Converge! NETWORK DIGEST | AvidThink®

SmartNICs and Infrastructure Acceleration Report 2022

Enabling the Next Generation of Digital Services

RESEARCH BRIEF

Table of Contents

© Copyright 2022 AvidThink LLC, All Rights Reserved

Table of Contents

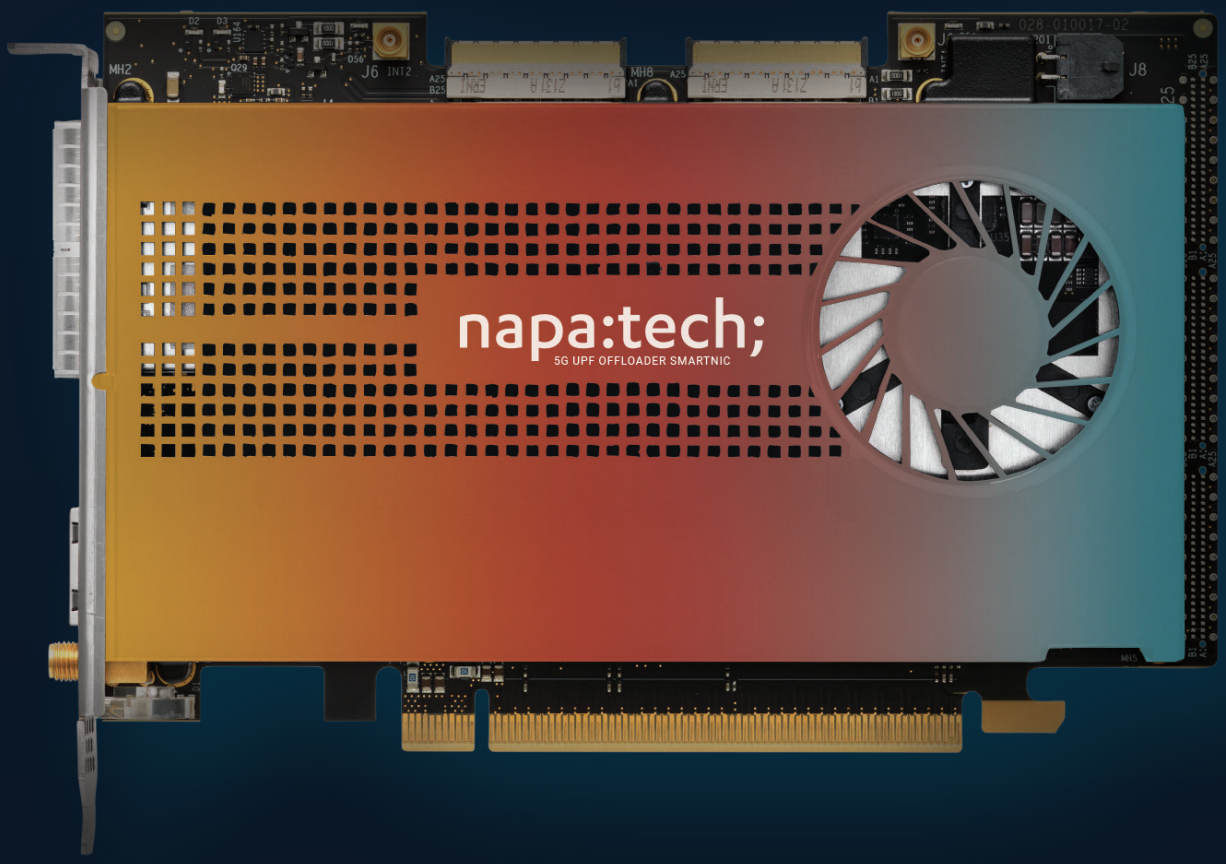
Research Briefs are independent content created by analysts working for AvidThink LLC. These reports are made possible through the sponsorship of our commercial supporters. Sponsors do not have any editorial control over the report content, and the views represented herein are solely those of AvidThink LLC. For more information about report sponsorships, please reach out to us at research@avidthink.com.

About AvidThink™

AvidThink is a research and analysis firm focused on providing cutting edge insights into the latest in infrastructure technologies. Formerly SDxCentral's research group, AvidThink launched as an independent company in October 2018. Over the last five years, over 110,000 copies of AvidThink's research reports (under the SDxCentral brand) have been downloaded by 40,000 technology buyers and industry thought leaders. AvidThink's expertise covers Edge and IoT, SD-WAN, cloud and containers, SDN, NFV, hyper-convergence and infrastructure applications for AI/ML and security. Visit AvidThink at www.avidthink.com.

© Copyright 2022 AvidThink LLC, All Rights Reserved

napa:tech;
SMARTNIC SOLUTIONS



Why are you wasting money on your 5G packet core?

Napatech's integrated hardware/software SmartNIC solution delivers **industry-leading performance and efficiency** for 5G User Plane Function deployments.

In a typical metro edge data center supporting 50,000 5G users, Napatech's solution enables you to support **7x more users per server** than an ASIC-based NIC, with **five-year CAPEX and OPEX savings of over 80%.**

When would you like to start saving money?

WWW.NAPATECH.COM

SMARTNIC SOLUTIONS

RESEARCH BRIEF

SmartNICs and Infrastructure Acceleration Report 2022

Enabling the Next Generation of Digital Services

Introduction

Accelerated digitization of businesses, governments, and consumers has driven growth in the consumption of computing, networking, and storage resources. Evidence lies in the ongoing growth in cloud computing which Gartner estimates at 21.7% in 2021 [1], where business-to-business (B2B) and business-to-consumer (B2C) applications and data are now increasingly being hosted. Likewise, mobile devices and internet-of-things (IoT) are seeing rapid growth, with an estimated 3.6 connected devices per capita worldwide, while broadband speeds worldwide continue to increase to an estimated 110.4Mbps in 2023 [2].

Meanwhile, application architectures and the underlying infrastructure that hosts and transports applications and data are shifting to accommodate the load and more stringent demands of real-time experiences expected as we embrace the all-digital metaverse and omniverse (and other related virtualized worlds).

Previous and current generations of data center and computing architecture, constrained by technology and physical limits, strain to meet these new requirements. Consequently, the technology industry, led by hyperscalers, infrastructure vendors, and carriers, has embarked on efforts to adopt infrastructure acceleration technologies. With the recent news that AMD will acquire Pensando Systems, a high-profile infrastructure acceleration startup, for $1.9B, the broader market has awoken to the importance of acceleration hardware and software.

With the recent news that AMD will acquire Pensando Systems, a high-profile infrastructure acceleration startup, for $1.9B, the broader market has awoken to the importance of acceleration hardware and software.

