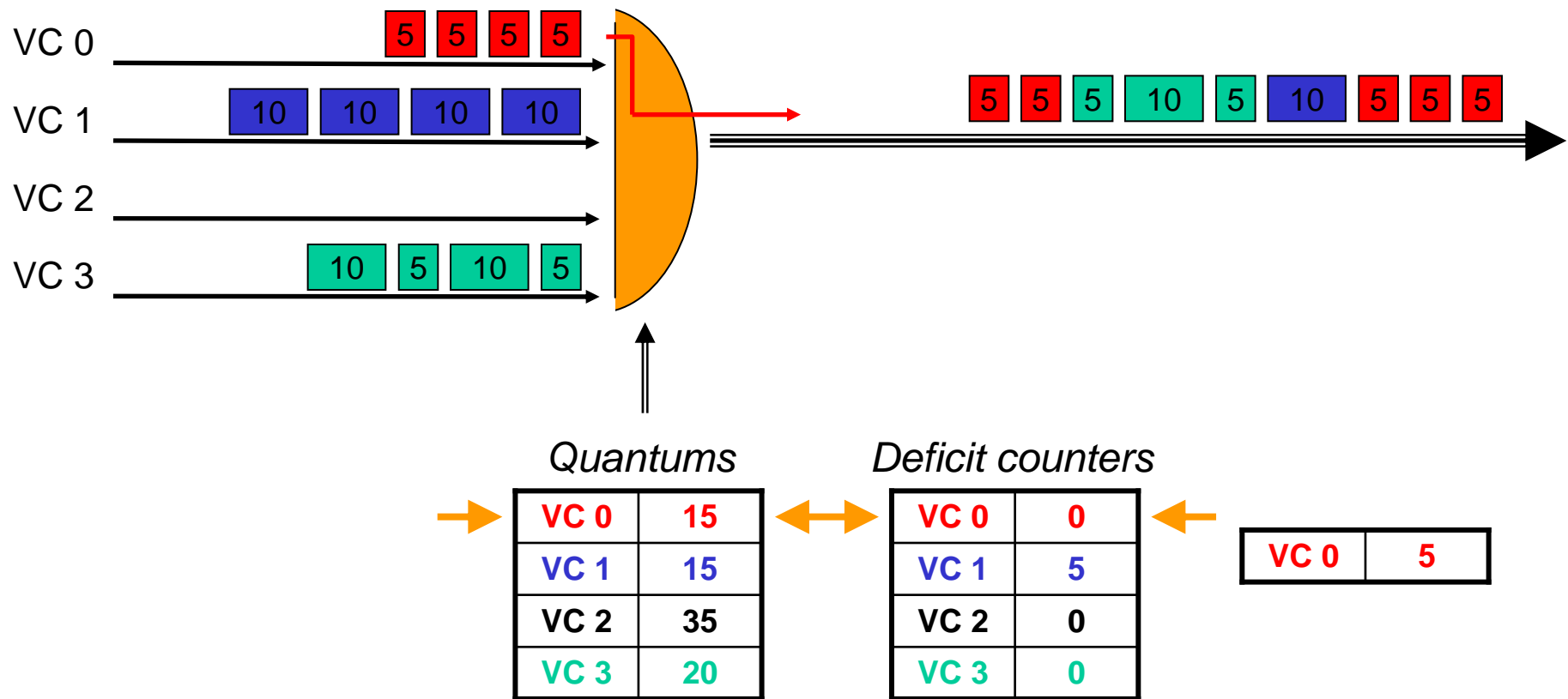
# Deficit Round Robin (DRR)



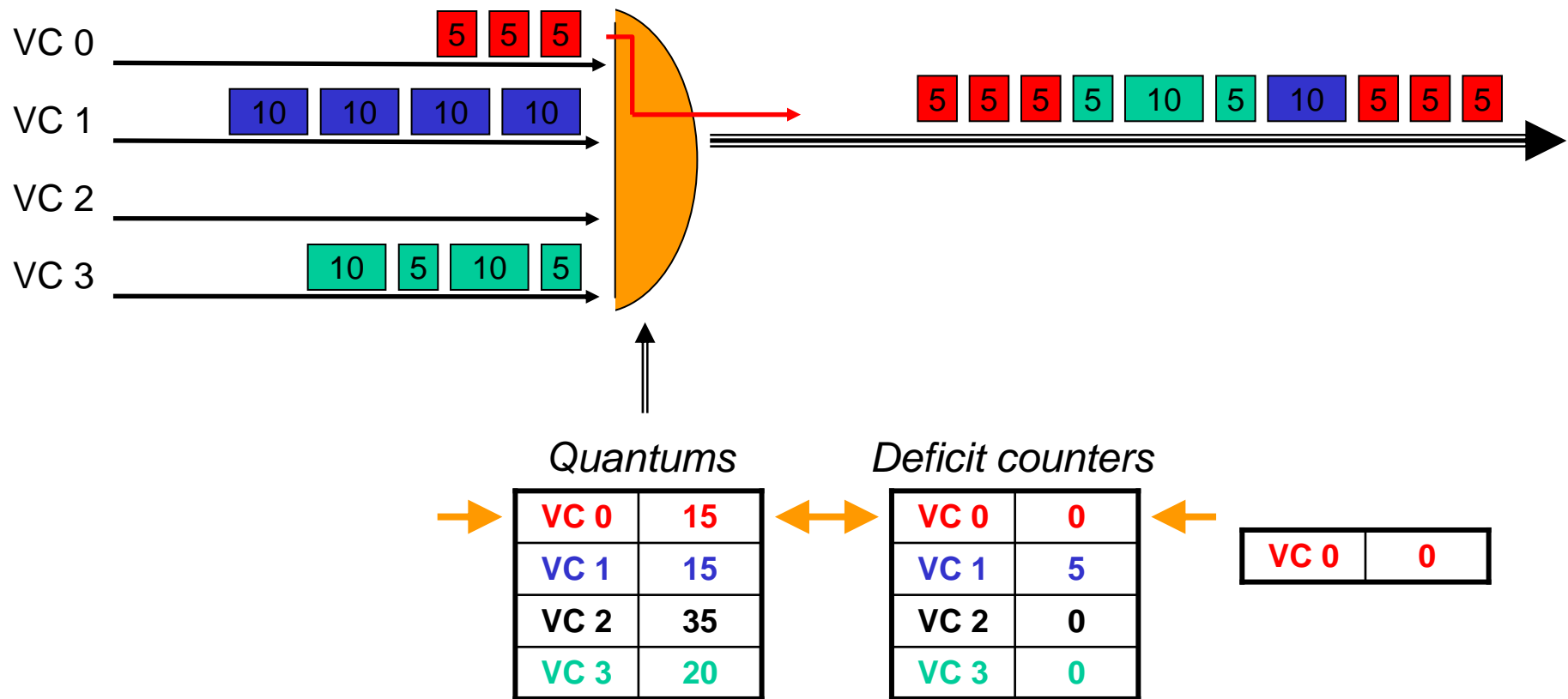
# Deficit Round Robin (DRR)



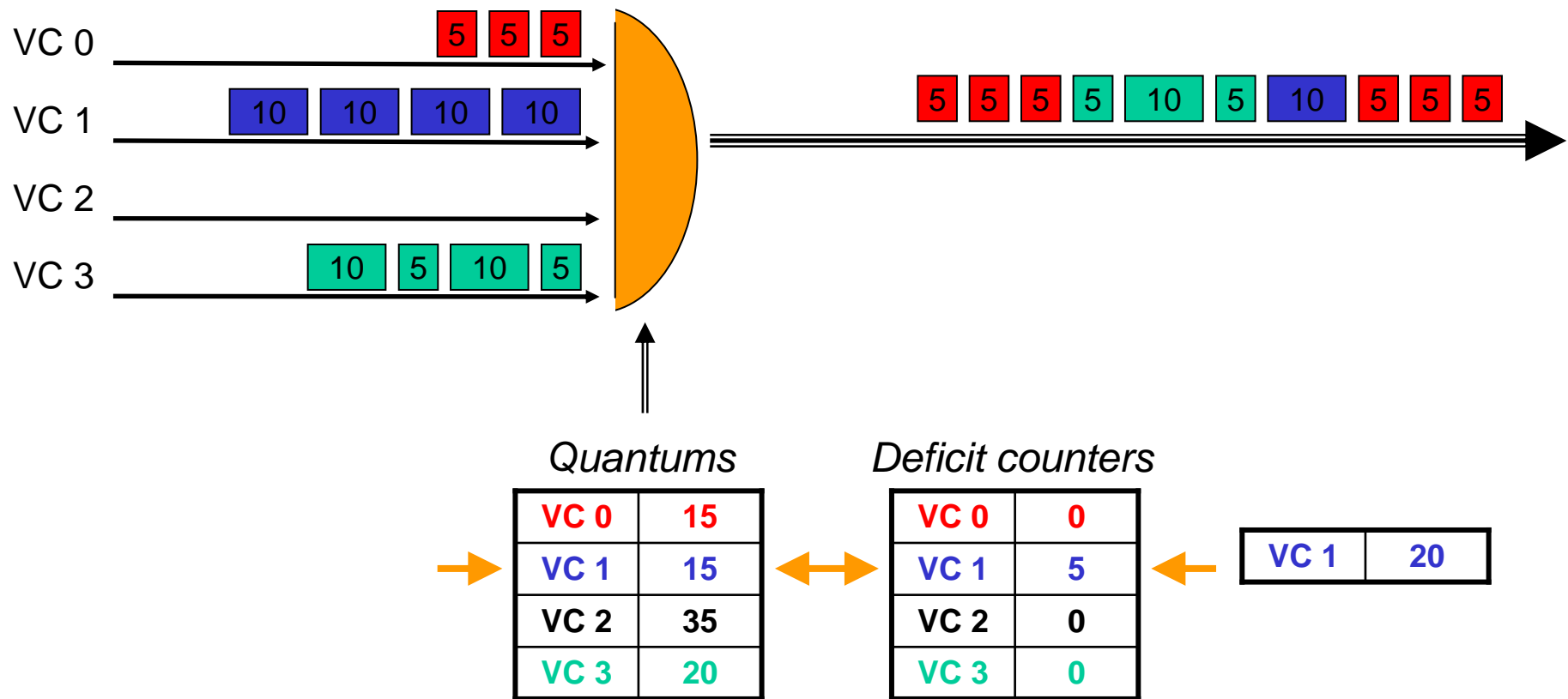
# Deficit Round Robin (DRR)



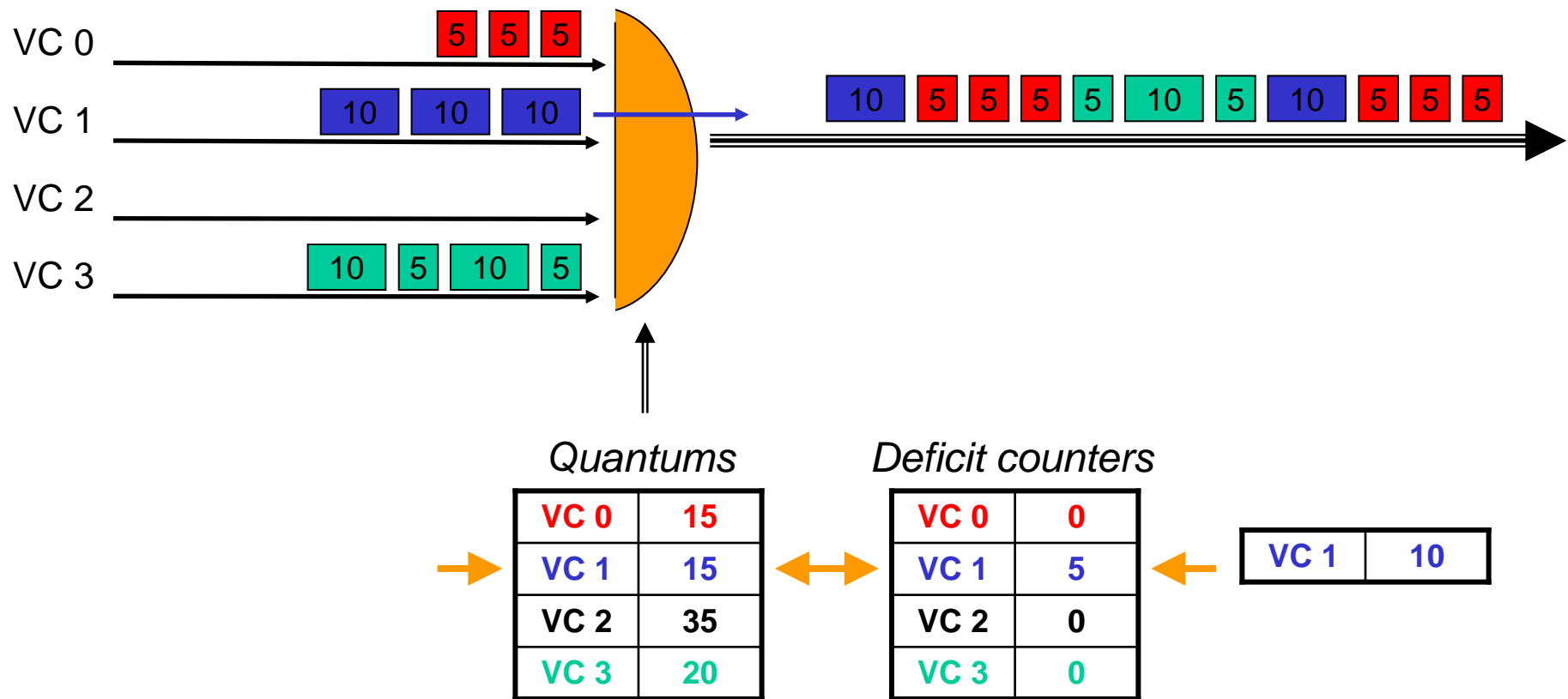
# Deficit Round Robin (DRR)



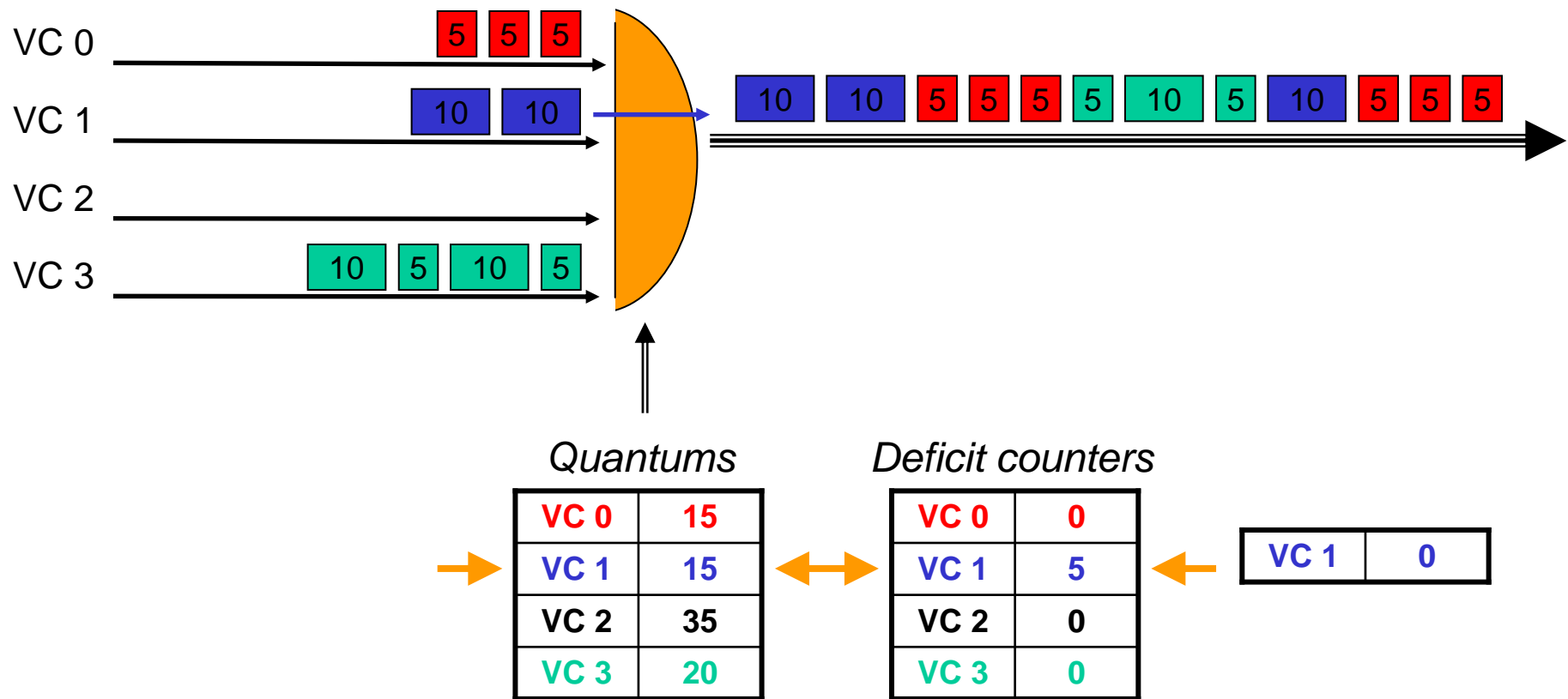
# Deficit Round Robin (DRR)



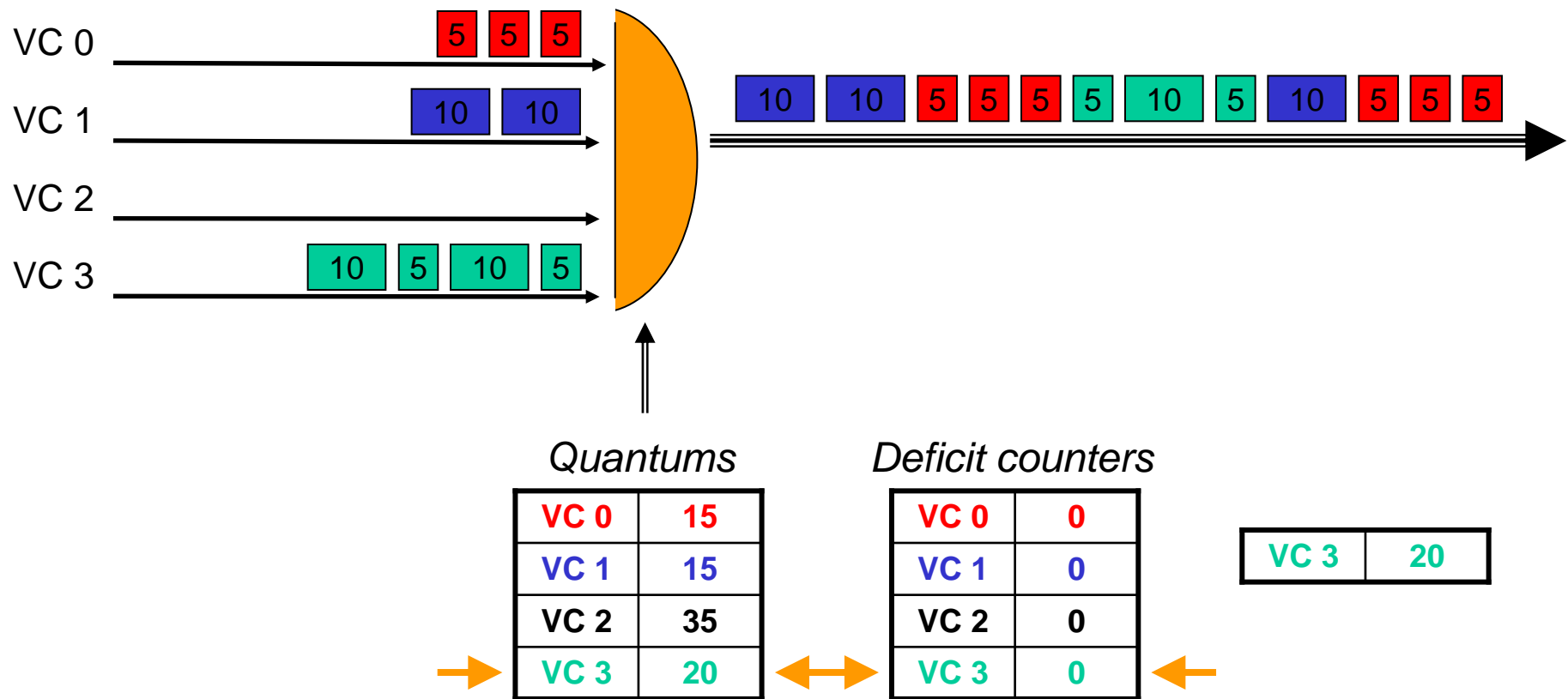
# Deficit Round Robin (DRR)



