Impossibility Results for Data-Center Routing with Congestion Control and Unsplittable Flows

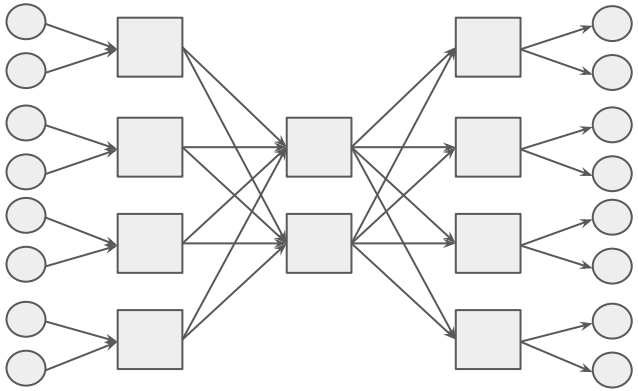
Miguel Alves Ferreira
Carnegie Mellon, Instituto de Telecomunicações, and Instituto Superior Técnico

Nirav Atre, Justine Sherry
Carnegie Mellon

João Luís Sobrinho
Instituto de Telecomunicações, and Instituto Superior Técnico

Most data-centers are architectured after (folded) Clos networks

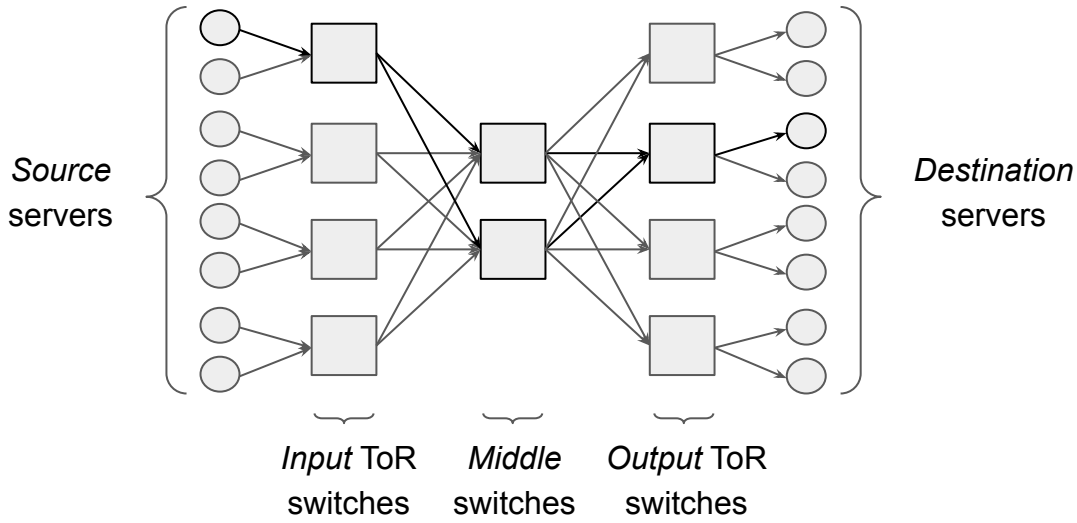
[Greenberg et at. 09, Roy et al. 15, Singh et al. 15]



Clos network

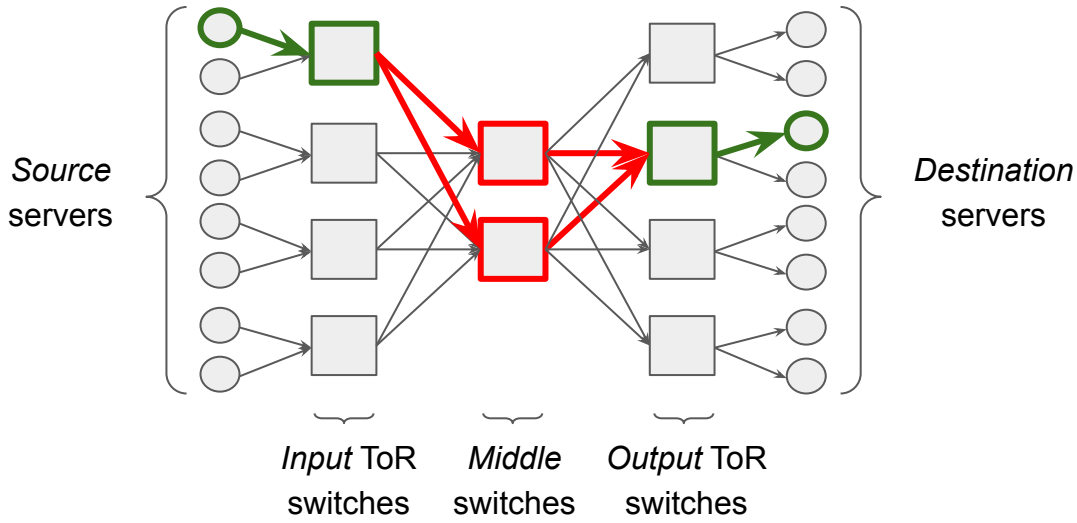
Most data-centers are architectured after (folded) Clos networks

[Greenberg et at. 09, Roy et al. 15, Singh et al. 15]



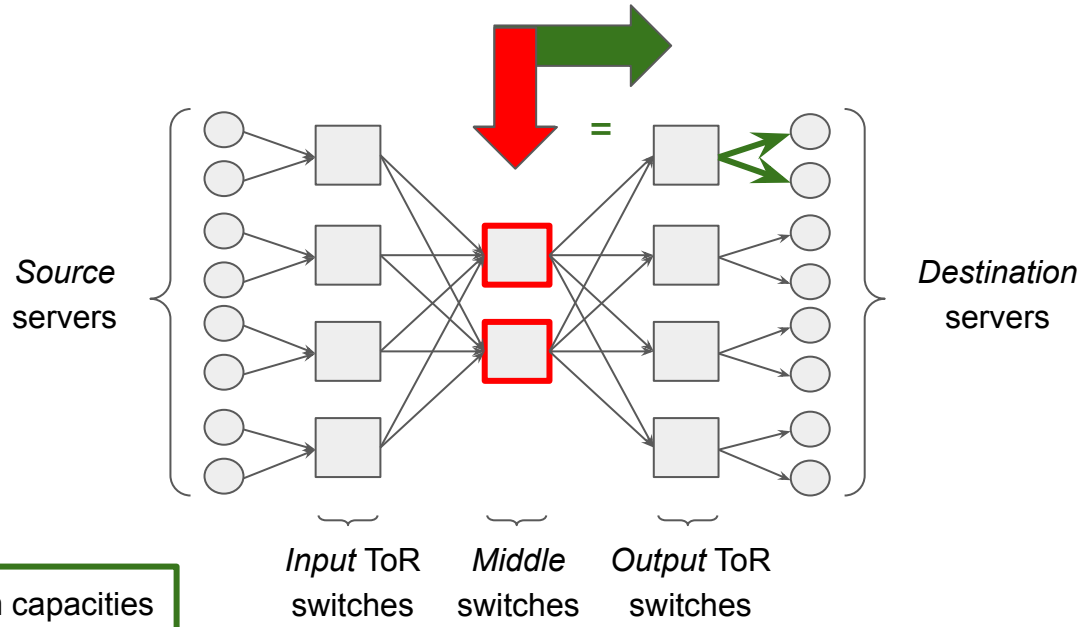
Most data-centers are architectured after (folded) Clos networks

[Greenberg et at. 09, Roy et al. 15, Singh et al. 15]



Most data-centers are architectured after (folded) Clos networks

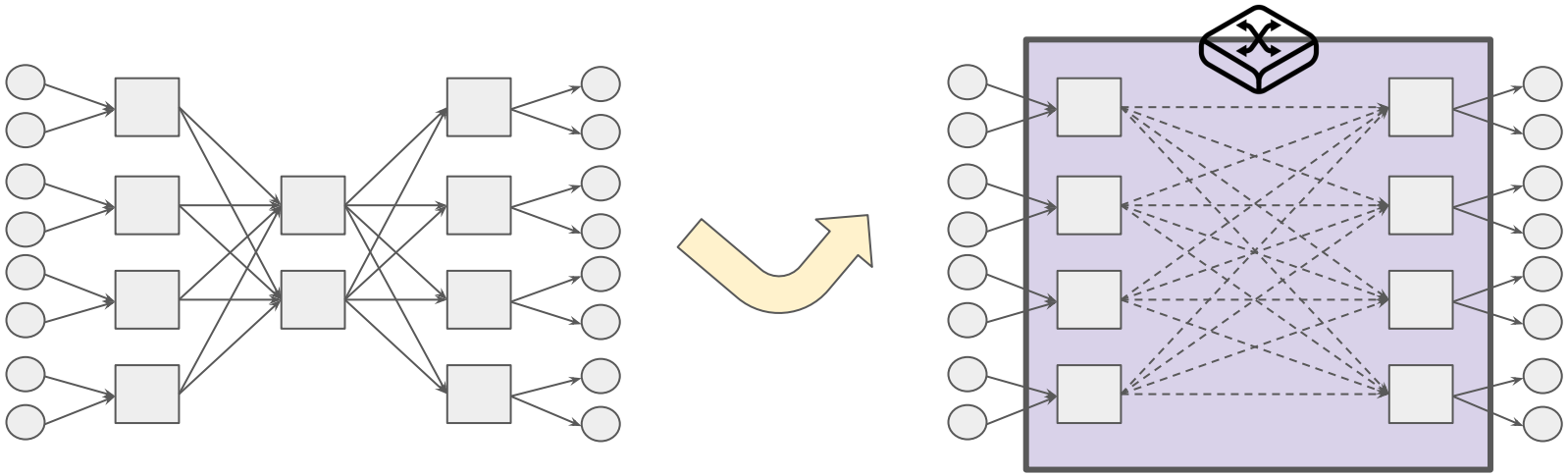
[Greenberg et at. 09, Roy et al. 15, Singh et al. 15]



Links have uniform capacities

Under certain traffic assumptions, Clos networks are *equivalent* to a macro-switch (hose model, fully-non-blocking switch)

[Duffield et at. 99, Alizadeh et al. 12, Namyar et al. 21]



