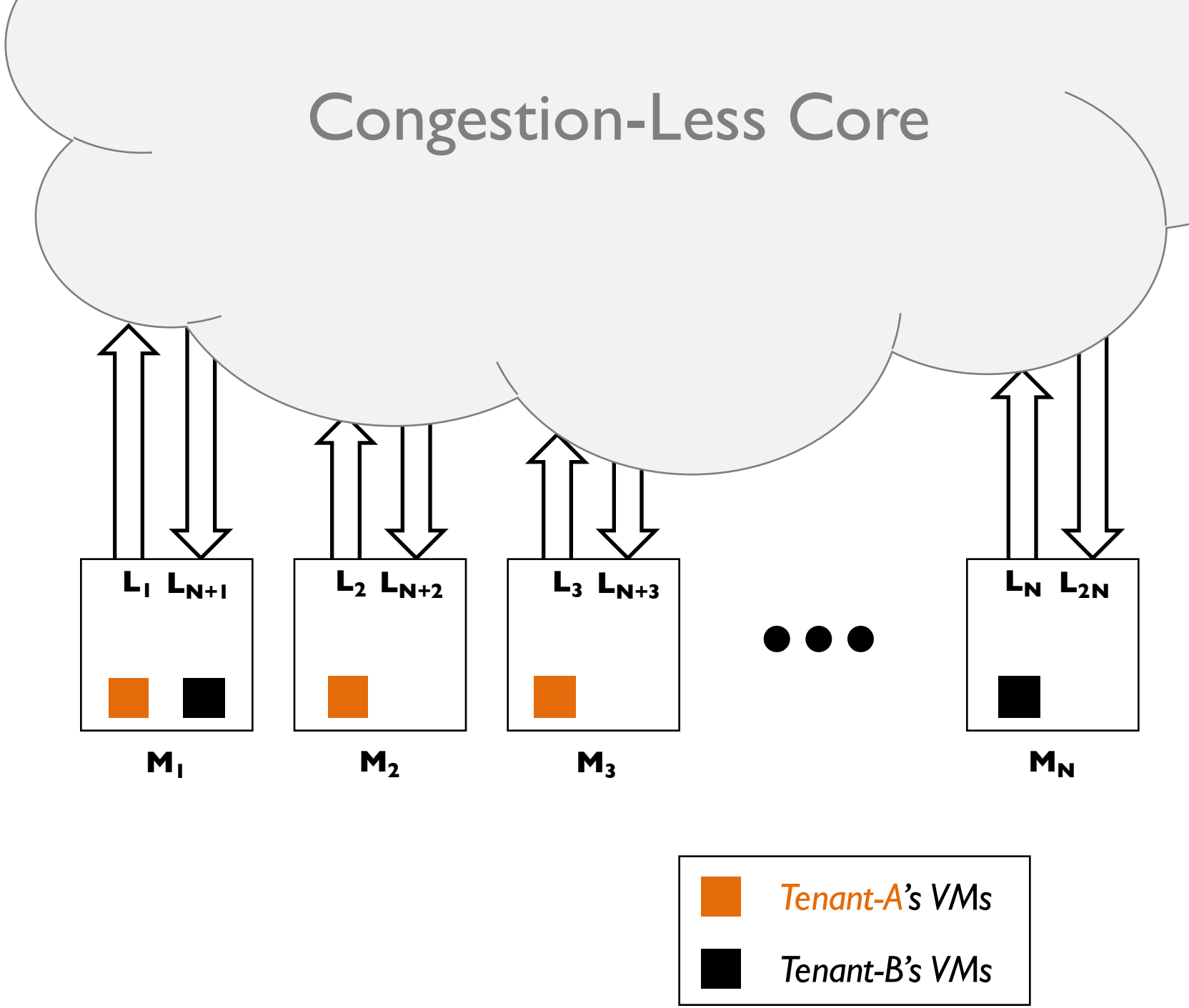
HUG

Multi-Resource Fairness for Correlated and Elastic Demands

Mosharaf Chowdhury, Zhenhua Liu
Ali Ghodsi, Ion Stoica

How to share the links between multiple tenants to provide

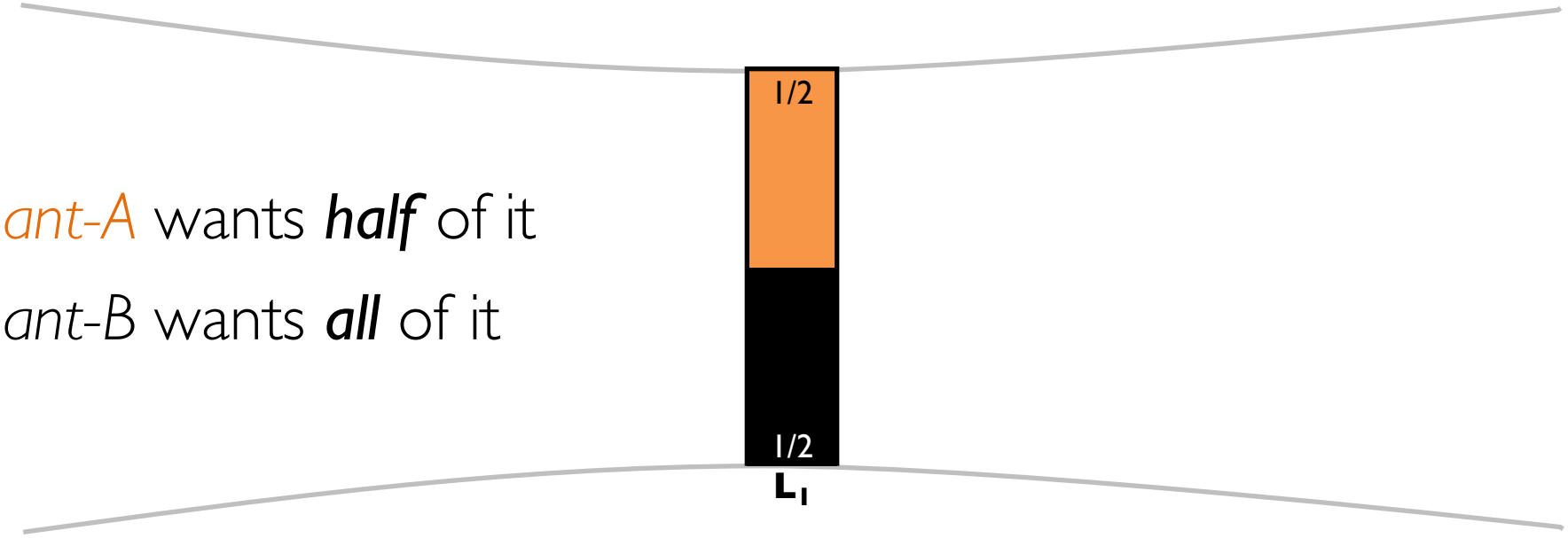
1. optimal performance guarantees and
2. maximize utilization?



