Fabric Capacities, beyond the obvious

Benni De Jagere
Daan Humblé

Slides

CAT

Microsoft

Benni De Jagere

Principal Program Manager | Azure Data Customer Advisory Team (Azure Data CAT)

Fabric CAT

dataMinds .be Member

/bennidejagere

sessionize /bennidejagere

/bennidejagere

#SayNoToPieCharts

Microsoft

Daan Humblé

Cloud Solutions Architect | Customer Success Unit Cloud & AI

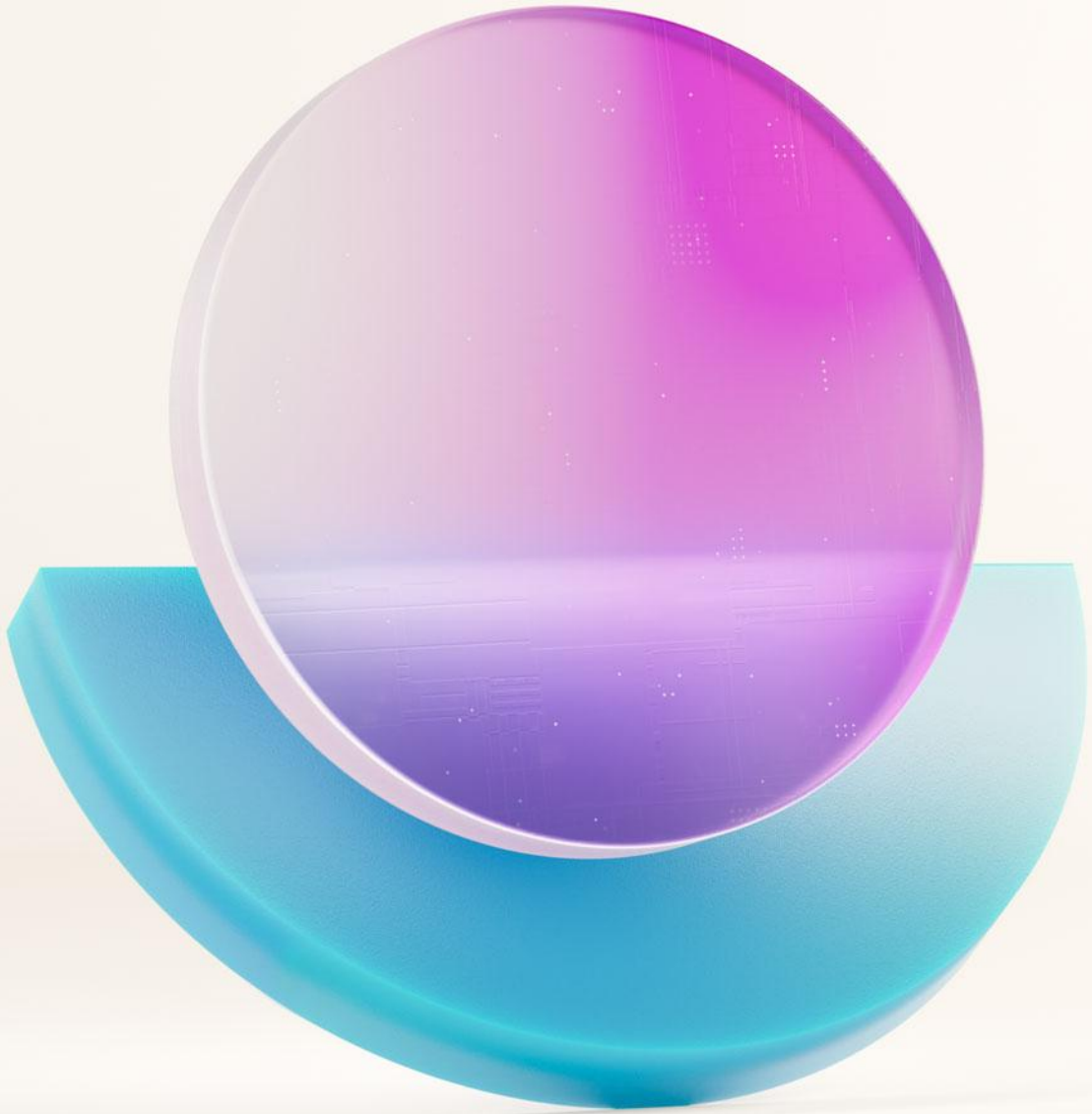
Microsoft NL

/daanhumble

sessionize /daan-humble

/HumbleDaan

Microsoft Fabric
Community Conference



Assumptions

You read the session abstract ..

So you know about bursting, smoothing, and throttling ..

Right? 😼

Fabric Capacity Core Concepts

Scalable Compute

- **Capacity Unit Seconds (CUs)** are the base compute unit for all Capacities
- Your SKU Size determines the number of CUs you have available
- **Multiple workloads** can use the same capacity at the same time

Resizeable, Pausable

- Fabric SKUs (F SKUs) enable **added flexibility**
- **Resize** to increase or decrease the SKU to meet your compute needs
- **Pause** and **Resume** the capacity

Self-Managing

- **Self-Managing** with Bursting and Smoothing
- **Bursting** allows jobs to run at peak performance
- **Smoothing** reduces the impact of spikes in compute

But as with any resource, you still can push them too far (Throttling)

- Capacities offer **built-in resource governance**
- When there's too much smoothed usage, **throttling is applied**
  - **Interactive jobs Delay –** 20s delay, when 10 min > Usage <=60 min
  - **Interactive jobs Rejections –** Rejection, when 60 min > Usage <= 24 hours
  - **Background Rejections –** Rejection, when Usage > 24hrs

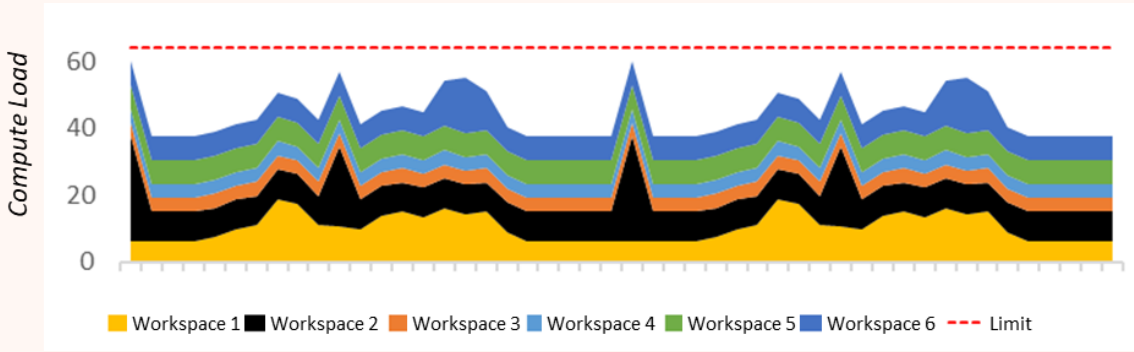
Capacities are a shared resource

Shared across workloads
A single capacity is providing the compute power for all Fabric workloads.

There is no need to allocate compute for each workload separately.

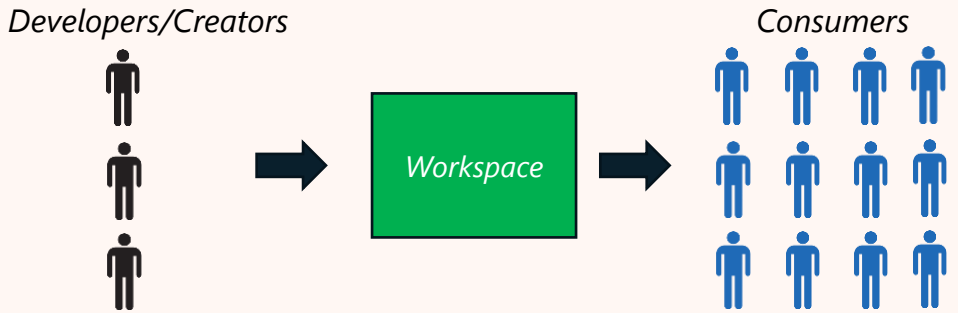
Shared Across Projects
A single capacity typically supports dozens of separate projects simultaneously, each managed in its own workspace.

It is rare to have a capacity dedicated to a single project



Shared across users
For each project, many developers will share a workspace where collaborative development and consumption at scale is managed.

Each creator can provision any artifact and run any job without the need for any pre-approval or planning

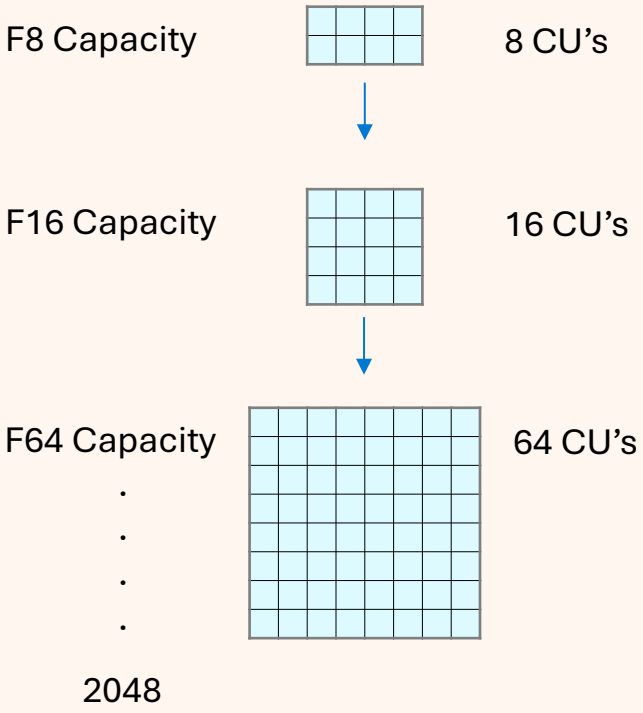


Capacities are flexible building blocks for growth

Capacities can be configured in endless ways to meet scale, usage and governance requirements while tuning to minimize TCO and performance goals

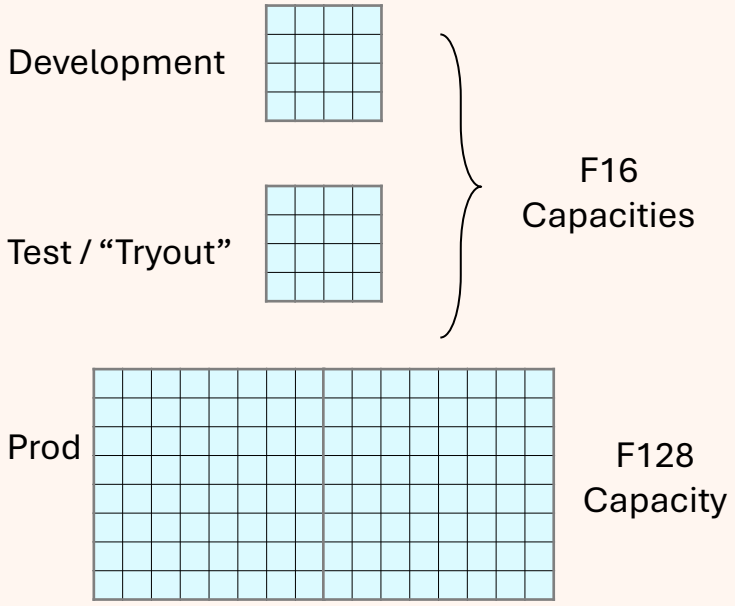
Scale Vertically

Increased capacity size provides more throughput



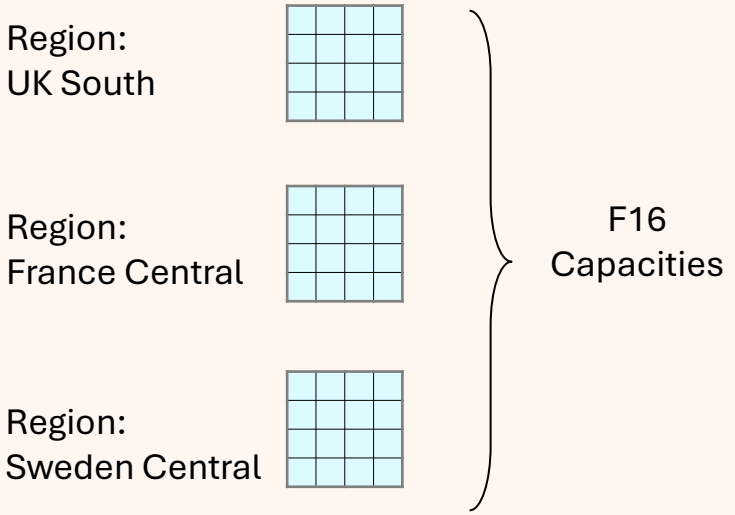
Scale Horizontally

Scale horizontally using the benefits of modular design for hardened isolation and governance



Regional Availability

Use different capacities for different regions to support GDPR / Data residency requirements



Provisioning and Deploying Capacities

Purchased in Azure

- **Purchased** either as a PAYG or RI resource
- **Provisioned with a certain amount of compute** units, analogous to CPU cores.
- The **more capacity units are provisioned, the more load** the capacity can support
  - Multiply SKU size by 30s to match platform evaluation in metrics app
- Capacities are **priced at a fixed hourly rate**, based on capacity units provisioned
- The RI commitment (1-year reserved instance) enjoys a **41% discount**

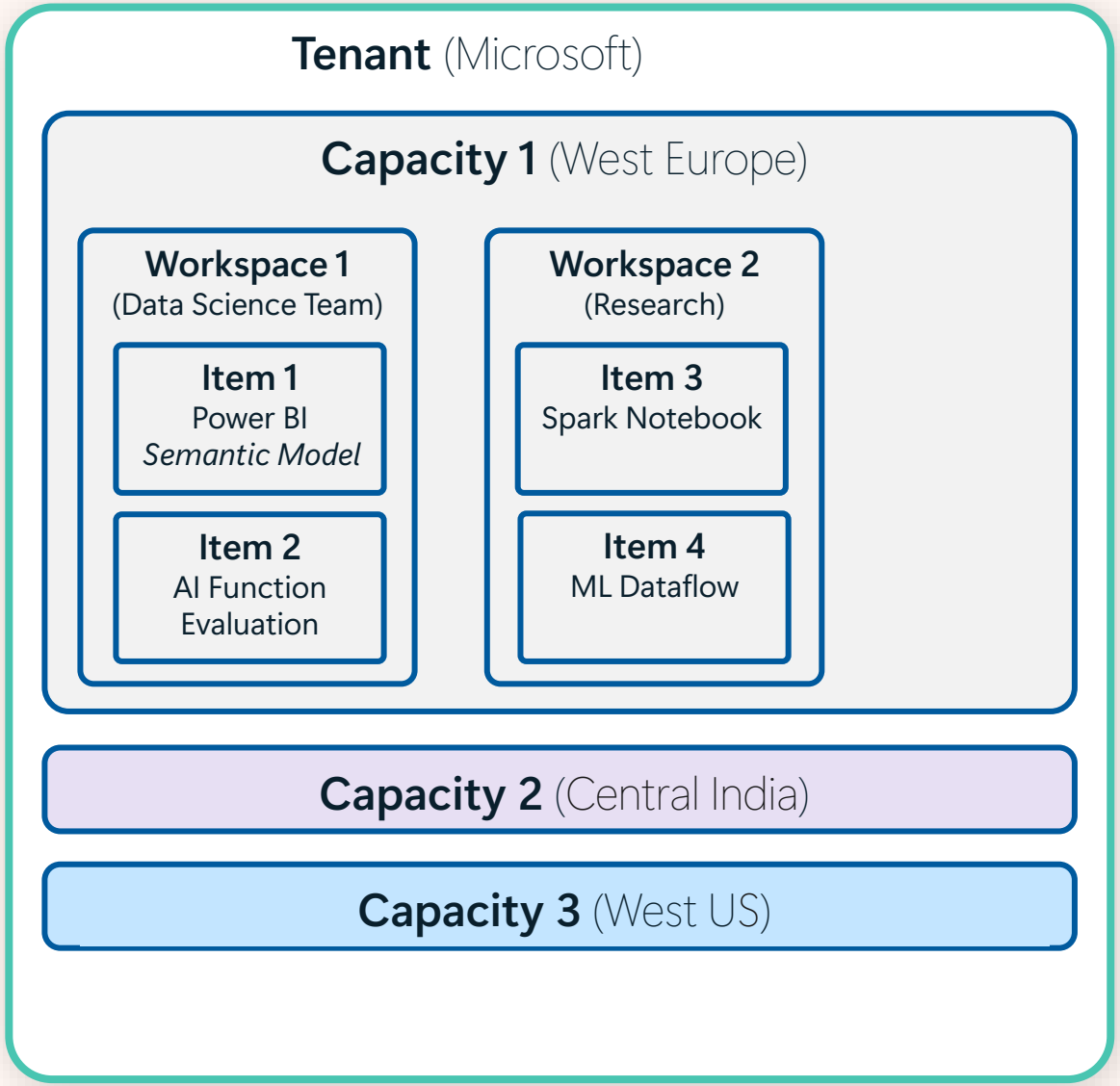
Universal Compute Capacities SKU Sizing

SKU	Capacity Units (CU)	CU's (per 30s)	Power BI SKU	Power BI V-cores
F2	2	60	-	0.25
F4	4	120	-	0.5
F8	8	240	A1	1
F16	16	480	A2	2
F32	32	960	A3	4
F64	64	1920	P1	8
F128	128	3840	P2	16
F256	256	7680	P3	32
F512	512	15360	P4	64
F1024	1024	30720	P5	128
F2048	2048	61440	-	256

Provisioning and Deploying Capacities

Deployed to Regions

- Each capacity **resides in a specific region of the buyers' choice** where both the data & compute reside
- **Workspaces are assigned to a capacity** that provides the compute and storage for all the workspace artifacts
- Multiple capacities can be purchased, deployed and managed by **different owners** residing in a single tenant allowing each business unit to pay for their own consumption

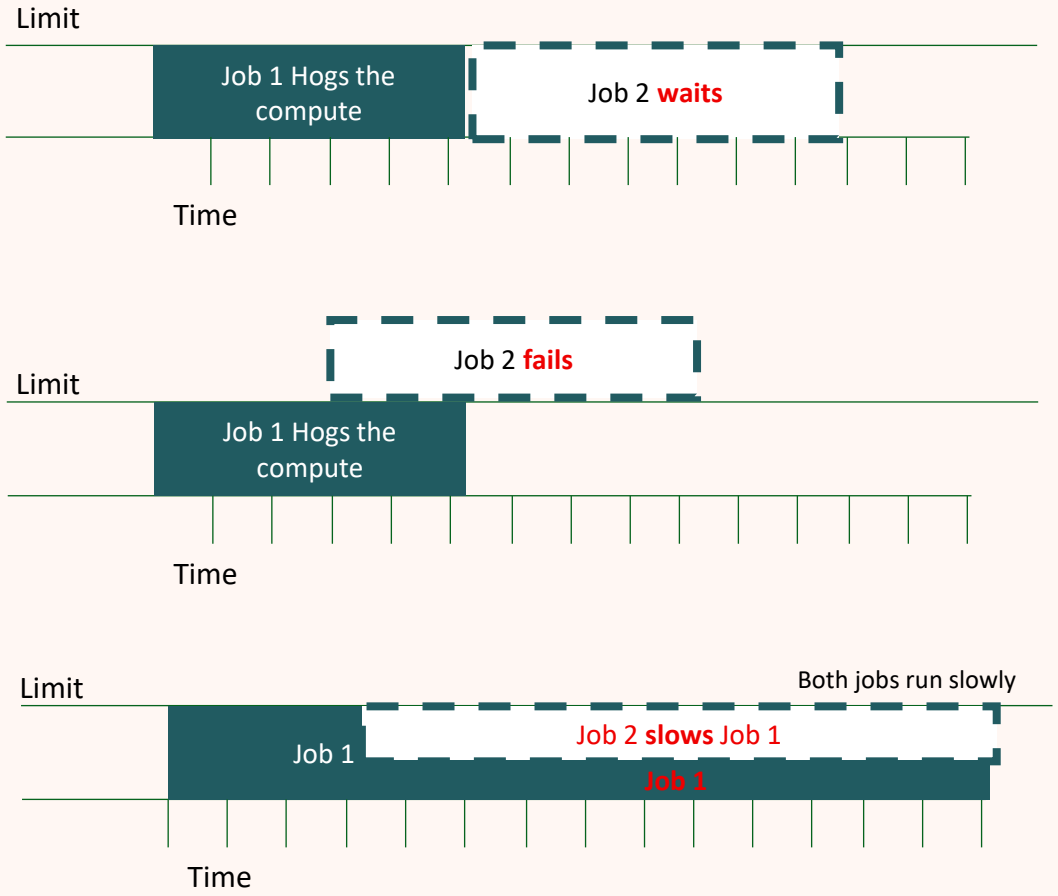


Microsoft Fabric
COMMUNITY CONFERENCE

Bursting and Smoothing

Bursting and smoothing | before and after

‘Traditional’ Systems stay within a hard compute limit

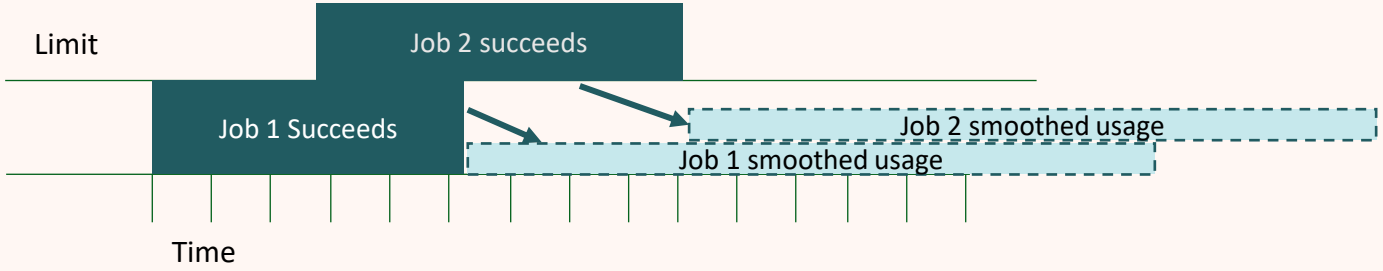


Bursting and smoothing | before and after

Bursting and smoothing help to solve these problems

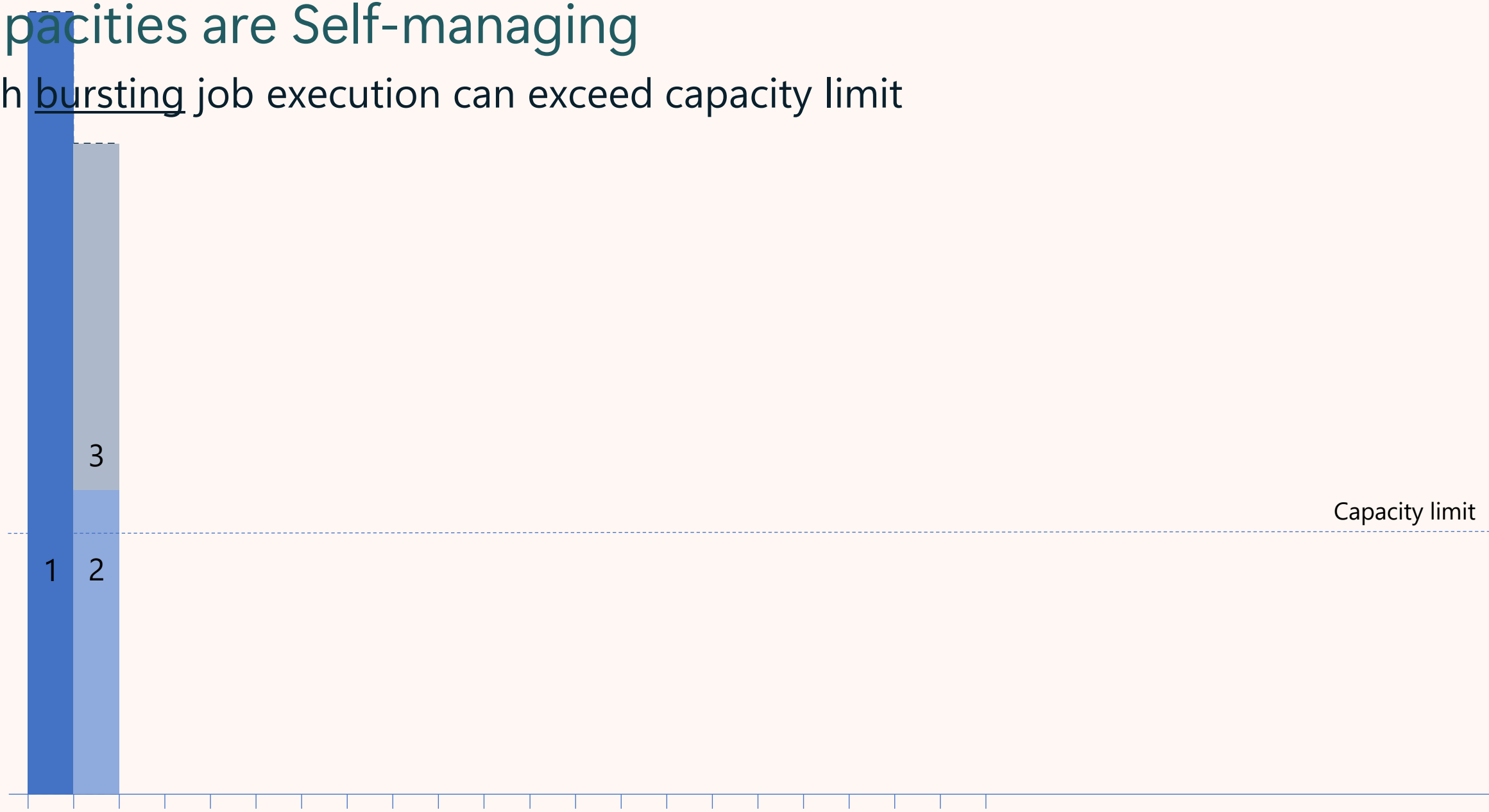
Bursting lets both jobs **succeed**, running at **peak performance** to finish fast.

Smoothing distributes the **cost** of jobs over a longer time-period preventing scheduling issues.



