A SMARTER NETWORK: CREATING A PLATFORM FOR INNOVATION WITH EDGE COMPUTING

CONTENTS

MOVING COMPUTE TO THE EDGE

Some of the most exciting technologies being piloted today require ultra-low latency. Self-driving cars need to be able to react quickly to what's happening on the road. Augmented reality (AR) and virtual reality (VR) need to respond in less than 15-20 milliseconds[1] to create a seamless experience and avoid giving users the equivalent of motion sickness. Many Internet of Things applications exist for safety and quality control purposes, and a delayed response could be dangerous or may result in damage.

Latency-sensitive applications such as these may prove unviable if the application is hosted in a centralized cloud. The time taken to send data across the network and the response back again can be significant when there are huge volumes of data involved. As Figure 1 shows, an application in the cloud might have latency of 100ms, 20 times greater than the latency of an on-premise application.

To reduce latency, communications service providers (CoSPs) and cloud over-the-top (OTT) service providers are increasingly looking at using edge computing to process workloads closer to end user devices. Not only do you get lower latency the closer to the user device you can deploy the workload (see Figure 1), but you also get more bandwidth at the edge of the network. This is essential for some AR, VR, and video analytics applications.

When we talk about edge computing, we mean the placement of data center-grade network, compute and storage resources at the edge of the network. We think of the network edge as being outside the network core, and covering locations such as regional data centers, next generation central offices (NGCOs), wireline fixed access points, wireless access base stations,

and radio access networks (RANs). Our definition also includes on-premise locations. One example would be universal customer premises equipment (uCPE) devices where multiple workloads such as a software defined wide area network (SD-WAN) and enterprise applications can be hosted in a single edge computing device.

Network functions virtualization (NFV) at the edge is helping to cut network latency, and is driving a heightened interest in hosting customer-facing applications at the edge using Multi-Access Edge Computing (MEC). Security applications can reduce the load on the network by, for example, blocking distributed denial of service (DDOS) attacks at the point where they enter the network. A user-facing application, such as VR gaming, can store and serve data locally to increase responsiveness. In some cases, hosting a user application at the edge will require network workloads (such as firewalls) to be hosted at the edge too, to ensure latency remains low.

As 5G and wireline fiber to the home increase the bandwidth and speed of network access, the backhaul across the network risks throttling innovation, and reducing the performance that applications can achieve. Hosting applications at the edge can help to deliver on the potential of 5G to offer high bandwidth, low latency experiences.

