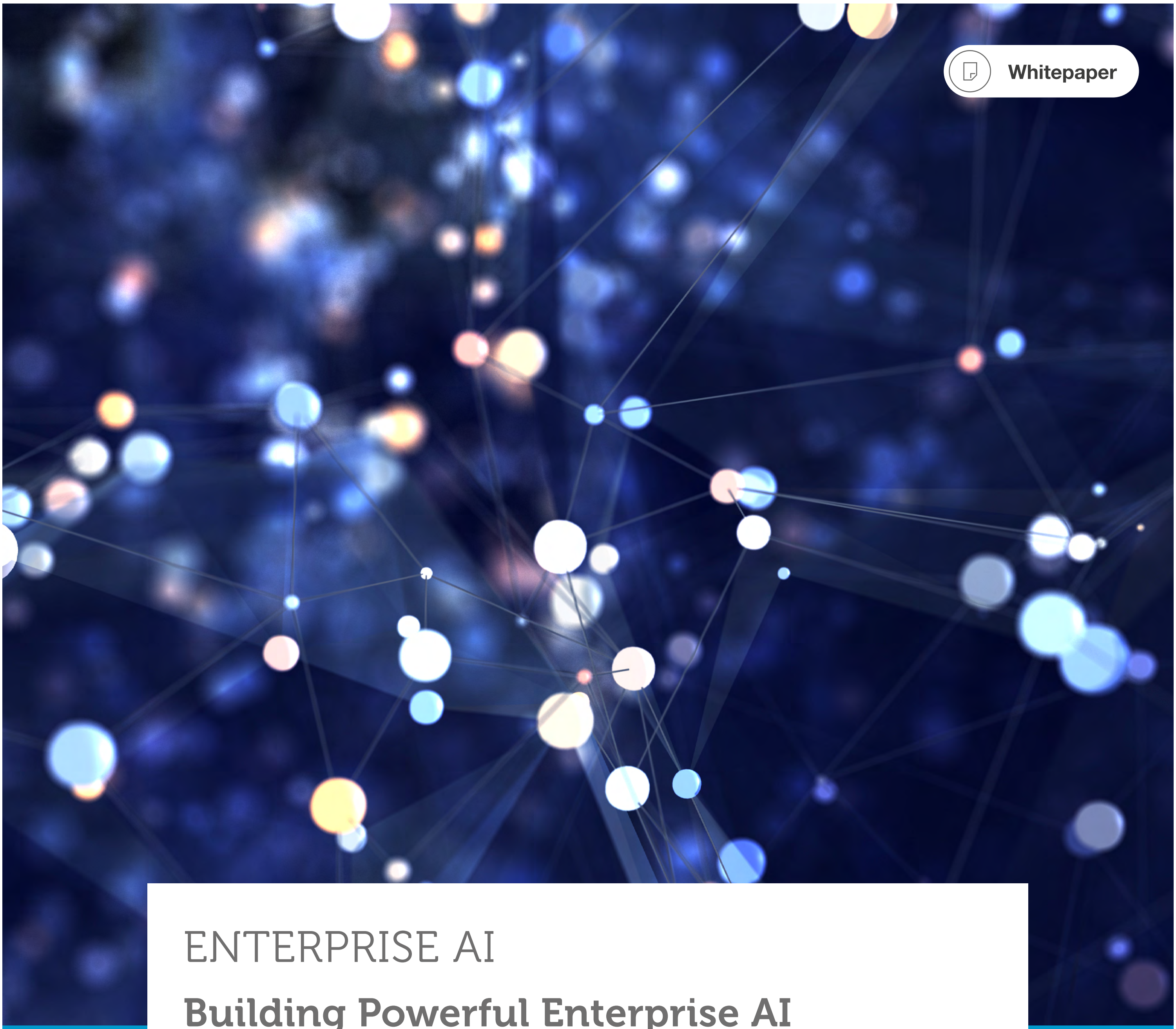
interxion
A DIGITAL REALTY COMPANY



Whitepaper

ENTERPRISE AI

Building Powerful Enterprise AI Infrastructure: How to design enduring infrastructure for AI

Patrick Lastennet, Director of Enterprise

Bryan Hill, Director of Platforms

In partnership with NVIDIA, Interxion offers everything enterprises need to develop and scale their AI infrastructure, from advanced power and cooling technology to access to a thriving community of connectivity and cloud providers.

