Point of View

Supercharge your AI applications

Alleviate data overload and realise the potential of your artificial intelligence (AI) initiatives with the right infrastructure.

verizon business

Foreword

Supercharge your artificial intelligence (AI) ecosystem with a network that has all the components you need to realise the potential of your AI initiatives.

David Bailey
Global Solutions Executive

David has over 20 years of experience working with large global enterprises and public sector organisations to define and meet their infrastructure and security needs. This has included wired and wireless networks for Internet of Things (IoT) applications and agile, intelligent infrastructure to support demanding applications like artificial intelligence (AI).

This short paper discusses the rapid pace at which AI is evolving and how supporting technology layers need to keep pace with AI models and compute power. This includes high-performance, low-latency transport (networking), which can be crucial to successful business outcomes.

We've seen a phenomenal acceleration in the development and use of artificial intelligence and machine learning (AI/ML) technologies. How AI could help companies be more competitive, increase profits and transform operations and customer experiences is on the agenda of many, if not most, boardrooms across all industries. Faster hardware and more advanced software have been developed and will continue to be developed to support the increasing expectation for what AI can deliver.

Chip manufacturers—including Nvidia, AMD, Intel and Qualcomm—are competing for supremacy in the production of chips optimised for AI workloads. Google, Amazon and Microsoft have also developed their own processors to enable them to offer AI services at scale without becoming reliant on these manufacturers. Using this bespoke hardware, these cloud operators now offer public AI services that can process workloads in milliseconds.

This can be an effective way to quickly build and scale a solution, but if you rely on processing data in the cloud, the network is also critical. The network—often the internet—used to transport the data from where it's gathered to the cloud or data centre will dictate the responsiveness of the application. If the network is sluggish, the gains made by investing to increase the speed of the AI chipset may be largely, or even entirely, negated.

Consider the analogy of ordering a burger online: the restaurant could have it ready for you in minutes (the parallel of the super-powered chipset). However, if the delivery rider (the parallel of the internet), takes 30 minutes to get it to you, how long it took to cook is irrelevant, it still arrives cold.

In this paper I'll discuss the importance of latency (the delivery time) and what you can do to help reduce it and make your AI initiatives successful.

There's been a rise of organisations developing custom AI chips to improve performance. Examples include Meta's Training and Inference Accelerator (MTIA). This second-generation chip is performing better in terms of compute and memory bandwidth, improving the performance of Meta's AI applications, such as ranking and recommendation engines on its Facebook and Instagram platforms.

The use of custom-built AI hardware in discrete technology (i.e., technology not hosted in the cloud or in a data centre) also continues to evolve. Examples include:

- In healthcare, advanced scanners (e.g., CT, MRI, PET) that need to process huge amounts of data very quickly.
- Anyone who has recently bought a new vehicle will have noticed that there are sensors all over it: many enabling collision detection and other safety features. These sensors are connected to extremely powerful chips that enable the vehicle to make critical decisions in near-real-time.

In the examples above, most of the processing takes place locally—onboard the vehicle itself or within the scanner—notwithstanding that some of this data might be sent to the cloud after processing.

There are many other use cases where the processing typically takes place remotely, such as:

Finance

- Aggregation and analysis of vast quantities of historical and real-time data to make decisions where milliseconds can count. This can help improve financial results.
- Analysis of behavioural and biometric data to detect potential fraud. This can be much more effective and better for the user experience than traditional methods like "security questions."

Manufacturing

- Automated analysis of real-time video data from production lines, enabling every item to be checked at multiple points during the process—replacing traditional spot checks at the end. This can identify problems much more quickly, reducing waste and potentially mitigating damage to machinery.

Retail

- Analysis of behaviour and customer history to deliver better, more personalised in-store shopping experiences. This can improve the customer experience and increase customer lifetime value.

There are many more examples in manufacturing, agriculture and service industries. AI hardware can be seen everywhere. Where organisations have made the decision to move applications and compute power into the cloud, a network sits in the middle of where the data is collected, where it is processed, and where it is stored.

