Dweve
Weaving Intelligence

ENGLISH EDITION · JULY 2026

MARKETING BROCHURE · 16 PAGES

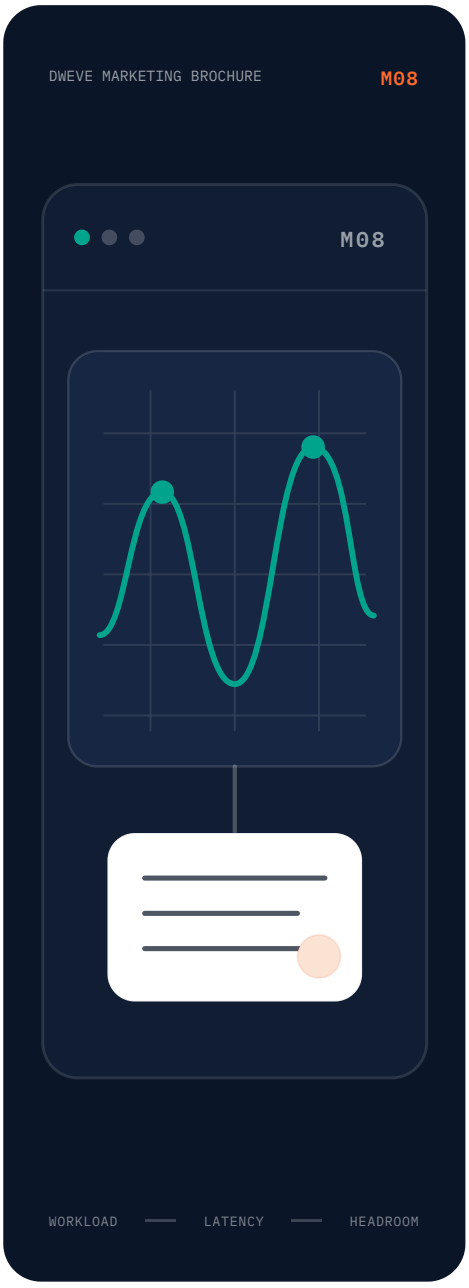
CPU-first AI efficiency

Match compute, energy, deployment, and cost to the work that needs doing.

VISUAL BRIEFING

TECHNOLOGY, INFRASTRUCTURE, FINANCE, AND SUSTAINABILITY TEAMS

ENGLISH



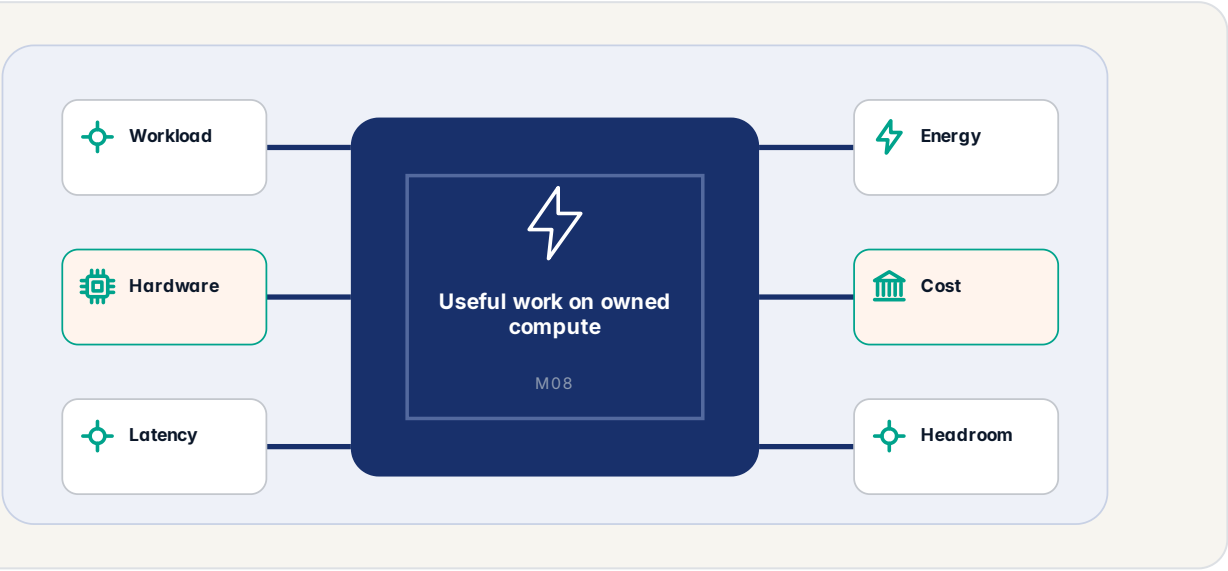
Ask better infrastructure questions before accepting an AI cost curve.

M08

THE PROPOSITION

Useful work per owned resource

~408K implementations, 2,900+ operations, 12 backends, 96% less energy.



How to read this visual: follow useful work on owned compute through the proposition.

MECHANISM

Read the promise against the operating reality

Most AI products are only the part you see. You open the app, type a question, and receive an answer. Behind that simple screen, the work may pass through software, computers, and services owned by several different companies, and the app works only while that whole chain continues to work. Dweve built Core differently. Core is the complete AI machine underneath Dweve's products: the many small operations, calculations, methods, and hardware paths required to make the AI work. That is why Dweve can decide how the work should run instead of sending every task through the same distant system.

