

Our 2021 Server Trends & Observations

Earlier this year, we put out our annual list of [industry trends](#) we feel will be the most impactful to our server innovation workstreams. In this paper, Dell Technologies' senior technologists provide their insight on the trends, their influences, and how we're addressing the challenges and opportunities.

1 aaS Becomes the Enterprise Theme

Authors: Bill Dawkins (Fellow), Bhyrav Mutnury (Sr. Distinguished Eng.)

Cloud is an operating model, not a destination. The journey to the adoption of the cloud operating model began several decades ago and the revenue from services that use this model is expected to touch about 365 billion dollars by 2022 according to Gartner.¹ As the cloud operating model continues to be adopted on premises, as-a-Service (aaS) is becoming more important for the enterprise.

The on-premises aaS concept refers to bringing tools, technologies and products to enterprise customers in a subscription- or consumption-based paradigm. As-a-Service provides customers with features like flexibility and scalability while removing the burden of managing part or all of the infrastructure, ceding many responsibilities to the aaS provider. Also, customers don't incur the upfront burden for the complete infrastructure tools or technologies. These are managed by a service provider, and the end customer costs are based on subscription and usage. As-a-Service reduces the resources and skills needed by the customer to optimally run the datacenter. This model can also have the least maintenance overhead and may result in the best price-for-performance based on infrastructure utilization.

There are a many flavors of aaS, including Infrastructure as a Service (IaaS), Bare Metal as a Service (BMaaS), compute as a service, storage as a service (STaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Business Process as a Service (BPaaS). While cloud services continue to have their value, some enterprise customers will want to adopt these services offerings while retaining the benefits of on-premises locality. These include increased performance due to low-latency, control of the location of business-critical and valuable data, and increased security of infrastructure. In addition, on-premises aaS provides better customization (both software and hardware) to the enterprise and can potentially have lower total cost of ownership (TCO) over time.

aaS as an enterprise theme has implications for server technologies. aaS places greater emphasis on managing the infrastructure in a way that supports the changes in the customer's consumption model. Accurately forecasting the usage and deploying the resources with quick turnaround time is critical as reducing consumption of hardware or storage might not be easy later, so having an optimal configuration ahead of time is important.

Additionally, role-based management capabilities need to be enhanced as some aaS models will allow for a hybrid of customer managed and service provider managed infrastructure. A hybrid solution should offer consistent infrastructure and operations to eliminate silos and drive efficiency and agility. RESTful APIs, like DMTF's Redfish are required to allow the infrastructure to interact with the various levels of IaaS, PaaS, and SaaS. Dell EMC PowerEdge servers began the journey of adopting Redfish several generations ago to ensure customers have the most modern industry management standard.

¹ <https://www.gartner.com/en/newsroom/press-releases/2020-07-23-gartner-forecasts-worldwide-public-cloud-revenue-to-grow-6point3-percent-in-2020>

Rich telemetry for aaS infrastructure components becomes even more important as the service provider needs to monitor CPU, storage, network, applications, response times and other key performance metrics to ensure the customer gets the expected performance and capabilities of the services to which they have subscribed. Rich telemetry will also provide the means for the service provider to remediate infrastructure issues preemptively. With the emergence of AIOps, AI/ML techniques play an important role in failure prediction, workload mapping and forecasting. Security and Lifecycle management of data and infrastructure become foundational tenets for on-premises aaS and enhanced management capabilities become required by the service provider to perform infrastructure maintenance tasks (e.g., Firmware and hardware updates/upgrades) with minimal impact to the customer workloads and operations.

Dell Technologies is integrating the necessary capabilities for an on-premises architecture to address the aaS Enterprise theme with Project [APEX](#).

2 Server Growth Is Building Vertically

Authors: Alan Brumley (Sr. Distinguished Eng.), Joe Vivio (Fellow)

Customers have begun embracing the scalable software design methodology of using containers and microservices as a core design philosophy. This method gives customers the flexibility to quickly integrate with various services as needed and changes how servers are consumed.

