L3C

SCALING YOUR AI APPLICATIONS

A Cloud versus on-premise discussion

Cloud or On-premise: scaling your AI applications

Introduction

Companies are increasingly using AI to achieve their strategic initiatives, whether new revenue generating activities or improved process optimisation. AI has the ability to analyse and make sense of the ever increasing data, whether historical corporate data, new data feeds and sources or the ever increasing abundance of IoT (Internet of Things) sensor data. AI can process this effectively through the use of complex algorithms to enable better decision making.

The explosion of AI into the public consciousness and now at board level discussions in enterprises has been driven over the past 2-3 years by the abundance of data that is available to be processed and the rise of GPU (Graphical Processing Unit) technology. The development and training of AI models needs data and compute/infrastructure resources, both traditional technology and GPU technology.

Many organisations have often accessed this technology for initial models and Proof of Concepts (PoCs) using resources from public cloud providers such as AWS, Google Cloud, Microsoft Azure and IBM Cloud. However as AI initiatives are scaled throughout the organisation we should consider whether this approach is the most effective. The public clouds provide similar technology platforms (essentially x86 based) which are usually shared environments and enterprises often start using the basic level of GPU technology in these clouds. However as AI applications expand in organisations, more specialist technology may be required, whether optimised for image processing, or more advanced GPUs where usage in public clouds can become very expensive, very quickly.

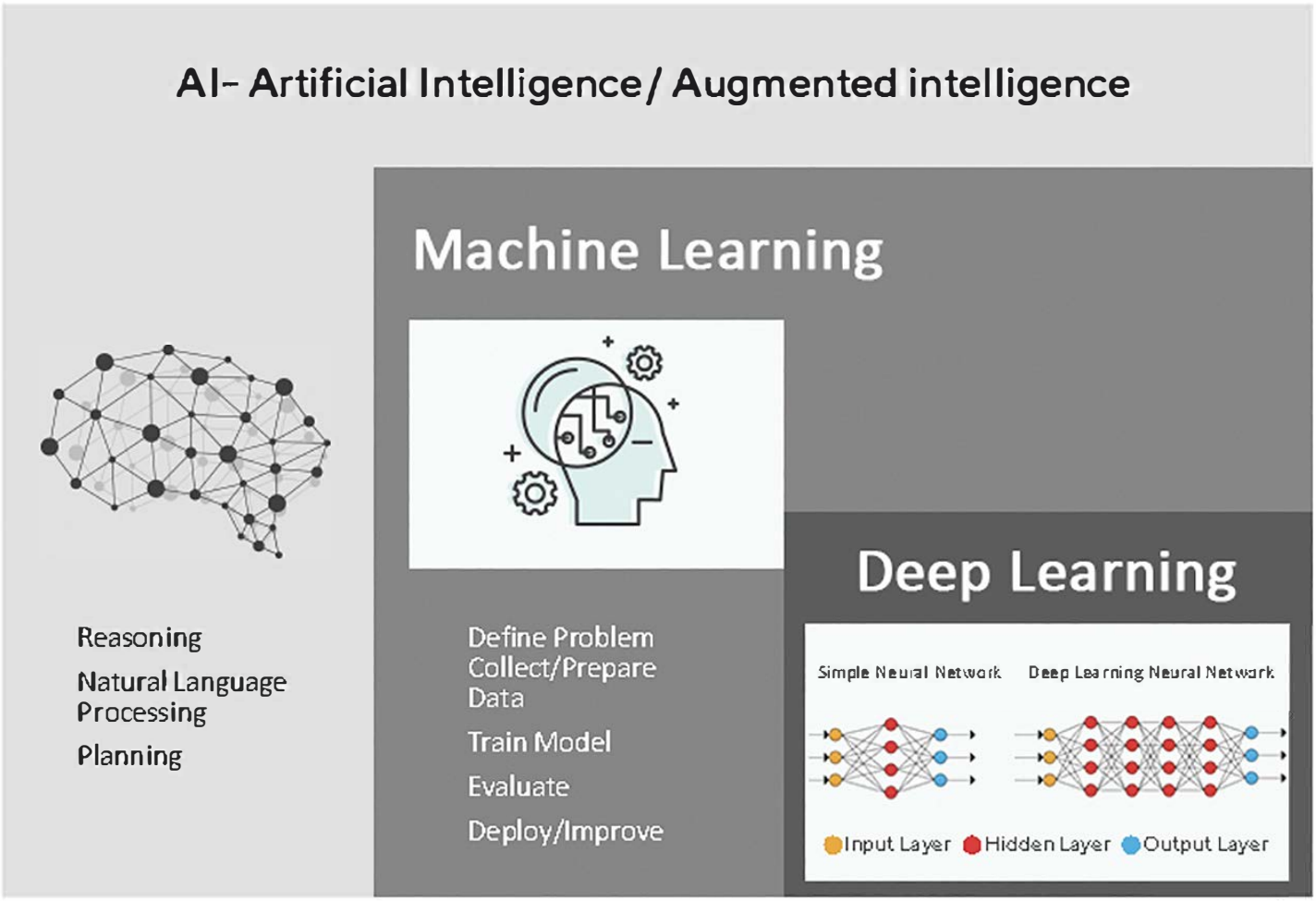
Additionally, corporate guidelines on data sovereignty or industry regulation may also need to be adhered to as well as considerations around data transfer charges that apply in public clouds. Hence as AI models move from the PoC phase to being applied in the enterprise, on premise technology should at least be evaluated as an option.

