PERFORMANCE-INTENSIVE BUSINESSES NEED A NEW CLASS OF INFRASTRUCTURE

Authors:

Andrew Buss
Luis Fernandes

February 2023

An IDC Spotlight sponsored by AMD

IDC #EUR150224123

Performance-Intensive Businesses Need a New Class of Infrastructure

Introduction

The global economy is undergoing a marked shift to digital. IT infrastructure has moved on from being an enabler of back-office efficiency, to being a key driver of competitive advantage and differentiation, integrating across supply chain, product development, operations, sales and marketing, alliance and partnering, go-to-market, and customer experience.

This means that the application portfolio of a modern digital business is evolving. Most enterprise applications remain suited to run on a typical IT infrastructure. However, new classes of real-time data-centric workloads — such as Big Data and analytics (BDA), high-performance computing (HPC), and artificial intelligence (AI) — are generating significantly higher demands on infrastructure. This requires new architectural approaches to deliver the performance required.

To deliver this performance, many companies are choosing to have part of their IT infrastructure dedicated to delivering higher performance — and they are choosing to do this with a new generation of accessible enterprise-class performance-intensive computing (PIC) infrastructure.

AT A GLANCE

PIC WORKLOADS ARE EXTENSIVELY ADOPTED

» 56% run AI as a major workload.

» 34% run HPC as a major workload.

» 31% run DBA as a major workload.

PIC NEEDS DEDICATED INFRASTRUCTURE

» 93% say AI needs dedicated performance-oriented infrastructure.

» 83% say BDA and HPC need dedicated performance-oriented infrastructure.

KEY TAKEAWAYS

» PIC workloads are increasingly critical in enabling business innovation, operational efficiency, and enhanced customer experience.

» The key areas to focus on to ensure successful PIC implementations are data quality, infrastructure skills, and data scientists.

PIC-focused infrastructure integrates high-end scalable CPUs, large memory support, and high-performance disk and network IO, together with integrated support for workload-specific compute accelerators such as graphical processing units (GPUs), field programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs).

PIC Critical to Digital Success

IDC's May 2022 *European Performance-Intensive Computing Survey* of 280 midmarket (500–999 employees) and enterprise (1,000 or more employees) businesses showed that three-quarters of businesses are already running a PIC workload, be it AI, BDA, or HPC, and that for over half this represents a major workload for them.

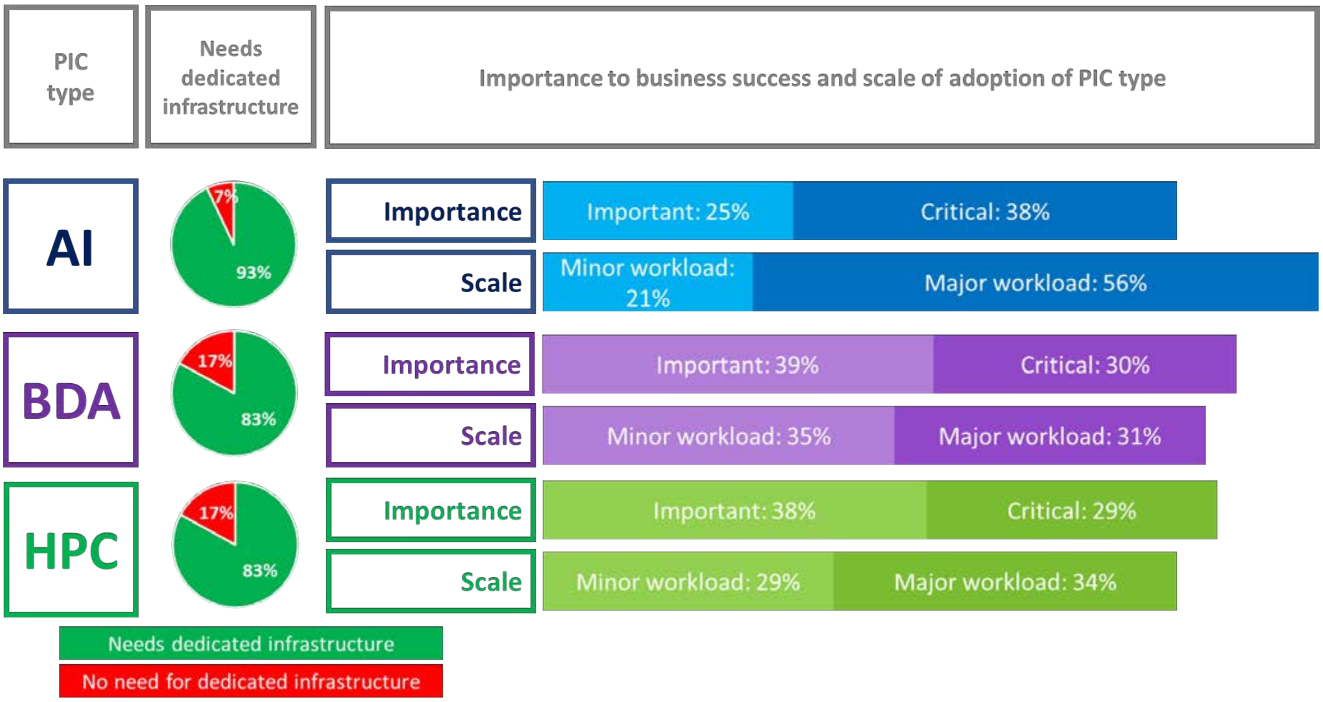
When we looked at how important each of these PIC types is in enabling the overall success of the company, AI was seen to be the most important PIC workload, with 38% of respondents feeling that AI is critical to the success of the company, with a further 26% recognizing that it is an important ingredient for success.