Energy consumption is only part of the story. Datacenter GPU lifespans can be as short as 1 to 3 years as AI models advance, creating a hardware replacement cycle that traditional IT equipment never faced. Each DGX H100 server weighs 130 kg, and 26% of companies do not fully recycle their IT assets. Dweve runs on standard servers with longer useful lives and far less waste.

- Workload
- Hardware
- Latency
- Energy

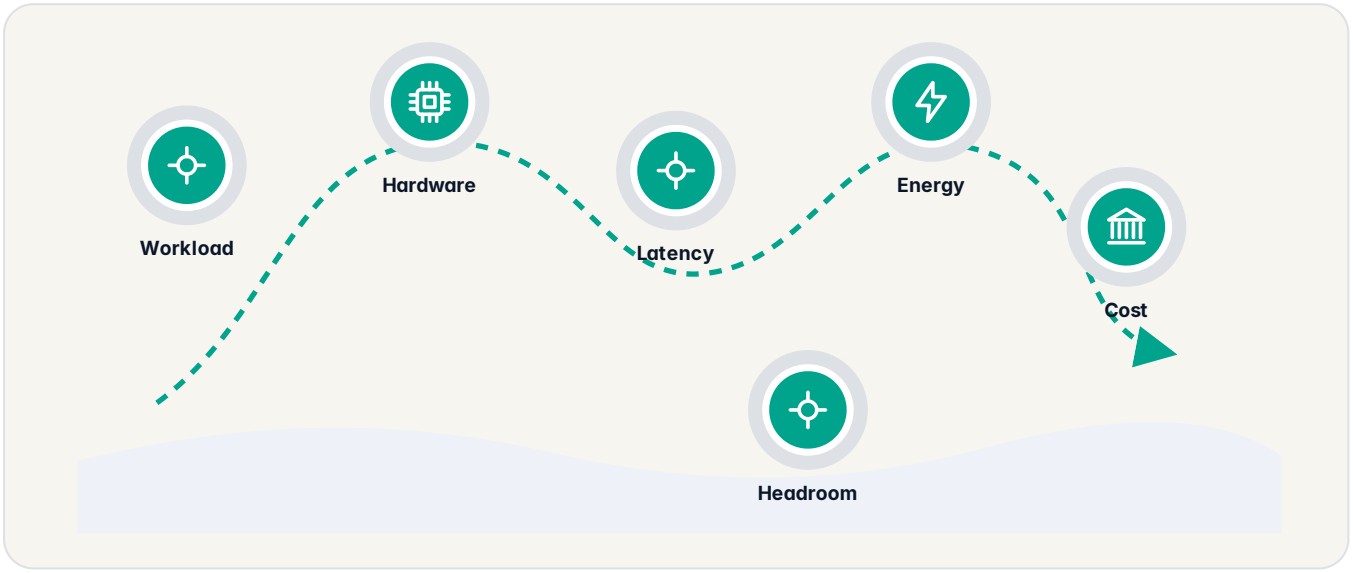
QUESTION TO CARRY

Which part of this proposition changes the outcome people receive today?

USE AND FIT

Where this proposition can earn its place

Start with responsibilities that have a real owner, useful sources, and a consequence worth improving.



How to read this visual: follow useful work on owned compute through use and fit.

DECISION

Separate useful scope from attractive distraction

That is the whole story. Useful work, far less power, kept close to home, and kinder to the world your grandchildren will live in.

When useful AI can run on the machine in front of you or on servers controlled nearby, less work has to travel to a distant data centre. Not every AI task will always run entirely on one laptop. But Core gives Dweve far more freedom to keep work local than a system built only for remote accelerator clusters. Your files can remain on your device for supported local tasks. An organisation can keep processing inside its own data centre. A hospital or public service can run a system inside the environment where its rules already apply.

