Cyxtera

Scaling AI: Solving the Hidden Challenges

Cyxtera + NVIDIA Partnership

Global Data Center, Interconnection & Colocation Services

Global Headquarters

Cyxtera Technologies
BAC Colonnade Office Towers
2333 Ponce De Leon Blvd, Suite 900
Coral Gables, FL 33134

Colocation Support

United States/Canada:
1-800-884-3082 (303-738-2008)

EMEA:
0800-028-8563

Asia Pacific:
00531-13-0249

Online

Customer: CustomerCare@cyxtera com
Sales: sales@cyxtera.com

Website: www.cyxtera.com

Table of Contents
Whitepaper

The hidden challenges of AI

01

It is **easy to take your first steps in Artificial Intelligence (AI)**. Popular frameworks and pretrained models provide a foundation you can build on so that often, your existing hardware, using general purpose processors, can give you the performance you need for those early experiments and initial deployments.

But there is a **hidden challenge in AI: Scaling it**. As your experience and ambition grow, so too will the complexity and scope of your AI projects. However, using existing infrastructure to incrementally scale your AI projects can be inflexible, costly and cumbersome.

There are new challenges that arise as AI deployments grow:

Larger, More Complex Data Sets Require More Resources:

Your early experiments can use small data sets to create a working model, or perhaps use islands of data to prove a particular use case. But, as you scale AI, you'll need to be able to access and process real business data for each use case, which can be problematic if your training data is stored in the public cloud. Without a dedicated server, you run the risk of slower development iterations and insufficient performance, which can hinder the success of your AI programs.

AI Becomes Too Essential to Fail:

Once an organization successfully implements AI in one area of the business and sees its transformative impact, expanding the scope becomes a natural next step. As AI becomes a part of essential business functions such as quality control or customer communications, it no longer has room to fail and any lags in performance can have a massive impact. For example, if you have a quality control solution for the factory floor, or a fraud detection solution, it needs to be fast enough to intervene before damage is done. If AI is used for customer-facing processes, such as chatbots or recommendation engines, it needs to operate quickly to avoid driving customers away. What works in a test environment may not scale up to deliver the performance you need at the speed your business requires.

The Cost of Cloud Infrastructure Can Be Prohibitively Expensive:

Scaling AI involves large data streams, sophisticated AI models and real-time results, AI is incredibly compute and storage intensive. While a cloud provider might make sense for a company looking to get its first model running on a Graphic Processing Unit (GPU), once an organization's needs grow from one GPU to dozens of multi-GPU systems, leasing these resources from the cloud providers quickly becomes impractical from a cost perspective, as does the cost of storage and data transport.

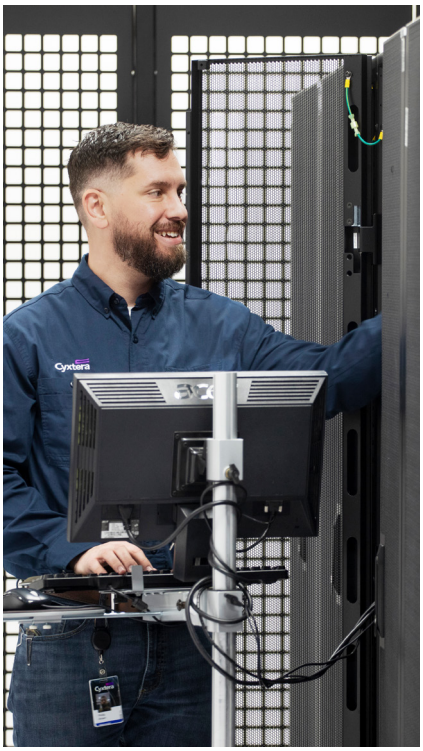
AI Innovation Outpaces Governance Strategy:

As AI becomes more accessible, organizations are starting to see individual teams working to develop and implement AI solutions within their own departments. While this bottoms-up approach can be positive, a lack of a unified organizational AI strategy can lead to disconnected solutions that are difficult for IT to monitor and control. Managing the rise of AI silos or "Shadow AI" requires a centralized AI strategy, organization-wide governance and a comprehensive infrastructure solution that that meets the needs of many groups within the enterprise.

What is Shadow AI?

Shadow AI refers to AI solutions that are not managed by the IT department. In AI, this phenomenon occurs when multiple teams are independently engaged in AI projects, adding platforms and infrastructure to build AI outside of IT. While this democratized approach can initially be positive for innovation, if governance does not keep up with the pace of change, organizations may see challenges in security, access, reliability, privacy and governance.

