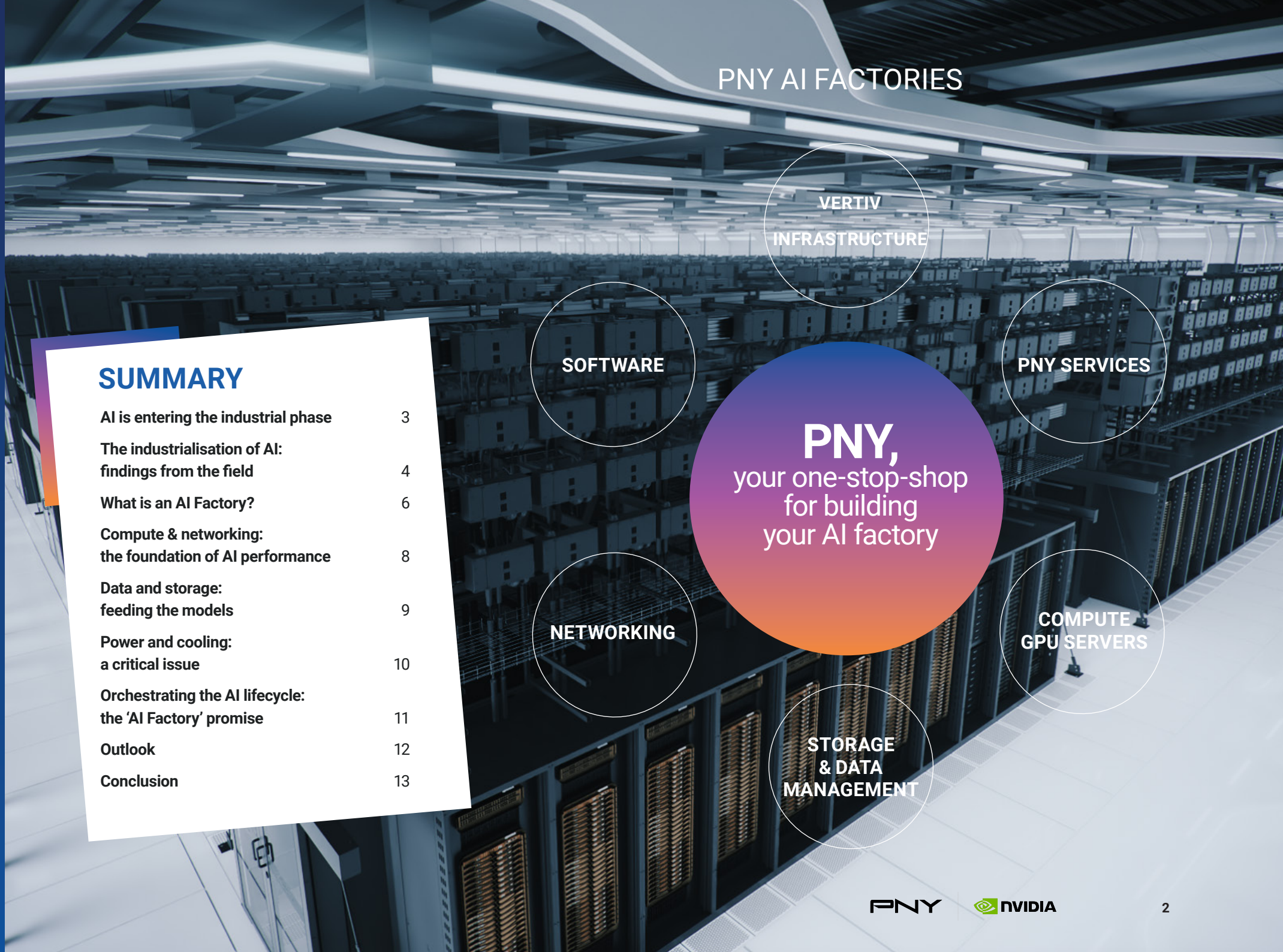
Building a high-performance AI Factory

WHITE PAPER

PNY | NVIDIA



SUMMARY

AI ENTERS THE INDUSTRIAL PHASE

ARTIFICIAL INTELLIGENCE HAS CHANGED STATUS. In business, the challenge is no longer to experiment, but to deploy and scale AI, with requirements comparable to those of a production line: performance, security, governance and continuity. Martin Jezequel sums up the dynamic: *« Organisations now want to industrialise AI — either to develop their own models using their own data, or to offer their customers computing power and AI services. »*