Clos network

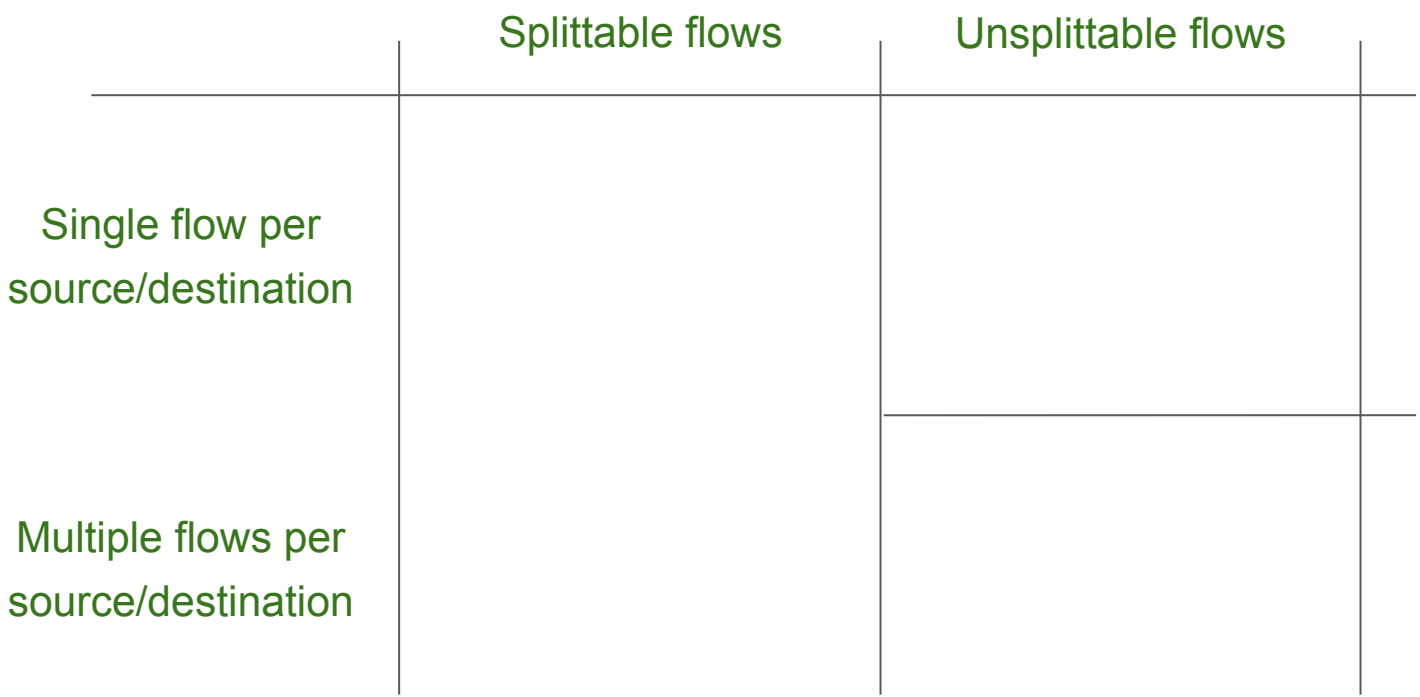
Macro-switch abstraction

Traffic nomenclature

- Flow: Source-destination pair
  - Possibly multiple flows mapping to same source-destination pair
- Routing: Assignment from flows to source-destination paths
- Rate allocation: Assignment from flows to non-negative rates
- Throughput: Total rate over all input flows

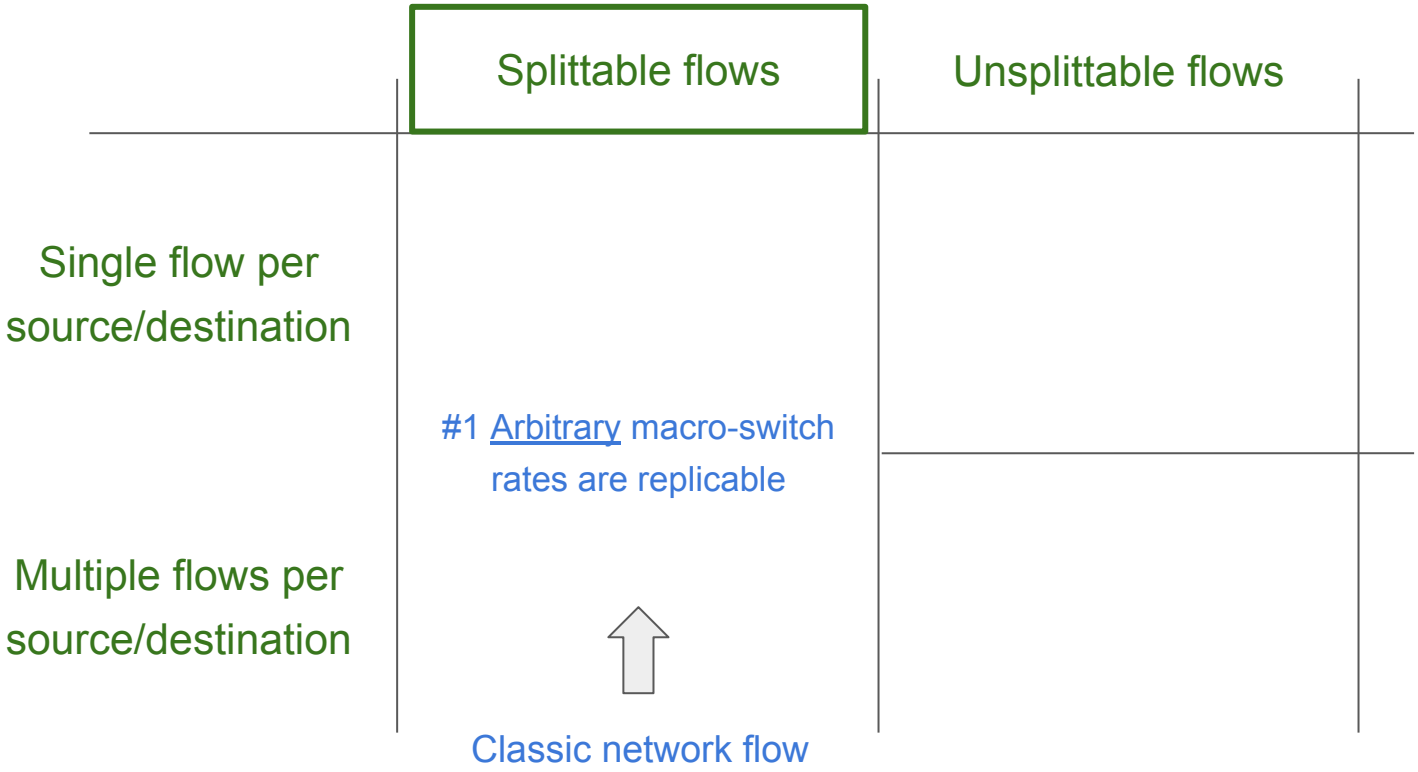
Prior results for splittable flows and admission control

[Chiesa et at. 17, Hwang 83]



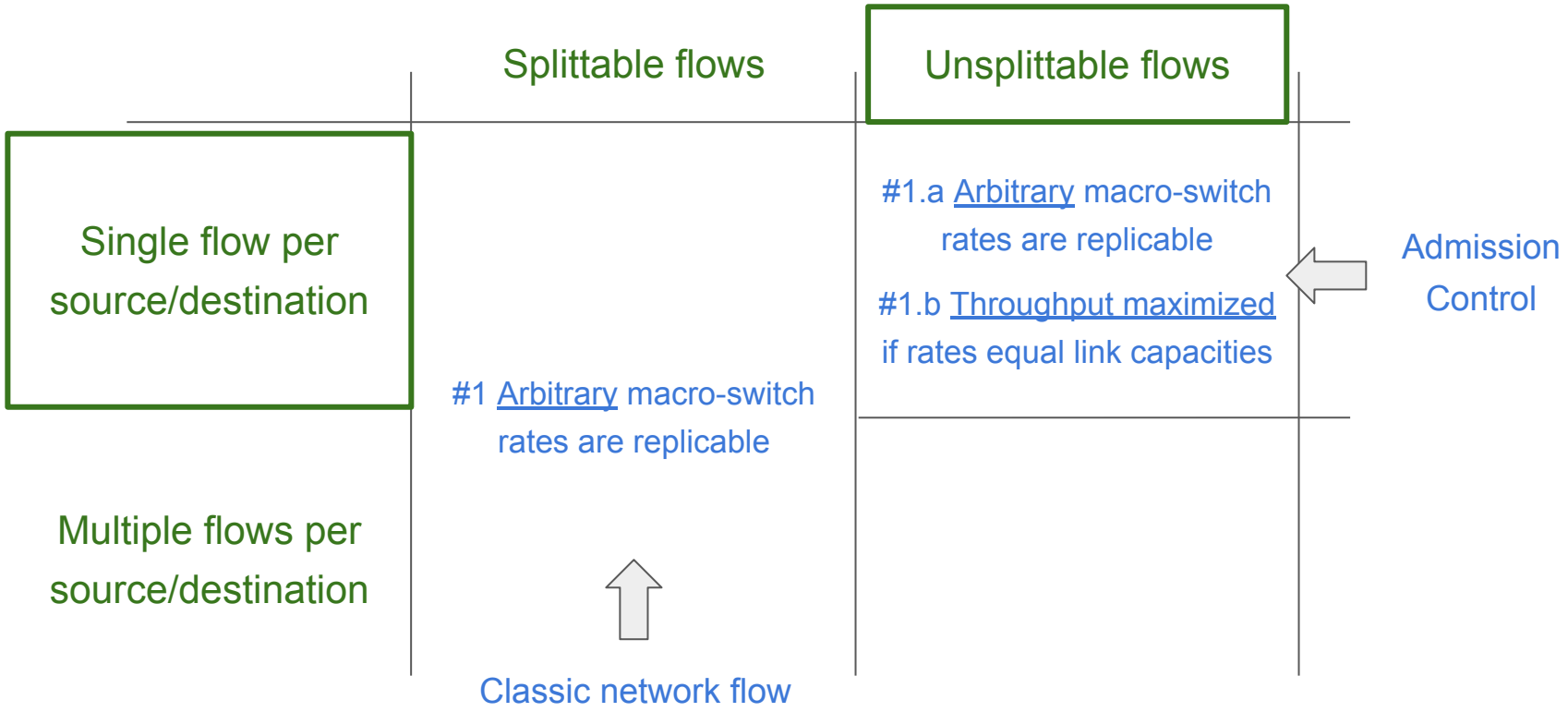
Prior results for splittable flows and admission control

[Chiesa et at. 17, Hwang 83]



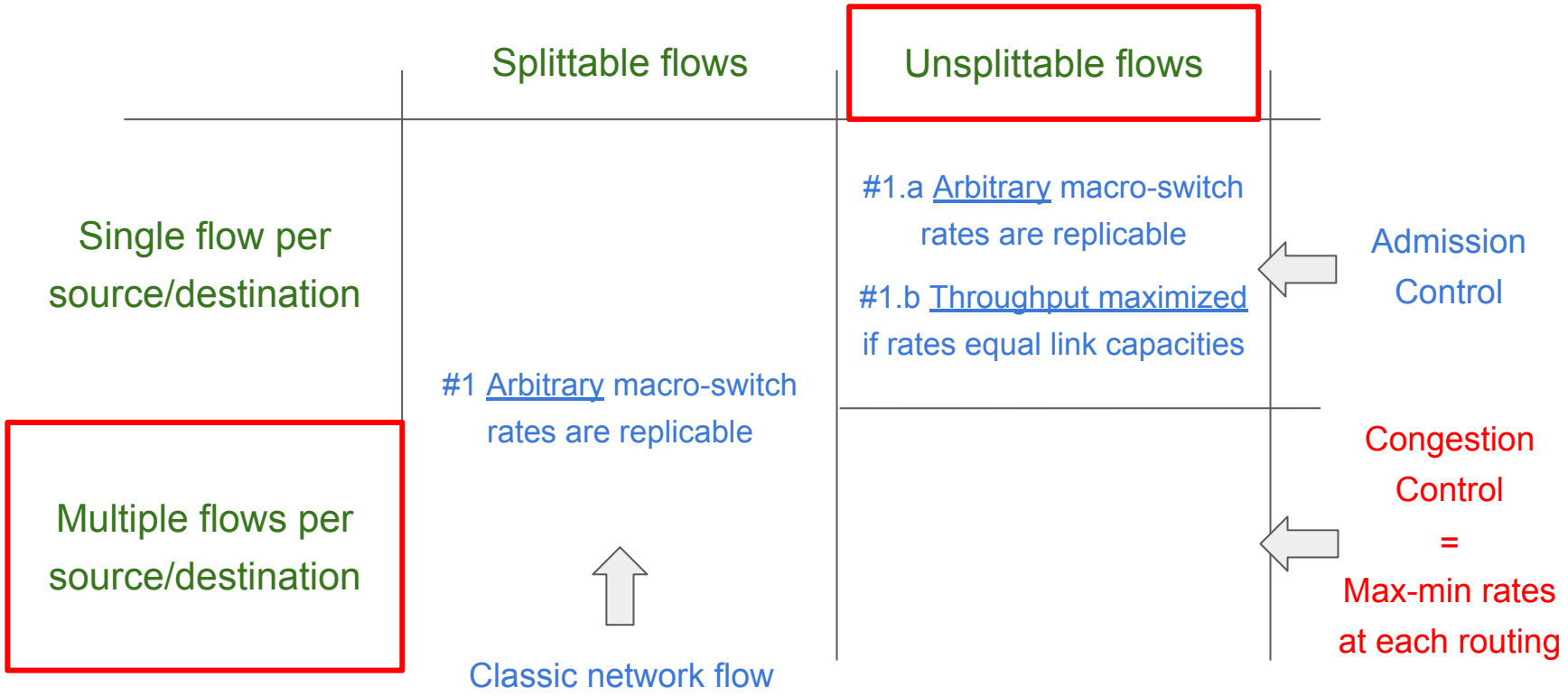
Prior results for splittable flows and admission control

