Intelligent Internet

Intelligent Silicon

The Silicon Thesis

The physical case for sovereign intelligence.

July 2026

CONFIDENTIAL DRAFT

Contents

How to read this pair. The Strategic Thesis and this document make one argument at two altitudes. The first identifies the institution, the Champion, a utility-class company a jurisdiction can license, own, and hold to account. This one identifies the artifact that institution dispenses, a model manufactured into silicon, certified, escrowed, and owned. The utility is the institution; the good is what it meters. Policy and procurement readers should read the Strategic Thesis first, then Sections 6--7, 16, and 21 here. Technical readers should go straight to Sections 10--14 and Appendix A. Investors should read Sections 13, 17--19, and the gates in Section 23, and every measured claim's anchor in Appendix B. Key terms, including those inherited from the Strategic Thesis, are defined in Appendix C.

Part I

The Diagnosis

Before the answer, the physics. A token has a manufacturing cost measured in joules, and almost none of what is paid for a token pays for joules.

1 The Core Diagnosis

Every strange fact about the AI economy, the collapsing prices beside the widening losses, the trillion-dollar buildouts financed against uncertain demand, the geopolitics of a memory component, all of it resolves under a single observation. Inference is financed, priced, and regulated as a cloud service. It is actually a manufactured good.

This is the Strategic Thesis's diagnosis carried one layer down. That document argued the institution is miscategorised, a regulated utility financed as enterprise software. This one argues the product is miscategorised too. The utility is the institution; the manufactured good is what it dispenses, and both corrections are needed for either to hold.

Treated as: a cloud service	**Actually is: a manufactured good**
Rented per API call	Produced at marginal cost
Priced on scarcity of access	Priced on silicon and power
Assumed never to depreciate, like software	Depreciates like equipment, on a schedule
Financed on hyperscaler balance sheets	Financeable like plant, against contracts
Supply constrained by HBM allocation	Supply constrained by logic wafers, which are abundant
Quality means access to a model	Quality means a certified artifact

Read the market through the left column and it is incomprehensible. The price of inference at fixed quality falls at a median of **fifty-fold per year**, halving roughly every two months. Across model classes the decline ranges from nine-fold to nine-hundred-fold, faster than any comparable technology transition on record; the older estimate of ten-fold per year now reads as the conservative bound. Yet the providers of this collapsing-price good lose money, because per-token prices fell ten-fold while reasoning workloads raised token consumption a hundred-fold. Supply is rationed not by fabrication capacity but by the allocation politics of high-bandwidth memory. And an entire leveraged financing structure, the neocloud, exists for one reason: you can only lease what you cannot manufacture.

Read the market through the right column and each anomaly becomes ordinary industrial economics. This is a good whose input costs are silicon and joules, sold through the pricing conventions of software, by institutions structured for rent extraction rather than production. The dysfunction is not mysterious. It is a category error, priced.

One objection follows immediately, and it deserves to be posed at full strength before anything is built on this diagnosis. If the spot price of inference halves every two months, how can any fixed artifact, a model frozen into hardware, ever pay for itself before the market undercuts it? That objection is the payback clock, and it ticks over every page of this document until Section 18 answers it.

The correction is the right industrial form for a manufactured good, not a cheaper service.

2 The Hidden Mechanics of a Token

What does a token actually cost to make? Follow the joules. A low-precision multiply-accumulate, the atom of inference arithmetic, costs on the order of a picojoule. Fetching the weight that feeds it from off-chip memory costs on the order of hundreds of picojoules. The ratio, not the absolute values, is what matters, and it is measured rather than modelled: studies of production inference confirm that data movement, not arithmetic, dominates the energy of serving a language model. A modern GPU serving a large model spends the great majority of its power budget not computing the answer but transporting the model to the computation, every layer, every token, every time.

Everything else follows from that one imbalance. Because weights must stream from memory, the memory's bandwidth is the binding constraint: the memory wall. Because the wall binds, the industry buys the fastest memory that exists, stacked, exotic, supply-constrained HBM, and the memory becomes the geopolitical chokepoint. Because weight movement is amortised across users, providers batch aggressively, sacrificing per-user speed for fleet throughput: the batching tyranny. Because the economics only close at high utilisation, the fleet must run hot around the clock, and the financing structures of Section 1 are erected on that requirement. The whole tower stands on moving data that never changes.

The category error underneath it

Underneath the tower sits the buried absurdity: inference weights are **constants**. They are written once, at the end of training, and then read, billions of times, unmodified, for the life of the deployment. The industry stores its one great constant in the most expensive variable storage ever manufactured, and pays refresh power, leakage power, and allocation politics for a mutability it never once exercises. The HBM shortage is a tax on mutability that inference never uses.

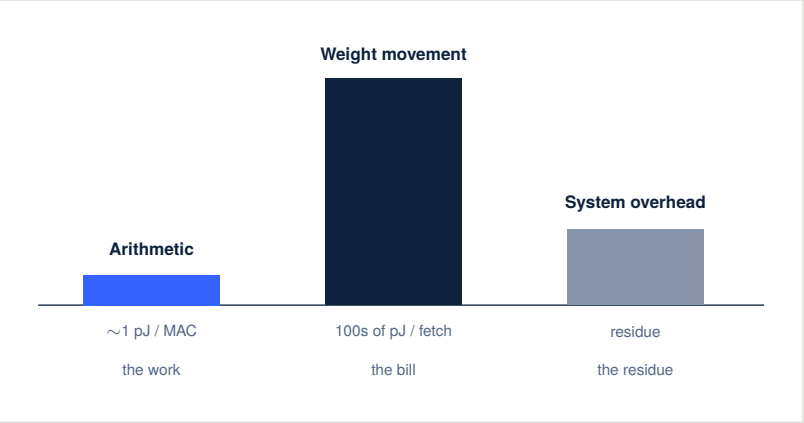

Figure 1 The joule waterfall of a served token. Arithmetic (~1 pJ) is the work; weight movement (hundreds of pJ) is the bill; system overhead is the residue. The ratios are modelled; that movement dominates is measured.

Stand back from the mechanism and the shape of the problem is simple. The expensive part of a token is not thinking; it is shipping. The thing being shipped never changes. And a constant that never changes does not need a vehicle; it needs a location. The GPU does not compute intelligence expensively; it ships intelligence to the compute expensively, forever, one token at a time.

3 The Token Wave Meets the Grid

The demand side makes the cost structure a civilisational problem rather than a commercial one. Token consumption is not growing along the gentle curve of past internet services. Agentic workloads multiply the tokens per task by orders of magnitude, and reasoning models alone have raised consumption on some workloads a hundred-fold. The wave this document plans for, a thousand to ten thousand times today's volume, is the Strategic Thesis's base case, not its extreme.

Run the arithmetic at wave scale and rented FLOPs fail on their own terms. At the tens-to-hundreds of millijoules per token that general-purpose serving is estimated to cost today, wave-scale volume collides with national grids: projected datacenter electricity demand already strains against generation buildout, and inference is the compounding term. Either the joules per token fall by orders of magnitude, or the answers get rationed, by price, by queue, or by jurisdiction.

The geography follows the same forcing. Computation migrates to where power is cheap; the Strategic Thesis draws that map for training. But a **manufactured** token obeys a different logistics: it moves to wherever the certified silicon sits, a national datacenter, a workstation, and, because a printed model can run from a solar panel, places the grid has never reached at all (§16). Sovereign buyers have already begun procuring on just these terms, with custody language attached and money committed; the evidence is tabled once, in Section 21. The wave does not negotiate with the grid.

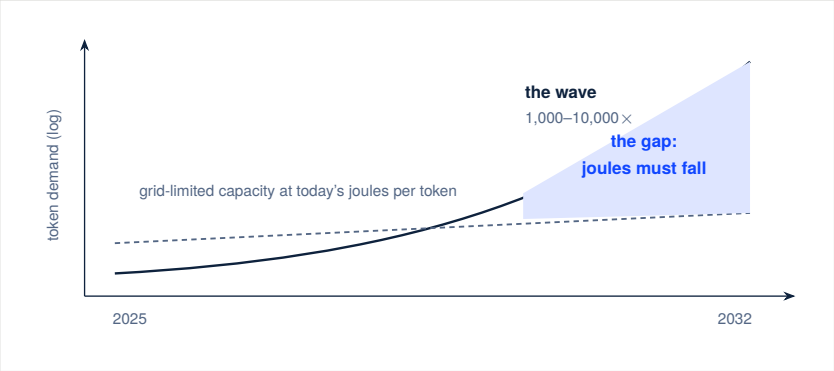

Figure 2 The wave meets the grid. Token demand (the Strategic Thesis's 1,000–10,000× base case, log scale) against grid-limited serving capacity at today's joules per token. The shaded gap is not optional: either joules per token fall by orders of magnitude, or answers are rationed by price, queue, or jurisdiction.

4 The Fixed-Function Precedent

The semiconductor industry has run this experiment before, at scale, with public results. When Bitcoin's proof-of-work froze a single function, SHA-256, into permanent economic importance, silicon specialised against it, and the efficiency of computing that function improved by orders of magnitude within a decade, the exact multiple depending on which ASIC generation is the baseline. GPU mining, the general-purpose incumbent, did not decline gracefully; it went extinct within months of the first ASICs shipping. The same fleets, the same operators, and the same power-site discipline that mining built are now converting to AI workloads in a multi-billion-

dollar wave, proof that fixed-function infrastructure converts, and that its operators outlive any one function.

The market has, in the past year, begun to price specialisation itself: an inference-hardware asset sale to the largest chipmaker on earth, a record public offering, a billion dollars of contracts signed against unshipped silicon. The rows are tabled once, in Section 18. What the market has not yet priced is the step this document describes.

The standing objection to the precedent is familiar: AI workloads don't freeze. Models improve monthly; architectures turn over yearly; freezing a model into hardware means shipping obsolescence. Part of the answer is economic, and its one-line form can be given now: a weight update is a metal-layer respin costing about a million dollars and taking weeks, not a new chip; obsolescence becomes a product line, not a stranding event (§15). But the deeper answer is conceptual, and it changes what kind of object a model is. The next section argues something stronger: a deterministic model is already a frozen function; it just hasn't been printed yet.

5 The Learned Function

A Bitcoin ASIC prints a human-written function. An etched model prints a learned one. The difference is semantic, not industrial.

Consider what a trained model actually is, once training ends. Gradient descent has searched a space of programs and returned one: a fixed composition of linear maps and nonlinearities, specified to the last parameter. Fix the weights, the sampling seed, the step schedule, and the conditioning interface, and the model is a **deterministic function.** The same input produces the same artifact, bit for bit, forever. It differs from SHA-256 in what it computes, not in what kind of thing it is. One maps a message and a nonce to a digest. The other maps noise, a prompt, and a schedule to a paragraph, a triage decision, a translation. Both are fixed functions repeated at civilisational scale, and every economically important fixed function in computing history has ended the same way.

Function	Input	Repeated operation	Output	Silicon form
SHA-256	message + nonce	compression rounds	digest	mining ASIC
Video codec	frame stream	signal transform	media	codec block, in every phone
Seeded diffusion	noise + prompt + schedule	denoising rounds	image, latent, artifact	learned-function ASIC
Competent inference	context + retrieval + policy	neural block transition	answer, action, route	the etched model

The codec row is the quiet proof that this transition happens even to functions committees revise every few years. The revision cadence did not keep video decoding in software; it created gener-

ations of codec silicon. The learned function's revision cadence is faster, and its silicon revision cost is smaller too, because a model update touches masks, not architecture (§15).

The rule the table implies is mutability discipline, not *etch everything*. A function earns its silicon to the degree that it is stable and repeated; what changes with every input stays in software; what is genuinely new is sent elsewhere. Three words carry the architecture of everything that follows: **print the prior, stream the state, route the novelty.** The prior is the stable, expensive, endlessly reused operator, a language, a procedure, in time a world model. The state is the live, compact, per-request context. The novelty is the case the prior has never seen. Every domain this document touches, from the citizen's question to the sovereign's expert library, is that same three-way split wearing a different interface. (A disambiguation: the learned *function* here is distinct from a learned *hash*, how a mixture-of-experts router behaves over pattern space.)

Once a deterministic function becomes economically important enough, it stops being software and becomes silicon. Gradient descent has been manufacturing such functions for a decade; lithography simply hasn't caught up.

