Maximising GenAI Performance:
The GPU & Beyond

Karl Havard
COO @ Nscale
September 11, 2024

NSCALE

AI HARDWARE
& EDGE AI
SUMMIT

Context

01 Factors

The key elements of GenAI performance, it's not just the GPU.

02 Maturity

As the market matures, values, priority and focus will change

03 Correlation

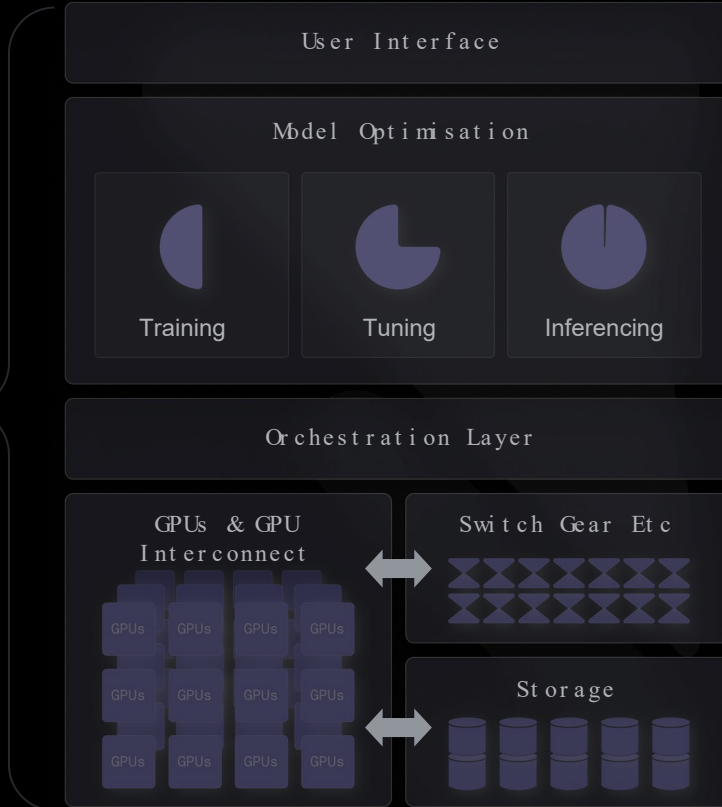GenAI performance will directly correlate to business performance

01 Factors
02 Maturity
03 Correlation

The key elements of GenAI performance, it's not just about the GPU.

The Tech Stack

External Factors

User Interface

Model Optimisation

| Training | Tuning | Inferencing |

Orchestration Layer

GPUs & GPU Interconnect

GPUs GPUs GPUs GPUs
GPUs GPUs GPUs GPUs
GPUs GPUs GPUs GPUs

Switch Gear Etc

Storage

Expertise

| Specialism | Training | In-house | External |

Networking

| T1 Location | Cost | Capacity | Latency |

Cooling

| Type | Cost | Bi-product | Efficiency |

Power

| 100% Renewable | Cost | Scale | Efficiency |

Location

| Natural Sources | Sovereignty | Community | Mobility |

?

01 Factors
02 Maturity
03 Correlation

As the market matures, values, priority and focus will change

Today

Focused on

How many GPUs do you have?

Scale

When are they available?

Time

How much are they?

Cost

The Model Maker

Tomorrow

AI HARDWARE & EDGE AI SUMMIT

Enterprise Leadership

Can you help us use GenAI to create new products and services?

Can we create our own GenAI Private Cloud with Tier 3 capability?

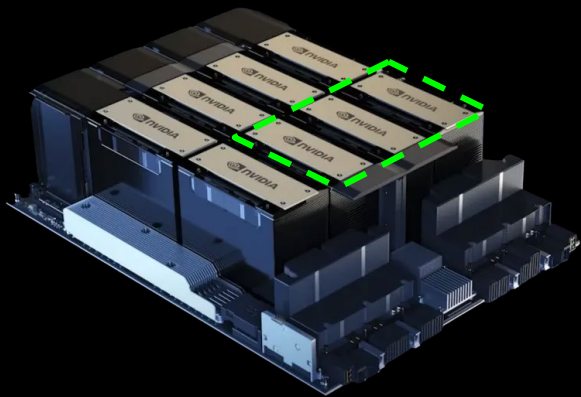What about helping us to achieve our ESG goals?

