White Paper

Cloud Service Providers
Infrastructure Processing Unit (IPU)

IPU-Based Cloud Infrastructure: The Fulcrum for Digital Business

As Cloud Service Providers consider their investment strategies and technology plans for the future, IPUs offer a path to accelerate and financially optimize cloud services.

Authors

Andrew Moore
CEO/Digital Officer - Digital Nexus Associates

Jim Henrys
CEO/Strategy Officer - Digital Nexus Associates

Table of Contents

The Unstoppable Rise of Digital Business

2020 will go down in history as the year the world met with one of its most significant challenges in nearly a century – the great coronavirus pandemic. The resulting situation can be aptly described by the term "VUCA," an acronym standing for Volatility, Uncertainty, Complexity and Ambiguity. In other words, the world faced a crisis.

Against such a background, businesses might reasonably have been expected to batten down the hatches, retrench, and cut their investments. However, in an environment defined by lockdowns and social distancing, enlightened businesses the world over have realized that the best way to overcome crises, and continue trading, is through virtual means. As a result, they instead have begun to accelerate their digital efforts. Co-founder and former Intel CEO Andy Grove summed it up well: "Bad companies are destroyed by crises; Good companies survive them; Great companies are improved by them" - a truism more relevant now than ever.

Put simply, the speed at which this move to digital is happening would not have been possible without the Cloud and the Cloud Service Provider (CSP) community. This indispensability is reflected in market data: spending on public cloud services grew to $270 billion in 2020.[1] Looking forward, Gartner predicts a 23 percent year-on-year increase to $332 billion in 2021, and a nearly $400 billion spend in 2022.[2]

It is also worth taking into account how the uptake of next-generation technologies will likely further fuel innovation in the enterprise and corresponding demand for cloud capacity. Global businesses have seen an economic shift from products to services, to digitized services, and now to on-demand, personalized, intelligent services that blend the physical and digital worlds. This is not limited to business-to-consumer, but also encompasses business-to-business along with new styles of distanced working methods.

Much of this is and will continue to be made possible through "innovation accelerants". For example, 5G, artificial intelligence, mixed realities, and the edge / Internet of Things coming together to form brand new business and workplace models. These models are ultimately developed in, and delivered from, the Cloud. New business models range from the simple digitization of products and services to sharing, on-demand, co-creation, digital twins, mixed reality experiences and more.

