A Perspective on the Emerging Technology Landscape for Enterprise Decision Makers

Nine Critical Technology Inflections for Enterprises to Understand while Digitally Transforming their Business

Author:
John Roese, Global Chief Technology Officer Products & Operations, Dell Technologies

Contents

Introduction

The enterprise information technology environment is changing dramatically with transformation occurring at many different layers across the information technology stack. New technologies are emerging at each of these different layers that collectively predict what the future enterprise is going to look like within the next three to five years.

Conceptualized as data, cloud, and infrastructure technology inflections – these technology inflection points interact to create the connectivity, performance, capacity, orchestration, and insights that support the data-driven enterprise and digital transformation.

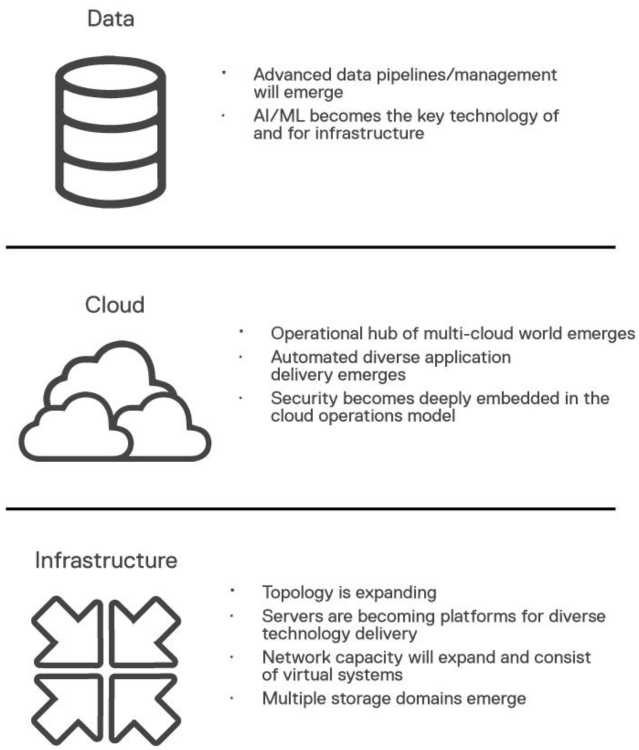

Diagram 1. Major data, cloud, and infrastructure technology inflections

This paper outlines nine critical technology inflections that are important for enterprises to comprehend as they make technology choices that enact their digital future. Understanding the interaction between these inflections is essential to making aligned decisions that provide the infrastructure, cloud, and data foundation for digital transformation.

Topology is expanding

The enterprise topologies of today are expanding to include technology elements that have not traditionally been thought of as constituents of an enterprise data center.

Going forward, we believe every large enterprise will have a topology similar to that depicted in Diagram 2, including public clouds, enterprise data centers, branch offices, remote manufacturing/R&D centers, SaaS environments, mobile users, and cloud data colocation facilities. These environments will be connected as part of a single topology, with unique applications, data, and infrastructure deployed to each environment, and they will operate seamlessly together.

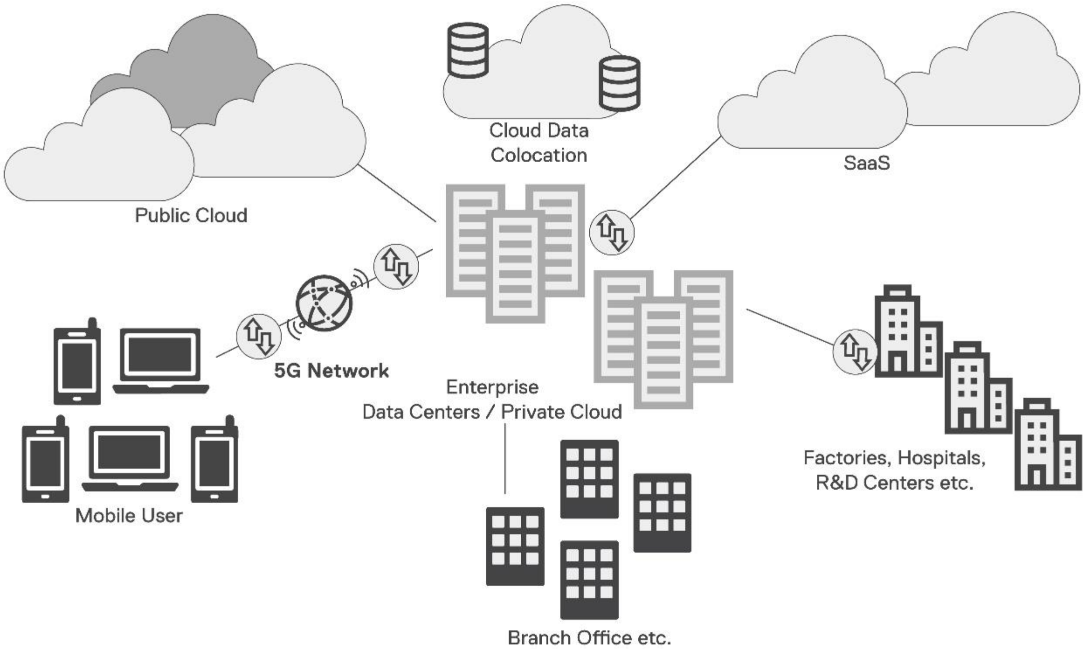

Diagram 2. A modern enterprise topology

Within this topology, we have identified three emerging technology trends that have not traditionally been considered part of the enterprise. All of these connect users, applications, and data processing facilities to data, for the purposes of supporting the digital enterprise:

1. 5G cellular networks

   4G provided faster connectivity, but 5G moves beyond connectivity by providing the capability to move compute, applications and data processing into the cellular environment itself. This is enabled through features such as very low latency, massive wireless bandwidth, network slicing, and quality-of-service (QoS) controls. 5G will provide connectivity across the enterprise topology, connecting data centers to large sites such as factories and R&D centers. This has radical implications for the ability to create responsive, low-latency data processing frameworks.