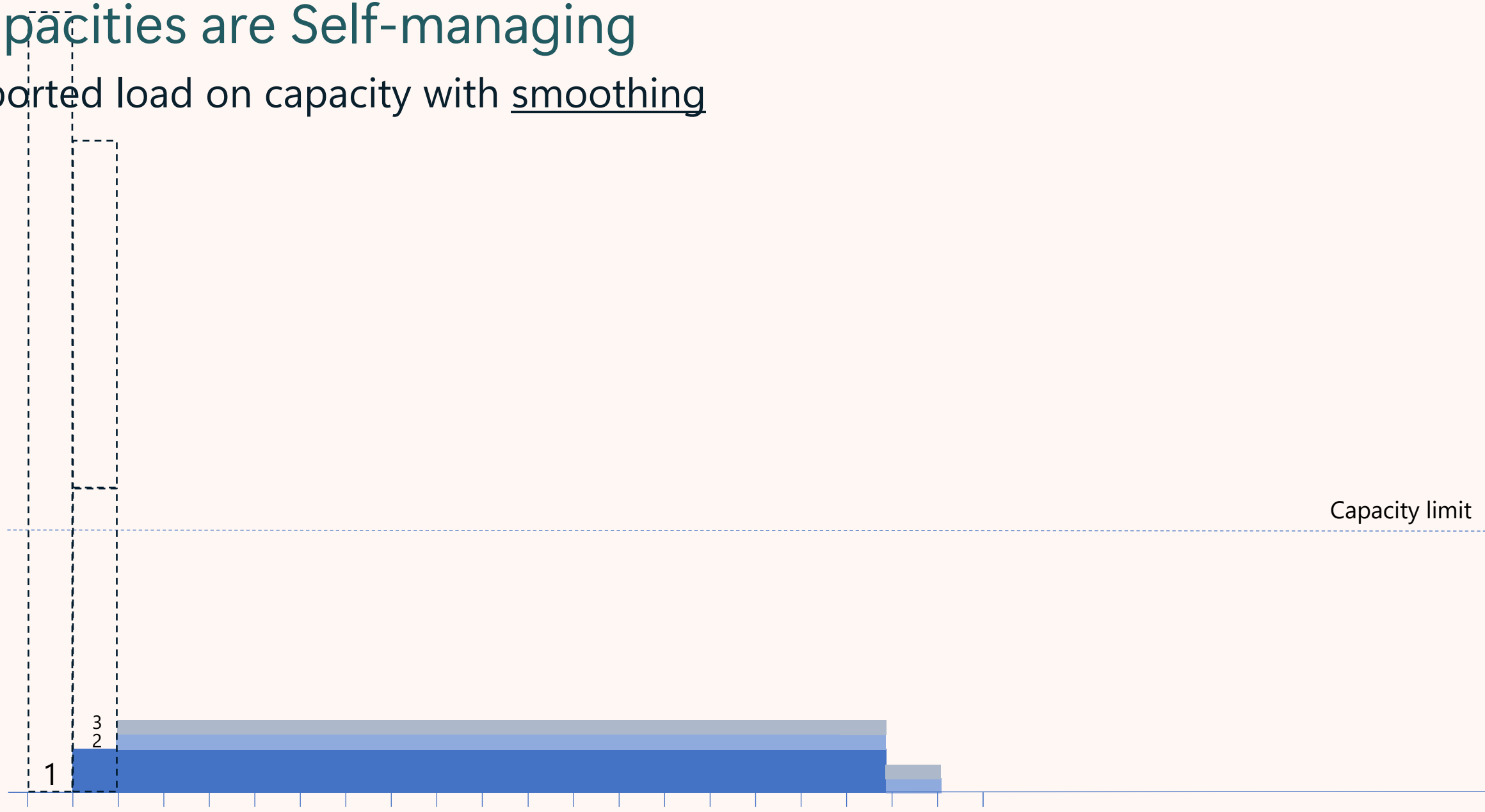
Capacities are Self-managing

With bursting job execution can exceed capacity limit



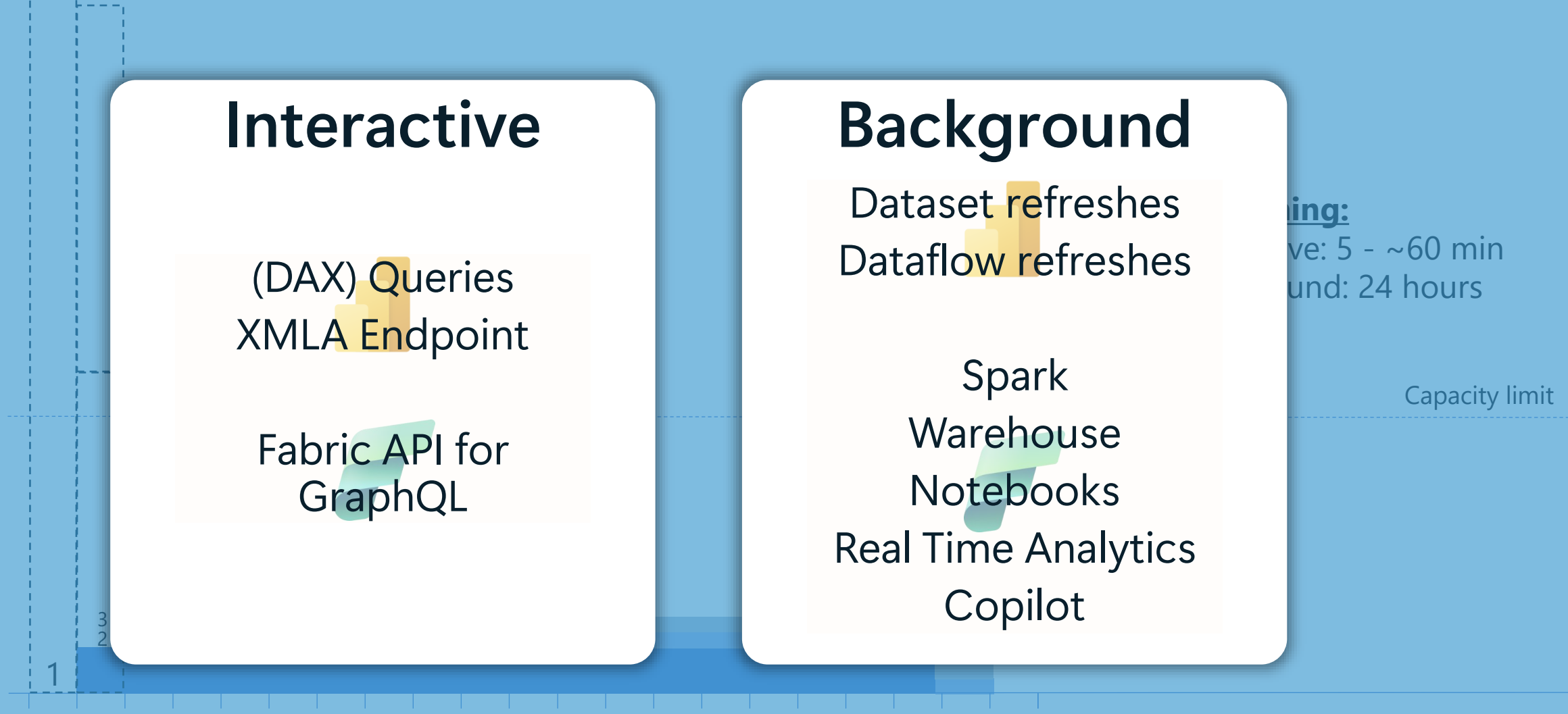
Capacities are Self-managing

Reported load on capacity with smoothing



Capacities are Self-managing

Reported load on capacity with smoothing



Interactive

(DAX) Queries
XMLA Endpoint

Fabric API for GraphQL

Background

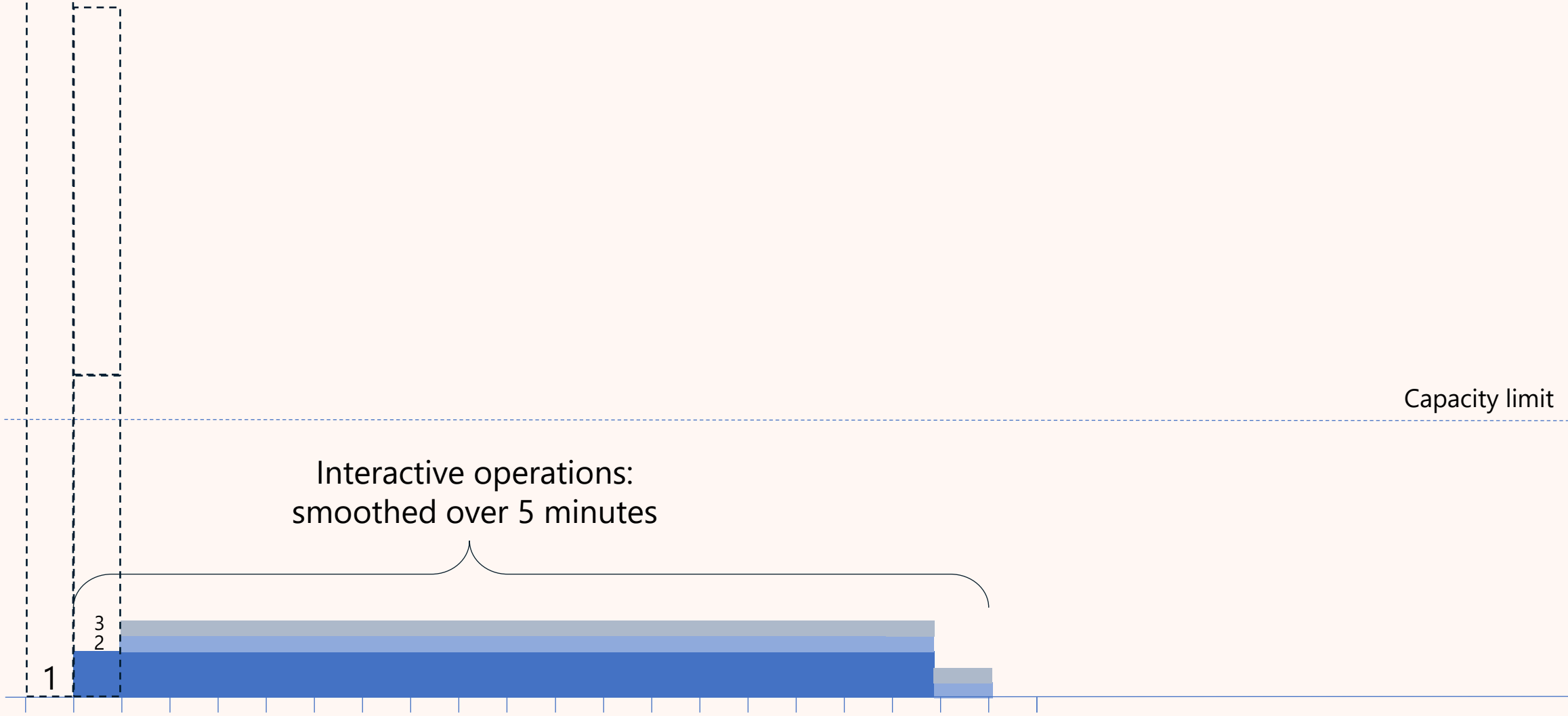
Dataset refreshes
Dataflow refreshes

Spark
Warehouse
Notebooks
Real Time Analytics
Copilot

https://learn.microsoft.com/en-us/fabric/enterprise/fabric-operations

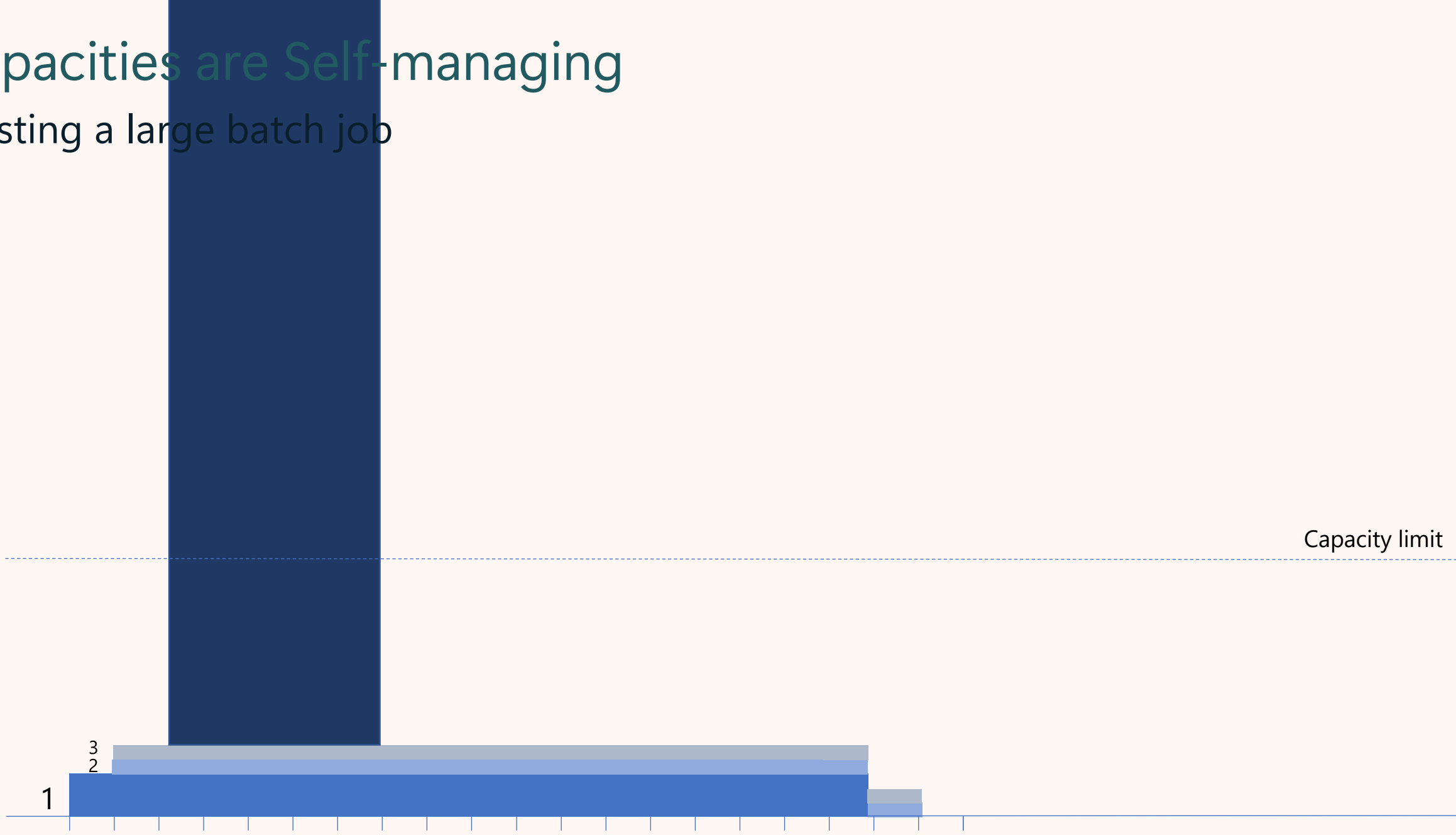
Capacities are Self-managing

Reported load on capacity with smoothing



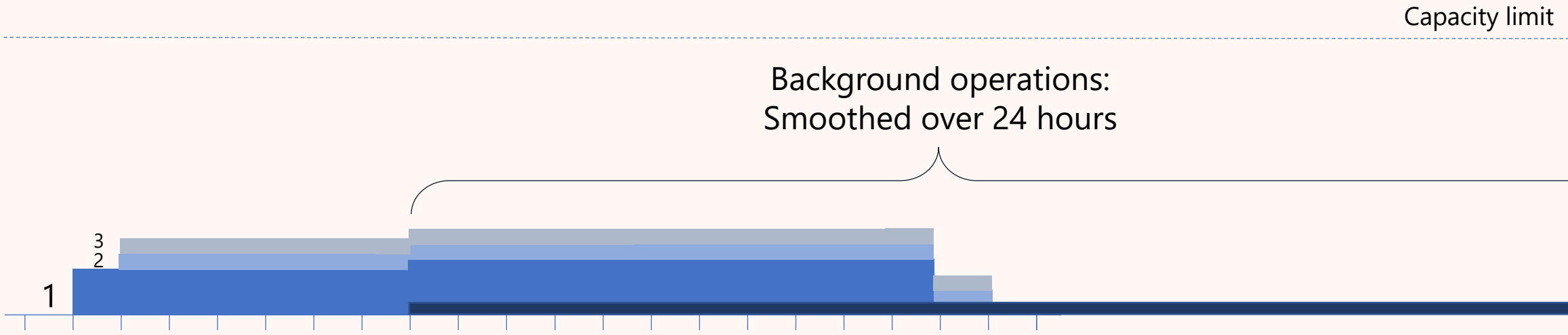
Capacities are Self-managing

Bursting a large batch job



Capacities are Self-managing

Smoothing a large batch job

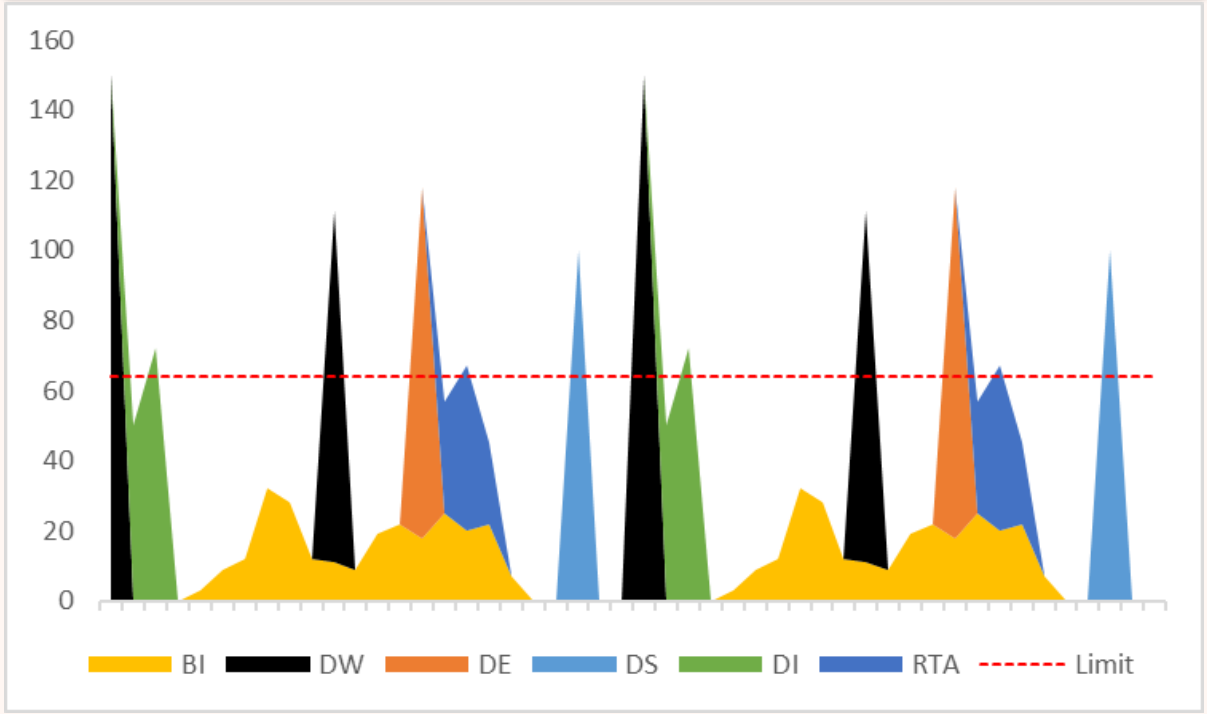


Bursting and smoothing | before and after

Looking at an example of a 64 CU capacity, running multiple workloads over a couple of days…

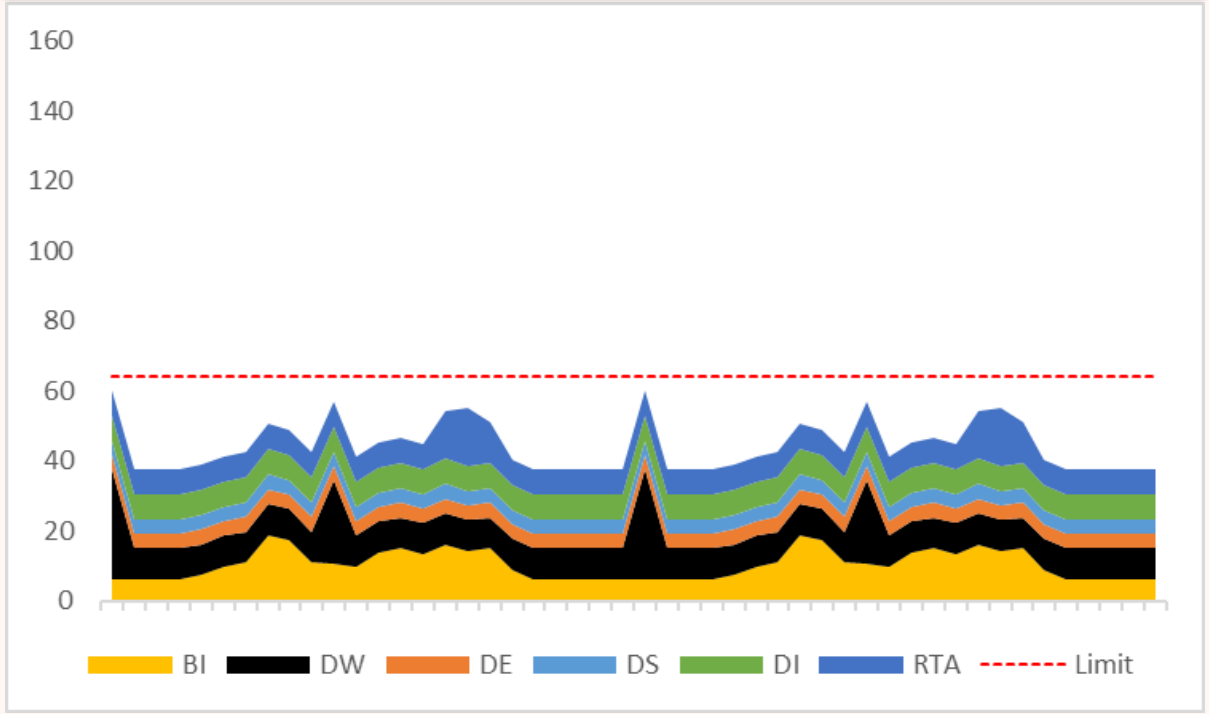
Before Smoothing

- Actual load as executed on the capacity before smoothing
- *Bursting* accelerates jobs execution by resource boosting
- The capacity could be overloaded 25% of the time
- Some of the overloads are more than 2x the limit
- There are periods of no/low usage



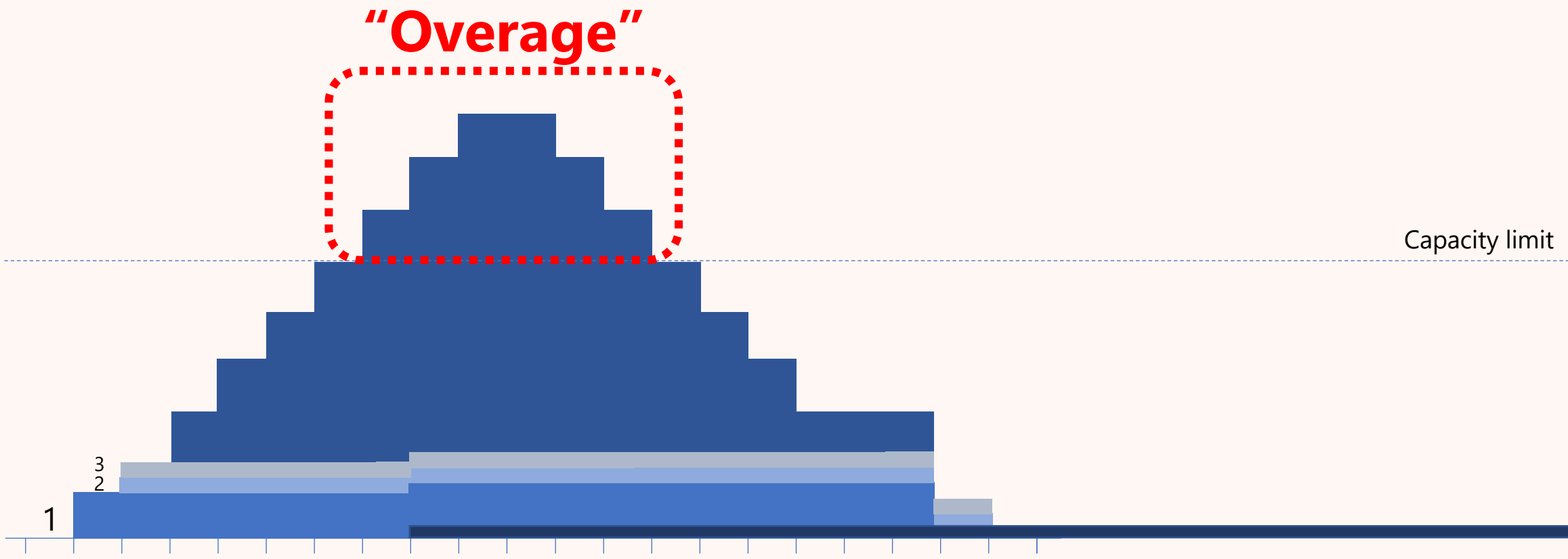
After Smoothing

- Shows the reported load (not runtime execution) against the capacity limits
- There is NO overload, and consumption is more stable
- The smoothing of usage fills in gaps



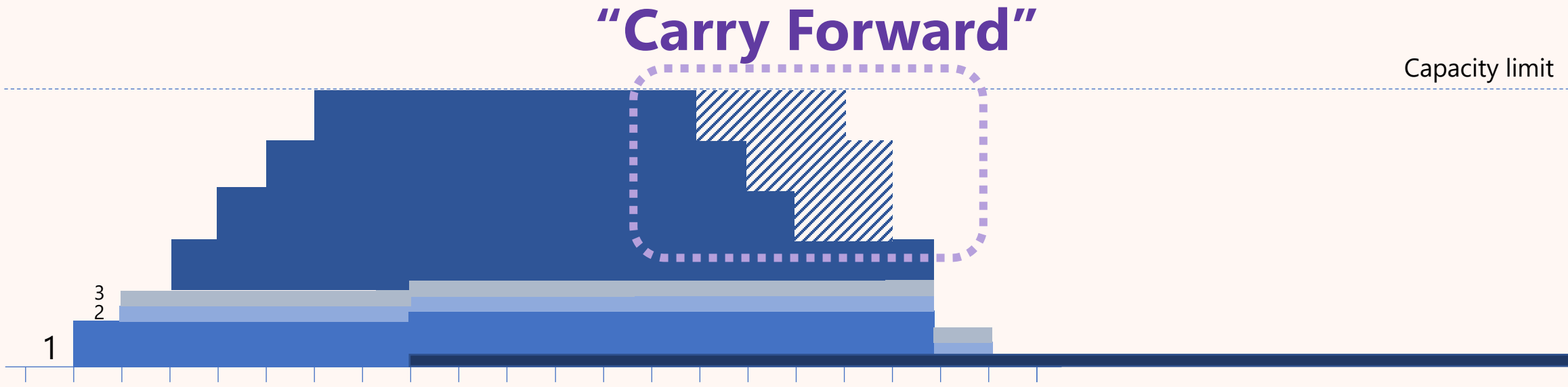
Capacities are Self-managing

What if you keep going over the limit?



Capacities are Self-managing

What if you keep going over the limit?



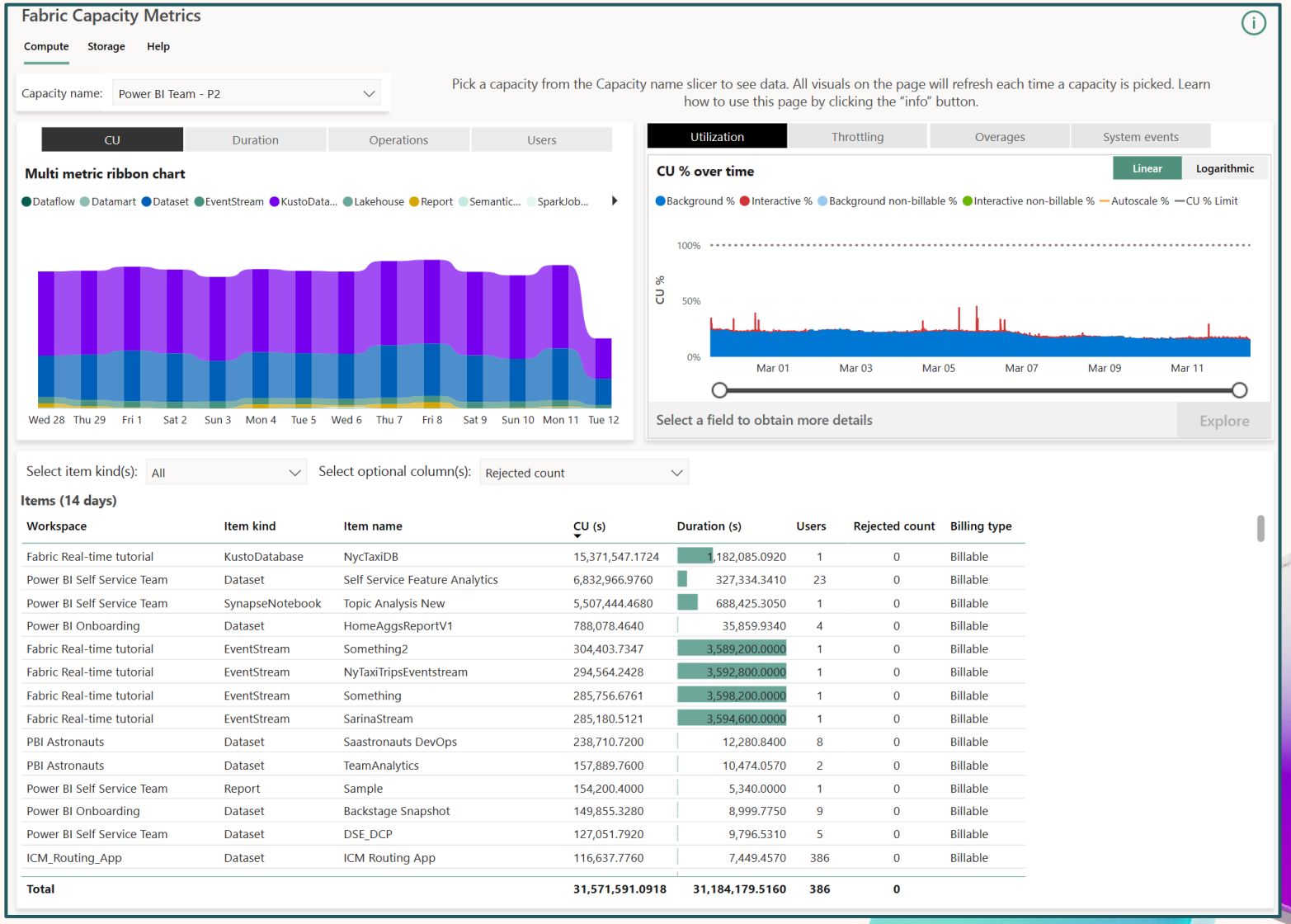
Microsoft Fabric
COMMUNITY CONFERENCE

Monitoring with Capacity Metrics

Capacity Metrics

Monitor Capacities and Plan capacity scale-up with confidence

- Tenant wide visibility into capacity usage for all Fabric experiences
- Identify resource usage trends and their impact to autoscale & throttling
- View preview workload usage alongside production workloads to make data-driven capacity sizing decisions



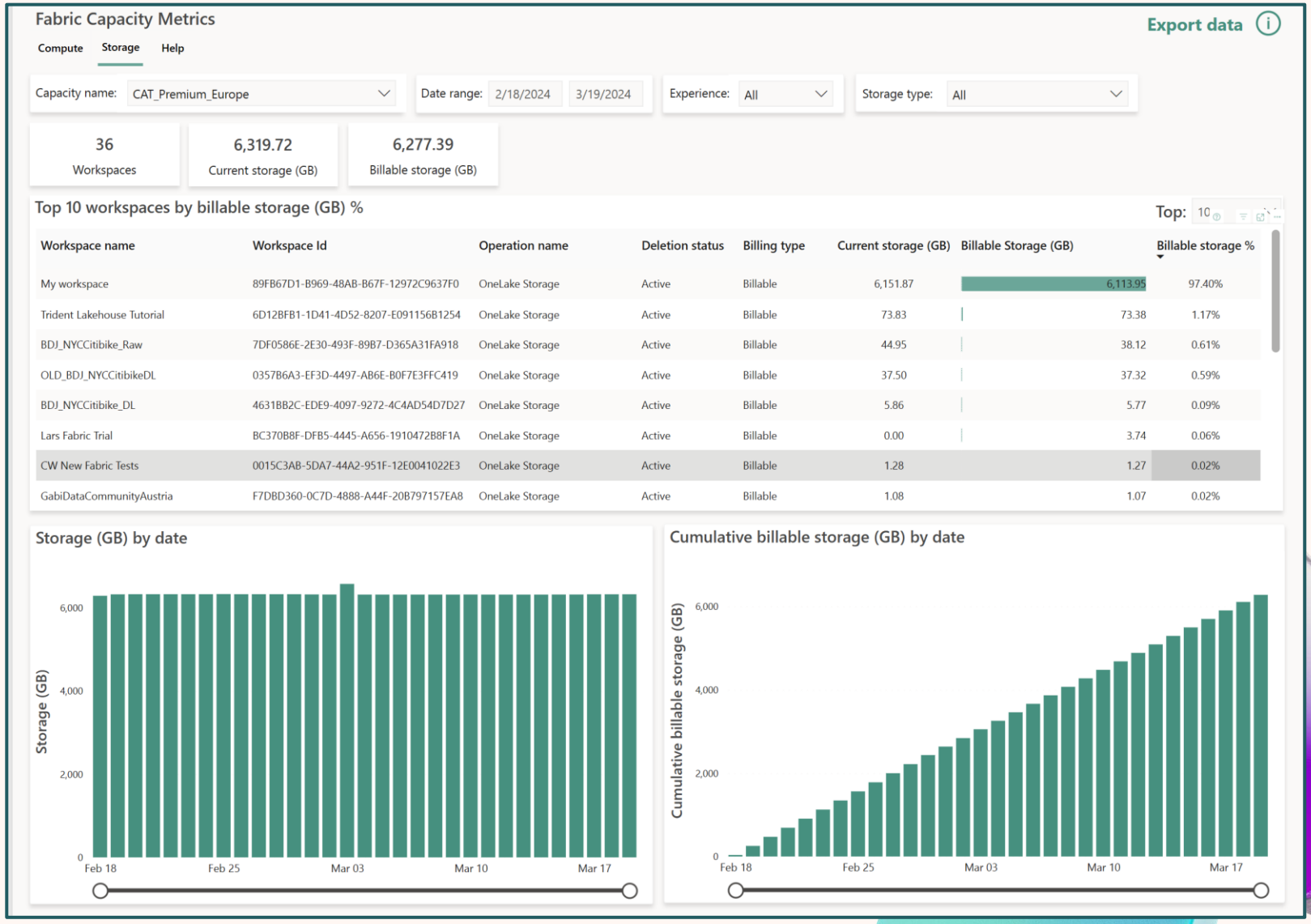
Workspace	Item kind	Item name	CU (s)	Duration (s)	Users	Rejected count	Billing type
Fabric Real-time tutorial	KustoDatabase	NycTaxiDB	15,371,547.1724	1,182,085.0920	1	0	Billable
Power BI Self Service Team	Dataset	Self Service Feature Analytics	6,832,966.9760	327,334.3410	23	0	Billable
Power BI Self Service Team	SynapseNotebook	Topic Analysis New	5,507,444.4680	688,425.3050	1	0	Billable
Power BI Onboarding	Dataset	HomeAggsReportV1	788,078.4640	35,859.9340	4	0	Billable
Fabric Real-time tutorial	EventStream	Something2	304,403.7347	3,589,200.0000	1	0	Billable
Fabric Real-time tutorial	EventStream	NyTaxiTripsEventstream	294,564.2428	3,592,800.0000	1	0	Billable
Fabric Real-time tutorial	EventStream	Something	285,756.6761	3,598,200.0000	1	0	Billable
Fabric Real-time tutorial	EventStream	SarinaStream	285,180.5121	3,594,600.0000	1	0	Billable
PBI Astronauts	Dataset	Saastronauts DevOps	238,710.7200	12,280.8400	8	0	Billable
PBI Astronauts	Dataset	TeamAnalytics	157,889.7600	10,474.0570	2	0	Billable
Power BI Self Service Team	Report	Sample	154,200.4000	5,340.0000	1	0	Billable
Power BI Onboarding	Dataset	Backstage Snapshot	149,855.3280	8,999.7750	9	0	Billable
Power BI Self Service Team	Dataset	DSE_DCP	127,051.7920	9,796.5310	5	0	Billable
ICM_Routing_App	Dataset	ICM Routing App	116,637.7760	7,449.4570	386	0	Billable
Total			**31,571,591.0918**	**31,184,179.5160**	**386**	**0**	

Capacity Metrics

Monitor OneLake consumption

Measure the trends of workspace storage consumption against capacity limits, by day or hour

Reconcile costs with internal chargeback processes



Fabric Capacity Metrics

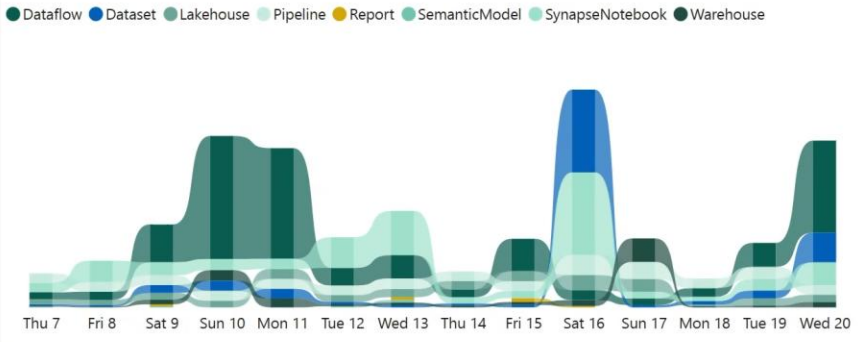
Compute Storage Help

Capacity name: CAT_Premium_US

Pick a capacity from the Capacity name slicer to see data. All visuals on the page will refresh each time a capacity is picked. Learn how to use this page by clicking the "info" button.

