emma

Ebook

AI for SMBs: Accessible, Affordable, and Unconstrained

How small and medium-sized businesses can build and deliver AI-powered services — from personalized customer experiences to predictive maintenance, industrial automation, and real-time analytics — without GPU shortages, cloud lock-in, or runaway costs.

IN THIS EBOOK

Table of Contents

CHAPTER 04

Sovereign AI on a Budget

> **For European SMBs handling sensitive data or operating in regulated industries, compliance requirements often introduce additional infrastructure constraints.**

Sovereignty Without the Tradeoff

Maintaining strict data residency or sovereign control doesn't just limit regions where workloads can run, it also restricts which cloud providers can be used. That often means reduced flexibility to tap into hyperscaler-grade services, global GPU availability, or advanced AI tooling. It creates a tradeoff between sovereignty and innovation.

emma offers a unique advantage for sovereignty-bound organizations. It gives the flexibility to orchestrate sensitive data and regulated workloads on EU-native sovereign GPU infrastructure, while allowing non-sensitive training or inference workloads to scale across hyperscalers and other providers where GPU pricing or availability is more favorable.

emma powers several architectural capabilities that make sovereignty economically viable:

- **Split inference** and federated deployments keep sensitive data in EU-based sovereign cloud providers while routing non-sensitive processing to best-price cloud GPUs.
- Dedicated **bare metal GPU clusters** in a Luxembourg-based data center provide full hardware isolation, predictable performance, and EU jurisdiction alignment ideal for highly regulated or IP-sensitive workloads.
- **Optimized private networking** reduces egress charges and avoids non-optimal data paths in hybrid and multi-cloud AI deployments.
- **Right-sized GPU allocation** prevents overprovisioning expensive hardware for lighter workloads.
- **Centralized governance** and orchestration eliminate the need for separate management stacks per environment, saving time and effort.

Because workloads are intelligently distributed instead of confined to a single sovereign environment, organizations avoid the common cost traps of sovereignty. SMBs can meet regulatory obligations without isolating themselves from innovation. This makes sovereignty operational, not just a legal constraint.

> **Because workloads are intelligently distributed instead of confined to a single sovereign environment, organizations avoid the common cost traps of sovereignty. This makes sovereignty operational, not just a legal constraint.**

CHAPTER 02

Flexible Infrastructure for Every AI Workload

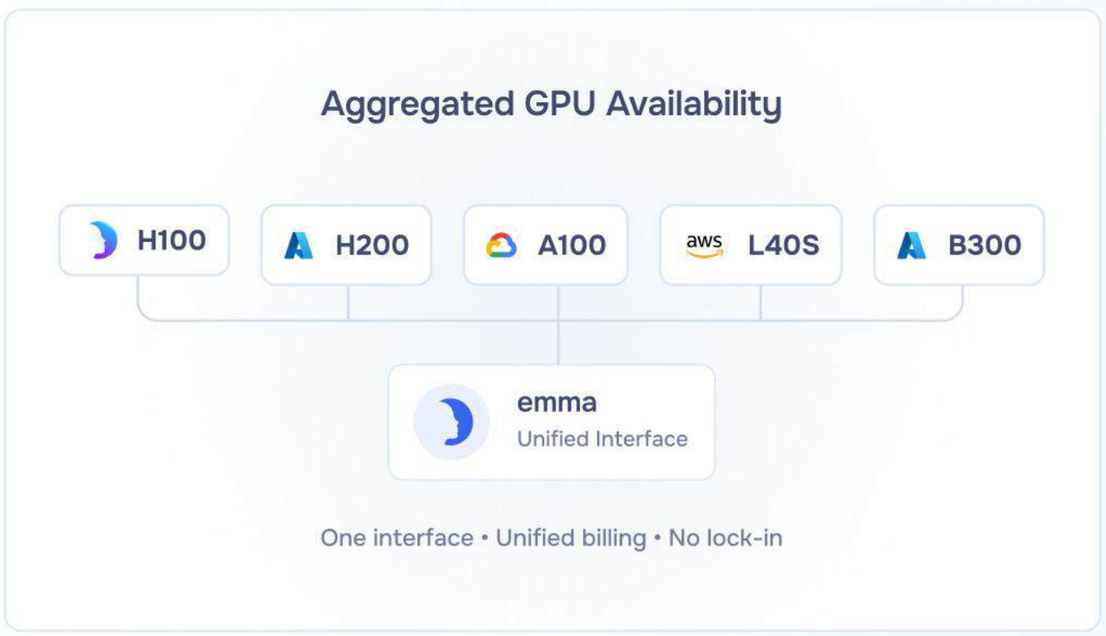
> **To meet the challenges SMBs face, we've structured the emma cloud operations platform to support several infrastructure modalities.**

