More for less: Cost optimizing your Azure SQL databases

Aditya Badramraju, **Arvind** Shyamsundar

Principal Product Managers, Microsoft Azure SQL DB

sqlbits ***2024***

About us...

Arvind Shyamsundar

Principal Product Manager, Microsoft

LinkedIn: **ArvindSh**

@arvisam (Bluesky, Mastodon and X)

Aditya Badramraju

Principal Product Manager, Microsoft

LinkedIn: **AdityaBadramRaju**

@aditya_feb22 (X)

About you…

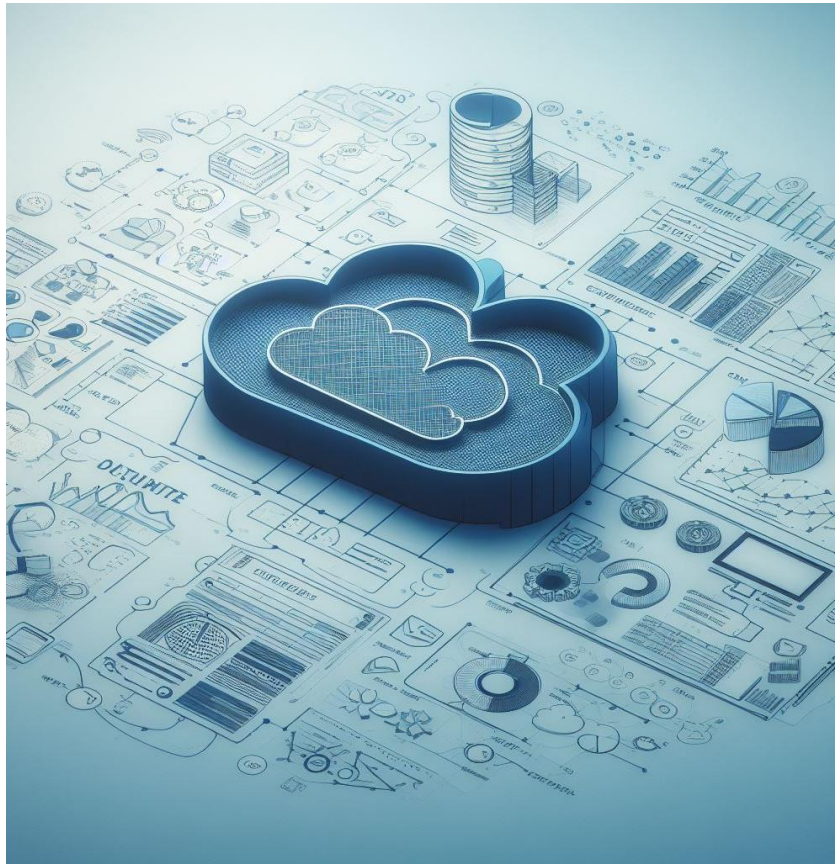
Are you actively using Azure SQL Database?

Are you responsible for cost management?

Any conceptual questions you'd like us to address today?

A **Blueprint** for cost optimization

- **Categorize** your databases based on criticality, scale and performance needs.
- **Choose the right** Azure SQL DB **service tier** (“edition”) to fit the database category.
- **Pick** the **best configuration** that fits your database need.
- **Apply** additional **fine tuning** to the configuration.
- **Avail** relevant **discounts**.



Disclaimer: Service tier and configuration recommendations are for illustration purposes only. All costs shown in this presentation are estimates and based off retail prices published in the Azure SQL Database pricing page and in the Azure Pricing Calculator. Numbers subject to change. Information in this session is provided as-is, where-is, without any warranty of any kind.

#1: Categorize your databases

A typical **categorization** of databases

Tier One Production

Tier Two Production

Tier Three Production

Low tier Production

Dev / Test

Integration
Test / Perf
Test / UAT

Requirements for database categories

Characteristic	Tier 1 Prod	Tier 2 Prod	Tier 3 Prod	Low Tier Prod	Dev/Test	Integration / Perf / UAT
Performance	Highest	High	Medium	Medium/Low	Low	Same as Prod
Usage	24/7	24/7	24/6	12/6	10/5	Infrequent
Database Size	Large	Large	Medium	Small	Tiny	Same as Prod
Vertical Scalability	Needed	Needed	Needed	Minimal need	Not Needed	Needed
Horizontal Scalability	Needed	Needed	Minimal need	Minimal need	Not Needed	Needed
Cost Profile	Highest	High	Medium	Medium / Low	Low	Low
Disruption Tolerance (RTO)	Seconds	Low Minutes	High Minutes	Hours	Hours	Hours
Data Loss Tolerance (RPO)	Near zero	2 Minutes	5 Minutes	10 minute	1 hour	N/A
Zone Redundancy	Yes	Yes	Yes	No	No	No
Geo Replication	Yes	Yes	No	No	No	No

Related session: “Perfecting business continuity for Azure SQL DB” on Saturday @ 10:10AM.

#2: Select the right service tier

Azure SQL Database Services

Infrastructure-as-a-Service

Best for lift and shift and/or workloads requiring OS-level access

SQL Server on Azure VMs

Platform-as-a-Service

Best for modernizing existing apps

Azure SQL managed instance

Best for supporting modern cloud apps

Azure SQL database

Azure SQL Database - options

Service tiers / edition

- Hyperscale
- Business Critical
- General Purpose

Compute tiers

- Provisioned
- Serverless

Deployment models

- Single (standalone) database
- Elastic pool

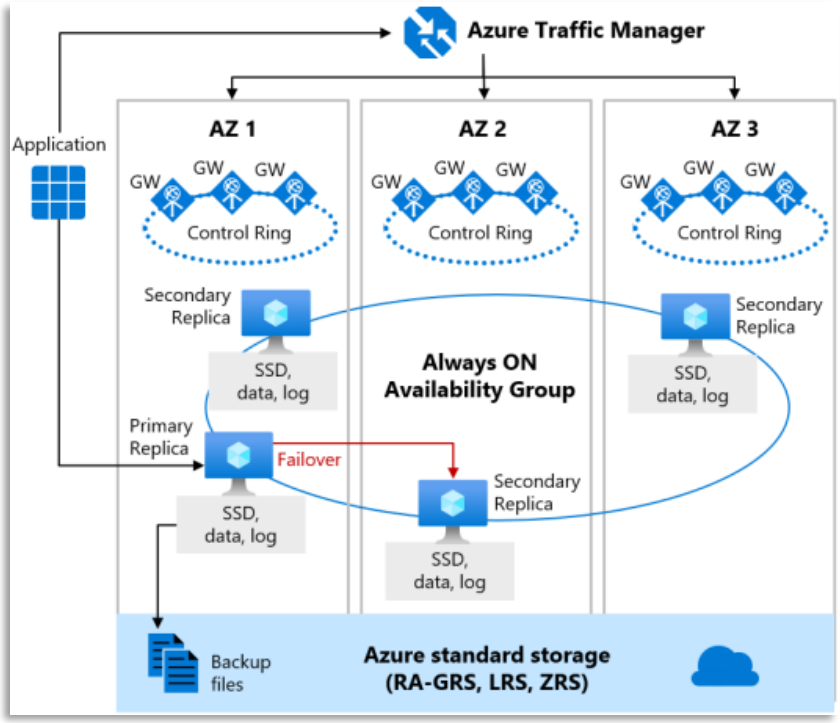
Hardware

- All service tiers: standard-series (“Gen 5”)
- Hyperscale only: premium-series and memory-optimized premium-series

