AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

DAT322

Deep dive into Amazon Neptune Serverless

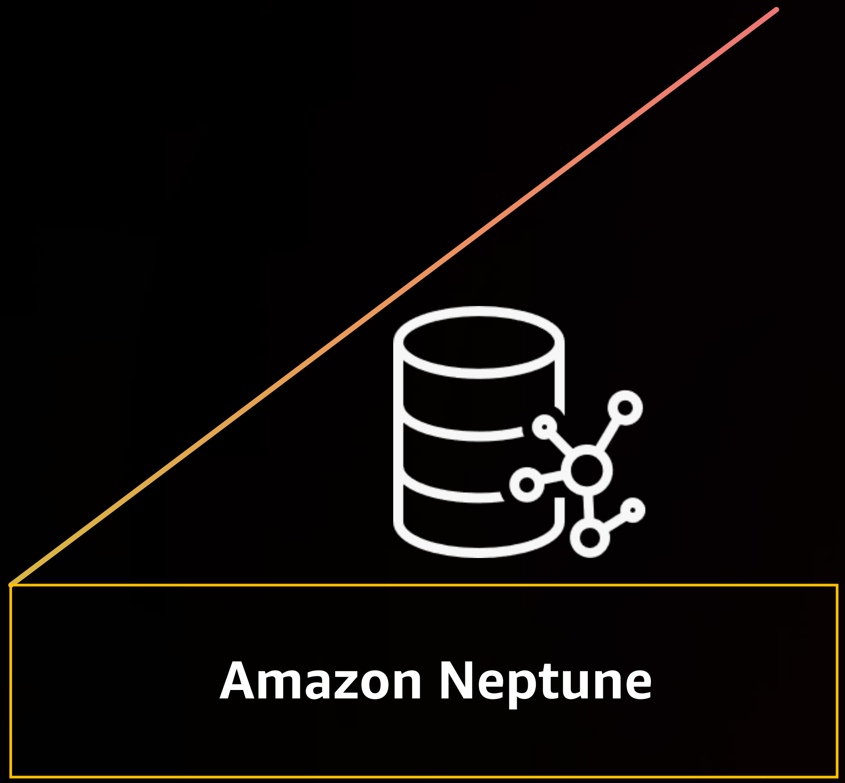
Brad Bebee (he/him)

General Manager, Amazon Neptune
AWS

Ian Robinson (he/him)

Principal Graph Architect, Amazon Neptune
AWS

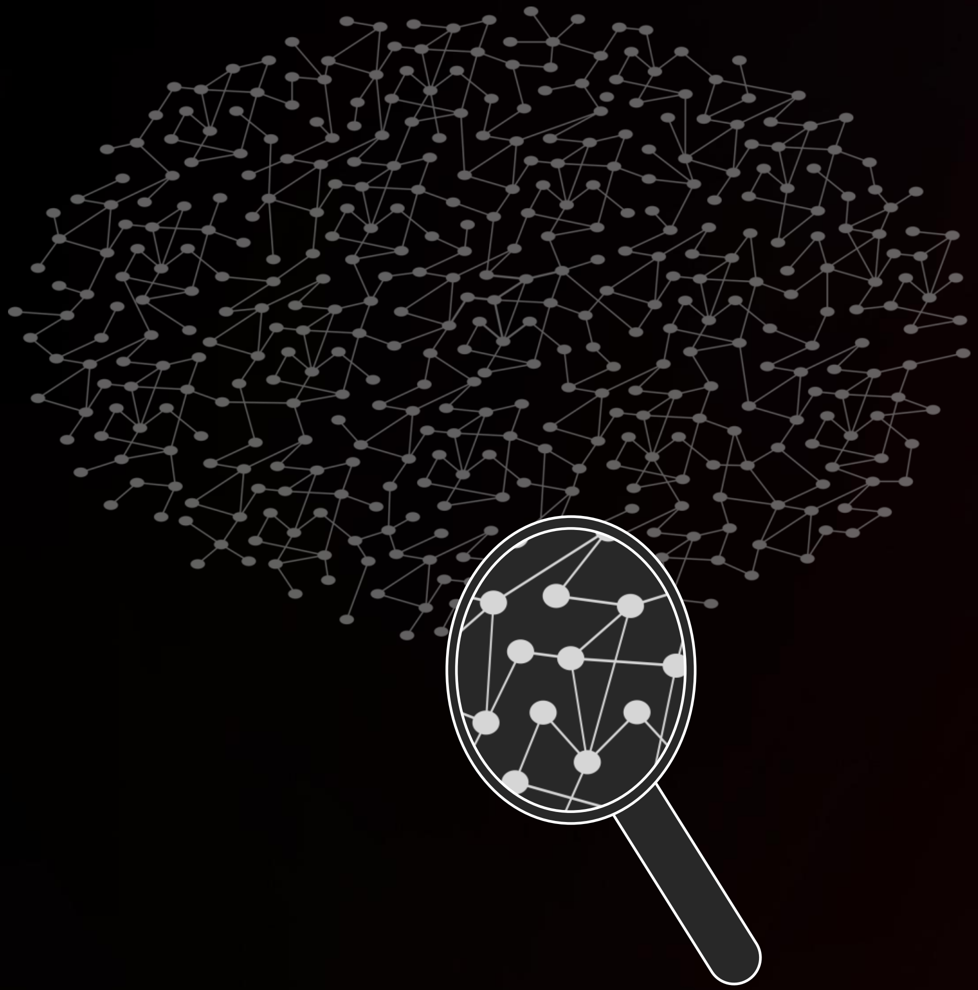
Agenda



Amazon Neptune

- Quick review of graphs and use cases
- Understand how recent Neptune features work
- Understand how Neptune Serverless optimizes CPU and memory
- Show you a Global Serverless Graph Database Cluster
- Take questions!

Graphs are awesome!



1. Model data based on relationships
2. Applications explore connections and patterns in connected data
3. Processing graphs is hard due to random data access
4. Generalized graph operations require purpose-built processing

Amazon Neptune (Now Serverless and Global too!)

FULLY MANAGED, PURPOSE-BUILT GRAPH DATABASE IN THE CLOUD

Cost-effective

No hardware management

Instant provisioning

Security and compliance

Serverless

Global

- Optimized to **store and map billions of relationships**
- Enables **real-time navigation of connections with millisecond query** response time
- Supports **open standard query languages** openCypher, Gremlin, and SPARQL

Every day thousands of customers use Neptune

Amazon Neptune

Customers across different verticals and use cases use Amazon Neptune in production today

WIZ
FINRA
SIEMENS
AstraZeneca
JupiterOne
freshworks

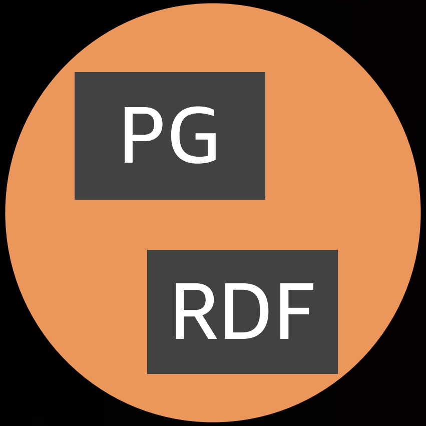
amazon alexa
intuit
FACTSET
SAMSUNG
Blackfynn
Rappi

Uber ATG
HUUUGE
NBCUniversal
NETFLIX
asurion
COX AUTOMOTIVE

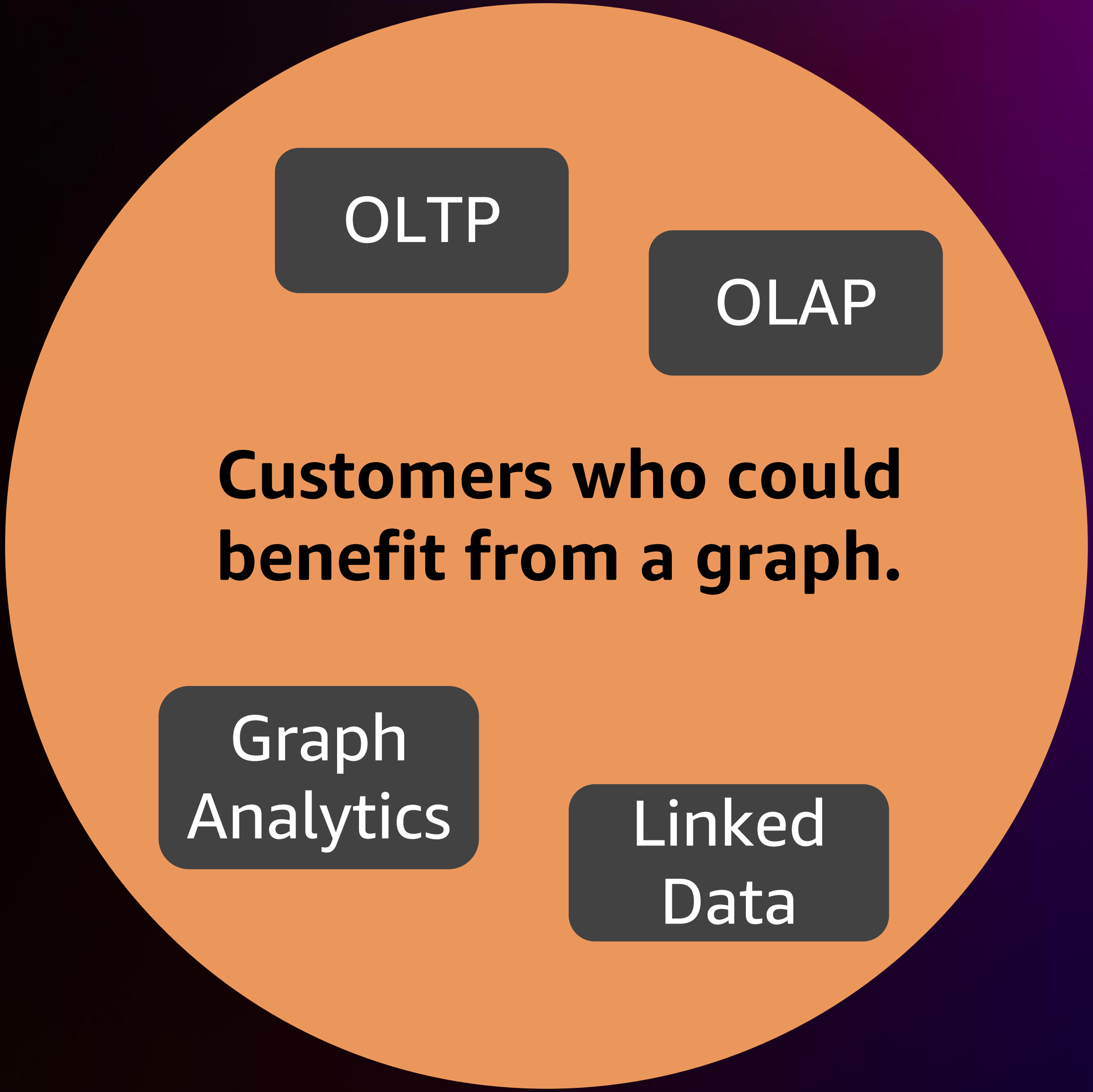
PaySense
zeta
Pearson
MARINUS ANALYTICS
noonum
THOMSON REUTERS