Multi-Cloud GPU Infrastructure

Instead of being limited by a single cloud's regional quotas or waitlists, emma aggregates GPU availability across multiple providers and geographies. If one provider is constrained or backlogged, capacity can be provisioned elsewhere without requiring new skills, manual configurations, or architectural changes. This removes one of the biggest bottlenecks in AI infrastructure — waiting weeks for inventory opening up and subsequent approvals.

Teams can provision H100, H200, A100, L40S, or B300 GPUs on demand, scale into NVLink-ready 8-GPU training nodes, or reserve capacity for predictable workloads, all from the same interface with unified billing.

Logical GPU clusters can span several clouds and regions, allowing training and inference workloads to run wherever availability, cost, or latency is optimal. Combined with emma's private inter-cloud backbone, which minimizes egress costs by up to 70% and optimizes cross-cloud traffic, data movement never becomes a cost barrier, regardless of how data-intensive workloads are.



The Egress Problem in Multi-Cloud GPU

In most environments, egress is where multi-cloud GPU strategies break down. Even when compute pricing is transparent, egress charges come as a surprise whenever you synchronize model checkpoints between training nodes, replicate datasets across regions, or distribute inference workloads across clouds. Cloud providers have little incentive to make outbound traffic inexpensive, which is why multi-cloud GPU clusters frequently become financially unpredictable. emma largely neutralizes this friction by bypassing the public internet and provider-specific interconnects.

Bare Metal GPUs

For workloads that demand consistent performance and full hardware control, the emma platform gives access to dedicated bare metal GPUs hosted in Luxembourg. Unlike shared cloud instances, each node provides full physical GPU access, free from noisy neighbors whose heavy workloads can compete for compute, storage, and network resources.

Shared resources can sometimes lead to inconsistent performance and slower execution, but bare metal GPUs are exclusive to your team. Your workloads get predictable, guaranteed compute performance and can also leverage high-performance networking fabrics for latency-sensitive workloads.

Luxembourg-based bare metal nodes also support strict data residency requirements, making them ideal for regulated industries or workloads that require keeping sensitive data in EU jurisdiction.

All bare metal deployments are integrated with emma's orchestration layer, allowing teams to manage bare metal and other cloud GPUs, VMs, K8s clusters, storage, and networking from the same unified interface. This means less context switching, fewer manual interventions, and lower operational overhead, despite multiple environments.

Hybrid Deployments

When AI projects need both elasticity and control, emma lets teams combine multiple infrastructure types into a single, orchestrated environment: cloud GPUs for burst capacity, reserved instances for predictable workloads, and bare metal nodes for steady-state or regulated workloads.

With hybrid deployments, teams can store, process, and train critical models where data residency and compliance requirements demand it, and burst into the cloud when training demand spikes. They can spin up H100s, H200s, or A100s instantly to handle peak workloads without overprovisioning or waiting for availability.

This model also helps cost management as teams can run inference or development workloads on fractional or virtualized GPUs while reserving bare-metal nodes for performance-intensive training. Teams can orchestrate a balanced AI architecture, which is flexible, consistent, and fully managed from emma's unified interface.