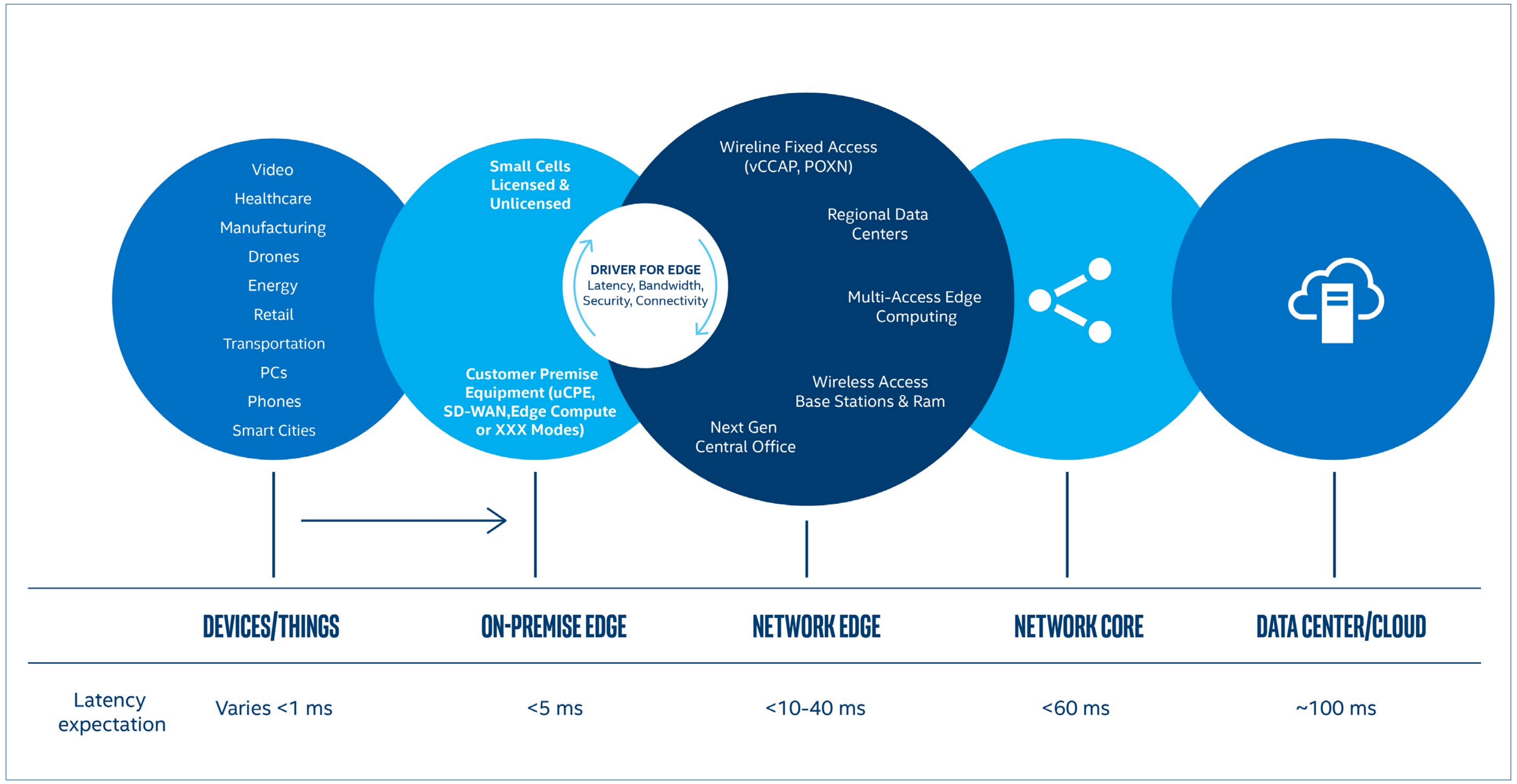

Figure 1: Edge computing covers a continuum of application locations, including where devices connect to the network (the network edge), on-premises equipment (such as uCPE), and devices themselves. Greenfield networks based on fiber may offer lower latency, but these estimates are typical of Europe and North America.

There are other benefits to edge computing too, beyond reduced latency. Edge computing can help to:

- **Optimize total cost of ownership (TCO):** It may be costly to send huge amounts of data across the network that don't need to take that trip. Processing data locally can cut the cost of transmission, and cut the cost of centralized storage. Applications can be hosted where they deliver the highest return. Those requiring lower latency and requiring higher bandwidth can be hosted at the base station or Next Generation Central Office (NGCO). Applications with a little more tolerance for latency could be hosted in a regional data center. Network workloads (such as firewalls) can be hosted at optimal points in the network to strike a balance between performance (closer to the user) and scale economies (closer to the centralized cloud). The network functions virtualization infrastructure (NFVI) at the edge can be shared between fixed and mobile services, so more resource could be provisioned to broadband gateways when customers are at home, and to mobile gateways when they are at work and using their mobile devices.
- **Increase agility:** Replacing fixed function customer premises equipment (CPE) with edge computing platforms enables CoSPs to provision services more quickly, offer flexible trial periods on new services, and upgrade the CPE without a site visit or hardware upgrade. With dynamic service provisioning at the edge, CoSPs can prioritize the most essential services during periods of network disruption to enable a quicker recovery. Decoupling the hardware and software at the edge enables the CoSP to upgrade capacity and capabilities more flexibly at any time.
- **Generate new revenue streams:** By introducing edge computing, CoSPs can establish a platform for innovation. They can create their own services and resell them, but, importantly, they can also open up their network to host third-party applications. Physical network locations provide a key competitive advantage that can help CoSPs to compete with over-the-top providers, or to win more business from them by renting out low-latency compute capacity within the network.
- **Comply with data locality requirements:** Whether it is required by legislation or by the sensitivity of a customer's business, hosting applications on-premise can help to meet data sovereignty requirements.
- **Enable context awareness:** Applications that are hosted at the edge can have a layer of context awareness that may be difficult to achieve in a centralized cloud. For example, an AR application could have data and imagery related to the location where it is hosted. Decisions can be taken in a real time context, for example in security applications for which there would not be time to go back to the cloud for a decision.
- **Cope without connectivity:** If there is a weak or intermittent connection between the edge and the centralized cloud, edge computing can enable the application to run without interruption. In some cases, connectivity may be undesirable or unnecessary, for example in an internet of things (IoT) monitoring application where the only outbound communication needed is an alert when there's an abnormal measurement.

The innovation in data center technology we've seen over the last few decades has made edge computing possible. We can now deploy servers that are small enough and tough enough to be hosted in on-the-street locations such as central offices, with viable space, power, and cooling requirements.

EDGE COMPUTING IN THE REAL WORLD

Edge computing is perhaps most powerfully demonstrated by media use cases, where large amounts of visual data must be processed in real time. Intel has worked on pilot projects involving broadcast video, live streaming of events and gaming.

STREAMING LIVE GOLF ACTION

Intel and AT&T* are working together to research and demonstrate how 5G wireless communications can be enhanced with edge computing. In the media industry, one of the challenges is that video might be captured in 4K+ and must currently be transmitted using wired cameras. The wiring is costly and difficult to implement in many environments, increasing the time required for set-up and tear-down. Intel, AT&T and Ericsson* worked with Fox Sports* to stream live golf action from the 2018 US Open over a 5G link. Video was transmitted from the cameras over 5G to a media encoder/decoder based on Intel® architecture which was positioned at the edge of the network (see Figure 2). Decoding the camera feed at the edge of the network creates the potential for video to be shown to athletes and audience members on site, and in the future could be enhanced with AR or multi-camera stitching. Content in the pilot was also shared with golf fans on DIRECTV*, AT&T's broadcast satellite service.

STREAMING HD VIDEO AT A CONCERT

China Unicom*, Nokia*, Intel, and Tencent Cloud* worked together to implement an edge computing trial at the Mercedes-Benz* Arena in Shanghai. This test was carried out during a pop concert, with the arena at full capacity. The trial used an application installed on attendees' devices to watch HD video streamed from cameras around the arena, ensuring they had the best possible view of the show. Using this application, fans could choose which of four camera channels they wished to view. The trial was carried out using today's 4G technology, but when 5G becomes available, the extra bandwidth could be used to support VR or augmented reality experiences, creating new ways for fans to interact with their favorite bands or sports teams. In the Shanghai trial, live video latency was cut from the typical 30 seconds for Internet video to 0.5 seconds[2].