Single-Resource Max-Min Fairness

Tenant-A wants ***half*** of it

Tenant-B wants ***all*** of it



1. Optimal Isolation Guarantee

1. Each tenant can create exactly one flow in each link

Single-Resource Max-Min Fairness

Tenant-A wants ***half*** of it

Tenant-B wants ***all*** of it

Progress (*M*) of a tenant is its demand-normalized allocation

Isolation Guarantee is the minimum progress across all

$d_A = 1/2 \quad a_A = 1/2 \quad M_A = \frac{a_A}{d_A} = 1$

$d_B = 1 \quad a_B = 1/2 \quad M_B = \frac{a_B}{d_B} = 1/2$

$\Big\}$ $\text{Min}(M_A, M_B) = 1/2$

Single-Resource Max-Min Fairness

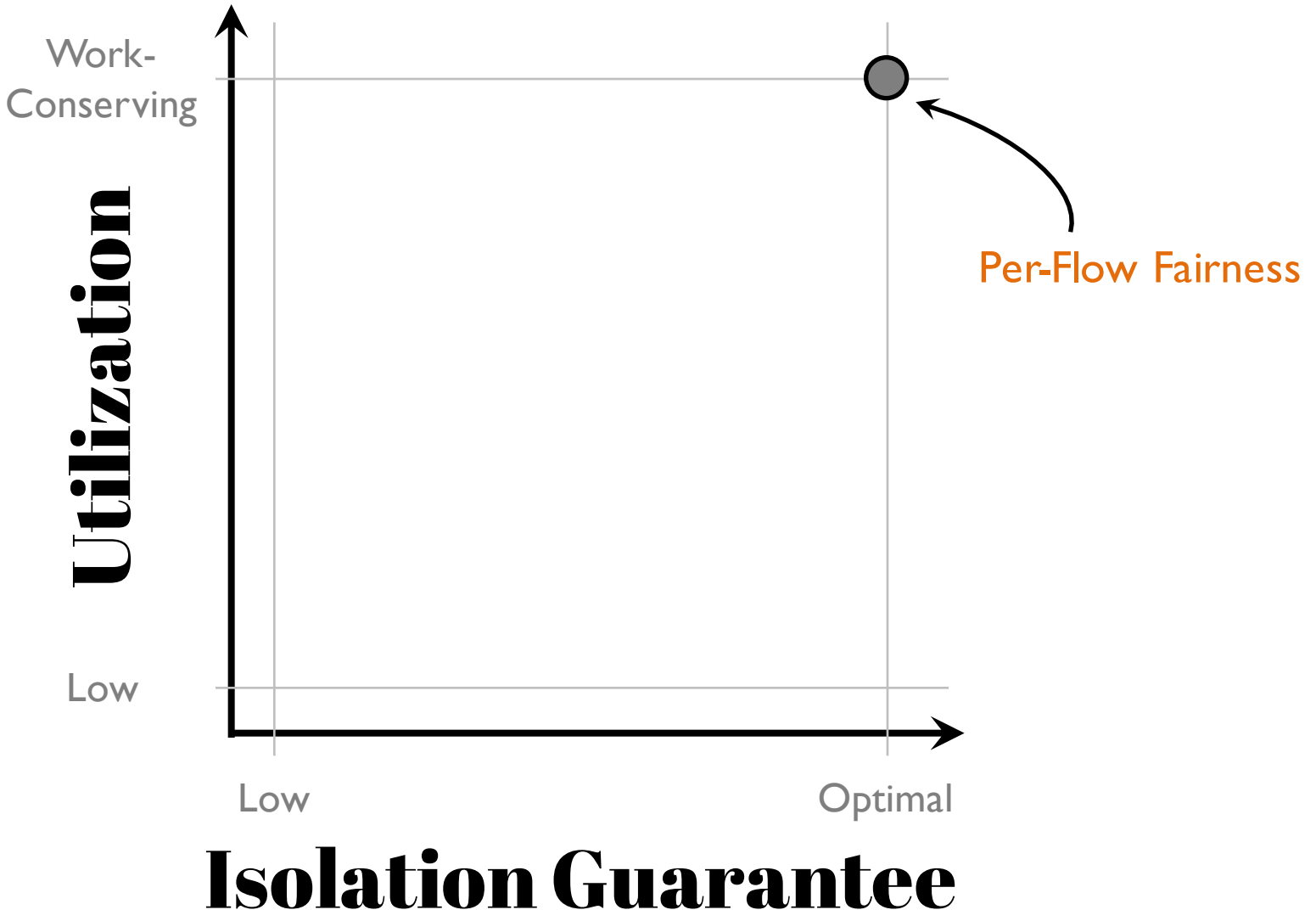
Tenant-A wants ***half*** of it

Tenant-B wants ***all*** of it

1. **Optimal Isolation Guarantee**
2. **Work Conservation**

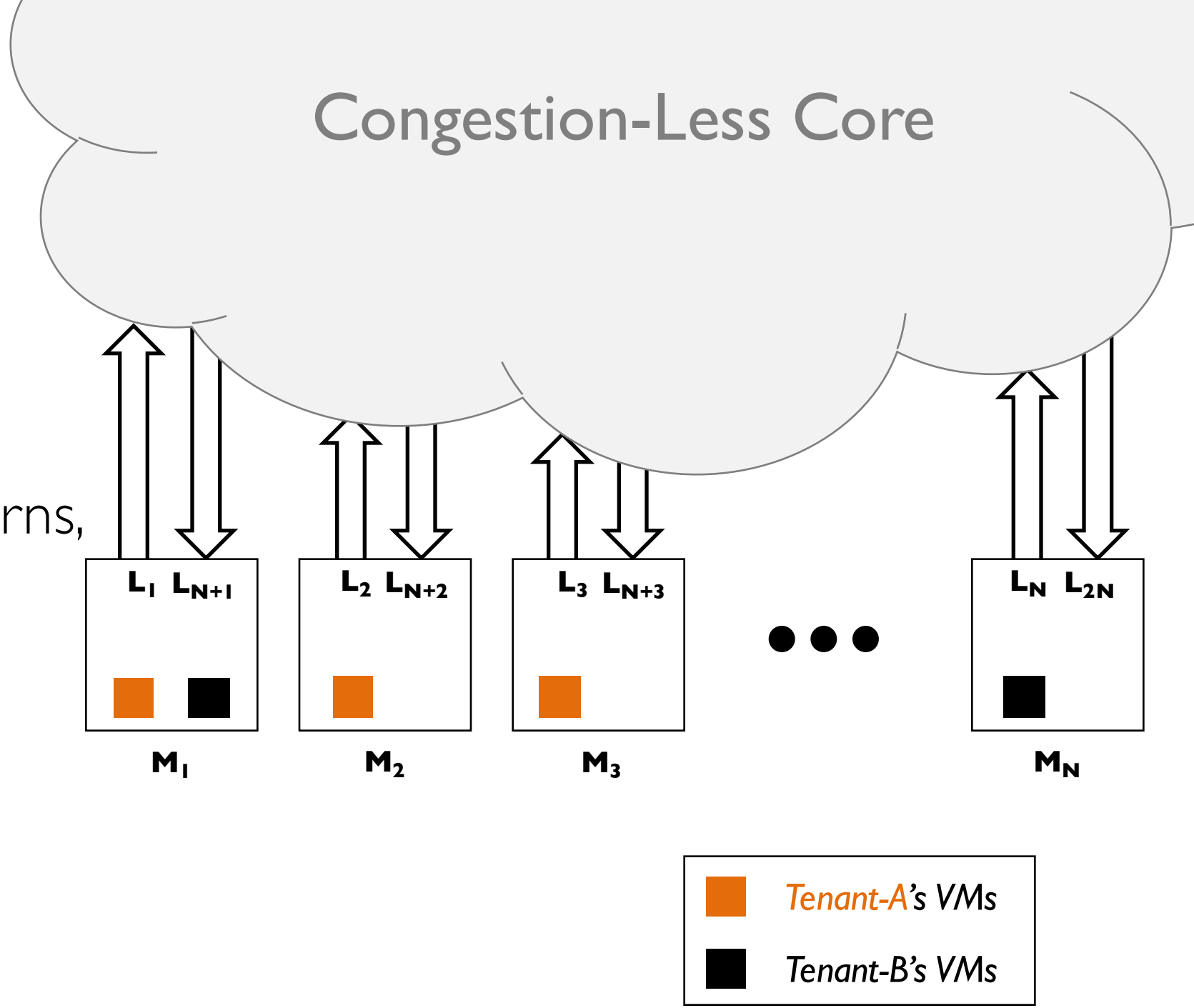
No Tradeoff for Single Resource

1. **Optimal Isolation Guarantee**
2. **Work Conservation**
3. **Strategyproof**



Tenants have different

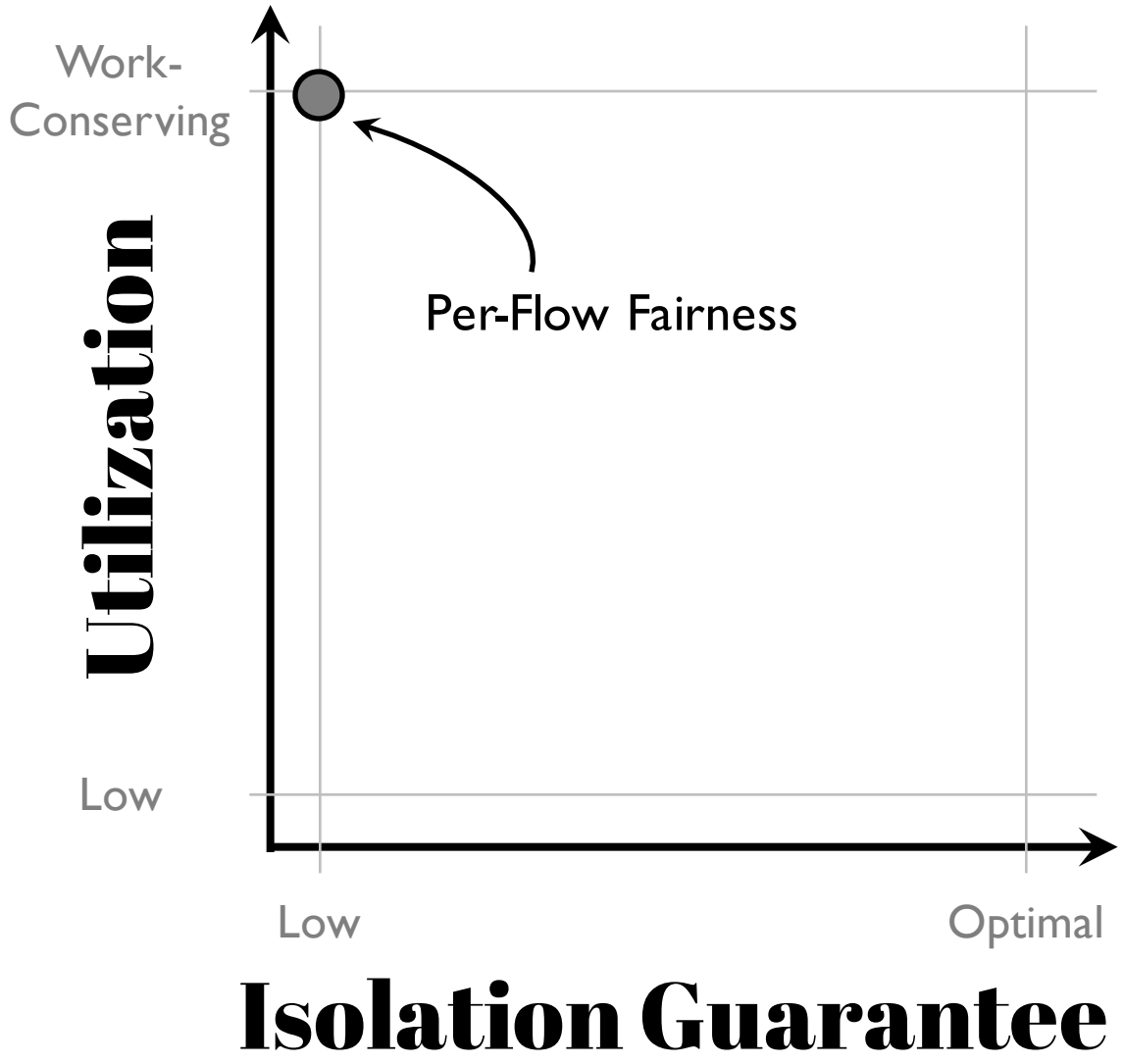
1. placements,
2. communication patterns,
3. demand correlations,
4. …



Per-Flow Fairness For Multiple Resources

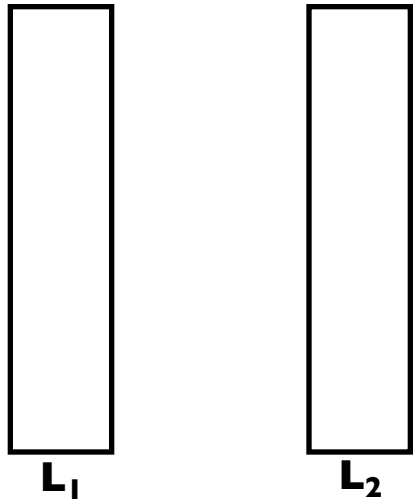
Low

1. ~~**Optimal** **Isolation Guarantee**~~
2. **Work Conservation**
3. ~~**Strategyproof**~~



Elastic Demands[1]

Tenant-A wants ***all*** of L_1 and ***all*** of L_2

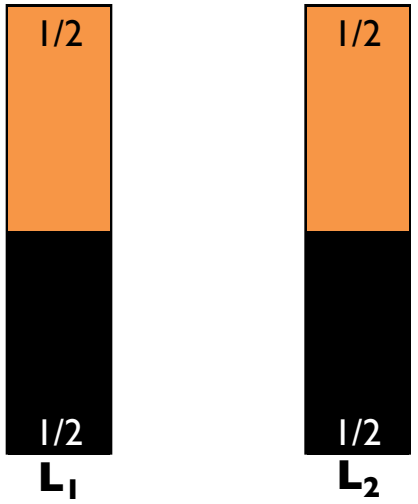


Tenant-B wants ***all*** of L_1 and ***all*** of L_2

1. *FairCloud: Sharing the Network in Cloud Computing, SIGCOMM'12*

Tenant-Level Max-Min Fairness (PS-P)