In this analysis we consider the advantages and disadvantages when utilising public clouds compared to on premise private cloud from the perspective of:

We will also consider the technology enterprises should consider and perform a benchmark test comparing the use of Nvidia v100 GPUs in AWS, within IBM's flagship Power AC922 and Nvidia's DGX Station.

De-mystifying AI

AI is an all encompassing term and heavily overused to describe anything from a simple analytics model (the kind of dashboards companies have been using for 10-20 years) to very complex neural networks, robotics and machines taking over the world and subjugating humans. From a business leader's perspective consider the following schematic. Essentially the concept of AI has been around for at least 50 years but the rise of GPU technology has allowed Deep Learning to become a reality. Essentially AI can be considered as a class of software applications that historically have been very hard for computers and easy for humans. With an abundance of data and advances in processing technology, computers can now perform computational tasks, process data and identify patterns far quicker than humans making AI models meaningful and pertinent.



Further to de-mystifying AI and ensuring correct decisions are made regarding infrastructure there are three essential concepts; data preparation, training and inference. Before any training can begin the dataset needs to be prepared and ensured it is of a suitable standard to enable the algorithms to make use of the data. This work can involve internal teams and subject matter experts from within the organisation or the use of external data scientists. Once the dataset is ready training can begin.

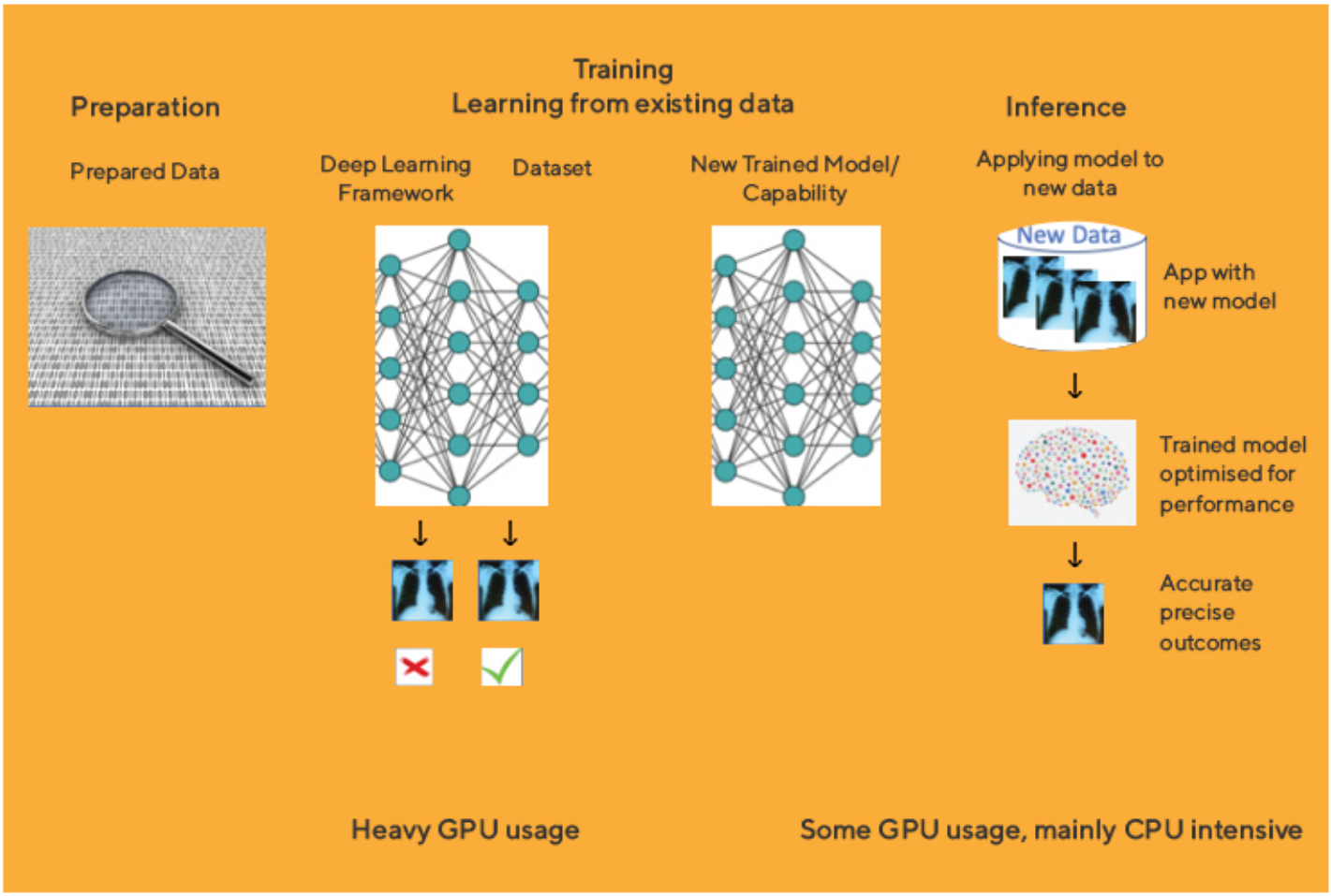