CU | Duration | Operations | Users

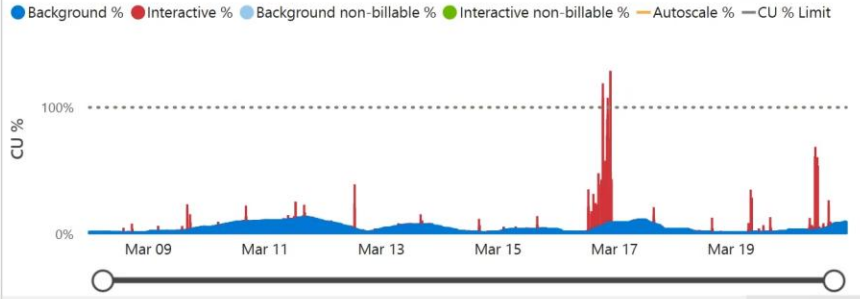
Multi metric ribbon chart



Utilization | Throttling | Overages | System events

CU % over time

Linear | Logarithmic



Select a field to obtain more details | Explore

Select item kind(s): All

Select optional column(s): Multiple selections

Items (14 days)

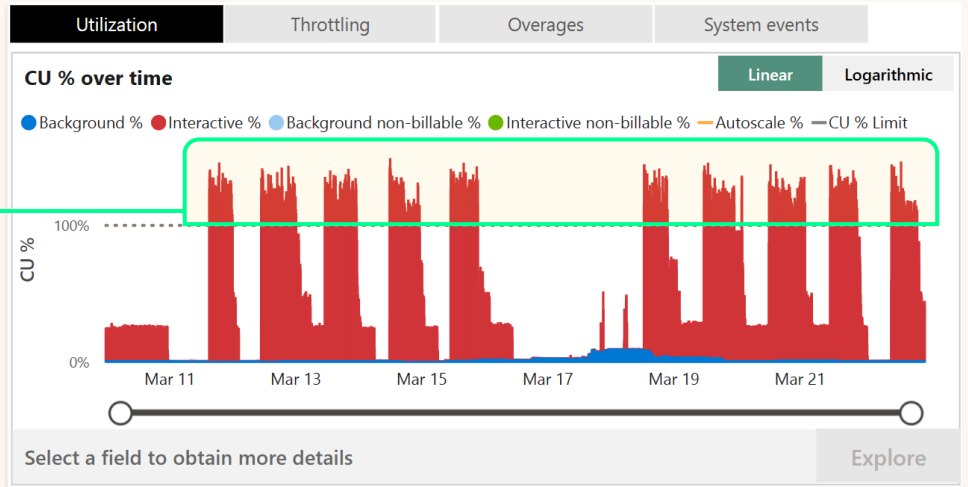
Workspace	Item kind	Item name	CU (s)	Duration (s)	Users	Rejected count	Overloaded minutes	Successful count	Performance delta	Billing type
Fabric for Power BI users (alpowe...	Dataflow	Dataflow 1Nashville	902,712.8000	14,585.3890	1	0	0.0000	97		Billable
Data Factory in an Hour (old)	Pipeline	dfPipeline	475,200.0000	19,858.7140	1	0	0.0000	1247		Billable
Analytics at the speed of Direct L...	Dataset	Metrics	359,352.5440	11,100.0940	3	0	1.3333	1095		Billable
Day After DIAD (alpowers)	Dataflow	OnlineSalesDataflow	355,036.3520	14,154.7430	1	0	0.0000	12		Billable
DADIAD Sample	Dataflow	DataflowTest	343,605.4880	15,609.4410	1	0	0.0000	12		Billable
Survey_test	SynapseNotebook	Survey_Notebook	214,899.6685	24,328.8260	2	0	0.0000	42		Billable
Mahoney_MetricsSnapshots	SynapseNotebook	MetricsAppSnapshots	184,471.7595	20,047.6250	1	0	0.0000	0		Billable
Analytics at the speed of Direct L...	SynapseNotebook	Notebook 1	165,965.6310	19,895.5380	2	0	0.0000	0		Billable
BDJ_NYCCitibike_Base	Warehouse	NYCCitibike_BASE	87,065.7860	50,126.5260	3	0	0.0000	3164		Billable
Lars - ready for demo	SynapseNotebook	Framing Demo	86,681.3000	8,158.5510	1	0	0.0000	0		Billable
CW Direct Lake Paging	Dataflow	CollectDMVData	83,097.0720	7,427.1870	1	0	0.0000	52		Billable
Fabric_Demo	Dataflow	lh_Sample	82,095.6800	4,454.1890	2	0	0.0000	73		Billable
Fabric for Power BI users (alpowe...	Dataset	WeirdDemo	77,008.5280	1,806.6590	2	0	0.0000	540		Billable
Mahoney_MetricsSnapshots	SynapseNotebook	Notebook 1	63,871.2960	7,408.3180	1	0	0.0000	0		Billable
Total			**4,907,150.8628**	**461,929.2860**	**4**	**0**	**2.0000**	**1662399**		

Microsoft Fabric
COMMUNITY CONFERENCE

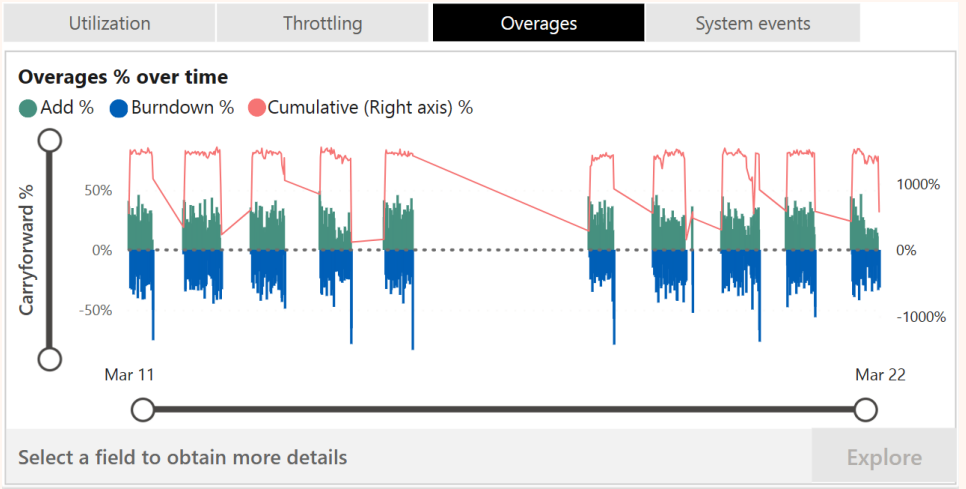
Capacity Throttling Policies

Throttling intro

- Throttling is the platform policy for managing consumption that exceeds throughput is provided by SKU choice
- When workloads exceed the throughput of a capacity a cumulative debt is tracked to be burned down
- Cumulative debt is used to determine throttling policies and is burned down when resources are free

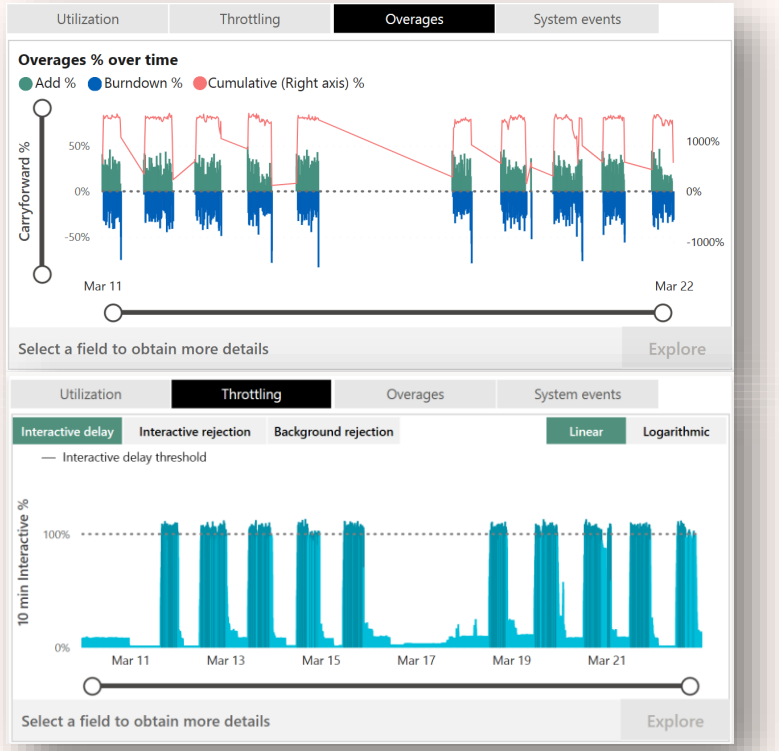


Overage Operation	Description
Overages - Added	• Timepoint when job requests exceed the throughput of a capacity, overages are was added to the cumulative buffer to burn down. • This graph simplifies identification of the optimal timepoint to load timepoint drill to analyze the user operations that contributed to an overage.
Overages - Burndown	• Overages being reconciled when future capacity is free to burn down
Overages - Cumulative	• The total amount of queued work on the capacity to be burned down in the future when the capacity is not fully utilized



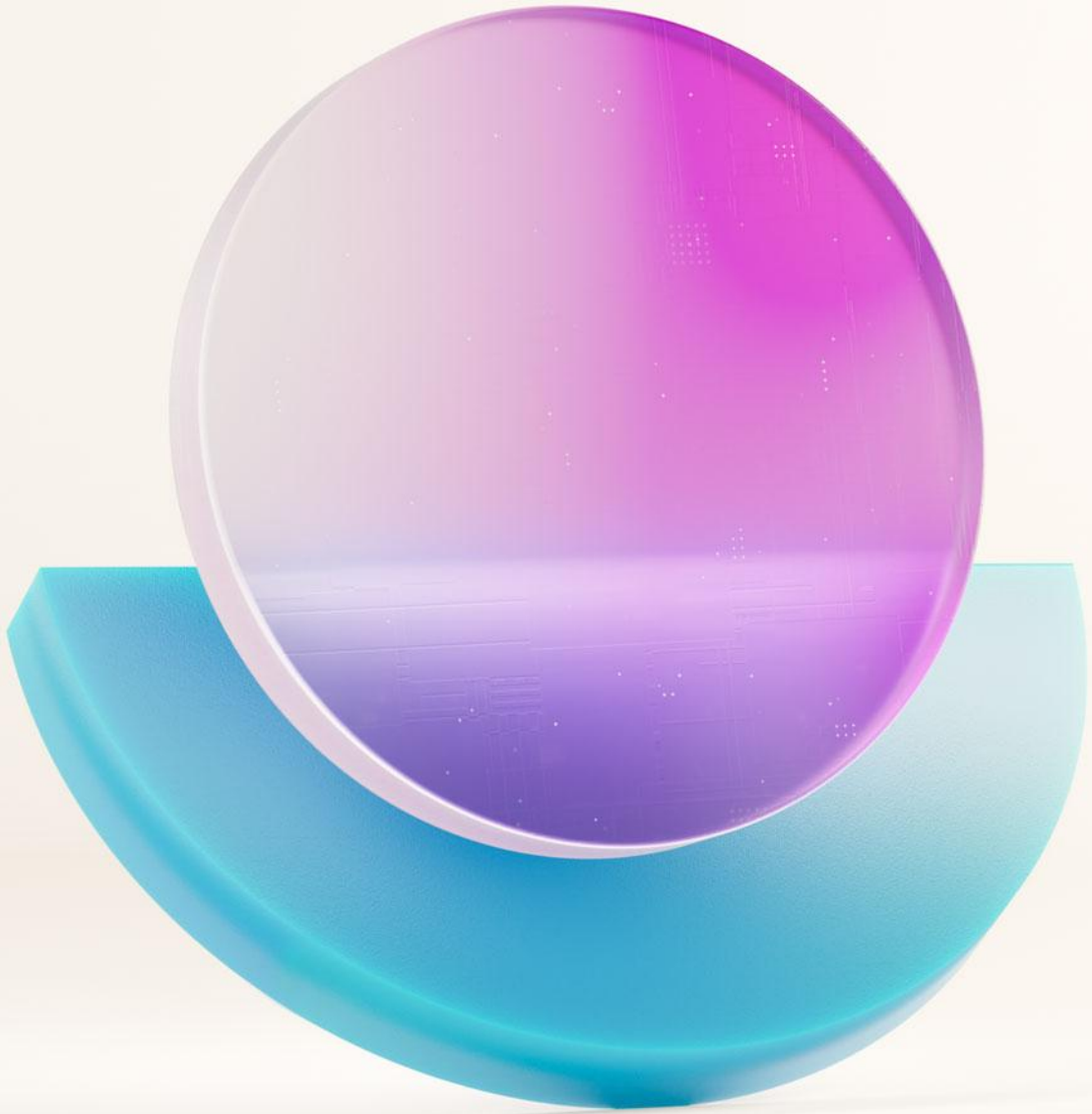
Capacity throttling evolution for Fabric

- For Fabric, throttling policies were refined to deliver multiple benefits
  - **Reduced throttling** for capacities that only experience occasional spikes
  - **Added overage protection** – rejection policies prevent overloaded capacities from irrecoverable overload
  - **Optimizations for long-running jobs:** We're optimizing the platform for long-running jobs, so if a job exceeds capacity limits, it will run to completion and the overage will be burned down against future capacity



Smoothed Capacity - Future Use	Platform Policy	Customer Impact
<= 10m	Overage Protection	Jobs can consume 10 minutes of future capacity use without throttling
> 10m → <= 60m	Interactive Delay	User requested interactive type jobs will be throttled
> 60m → <= 24h	Interactive Rejection	User requested interactive type jobs will be rejected
> 24h	Background Rejection	User Scheduled background jobs will be rejected from execution

Microsoft Fabric
COMMUNITY CONFERENCE



Capacity Planning

https://aka.ms/FabricCapacityPlanning

Fabric SKU Estimator

GA

Enabling customers to better estimate their SKU before purchase

Online calculator

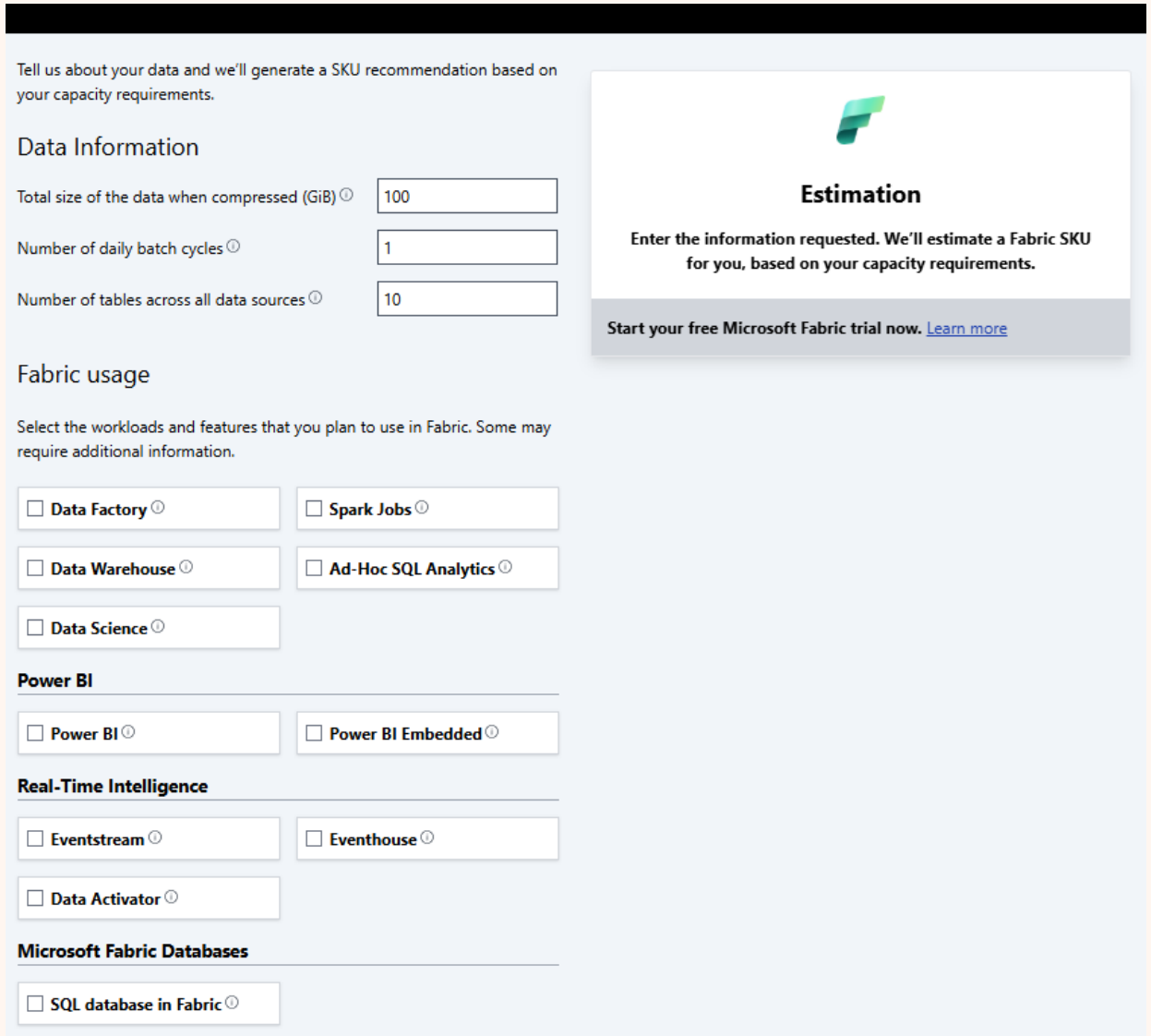
- Provides capacity estimations customized to your unique requirements
- Results vary depending on your input

Impact

- Recommended SKU when starting on Fabric
- Test solutions to ensure they're correctly sized.

Learn more?

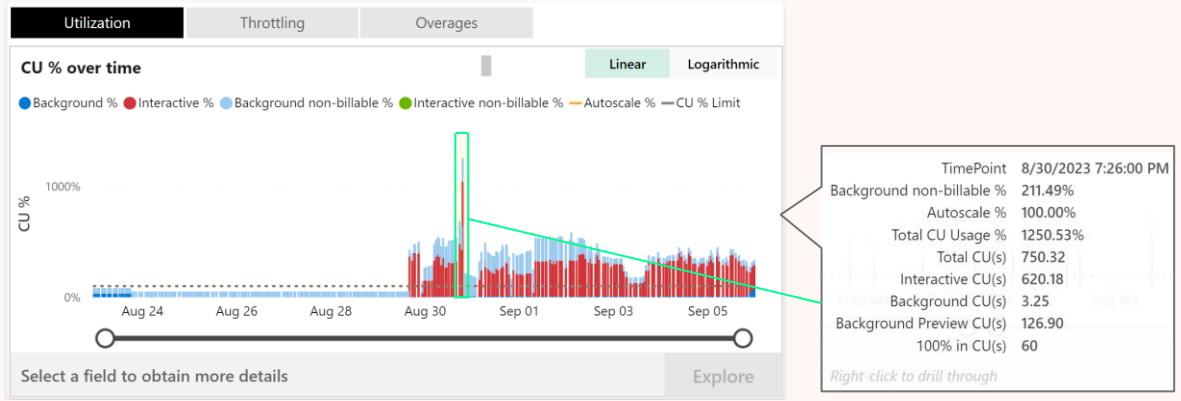
- Fabric Tech Talks - Estimating Capacity in Microsoft Fabric
- Blog – Master SKU Estimations



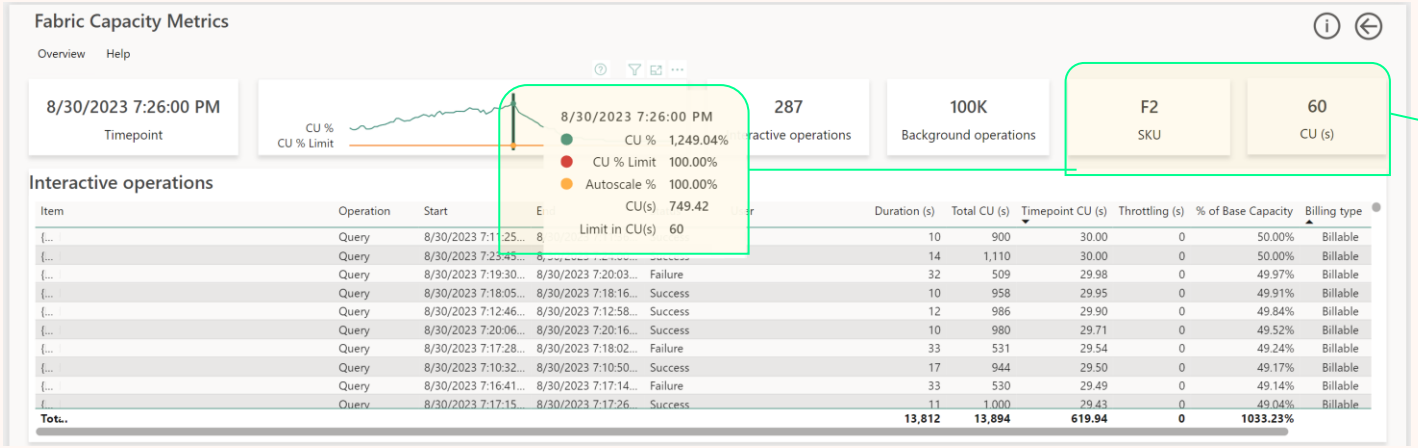
https://www.microsoft.com/en-us/microsoft-fabric/capacity-estimator

Capacity planning case study - measurement

Start with a test or trial capacity to evaluate the load of specific Fabric Experiences i.e., Power BI Datasets, Spark Notebooks or a Datawarehouse



If usage is above the current capacity limits , choose the desired utilization rate to accommodate via capacity scale up



Load Capacity Metrics timepoint drill to analyze :

- Total CU's consumed : 749 CU(s)
- Capacity Size : (F2)
- CU(s) available on your capacity : 60 CU(s)

Capacity planning case study – SKU selection

To accommodate a 749 CU(s) load the admin can purchase an F32 capacity providing 960 CU(s) of throughput

Universal Compute Capacities SKU Sizing

SKU	Capacity Units (CU)	CU's (per 30s)	Power BI SKU	Power BI V-cores
F2	2	60	-	0.25
F4	4	120	-	0.5
F8	8	240	A1	1
F16	16	480	A2	2
F32	32	960	A3	4
F64	64	1920	P1	8
F128	128	3840	P2	16
F256	256	7680	P3	32
F512	512	15360	P4	64
F1024	1024	30720	P5	128
F2048	2048	61440	-	256

Fabric Capacity Reservations

Commit to spending for 1-year periods, to get a 41% discount

Existing Azure Concept

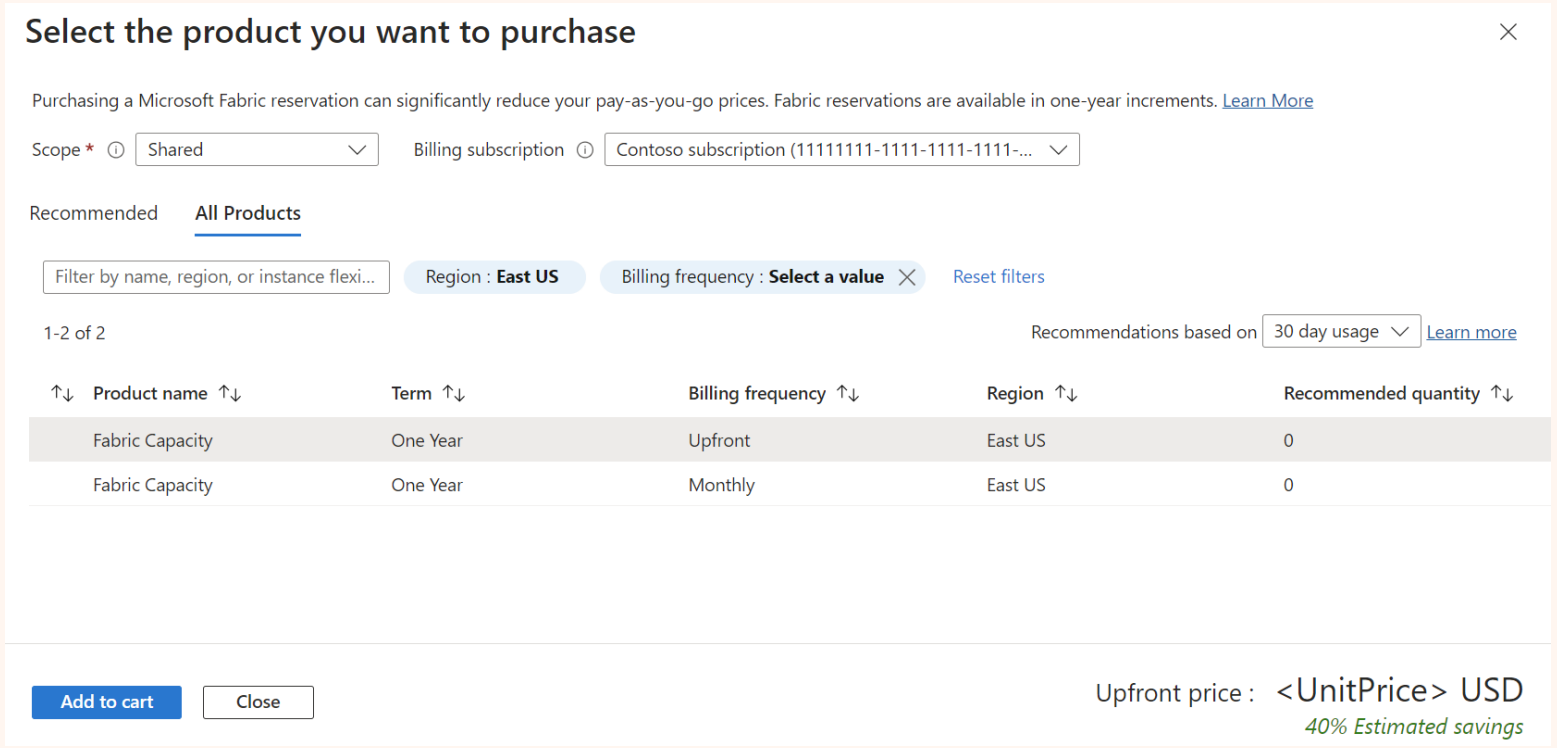
- Cancel, Refund, Exchange by policies

Reservations can be configured by

- Scope:
  - Billing Account
  - Subscription(s)
  - Resource Group(s)
  - Region
- Management Group

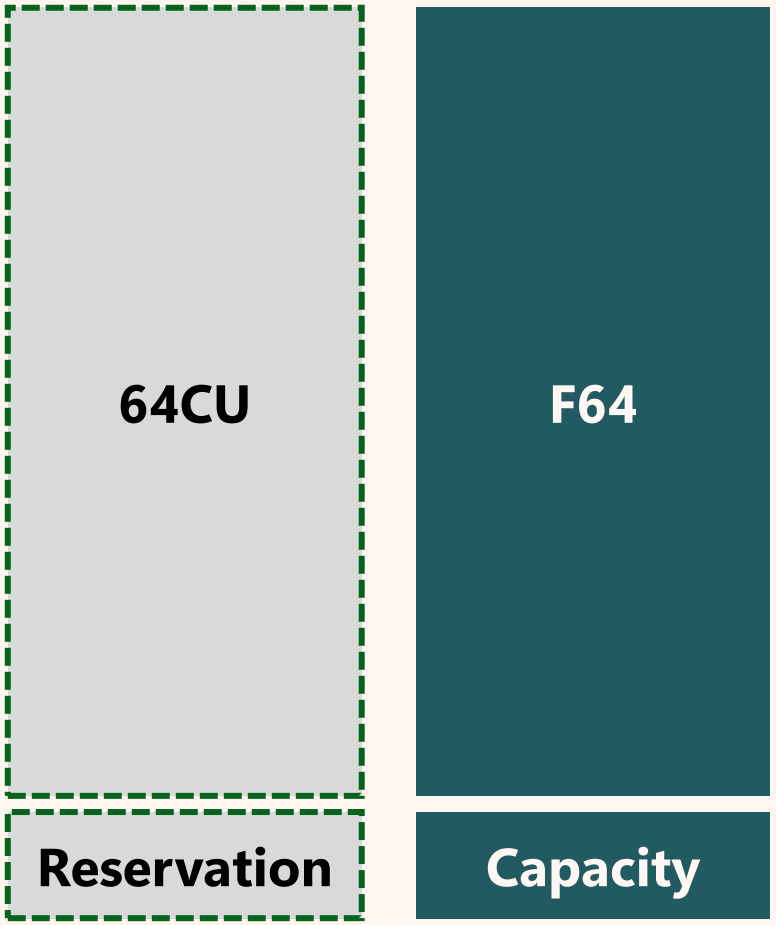
Reservation that is enforced

- Even when no capacity matches the scope, billing happens
- Not automatically renewed, unless configured
- Upon expiry, capacities impacted automatically switch to PayGo



Fabric Capacity Reservations

Example 1 – Single Reservation matching a single capacity



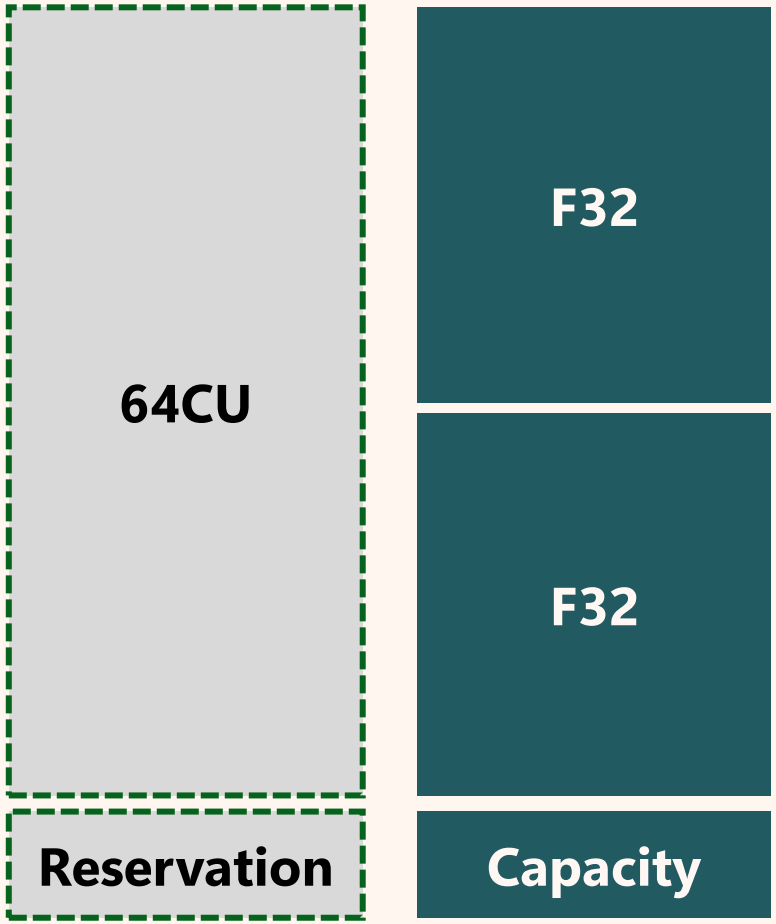
Without a reservation, you pay-as-you-go

