An IDC White Paper, sponsored by IBM

IDC

Tri-Delta Resources

The Sweet Spot of Modern Enterprise Computing

Research by:

Peter Rutten
Research Director, Infrastructure Systems, Platforms and Technologies Group, Performance Intensive Computing Solutions
Global Research Lead, IDC

September 2021 | IDC Doc. #US48197921

Navigating this White Paper

Click on titles or page numbers to navigate to each section.

IDC Opinion 3

Situation Overview 4

- Security as an Imperative Requirement 4
- The Reliability Mandate 5
- The Need for Scalability and Sustainability 7
- The Right Hybrid IT Infrastructure 8
  - Gravitating Toward Hybrid Cloud 8
  - Hybrid Cloud and Cloud-Native Applications 10
- The Importance of AI and Where to Run It 10

IBM Power10 and IBM Power E1080 12

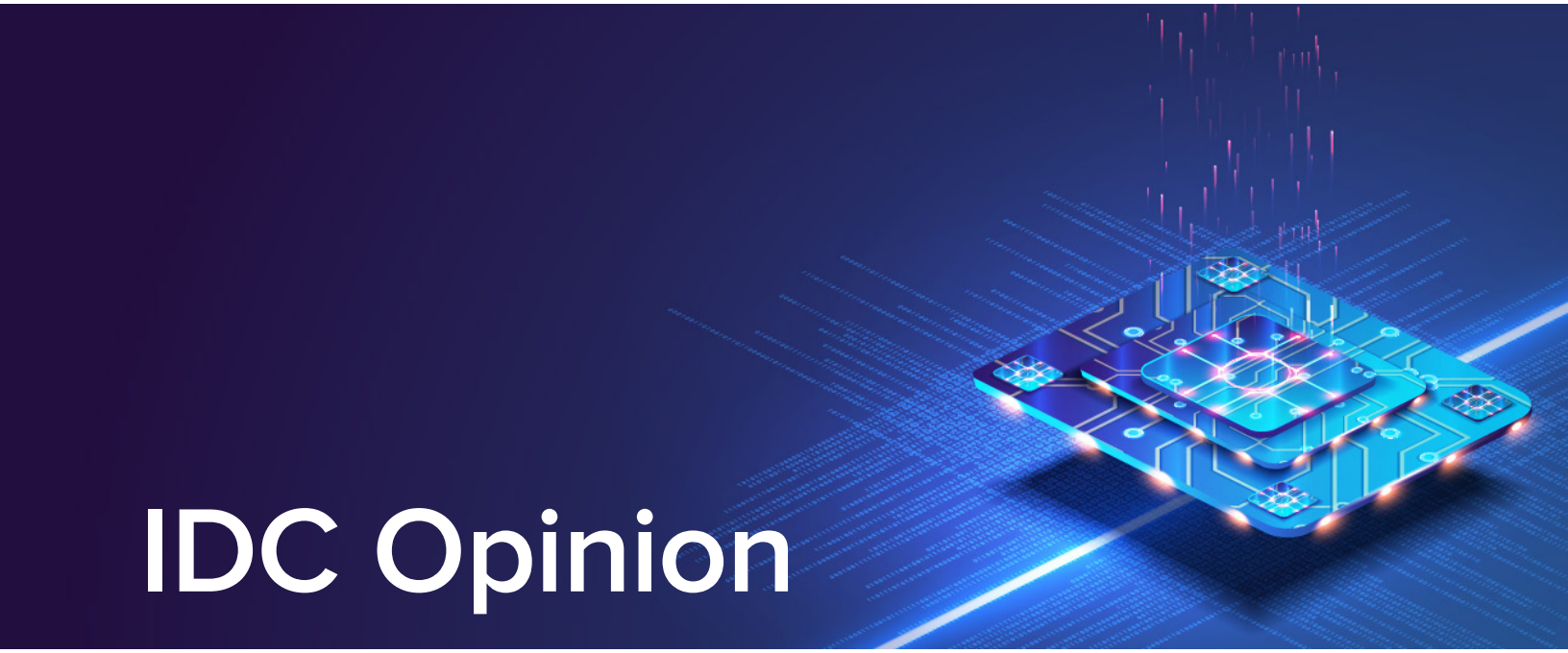
- The New Power10 Processor 12
- The IBM Power E1080 12
  - Security 12
  - Resiliency 13
  - Scalability and Sustainability 13
  - Hybrid Cloud 13
  - Artificial Intelligence 15

Challenges/Opportunities 16

- For Businesses 16
- For IBM 16

Conclusion 17

About the Analyst 18



IDC Opinion

Today's IT landscape can appear as a conundrum. In the drive toward becoming a digital enterprise and satisfying the needs of hyper-demanding customers, businesses find themselves attempting to achieve the near impossible.

- Markets can change on a whim, causing spikes or depressions, and this volatility cannot be interpreted as an exception. Volatility is today's baseline.
- To fulfill the workload ebb and flow of demand, systems must scale flawlessly and dynamically without requiring a massive, costly, and energy-consuming datacenter buildout just for the spikes. Sustainability is not just a marketing gimmick anymore.
- The complexity of these markets can also no longer be analyzed and leveraged with common human experience and intelligence. Much of the intelligence must now be artificial, operating in real time and juggling countless variables while taking in vast amounts of data. Artificial intelligence (AI) will increasingly infuse everything everywhere, and AI requires purpose-built hardware capabilities.
- Given the demand for permanent and perpetual availability, the workloads that underpin the digital enterprise cannot be slowed down or impeded, let alone go down entirely. In today's always-on world, any downtime can be catastrophic.
- With everything digital and connected to enable the digital enterprise, everything has also become exposed to, and can be compromised by, new types of attacks. Entire communities of ill-intentioned people have coalesced into a netherworld that wages permanent war on businesses around the globe using a vast arsenal of cyberattack tools and strategies. As a result, everything now must start with comprehensive, watertight security.

