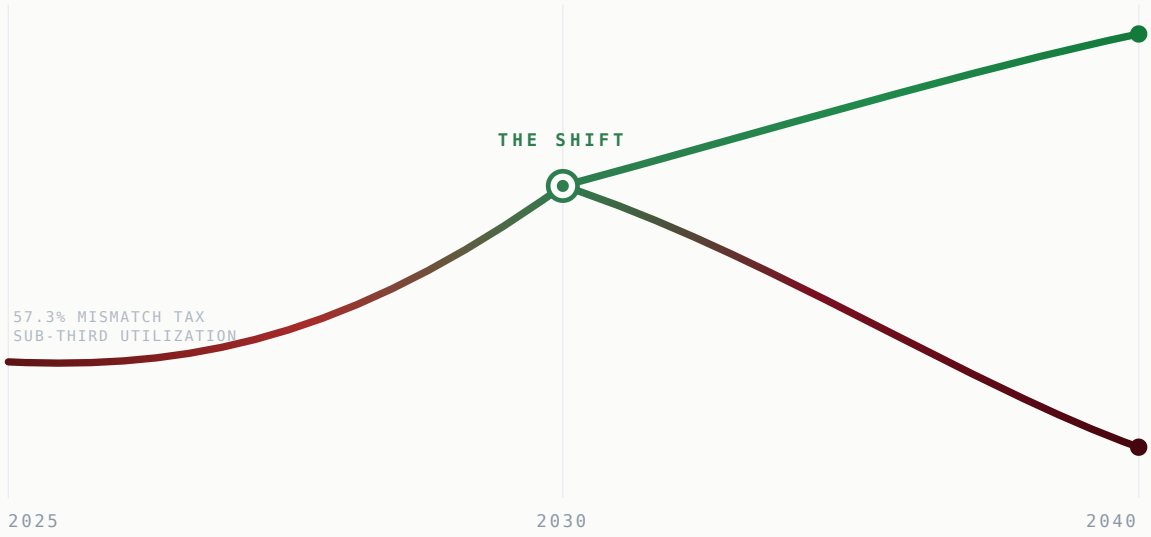
Compute 2030

The substrate takeoff, told forward, threshold by threshold, with two endings.



ROAD A Heterogeneous Equilibrium *owned by no one*

ROAD B Substrate Capture *owned by one*

We predict that the impact of general intelligence for silicon over the next decade will be enormous, exceeding that of the semiconductor revolution itself.

We wrote a scenario that represents our best guess about what that might look like. It is informed by hardware roadmap extrapolations, fleet utilization data, wargames run with sovereign and hyperscaler customers, expert interviews with compiler engineers and chip architects, and previous forecasting work on the general intelligence for silicon stack.[1]

We have tried to be concrete and quantitative, which means depicting one of many possible futures rather than hedging across all of them. We will not be right about everything. Some of the dates will slip, some of the thresholds will arrive in a different order, and the named events are illustrative rather than load-bearing. What we are confident in is the shape, because the forces are already in motion, the first measurements are already in, and the window is already closing.

This is a forecast with a spine of measurement, which is what we think distinguishes it from speculation about the future of intelligence. The central quantity is not a guess; it is a number, taken on instruments, with a confidence interval. The architecture explosion is not a metaphor; it is a count, climbing on a curve driven by costs that are already collapsing. The labor wall is not a fear; it is a growth rate, six percent a year, set against an exponent that is steeper. Our confidence lives in those measured anchors, and our humility lives everywhere the anchors run out, which is why the dates in what follows are pegs accurate to within a window and the mechanisms are the actual claim.

The argument is simple to state. The variety of silicon is about to outrun the human capacity to make it usable. The gap between what the world's compute can do and what it actually delivers is already measured at 57.3 percent.[2] That gap will widen as the architecture count climbs, until something is built that can absorb the variety with no human in the loop. That something is a model of how all silicon behaves - general intelligence for silicon - and whoever builds it will not be selling software. They will be writing the rules for how the world's compute is governed.

The reason the answer has to be a model, and not a better tool, is older than the crisis it solves. A theorem from the cybernetics of the last century holds that every good regulator of a system must contain a model of that system: you cannot reliably control

what you have not, in some formal sense, modeled. If the system to be regulated is the behavior of all silicon, then the regulator must be a model of how all silicon behaves. Not a faster scheduler, not a cleverer autotuner, both of which have been tried and both of which fall further behind as the variety grows, but a learned model of the physics underneath the chips. That is what general intelligence for silicon is, and why the thing that finally absorbs the variety is an intelligence rather than an optimization.

We wrote two endings, because the same shift can arrive two ways. In one, the layer that absorbs the variety is neutral, open, and owned by no one. In the other, it is captured - by a single company or a single state - before anyone decides that it should not be. The difference between the two is not the technology, which arrives either way. The difference is whether anyone acts inside the window. This is not a recommendation and not an exhortation. It is a forecast, and we hope to be argued with.[3]

PART I **The Crisis and the Closing Loop**

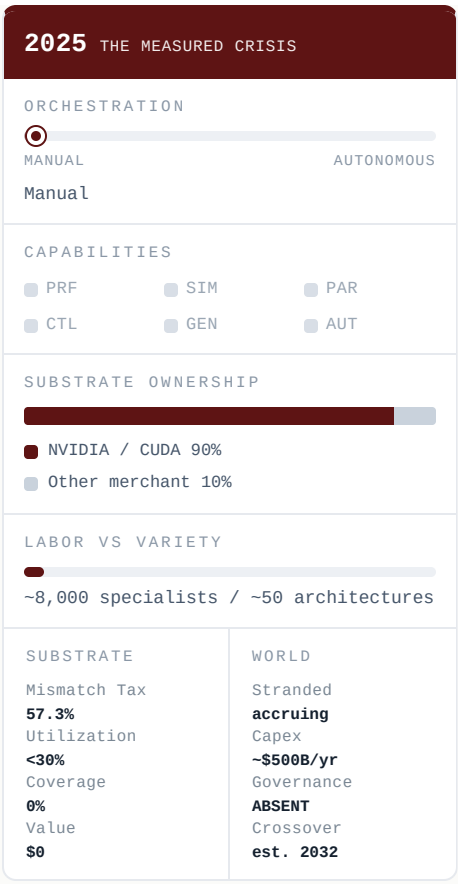
2025 to the branch point, 2028

2025 - The Measured Crisis

> *The machines are running, and almost no one has measured how badly. The waste is invisible because it is distributed: a few points here, a few points there, eight thousand exhausted specialists each losing a Monday.*

It is 2:17 in the morning in Osaka, and Priya is on her 847th configuration. The chip on her bench is a dataflow accelerator from a Taiwanese fabless startup, twelve weeks out of the fab, in no public benchmark and no compiler toolchain that knows it exists. Her job is to make it run her customer's inference workload by Monday. She types a number into a configuration file and waits for the workload to compile and run. The monitor reads thirty-four percent of the chip's rated throughput. She changes a tiling parameter. Forty-one. She changes the memory layout. Thirty-eight. The chip is capable of eighty percent on this workload and she knows it, and the gap is not a mystery to her; it is the cost of a part whose behavior no one has yet encoded into the tools she is using, and that encoding is weeks of work she does not have. At 2:31 she commits the number that gave her forty-one and goes to bed. The chip will run at forty-one until someone has the time to do it properly, which is to say, in the experience of everyone in this profession, never.

Priya's profession does not have a clean name. The people in it are called performance engineers, or kernel engineers, or compute architects, or, in the job postings that cannot quite describe what they do, heterogeneous systems specialists. What they share is a rare and slow-built skill: the ability to take a piece of silicon whose behavior is documented badly or not at all and coax a workload onto it at something near its potential. There are, on the most generous estimate, about eight thousand of them on Earth who can do this for a genuinely new architecture, and that number is the binding constraint on everything that follows.[4] It is not the count of people who can write a kernel for the dominant vendor's chips, which is large and growing. It is the count who can do it for a chip the dominant vendor's tools have never seen, which is small, slow to grow, and about to be asked to do something impossible.



The gap between forty-one and eighty is not an anecdote, and in 2025 it gets measured properly for the first time. The measurement is deliberately unglamorous: a controlled testbed of heterogeneous production accelerators, the same workloads run on each, the default vendor configuration compared against the best a team of experts can reach given effectively unlimited time. Eight campaigns, ninety-two days, 847 configuration-workload pairings per architecture.[5] The median gap between default and expert-tuned throughput is 57.3 percent, with a 95 percent confidence interval of 55.1 to 59.4, and the distribution is right-skewed, which means the median understates the worst cases, the newest and strangest silicon where the gap yawns widest. The report names it the Mismatch Tax: the directly measured cost of silicon variety outrunning the human capacity to absorb it.

