White Paper

Telecommunications

WWT, Druid, and Intel Evaluate Feasibility of Public Cloud-Hosted Private Wireless 5G Core

This joint initiative evaluates feasibility and validates the performance of the private 5G control plane functions of Druid's Raemis™ 5G core and IP Multimedia Subsystem (IMS) software running in the public cloud

The primary goal of this study was to assess the feasibility of the public cloud for hosting 5G core control plane functions, ensuring it can deliver the performance metrics demanded by the market.

The virtualization and disaggregation of mobile network systems is leading to mobile network operators (MNOs) trialing and deploying an increasing number of network elements in the public cloud. Using the public cloud may bring a more flexible and beneficial cost structure for workloads thanks to the ability to scale these network elements to meet variable demands.

The decision to move any network element to the cloud, however, comes with a raft of questions to resolve including:

- Is the workload ready for public cloud?
- Is the public cloud network ready for telco specific protocols?
- Can the cloud service meet an uptime commitment of 99.999%?
- Does the cloud service introduce any new security or regulatory issues?
- What are the cloud instance characteristics?

The tests outlined in this paper attempt to answer that last question for private cellular 5G core services. Public cloud services are configurable as a "computing resource" that includes CPU cores, RAM capacity, disk space, and network throughput. Knowing the computing resource capacity for your workload helps to optimize the resources needed for that network element.

Intel Network Builders ecosystem members WWT and Druid Software worked with Intel to define computing resources for 5G core services based on Druid's Raemis software platform. The use case was a disaggregated private wireless deployment where the user plane function (UPF) and mobile RAN are distributed on site but signaling and control plane functions are centralized in the public cloud. This network design enables a centralized signaling and control plane to manage a distributed mobile RAN across multiple sites.

The objectives of these tests were to measure the reasonable and maximum number of 5G UEs that could be serviced per computing resource. In this paper, four scenarios are tested with the minimum computing resources defined for each of the scenarios.

Table of Contents

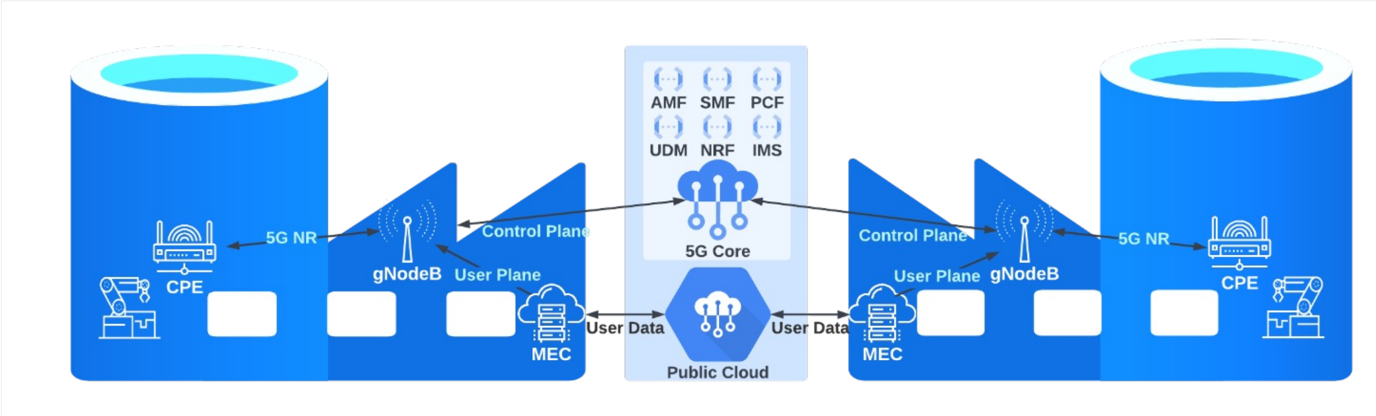

Figure 1. Block illustration of the private 5G use case discussed in this paper. UPF and RAN are located at the network edge, but the control plane signaling is in the public cloud.