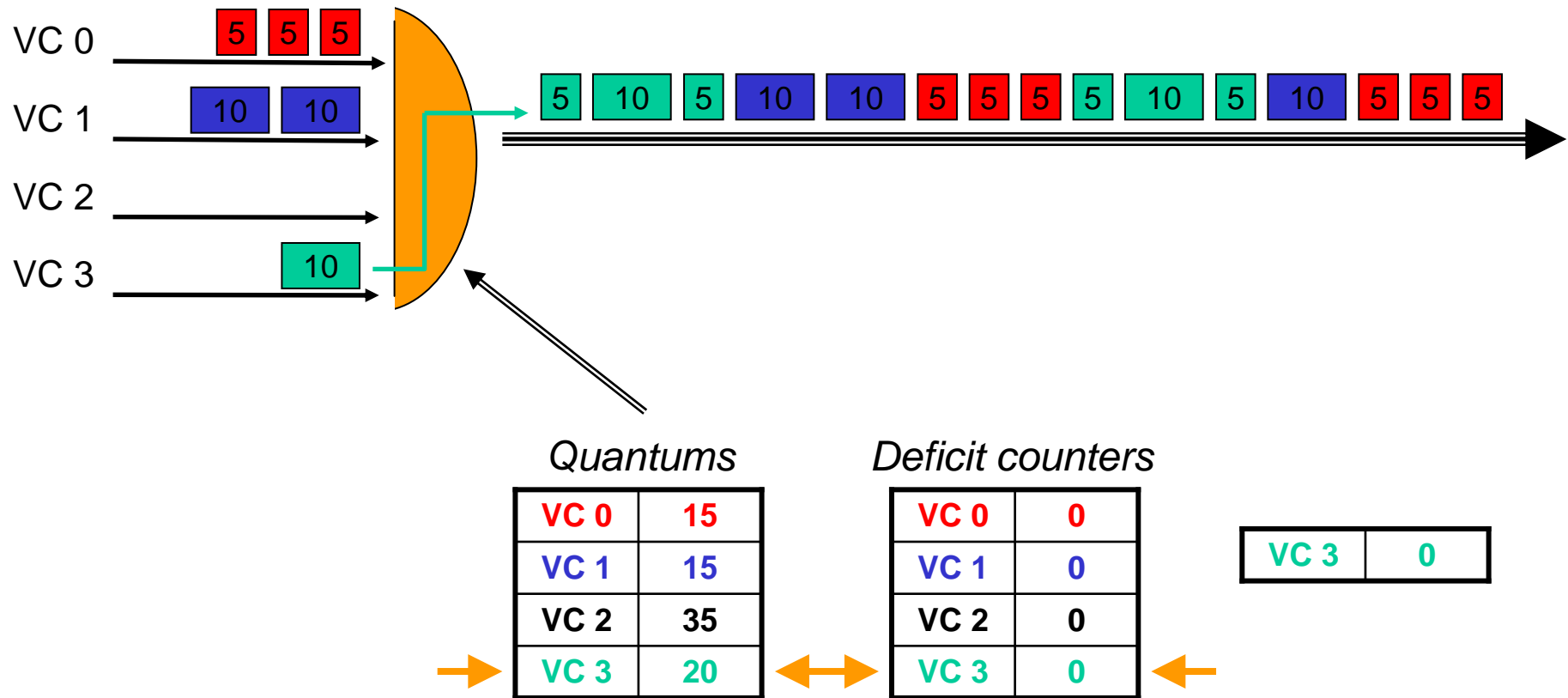
# Deficit Round Robin (DRR)



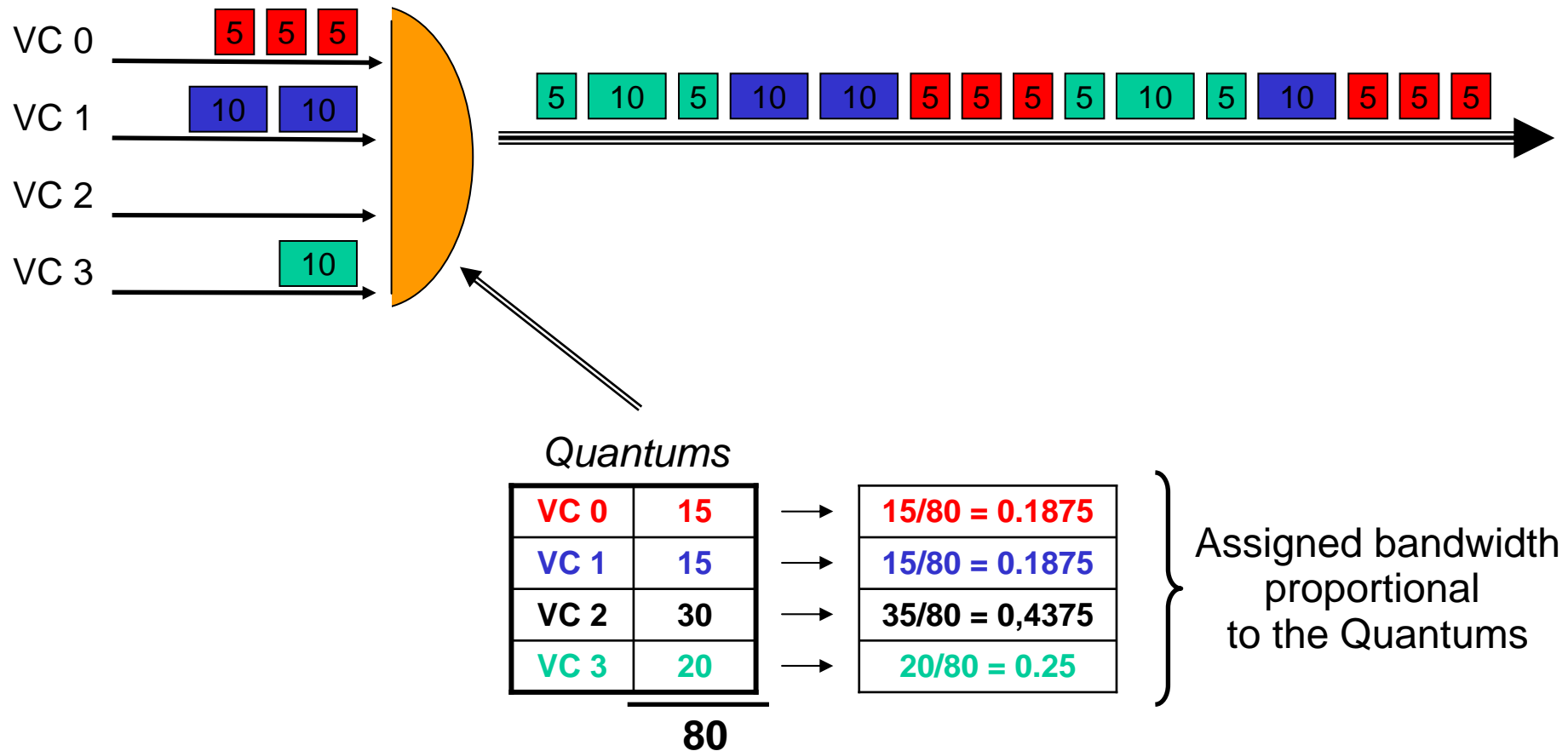
# Deficit Round Robin (DRR)



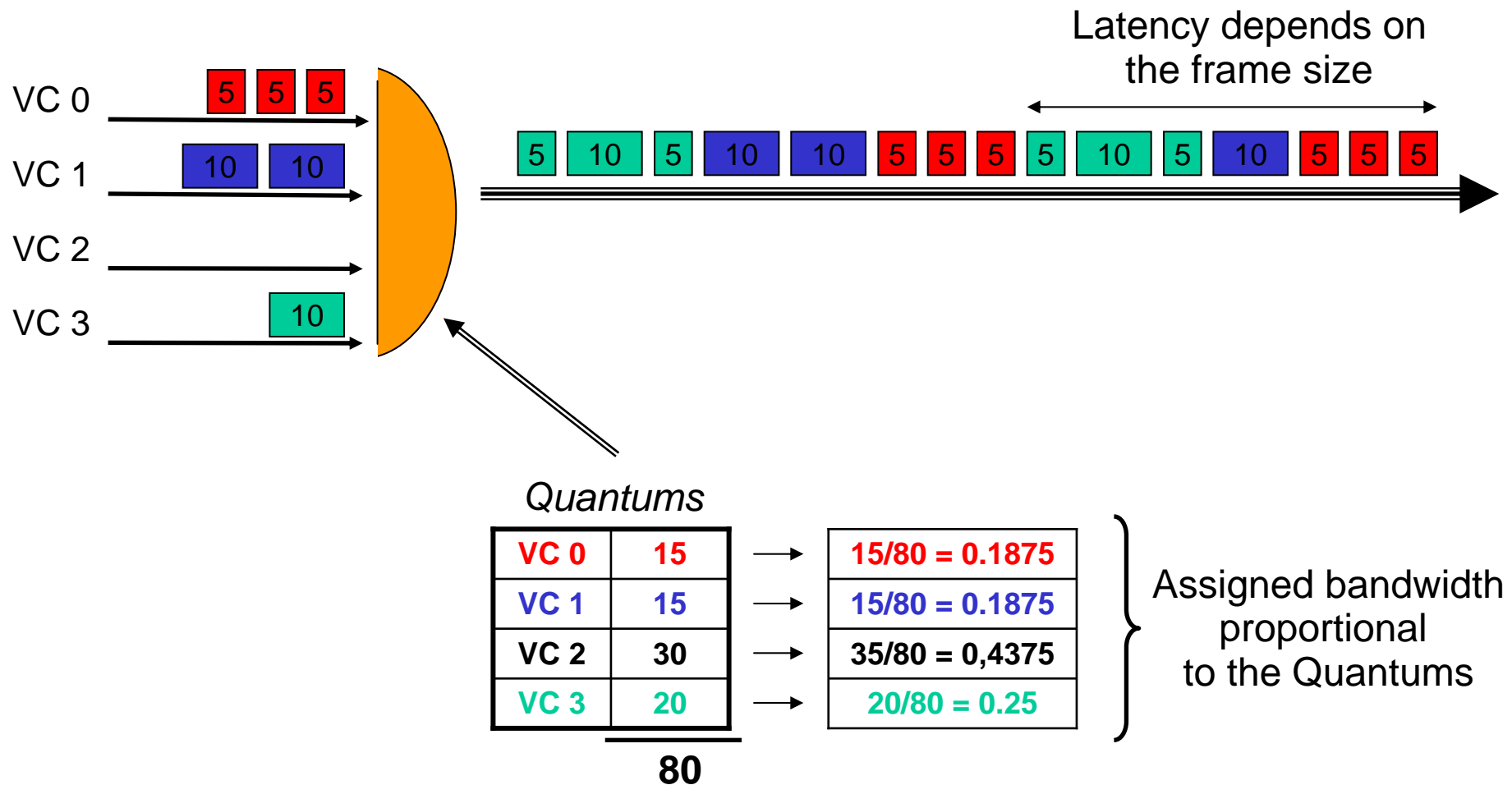
# Deficit Round Robin (DRR)



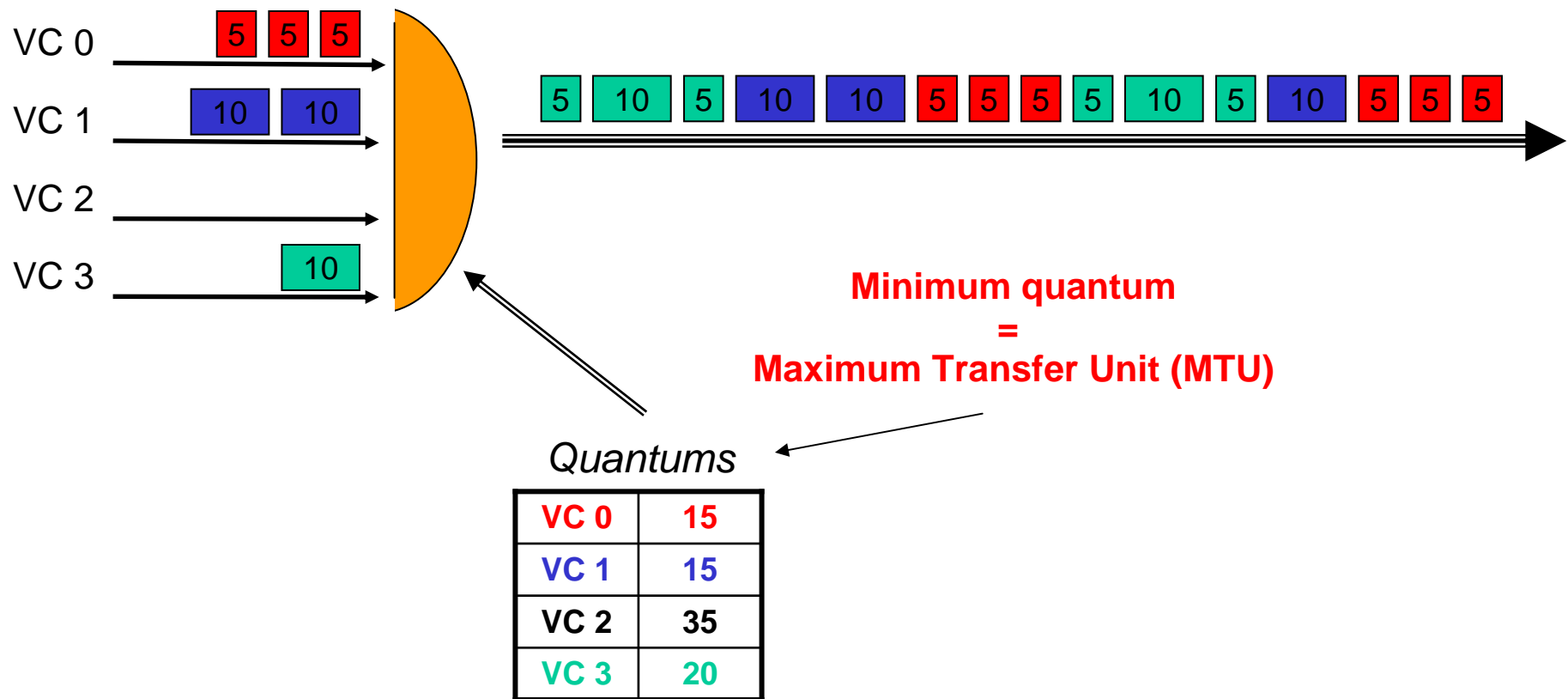
# Deficit Round Robin (DRR)



# Deficit Round Robin (DRR)



# Deficit Round Robin (DRR)

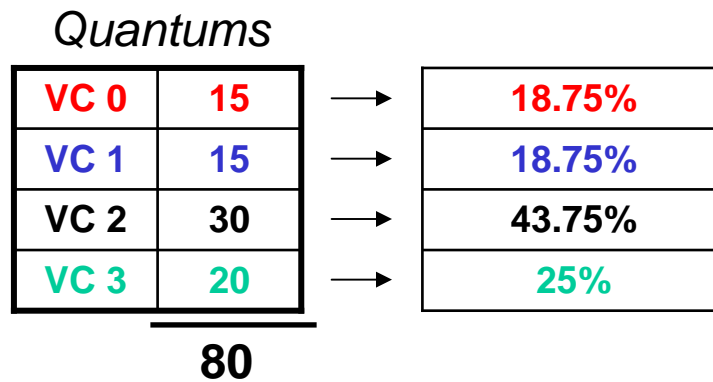


# Deficit Round Robin (DRR)

*Quantums*

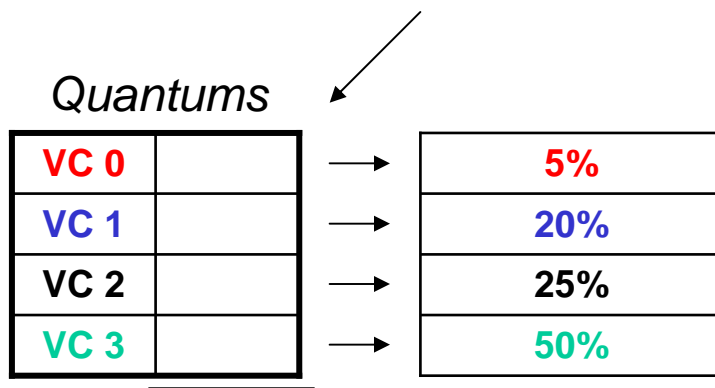
VC 0	15	→	18.75%
VC 1	15	→	18.75%
VC 2	30	→	43.75%
VC 3	20	→	25%

**80**



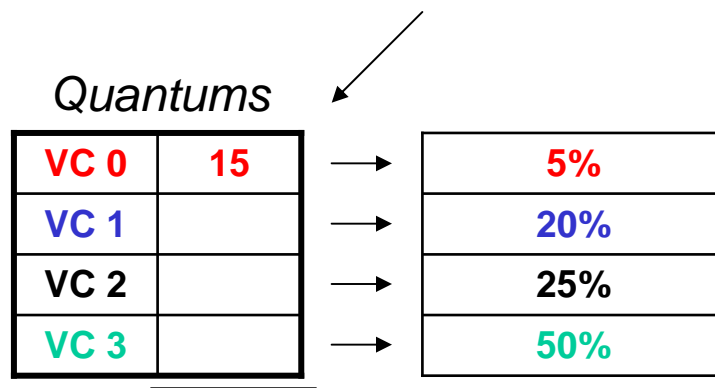
# Deficit Round Robin (DRR)

Minimum quantum = Maximum Transfer Unit (MTU)



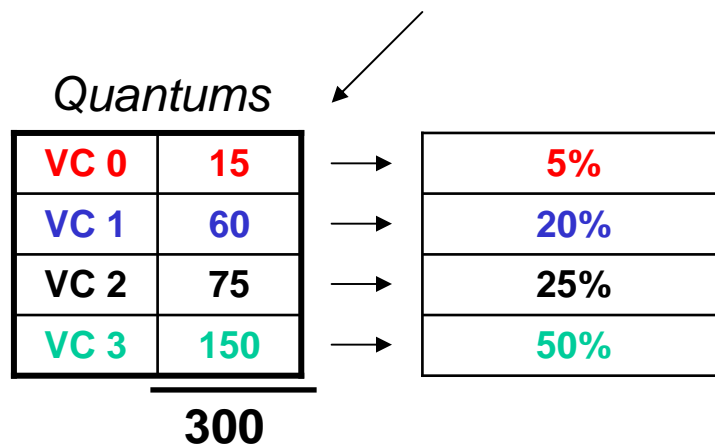
# Deficit Round Robin (DRR)

Minimum quantum = Maximum Transfer Unit (MTU)