Training typically involves using complex algorithms to find patterns in the prepared dataset which are then tested for accuracy. These accuracy levels can be pre-defined by the organisation to ensure the solution is better than the current method of determining action. Training generally involves the heavy use of GPU resources to 'train' data using neural.

De-mystifying AI

Further to de-mystifying AI and ensuring correct decisions are made regarding infrastructure there are three essential concepts; data preparation, training and inference. Before any training can begin the dataset needs to be prepared and ensured it is of a suitable standard to enable the algorithms to make use of the data. This work can involve internal teams and subject matter experts from within the organisation or the use of external data scientists. Once the dataset is ready training can begin.

Training typically involves using complex algorithms to find patterns in the prepared dataset which are then tested for accuracy. These accuracy levels can be pre-defined by the organisation to ensure the solution is better than the current method of determining action. Training generally involves the heavy use of GPU resources to 'train' data using neural

Deep learning Phases, Data, Training and Inference



Example Use Cases by Industry

Healthcare	Retail	Financial
Radiology - improve workflow, chest x-rays, mammograms	Product recommendations based on buying behaviour. Customer behaviour prediction	Customer identification and authentication. Chatbots to improve customer service
Medical image analysis, 2D,3D,4D images to improve disease detection	Customer service and complaint resolution	Detection and prevention of fraud (payment fraud, insurance fraud)
Improve patient outcomes	Queue Management	Credit risk analysis
Pathological image analysis	Video analysis, content activation in video	Algorithmic Trading
Convert handwritten documents to digital, referenceable data	Improve or control the quality of product images uploaded to websites and e-commerce platforms	Personalised product recommendations for clients