- Workload
- Hardware
- Latency
- Energy

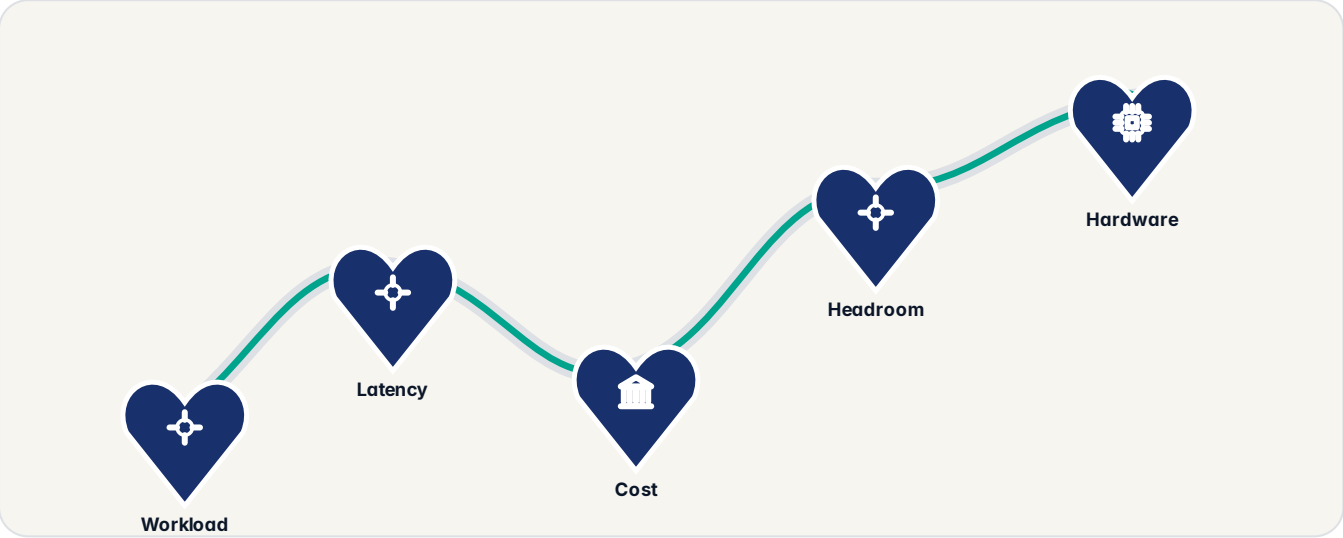
QUESTION TO CARRY

Where does this earn a place, and where should it deliberately not be used?

THE PATH

From first question to an informed decision

A focused engagement should make the next decision easier, not create a larger promise.



How to read this visual: follow useful work on owned compute through the path.

WHY THIS SHAPE WORKS

The workflow, operating boundary, evidence, and decision criteria are considered together from the start.

MECHANISM

Make every stage earn the next one

The best path can differ by operation, datatype, packing, tensor shape, alignment, policy, and available instruction set. Core's dispatcher resolves the implementation at runtime according to the current host and execution contract. Dispatch is not a one-time declaration that the whole model belongs to one device. It is a decision attached to the operation. The selected path can be recorded and verified. The same process may use an AVX-512 binary kernel for one operation, an AVX2 integer path for another, a scalar exact path for a deterministic check, and a GPU backend for a larger floating-point workload.

The 96% is not a tuning trick, it is a different unit of work. A GPU inference is billions of FP16 multiply-accumulates with every parameter firing. The substrate decides in binary constraints, routes to 4 to 8 experts per query, and runs fixed-point integer math on SIMD lanes. Read it axis by axis: the dense matrix multiply that burns the power simply never runs on the hot path.

- Workload
- Latency
- Cost
- Headroom

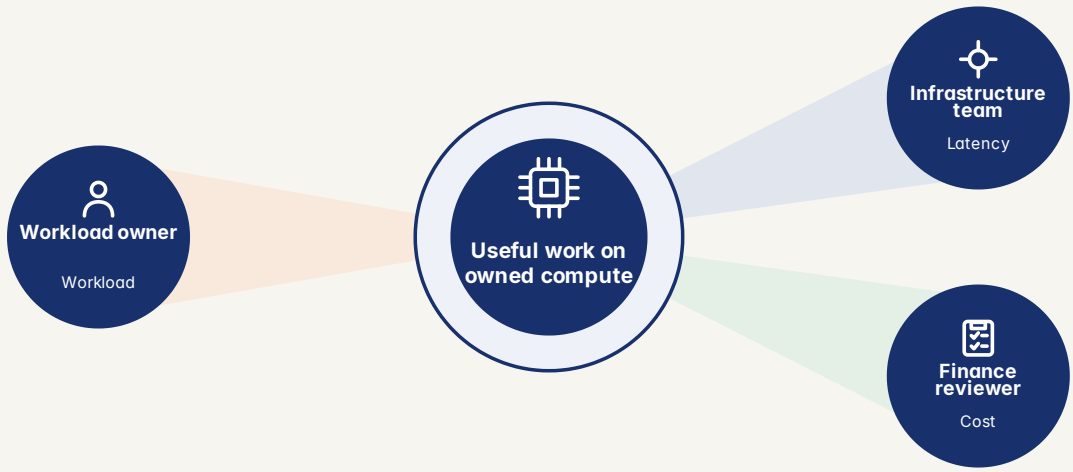
QUESTION TO CARRY

What evidence must exist before the scope is allowed to grow?

THE VALUE PICTURE

One outcome, seen from three responsibilities

Good AI work must make sense to the person using it, the team operating it, and the people reviewing it.



How to read this visual: follow useful work on owned compute through the value picture.

PRIMARY READER

Technology, infrastructure, finance, and sustainability teams

The first conversation.

DECISION SUPPORTED

Ask better infrastructure questions before accepting an AI cost curve.

The next useful step.

DECISION

Read the same outcome from several responsibilities

Big picture aside, here is what a 96% energy reduction means in practice, right now. Plug into a standard office outlet, power serious workloads from a small solar installation, deploy anywhere humans can work, and put advanced AI within reach of every organisation, not just the tech giants.

The 96% energy reduction is not a procurement slogan, it is a property of the substrate. Binary-first compute and correctly-rounded fixed-point arithmetic replace the dense FP16 matrix multiply that makes a GPU stack draw 9,150W. The same work lands at 366W on commodity x86, ARM, RISC-V, or FPGA, air cooled, with no custom accelerator. Here are the per-server figures and the substrate behind them.