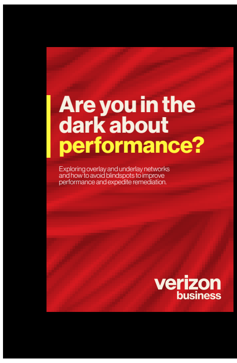
The problem

The changing economics of IT, largely driven by the rapid evolution of cloud computing, has led to in-house IT and procurement teams focusing on optimising the costs of networking. In many cases, these organisations have sourced network access (underlay) and the routers and switching (overlay) discretely, sometimes ending up using dozens of different local providers in the hunt for savings.

This focus on cost has often been at the expense of performance and quality of service. This has partly been driven by the misconception that all internet access is the same.

The reality is that the reliability and performance, especially latency, of access can vary hugely.

I cover this topic in detail in the companion paper below.



Are you in the dark about performance?

Exploring overlay and underlay networks and how to avoid blindspots to improve performance and expedite remediation.

verizon business

To split or not to split?

This other paper from this series looks into the perennial question of whether it's better to procure overlay and underlay separately or as a package.

15-minute read

Why latency matters

Latency, the delay between making a request and receiving the response, has three components (see below) and the perceived latency is the sum of these.

High latency could mean that a trade is missed, it's too late to block a fraudulent transaction or the opportunity for a customer sale is lost. In an autonomous vehicle or medical application, lives might be at risk.

Where AI is "self contained", such as onboard a vehicle or within an MRI scanner, the network is built into the appliance. The distance data needs to travel is minimal and latency is not dependent upon anyone else's infrastructure, so is minimal.

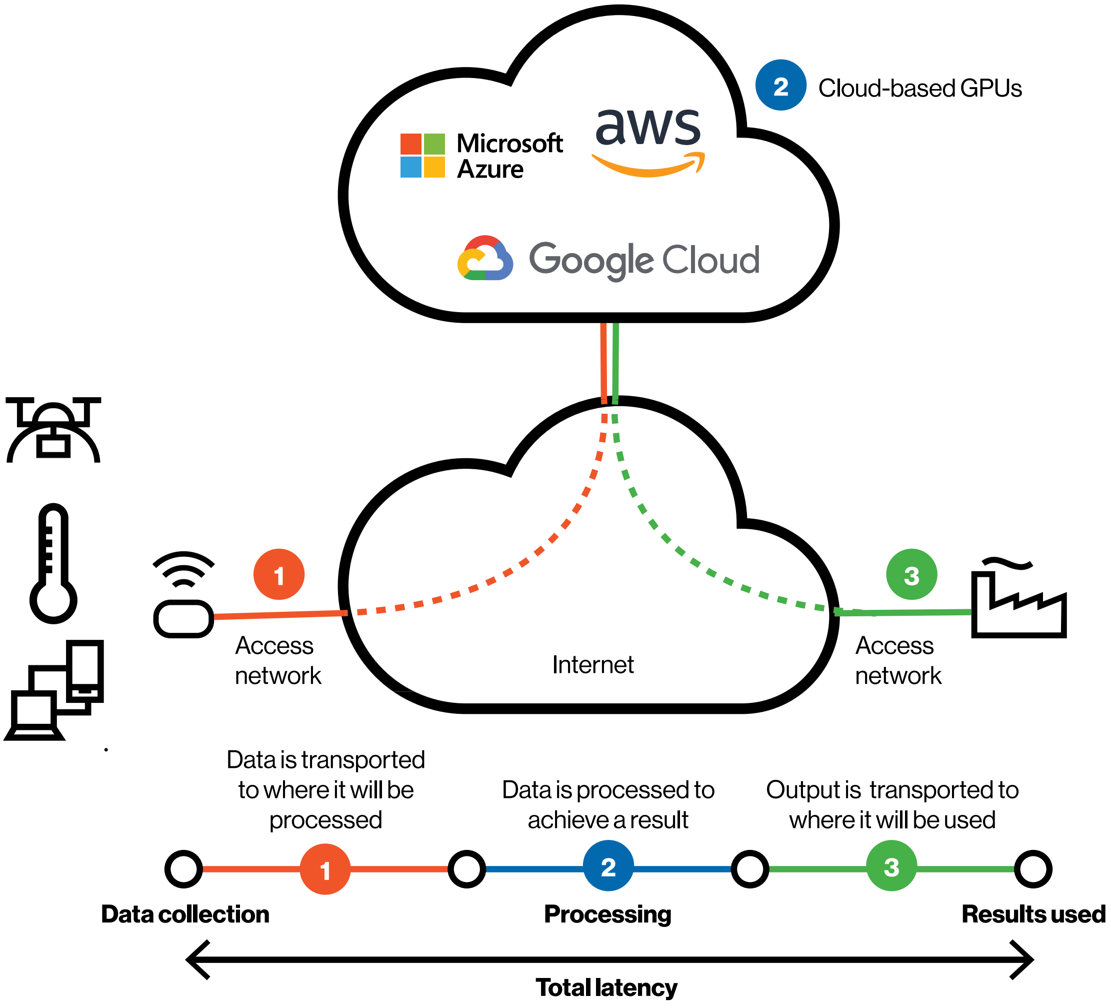
In scenarios where AI is incorporated in a single site with compute power contained in a single location such as across machines in a factory, then there is a LAN infrastructure (whether this is fixed or wireless) that needs to be carefully designed and managed to reduce the latency.