Figure 2: A remote 4K video camera connected to the Fox Sports* network over a 5G link using spectrum provided by AT&T*. Ericsson's* 28GHz radio and antenna are located at the tower seen in the distance.

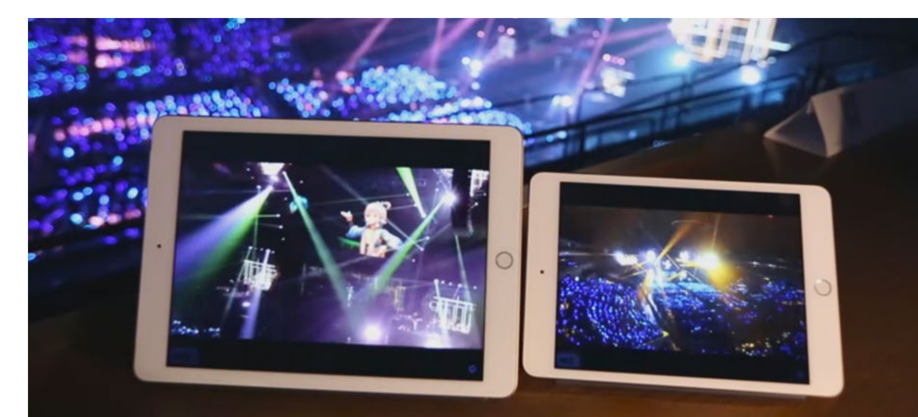

Figure 3: Concert attendees at the Mercedes-Benz* Arena in Shanghai used a mobile app to view HD video streamed from four camera channels.

ENABLING MIXED REALITY GAMING

Intel, AT&T, Ericsson, Warner Bros.* and DC* collaborated to create a mixed reality experience that immerses fans wearing VR headsets in the world of DC's tech-savvy Super Hero Batman* as he takes on Super-Villain The Scarecrow*. The experience builds on a proof-of-concept demonstration conducted in December 2018 with the University of Southern California's School of Cinematic Arts, and was shown at Mobile World Congress 2019 (see Figure 4). The demonstration showcased the latest evolution in mobile broadband technology and distributed cloud remote rendering. It ran on Intel® Xeon® Scalable processors in Ericsson radio base stations with the Intel® 5G Mobile Trial Platform.

RE-ARCHITECTING THE CENTRAL OFFICE

Telefonica* has been working with Intel on a pilot project to make the central office fully software defined. The pilot is running on servers based on the Intel Xeon Scalable processor and an architecture based on open source software. The initiative puts compute resources closer to Telefonica's customers. The new architecture:

- Enables the CoSP to help reduce transport costs across the network with edge processing and traffic filtering;
- Creates an opportunity to monetize new edge services provided by Telefonica and its partners; and
- Enables customers to be served with low latency and personalized apps with excellent performance.

For example, a video application might suffer from unpredictable and large (50ms) latency when served from the cloud. Using edge computing, Telefonica estimates that latency can be cut to less than 4ms, and latency can be guaranteed below a certain level.

Figure 4: Visitors to Intel's booth at MWC 2019 are part of a mixed reality experience that immerses fans wearing VR headsets in the world of DC's tech-savvy Super Hero Batman as he takes on Super-Villain The Scarecrow.

STANDARDIZING THE SOFTWARE ENVIRONMENT AT THE EDGE

One of the challenges of deploying applications at the edge is that there are diverse underlying platform architectures (uCPE, NGCO, vRAN) and different access termination methods (S1, LTE CUPS, SGI, 5G UPF). None of this complexity exists in cloud applications, so it presents a significant barrier to software developers who might otherwise be able to create innovative apps for the network. The complexity also complicates the development and testing cycle for applications developed by the CoSP in-house and restricts their portability across the network.

The Open Network Edge Services Software* (OpenNESS*) is a software toolkit that abstracts away this complexity, enabling developers to create apps that run unchanged in any edge location or in the centralized cloud. The platform has standard APIs from 3GPP and the ETSI Multi-access Edge Computing (MEC) industry group. It also offers cloud adapters to connect to Amazon Web Services* (AWS*) and Microsoft Azure*, among others. Applications can move data seamlessly from the devices, to the edge, and to the cloud, and back again.

OpenNESS integrates with the Intel® Media SDK and Intel® Distribution of OpenVINO™ toolkit, to streamline the development and help accelerate the performance of visual applications. Figure 5 shows the architecture. OpenNESS can be used with Intel® FPGAs and artifical intelligence (AI) accelerators to deliver higher performance applications.

Mobile World Congress 2019 included a demonstration of OpenNESS, showing how applications can be orchestrated, managed and run across vRAN, MEC, NGCO and cloud locations. Using OpenNESS, the same application could be written for the cloud and run on any edge location.

For more information, visit www.openness.org.