Focused on

Outcomes & Business Value

Security & Resilience

Sustainability & Compliance

As an aside….

The performance conversation will change; transparency will be key.

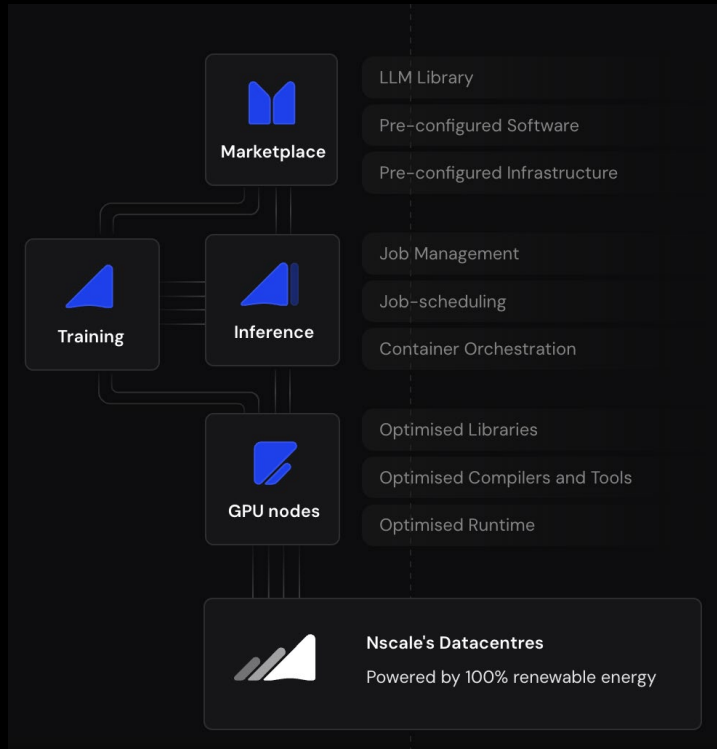**AI HARDWARE & EDGE AI SUMMIT**

User Interface

Model Optimisation

Training

Tuning

Inferencing

Orchestration Layer

GPUs & GPU Interconnect

GPUs	GPUs	GPUs	GPUs
GPUs	GPUs	GPUs	GPUs
GPUs	GPUs	GPUs	GPUs

Switch Gear Etc

Storage

Security & Resilience

■ **Essential**

■ **Preferred**

■ **Bonus**

Expertise

| Specialism | Training | In-house | External |

Networking

| T1 Location | Cost | Capacity | Latency |

Cooling

| Type | Cost | Bi-product | Efficiency |

Power

| 100% Renewable | Cost | Scale | Efficiency |

Location

| Natural Sources | Sovereignty | Community | Mobility |

01 Factors
02 Maturity
03 Correlation

GenAI performance will directly correlate to business performance

GenAI Performance

Marketplace
- LLM Library
- Pre-configured Software
- Pre-configured Infrastructure

Training

Inference
- Job Management
- Job-scheduling
- Container Orchestration

GPU nodes
- Optimised Libraries
- Optimised Compilers and Tools
- Optimised Runtime

Nscale's Datacentres
Powered by 100% renewable energy

Business Performance

- High performance training means faster time to deployment

- High performance inferencing means a much better user experience

- Delivered **securely**, ensuring **compliance**, in region and with **100% renewable energy** means you achieve your ESG goals

- **Evergreen technology** with scale to grow provides peace of mind for your future needs

- **Transparency**, so you know exactly what you are committing to; nothing hidden

Glomfjord, Norway
60MW of Contiguous Capacity.

