The future of cloud computing

HIGHLY DISTRIBUTED WITH HETEROGENEOUS HARDWARE

With a vastly distributed system (the telco network) already in place, the telecom industry has a significant advantage in the transition toward distributed cloud computing. To deliver best-in-class application performance, however, operators must also have the ability to fully leverage heterogeneous compute and storage capabilities.

WOLFGANG JOHN, CHANDRAMOULI SARGOR, ROBERT SZABO, AHSAN JAVED AWAN, CHAKRI PADALA, EDVARD DRAKE, MARTIN JULIEN, MILJENKO OPSENICA

The cloud is transforming, both in terms of the extent of distribution and in the diversity of compute and storage capabilities. On-premises and edge data centers (DCs) are emerging, and hardware (HW) accelerators are becoming integral components of formerly software-only services.

■ One of the main drivers into the age of virtualization and cloud was the promise of reducing costs by running all types of workloads on homogeneous, generic, commercial off-the-shelf (COTS) HW hosted in dedicated, centralized DCs. Over the years, however, as use cases have matured and new ones have continued to emerge, requirements on latency, energy efficiency, privacy and resiliency have become more stringent, while demand for massive data storage has increased.

To meet performance, cost and/or legal requirements, cloud resources are moving toward the edge of the network to bridge the gap between resource-constrained devices and distant but powerful cloud DCs. Meanwhile, traditional COTS HW is being augmented by specialized programmable HW resources to satisfy the strict performance requirements of certain applications and limited energy budgets of remote sites.

The result is that cloud computing resources are becoming increasingly heterogeneous, while simultaneously being widely distributed across smaller DCs at multiple locations. Cloud deployments must be rethought to address the complexity and technical challenges that result from this profound transformation.

In the context of telecommunication networks, the key challenges are in the following areas:

1. Virtualization of specialized HW resources
2. Exposure of heterogeneous HW capabilities
3. HW-aware workload placement
4. Managing increased complexity.

Getting all these pieces right will enable the future network platform to deliver optimal application performance by leveraging emerging HW innovation that is intelligently distributed throughout the network, while continuing to harvest the operational and business benefits of cloud computing models.

Figure 1 positions the four key challenges in relation to the orchestration/operations support systems (OSS) layer, the application layer, run-time and the operating system/hypervisor. The lower part of the figure provides some examples of specialized HW in a telco environment, which includes domain-specific accelerators, next-generation memory and storage, and novel interconnect technologies.

Compute and storage trends

With the inevitable end of Moore's Law [2], developers can no longer assume that rapidly increasing application resource demands will be addressed by the next generation of faster general-purpose chips. Instead, commodity HW is being augmented by a very heterogeneous set of specialized chipsets, referred to as domain-specific accelerators, that attempt to provide both cost and energy savings.

For instance, data-intensive applications can take advantage of the massive scope for parallelization

