ORACLE

Oracle Cloud Scale Charging 100 Million Subscriber Performance Test

Summary of Large Scale Performance Testing

Technical Brief

January 2024, Version 1.0.1

Public

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

This document is provided for information purposes and should not be relied upon in making a purchasing decision. The contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose.

This document is not part of a license agreement nor can it be incorporated into any contractual agreement with Oracle Corporation or its subsidiaries or affiliates.

Failure to adhere to these benchmarks does not constitute a breach of Oracle's obligations. We specifically disclaim any liability with respect to this document and no contractual obligations are formed either directly or indirectly by this document.

Table of contents

Cloud Scale Charging from Oracle Communications

Oracle's Cloud Scale Charging, powered by industry leading in-memory grid technology, has been designed from the outset to support the technical and business monetization demands for hyperscale digital communications providers. It is a digital experience engine for the 5G era, providing 3GPP aligned real-time converged data and communications session charging and balance management, with native integration into Oracle's full suite of billing and revenue management capabilities designed in accordance with TM Forum principles.

Built around network and IT industry standards, Oracle Cloud Scale Charging uses an innovative high performance and coherent data management architecture to support near linear scalability, low latency, and highly available geographical redundant deployments with transactional consistency (figure 1).

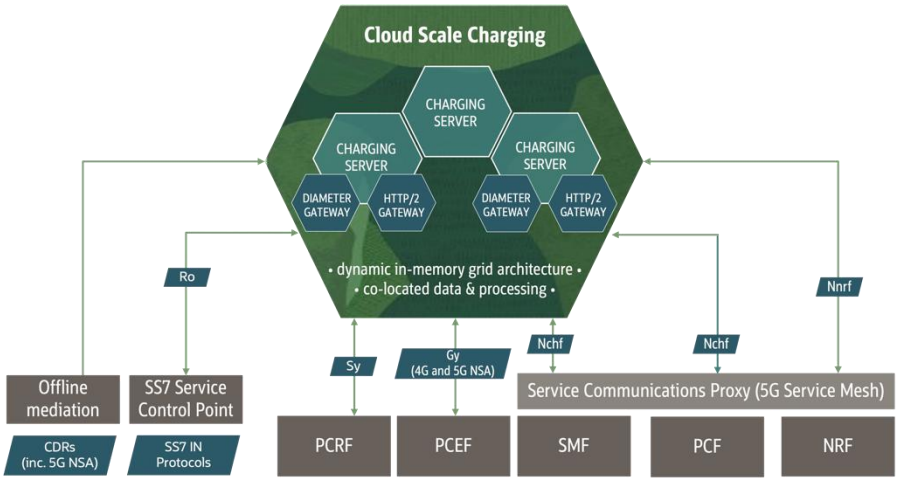


Figure 1 – Cloud Scale Charging from Oracle Communications

The real-time rating and balance management functions are underpinned by the industry leading Oracle Coherence in-memory data grid technology, forming a high performance and resilient **charging grid**. Coherence has a **dynamic mesh-based architecture** that provides fast data access and enables predictable scalability for mission critical applications.

The use of in-memory technology in modern network charging applications is essential to support the very low latency service authorization and re-authorization network requests required by service providers, typically specified in the order of milliseconds.

The Oracle Cloud Scale Charging grid adopts an innovative approach that co-locates the processing and data, offering high degrees of parallelism, with events persisted asynchronously to an enterprise class database ensuring efficient processing and low latencies.

The converged charging system can be deployed as a cloud native application in a containerized and orchestrated environment, taking advantage of modern cloud infrastructure and DevOps CI/CD tooling to enable service providers to design, test, and deploy services more quickly, operate more efficiently, and scale as business needs require.

Network grade performance, cloud scale operational experience

The Oracle Cloud Scale Charging grid stores customer data (including active session details and balances) and pricing data using in-memory cache technology distributed across a cluster of grid members (realized as JVM nodes), with data entries serialized in key-value pairs. Read and write latencies are extremely small, supporting very low end-to-end charging transaction response times for data session initiate, update, and terminate requests. The Oracle Cloud Scale Charging system uses Coherence distributed caching for storing customer objects across members of the charging grid with automatic partitioning and rebalancing of data as new members are added or removed from the grid.