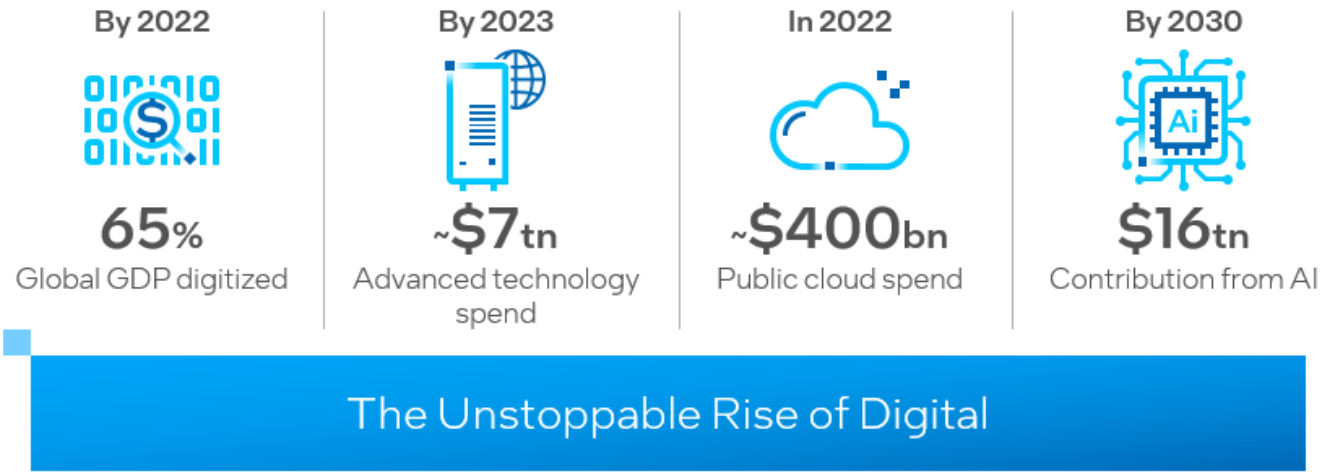

Figure 1. The CSP opportunity – digital business growth

Alongside new business models are new workplace practices, which include virtual presence, remote collaboration, and digital "white boarding." Looking forward, innovations such as voice recognition for meeting transcription could be included as well.

As the ripple effects of the pandemic fan out, it is becoming clearer that this likely isn't just a "digital blip," but part of a broader, sustained inflection point towards an even more rapid adoption of digital. This point of view is further supported by a forecast ~$7 trillion technology-related spend by 2023[7], which represents an eye-wateringly large opportunity for CSPs ready to invest and ride the digital wave.

Pandemic Inflection Point – examples of accelerated digital services:

- Virtual and video presence
- Online retail and 'click and collect'
- Streaming entertainment
- Logistics and transport (from more intelligent supply chains today to the use of autonomous vehicles tomorrow)
- Food delivery services (order digitally and potential for driverless vehicles and drones)
- Remote medical diagnosis
- Cash to digital payments
- BYOD as employees work from home

From Technology Provider to Key Strategic Business Partner

In a post-pandemic world, one thing has become clear as businesses look to reinvent and rebound: Technology and architecture choices matter more than ever. Accenture, in their Technology Vision 2021 report, states that 89 percent of executives believe their organization's ability to generate value will increasingly be based on the limitations and opportunities of their technology architecture.[8] In other words, leading and winning in this new world has as much to do with the ingenuity of technology choices as it historically has with the business plan.

For enterprises, the ability to develop, deploy and scale new services, using next-generation capabilities, at light speed becomes an imperative.

Notably, this pertains not only to the so called 'unicorns' but equally to all businesses. Comprehending what these leaders are doing, and how, provides insight into market shifts and sources of future disruption.

For the CSP, this understanding provides an opportunity to take on an increasingly critical role - not just as a technology provider, but as a key strategic business partner. In this situation, anticipating market demands, and being prepared, can be key delineating factors between success and mediocrity.

A new landscape has emerged, and with it a new axiom for business takes root - that nearly every business is becoming a technology business. Technology is no longer just one determinant of success - it is becoming the key determinant of success.

Transformational IT – The DevOps and Microservices Double Whammy

As enterprises strive for first-mover advantage in a digital world, the domains of Business and IT don't simply align, they blend. With this blending comes an increasingly important seat at the executive table for CDOs/CIOs as the role of corporate IT shifts from support and enabling to one of innovation, integration, and customer centricity.

With this in mind, it is worth highlighting two key IT strategies gaining increasing prominence in the enterprise as organizations seek to develop, deploy, and scale new services with velocity – DevOps and Microservices Architecture:

