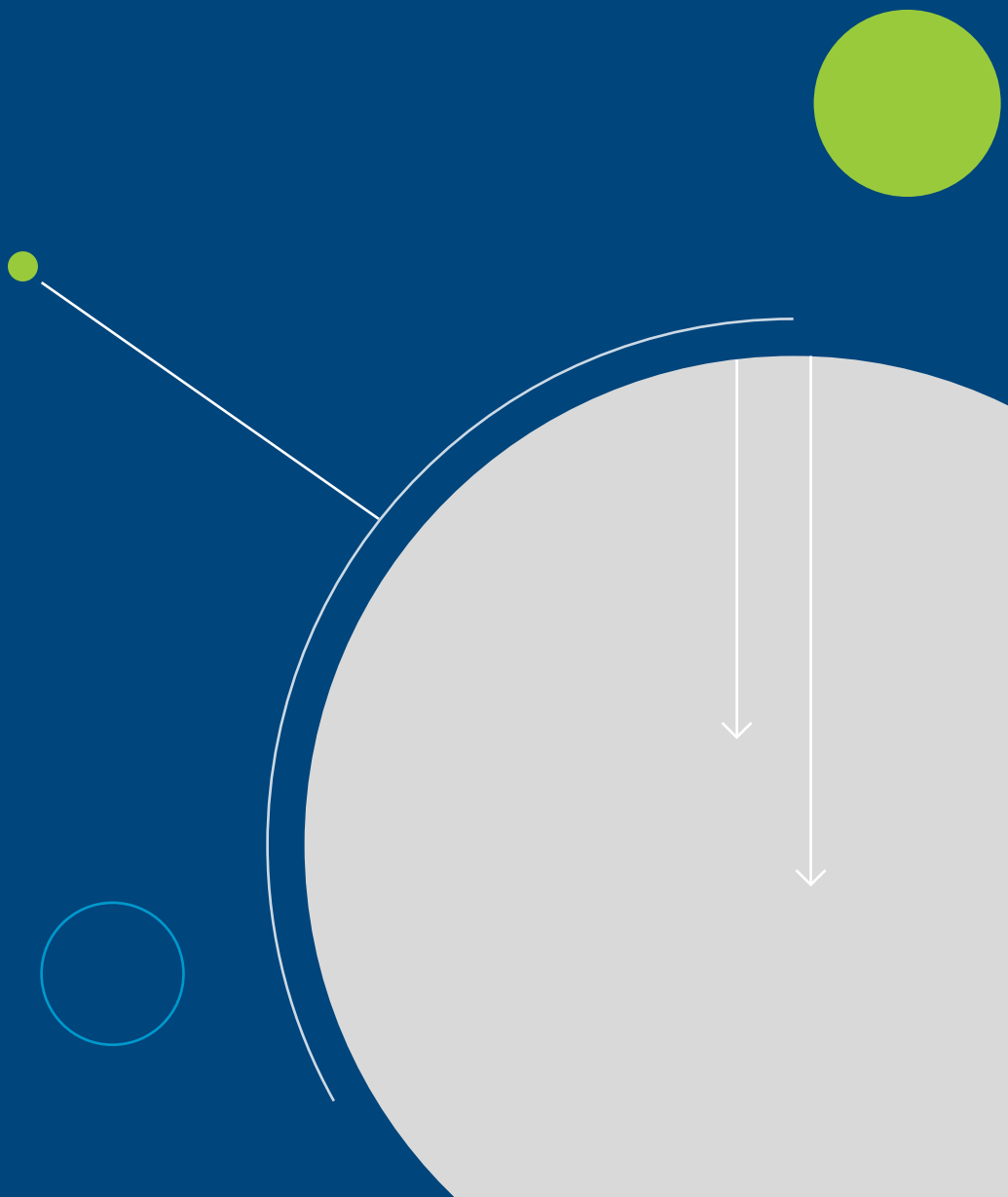
interxion™

Success in AI, Deep Learning and HPC requires a different kind of data centre



Success in AI, Deep Learning and HPC requires a different kind of data centre

Artificial Intelligence (AI), deep learning (DL), machine learning (ML), and high-performance computing (HPC) have been growing in popularity over the last five years. Unlike many areas of computing, these three are not only bound together but have very specific requirements. Organisations looking to build their own cloud services delivering these technologies need to think carefully about the technical challenges that they present.