Tenant-A wants ***all*** of L_1 and ***all*** of L_2



Tenant-B wants ***all*** of L_1 and ***all*** of L_2

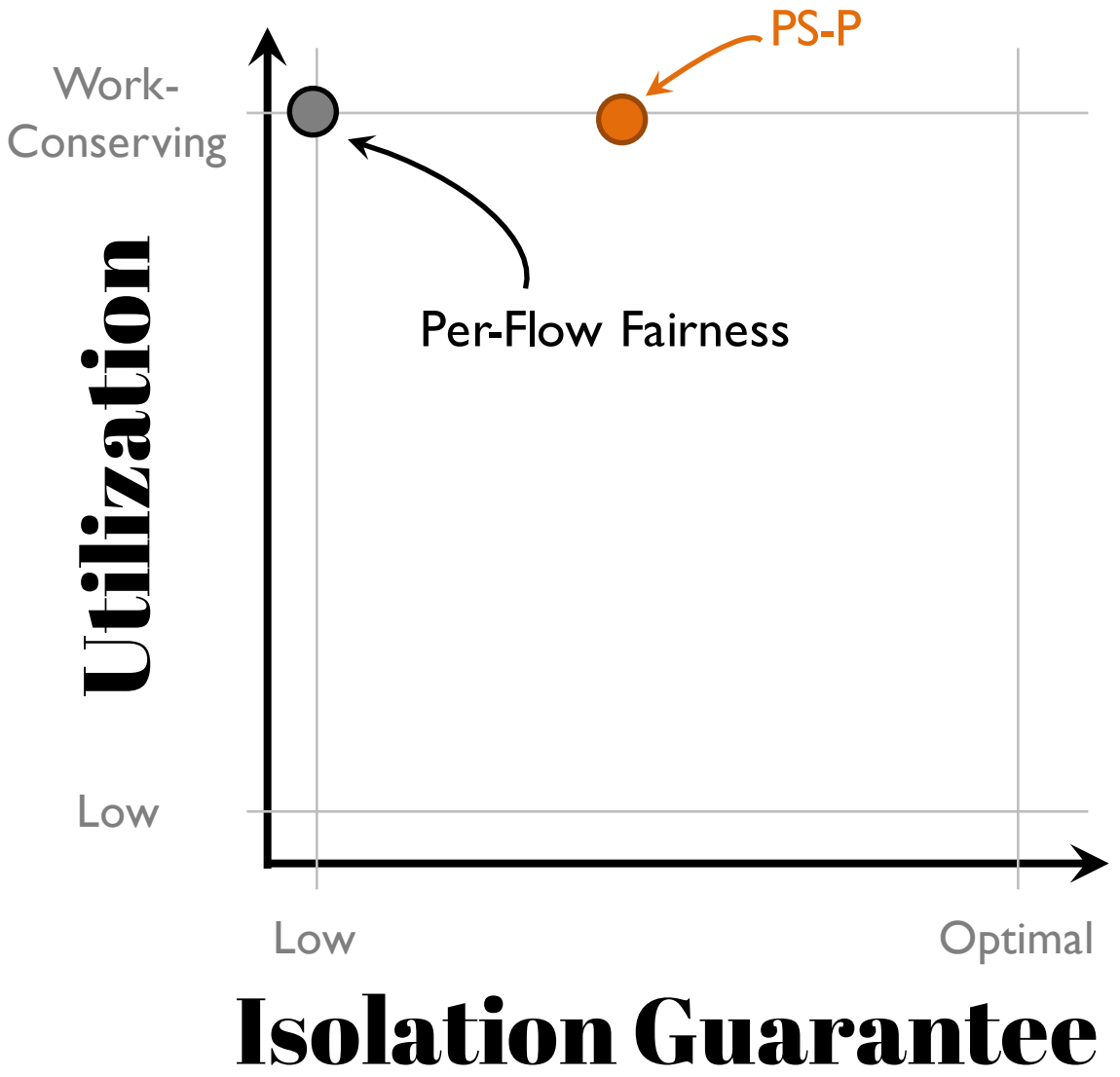
$\boldsymbol{d_A} = \langle 1, 1 \rangle \quad \boldsymbol{a_A} = \langle ½, ½ \rangle \quad \boldsymbol{M_A} = \min\left(\frac{\boldsymbol{a_A^i}}{\boldsymbol{d_A^i}}\right) = ½$

$\boldsymbol{d_B} = \langle 1, 1 \rangle \quad \boldsymbol{a_B} = \langle ½, ½ \rangle \quad \boldsymbol{M_B} = \min\left(\frac{\boldsymbol{a_B^i}}{\boldsymbol{d_B^i}}\right) = ½$

$\Big\}$ $\text{Min}(\boldsymbol{M_A}, \boldsymbol{M_B}) = ½$

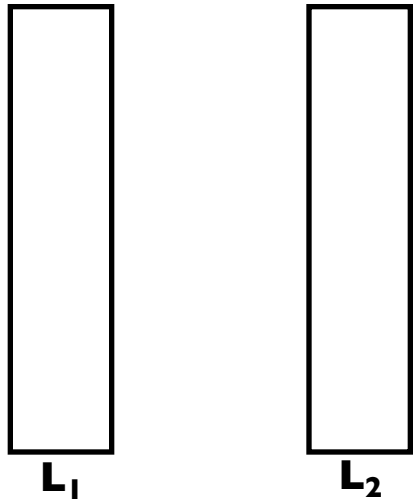
Tenant-Level Max-Min Fairness (PS-P)

1. **Suboptimal Isolation Guarantee**
2. **Work Conservation**



Correlated Demands[1]

Tenant-A wants ***some*** of L_1 and ***all*** of L_2

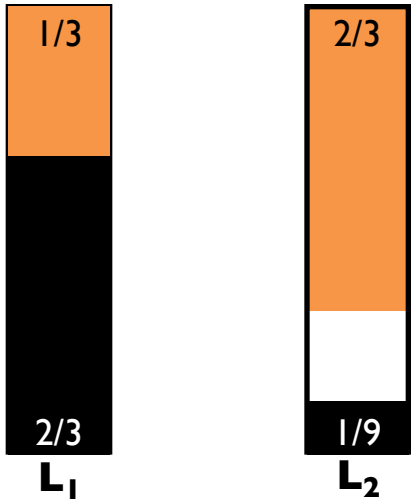


Tenant-B wants ***some*** of L_2 and ***all*** of L_1

1. Dominant Resource Fairness: Fair Allocation of Multiple Resource Types, NSDI'11

Dominant Resource Fairness (DRF)

Tenant-A wants ***exactly half*** unit of L_1 for ***each*** of L_2



Tenant-B wants ***exactly 1/6*** unit of L_2 for ***each*** of L_1

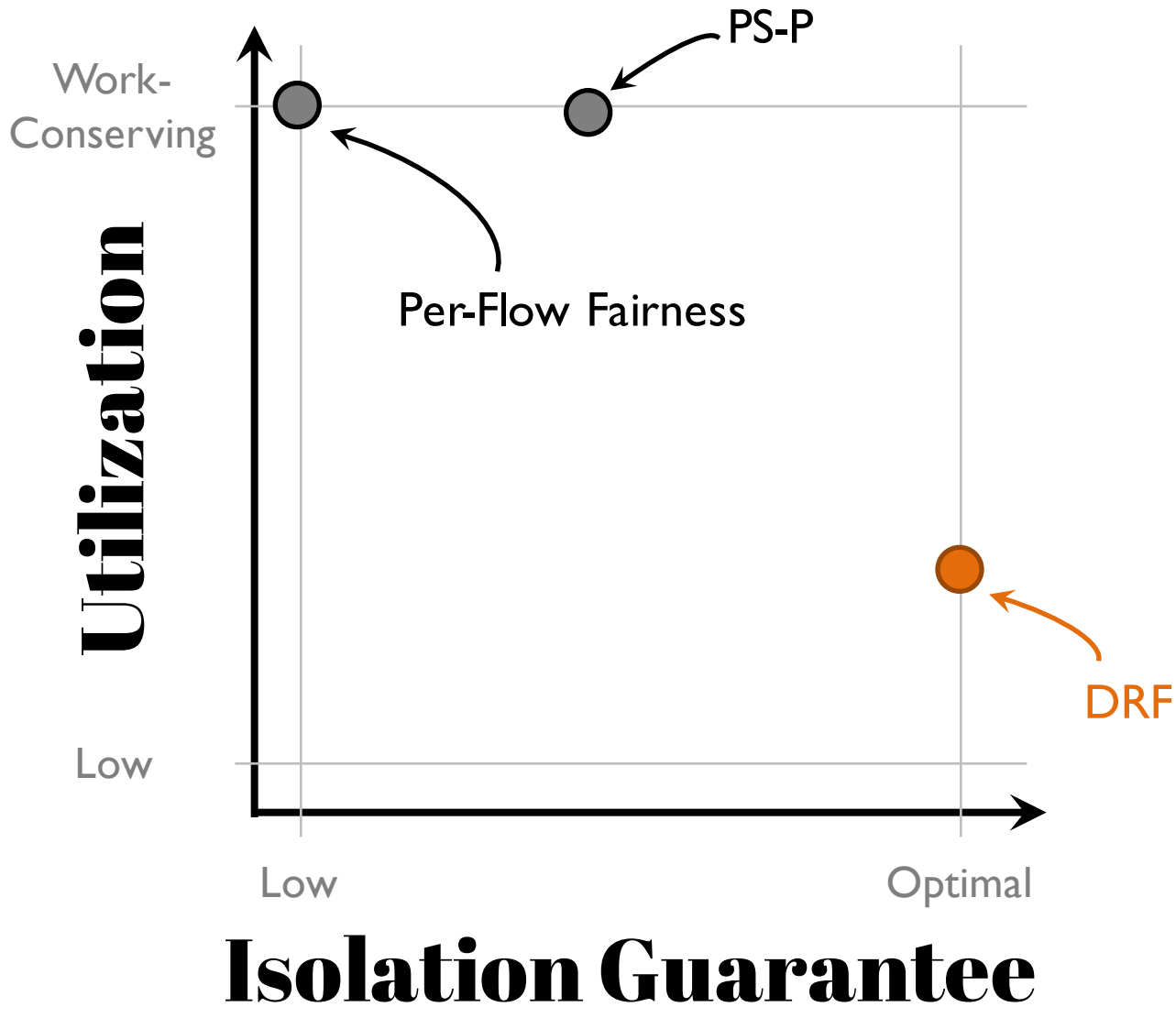
$\boldsymbol{d_A}$ = <1/2, 1> $\boldsymbol{a_A}$ = <1/3, 2/3> $\boldsymbol{M_A} = \min\left(\frac{\boldsymbol{a_A^i}}{\boldsymbol{d_A^i}}\right) = 2/3$

$\boldsymbol{d_B}$ = <1, 1/6> $\boldsymbol{a_B}$ = <2/3, 1/9> $\boldsymbol{M_B} = \min\left(\frac{\boldsymbol{a_B^i}}{\boldsymbol{d_B^i}}\right) = 2/3$