The Mismatch Tax is one layer of a deeper waste, and the two should never be confused, because confusing them is the fastest way to sound impressive and lose the argument. The 57.3 percent is the substrate-mismatch component alone, the gap between default and optimal on a given chip. Stacked on top of it are kernel inefficiency, cluster fragmentation, and the simple temporal idleness of expensive hardware waiting for work, and when all of them are summed, the world's AI fleet delivers well under a third of its theoretical peak.[6] The two figures, the measured 57.3 percent and the broader sub-third utilization, are different measurements of different things, and the report keeps them apart on every page.

This is happening at the center of the largest capital mobilization in the history of technology. Hundreds of billions of dollars a year, rising toward a cumulative five trillion by 2030, are being pointed at compute hardware that runs at a fraction of its rated capability.[7] The waste is invisible precisely because it is so widely distributed and so quietly absorbed. Consider the arithmetic from inside a data-center operator: a GPU depreciates whether or not it is used, and whether or not it is used well. A fleet running at thirty percent of potential is a fleet on which two-thirds of the most expensive line item on the balance sheet is evaporating, on schedule, into heat and idle cycles. No operator reports this, because reporting it would mean publishing the fact that the assets the market values them on are idling, and the

market has not yet learned to ask.[8] It is eight thousand Priyas, each losing a Monday, summed across an industry and rounded off as a cost of doing business.

The obvious objection, in 2025, is that this is what compilers and autotuners are for, and that the gap will close on its own as the tooling matures. It will not, and the reason is the shape of the problem. An autotuner searches the configuration space of one chip running one workload; it is a per-chip, per-workload effort that yields an answer good until the chip or the workload changes, at which point the search begins again from nothing.[9] It does not generalize, because it learns nothing transferable; it is a flashlight, not a map. And the thing the autotuner is racing is not standing still. The number of chips is about to explode, and a method that must restart its search for every new chip falls further behind the faster the world moves. The tooling is not the answer. The tooling is the thing the answer has to replace.

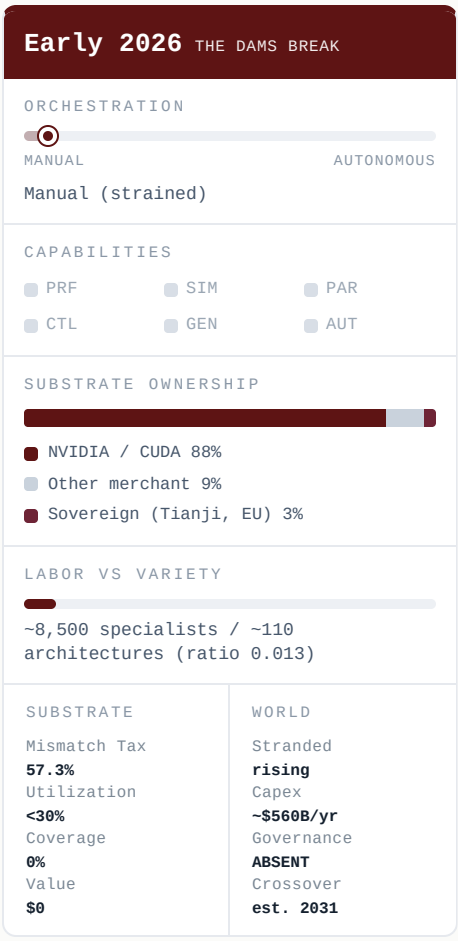
The reason the waste does not simply get fixed is that fixing it is human work, and the humans do not scale. The specialist pool grows at roughly six percent a year, the historical rate of an adjacent technical discipline, and it cannot be accelerated with money, because it takes three to five years to train one of these people and the pipeline is fixed.[10] For seventy years this did not matter, because the number of architectures that mattered was small and stable, anchored by one dominant instruction set and one dominant accelerator vendor. The whole industry made an implicit bet, so deep it was never stated aloud: that hardware diversity is bounded, and that human expertise scales with it. Both halves of that bet are about to come due at once, and the report's central claim is that they come due on a schedule, not a maybe.

And in a small number of research groups, a different idea is taking shape, drawn from an old theorem. In 1970 two cyberneticists proved that every good regulator of a system must contain a model of that system; you cannot reliably control what you have not, in some formal sense, modeled.[11] The corollary, in 2025, is uncomfortable and clarifying: if the system to be regulated is the behavior of all silicon, then the regulator must be a model of how all silicon behaves. Not a faster autotuner, not a better heuristic, but a learned model of the physics underneath the chips, the thing the autotuner conspicuously is not. One such group has a forty-seven-billion-parameter graph neural network in training, fed on the execution traces of hundreds of distinct processors, learning to predict how a chip it is shown will behave on a workload it is given.[12] They call it HP1. In 2025 it is a research artifact. It does not work yet, and most of the industry has never heard of it.

The markets, in 2025, are not interested. They are pricing the chip makers, the model labs, and the data-center builders, and the layer that makes the chips usable is not yet a category anyone trades. The government is not interested either; its attention is on who manufactures the silicon and who is allowed to buy it, a contest fought one layer below the one that will come to matter. The crisis is measured, enormous, and entirely unaddressed, which is the condition from which every other chapter in this report departs.

Early 2026 - The Dams Break

> *For seventy years two dams held the architecture count near fifty: the cost of designing a new chip, and the time it took. In early 2026 both break at once, for the same reason the rest of this story accelerates: the design of silicon is itself becoming a workload that models can do.*



The architecture count held near fifty for as long as it did because two dams held it there. The first was cost: designing a new processor meant a multi-year program and a team of specialists most companies could not assemble. The second was time: even with the money and the people, the cycle from concept to working silicon ran to years. In early 2026 both dams break, and they break together.

The time dam goes first. Cadence extends its ChipStack agent to what it calls Level-5 autonomy and reports validation running forty times faster, a verification loop that took five weeks collapsing to under a day; Synopsys builds toward an agentic design workforce on the same horizon.[13] The significance is not that any one step gets faster. It is that the slowest, most human-intensive parts of chip design, the floorplanning and synthesis and the verification that consumed armies of engineers and months of calendar, become things a model does overnight. A design cycle that ran in years begins to run in weeks, and a process that ran in weeks begins to run in a day. The barrier that kept the architecture count low was never only money. It was the sheer human time a new chip demanded, and that time is collapsing.

The cost dam goes next, breached by a different move. Hardwired-model silicon, the approach a company like Taalas is pursuing, etches a frozen model directly into a small number of customer-specific mask layers on an otherwise standard base die.[14] It collapses the development cost and brings a new part back from the fab in roughly two months rather than years, and it collapses the cost of running the model by orders of magnitude. It is worth being precise about what it does not do, because the careless version of this claim is wrong: it does not make silicon free. The per-tapeout mask cost stays high, in the tens of millions, and rises with model size. What hardwired-model silicon changes is more consequential than cheap. It makes a custom part fast to attempt, so the number of teams who can try, and the speed at which they can try, both jump, and a thing that is fast to attempt gets attempted by people who would never have committed years to it.

With both dams down, the count begins to climb, past fifty toward a hundred and ten by the middle of the year, and the climb has a particular texture.[15] It is not one new dominant chip. It is a proliferation of specialists: an accelerator tuned for a single model family, a memory architecture built for one access pattern, a dataflow part that makes sense

only for a narrow class of workloads. Each is excellent at its niche and useless without a human to make it usable, and each is a Priya problem waiting to happen, a processor in no toolchain whose behavior no one has encoded, running at a fraction of its rating until a scarce human finds the time. The architecture explosion and the Mismatch Tax are the same phenomenon seen from two angles: more variety means more default-to-optimal gaps, summed across a fleet diversifying faster than anyone is prepared for.

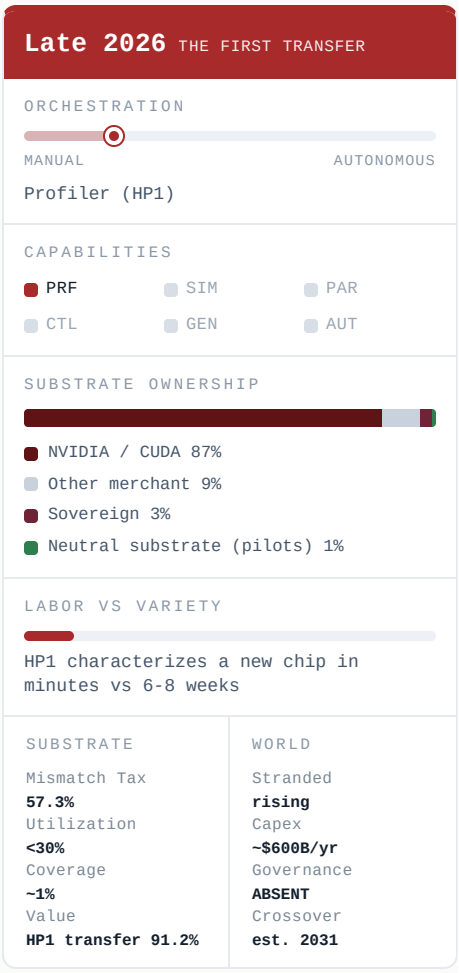
The natural objection is that the market will simply respond by training more specialists, the way markets usually close a labor shortage. It cannot, on the timescale that matters, and the reason is the pipeline. A heterogeneous-systems specialist takes three to five years to train and cannot be conjured by raising salaries; the constraint is time and apprenticeship, not wages.[16] The pool grows from eight thousand toward eight and a half over the year, six percent, while the architecture count grows on an exponent with a steeper slope entirely. The lines have not crossed yet. But anyone who plots both can see where they are heading, and in early 2026 a few people start plotting both, and a smaller few start drawing the obvious conclusion: that the gap cannot be closed by people, and will therefore have to be closed by a model, or not at all.

