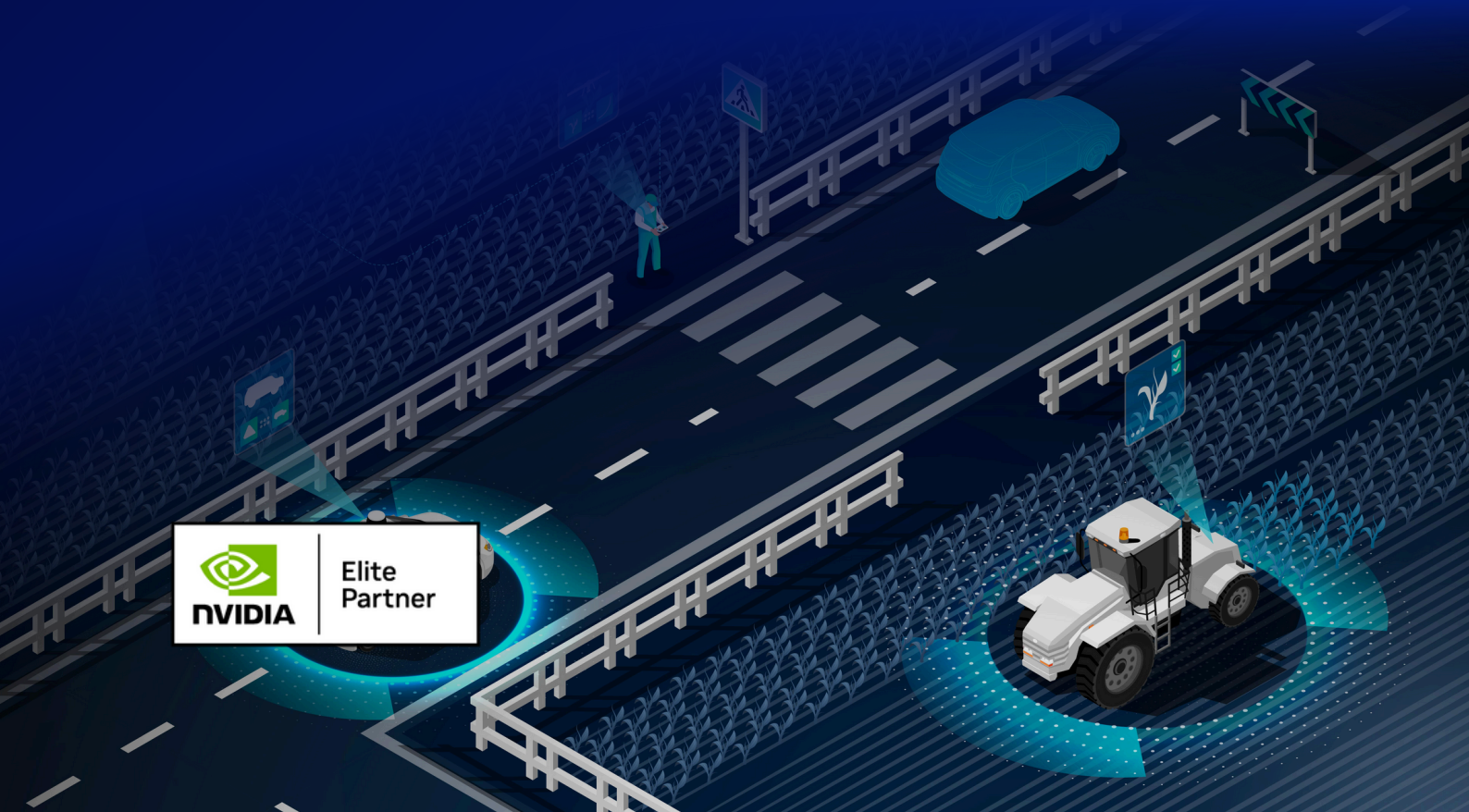
AMBER | LIGHTLY

CASE STUDY

Sustainable AI at Lightly - Performance Powered by NVIDIA DGX™

From cloud to co-location in Norway



COMPANY BACKGROUNDS

Pioneers in AI Technology

AMBER and Lightly: Transforming industries through advanced computing and vision AI

LIGHTLY

Lightly is the computer vision suite for ML teams building real-world vision systems. Founded in Zurich by ETH and Harvard alumni, Igor Susmelj and Matthias Heller, Lightly provides an end-to-end platform for managing datasets, preparing training data, and training better AI models.