[Chiesa et at. 17, Hwang 83]



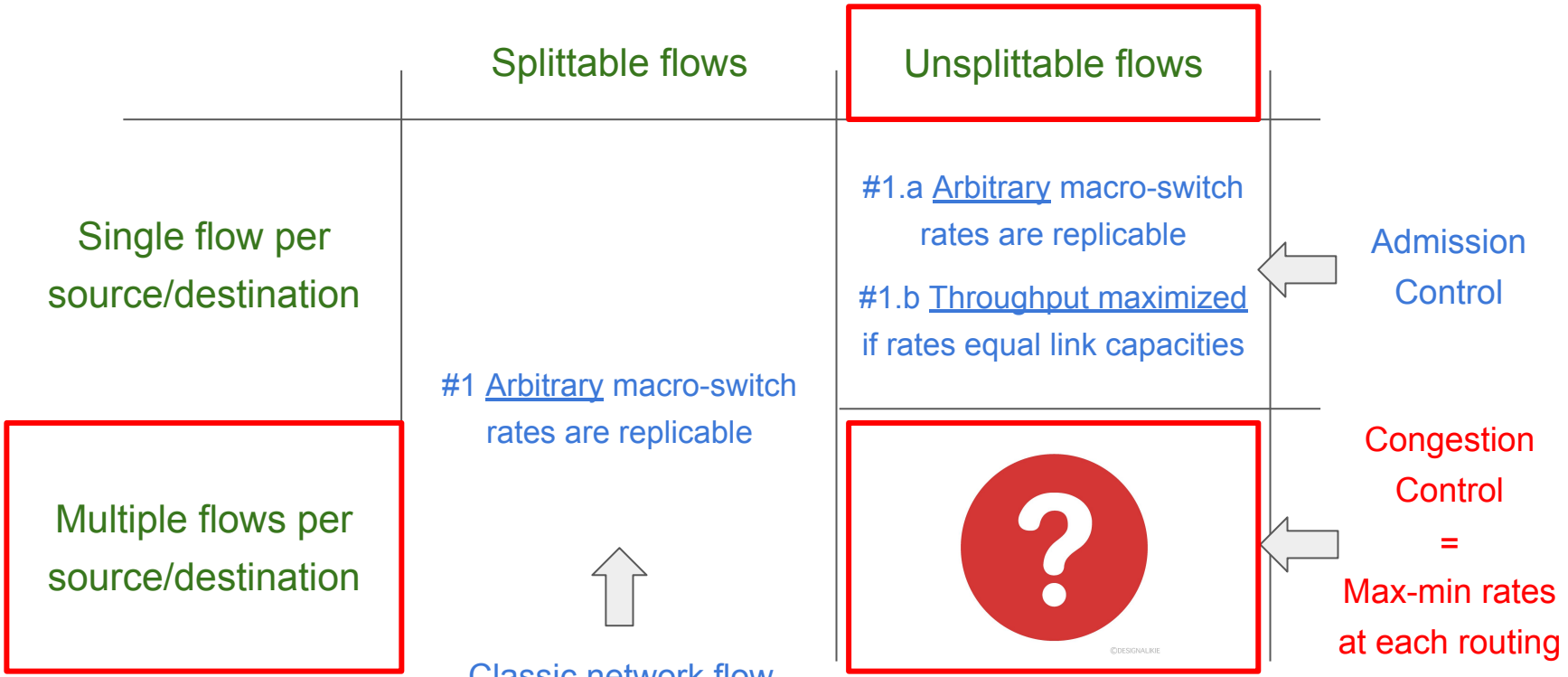
Prior results for splittable flows and admission control

[Chiesa et at. 17, Hwang 83]



Prior results for splittable flows and admission control

[Chiesa et at. 17, Hwang 83]



Impossibility results for unsplittable flows and congestion control

Impossibility results for unsplittable flows and congestion control

- **Result #1:** In macro-switch, imposing max-min fair rates reduces throughput by ½ relative to maximum throughput (without max-min fair rates)
  - **Implication #1:** Congestion control introduces significant throughput loss

Impossibility results for unsplittable flows and congestion control

- **Result #1:** In macro-switch, imposing max-min fair rates reduces throughput by ½ relative to maximum throughput (without max-min fair rates)
  - **Implication #1:** Congestion control introduces significant throughput loss

- **Result #2:** Routing for max-min fairness reduces max-min fair rates of some flows by **1/n** relative to macro-switch, where n is number of middle switches
  - **Implication #2:** Optimizing fairness cannot ensure macro-switch abstraction for fairness

Impossibility results for unsplittable flows and congestion control

- **Result #1:** In macro-switch, imposing max-min fair rates reduces throughput by ½ relative to maximum throughput (without max-min fair rates)
  - **Implication #1:** Congestion control introduces significant throughput loss

- **Result #2:** Routing for max-min fairness reduces max-min fair rates of some flows by **1/n** relative to macro-switch, where n is number of middle switches
  - **Implication #2:** Optimizing fairness cannot ensure macro-switch abstraction for fairness

- **Result #3:** Routing for throughput increases throughput by **2** relative to macro-switch, but reduces max-min fair rates of most flows to **0**
  - **Implication #3:** Throughput should not be primary routing objective

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

- First, maximizes lowest rate assigned to some flow. Second, assuming lowest rate is fixed, maximizes second lowest rate assigned to some flow. And so on

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

- First, maximizes lowest rate assigned to some flow. Second, assuming lowest rate is fixed, maximizes second lowest rate assigned to some flow. And so on

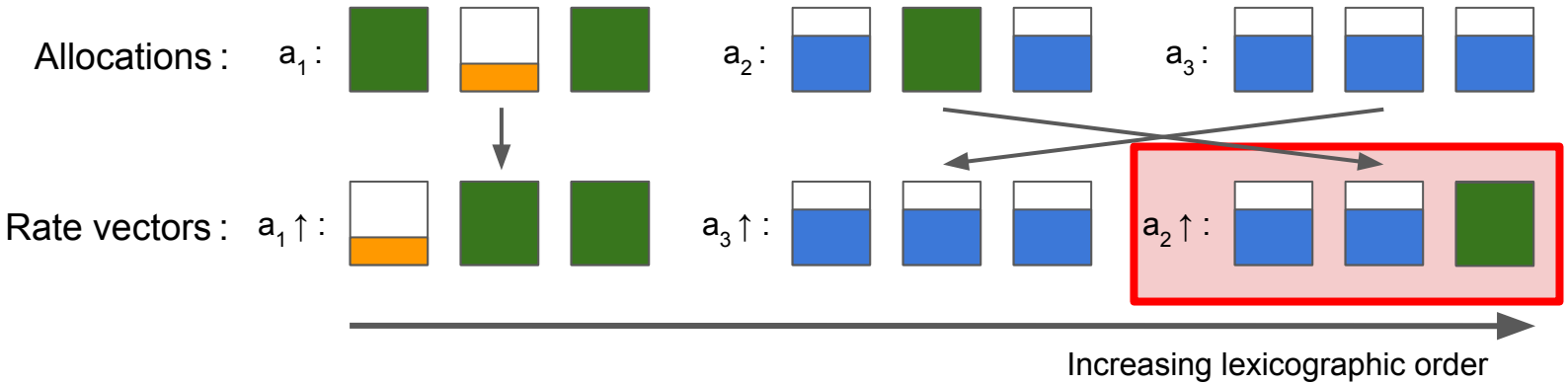
- **Definition**: The max-min fair allocation (MmF) maximizes in lexicographic (lex.) order vectors whose components are flow rates ordered from lowest to highest

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

- First, maximizes lowest rate assigned to some flow. Second, assuming lowest rate is fixed, maximizes second lowest rate assigned to some flow. And so on

- **Definition**: The max-min fair allocation (MmF) maximizes in lexicographic (lex.) order vectors whose components are flow rates ordered from lowest to highest

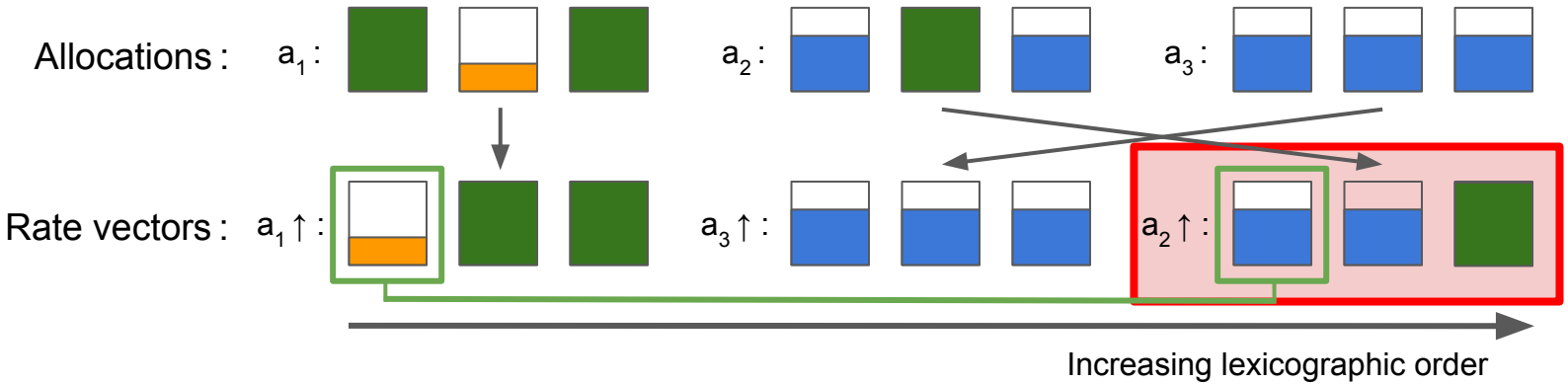


Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

- First, maximizes lowest rate assigned to some flow. Second, assuming lowest rate is fixed, maximizes second lowest rate assigned to some flow. And so on