Using storage as an example, customers are deploying software defined storage solutions, which are based on servers, and deploying them as vertical chunks in the datacenter. As the storage needs grow, more servers or storage nodes are added to increase capacity and scale performance. The storage service might require a protocol offload, via an FPGA or SOC enabled SmartNIC, to increase the performance of the individual nodes. Given the storage is being consumed as a service, the innovation is added without disrupting the core business software.

AI and ML are good examples of this freedom to innovate independently and provide services back to the datacenter. These AI services are provided via domain-specific silicon built and deployed on standard servers. Deployed as vertical solution stacks, they allow customers to choose the size of the AI function needed in the datacenter.

At a smaller scale, we expect similar deployment strategies at the edge, outside of the datacenter. Using 5G as an example, we are seeing the architecture allow scale up and scale down in certain networking functions inside of the 5G solutions. The functions are built around server technology as well as specific accelerators to maximize efficiency, all hardened to deal with the environmental challenges outside of the datacenter. Each of these functions can be independently scaled to meet the communication requirements of the deployment with available bandwidth, cell size, or customer usage expectations.

Farther away from the datacenter and in other edge workloads, we see the same solution architecture principles, but hyperconverged solutions become even more critical in server-based workloads. The ability to load SDS, along with compute and acceleration, into a small team of servers allows the flexibility to adopt to changing needs out at the edge.

These solutions will be offered prepackaged and turnkey by Dell, and in many cases delivered as racks to the data center. We can also package them in their own modular data center, ready to drop where the services are needed. We can deliver partial racks, self-contained and remotely managed at the edge, or multi-megawatt data centers ready to serve traffic.

3 More Data, More Smarts

Authors: Shawn Hoss (Sr. Distinguished Eng.), Bhavesh Patel (Distinguished Eng.),
Michael Bennett (Distinguished Member, Technical Staff)

Data continues to grow at an exponential rate with 5G, IoT and the need for real-time intelligent data curation and analysis at the source.

With the increase of data collection, transport cost, and complexity, the data generated is at risk of being siloed or trapped in the cloud. Siloed data cannot be efficiently used which reduces productivity and efficiency of operations. Real-time intelligent analytics and curation of the data at the source with flexible cost-effective low latency solutions are needed.

With this explosive growth in data, the need for [AI/ML solutions](#) will continue to grow across the industry to combat these challenges. Today we see more and more applications powered by artificial intelligence being deployed at the edge to analyze and curate the data at the source. Edge customers will face challenges in power, storage, and bandwidth, which require important choices regarding the operational flow of their AI-powered applications. A growing trend is to limit the retention of original data and instead store metadata that captures the features, values, and insights obtained from that data in a more concise format. This metadata also offers greater privacy and security because metadata can be anonymized.

Edge proliferation drives core data center/cloud growth: The 2021 growth in AI deployed at the edge will be supported by a large amount of compute and storage capacity hosted in a core datacenter environment where models are trained, updated, and tested against historical data. More AI-powered services such as image detection, natural language processing, and recommendation mean different hardware resource requirements. System administrators will become more knowledgeable on AI workloads, GPU acceleration, and the different requirements of training vs inferencing.