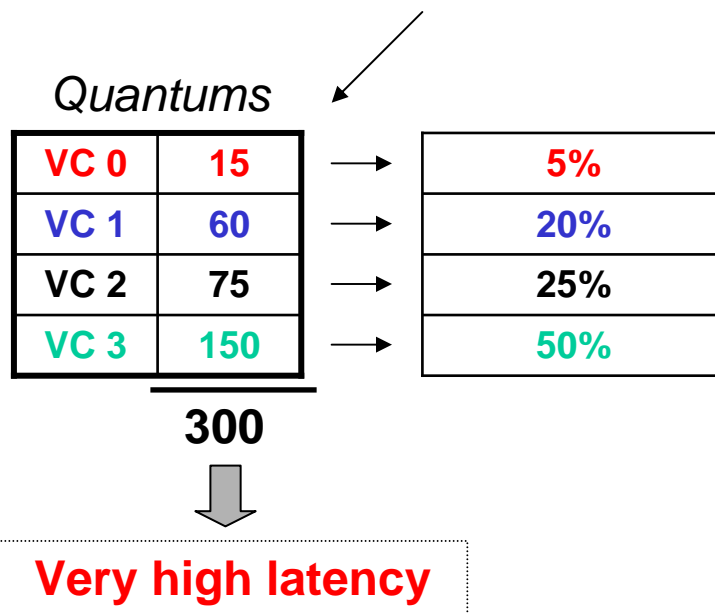
# Deficit Round Robin (DRR)

Minimum quantum = Maximum Transfer Unit (MTU)



# Deficit Round Robin (DRR)

Minimum quantum = Maximum Transfer Unit (MTU)

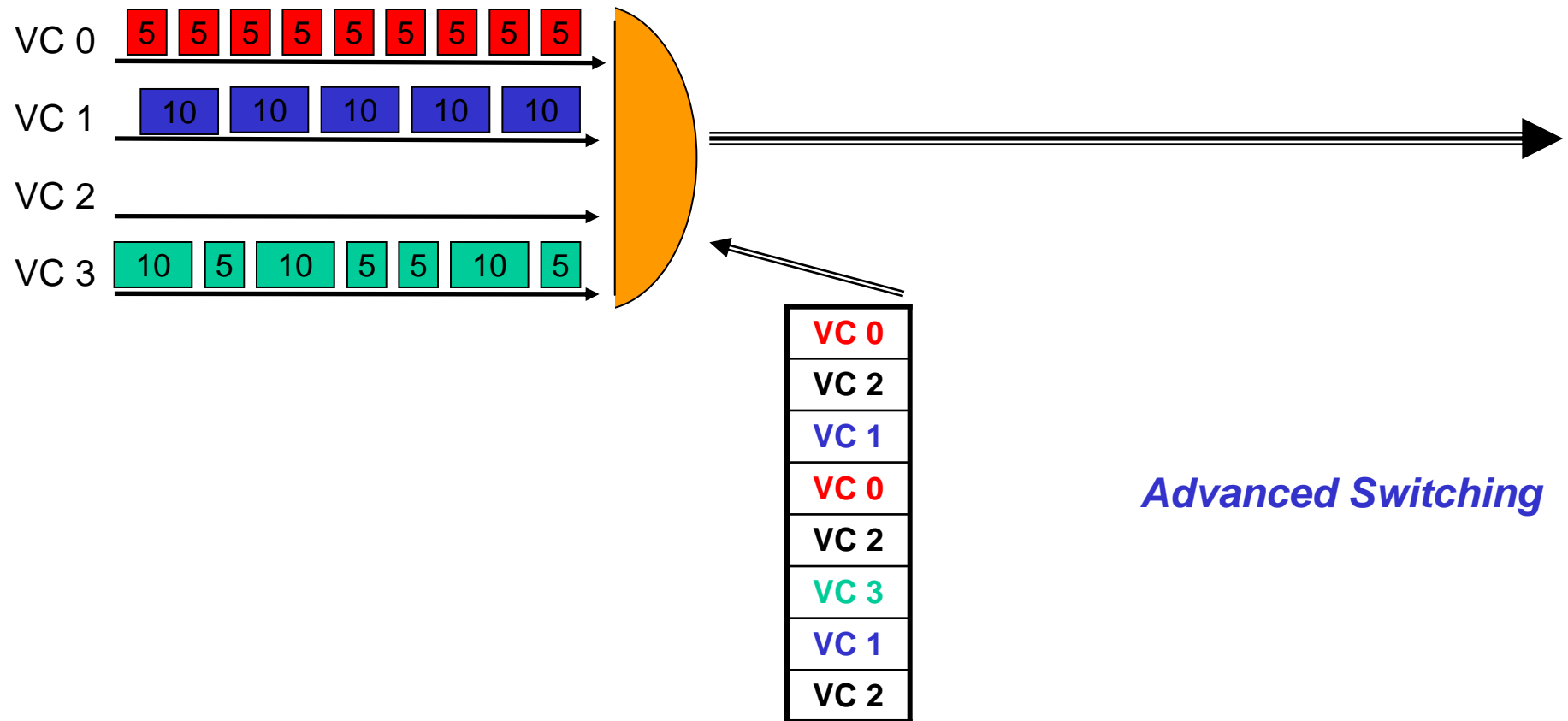


# Outline

- The Deficit Round Robin (DRR) scheduler
- **The Deficit Table (DTable) scheduler**
  - **The DTable scheduling mechanism**
  - Configuring the DTable scheduler
- Performance evaluation
- Conclusions

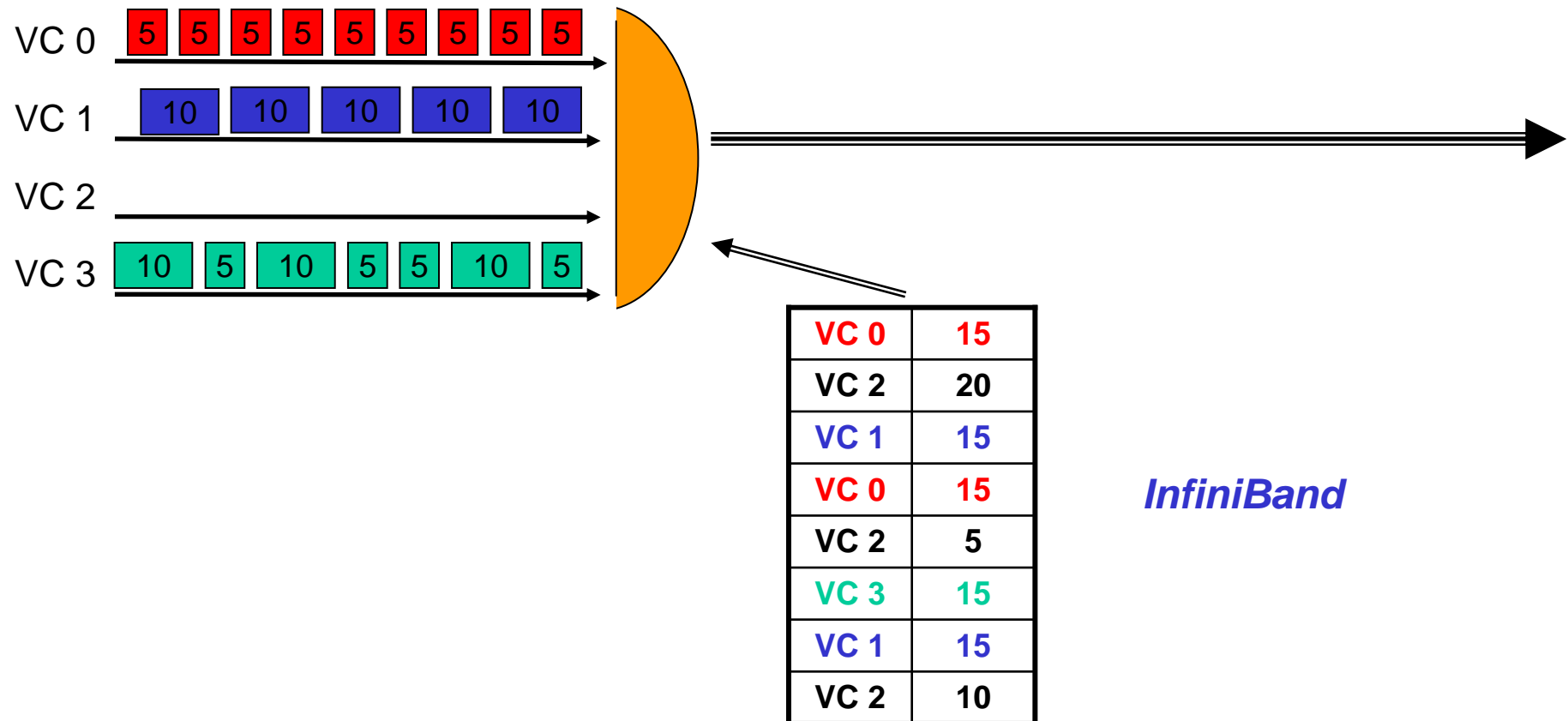
# Deficit Table (DTable)

Scheduling mechanism



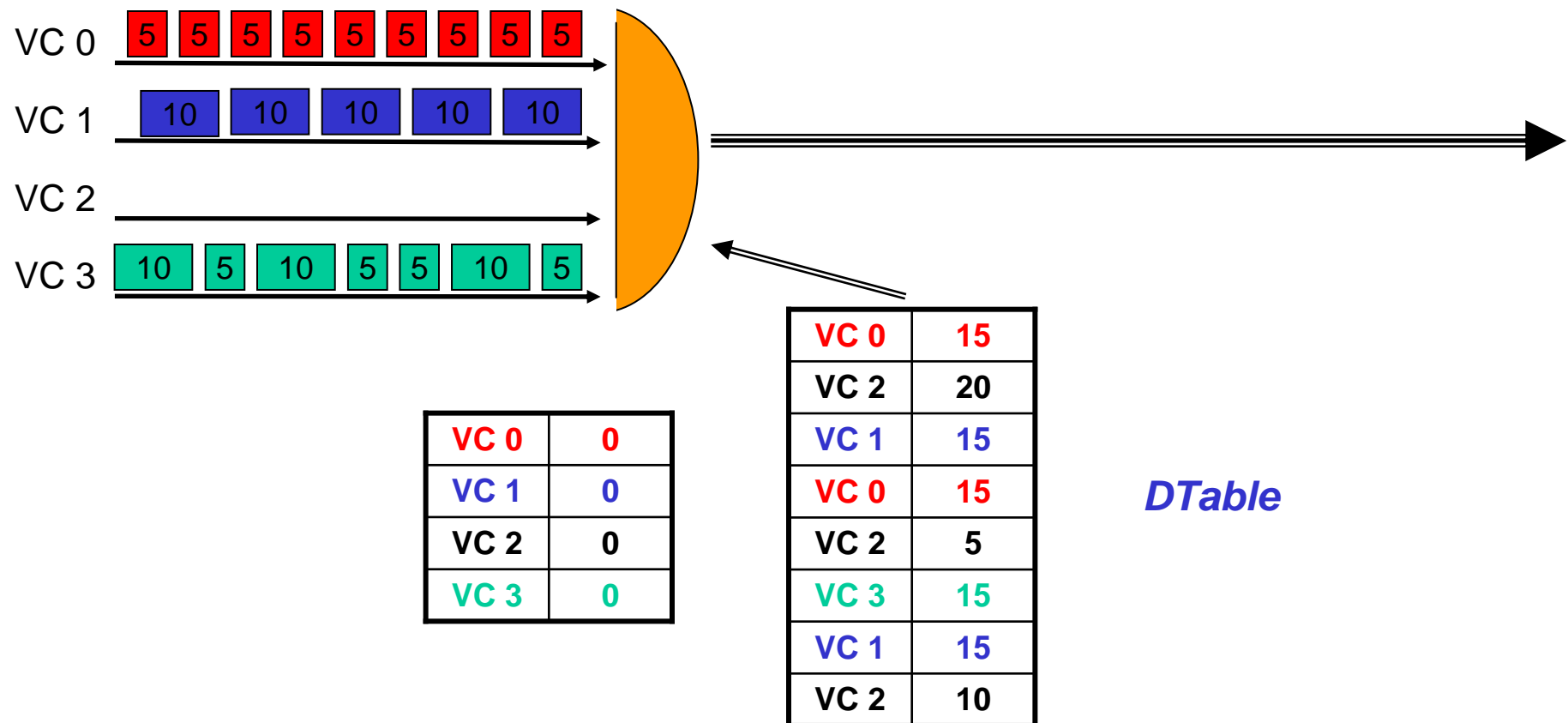
# Deficit Table (DTable)

Scheduling mechanism



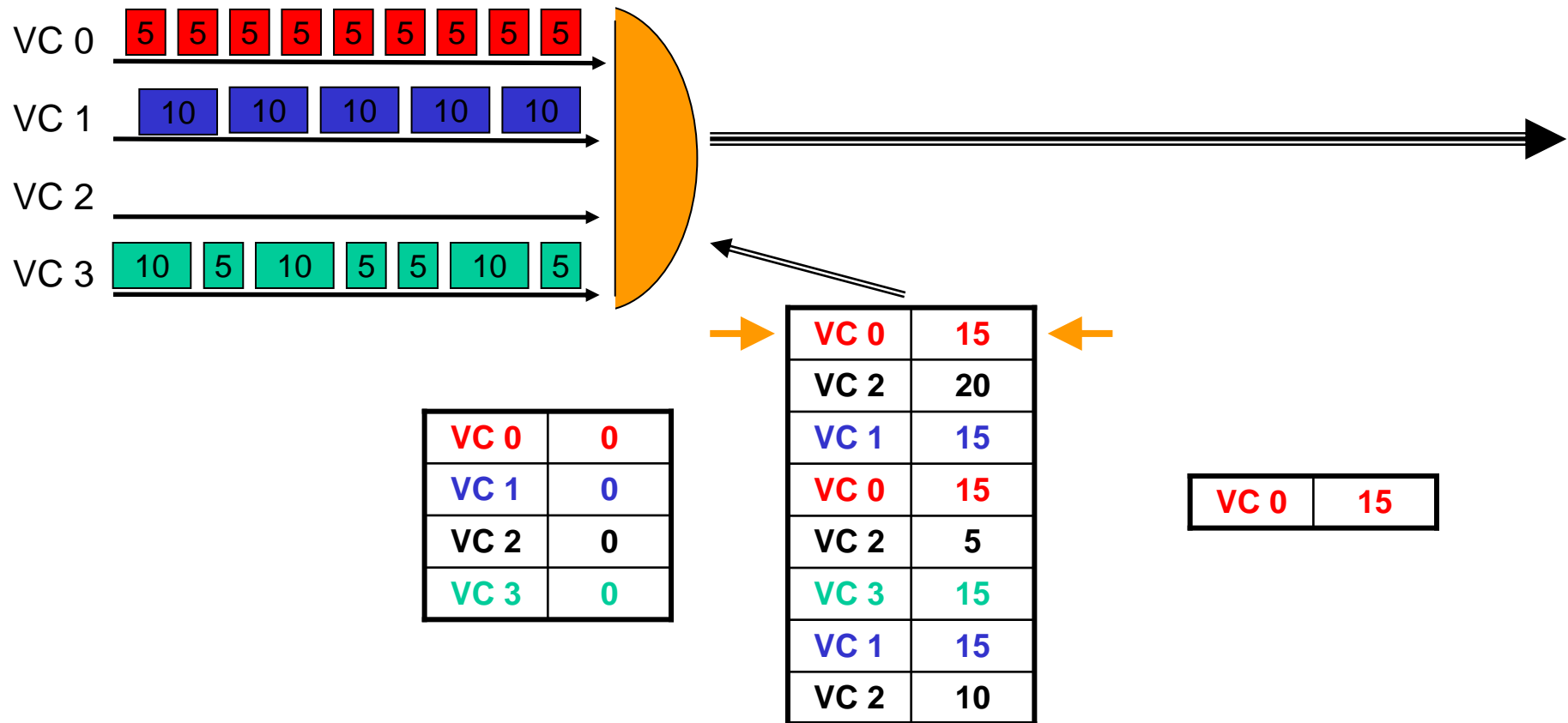
# Deficit Table (DTable)

## Scheduling mechanism



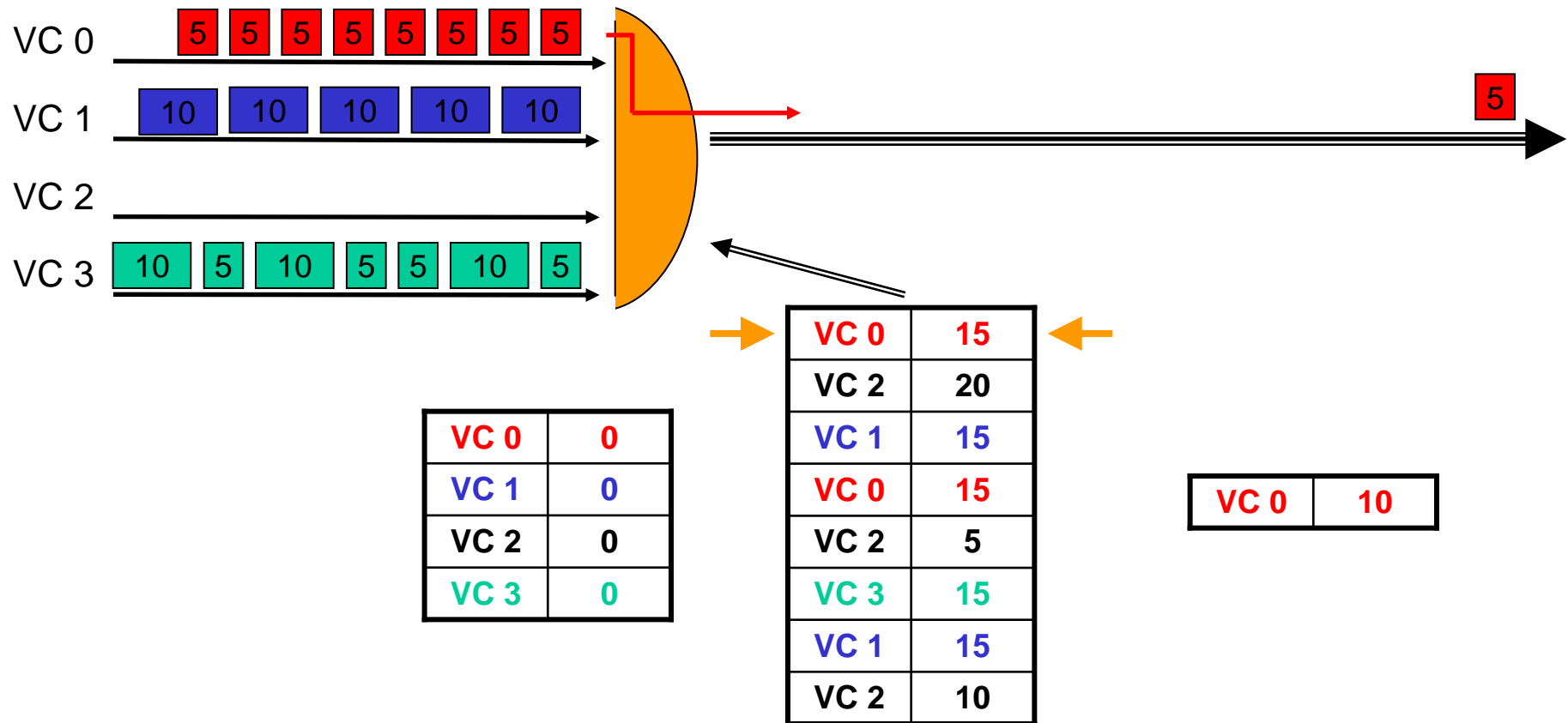
# Deficit Table (DTable)