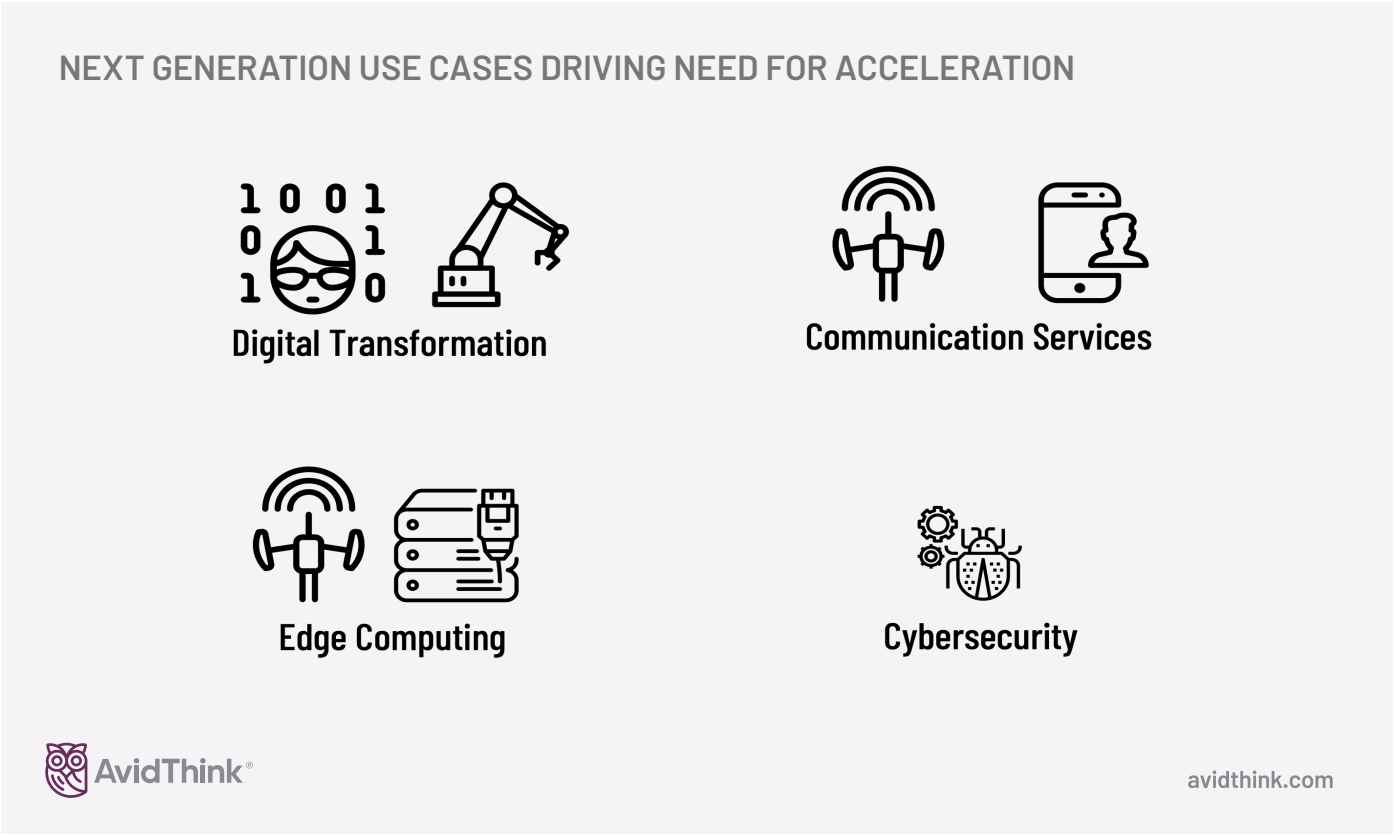
This AvidThink research brief, which updates and expands the previous version, aims to educate and update executives and technologists at communication service providers (CSPs) and enterprises on recent efforts in this field, particularly around networking acceleration in the form of smart network interface cards (SmartNICs), data processing units (DPUs)/infrastructure processing units (IPU), field-programmable gate arrays (FPGAs), and related technologies. We believe that understanding the foundational elements, tradeoffs, and landscape for infrastructure acceleration can help our readers make more informed and better business and technical decisions. As always, we welcome reader feedback at research@avidthink.com.

Next-Generation Use Cases Drive Next-Gen Workloads

To understand the pressures driving changes and adaption in the data center, at the edge, and throughout the computing complex, we first need to capture the business and societal drivers that impact applications and data. We've covered a number of these in the previous report but will revisit and expand them.

[1] Gartner Research, "Four Trends Are Shaping the Future of Public Cloud"
[2] Cisco, "Cisco Annual Internet Report"

© Copyright 2022 AvidThink LLC, All Rights Reserved



Digital Transformation, Data Analytics, and the Experience Economy (X-verse)

The ongoing digitization of our business and personal lives promises improved efficiency, agility, innovation, and experience. Companies that have successfully transformed physical processes into the digital world across retail, transportation, media, finance, healthcare, manufacturing, and other verticals have managed to reap the rewards of more significant return on capital investment while upgrading the user experience — think Uber, Lyft, Amazon, Netflix.

And the next step in transformation is ensconced within high-level constructs like Meta's metaverse and Nvidia's omniverse, which embrace experiential technologies like augmented and virtual reality (AR/VR) as part of extended reality (XR), and the ongoing digitization of real-world processes and devices like digital twins. Simultaneously, businesses are adopting advanced data analytics, artificial intelligence, and machine learning (AI/ML) to glean business insights and improve end-user experience.

The concomitant explosion in data that comes with digitization and the journey towards the X-verse concept will drive more network traffic, increased input/output (I/O) demands, more processing, memory bandwidth and capacity, and storage needs.

Communications Services (5G, Fiber, Network Virtualizations)

A big part of enabling digitalization is the underlying communications infrastructure which continues to get upgraded — in particular, 5G rollouts abound worldwide. Mobile network operators (MNO), while working out their 5G monetization strategies, continue to pump billions of dollars into upgrading their existing infrastructure to accommodate 5G — higher data rates, increased device density, and lower latencies. Meanwhile, global fiber buildout continues as we attempt to connect more homes, more businesses, and more people to the worldwide internet.

© Copyright 2022 AvidThink LLC, All Rights Reserved

For communication service providers (CSPs), there are a few trends of note. First, network functions virtualization (NFV) and its successor, cloud-native network functions (CNF), which are moves from physical dedicated hardware appliances to software-only versions of network functions on commercial-off-the-shelf (COTS) servers. This brings operators greater flexibility and agility and the opportunity to standardize hardware platforms, with the attendant benefits.

Second, operators are pushing disaggregation of the networking stack. For instance, the Open RAN initiative from the Telecom Infra Project (TIP), the ORAN Alliance, and the Open Networking Foundation (ONF) uses commercial-off-the-shelf (COTS) servers to process traffic from 5G radio networks. 5G open RAN (O-RAN), and virtual RAN (vRAN) workloads are growing in importance, and virtualized distributed units (vDU) and centralized units (vCU) are viewed as essential network functions to be deployed in telco networks. Similar disaggregation and open networking initiatives in high-speed wireline networks and optical transport are underway.

Third, both cloud data centers and telecom operators have adopted network virtualization, which represents a way of abstracting networks from their underlying physical connections, bringing more agility and flexibility to businesses and telecom operators.

The end result is a need to process all this network traffic, virtual and physical, at high speeds and at lower cost-per-bit as traffic scales up.

Edge Computing

Related to 5G networks is edge computing (sometimes called multi-access edge computing or MEC). Edge computing existed prior to 5G, but it's become prominent because 5 G's capabilities like ultra-low latency rely on the use of edge computing. CSPs worldwide are exploring edge computing applications in mobile and fixed-line networks that represent new business opportunities. IoT processing, analytics, big data filtering, and video and image processing are examples of edge applications that will require increased processing capabilities, often supported by graphic processing units (GPU). Similarly, AI/ML are critical to edge-powered use cases such as assisted driving, AR, video surveillance, and voice control and may require specialized hardware.

An additional challenge with the edge is constraints around power, space, and cooling in many edge locations, coupled with remoteness and lack of hands-on technical support..

Cybersecurity

Another trend that consumes significant computing power is the ongoing protection of our physical and digital infrastructure. While we transport bits across the network edge to the network core, they must be adequately protected — encryption, digital signatures and attestation, monitoring, and active defense against malicious eavesdropping and attacks. The same applies to data being processed and stored.

Security-related processing cycles represent a significant overhead that, unfortunately, will likely increase instead of leveling off or decreasing.

The Shifting Sands

Silicon has powered computing in our lives for more than half a century since it became the industry's preferred semiconductor substrate in the late 1950s. Transistor count, clock frequency, die area, feature size, core count, and other vital metrics have dominated our conversations and driven system design architecture. An industry, led by a handful of chip design and manufacturing firms offering a limited set of instruction set architectures (ISAs), has transformed in the last decade from being driven by general-purpose central processing units (GP-CPUs) to multiple silicon domains across graphics processing (GPUs) for gaming and simulation, artificial intelligence and machine learning (AI/ML), and network and data processing (NPU, DPU).

© Copyright 2022 AvidThink LLC, All Rights Reserved

At a higher level, system and data center architecture is undergoing transformation into multi-core systems, domain-specific hardware, composable hardware, data center-scale distributed computing, and storage systems. What's driving these changes?

The M.A.D. Laws

The changes in silicon, system, and data center hardware and software architecture are driven by both shifts in workload and constraints and innovation in technology. We'll discuss workloads and use cases shortly, but first, let's visit a couple of technology "laws," or observations and predictions as they are sometimes called.