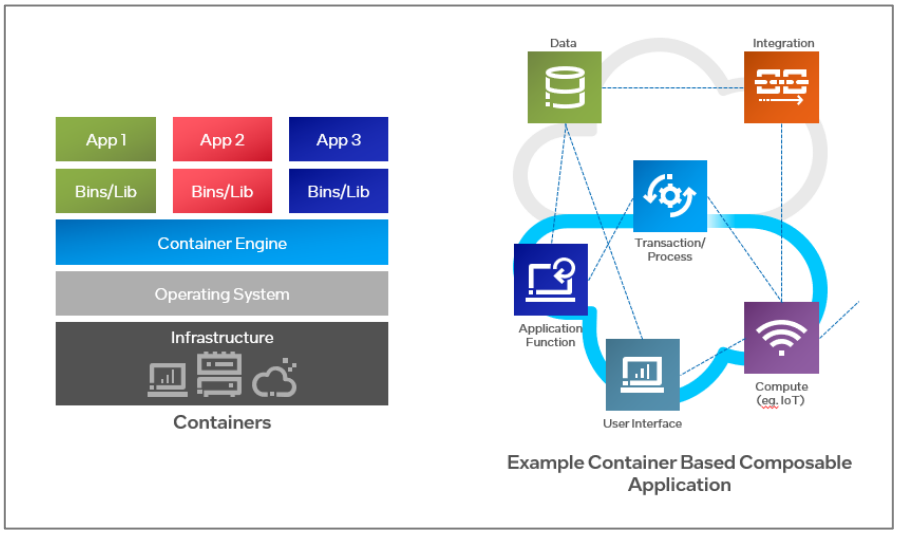

Figure 2. Container-based microservices applications

- **DevOps** combines software development and IT operations best practices to fast-track the service development lifecycle. Virtualization has been a key enabler for the automation of IT operations, meaning system configuration, management and provisioning tasks can be combined with software development activities, reducing time to market for solutions from weeks to days to (in some cases) minutes, while also making continuous service improvements possible.
- **Microservices-based software architecture** – Closely related to DevOps best practice is the move to a new software development paradigm based on the idea that monolithic applications that would previously have been run in virtual machines are broken down into a set of API-driven micro-services, with each instantiated in its own lightweight container.

The combined benefits of these approaches include:

- The ability to make continuous service improvements with minimal impact to production systems.
- The freedom to develop each micro-service independently using whatever development environment is the most pertinent.
- A considerably more efficient way to scale applications by adding containers at the micro-service level versus standing up many virtual machines, each with its own resource-hungry operating system.

The resulting "cloud-native" applications are lightweight in terms of resource utilization, nimble, cost-effective, and infinitely adjustable.

To punctuate this further, a recent O'Reilly Survey on Microservices Adoption in 2020 with input from over 1,500 technical leaders from around the world, found that 77 percent of companies said they are already doing it. 92 percent reported success and 89 percent of leaders believe companies that don't adopt a microservice approach will be unsuccessful.[9]

Infrastructure Matters!

Previous sections of this paper have highlighted the accelerated shift to new digital services in Enterprise, the way in which IT provides competitive changes to drive transformation and time-to-market competitive advantage, and the potential opportunity this presents to CSPs.

Together, those earlier discussions shine a spotlight on the key dependency for all of this - the foundation for digital services, DevOps, virtualization, and microservices: the Cloud Data Center.

Starting in the 2000s, big iron and proprietary infrastructure has, over time, largely been displaced by a software-defined, utility compute model. The de-facto basic building block of this new model, or "unit of compute," is the x86-based industry standard server. In this model, the server, in conjunction with various virtualization techniques, is deployed to provide compute, network, storage, security, and management functions.

Recently, this has been developed further with the arrival of converged and hyper-converged infrastructure driving even greater levels of automation and efficiency.

This approach to infrastructure is based on a philosophy of optimizing at the server node level. There, a general-purpose CPU is utilized to run all the software - from hypervisors and containers through to operating systems and applications.

However, as businesses move at scale to the CSP-provided cloud, there is need to diverge away from this server optimized architecture to one better designed to meet the needs of the CSP. This evolution should drive superior levels of optimization to increase both performance and profit.

To this end, Intel is bringing to market the Infrastructure Processing Unit (IPU).

IPUs – An Evolutionary Leap

In a typical "server optimized" enterprise data center, systems are designed for use by a single party, that is, the enterprise itself. However, in a CSP cloud data center, the workload is owned by the tenant, while the systems themselves are owned by the Service Provider.

