ORACLE
Communications

Oracle CCS Performance Test March 2021

Summary of 50 Million Subscriber Performance Test

Technical Brief

March 2021 | Version 1.0

SAFE HARBOR STATEMENT

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

CONVERGED CHARGING FROM ORACLE

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

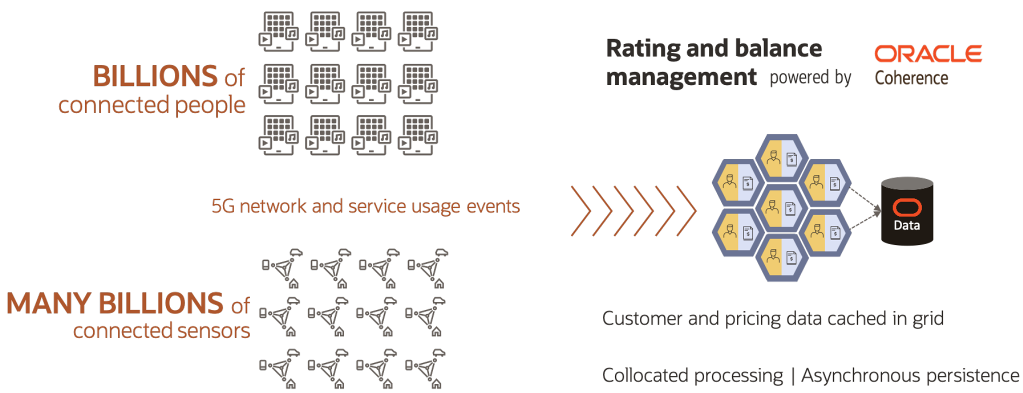


Figure 1 – Oracle's charging grid Technology

Built around network and IT industry standards, the Oracle CCS uses an innovative high performance and coherent data management architecture to support near linear scalability, low latency, and highly available multi-site deployment with transactional consistency (figure 1).

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** which enables a webscale experience. 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, typically specified in the order of milliseconds.

The Oracle CCS 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 CCS can be configured to run as a cloud native application in a containerized and orchestrated deployment architecture, taking advantage of cloud native 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 grow.

Network Grade Performance, Web Scale Operational Experience

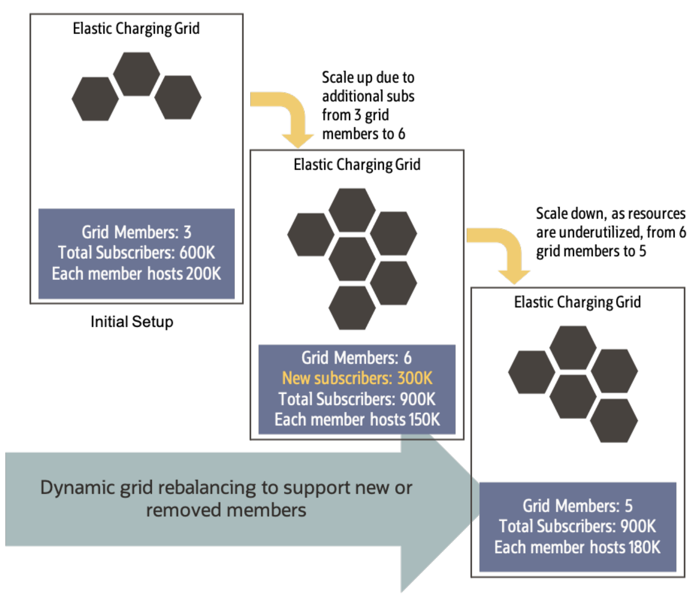
The Oracle CCS charging grid stores customer data (including active session details and balances) and pricing data using in-memory cache technology distributed across a cluster of grid members (realized as JVM nodes), with data entries serialized in key-value pairs. Read and write latencies are extremely small, supporting very low end-to-end charging transaction response times for data session initiate, update, and terminate requests. The Oracle CCS 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 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 Oracle CCS 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 – CCS 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.

SCALABILITY AND PERFORMANCE TEST RESULTS

In a recent test conducted on Oracle Cloud Infrastructure with a provisioned test subscriber base of **50 million mobile subscribers**, the Oracle Converged Charging System demonstrated compelling performance and scalability characteristics.