As a result of the critical role that it plays, AI has moved beyond its infancy to become the most adopted PIC workload by some margin, with 56% of respondents seeing AI as a very important workload, with HPC following at 34% and BDA at 31%. Crucially, only 10% of companies had no plans to run HPC or BDA workloads, and just 6% for AI.

BDA is not far behind, with 30% of companies saying it is critical to success and 39% saying it is important, while HPC is almost the same at 29% and 38% respectively.

FIGURE 1

The Importance and Scale of PIC Workloads



Source: IDC PIC Survey Europe, May 2022, n = 280

Different industries have different approaches to different PIC types, with AI being most critical to retail, resource industries (such as gas and oil, and mining), and telecommunications. For BDA, it is resource industries, retail, and financial services, while for HPC it is resource industries, retail, and manufacturing.

We saw as well that as the PIC workload becomes more critical to the business it becomes a more significant workload running on the IT infrastructure, with a close correlation of the workload moving from being minor in scale as it initially becomes important to the business, to expanding greatly in scale to a major workload once it becomes critically important.

The key takeaway is that the scale of a PIC workload's infrastructure demands will track with how critical it is to the success of the business, and as a result we took a closer look at how companies are approaching the infrastructure required to deliver PIC workloads.

PIC Demands Drive the Need for Dedicated Infrastructure

The demands of PIC are high, with extreme demands across CPU, memory, accelerators, disk, and networking. What may work very well for typical enterprise applications will not deliver a good, let alone great, experience when it comes to delivering value from PIC workloads. Each type of PIC workload has a different set of requirements and bottlenecks to balance and resolve, with the biggest common challenges being memory size and interconnect bandwidth:

- **AI:** memory size, interconnect bandwidth, co-processor
- **BDA:** memory size, CPU security, interconnect bandwidth
- **HPC:** memory size, interconnect bandwidth, software stack

As a result, what we saw is that for all PIC workloads there was a strong preference across all respondents and industries to have dedicated infrastructure that is focused first and foremost on performance for that workload. This was very strongly felt for AI workloads, where 93% of respondents indicated that AI needs dedicated infrastructure focused on performance, with both BDA and HPC at 83% following closely behind.

