RESEARCH

Hewlett Packard Enterprise

intel

analysys mason

White paper

Data plane acceleration technologies: realising the potential of network virtualization

February 2019

Gorkem Yigit and Caroline Chappell

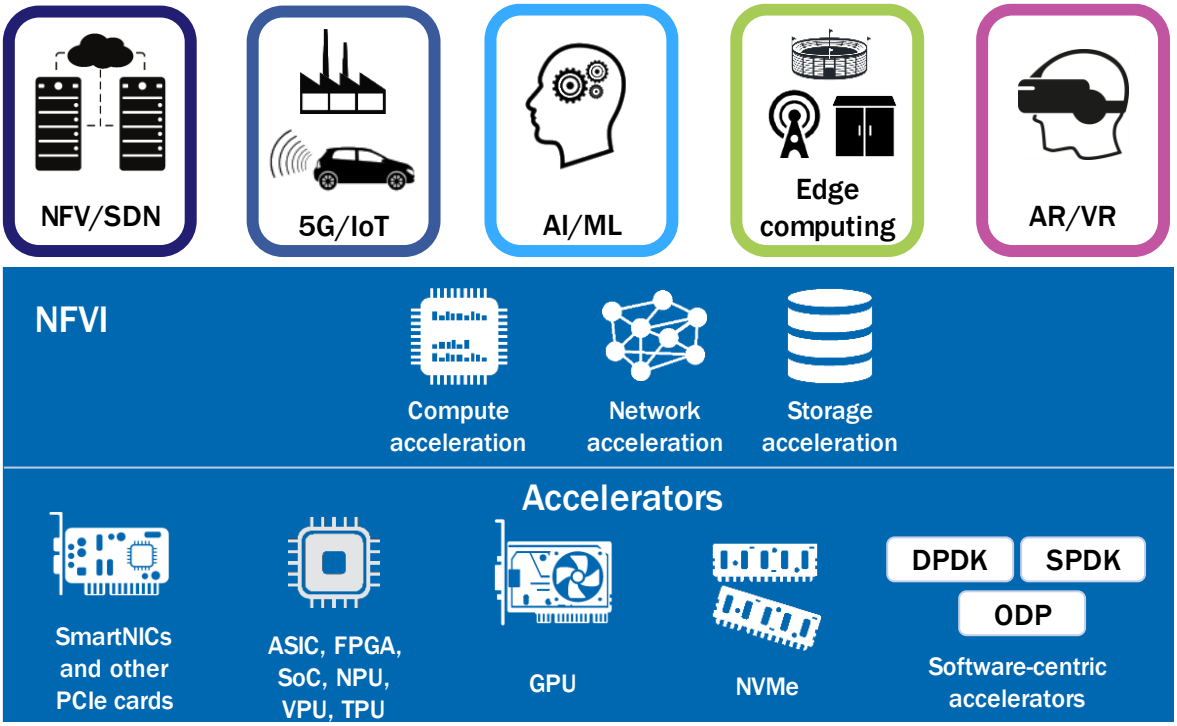
Contents

List of figures

1. Executive summary

Network operators have virtualized many control plane network functions but they face technical limitations and cost disincentives when they migrate data plane functions to general-purpose network function virtualization infrastructures (NFVI) that are based on x86 platforms. Many acceleration technologies are available to address these challenges, from software-centric acceleration approaches to hardware-centric solutions. These can be used to offload compute-intensive virtual network function (VNF) workloads and reclaim virtualization overhead. For some use cases, such as AI/ML and video processing, graphics processing units (GPUs) and vision processing units (VPUs) can provide significant performance enhancements. However, operators are confused about the benefits of, and the dependencies between, each approach, as well as when to use each to support specific network functions and use cases.

Figure 1.1: Acceleration technologies and use cases [Source: Analysys Mason, 2019]



This white paper aims to improve operators' understanding and awareness of acceleration technologies in order to maximise the performance and efficiency of their NFV/SDN, 5G and AI/ML applications. The paper outlines the key drivers for acceleration technologies and identifies the main use cases and functions that require acceleration. It introduces a range of software-centric and hardware-centric acceleration technologies, analyses their advantages and challenges, as well as their suitability for different VNF and NFVI use cases. It also presents performance results for accelerating virtualized security functions. Finally, it discusses the need for an acceleration technology abstraction layer and a uniform management framework to simplify and automate the operations of heterogenous acceleration resources with an 'as-a-service' model. It highlights industry initiatives such as OpenStack Cyborg and OPNFV DPACC, which aim to help operators achieve these goals.

2. Drivers for acceleration technologies

This section discusses the benefits of having a common infrastructure and operations for all VNF workloads based on general purpose hardware and software. It explains the challenge that this poses for VNFs that require high performance and throughput to perform data plane tasks. It suggests that an NFVI based on general-purpose CPUs should be augmented with acceleration technologies that offload compute-intensive VNF and NFVI workloads, and then identifies the use cases and functions that need this capability.

2.1 Acceleration technologies are needed to move network virtualization forward

NFV and software-defined networking (SDN) underpin operators' aspirations to become more efficient, competitive and innovative. Many operators have already passed the early stages of network virtualization, having implemented NFV/SDN within specific domains that are limited in scope and size. Operators are now looking to scale up these deployments to support the delivery of dynamic, on-demand network services (including vCPE/uCPE, SD-WAN and managed security) and to lay the groundwork for high-capacity and low-latency 5G networks (such as edge, cloud RAN and vEPC). Analysys Mason forecasts that operator spending on NFV and SDN will increase from USD3.7 billion in 2017 to USD23.8 billion in 2022, growing at a CAGR of 45%.[1]

Network virtualization is moving from a phase of closed, vertically siloed, custom appliances to one where multiple VNFs from different vendors run on a horizontal, open and highly automated platform. This will require network functions to be fully decoupled from specialised appliances and implemented as cloud-native software components (VNFs) on a common NFVI. The NFVI should consist of standard, off-the-shelf compute and memory resources and a virtualization layer. Eliminating interdependencies between hardware and software is essential if operators are to realise the financial and operational benefits of network virtualization. Operators can lower capex using inexpensive commercial off-the-shelf (COTS) hardware and white boxes, which reduce vendor lock-in and leverage IT market commoditization. They can also shrink opex through the use of low-cost IT automation tools. Together, these capabilities deliver a simpler, more flexible and programmable infrastructure on which to run the network.

Achieving disaggregation of network hardware and software is a key focus for operators during their network virtualization journey, but it is not without challenges. Standard, general-purpose hardware platforms are usually adequate for handling many use cases. However, some network functions and workloads (especially those that are related to the data plane) require a high level of QoS (for areas such as throughput, latency and jitter), predictable performance and security. When implementing such VNFs on COTS hardware, it can be difficult to achieve parity in these respects with custom network appliances. There are several reasons for this.

- VNFs are not yet fully optimised to handle these tasks as efficiently as physical functions and they consume large amounts of CPU resource.
- The virtualization layer in the NFVI adds further overhead.
- Moore's Law gains are slowing, pushing up the cost of CPU cycles at a time when growing network bandwidth and workload requirements are maxing out CPU clock speeds.