Moore's Law

First, Moore's law, originally formulated in 1965, which is Gordon Moore's observation that the number of transistors in integrated circuits doubles about every two years[3]. It's more of an observation and historical trend projection than a physical law, but it's been a mainstay in semiconductor density and performance projections

Amdahl's Law

Next, we have Gene Amdahl's law, in which Amdahl, in a paper presented in 1967[4], observes the performance improvement gained by optimizing a single part of a system is limited by the fraction of time that the improved part is used. That means the maximum theoretical speedup of a workload is tied to the portion of the workload that is parallelizable.

Dennard Scaling Principles

Subsequently, Robert H. Dennard and co-authors described the relationship between transistor density, switching speed, and power dissipation in the paper "Design of Ion-implanted MOSFETs with Very Small Physical Dimensions" (1974)[5]. As we shrink transistors, power density stays constant, and power use stays in proportion with area. Meanwhile, both voltage and current scale downward with length, which should mean that chip designers can increase clock frequencies each technology generation without increasing power consumption.

Single Core to Multiple Cores to Diverse Cores

For decades, Moore's Laws and Dennard Scaling drove silicon roadmaps, with computing focused on leveraging ever-improving CPU performance driven by increasing transistor counts and clock frequencies increases. Unfortunately, due to current leakage, which causes increasing power consumption, and thermal effects, chip designers could not keep pushing the limits. Dennard Scaling broke down around the 2005-2007 timeframe, resulting in a limitation on CPU clock frequencies.

The diagram on the next page, from researcher Karl Rupp's open-source plot[6], provides visual evidence of this breakdown. The version below shows 50 years of data and is an update to a popular plot originally by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten titled "35 years of microprocessor trend data."

---

[3] 50 Years of Moore's Law. IEEE Spectrum Special Report 2015

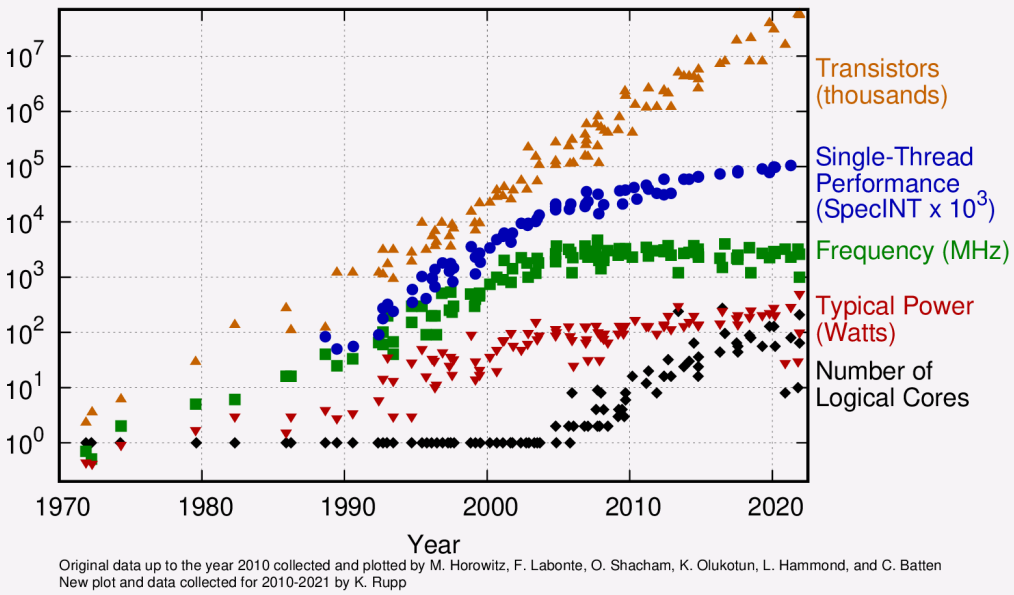
[4] G. M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities, Reprinted from the AFIPS Conference Proceedings, Vol. 30 (Atlantic City, N.J., Apr. 18–20), AFIPS Press, Reston, Va., 1967, pp. 483–485, when Dr. Amdahl was at International Business Machines Corporation, Sunnyvale, California," in IEEE Solid-State Circuits Society Newsletter, vol. 12, no. 3, pp. 19-20, Summer 2007, doi: 10.1109/N-SSC.2007.4785615.

[5] R. H. Dennard, F. H. Gaensslen, H. -N. Yu, V. L. Rideout, E. Bassous and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," in IEEE Journal of Solid-State Circuits, vol. 9, no. 5, pp. 256-268, Oct. 1974, doi: 10.1109/JSSC.1974.1050511.

[6] K. Rupp, "Microprocessor Trend Data GitHub repository"

© Copyright 2022 AvidThink LLC, All Rights Reserved

50 YEARS OF MICROPROCESSOR TREND DATA



Source: K. Rupp, "Microprocessor Trend Data GitHub repository"

The orange triangles show Moore's Law is hanging on, but the green triangles and red inverted triangles show Dennard Scaling breaking down around the mid-2000s, with the implications on single-threaded performance plateauing captured by the blue circles. Meanwhile, the move towards multi-core is shown by the black diamonds.

Remarkably, if processors had continued to double in performance every 18 months since 2003, as opposed to what's show above, then single-core systems would be 20 times more powerful than those today[7]. To compensate for the breakdown of Dennard Scaling, the industry shifted towards multi-core processors and multi-CPU systems. However, these multi-core systems are constrained by how much performance speed-up they can provide to the workloads running on them — as governed Amdahl's law.

In response to the slowing growth of single-threaded core, limitations of multi-core, and demands from new workloads (I/O-centric, AI/ML, graphics), silicon designers and system builders are taking different actions.

In response to the slowing growth of single-threaded core, limitations of multi-core, and demands from new workloads (I/O-centric, AI/ML, graphics), silicon designers and system builders are taking different actions, including innovating through heterogenous

[7] M. D. Hill and M. R. Marty, "Retrospective on Amdahl's Law in the Multicore Era," in Computer, vol. 50, no. 6, pp. 12-14, 2017, doi: 10.1109/MC.2017.164.

© Copyright 2022 AvidThink LLC, All Rights Reserved

multi-core designs, hardware customization and specialization, inventing new techniques in energy efficiency and power management, employing better orchestration in data movement, improving interconnect technologies, and co-evolving software and hardware coupling[8].

Application Architecture Evolution

Separately, application and software architecture has evolved, changing from monolithic all-in-one to component-based designs to distributed multi-tier applications hosted in centralized data centers serving web-based UIs or mobile apps via RESTful APIs. This evolution was in response to the need to develop application software faster, driving efficiency and productivity with larger teams and serving more application users from centralized cloud infrastructure — improving scaling, resiliency, and cost-efficiency.

Breakdown of Applications and Rise of Microservices

Today's application developers are focused on breaking down application components into smaller micro-services, each of which performs a limited set of functions well and adhere to distributed application principles of isolating state into a few components leaving the majority stateless for scalability and resiliency. Linux Containers are used for packaging and deployment of these micro-services, with Kubernetes as the de facto cluster and orchestration platform to manage these containers.

This micro-services architecture aligns with today's agile, continuous integration/continuous deployment (CD/CD) development pipelines and helps enable faster time to market with higher code quality and resiliency.

Simultaneously, there's an arc within application architecture towards an end goal of a serverless approach (which is a misnomer since servers are involved). The move towards serverless or function-based programming provides vital portability and scaling and distribution of application logic.

North-South, East-West Patterns

Earlier moves towards multi-tiered cloud-hosted applications had already driven traffic patterns from North-South (end-user client to web-hosted server) to East-West (between components within a data center). The micro-services architecture and serverless push are driving this even further. Researchers differ on the ratio of East-West to North-South traffic, with **older Cisco studies estimating roughly an 85%/15% split by 2020**, and others in the industry using 70%/30% or 80%/20% splits and others in the industry using 70%/30% or 80%/20% splits in their analysis today. Regardless, because of the changes in application workloads, East-West flows will continue to dominate data center traffic.

Historically, N-S traffic was processed by a conga line of dedicated network appliances that handled load-balancing, firewall and other security, encryption and decryption, and caching functions. Meanwhile, E-W traffic was relatively unprotected. Today's micro-service-heavy E-W traffic requires acceleration, scale, and security capabilities, often with micro-services communicating via service mesh or similar proxy mechanisms that handle discovery, load distribution, and security (encryption, mutual authentication).