Active capacity is covered by Reservation*, discount (~41%) applies

*) Billing Account / Subscription / Resource Group / Region Need to align

Fabric Capacity Reservations

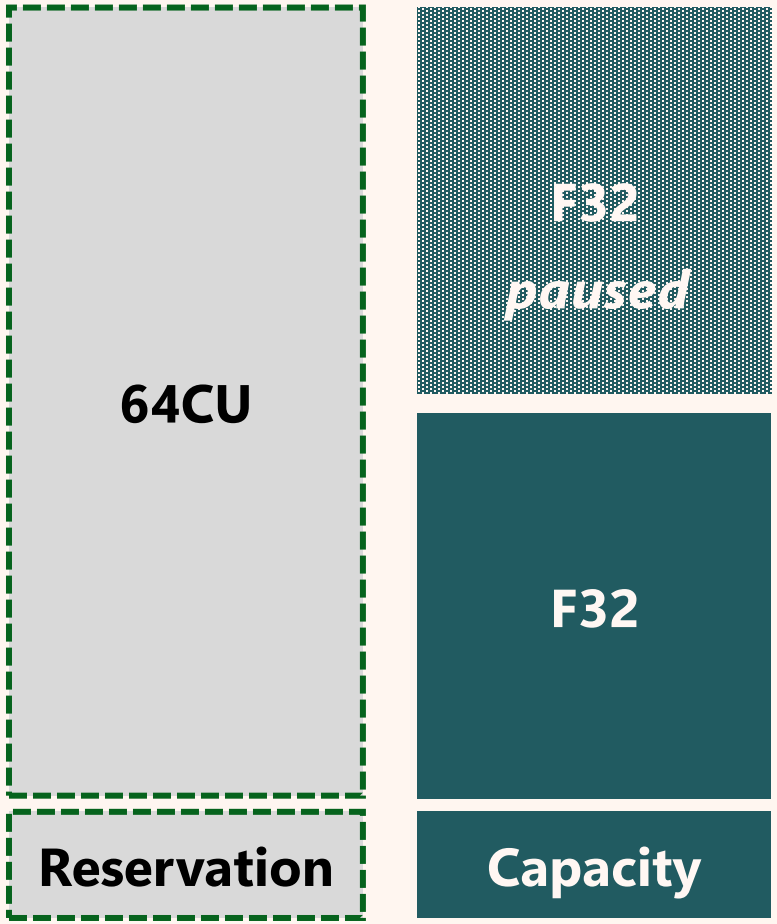
Example 2 – Single Reservation matches active capacities



All active capacities are covered by Reservation, discount applies

Fabric Capacity Reservations

Example 3 – Single Reservation exceeds Active Capacities



Reservation only applies to active capacity, 32CU are "wasted"

Reservation sets the floor

Fabric Capacity Reservations

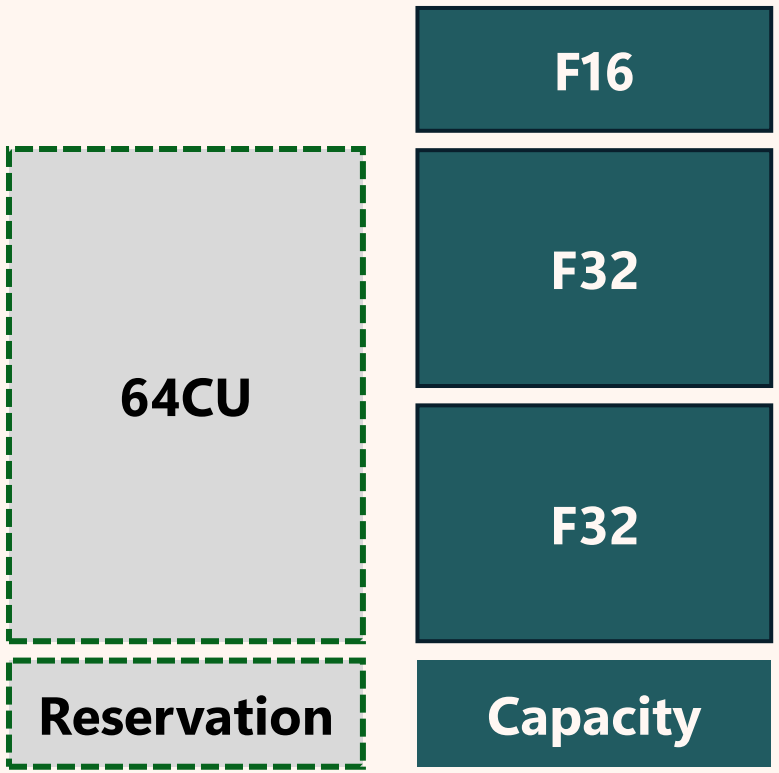
Example 4 – Single SKU exceeding the reservation amount

Time scaling up	PAYGO	F64 Reservation	PAYGO + F64 RI	F128 Reservation
10%	€ 1.000	€ 6.000	€ 7.000	€ 12.000
20%	€ 2.000	€ 6.000	€ 8.000	€ 12.000
30%	€ 3.000	€ 6.000	€ 9.000	€ 12.000
40%	€ 4.000	€ 6.000	€ 10.000	€ 12.000
50%	€ 5.000	€ 6.000	€ 11.000	€ 12.000
60%	**€ 6.000**	**€ 6.000**	**€ 12.000**	**€ 12.000**
70%	€ 7.000	€ 6.000	€ 13.000	€ 12.000
80%	€ 8.000	€ 6.000	€ 14.000	€ 12.000
90%	€ 9.000	€ 6.000	€ 15.000	€ 12.000
100%	€ 10.000	€ 6.000	€ 16.000	€ 12.000

ed billing for F128
city, 64CU reserved,
64CU PaYGo

Fabric Capacity Reservations

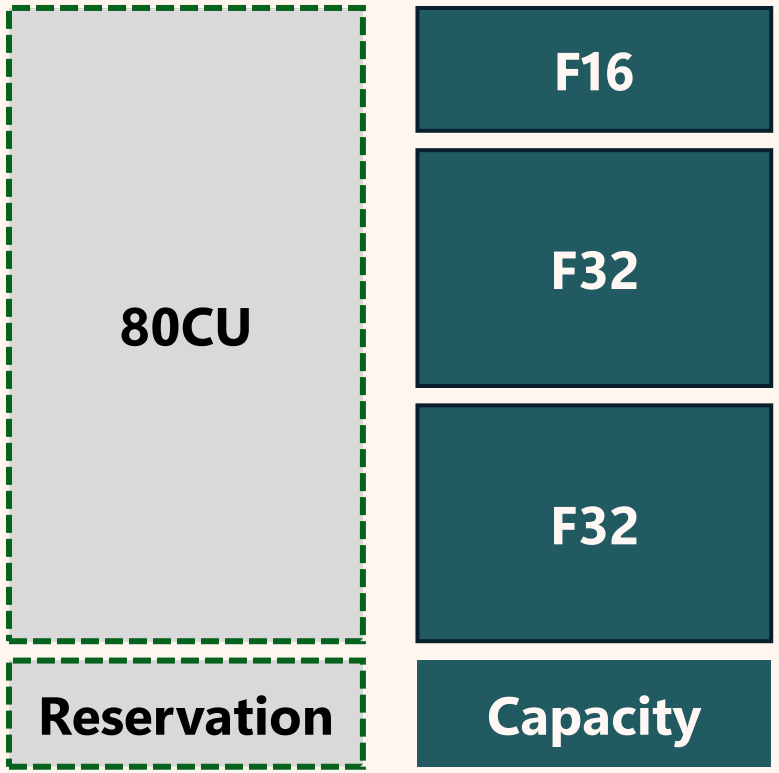
Example 5 – Multiple SKUs exceeding reservation amount



Billing can apply to all the capacities, is only shown in Billing

Fabric Capacity Reservations

Example 5 – Multiple SKUs exceeding reservation amount

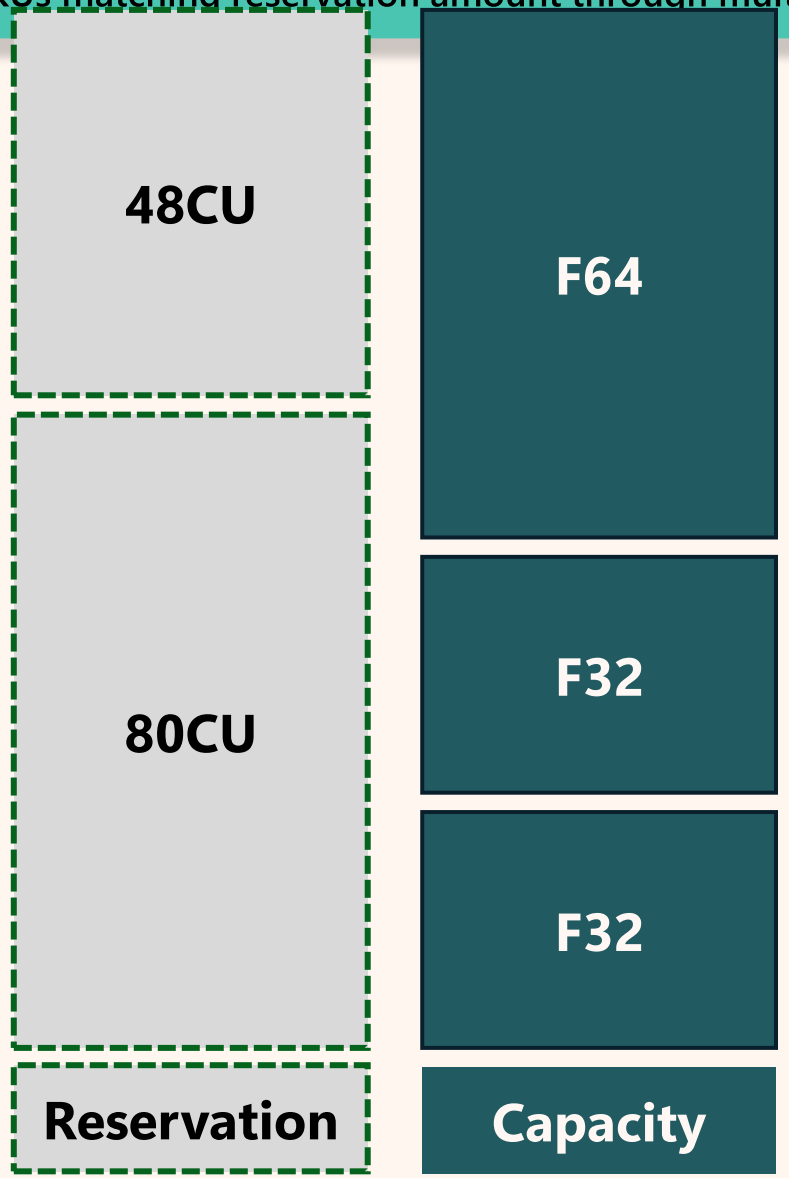


Reservations are based on CUs, not SKUs

Can be any (?) even number

Fabric Capacity Reservations

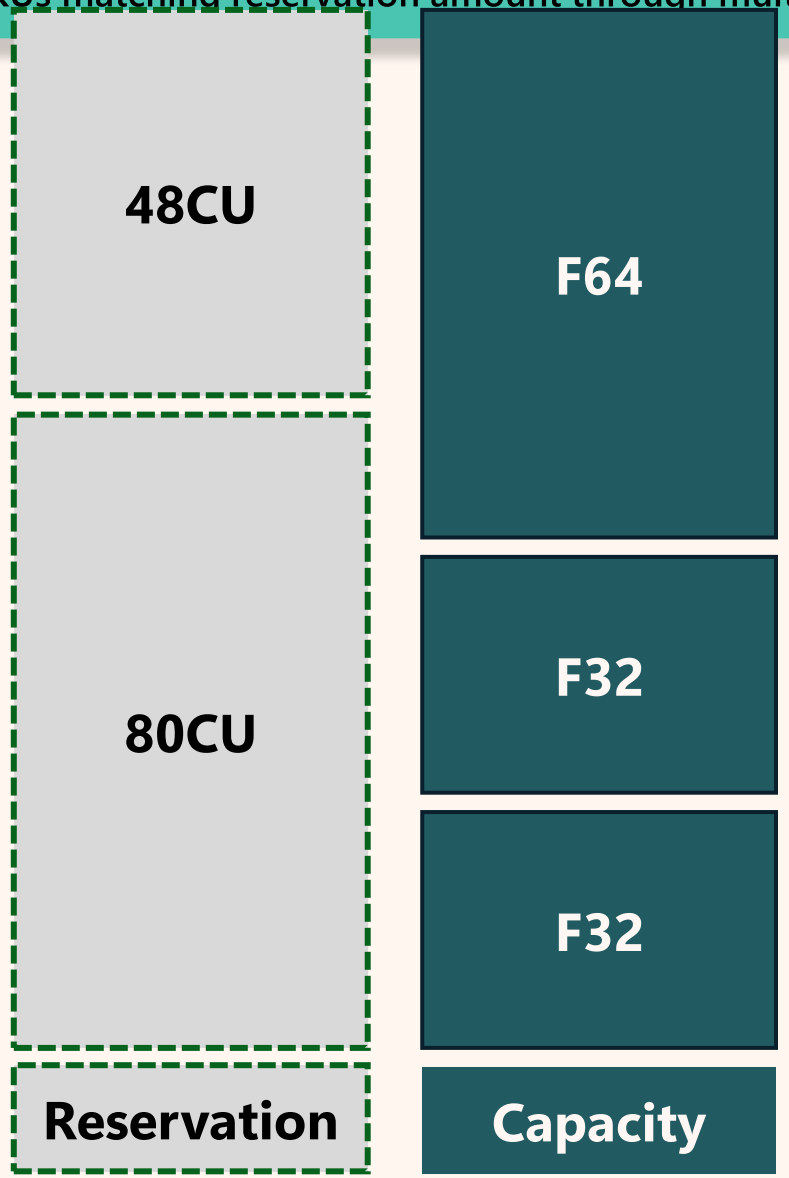
Example 6 – Multiple SKUs matching reservation amount through multiple reservations



Only when second Reservation is added, all Capacities are covered

Fabric Capacity Reservations

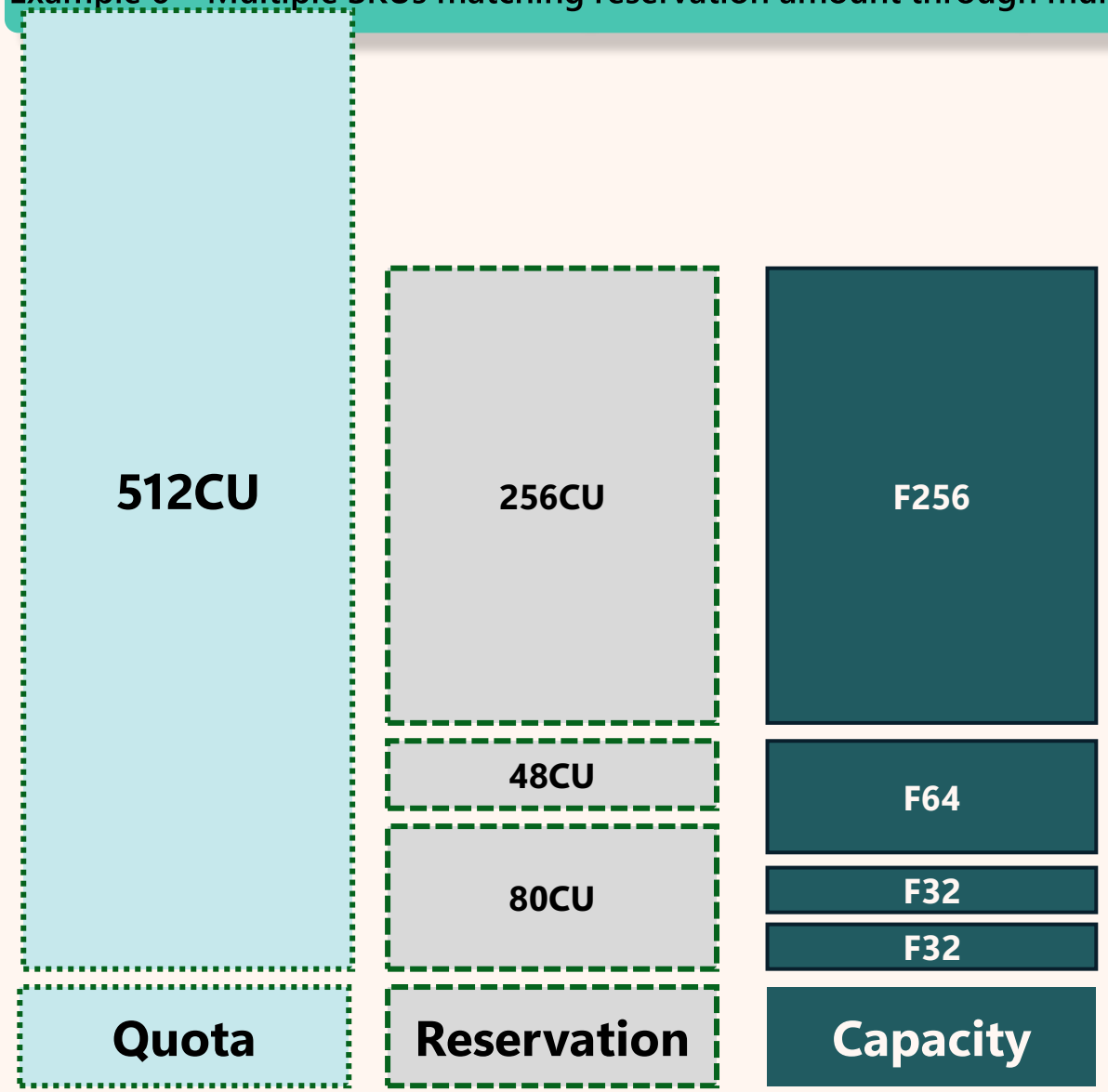
Example 6 – Multiple SKUs matching reservation amount through multiple reservations



Reservation sets the floor

Fabric Capacity Reservations

Example 6 – Multiple SKUs matching reservation amount through multiple reservations



Reservation sets the floor

Quota determines the ceiling*

*) per Subscription/Region

Azure Quota Management Service Integration

GA (March 2025)

Better resource allocation to meet Microsoft's customer capacity needs (blog || docs)

Fabric Quota limits

- Limits the number of Capacity Units (CUs) you can provision across multiple capacities in a subscription, based on subscription type and region

Customers can request a quota adjustment

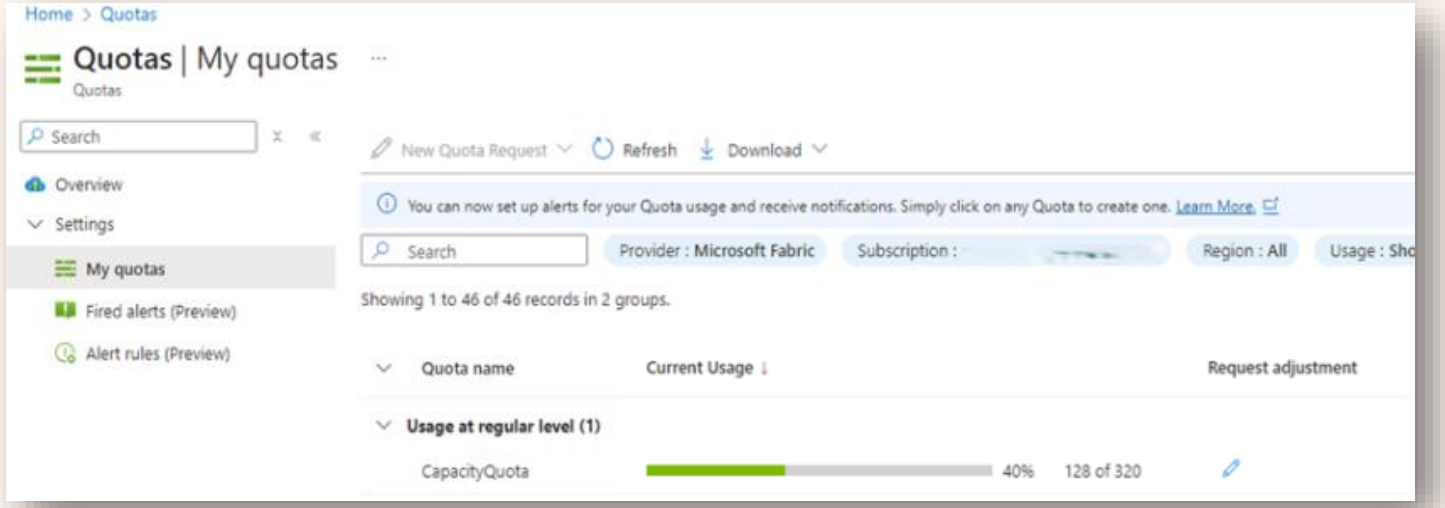
- Auto approved up to specific limits
- Customers can request additional quota through Microsoft customer support

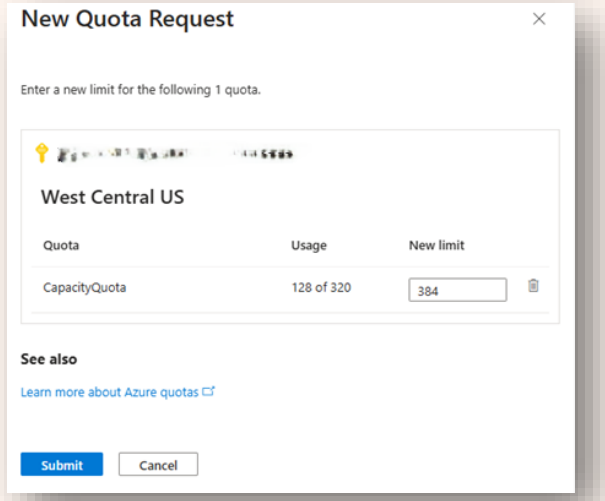
Update automation scripts

- Customers who provision capacity dynamically should check quota first

Impact

- Better resource allocation
- Security and compliance – reduce risk of unauthorized excessive usage





Azure Quota Management Service Integration

GA (March 2025)

Better resource allocation to meet Microsoft's customer capacity needs (blog || docs)

Fabric Quota limits

- Limits the number of Capacity Units (CUs) you can provision across multiple capacities in a subscription, based on subscription type and region

Customers can request a quota adjustment

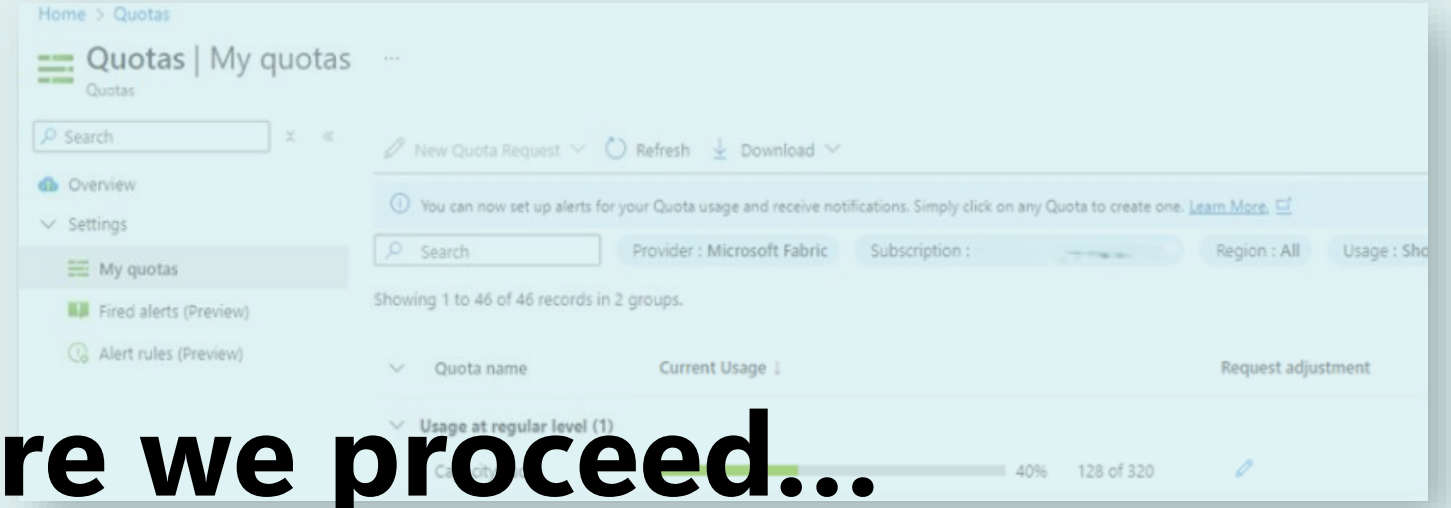
- Auto approved up to specific limits
- Customers can request additional quota through Microsoft customer support

Update automation scripts

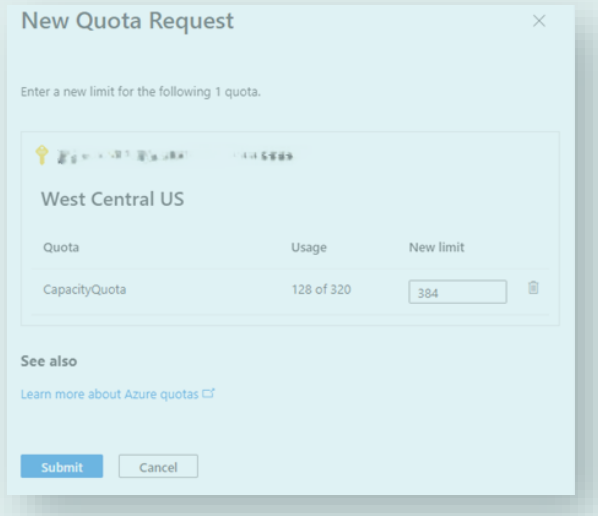
- Customers who provision capacity dynamically should check quota first

Impact

- Better resource allocation
- Security and compliance – reduce risk of unauthorized excessive usage



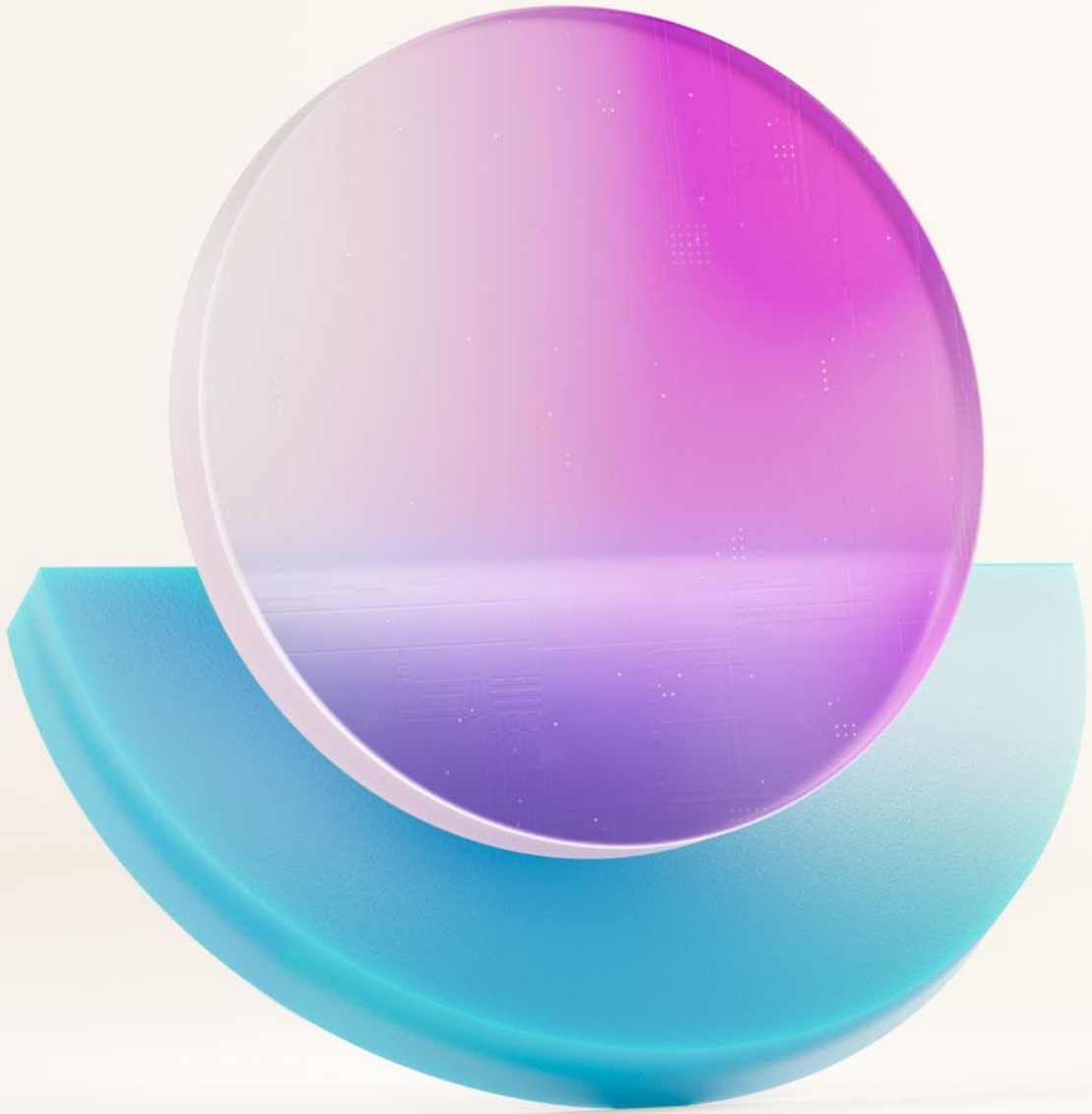
Before we proceed…



A big thank you to our partners

PBIG
POWER BI GEBRUIKERSGROEP

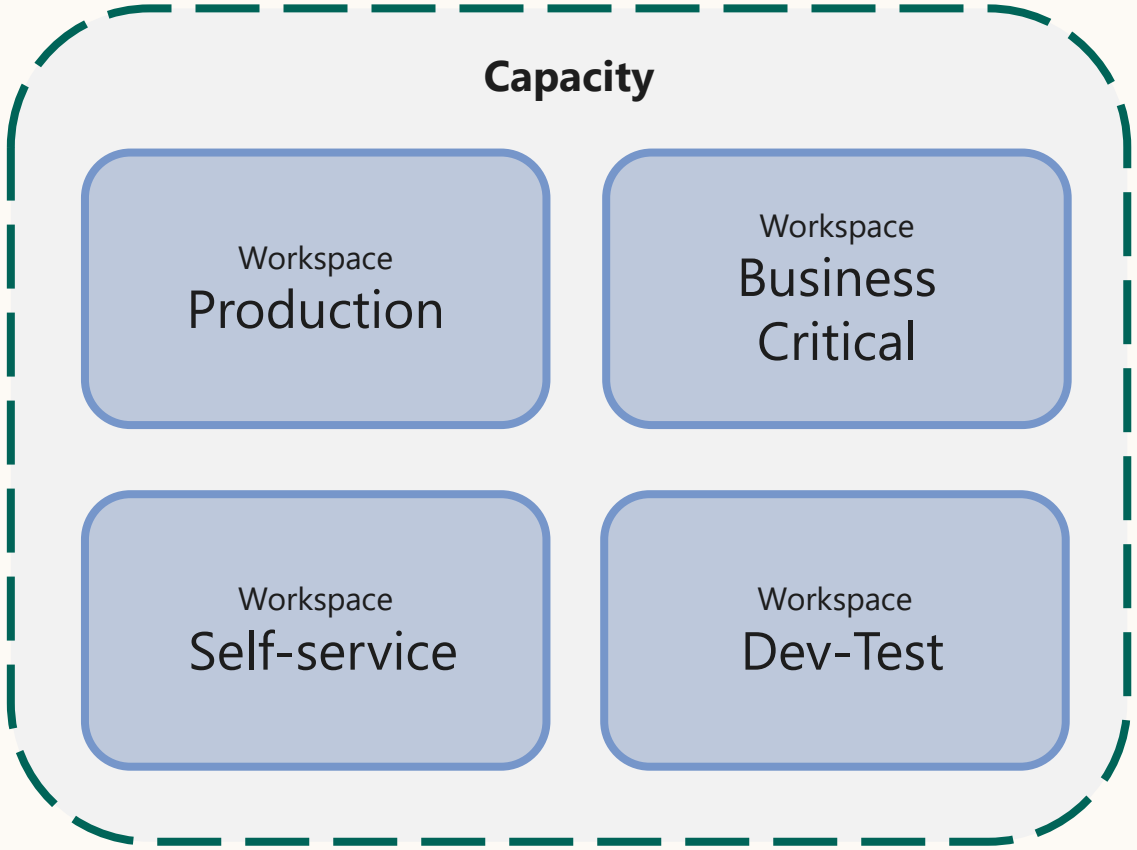
Microsoft Fabric
COMMUNITY CONFERENCE



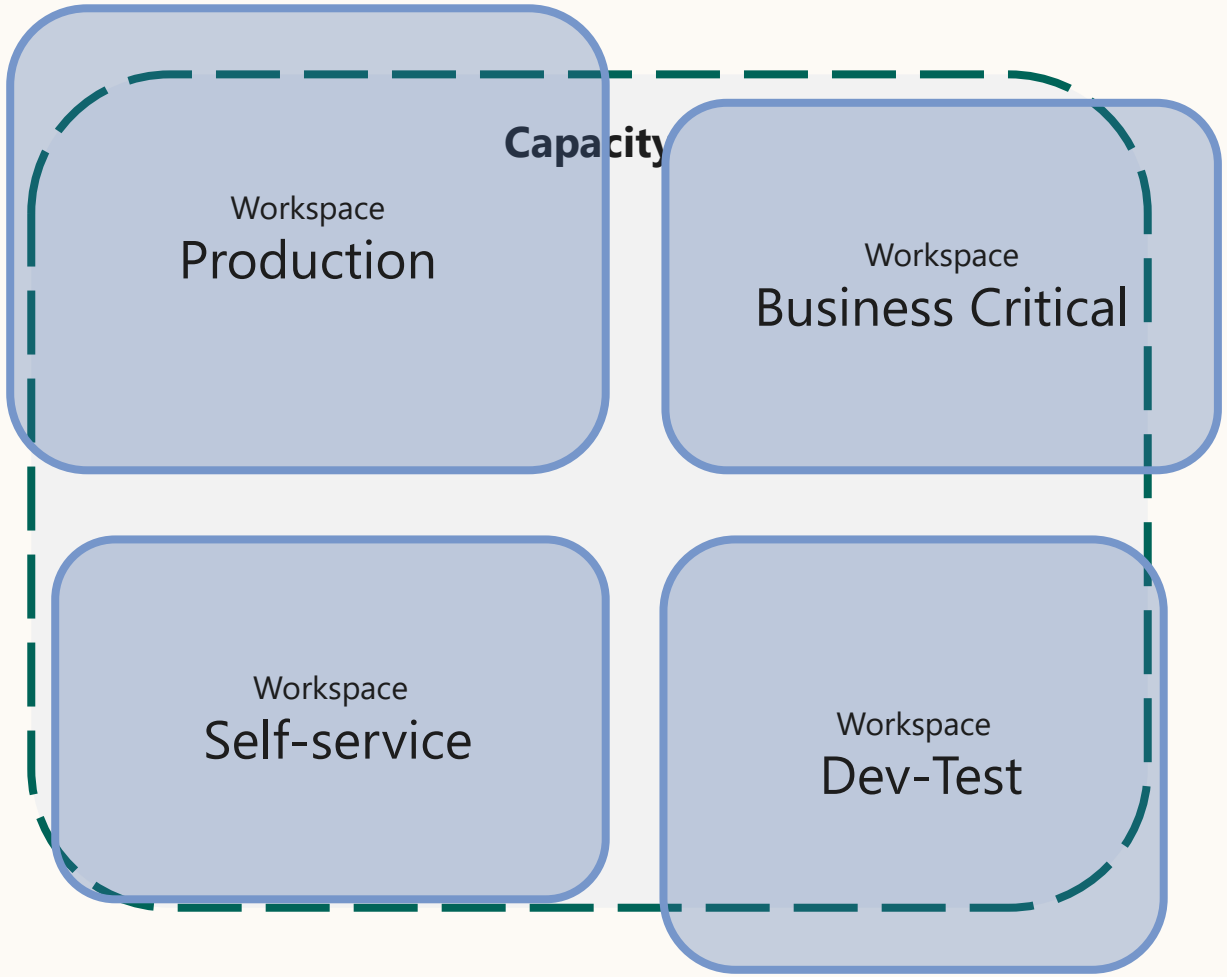
Workspace & Capacity Strategies

When 💩 hits the fan ..