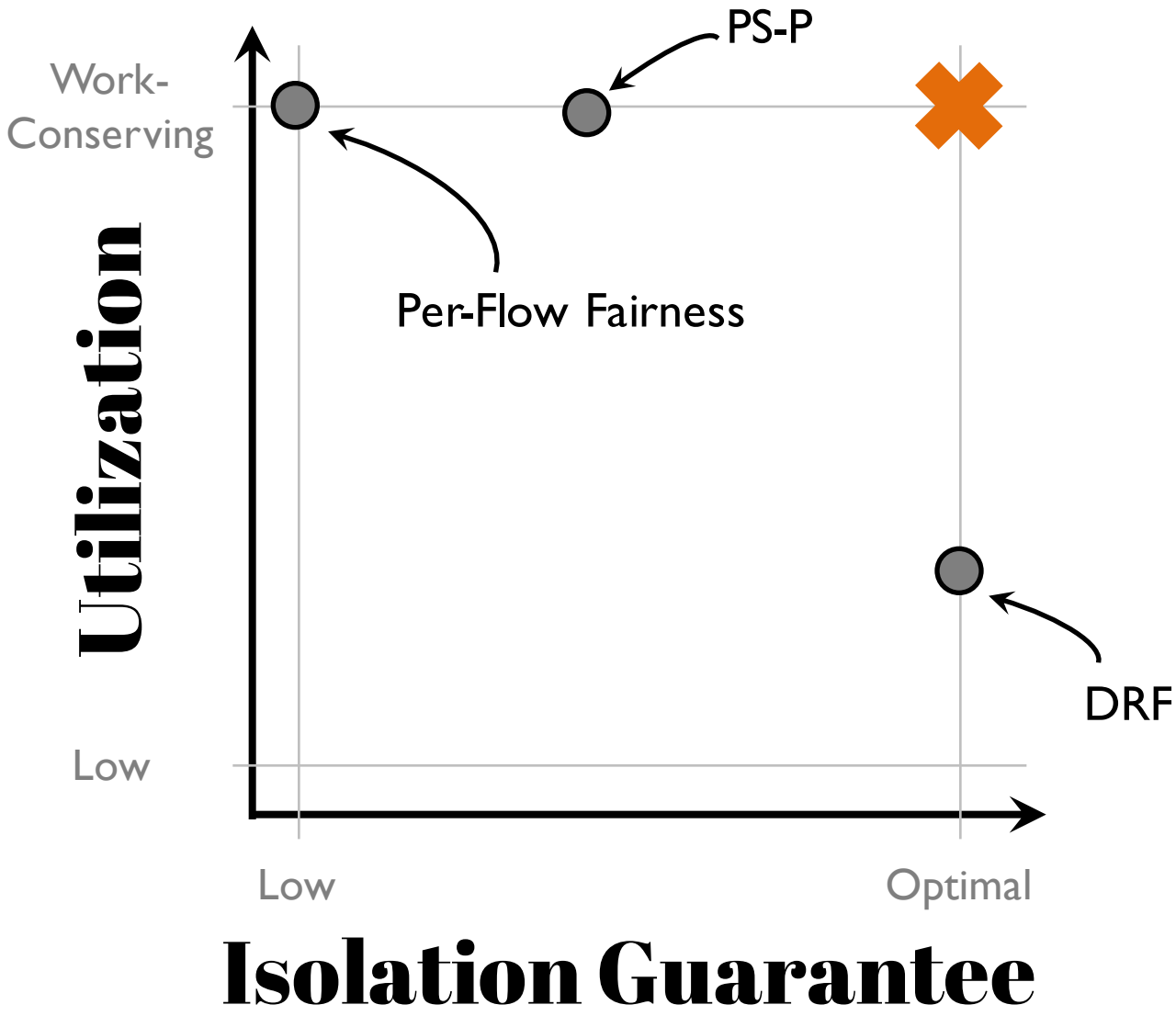
} **Min(M_A, M_B) = 2/3**

Dominant Resource Fairness (DRF)

1. **Optimal Isolation Guarantee**
2. **Arbitrarily Low Utilization**
3. **Strategyproof**



For elastic and correlated demands, can we ***simultaneously*** *achieve* **optimal isolation guarantee** *and* **maximum utilization**?



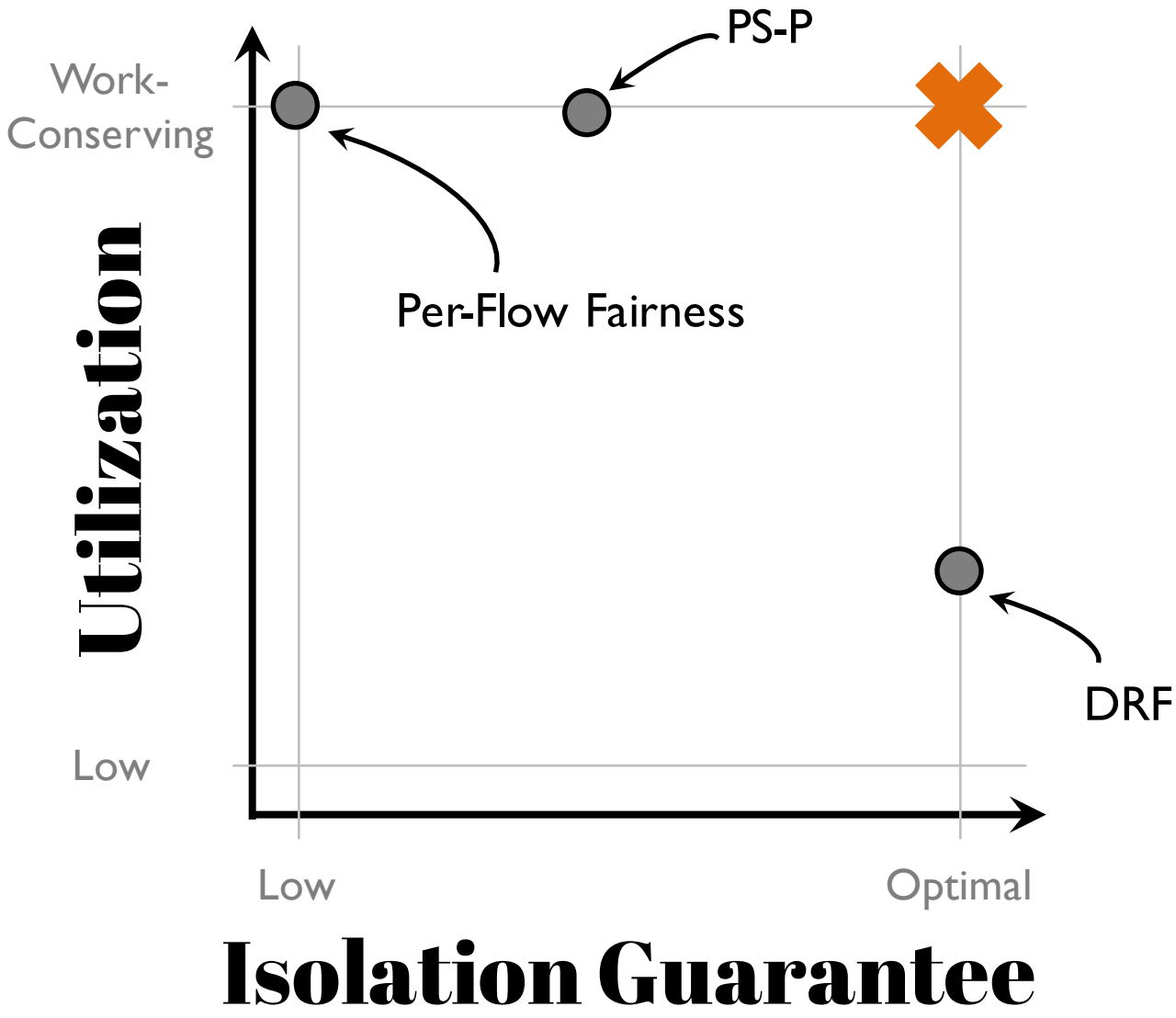
For elastic and correlated demands, can we ***simultaneously*** *achieve* **optimal isolation guarantee** *and* **maximum utilization**?

NO

1. Why not?
2. What's the best we can achieve?
3. How can we achieve that?
4. Does it matter?

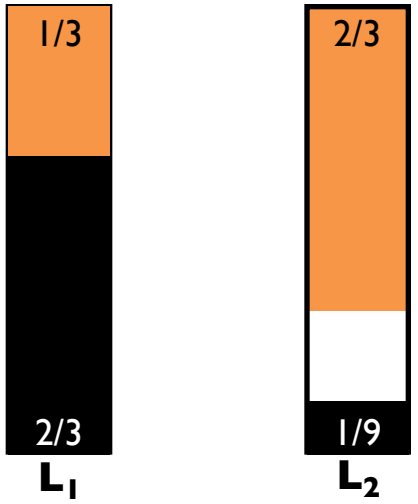
NO

1. **Why not?**
2. What's the best we can achieve?
3. How can we achieve that?
4. Does it matter?



Elastic and Correlated Demands

Tenant-A wants ***at least half*** unit of L_1 for ***each*** of L_2



Tenant-B wants ***at least 1/6*** unit of L_2 for ***each*** of L_1

$\boldsymbol{d_A}$ = <1/2, 1> $\boldsymbol{a_A}$ = <1/3, 2/3> $\boldsymbol{M_A}$ = 2/3

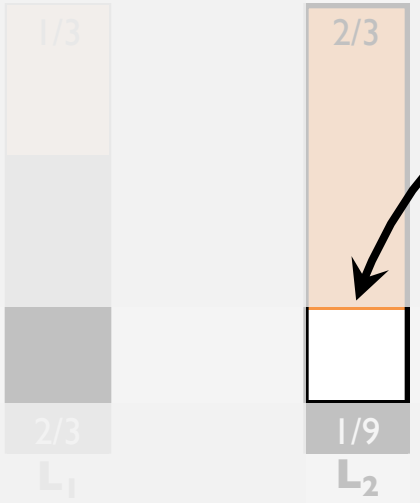
$\boldsymbol{d_B}$ = <1, 1/6> $\boldsymbol{a_B}$ = <2/3, 1/9> $\boldsymbol{M_B}$ = 2/3

} **Min($\boldsymbol{M_A}$, $\boldsymbol{M_B}$) = 2/3**

Elastic and Correlated Demands

Tenant-A wants ***at least half*** unit of L_1 for ***each*** of L_2

Tenant-B wants ***at least 1/6*** unit of L_2 for ***each*** of L_1



Who gets this?

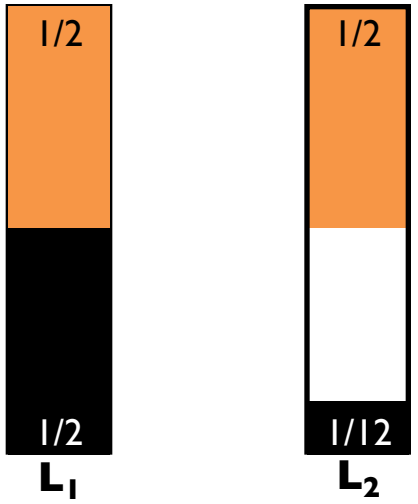
d_A = <1/2, 1> a_A = <1/3, 2/3> M_A = 2/3

d_B = <1, 1/6> a_B = <2/3, 1/9> M_B = 2/3

} Min(M_A, M_B) = 2/3

Work Conservation Doesn't Work!

Tenant-A ***lies*** and asks for ***one*** unit of L_1 for ***each*** of L_2



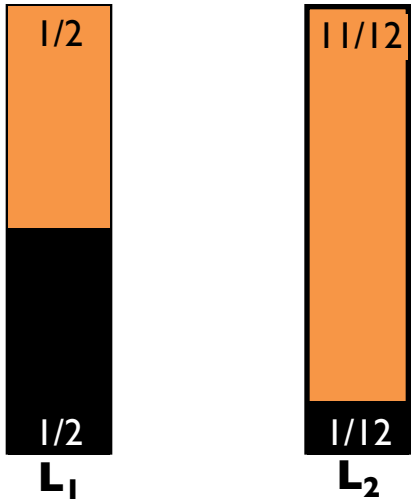
Tenant-B wants ***at least 1/6*** unit of L_2 for ***each*** of L_1

$\boldsymbol{d'_A} = \langle 1, 1 \rangle \quad \boldsymbol{a'_A} = \langle ½, ½ \rangle$

$\boldsymbol{d_B} = \langle 1, 1/6 \rangle \quad \boldsymbol{a_B} = \langle ½, 1/12 \rangle$

Work Conservation Doesn't Work!

