VERNE

Artificial Intelligence in the data center

Anticipating infrastructure requirements

Artificial Intelligence is everywhere

Artificial Intelligence (AI) is a key component of the digital transformation journey upon which many organisations have embarked.

Fundamental to all AI applications is the ability to capture, process and understand huge data sets. AI can analyse data faster than humans, therefore, improving overall project management performance. The resulting information, whether produced in near real-time or over a longer period, provides the basis for making crucial business decisions. And these decisions can be made much faster and smarter using AI than by relying on traditional, human-intensive data logistics methods. So, where is AI making an impact in the business world? And, importantly, what impact will the AI 'explosion' have on the data center and IT infrastructure on which it relies so heavily?

The simple answer to the question 'Where is AI making an impact in the business world?' is: EVERYWHERE.

According to Mckinsey's 2022 State of AI report, AI adoption globally is 2.5 times greater than in 2017 with AI capabilities predominantly focused on ways to optimise service operations. Bottom-line value realised from AI is strong with the biggest reported revenue effects are found in marketing and sales, product and service development, and strategy and corporate finance, with the highest cost benefits from AI in supply chain management.

With these business benefits likely to only exponentially increase over time, we are going to continue to see rapid adoption of AI in the workplace.

AI: from general to niche applications

AI is being used across all organisations, no matter their size or industry sector. From project management to customer service AI is being used to offload repetitive, tedious and relatively low-value tasks from humans to machines, such as checking online applications and/or order forms, for example. AI can be used across the finance department to give expense claims the once over to spot any anomalies or unusual items. Similarly in HR, AI can automate the keeping and updating of personnel records. In all these cases, the suggestion is that AI will perform these tasks at least as well as humans, and at a significantly lower cost.

This ebook's remit does not cover the ethics of AI, nor the necessary evaluation process as to how customers will react when their human contact is replaced by a robot. Nonetheless, organisations should be aware of the overall AI ecosystem to avoid unpleasant surprises, like workforce resistance, customer dissatisfaction, etc., along the way.

Where AI is undoubtedly a win/win for the enterprise is in the performance of tasks which are, quite simply, too complex, large and/or expensive for humans to undertake. Additionally, speed can be a crucial factor. For example, research tasks which might take humans months or years to perform, can be carried out in minutes, hours or days. The result is products and services that are brought to market far faster than was previously possible – great for commercial organisations and the health sector, for example. Customers can also be offered highly (hyper) personalised engagement in real time on the high street – helping bricks and mortar businesses compete with their online competition.

Sectors leading AI adoption

Finance

Customer Success

Healthcare

Media

Peptone is accelerating protein-based Drug discovery by using AI to analyse the anomalous behaviour of proteins and identify the most desirable protein variants for further drug development research.

This approach helps scientists identify the right proteins quickly and is a more time and cost-effective method over wet lab approaches.

For more information visit Verne's case study library.

AI and the data center

AI can monitor, analyse and make continuous recommendation on improvements that can be made to the mechanical and electrical aspects of a data center.

AI can learn about a data center environment and, as and when a fault occurs, the circumstances which led up to such a fault can be analysed and stored.

AI is already having an impact on the data center in two crucial ways. Firstly, AI is being harnessed to improve the performance of data centers. Secondly, data center design is changing, and will continue to evolve, to ensure that it can provide the optimum environment in which to host AI applications.

Let's examine the possibilities of AI, in terms of how it can help to improve the performance of a data center in more detail. There are plenty of software programs that can help humans design a data center to try and ensure optimum efficiency in terms of space, utilisation, energy consumption, overall build cost and the like. However, many of the figures upon which a data center design is based are, from day one of the actual running of the data center, inaccurate. Assumed workloads, predicted temperatures (both inside and outside of the building), for example, will almost certainly not be the same as the actual operating numbers.

Furthermore, a data center is, now more than ever, a dynamic, 'living' building, where the only constant is, or should be, change. For one or more humans to continuously optimise such an environment, with so many variables, is all but impossible. By the time the various calculations have been carried out, and a slight tweak to the data center layout is recommended, a new set of variables will have been introduced as more cabinets and racks are installed and populated. AI can monitor, analyse and make recommendations continuously on improvements that can be made to the mechanical and electrical aspects of a data center, not to mention the provisioning and maintenance of the critical IT kit (servers, storage, networks) which it houses.

Additionally, thanks to the underlying machine learning and neural networks, AI can learn about a data center environment and, as and when a fault occurs, the circumstances which led up to such a fault can be analysed and stored. Analysed so that they can be avoided in future; stored so that, if the same set of circumstances does occur in the future, this can be flagged up before the actual fault occurs, and maintenance carried out to avoid the fault. Such an AI tool is powerful enough in a single data center, but imagine its impact across a substantial data center estate. All the data centers can learn from each other and help to continuously improve their reliability and performance.