Druid Raemis Virtualized Cellular Network Software

The software used in the system under test (SUT) was an instance of Druid Software's Raemis™ 5G core and IP Multimedia Subsystem (IMS) software which supports IP-based voice and other multimedia applications. Raemis is a virtualized software suite that mobile network operators can use to deliver 5G, 4G, 3G, 2G and Wi-Fi services working with a wide array of radio vendors. The software supports O-RAN Alliance standard interfaces.

Raemis 5G mobile networking solutions delivers a full 5G core solution including access and mobility function (AMF), session management function (SMF), policy control function (PCF) and the short message service function (SMSF). All together this delivers the essential services for 5G data networks; taking data from the radio access network (RAN) and preparing it to go the Internet. This 5G core software is supported by the Raemis user plane function (UPF) that provides data throughput, and also the IMS solution which delivers voice and other multimedia services.

The solution features a built in REST API to simplify custom integration with third-party software. Druid also makes a 4G Raemis enhanced packet core (EPC) that supports non-standalone 5G services with just a software upgrade and the same software can be used for standalone 5G networks. The Raemis platform features extensive network slicing capabilities for 5G networks that add additional levels of separation for QoS and security.

Testing Hardware and Expertise from WWT

The test server and services were contributed by World Wide Technology (WWT). WWT partners with MNOs and large enterprises worldwide to provide innovative technology and supply chain solutions. The company's cloud consulting helps organizations secure and optimize cloud environments using its vast resources including 5,000 experts and more than 4 million square feet of warehousing, distribution and integration space throughout the world.

The load generation infrastructure, UEs and the mobile RAN were setup to run on a server in the WWT Advanced Technology Center (ATC). The Advanced Technology Center provides the perfect environment to conduct communications technology proofs of concept in a suite of collaborative technology labs, to evaluate reference architectures, and to analyze and compare competing solutions.

The center is situated on more than 45,000 square feet of secure product integration space and has the capacity to configure and test thousands of communication system projects per week. More than 3,000 engineers and technical personnel are also available on site to perform product demonstrations in person or in a virtual environment accessible from anywhere in the world. The center's engineers and technical personnel stay current with the latest technology developments through continual training and close working relationships with both original equipment manufacturers (OEM) and independent software vendors (ISV) personnel.

Powered by a multi-tenant private-cloud infrastructure, the Advanced Technology Center is organized into four levels of research and development capabilities to allow for efficient and cost-effective testing, training, and integration:

- On-site technology infrastructure that meshes with the networks of all leading OEMs, with more than 500 racks of equipment, sandbox testing environments, and a hands-on solution showcase center that reduces technology evaluation time from months to weeks or days;
- Multiple software- and hardware-based demonstration projects to assist in the design and procurement of deployment solutions;
- Functionality for remote or on-site access; and
- Space for individual test subjects to conduct evaluations, proofs of concept, and pilot projects.

Figure 2. Lobby view of the WWT Advanced Technology Center.

Cloud Servers Based on Intel® Xeon® Scalable Processors

Raemis was tested in both Azure and AWS. Note that our goal/purpose was not to compare and contrast the two cloud services but to characterize the performance on a given platform – this report should not be viewed as a competitive comparison; each cloud should be viewed on its own.

On AWS, two EC2 cloud instances were selected for the test (see Table 1): m5zn, which is a high-frequency larger memory instance offering 100Gbps networking and based on a 2nd Gen Intel® Xeon® Scalable Processor. The other instance was the m6i larger memory instance based on a 3rd Gen Intel® Xeon® Scalable Processor.

The Azure Standard_d16d_v5 service was chosen. In Azure it isn't possible to choose a specific CPU SKU. Each instance is configured to draw from multiple CPU generations and the CPU features are masked to the common feature set for that instance (v3, v4, v5, etc.) chosen when the instance was created.

Cloud Test Set Up

Figure 3 shows the test environment architecture. The RAN and UEs are simulated in a server that is part of the WWT ATC, whereas the Raemis 5G core is running in a number of cloud instances (see Table 1 for cloud configurations).