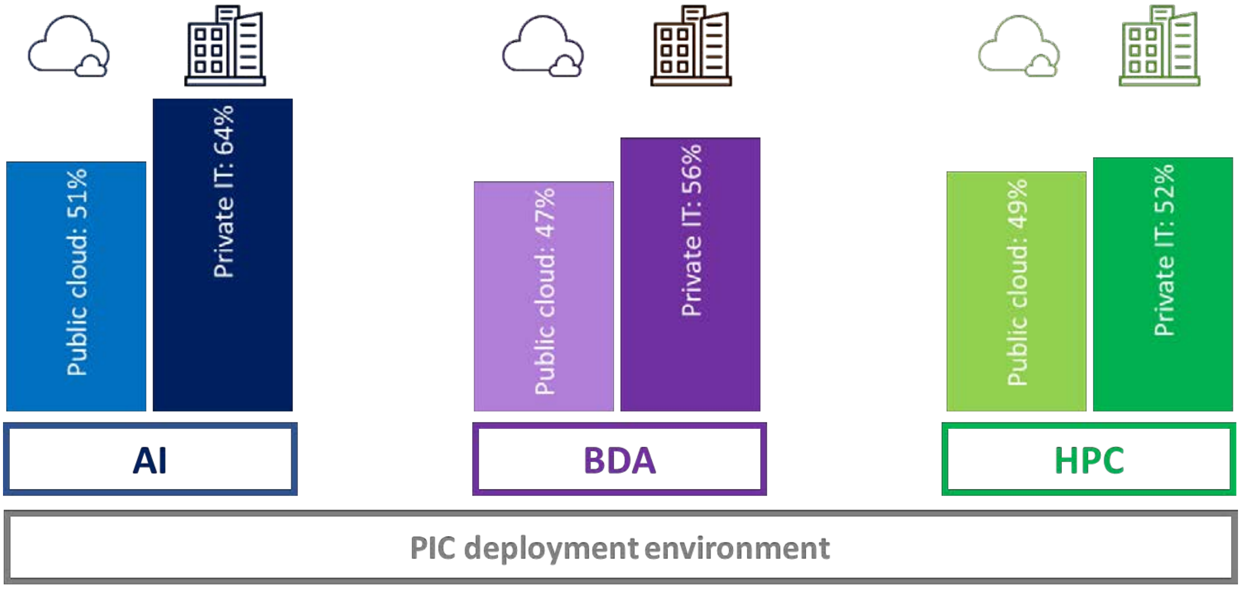
93% of respondents indicated that AI needs dedicated infrastructure focused on performance.

With a need for dedicated performance-centric infrastructure for each of the PIC workload types, it becomes a question of how to approach this, and in particular whether it is worth building and running this in private IT infrastructure or to run PIC workloads in the public cloud.

Deployments Are Focused on Private IT Infrastructure, But Hybrid Approaches Mean Public Cloud Is an Important Consideration for PIC

Due to the high-end performance requirements and scale required, PIC is not a low-cost solution. This would typically result in public cloud services being a great fit. Indeed, we do see significant adoption of public cloud for AI (51%), BDA (47%), and HPC (49%). Key reasons for choosing public cloud for AI are a cloud-first IT strategy and built-in security. For BDA, respondents valued easier data access and the global scale and availability of public cloud infrastructure. Public cloud HPC users meanwhile focused on better resilience and disaster recovery solutions, and built-in security.

FIGURE 2
PIC Deployment Environment



Source: IDC PIC Survey Europe, May 2022, n = 280

However, even though public cloud adoption for PIC is strong, it is private IT infrastructure that remains the most popular deployment type for PIC workloads, with AI at 64%, BDA at 56%, and HPC at 52%.

There are a number of reasons for preferring private IT for PIC deployments. For AI, the primary driver is to keep data in private datacenters as well as having greater control over the actual infrastructure architecture. For BDA, control of the infrastructure architecture is the main reason, followed by the ability to optimize the infrastructure for performance, while HPC values controlling the infrastructure architecture and having a free choice of platforms.