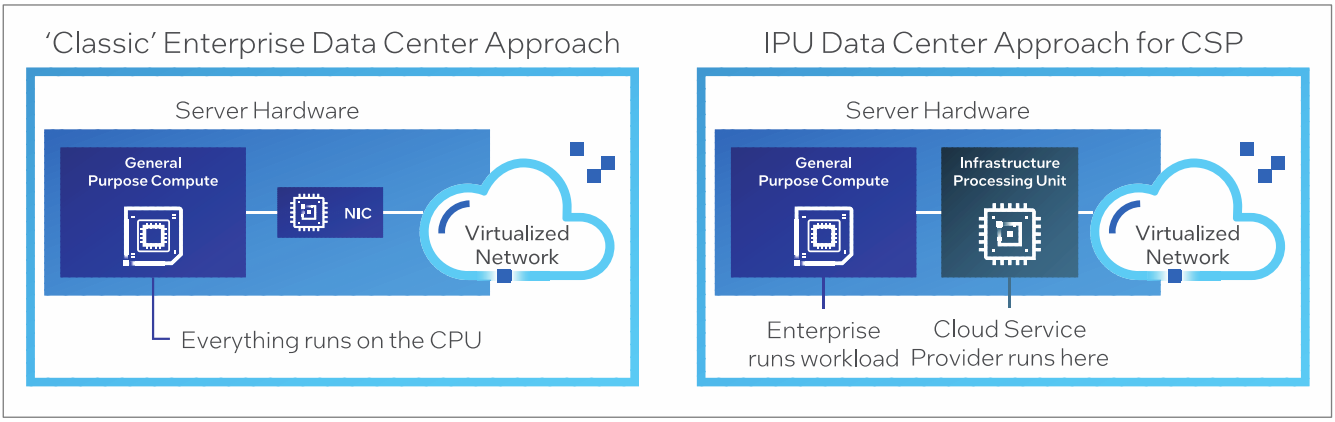

Figure 3. IPU 'disaggregation' in the CSP data center

Guido Appenzeller, CTO of Intel's Data Platforms Group, uses a simple metaphor comparing hotels to homes to explain this. In a home, it is convenient to have the kitchen close to the living room. However, in a hotel, the kitchen, where the food is prepared, and the dining room, where the guests eat, are clearly separated. In general, the areas where the staff work (the CSP) and those where the guests mingle (the tenant) are different.

Further, in highly virtualized environments, significant amounts of server resource are expended processing tasks beyond user applications, such as hypervisors, container engines, network and storage functions, security, and vast amounts of network traffic.

To address this challenge Intel has introduced a new class of product called the IPU. An IPU is an advanced networking device with hardened accelerators and Ethernet connectivity that accelerates and manages infrastructure functions using tightly coupled, dedicated, programmable cores. An IPU offers full infrastructure offload and provides an extra layer of security by serving as a control point of the host for running infrastructure applications.

By using an IPU, the overhead associated with running infrastructure tasks can be offloaded from the server (Figure 3). In other words, the CSP software runs on the IPU itself, while the tenant's applications run on the server CPU. This not only frees up resources on the server, whilst optimizing overall performance, but provides the CSP with a separate and secure control point.

This workload disaggregation is analogous to the separation of guests and staff in a hotel.

It's also important to note the difference between an IPU and a SmartNIC. A SmartNIC is a programmable network adapter that can accelerate infrastructure applications, however, unlike an IPU it does not provide offload capability to run the entire infrastructure stack and therefore does not give the service provider an extra layer of security and control, enforced in hardware.

The IPU in Action

As data center networking marches forward from 25 GbE, to 50 GbE and into the realm of Terabit Ethernet (100+ GbE), it creates unprecedented volumes of network traffic. The net result is an exponential increase in the number of packets transferred per second putting incremental strain on the capabilities of a traditional Network Interface Card (NIC).

Additionally, the advent of software-defined networking (SDN) puts more load onto servers as CPU cores are swallowed up with virtual switches, load balancing, encryption, deep packet inspection, and other I/O intensive tasks.

Add into the mix the increasing sophistication of management software running on servers, and it becomes evident that there is a genuine need to manage the explosive growth in network traffic while also offloading "infrastructure" workloads from server CPUs to enable more resources to be dedicated to mission-critical application processing.

To put this into context, studies have shown that networking in highly virtualized environments can consume upwards of 30 percent of the host's CPU cycles.[10]

IPUs combine hardware-based data paths, which can include FPGAs, with processor cores. This enables infrastructure processing at the speed of hardware to keep up with increasing network speeds and the flexibility of software to implement control plane functions.