This shift can also be attributed to a gradual standardisation of approaches: after years of ad hoc set-ups (home-made servers and stacks), AI is becoming industrialised around more codified platforms that combine hardware and software.

WHY TRADITIONAL INFRASTRUCTURES ARE REACHING THEIR LIMITS

AI workloads — and in particular generative AI and agentic AI — are changing the nature of the constraints. It is no longer simply a question of having computing power: organisations must also ensure sufficient density, bandwidth and low latency, as well as a data pipeline capable of feeding the GPUs, and suitable physical conditions (power, cooling). Martin Jezequel describes this evolution: The underlying implication is that more advanced models require more compute and infrastructure bringing issues such as liquid cooling to the fore.

In other words, AI isn't just about the GPU: it's about the entire platform and the ability to utilise it on a daily basis.

THE PNY AI FACTORY PROMISE

PNY AI Factory is positioned as a unified approach: a software-defined environment where data, compute and orchestration converge, to accelerate the AI lifecycle and scale up with greater confidence.

As for PNY, the aim is clear: not to simply supply individual components, but to ensure a seamless production workflow—from sizing through to integration and orchestration—by leveraging the NVIDIA ecosystem and infrastructure partners, in order to minimise the typical risks associated with the transition from « POC to production ».

« The challenge is no longer about experimenting with AI, but about producing it and running it at scale. »

MARTIN JEZEQUEL,
Product Manager
Datacenter
Solutions EMEA

THE INDUSTRIALISATION OF AI: PRACTICAL INSIGHTS

EXPLOSION IN COMPUTING AND DATA REQUIREMENTS

In recent months, the trend has not been limited to a mere craze for artificial intelligence: above all, it has resulted in a significant increase in scale, with larger projects and rapidly growing volumes of data.

As Martin Jezequel points out: *« What has changed, above all, is the scale of the infrastructure, as well as the volumes of data being used. »*

The reason is simple: the more organisations demand precision and power from their models (particularly generative ones), the greater the demand for computing power becomes, and the more densely packed the infrastructure grows. Martin Jezequel directly links this surge in power to the rise of next-generation AI data centres — as well as to the emergence of technical challenges that were once considered secondary, such as liquid cooling.

SCALING UP FROM POC: THE BREAKING POINTS

Many organisations know how to launch a POC or a small test environment. The main challenge arises when moving to production, as scaling up requires meeting three requirements simultaneously :

1. **RIGOROUS SIZING :** computing power, as well as the network, storage and software layer, must be aligned with the workloads (training, inference, RAG, agent-based AI).

2. **ARCHITECTURAL CONSISTENCY :** a « low-cost » assembly, built up in successive stages, may work for a while... but will break down when you scale up. Martin Jezequel highlights the risk of an « ad hoc assembly » : doing the integration yourself, adding disparate building blocks, choosing components without an overall vision.

3. **REALISTIC DATA CENTRE PLANNING :** at high density, the equation becomes a physical one (power, cooling, electrical availability). This aspect must be anticipated in advance, otherwise it risks becoming a bottleneck when scaling up.

« Experience shows that implementing AI in a business is a demanding process, which requires a precise understanding of needs before any implementation. »

MARTIN JEZEQUEL

THE INDUSTRIALISATION OF AI: PRACTICAL INSIGHTS

WHY PNY IS INVOLVED AT THE HEART OF INDUSTRIALISATION

The unique selling point highlighted by PNY is not simply to supply a component, but to ensure the whole system works: design, validation, integration and deployment.