[1] For more information, see Analysys Mason's *Digital infrastructure: worldwide forecast 2018–2022*. Available at: www.analysysmason.com/Research/Content/Reports/digital-infrastructure-forecast-rma16.

As the scale and scope of virtualized networks grows, adding CPUs and servers is a suboptimal way to tackle this problem. It is likely to result in high total cost of ownership (TCO) due to additional hardware, power, space and cooling costs. It also fails to address the technological limitation imposed by virtualization overheads from the multiple layers of packet processing needed as traffic flows from Network Interface Cards (NICs) to VNFs. This limitation makes it difficult to satisfy SLA and QoS requirements, especially for ultra-low-latency 5G use cases.

To match, or even surpass, the performance and latency levels of traditional appliances without negating the capex and opex benefits of virtualization, as well as to avoid virtualization overheads, NFV/SDN networks need to offload key compute-intensive VNF workloads from general-purpose CPUs to additional hardware and software acceleration components in the NFVI. These components help operators to address increasing overhead costs (which cannot be billed to end customers) by reducing the cost-per-bit. They also help operators to achieve the optimal allocation of computing, networking and storage for different types of VNF.

2.2 Which virtualization use cases require acceleration?

VNFs responsible for executing data plane functions fulfil a variety of specific networking, security and media-related tasks such as switching, routing, traffic management, cryptography (SSL, IPSEC), compression and transcoding. Data plane functions need to process and forward traffic at near-line rates, that is, at the same speed as their network interfaces (for example, 40G or 100G), so they suffer from NFVI and VNF performance bottlenecks and cost-effective scalability if they rely solely on general-purpose hardware and software-based resources to do this. Use cases that involve the service chaining of multiple VNF components can add further load on the infrastructure. In these cases, the throughput requirement is amplified and the tunnel processing associated with overlay networks (GRE, VXLAN) may add further overhead. The virtualization use cases and workloads that would benefit most from acceleration are detailed in Figure 2.1 below.

Figure 2.1: Virtualization use cases and their acceleration requirements [Source: Analysys Mason, 2019]

Use cases	Functions	Example acceleration points
vBNG/BRAS	VPN, firewall, DPI, multicast and service chaining	• Layer 2/3/4 packet forwarding (high throughput; for example, 1000+ Gbps) • Overlay networking, Layers 4–7 traffic management (QoS/traffic shaping), pattern matching
vCPE, SD-WAN and vSD-WAN (centralised SD-WAN capabilities as VNFs)	Router, VPN, firewall, DPI and service chaining	• IPsec, encryption/decryption, pattern matching, overlay networking (VXLAN/GRE/MPLS), SD-WAN security monitoring • Centralised, higher-density vSD-WAN nodes that amplifies requirements for all tasks above
vRAN	Baseband PHY layer (Layer 1) for signal processing	• Layer 1 and potentially Layer 2/3 packet processing
vEPC	P/S GW, deep packet inspection (DPI); in 5G core UPF	• Layer 2/3/4 forwarding, 5G core increased throughput requirements (for example, 200+ Gbps) • Overlay networking (VXLAN/GRE/MPLS) • Layers 4–7 traffic management (hierarchical QoS)
vIPsec	Aggregation points (Wi-Fi hotspot, vCPE, service provider edge) and enodeB backhauling	• Cryptography (encryption, decryption), IPsec protocol, SSL record layer processing • Authentication processing
vNGFW	Firewall, Intrusion Prevention Systems (IPS),	• Layer 2/3 forwarding, network address and port translation (NAPT), load balancing, pattern matching

Use cases	Functions	Example acceleration points
	SSL VPN and Deep Packet Inspection (DPI)	• Cryptography (encryption, decryption) • Compression/decompression
vIMS	SBC, MRF	• Media/audio transcoding
AI/ML	Deep learning, neural networks, big data analytics	• Parallel processing, image/face recognition, natural language processing • Data compression/decompression for analytics
Video and graphic processing	IPTV/OTT/cable head-end, edge, cloud gaming and AR/VR	• Video encoding/transcoding, compression
IoT	URLLC use cases: factory automation, robotics, health care, smart transportation	• Latency – each application requires a different range of latency, some as low as <1–10ms • Most of the above functions/workloads also apply because they will be deployed to deliver these IoT applications.

These use cases and functions will be deployed in different parts of operators' networks including 4G/5G core, edge locations (multi-access edge, central offices) and enterprise and video networks. Operators will need to identify the functions/workloads that make most sense to accelerate and choose the most-suitable acceleration/offloading technologies for these functions and their deployment scenarios.

3. Introduction to acceleration technologies

This section introduces the range of acceleration technologies available to NFVI builders and describes their capabilities. It divides acceleration technologies into two categories: software-centric acceleration and hardware-centric and discusses the advantages and challenges of solutions in each category, as well as the potential of hybrid architectures. It summarizes the suitability of the various acceleration solutions for different VNF and NFVI use cases.

3.1 Software-centric acceleration

Software-centric acceleration solutions can be implemented as an additional layer in various parts of virtualized networks, for example, in the CPU, hypervisor and VNFs themselves to augment VNF and NFVI performance. Software-centric acceleration solutions are based on acceleration frameworks, such as the Data Plane Development Kit (DPDK), Open Data Plane (ODP) for SoC (System on Chip) and the Storage Performance Development Kit (SPDK) for storage applications. These frameworks leverage the capabilities of underlying chipsets including CPUs and SoCs and provide a set of libraries, drivers and interfaces that enable developers to build acceleration solutions on top of the underlying hardware for specific network virtualization demands.

Data Plane Development Kit (DPDK)

DPDK is a widely used software-centric acceleration framework created by Intel in 2010. Its source code was made available to developers under the Open Source BSD License in 2017. DPDK provides data plane libraries to accelerate Layer 3 packet processing and throughput performance of virtualized resources on various CPU architectures (such as Intel x86, ARM and POWER) and Network Interface Cards (NICs) from multiple

vendors. Through its developer ecosystem, DPDK supports a growing variety of data plane functions and use cases in virtualized networks (including vSwitch, crypto, compression and baseband acceleration). vSwitch acceleration is one of its most common use cases.

NFVI virtual networking acceleration: accelerated vSwitch, SR-IOV and PCI passthrough

Operators are presented with various NFVI networking options for forwarding traffic from north to south - from network interfaces (for example, NIC) to virtual machines (VMs), as well as east to west and between VNFs and their component VMs or containers. The main approaches to traffic forwarding use one or more of the following: a software-based virtual switch (vSwitch), single root I/O virtualization (SR-IOV) and/or PCI passthrough. There are distinct advantages and disadvantages associated with each approach to NFVI networking and these approaches can be combined in hybrid architectures depending on use case demands.