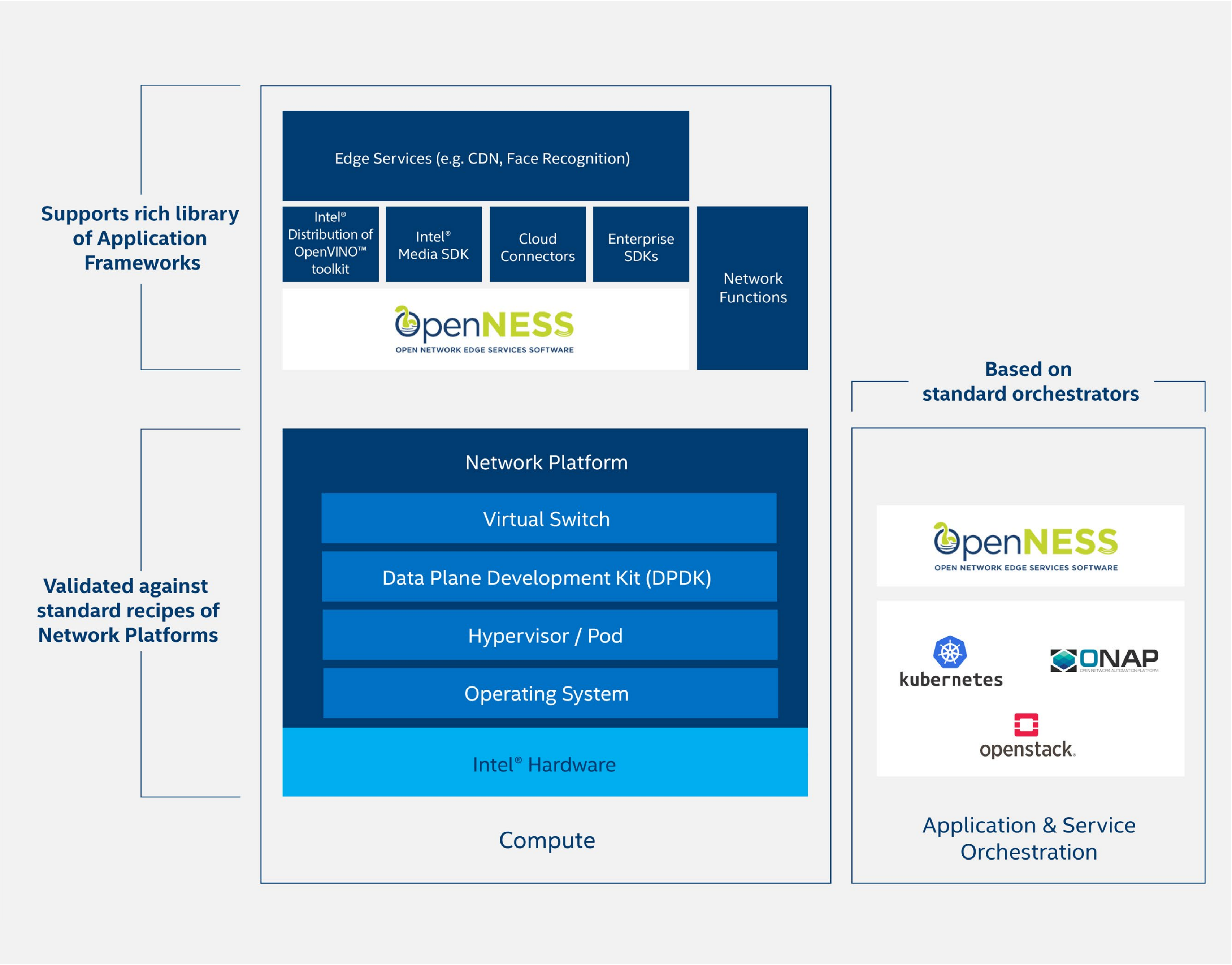

Figure 5: OpenNESS* enables the same applications to run unmodified in different edge locations, and in the cloud.

DELIVERING THE PERFORMANCE REQUIRED AT THE EDGE

Intel offers a range of processor technologies and reference architectures to help you deliver the performance required at the edge of the network.

INTEL® PROCESSORS

Intel provides a portfolio of processor technologies, so you can right-size your platform and your investment to your application and network requirements.

These processors (see Figure 6) are:

- Intel Atom® processors, which offer up to 40Gb/s packet processing with a low power requirement of 4W+[3].
- Intel® Core™ processors, which are well suited to applications that require higher media performance.
- Intel® Xeon® D processors, which offer up to 190Gb/s packet processing and have integrated Ethernet and acceleration[4].
- Intel® Xeon® Scalable processors, which offer up to 580 Gb/s packet processing on a dual socket platform[5].

The 2nd generation Intel® Xeon® Scalable processor offers higher per-core performance than the previous generation, more cores, and enhanced internal bandwidth. There are also specialized processors for NFV, which offer an improvement in the NFV workload performance, compared to the previous generation Intel Xeon Scalable processor[6]. The Intel® Xeon® Scalable 6252N processor, for example, has 24 cores and consumes up to 150W, the same as the Intel® Xeon® Scalable 6252 processor. However, it operates at a frequency that is 10 percent higher, 2.3GHz. Utilizing Intel® Speed Select Technology can optimize NFV performance and power consumption.

The 2nd generation Intel Xeon Scalable processor also introduces support for Intel® Optane™ DC persistent memory, which can be used to store and process data more quickly in applications such as content delivery networks (CDNs), VR, AR, and image recognition.

INTEL® SELECT SOLUTIONS FOR UCPE

Intel® Select Solutions for uCPE provide a solution reference design for uCPE products, with commercial solutions available that have been verified to deliver the performance required. The solutions are based on the Intel® Xeon® D processor, which brings powerful performance to the edge of CoSP networks. The reference design includes 16GB of DDR4 2133 MHz memory, 2 x 10GbE integrated Ethernet ports and Intel® SSDs for storage. Optionally, Intel® QuickAssist Technology (Intel® QAT) can be integrated to accelerate operations such as lossless compression and encryption/decryption.

Vendors providing Intel Select Solutions for uCPE include Advantech*, Caswell*, Lanner*, Nexcom*, Premier*, Silicom* and SuperMicro*.