The road from here

The diagnosis leaves four questions standing, and the rest of the document takes them in order. What physical form corrects the category error, and why will regulated buyers demand that form, Part II. Which models are safe to freeze, and how one is built for freezing from the first training token, Part III. What the silicon actually is, rung by rung, and what it costs, Parts IV and V. And who sells it, who buys it, and why the window is now, Parts VI and VII. One objection travels the whole road: the payback clock of Section 1, which Section 18 stops.

Part II

The Physical Answer

The institutional form of the Strategic Thesis has a physical form. A model whose weights are lithographically fixed is a utility asset: certifiable, auditable, escrowable, and owned.

6 The Model as Manufactured Good

Etch the weights into mask ROM and the object changes category. A GPU **stores** a model: the model is data, endlessly transported to the compute that consumes it, and Section 2's bill is the freight. A printed model chip **is** the model: the weights are a spatial arrangement of the compute itself, a pattern in metal and vias, and there is nothing left to transport.

The memory hierarchy, HBM, DRAM controllers, cache towers, the machinery that exists solely to move the unmoving, is not optimised; it is deleted. Two memories leave the die, not one. The weights become lithography, a pattern in metal rather than a payload in DRAM. And the history, the growing per-token cache that an autoregressive machine accumulates to avoid recomputing its own sequence, is gone too, because a bounded-state model has no such cache to keep. Long context arrives instead as conditioning, compressed by the host and Radiant, the Strategic Thesis's retrieval layer, the governed store of facts and records the model consults rather than memorises, into a bounded object and streamed in. That is work, but bounded work, and it never grows a structure on the die. The only memory left on silicon is the bounded state of the current thought.

This deletion is a gain in governance, not only in silicon. The cache it replaces was never institutional memory anyway, only opaque per-user activation. History held in Radiant is cited, versioned, and auditable, which is what a certified system requires and a cache can never be. And with the hierarchy goes the right unit of comparison: not FLOPs, which measure a general-purpose machine's potential, but **lifetime useful outputs per wafer-dollar and watt**, which measure what the artifact was built to do.

The density physics is now independently anchored. A 2026 ternary read-only-memory accelerator study, the ternary-ROM measurement, cited throughout this document, measures **15.0 MB/mm² at a ~70% zero-weight ratio, 3.3× the density of SRAM on the same node, with 200 TB/s of on-chip bandwidth**, against an SRAM baseline of 0.021 μm² per bit. This document's working figure of 60–80 Mbit/mm² is conservative against that measurement by a factor of 1.5–2×, and is retained as margin, part of which prices the gap between the study's ~70% zero ratio and the ~50% a trained model shows (E.13). For calibration: 200 TB/s is roughly forty times the bandwidth of the memory stack the incumbent architecture queues for, available here without a single off-chip wire, because the "memory read" and the computation are the same physical event.

Every bit in the cheapest technology its mutability permits

The design rule that organises the whole machine is a mutability audit. Sort every bit the system holds by how often it changes, and place it accordingly:

What the bit is	How often it changes	Where it lives
The operator: the model's weights	Per generation	Mask ROM, shell and residual planes
The routing (sparse generations)	Per generation	Wires and an argmax, no memory at all
State in flight: the block buffer	Every cycle	Megabytes of on-die SRAM
Task adapters	Per signed update	A small, signed SRAM island
Conversation context (KV)	Per request	Commodity DDR on the host
Knowledge: facts, records, curricula	Continuously	Retrieval, via Radiant in the datacenter and signed packs in the appliance (§16)

Nothing on the die is writable except the block buffer and the signed island. There is no weight state to poison, no exfiltration path wider than the answer itself, and no idle draw, a ROM holds its bits for free, so the artifact costs nothing between questions and is instant-on when one arrives.

The audit is not particular to language. Run the same sort over a generative world model and the proportions only sharpen: the mutable scene state is megabytes, while the world prior it is decoded against runs to terabytes per second of weight traffic on the program's modelling, the identical tax, one domain over. The near-term artifact is the language model in front of us; the law it obeys is general.

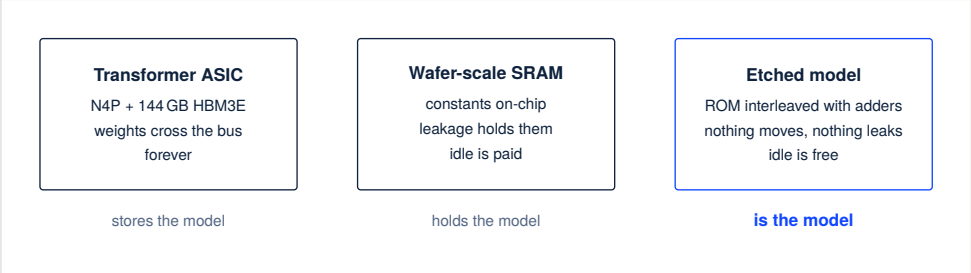

Figure 3 Three ways to hold a constant. The incumbent transformer ASIC keeps the hierarchy (N4P plus 144 GB of HBM3E, by the vendor's spec): weights cross the bus forever. Wafer-scale SRAM holds the constants on-chip but pays leakage to keep them. The etched model, ROM interleaved with adders, deletes the question: nothing moves, nothing leaks, idle is free.

7 Certification Inevitability

Regulated intelligence will be bought the way regulated everything is bought: as a certified, version-locked artifact. Five structural conditions force this, and none of them is speculative. Conformity assessment requires an **unchanging object** to assess. Procurement law requires an **auditable artifact** to specify. Liability requires **determinism**, the same input yielding the same output at the same version. Sovereignty requires **physical custody** of the thing licensed. And post-market surveillance requires a **version lock**, because an artifact that silently improves is also an artifact that silently changes. A hosted model, updated at the provider's discretion, fails all five by construction. A lithographically frozen model is the only form of the technology that cannot drift, because drift would require a fab.

One clarification keeps the claim honest. Immutability satisfies the artifact-side conditions; it does not complete a dossier. The process obligations of a conformity regime, data governance, human oversight, post-market monitoring, live on in the host and Radiant layers, where they belong. The frozen artifact removes drift from the assessment problem; it does not remove the assessment.

The regulatory calendar is now legible. The EU AI Act's prohibitions took effect in February 2025 and its general-purpose obligations in August 2025; transparency duties and the general-purpose penalty regime, fines to €35M or 7% of turnover, arrive on 2 August 2026. The Digital Omnibus agreed in May 2026 deferred the high-risk conformity deadlines to December 2027 for stand-alone systems and August 2028 for product-embedded ones. Honesty requires stating what the deferral means: it lengthens the build window and softens any claim that certification forces freezing *now*. What replaces the deadline as the forcing function is already visible, sovereign procurement running ahead of regulation with custody language and money attached (§21), and the category-default dynamic: the first certified artifact defines what conformity assessment looks like for everything after it.

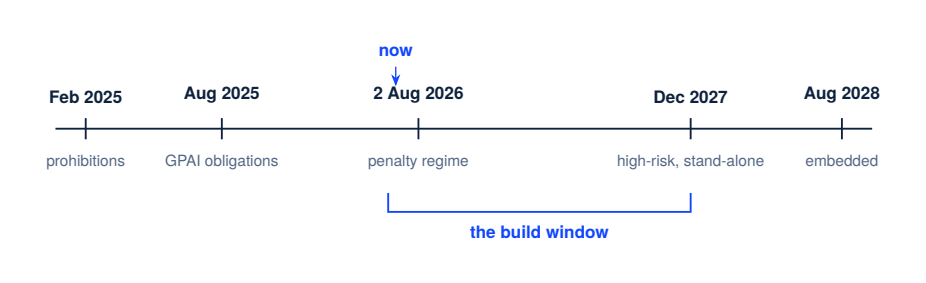

Figure 4 The certification calendar and the build window. The EU AI Act's dates: prohibitions (Feb 2025), GPAI obligations (Aug 2025), the penalty regime (2 Aug 2026), and the deferred high-risk conformity deadlines (Dec 2027 stand-alone; Aug 2028 embedded). The Digital Omnibus lengthened the window; sovereign procurement, not the deadline, is the near-term forcing function.

One mechanism is honestly open: whether a weight-changing metal respin constitutes a "substantial modification" requiring re-assessment has not been clarified by any notified body. The question is two-sided. If yes, a recall is a respin plus a re-certification, a budgeted operational event (§15). If no, the incremental-conformity design of Section 13, re-certifying only the changed expert banks against an unchanged base, becomes a decisive advantage. The contrast case is instructive either way. The FDA's predetermined change-control pathway exists because adaptive models strain assessment regimes built for fixed devices. The etched model does not strain them. It is what they were built for.

Two consequences follow. First, the certification dossier for an etched generation classifies every design variable as forced, interior, plan-of-record, or free, each with its derivation, because a design whose every choice has a derivation is a design whose every choice can be audited. Second, the protocol health aide of Section 16 is plausibly a high-risk system under Annex III, which is a feature: the appliance **inherits** the conformity machinery rather than evading it, and the machinery is the moat. Recall equals respin, a corrective action is a metal-layer event, budgeted as operations, not an existential one.

8 The Historical Analog

The etched card is the same thermodynamic object as a mining ASIC, a fixed function, printed, run at the physics floor, and the opposite business. The comparison repays precision, because the mining industry already trained the operators, built the power-site discipline, and proved the fleet economics this document borrows. What it never had was a durable claim on the value its silicon produced.

Three inversions separate the two. The miner earns a **commodity reward** set by a global market; the etched card earns **contracted revenue** set by a utility agreement. The miner's gains are **confiscated by difficulty adjustment.** The protocol raises the bar as efficiency improves; the card's gains are **protected by certification.** The conformity assessment raises the bar against everyone else. And for the miner, **the next chip devalues the fleet**; for the card, the next chip is the next contract, because the installed base upgrades by respin rather than replacement.

Figure 5 The lineage and the inversion. Same physics, same fleets, same power sites; opposite claim on value. The respin is a revenue event.

The secondary analogs each contribute one load-bearing precedent. Secure elements, for silicon whose **immutability is the product.** Billions of chips bought because they cannot be rewritten. Avionics, for version-locked certification as a priced, survivable process. And the mask-ROM game cartridge, for an industry that shipped fixed code at consumer scale, updated it by revision, and routinely shared ROM content across titles, the direct ancestor of the banked base wafers that serve many models from one design (§15).

Societies already buy frozen silicon, in volume, on purpose. They have never yet been offered frozen intelligence.

Part III

What Gets Etched

The recipe, each step with published precedent, ordered so that every failure is discovered in software, and the mathematics that says why each step is the forced one.

9 The Competent/Frontier Split Is the Selection Principle

The Strategic Thesis's central economic engine, competent intelligence for the routine majority of work, frontier intelligence routed in for the exceptional remainder, turns out to be a hardware specification. Frontier capability changes quarterly; it stays on GPUs and APIs, reached through II-Agent, the Strategic Thesis's orchestration layer, which routes each request to the cheapest tier that can answer it, earning distribution margin. Competent capability is the part that is stable, certifiable, and therefore **freezable.** And stability is the property that makes a workload manufacturable. Far from a compromise the silicon imposes, the split is the selection principle that tells you which tokens to manufacture.

This is where enthusiasts overclaim, so the load-bearing number is NVIDIA's own, from its argument that small models are the future of agentic AI: "Serving a 7bn SLM is 10–30× cheaper (in latency, energy consumption, and FLOPs) than a 70–175bn LLM." The popular stronger claim, that 80–95% of deployed tokens need only competent capability, is a practitioner's rule of thumb, a **dial** rather than a fact. The routing threshold γ is measured per deployment against its own workload, and the same γ turns out to be the machine's one hardware knob (§14).

The architecture of the split is a three-layer decomposition that recurs at every scale of this document: **etched silicon is procedure; Radiant is memory; frontier APIs are the tail.** The etched card carries the stable procedural operator, reasoning patterns, language competence, protocol-following. Radiant retrieval carries the knowledge, facts, records, curricula, the content that low-bit compression handles worst and that changes too fast to print. The frontier API carries the capability tail. Formally, the Champion stack is a system-level mixture-of-experts whose largest expert costs disk instead of HBM: capacity scales with the total, energy scales with the active. Sovereign workloads, being retrieval-heavy by nature, the citizen's question is about *their* records, *their* law, *their* curriculum, sit exactly where this decomposition is strongest. On the etched fraction, the manufacturing cost target is ~$0.004 per million tokens: where the Champion's competent tier already enjoys a 5–20× cost advantage over frontier serving, the split's silicon layer extends it toward 50–100×.