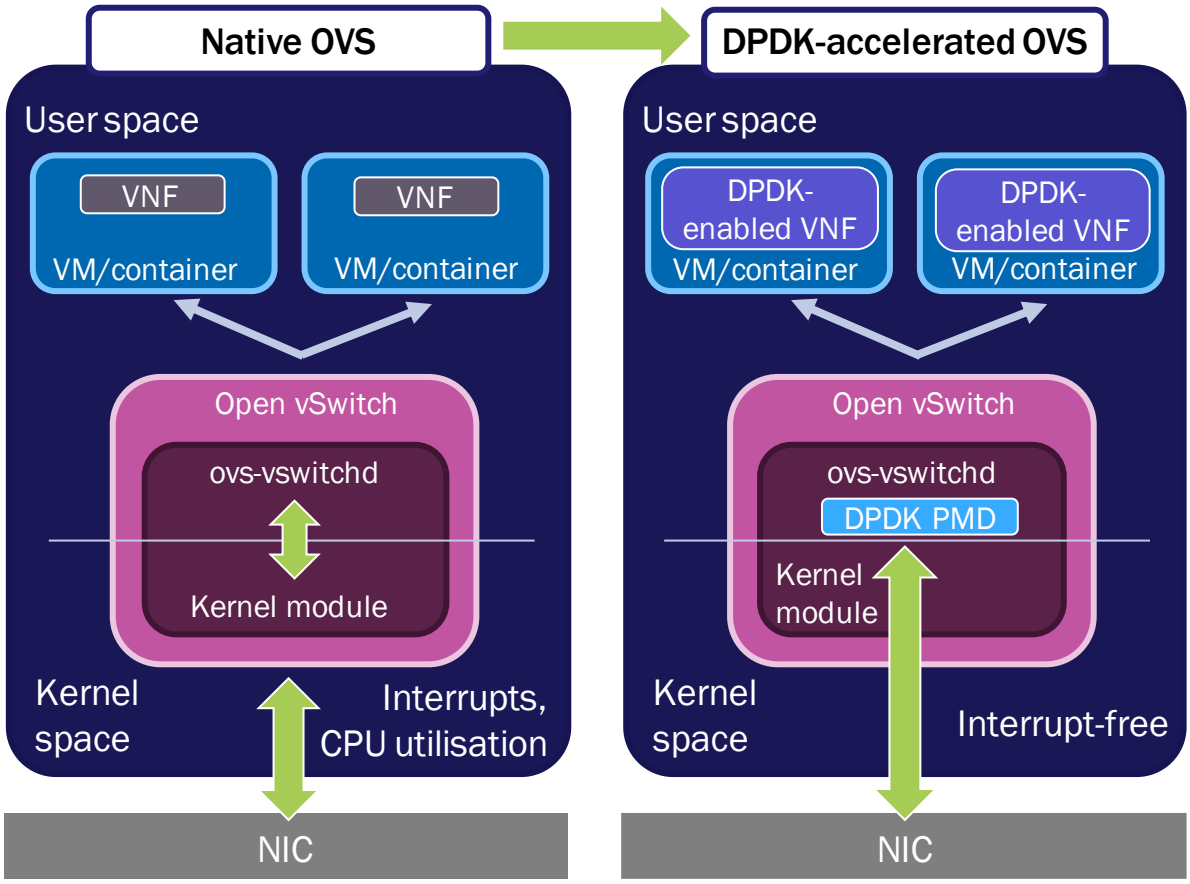
vSwitch is a software layer that sits in the hypervisor, where it aggregates and controls NFVI traffic. It provides isolation from the underlying hardware for network I/O and manages network traffic to and from VNFs. The main advantage of vSwitch is that it is abstracted from the hardware layer and, when coupled with an SDN controller and orchestration, provides a highly flexible and programmable infrastructure. However, its native software implementation without acceleration may not meet the requirements of high-packet processing and throughput use cases. DPDK plays a crucial role in addressing this problem by optimising the way vSwitches process and forward packets. An example of this is DPDK Accelerated Open vSwitch (OVS).[2]

OVS is a vSwitch for Linux-based hypervisors (such as KVM) and consists of two main components: kernel space and user space. Native deployment of OVS introduces performance bottlenecks due to ‘interrupts’ that occur when a packet received from the NIC is first processed in kernel space and then moved to the application in the user space. This process increases CPU utilization and creates overhead. An OVS with DPDK removes this overhead by bypassing the kernel space with an interrupt-free approach. Packets are polled (using Poll Mode Driver (PMD)) and moved directly to the application in user space. DPDK also accelerates the forwarding between VNFs and vSwitches. Figure 3.1 below illustrates OVS acceleration with DPDK. Intel tests show that DPDK can improve OVS performance by around 12 times and continues to improve.[3]

[2] DPDK supports various vSwitch offerings in the market including open-source (OVS, VPP) and vendor-specific solutions (VMware, Cisco). OVS is a widely adopted open-source vSwitch solution and plays an important role in OpenStack and OpenDayLight implementations.

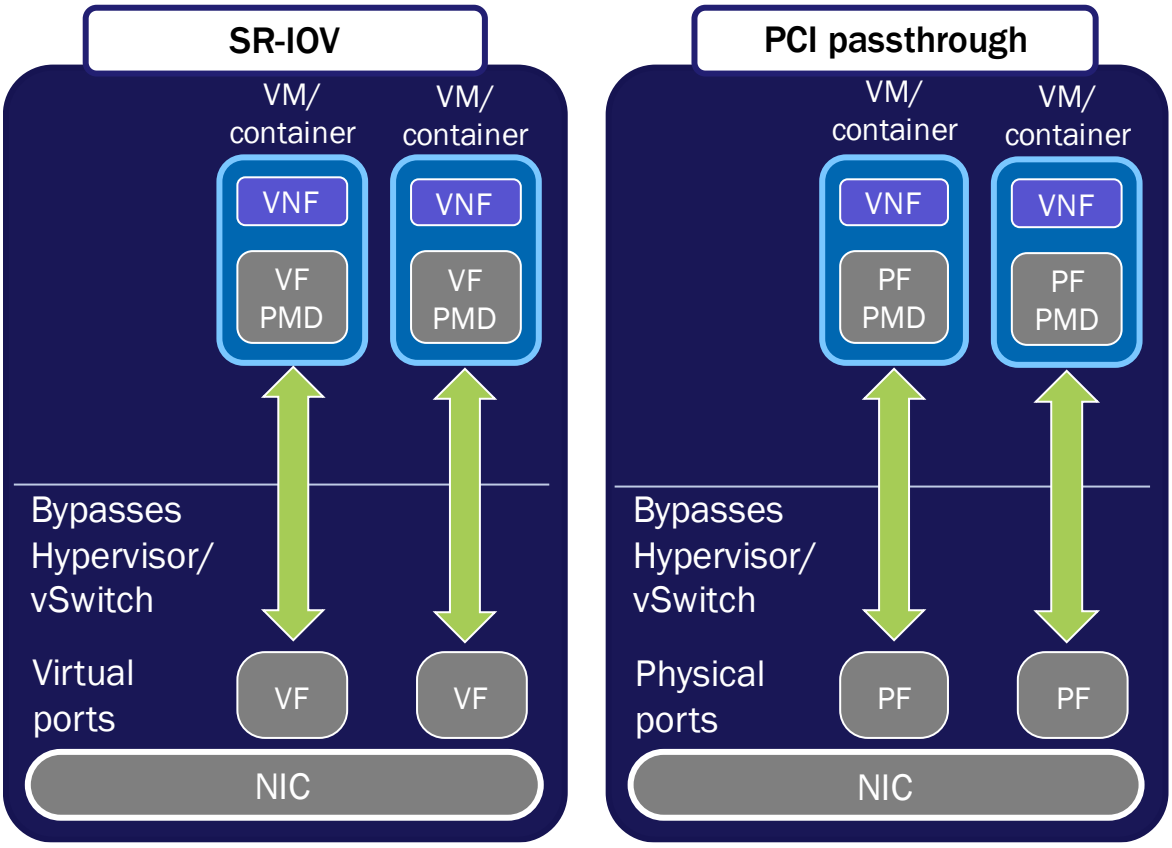
[3] Intel Software Developer Zone (19 December 2016), *Open vSwitch* with DPDK* Overview. Available at: https://software.intel.com/en-us/articles/open-vswitch-with-dpdk-overview.