2. Cloud data co-location

   Cloud data co-location solves a specific problem for enterprise organizations by providing a neutral interconnection between one or more public clouds and private enterprise data centers for the purposes of making enterprise data accessible to multiple clouds. Co-location provides organizations with a way to retain control of their data and the environment it is hosted in – supporting data governance and data sovereignty – while situating the data in an environment that delivers optimal performance and accessibility for cloud or enterprise workloads.

3. Edge computing[1]

   As enterprises move into the era of machine learning, artificial intelligence, real-time analytics, and high-performance telemetry management, physical and operational constraints make it desirable to locate data and processing close to the consumers of that data (users and systems). Cost is one such concern, because it can be cost prohibitive to move large amounts of data between clouds and across disparate networks for more centralized processing.

These considerations have resulted in the proliferation of edge compute environments, with at least four edges emerging:

- **The mobile user edge** in the cellular site itself – bringing an aggregate set of compute and processing capability into the mobile experience
- **The enterprise (private) data center edge** – close to systems of record and other aggregate data sources
- **The large site edge (factories, R&D centers, etc.)** – hosting intelligent devices that generate massive amounts of data
- **The public cloud edge (in private data centers)** – giving public clouds the ability to support real-time behavior through vehicles like Amazon Outpost

Enterprises will need to devise an edge computing strategy and implement technology that distributes compute – in the right form factors – close to where data consumers can generate real-time outcomes for the business, leveraging model-based inferencing, for example.

[1] Edge computing is a distributed computing paradigm which brings computation and data storage closer to the location where it is needed to improve response times and save bandwidth.

Servers are becoming platforms for diverse technology delivery

Compute is rapidly evolving to become a platform of advanced technology processing capacity for new and emerging workloads (including Artificial Intelligence/Machine Learning (AI/ML) workloads). In this world, processing capabilities are no longer held captive to the notion of a traditional server, and new and diverse form factors support the delivery of accelerators, along with new media and memory technologies.

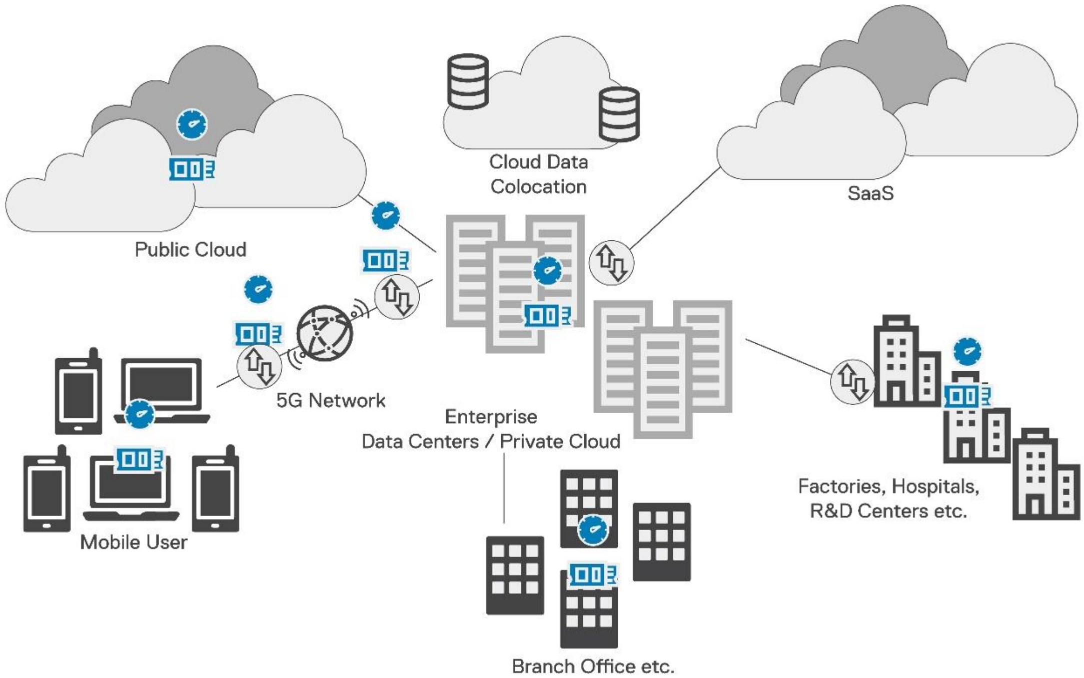

Diagram 3. Distribution of accelerators and advanced media

First and foremost, dense compute acceleration platforms are being created across the enterprise environment with the following form factors:

- Dense acceleration platforms – Enterprises are building clusters of dense acceleration platforms that can land lots of accelerators with a minimal set of x86 cores
- Dense acceleration clouds – Cloud providers are building dense acceleration clouds (using Tensorflow Processing Units (TPUs), for example)
- Edge compute platforms – Enterprise edge compute modules are being built that have the right form factor for edge environments (for example, for a cell site or modular data center). These can be populated to:
  - Near edge compute environments – in which accelerators are populated to micro data centers
  - Far edge compute environments – where accelerators are made accessible via radio access networks

These compute form factors are being created to address the interplay between edge computing and dense acceleration, where compute provides not just x86 technology, but x86 plus a range of accelerators that are biased towards specific types of training and inferencing. These accelerators include GPUs, FPGAs, TPUs, specialized AI processors (such as Graphcore IPUs), and neuromorphic processors.

Advanced media is also part of this picture. Non-volatile memory potentially speeds performance for computationally-intensive workloads (such as deep neural networks and classification algorithms), while the latest flash technology supports high performance (I/Os per second), low latency (microsecond) and high throughput (GB/s) for these types of workloads.