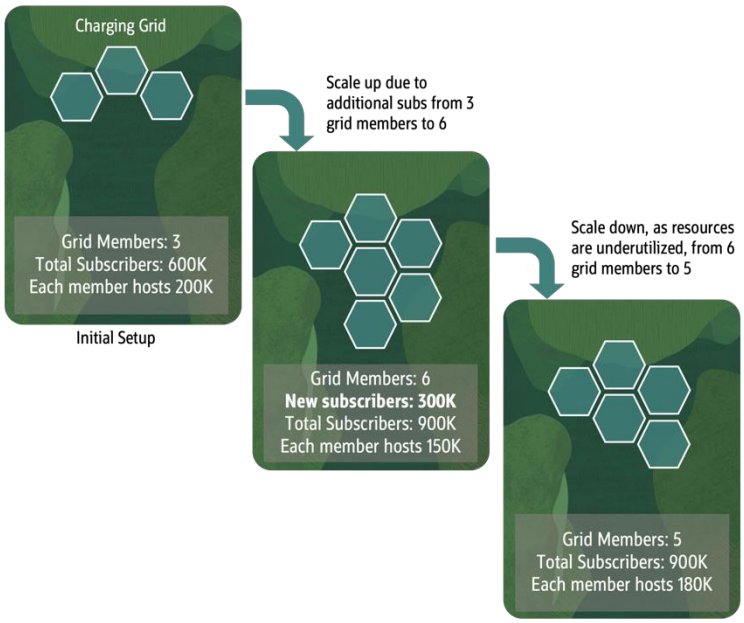
Rather than taking the approach of fetching data from a remote store, performing processing, and then writing the data back to the remote store, the Oracle Cloud Scale charging grid processes all charging transaction

requests directly where the data entries are managed in the cluster. This co-located data and processing affinity architecture offers the following benefits:

- Processing is extremely fast as all objects are held in-memory, ensuring low latency and cost-efficient compute resource utilization and high charging transaction throughput
- Data access times are close to zero, with processing invoking optimized HashMap lookups
- Almost zero cost locking, retaining transactional data consistency and ensuring no revenue leakage within the charging system

Asynchronous persistence of the grid cache ensures high performance without compromising business-critical data availability. Rated events are offloaded asynchronously to revenue management functions providing a near real-time event flow that does not impact the core network charging processing. Off grid persistence of customer account and pricing data is stored in an Oracle database that underpins a complete suite of pre-integrated billing and revenue management applications.

The grid is fully distributed, with no single point of contention, supporting independent scalability for large and growing customer data sets. The charging grid supports near linear scalability due to the automatic partitioning of customer data objects across the grid members. Coherence detects new grid members and automatically re-balances the cache data so that it is spread evenly across the grid (figure 2).



- Coherence based charging grid
- **Collocated data and operations** for increased efficiency and throughput, with low latency
- Configured to **keep-a-copy** for high availability of data
- Supports **scaling up or down** dynamically
- Grid **rebalances** automatically every time a new member is added or removed to distribute subscribers evenly across the grid members

Figure 2 – Oracle Cloud Scale Charging elastic scaling model

Dynamic scaling up or down can be handled "in-flight" to support changes in presented traffic load, subscriber growth, or compute availability, for example to change roles between test and production compute for efficient resource utilization.

Business continuity architecture

In a geographical redundancy deployment model, charging engine sites can concurrently process network charging requests across two or three geographic locations [1]. This architecture is designed to provide a very high level of resiliency and service continuity in the event of unplanned outages. All updates that occur in the charging cluster in a site are replicated to other sites using Coherence cache federation (figure 3).

[1] 3-site deployment architectures are subject to validation by Oracle based upon the specific customer deployment requirements.