Lightly's flagship products, LightlyStudio and LightlyTrain, cover the full lifecycle of computer vision development. LightlyStudio supports dataset curation, labeling, quality assurance, visualization, and dataset management in a single workflow. LightlyTrain enables self-supervised pretraining and fine-tuning to improve model performance while significantly reducing labeling effort.

AMBER

AMBER AI & Data Science Solutions GmbH is a leading provider of cutting-edge full-stack AI solutions and high-performance computing. As an NVIDIA Elite Partner, AMBER collaborates with industry leaders to deliver state-of-the-art, tailored solutions that accelerate innovation and reduce costs. With ISO-certified quality standards, AMBER ensures reliability and excellence in every project. Founded in 2006 as FluiDyna, AMBER has been at the forefront of artificial intelligence innovation since 2008.

COST & SUSTAINABILITY DILEMMA

Challenges Facing the Industry

The vision AI industry stands at a critical inflection point. With market projections exceeding earlier estimations over and over again, organizations face a dual challenge: managing spiraling costs while addressing growing environmental concerns.

Vision AI's appetite for computing power has become insatiable. Training advanced models demands thousands of GPU hours, translating to monthly cloud computing bills that can reach hundreds of thousands of dollars for companies in autonomous driving, robotics, and industrial inspection (NVIDIA, 2022). This computational burden creates a significant barrier to entry for smaller organizations and limits innovation across the sector.

These computational demands carry severe environmental consequences. Research by Strubell (2019) revealed that training a single large AI model can generate carbon emissions equivalent to five cars' lifetime output. The situation has worsened dramatically, with MIT (2023) documenting a 300,000-fold increase in energy consumption for AI training since 2012. Vision models, with their complex architectures and massive datasets, rank among the highest consumers of energy.

Compounding these challenges is widespread data inefficiency. Based on Lightly's own research, organizations typically utilize only 20% of their collected visual data effectively. The remaining 80% - often redundant or low-quality samples - consumes valuable computational resources without improving model performance.

The use cases justify the significant energy consumption, yet front-runner companies like Lightly are actively working to minimize their carbon footprint. Through smart hardware choices and strategic partnerships, they are able to achieve both: outstanding performance and sustainable energy usage.

"Renewable-powered co-location is the answer to AI's twin challenges: it cuts both carbon emissions and costs while delivering the performance vision AI demands."

Michael Rechenmacher, CEO of AMBER

CHALLENGES

Why Lightly was in Need for a Cloud Alternative

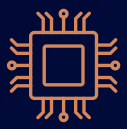
Cost, performance, and control made self-hosting the clear choice

Unsustainable Cloud Expenses

Lightly's GPU cloud costs were reaching $17,000-$70,000 monthly ($16.10 per GPU hour at peak), creating an immense business challenge as computational needs grew. This expense trajectory urged the company's management to consider other options.

Unpredictable Cloud Performance

Cloud-based GPU instances introduced variability due to shared resources, virtualization overhead, and infrastructure contention. For Lightly’s production-grade AI workloads, this translated into unpredictable throughput, undermining the consistency required for continuous training and experimentation.



Limited Hardware Access

Limited access to high-performance computing resources was restricting Lightly's ability to train larger models and process more complex datasets, directly impacting the company's capacity to serve high-value enterprise clients and expand into new market segments.

NEEDS

Building the Infrastructure Vision AI Demands

How Lightly achieves performance, efficiency, and sustainability at scale

Sustainable High-Performance Infrastructure

Vision AI's energy demands are surging. Lightly required computing infrastructure that could deliver exceptional performance for large-scale model training, while dramatically reducing the environmental footprint through energy-efficient design and renewable-powered colocation.