Figure 3.1: Native OVS and DPDK accelerated OVS [Source: Analysys Mason, Intel 2019]



PCI passthrough and SR-IOV are alternative options to vSwitch. They are similar technologies that enable operators to assign NIC resources directly to specific VMs and containers, bypassing the hypervisor and vSwitch. **PCI passthrough** allows an entire single physical NIC port (PF) to be dedicated to a specific VM. **SR-IOV** is a more-advanced standard than PCI passthrough. It exposes the ports of a single NIC device as shared resources, in other words, as virtual ports (VFs) that can be shared by multiple VMs and containers. Both technologies require hardware drivers that sit inside VMs. Figure 3.2 below provides a description of these approaches.

Figure 3.2: SR-IOV and PCI- passthrough models [Source: Analysys Mason, 2019]



SR-IOV and PCI passthrough have been widely used in virtualized networks because they enable high throughput (near line-rate) and low-latency packet processing at low or no CPU utilization. However, these benefits may come at the expense of hardware independence, flexibility and portability, which are the main goals of NFV/SDN deployments. This is because these approaches tie VNFs rigidly to the infrastructure equipped with these capabilities and bypassing the virtualization layer introduces additional complexity in the management layer. Figure 3.3 below summarises the main advantages and disadvantages of accelerated vSwitch and SR-IOV and PCI passthrough.

Figure 3.3: Comparison of accelerated vSwitch, SR-IOV and PCI passthrough [Source: Analysys Mason, Intel 2019]

Criteria	Accelerated OVS	SR-IOV/PCI passthrough
North/south traffic (maximum throughput to VM/containers)	Lower performance than SR-IOV and PCIpassthrough	Higher (near-line rate performance)
East/west traffic in the same node (VM to VM, for example, service chaining)	Higher performance	Lower performance than accelerated vSwitch because packets need to traverse the network interface in each hop, adding overhead
Latency	Higher than SR-IOV and PCI passthrough	Low
CPU utilization	Higher than SR-IOV and PCI passthrough but significantly lower than native OVS	Low or no impact

Criteria	Accelerated OVS	SR-IOV/PCI passthrough
Hardware dependency/lock-in	No, decoupled from underlying hardware with common interfaces (such as VirtIO)	High, tightly coupled; may require vendor-specific network hardware interfaces and drivers to reside in VMs
SDN programmability	High – thanks to centralised management	Limited – despite being supported by several SDN controllers and orchestrators, its management and control is complex unlike vSwitch
VNF portability	High, supports live migration	Limited – VMs are rigidly tied to specific hardware ports

OVS-DPDK has other limitations beyond providing lower performance than SR-IOV. For instance, because it bypasses the kernel space, it presents certain security challenges. It has been found to reduce performance in cases where overlay networking (VXLAN, GRE) is needed because these packets still need to traverse kernel space, which results in inefficiencies.[4] Other solutions are introduced to tackle these issues in accelerating OVS including TC flower-based OVS, which uses the TC flower classifier in the Linux subsystem to offload packet processing, as well as vendor-specific offerings.

We expect both types of software-centric acceleration to co-exist in the market. Hybrid acceleration architectures that combine vSwitch and SR-IOV/PCI passthrough in the same NFVI create additional complexity, however. Operators may prefer to use the former for use cases that involve service chaining/high inter-VM traffic (such as vBNG, vCPE and Gi-LAN) and use the latter for use cases that require best possible performance and lowest latency (such as vRouters). Operators are increasingly adopting accelerated vSwitch in their NFV/SDN deployments in conjunction with hardware-centric accelerators (for example, SmartNIC) to further enhance NFVI virtual networking performance and achieve a more-complete offload/acceleration of the OVS data path.

3.2 Hardware-centric acceleration

Hardware-centric acceleration solutions consist of specialised hardware components that enable higher performance and efficiency than general-purpose computing, networking and storage resources in an NFVI. Such components provide pure hardware acceleration by augmenting general-purpose resources (working in a complementary way or independently) and can be used in combination with software acceleration technologies (including DPDK, SR-IOV and TC flower) as discussed in section 3.1.

Examples of domains that benefit from offloading tasks to more-efficient hardware-based acceleration components include:

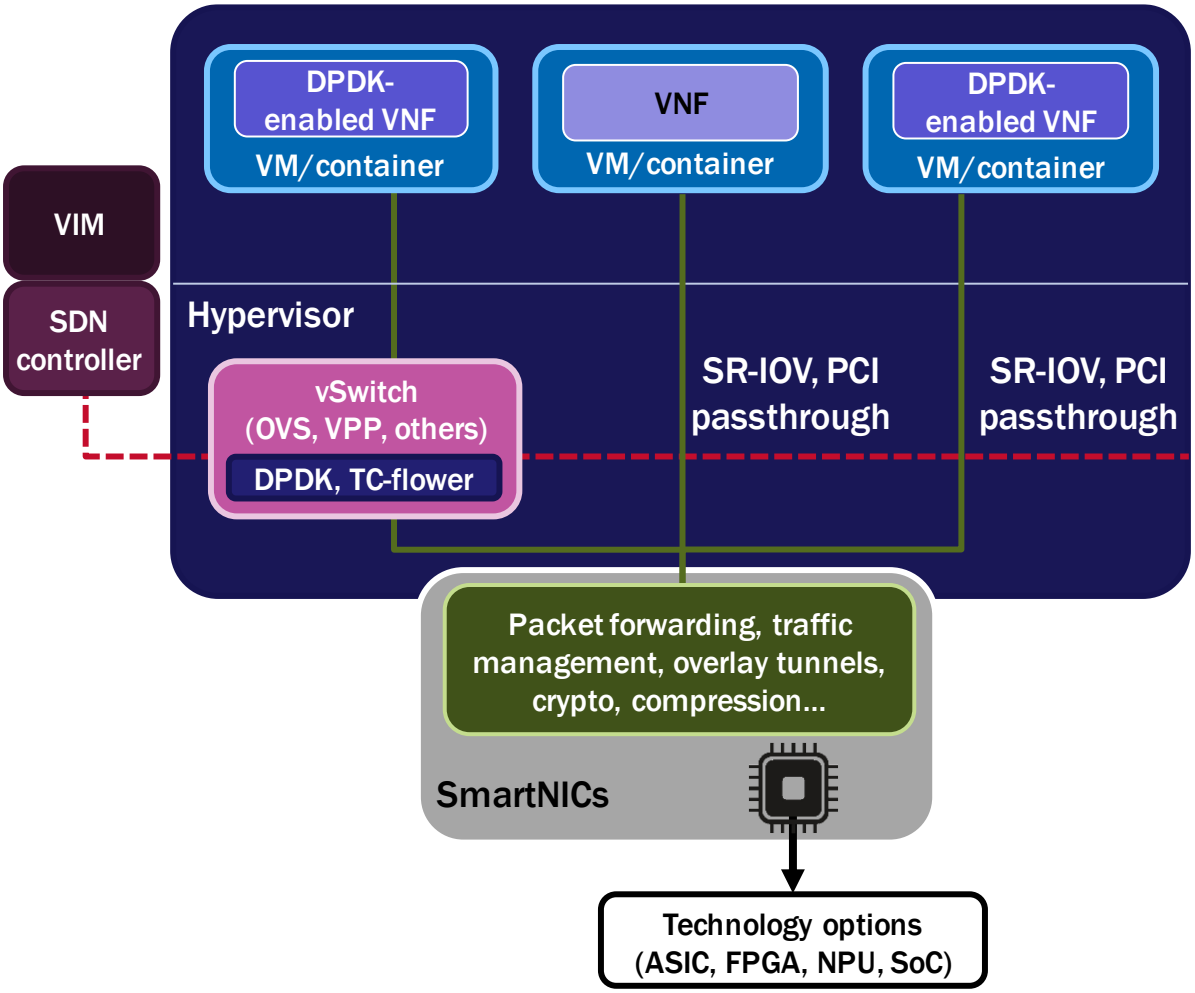
- NFVI networking offload, for example, vSwitch packet processing
- VNFs with complex, compute-intensive network and non-network functionalities (including IPsec, encryption, transcoding and tunnel processing)
- storage (storage virtualization, networking, software-defined storage and hyperconverged infrastructure).