Key Takeaway

Large model training	Speed
Ultra-low latency inference	Control
Performance-intensive AI pipelines	Scale

> **Imagine running inference in one cloud, training in another, and orchestrating a logical GPU cluster across multiple providers — without waiting, without lock-in, without fragmentation, and with up to 70% lower data transfer costs.**

CHAPTER 03

Engineering AI on a Budget

> **The emma platform gives your teams the levers to match workload to infrastructure, which helps control costs without slowing down innovation.**

Right-sizing GPU Workloads

Different tasks demand different GPU capacity. With emma, teams can instantly find and choose the right infrastructure for the job. Practically, it can look like:

Picking the right GPU for the job avoids overspending while keeping performance intact, as needed.

If a workload isn't running on the ideal GPU, emma can analyze historical usage and real-time demand to recommend a better configuration. Switching is simple — one click, and the workload is running on the right GPU, instantly optimizing cost and performance.

Inference ↓ **L40S/RTX5000**	Fine-tuning ↓ **A100**
100B+ training ↓ **H200**	Trillion-parameter reasoning ↓ **B300**

Bare Metal GPUs

emma makes it virtually seamless to keep predictable, steady workloads on reserved cloud GPUs or bare-metal nodes, and tap on-demand cloud GPUs only when demand spikes. This way, surplus capacity is always available when needed, without paying for idle GPUs.

Transitioning between environments is seamless. No extra skills, scripts, or complex coordination required. Teams manage everything from a single control plane, where all data, usage stats, and monitoring are centralized. Scaling up or down is simple, giving SMBs predictable performance and cost control without operational headaches.

HOW EMMA HELPS

Workload-to-Infrastructure Matching

emma gives teams the levers to match workload to infrastructure

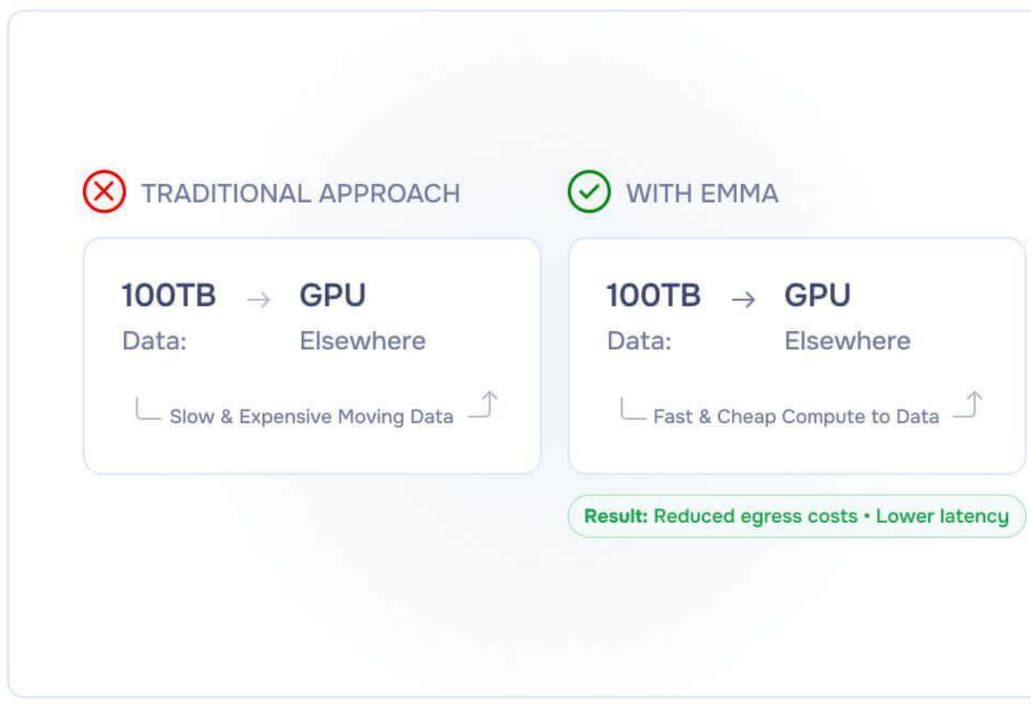
- right-sizing GPU allocation
- burst/steady-state modeling
- data locality control,
- multi-cloud arbitrage

All from a single control plane with unified billing.