Tenant-A ***lies*** and asks for ***one*** unit of L_1 for ***each*** of L_2

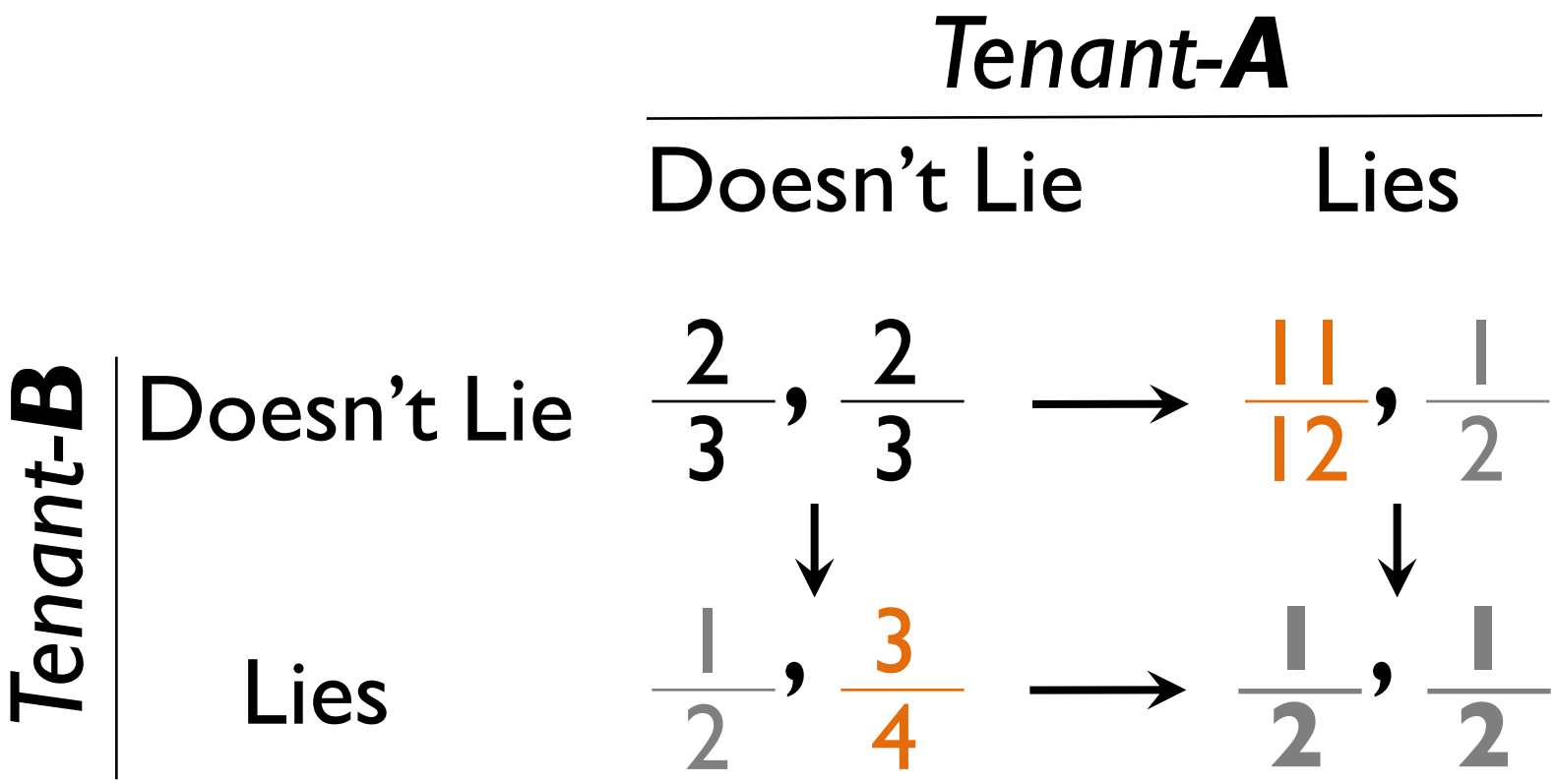


Tenant-B wants ***at least 1/6*** unit of L_2 for ***each*** of L_1

$\boldsymbol{d'_A} = \langle 1, 1 \rangle$ $\boldsymbol{a'_A} = \langle ½, ½ \rangle$ $\boldsymbol{a'_A} = \langle ½, 11/12 \rangle$ $\boldsymbol{M'_A} = 11/12$

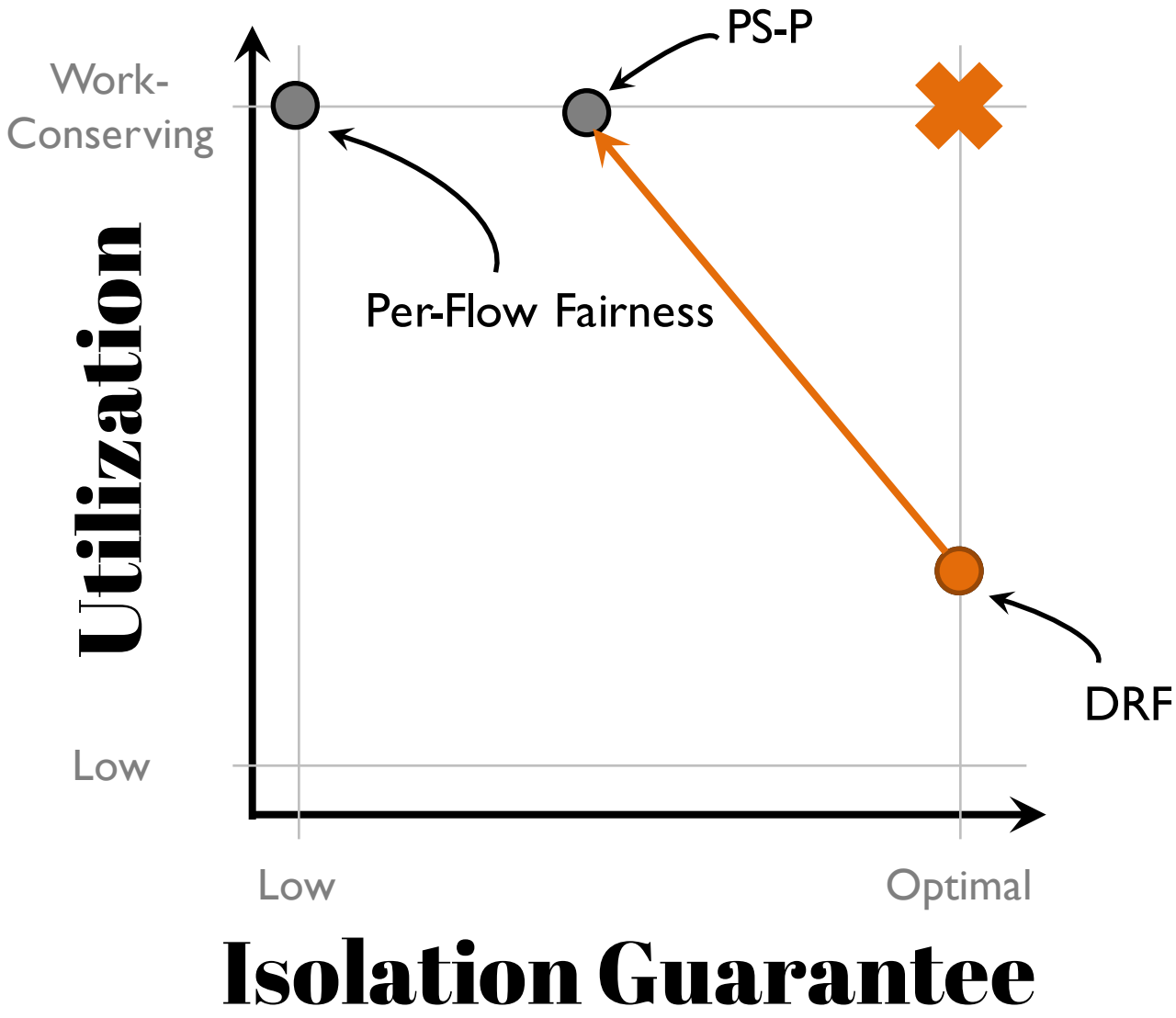
$\boldsymbol{d_B} = \langle 1, 1/6 \rangle$ $\boldsymbol{a_B} = \langle ½, 1/12 \rangle$ $\boldsymbol{M'_B} = 1/2$

Prisoner's Dilemma

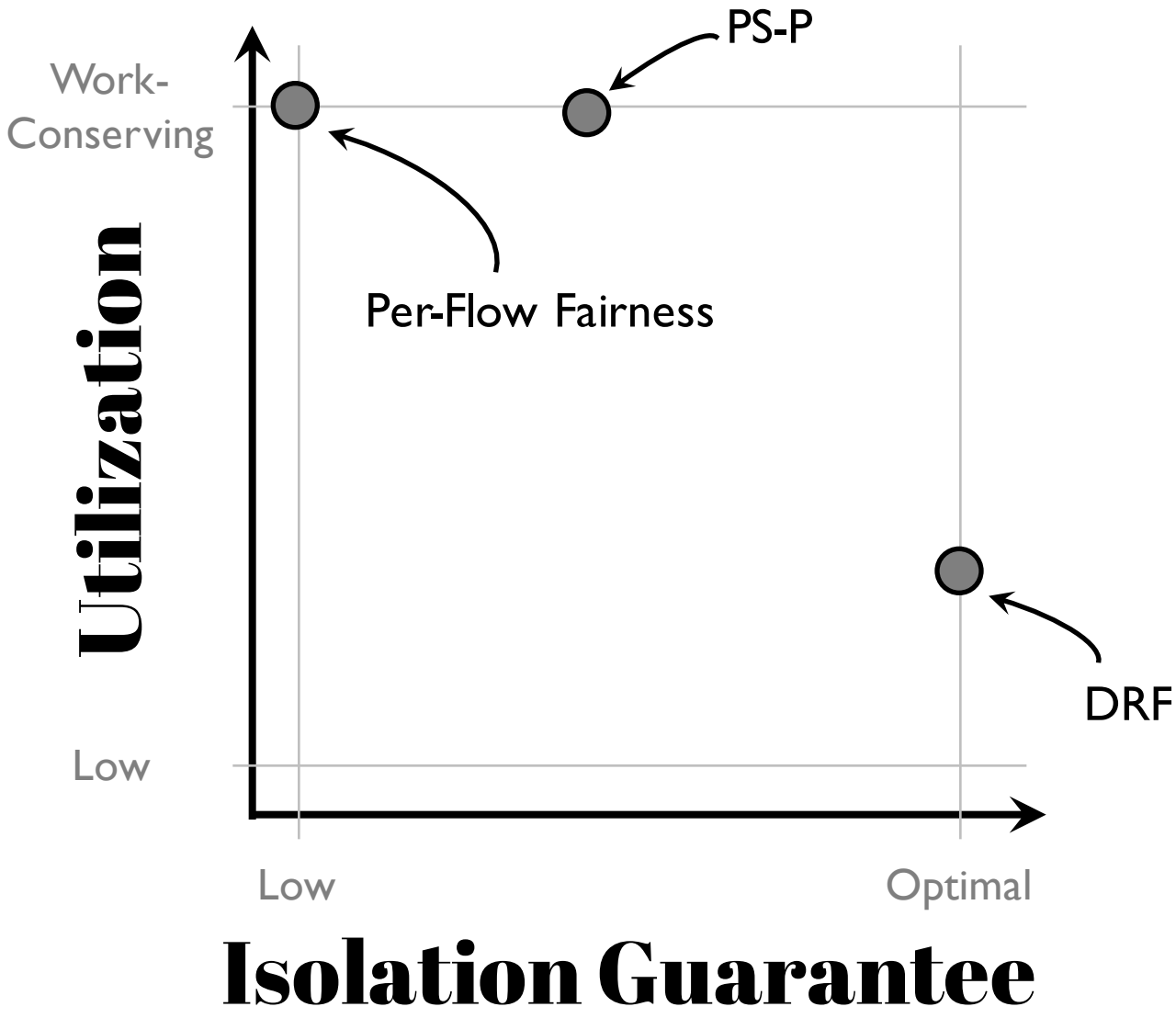


1. Why not?

Optimal isolation guarantee depends on being strategyproof, but work conservation cannot coexist with strategyproof-ness