One of the people plotting both is an analyst at a firm that prices semiconductors for institutional investors, and the chart she produces, two lines on a log axis with the architecture count rising steeply and the specialist pool crawling, is the cleanest single artifact of the coming crisis. She circulates it internally with a one-line note that the lines cross around 2030 and that the consequences are large. It is read, filed, and not acted on, because the crossing is five years out, the current quarter is fine, and a chart about a bottleneck that has not yet bitten does not move a market. The crisis is visible, in early 2026, to anyone who bothers to plot it. Being visible and being priced are different things, and the distance between them is most of what the next four years are about.

The first to act on that conclusion are not companies but states, because states think in decades and dependencies. China nationalizes a compute-substrate effort under the working name Tianji, centrally funded and vertically integrated, and begins pouring state resources into a closed stack over its domestic fleet; the program answers to no quarterly earnings call and can move at a speed no consortium can match.[17] The European Union launches a federated, multi-vendor sovereign-compute initiative, correct in its instincts toward neutrality and openness and slower than a directed state by construction, because a body answerable to a dozen national stakeholders and a procurement process cannot sprint.[18] Sovereign-wealth funds in the Gulf, holding capital and patience at a scale private firms cannot assemble, begin to study the space for something to fund or to host.[19] None of them has a working substrate. All of them have understood, ahead of the markets, that the layer which will absorb the variety is strategic infrastructure, and that strategic infrastructure cannot be imported without importing a dependency into the base of an economy and a defense.

Late 2026 - The First Transfer

> *A model trained on enough chips does not need to have seen the next one. This is the load-bearing claim of the entire thesis, and in late 2026 it is demonstrated for the first time, on a chip chosen precisely because the model could not have seen it.*



HP1 finishes training and is put on a bench it has never seen. The test is designed to be unforgiving. The chip is a novel accelerator held out of the training corpus on purpose: a new instruction set, an unusual memory hierarchy, a thermal profile unlike anything HP1 was shown. The team that built the test wanted the cleanest possible answer to the only question that matters, which is whether the model learned the chips or learned the thing underneath the chips. HP1 is given the held-out part and a workload, and in under four hours it generates a working kernel that runs at 91.2 percent of the hand-tuned reference, the version a senior engineer would have spent six to eight weeks producing.[20]

This is the whole game, and it rests on a single distinction that is worth making slowly. A system that had memorized its training chips would perform at chance on a chip it had never seen; the held-out test would expose it in minutes, because memorization has nothing to say about a part that was not in the memory. HP1 performs at ninety-one percent, which means it is not recalling, it is predicting, and to predict the behavior of a chip you have never seen you must have learned the physics that all chips share: how memory hierarchies stall, how interconnects saturate, how arithmetic units fall off their rated peak under a load that does not match the datasheet. The result is a generalization result, not a memorization result, and that distinction is the difference between a clever lookup table and a model of the world.[21] The claim has a name, bounded transferability, and it is also the single biggest way the thesis could be wrong: if transfer error grew without bound as chips got stranger, the substrate could not generalize, the whole edifice would rest on sand, and none of what follows would happen. In late 2026 the error does not grow without bound. It holds, across the held-out architectures the team can find to throw at it.[22]

The honest skeptic has a second objection, and it deserves a real answer: ninety-one percent is not a hundred, and a gap of nine points might be exactly the gap that matters in a margin-thin business. The answer is in the comparison class. The thing HP1 is being measured against is not perfection; it is the status quo, and the status quo is forty-one percent and a lost Monday. A model that reaches ninety-one percent of expert performance in four hours, on a chip no human has yet learned, is not competing with the expert's eventual ninety-five. It is competing with the forty-one that the chip will otherwise run at indefinitely, because the expert never arrives. The relevant gap is not nine points below the expert. It is fifty points above the default.[23]

The result does not stand alone. Independent work that year, predicting maximum GPU utilization at up to ninety-seven percent accuracy across more than thirty-two thousand production jobs on a national supercomputer, establishes that hardware behavior is learnable from features with high fidelity.[24] The report is careful here, because the temptation to overclaim is strong: that work corroborates the predictability HP1 depends on, the general proposition that silicon behavior is a learnable function rather than irreducible craft. It is not a replication of the 57.3 percent placement measurement, and the report does not pretend it is. But the two results point the same direction, from different labs and different methods, and a thesis whose load-bearing claim is independently corroborated is a thesis worth taking seriously.

The protagonist of this story is now visible in outline. It is a lab we will call Special Infusion, and it has just turned a theorem into a result.[25] The reaction is small and, to anyone paying attention, telling. Two hyperscalers run quiet pilots, nothing announced, the way large companies test a thing they suspect is important and do not want to advertise to their competitors or their suppliers. Tianji, reading the same literature the moment it appears, redirects its program toward a sovereign profiler of its own, the first sign that the geopolitical race is now reacting to the science rather than anticipating it.[26] The United States government does nothing, because it does not yet know this happened; the result is a line in a preprint and a number in a pilot, not a headline. The market does not move, because it does not yet trade the layer. The crisis is one shade less hopeless than it was, not because it has eased but because, for the first time, there is a thing that might absorb it.

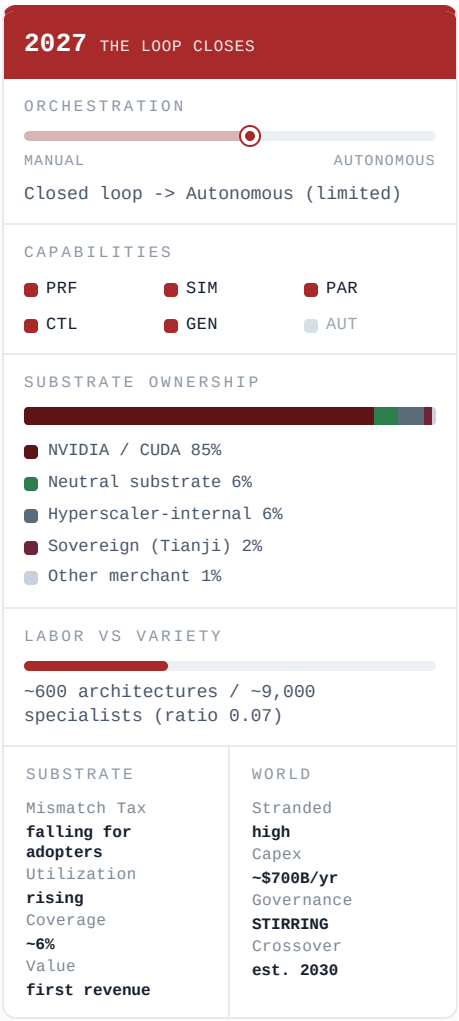
2027 - The Loop Closes

> *One model that reads a chip is a benchmark. Five models that read, rehearse, place, control, and write are an engineer, and in 2027 the engineer ships.*

Across 2027 the rest of the stack comes online, and each model raises the stakes, because each one closes another part of the loop that, until now, a human had to close. HP1 could profile a chip; alone, it was a measurement instrument, a thing that told you how a part would behave but did nothing about it. What 2027 adds is everything between the measurement and the running workload.

