

Understanding Fabric Capacities

Benni De Jagere



Slides





Benni De Jagere

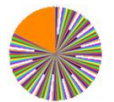
Senior Program Manager | Fabric Customer Advisory Team (FabricCAT)



dataMinds



sessionize



Fabric CAT

.be Member

@BenniDeJagere

/bennidejagere

/bennidejagere

/bennidejagere

#SayNoToPieCharts





Fabric Capacities Introduction



Microsoft Fabric



Data
Factory



Synapse Data
Engineering



Synapse Data
Science



Synapse Data
Warehouse



Synapse Real
Time
Intelligence



Power BI



Data
Activator

AI Assisted

Shared Workspaces

Universal Compute Capacities

One Security

OneLake

Intelligent data foundation

Single...

Onboarding and trials
Sign-on
Navigation model
UX model
Workspace organization
Collaboration experience
Data Lake
Storage format
Data copy for all engines
Security model
CI/CD
Monitoring hub
Governance & Capacity Metrics
Data Hub

Capacities are a shared resource

Shared across workloads

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

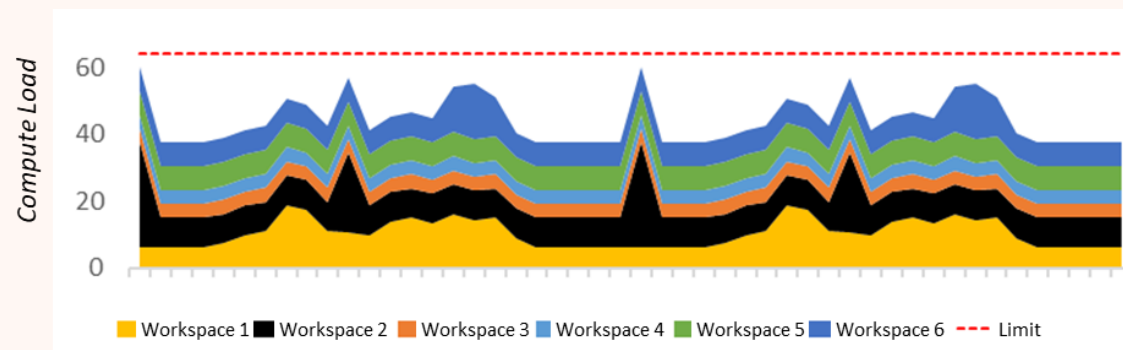
There is no need to allocate compute for each workload separately.



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

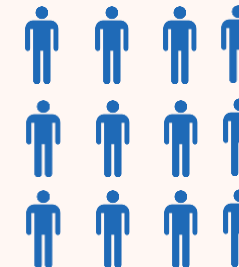
For each project, many developers will share a workspace where collaborative development and consumption at scale is managed.

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

Developers/Creators



Consumers

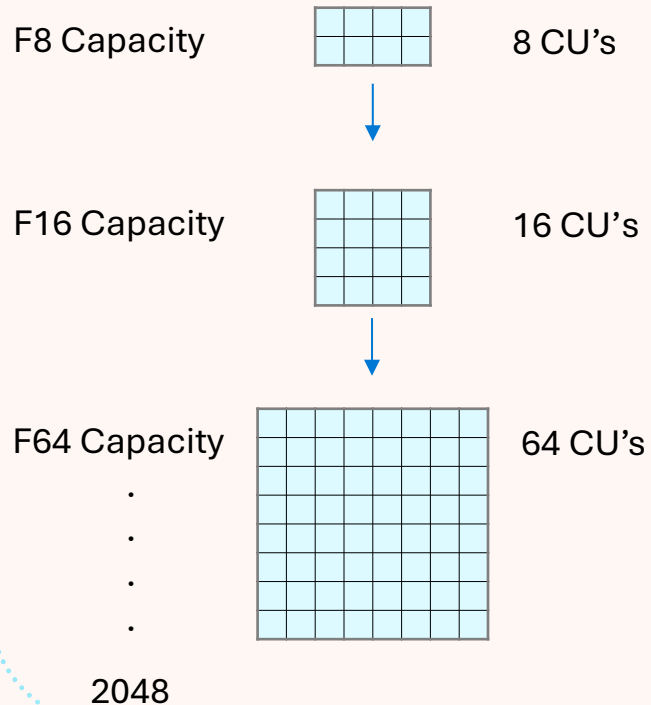


Capacities are flexible building blocks for growth

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

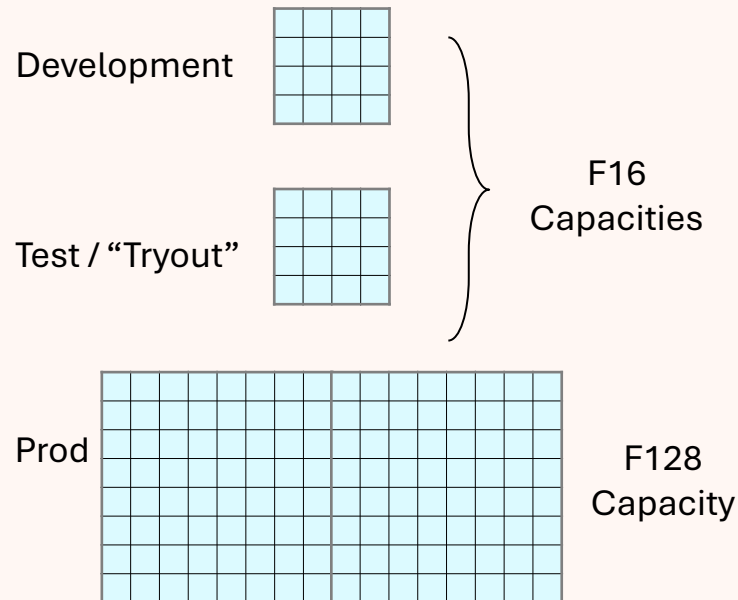
Scale Vertically

Increased capacity size provides more throughput



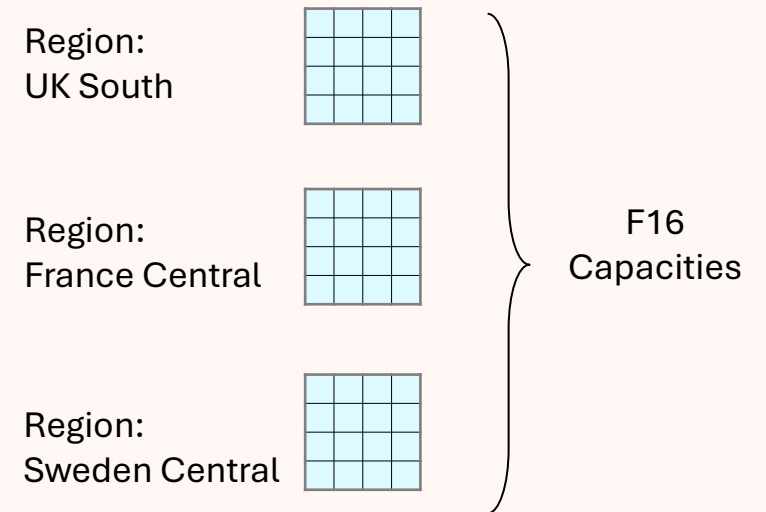
Scale Horizontally

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

Tenant (Microsoft)

Capacity 1 (West Europe)

Workspace 1 (Data Science Team)

Item 1
Power BI
Semantic Model

Item 2
AI Function
Evaluation

Workspace 2 (Research)

Item 3
Spark Notebook

Item 4
ML Dataflow

Capacity 2 (Central India)

Capacity 3 (West US)



Bursting and Smoothing

Smoothing intro and benefits

Load stabilization

Smoothing helps capacities self-stabilize by flattening large spikey loads into a smooth load profile, eliminating temporal spikes

Eliminates Scheduling contention

Large/scheduled Jobs usage (not execution) are smoothed over 24 hours, eliminating the need to decide the timing and order of job execution

Surge protection

Interactive operations smoothed over several minutes, preventing a single user with a very demanding query from hogging the entire capacity



What is Bursting?

Job acceleration

Bursting provides extra compute resources to jobs and queries to accelerate their completion

Go beyond

The extra resources of bursting allow jobs to **utilize far more resources than “face value”**

Instead of running a job on 64 CU and completing in 60 seconds, bursting could use 256 CUs to complete the job in 15 seconds.

Same amount of work, just completed faster

No hassle, No overload

Bursting is automatic when the system reasons it can accelerate the job by applying extra resources. No settings are required.

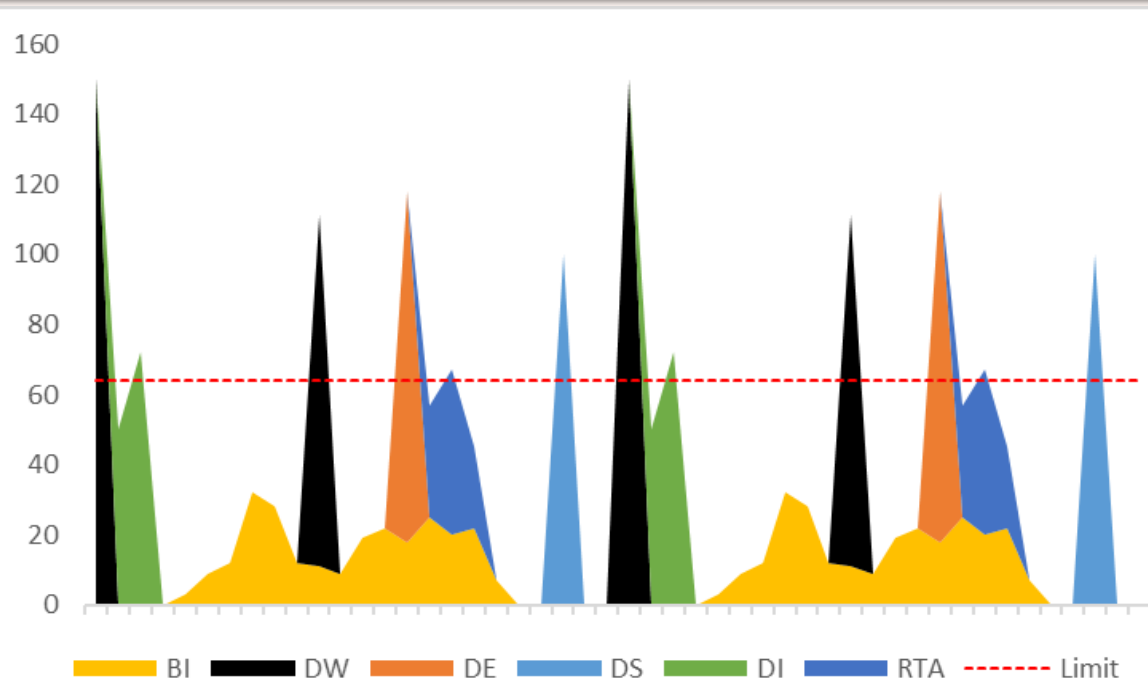
Bursting prevents an overload as the *smoothing* mechanism will always flatten the resource burst