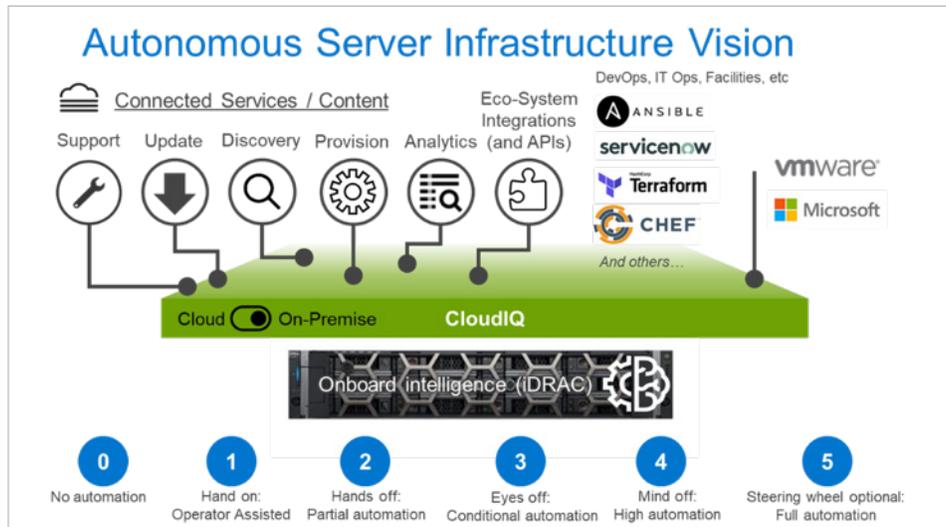
The core datacenter environment will also ingest the telemetry used to keep a handle on all the AI solutions running at the edge. Administrators will begin to deploy control planes in public or private cloud environments that enable them to manage and monitor the accelerator devices and models running on them. Public cloud will provide the necessary transport services from edge to core such as SD-WAN. For some industries such as ecommerce and social media, public cloud has become the AI edge. Customers in this situation will benefit from hybrid cloud solutions that allow them to store data at facilities with high-speed connections to cloud resources. These solutions offer an attractive option for making CapEx investments to reduce the monthly cloud storage bills that drive up OpEx.

AI applications will continue to be largely powered by CPUs, GPUs, FPGAs, and domain-specific [accelerators](#). We are seeing devices bifurcate to both low and high-power configurations. At the edge, we will likely see more very low-power versions of these processors. For the most part, high-power accelerators will reside in the core and multi-cloud environments. The hardware design to support these devices will need to support multiple environment and form-factor requirements. We will see AI/ML devices range from small devices like M.2 to form factors optimized for power, cooling, and bandwidth in the data center and cloud environments. With edge storage being limited, the core and cloud environments will require large storage footprint along with edge AI/ML for a scaling approach to solve the data growth problem.

Data is key to solve business problems, but a single paintbrush cannot be used when addressing AI/ML, and as a system designer you need to consider the type of accelerator, storage, high-speed networking, power, thermal, and host processor. AI/ML solutions utilizing SW and HW optimized for tasks and environments are key to managing the onslaught of data and getting smarter about how and where this data is managed.

4 The Emergence of the Self-Driving Server

Authors: Elie Jreij (Fellow), Jon Hass (Sr. Distinguished Eng.)



Managing the increasing deployment of servers outside the data center is a challenge. This presents a need for autonomous, or self-driving servers, that can manage themselves based on pre-defined and dynamically adjusted policies. To achieve a self-driving server, the following steps must be taken:

1. Configure and optimize telemetry data acquisition
2. Analyze the data for patterns and trends using AI/ML and statistical mechanisms
3. Apply analytics insights and policies through recommendation engines

Configure and optimize telemetry data acquisition

The trend is to support multiple types of streaming and sampling frequency that are optimized for use cases such as power consumption optimization, performance tuning, predictive failure, etc.

System components are getting smarter and have more inherent instrumentation that can be delivered via sideband interfaces, independent of the host processor being involved (e.g., PSUs, storage devices, processor chipsets).

Baseboard Mgmt. Controllers (BMCs) are more powerful with specialized capabilities like accelerators and encryption circuits. These BMCs cost less and are capable of running virtualized and containerized services and therefore enable embedded curation and early-stage processing of the telemetry, allowing:

- Customizable aggregation and interpolation of the raw telemetry
- Subsequent application of statistical and machine learning analytics
- Use of inferencing engines at the sources of the telemetry

Analyzing the data

The critical need to effectively analyze the operational data is moving towards three points of telemetry acquisition and processing:

1. Embedded in systems and components, early-stage processing, and analytics capabilities at the initial point of telemetry sampling.
2. In on-prem contexts, an aggregation entity processing telemetry acquired over local high-bandwidth, low-latency communications fabrics.
3. In cloud contexts, incorporating telemetry from many sites acquired over wide-area communications infrastructures and typically having lower bandwidth and potentially intermittent connectivity.