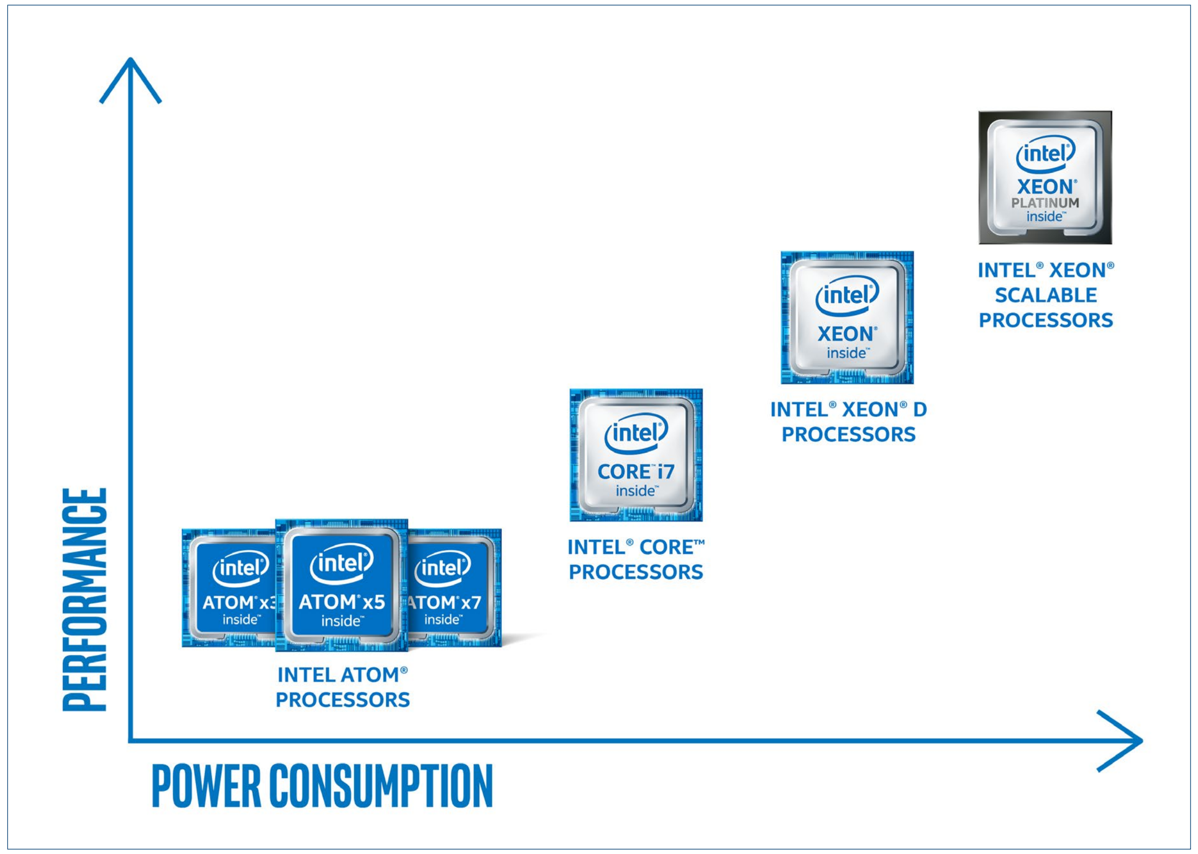

Figure 6: Intel offers a portfolio of processors, with a range of performance and power consumption characteristics.

CONVERGED EDGE REFERENCE ARCHITECTURE FOR ON-PREMISE

Intel has published a reference solution for converging and consolidating multiple workloads in a single edge platform, called Converged Edge Reference Architecture for On-Premise (see Figure 7). It helps enable a combination of multiple workloads such as WiFi/LTE/5G and CBRS radio access; analytics, AI, and media network and application workloads running in VMs or containers; and the OpenNESS SDK to accelerate the development and deployment of software. The aim is to accelerate time to market for systems integrators and CoSPs and smooth the deployment path for end customers looking to introduce edge computing. By consolidating workloads at the edge, organizations can reduce the space and power required and optimize their TCO.

The uCPE hardware platform with SD-WAN, Network Analytics, and Network Security makes up the base configuration of the Converged Edge Reference Architecture for indoor on-premise form factors. Building on that, the Converged Edge Reference Architecture supports different wireless access capabilities, such as WiFi, LTE/5G, and private wireless where required. On top of either the base configuration or the configuration with wireless access, the platform can add platform software stacks to address different use cases for vision, media and analytics (see Figure 7).

The solution design is based on the Intel Xeon D processor, and supports Intel Xeon Scalable processors to deliver enhanced performance for more demanding applications. Accelerators such as Intel FPGAs and Intel® Movidius™ technology products for AI can be used to further accelerate performance. A "virtual safety fence" application for Intel Movidius technology was demonstrated at Mobile World Congress 2019. In the demonstration, when somebody was detected entering an industrial area, machinery was stopped automatically to ensure safety.