## Scheduling mechanism



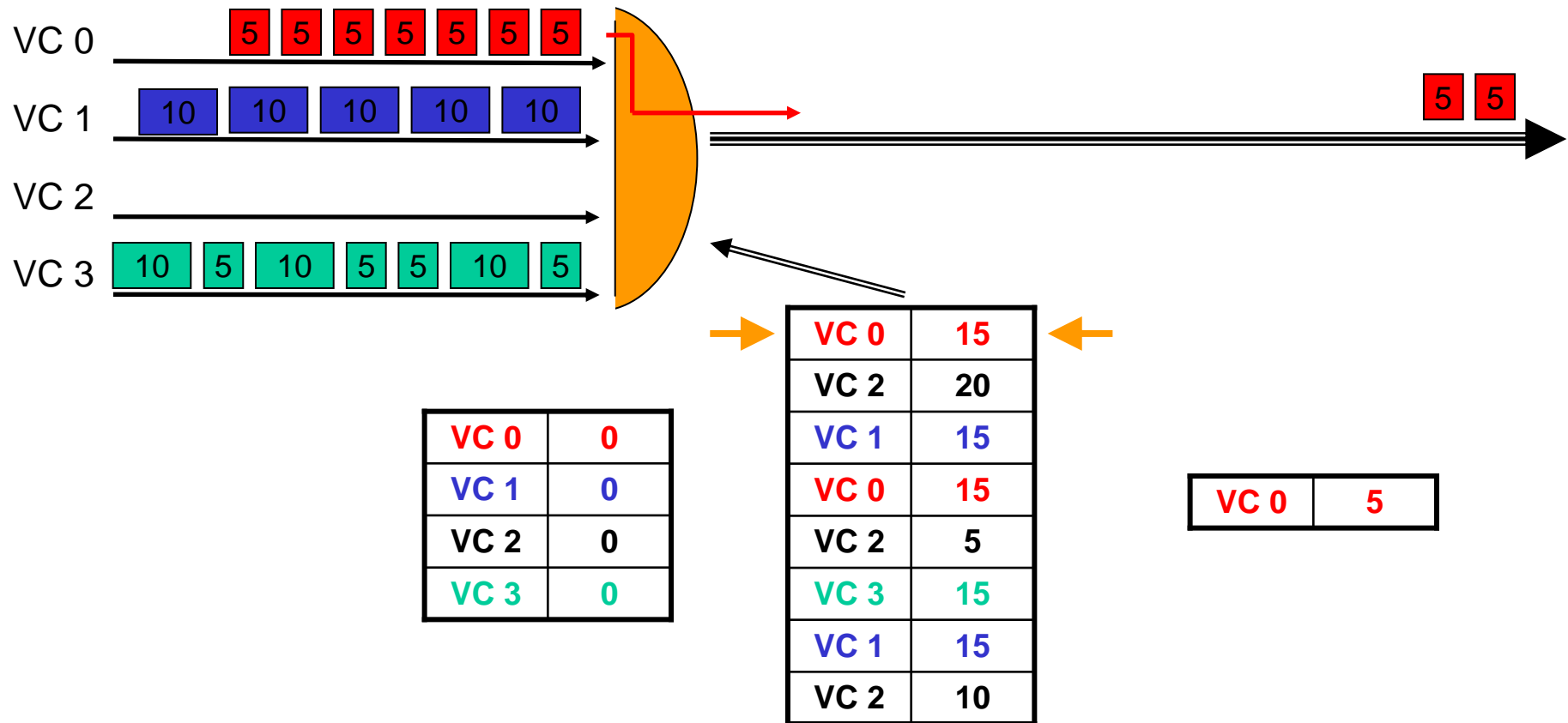
# Deficit Table (DTable)

## Scheduling mechanism



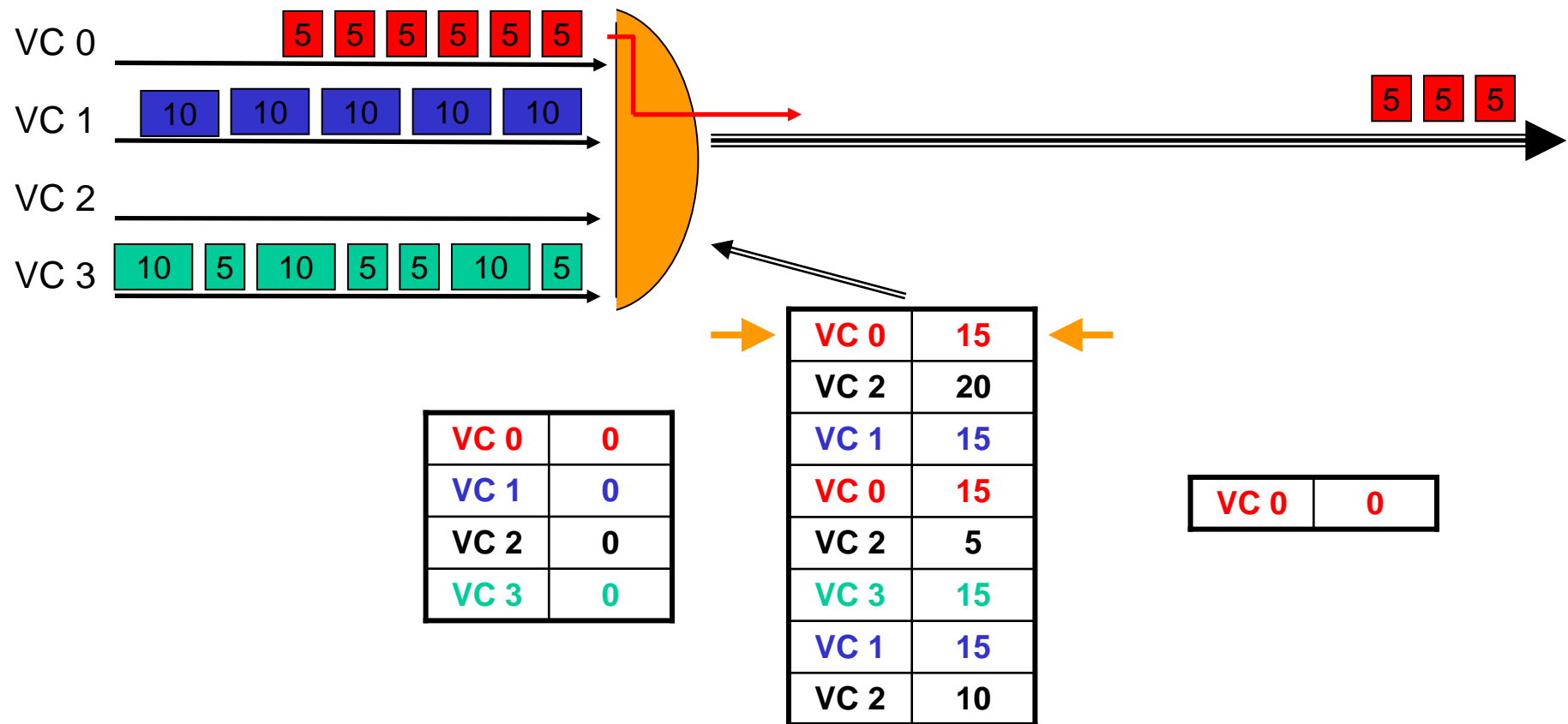
# Deficit Table (DTable)

## Scheduling mechanism



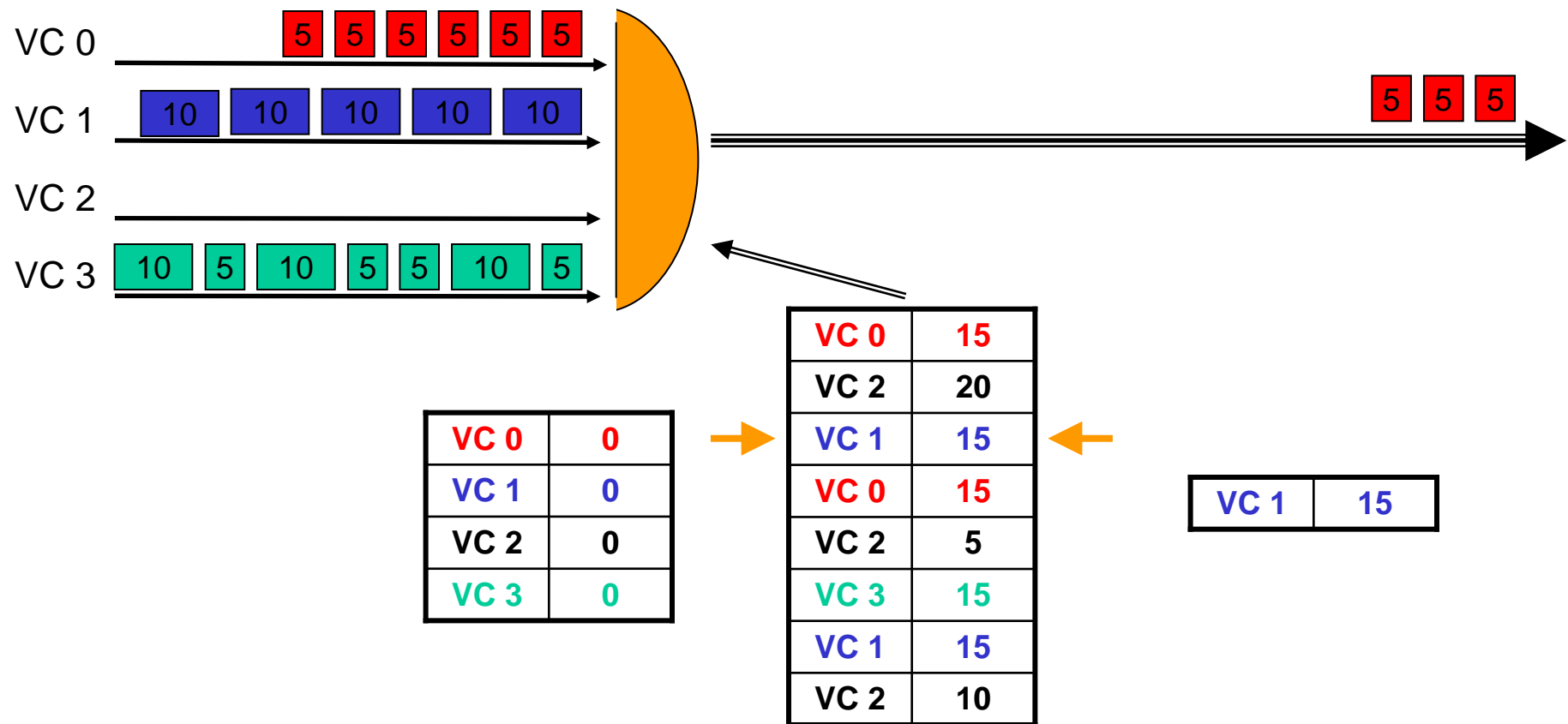
# Deficit Table (DTable)

## Scheduling mechanism



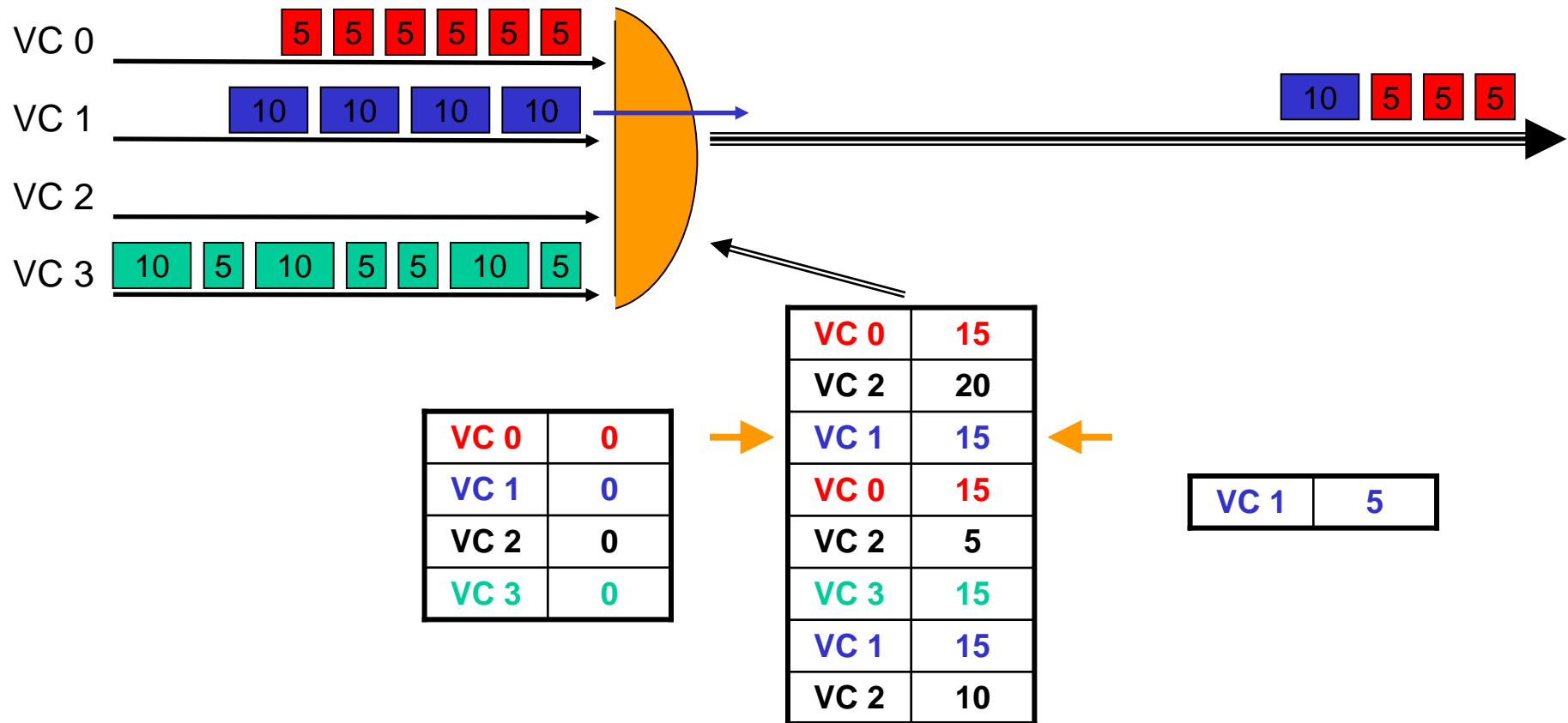
# Deficit Table (DTable)

## Scheduling mechanism



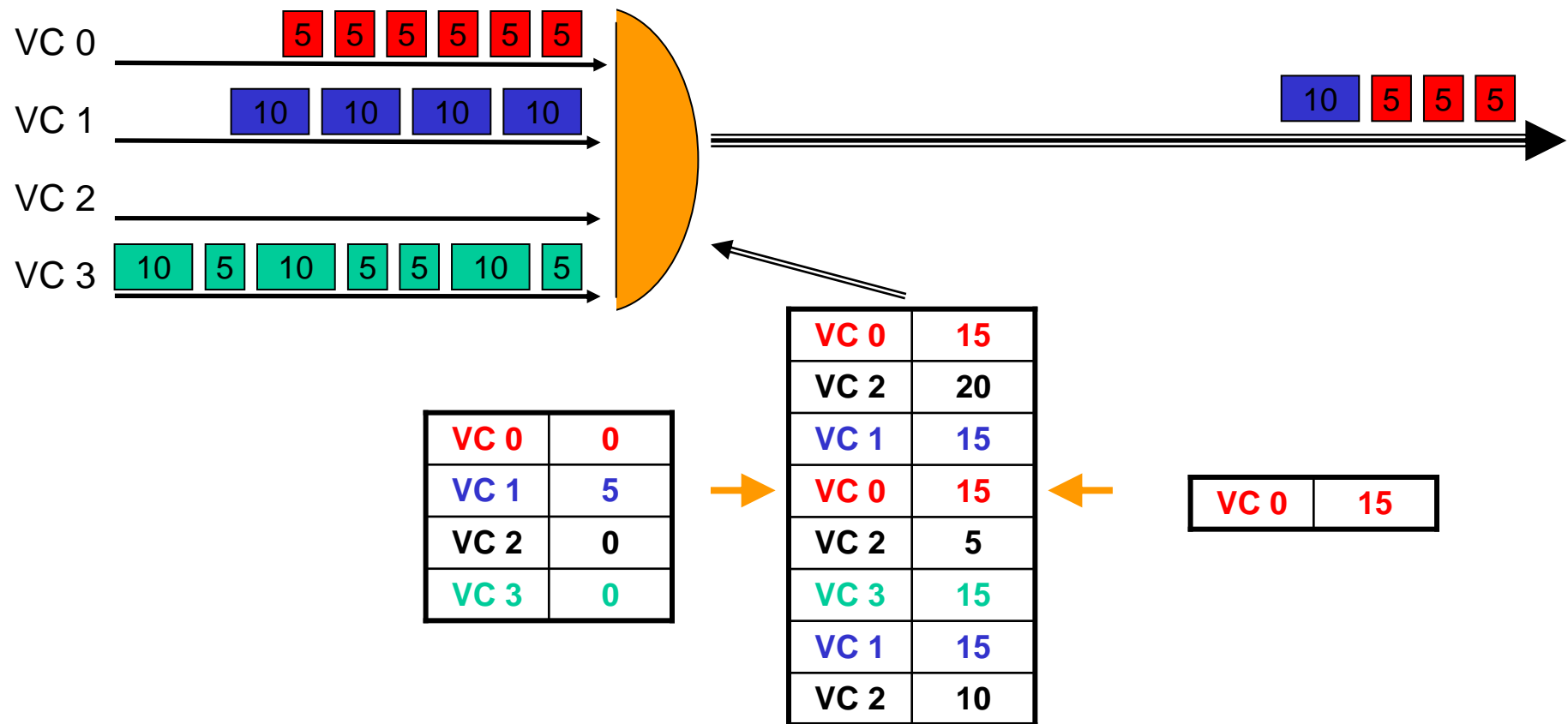
# Deficit Table (DTable)

## Scheduling mechanism



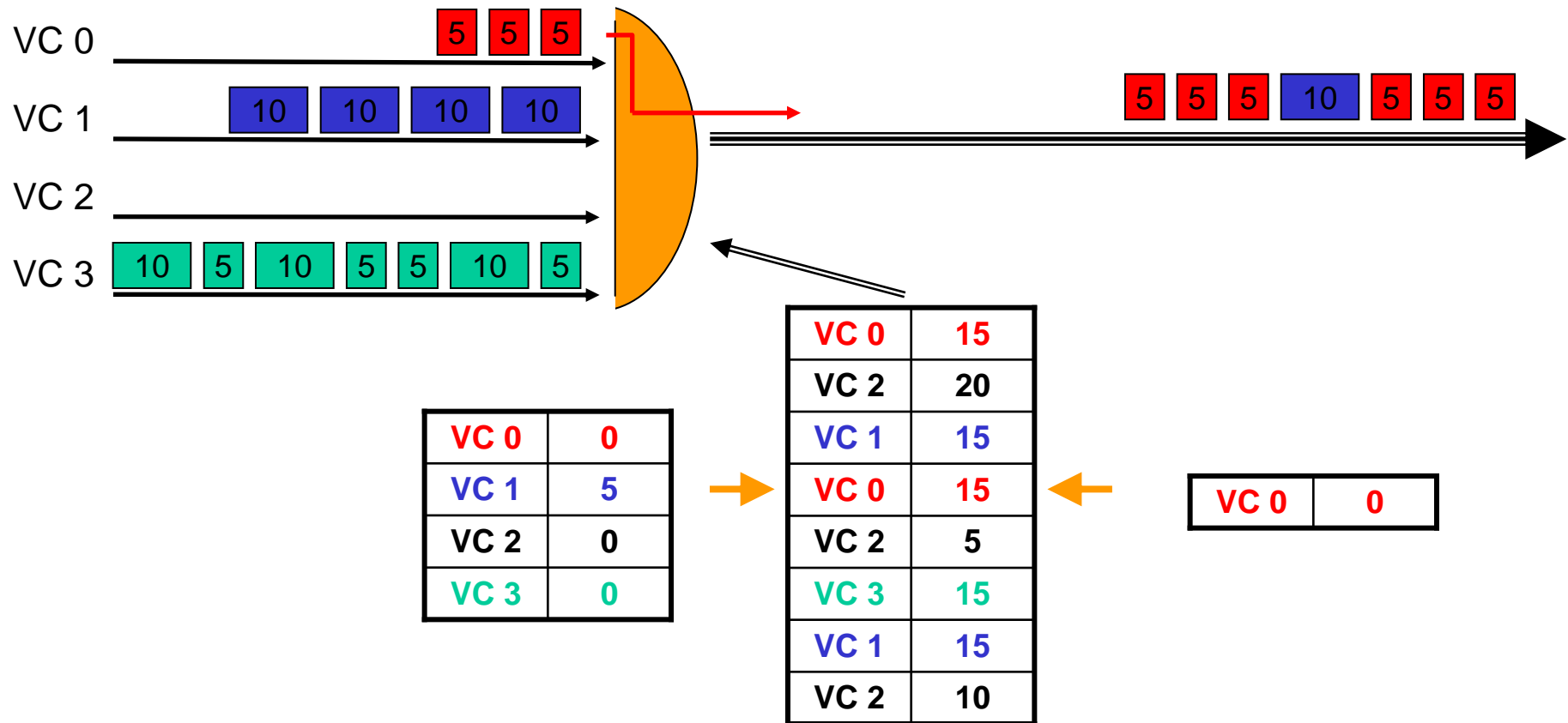
# Deficit Table (DTable)