Autonomous server solutions are trending towards locating varying aspects of automation based on analytics and policy at all three points of telemetry acquisition. Processing and placement will depend on timeliness, robustness requirements, and self-driving goals. Tendencies are towards automating operations based on analytics outcomes as much as possible, coupled with visualization and indication techniques that allow administrators to inspect the automation analysis mechanisms and enable more autonomy as confidence in self-driving capabilities grows.

Managing the big data aspects of telemetry acquisition and processing will continue to be a major challenge in the various solution architectures. Open and closed source solutions for telemetry acquisition, analytics, visualization, and policy execution will continue to proliferate with offerings leveraging combinations to target specific situations, robustness, and time to market considerations.

Applying analytics insights and policies

Once the data is curated and stored, a machine learning process can analyze the data, observe trends, and correlate them with events such as component failures, network intrusions, and other anomalies. Once these trends are analyzed and associated with resulting events, an AI model can be developed and pushed into the server. The server will then monitor events and anticipate problem components or practices. A policy-based engine can be configured to make recommendations to administrators; or, once recommendations are trusted, automate decisions when trends are observed that are known to result in issues. The configured policy can decide the level of autonomy the administrator endorses.

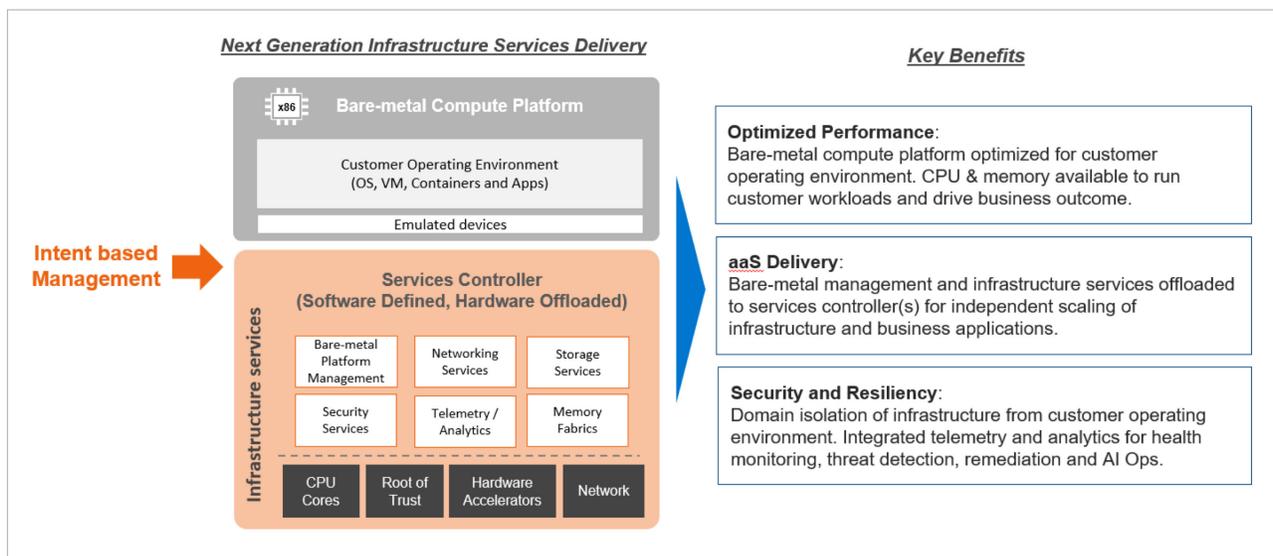
The earliest areas of focus for self-driving servers are:

- Capacity and utilization prediction including processor, memory, storage, communications (latency, bandwidth, errors), power, cooling/air flow, and failures
- Workload performance optimization
- Security compliance and compromises
- Anomaly detection and remediation

The emergence of the autonomous server will be a journey of creative and helpful administrative capabilities in which Dell will continue to lead by working closely with our customers to ensure the most valued features are the priority.