The distribution boundary is the economic boundary

The split has a sharper statement than *competent versus frontier*, because capability is not the axis that sets cost. Distribution is. Work deep in the model's training distribution is answered by the etched operator at the physics floor. Work whose **facts** have moved but whose **procedure** has not is still in-distribution for the operator, and retrieval returns it there, which is how Radiant converts factual novelty into a cheap etched answer rather than an expensive routed one. Only genuine out-of-distribution work, a capability the prior lacks, leaves for the frontier. Out-of-distribution is the routing condition, not the error condition. And what routes *repeatedly* is not waste; it is the specification for the next generation's training set.

This is the exact shape of the Champion's advantage, and it is informational rather than technical. A Champion does not guess what to etch; it measures what repeats. Sitting inside a jurisdiction's workload, it observes the local distribution, the questions a health system actually asks, the forms a ministry actually files, and manufactures against the mass of it. A general-purpose vendor, blind to that distribution, cannot know which functions have frozen there. The moat is not the mask. It is knowing which mask to cut.

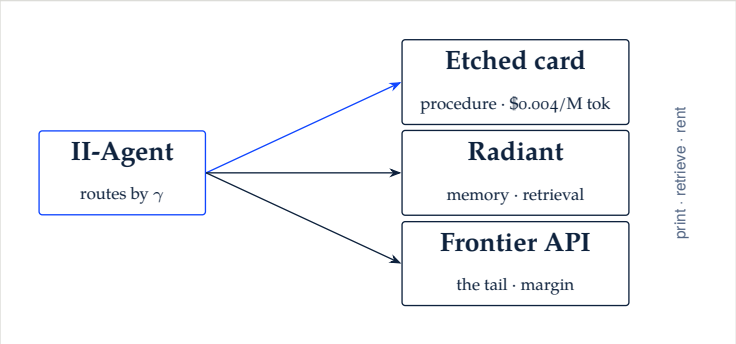

Figure 6 The split as a router. II-Agent routes by γ. Procedure is etched (~$0.004/M tokens, modelled); memory is retrieved (what low-bit compresses worst); the tail is rented (margin, not manufacture).

Anything stable enough to certify is stable enough to etch, and nothing else should be.

10 The Recipe

What gets printed is no transformer downcast to hardware but a model shaped, from the first training token, for the machine it will run on (§14), and the shape has a name: autoregressive inference is a loop; seeded diffusion is a circuit. An autoregressive decoder is a data-dependent loop whose state grows with every token. A seeded block-diffusion decoder is K applications of one fixed operator to a bounded buffer, with the randomness supplied by a keyed counter, deterministic end to end. This program was therefore never "an LLM, converted." The block-diffusion language model *is* the diffusion-shaped design, and the first chip of Section 13 is its smallest instance.

One distinction before the recipe, because it sets the load-bearing status of everything in it. Two of its choices are bets with floors, the precision ladder and the seeded road. Two are not choices at all: bounded state is forced by the machine, because a cache that grows with the conversation is the one thing a fixed circuit cannot hold (§14), and block diffusion is, today, the only published route to a quality language model with bounded state. The rival's shipped chip proves hardwired weights; it does not prove the printed model, because its silicon still carries the sequence's growing state in SRAM. The category this document claims begins where that state ends.

Five steps, each with published precedent

Pretrain in NVFP4. Four-bit native pretraining is production fact. The recipe is published and validated at scale, with independent reproductions (E.4), and its own precision partition, matrix layers in four bits, attention held high because softmax "exponentially amplifies quantisation noise", is the same partition this document's Part IV derives from first principles.

Convert to block diffusion. Demonstrated at 100B scale by continual pretraining. The plan of record is single-tower joint adaptation with an autoregressive anchor (E.1, E.2); the fallback, freezing the base and training a coupled tower, is reported at 30B with 98.7% quality retention on the developer's own benchmark. Commercial diffusion LMs already serve, by their makers' accounts, at 1,000+ tokens per second at near-parity quality.

Distill the sampler. Few-step consistency distillation collapses the denoising schedule to K≈4–8 fixed steps. The surviving in-loop noise is generated by a keyed counter-mode PRNG at an amplitude the mathematics pins to the quantisation lattice (§14, E.3).

Co-train the ternary shell. The step that used to be the risk, restructured below.

Co-train corrupt-and-predict. Teaching the model to repair corrupted context makes committed tokens revisable at inference, Section 12's editing, and the technique's value grows with quantisation coarseness; no chip needs it more than this one.

The nesting: the risk becomes a curve

Ternary weights $\{-1, 0, +1\}$ are not a different species from NVFP4; they are NVFP4 with the magnitude field collapsed to a single bit, under the *same block scales*. That identity gives a seven-candidate closed-form projection per weight block (E.5), which makes matryoshka co-training cheap: the ternary shell is trained *inside* the NVFP4 parent throughout pretraining, and its quality is a curve watched during the run rather than a conversion gambled after it. The parent had to be NVFP4 specifically, the projection gap to ternary is 2.25×, against FP8's 5× and FP16's 10×, the unique format close enough to nest. The competing four-bit format prices the choice: MXFP4 requires ~36% more training tokens to match NVFP4's loss, a footnote that matters because the nearest competitor's next chip is MXFP4-based (§19). Native low-bit training at scale is meanwhile established on both ends, ternary at 2B parameters over 4T tokens and beyond, four-bit at 12B over 10T within ~1% of FP8. The nesting itself, ternary co-trained inside NVFP4, is the one step with no published precedent. It is stated plainly as the program's principal novel claim, and Stage-0's principal experiment. The curve decides.

Two roads to the checkpoint

From scratch, the recipe costs ~25T tokens. The seeded road starts from the Intelligent Internet's own open mixture-of-experts checkpoints and reaches a dense student in about 4–5T: fold the router into the expert keys (the router is linear, so it composes), merge the duplicate units that balanced routing provably creates, compress against the measured sovereign workload, and distill layer-aligned. The arithmetic of what survives is a power mean between total and active parameters (E.9): a 30B-total, 3B-active teacher lands a ~14B dense student at parity-with-margin on procedural and retrieval-backed work, and irreducibly below the teacher on closed-book recall, the gap Radiant exists to cover. And because the teachers are the Intelligent Internet's own weights, the one thing an immutable artifact could never renegotiate, the licence, never arises.

What is not yet true

Zero-shot diffusion models still trail autoregressive ones; parity depends on the conversion-and-alignment stage. Quality-preserving parallel commitment currently averages, on the vendor's own benchmark, ~2.4× effective speedup at 30B, the aggressive step counts this document budgets are targets, not results. And no fold-initialised sparse-to-dense distillation has been published at scale. Every one of these is a Stage-0 measurement, made in software, before any mask is cut.

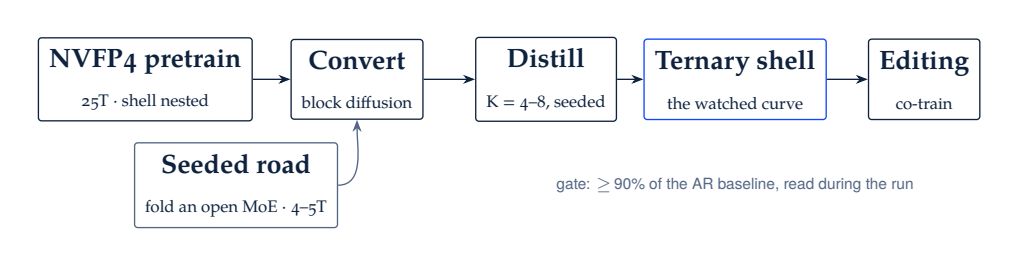

Figure 7 The pipeline. NVFP4 pretrain (25T, shell nested) → block-diffusion conversion (single tower, AR anchor) → sampler distillation (K = 4–8, seeded) → ternary shell (the watched curve) → editing co-train. The seeded road, fold an open MoE, merge, compress, distill (~4–5T tokens), joins at step two. Gate: ≥90% of the autoregressive baseline, read during the run. The dress-rehearsal checkpoint at small scale is also the first product (§13).

The recipe's discipline is its ordering. Each step has a published reference; the one novel step is watched as a curve during a run already being paid for, and every failure mode surfaces in software, where discovering it costs a checkpoint, not a mask set.

11 The Capacity Argument and the Option

The standing scientific objection to low-bit deployment is the overtraining penalty: heavily trained models quantise badly, and the models worth etching are heavily trained. The objection is real, and it is a property of **projection**, not of precision. Post-training quantisation takes a model converged off-lattice and projects it onto the lattice; the loss penalty is the curvature term $\frac{1}{2}\,\delta^{\top} H\,\delta$, and overtraining sharpens H (E.6). Native low-bit training converges *on* the lattice; $\delta \to 0$, and the penalty term dissolves with it, a result the measurements support even as the projection law's exact scope stays modelled. What remains is a capacity question, and the arithmetic closes: language models utilise roughly 2 bits of knowledge per parameter, and a deployed ternary container holds ≈2.0 effective bits. The container matches the contents. The knowledge that doesn't fit is the knowledge Section 9 already routed to retrieval.

Capability surplus becomes quantisation margin

The other objection, that models keep improving, so why freeze one, is usually offered as a reason against etching. It is in fact the reason etching gets easier. A model that clears the task threshold by a hair cannot be frozen safely: every bit of precision surrendered to ternary, every degree of freedom given up to deterministic decoding, every constraint imposed by certification spends margin it does not have. A model far above the threshold can afford all three and still clear the bar. Frontier progress does not race the etched model toward obsolescence; it lifts the whole field above the competence line, and past that line the rational act is to manufacture. The better the models get, the more of the public workload becomes safe to print.

The program structure that follows is an option, not a bet. Its floor is the NVFP4-precision model on the same machine: the converted, distilled operator carried at four bits by shell plus residual plane, two dies, one if distilled smaller. The floor is a precision floor, never an architecture floor; nothing in this program etches an autoregressive checkpoint, because the resulting object would be an accelerator beside a memory, not a printed model (§10, §14). The ternary shell, co-trained for ~4% additional tokens, is a call option on that finished asset. If the watched curve clears the gate, the payoff is 2.6× weight density and roughly 2× energy, in wafers, roughly 85–90 yielded models per 300mm start against 40–45 at the floor (E.11); if it doesn't, the floor ships. That decision stays visible during the run and is exercised at the packaging line (§15), the latest, cheapest, most reversible point a silicon decision can occupy.

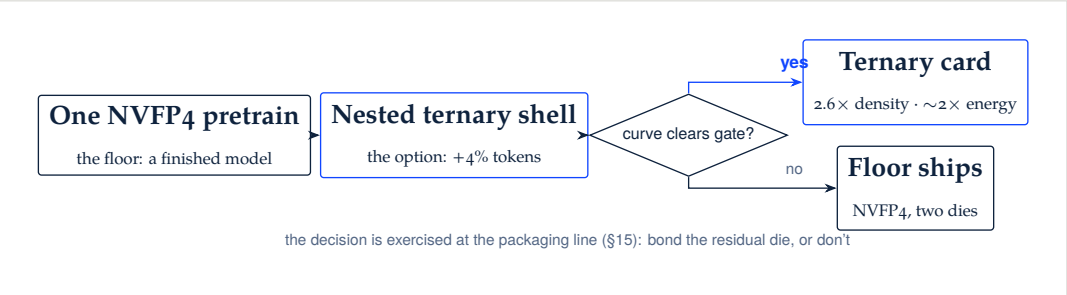

Figure 8 The option, drawn. One pretraining run buys a finished NVFP4 model (the floor) and a co-trained ternary shell (the call option, ~4% more tokens). The watched curve decides at the packaging line: clear the gate and the payoff is 2.6× weight density and roughly 2× energy; miss it and the floor ships. No mask is hostage to the bet.