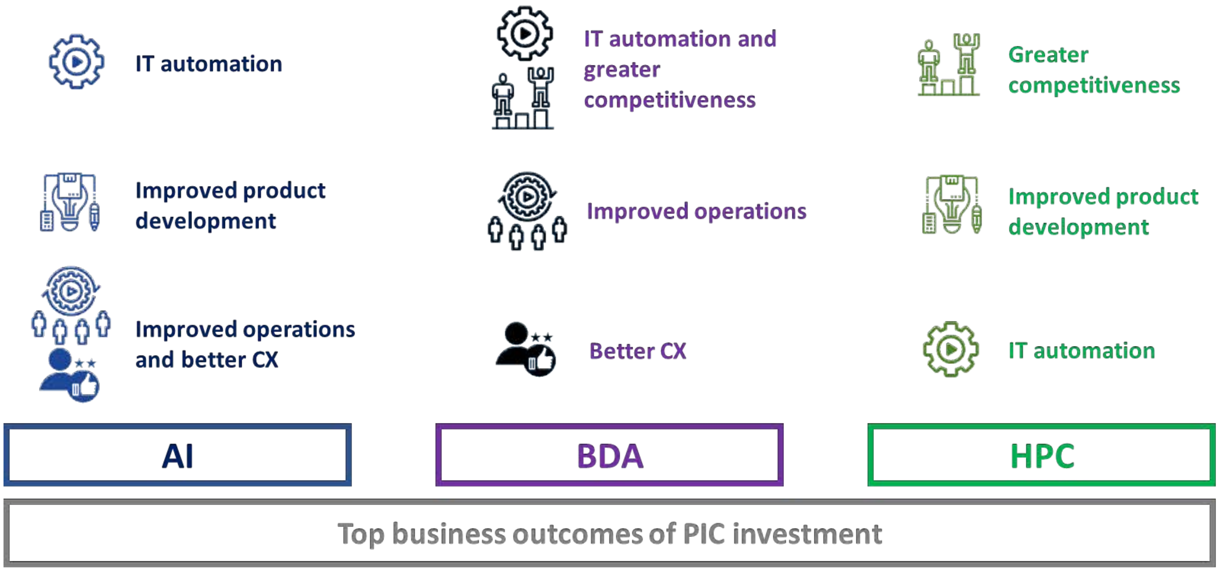
Private deployments of PIC infrastructure will continue to remain sizeable for the foreseeable future. One of the biggest ongoing concerns, because of the significant cost and scale, has been the upfront capex investment required. However, the rise of flexible consumption models from the major IT infrastructure vendors — such as HPE GreenLake, Dell APEX, and Lenovo TruScale — means that PIC solutions can be paid for in smaller monthly payments for the capacity that is used rather than having to commit to a large amount upfront.

Business Benefits of PIC Investment

We've seen how critical PIC workloads are becoming to a modern data-driven business, but what are the business benefits of investing in these capabilities?

FIGURE 3

Top Business Outcomes of PIC Investment



AI	BDA	HPC
IT automation	IT automation and greater competitiveness	Greater competitiveness
Improved product development	Improved operations	Improved product development
Improved operations and better CX	Better CX	IT automation

Top business outcomes of PIC investment

Source: IDC PIC Survey Europe, May 2022, n = 280

What we saw was a combination of benefits across both the IT organization itself and the overall business. For the IT organization, improvements to IT automation featured prominently across all three types of PIC, while also helping digital operations.

For the broader business, PIC helps the business develop better products, have greater competitiveness to outperform competitors, and improve the overall customer experience to retain and develop customers.

Keys to Successful PIC Solutions

When it comes to successful PIC solutions, we saw that there was a combination of higher-level issues such as skills and data issues, as well as lower-level technical areas, that need focused attention and investment to ensure they don't become limiting factors.

On the higher-level PIC issues, there were two issues that were equally felt to be the most significant: IT skill sets and data quality.

While building a general-purpose IT infrastructure is reasonably well understood, the performance requirements and associated specialized architectures of PIC require up-to-date and comprehensive knowledge of the latest workloads, development environments, architectures, and accelerators. There is an overall shortage of IT skills in the European market, and this is more acutely felt in the PIC market. While some larger companies may be able to recruit or train the required talent, many companies will be better off working with the PIC vendors or specialist VARs or SIs to scope, design, build, and operate their PIC solutions.