5 Goodbye, SW Defined. Hello, SW Defined with HW Offload

Author: Gaurav Chawla (Fellow)



In the last few years, we have seen the emergence of AI, an increase in network speeds to 25G/100G, and the transition of storage to flash and persistent memory. A new class of accelerators emerged to assist with data transformation and processing for AI/ML, networking, storage, and security. It includes GPUs, FPGAs, SmartNICs, and custom ASICs. We also saw a shift from virtualized applications to distributed containerized services and emergence of edge computing for distributed processing.

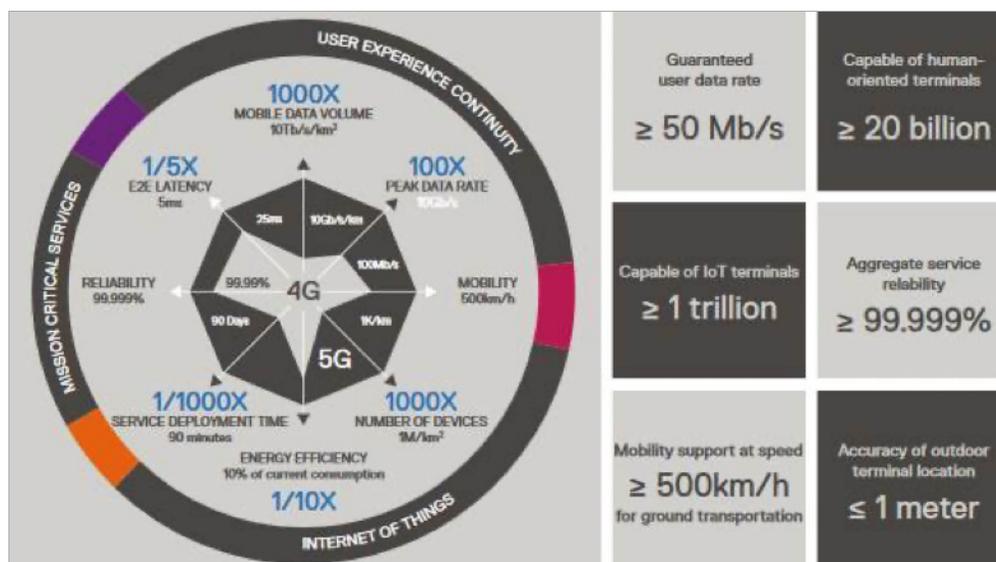
In 2021, we will see a transition of infrastructure services to “as-a-Service” delivery model. Software defined architectures will evolve to “Software Defined and HW Accelerated” with control plane and data plane separation. The control plane will run on CPU cores and the data plane will be embedded in the programmable hardware. Industry consortiums have started to standardize the interface to programmable hardware via [P4 for networking](#), and [SDXI \(Storage Data Accelerator Interface\)](#) for storage. SmartNICs have been a starting point on this journey and this architecture will continue to evolve in 2021, laying the foundation for disaggregated composable infrastructure, wherein infrastructure services are offloaded and composed with the bare-metal compute running containerized customer applications.

Such an architecture enables system-level optimizations in which data flows directly from the network to other system components like GPUs, FPGAs, storage, and persistent memory without requiring intervention by the host x86 compute running customer applications. The infrastructure services will be dynamically provisioned via an intent-based interface and composed with the bare-metal compute. Such an architecture disaggregates and offloads infrastructure services from customer applications and enables dynamic composability and independent scaling of business applications from infrastructure software. These architectures will come to fruition in 2021 and will further evolve in next 2-3 years with memory-based fabrics like CXL and Gen-Z.