Data Locality Control

In modern AI workloads, data gravity often outweighs model size. A trained model might be a few GBs, but the datasets behind it can be hundreds of GBs, terabytes, or even petabytes. However, GPU capacity may not always be available in the same location where data resides. And moving that data across regions quickly becomes the real bottleneck.

With emma, teams can provision GPUs from other cloud providers that have availability within the same region where the data already resides. Instead of relocating datasets, compute is brought closer to the data.



Combined with emma's optimized private inter-cloud networking backbone, this setup minimizes egress costs, and helps preserve regulatory boundaries when data is already stored in compliant zones.

Multi-Cloud Arbitrage

With emma, teams can deploy workloads wherever required GPUs are available and pricing is optimal, without changing tooling, rewriting infrastructure code, or rearchitecting their stack. Instead of being locked into a single provider's pricing model, SMBs can build logical GPU clusters that span multiple clouds and shift workloads when market conditions change. If GPU prices spike in one region, workloads can move to a more cost-efficient provider or region. If capacity tightens, additional GPUs can be provisioned elsewhere and orchestrated as part of the same distributed environment.

Because every environment runs through the same orchestration and networking layer, deployments feel identical across providers. The control plane remains the same, the management experience remains the same, and operational overhead doesn't increase.

Key Takeaway

Instead of relocating datasets, emma brings compute closer to the data:

- **Minimizing egress costs**
- **Reducing latency**
- **Preserving regulatory boundaries.**

> **"Instead of being locked into a single provider's pricing model, SMBs can build logical GPU clusters that span multiple clouds and shift workloads when market conditions change.**

CHAPTER 04

Sovereign AI on a Budget

> **For European SMBs handling sensitive data or operating in regulated industries, compliance requirements often introduce additional infrastructure constraints.**

Sovereignty Without the Tradeoff

Maintaining strict data residency or sovereign control doesn't just limit regions where workloads can run, it also restricts which cloud providers can be used. That often means reduced flexibility to tap into hyperscaler-grade services, global GPU availability, or advanced AI tooling. It creates a tradeoff between sovereignty and innovation.

emma offers a unique advantage for sovereignty-bound organizations. It gives the flexibility to orchestrate sensitive data and regulated workloads on EU-native sovereign GPU infrastructure, while allowing non-sensitive training or inference workloads to scale across hyperscalers and other providers where GPU pricing or availability is more favorable.

emma powers several architectural capabilities that make sovereignty economically viable:

- **Split inference** and federated deployments keep sensitive data in EU-based sovereign cloud providers while routing non-sensitive processing to best-price cloud GPUs.
- Dedicated **bare metal GPU clusters** in a Luxembourg-based data center provide full hardware isolation, predictable performance, and EU jurisdiction alignment ideal for highly regulated or IP-sensitive workloads.
- **Optimized private networking** reduces egress charges and avoids non-optimal data paths in hybrid and multi-cloud AI deployments.
- **Right-sized GPU allocation** prevents overprovisioning expensive hardware for lighter workloads.
- **Centralized governance** and orchestration eliminate the need for separate management stacks per environment, saving time and effort.

Because workloads are intelligently distributed instead of confined to a single sovereign environment, organizations avoid the common cost traps of sovereignty. SMBs can meet regulatory obligations without isolating themselves from innovation. This makes sovereignty operational, not just a legal constraint.