Business Critical and Premium

- High performance with **low commit latency** courtesy of **local SSD storage.**
  - Scaling operations can be “size of data” operations.
- Resilience through 4-way synchronous replication. **Zone redundancy** and **one read replica** is **included** in the price.
- **Highest price** amongst the service tiers / edition.
- Limited to **max 4TB**; max DB size also **limited by vCores**:

vCores	Max db size GB
2-4	1024
6	1536
8-10	2048
12-20	3072
20+	4096

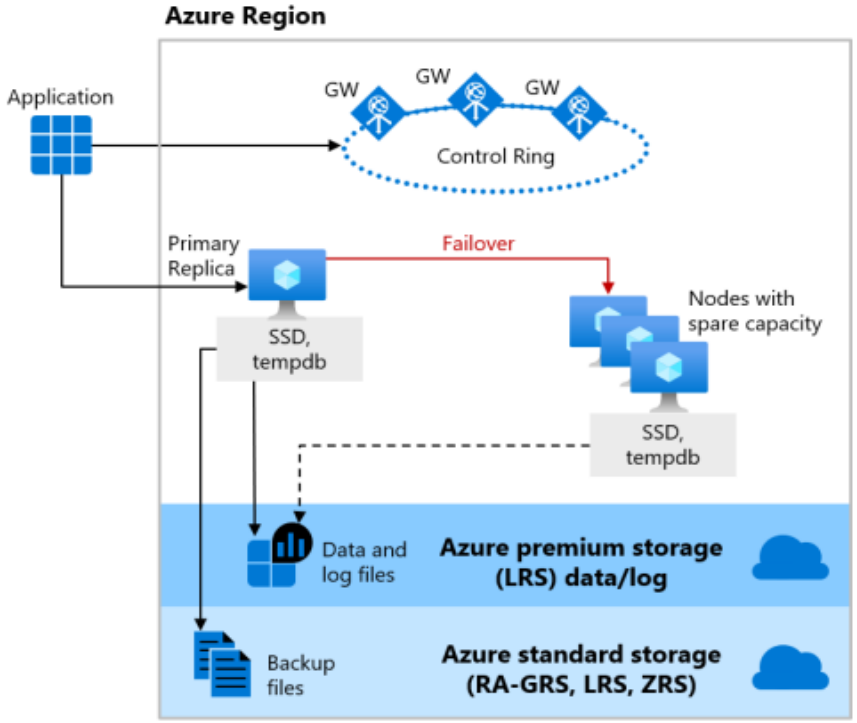


vCORE	Memory (GB)	Pay as you go
2	10.2	$1.359/hour
Storage GB		$0.25/month

General Purpose and Standard

- Uses **multi-node persistent remote storage**.
  - Scaling operations are generally independent of the size of data.
  - Higher latency and lower performance.
- **No read replica**; **zone redundancy** costs extra.
- Lower price than Businesss Critical.
- Limited to **max 4TB**; max DB size also **limited by vCores**:

vCores	Max db size GB
2-4	1024
6	1536
8-10	2048
12-20	3072
24+	4096



vCORE	Memory (GB)	Pay as you go
2	10.2	$0.505/hour
Storage GB		$0.115/month

Hyperscale

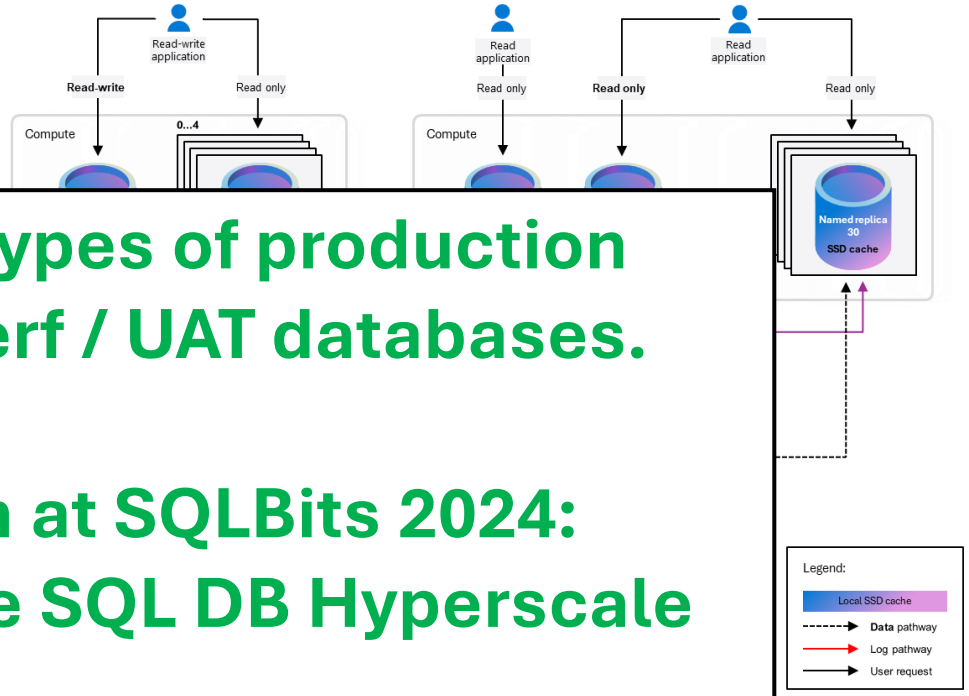
- **Built ground up** based on a **cloud native architecture.**
  - **Independent scalability** of **compute** and **storage**.
- **Z**
- **E**
- **L**
- **U**

Replicas and HA replicas.

- **Predictably quick** scaling, DB copy and restore operations.
- **Memory Optimized** hardware option with 2x RAM / vCore.

Hyperscale is a great option for all types of production databases as well as Integration / Perf / UAT databases.