Test Setup

Methodology

Mixed charging traffic generated from a core network simulator (Seagull) was presented to a cloud native CCS deployment with 50 million provisioned accounts, consisting of 90% prepaid and 10% postpaid user profiles, with recorded observations of charging transaction throughout, latency and resource utilization. Additionally, in order to demonstrate linearity of scaling, tests were conducted on a smaller test base of 12.5 million subscribers for comparative purposes.

The 50 million accounts were provisioned across 4 database schemas hosted by 4 RAC nodes, with the 12.5 million subscriber test case accounts provisioned on a single schema and RAC node.

Configured Price Plans

Two price plans (one for prepaid and one postpaid) were deployed in the test configured based on 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 client towards the CCS network gateway PODs simulated a mixture of initiate, update and terminate charging operations across data, voice and messaging, including balance enquires and triggered notifications (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

Software Environment

The application software under test consisted of:

- BRM 12.0.0.3.4 (cloud native deployment)
- ECE 12.0.0.3.4 (cloud native deployment with cache persistence enabled)
- Oracle Database 19.8.0.0 (4 node RAC deployment for the 50 million subscriber test case, single node RAC deployment for the 15 million subscriber test case)

Deployed on:

- Kubernetes 1.17.9
- Docker 19.03.11
- Helm 3.0.1
- Oracle Linux 7.8

Hardware environment and deployment architecture

The test driver, generating charging operation requests, was deployed on two standard 8-core virtual machine shapes (VM.Standard.B1.8), with the Oracle converged charging system Kubernetes cluster (Oracle Cloud Infrastructure Container Engine for Kubernetes) deployed on 51 standard 16-core virtual machine shapes (VM.Standard.B1.16). The revenue management layer persistence database was deployed in a full rack Exadata X7-2 hosting four RAC nodes. The details of the infrastructure setup for the 50 million subscriber test are summarized in table 2.

	PROCESSOR	SHAPE	QUANTITY	NOTES
Test Client	X6	VM.Standard.B1.8 8 cores, 96 GB RAM, 4.8 Gbps	2	Intel(R) Xeon(R) CPU E5-2699C v4 @ 2.20GHz Total: 16 cores, 192 GB RAM
CCS Application Under Test	X6	VM.Standard.B1.16 16 cores, 192 GB RAM, 9.6 Gbps	51 (running 186 Elastic Charging Server pods, 48 network gateway pods, 3 Kafka pods and additional BRM pods)	Intel(R) Xeon(R) CPU E5-2699C v4 @ 2.20GHz Total: 816 cores, 9792 GB RAM
Persistence Database	X7-2	Exadata X7-2	1	4 out of 8 available RAC nodes used Total: 184 cores, 2880 GB RAM

Table 2 – Scalability and Performance test cloud infrastructure details (50 million subscribers)

The deployment architecture, hosted on Oracle Cloud Infrastructure, is depicted in figure 3.

