2DFQ: Two-Dimensional Fair Queuing for Multi-Tenant Cloud Services

Jonathan Mace
Brown University

Peter Bodik
Microsoft

Madanlal Musuvathi
Microsoft

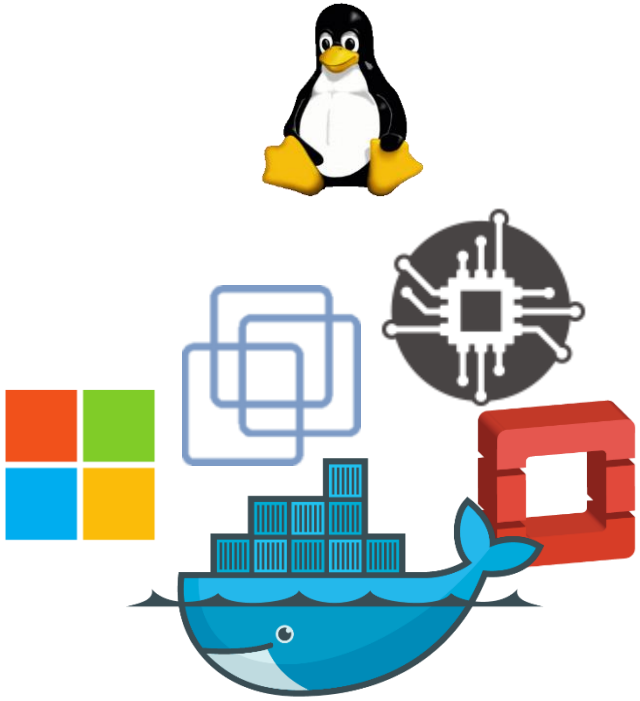
Rodrigo Fonseca
Brown University

Krishnan Varadarajan
Microsoft

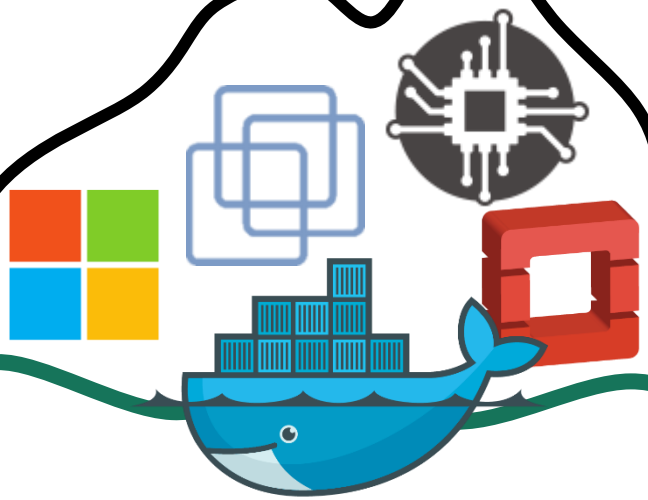
BROWN UNIVERSITY

Microsoft Research

Containers / VMs

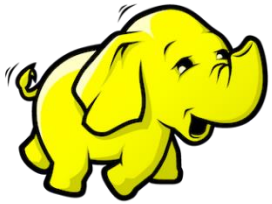


Containers / VMs

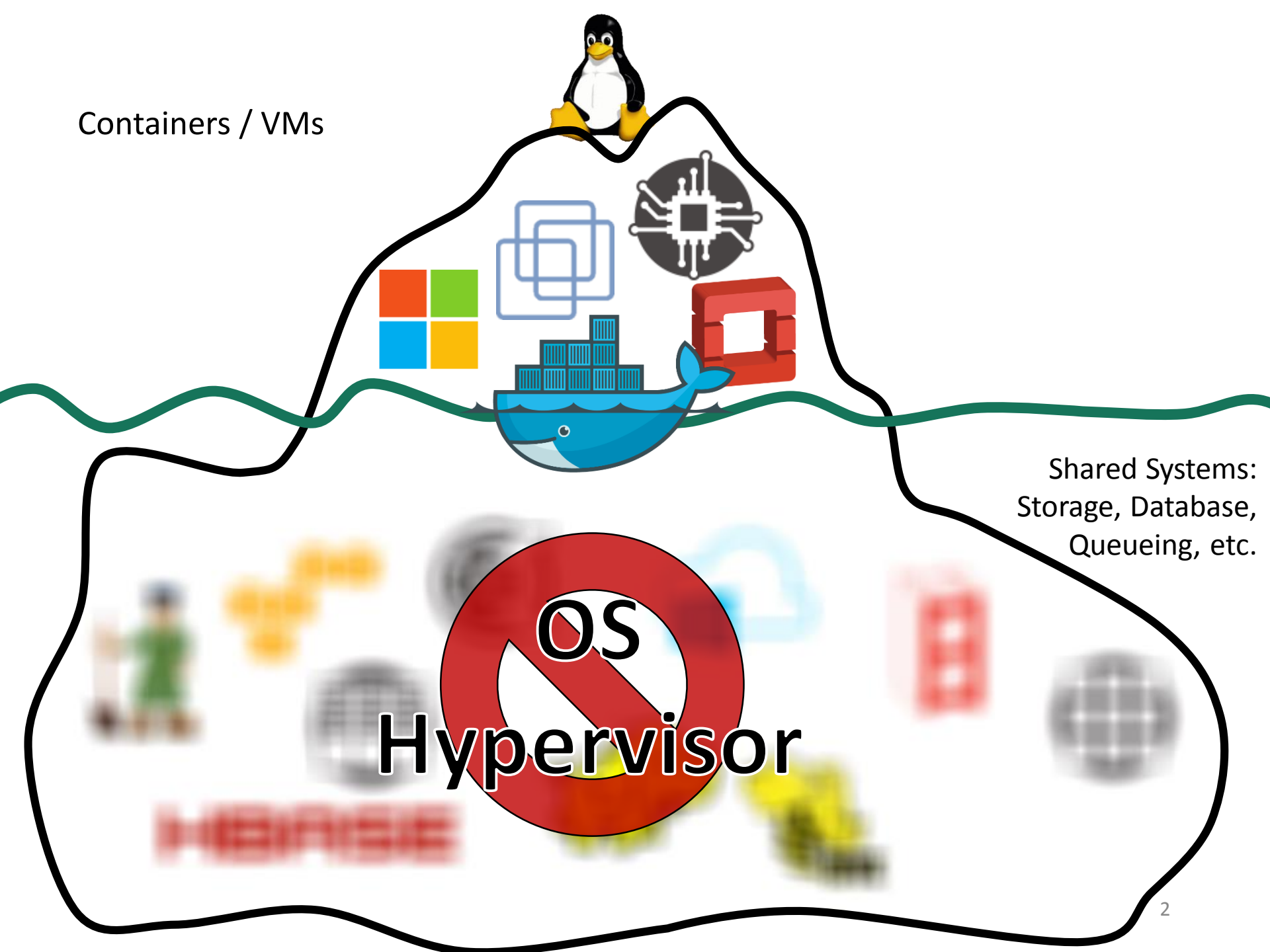


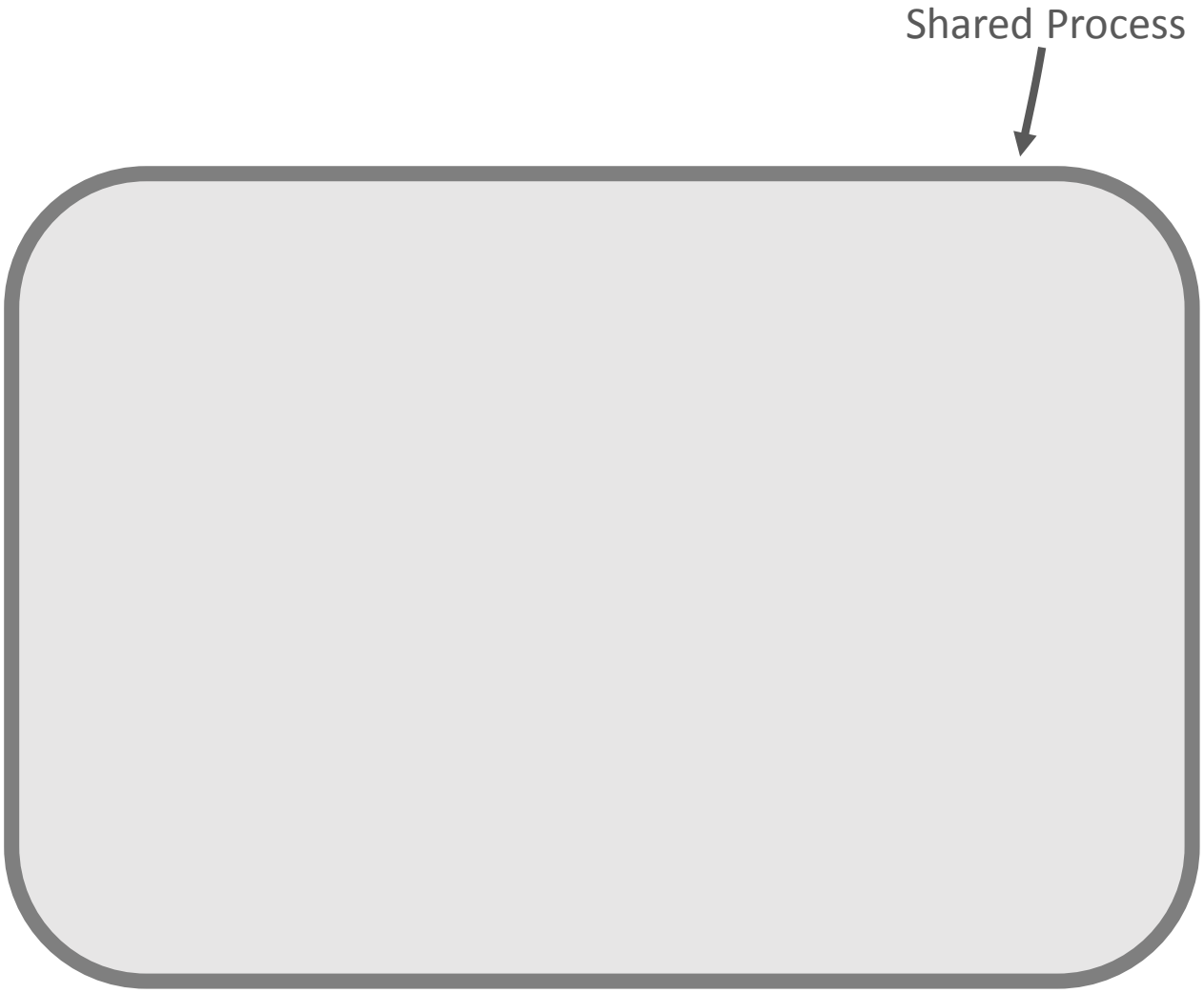
Shared Systems:
Storage, Database,
Queueing, etc.

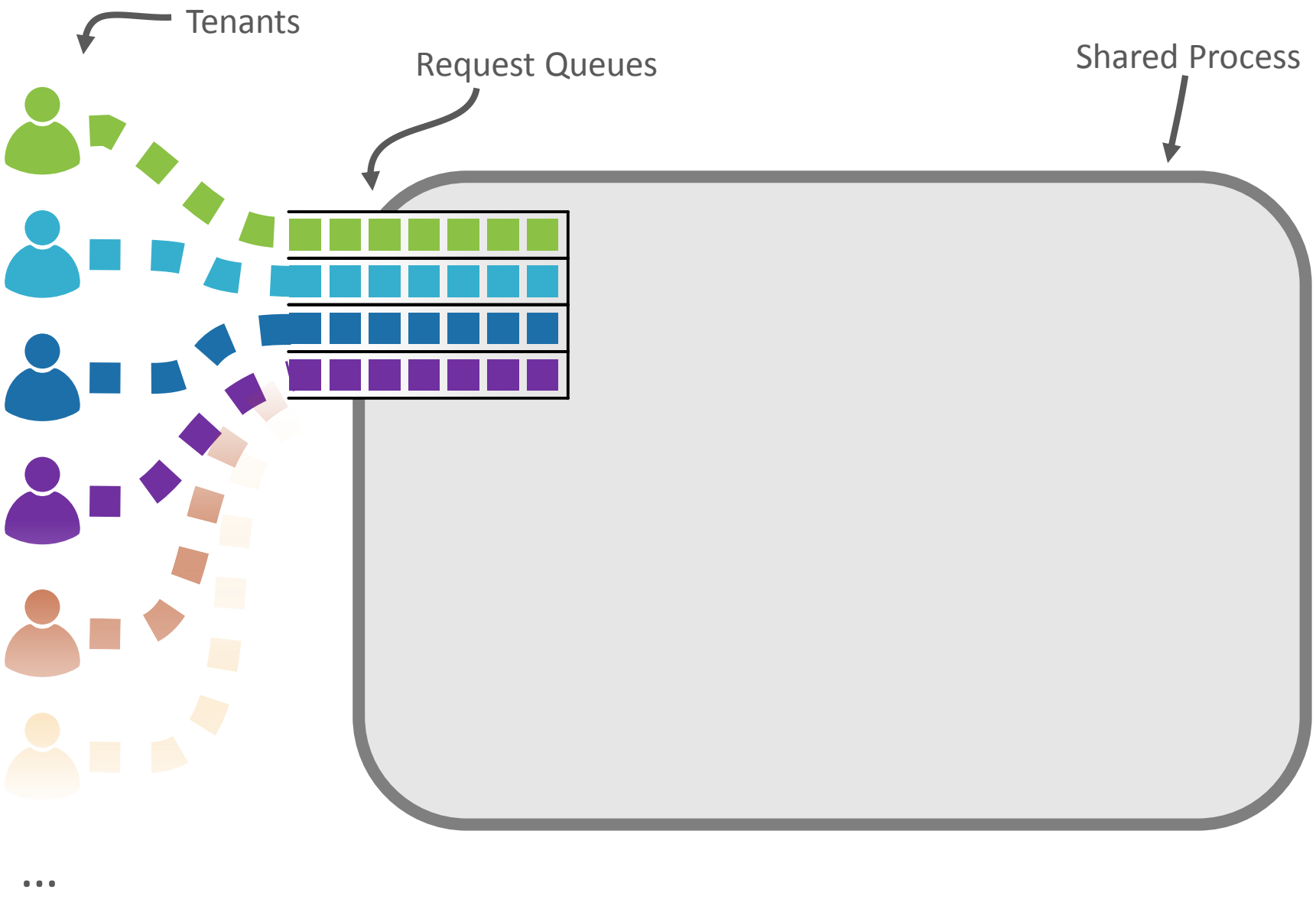
HBASE

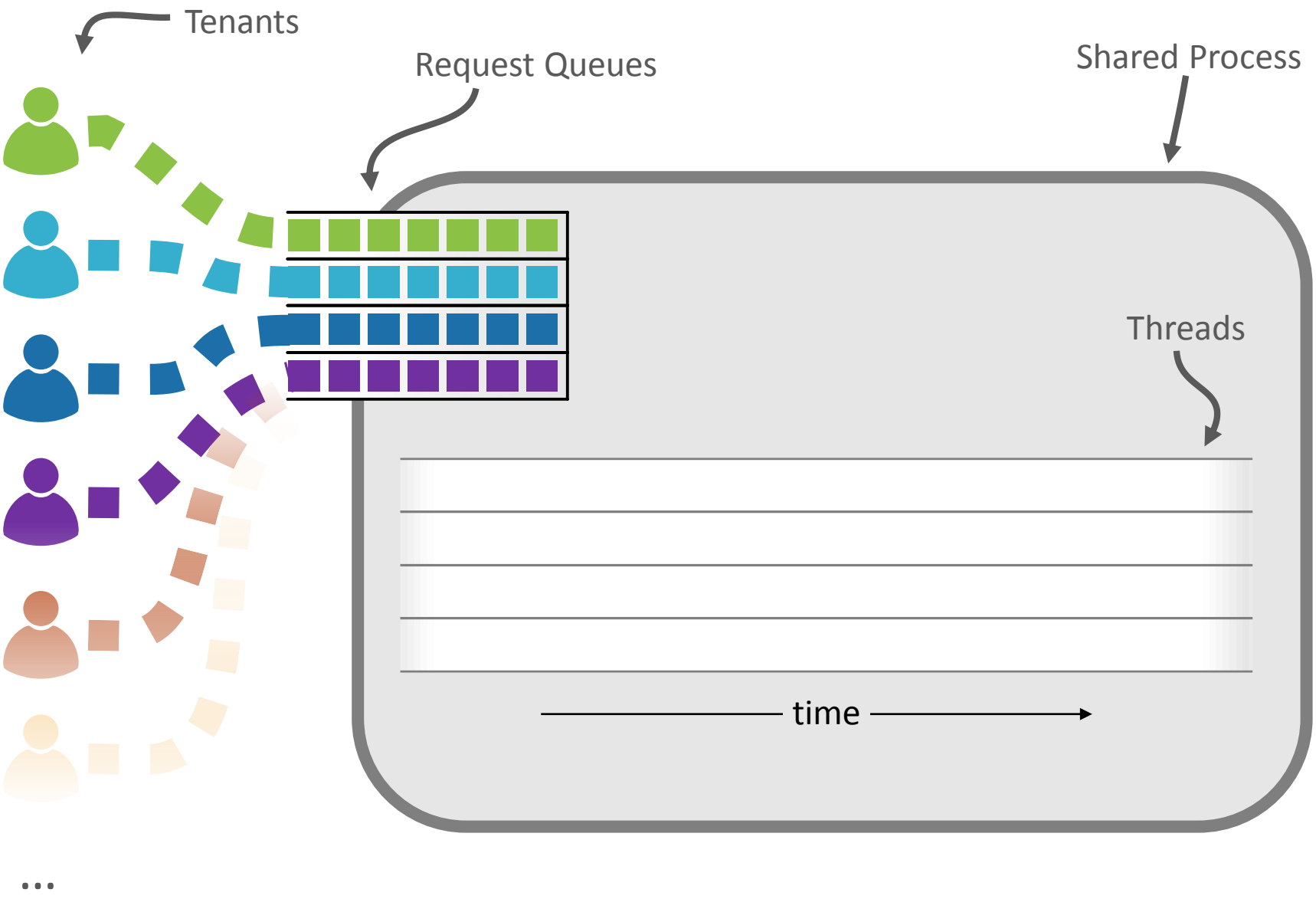


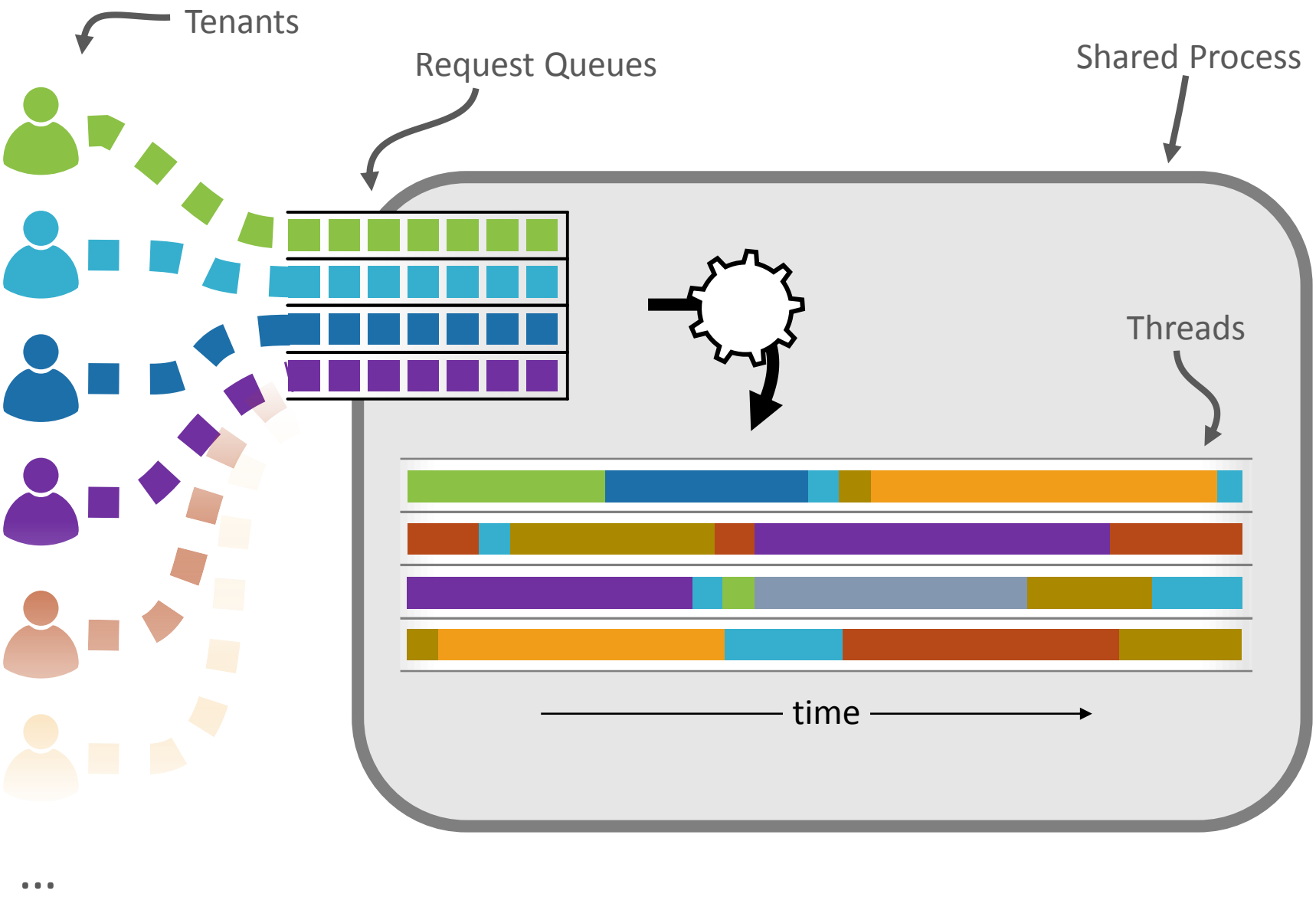
HIVE

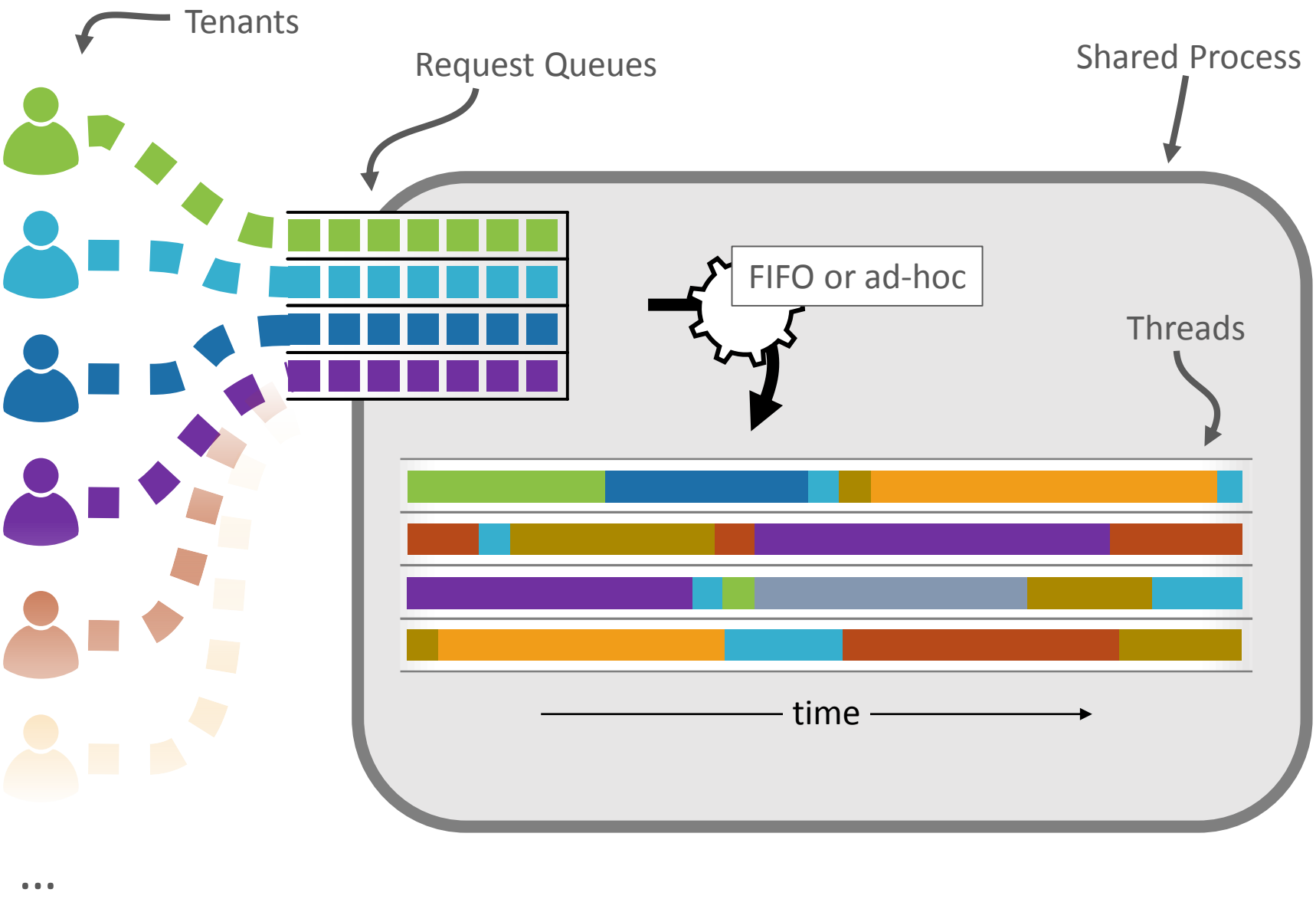

Containers / VMs
Shared Systems:
Storage, Database,
Queueing, etc.
OS
Hypervisor

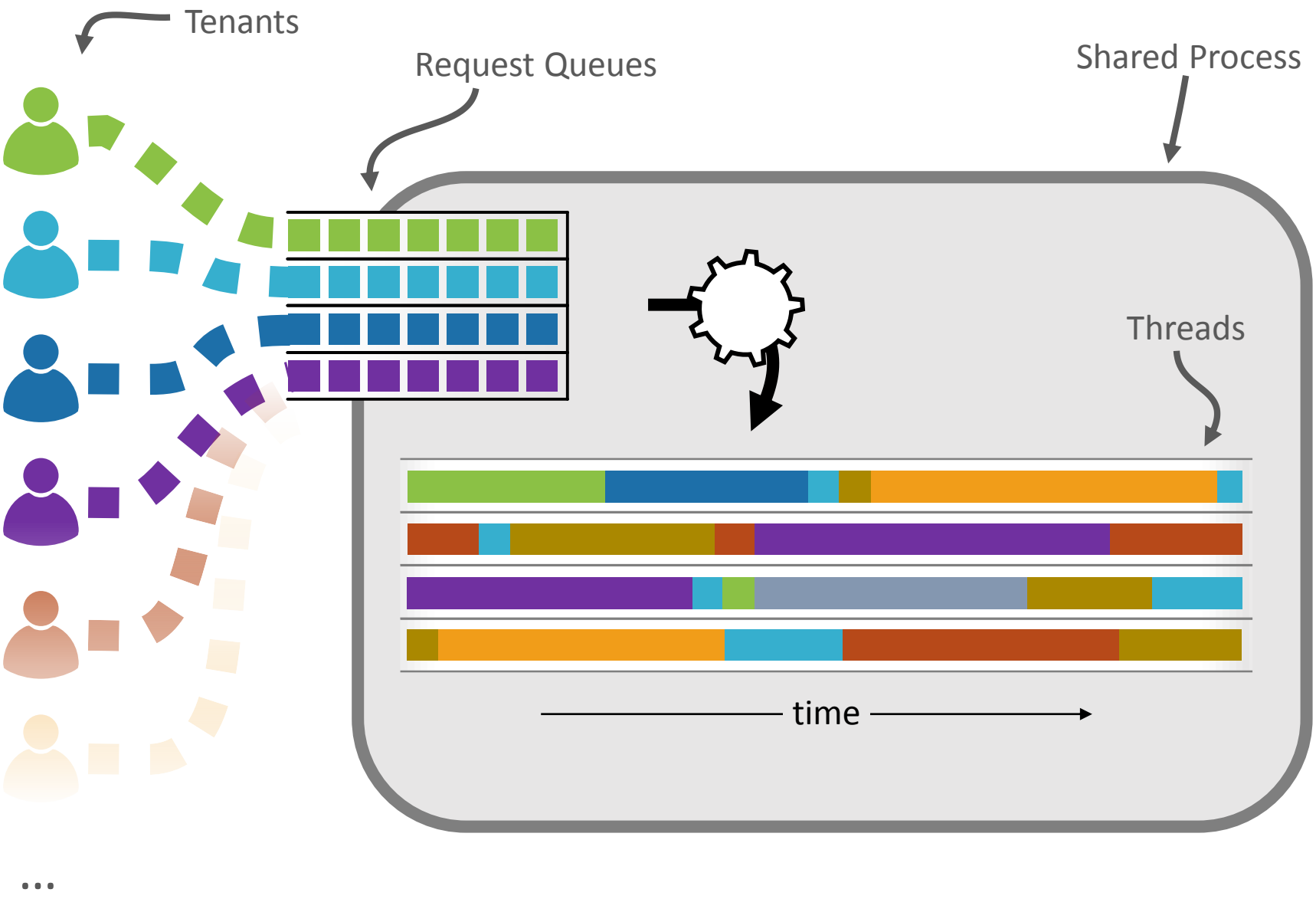

Shared Process

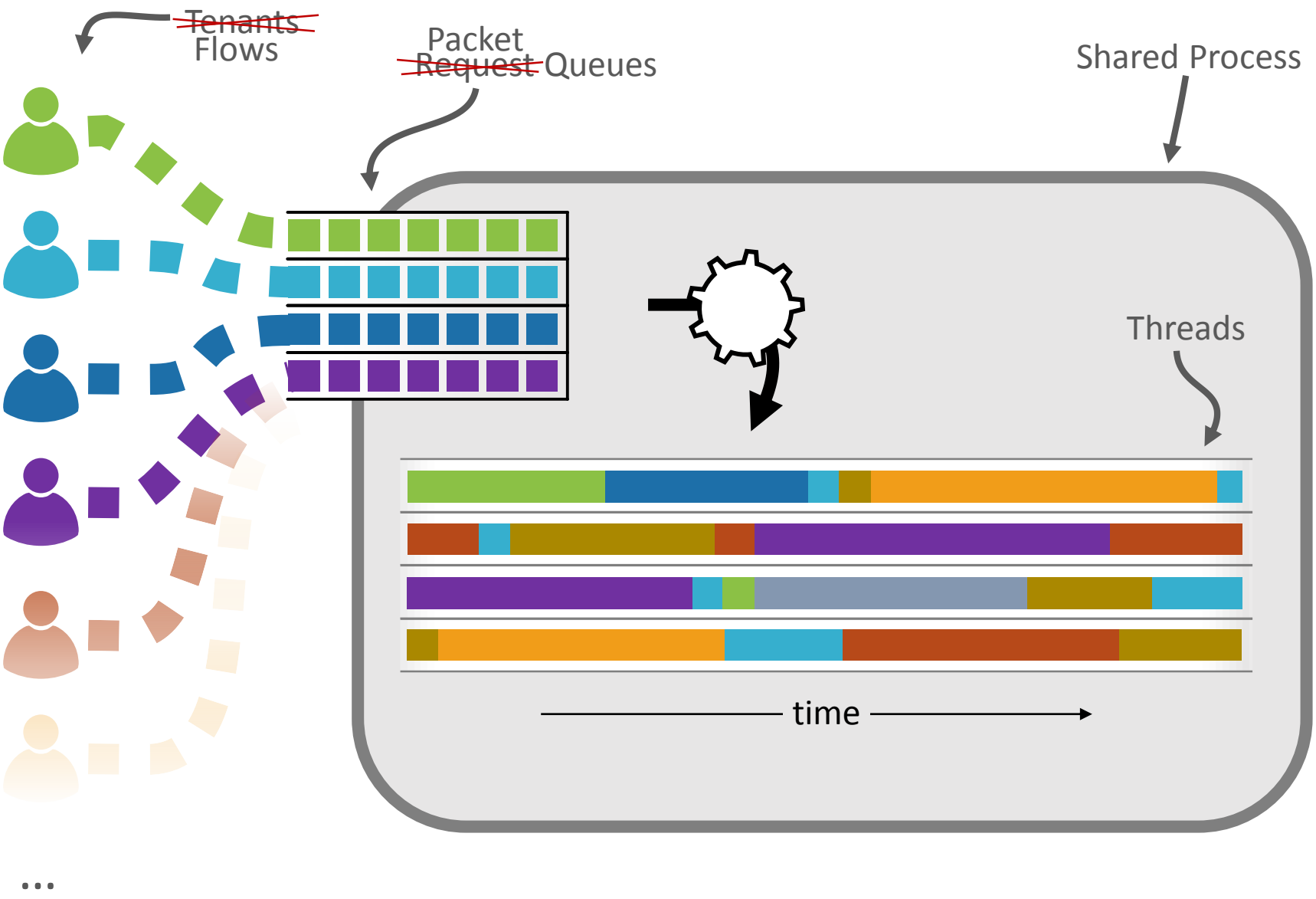

Tenants
Request Queues
Shared Process
...

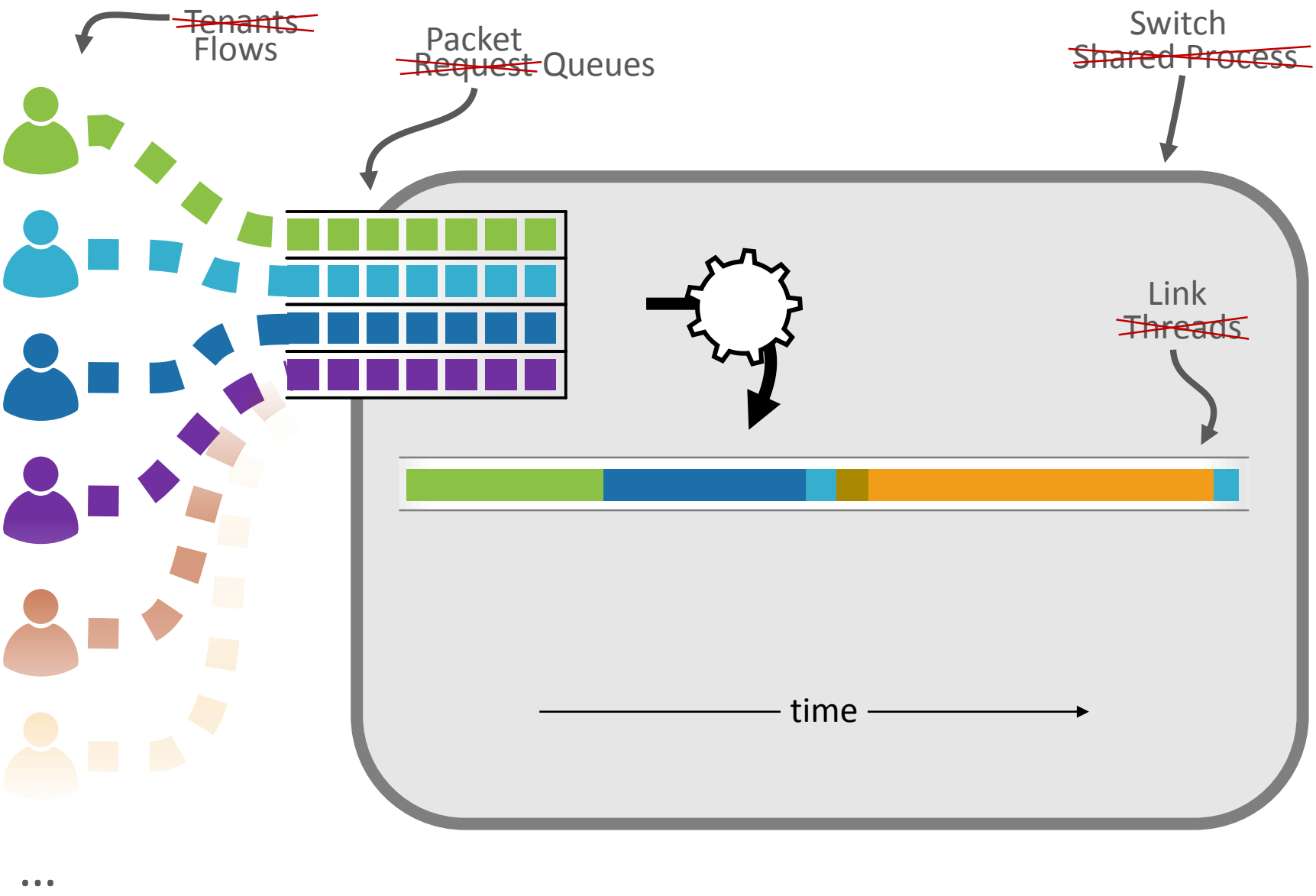

Tenants
Request Queues
Shared Process
Threads
time
...

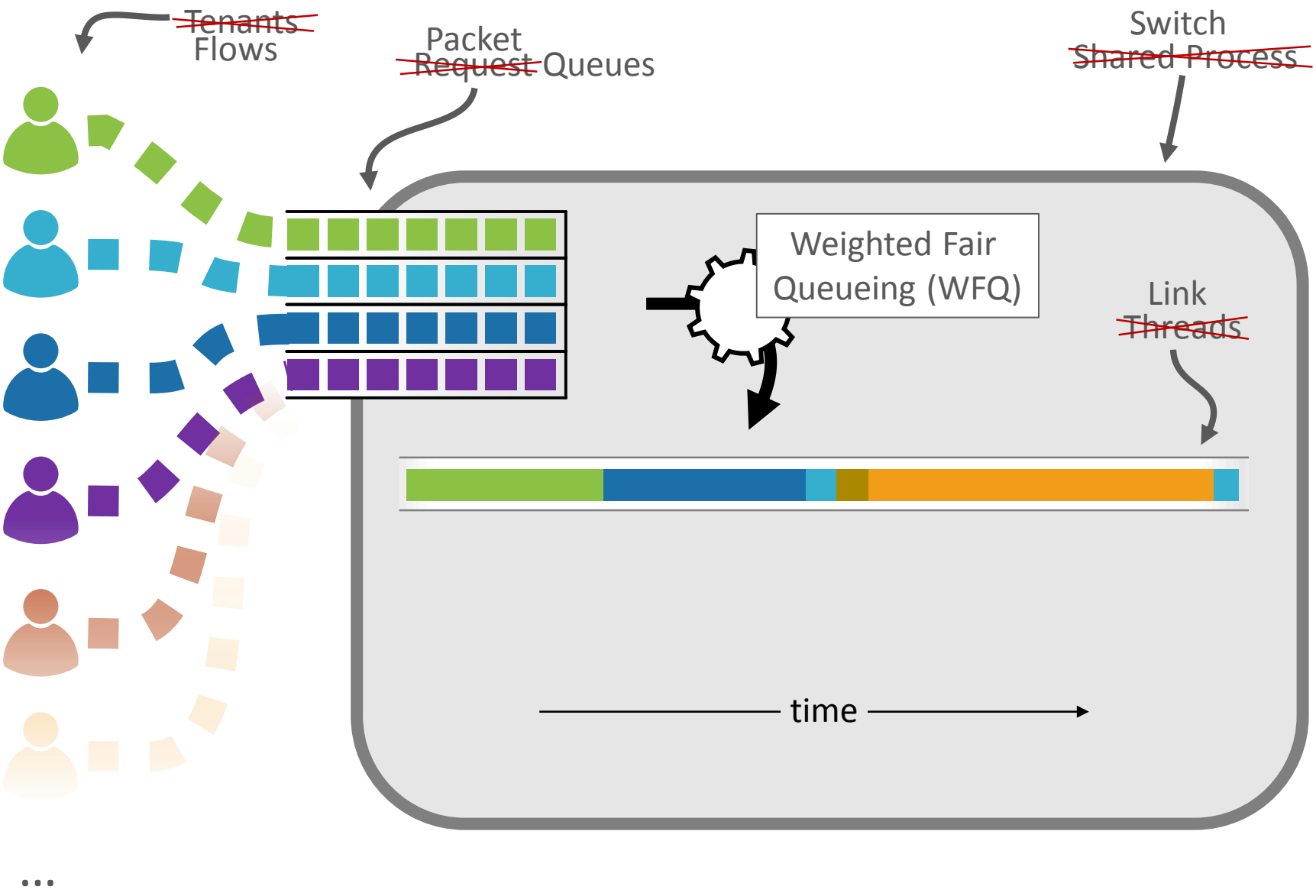

Tenants
Request Queues
Shared Process
Threads
time
...

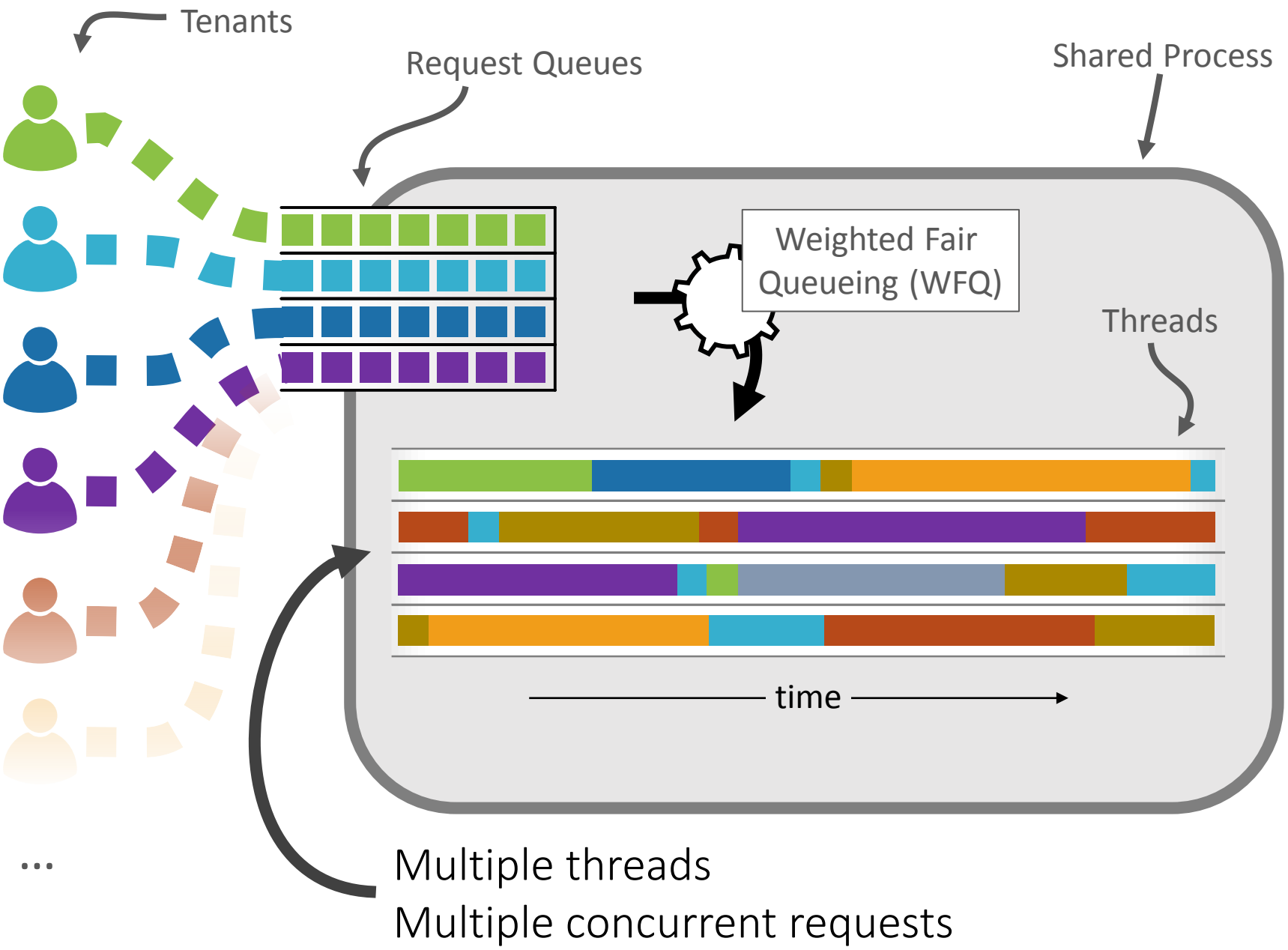

Tenants
Request Queues
Shared Process
FIFO or ad-hoc
Threads
time
...

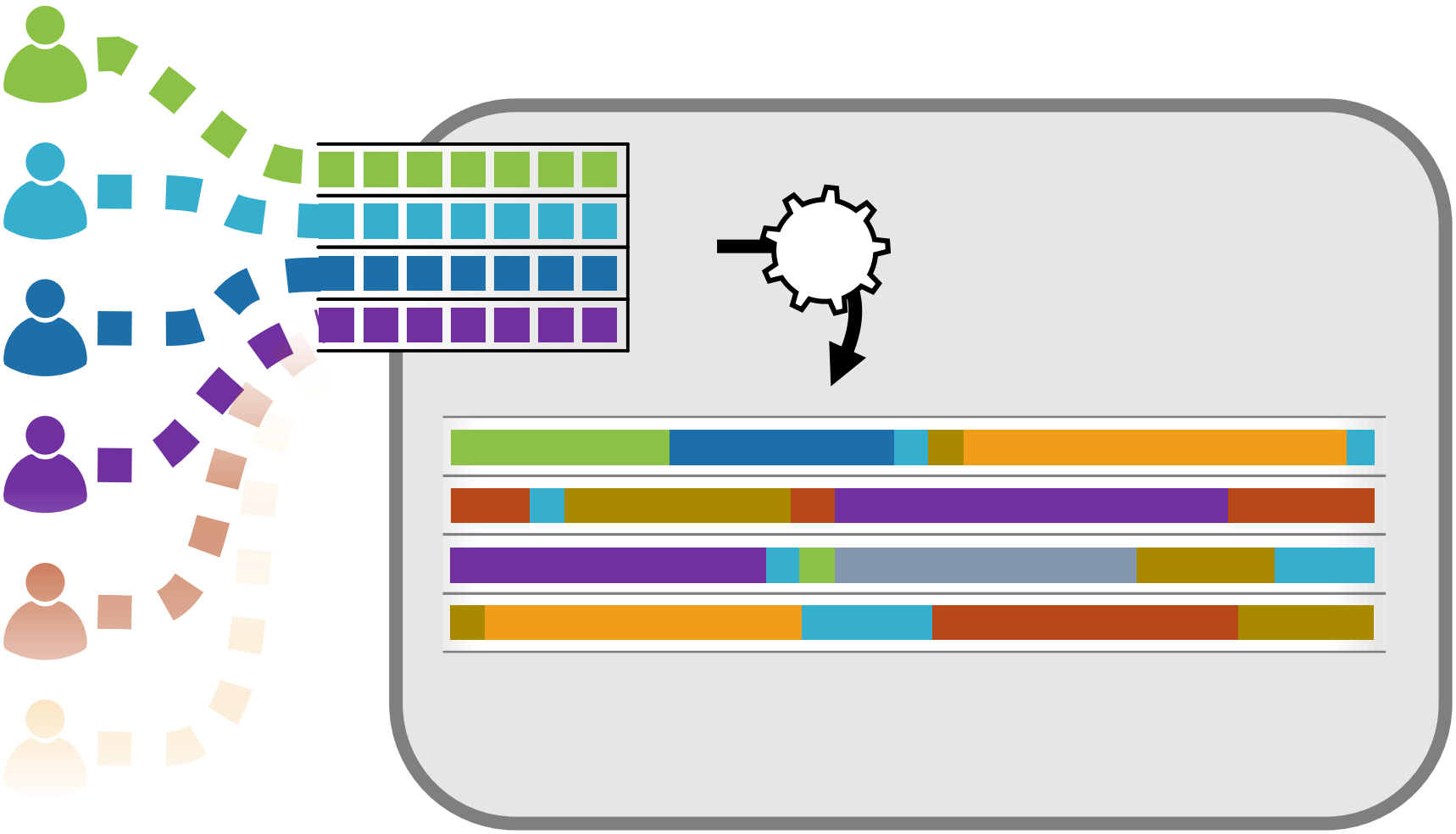

Tenants
Request Queues
Shared Process
Threads
time
...


~~Tenants~~
Flows
Packet
~~Request~~ Queues
Shared Process
Threads
time
...

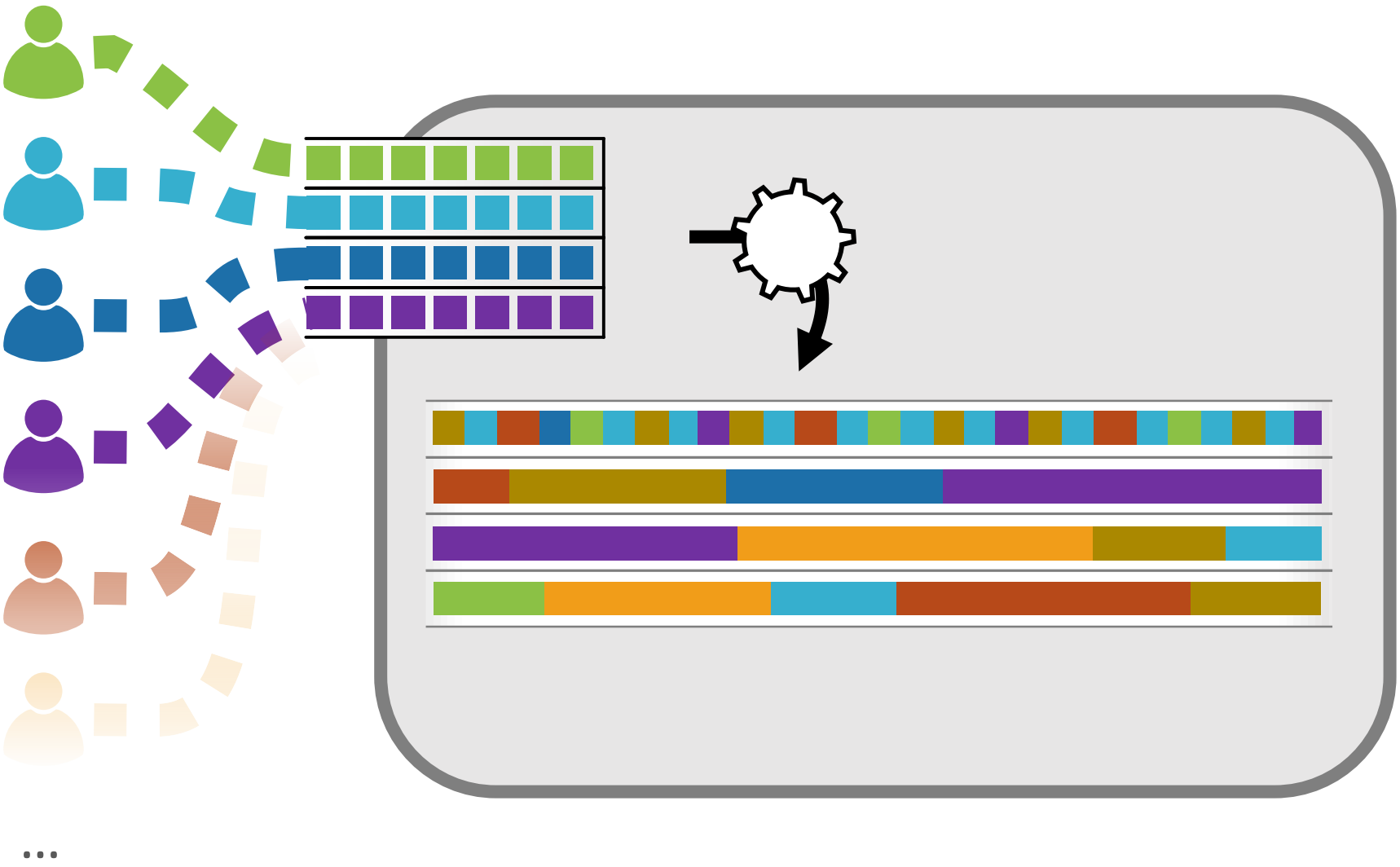

~~Tenants~~
Flows
Packet
~~Request~~ Queues
Switch
~~Shared Process~~
Link
~~Threads~~
time
...


~~Tenants~~
Flows
Packet
~~Request~~ Queues
Switch
~~Shared Process~~
Weighted Fair
Queueing (WFQ)
Link
~~Threads~~
time
...

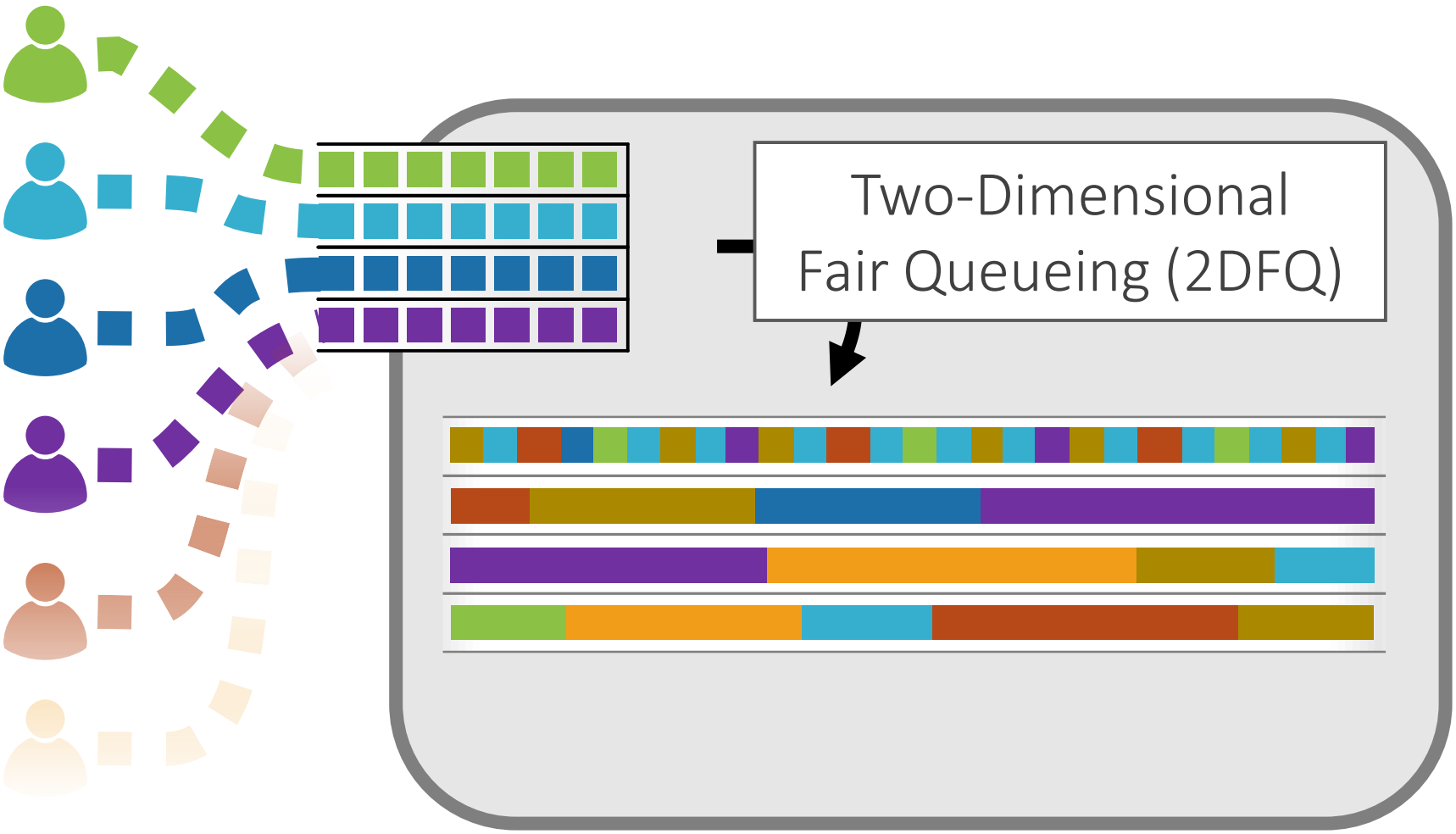

Tenants
Request Queues
Shared Process
Weighted Fair Queueing (WFQ)
Threads
time
...
Multiple threads
Multiple concurrent requests



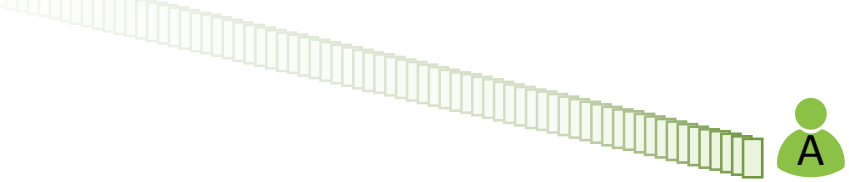
...



...

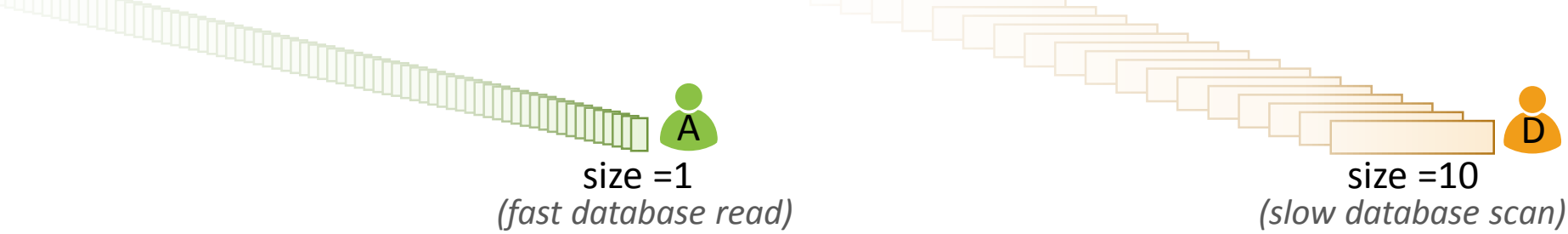

Two-Dimensional
Fair Queueing (2DFQ)

...