Martin Jezequel clearly describes the proposition: *« investing in something that is guaranteed to work »* rather than attempting to build an ad hoc infrastructure in-house. This approach is based on « turnkey » designs that are advertised as « validated and proven » with the NVIDIA ecosystem, in order to reduce integration risk and speed up time-to-market — thereby enabling the company to develop its use cases more quickly and accelerate return on investment.

Finally, Martin Jezequel highlights a real shift in perception: on complex projects, clients are realising that what is at stake goes far beyond simply supplying hardware. *« They are gradually coming to understand that PNY is not just a distributor, » he explains. « Our technical expertise, our bespoke support and our ability to design solutions tailored to every need make us a strategic partner, capable of turning ambitions into tangible achievements. »*

WHY PNY ?

- **Avoid the « cheapest option » that undermines the project.**
- **Rely on a validated / proven / tested design (PNY + NVIDIA + partners) to reduce integration risk.**
- **Accelerate the « POC to production » path (objective: faster time-to-value).**

WHAT IS AN AI FACTORY?

AN AI FACTORY IS NOT JUST A « LARGE GPU SERVER » : it is a complete, end-to-end solution designed to produce AI in a repeatable and scalable manner. It can be defined as a 'turnkey' infrastructure solution for developing and deploying AI: the idea is not just to have power, but to have a coherent environment where compute, data, network and software work together.

In practice, the AI Factory meets a very specific market need: businesses want to train or run their models using their own data, whilst cloud service providers are looking to offer their customers dedicated AI computing capabilities. In both cases, the challenge is not limited to GPUs: it lies in setting up and operating a comprehensive infrastructure capable of managing data, workloads and the scale of the models.

THE PILLARS OF A HIGH-PERFORMANCE AI FACTORY

A high-performing AI Factory is built on five inseparable pillars

1. **COMPUTE (GPU/SERVERS) :** scaled according to workloads (training, inference) and model size.

2. **NETWORK (LOW LATENCY, HIGH BANDWIDTH) :** essential for effectively interconnecting compute nodes and preventing the network from becoming the bottleneck of the cluster.

3. **STORAGE / DATA PIPELINE :** fast data access, performance and capacity; Without this, GPUs are left idle.

4. **POWER / COOLING :** density requires anticipating physical constraints and, where necessary, moving towards liquid cooling.

5. **SOFTWARE & ORCHESTRATION :** cluster management, control, allocation, tools; this is essential for real-world use and scaling.

« You can have the best hardware in the world; without proper control, it's useless. »

MARTIN JEZEQUEL

WHAT IS AN AI FACTORY ?

DIFFERENCE FROM A « TRADITIONAL » DATA CENTRE

The difference is not visual - it lies in its purpose. A 'traditional' data centre is often thought of as a facility for hosting and general-purpose processing. Conversely, an AI data centre — and therefore an AI Factory — is designed as a computing and processing engine dedicated to highly intensive workloads.

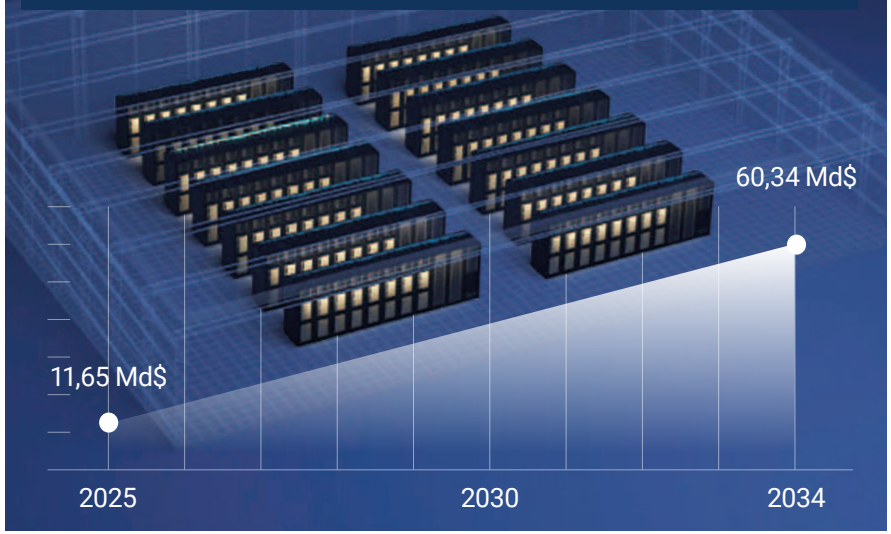
Martin Jezequel puts it in deliberately simple terms: a traditional data centre 'transports', whilst an AI data centre « propels ».