- Workload
- Latency
- Cost
- Hardware

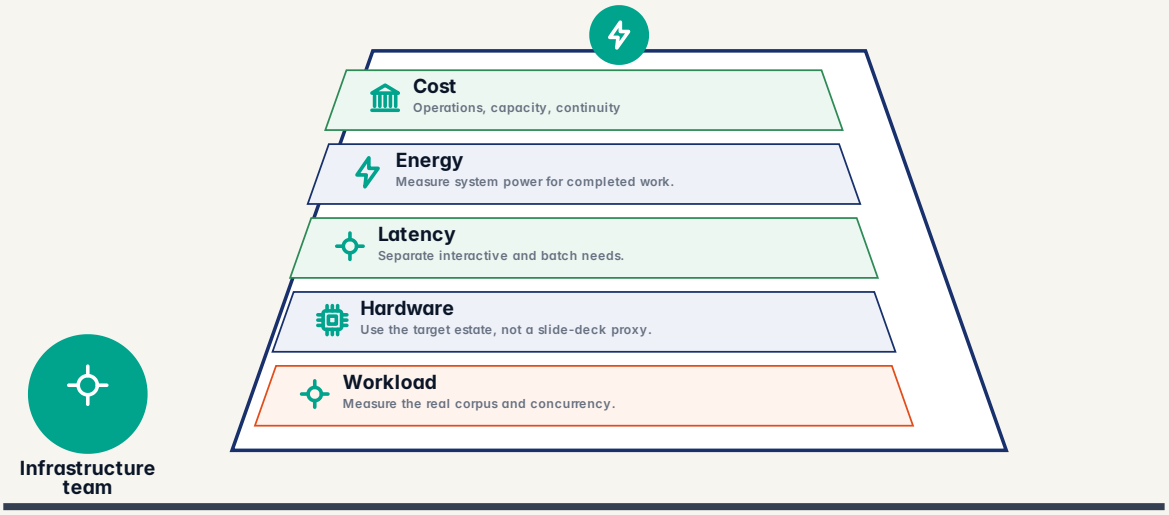
QUESTION TO CARRY

Can the same result satisfy its user, operator, and reviewer?

THE OPERATING MODEL

Control follows the work, from outcome to exit

A credible proposition connects the visible result to the less visible responsibilities underneath it.



How to read this visual: follow useful work on owned compute through the operating model.

MECHANISM

Name the controls behind the simple interface

Most AI platforms cover a slice. One framework defines the model, another library handles the numbers, another system trains it, another tool compresses it, another runtime serves it, and another vendor translates it for the processor underneath. Every missing capability becomes another product to buy. Every handover becomes another place where cost, behaviour, security, and responsibility can drift apart. Dweve Core implements the complete machine beneath AI as one coherent system. Every operation. Every representation. Every machine.

The reduction is not a tuning trick, it is the compute model. Binary-first compute and correctly-rounded fixed-point arithmetic remove the dense FP16 matrix multiply that dominates a GPU stack. The result runs unchanged on commodity x86, ARM, RISC-V, or FPGA, air cooled, with no custom accelerator to fabricate or replace.

- Workload
- Hardware
- Latency
- Energy

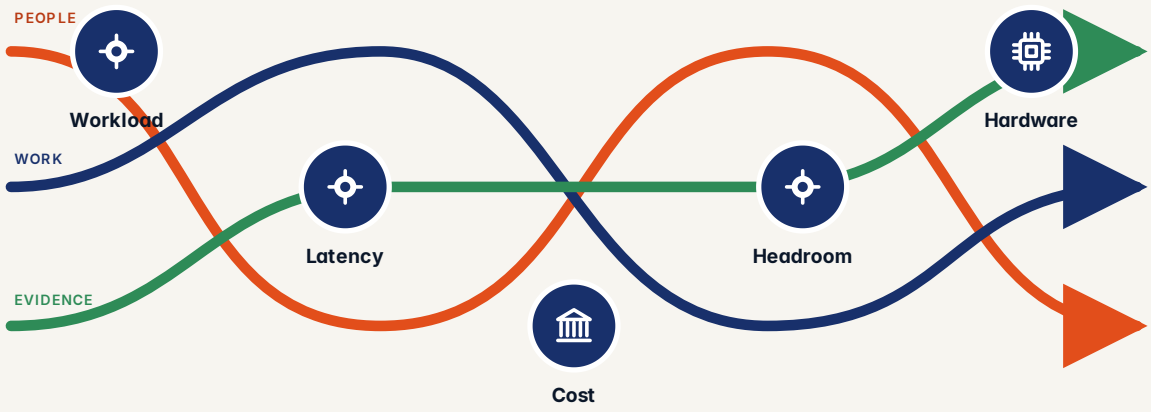
QUESTION TO CARRY

Who owns each control when the interface appears simple?

THE RESPONSIBILITY FLOW

People, platform, and evidence move together

The work becomes easier to govern when each stage leaves a clear hand-off.



How to read this visual: follow useful work on owned compute through the responsibility flow.

DECISION

Keep every hand-off attached to an owner

It is hard to picture how hungry the big AI sheds are, so here is a simple way to see it. One large AI company keeps about 125,000 machines running. Together they pull as much electricity, every hour, as millions of homes. The gentle kind needs about as much as a single household. Tap each one and watch the little houses light up.

Each H100 runs continuously near 700W. The same work maps to SIMD lanes on commodity CPUs or FPGAs, so there is no proprietary accelerator to power, cool, or replace on a 1 to 3 year treadmill.