Druid provided the test scripts that simulate the 5G radios (vGNBs) and the end user equipment (UE). Druid Stats Collector is also running on the server and is monitoring the Raemis threads, CPU usage, memory usage and network I/O and then logs that information to be presented using a Grafana dashboard.

This test setup was designed to determine the maximum number of UEs that can be serviced by a computing resource, which is defined as the amount of CPU compute cycles, RAM, disk space and network I/O required to support a given number of UEs. The tests were constrained by core usage levels, running at either 90% (maximum) or 60% (safety) as determined by Druid based on their customer experiences. The safety level was chosen to allow for core usage to spike in response to a jump in number of UEs attaching to a core.

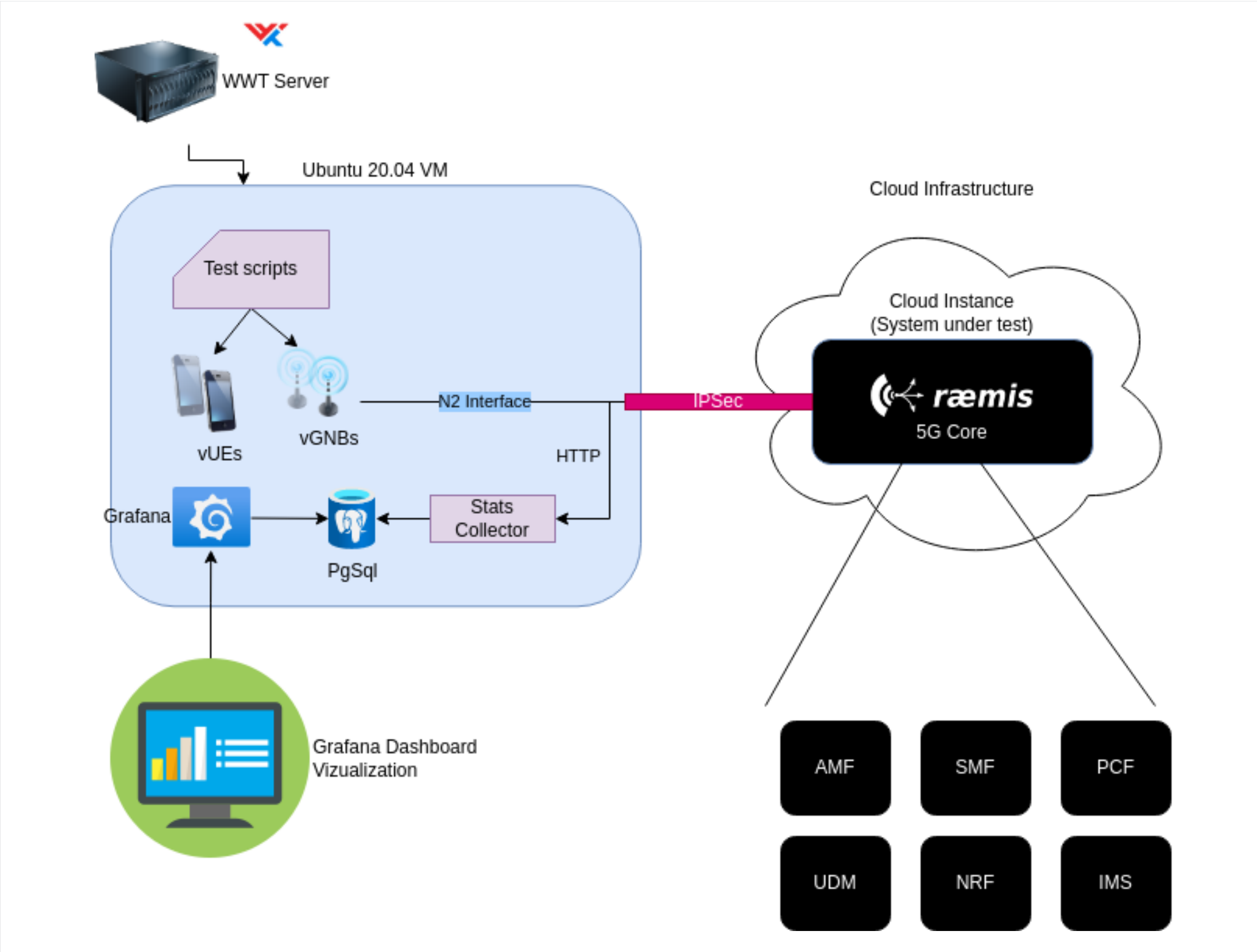

Figure 3. Raemis cloud test testbed.