Bursting and smoothing | before and after

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

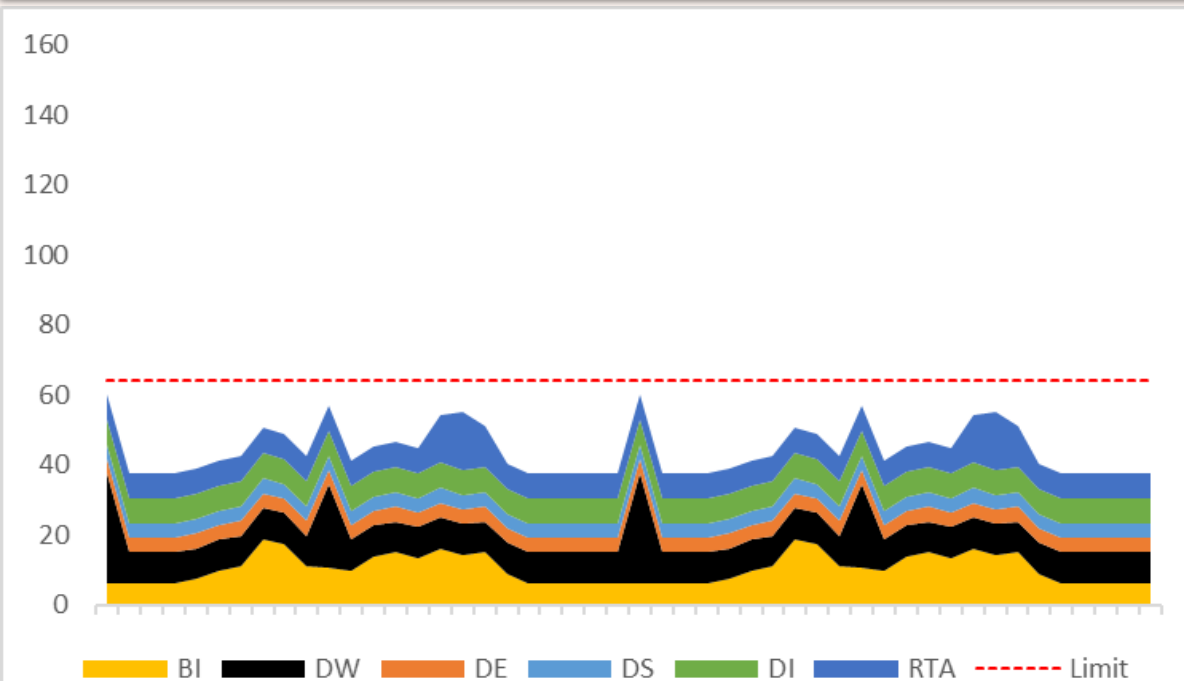
Before Smoothing

- 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

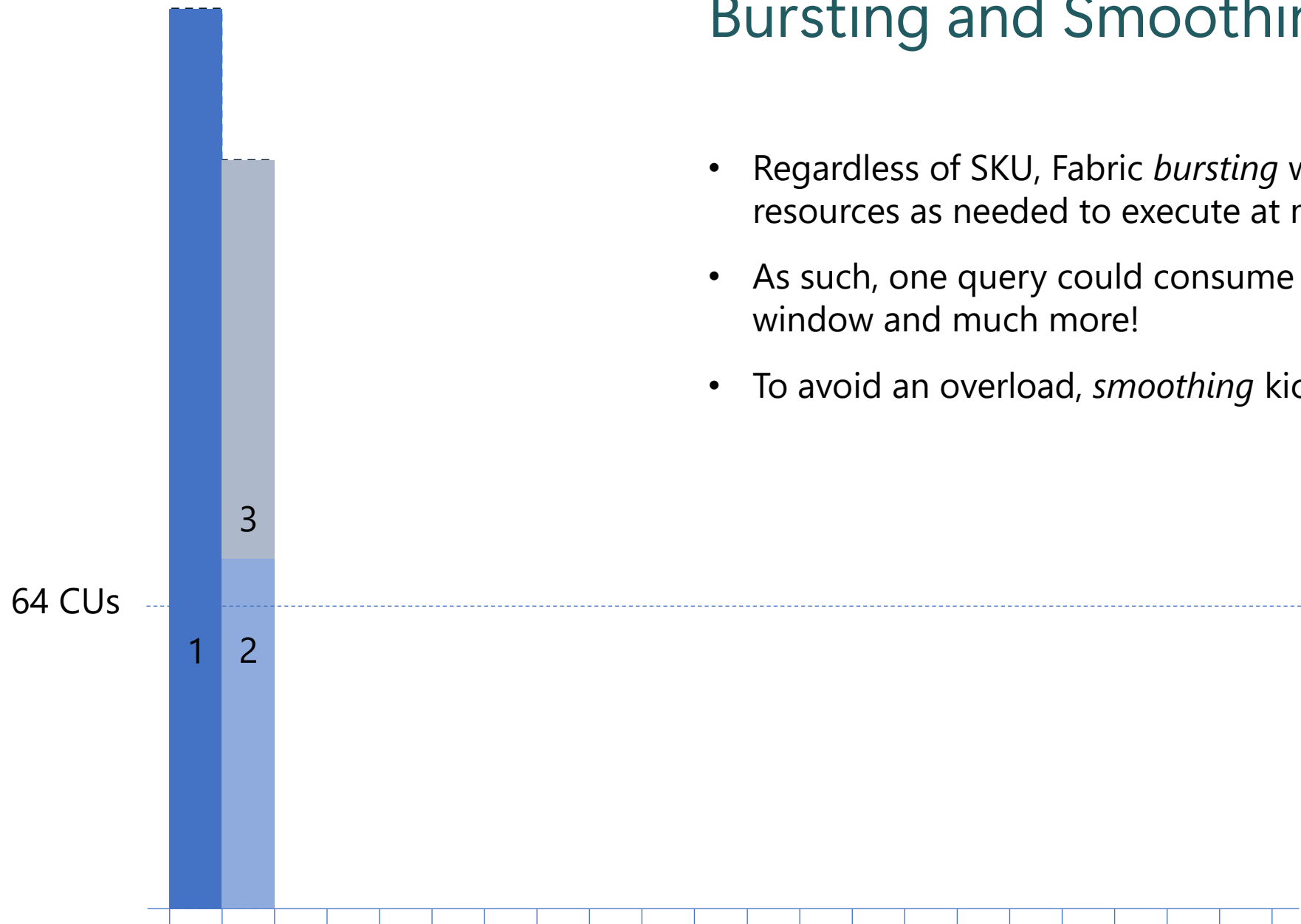


After Smoothing

- 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



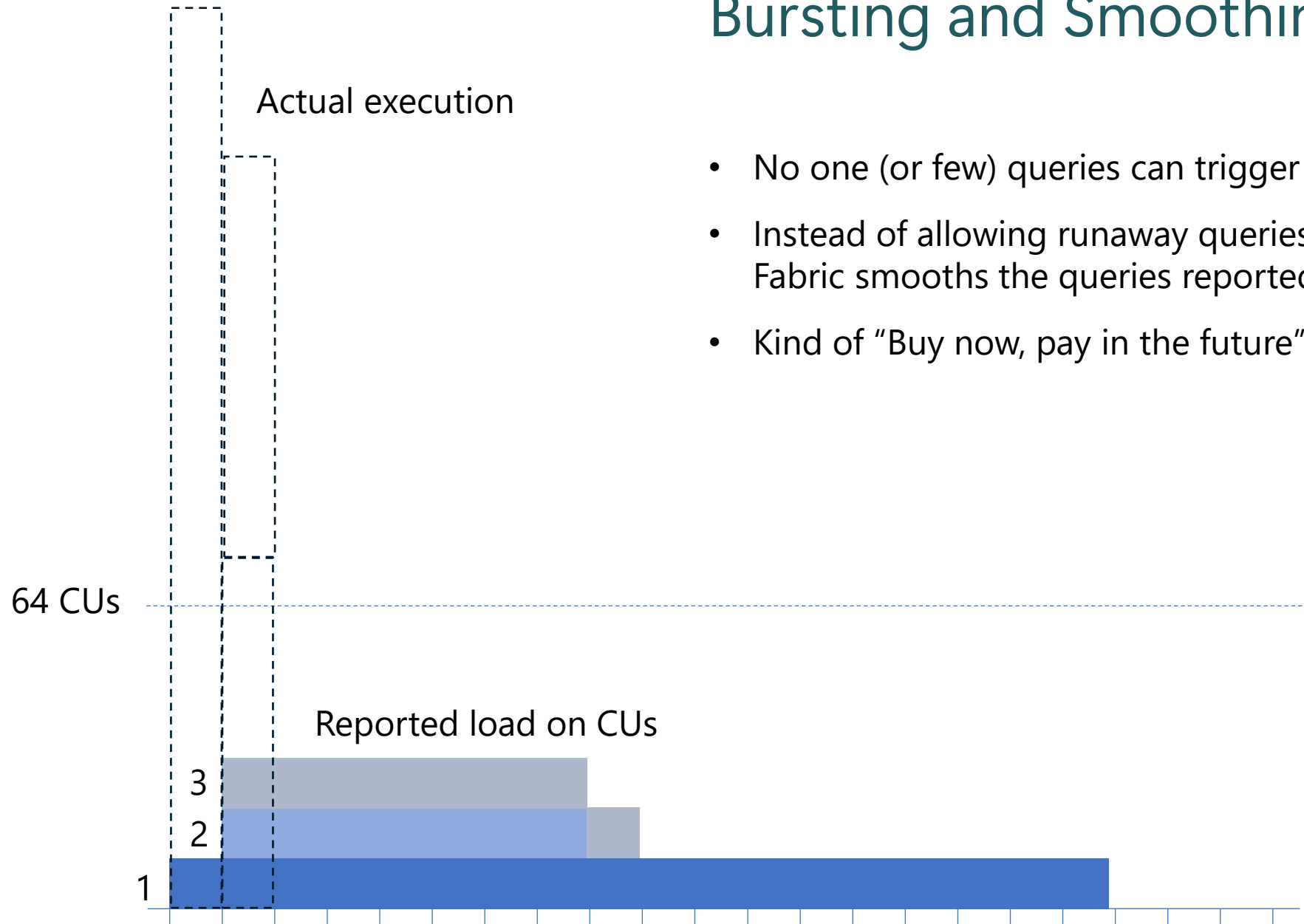
Jobs Executed



Bursting and Smoothing

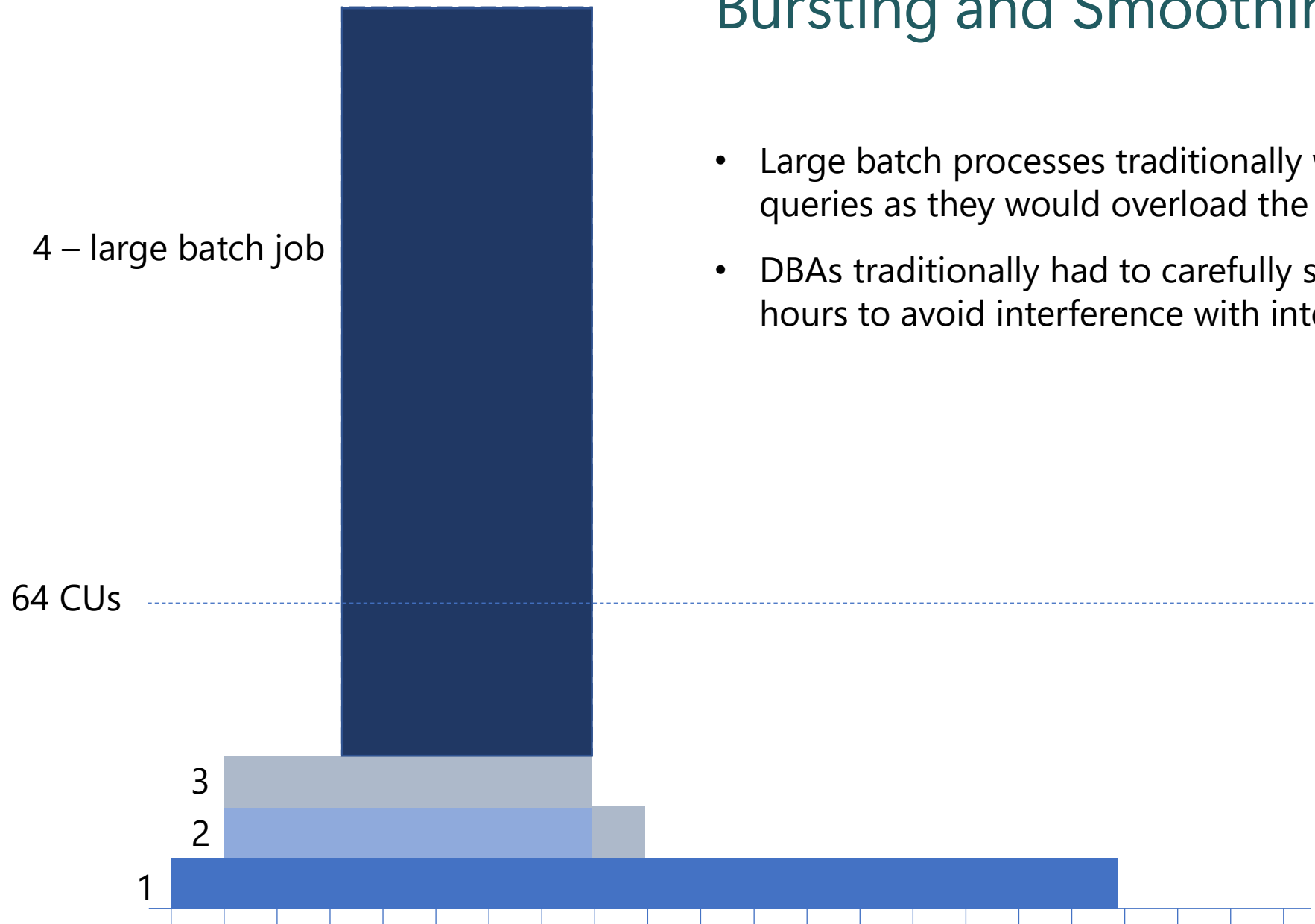
- Regardless of SKU, Fabric *bursting* will automatically allocate resources as needed to execute at maximum performance
- As such, one query could consume all the quota of a single time window and much more!
- To avoid an overload, *smoothing* kicks in

Bursting and Smoothing



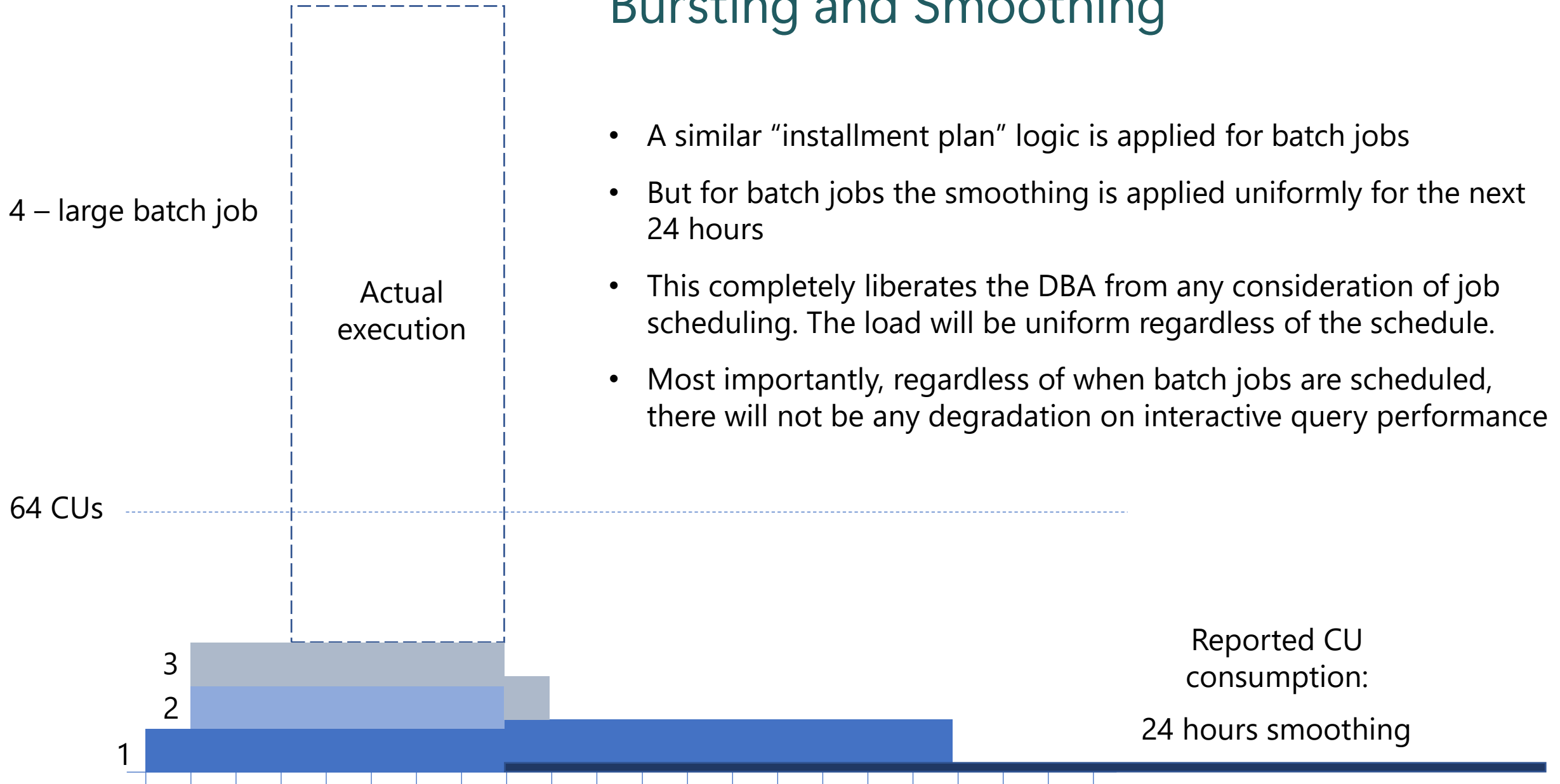
- No one (or few) queries can trigger an overload
- Instead of allowing runaway queries to create a local overload, Fabric smooths the queries reported usage to future time windows
- Kind of "Buy now, pay in the future" installment plan