In short, servers are no longer just a delivery vehicle for x86 cores – they are now a delivery vehicle for this diverse set of technology componentry that can be packaged and delivered to enterprises to support heterogenous processing outcomes. When these technology components are put together in the right way and exposed to applications and algorithms, this can lead to orders of magnitude improved performance and time to results for businesses.

Network capacity will expand and consist of virtual systems

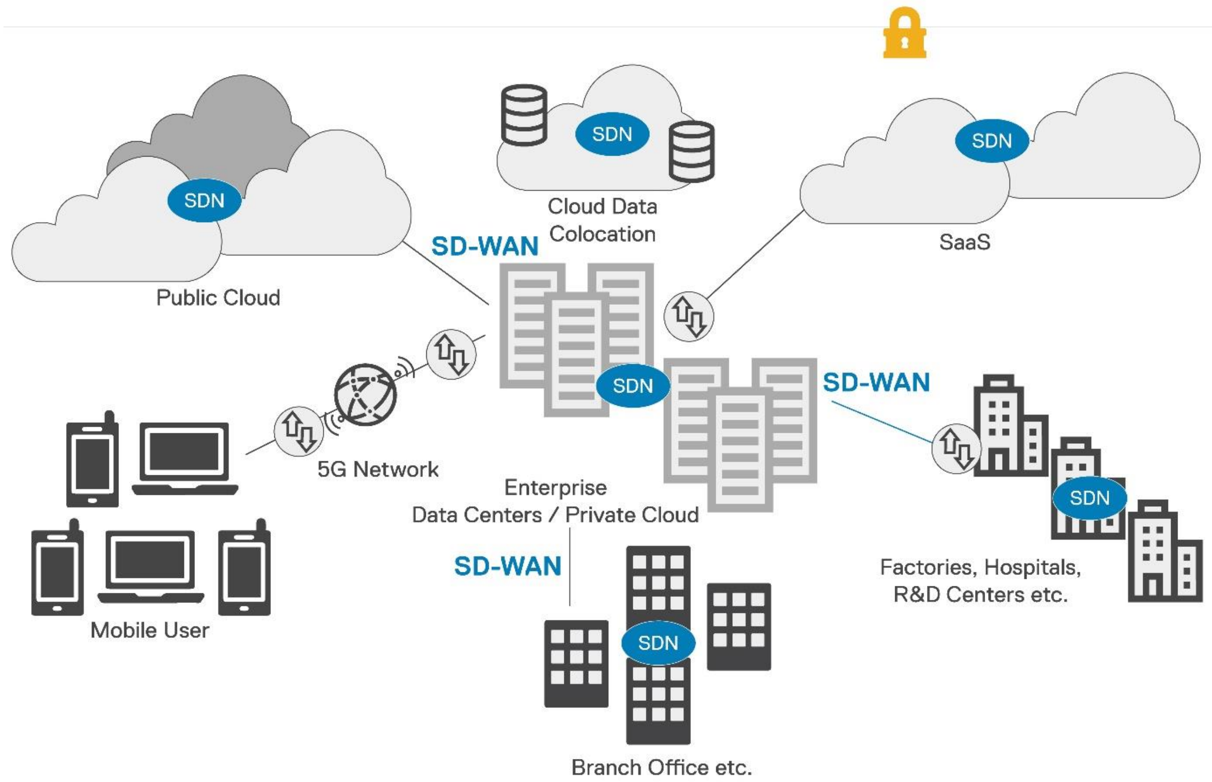

Diagram 4. SDN and SD-WAN in the modern enterprise networking environment

We expect significant changes that will define modern enterprise networking over the next 5 years. Most significantly, the mobile networking environment is going to change with the evolution towards 5G. While 4G provided broadband over the air, 5G will provide faster broadband, low latency connectivity, and an abundance of bandwidth, removing a lot of the bottlenecks for mobile users to services provided in public or private clouds. In this world, 5G is not just in the realm of service providers, but becomes a logical extension of enterprise connectivity and a part of overall enterprise architecture.

Perhaps the most important outcome that 5G will deliver is the ability to allocate network slices, where enterprises will be able to dictate the quality of experience (including latency and bandwidth) for a specific application running on a particular mobile network somewhere in the world. This will allow enterprise policy decisions and a consistent service level experience to be enforced all the way out to the mobile user, making 5G an integral part of the enterprise application experience.

In addition to the shift to 5G, data center fabrics are going to go through a revolution. Today's fabrics are primarily legacy in nature – consisting of Fiber Channel, InfiniBand, and Ethernet-based fabrics, but next generation fabrics will materialize, facilitated by industry consortium activity. Fabrics such as Gen-Z, OpenCAPI, and CXL are good examples of this. These next generation fabrics will deliver the same advantages as 5G, delivering an abundance of bandwidth to enterprise and public cloud data centers for the purposes of interconnecting devices. This will enable the convergence of memory, storage, fabric interconnects and network fabrics into sophisticated compute and fabric architectures. These next generation fabrics are expected to emerge over the next 5-10 years.

Advanced optical innovation will continue providing more capacity over fiber optic infrastructure, whether in the data center or long-haul environments, enabling the hybrid and multi-cloud experience. There will be more bandwidth within mobile environments, within data centers and interconnecting data centers and clouds, but this bandwidth is not going to be infinite and is always at risk of being overwhelmed by the massive amounts of data expected to be delivered via this infrastructure. For this reason, edge computing will be leveraged to push compute closer to where real-time outcomes are delivered, and away from the long distance reaches that are necessary today. Bandwidth will continue to be a reasonably scarce commodity, even as it becomes more abundant.

To support multi-cloud and multi-datacenter connectivity, the industry is rapidly separating the hardware and software experiences of networking, so that networks can be managed more intelligently. The shift to software-defined networks and software-defined WANs makes the control plane – which defines the logical connectivity and behavior of the network – into a first-class citizen in the cloud technology stack. This transforms the network from being just a pipe to an environment where decisions about packet routing, access, and consumption are no longer a network decision, but instead an application and cloud decision executed through SDNs and SD-WANs. Ultimately, this capability provides the opportunity to optimize bandwidth utilization by ensuring that the right amount of bandwidth is allocated to the right applications in the right manner.