There were two phases in the test setup:

- **Registration:** This is when all virtual UEs attach themselves to the gNBs with an even distribution. The UE then requests a connection to a data network or networks using a data network name (DNN) which is a unique identifier used by the 5G core network to route traffic to a specific network slice. During this phase the UEs are fully authenticated through the encryption mechanisms built into the network core.
- **UE Mode Transition:** The UEs are now programmed to emulate user activity by cycling between idle mode and connected mode every 30 seconds. Because in the real world each user's average rate of mode transitions is different and varies throughout the day, this cycling frequency will not hold for all users.

The testing team ran the tests described above in addition to two variants of the tests:

- Enabling or disabling non-access stratum (NAS) signaling encryption. Enabling NAS encryption consumes more CPU cycles per signaling procedure but protects signaling from third-party snooping.
- Registration to IMS DNN enabled / disabled. Connection to the built-in Raemis IMS DNN provides support for native telephony services on the UEs, such as voice calls. However, this requires additional signaling which should result in each fewer UEs per computing resource.

A cloud optimized version of the Raemis 5G core was used throughout the tests. This version of Raemis was fine tuned to cater to typical conditions in cloud environments such as slow access to physical volumes.

Cloud Image	Service provided	NAS Encryption	Safe number of serviceable UEs[1]	Max number of serviceable UEs[2]
AWS m5zn	Data only	Disabled	90K	130K
AWS m6i	Data only	Disabled	90K	130K
Azure Standard_ D16d_v5	Data only	Disabled	180K	260K
AWS m5zn	Data + telephony	Disabled	35K	50K
AWS m6i	Data + telephony	Disabled	35K	50K
Azure Standard_ D16d_v5	Data + telephony	Disabled	105K	150K
AWS m5zn	Data only	Enabled	70K	105K
AWS m6i	Data only	Enabled	70K	105K
Azure Standard_ D16d_v5	Data only	Enabled	140K	210K
AWS m5zn	Data + telephony	Enabled	20K	30K
AWS m6i	Data + telephony	Enabled	20K	30K
Azure Standard_ D16d_v5	Data + telephony	Enabled	80K	120K

Table 1. Test set up executive summary.

The cloud instances covered in this testing are seen in table 2.

Instance	CPU	CPU Details	# Cores	RAM	Disk space
AWS EC2 m5zn-large	Intel® Xeon® Platinum 8252C Processor	Speed: 3.8GHz Turbo speed: 4.5GHz L1d cache: 192KB L1i cache: 128KB L2 cache: 4MB L3 cache: 24.8MB	8 logical, 4 physical	32GB	128GB
AWS EC2 m6i-large	Intel® Xeon® Platinum 8375C Processor	Speed: 2.9GHz Turbo speed: 3.5GHz L1d cache: 192KB L1i cache: 128KB L2 cache: 5MB L3 cache: 54MB	8 logical, 4 physical	32GB	128GB
Azure Standard_D 16d_v5	Intel® Xeon® Platinum 8370C Processor	Speed: 2.8GHz Turbo speed: 3.5GHz L1d cache: 384 KiB L1i cache: 256 KiB L2 cache: 10 MiB L3 cache: 48 MiB	16 logical, 8 physical	64GB	600GB

Table 2. Cloud instance descriptions.

Note: These cloud instances were provisioned via default on-demand templates (i.e., shared hardware) and not dedicated instances.

Test Results

Tests were run with four configurations. In the first two tests each UE was registered to only one DNN for data services and performance tests were run with and without NAS encryption. In additional tests, each UE registered to two DNNs – one for data services and the other for IMS. These configurations were tested with and without NAS encryption.