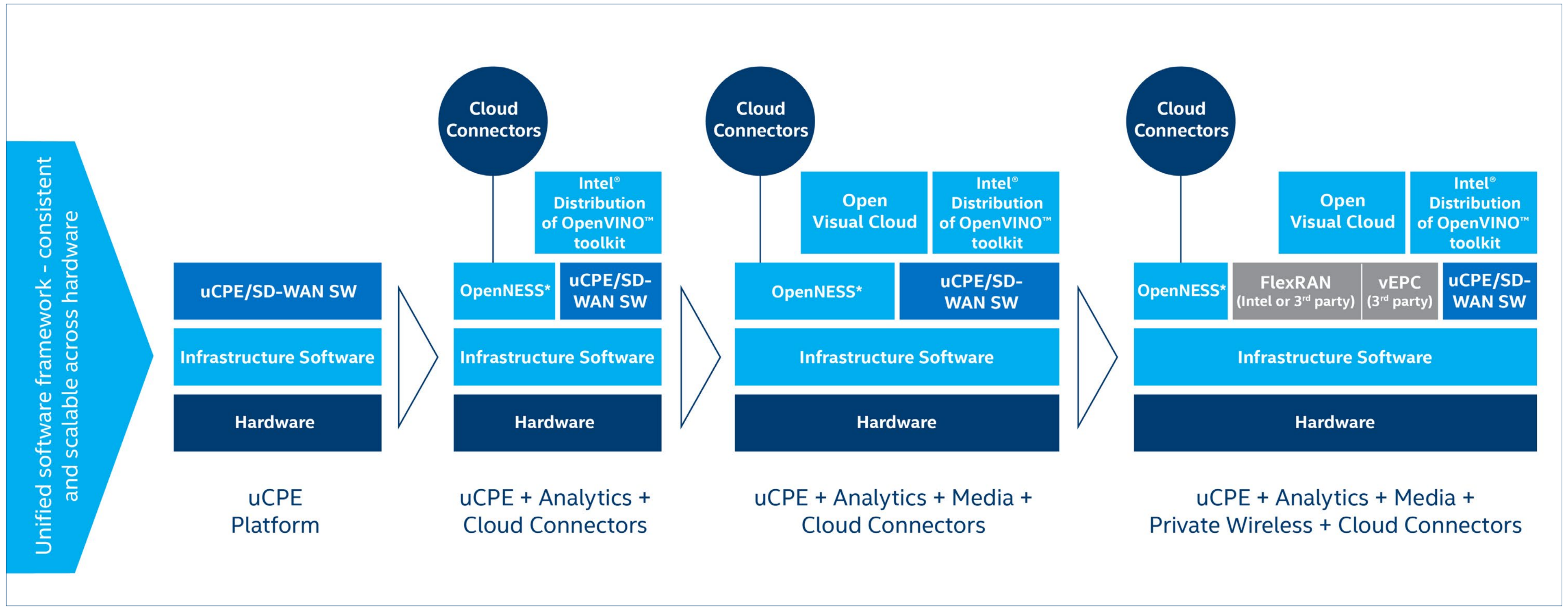

Figure 7: The Converged Edge Reference Architecture for On-Premise expands uCPE into new services and private wireless use cases for the retail, industrial and transportation sectors.

NEXT GENERATION CENTRAL OFFICE (NGCO)

As copper-wired central offices are upgraded for fiber and passive optical networking (PON) technologies, it becomes possible for fewer central offices to serve the same population. CoSPs are looking at the opportunity to consolidate their central office locations, and introduce an edge platform for innovation at the same time.

Intel has created a reference architecture for the Next Generation Central Office (NGCO) and is working with Quanta Cloud Technology* (QCT) to implement the framework, shown in Figure 8. In this platform, a number of Enhanced Platform Awareness (EPA) features from Intel are implemented using OpenStack* to boost data plane performance. These include memory huge pages in the compute nodes to enable a fast data path; pinning virtual CPUs to physical CPUs to keep data in the cache; and awareness of non-uniform memory access (NUMA), to ensure VMs work with memory on the same processor.

Intel and QCT have worked with independent software vendors (ISVs) to deploy their virtual network functions (VNFs) in the reference architecture. The VNFs have been tested to ensure that they meet the CoSP's expectations for performance at the edge.

The integrated applications are:

- ASTRI's* next-generation mobile core (NGMC) software, which supports separate control and data plane VMs to enable them to be scaled independently. A single data plane VM can handle more than 40 Gbps of traffic with a packet size of 768 bytes[7].
- netElastic's* broadband network gateway (BNG), which also enables control and user plane separation.
- Advanced Firewall Manager* (AFM) from F5 Networks*, to protect network service and subscribers.
- A CDN from Qwilt* Solutions, to help increase the speed and quality of video content distributed through the network.

Open Network Automation Platform (ONAP) is used for management and orchestration, and the QCT telemetry framework enables additional data plane VMs to be automatically launched when the network demand requires them.