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

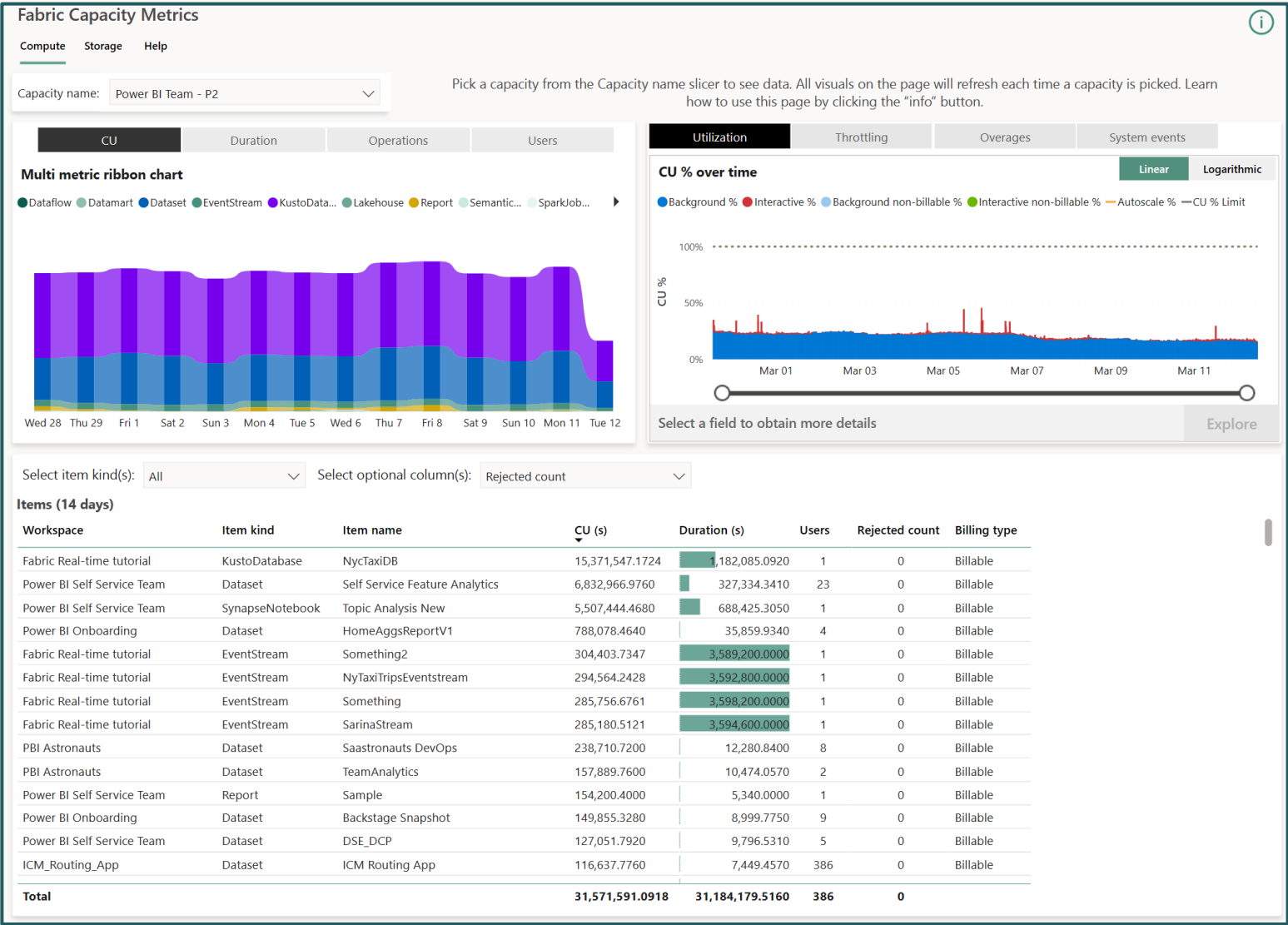


Monitoring with Capacity Metrics

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

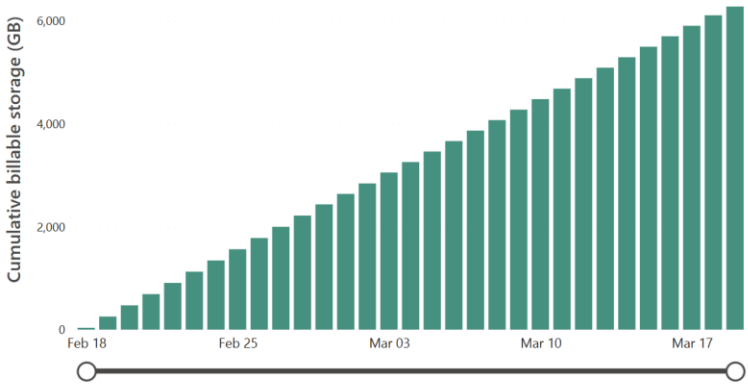
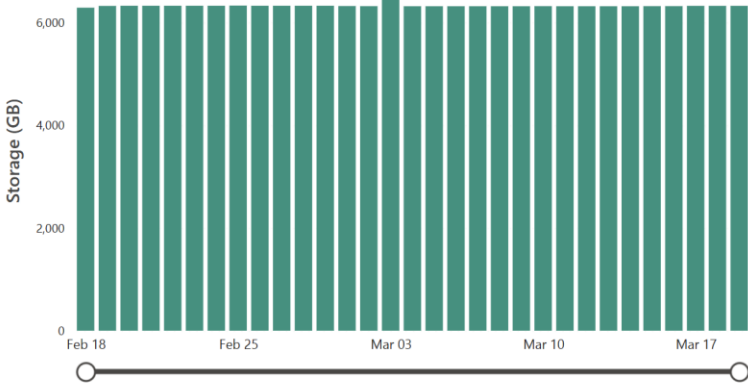
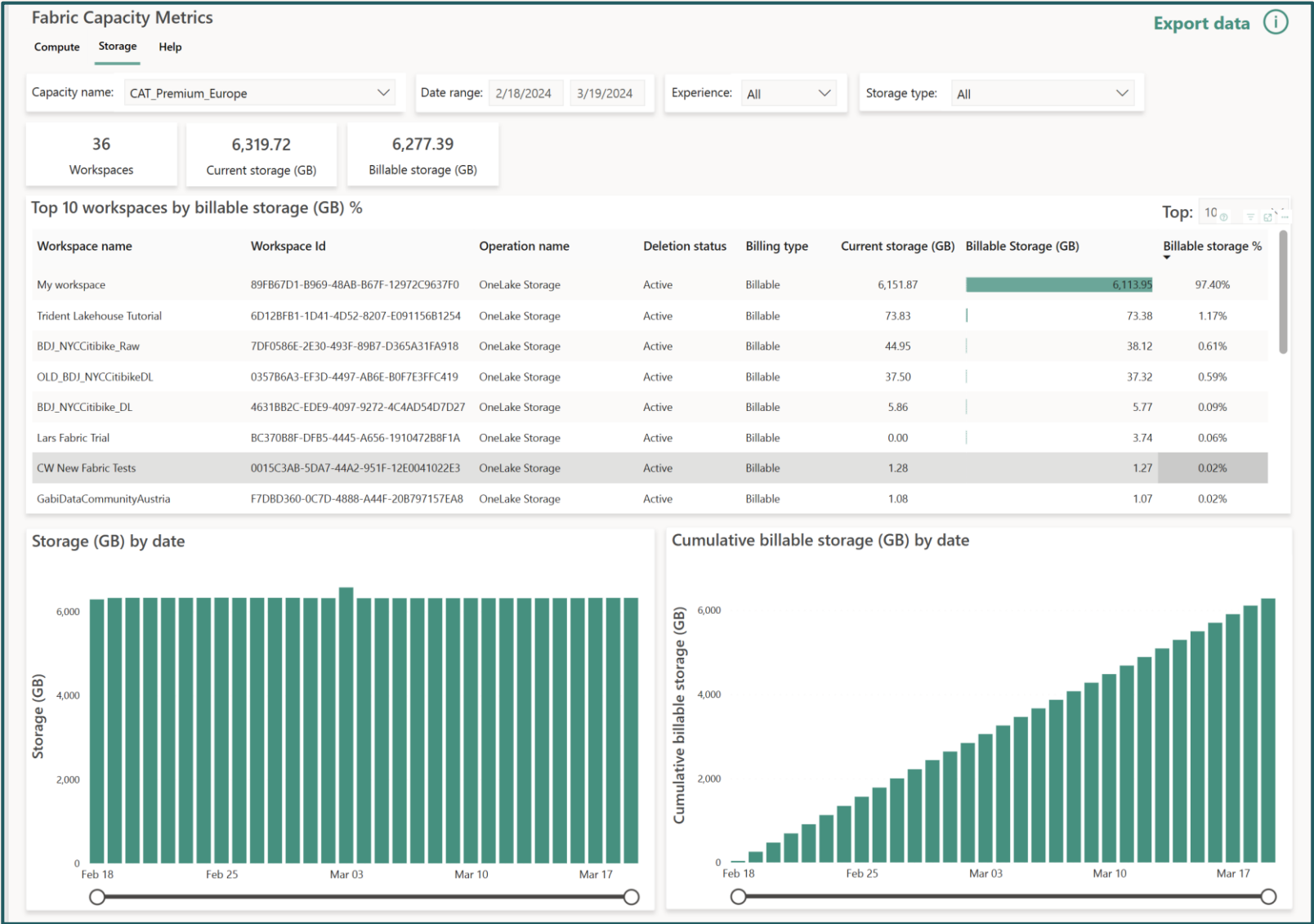


Capacity Metrics

Monitor OneLake consumption

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

Reconcile costs with internal chargeback processes






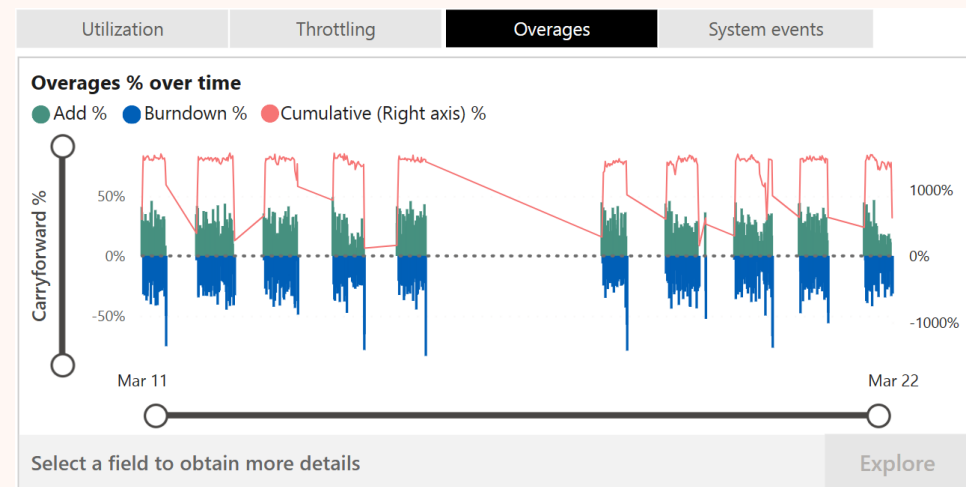
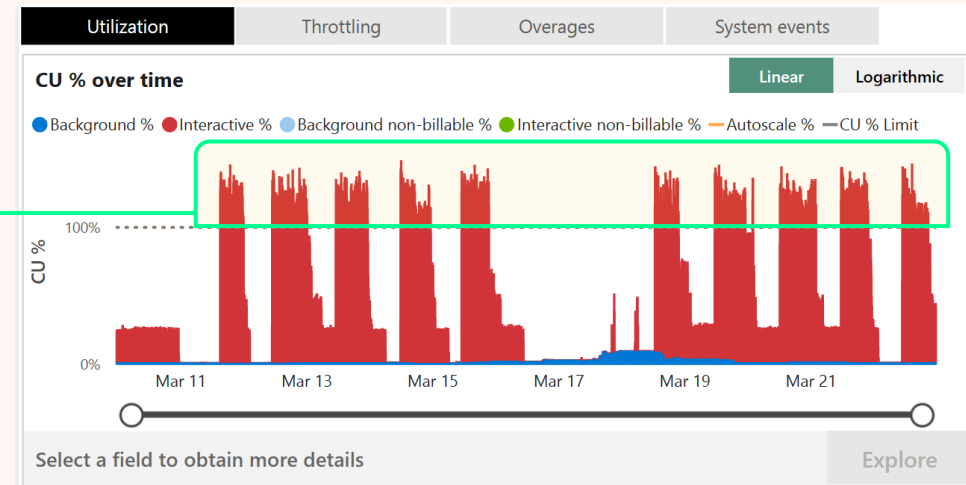


Capacity Throttling Policies

Throttling intro

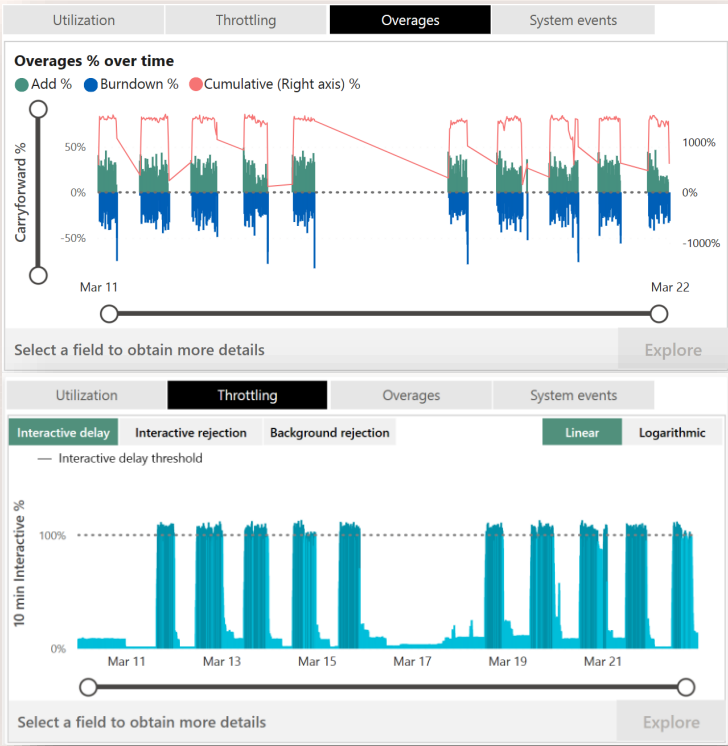
- Throttling is the platform policy for managing consumption that exceeds throughput is provided by SKU choice
- When workloads exceed the throughput of a capacity a cumulative debt is tracked to be burned down
- Cumulative debt is used to determine throttling policies and is burned down when resources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

Overage Operation	Description
 Overages - Added	<ul style="list-style-type: none">• Timepoint when job requests exceed the throughput of a capacity, overages are 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	<ul style="list-style-type: none">• Overages being reconciled when future capacity is free to burn down
 Overages - Cumulative	<ul style="list-style-type: none">• 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



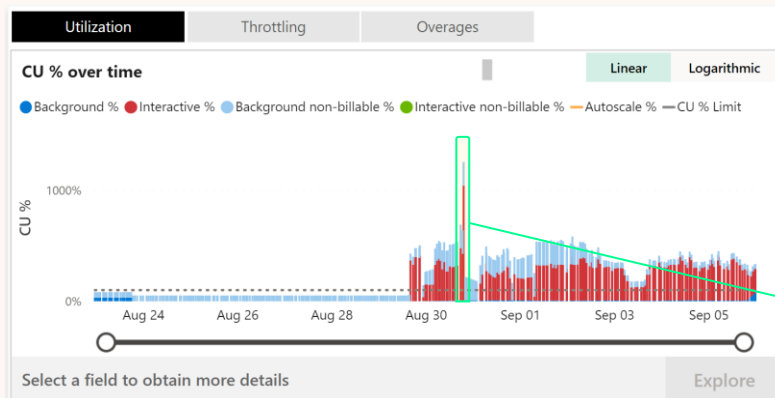
Smoothed Capacity - Future Use	Platform Policy	Customer Impact
$\leq 10m$	Overage Protection	Jobs can consume 10 minutes of future capacity use without throttling
$> 10m \rightarrow \leq 60m$	Interactive Delay	User requested interactive type jobs will be throttled
$> 60m \rightarrow \leq 24h$	Interactive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



Capacity Planning with Capacity Metrics

Capacity planning case study - measurement

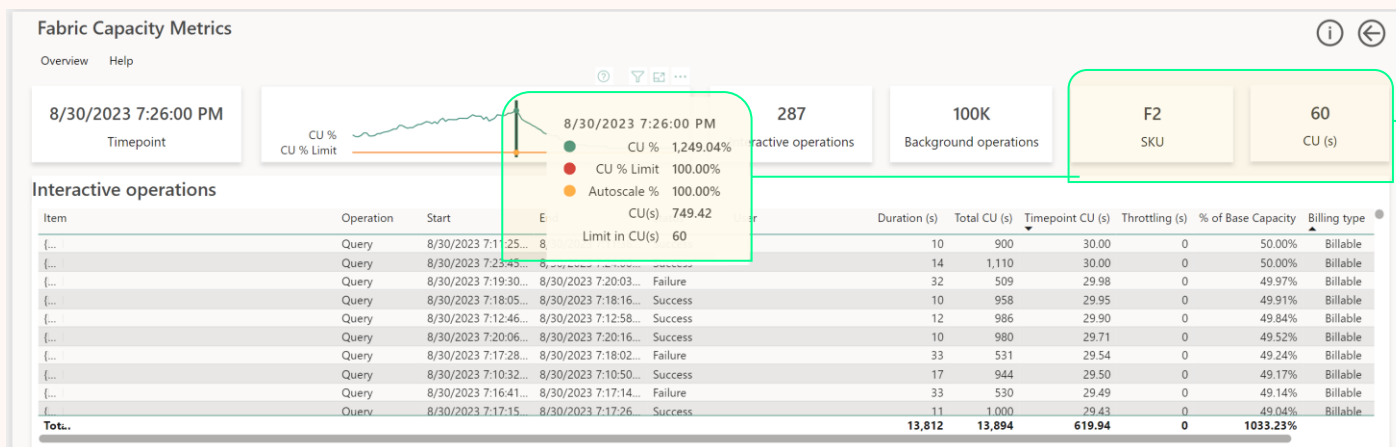
Start with a test or trial capacity to evaluate the load of specific Fabric Experiences i.e., Power BI Datasets, Spark Notebooks or a Datawarehouse