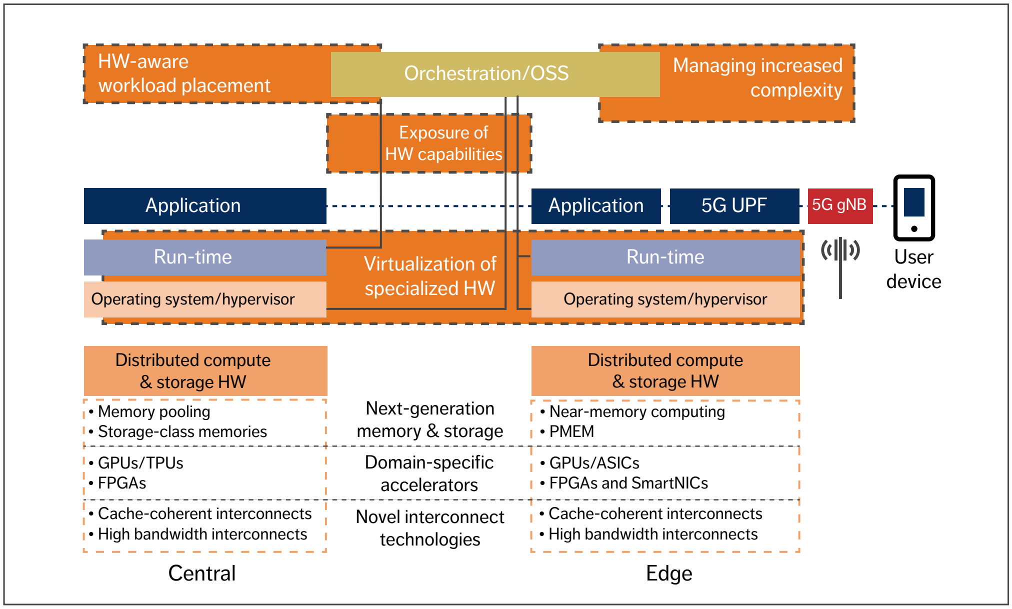
HW-aware workload placement
Orchestration/OSS
Managing increased complexity
Exposure of HW capabilities
Application
Application
5G UPF
5G gNB
User device
Run-time
Virtualization of specialized HW
Run-time
Operating system/hypervisor
Operating system/hypervisor
Distributed compute & storage HW
Distributed compute & storage HW
• Memory pooling
• Storage-class memories
Next-generation memory & storage
• Near-memory computing
• PMEM
• GPUs/TPUs
• FPGAs
Domain-specific accelerators
• GPUs/ASICs
• FPGAs and SmartNICs
• Cache-coherent interconnects
• High bandwidth interconnects
Novel interconnect technologies
• Cache-coherent interconnects
• High bandwidth interconnects
Central
Edge

Figure 1 Impact of the four key challenges on the stack (top) and heterogeneity of HW infrastructure (bottom)

in graphics processing units (GPUs) or tensor processing units (TPUs), while latency-sensitive applications or locations with limited power budgets may utilize field-programmable gate arrays (FPGAs). These trends point to a rapidly increasing adoption of accelerators in the near future.

The growing demand for memory capacity from emerging data-intensive applications must be met by upcoming generations of memory. Next-generation memories aim to blur the strict dichotomy between classical volatile and persistent storage technologies – offering the capacity and persistence features of storage, combined with the byte-addressability and access speeds close to today's random-access memory (RAM) technologies. Such persistent memory (PMEM) technologies [3] can be used either as large terabyte scale volatile memory, or as storage with better latency and bandwidth relative to solid-state disks.

> “ THESE TRENDS POINT TO A RAPIDLY INCREASING ADOPTION OF ACCELERATORS... ”

3D silicon die-stacking has facilitated the embedding of compute units directly inside memory and storage fabrics, opening a paradigm of near-memory processing [1], a technology that reduces data transfer between compute and storage and improves performance and energy consumption.

Finally, advancements in interconnect technologies will enable faster speeds, higher capacity and lower latency/jitter to support communication between the various memory and processing resources within nodes as well as within clusters. The cache coherency properties of modern interconnect technologies, such as Compute Express Link [4] and Gen-Z, can enable direct access to configuration registers and memory regions across the compute infrastructure. This will simplify the programmability of accelerators and facilitate fine-grained data sharing among heterogeneous HW.

Supporting heterogeneous hardware in distributed cloud

While the combination of heterogeneous HW and distributed compute resources poses unique challenges, there are mechanisms to address each of them.

Virtualization of specialized hardware

The adoption of specialized HW in the cloud enables multiple tenants to use the same HW under the illusion that they are the sole user, with no data leakage between them. The tenants can request, utilize and release accelerators at any time using application programming interfaces (APIs). This arrangement requires an abstraction layer that provides a mechanism to schedule jobs to the specialized HW, monitor their resource usage and dynamically scale resource allocations to meet performance requirements. It is pertinent to keep the overhead of this virtualization to a minimum. While virtualization techniques for common COTS HW (x86-based central processing units (CPUs), dynamic RAM (DRAM), block storage and so on) have matured well during recent decades, corresponding virtualization techniques for domain-specific accelerators are largely still missing for production-grade systems.