THE BENEFITS OF AI/DL FOR THE ENTERPRISE

Enterprises today are increasingly gaining value from artificial intelligence (AI) and deep learning (DL) applications that, whether they're accelerating innovation, increasing efficiency or improving their bottom line, leverage automation to extract useful patterns from data. As a subcategory of AI, DL is a cutting-edge technique conducted through the use and optimisation of neural networks. Even though some of these ideas have been around for a long time, the past decade has seen a major breakthrough in the application of neural networks, due to the digitisation of information, advanced tooling, and the availability of hardware that enables large-scale computation.

AI has become a game changer in highly competitive markets like healthcare, academic research, supply and logistics, robotics, financial services, media and entertainment, retail, and autonomous vehicles (AVs). By automating repetitive processes, delivering new strategic insights, and accelerating innovation, AI has the power to revolutionise business.

Take a look at the world of AVs, for example. In order for AVs to become a reality, the software needs to be able to make crucial, real-time decisions. The more data analysed, the safer those decisions – and therefore, AVs – will be. But until the recent application of AI and DL, data analysis wasn't quick enough or powerful enough. Powered by DL and supported by AI-ready infrastructure, software is making AVs not only possible but smarter and safer than ever before.

Across all industries, enterprises adopting AI are already seeing the value it brings, from increased productivity and efficiency to improved customer experience. Early adopters aren't only gaining value from the technical implications, however. They're also gaining a competitive business edge over companies not yet incorporating AI.

AI adoption helps 44 percent of enterprises reduce costs, particularly in areas like manufacturing and supply-chain management. From predictive maintenance to logistics network optimisation, AI provides the analysis enterprises need to realise savings. AI also helps accelerate time to market, by advancing R&D at a faster pace.

At the same time, 63 percent of enterprises report increased revenue as a result of incorporating AI. Marketing and sales teams generate new revenue through AI use cases like customer service analytics, prediction of likelihood to buy, and pricing and promotion. On the product development side, creating new AI-based products and enhancements also brings new opportunities for revenue.

As AI use continues to evolve, making sure your enterprise infrastructure is also evolving to properly leverage the technology will be crucial to future success.

40%
of enterprises agree that lack of IT infrastructure is the primary barrier to AI implementation

89%
of enterprises expect the volume of data used in AI workloads to increase within the next year

45%
of enterprises say their current infrastructure is not capable of meeting future demands for AI workloads

THE AI INFRASTRUCTURE CHALLENGE

Building the infrastructure needed to support AI deployment at scale is a growing challenge.

Traditional AI as machine learning (ML) doesn't necessarily require a ton of data. But with DL and the emergence of IoT/5G, there's going to be a huge amount of data generated from factories, smart cities, driverless cars, edge devices, and so on.

Designing an infrastructure capable of leveraging that data for AI is complex. Decision makers need to understand all AI testing and deployments happening in their company in order to choose the infrastructure that provides the scale and performance needed for the long-term. There are significant costs – time, money and resources – involved in having to rearchitect or move the AI deployment, so it's important to get the infrastructure right from the start.

What are the requirements of an ideal infrastructure when it comes to leveraging the growing volume of data and enabling AI at scale?

Accessing the data

Interconnection is key, especially as some AI applications move to the edge. High connectivity enables enterprises to bring data from the edge into data centres and send models and data back to the edge to optimise inference. This entails:

- **Proximity to 5G core nodes in data centres:** These nodes bring back data from devices in the field.
- **Proximity to Fixed line core nodes:** The nodes bring back data from fixed points, from offices to manufacturing facilities.
- **Direct cloud access:** Some workloads and use cases will be optimised for cloud, and this should be managed in a secure performant way.
- **Enterprise data transfer:** Connectivity hubs draw in enterprise data for processing.
- **Geographic scale:** Scalability allows enterprises to enable AI workloads in different locations.

Computing and processing the data

Accessing data is just the first step of enabling AI, however. Once the data enters the data centre, here's what's needed to support computation for training of models:

- **High density support:** The data centre must be able to handle high density now, but also in the future – well beyond 15 kW/rack, which is the declared maximum for most enterprises. In fact, 60% of enterprises say their highest density rack is running at 15 kW or less, and they're unable to manage more. NVIDIA has already developed architectures than run beyond 40 kW/rack and as technology evolves, densities will accelerate over the next few years.
- **Size and scale:** Key to leveraging the benefits of AI is doing it at scale. The ability to run at scale hardware (GPU) enables the effect of large-scale computation.

ENABLING AI WORKLOADS WITH A POWERFUL, HOLISTIC INFRASTRUCTURE

Based on these requirements for AI infrastructure, there are a few options for deployment.

An on-premises solution may be relatively inexpensive, but there are significant challenges when it comes to power density and scale. Most enterprise data centres simply aren't capable of handling the scale required to leverage the benefits of AI.