With the development of its first IPU, Intel has combined onto a single card an Intel® Stratix® 10 FPGA, through which a high-speed Ethernet controller and programmable data path is implemented, along with an Intel® Xeon® D processor for the control plane functions.

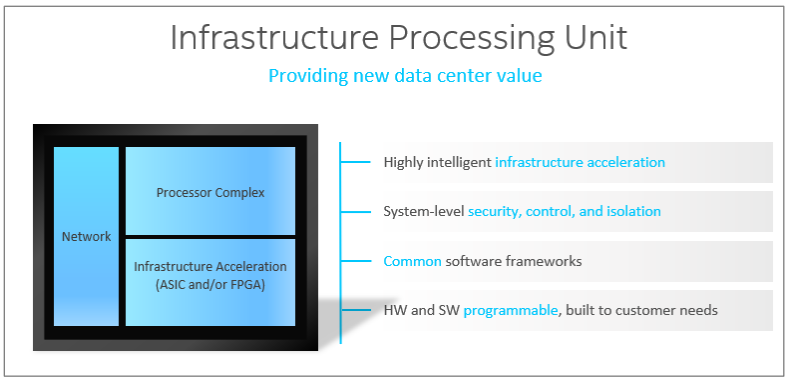

Figure 4. IPU conceptual architecture

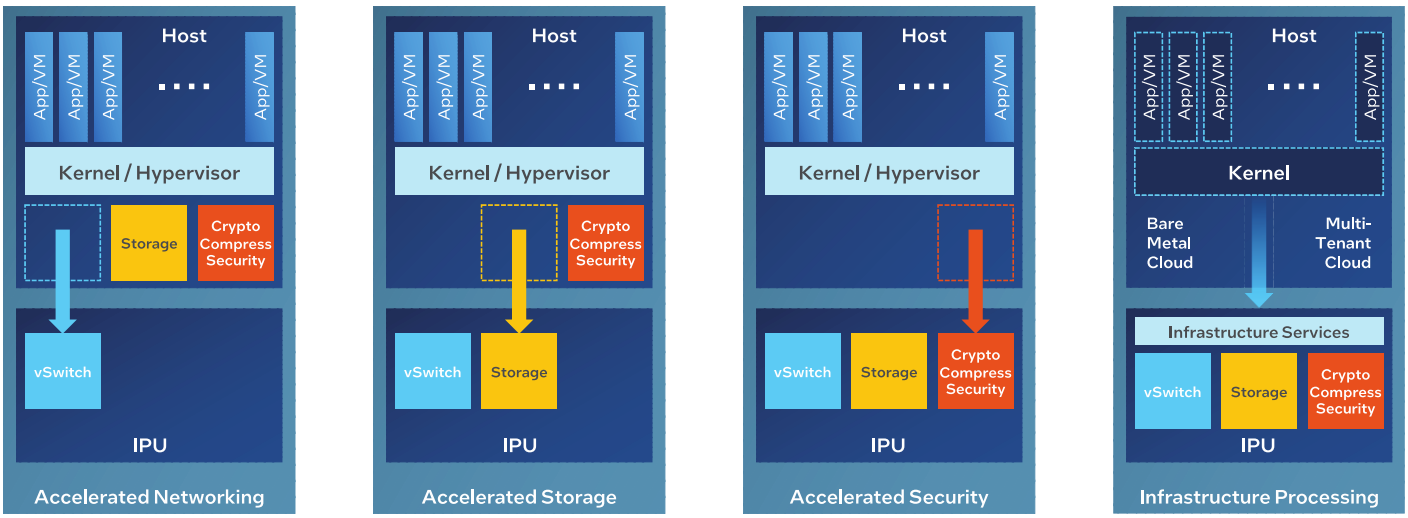

Figure 5. IPU 'offloading' use cases

Blending this capability with the ongoing trend in microservices development offers a unique opportunity for function-based infrastructure—achieved through matching optimal hardware components and common software frameworks to each application or service.