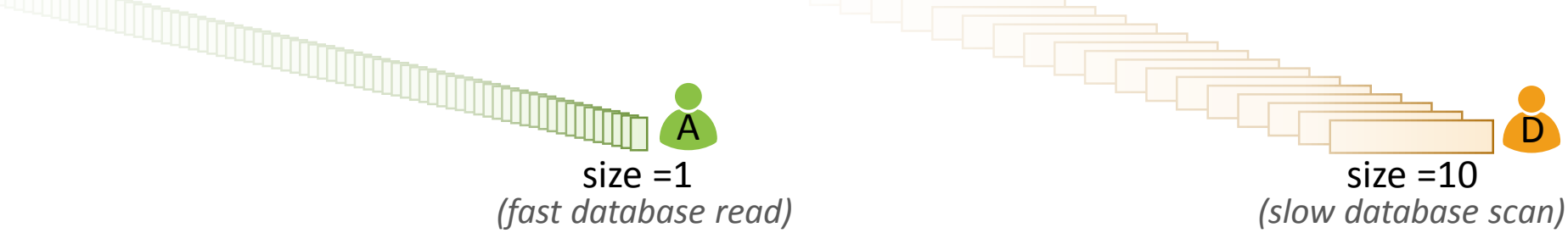
size =1

(fast database read)

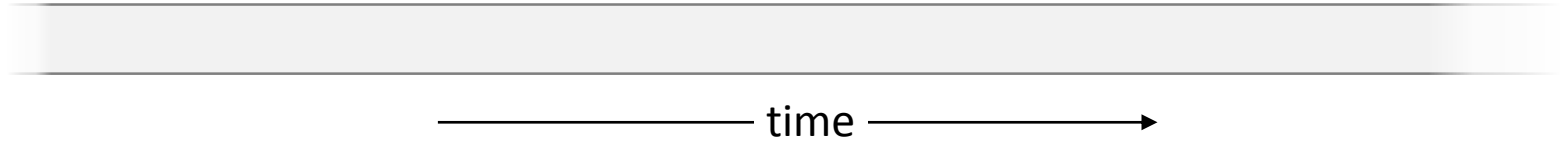


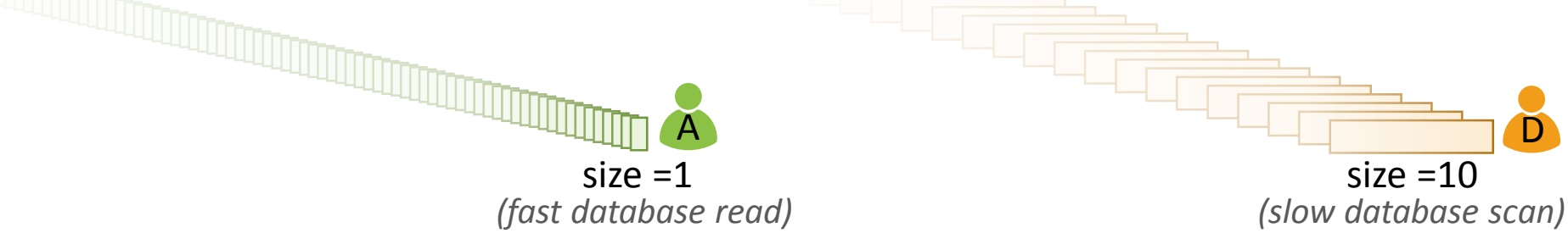
size =1
(fast database read)

size =10
(slow database scan)

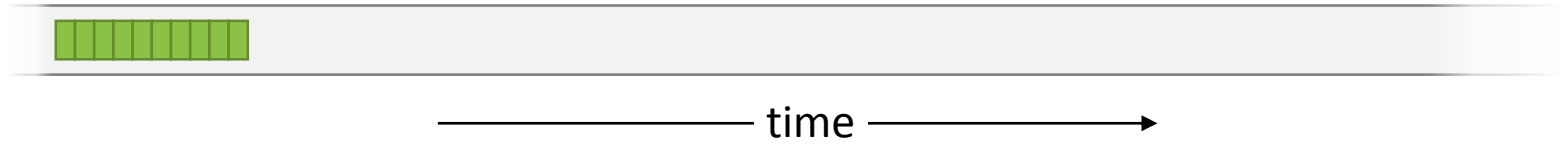


One thread:





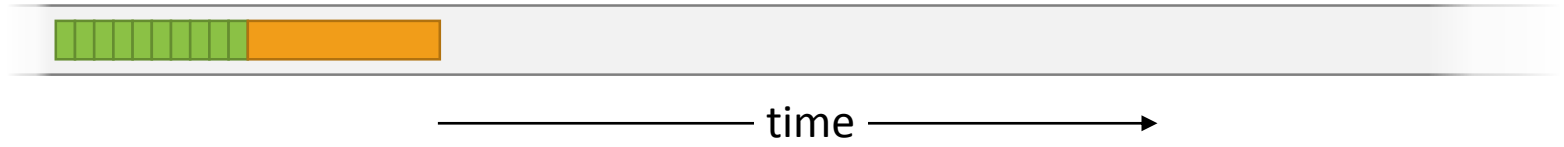
One thread:

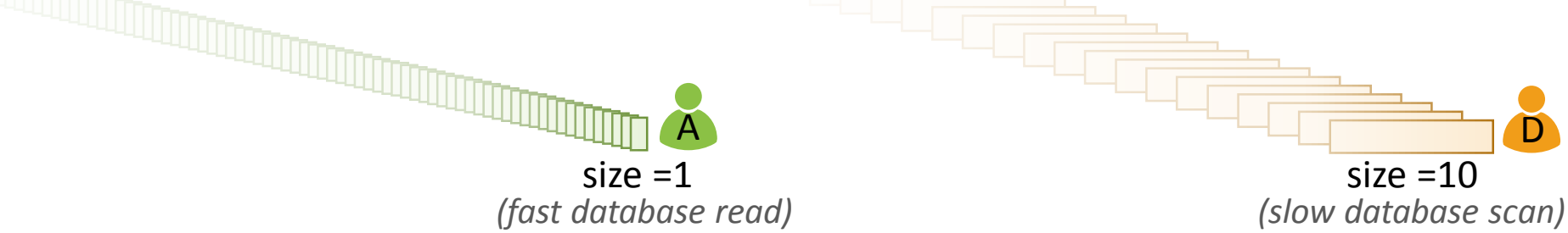


size =1
(fast database read)

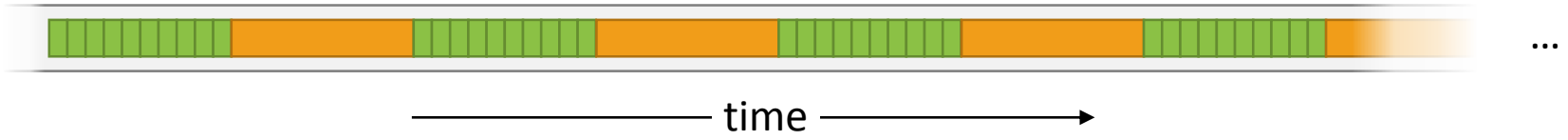
size =10
(slow database scan)

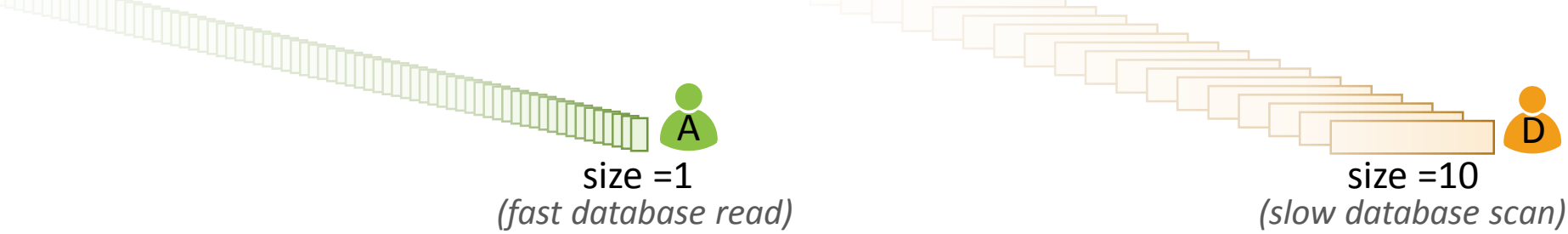
One thread:



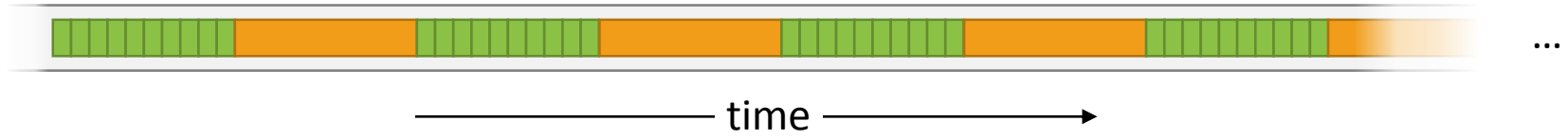


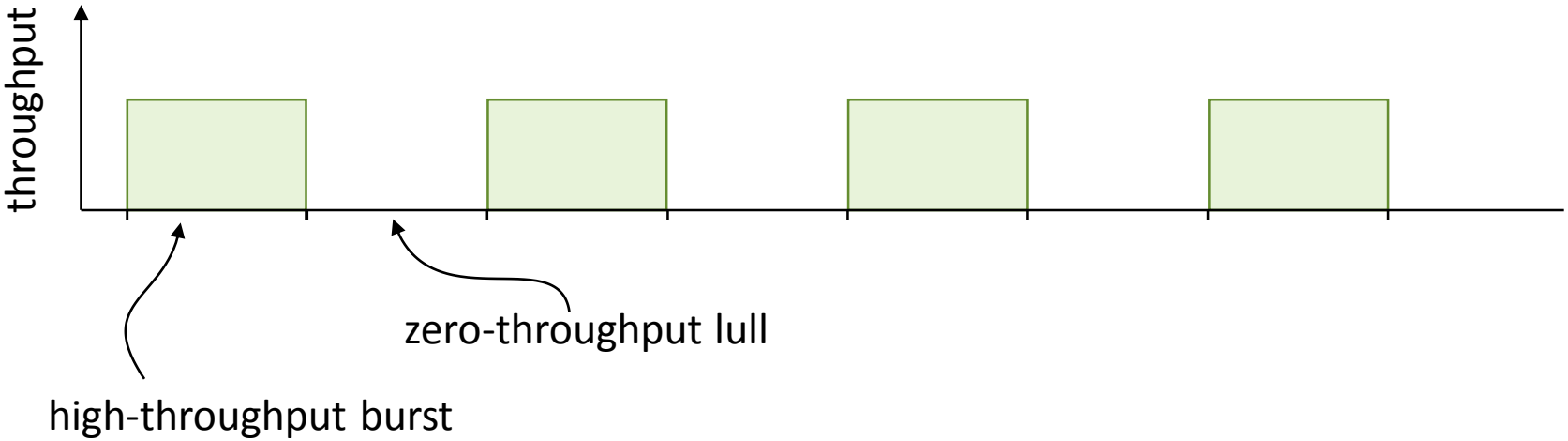
One thread:

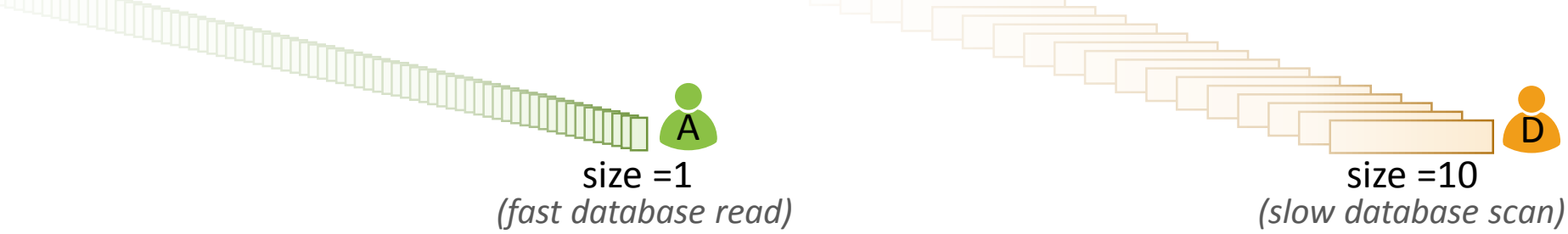



A
size =1
(fast database read)
D
size =10
(slow database scan)

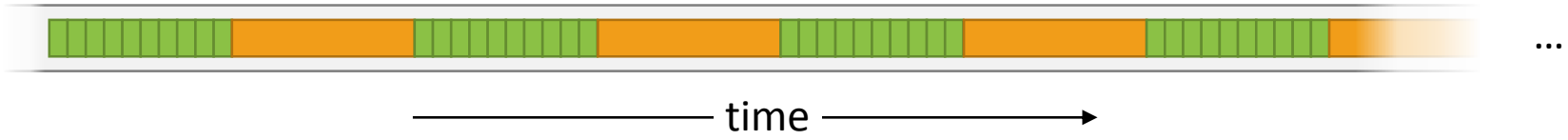
One thread:

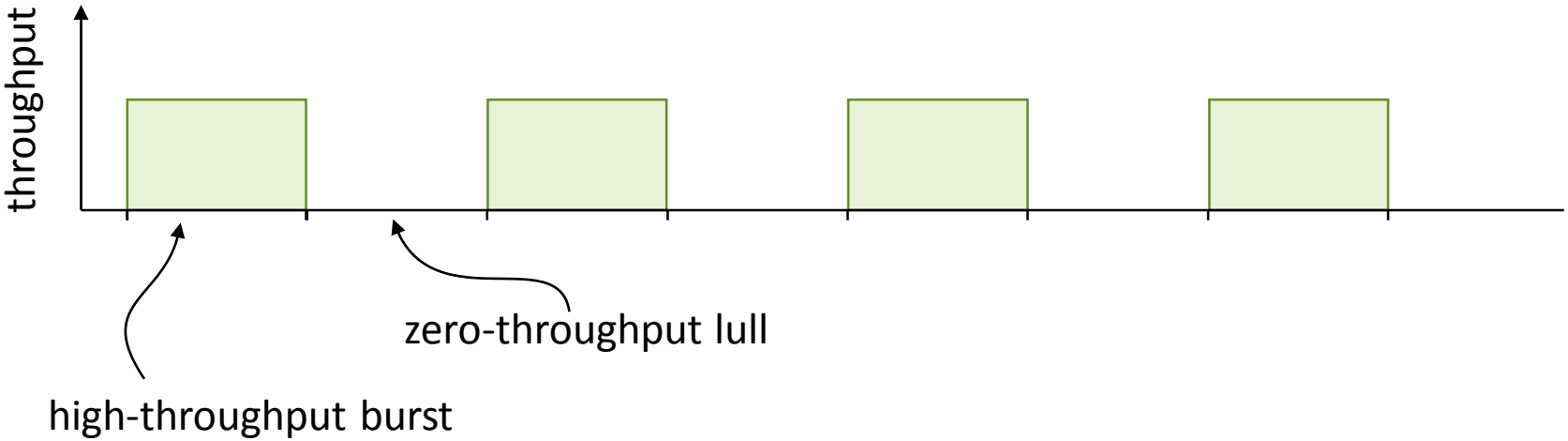

...
time


throughput
zero-throughput lull
high-throughput burst

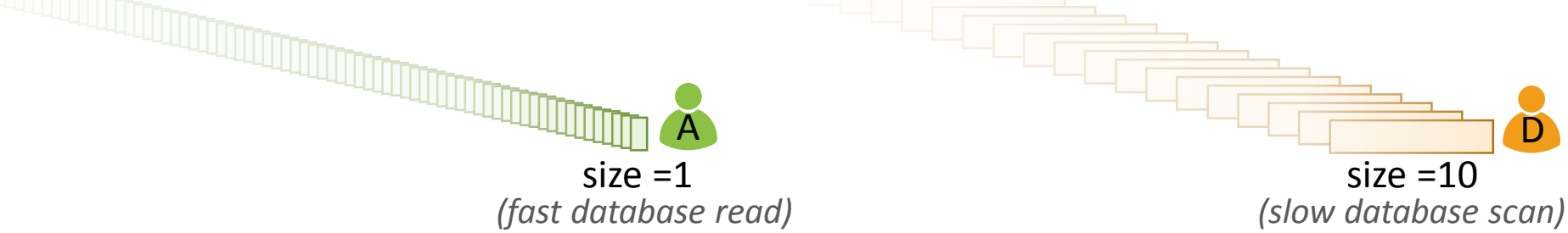


One thread:

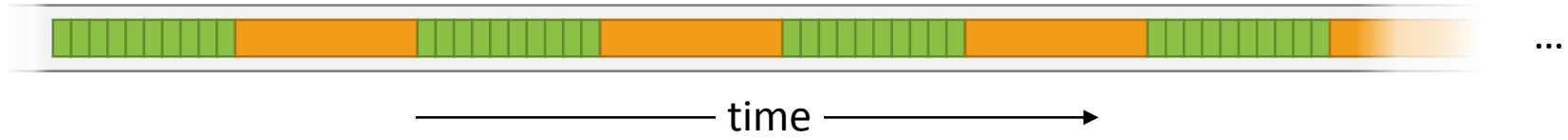


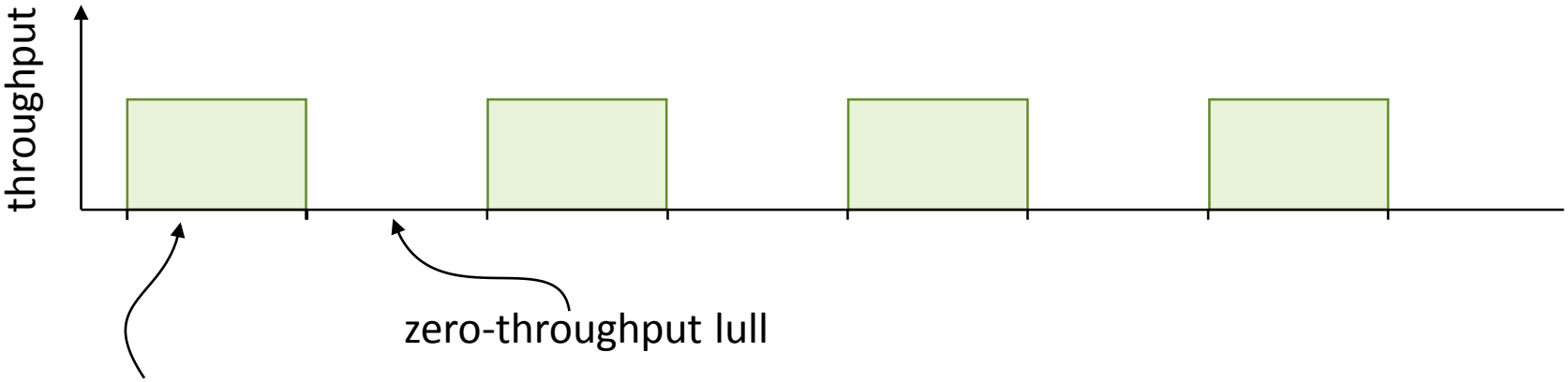


Ideal:

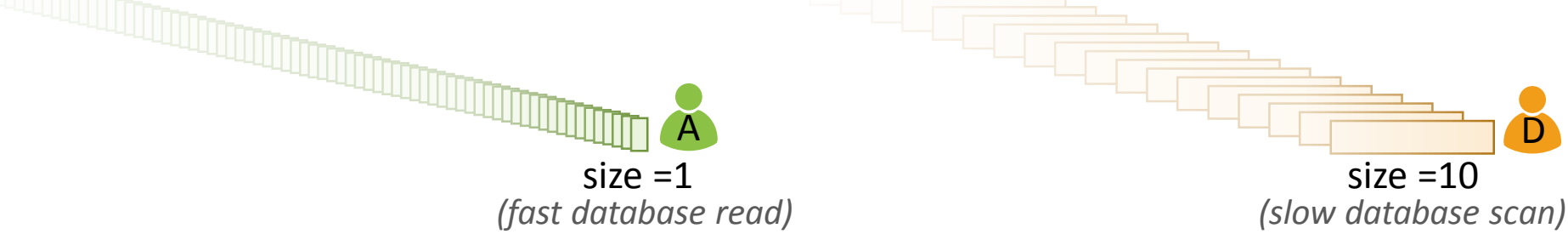


One thread:

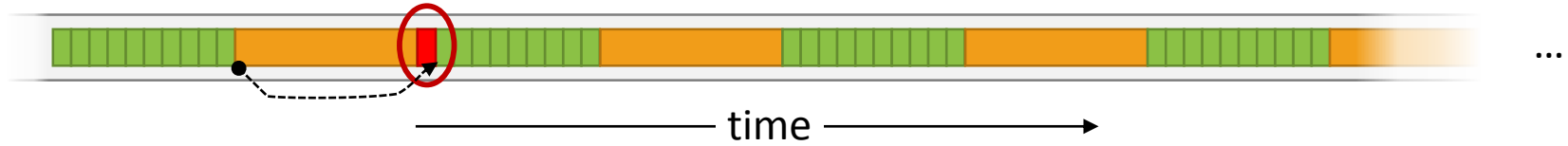


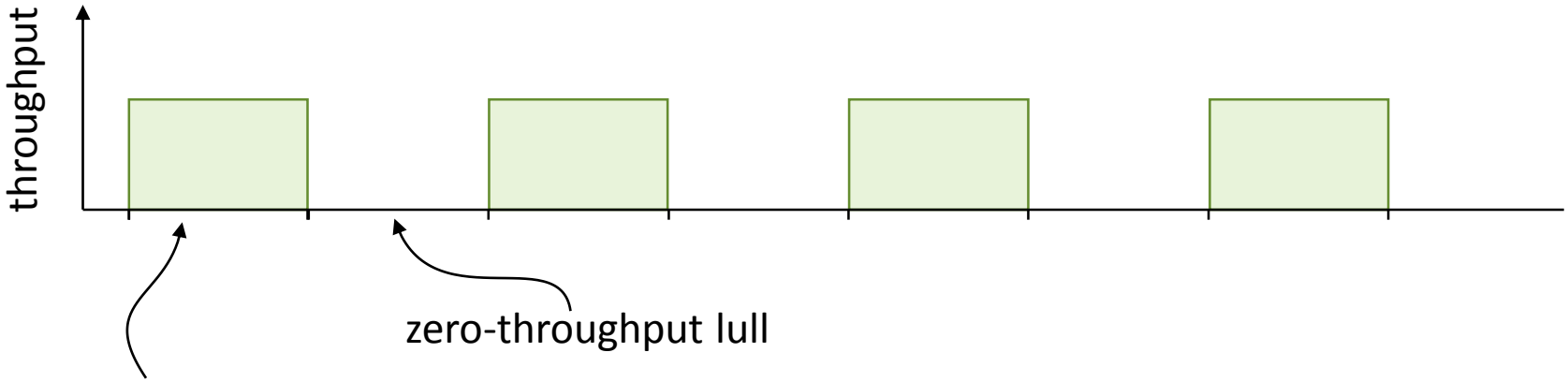


Ideal:

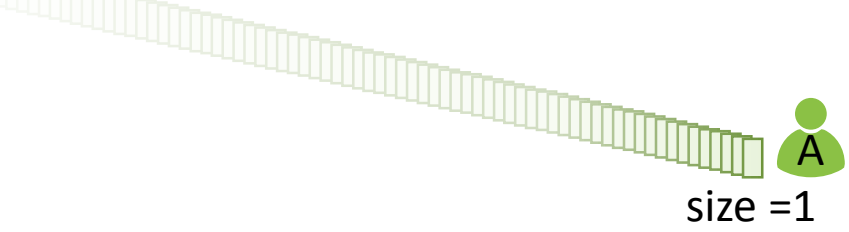


One thread:

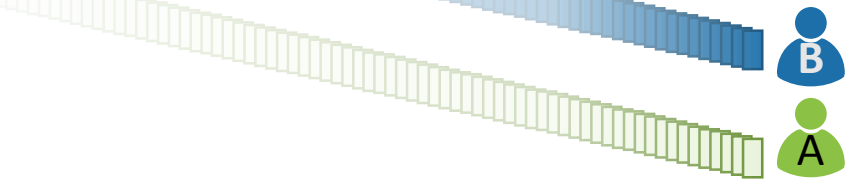




Ideal:


A
size =1

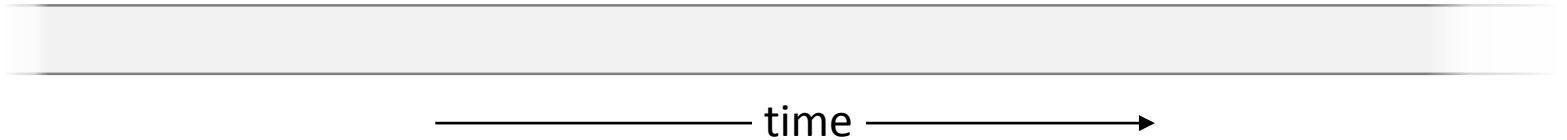

D
size =10



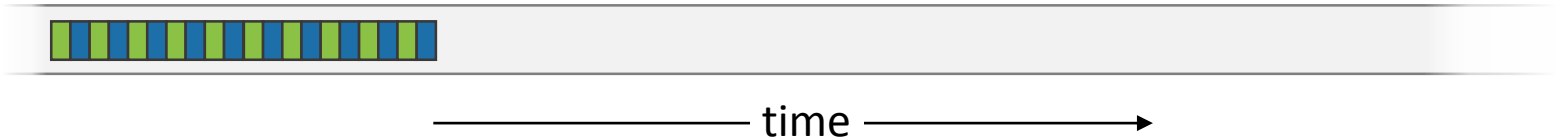
size =1

size =10

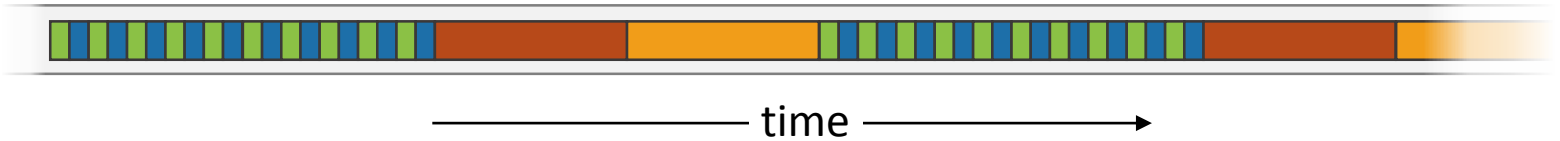

B
A
C
D
size =1
size =10


time


B
A
C
D
size =1
size =10


time


B
A
C
D
size =1
size =10

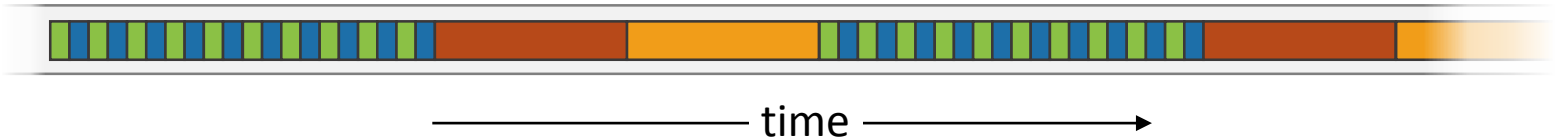

time

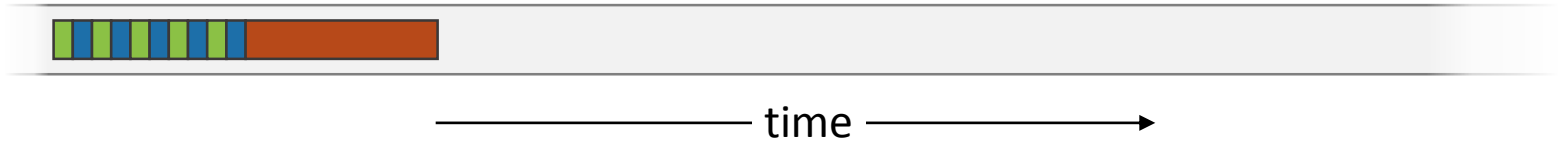

B
A
C
D
size =1
size =10


time


time

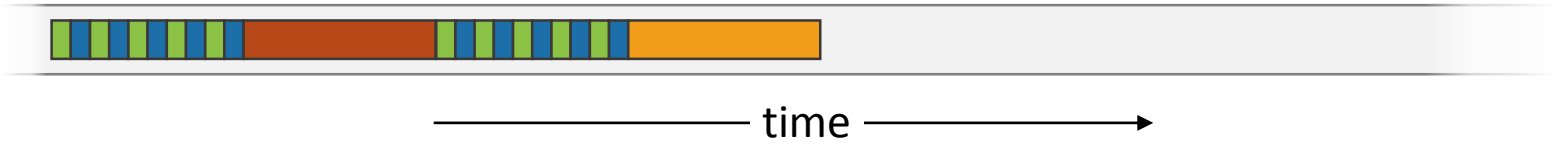

B
A
C
D
size =1
size =10


time


time


B
A
C
D
size =1
size =10


time

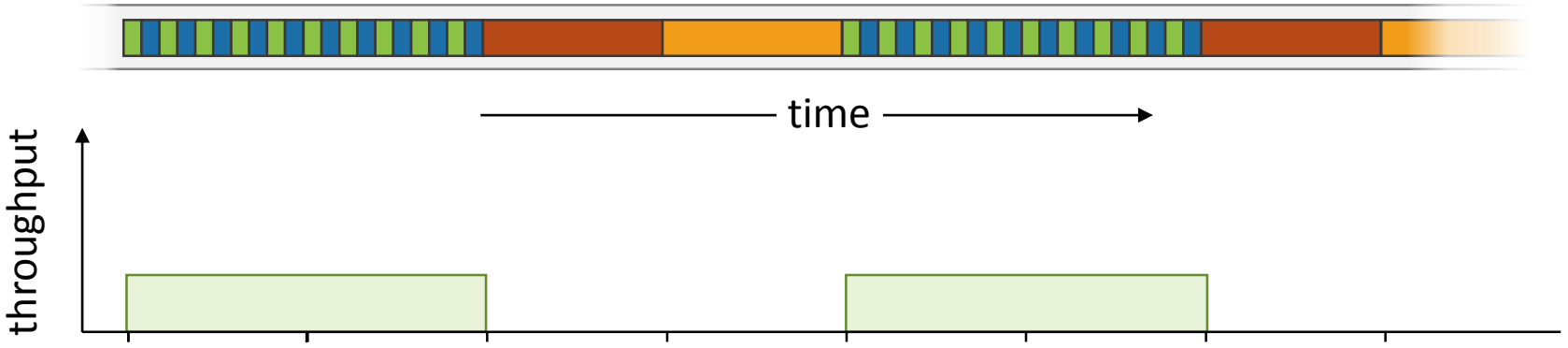

time

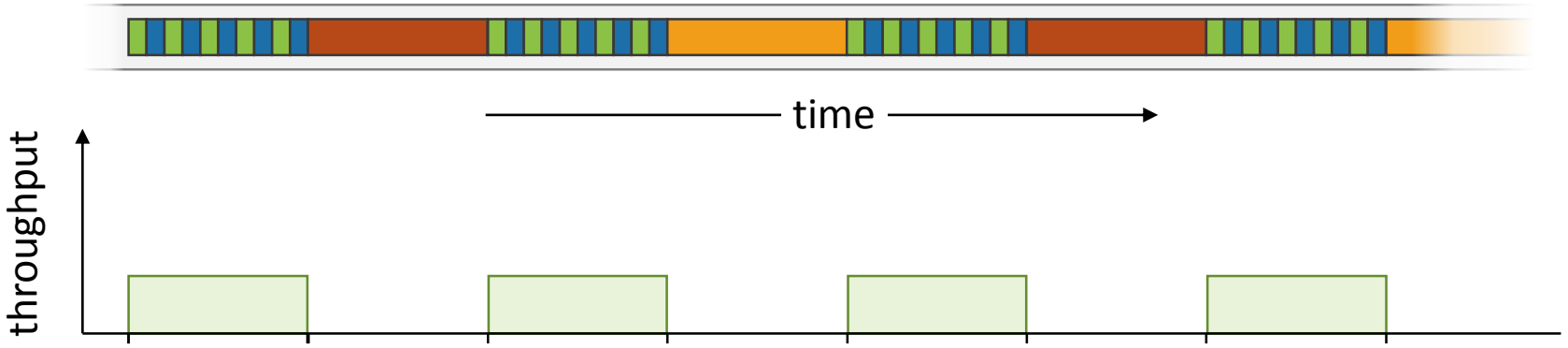

B
A
C
D
size =1
size =10


time

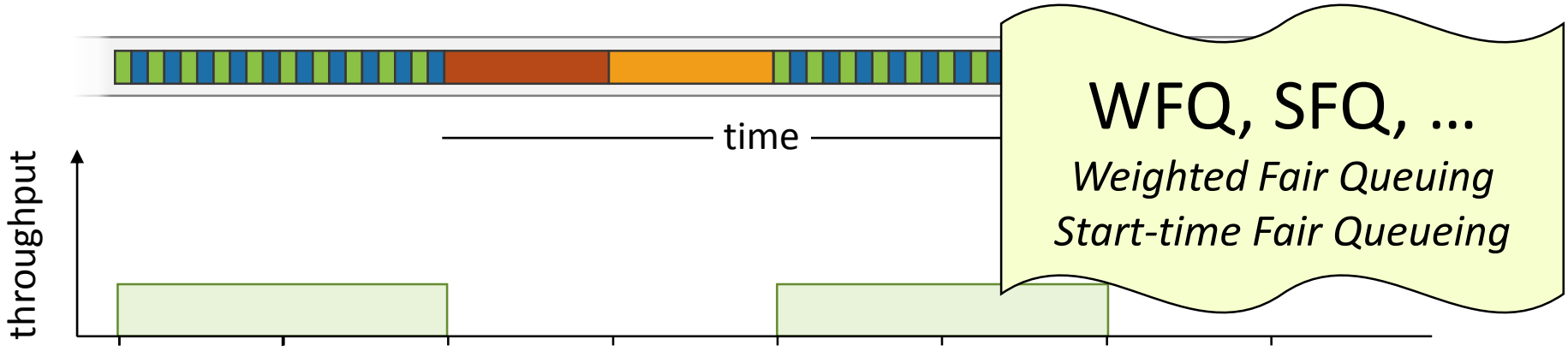

time

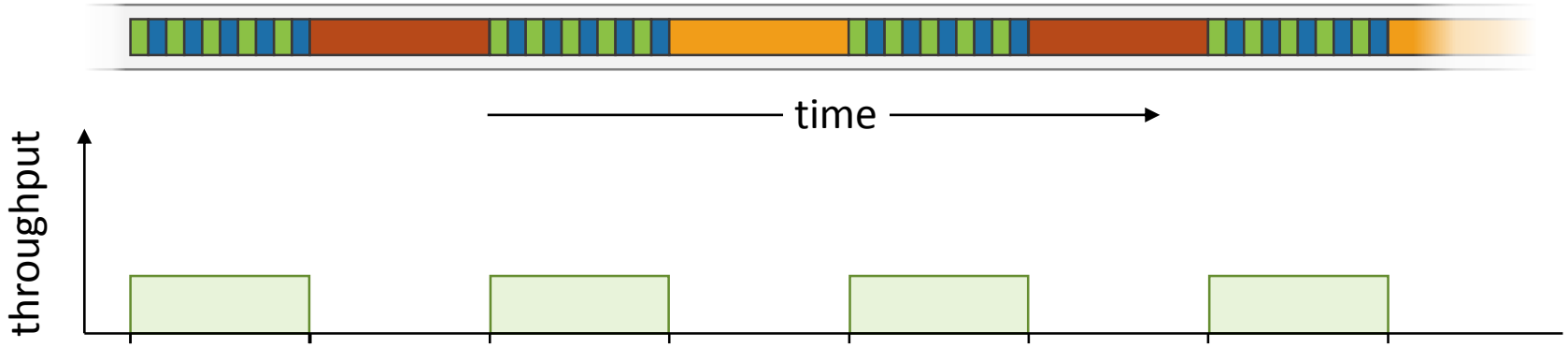

B
A
C
D
size =1
size =10

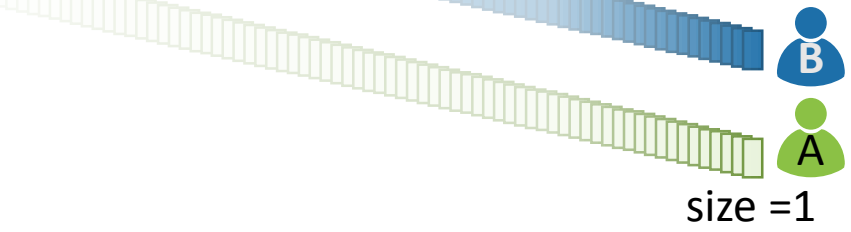

time
throughput


time
throughput

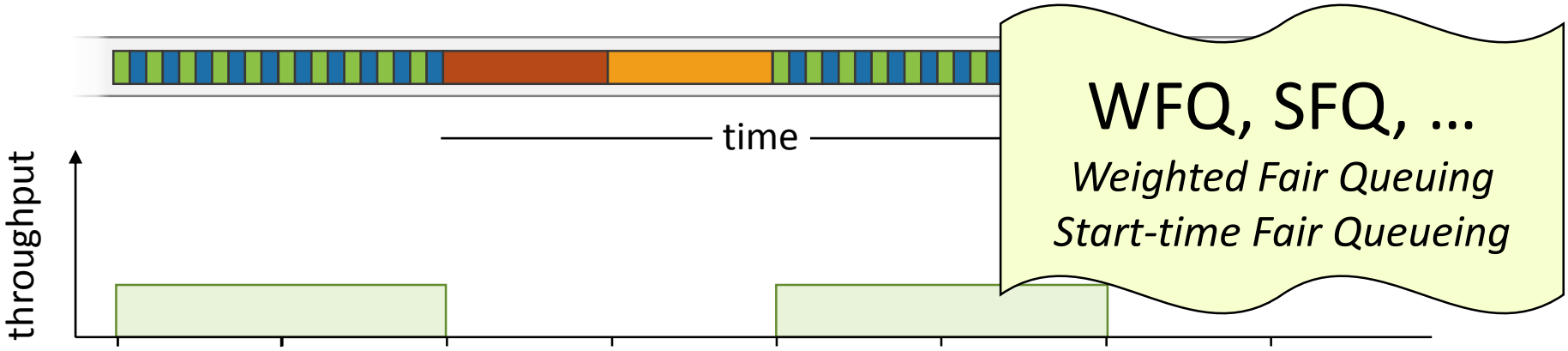

B
A
size =1
C
D
size =10

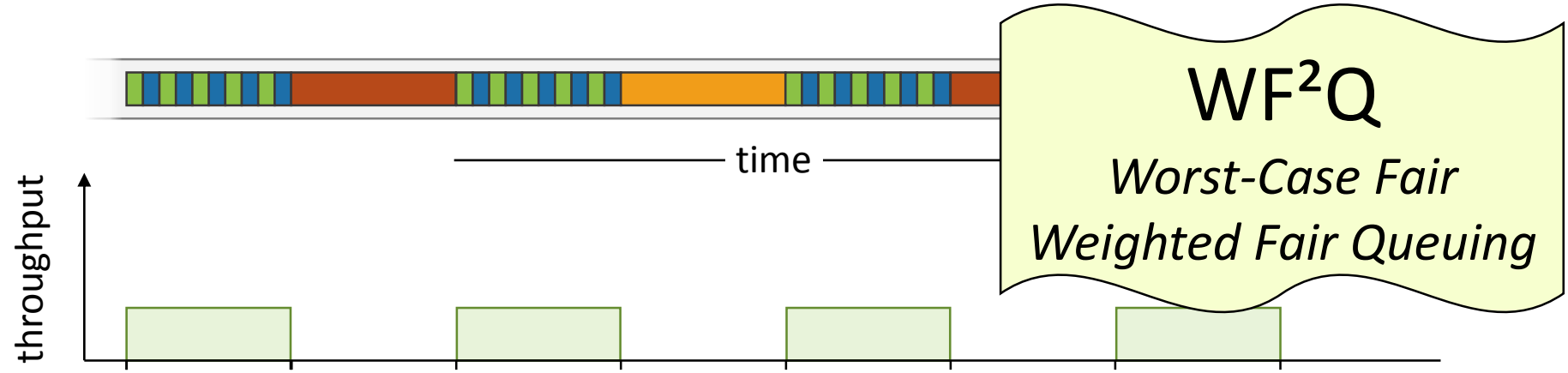

time
throughput
WFQ, SFQ, …
Weighted Fair Queuing
Start-time Fair Queueing

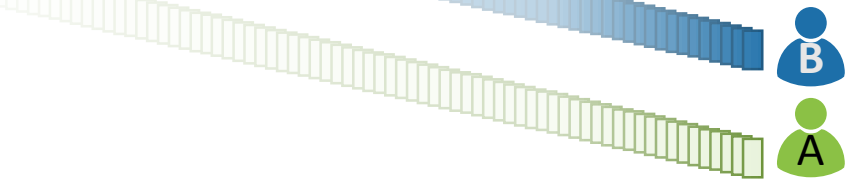

time
throughput


B
A
size =1


C
D
size =10


time
throughput
WFQ, SFQ, …
Weighted Fair Queuing
Start-time Fair Queueing


time
throughput
WF²Q
Worst-Case Fair
Weighted Fair Queuing



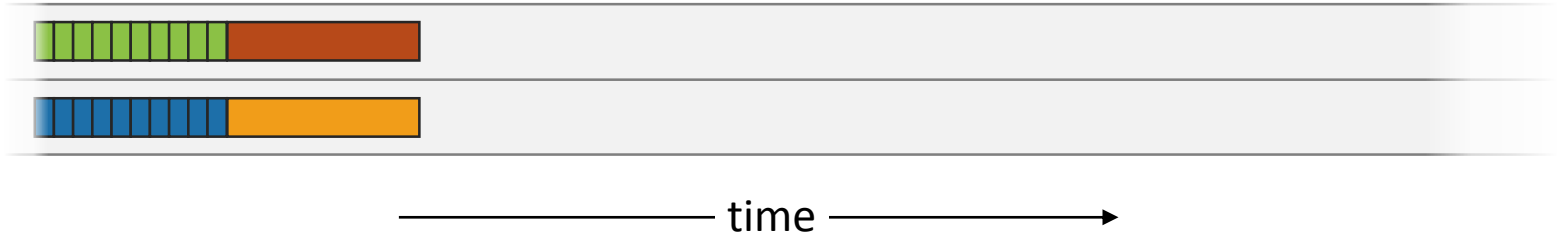
size =1

size =10

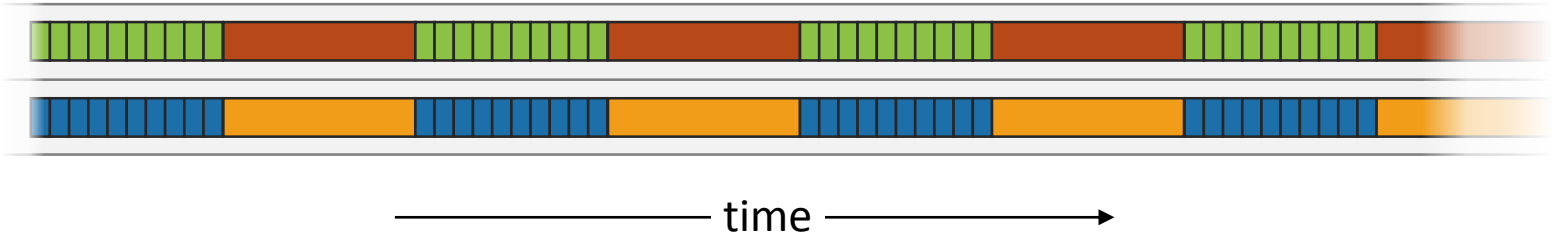
time

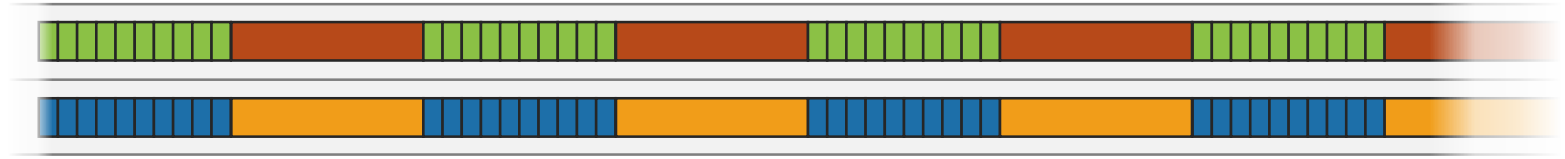
time


B
A
C
D
size =1
size =10


time

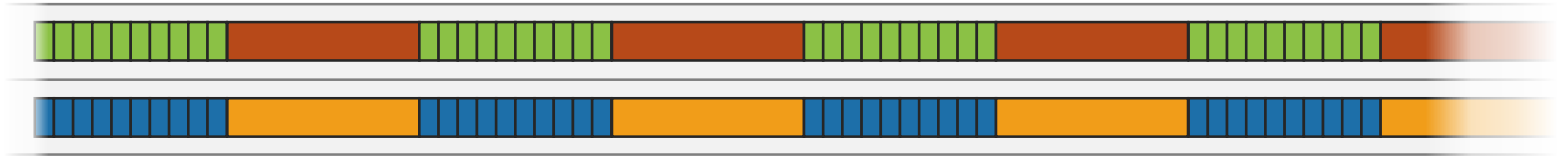

B
A
C
D
size =1
size =10


time



time

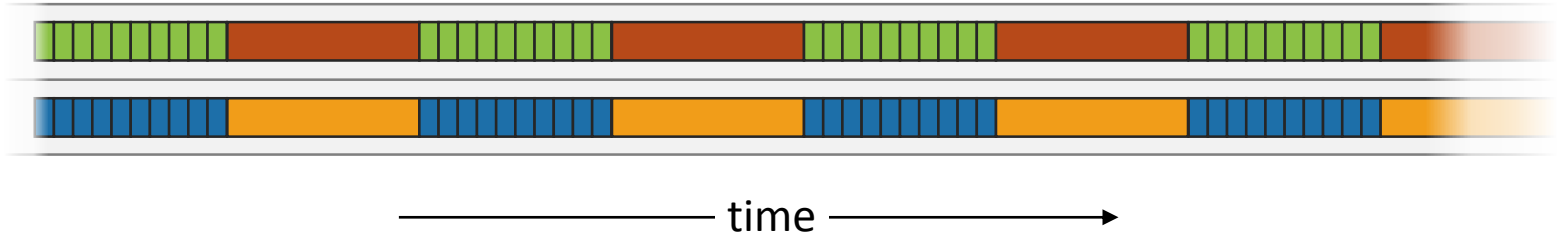
time

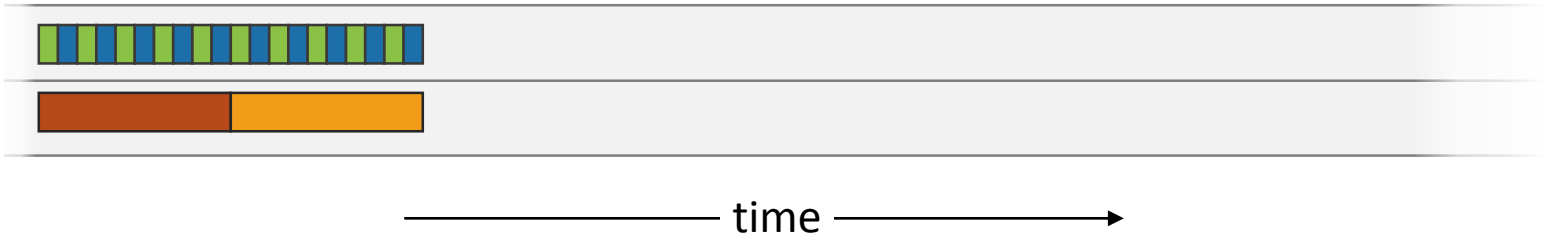


time

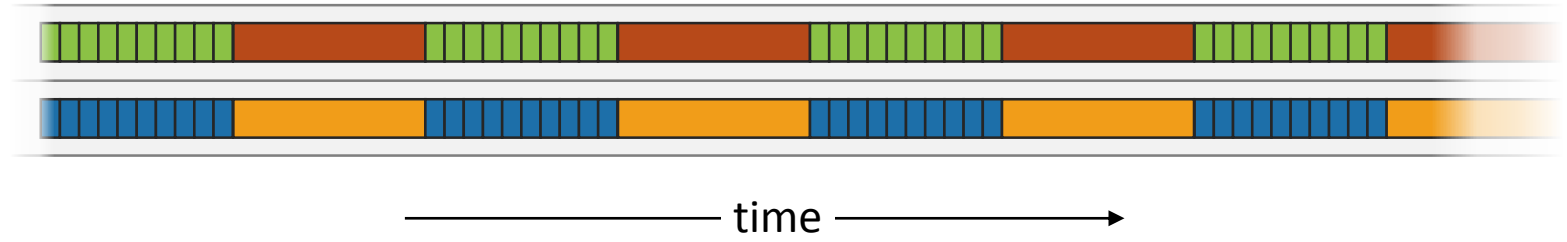
time

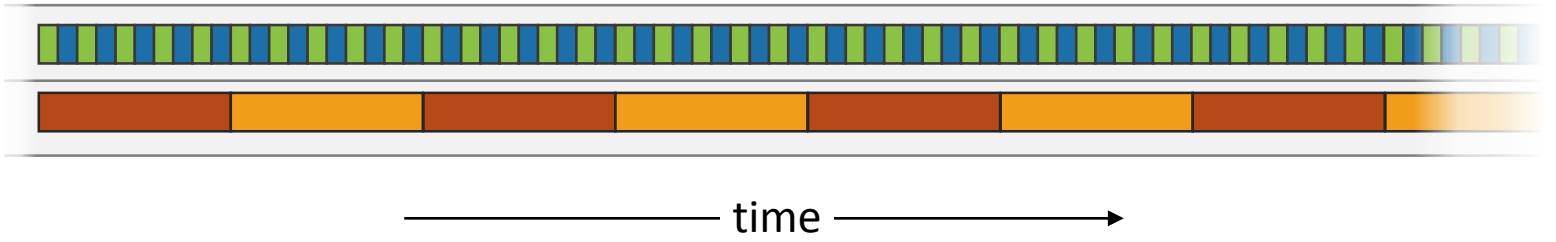

B
A
C
D
size =1
size =10

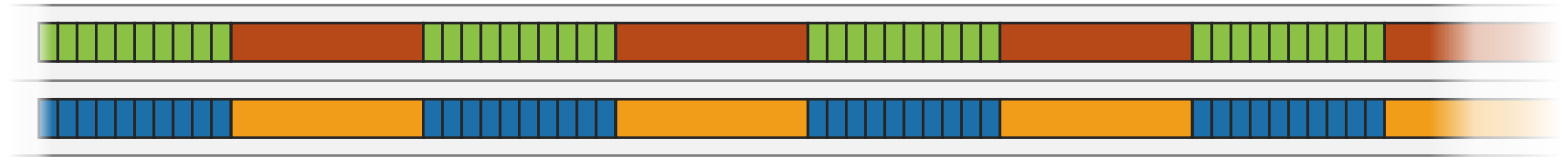

time


time


B
A
C
D
size =1
size =10

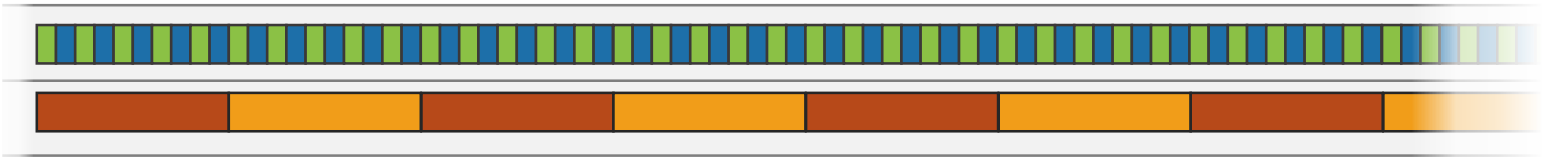

time


time

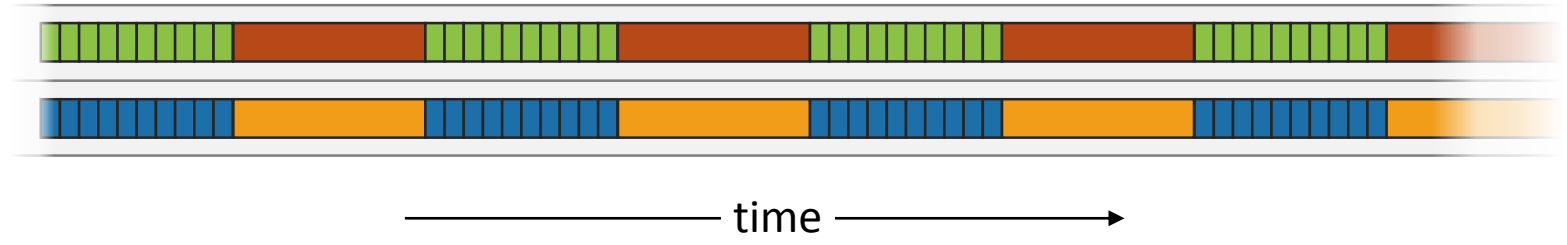


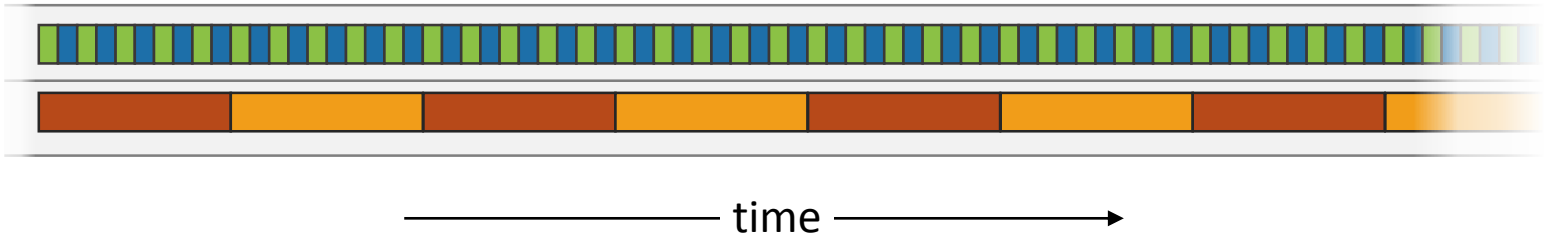
time

Ideal:

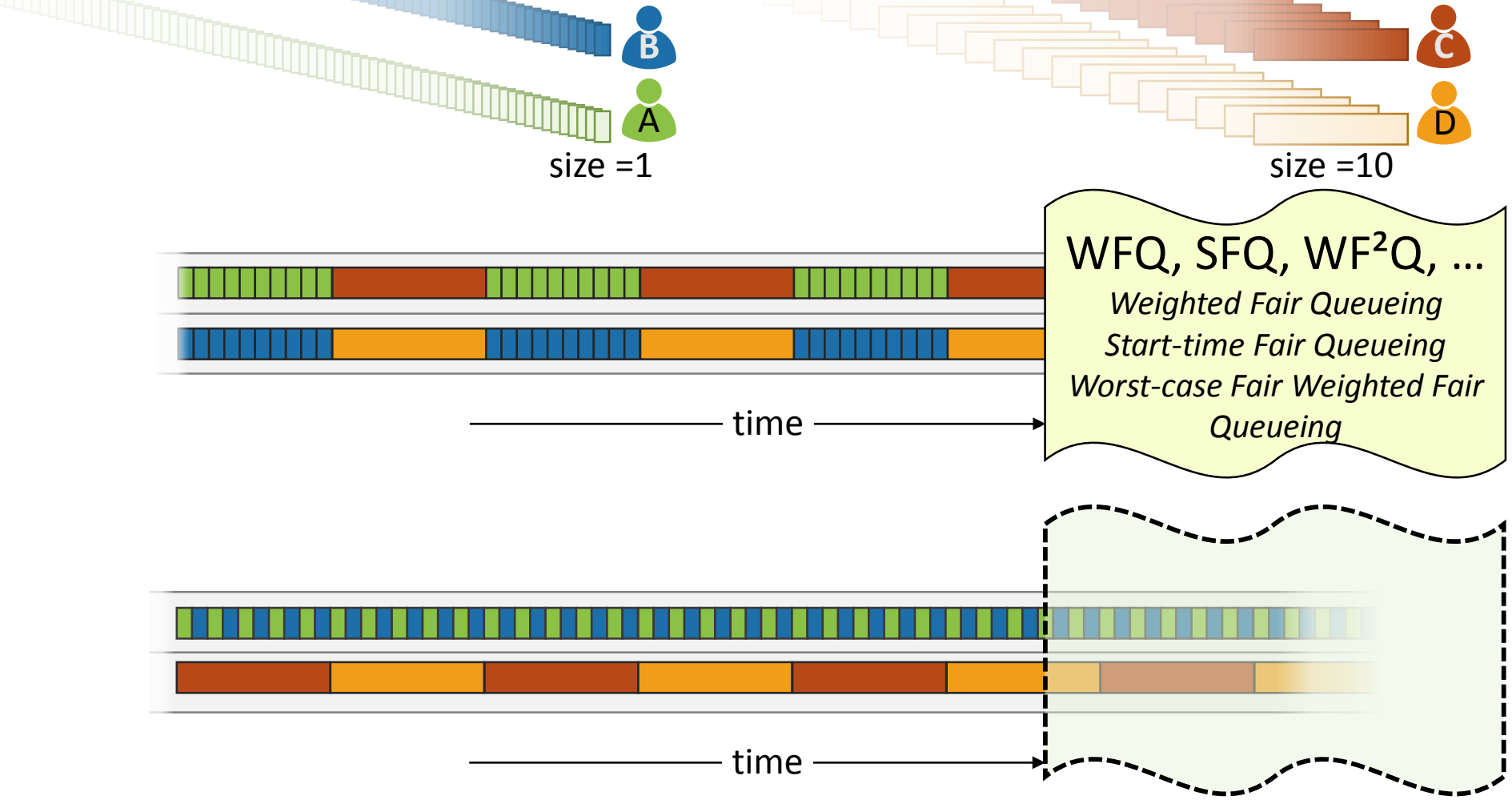


time

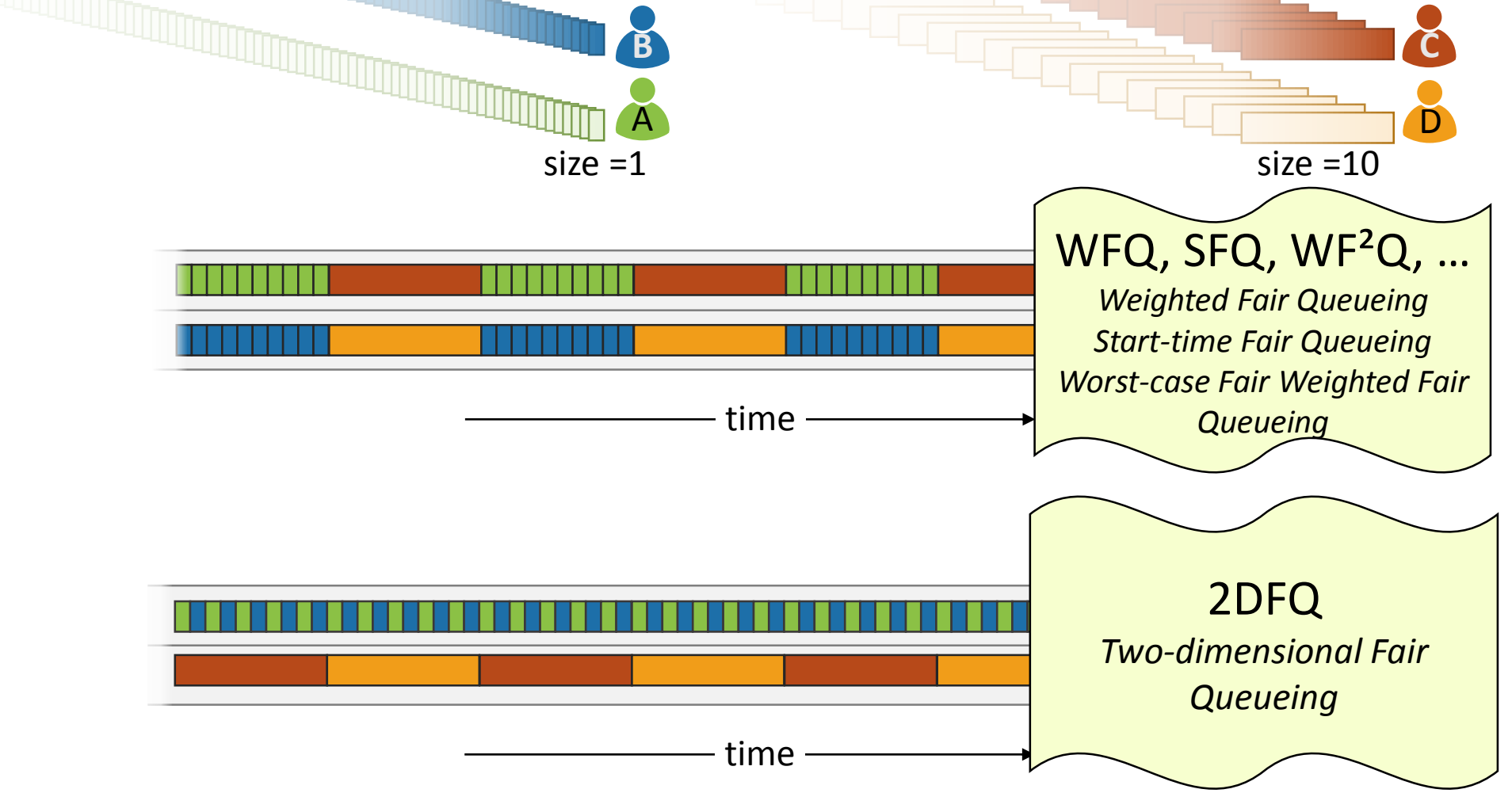




More threads → Opportunity to reduce burstiness



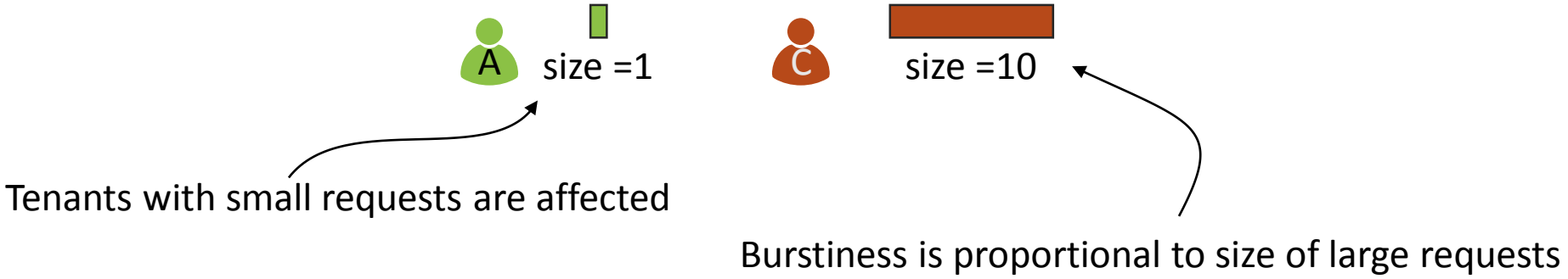
More threads → Opportunity to reduce burstiness

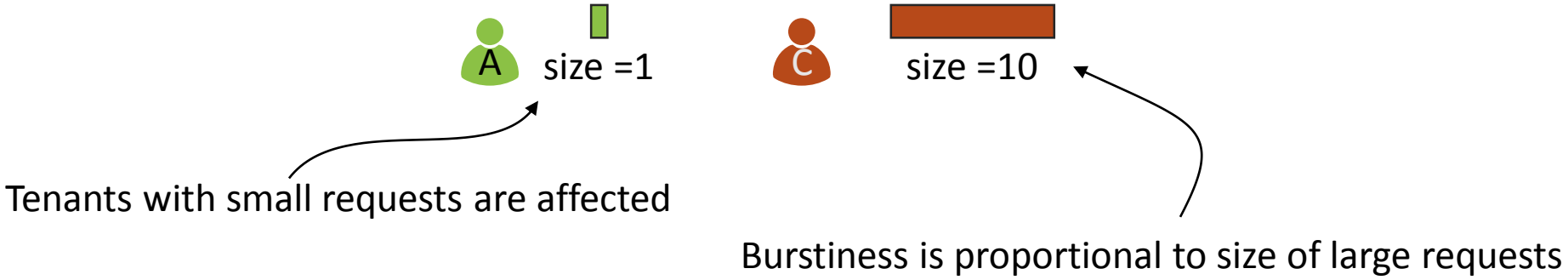


More threads → Opportunity to reduce burstiness

Challenges

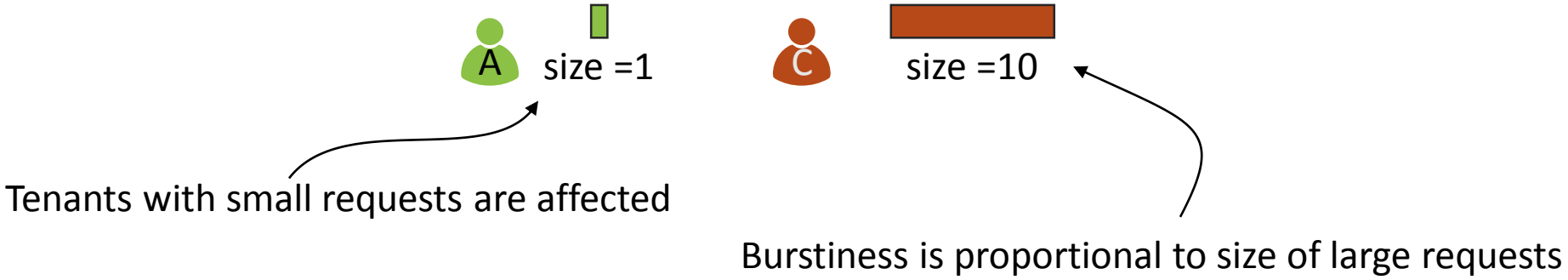
Tenants with small requests are affected


A
size =1
C
size =10
Tenants with small requests are affected
Burstiness is proportional to size of large requests



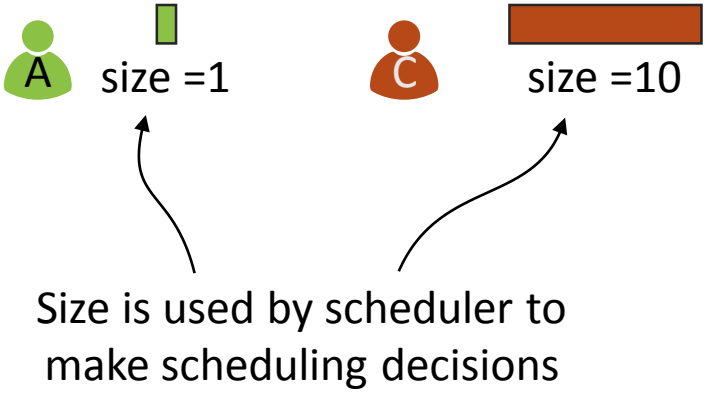
Cloud services:

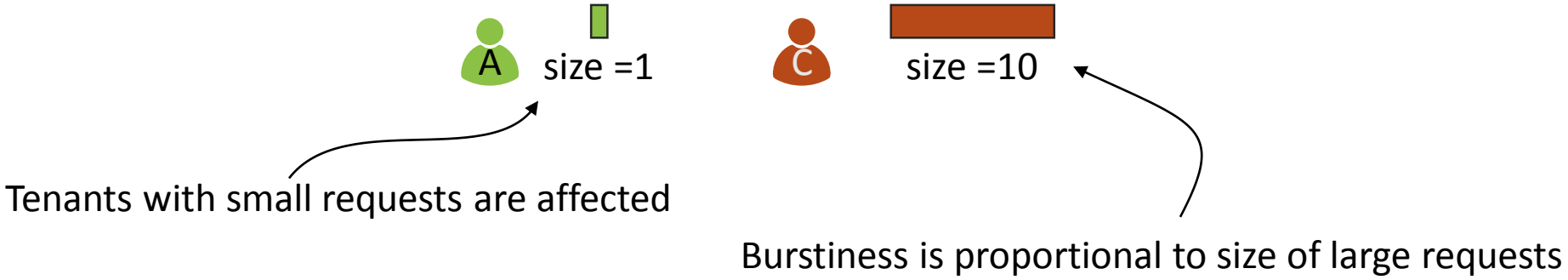
4+ orders of magnitude variation in cost



Cloud services:

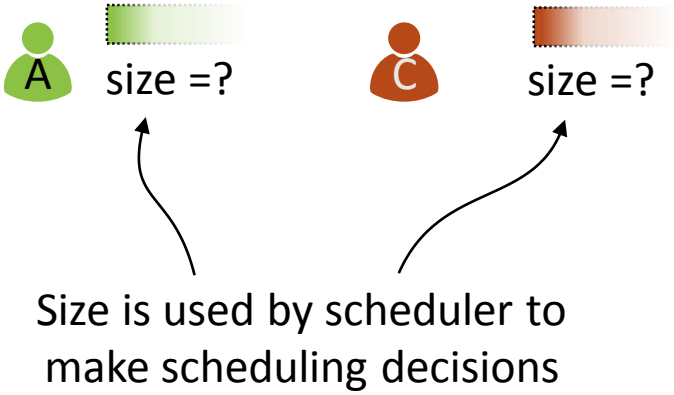
4+ orders of magnitude variation in cost

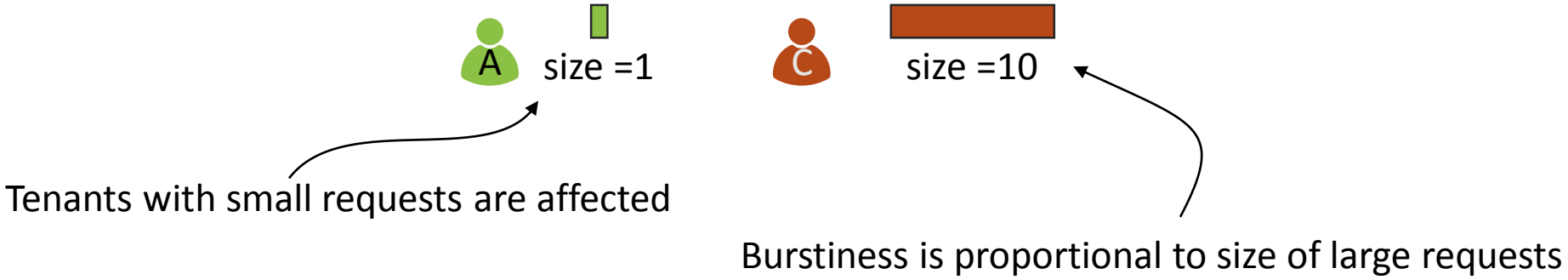




Cloud services:

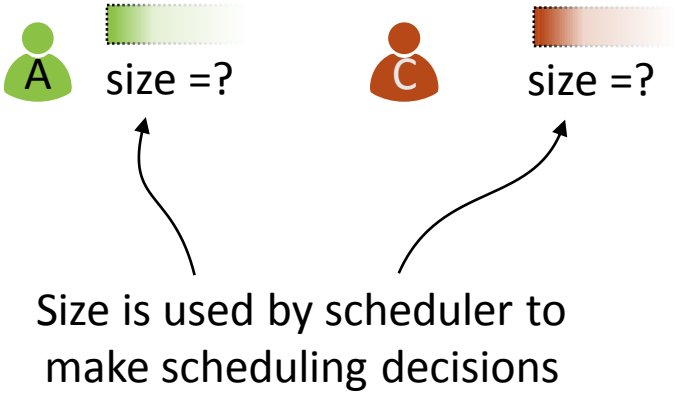
4+ orders of magnitude variation in cost





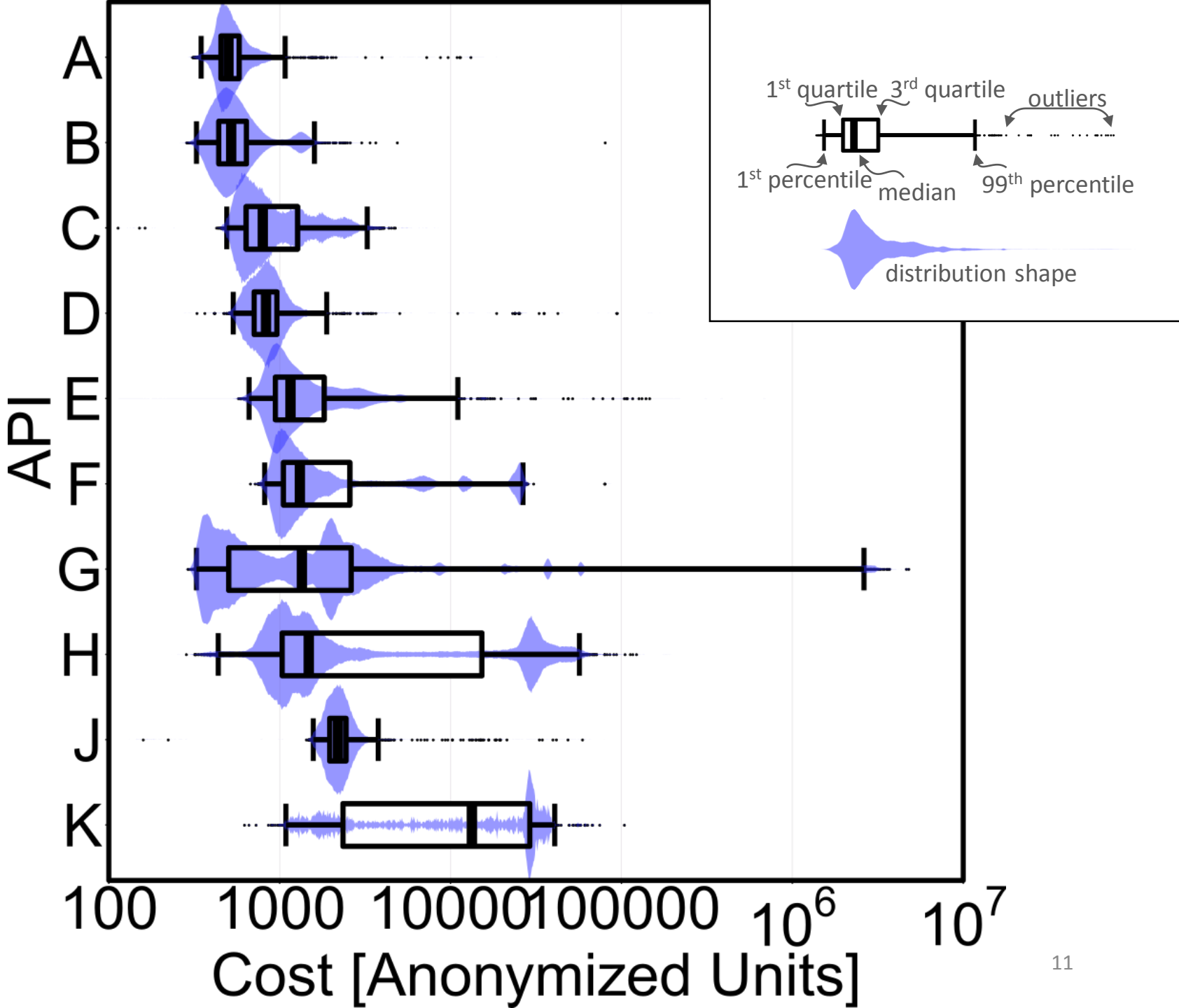
Cloud services:

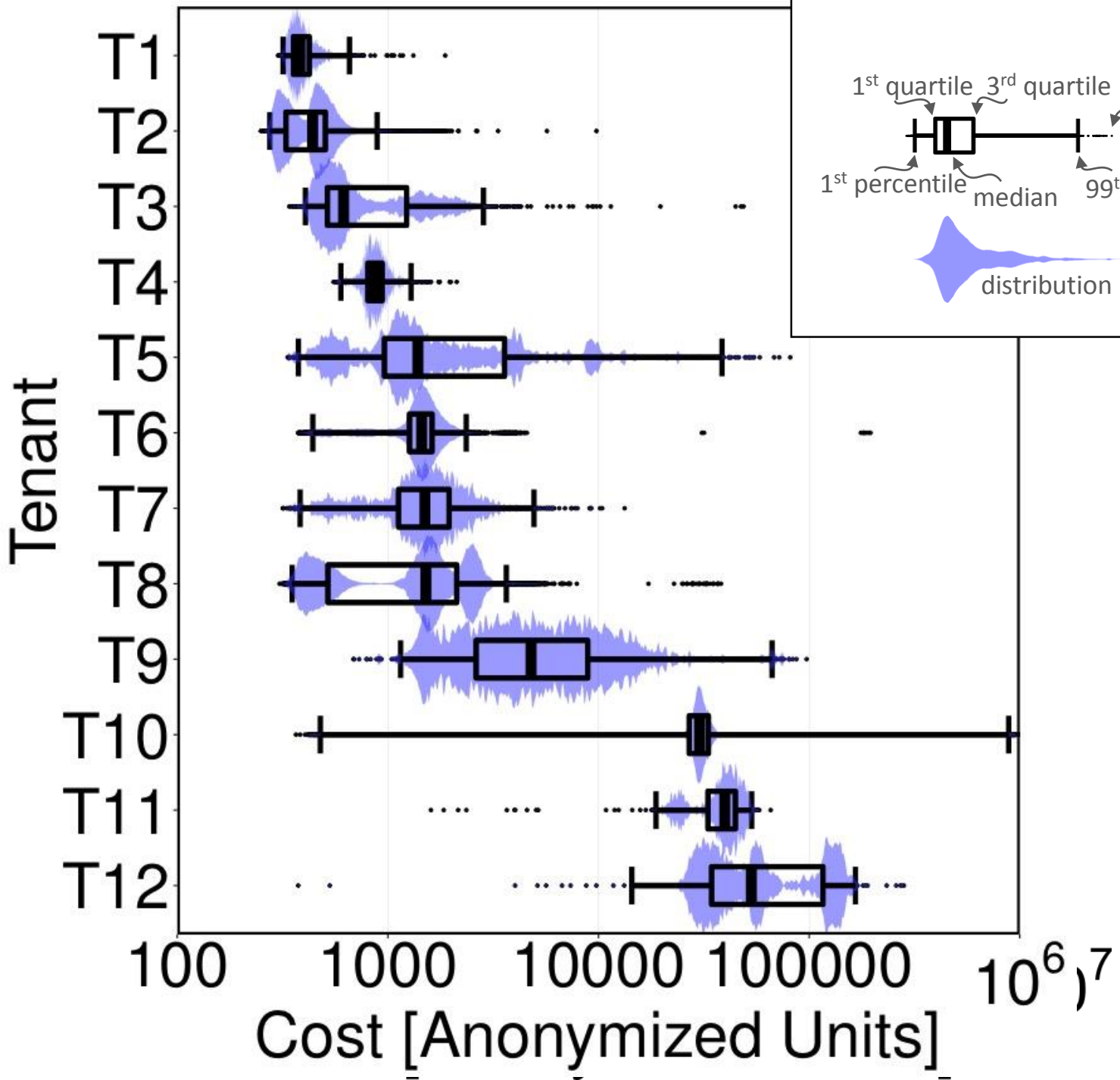
4+ orders of magnitude variation in cost

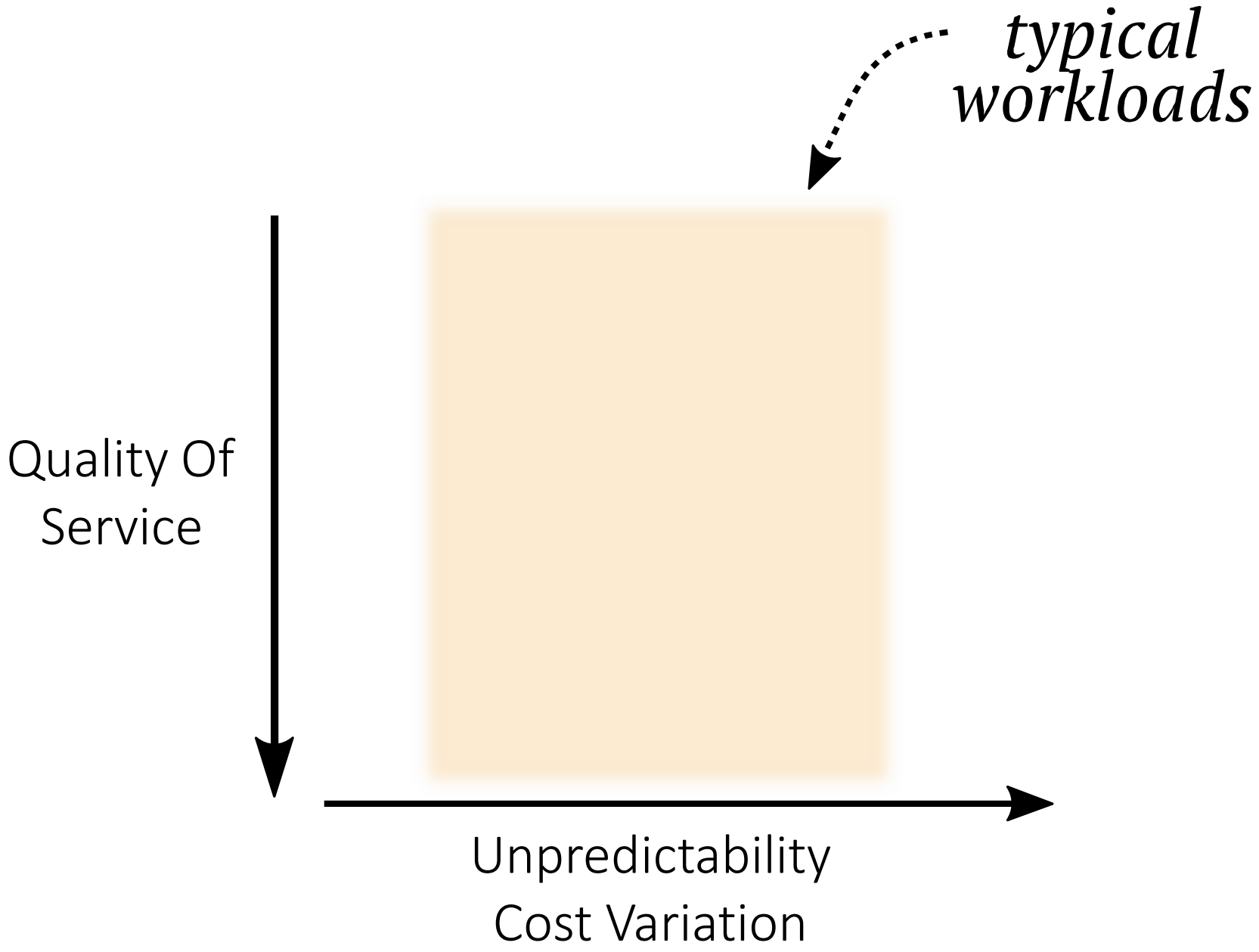


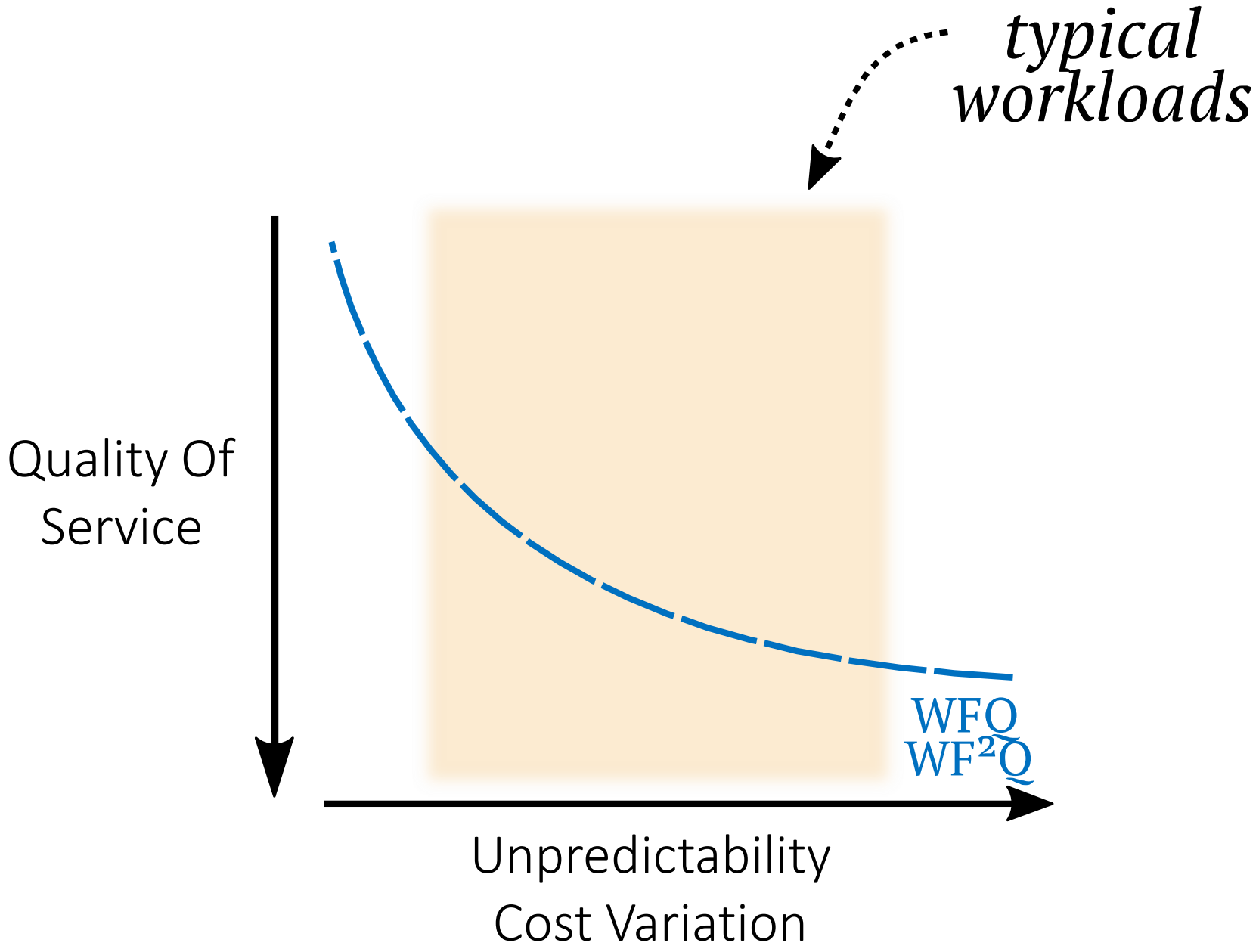
Cloud services:

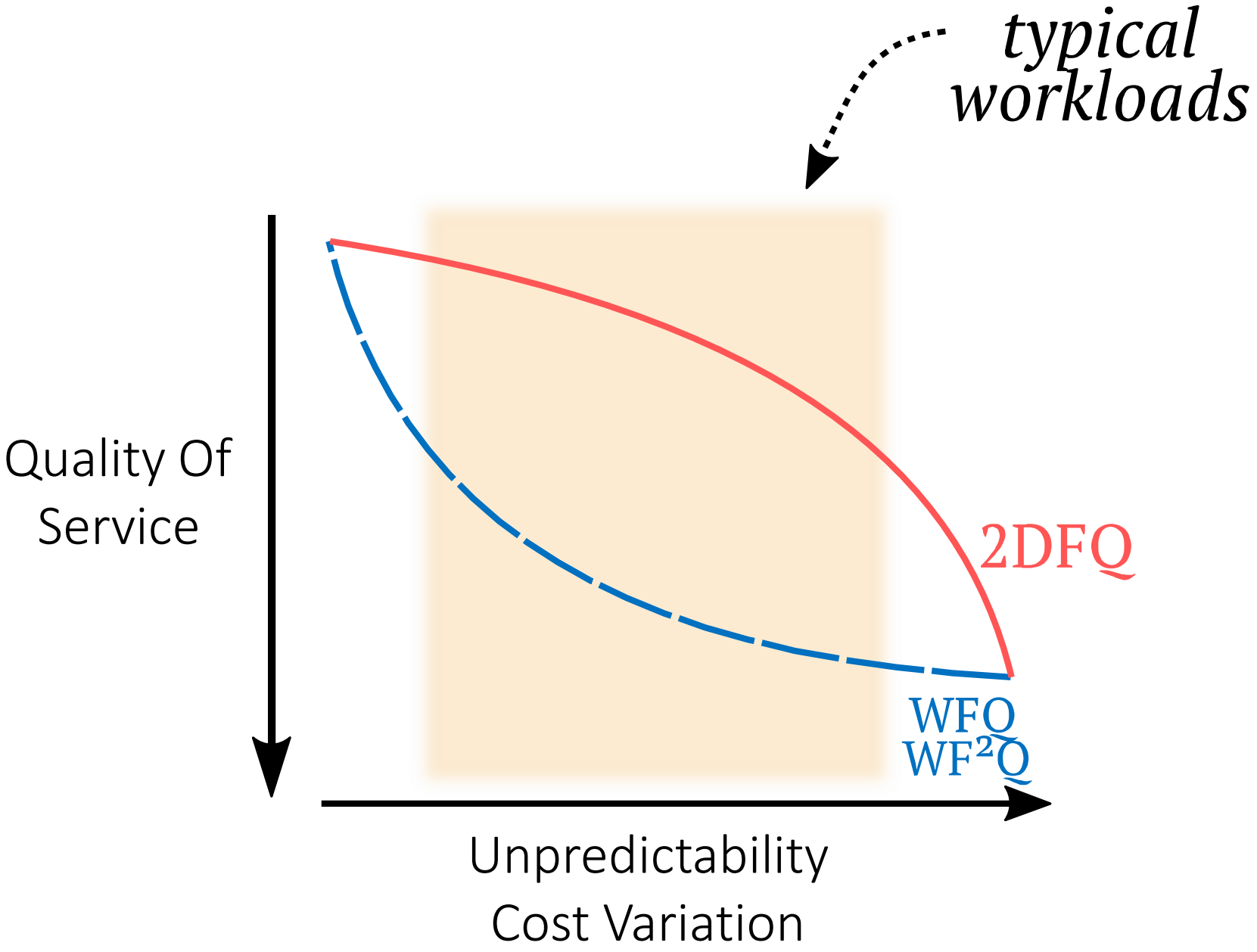
Estimation using model or moving averages


A
B
C
D
E
F
G
H
J
K
API
100
1000
10000
100000
10^6
10^7
Cost [Anonymized Units]
1st quartile
3rd quartile
outliers
1st percentile
median
99th percentile
distribution shape


T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
Tenant
100
1000
10000
100000
10^6
10^7
Cost [Anonymized Units]
1st quartile
3rd quartile
outliers
1st percentile
median
99th percentile
distribution shape

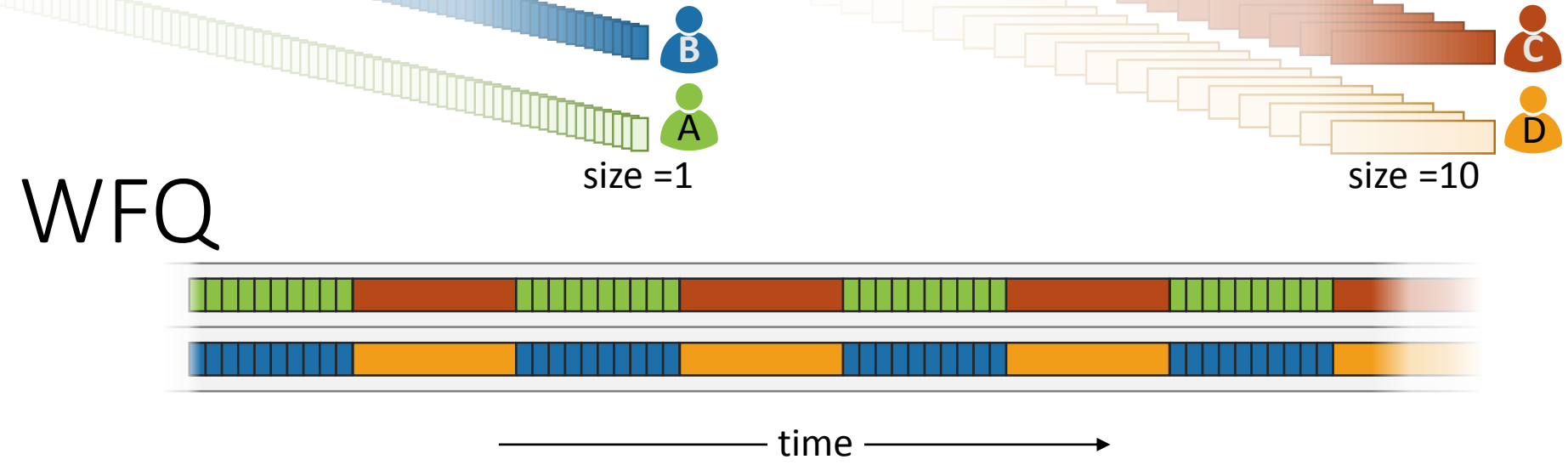

typical
workloads
Quality Of
Service
Unpredictability
Cost Variation


typical workloads
Quality Of Service
WFQ
WF²Q
Unpredictability
Cost Variation

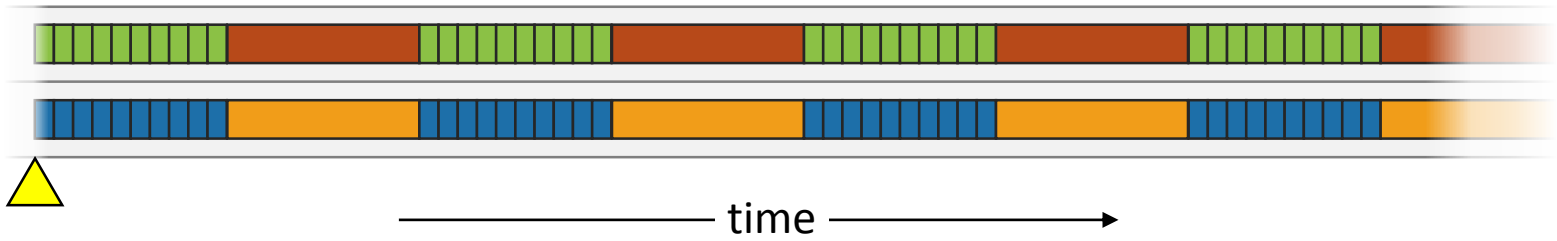

typical workloads
Quality Of Service
2DFQ
WFQ
WF²Q
Unpredictability
Cost Variation