PUTTING THE ENTIRE PLATFORM BACK AT THE HEART OF THE PROJECT

Thanks to the recognition of NVIDIA's technological expertise, many requests are now made primarily in terms of GPUs (« I want an RTX PRO », « I want a GB300 »). PNY's approach is precisely to refocus the discussion: these components alone are not enough. To ensure reliable operation in a production environment, they must form part of a coherent system that integrates storage, networking, rack infrastructure and power supply.

MARKET FIGURES

The « AI orchestration » market is reported to be growing strongly by several firms: Fortune Business Insights estimates it at $11.65 billion in 2025, projected to reach $60.34 billion by 2034 (CAGR ~20%).



« A traditional data centre is designed to host and process data. An AI-dedicated infrastructure, on the other hand, is designed to deliver massive computing power and support much more intensive workloads. »

MARTIN JEZEQUEL

COMPUTE & NETWORKING: THE FOUNDATION OF AI PERFORMANCE

COMPUTING: GETTING THE SIZE RIGHT FROM THE START

The success of an AI Factory starts with a very practical question: what workload do you need to run, and what are the constraints? In practice, the equation is different depending on whether you are looking to train a model, fine-tune it, perform large-scale inference, or support hybrid use cases (RAG, agents).

Martin Jezequel emphasises this fundamental point: « Accurate server sizing is the foundation. »

This sizing phase is not limited to the number of GPUs: it must take into account memory, server type, the envisaged level of parallelism, data consumption patterns, and the scaling trajectory. The aim is simple: to avoid unnecessary oversizing, but above all to avoid undersizing, which would cause the project to hit a ceiling as soon as it goes into production.

THE NETWORK : THE BACKBONE OF CLUSTERS

In distributed AI environments, the network is often the factor that separates theoretical power from actual performance. As model sizes and cluster densities increase, interconnections become critical: GPUs must be fed with data, and nodes must communicate quickly and efficiently. Martin Jezequel illustrates this reality by pointing to the continuous rise in data rates observed in the market: 'data rates are rising' — without 'suitable interconnections', data cannot be transmitted correctly.

It is precisely here that many « Ad hoc clusters » architectures start to malfunction: a single component may be excellent (a top-of-the-range GPU), but without a suitable network, the cluster loses efficiency, training times increase, and resources are costly when only partially utilised.

« The first step remains the precise sizing of the servers. »

MARTIN JEZEQUEL

THE VALUE OF PNY: ENSURING CLUSTER CONSISTENCY

PNY is explicitly focusing on platform design and the practical integration of AI environments. The aim is to bring together compute, networking and data within a coherent architecture that remains stable as the cluster scales up.

Martin Jezequel explains that, in an AI Factory project, PNY can act as a 'central hub' between partners, particularly by addressing the 'white space': servers, cabling plans and rack integration. This approach has a very practical aim: to prevent every cluster expansion from turning into a major redesign project, and to ensure a smooth path to scalability with an architecture designed from the outset to evolve.

NVIDIA

END-TO-END AI ECOSYSTEM

PNY relies on the NVIDIA ecosystem to integrate compute, networking and software layers (orchestration/ops), in order to ensure a smooth transition to production and operation at scale (e.g. Run:ai, Mission Control).