This is a significant shift because today's networks operate for the most part as an independent layer that provides bulk bandwidth capacity to upstream environments without understanding their intention. Going forward, network behavior will become much more tightly coupled with the overall cloud stack and will be driven by the same cloud orchestration capabilities that coordinate infrastructure resources, workload placement, data protection, disaster recovery, and other aspects of application delivery today.

Multiple Storage Domains Emerge

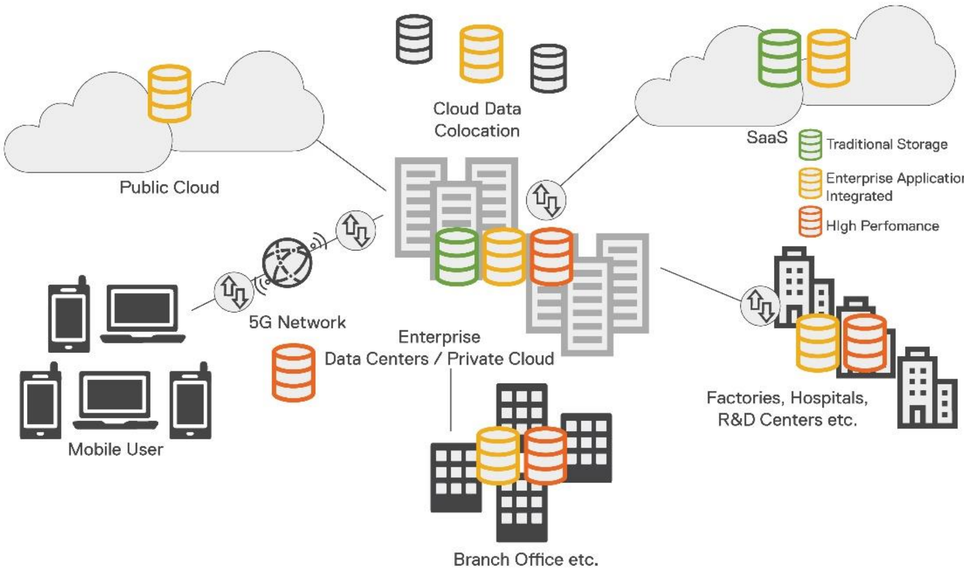

Diagram 5. A modern storage environment

Storage and data protection will continue to be incredibly strategic as organizations transition to the data era. Data is the fuel of modern enterprises, and it will be created, maintained, consumed, persisted, and protected in a diverse set of environments. In fact, there are at least 3 distinct environments in which data will be stored and protected going forward:

1. Traditional Storage – First, the traditional environments or enterprise systems of record including mainframes, open systems and databases, need a hardware-resilient storage architecture to run on. These systems are typically based on a single vendor platform and are not going away, so the objective is to abstract them via an API so that they are consumable by new applications. They also need to be optimized and made operationally efficient.

2. Enterprise application-integrated storage – This type of storage, much like networking, is being integrated into cloud orchestration workflows and is rapidly becoming part of the cloud stack. Multiple ways exist to integrate this type of storage to support the desired performance, scale, and functionality:
   – Hyperconverged infrastructure – in which the HCI system as designed as a turnkey hardware/software solution comprised of compute, storage, networking, data protection, virtualization, and application management.
   – Storage arrays – storage arrays that can be integrated into cloud stacks such as midrange or high-end storage platforms
   – Public cloud storage – storage services provided by the public clouds, such as Amazon EBS and S3, as a component of a cloud-orchestrated environment. This also picks up co-located cloud storage services that are consumed through a public cloud API.
   – Software-defined storage – software-defined storage architectures implemented as software layers in an overarching cloud stack.

What is consistent across these different types of storage – even though they provide different levels of performance, scale, and functionality – is that they are each, in effect, subordinated to a particular cloud architecture, and get connected to enterprise workloads through the orchestration of compute, network, and storage at the cloud layer.

3. High-performance compute/storage – the third category of storage is a new and emerging class of advanced analytics (AI/ML) performance compute/storage. This type of storage is attached over high-performance networks (10G+) and needs to support the consumption of gigabits per second of unstructured data by 1000's of accelerators at petabyte- and exa-scale. The performance expectations for this type of storage are significant.

Each of these types of storage solve slightly different problems, but most modern enterprises will need to develop strategies for all three, albeit at different locations in the overall enterprise topology. For example, enterprises will likely use high performance storage at the edge or enterprise data center for AI/ML, and application integrated storage in the cloud or enterprise data center for modern applications

The operational hub of the multi-cloud world emerges

To manage this diverse infrastructure across enterprise data centers, edge environments, hybrid clouds and multi-cloud environments – the diverse topology previously depicted – enterprises will need an operational hub that makes these diverse environments work as a system.