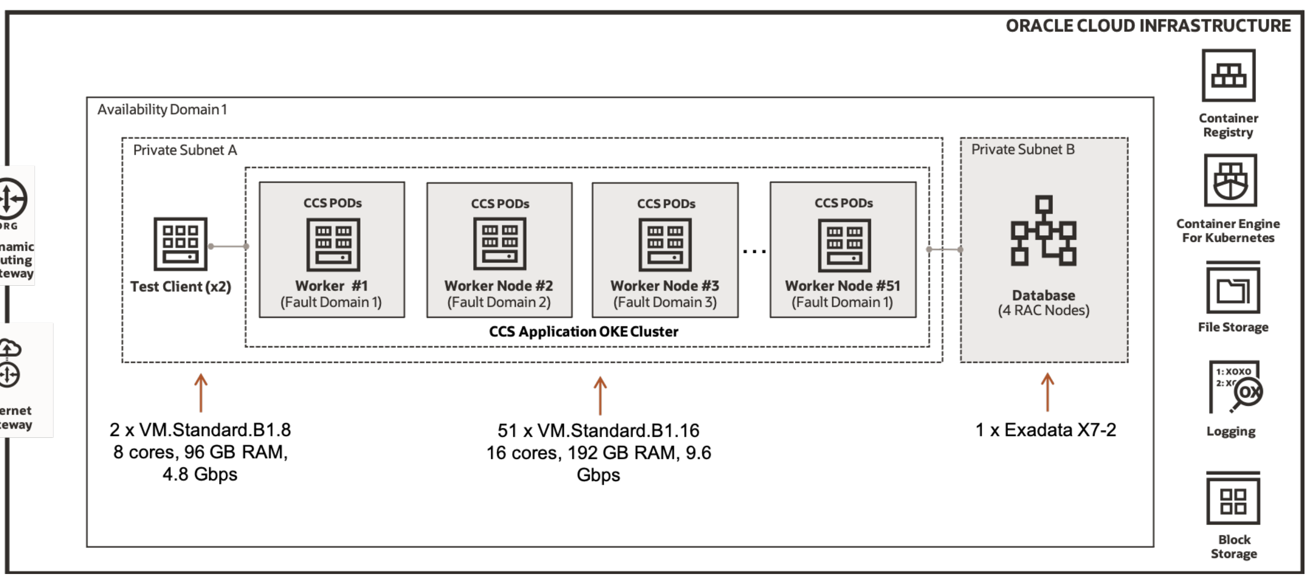


Figure 3 – Scalability and Performance cloud infrastructure test environment

Test Results

The Oracle CCS demonstrated single digit core charging latencies, high transaction throughput, efficient resource utilization and linear scalability. A 100% success rate was achieved for all traffic presented to the system.

High throughput, Low latency

Table 3 summarizes the 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 charging server (ecs) instances. Note that latency is not applicable for the notification traffic, as this was initiated by the Oracle Converged Charging System.

TRAFFIC TYPE	TRANSACTIONS PER SECOND (TPS)	AVERAGE LATENCY (MILLISECONDS)
SMS	2,884	5.75
Voice	21,200	6
Data	76,800	6.96
Notifications	30,200	N/A
Balance Queries	5,592	2

Table 3 – Observed throughput and average latencies (50 million subscribers)

Efficient Resource Utilization

Observed resource utilization across the core converged charging system application and the Oracle database used for persistence is shown in table 4, showing compelling results. Note that the utilization values shown in the table were observed during the steady state (maximum) traffic phase of the test.

RESOURCE UTILIZATION	
Average App CPU utilization	46%
Average DB CPU utilization	12%
Average App memory utilization	46%
Average DB memory utilization	54%
Max. App IOPS	300
Max. DB IOPS	7,360

Table 4 – Observed resource utilization (50 million subscribers)

Additional measurements showed consistent utilization results achieved across the ramp-up and ramp-down phases of the testing.

Linear Scalability

When comparing resource utilization measurements taken during the separate runs for 12.5M provisioned test subscribers and 50M provisioned test subscribers, linear scaling characteristics were observed (figure 4).