TimePoint	8/30/2023 7:26:00 PM
Background non-billable %	211.49%
Autoscale %	100.00%
Total CU Usage %	1250.53%
Total CU(s)	750.32
Interactive CU(s)	620.18
Background CU(s)	3.25
Background Preview CU(s)	126.90
100% in CU(s)	60

Right-click to drill through

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)

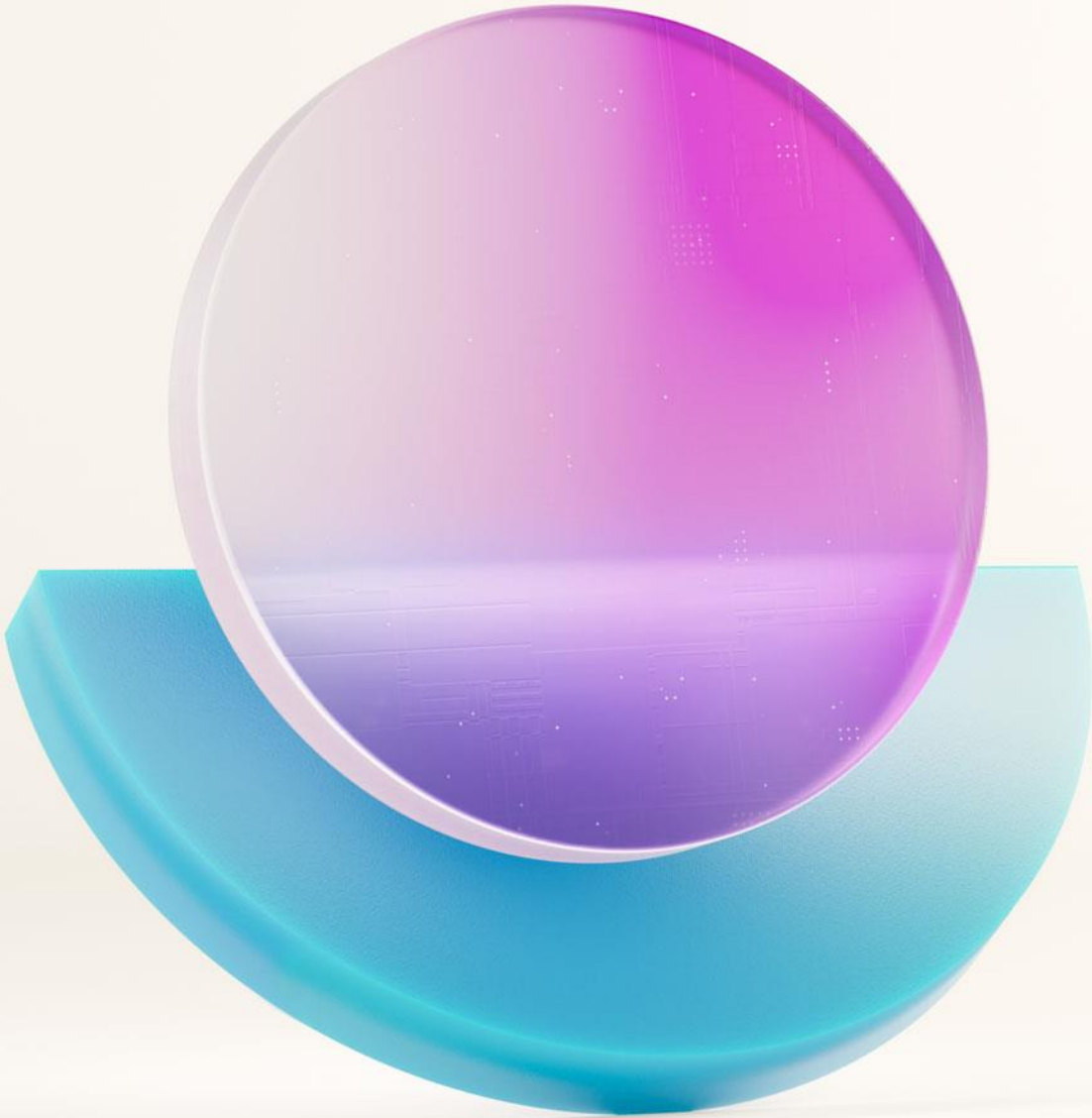
Capacity planning case study – SKU selection

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

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

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) [?]

Number of daily batch cycles [?]

Number of tables across all data sources [?]

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-...

Recommended All Products

Region : East US Billing frequency : Select a value Reset filters

1-2 of 2 Recommendations based on 30 day usage [Learn more](#)

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

Add to cart Close

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)

Active capacity is 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

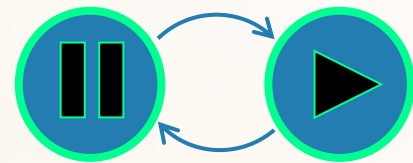
Example 6 – Multiple SKUs matching reservation amount through multiple reservations

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



Pausing and Resuming Capacities



Pausing and Resuming Capacities

Why pause capacities?

- 1) It helps manage compute costs.
- 2) It clears any debt that has accumulated. Use it to quickly resolve throttling.

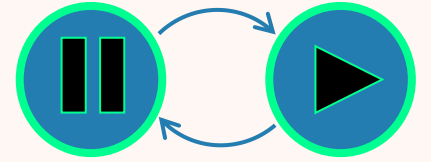
What does it do?

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

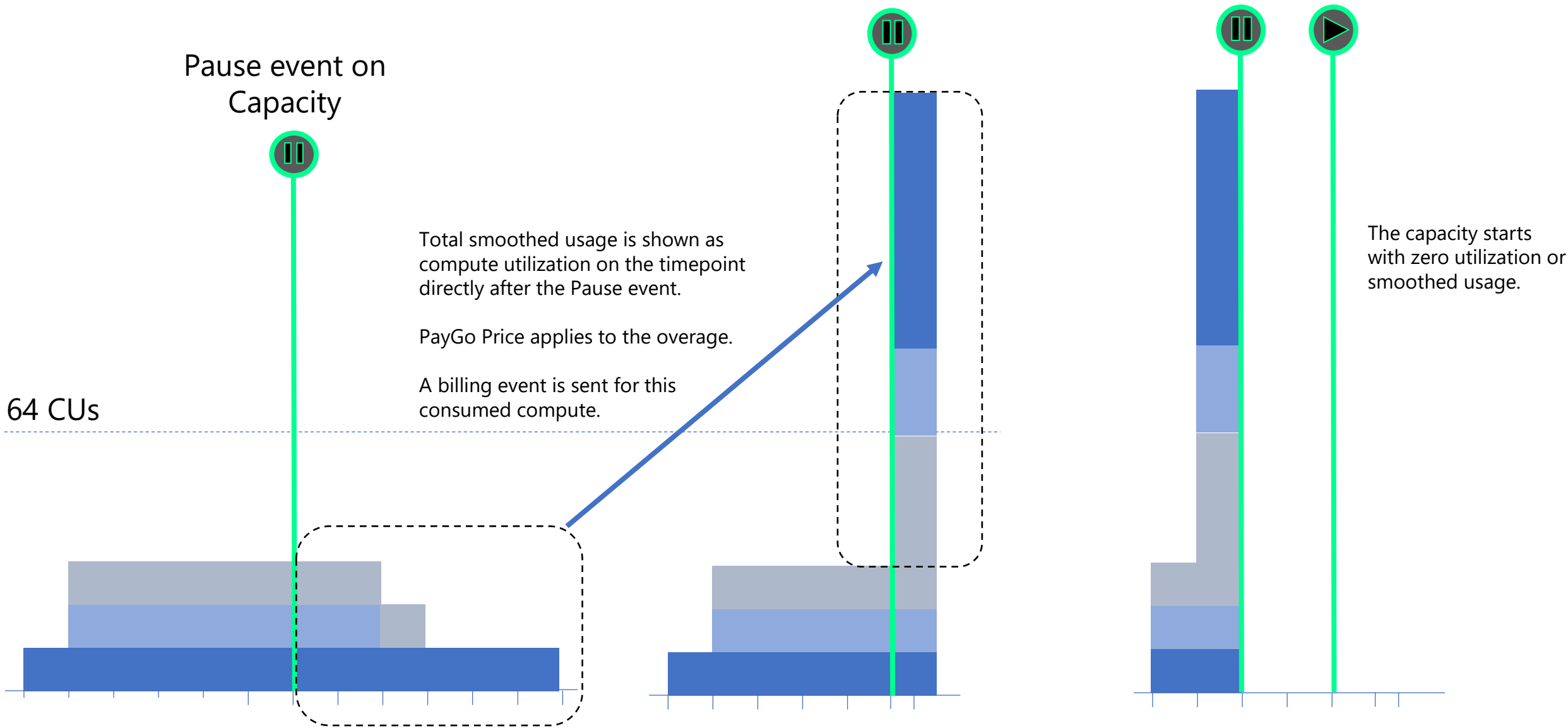


How Capacity Pause & Resume works

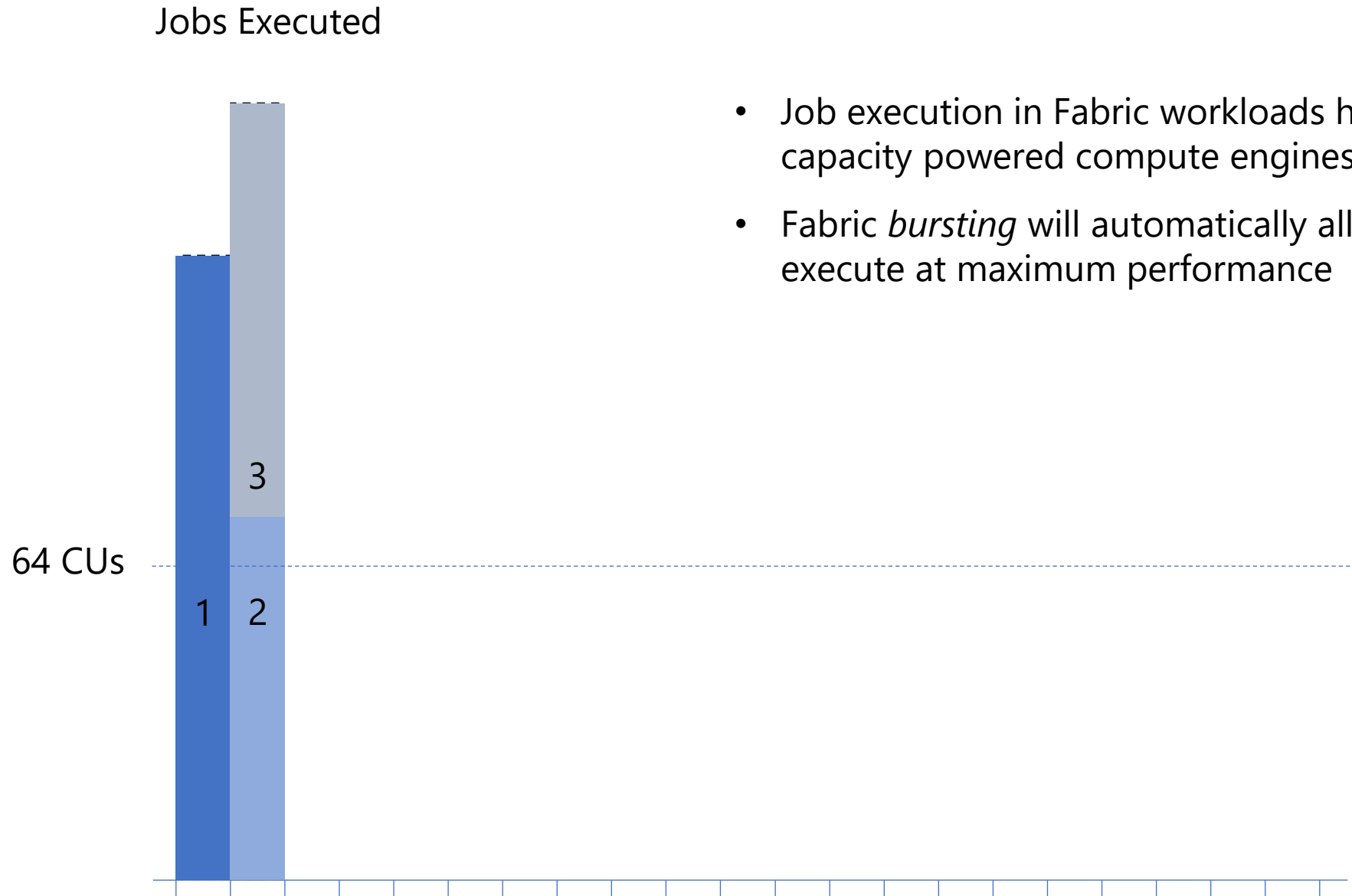
When a capacity is **paused**...

Smoothed usage is **reconciled**.

Later, it can be **resumed**.