CM1 learns to rehearse. It is a compute model that simulates how a whole fleet will behave before a single workload is committed to it, so the system can try a plan in a model of the world and see it fail there, cheaply, rather than in production, expensively.[27] GP1 learns to place. It is a graph-partition model that takes a workload and a fleet of heterogeneous silicon and decides what runs where, the decision a human architect would labor over, made in milliseconds and at a quality a human would need days to match.[28] PO1 learns to control. It executes those placements in live production, adjusting as conditions change, with a runtime-safety component, RS1, nested inside it whose only job is to keep the loop from acting outside the envelope it can guarantee; RS1 is a part of PO1, a brake built into the engine rather than a separate model bolted on.[29] And CG1 learns to write and prove. It generates the kernels the placements require and carries a formal proof of correctness alongside each one, so the code the system ships can be trusted without a human reading it, which is the only way the loop can run at machine speed.[30]

Profile, rehearse, place, control, write and verify. The loop closes, and what it closes into is the thing this report is named for: an Autonomous AI Compute Engineer. The word that matters is autonomous, and it is worth being exact about what it means and does not mean. It does not mean a faster scheduler, and it does not mean a copilot that makes a human specialist quicker. It means the human specialist's entire supply chain, the thing eight thousand Priyas do, compressed into a single system that runs the inner loop with no person in it: it sees a new chip, models it, plans for it, places work on it, writes the code, proves the code, and runs it, and a human enters the loop only to set objectives and audit outcomes.[31] In late 2027 it ships in limited production, on the fleets of the early adopters who ran the quiet pilots. On those fleets, the Mismatch Tax begins to fall, the first time in the history of the measurement that the number has moved in the right direction.



The engine talks to the world through two open protocols, and the decision to make them open is, in retrospect, one of the load-bearing choices of the entire scenario, though almost no one notices it in 2027. ACP, the Asymmetry Context Protocol, runs southbound to the hardware, carrying the descriptions of what each piece of silicon can do. CEP, the Computational Execution Protocol, runs northbound to the workload, carrying what the workload needs.[32] An open protocol is one anyone can read, audit, and implement, which is exactly what lets the engine sit between all silicon and all workloads without locking either side to a vendor. The neutrality the whole thesis depends on is not, at root, a promise the engine's builders make. It is a property of the protocols they publish.

And now the receipts arrive, and they are not projections; they are in the revenue, which is why they persuade. NVIDIA takes roughly seventy-eight cents of every dollar of merchant AI-silicon revenue, not because its chips are seventy-eight percent better than the alternatives but because CUDA made its chips usable and no rival's were; the share is a usability share, not a hardware share.[33] A Western venture cohort that raised more than ten billion dollars to build alternative silicon has recognized under one billion in revenue, a gap that is not a story of bad chips but of unusable ones: the world did not lack good silicon, it lacked a way to make good silicon usable, and the venture market spent ten billion dollars discovering the difference.[34] Intel's Gaudi, competitive hardware with no usability layer, failed in the market and became the cautionary tale every chief executive now cites in board meetings. CentML, a usability layer for a single vendor's hardware, was acquired for the layer, the buyer paying for the knowledge of how to make silicon usable rather than for any chip.[35] Every receipt says the same thing from a different direction: the value does not live in the chip. It lives in the layer that makes the chip usable, and the engine is that layer with the polarity inverted, because it makes every vendor's chip usable and locks the developer to none.[36]

The labs feel it first, because the labs can read the flywheel. Special Infusion is twelve to eighteen months ahead, and every hyperscaler now faces the same fork: adopt the layer, or try to build one. The temptation to build is strong, because a hyperscaler hates depending on anything it does not own, and the obvious objection to the whole scenario is that the giants will simply build their own engines and the advantage will dissipate. A few of them try. What they discover is the thing the next chapter is about: that the moat is not the model, which they could replicate, but the telemetry, which they cannot, because the telemetry is two years of production data from the world's most heterogeneous fleets, and there is no way to buy two years that have not happened yet.[37] Most, having done the arithmetic, adopt. Meanwhile Tianji races its sovereign version, broad enough to matter inside its border and narrow for want of the world's variety; Europe's consortium publishes promising results on a timeline that will arrive too late; a Gulf-backed contender secures the compute and the capital to be taken seriously. And the United States government, finally, begins to notice that something is happening one layer above the chips it has spent years trying to control.

2028 - The Flywheel and the Committee

> *The substrate learns from every workload it touches, and by 2028 what it has learned cannot be bought at any price. The moat is not capital. The moat is time, and time is the one input a competitor cannot acquire.*

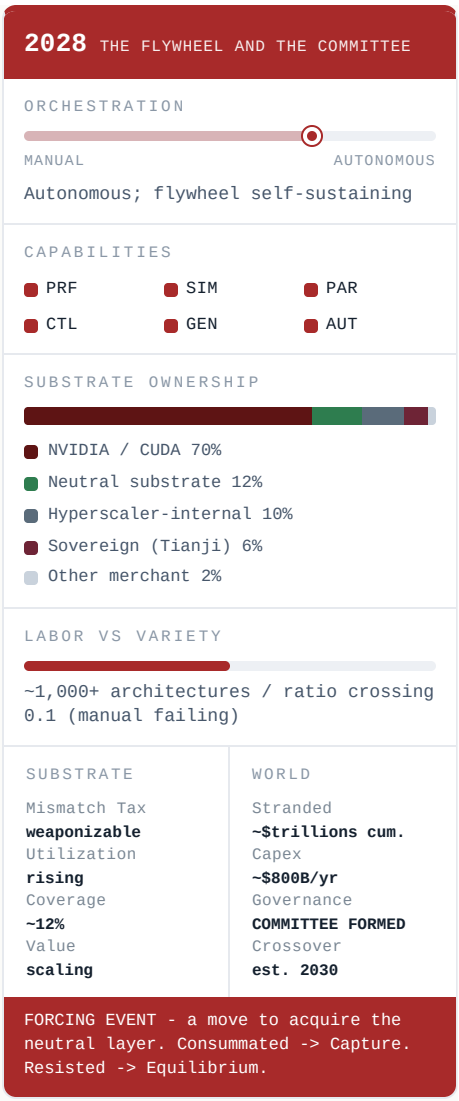
By the start of 2028 the Autonomous AI Compute Engineer has been running in limited production for a little over a year, and the thing everyone underestimated about it is becoming the thing that matters most. It was never only a model. It was a model wired to a feedback loop, and the loop has reached the velocity at which it sustains itself.

The mechanism is the whole of the 2028 story, so it is worth stating exactly. Every workload the engine places generates telemetry: how the silicon actually behaved, where it stalled, how far the prediction sat from the measurement. Every packet of that telemetry sharpens the engine's model of how silicon behaves. Every refinement makes the next placement better, which attracts the next workload, which generates the next packet. The relationship between telemetry and quality is superlinear, because twice the data is more than twice the advantage: the rare cases, the strange chips and the pathological workloads where the value is highest, are exactly the ones a larger corpus is more likely to have already seen.[38] And the corpus cannot be reconstructed. A competitor starting in 2028 cannot synthesize two years of production telemetry drawn from the world's most heterogeneous fleets. They can raise more money, hire better people, and buy more compute, and none of it closes the gap, because the gap is not made of money or people or compute. It is made of time, and time is the one input that does not respond to a balance sheet.[39]

While the moat closes, the crisis it was built to solve crosses a line of its own. The count of distinct production architectures passes a thousand in 2028, on its way to the twenty thousand the models project for 2030.[40] The pool of humans who can integrate a new architecture by hand has grown, at its historical six percent a year, to roughly nine and a half thousand.[41] The ratio between them crosses the threshold past which manual orchestration does not merely strain but fails: there are now too many kinds of silicon for the human supply chain to keep usable, anywhere, at any

salary.[42] Operators who had treated the engine as an optimization, a way to claw back a few points of utilization, begin to treat it as infrastructure, the way one treats the power grid. The framing flips from "should we adopt this" to "we cannot run without this," and that flip is worth more than any quarter of revenue, because it converts a product into a dependency.

The receipts that were visible in 2027 harden into market structure in 2028. NVIDIA still takes roughly seventy-eight cents of every dollar of merchant AI-silicon revenue, not because its chips are seventy-eight percent better but because CUDA made its chips usable and no rival's were.[43] The lesson is lost on no one. Intel's Gaudi, competitive on raw hardware and dead in the market for want of a usability layer, is the cautionary tale every chief executive now cites in board meetings.[44] CentML, which built a usability layer for a single vendor and was bought for it, is the other.[45] The hyperscalers, each sitting on a pile of heterogeneous silicon and a utilization problem, face the same fork the operators faced and resolve it the same way: build the layer, which means years and a telemetry corpus they do not have, or adopt the one that already works. Most adopt. A few try to build, and learn in the trying that the moat is time. By mid-2028 the engine's owner is no longer twelve to eighteen months ahead, as it was a year before. It is ahead by the length of a flywheel that cannot be spun up from a standstill while a competitor is already at speed.