Case studies: https://aws.amazon.com/solutions/case-studies/?customer-references-cards.q=neptune

It’s still Day 1 for graphs



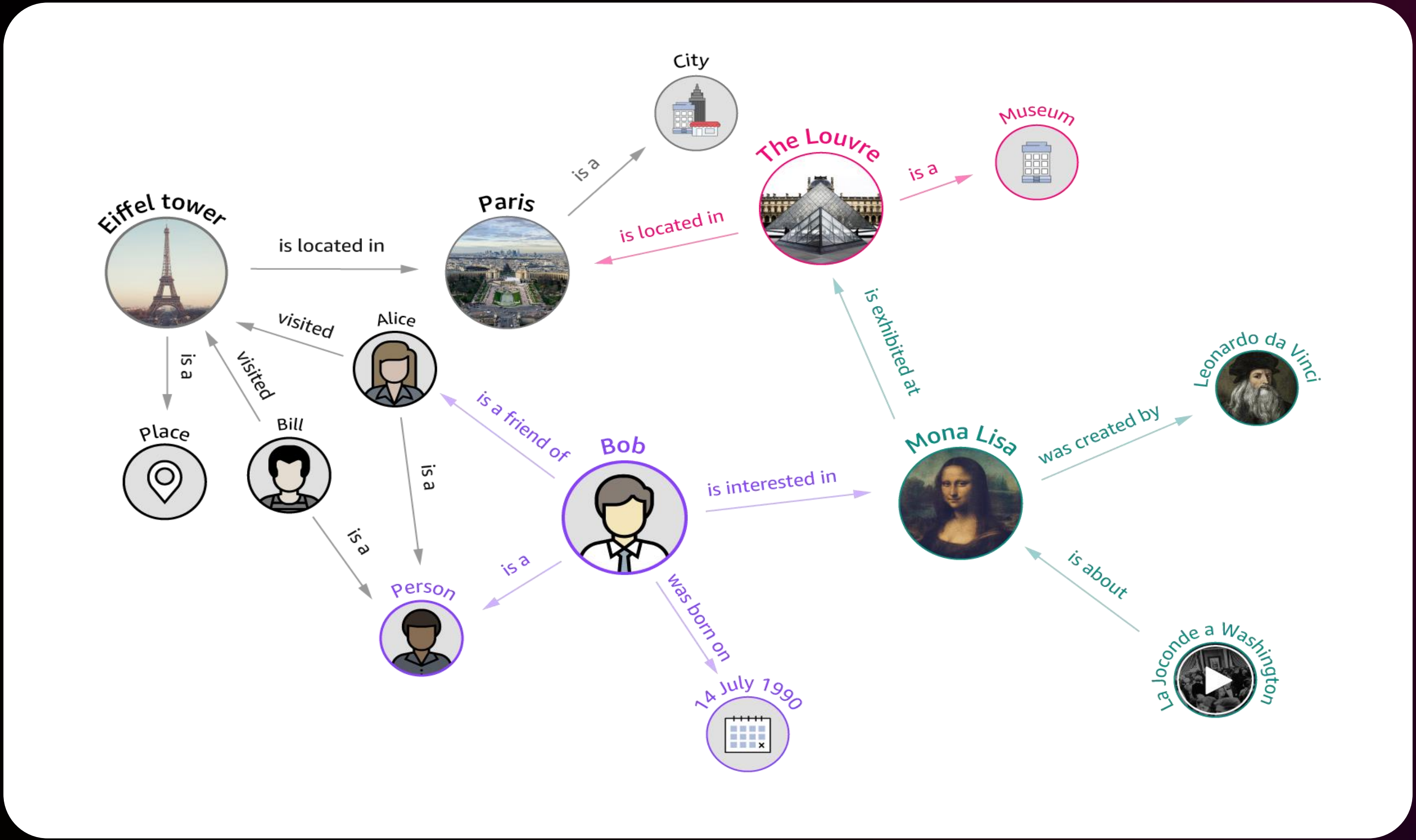
Customers who know they want a graph database.



Customers who could benefit from a graph.

Knowledge graphs

UNDERSTANDING THE WHO, WHAT, WHEN, AND WHERE



https://aws.amazon.com/neptune/knowledge-graphs-on-aws/

SIEMENS
Ingenuity for life

Siemens is a global powerhouse focusing on the areas of electrification, automation, and digitalization

Challenge:

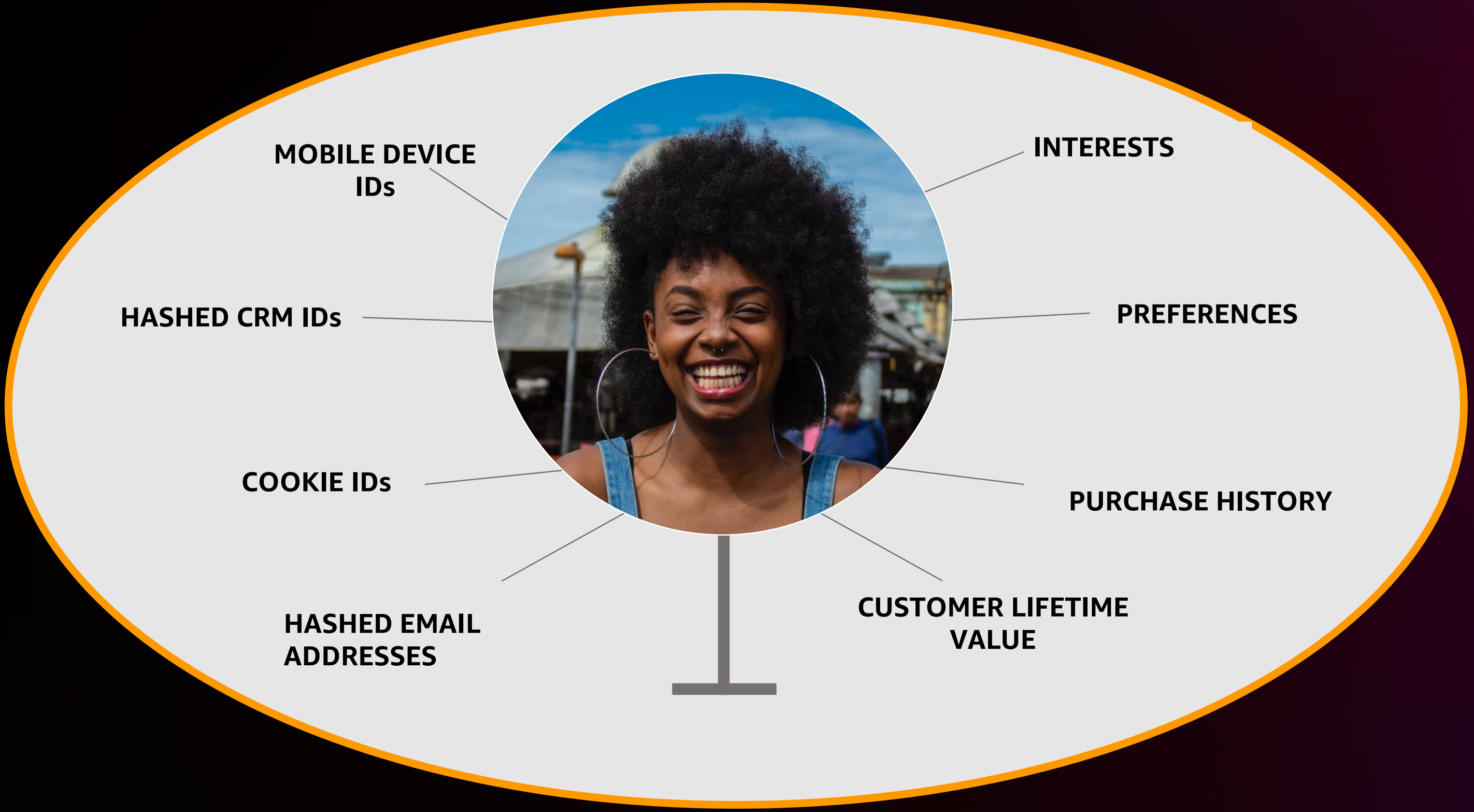
They were faced with isolated data silos from different departments that resulted in data inaccessibility, inefficient workflows, and low data quality

Solution:

- Industrial knowledge graphs for capturing Siemens Domain Knowledge
- Providing knowledge graphs as a service

