White Paper

# Project Monterey: Evolving Infrastructure to Accelerate and Secure All Your Workloads

Sponsored by: VMware

Ashish Nadkarni
March 2022

## IDC OPINION

As organizations pursue a holistic digital strategy for people, processes, technology, data, and governance, IDC estimates that global spending on digital transformation (DX) of business practices, products, and services to reach $2.8 trillion in 2025, more than double the amount allocated in 2020. Organizations allocate their DX investments toward two key groups of strategic objectives that align with what they expect to accomplish over an extended period in pursuit of their digital mission:

- Operational objectives, including back-office support and core business functions such as accounting and finance, human resources, legal, security and risk, and enterprise IT (These are implemented via current generation workloads for revenue operations.)
- Innovate and scale, which include the development of new products and services and the delivery of new customer experiences at unprecedented scale (These are implemented via next-generation workloads to deliver differentiation for future enterprise.)

A modern IT infrastructure — built using the appropriate computing platforms and systems — provides a singular foundation for the seamless execution of operational as well as innovate and scale objectives. When designed appropriately, it can deliver the agility and elasticity for compressed and secure time to insights from vast amounts of data sets that are necessary for a complete digital experience.

We are on the cusp of a major systems architecture overhaul, thanks to the advent of data processing units (DPUs) and work done by vendors like VMware to build an ecosystem around Project Monterey, a distributed infrastructure solution. [Note that DPUs are also called SmartNICs. IDC uses the term *function-offload accelerators (FAs)*]. This new architecture — which has already been adopted at scale by some of the largest service providers — augments the current computing platform architecture and lays the groundwork for infrastructure services offload, full software composability, and eventual hardware disaggregation. It is designed to support an era of hyperscale, multicloud, shared-everything, and zero trust computing.

In the foreseeable future, a collection of DPUs, enabled by VMware's Project Monterey, running in servers could create a unified datacenter backplane. This approach could offer a consistent software-defined but hardware-controlled security and monitoring network across the entire datacenter for configuring, deploying, and managing bare metal, virtualized, and containerized workloads. It could provide a consistent and simple but no-compromises operations environment that limits the ability for "nonoperator-approved entities" (humans or applications) to access the control environment and thus limit the impact of exploits on low-level hardware vulnerabilities.

## SITUATION OVERVIEW

Digital transformation changes the way businesses view IT infrastructure. It shifts from being a necessary overhead to a strategic investment area. As businesses overhaul their business practices, products, and services to maintain their differentiation now and into the future, their pact with IT infrastructure is fundamentally altered.

Organizations that are digitally transforming themselves are also modernizing their IT infrastructure, which — using the appropriate mix of computing platforms and systems — is designed to provide a singular foundation for the seamless execution of operational as well as innovate and scale objectives. It is designed to support a new class of DX workloads (called next-generation or cloud-native workloads) in addition to revenue-generating operations workloads (called current generation workloads). A modern infrastructure must:

- Deliver a seamless experience to internal and external stakeholders and customers by tying together dedicated (private) and shared (public) resources.
- Deliver the scale necessary for timely insights from vast amounts of data sets that are necessary for a complete digital experience.
- Enable agility and elasticity necessary for all workloads to run in a variety of cloud environments (multicloud) with consistent infrastructure services and deployment options.

### Current Infrastructure Architecture

In the current datacenter architecture, the central processing unit much like the human brain maintains centralized control of all privileged operations. It controls most — if not all — of the hardware functionality and software-defined infrastructure services delivered by the computing platform itself.

This computing platform architecture has been the de facto model for all infrastructure elements, regardless of their size and usage, deployed in datacenter or edge deployment or consumer handhelds. This model was designed for simplicity and vertical scaling for monolithic workloads. No matter what the use of the computing platform — as a general-purpose server, a security or network appliance, or a storage system controller — its architecture remains the same.

There are opportunities to enhance this architecture using software defined but hardware assisted approaches with the ultimate objective of gaining consistent service quality in multicloud environments.