Exposure of hardware capabilities

Current cloud architectures are largely agnostic to the capabilities of specialized HW. For example, all GPUs of a certain vendor are treated as equivalent, regardless of their exact type or make. To differentiate them, operators typically tag the nodes equipped with different accelerators with unique tags and the users request resources with a specific tag. This model is very different to general-purpose CPUs and can therefore lead to complications when a user requires combinations of accelerators.

Current deployment specifications also do not have good support for requesting partial allocation of accelerators. For accelerators that can be partitioned today, the decomposition of a single physical accelerator into multiple virtual accelerators must be done manually. Addressing these issues will require appropriate abstractions and models of specialized HW, so that their capabilities can be interpreted and incorporated by orchestration functions.

The need for appropriate models will be further amplified in the case of distributed compute and storage. Here, the selection of the optimal site location will depend on the application requirements (bounded latency or throughput constraints, for example) and the available resources and HW capabilities at the sites. The programming and orchestration models must be able to select appropriate software (SW) options – SW only in the case of moderate requirements, for example, or SW complemented with HW acceleration for stringent requirements.

As SW deployment options with or without HW acceleration may have significantly different resource footprints, sites must expose their HW capabilities to be able to construct a topology map of resources and capabilities. During exposure and abstraction, proprietary features and the interfaces to them must be hidden and mapped to (formal or informal) industry standards that are hopefully coming soon. Modeling and abstraction of resources and capabilities are necessary prerequisites to be able to select the appropriate location and application deployment options and flavors.

Orchestrating heterogeneous distributed cloud

Based on a global view of the resources and capabilities within the distributed environment, an orchestration system (OSS in telco terminology) typically takes care of designing and assigning application workloads within the compute and storage of the distributed environment. This means that decisions regarding optimal workload placement also should factor in the type of HW components available at the sites related to the requirements of the specific application SW.

Due to the pricing of and power constraints on existing and upcoming HW accelerators,

Definition of key terms

Edge computing provides distributed computing and storage resources closer to the location where they are needed/consumed.

Distributed cloud provides an execution environment for cloud application optimization across multiple sites, including required connectivity in between, managed as one solution and perceived as such by the applications.

Hardware accelerators are devices that provide several orders of magnitude more efficiency/ performance compared with software running on general purpose central processing units for selected functions. Different hardware accelerators may be needed for acceleration of different functions.

Persistent memory is an emerging memory technology offering capacity and persistence features of block-addressable storage, combined with the byte-addressability and access speeds close to today's random-access memory technologies. It is also referred to as storage-class memory.

Moore's law holds that the number of transistors in a densely integrated circuit doubles about every two years, increasing the computational performance of applications without the need for software redesign. Since 2010, however, physical constraints have made the reduction in transistor size increasingly difficult and expensive.