Identity graphs

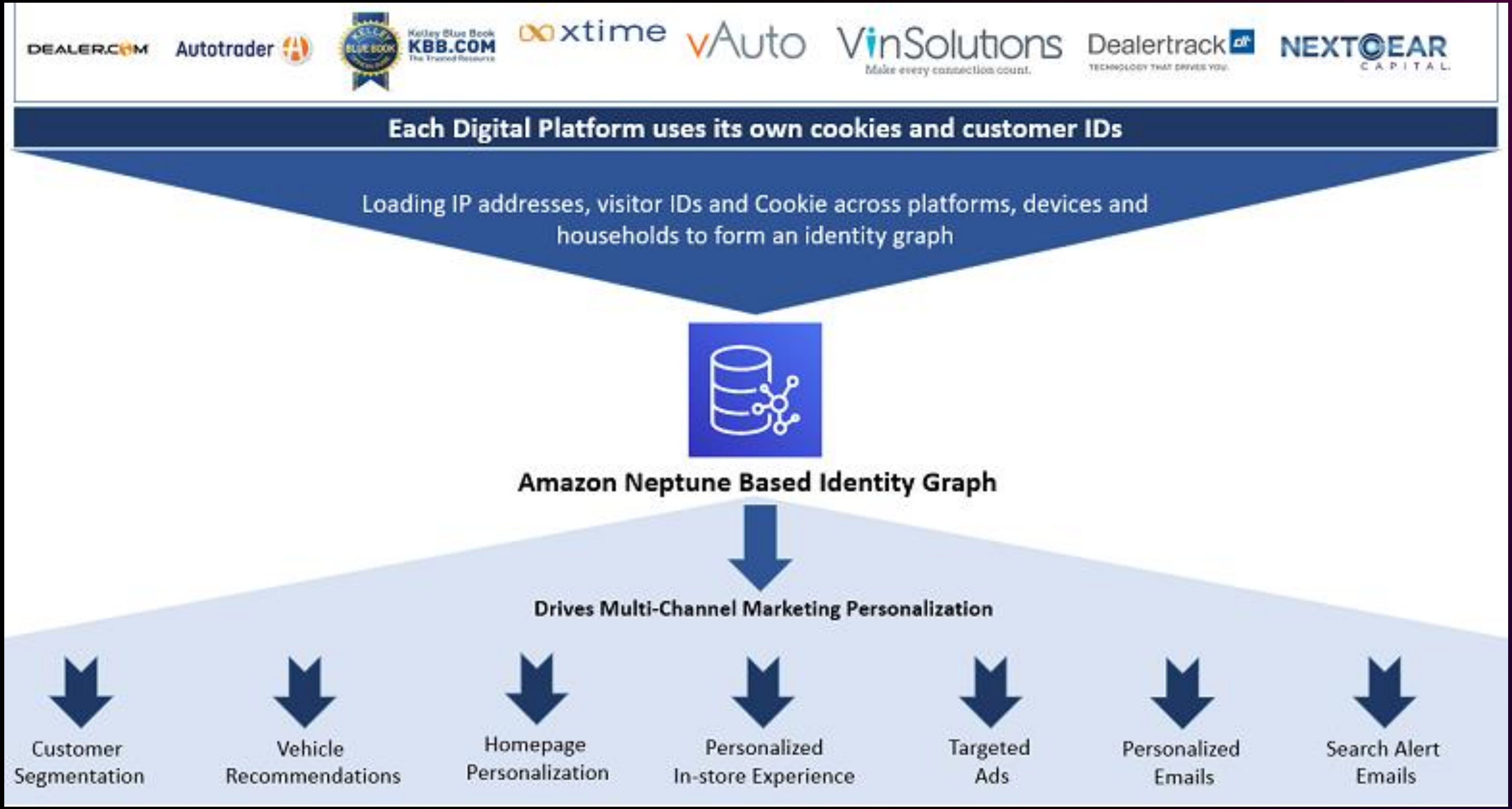
UNIFIED 360° VIEW OF THE CUSTOMER



https://aws.amazon.com/neptune/identity-graphs-on-aws/

Identity graph with Amazon Neptune

COX AUTOMOTIVE MAKES BUYING, SELLING, OWNING, AND USING CARS EASIER FOR EVERYONE


DEALER.COM
Autotrader
Kelley Blue Book KBB.COM The Trusted Resource
xtime
vAuto
VinSolutions Make every connection count.
Dealertrack TECHNOLOGY THAT DRIVES YOU.
NEXTGEAR CAPITAL

Each Digital Platform uses its own cookies and customer IDs

Loading IP addresses, visitor IDs and Cookie across platforms, devices and households to form an identity graph

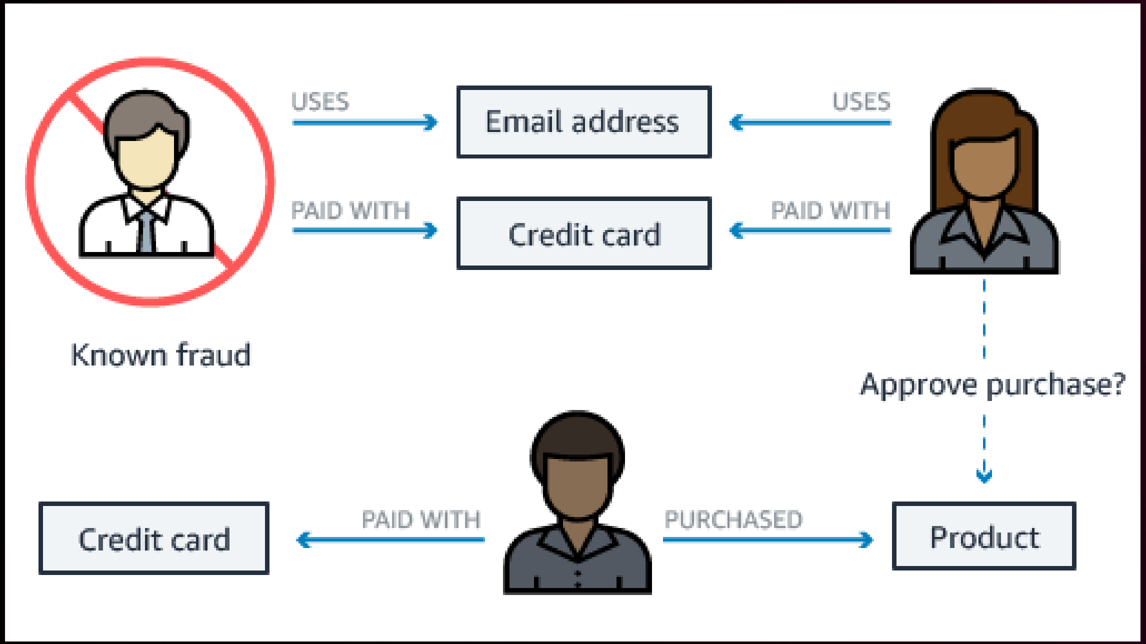
Amazon Neptune Based Identity Graph

Drives Multi-Channel Marketing Personalization

- Customer Segmentation
- Vehicle Recommendations
- Homepage Personalization
- Personalized In-store Experience
- Targeted Ads
- Personalized Emails
- Search Alert Emails

Fraud graphs

DETECTING FRAUD AS IT HAPPENS USING RELATIONSHIPS



https://aws.amazon.com/neptune/fraud-graphs-on-aws/

GAMES 24 SEVEN

As India's leading gaming company, Games24x7 is known for its flagship products like RummyCircle, which offers online rummy, and My11Circle, which offers fantasy sports.

Challenge:

As the game of Rummy involves real money, Games24x7 has to stay vigilant to prevent fraud and collusion during tournaments.

Solution:

It uses the Amazon Neptune graph database to detect if two players in a game are colluding to beat the other four players. This is accomplished by assigning a table in the database to each player when they log in.

RummyCircle.com™

apna GAME dikhao

https://aws.amazon.com/solutions/case-studies/games24x7/

Security graphs

UNDERSTAND SECURITY VULNERABILITIES ACROSS LAYERS

1. Cloud Security Posture Management
2. Data Flow/Exfiltration
3. Identity and Access Management

https://aws.amazon.com/neptune/security-graphs-on-aws/

Wiz security graphs: Detecting critical risks

Workload context

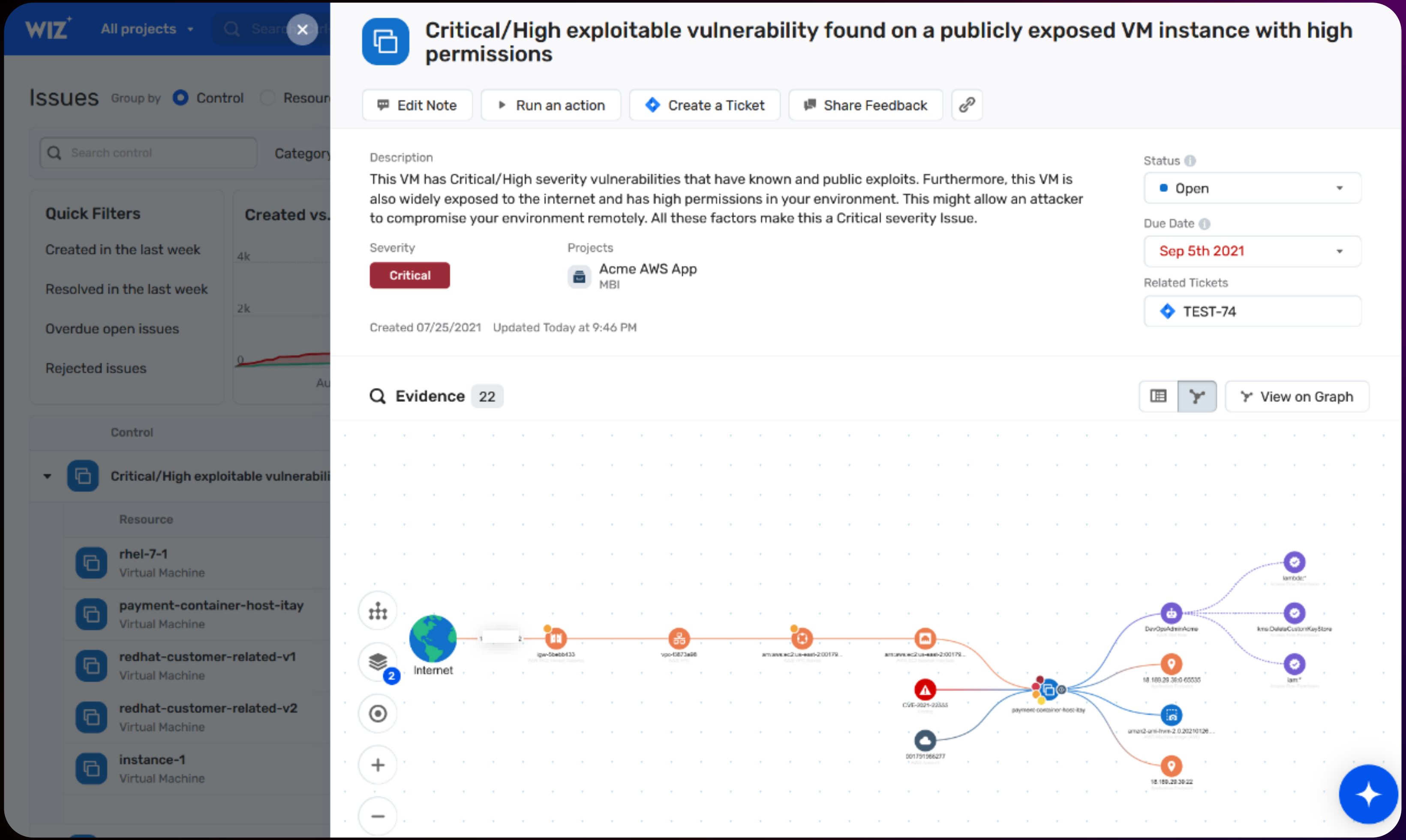
- Vulnerabilities
- Inventory
- Exposed secrets

Cloud context

- Resource configuration
- Networking
- Identities

Business context

- Tags
- Environment
- Business team



Critical/High exploitable vulnerability found on a publicly exposed VM instance with high permissions

Description

This VM has Critical/High severity vulnerabilities that have known and public exploits. Furthermore, this VM is also widely exposed to the internet and has high permissions in your environment. This might allow an attacker to compromise your environment remotely. All these factors make this a Critical severity Issue.

Severity: Critical

Projects: Acme AWS App MBI

Created 07/25/2021 Updated Today at 9:46 PM

Status: Open

Due Date: Sep 5th 2021

Related Tickets: TEST-74

Evidence 22

Launched this year

IN CASE YOU MISSED IT!