Media	Industrial	Utilities
Analysis of advertising effectiveness	Autonomous vehicles	Predict wind and weather patterns to optimise resources
Sentiment Analysis	Improve maintenance/ predictive maintenance, Processing of IoT data	Remote inspection of infrastructure (turbines, pylons etc) using drones
Label and classify content libraries to promote new and hidden content and annotate content with potential to offend	Improved raw material planning and scheduling	Chatbots to improve customer access and satisfaction
Identify social media trends in near real time	Process optimisation, vision control, quality control	Demand forecasting to improve resource allocation
Natural language processing for real time translation	Seismic image processing - fault lines, hydrocarbon detection	Enhanced processing of IoT data

The recent NHSX report on AI in healthcare highlighted several areas where AI is being utilised to improve effectiveness and ultimately patient outcomes.[https://www.nhsx.nhs.uk/assets/NHSX_AI_report.pdf] Processing of medical images, particularly xrays and scans can benefit significantly from AI models, whether improving the workflow of radiologists with a chest xray triage system[Dr Shah Islam, Clinical Research Fellow, Neuroradiology Fellow, Division of Brain Sciences, Imperial College London http://bit.ly/L3CRadiology] or more complex applications.
PathAI[https://www.pathai.com/what-we-do/] PathAI is developing technology that assists pathologists in making rapid and accurate diagnoses.

In retail, we have all seen on commercial sites recommendations based on our previous buying behaviour or viewing habits (for example Netflix[Netflix Is Using AI to Conquer the World... and Bandwidth Issues.", https://www.fool.com/investing/2017/03/21/netflix-is-using-ai-to-conquer-the-worldand-bandwi.aspx.] which suggests content to us based on our viewing habits while also using AI to manage bandwidth), while the visual recognition of consumers entering shops and making recommendations accordingly is seen by some as the next phase despite the obvious privacy concerns. More specific examples have seen solutions improve conversion rates when re-contacting consumers regarding incomplete checkout of web baskets.

The financial industry adopted chatbots to improve customer service, particularly when clicking the 'let's talk or chat' button on a website. AI models are increasingly using to analyse and assess credit risk and are heavily using within cyber fraud solutions. Insurance companies process images using AI models to help settle claims and prevent fraud.

Geoscience clients are analysing seismic images and geospatial images using AI while industrial applications range from improving maintenance and service programmes on premise.

Content providers and media companies have a wealth of historical video content that will need to be analysed and maybe tagged with appropriate notifications before being released for use for campaigns or just pure viewing (some terminology that was prevalent in the 60s and 70s is certainly not acceptable today). Corporate retail brands are keen to activate their content in video (80% of internet traffic is video), not just product placement but have the ability for consumers to purchase in real time the brand of trainers, jeans, watch etc in say the latest music video or clip.

What is clear from the highlighted examples is that enterprises are using AI to solve their real world issues. It is being applied to practical issues to enhance current products and solutions, optimise operations and improve decision making accuracy. This is giving a more recognisable return and can operate within the budget constraints of most enterprises.

Top 5 Benefits of AI

(percentage of survey respondents who rate each benefit in the top 3 for their company)

What is also apparent in the uses cases is that these approaches currently use and will increasingly use image processing, whether 2D/3D/4D medical images, geospatial, industrial and retail products. Video is also becoming a significant data input to be analysed.

It also means there is the opportunity to consider the infrastructure requirements as usage scales and whether cloud or on premise is the most appropriate option to avoid capacity or financial constraints later.

