Fabric Capacities, beyond the obvious

Benni De Jagere

Slides

CAT

Premium sponsors

dataon
Power your data

twoday

INVIXO

TIMEXTENDER

Standard sponsors

cognitech
Din vej til bedre beslutninger

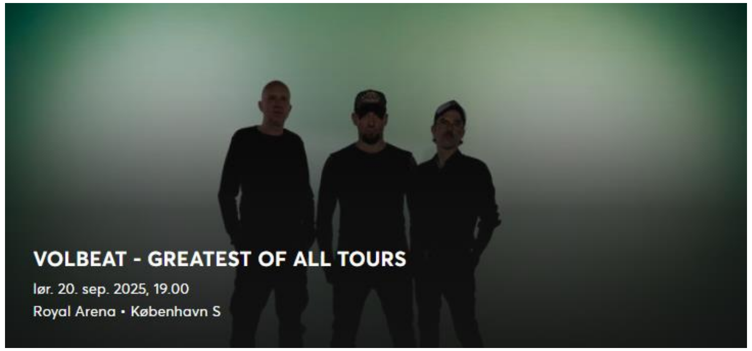
itm8®

ANALYTICS
masterminds

Eddytor

Columbus

Raffle Prizes



VOLBEAT - GREATEST OF ALL TOURS

lør. 20. sep. 2025, 19.00

Royal Arena • København S

EUROPEAN
MiCROSOFT
FABRiC
Community Conference

Microsoft

Benni De Jagere

Principal Program Manager | Fabric Customer Advisory Team (FabricCAT)

Fabric CAT

dataMinds.be Member

/bennidejagere

sessionize /bennidejagere

/bennidejagere

#SayNoToPieCharts

Fabric Capacity Core Concepts

Scalable Compute Units

- **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

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

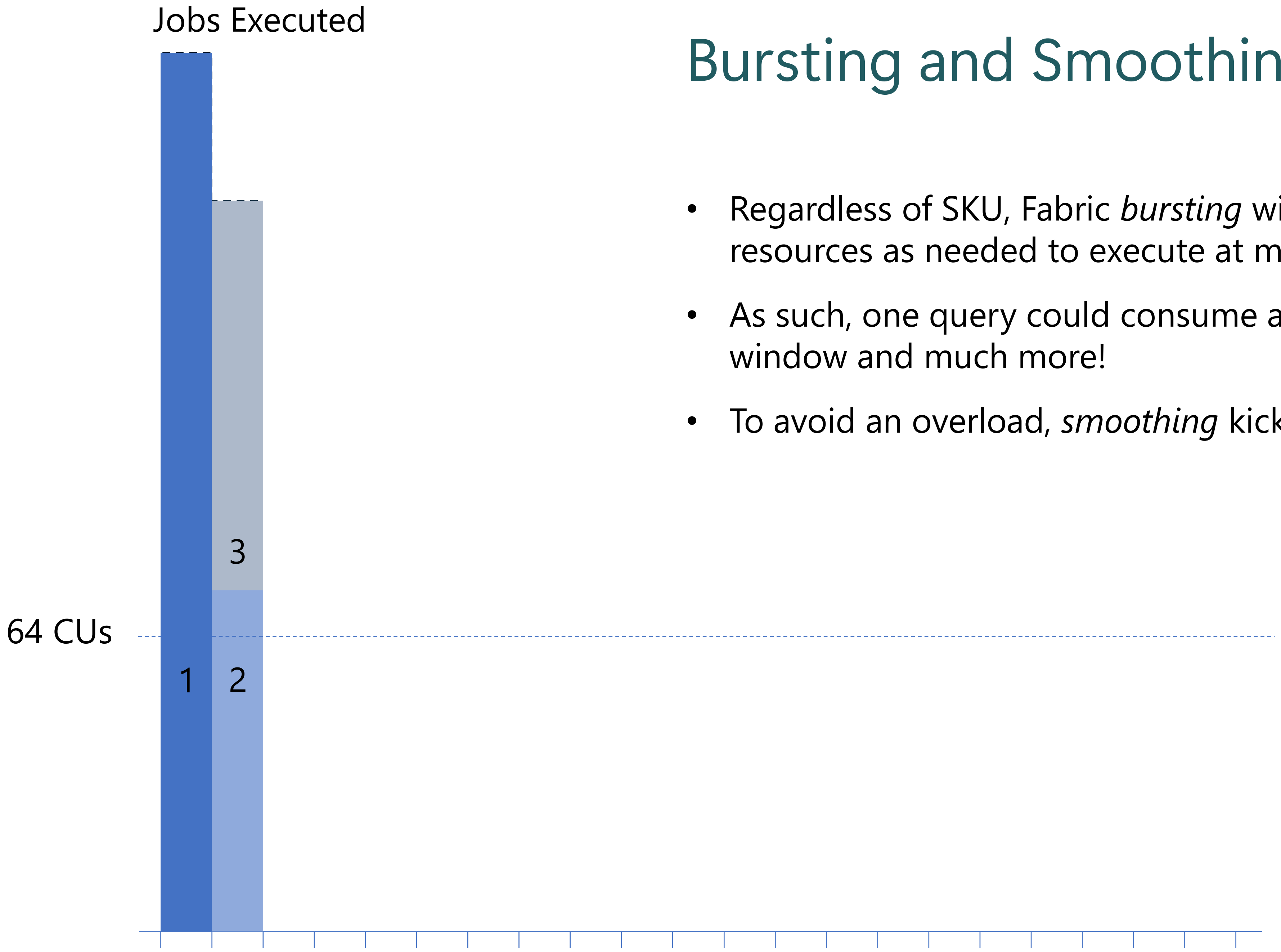
Self-Managing with Bursting and Smoothing

- **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

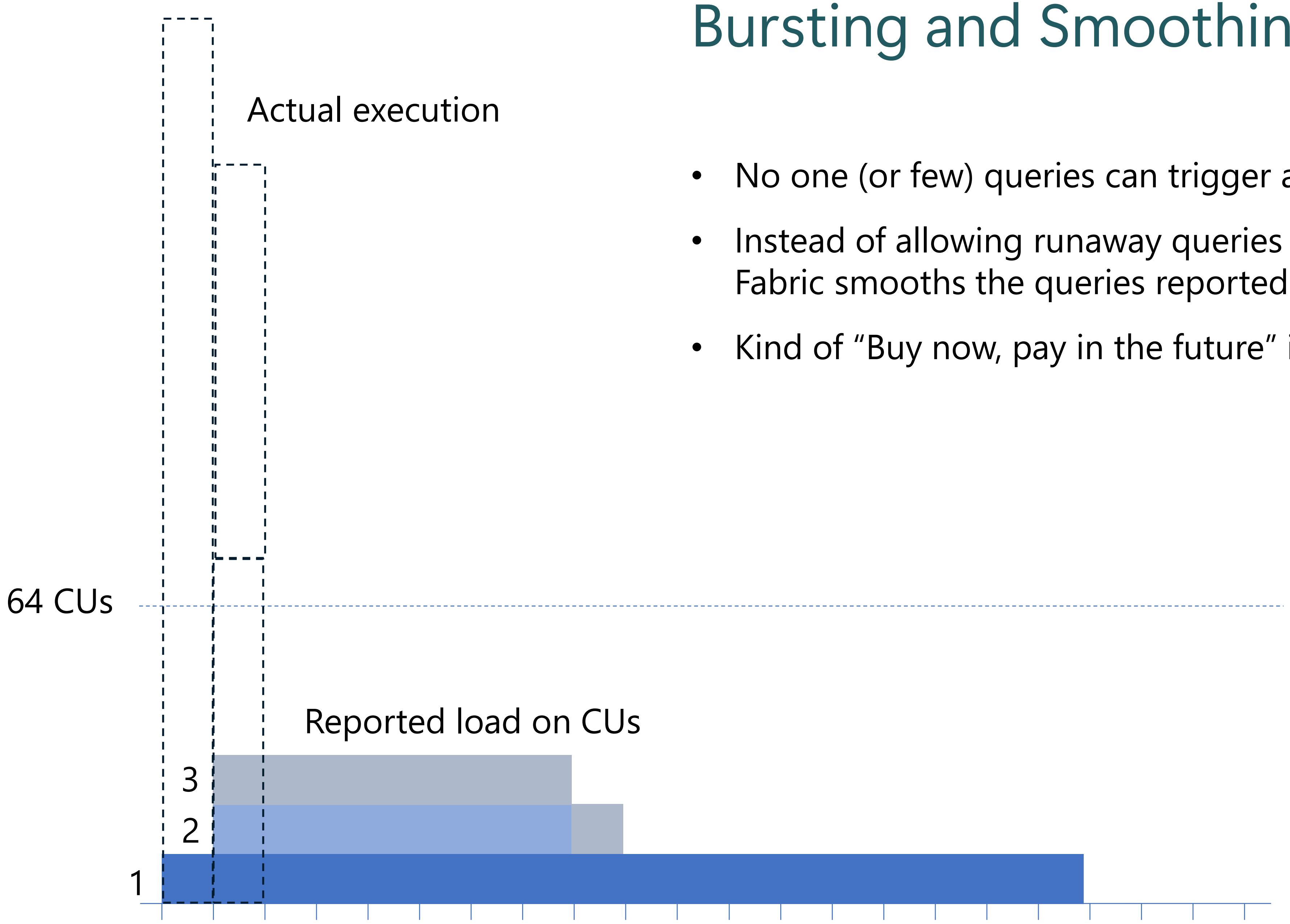
Bursting and Smoothing



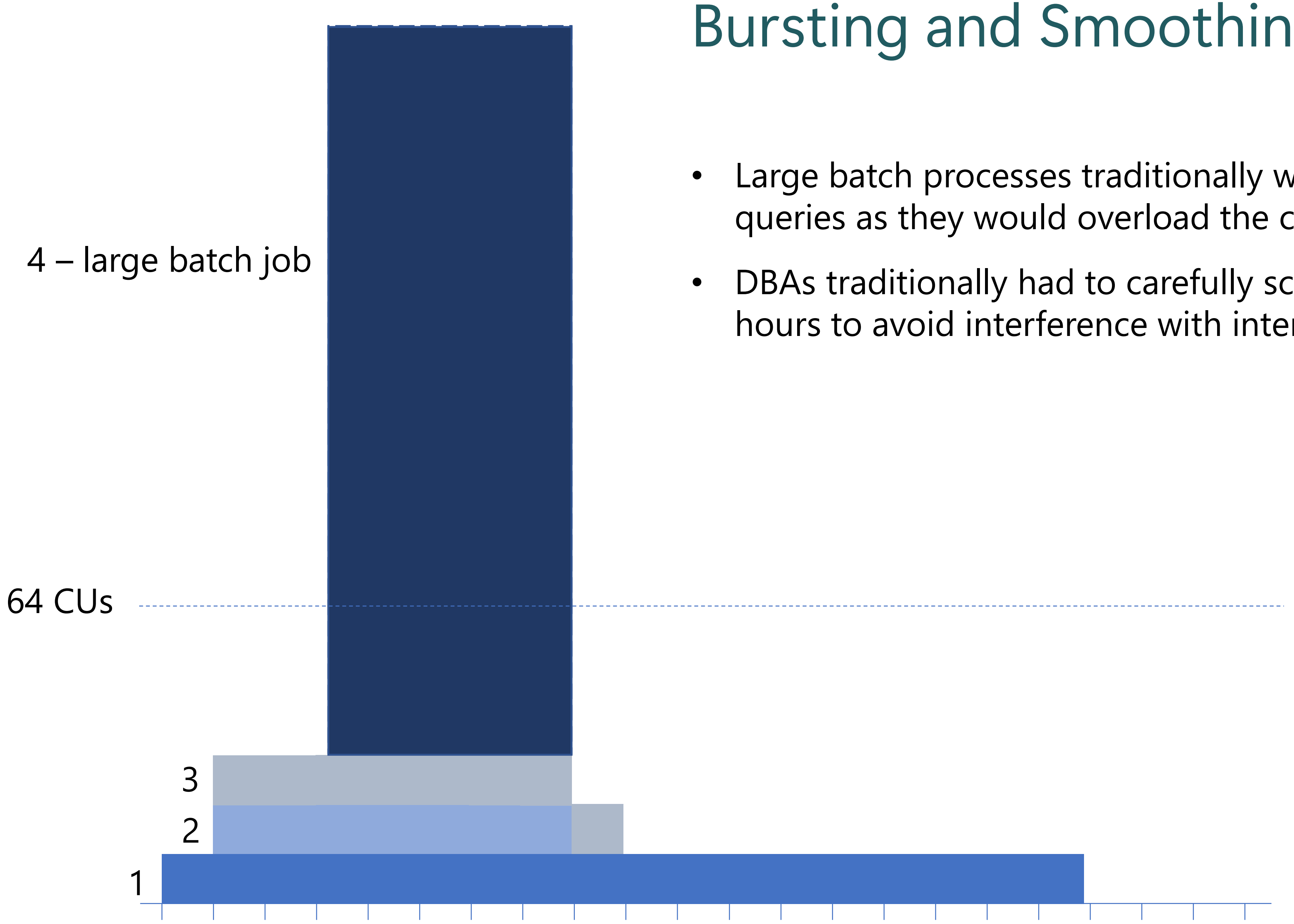
- Regardless of SKU, Fabric *bursting* will automatically allocate resources as needed to execute at maximum performance
- As such, one query could consume all the quota of a single time window and much more!
- To avoid an overload, *smoothing* kicks in

Bursting and Smoothing

- No one (or few) queries can trigger an overload
- Instead of allowing runaway queries to create a local overload, Fabric smooths the queries reported usage to future time windows
- Kind of "Buy now, pay in the future" installment plan



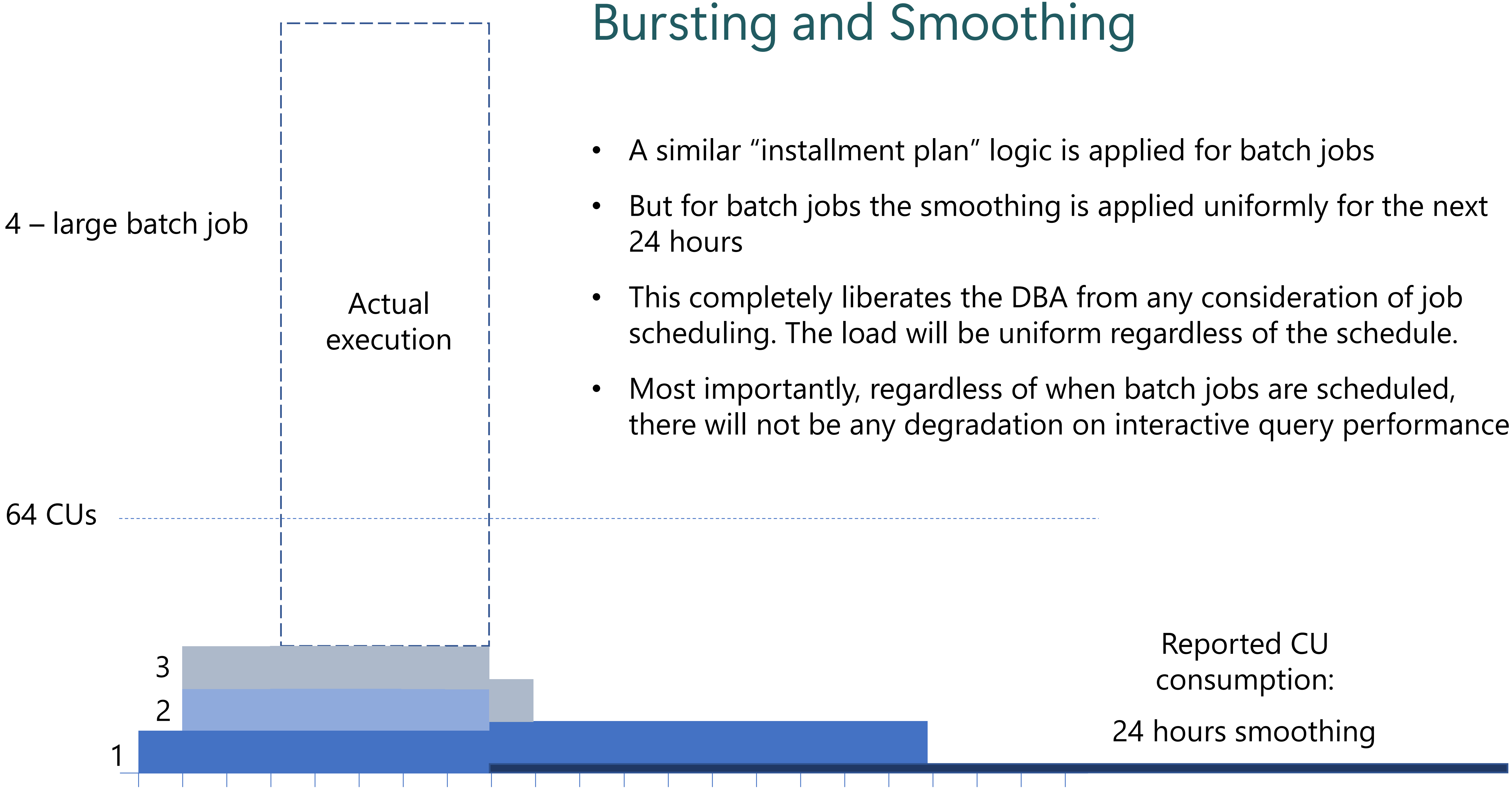
Bursting and Smoothing



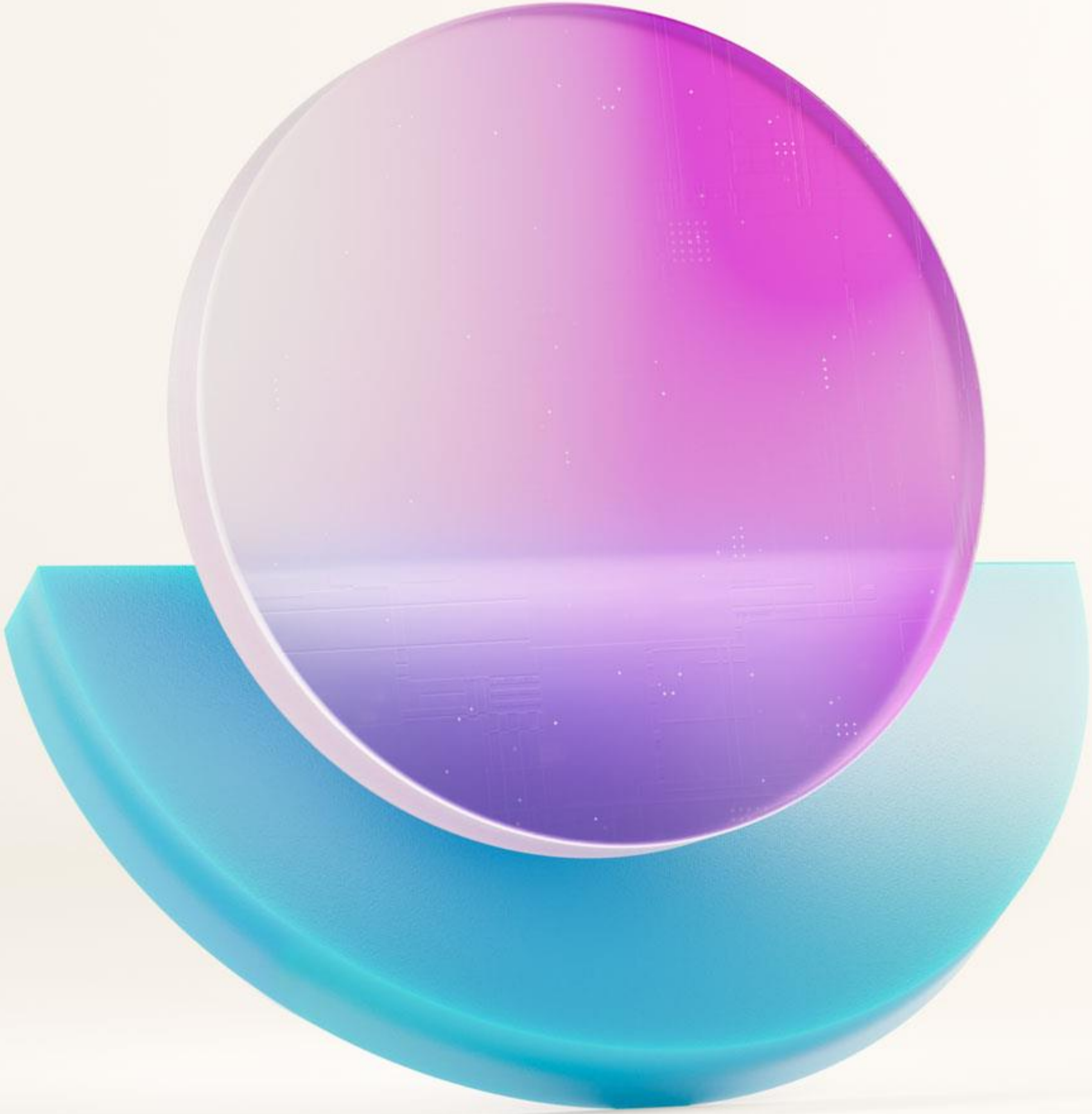
- Large batch processes traditionally were a threat to interactive queries as they would overload the compute resource
- DBAs traditionally had to carefully schedule these jobs to off-hours to avoid interference with interactive user experiences