Neptune ML Inductive Inference

openCypher GA

Global Database

Serverless

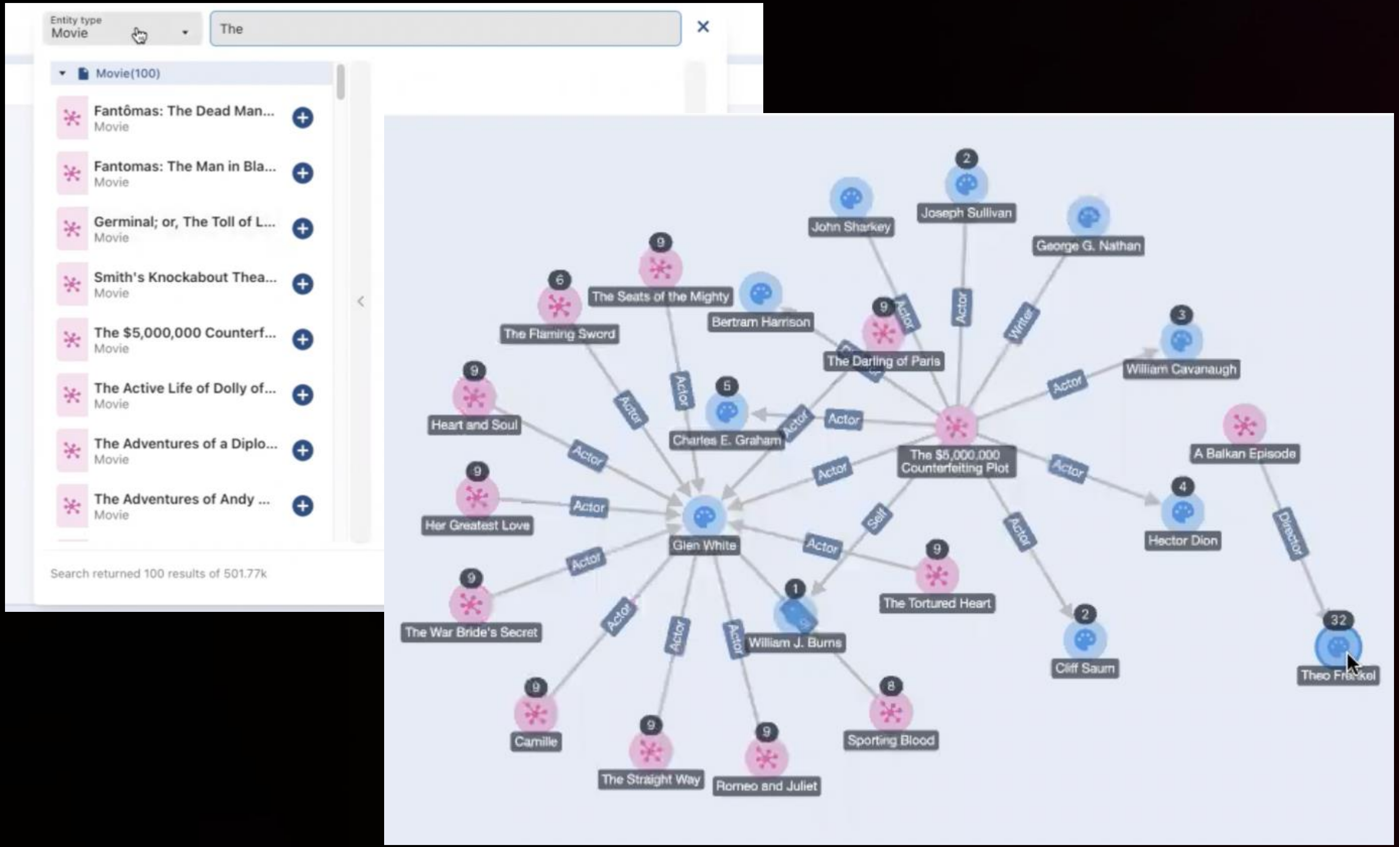
Plus 14 engine releases year-to-date to improve performance, features, reliability, and availability…

Customers said they wanted an easy way to explore their graph data

Announcement!

Graph Explorer: Open-source graph exploration

A simple low-code web application that makes graphs easier to understand by presenting the information in an interactive and visual explorer



Entity type: Movie

The

Movie(100)

- Fantômas: The Dead Man... Movie
- Fantomas: The Man in Bla... Movie
- Germinal; or, The Toll of L... Movie
- Smith's Knockabout Thea... Movie
- The $5,000,000 Counterf... Movie
- The Active Life of Dolly of... Movie
- The Adventures of a Diplo... Movie
- The Adventures of Andy ... Movie

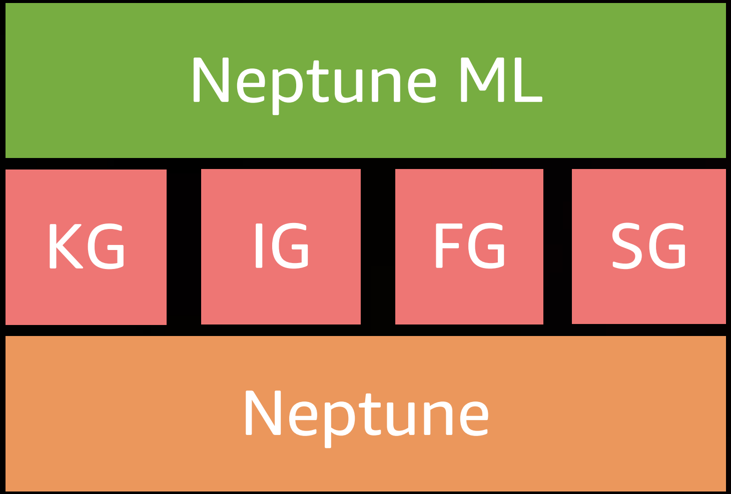
Search returned 100 results of 501.77k

- Start exploring using faceted search
- Click to expand and customize visualizations
- Supports RDF and property graph data
- Will be available under an Apache 2 license in December 2022

New!

Amazon Neptune ML: Fast and accurate predictions on graphs

Easy, fast, and accurate predictions on graphs with graph neural networks (GNNs), powered by the Deep Graph Library (DGL) and Amazon SageMaker



Now supports Online Inductive Inference (OII) for dynamic graph predictions

- With inductive inference, the GNN model applies data processing and model evaluation in real time.
- Expands Neptune ML to use cases like fraud and recommendations that require predictions based on the current state of the graph.
- Available in Engine Versions 1.2.0.2+

https://aws.amazon.com/neptune/machine-learning/

openCypher for Amazon Neptune

Developers can now use openCypher, a popular graph query language, with Amazon Neptune, providing them the most choice to build or migrate graph applications

Generally Available

A declarative query language for property graph data

- openCypher allows customers to draw on their SQL knowledge to help power their businesses with graph applications

Data interoperability

- Customers can use the openCypher and Apache TinkerPop Gremlin query languages over the same property graph data

Compatible with Bolt Protocol

- Allows customers to leverage familiar and existing tooling to migrate workloads

Avoid expensive commercial licensing

Comparing Gremlin and openCypher

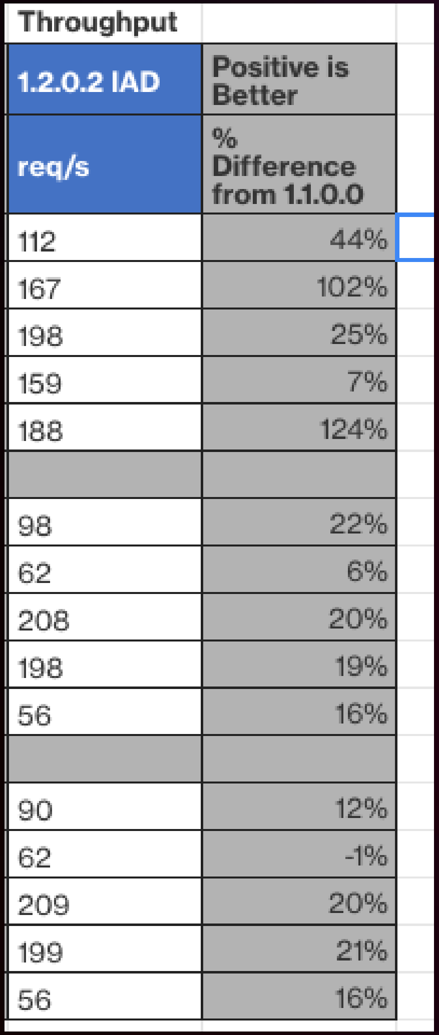
	openCypher	Gremlin
Style	Declarative	Imperative

Syntax	Pattern matching
	MATCH p=(a)-[:route]->(d)
	WHERE a.code=‘ANC’
	RETURN p
		Traversal based
		g.V().has(‘code’, ‘ANC’).
		out(‘route’).path().
		by(elementMap())

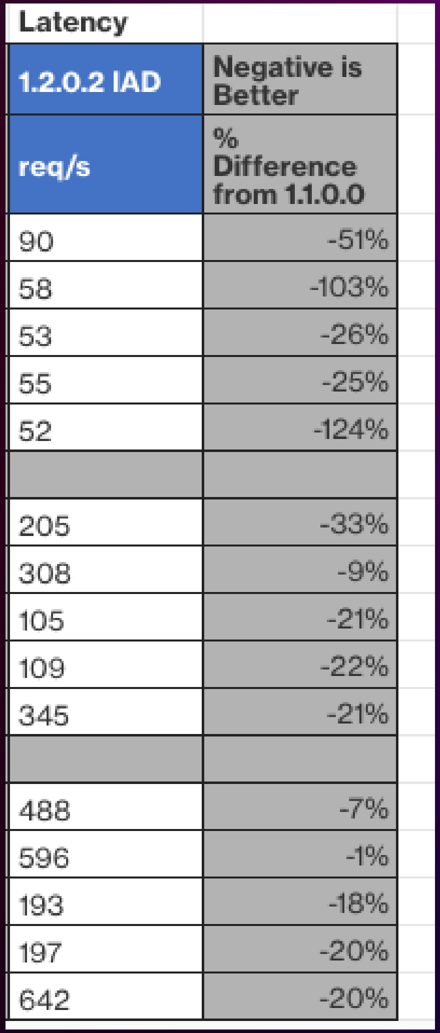
Ease of use	Easy to learn, SQL-inspired readable by non-programmers	Steeper learning curve, similar to stream processing languages
Flexibility	Low	High
Query support	String based queries	String based queries or programmatic based GLVs
Clients	HTTPS and Bolt	HTTPS and Websockets

openCypher tips and roadmap

- We're iterating rapidly – keep your engine version up-to-date.
  - Last week's 1.2.0.2 release included significant performance improvements for variable length path (VLP) queries
  - New features (user-specified IDs)
- Don't be a stranger – we can help!
  - Let us know how openCypher is working for you.
  - Check out the latest documentation for tips and best practices.