With this in mind, for any enterprise-class compute platform to be able to function as the engine of the digital enterprise, it must be unequivocally secure, reliable, scalable, sustainable, integrative with cloud as part of a hybrid approach, and built for AI. This white paper will dig deeper into these considerations from an infrastructure and deployment perspective, and it will take a look at how the new IBM Power10 processor and the new IBM Power enterprise-class platform, the E1080, executes on them.

Situation Overview

IDC believes that for a digital enterprise to succeed in today's challenging, multifaceted environment, the following are critical considerations:

- **Security as an imperative requirement**
- **The reliability mandate**
- **Scalability and sustainability**
- **The right hybrid IT infrastructure (hybrid cloud and cloud-native applications)**
- **The importance of AI and where to run it**

The next sections will go deeper into each of these considerations.

Security as an Imperative Requirement

Security has become the single most important requirement of a digital enterprise. When IDC surveys organizations about their priorities, security is invariably at or near the top of the list. Indeed, when asked, for example, to select the top AI infrastructure items that businesses feel are not optimal within the offerings of their server and storage vendors or providers, security scores highest, with 30% saying that they are unhappy with the security features.[1]

This discontentedness also shows in the fact that many businesses do not allow the storage devices that contain the data for their AI workloads to be used by other workloads. The reason most often given for this (45%) is security and data privacy. Furthermore, IDC research has found that security is a top concern in public cloud infrastructure as a service (IaaS), with 37% of enterprises saying that security is their biggest concern in those deployments.[2] Businesses are also increasingly infusing their security workloads with AI, more than any other workload, to make them more capable of predicting and acting on breaches.

Currently, most of the attention and investment are going toward the security of application and networking stacks. A large number of attacks, however, are low-level and hardware-centric. They are often launched by taking advantage of vulnerabilities in the processors and/or microcode. These attacks are sophisticated and hard to detect.

IDC is therefore seeing businesses become increasingly interested in "confidential computing" for their critical business platforms. Confidential computing enables the isolation of sensitive data to a designated and protected processor subsystem (sometimes referred to as a "secure processor enclave") for processing. Today, data is often encrypted at rest in storage and in transit across the network, but not while in use in memory.

[1] Source: IDC AI Infrastructure View 2021

[2] Source: IDC IaaSView 2020

The ability to protect data and code while it is in memory is limited in many computing platforms. Yet organizations that handle sensitive data such as personally identifiable information (PII), financial data, or health information need to mitigate threats that target the application or the data in system memory.

In confidential computing, the contents of the subsystem, which may be encrypted at the hardware level, are accessible only to authorized code within a program. The contents are inaccessible to anything outside of it, including other code, other systems, or other operators. Unauthorized entities cannot view or tamper with the data or with the authorized code execution process. A comprehensive confidential computing solution will secure data in use as well as at rest, this can be enabled by the encryption of contents in volatile or non-volatile system memory and persistent data stores, either on flash or rotational media.

Modern confidential computing infrastructures – especially those implemented in shared, multi-tenant environments – make use of discrete coprocessors to offload privileged processor operations that can be compromised by low-level code execution vulnerabilities. This is not a common approach yet, but for core enterprise workloads it holds significant promise. In the meantime, businesses are leveraging various security strategies simultaneously, with both hardware and software.

The Reliability Mandate

While security strategies are critically important in protecting data, applications, and hardware from attacks, another crucial aspect of the digital enterprise is unmitigated reliability of the IT environment, including the infrastructure. High availability is hardly a new concept, and businesses can choose compute platforms with up to 99.999% availability and storage platforms with 99.99999%. But these numbers are only achieved with the right hardware, software, and policies. IDC has designated only nine server platforms from six vendors in the server market as Availability Level 4 (AL4)[3], which is the highest level and represents full fault tolerance.

- IDC research[4] shows that the top three causes of application downtime are failure in the network (16.2%), failure in the servers (15.5%), and malware (10.3%). Among the most common causes of server failure are overload on memory (DRAM) or CPUs and memory failure or corruption.
- Transaction volumes are increasing dramatically, and businesses need ever-faster transaction speeds to satisfy their customers.
- Mission- and business-critical workloads are growing, and business support functions that previously could be run on a low availability tier — for example, through virtualization or clustering — are increasingly deemed business-critical.
- The cost of downtime is increasing as businesses become more and more dependent on their infrastructure for daily operations. IDC research shows that for 20.7% of organizations, the cost of downtime is $5,000–10,000 per hour; for 18.4%, it is $10,000–25,000 per hour; for 17%, it is $25,000–100,000 per hour; and for some businesses (1.4%), it is $500,000 per hour.
- The end of "regular business hours," with businesses' applications now required to be available to customers at all times, has put tremendous pressure on the infrastructure that supports those applications, allowing for little if any scheduled or unscheduled downtime.
- Tolerance for outages, delays, data loss, and data corruption is zero – from both businesses and consumers – and any breaches or errors can have catastrophic consequences for an organization's reputation.