## Scheduling mechanism



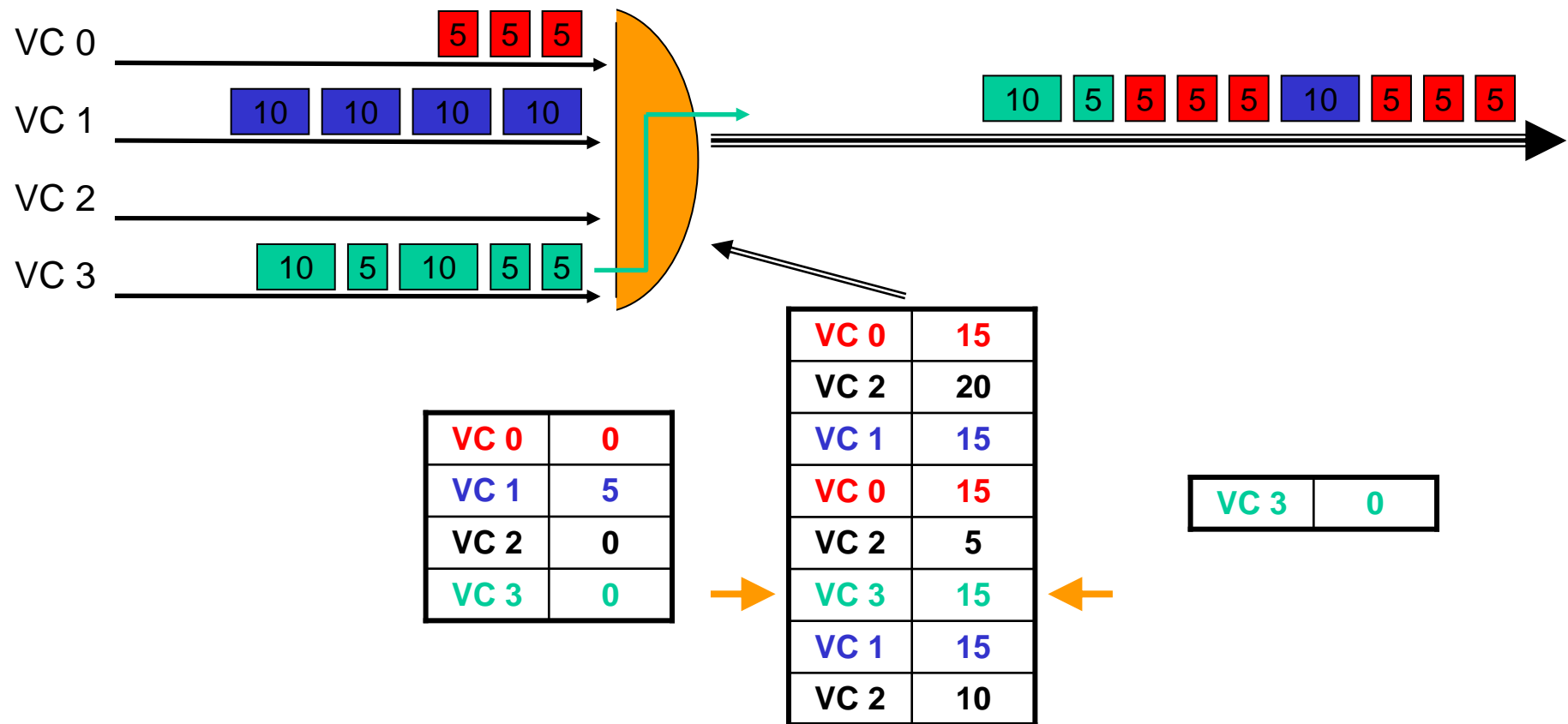
# Deficit Table (DTable)

## Scheduling mechanism



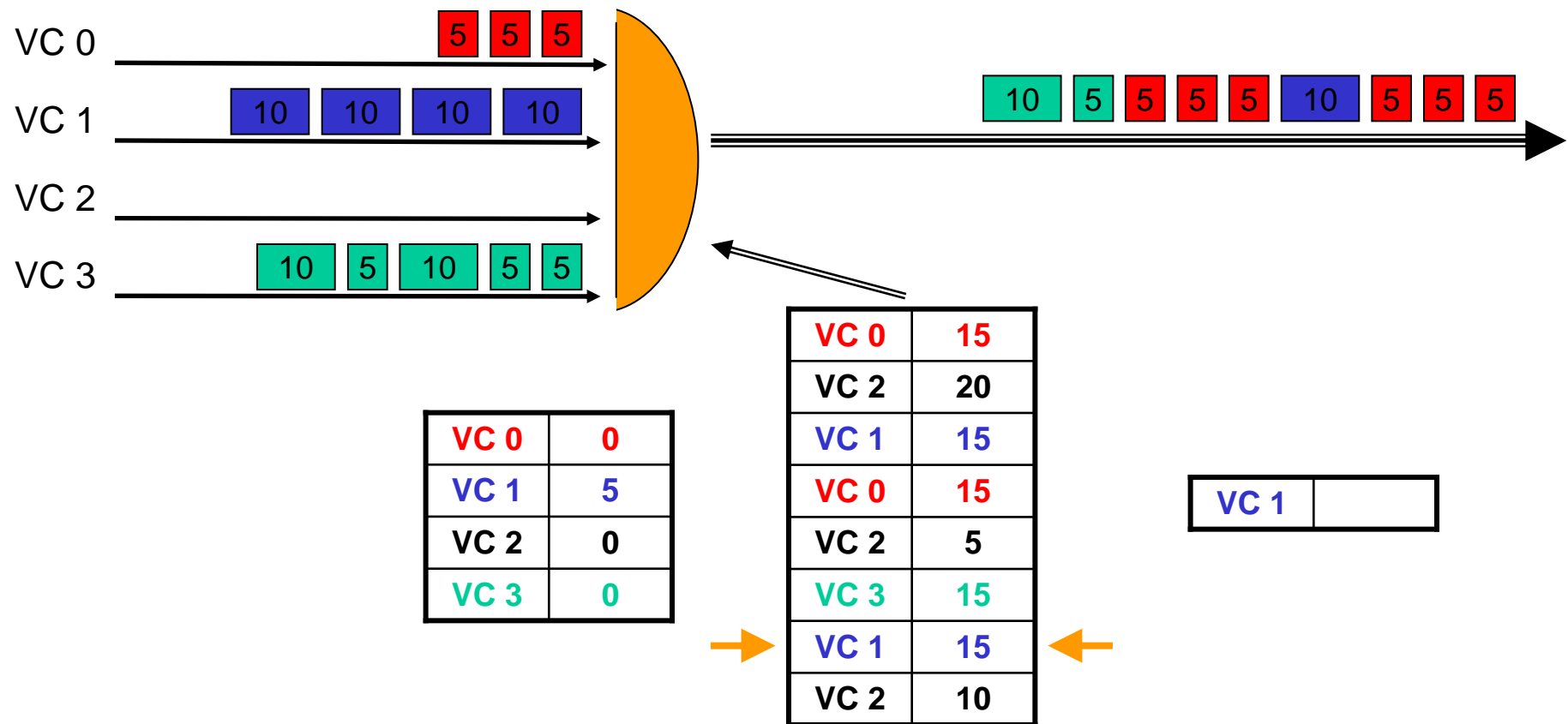
# Deficit Table (DTable)

## Scheduling mechanism



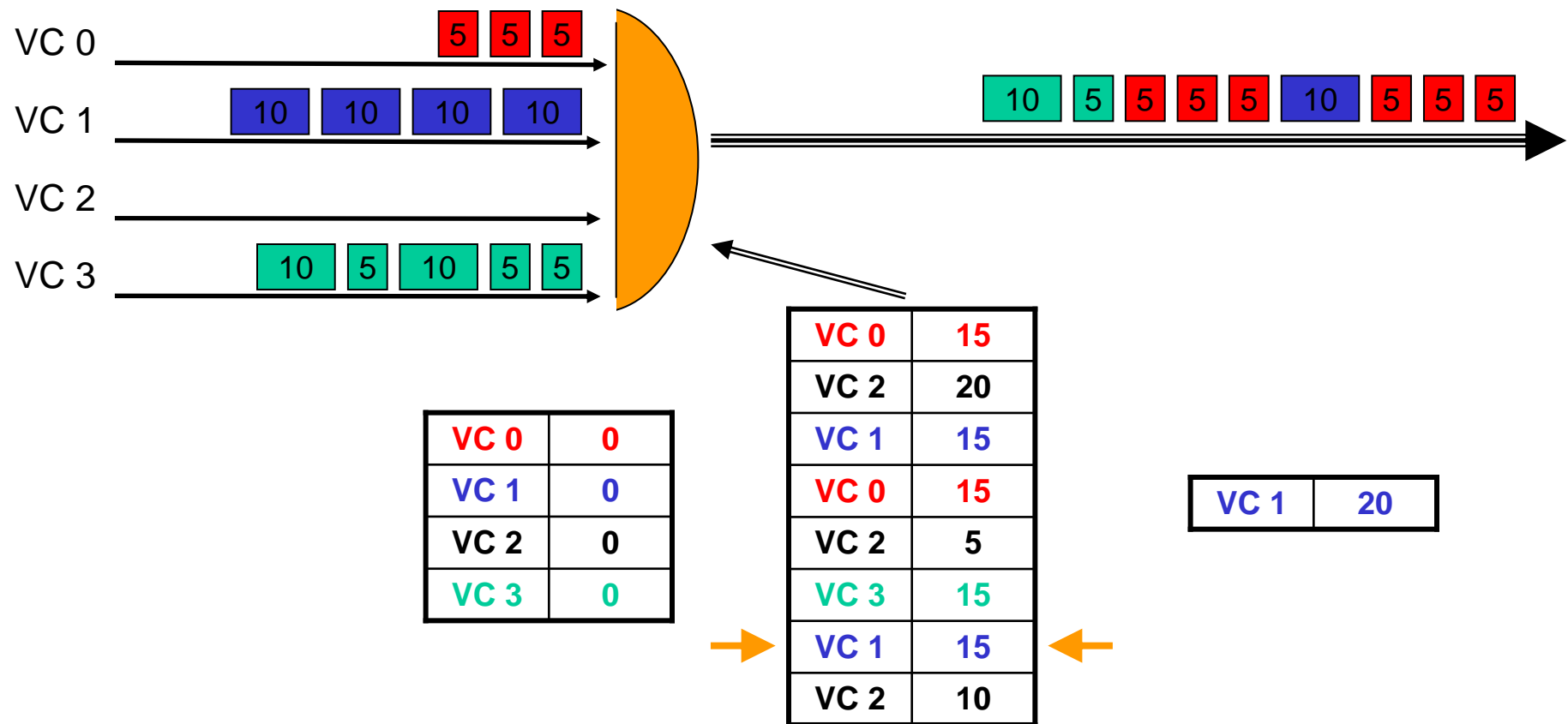
# Deficit Table (DTable)

## Scheduling mechanism



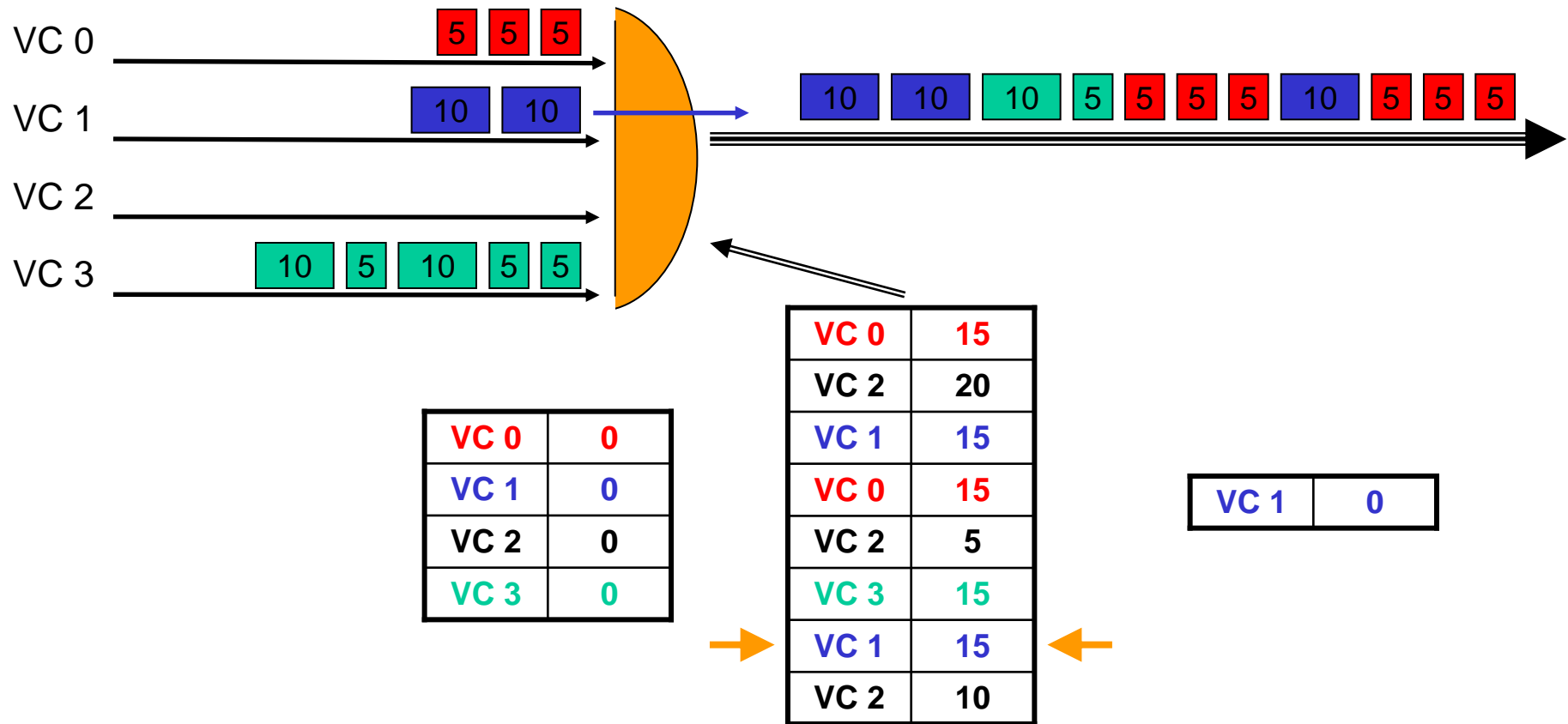
# Deficit Table (DTable)

## Scheduling mechanism



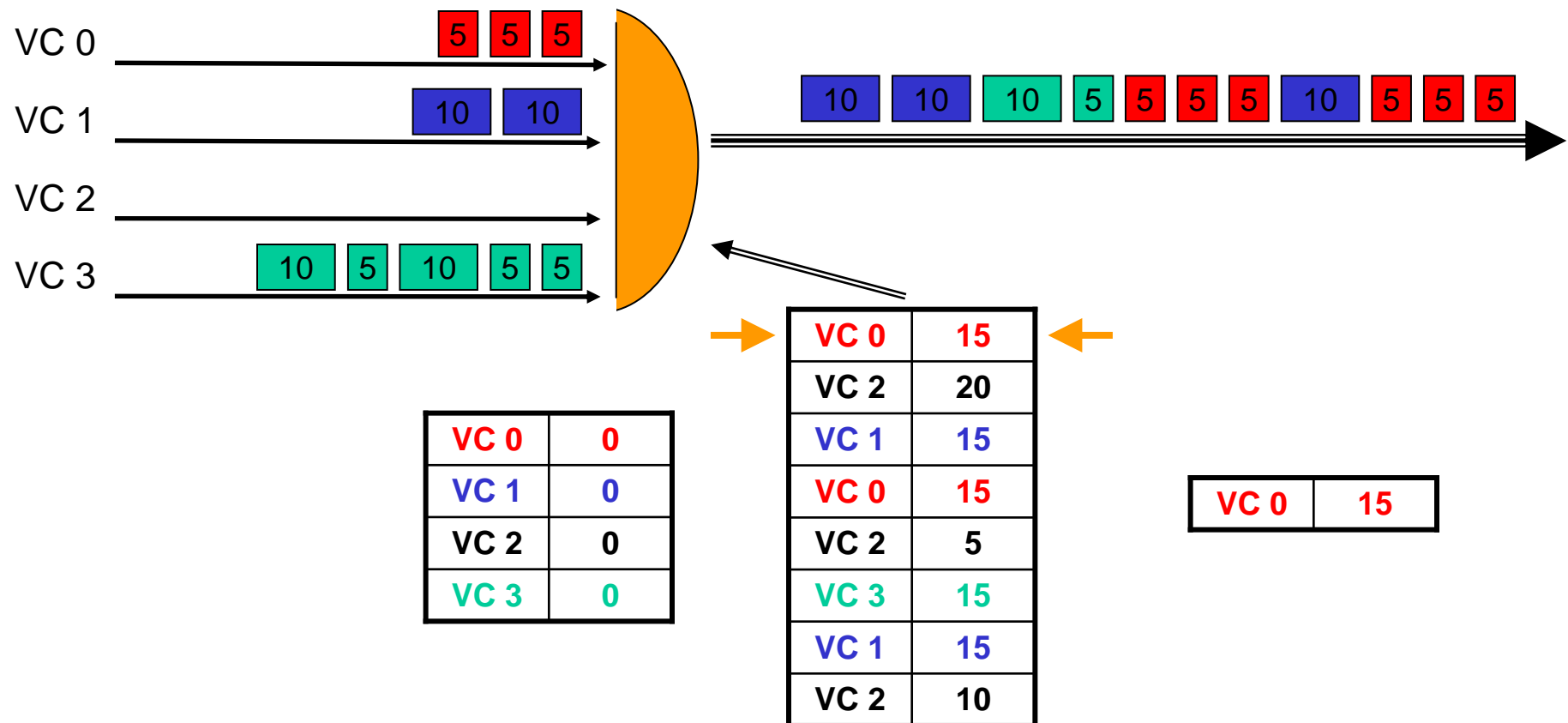
# Deficit Table (DTable)