> **Because workloads are intelligently distributed instead of confined to a single sovereign environment, organizations avoid the common cost traps of sovereignty. This makes sovereignty operational, not just a legal constraint.**

CHAPTER 05

Accessible, Affordable AI, Anywhere

> **AI becomes truly accessible when infrastructure isn't locked into a single provider, pricing model, or deployment type.**

Unified Multi-Cloud Orchestration

By combining multi-cloud and hybrid options, teams can choose the right environment for every workload, whether that's any public cloud with available capacity, bare-metal infrastructure for high performance and isolation, or hyperscalers for the latest generation GPUs.

This flexibility directly impacts cost and availability. Instead of overpaying for a hyperscaler region or waiting on limited GPU supply, teams can distribute workloads wherever demands are met.

However, managing multiple infrastructure environments can get complicated fast. Each has their own setup, provisioning, and orchestration requirements. Without a unified approach, teams spend more time wrangling infrastructure than building AI. At present, 78% of engineering groups say they wait a day or more for infrastructure support from SRE/DevOps teams, and 16% report waits of a week or longer.

Wait times for infra support

78%
Wait a day or more

16%
Wait a week or longer

EMMA SOLUTION

Unified orchestration

- One dashboard for all stakeholders
- No more infrastructure wrangling

Source: 2025 State of internal developer portals

emma combines all kinds of AI resources into a single, orchestrated environment behind one powerful, intuitive dashboard suitable for every stakeholder — infrastructure teams, platform engineers, administrators, and finance.

Platform Benefits

- **Unified management:** All cloud environments and bare metal GPUs are visible, controllable, and billed in one place.
- **Fast deployment:** Hybrid environments can be set up and expanded quickly, without complex manual integration.
- **Cost optimization:** Match infrastructure and configurations to precise workload needs, minimizing idle or overprovisioned resources.
- **Growth-ready:** Easily scale deployments across environments as AI projects grow, without re-architecting pipelines.

FINAL WORD

The Future of AI Infrastructure

Flexibility

AI infrastructure doesn't have to be expensive, fragmented, or locked into a single provider. With platforms like emma, organizations gain the flexibility to run workloads wherever it makes the most operational and financial sense, across clouds, bare metal, and hybrid environments.

Control

With a unified orchestration and management layer, even SMBs with lean teams can strategically design architectures that balance compliance, innovation, and affordability without adding operational complexity.

- No tradeoffs
- Compliance
- Innovation
- Predictable costs

GET STARTED

Ready to Build Real AI Workloads?

SPECIAL PROGRAM

Up to $300K in Cloud Credits

For teams actively building and preparing production-ready AI workloads, emma is currently offering access to up to $300K in cloud credits to help accelerate real deployments – no complicated hoops, no hidden strings.
If you're ready to scale something real, apply and put the credits to work.

Apply for Credits →

DEMO ACCESS

Try the Platform First

Spin up multi-cloud and baremetal GPUs yourself in the demo portal and experience the deployment workflow hands-on – **no sales calls, no pressure, just infrastructure at your fingertips.**