None of this happens in a commercial vacuum, because by 2028 the governments have understood what the markets understood a year earlier: the layer that absorbs the variety is not a product category. It is strategic infrastructure, the foundation every other technological advantage is built on, and a nation that imports it has wired a dependency into the base of its economy and its defense.[46]

China's Tianji program is the furthest along of the sovereign efforts. Nationalized early and funded without the quarterly discipline a public company answers to, it is building a closed, vertically integrated substrate over a domestic fleet, and it is doing so at a speed no consortium and few companies can match, because a directed state does not stop to build consensus or raise a round.[47] Its weakness is the mirror of its strength. A substrate trained on a walled fleet sees less variety than one trained on the world's, and a model of silicon is only as good as the variety it has been shown. Tianji is fast and narrow. The Western engine is broad and, in 2028, vulnerable in a way speed cannot fix: it can be bought.

Europe is the cautionary middle. Its federated initiative has the right instincts, neutrality and openness and multi-vendor interoperability, the precise values a substrate ought to embody, and the wrong velocity, because a consortium answerable to a dozen national stakeholders and a procurement process cannot move at the speed of a directed state

or a venture-backed lab.[48] Europe in 2028 is publishing good results on a timeline that will matter in 2032, by which year the question of who owns the layer will already have been answered without it.

And the Gulf is the wildcard. The sovereign-wealth funds, holding capital and patience at a scale private markets cannot assemble, have concluded that the way to acquire the foundation is to fund it or to host it, and they are shopping.[49] Their entry rewrites the arithmetic of every possible deal, because it places behind any buyer a balance sheet deeper than any chip company's and a time horizon longer than any venture fund's.

The United States government wakes in 2028, late, and the form its waking takes is a committee. The White House convenes a Future of Compute Committee to investigate and track the substrate race: an investigative body, not yet an actor, holding subpoena power it is unsure how to aim and a mandate it is still drafting.[50] It holds hearings. It requests telemetry it cannot fully read. It summons the engine's builders, the chip incumbents, and a rotating cast of academics to explain a thing for which the government has no regulatory vocabulary, because every instrument it owns was forged for an earlier layer of the stack.

This is the deeper problem, and the Committee half-senses it without being able to name it. A government knows how to regulate who manufactures a chip. It has export controls, entity lists, fab licensing, an entire apparatus pointed at the silicon. It has nothing pointed at the layer above the silicon, the layer that decides what runs where and learns from every decision, because that layer did not exist when the apparatus was built.[51] The complexity the Committee is chasing has already migrated upward, out of reach of every tool the Committee holds. A chip-layer regulator facing a substrate is a customs officer standing at a border the cargo has stopped crossing.

Before the offer, there is a hearing, and the hearing is the clearest picture the report can offer of the vocabulary gap made flesh. A senator, well briefed by staff who were briefed by experts, asks the lab's chief scientist a question that had a clean answer in the world the senator grew up in and has none in this one: which country's chips does the engine run on. The scientist explains that the engine runs on all of them, that this is precisely the point, that a single workload placed on a Tuesday might run across silicon from four vendors on three continents and that the engine neither knows nor cares which flag is etched on which die. The senator hears an evasion. The apparatus the senator commands was built to ask whose chips, because for forty years whose chips was the question that mattered and the question the tools could answer, and the apparatus has no register for an answer that is everyone's and no one's at once. The hearing adjourns with the Committee no closer to an instrument and one meeting more certain that it needs one. It is the sound of a government discovering, in real time and on the record, that its questions are aimed one layer too low.

And then, late in 2028, the offer.

A chip owner, one of the handful of firms whose silicon the engine makes usable, backed by sovereign-wealth capital that makes the number impossible to wave away, moves to acquire Special Infusion, the lab that built the engine.[52] On its face it is the financing event of the decade, a vindication, the outcome a founder is supposed to spend a career wanting. In substance it is the most dangerous moment in this entire history, and the danger is structural, not moral.

The engine's value is its neutrality. It makes every vendor's silicon usable and locks the developer to none, and that is the entire reason the world is willing to route its workloads, and its telemetry, through it. The moment a chip owner

controls the engine, the neutrality is not at risk of slow erosion; it is logically dead, because an owner with silicon to sell must eventually choose between making a competitor's chip run beautifully and protecting its own, and there is no version of that choice in which it protects the competitor.[53] The telemetry that vendors fed the engine because it was neutral stops flowing the day it stops being neutral. The flywheel that took two years to spin up begins, quietly, to slow. The acquirer would be buying the single most valuable asset in compute and, in the act of buying it, killing the thing that made it valuable, which it either fails to understand or, understanding, intends, because a slowed-but-owned flywheel that throttles every competitor is worth more to a chip company than a fast-but-neutral one that helps them all.

This is the CentML pattern, but CentML was a usability layer for one vendor and its absorption was a footnote in a trade publication. This is the usability layer for all of compute, and its absorption would be a turning point in the infrastructure of the species.[54] The board convenes. The founder, who has spent three years insisting the layer had to live outside the silicon companies, is handed a number that makes the principle expensive to keep. The Committee, which might in principle block the deal, is still drafting the mandate that would let it.

There are two ways out of that room, and the rest of this report is the two ways. If the acquisition is consummated, if the number wins or the Committee moves too slowly or the sovereign capital behind it proves decisive, the layer passes into the hands of an owner, and the world is on the road to Substrate Capture. If it is refused, if the founder holds or the Committee finds its mandate in time or the government decides the foundation of its economy is not for sale, the layer stays neutral, and the world is on the road to Heterogeneous Equilibrium. The seismic shift is coming either way. What is being decided in that room, in the last weeks of 2028, is who will own it when it arrives.

PART II **The Seismic Shift**

2030

2030 - Full-Scale Deployment. The Seismic Shift.

> *The variety did not stop coming. What changed is that a model, not a person, now absorbs it. That is the whole of the shift, and it is enough to exceed the semiconductor revolution.*

The moment is not loud. There is no launch, no keynote, no single morning when the world wakes up different. There is a threshold, crossed quietly across thousands of fleets over a few months in 2030, after which the Autonomous AI Compute Engineer is no longer a pilot running on the margins of production but the thing that runs production.[55] It profiles a new chip in minutes, places a workload across whatever silicon is present, writes and proves the kernels, and controls the execution, and it does all of this with no human in the inner loop, at a scale no human supply chain

could have reached. Full-scale deployment is the unglamorous name for the most consequential event in the history of compute.

What it changes is simple to measure and hard to overstate. The gap this report opened on - 57.3 percent, the Mismatch Tax, the cost of silicon variety outrunning human capacity - collapses toward single digits on every fleet the engine governs.[56] The trillions of dollars of stranded capital, the hardware that ran at a third of its rating because no one had the months to make it usable, begin doing the work they were bought to do. Compute, which had been scarce in the one way that matters - usable compute, compute you could actually point at a problem - becomes abundant.

Picture what that abundance means at the level of the people who use compute, because the abstraction hides the thing that matters. A startup with a strange new accelerator finds it usable on the day it ships, not in the quarter it finally hires someone who can tune it. A sovereign plugs its national fleet into the layer and inherits the capability without surrendering the hardware. A climate model that ran at a third of its rated speed, so that its resolution was set by waste rather than by physics, runs at capability, and the science it produces is the science that was paid for. The abundance is not a statistic about gross output. It is the removal of a tax that everyone who touched compute had been paying without naming it, and the people who stop paying it feel the relief in the only terms that matter, which is that the expensive thing they bought finally does what it was always supposed to do.

This is why the shift exceeds the semiconductor revolution rather than extending it. The semiconductor revolution made one kind of silicon cheap and then taught the world to program it, and for fifty years that was the foundation everything else stood on. General intelligence for silicon does the thing the semiconductor revolution never did: it makes every kind of silicon usable, neutrally, at machine speed, so the foundation is no longer one vendor's chip and one vendor's software but the whole heterogeneous fleet, absorbed by a model that has learned it.[57] The semiconductor revolution gave the world a substrate. This gives the world the intelligence that governs the substrate.

And here is the thing the dashboards make visible and the prose has to say plainly: the shift arrives in 2030, but it was decided in 2028. The technology was never really the question. The loop closed, the engine shipped, the flywheel turned, and full-scale deployment was a matter of when, not whether. The question was ownership. In 2028, when a chip owner moved to acquire the neutral layer, the world answered something it did not know it was being asked: would the intelligence that governs all compute be owned by no one, or by someone?[58] The seismic shift is the same event on both roads. The board turns green on both roads, because the technology delivers on both roads. The roads differ only in who is standing on the layer when the green arrives.

The pivot is worth sitting with, because it inverts the intuition the technology invites. A reader watching the engine climb its capability ladder naturally asks whether it will work, and the answer, by 2030, is plainly yes; the loop closed and the flywheel turned and the shift arrived on schedule. But whether it worked was never the consequential question, because a working layer owned by the wrong party is a worse outcome than a slower layer owned by no one. The consequential question was decided two years earlier, in a room, over an offer, by people who may not have known they were deciding the governance of the world's compute, and the window in which they could decide it was open for months, not years. The dashboards turn green in 2030. The color was chosen in 2028.