Dell is partnering with VMware on [Project Monterey to disaggregate virtualization and containerized workloads](#). Networking, storage, and security services will follow a similar model and will transform how data center and edge infrastructure is designed, delivered, and managed.

6 5G is Here! Seriously, it is this year

Authors: Andy Butcher (Distinguished Member Technical Staff), Gaurav Chawla (Fellow)



Large carriers are aggressively deploying 5G connectivity in many cities across the world to enable new devices. A new class of communication providers have emerged with focus on software defined RAN, which will be built using servers with purpose-built accelerators for low-level protocol processing, replacing custom appliances used in 4G networks and laying the foundation for evolution to 6G. In parallel, industry consortiums published specifications defining an end-to-end open disaggregated 5G architecture to ensure interoperability and a multi-vendor ecosystem.

This includes standardization efforts in 3GPP, O-RAN and TIP (Telecom Infrastructure Project), and these efforts have been accompanied by open source projects in ONF (Open Networking Foundation) and Linux Foundation.

These industry efforts provide for the disaggregation of RAN that will enable a common software-defined architecture for micro-cells, macro-cells, private wireless, and fixed wireless. Private wireless will enable enterprises to implement their own networks with well-defined quality of service and security using the same standards as the carriers with licensed or unlicensed spectrum. Fixed wireless access points will enable last-mile service to a wide geographic range with the throughput and latency of 5G.

With the industry laying the foundation for growth and interoperability, 5G will proliferate in 2021 to create efficiencies in Industry 4.0 use cases like manufacturing, healthcare, media and entertainment, and smart cities. 5G combined with edge computing will enable robots, drones, and connected devices in various deployments with local infrastructure for data processing. Remote locations will host edge-optimized servers to bring decision-making closer to the end devices. Content delivery and inferencing will branch out from large regional data centers to more distributed locations.

To support these new business opportunities, telcos and system vendors are building labs to enable industry innovation and new business models for monetization. This will enable an emergence of new startups focused on industry verticals and low-latency data processing at the edge, utilizing new capabilities such as network slicing and traffic steering with control and user plane separation across the 5G network. Additionally, the edge requirements for form factors and operating conditions will drive additional changes into the servers used for access network and edge data processing. Systems management will be enhanced to accommodate the scale of many distributed installations. Infrastructure provisioning, configuration, monitoring, and updates will be performed across these locations and regional data centers.

5G will begin to deliver its real promise in 2021, evolving from connectivity of users to always-on, self-optimizing connectivity of devices for mission-critical deployments. A common 5G architecture across macro-cells and private wireless, integration with edge compute infrastructure, and AI will enable a marketplace of applications running on the 5G edge.

Dell is focused on enabling software-defined 5G across the RAN, next generation core (NGC), and edge computing at on-premise and distributed edge cloud locations. We are also working closely with partners and the industry ecosystem to bring solutions to market. Expect reference architectures and new products to come in 2021.

7 Rethinking Memory and Storage to be Data Centric

Authors: Stuart Berke (Fellow), Bill Lynn (Sr. Distinguished Eng.)

Traditional information system architectures are based on a compute-centric mindset. Traditionally, applications were installed on a node, kept relatively static, updated infrequently, and utilized a fixed set of compute, storage, and networking elements to cope with a relatively small set of structured data. Over the past decade, data growth, particularly unstructured data growth, put new pressures on organizations, information architectures, and datacenter infrastructure. Today 80% of new data is unstructured and spread over large numbers of servers and storage nodes.²

Existing architectures are not designed to address service requirements at petabyte scale and beyond without significant performance limits. Traditional architectures fail to fully store, retrieve, move, and utilize that data due to limitations of hardware infrastructure as well as compute-centric systems' design, development, and management. And traditional CPU-based computing is unable to fully utilize the performance capabilities of a vast array of new optimized workload-specific accelerators.