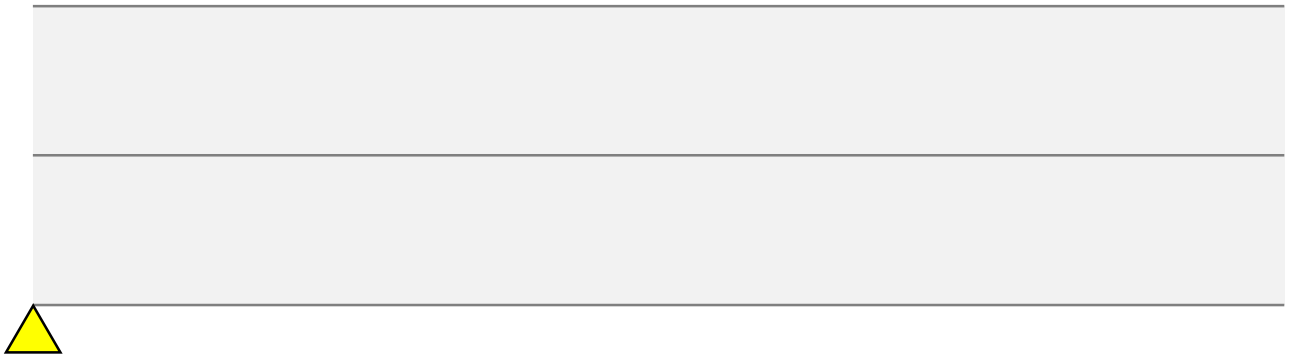
Two-Dimensional Fair Queueing

WFQ

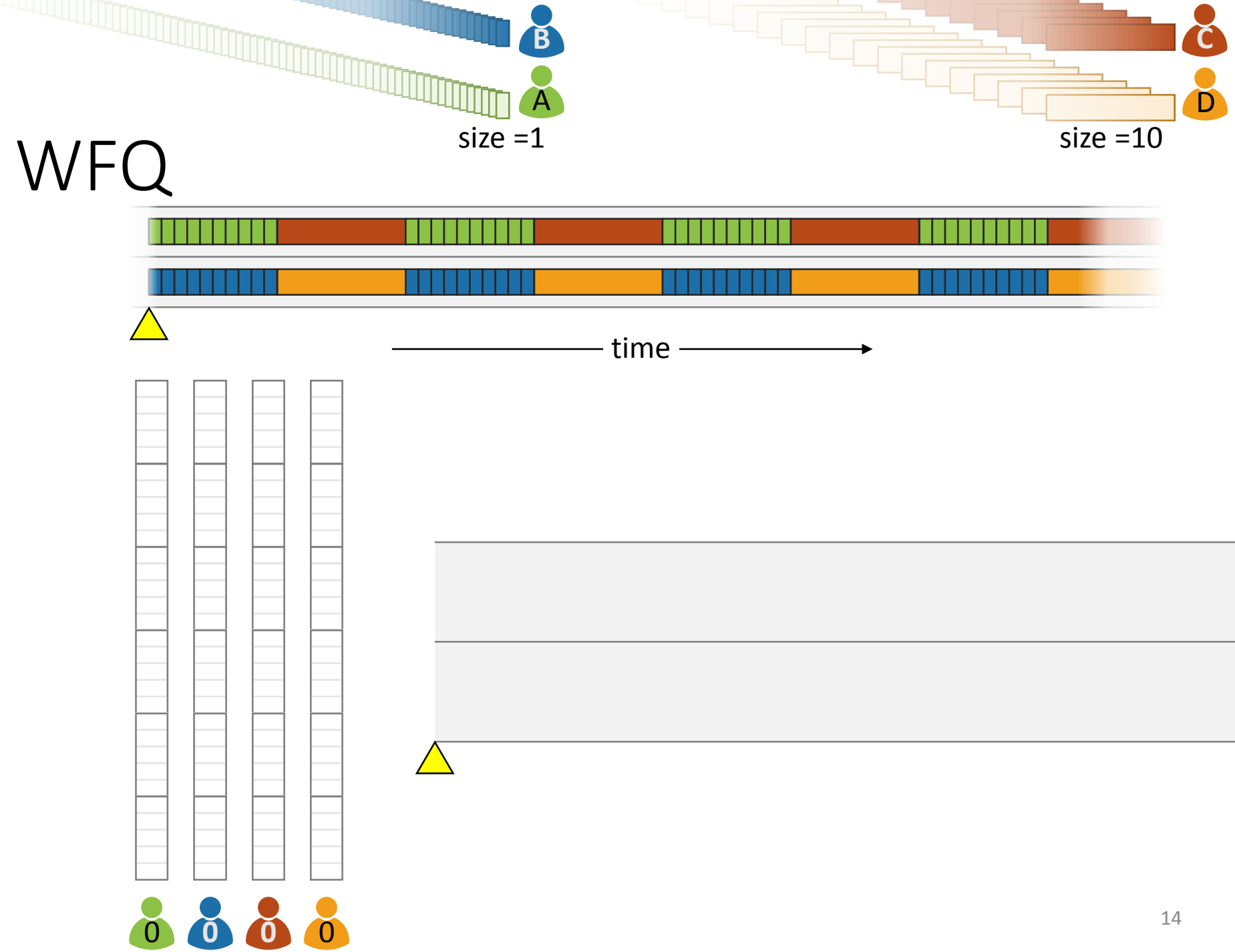


WFQ

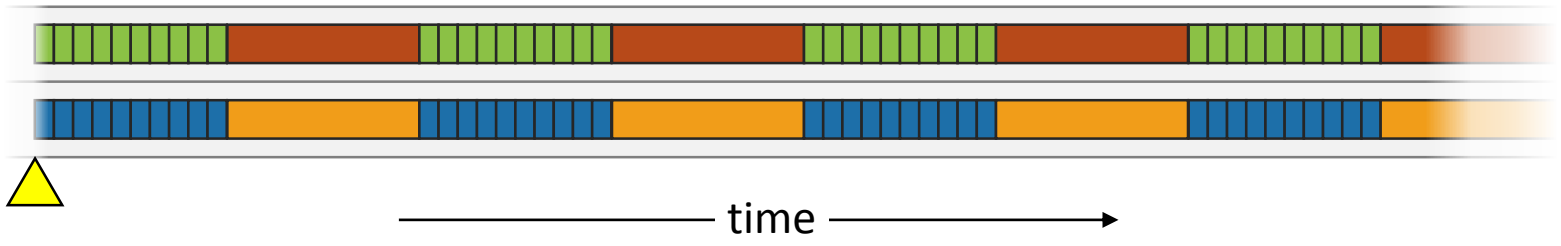


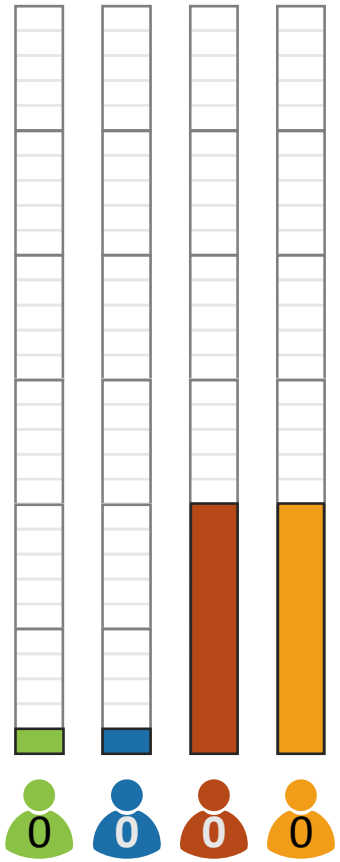


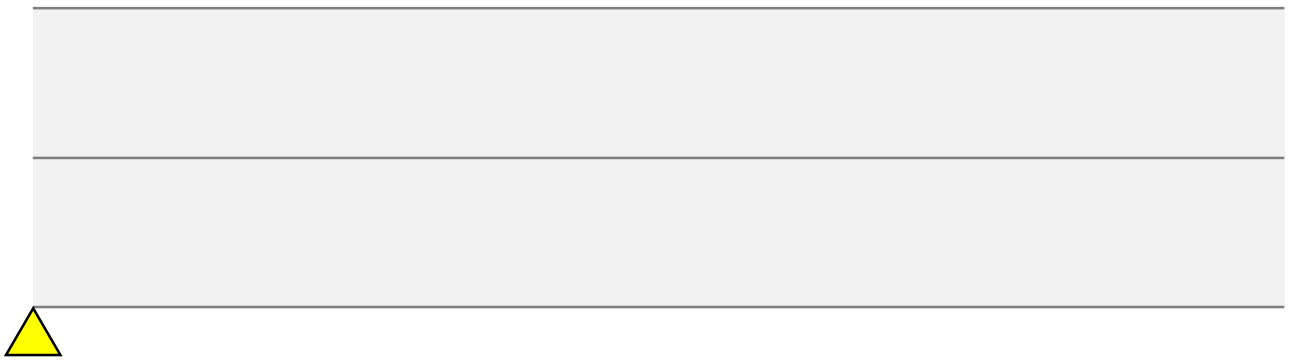
WFQ



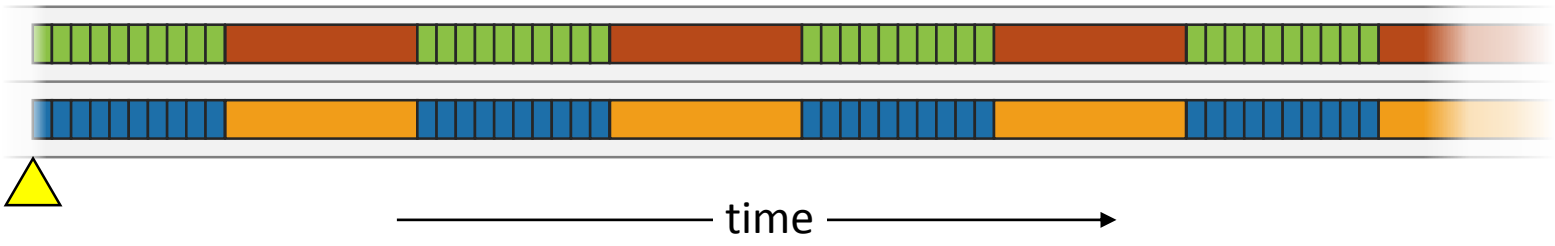
WFQ

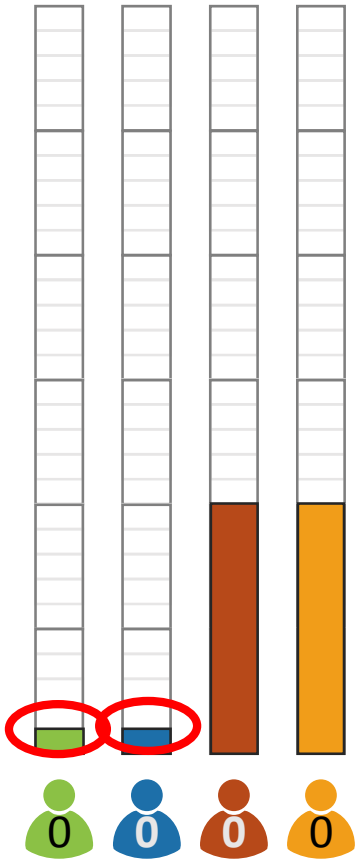


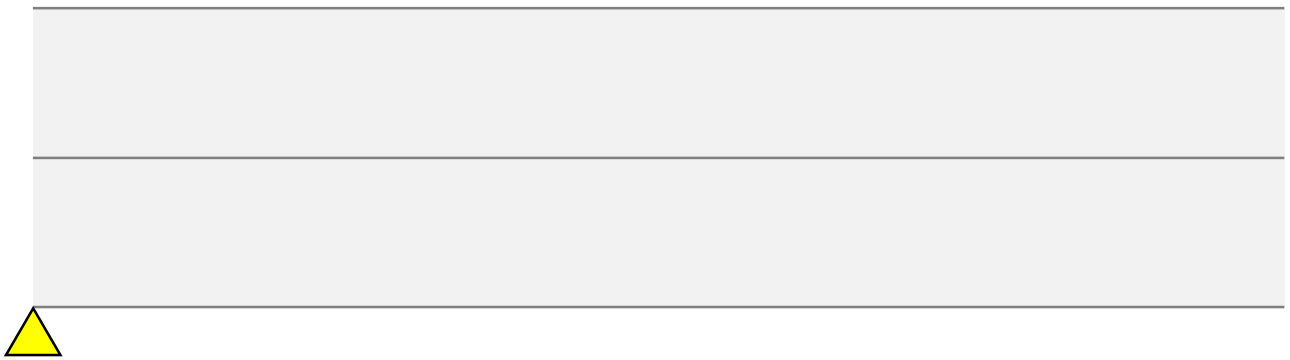




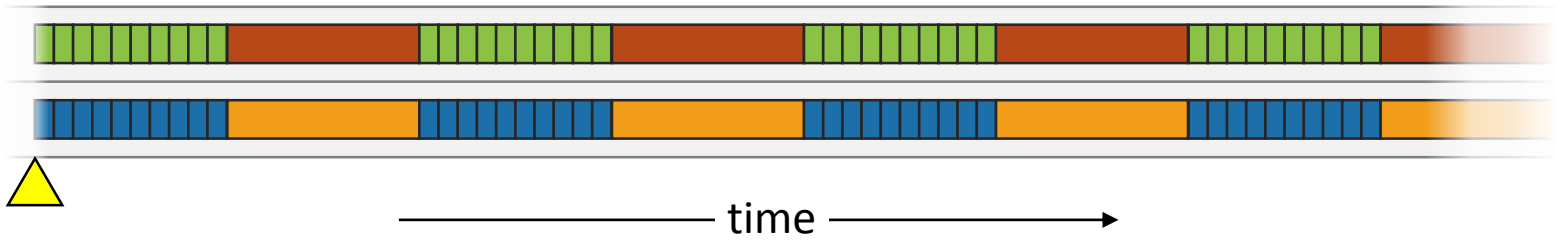
WFQ

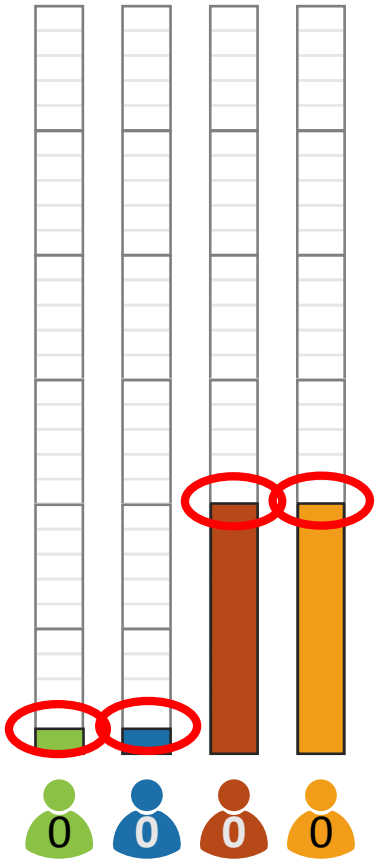


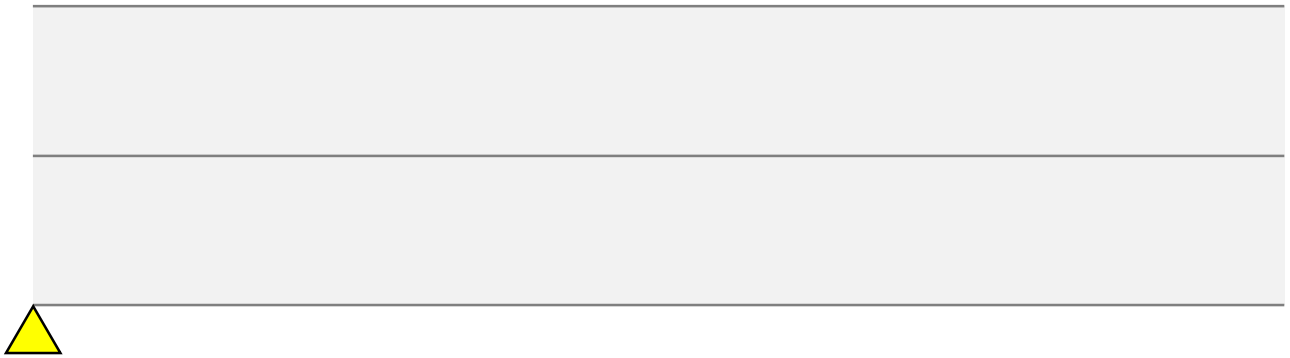




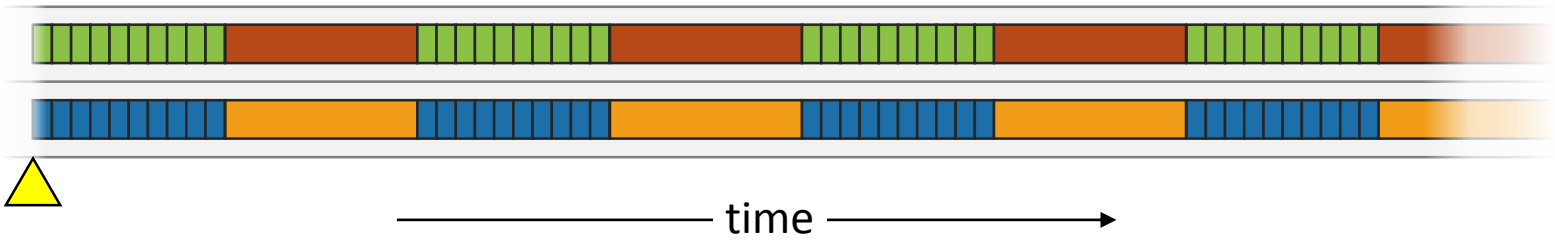
WFQ

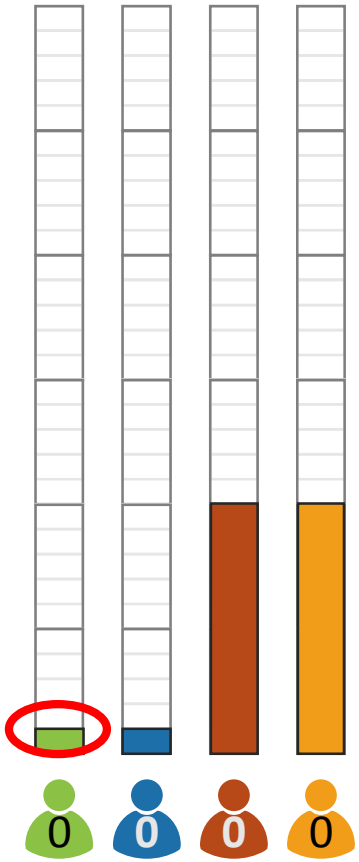


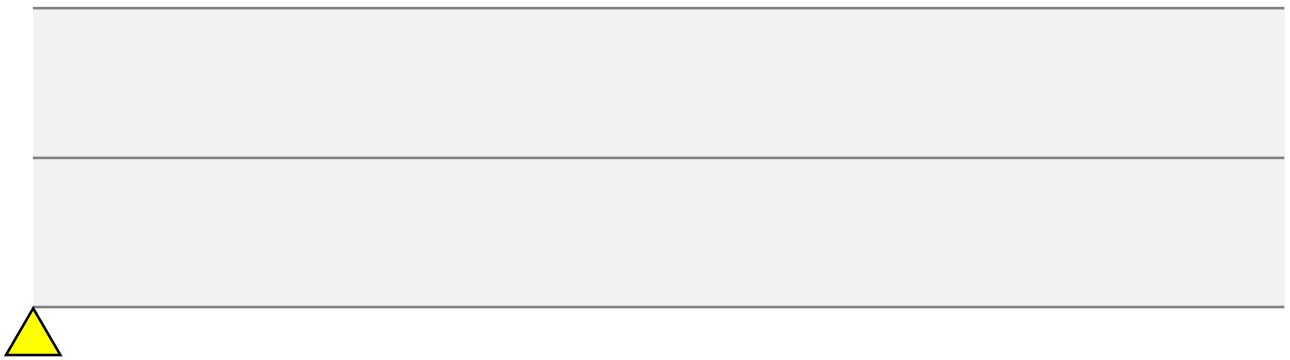




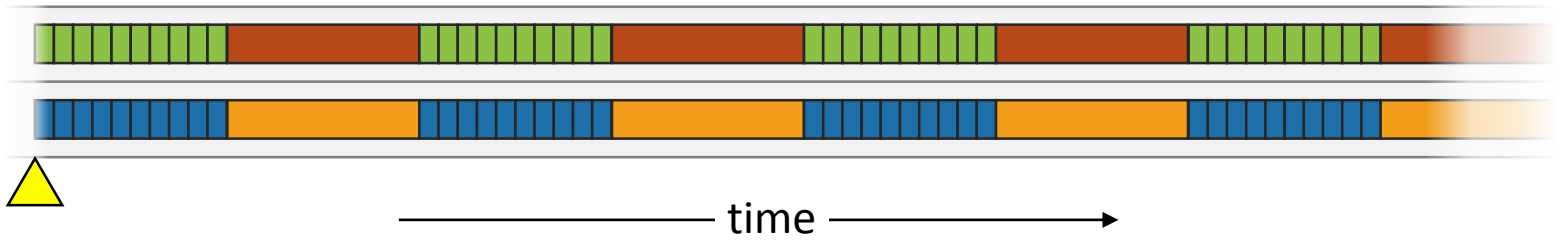
WFQ

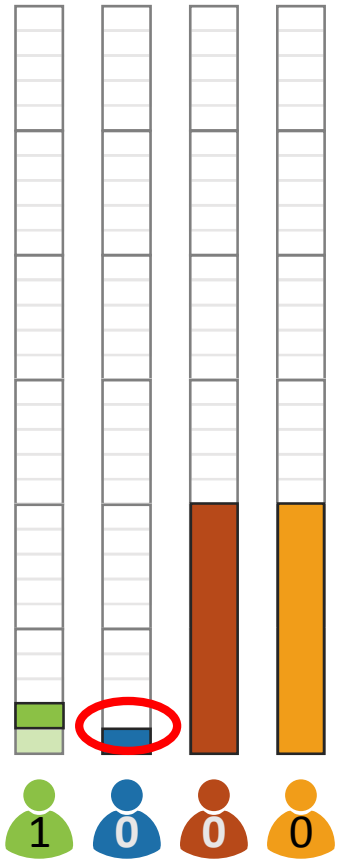


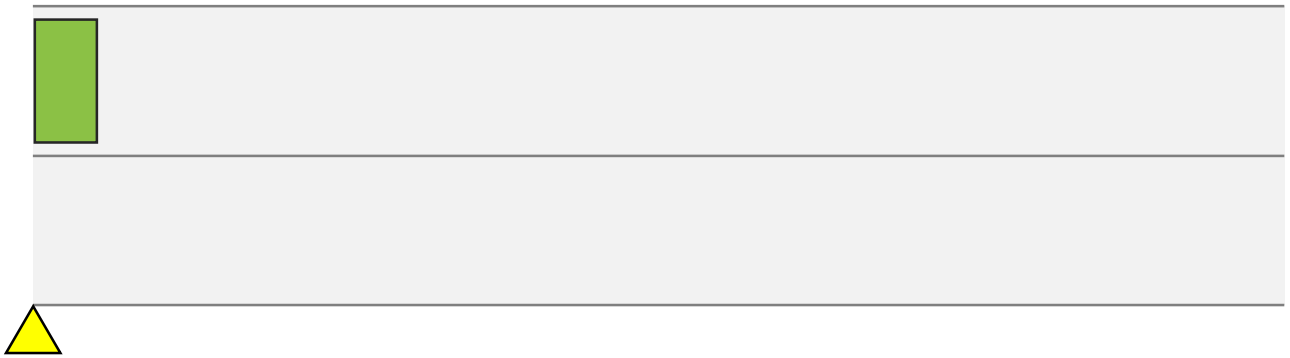




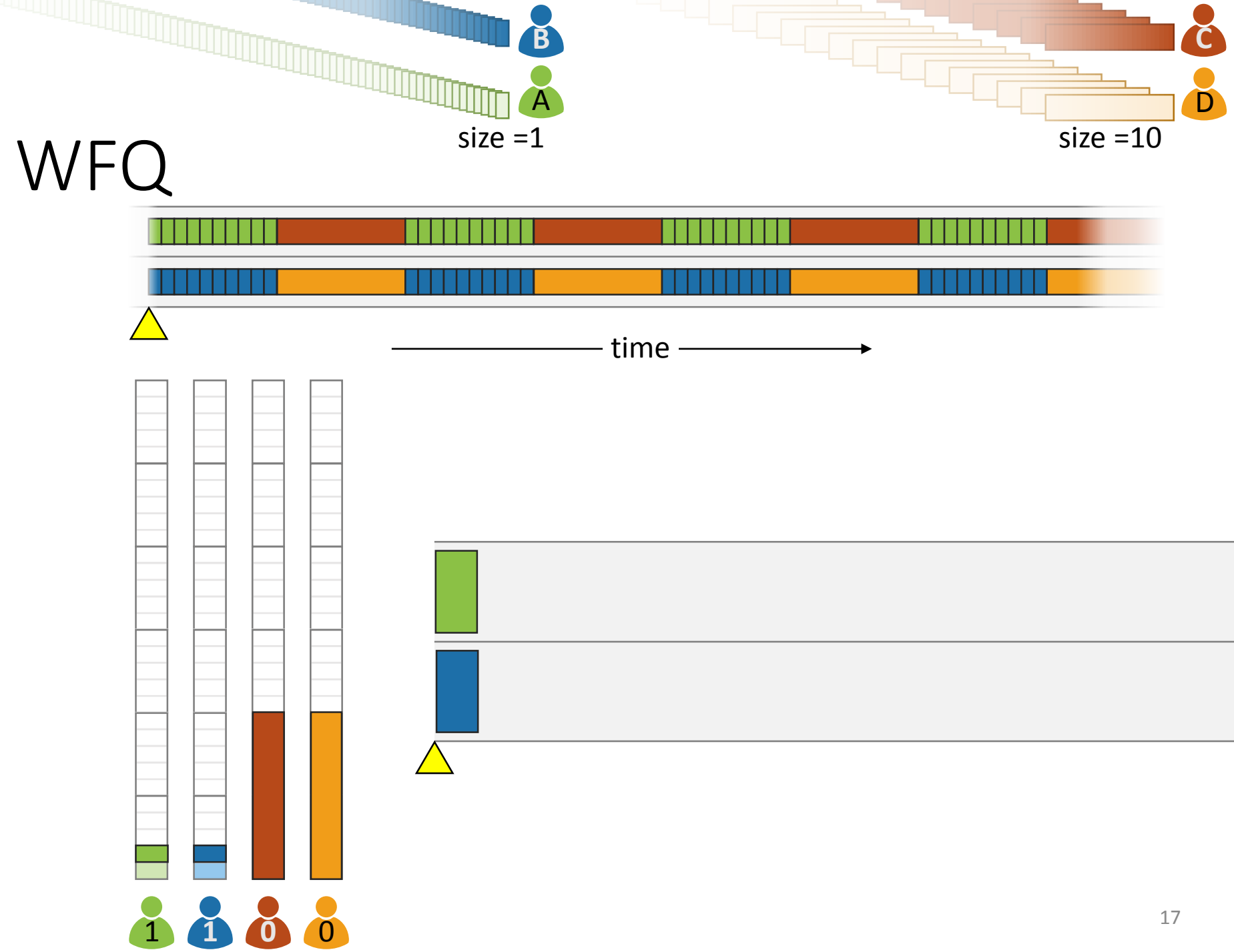
WFQ



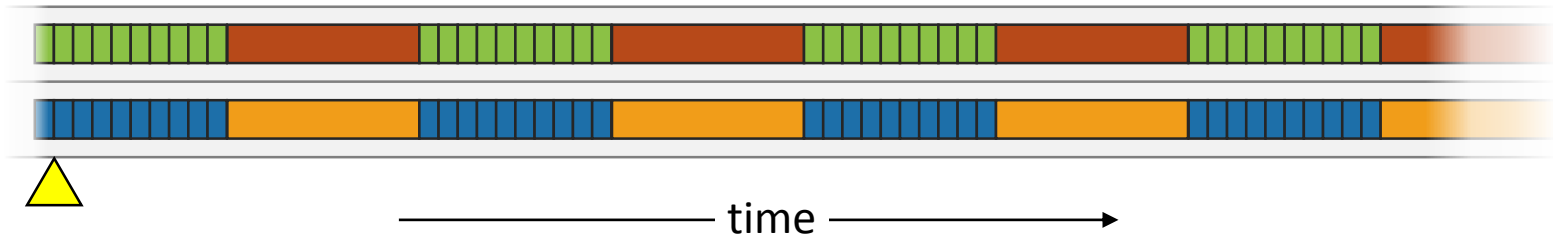


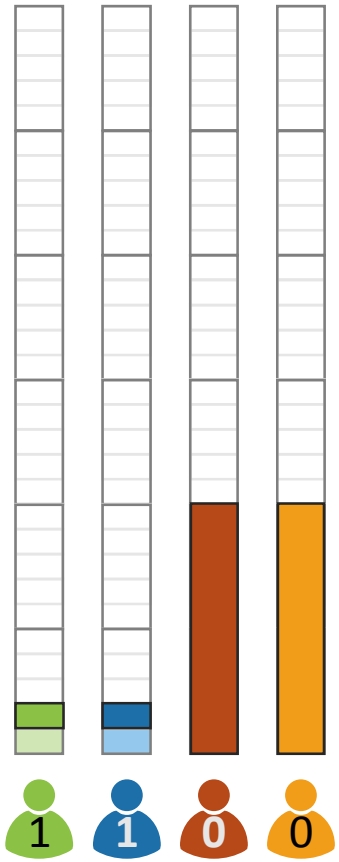


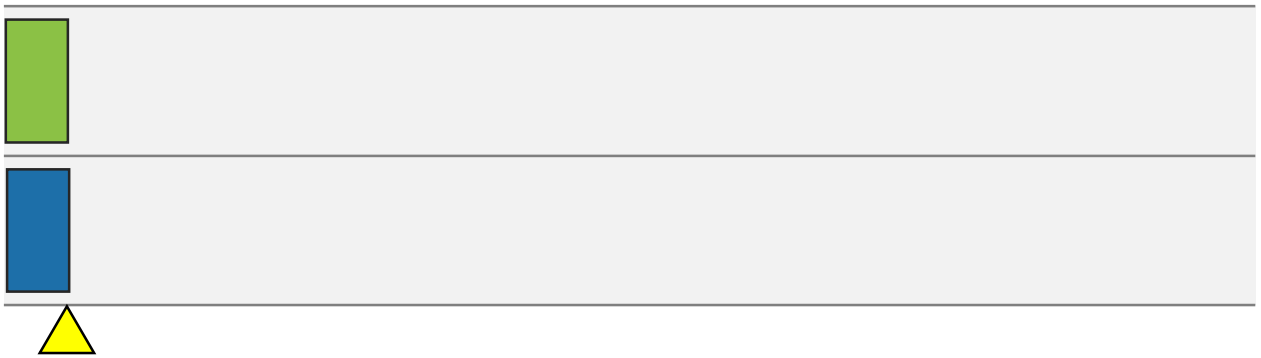
WFQ



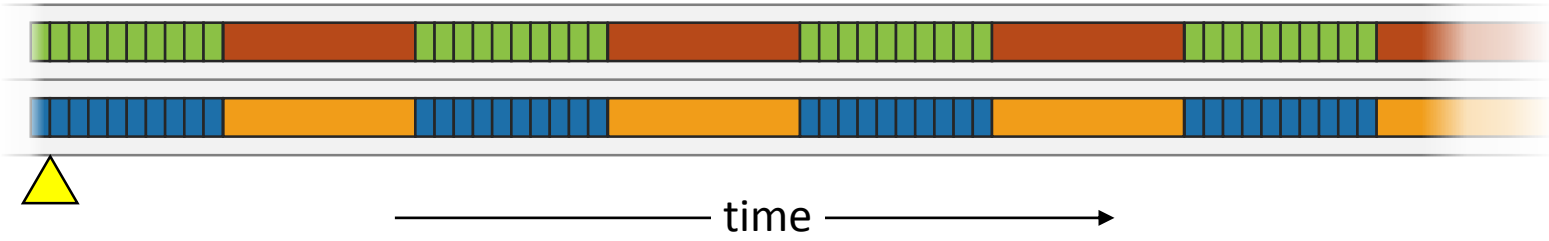
WFQ

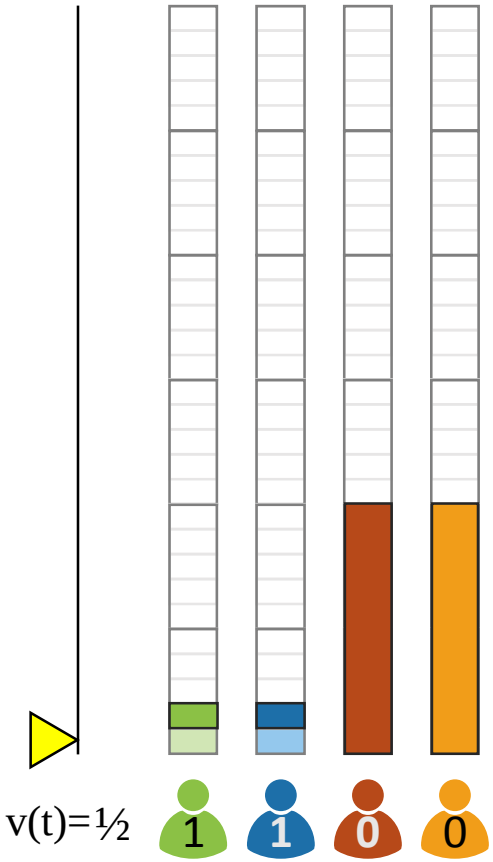


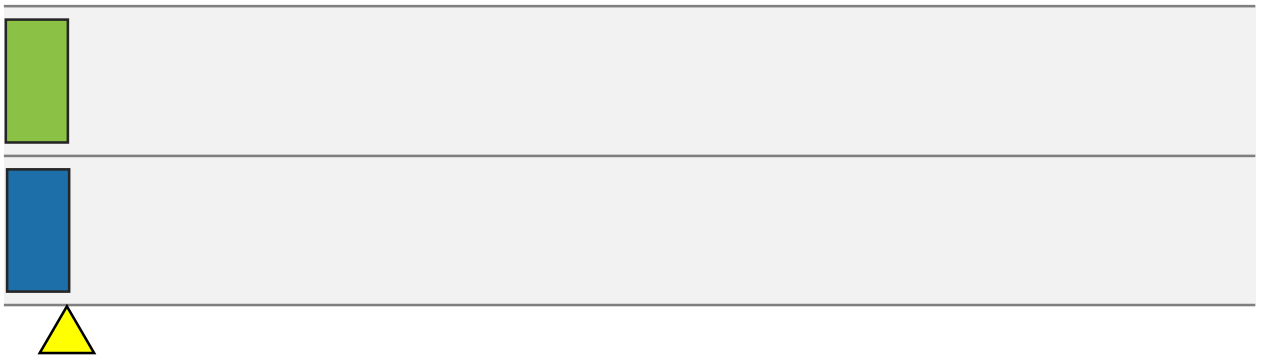




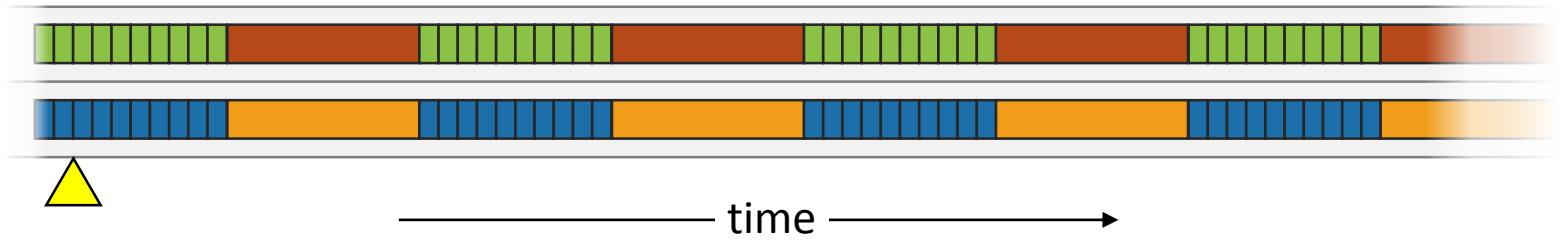
WFQ



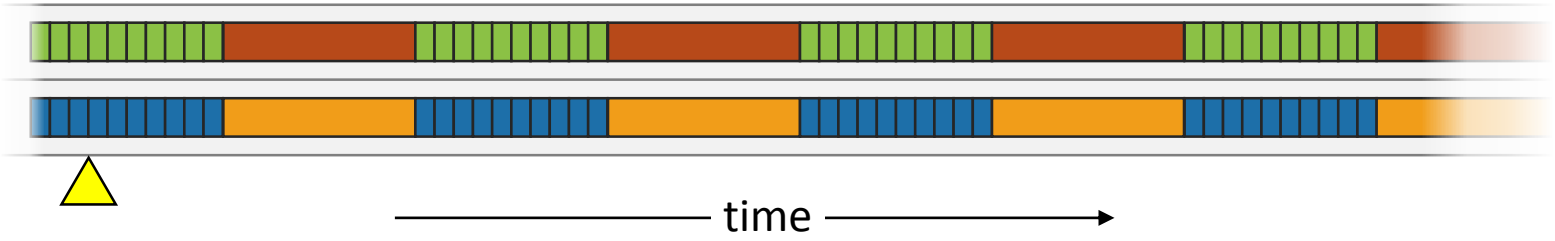


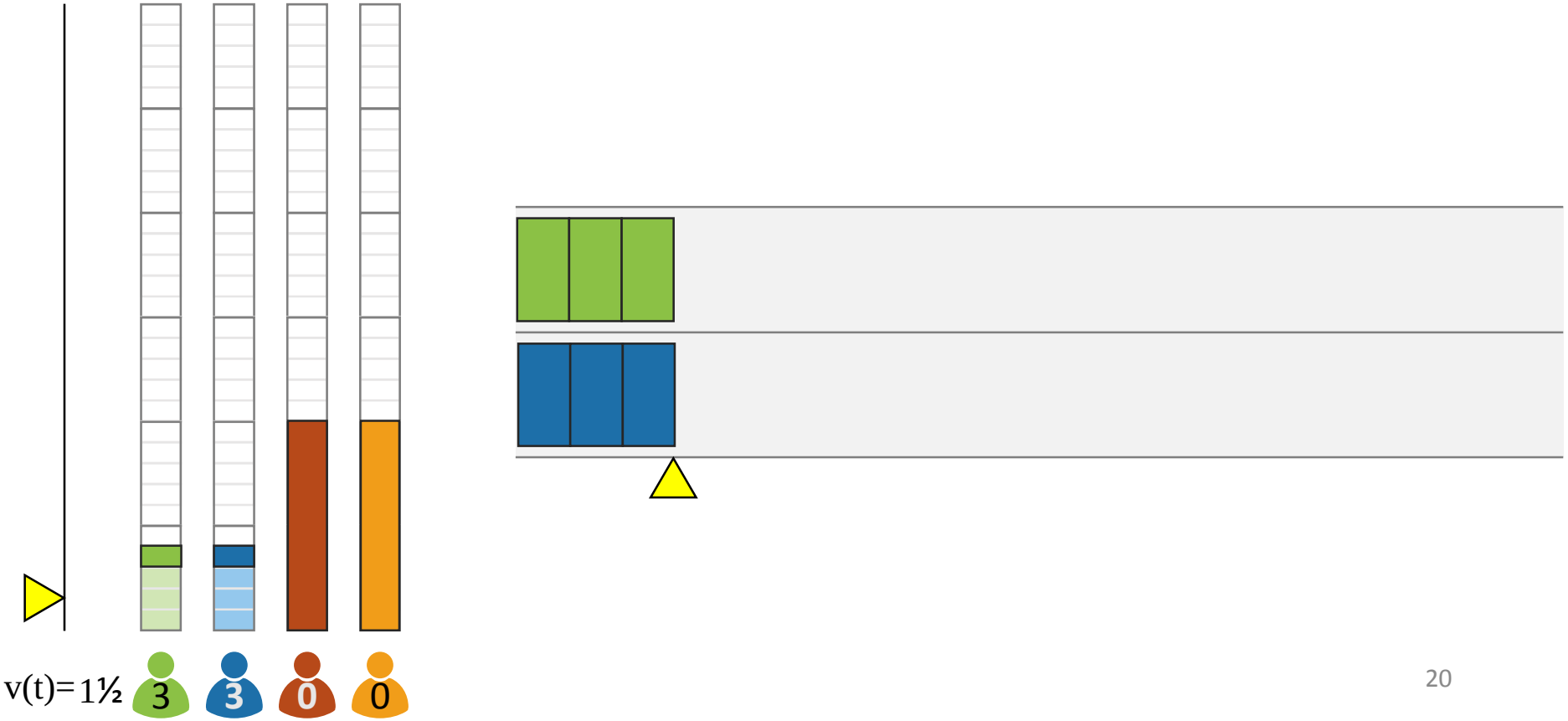


WFQ

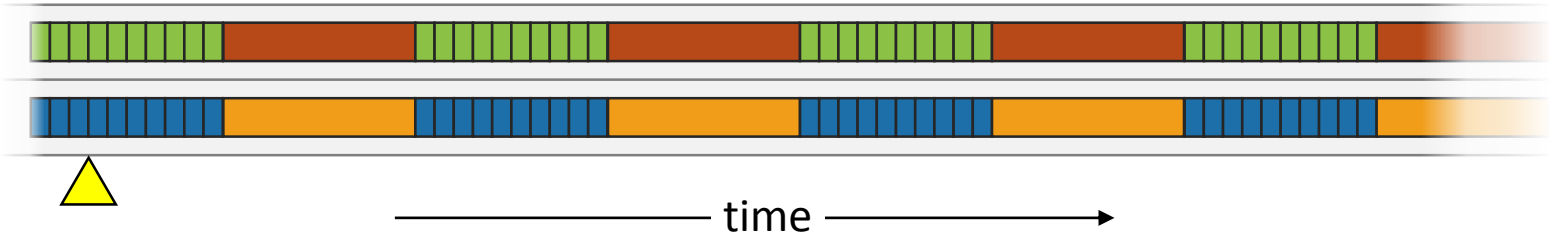


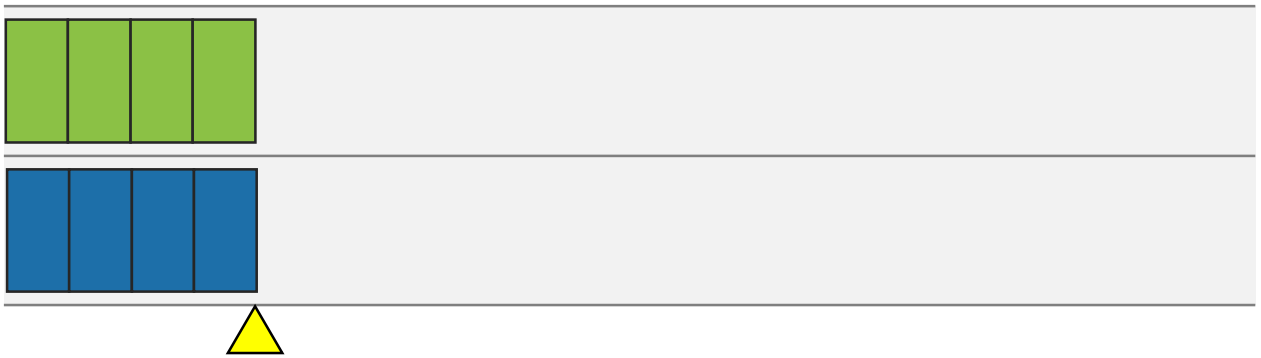
WFQ

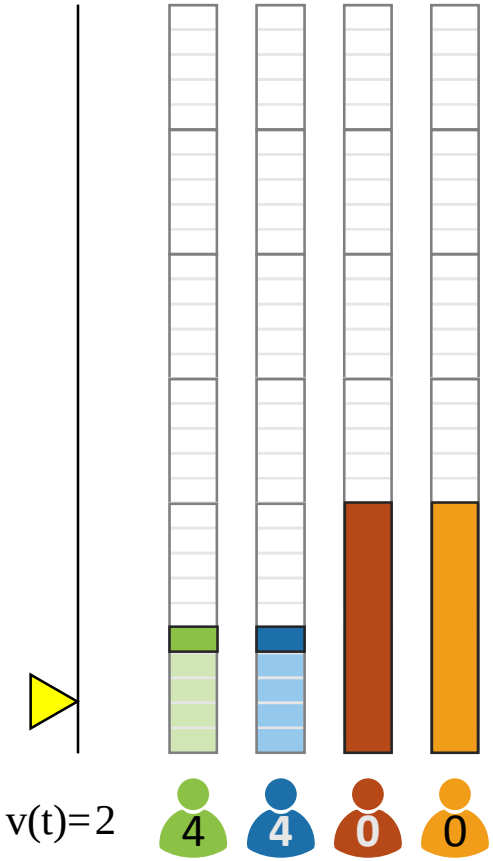




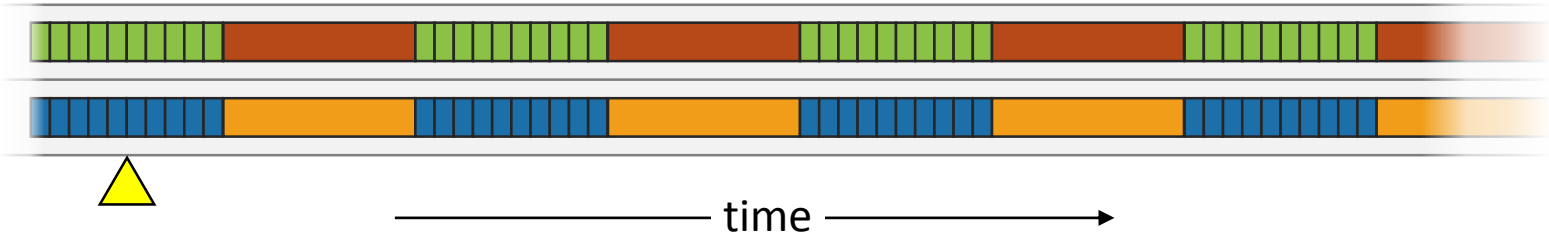
WFQ

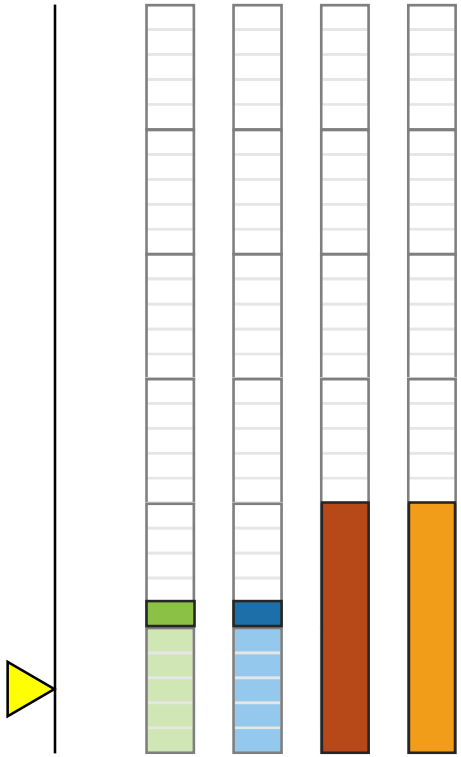


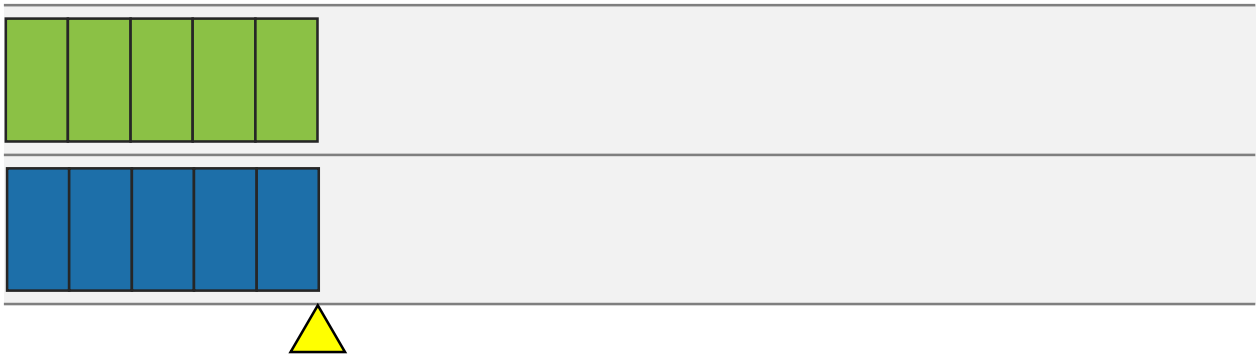




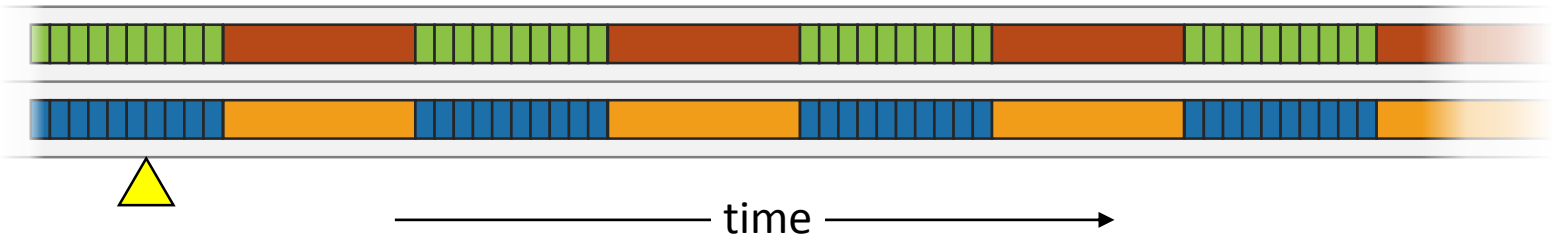
WFQ

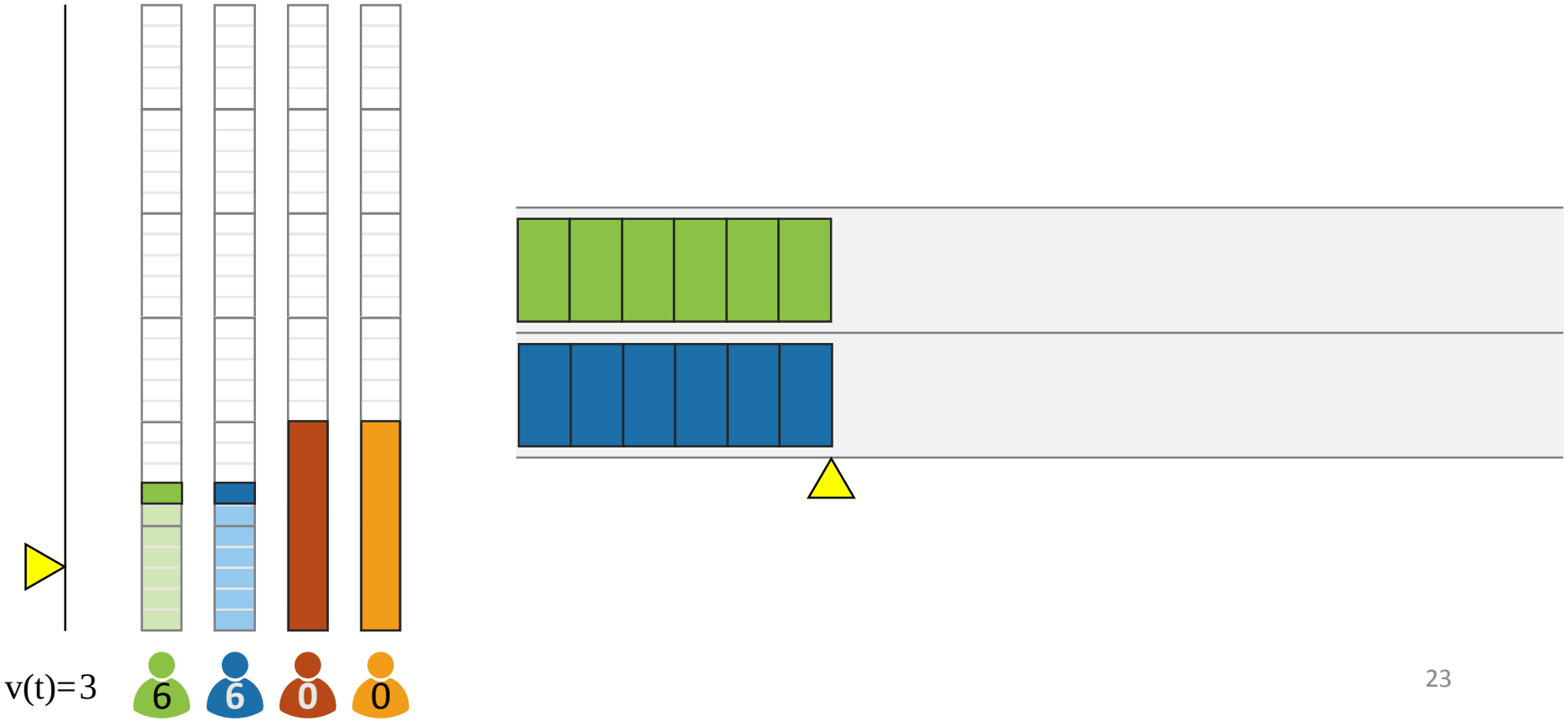




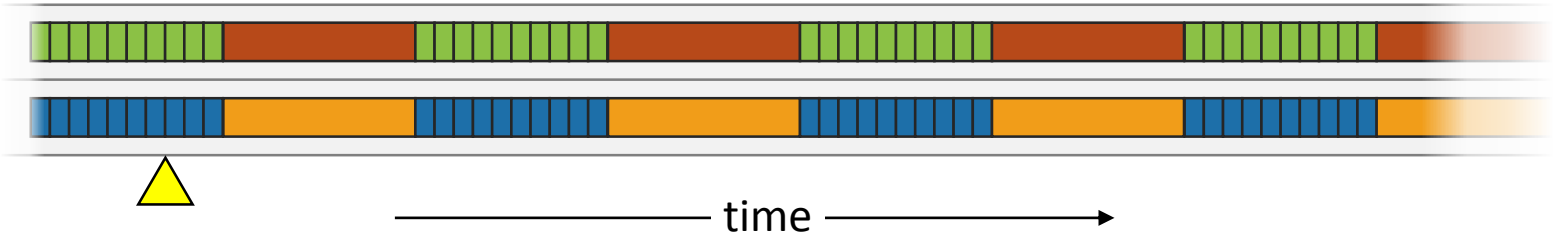


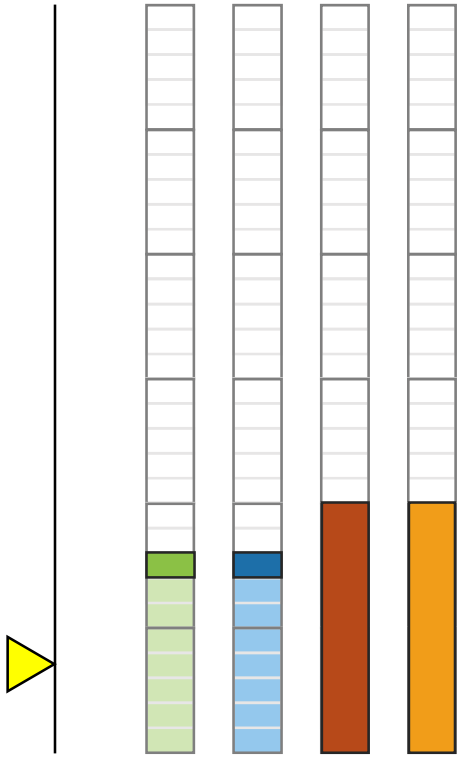
WFQ

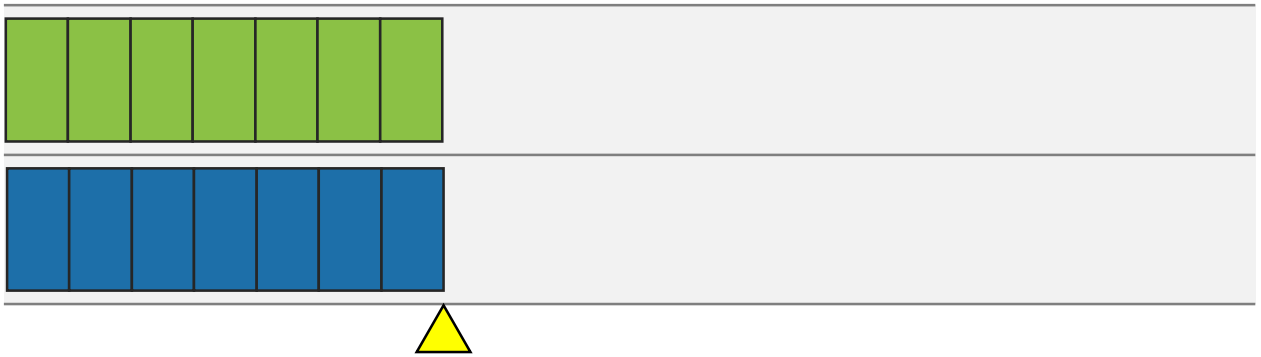




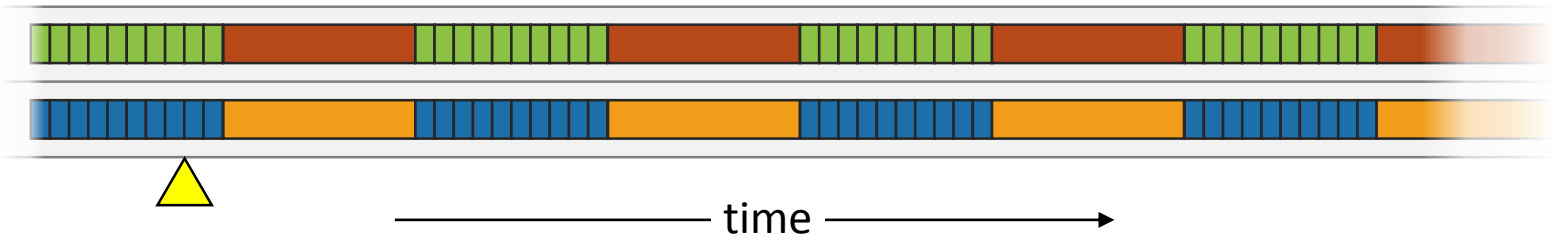
WFQ

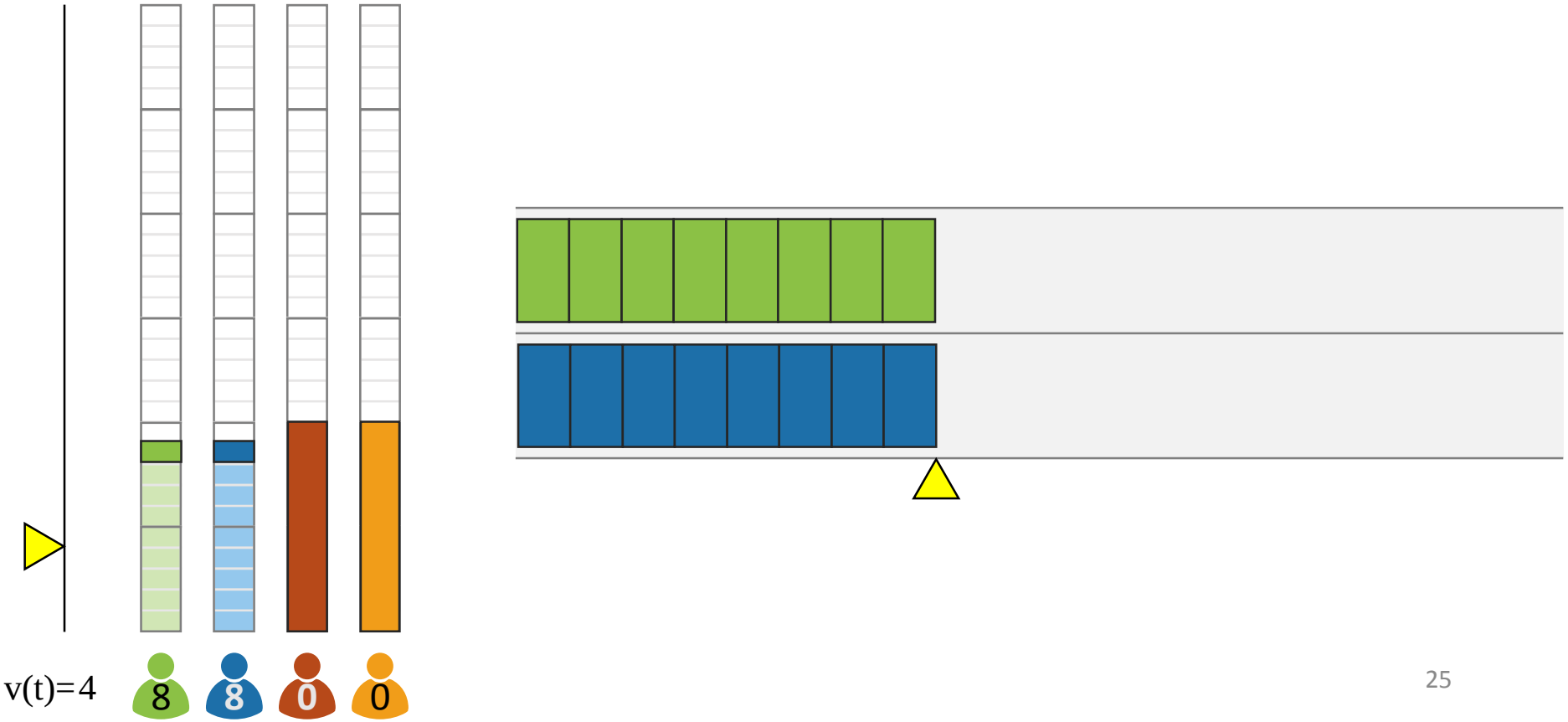




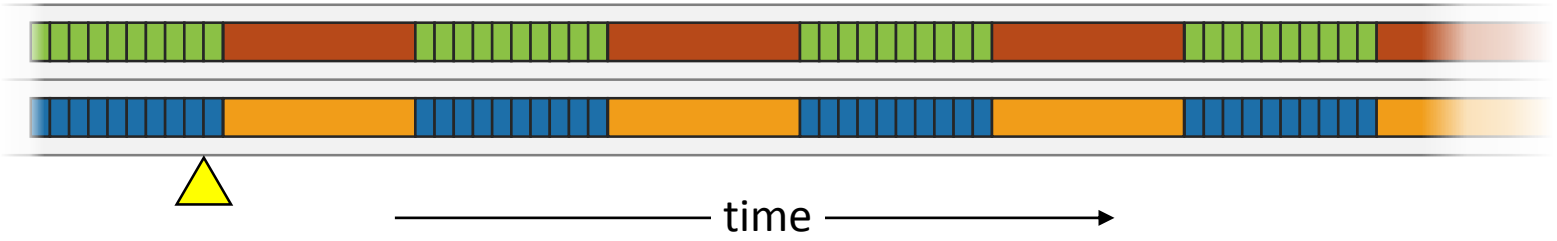


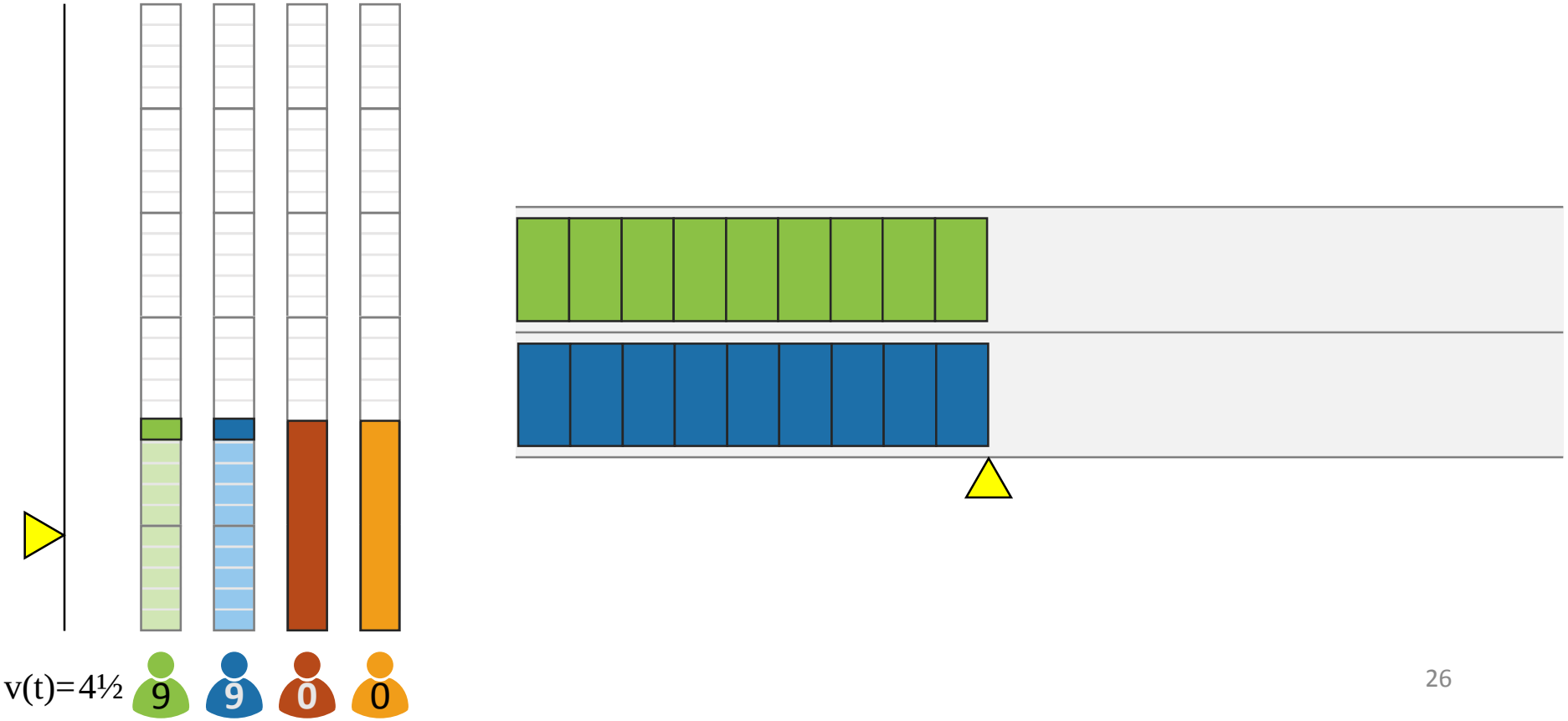
WFQ



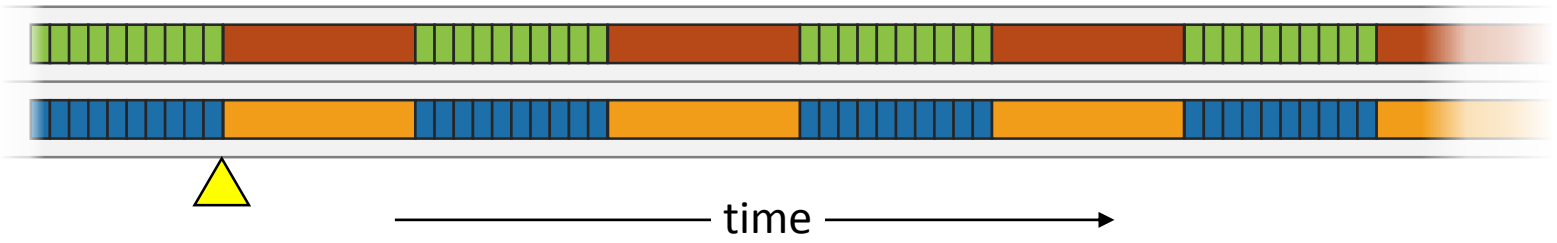


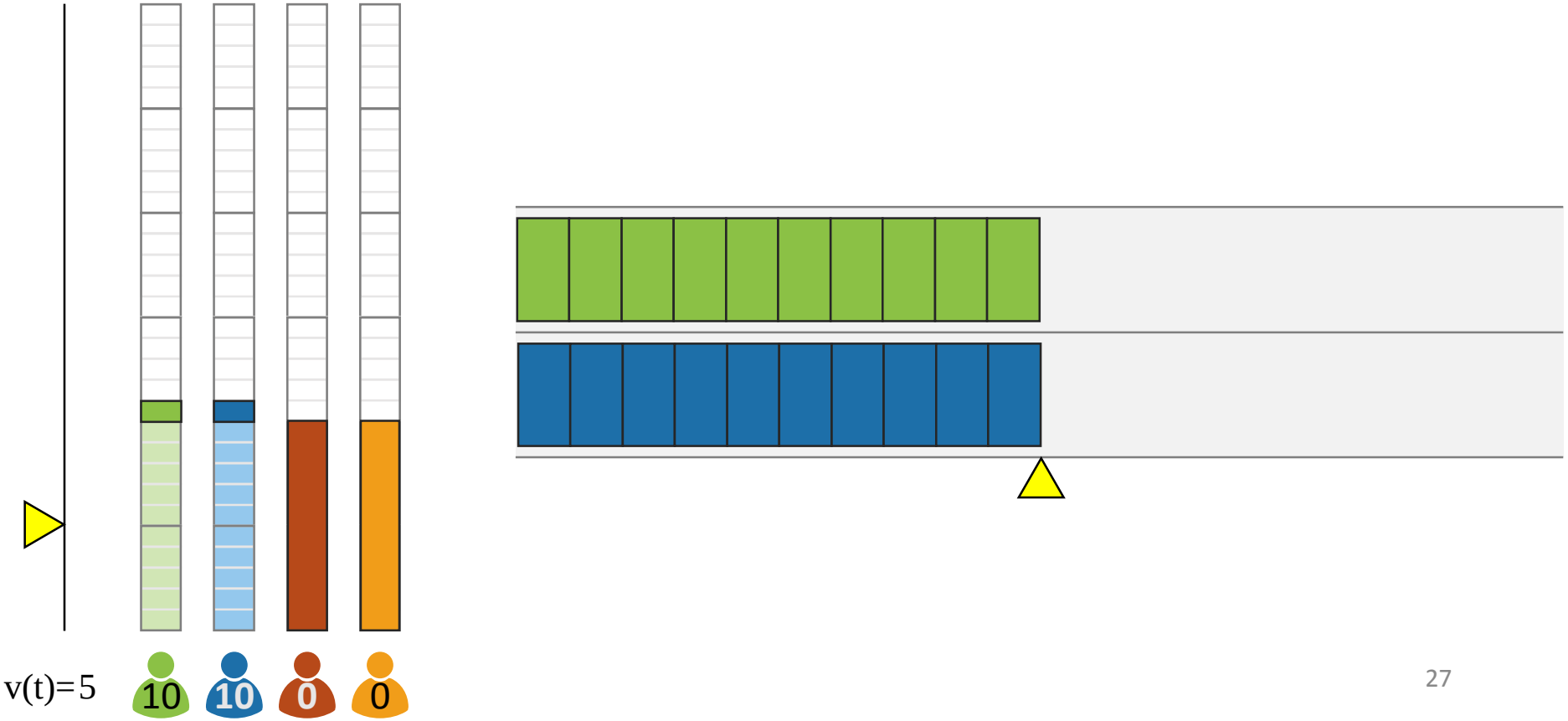
WFQ



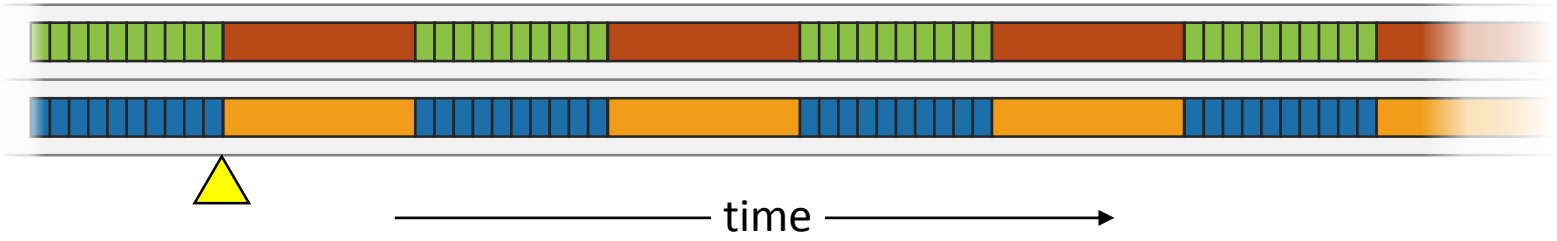


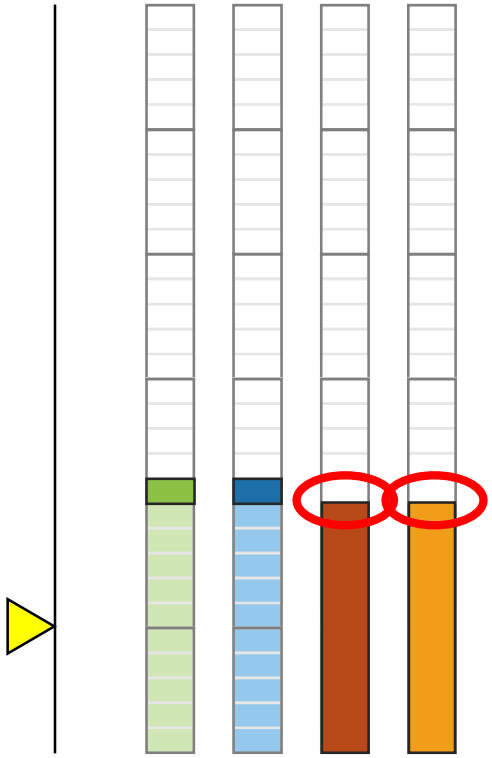
WFQ

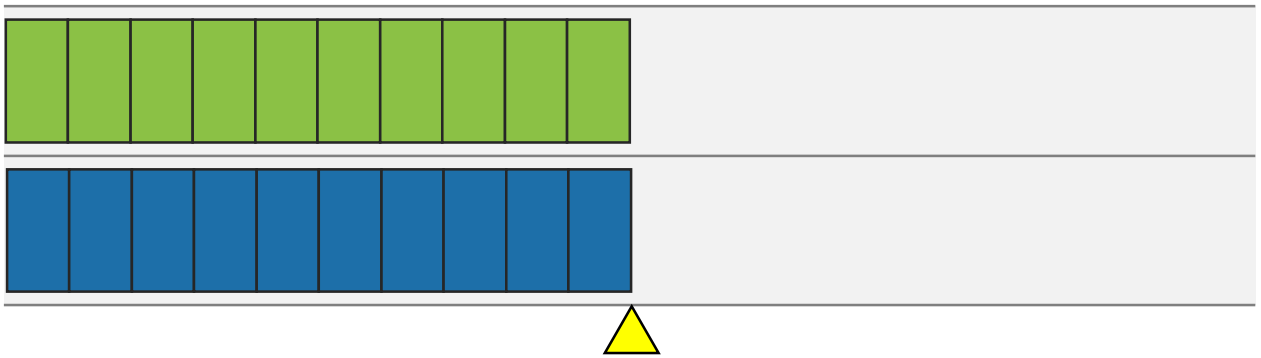




WFQ

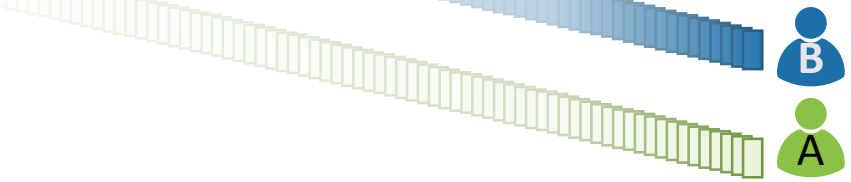






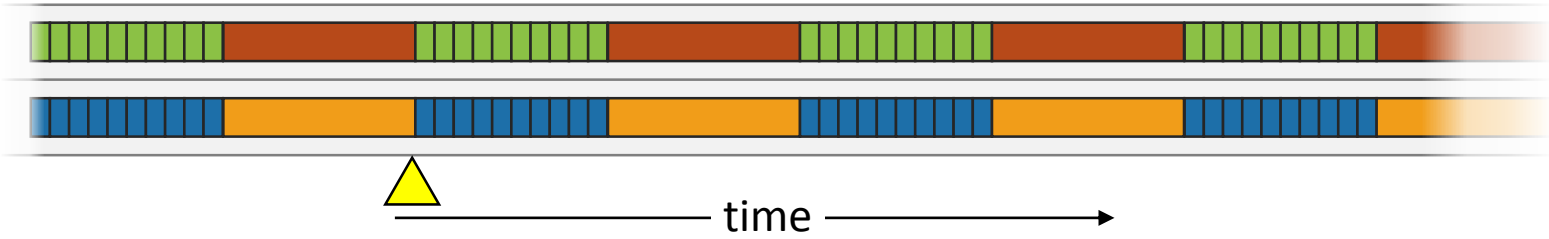
v(t)=5

WFQ

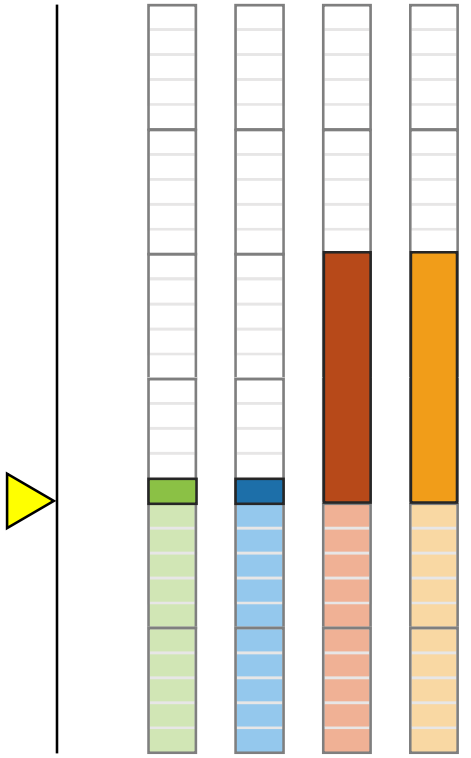


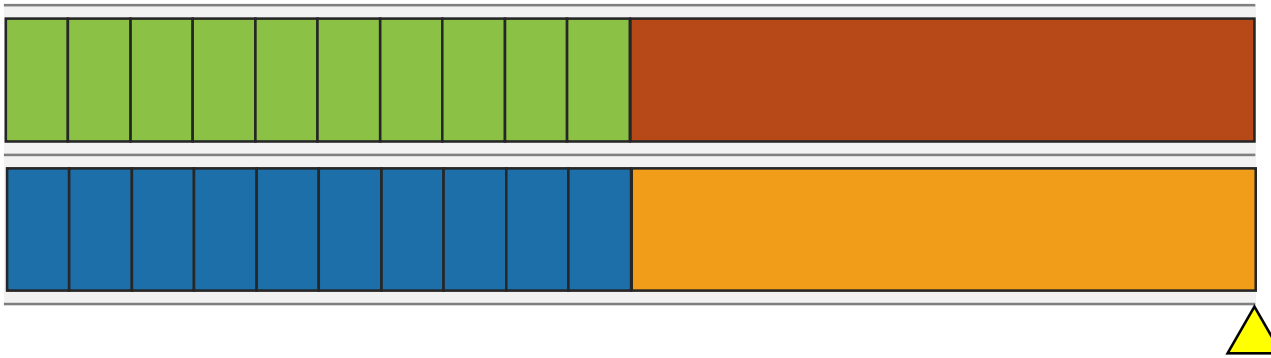
size =1

size =10



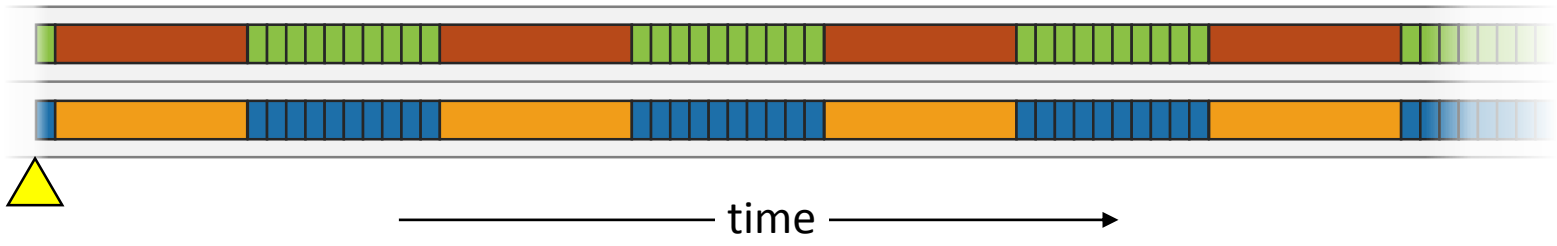
time

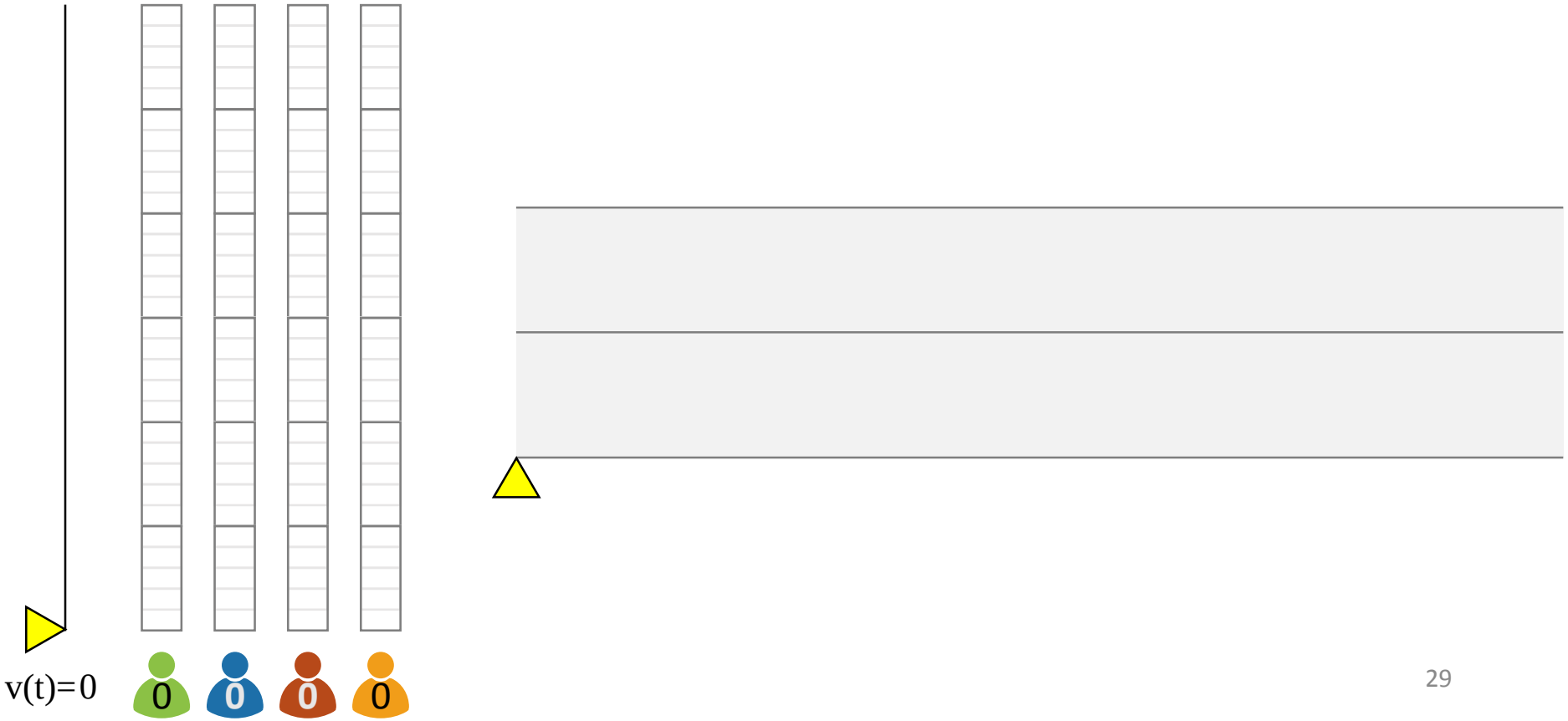




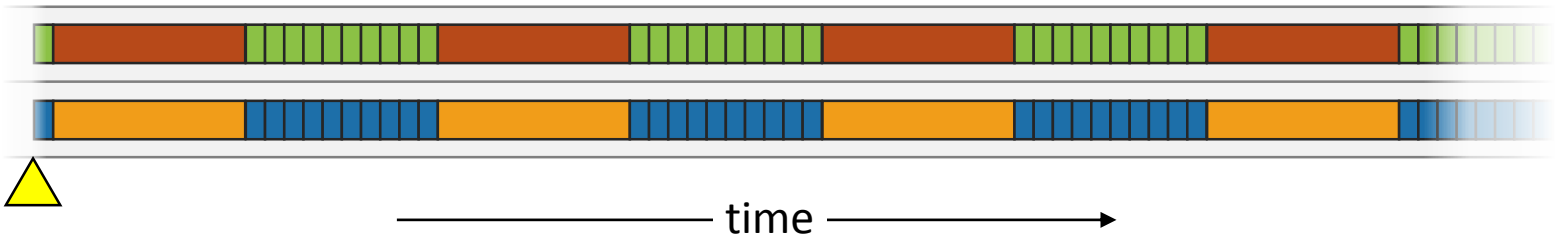
v(t)=10

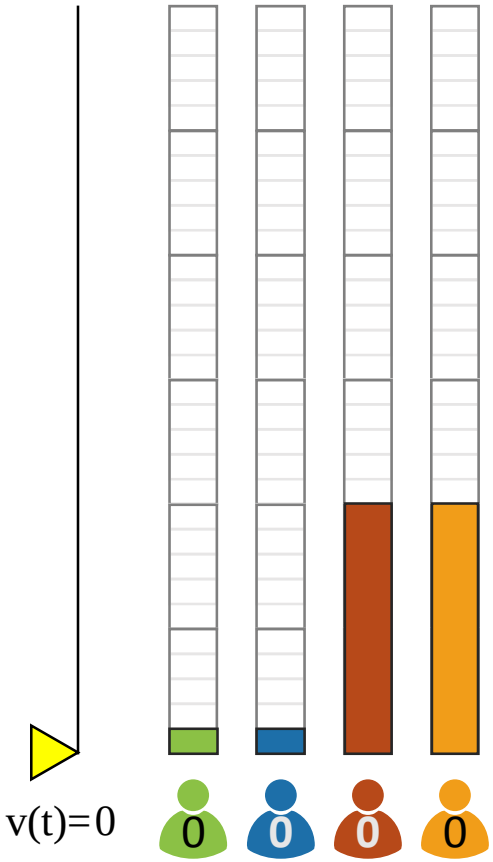
WF²Q

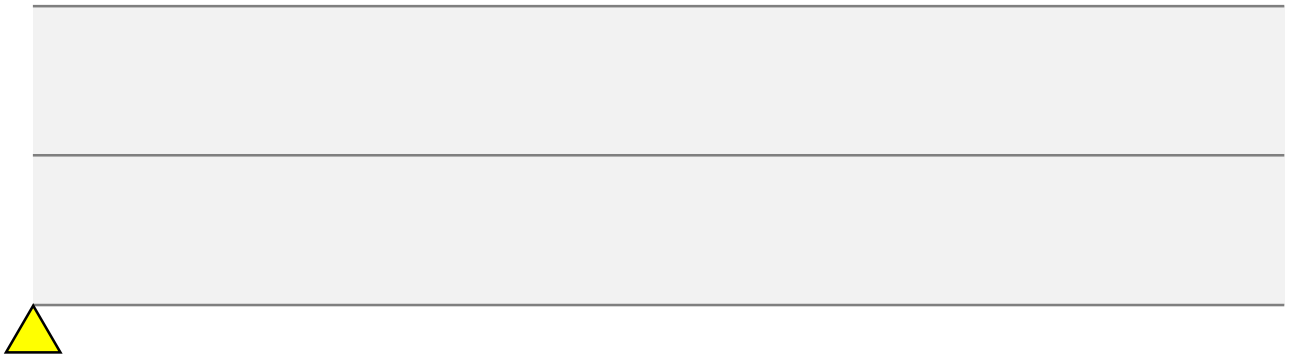




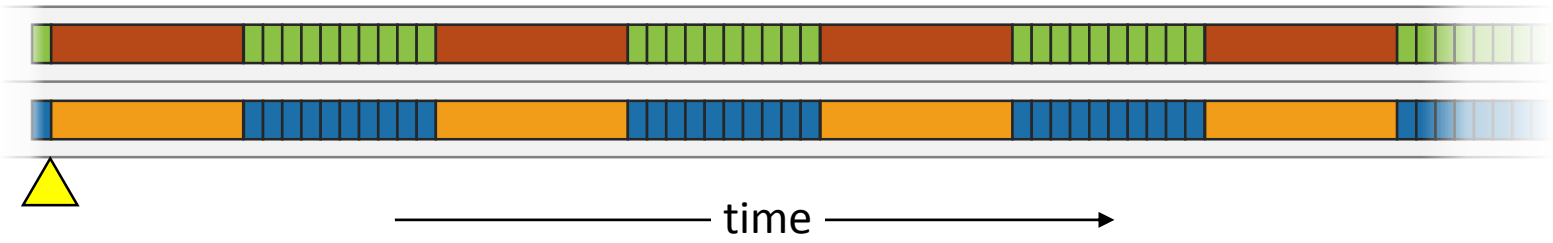
WF²Q

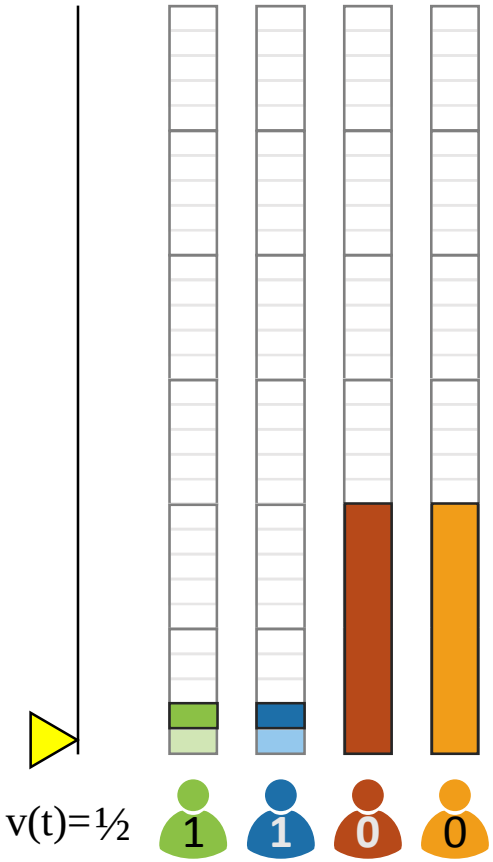


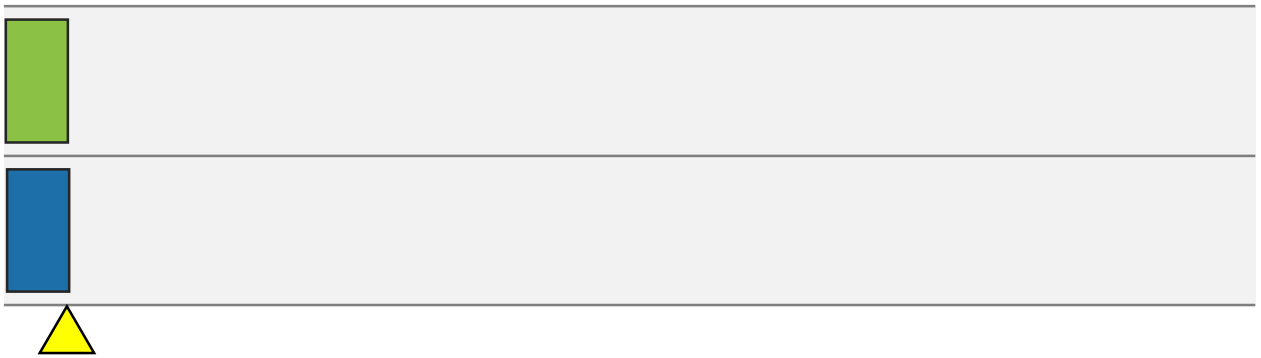




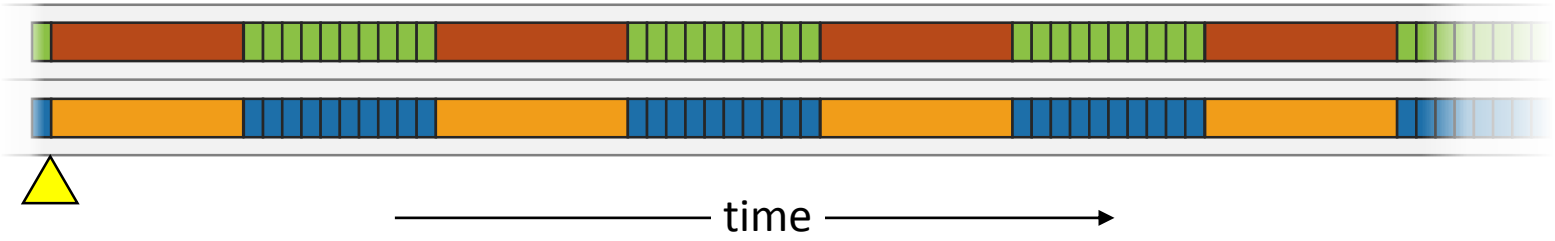