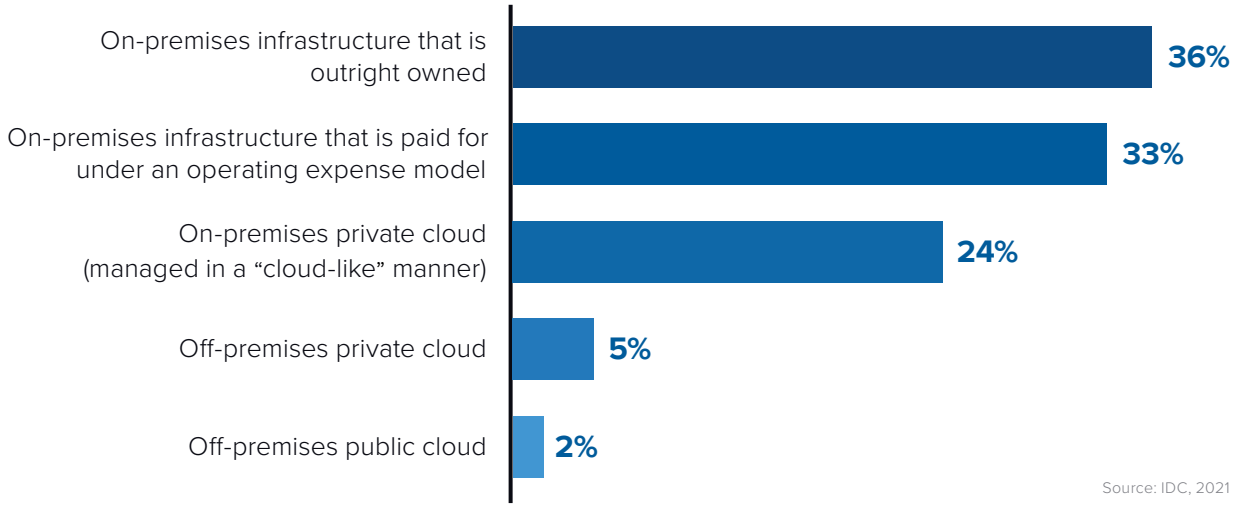
[3] Source: IDC Worldwide AL4 Server Market Shares, 2019: *Fault-Tolerant Systems Become Digital Transformation Platforms*

[4] Source: IDC *Server Storage Infrastructure Availability Survey,* 2018

- As businesses engage digitally with consumers or citizens and with other businesses more often and in many more diverse ways, compliance with national and international regulations on data availability, security, and privacy is of paramount importance.
- Even as availability and security in the public cloud have greatly improved, true fault tolerance continues to be seen as an on-premises or hybrid cloud capability, not as a public cloud capability (see **Figure 1**).

FIGURE 1

Infrastructure That Hosts the Highest Availability Tier

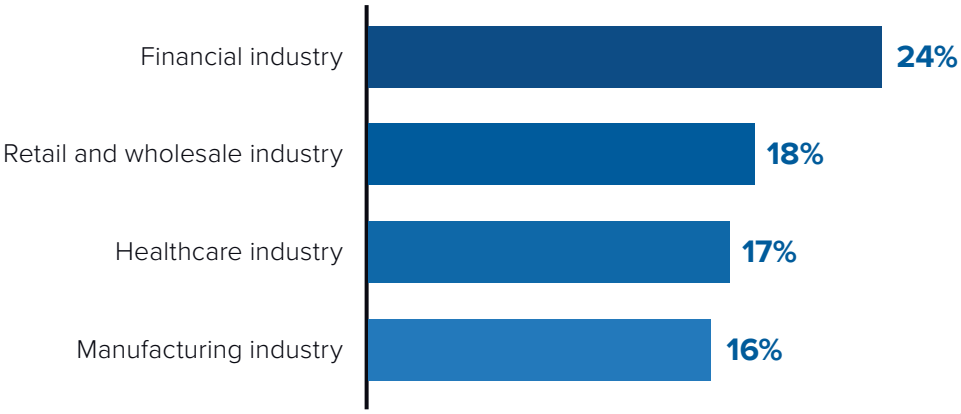


Source: IDC, 2021

The percentage of all systems that need to be highly available is growing as a result. Across all industries, more than 60% of businesses have 21–30% of all of their servers in the highest availability tier. **Figure 2** shows the percentage of systems that need to be highly available across various industries.

FIGURE 2

Percentage of Systems That Need to Be Highly Available, by Industry



Source: IDC, 2021

Today's dominant AL4 platforms have made major strides toward becoming fully integrated platforms in the datacenter that not only participate in an organization's digital transformation but in fact drive it. These systems process many enterprises' most critical and most valuable data, often in greater volumes than any other data types, and businesses need to unlock this data and leverage it to become digital enterprises.

The Need for Scalability and Sustainability

Businesses need to scale workloads that process ever-greater amounts of data on consistently expanding IT environments.

At the same time, they need to be able to rapidly scale up and down based on sometimes unpredictable waves of demand that can occasionally take the shape of severe spikes. All of this means larger datacenters, more equipment, more equipment renewal, and more energy to run the equipment while at the same time cooling it down.

AI workloads are the fastest-growing portion of workloads that consume data and drive the compute investments that businesses make. Currently, 21% of organizations say that they are investing in compute technologies that enable the parallel processing needed for training and inferencing on AI deep learning networks, and an additional 9% of businesses say they plan to do so in 2021. Also, 46% of businesses are investing in workload acceleration technologies such as GPUs, FPGAs, and ASICs, and an additional 7% plan to invest in 2021.[5] The latter, especially, have led to datacenter issues with regard to energy requirements and cooling. The most common use case for acceleration is AI deep learning inferencing (taking an AI model that has been developed with a deep neural network [DNN] into production). Currently, 38% of organizations use acceleration for AI inferencing, whereas only 27% use acceleration for the training of a DNN.[6] This trend, that AI inferencing compute investments start to exceed AI training, was expected. Moreover, AI is not the only workload that is driving investments in acceleration using such coprocessors. Data analytics, HPC, financial modeling, cybersecurity and fraud detection, and financial trading are additional examples of workloads that are increasingly running on GPUs, FPGAs, or ASICs, and a majority of businesses run these workloads on premises.

A major issue, however, is that most datacenters are not equipped to sustain multiple racks of accelerated compute nodes in terms of delivering the wattage they require and dissipating the heat they generate, which is much greater than with racks of non-accelerated servers. According to the U.S. Department of Energy (2020), datacenters are one of the most energy-intensive building types, consuming 10 to 50 times the energy per floor space of a typical commercial office building. IDC has found that, on average, 17.6% of the datacenter operating budget is spent on electricity, more than any other budget item. In the United States datacenters account for 2% of the total electricity use in the commercial sector.