In addition, AI silos within an organization can quickly create cost inefficiencies. As AI models increase in breadth and complexity, it will likely become significantly more economically and technically viable to deploy models across business units rather than having each team start the process from scratch. However, if an enterprise lacks a centralized approach to AI management, they may not be aware of the AI projects happening in "the shadows", making it difficult to take a strategic approach to resource management.



Complexity of Scaling AI in numbers

9 out of 10 C-suite executives in the UK believe they must leverage artificial intelligence (AI) to achieve their growth objectives, yet **87% report they struggle with how to scale.**[1]

The worldwide market for accelerated servers [will] **grow to $25.6 billion in 2022**, with a 31.6% CAGR.[2]

50% [of AI and machine learning developers] say **gathering or generating data is the most challenging** aspect of continually training and fine-tuning AI models.[3]

38% of AI developers state that the **complexity of managing operations is the top challenge** when developing AI applications.[4]

1) AI: Built to Scale - Accenture (2019)

2) Ready to Scale AI? Don't suffer from Core Starvation - IBM (2019)

3) What are the greatest challenges to developing quality AI apps? - Forbes (2019)

4) Artificial Intelligence and Machine Learning - Evans Data (2020)

02 Why is AI so hard to scale?

AI puts unprecedented compute demand on the data center. Gartner predicts that computational resources used in AI will increase by five times between 2018 and 2023[5]. In fact, the impact of AI is so substantial that in many cases infrastructure decisions will be made with AI in mind.

To scale up AI successfully, you need the right compute resources, hosted in a data center that can reliably deliver the power and cooling required.

5) Our Top Data and Analytics Predicts for 2019 - Gartner (2019)

2.1 The compute challenge

As your AI models grow in sophistication and they need to be "parallelized" over multiple training systems in order to achieve faster processing, many companies find that they outgrow the capabilities of their general-purpose infrastructure. As a result, **GPUs become the de facto standard for AI development, with mature enterprises leveraging multi-GPU systems that are interconnected, to enable AI at scale**. They are designed to carry out lots of calculations at the same time, which accelerates AI processing, and have a high memory bandwidth to handle incoming data required for the AI workload.

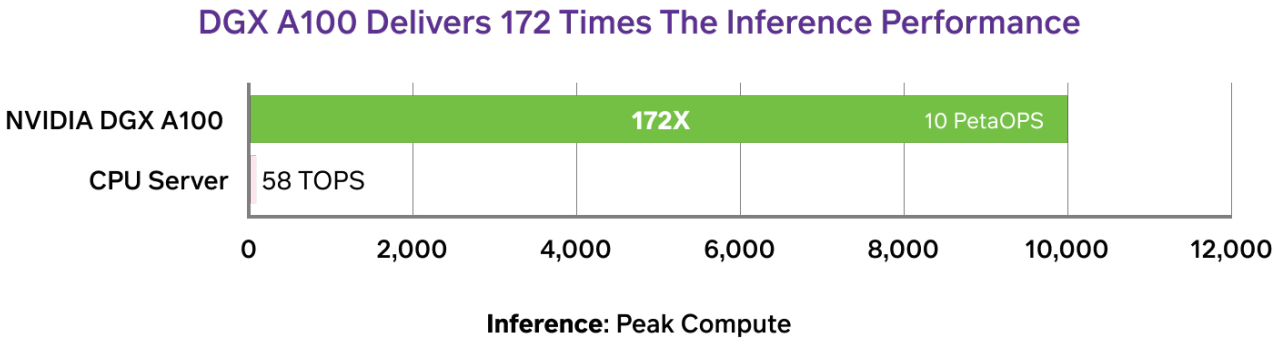
One way that businesses can deploy GPUs for AI is by using a platform especially designed for AI workloads. The world's leading solution, purpose-built for enterprise AI is the NVIDIA DGX™ A100 system, which features eight NVIDIA A100 Tensor Core GPUs. These systems are designed and optimized for easy deployment of AI workloads that are accelerated by the GPU . The GPU is optimized for matrix multiply and accumulate (MMA) calculations that are used in AI. The DGX A100 system delivers **172 times the inference performance** of a traditional CPU server and **13 times the data analytics performance** of a CPU cluster.

Time saved in model training and inference translates directly to business benefits:

- If you can train faster, you can **launch your application sooner**.
- If inference is faster, you can **make your application react quicker** to incoming data and do more inference work with the same resources.
- If you require fewer servers for your AI workloads, you can **save on space in the data center as well as cloud computing costs**.