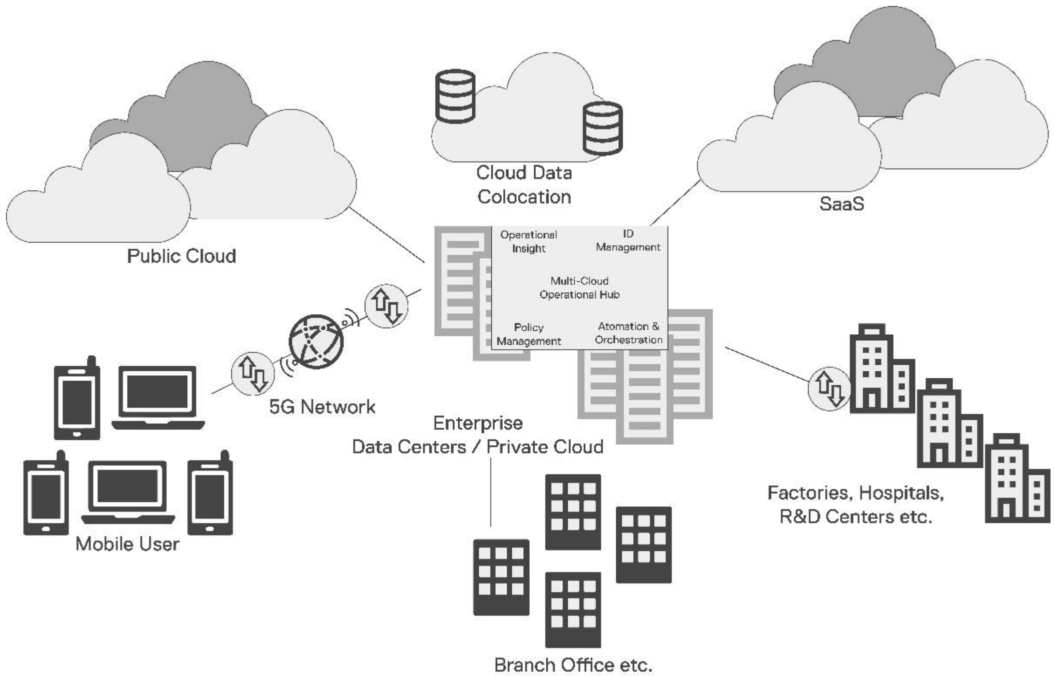

Diagram 6. A modern enterprise topology

Each of these environments needs to be governed by an authoritative center – an operational hub – that can orchestrate this collection of clouds into a coherent system. The results of having a single operational hub are extremely important – if a single operational experience can be created across these clouds, then it becomes possible to have consistent policy management, service-level orchestration, identity management and authentication, for example, by orchestrating behaviors across the entire operational environment. This entire experience can then be wrapped by intent-based interfaces that support the expression of application outcomes (such as service levels) while orchestrating underlying components to achieve that outcome. This not only becomes foundational for the reduction of operational complexity, but also facilitates self-service capabilities for developers and application owners.

The operational hub will need to provide insights across the entire environment, for the purposes of costing different cloud resources, comprehending the relative health and performance of different cloud environments, or monitoring the compliance of different clouds relative to pre-established SLAs. With a single pane of glass supporting the entire multi-cloud environment, subscriber management and billing can be done in a more comprehensive, coherent manner.

The operational hub is one of the missing links for creating a true multi-cloud, end-to-end management experience. Vendors that are already prominent in data centers across all of the environments described above, and end-to-end from infrastructure to applications, are strategically well-positioned to deliver on the promise of the single operational hub.

Automated diverse application delivery emerges

Modern applications will consume this diverse infrastructure and represent a significant shift in how enterprises will build, deploy, and run applications going forward.

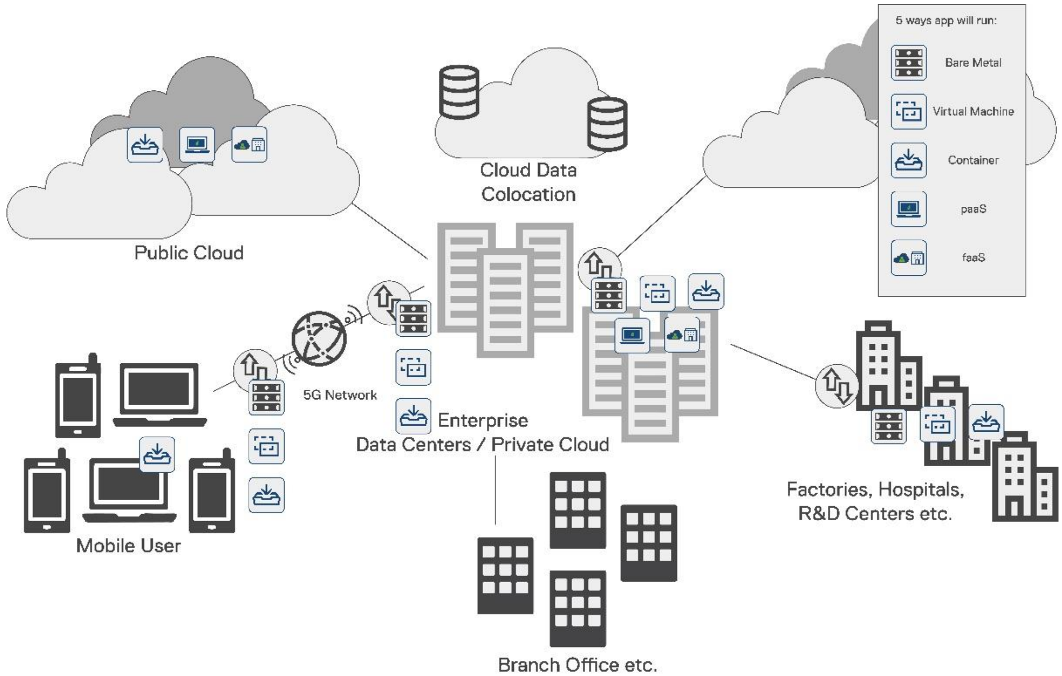

Diagram 7. Diversity of modern application runtime environments

Traditionally, application developers have written applications, installed them to an operating system or virtual machine, and then run them in a bare metal or virtualized runtime environment. Over the last several years, we've seen a diversification of code runtimes emerge. This reflects the diversity of application code itself – some kinds of code will be optimized for performance, some need to be highly automated, some code is highly distributed, some requires a lot of infrastructure resources for optimization and resiliency, and some is very fluid and can recover without a lot of hardware dependencies. In other words, adopting this philosophy allows application code to be matched to a runtime that best suits its deployment and runtime requirements.