What if the capacity runs out?



What if the capacity runs out?



Strategy 1: Optimize

Approach

- ❑ Work with content creators to optimize CU utilization
- ❑ Establish and follow best practices

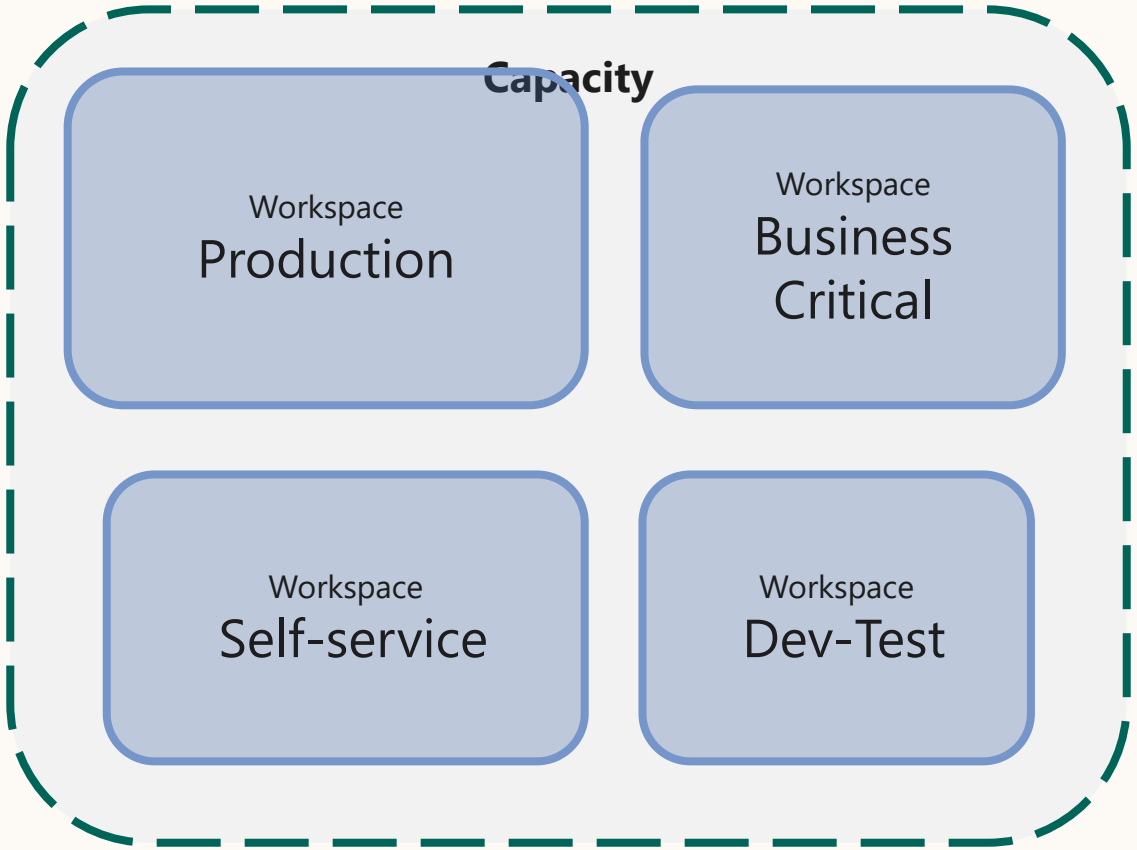
Pros

- ✓ Avoids increased cost
- ✓ Learning carries over

Consa

- x Can be difficult and time-consuming

What if the capacity runs out?



Strategy 1: Optimize

Approach

- ❑ Work with content creators to optimize CU utilization
- ❑ Establish and follow best practices

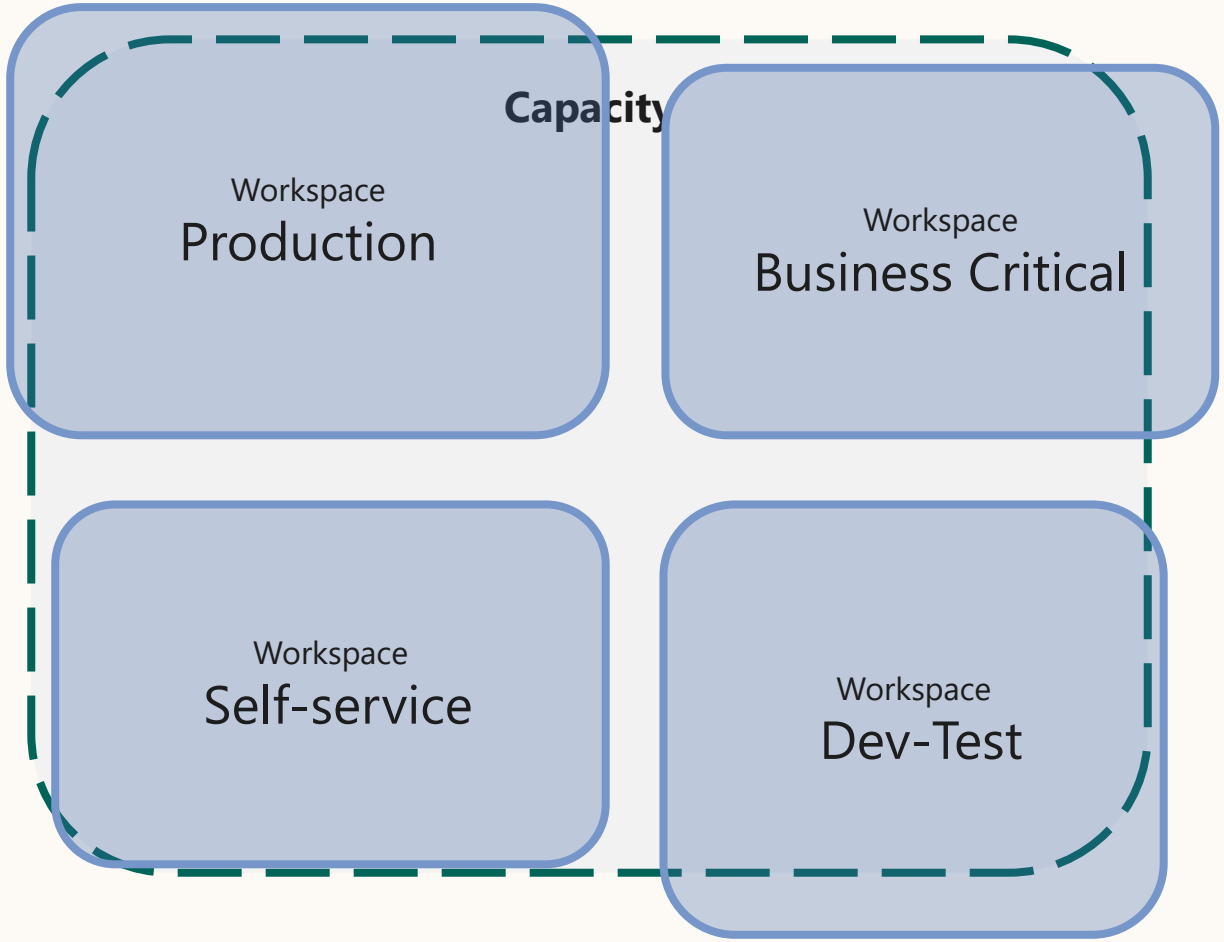
Pros

- ✓ Avoids increased cost
- ✓ Learning carries over

Cons

- x Can be difficult and time-consuming

What if the capacity runs out?



Strategy 2: Scale up

Approach

- ❑ Move to a bigger SKU size
- ❑ Turn on autoscale on a P SKU or setup size change on a F SKU

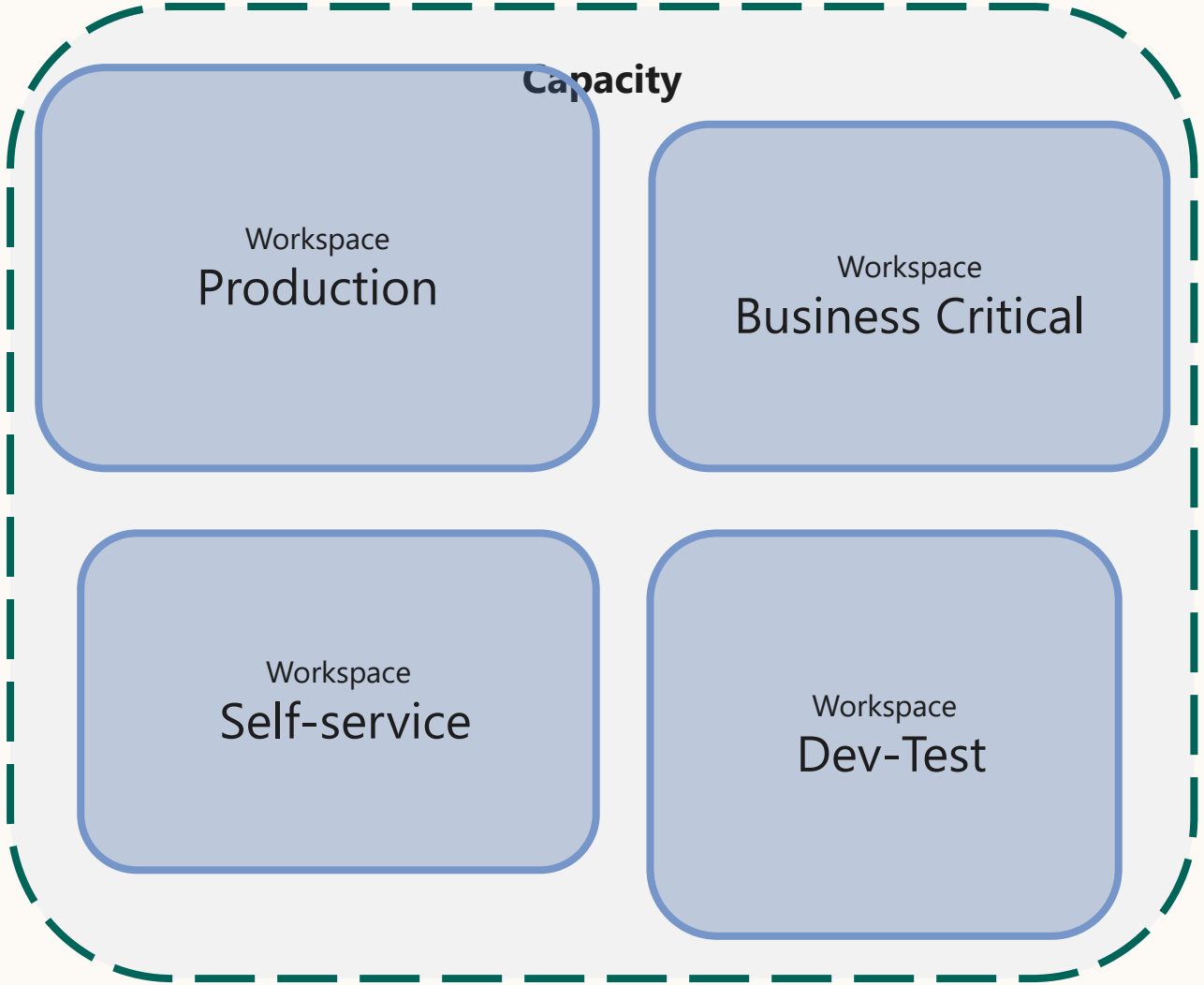
Pros

- ✓ Adds available CUs for every item
- ✓ Easy

Cons

- x Cost
- x Bad actors can still be a problem

What if the capacity runs out?



Strategy 2: Scale up

Approach

- ❑ Move to a bigger SKU size
- ❑ Turn on autoscale on a P SKU or setup size change on a F SKU

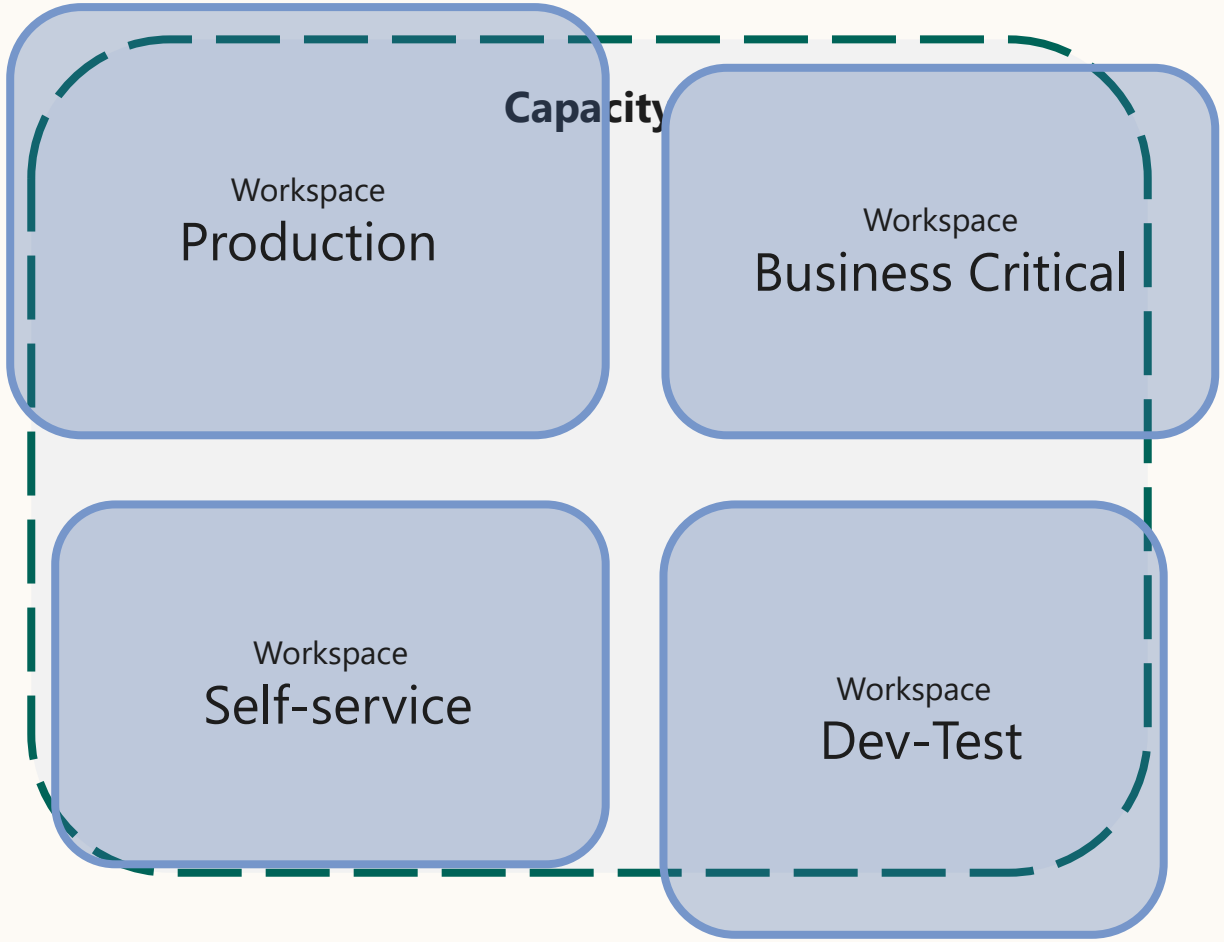
Pros

- ✓ Adds available CUs for every item
- ✓ Easy

Cons

- x Cost
- x Bad actors can still be a problem

What if the capacity runs out?



Strategy 3: Scale out

Approach

- ❑ Create multiple smaller SKUs based on workload, organization, type of work, etc..

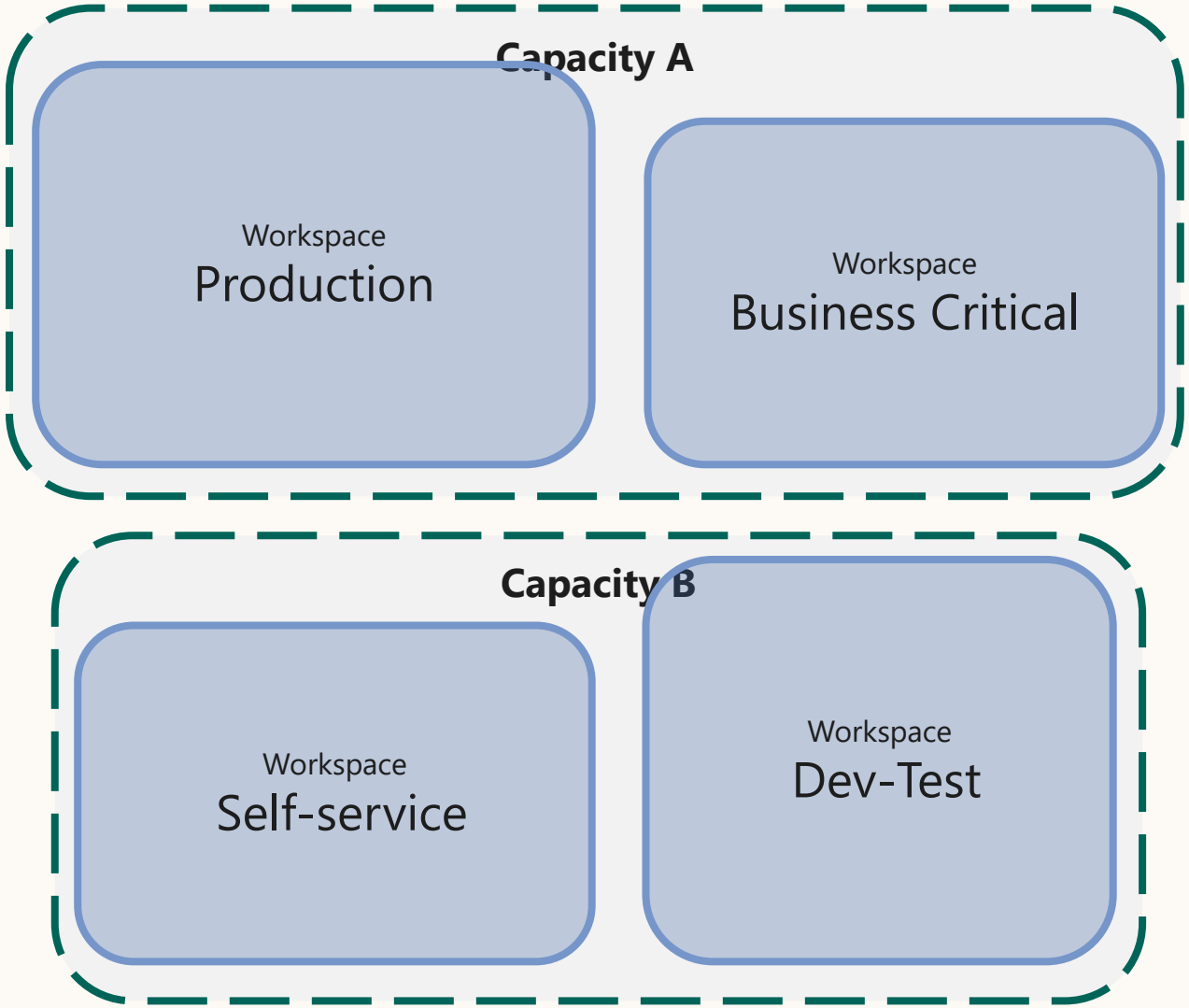
Pros

- ✓ Easy
- ✓ Provides some isolation from bad actors
- ✓ Flexibility

Cons

- x Cost
- x High CU items have increased change of throttling

What if the capacity runs out?



Strategy 3: Scale out

Approach

- ❑ Create multiple smaller SKUs based on workload, organization, type of work, etc..

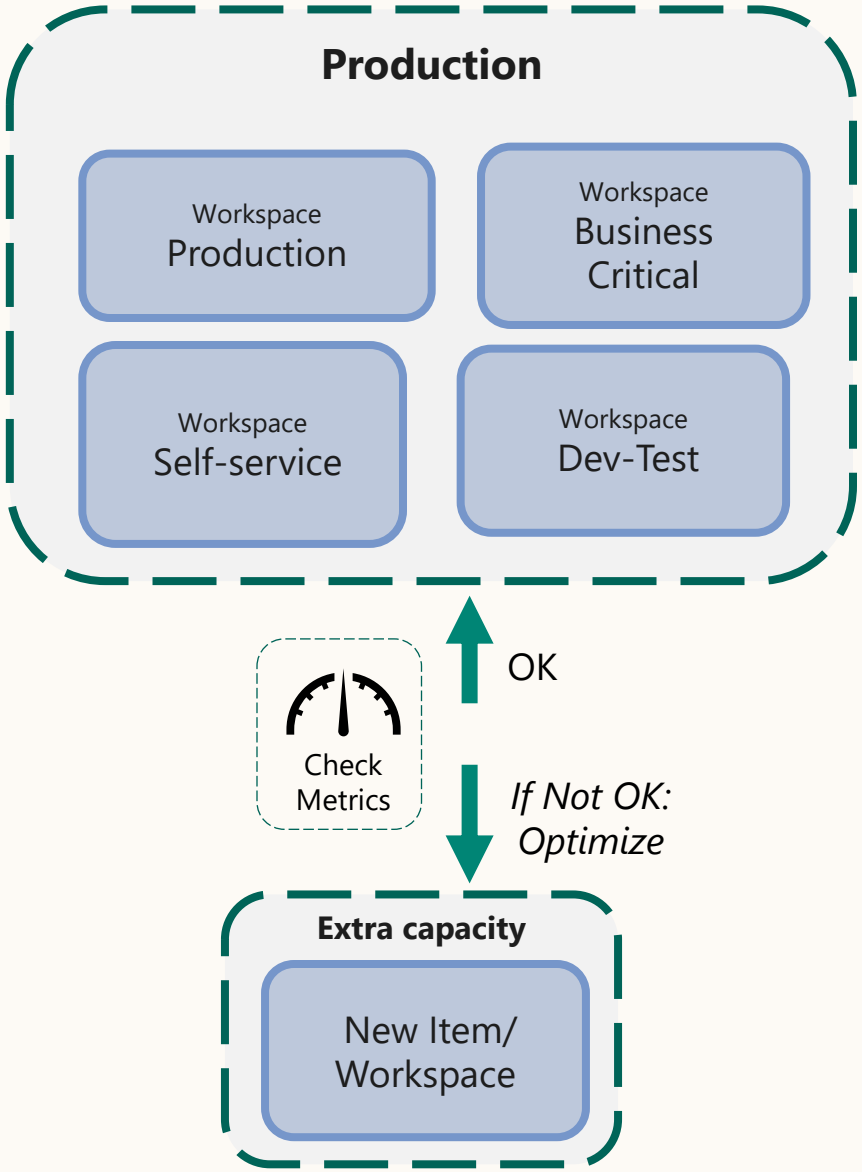
Pros

- ✓ Easy
- ✓ Provides some isolation from bad actors
- ✓ Flexibility

Cons

- x Cost
- x High CU items have increased change of throttling

What if the capacity runs out?



Strategy 4: Isolate

Approach

❑ Set up active monitoring and separate specific cases from other workloads in designated capacities

Pros

✓ Provides some isolation from bad actors

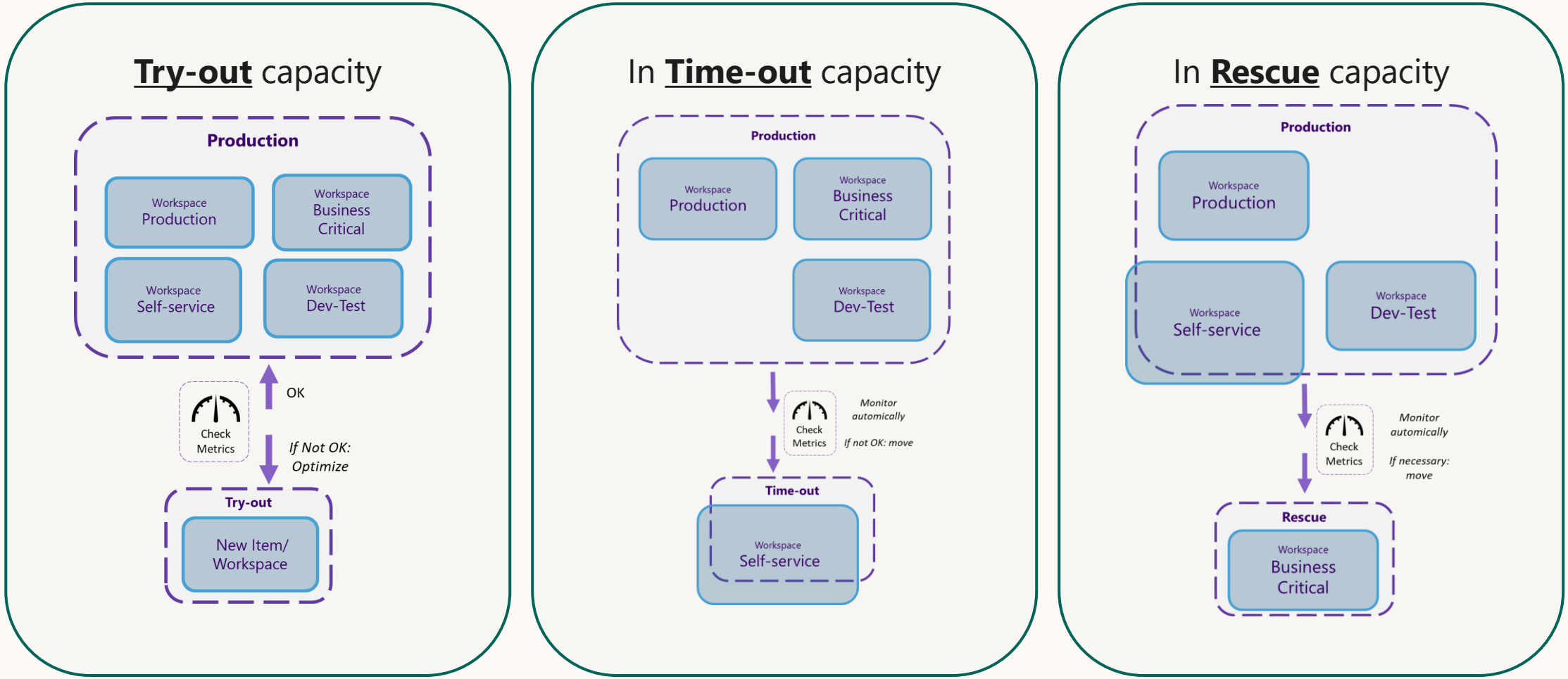
✓ Flexibility

Cons

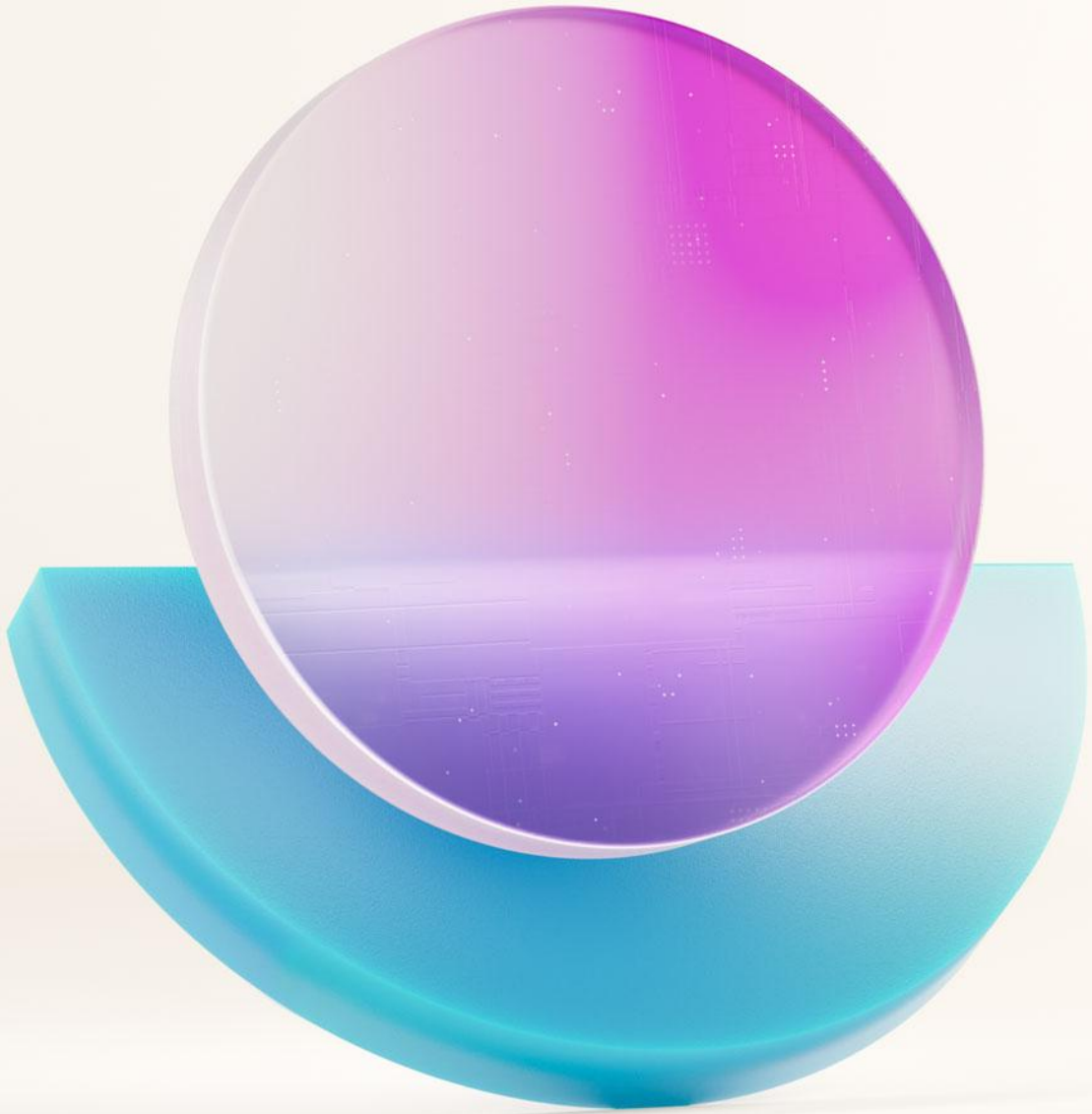
x Requires planning and maintenance

What if the capacity runs out?