The two phases of AI

- **Training** is where the AI model is created and learns how to perform the job it needs to do. This typically happens once at the start of the project, but the model may be retrained to improve accuracy at later stages.
- **Inference** is where the AI model is applied to incoming data. The inference is the operational use of AI and happens in real time.



CPU Server: 2x Intel Platinum 8280 using INT8 | DGX A100: DGX A100 with 8x A100 using INT8 with Structural Sparsity

Power and Cooling Challenges

2.2 Modern analytics workloads

Modern analytics workloads – particularly machine learning – require high levels of computational power and cooling management to support large data sets, complex training algorithms and real-time inputs. As a result, their energy needs far exceed those of traditional workloads. Without efficient power utilization and cooling technology, AI-intensive businesses risk long computation times, high costs and poor user experiences.

To gain access to the continuous, reliable energy and cooling capacity that AI applications require to run at scale, enterprises will want to partner with a colocation provider that has the expertise necessary to optimize AI workloads.. In addition, by substituting CPU servers with purpose-built systems designed to handle AI workloads at significantly lower power requirements, projects become substantially less resource intensive, improving processing speed and enhancing the end user experience.

03 What Infrastructure Works Best for AI?

There are four options for hosting your AI workloads: public cloud, on-premises, traditional colocation, and colocation with AI/ML Compute as a Service (CAAS).

Public cloud

The **advantages** of the cloud are well-documented:

- It is **flexible and scalable.**
- The **"as a service" consumption** model makes it easy to start and scale.
- The financial model aligns well with your growth and **does not require a huge upfront Investment**.
- You **do not have to worry about maintenance and management** of hardware resources. You pay for a service and the cloud provider delivers it.

The **disadvantages** are also well known:

- Cloud **may not be able to deliver the performance** you require.
- **With incremental fees for storing, downloading and processing data, costs add up fast as projects scale**.
- There may be questions around **security and data sovereignty**.

On-premises

Alternatively, you could buy your own servers and host your AI workloads in house.

The **advantages** are:

- **Complete control over your data**.
- **Dedicated hardware** to achieve the performance you require.
- You can **use the system as much as you want** without worrying about any additional costs.
- In highly regulated industries, **storing data in your dedicated infrastructure gives you the security assurances** you require.

The **disadvantages** can outweigh the advantages, though:

- Hosting AI on-premises requires a **significant capital expenditure** at the outset.
- Setting up the GPU cluster will **take time** and may require your team to acquire **new skills**.
- **Additional investments** in the data center's power, cooling, and networking capabilities may be required.
- There is **significant work involved** in maintenance and management, including keeping software secure and up to date.
- **Capacity planning can be difficult**. With a significant lead time in acquiring and installing additional hardware, it might be **difficult to be responsive** to the business's needs.
- It is **hard to innovate** when you need to plan data center capacity months ahead and need to make significant concrete investments to expand capacity.

Traditional colocation

Using a colocation model, you can have your own hardware managed by a colocation center.

This model solves the maintenance and management challenges of hosting on-premises, while giving you all the advantages of your own, dedicated hardware.

The **disadvantages** are also well known:

- You still have significant **capital expenditure** for your dedicated hardware.
- You're still tied to an **inflexible 3-6 month lead time** for acquiring new hardware.
- You do not have the flexibility to make changes on the fly, **restricting your ability to innovate**.

04 Cyxtera AI/ML Compute as a Service

Cyxtera provides a robust AI/ML Compute as a Service offering, powered by NVIDIA DGX A100 systems. It provides an alternative to the public cloud, on-premises, and traditional colocation options, combining their strengths.

It provides you with dedicated hardware for your AI workloads, but with the flexibility and simplicity of the cloud:

- **Best-in-class technology**, delivering powerful AI capabilities backed with high-end security.
- **Full control** over your dedicated hardware and the data on it, at every layer of the technology stack.
- **Rapid time to market**, with deployment time typically cut from 3-6 months to 1 day.
- **Business agility**, with the ability to quickly add resources as required.
- **Lower total cost of ownership**, with no need to invest in hardware at the outset. Cyxtera's expert team handles maintenance and management.