More details in our related session at SQLBits 2024: "Database of the future is here - Azure SQL DB Hyperscale Deep Dive"



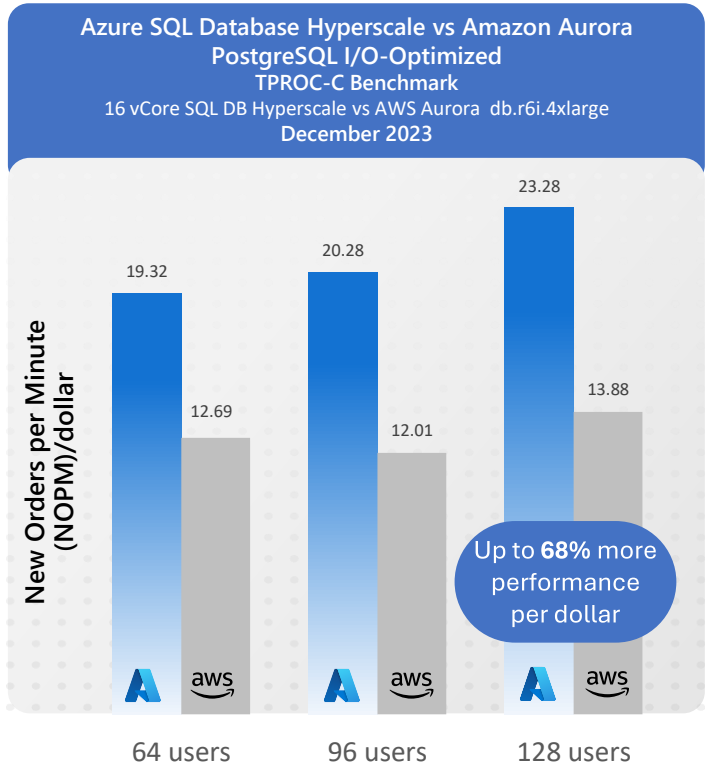
vCORE	Hardware	Memory (GB)	Pay as you go
2	Standard / Premium series	10.2	$0.366/hour
2	Premium series, memory-optimized	20.8	$0.512/hour
Storage GB			$0.25/month

Azure SQL Database Hyperscale outperforms

Principled Technologies published a study where they tested throughput performance between **Azure SQL Database Hyperscale** and **Amazon Aurora PostgreSQL**.

SQL Database **Hyperscale** emerged as the **price-performance leader** for mission-critical workloads, delivering up to 68 percent more performance per dollar than Amazon Aurora PostgreSQL.[1]

Blog: https://devblogs.microsoft.com/azure-sql/build-highly-scalable-ai-ready-applications-on-azure-sql-database-hyperscale/

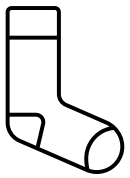


Database performance per dollar comparison.
Higher is better.

Price-performance claims based on data from a study commissioned by Microsoft and conducted by Principled Technologies in December 2023. The study compared performance and price performance between a 16 vCore and 32 vCore Azure SQL Database using premium-series hardware on the Hyperscale service tier and the db.r6i.4xlarge and db.r6i.8xlarge offerings for Amazon Web Services Aurora PostgreSQL I/O-Optimized (AWS Aurora). Benchmark data is taken from a Principled Technologies report which used the HammerDB TPROC-C benchmark. The TPROC-C workload is derived from the TPC-C Benchmark and results were obtained with the HammerDB TPROC-C workload . The HammerDB TPROC-C workload is derived from the TPC-C benchmark and is not comparable to published TPC-C Benchmark results, as this implementation does not comply with all requirements of the TPC Benchmark. Price-performance is calculated by Principled Technologies as the cost of running the cloud platform continuously divided by new orders per minute throughput, based upon the standard. Prices are based on publicly available US pricing in East US 1 for Azure SQL Database and US East for AWS Aurora as of December 2023. Performance and price-performance results are based upon the configurations detailed in the Principled Technologies report. Actual results and prices may vary based on configuration and region.

Quiz time!

What is the difference in compute costs between Hyperscale standard-series and Hyperscale premium-series?



#3: Pick the right configuration

Azure SQL Database - options

Service tiers / edition

- Hyperscale
- Business Critical
- General Purpose

Compute tiers

- Provisioned
- Serverless

Deployment models

- Single (standalone) database
- Elastic pool

Hardware

- All service tiers: standard-series (“Gen 5”)
- Hyperscale only: premium-series and memory-optimized premium-series

Price-perf optimized solution using serverless

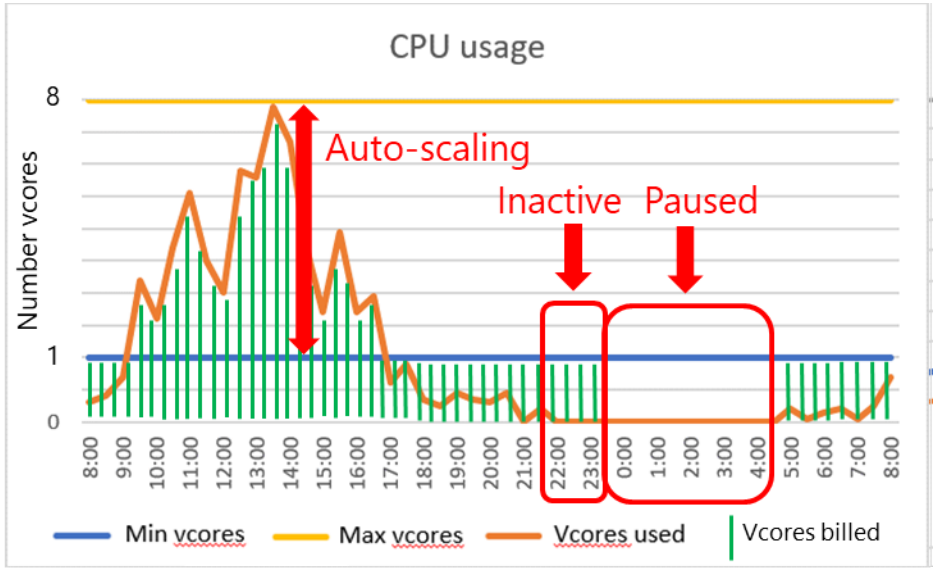
Scaling

Auto-scale

- Min-max vCores (and memory)

Auto-pause/resume

- Auto-pause delay (inactive period)