- Workload
- Latency
- Cost
- Headroom

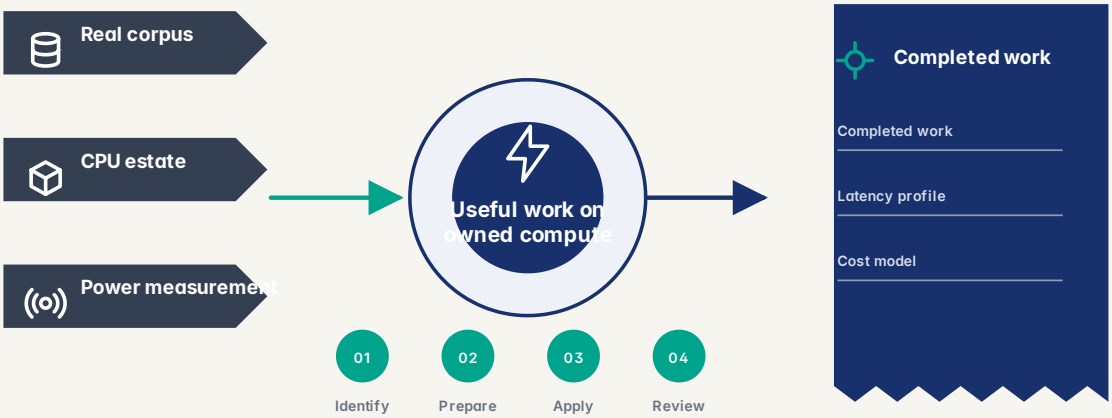
QUESTION TO CARRY

Is every hand-off visible, bounded, and assigned to an owner?

SOURCE TO RESULT

Every useful outcome begins with an authoritative basis

The work stays explainable when source identity, transformation, decision, and hand-off remain visible between systems.



How to read this visual: follow useful work on owned compute through source to result.

MECHANISM

Keep the basis visible as the work changes form

Right now, over 12,000 Dutch businesses cannot get power connections until the 2030s. AI data centres are consuming entire city grids. We built a different path. Neurosymbolic AI, combining neural learning with symbolic reasoning, uses 96% less energy and delivers explainable intelligence. This is how we keep the lights on for everyone.

In your office building. On a hospital floor. In a rural clinic. On a ship. If humans can work there, this AI can run there. No data centre required.

- Identify
- Prepare
- Apply
- Review

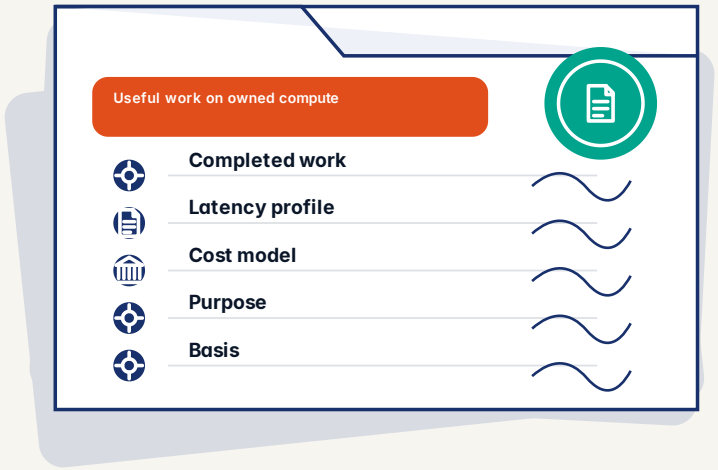
QUESTION TO CARRY

Can the result still be connected to the exact basis that shaped it?

EVIDENCE PACKET

The record should answer the next difficult question

Capture enough to explain the result, inspect authority, investigate failure, and make the next decision without rebuilding the story from memory.



How to read this visual: follow useful work on owned compute through evidence packet.

DECISION

Preserve the record needed by the next question

Lower board-level draw means standard air cooling is enough. No liquid loop, no operational water footprint, and one fewer class of mechanical failure point.

A system that can weave many capabilities must also know when none of them can support a defensible result. Loom can surface that state explicitly. It does not need to force every cognitive path into a fluent answer. The perception may be incomplete. The memory may not contain the required fact. Retrieval may return weak evidence. The constraints may conflict. The solver may return unknown. The experts may disagree.