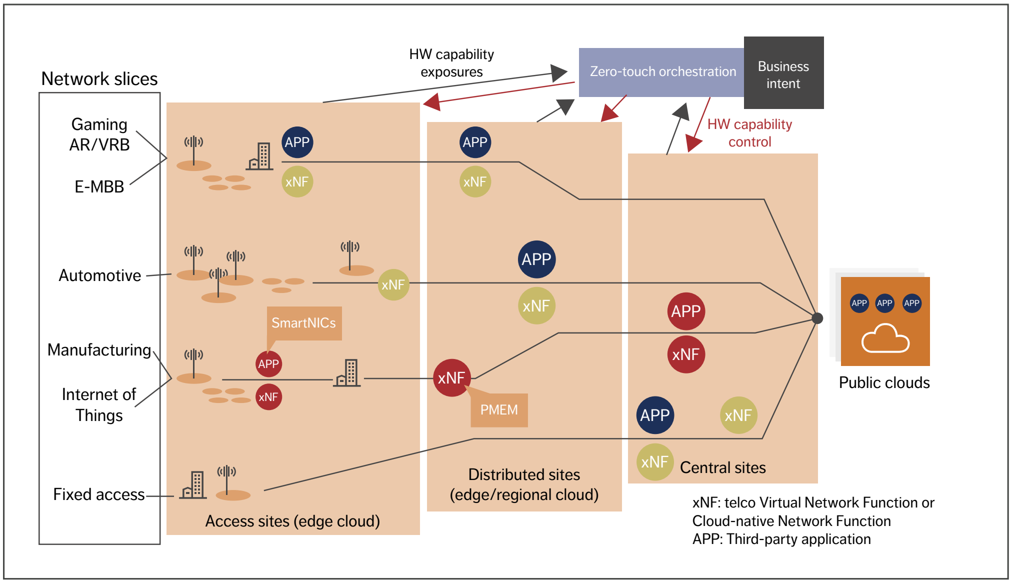

Figure 2 Integrated network slicing (telco) and third-party applications

they are expected to be scarce among edge-cloud sites, which in turn will require mechanisms to employ prioritization and preemption of workloads. Unlike conventional IT cloud environments, distributed cloud allows considerations of remote resources and capabilities.

Moreover, telco applications and workloads hosted in telco clouds may require much stricter Service Level Agreements (SLAs) to be fulfilled. Prioritization and preemption for new workloads may only be a viable option if capabilities or resources are already taken. However, it is important to migrate evicted workloads either to a new location, or to a new SW and HW deployment option to minimize service disruption during preemption.

Managing increased complexity

Traditional automation techniques based on human scripting and/or rule books cannot scale to address the complexity of the heterogeneous distributed cloud. We can already see a shift away from human-guided automation to machine-reasoning-based automation such as cognitive artificial intelligence (AI) technologies. Specifically, a paradigm is emerging where the human input to the cloud system will be limited to specifying the desired business objectives (intents). The cloud system then figures how best to realize those objectives/intents.

Figure 2 presents an exemplary distributed cloud scenario with access sites, regional and central DCs and public clouds. It is based on the assumption that the manufacturing network slice (red) includes both telco (xNF) and third-party workloads (APP), out of which one APP requires network acceleration (SmartNIC), while another xNF depends on PMEM.

Multiple network slices are created based on customer need. Network slices differ not only in their service characteristics, but are separated and isolated from each other. Aggregated views of HW accelerators per location are collected for the zero-touch orchestration service (gray arrows).

When a service request arrives, the orchestration service designs the service instance topology and assigns resources to each service component instance (red arrows). These actions are based on the actual service requirements, the service access points and the business intent.

Opportunities and use cases

In terms of the opportunities in support of the ongoing cloudification of telco networks, let us consider the case of RAN. The functional split of higher and lower layers of the RAN protocol makes it possible to utilize Network Functions Virtualization (NFV) and distributed compute infrastructure to achieve ease of deployment and management. The asynchronous functions in the higher layer may be able to be run on COTS HW.

However, a set of specialized HW will be required to meet the stringent performance criteria of lower-layer RAN functions. For instance, the time-synchronous functions in the medium-access control layer, such as scheduling, link adaptation, power control, or interference coordination, typically require high data rates on their interfaces that scale with the traffic, signal bandwidth and number of antennas. These cannot be easily met with current general-purpose processing capabilities.

Likewise, deciphering functions in the packet data convergence protocol layer, compression/decompression schemes of fronthaul links and channel decoding and modulation functions in the physical layer would all benefit from HW acceleration.

The security requirements for data flows across the backhaul for 4G/5G RANs mandate the use of IP security protocols (IPsec). By offloading encrypt/decrypt functions to specialized HW such as Smart Network Interface Controllers (SmartNICs), application-specific integrated circuits (ASICs) or FPGAs, the processing overhead associated with IPsec can be minimized. This is crucial to support higher data rates in the transport network.

The network data analytics function in 5G Core networks would benefit from GPUs to accelerate training of machine learning (ML) models on live network data. The enhancements to interconnects (cache coherency, for example) make it easier for the various accelerators and CPUs to work together. The interconnects also enable low latencies and high bandwidths within sites and nodes. There is increasing demand on memory from several core network functions (user-database functions, for example), both from a scale and a latency perspective. The scale of PMEM can be intelligently combined with the low latency of double data rate memories to address these requirements.