[4] European Telecommunications Standards Institute (December 2015), *ETSI GS NFV-IFA 001: Network Functions Virtualisation (NFV); Acceleration Technologies; Report on Acceleration Technologies & Use Cases*. Available at: https://docbox.etsi.org/ISG/NFV/Open/Publications_pdf/Specs-Reports/NFV-IFA%20001v1.1.1%20-%20GS%20-%20Acceleration%20-%20UCs%20report.pdf.

Many hardware-centric acceleration solutions are available for virtualized networks based on different types of chipset, including ASIC, FPGA, NPU and multicore processors, which may be implemented – for performance and low-power reasons – as Systems-on-a-chip (SoCs). Hardware-centric acceleration solutions are deployed in a wide variety of use cases and scenarios but also often for scenario-specific offloading purposes. GPUs, vision processing units (VPU) and tensor processing units (TPUs) can all give excellent results for specific acceleration cases including video processing, vision systems and AI/machine learning (ML) respectively. These hardware solutions come in different form factors and can reside in various parts of the NFVI.

Smart NICs are increasingly deployed by operators because they provide a natural evolution path from conventional NICs by adding programmable acceleration. SmartNIC solutions can be based on different chipsets and architectures (including ASIC, FPGA, SoC and NPU) and there is a large ecosystem of vendors that provide these solutions including Broadcom, Ethernity Networks, Marvell, Mellanox, Napatech, Netronome and Xilinx. Other options include PCI-linked devices (for example, Intel QAT accelerator, discussed in section 3.3.), network attached devices or integrated architectures (such as Intel's Xeon Gold 6138P processor with integrated Intel Arria 10 FPGA).

Figure 3.4: VNF and NFVI acceleration using SmartNICs [Source: Analysys Mason, 2019]



Hardware-centric accelerators provide a significant boost to VNF and NFVI performance, but they can also add to NFVI costs and operational complexity without standardization and management abstraction. Operators need

to carefully evaluate the cost/benefit analysis of additional hardware acceleration in NFVI and should be prepared to handle the additional management complexity of heterogenous, multi-vendor acceleration devices.

No single technology option for hardware-centric acceleration prevails; there are trade-offs in terms of performance, cost and programmability in each of the various options, as detailed in Figure 3.5.

Figure 3.5: Comparison of hardware acceleration options [Source: Analysys Mason, 2019]

Hardware acceleration	Performance	Price–performance	Programmability and Flexibility
ASIC	●●●● Specifically designed for an application/workload	●●●● Best cost-performance among all options	●○○○ Limited by the initial design capabilities
FPGA	●●●○ High performance but small trade-off with flexibility	●●●○ High performance at high cost; prices are decreasing	●●●○ Can be programmed on the fly, but requires specialist programming skills or vendor support
NPU	●●●○ High performance but small trade-off with flexibility	●●●○	●●○○ Can be programmed but has limited focus on network I/O, solutions are typically vendor-proprietary
SoC	●●○○ High flexibility - performance trade-off	●●●○ Good performance at moderate cost	●●●● General purpose, C language, Linux

Operators' choice of technology will be dependent on use case demands. For example, ASIC-based acceleration is more suitable to scenarios where acceleration requirements do not change often, and maximum cost–performance is desired (for example, vBNG in the fixed network core). FPGA would be ideal for edge deployments where application requirements are dynamic and require on-the-fly programmability (for example, perform DPI during the day and switch to transcoding to support a sport events broadcast in the evening). SoCs can be deployed for security applications such as encryption acceleration or transcoding use cases.

AI/ML, next-generation video and AR/VR services will need acceleration

GPUs, such as those provided by NVIDIA and AMD, are commonly used for multimedia processing workloads such as video encoding/transcoding and compression. New generations of codecs such as H.265, VP9 and AV1 are increasingly being adopted to deliver high-quality video services (4K/8K) and low-latency, high-bandwidth AR/VR services. Virtualized video networks and edge computing services would benefit from GPU acceleration and offload to meet the requirements of these services.

GPUs have become a popular solution for artificial intelligence (AI)/ ML demands in the past 2 years. The parallel processing capabilities of GPUs bring significant improvement over CPUs; GPUs typically comprise thousands of cores, which can perform millions of calculations in parallel as demanded by tasks such as deep learning, neural networks, image/face recognition and natural language processing. Other hardware acceleration options such as FPGAs, ASICs and SoCs, as well as specialised solutions such as Google's TPUs are also increasingly used to accelerate computationally intensive AI/ML workloads.

VPUs, such as Intel Myriad X VPU, are emerging as purpose-built processors for computer vision applications. They process real-world elements and images including AR/VR, robotics and drones. VPUs are specifically

designed for low power and high performance for mobility requirements, which may not be met with existing GPU architectures.

3.3 Acceleration for NFV/SDN security functions: Intel QAT and HPE performance results

NFV/SDN applications such as vCPE, SD-WAN, vADC, vIPsec and vNGFW need to perform compute-intensive security tasks (such as cryptography including encryption/decryption (SSL/TLS), authentication, public key functions) with high throughput and low latency. The performance, QoS and CPU utilization of these applications can be significantly improved by using hardware-centric accelerators, such as Intel QuickAssist Technology (QAT),[5] which can offload the CPU and accelerate and compress cryptographic workloads. Figure 3.6 provides an overview of performance improvements achieved using Intel QAT acceleration in several NFV/SDN use case tests.