For the CSP, this represents an opportunity to accelerate the cloud while hosting more services (apps/virtual machines) on a single machine, leading to improved service delivery and greater profit potential per server.

Intel Advantage

The programmable logic blocks of Intel® Stratix® 10 FPGAs, provide a highly performant way in which to accomplish tasks such as packet processing. Moreover, an open standard Intel® Toolkit and associated libraries provide a unified programming model that enables developers to more rapidly and easily deploy solutions. In addition, the Intel® Xeon® D processor, with its emphasis on efficiency and networking, coupled with the open source Data Plane Development Kit (DPDK) and Storage Performance Development Kit (SPDK), supports offloading of server CPU tasks while offering full x86 compatibility - removing the need for recoding and reducing risk.

Building on extensive research and experience, Intel continues to work cooperatively across the networking ecosystem to bring to market IPU products from companies such as Silicom and Inventec.

Realizing Value – Use Case Examples

Leading hyperscale CSPs are driving IPU "offloading" use cases and have been realizing the value in stages. On a typical server, a host processor dedicates CPU cycles to a hypervisor, which may contain a virtual switch, storage stack, and security functions, plus the various applications that may run in a virtual machine or containers.

- Stage 1: **Accelerated Networking** – To offload virtual switch functionality, a common function, from the host application processor onto the IPU.
- Stage 2: **Accelerated Storage** – To move the storage stack from the host application processor onto the IPU, increasing throughput and reducing complexity and overhead.
- Stage 3: **Accelerated Security** – To offload encryption/decryption, compression, and other security functions that would otherwise be CPU intensive from the host application processor. (These functions are often paired with the networking and storage functions that have been offloaded in Stages 1 & 2).
- Stage 4: **Infrastructure Processing** – To offload the hypervisor services management functions from the host application processor down to the IPU.

Stage 4 delivers value in two ways:

Firstly, there is the **bare metal scenario** where a tenant rents an entire physical server while the CSP is able to manage its operation through the IPU offload approach, providing a separate security domain and physical isolation from the application server itself. This provides a full bare metal experience for the tenant with a separate platform (via the IPU) for the CSP to provision and secure the service.

Secondly, in a **multi-tenant virtualized environment** (Figure 5) where a hypervisor has many applications running in VMs, the hypervisor functionality is offloaded onto the IPU. This provides better isolation for the CSP and the ability to fairly distribute the networking and storage infrastructure services to those virtual machines or containers.

It has been reported that hyperscale providers observe microservices communication overhead of between 22 percent and 80 percent of CPU cycles.[11] This means between 20 percent and 78 percent of expensive CPU resource is unavailable to "rent," hence the IPU does not just provide accelerated and more secure service, but also offers an opportunity to monetize more of a server's resources.

To put it another way, if in the previously mentioned highly virtualized environment a CPU can be unburdened of the 30 percent potential overhead, that CPU can be earning more revenue.

Summary: Moving Forward

The accelerative effect the pandemic has had on the shift to digital has created unprecedented opportunities for those who can innovate, with velocity, at scale, ahead of the market:

- ~$400Bn public cloud spend in 2022[12]
- ~$7Tn tech related spend by 2023[13]

A new axiom for business has taken root - that every business is a technology business. The ability to lead and win in this new world has as much to do with the ingenuity of your technology choices as it historically has with execution of the business plan. Technology is no longer just one determinant for success – it is becoming the key determinant of success.

Central to this shift are the IT strategies of DevOps and microservices. Collectively, these help organizations develop, deploy, and scale new services at velocity.

A key and frankly foundational dependency in this is the ongoing evolution of the Cloud Data Center, where the need to drive ever greater levels of optimization that increase both performance and profitability is a delineating factor between success and mediocrity.

To help providers take a leadership position in this race, Intel is bringing to market the IPU. Intel's IPU offerings move the needle on data center infrastructure. At both the hardware and software levels, they create a platform for optimization, innovation, and revenue growth.

To see how you can take advantage of these groundbreaking capabilities, contact your Intel Account Manager.