What do AI and ML need?

The connection between these three technologies is not accidental. AI and ML have been subjects of discussion for at least four decades. Standing in the way of delivery was the current state of computing technology.

To be effective, AI and ML both require large amounts of data. This is not about megabytes, gigabytes or even terabytes. The more data that can be ingested, the more effective the solution. This means that organisations are looking to work with hundreds of terabytes or, in some cases, exabytes of data. A further complication is that the data often comes from multiple sources. This requires a diverse access capability to upload and maintain the data sets.

Once ingested, the data has to be collated and acted upon by the software. Traditional computing which relies on CPUs is limited in the number of cores and threads it possesses. The more it has of each, the more operations it can carry out in parallel on the data.

Enter the Graphics Processing Unit (GPU). The GPU is made up of a very large number of cores. Each core can deal with multiple threads. More importantly, the GPU is capable of doing extremely large amounts of parallel processing, far more so than a CPU. Look at the two largest supercomputers today, Summit and Sierra. Both of these rely on massive arrays of GPUs that process vast amounts of data.

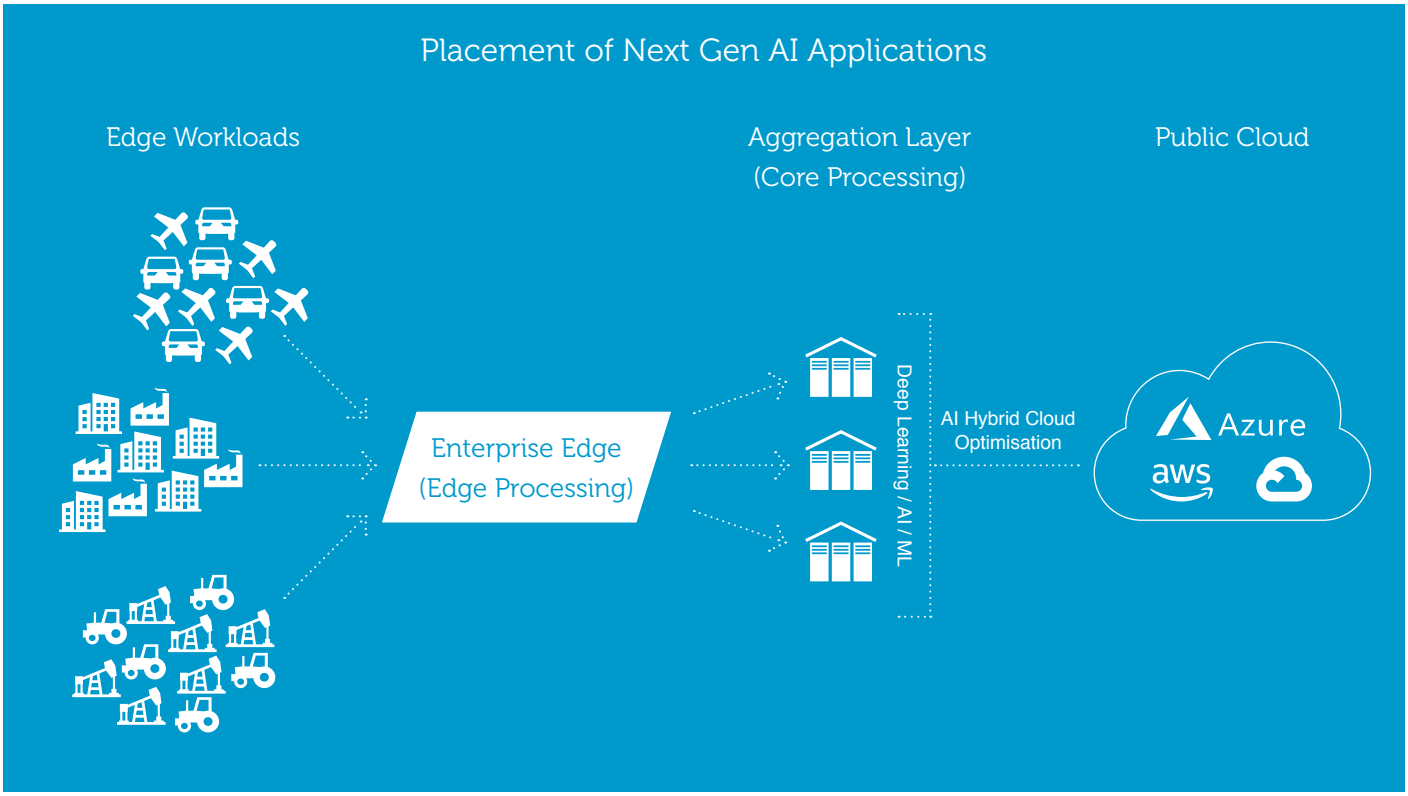
Connectivity – the non trivial challenge