[8] S. Borkar and A. A. Chien. "The Future of Microprocessors" in Communications of the ACM, Vol. 54 No. 5, Pages 67-77 doi:10.1145/1941487.1941507.

© Copyright 2022 AvidThink LLC, All Rights Reserved

Infrastructure Acceleration for Data Centers

The combination of changes in software and silicon architectures, along with the centralization of workloads in cloud data centers, has led to new challenges in these data centers. Google published a paper in 2015 that measured 20,000 machines in their data centers over three years, running thousands of different application workloads[9]. They found that the diversity of workloads supported a need for flexible architectures that can accelerate performance and identify a "datacenter tax" in lower layers of the software stack that comprise nearly 30% of cycles across jobs and that are prime candidates for hardware specialization and acceleration.

Another study by Facebook published in 2020[10] found that microservices spend as few as 18% of CPU cycles executing core application logic such as executing a key-value store. The remaining cycles were spent in common operations not core to the application logic, including I/O processing, logging, and compression). FaceBook believed that accelerating standard building blocks can significantly improve data center performance and built a model to project possible hardware speedup in microservices.

Unsurprisingly, hyperscalers were among the first to explore specialization and acceleration for data center workloads. Microsoft's Catapult project [11] was one of the first to reveal a hyperscaler's use of FPGAs in a SmartNIC form-factor for network acceleration. The use of network acceleration hardware is now known to be common practice across the major hyperscalers.

Before we jump into SmartNICs and the DPU, let's explore some of the other elements in accelerating workloads.

Key Infrastructure Acceleration Technologies

There are two main classes of network acceleration technology. The first is software-based, focused on providing a fast path for packets within a server while taking advantage of unique CPU and server architecture features. The second class is hardware-based, utilizing different processing architectures more suited to parse and dispatch network packets than general-purpose CPUs.

Software Accelerators

Packet Processing Acceleration

Generically, several companies provide software that accelerates network packet processing by, for example, a fast path architecture in which the data plane is split into two layers.

The lower layer, the fast path, processes the majority of incoming packets outside the OS environment and without incurring any of the OS overheads that degrade overall performance. Only those packets that require complex processing are forwarded to the OS networking stack (the upper layer of the data plane), which performs the necessary management, signaling, and control functions.

Some of the more open frameworks in this space include using eBPF (extended Berkeley Packet Filter), which allows user applications to compile packet handling code to run within the kernel and take appropriate actions, which might include invoking the OS networking stack if needed.

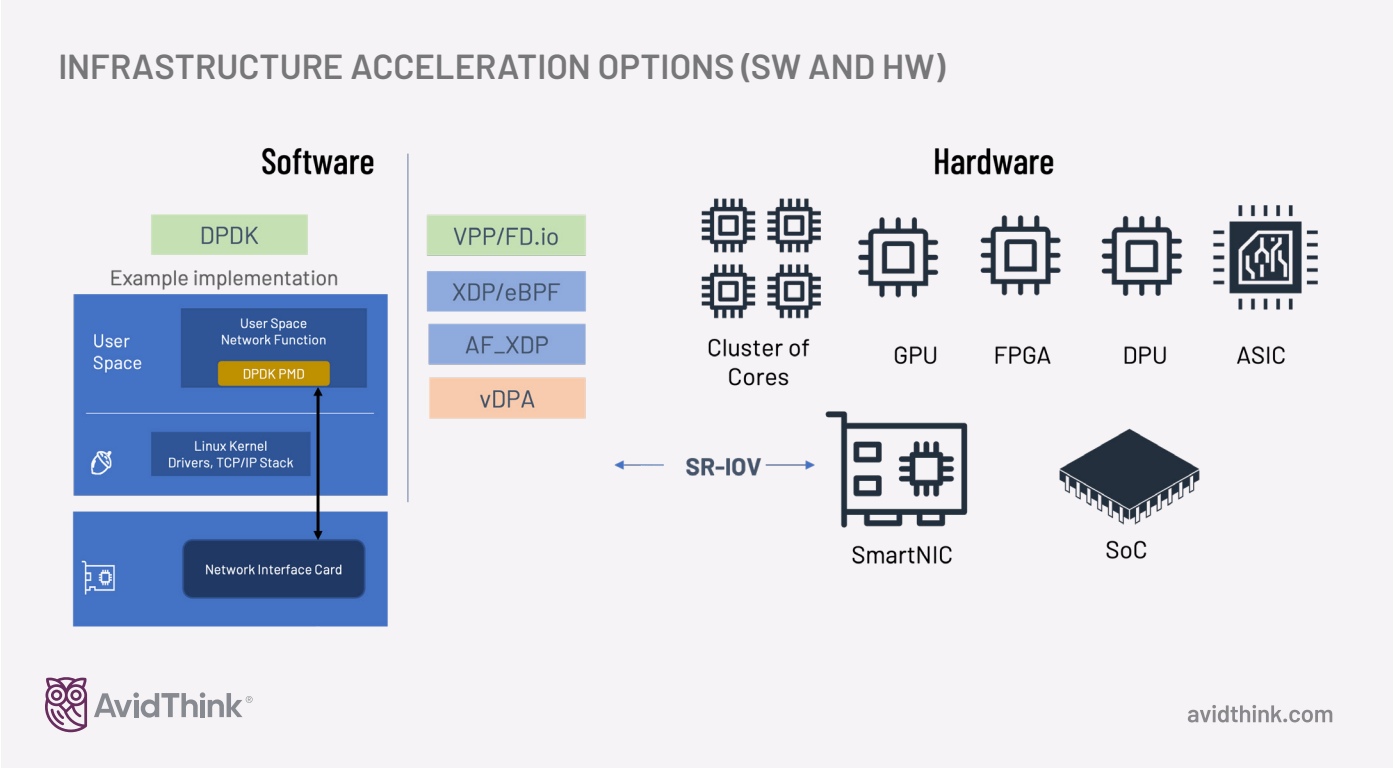
---

[9] S Kanev, J P Darago, K Hazelwood, P Ranganathan, T Moseley, G Wei, and D Brooks. "Profiling a warehouse-scale computer" in SIGARCH Comput. Archit. News 43, 3S (June 2015), 158–169. doi:10.1145/2872887.2750392

[10] A. Sriraman and A. Dhanotia. "Accelerometer: Understanding Acceleration Opportunities for Data Center Overheads at Hyperscale" in Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems. Association for Computing Machinery, New York, NY, USA, 733–750. doi:10.1145/3373376.3378450

[11] A. M. Caulfield et al., "A cloud-scale acceleration architecture," 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), 2016, pp. 1-13, doi: 10.1109/MICRO.2016.7783710

© Copyright 2022 AvidThink LLC, All Rights Reserved



DPDK

The Data Plane Development Kit (DPDK) software is a set of Linux user space libraries and drivers that accelerate packet processing workloads running on all major CPU architectures. DPDK is the most popular software-based acceleration solution for CSP workloads in NFV. It works by bypassing the OS kernel and handling the packets in the application user space.

An open-source project run by the Linux Foundation, DPDK is used for specific networking functions where high throughput and/or low latency are critical. It features in most NFV solutions.

VPP

Vector Packet Processing (VPP) is an extensible packet processing framework for network-intensive applications. It comes with a set of layer 2/3 switching and routing functionality built-in. Initially developed by Cisco Systems, it is now part of the open-source project FD.io hosted by the Linux Foundation. VPP is used in many layer 4-7 network functions to achieve higher throughput and better CPU efficiency.

VPP can run on DPDK so that instead of using kernel drivers to get packets from the hardware, it takes direct hardware control to speed up the packet path, resulting in faster processing. VPP ensures that the fewest cycles possible are spent on packet processing by processing multiple packets in batches, so the CPU's caches remain hot, and cache misses are avoided.

There are other software approaches, including eXpress Data Path (XDP/eBPF), address family XDP (AF_XDP), and virtio data path acceleration (vDPA), for which we refer you to our report "**Myth-busting DPDK in 2020**" commissioned by the Linux Foundation, for details. These approaches continue to be worked on today, with Red Hat recently making a larger push for the use of vDPA[12].