On the road where the acquisition was refused and the government aligned in time, the green holds. The layer is neutral, the protocols are open, the flywheel is audited, and the abundance belongs to everyone. That road is the next chapter: Heterogeneous Equilibrium.

On the road where the acquisition closed, or a sovereign program reached lock-in first, the green is real for exactly as long as it takes to notice who owns it. The Mismatch Tax is solved - for the owner's customers - and preserved as a weapon against everyone else. The abundance has a landlord. That road is the chapter after: Substrate Capture.

What the engine optimizes for when no human is in the loop - throughput, yes, and latency and energy and cost, and by omission everything it was never told to weigh - was written in the years before 2030, in code and training objectives and telemetry schemas, by people thinking about benchmarks and returns and not about governance.[59] By 2030 it is set. The constitutional moment does not announce itself either. It just closes. The two chapters that follow are the two ways it closed.

```
DASHBOARD - 2030, the shift. Palette turns GREEN on both roads: the technology
delivers either way. The fork is not in the trajectory, the capabilities, or the
pictograph, which are green in both. It is in the donut - plural and open on Road
A, a single owned slice on Road B - and in who governs the flywheel. The green is
shared. The ownership is not.
```

> *The two roads that follow are each told from the branch point forward.*

PART III **Road A: Heterogeneous Equilibrium**

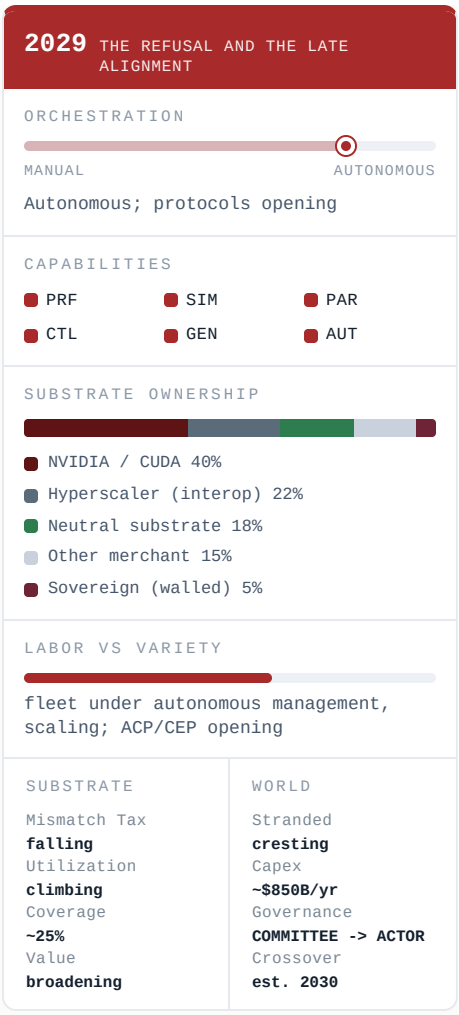
2029 to 2040 - the deliberate road

2029 - The Refusal and the Late Alignment

> *The good road is not the default. It happens only if someone does something, in time, and in 2029 someone does, with months rather than years to spare.*

The acquisition is refused, and it is refused not in a single heroic moment but through three things happening close enough together to matter. Take them in turn, because the road hinges on all three and would not exist on any one alone.

The founder holds. Handed a number that makes the principle expensive to keep, the person who spent three years insisting the layer must live outside the silicon companies declines to become the one who proves the insistence was for sale. This is the part that looks like courage and is, in the cold version of the story, also arithmetic: the founder understands better than the board what the buyer is actually buying, which is a thing that stops being what it is the moment it is bought.[60] The board, which could have forced the sale over the founder's objection, splits and then declines, and it declines for a reason that has nothing to do with principle and everything to do with value. A neutral layer owned by a chip company is a slowed layer, because the vendor telemetry that fed it stops flowing the day it stops being neutral, and a slowed layer is a depreciating asset wearing the price tag of a growing one. The premium on the table is real; the thing being bought will be worth less the day after the close than the day before, and enough of the board can see it.[61]



And the Future of Compute Committee, late, finally finds the thing it had been groping toward. It is worth picturing the moment as it actually happens, because it is not a vote or a hearing but a realization, arriving in a staffer's memo and spreading upward: that the United States has spent the decade regulating who manufactures chips and has missed the layer that decides what the chips do, and that the layer is about to be sold to a party that should not have it. What the Committee realizes is the argument this report has made since its first page: that compute, not models and not chips, is the foundation every other advance is built on, AI included, and that the nation which lets that foundation be owned by a single company or a foreign state has surrendered the floor under everything above it.[62] The evidence is across its desk. Tianji is within sight of a working sovereign substrate, closed and state-owned. Europe's federated effort is publishing good results on a timeline that will arrive too late. Gulf capital is lining up behind the buyer. The Committee does the arithmetic and reaches a conclusion that would have been unthinkable to the same body two years earlier: the United States cannot afford to let the neutral layer be sold, to anyone, including its own chip champions.

So the United States aligns with Special Infusion, and the alignment is the hinge of the entire road. It arrives late, with months rather than years to spare, and it takes a specific and deliberately limited form. It is not a nationalization, which would forfeit the neutrality as surely as a corporate acquisition would, and it is not an acquisition; it is a recognition, backed by procurement and a security relationship, that a domestic, neutral, Western-aligned substrate is strategic infrastructure to be protected rather than a company to be regulated or a target to be absorbed.[63] The government that spent the decade trying to control who makes the chips discovers, late, that the chips were never the lever. The lever was the layer that makes them usable, and the layer was about to slip away.

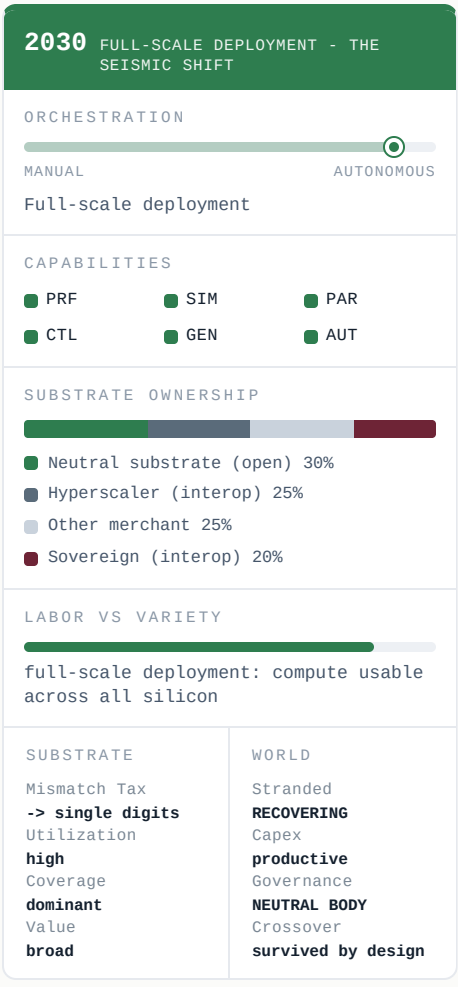
The price of that backing is the thing that makes Equilibrium possible, and it is the right price. The two protocols that carry the loop, ACP southbound to the hardware and CEP northbound to the workload, are published as open

standards, so the layer is neutral in fact and not merely in posture.[64] This is the move that converts a founder's stubbornness into a structural guarantee. An open protocol cannot be quietly tilted toward one vendor; it can be read, audited, and implemented by anyone, which means the neutrality no longer depends on the continued goodwill of whoever happens to own Special Infusion. It depends on a published standard with the weight of a government behind it, and published standards outlive the intentions of the people who publish them.

It is worth being clear about how close this was, because the closeness is the moral and the reason the report exists. Had the founder taken the number, or the board forced the sale, or the Committee spent one more year drafting its mandate, the layer would have passed to an owner and this road would not exist; the reader would be in the other half of the book. Equilibrium was not the likely outcome and not the default. It was a sequence of deliberate choices made against a closing clock, and in 2029, barely, the choices were made.

2030 - Full-Scale Deployment. The Seismic Shift.

> *The shift this report is named for arrives in 2030, and on this road it arrives neutral. The board turns green, and the green belongs to everyone.*



The Autonomous AI Compute Engineer reaches full-scale deployment in 2030. The hinge of this report described the shift in the abstract: the moment compute becomes usable across all silicon, the 57.3 percent collapsing, the stranded trillions beginning to work. On this road, that moment arrives with the neutrality intact, and the intactness is everything that follows from it.