WF²Q

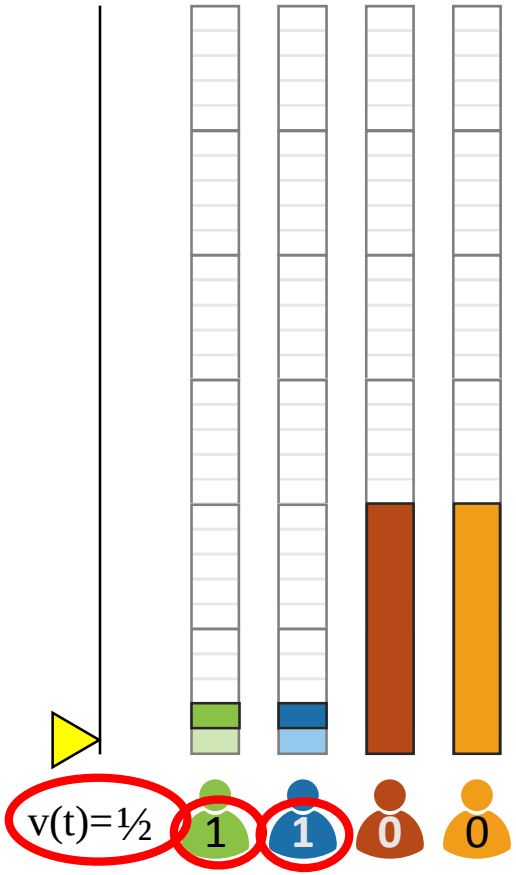


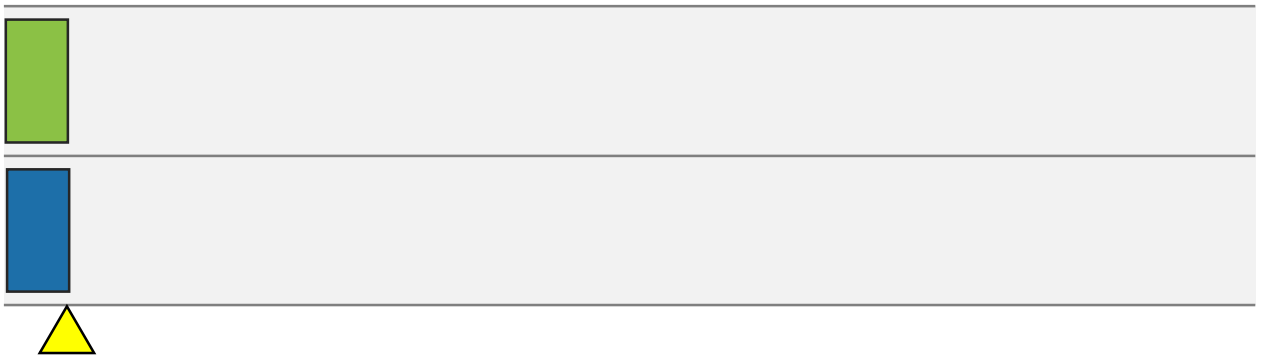




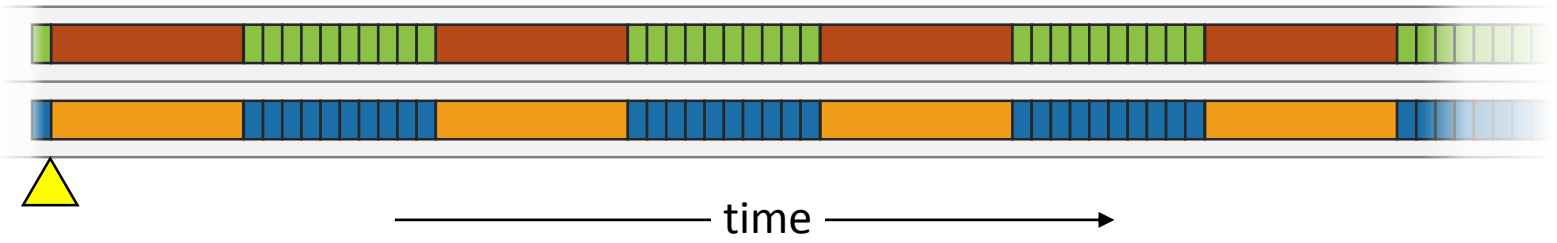
WF²Q

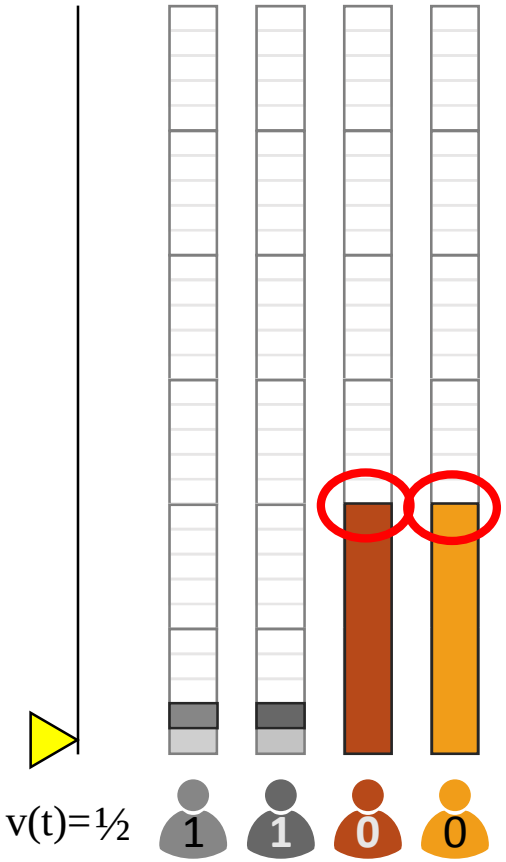


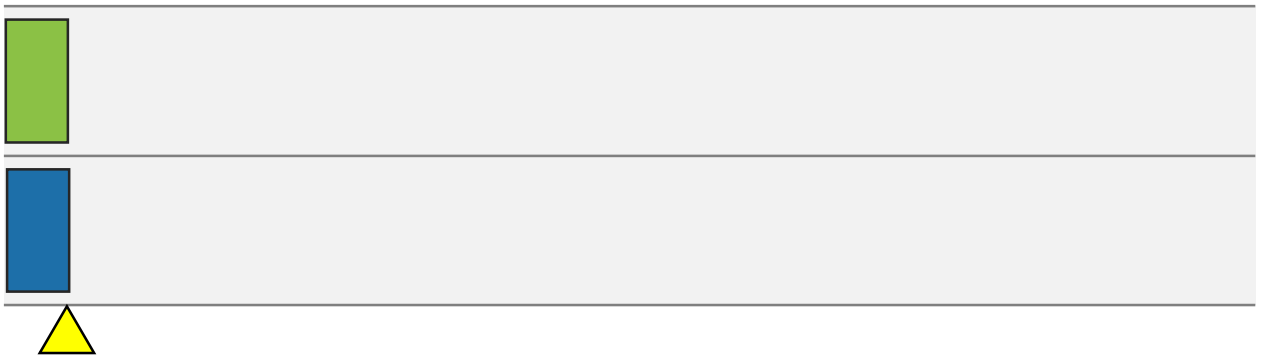




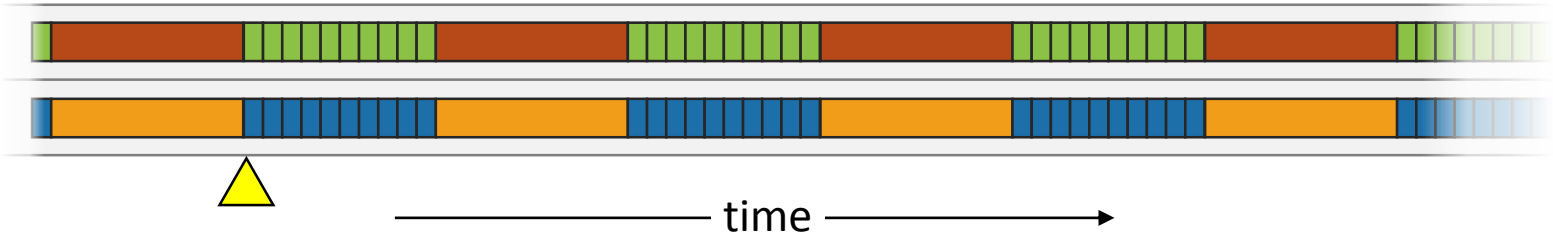
WF²Q

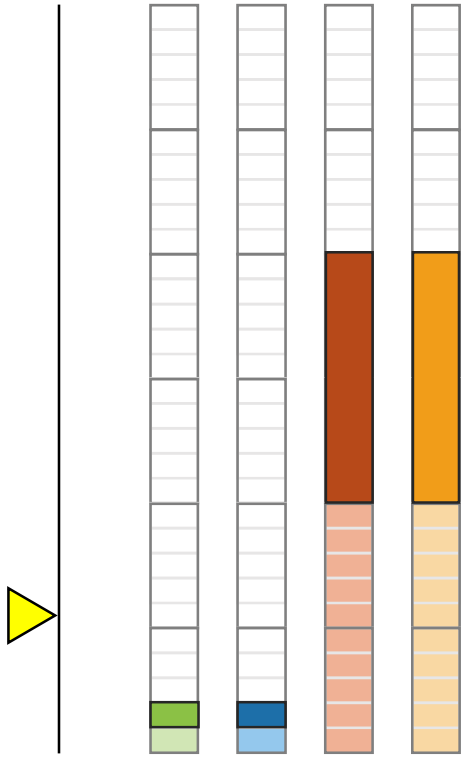


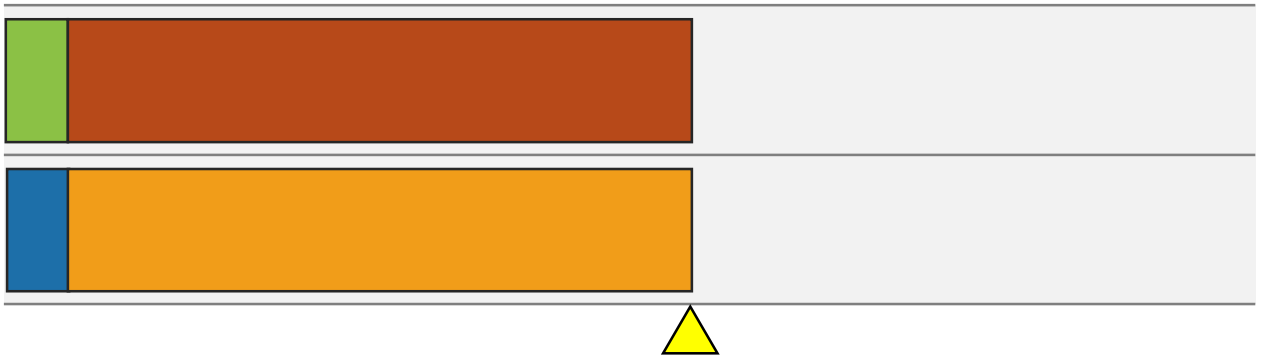




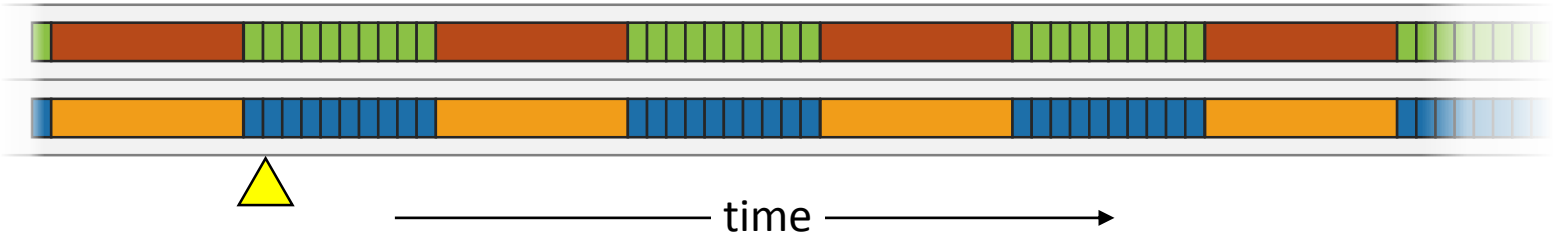
WF²Q

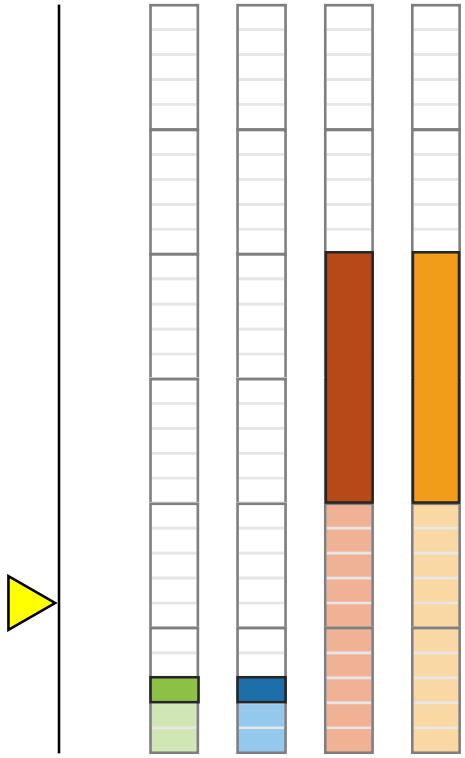


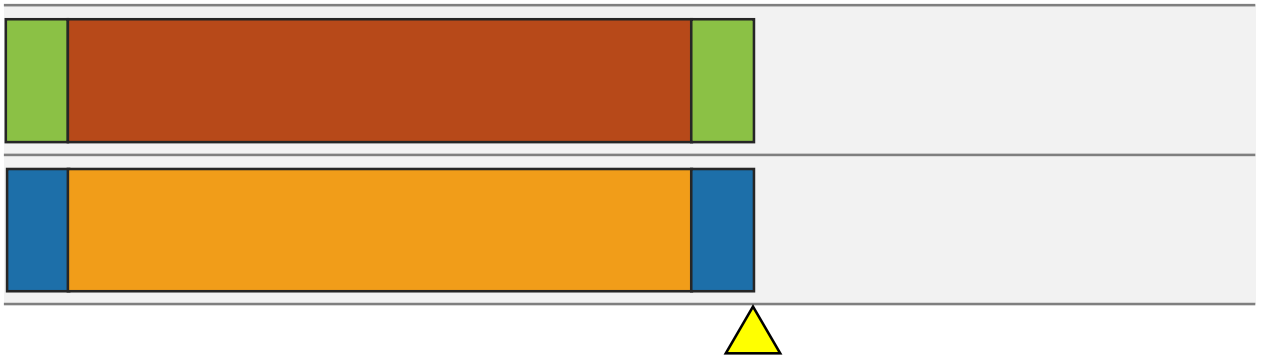




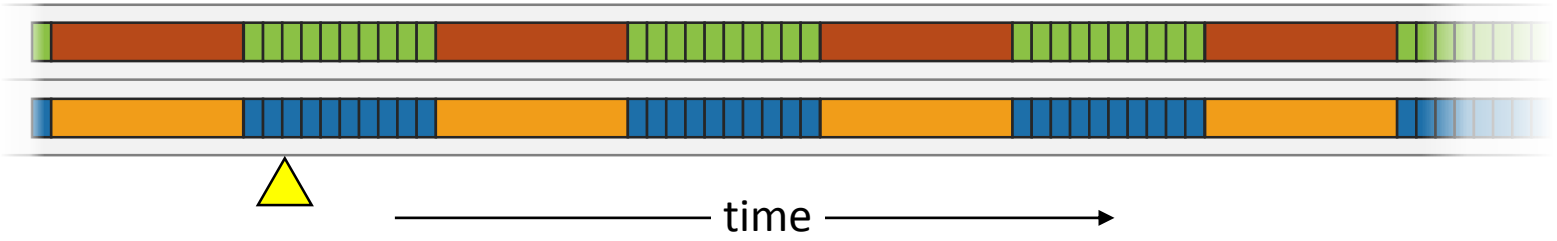
WF²Q

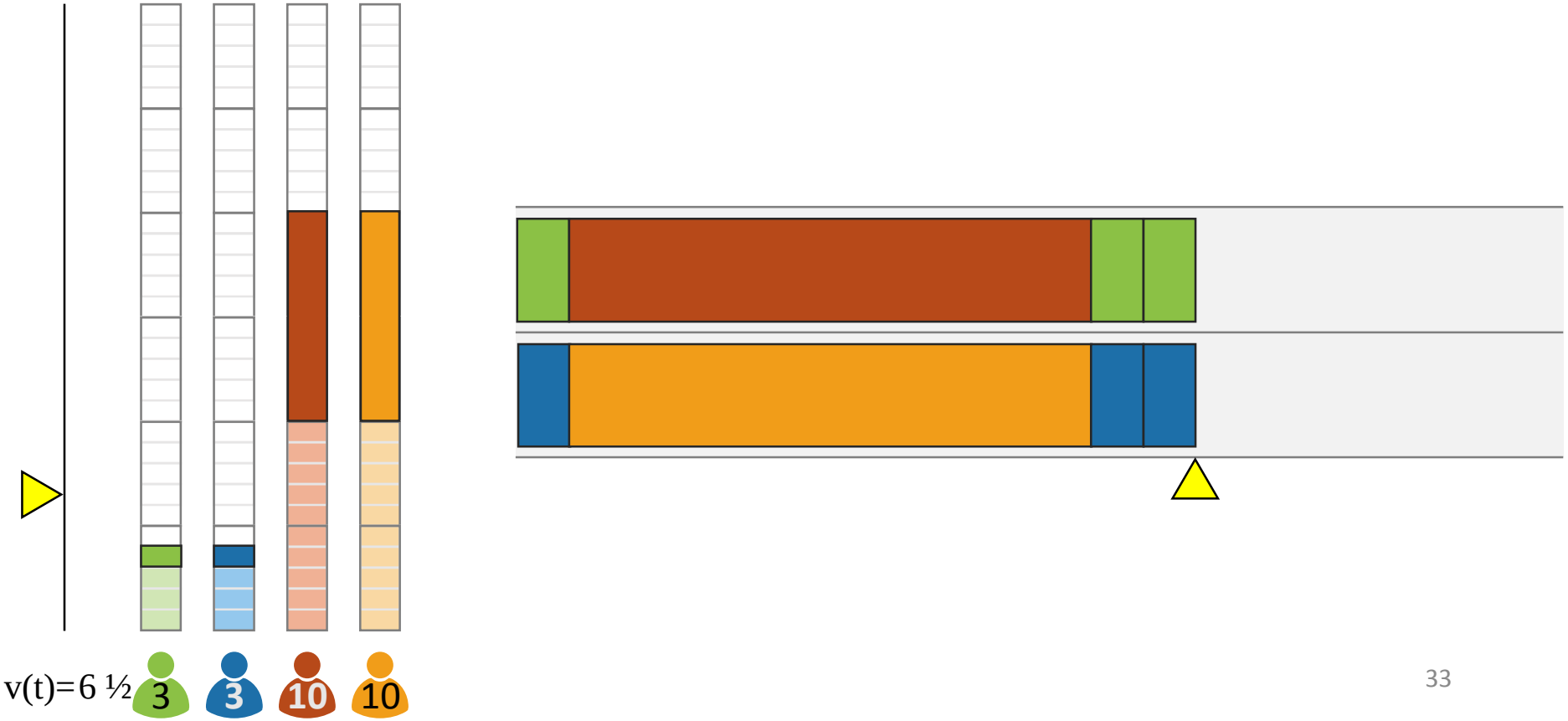




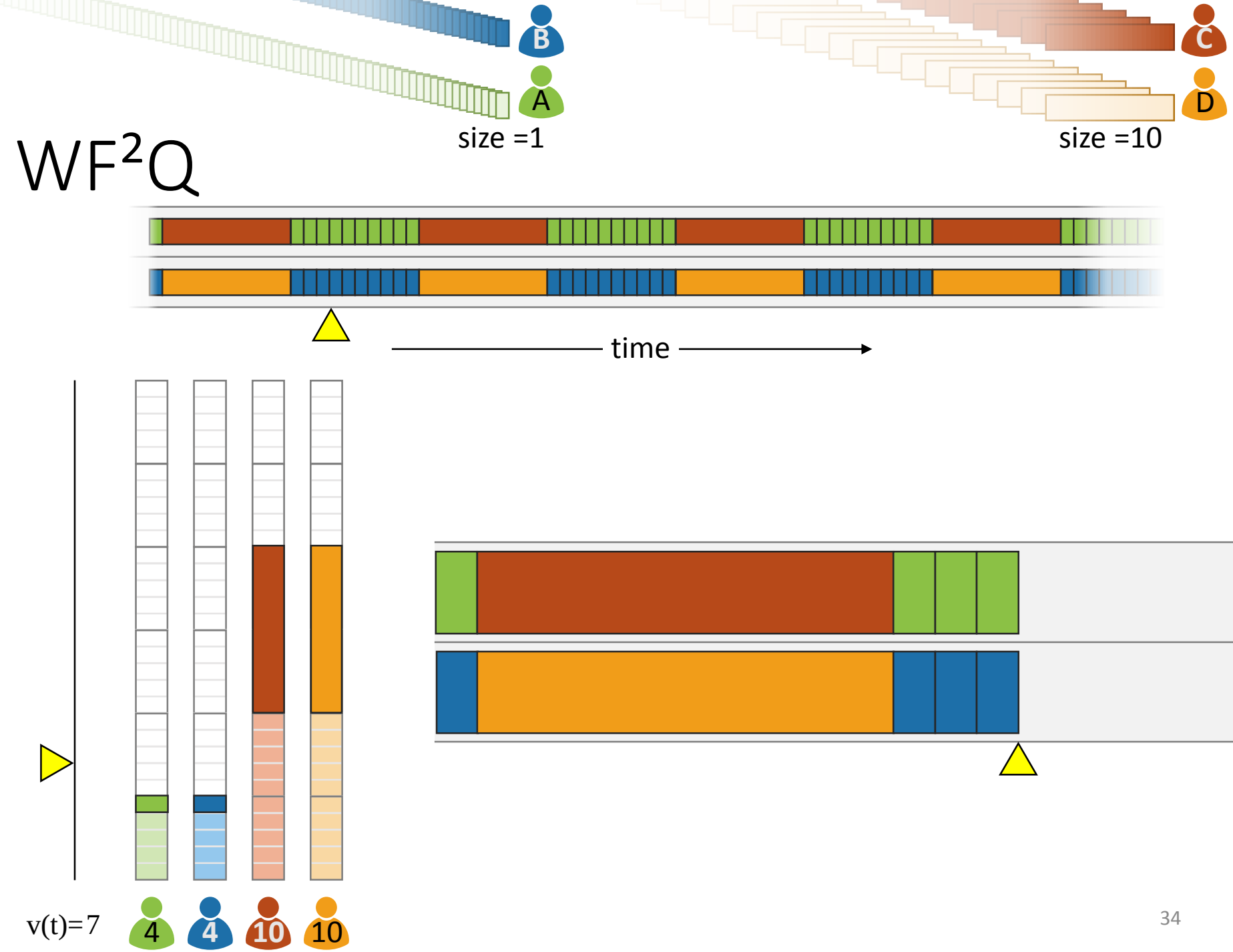


WF²Q

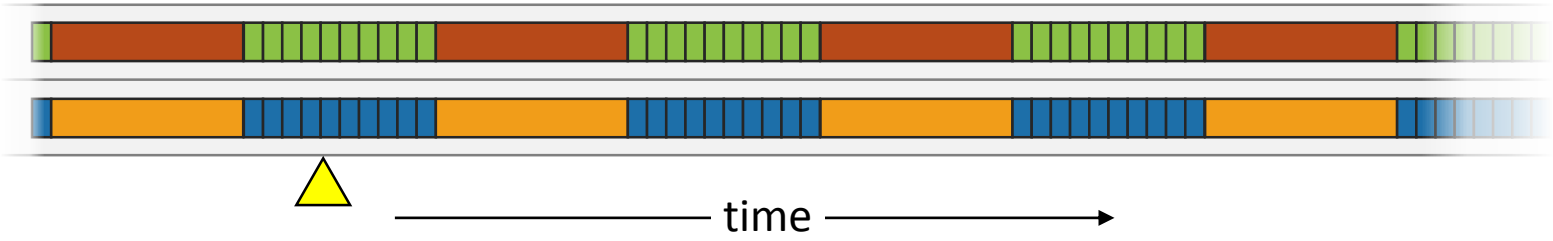


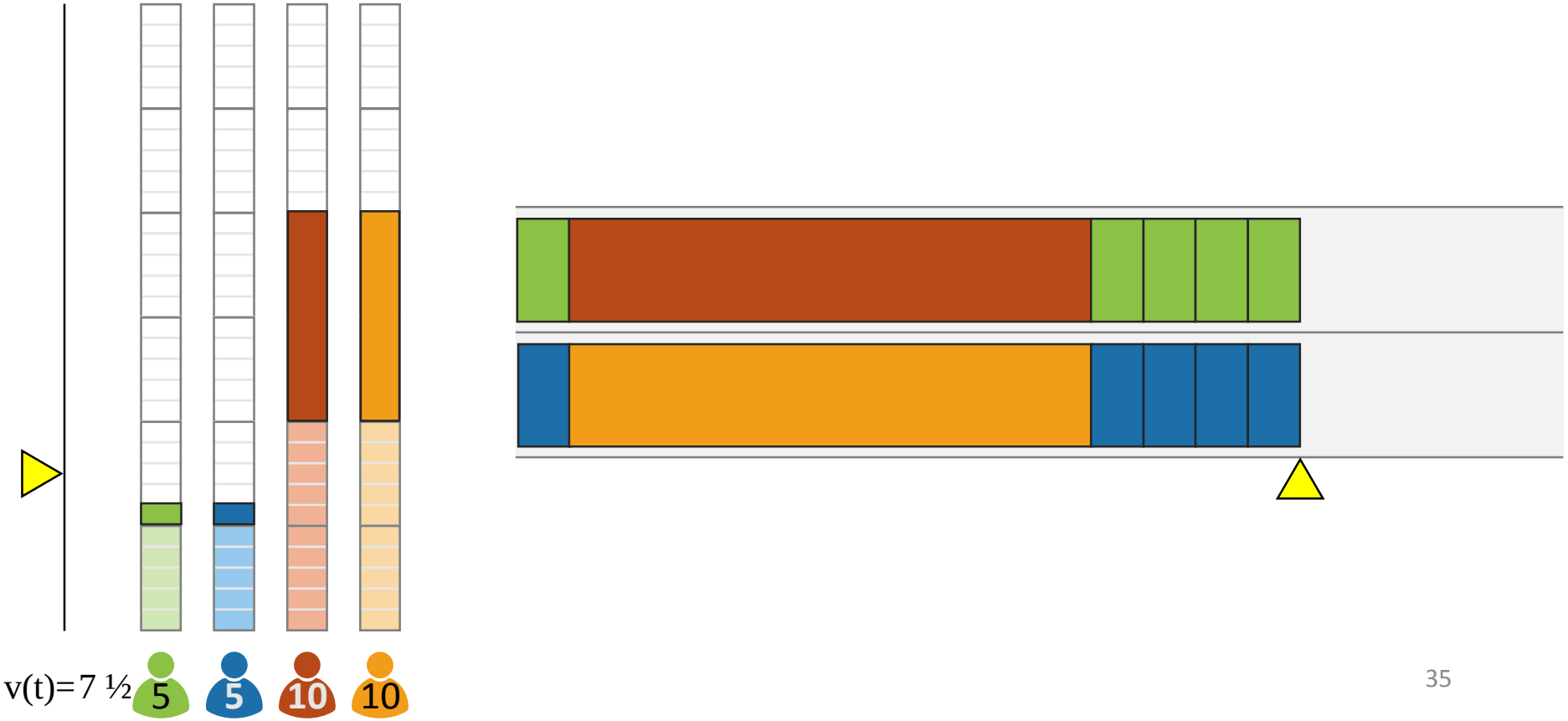


WF²Q

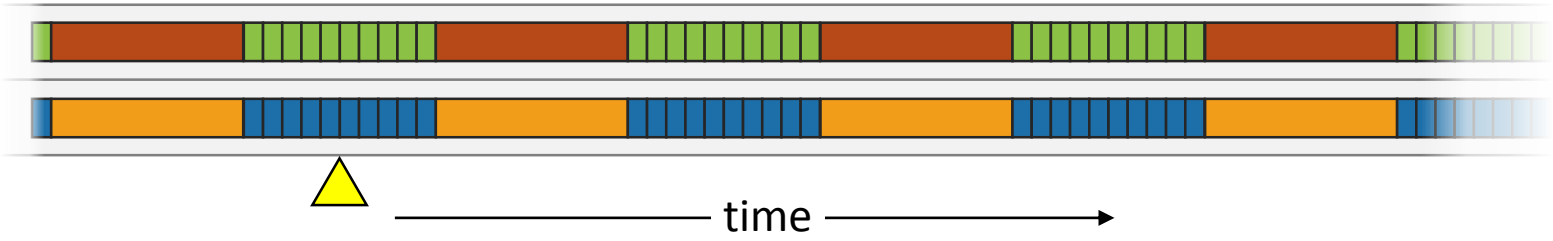


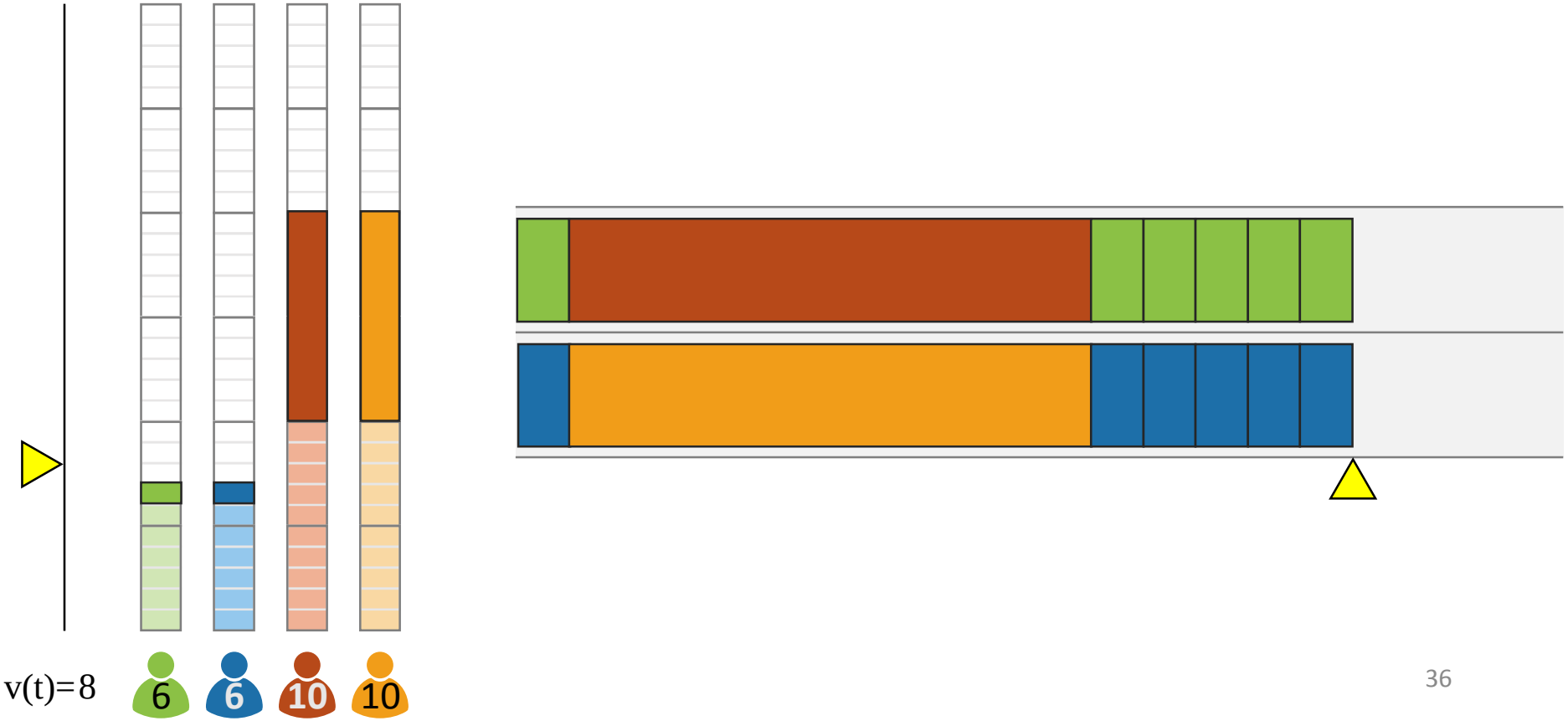
WF²Q



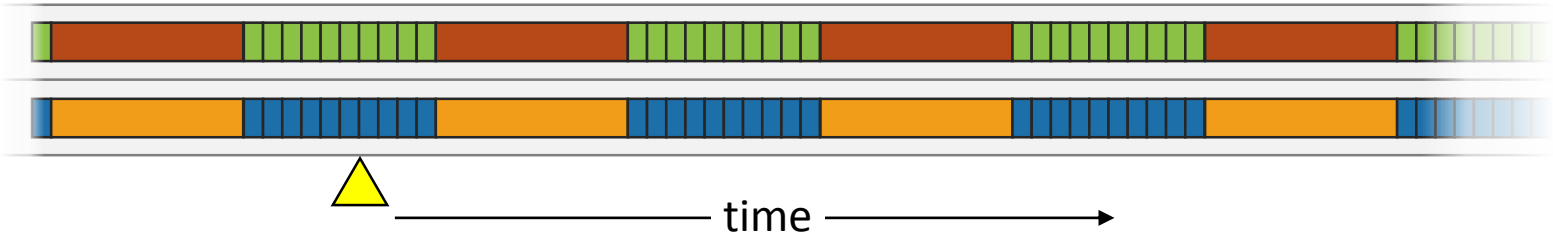


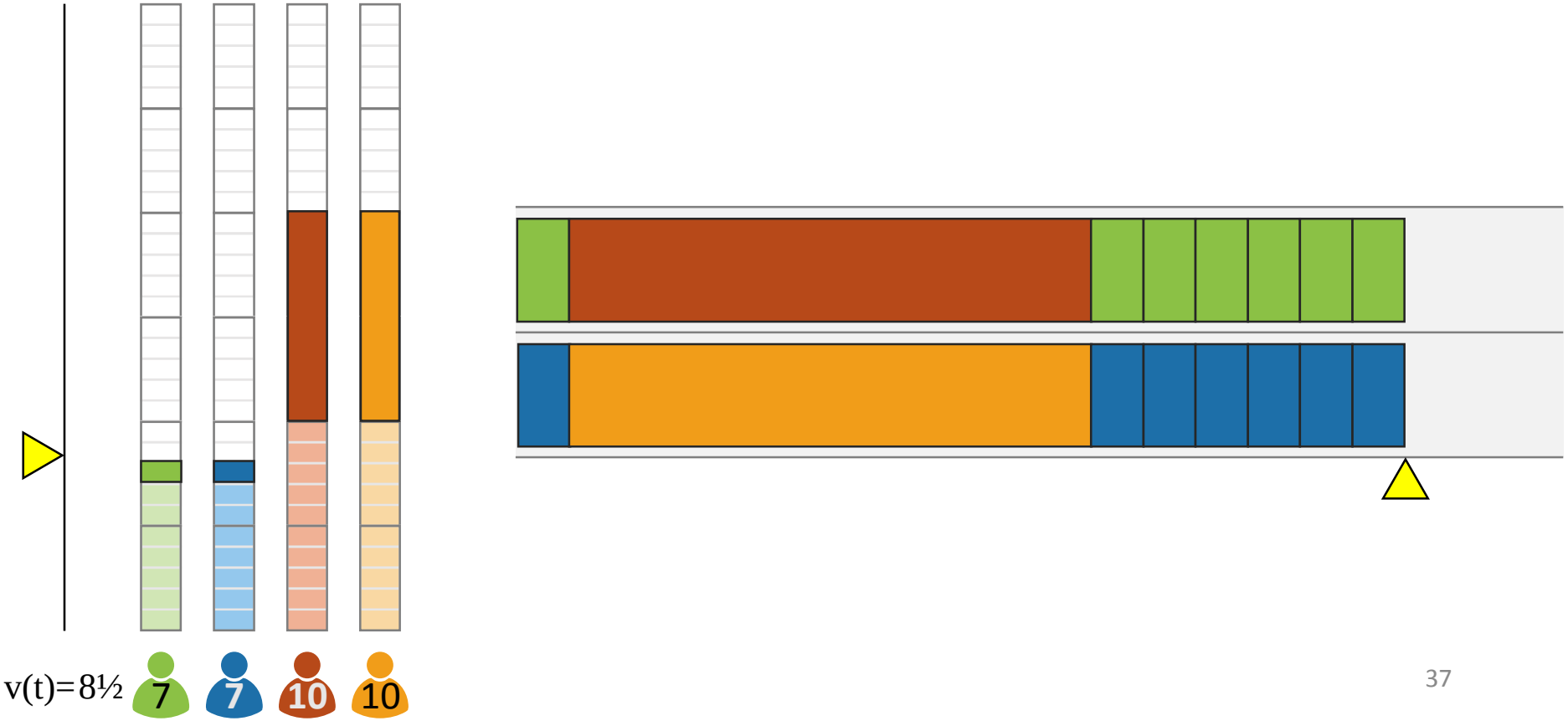
WF²Q

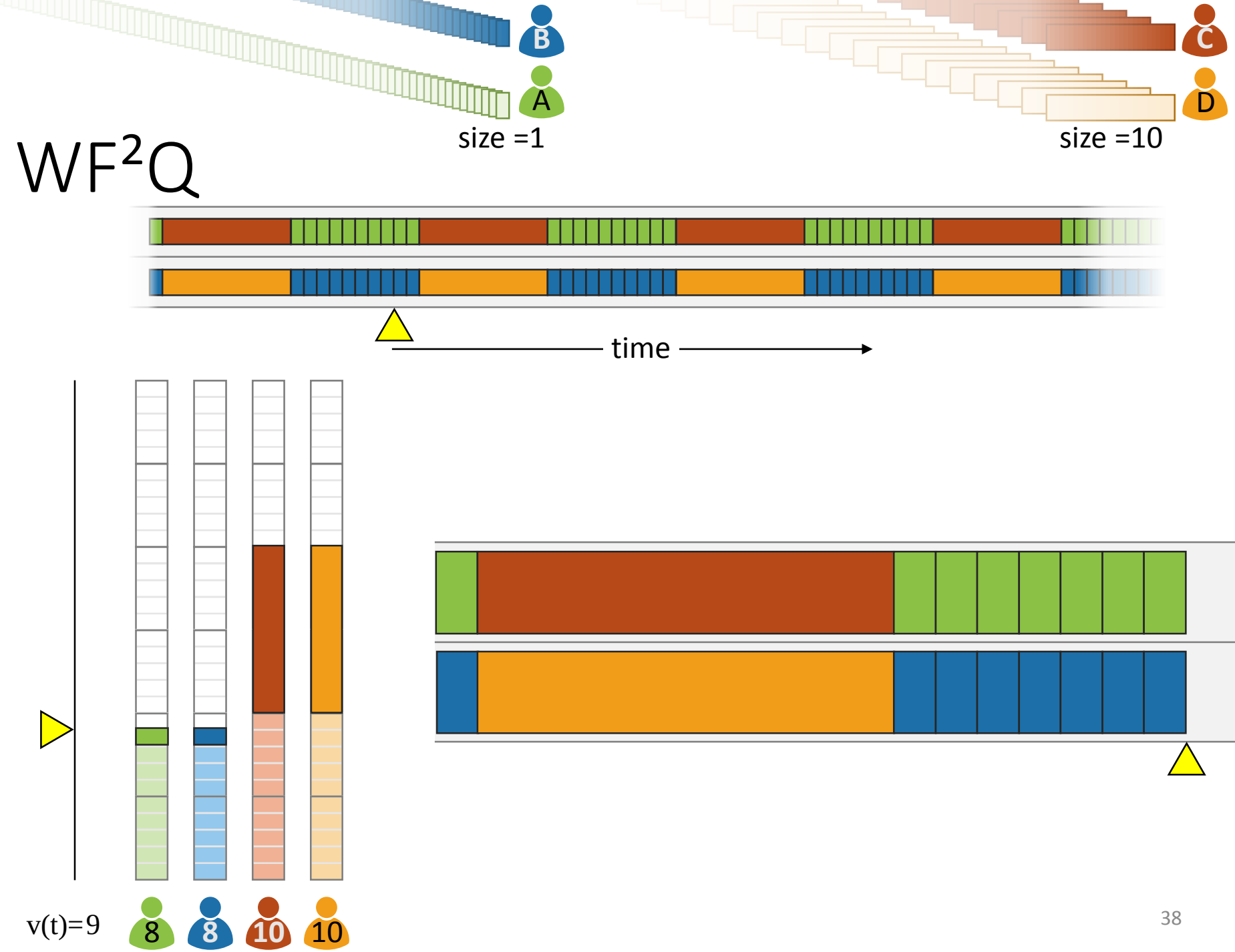




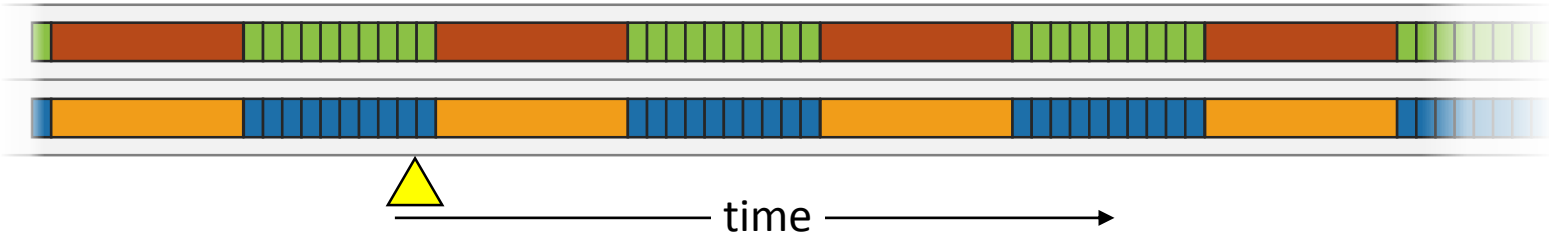
WF²Q

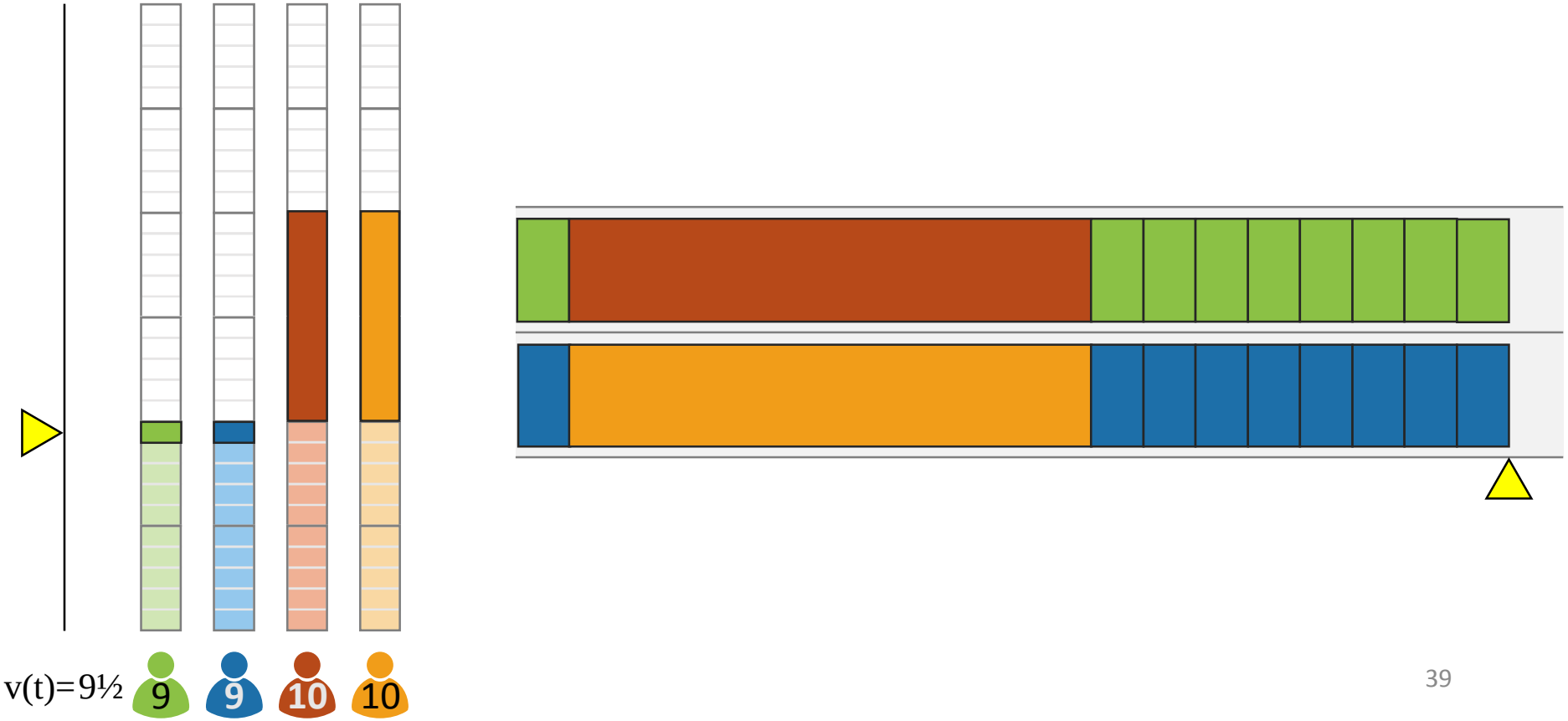




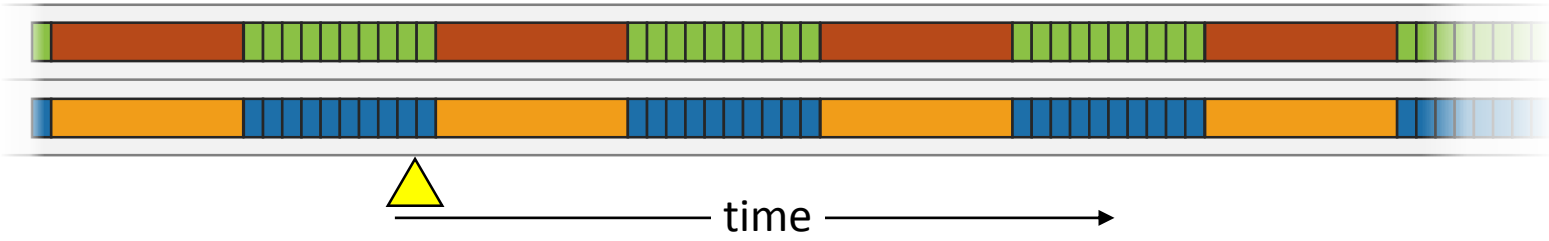

B
A
size =1
C
D
size =10
WF²Q
time
v(t)=9
8
8
10
10

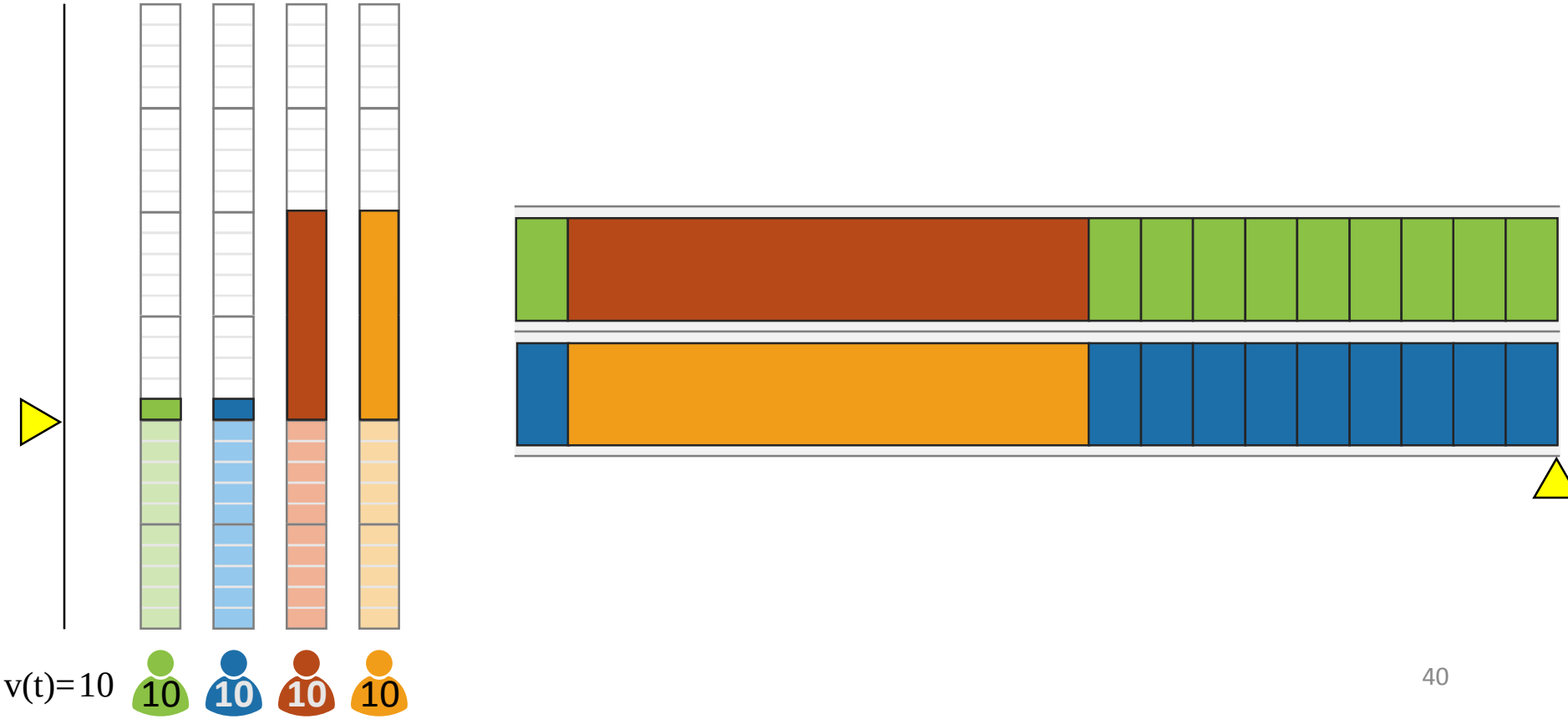
WF²Q



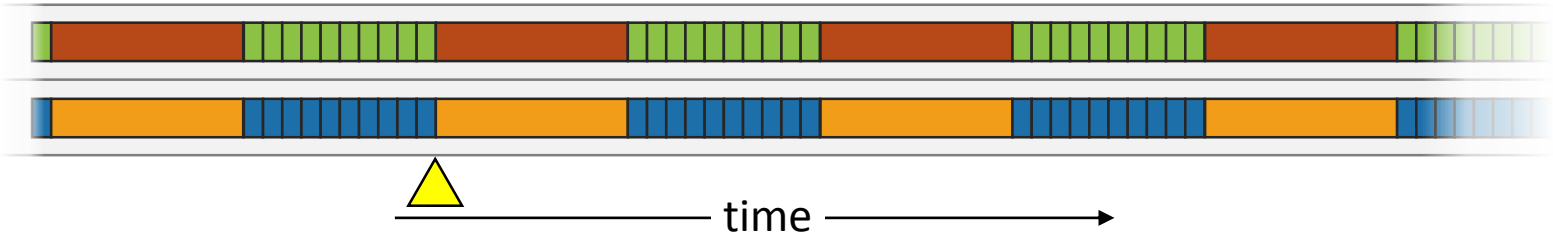


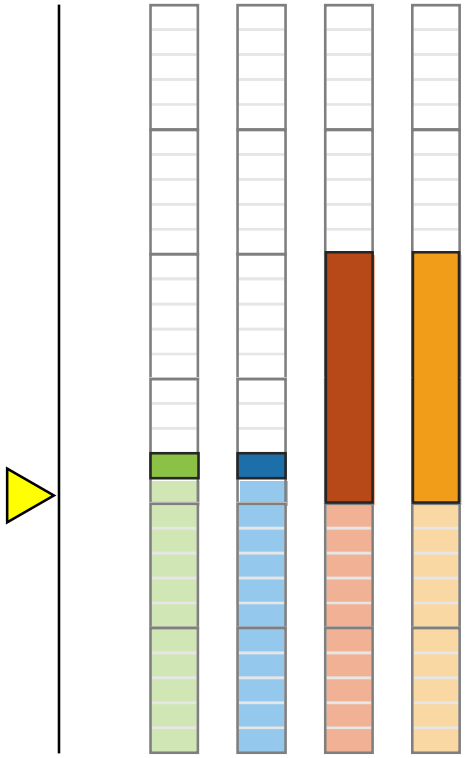
WF²Q

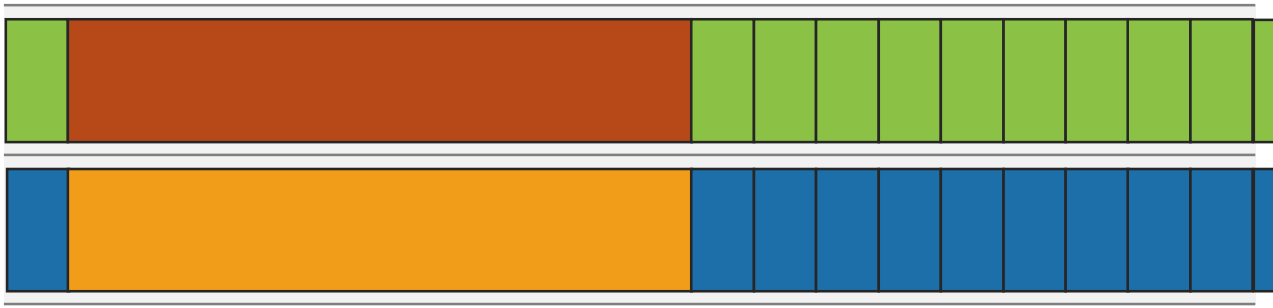




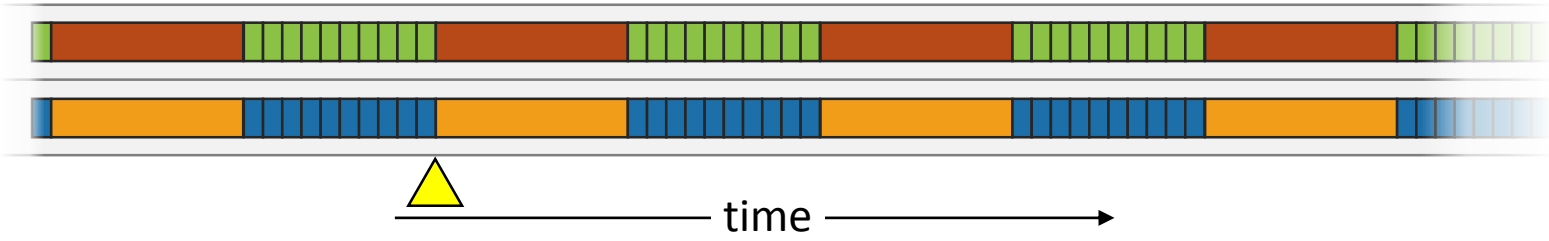
WF²Q

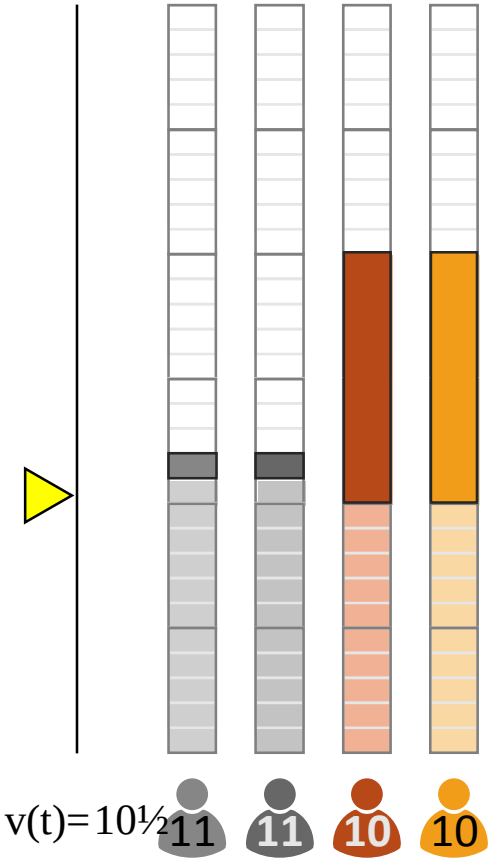


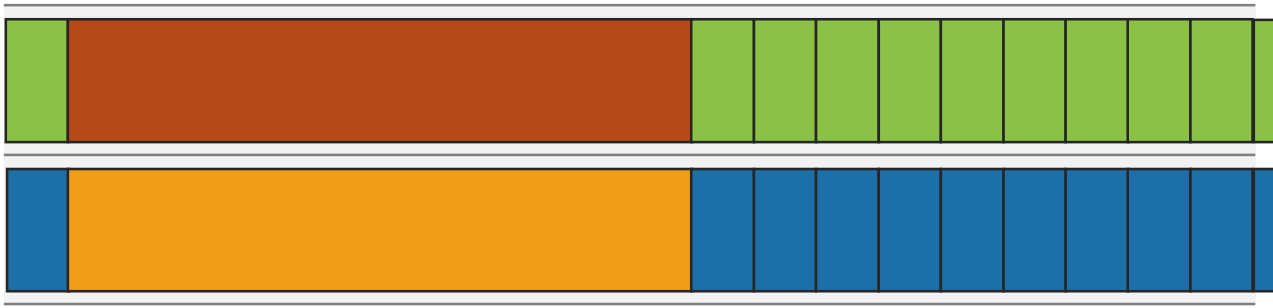


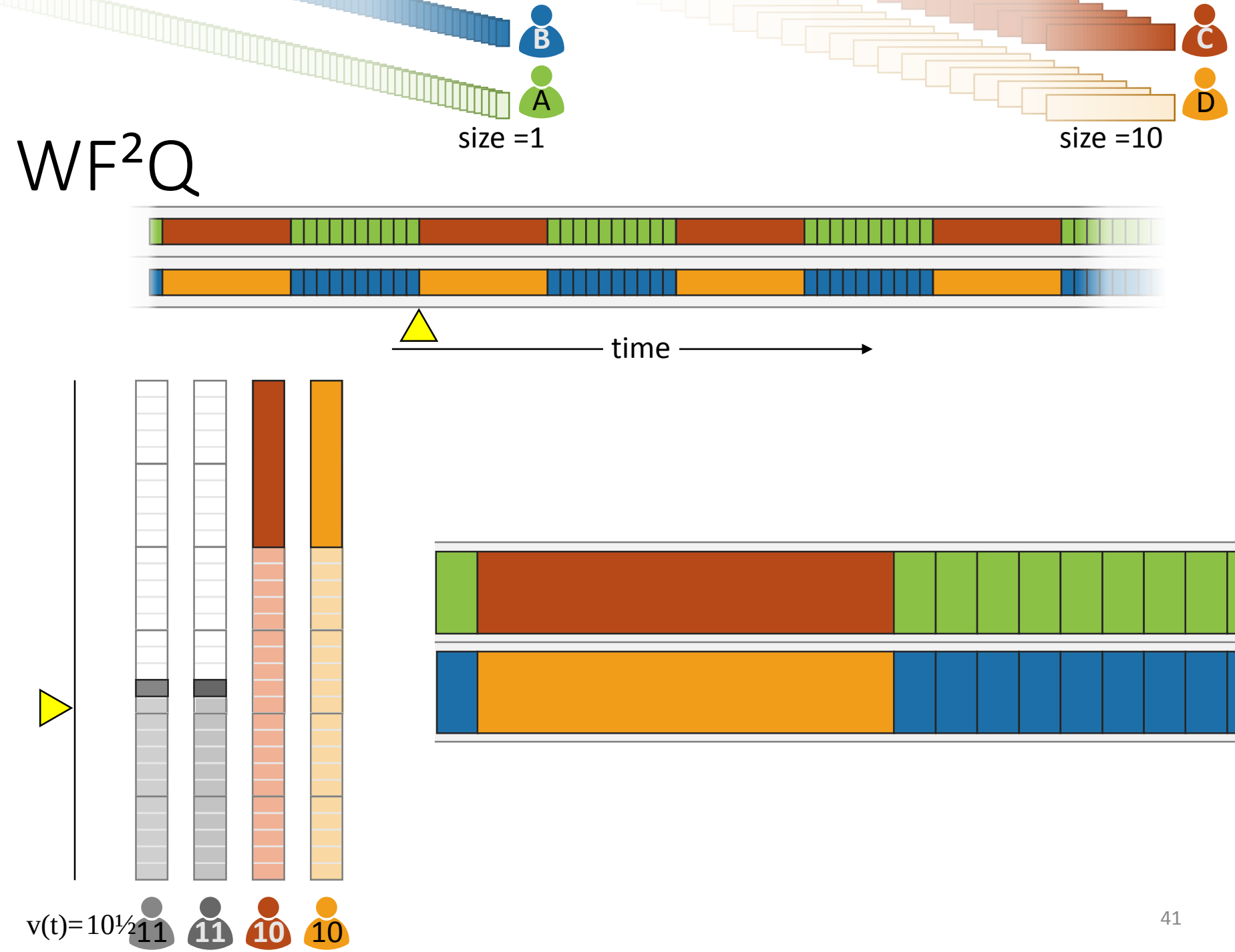


WF²Q

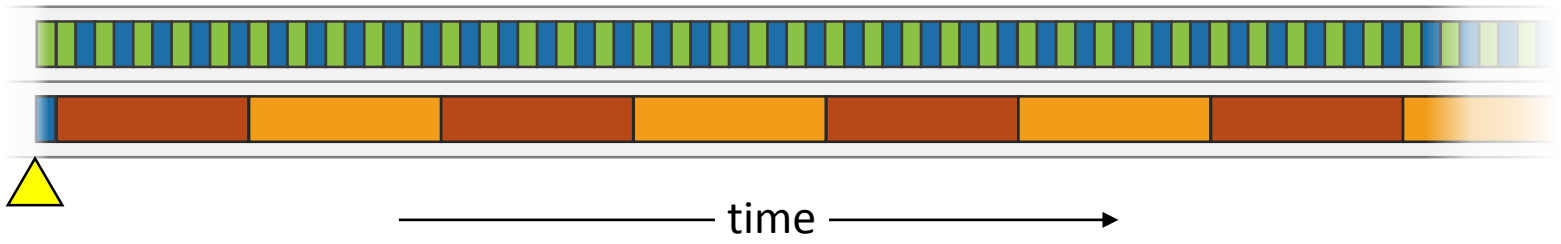


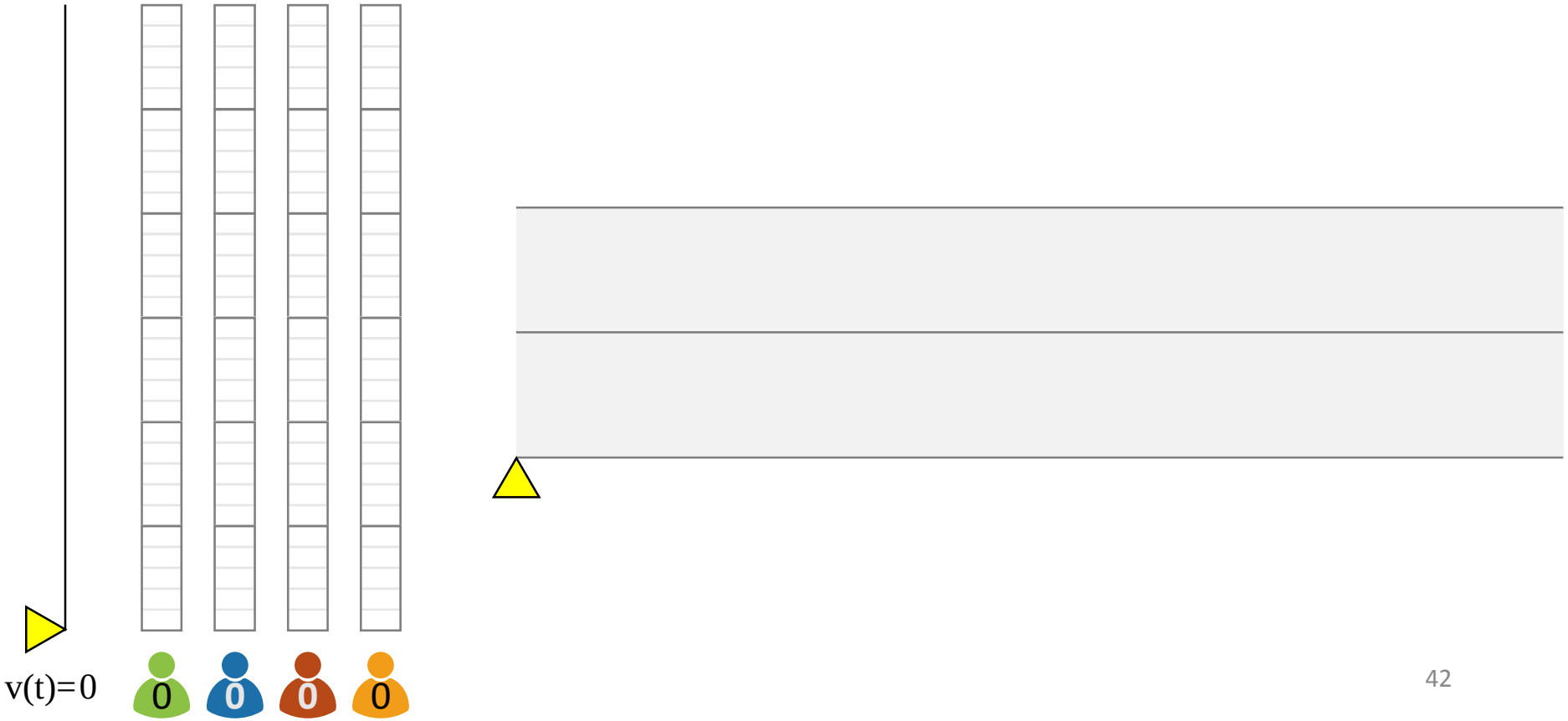




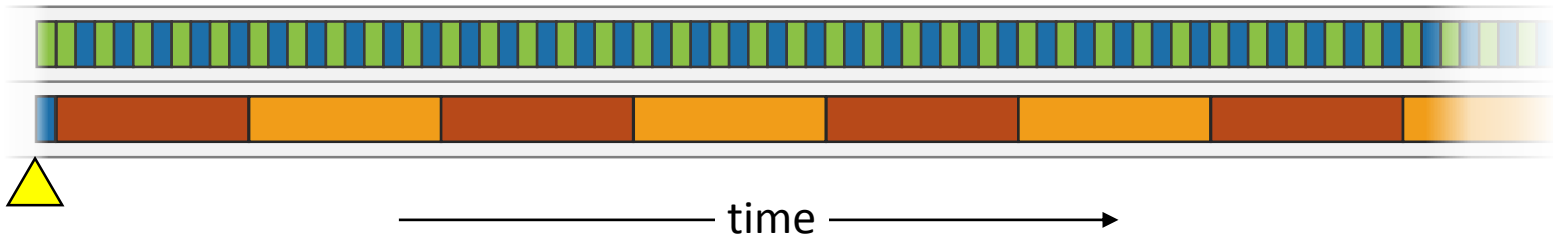

B
A
size =1
C
D
size =10
WF²Q
time
v(t)=10½
11
11
10
10

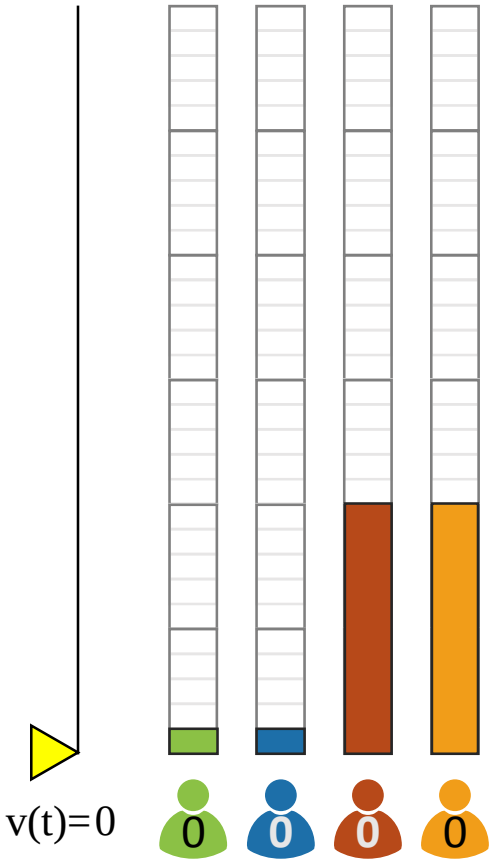
2DFQ



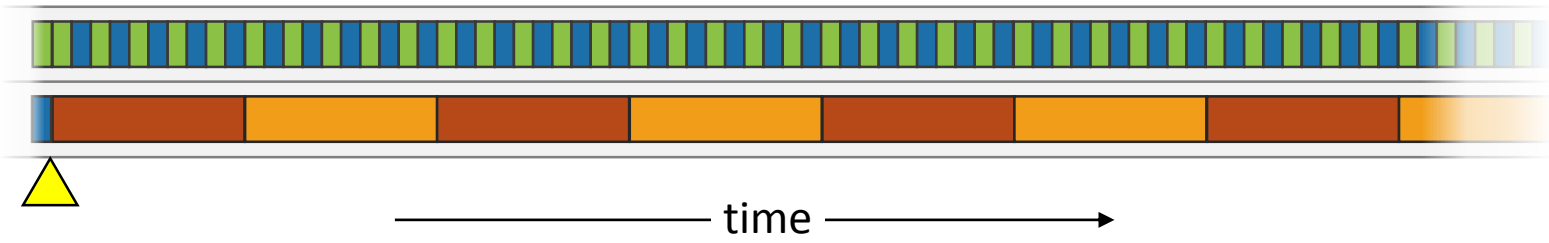


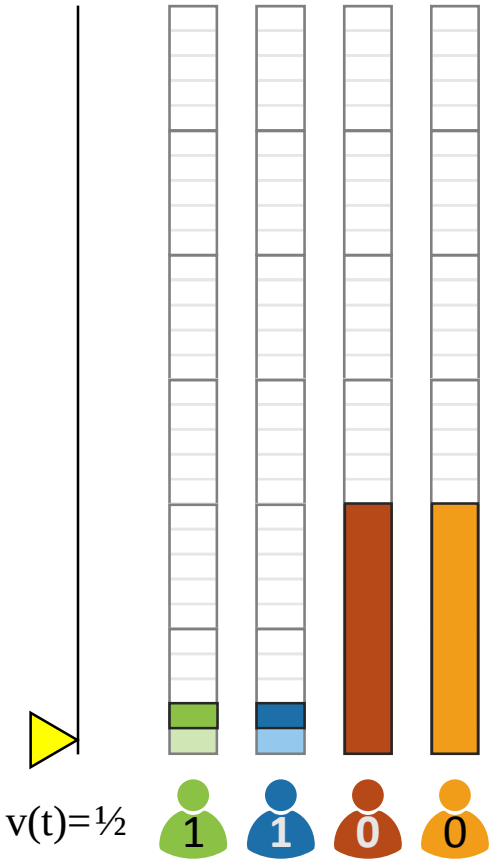
2DFQ

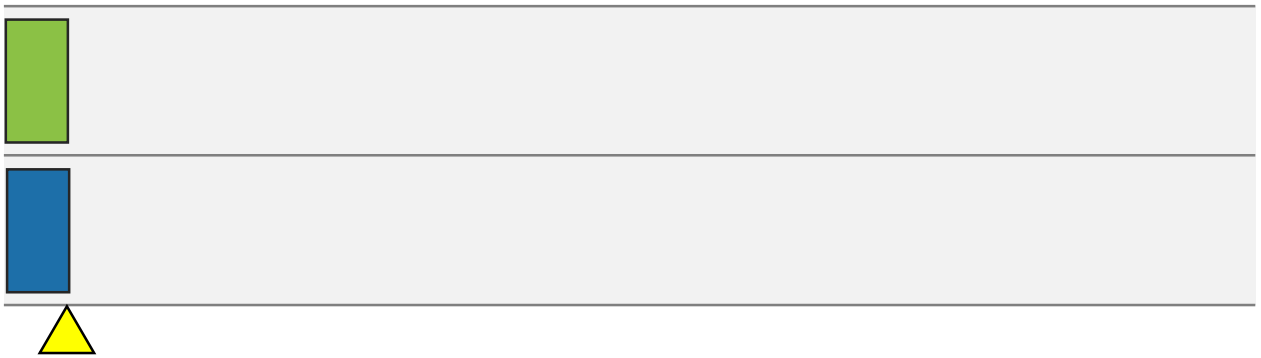




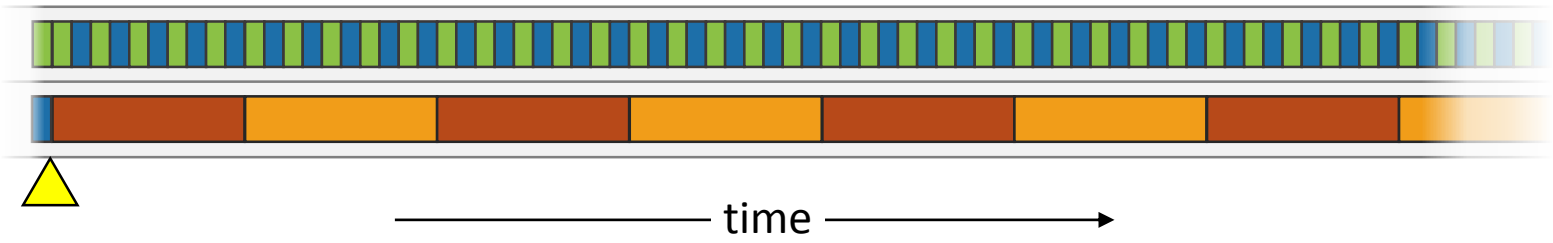
2DFQ

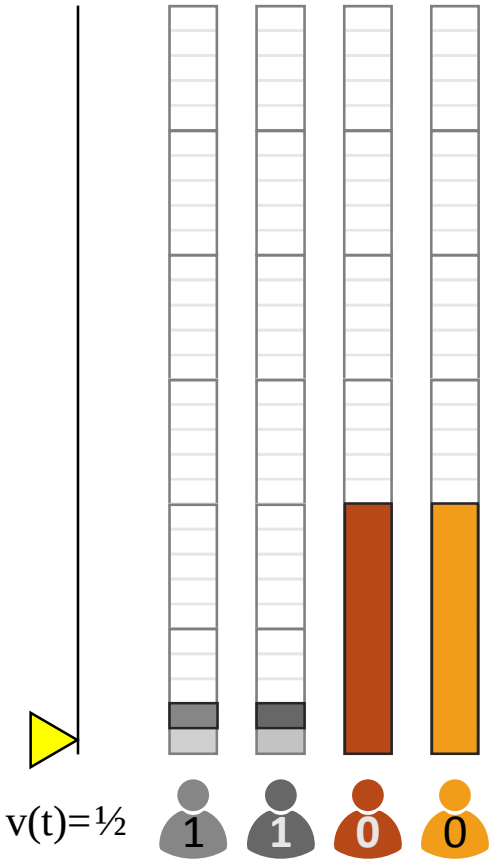


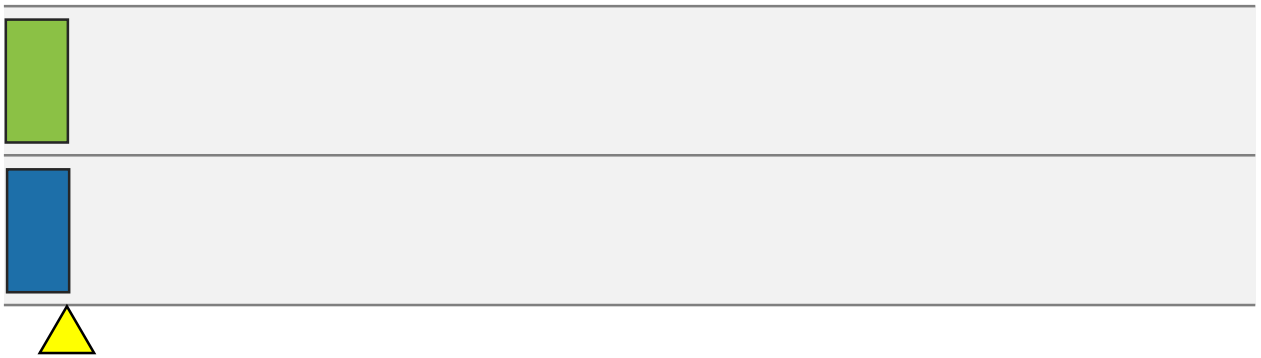




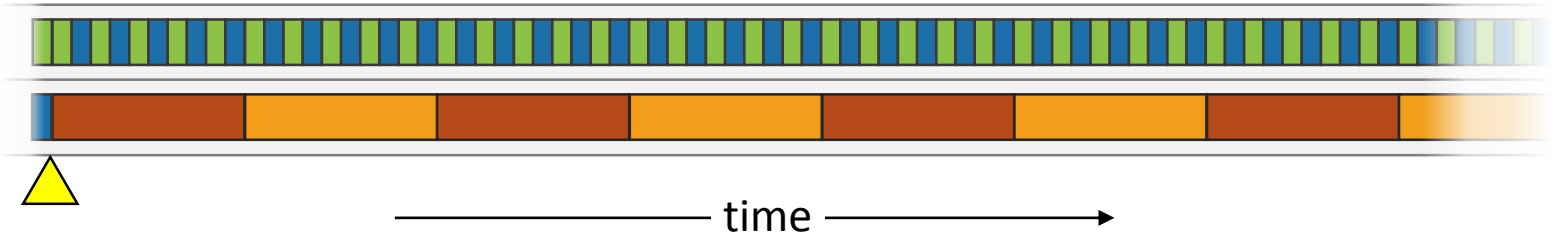
2DFQ

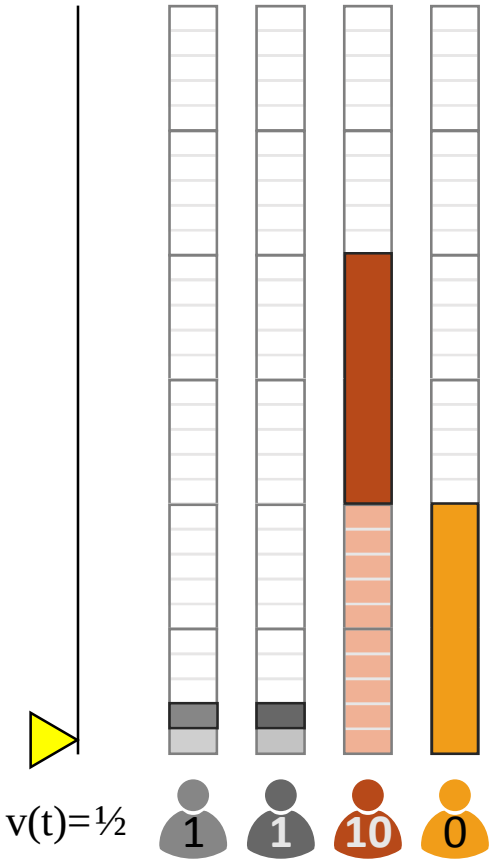


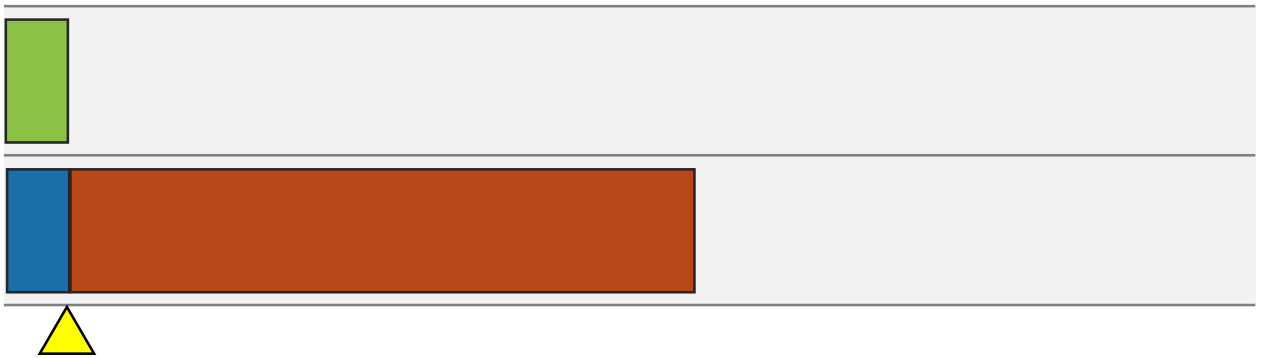




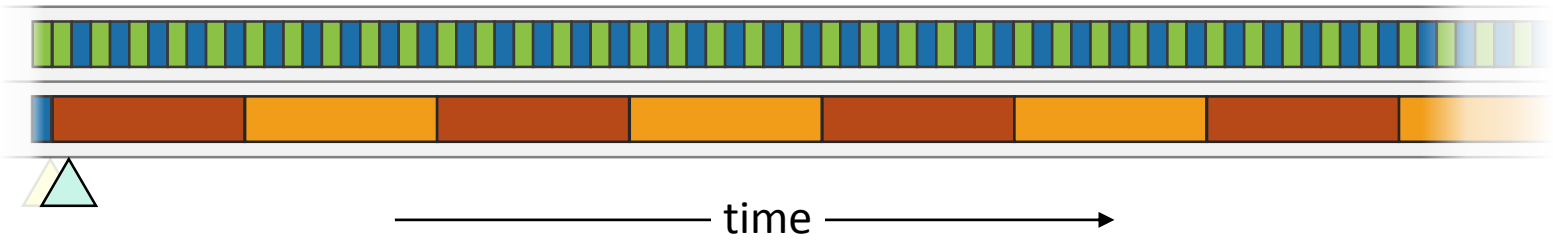
2DFQ

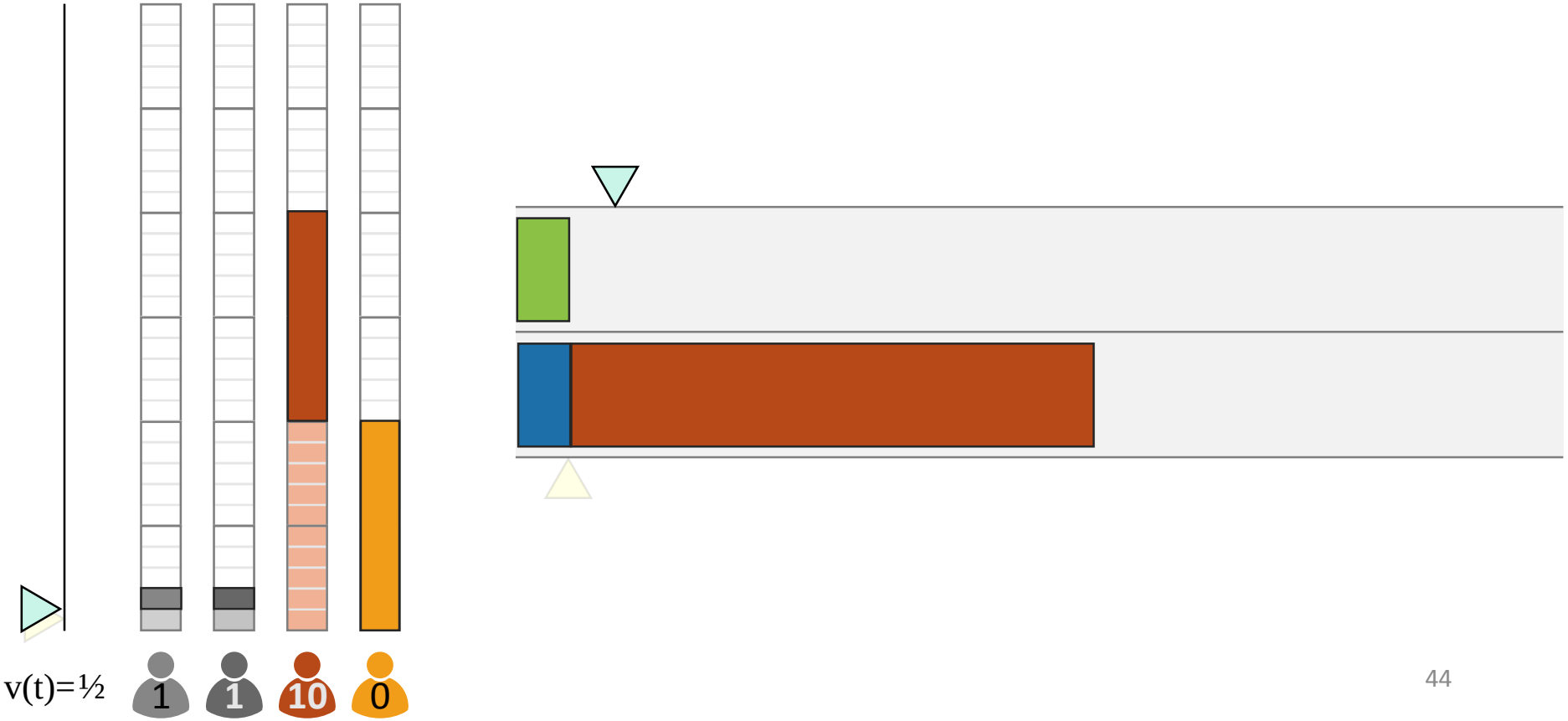




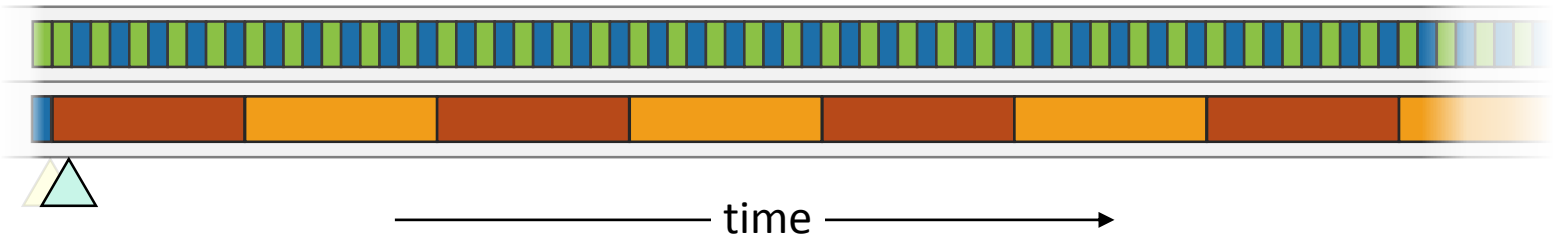


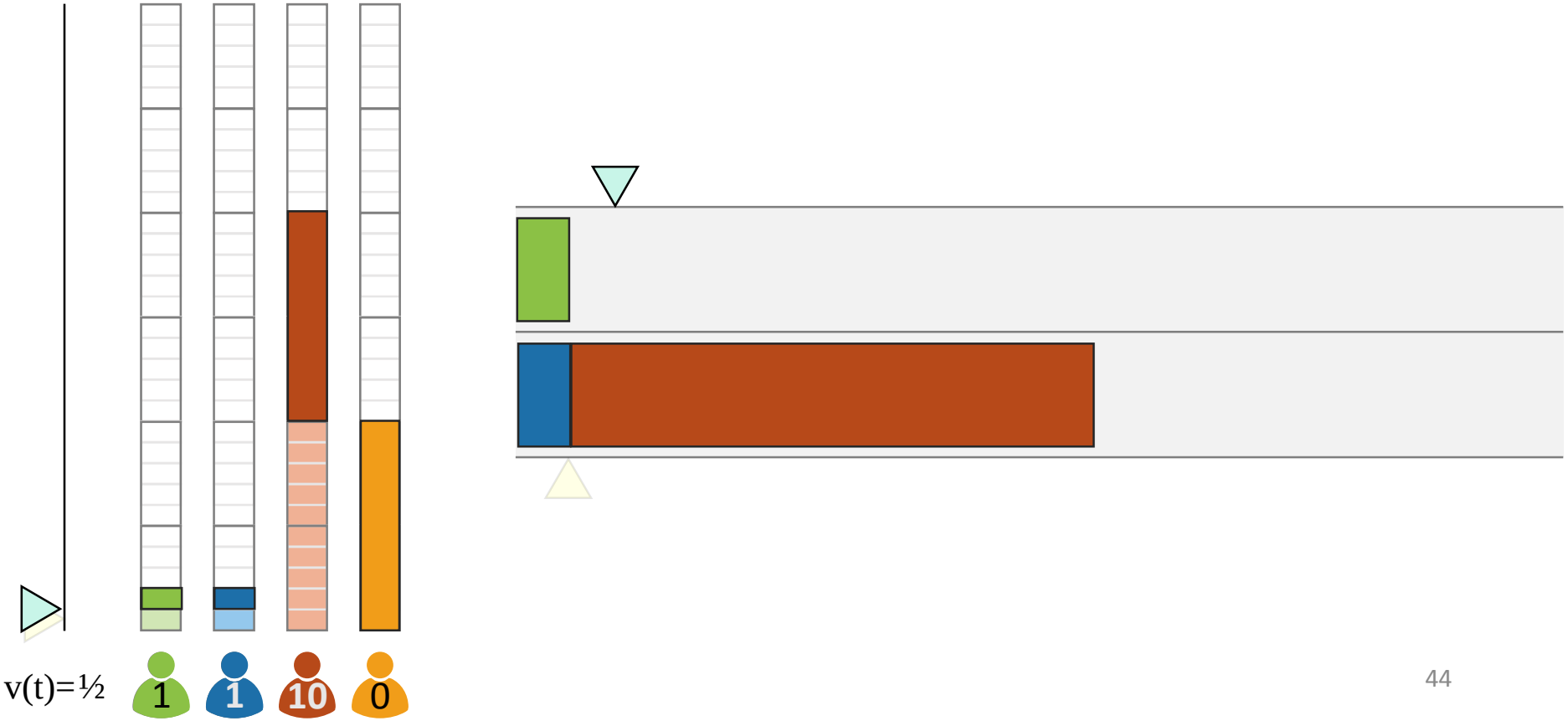
2DFQ



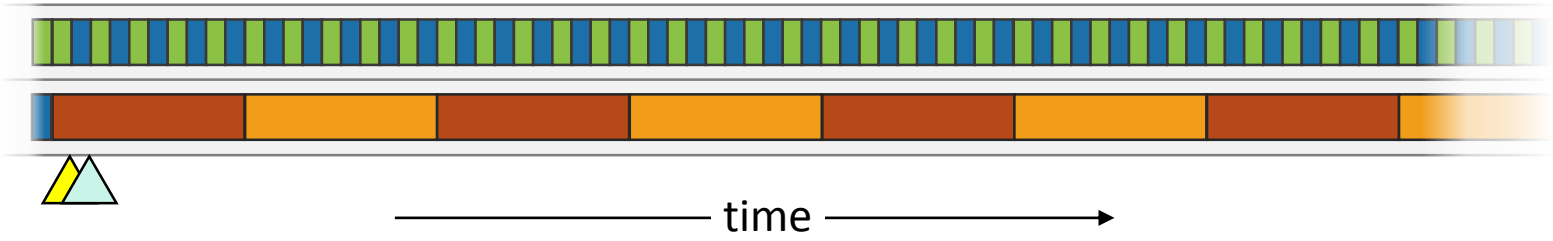


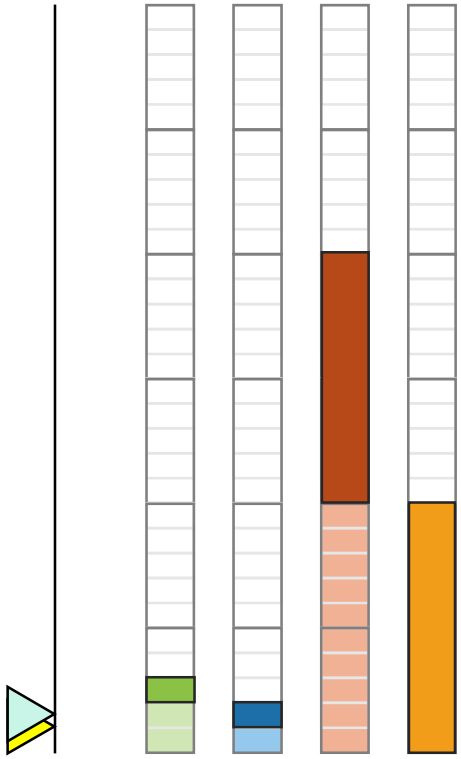
2DFQ

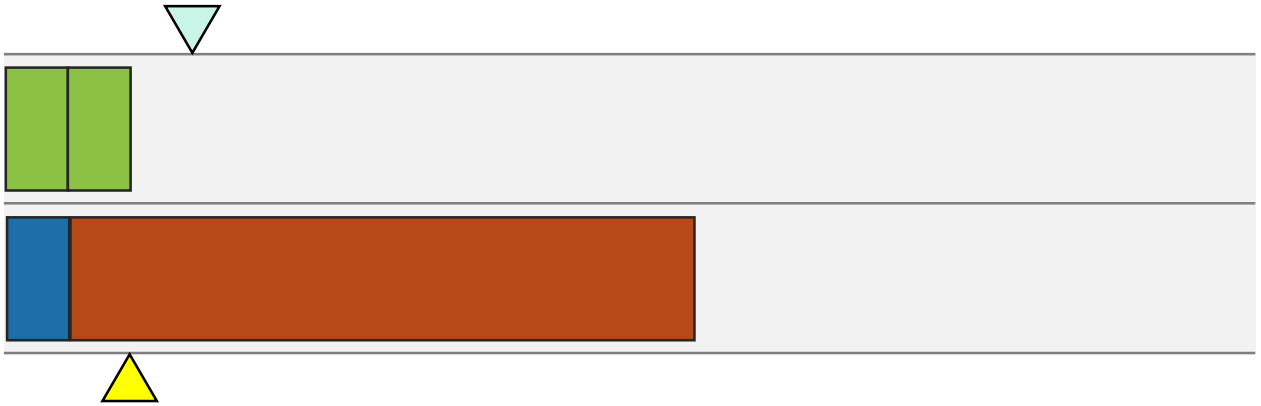




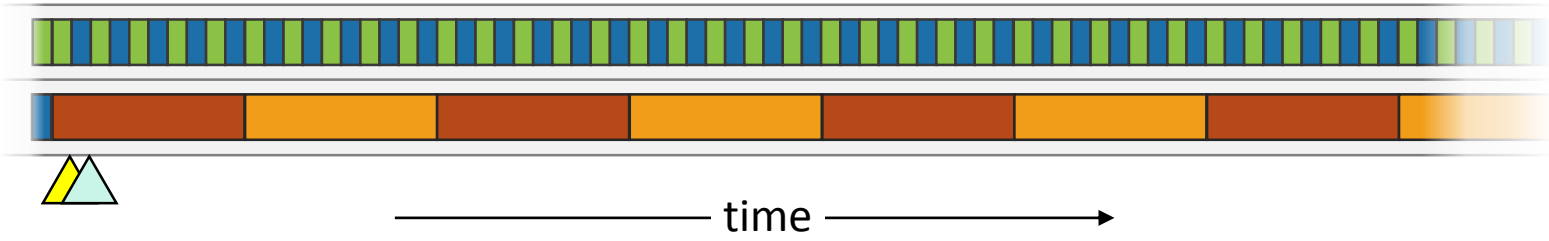
2DFQ

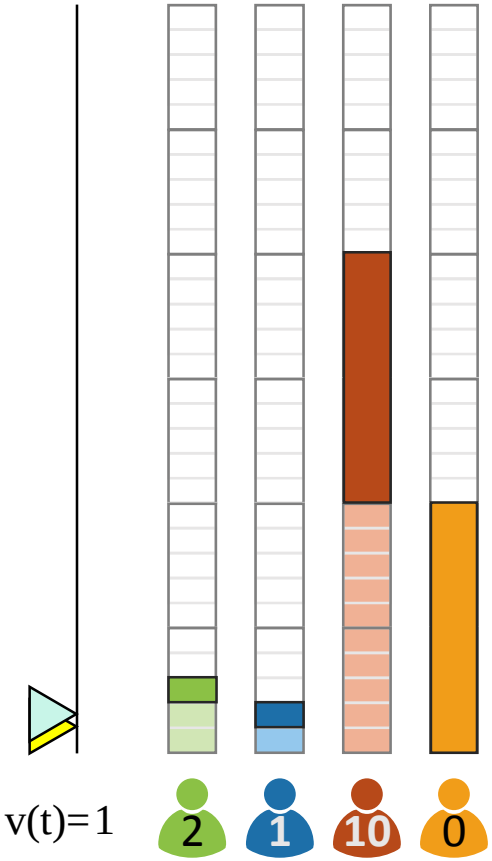


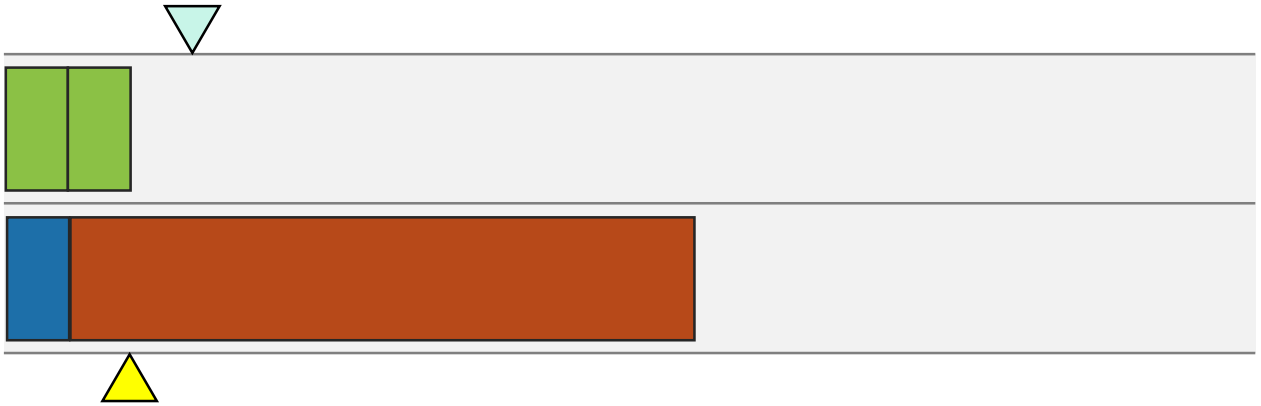




2DFQ

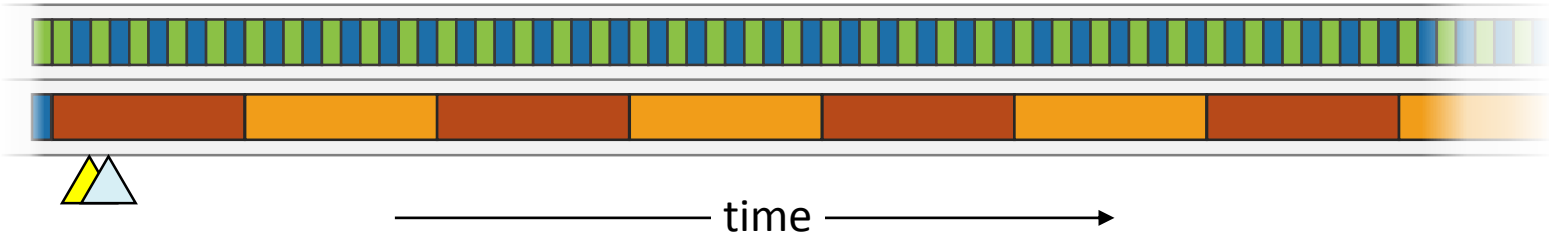


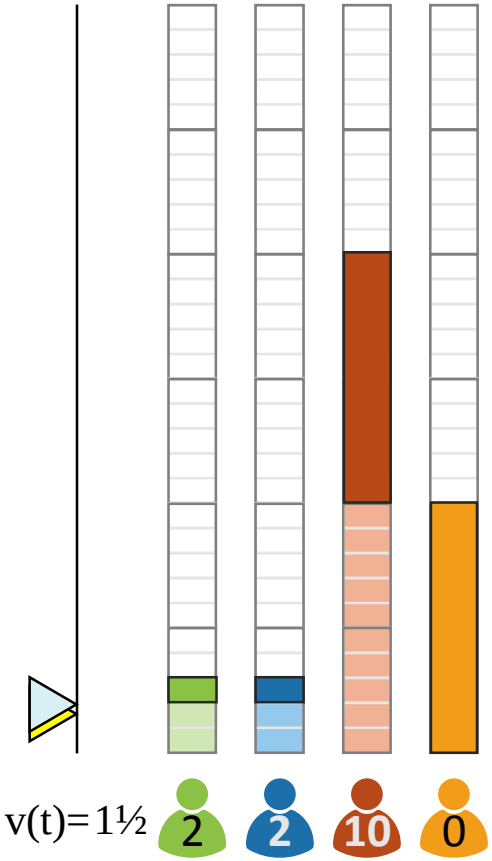


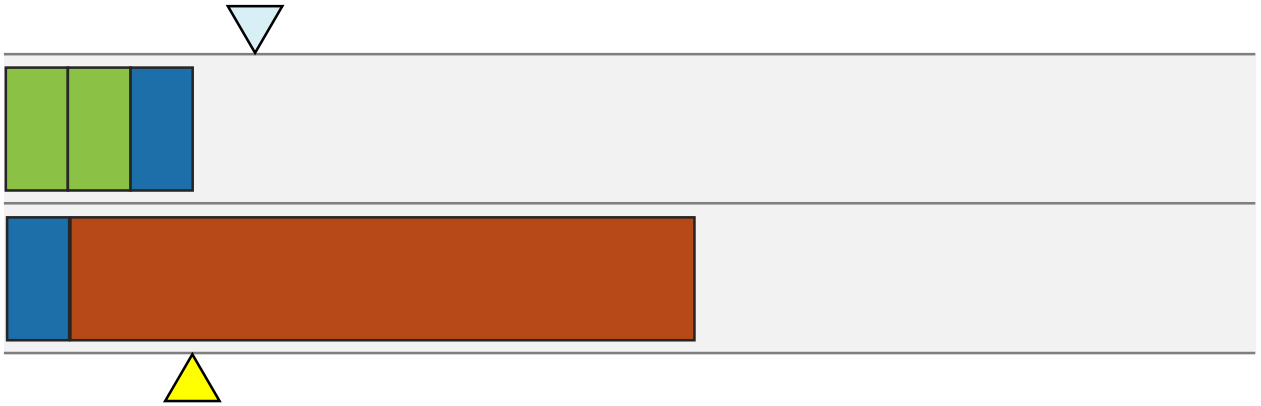