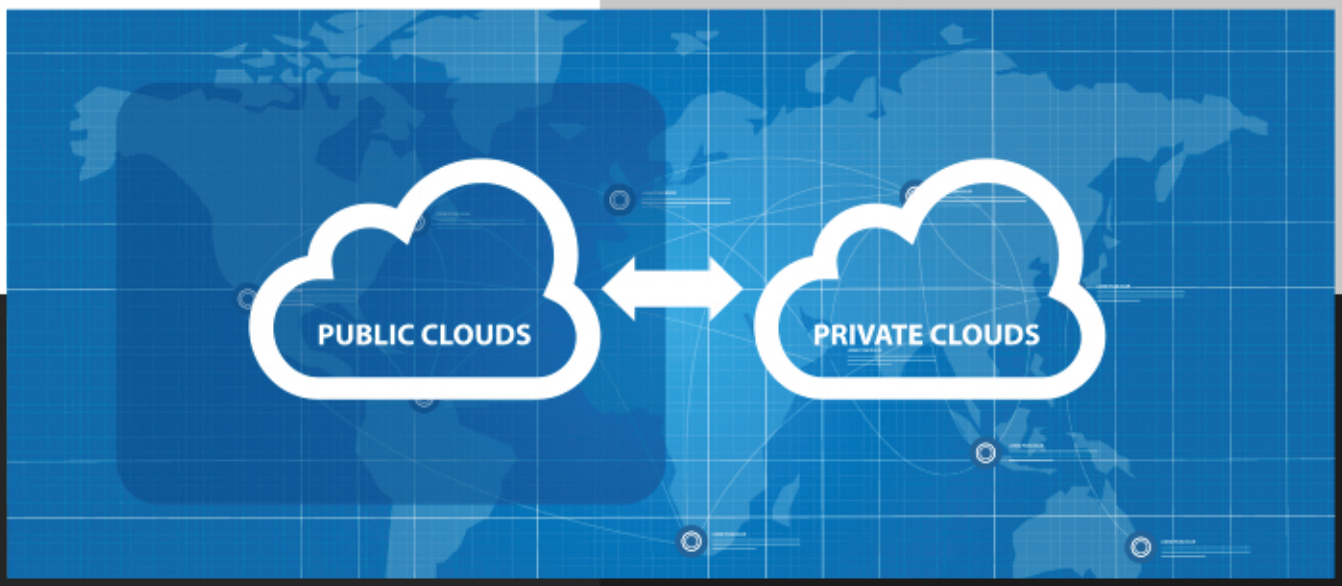
Public Cloud or Private On Premise Solution

Before evaluating GPU based technology and benchmarking alternative solutions there are a number of considerations that come into play when considering public clouds such as Azure, AWS, Google Cloud Platform or IBM Cloud compared to an on premise solution.

Many enterprises have developed their initial PoCs and maybe continue to train their models in public clouds or use a specialist provider. However as more applications begin to use the trained datasets and models and need to scale through the organisation this is not necessarily the most appropriate approach.

Organisations using public cloud providers have often fallen into the data transfer trap whereby unexpected costs are incurred when downloading data from the public cloud. This is often not accounted for in original budgets as has been a concern of many financial officers. Public clouds also often store data outside of the UK which could become a concern as post-Brexit regulatory frameworks evolve but can also be a compliance issue in regulated industries today.

As usage scales, cost can also become a concern. While enterprises have often adopted a cloud first policy for flexibility reasons and to move from Capex to Opex models for infrastructure deployment the cloud approach may not always be pertinent for GPU and compute intensive models and applications. Yet public clouds do have advantages, consider the table below to help evaluate the most appropriate approach.



	Public Cloud	On Premise
Performance	Hundreds of different configurations are available offering choice. A range of GPUs can be selected up to Nvidia Tesla v100. But all are x86 based and do not offer high speed data transfer between CPU and GPU memory meaning training time can increase	More likely to utilise specialist infrastructure optimised for performance such as IBM Power AC922 or Nvidia DGX. Performance likely to be better for this reason and proximity to dataset
Price	Can be complex to understand but offered in very granular units (often by the minute). Storage and bandwidth charges may be extra	Predictable without hidden costs. Some providers offer models whereby specialist technology can be consumed on premise in Opex model
Data Transfer Out Costs	Additional when transferring data out of the public cloud	N/A
Technical Support (including advice and guidance)	Extra Cost	In house team (or 3rd party provider)
Data location	Can be outside UK	In your own data centre
Usage	Good for short training bursts,, infrequent requirements. Multinational requirements. Consider niche providers for specific technology or SLA requirements	Better when large amounts of data are required that are not feasible to upload to cloud. Also when latency and regulatory concerns
Convenience	Very easy to acquire. Portal interfaces, sophisticated scripting and tooling, multiple tool sets available. Removing resources needs attention to avoid orphaned storage and IP addresses that could continue to be charged	Less granularity than public clouds and probably less automation and scripting but enterprises should consider if this is needed when training models. Has the advantage of fitting into existing corporate processes
Service Levels	Minimal and likely to be fixed	Depends on availability and skills of local teams
Terms and Conditions	Heavily in favour of the cloud provider. Generally inflexible 'take it or leave it'	N/A
AI Frameworks and Libraries	Generally all available (though may have to manually add to your model)	Depends on solution, IBM AC922 and Nvidia DGX come with a set of leading frameworks included

Public cloud providers have an important role to play in helping enterprises develop and apply AI solutions. They provide easy access to technology for initial PoCs and depending on the frequency of how often a model needs training (i.e. needs significant GPU access) can be an effective platform for inference. Local niche cloud providers can provide access to the enterprise class GPUs and specialist technology such as the IBM AC922 and Nvidia DGX. They are also most likely to provide customised Service Levels and a more client friendly support model as well as lower costs without hidden or unexpected add ons.