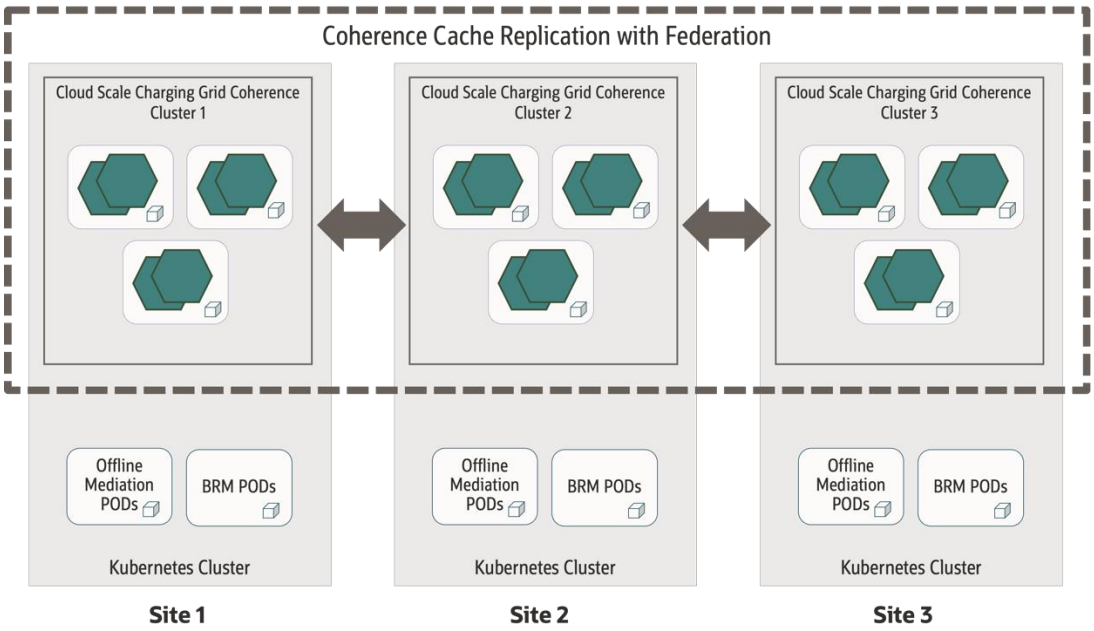


Figure 3 – Charging grid cache federation across multiple sites

In this configuration, all sites are active with each site housing complete customer and configuration data. Requests can be sent to any site; the core network functions need not know the preferred site for a subscriber. Responses are routed back through the site that originally received the request from the core network.

In case of a charging grid outage at a specific site, the remaining sites can continue to process the usage requests and share the load of the failed site, until the site is brought back up and ready to process again.

In an ideal geographically redundant deployment, each charging grid should handle an evenly balanced traffic load, for example in a dual site configuration each charging grid should handle approximately 50% of the traffic and in a three-site configuration each grid should handle approximately 33% of the traffic.

100 Million subscriber performance test

Building on the previous 50 million subscriber test, Oracle conducted single and dual site charging system performance testing on Oracle Cloud Infrastructure for a test subscriber base of 100 million subscribers. The single site test was conducted against a charging cluster deployed in an east coast US OCI region (**US East – Ashburn, Virginia**), with the dual site test conducted against a converged charging system deployment consisting of two active charging clusters deployed in west and east coast OCI regions (**US West – Phoenix, Arizona** and **US East – Ashburn, Virginia)**.

Oracle Cloud Scale Charging demonstrated compelling performance characteristics for both single and dual site deployments.

Test setup

Methodology

Mixed charging traffic generated from a core network simulator (Seagull) was presented to a Cloud Scale Charging deployment with 100 million provisioned accounts, consisting of 90% prepaid and 10% postpaid user profiles, with recorded observations of charging transaction throughput, latency, and resource utilization. Additionally, to demonstrate successful charging grid federation at scale, simulated network charging traffic was presented to a dual site deployment spanning the east and west coast United States. In both test runs, rated events were cached in memory, written to persistent storage, and transformed into CDR files.