#### *Computing Platform Payloads*

In a traditional infrastructure environment, the computing platform executes all three payloads within the same processing unit. These are:

- Embedded payloads that are part of the core operating system and often run as privileged operations in kernel space
- User space payloads (i.e., programmatically assigned portions of workloads that comprises an application and its accompanying data set)
- Management payloads that straddle between embedded and user space payloads and are gaining importance in cloud-native and scale-out deployments

Due to the nuances and architectural elements of a traditional operating system, all these payloads are executed in a time-sliced and dynamically prioritized environment. Any imbalance in the execution environment can lead to service quality issues at scale.

### Datacenter Scaling

At the same time, modern datacenter workloads are pushing the design envelope for network-based scaling. Scale-out and distributed workloads significantly increase the number of CPU cycles spent on sending and receiving workload and cluster-related network traffic. As more of these workloads are deployed, the efficiency of current infrastructure continues to decrease, leave alone the fact that it is vulnerable to security incidences. Datacenter operators at IT organizations and service providers require a new infrastructure architecture that can scale as their stakeholders and customers bring new workloads on board.

### Platform Security

Any system designed to deliver consistent service quality in the datacenter must provide intrinsic security. The integrity of applications and data hosted in a datacenter or in the cloud depends on the computing architecture of a server, storage, and networking infrastructure. The more centrally placed this system (e.g., a server in a public cloud or a server used in a virtualization cluster), the more crucial this situation as it can impact several businesses or tenants simultaneously.

Intrinsic platform security must therefore be as important to the infrastructure as other operational service-level objectives such as reliability, availability, and serviceability. Any solution to bolster the security posture of current architecture deployed in most servers and storage and networking systems today must be linearly scalable at the very minimum.

Further, the entire hardware stack is tightly coupled with the operating system environment to deliver a continuously and consistently enforceable security paradigm across the entire infrastructure. This limits options with hardware life-cycle management. The all-or-nothing process is rigid and expensive, requiring the entire server or cluster of servers hosting the workload — with all its components — to be replaced at the same time.

### Requirements for a New Datacenter Architecture

Any approaches to deliver consistent service quality in a multicloud environment must overcome the two challenges mentioned above: computing platform payload and datacenter scaling. Newer approaches must support expanded enterprise perimeter, with zero trust computing as the foundation. They must deliver:

- Infrastructure services (e.g., data, orchestration, security, and network services), regardless of whether the workload runs on bare metal or is virtualized or containerized and across multiple deployment locations (multicloud)
- Infrastructure as code to workloads that are designed to consume infrastructure resources on demand via automated API calls and with limited operator intervention
- Apportioned heterogeneous computing resources such as workload accelerators (e.g., GPUs) and persistent (storage-class) memory, especially to containerized workloads

There are opportunities for infrastructure stack providers to enhance the capabilities of the current CPU-centric architecture:

- A portion of key infrastructure services required to deliver a rich multicloud experience could be offloaded, freeing up processor and memory cycles required to deliver high service quality to the application workloads.

- Build in hardware-based mitigation capabilities to minimize the impact of any low-level security vulnerabilities in the central execution environment.
- Create a consistent operating model across multicloud environments, addressing service quality issues that are difficult to troubleshoot and isolate.
- Build in a hardware-based separation or isolation of security and supervisory functions making it harder for network-based attacks to compromise the operating system or workload execution environment.

## A New Approach for Infrastructure Architecture for the Multicloud Era

Platform architecture is as important to the infrastructure as other operational service-level objectives such as reliability, availability, and serviceability. Finally, the entire hardware stack should be tightly coupled with the operating system or hypervisor environment to deliver a continuously and consistently enforceable security paradigm across the entire infrastructure.

Decentralized or disaggregated architectures enabled by data processing units (aka function-offload accelerators) borrow an approach (known as accelerated computing) that is gaining traction in the industry. Specially designed processors (also known as accelerators, of which coprocessors are a variant) are used to offload specific portions of a user space payload (e.g., GPUs for mathematically intensive functions) to accelerate outcomes of the workload. However, even in most accelerated computing deployments, the CPU is still in command of the platform and runs all the embedded and management payloads. On the other hand, DPUs function as an autonomous coprocessor for the central processing unit, taking over many of the core functions that are performed by the CPU (see the Benefits of DPUs and DPU-Based Platforms sidebar).