Throughput	
1.2.0.2 IAD	Positive is Better
req/s	% Difference from 1.1.0.0
112	44%
167	102%
198	25%
159	7%
188	124%
98	22%
62	6%
208	20%
198	19%
56	16%
90	12%
62	-1%
209	20%
199	21%
56	16%



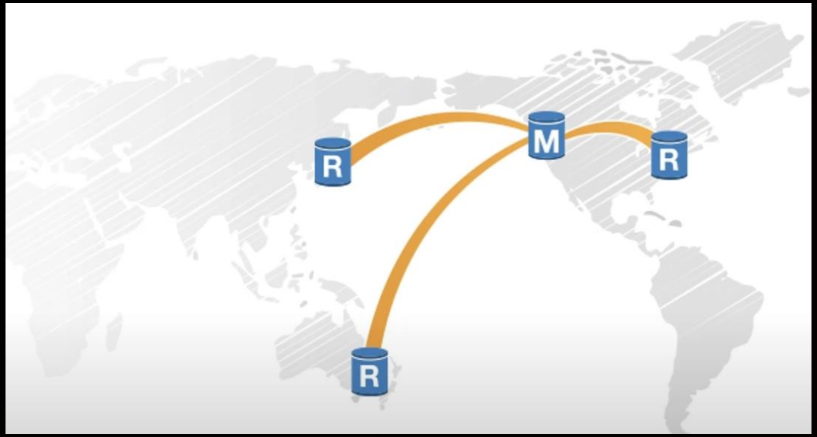
Latency	
1.2.0.2 IAD	Negative is Better
req/s	% Difference from 1.1.0.0
90	-51%
58	-103%
53	-26%
55	-25%
52	-124%
205	-33%
308	-9%
105	-21%
109	-22%
345	-21%
488	-7%
596	-1%
193	-18%
197	-20%
642	-20%

neptune-opencypher-feedback@amazon.com

Amazon Neptune Global Database

Generally Available

Deploy Neptune clusters across multiple AWS Regions for fast cross-region disaster recovery and low-latency global reads



Disaster recovery

- Maintain business continuity in the event of regional outages with fast global failover to secondary AWS Regions

Low latency reads

- Connect to the Neptune cluster closest to your applications

Fast cross-Region migrations

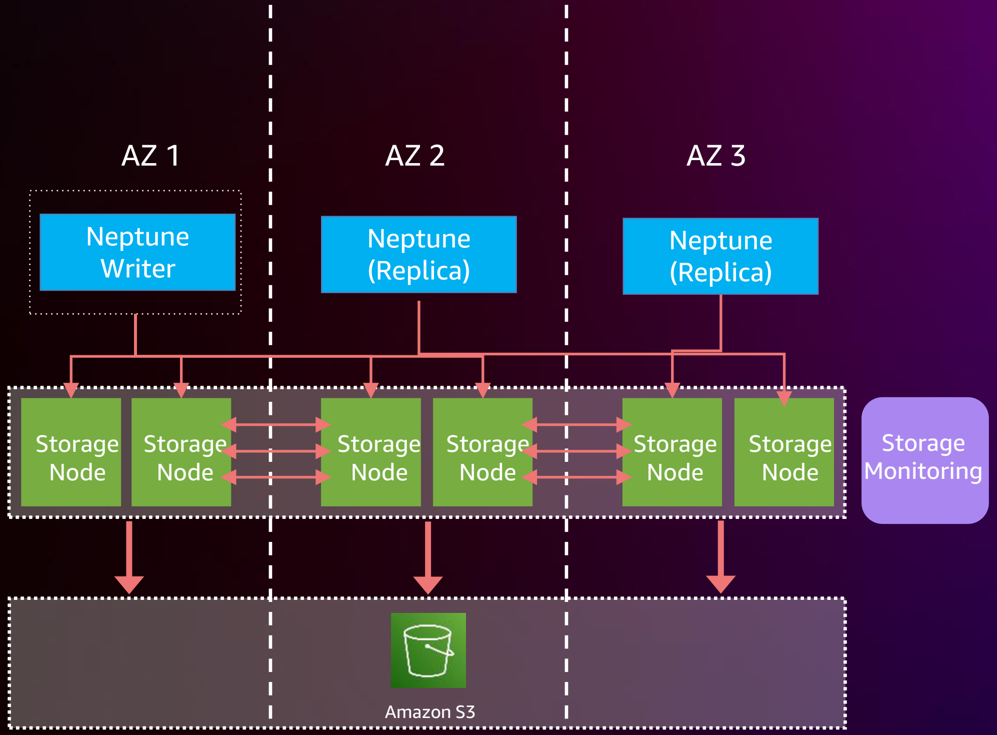
- Migrate primary clusters to new AWS Region

Low replication lag

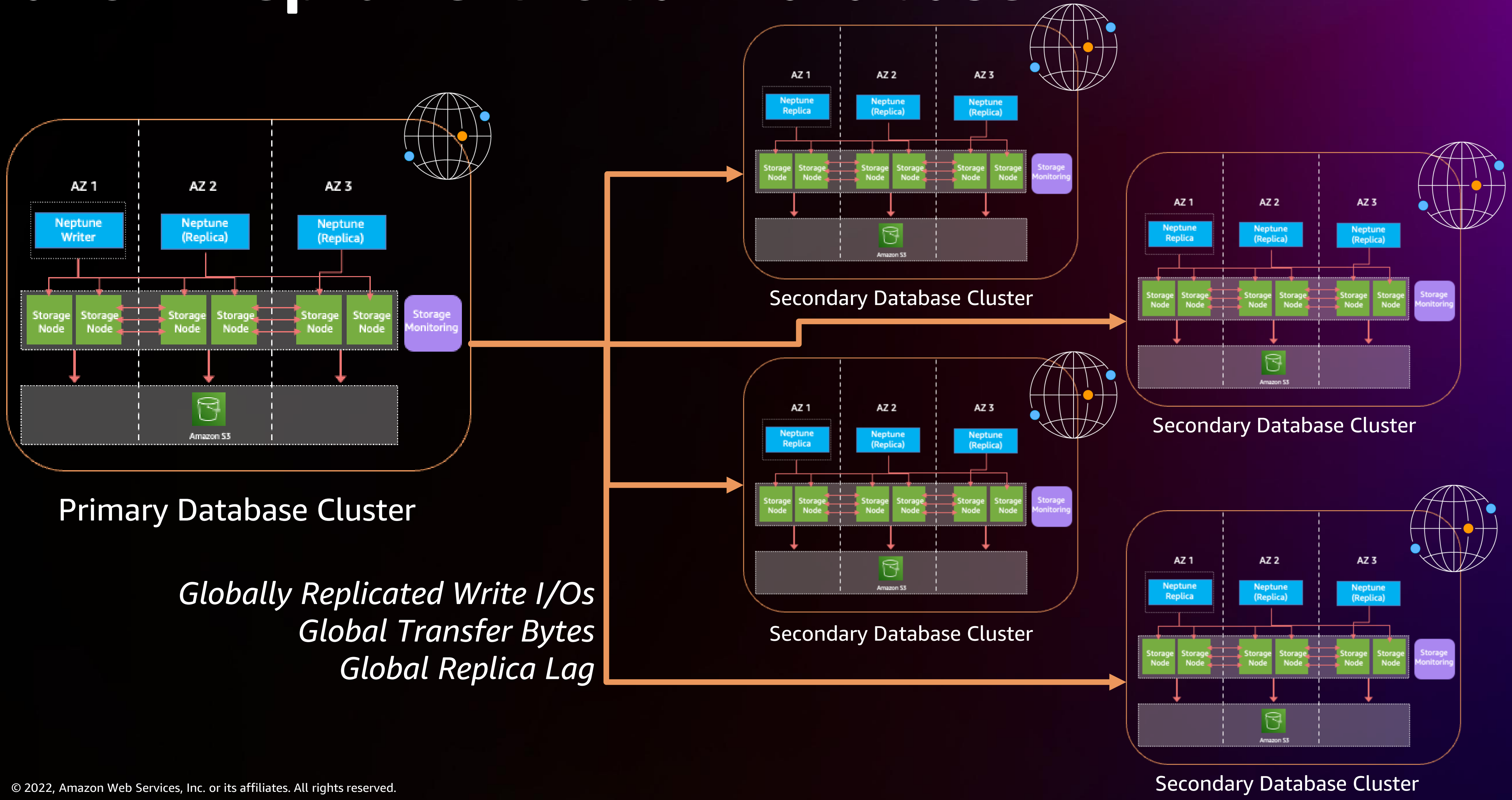
- Fast replication between AWS Regions

Amazon Neptune clusters – A quick review!

- Scale up query evaluation
  - 1 Writer; Up to 15 replicas
- Scale out storage
- Storage volume automatically grows up to 128 TiB (new in 2022!)
- Data is replicated 6 times across 3 AZs
- Continuous monitoring of nodes and disks
- 10 GB segments as unit of repair or hotspot rebalance
- Quorum system for read/write; latency tolerant
- Quorum membership changes do not stall writes
- Continuous backup to Amazon S3
  - *Built for 11 9s durability*