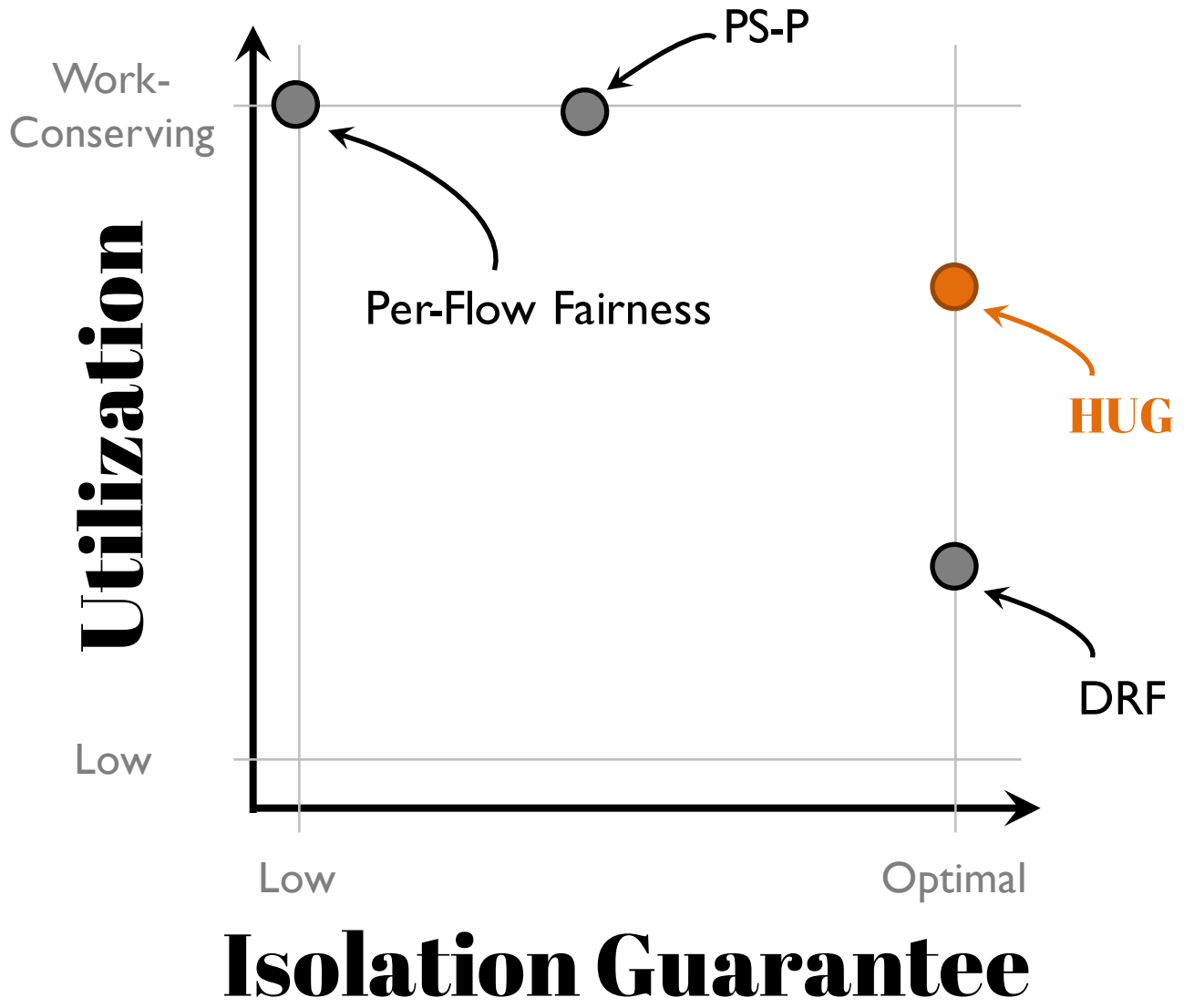


1. Why not?
2. **What's the best we can achieve?**
3. How can we achieve that?
4. Does it matter?



HUG in Non-Cooperative Setting

1. Optimal Isolation Guarantee
2. Highest Utilization
3. Strategyproof



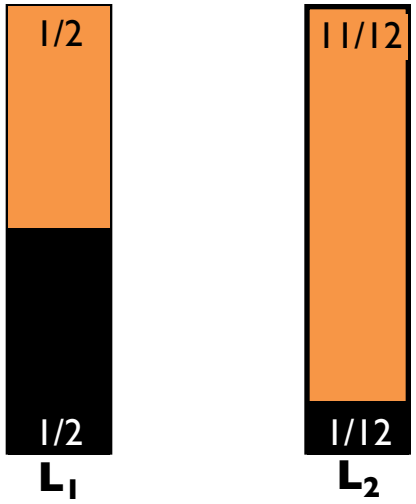
HUG

Highest Utilization with the Optimal Isolation Guarantee

Restrict a tenant's allocation in any link to its allocation in the bottleneck link

Tenant-A Lied

Tenant-A ***lies*** and asks for ***one*** unit of L_1 for ***each*** of L_2



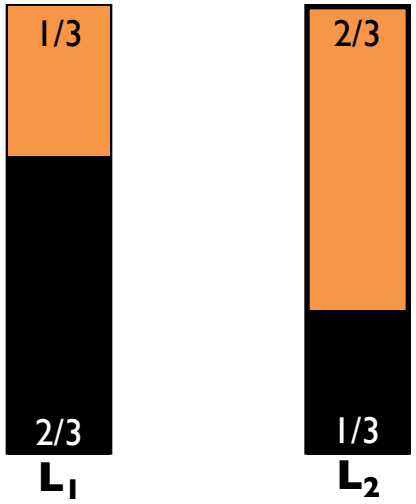
Tenant-B wants ***at least 1/6*** unit of L_2 for ***each*** of L_1

$\boldsymbol{d'_A} = \langle 1, 1 \rangle$ $\boldsymbol{a'_A} = \langle ½, ½ \rangle$ $\boldsymbol{a'_A} = \langle ½, 11/12 \rangle$ $\boldsymbol{M'_A} = 1/2$

$\boldsymbol{d_B} = \langle 1, 1/6 \rangle$ $\boldsymbol{a_B} = \langle ½, 1/12 \rangle$ $\boldsymbol{M'_B} = 1/2$

Everyone is Forced to Tell the Truth

Tenant-A wants ***at least half*** unit of L_1 for ***each*** of L_2



Tenant-B wants ***at least 1/6*** unit of L_2 for ***each*** of L_1

$d_A = \langle 1/2, 1 \rangle$ $a_A = \langle 1/3, 2/3 \rangle$ $M_A = 2/3$

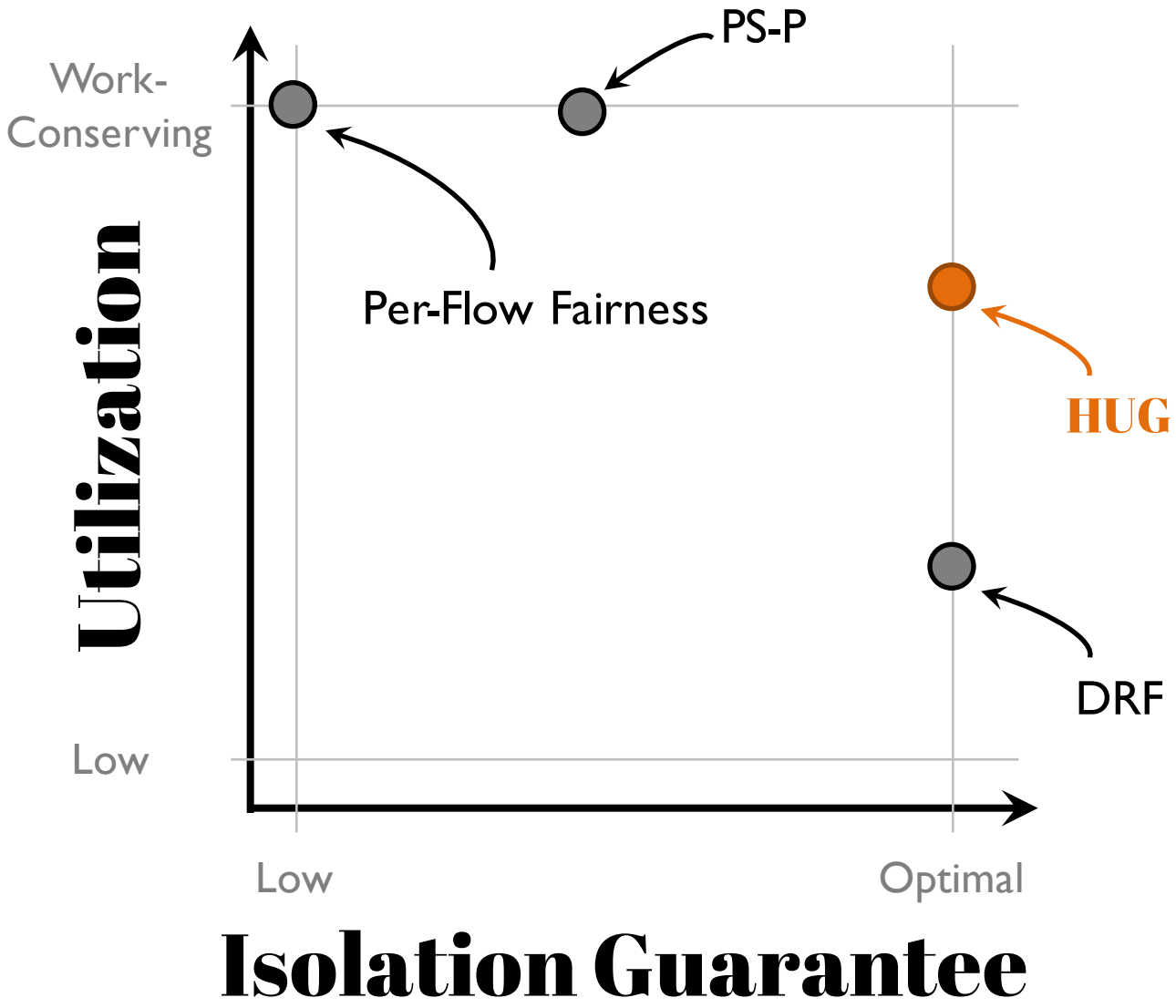
$d_B = \langle 1, 1/6 \rangle$ $a_B = \langle 2/3, 1/9 \rangle$ $a'_B = \langle 2/3, 1/3 \rangle$ $M_B = 2/3$