Some data facilities already operate in a 'lights out' environment – with very little human intervention. These are already leveraging AI to ensure smooth, optimised operations. Looking to the future, such AI-enabled data center monitoring and management will become the default standard.

And there's also talk of designing small footprint, high rise data centers, where the racks and cabinets will be stacked up to the roof. Equipment installation, replacement and maintenance will be in the hands of robots. Automated, high-bay warehouses have been in existence for a long time and it's not difficult to imagine similarly styled data centers becoming more common over time.

Optimisation is the key word when it comes to the data center and AI. And that's not just the optimum design, operating and maintenance environment of the facilities. AI has a role to play when it comes to improving the performance and usage of the servers, storage and networking infrastructure housed within the data center. Not to mention the security. AI can examine, profile and recognise cyberattacks at a speed and scale that's impossible for humans to achieve.

Some data facilities already operate in a 'lights out' enviroment - with very little human intervention. These are already leveraging AI to ensure smooth, optimised operations.

AI can examine, profile and recognise cyberattaks at a speed and scale that's impossible for humans to achieve.

AI demands a data center design change

As well as having a role to play within the data center, AI is set to make its mark on data center design. Most obviously, underpinning virtually all AI activities are increasingly large, complex data sets. The infrastructure needed to house, transport and analyse this data needs to be able to cope with the demands of speed, agility and scalability.

Speed because data needs to be processed in real-time in many applications, and so that tasks that previously took weeks and months can be performed in hours and days.

Agility because the demands made on a data center will vary as new applications need developing, old applications are retired, workloads are moved around in a virtual environment – everything designed to give the optimum performance for any application.

And scalability because, when a project is started, there's no telling how big it might grow in terms of data volume. And having to migrate an application and its data 'foundation' because you've run out of compute power and/or storage and/or networking capacity, is not a great scenario.

Armed with these requirements, more and more organisations are realising that, at some stage, they will need to invest in and/or have access to IT and facilities hardware that can meet these demands for speed, agility and scalability. With the prospect of faster, hotter servers leveraging technologies such as Graphical Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs), data centers will need to be able to cope with this significantly increased heat – not to mention even higher density environments!

Faster, smarter servers will require faster, smarter storage and networks. The connectivity vision is that data centers will need to be super-efficient, highly fibre-meshed, ultra-low latency, east-west spine and leaf networks, capable of supporting not just production traffic, but also supporting the machine learning training part of AI. Faster storage could well mean more and more flash-based systems.

Agility: Development of a new applications need to be fast and resources need to be available almost immediately
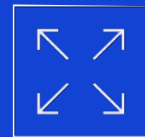
Scalability: Flexible and temporary access to data center and IT resources

Resiliency: Customers and potential customers expect real-time access to AI and VR and any application

Speed: Connectivity and application location for the appropriate infrastructure are crucial

Colocation provides an AI-ready environment

The move to AI is all but a given for the digital enterprise. What's less easy to resolve is the best approach to take to develop and leverage the AI applications that make sense for your business. As with the cloud, there are several approaches, ranging from doing everything yourself, right through to outsourcing everything.

Building out your own AI-ready infrastructure, investing in the IT hardware alongside a major data center re-design, is an expensive undertaking. And, make no mistake, if the full benefits of AI are to be realised, there can be no half-hearted tinkering with either the IT kit or the facilities infrastructure. Nothing less than a total data center re-design, in parallel with a major IT refresh, will deliver optimised AI solutions. Of course, these can be introduced over time, but there's no point in starting out with anything less than a total commitment to the process – and cost – of delivering on the promise of AI.

Handing over everything to Cloud and managed services providers has major attractions and benefits – there's no massive AI infrastructure CAPEX requirement for starters. And there's the required flexibility and scalability to grow as your AI requirement does . However, for many, the long-term cost and security of the Cloud and managed services approach gives pause for thought.

Sitting halfway between the two extremes is colocation – perfectly positioned to deliver all the benefits of AI applications, without the drawbacks of the in-house or fully outsourced approaches outlined above. Colocation can provide an AI-ready environment for companies to use to house their GPU investment, for example, allowing them to get going quickly, with more control and less expense than the alternatives. No need for an organisation to struggle with the expense and complexity of re-designing a data center to accommodate AI hardware – a modern colocation facility will do the job for you. And no need to abandon control of both costs and performance by putting yourself in the hands of a Cloud or managed services provider. Colocation and AI applications are a perfect fit.

In the short-term, getting up to speed on AI, what it can and can't do now, what it may well be able to offer in the future, is the place to start. Armed with this knowledge, sensible decisions around the data center can be made.