- **Definition**: The max-min fair allocation (MmF) maximizes in lexicographic (lex.) order vectors whose components are flow rates ordered from lowest to highest

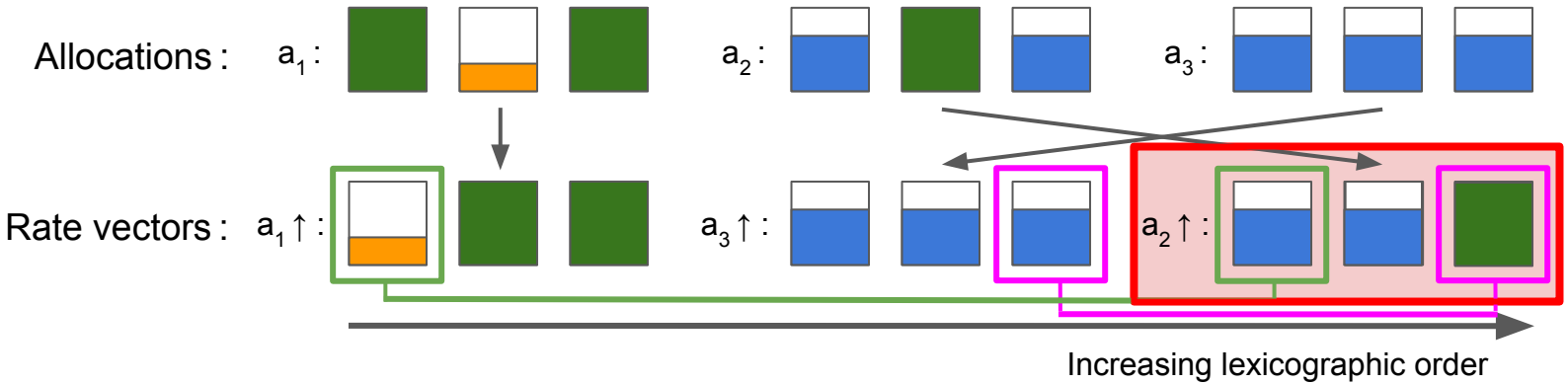


Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

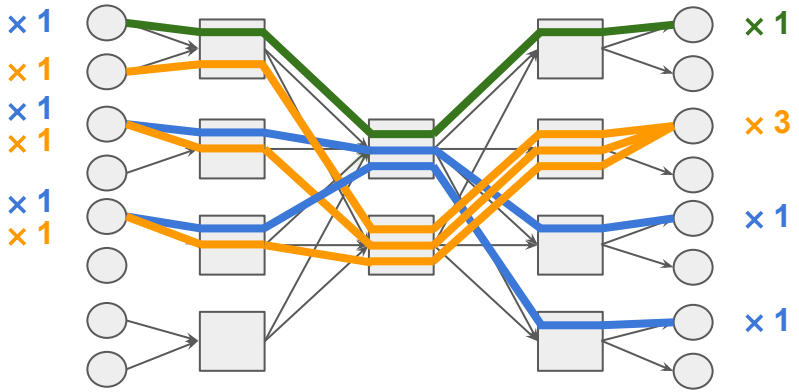
- First, maximizes lowest rate assigned to some flow. Second, assuming lowest rate is fixed, maximizes second lowest rate assigned to some flow. And so on

- **Definition**: The max-min fair allocation (MmF) maximizes in lexicographic (lex.) order vectors whose components are flow rates ordered from lowest to highest



Max-min fair allocations for *fixed* routing

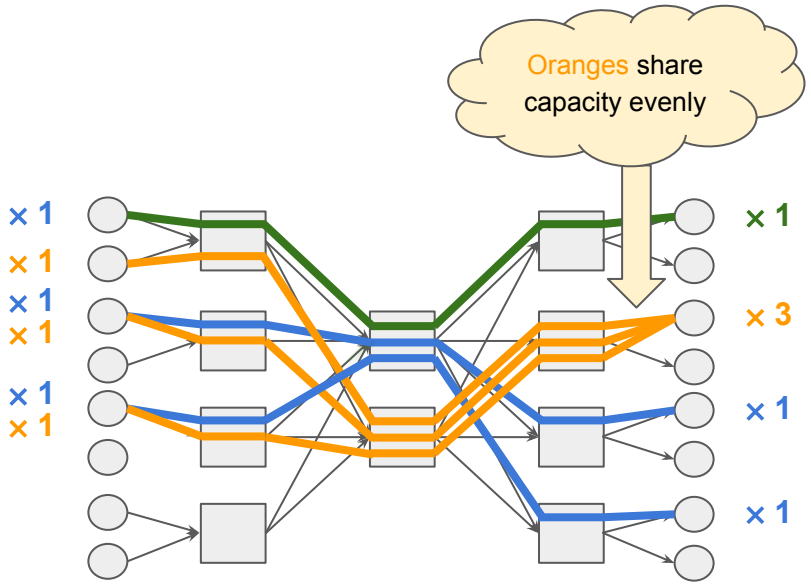
[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]



Links have capacity 1

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

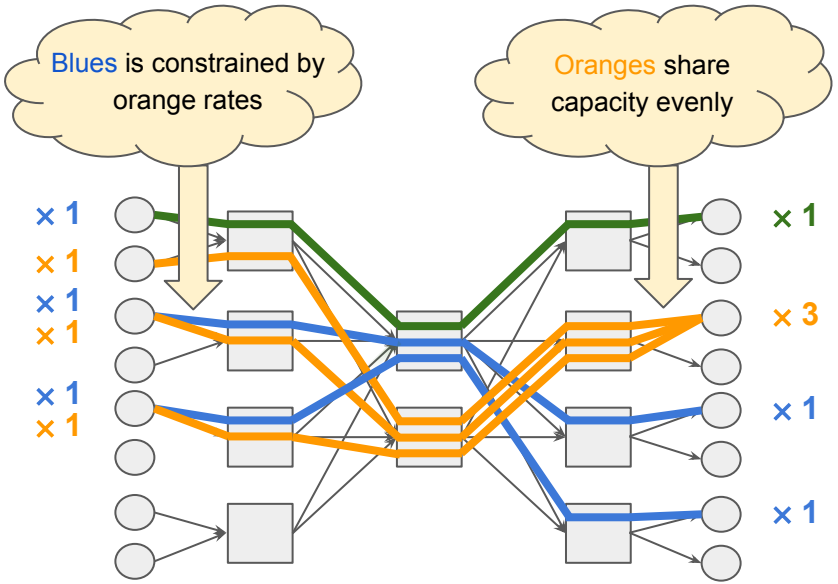


Flows	Rate
O:	1/3

Links have capacity 1

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

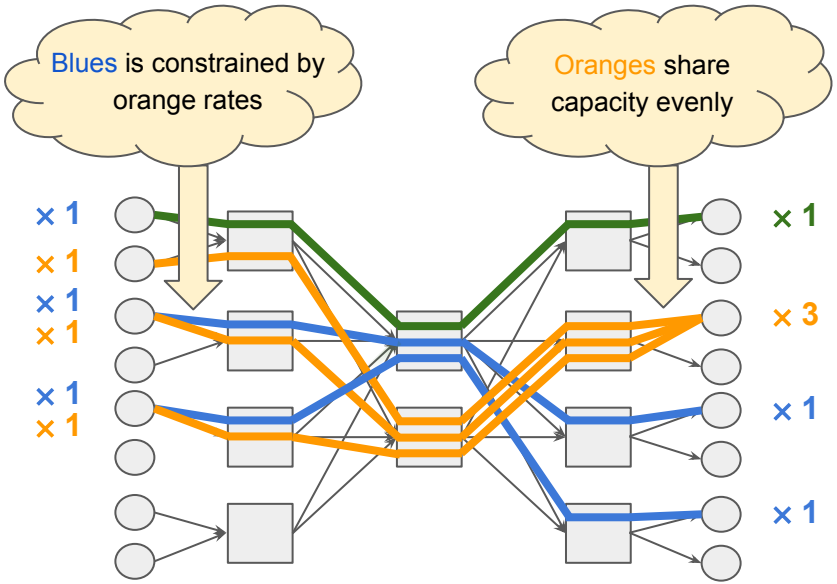


Flows	Rate
O:	1/3
B:	2/3

Links have capacity 1

Max-min fair allocations for *fixed* routing

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]



Flows	Rate
O:	1/3
B:	2/3
G:	1

Links have capacity 1

Bottleneck property of max-min fairness

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

Bottleneck property of max-min fairness

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

- **Definition**: A link is a *bottleneck* for a flow if:
  - (1) Link is *saturated* (total rate over all flows traversing the link is 1);
  - (2) Flow rate is at least flow rate of every flow traversing the link

Bottleneck property of max-min fairness

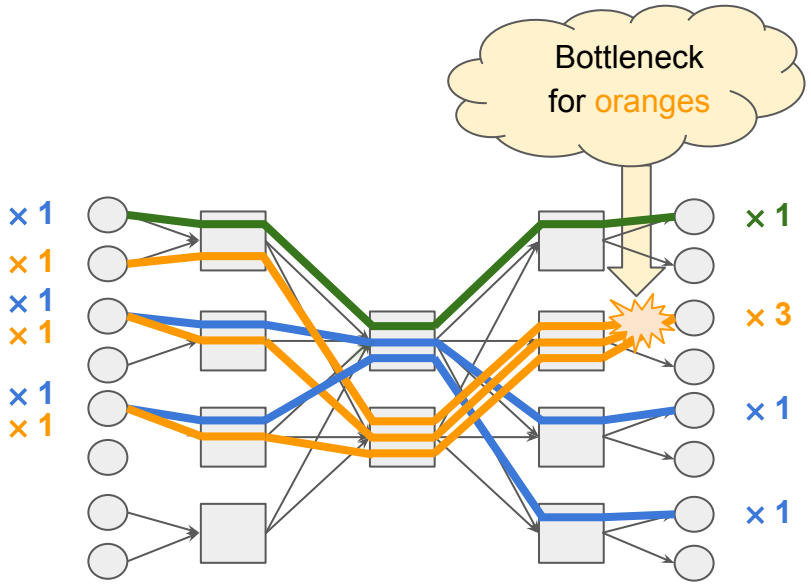
[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]

- **Definition**: A link is a *bottleneck* for a flow if:
  - (1) Link is *saturated* (total rate over all flows traversing the link is 1);
  - (2) Flow rate is at least flow rate of every flow traversing the link

- **Lemma**: An allocation is MmF if and only if all flows have a bottleneck link