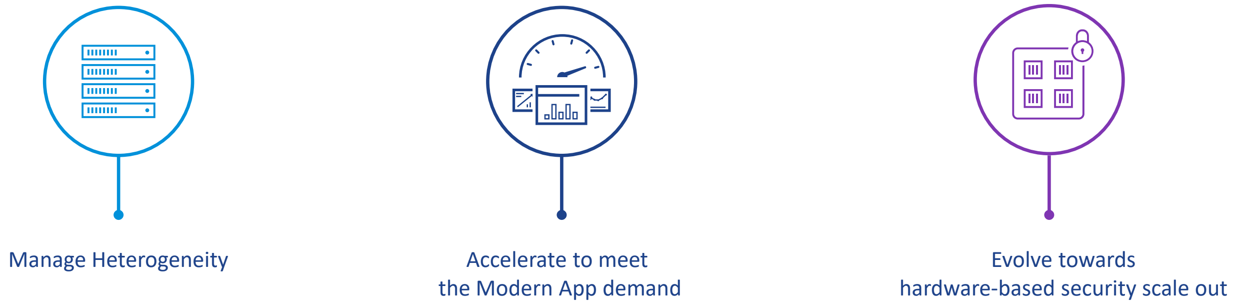
In a system that makes use of these accelerators, the central processing unit is basically dedicated to run application workloads. This provides three immediate benefits to a modern datacenter architecture in which multicloud infrastructure services are as critical as the workloads that run in the environment:

- **Reduced CPU and memory overhead.** Core infrastructure services no longer need to fight alongside application workloads for shared processor and memory resources. Operators have the option to run them on a separate processor subsystem, with an independent network and control plane.
- **An out-of-band control plane.** The presence of a transparent software defined I/O layer that runs on the coprocessor provides an additional layer of access control for physical and/or shared and networked resources (e.g., storage resources from a shared storage controller).
- **Zero Trust Computing.** By offloading infrastructure services to the DPU, the operating system or hypervisor environment gains an additional layer of immunity from rogue code that targets low-level processor vulnerabilities or tries to bypass the checks present in the central software and hardware stack.

Figure 1 illustrates the key tenets of an approach that enhances the current computing platform architecture. This approach provides a consistent management framework, zero trust computing foundation and the necessary scale required for a multicloud era.

## FIGURE 1

### Key Principles of Data Processing Unit-Based Modern Datacenter Architecture



Manage Heterogeneity

Accelerate to meet
the Modern App demand

Evolve towards
hardware-based security scale out

Source: VMware, 2021

### Benefits of DPUs and DPU-Based Platforms

Data processing units (or function-offload accelerators) are a special kind of processors usually installed in servers on a PCIe card or directly on a motherboard (see Figure 2). They are designed to:

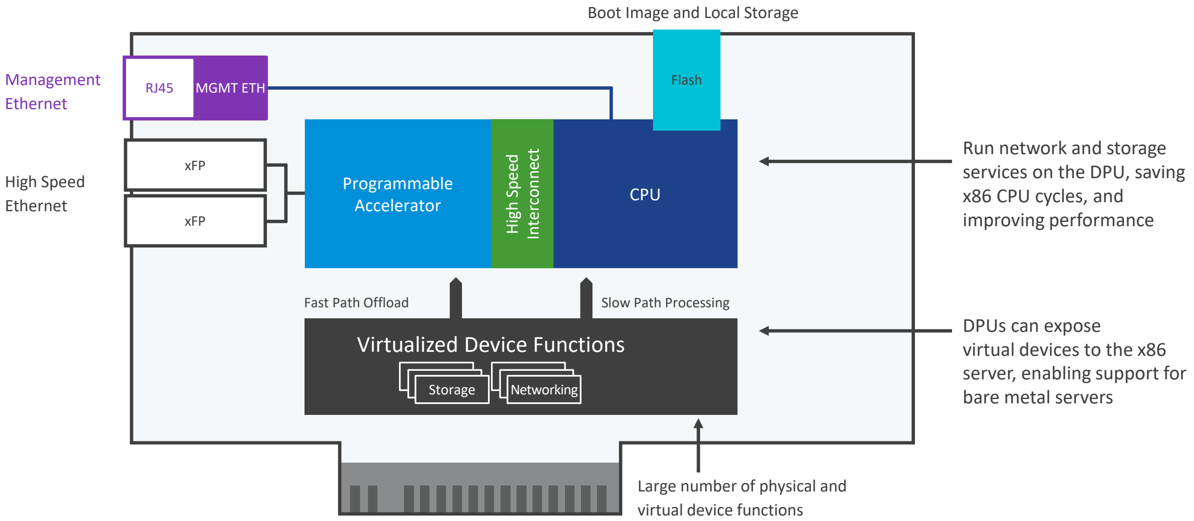
- Operate independent of the CPU and outside the control of the main processing system from a code execution perspective. In other words, the CPU is aware of their presence but cannot control them.
- Introduce an additional abstraction layer in the bootup and operational state of the platform. They boot up using their own independent microcode, firmware, or a lightweight hypervisor that treats the CPU subsystem as one big "virtual machine" running in a reduced privilege mode.
- Control access to physical resources like data persistence and network interfaces through which sensitive data can be accessed. Any payload executed on the CPU including the kernel itself that needs access to these resources must go through function-offload interfaces, which is presented in a software-defined manner to the operating system environment running on the CPU.
- Take over direct execution of crucial portions of the embedded and management payload functions such as security, hypervisor root partition, and virtual networking. Such kernel space functions are run and/or require the process to run in a privileged mode on the host operating system.

Accordingly, FA-based computing platforms offer several advantages over traditional (monolithic) platforms. These include:

- **Tamper-proof architecture.** Executing embedded and management components like root of trust, encryption, data, and network access on the same processor subsystem introduces inherent limits to the extent to which each payload code can be isolated from accessing or tampering with other components executing on the processor. On the other hand, executing management components in a separate physical device introduces physical isolation between the management and payload components.
- **Uniform resource utilization.** Offloading the embedded and management components frees up processing clock cycles for user space workloads. Even with very capable CPUs and platforms built with two or more CPUs per platform, it is not uncommon for users to experience symptoms due to resource constraints — much of which manifest themselves in the form of slow network or data access. By offloading the compute-intensive control functions for the physical interfaces from the CPU, the core user space payload has more room to breathe.
- **Ability to enable offloading of additional functions and payloads.** Finally, a key benefit of using DPUs is to start to offload management functions, thus creating a management fabric that transcends physical, virtual, and containerized compute and can be handled in a CPU-agnostic manner, regardless of the operational state of the platform itself. DPUs that are autonomous in nature (i.e., can be bootstrapped using a unique operating system or hypervisor instance) are especially attractive as they can create their own function fabric that decouples privileged execution from the operating system or hypervisor instance running on the CPU. FA-based platforms are especially valuable in shared infrastructure environments (such as cloud environments) that host a heterogenous mix of traditional and cloud-native workloads belonging to a variety of workload owners (tenants). In such environments, a high volume of intra-datacenter (or even intra-server) traffic may be expected.