To get the data into the data centre where the AI and ML is located is no trivial task. If the data is coming from a large number of disparate sources, think connected cars, IoT and IIoT, then connectivity must scale to cover the data being acquired. 5G will be a key enabler of an IoT/AI driven world. Edge processing will involve hyper local compute at the cabinet, container or mobile mast locations.

Telecommunications providers will then connect these back over terrestrial links to their ‘core’ nodes which exist in highly connected data centres. This is where core processing will exist and (where required) will securely connect into the major public cloud environments over private access nodes like Google Cloud Interconnect, AWS Direct Connect, and Microsoft Azure ExpressRoute.

Once data has been imported it needs to be accessed by applications. The location of the data centre and its communications links are critical here as well – as is the proximity to cloud providers who will handle the bulk of the compute. As with ingesting the data, latency can be key to the use of the data. Autonomous vehicles will be creating and using large amounts of data to be safe, which requires near real-time processing at the edge. However, this data also needs to be sent to the core processing center several times a day as well where all the aggregated data from thousands of vehicles can be inputted into the models and refined for better outcomes.

Due in large part to the unique processing needs, we expect to see communities of interest forming around the AI and ML which all come together in the data centre.



Why is HPC a foundation for AI and ML?

HPC has a long history of using both GPUs and working with very large amounts of data. It is used to understand fluid dynamics in automotive, aerospace and aircraft design. Drug companies use it to map the interactions and behaviour of compounds, as do chemical engineers.

The evolution of computer architectures to support more GPUs and allow them direct access to memory has changed HPC. It has created a platform on which AI and ML can function. Those same organisations that have used HPC for modelling are at the forefront of adopting AI to improve their ability to deliver new solutions.

For many of these organisations, the solutions are all run inside their own data centres. There are reasons for this, not least the colocation with massive data sets but also the security of the systems. These hold highly sensitive commercial data which has to be protected. However, as the demand for greater use of AI grows, there is an increasing interest in specialist cloud providers capable of delivering HPC and AI.

The challenge of building a data centre to deliver AI, ML and HPC

There are a number of challenges, some already mentioned, to building an HPC system capable of supporting AI and ML. For cloud service providers looking to build specialist services in this space, it is important to ensure that your chosen data centre partner can support these six requirements.

Power:
Racks used for HPC require far more power than those supporting general purpose computing. The high-density rack environments support this such as NVIDIA's reference architecture DGX1 and DGX2 PODs to requires up to 35kw per rack and multiples of this. This means that any facility in which an HPC and AI cloud is built must be able to support this level of power in terms of delivering this across the whole rack footprint. Traditional on-premise enterprise data centres are not built to support this, and aren't cost effective environments for such deployments either.

Cooling:
Increase the power consumption and you increase cooling requirements which also consume power. It is important that the facility is capable of delivering the cooling demands of high-dense HPC racks, otherwise failure rates will rocket. Tier3+ or equivalent data centres offer uptime SLAs of 99.999%.

Cloud Connectivity:
AI consumes vast amounts of data. Much of that is likely to be stored in systems around the world. Having multiple direct connections to the large public cloud providers to allow customers to move their data to the HPC and AI solution is critical. This is more than just moving data from Systems of Record and Systems of Engagement. It includes new platforms such as IoT that are gathering very large qualities of data on a daily basis.