Bottleneck property of max-min fairness

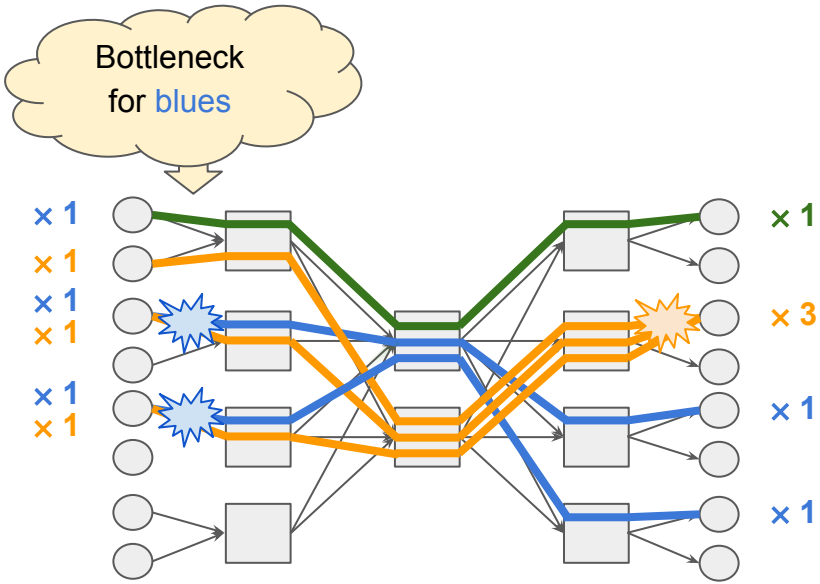
[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]



Flows	Rate
O:	1/3
B:	2/3
G:	1

Bottleneck property of max-min fairness

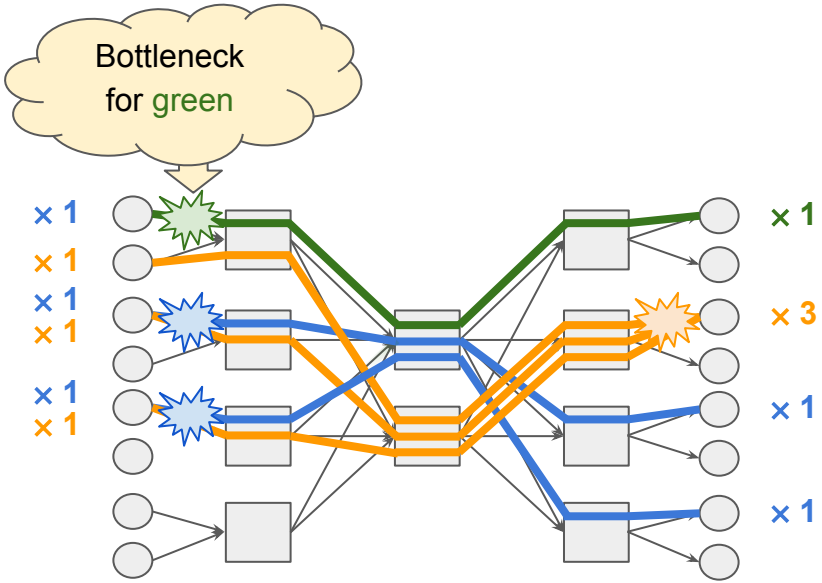
[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]



Flows	Rate
O:	1/3
B:	2/3
G:	1

Bottleneck property of max-min fairness

[Bertsekas/Gallagher 92, Radunovic/Le Boudec 07]



Flows	Rate
O:	1/3
B:	2/3
G:	1

Imposing max-min fairness halves throughput

Imposing max-min fairness halves throughput

- Expectation: Throughput loss is *small*

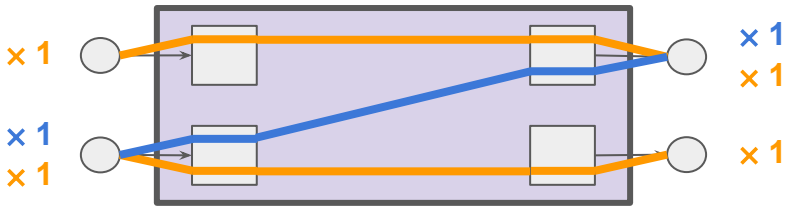
Imposing max-min fairness halves throughput

- Expectation: Throughput loss is *small*

- **Theorem #1:** For every macro-switch, throughput of MmF allocation is **≥ 1/2 x** throughput of maximum throughput allocation. This bound is tight.

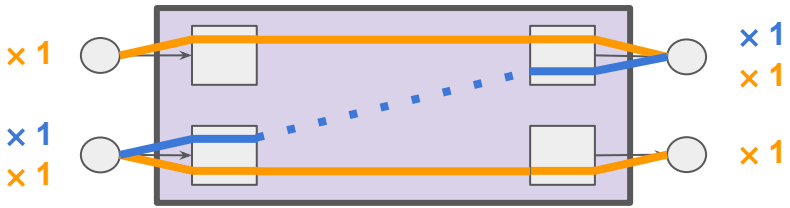
Imposing max-min fairness halves throughput

- **Proof of tightness:**



Imposing max-min fairness halves throughput

- **Proof of tightness:**



Maximum throughput: Flows Rate

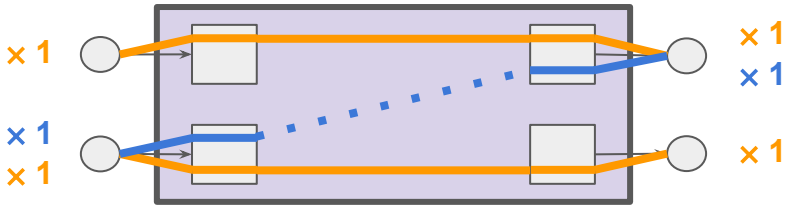
R: 1 (x 2)

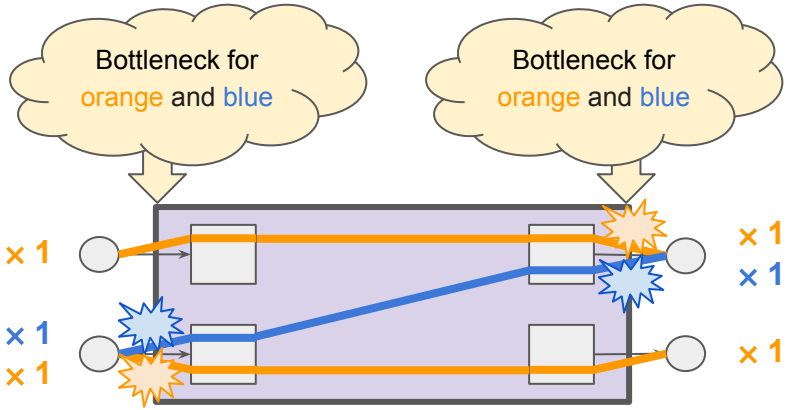
B: 0 (x 1)

Throughput	2

Imposing max-min fairness halves throughput

- **Proof of tightness:**



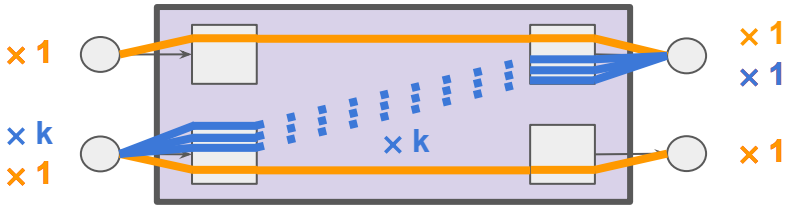


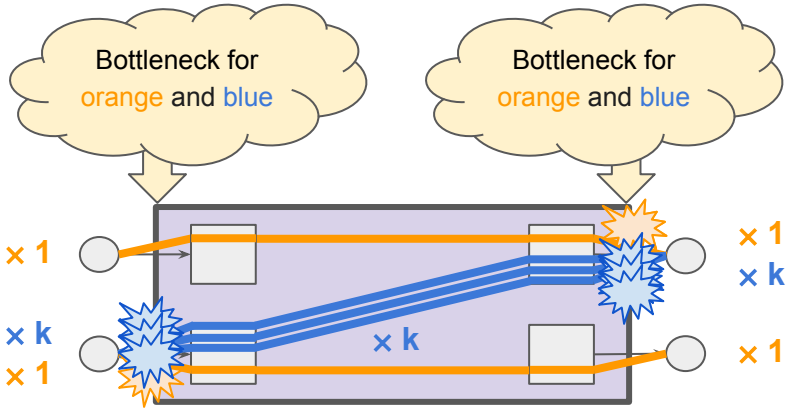
Maximum throughput:	Flows	Rate
	R:	1 (x 2)
	B:	0 (x 1)
Throughput		2

Max-min fair:	Flows	Rate
	R:	1/2 (x 2)
	B:	1/2 (x 1)
Throughput		1.5

Imposing max-min fairness halves throughput

- **Proof of tightness:** $k \to \infty$





Maximum throughput:

	Flows	Rate
R:	1 (x 2)	
B:	0 (x k)	
Throughput	2	

Max-min fair:

	Flows	Rate
R:	1/k (x 2)	
B:	1/k (x k)	
Throughput	1+1/k $\to$ 1	

Takeaways from Theorem #1

- For every network that implements macro-switch abstraction (not necessarily Clos), there is substantial throughput price to transmitting flows at MmF rates

Takeaways from Theorem #1

- For all interconnection networks (not necessarily Clos), there is substantial throughput price to transmitting flows at MmF rates
- Simulation-based evaluation (extended version) shows that this price is still substantial in stochastic setting

Extending max-min fairness to variable routing leads to starvation

Extending max-min fairness to variable routing leads to starvation

- **Definition** [Kleinberg/Rabani/Tardor 99]: The lex-max-min fair allocation (L-MmF, fairest network rates in max-min fair sense) maximizes in lex. order vectors corresponding to max-min fair allocations at each routing

Extending max-min fairness to variable routing leads to starvation

- **Definition** [Kleinberg/Rabani/Tardor 99]: The lex-max-min fair allocation (L-MmF, fairest network rates in max-min fair sense) maximizes in lex. order vectors corresponding to max-min fair allocations at each routing

- Expectation: L-MmF rates in Clos network and MmF rates in macro-switch are *close*

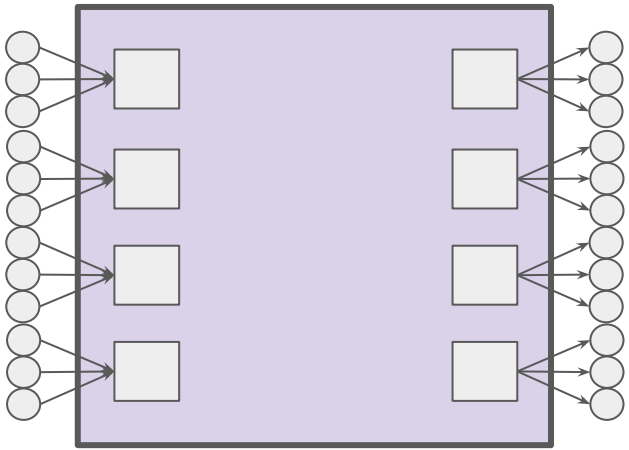
Extending max-min fairness to variable routing leads to starvation

- **Definition** [Kleinberg/Rabani/Tardor 99]: The lex-max-min fair allocation (L-MmF, fairest network rates in max-min fair sense) maximizes in lex. order vectors corresponding to max-min fair allocations at each routing

- Expectation: L-MmF rates in Clos network and MmF rates in macro-switch are *close*