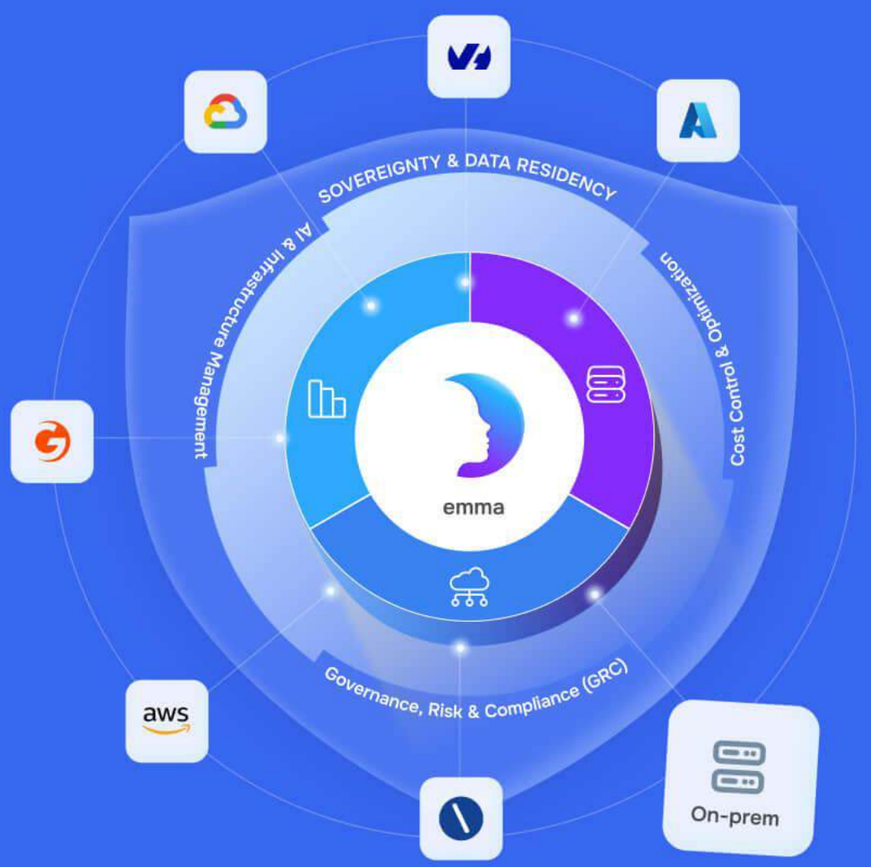
Explore the Demo Portal →

ABOUT EMMA

emma is built for organizations running distributed workloads across hybrid and multi-cloud environments – where operational complexity, fragmented tooling, and unpredictable costs get in the way of building.

emma provides a single policy-driven operating layer that spans hyperscalers, regional European cloud providers, AI-optimized infrastructure, and on-premises environments. It gives engineering, platform, and finance teams unified deployment, governance, policy and cost control from one place.

Sovereignty and compliance are built in through proactive guardrails, not bolted on. With a vendor-neutral architecture, emma ensures organizations can choose the right infrastructure for cost, performance, and regulatory requirements – without lock-in or operational trade-offs.



15+	~90	2021	NO
Cloud providers	Cloud engineers	Founded in Luxembourg	US Cloud Act applicability

CERTIFICATIONS & FRAMEWORKS

emma operates under internationally recognized security and compliance frameworks, including ISO-certified security management and SOC 2 Type II audited controls, with data protection aligned to GDPR and resilience aligned with NIS2 and DORA.

ISO/IEC 27001:2022 CERTIFIED · AICPA SOC · NIS2 · DORA · GDPR

Unified Operations

The entire distributed stack – hyperscalers, sovereign clouds, AI providers, on-premises – operated from a single interface. No context switching. No blind spots.

Automated Governance

The same policies, governance, and deployment logic apply across every environment. Compliance holds everywhere, by design.

Intelligent Workload Orchestration

Workloads are placed across providers based on performance, economics, and policy – not vendor incentives. No commercial bias toward any hyperscaler, neocloud, or AI provider.

Integrated Networking

emma's private networking backbone is built for data-intensive multi-cloud operations. When data moves over emma's private networking backbone, egress costs become predictable compared to the public internet.

Continuous Cost & Performance Optimization

Unified visibility into cloud usage, performance, and cost across every environment. Workload placement decisions that optimise economics without compromising governance.

Extensible by Design

New providers are added without rebuilding how everything is operated. emma's operating layer absorbs the new environment, while existing governance, cost model, and deployment patterns stay intact.

Contact us

info@emma.ms

emma.ms

19-21 Rte d'Arlon, 8008 Strassen, Luxembourg

emma | Make Cloud Work for You