Billing

Pay for compute used per second

- Billing based on vCores and memory used
- Only storage billed when paused

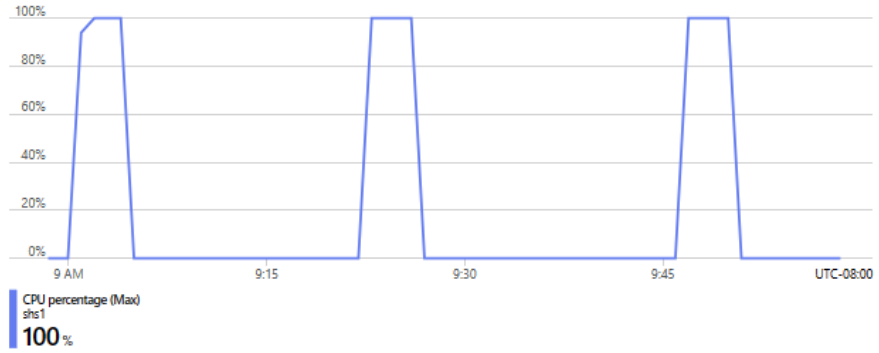
http://aka.ms/sqldb_serverless

Microsoft Azure

Usage patterns

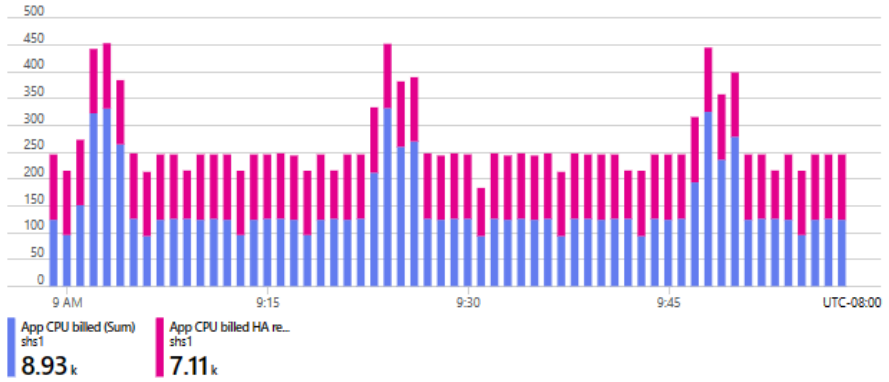
Typical serverless patterns

- New databases **without usage history.**
- **Intermittent**, **unpredictable usage** and lower average compute utilization.
- **Frequently rescaled** databases.



Serverless *anti*-patterns

- More **regular, predictable usage** and higher average compute utilization.
- Databases **sensitive to performance** trade-offs from scaling.
- **Multiple databases** that can be consolidated into elastic pools.



Azure Pricing calculator https://azure.microsoft.com/en-us/pricing/calculator/

Azure SQL DB Hyperscale serverless

GA

Automatic scaling of compute and memory; log throughput independent of compute

Auto-scaling independence of the primary replica, high availability replicas, and named replicas

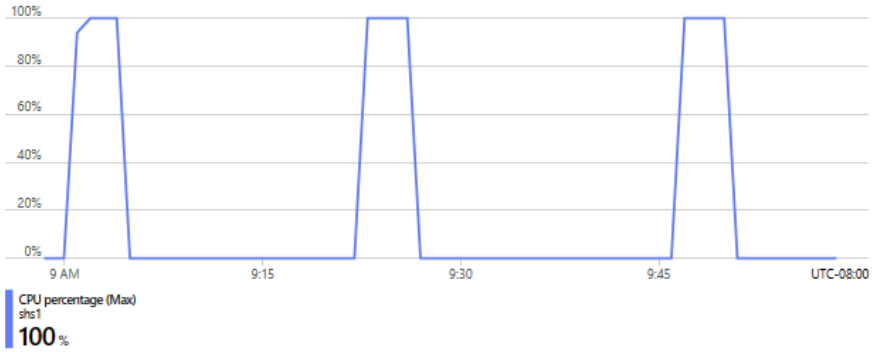
Automatic scaling of database storage up to 100 TB

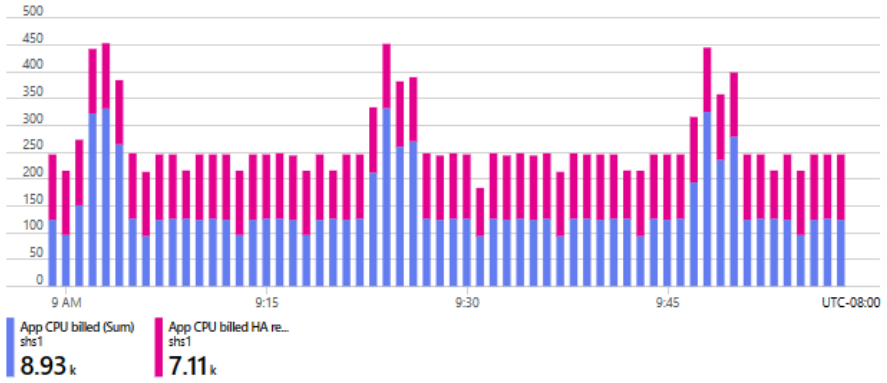
Auto-scaling independence of CPU and memory to match workload demand

Per-Second Billing of used resources

Hyperscale serverless compute cost example

Serverless Hyperscale database





Primary replica

Compute tier	**Serverless**	**Provisioned**
Max vcores configured	16	16
Vcore seconds billed	8.93k	Independent of usage
Compute unit price [1]	$0.000105/vcore-second	$0.183/vcore-hour
Compute cost for 1 hour [2]	**$.94**	**$2.93**

[1] Prices for the primary replica are the same price as for HA replicas.

HA replica

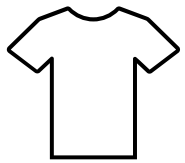
Compute tier	**Serverless**	**Provisioned**
Max vcores configured	16	16
Vcore seconds billed	7.11k	Independent of usage
Compute unit price	$0.000105/vcore-second	$0.183/vcore-hour
Compute cost for 1 hour [2]	**$0.75**	**$2.93**

[2] Intermittent, bursty usage pattern is important to achieve cost savings using serverless.

Microsoft Azure

Quiz time!

Cricket fans – imagine you are **developing** a website to sell tickets for The Ashes (match series), what compute tier would you pick?



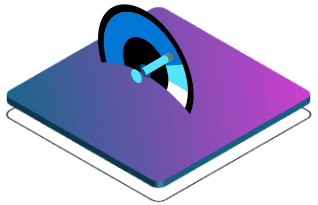
Try Azure SQL Database Free of Charge

Get 100,000 vCore seconds of serverless compute and 32 GB of storage every month!

Azure SQL Database with Serverless Compute

Flexible compute automatically scales to meet demand.

No Time Limits



Apply this free offer for the life of your subscription.

Need more? No problem.

Stick with the default auto-pause option or continue usage for additional charges.

What's included:

One Azure SQL Database with serverless compute per Azure Subscription with 100,000 vCore seconds every month.