At the same time, though, many organizations, especially in the tech industry, are trying to improve their carbon footprint. Tech firms lead the Environmental Protection Agency (EPA) list of green companies, and IDC has seen massive investments in renewable energy in the tech sector as well as investments in more energy-friendly hardware and software that help reduce energy consumption. IDC has found that the latter has helped reduce energy consumption on average by 26%.

Currently, 21% of organizations say that they are investing in compute technologies that enable the parallel processing needed for training and inferencing on AI deep learning networks.

[5] Source: IDC IT Infrastructure Plans for 2021 Survey, 2020

[6] Source: IDC IT Infrastructure for Compute Survey, 2021

Many enterprises have taken a cue from cloud service providers for a more sustainable approach to their IT, namely by reusing and recycling their equipment; 33% of respondents to an IDC survey[7] said they believed that this plays a role in achieving greater sustainability. Reuse and recycling of equipment can indeed contribute significantly to the overall carbon footprint of a datacenter. There may be reasons to upgrade certain components of a server, but the amount of must-have new components between two server generations does not exceed the number of components that could simply remain in place and be reused.

More awareness is emerging around this reuse opportunity to reduce the environmental footprint, and IDC has predicted that by 2025, 90% of G2000 companies will mandate reusable materials in IT hardware supply chains, carbon neutrality targets for providers' facilities, and lower energy use as prerequisites for doing business.[8] Such measures also help reduce costs for businesses, be it from lower energy use or reduced hardware investments.

The Right Hybrid IT Infrastructure

Gravitating Toward Hybrid Cloud

Today, 54% of organizations' applications are still deployed on premises.[9] IDC does not see this percentage going down significantly; businesses say that in two years, they expect to still run 52% of their applications on premises. Of those on-premises applications, 56% run as a private cloud, a figure which is expected to increase to 60% in two years. As to whether private cloud meets their goals, 61% of organizations say that it not only meets but exceeds their expectations.

Many of these applications, especially the critical business applications, have complex interdependencies. On average, businesses say that 49% of their business applications have some dependencies and 27% have complex interdependencies. Today, only 18% of all applications are considered to be "cloud-native," meaning they are modular, disaggregated microservices that represent suites of independently deployable services. In contrast, 32% of applications continue to be monolithic. This will change very rapidly, however. Businesses say that in two years, only 21% of critical business applications will be monolithic, while 44% will be cloud-native.

At the same time, businesses expect that they will be leveraging different on-premises and off-premises cloud deployments – what is often referred to as a "hybrid" cloud, which IDC sees as a fast-growing scenario. **Figure 3** shows that the most common cloud combination today is having multiple clouds to migrate workloads and data between. For the private cloud/public cloud scenario, about 40% of organizations say that these two deployments interoperate in their organizations – in other words, serving as a more-or-less integrated hybrid cloud.

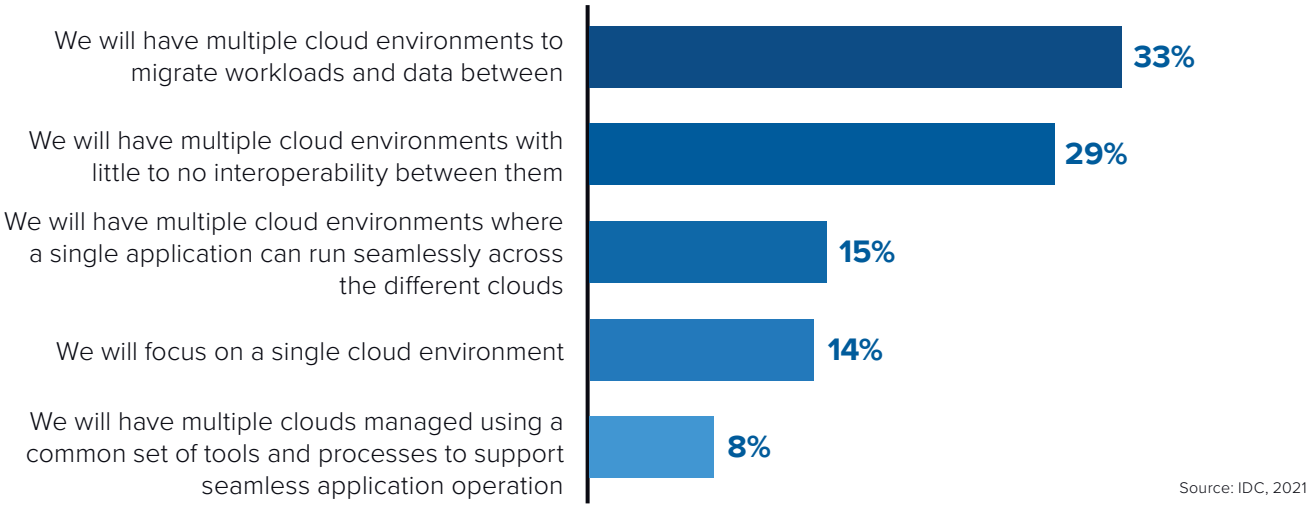
Note that for the on-premises portion of a hybrid cloud, businesses overwhelmingly (84%) want to move from a capex to an opex model. Currently, 42% of businesses' IT budgets are financed with an opex approach; three years ago, this figure was 36%.

Note that for the on-premises portion of a hybrid cloud, businesses overwhelmingly (84%) want to move from a capex to an opex model.