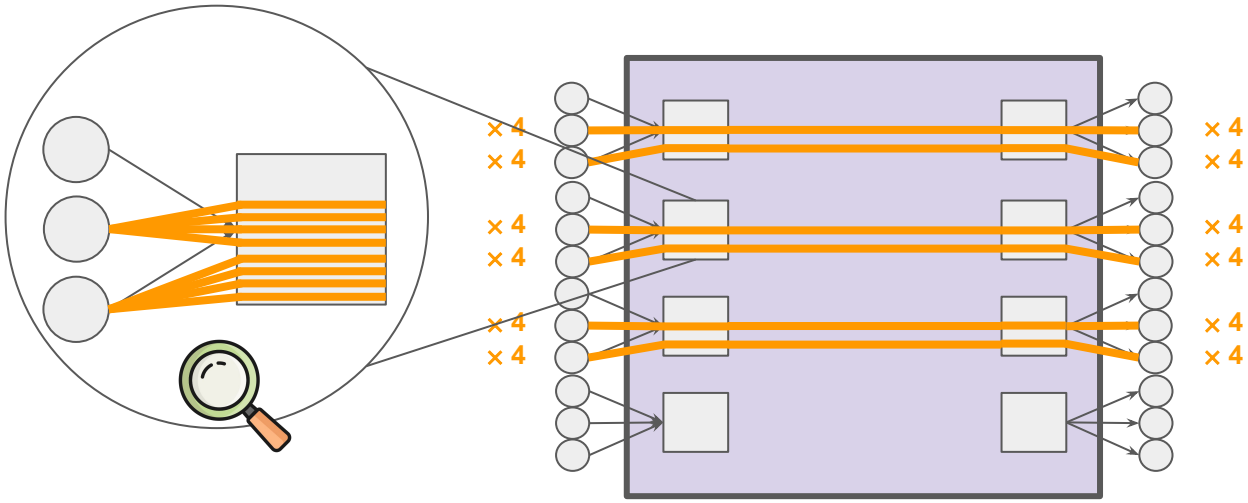
- **Theorem #2:** For every Clos network, there are input flows for which L-MmF rates of some flows in Clos network are smaller by **≥ 1/n x** their MmF rates in macro-switch, where n is number of middle switches

Extending max-min fairness to variable routing leads to starvation



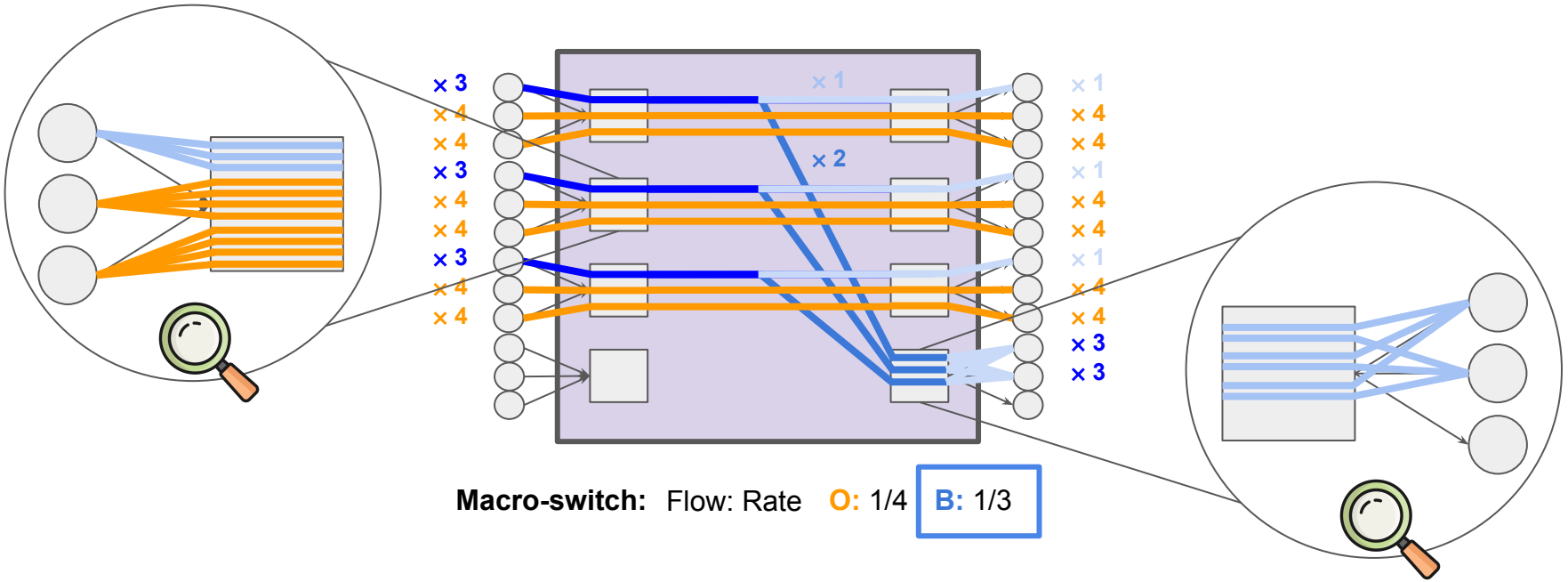
Macro-switch: Flow: Rate

Extending max-min fairness to variable routing leads to starvation



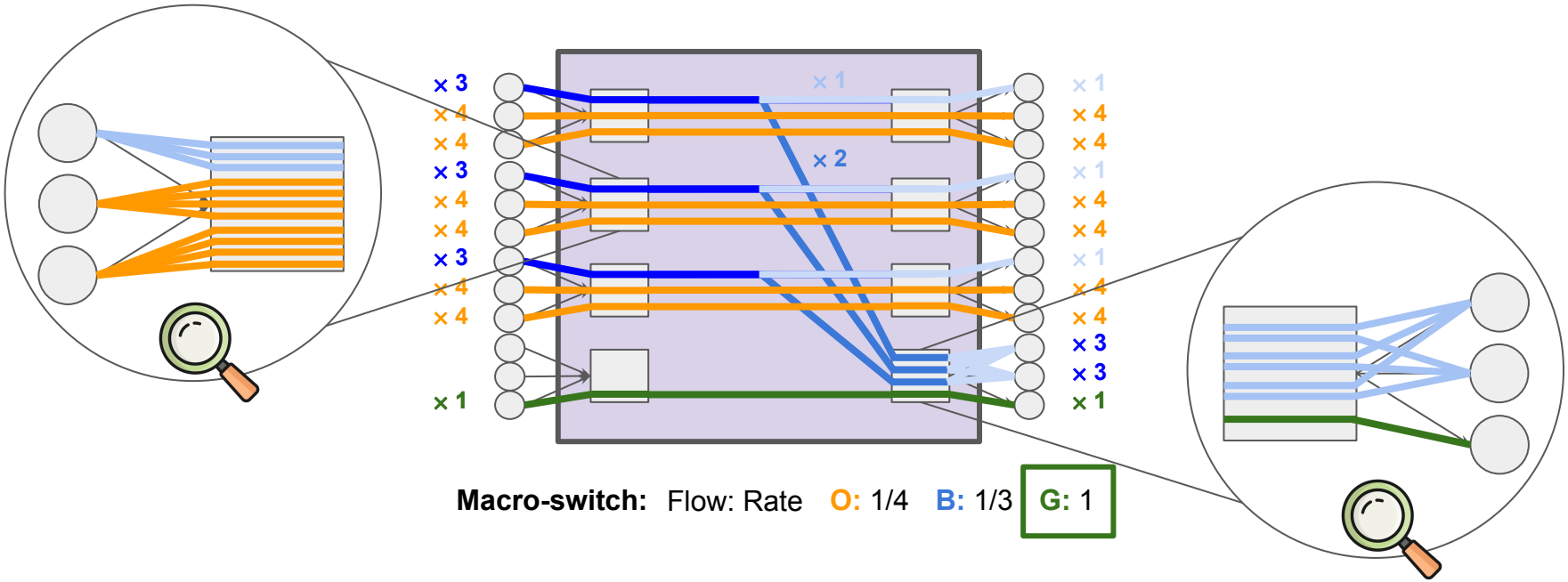
Macro-switch: Flow: Rate **O:** 1/4

Extending max-min fairness to variable routing leads to starvation



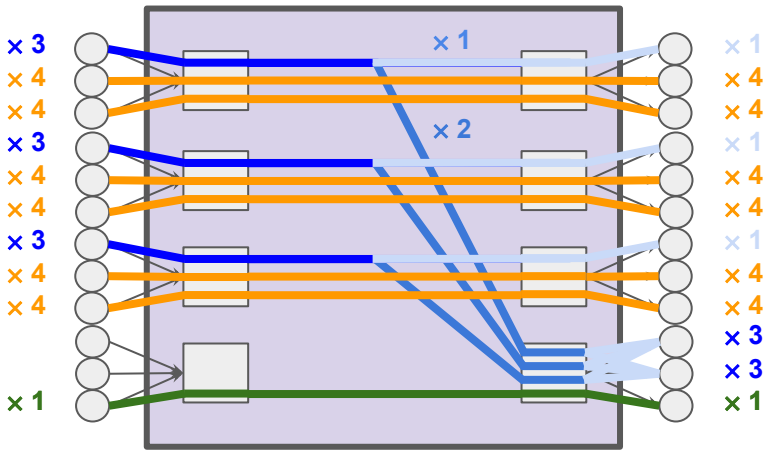
Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3

Extending max-min fairness to variable routing leads to starvation



Extending max-min fairness to variable routing leads to starvation

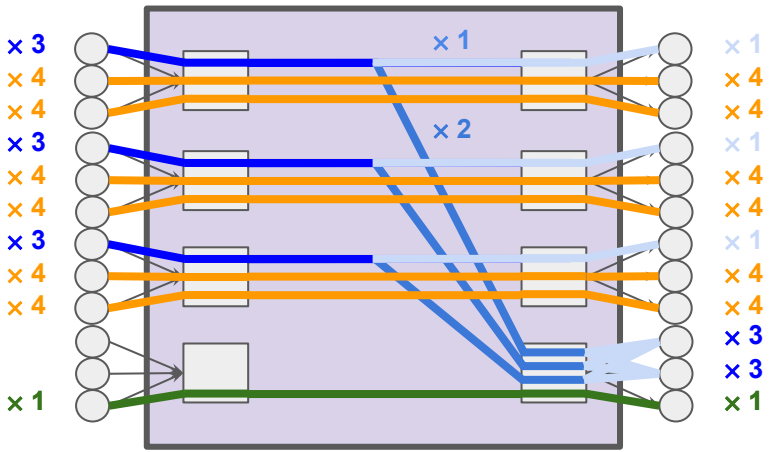
- Goal: In L-MmF allocation in Clos network, orange and blues uphold their macro-switch rates, but green decreases its rate from 1 to ⅓ (1 to 1/n)



Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1

Extending max-min fairness to variable routing leads to starvation

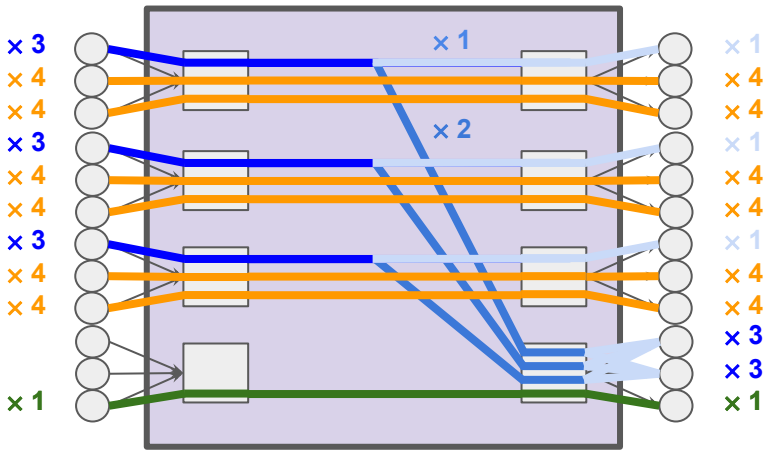
- Key idea #1: Some flows cannot uphold their macro-switch rates



Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1

Extending max-min fairness to variable routing leads to starvation

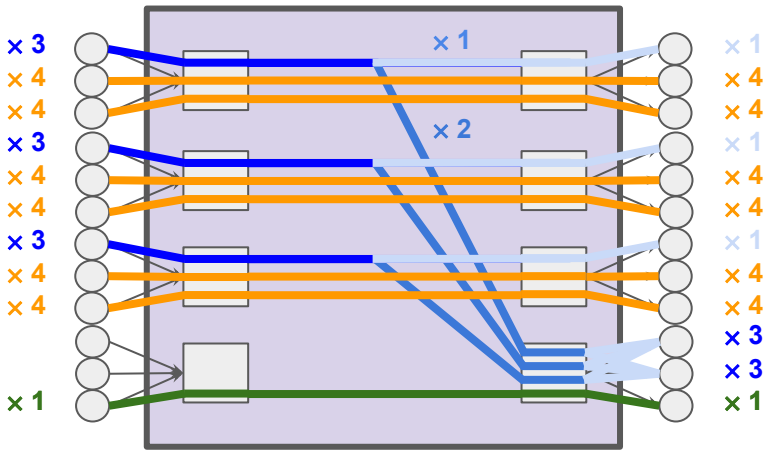
- Key idea #2: L-MmF upholds lower macro-switch rates of oranges and blues by decreasing higher macro-switch rate of green



Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1

Extending max-min fairness to variable routing leads to starvation

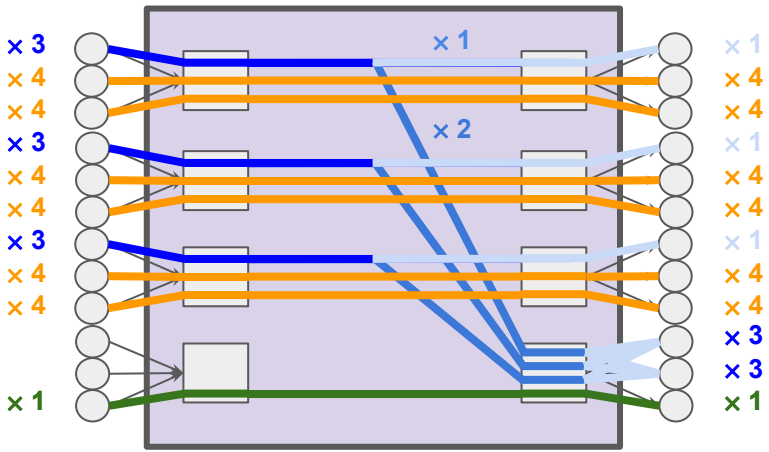
- Key idea #3: All routings upholding macro-switch rates of oranges and blues decrease green from 1 to 1/n



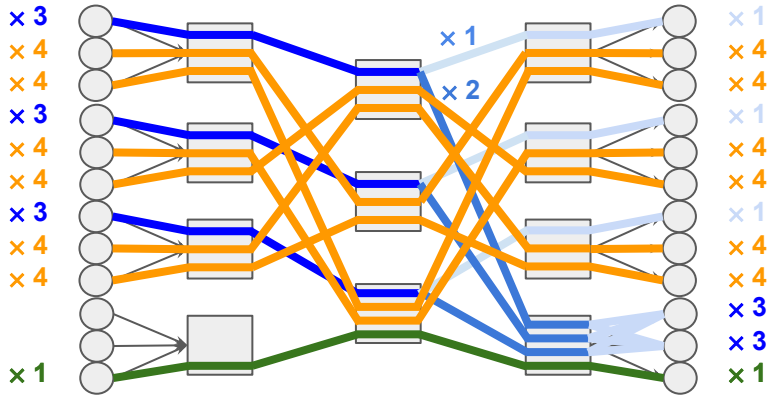
Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1

Extending max-min fairness to variable routing leads to starvation

- Key idea #3: All routings upholding macro-switch rates of oranges and blues decrease green from 1 to 1/n



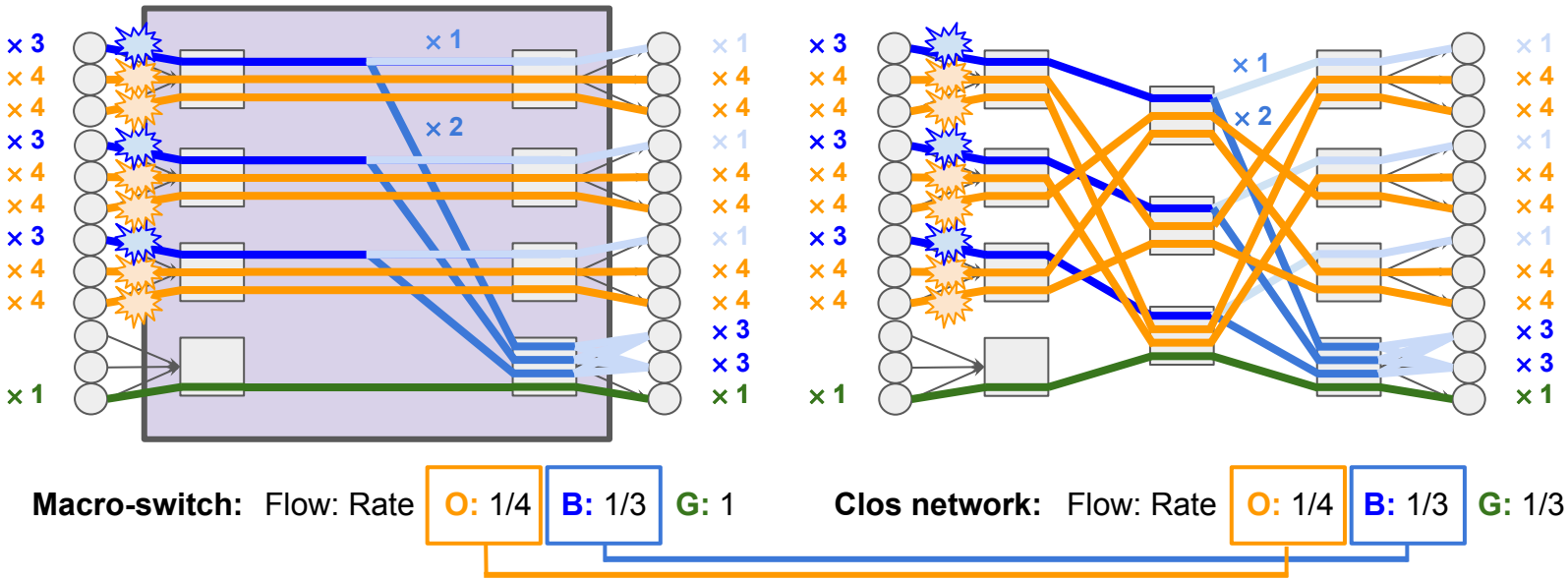
Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1



Clos network: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1/3

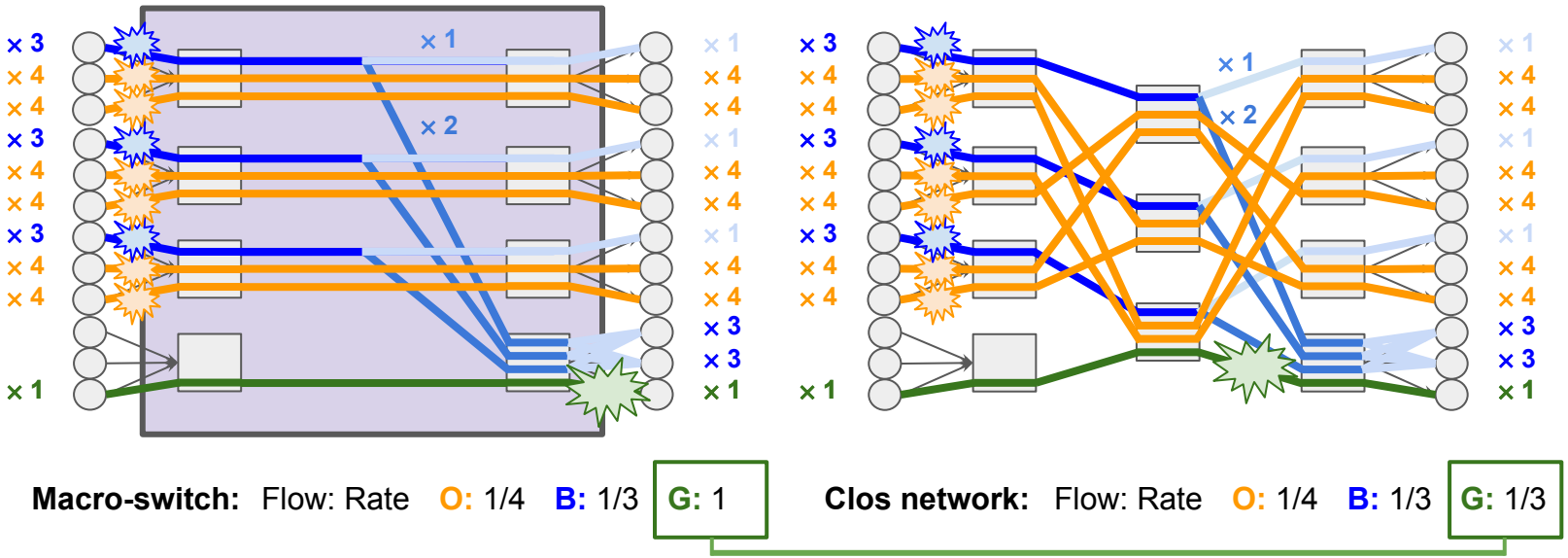
Extending max-min fairness to variable routing leads to starvation

- Key idea #3: All routings upholding macro-switch rates of oranges and blues decrease green from 1 to 1/n



Extending max-min fairness to variable routing leads to starvation

- Key idea #3: All routings upholding macro-switch rates of oranges and blues decrease green from 1 to 1/n



Macro-switch: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1

Clos network: Flow: Rate **O:** 1/4 **B:** 1/3 **G:** 1/3

Takeaways from Theorem #2

- If data-centers wish to approximate macro-switch abstraction for fairness, then lex-max-min fairness (natural objective) may not be the *right* routing objective

Takeaways from Theorem #2

- If data-centers wish to approximate macro-switch abstraction for fairness, then lex-max-min fairness (natural objective) may not be the *right* routing objective

- Simulation-based evaluation (extended version) shows that, while some existing routing algorithms can approximate macro-switch abstraction for fairness in stochastic setting, they fail in worse-case setting

Sacrificing max-min fairness with variable routing doubles throughput

Sacrificing max-min fairness with variable routing doubles throughput

- Expectation: Maximum throughput (with MmF rates) in Clos network at most throughput of MmF allocation in macro-switch, with MmF rates in Clos network somewhat fair relative to macro-switch