Bursting and Smoothing

- A similar “installment plan” logic is applied for batch jobs
- But for batch jobs the smoothing is applied uniformly for the next 24 hours
- This completely liberates the DBA from any consideration of job scheduling. The load will be uniform regardless of the schedule.
- Most importantly, regardless of when batch jobs are scheduled, there will not be any degradation on interactive query performance



Microsoft Fabric
Community Conference



Capacity Planning

Fabric SKU Estimator

Public Preview

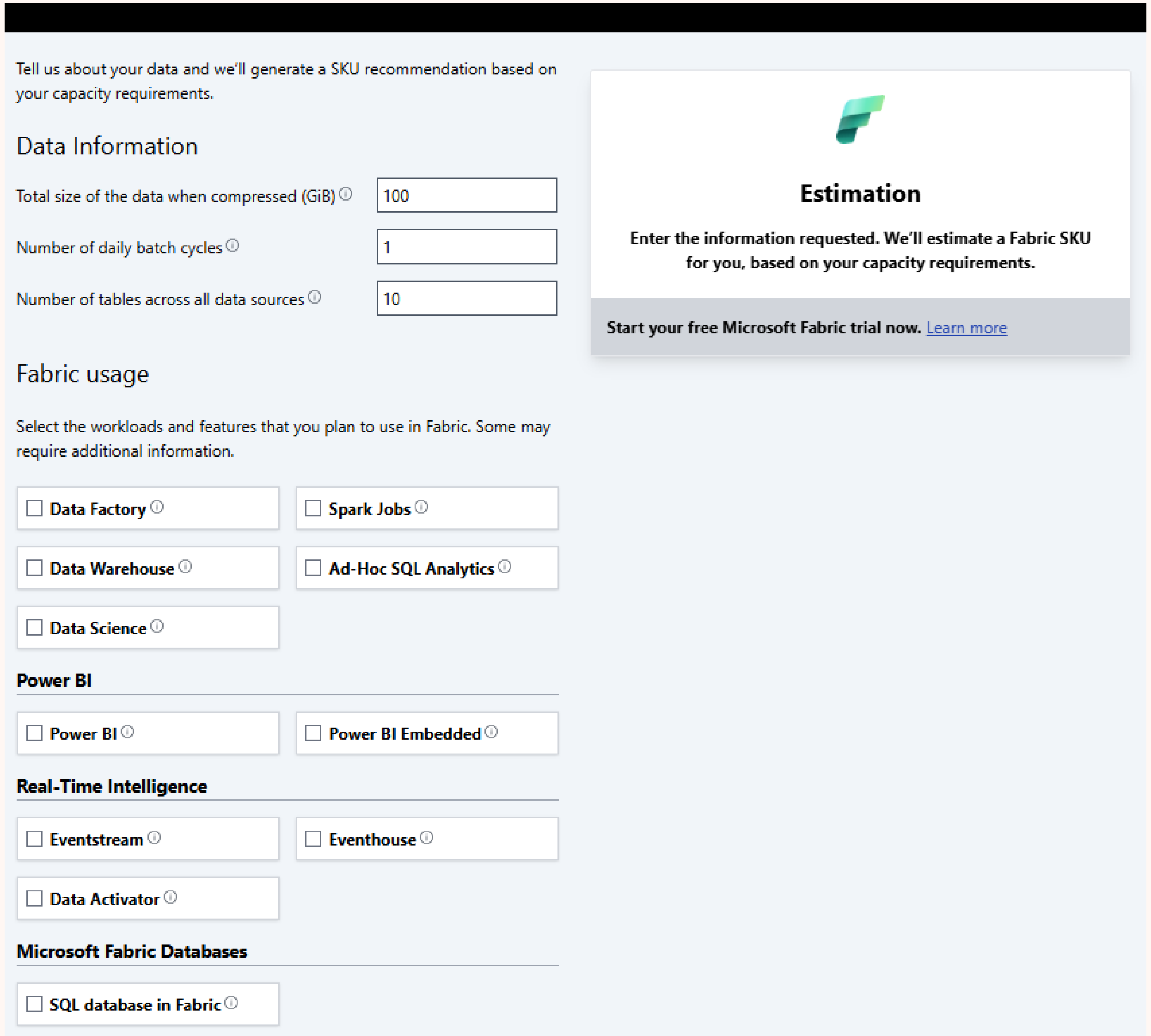
Enabling customers to better estimate their SKU before purchase

New online calculator

- Provides capacity estimations customized to your unique requirements
- Help businesses optimize their data infrastructure plan

Impact

- It’s easier to estimate a recommended SKU when starting on Fabric
- However, customers still should test their solutions to ensure they’re correctly sized.



Tell us about your data and we’ll generate a SKU recommendation based on your capacity requirements.

Data Information

Total size of the data when compressed (GiB) ⓘ	100
Number of daily batch cycles ⓘ	1
Number of tables across all data sources ⓘ	10

Fabric usage

Select the workloads and features that you plan to use in Fabric. Some may require additional information.

□ Data Factory ⓘ
□ Spark Jobs ⓘ
□ Data Warehouse ⓘ
□ Ad-Hoc SQL Analytics ⓘ
□ Data Science ⓘ

Power BI

□ Power BI ⓘ
□ Power BI Embedded ⓘ

Real-Time Intelligence

□ Eventstream ⓘ
□ Eventhouse ⓘ
□ Data Activator ⓘ

Microsoft Fabric Databases

□ SQL database in Fabric ⓘ

Estimation

Enter the information requested. We’ll estimate a Fabric SKU for you, based on your capacity requirements.

Start your free Microsoft Fabric trial now. Learn more

Fabric Capacity Reservations

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

Existing Azure Concept

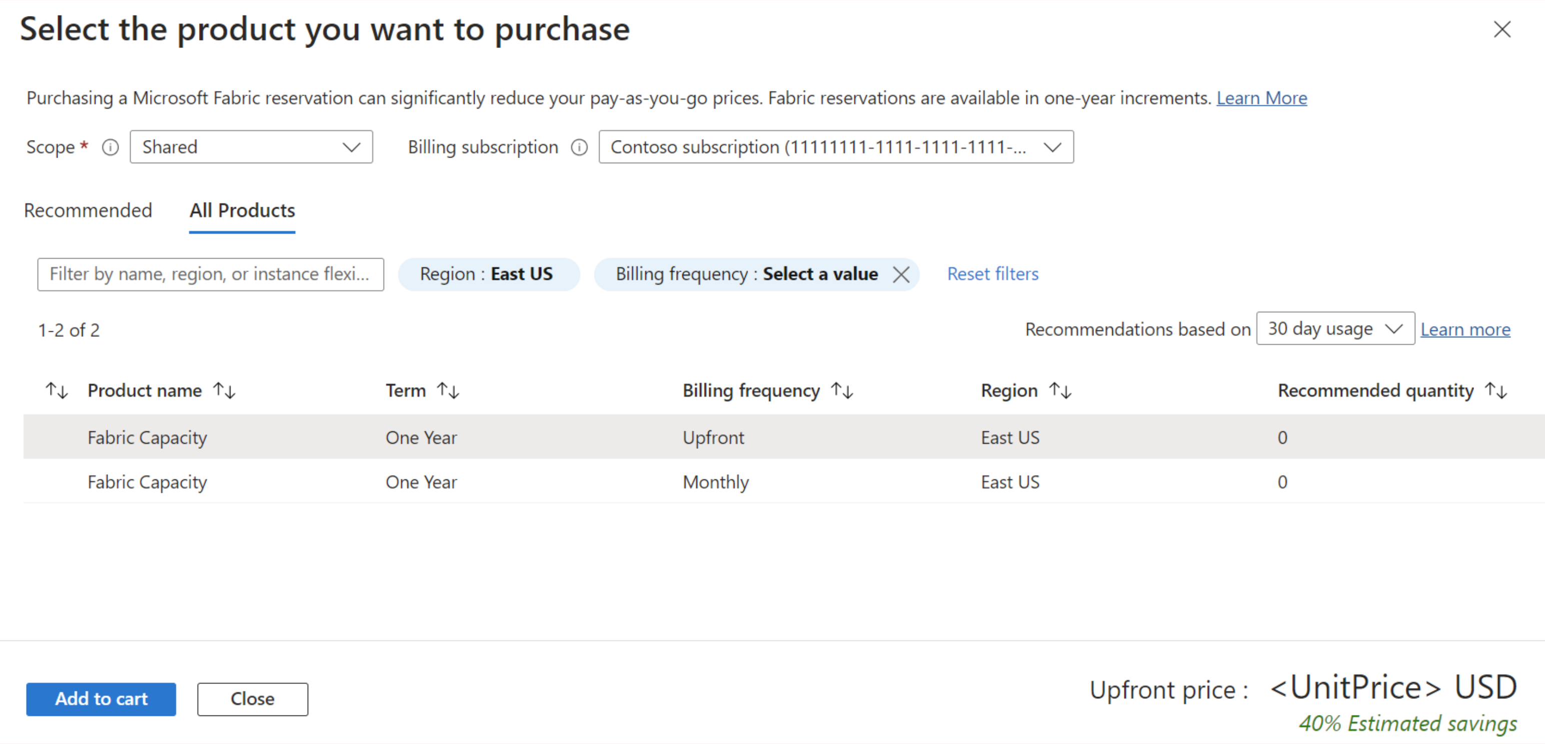
- Cancel, Refund, Exchange by policies

Reservations can be scoped by

- Billing Account
- Subscription(s)
- Resource Group(s)
- Region

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 PayG



Select the product you want to purchase

Purchasing a Microsoft Fabric reservation can significantly reduce your pay-as-you-go prices. Fabric reservations are available in one-year increments. Learn More

Scope * : Shared

Billing subscription : Contoso subscription (11111111-1111-1111-1111-...

Region : East US

Billing frequency : Select a value

1-2 of 2

Recommendations based on 30 day usage

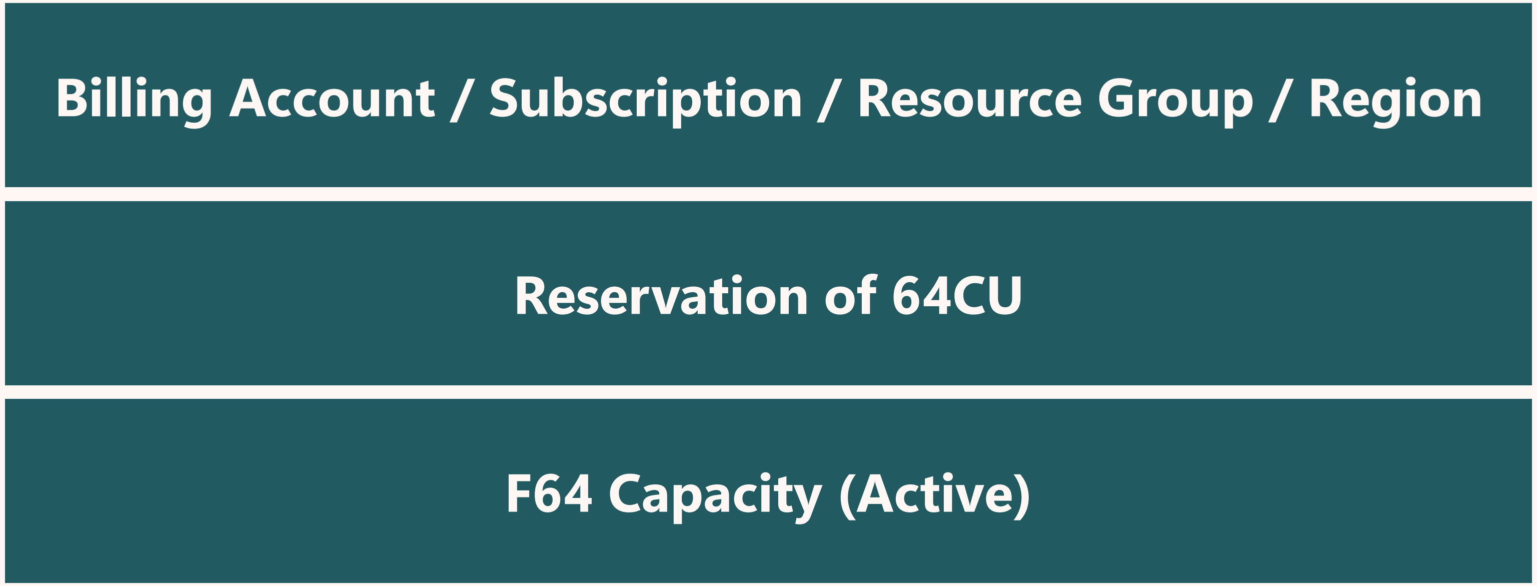
Product name	Term	Billing frequency	Region	Recommended quantity
Fabric Capacity	One Year	Upfront	East US	0
Fabric Capacity	One Year	Monthly	East US	0

Upfront price : <UnitPrice> USD

40% Estimated savings

Fabric Capacity Reservations

Example 1 – Single Reservation matching a single capacity



Billing Account / Subscription / Resource Group / Region

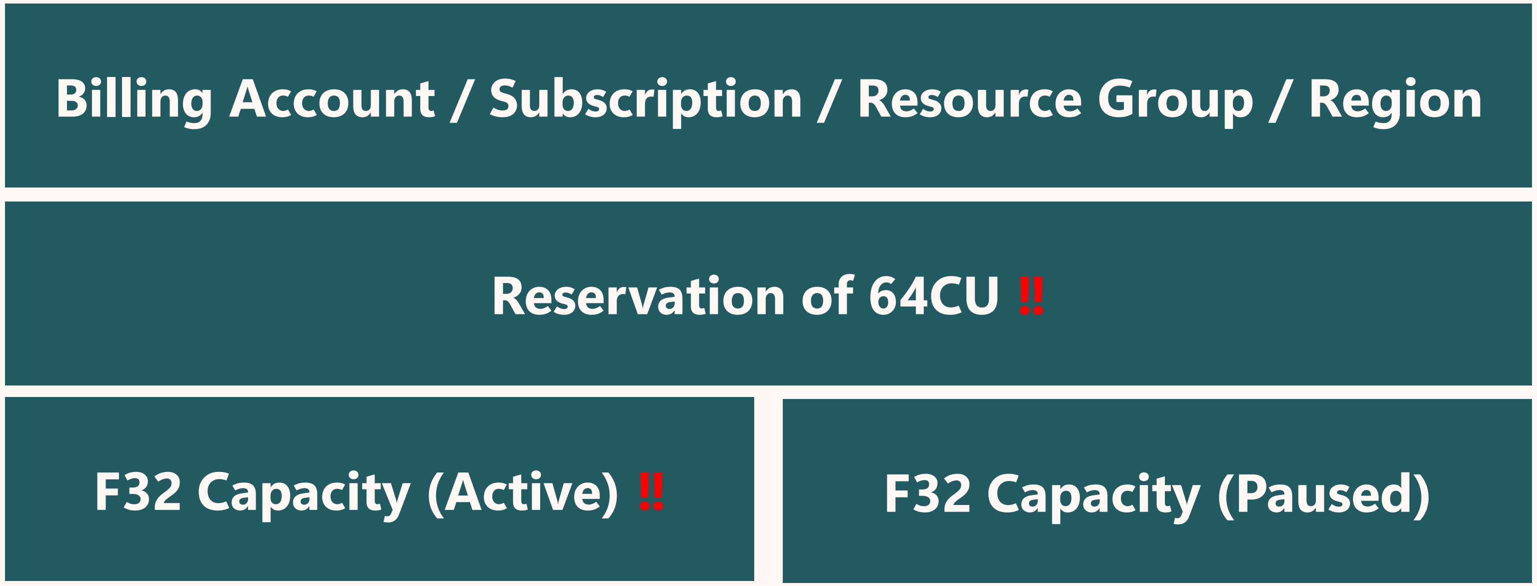
Reservation of 64CU

F64 Capacity (Active)

All active capacities are covered by Reservation, discount applies

Fabric Capacity Reservations

Example 2 – Single Reservation exceeds Active Capacities



Billing Account / Subscription / Resource Group / Region

Reservation of 64CU ‼

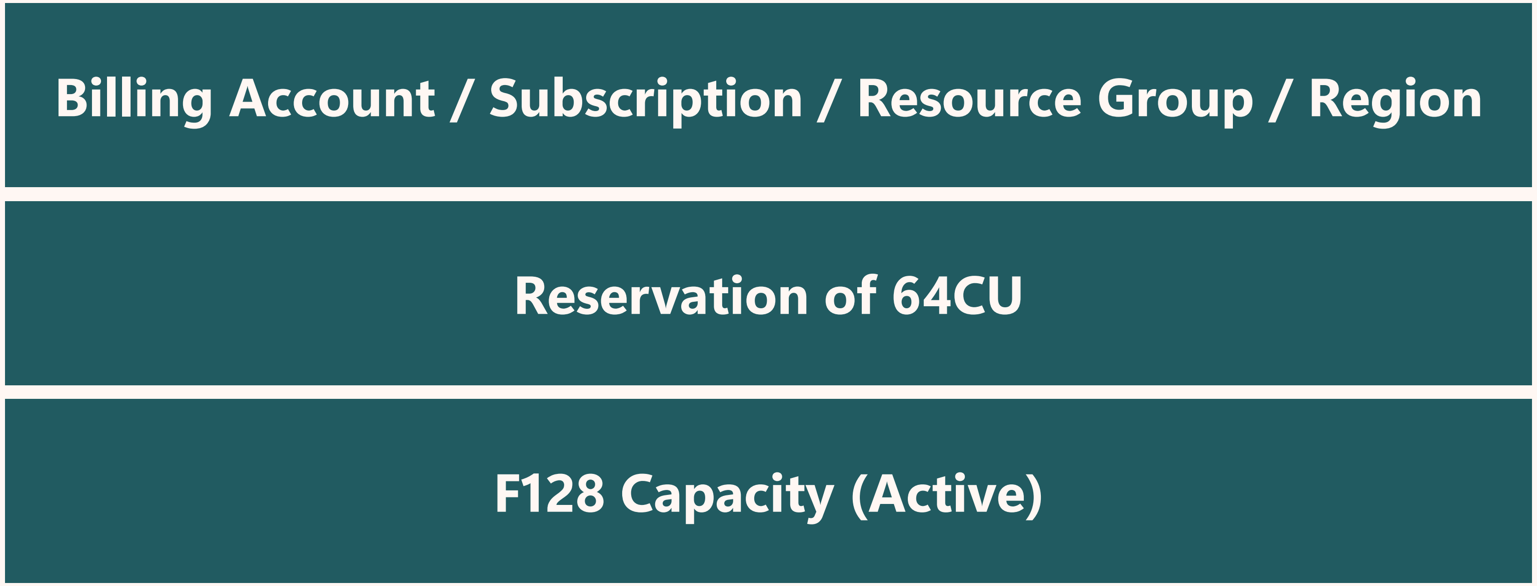
F32 Capacity (Active) ‼

F32 Capacity (Paused)

Reservation only applies to active capacity, 32CU are “wasted”