However, on premise technology has a role to play particularly where large amounts of enterprise data make uploading to clouds prohibitive, where corporate guidance predicates on premise solutions or where regulatory compliance dictates. On premise solutions don't always have to be consumed in a CAPEX model, finance models are available and niche service providers provide an 'appliance' model, delivering the technology solution to the enterprise data centre, providing technical support, yet allowing the enterprise to consume in a monthly cloud OPEX model.

Technology Really Matters

A Short Technology Overview in relation to AI

1

Generic x86 platform with GPUs

Most popular cloud providers offer x86 (virtual) hardware. This can come in different configurations with GPUs of differing performance (Nvidia are the leading GPU manufacturer and offer a range of GPU technology). Typically in the major public clouds using x86 platforms moving data between CPU memory and GPU memory is relatively slow which adds to training time, increasing the use of compute and GPU resources as well as data science time.

2

Nvidia DGX

Nvlink DGX can be purchased in several different configurations that offer up to 16 Nvidia Tesla v100 installed. DGX is specifically designed for AI and Machine Learning loads. It offers excellent performance and up to 16 Nvidia Tesla v100 GPUs. An important feature of DGX systems is NVLink and NVSwitch high speed data transfer between individual GPU memory. To our knowledge no other x86 based system offers this.

3

IBM Power9 AC922

IBM Power9 AC922 is IBM's high-end offering in the AI and Machine Learning segment. In itself AC922 is an impressively powerful machine and uses a different processor technology (IBM Power chip) than the x86 based environments. It contains up to 6 Nvidia Tesla v100 GPUs and utilises NVLink2.0. NVLink is a communications protocol developed by Nvidia but available to other providers, that can be used for data transfers between CPU and GPU and GPU to GPU. This is relevant to AI and deep learning models as fast communication between processor resources is an important factor in reducing model training time and improving accuracy.

4 IBM Large Model Support (LMS)

The IBM AC922 contains LMS, a software library written by IBM but fully open source and available to other providers. LMS allows the successful training of deep learning models that would otherwise exhaust GPU memory and therefore would abort. Combined with NVLink2.0 it essentially allows data models to be created that far exceed the size of the GPU memory which has been a traditional restriction. With LMS and NVLink2.0 as part of the IBM AC922, model sizes can be CPU+GPU memory combined which proves to be a major advantage when processing high resolution images.

---

5 Vision labelling Products - Enabling your own Subject Matter Experts

As enterprises expand their AI initiatives then as well as the cost of the technology the cost and effective utilisation of skills and resources is equally as important. Time spent preparing the dataset will save frustration, iterations and resources used in training models and yet when it comes to training models not all enterprises will have access to experienced data scientists.

Tools that enable industry and enterprise subject matter experts can have a major impact on the applicability and relevance of an AI model. Something created by an industry specialist or specialist within your own organisation who really understands the issues that need addressing is likely to be effective and trusted.

Dr Islam, referred to earlier[Dr Shah Islam, Clinical Research Fellow, Neuroradiology Fellow, Division of Brain Sciences, Imperial College London at Imperial College used IBM's AI Vision product as part of the IBM AC922 product to use his subject matter expertise as a radiologist to label x-rays as normal/abnormal. He was able to use the technology intuitively and needed no previous data science or programming expertise to create a model in 30 minutes to analyse a dataset of 60,000 x-rays. Therefore when considering the infrastructure technology to use the ease of use also needs to be considered and whether this comes included or at an additional cost as enterprises will increasingly look to enable their own teams.