## Scheduling mechanism



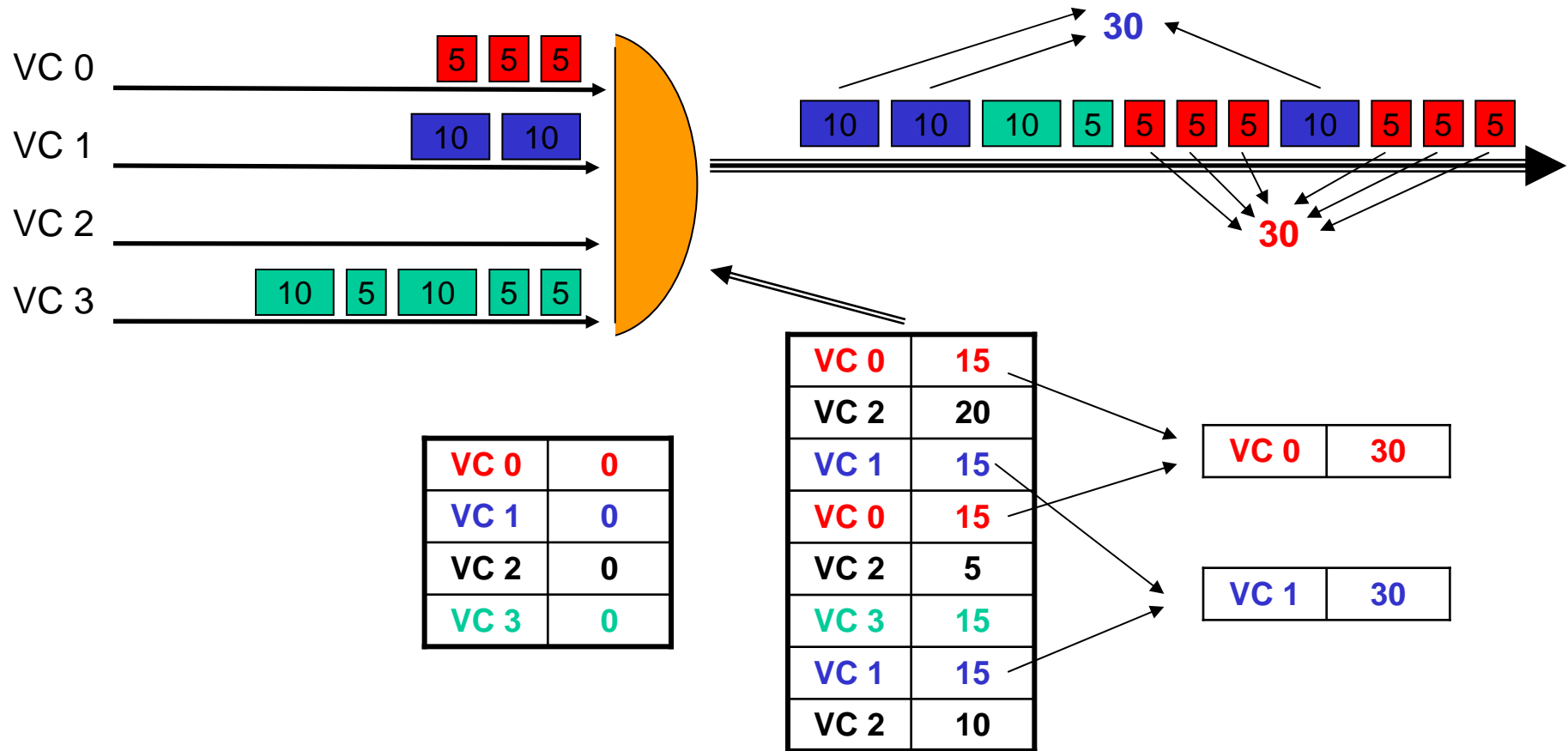
# Deficit Table (DTable)

## Scheduling mechanism



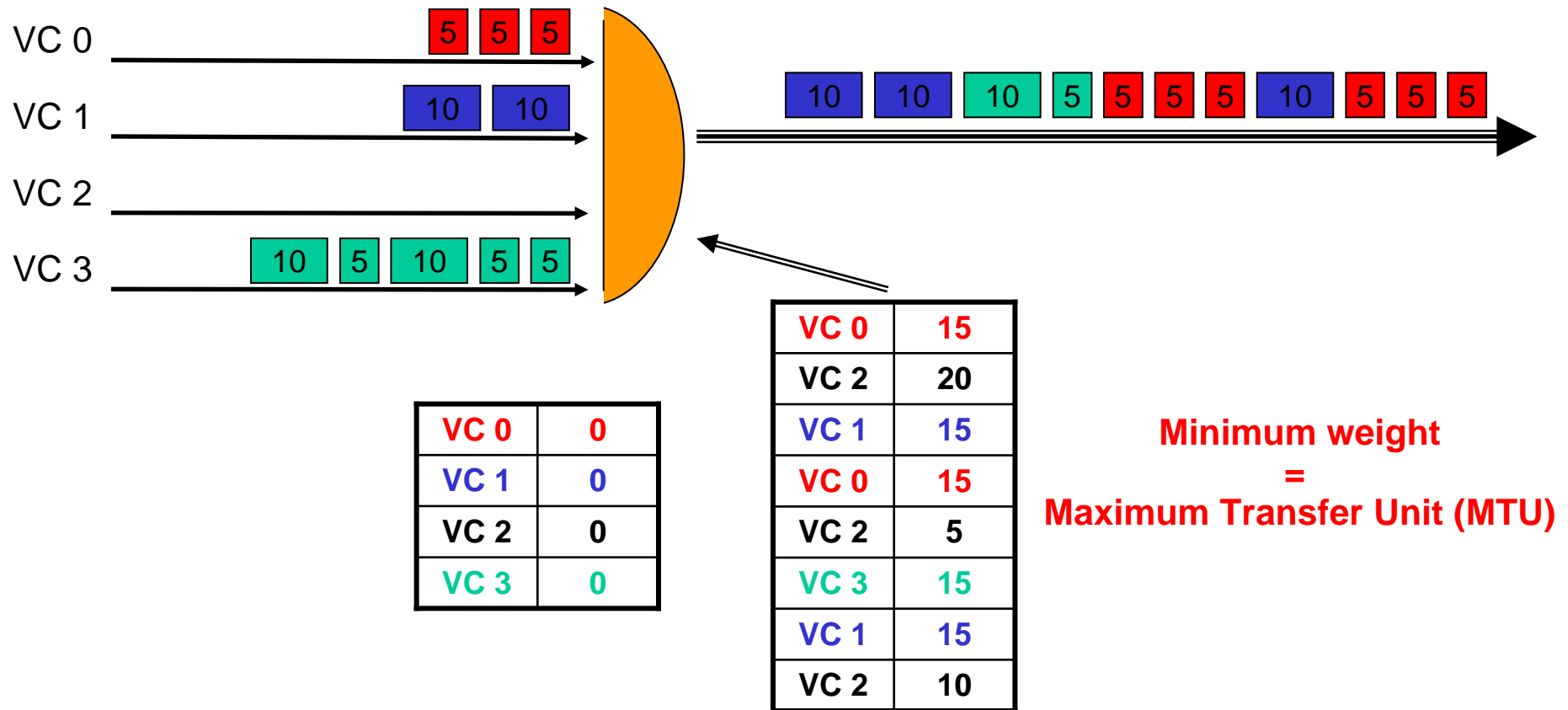
# Deficit Table (DTable)

## Scheduling mechanism



# Deficit Table (DTable)

## Scheduling mechanism



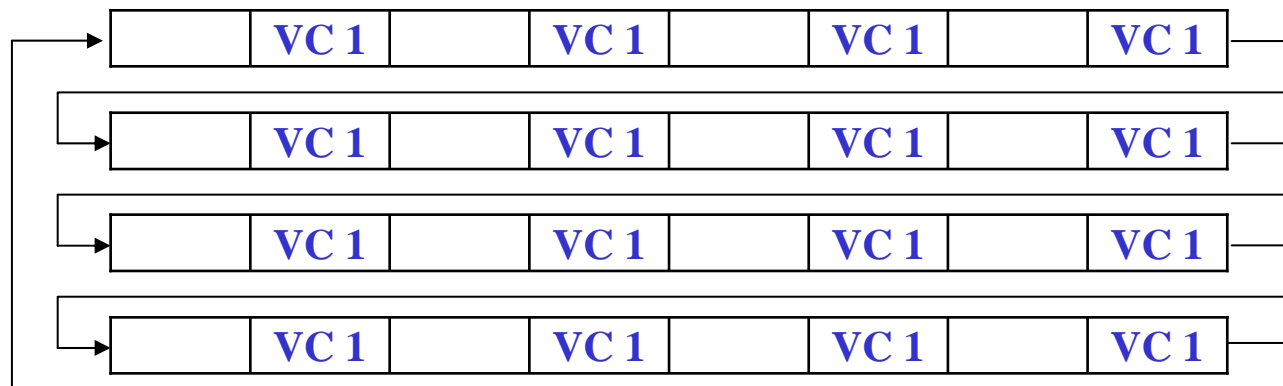
# Outline

- The Deficit Round Robin (DRR) scheduler
- **The Deficit Table (DTable) scheduler**
  - The DTable scheduling mechanism
  - **Configuring the DTable scheduler**
- Performance evaluation
- Conclusions

# Deficit Table (DTable)

## Configuration methodology

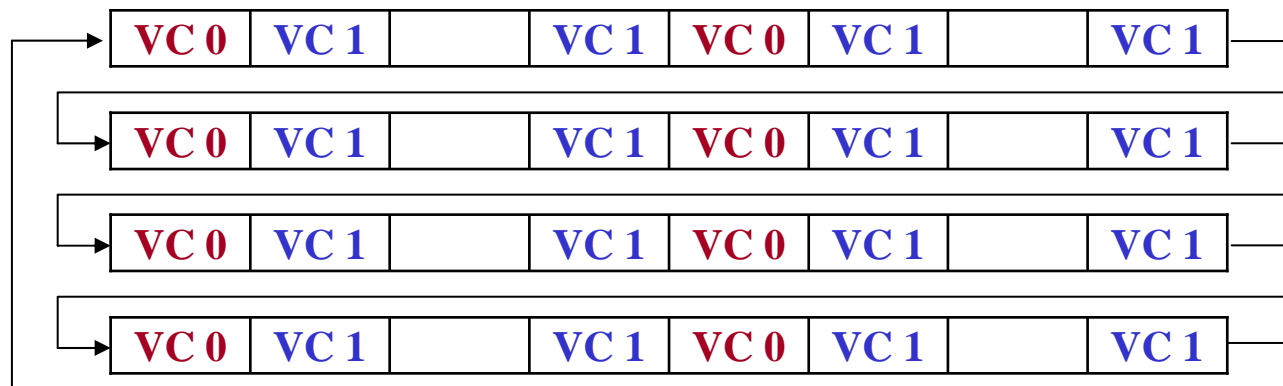
- With this kind of table-based schedulers it is possible to control the **latency** of a flow or aggregated of flows by controlling the **maximum separation between any consecutive pair of entries** assigned to that flow.
  - This distance determines the maximum time that a packet at the head of a flow queue is going to wait until being transmitted.



# Deficit Table (DTable)

## Configuration methodology

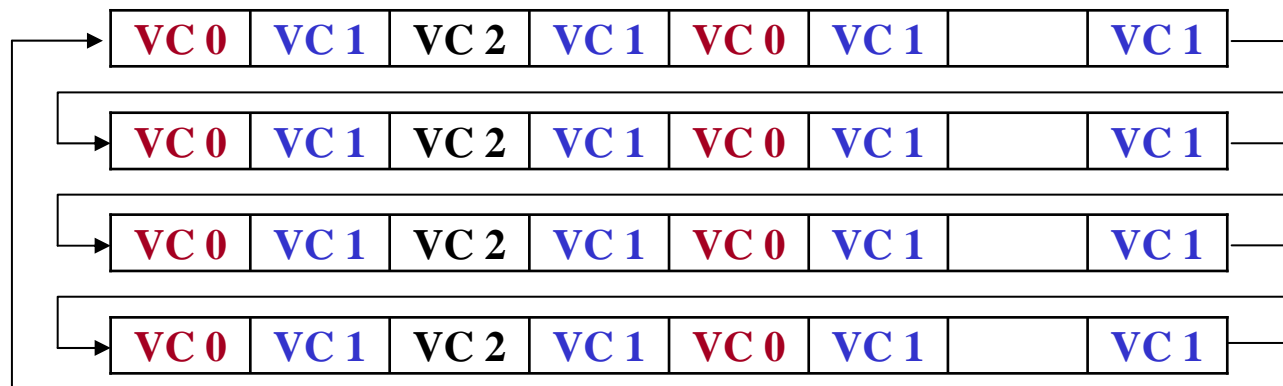
- In this kind of table-based schedulers is possible to control the **latency** of a flow or aggregated of flows by controlling the **maximum separation between any consecutive pair of entries** assigned to that flow.
  - This distance determines the maximum time that a packet at the head of a flow queue is going to wait until being transmitted.



# Deficit Table (DTable)

## Configuration methodology

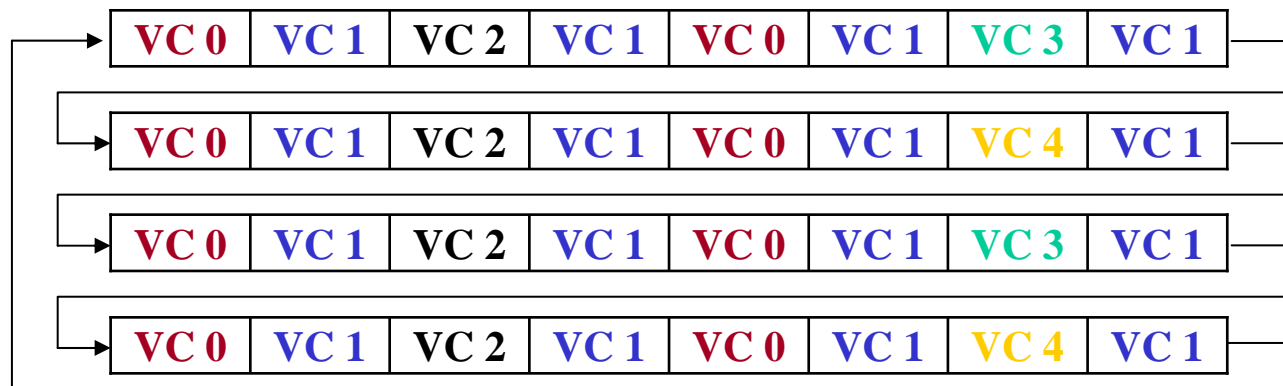
- In this kind of table-based schedulers is possible to control the **latency** of a flow or aggregated of flows by controlling the **maximum separation between any consecutive pair of entries** assigned to that flow.
  - This distance determines the maximum time that a packet at the head of a flow queue is going to wait until being transmitted.



# Deficit Table (DTable)

## Configuration methodology

- In this kind of table-based schedulers is possible to control the **latency** of a flow or aggregated of flows by controlling the **maximum separation between any consecutive pair of entries** assigned to that flow.
  - This distance determines the maximum time that a packet at the head of a flow queue is going to wait until being transmitted.



# Deficit Table (DTable)

## Configuration methodology

Maximum separation



Number/proportion of table entries



Assigned bandwidth

- If the flow requires more bandwidth, we can assign it more table entries.
- However, if the flow require less bandwidth we are wasting resources.
  - Flows which require a low latency usually require also little bandwidth.

# Deficit Table (DTable)

## Configuration methodology

<b>N</b>	Number of entries of the arbitration table
<b>MTU</b>	Maximum Transfer Unit of the network
<b><math>n_i</math></b>	Number of entries assigned to the $i^{\text{th}}$ flow
<b><math>\phi_i</math></b>	Bandwidth actually assigned to the $i^{\text{th}}$ flow
<b><math>\min\phi_i</math></b>	Minimum bandwidth assignable to the $i^{\text{th}}$ flow
<b><math>\max\phi_i</math></b>	Maximum bandwidth assignable to the $i^{\text{th}}$ flow
<b><math>w</math></b>	Maximum weight decoupling parameter
<b>M</b>	Maximum weight per table entry
<b>pool</b>	Bandwidth pool
<b>k</b>	Bandwidth pool decoupling parameter

<b>VC ID<sub>0</sub></b>	[MTU , M]
<b>VC ID<sub>1</sub></b>	[MTU , M]
<b>VC ID<sub>2</sub></b>	[MTU , M]
...	
<b>VC ID<sub>N-2</sub></b>	[MTU , M]
<b>VC ID<sub>N-1</sub></b>	[MTU , M]

[(N x MTU) , (N x M)]

$$pool \leq N \times M$$

$$M = MTU \times w$$

$$(k \leq w)$$

$$pool = N \times MTU \times k$$

$$\min \phi_i = \frac{n_i \times MTU}{pool} = \frac{n_i \times MTU}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{1}{k}$$

$$\max \phi_i = \frac{n_i \times MTU \times w}{pool} = \frac{n_i \times MTU \times w}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{w}{k}$$

# Deficit Table (DTable)

## Configuration methodology

<b>N</b>	Number of entries of the arbitration table
<b>MTU</b>	Maximum Transfer Unit of the network
<b><math>n_i</math></b>	Number of entries assigned to the $i^{\text{th}}$ flow
<b><math>\phi_i</math></b>	Bandwidth actually assigned to the $i^{\text{th}}$ flow
<b><math>\min\phi_i</math></b>	Minimum bandwidth assignable to the $i^{\text{th}}$ flow
<b><math>\max\phi_i</math></b>	Maximum bandwidth assignable to the $i^{\text{th}}$ flow
<b>w</b>	Maximum weight decoupling parameter
<b>M</b>	Maximum weight per table entry
<b>pool</b>	Bandwidth pool
<b>k</b>	Bandwidth pool decoupling parameter

<b>VC ID<sub>0</sub></b>	[MTU , M]
<b>VC ID<sub>1</sub></b>	[MTU , M]
<b>VC ID<sub>2</sub></b>	[MTU , M]
...	
<b>VC ID<sub>N-2</sub></b>	[MTU , M]
<b>VC ID<sub>N-1</sub></b>	[MTU , M]
<hr/> [(N x MTU) , (N x M)]  $pool \leq N \times M$	

$$M = MTU \times w$$

$$(k \leq w)$$

$$pool = N \times MTU \times k$$

$$\min \phi_i = \frac{n_i \times MTU}{pool} = \frac{n_i \times MTU}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{1}{k}$$

$$\max \phi_i = \frac{n_i \times MTU \times w}{pool} = \frac{n_i \times MTU \times w}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{w}{k}$$

# Deficit Table (DTable)

## Configuration methodology

<b>N</b>	Number of entries of the arbitration table
<b>MTU</b>	Maximum Transfer Unit of the network
<b><math>n_i</math></b>	Number of entries assigned to the $i^{\text{th}}$ flow
<b><math>\phi_i</math></b>	Bandwidth actually assigned to the $i^{\text{th}}$ flow
<b><math>\min\phi_i</math></b>	Minimum bandwidth assignable to the $i^{\text{th}}$ flow
<b><math>\max\phi_i</math></b>	Maximum bandwidth assignable to the $i^{\text{th}}$ flow
<b>w</b>	Maximum weight decoupling parameter
<b>M</b>	Maximum weight per table entry
<b>pool</b>	Bandwidth pool
<b>k</b>	Bandwidth pool decoupling parameter

<b>VC ID<sub>0</sub></b>	[MTU , M]
<b>VC ID<sub>1</sub></b>	[MTU , M]
<b>VC ID<sub>2</sub></b>	[MTU , M]
...	
<b>VC ID<sub>N-2</sub></b>	[MTU , M]
<b>VC ID<sub>N-1</sub></b>	[MTU , M]
-----	
[(N x MTU) , (N x M)]	
<b>pool ≤ N x M</b>	

$$M = MTU \times w$$

$$(k \leq w)$$

$$pool = N \times MTU \times k$$

$$\min \phi_i = \frac{n_i \times MTU}{pool} = \frac{n_i \times MTU}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{1}{k}$$

$$\max \phi_i = \frac{n_i \times MTU \times w}{pool} = \frac{n_i \times MTU \times w}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{w}{k}$$

# Deficit Table (DTable)

## Configuration methodology

<b>N</b>	Number of entries of the arbitration table
<b>MTU</b>	Maximum Transfer Unit of the network
<b><math>n_i</math></b>	Number of entries assigned to the $i^{\text{th}}$ flow
<b><math>\phi_i</math></b>	Bandwidth actually assigned to the $i^{\text{th}}$ flow
<b><math>\min\phi_i</math></b>	Minimum bandwidth assignable to the $i^{\text{th}}$ flow
<b><math>\max\phi_i</math></b>	Maximum bandwidth assignable to the $i^{\text{th}}$ flow
<b><math>w</math></b>	Maximum weight decoupling parameter
<b>M</b>	Maximum weight per table entry
<b>pool</b>	Bandwidth pool
<b>k</b>	Bandwidth pool decoupling parameter

<b>VC ID<sub>0</sub></b>	[MTU , M]
<b>VC ID<sub>1</sub></b>	[MTU , M]
<b>VC ID<sub>2</sub></b>	[MTU , M]
...	
<b>VC ID<sub>N-2</sub></b>	[MTU , M]
<b>VC ID<sub>N-1</sub></b>	[MTU , M]

[(N x MTU) , (N x M)]

$pool \leq N \times M$

$$M = MTU \times w$$

$$(k \leq w)$$

$$pool = N \times MTU \times k$$

$$\min \phi_i = \frac{n_i \times MTU}{pool} = \frac{n_i \times MTU}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{1}{k}$$

$$\max \phi_i = \frac{n_i \times MTU \times w}{pool} = \frac{n_i \times MTU \times w}{N \times MTU \times k} = \frac{n_i}{N} \times \frac{w}{k}$$

# Deficit Table (DTable)

## Configuration methodology

- We fix the maximum separation between any consecutive pair of entries of each flow attending to the latency requirements of the flows.
- We fix the weights assigned to each table entry attending to the bandwidth requirements.
  - This methodology allows us to assign each flow a bandwidth that depends not only on the assigned proportion of table entries, but also, on two configuration parameters.