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

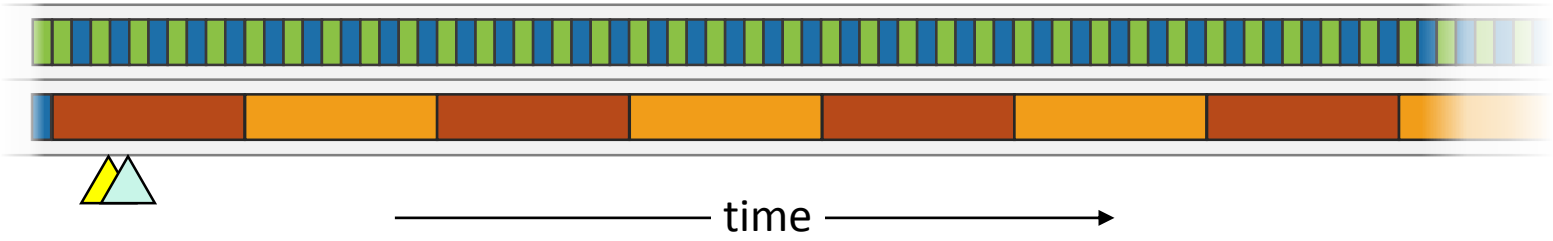


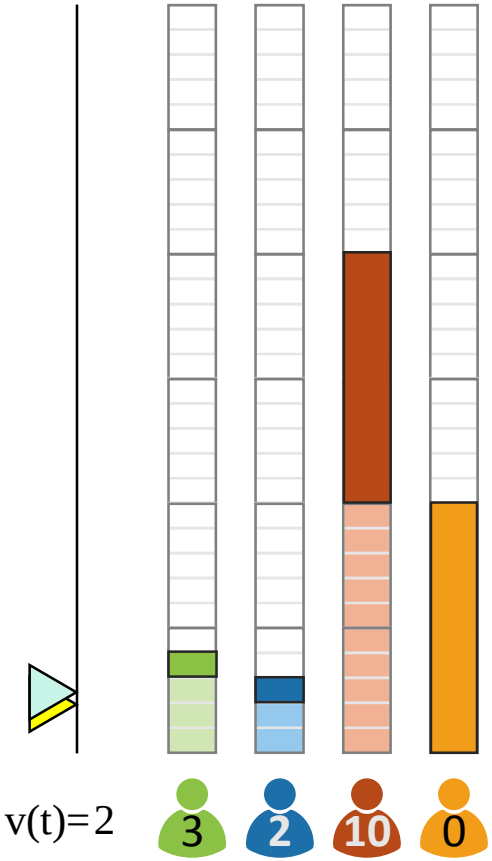


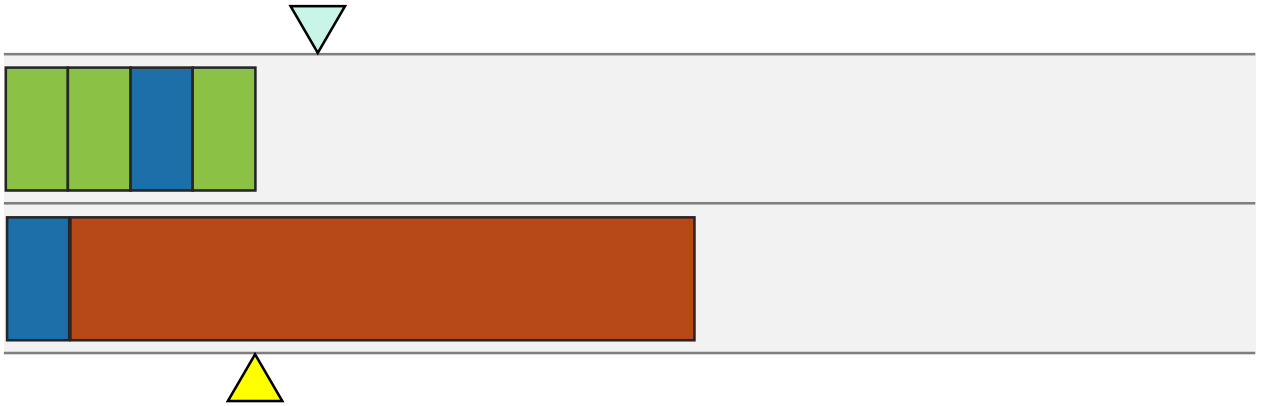


$$\text{Eligible}\left(r^j, i\right) = \text{Start}\left(r^j\right) - \frac{i}{n} \times \text{size}\left(r^j\right)$$

2DFQ

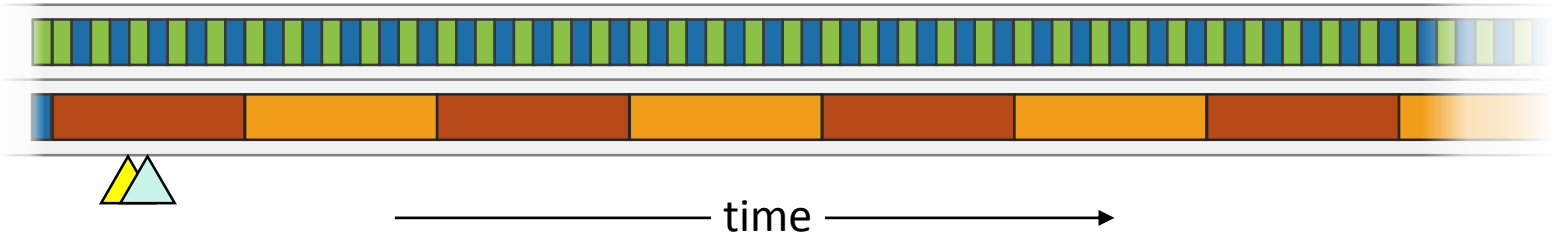


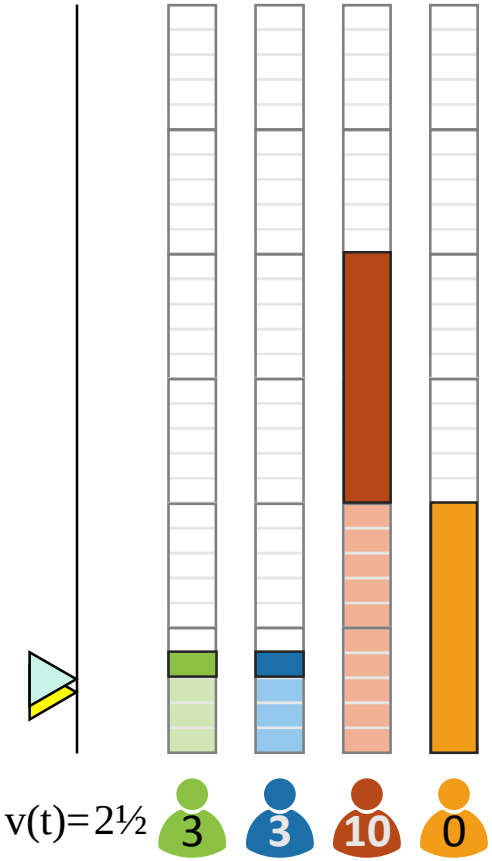


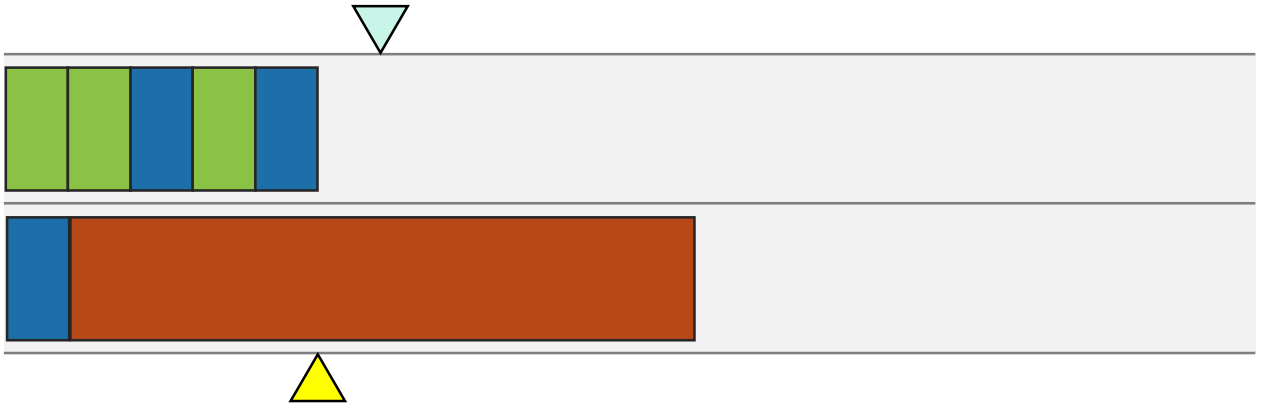


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

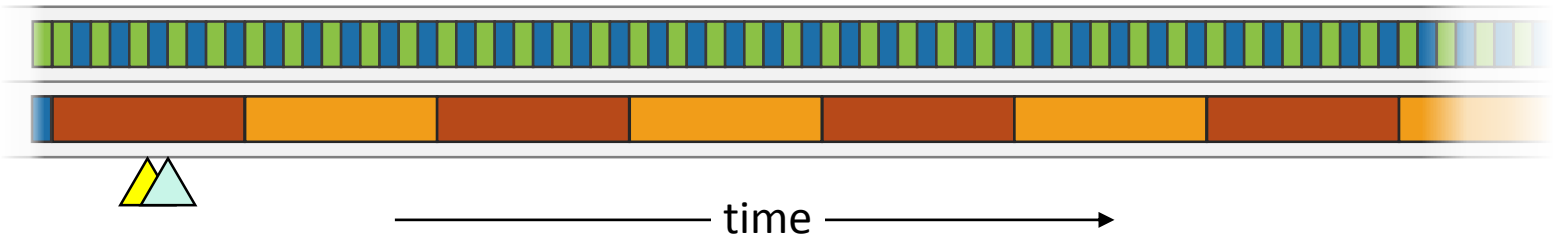


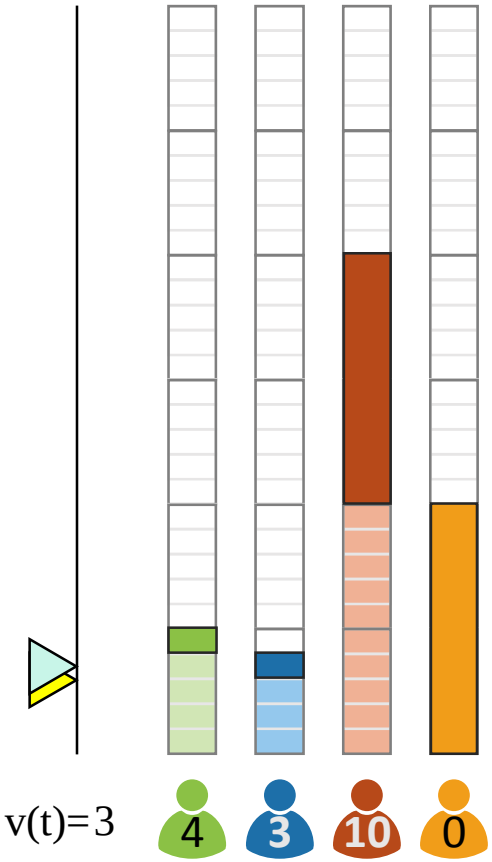


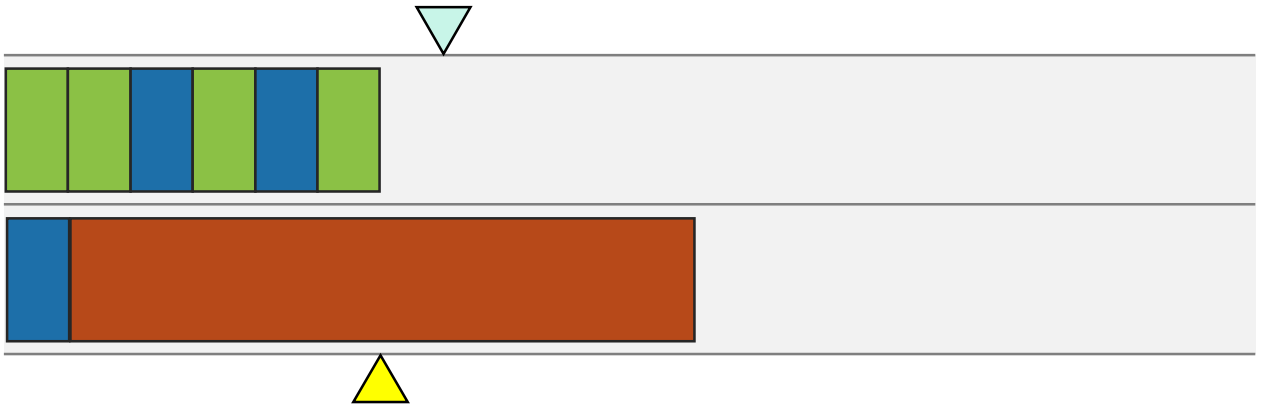


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

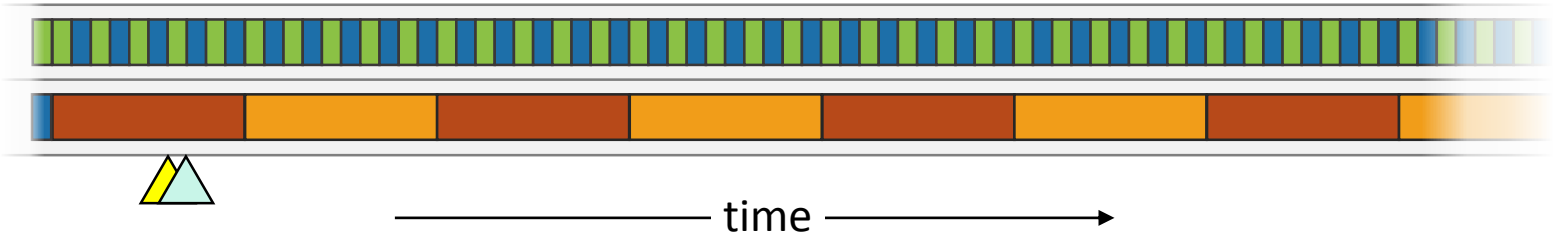


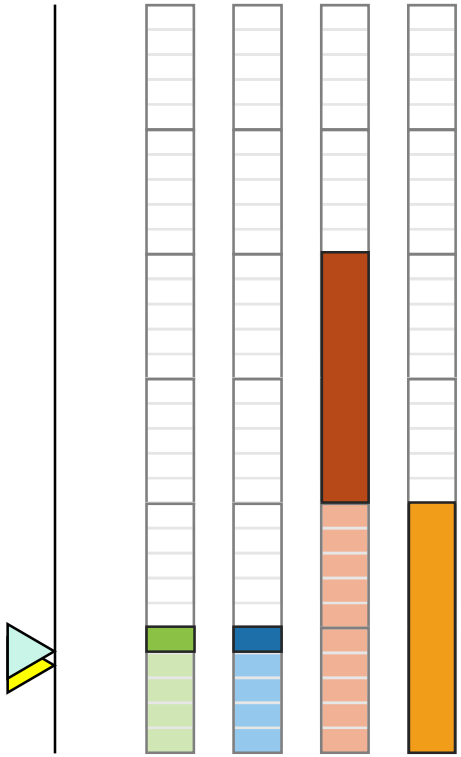


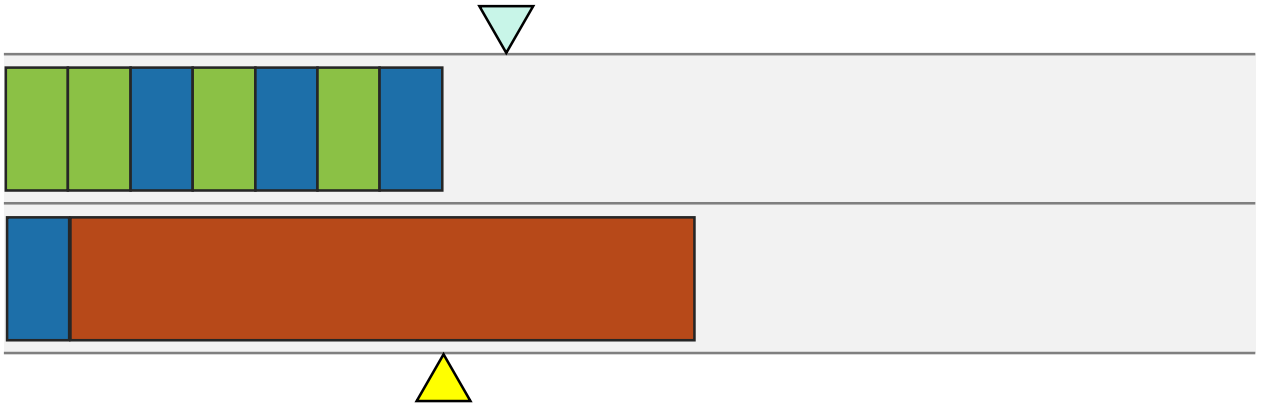


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

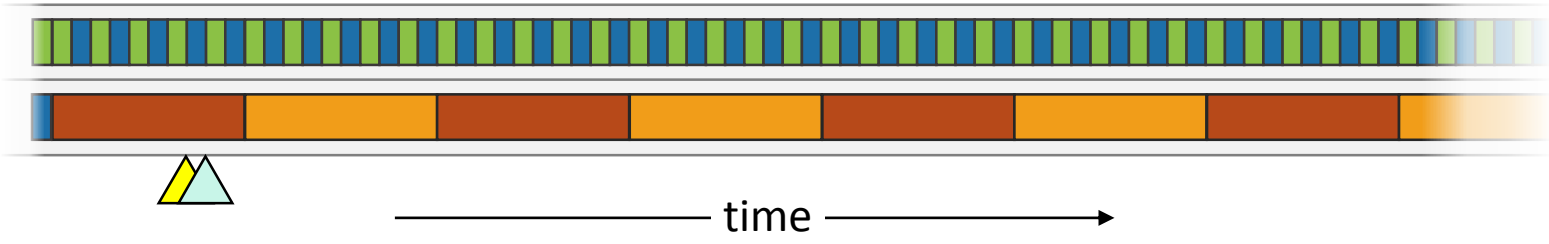


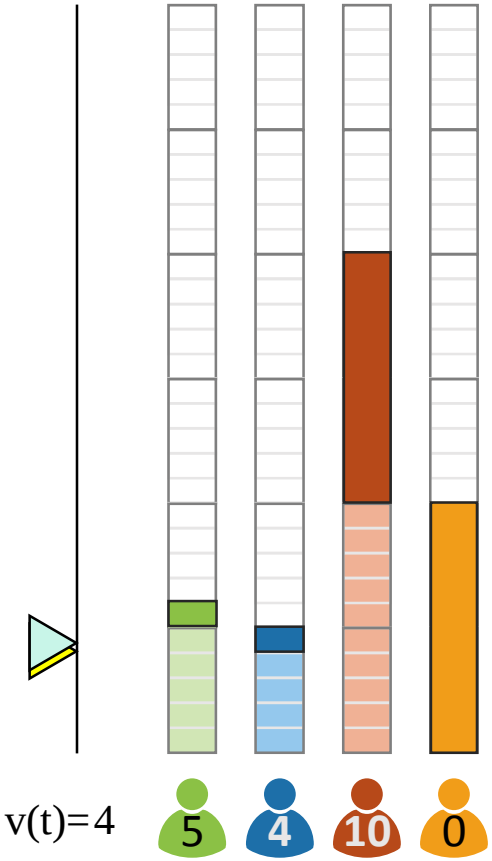


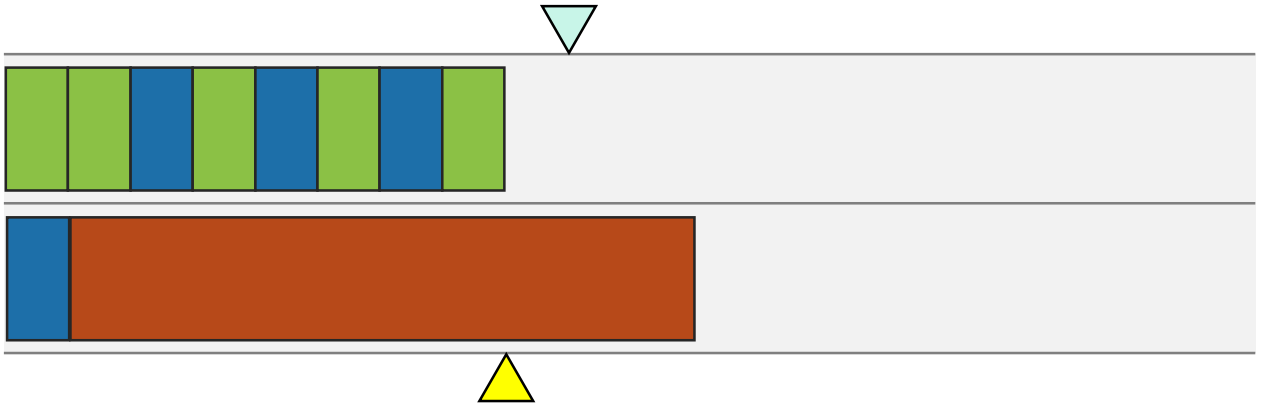


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

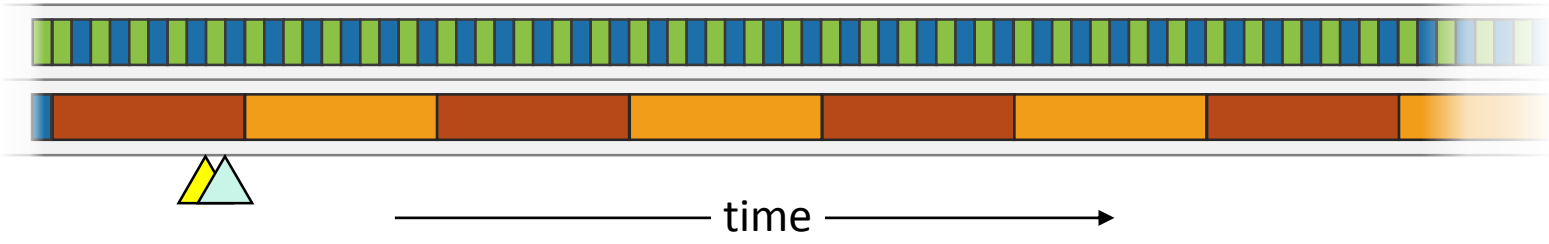


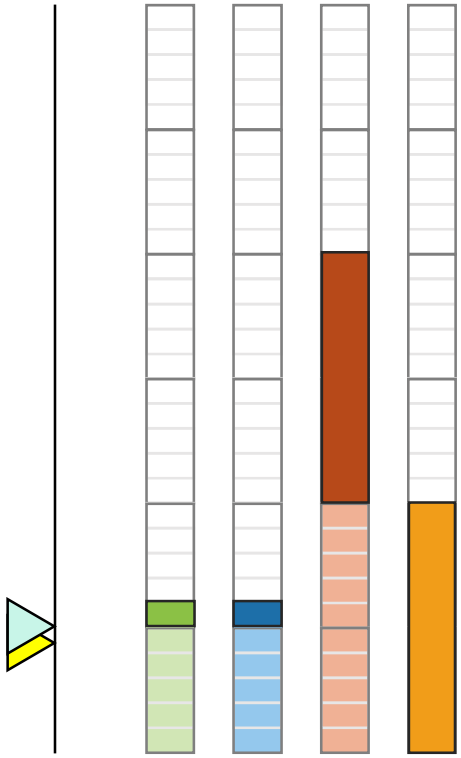


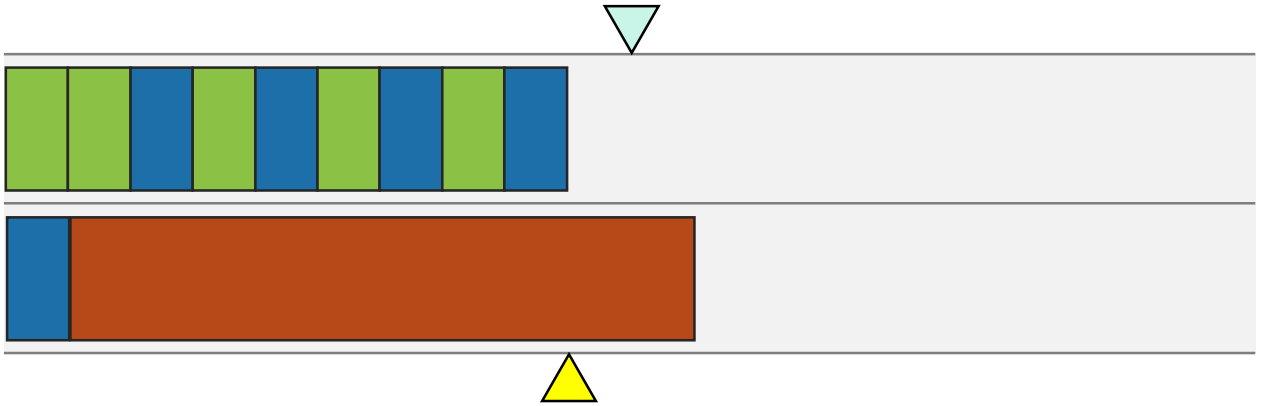


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

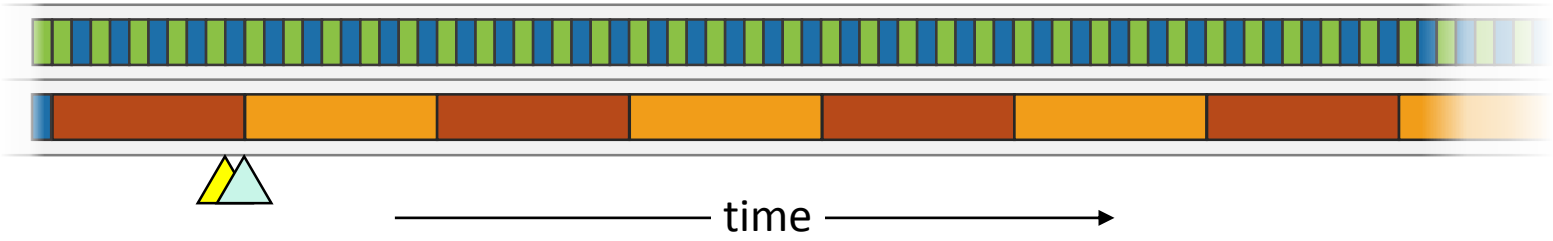


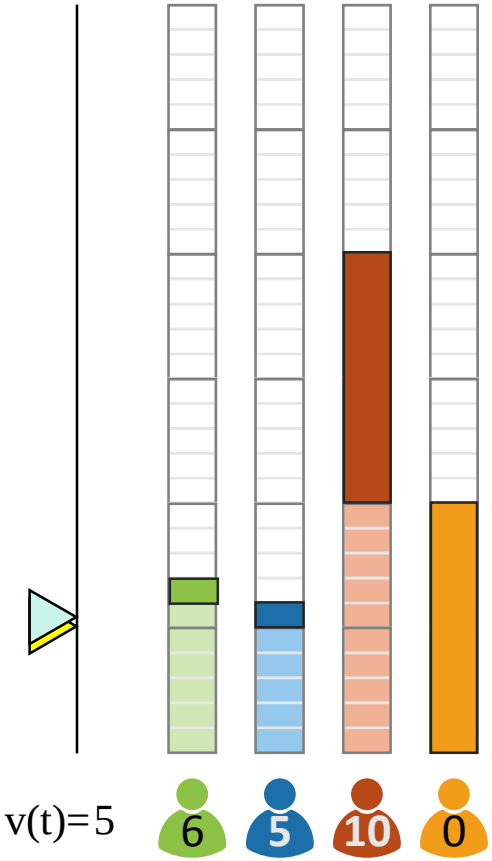


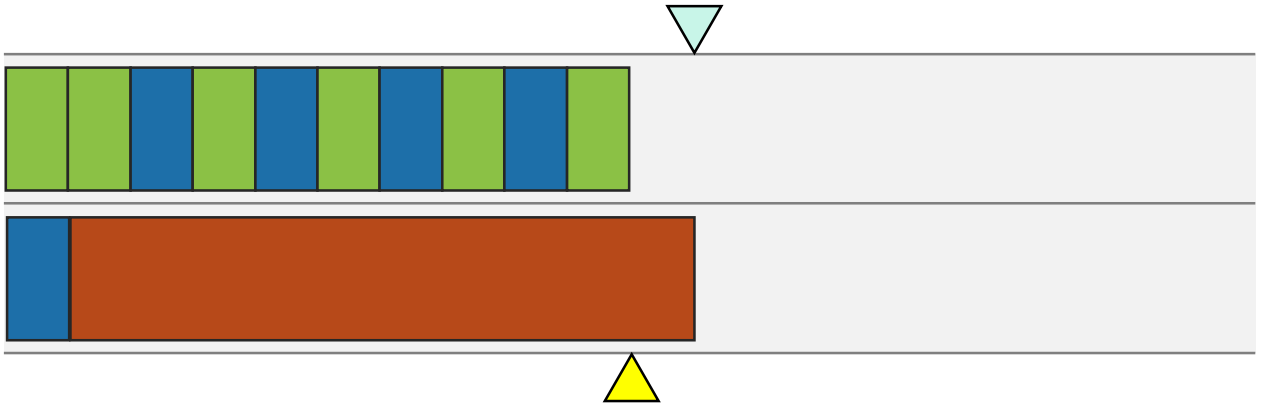


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

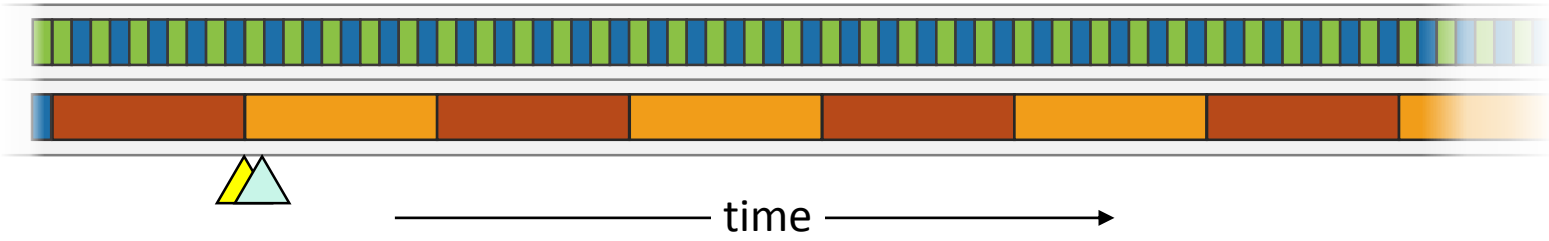


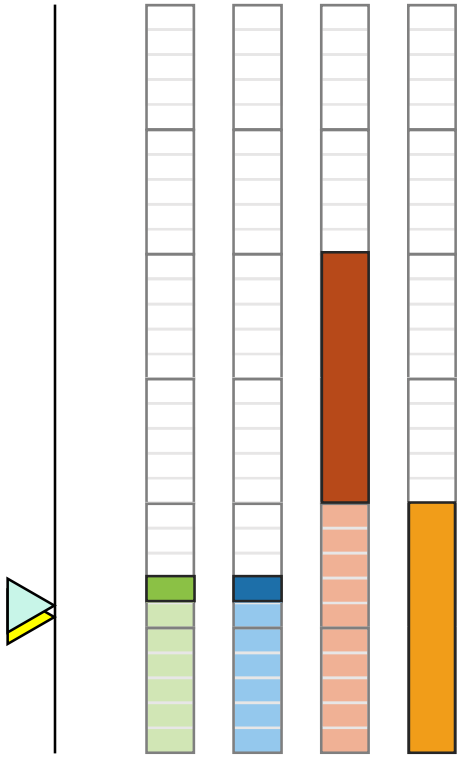


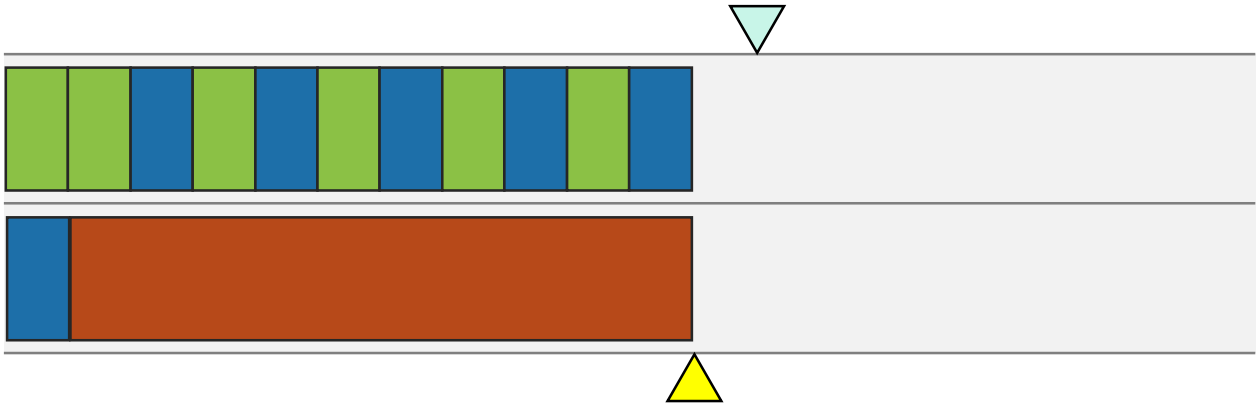


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

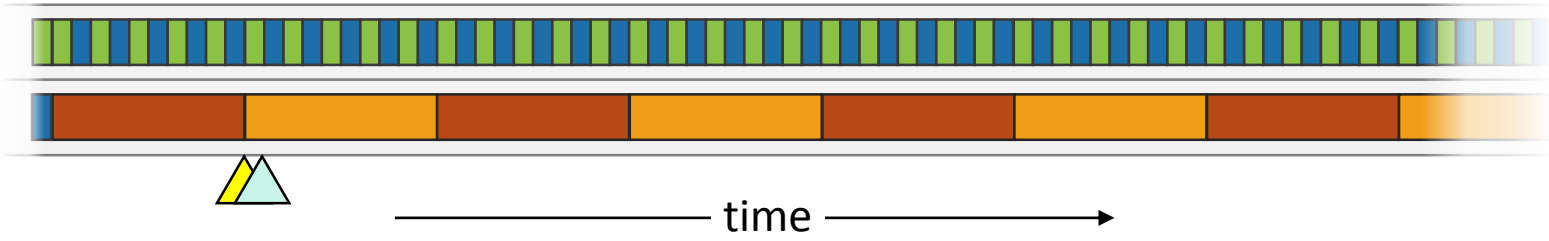


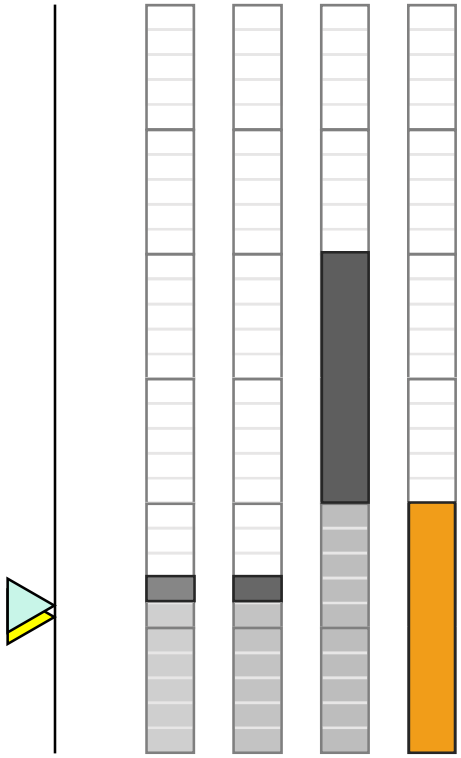


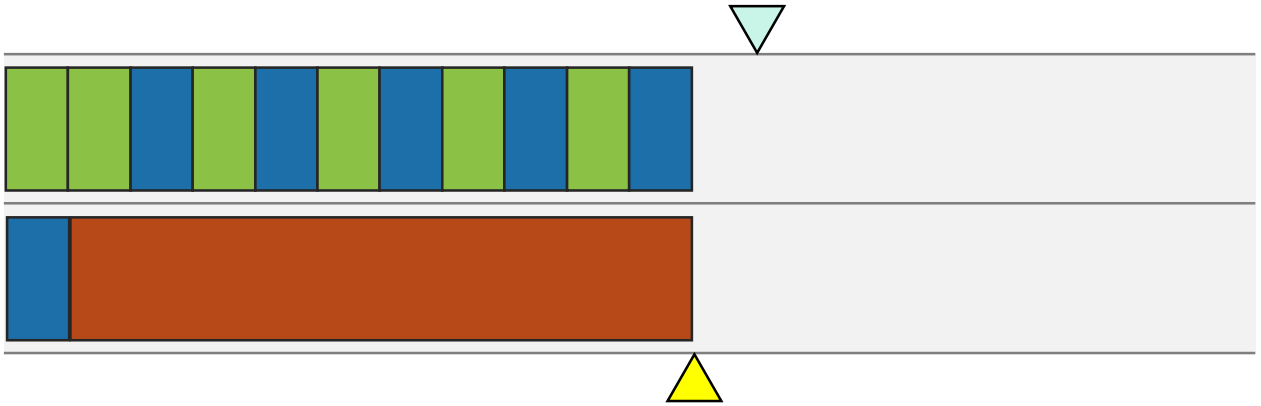


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

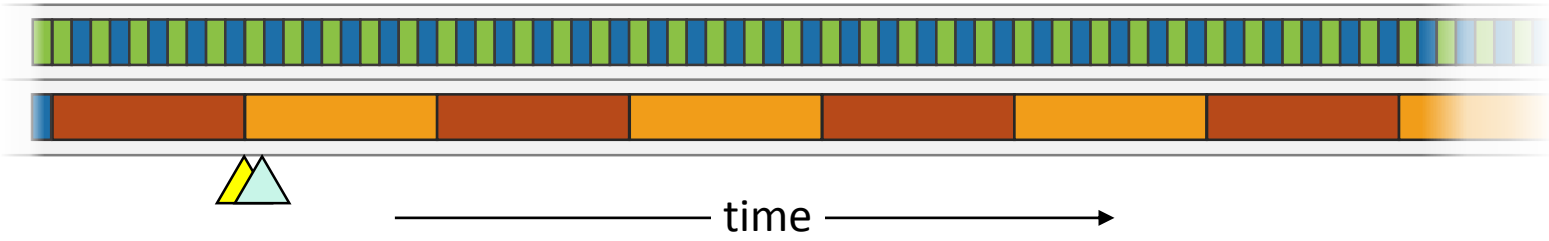


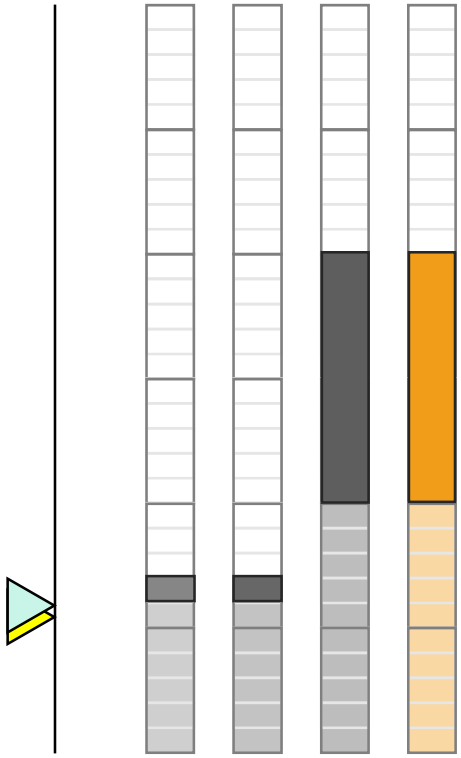


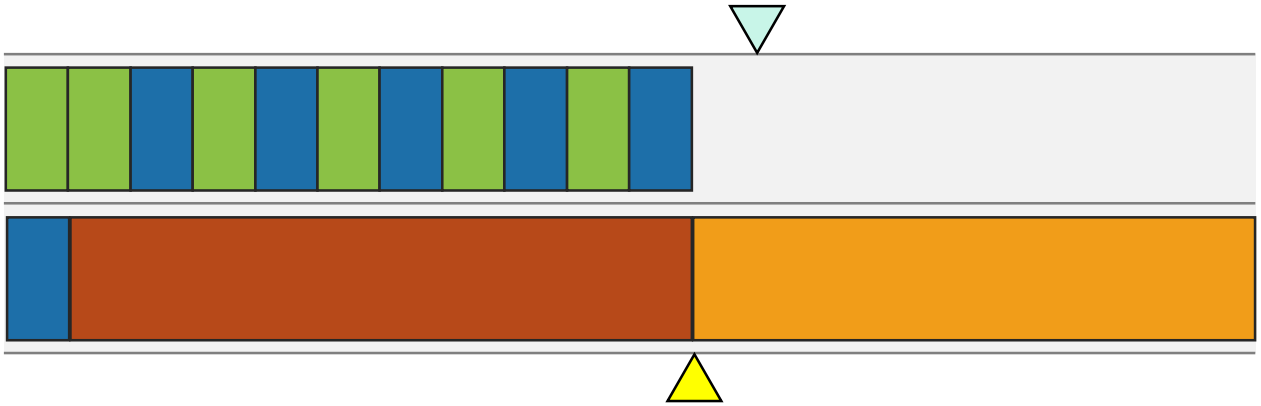


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

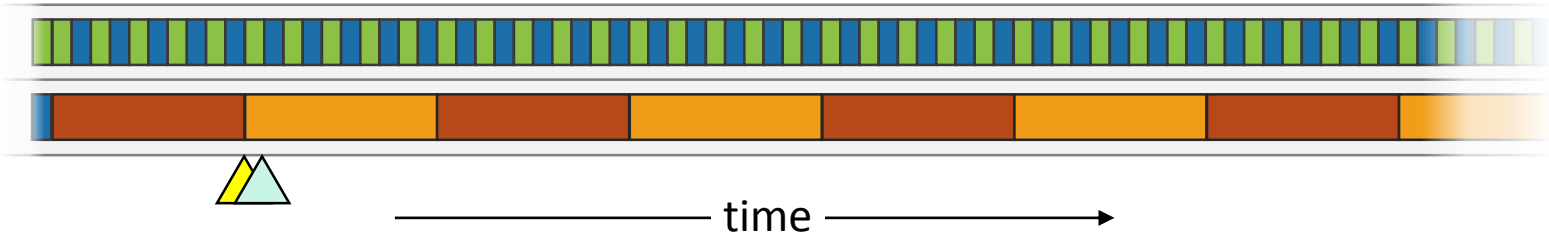


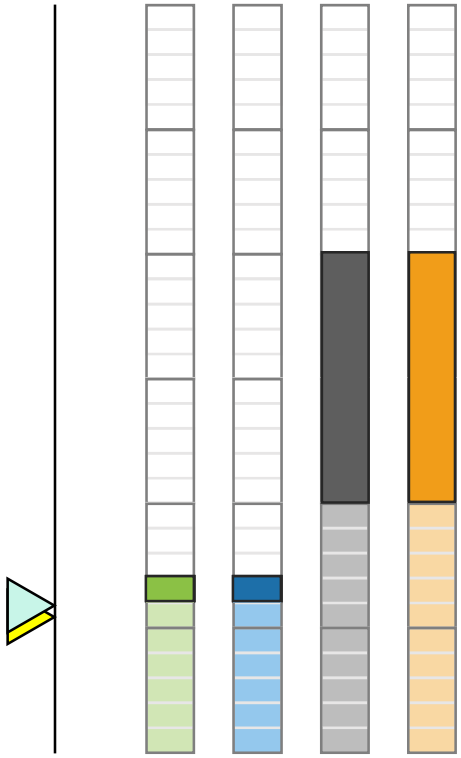


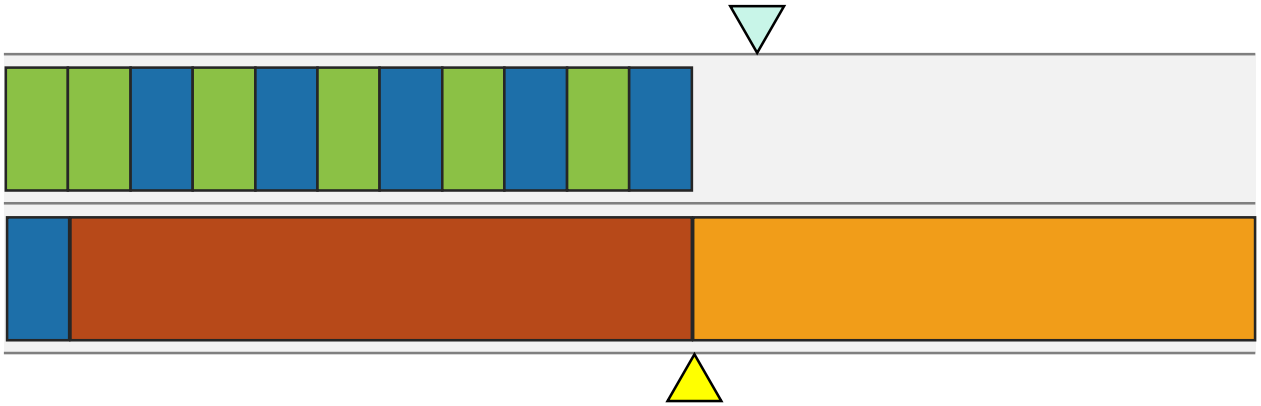


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

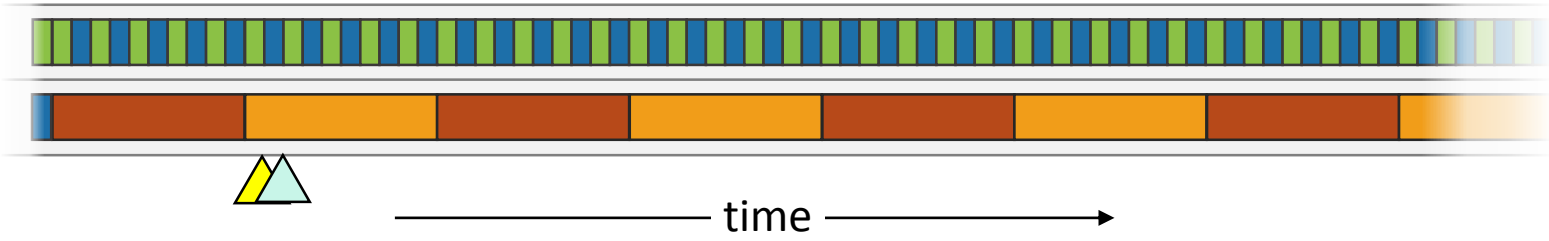


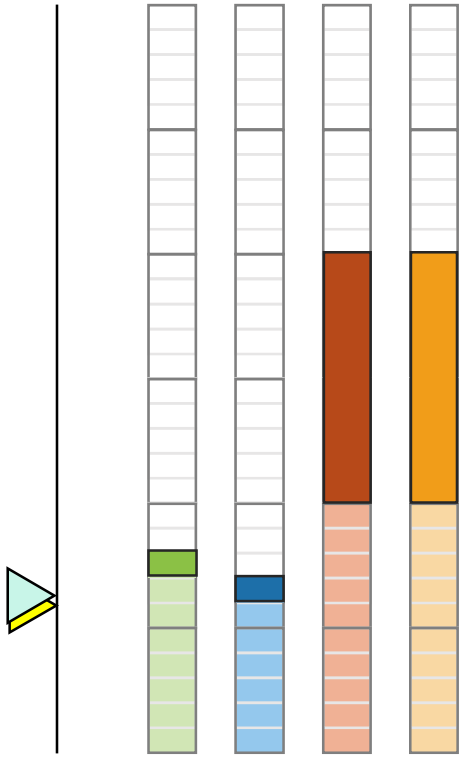


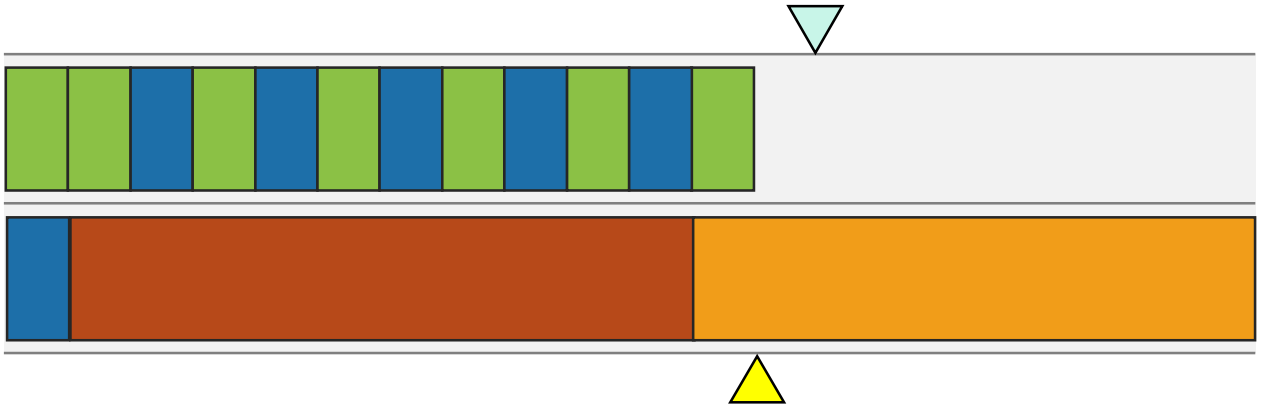


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

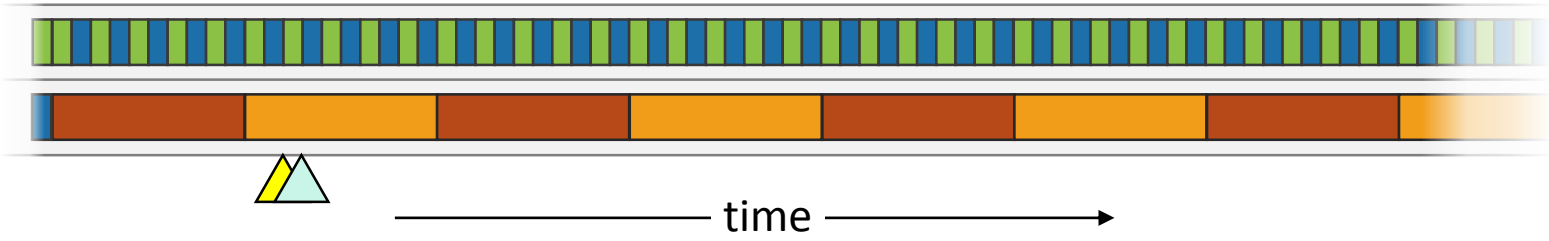


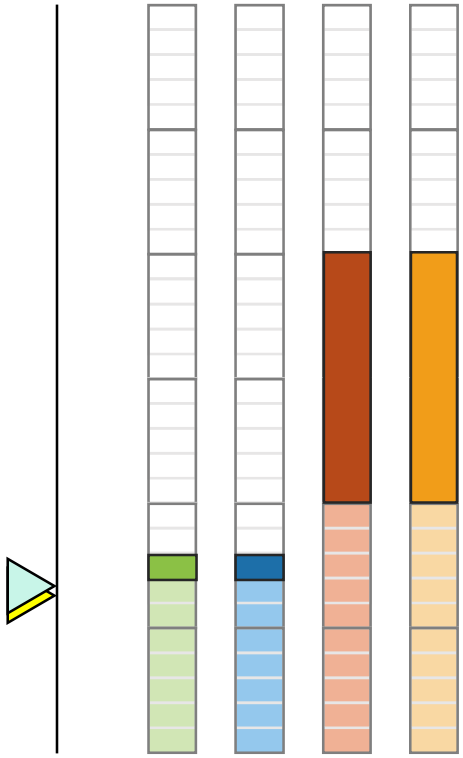


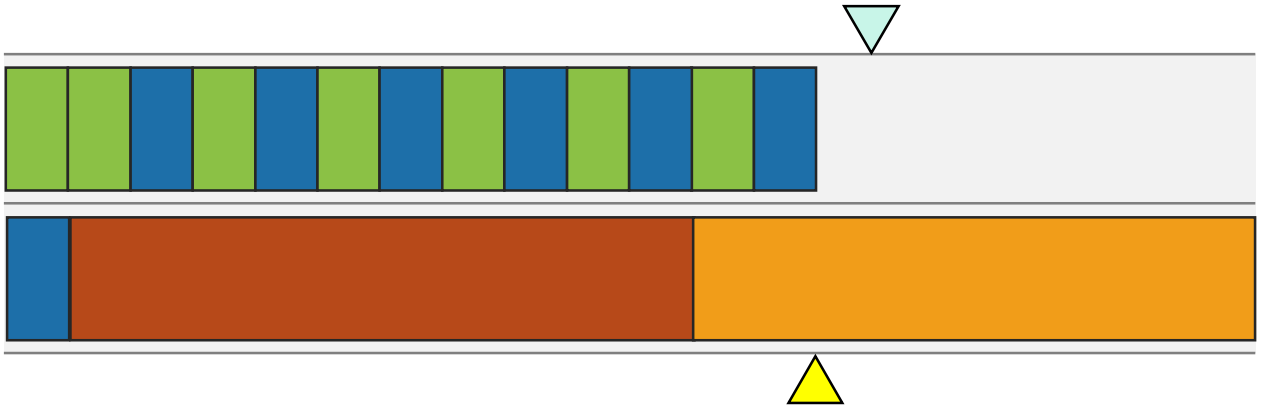


$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

2DFQ

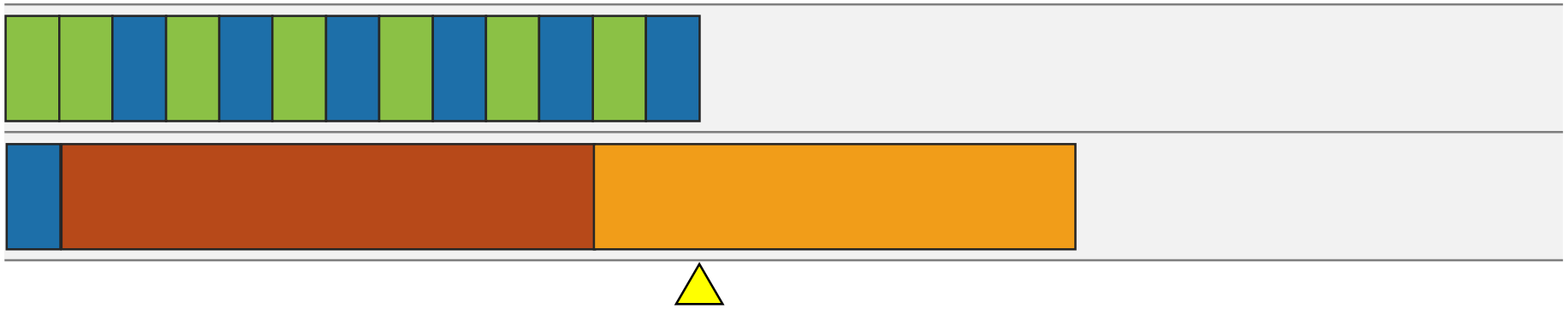




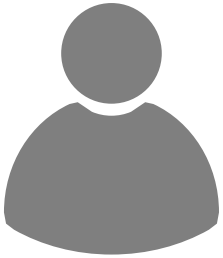


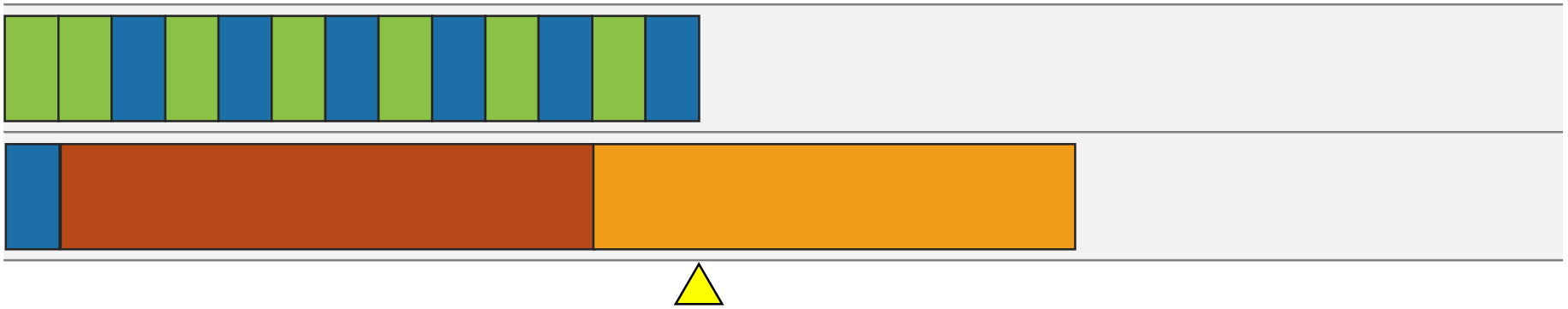
$$\text{Eligible}(r^j, i) = \text{Start}(r^j) - \frac{i}{n} \times \text{size}(r^j)$$