Public cloud, meanwhile, offers the path of least resistance, but it isn't always the best environment to train the models at scale or deploy them in production due to either high costs or latency issues. Training models at scale requires compute to constantly process and reprocess large data sets over a long period of time. High utilisation such as this, coupled with data egress, is not an optimum use case for public cloud, although public cloud will certainly play a part in an enterprise AI world, especially for test/dev and inference.

The importance of reference architectures

Many companies try and fail by going the "build your own" way. A fully-supported hardware and software stack is essential in keeping the most expensive resource – the data scientists – doing data science, not data-centre-maintenance and support. This is where NVIDIA DGX™ systems and its reference architectures become key success factors.

There are important challenges outside of Compute, Storage, Fabric, and Software, such as tools for optimising the resources (orchestration/scheduling). These challenges need to be addressed, not only in the basic infrastructure plan but also when it comes to data management and data-pipeline. Data science expertise and valuable expertise within project management for AI-projects is also critical for success.

Hyperscalers lead the way

Instead of building your own, take a look at how platforms that are already gaining value from AI have chosen to deploy their infrastructure. Hyperscalers like Google, Amazon, Facebook, and Microsoft have been successfully deploying AI at scale for their own use for some time now, with their own core and edge infrastructure often deployed in highly connected, high-quality data centres. They use colocation heavily around the globe because they know it can support the scale, high-density and connectivity they need.

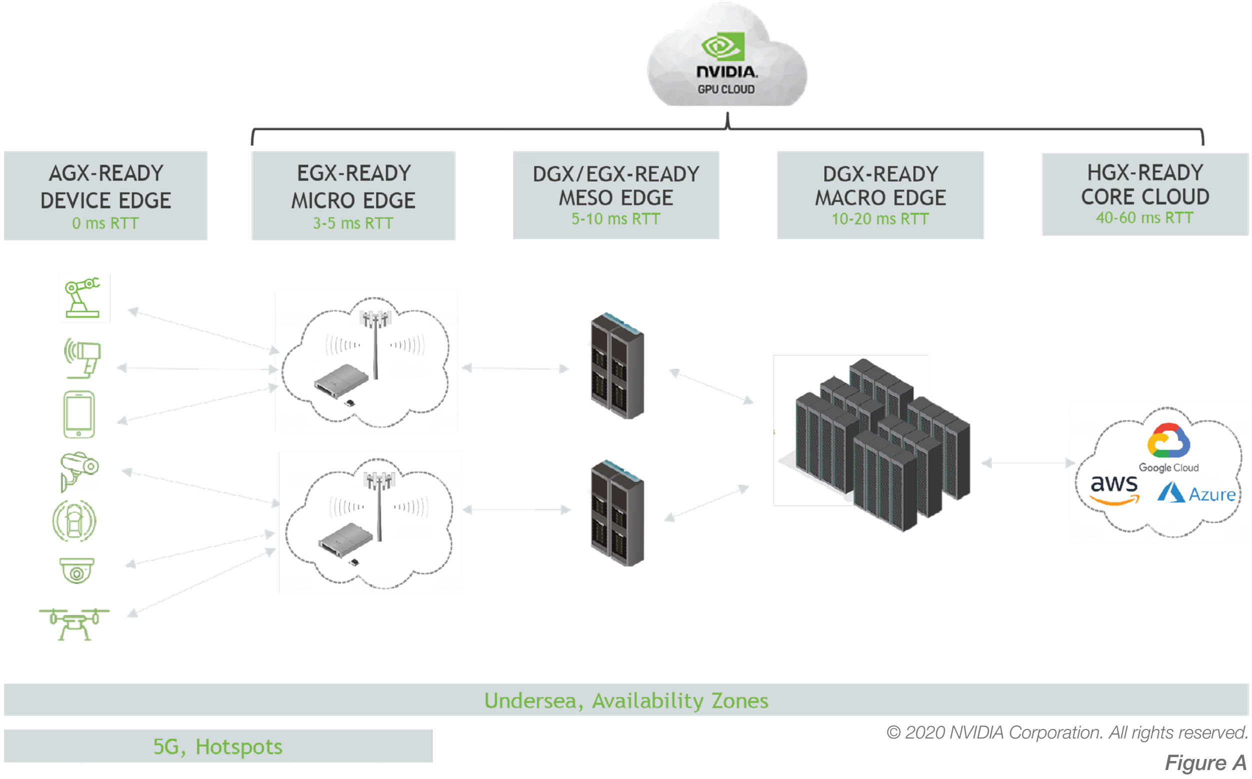
By leveraging the knowledge and experience of these AI leaders, enterprises will be able to own their own destiny when it comes to AI. Inference happens at the edge, but some of that data needs to come into core data centres that are capable of supporting AI and DL at scale. It's time for enterprises to take a page out of the hyperscalers' playbook and invest in colocation.