# Deficit Table (DTable)

## Configuration methodology

VC	max. dist	#entries	%entries	$\min\phi_i$	$\max\phi_i$
D2	2	32	50	25	100
D4	4	16	25	12.5	50
D8	8	8	12.5	6.25	25
D16	16	4	6.25	3.125	12.5
D32	32	2	3.13	1.5625	6.25
D64	64	1	1.56	0.78125	3.125
D64'	64	1	1.56	0.78125	3.125
Total		64	100	50	200

$$k = 2, w = 4$$

# Deficit Table (DTable)

## Configuration methodology

VC	max. dist	#entries	%entries	$\min\phi_i$	$\max\phi_i$
D2	2	32	50	25	100
<b>D4</b>	<b>4</b>	<b>16</b>	<b>25</b>	<b>12.5</b>	<b>50</b>
D8	8	8	12.5	6.25	25
D16	16	4	6.25	3.125	12.5
D32	32	2	3.13	1.5625	6.25
D64	64	1	1.56	0.78125	3.125
D64'	64	1	1.56	0.78125	3.125
Total		64	100	50	200

$$k = 2, w = 4$$

# Deficit Table (DTable)

## Configuration methodology

VC	max. dist	#entries	%entries	$\min\phi_i$	$\max\phi_i$
D2	2	32	50	25	100
<b>D4</b>	<b>4</b>	<b>16</b>	<b>25</b>	<b>12.5</b>	<b>50</b>
D8	8	8	12.5	6.25	25
D16	16	4	6.25	3.125	12.5
D32	32	2	3.13	1.5625	6.25
D64	64	1	1.56	0.78125	3.125
D64'	64	1	1.56	0.78125	3.125
Total		64	100	50	200

$$k = 2, w = 4$$

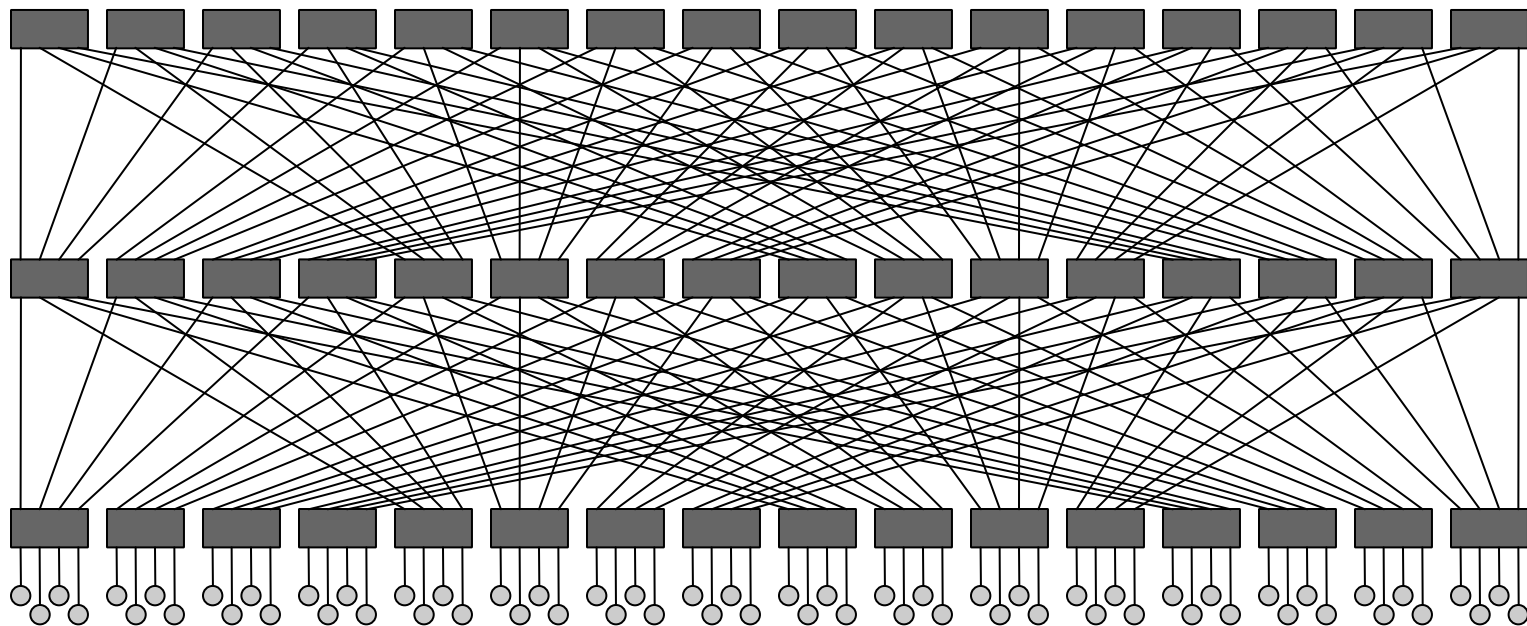
*We have partially decoupled the bandwidth and latency assignation*

# Outline

- The Deficit Round Robin (DRR) scheduler
- The Deficit Table (DTable) scheduler
  - The DTable scheduling mechanism
  - Configuring the DTable scheduler
- **Performance evaluation**
- Conclusions

# Performance evaluation

- We have employed a network simulator based on the Advanced Switching technology.
- Topology: Multistage with 64 nodes.



# Performance evaluation

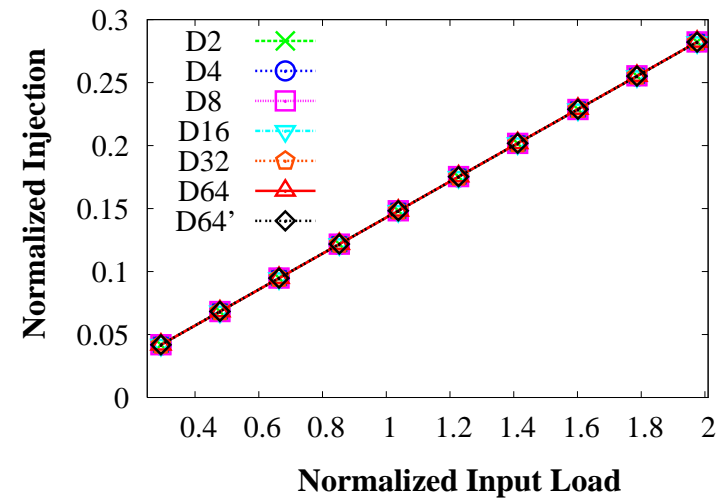
VC	$\phi_i$	DTable				DRR
		Max. dist.	%entries	Total weight	Weight per entry	Quantum
D2	25	2	50	1024	32	256
D4	25	4	25	1024	64	256
D8	25	8	12.5	1024	128	256
D16	12.5	16	6.25	512	128	128
D32	6.25	32	3.13	256	128	64
D64	3.125	64	1.56	128	128	32
D64'	3.125	64	1.56	128	128	32
Total	100		100	4096		1024

# Performance evaluation

VC	$\phi_i$	DTable				DRR
		Max. dist.	%entries	Total weight	Weight per entry	Quantum
D2	25	2	50	1024	32	256
D4	25	4	25	1024	64	256
D8	25	8	12.5	1024	128	256
D16	12.5	16	6.25	512	128	128
D32	6.25	32	3.13	256	128	64
D64	3.125	64	1.56	128	128	32
D64'	3.125	64	1.56	128	128	32
Total	100		100	4096		1024

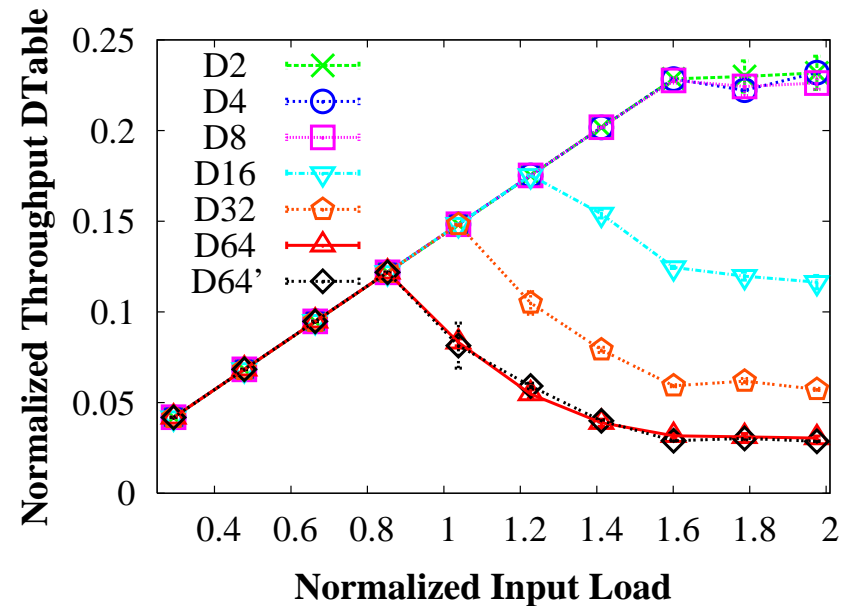
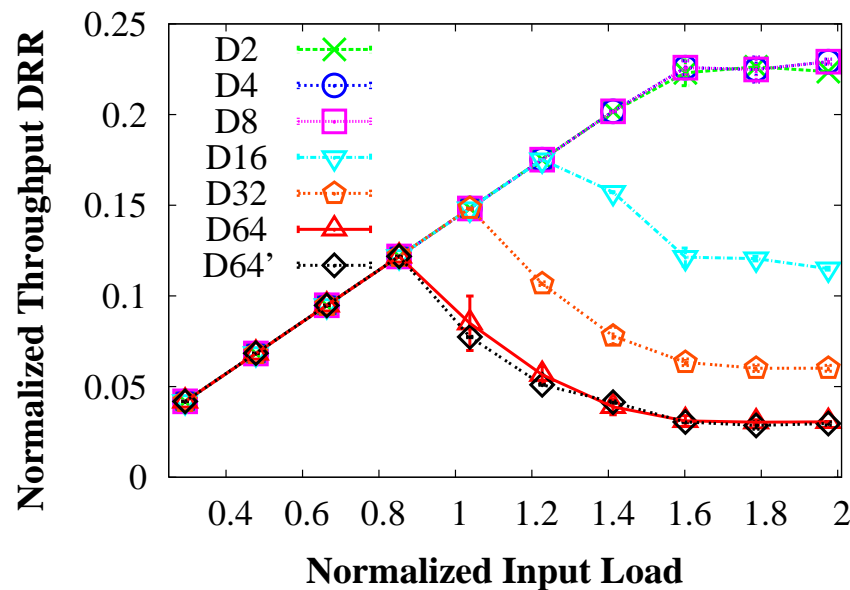
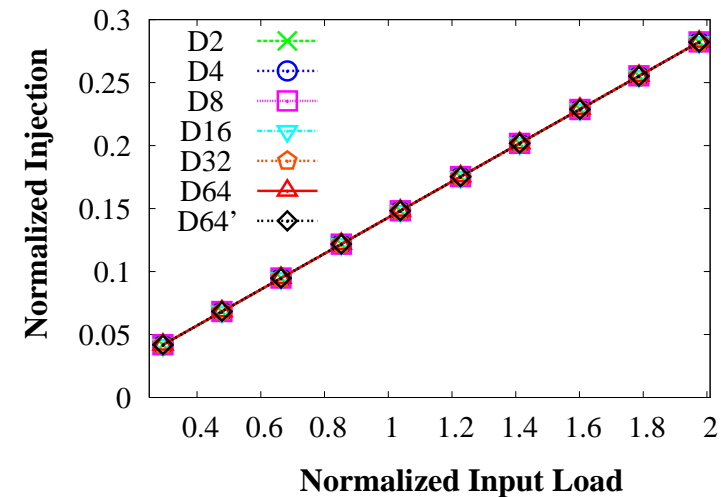
# Performance evaluation

VC	$\phi_i$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



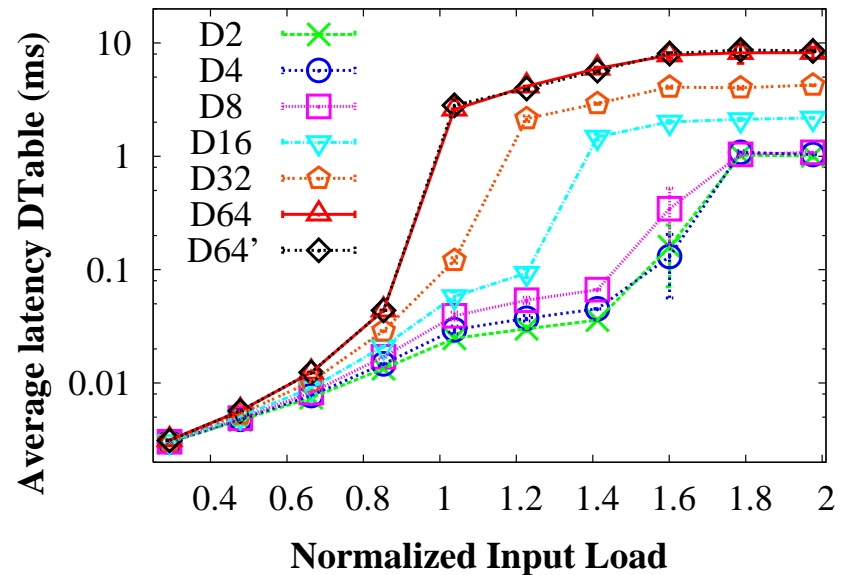
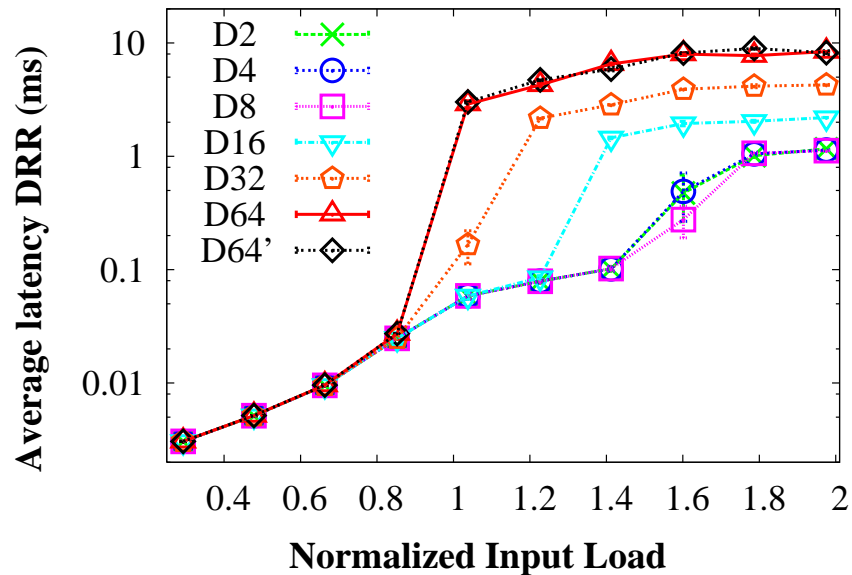
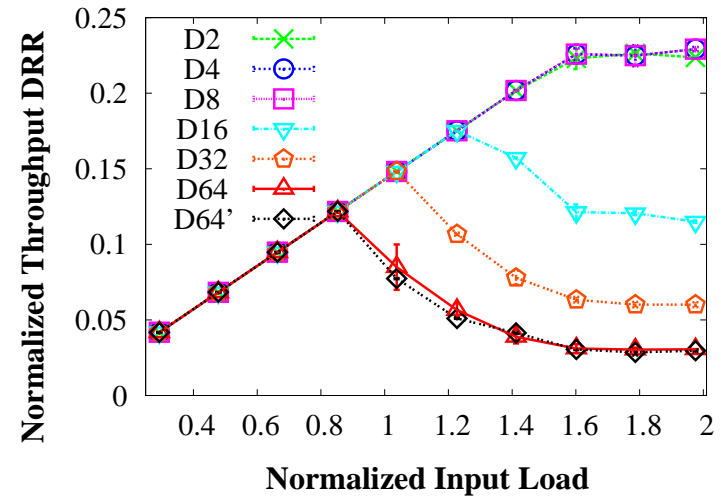
# Performance evaluation

VC	$\phi_i$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



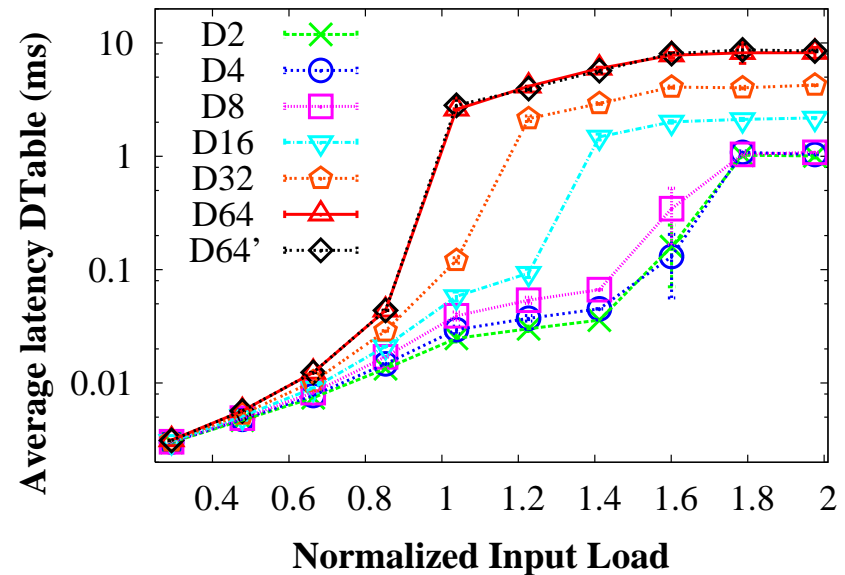
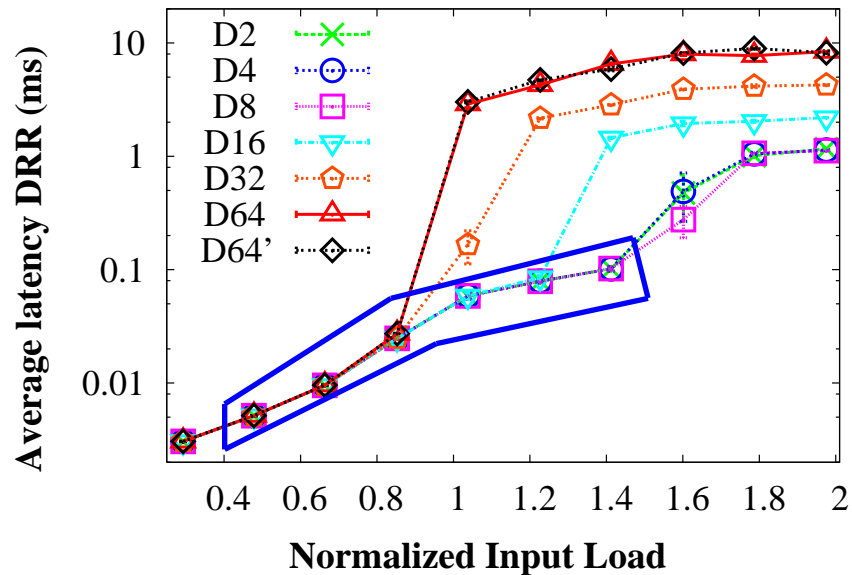
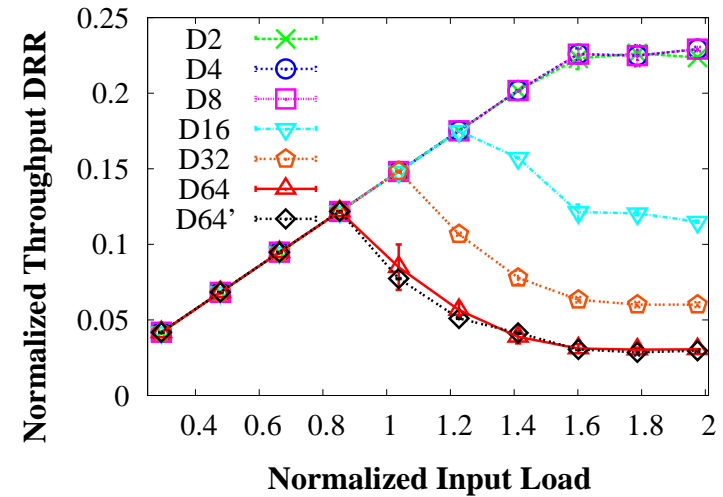
# Performance evaluation