Fabric Capacity Reservations

Example 3 – Single SKU exceeding the reservation amount



Billing Account / Subscription / Resource Group / Region

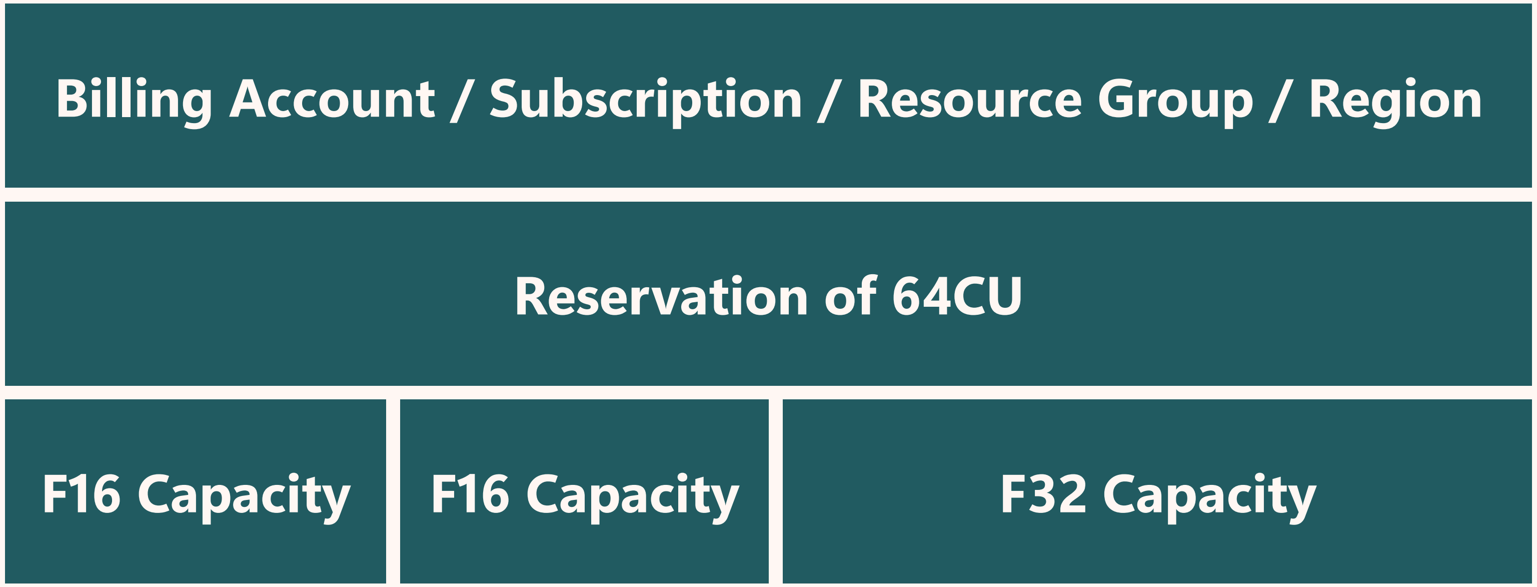
Reservation of 64CU

F128 Capacity (Active)

Mixed billing for F128 Capacity, 64CU reserved, 64CU PaYGo

Fabric Capacity Reservations

Example 4 – Single Reservation matches active capacities



Billing Account / Subscription / Resource Group / Region

Reservation of 64CU

F16 Capacity

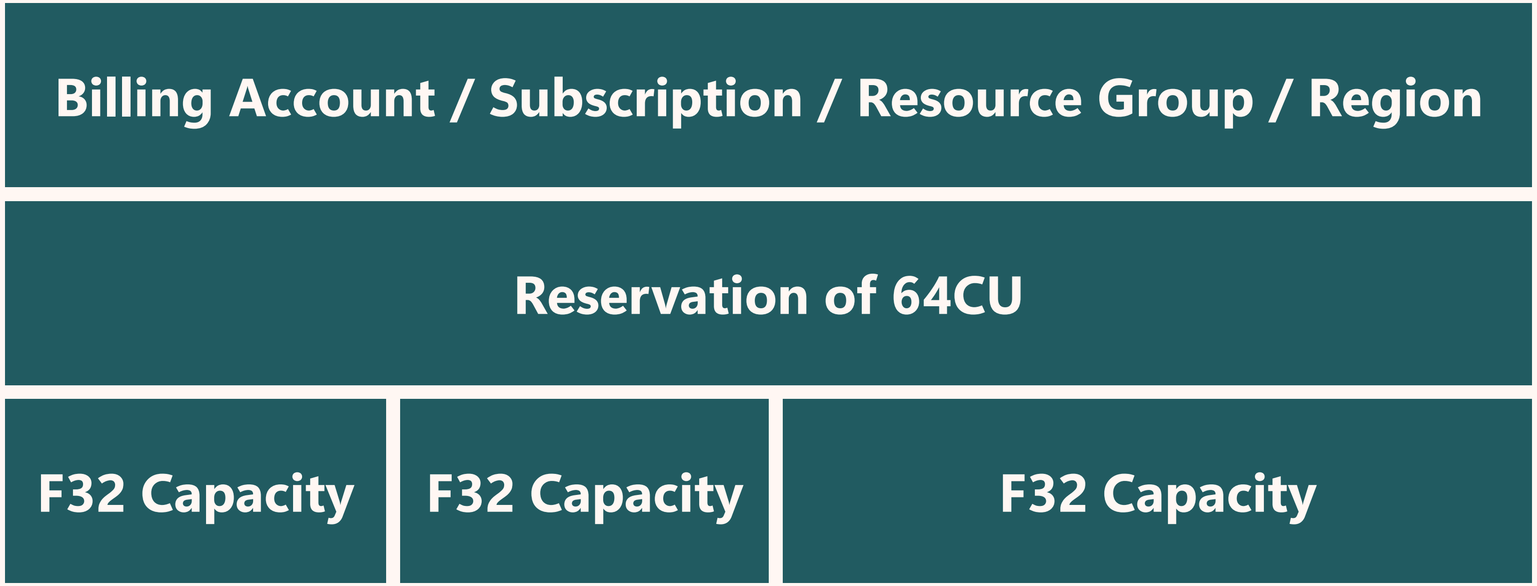
F16 Capacity

F32 Capacity

All active capacities are covered by Reservation, discount applies

Fabric Capacity Reservations

Example 5 – Multiple SKUs exceeding reservation amount



Billing Account / Subscription / Resource Group / Region

Reservation of 64CU

F32 Capacity

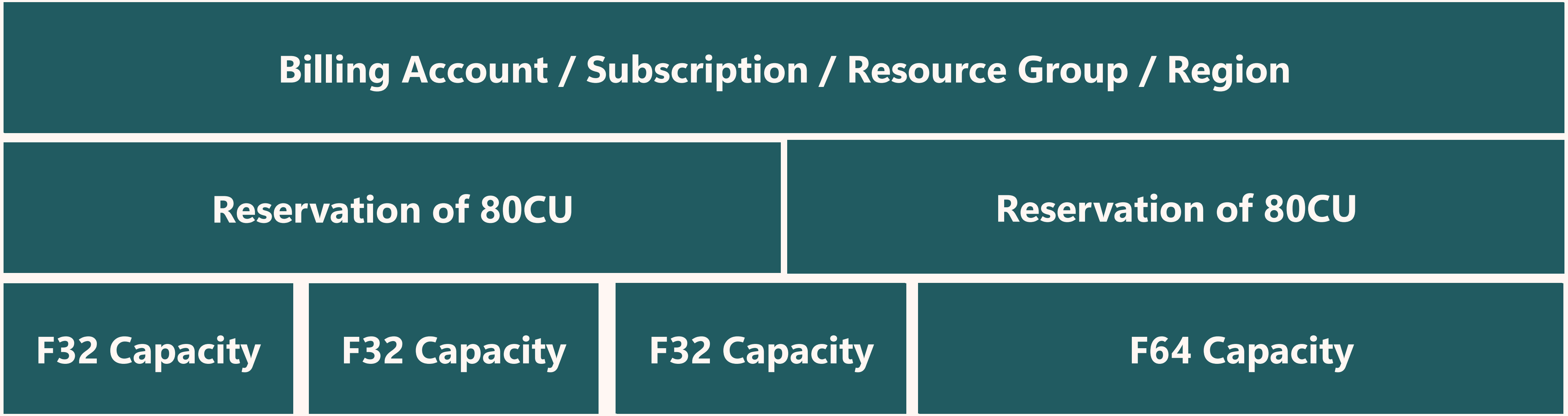
F32 Capacity

F32 Capacity

Billing can apply to all of the F32 capacities, is only shown in Billing

Fabric Capacity Reservations

Example 6 – Multiple SKUs matching reservation amount through multiple reservations



Billing Account / Subscription / Resource Group / Region

Reservation of 80CU

Reservation of 80CU

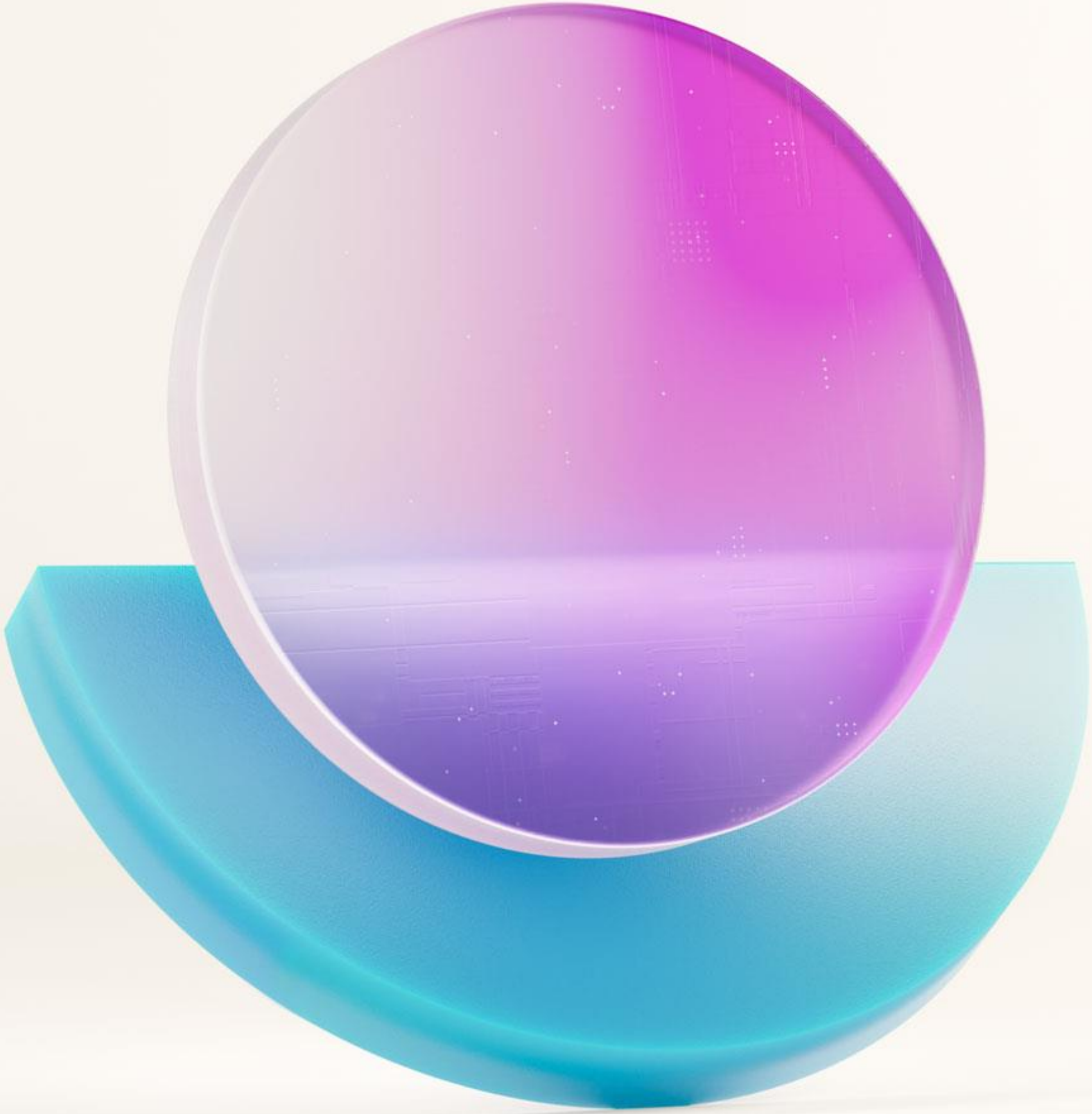
F32 Capacity

F32 Capacity

F32 Capacity

F64 Capacity

Only when second Reservation is added, all Capacities are covered



Dealing with Throttling

Microsoft Fabric
COMMUNITY CONFERENCE

Pausing and Resuming Capacities

Pausing and Resuming Capacities

Why pause capacities?

1) It **can** 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

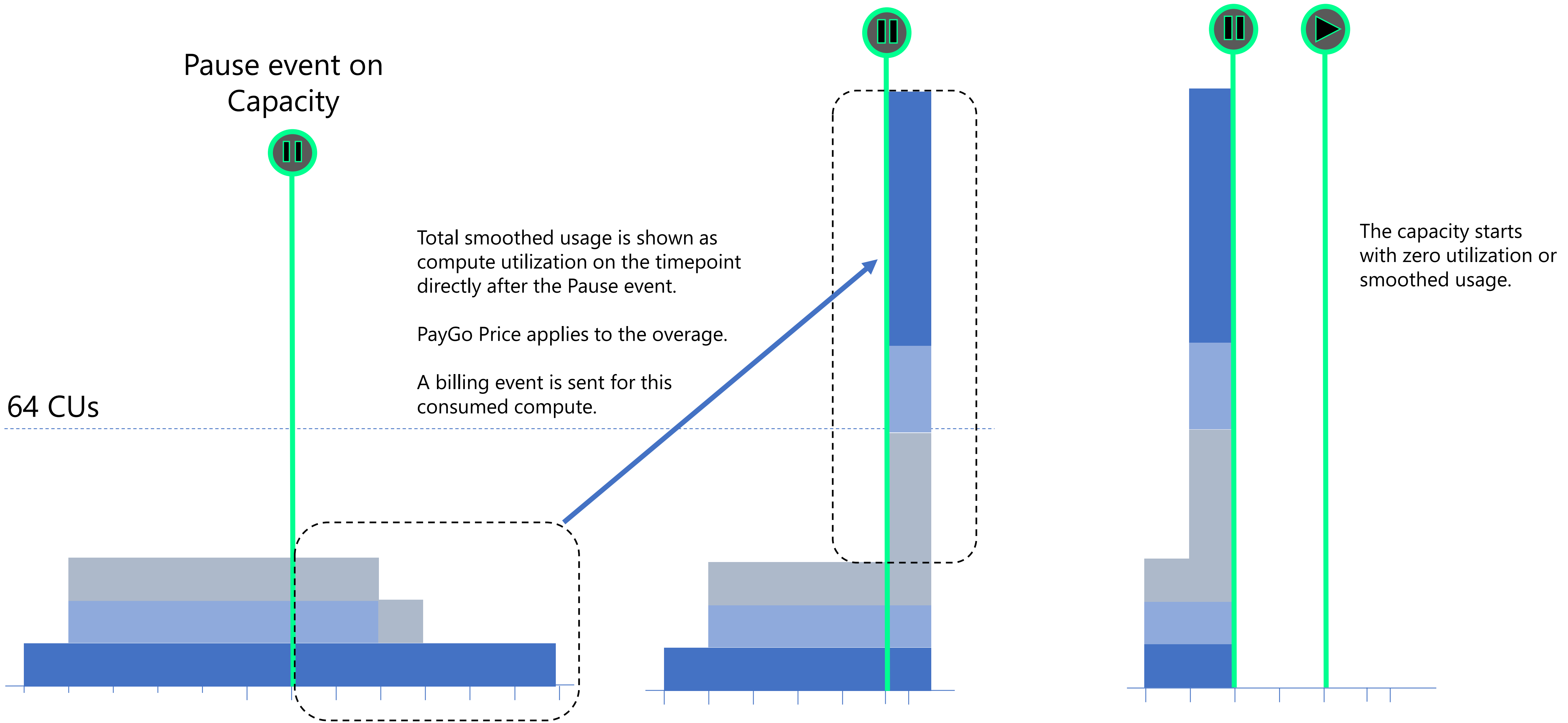
Note: OneLake storage costs continue to be billed while a capacity is paused

How Capacity Pause & Resume works

When a capacity is **paused**...

Smoothed usage is **reconciled**.

Later, it can be **resumed**.



Pause event on Capacity

64 CUs

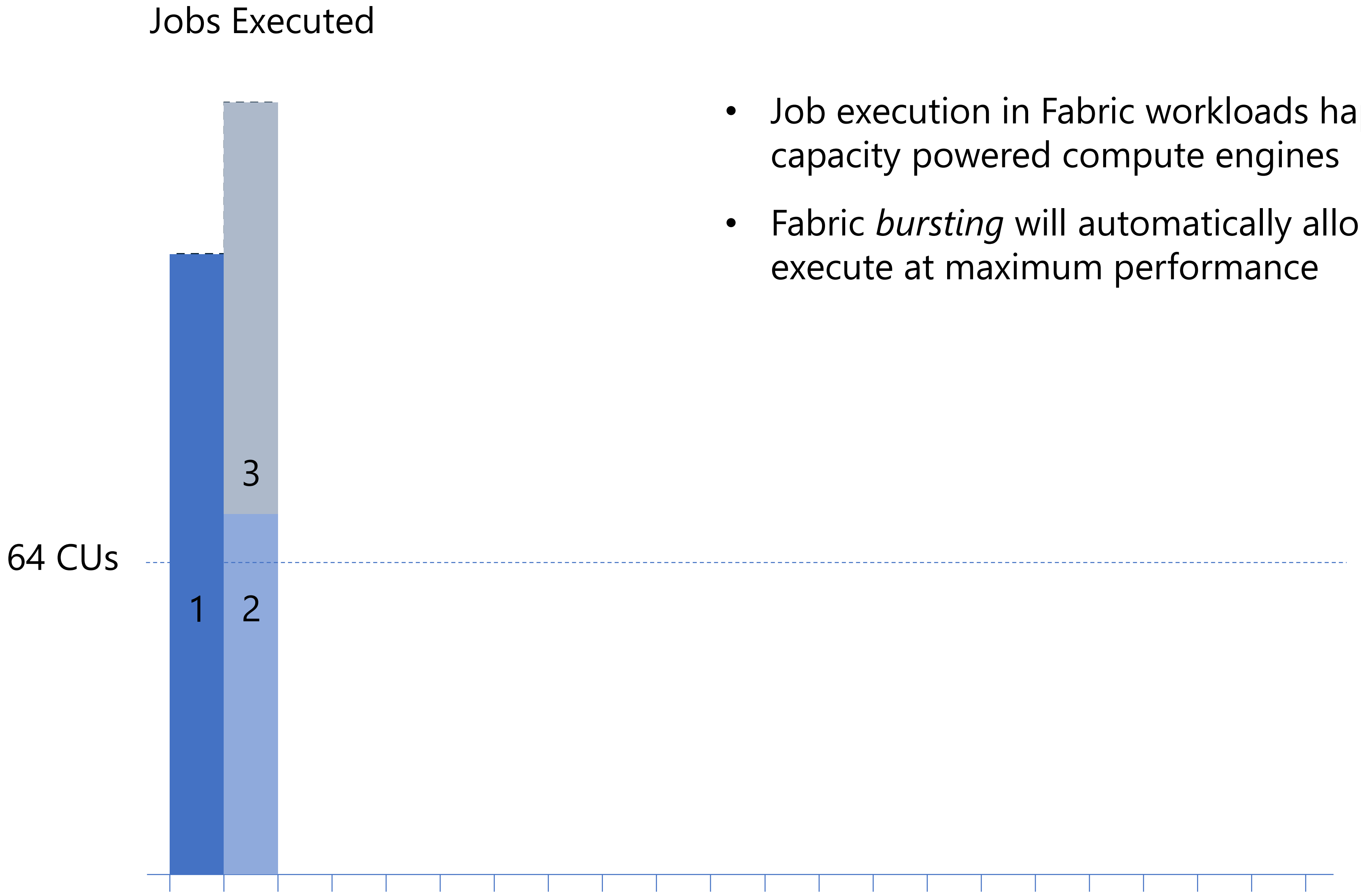
Total smoothed usage is shown as compute utilization on the timepoint directly after the Pause event.

PayGo Price applies to the overage.

A billing event is sent for this consumed compute.

The capacity starts with zero utilization or smoothed usage.