For organisations where AI might be spread across multiple locations, there's an additional dependency on the WAN. In this instance, the latency of your AI applications is in the hands of the suppliers of this infrastructure.

To support these different applications effectively, enterprises need to consider many factors, including:

- How much data is processed in the application, determines the total capacity of the computing architecture
- Where the data originates and how far it needs to travel for processing will determine the network latency
- Whether the application is human-facing or system-facing, and which use cases it's intended to support will determine how latency impacts output

At this point, it should be noted that the internet is made up of a collection of different networks. Data will take multiple different routes through the network and because of this there are typically no guarantees of latency through the internet. Organisations that use different suppliers for access in different countries or locations are even further at risk of poor latency performance. Where there are different providers of underlay (access) and overlay (managed routers and switches), there is no linkage between management of latency performance between the different providers.



Anyone involved in managing or operating an IT environment will know that when an application is performing poorly, the end user doesn't care where the issue is, they just want it fixed. Likewise, when a business has invested heavily in using expensive AI chips to increase the performance of its AI apps, but they aren't performing as expected, knowing why is of little comfort, the business will want it fixed so that it realises the outcomes it has invested in achieving.

The next problem

You can upgrade servers, chipset or procure more compute power almost instantaneously if they are not performing. Upgrading a network can take months–longer if it's a WAN spread across multiple locations.

46%

Nearly half of successful AI proofs-of-concept fail to make it to full production.[1]

You might consider how we occasionally find a glitch on a video conference when working from home, usually down to the performance of a lower-grade internet backbone you are connected to. If you are using the internet to transport critical AI data, you may have little or no visibility of routing issues or bottlenecks that are compromising the performance of your AI applications—see the paper mentioned below for more on this topic. All the effort you put into optimising your AI app and investing in specialist hosting using custom chipsets could be for nothing if network performance is poor and resolution of network issues is troublesome and time-consuming.

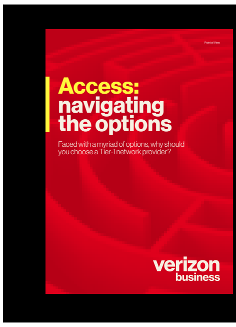
The more data that goes into the AI model, the more compute power and storage is needed for it, but what goes hand in hand with this is that more and more high-quality networking is required. You can start small, but then every time you want to do something, you must make sure the networking is able to support the workloads that your applications need.

It's important to carry out thorough due diligence on your network providers. Ask each potential provider:

- Are you directly connected to a Tier-1 backbone at the core of the internet? Or do you just provide access to the edge of the network and then leave the routing of data to chance?
- Are you directly connected with leading hyperscalers like AWS, Microsoft Azure and Google Cloud Platform?
- Do you offer latency and performance guarantees?
- Do you have the capability to prioritise specific applications, like many AI use cases, that are sensitive to latency?
- Can you provide end-to-end visibility of performance—from the customer site to the hyperscaler?

For peak performance, the network must perform as effectively as the high-performance chipsets in your data centre or cloud operator.

Verizon 's global network is built to support AI applications. Our experts can help you design and build the appropriate network infrastructure to help make your the performance of your AI applications meet your needs and your initiatives successful.