Cost-Efficient Scalability at Peak Demand

With GPU costs escalating rapidly, Lightly needed a setup that could scale efficiently under continuous workload pressure. The goal: eliminate unpredictable cloud billing by gaining full cost control and maximizing GPU utilization in-house.

Purpose-Built Systems for Vision AI Workloads

Vision models are uniquely compute-intensive and memory-hungry. Lightly needed dedicated systems optimized for pretraining and inference tasks like object detection and segmentation, ensuring fast iteration and throughput without cloud-related bottlenecks.

Reliable Partners for Implementation and Support

To ensure a smooth transition from cloud dependency to in-house GPU clusters, Lightly needed expert partners for sourcing, deployment, and ongoing support. The right infrastructure partner would bring not just hardware, but deep technical insight and operational reliability.

AMBER'S SOLUTION

From Cloud Variance to Predictable Throughput on NVIDIA DGX™

Co-located. renewable-powered, and enterprise-supported

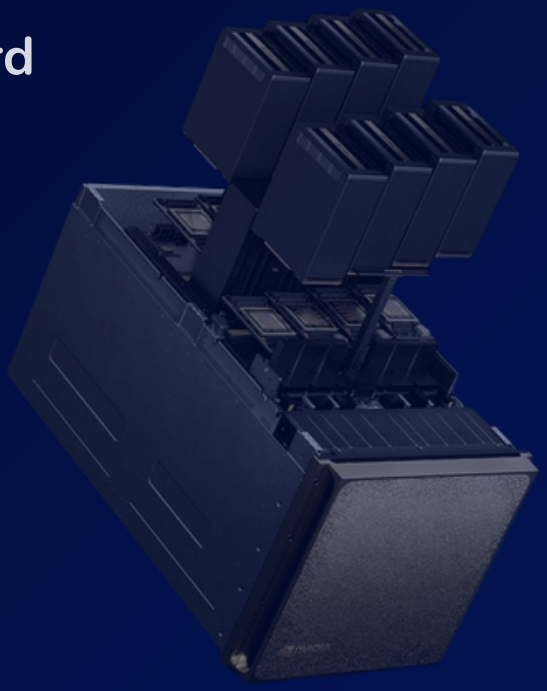
AMBER moved Lightly from variable cloud instances to a dedicated NVIDIA DGX™ B200 cluster in Norway. The target was simple: predictable performance and linear costs. Dedicated capacity removes noisy-neighbor effects and quota constraints while keeping data on owned infrastructure.

Each B200 provides 192 GB HBM3e to lift batch sizes and stabilize scaling. NVLink sustains multi-GPU throughput across CV training and LLM inference. Operating expenses shift from volatile per-hour fees to a fixed run-rate aligned to utilization.

Inference leadership validated by new InferenceMAX v1 results underpins the long-term efficiency path at scale. Net effect. faster epochs. stable queues. predictable capacity. Powered by 100% renewable energy at Green Mountain in Norway.

Blackwell B200 sets a new performance standard

- NVFP4 native low-precision path
- Fifth-gen NVLink + Switch
- TensorRT-LLM optimization stack
- Higher tokens/sec at scale
- Better cost per token
- InferenceMAX leadership results



BENEFITS

Operational Benefits at Production Scale

Lower OpEx, faster training cycles, predictable throughput, sustainable power, and data control

Predictable Spend

Lightly replaces variable cloud pricing with a predictable on-prem run rate, aligning spend with utilization rather than spot markets. The mechanism is dedicated DGX capacity in renewable co-location plus steady scheduling.

Faster Training Cycles

Jobs start immediately and finish consistently, so retrains and evaluations align with release cycles; mechanism is queue-free access on a local DGX fabric with tuned orchestration and no multi-tenant interference.

Data Control

Data stays inside Lightly's environment for compliance and privacy. The mechanism is on-prem storage, enterprise access controls, and integration with the DGX software stack across training and inference.

Predictable Throughput

When experiments or user load grow, Blackwell B200 keeps throughput high and latency steady. InferenceMAX validates step-change efficiency. Mechanism. NVFP4 precision, fifth-gen NVLink, and TensorRT-LLM optimizations proven in production stacks.