Configured price plans

As for the previous 50 million subscriber test, two price plans (one for prepaid and one postpaid) were deployed representative of realistic commercial offers covering voice, data and messaging including in-bundle and out-of-bundle tariffs, special rate friends and family numbers, a postpaid monthly cycle forward fee and add-on bundles with validity dates.

Traffic profile

The charging traffic generated from the test clients towards the Cloud Scale Charging network gateway PODs simulated a mixture of initiate, update and terminate charging operations across data, voice and messaging, including balance enquires and also triggered notifications from the charging system (table 1).

TRAFFIC TYPE	DESCRIPTION
Voice 1	**Initiate** and **Terminate** operations (60 seconds duration)
Voice 2	**Initiate**, **Update** and **Terminate** operations (180 seconds duration, 90 seconds between operations)
Data	**Initiate**, **Update** and **Terminate** operations (24 minutes duration, 180 seconds between operations)
SMS	**Terminate** operation only
Balance enquiries	Generated from the Seagull Diameter client
Notifications	Pre and post call notifications published on Kafka topics

Table 1 –Test traffic profile summary

Diameter traffic was used to reflect the predominant charging traffic deployed in today's mobile networks. Similar performance is expected for 5G standalone core network charging.

Software environment

The application software under test consisted of:

- BRM 12.0.0.4.0 (cloud native deployment)
- ECE 12.0.0.4.0 (cloud native deployment with cache persistence enabled)
- Oracle Database 19c EE Extreme Perf Release 19.0.0.0.0 – Production Version 19.11.0.0.0 (Ashburn OCI region)
- Oracle Database 19c EE Extreme Perf Release 19.0.0.0.0 – Production Version 19.12.0.0.0 (Phoenix OCI region)

Deployed on:

- Kubernetes 1.17.9 (Ashburn OCI region)
- Kubernetes 1.18.10 (Phoenix OCI region)
- Docker 19.03.11-ol
- Helm 3.0.1
- Oracle Linux 7.8

Hardware environment and deployment architectures

Single Site Test – Ashburn OCI region

The Ashburn site included:

- The test driver, responsible for generating charging operation requests, was deployed on four standard 8-core virtual machine shapes (VM.Standard.B1.8[2])
- The Oracle converged charging system Kubernetes cluster (Oracle Cloud Infrastructure Container Engine for Kubernetes) deployed on 102 standard 16-core virtual machine shapes (VM.Standard.B1.16[1])

[2] https://docs.oracle.com/en-us/iaas/Content/Compute/References/computeshapes.htm#

- The revenue management layer persistence database deployed in a full rack Exadata X7-2 hosting eight RAC nodes

The details of the infrastructure setup for the single site 100 million subscriber test are summarized in table 2.

	HARDWARE PLATFORM	SHAPE	QUANTITY	NOTES
Test Client	X6	VM.Standard.B1.8 8 cores, 96 GB RAM, 4.8 Gbps maximum network bandwidth	4	Intel(R) Xeon(R) CPU E5-2699C v4 @ 2.20GHz base frequency Total: 32 OCPUs (64 vCPUs), 384 GB RAM
Charging Application Under Test	X6	VM.Standard.B1.16 16 cores, 192 GB RAM, 9.6 Gbps maximum network bandwidth	102 Consisting of 99 VMs running 371 Elastic Charging Server pods and 96 network gateway pods plus 3 VMs running Kafka, BRM, PDC and Rated Event Formatter pods	Intel(R) Xeon(R) CPU E5-2699C v4 @ 2.20 GHz base frequency Total: 1,632 OCPUs (3264 vCPUs), 19,584 GB RAM
Persistence Database	X7-2	Exadata X7-2	1	8 out of 8 available RAC nodes used Total: 368 OCPUs (736 vCPUs), 5760 GB RAM

Table 2 – Performance test cloud infrastructure details (100 million subscribers, single site test run)

The deployment architecture, hosted on Oracle Cloud Infrastructure, is depicted in figure 4.