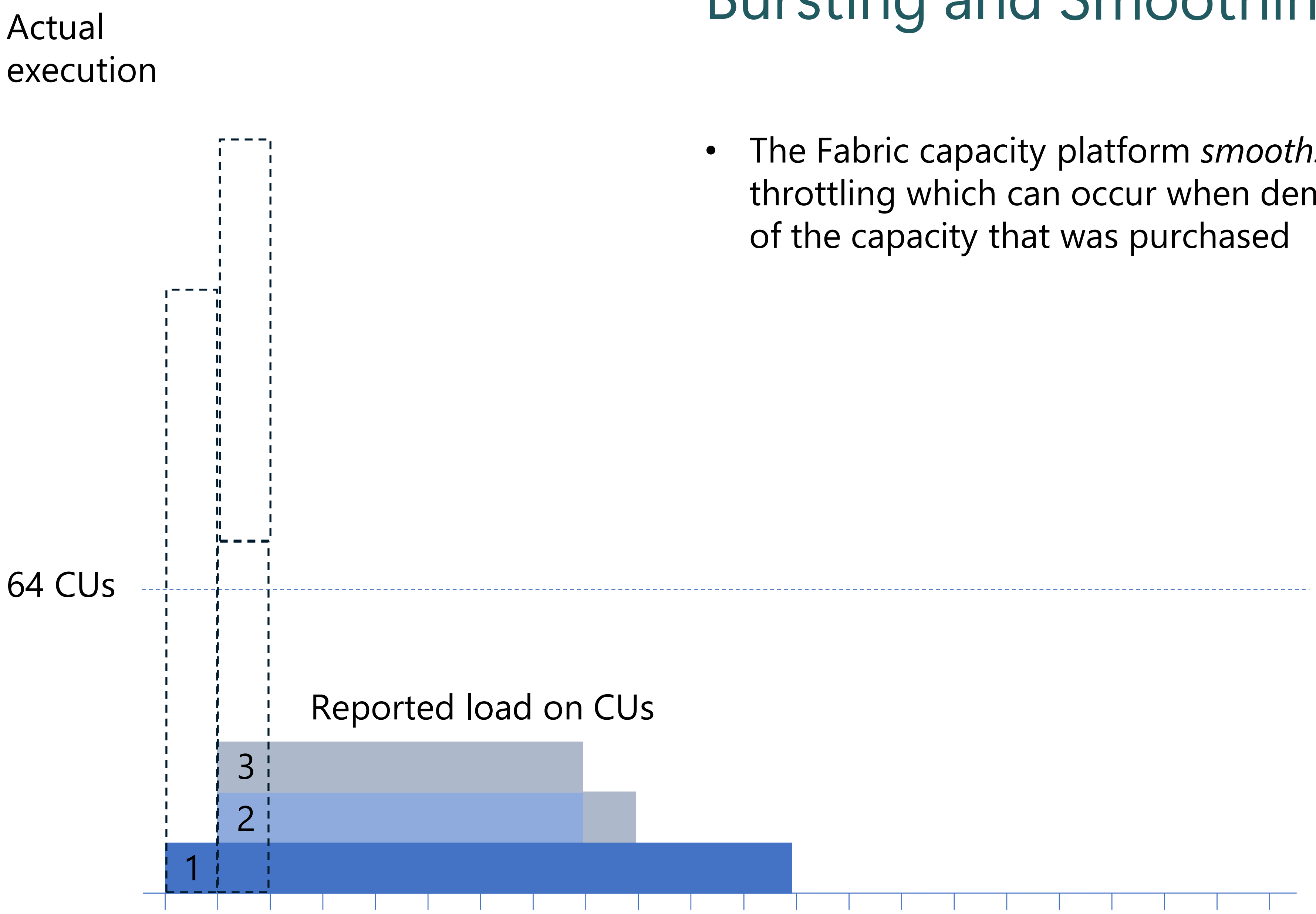
Bursting and Smoothing



- 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

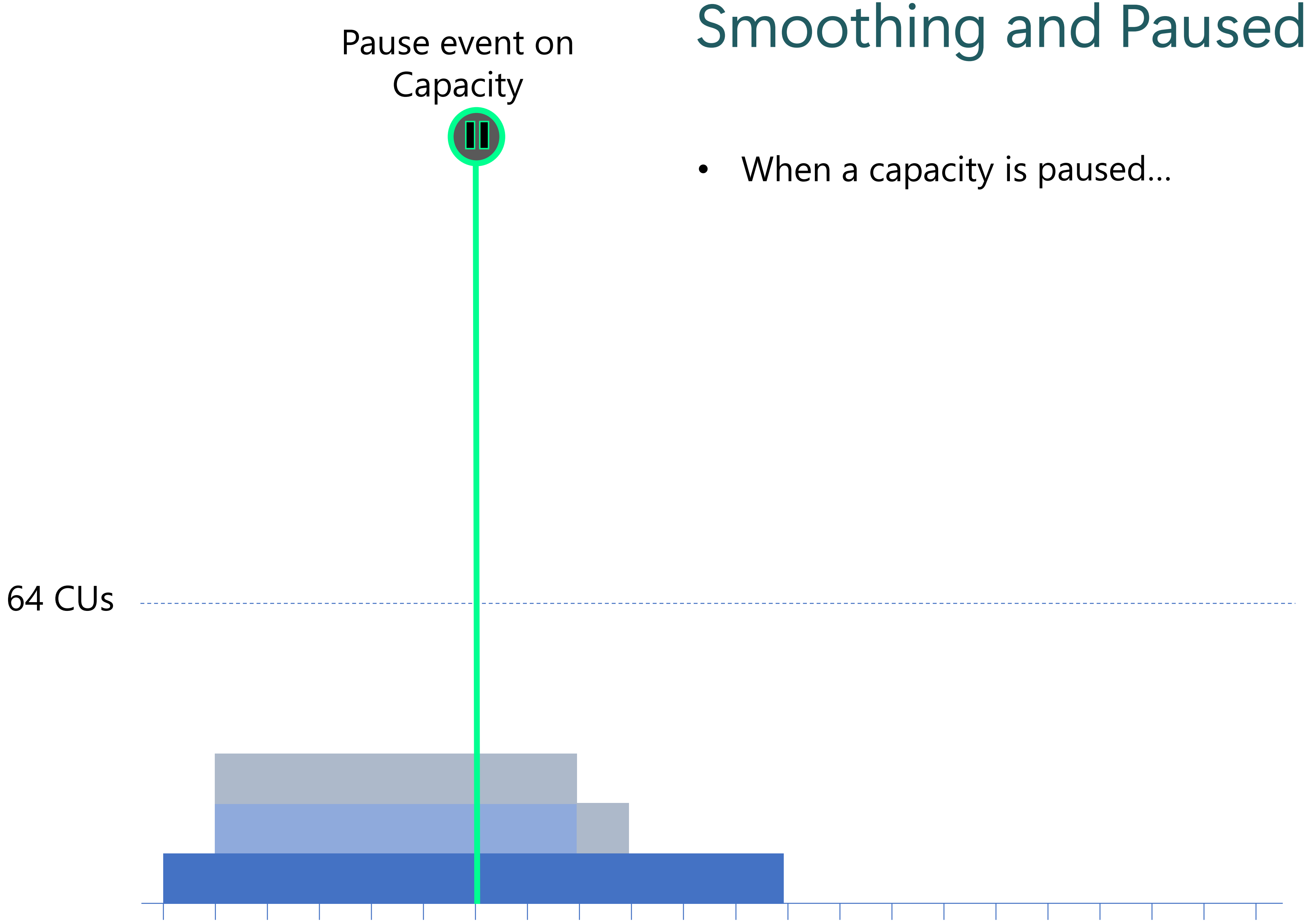
Bursting and Smoothing

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



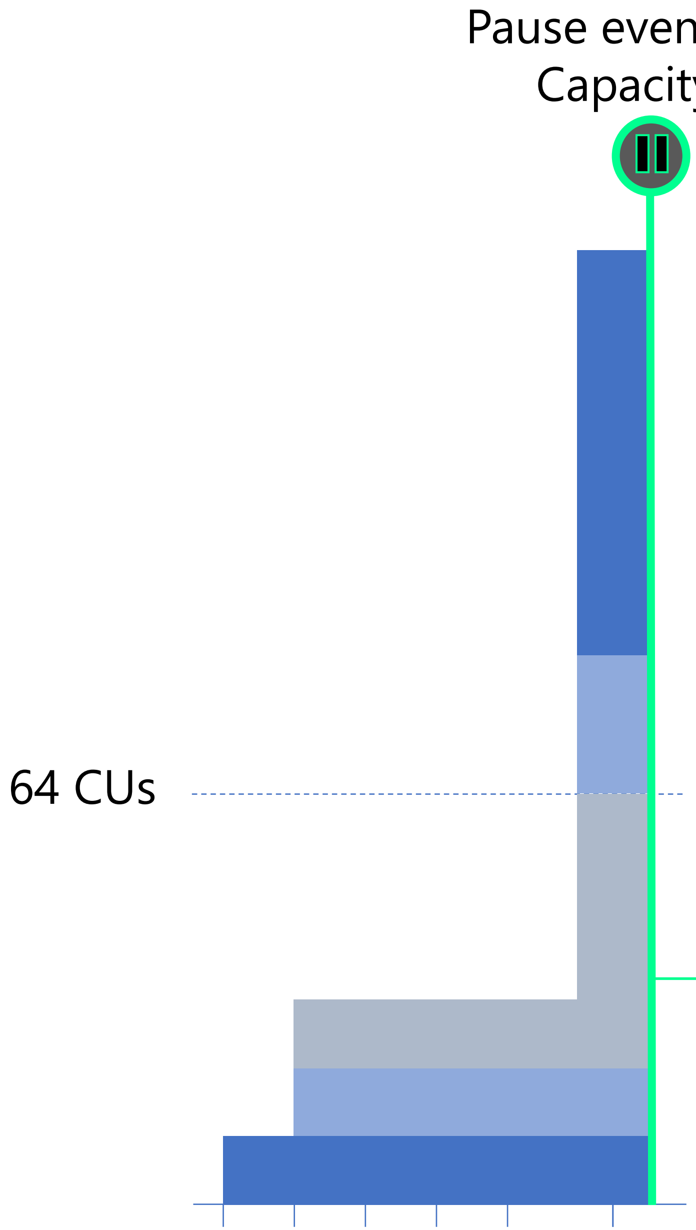
Smoothing and Paused Capacities

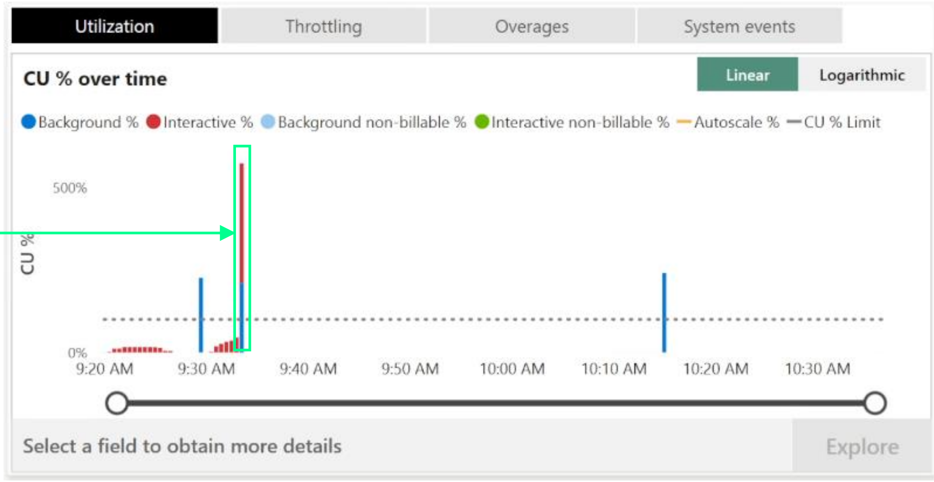
- When a capacity is paused…



Smoothing and Paused Capacities

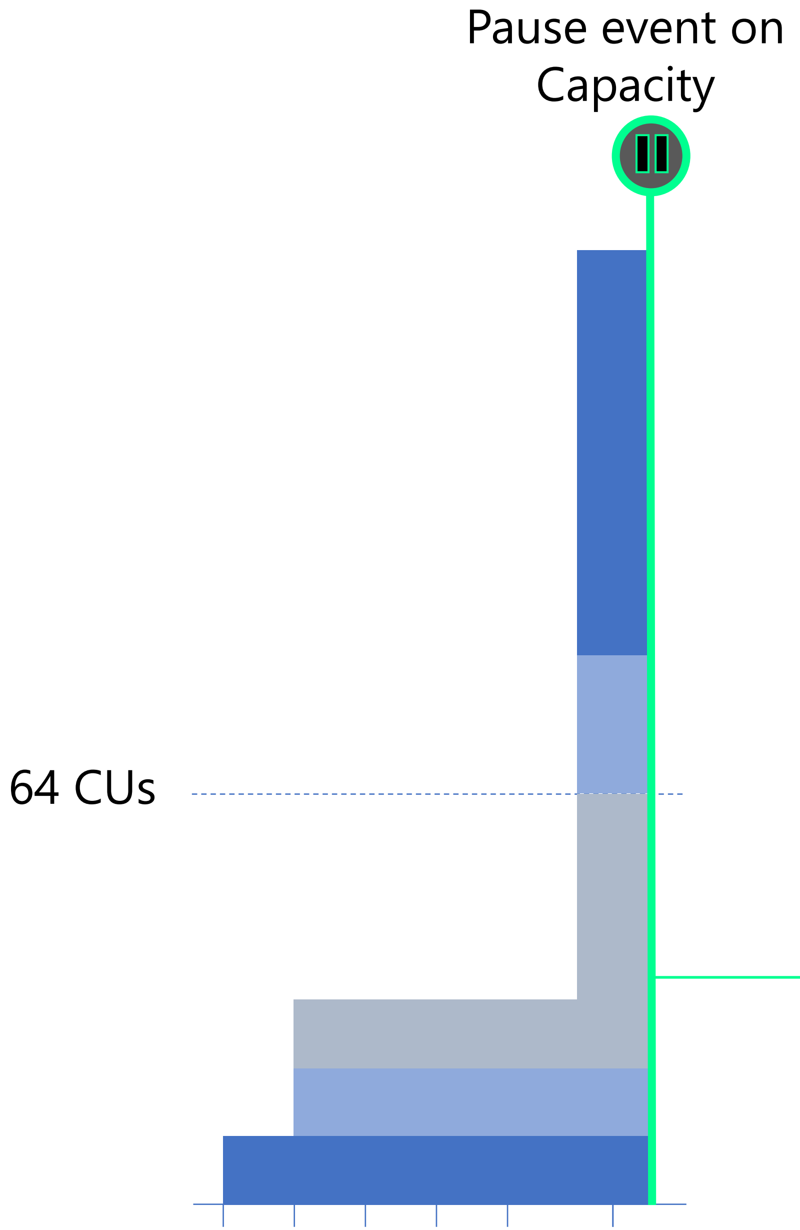
- When a capacity is paused…
- Usage that was smoothed into the future will be “reconciled” and charged against the capacity at the timestamp the capacity was paused
- Reconciled usage will show up as a spike in capacity metrics



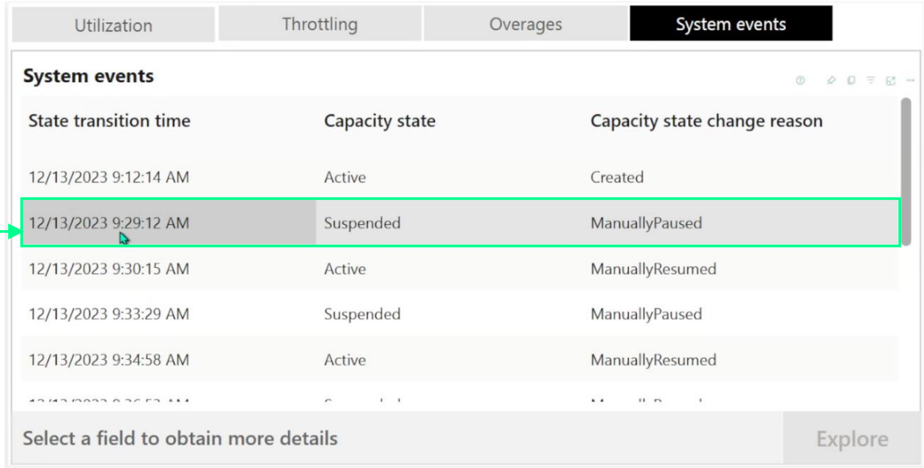


Select a field to obtain more details

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



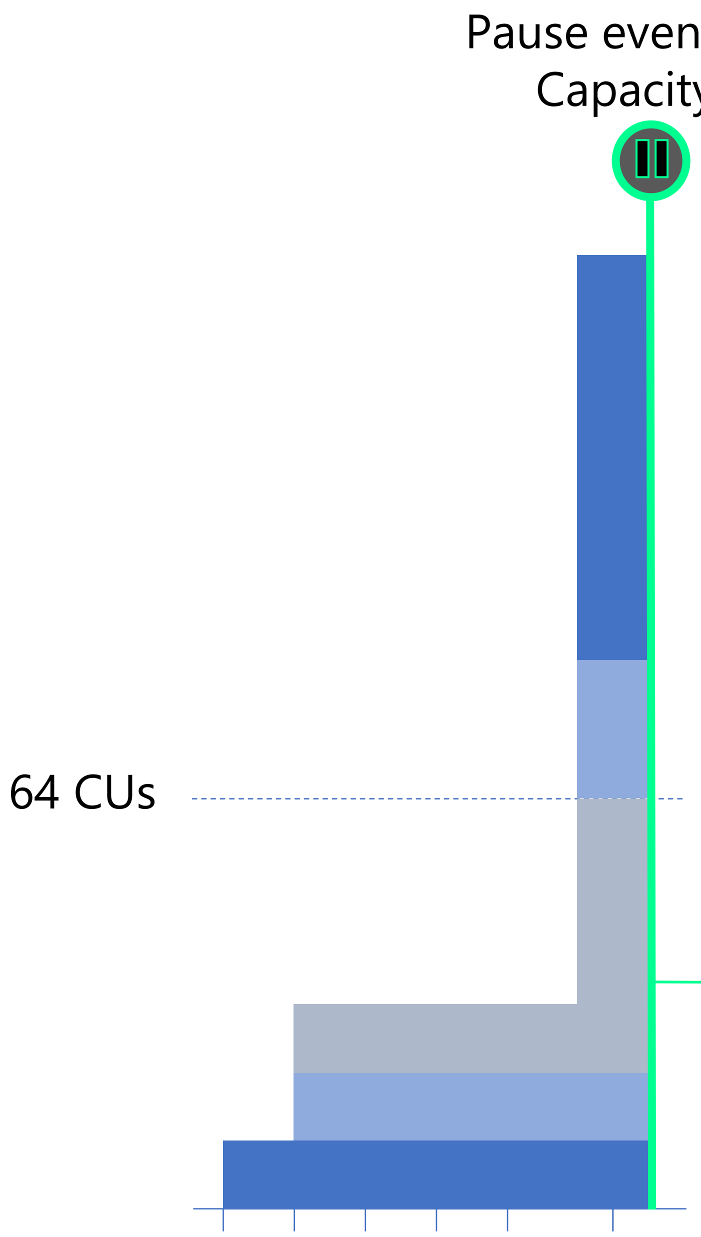
System events

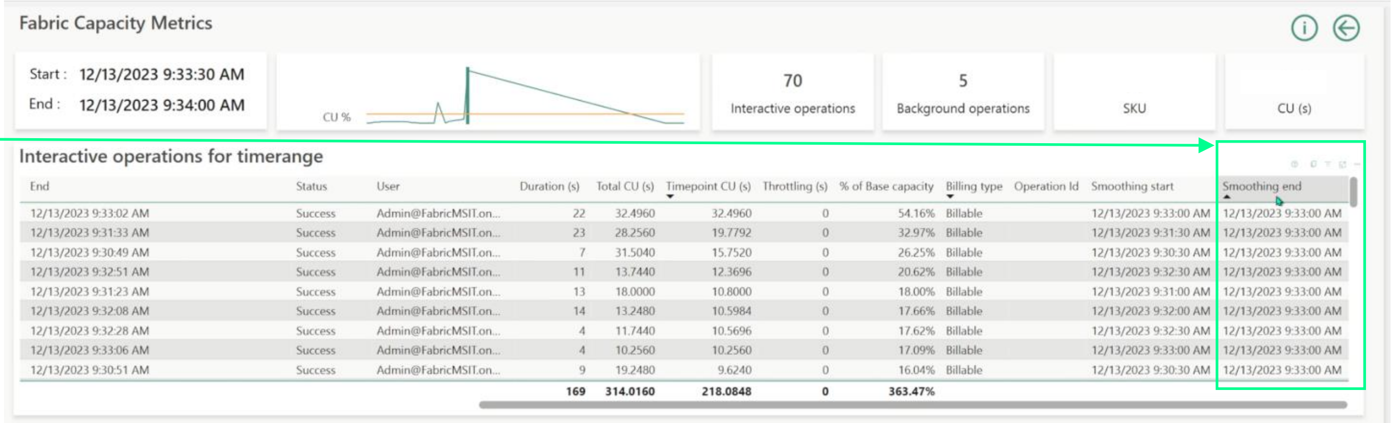
State transition time	Capacity state	Capacity state change reason
12/13/2023 9:12:14 AM	Active	Created
12/13/2023 9:29:12 AM	Suspended	ManuallyPaused
12/13/2023 9:30:15 AM	Active	ManuallyResumed
12/13/2023 9:33:29 AM	Suspended	ManuallyPaused
12/13/2023 9:34:58 AM	Active	ManuallyResumed

Select a field to obtain more details

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




Fabric Capacity Metrics

Start : 12/13/2023 9:33:30 AM
End : 12/13/2023 9:34:00 AM

70 Interactive operations

5 Background operations

SKU

CU (s)