DATA AND STORAGE: FEEDING THE MODELS

STORAGE: THE SILENT BOTTLENECK

In AI projects, computing power naturally attracts attention — but it is often the data that sets the real limit. Martin Jezequel's message is clear: beyond the GPU, you need 'a storage server' and, above all, fast access to data; otherwise, the AI platform loses its raison d'être.

In production, this becomes a critical factor: powerful GPUs that are starved of data result in slower cycles (training, iterations), reduced efficiency, and a diluted ROI. Storage is therefore not merely a 'support' component: it is a central component of performance.

BALANCING FLASH AND MASS STORAGE: THE 'PIPELINE' APPROACH

An AI Factory needs to support different AI workflows, each with its own specific requirements. For training (and rapid iterations), performance is paramount: data must be delivered at the right throughput and with the right latency — which points towards flash-based architectures.

When it comes to capacity, the challenges are different: unstructured data, archives, large datasets, and multiple versions. The challenge then becomes to standardise an end-to-end data pipeline: access, management, governance, and the ability to scale up without having to redesign the architecture every time the system is expanded.

ECOSYSTEM: BUILDING A COHERENT CHAIN

In this area, the PNY AI Factory approach is based on a simple principle: avoiding the addition of components and building a coherent chain. The aim is not to optimise a single component (storage, networking or compute) but to ensure that all three are aligned, with orchestration and operations that enable the platform to remain in production.

This approach is also a response to 'in-house' projects: you can do a lot yourself, but as usage grows and volumes skyrocket, heterogeneity becomes a risk (vulnerability, bottlenecks, constant redesign). PNY, on the other hand, advocates a more industrialised approach, drawing on an ecosystem of solutions and partners.

« The storage solution does not tolerate mediocrity. Recycling old equipment costs more than buying new.»

MARTIN JEZEQUEL

ENERGY AND COOLING: A CRITICAL ISSUE

ENERGY BARRIER: THE REAL STICKING POINT

As AI infrastructure becomes increasingly dense, the challenge becomes physical constraints available : electrical power, cooling capacity, site limitations, and lead times.

Martin Jezequel puts it plainly: on the most ambitious projects, the difficulties shift — « the challenges are now funding and operational infrastructure ». In other words, infrastructure is not just a 'purchase': this refers to a set of prerequisites (power, distribution, cooling) that determine the feasibility and speed of deployment. Many projects are held back not by the choice of GPU, but by the data centre's actual capacity to handle the density.

AIR COOLING VS LIQUID COOLING: DECIDING BASED ON DENSITY

Cooling is becoming one of the key themes at the AI Factory. Martin Jezequel links this trend directly to the increasing power of servers: the more powerful the servers are, the more heat management is required to avoid a loss of efficiency. He emphasises a simple idea: if we do not want to waste computing power unnecessarily, liquid cooling is essential for the densest configurations.

The operational conclusion is clear: the decision between air and liquid cooling must be based on density levels and scaling trajectories, not on theoretical considerations. And in many cases, hybrid approaches (air and liquid cooling depending on the area or equipment) are becoming the norm.

PNY'S CLAIMED ROLE: FROM WHITE SPACE TO COOLING

When it comes to 'critical infrastructure', PNY takes a pragmatic approach: it offers extensive expertise in the 'white space' (servers, cabling, integration) and is increasingly active in the 'gray space', particularly in the area of cooling (e.g. CDU, heat exchangers), working alongside qualified partners.

Martin Jezequel clarifies the scope, however: PNY is not the building contractor and does not manage RTE-type connections; on the other hand, PNY can play a key role in the design and installation of the 'AI data centre' technical environment and in the coherent assembly of the components required for its operation.

VERTIV.

POWER & COOLING INFRASTRUCTURE

Vertiv and PNY have announced a partnership/ collaboration aimed at facilitating the deployment of high-density AI infrastructure in EMEA, combining infrastructure solutions (power, cooling) and NVIDIA platforms, with validated and industry-ready designs.