PIC solutions live or die based on the data they are able to work on. Many companies still struggle to have an information life-cycle management process that helps with the integration, classification, and management of their data resources, and the end result is poor quality data that no amount of technical investment can overcome. Investing in data skills will be a key factor in determining which companies will be winners in the data-driven market.

On the technical side, there was a fair amount of alignment that compute performance, network performance, and security were the most significant success factors overall, although the exact priority order did vary among the different PIC types, and in HPC security was not so much a concern as the software stack and operational costs:

- **AI:** security, compute performance, network performance
- **BDA:** compute performance, security, network performance
- **HPC:** network performance, compute performance, software stack and operational costs

Considering AMD for Your PIC Needs

AMD is a leading provider of silicon products with a broad portfolio covering compute, graphics, and networking. AMD has been broadly adopted across client devices such as PCs and laptops, to embedded devices for industrial, IoT, and edge workloads, through to scalable datacenter workloads and the largest high-performance compute installations.

AMD's broad portfolio helps companies accelerate a full range of datacenter workloads — from general-purpose computing, to technical computing, cloud-native computing, and accelerated computing. This gives scientists, engineers, and designers fast insights and accurate results.

AMD revolutionized the x86/64 server market with the AMD EPYC server CPU portfolio based on multiple chiplets within a single CPU package. This has enabled AMD to provide an industry-leading 96 high-performance cores within a single socket. In early 2022 AMD expanded this packaging technology to include 3D stacking of cache memory to bring even more performance to memory-dependent workloads. Having this many cores active would lead to bottlenecks without the ability to feed them data. The fourth-generation AMD EPYC 9000 series platform, launched in November 2022, includes a dedicated IO chiplet in the CPU package that has 12 DDR5 memory channels as well as 128 PCIe 5.0 lanes for IO such as storage and networking.

AMD also has dedicated datacenter workload accelerators, with the Instinct MI200 series of accelerator cards based on the AMD CDNA2 architecture, topped by the flagship multidie MI250 accelerator enabled by AMD's high-performance Infinity Fabric. This performance will continue to increase with the upcoming MI300, which supports coherent memory across CPU and accelerator for even more scalable performance on large systems.

AMD has also invested in a software ecosystem to support these CPUs and accelerators. It has developed ZenDNN — providing a platform for AI and deep learning workloads on EPYC CPUs — and AMD ROCm for targeting workloads to run on the AMD Instinct MI platform. AMD also provides the Infinity Hub, which groups together various AI applications and utilities.

AMD has been expanding its portfolio in recent years beyond its core markets of CPUs and GPUs, to include FPGAs and adaptive SoCs — through the acquisition of Xilinx — and next-generation networking with data-processing units (dPUs) through the acquisition of Pensando.

When datacenters are energy efficient, enterprise computing and scientific research can thrive. AMD has committed to environmental sustainability goals, publicly reporting on its progress, and pushing the limits of high-performance computing. It has set an ambitious "30 x 25" goal, which involves achieving a 30-fold improvement in energy efficiency for processors and accelerators powering HPC and AI-training workloads by 2025 over 2020 (see https://www.amd.com/en/corporate-responsibility/data-center-sustainability).

Accomplishing this ambitious goal will require AMD to increase the energy efficiency of a compute node at a rate that is more than 2.5x faster than the aggregate industrywide improvement made during the past five years. AMD has made good progress on delivering on this goal, and the results show this. In the November 2022 Green500 list, it leads efficiency in the supercomputer space with four of the top 5 most efficient supercomputers in the world. Beyond that, AMD products are in 15 of the top 20 most efficient supercomputers.

AMD is also very focused on its overall ESG strategy and has set a target of a 50% reduction in 2020-level greenhouse gas emissions by 2030, and requires its manufacturing partners to have their own publicly committed greenhouse gas emissions targets.

AMD has shown over the past five years that it can design cutting-edge innovation and deliver this on time in a predictable manner over multiple generations, making it a dependable partner for performance-intensive computing solutions both now and in the future.