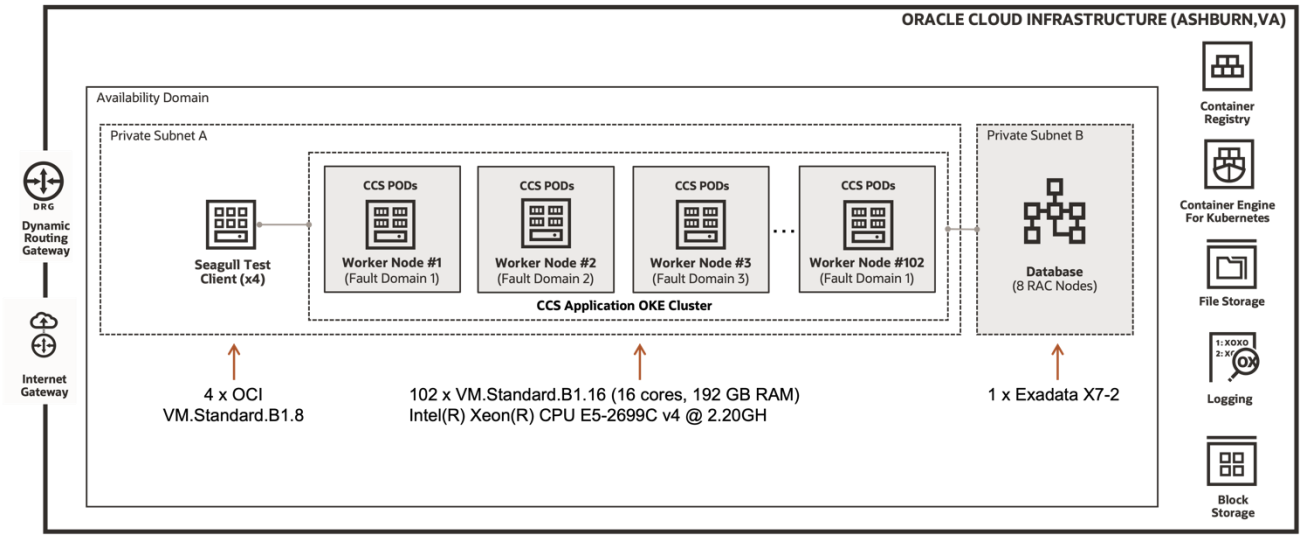


Figure 4 –Performance test cloud infrastructure single site deployment architecture

Dual site test

US East – Ashburn OCI region

The Ashburn OCI dual site test configuration was the same as for the single site test, with the exception that two test client VM's were deployed instead of four.

US West – Phoenix OCI region

The Phoenix site included:

- The test driver, responsible for generating charging operation requests, deployed on two standard 8-core virtual machine shapes (VM.Standard2.8[1])
- The Oracle converged charging system Kubernetes cluster (Oracle Cloud Infrastructure Container Engine for Kubernetes) deployed on 64 standard 24-core virtual machine shapes (VM.Standard2.24[1])

- The revenue management layer persistence database deployed in a full rack Exadata X7-2 hosting eight RAC nodes

The details of the infrastructure setup for the Phoenix site are summarized in table 3.

	HARDWARE PLATFORM	SHAPE	QUANTITY	NOTES
Test Client	X7	VM.Standard2.8 8 cores, 120 GB RAM, 8.2 Gbps maximum network bandwidth	2	Intel(R) Xeon(R) Platinum 8167M CPU @ 2.00GHz base frequency Total: 16 OCPUs (32 vCPUs), 240 GB RAM
Charging Application Under Test	X7	VM.Standard2.24 24 cores, 320 GB, 24.6 Gbps maximum network bandwidth	64 Consisting of 61 VMs running 371 Elastic Charging Server pods and 96 network gateway pods plus 3 VMs running Kafka, BRM, PDC and Rated Event Formatter pods	Intel(R) Xeon(R) Platinum 8167M CPU @ 2.00GHz base frequency Total: 1536 OCPUs (3072 vCPUs) 20,480 GB RAM
Persistence Database	X8-2	Exadata X8-2	1	8 out of 8 available RAC nodes used Total: 400 OCPUs (800 vCPUs), 5760 GB RAM

Table 3 – US West (Phoenix) OCI region test cloud infrastructure details (dual site test run)

Dual site deployment architecture

In the dual site active-active test architecture, each site had 100% of the test subscriber account data maintained in the charging grid Coherence cache via bi-directional federation. Each site processed traffic for 50% of the subscriber base (50M subscribers). Figure 5 shows the end-to-end dual site test deployment architecture.