HUG

Highest Utilization with the Optimal Isolation Guarantee

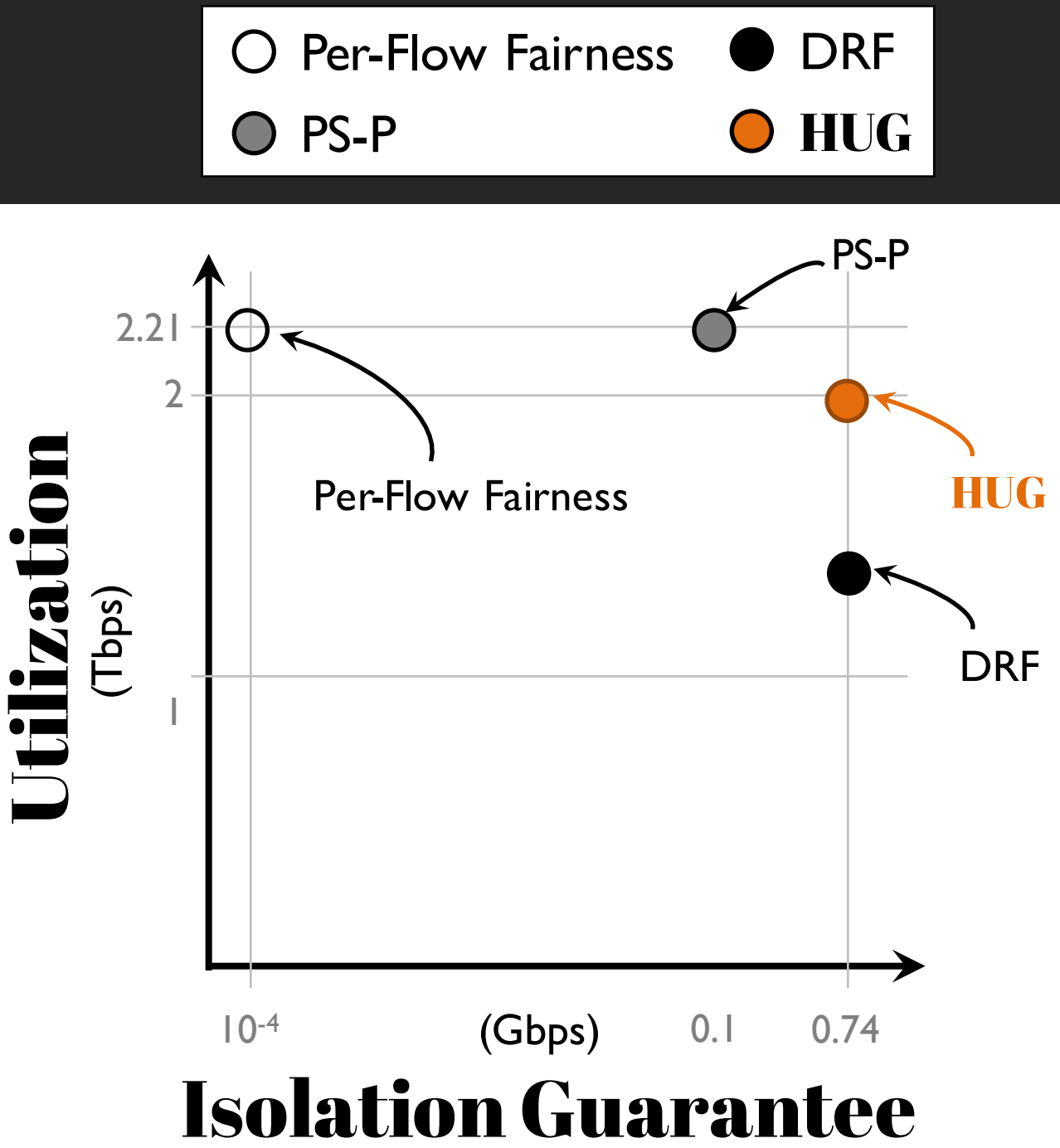
1. Tenants update correlation vectors through an API
2. Operators calculate HUG centrally and enforce it locally

1. Why not?
2. What's the best we can achieve?
3. How can we achieve that?
4. **Does it matter?**



Evaluation

- 100 concurrent tenants
- 3000 machines with 3Tbps total capacity
- Original placement and communication patterns from the Facebook trace



#1

Bursty Demands

Periodic demand bursts in Spark streaming

#2

Long-Term Guarantees

Predictable performance guarantees over time

#3

Decentralized Algorithms

Survive master failures and enable low response times

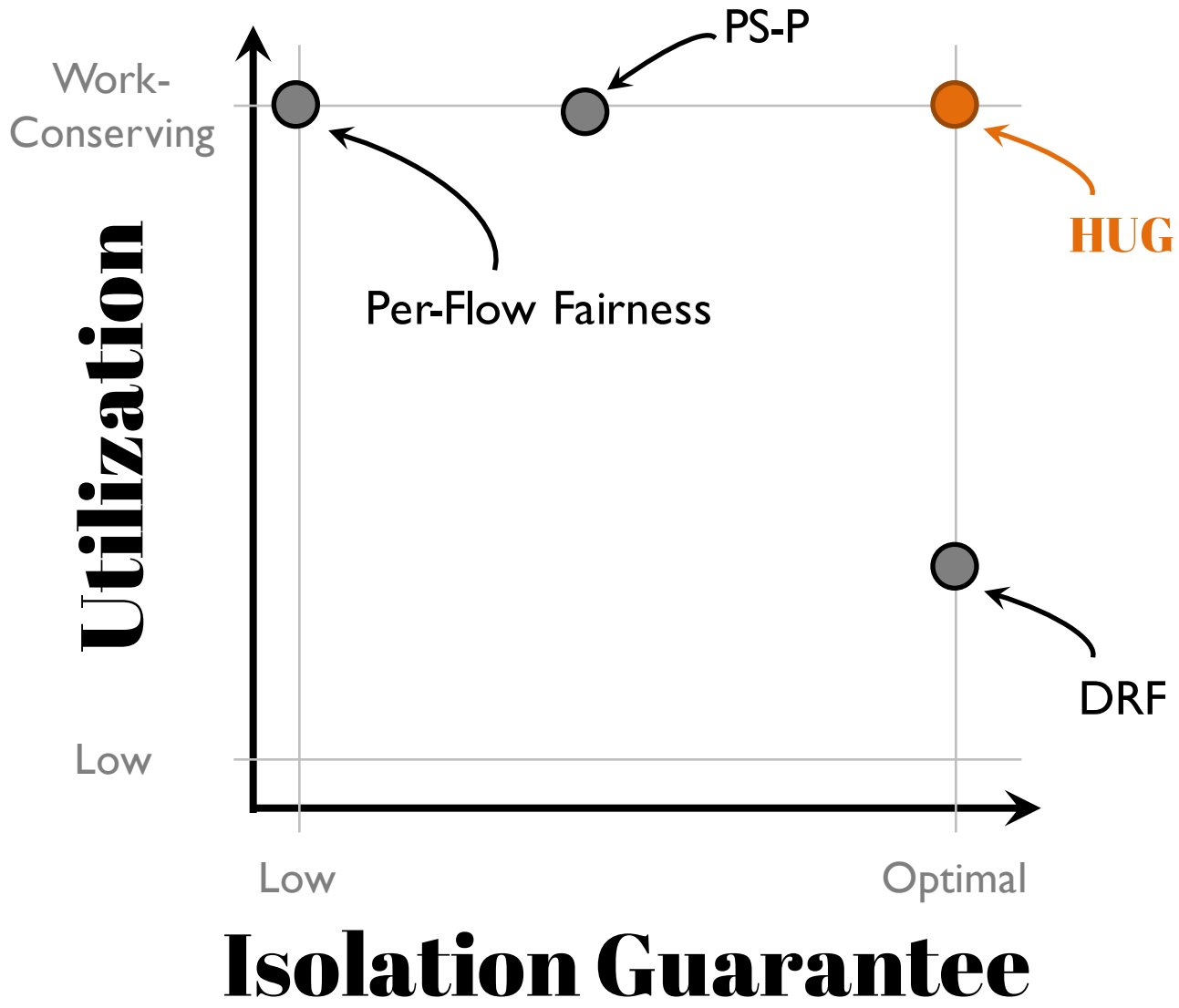
HUG

Highest Utilization with the Optimal Isolation Guarantee

- Generalizes single- and multi-resource fairness schemes
- Optimal worst-case performance guarantees for tenants
- Highest utilization for operators

HUG in Cooperative Setting

1. Optimal Isolation Guarantee
2. Work Conservation



Evaluation

A 3000-machine trace-driven simulation based on a snapshot of Facebook production trace

1. Does it provide isolation guarantee?
2. Does it improve utilization?
3. Is it practical?

YES

Optimal Progress for ALL

	Per-Flow Fairness	PS-P[2]	DRF[3]	HUG
Max	1	1	0.74	0.74
Min	0.0001	0.10	0.74	0.74
Max-to-Min Progress Ratio	10000X	10X	1X	1X