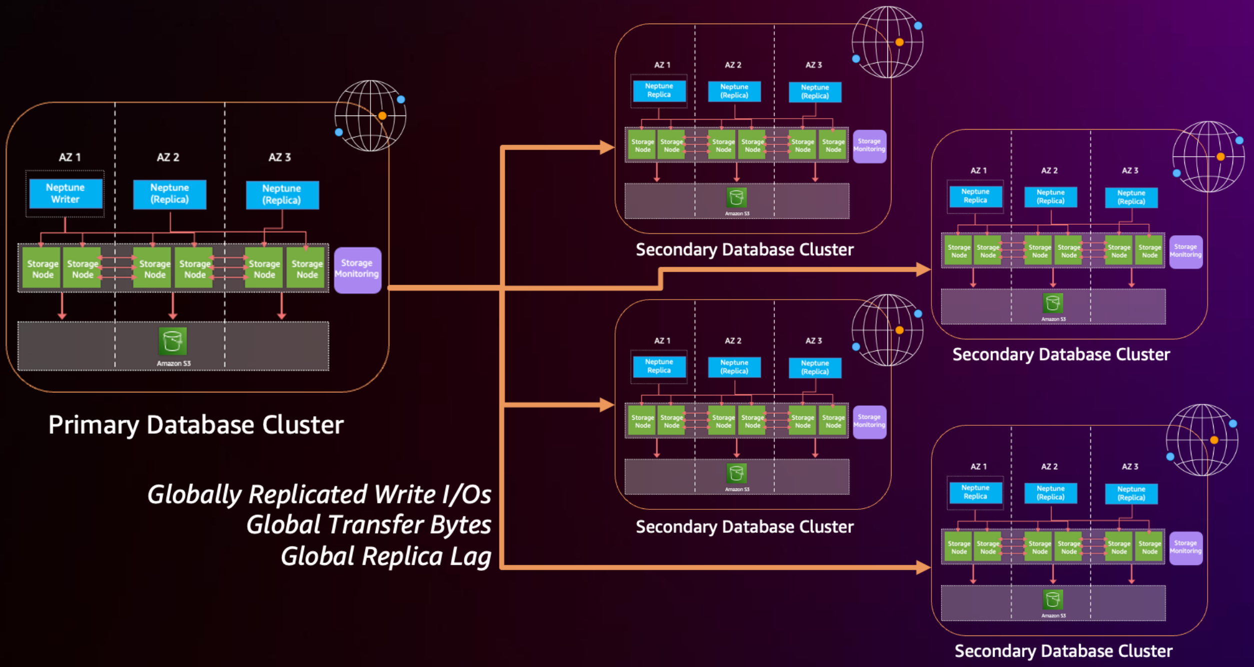
Amazon Neptune Global Database



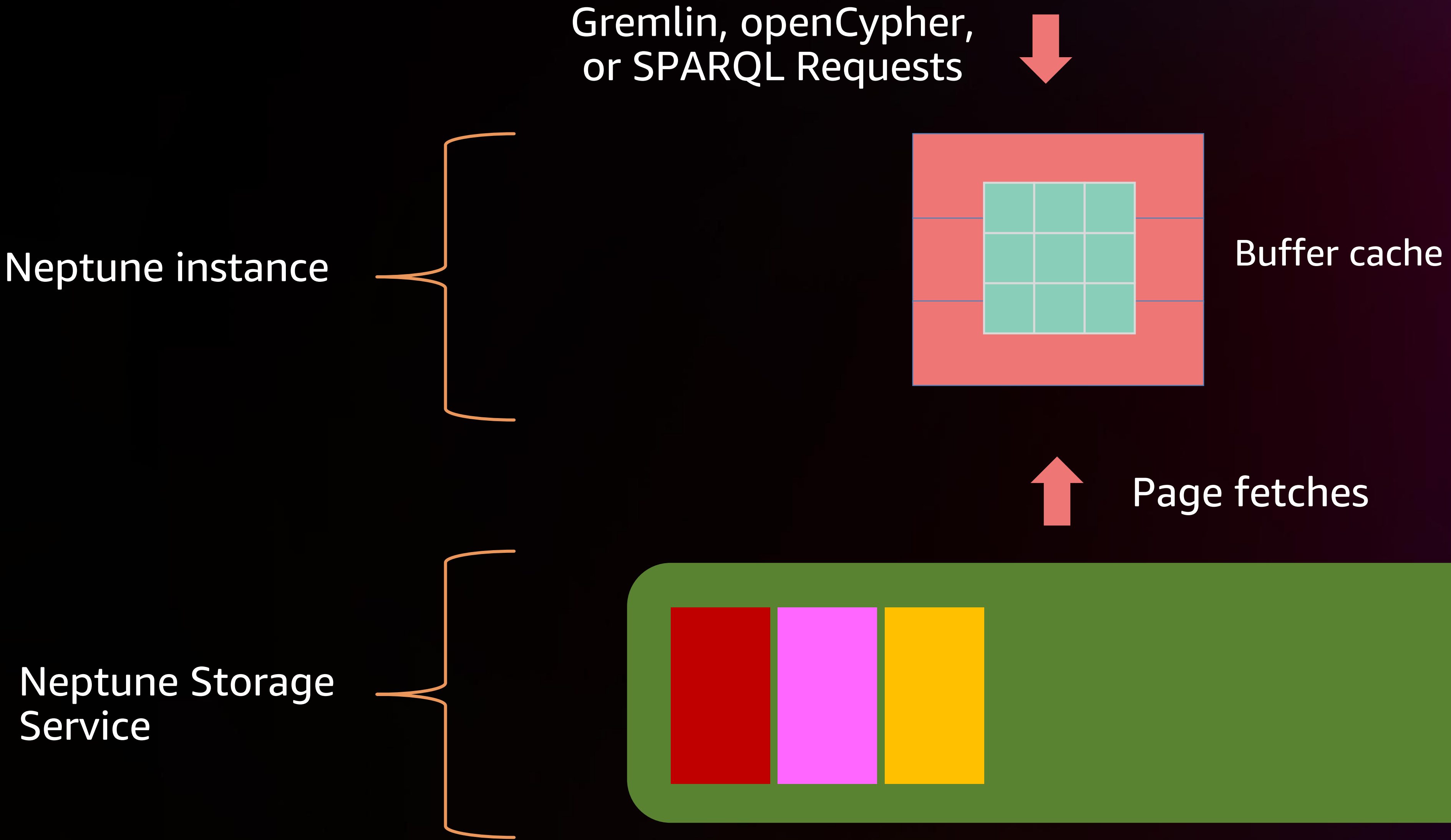
Globally Replicated Write I/Os
Global Transfer Bytes
Global Replica Lag

Amazon Neptune Global Database

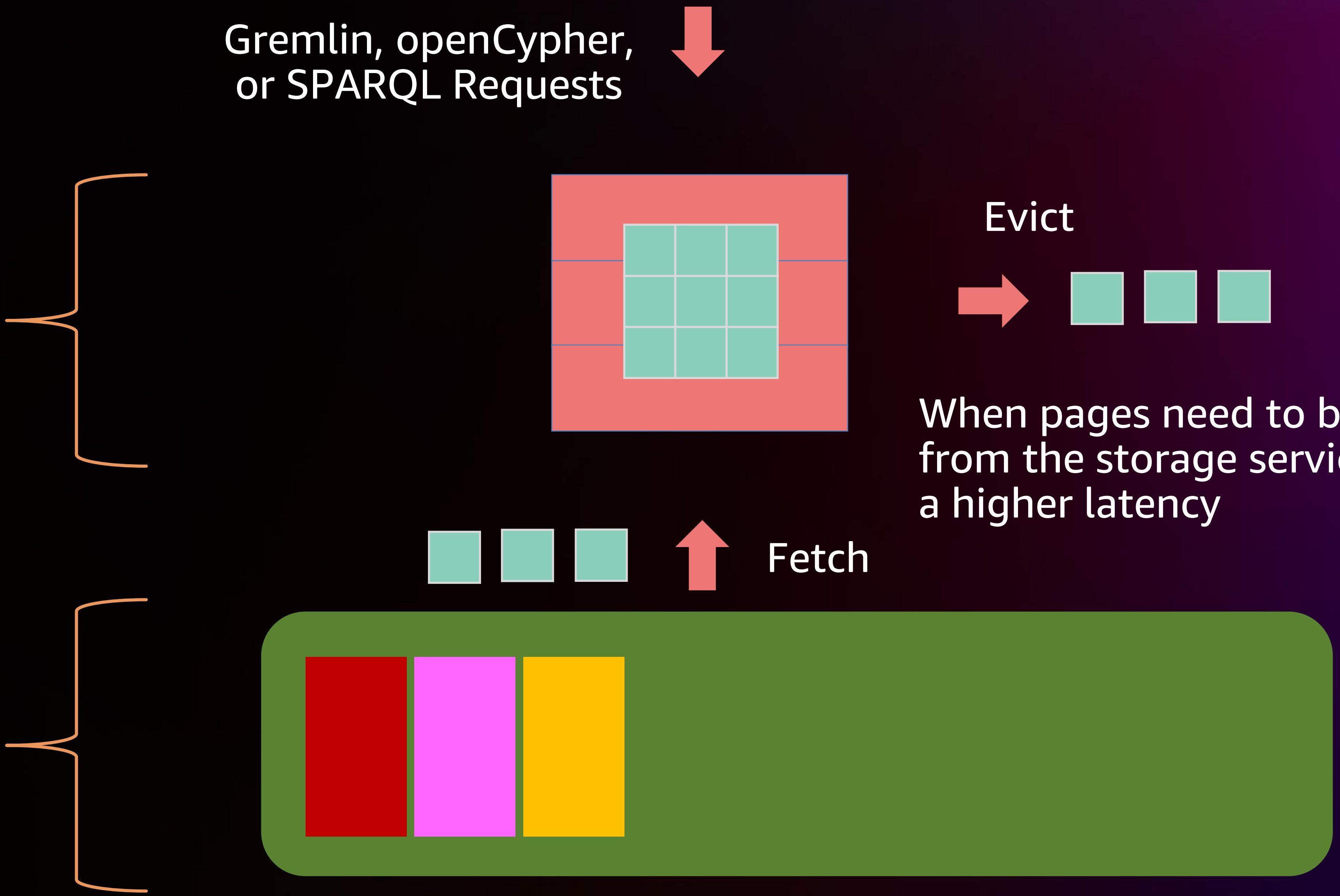
- Up to 5 Secondary Clusters in supported AWS Regions
  - Each secondary DB cluster is like a read-replica
- Supports planned and unplanned failover modes
  - Detach and promote (unplanned)
  - Managed planned failover
- Engine Release 1.2.0.0+
- 7 AWS Regions: US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Europe (Ireland), Europe (London), and Asia Pacific (Tokyo)



Evaluating queries on Neptune: Buffer cache

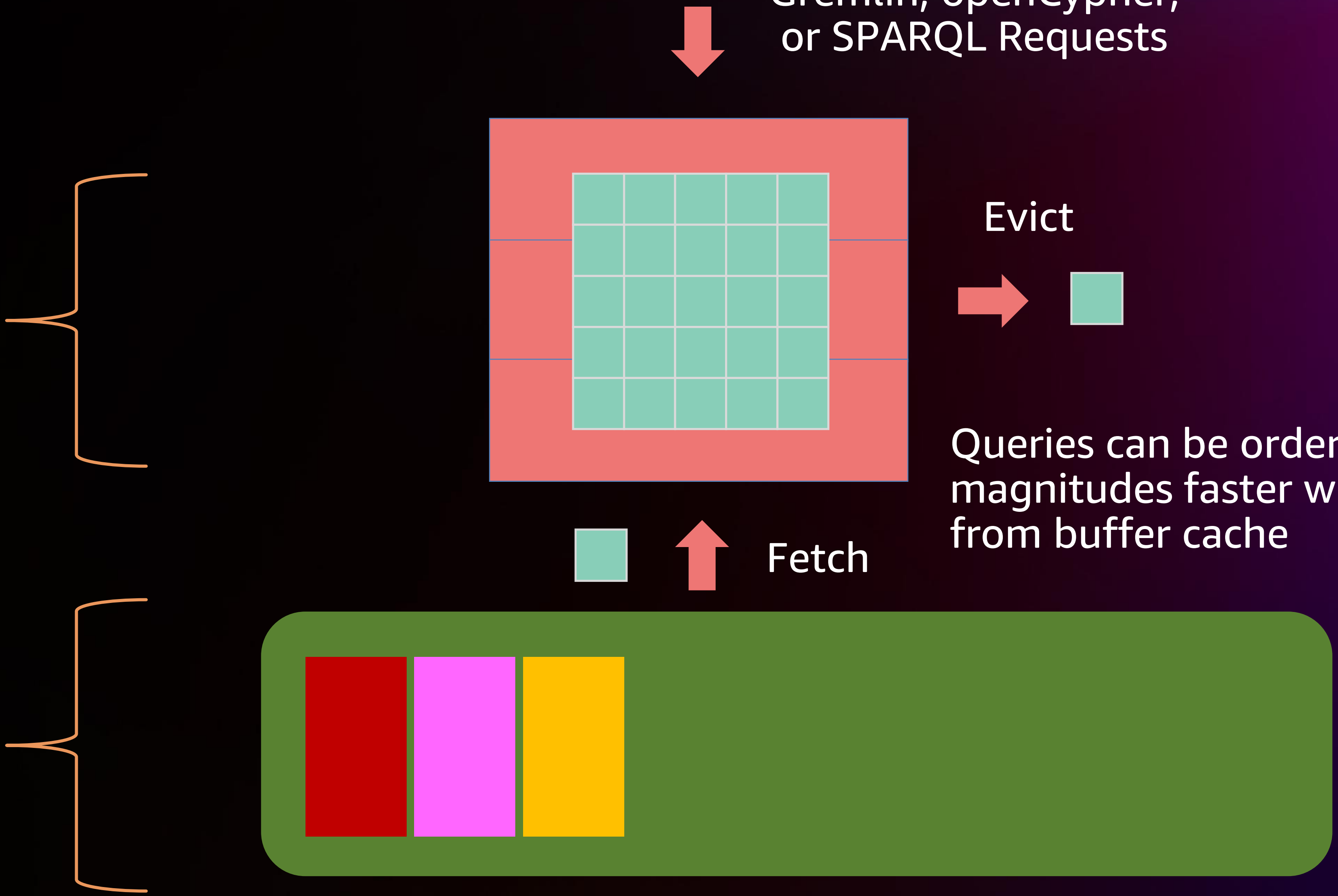


Buffer cache: Cache miss



When pages need to be retrieved from the storage service, there is a higher latency