Interactive operations for timerange

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

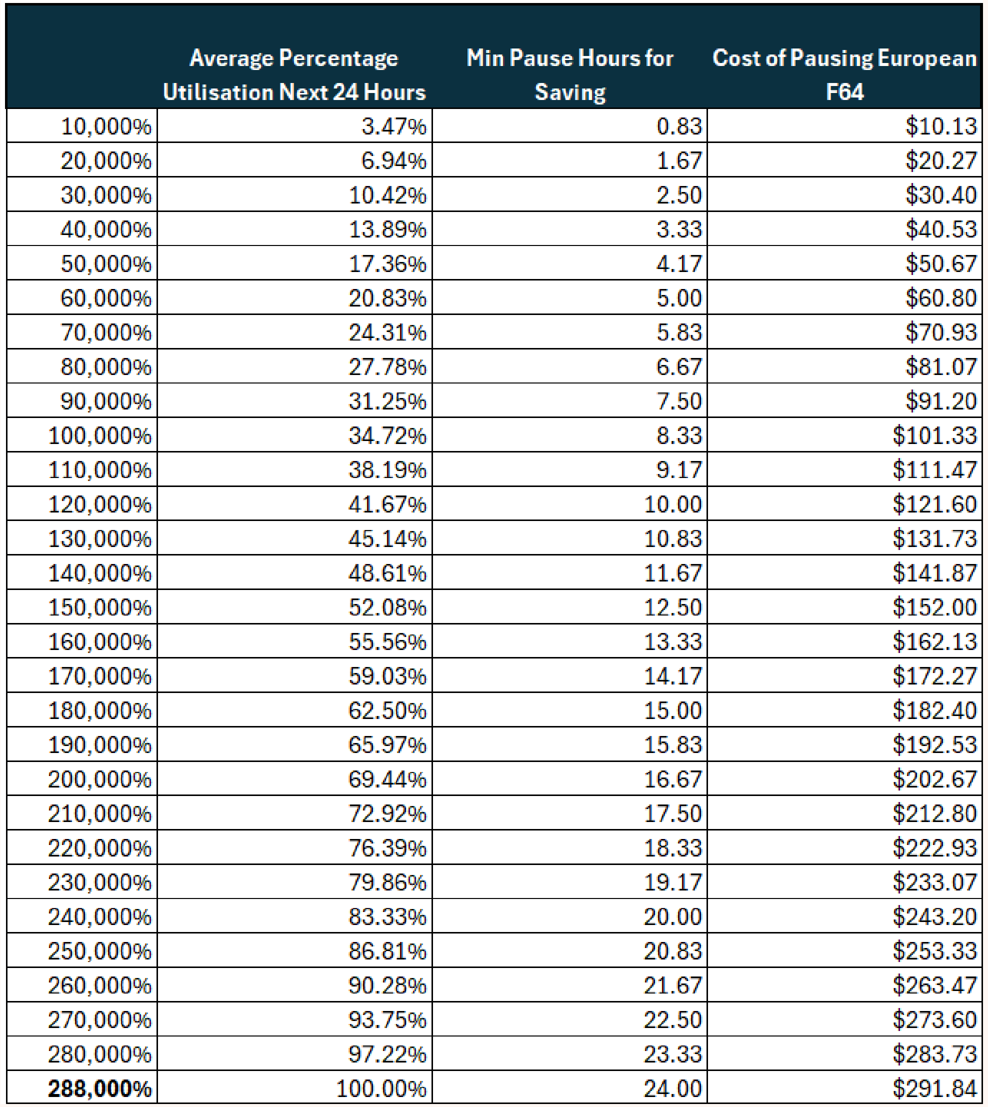
It might actually cost you more ..

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

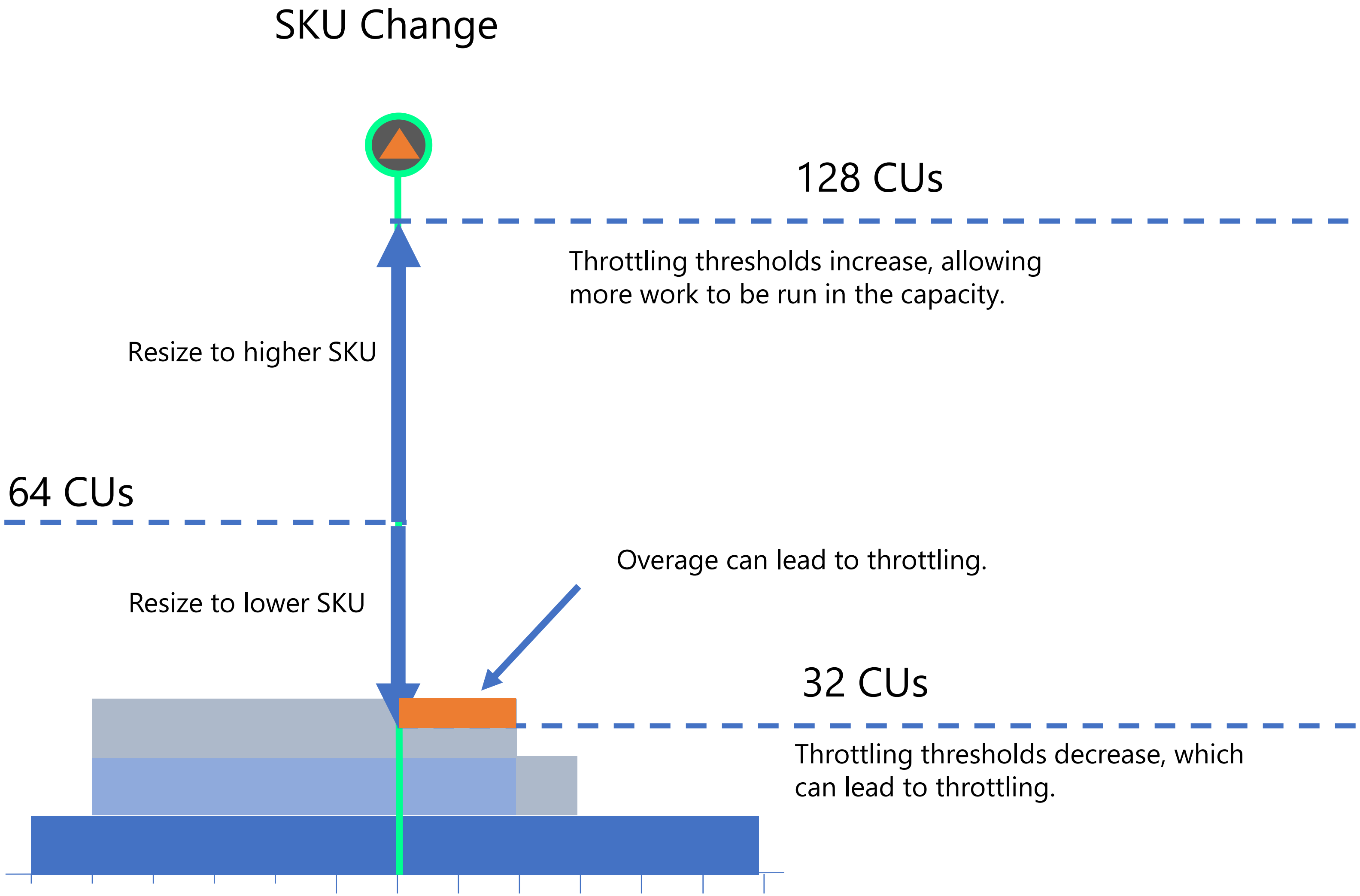
https://www.linkedin.com/pulse/fabric-billing-part-4-implications-pause-restart-matthew-farrow-fznse/

How Capacity Resize works

When a capacity is **resized**…

The allowed CUs per timepoint increase or decrease.

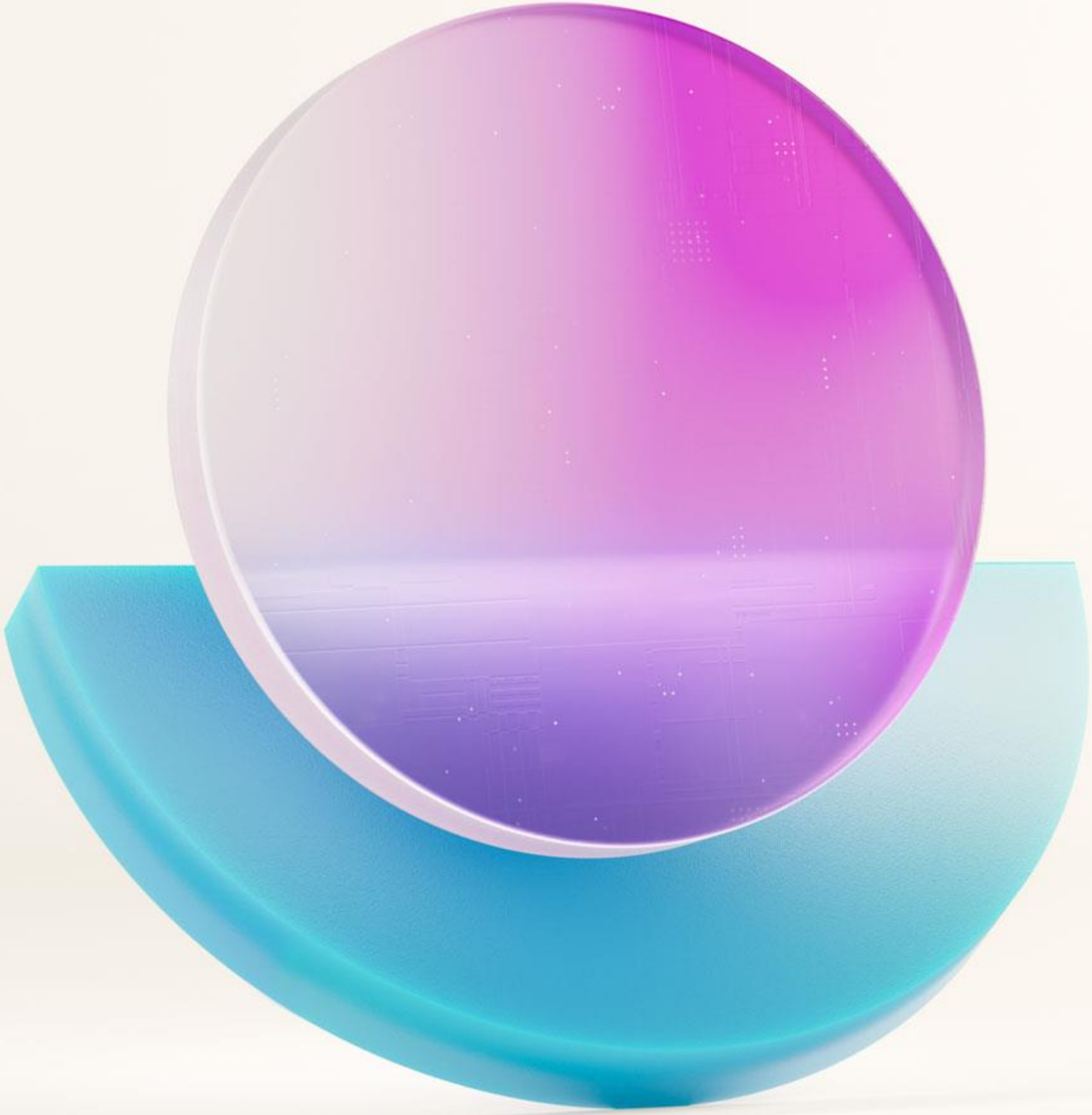
This changes the throttling allowed limits based on the new SKU’s CUs and the throttling windows.


SKU Change
128 CUs
Throttling thresholds increase, allowing more work to be run in the capacity.
Resize to higher SKU
64 CUs
Resize to lower SKU
Overage can lead to throttling.
32 CUs
Throttling thresholds decrease, which can lead to throttling.

Key Insights

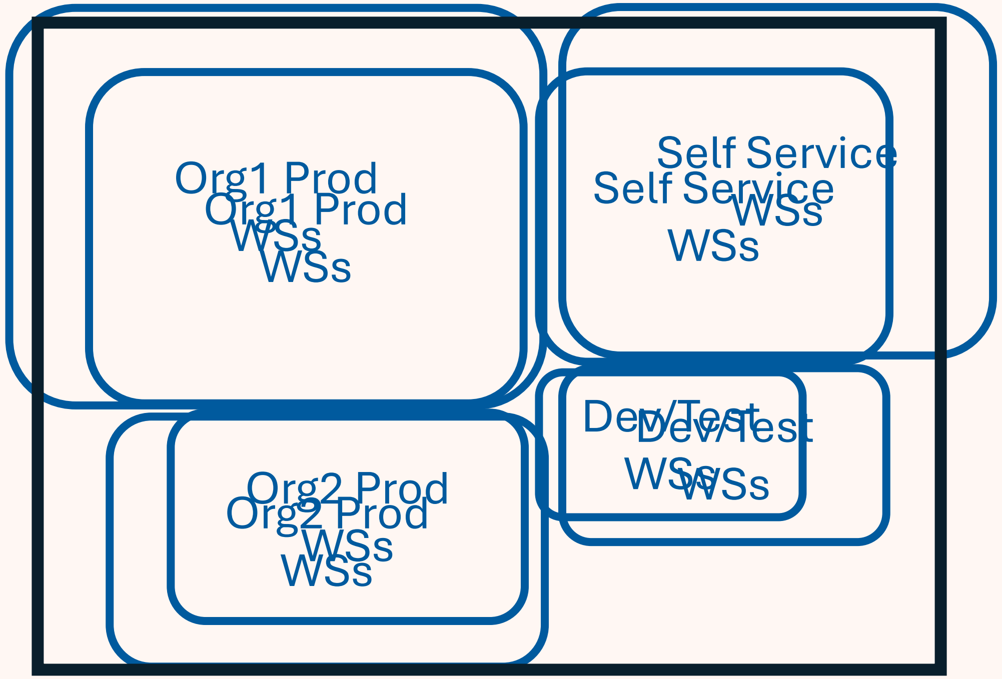
- Sizing up will incur the cost of the new SKU
- Sizing down could lead to more throttling
- Review your Throttling Thresholds before sizing down your SKU.

Microsoft Fabric
COMMUNITY CONFERENCE



Workspace planning

When Capacity Units Run Out
Option 1 – Optimize



WSs = Workspaces

Capacity

Approach

- Work with content creators to follow best practices and reduce CU consumption

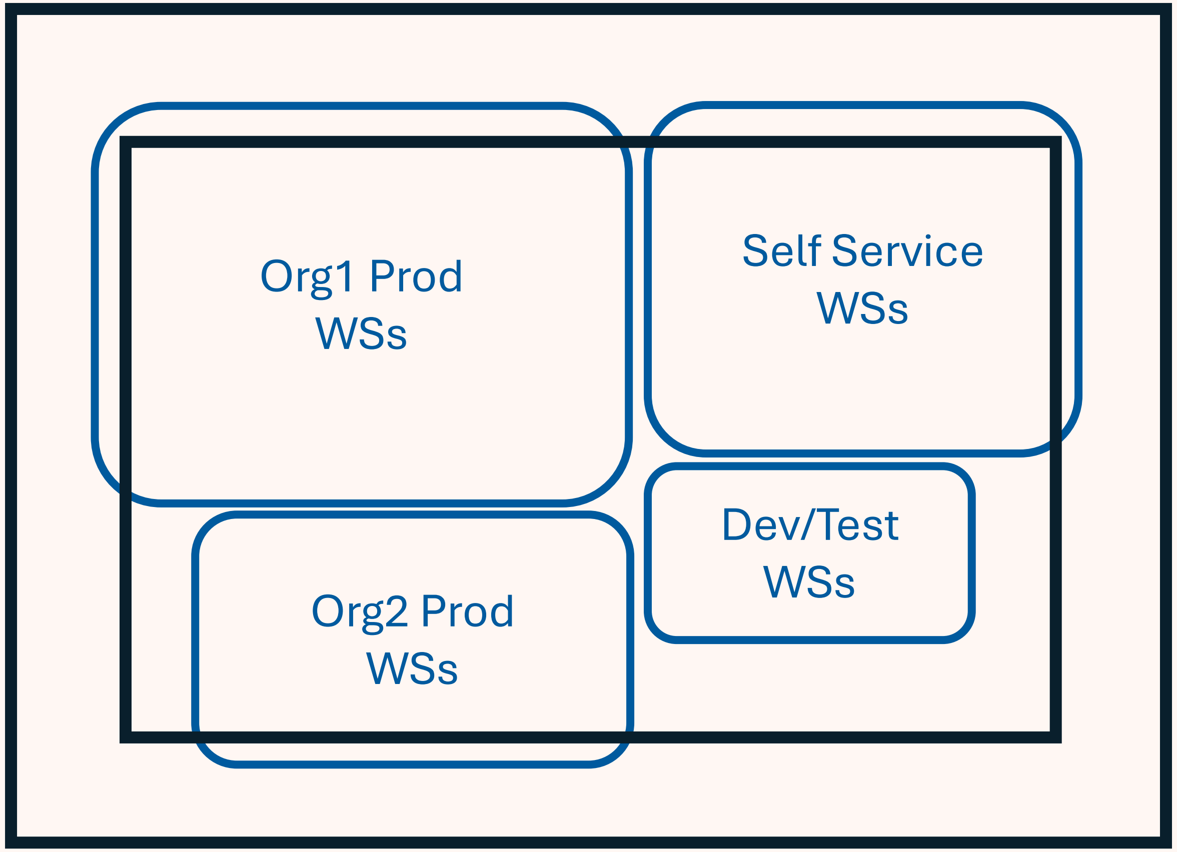
Pros

- Avoids increased cost
- Learning carries over to future content

Cons

- Can be difficult/time consuming

When Capacity Units Run Out
Option 2 – Scale Up



WSs = Workspaces Capacity

Options to add compute

- Move to a bigger P SKU or RI F SKU
- Turn on autoscale (P SKU)
- Manual/Dynamic change size (F SKU)

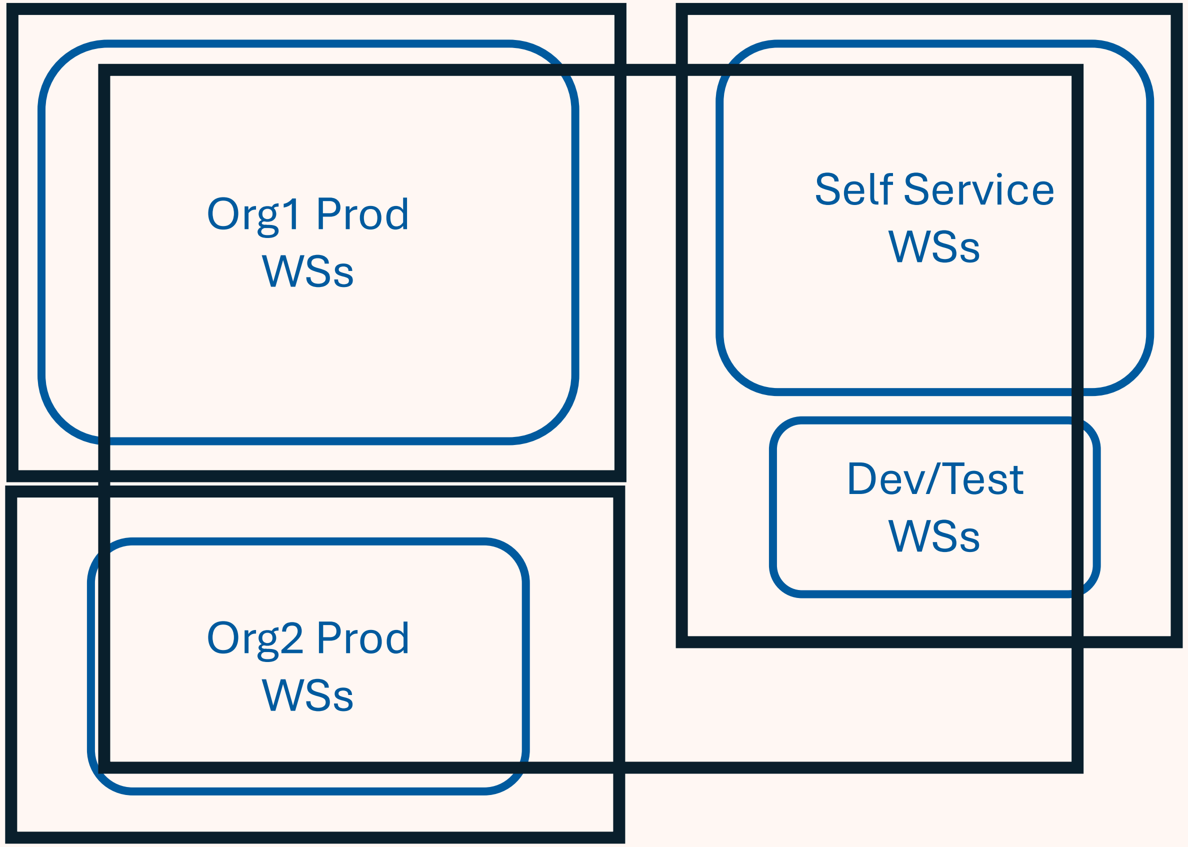
Pros

- Add CUs for all items
- Easy

Cons

- Cost
- Bad actors (items with unintentionally high CU burn) can still be a problem

When Capacity Units Run Out
Option 3 – Scale Out



WSs = Workspaces

Capacity

Options

- Create multiple smaller P or F SKUs based on organization, type of work, etc.