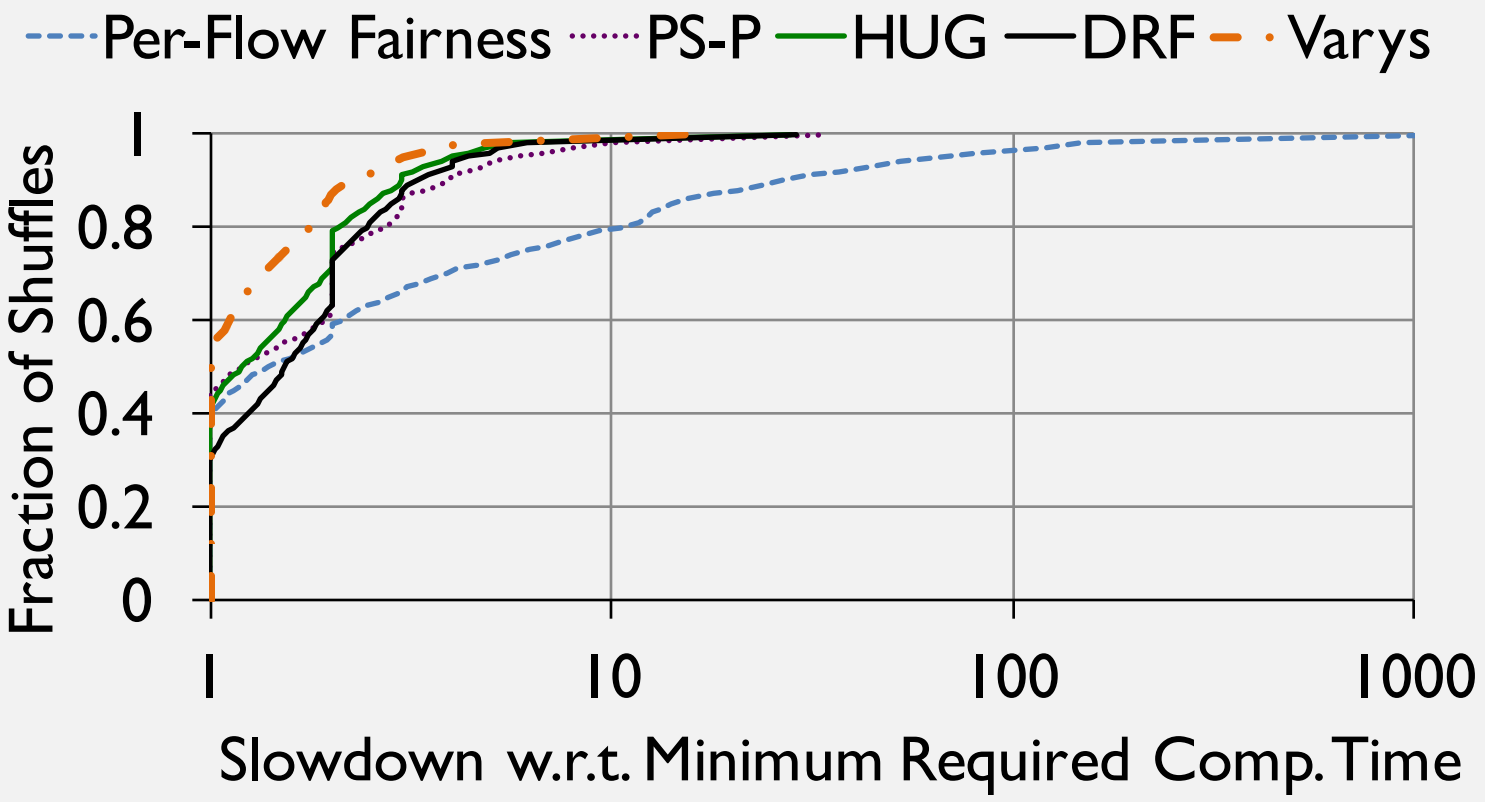
1. 100 tenants in this particular snapshot. The unit of progress is Gbps.

Higher Network Utilization

	Per-Flow Fairness	PS-P[2]	DRF[3]	HUG
Total Utilization (Tbps)	**2.21**	**2.20**	**1.42**	**2.00**
Max-to-Min Progress Ratio	**10000X**	**10X**	**1X**	**1X**

1. 100 tenants in this particular snapshot. The unit of progress is Gbps.
2. FairCloud: Sharing the Network in Cloud Computing, SIGCOMM'12
3. Dominant Resource Fairness: Fair Allocation of Multiple Resource Types, NSDI'11

Long-Term Performance



	Average Time
Per-flow Fairness	1.49X
PS-P	1.14X
DRF	1.14X
HUG	1X
Varys[1]	0.69X

1. Efficient Coflow Scheduling with Varys, SIGCOMM'14

Coordination Overheads and Scalability

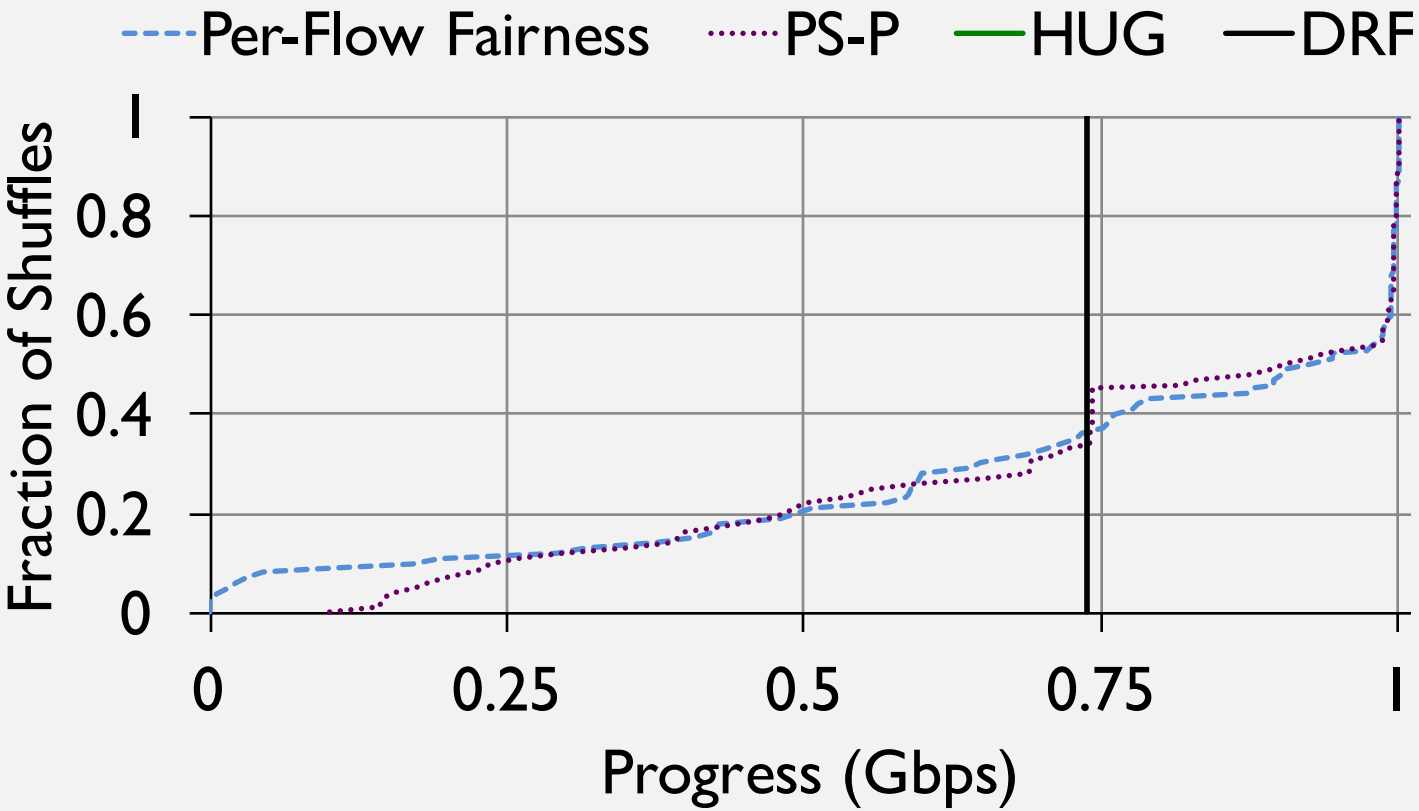
Computation overheads

- Less than 5 μs for 100-machine cluster
- Less than 10 ms for 100,000 machines

Communication overheads

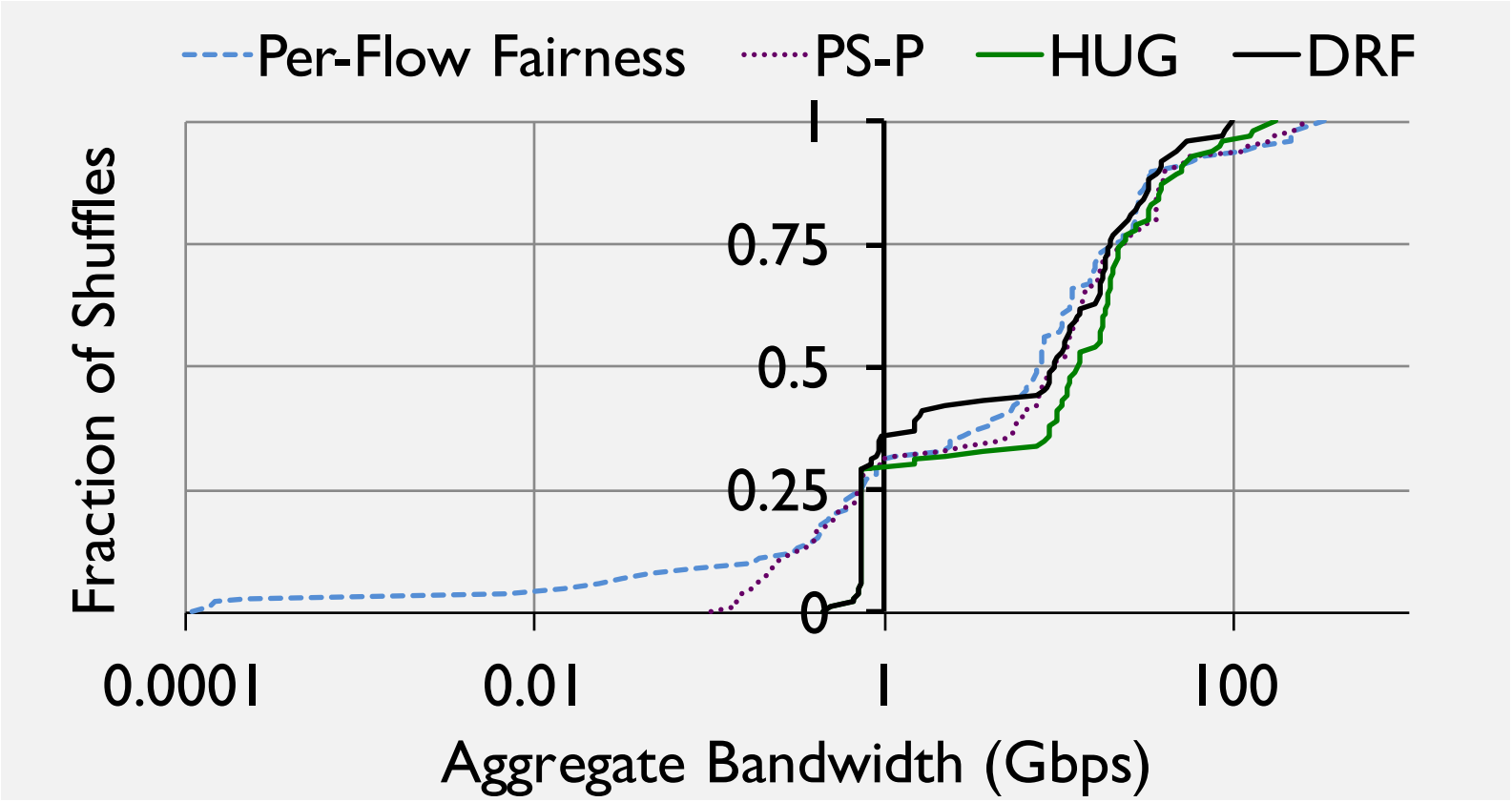
- Less than 10 ms for 100-machine cluster
- Less than 1 second for 100,000 machines

Optimal Progress for ALL



	Max/Min Ratio
Per-flow Fairness	10000X
PS-P	10X
DRF	1X
HUG	1X

Higher Utilization



	Max/Min Ratio
Per-flow Fairness	3240000X
PS-P	2590X
DRF	196X
HUG	340X