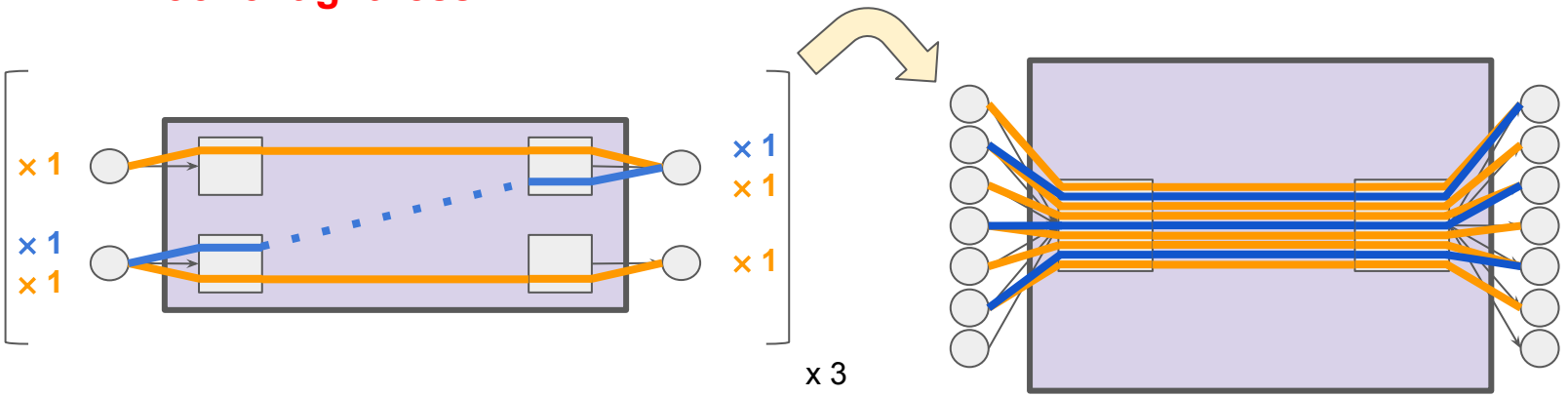
Sacrificing max-min fairness with variable routing doubles throughput

- Expectation: Maximum throughput (with MmF rates) in Clos network at most throughput of MmF allocation in macro-switch, with MmF rates in Clos network somewhat fair relative to macro-switch

- **Theorem #3:** For every Clos network, maximum throughput over all MmF allocations is **≤ *2 x*** throughput of MmF in macro-switch. This bound is tight (as network grows large), with MmF rates of most flows to **0**.

Sacrificing max-min fairness with variable routing doubles throughput

- **Proof of tightness:**

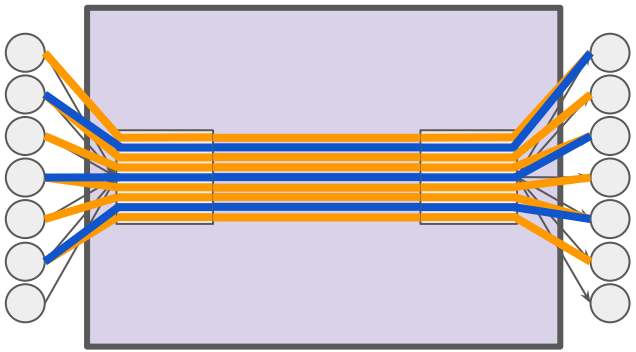


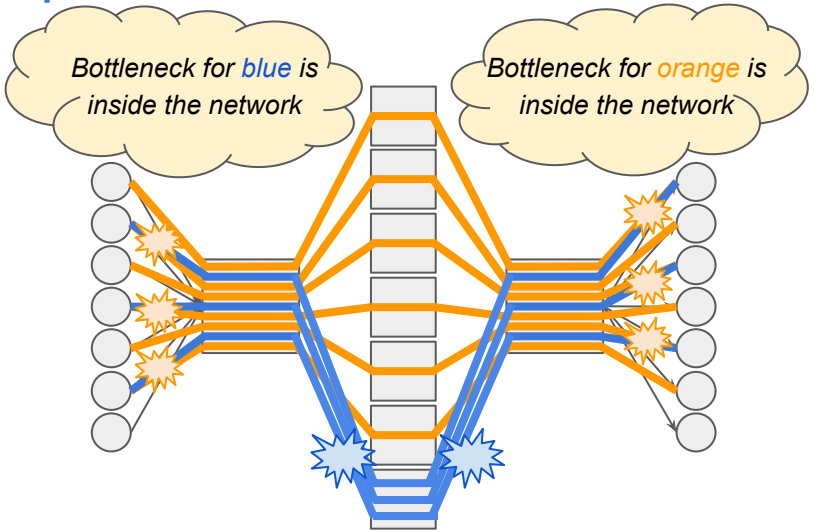
Macro-switch:	Flows	Rate
	R:	1/2 (x 2)
	B:	1/2 (x 1)
Throughput		1.5

Macro-switch:	Flows	Rate
	R:	1/2 (x 6)
	B:	1/2 (x 3)
Throughput		4.5

Sacrificing max-min fairness with variable routing doubles throughput

- **Proof of tightness:**



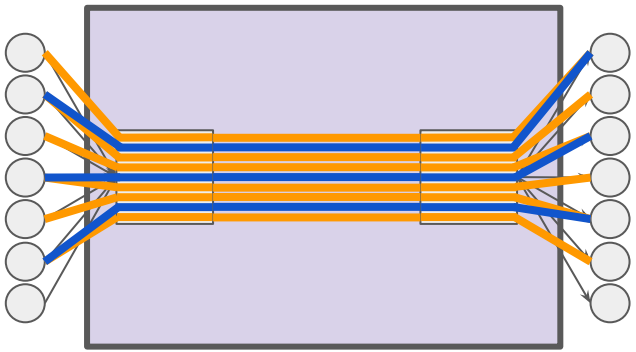


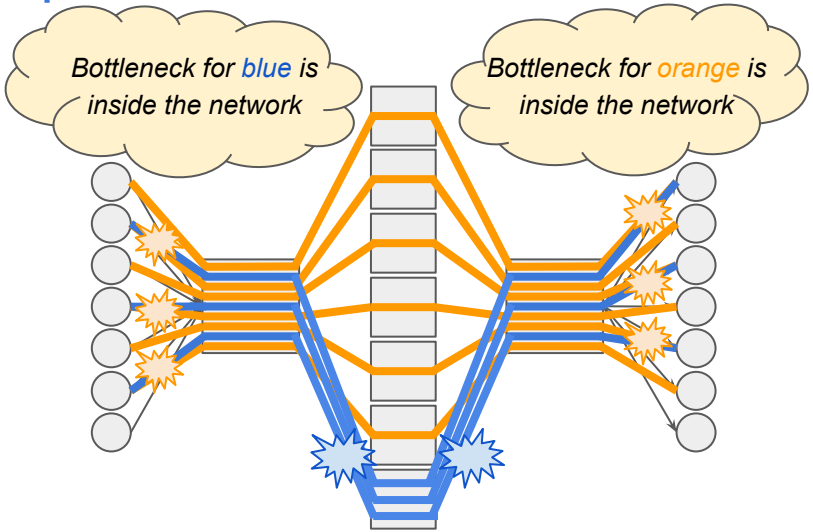
Macro-switch:	Flows	Rate
	R:	1/2 (x 6)
	B:	1/2 (x 3)
Throughput		4.5

Clos network:	Flows	Rate
	R:	2/3 (x 6)
	B:	1/3 (x 3)
Throughput		5

Sacrificing max-min fairness with variable routing doubles throughput

- **Proof of tightness:** n, k → ∞





Macro-switch:	Flows	Rate
	R:	1/k (x 2n)
	B:	1/k (x nk)
Throughput		n

Clos network:	Flows	Rate
	R:	1 (x 2n)
	B:	0 (x nk)
Throughput		2n

Takeaways to Theorem #3

- Variable routing can subvert max-min fair constraints to overall throughput gain

Takeaways to Theorem #3

- Variable routing can subvert max-min fair constraints to overall throughput gain
- Proof of upper bound introduces routing algorithm for approximating maximum throughput (with MmF rates) in Clos network

Conclusion

- We initiate rigorous study of the performance properties of data-centers under joint routing and congestion control

Conclusion

- We initiate rigorous study of the performance properties of data-centers under joint routing and congestion control
- We motivate the development of new theory and algorithms supporting the design of data-centers from foundational principles

Conclusion

- We initiate rigorous study of the performance properties of data-centers under joint routing and congestion control
- We motivate the development of new theory and algorithms supporting the design of data-centers from foundational principles

Miguel Alves Ferreira, maferrei@andrew.cmu.edu

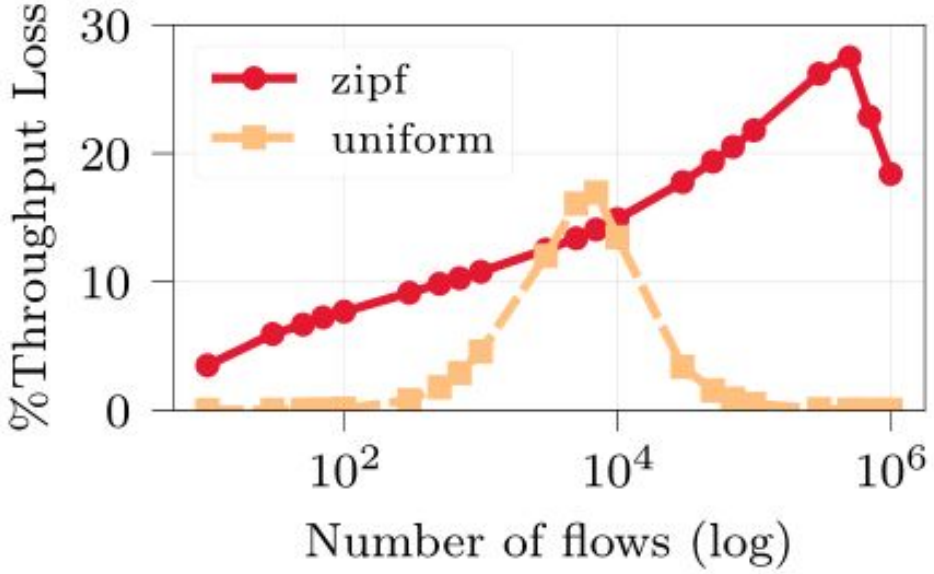
Carnegie Mellon, Instituto de Telecomunicações, and Instituto Superior Técnico

Follow-ups to Theorem #1 and #2

- If data-centers wish to minimize flow completion times, then they may benefit from avoiding fairness constraints, for instance via *scheduling* (selectively delaying some flows). Can scheduling improve on congestion control?

- If data-centers wish to approximate macro-switch abstraction for fairness, then an alternative objective is *proportional max-min fairness* (network rate of each flow proportional to macro-switch rate) [Jyothi et al. 14, Namay et al. 21]. Can proportional max-min fairness approximate fairness abstraction?

In a stochastic setting, throughput loss across macro-switch due to congestion control is still significant



In a stochastic setting, incorporating macro-switch rates into routing decisions allows to closely replicate macro-switch rates