Figure 3.6: Intel QAT acceleration performance results on HPE servers [Source: Intel, HPE, 2019]

Use case	VNF	Configuration	Performance improvements
Application delivery control (traffic management (for example, load balancing, DNS), security (such as firewall))	F5 BIG-IP Virtual Edition	VNF with 8 vCPUs on HPE servers with QAT providing SSL/TLS encryption/decryption	Higher SSL performance: • **5x** greater SSL transactions per second (TPS) (TLS1.2 AES128-SHA 2K Key): from 7034 TPS to 34173 TPS • **2x** greater SSL bulk encryption throughput: from 7744 Mbps to 15360 Mbps The performance boost with QAT is achieved at 56% CPU utilization while the baseline is at 99.9%.
vCPE and SD-WAN	Nuage Networks SD-WAN vCPE (7850 Network Services Gateway)	HPE servers (Intel® Xeon® Gold 6152 processor, with 22 Cores (44 threads), 2.1 GHz) with Intel QAT providing acceleration for SD-WAN (Router, firewall, Service Chain) and vBNG (VPN, firewall, DPI) service.	Higher IPsec performance: • **1.5x** greater encryption throughput • **2x** greater CPU capacity for application processing, VNF Density and value-added services on the same per server node.

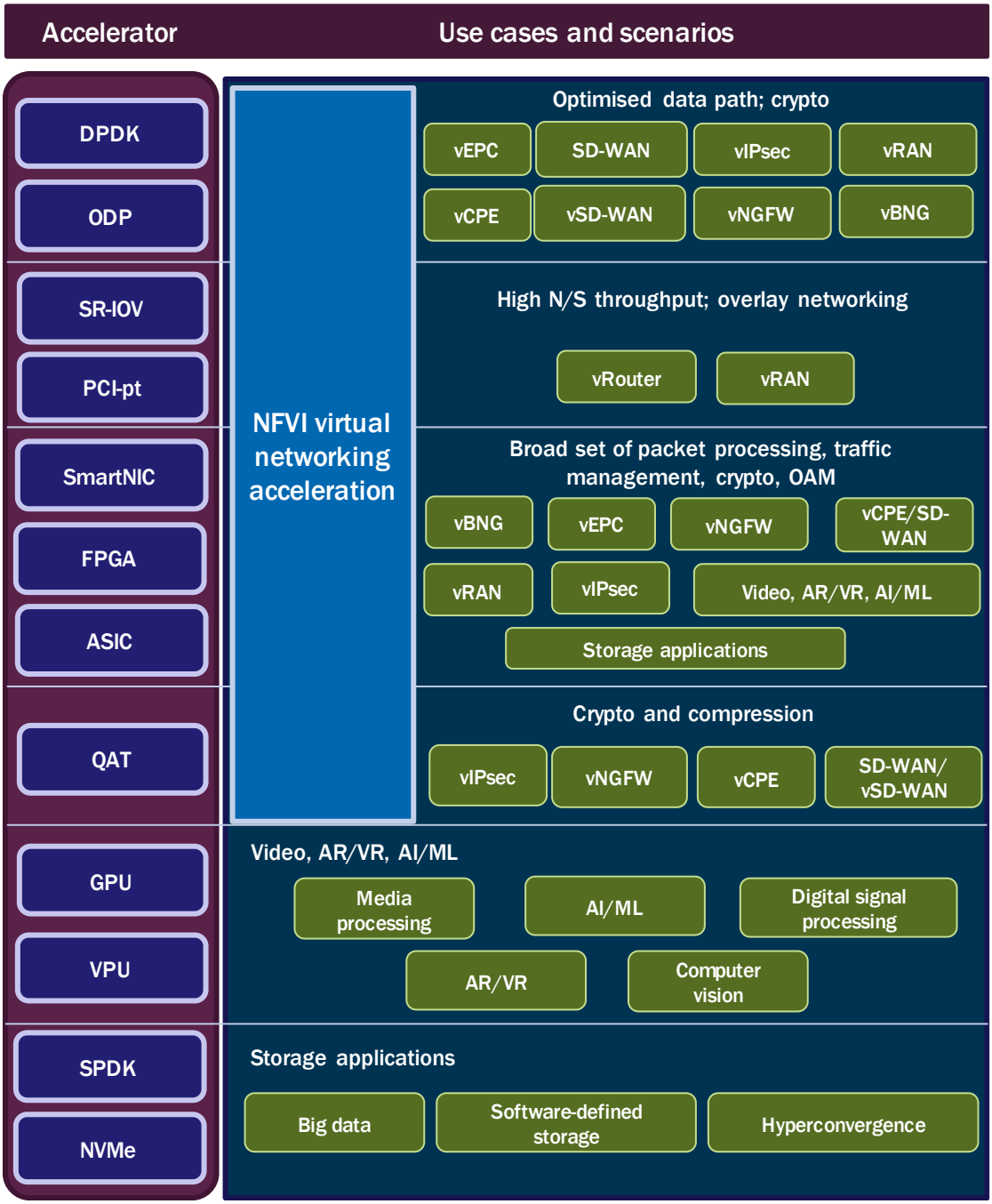
3.4 Summary comparison of acceleration technologies

Operators implementing NFVIs that will support multiple VNFs and digital services, both today and in future, will need to understand the range of available acceleration technologies and their applicability to individual NFV and SDN use cases. Figure 3.6 summarizes the different types of acceleration technology and the use cases for which they are suitable. It should be noted that this is not an exhaustive list of use cases/scenarios for each of

[5] Intel QAT can take various forms, for example, integrated into CPU (Intel Xeon), in a System on Chip package (Intel Atom SoC), or as PCI linked accelerator cards (Intel QuickAssist Adapter).

the acceleration options. In addition, these are not mutually exclusive (multiple options can be combined for a specific use case).

Figure 3.7: Summary comparison of acceleration technologies (Source: Analysys Mason, 2019]



4. Acceleration-as-a-service is crucial to abstract the complexity of acceleration resources

As the number and type of acceleration technologies needed in an NFVI grows, operators face a substantial challenge in managing both the technologies themselves and VNF access to them. As explained previously, acceleration technologies come in different forms (software, hardware, hybrid), are implemented in different devices (smartNICs, crypto cards) in different components of the NFVI (PCI linked, integrated with the CPU, or even remote) and are supplied by a broad set of vendors. The requirement for a heterogeneous NFVI that is able to support multiple use cases and scenarios with high-performance demands is spurring a strong, parallel need for an acceleration technology abstraction layer. The NFVI would use such a layer to manage a complex set of acceleration resources and expose their capabilities through a common API in an 'as-a-service' model to virtualized networks.

Such an acceleration abstraction layer should:

- ensure the complete independence of VNFs from the NFVI. It should provide a common interface across all acceleration technologies, regardless of type or vendor, to prevent vendor lock-in and eliminate the need to re-write/modify VNFs to work with a specific acceleration technology.
- streamline and automate the lifecycle management of acceleration resources (identify/discover, provision, configuration, maintenance and monitoring), and acceleration scenarios, including assigning the right acceleration resources dynamically to adapt to workloads/use cases in edge infrastructure.