PARTNERS WITH PROVEN, REAL-WORLD AI EXPERIENCE

There are very few companies that have built AI infrastructure at scale in the real world, however. Proven experience building AI infrastructure is a key factor for a successful AI project, which is why Interxion has partnered with providers like CGit, Scan UK, A.P.Y., and Altair Engineering.

For example, CGit, an AI infrastructure leader and NVIDIA Elite Partner, has designed, implemented and provided support to some of the largest AI infrastructure projects in Scandinavia, spanning industries like automotive, medical research and telecommunications. CGit is also recognised as the Core Technology Partner to AI Innovation Of Sweden, Sweden's national AI-centre for applied AI research and innovation.

NVIDIA launched its DGX-Ready Data Centre Program to help enterprises solve the challenges involved in building and deploying an AI-ready infrastructure. Through the programme, NVIDIA offers a qualified network of AI-ready colocation partners for enterprises running AI infrastructure built on NVIDIA DGX systems. Designed to meet the demands of AI and analytics, the programme helps enterprises accelerate their AI mission and see ROI right away. It's the ideal solution for enterprises who lack the expertise or resources to build a future-proof data centre for GPU compute at scale.


NVIDIA GPU CLOUD
AGX-READY DEVICE EDGE 0 ms RTT
EGX-READY MICRO EDGE 3-5 ms RTT
DGX/EGX-READY MESO EDGE 5-10 ms RTT
DGX-READY MACRO EDGE 10-20 ms RTT
HGX-READY CORE CLOUD 40-60 ms RTT
Google Cloud
aws
Azure
Undersea, Availability Zones
5G, Hotspots

© 2020 NVIDIA Corporation. All rights reserved.

Figure A

Figure A, above, from NVIDIA shows the ideal AI topology from edge-core and outlines the key elements enterprises should consider in order to deploy successful AI initiatives.

Inference will happen at low-latency within the edge device or at an Enterprise's manufacturing facility. The Meso and Macro edge data centres are where deep learning (DL) and training will take place for model optimization.

In order to create a seamless enterprise edge-core AI workflow, it is clear that these data centres will need to fulfill certain criteria:

1. High density support to enable enterprises to take advantage of the latest high performance GPU cluster architectures

2. Dense interconnection and connectivity

    a. Presence of carrier connectivity to transfer large amounts of data from the Enterprise edge into the DL clusters and to transfer optimised models back to the fixed edge

    b. Presence of mobile connectivity providers/4 and 5G core nodes to transfer data and models from and back to the mobile edge

    c. Direct connections to all the public cloud providers to enable fast, secure, performant and cost-effective hybrid cloud AI environments

3. Large highly distributed data centres that can provide both the infrastructure scale that enterprises will need within the latency requirements outlined

Therefore, enterprise AI architects and CIOs need to carefully consider their data centre choice against these criteria to ensure their AI workflows are future-proof from the beginning.

AI-READY DATA CENTRE SOLUTIONS

Because most enterprises don't have the facilities to build the powerful, connected, highly-performant infrastructure needed to support accelerated computing operations, they choose to partner with third party colocation data centres instead.

Interxion is an AI-ready colocation data centre partner for NVIDIA DGX systems, helping enterprises who are held back by the challenges of facilities planning.

Together with NVIDIA, Interxion provides the scalable and flexible environments enterprises need to develop and scale their AI programmes.

Here's how Interxion solves some of the major IT infrastructure challenges.

Support for high-performance computing (HPC)

Interxion supports high density workloads with Tier3+ Resilient facilities that already serve all the major hyperscale platforms. Interxion is a certified DGX-Ready Data Centre Partner.

Scale

Interxion operates 53 data centres in 13 cities in 11 countries, with nearly 300 MW of equipped capacity across its footprint.

Connectivity

Interxion hosts all the major ISPs where 5G core nodes will reside and has over 700 connectivity partners available across its data centres. In addition, Interxion provides an end-to-end AI topology with all the private interconnection points for the clouds.

Working with Interxion enables enterprises to focus on the activities that drive value for them – gaining insights from their data and innovating with AI. Powered by Interxion and NVIDIA DGX systems, enterprises can drive their AI initiatives forward.