[12] Red Hat Blog Post "Hyperscale virtio/vDPA introduction: One control plane to rule them all"

© Copyright 2022 AvidThink LLC, All Rights Reserved

Silicon-based Platforms

The components that are used to build hardware acceleration are manifold, from system-on-chips (SoCs) that incorporate multiple CPU cores (usually Arm architecture) with hardware accelerators for regular expression (regex) matching, encryption/decryption, compression/decompression, to FPGAs, to GPUs, to various flavors of custom application-specific integrated circuits (ASICs). Further, the packaging or configuration that they take may vary — often as a server NIC, but sometimes in a dedicated appliance, or as an augmentation to a networking switch.

These domain-specific hardware solutions are designed to outperform CPUs for network I/O tasks in both sheer performance and in both costs per bit and energy consumed per bit processed. Let's examine standard components and technologies used for network acceleration and work our way up into packaging form factors.

FPGAs

FPGAs have been used as a preliminary step in designing Application-Specific Integrated Circuits (ASICs), accelerating time-to-market while volumes are low. They have also been used for implementing fast custom logic in specialized fields, such as high-frequency trading (HFT), where nanoseconds matter in tick-to-trade situations.

FPGAs are developed using the same tools as those used to design ASICs but can be rewritten or reconfigured on the fly. FPGAs can be used for the acceleration of network traffic and video processing and also for AI and ML. FPGAs combine dedicated hardware acceleration with software-like flexibility, but usually at a higher price-point and increased power consumption when compared to ASICs.

GPUs

As their name implies, GPUs were originally designed to render computer graphics. However, their architectures, comprising multiple processor cores working in tandem, also make them valuable as coprocessors for workloads that are inherently parallelizable, like networking. GPUs are used in network equipment to boost line rates while minimizing software development costs by using existing GPU programming models.

GPUs are used for training in AI and ML applications (and for inferencing) and can be used in processing wireless RAN traffic.

ASICs

ASICs implement custom logic and are more cost-effective, power-efficient, and performant than an FPGA. However, the upfront development and fabrication costs run into millions of dollars for a moderately-sized ASIC, so ASICs are typically used only for high-volume applications where the design is stable.

A key advantage of ASICs for network infrastructure applications is that, unlike FPGAs, they can include analog circuits such as transceivers on the same die as CPU cores.

Further, ASICs don't always have to be fixed functions and can implement a certain amount of programmability, as evidenced by today's merchant silicon used in chips as well as ASIC-powered SmartNICs that are programmable via languages like P4.

DPU

The data processing unit (DPU) is not well-defined. Leading vendors in the space, including Pensando/AMD, Fungible, Marvell (who indicate Cavium OCTEON was the first DPU), Nvidia (Mellanox/Tilera), and those from the network processing unit (NPU) lineage, including Napatech, Netronome, debate the details. Intel has decided to use the term infrastructure processing unit (IPU), which they believe is more apt, to further stir this pot.

Regardless, the DPU is viewed as the next generation of silicon specialization that is focused on processing data (in the form of network packets) more efficiently and at higher speeds than CPUs and may rely on varying architectures, including using embedded CPU cores inside the DPU in combination with custom logic to create a programmable silicon component designed for optimizing I/O.

Despite the attempt to cleanly categorize each of the technology classes used on SmartNICs, the lines are blurring.

© Copyright 2022 AvidThink LLC, All Rights Reserved

SmartNICs and Other Form Factors

A SmartNIC is a network adapter that offloads processing tasks from the main CPU. Using an onboard multi-core processor, DPU, FPGA, or ASIC (or combinations of any of these), the SmartNIC performs functions such as encryption, decryption, firewall, filtering, TCP/IP processing, HTTP processing, virtual switching, storage, AI and ML.

SmartNICs can be utilized directly by a hypervisor or host OS or by virtual machines and containers through sharing mechanisms like SR-IOV and/or one of the software approaches described earlier.

Like the DPU, definitions of SmartNICs vary, but most in the industry agree that the following components are usually present:

- **High-speed NIC:** 25Gps, 50Gbps, 100Gbps, and multiples of those speeds - allows it to be directly connected to a network, this includes the MAC/PHY etc.
- **Packet acceleration logic:** DPU, FPGA, cluster of CPU cores, custom ASICs
- **Other acceleration logic** (if not in DPU, FPGA, ASIC): crypto, compression, regex
- **GPU:** used for AI/ML, other workloads
- **CPU:** GP-CPU cores for running control plane, management logic, general purpose OS
- **Memory controllers/memory:** DDR, high-bandwidth memory (HBM) for storing state and other lookup tables
- **Software stacks:** TCP/IP and general networking, storage stack, security stacks (TLS)
- **Bus connectivity:** PCIe Gen 3,4,5, Compute Express Link (CXL)
- **Security subsystem:** including secure boot, root-of-trust (ROT), secure enclaves

This list is rapidly evolving, and form factors specializing. For example, **Dell's Open RAN Accelerator Card that uses Marvell's OCTEON Fusion platform** with a 5G layer-1 hardware accelerator for 5G open RAN workloads, and **a similar card from HPE powered by Qualcomm's technology**.

Storage is another major use case for hardware accelerators. Non-Volatile Memory Express (NVMe) is a crucial storage technology, and even more so remote storage protocols such as NVMe over Fiber (NVMe-oF), NVMe over TCP (NVMe/TCP), and Remote Direct Memory Access over Converged Ethernet (ROCEv2). Accelerating these protocols can dramatically improve application performance. And that's the driver behind packaging DPU into storage appliances (like **Fungible's FS1600 Node**) or **NetApp A400 storage appliance using Pensando chips**.

And in data center networking, DPUs and acceleration don't always have to reside on the server side; Pensando and HPE inserted DPU technology into the **Aruba CX 10000 switches**, which allows it to centrally run network workloads like firewalls or load-balancers.

Hyperscalers lead SmartNICs Revolution

CSPs and enterprises look to hyperscalers for cutting-edge best practices in data centers and computing technology. And unsurprisingly, SmartNICs were not even considered an option by CSPs until hyperscalers revealed their use — AvidThink's surveys of CSPs three years ago indicated resistance to the idea. Still, recent conversations indicate a solid willingness to adopt it today, especially for edge and 5G vRAN/O-RAN workloads.

Amazon AWS

Amazon's Nitro system is the underlying platform for their Elastic Compute Cloud (EC2) web service that offloads networking, storage, and management services from the host servers to dedicated hardware. The initial version of Nitro was based on

© Copyright 2022 AvidThink LLC, All Rights Reserved

commercial-off-the-shelf (COTS) silicon, but in 2015 AWS acquired Annapurna Labs to develop custom ASICs that improve hardware performance and move Elastic Block Store (EBS) processing from the main CPU and offload all I/O virtualization to separate cards.

The Amazon family of Nitro cards includes cards for Virtual Private Cloud (VPC), Elastic Block Store (EBS), instance storage, control plane, and security. Nitro now controls all AWS's compute infrastructure. The Nitro Hypervisor is a lightweight hypervisor that manages memory and CPU allocation, delivering an indistinguishable performance from bare metal. Nitro plays a heavy role in AWS' EC2 platform today, even on its on-premises platform, AWS Outposts. Nitro is used to power EC2 security capabilities, including Nitro Enclaves for processing sensitive data in an isolated environment and upcoming NitroTPM for trusted computing..

Google | Google Cloud

Within the Google Cloud platform, the Andromeda virtual network stack leverages NVIDIA GPU accelerator platforms for fast processing. Google positions its "100 Gbps Accelerator VMs" as ideal for high-throughput applications like scientific modeling, high-performance web servers, virtual network appliances, multiplayer gaming, video encoding services, distributed analytics, machine learning, and deep learning.

And Andromeda takes advantage of Intel QuickData DMA Engines to offload payload copies of larger packets. Driving the DMA hardware directly from the Andromeda stack, which bypasses the OS, enables the stack to spend more time processing packets rather than moving data around. The Andromeda architecture allows Google to offload other virtual network processing to hardware opportunistically, improving performance and efficiency without requiring the use of Single Root Input / Output Virtualization (SR-IOV) or different approaches that tie a VM to a physical machine for its lifetime.