Strategically located near the Arctic Circle in Norway, Nscale leverages the environment's natural cooling. The data centre is powered by 100% renewable energy and features energy-efficient adiabatic cooling, streamlined operations, and scalability.

Nscale benefits from some of the lowest cost-renewable power on earth making it one of the lowest cost AI training hubs in the world.

Our vertical integration and low-cost power enables us to deliver sustainable, high-performance AI infrastructure more quickly and more cost-effectively than our competitors.

AI HARDWARE & EDGE AI SUMMIT

Building a pipeline of sites
across Europe and the US

Operational
- 🔵 Glomfjord, Norway: 60MW
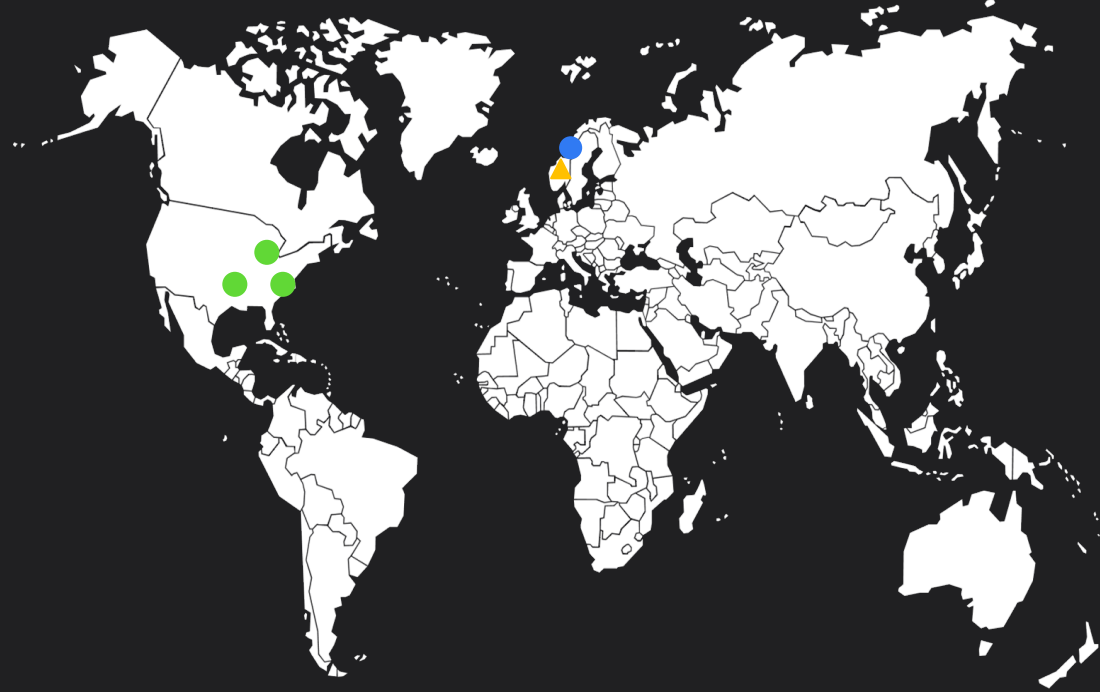
Greenfield Pipeline
- 🟢 Texas, USA: Up to 400 MW
- 🟢 Ohio, USA: Up to 333 MW

Colocation Partners
- 🔺 Stavanger, Norway: 6 MW
- 🔺 Oslo, Norway: 6 MW

Total Capacity of 0.805 GigaWatts

With more on the horizon.

AI HARDWARE
& EDGE AI
SUMMIT

Thank you

To learn more about our platform, visit nscale.com and:
- Elio's workshop is later today @ 4.15PM - 5.15PM in the Valley Room
- Followed by our Networking event at SP2 Bar & Restaurant from 6.30-8.30pm.
- And please visit us on Booth 43 & 44