Associated with this, one significant shift is that code will now be built without an intention of where it is going to be run, post-development. Most enterprises are now leveraging modern design patterns such as micro services or are building functions for later deployment and are storing these services or functions in repositories for later deployment. Enterprise developers are now building code for enterprise software libraries and storing it in containers and repositories for later deployment. Once built, the code might be deployed to different environments in a number of ways:

- Bare Metal – Bare metal high performance apps will run in enterprise data centers and at the mobile user edge, --for example, AI/ML tasks that need to run close to hardware.
- Virtual machines – VMs will run everywhere – clouds, data centers, the enterprise edge, and in the 5G near edge. They provide a stable, predictable substrate for applications.
- Containers – Containers will run everywhere – in public clouds, edge devices, mobile user devices, branch offices as containerized runtimes.

- Platform-as-a-Service – PaaS applications run in the public cloud and enterprise data centers delivering composite applications as one system that can be dynamically scaled, instrumented, and curated.
- Function-as-a-Service (Serverless) – FaaS is likely to run in the enterprise data center and in edge computing environments. Functions are particularly suited for building data processing pipelines.

Containers have become ubiquitous and a common denominator across all environments because they support the ability to run code in a lightweight model. They tend to support more ephemeral and fluid applications, in which the ability to continuously and dynamically deploy and scale applications (via orchestrators like Kubernetes), is a significant advantage. Many organizations choose to run containers in virtual machines so that they can benefit from a common virtual hardware substrate with the container expression and orchestration logic that allows code to be built and deployed in a more automated fashion.

The decision about which application delivery and code execution environment to use to run the code in a multi-cloud environment will be complex and will likely be orchestrated with AI-driven orchestration providing the connective tissue across the entire ecosystem.

Security becomes deeply embedded in the cloud operations model

As application delivery models proliferate and topology expands, security will be embedded in the cloud operations model and in the traffic flow of software-defined networks (SDNs).

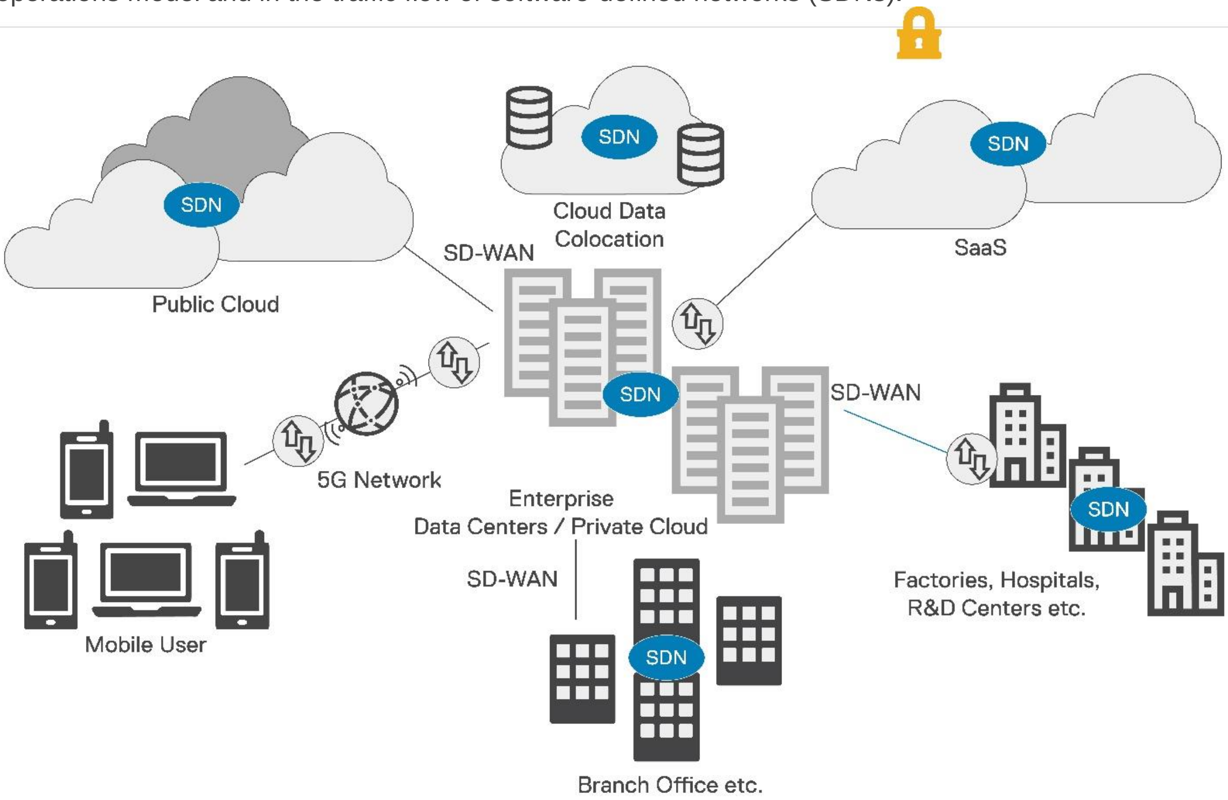

Diagram 8. Intrinsic security

Today, security is largely enacted through the deployment of point systems bolted onto solutions within an infrastructure--for example, through the deployment of firewalls. This model has been challenged by the scale and diversity of the current infrastructure environments. In the future, security functions will be embedded as logical functions within the overall cloud topology and cloud activity, with cloud orchestrators coordinating security rules and behaviors as part of a larger orchestrated application and infrastructure deployment scenario.

In this world, security functions will manifest as software functions that get deployed from a repository at the same time the application is deployed, either to a containerized environment or to a platform-as-a-service environment (PaaS). In this case, one of the micro services that gets deployed in a multi-cloud environment might be a code inspection service, or a service that performs deep packet inspection, or a cryptographic function that is part of an overall orchestrated application.