32 GB data storage +

32 GB backup storage.

Learn More: aka.ms/sqlfreeoffer

Cost optimization, for a group of DBs

Azure SQL Database - options

Service tiers / editions

- Hyperscale
- Business Critical
- General Purpose

Compute tiers

- Provisioned
- Serverless

Deployment models

- Single (standalone) database
- Elastic pool

Hardware

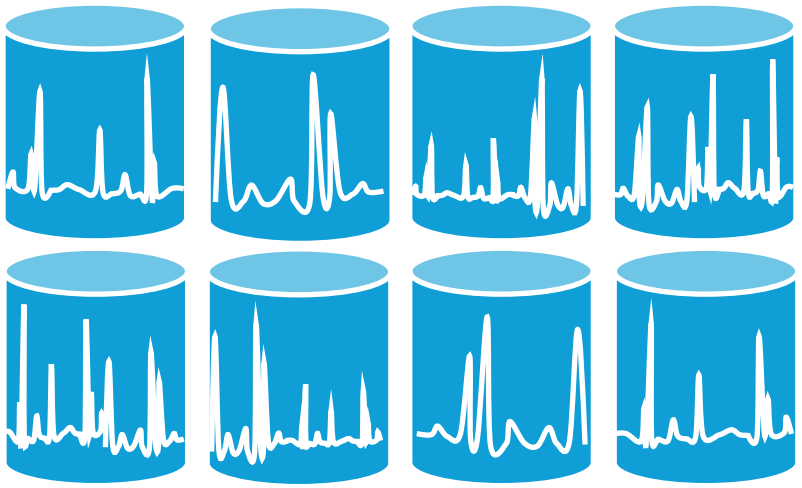
- All service tiers: standard-series (“Gen 5”)
- Hyperscale only: premium-series and memory-optimized premium-series

Azure SQL DB **Elastic Pools**

Real-world analogy			
	Single-family home + High flexibility / freedom + Customize to needs - Costs (down payment, mortgage, maintenance, etc.) - Oversized home can be expensive for a small group	**Rentals / hotels** + Mostly, pay as you use + Right-size per your needs - Can be expensive for "heavy use"	**Corporate serviced apartments / Dorms / hostels** + Most economical for larger groups + Can be comfortable (if tenants cooperate ☺) - Potential for "noisy tenants" to cause problems
Azure SQL equivalent	**Single DB (Provisioned Compute)**	**Single DB (Serverless Compute)**	**Elastic Pool (Provisioned Compute)**

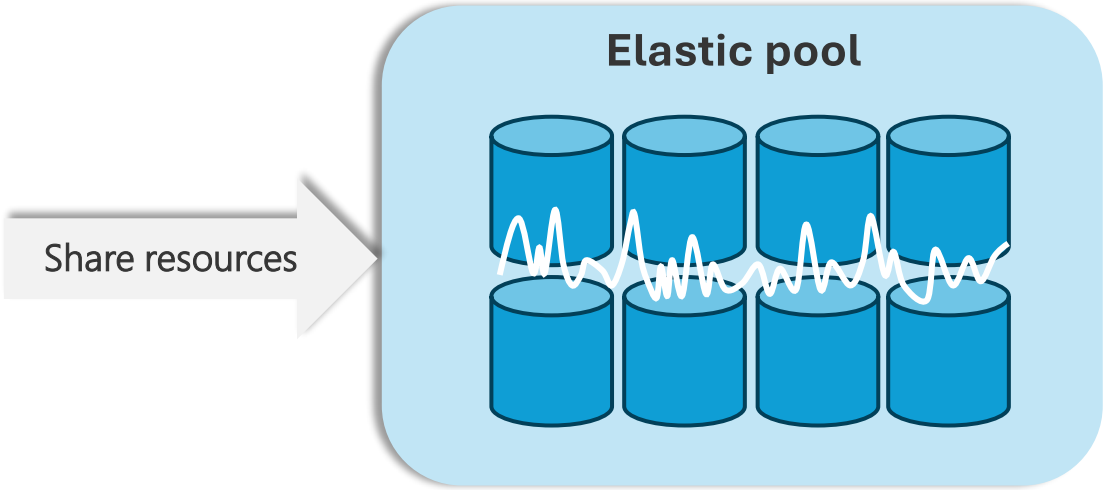
Elastic pools are great for multiple DB workloads

Cost = ∑ (Max for each DB)



Must provision each DB for individual peaks

Cost = Max (∑ (DB resource used at any given time))



Collectively provision for max combined load

Azure SQL DB Hyperscale elastic pools

Public preview

Increased storage limits

Storage up to 100 TB per elastic pool

Higher log generation rates

Log throughput independent of compute

Predictable scaling

Scale the compute for the elastic pool up or down in minutes, regardless of how big the pooled DBs are!

Pool-level read scale

Optional pool replicas to host read-only workload

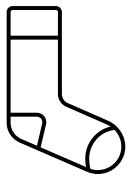
https://aka.ms/hsepvideo (4 minute video); https://aka.ms/hsep (docs)
https://aka.ms/hsep-public-preview-tech-blog (public preview announcement).

Costs associated with an elastic pool

	General Purpose	Business Critical	Hyperscale
Pool compute	Fixed cost **per pool**	Fixed cost **per pool**	Fixed cost **per pool**
Zone redundancy	Add-on cost	Included	Additional cost proportionate to # of HA replicas)
Software (license)	Fixed cost **per pool**	Fixed cost per **pool**	N/A
Storage	Fixed cost **per pool** (depends on **max size**)	Fixed cost **per pool** (depends on **max size**)	Based on actual allocated size **per database**.
Backup	**Tracked per database**, but aggregated and **billed** at the **pool level**.	**Tracked per database**, but aggregated and **billed** at the **pool level**.	**Tracked** and **billed per database.**

Quiz time!

I have 4 DBs storing live stock price data (UK, US, Japan, India). Would elastic pools be a good choice here?



HyperscalePool1 (mylogicalserver/HyperscalePool1) | Database Resource Utilization

SQL elastic pool

Search

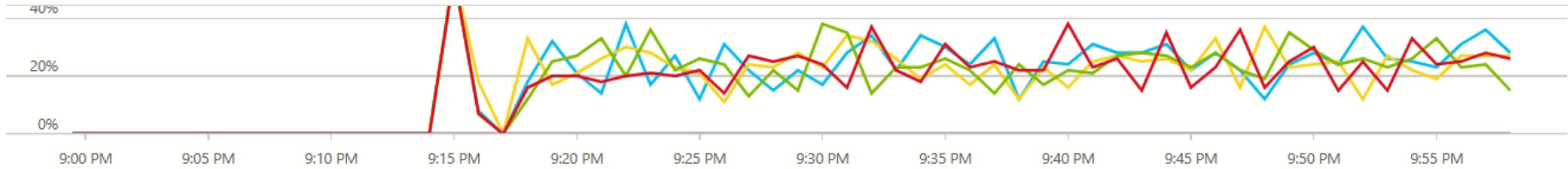
Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems

Settings
Quick start
Configure
Locks

Monitoring
Database Resource Utilization
Alerts
Metrics
Diagnostic settings
Logs

Automation
Tasks (preview)
Export template

Feedback



Select additional metrics to display below:

0 selected

Elastic databases within the pool (select up to five)

Search to filter databases...