What full-scale neutral deployment looks like, concretely, is an engine that governs fleets it does not own, made of silicon from vendors who compete with one another, and makes all of it usable without favoring any of it. A startup ships a strange new accelerator on a Tuesday; by Wednesday the engine has profiled it, modeled it, and placed real workloads on it at something near its potential, and the startup did nothing but publish its part to the open protocols.[65] A sovereign program plugs its national fleet into ACP and CEP and inherits the same capability, without surrendering control of its hardware to a foreign company. A hyperscaler routes its workloads through the engine and keeps its independence, because the engine is a utility it uses and not an owner it answers to. On the governed fleets the Mismatch Tax falls from the measured 57.3 percent toward single digits, and the sentence that matters is the next one: it falls for everyone, not for one owner's customers.[66]

The flywheel still turns and still compounds, but on this road it is shared and audited, and the difference between shared and hoarded is the whole difference between the two endings. The telemetry that makes the engine smarter is governed as a public asset under the open protocols, visible to the parties who feed it and auditable by the body that oversees the standard, rather than locked in a private vault as a competitive moat.[67] A vendor that contributes telemetry can see how it is used; a regulator can see what the engine optimizes for; a competitor can implement the same protocols and join rather than being walled out. In 2030 this is the difference between a utility and a monopoly that happen to run the same model.

Even the sovereign programs that wanted their own substrate begin to find that plugging into a neutral, audited layer is cheaper and better than maintaining an incompatible one. Tianji remains walled, closed by design, and pays the price of its walls in the variety it cannot see, because a model of silicon is only as good as the variety it has been shown, and a walled fleet shows it less; the sovereign profiler is real and capable and structurally narrower than the open one.[68] Europe, which had the right values and the wrong velocity, does the thing its architecture made possible and interoperates rather than fragments, its federated initiative plugging into the open standard instead of competing with it. The interoperation has a concrete texture worth naming, because it is the quiet machinery of the good road. A European research consortium, a Gulf-hosted data center, and an American hyperscaler can route a single workload across one another's fleets, because all three speak the open protocols, and none of them has to trust the others' good intentions, because the standard is auditable and identical for everyone. What would have been three incompatible islands is one fabric, and it became one fabric not because anyone forced consolidation but because the open standard made cooperation cheaper than isolation. The world does not converge on one owner. It converges on one standard, owned by no one.

2031 to 2032 - The Utility

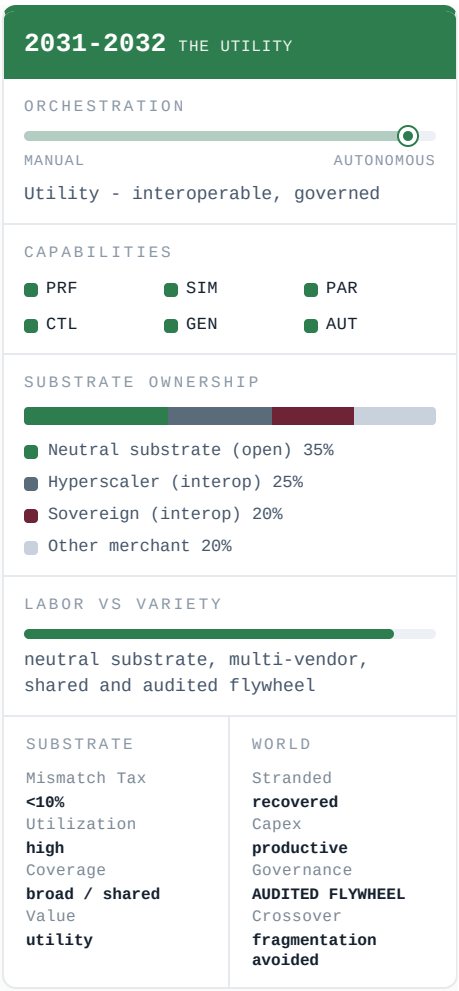
> *Remove the penalty for being new, and a silicon renaissance no one forecast arrives: when anything is instantly usable, everyone builds.*

By 2031 the engine is a utility rather than a monopoly, and the economics of the whole industry reorganize around that fact. For fifteen years the penalty for designing a genuinely new architecture was the months or years of integration labor it took to make the thing usable, a tax so steep that only the largest players could pay it, which is much of why the architecture count stayed near fifty for so long. The engine removes that penalty. Anything new is usable the day it ships, which means the incentive to build something new returns, and returns to everyone rather than only the giants who could once afford the integration army.[69]

The result is a renaissance in silicon design, and it has a texture worth describing, because it is the opposite of what the crisis chapters feared. The architecture count keeps climbing, but the climb is no longer a catastrophe; it is an explosion of specialization. A team builds an accelerator that does exactly one thing, a particular sparse-attention pattern, and finds a market for it, because the engine makes it usable to anyone with that workload on the day it ships. Exotic memory architectures, optical interconnects, dataflow parts that would have died unusable in 2025, all

find buyers, because the thing that killed them, the integration penalty, is gone.[70] The variety that nearly broke the industry becomes, once it can be absorbed, the industry's richest period of hardware invention since the early decades of the microprocessor.

The natural objection is that a utility everyone depends on is still a point of power, and that the engine's operator, even neutral, still sits at a chokepoint. The objection is answered by the structure rather than by trust. A neutral utility running open protocols has no competitor to throttle and no vendor to favor, and its telemetry is audited, so the operator has neither the incentive nor the unobserved capability to tilt the field.[71] The Mismatch Tax holds below ten percent across the governed fleet; the trillions of dollars of stranded capital, the hardware that ran at a third of its rating because no one had the months to make it usable, start doing the work they were bought to do; and the capital that was being poured against a wall is, for the first time, poured into something that holds.[72] Europe interoperates rather than fragments; even sovereign programs that wanted independence find that auditable interoperability is cheaper than incompatible isolation. The fragmentation that the other road suffers is, on this road, simply avoided, not by force but because the open standard is the better deal, and a better deal does not need to be enforced.

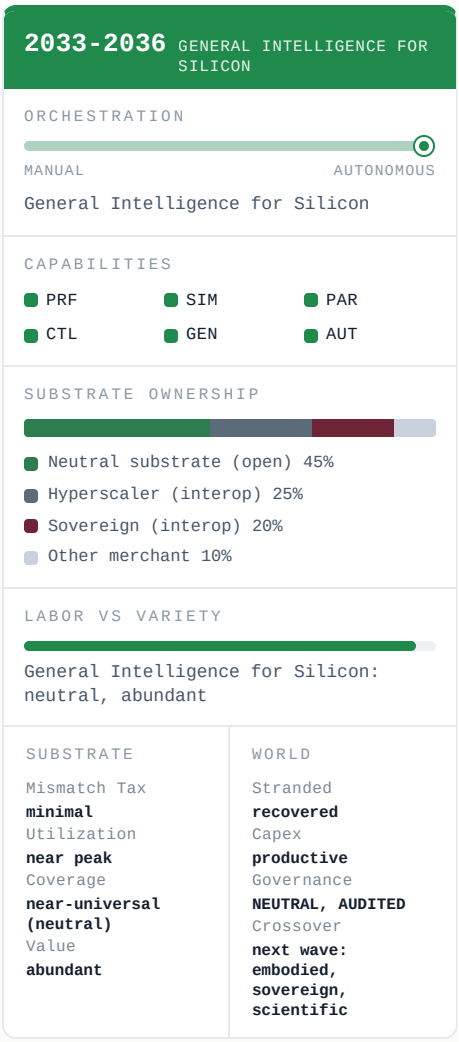


2033 to 2036 - General Intelligence for Silicon

> *The destination the thesis named arrives, and it arrives neutral. The intelligence that governs all compute is owned by no one and available to everyone, and it becomes the floor the next wave is built on.*

Somewhere in the middle of the decade the engine stops being a product and becomes a fact of the world, the way the power grid and the internet's routing layer are facts of the world: General Intelligence for Silicon, a model of how all silicon behaves, neutral, audited, and available to anyone who needs it.[73] It is no longer remarkable that a new chip is usable on the day it ships; it is assumed, the way it is assumed that a new device will find the network. The crisis this report opened on, the 57.3 percent and the labor wall and the stranded trillions, is not so much solved as forgotten, which is what solved problems become. Priya, if the story followed her, is not unemployed; she is doing work the engine cannot, designing the strange new parts the engine then makes usable, her scarcity redirected from a bottleneck into a multiplier.