Strategy 4: Isolate



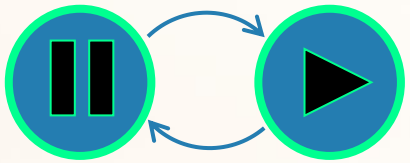
Microsoft Fabric
Community Conference



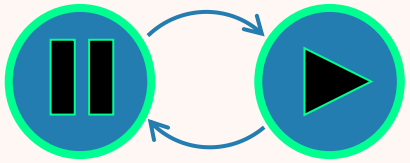
Dealing with Throttling

Microsoft Fabric
COMMUNITY CONFERENCE

Pausing and Resuming Capacities



Pausing and Resuming Capacities



Why pause capacities?

1) It **<u>can</u>** help manage compute costs.
2) It clears any debt that has accumulated. Use it to quickly resolve throttling.

What does it do?

Workloads stop execution within 10 minutes of Pause action	New requests are not allowed to Start	Smoothed usage will be reconciled

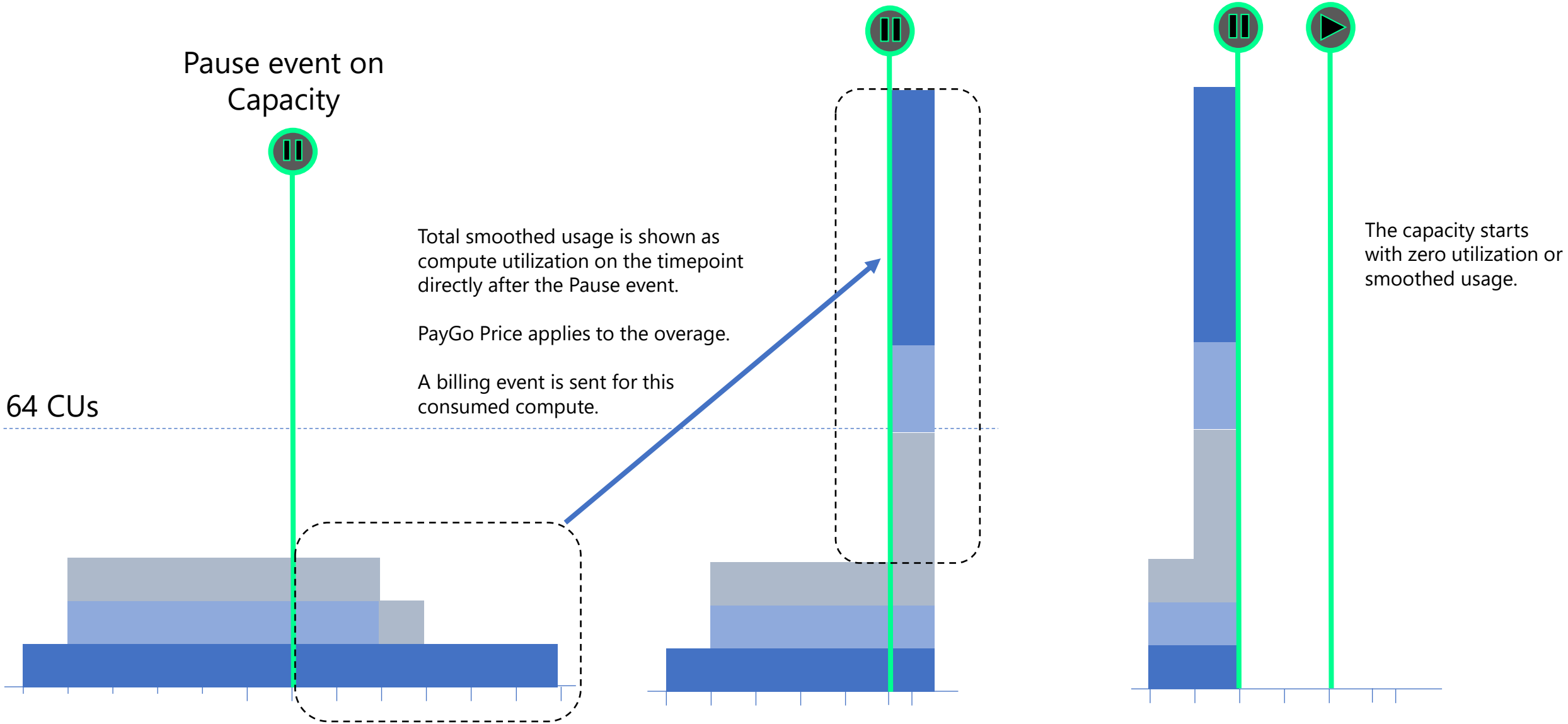
Note: OneLake storage costs continue to be billed while a capacity is paused

How Capacity Pause & Resume works

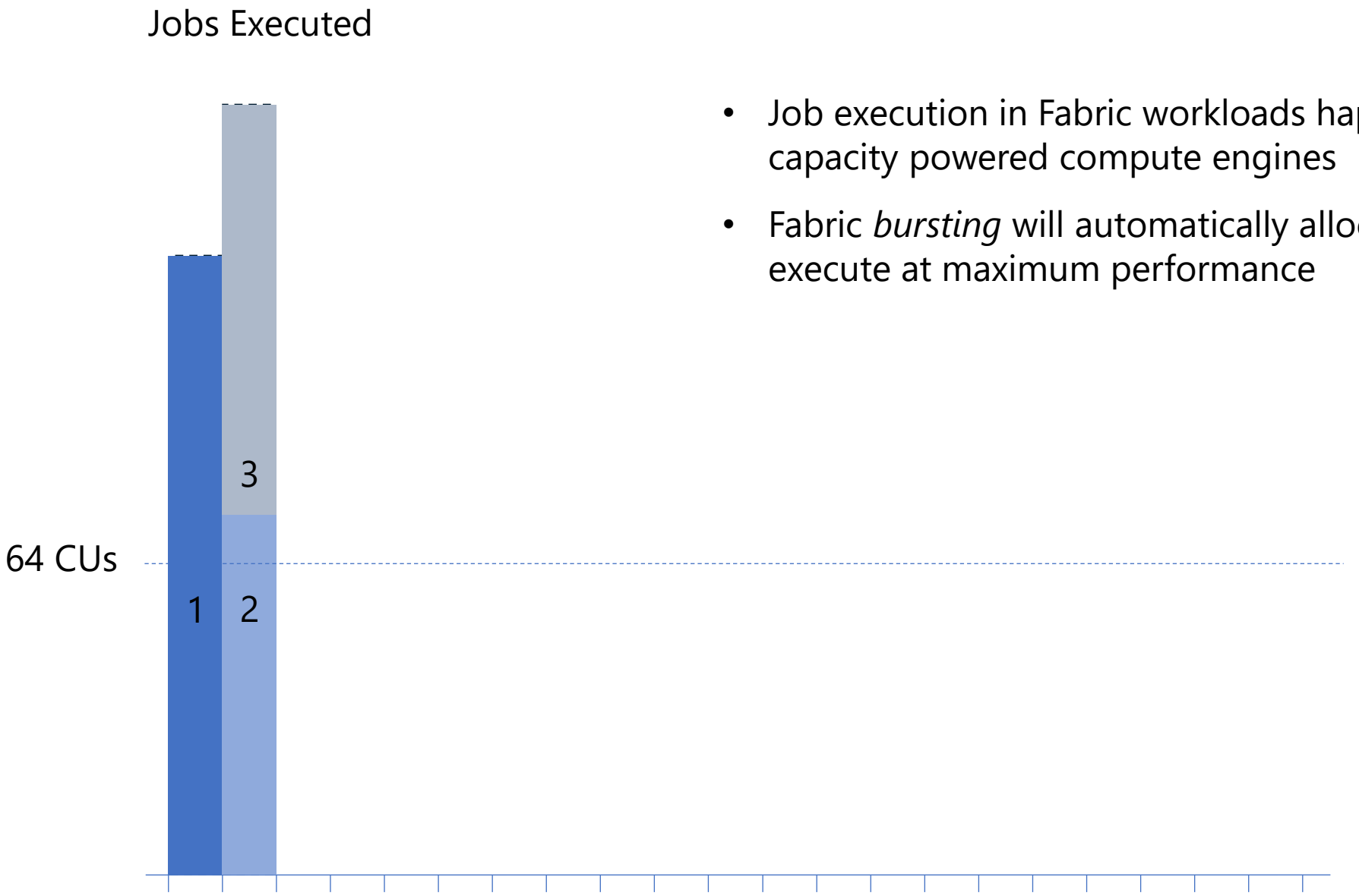
When a capacity is **paused**…

Smoothed usage is **reconciled**.

Later, it can be **resumed**.



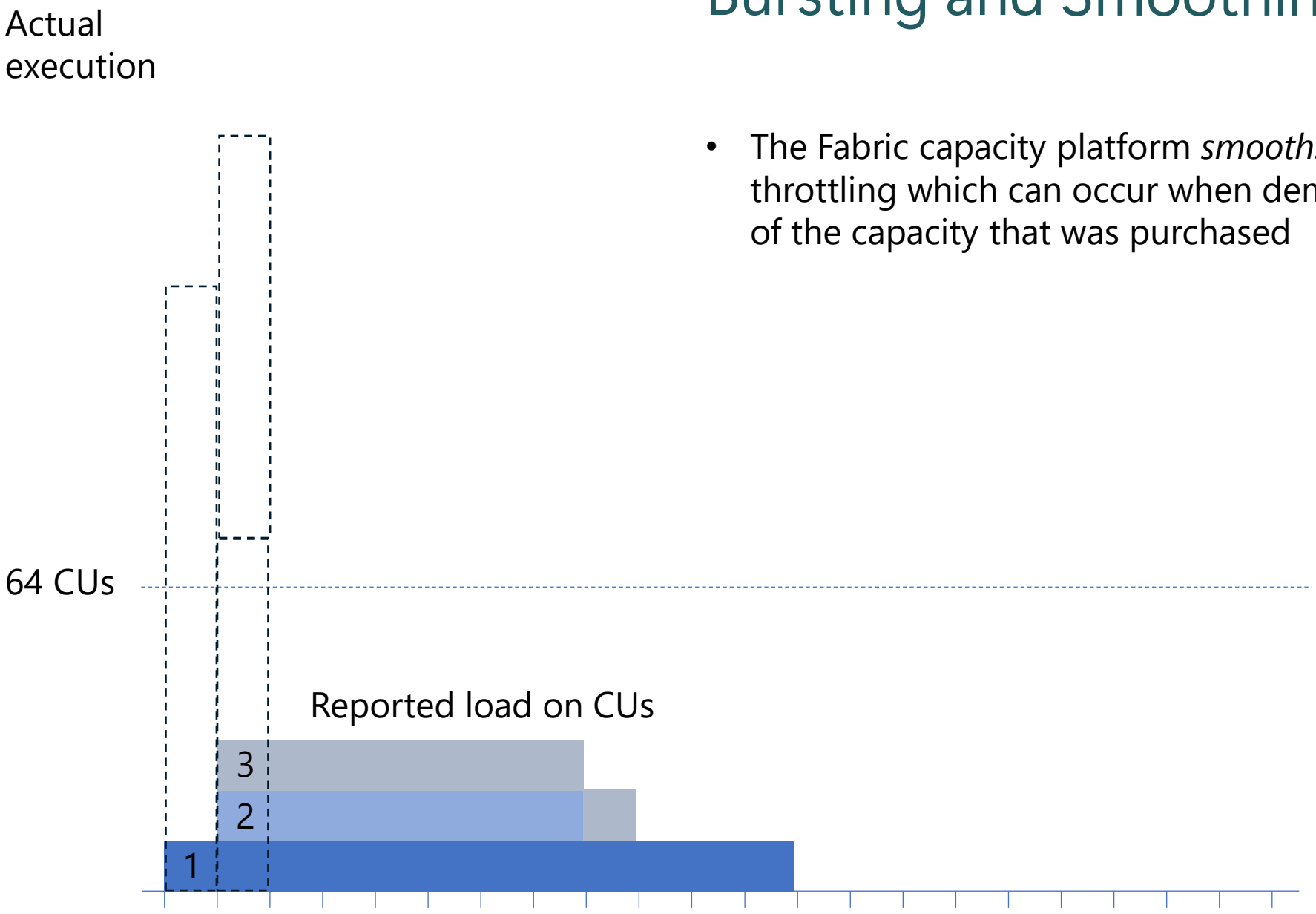
Bursting and Smoothing



- Job execution in Fabric workloads happens on-demand via capacity powered compute engines
- Fabric *bursting* will automatically allocate resources as needed to execute at maximum performance

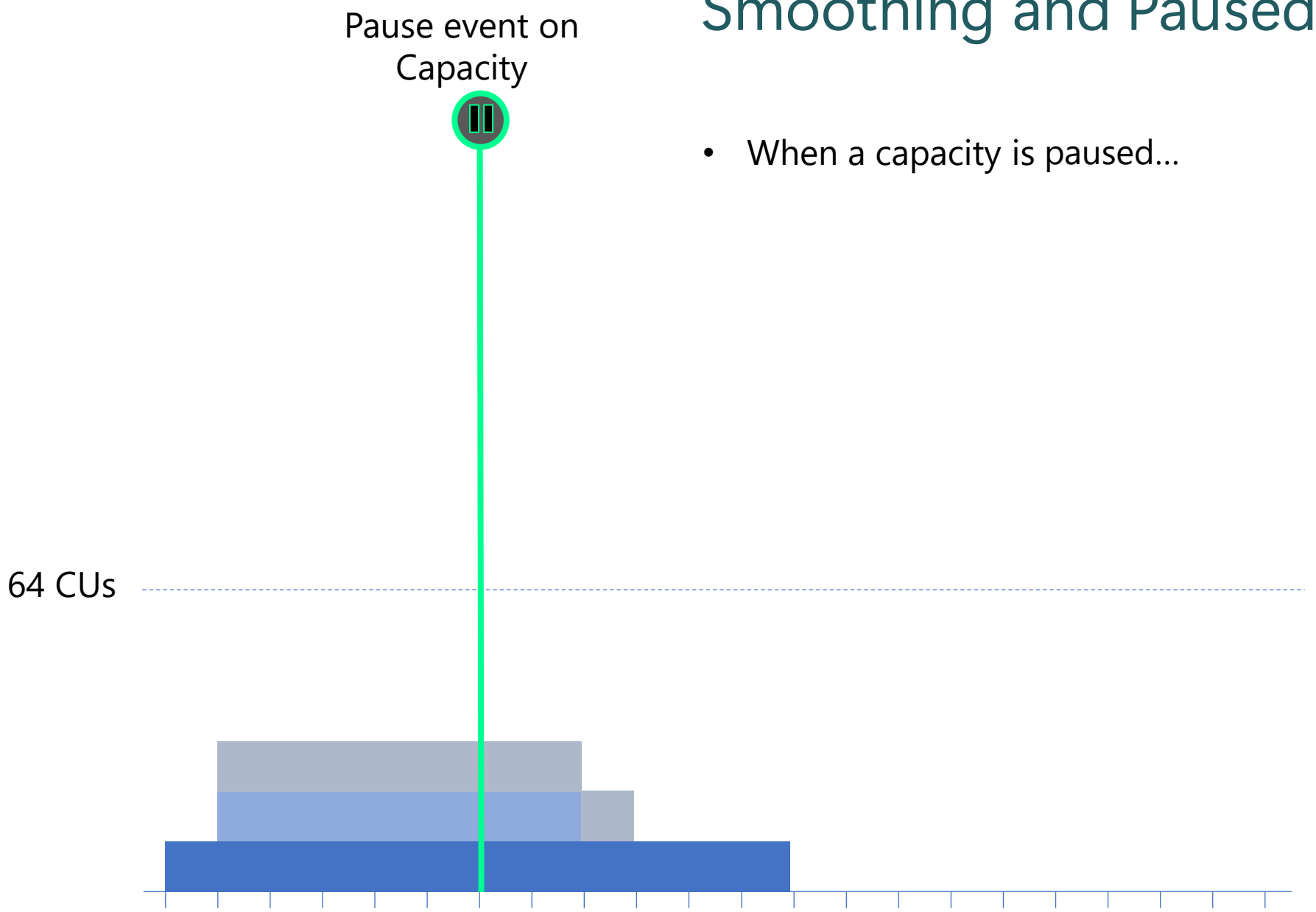
Bursting and Smoothing

- The Fabric capacity platform *smooths* usage out to reduce throttling which can occur when demand exceeds the throughput of the capacity that was purchased

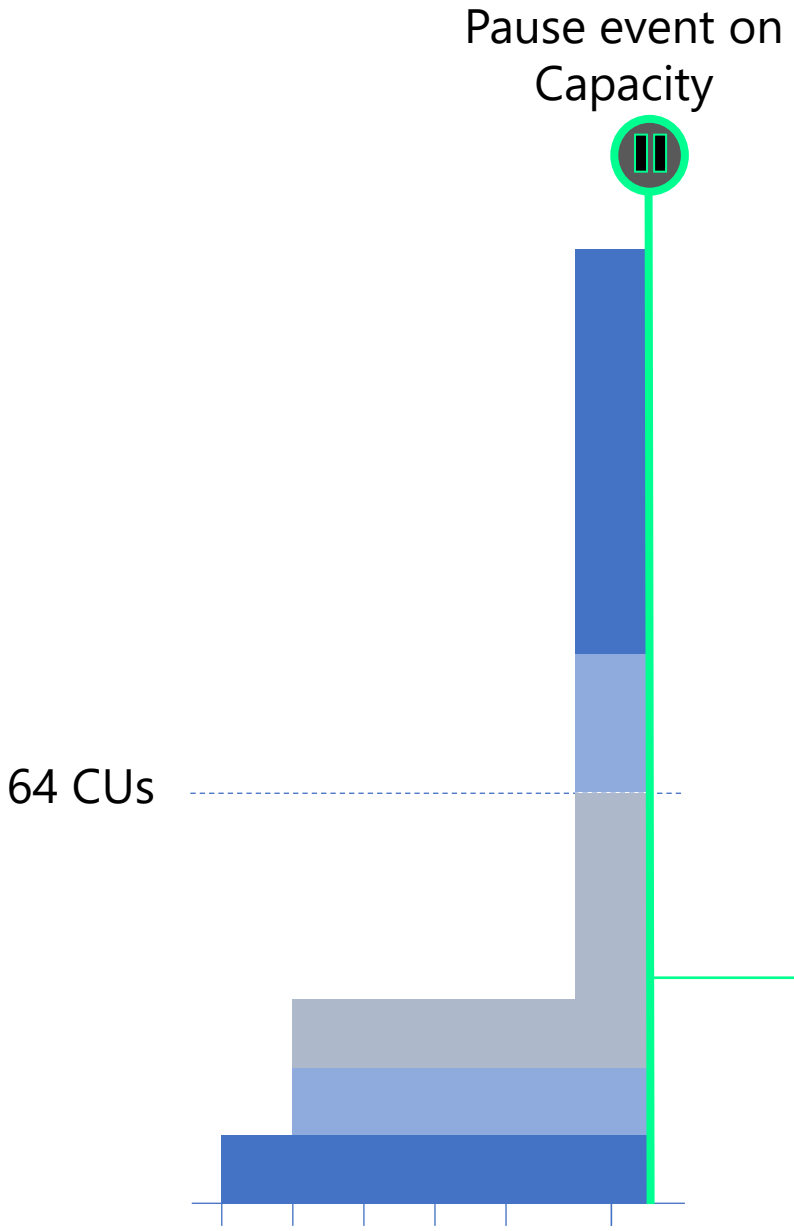


Smoothing and Paused Capacities

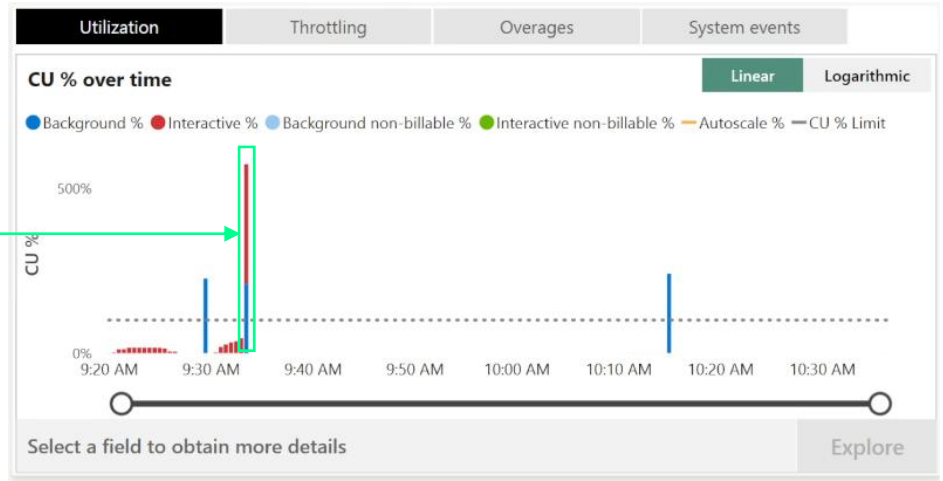
- When a capacity is paused…



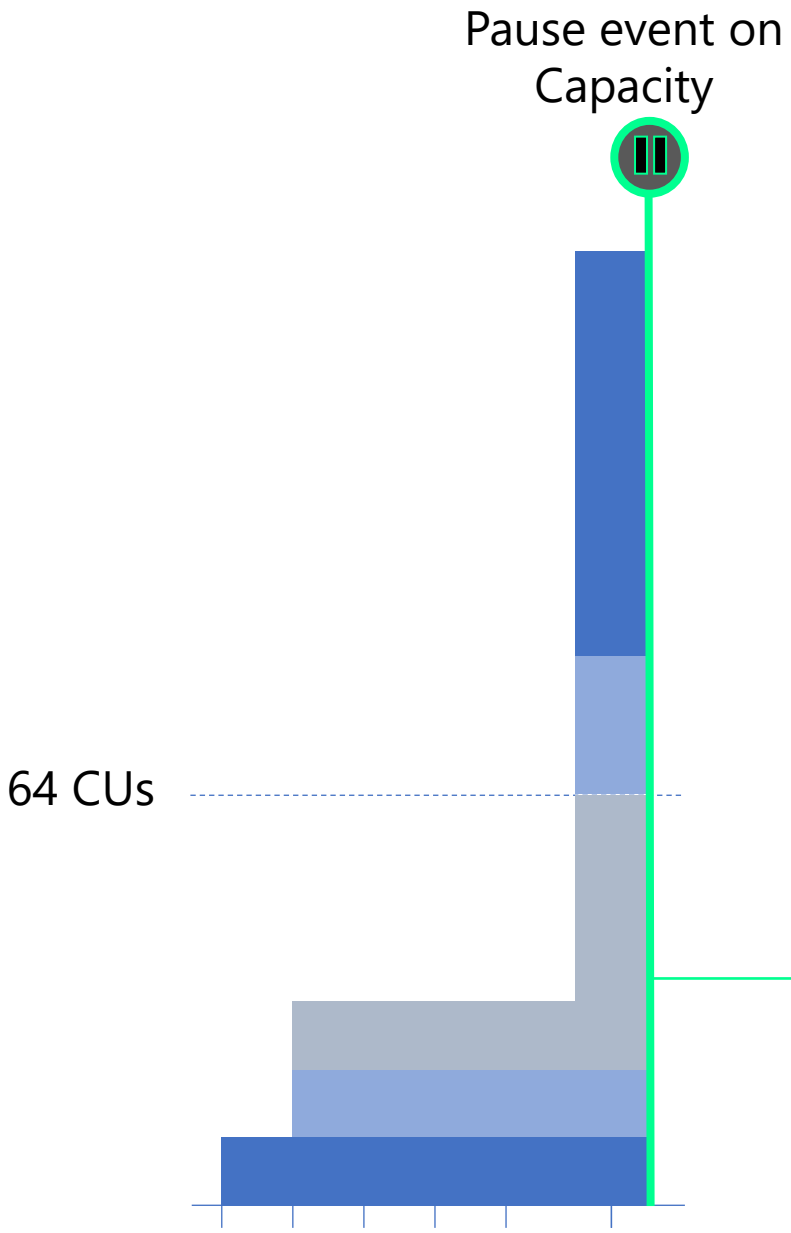
Smoothing and Paused Capacities



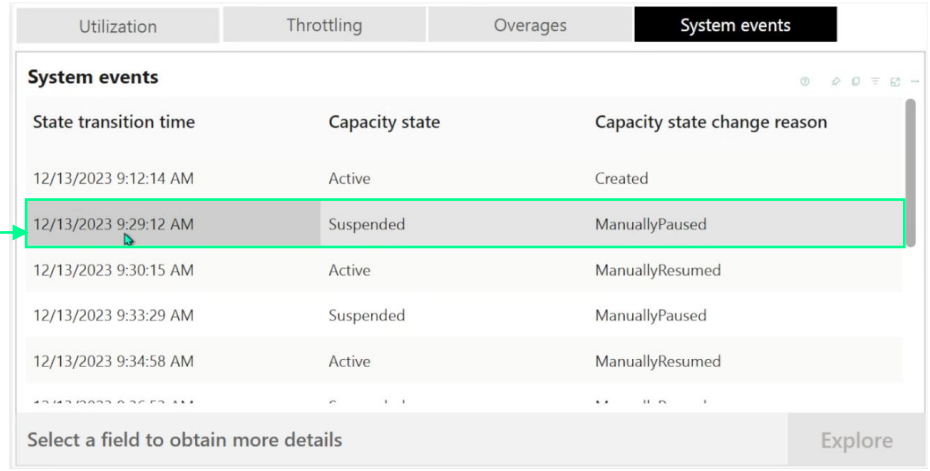
- When a capacity is paused…
- Usage that was smoothed into the future will be “reconciled” and charged against the capacity at the timestamp the capacity was paused
- Reconciled usage will show up as a spike in capacity metrics



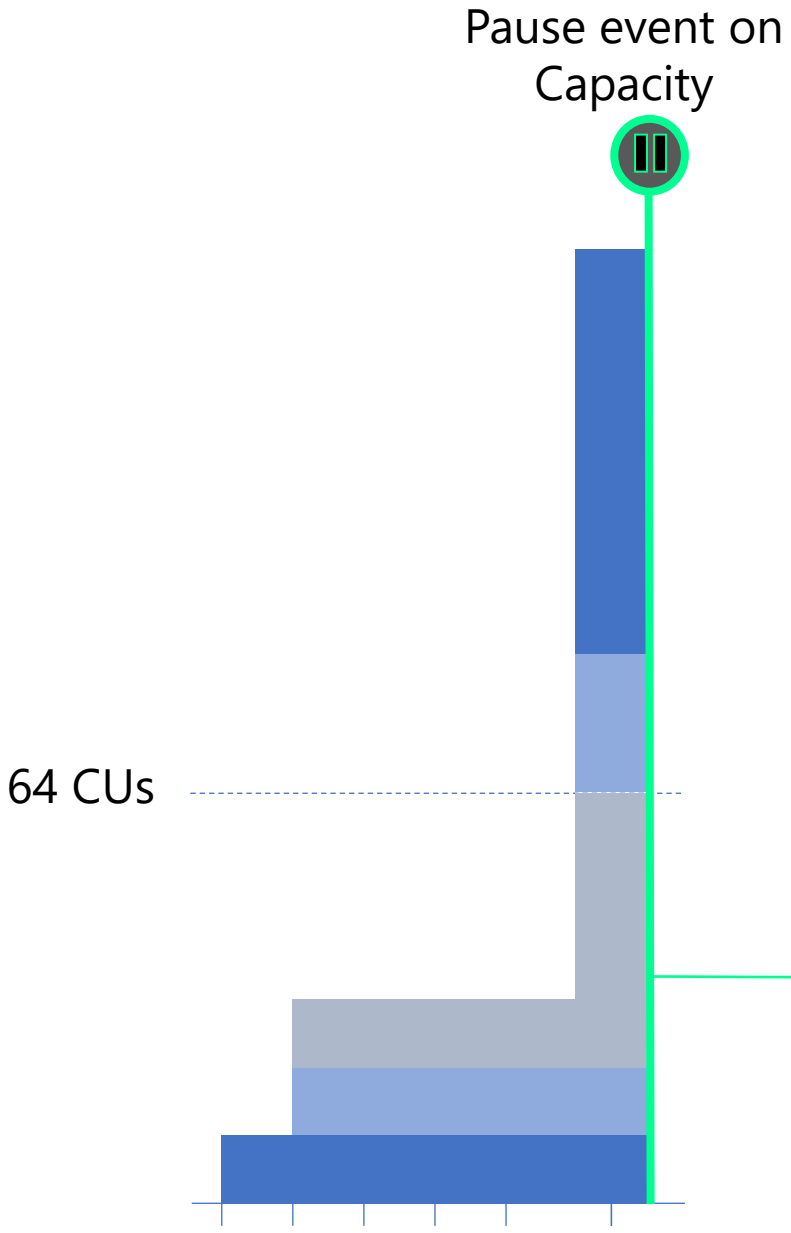
Smoothing and Paused Capacities



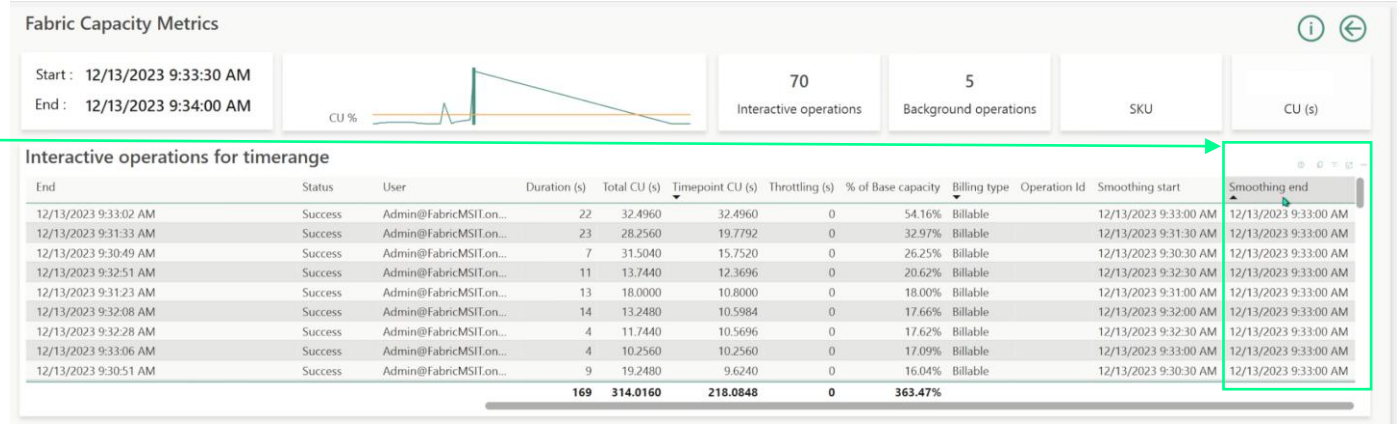
- When a capacity is paused…
- Usage that was smoothed into the future will be “reconciled” and charged against the capacity at the timestamp the capacity was paused
- Pause events can be viewed in the new System events tab



Smoothing and Paused Capacities



- When a capacity is paused…
- Usage that was smoothed into the future will be “reconciled” and charged against the capacity at the timestamp the capacity was paused
- Pause events timestamp is shown in the smoothing end field in timepoint drill views



End	Status	User	Duration (s)	Total CU (s)	Timepoint CU (s)	Throttling (s)	% of Base capacity	Billing type	Operation Id	Smoothing start	Smoothing end
12/13/2023 9:33:02 AM	Success	Admin@FabricMSIT.on...	22	32.4960	32.4960	0	54.16%	Billable		12/13/2023 9:33:00 AM	12/13/2023 9:33:00 AM
12/13/2023 9:31:33 AM	Success	Admin@FabricMSIT.on...	23	28.2560	19.7792	0	32.97%	Billable		12/13/2023 9:31:30 AM	12/13/2023 9:33:00 AM
12/13/2023 9:30:49 AM	Success	Admin@FabricMSIT.on...	7	31.5040	15.7520	0	26.25%	Billable		12/13/2023 9:30:30 AM	12/13/2023 9:33:00 AM
12/13/2023 9:32:51 AM	Success	Admin@FabricMSIT.on...	11	13.7440	12.3696	0	20.62%	Billable		12/13/2023 9:32:30 AM	12/13/2023 9:33:00 AM
12/13/2023 9:31:23 AM	Success	Admin@FabricMSIT.on...	13	18.0000	10.8000	0	18.00%	Billable		12/13/2023 9:31:00 AM	12/13/2023 9:33:00 AM
12/13/2023 9:32:08 AM	Success	Admin@FabricMSIT.on...	14	13.2480	10.5984	0	17.66%	Billable		12/13/2023 9:32:00 AM	12/13/2023 9:33:00 AM
12/13/2023 9:32:28 AM	Success	Admin@FabricMSIT.on...	4	11.7440	10.5696	0	17.62%	Billable		12/13/2023 9:32:30 AM	12/13/2023 9:33:00 AM
12/13/2023 9:33:06 AM	Success	Admin@FabricMSIT.on...	4	10.2560	10.2560	0	17.09%	Billable		12/13/2023 9:33:00 AM	12/13/2023 9:33:00 AM
12/13/2023 9:30:51 AM	Success	Admin@FabricMSIT.on...	9	19.2480	9.6240	0	16.04%	Billable		12/13/2023 9:30:30 AM	12/13/2023 9:33:00 AM
			169	**314.0160**	**218.0848**	**0**	**363.47%**				

Pausing a Fabric Capacity

It might actually cost you more ..