[7] Source: IDC 2021 Datacenter Operational Survey
[8] Source: IDC Worldwide Future of Digital Infrastructure 2021 Predictions
[9] Source: IDC 1Q21 Cloud Pulse Survey, May 2021

FIGURE 3

Use of On-Premises and Off-Premises Cloud Environments

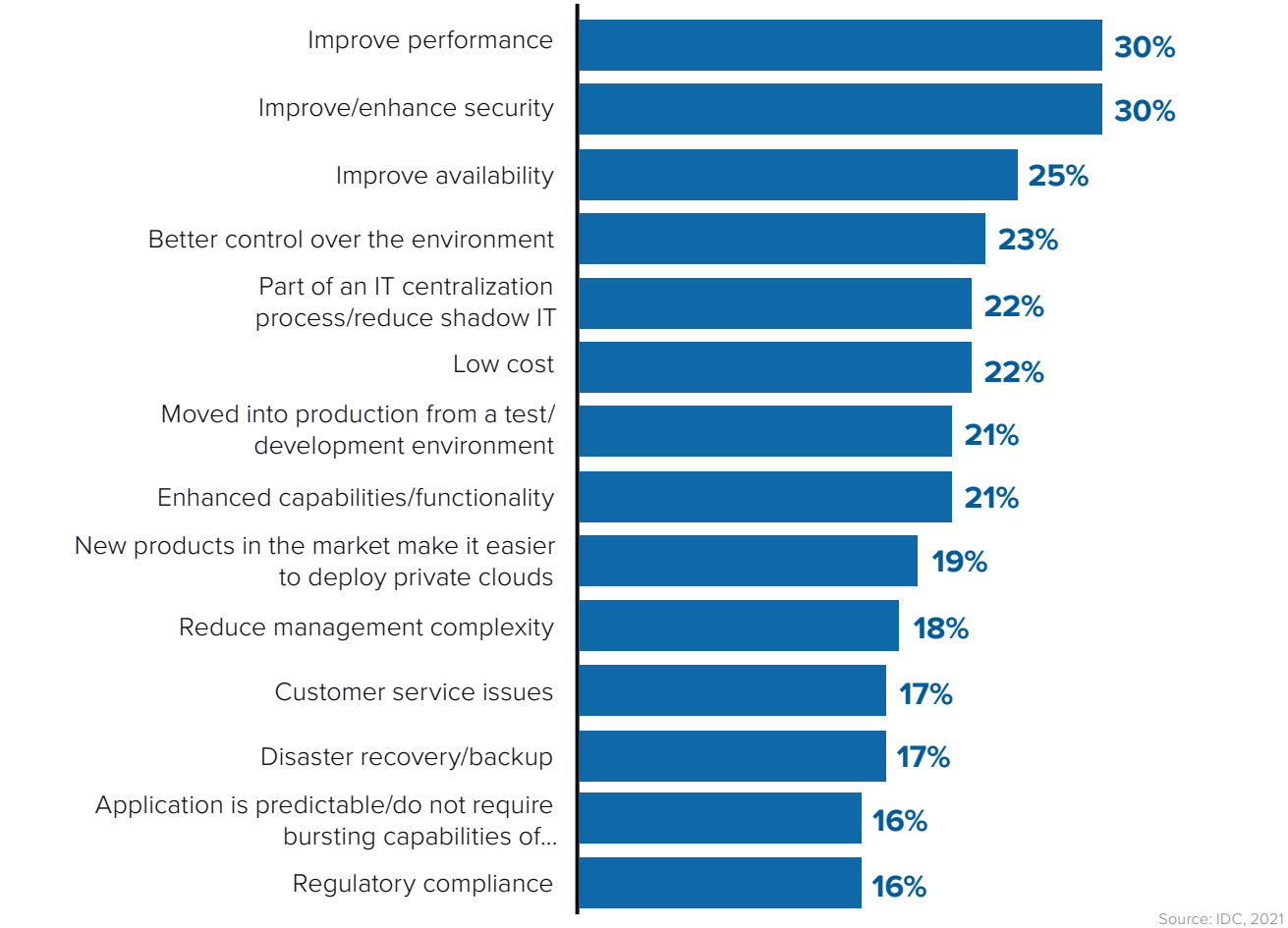


Source: IDC, 2021

As hybrid cloud becomes more prevalent, repatriation from a public cloud to a private cloud is also very common: 66% of businesses say they move applications to their private cloud or non-cloud environments for various reasons, with performance, security, and availability being the top three (see **Figure 4**).

FIGURE 4

Reasons for Moving Applications from IaaS to Private Cloud or Non-Cloud



Source: IDC, 2021

Hybrid Cloud and Cloud-Native Applications

A hybrid cloud that has been designed correctly is an ideal platform for developing and running cloud-native applications, which more and more businesses consider to be an important capability for their digital transformation. IDC has found that a majority of businesses consider implementing various capabilities to be "important" to "extremely important" for meeting their business needs as they invest in a cloud strategy that is suitable for developing and running cloud-native applications. These capabilities include:

- Greater application performance, availability, portability, and management
- Improved data integration, orchestration, observability, API management, and AIOps across cloud environments
- Faster development cycles and time to market with CI/CD (continuous development and deployment) and automation
- Comprehensive security policies, risk management, disaster recovery strategies, and regulatory compliance
- An opex rather than a capex model, including charge-back capabilities
- Optimized staff productivity, efficiency, and skill sets

Businesses that want to increase their investments in a hybrid cloud need to "put checkmarks" next to these items to ensure that they achieve the ROI they are anticipating.

The Importance of AI and Where to Run It

IDC expects that the worldwide artificial intelligence server platforms market will grow to $27 billion by 2025.[10]

This growth will be driven by increasing adoption of conversational technologies, natural language processing (NLP), image and video analytics, deep learning, machine learning (ML), hypothesis generation, and predictive analytics. As a result, AI server platforms will constitute 21% of the total worldwide server market by 2025.

In a previous section, we discussed the growing need for coprocessors in order to run both AI training and AI inferencing workloads. Given that on-premises private cloud is the preferred deployment scenario for AI, and on-premises non-cloud the second most common, this translates directly into significant investments for businesses in terms of added GPUs, FPGAs, and ASICs. For AI training, these investments are more or less unavoidable — training a DNN algorithm simply cannot be done on a host processor. For AI inferencing, however, there are many AI models that will run very well on an advanced host processor or on a host processor with an integrated specialty AI processor. These scenarios have a distinct cost advantage for businesses, given that adding a few GPUs to a server can quickly double the price of the total package.