Why ternary and not binary

Binary weights minimise the bit. Ternary weights minimise the **useful circuit**, and the difference is the zero. A ternary zero is not another symbol to store; it is an omitted operation, an absent wire in the hardwired limit, a silent adder input, a smaller reduction tree, a bank that never switches. Binary forces every weight to participate in every result; ternary lets the learned function *say nothing*, and trained ternary models say nothing about half the time. The empirical signature is exactly where it should be: measured ternary-ROM density *rises* with the zero ratio (E.13). Binary remains a research adjacency (§13); it is not the product path.

The program does not bet the model on ternary. It buys a proven model, then spends 4% more to try to halve the silicon, and ternary, not binary, is where the halving is real.

12 The Update Path and the Two Modes

A frozen operator does not mean a frozen product. The update path has three layers, each matched to a rate of change.

Adapters are megabyte-scale task modules, cryptographically signed, loaded into the SRAM island, and portable across silicon generations, so the task library outlives any chip. **The delta plane** is a bounded, norm-constrained divergence from the backbone, held beside the residual weights: one backbone ROM at ~1.1× area rather than two towers at 2×, with the constraint enforced during adaptation so the divergence stays printable (E.2). **Token editing** is the corrupt-and-predict co-training of Section 10 at work: the sampler revisits committed tokens above a confidence threshold,

restoring the reversibility that few-step distillation spends (E.7), and in silicon it is firmware on the block buffer, at zero datapath cost.

Above the update path sit two product modes, and a firewall between them. In **certified oracle mode**, the chip is the artifact that conformity assessment attaches to, the competent tier's system of record, serving answers on its own authority within its certified envelope. This is what the sovereign buyer is buying. In **draft mode**, the ternary path runs as a cheap proposal distribution in front of a high-precision verifier, the hardware analogue of speculative decoding, and is priced by its acceptance rate: above ~95% category-breaking, above ~80% strong, above ~50% still useful. Draft mode widens the addressable workload to tasks the certified envelope excludes. The firewall is that it must never become the identity. A speculative-decoding coprocessor is a feature an incumbent can capture in a driver update, and it surrenders the certification story that Part II built; the certified tier's correctness burden is met by co-design, editing, and the residual step, never lowered to "proposes acceptably." Draft mode is a mode, never the company.

One objection deserves answering here, because it recurs wherever frozen weights meet a changing world: surely open, shifting environments demand mutable models. They do not; the objection mistakes context for weights. A self-driving stack meets a road it has never seen on every drive, yet its deployed weights are versioned between releases. What changes continuously is the *context*, position, map, route, the sensor field, applied to a stable learned function, while the logs of the genuinely novel accumulate into the next release. Public-sector intelligence is the easier case of the identical pattern. The citizen's records, the jurisdiction's law, the local curriculum are the live context; the etched model is the policy, and the out-of-distribution case that repeats is next year's respin (§15). Fixed weights, live context, fleet-learned generations: the same three-way split of Section 9, now answering the objection that first motivated it.

Part IV

The Silicon

One ladder, from a thumbnail of 16-nanometer ROM to the sovereign stack, each rung a smaller bet than the last one retired.

13 The Proof Sequence and the G-Ladder

The first chip is a printed learned function rather than an AI accelerator: the smallest artifact that proves the thesis, and along the proof path it is deliberately austere. One model. One graph. One precision regime. One shape family. One host interface. No HBM, because nothing here uses HBM; no 3D stacking; no mixture-of-experts; no photonics; no frontier claim. The first tapeout proves the one thing the entire stack depends on: that a useful low-bit model can be compiled into a cheap fixed circuit by a repeatable toolchain.

What the first rungs uniquely prove matters. That weights hardwire into silicon is no longer in question, a hardwired 8-billion-parameter model has been running on TSMC N6 since February, on the company's telling (§19). What has never shipped is the printed model itself, and five claims are ours to take: the **printed model**, bounded state, both memories deleted, the only mutable silicon a block buffer (§10, §14); the first **ternary** one; the first **co-designed** one, trained for the die from token zero and so paying no post-training quantisation tax; the **appliance-class cost floor**, an order of magnitude below an 815mm² leading-edge monster; and the **model-to-GDS compiler for co-designed models.** The rival's foundry proves a checkpoint can become masks in weeks, but a toolchain that compiles a model *trained for its own lattice*, and wraps the output in a certification dossier and an escrow chain, does not yet exist, and it is Intelligent Silicon's crown-jewel IP.

Gen	Artifact	Node / configuration	Cost, modelled	Proves / serves
G0	MPW block test: ternary ROM macro, adder fabric, keyed PRNG, compiler output	N12/16 shuttle	$1–5M	model-to-GDS; the operator in silicon
G1	Minimum printed model: 0.5–3B ternary block-diffusion LM, the Stage-0 checkpoint, etched	16nm-class, 2D, no bonding; 45–200mm²	$10–30M	the printed ternary learned function; appliance economics
G1a/b	Solar voice teacher; protocol health aide: modules and reference designs	G1 die + host MCU/NPU	module-level	universal service made physical (§16)
G2	Champion competent-inference card, 14B dense	N6; 2D standard / 3D premium SKU	$60–90M	the certified bulk-token utility layer
G2S	Sovereign escrow card	G2 + via-mask escrow + dossier	·	custody chain to lithography (§21)
G3	MoE expert-plane silicon, dynamic top-k	3D stack	·	jurisdictional expert libraries
GMax	Fully hardwired long-life artifact	any generation, frozen	·	ten-year certified deployments

Two corrections to the obvious plan make the ladder cheap. First, the node. The proof rungs do not need N6: at 16nm-class, a mask set costs an estimated $3–5M against ~$15M at N6, wafers run ~$4,000, capacity is utterly uncontended, and the process development is not even a detour, because 16nm-class is the plan-of-record node for G2's bottom die. The cheap chip subsidises

the sovereign stack's engineering. A 1B-parameter ternary shell fits in roughly 45–70mm²; 3B in 130–200mm². Second, the volume threshold, stated honestly because it reorganises the business logic. Mask amortisation dominates small-chip economics: four million dollars of masks over ten thousand units is $400 per chip, dead; over a million units it is $4, appliance-grade. **The appliance is economical only at Champion-procurement volume.** National school systems, clinic networks, the universal-service quota, which means the cheap path does not bypass the Champion model; it requires it. The aggregated sovereign buyer is what amortises the masks, and wafer banking (§15) serves the sub-threshold curriculum variants from one base set.

Discipline is a list of refusals. Not first: a wafer-scale chip; a 3D hybrid-bonded stack; a general LLM accelerator; a dynamic-top-k MoE chip; a photonic co-processor; a doctor replacement; II-Agent-in-silicon; a screen appliance; a full sovereign certification package. First: one fixed-shape, two-dimensional, ternary learned-function chip. The minimum chip is the *proof* that learned functions print cheaply; everything in this document above the first rung, certified cards, sovereign escrow, Radiant-backed retrieval, public-benefit appliances, is the path that follows from the proof.

Beyond G2: the trajectory

G3, the mixture-of-experts stack, is where read-only-memory economics force the architecture everyone else chooses for other reasons. In ROM, dark experts are literally dark. Capability scales with *total* parameters, which are mask-programmed area and nearly free, while energy scales with *active* parameters, the only banks that switch. That proportionality is circuit physics, unconditional on any claim about what experts "mean." The router itself becomes the address decoder, expert selection is word-line enablement, and the memory read *is* the inference. Three router designs fork the product: a wired hash (zero weights, a printable truth table, maximal certifiability), a learned linear router (maximal quality, semantically illegible, thermally flat), and imposed domain routing. That last is the certifiable route, because the honest reading of the literature is that *emergent* expert legibility is architecture-dependent, strong in fine-grained designs, absent in coarse ones, so a certification story must be imposed, never assumed.

Imposed routing buys the strategic product: in **sovereign expert planes**, a jurisdiction's personalisation is which expert library gets printed in its via masks, and the payoff is partial respins, incremental conformity, a national curriculum as a mask layer. Expert granularity is then a computed optimum (bank periphery against conditionality gain), and the quality–energy–area triangle explains the ladder's order: G2 is area-bound, where dense wins; the 3D stack relaxes area, energy becomes binding, and the mixture is forced. **GMax** is the endpoint artifact for locked decade-long deployments: zeros become absent wires, nothing on the die is writable, and the chip inherits the memory model of a hardware security module.

The research adjacencies stay adjacencies, and are flagged as speculative: XNOR-binary arithmetic; a photonic attention co-processor; the attention product is activation-times-activation, the one computation in the stack with no stored weights and therefore the one that cannot be etched; optical interconnect bought from its industry, not built; full-optical weight storage off the roadmap on density grounds. One line for the wafer-scale aside: a single 300mm wafer holds ~1.8 terabits of NVFP4 ROM, so the physics permits a frontier-scale printed model. The competent product simply doesn't need it.

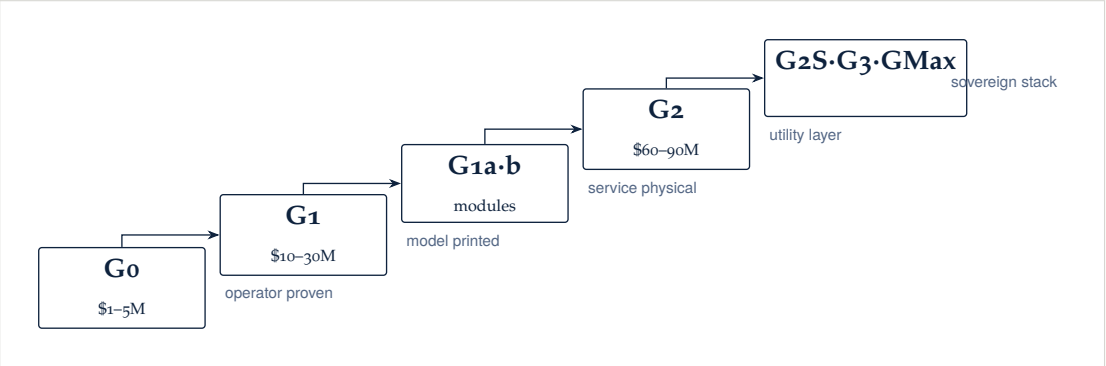

Figure 9 The G-ladder. G0 ($1–5M, operator proven) → G1 ($10–30M, model printed) → G1a/b (modules, service physical) → G2 ($60–90M, utility layer) → G2S/G3/GMax (sovereign stack). Each rung is retired before the next is climbed; the checkpoint that gates the program is the die that starts it.

14 One Stack

The chip is a fixed-depth function machine: noise in, tokens out. Block generation is K applications of one stateless operator to a bounded buffer, no data-dependent control flow, no state that grows, nothing resident between requests.

The plain mechanism first, because the whole architecture follows from it. Autoregression turns time into memory: it appends a token, grows a cache, and must write and re-read that cache every step, so its silicon is a memory machine whose bandwidth scales with the length of the conversation. Block diffusion turns time into recirculation: it reuses one fixed operator over a bounded buffer K times, so its silicon is a function machine with a static latency envelope and a working set that never grows. Memory that grows with use is what kills fixed-function economics; a circuit reused is what makes them. This is why the architecture is forced, not preferred.

The contrast is also a utilisation theorem (E.8). An autoregressive decoder at batch one drains an L-stage pipeline once per token, running the silicon at roughly 1/L of capacity, the batching tyranny rebuilt on-die. A diffusion block fills the same pipeline width-wise and recirculates it K times back-to-back: near-total utilisation at batch one, with cycle-accurate latency. Single-user speed stops being the thing sacrificed for economics and becomes the thing the geometry produces.

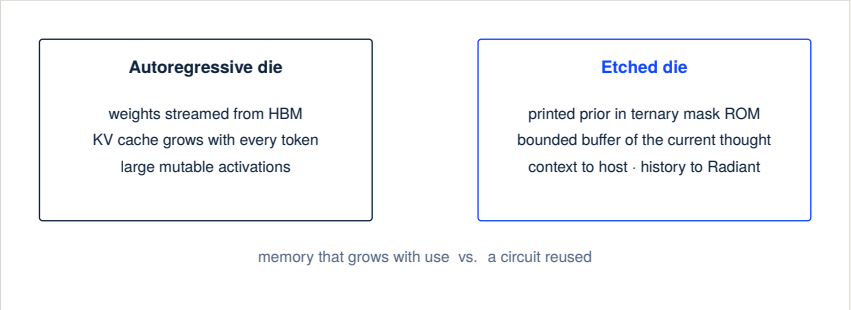

Figure 10 Two memories, one deletion. The autoregressive die carries weights streamed from HBM, a KV cache that grows with every token, and large mutable activations. The etched die carries a printed prior in ternary mask ROM and the bounded buffer of the current thought, pushing context to the host and history to Radiant, where it becomes a cited record rather than an opaque cache.