Buffer cache: Cache hit



Neptune Instance

Neptune Storage Service

Gremlin, openCypher, or SPARQL Requests

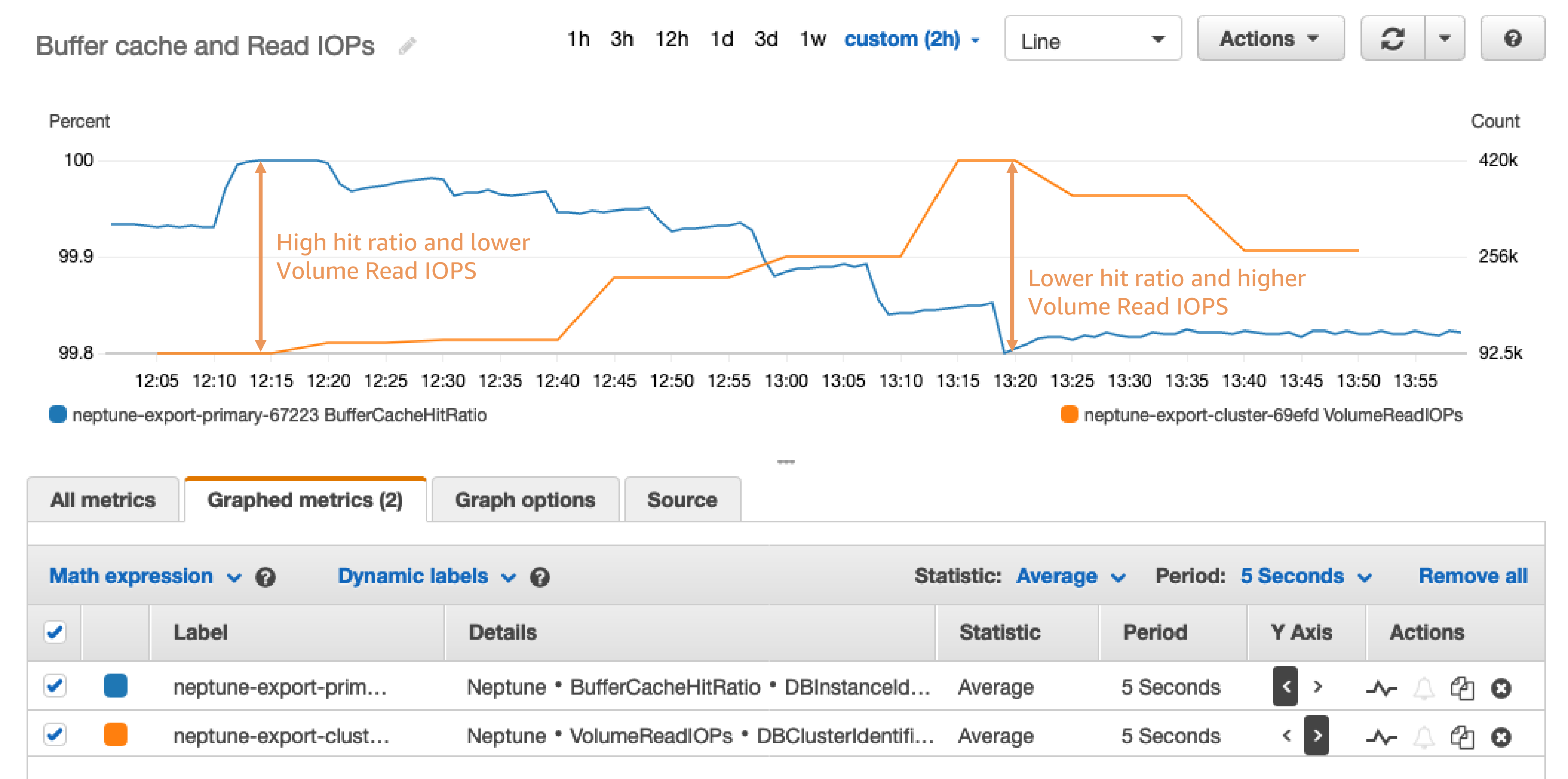
Evict

Fetch

Queries can be orders of magnitudes faster when served from buffer cache

Using the BufferCacheHitRatio for Scaling

If the cache hit ratio is below 99.9%, more memory may improve performance, e.g., larger instance.



High hit ratio and lower Volume Read IOPS

Lower hit ratio and higher Volume Read IOPS

Statistic: Average

Period: 5 Seconds

☑		Label	Details	Statistic	Period	Y Axis	Actions
☑	🟦	neptune-export-prim...	Neptune • BufferCacheHitRatio • DBInstanceId...	Average	5 Seconds	< >	
☑	🟧	neptune-export-clust...	Neptune • VolumeReadIOPs • DBClusterIdentifi...	Average	5 Seconds	< >	

We heard from customers that managing capacity was hard

Amazon Neptune Serverless

10/26/22 GA

The first serverless graph database that automatically scales database capacity up or down to optimize cost and performance.

Amazon
Neptune
Serverless

Scale Instantly

- Instantly scale capacity in a fraction of a second to meet workload demands.

Optimize performance for demanding workloads

- Scales capacity in fine-grained increments. Eliminate the complexity of configuring capacity for unpredictable or variable workloads.

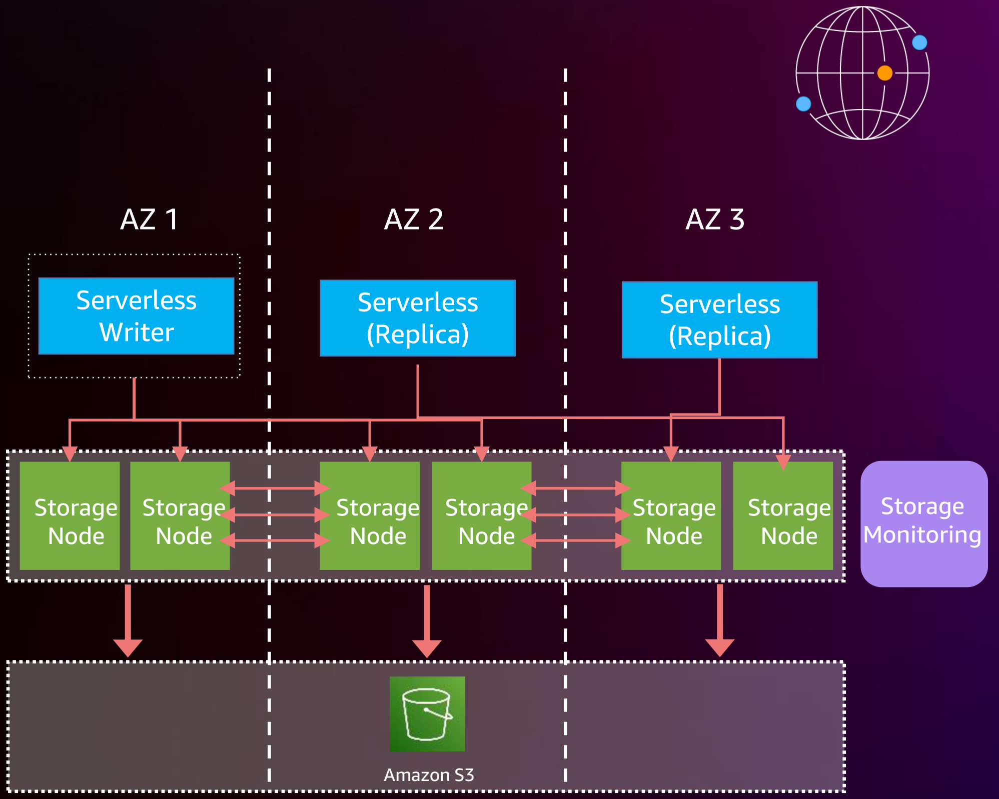
Save up to 90% on database costs

- Reduce costs by up to 90% compared to provisioning for maximum database capacity.

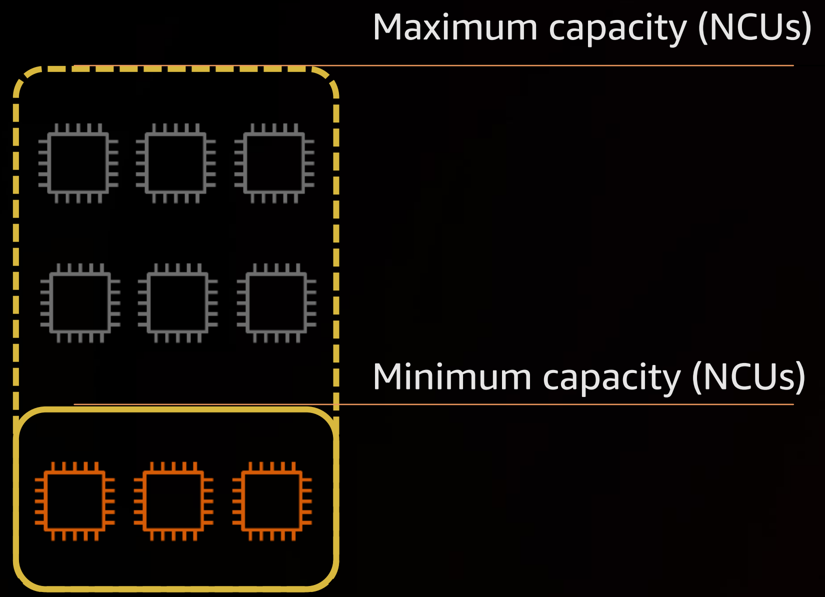
https://aws.amazon.com/neptune/serverless/

Serverless scaling and availability

- Amazon Neptune Serverless has the same query, loading, availability, and scalability features
  - Add a db.serverless instances to your cluster
  - When using a Serverless Writer, Serverless replicas in Tier 0 or Tier 1 scale with the writer to be available as fail-over targets
  - You can combine serverless and non-serverless instances in the same cluster
  - Serverless instances can be part of Neptune Global Database clusters



How is Serverless capacity managed?