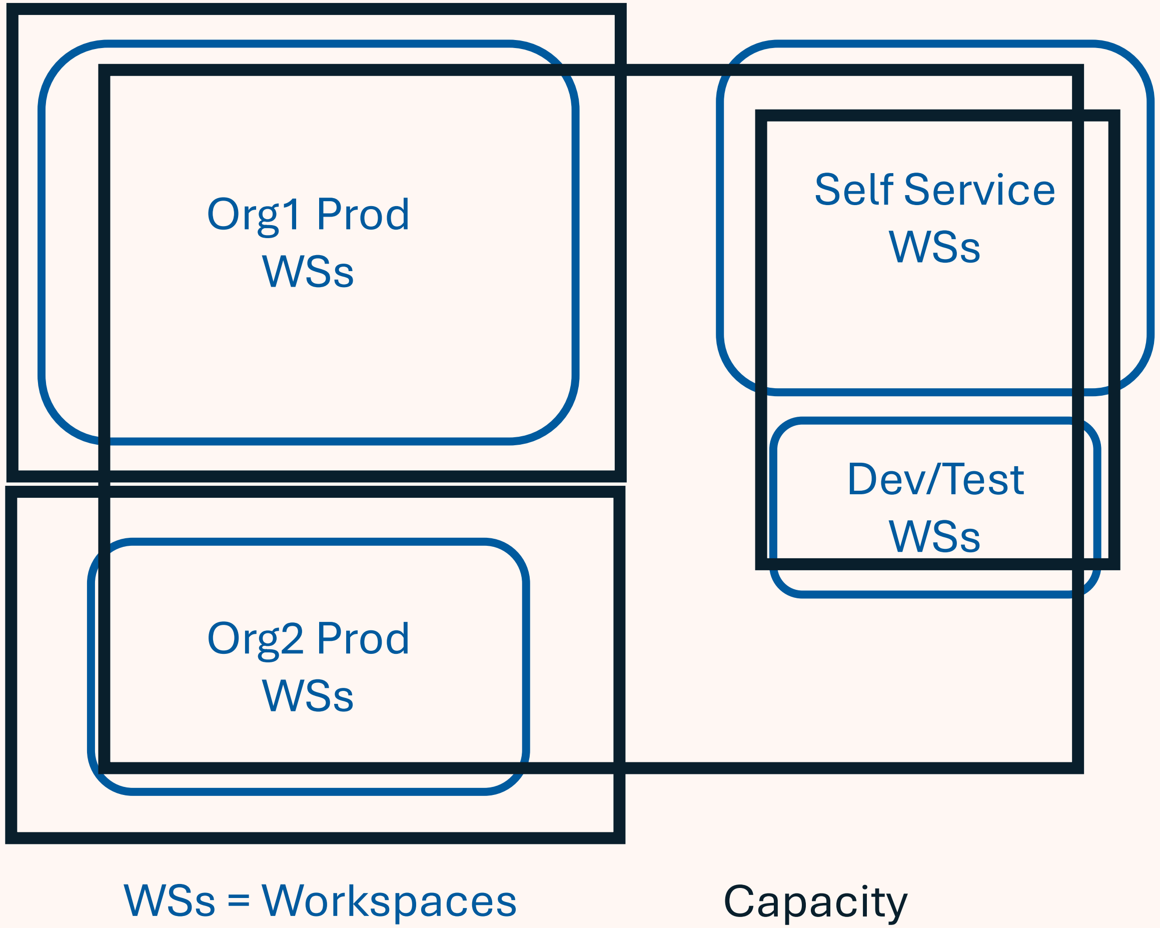
Pros

- Easy
- Provides some isolation from bad actors (items with unintentionally high CU burn)
- Flexibility in capacity settings/governance

Cons

- Cost
- High CU items have increased chance of throttling

When Capacity Units Run Out
Option 4 – Isolate

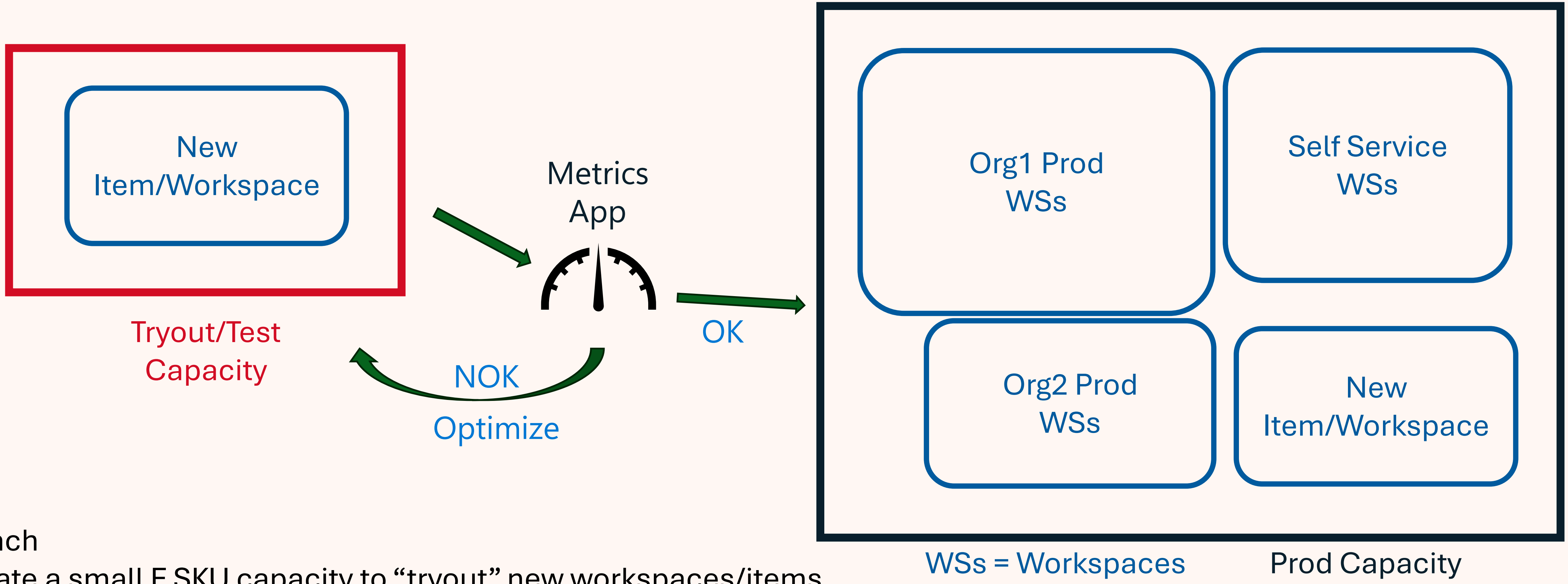


Approach
- Provide isolated capacity for key items built by experienced developers

Pros
- Easy
- Provides isolation from items built by inexperienced developers and/or rapid unplanned usage growth
- Flexibility in capacity settings/governance

Cons
- Cost
- May lead to frustration of lower priority content developers/consumers

Isolation Strategy #4a – Tryout Capacity



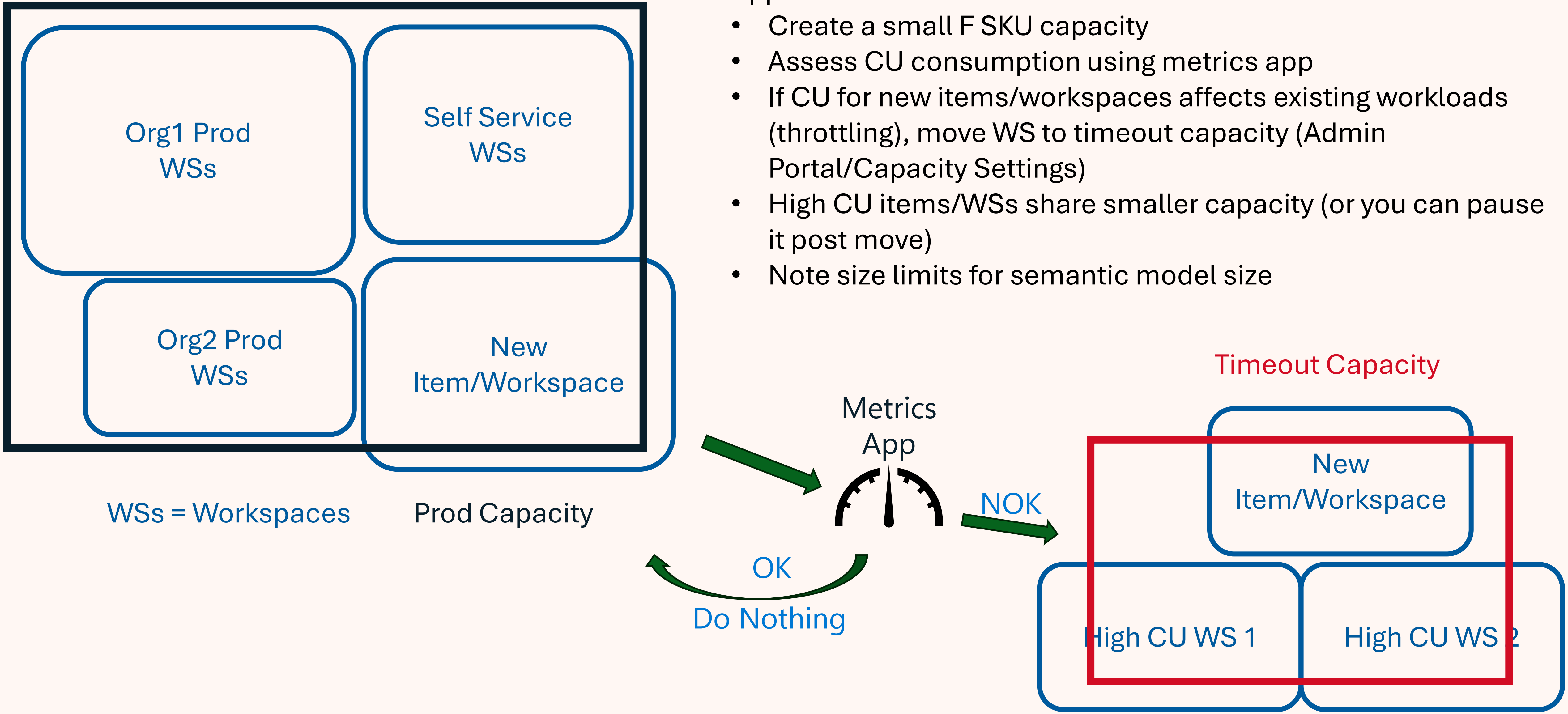
Approach

- Create a small F SKU capacity to “tryout” new workspaces/items
- Assess CU consumption using metrics app
- If acceptable, move to prod capacity
- If not, optimize
- Pause tryout capacity when not in use, if possible
- Note size limits for semantic model size

Isolation Strategy #4b – Timeout Capacity

Approach

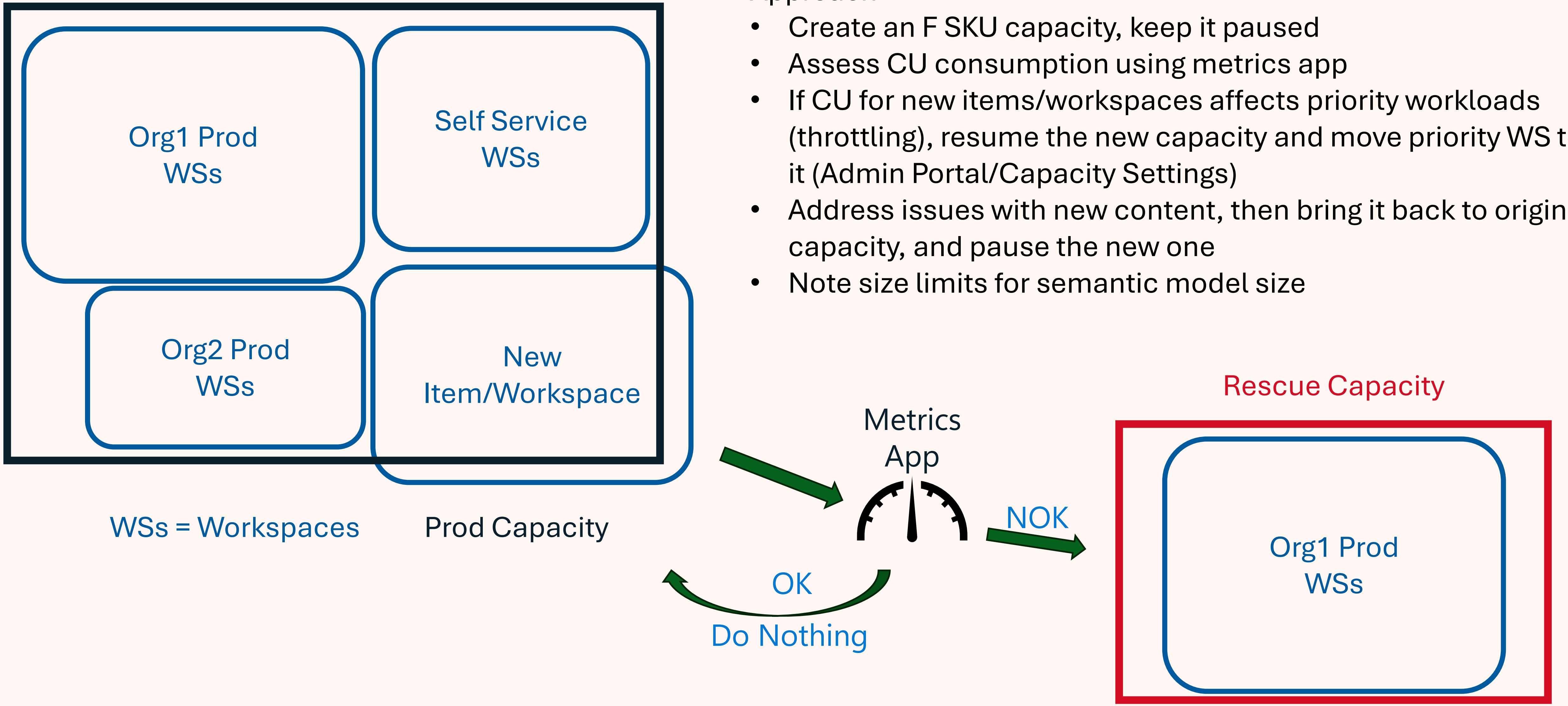
- Create a small F SKU capacity
- Assess CU consumption using metrics app
- If CU for new items/workspaces affects existing workloads (throttling), move WS to timeout capacity (Admin Portal/Capacity Settings)
- High CU items/WSs share smaller capacity (or you can pause it post move)
- Note size limits for semantic model size



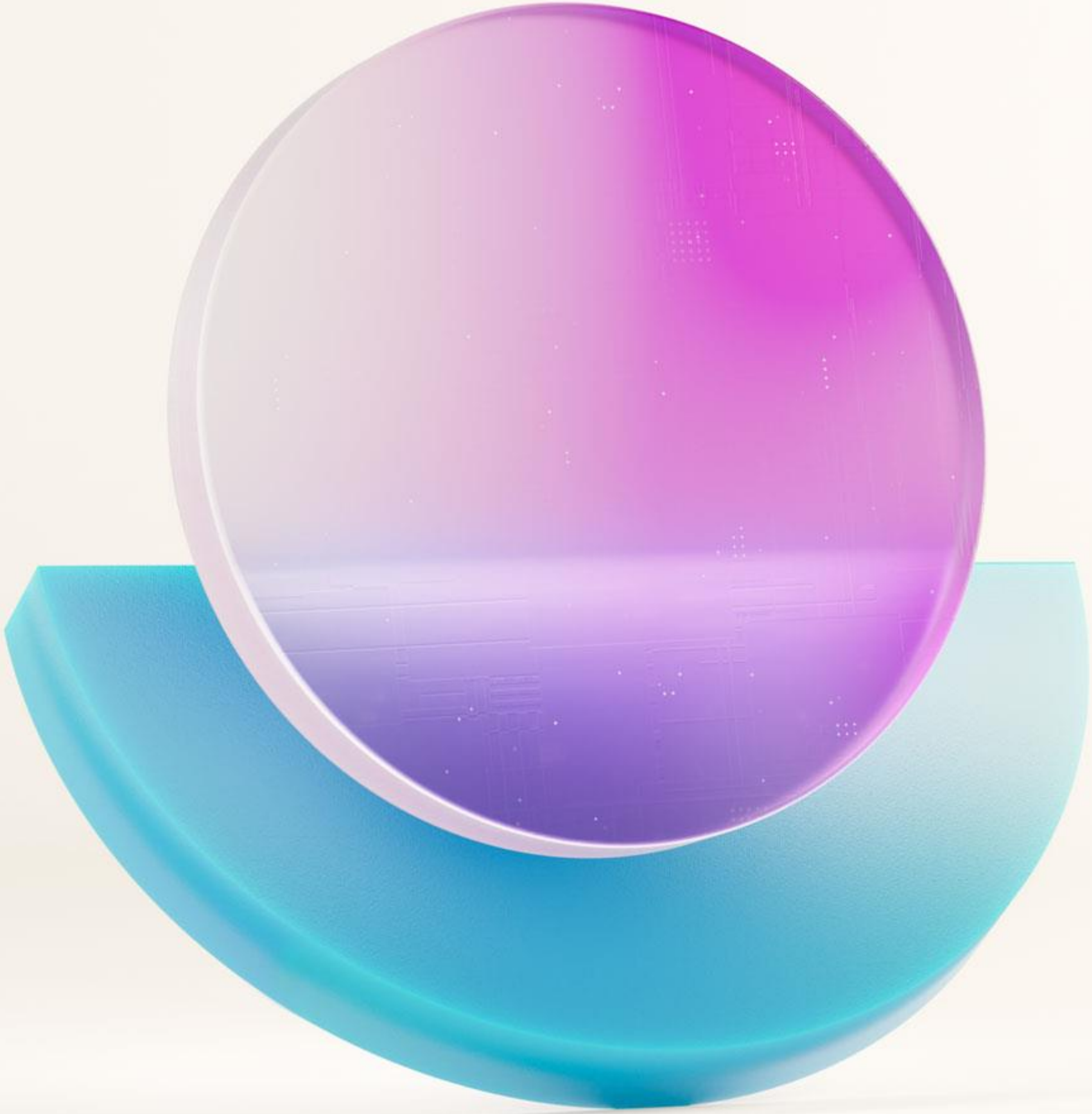
Isolation Strategy #4c – Rescue Capacity

Approach

- Create an F SKU capacity, keep it paused
- Assess CU consumption using metrics app
- If CU for new items/workspaces affects priority workloads (throttling), resume the new capacity and move priority WS to it (Admin Portal/Capacity Settings)
- Address issues with new content, then bring it back to original capacity, and pause the new one
- Note size limits for semantic model size



Microsoft Fabric
Community Conference



Protecting Capacities

Surge Protection v1 – Background usage limits

In Public Preview since Jan 2025

Enabling Capacity Admins to get ahead of throttling.