## FIGURE 2

### A Data Processing Unit Card


Boot Image and Local Storage
Management Ethernet
RJ45
MGMT ETH
Flash
High Speed Ethernet
xFP
xFP
Programmable Accelerator
High Speed Interconnect
CPU
Run network and storage services on the DPU, saving x86 CPU cycles, and improving performance
Fast Path Offload
Slow Path Processing
Virtualized Device Functions
Storage
Networking
DPUs can expose virtual devices to the x86 server, enabling support for bare metal servers
Large number of physical and virtual device functions

Source: VMware, 2021

## Composable Infrastructure for Multicloud Environments

Hardware-based disaggregated or decentralized approaches discussed previously can only be successful in deployments wherein the infrastructure software stack (aka the operating environment) is tightly coupled to the hardware stack. The resulting composable infrastructure system must be designed such that the software stack complements and benefits from the specific capabilities offered by the DPUs that enable the hardware disaggregation. This is because the capabilities of DPUs vary from vendor to vendor and product to product. Some are designed with networking and security in mind, others to offload data services, and some offer management offload. Whatever the functionality, the use of DPU enables the systems vendor to:

- Offer a consistent operating environment for bare metal, virtualized, and containerized workloads.
- Provide a consistent way to deploy and consume heterogeneous (and accelerated) computing across multicloud.
- Provide hardware-based security and access control, with a full software-defined conversion/offload for hardware access.
- Offer a system for modern application delivery with a management layer that is independent of the CPU subsystem.

Composable infrastructure systems build on the many benefits of hyperconverged infrastructure but add the ability to manage bare metal workloads, deploy a distributed hardware-based security, and expand the multicloud experience.

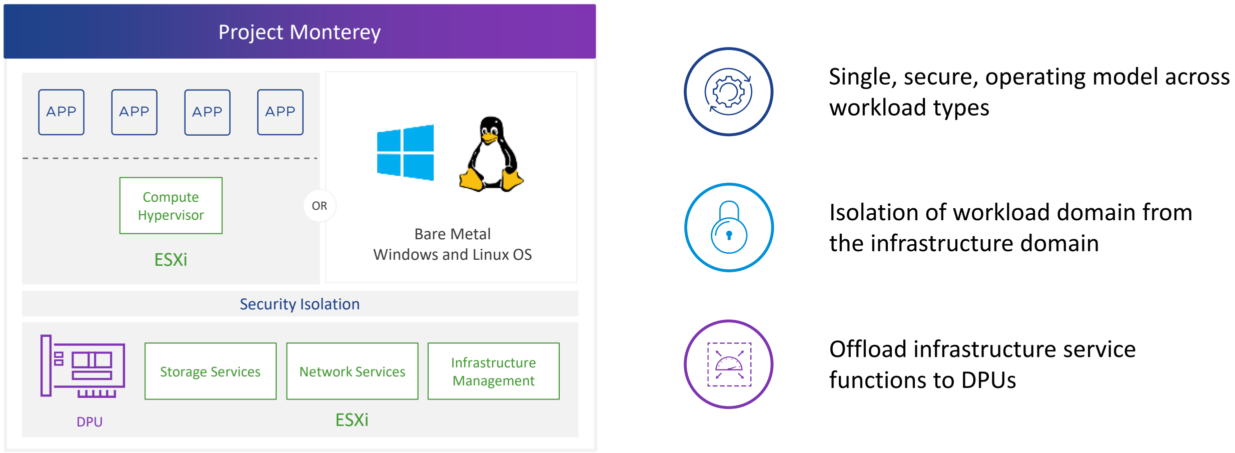
## VMWARE'S PROJECT MONTEREY

VMware’s Project Monterey is an evolutionary architectural approach for the datacenter, cloud, and edge to address the changing requirements of workloads, now and in the future. As illustrated in Figure 3, Project Monterey will usher in a new era of infrastructure management with:

- A single, secure operating model across workload types, regardless of whether they are bare metal, virtualized, or containerized
- The ability to isolate application workload domain from the infrastructure services domain communicating with each other via a software-defined connection
- Infrastructure services that can be offloaded and run on an independent control plane comprised of DPUs

## FIGURE 3

### Project Monterey


Project Monterey
APP
APP
APP
APP
Compute Hypervisor
ESXi
OR
Bare Metal
Windows and Linux OS
Security Isolation
DPU
Storage Services
Network Services
Infrastructure Management
ESXi
Single, secure, operating model across workload types
Isolation of workload domain from the infrastructure domain
Offload infrastructure service functions to DPUs