Point of View

Access: navigating the options

Faced with a myriad of options, why should you choose a Tier-1 network provider?

verizon business

Access: Navigating the options

This other paper from this series gives insight into how the choice of provider can have a huge impact in terms of performance and outcomes for your business. This is especially important for latency-sensitive applications, such as many AI use cases.

15-minute read

Verizon's AI-enabled multi-service backbone

We continue to invest in our network. We have invested $176 billion since 2000 and anticipate spending a further $17.5–18.5 billion in the current fiscal year. Verizon's networks are operated across our global multi-service backbone (MSB), which enables granular prioritisation of applications, including the whole internet network.

This enables us to provide service-level agreements (SLAs) with latency guarantees and contracted performance characteristics across the internet. That's right, we offer internet access designed and operated to be in lock-sync with your AI compute power in the cloud.

And with our extension into the advanced operating systems of vendors such as Cisco, Juniper and Aruba, this control can be extended to the edge of the network with managed routing and LAN technologies. Our network also incorporates its own AI engine to dynamically and intelligently move its resources between cloud operators as your applications shift their demand for compute power between operators.

We have built capabilities into the core of our MSB that enable us to prioritise latency-sensitive applications and provide meaningful SLAs. We're able to offer these "latency-reducing performance boosters" due to a combination of:

- Owning and operating our own infrastructure
- Our ability to load balance across cloud operators
- Our direct peering with cloud operators

Earlier I noted how Meta, Google, Amazon and Microsoft have developed their own custom chips optimised for AI. This is a substantial investment designed to cut latency and improve performance. These businesses also consider their network backbone essential to their AI success. They have also invested in gold standard network architectures to help ensure the success of their AI offerings.

Verizon employs many highly experienced network specialists, including some of the world's leading experts in the field. They are working to continually develop our global networks to not just keep pace with demand, but be ready for the demands of future applications, too. We're facilitating and enabling applications to perform in a way that AI platforms demand.

Through the latency-reducing performance boosters built into the fabric of our global architecture — we can deliver a fast, secure and resilient network that can help alleviate the performance headaches and accelerate AI applications.

Let's talk

Our teams of specialists are at the ready to provide you with the expertise you need to help make your AI applications work at their best and make your initiatives successful.

From IoT to real-time analytics and AI, Verizon can help you stay at the forefront of using technology to monitor, manage and improve operations. If you still have questions after reading this paper, get in touch with us at:

verizon.com/business/en-gb/contact-us

The Network Procurement series

This paper is one of a series exploring the growing demands on enterprise networks and important questions companies should ask during the procurement process to help ensure that the solution they chose are truly enterprise-grade and will meet their current and needs.

Something big is coming

Data | IoT | AI

This paper explores some of the key drivers behind the explosive growth in the volume of data enterprises are gathering and what that means for network planning.
verizon.com/business/resources/articles/iot-genai-data-explosion.pdf

Access: navigating the options

Performance

There are many decisions to make when buying networking. Understanding the three tiers of the internet is critical to thoroughly evaluating the options. This paper explains what they mean for network performance and security.
verizon.com/business/resources/articles/tier-1-isp-enterprise-connectivity.pdf

Network peering

Cloud | Performance | Reliability

Peering is fundamental to network performance and consequently enterprise applications, particularly ones based in the cloud. Despite this, it's rarely discussed during procurement. Read this short paper and put that right.
verizon.com/business/resources/articles/network-peering.pdf

Are you in the dark about performance?

Data | Performance | Manageability

Read this paper to learn how the decision to split the procurement of physical (underlay) and logical (overlay) networks can affect network performance, visibility and manageability.
verizon.com/business/resources/articles/overlay-underlay-network-procurement.pdf

Better together

Security | Performance | Manageability

Cyberthreats continue to grow in volume and sophistication. This short paper offers six reasons to consider greater integration between cybersecurity and networking to improve protection while reducing workload and cost.
verizon.com/business/resources/articles/unified-network-security-services.pdf

Supercharge your AI applications

AI | Performance

Artificial intelligence (AI) promises to be the most disruptive technology since the internet became mainstream around 30 years ago. This paper explains why network performance is critical to the performance of many AI applications and realising the anticipated benefits.
verizon.com/business/resources/articles/network-infrastructure-ai-platforms.pdf

1. Verizon, Deploying AI at scale, 2025

verizon
business