The temperature split

Where does randomness live in a deterministic machine? The design rule: print the deterministic learned operator; leave decode-time temperature in software; keep lattice-temperature noise in the loop, keyed. The host owns everything that is policy: seeds, schedules, sampling temperature, top-p, guidance weights, safety filters, retrieval, decoding to text or audio, the audit trail. The card owns the frozen operator, and one thing more: the in-loop noise injected between denoising steps. That noise is not optional while the score carries error, and a ternary operator carries permanent score error; the mathematics pins its amplitude to the quantisation lattice, the calibrated warmth that anneals lattice roughness out of the trajectory. It is generated by a keyed counter-mode PRNG, a few thousand gates of the circuit family mining silicon perfected: bit-exact, golden-vector testable, certifiable. The ASIC is the cold path; the host is the heat bath; the cold path carries its own calibrated thermometer.

The card–host boundary, budgeted

The boundary is disciplined by the skeleton itself. In the reference hybrid architecture only 6 of 62 blocks are attention; the state-space layers that dominate the depth carry fixed, small recurrent state and etch well. The host runs what needs mutability and high precision: commitment, ordering, the conversation's memory, attention, softmax, normalisation, embeddings, the output head, at INT8/FP8. The card runs the recirculating denoiser in ternary ROM. Across the PCIe link, only projections cross: activations, never weights.

That link is the machine's one off-die data path, so it gets a budget rather than an assurance. On the reference shapes, a 14B-class trunk, a block of 32–64 tokens, FP8 activations, each crossing carries a few hundred kilobytes; six attention blocks crossed K times in both directions puts a full block generation in the tens of megabytes of link traffic. Against a PCIe Gen5 x16 link's ~64 GB/s, that is well under a millisecond per block, inside the latency envelope the committed-speed targets of Section 17 require, with margin. The figures are modelled from the reference shapes; the full budget, with latency terms, is E.14, and the measurement itself is a Stage-0/G0 item.

For readers who want the physics framing: the same boundary can be read as a phase diagram. The host side is the broken phase, where commitment, ordering, and the metric's pairing live; the card is the symmetric phase, the recirculating denoiser; the PCIe link is the domain wall that only projections cross. Precision is spent in forcing order, metric first, measure second, operator last, which *derives* the ~5% of layers held at FP8 rather than sweeping for them; the published four-bit pretraining recipe documents the same partition from noise analysis.

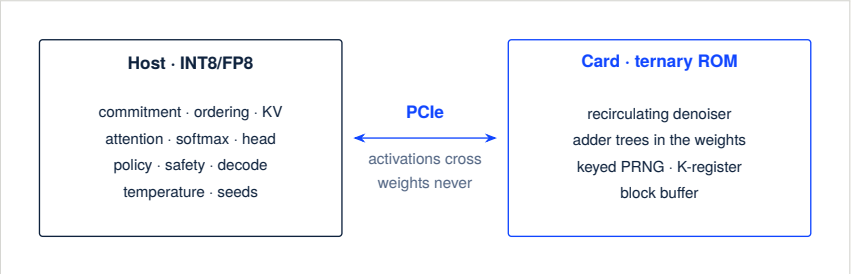

Figure 11 The floorplan is the boundary. Host: commitment, order, KV, attention, softmax, head, policy, safety, decode, temperature, seeds, INT8/FP8. Card: recirculating denoiser, ternary ROM with adder trees, keyed PRNG at lattice temperature, K-register, block buffer, weights never leave the die. PCIe between them: activations cross, weights never.

The allocation law: one knob

Token difficulty is heavy-tailed, so spending equal compute on every token is provably wasteful. The optimal policy is water-filling: allocate marginal compute where marginal loss-reduction is highest, until every active margin equals a single shadow price. The machine already has the estimator, its own confidence, and, it turns out, already has the price. **The commitment threshold γ is the Lagrange multiplier** (E.10). The speed–quality dial that every deployment tunes empirically *is* the price of a joule, and every other threshold in the machine, routing scores, speculative acceptance, is denominated in the same currency.

Four axes of conditional compute follow. *Iterations* adapt through confidence-selected commitment at fixed step count, giving static latency with adaptive content. *Precision* follows the matching law $\sigma_{\text{quant}} \leq \sigma_{\text{sample}}$: the ternary shell serves the noisy early steps; the residual plane joins for the final step or two at a duty cycle of order 1/K, derived rather than tuned, with shell-drafts/residual-verifies self-speculation as Section 12's draft mode on-package. *Width*, at G3, comes through dynamic top-k over parallel physical adder banks, latency set by depth and energy by the sum of k. And *depth* is rejected, because per-token layer-skipping breaks the recirculating lockstep; the machine takes fewer whole passes instead.

The stack, the tiers, and the register

The G2 premium SKU is a two-die stack. The N6 top die carries logic and the ~24 Gbit ternary shell, adder trees interleaved with the weights they read; the 16nm-class bottom die carries the ~40 Gbit residual plane, the delta plane, and the adapter SRAM. Hybrid bonding at 6µm pitch is in high-volume manufacturing; the bond adds ~$120 per unit. The stack collects a thermal gift: the bonded layer is ROM, no refresh, no switching at rest, so the model sits beneath the logic as a thermally inert substrate, taking 3D's density while dodging 3D's defining problem. The honest risk goes next to it: leading-edge hybrid bonding is foundry-captive, so the 3D SKU is TSMC-tied. The mitigation is structural, because the 2D standard SKU and the entire proof path are bonding-free and second-sourceable.

Power is then a product decision on one architecture: Sovereign 250W, all on-card and escrowed; Lean 40–80W on bus power; Workstation 8–15W on an M.2 stick, passive; Edge 3–5W; Appliance sub-watt on a solar duty cycle (§16). Idle is ~zero everywhere, the model costs nothing between questions. And the recirculation count K is a register, programmable 1–8 and certified per artifact, so every step the training side removes lands on the installed base as a field upgrade: free in silicon, with its re-assessment treatment inheriting the open question of Section 7.

The optimal inference engine has one knob, and it is the price of a joule.

15 The Respin Economy

The expensive act is designing and masking the learned function. The cheap act is copying it. "Engraving the model" is paid as masks, process steps, and wafer starts, never per answer, and never meaningfully per die; once the masks exist, each additional copy costs wafer area, yield, package, and test. This is the economic discontinuity the whole document turns on. On a GPU, the model is a recurring rent paid to the memory hierarchy on every token. Etched, the model is one-time lithographic capital, amortised over every token the artifact will ever produce.

Change	Cost mode
Same design, more units	wafer + package + test; copying is free of the model
New model on the same base wafers	via-mask respin against banked wafers, weeks
New architecture	full design and mask cycle, a generation
Adapter or curriculum update	signed payload or storage pack, no mask at all
Retrieval update	software: Radiant's problem, by design

The grades, plainly. Weights compile to GDSII in about a week, by the toolchain vendor's account. A weight respin runs ~$1M and ~2 months, an inference, but a defensible one, since ~99% of a mask set is unchanged in a respin and the physical mask writing takes days. A full N6-class set is ~$15M; a 16nm-class set, $3–5M. The deeper flow is via-ROM with wafer banking: base wafers, identical and model-agnostic, are fabricated ahead and parked before the final via layers, and the model itself lives in one or two via masks applied at the end. One correction belongs in the record: the structured-ASIC vendor that industrialised this playbook is no longer available for it at this class of node, so the banking flow is implemented directly with the foundry, the stronger position anyway, since it is masks and logistics, not a licensed product. Masks are cheap; being wrong in masks is not.

From there the consequences compound. Personalisation per model, customer, or jurisdiction becomes a weeks-scale packaging-line event. The sovereign escrow artifact of Section 21 shrinks to the via masks. And the ternary option of Section 11 completes its migration into physics: the same banked wafer ships as a cheap ternary card or a dual-precision sovereign part depending on whether a residual die is bonded, the decision moved to the latest, cheapest, most reversible point it can occupy. The supply logistics cooperate: mature-node capacity is genuinely uncontended, with analysts describing idling N6/N7 lines being repurposed as the leading edge moves on, and wafers at ~$9,500 (N6) and ~$4,000 (16nm-class). The design's scarce inputs are the ones nobody is fighting over.

Engraving the function is capital. Copying it is free. That is the whole inversion.

16 The Last-Mile Intelligence Appliance

The Strategic Thesis obligates every Champion to honour a daily inference quota and gives every citizen an II-Account, universal-service commitments that today have no physical form beyond datacenters and phones. This section supplies the missing artifact: voice-first, screen-optional, grid-optional devices whose intelligence is printed, whose knowledge is a signed pack, and whose custody is local.

The product family, in the order a Champion would deploy it. A **teacher in a box**: a patient, tireless voice tutor in the local language. A **protocol health aide.** And the name is chosen with regulatory care: not a doctor replacing doctors, but a protocol-following clinical aide that never leaves, never forgets, speaks the local language, and knows when to escalate; its plausible high-risk classification is inherited and budgeted, per Section 7. A **civic services box** for forms, entitlements, and administration. A **disaster and refugee field node** that works when infrastructure doesn't. And

the **rural Champion edge node** serving a school or clinic. The operating model keeps the fabless discipline: Intelligent Silicon ships modules and reference designs; Champions and ODM partners build, deploy, and support the boxes, with the Strategic Thesis's forward-deployed engineers as the field organisation.

The bill of materials is the argument, on modelled targets. Ternary ASIC $5–30; microcontroller and audio $1–5; mic and speaker $1–5; battery or supercap $3–8; solar panel $3–10; enclosure $3–10; optional storage and connectivity $1–8; assembly and test $3–10. **$20–75 all in.** The lifetime arithmetic follows: a $50 device serving 100 interactions a day for three years delivers ~110,000 answers at **$0.00046 of capital per interaction.** No API bill, no network dependency, and near-zero idle draw, which is the property that makes solar viable at all. One design fork is left open deliberately: speech via a host NPU around a text-token die, or speech-token-native block diffusion with voice handled end-to-end on the ASIC, the cleaner appliance, decided by the Stage-0 voice suite.

The knowledge architecture answers the stale-weights objection by construction, because it is Section 9's split miniaturised. The fixed model is the *procedural* tutor. Curriculum, health protocols, and local information live in signed, updatable storage packs, the appliance's Radiant. Session memory stays local and encrypted; governed reference syncs when connectivity exists. Procedure is etched; memory is packed; the tail escalates. The selection principle then becomes measurable through load residency, the fraction of work the local artifact answers without escalation, with effective cost $\approx$ ASIC + host + $(1-p)\cdot$verifier (E.12). Residency targets, modelled and still to be measured: 85–98% for classification and routing; 70–95% for the voice tutor; 70–90% for translation and summarisation; 50–85% for the protocol-bound health aide; 40–70% for legal or medical drafting, which runs in draft mode by design. Frontier reasoning, at 0–30%, is not a target; it is what the tail is for.

Device economics span the family, dongle $10–40, voice teacher $20–75, health aide $30–100, screen terminal $75–200, edge node $100–500, all modelled, at prices that exist only at Champion-procurement volume, per Section 13's threshold. The appliance and the utility layer are one business, and this is the page where the Strategic Thesis's most abstract commitment, intelligence as universal service, acquires a housing, a speaker, and a panel of silicon that drinks sunlight.

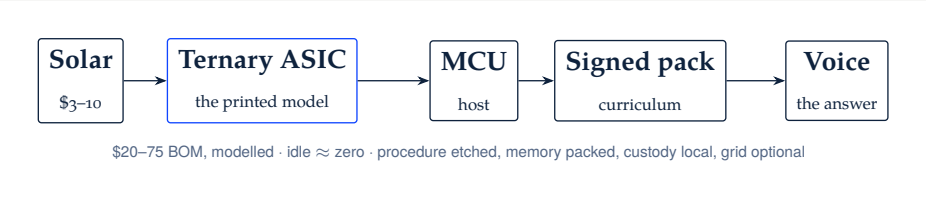

Figure 12 The last-mile appliance. Solar panel ($3–10) → ternary ASIC → MCU → signed pack → voice → the answer. $20–75 BOM, modelled; idle ≈ zero. Procedure etched, memory packed, custody local, grid optional.