Source: VMware, 2021

Project Monterey comprises three key elements:

- **Hardware-based acceleration:** VMware Cloud Foundation will be able to maintain compute virtualization on the server central processing unit while offloading networking, security, and storage functions to a specially designed coprocessor on a variety of DPUs. This will allow applications to maximize the use of the available network bandwidth while saving server CPU cycles for top application performance.
- **Composable infrastructure:** Project Monterey will extend management support for bare metal servers. This will enable physical resources to be dynamically accessed based on policy or via software API, tailored to the needs of the application. In addition, organizations will be able to use a single management framework to manage all their compute infrastructure, whether virtualized or bare metal. The decoupling of networking, storage, and security functions from the main server allows these functions to be patched and upgraded independently from the server.

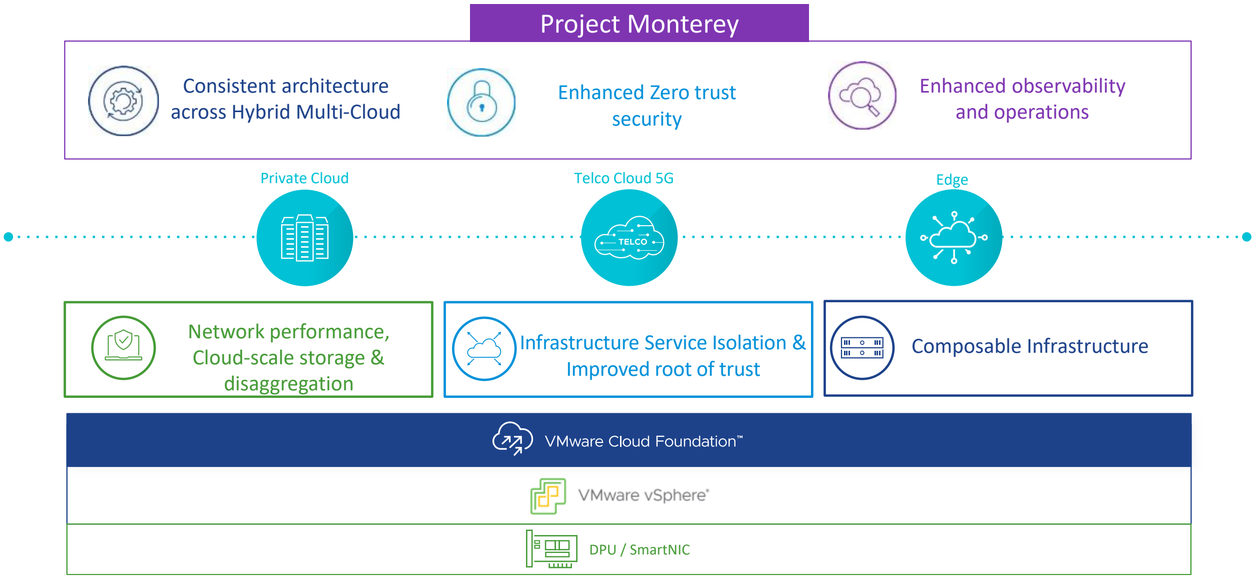
- **Isolation of infrastructure services:** With Project Monterey, VMware can implement intrinsic security. Each DPU can run a full-featured stateful firewall and advanced security suite. Since this will run on a card and not on the CPU, several network-based security services can be deployed and automatically tuned to protect specific application services in a transparent manner. Wrapping each service with intelligent defenses can shield any vulnerability of that specific service. This will enable a custom-built defense that can be automatically tuned and deployed across tens of thousands of application services. In addition, Project Monterey will enable enterprises or service providers supporting multiple tenants to isolate them from the core infrastructure.

As illustrated in Figure 4, with these changes, Project Monterey adds a new way in which VMware Cloud Foundation (VMware vSphere, VMware vSAN, and VMware NSX) can make use of additional DPU (function-offload accelerators or coprocessors) cards in a server.