Note that all the OCI commercial regions are interconnected by the Oracle Cloud Infrastructure backbone designed to allow customer workloads to move between regions in diverse geographic locations through an encrypted and reliable connection. The backbone network provides privately routed inter-region connectivity with consistent inter-region performance for bandwidth, latency, and jitter.

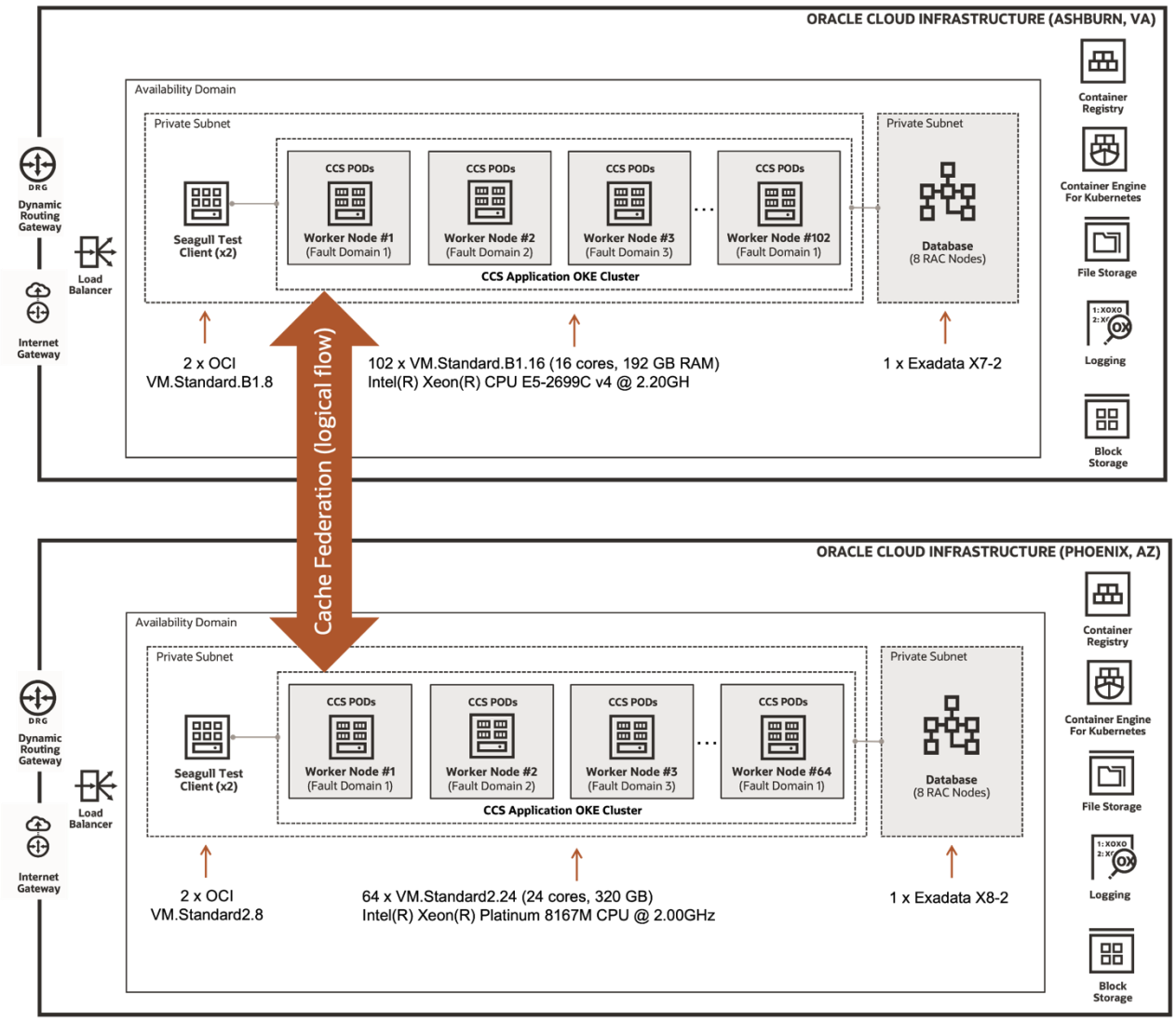


Figure 5 – Dual site test end to end deployment architecture

Test results

In both test runs Oracle Cloud Scale Charging demonstrated single digit millisecond core charging latencies, high transaction throughput and efficient resource utilization. A 100% success rate was achieved for all traffic presented to the system. The dual site test demonstrated successful charging grid bi-directional federation whilst retaining the throughput and latency performance characteristics of the single site test. For both tests rated events were successfully formatted into CDRs (call detail records).

Single site

Throughput and Latency

Table 4 summarizes the single site transaction throughput and average observed latencies (all in single digit milliseconds), measured as a roundtrip between the network gateway internal charging requests and the core elastic charging server (ecs) instances. Note that latency is not applicable for the notification traffic and rated events, as these were initiated by the Oracle Cloud Scale Charging system.

TRAFFIC TYPE	TRANSACTIONS PER SECOND (TPS)	AVERAGE LATENCY (MILLISECONDS)
SMS	5,768	6.0
Voice	41,900	7.0
Data	152,000	8.6
Notifications	40,200	N/A
Balance queries	11,200	2.5
Rated events generated	41,400	N/A

Table 4 – Observed throughput and average latencies (100 million subscribers – single site)

Resource utilization