While these opportunities are specific to telecommunication providers, there are also several classes of third-party applications that would benefit from distributed compute and storage capabilities within the telco infrastructure. Industry 4.0 includes several use cases that could utilize HW-optimized processing. Indoor positioning typically requires the processing of high-resolution images to accurately determine the location of an object relative to others on a factory floor. This is computationally intensive and GPUs/FPGAs are typically used. Likewise, the application of augmented reality (AR)/virtual reality (VR) technologies in smart manufacturing for remote assistance, training or maintenance will rely significantly on HW acceleration and edge computing to optimize performance and reduce latencies.

The gaming industry is also witnessing significant technology shifts – specifically, remote rendering and mixed-reality technologies will have a profound impact on the consumer experience. These technologies rely on an underlying distributed cloud infrastructure that has HW acceleration capabilities at the edge to offload the processing from consumer devices, while maintaining strict latency bounds.

Furthermore, several use cases in the automotive industry involve strict latency requirements that demand HW acceleration in the form of GPUs and FPGAs at remote sites. Examples include real-time object detection in video streams that are processed by either vehicles or road-side infrastructure.

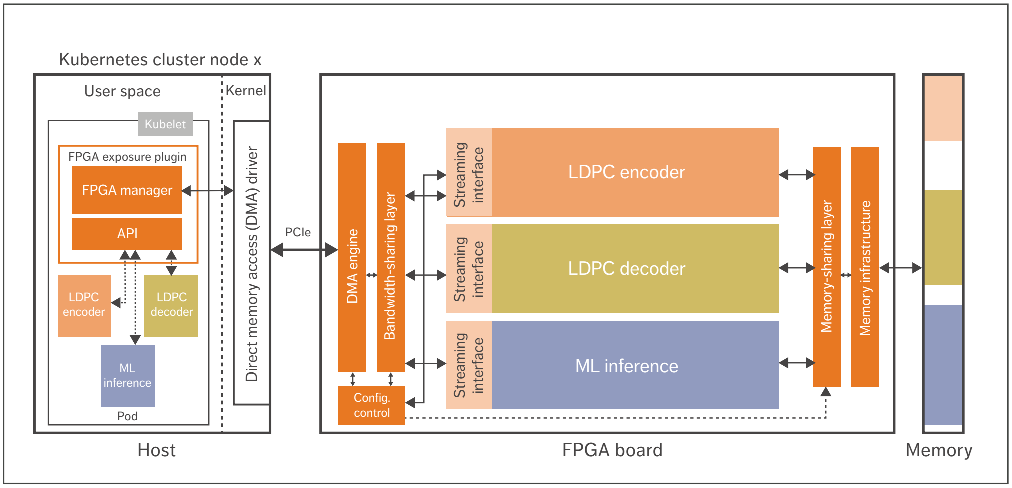

Figure 3 FPGA sharing solution

Initial proof points

Our ongoing research in the area of distributed cloud has yielded several initial proof points that demonstrate Ericsson's leadership in terms of fully leveraging heterogeneous compute and storage capabilities to deliver best-in-class application performance to our customers.

Virtualized heterogeneous hardware

The 3GPP network evolution requires a dramatic increase in compute capacity when increasing carrier bandwidths up to terahertz level and the addition of more MIMO (multiple-input, multiple-output) layers and antenna ports. High-end GPUs could provide the required compute capacity. To understand how 5G baseband can be implemented on GPUs, Ericsson has entered into a collaboration with NVIDIA. Early findings show that the GPU instruction set and CUDA (Compute Unified Device Architecture) SW design patterns for key receiver algorithms excel in comparison with CPU-only solutions, indicating that further investigation of GPUs for this purpose is indeed a viable direction.

As high-end GPUs tend to be expensive and may not be available in every node of the distributed cloud, Ericsson Research has also devised a solution to enable GPU access by applications hosted on nodes without local GPUs, both enabled by OpenStack and Kubernetes. Given a high-speed network between the nodes, GPU requests are locally intercepted and redirected to the remote GPU servers for execution.

FPGAs may be another alternative to meet the requirements of RAN functions and third-party applications hosted at the telco edge. At Ericsson Research, we have enabled multi-tenancy support on FPGAs through Kubernetes, so that multiple-application containers that need HW acceleration can share the FPGA's internal resources, off-chip DRAM and peripheral component interconnect express (PCIe) bandwidth, as shown in *Figure 3*.