Amazon Neptune Serverless (db.serverless)

- A Neptune Capacity Unit (NCU) is the measure of scaling
  - 1 NCU = 2GB RAM of capacity and proportionate CPU and network bandwidth
  - User specified min. and max.
  - System min = 2.5 NCU and max = 128 NCU (equiv. r6g.8xlarge)
- Minimum capacity determines the starting capacity of the instance
- Maximum capacity is a budget control

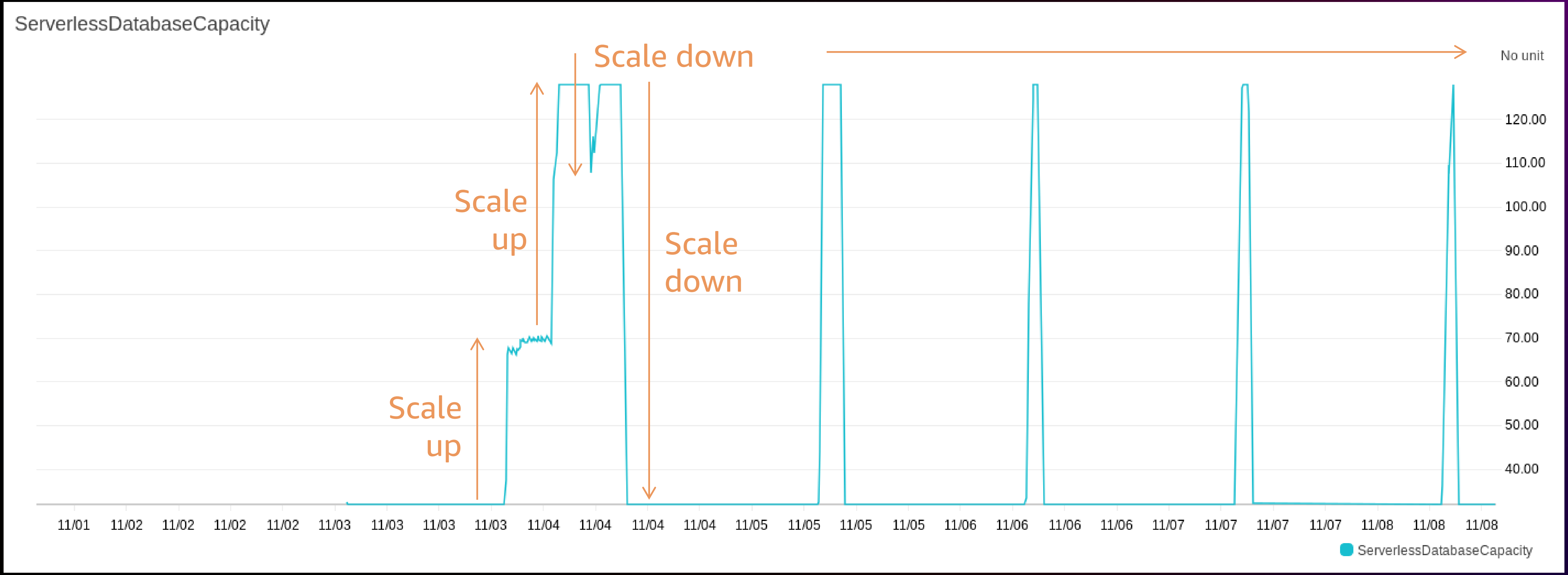
How does Neptune Serverless scale capacity?

Amazon Neptune Serverless (db.serverless)

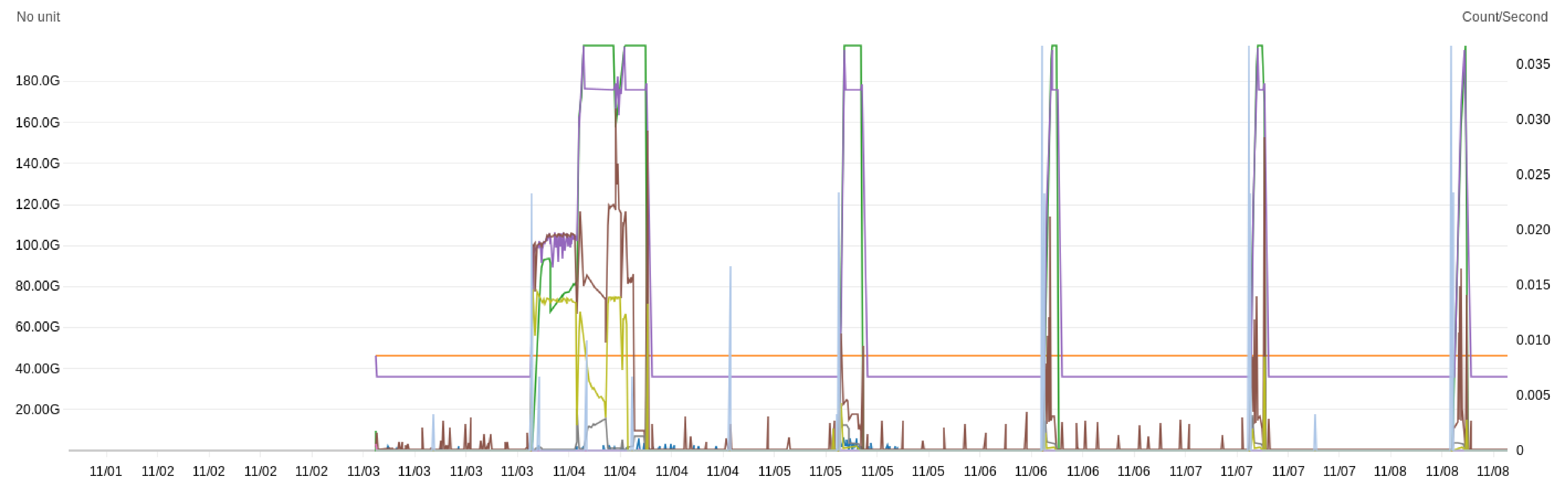
- CPU utilization of both foreground and background processes
- Memory utilization of internal data structures (e.g., buffer pool)
- Network throughput is proportional to capacity – capacity is scaled to match network throughput needs
- Scale-up rate is predictable and proportional to current capacity – larger instances scale up faster

Neptune Serverless customer scaling behavior

Scaling over time saving costs
vs. peak provisioning



How do we know when and how to scale?

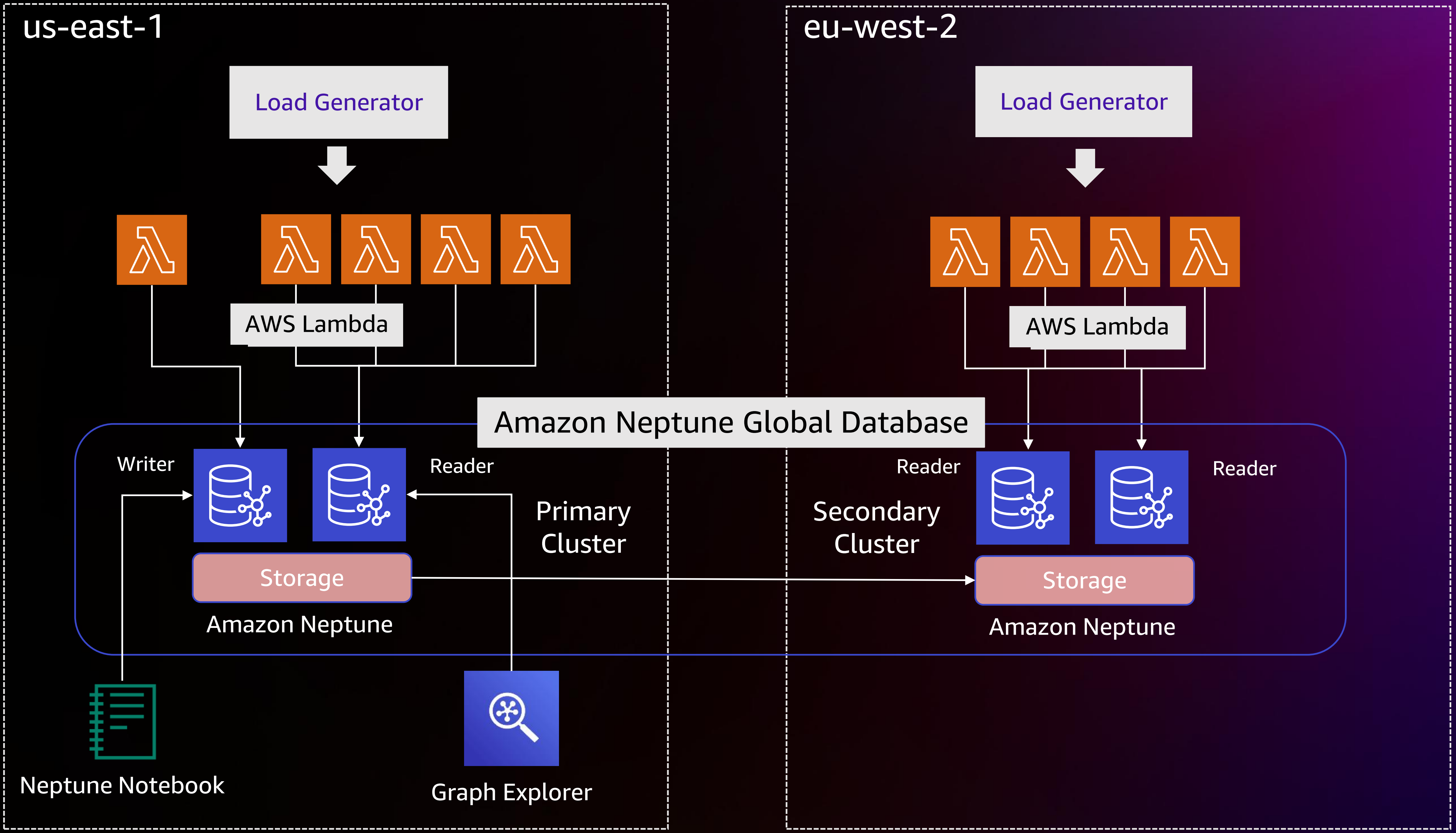


Colored lines represent different decision making processes that are monitoring memory, CPU, network utilization, min/max NCU.

A few more things on Serverless…

- When is Serverless not a good fit?
  - Memory intensive workloads that require more RAM than available in 128 NCUs / r6g.8xlarge
  - Have a highly predictable, steady-state workloads and don’t need the ability to scale based on demand
- On the horizon…
  - AWS Cloudformation and AWS Cloud Development Kit (CDK) support is coming very soon!
  - Support for lower minimum NCUs
  - Expanded Regional availability
  - And more… let us know how you’re using Serverless

Demo: Global Serverless Graph Database Cluster!


us-east-1
Load Generator
AWS Lambda
Amazon Neptune Global Database
Writer
Reader
Primary Cluster
Storage
Amazon Neptune
Neptune Notebook
Graph Explorer
eu-west-2
Load Generator
AWS Lambda
Reader
Reader
Secondary Cluster
Storage
Amazon Neptune

Related Sessions

WPS303	Secure public sector environments using graph technology	11/30 – MGM Grand – 11:30 – 12:30
DAT205	Build your first graph application with Amazon Neptune	11/30 – MGM Grand – 12:15 – 14:15
DAT302	How Wiz uses graphs to gain security insights with Amazon Neptune	11/30 – MGM Grand – 13:00 – 14:00

Additional resources

Neptune Serverless

Neptune Developer Resources

Neptune openCypher

Neptune Release Notes – Latest Features & Improvements

Neptune Global Database

Neptune Free Trial FAQs

Thank you!

Brad Bebee
beebs@amazon.com

Ian Robinson
ianrob@amazon.co.uk

Please complete the session survey in the **mobile app**