A cloud model answers only where the internet reaches. A solar etched model answers wherever sunlight reaches.

Part V

The Economics

Mining cost of goods, software pricing, underwritten against a market whose spot price halves every two months.

17 Unit Economics: from Die to Card to Appliance

The program's capital requirements come in tiers, every figure modelled: G0, the block-test shuttle, at $1–5M; G1, the minimum printed model, at $10–30M (conservatively $20–50M); a G1 production generation with packaging, test, and a respin budget at $30–70M, and the G2 Champion generation at **$60–90M**. That last figure sits well below published greenfield ASIC benchmarks, industry models put a 7nm-class SoC at $217–300M, with realistic practitioner estimates near $160M, and the gap is defensible only for the reasons this design exists: fixed function, mature node, heavy IP reuse, and no software stack to amortise. One to two respins are budgeted in every tier, and the tiers absorb their own training: on current rental arithmetic, compute for even the full 25T-token pretrain prices in the low single-digit millions, inside the G1 envelope, and the seeded road is a fraction of that.

Below the program tiers, the die economics are almost embarrassingly simple, because after masks a die costs wafer price divided by good dies. The discipline is to run that arithmetic *forward*, from the die-area targets the products require, to the model sizes they permit. 25mm² is the voice-tutor primitive at roughly $3–6 of raw silicon; 50mm², a useful tutor in single-digit dollars; 100mm², a strong appliance model in the tens; 200mm², a premium card; anything past 400mm² is not a first product, and near-reticle dies are strategic instruments for G2 alone. The first model is selected by die area, not benchmark vanity: the model must fit the die, not the die the model.

The G2 card, on this arithmetic: die cost an estimated $131–168, plus ~$120 of bonding for the premium two-die SKU. Energy lands at **4–8 millijoules per token**, modelled to within a factor of two to three, a best case of 1.7, a compound-worst of 26, against the tens-to-hundreds of millijoules of general-purpose serving. The range deserves its honest reading: at the compound-worst against an aggressive incumbent deployment, the headline gap compresses from orders of magnitude toward small multiples, which is why measured GPU joules-per-token is gate (e) of Section 23 rather than an assumption. Committed speed is **10–30 thousand tokens per second per user**, on the program's model; the external anchor is a shipped hardwired chip demonstrated at 16,960 (§19), and the upside path, four-step sampling, editing, dynamic width, points, on the same model, toward 80K. Divide through and the manufacturing target on served tokens is ~$0.004 per million; the appliance's half-a-millicent per interaction was Section 16's version of the same fact. Physics sets the floor; the market sets the clock.

One KPI unifies die, card, and box: **lifetime useful outputs per wafer-dollar-watt**. And one supply-chain sentence completes the picture: logic wafers plus commodity flip-chip, with bonding only on the premium SKU, no CoWoS queue, no HBM allocation fight. The scarce inputs of the incumbent stack are avoided by construction.

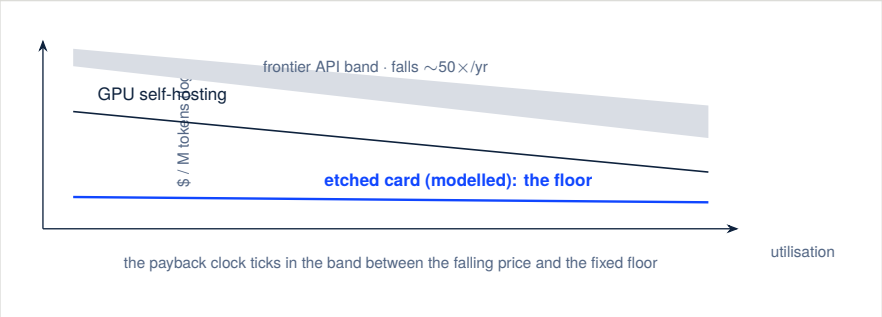

Figure 13 Cost per million tokens against utilisation, log scale. The frontier API band falls ~50× per year; GPU self-hosting sits below it; the etched card (modelled) is the floor. The gap is structural, the joules of Section 2, not a discount, and the payback clock of Section 1 ticks in the shaded band between the falling price and the fixed floor.

18 The Breakeven, the Fundability, and the Payback Clock

The breakeven arithmetic is short. A G2 card selling manufactured tokens at $0.15 per million, an order of magnitude under the frontier band and well over an order above its own manufacturing target, clears roughly $9,500 of gross margin per card-year at sustained utility-class utilisation, the deployment mode the certified tier is contracted for; between 2,400 and 4,700 deployed cards retire the entire generation's NRE in two years on this arithmetic. One mid-sized national deployment is that number, G1 retires on a single national education or health procurement at appliance volume per Section 13's threshold, and for scale, the entire G2 generation costs roughly one week of a frontier lab's compute bill.

Two clauses keep the arithmetic honest. The split of that margin between artifact price, per-device royalty, and operator return is a term sheet rather than a thesis claim; the arithmetic prices the system, and Section 20 names the layers that share it. And the volume assumption is stated as its own implication: roughly 65 billion served tokens per card-year, about two thousand tokens a second averaged around the clock, an order of magnitude beneath the committed per-user peak of Section 17. The breakeven underwrites a tenth of what the silicon can do; utilisation above the assumption shortens the clock, and the headroom is the conservatism.

The buyer base exists, and what it lacks is this asset class: the Strategic Thesis's patient pools, pensions, insurers, sovereign funds, cannot buy frontier-lab equity at reasonable terms, but infrastructure priced on contracts, plant, and certified assets is what they are built to hold.

Now the objection Section 1 posed, at full strength. Spot inference prices fall ~50× per year. A frozen artifact underwritten against spot prices is underwater within a year, so how is any of the above real? Three structural answers, in ascending order of importance.

First, the contract. The certified tier does not sell spot tokens; it sells custody and conformity at fixed quality on utility terms, and the buyer of a certified artifact is buying the one property the spot market cannot offer: that the artifact does not change under it.

Second, the respin. The installed base is not stranded by the improvement curve; it rides it, weight respins at ~$1M and weeks (§15), and the K-register collecting every training-side improvement as a free field upgrade (§14).

Third, the floor. What is falling is the *price* of inference; its destination is the *cost* of inference, and the cost floor is the physics of Section 2. The spot market is racing toward manufactured-good economics. The manufacturer is already standing there. When the price war ends, it ends at the physics floor, where the product already sits.

One distinction keeps the demand claim honest, and it matters enough to state twice in this document (here and in §21). What sovereign buyers have demonstrably paid for, at scale, is *custody of compute*: capacity inside the jurisdiction, on contract terms, with sovereignty language attached. What this program sells is one step further: *immutability of the model itself*, as the certified artifact. The first is the revenue base of the category's largest public company; the second is not yet a demonstrated market, which is why a contracted sovereign purchase of a version-locked artifact is gate (g) of Section 23, the commercial gate beside the six technical ones.

The surplus buys judgment, not only savings

There is a second-order consequence the price war misses entirely. A substrate fifty times cheaper can be spent two ways. Taken as savings, it is fifty times lower cost, the axis the incumbents compete on. Spent as *thought*, it is ten independent agents on one task at five times lower cost: a drafter, a critic, a citation checker, a fairness reviewer, an appeal-rights advocate, an escalation monitor, the deliberation a public institution owes a citizen and cannot today afford at scale. The unit of public-sector intelligence is not the token; it is the deliberated task, and cheap determinism is what turns deliberation from an exception into the default. This is where the silicon meets the rest of the stack. II-Agent orchestrates the panel; Radiant records its evidence and its disagreement; the etched substrate makes the whole committee cheap enough to convene for every decision. And because the substrate is deterministic, fixed seeds, versioned weights, logged retrieval, the committee is reproducible: a citizen's decision is not one sampled response but a recorded procedure, which is what an auditor, an appeal, or a court requires.

The market has meanwhile validated the category with money, in three rows tabled once here. The largest chipmaker on earth agreed to acquire an inference-hardware company's assets for ~$20B. The wafer-scale incumbent raised $5.55B in the largest US listing of 2026, on $510M of revenue growing 76%, with 86% of it from two sovereign-linked buyers: sovereign demand as the revenue base of the category's largest public company. And a transformer-ASIC startup exited stealth with $800M raised and $1B in contracts against unshipped silicon, on its own disclosures. The market already prices specialisation. It has not yet priced custody, and it has not yet been offered immutability at all.

19 Against the Field

The field is best organised by a single column: what stays mutable.

Player	What they sell	What stays mutable	Distinction
GPU / TPU	general AI compute	everything	the incumbent; pays Section 2's bill forever
Etched (Sohu)	transformer ASIC on N4P with 144 GB of HBM3E, per its published spec	the model, in HBM	$800M raised, $1B contracts, unshipped; kept the hierarchy
Groq → NVIDIA	deterministic inference computers	programmable execution	still a computer; assets acquired ~$20B
Cerebras	wafer-scale systems	the model, in SRAM	leakage holds the constants; sovereign-heavy revenue
Taalas	hardwired model chips	the sequence state, in SRAM	proves hardwired weights; the chief rival
Apple / Qualcomm NPUs	general edge inference	model loaded from memory	the ideal *host*, not the printed model
Intelligent Silicon	printed learned functions	only host entropy and policy	the lowest fixed-function floor, plus custody

Taalas deserves its specification in full, because it is both validation and chief rival. By the company's own figures it runs on TSMC N6 at 815mm² and 53 billion transistors, Llama-3.1-8B fixed in mask ROM, 16,960 tokens per second per user, ~250W air-cooled, no HBM anywhere, $0.0075 per million tokens against the company's own B200 comparison of $0.0379, $219M raised, with a successor in development at 20B parameters in MXFP4, its SRAM moved to a separate die. The differentiation is then precise, and the two-die resemblance is superficial: their second die holds mutable state, this design's holds more printed model. Taalas proves hardwired inference; the minimum ternary chip proves printed models. They hardwire someone else's checkpoint and pay the post-training quantisation tax, and public reporting of the first silicon notes visible quality loss from its aggressive 3-to-6-bit quantisation, independent evidence for the tax this program's co-design exists to avoid. This program trains for the die from token zero, on the four-bit parent whose competitor format needs ~36% more tokens (§10), and the rival's next chip is built on that competitor format. And the state: their silicon still carries the conversation's growing memory in SRAM, the successor moves that SRAM to its own die, which is the memory relocating, not leaving, while the printed model's only mutable silicon is a bounded block buffer. They sell speed to the market that wants tokens; this sells custody to the market that wants artifacts. Their stated adoption obstacle, software flexibility, is the one thing the certified fixed-function tier doesn't need.

Two residual objections close the Part. The graveyard: Graphcore, Wave, and Nervana died rebuilding CUDA, a software-ecosystem war that fixed-function silicon never enters, because a printed model has no kernels to port. The graveyard's surviving lesson, that architectures move, is answered by the respin economy, the K-register, and the selection principle that etches only

what is stable. And the incumbent's counterattack: GPU four-bit formats improve the small term in Section 2's waterfall, arithmetic, and cannot touch the large one, because the weights still cross the bus.

The last honest sentence cuts against romance about the recipe itself. The pipeline of Section 10 is published, by the largest company on earth, in the open; training risk and training moat collapsed together. The recipe was never the moat. The artifact, frozen, certified, escrowed, is, and the rest of this Part has priced what it costs to make and what it earns.

Part VI

The Foundry and the Network

Intelligent Silicon fabricates nothing and operates nothing. It does not even sell chips. It sells printed models.

20 Printed Models: Where It Sits in the Stack

Intelligent Silicon does not sell accelerators. It sells printed learned functions: certified, escrowed, refreshed by respin, royaltied per device. Across domains the deeper noun is **printed priors**, a family that runs from a language prior and a procedure prior to, in time, a world prior and an affordance prior: stable functions reused across countless live states. *Printed learned functions* stays the public phrase; *printed prior* is the technical generalisation, and it is what lets the same company extend from the citizen's answer to the robot's reflex without changing its business.