- Purpose
- Basis
- Authority
- Action

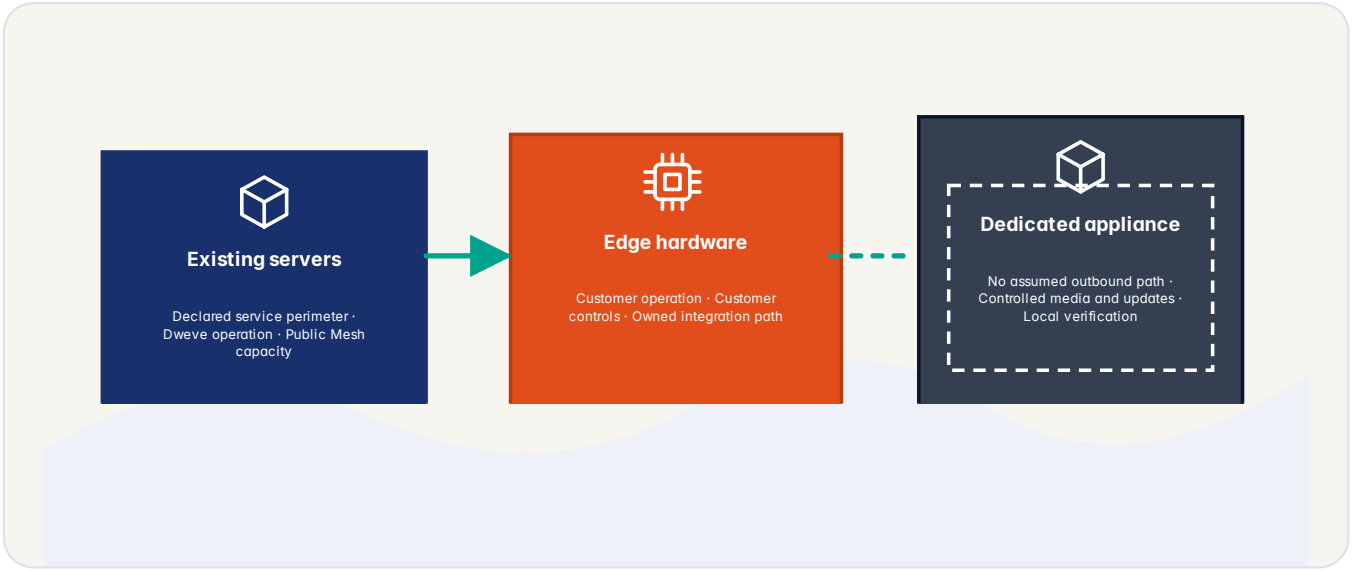
QUESTION TO CARRY

Does the retained record answer the next difficult question without reconstruction?

DEPLOYMENT CHOICE

Choose the operating posture around the work

Location is only one dimension. Decide who operates, who holds keys, which dependencies exist, how updates arrive, and what continuity requires.



How to read this visual: follow useful work on owned compute through deployment choice.

MECHANISM

Compare operating responsibility, not location alone

The work is the same either way. The only difference is how much power gets used up, and how much is left over for the homes, schools, and shops around you.

Infrastructure decisions should not force a model rewrite. Core keeps the model, operator surface, numerical design, and execution contract inside the same system while the deployment posture changes. The selected backend may differ. The organisation does not need to reconstruct the AI stack around it.

- Begin through managed Fabric
- Run on customer infrastructure
- Close the environment
- Workload

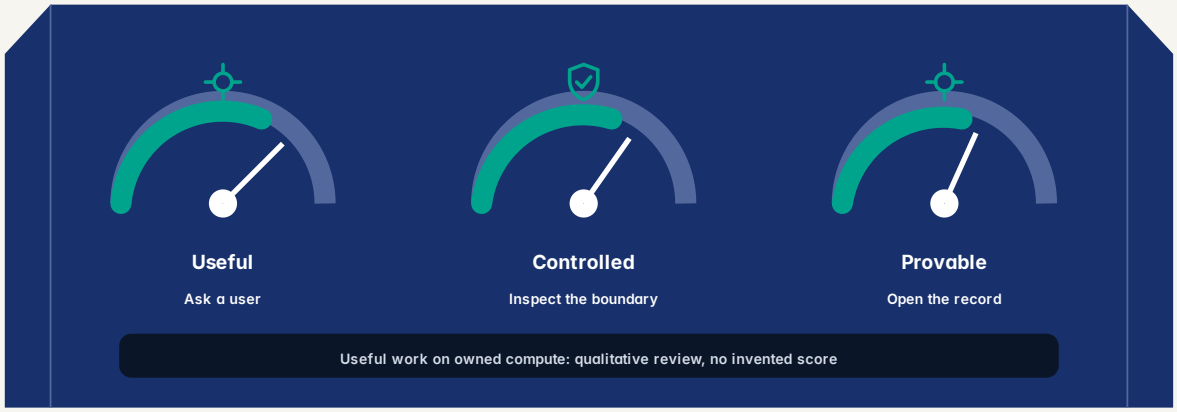
QUESTION TO CARRY

Who holds infrastructure, keys, updates, logs, recovery, and the exit path?

WHAT TO LOOK FOR

Three signals that the proof is earning trust

Progress is a useful outcome under understood control, supported by reviewable evidence.



How to read this visual: follow useful work on owned compute through what to look for.

DECISION

Use signals that can change the next decision

Sovereignty is an engineering property here. With no FP16 accelerator or foreign dependency in the hot path, the full stack runs on-premise from a standard outlet, air-gapped without a network, in an EU sovereign cloud, or Mesh-distributed across federated EU nodes. The measured 366W per server is what lets advanced AI stay inside a building you control.

It is built and run in Europe, often close to where you live, under the rules you already trust to look after your privacy.

- Useful
- Controlled
- Provable
- Workload

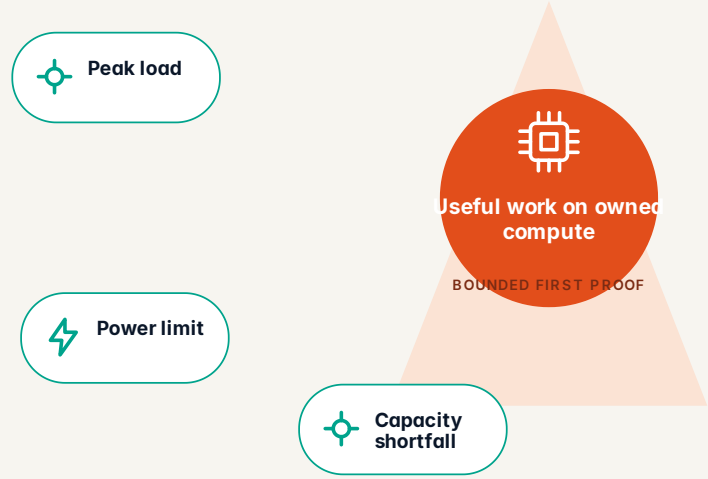
QUESTION TO CARRY

Which signals would change the next decision rather than merely impress the room?