Database name	Status	Avg CPU (%)	Peak CPU (%)
tenant-db-06	Online	12.08	55
tenant-db-08	Online	12.067	55
tenant-db-03	Online	11.962	55
tenant-db-07	Online	11.769	55
tenant-db-02	Online	11.689	55
tenant-db-01	Online	11.529	55
tenant-db-04	Online	11.489	55
tenant-db-05	Online	0	0

HyperscalePool2 (mylogicalserver/HyperscalePool2)

SQL elastic pool

Search

Overview
Activity log
Access control (IAM)
Tags
Diagnose and solve problems

Settings
Quick start
Configure
Locks

Monitoring
Database Resource Utilization
Alerts
Metrics
Diagnostic settings
Logs

Automation
Tasks (preview)
Export template

Configure | Delete | Create database | Feedback

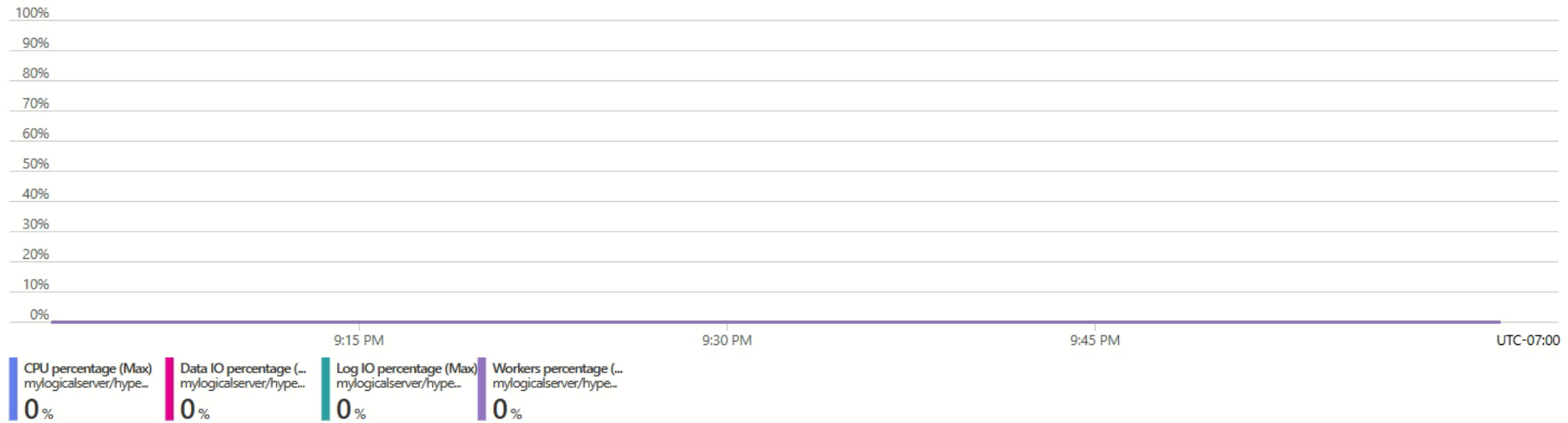
Essentials

View Cost | JSON View

Show data for last: 1 hour | 24 hours | 7 days

Aggregation type: Max

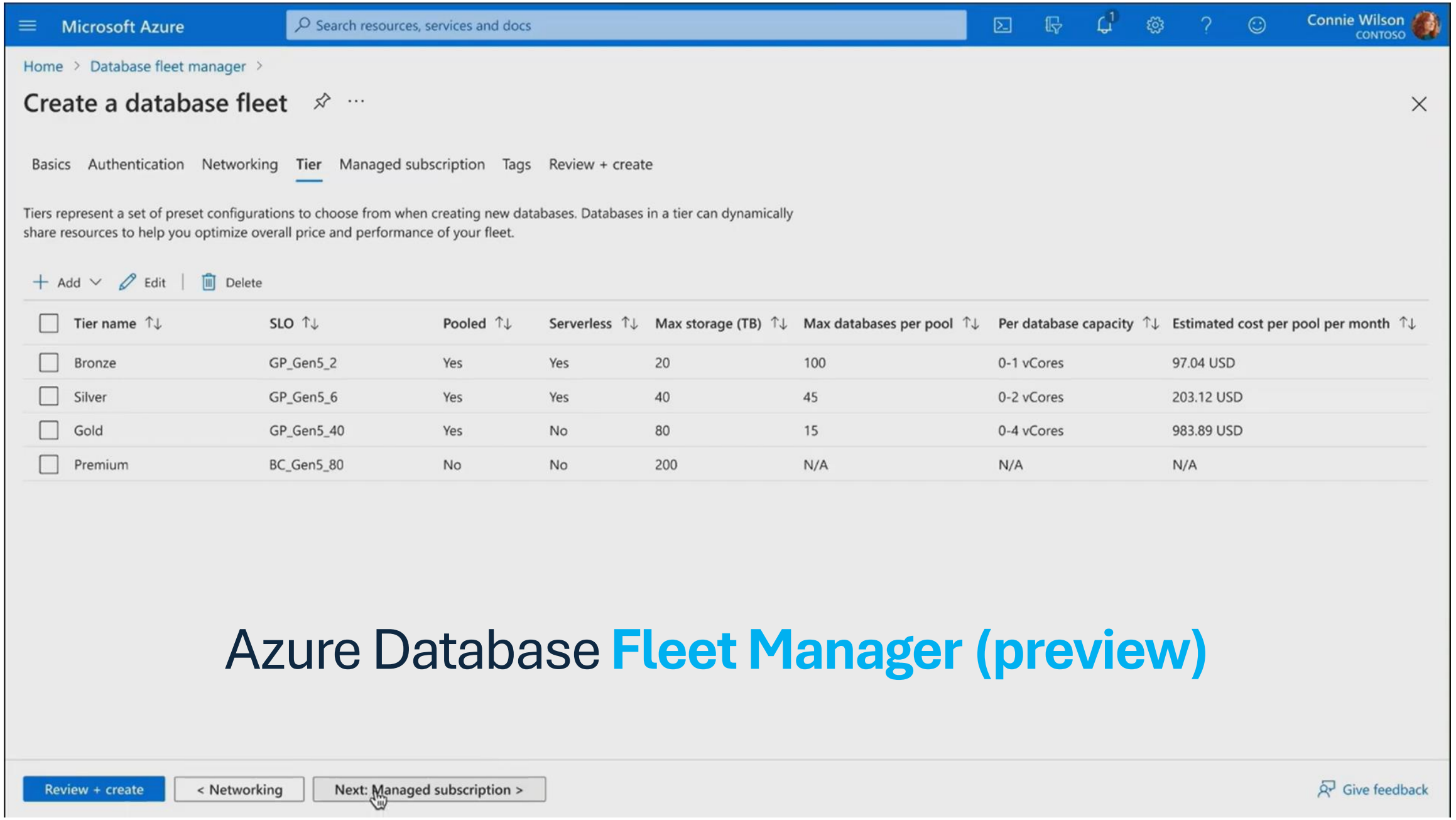
Resource utilization (HyperscalePool2)