Unknown Costs

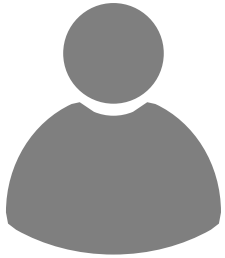


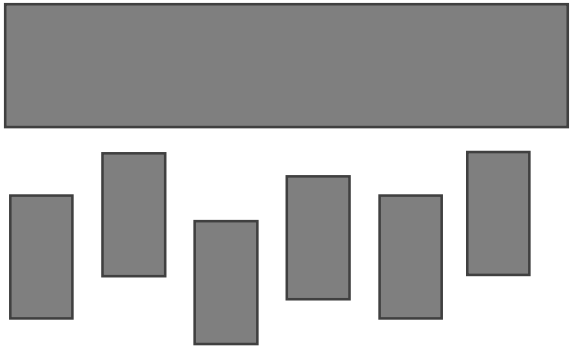
Unknown Costs

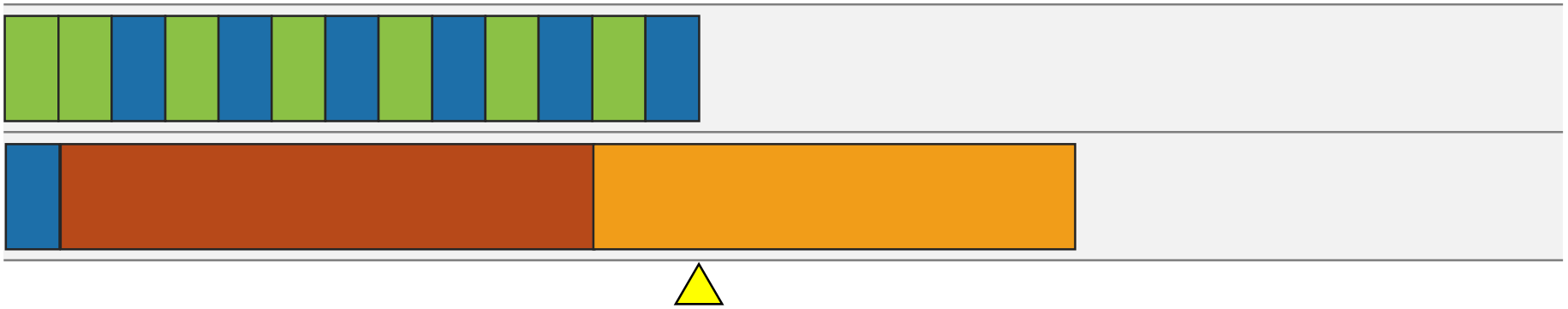




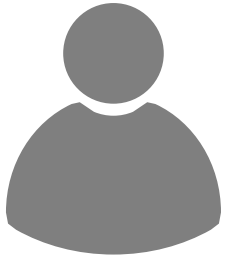
Unknown Costs

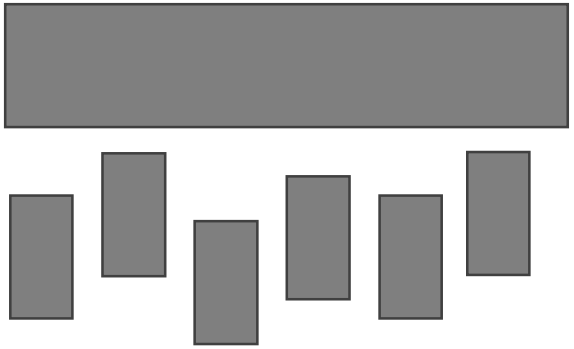


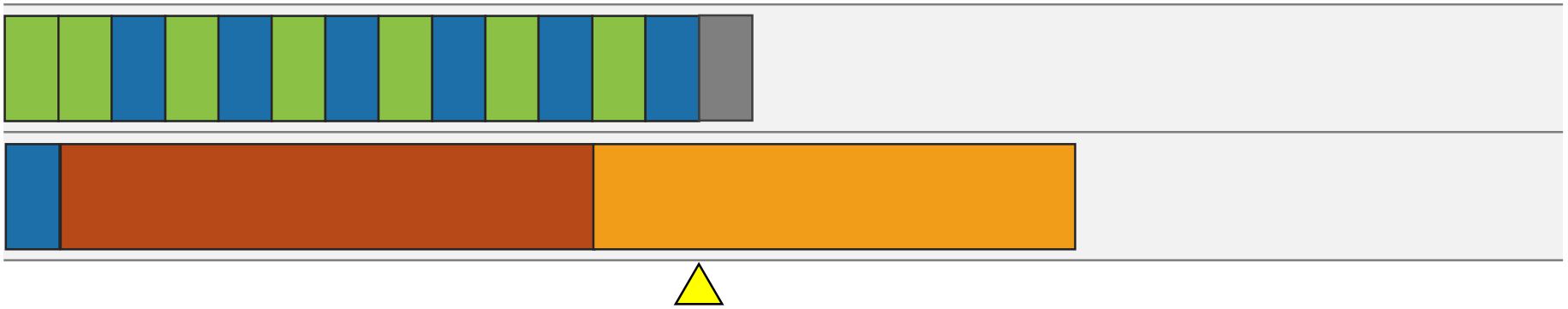




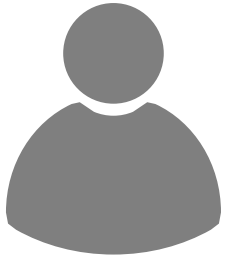
Unknown Costs

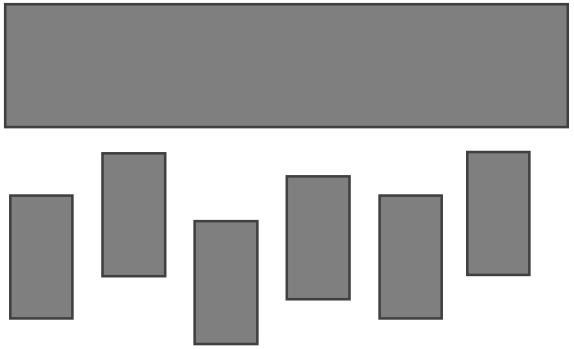


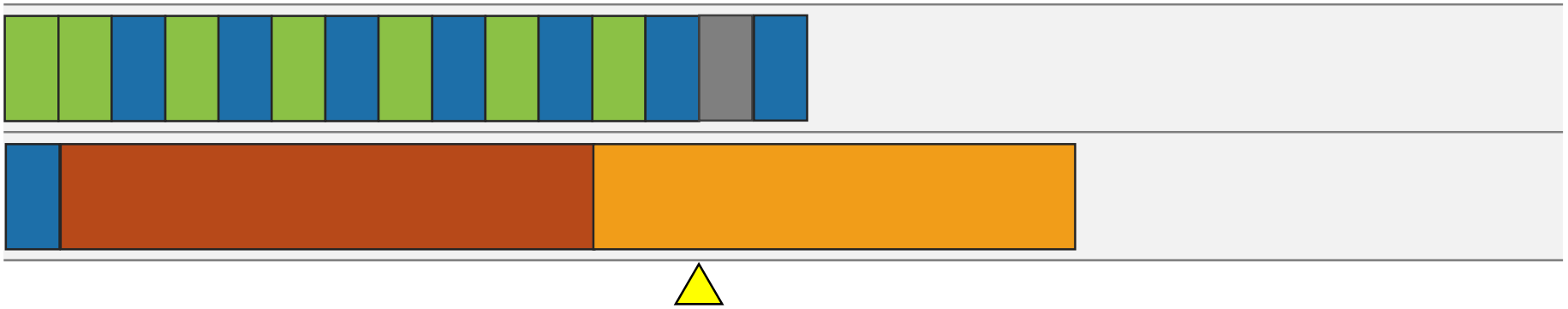




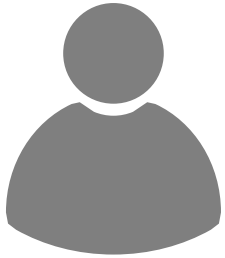
Unknown Costs

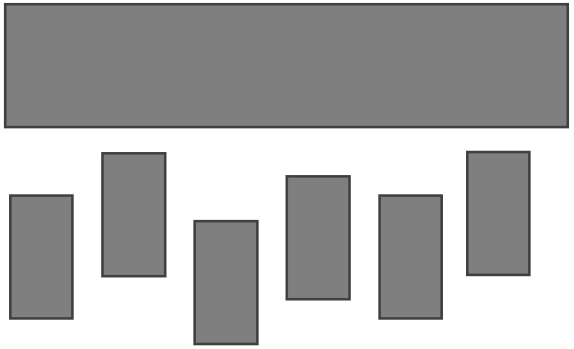


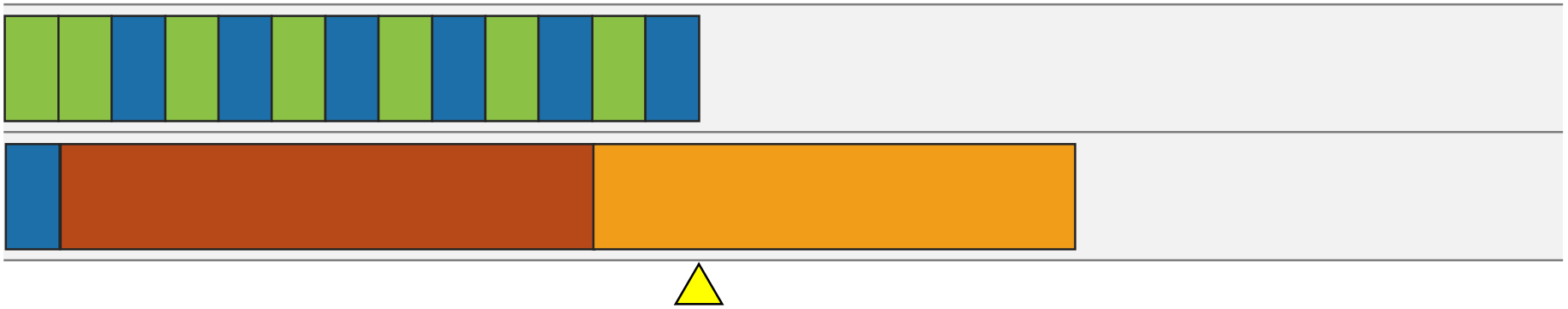




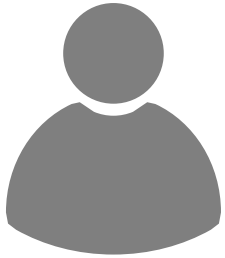
Unknown Costs

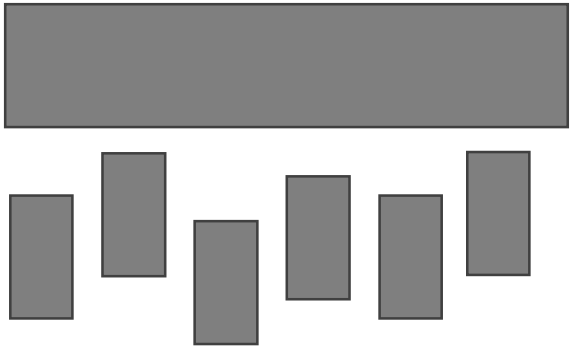


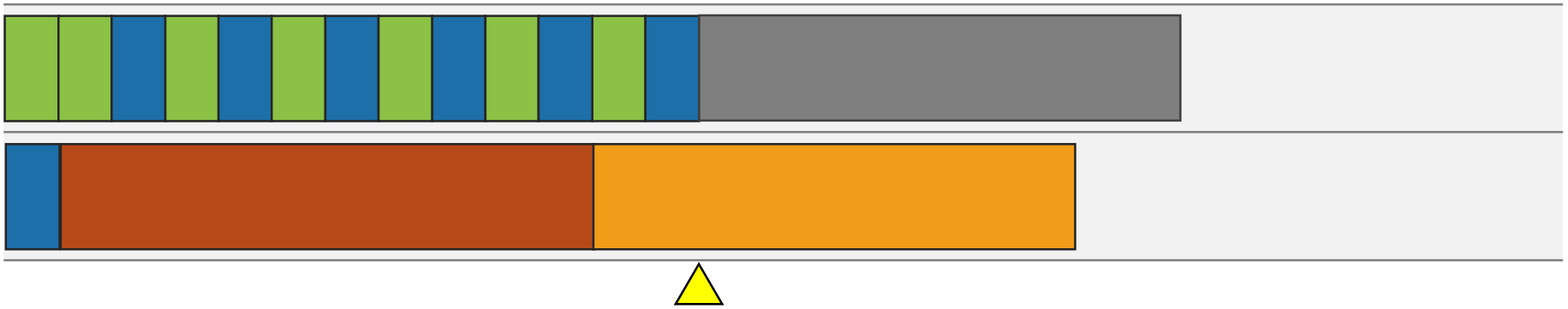




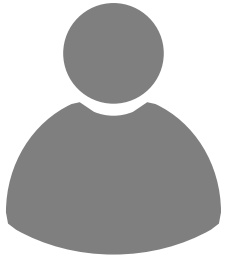
Unknown Costs

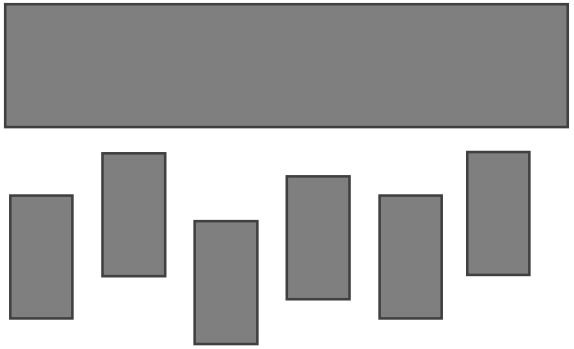


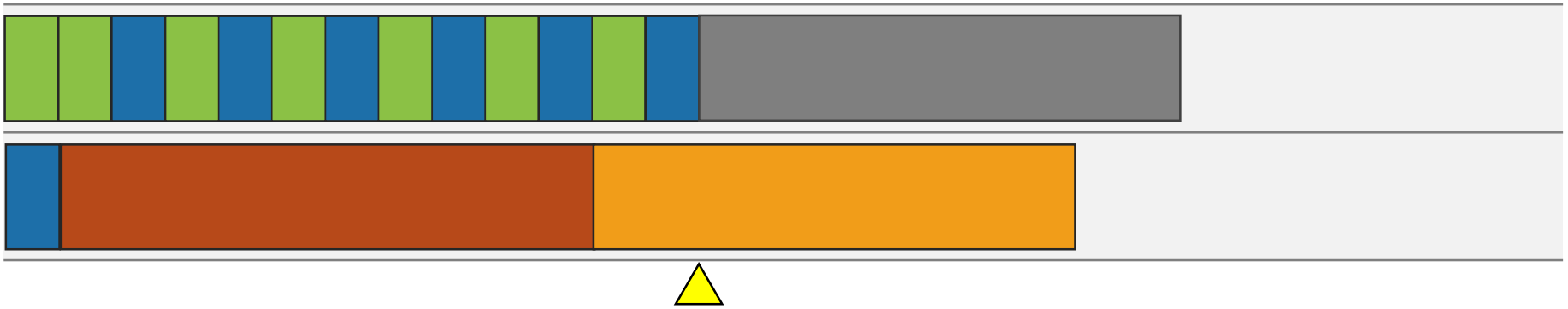




Unknown Costs







Pessimistic cost estimation

Evaluation

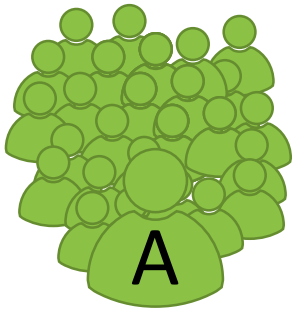
Compare 2DFQ to WFQ and WF²Q

Discrete event simulator with Azure Storage workloads

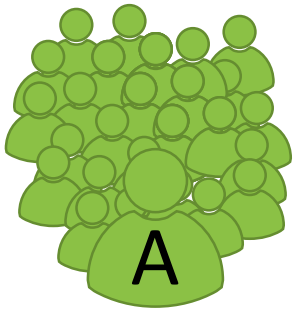
More experiment results in the paper, evaluating:

- Burstiness
- Fairness
- Tail latency

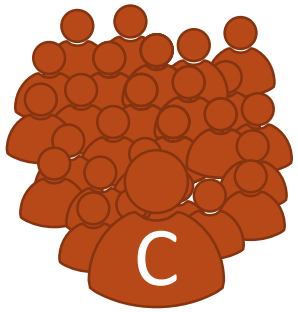
50 tenants with size ≈ 1



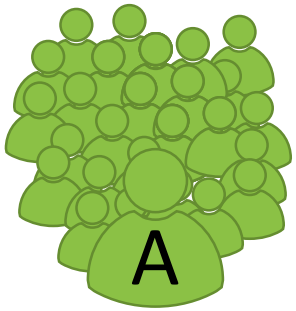
50 tenants with size ≈ 1



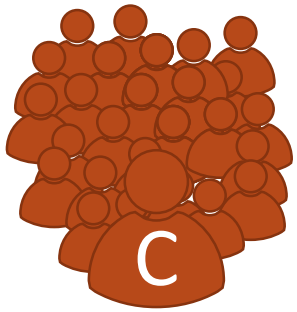
50 tenants with size ≈ 1000



50 tenants with size ≈ 1

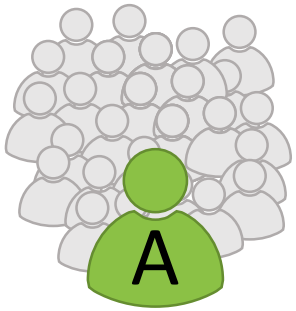


50 tenants with size ≈ 1000

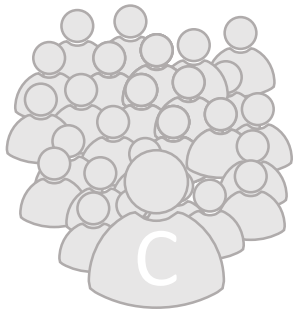


16 threads
1000 units/second
Costs known by scheduler

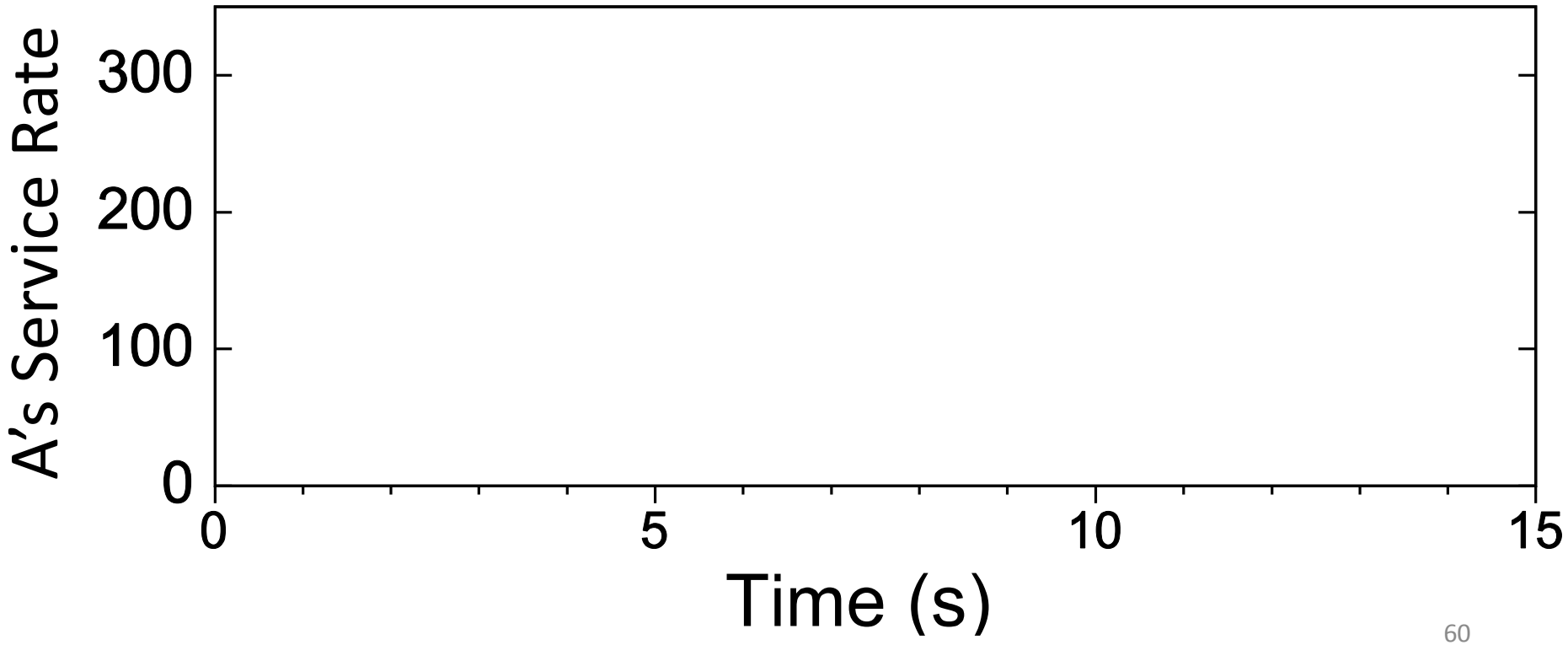
50 tenants with size ≈ 1



50 tenants with size ≈ 1000



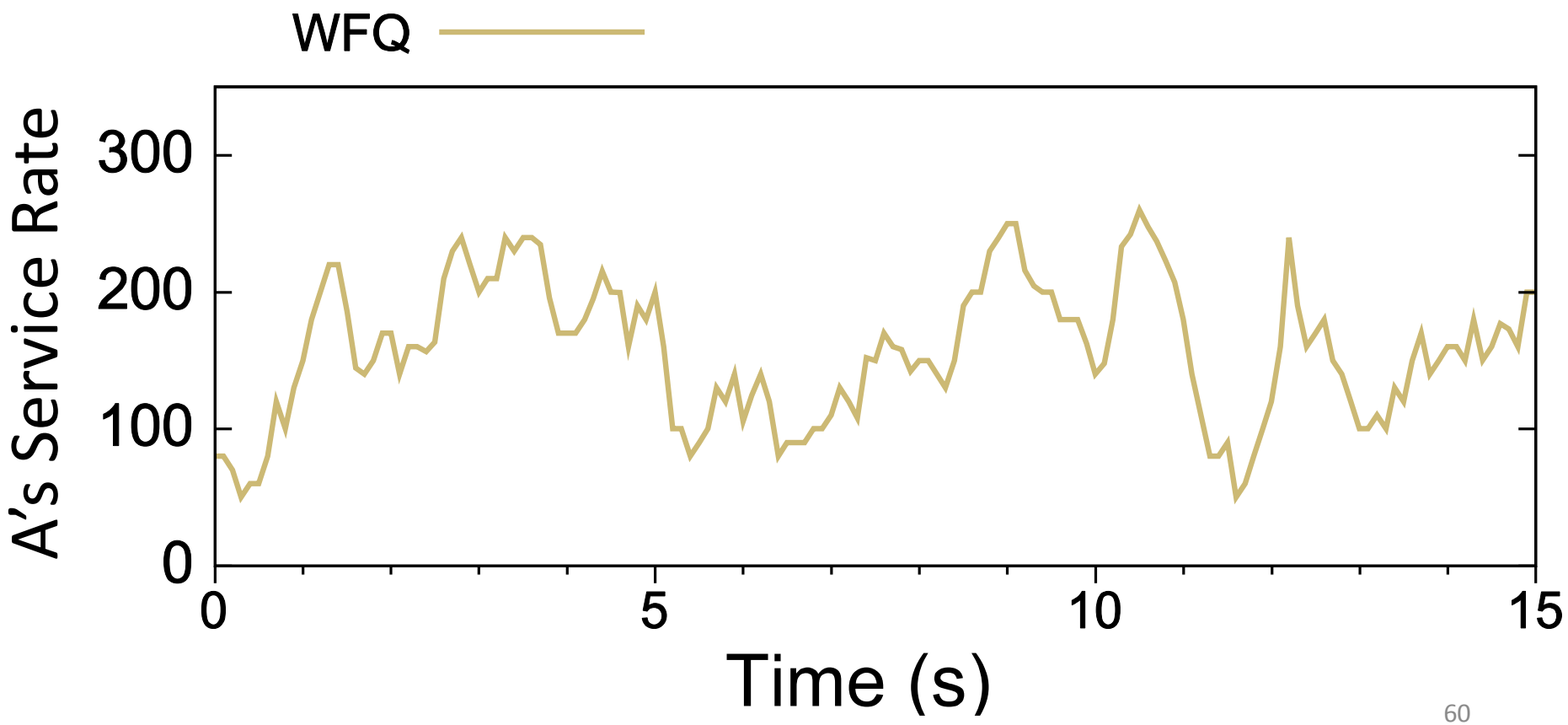
16 threads
1000 units/second
Costs known by scheduler



50 tenants with size ≈ 1

50 tenants with size ≈ 1000

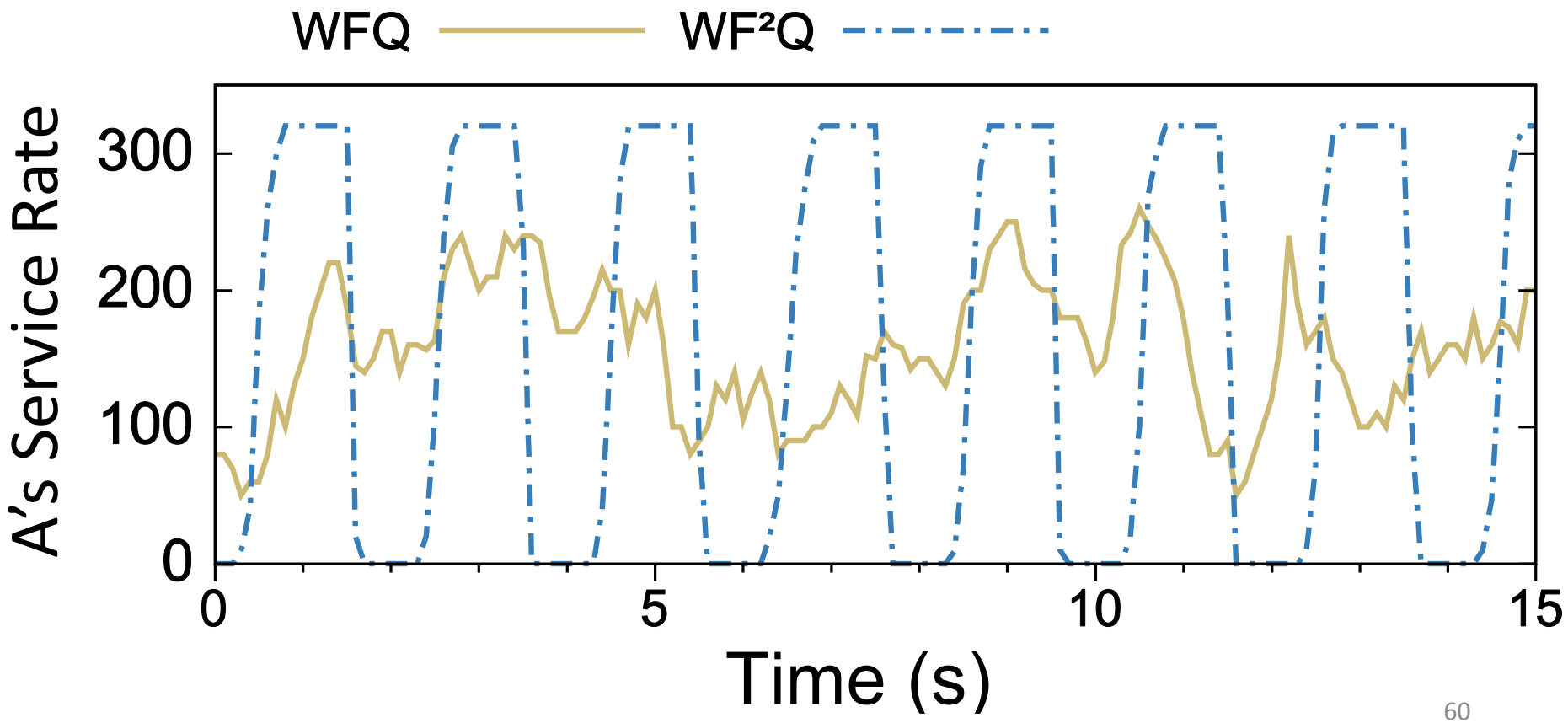
16 threads
1000 units/second
Costs known by scheduler

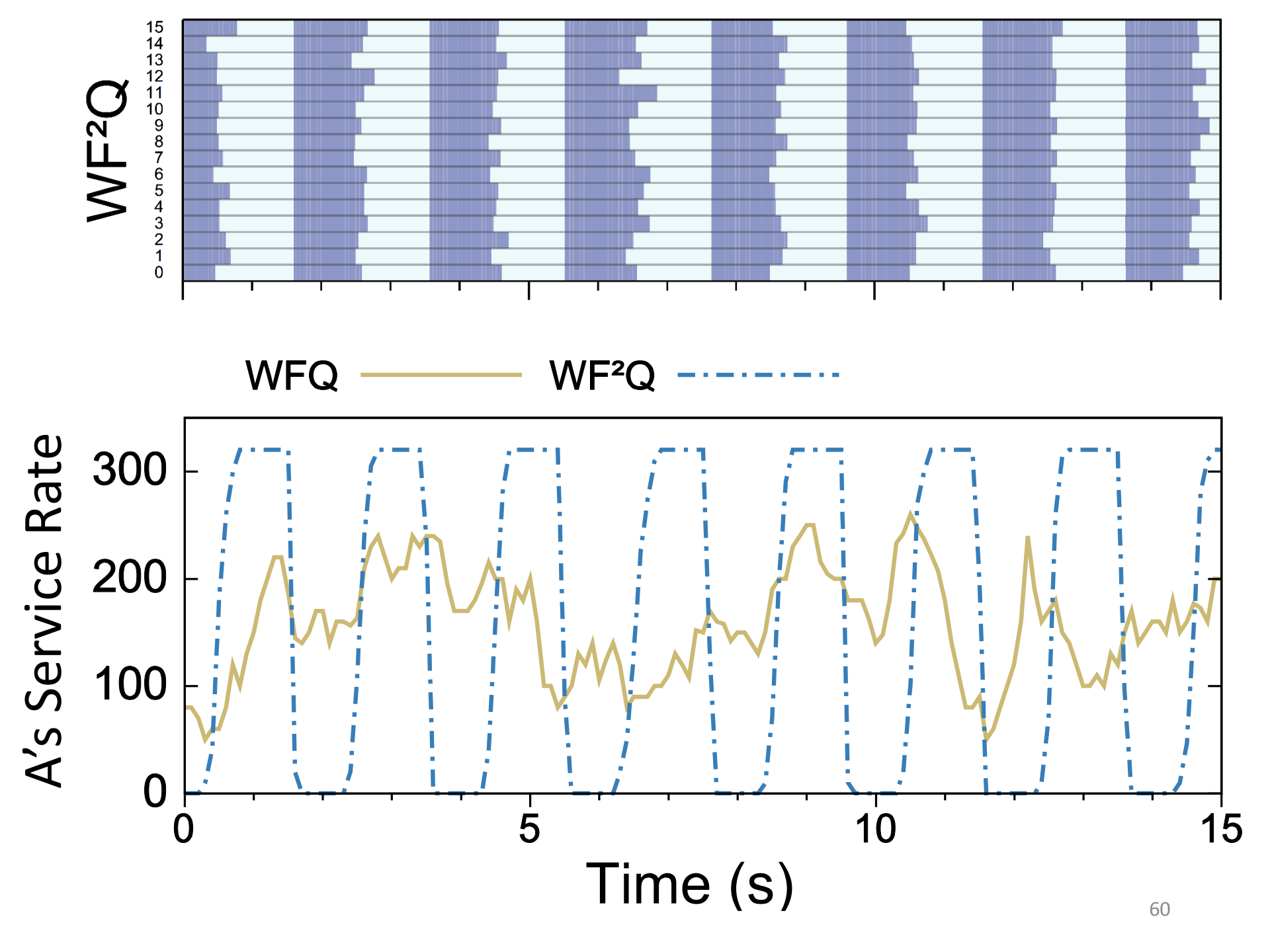


50 tenants with size ≈ 1

50 tenants with size ≈ 1000

16 threads
1000 units/second
Costs known by scheduler

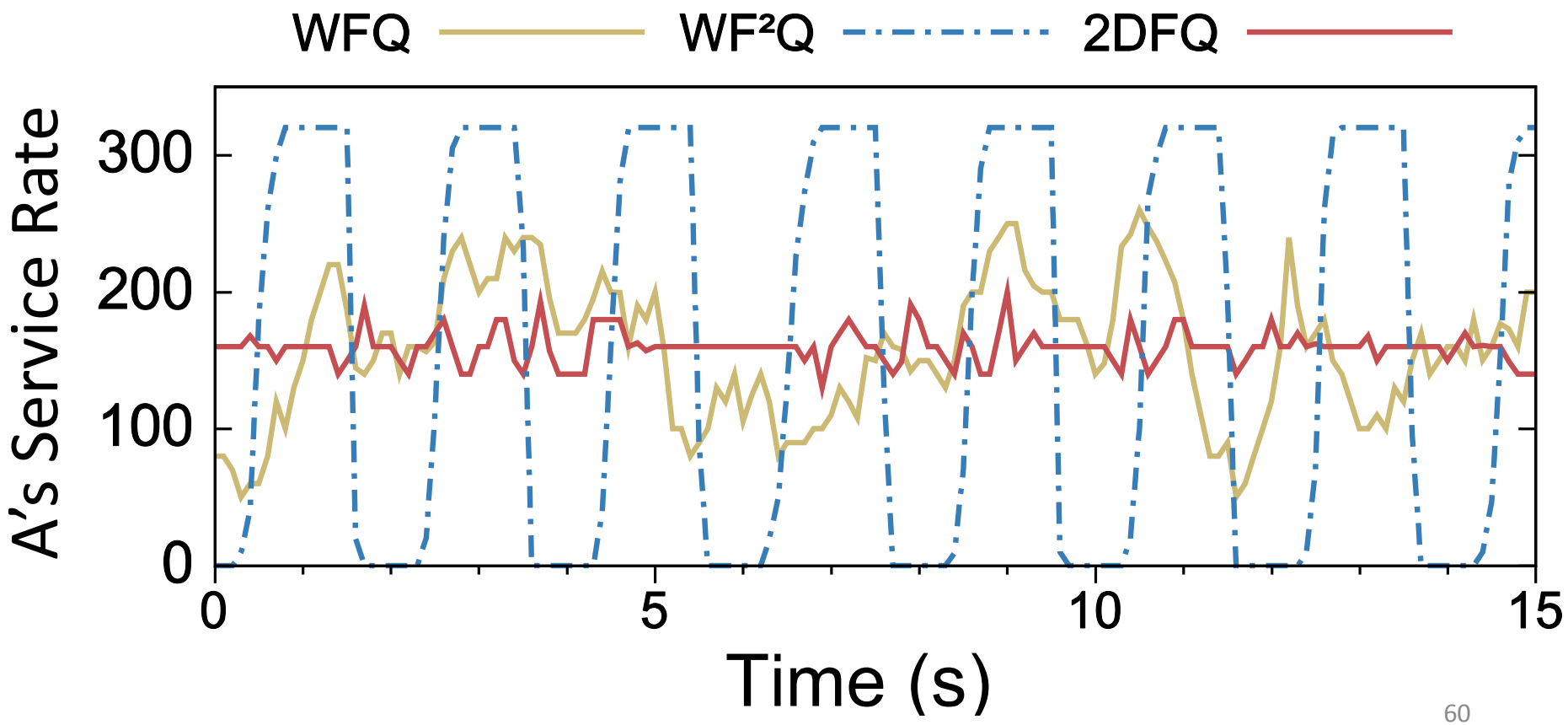


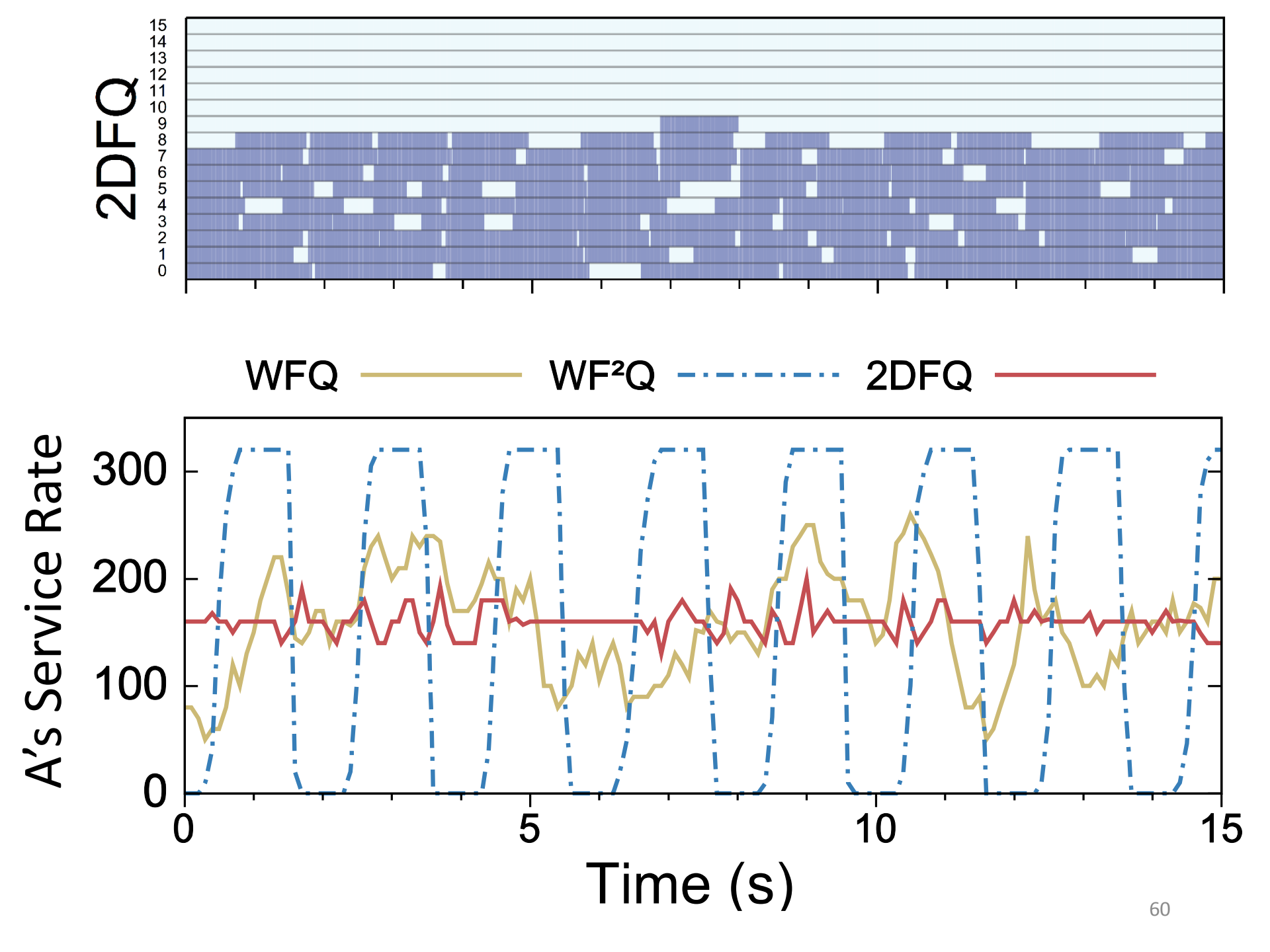

WF²Q
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
WFQ
WF²Q
A's Service Rate
300
200
100
0
0
5
10
15
Time (s)

50 tenants with size ≈ 1

50 tenants with size ≈ 1000

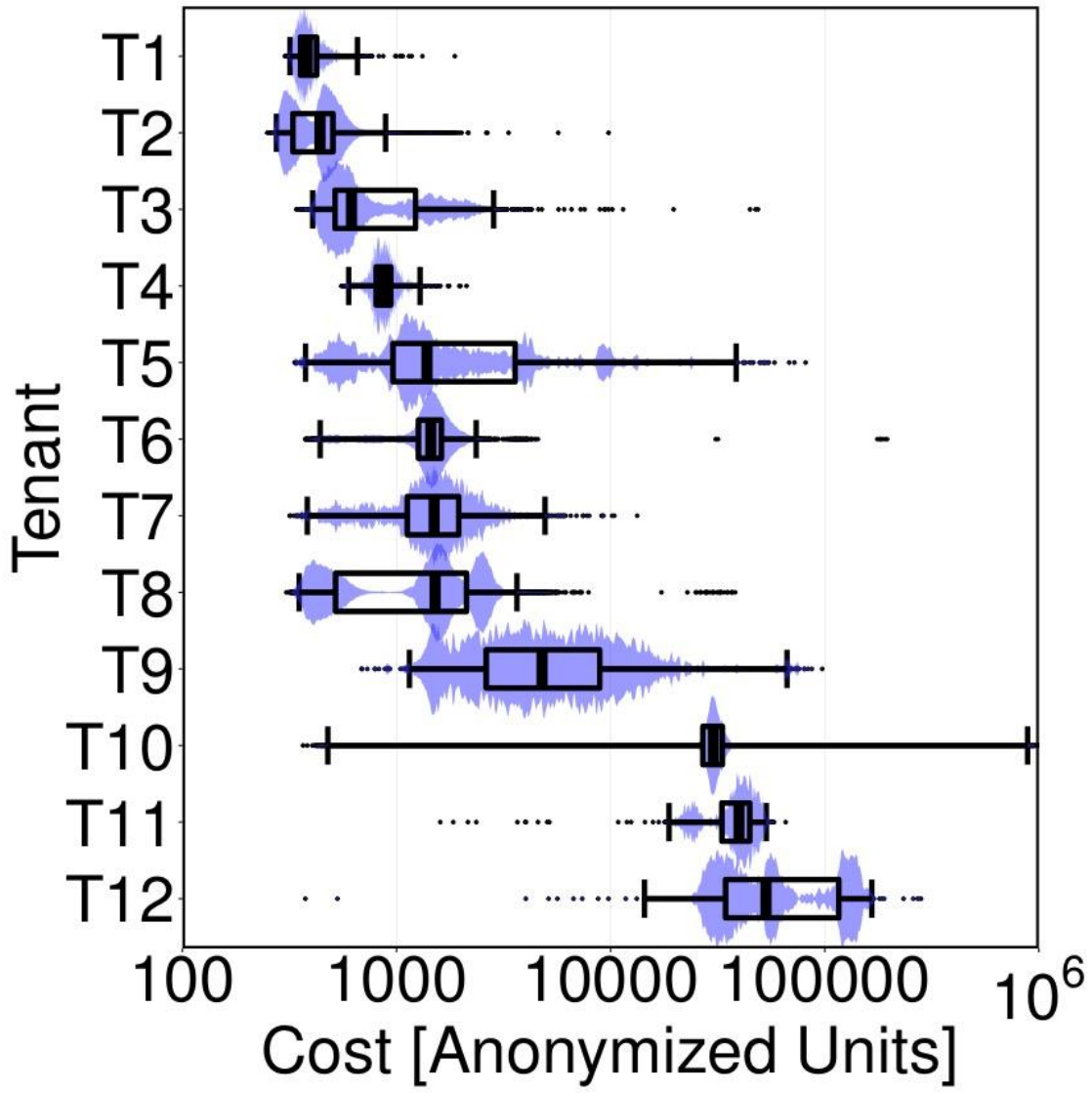
16 threads
1000 units/second
Costs known by scheduler



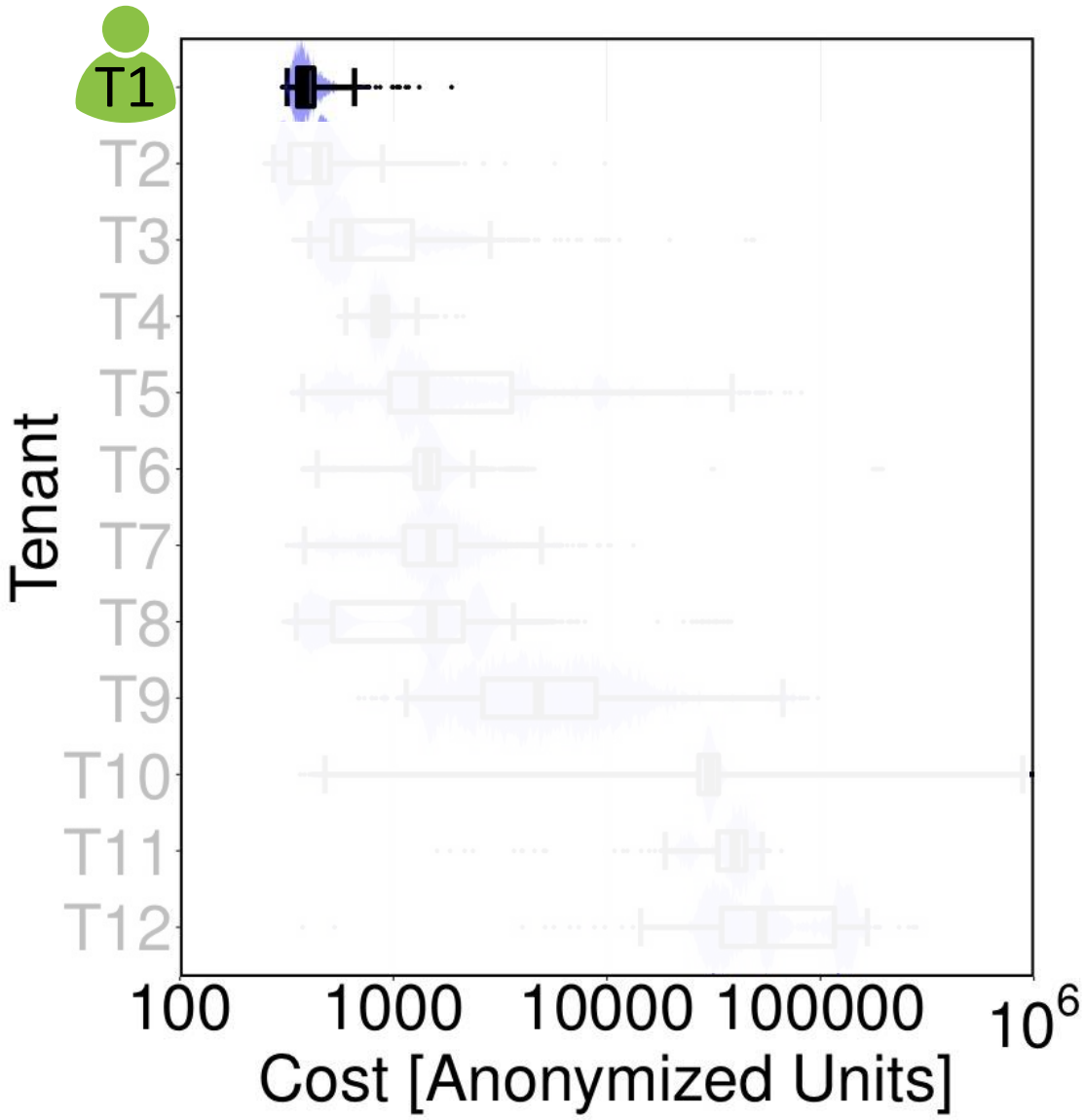

2DFQ
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
WFQ
WF²Q
2DFQ
A's Service Rate
300
200
100
0
0
5
10
15
Time (s)

250 Azure Storage tenants
32 threads
1 million units/second
Costs known by scheduler

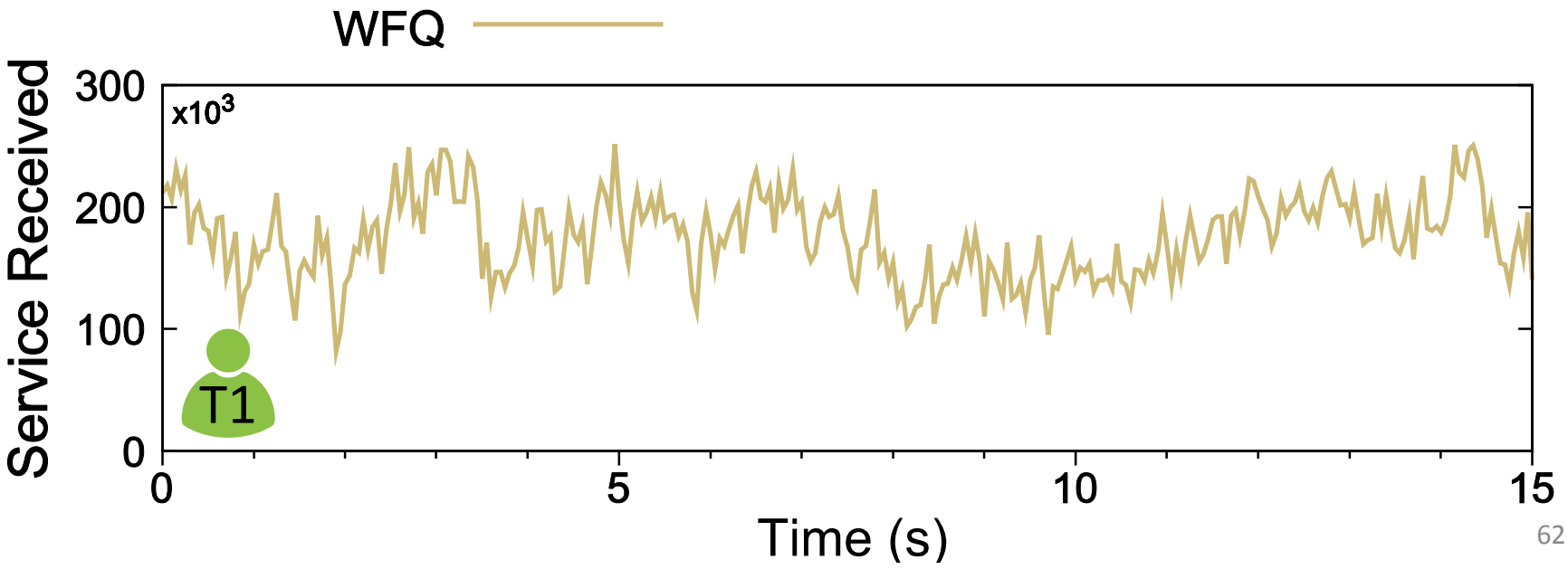
250 Azure Storage tenants
32 threads
1 million units/second
Costs known by scheduler



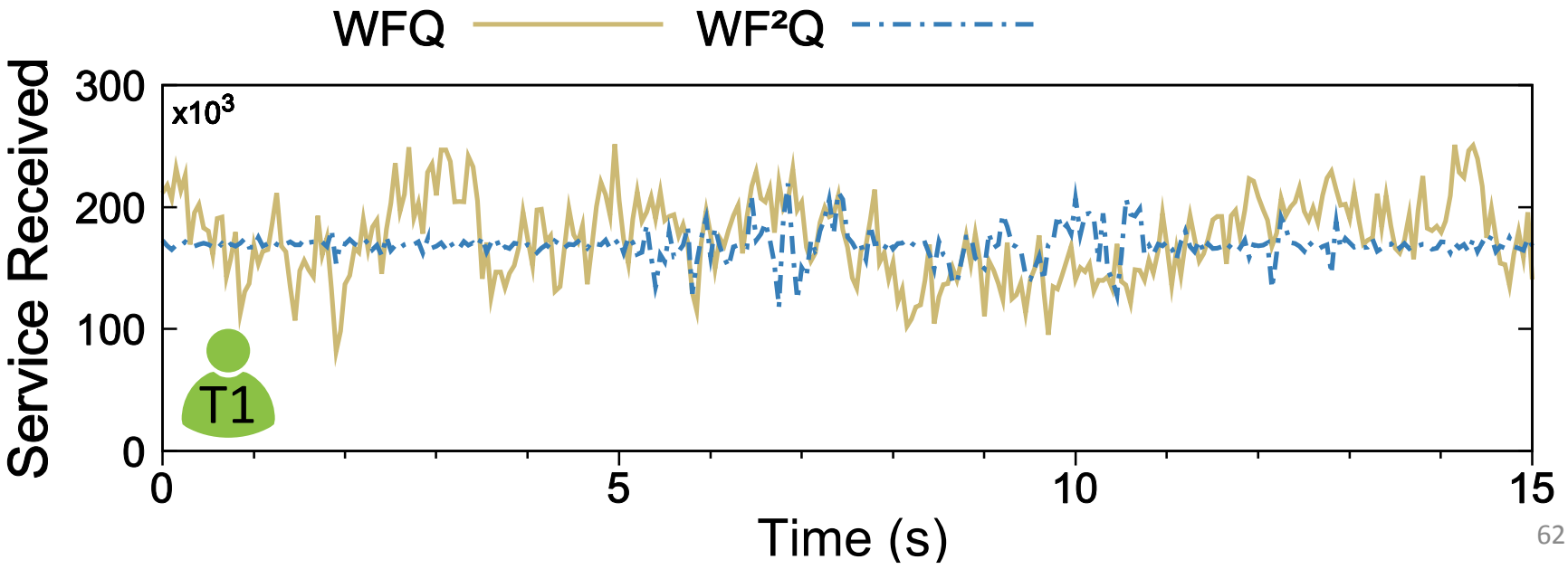
250 Azure Storage tenants
32 threads
1 million units/second
Costs known by scheduler



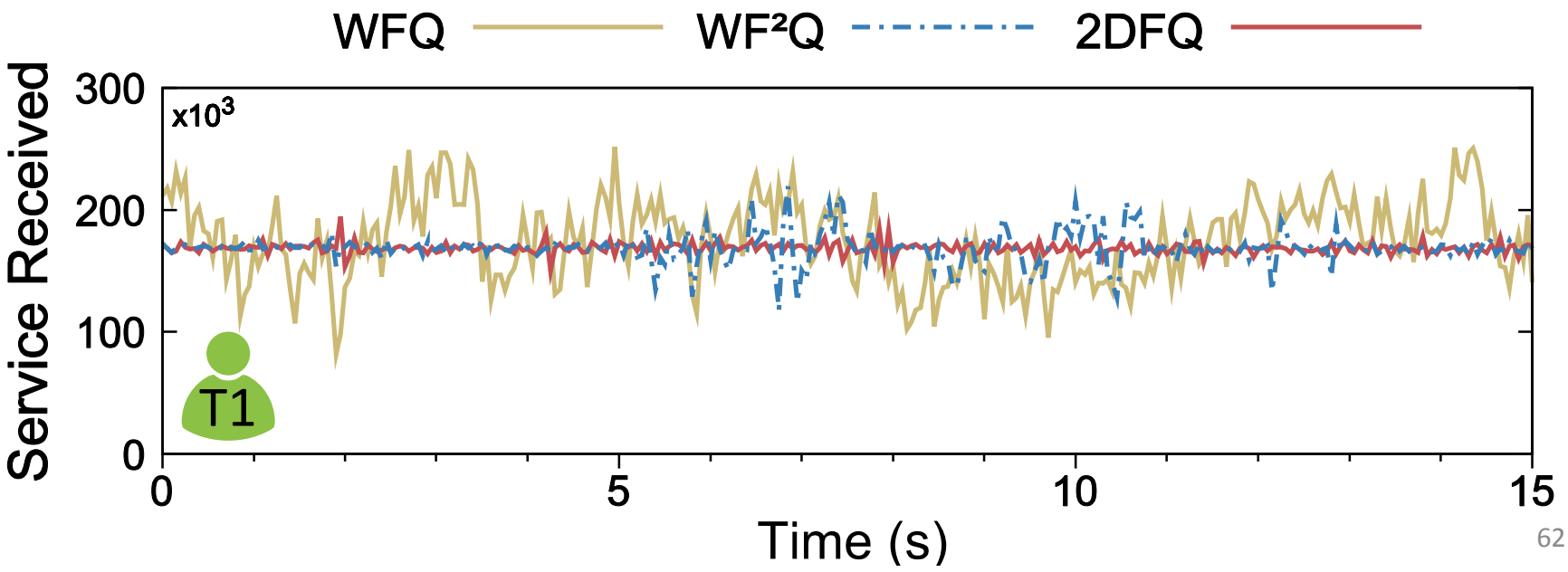
250 Azure Storage workloads
32 threads
1 million units/second
Costs known by scheduler



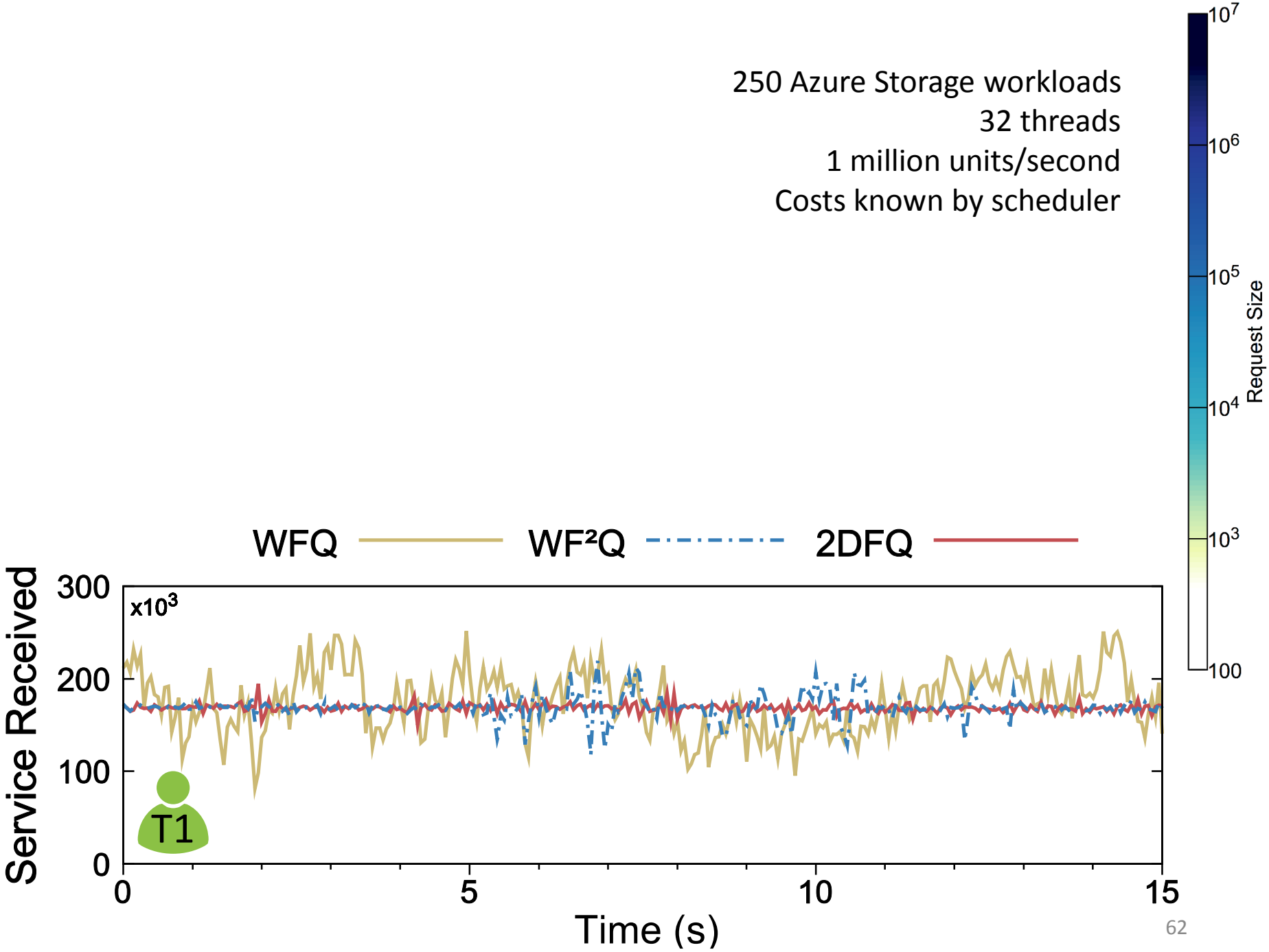
250 Azure Storage workloads
32 threads
1 million units/second
Costs known by scheduler