HONEST LIMITS

Design failure before the first success demo

Visible failure is more useful than a polished result whose limits remain unknown.



How to read this visual: follow useful work on owned compute through honest limits.

MECHANISM

Design the failure before presenting the success

One path leads to limits and rationing. The other leads to abundance and accessibility. The technology exists for both, and the choice is ours. Efficiency and capability are not opposites. With a 25x efficiency gain, you can deploy AI 25x more widely with the same infrastructure.

Let us look at the numbers, not to scare you, but because understanding the problem is the first step to fixing it. Here is what one NVIDIA DGX H100 server actually draws, and what that means once you multiply it across an industry.

- Peak load
- Power limit
- Capacity shortfall

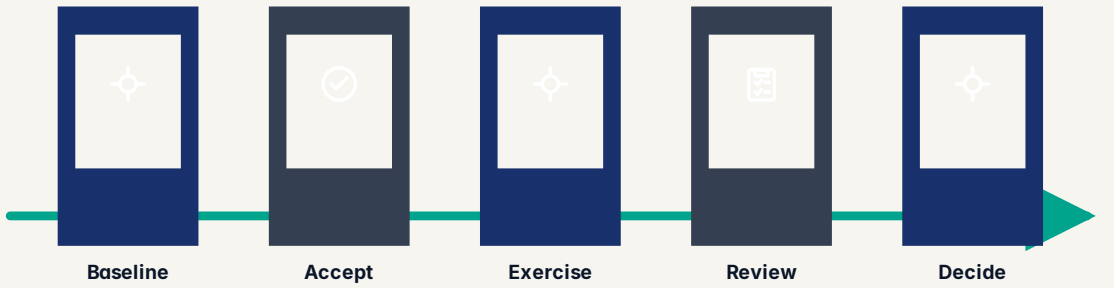
QUESTION TO CARRY

What happens when a source, permission, target, model, or supplier is unavailable?

ACCEPTANCE PATH

A proof should close with a decision, not applause

Agree the decision path before the demonstration so a strong or weak result can change what happens next.



PROCEED · NARROW · HOLD · REJECT

How to read this visual: follow useful work on owned compute through acceptance path.

DECISION

Close the proof with an owned decision

A creative request can allow variation. A test can use a fixed seed. A supported hard-deterministic operation can use the strict path required for replay. The product does not have to accept accidental variation merely because the underlying AI software was never designed to control it.

One NVIDIA DGX H100 draws 9,150W: roughly 5,600W across eight H100 GPUs, around 500W for CPU, memory, and storage, and a further 3,050W of datacenter overhead at PUE 1.5x. At 60% utilisation that is 48,139 kWh a year. Almost all of it goes into dense FP16 matrix multiply and the cooling needed to survive it. Remove that workload from the hot path and the entire power and thermal budget collapses. That is the lever the substrate pulls.

- Baseline
- Accept
- Exercise
- Review

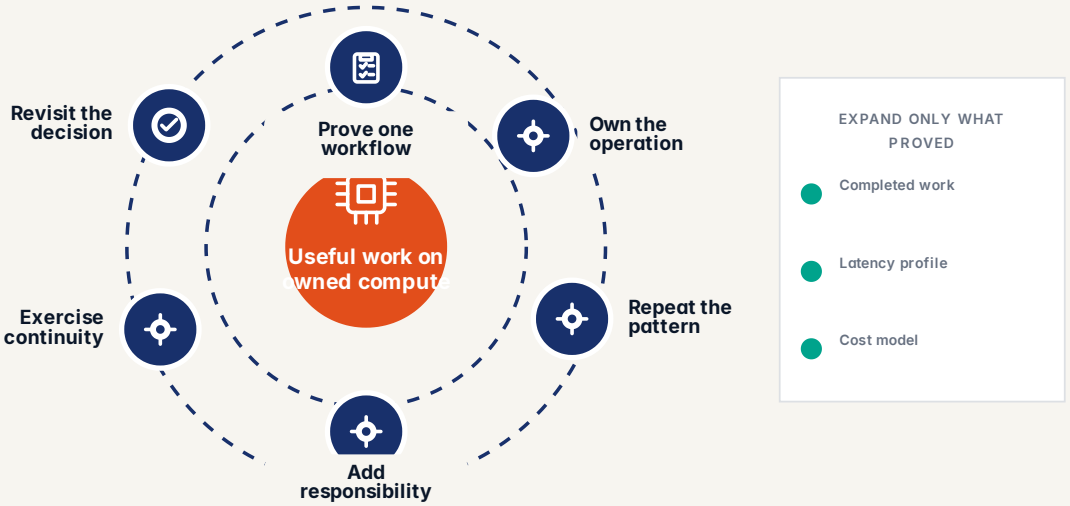
QUESTION TO CARRY

Are success, counter-metrics, failure cases, and stop conditions agreed first?

ADOPTION

Expansion is a sequence of owned choices

Move from one bounded result to a repeatable operating capability without making the scope vague.



How to read this visual: follow useful work on owned compute through adoption.

MECHANISM

Expand only what proved reusable

Here is the good news. Efficiency and capability are not opposites. With a 25x efficiency gain, you can deploy AI 25x more widely with the same infrastructure. We are not choosing between progress and sustainability. We are choosing between

A licensed team may operate Core on organisation-controlled clusters or workstations. A regulated programme may require a fully isolated estate. A product may move part of its inference to the edge, while a browser application may need a local execution target. Core keeps those licensed operating postures inside one machine without claiming that every backend supports every cell. Core is not a managed-service product.