VMware announced that it is a taking an ecosystem approach with Project Monterey, which means that it will support DPUs from several providers such as Intel, NVIDIA, and Pensando Systems inserted into or integrated with server solutions from OEM vendors like Dell Technologies, Hewlett Packard Enterprise, and Lenovo. As the program matures, VMware will onboard more partners and increase the diversity of their ecosystem.

## FIGURE 4

### Project Monterey for Multicloud Environments


Project Monterey
Consistent architecture across Hybrid Multi-Cloud
Enhanced Zero trust security
Enhanced observability and operations
Private Cloud
Telco Cloud 5G
TELCO
Edge
Network performance, Cloud-scale storage & disaggregation
Infrastructure Service Isolation & Improved root of trust
Composable Infrastructure
VMware Cloud Foundation™
VMware vSphere®
DPU / SmartNIC

Source: VMware, 2021

## Business Value of Project Monterey

IT Organizations stand to benefit from VMware Cloud Foundation (enabled by Project Monterey) as the de facto operating stack for multicloud environments. Key benefits include:

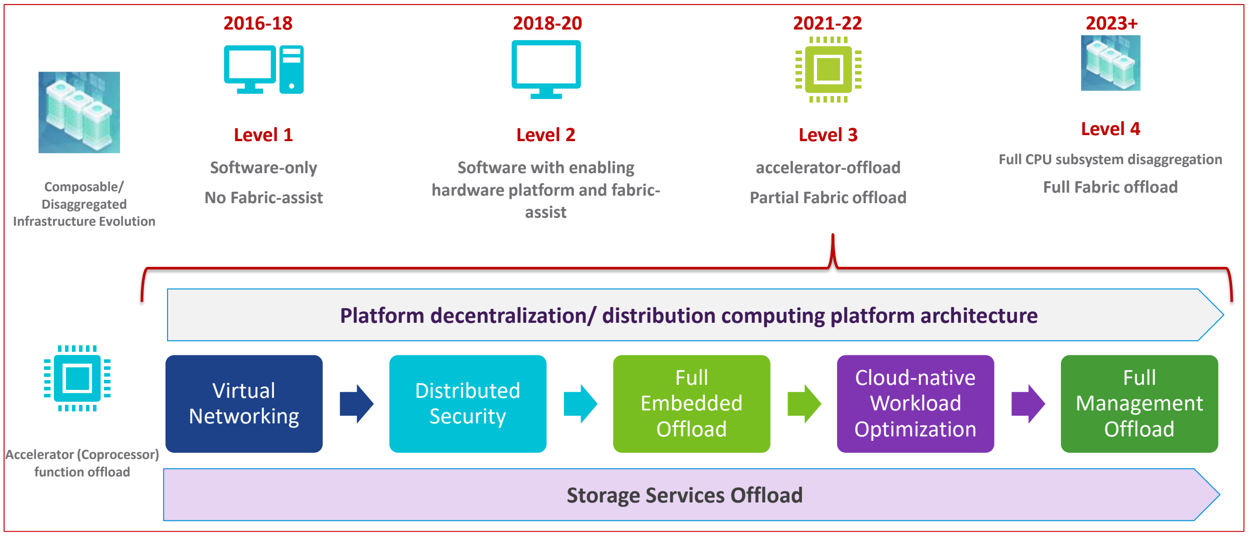
- Flexible and programmatic consumption of compute, storage, and networking resources, leading to improved and consistent resource utilization
- Reduced cost of building and delivering new services, with infrastructure that is configured optimally and utilized fully
- Flexible deployment options ranging from standard general-purpose servers to specialized modular options, depending on the workloads
- Reduced provisioning times with ease of management driven by adaptability to on-demand workload capacity
- A shift toward secure, application-centric infrastructure accelerating rapid development and deployment

## FUTURE OUTLOOK

Use cases for data processing units (aka function-offload accelerators) are still up and coming, but the platform architecture does hold out promise in the long run, as enterprises and service providers discard their inhibitions and eventually warm up to a modular and distributed computing architecture. Figure 5 illustrates IDC's view of how it anticipates use cases to evolve with DPUs.