The industry is making significant efforts to achieve these goals. For example, ETSI NFV ISG defines an acceleration abstraction layer (AAL), which enables the decoupling of VNFs from the underlying hardware accelerators and provides abstraction interfaces for VNFs and supports for the management of these acceleration resources in the VIM.[6] Open-source initiatives are also following this course. OPNFV has the Data Plane Acceleration Project (DPACC), which aims to provide a common suite of abstraction APIs to enable VNF portability and resource management for integrated SoC accelerators. OpenStack has responded to industry demands with its Cyborg project. OpenStack Cyborg[7] is an emerging general-purpose management framework for the automated lifecycle management of software and hardware acceleration resources (including FPGA, GPU, crypto cards, DPDK, ODP and storage). Currently, management of these resources can be complex and time-consuming as each may require specific tuning and configuration. Network engineers need to develop their own Python scripts or Ansible playbooks to carry out these processes. OpenStack Cyborg aims to simplify these processes with a set of software components, RESTful APIs and generic drivers, which enable dynamic resource discovery, attachment/detachment of acceleration devices and driver management, and work in conjunction with OpenStack Nova or standalone in bare metal (for example, OpenStack Ironic).

Acceleration-as-a-service is a key step for the movement towards complete hardware and software independence in virtualized networks. Building open, unified management platforms with common, standardised interfaces will be critical and it will require close collaboration between operators, vendors and industry groups to achieve it.

[6] ETSI GS NFV-IFA 001, ETSI GS NFV-IFA 002, ETSI GS NFV-IFA 004.

[7] For more information, see https://wiki.openstack.org/wiki/Cyborg.

5. Conclusion and recommendations

Operator networks are transforming into software-defined, automated, cloud-native NFV infrastructures enhanced by advanced analytics, AI/ML techniques to manage the rapid growth in traffic and bandwidth demands and to achieve dynamic, on-demand delivery of network services. 5G and edge computing will usher in the new era of cloud-based services such as AR/VR, robotics and many other URLLC applications that will be developed and delivered through these new infrastructures. Acceleration technologies will be pivotal to this transformation by enabling operators to meet the performance, latency, QoS, subscriber density and security requirements of existing and future applications with optimum TCO.

As presented in this paper, many network virtualization, AI/ML, 5G and IoT use cases will require operators to take advantage of acceleration technologies. There are many software- and hardware-centric acceleration solutions available for these use cases, which can be deployed in different parts of the NFVI in different form factors and can be supplied by different vendors or through open-source initiatives. Operators will need to identify and deploy the most-suitable acceleration and offloading technologies for their application/workload needs, networks (including fixed, wireless, core and access), deployment scenarios and operational capabilities. Moreover, operators need to carefully evaluate the cost, performance, programmability and flexibility of the acceleration solutions because there is no one-size-fits-all approach as they often represent a trade-off. For example, fixed-function (ASIC) accelerators are preferable in scenarios where acceleration requirements are constant and maximum price–performance is desired (for example, 24/7 active security workloads), while FPGA or SoC-based accelerators provide lower price-performance than ASIC but they are more programmable and can adapt to dynamic workloads (for example, in edge infrastructure).

The performance and cost benefits of the acceleration technologies should not come at the expense of hardware and software independence in virtualized networks. The sheer multitude of use cases and technologies, as well as the current lack of standardization across these technologies and their suppliers, will inevitably increase the heterogeneity in NFVI, which can lead to additional operational complexity and negate the automation and agility benefits of virtualization. To mitigate these risks, operators need an acceleration technology abstraction layer on NFVI to decouple VNFs from underlying acceleration resources in order to independently accommodate all use case needs of operators and to automate the lifecycle management of acceleration resources and scenarios with an ‘as-a-service’ model. Strong industry collaboration to set a common framework for acceleration standardization and support for open-source initiatives such as OpenStack Cyborg and OPNFV DPACC will be crucial in order to achieve operators’ goal to create horizontal and highly automated NFVIs.

6. About the authors

Gorkem Yigit (Senior Analyst) is the lead analyst for the Video and Identity Platforms programme and a contributor to the Digital Infrastructure Strategies and Network Automation and Orchestration programmes, focusing on producing market share, forecast and research collateral. He has published research on NFV/SDN services business cases, identity management in the digital economy, and has been a key part of major consulting projects including Telco Cloud Index and IPTV/OTT procurement. He holds a cum laude MSc degree in Economics and Management of Innovation and Technology from Bocconi University (Milan, Italy).

Caroline Chappell (Research Director) is the lead analyst for Analysys Mason's Digital Infrastructure Strategies research programme. Her research focuses on service provider adoption of cloud, and the application of cloud technologies to fixed and mobile networks. She is a leading exponent of SDN and NFV and the potential that these technologies have to enhance business agility and enable new revenue opportunities for service providers. Caroline investigates key cloud and network virtualization challenges and helps telecoms customers to devise strategies that mitigate the disruptive effects of cloud and support a smooth transition to the era of software-controlled networks.

This white paper was commissioned by HPE and Intel. Analysys Mason does not endorse any of the vendor's products or services.

Published by Analysys Mason Limited • Bush House • North West Wing • Aldwych • London • WC2B 4PJ • UK
Tel: +44 (0)20 7395 9000 • Email: research@analysysmason.com • www.analysysmason.com/research

Registered in England No. 5177472

7. Analysys Mason's consulting and research are uniquely positioned

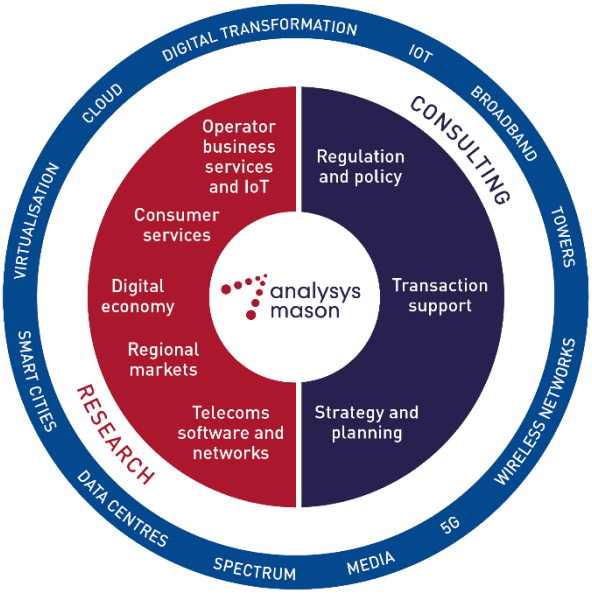
Analysys Mason is a trusted adviser on telecoms, technology and media. We work with our clients, including communications service providers (CSPs), regulators and end users to:

- design winning strategies that deliver measurable results
- make informed decisions based on market intelligence and analytical rigour
- develop innovative propositions to gain competitive advantage.

We have around 220 staff in 14 offices and are respected worldwide for the exceptional quality of our work, as well as our independence and flexibility in responding to client needs. For over 30 years, we have been helping clients in more than 110 countries to maximise their opportunities.

Consulting

- We deliver tangible benefits to clients across the telecoms industry:
  - communications and digital service providers, vendors, financial and strategic investors, private equity and infrastructure funds, governments, regulators, broadcasters, and service and content providers.
- Our sector specialists understand the distinct local challenges facing clients, in addition to the wider effects of global forces.
- We are future-focused and help clients understand the challenges and opportunities that new technology brings.