This, of course, invites the question of why businesses continue to run their AI applications on premises in the first place. Why not run them in the cloud, for example, and avoid the capex altogether? Certainly, some amount of AI training is done in public clouds on the providers' AI platforms, and once developed, these models sometimes remain in the cloud as production workloads.

Given that on-premises private cloud is the preferred deployment scenario for AI, and on-premises non-cloud the second most common, this translates directly into significant investments.

[10] IDC Worldwide AI Server Forecast, 2021–2025, July 2021

The most important factor that determines cloud versus on-premises is data, and the following questions are the underlying reason for this:

What is the data needed to develop the model?

If it's data from core enterprise applications such as transactional data, remaining on the transactional platform is preferred, including for latency reasons.

How sensitive is that data?

If the data is sensitive, meaning that it must be vigorously protected, it will be undesirable to move it to the cloud, whether for training or inferencing.

What is the regulatory framework around the data?

Some data can legally not be moved to a public cloud, and this is true for core enterprise data more often than not. Businesses are subject to everything from in-country data protection regulations to GDPR to industry regulations such as HIPAA to ISO regulations to the California Consumer Protection Act.

What can and cannot be done with the data in order to remain compliant?

Once data is starting to get moved around, it becomes difficult to make sure that the organization remains compliant.

How voluminous is the data?

The more data that is required for training, or the more data that the AI model is inferencing on, especially if that inferencing is in near-real time, the harder it becomes to do so in the cloud.

How deeply integrated are the applications that leverage the data?

The platform that executes the transactions will most likely have multiple applications deeply integrated with the database to perform analytics and other functions, making it difficult to move that to the cloud.

How costly is the storage for the data?

Storage in large volumes in the cloud can quickly exceed any capital expenses that would be required for storage on premises.

In aggregate, these considerations lead many organizations toward remaining on premises with their AI training and inferencing workloads. They may still train on a separate compute environment in the datacenter behind their firewall, but then move the trained model back to the platform that runs the core enterprise applications for inferencing. If the platform enables robust inferencing, this allows businesses to use AI on core data that had been off-limits in the past.

IBM Power10 and IBM Power E1080

To successfully transform themselves toward the digital enterprise, businesses need compute platforms that can absorb any kind of market volatility, are secure without compromise, scale effortlessly while reducing the businesses' physical and carbon footprints, provide the highest levels of resiliency, and can run real-time AI on vast numbers of transactions — all as part of a seamless hybrid cloud. IBM's new Power10 processor and the enterprise-class IBM Power E1080 platform based on Power10 deliver a range of innovations that address these requirements in interesting new ways.

The New Power10 Processor

IBM's new IBM Power10 architecture and processor features important new technologies that will help businesses with compute, memory, and bandwidth-demanding workloads, including new technologies for fast AI inferencing on the chip without additional hardware, based on a built-in purpose-built matrix math accelerator (MMA).

From a security perspective, Power10 implements memory encryption without performance degradation (as opposed to software-based memory encryption), provides hardware/software co-optimized container security for container isolation; and includes security features to preempt the imminent capability of quantum computing to break traditional encryption keys.

Scalability with Power10 is expanded to new levels with several bandwidth innovations. IBM has enhanced the POWER AXON connectivity technology and added Open Memory Interface (OMI), both running at 32 GT/s. The Power10 AXON interface connects up to 16 sockets into a large, scalable system. The OMI communicates with DDR4 DRAM memory via 16 DDR ports per socket, providing bandwidth up to 409 GB/s per socket. These two interfaces can be used to provide very flexible, and even composable, compute solutions.

This is IBM's first 7-nanometer processor, and IBM claims a 3x efficiency gain compared to IBM Power9 in terms of compute power (number of users, number of transactions) and energy.[11] With IBM's ongoing focus on hybrid cloud, this translates directly into a smaller footprint in the datacenter and significantly reduced energy. There are 15 processor cores on the chip, and Power10 will feature PCI Gen5, which is beginning to emerge in the industry.

The IBM Power E1080

The IBM Power E1080 is IBM's first enterprise-class platform built with the Power10 processor. The system scales up to as many as 16 processors and is distinctly focused on top IT considerations for organizations that must meet the demands of the digital enterprise.

Security

To make security persistent and penalty-free, IBM has built encryption into the Power10 processor. This allows data to be encrypted without compromising system performance. The system has further been equipped with additional security features to protect against return-oriented programming attacks, a technique in which an attacker can

[11] 3X performance is based upon pre-silicon engineering analysis of Integer, Enterprise and Floating Point environments on a POWER10 dual socket server offering with 2x30-core modules vs POWER9 dual socket server offering with 2x12-core modules; both modules have the same energy level.

execute malicious code in the presence of security defenses. The Power E1080 provides advanced data protection with transparent memory encryption, the type of hardware-level security for data in use that confidential computing is based on, and features four times as many cryptographic encryption accelerators as its predecessor. Partitions on the platform have improved isolation, and the system is protected from future quantum-based threats with post quantum crypto (PQC) as well as fully homomorphic encryption (FHE), a technology in which inputs into the system don't need to be decrypted, which means it can be run by an untrusted party without revealing those inputs.

Resiliency

IDC considers the enterprise-class Power family of servers as having AL4 — in other words, fully fault-tolerant and therefore providing 99.999% or greater availability. With Power10, the IBM Power E1080 goes a step further than its predecessor in delivering very high bandwidth and memory reliability, availability, and serviceability (RAS) with the new Open Memory Interface. The processor can automatically detect, isolate, and recover from soft errors without an outage or without relying on the operating system to manage faults and self-heal recoverable errors. The system also features enhanced concurrent repair capabilities such as inter-node sub miniature push-on (SMP) cables to reduce application downtime.