Data-centric computing is an approach that merges new hardware and software architectures to treat data as the permanent source of value. New system interfaces and data centric fabrics standards such as Compute Express Link (CXL), Gen-Z, NVMe over Fabric (NVMeoF) are coming which will allow data to be "central" to the system, processed and transformed as needed by a variety of fabric-accessible computing elements such as CPUs, accelerators, networking, and storage engines.

² Carmen DeCouto, Understanding Structured and Unstructured Data, April 2020.

Storage class memory (SCM), a new class of persistent media at performance close to DRAM, enables new power-loss protected stores for data and metadata within server and storage nodes, or on the fabric itself. These SCMs may be standalone or utilized within memory and storage tiers and include new data-centric models for memory encryption, resilience, security, and accessibility.

Dell has taken leadership positions across the industry consortiums driving the foundations and specifications for data-centric architectures. We are also partnering with suppliers to ensure the entire ecosystem is in place to realize novel data-centric systems that provide better data analysis, security, scale and connectivity.

8 Adopting new server technology while being remote

Author: Ty Schmitt (Fellow)

Consumers and providers must continue to adapt to enable remote workforces to be effective. While 2020 changed much of our operational lives, many companies were well on their way to the digital transformation that allowed for quick adoption of a completely remote workforce. At Dell, we have looked at the remote workforce as an example of an edge ecosystem. It's distributed, decentralized, and spread across the world ... a connected extension of core and cloud. Remote workforce ecosystems are constraint driven and must adapt to change quickly. These constraints and changes are determined by consumption and usage models. Data is the core of this ecosystem and must be captured, processed, managed, moved, and accessed in a fast, flexible, and secure way. With growth in 5G, the industry will be fueled by faster, larger data pipelines which will enable 5G connected technology to more effectively monetize data. Automation and cloud/edge native application development will accelerate and with this the need to deliver software faster and hardware to be more flexible based on unknown changes in workload requirements.

The data volatility, volume, and variety, along with the speed of technology and the transformation it will enable, comes with ambiguity and chaos. It's a forcing function that demands multiple layers of the ecosystem to be understood, connected, and solved for simultaneously. It will demand flexible and adaptive IT, data/operational management, datacenter infrastructure, and associated delivery and financial consumption model solutions to maintain business continuity in dynamic environments.

Whether they are at a retail store, in a vehicle, at an industrial plant, at the base of a cell tower, or in a home office, new workloads will demand an adaptive approach to provide flexible and optimized solutions. Dell Technologies understands this ecosystem ambiguity and connects with its customers and partners to understand and solve for edge. Our approach enables faster, more flexible, and cost-effective digital transformation by utilizing common frameworks for management, analytics and service to seamlessly connect edge-core-cloud.

9 It's not a CPU competition, it's a recipe bake-off

Author: Stuart Berke (Fellow)

We are at the start of an exceptionally exciting era of competition in computation, across CPUs and accelerators. Multiple CPU suppliers spanning x86 and ARM architectures provide industry leadership in one or more segments and workloads. Equal accessibility to leading edge silicon process fabrication capability has provided a backdrop of instruction set and architectural innovation in order to differentiate at the hardware level. Historically, this has created a massive competition between CPU vendors based on cores, frequency, cache sizes, memory channels, optimized instruction sets, etc. Although the CPU competition is still fierce even with workload-optimized instruction sets and other incremental features, CPU generational performance scaling is slowing, setting the stage for a vibrant and varied hardware acceleration ecosystem to emerge. Specialized accelerators for fixed- and floating-point vector and matrix operations and those supporting new AI/ML tensor operations are scaling at 2x or more performance per generation.

Accelerators for networking, storage, security and crypto, encryption and compression, and many others are now readily available stand alone, or in some case integrated into CPUs and GPUs.