VC	$\phi_1$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



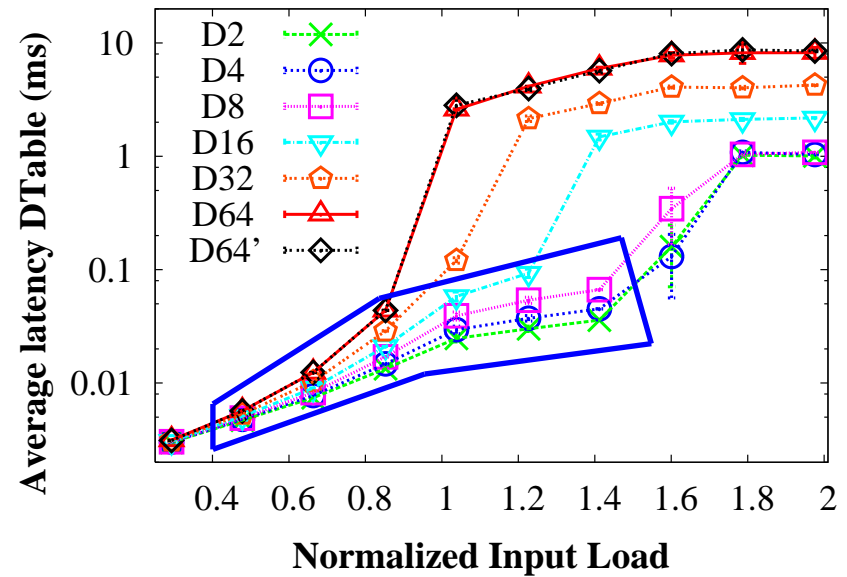
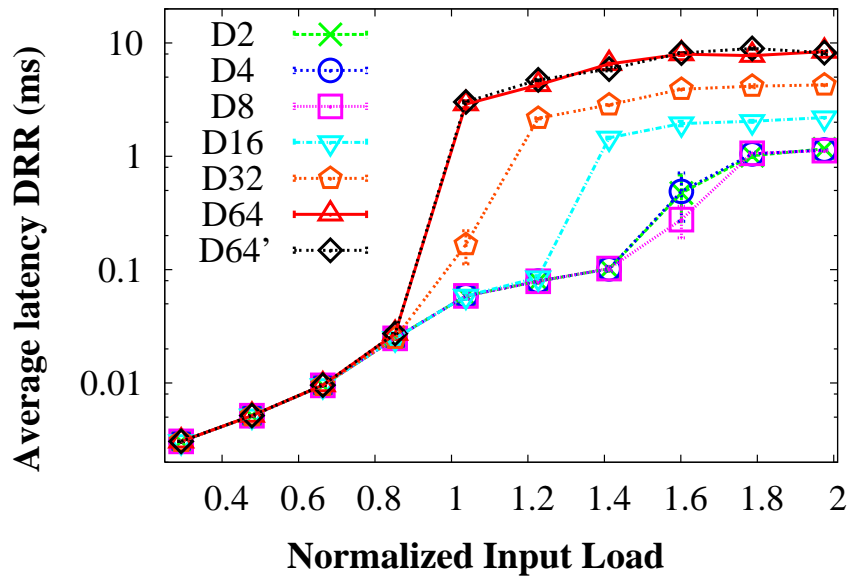
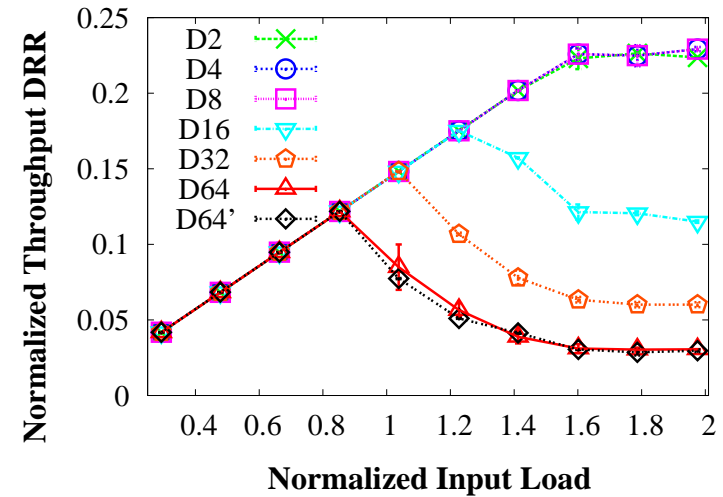
# Performance evaluation

VC	$\phi_1$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



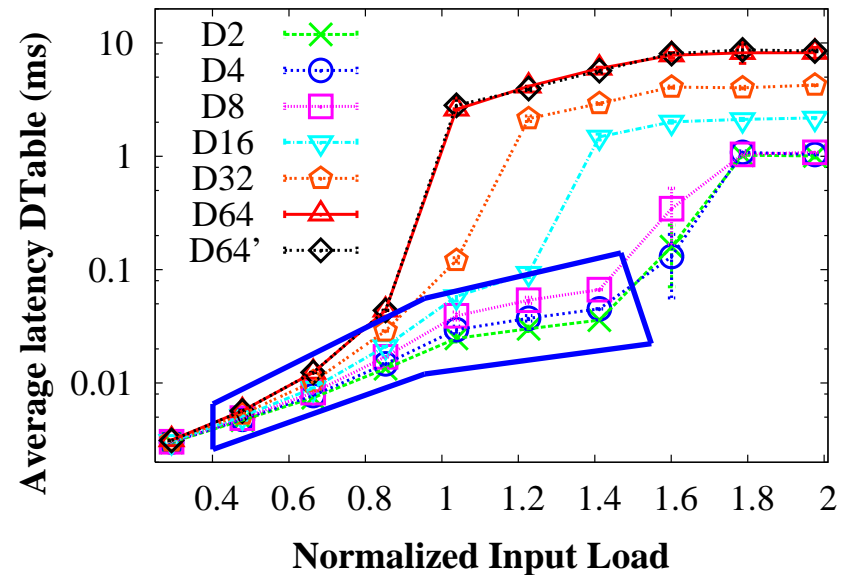
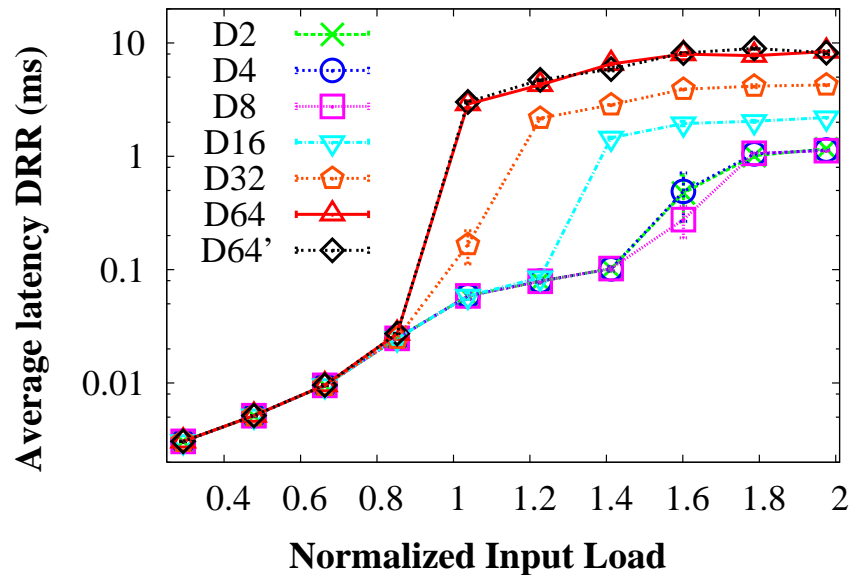
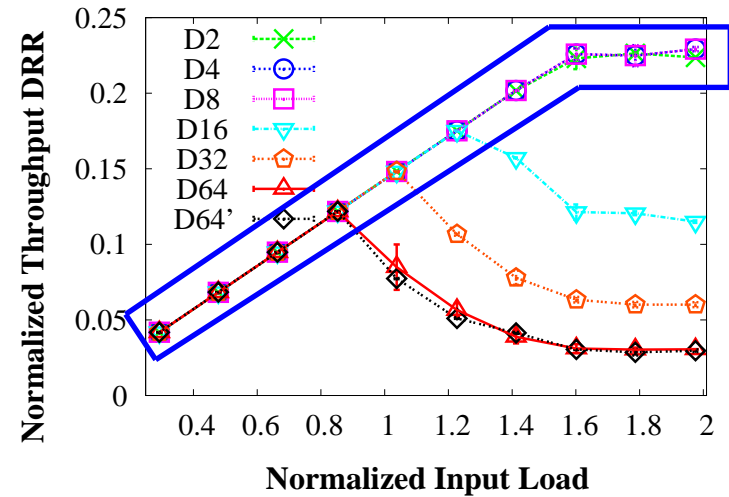
# Performance evaluation

VC	$\phi_1$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



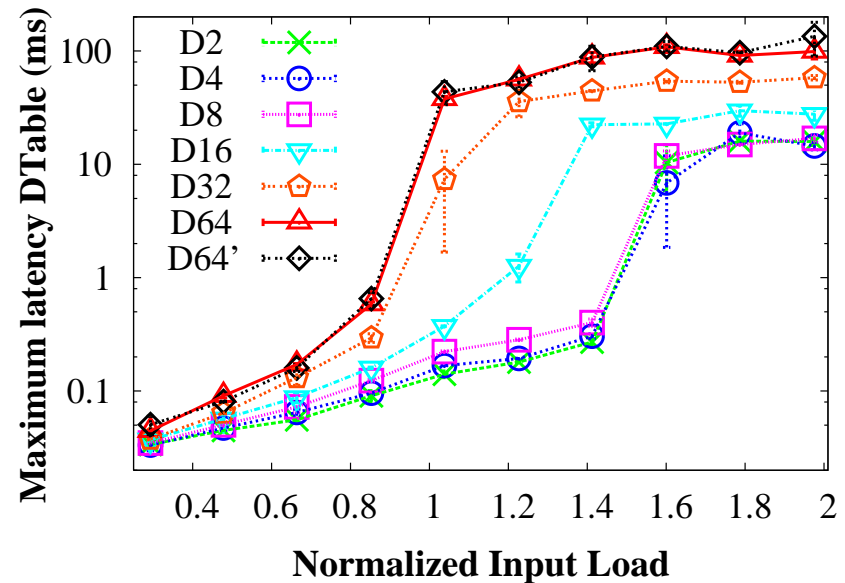
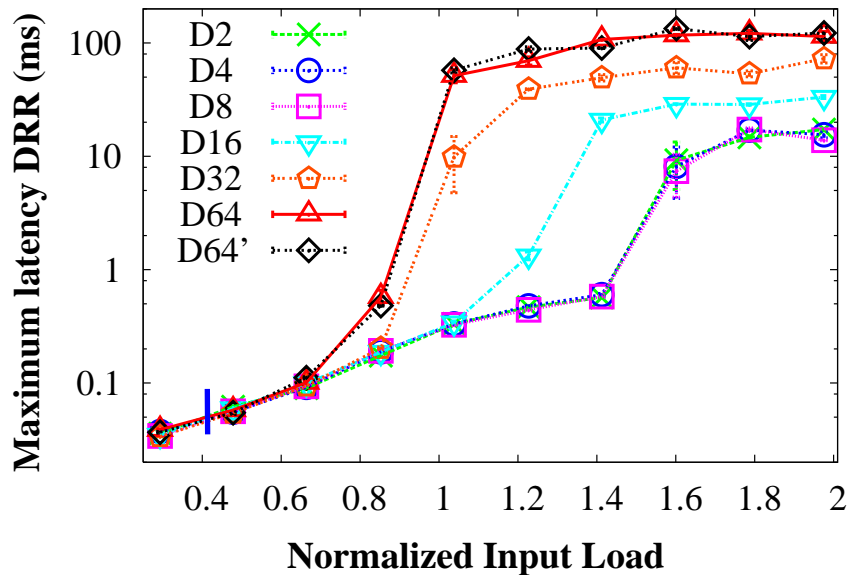
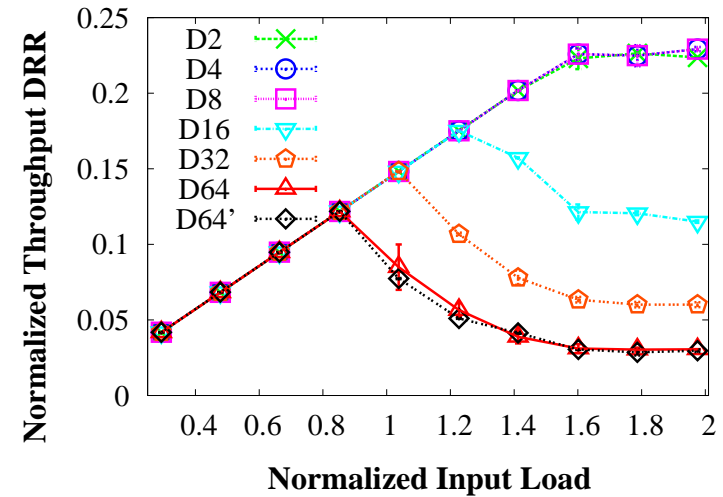
# Performance evaluation

VC	$\phi_1$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



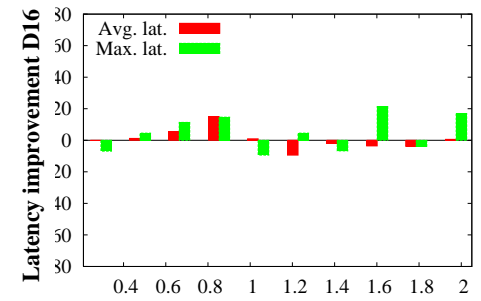
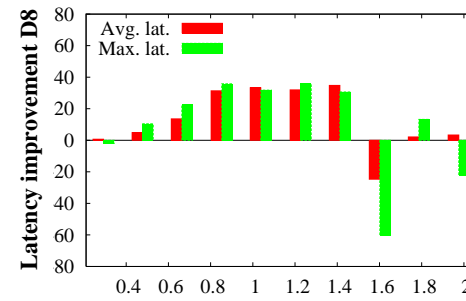
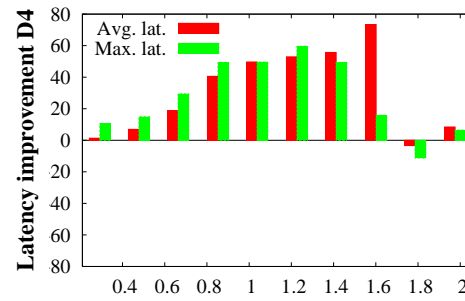
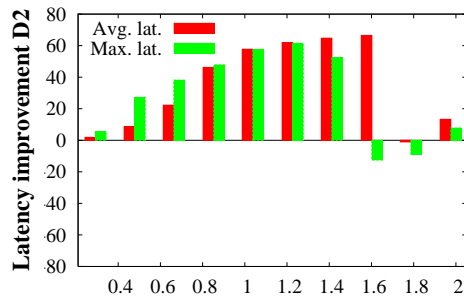
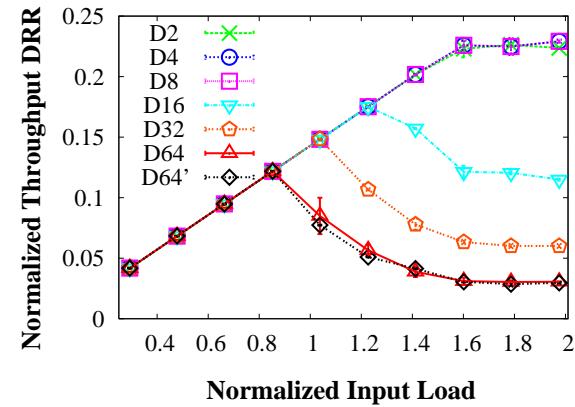
# Performance evaluation

VC	$\phi_i$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32



# Performance evaluation

VC	$\phi_1$	DTable		DRR
		Max. dist.	Total weight	Quantum
D2	25	2	1024	256
D4	25	4	1024	256
D8	25	8	1024	256
D16	12.5	16	512	128
D32	6.25	32	256	64
D64	3.125	64	128	32
D64'	3.125	64	128	32

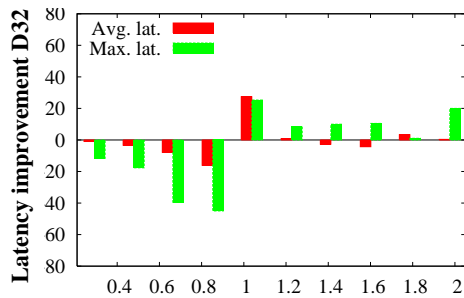


Normalized Input Load

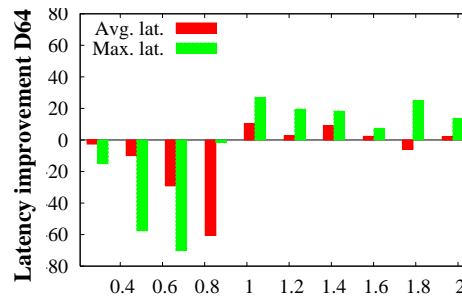
Normalized Input Load

Normalized Input Load

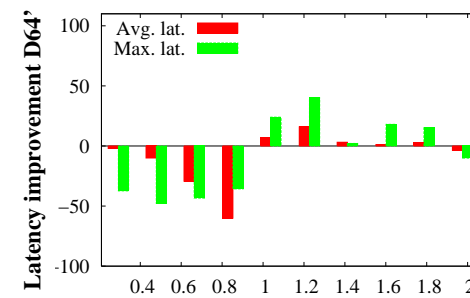
Normalized Input Load



Normalized Input Load



Normalized Input Load



Normalized Input Load

# Outline

- The Deficit Round Robin (DRR) scheduler
- The Deficit Table (DTable) scheduler
  - The DTable scheduling mechanism
  - Configuring the DTable scheduler
- Performance evaluation
- **Conclusions**

# Conclusions

- Deficit Round Robin →

-Very low complexity  
-Very high latency

- Table based schedulers

- Advanced Switching
- InfiniBand
- **Deficit Table** →

Works in a proper way  
with variable packet size

- Configuration methodology →

Partial decoupling between:  
**Bandwidth** and **Latency**

- Latency performance comparison:



-DTable provides a better latency performance to those flows with higher latency requirements  
- It provides a worse latency performance to those flows with lower latency requirements

# Comparing the latency performance of the DTable and DRR schedulers

Raúl Martínez

Francisco J. Alfaro

José L. Sánchez