Resource utilization observed across the core converged charging system application and the Oracle database used for persistence during the steady state (maximum) traffic phase of the test is shown in table 5.

Average App CPU utilization	53%
Average DB CPU utilization	21%
Average App memory utilization	52%
Average DB memory utilization	55%

Table 5 – Observed resource utilization (100 million subscribers – single site)

Dual site

Throughput and latency

Table 6 summarizes the dual site transaction throughput and average observed latencies (all in single digit milliseconds), measured as a roundtrip between the network gateway internal charging requests and the core elastic charging server (ecs) instances.

Phoenix			Ashburn			Both Sites
TRAFFIC TYPE	TRANSACTIONS PER SECOND (TPS)	AVERAGE LATENCY (MILLISECONDS)	TRAFFIC TYPE	TRANSACTIONS PER SECOND (TPS)	AVERAGE LATENCY (MILLISECONDS)	TRANSACTIONS PER SECOND (TPS)
SMS	2,900	5.98	SMS	2,900	4.85	5,800
Voice	21,000	6.40	Voice	21,000	5.31	42,000
Data	75,500	6.52	Data	75,500	5.95	151,000
Notifications	30,081	n/a	Notifications	30,102	n/a	60,183
Balance Queries	5,591	2.53	Balance Queries	5,591	1.89	11,182
Rated Events Generated	20,700	n/a	Rated Events Generated	20,700	n/a	41,400

Table 6 – Observed throughput and average latencies (100 million subscribers – dual site)

The results compare favorably with the single site test measurements when combined across both sites (the results shown in green in table 6). Also, each site is comparable to the results achieved in the previous 50M test (published in March 2021).

The latencies for the Phoenix site are slightly higher than Ashburn as fewer worker nodes were deployed (64 vs 102, on newer generation hardware) which were subject to higher load levels.

Resource utilization

Resource utilization observed across the core converged charging system application and the Oracle database used for persistence during the steady state (maximum) traffic phase of the test is shown in table 7.