Simple experience that limits overuse by background jobs

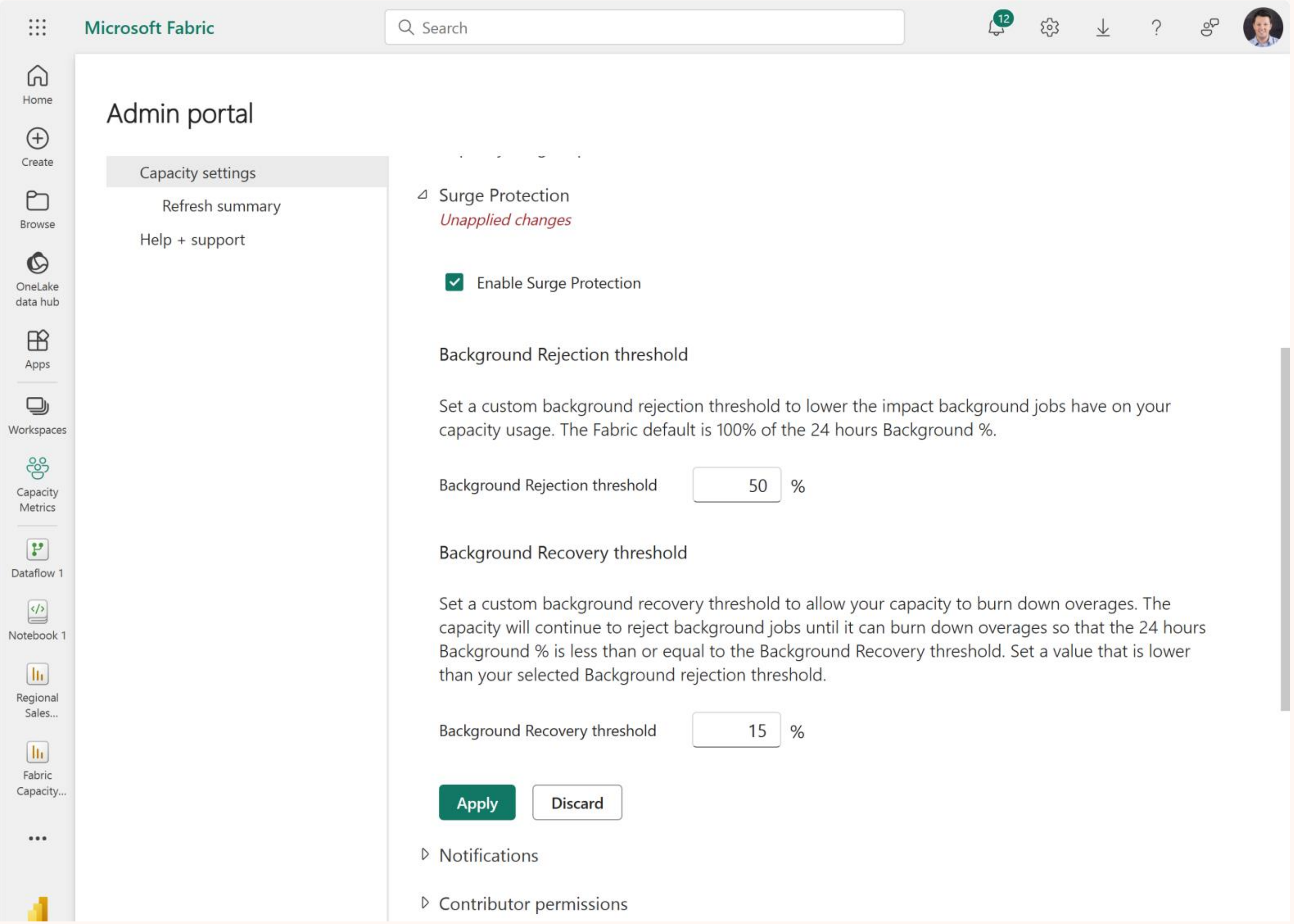
- Throttles background jobs before a full 24-hours of CUs is consumed.
- Helps to **protect interactive** usage like Power BI

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

Unapplied changes

☑ Enable Surge Protection

Background Rejection threshold

Set a custom background rejection threshold to lower the impact background jobs have on your capacity usage. The Fabric default is 100% of the 24 hours Background %.

Background Rejection threshold 50 %

Background Recovery threshold

Set a custom background recovery threshold to allow your capacity to burn down overages. The capacity will continue to reject background jobs until it can burn down overages so that the 24 hours Background % is less than or equal to the Background Recovery threshold. Set a value that is lower than your selected Background rejection threshold.

Background Recovery threshold 15 %

Autoscale billing for Apache Spark

Public Preview

Enable everyone in your org to use Spark and manage its cost

Serverless style billing for Spark jobs

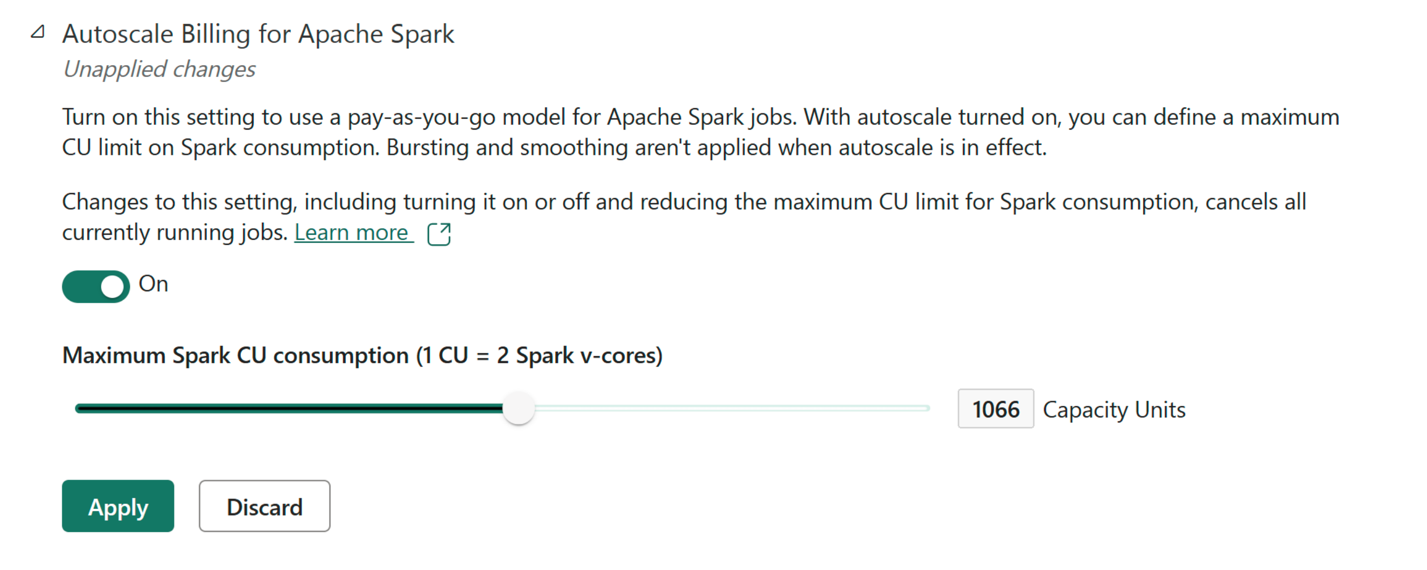
- Capacity admins can opt-in
- Set a max limit on CU used by Spark
- Only pay for what you use
- Spark manages the limit ensuring pools don't over consume
- Observability in a new metrics app page

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



Autoscale Billing for Apache Spark

Unapplied changes

Turn on this setting to use a pay-as-you-go model for Apache Spark jobs. With autoscale turned on, you can define a maximum CU limit on Spark consumption. Bursting and smoothing aren't applied when autoscale is in effect.

Changes to this setting, including turning it on or off and reducing the maximum CU limit for Spark consumption, cancels all currently running jobs. Learn more

On

Maximum Spark CU consumption (1 CU = 2 Spark v-cores)

1066 Capacity Units

Autoscale billing for Apache Spark

Public Preview

Monitor consumption in Metrics app

New Autoscale Compute for Spark page

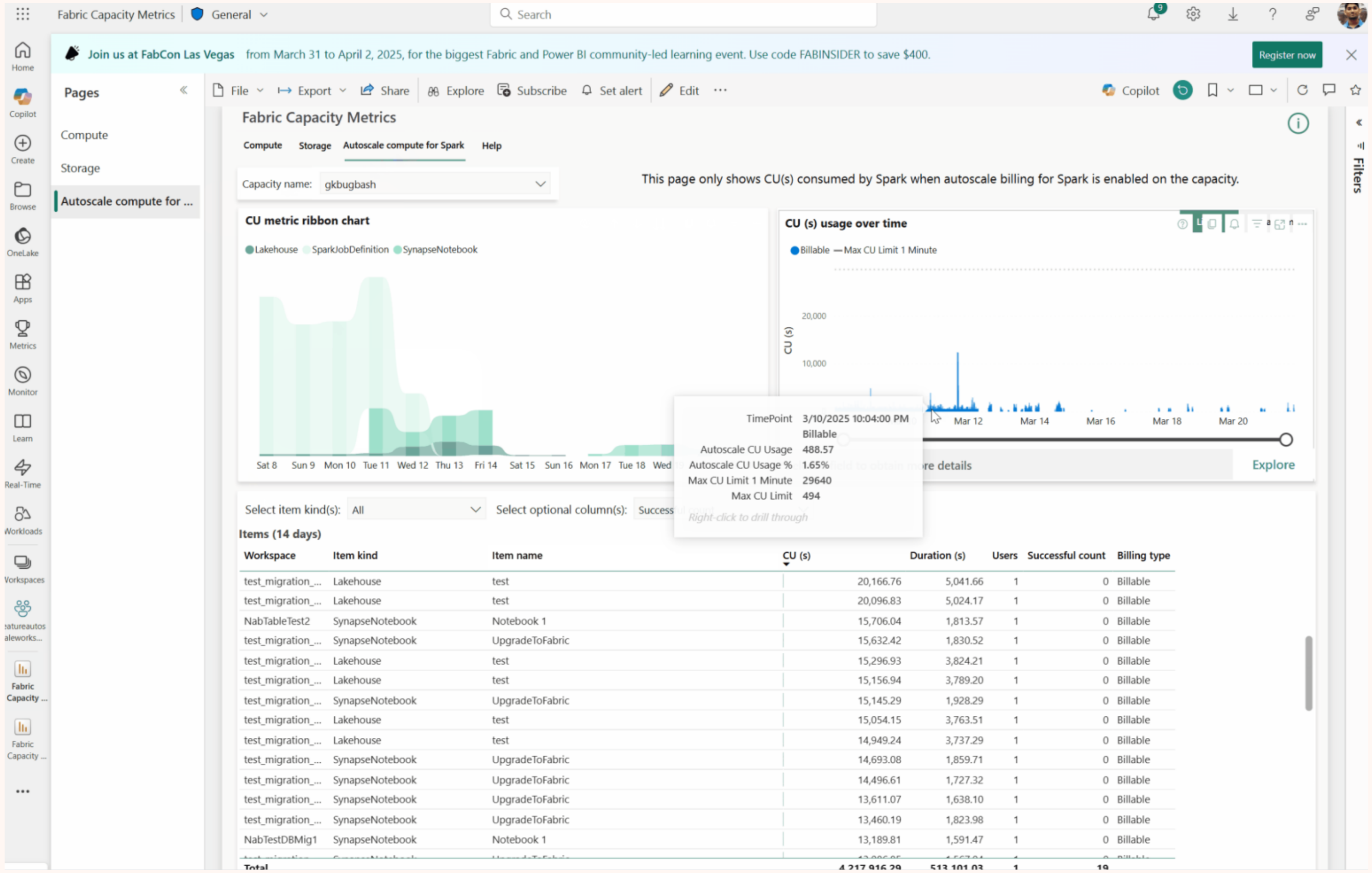
- Shows Spark CU consumed through Autoscale
- Easy to track to the configured autoscale limit.

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



Join us at FabCon Las Vegas from March 31 to April 2, 2025, for the biggest Fabric and Power BI community-led learning event. Use code FABINSIDER to save $400.

Fabric Capacity Metrics

Capacity name: gkbugbash

This page only shows CU(s) consumed by Spark when autoscale billing for Spark is enabled on the capacity.

TimePoint 3/10/2025 10:04:00 PM
Billable
Autoscale CU Usage 488.57
Autoscale CU Usage % 1.65%
Max CU Limit 1 Minute 29640
Max CU Limit 494
Right-click to drill through

Select item kind(s): All

Select optional column(s): Success

Items (14 days)

Workspace	Item kind	Item name	CU (s)	Duration (s)	Users	Successful count	Billing type
test_migration_...	Lakehouse	test	20,166.76	5,041.66	1	0	Billable
test_migration_...	Lakehouse	test	20,096.83	5,024.17	1	0	Billable
NabTableTest2	SynapseNotebook	Notebook 1	15,706.04	1,813.57	1	0	Billable
test_migration_...	SynapseNotebook	UpgradeToFabric	15,632.42	1,830.52	1	0	Billable
test_migration_...	Lakehouse	test	15,296.93	3,824.21	1	0	Billable
test_migration_...	Lakehouse	test	15,156.94	3,789.20	1	0	Billable
test_migration_...	SynapseNotebook	UpgradeToFabric	15,145.29	1,928.29	1	0	Billable
test_migration_...	Lakehouse	test	15,054.15	3,763.51	1	0	Billable
test_migration_...	Lakehouse	test	14,949.24	3,737.29	1	0	Billable
test_migration_...	SynapseNotebook	UpgradeToFabric	14,693.08	1,859.71	1	0	Billable
test_migration_...	SynapseNotebook	UpgradeToFabric	14,496.61	1,727.32	1	0	Billable
test_migration_...	SynapseNotebook	UpgradeToFabric	13,611.07	1,638.10	1	0	Billable
test_migration_...	SynapseNotebook	UpgradeToFabric	13,460.19	1,823.98	1	0	Billable
NabTestDBMig1	SynapseNotebook	Notebook 1	13,189.81	1,591.47	1	0	Billable
Total			4,217,916.29	513,101.03	1	19	

Fabric Copilot Capacity

Nov 2024

Enable everyone in your org to use Copilot and manage its cost

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

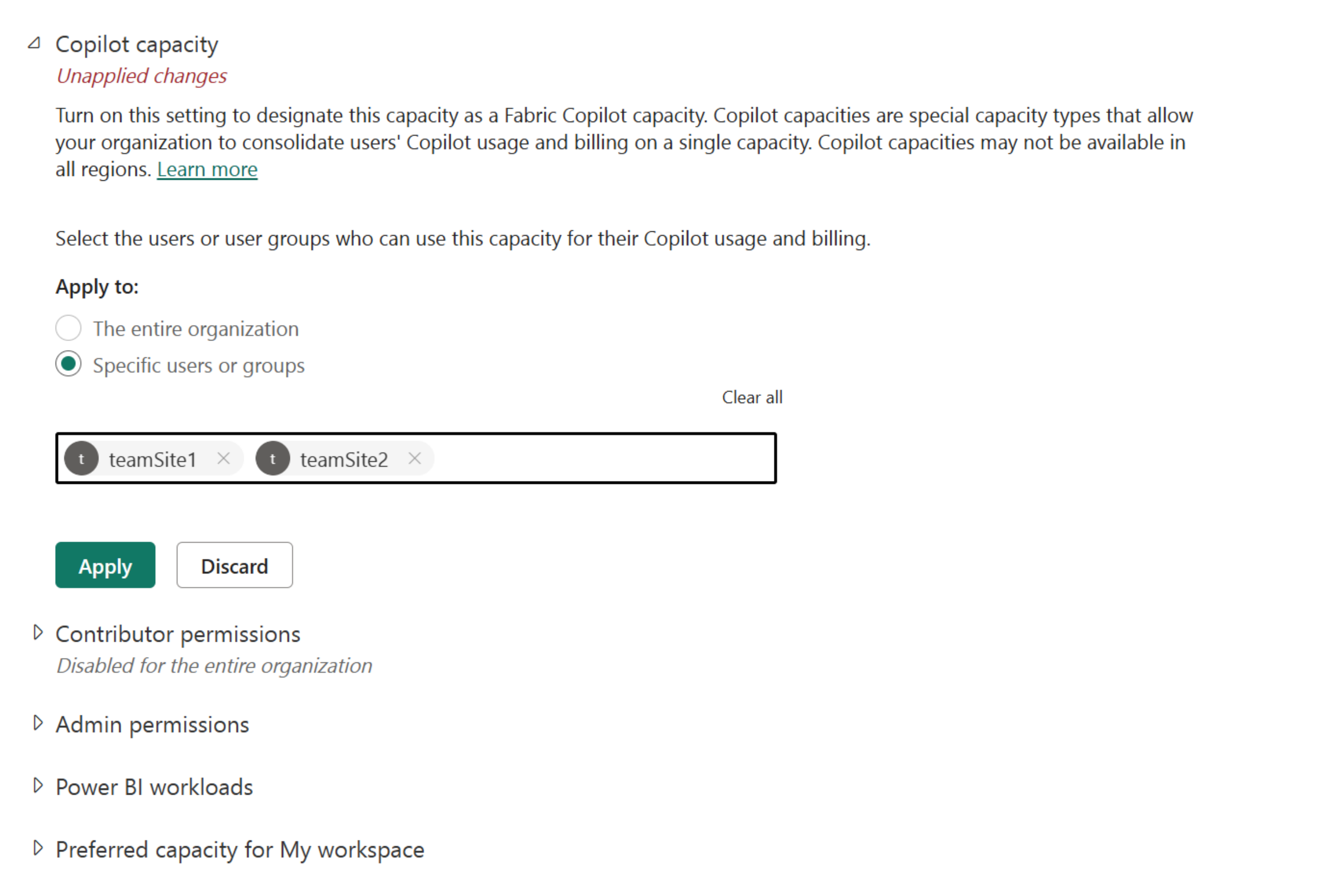
- Select the Users or groups who use a specific Copilot capacity
- A user can have only one Copilot capacity
  - Newest one matters..

No longer just P & F SKUs

- Pro, Premium Per User, and Trial

Tenant Setting

- Restrict who can configure
- "[Capacities can be delegated ..](#)"



◿ Copilot capacity

Unapplied changes

Turn on this setting to designate this capacity as a Fabric Copilot capacity. Copilot capacities are special capacity types that allow your organization to consolidate users' Copilot usage and billing on a single capacity. Copilot capacities may not be available in all regions. Learn more

Select the users or user groups who can use this capacity for their Copilot usage and billing.