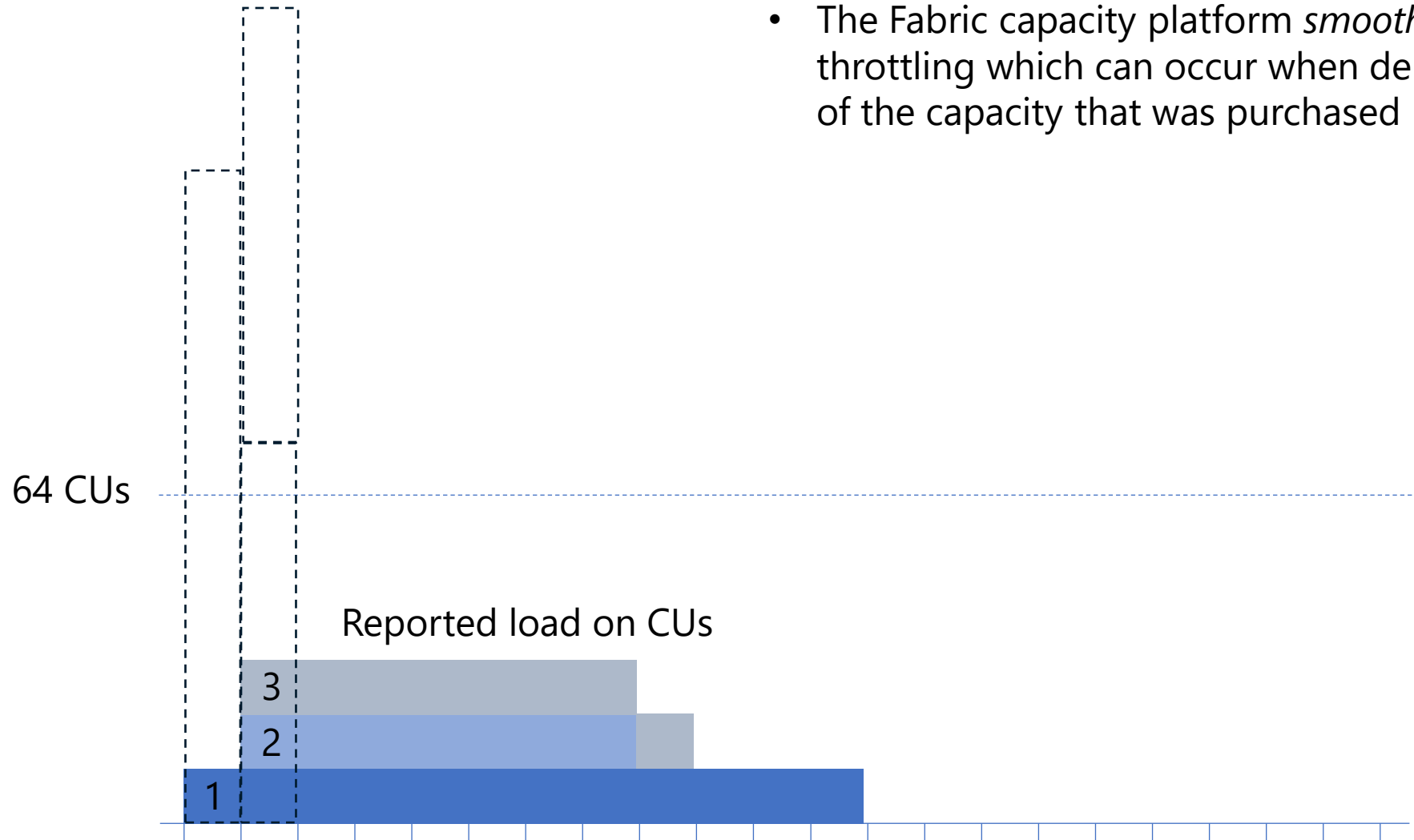
Bursting and Smoothing



- 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

Actual
execution



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

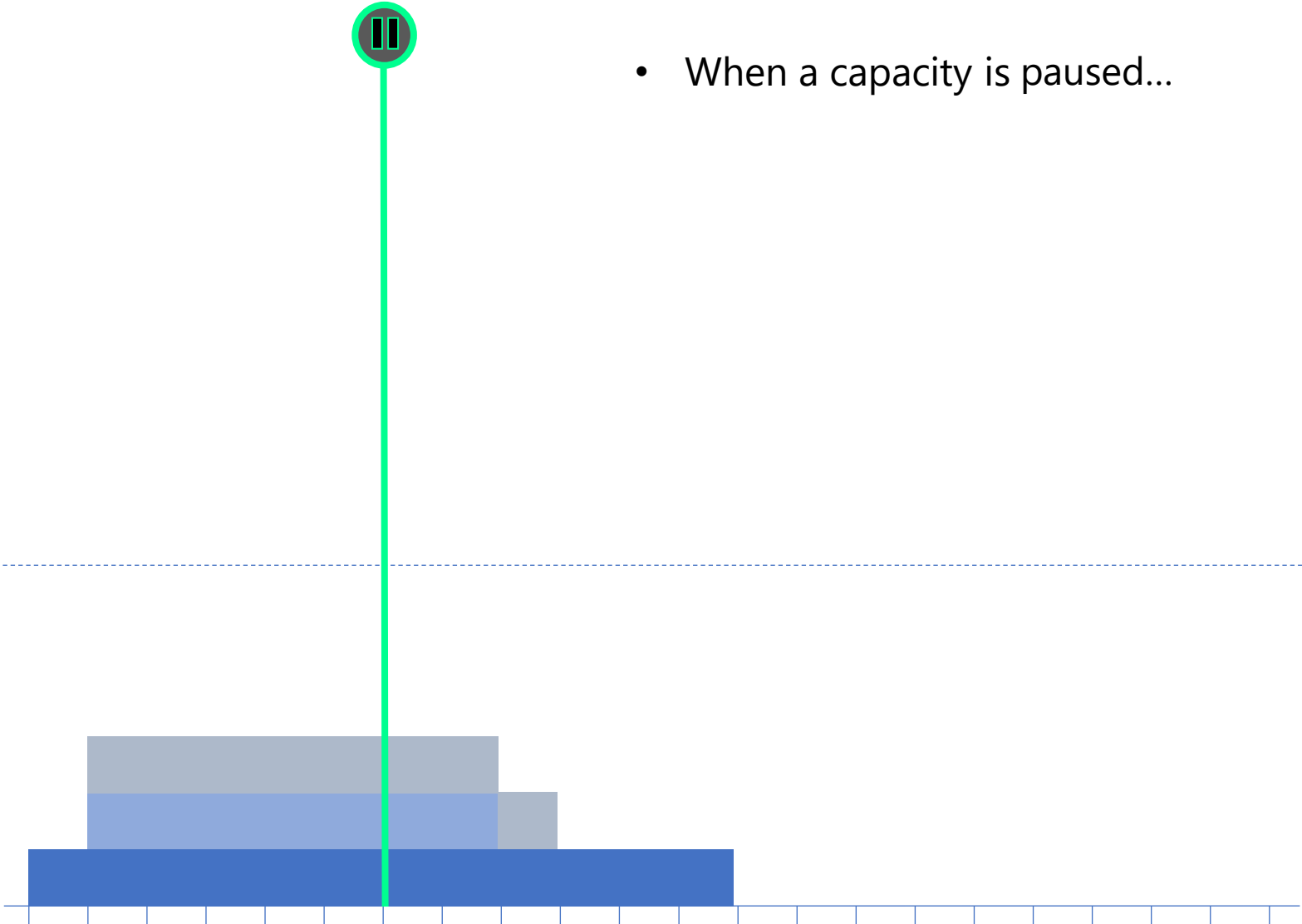
Smoothing and Paused Capacities

Pause event on
Capacity



- When a capacity is paused...

64 CUs

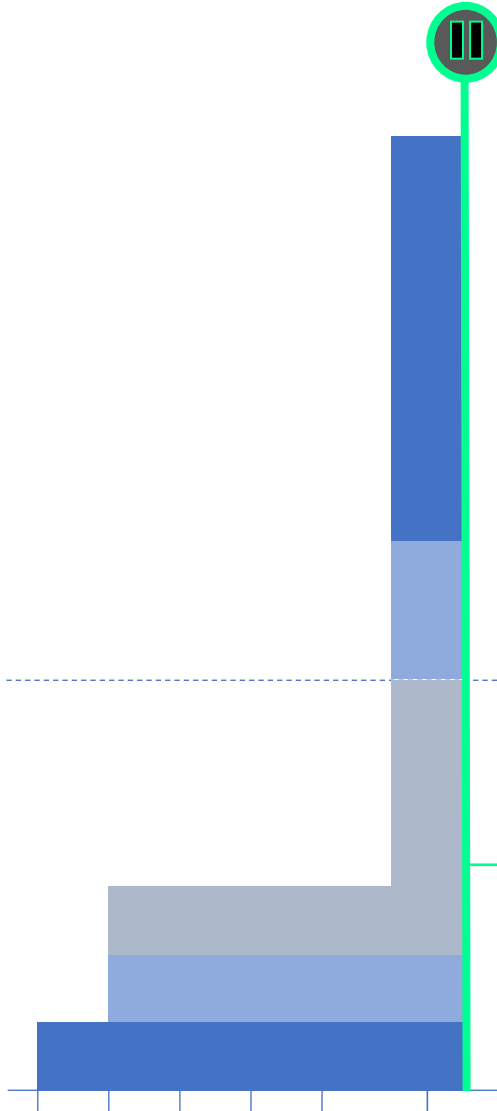


Smoothing and Paused Capacities

Pause event on
Capacity



64 CUs



- 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

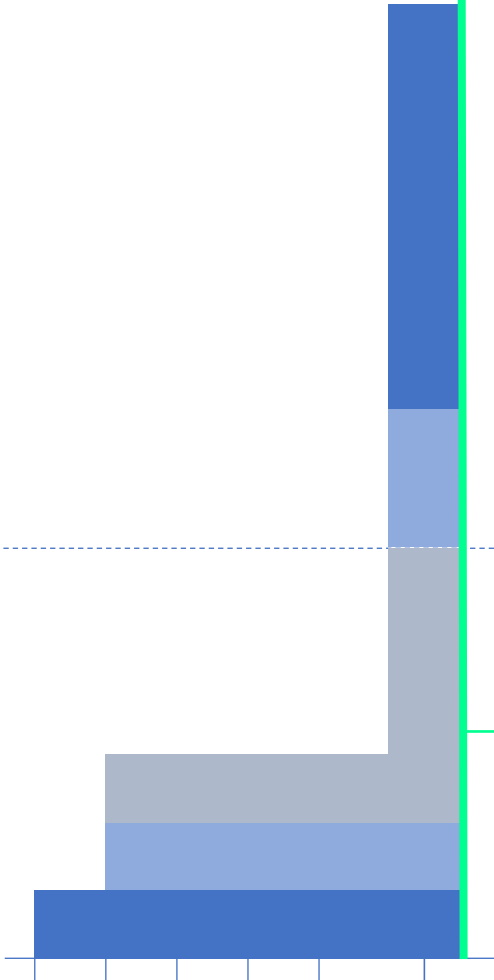


Smoothing and Paused Capacities

Pause event on
Capacity



64 CUs



- 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

Utilization	Throttling	Overages	System events
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	
12/13/2023 9:35:11 AM	Suspended	ManuallyPaused	
Select a field to obtain more details			Explore

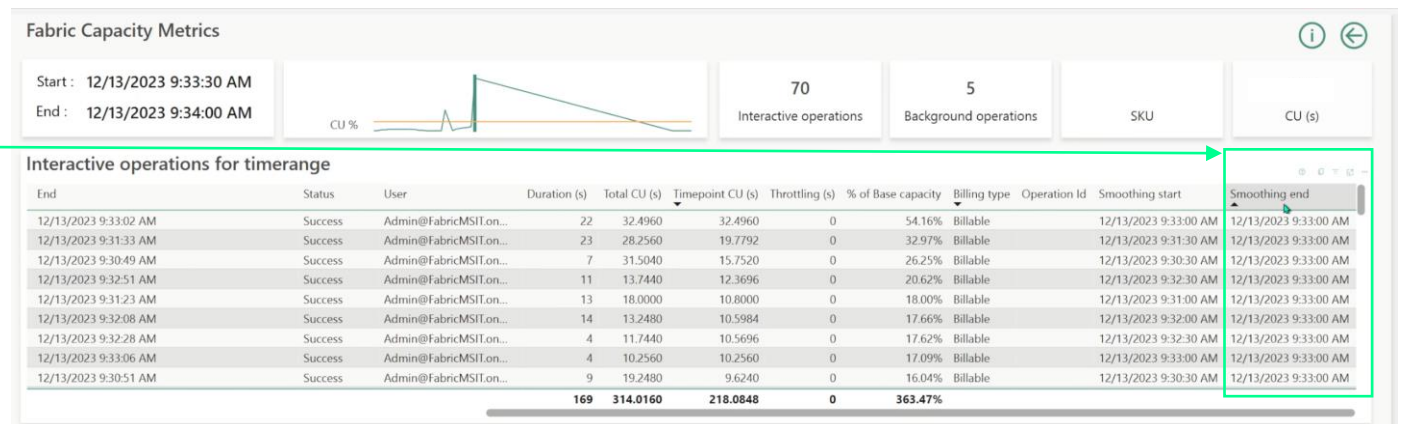
Smoothing and Paused Capacities

Pause event on
Capacity



64 CUs

- 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



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

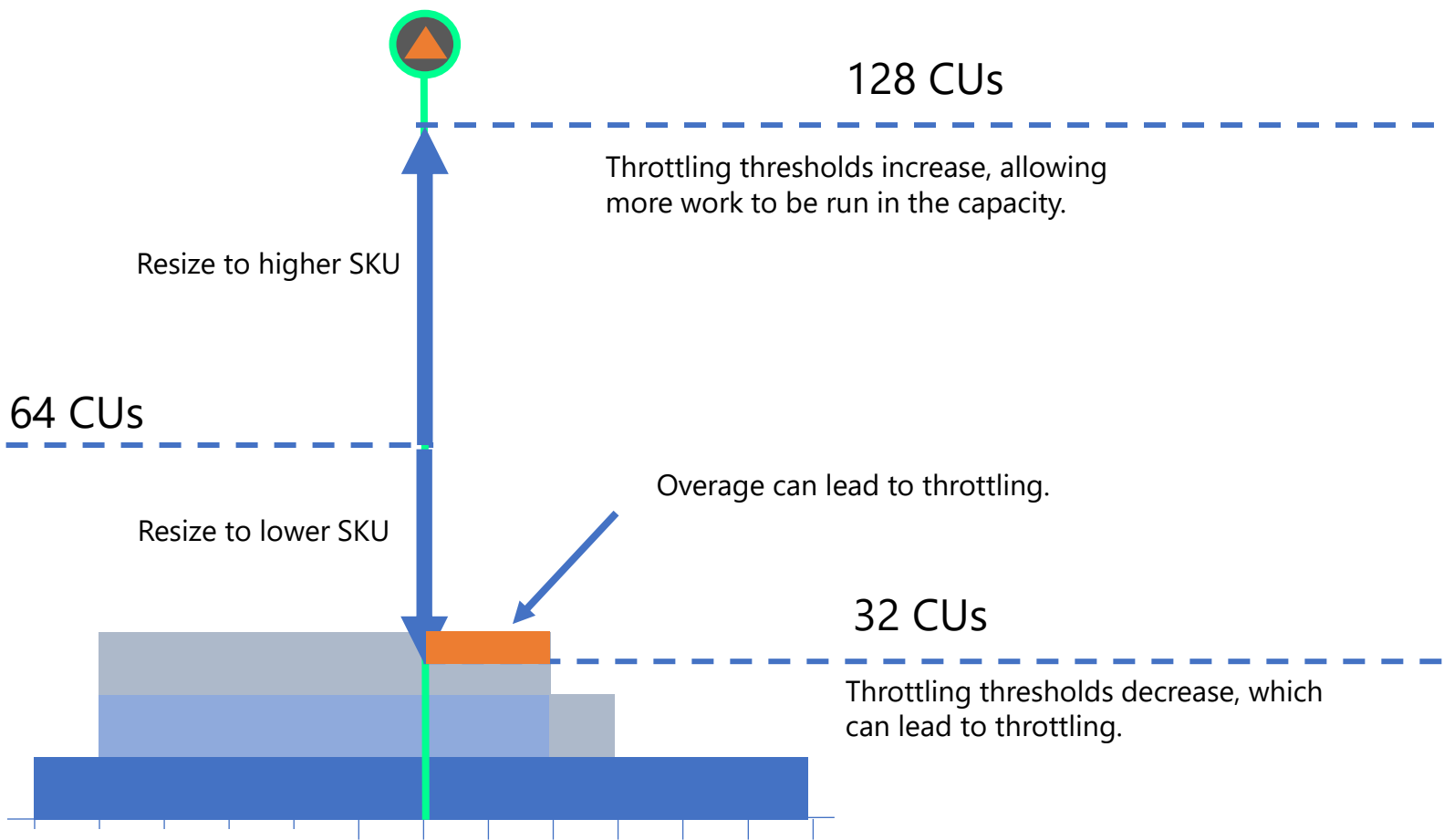
How Capacity Resize works

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



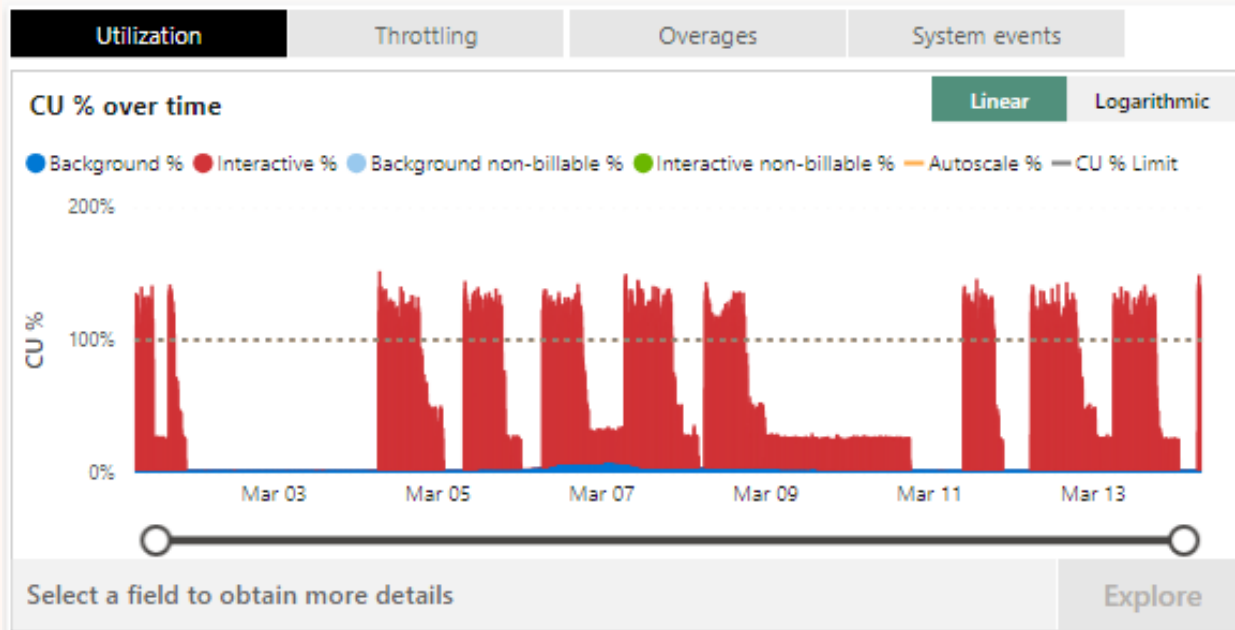
Key Insights

- 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.