Scalability and Sustainability

In terms of scalability and sustainability, the IBM Power E1080 benefits tremendously from the fact that the Power family of servers is exceptionally well integrated from processor to firmware to OS to hardware, as these are all IBM components. The software and OpenShift container efficiency of the platform is exceptional, according to IBM. As a result, the platform, with the new Power10 processor, achieves 50% more performance in the same space and energy footprint as compared with the Power E980.[12] This also translates to 33% lower energy consumption for the same workload, states IBM.[13] The greater efficiency helps businesses to significantly reduce their carbon footprint and potentially consolidate workloads, saving on both hardware and software costs.

Hybrid Cloud

The Power E1080 supports three operating environments — AIX, IBM i, and Linux — on the same platform, and is designed to support businesses' hybrid cloud adoption for all three operating environments. AIX is, of course, IBM's comprehensively modernized Unix operating system that continues to be a preferred platform for the enterprise-class, scale-up Power platform. IBM i is IBM's operating environment that integrates the database and other enterprise software into the operating system, greatly simplifying the platform's management — for many midsize businesses, IBM i is the heart of their operations. AIX and IBM i are extremely open source–friendly, support modern and preferred developer languages, and are fully operated as a hybrid cloud. As with previous generations, the Power E1080, can also run entirely or partially on Linux with the same security, availability, and scalability features, representing an opportunity for businesses to move their transactional and analytical workloads to a fully open source platform.

The following IBM Power software components play an important role in enabling businesses to leverage their enterprise-grade Power platform with AIX, IBM i, and Linux for secure, highly available, cloud-based workload modernization:

IBM PowerVM

- IBM Power server workloads are virtualized, mobile, and fully cloud-enabled with PowerVM, which was recently enhanced with multiple new features, including Compression and Encryption of Live Partition Mobility (LPM) Data, meaning that when an active partition is migrated from one Power server to another, which occurs with zero downtime, the data will be automatically encrypted and compressed — an important security and performance feature.

IBM PowerVC

- PowerVC is the virtualization management tool that is built on OpenStack, simplifying the management of virtual resources in Power environments. The software has recently been improved with multiple new features, including an export/import capability to share VM images across datacenters.

[12] Information provided by IBM. Based on published rPerf results for Power E980/12 core compared to IBM Internal rPerf measurements (using the same methodology) for Power E1080/15 core.

[13] Power9 (12c) is 5081 rPerf @ 16,520 Watts (0.31 rPerf/Watt), Power10 (15c) is 7998 rPerf @ 17,320 Watts (0.46 rPerf/Watt) 0.46 / 0.31 = 1.48 More rPerf/Watt

IBM PowerSC

PowerSC is the platform's security portfolio, simplifying security and compliance management, featuring compliance automation, malware intrusion detection, patch management, and more. It has been enhanced with various features or even new offerings, including multifactor authentication (MFA) enablement, another important security feature. In general, security on IBM Power with AIX is achieved with a comprehensive solution that includes the processor, firmware, hypervisor, and the countless security features of the operating system itself to protect data at all levels.

IBM PowerHA and VM Recovery Manager HA and DR

PowerHA is a high-availability technology that helps provide near-continuous application availability and improves service reliability. It is a key contributor to IBM Enterprise Power being characterized as fault-tolerant (AL4) by IDC and has been improved with various features such as enhanced failover metrics and cross-cluster verification (for example, to compare a development with a test cluster). VM Recovery Manager (VMRM) is a simplified HA/DR solution based on VM replication and restart that is operating system–agnostic and includes application monitoring agents such as for DB2, Oracle, and SAP HANA.

Cloud Management Console

The Cloud Management Console (CMC) provides a complete view on performance, inventory, and logging of on-premises and off-premises Power infrastructure. CMC is hosted on the IBM Cloud, thereby freeing businesses from having to maintain software to monitor their infrastructure and helping to simplify management of hybrid cloud deployments and to simplify the monitoring and management their infrastructure.

Enterprise Cloud Edition 2.0

Enterprise Cloud Edition brings together all of the key components of a simplified cloud management infrastructure on top of PowerVM, including PowerSC, MFA, PowerVC, CMC, VMRM, and Aspera. It enables rapid deployment and management of a private cloud; simplified security and compliance management; simplified high availability; and accelerated large-file transfers across clouds. Enterprise Cloud 2.0 can be purchased with AIX 7.2 built in.

Red Hat Ansible Automation Platform

Red Hat Ansible Automation Platform enables scalable and secure automation of various aspects of enterprise IT operations, including resource provisioning, application life-cycle management, and network operations. It consists of Ansible Engine, Ansible Tower, and Ansible Hosted Services. All other products within the Red Hat portfolio can be integrated using the Red Hat Ansible Automation Platform. Red Hat Ansible Automation Platform enables consistency in the datacenter by providing programmatic methods to deploy, manage, and secure infrastructure resources.

Red Hat OpenShift

Red Hat OpenShift is an enterprise-grade, certified Kubernetes (a container orchestration) platform to build, deploy, and manage containerized applications. Red Hat OpenShift can be consumed as a fully managed service on different cloud providers, or customer-managed using Red Hat OpenShift Container Platform or Red Hat OpenShift Kubernetes Engine. It can be deployed on premises on bare metal servers, virtualization platforms (Red Hat Virtualization, VMware, or Red Hat OpenStack), or major cloud providers such as IBM Cloud, AWS, Google, or Azure. In addition, Red Hat Advanced Cluster Management for Kubernetes can be used to manage multiple Red Hat OpenShift clusters and applications from a single console, with built-in security policies, enabling customers on open hybrid cloud. Red Hat OpenShift is supported across IBM Power, IBM Z, and x86-based platforms and can be used with AIX, IBM i, and Linux.