On the left side, three application containers (pods) on a node are managed through Kubernetes. Our FPGA exposure plugin supports the pods by offering three isolated, virtual FPGAs. On the right side, three separate FPGA regions (one per application container) are managed and exposed through our HW management layer, comprising of PCI bandwidth and memory-sharing functions (the orange boxes in the FPGA part of the figure).

An HW management layer uses partial reconfiguration technology to support spatial partitioning of one FPGA to share its resources. It is comprised of a bandwidth-sharing layer for dynamic allocation of PCIe bandwidth. The memory-sharing layer protects tenants from data leaks on the off-chip DRAM. It adds a protection layer for internal reconfiguration to prevent unintended configurations. This solution has been validated using multiple regions, hosting 5G low-density parity-check (LDPC) code encoder and decoder accelerators next to an ML inference function. However, essentially any telco or third-party application accelerator could be deployed within these partial regions, whose number could vary depending on the FPGA size and the application's resource requirements.

Virtual switches form an integral part of distributed cloud infrastructure, providing network access to virtual machines by linking the virtual and physical network interfaces. The overhead of these virtual switches is one of the main obstacles to achieving high throughput in the packet-processing functions. Through our solution to off-load the data plane of a virtual switch onto the specialized HW (FPGA-based SmartNICs in this case), we achieve the highest possible network input/output performance in a virtualized environment and free up CPUs for the execution of other functions and applications.

Orchestrating heterogeneous distributed cloud

There is a need for intelligent workload placement schemes to meet the diverse acceleration needs of 5G use cases that require domain-specific HW. We have two proof points related to this issue.

The first is our development of an HW-aware service instance design and workload placement to optimize the deployment of workloads at the edge. We have demonstrated an on-demand service instance design, prioritization and preemption for the telco Packet Core Gateway (PCG) user-plane function (UPF) over a Kubernetes edge cluster, in which only some nodes are equipped with PMEM. The PCG-UPF is considered a high-priority workload, which could use PMEM to distribute database instances needed for resiliency.

At the instantiation time of the PCG-UPF, another workload using the PMEM was evicted. The challenge was to migrate the lower priority workload to a new host, considering its original service requirements. Here, the lower priority workload shared an affinity with other components,

Terms and abbreviations

AI – Artificial Intelligence | **API** – Application Programming Interface | **AR** – Augmented Reality | **ASIC** – Application-Specific Integrated Circuit | **COTS** – Commercial Off-the-Shelf | **CPU** – Central Processing Unit | **DC** – Data Center | **DRAM** – Dynamic Random-Access Memory | **E-WBB** – Enhanced Mobile Broadband | **FPGA** – Field-Programmable Gate Array | **GNB** – GNodeB (3GPP next-generation base station) | **GPU** – Graphics Processing Unit | **HW** – Hardware | **IPsec** – IP Security Protocols | **LDPC** – Low-Density Parity-Check | **ML** – Machine Learning | **NFV** – Network Functions Virtualization | **OSS** – Operations Support Systems | **PCG** – Packet Core Gateway | **PCIe** – Peripheral Component Interconnect Express | **PMEM** – Persistent Memory | **RAM** – Random-Access Memory | **SLA** – Service Level Agreement | **SmartNIC** – Smart Network Interface Controller | **TPU** – Tensor Processing Unit | **UPF** – User-Plane Function | **VR** – Virtual Reality | **xNF** – network functions, both telco and third party which were automatically migrated together with the workload blocking the PMEM needed by the PCG-UPF. The service disruption time for the evicted traffic was minimized with the instant triggering of the migration workflow.

Our second proof point demonstrates how we can enable distributed applications to utilize the enhanced capacity and persistency of non-volatile memories in a transparent fashion. While PMEM could be used to make up for the slow growth of DRAM capacity in recent years, it can lead to performance degradation due to PMEM's slightly higher latencies.

The memory-tiering concept developed by Ericsson Research enables the dynamic placement/migration of application data across DRAM and PMEM, based on observed application behavior. This infrastructure, using low-level CPU performance counters to drive placement decisions, achieves performance similar to DRAM, without any changes to the application, while using a mix of PMEM and DRAM.