Contributors
Graham McKenzie, Field Application Engineer, Intel Programmable Solutions Group
Natalia Poliakova, Technical Solutions Sales, Asia Pacific Territory, Intel Programmable Solutions Group

References

[1] Gartner, Gartner Forecasts Worldwide Public Cloud End-User Spending to Grow 23% in 2021, https://www.gartner.com/en/newsroom/press-releases/2021-04-21-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-grow-23-percent-in-2021
[2] Gartner, Gartner Forecasts Worldwide Public Cloud End-User Spending to Grow 23% in 2021, https://www.gartner.com/en/newsroom/press-releases/2021-04-21-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-grow-23-percent-in-2021
[3] IDC Reveals 2021 Worldwide Digital Transformation Predictions; 65% of Global GDP Digitalized by 2022, Driving Over $6.8 Trillion of Direct DX Investments from 2020 to 2023, 29 Oct 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46967420
[4] IDC, IDC Reveals 2021 Worldwide Digital Transformation Predictions; 65% of Global GDP Digitalized by 2022, Driving Over $6.8 Trillion of Direct DX Investments from 2020 to 2023, 29 Oct 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46967420
[5] Gartner, Gartner Forecasts Worldwide Public Cloud End-User Spending to Grow 23% in 2021, https://www.gartner.com/en/newsroom/press-releases/2021-04-21-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-grow-23-percent-in-2021
[6] RT News, Artificial Intelligence to Contribute $16 Trillion to Global GDP by 2030, 17 Nov, 2018, https://www.rt.com/business/444240-ai-global-gdp-growth/
[7] IDC, IDC Reveals 2021 Worldwide Digital Transformation Predictions; 65% of Global GDP Digitalized by 2022, Driving Over $6.8 Trillion of Direct DX Investments from 2020 to 2023, 29 Oct 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46967420
[8] Accenture, Accenture Technology Vision 2021, https://www.accenture.com/us-en/insights/technology/technology-trends-2021?c=acn_us_technologyvisiogoogle_11975684&n=psgs_0221&gclid=CjwKCAjw55-HBhAHEiwARMCszqEmFTnmawv74vJVVtQooLWqo8Q0OFl50ADQgttucRSSB5TlSP8pSRoCfoIQAvD_BwE&gclsrc=aw.d
[9] Business Wire, O'Reilly's Microservices Adoption in 2020 Report Finds that 92% of Organizations are Experiencing Success with Microservices, https://www.businesswire.com/news/home/20200716005101/en/O%E2%80%99Reilly%E2%80%99s-Microservices-Adoption-in-2020-Report-Finds-that-92-of-Organizations-are-Experiencing-Success-with-Microservices#:~:text=The%20report%20found%20that%2077.92%25%20experiencing%20success%20with%20microservices.&text=The%20report%20surveyed%201%2C502%20software,makers%20from%20around%20the%20globe
[10] Evaluation Engineering, SmartNIC Architectures: A Shift to Accelerators and Why FPGAs are Poised to Dominate, 18 Oct, 2020, https://www.evaluationengineering.com/industries/communications/wireline-fiber-optic-ethernet-pcie-usb-etc/article/21158389/smartnic-architectures-a-shift-to-accelerators-and-why-fpgas-are-poised-to-dominate
[11] Patrick Moorhead, Forbes, Intel Announces The 'Infrastructure Processing Unit At The Six Five Summit 2021,' https://www.forbes.com/sites/patrickmoorhead/2021/06/14/intel-announces-the-infrastructure-processing-unit-at-the-six-five-summit-2021/?sh=4d1f3aa11bb5
[12] Gartner, Gartner Forecasts Worldwide Public Cloud End-User Spending to Grow 23% in 2021, https://www.gartner.com/en/newsroom/press-releases/2021-04-21-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-grow-23-percent-in-2021
[13] IDC, IDC Reveals 2021 Worldwide Digital Transformation Predictions; 65% of Global GDP Digitalized by 2022, Driving Over $6.8 Trillion of Direct DX Investments from 2020 to 2023, 29 Oct 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46967420