Notifications (0)

All | Alerts (0) | Recommendations (0) | Info (0)

#4: Fine tuning



Azure Database **Fleet Manager (preview)**

https://techcommunity.microsoft.com/t5/azure-sql-blog/introducing-azure-database-fleet-manager/ba-p/3979576

Resource management and optimization with Azure database fleet manager

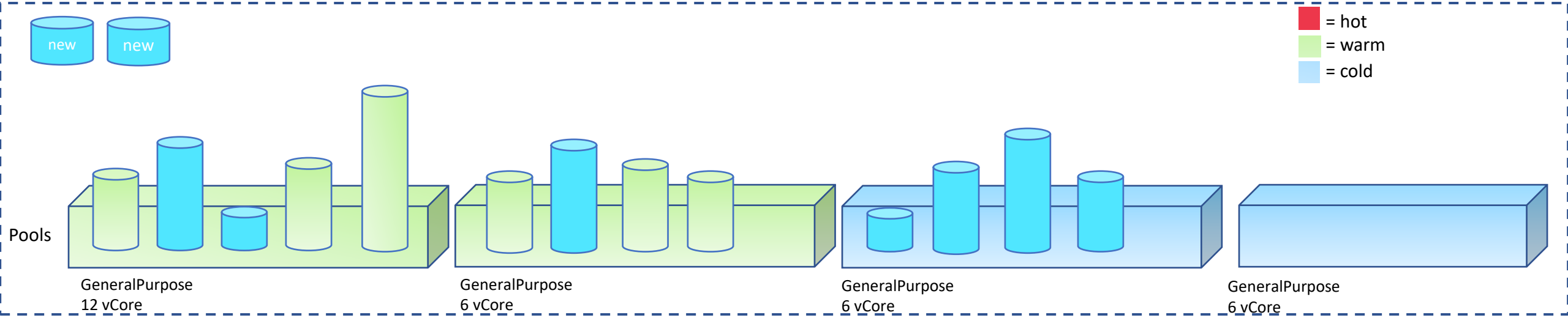
Capacity Management
Provisioning servers and pools ahead of time

Database placement
Fast (with pre-provisioning) and intelligent placement across pools based on tier definition policy

Tier re-balancing
Re-balance databases across elastic pools to keep overall temperature under control and mitigate HOT dbs

Auto-scaling
Scale up/down individual pools to accommodate peaks based on tier policy definition

Defragmentation
Actively try to reduce overall number of pools



More details: “Modern models of managing database fleets in Azure PaaS” session @ SQLBits 2024.

Azure SQL DB **Backup costs**

- Select the right storage redundancy:
  - Locally redundant (LRS)
  - Zone redundant (ZRS)
  - Geo redundant (RA-GRS)
  - Geo zone redundant (GZRS)
- By default, RA-GRS is chosen as the backup storage with 7 days of retention for Point in Time Restore.

Point in Time Restore (PITR) backup rates

Redundancy	Price
LRS	$0.10/GB/month
ZRS	$0.125/GB/month
RA-GRS*	$0.20/GB/month

Long Term Retention (LTR) backup rates

Redundancy	Price
LRS	$0.025/GB/month
ZRS	$0.0313/GB/month
RA-GRS	$0.05/GB/month
RA-GZRS	$0.0845/GB/month

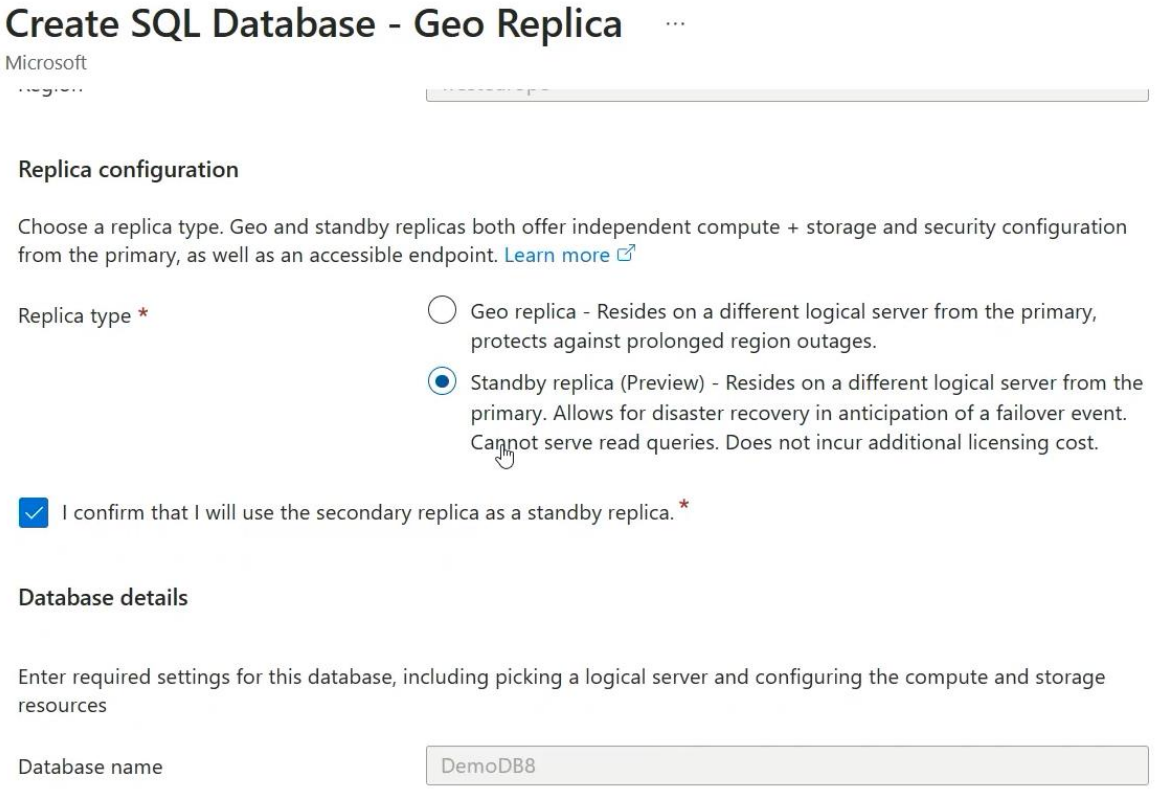
Possible configuration for optimal **Backup costs**

Database Category	PITR Storage	PITR Retention (Days)	LTR Storage
Tier 1 Production	RA-GRS	**35**	**RA-GRS**
Tier 2 Production	RA-GRS	**15**	**RA-GRS**
Tier 3 Production	RA-GRS	7	Not Needed
Low Tier Production	**ZRS**	7	Not Needed
Dev/Test	**LRS**	**3**	Not Needed
Integration/Perf/UAT	**LRS**	**1**	Not Needed

Related session: “Perfecting business continuity for Azure SQL DB” on Saturday @ 10:10AM.