Research

Our dedicated team of analysts track and forecast the different services accessed by consumers and enterprises.

We offer detailed insight into the software, infrastructure and technology delivering those services.

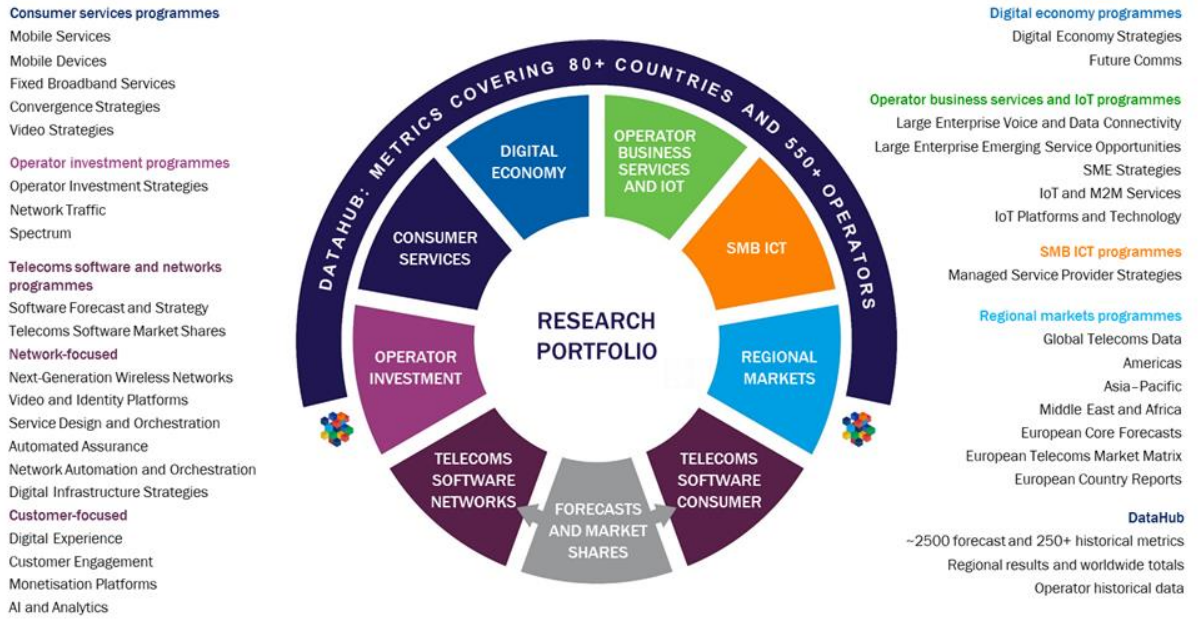
Clients benefit from regular and timely intelligence, and direct access to analysts.

8. Research from Analysys Mason

We provide dedicated coverage of developments in the telecoms, media and technology (TMT) sectors, through a range of research programmes that focus on different services and regions of the world.

The division consists of a specialised team of analysts, who provide dedicated coverage of TMT issues and trends. Our experts understand not only the complexities of the TMT sectors, but the unique challenges of companies, regulators and other stakeholders operating in such a dynamic industry.

Our subscription research programmes cover the following key areas.



Each subscription programme provides a combination of quantitative deliverables, including access to more than 3 million consumer and industry data points, as well as research articles and reports on emerging trends drawn from our library of research and consulting work.

Our custom research service offers in-depth, tailored analysis that addresses specific issues to meet your exact requirements.

Alongside our standardised suite of research programmes, Analysys Mason's Custom Research team undertakes specialised, bespoke research projects for clients. The dedicated team offers tailored investigations and answers complex questions on markets, competitors and services with customised industry intelligence and insights.

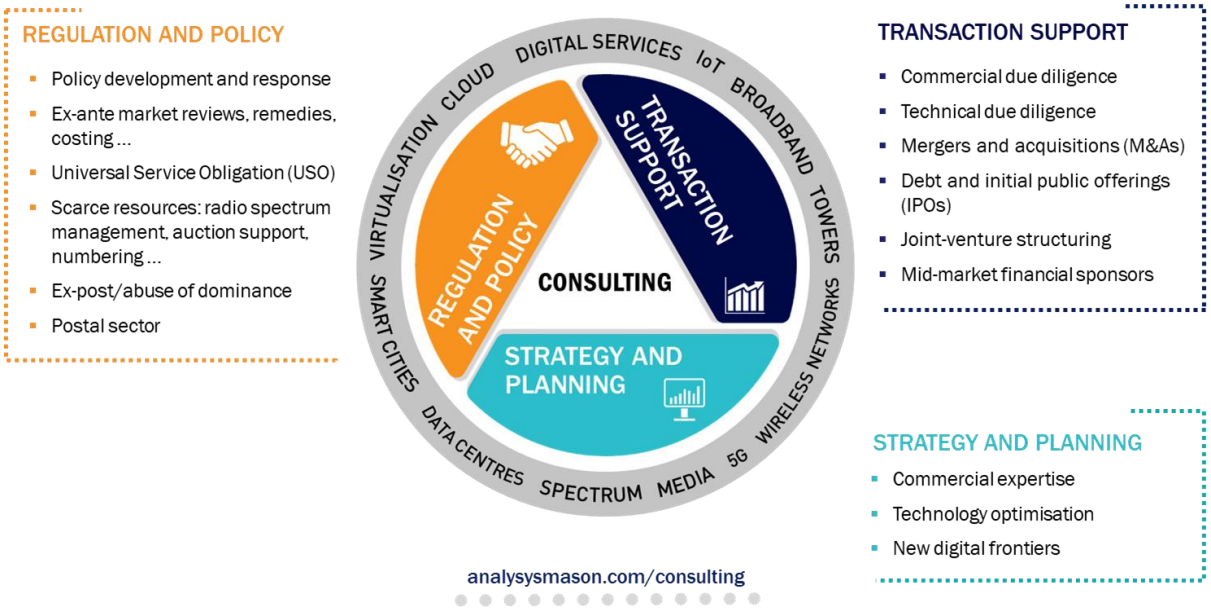
For more information about our research services, please visit www.analysysmason.com/research.

9. Consulting from Analysys Mason

For more than 30 years, our consultants have been bringing the benefits of applied intelligence to enable clients around the world to make the most of their opportunities.

Our clients in the telecoms, media and technology (TMT) sectors operate in dynamic markets where change is constant. We help shape their understanding of the future so they can thrive in these demanding conditions. To do that, we have developed rigorous methodologies that deliver real results for clients around the world.

Our focus is exclusively on TMT. We advise clients on regulatory matters, help shape spectrum policy and develop spectrum strategy, support multi-billion dollar investments, advise on operational performance and develop new business strategies. Such projects result in a depth of knowledge and a range of expertise that sets us apart.



We look beyond the obvious to understand a situation from a client's perspective. Most importantly, we never forget that the point of consultancy is to provide appropriate and practical solutions. We help clients solve their most pressing problems, enabling them to go farther, faster and achieve their commercial objectives.

For more information about our consulting services, please visit www.analysysmason.com/consulting.