Don’t blindly pause

- Especially if you’re hoping it will reduce costs
- Be mindful of ‘open balance’
- How long would you need to pause for it to be beneficial?

Throttling

- In a throttled state, this can add up
- What is the price for business continuity?

	Average Percentage Utilisation Next 24 Hours	Min Pause Hours for Saving	Cost of Pausing European F64
10,000%	3.47%	0.83	$10.13
20,000%	6.94%	1.67	$20.27
30,000%	10.42%	2.50	$30.40
40,000%	13.89%	3.33	$40.53
50,000%	17.36%	4.17	$50.67
60,000%	20.83%	5.00	$60.80
70,000%	24.31%	5.83	$70.93
80,000%	27.78%	6.67	$81.07
90,000%	31.25%	7.50	$91.20
100,000%	34.72%	8.33	$101.33
110,000%	38.19%	9.17	$111.47
120,000%	41.67%	10.00	$121.60
130,000%	45.14%	10.83	$131.73
140,000%	48.61%	11.67	$141.87
150,000%	52.08%	12.50	$152.00
160,000%	55.56%	13.33	$162.13
170,000%	59.03%	14.17	$172.27
180,000%	62.50%	15.00	$182.40
190,000%	65.97%	15.83	$192.53
200,000%	69.44%	16.67	$202.67
210,000%	72.92%	17.50	$212.80
220,000%	76.39%	18.33	$222.93
230,000%	79.86%	19.17	$233.07
240,000%	83.33%	20.00	$243.20
250,000%	86.81%	20.83	$253.33
260,000%	90.28%	21.67	$263.47
270,000%	93.75%	22.50	$273.60
280,000%	97.22%	23.33	$283.73
288,000%	100.00%	24.00	$291.84

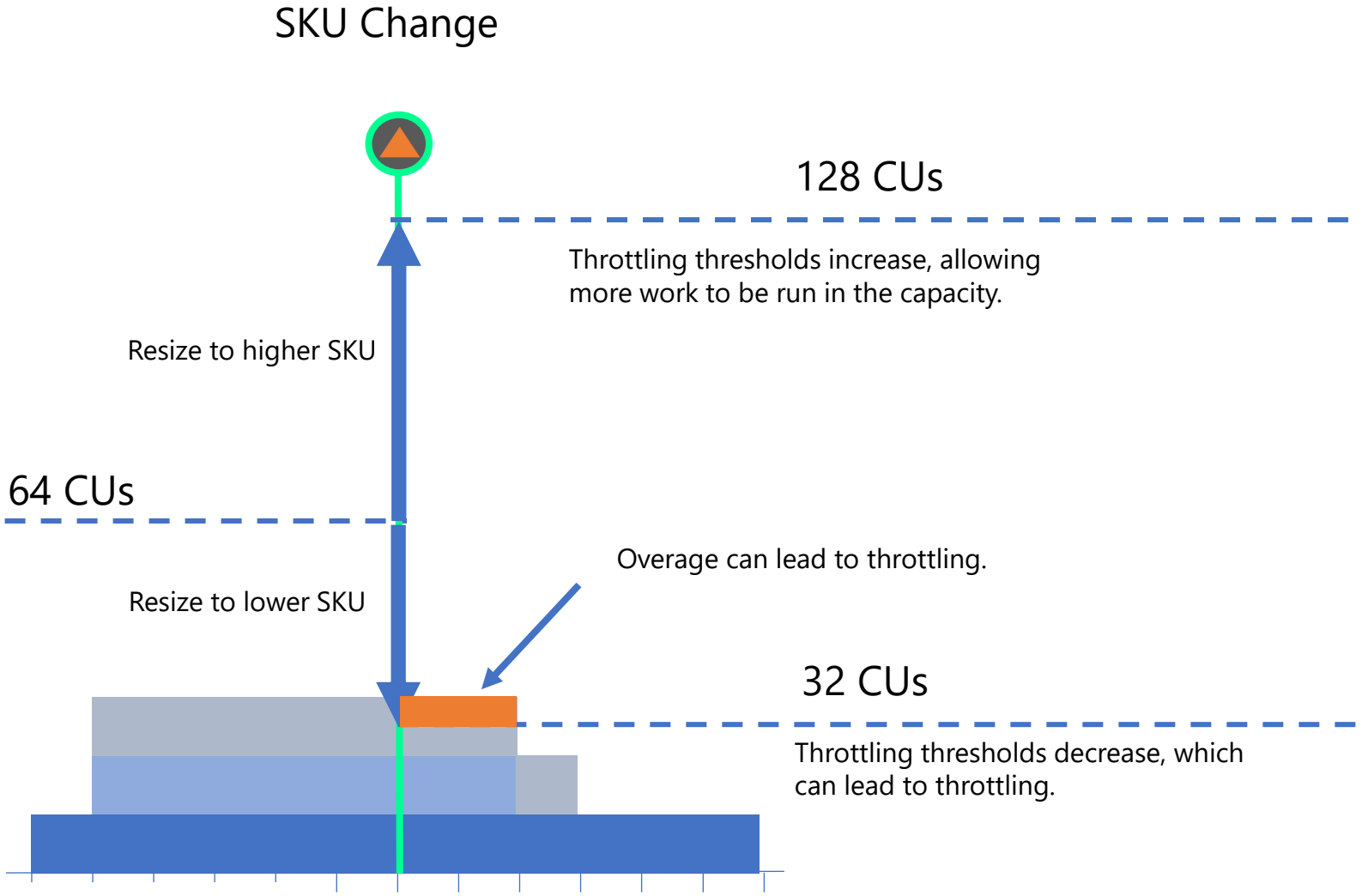
https://www.linkedin.com/pulse/fabric-billing-part-4-implications-pause-restart-matthew-farrow-fznse/

How Capacity Resize works

When a capacity is **resized**…

The allowed CUs per timepoint increase or decrease.

This changes the throttling allowed limits based on the new SKU's CUs and the throttling windows.



Key Insights

- Sizing up will incur the cost of the new SKU
- Sizing down could lead to more throttling
- Review your Throttling Thresholds before sizing down your SKU.

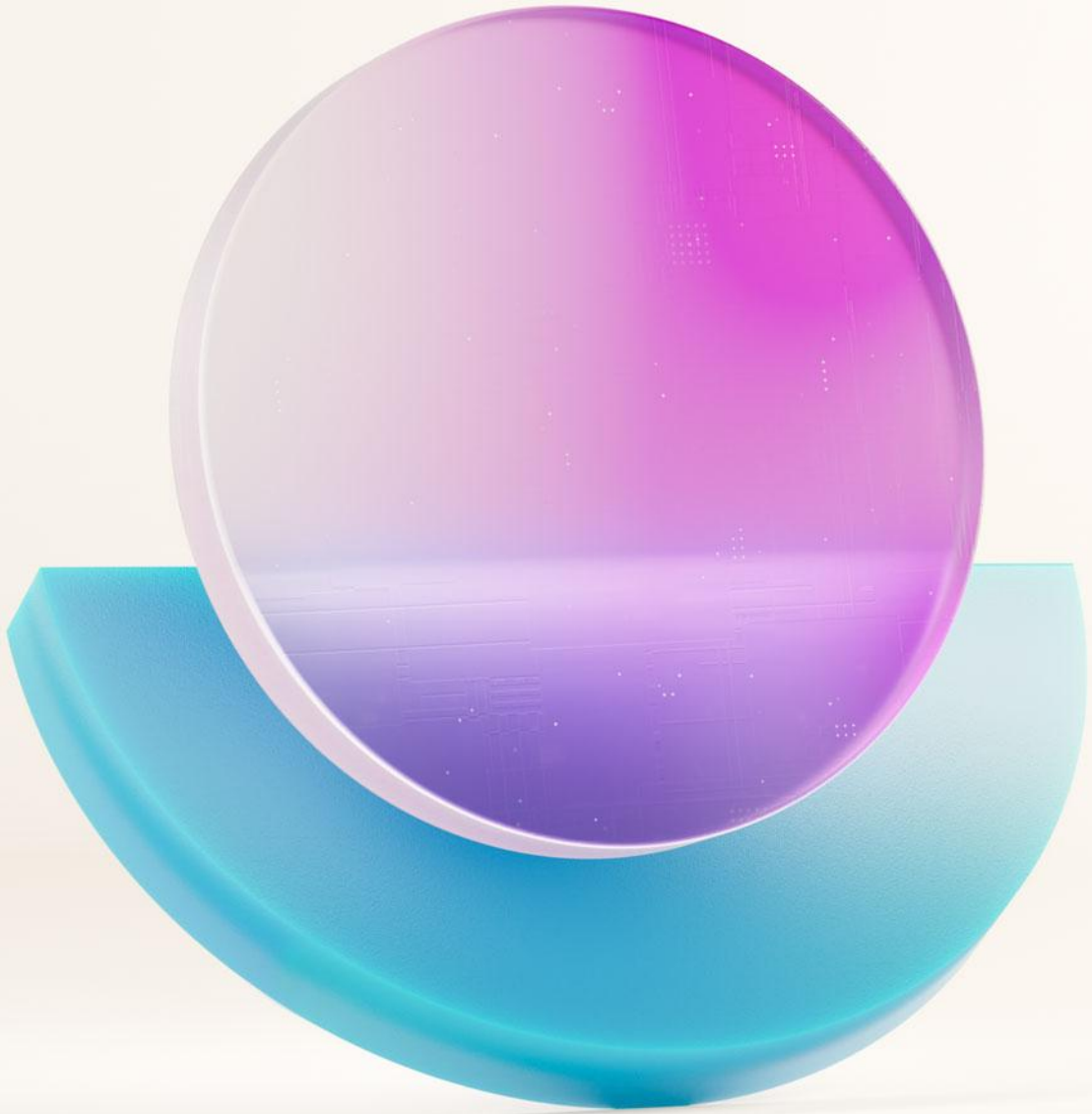
Scaling Capacities

Just be careful with this 😳

Scaling up a capacity that's smaller than F64 to a larger capacity happens almost immediately. **The capacity license updates usually take up to a day**, however some updates could take longer. During this time, Fabric Free users might see a request to upgrade to a Power BI Pro license when they try to access a Power BI report.

https://learn.microsoft.com/en-us/fabric/enterprise/scale-capacity#considerations-and-limitations

Microsoft Fabric
Community Conference



Protecting Capacities

Surge Protection v1 – Background usage limits

GA (Jul 2025)

Enabling Capacity Admins to get ahead of throttling. (blog || docs)

Simple experience that limits overuse by background jobs

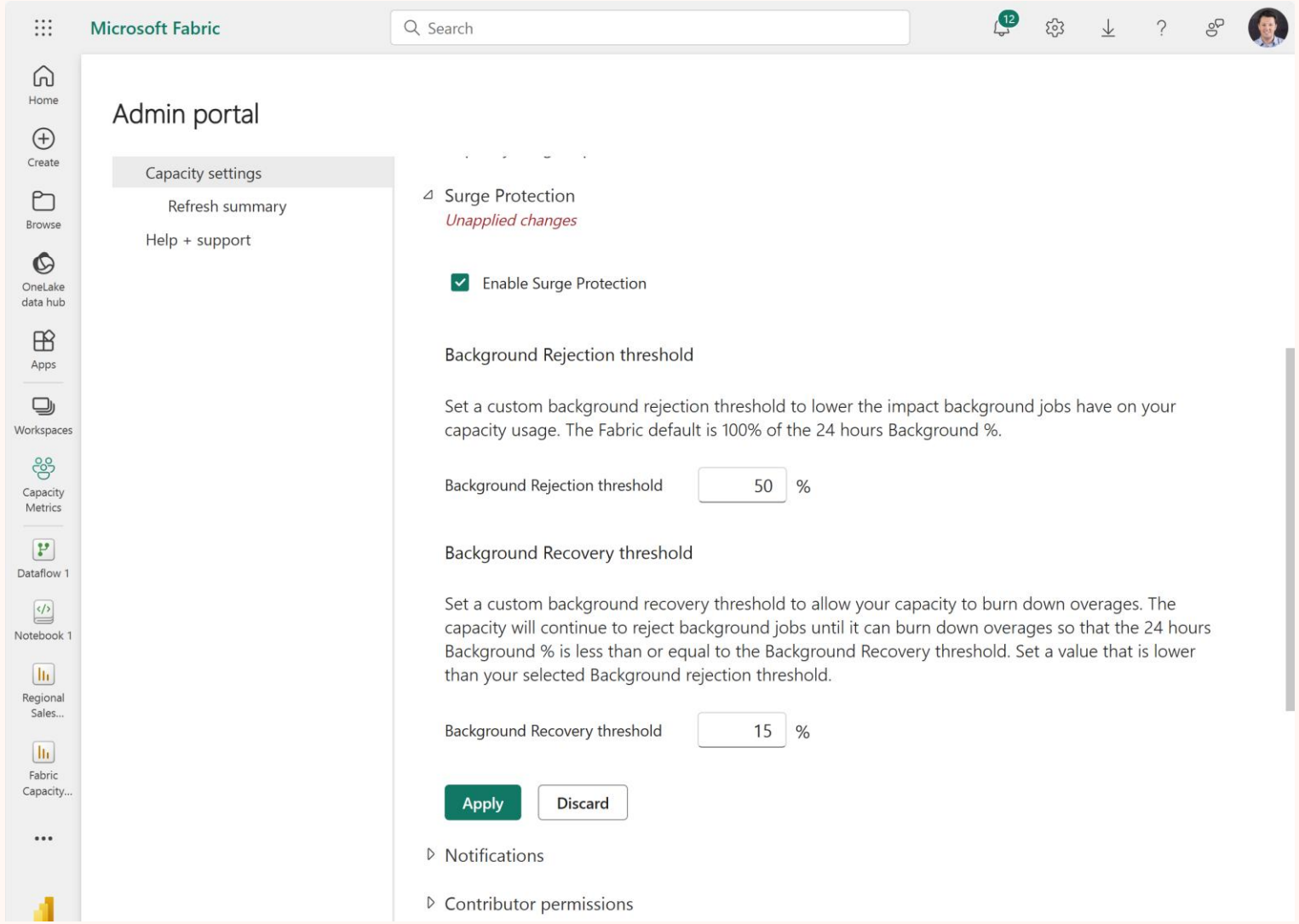
- Throttles background jobs before a full 24-hours of CUs is consumed.
- Helps to **protect interactive** usage like Power BI

Recovery limit

- Keep throttling until the capacity is 'healthy' as defined by the customer
- Helps prevent a capacity from immediately being throttled again

Impact

- Throttling background jobs will help 40-60% of capacities experiencing Interactive Rejections.



Surge Protection V2 – Workspace limits

PuPr (Jan 2026)

Enabling Capacity Admins to get ahead of throttling.

Simple experience that limits overuse by workspaces

- **Blocks a workspaces** for 24-hours if it exceeds a limit set by the Capacity Admin
- **Mission critical workspaces** are excluded from surge protection limits

Self-healing

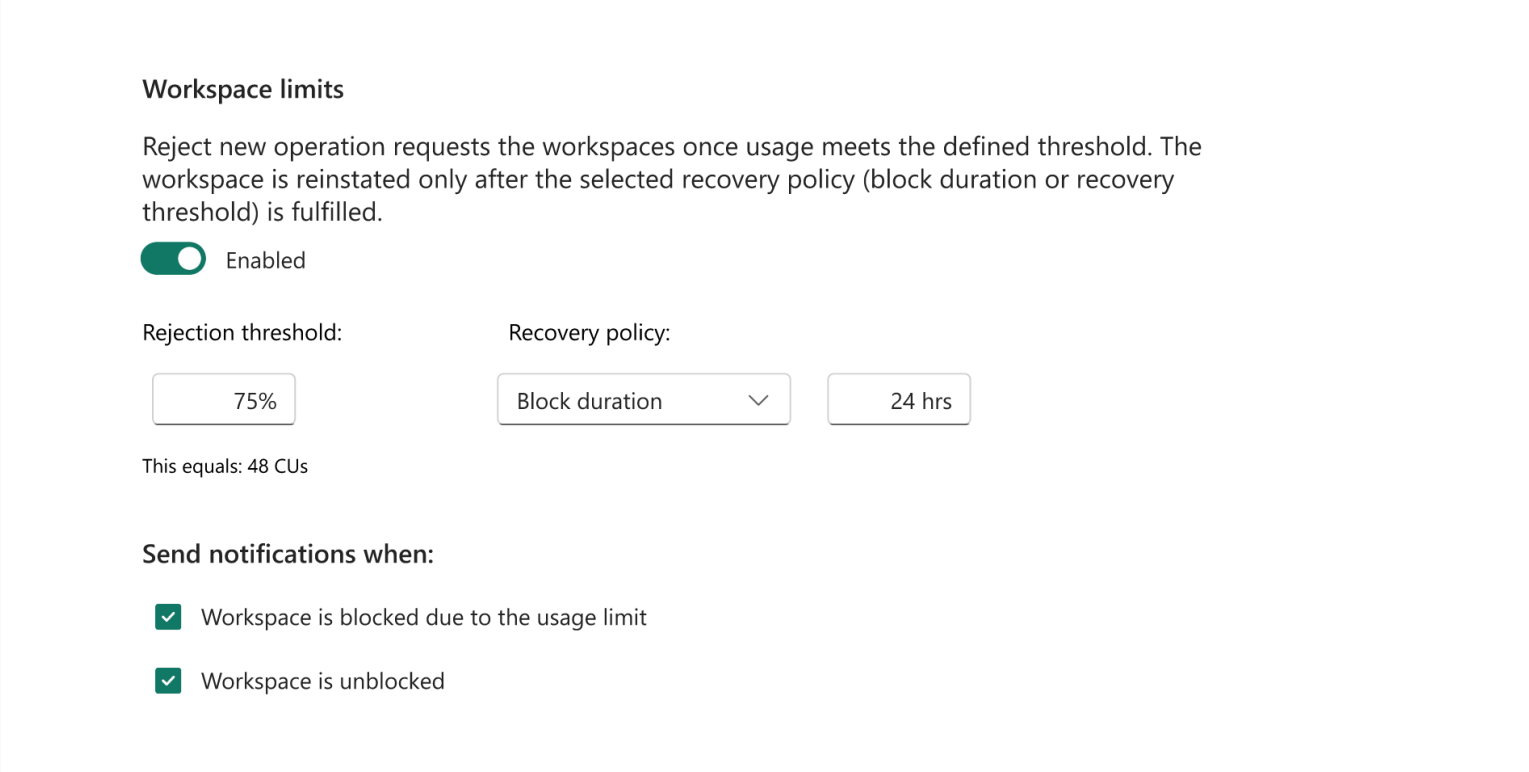
- Automatically restores workspace after 24 hours.

Capacity Admins are in control

- Can add / remove blocked workspaces
- APIs are provided

Impact

- Limit blocking to mis-behaving workspaces allowing the rest of the org to continue working without interruption.



Autoscale billing for Apache Spark

GA (July 2025)

Enable everyone in your org to use Spark and manage its cost (blog || docs)

Serverless style billing for Spark jobs

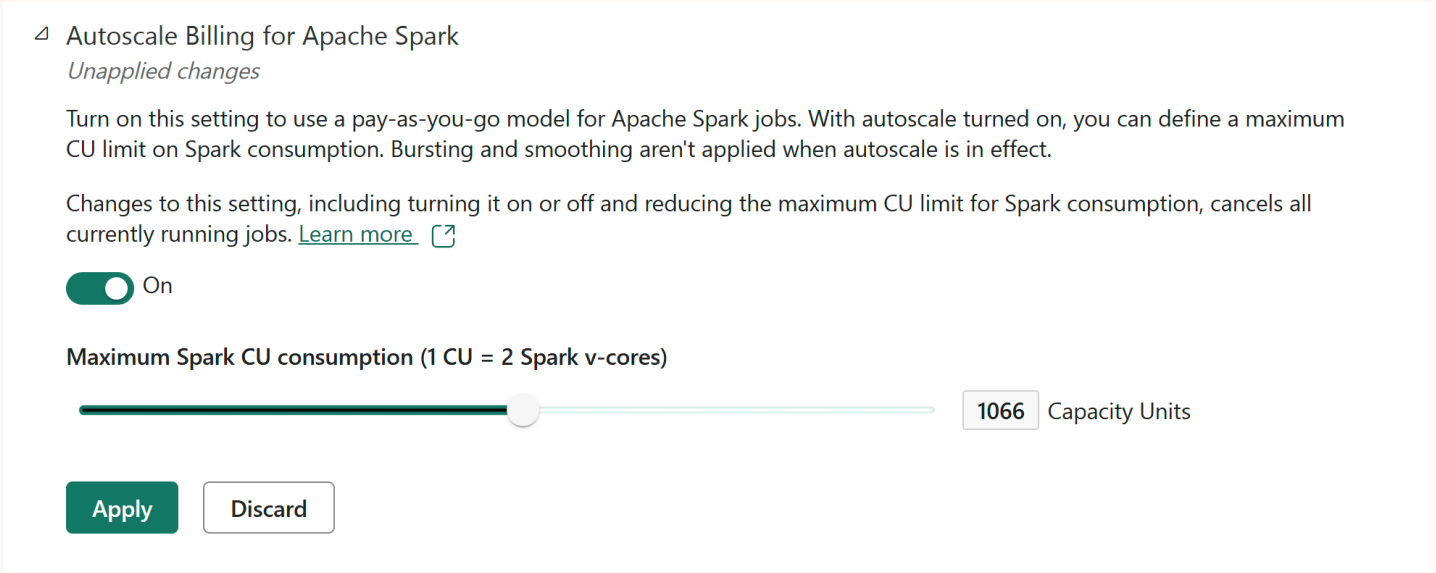
- Capacity admins can opt-in
- Set a max limit on CU used by Spark
- Only pay for what you use
- Spark manages the limit ensuring pools don't over consume
- Observability in a new metrics app page

Spark jobs are billed separately

- Jobs are billed when they execute
- Cost is at **Pay-as-you-go rate**
- Must also have an active capacity
- If Spark calls other workloads, like OneLake, those costs are billed to the capacity.

Impact

- Isolate your spiky Spark jobs from the rest of capacity compute
- Helps save on costs and reduce throttling



But .. Is it right for you?

- Part 1
- Part 2
- Part 3

Autoscale billing for Apache Spark

GA (July 2025)

Monitor consumption in Metrics app

New Autoscale Compute for Spark page

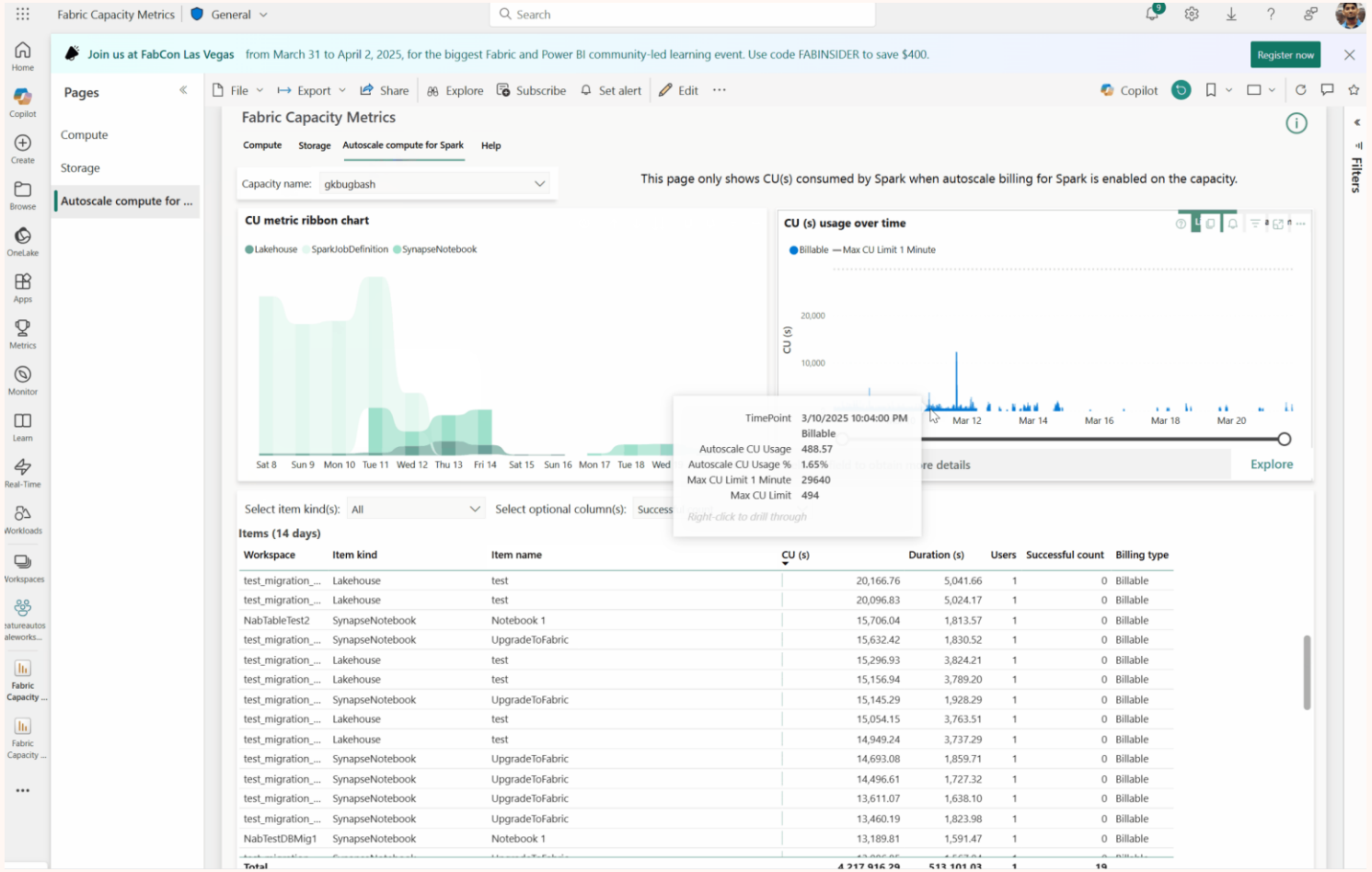
- Shows Spark CU consumed through Autoscale
- Easy to track to the configured autoscale limit.