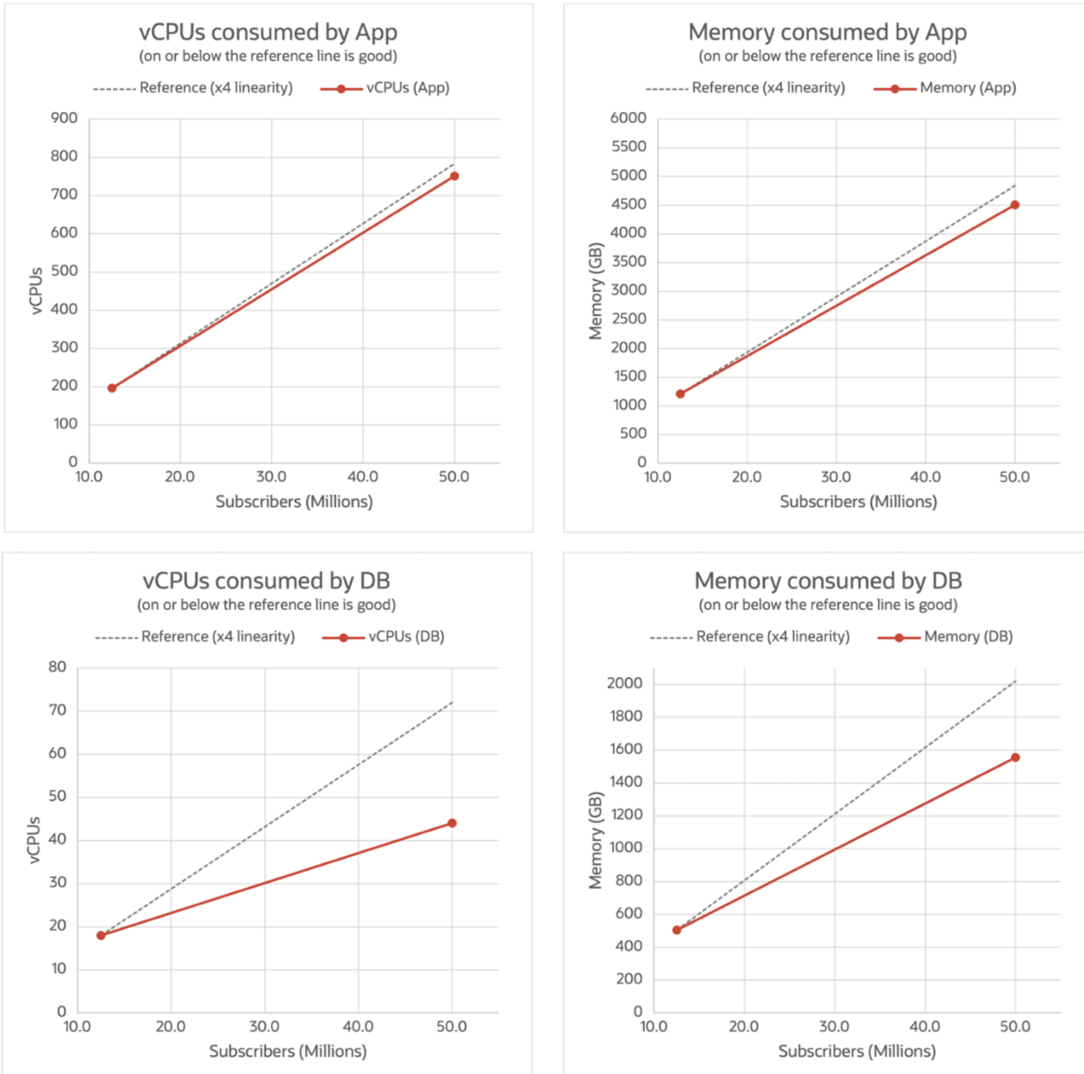


Figure 4 – Linear scaling results observed between 12.5 and 50 million subscribers

Table 5 summarizes the linearity characteristics, showing the scaling factors between the two subscriber volumes.

	12.5M	50M	SCALING FACTOR
Subscribers	12,500,000	50,000,000	4.00x
Total TPS (achieved)	33,771	136,676	4.05x
vCPUs Apps (consumed)	196	751	3.83x
Memory Apps (consumed)	1210 GB	4505 GB	3.72x
vCPUs DB (consumed)	18	44	2.44x
Memory DB (consumed)	505 GB	1555 GB	3.08x

Table 5 – Linear scaling results showing scaling factors

The results show that the CCS application demonstrates linear scaling characteristics, showing slightly increased efficiency at the higher level of throughput. The Database vCPU and Memory are not linear, with less resources required per subscriber for the 50 million subscriber test. This is due to database batch operations achieving higher efficiency as the batch size increases.

SUMMARY

The 5G era presents new challenges for digital service providers to effectively 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 webscale operational efficiencies and at the same time function as a real-time experience engine for end users. The Oracle Converged Charging System (CCS), 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.

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

The performance test described in this technical briefing demonstrates the inherent high throughput, low latency and linear scaling characteristics of the Oracle Converged Charging System. Using a provisioned base of 50 million subscribers (90% prepaid, 10% postpaid) provisioned with industry realistic price plans, the test, undertaken on Oracle Cloud Infrastructure, demonstrated single digit core charging latencies, high transaction throughput, efficient resource utilization and linear scalability.

CONNECT WITH US

Call +1.800.ORACLE1 or visit oracle.com.
Outside North America, find your local office at oracle.com/contact.

blogs.oracle.com facebook.com/oracle twitter.com/oracle