ORCHESTRATING THE AI LIFECYCLE: THE « AI FACTORY » PROMISE

At this stage, the challenge is no longer simply to add up resources (GPUs, networking, storage), but to manage them: prioritising usage, avoiding downtime, ensuring secure operation and speeding up iterations. Martin Jezequel explicitly emphasises this point: after sizing and interconnection, the software layer becomes critical.

The underlying concept is simple: a high-performing AI Factory is not a static 'stack'; it is an operational system (observability, governance, scheduling, operational maintenance) — and it is this ability to manage it that transforms the infrastructure into a production tool.

MAXIMISING GPU ROI: SCHEDULING, VISIBILITY, CONTROL

In AI environments, value lies in the actual utilisation rate of GPUs and in the ability to accommodate multiple teams and workloads (R&D, training, fine-tuning, inference, testing). This is where the 'AI Factory' approach comes in: providing end-to-end management tools to prevent powerful yet underutilised clusters.

In this context, PNY relies on the NVIDIA ecosystem to orchestrate and operate its solutions, notably :

› **NVIDIA AI ENTERPRISE:** a software suite designed for the deployment, operation and industrialisation of AI workloads in production.

› **NVIDIA NIM AND NVIDIA INFERENCE MODELS:** application building blocks designed to accelerate the implementation of inference use cases and simplify their deployment at scale.

A STANDARDISED APPROACH TO ACCELERATE DEPLOYMENT

Another key point: the ability to respond quickly and to scale up a solution (rather than reinventing each project). Martin Jezequel speaks of a "fairly standardised" approach, and of experience already gained across deployments of various scales.

« You can have the best hardware in the world... but without mastering how to control it, it becomes useless. »

MARTIN JEZEQUEL

OUTLOOKS

AI FACTORY: A STRATEGIC FOUNDATION, BEYOND JUST THE GPU

The industrialisation of AI is not simply a matter of adding computing power. It requires the design of a sustainable, usable foundation capable of evolving: computing, networking, data pipelines, orchestration… and physical constraints (power/cooling). Martin Jezequel outlines a very clear trajectory: from ad hoc infrastructure setups towards more codified architectures, driven by the demand for ever-more-resource-intensive models and increasingly dense deployments.

A STRONG TREND: INCREASING DENSITY, ENERGY CONSTRAINTS, LIQUID COOLING

As projects progress, the issue of energy and cooling is becoming a key decision-making factor in its own right. Martin Jezequel emphasises that energy is a constraint that hampers development, and that the latest high-performance platforms are driving the move towards liquid cooling in 'large AI factories' (particularly in the field of generative AI).

THE USUAL PITFALLS (AND HOW TO AVOID THEM)

Underestimating operations (run)

› A powerful infrastructure can remain underutilised without proper management.

LEVERAGE : orchestration, visibility, control.

SOLUTION : a coherent, validated and documented design.

Building a fragmented architecture on an ad hoc basis.

› Risk of underperformance and lengthy integration.

› A powerful infrastructure may remain underutilised due to a lack of management.

Lever: orchestration, visibility, control.

Discovering power/cooling constraints too late

› Delays, resizing, unforeseen costs.

SOLUTION : prepare the site, anticipate density and scaling trajectory from the outset.

REDUCING RISK: VALIDATED ARCHITECTURE, COHERENT DESIGN, SCALABILITY

The key promise highlighted by PNY is to ensure a smooth transition to production: avoiding 'low-cost' heterogeneous assembly, reducing integration risks and accelerating time-to-value with validated, industrially viable designs (PNY + NVIDIA + partners).

On scalability, the approach is pragmatic: scaling up without endless 're-wiring' or complete redesigns, to get back up and running as quickly as possible.

« Energy is an issue because it's a constraint; it's what holds back development. »

MARTIN JEZEQUEL

CONCLUSION

PNY AI Factory positions itself as a structured solution to the most challenging question: **moving from 'project' AI to 'production' AI**, with an environment that is consistent, scalable and suitable for day-to-day use, building on the NVIDIA ecosystem and data centre partners.

PNY | NVIDIA