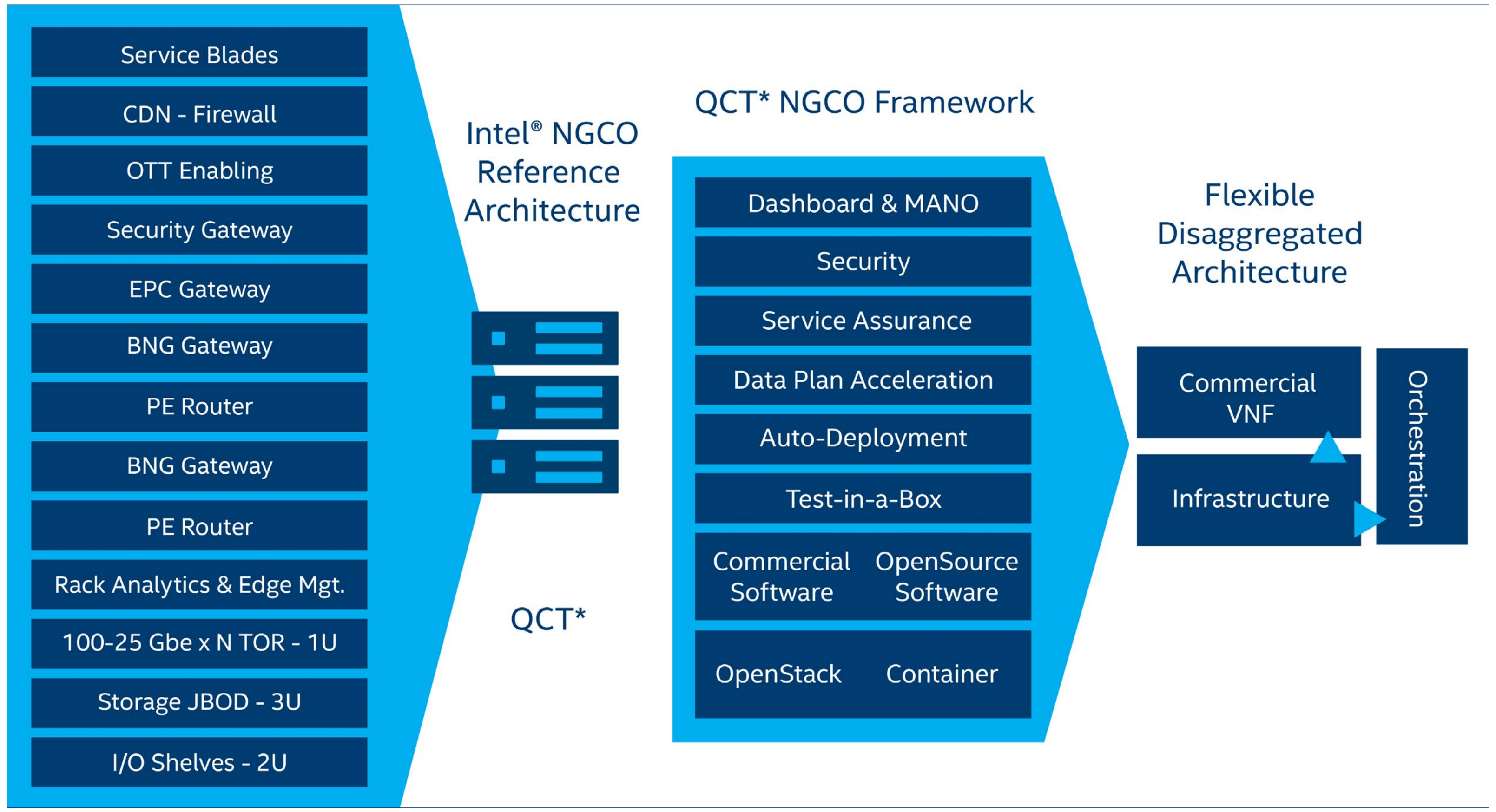

Figure 8: A reference architecture for the Next Generation Central Office (NGCO).

INTEL® 5G INFRASTRUCTURE REFERENCE DESIGN

For 5G wireless aggregation hubs, Intel has created a reference design that has a half-rack form factor (see Figure 9). The reference design is based on a hardened virtualization layer from Wind River* based on OpenStack; OpenNESS to simplify application deployment; virtual network functions from Mavenir* and Radisys*; and a dynamic orchestration layer for both applications and network functions, provided by Amdocs*.

This reference design can be deployed in the network, or on-premise for the lowest latency. One example application of this stack is in enabling the smart stadium, with Intel® True VR technology delivering a multi-camera experience using a helmet-mounted display.