Apply to:

○ The entire organization

● Specific users or groups

Clear all

teamSite1 × teamSite2 ×

▷ Contributor permissions

Disabled for the entire organization

▷ Admin permissions

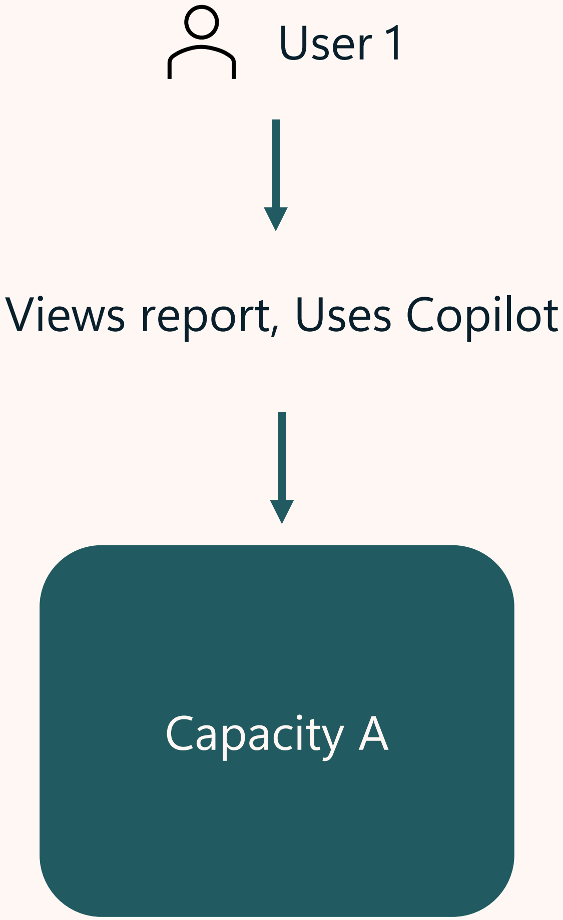
▷ Power BI workloads

▷ Preferred capacity for My workspace

Fabric Copilot capacity

Nov 2024

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



User 1

Views report, Uses Copilot

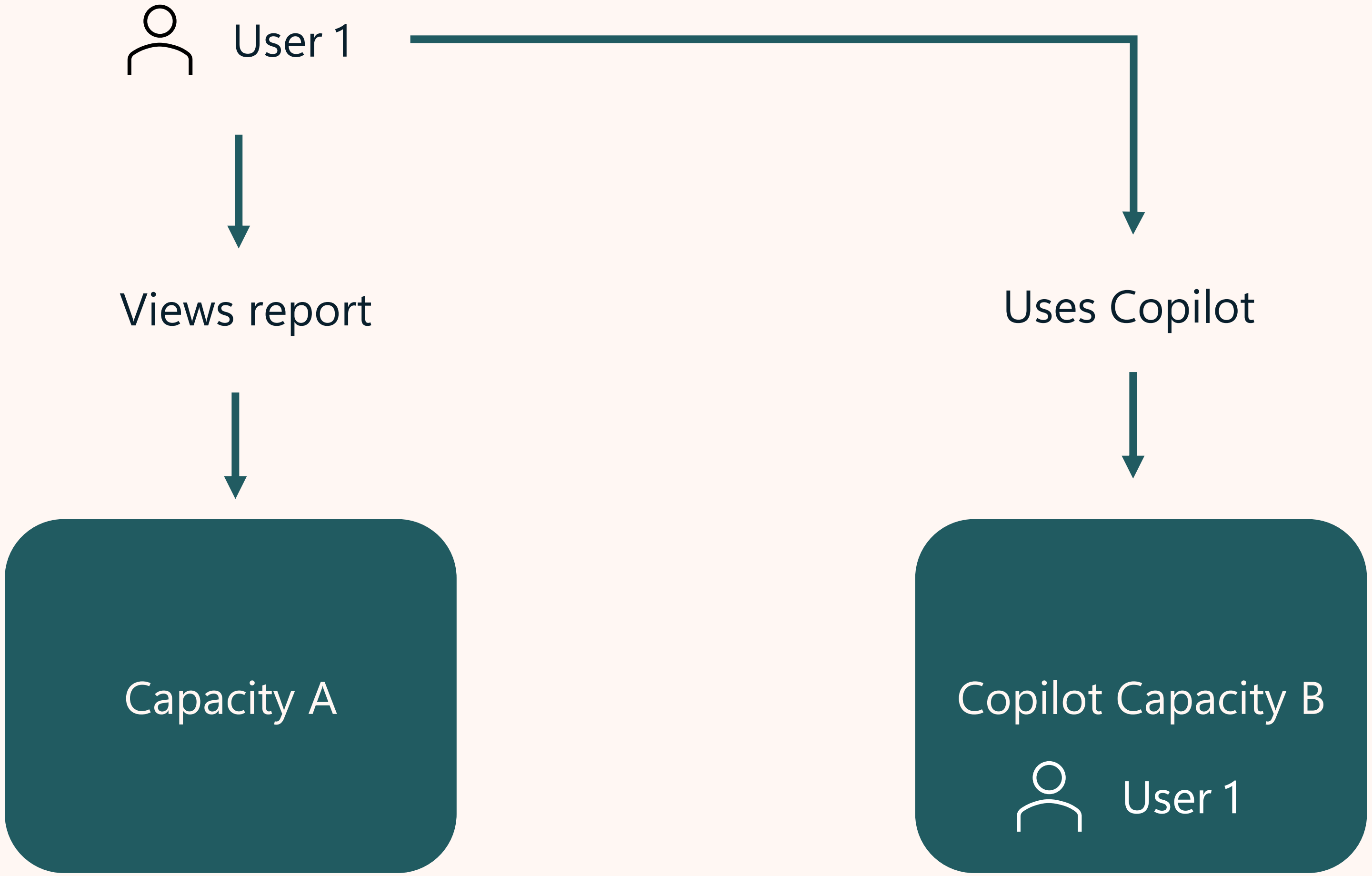
Capacity A

User 1

Fabric Copilot capacity

Nov 2024

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



User 1

Views report

Uses Copilot

Capacity A

Copilot Capacity B

User 1

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

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 .. 😼

Azure Quota Management Service Integration

GA (March 2025)

Better resource allocation to meet Microsoft's customer capacity needs

New 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

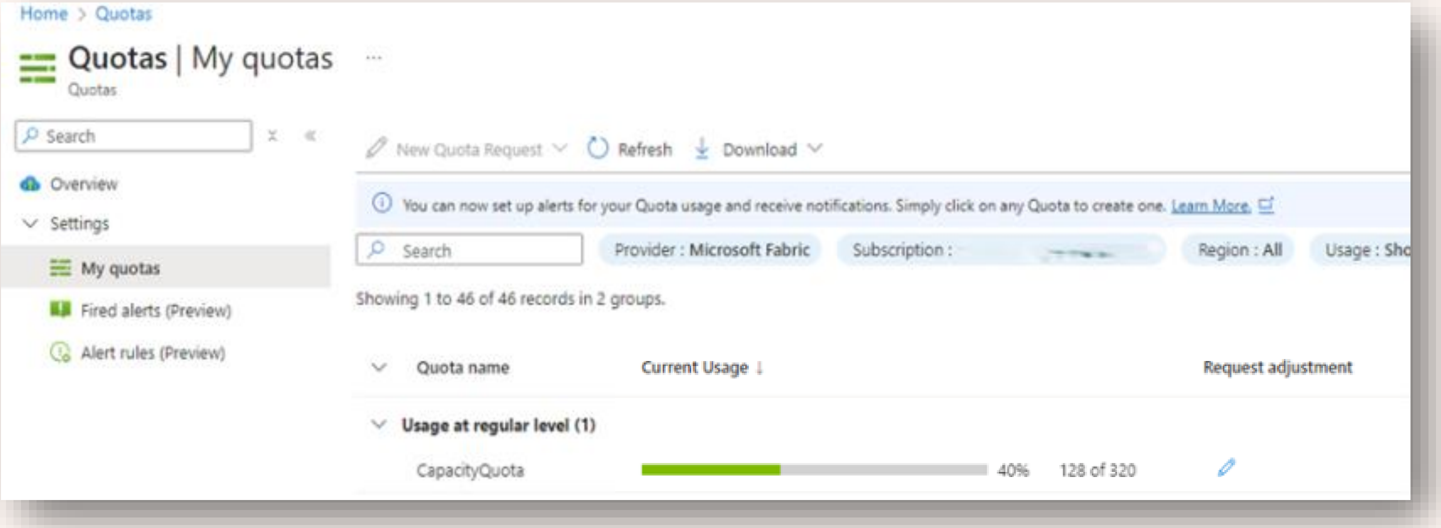
- Auto approved up to specific limits
- Customers can request additional quota through Microsoft customer support

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

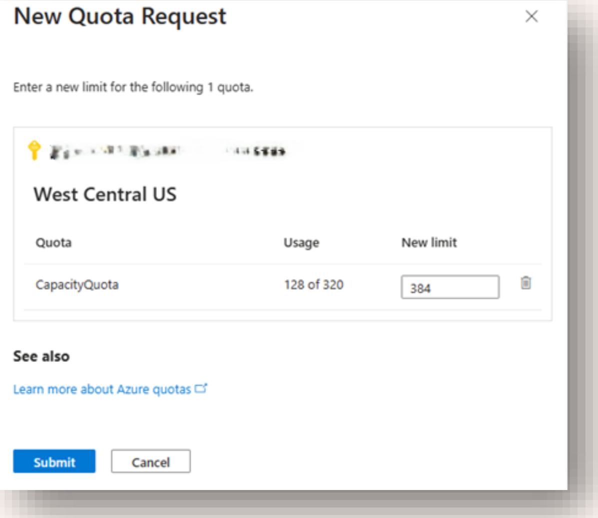


You can now set up alerts for your Quota usage and receive notifications. Simply click on any Quota to create one. Learn More.

Search | Provider : Microsoft Fabric | Subscription : [redacted] | Region : All | Usage : Sho

Showing 1 to 46 of 46 records in 2 groups.

Quota name	Current Usage ↓		Request adjustment
Usage at regular level (1)			
CapacityQuota	40%	128 of 320	✎


New Quota Request

Enter a new limit for the following 1 quota.

[redacted]

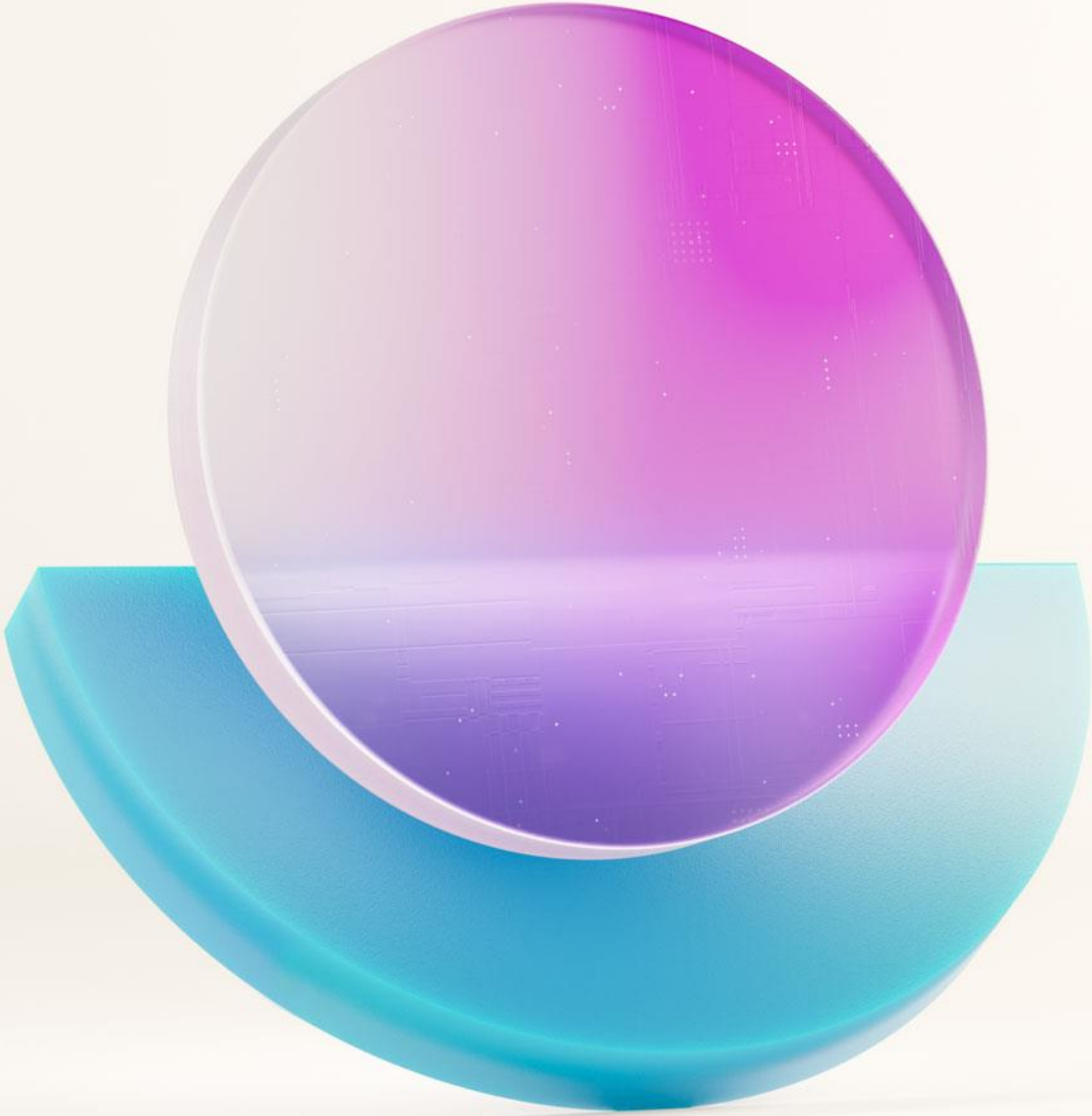
West Central US

Quota	Usage	New limit
CapacityQuota	128 of 320	384

See also

Learn more about Azure quotas

Microsoft Fabric
Community Conference



Monitoring Capacities

Capacity Chargeback Reporting

Public Preview

Allocate costs to those who use your capacity

Helps allocate costs across your org

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

Focuses on % utilization

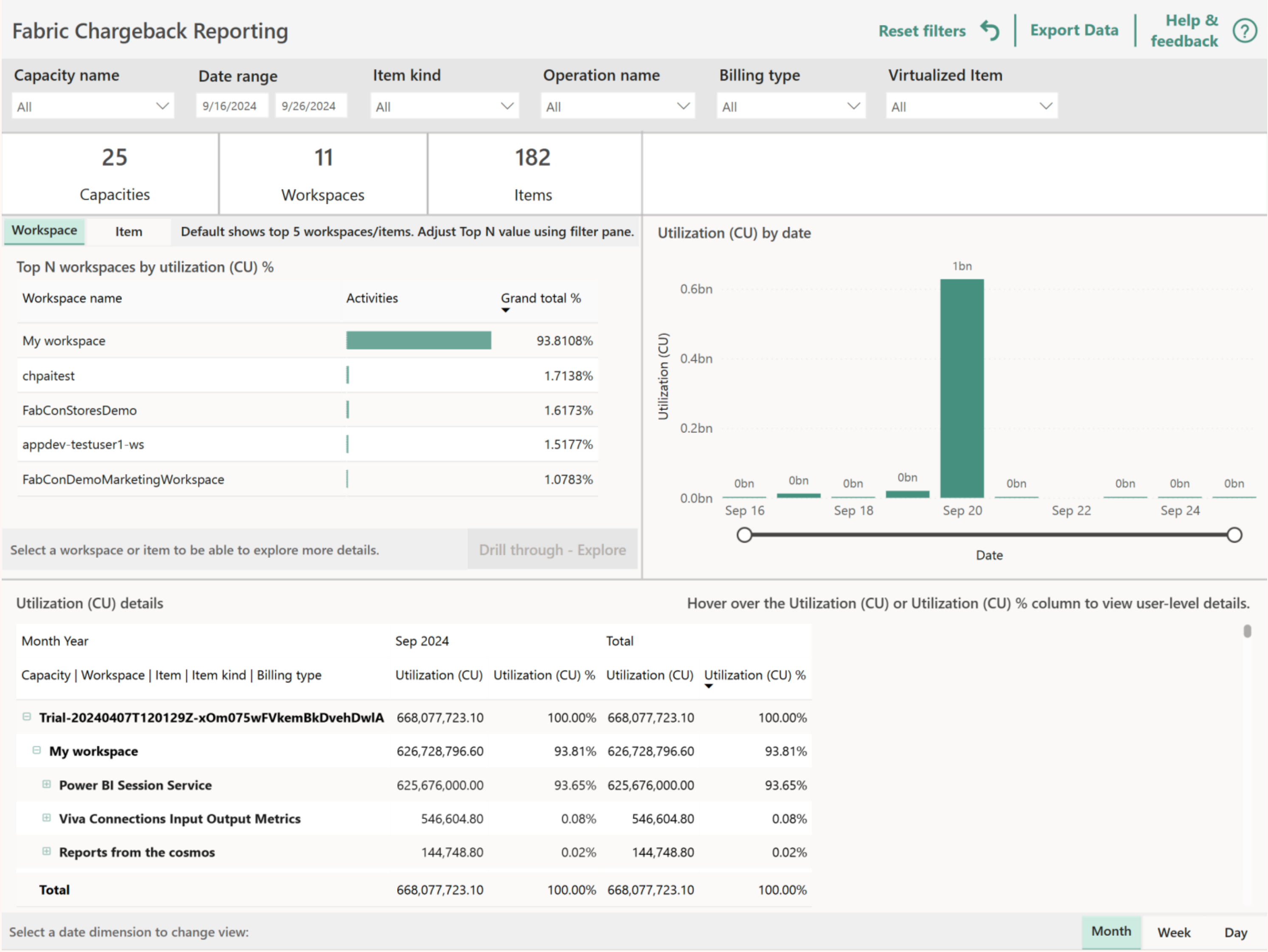
- Orgs need to look at cost in Azure billing and then allocate that cost to their content owners.

Impact

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

Future

- Domain support
- Tags support



Fabric Chargeback Reporting

Capacity name: All
Date range: 9/16/2024 9/26/2024
Item kind: All
Operation name: All
Billing type: All
Virtualized Item: All

25	11	182
Capacities	Workspaces	Items

Default shows top 5 workspaces/items. Adjust Top N value using filter pane.

Top N workspaces by utilization (CU) %

Workspace name	Activities	Grand total %
My workspace		93.8108%
chpaitest		1.7138%
FabConStoresDemo		1.6173%
appdev-testuser1-ws		1.5177%
FabConDemoMarketingWorkspace		1.0783%

Select a workspace or item to be able to explore more details.

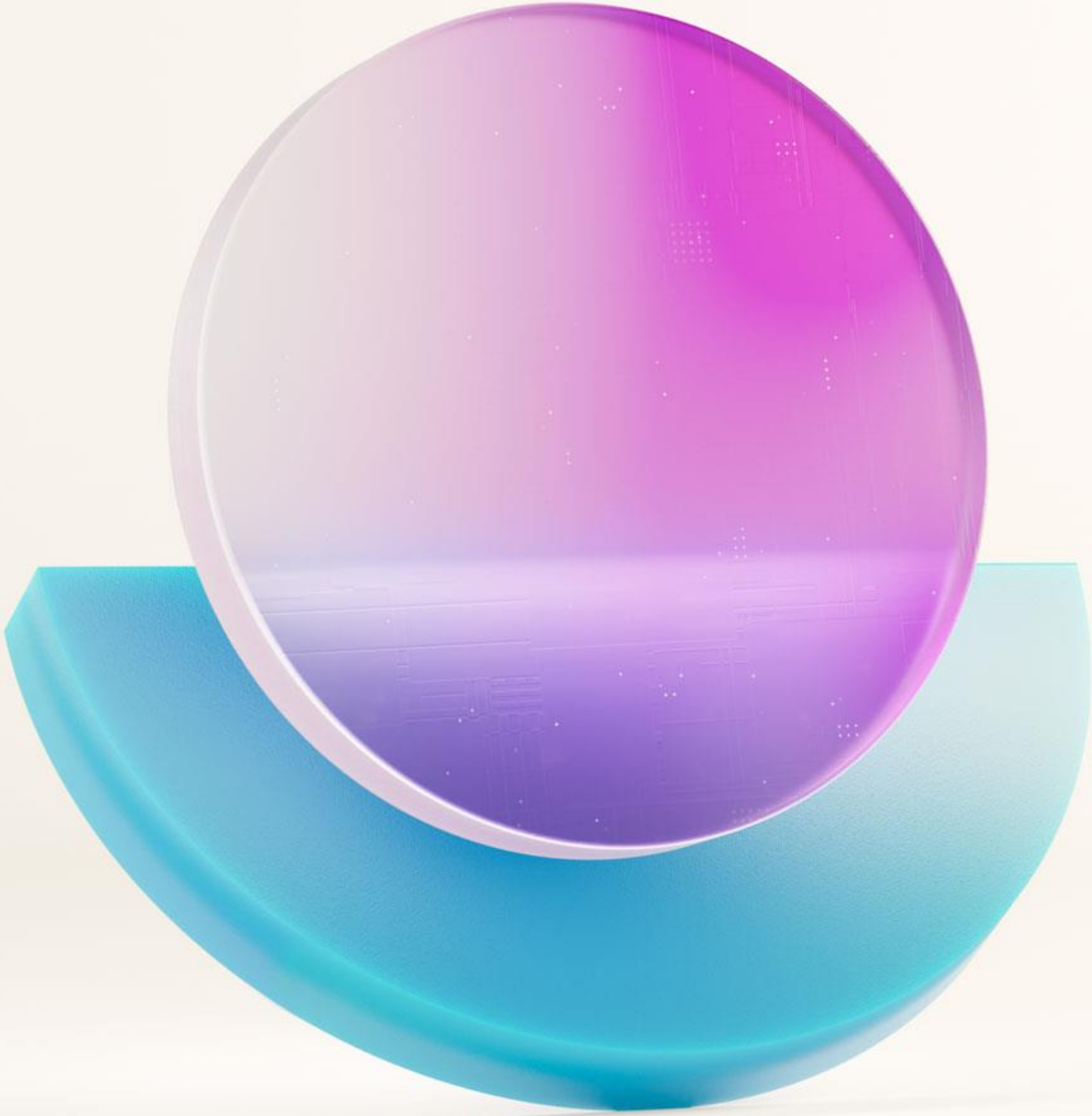
Utilization (CU) details

Hover over the Utilization (CU) or Utilization (CU) % column to view user-level details.

Month Year	Sep 2024		Total	
Capacity | Workspace | Item | Item kind | Billing type	Utilization (CU)	Utilization (CU) %	Utilization (CU)	Utilization (CU) %
⊟ **Trial-20240407T120129Z-xOm075wFVkemBkDvehDwlA**	668,077,723.10	100.00%	668,077,723.10	100.00%
⊟ **My workspace**	626,728,796.60	93.81%	626,728,796.60	93.81%
⊞ **Power BI Session Service**	625,676,000.00	93.65%	625,676,000.00	93.65%
⊞ **Viva Connections Input Output Metrics**	546,604.80	0.08%	546,604.80	0.08%
⊞ **Reports from the cosmos**	144,748.80	0.02%	144,748.80	0.02%
Total	668,077,723.10	100.00%	668,077,723.10	100.00%

Select a date dimension to change view:

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

For general purpose compute needs

Sized for typical needs

Capacity B

For Self-Service Reporting workloads

Capacity C

For large periodic workloads

Paused when not needed

Capacity D

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

Slides

Eval

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