SDN is part of this picture because it supports the ability to perform micro-segmentation, confining which consumers are able to access the compute environment. This security model is likely to become more pervasive. As security gets embedded as functions in an overall traffic flow – distributed across a diverse topology – it is the orchestrated SDN-based cloud environment that becomes the platform for security, and not individual products.

Quantum computing is another technology that is garnering attention recently in the security space. Quantum has the ability to run algorithms that cannot be run on traditional computing architectures – for example, algorithms that can be used to break traditional cryptographic algorithms, such as Shor's algorithm[2]. This has implications for the security of the key management schemes that are used to secure much of internet communications around the globe, such as RSA. To address this risk, the industry has taken steps – in conjunction with NIST and other standards bodies – to rethink key management in a post-quantum world. It is likely that in the next several years, a secondary key management protocol will be made available to systems to render them safe in the post-quantum era.

[2] Shor's algorithm is a quantum computer algorithm for integer factorization that was invented in 1994 by mathematician Peter Shor. On a quantum computer, Shor's algorithm can potentially be used to break cryptographic algorithms that are based on the assumption that factoring large integers is computationally intractable.

Advanced data pipelines/management will emerge

Data is the fuel of the modern enterprise. There is no AI/ML, no advanced analytics, no self-driving enterprise without a vast amount of data that can be successfully consumed, processed, and used to influence outcomes.

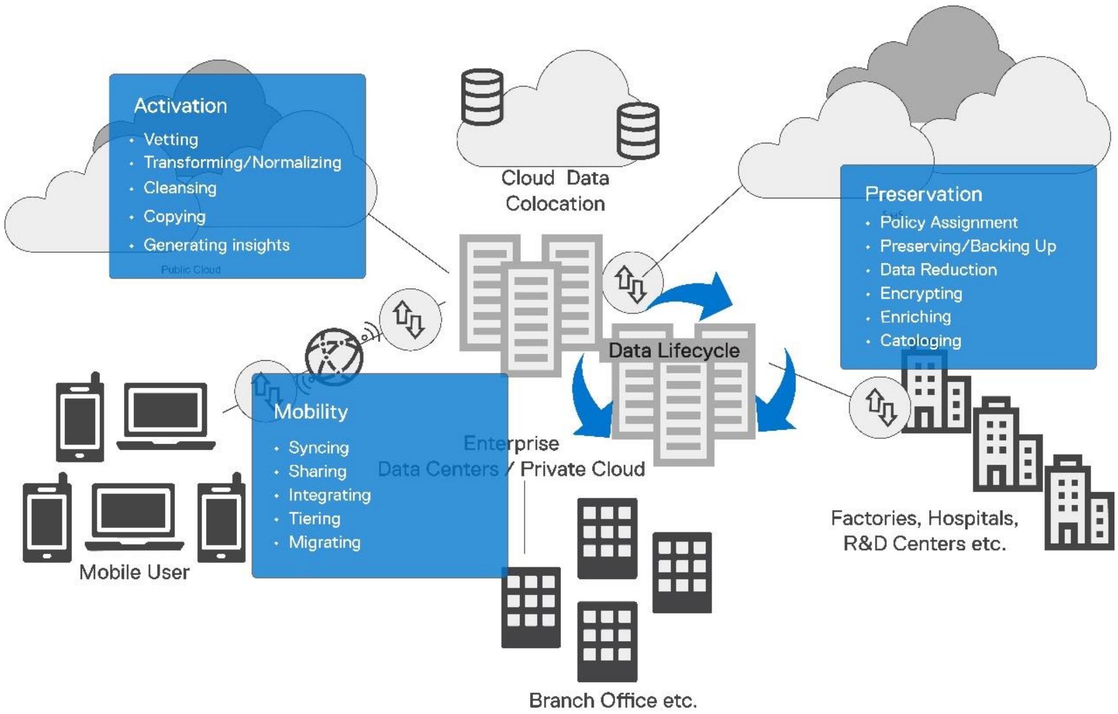

Diagram 9. The data lifecycle occurs in all aspects of the enterprise environment

The data pipelines necessary to move data from source to impact are much more sophisticated than traditional enterprise data sources. Analytics and data science at enterprise-wide scale require iterative and rapid access to integrated, high-quality data. Two parallel paradigms have matured to meet this need: the traditional extract/transform/load (ETL) + data warehouse + business intelligence (BI) approach, which is typically an IT-driven process, and the newer self-service data preparation + visualization tool (or data science tool) approach, which meets the needs of business users. Although ETL and data warehouses long predate the self-service ecosystem of tools, most organizations still need both approaches. However, because ETL has traditionally been batch-oriented and highly dependent on technical skill, there is a push for more agile and managed capabilities that allow for more responsive leverage of data. Widespread adoption of enterprise cloud infrastructure has made this increasingly possible and has put pressure on organizations to modernize their ETL workflow. Future scenarios will accelerate this transition, like the autonomous vehicle world where hundreds of millions of connected vehicles, with thousands of sensors, are each producing zettabytes of data, flowing across a global infrastructure. The ultimate goal of these modern data pipelines is to solve complex problems for industries such as the autonomous vehicle industry, for example, by supporting the creation of intelligent point maps that describe the physical topology of every road in the world, with the objective of influencing traffic patterns and fleet efficiency.

The answer to creating these kind of data pipelines doesn't lie in a single product. Vendors will need to create next generation data pipeline tools that support bringing together disparate cloud and enterprise data sources in a way that synchronizes them and coordinates them, transforming the semantics of the data as it is stored so that it is accessible in a consistent way across environments. These tools will need to solve specific problems such as discovery, visualization, and conditioning of data from different data sources. In attempting to become more data driven, many organizations are investing in machine learning (ML) tools and developing ML-driven applications. The success of these projects depends on the ability of the organization to operationalize experimental data science projects through model deployment and management. This is important not just from a regulatory perspective, but also to increase data pipeline performance. Without data visibility, for example, it is difficult to organize data so that it is in the right place at the right time.