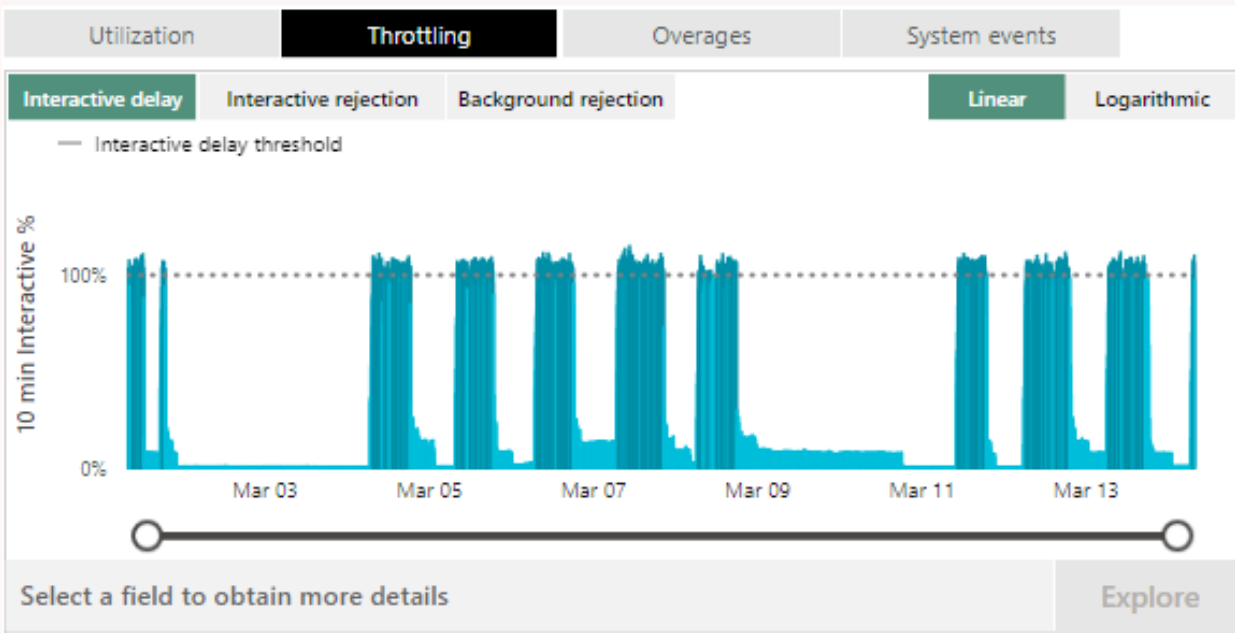


Bonus: Tips and Tricks
for capacity
management and
monitoring

My capacity is being throttled! What can I do?



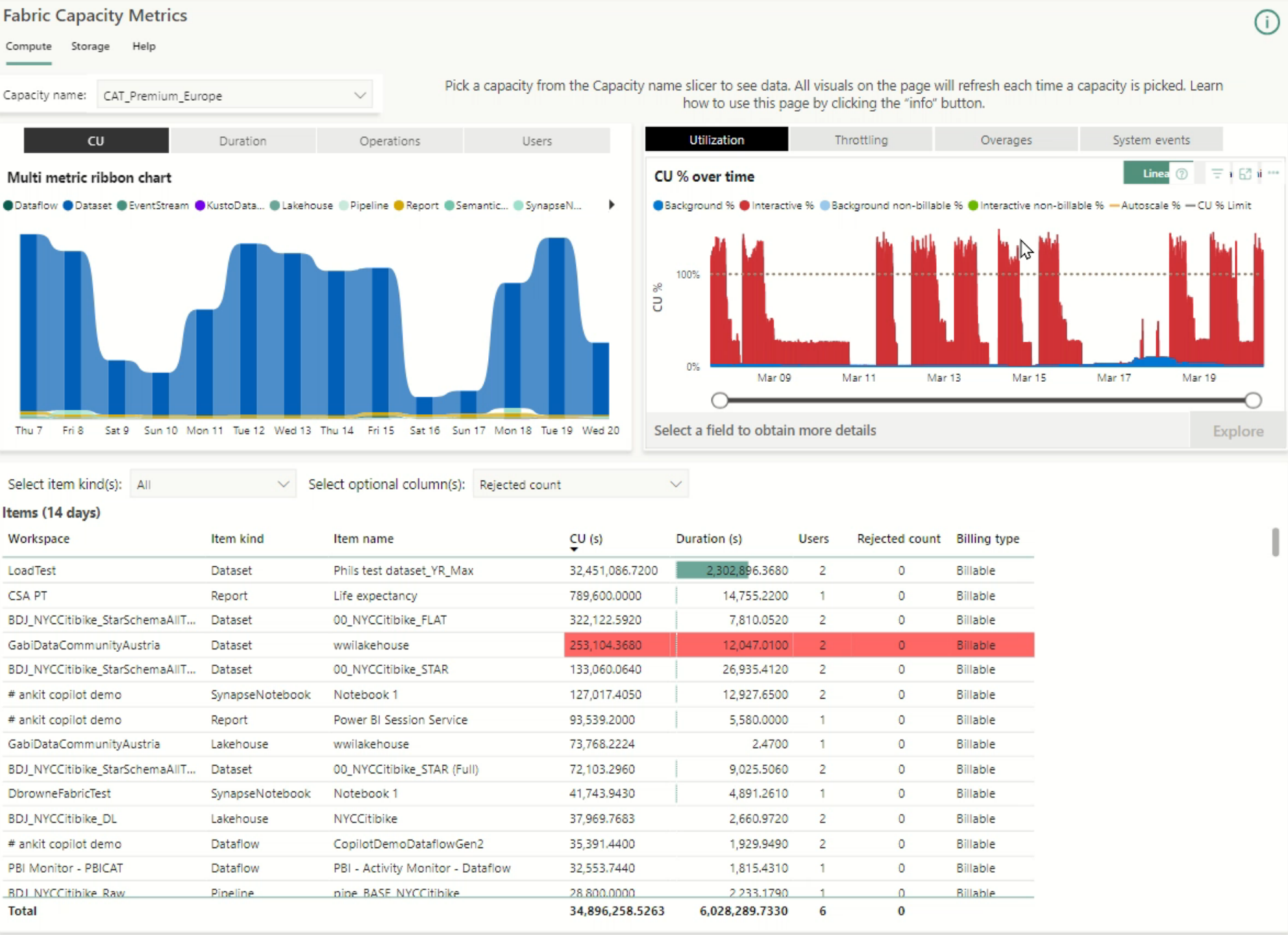
Over 100% utilization doesn't always result in throttling



No penalty until you hit 100% on one of the throttling tabs

Note: For F SKU, if throttled, you can pause/resume to pay now and clear the carry forward, but that is not a long-term solution

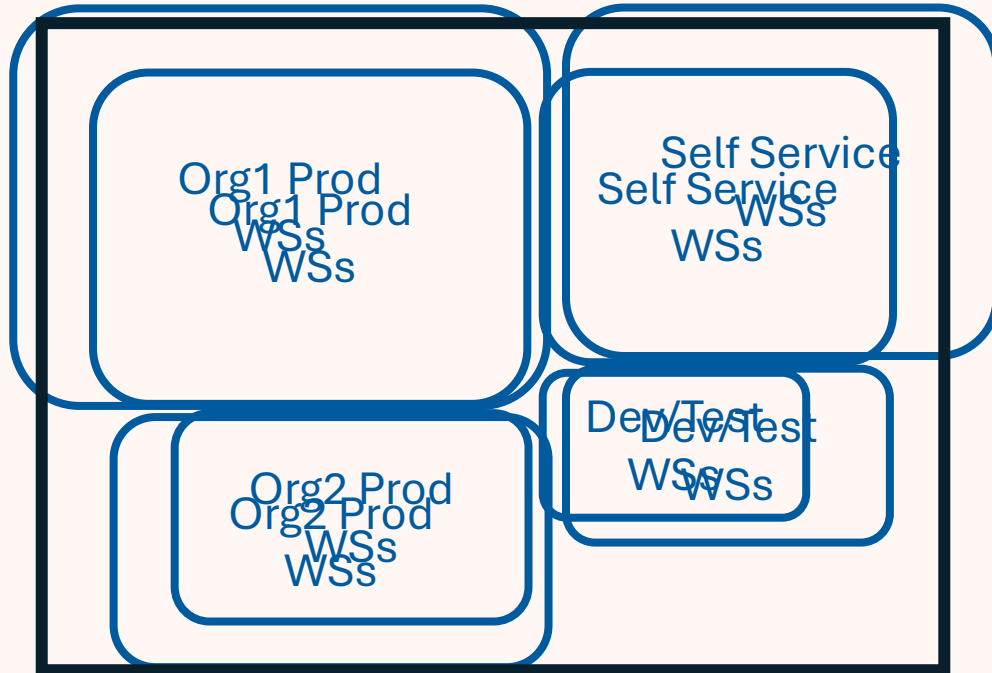
Metrics App Tips



Note: Didn't even go to Timepoint Detail page. Useful for usernames and individual operation time/CU

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

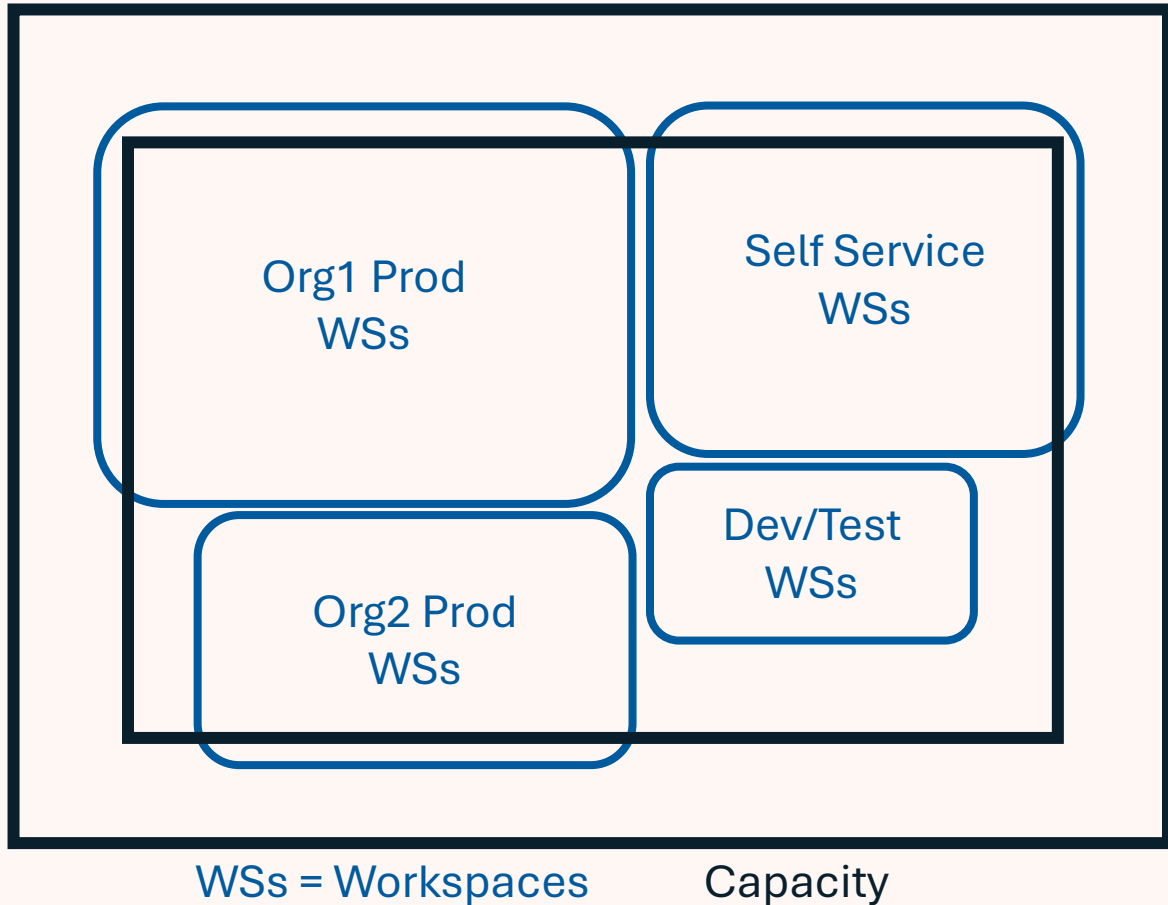
- Avoids increased cost
- Learning carries over to future content