Familiar experience

- Same experiences as for the capacity compute page.
- Provides drill down experience to see operation details

Impact

- Clearly understand the compute specific to Spark that will be reflected on your bill



Fabric Copilot Capacity

GA (Nov 2024)

Enable everyone in your org to use Copilot and manage its cost (blog || docs)

Enable everyone to use Copilot

- All users can use Copilot experiences
- Consumption of Copilot goes to only the selected capacity

Select who can use a Copilot capacity

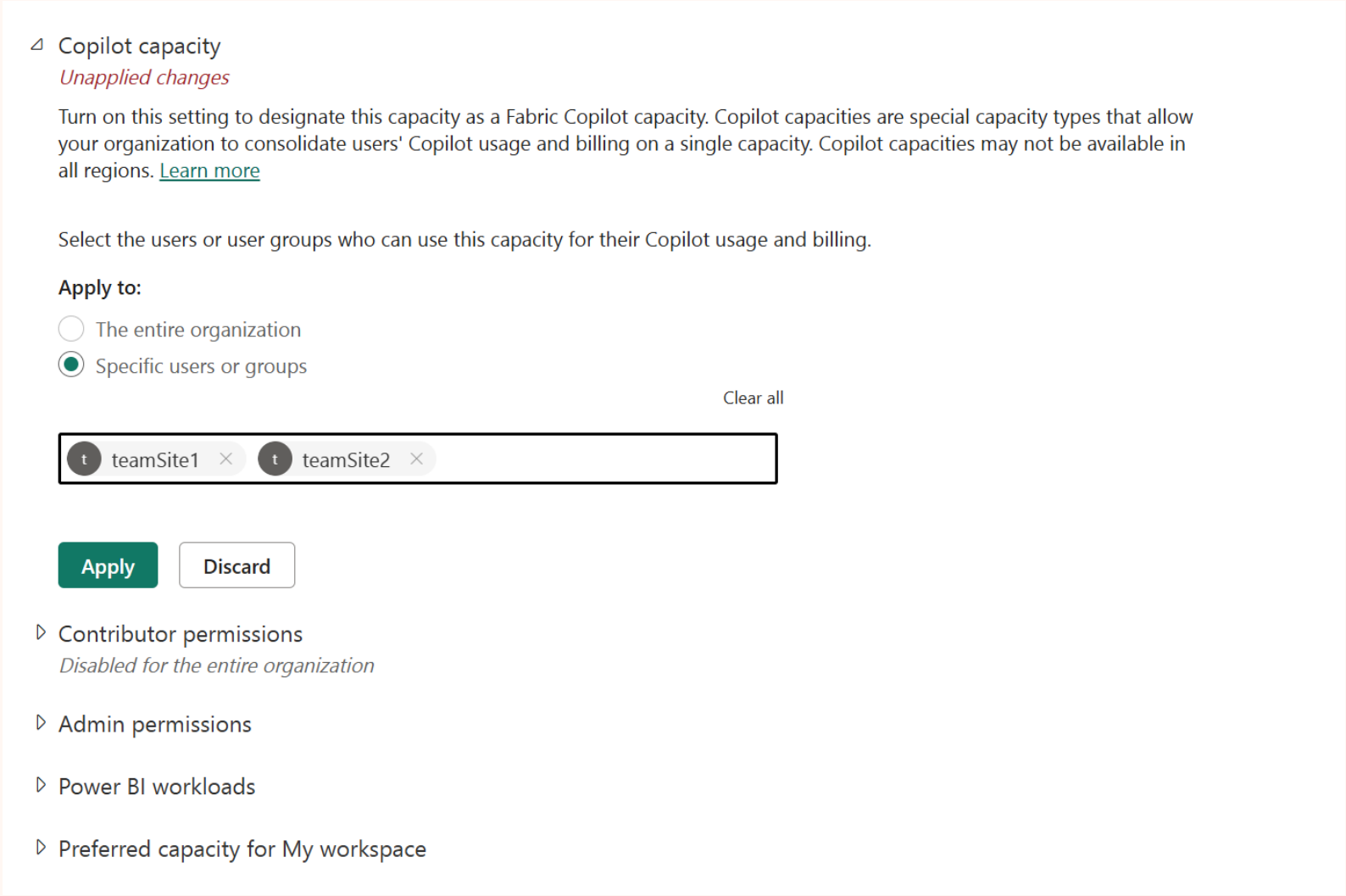
- Select the Users or groups who use a specific Copilot capacity
- A user can have only one Copilot capacity
  - Newest one matters..

No longer just P & F SKUs

- Pro, Premium Per User, and Trial

Tenant Setting

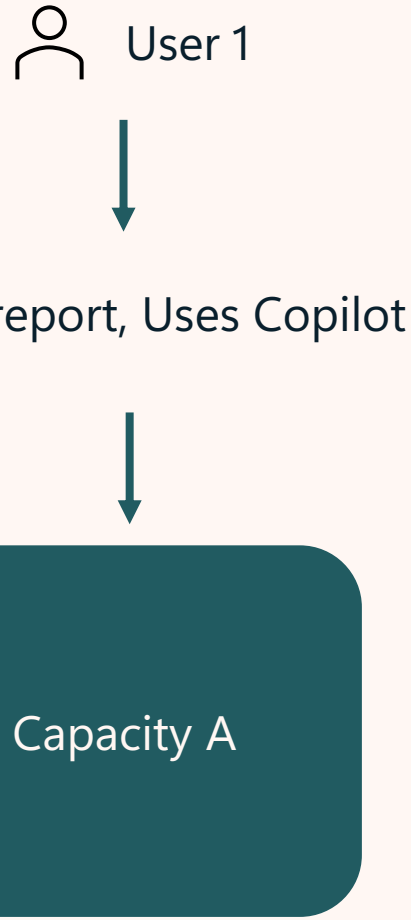
- Restrict who can configure
- "Capacities can be delegated .."



Fabric Copilot capacity

GA (Nov 2024)

Normally, copilot usage applies to the capacity the content is in.

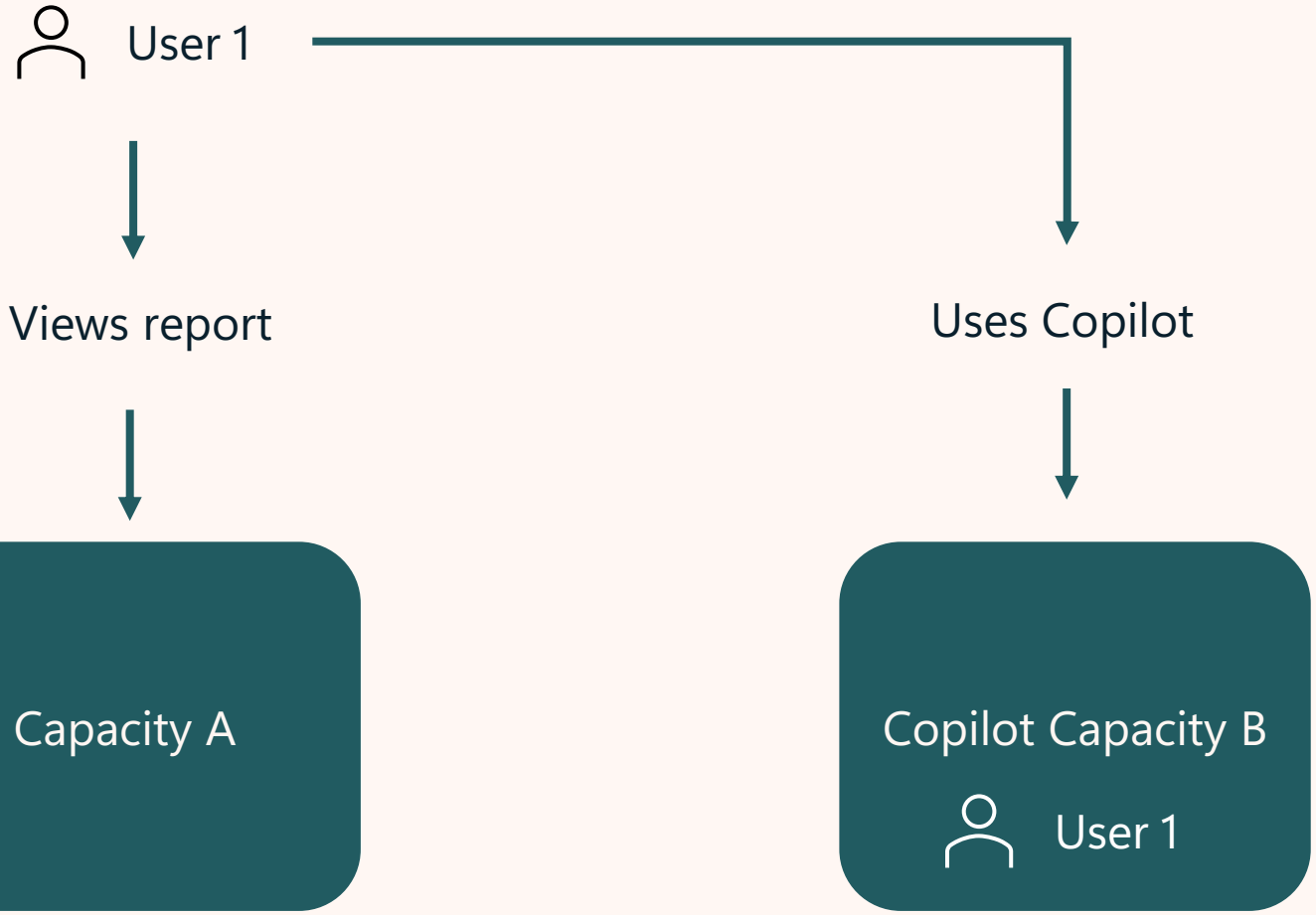


User 1

Fabric Copilot capacity

GA (Nov 2024)

After a user is added to a copilot capacity, the user's copilot usage is consumed from their copilot capacity



Protecting your capacity from Copilot usage

Copilot in Fabric operations are background, so no immediate spikes

- But when they do .. it has a 24h impact

Be mindful of who has access to Copilot skills and educate!

- Copilot requests trigger other operations too!

Track usage for Power BI Desktop, ensuring usage doesn't interfere with key workloads

- ℹ️ Daily check for "Power BI Session Desktop" item name in Metrics App
- Any workspace with permissions (Contributor), on Capacity that allows Copilot usage
- If user is assigned a Copilot Capacity, it automatically goes here

Track usage for Copilot in Service, ensuring usage doesn't interfere with key workloads

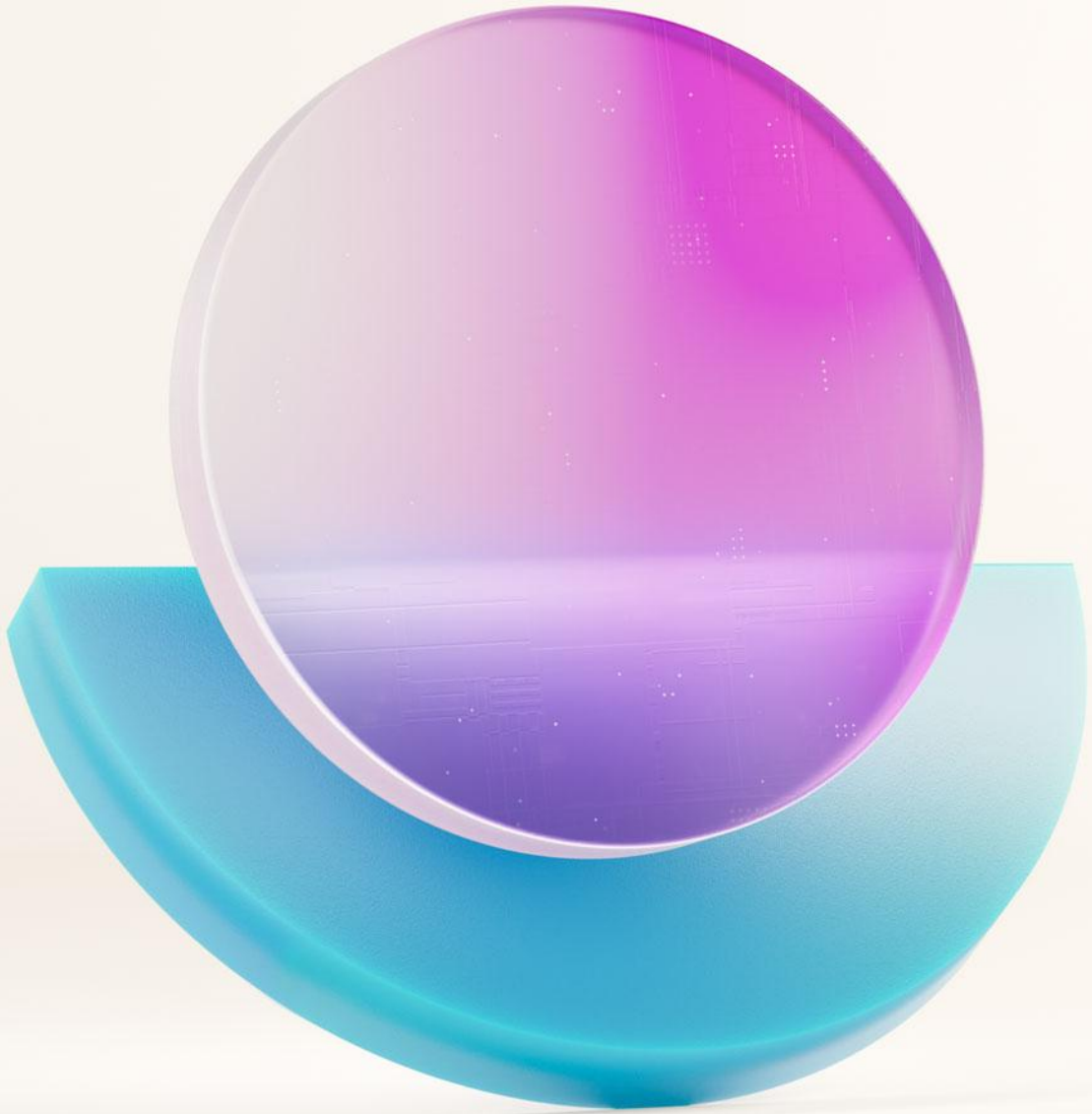
- ℹ️ Daily check for "Power BI Session Web" item name in Metrics App
- Any workspace with permissions (Viewer), on Capacity that allows Copilot usage
- If user is assigned a Copilot Capacity, it automatically goes here

Protecting your capacity from Copilot usage

Options to ensure the health of your capacity

- Fabric Copilot Capacity enabled for user base (by region, department, ..)
  - If possible, set up new Security Groups (avoid cross pollination)
- When Capacity throttles/rejects, Copilot no longer works
  - But everything else does .. 😼

Microsoft Fabric
Community Conference



Monitoring Capacities

Capacity Chargeback Reporting

Public Preview

Allocate costs to those who use your capacity (docs)

Helps allocate costs across your org

- Built-in turnkey reporting
- Rolls up usage per workspace / item / user

Focuses on % utilization

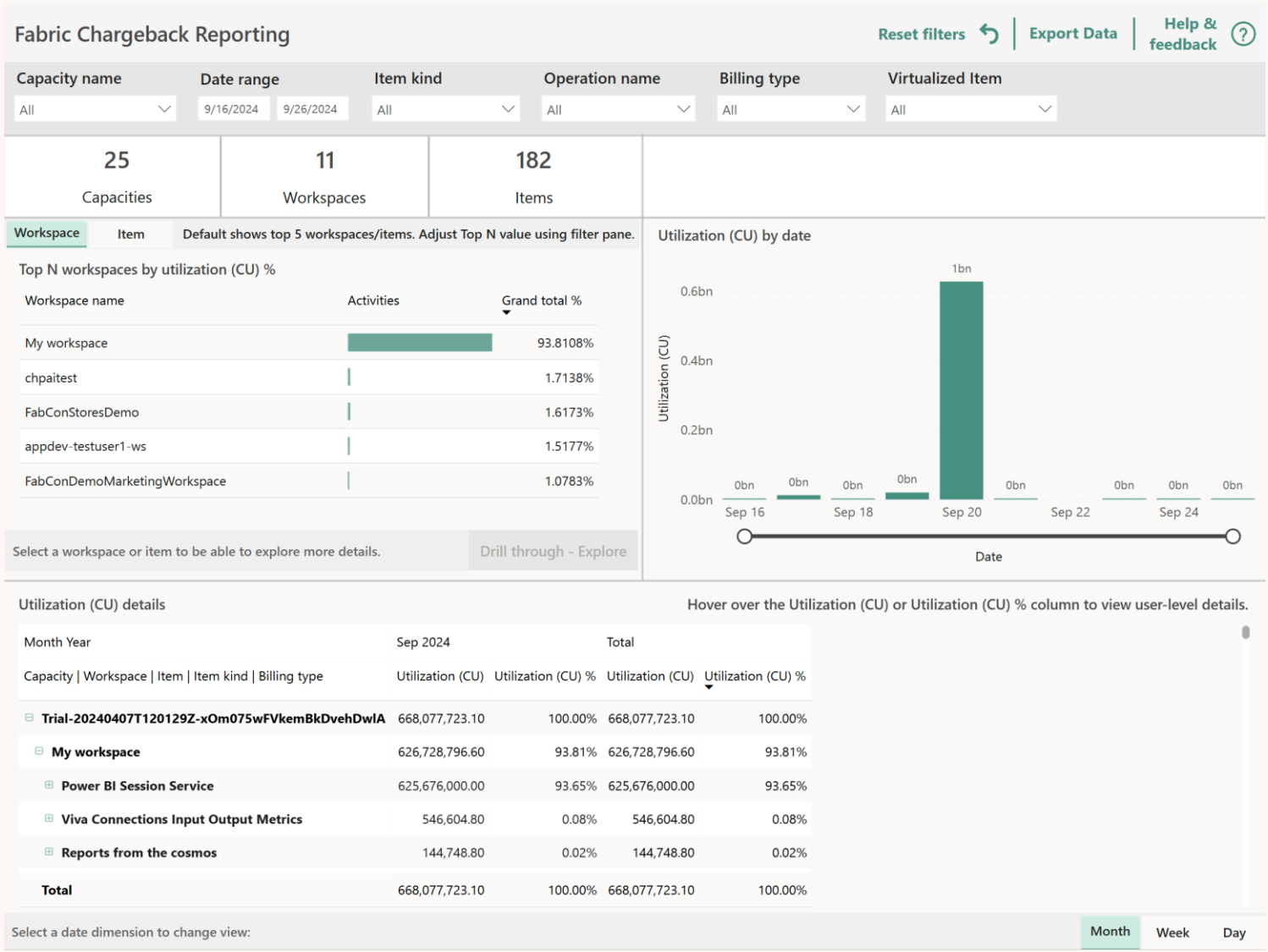
- Orgs need to look at cost in Azure billing and then allocate that cost to their content owners.

Impact

- Turn-key solution for Fin-ops with reduction in need to build custom solutions

Future

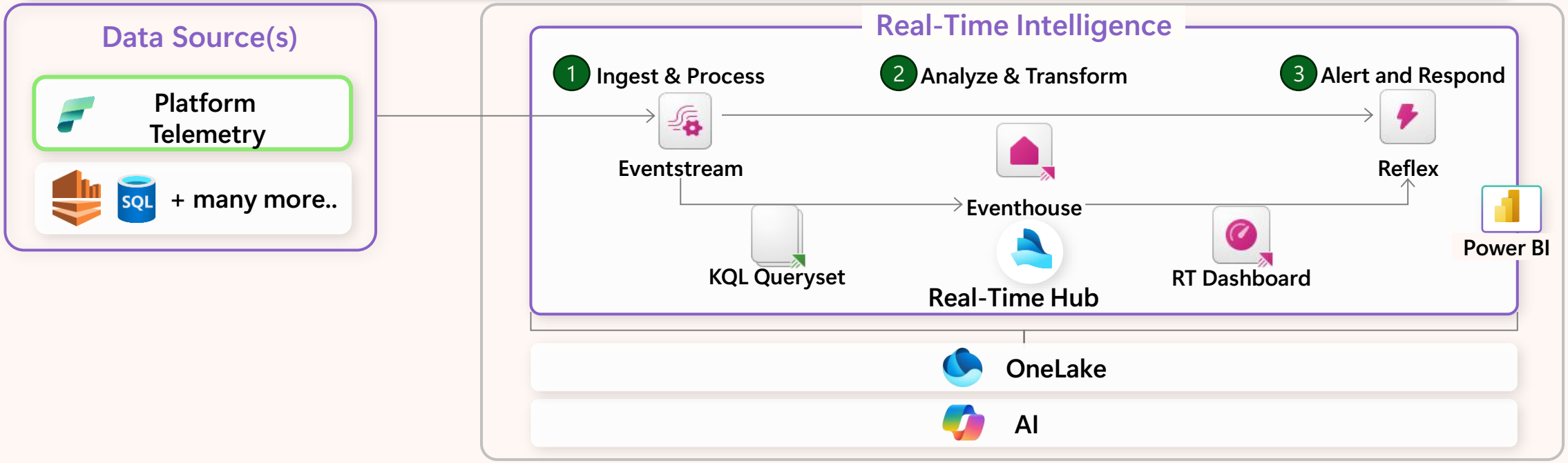
- Domain support
- Tags support

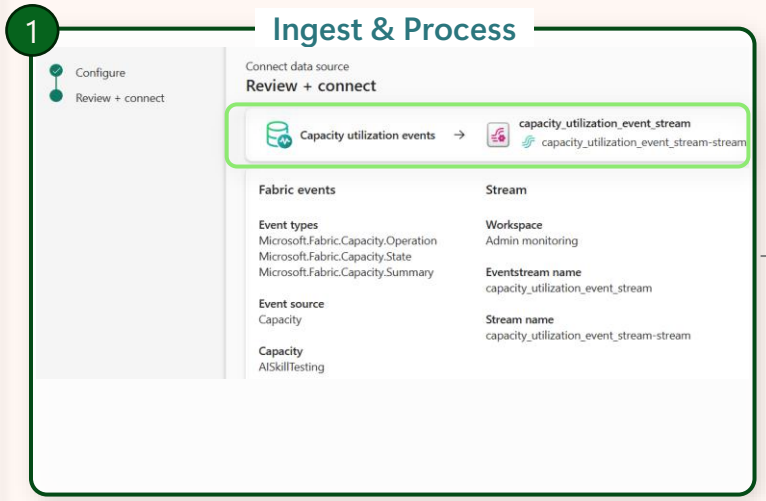


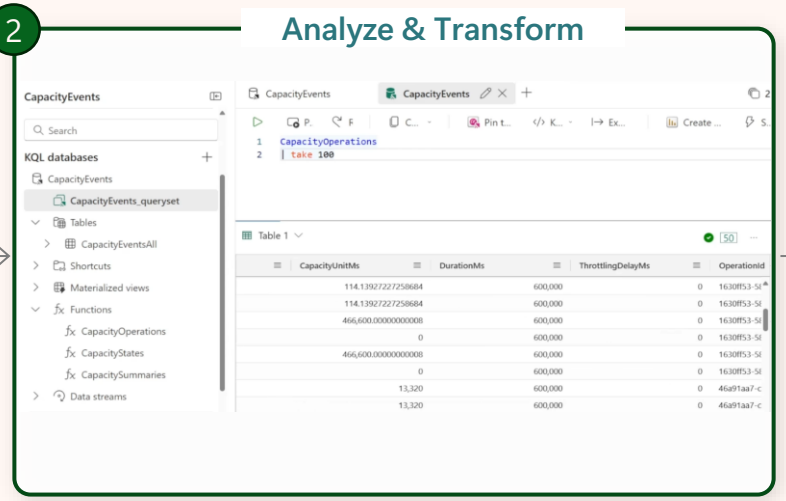
Capacity Platform Data in Realtime Hub (RTH)

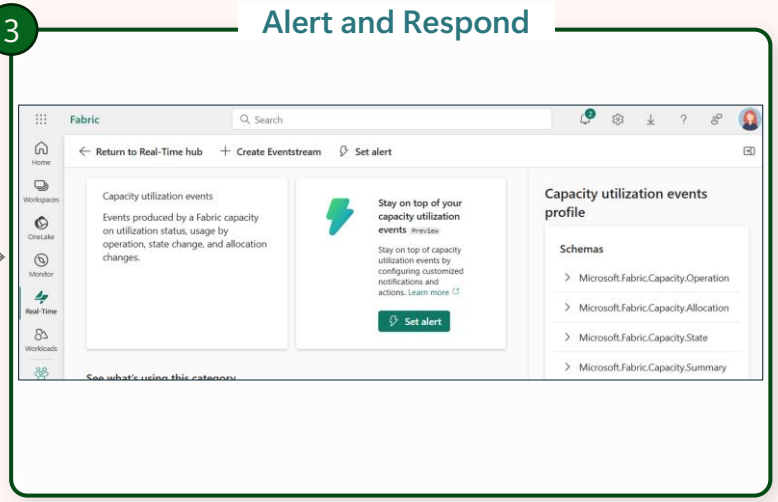
Sneak Peek

Direct access to the same Capacity Platform data used in Capacity Metrics with flexible routing and alerting capabilities

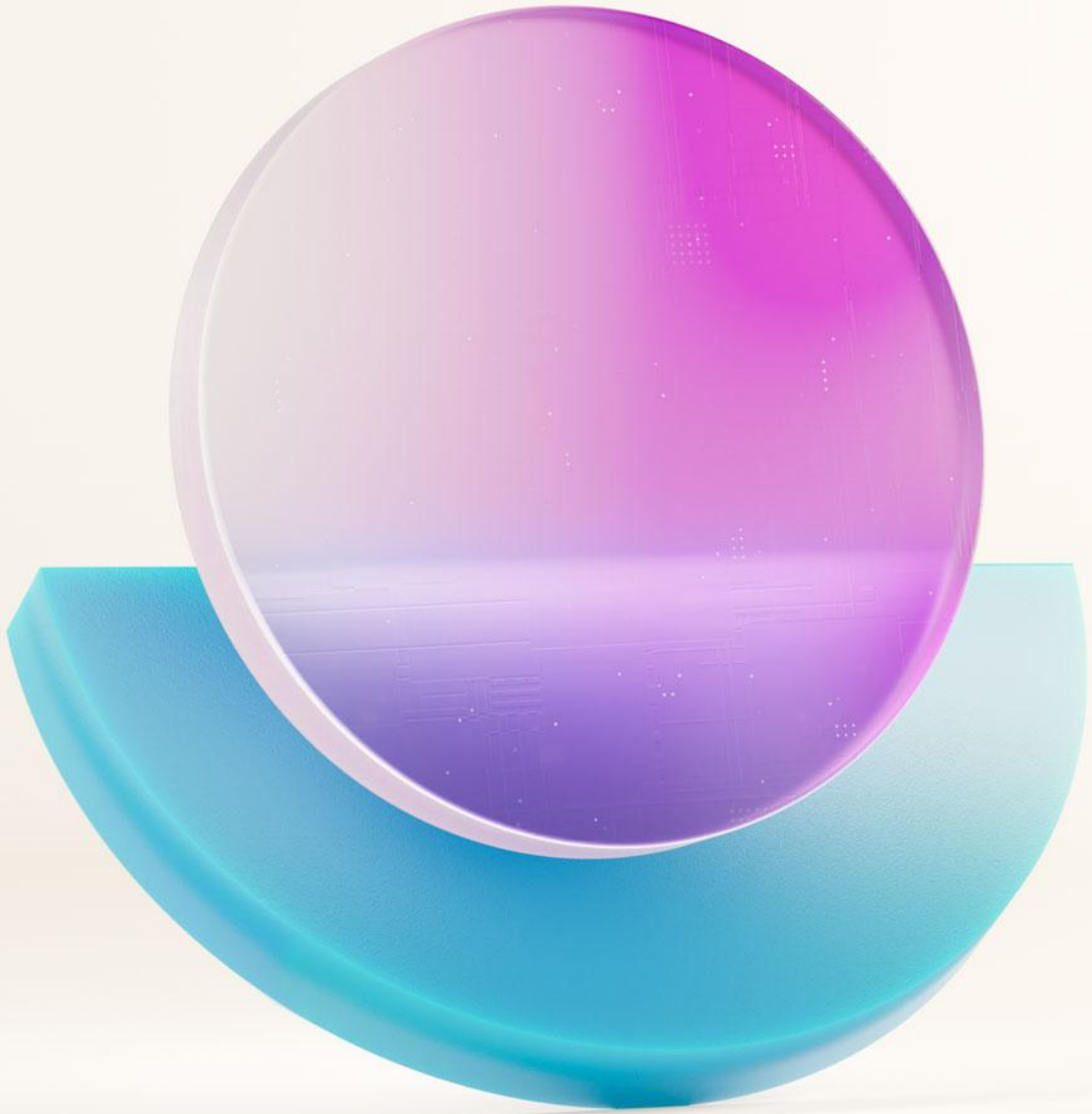






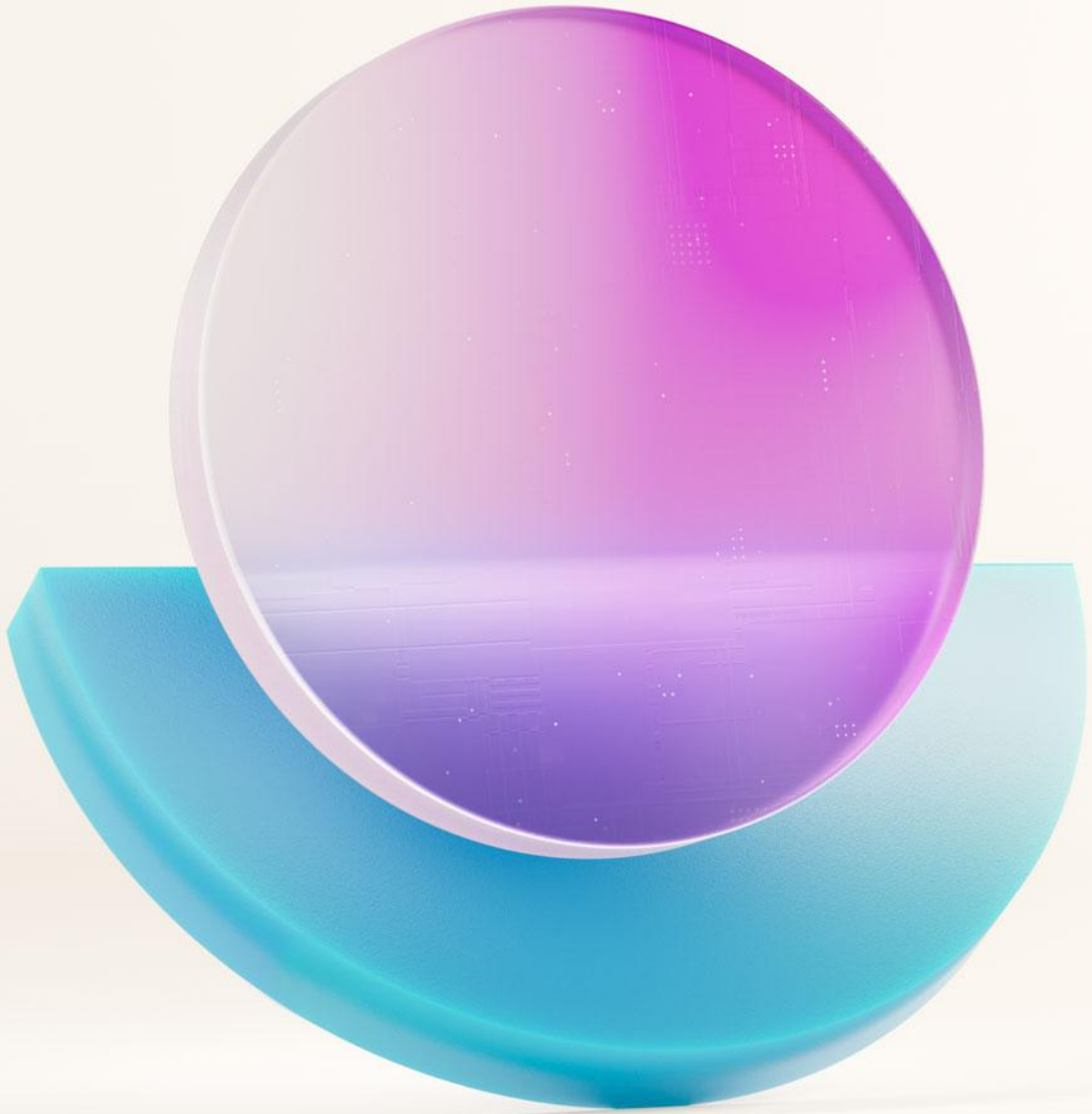


Microsoft Fabric
Community Conference



Exercise

Microsoft Fabric
Community Conference



Wrapping up

Best practices for deploying your capacities

Plan your capacities

Ensure your capacities are correctly sized

Use dedicated capacities to optimize quality of experience and costs

Isolate production, development, testing in separate capacities

Budget for variability

Manage Resources

Enable Surge Protection

Monitor usage using metrics app

Adjust workload limits like pools, memory, and timeouts

Share best practices with colleagues

Optimize experiences and costs

Consider Autoscale billing for Spark

Consider Fabric Copilot Capacities

Leverage pause/resume **appropriately**

Resize capacities as needed

Move problematic content to rescue, time-out, or testing capacities

How do you prevent overloading your capacities...

Use multiple capacities and strategies to operate your capacities

Capacity A	Capacity B	Capacity C	Capacity D
For general purpose compute needs Sized for typical needs	For Self-Service Reporting workloads	For large periodic workloads Paused when not needed	For testing newly built content Small size to avoid large costs

- Surge protection
- Resize
- Pause and Resume
- Autoscale Billing for Spark, Copilot Capacity
- OneLake Shortcuts

Loved it? Learned something? Tell us! Share your feedback in just 1 minute

Fabric Capacities, beyond the obvious

Want to check the slides again?

https://github.com/BenniDeJagere/Presentations
/{Year}
/{YYYYMMDD}_{Event}

PBIG
POWER BI GEBRUIKERSGROEP

Slides

https://github.com/BenniDeJagere/Presentations/{Year}/{YYYYMMDD}_{Event}