> “... PAVING THE WAY TOWARD AN INTEGRATED NETWORK COMPUTE FABRIC THAT IS UNIVERSALLY AVAILABLE ”

Complexity management

A heterogeneous and distributed cloud implies high complexity in service assurance, and more specifically, high complexity in terms of continually finding optimal configurations in dynamically changing environments. Ericsson's cognitive layer has demonstrated cloud service assurance while satisfying contracted SLAs.

This cognitive process evaluates the effect of a proposed new service deployment on all existing services and their SLA fulfilment. Furthermore, the cognitive layer continuously reevaluates whether the current deployment of all services is still optimal, and searches for improved configurations and uses predictive models to drive proactive actions to be taken, thereby enabling intelligent autonomous cloud operations.

Conclusion

The cloud is becoming more and more distributed at the same time that compute and storage capabilities are becoming increasingly diverse. Datacenters are emerging at the edges of telco networks and on customer premises and hardware accelerators are becoming essential components of formerly software-only services. Due to the inevitable end of Moore's law, the importance and use of hardware accelerators will only continue to increase, presenting a significant challenge to existing solutions for exposure and orchestration.

To address these challenges, Ericsson is innovating in three key areas. Firstly, we are using virtualization techniques for domain-specific accelerators to support sharing and multi-tenant use of specific hardware. Secondly, we are using zero-touch orchestration for hardware accelerators, which includes hardware capability exposure and aggregation for the orchestration system, as well as automated mechanisms to design and assign service instances based on the abstract map of resources and accelerator capabilities. And thirdly, we are using artificial intelligence and cognitive technologies to address the technical complexity and to optimize for business value.

Redefining cloud to expose and optimize the use of heterogeneous resources is not straightforward, and to some extent goes against the centralization and homogenization trends. However, we believe that our use cases and proof points validate our approach and will gain traction both in the telecommunications community and beyond, paving the way toward an integrated network compute fabric [5] that is universally available across telco networks.

References

1. **ArXiv, Near-Memory Computing: Past, Present, and Future, August 7, 2019, Gagandeep Singh et al., available at:** *https://arxiv.org/pdf/1908.02640.pdf*
2. **Nature, The chips are down for Moore's law, February 9, 2016, M. Mitchell Waldrop, available at:** *https://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338*
3. **Admin magazine, How Persistent Memory Will Change Computing, Jeff Layton, available at:** *https://www.admin-magazine.com/HPC/Articles/Persistent-Memory*
4. **CXL, Compute Express Link, Breakthrough CPU-to-device interconnect, available at:** *https://www.computeexpresslink.org/*
5. **Ericsson, Network compute fabric, available at:** *https://www.ericsson.com/en/future-technologies/network-compute-fabric*

Further reading

- **Ericsson blog, How will distributed compute and storage improve future networks, available at:** *https://www.ericsson.com/en/blog/2020/2/distributed-compute-and-storage-technology-trend*
- **Ericsson white paper, Edge computing and deployment strategies for communication service providers, available at:** *https://www.ericsson.com/en/reports-and-papers/white-papers/edge-computing-and-deployment-strategies-for-communication-service-providers*
- **Ericsson blog, Cloud evolution: the era of intent-aware clouds, available at:** *https://www.ericsson.com/en/blog/2019/5/cloud-evolution-the-era-of-intent-aware-clouds*
- **Ericsson blog, What is network slicing?, available at:** *https://www.ericsson.com/en/blog/2018/1/what-is-network-slicing*
- **Ericsson blog, Virtualized 5G RAN: why, when and how?, available at:** *https://www.ericsson.com/en/blog/2020/2/virtualized-5g-ran-why-when-and-how*
- **Ericsson Technology Review, Cognitive technologies in network and business automation, available at:** *https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/cognitive-technologies-in-network-and-business-automation*
- **Ericsson, A guide to 5G network security, available at:** *https://www.ericsson.com/en/security/a-guide-to-5g-network-security*
- **Ericsson, 5G Core, available at:** *https://www.ericsson.com/en/digital-services/offerings/core-network/5g-core*
- **Ericsson, Edge computing, available at:** *https://www.ericsson.com/en/digital-services/trending/edge-computing*