Sustainable Performance

Lightly's self-hosted OPEX runs near $0.51 per GPU-hour while cloud H100 ranges $2.95–$16.10 per GPU-hour. Mechanism. Amortized hardware, efficient power, and zero virtualization tax under sustained usage.

Sharper Data Efficiency

B200 delivered up to 57% faster training than H100 in Lightly's tests, enabling larger batches and tighter iteration; mechanism is more memory, bandwidth, and higher compute density.

RESULTS

Measurable Gains. Same Workloads.

Training speed, batch scaling, and unit economics vs. shared cloud caselines

Lightly's baseline was cloud spend ranging $17,000–$70,000 per month and $2.95–$16.10 per GPU-hour, with unpredictable performance that broke continuous training. AMBER's DGX deployment shifted operations to a self-hosted model with OPEX around $0.51 per GPU-hour under continuous use. Training throughput improved significantly.

Lightly's benchmarks show up to 57% faster model training on B200 versus H100 for computer vision workloads, driven by larger batches and memory bandwidth. For the next inference upgrade, NVIDIA's InferenceMAX v1 results set the marker.

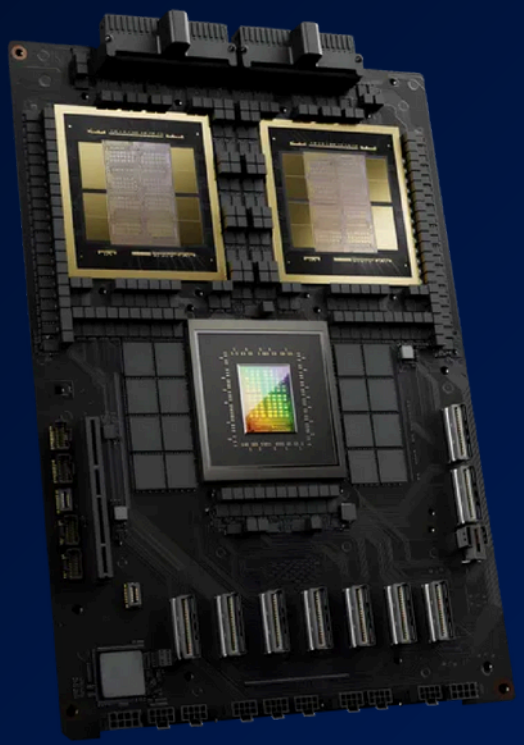
Blackwell reduces cost per million tokens by ~15× versus the prior generation and reaches up to 60,000 tokens per second per GPU on gpt-oss with the latest TensorRT-LLM stack. These external results define the ceiling Lightly can target as the software stack matures, enabling fewer nodes per SLA and clearer economics at scale.

Read more about the NVIDIA Blackwell B200 vs H100 on this Lightly Blog Article

60,000 TOKENS/SEC PER GPU ON GPT-OSS WITH TENSORRT-LLM

15× LOWER COST PER MILLION TOKENS VS H200 AT SIMILAR LATENCY

8 BLACKWELL GPUS PER DGX B200 SYSTEM



TESTIMONIALS

Voices

Insights from Lightly & AMBER on the transformative impact of NVIDIA DGX™ technology:

LIGHTLY

"Cloud felt like a moving target. On DGX, our jobs start when we need them and finish on time, and the bill finally makes sense. Our datasets stay in our environment. The Blackwell path gives us headroom without surprise trade-offs."

Igor Susmelj, co-Founder, Lightly

AMBER

"We delivered a clean DGX foundation with a direct lane to B200. Lightly now iterates at its own pace with predictable access and stable spend. When they enable Blackwell inference, throughput and unit-cost gains show up immediately in production."

Michael Rechenmacher, Founder, CEO

CONTACT

Your Contact for More Information

AMBER AI & Data Science Solutions GmbH

Grünwalder Weg 32
D-82041 Oberhaching

T +49 89 413 2733-00
F +49 89 413 2733-09
E info@amber.eu