	RESOURCE UTILIZATION (PHOENIX)	RESOURCE UTILIZATION (ASHBURN)
Average App CPU utilization	51%	44%
Average DB CPU utilization	23%	20%
Average App memory utilization	54%	61%
Average DB memory utilization	54%	53%
Average Coherence federation bandwidth	4.92 Gbits/sec	4.85 Gbits/sec

Table 7 – Observed resource utilization (100 million subscribers – dual site)

Rated events

During the steady state period captured, each charging cluster locally created 20.7K rated events per second during usage event processing. These rated events were cached in memory, written to persistent storage, and transformed into CDR files. In addition, each charging cluster received an additional 20.7K rated events per second (federated from the other charging cluster) which were cached in memory and written to persistent storage for high availability purposes.

Summary

The 5G era presents new challenges for digital service providers to efficiently monetize high volumes of communications, data and media traffic and at the same time provide a compelling customer experience. Modern charging systems will be required to deliver high degrees of operational efficiency while at the same time functioning as a real-time experience engine for end users. Oracle Cloud Scale Charging, designed from the ground up to support the future needs of hyperscale service providers, is a cloud native 5G ready monetization platform that meets these needs (figure 6).

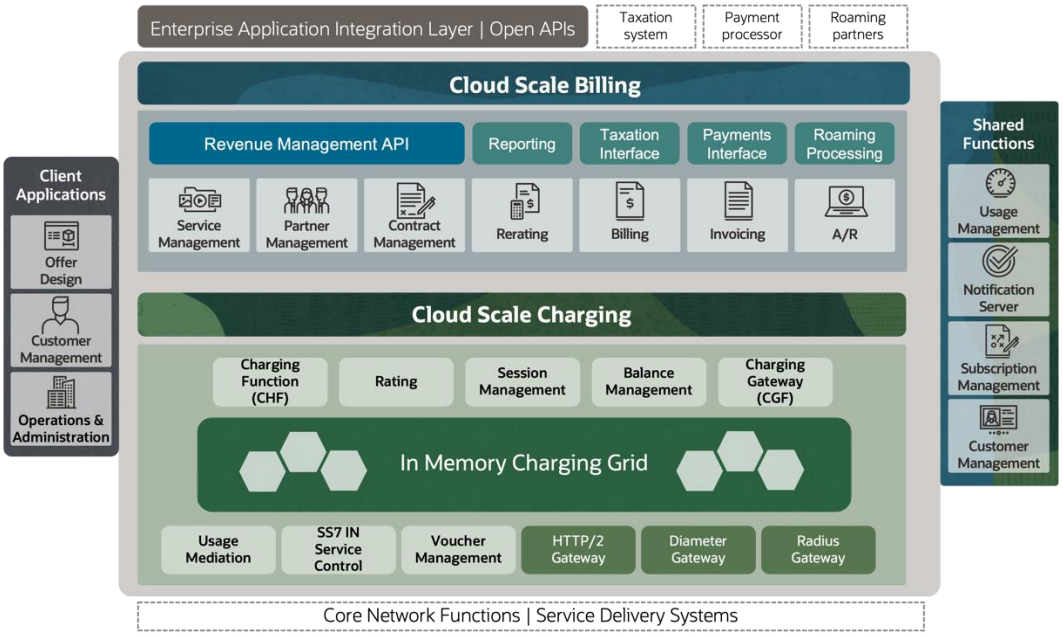


Figure 6 – Cloud Scale Charging and Billing from Oracle Communications

Designed to operate at the intersection of core network and IT domains, the Oracle Cloud Scale Charging uses mesh-based in-memory technology to provide high performance, resilient and linearly scalable charging, with pre-integrations available to advanced revenue management capabilities.

Using a provisioned base of 100 million subscribers (90% prepaid, 10% postpaid) provisioned with industry realistic price plans, this Oracle Cloud Scale Charging test, undertaken on Oracle Cloud Infrastructure,

demonstrated low core charging latencies, high transaction throughput and efficient resource utilization. Specifically, this test highlighted:

- Single digit millisecond latency for 100 million subscribers in a single charging grid site
- Active-active wide area, high performance deployment
- Large scale rated event generation