IBM Cloud Paks

IBM Cloud Paks are increasingly popular software products prepackaged in containers and highly integrated into various OpenShift services for fast and easy deployment onto OpenShift. IBM Cloud Paks offer developer tools, data, and artificial intelligence services, and open source middleware software. They run on the Red Hat

OpenShift cloud platform. Some Cloud Paks that are particularly relevant for IBM Power are:

- ***Cloud Pak for Data:*** helps customers with expanding insights from data and AI capabilities
- ***Cloud Pak for Integration:*** consists of integration tools for data, application services, and cloud services to help integrate apps, data, cloud services, and APIs
- ***Cloud Pak for Watson AIOps:*** offers multicloud visibility, governance, and automation, given the common use of multicloud deployments

Artificial Intelligence

IBM states that the Power E1080 speeds up AI inferencing performance by an order of magnitude compared to its predecessor. This does not require any specialized hardware such a coprocessor (GPU, FPGA, or ASIC). Instead, the inferencing takes place on a matrix math accelerator (MMA). Every core of the Power10 chip has a built-in MMA for efficiently performing matrix math operations. These operations have been optimized across a wide range of data types for various precisions, which are important for deep learning — from double precision and single precision to two types of half precision, including Bfloat-16 as well as Int-16, Int-8, and Int-4. AI inferencing performance has been infused into every layer of the processor. The L2 cache was quadrupled: the load store units and the SIMD doubled. This means that a transactional workload that has embedded AI components can run the transactions and the AI inferencing on the same Power10 processor without requiring a coprocessor.

Inferencing on the chip also means that all the processor's and system's security features are available to protect the data that is being inferenced on. Furthermore, the platform is Open Neural Network Exchange– (ONNX-) friendly. ONNX is an open source AI ecosystem of technology companies and research organizations working to establish open standards for representing AI algorithms and tools in order to promote innovation and collaboration in the AI sector. Businesses with IBM Power E1080 can bring ONNX models to the platform unchanged and run them, taking advantage of the platform's RAS features during the inferencing.

Challenges/Opportunities

For Businesses

Enterprise-class platforms that run an organization's core transactional and analytical workloads tend to be treated as silos in the datacenter, even if they are designed and built with comprehensive features and technologies to avoid this. These platforms are often being "protected" from new technologies by IT staff who have deep expertise with the system but who are wary of exposing the data, integrating the platform with the cloud, running open source on the platform, and executing AI models on real-time data. For businesses, the challenge is to break with this culture of hesitancy as soon as possible. It is absolutely critical to allow enterprise-class platforms to be appreciated as the open systems they are — this will allow businesses to fully leverage them as digital transformation platforms that drive new revenue opportunities. At the same time, these platforms offer an opportunity to begin seriously addressing sustainability issues, reducing the organization's carbon footprint. And running AI on an enterprise platform without the need for expensive, energy-consuming coprocessors is an important requirement, as more and more core applications are being infused with AI functionality.

For IBM

With the new Power E1080 platform, IBM continues to drive businesses toward openness, hybrid cloud, AI, and sustainability in a highly secure, performant, and reliable platform. IBM tends to meet innovation challenges with interesting new technologies that in some cases are groundbreaking and ahead of the pack: Take, for example, the MMA in the new Power10 processor. Innovation is not IBM's greatest challenge. The real challenge for IBM is to shift a portion of its customers' mindset from treating their enterprise platform as a siloed system, or perhaps a carefully opened-up system, toward an aggressively integrated platform with the rest of the datacenter and with the cloud, fully leveraging all of its capabilities to do new things and generate more revenue with the core data that lives on the platform. IBM needs to continue encouraging its customers to be bold and creative with its enterprise platform through education, incentives, and ROI studies.

Conclusion

Modern businesses need compute platforms that can handle extreme market volatility, provide unyielding security, scale effortlessly and sustainably, deliver utmost resiliency, execute real-time AI, and operate as a hybrid cloud. IBM's new Power10 processor and the enterprise-class IBM Power E1080 platform based on Power10 address these requirements head-on. IBM's new-generation Power processor is anything but an incremental step and ventures into important forward-looking territory.

The processor enables confidential computing technology for hardware-based encryption that secures data in flight. Bandwidth on Power10 has been greatly increased to enable powerful 16-socket scalability. Resilience is further enhanced with the ability to automatically detect, isolate, and recover from soft errors without an outage or without relying on the operating system. The MMA on the chip powers real-time AI inferencing without the need for a coprocessor. And between Red Hat solutions and IBM Cloud software, the ability to fully operate as a hybrid cloud is a given. With the Power10 chip as the engine of the new Power E1080 platform, IBM continues to push enterprise computing toward a sweet spot where the best of all worlds come together: openness, compute power, hybrid cloud, AI, security, scalability, sustainability, and reliability in a single platform.

About the Analyst

Peter Rutten

Research Director, Infrastructure Systems, Platforms and Technologies Group,
Performance Intensive Computing Solutions Global Research Lead, IDC

Peter Rutten is a Research Director within IDC's Worldwide Infrastructure Practice, covering research on computing platforms. Mr. Rutten is IDC's global research lead on performance intensive computing solutions and use cases. This includes research on Artificial Intelligence. (AI), Modeling and Simulation (M&S), and Big Data and Analytics (BDA) infrastructure and associated solution stacks. His coverage of performance intensive computing includes supercomputing, high-end, accelerated, in-memory and heterogeneous computing infrastructure systems, platforms, and technologies. It includes computing platforms with GPUs, FPGAs, ASICs, and other accelerators that are deployed in the cloud as well as on-premises. It also includes research on mission-critical x86 platforms, mainframes, and RISC-based systems as well as their operating environments (Linux, z/OS, Unix). Mr. Rutten also examines emerging technologies and platforms such as quantum computing, neuromorphic computing and technologies that are potentially disruptive to mature infrastructure markets. As part of his role, Mr. Rutten performs quantitative (market sizing and forecasting) and qualitative (primary research based) analysis as well as custom market sizing for IDC's clients.

More about Peter Rutten

IDC Custom Solutions

This publication was produced by IDC Custom Solutions. As a premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets, IDC's Custom Solutions group helps clients plan, market, sell and succeed in the global marketplace. We create actionable market intelligence and influential content marketing programs that yield measurable results.