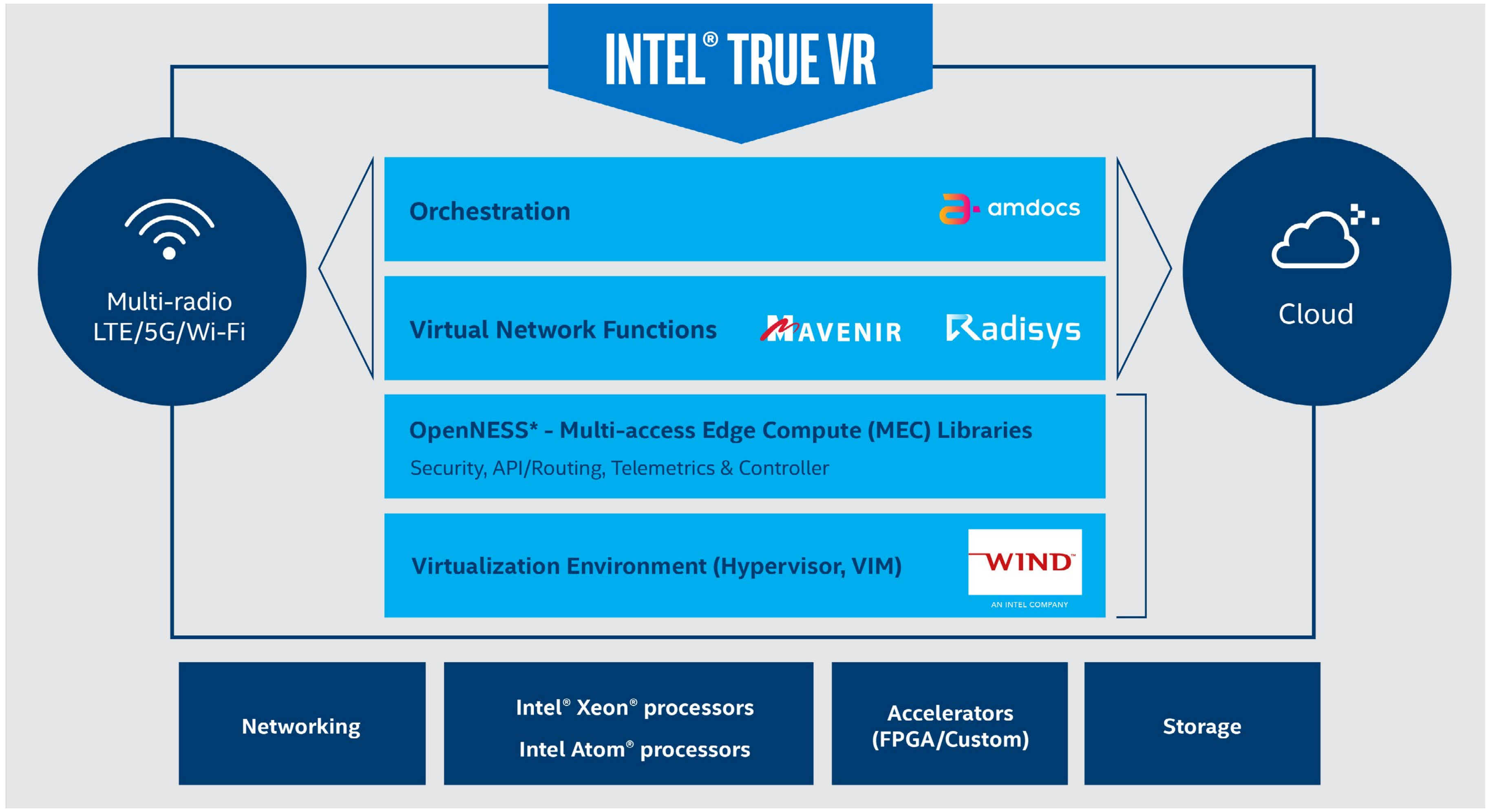

Figure 9: The Intel® 5G Infrastructure Reference Design supports edge applications including virtual reality in the smart stadium.

CONCLUSION

Edge computing provides a powerful opportunity for CoSPs to enhance their competitiveness, both by consolidating network functions and by enabling new revenue streams with innovative new applications, hosted in the network. Intel is working with CoSPs and technology providers to deliver technologies and reference designs that help CoSPs to achieve the performance required and to accelerate the deployment of edge computing solutions.

FIND OUT MORE >

LEARN MORE >

WHITE PAPERS AND VIDEOS:

- A front row seat for everyone: Mobile HD video enhances the show >
- 5G Using Edge Computing Offers Low Latency, and Cost-Effective Innovation Opportunities >
- Watch the video: Intel helps bring Batman to life with mixed reality >

SOFTWARE SOLUTIONS:

- Intel® Distribution of OpenVINO™ toolkit >
- Intel® Media SDK >
- OpenNESS: Unleashing Cloud and IOT Developer Ecosystem on the Edge with Easy "Network" Button >
- OpenNESS* website >

HARDWARE SOLUTIONS:

- Intel Atom® processors >
- Intel® Core™ processors >
- Intel® Xeon® D processors >
- Intel® Xeon® Scalable processors >
- Intel® FPGAs >
- Intel® Movidius™ technology >
- Intel® True VR >

ARCHITECTURES:

- Next Generation Central Office (NGCO) Architecture >
- Intel® 5G Infrastructure Reference Design >

INTEL® SELECT SOLUTIONS:

- Intel® Select Solutions for uCPE >

[1] http://about.att.com/innovationblog/5g_for_ai_and_vr

[2] China Unicom Edge Computing White Paper

[3] Up to 40Gb/s Packet Processing: Results based on internal Intel testing as of 8/14/2017. Intel Atom® Processor C3958 @2.0GHz, 2x Intel® X710-DA2 PCI Express Gen Dual Port 10GbE Ethernet controller (2x10GbE/card). Benchmark: DPDK v17.02 L3fwd sample application (IPv4, LPM, 1024 flows). Score: 40Gbits/s packet forwarding at 512B packet size.

[4] Up to 191Gb/s Packet Processing: Results based on internal Intel testing as of 5/1/2018. Intel® Xeon® D-2187NT CPU @ 2.0GHz, 4x Intel® XXV710-DA2 PCI Express Gen Dual Port 25GbE Ethernet controller (2x25GbE/card). Benchmark: DPDK v17.11 L3fwd sample application (IPv4, LPM, 2048 flows). Score: 191Gbits/s packet forwarding at 512B packet size.

[5] Up To 586Gb/s Packet Processing on a dual socket Platform: Results based on internal Intel testing as of 8/2/2018. Intel® Xeon® Platinum 8160 CPU @ 2.10GHz (DP), 12x Intel® XXV710-DA4 PCI Express Gen Dual Port 25GbE Ethernet controller (4x25GbE/card). Benchmark: DPDK v17.11 L3fwd sample application (IPv4, LPM, 3750000 flows). Score: 586Gbits/s packet forwarding at 512B packet size.

[6] Performance results are based on estimates as of 12/11/2018 and may not reflect all publicly available security updates. Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

[7] See the white paper, Next-Generation Central Office (NGCO). Configuration: Performance tested on a single QCT QuantaGrid* D52BQ-2U node configured with 8 VMs. Dual Intel® Xeon® Gold 6152 processors with six PCI-Express* (PCI-e) Gen3 x8 slots to CPU1 and two PCI-e Gen3 x8 slots with one Open Compute Project* (OCP) Mezzanine slot to CPU0.

0719/FP/CAT/PDF 340926-001EN