Cons

- Can be difficult/time consuming

When Capacity Units Run Out

Option 2 – Scale Up



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

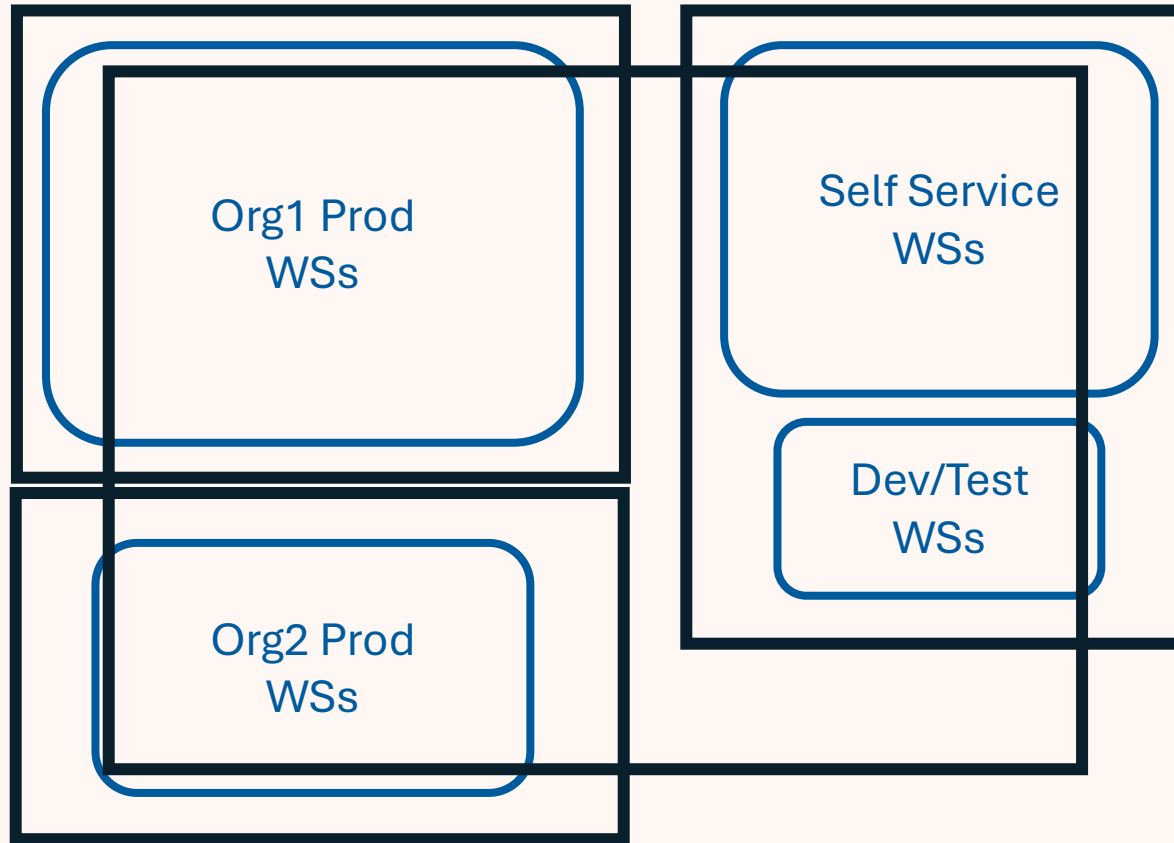
- Add CUs for all items
- Easy

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

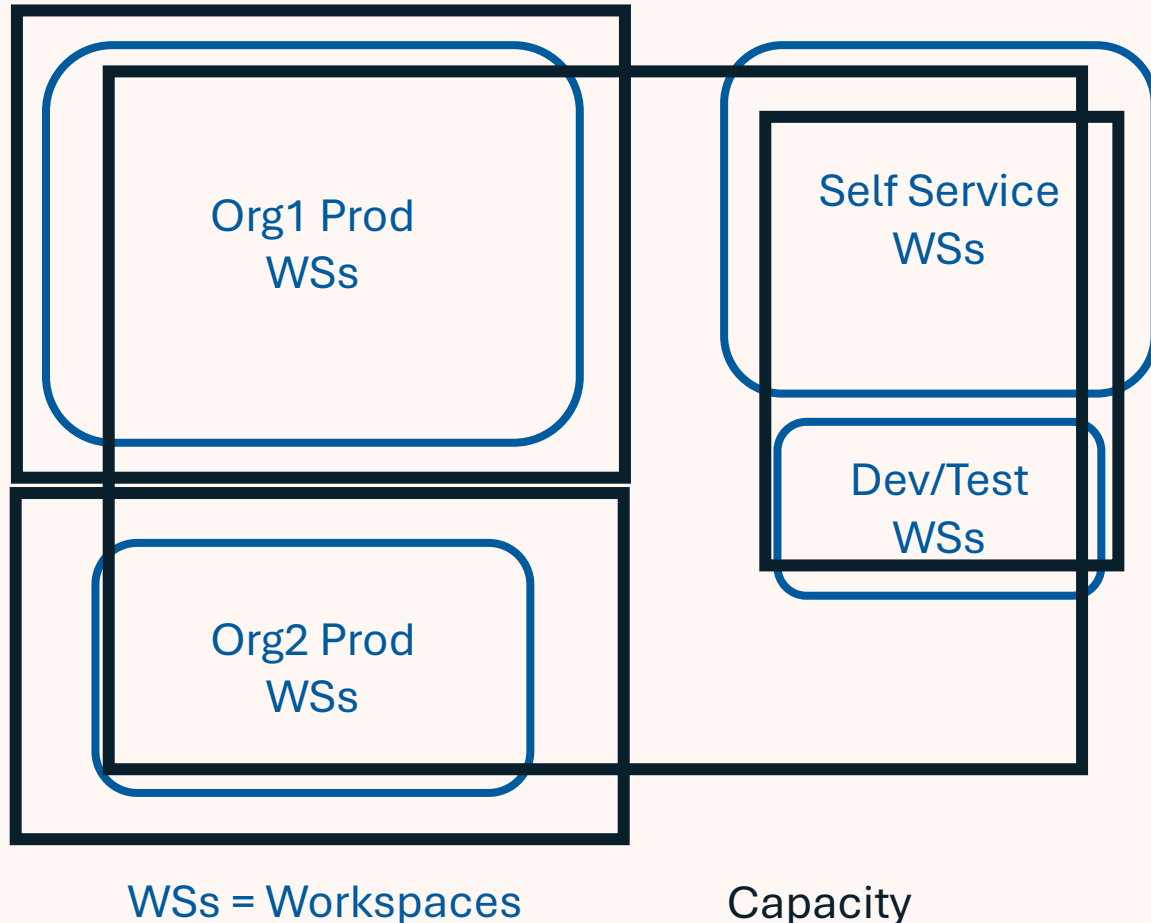
- Easy
- Provides some isolation from bad actors (items with unintentionally high CU burn)
- Flexibility in capacity settings/governance