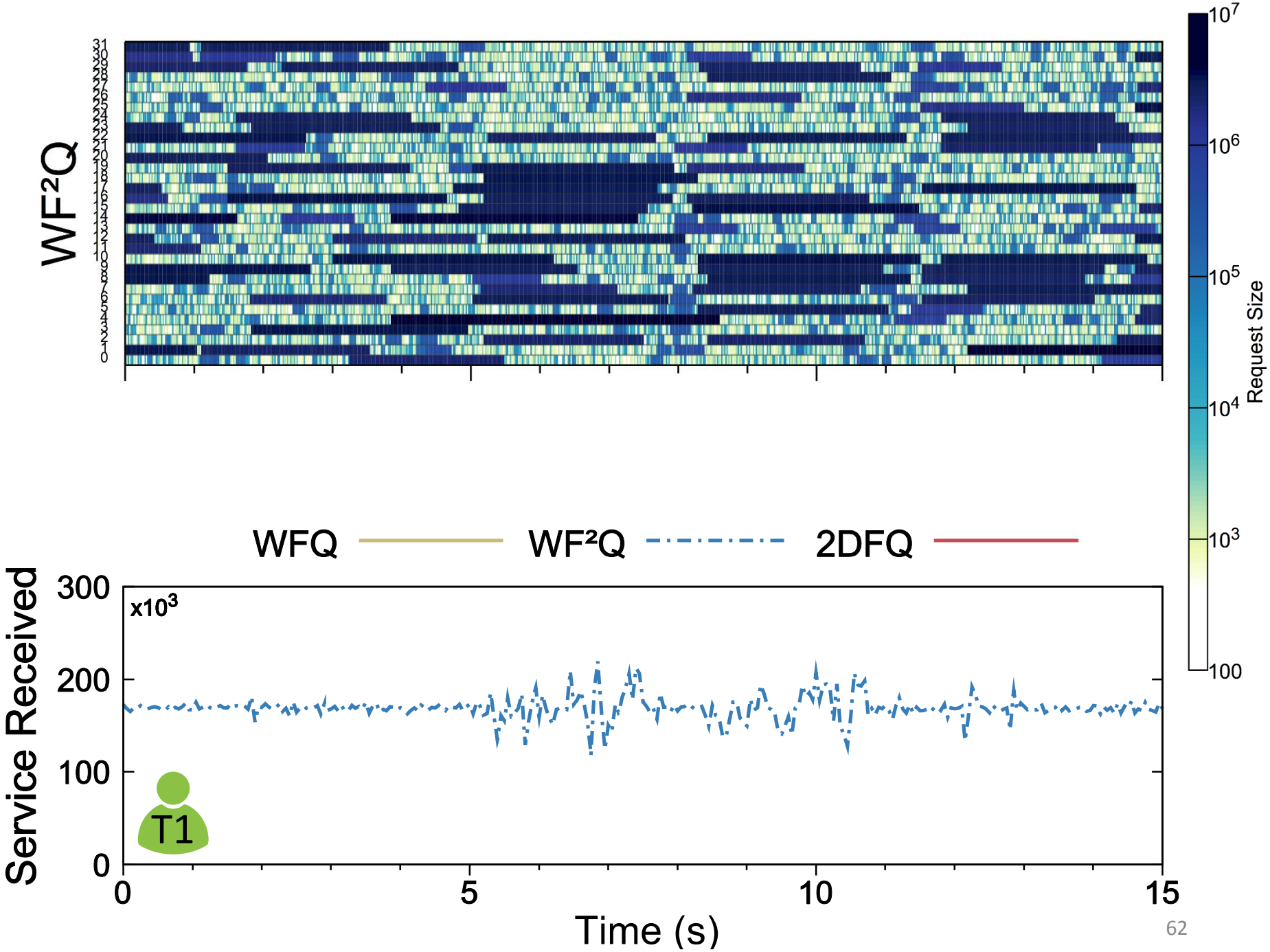


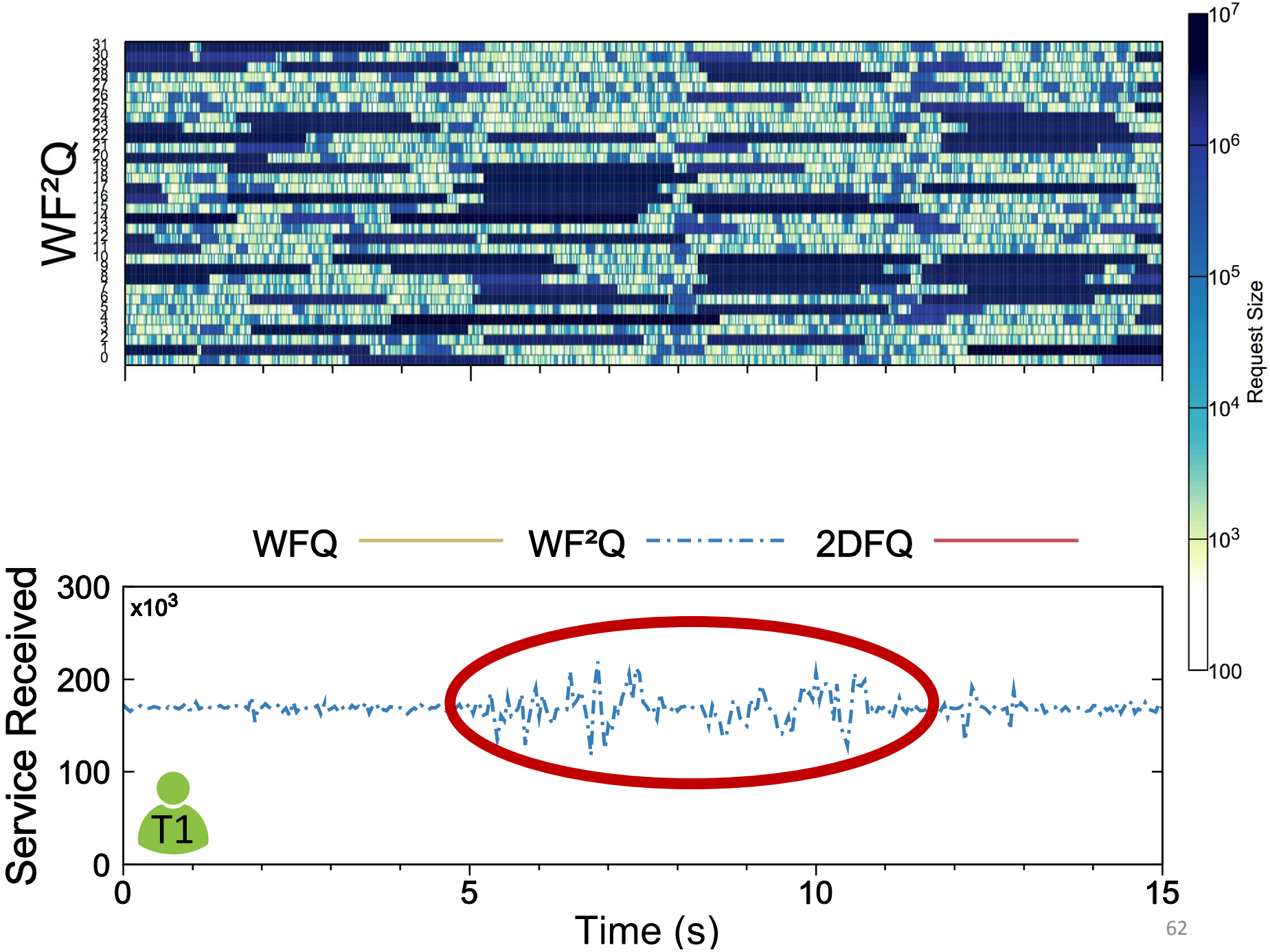
250 Azure Storage workloads
32 threads
1 million units/second
Costs known by scheduler

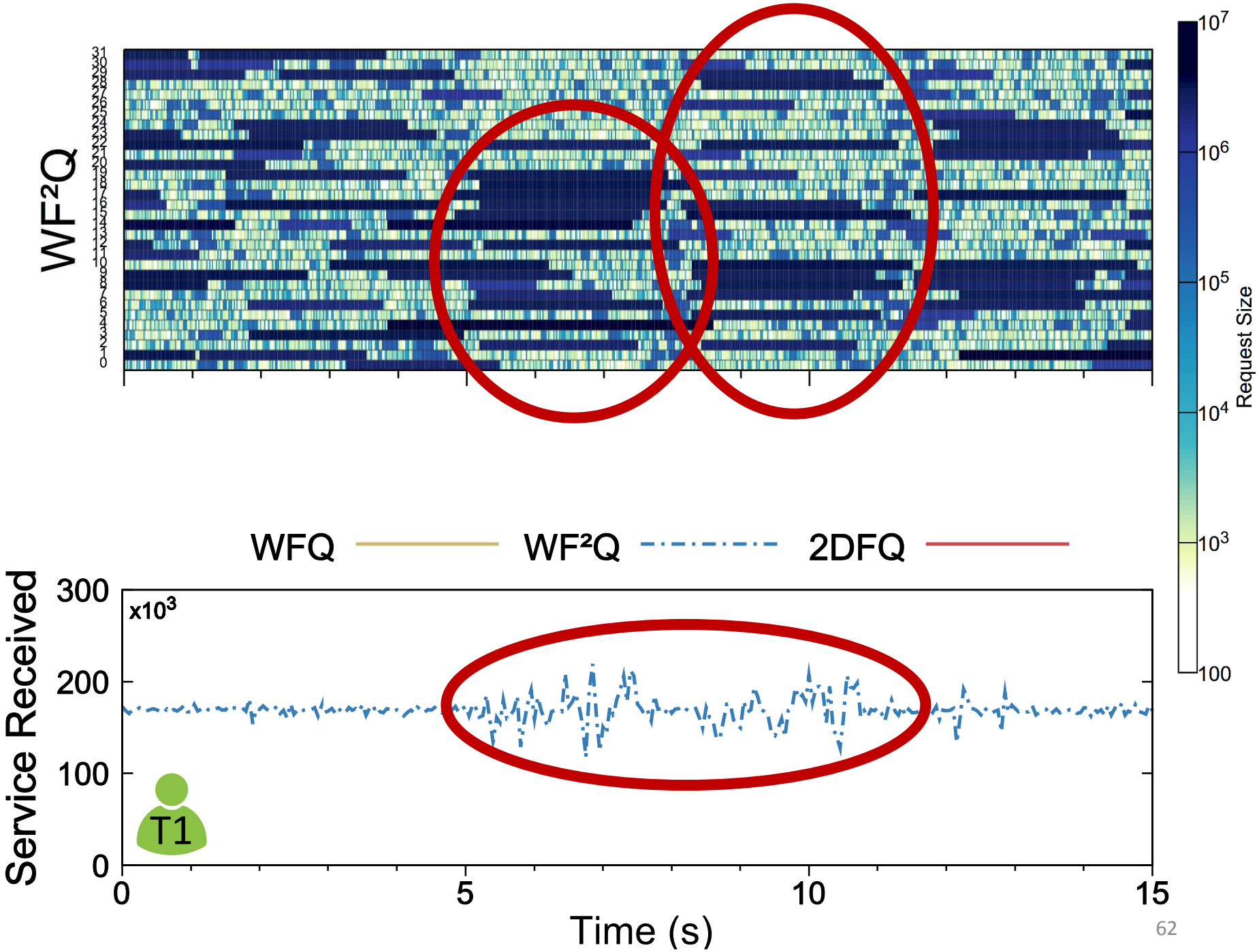


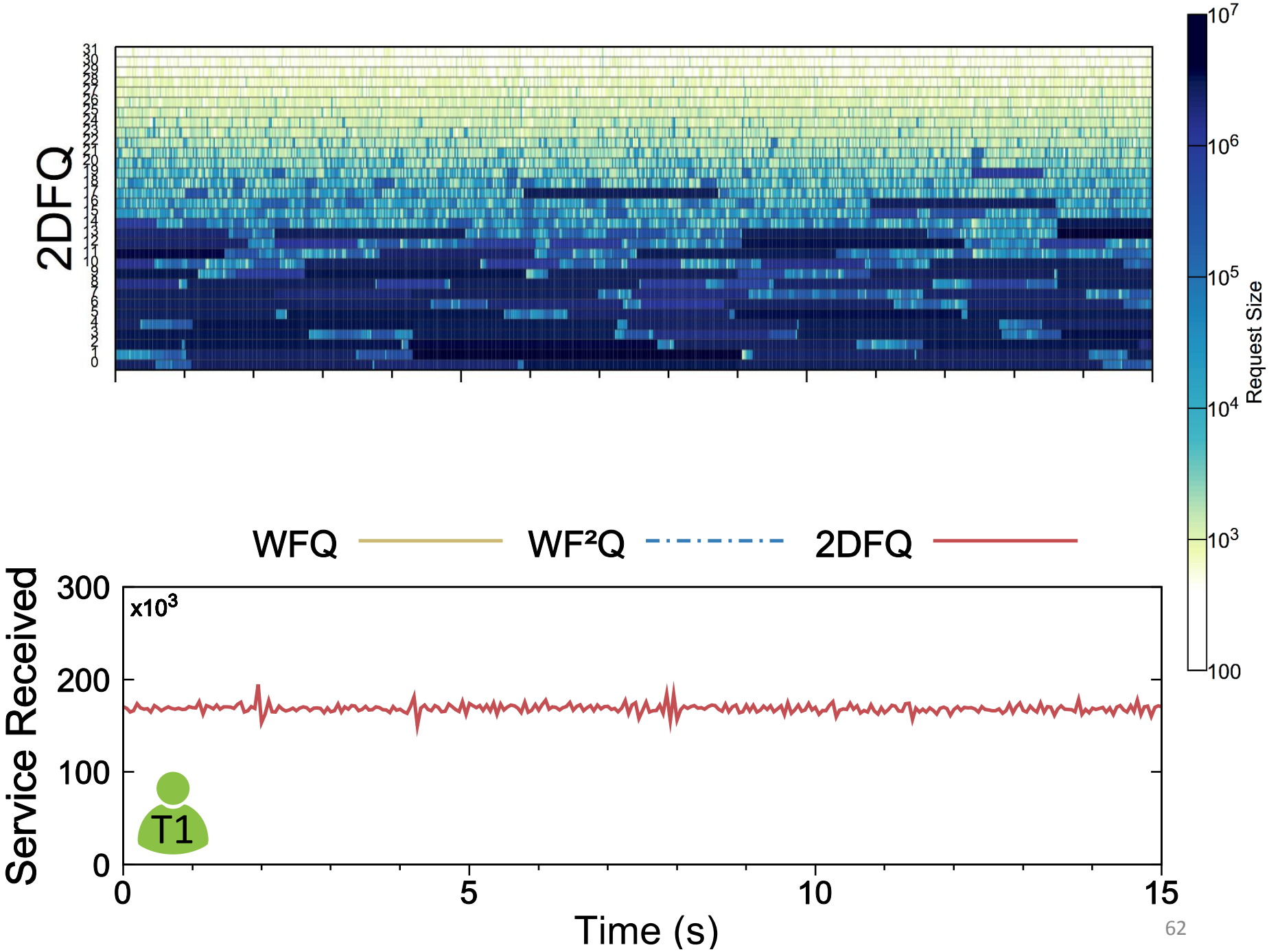
250 Azure Storage workloads
32 threads
1 million units/second
Costs known by scheduler

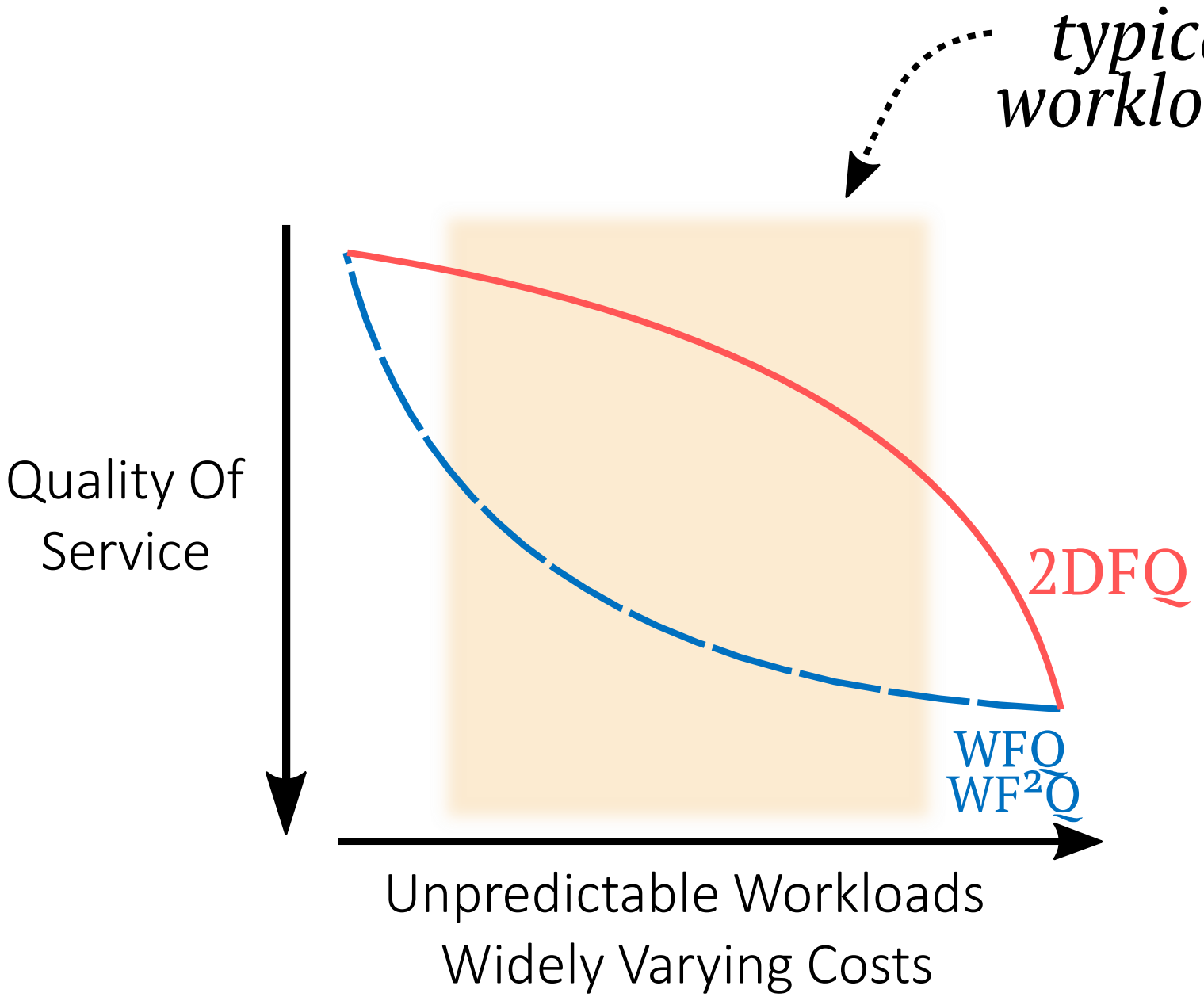


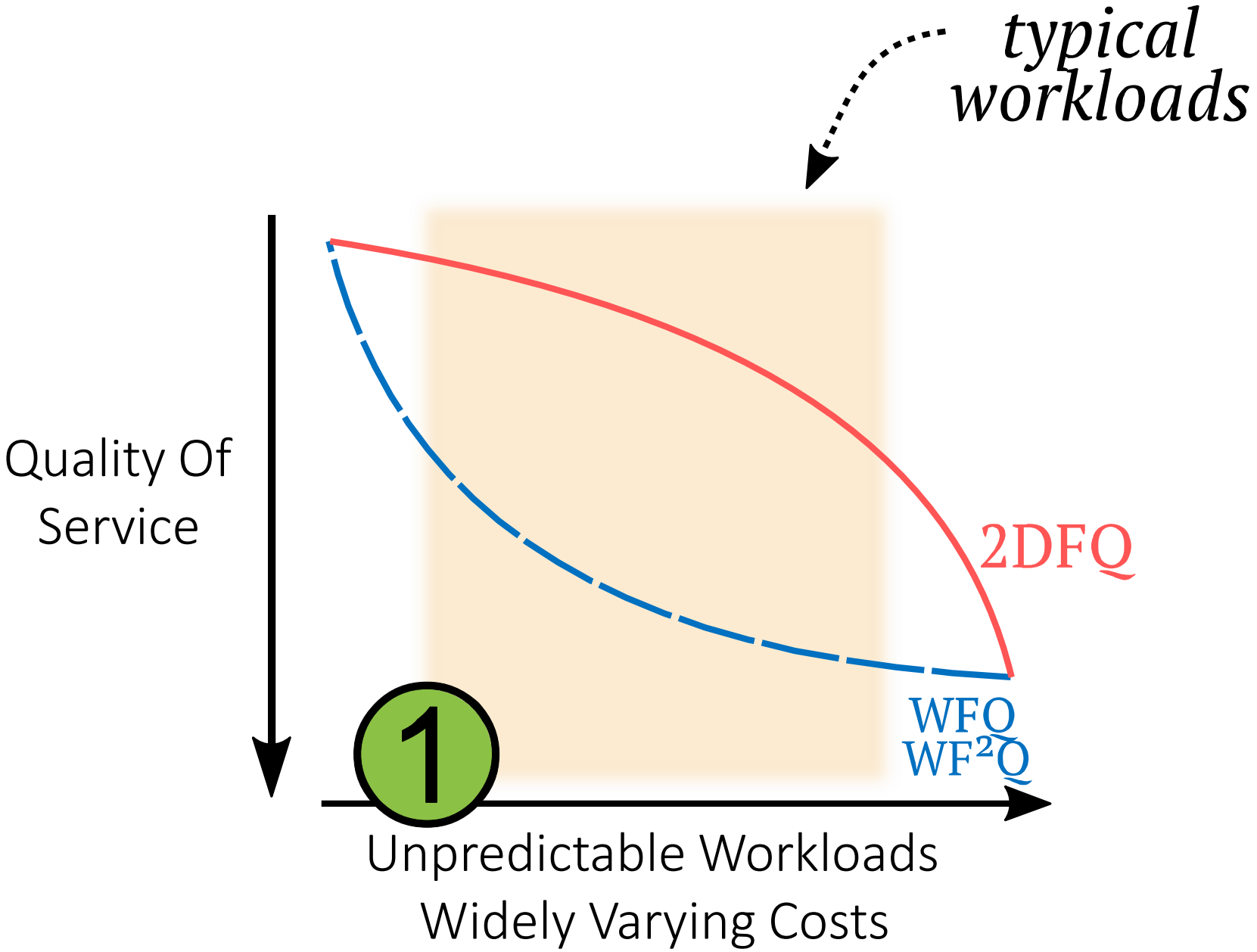

WF²Q
31
30
29
28
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
10^7
10^6
10^5
10^4
10^3
100
Request Size
WFQ
WF²Q
2DFQ
Service Received
300
200
100
0
x10^3
T1
0
5
10
15
Time (s)

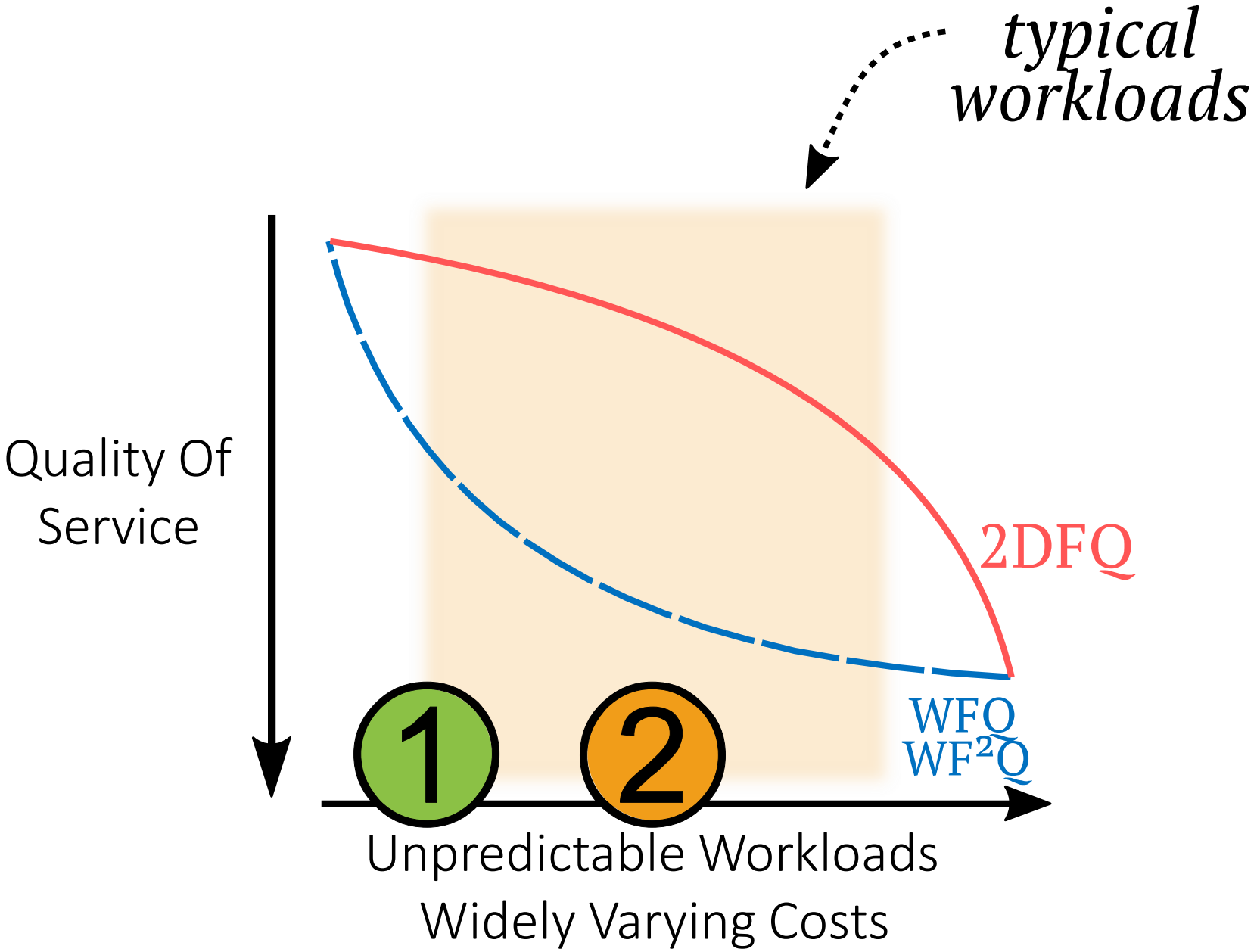

WF²Q
Request Size
10^7
10^6
10^5
10^4
10^3
100
WFQ
WF²Q
2DFQ
Service Received
x10^3
300
200
100
0
T1
0
5
10
15
Time (s)

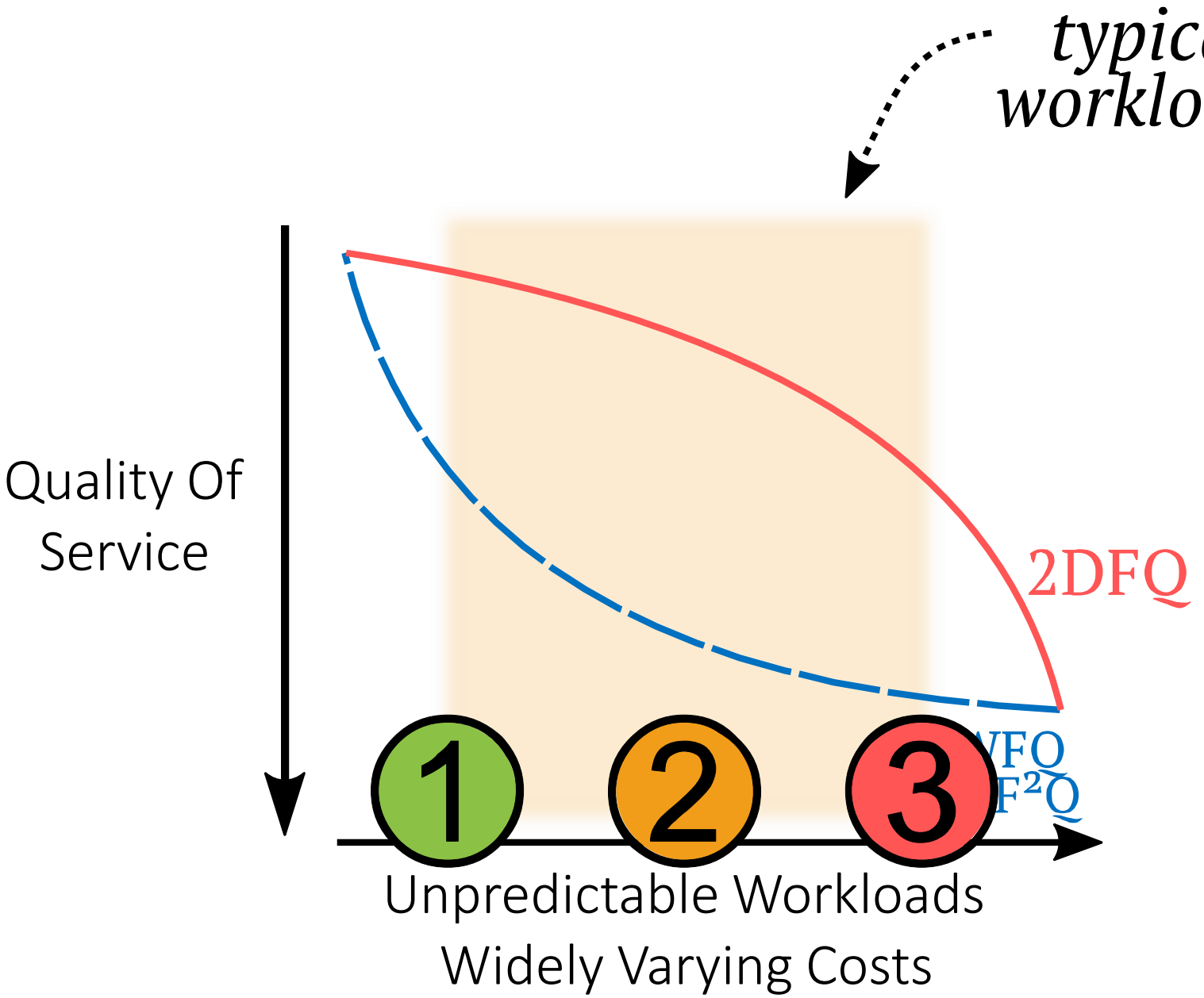

WF²Q
Request Size
10^7
10^6
10^5
10^4
10^3
100
WFQ
WF²Q
2DFQ
Service Received
x10^3
300
200
100
0
T1
0
5
10
15
Time (s)

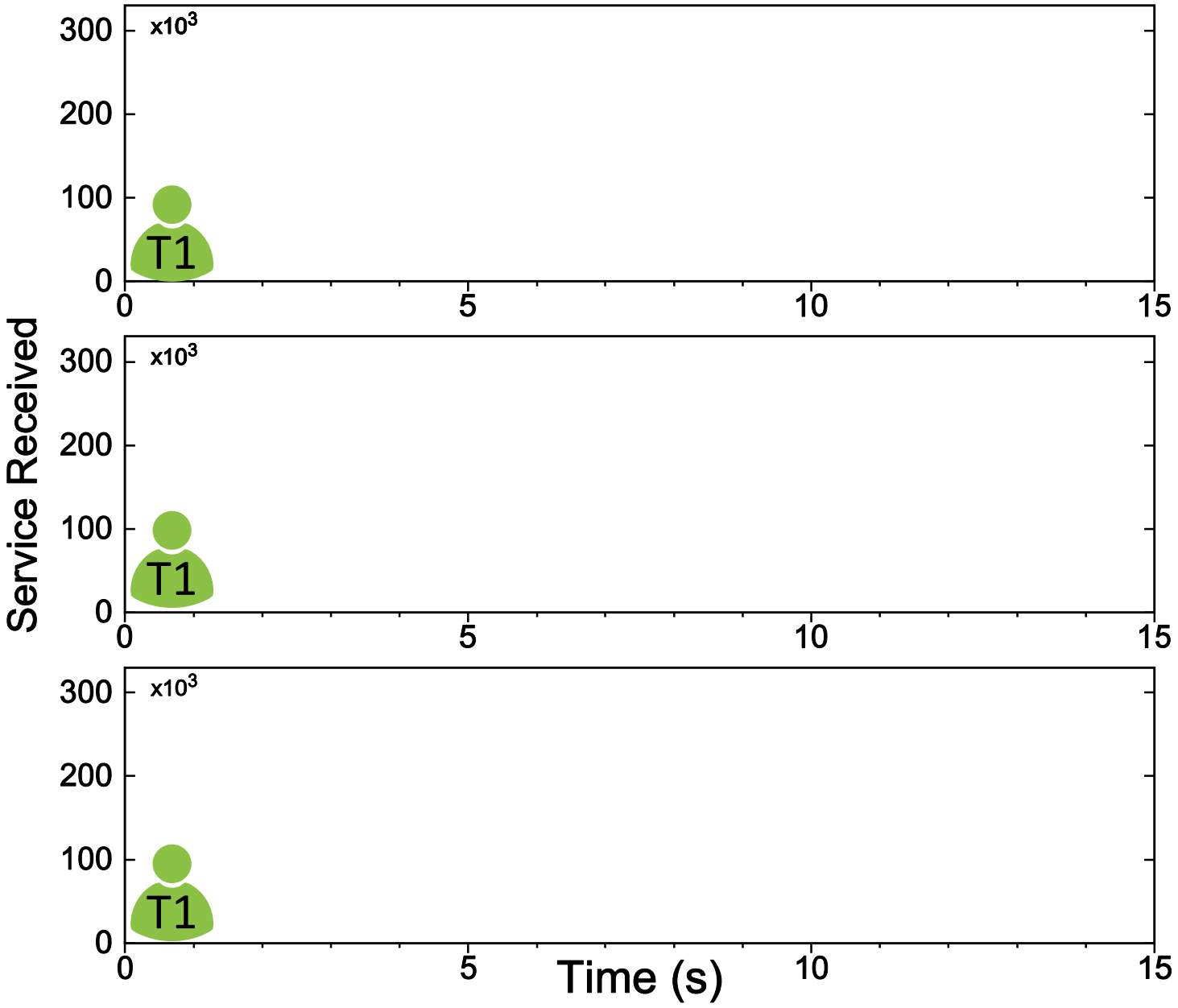

2DFQ
Request Size
10^7
10^6
10^5
10^4
10^3
100
WFQ
WF²Q
2DFQ
Service Received
x10³
300
200
100
0
T1
0
5
10
15
Time (s)


typical workloads
Quality Of Service
2DFQ
WFQ
WF²Q
Unpredictable Workloads
Widely Varying Costs


typical workloads
Quality Of Service
2DFQ
WFQ
WF²Q
1
Unpredictable Workloads
Widely Varying Costs


typical workloads
Quality Of Service
2DFQ
WFQ
WF²Q
1
2
Unpredictable Workloads
Widely Varying Costs


typical workloads
Quality Of Service
2DFQ
1
2
3
Unpredictable Workloads
Widely Varying Costs



1

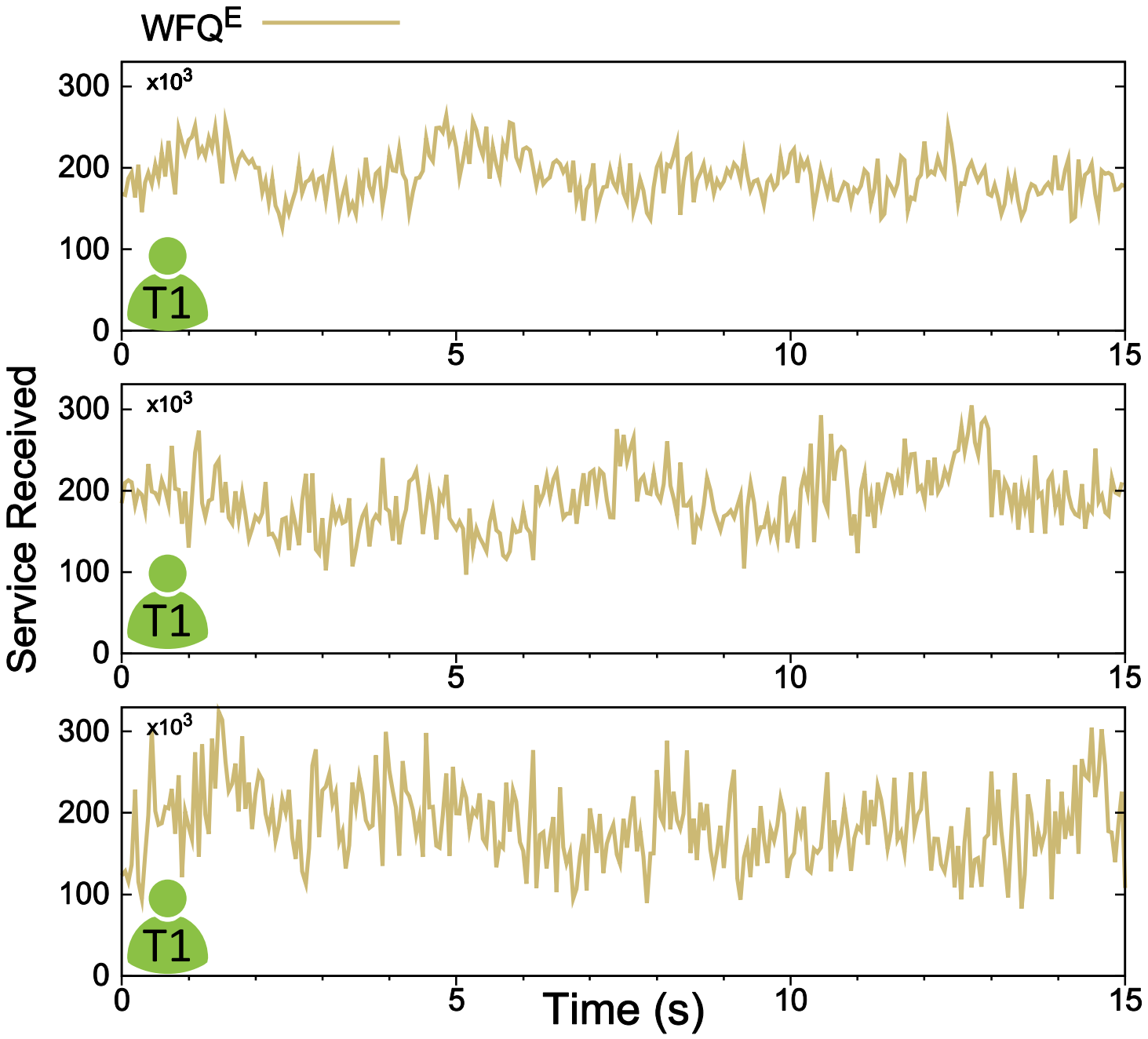
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

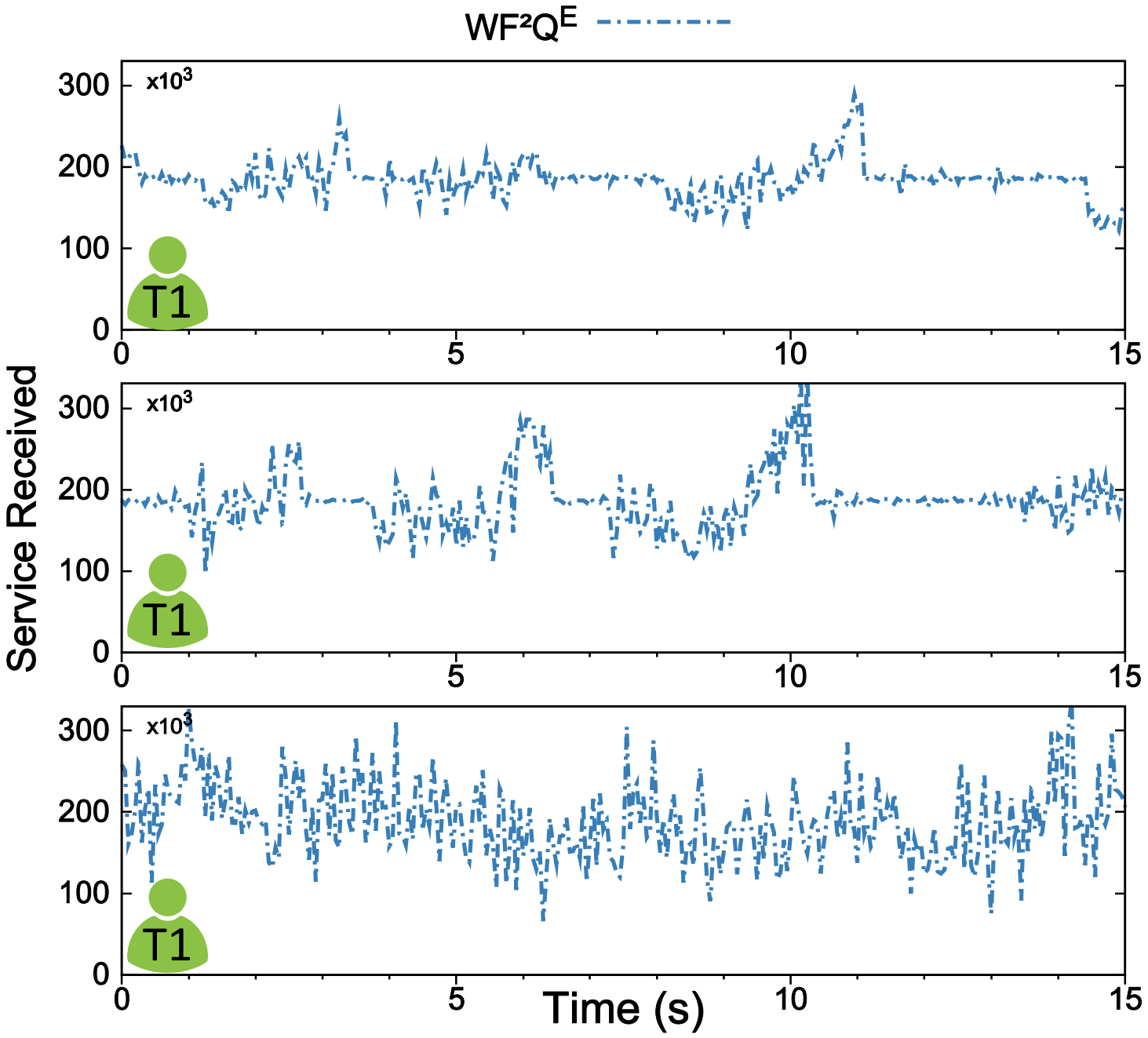
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

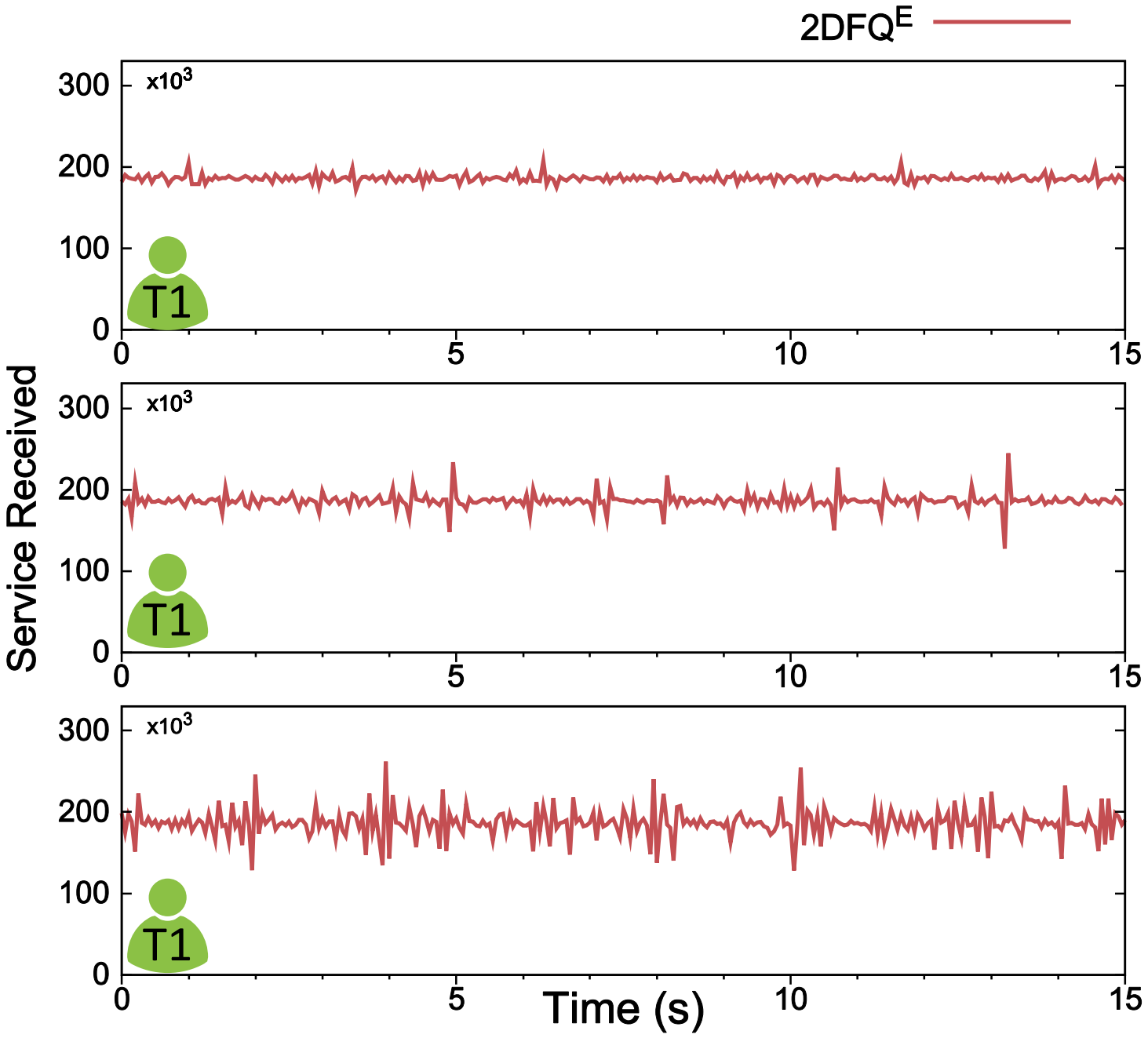
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

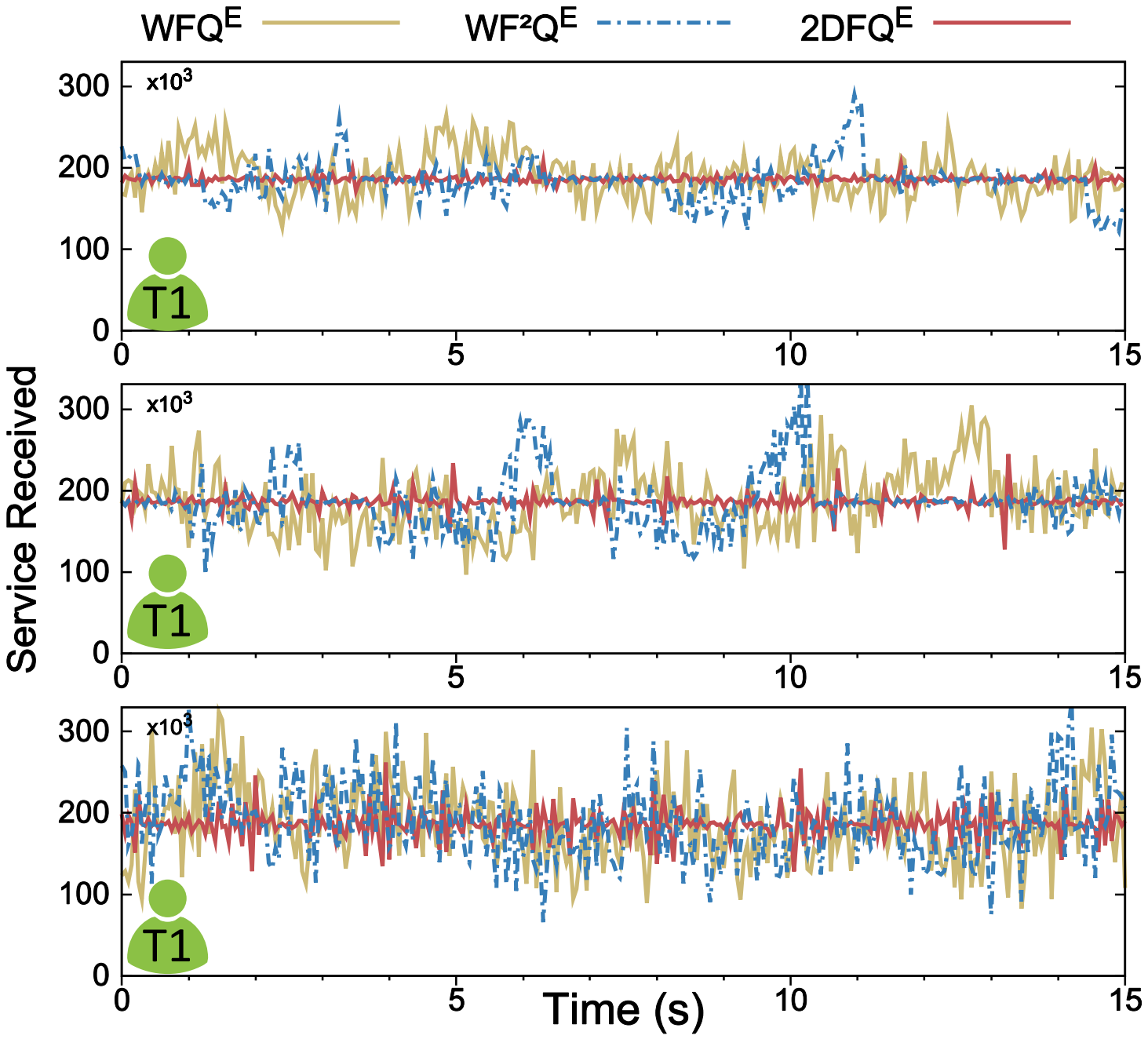
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

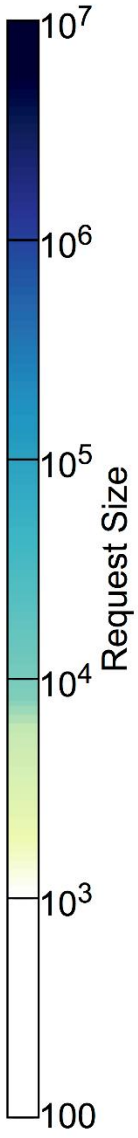
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

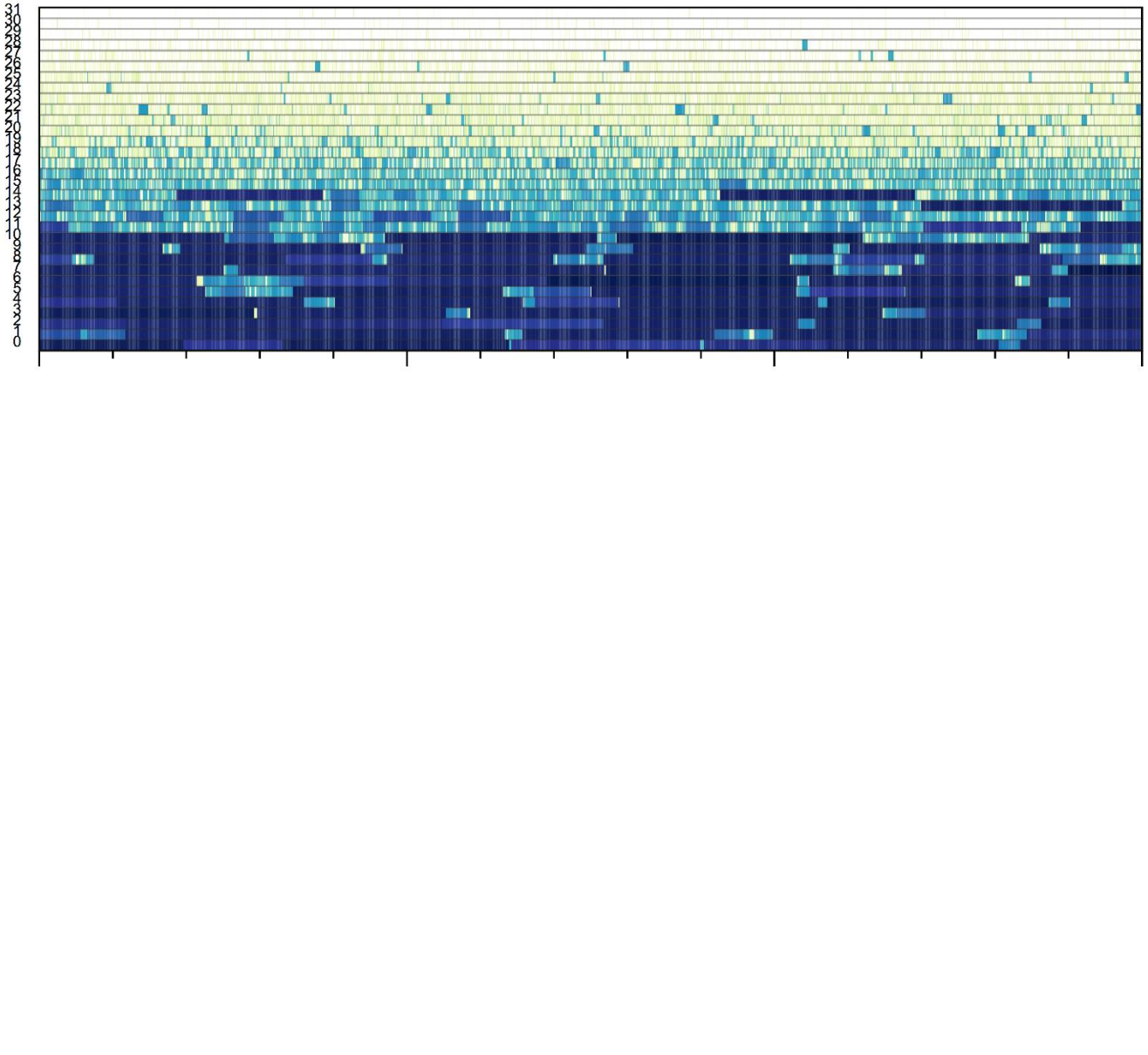
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

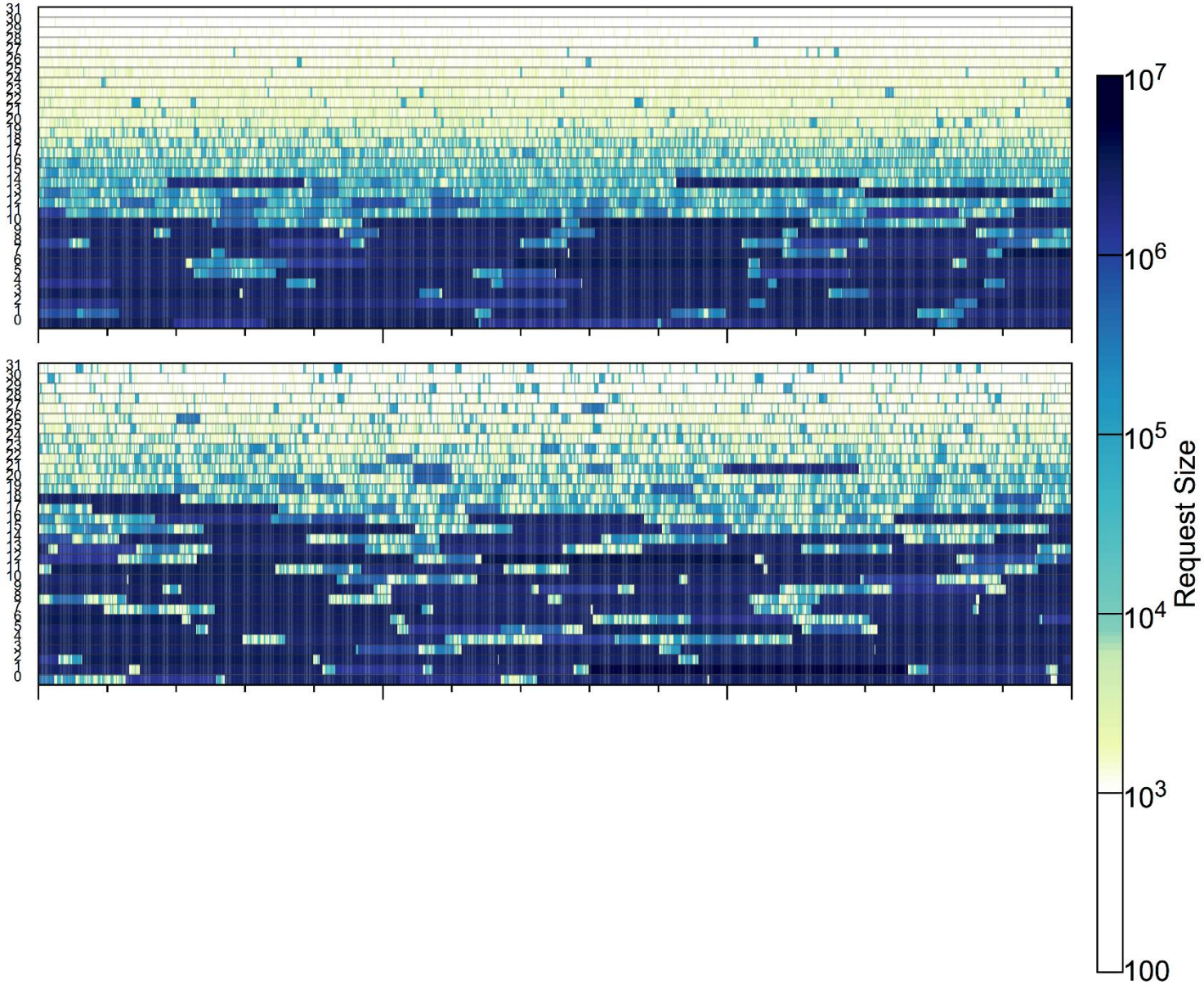
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

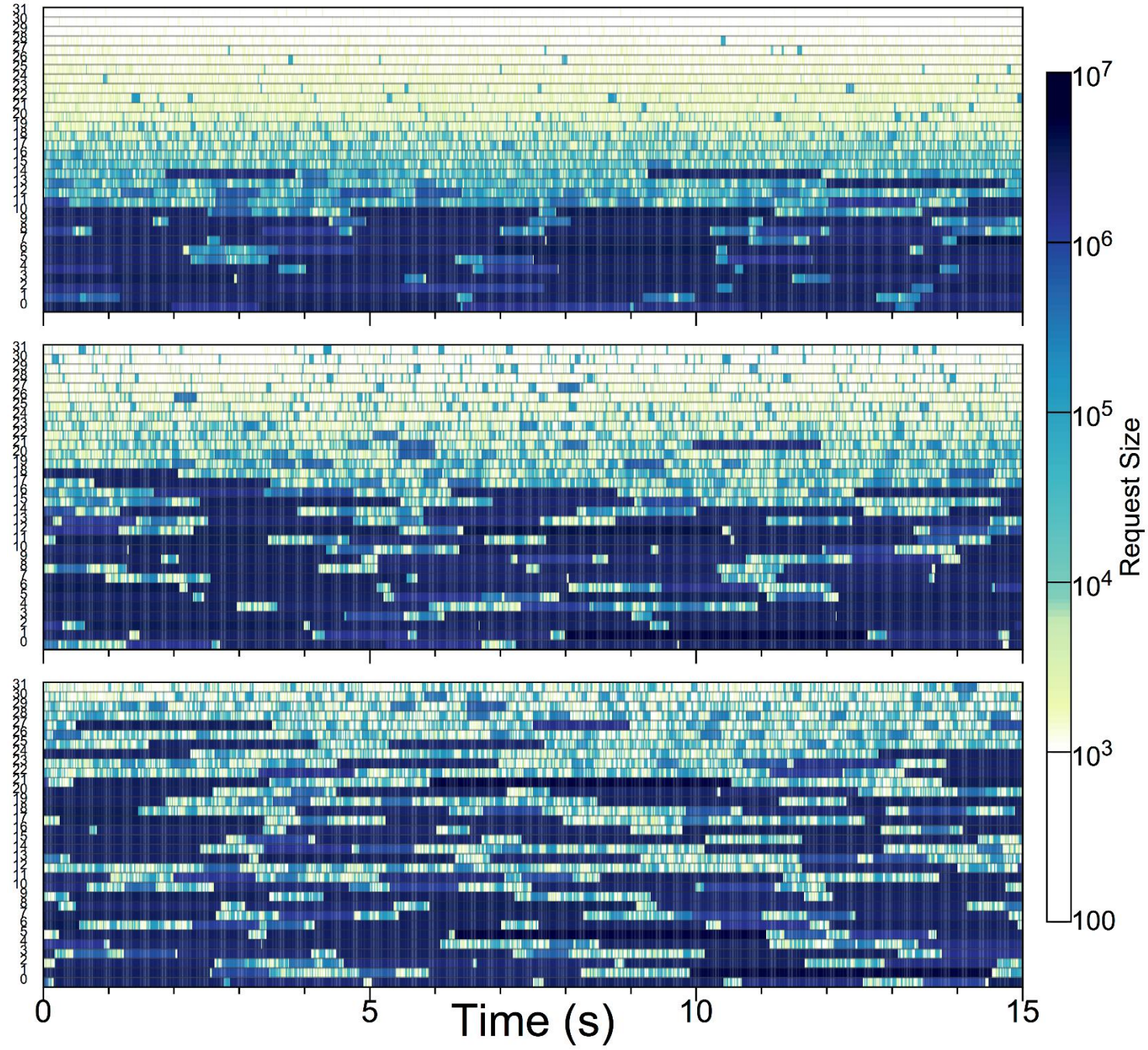
Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable



1

Predictable

2

2/3 predictable
1/3 unpredictable

3

1/3 predictable
2/3 unpredictable

Two-Dimensional Fair Queueing

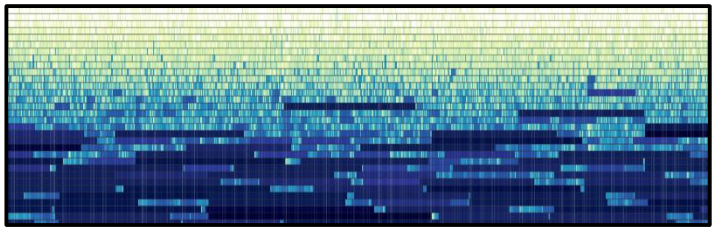
Two-Dimensional Fair Queueing

More threads → Opportunity to reduce burstiness

Two-Dimensional Fair Queueing

More threads → Opportunity to reduce burstiness

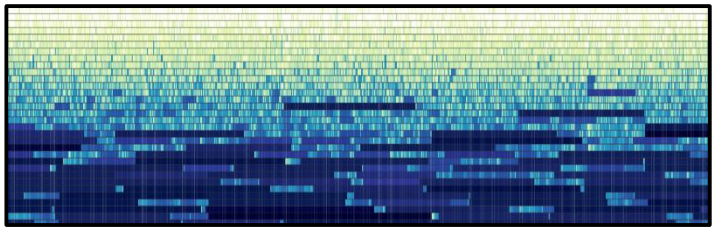
Partitions requests
across threads by size



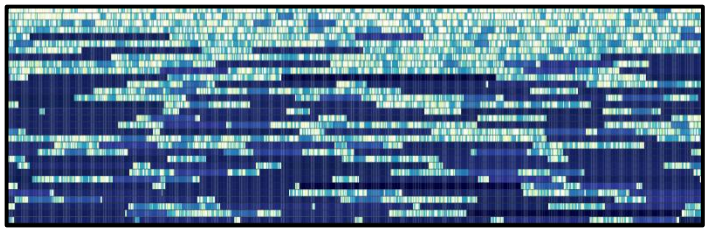
Two-Dimensional Fair Queueing

More threads → Opportunity to reduce burstiness

Partitions requests across threads by size



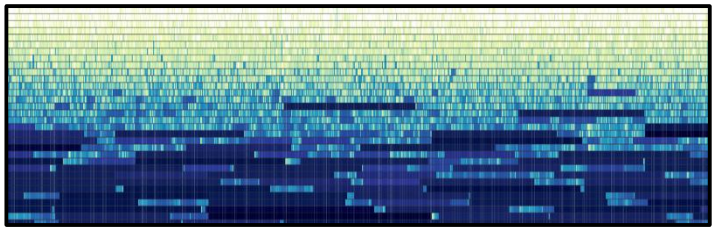
Co-locates unpredictable and expensive workloads



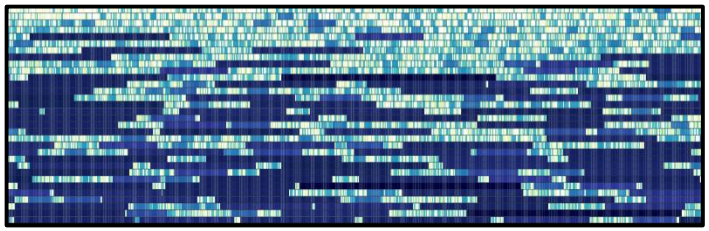
Two-Dimensional Fair Queueing

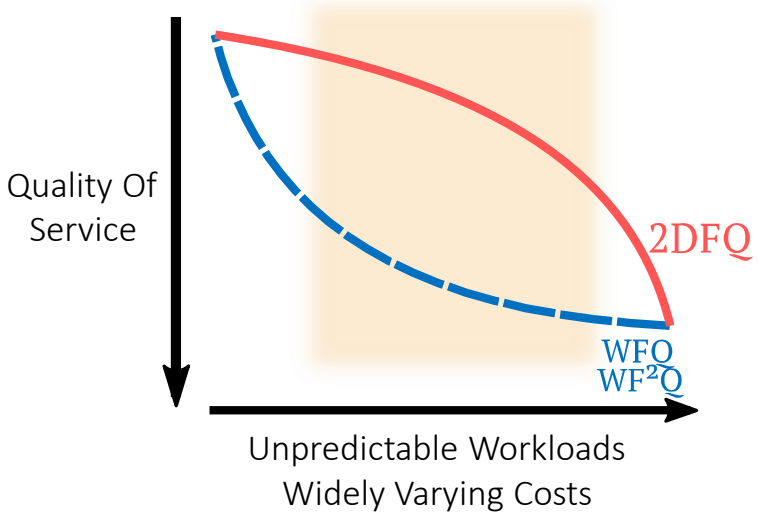
More threads → Opportunity to reduce burstiness

Partitions requests across threads by size



Co-locates unpredictable and expensive workloads

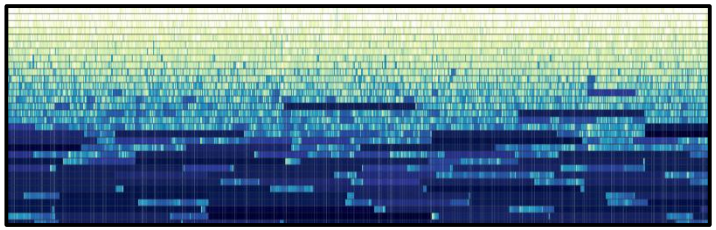




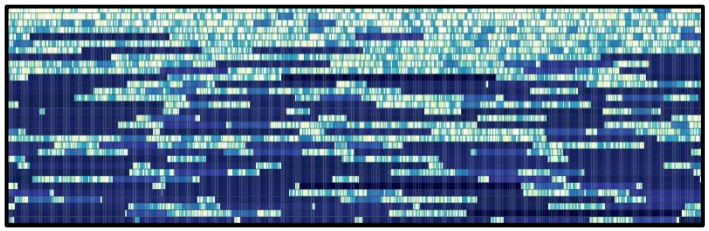
Two-Dimensional Fair Queueing

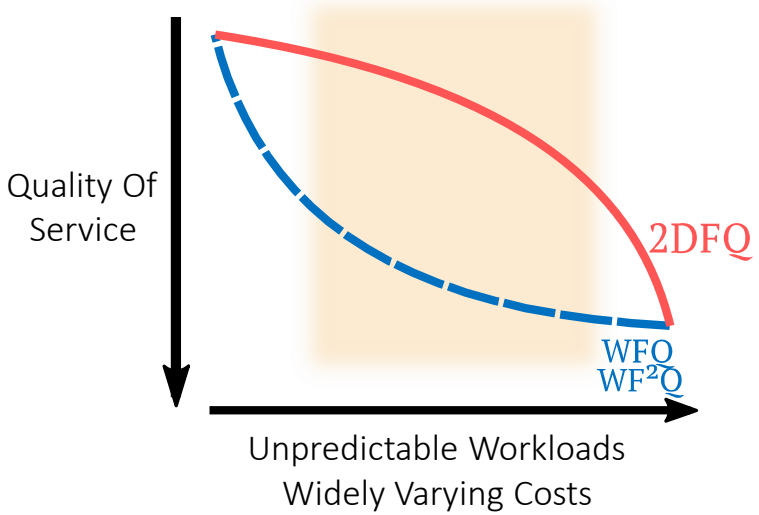
More threads → Opportunity to reduce burstiness

Partitions requests across threads by size



Co-locates unpredictable and expensive workloads





Reduced tail latency

Less burstiness