## FIGURE 5

### Evolution of Use Cases for Data Processing Units (Function-Offload Accelerators)


2016-18
2018-20
2021-22
2023+
Composable/ Disaggregated Infrastructure Evolution
Level 1
Software-only
No Fabric-assist
Level 2
Software with enabling hardware platform and fabric-assist
Level 3
accelerator-offload
Partial Fabric offload
Level 4
Full CPU subsystem disaggregation
Full Fabric offload
Platform decentralization/ distribution computing platform architecture
Accelerator (Coprocessor) function offload
Virtual Networking
Distributed Security
Full Embedded Offload
Cloud-native Workload Optimization
Full Management Offload
Storage Services Offload

Source: VMware, 2021

### Stage 1: Virtual Networking

In this stage, networking functions are offloaded. This includes compute-intensive networking operations like authentication, routing, virtualization, firewalling, and load balancing. The net effect of these offloads is a distributed host-to-network virtual networking paradigm with low CPU overhead. A lot of DPUs that offer Stage 1 functions are also called "SmartNICs" or "data processing units."

### Stage 2: Distributed Security

In this stage, security functions such as isolation, encryption, and root of trust are offloaded. A physically isolated root of trust enables the execution of embedded functions in a tamperproof manner. It also starts to lay the groundwork for the delivery of private cloud as a service, with tamper protection for key capabilities like metering.

### Stage 3: Full Embedded Offload

This stage enables full embedded offload. In addition to networking and security functions, storage access and persistence functions are offloaded. This stage enables the operator to achieve the benefits of a distributed computing platform architecture at scale:

- Offload of storage, networking, and network virtualization functions onto the FA
- Offload of hypervisor root partition from the CPU
- Security and isolation of the root of trust from the CPU

### Stage 4: Cloud-Native Workload Optimization

In this stage, the operator starts to offload the management control plane onto the FA as well. This means managing intra-datacenter and intra-server communication in cloud-native environments (like execution of the Envoy control plane). It also enables linear scaling of microservices in cloud-native environments, without saturating the mesh connectivity.

### Stage 5: Full Management Offload

This stage enables a full management offload in which the combination of Stage 3 and Stage 4 enables the creation of a distributed data-centric management and control layer, with hardware assisted pooled compute and storage resource sharing.

## OPPORTUNITIES FOR VMWARE

As Project Monterey starts to permeate datacenters and multicloud environments, IDC expects that eventually all of VMware's OEM partners will introduce a new breed of appliances and certified reference architectures. This is no doubt going to enhance the value proposition of converged and hyperconverged infrastructure.

With its extensive converged and hyperconverged ecosystem, VMware can play a role in bridging on-premises and cloud-based environments; and this represents a natural launch point for the company to expand into a new market segment for composable infrastructure, specifically for multicloud environments.

Longer term, this is where much of the value lies for Project Monterey: introducing a refined approach for executing the composable infrastructure at scale. As VMware pulls together its network of OEMs, accelerated compute providers, and the infrastructure software stack, IDC can see VMware defining a de facto path forward in which entities with disparate interests can all pull together to bring solutions to the market that redefine (and ultimately blur) the way the datacenter delivers its services to end users.

Project Monterey will enable organizations to adapt datacenter, cloud, or edge environments for application-specific performance, availability, and security needs. In addition, the initiative will extend VMware infrastructure and operations for all applications — reducing the need for specialized systems, teams, and management tools — which in turn will be able to reduce overall complexity and TCO.

## CONCLUSION

The current central processing unit-based architecture deployed in datacenters, edge devices, and consumer handhelds needs to be enhanced to work the era of hyperscale, multitenancy, and shared-everything computing. VMware's Project Monterey — by making use of function-offload accelerator-based computing platforms — provides a robust approach for offloading embedded and management functions. By embracing Project Monterey as part of their VMware Cloud Foundation and VMware vSphere deployments, customers can gain a consistent operating experience for their multicloud environments.

## About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-community.com
www.idc.com