THE AUTHORS

Wolfgang John

◆ is a research leader and scientist at Ericsson Research in Stockholm. His current research focuses primarily on distributed cloud computing systems and platform concepts for both telco and IT applications. Since joining Ericsson in 2011, he has also done research on NFV, software-defined networking and network management. John holds a Ph.D. in computer engineering from Chalmers University of Technology in Gothenburg, Sweden, and has coauthored more than 50 scientific papers and reports, as well as several patent families.

Chandramouli Sargor

◆ currently heads the AI-inspired Design team within the Ericsson Global AI Accelerator in Bangalore, India. He joined Ericsson in 2007 and previously headed the Ericsson Research team in Bangalore with a focus on advanced cloud and AI technologies, including the use of emerging HW to build disruptive cloud compute solutions. Sargor holds a B. Tech. in electrical engineering from the Indian Institute of Technology in Mumbai and an M.S. in computer engineering from North Carolina State University in the US. He has coauthored more than 25 patents and several publications.

Robert Szabo

◆ joined Ericsson in 2013 and currently serves as a principal researcher in Cloud Systems and Platforms at Ericsson Research in Hungary. At present, his work focuses primarily on distributed/edge cloud, zero-touch automation for orchestration and NFV. Szabo is the coauthor of more than 80 publications and holds a both a Ph.D. in electrical engineering and an MBA from the Budapest University of Technology and Economics, Hungary.

Ahsan Javed Awan

◆ is an experienced researcher of warehouse-scale computers at Ericsson Research. Prior to joining Ericsson in 2018, he worked at Imperial College London, IBM Research – Tokyo in Japan, and Barcelona Supercomputing Center in Spain. Awan holds an Erasmus Mundus joint Ph.D. from KTH Royal Institute of Technology in Stockholm, Sweden and the Universitat Politècnica de Catalunya in Barcelona, Spain, for his work on performance characterization and optimization of in-memory data analytics on a scale-up server.

The authors would like to thank Jörg Niemöller, Fetahi Wuhib, Andrew Williams, Torbjörn Keisu, Daniel Seiler, Jonas Falkenå, Jonas Bjurel, Tobias Lindqvist and Azimeh Sefidcon for their contributions to this article.

Chakri Padala

◆ joined Ericsson in 2007 and currently serves as a master researcher within Cloud Systems and Platforms at Ericsson Research in Bangalore, India. His research interests include new memory/storage technologies, acceleration of SW functions and operating system stacks. Padala has an M.S. from the University of Louisiana at Lafayette in the US, and a B.Tech. from the National Institute of Technology in Warangal, India.

Edvard Drake

◆ joined Ericsson in 1993. Since the early 2000s, he has been deeply engaged in technology relations with many of the major technology vendors, primarily as part of the Multimedia, Operations Support Systems/Business Support Systems and Digital Services parts of the Ericsson organization, gathering insights into many aspects of technology evolution. Drake currently serves as a technology expert in the area of platform technologies, working with both technology intelligence/scouting as well as with architecture. He holds a B.Sc. in computer science from Umeå University in Sweden.

Martin Julien

◆ joined Ericsson in 1995 and currently serves as a senior specialist in cloud systems and platforms. With deep expertise in distributed systems, networking and optical interconnects, he has played a significant role in the development of innovative cloud system infrastructure products. Julien's current work mainly focuses on cloud acceleration and cloud intelligence, taking advantage of advanced hardware offload capabilities and AI technologies. He holds a B. Eng. from Sherbrooke University in Canada.

Miljenko Opsenica

◆ joined Ericsson in 1998 and currently serves as a master researcher at Ericsson Research in Finland (NomadicLab), where he is working on cloud architectures and technologies, orchestration frameworks and automation. Opsenica also leads Ericsson Research's integrated connectivity and edge program, which focuses on integrated edge architecture, cross-resource domain interactions and performance optimization management. He holds an M.Sc. in electrical engineering and computing from the University of Zagreb in Croatia.