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

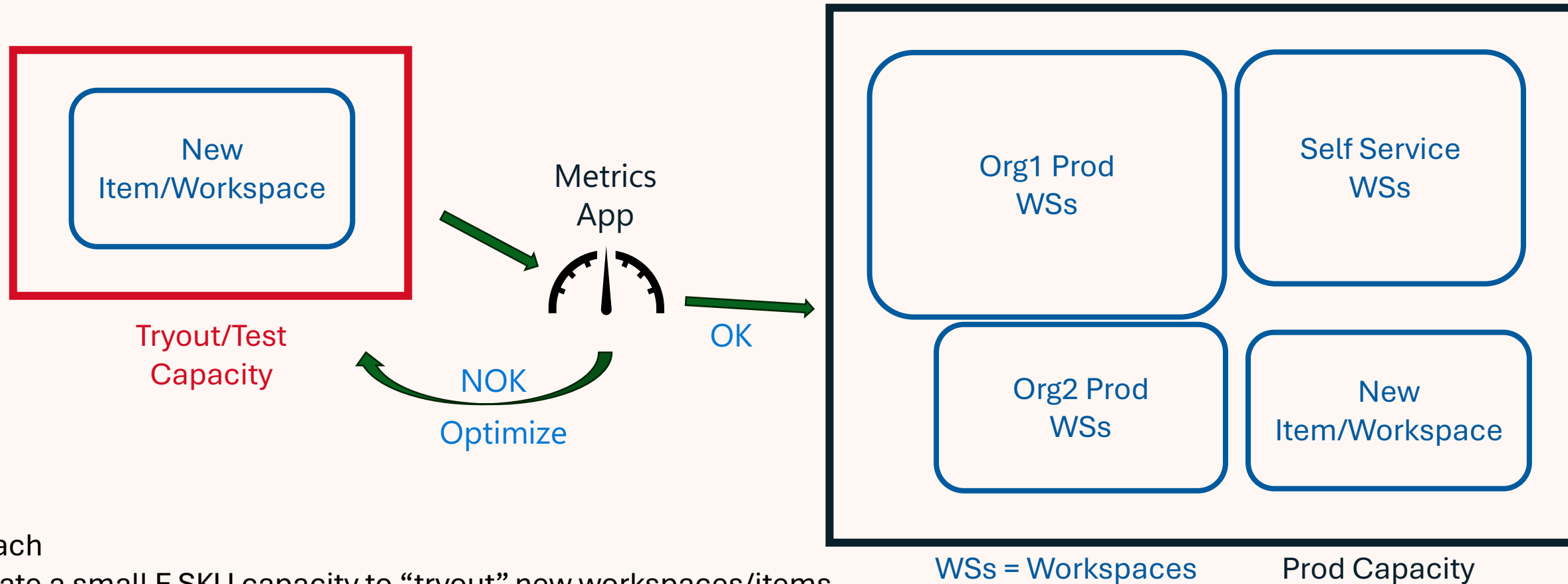
Pros

- 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



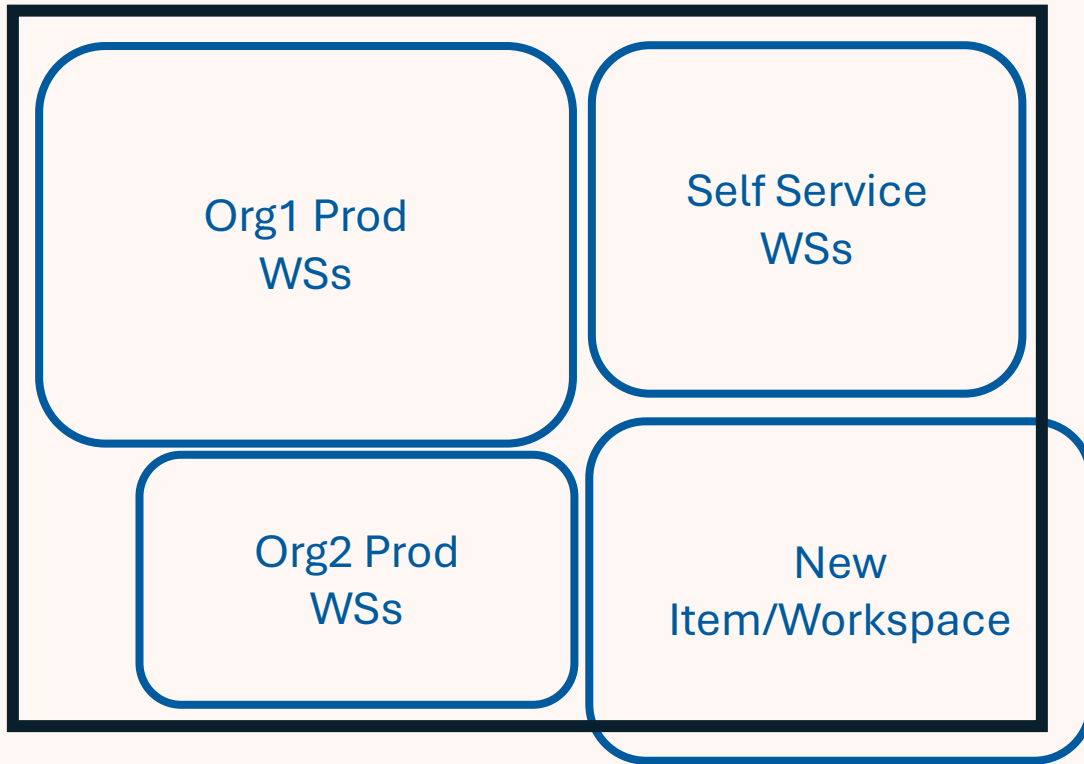
Approach

- 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

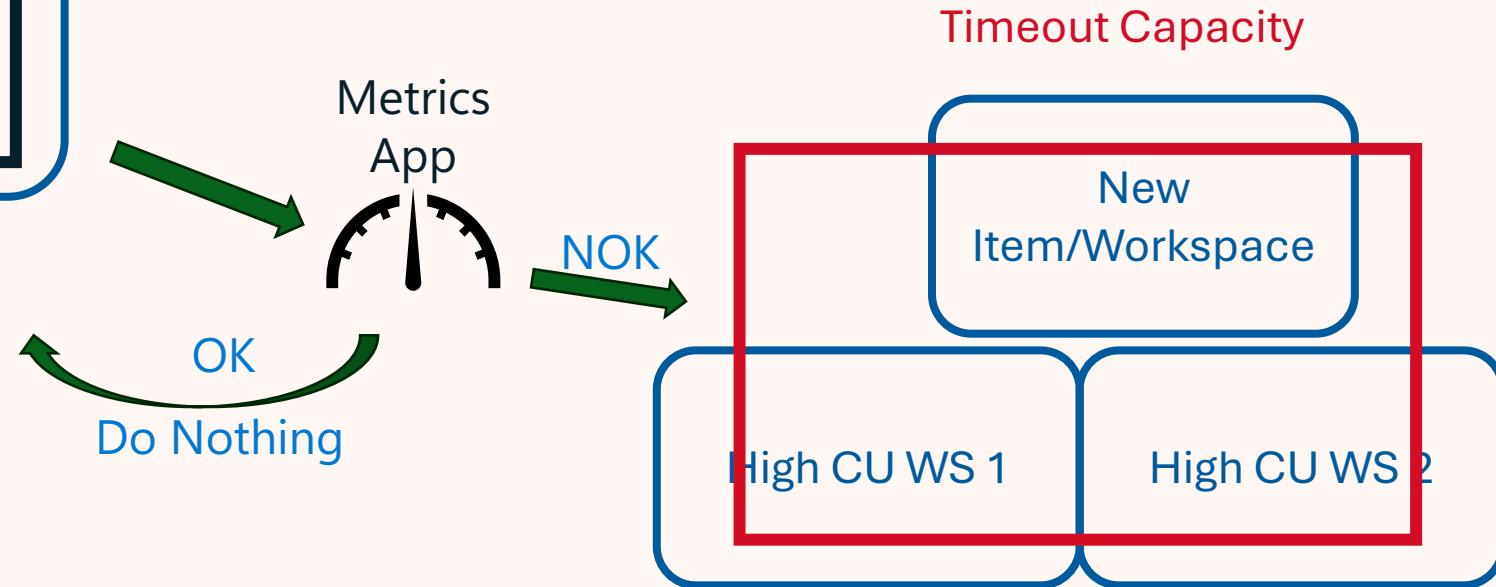
Approach

- 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



WSs = Workspaces

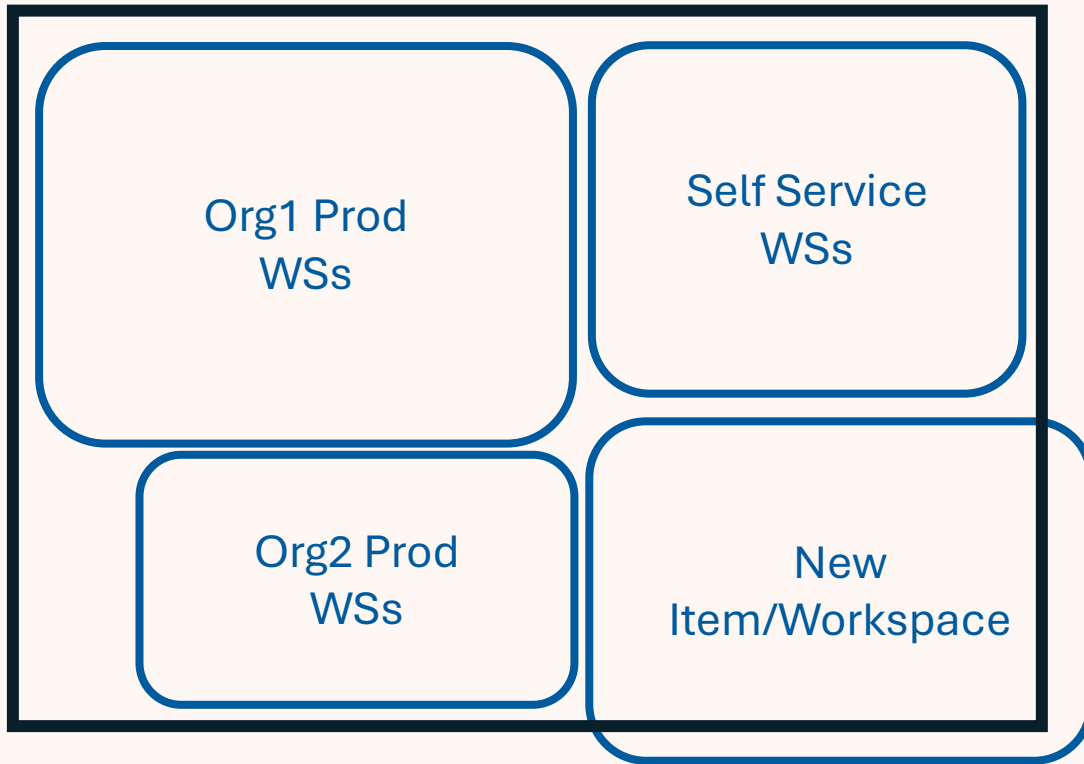
Prod Capacity



Isolation Strategy #4c – Rescue Capacity

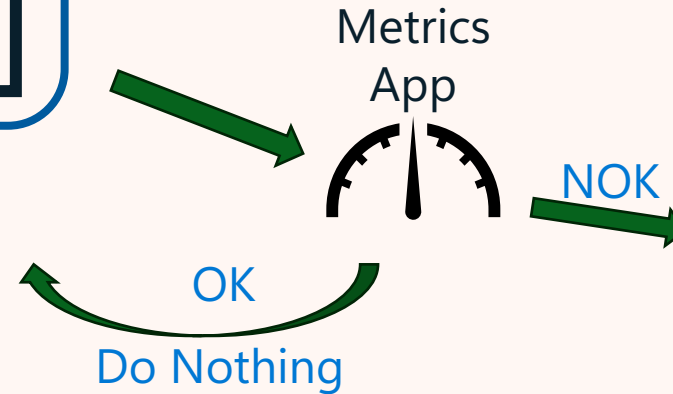
Approach

- 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



WSs = Workspaces

Prod Capacity



Rescue Capacity

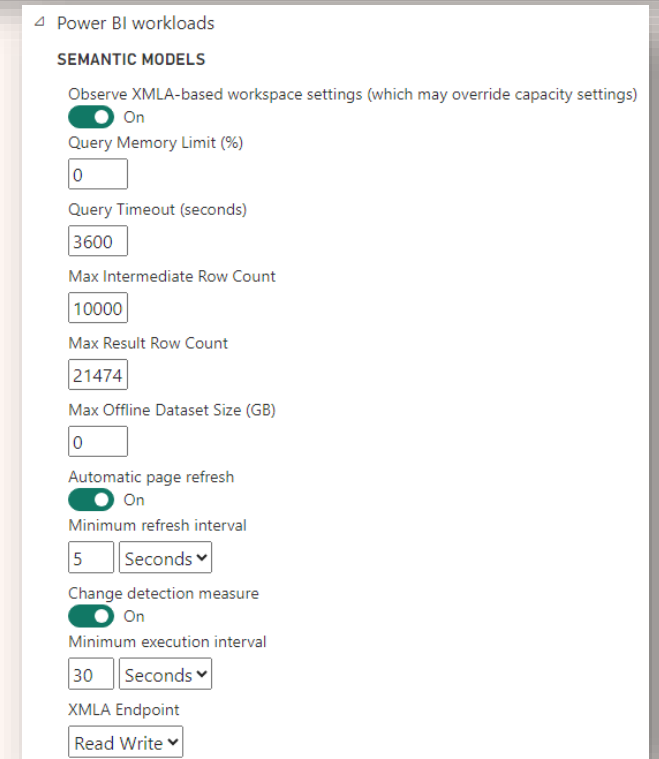
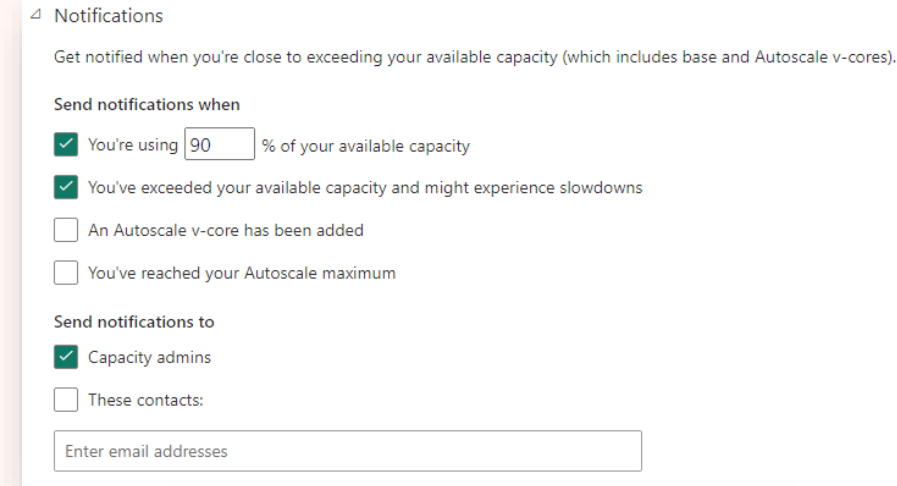
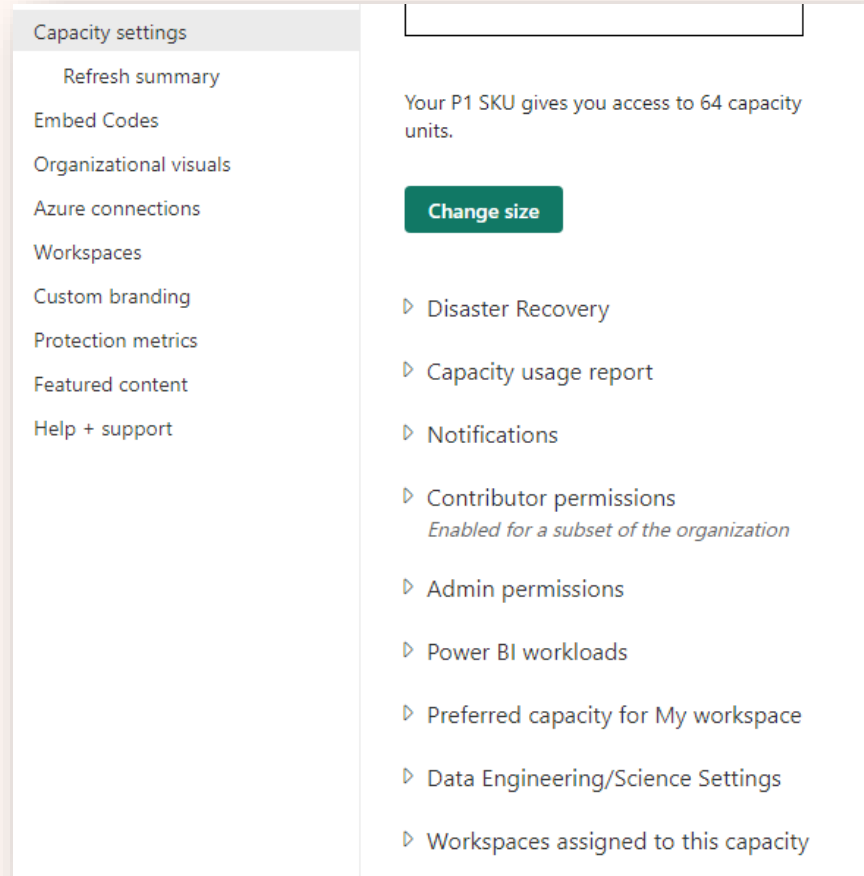


Recommendations for Cost/CU Savings

- Invest in education, knowledge/best practice sharing, COE, etc. for creators and consumers (proactive optimization)
- Avoid data/report sprawl (leverage certified/promoted models, OneLake shortcuts, etc.)
- Leverage a multi-capacity strategy (isolate, tryout, timeout, etc.)
- Right size your capacities and leverage F SKUs for pause/resume/resize, or reserved instances for discounts
 - Consider a combo of RI and PAYGO (for predictable surge activity)
- Choose the right tool for the job and stay up to date on Fabric feature releases
 - High concurrency mode for notebooks

Leverage the capacity settings in the UI

- Notifications on CU overuse
- Power BI workloads settings (e.g., query limits, page refresh)



DIY Autoscale – Fabric Notebook

(Bret Myers)

Set SKU Ranges and Values

```
1 # Parameters to be passed in from pipeline.
2 minSku = 'F2' # min sku size we can scale down to
3 maxSku = 'F128' # max sku size we can scale up to
4 utilizationTolerance = 90 # Percentage of CU used to st
5 capacityName = 'fabricbamdemo' #capacity name to be mon
6 subscriptionId = '...'
7 metricsAppWorkspaceName = 'WS_FabricCapacityMetrics' #
8 metricsAppModelName = 'Fabric Capacity Metrics' # name
9 alertEmail = '' # email address to send alert that we s
```

Get credentials

```
1 tenantId = mssparkutils.credentials.getSecret('keyVaultEndpoint', 'secretName_tenantId')
2 clientId = mssparkutils.credentials.getSecret('keyVaultEndpoint', 'secretName_clientId')
3 secret = mssparkutils.credentials.getSecret('keyVaultEndpoint', 'secretName_clientSecret')
4
5 api_pbi = 'https://analysis.windows.net/powerbi/api/.default'
6 api_azuremgmt = 'https://management.core.windows.net/.default'
```

Not all code shown

FabricTools/CapacityAutoScale at main · bretamyers/FabricTools · GitHub

Query metrics app model

```
1 from azure.identity import ClientSecretCredential
2 import requests, json, math
3 from pyspark.sql.functions import explode
4
5 auth = ClientSecretCredential(tenant_id=tenantId, client_id=clientId, client_secret=secret)
6 access_token = auth.get_token(api_pbi)
7
8 header = {'Authorization': f'Bearer {access_token.token}', 'Content-type': 'application/json'}
9
10 body = {
11     "queries": [
12         {
13             "query": f"""
14                 DEFINE
15                     MPARAMETER 'CapacityID' = "{capacityId}"
16
17                     VAR __DS0FilterTable =
18                         FILTER(
19                             KEEPFILTERS(VALUE('TimePoints'[TimePoint])),
20                             'TimePoints'[TimePoint] >= NOW() - 1
21                         )
22
23                     VAR __DS0FilterTable2 = TREATAS({{"{capacityId}"}}, 'Capacities'[capacityId])
24
25                     VAR __DS0Core =
26                         SELECTCOLUMNS(
27                             KEEPFILTERS(
28                                 FILTER(
29                                     KEEPFILTERS(
30                                         SUMMARIZECOLUMNS(
31                                             'Capacities'[capacityId],
32                                             'Time'f'BillableTime'
```

Change SKU Size

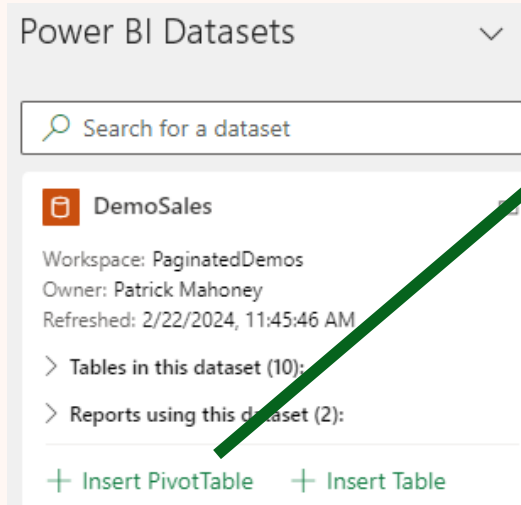
```
19 url = f'https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{res
20
21 body = {
22     "sku": {
23         "name": f"{scaleSku}",
24         "tier": "Fabric"
25     }
26 }
27
28 response = requests.patch(url, headers=header, data=json.dumps(body))
```

Most Common Capacity Issues (Power BI)

Bad Practice	Recommendations/Typical Resolution
Model issues (M:M, bi-di, snowflake, etc.) and/or inefficient DAX	Follow best practices (e.g., BPA), star schema
Too many visuals	Multi card, small multiples, Deneb, PowerPoint background, etc.
Big single visual (i.e., matrix with lots of rows, columns, and/or measures)	Improve report design (e.g., drillthrough, apply all Slicers, report page tooltip), field parameters, calc group guardrails, etc.
Complex RLS	Remodel to enable simple filter like Table[Email] = USERPRINCIPALNAME()
Very high concurrency	Optimize reports, DAX, etc. (big multiplier) Consider QSO, data subsets
Direct Query	Switch to import or Direct Lake, if possible. Aggregations, hybrid tables, etc.
Analyze in Excel	Automate downstream analytics with a Power BI report instead, subscriptions, DAX connected table, slicers/measures first, etc.
Excessive refresh	Don't "break the fold", incremental refresh, reduce frequency, optimize M code

Save Those CUs – Getting Data Into Excel

Analyze in Excel



Connected Table

Key Takeaways

- How you build it matters
 - Filters & measures first!
- This shows durations but it's CU that matters (test your use cases/models)
- Opt for DAX Connected Tables
 - Create pivot table from that, if needed

✗ Rows, Measures, Filter

StartTime	Type	Duration	User	Database	Query
11:49:30	MDX	2,328ms	Power BI...	DemoSales	SELECT {[Measures].[To
11:49:26	MDX	0ms	Power BI...	DemoSales	SELECT {AddCalculated
11:49:23	MDX	0ms	Power BI...	DemoSales	SELECT {AddCalculated
11:49:17	MDX	1,875ms	Power BI...	DemoSales	SELECT {[Measures].[To
11:49:03	MDX	4,469ms	Power BI...	DemoSales	SELECT {[Measures].[To
11:48:54	MDX	3,938ms	Power BI...	DemoSales	SELECT {[Measures].[To

✓ Filter, measures, rows

StartTime	Type	Duration	User	Database	Query
10:06:13	MDX	1,625ms	Power BI...	DemoSales	SELECT {[Measu
10:06:03	MDX	781ms	Power BI...	DemoSales	SELECT {[Measu
10:05:49	MDX	109ms	Power BI...	DemoSales	SELECT {[Measu
10:05:46	MDX	312ms	Power BI...	DemoSales	SELECT {[Measu
10:05:43	MDX	234ms	Power BI...	DemoSales	SELECT FROM [M
10:05:14	MDX	0ms	Power BI...	DemoSales	SELECT {AddCal

Refresh (same for both)

StartTime	Type	Duration	User	Database	Query
11:50:30	MDX	2,234ms	Power BI...	DemoSales	SELECT {[Measures].[To

✗ Rows, Measure, Filter

StartTime	Type	Duration	User	Database	Query
01:28:50	DAX	31ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:41	DAX	1,516ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:40	DAX	16ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:34	DAX	156ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:33	DAX	16ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:31	DAX	0ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:30	DAX	141ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:15	DAX	2,047ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:11	DAX	1,797ms	Power BI...	DemoSales	DEFINE VAR __C
01:28:08	DAX	594ms	Power BI...	DemoSales	DEFINE VAR __C
01:27:56	DAX	281ms	Power BI...	DemoSales	DEFINE VAR __C
01:27:50	DAX	16ms	Power BI...	DemoSales	DEFINE VAR __C

✓ Filter, measures, rows

StartTime	Type	Duration	User	Database	Query
09:14:20	DAX	16ms	Power BI...	DemoSales	DEFINE VAR __DS0Filt
09:14:07	DAX	1,000ms	Power BI...	DemoSales	DEFINE VAR __DS0Filt
09:14:02	DAX	1,188ms	Power BI...	DemoSales	DEFINE VAR __DS0Filt
09:13:59	DAX	594ms	Power BI...	DemoSales	DEFINE VAR __DS0Filt
09:13:51	DAX	531ms	Power BI...	DemoSales	DEFINE VAR __DS0Filt
09:13:50	DAX	0ms	Power BI...	DemoSales	DEFINE VAR __DS0Cor

Refresh (same for both)

StartTime	Type	Duration	User	Database	Query
11:54:49	DAX	1,969ms	Power BI...	DemoSales	DEFINE VAR __DS0FilterTable = TREATA



Slides



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