- Prove one workflow
- Own the operation
- Repeat the pattern
- Add responsibility

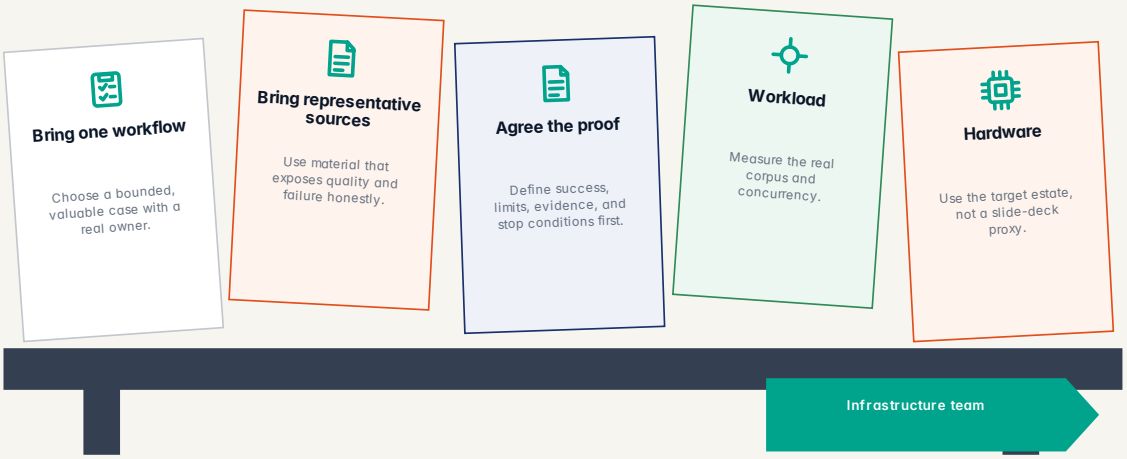
QUESTION TO CARRY

What can be reused safely, and what requires a new decision and proof?

WHAT TO PREPARE

Bring the ingredients for a useful first session

A small amount of real context produces a better conversation than a broad catalogue of hypothetical features.



How to read this visual: follow useful work on owned compute through what to prepare.

KEEP IT BOUNDED

Sensitive production data is not required for the first conversation. Bring representative material and the rules that define a useful result.

DECISION

Bring enough reality to make the session useful

Grid congestion already costs the Dutch economy EUR 40 billion every year. Upgrading Europe's grid to handle the current trajectory of AI growth will require more than EUR 200 billion by 2040. Meanwhile, hospitals, schools, and new homes wait years for power connections while AI data centres get priority access to shrinking grid capacity. The numbers tell a story we need to understand, and they are unfolding across Europe today.

There are two ways to build AI. One way is hungry: it lives in giant sheds full of machines that drink as much electricity as a small town. The other way is gentle: it does the same helpful work on an ordinary computer in an ordinary room. We build the gentle kind. Here is the difference, side by side.

- Bring one workflow
- Bring representative sources
- Agree the proof
- Workload

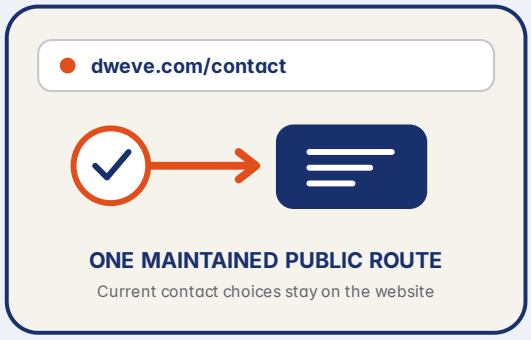
QUESTION TO CARRY

Which real sources, rules, owners, and examples will make the first session honest?

THE INVITATION

Continue with the question that matters

Scope a briefing, demonstration, or proof around the workflow and decision that matter to your organisation.



CONTINUE ONLINE

dweve.com/contact

Current route and contact choices

START A CONVERSATION

Continue through the current contact page

Scope a workflow or ask a technical question, then use the contact page to begin.

START	**dweve.com/contact**	Use the maintained public contact route
BRING	**One bounded question**	Workflow, audience, decision, and desired evidence
CHOOSE	**Briefing · demonstration · PoV**	Select the smallest useful next step
VERIFY	**Current scope and terms**	Confirm availability, pricing, and responsibilities before commitment

MECHANISM

Turn the brochure into one bounded next step

The per-server figure only matters because it multiplies. As of October 2025, one operator alone runs about 125,000 of these dense FP16 servers, drawing roughly 1.14 gigawatts continuously, enough for 3.8 million Dutch households against the country's 7.9 million. The arithmetic is linear: cut each server 96% and the datacenter count, the grid load, and the e-waste fall with it.

There is only so much electricity to go around. When giant AI sheds grab a big share of it, there is less left for everything else, and the price of what remains goes up. In the Netherlands, more than 12,000 businesses are already waiting years for a power connection, and new homes and hospitals wait behind them. The good news is simple. If AI uses far less power, this problem gets much smaller, and that is exactly what we set out to do.

- Start: dweve.com/contact
- Bring: One bounded question
- Choose: Briefing · demonstration · PoV
- Verify: Current scope and terms

QUESTION TO CARRY

What is the smallest useful next conversation Dweve should prepare?