Google's analysis shows that these capabilities have allowed them to seamlessly upgrade their network infrastructure across five generations of virtual networking, increasing VM-to-VM bandwidth by nearly 18x while reducing latency by 8x

Microsoft | Microsoft Azure

The Microsoft Azure Accelerated Networking[13] (AccelNet) solution offloads host networking to custom-developed SmartNICs implemented in FPGAs. Microsoft selected an FPGA approach because they found that ASICs would not provide sufficient

Building the Smartest SmartNIC

There's ongoing debate between the vendors on what's the best technology to accelerate packets. The three main choices (with obvious combinations in between) are:

- Cluster of increasing number of GP-CPU cores with specialized accelerators — more flexibility, easier programmability
- FPGA — programmable but arguably harder that GP-CPUs, low latency and jitter
- ASICs — best price/performance, lower power consumption, more hard-coded (though can implement programmability like P4)

Our expectation is that many SmartNICs will incorporate more than one of these technologies, sometimes in concert with specialized silicon that's use-case specific (e.g. 5G RAN), and there's already blended architectures with custom ASICs with embedded CPU cores, FPGAs with embedded CPU cores.

© Copyright 2022 AvidThink LLC, All Rights Reserved

programmability, limiting their adaptability over time, while embedded CPU cores in an ASIC would not provide scalable performance, especially on single network flows. Multicore, System-on-Chip (SoC)-based SmartNICs were also evaluated, but considered to be inferior in terms of latency, price and power, especially at high bandwidths.

Azure SmartNICs implementing AccelNet have been deployed on all new Azure servers since late 2015, with AccelNet service available to customers since 2016, providing consistent sub-15µs VM-to-VM TCP latencies and 32 Gbps throughput (performance is higher now, based on publicly-available test results published by third-parties, but not endorsed by Microsoft). AccelNet supports major Microsoft services such as Bing and Office365.

Architecturally, AccelNet enables SR-IOV to a VM as a high-performance data path bypassing the host and thereby reducing latency, jitter, and CPU utilization. All the AccelNet SmartNICs in Azure are upgradeable, with the rules tables and/or configuration typically being upgraded quarterly with new functionality.

Current Landscape

With the ongoing demand for SmartNICs and acceleration technologies by the hyperscalers and carriers rolling out 5G initiatives, the market, since our last report in 2020, has responded with increased activity. Notably, major semiconductor players, including AMD, Broadcom, Intel, Marvell, and Nvidia, all have irons in the fire. The Pensando Systems acquisition and the recent closing of the Xilinx acquisition by AMD will drive greater market interest. Likewise, VMware's ongoing initiatives in this space around Project Monterey will bring the software and hypervisor elements into play.

All this points to a robust SmartNIC market, which has been recently estimated by Dell'Oro Group, a market research firm, to be $1.6B by 2026[14], with more than half of new servers deployed at hyperscaler cloud service providers equipped with SmartNICs.

Vendors

Select vendors are discussed in this section, and this is not an exhaustive list of vendors. These are vendors that were mentioned by network operators in both the cloud and telecommunications space during AvidThink's recent research. This space continues to include a number of proprietary vendors who focus on developing unique IP, sometimes for niche applications — such as Algo-Logic Systems, which provides the financial services with low-latency, high-performance solutions for trading and other use cases. In the interest of brevity, we will not be able to cover all those players. If you believe there's a vendor that we've not listed below but that we should have, feel free to reach out to us at **research@avidthink.com**.

AMD/Pensando

PENSANDO

Recently acquired by AMD for $1.9B, Pensando Systems was founded by the famed "MPLS" team (Mario Mazzola, Prem Jain, Luca Cafiero, and Soni Jiandani) from Cisco, with former Cisco CEO John Chambers, as Chairman.

Pensando Systems has built its Distributed Services Platform (DSP) that includes key components, including a programmable P4 processor built around an Arm core that consumes only 30W at 100Gbps throughput, a Distributed Services Card (DSC) using this processor, and a Policy and Services Manager (PSM).

The DSC provides software-defined services at the server edge, eliminating an assortment of discrete appliances throughout the data center and simplifying IT operations. The Pensando DSC additionally enables pervasive network visibility using its

[13] D. Firestone et al. "Azure accelerated networking: SmartNICs in the public cloud" in NSDI'18: Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation, April 2018 were also evaluated, but considered to be inferior in terms of latency, price and power, especially at high bandwidths.

[14] B. Fung Dell'Oro Group "Market Research Reports on Ethernet Adapter & Smart NIC"

© Copyright 2022 AvidThink LLC, All Rights Reserved

hardware bi-directional flow streaming and traffic mirroring capabilities. Pensando aims to bring AWS Nitro-type ability to the non-hyperscaler market at a performance that's an order of magnitude better than AWS (based on a Pensando-sponsored benchmark).

At the time of acquisition, Pensando had taken a data center-centric approach, focusing on network edge security services and accelerating network I/O workloads. Pensando's customers include Microsoft Azure, Oracle Cloud, Equinix, Goldman Sachs, and NetApp (also a partner). HPE was both an investor and a partner, distributing Pensado-powered servers and incorporating Pensando into an Aruba switch..

AMD/Xilinx

AMD recently completed its acquisition of Xilinx, wrapping up the deal first announced in Oct 2020.

Xilinx provides both standalone FPGAs and SmartNICs that accelerate networking infrastructure. Through the 2019 acquisition of Solarflare, Xilinx offers the XtremeScale X2 series of SmartNIC Ethernet adapters with Onload kernel bypass technology. X2 SmartNICs provide real-time packet and flow information across thousands of virtual NICs.

Beyond the X2 series, Xilinx has launched a 25GbE solution, the Alveo U25, which is a PCIe Gen3 x8 NIC with dual 25GbE. The U25 has an FPGA with over half-a-million lookup tables (LUTs) and a quad-core Arm A53 processor for improved programmability. It also has 6 GB of DDR4 memory.

More recently, Xilinx's recent SN1000 Alveo 100GbE SmartNIC series brings on a 16-core Arm A72 NXP processor and can process traffic at 100 million packets per second (Mpps) in a 75W small form-factor with PCIe Gen4 and HBM2 memory.

Xilinx's solutions address a wide range of solutions, from custom logic in HFT and other financial applications to AI/ML, video analytics, data analytics, and networking and security use cases.

Ethernity Networks

Ethernity Networks provides networking and security solutions on programmable hardware for accelerating telco/cloud edge networks. Ethernity's FPGA logic offers complete data plane processing with a rich set of networking features, security, and a wide range of virtual function accelerations for optimizing edge networks.

Ethernity's advantage comes from its Router-on-NIC offering, which provides switch/router functionality and an NFVI gateway, delivered on their own FPGA-based SmartNIC while shipping an SoC version. With a programmable data plane and features such as load balancing, monitoring, and SLA, Ethernity's Router-on-NIC enables acceleration of SD-WAN, vRouter, vBNG, and numerous other edge applications.

Ethernity is focusing on 5G deployments, using their routing IP for a 5G DU vRouter, leveraging their unique wireless bonding for wireless backhaul, and adding support for 5G User Plane Function (UPF) acceleration, including traffic management, deep packet buffers, classification, tunneling, flow counters, and IPsec..

[14] D. Firestone et al. "Azure accelerated networking: SmartNICs in the public cloud" in NSDI'18: Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation, April 2018

© Copyright 2022 AvidThink LLC, All Rights Reserved

Fungible

Fungible was founded by Pradeep Sindhu (formerly founder and CEO at Juniper Networks) and Betrand Serlet (formerly SVP of Software Engineering at Apple).

Fungible positions its Data Processing Unit (DPU) as the "third socket" in data centers, complementing the CPU and GPU and promising to deliver benefits not just in performance per unit of power and space but also in strengthening reliability and security. The Fungible DPU is purpose-built to address two of the biggest challenges in scale-out data centers – inefficient data interchange between nodes and poor execution of data-centric computations. Fungible's end goal is to provide "fungibility" of resources across an entire data center, driving towards composable infrastructure at data center-scale.

Fungible has been pushing storage-centric use cases in data centers, especially with their dedicated storage appliance built using their DPU, the FS1600. Fungible also ships Fungible Accelerator Cards powered by their DPUs in 200/100/50GbE configurations to improve storage performance on servers. More recently, Fungible announced their GPU-Connect solution that allows GPUs in a data center to be pooled and connected over Ethernet to CPUs not on the same server.