The stack position makes that sentence precise. Models and adapters originate with the Intelligent Internet (II), the architecture-licence analog. Intelligent Silicon owns the layer from weights to masks to certified printed models: the compiler, the mask sets, the banked-wafer rights, the golden cards, the dossiers. Fabrication is TSMC's, with a front-end second source available and one honest captivity noted at the bonding step (§14). Packaging is commodity flip-chip, with the premium stack reserved for the sovereign SKU. Fleets and field are run by others; miner-class operators, sovereign datacenters, telecoms, ODM appliance partners. Champions deploy, with the forward-deployed engineering corps as the field organisation. Nothing in the company touches a fab tool or an end customer's rack.

The revenue stack follows the layer: a generation licence; a **per-device royalty.** And the royalty template is proven at scale, with Arm's newest compute-subsystem rates running above 10% of a chip's selling price at near-100% gross margin, on royalty revenue of $620M+ a quarter and rising; mask escrow fees; certification dossiers; respin fees; appliance modules; and, gated, the toolchain licence. The gate is deliberate. The model-to-GDS compiler is the crown-jewel IP, licensed late and only to sovereign-grade partners, because licensing it early manufactures competitors.

One novel legal layer is flagged as theory rather than asserted as fact: mask-work protection. The US Semiconductor Chip Protection Act grants a ten-year sui generis right over mask works, with European topography parallels, and weights etched as the physical via pattern of a ROM array are arguably a protectable mask work *independent of the contested copyright status of weights as data*, and the theory is testable at first silicon: mask-work registration is an administrative filing, and the Go die is the test article. If the theory holds, the printed form of a model enjoys an IP protection that no hosted or downloaded form can, a protection layer that exists only for the artifact. The same physics cuts the other way, and it is priced in: a printed model is a readable one, since a delayered ROM surrenders its pattern, so the moat was never parameter secrecy. It is the co-design pipeline, the dossier, and the licence, which is why the compiler, not the checkpoint, is the crown jewel (§13). The historical frame belongs beside it: "real men have fabs" aged into its own author's company going fabless, and the industry's admission arrived as early as 1999: "the delta is in our intellectual property." Value migrated from fabs to instruction sets to IP cores. The printed model is the next layer in that same migration.

21 The Escrow Clause and Edge Sovereignty

The escrow clause is the Strategic Thesis's sovereignty commitment carried to lithography. Each jurisdiction's deployment escrows its via masks and model hash against globally standard base wafers. Because the model lives in one or two via layers (§15), the sovereign artifact is the *thinnest* layer in the stack and the cheapest to re-fabricate: a nation's certified intelligence, reproducible at any qualified foundry from a vault it controls, independent of any company's survival, including

this one's. Under the G3 architecture the escrowed object becomes the jurisdiction's expert library, its law, its language, its curriculum, printed as mask layers and re-certified incrementally as they change.

Figure 14 The custody chain. From protocol to lithography to schoolroom: the Champion's licence, the via masks and model hash in a sovereign vault, re-fabrication at any qualified foundry, the certified card and appliance in the field. The thinnest layer in the stack is the one the jurisdiction holds, and it is reproducible independent of any company's survival, including Intelligent Silicon's.

Edge sovereignty is the same property at the other end of the wire. The sovereign artifact is not only a card in a national datacenter; it is a voice box in a school, a clinic, a village, that keeps working when the network fails and the grid follows it. The fixed-weight property that makes certification possible is the identical property that makes last-mile independence possible: nothing to update is also nothing to cut off.

The demand evidence is tabled here, once, with its scope stated precisely. The flagship sovereign-AI buyer's language for its in-Kingdom deployments, "This is what AI sovereignty looks like", and the fact that the category's largest public company earns 86% of its revenue from sovereign-linked buyers, establish that sovereigns pay, today, for custody of compute inside their borders. That demand is not a forecast; it is the current customer. The step this document sells, custody of the *model*, as an immutable certified artifact, is the same buyer one clause further, and closing that clause is the program's commercial gate (§23).

The network effects close the Part, because generations compound the way the Strategic Thesis says reference compounds. Every certified deployment enlarges the shared evidence base the next conformity assessment cites. Every generation's shell recipe improves the next generation's training. Adapter task libraries port across silicon generations and across Champions. Sovereign expert libraries and signed curriculum packs accumulate into governed catalogs; every appliance that syncs feeds Radiant. The appliances are the universal-service obligation made physical, and Intelligent Silicon's own reference, in the Strategic Thesis's sense of the word, is the accumulating certified proof that frozen intelligence works. One custody chain, from protocol to lithography to schoolroom.

Part VII

Timing

The adjacent category was proven this year on someone else's silicon, and the printed model is still unclaimed; the recipe was published by the largest company on earth; the regulator moved the deadline, the sovereigns didn't wait, and the first proof now costs less than a Series A.

22 Why Now

Seven reasons, each one sentence, each pointing at evidence already tabled. The adjacent category shipped: hardwired-weight silicon is running with demonstrated economics, and the printed model, bounded state, both memories gone, remains unclaimed (§13, §19). The recipe is published: every step of Section 10 now has an open reference, which collapsed training risk for everyone, including us. The capacity is uncontended: mature-node lines are idling as the leading edge moves on, and this design needs no HBM allocation at all (§15, §6). Sovereign procurement is live ahead of regulation, with custody language and committed money (§21). Certification arrives on a legible schedule, and the first certified artifact will define what conformity looks like for the category (§7). The first proof is cheap: G0 costs eval-branch money, G1 less than a Series A, and both run on idling capacity (§13, §17). And every month the position stays unclaimed, the category's eventual first mover hardens into its default.

The window is a gap between proofs rather than a moment, and the first proof is already priced.

23 What Must Be True

Honesty about fracture points is the discipline that separates a thesis from a pitch. Seven things must be true, and each is gated before any mask is cut, six in the lab, one in the market:

(**a**) the ternary shell holds quality at 14B on long-form generative work; (**b**) converted diffusion reaches parity after alignment at quality-preserving commitment rates; (**c**) the step count K falls as far with editing and distillation as the speed targets assume; (**d**) the fold retains what naive sparse-to-dense distillation loses; (**e**) measured GPU joules-per-token leave the modelled energy gap intact; (**f**) the appliance clears acceptance and escalation-safety thresholds under protocol constraints, and (**g**) a sovereign buyer contracts for a version-locked artifact on utility terms, the demonstrated demand for custody of compute (§21) converting into demand for immutability of the model.

The nesting reshapes the list's character: the shell-quality question is a curve watched *during* pretraining rather than a cliff discovered after it, and the NVFP4 floor exists whatever the curve shows. Gate (b) is different in kind from the precision gates: the precision bets have floors on the same machine, but bounded-state decoding *is* the machine. If conversion parity is never reached at deployment scale, the failure is discovered in software, where it costs a checkpoint, and what remains without it is the hardwired-accelerator category, which is already occupied and which this document deliberately does not claim. The program's instrument is a four-track experimental register, model, silicon, system, appliance, whose spine is a one-to-three-billion-parameter full-pipeline dress rehearsal, gated at ninety percent of the autoregressive baseline. And the register's defining economy is the one Section 13 built: the dress-rehearsal checkpoint is G1. The program's cheapest experiment and its first product are the same object.

The program's technical fracture points are placed where they are cheapest to discover, in software, during a run already being paid for, before any mask is cut. The commercial one is placed where it is most visible: in a named contract, before the volume rungs are climbed. And the first mask now costs less than the experiment used to.

24 Conclusion

The document reduces to one observation and one consequence. The observation: intelligence is a manufactured good being sold as a service. The consequence: whoever manufactures the certified good, and escrows it to the jurisdictions that depend on it, owns the physical layer of the intelligence age.

The first proof is smaller than the thesis. A single printed learned function, running from sunlight, speaking through a speaker, answering without a network, proves the category more cleanly than a datacenter card ever could. From that artifact the stack scales upward through teacher box, health aide, Champion card, sovereign mask, expert plane, and utility layer, each rung a smaller bet than the last one retired, each artifact inheriting the same three properties: it cannot drift, it cannot be cut off, and it cannot be taken back.

The law beneath all of it is three words wide. Print the prior, stream the state, route the novelty. The token was its first unit and the answer its first product; the deliberated decision, the embodied skill, and, further out than this document reaches, the generated world-second are the same architecture under later interfaces. What holds at every scale is the single claim argued here from the first page: the expensive, stable, repeated part of intelligence belongs in silicon a jurisdiction can hold.

The Strategic Thesis asked who will own the intelligence age. This document answers where that ownership physically lives: in weights that never move, on silicon a jurisdiction can hold, from the national datacenter to the village school.

The weights never move. The value never leaves. And the answer reaches wherever sunlight does.

A The Mathematics

The results the argument rests on, in compact form: the objectives, the lattice, and the forcing that fixes each choice. Each result is given with enough structure to check, and with the section it supports.

E.1 · Two traversals, one operator. The autoregressive factorisation computes the exact likelihood token by token; masked block diffusion optimises a weighted bound over parallel unmaskings of a bounded block. The operator is shared; only the traversal measure differs, the formal ground for conversion by continual training (§10).

E.2 · The coupling objective. Conversion trains $\mathcal{L}(\theta, \Delta) = \mathcal{L}_{\text{AR}}(\theta) + \lambda\, \mathcal{L}_{\text{diff}}(\theta + \Delta) + \mu\|\Delta\|$. The frozen-tower method is the $\mu \to 0$ corner with θ fixed (Δ a full second tower, $2\times$ parameters); naive conversion is the anchorless opposite corner. The plan of record is the interior, one tower, a norm-bounded Δ, which is also the delta plane the bottom die reserves (§12). The Stage-0 sweep over (μ, λ) both selects the method and dimensions the silicon.

E.3 · The seeded sampler. Consistency distillation collapses the schedule to K steps; with a quantised score, residual error is annealed by in-loop noise whose amplitude matches the lattice, the "temperature" the card's keyed PRNG generates (§14). Determinism survives because the noise is a keyed counter stream: same key, same answer.

E.4 · NVFP4. E2M1 elements under two-level scaling (16-element E4M3 block scales beneath a tensor scale), stochastic rounding on gradients, Hadamard rotation on the gradient path, the change of basis onto the Gaussian form the format assumes; attention held high-precision because softmax exponentially amplifies quantisation noise.

E.5 · The nesting theorem. With shared block scales, each NVFP4 weight projects to $\{-1, 0, +1\}$ among seven candidates with a closed-form block-optimal choice; matryoshka co-training keeps the ternary child on the loss surface throughout. Parent ladder, the projection gap: FP16 $\approx 10\times$, FP8 $\approx 5\times$, **NVFP4** $\approx 2.25\times$, the unique format close enough to nest. MXFP4's ∼36% token handicap prices the alternative.

E.6 · The capacity fixed point. PTQ damage is the projection term $\Delta\mathcal{L} \approx \frac{1}{2}\,\delta^{\top} H\,\delta$, sharpened by overtraining; native training sends $\delta \to 0$. Utilisation is an estimated 2 bits/param; the ternary container holds ≈2.0 effective bits, container sized to contents, the overflow routed to retrieval by design (§9).

E.7 · Open loop vs closed loop. Distilled open-loop error accumulates as $O(L\,\varepsilon)$ over commitment length; editing restores contraction toward $\varepsilon/(1-\rho)$. Distillation spends robustness; editing buys it back, and its value grows with lattice coarseness, which is why it is co-trained for this chip in particular (§10, §12).

E.8 · Pipeline utilisation. Autoregressive decode at batch 1 uses $\sim 1/L$ of an L-stage pipeline; a width-S block recirculated K times approaches full utilisation at batch 1. The theorem behind "AR is a loop; diffusion is a circuit" (§14).

E.9 · The fold and the power mean. A linear router composes into expert keys: $\hat{k}_j = k_j + \lambda\, w_r$, values rescaled by calibrated expected gates, a near-function-preserving dense initialisation from a sparse teacher. After merging duplicates (redundancy η) and compressing on the measured

workload, the dense-equivalent size is a power mean $M_{-\alpha}(\eta N_{\text{tot}}, N_{\text{act}})$. For a 30B-total/3B-active teacher, a task-mix band of ∼7B (procedural) to ∼24B (closed-book recall), the 14B student at parity-with-margin on the former, covered by retrieval on the latter (§10).