Performance Test Case

To help evaluate a public cloud solution compared to an on premise solution based on IBM Power AC922 or Nvidia DGX Station (the DGX Station was chosen as it addresses a similar market and price point as the IBM Power AC922) we considered the analysis performed by Imagga[https://imagga.com] a leader in image recognition applications and established AI company. They used the following similar configurations for the exercise.

- AWS p3.8xlarge instance with 4 x Nvidia v100 GPUs
- IBM Power9 AC922 (accessed from L3C AI Cloud[www.l3c.cloud]) with 4 x Nvidia v100 GPUs
- Nvidia DGX Station with 4 x Nvidia v100 GPUs accessed on premise at Imagga

The full analysis is available for further reading and assessment [imagga.com/static/pdf/Semantic-Segmentation-of-Cityscape-and-Waste-Images-a-Comparison-between-IBM-Power-AC922-NVIDIA-DGX-Station-AWS-p3_8xlarge.pdf] but the highlights are summarised as;

- Publicly available data sets and models are used so anyone can replicate the study.
- The IBM Power AC922 was clearly more productive and outperformed the public cloud and DGX Station by nearly 2 times.
- Models were trained to comparatively low accuracy of 70-74%. This is a low value in practice but for comparing performance it makes perfect sense. We presume the study authors were driven by practical restrictions. Even on IBM Power AC922 the image heavy model training took nearly 4 days. On the other platforms it was closer to a week. Training to production level accuracy of 90% and above would have required required longer but we have no reason to believe that results would be fundamentally different.
- Large Model Support, available only on IBM's Power machines, makes a lot of difference with image intensive, high resolution data. Large Model Support is an open source library and is available on most platforms but becomes highly effective when combined with NVLink2.0 to facilitate rapid data transfer between CPU and GPU RAM.

The results are summarised below.

	Total Training Time	Accuracy
AWS Instance p3.8xlarge	6d 14hr	70.6%
IBM Power AC922	3d 17hr	72.6%
NVIDIA DGX Station	6d 11hr	73.3%

The IBM Power AC922 completed training the model 48% faster than the public cloud instance. This is very significant when resource costs, including data science resources are factored in.

Given the major public clouds use the same underlying x86 technology we see no material reason why the performance would be any different on other major public clouds using the same publicly available dataset and compute/GPU resources. However does this performance come at a cost? Given the IBM Power AC922 was provided by specialist AI Cloud provider L3C Limited we asked them to provide their monthly rolling and monthly (12 month commitment) costs to give a comparison with the publicly available pricing from the major public clouds. Their minimum charging unit is per week so they don't have the charging granularity of the major public clouds or many of the self service portal features but the performance benefit of the IBM Power AC922 does not come at an additional cost.

	Azure	AWS	Google	IBM AC922
Environment	NC24sv3 4 x v100 GPU	p3.8xlarge 4 x v100 GPU	n2-highmem-32 4 x v100 GPU	L3C Cloud 4 x v100 GPU
Minimum Charging Units	minute	minute	minute	week
Monthly Charge (rolling)	$10,480	$10,678	$7,906	$6,687
Monthly Charge (12mth commit)	$6,676	$7,495	$6,639	$5,937
Data Transfer Costs	Additional	Additional	Additional	included
Technical Support (including advice and guidance)	Additional	Additional	Additional	included
Data location	UK South	London	Netherlands	UK/London
Scale	Global	Global	Global	UK
SLA and Terms and Conditions	Non customisable	Non customisable	Non customisable	Customisable

Conclusion

Enterprises are now looking beyond PoCs and applying AI models within their organisations. Public clouds have a valid role to play based on usage requirements but as applications using AI scale, predictability of cost, proximity to data and technical support become increasingly important factors in driving organisations to consider on premise models.

The performance analysis demonstrated the IBM Power AC922 having a clear advantage over a public cloud instance and Nvidia DGX Station for a model requiring processing of high resolution images, the type prevalent across many industries including medical imaging, retail, geoscience and several financial services applications.

AI has the capability to enhance human expertise and traits such as compassion, imagination, common sense and abstraction through its ability to identify patterns, locate knowledge and almost endless capacity to process data. To maximise the benefits of this within cost, organisational, regulatory and data location constraints the default option should not necessarily be public cloud.

L3C

2 Portman Street London W1H 6DU