Established and emerging software, frameworks, and ecosystems are competing to provide users with simple yet powerful tools in order to take advantage of the increasing complex hardware elements. These include libraries that abstract away the CPU, GPU, and accelerator intricacies, while providing high efficiency of computing resources. This has resulted in many of these traditional compute suppliers expanding their portfolio of IP to include not just traditional compute silicon, but also IP around specific hardware acceleration and the SW libraries used to optimally use them in application environments. Each vendor is trying to lessen the customer burden of finding the right mix of technology and supporting SW by creating a “better together” integrated set of their own traditional compute, targeted acceleration, and performance optimized software – creating a technology recipe from their own portfolio of ingredients.

But an even higher level of integrated offerings is needed to provide customers with fully optimized solutions for their businesses. Customer-winning differentiation will move from the hardware and software to highly integrated workload-and segment-specific solutions such as those tailored to telco and 5G, HPC, database and analytics, cloud and hosting, and other segments. This is where Dell provides optimal integrated platforms and offerings based on industry suppliers along with the workload characterization analysis and, in some cases, the full solution stack. So, as we all continue to watch how the CPU landscape evolves, the more interesting competition will be how the various suppliers bring together their portfolio of IP in the most optimized and easy to integrate technology recipe.

10 Measure your IT Trust Index

Author: Mukund Khatri (Fellow)

While many had predicted 2020 to be a year of rapid acceleration of enterprises embracing digital transformation, no one could have dreamt of the catalyst the pandemic would bring. Almost overnight, organizations were forced to transition their workforce to telework while enabling massive migration of services to cloud. As feared, this was accompanied by fragility visible through the security lens and serving a new haven for bad actors.

As we look to this year and beyond, we will see fundamental shifts in security posture by enterprises and customers. There will certainly be a broader mandate for built-in resilience against a growing number of advanced persistent threats along with innovative and automated options to stay current with patch management tools across all aspects of infrastructure deployments. As hybrid cloud rapidly extends to the edge, bringing compute close to where data is being generated, we will see innovations leveraging new technologies like AI, ML and principles of zero trust across various security solution spaces, including supply chain risk management, advanced threat monitoring and management solutions, and enhanced identity and access management. Increasing adoption of newer technologies like SmartNICs and other accelerators in modern architectures will facilitate new sets of intrinsic and resilient security solutions. A rapid rise in awareness and visible activities tied to quantum computing along with need for quantum safe crypto algorithms will also intensify towards protection of critical assets and data that will be accessed well into the next decade.

Going forward, supply chain security considerations will be one of the top criteria driving purchase decisions for IT devices. Transparency in supply chain of all infrastructure components, inclusive of hardware and software, around how products are developed, delivered, deployed and managed will be critical to all enterprises.

In the backdrop of increasing regulatory and compliance frameworks, a need to measure integrity and trustworthiness of both infrastructure devices and data for effective identification of digital risks is going to be increasingly critical, not only during deployment but across the end-to-end lifecycle. Enhanced solutions such as the Dell Secured Component Verification that enable access to provenance for new computing devices will need to be broadly deployed across the broader supply chain ecosystem. Real-time observability of the state of integrity and compliance drift built on current cryptography standards for all layers of the solution stack will be a key mandate for risk-based trusted lifecycle management and operations. IT administrators can be expected to increasingly adopt products and offerings that effectively enable these capabilities across on-prem, cloud and edge environments.

Transparency of the state of IT infrastructure will be a crucial mandate of this data decade. Additionally, a company's ability to quantify their security or Trust Index across their edge--core-cloud infrastructure and supply chain will provide them better awareness of possible threat vectors and data risk. At Dell, we work closely with our customers to help them architect and implement strong IT security from delivery, to deployment and through operations and decommission to ensure the highest confidence in their Trust Index.

Discover more about PowerEdge servers



[Learn more](#) about our PowerEdge servers



[Learn more](#) about Dell Technologies Services



[Follow](#) PowerEdge servers on Twitter



Contact a Dell Technologies Expert for [Sales](#) or [Support](#)