E.10 · The allocation law. Maximise $\sum_i$ quality subject to $\sum_i c_i \le C$: water-filling, $\partial \mathcal{L}_i / \partial c_i = -\lambda^*$. The model's confidence estimates the marginal; the commitment threshold γ *is* λ^* (§14). Iterations and precision allocate on-die today; width joins at G3; depth is rejected; lockstep is worth more than the skip.

E.11 · The wafer table. At ∼70 Mbit/mm², modelled and anchored to the ternary-ROM measurement of §6, the NVFP4 floor (∼63 Gbit) needs two dies, ∼40–45 yielded models per 300mm wafer; the ternary shell (∼24–28 Gbit) fits one die, ∼85–90 yielded. The option's payoff, in wafers.

E.12 · Load residency. Effective cost $= c_{\text{ASIC}} + c_{\text{host}} + (1-p)\, c_{\text{verifier}}$, with acceptance p a function of the price γ, Section 16's selection principle as the system-level image of E.10.

E.13 · Binary vs ternary, formalised. Expected switching energy $\propto$ nonzero density; reduction-tree depth falls with sparsity; the zero is an omitted operation. Ternary's ∼50% trained zero-rate is a circuit resource, the measured density-vs-zero-ratio curve is its signature (§11).

E.14 · The domain-wall budget. [Modelled; measured at Stage-0/Go.] Per-crossing traffic is $S \cdot d \cdot b$ bytes for block width S, model width d, activation precision b. On the reference shapes ($d \approx 5{\times}10^3$, FP8), a crossing is ≈0.16 MB at $S = 32$ and ≈0.33 MB at $S = 64$; with A attention blocks ($A = 6$ of 62 in the reference skeleton) crossed K times bidirectionally, total traffic is $2AK$ crossings, at the ceiling ($S = 64$, $K = 8$), 96 crossings and ≈31 MB per block generation, ≈0.5 ms of transfer at PCIe Gen5 x16 (∼64 GB/s). The latency terms: per-crossing initiation overhead of 5–10 µs adds 0.5–1.0 ms at the naive ceiling, and falls by an order of magnitude when the twelve crossings of each recirculation step are coalesced into one transfer per direction per step. Against the block budget the committed speed implies, ≈3.2 ms per 64-token block at 20K tokens/second/user, the naive ceiling fits with $\ge 2\times$ margin at $K = 8$ and $\ge 4\times$ at $K = 4$; Gen6 doubles the transfer margin again.

B The Evidence Register

Every claim in this document tagged as measured or reported, mapped to its anchor. Anchors are stated at the level needed to locate the source. Claims resting on a company's own disclosures carry that tag here and in the text; nothing in this register is load-bearing beyond the weight the text gives it.

Measured on silicon

The ternary-ROM measurement (§6, E.11): a 2026 ternary read-only-memory accelerator study; 15.0 MB/mm² at ∼70% zero weights, 3.3× SRAM density on the same node, 200 TB/s on-chip, against an SRAM baseline of 0.021 µm² per bit. The Taalas HC1 (§13, §19): launch materials and independent press measurement, February 2026; TSMC N6, 815 mm², 53 billion transistors, Llama-3.1-8B in mask ROM, 16,960 tokens/second/user, ∼250 W, $0.0075 per million tokens against the company's own B200 comparison of $0.0379; successor at 20B parameters in MXFP4 with SRAM on a separate die, the sequence state relocating, not leaving; press-reported visible quality loss from 3–6-bit post-training quantisation; weights-to-GDSII in about a week, vendor-reported.

Published training results

NVFP4 native pretraining (§10, E.4): the vendor's technical report; 12B parameters over 10T tokens within ∼1% of FP8; attention held high-precision; the two-level scaling, stochastic rounding, and Hadamard rotation of E.4; MXFP4 requiring ∼36% more tokens to match. Ternary native training (§10): the open 2B-parameter/4T-token ternary LM line and successors. Block-diffusion conversion (§10): conversion by continual pretraining demonstrated at 100B; the coupled-tower fallback at 30B with 98.7% retention, developer-reported; commercial diffusion LMs serving at 1,000+ tokens/second at near-parity, maker-reported; quality-preserving parallel commitment averaging ∼2.4× at 30B on the vendor's benchmark. Knowledge capacity ∼2 bits per parameter (§11): published capacity-scaling result.

Market record

Groq → NVIDIA (§18, §19): transaction coverage, December 2025; assets acquired for ∼$20B in a licensing-and-assets structure. Cerebras (§18, §19, §21): listing prospectus and coverage, May 2026; $5.55B raised in the largest US listing of 2026, on $510M revenue growing 76%, with 86% from two sovereign-linked buyers. Etched (§18, §19): the company's own disclosures; $800M raised, $1B in contracts, unshipped; used only with the company-reported tag. Inference price decline (§1): published price-trend analysis; median ∼50× per year at fixed quality, a 9–900× range across model classes. The Arm royalty template (§20): quarterly royalty disclosures; compute-subsystem rates above 10% of chip price, royalty revenue above $620M per quarter. Mining precedent (§4): the public ASIC efficiency record; orders-of-magnitude J/TH improvement across the ASIC decade, baseline-dependent, and GPU mining's exit within months of first silicon. Mature-node capacity and wafer pricing (§15): industry analyst reporting of idling N6/N7 lines; wafer prices ∼$9,500 (N6) and ∼$4,000 (16nm-class) as modelling anchors. ASIC cost benchmarks (§17): published industry models at $217–300M for a 7nm-class SoC, with practitioner estimates near $160M. The small-model economics quote (§9): NVIDIA's position paper on small language models as the future of agentic AI. The sovereignty quote (§21): the Saudi sovereign-AI company's public language for its in-Kingdom

deployments on the deterministic-inference vendor's hardware, 2025; a distinct anchor from the revenue-concentration figure beside it, which is Cerebras's.

Regulatory and legal

The EU AI Act calendar (§7): Regulation (EU) 2024/1689; prohibitions February 2025, GPAI obligations August 2025, penalties 2 August 2026, and the Digital Omnibus agreed May 2026, deferring high-risk conformity to December 2027 (stand-alone) and August 2028 (embedded). Mask-work protection (§20): the US Semiconductor Chip Protection Act, 17 U.S.C. ch. 9, with European topography parallels; advanced as theory rather than asserted, as the text states.

C Key Terms

Champion The Strategic Thesis's central institution: a utility-class AI company a jurisdiction licenses, owns a stake in, and holds to account; the buyer and deployer of the artifacts this document describes.

Intelligent Silicon The fabless entity this document describes. It compiles co-designed models to masks and sells printed learned functions: certified, escrowed, respin-refreshed, royaltied per device.

II-Account / II-Agent / Radiant The Strategic Thesis's citizen account (the universal-service entitlement the appliance makes physical), its orchestration layer (routes each request to the cheapest tier that can answer it; the router whose threshold is γ), and its retrieval layer (the governed, versioned, citable store of facts, records, and curricula, the memory the etched model consults instead of memorising).

Stage-0 The software-only experimental phase before any mask is cut: the checkpoint runs, gate measurements, and watched curves whose passing releases G0/G1.

Adapter A megabyte-scale, cryptographically signed task module loaded into the die's SRAM island; portable across silicon generations.

Block diffusion Decoding a block of tokens in parallel by K repeated applications of one fixed denoising operator, rather than one token at a time; currently the only published route to quality generation with bounded state, which is why the architecture is forced rather than preferred (§14).

Delta plane A bounded, norm-constrained divergence from the printed backbone, held beside the residual weights, that lets a frozen operator be adapted without a second tower.

Draft mode Running the ternary path as a cheap proposal in front of a high-precision verifier (hardware speculative decoding), priced by acceptance rate.

The fold Collapsing a mixture-of-experts teacher into a dense student by composing the linear router into the expert keys, then merging, compressing, and distilling.

γ **(gamma)** The commitment threshold: the confidence level at which the machine commits a token; mathematically the shadow price of a joule (E.10), operationally the deployment's one quality–speed knob.

HBM High-bandwidth memory, the stacked DRAM whose supply allocation constrains today's inference fleet.

K / K-register The number of recirculations of the fixed operator per block; programmable 1–8 and certified per artifact, so training-side improvements arrive as field upgrades.

Mask ROM Read-only memory whose contents are set by the photolithographic masks at manufacture; the weights' final form.

The nesting Co-training a ternary child model inside an NVFP4 parent under shared block scales (matryoshka), so shell quality is a curve watched during pretraining; the program's principal novel claim.

Printed model A model manufactured as silicon: weights in mask ROM, bounded state as the only mutable memory, so both the weight hierarchy and the growing sequence cache are deleted. Distinct from a hardwired accelerator, which etches the weights but keeps the conversation's state in mutable memory beside them.

NVFP4 / MXFP4 Competing four-bit floating-point training formats; NVFP4's two-level block scaling makes it the unique parent close enough to nest ternary (E.5).

PTQ Post-training quantisation: projecting an already-trained model onto a coarser lattice, with a loss penalty native low-bit training avoids.

Residency The fraction of a workload the local artifact answers without escalating to a verifier or the frontier (E.12).

Respin Re-fabricating only the changed mask layers of an existing design; for a via-ROM model, a weight update at ∼$1M and weeks.

Ternary Weights restricted to $\{-1, 0, +1\}$; the zero is an omitted operation, which is why ternary, not binary, minimises the useful circuit.

Via-ROM / wafer banking Building identical, model-agnostic base wafers ahead of demand and parking them before the final one or two via layers, which carry the model; the escrowed sovereign artifact is those via masks.

The wave The Strategic Thesis's base-case token-demand growth of 1,000–10,000× by 2032.

Intelligent Internet · Intelligent Silicon: The Silicon Thesis · July 2026

First Edition · Companion to the Strategic Thesis

Confidential draft, not for distribution

Disclaimer

Date: July 2026

This memorandum is issued by L42 Ltd., trading as Intelligent Internet (the "Company"), and is directed only at persons to whom it may lawfully be communicated. It has not been approved by an authorised person for the purposes of section 21 of the Financial Services and Markets Act 2000 ("FSMA"). In the United Kingdom, it is directed only at persons falling within one or more of the following categories under the Financial Services and Markets Act 2000 (Financial Promotion) Order 2005 (the "FPO"): (i) investment professionals (Article 19 FPO); (ii) high net worth companies, unincorporated associations, or partnerships (Article 49 FPO); (iii) certified or self-certified sophisticated investors (Articles 50 and 50A FPO); or (iv) certified high net worth individuals (Article 48 FPO).

In the United States, this memorandum is directed solely at accredited investors as defined under Rule 501 of Regulation D of the U.S. Securities Act of 1933, as amended (the "Securities Act"). The securities described herein have not been and will not be registered under the Securities Act or any U.S. state securities laws, and may not be offered or sold in the United States absent registration or an applicable exemption therefrom.

In all other jurisdictions, this memorandum is directed only at persons to whom it may be lawfully communicated under applicable local laws and regulations. It is the sole responsibility of any recipient outside the United Kingdom and the United States to satisfy themselves that they may lawfully receive and act on this memorandum in their jurisdiction, and the Company makes no representation that this memorandum complies with the laws of any jurisdiction other than England and Wales.

This memorandum is provided on a strictly confidential basis for informational purposes only in connection with evaluating the business and prospects of the Company. It does not constitute an offer to sell, or a solicitation of an offer to buy, any securities, nor a recommendation to invest. It must not be copied, reproduced, or distributed to any other person without the Company's prior written consent.

The information contained herein does not purport to be complete and has not been independently verified. No representation or warranty, express or implied, is made as to its accuracy or completeness, and no responsibility or liability (whether direct, indirect, or consequential) is accepted for any errors, omissions, or misstatements. Recipients should conduct their own independent investigation, due diligence, and analysis and should obtain independent financial, legal, and tax advice before making any investment decision. The Company is not acting as financial adviser or fiduciary to any recipient and undertakes no obligation to update or correct information herein.

This memorandum contains forward-looking statements involving known and unknown risks and uncertainties, including those specific to the rapidly evolving AI regulatory, technological, and competitive landscape. Actual results may differ materially from those expressed or implied. Forward-looking statements reflect the Company's views only as at the date of this memorandum, and the Company undertakes no obligation to update them except as required by law.

This memorandum is governed by the laws of England and Wales.

L42 Ltd.

Suites 404/405, Pavilion Club, 96 Kensington High Street, London W8 4SG