#5: Apply discounts

Standby replica (only for General Purpose / Business Critical only)

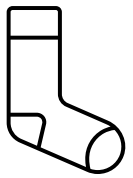


Deployment model	Compute tier	Service tier	Standby replica supported	Hardware
Single database	Provisioned	General Purpose	Yes	Standard-series (Gen5), FSv2-Series, DC-Series
Single database	Provisioned	Business Critical	Yes	Standard-series (Gen5), DC-Series
Single database	Provisioned	Hyperscale	N/A*	N/A
Single database	Serverless	All	No	N/A
Elastic pool	All	All	No	N/A

Related session: “Perfecting business continuity for Azure SQL DB” on Saturday @ 10:10AM.

Quiz time!

Why is the Standby Replica N/A for Hyperscale?



Azure Hybrid Benefit

- Exchange your existing licenses for discounted rates on Azure SQL Database by using your Software Assurance-enabled SQL Server licenses on Azure.
- Azure Hybrid Benefit (AHB) can be applied to Business Critical as well as General Purpose service tiers.

General Purpose

vCORE	Memory (GB)	Pay as you go	Azure Hybrid Benefit[1] Price
2	10.2	**$0.505**/hour	**$0.305**/hour ~40% savings
4	20.4	**$1.009**/hour	**$0.609**/hour ~40% savings

Business Critical

vCORE	Memory (GB)	Pay as you go	Azure Hybrid Benefit[1] Price
2	10.2	**$1.359**/hour	**$0.609**/hour ~55% savings
4	20.4	**$2.718**/hour	**$1.218**/hour ~55% savings

Azure **Reserved Capacity**

- **Commit** upfront to Azure SQL Database resources for either one or three years and get a discount on provisioned compute. Huge savings, especially when combined with AHB.
- Azure Reserved Capacity discount applies to Business Critical, General Purpose and Hyperscale service tiers.

General Purpose

vCORE	Memory (GB)	Pay as you go	Azure Hybrid Benefit[1] Price	1 year reserved capacity [2]	3 year reserved capacity [2]	3 years reserved with Azure Hybrid Benefit [2]
2	10.2	$0.505/hour	$0.305/hour ~40% savings	$0.398/hour ~21% savings	$0.337/hour ~33% savings	$0.137/hour ~73% savings
4	20.4	$1.009/hour	$0.609/hour ~40% savings	$0.796/hour ~21% savings	$0.674/hour ~33% savings	$0.274/hour ~73% savings

Business Critical

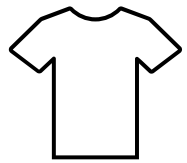
vCORE	Memory (GB)	Pay as you go	Azure Hybrid Benefit[1] Price	1 year reserved capacity [2]	3 year reserved capacity [2]	3 years reserved with Azure Hybrid Benefit [2]
2	10.2	$1.359/hour	$0.609/hour ~55% savings	$1.146/hour ~16% savings	$1.024/hour ~25% savings	$0.274/hour ~80% savings
4	20.4	$2.718/hour	$1.218/hour ~55% savings	$2.292/hour ~16% savings	$2.048/hour ~25% savings	$0.548/hour ~80% savings

Hyperscale

vCORE	Memory (GB)	Pay as you go	3 year reserved capacity [2]
2	10.2	$0.366/hour ~35% savings	$0.165/hour ~71% savings
4	20.4	$0.731/hour ~35% savings	$0.329/hour ~71% savings

Quiz time!

Is Reserved Capacity applicable to all Compute Tiers?



Sample cost comparisons for Production DBs

Assumptions

- Assume the database is sized with 4 vCores.
- Database has 500 GB data.
- The vCore costs for Business Critical and General Purpose in the upcoming slides include Azure Hybrid Benefit (AHB) discount and 3-year Reserved Capacity (RI) discount.

Account for Hyperscale HA replicas

- The Hyperscale HA replica effectively adds 4 more available vCores.
- So, we can can safely expect better performance than corresponding General Purpose database even at 50% of General Purpose vCores.
- So the following slides additionally includes costing for Hyperscale at 50% vCores.

Final **Cost comparison** with **Hybrid Benefit** and **Reserved Capacity** discounts

- Combining AHB and RI gives you the maximum possible discount but does impact TCO due to the cost of SQL Server licenses / Software Assurance (SA).
  - Approximate SA cost is shown in the calculation for the primary and Geo Replica (as applicable).
  - There is no license or SA requirement and no Azure Hybrid Benefit discount for Hyperscale, Hyperscale Serverless or General Purpose Serverless.

Tier 1 Production

	Business Critical	MO Hyperscale
4 vCore	**$394.56**	**$296.64**
Storage	$125.00	$125.00
Zone Redundancy	$0.00	**$296.64**
Geo Replica	**$394.56**	**$296.64**
Geo Replica Storage	$125.00	$125.00
Software Assurance (SA) Approx.	$1,260.00	N/A
Total Cost	**$2,299.12**	**$1,139.92**

Tier 2 Production

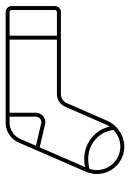
	Business Critical	Hyperscale
4 vCore	**$394.56**	**$236.88**
Storage	$125.00	$125.00
Zone Redundancy	$0.00	**$236.88**
Geo Replica Standby	**$394.56**	**$236.88**
Geo Replica Storage	$125.00	$125.00
Software Assurance (SA)	$630.00	N/A
Total Cost	**$1,669.12**	**$960.64**

Tier 3 Production

	Hyperscale	General Purpose	Hyperscale 50%
4 vCore	**$236.88**	**$197.28**	**$118.80**
Storage	$125.00	$57.50	$125.00
Zone Redundancy	**$236.88**	**$88.77**	**$118.80**
Geo Replica	N/A	N/A	N/A
Geo Replica Storage	N/A	N/A	N/A
Software Assurance (SA)	N/A	$366.00	N/A
Total Cost	**$598.76**	**$709.55**	**$362.60**

Quiz time!

In the previous comparisons, what was usually the biggest element of cost?



sqlbits
★★★ 2024 ★★★

Most **cost optimization** roads lead to Azure SQL DB **Hyperscale!**

References

- Azure SQL DB pricing: https://aka.ms/sqldbprice
- Azure pricing calculator: https://aka.ms/azurecalc
- More about Hyperscale: https://aka.ms/hs
- Hyperscale simplified and lower pricing: https://aka.ms/hsignite2023 and https://aka.ms/hsprice2023
- Hyperscale serverless: https://aka.ms/sqldbserverless
- Hyperscale elastic pools: https://aka.ms/hsep

Please do provide feedback!

- Short URL: https://sqlb.it/?12759

Thank you!

@aditya_feb22 (X)

@arvisam (Bluesky, Mastodon and X)