Intel

Intel has embraced non-x86 CPU processing of I/O workload, leveraging their FPGA capabilities (from their Altera acquisition in 2015), and has chosen to use the infrastructure processing units (IPUs) to refer to their new acceleration SoC and cards, reserving the SmartNICs moniker for their older lines of products. SmartNICs include the N6000-PL (code-named Arrow Creek), built with Agilex FPGA and targeted at telco workloads, and Silicom FPGA SmartNIC N5010 (developed with partner Silicom), and their end-of-life PAC N3000.

In 2H2021, Intel announced a suite of recent IPU products that include Oak Springs Canyon, built using Agilex FPGA and a Xeon-D SoC that provides 2x100GbE and supports PCIe Gen4. Oak Springs Canyon has virtual switch acceleration, NVMe over Fabric, and RoCE support for storage acceleration.

More significant was their first ASIC-based P4-programmable IPU code-named Mount Evans, a 200Gbps SoC that was the result of co-design efforts with a hyperscaler. Mount Evans includes 16 Neoverse N1 high-frequency cores from Arm and crypto and compression engines. The fact that Intel's IPU SoC has Arm cores is significant and speaks to the Arm ISA dominance in network acceleration offload cards.

Across their IPU and SmartNIC product line, Intel supports Open Programmable Acceleration Engine (OPAE), a consistent software programming and API layer across FPGA product generations and platforms. To encourage the use of FPGA acceleration for data center workloads, Intel has open sourced OPAE

Marvell

Marvell's DPU heritage comes from their Cavium acquisition (closed in 2018). The Cavium OCTEON platform claims to be the original DPU, tying together MIPS64 cores with acceleration logic. OCTEON became a mainstay of security appliances across the industry. Since then, the OCTEON DPU has evolved and now is an Arm-based SoC.

The most recent in the line is the OCTEON 10, which includes 8 to 36 Arm Neoverse N2 cores that utilize the latest Arm9 architecture. OCTEON 10 is supposed to be up to 3X faster than the previous generation OCTEON TX2 while drawing half the power. OCTEON 10 also includes AI/ML acceleration, along with its crypto accelerators, packet parsers, vector packet processing, and supports DDR5 memory and PCIe Gen5. OCTEON 10 supports up to 400GbE.

Marvell also sells their LiquidIO SmartNIC product that uses the OCTEON DPUs. LiquidIO III is still the most current solution, which uses the previous generation OCTEON TX2 DPU with 36 ARM V9 N2-based cores, and 16 GB of DDR memory.

© Copyright 2022 AvidThink LLC, All Rights Reserved

Additionally, Marvell has a family of OCTEON Fusion DPUs built on the TX2 platform and specialized for cellular base stations. In addition to the TX2 platform, OCTEON Fusion adds programmable DSP cores and baseband accelerators to handle the layer 1 functions in a RAN network. It is targeted at all-in-one base stations and DU processing in disaggregated RAN architectures (like O-RAN).

Marvell's DPU and SmartNICs solutions come with an SDK, supporting popular networking software and protocol stacks (most notably DPDK).

Nvidia Networking

Nvidia's networking and SmartNICs solutions come from their Mellanox acquisition which was initiated in April 2019 and closed in April 2020. Nvidia networking's ConnectX SmartNIC and BlueField DPU provide a range of acceleration options that accelerate workloads in the cloud, storage, security, broadcasting, AI, edge, and telco markets.

The ConnectX SmartNICs leverage built-in acceleration engines for RDMA over Converged Ethernet (RoCE), TLS/IPsec crypto offloads, accelerated switch, and packet processing (ASAP2) for virtual switching/routing, and NVMe over Fabrics storage, in addition to traditional networking offloads. ConnectX-7 is the most recent release in the ConnectX family and supports 400Gbps of throughput.

Concurrently released with ConnectX-7 is the BlueField-3 DPU that features 16 64-bit Armv8 A78 Hercules cores in a single SOC. The DPU offers either 400GbE or NDR 400Gbps Infiniband connectivity. The DPU also features 16 GB onboard DDR5 memory and supports PCIe Gen 5. BlueField accelerates control and data plane performance and functionality for cloud and edge, such as NVMe SNAP storage virtualization, bare metal networking, or cloud-native and security applications, including analytics, micro-segmentation, and firewalls.

Nvidia also provides their DOCA framework for programming their DPU platform. DOCA software consists of an SDK and a runtime environment. DOCA SDK provides industry-standard open APIs and frameworks, including DPDK for networking and security and Storage Performance Development Kit (SPDK) for storage. DOCA-based services are exposed as industry-standard input/output (IO) interfaces, enabling infrastructure virtualization and isolation.

Napatech

Napatech delivers integrated hardware/software solutions based around SmartNICs, which accelerate compute-intensive applications running on standard servers. Since 2003, their packet capture and security offload solutions have been deployed in applications like cybersecurity, financial systems, telecom infrastructure, data centers, and monitoring.

Their new 5G User Plane Function (UPF) offload solution comprises a complete UPF fast path implemented as part of their cloud-native software stack, compatible with standard APIs, and running on selected SmartNICs within their portfolio. Using a single 200G SmartNIC to sustain 100G of full-duplex traffic, this solution processes up to 100 million concurrent flows and achieves a total throughput of up to 85 million packets per second, which enables full wire-speed operation for typical packet sizes.

In a deployment scenario such as a metro edge data center running a 5G packet core to support fifty thousand users, Napatech's analysis shows that their solution enables carriers to support seven times more users per server than an ASIC-based NIC, with reductions of over 80% in both CAPEX and OPEX. Napatech integrates its solution with leading packet core software suppliers, ensuring that it is available as part of end-to-end, pre-validated software ready for deployment in carrier networks.

© Copyright 2022 AvidThink LLC, All Rights Reserved

Netronome

Netronome's Agilio SmartNICs offer a network data plane acceleration platform with support for accelerating Open vSwitch, Contrail vRouter / Tungsten Fabric, and eBPF- based data planes. The Agilio family includes low-profile CX for computing nodes running virtual network functions, FX for bare metal servers, and LX for service nodes that serve mobile core and gateway operations.

Netronome's Express VirtIO (XVIO) supports VM migration while offloading the hypervisor's networking functions. Acceleration extends to security features like connection tracking, DoS prevention, and IPsec / TLS encryption and as QoS features like metering and scheduling/shaping. Transparent inline TLS / SSH decryption with built-in load balancing enables scaling of content-aware security and visibility applications.

VMware and Project Monterey

Project Monterey is a notable initiative that represents VMware's move into SmartNICs and involves Intel, Pensando, Nvidia as SmartNIC/DPU vendors and Dell, HPE, and Lenovo as server vendors. Project Monterey enables VMware's ESXi hypervisor to run on SmartNICs. VMware envisions deployments where the hypervisor on the SmartNIC controls the host system, which may be bare metal or which may itself run another instance of the ESXi hypervisor.

The goals behind this project include:

- Improving security and manageability by using the SmartNIC hypervisor to act as an air gap and secure supervisor to the host system
- Enhancing storage and network I/O performance, offloading these to the SmartNIC
- Providing bare metal OS support, allowing the SmartNIC hypervisor to manage and deploy the OS on the host

It aims to provide a level of composability and fluidity around making different resources like FPGA or DPU acceleration units available to host VMs and containers.

The Future of SmartNICs and Infrastructure Acceleration

SmartNICs and DPU have gone from being esoteric and specialized hardware elements used only by hyperscalers and for unique workloads to increasing adoption by enterprises and CSPs. There's little debate today that as we move towards higher networking and I/O speeds, analyze growing amounts of data, and demand encryption for all traffic that SmartNICs and DPUs are necessary to achieve the performance required at the costs and power budget available.

As we add more workloads at edge locations (5G RAN, edge computing), these domain-specific approaches will be necessary to eke out the maximum performance given constrained footprints and power.