Access networks:
Getting data into the AI solution either as part of a private or hybrid cloud is just part of the demand. Customers want to access their data from wherever their staff are located. This means access to local points of presence that can support both high-speed data transfer and encrypted data transfer.

Security:
The sensitivity of the data being stored, used and generated by HPC and AI is critical to those customers using the service. High levels of both physical and IT security are necessary, ISO27001 is a start point.

Conclusion

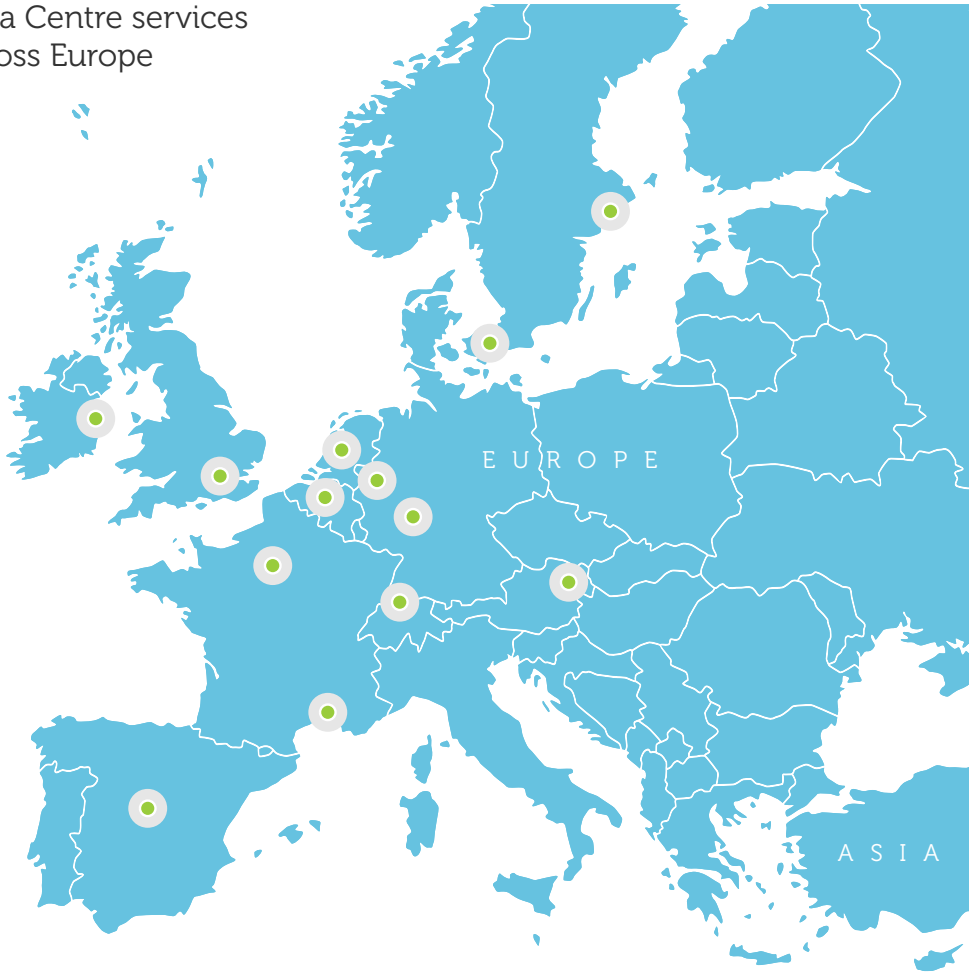
AI is finally becoming accessible to most organisations. This is starting to drive demand for public cloud AI services as well as hybrid and private cloud offerings. Choosing a data centre partner, especially in Europe, who can provide all of the key requirements is no small feat. There are plenty of smaller providers who can provide some but not all of these.

As the race heats up among digitally oriented enterprises to develop AI/Deep Learning for competitive differentiation, choosing the right future proof and flexible locations for dedicated and hybrid cloud environments is a critical part of the CIO decision-making process. Making the right choice of data centre partner will be the difference between success and failure.

About Interxion

Interxion (NYSE: INXN) is a leading provier of carrier and cloud-neutral colocation data centre services in Europe, serving a wide range of customers through over 45 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