One cost saving result of the tests is that the servers each returned the exact same supported UEs per signaling thread. This means that the newer Intel® Xeon® Scalable processors deliver similar performance at 22% lower clock speeds which brings a significant cloud cost saving and reduces the energy consumption of the workload compared to cloud servers with higher clock speed processors.

1DNN no Encryption

Without NAS encryption, approximately 65,000 UEs can be serviced at peak CPU usage of ~88% for the main signaling core. In the safe mode tests – where the CPU usage for the main signaling core at ~60% - the computing resource supported ~45,000 UEs. Other elements of the computing resource included 6GB of RAM, total disk space of 10GB and only 10MB of network bandwidth.

These levels were supported by a computing resource with one full physical core for signaling with an additional 33% of another physical core for overhead such as database, etc.; 3GB RAM and 10Mbps network I/O.

Note that the instances used in the 1DNN without encryption tests provided four physical cores, three of which could be used for the Raemis workload. With this configuration a maximum threshold of 130,000 UEs were supported with encryption at peak CPU usage, and 70,000 UEs hosted with safe production.

1DNN with Encryption

In this test, encryption was enabled that places an additional processing requirement on the CPU. MNOs enable NAS encryption to prevent interception of signaling traffic carrying information normally sensitive to mobile operators, such as personal subscriber identification. With encryption, the server reached a peak of 35,000 UE serviced at 90% CPU usage for the main signaling thread.

These levels were supported by a computing resource with one full physical core for signaling with an additional 20% of another physical core for overhead; 2GB RAM and 10Mbps network I/O.

Note that the instances used in the 1DNN with encryption tests provided four physical cores, three of which could be used for the Raemis workload. With this configuration, 105,000 UEs can be supported with encryption at peak CPU usage, and 70,000 UEs hosted with safe production.

2DNN without Encryption

In these tests, the UEs register with the 5G core seeking two network connections: one for data and one for IMS (voice). In the first test, no encryption was used resulting in support for 50,000 UEs at 85% CPU utilization.

The computing resource needed to support this result included one full physical core for signaling, and an additional 1/2 physical core for overhead; 3GB RAM and 10Mbps network I/O.

While that was the computing resource, the actual instance used in the tests provided four physical cores, three of which could be used for the Raemis workload and one of which was dedicated to IMS. With this configuration, 50,000 UEs were supported with encryption at peak CPU usage, and 35,000 UEs could be hosted with safe production. Only 5GB of RAM is needed along with 10GB of disk space.

2DNN with Encryption

The dual DNN tests with encryption showed that 30,000 UEs can be serviced on both instances at 90% CPU utilization for the signaling thread. The computing resource needed to support this result included one full physical core for signaling, an additional 1/3 physical core for overhead; 3GB RAM and 10Mbps network I/O.

When testing the full four-CPU instance provided four physical cores, three of which could be used for the Raemis workload and one of which was dedicated to IMS. With this configuration, 30,000 UEs can be supported with encryption at peak CPU usage, and 20,000 UEs could be hosted with safe production. Only 4GB of RAM is needed along with 10GB of disk space.

Conclusion

The results outlined in this document provide guidance on the size of the computing resources needed for running Raemis 5G core services on public cloud instances. We demonstrated the testing using two different cloud services to show the importance of processor performance, RAM and networking on the ability to deliver 5G core control plane services from the public cloud. This knowledge provides the ability to order the right cloud capacity for a network and to save costs. This could be seen in the standard instances used for these tests, each of which came with four cores and more RAM and networking than what was needed.

Learn More

WWT

Druid Software Raemis platform

Intel® Xeon® Scalable Processors

Intel Network Builders

Characterizing Telco Workloads in Public Cloud Infrastructure

Scaling Telecom Core Network Functions in Public Cloud Infrastructure

Previous Work on this subject:

Characterizing Telco Workloads in Public Cloud Infrastructure

Scaling Telecom Core Network Functions in Public Cloud Infrastructure

[1] Safe, in this study, is defined as 60% compute capacity
[2] Max, in this study, is defined as 90% compute capacity

0923/DC/H09/PDF ♻Please Recycle 357180-001US