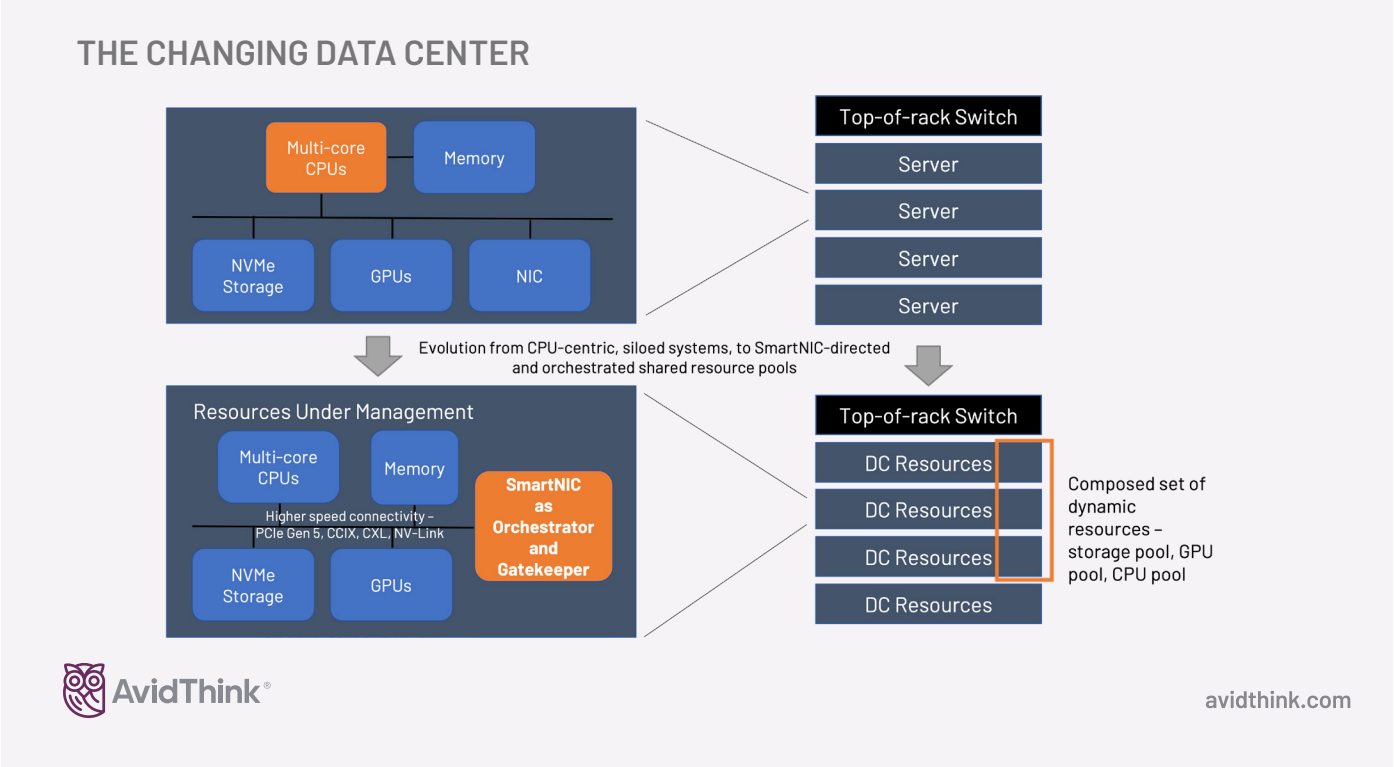
With regard to the architecture of SmartNICs, AvidThink believes that the technical elements will be blended with multi-core SoCs, FPGAs, programmable ASICs, custom acceleration, crypto logic blocks, and even GPUs coexisting on the card. The mix of which elements are used will depend on the targeted workloads.

Beyond workload acceleration, though, the isolation and supervisory capabilities of the SmartNIC point to a potential inversion of relationships between components in the standard server and across the data center.

Who's the Boss?

Multiple vendors, including Pensando Systems and VMware via Project Monterey, as well as the hyperscalers (AWS Nitro), are positioning the SmartNIC as the new supervisory element for a server. Fungible positions the DPU as the new supervisory and distribution element within the data center.

© Copyright 2022 AvidThink LLC, All Rights Reserved



The premise is that the SmartNIC is a secure (with HW root-of-trust) platform that sees all network traffic coming in and going out and can provide isolation or help create an air gap between the outside world and the CPU/memory/data storage complex. It can also inspect all traffic and block malicious attempts to compromise the system.

Beyond this, the SmartNIC can be used to deploy and provision host OS on a bare metal server (Project Monterey), acting as a new type of hypervisor. The SmartNIC can be counted on to aggregate and directly manage storage elements, GPUs, and likely host memory.

In the current and previous generations of server architecture, the server CPU is the primary gatekeeper and source of control and management. The arrival of the SmartNIC and DPU might usurp that position, relegating the server CPU into a pool of available CPUs to be composed into a virtual compute instance, along with memory and storage.

Software Innovation Needed

While promising, there's still a journey to effectively utilize SmartNICs, and software will play a huge role. Whether Nvidia DOCA or the P4 programming language adopted by most of today's SmartNICs and DPU vendors or the open-source SONiC OS used to manage switches that could be applied to SmartNICs, software libraries that unlock the processing power of SmartNICs will be critical to their success.

Unlike hyperscalers who may have the staff and expertise to do research on and exploit the unique capabilities of the SmartNIC and DPU, most enterprises and CSPs will primarily consume SmartNICs and expect that the software they load on these systems be already optimized. The load will fall on software stack and network functions vendors to ensure that their solutions detect and utilize the underlying SmartNICs to maximize performance.

This continues to be an area of active research and exploration. For instance, researchers like those at Rice University and the University of Washington are working on tools to analyze network functions and predict their performance speedup when ported a SmartNIC [15].

[15] Yiming Qiu, Qiao Kang, Ming Liu, and Ang Chen. 2020. Clara: Performance Clarity for SmartNIC Offloading. In Proceedings of the 19th ACM Workshop on Hot Topics in Networks (HotNets' 20). Association for Computing Machinery, New York, NY, USA, 16–22.

© Copyright 2022 AvidThink LLC, All Rights Reserved

OCP Open Domain-Specific Architecture (ODSA)

The other area of active exploration is how to accelerate the design and production of new domain-specific hardware. The Open Compute Project (OCP) Open Domain-Specific Architecture (ODSA) Sub-Project under the OCP Server Project aims to explore this area, taking an open approach to drive a standard that allows multiple vendors to provide IP blocks (chiplets) that can be combined and linked via high-speed interconnects within an SoC. ODSA aims to facilitate a marketplace where vendors can pick and choose key IP blocks to fit their unique needs, accelerating the creation of domain-specific acceleration hardware like SmartNICs.

Other Areas of Evolution

Other elements will have to be addressed as SmartNICs become more pervasive in the cloud, and edge deployments include improving troubleshooting and visibility. The ease of debugging packets being processed within a SmartNIC while maintaining isolation, security, and reasonable performance, will be a critical element of software development and adoption. Comprehensive visibility into the flows and accurate telemetry will be demanded by end-users who want to know what's going on with the I/O for their workloads.

Summary and Recommendations

With the constraints of growth of GP-CPUs, and the superior fit of DPU and SmartNICs for the new class of I/O-centric workloads demanded by today's use cases, DPUs and SmartNICs will continue to proliferate across data centers and edge locations. 5G, edge computing, AR/VR, and AI/ML use cases will drive increased uptake of these devices. Just as GPUs are now widely adopted for graphics processing and AI/MLworkloads , DPUs, and SmartNICs will be equally popular with networking and I/O workloads.

With this rising ecosystem and vendor innovation, enterprises and CSPs will want to explore SmartNICs, DPUs, and software technologies for infrastructure acceleration to improve the performance of their workloads and achieve better price/performance than possible with just using GP-CPUs.

© Copyright 2022 AvidThink LLC, All Rights Reserved

AvidThink, LLC
1900 Camden Ave
San Jose, California 95124 USA
avidthink.com

© Copyright 2022 AvidThinkI, LLC, All Rights Reserved
This material may not be copied, reproduced, or modified in whole or in part for any purpose except with express written permission from an authorized representative of AvidThink, LLC. In addition to such written permission to copy, reproduce, or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright notice must be clearly referenced. All Rights Reserved.

Rev A