In summary, enterprises will need to develop strategies for transitioning from static sources of data to dynamic, automated AI-driven impact of a collection of data that influences the outcomes of their business. The reason to build infrastructure therefore, shifts from being purely about cloud consumption models to creating an infrastructure that can support data pipelines and be consumed by AI/ML functions driving desired business outcomes.

AI/ML becomes the key technology of and for infrastructure

Today, the line between the tasks that are accomplished by human beings and those that are accomplished by technology is relatively low, and almost all thinking and cognitive tasks are done by human beings. Even in the analytics world, the primary activity is to organize data so that a human being can look at the data in a dashboard, derive insights, and make a decision. There are relatively few cognitive tasks that are performed by infrastructure today. Moving deeper into the AI era, this will change.

Three significant impacts will happen as we transition into the AI era. The first is that the industry will create many more smart products. Today, most products are passive unless instructed to perform a task by a human or by an application that sends them a command. By adding AI/ML to infrastructure we can deliver new capabilities and outcomes. One example of this is adding advanced machine learning algorithms to high-end storage products to achieve improvements in cache performance by improving the possibility of a cache hit. This can be done by leveraging AI to avoid adding hardware, changing the physical architecture, or adding cost.

In the future, potential evaluation criteria for smart products could include answers to questions such as:

- How many tasks can the smart product perform without human intervention?
- How often can the product operate and tune itself so that human beings don't have to be involved?
- What is the intended outcome of the tuning, optimization, and automation?
- Will the embedded intelligence dramatically change the performance, economics, and functionality of the system?

In the future, all product decisions will be based not just on price, performance, and feature set, but will also on how intelligent the product is in performing tasks autonomously.

The second significant area of impact for AI/ML is in absorbing most of the IT operational burden. Looking at industries like the autonomous vehicle industry, it is clear that some environments will generate zettabytes of data in their infrastructure to manage their autonomous vehicle fleet. The bigger problem with this volume of data is not just the cost of storing that data and processing it, it is scaling the operational burden associated with it. For example, consider the ratios used to determine how many storage administrators you need to manage petabytes of data in the storage industry, which today is roughly one storage administrator to ten petabytes of data. Extrapolated to zettabyte-scale, a million storage administrators would need to be on staff to run a zetta-scale environment.

Solving this scaling problem requires the ability to have one person expressing intent to an AI system that manages exabytes of data – essentially shifting most of the operational burden into the infrastructure by leveraging cognitive capabilities, such as AI/ML. Most cloud environments are moving in this direction, as are most IT operations environments. This capability will ultimately support the delivery of very large infrastructures by shifting most of the operations burden into fully autonomous, AI/ML-driven infrastructures.

The third area of impact is the shift to AI/ML as a significant consumer of technology, services, and infrastructure. AI demands very specialized processing capabilities, is a big consumer of data, creates massive amounts of data, and influences outcomes. In this paradigm, infrastructure becomes a facility for delivering services to AI/ML systems. In the future – perhaps five to ten years from now – the majority of demand for infrastructure will be to serve AI/ML systems, as opposed to human beings, in the service of cognitive tasks that allow systems, business, industries, and our world to operate optimally.

In Summary

In summary, what is going to happen in the enterprise over the next five to 10 years revolves around the following key trends:

1. Topology is expanding. The modern enterprise is expanding, and more topological elements will materialize including at the edge, with 5G, and in the new cloud model, such as cloud data colocation. This trend will continue to accelerate.
2. Servers are becoming platforms for diverse technology delivery. New and diverse compute and memory models are going to provide more capabilities that will need to be landed in new compute delivery models. AI/ML and advanced analytics tasks will benefit significantly from the incorporation of diverse, accelerated compute into these platforms.
3. Network capacity will expand and consist of virtual systems. There will never be an infinite supply of bandwidth – even with new advances in fabric and optical technology – so there will be a huge shift towards cloud-orchestrated network environments that can make intelligent decisions about how to allocate network bandwidth based on application requirements.
4. Multiple storage domains will emerge. At least three different storage domains will emerge (traditional systems of record, cloud-integrated storage, high-performance next-generation storage environments). The existence of these domains will require more sophisticated and highly orchestrated approaches to storage and storage management.
5. The operational hub of the Multi-Cloud world emerges. The emerging multi-cloud world needs a common operational hub. There will be a collection of clouds – in private data centers, public clouds, and across the edge including wide area and 5G cellular networks – within the enterprise topology, but there needs to be a common orchestrated user experience across them.
6. Automated diverse application delivery emerges. A diverse set of application delivery models will evolve, and the development of application code will be separated from the decision about which runtime to execute it on. Orchestrators will need to incorporate intelligence that helps them understand how best to optimize the code by running it on the right runtime, on the right hardware.
7. Security becomes deeply embedded in the Cloud operations model. Security will be injected into the overall cloud orchestration framework as software functions that run in an SDN or SD-WAN environment. Much like networking, these software-defined security functions will run under the control of a multi-cloud control plane.
8. Advanced data pipelines/management will emerge. These will carry data from its source (intelligent devices and people) all the way to its destination (advanced AI/ML frameworks), where the data is distilled into new models, behaviors, and insights that influence real-world outcomes. Building next generation data pipelines is probably the most strategic problem enterprises will face over the next five to10-year horizon.
9. AI/ML becomes the key technology of and for infrastructure. AI/ML is the key technology for and of infrastructure. AI/ML will manifest itself as smart products that can operate autonomously; it will be used to shift the operational burden of running large-scale infrastructure to machines; and it will be an increasingly significant consumer of IT infrastructure.

Enterprises will need to be proficient in navigating all of these technology inflections, to create a technology infrastructure that optimizes their business for a digital future, supports great customer outcomes, and minimizes the risk of digital disruption by competitors.