Learn more

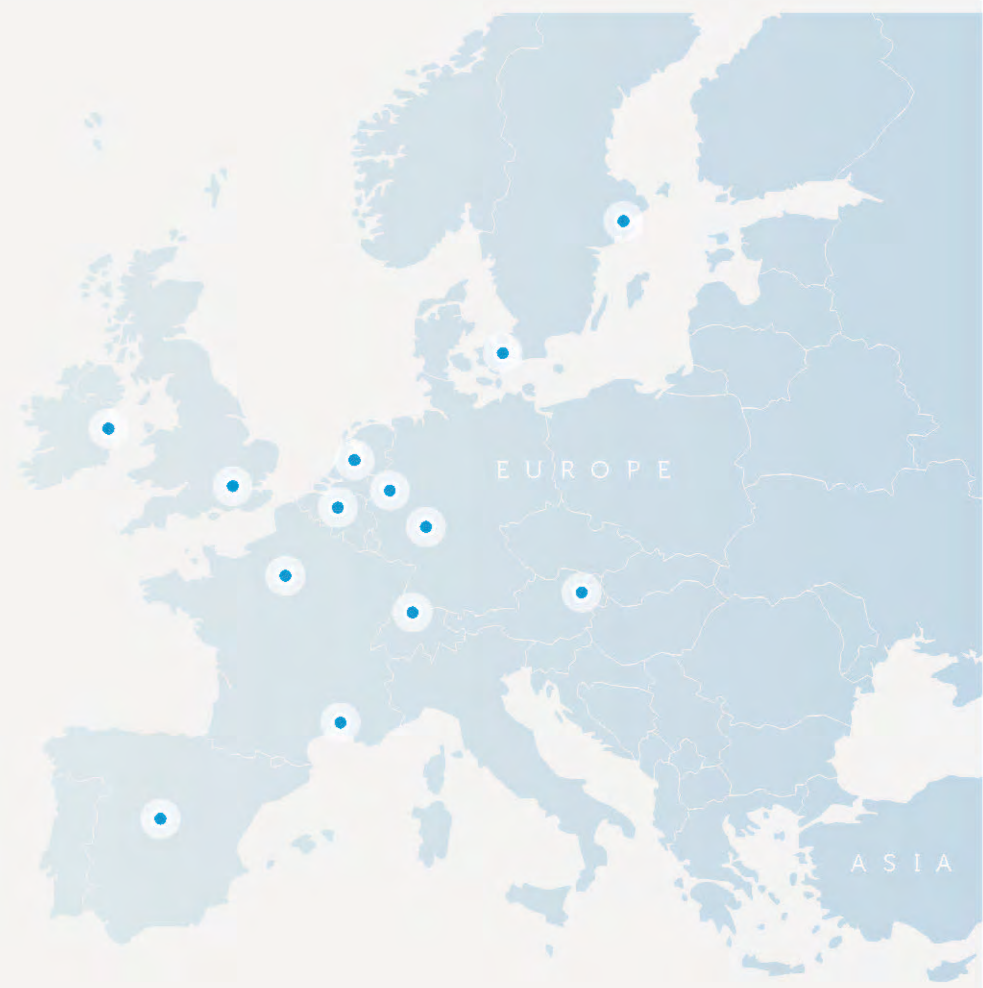
To find out how Interxion helps enterprises accelerate their AI mission, visit www.interxion.com

interxion™
A DIGITAL REALTY COMPANY

About Interxion

Interxion, a Digital Realty company, is a leading provider of carrier- and cloud-neutral colocation data centre services in Europe, serving a wide range of customers through more than 50 data centres in 11 European countries. Interxion's uniformly designed, energy-efficient data centres offer customers extensive security and uptime for their mission-critical applications. With over 700 connectivity providers, 21 European Internet exchanges, and most leading cloud and digital media platforms across its footprint, Interxion has created connectivity, cloud, content and finance hubs that foster growing customer communities of interest. For more information, please visit **www.interxion.com.**

Data Centre services across Europe



www.interxion.com
customer.services@interxion.com

International Headquarters
Main: + 44 207 375 7070
Email: hq.info@interxion.com

European Customer Service Centre (ECSC)
Toll free Europe: + 800 00 999 222 / Toll free US: 185 55 999 222
Email: customer.services@interxion.com

Cofounder: Uptime Institute EMEA chapter. **Founding member:** European Data Centre Association. **Patron:** European Internet Exchange Association. **Member:** The Green Grid, with role on Advisory Council and Technical Committee. **Contributor:** EC Joint Research Centre on Sustainability. **Member:** EuroCloud.

Interxion is compliant with the internationally recognised ISO/IEC 27001 (537141) certification for Information Security Management and ISO 22301 (BCMS 560099) for Business Continuity Management across all our European operations. © Copyright 2020 Interxion.