	On-Premises	Traditional Colocation	Public Cloud	Cyxtera AI/ML Compute as a Service
Dedicated hardware?	x	x		x
Control of data?	x	x		x
Nominal upfront investment?			x	x
Avoids Risk of Vendor Lock-in?	x	x		x
Hardware Maintenance Included?			x	x
Flexibility to experiment without additional cost?	x	x		x
Avoids Network Congestion?	x	x		x
Pay as You Go?			x	x
Rapid Time-to-Market?			x	x

The NVIDIA DGX A100 system is **purpose-built for AI** with a combination of **powerful GPUs** and **optimized AI software**.

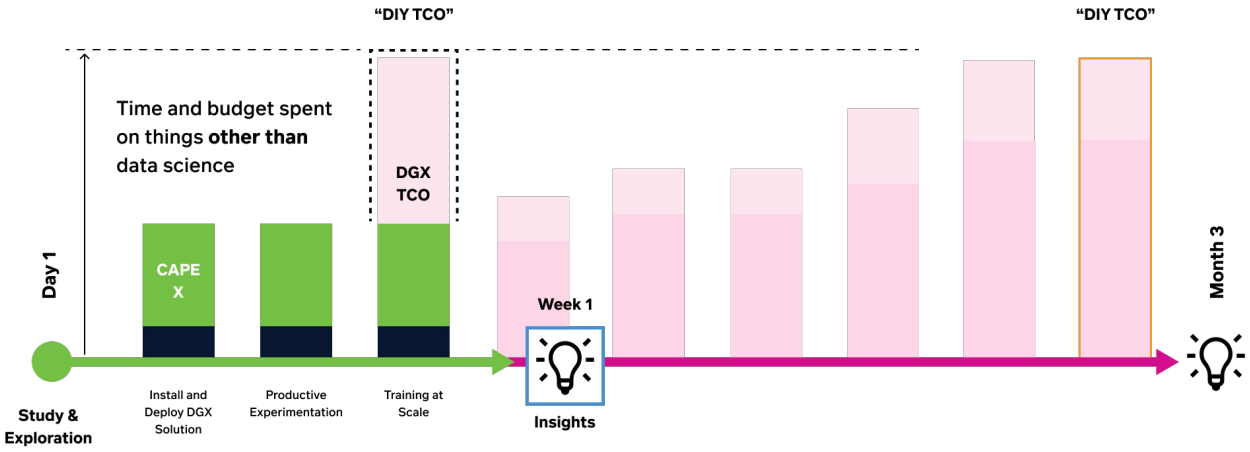
Cyxtera is the first global data center operator that can deliver access to subscription-based NVIDIA DGXA100 systems via our landmark DGX compute as a service offering. Cyxtera's solution offers enterprise customers greater agility and rapid deployments for their AI workloads.

With five petaFLOPS of performance, the DGX A100 system excels at analytics, training, and inference workloads, so you can use a **single system for all your AI needs**. That reduces complexity and improves scalability. You can divide each GPU into seven separate units, so you can run up to seven smaller workloads on a GPU at the same time; or you can combine GPUs for bigger jobs.

The software is designed to run AI applications at scale, and includes optimized versions of:

- PyTorch, MXNet, and TensorFlow for deep learning
- TensorRT for inference
- RAPIDS for data analytics

Pre-trained models, model scripts, and software development kits are provided to help you develop and launch your AI solutions faster.



Speed Up Time to Solution with Nvidia DGX A100

05 Access NVIDIA DGX A100 as a service with Cyxtera

Cyxtera's AI/ML Compute as a Service model gives you access to the DGX A100 in Cyxtera's data centers.

Cyxtera's data centers have the power, cooling, and networking capabilities required to support AI infrastructure, and are supported by a highly experienced technical team.

Cyxtera is a leading provider of interconnection services globally, with low latency direct connections to all leading cloud providers, including AWS, Google Cloud, and Microsoft Azure. That means you can get **rapid access to your data** for use in your AI applications running on the DGX A100.

Using the Cyxtera Platform, you can build and customize your data center stack, including AI/ML Compute as a Service, using **point-click provisioning**.

In addition, Cyxtera customers have access to the Cyxtera Marketplace, which enables you to expand the DGX A100 with direct access to a **rich ecosystem of managed services and technologies**, including Storage as a Service, interconnectivity, backup, disaster recovery, and security.

Test drive AI/ML Compute as a Service

Later this year, we are launching an on-demand test drive feature, so you can see the power of our new solution.

Reserve your spot here.
Call us at 1-855-699-8372 or
send us a note at sales@Cyxtera.com

Cyxtera

Highly Connected.
Hybrid Ready.

Scaling AI: Solving The Hidden Challenges

www.cyxtera.com