Challenges

The technology industry has suffered from component shortages and supply chain disruption since the pandemic. AMD relies on third-party manufacturing, primarily with TSMC and Global Foundries, to manufacture its portfolio of CPUs and GPUs, and its ability to supply the market is tightly coupled to its ability to secure manufacturing capacity on the appropriate process nodes in these foundries. AMD has been executing well and growing strongly in the past five years, particularly in the datacenter space with AMD EPYC CPUs, which are built on a flexible chiplet architecture to enable more diversity in product binning and targeting during assembly rather than at early stages of manufacturing.

Although a long-established and significant player in the semiconductor market, AMD is still a relative newcomer to the modern datacenter infrastructure with AMD EPYC CPUs, AMD Instinct GPU accelerators, and Xilinx FPGAs. It takes time for software providers to certify their software and solutions on new platform systems, and there may be a more limited choice of AMD systems that are certified to run various applications or software stacks. However, the testing and validation that occurs on modern enterprise server and infrastructure platforms means the risk of incompatibilities is much reduced compared with a decade ago, and the potential advantages in performance and overall power efficiency will likely justify extra efforts in testing to ensure the hardware can run the PIC workloads reliably and consistently.

Conclusion

Data-intensive workloads are increasingly vital for companies looking to grow revenues and succeed in a digital-first economy. As these workloads become more critical to the business, so do their demands put pressure on the IT infrastructure. To deliver a great experience, businesses need appropriate performance-centric infrastructure that can scale securely and predictably, or they risk falling behind in competitiveness.

PIC delivers great benefits for the business, but it is not cheap. Although many companies are increasingly using public cloud to deliver parts of their PIC needs, most are still building solutions based on private IT infrastructure. New flexible consumption models can help businesses to build and operate a PIC solution without a large upfront capex commitment, enabling the benefits of a PIC solution to be realized immediately while paying regularly only for what is used.

PIC solutions require specialized infrastructure, and as a result also require specialist knowledge and skills to understand the requirements and to design, build, and operate challenging performance-centric infrastructure. As these technical skills are in short supply, working with technology suppliers and VARs and SIs can help supply the required skills to implement a successful solution.

Gathering, understanding, integrating, and managing the underlying data will be critical for data-driven PIC workloads. Data-science skills are going to be a critical factor for long-term scalable solutions — and this needs a long-term plan for recruiting, training, and retaining staff.

MESSAGE FROM THE SPONSOR

The AMD Story

We make the world's most powerful processors, and power the new wave of data-first business, from supercomputers and AI through to always-on infrastructure and public cloud. It is companies like you that put high-performance and adaptive computing to work and spark new ideas to work and life. Our long-term relationships with industry-leading partners give us a deep understanding of the processing power that tomorrow will demand. This pushes us to innovate and helps to propel the computing industry forward. This is where the future starts. https://www.amd.com/en/corporate/amd-story

About the Analysts

Andrew Buss, Research Director, European Infrastructure Strategies

Andrew Buss is research director for IDC's European Enterprise Infrastructure program. He is based in London and his research area focuses on understanding the convergence of technologies and capabilities and how they need to integrate and work together to deliver efficient, effective, and agile IT services from the datacenter or cloud through to the end user. He works with global, multinational, and local companies to understand the dynamics of business technology desires and needs, technology purchasing and investment, organizational and operations structures, and customer mindsets and disconnects to help accelerate IT transformation.

Luis Fernandes, Senior Research Manager, European Infrastructure Strategies

Luis Fernandes is part of IDC's infrastructure research team in Europe, bringing deep technology skills to the European Infrastructure Strategies CIS and the Edge Strategies practice. He has more than 15 years' experience in the ITC sector, in technical sales, covering several vertical sectors with a focus on telco, the public sector, and defense. Before joining IDC, he worked for HPE as a senior solution architect for servers and storage, with a focus on hyperconverged infrastructures. He has a degree in computer science from the University of Lisbon.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications, and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

IDC UK

5th Floor, Ealing Cross,
85 Uxbridge Road
London
W5 5TH, United Kingdom
44.208.987.7100
Twitter: @IDC
idc-community.com
www.uk.idc.com

Global Headquarters

140 Kendrick Street,
Building B
Needham,
MA 02494
+1.508.872.8200
www.idc.com

Copyright and Restrictions