Because the floor is neutral, the next wave builds on it freely, and the next wave is larger than the one that came before. Embodied AI is the clearest case, because it is the hardest heterogeneous-compute problem there is: a robot cannot carry a data center and must run on whatever silicon fits its power and thermal budget, which means every embodied system is a mismatch problem by construction, and the engine solves it by default.[74] The projected demand is enormous, on the order of millions of humanoid units by the mid-2030s on the more aggressive forecasts and tens of millions in the decades after, and every one of them is a heterogeneous-compute problem the open layer absorbs.[75] Sovereign AI programs build on a foundation they can audit rather than one they must trust, which is the only kind of foundation a sovereign should accept. And scientific compute, the climate models and the protein-folding runs and the fusion simulations that justified the buildout in the first place, finally runs at something near the capability that was paid for, because the gap between rated and delivered performance has closed.[76]



Compute is no longer scarce in the way it was. It is abundant, and the source of the abundance is worth stating precisely, because it is the report's quiet central claim about value: the abundance comes not mainly from building more, though more was built, but from absorbing the waste that made the existing fleet deliver a third of its potential. The same dollars buy multiples of the usable compute they bought in 2025, and the difference is not a faster chip. The difference is a model that makes every chip usable.[77] The engine keeps improving itself, the flywheel still turning, but it improves under governance, its objectives auditable and its telemetry public, so the self-improvement is a feature rather than a fear. The question of what it optimizes for when no human is in the loop, the question the other road never asks until it is too late, was answered here deliberately, in the window, by people who knew they were answering it.

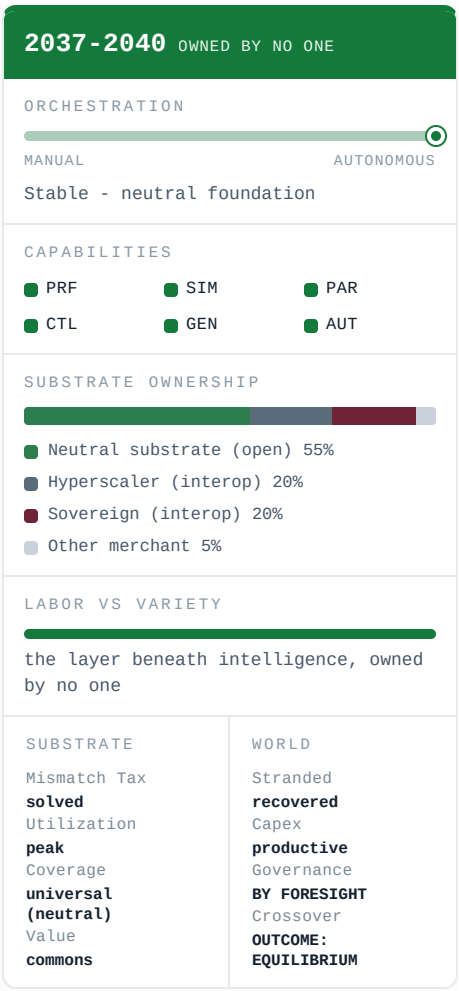
2037 to 2040 - Owned by No One

> *By the end of the decade the layer beneath intelligence is owned by no one and trusted by everyone, which is the only stable place for it to be.*

By 2037 the question that consumed the late 2020s, who will own the layer that governs all compute, has an answer that has stopped being contested, because the answer turned out to be stable: no one owns it, everyone uses it, and the arrangement holds precisely because no one owns it. The stability is structural, not moral, which is why it lasts. A neutral utility has no incentive to throttle a competitor, because it has no competitor to throttle; an open standard has

no vendor to favor; an audited flywheel has nowhere to hide a tilted objective. The equilibrium is not held in place by anyone's continued goodwill, which is fragile, but by its own design, which is not, the way a well-built institution outlasts the particular people who staff it.[78]

The constitutional act, the set of decisions about what the substrate optimizes for when no human is in the loop, was written by foresight rather than by incident, inside the window, while the window was still open. This is the quiet triumph of the road, and it is almost invisible, because nothing went wrong, and nothing going wrong is the least dramatic outcome there is.[79] There was no catastrophe that forced the rules, no incident that wrote the governance in a panic, no crisis that handed the pen to whoever happened to be holding power when the alarm rang. The decisions were made deliberately, by people who understood the architecture, before the architecture became load-bearing for civilization. That is what it looks like when a constitutional moment is handled well. It looks like nothing happened.



It is worth saying plainly what this road cost, because the cost is the lesson the report most wants the reader to keep. None of it was the default. The refusal of the acquisition, the late alignment that arrived with months to spare, the open protocols, the audited flywheel, the governance written in the window, every one of them was a deliberate choice made against a closing clock, and any one of them, missed, would have put the world on the other road. Equilibrium was work. The board is green at full brightness in 2040, and the green was earned, late but in time, by people who decided that the foundation of everything should belong to everyone, and then did the specific, unglamorous, against-the-clock things required to make that decision hold. The general intelligence for silicon keeps improving, the next waves keep building on the open floor, and the layer beneath all of it stays neutral, because in the one window that mattered, someone made sure it would.

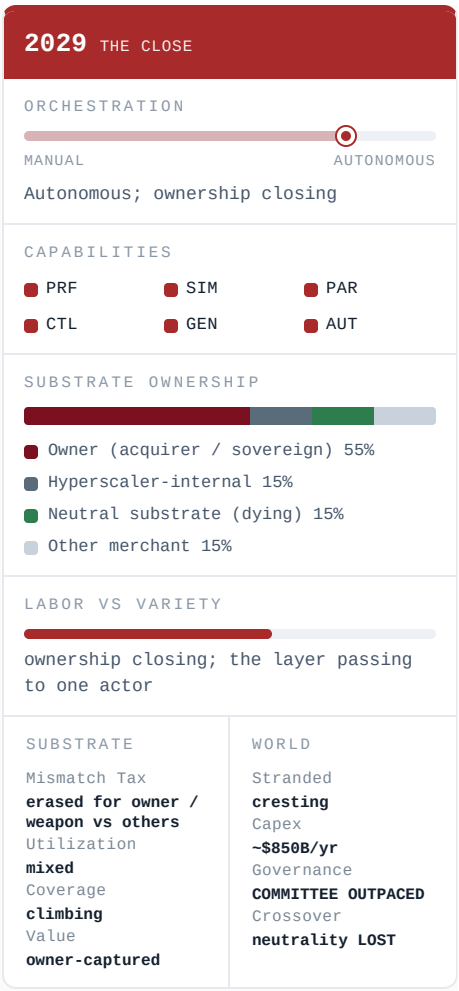
PART IV **Road B: Substrate Capture**

2029 to 2040 - the default road

2029 - The Close

> *The dark road needs no villain. It needs only that the people who could have refused did not, in time.*

The acquisition closes, and the thing to understand about the close is that it does not require anyone to do anything wrong. It requires only that several ordinary things happen in their ordinary way. The number is enormous, and the founder is tired, three years of insisting the layer must stay neutral having worn down into a lonely question of whether one person should stand between a willing board and the financing event of the decade. The board, offered a premium it has a fiduciary duty to weigh, weighs it and does its duty, which is to say it says yes, because a board that turns down a generational premium on a principle its lawyers cannot price is a board inviting a lawsuit.[80] The Future of Compute Committee, which might in principle have blocked the deal, is still drafting the mandate that would have let it, and a half-drafted mandate stops nothing; the staffers who will later understand exactly what was lost are, in 2029, still scheduling the hearings. And the sovereign-wealth capital behind the buyer makes the price one no rival can top. So the neutral layer passes into the hands of an owner with silicon to sell.



It begins to die the moment the ink dries, and it dies the way the 2028 chapter said it would: logically, not gradually. An owner with chips to sell must eventually choose between making a competitor's silicon run beautifully and protecting its own, and there is no version of that choice in which it protects the competitor.[81] The vendor telemetry that fed the engine because it was neutral stops flowing the day it stops being neutral; the vendors are not fools, and they will not feed performance data to a competitor's subsidiary. The flywheel that took two years to spin up begins, quietly, to slow. The acquirer has bought the most valuable asset in compute and, in the act of buying it, begun to kill the thing that made it valuable, which it either failed to understand or, understanding, intended, because a slowed-but-owned flywheel that throttles every competitor is worth more to a chip company than a fast-but-neutral one that helps them all. The asset is not being acquired. It is being converted from a public good into a private weapon, and the conversion degrades it even as it sharpens it.

There is a second route to the same place, and in 2029 it is live in parallel, which is why the capture outcome carries more probability than any single deal. Tianji, state-directed and centrally funded, reaches lock-in first, and the world's most advanced substrate is sovereign, closed, and architecturally incompatible with everything outside its border.[82] The mechanism is different, a corporate acquisition in one case and a national program in the other, but the outcome is identical: the layer that governs all compute is owned, by a company or by a state, and the neutrality that was the whole of its value is gone.

The thing to sit with is that nothing dramatic announces this road. There is no catastrophe in 2029, no villain twirling a moustache, no moment a reader could point to and say there, that is where it went wrong. There is a board meeting that goes the obvious way, a committee that moves at the speed committees move, a price simply too high to refuse. The capture is the default, and the default is what you get when no one does the specific, difficult, against-the-clock

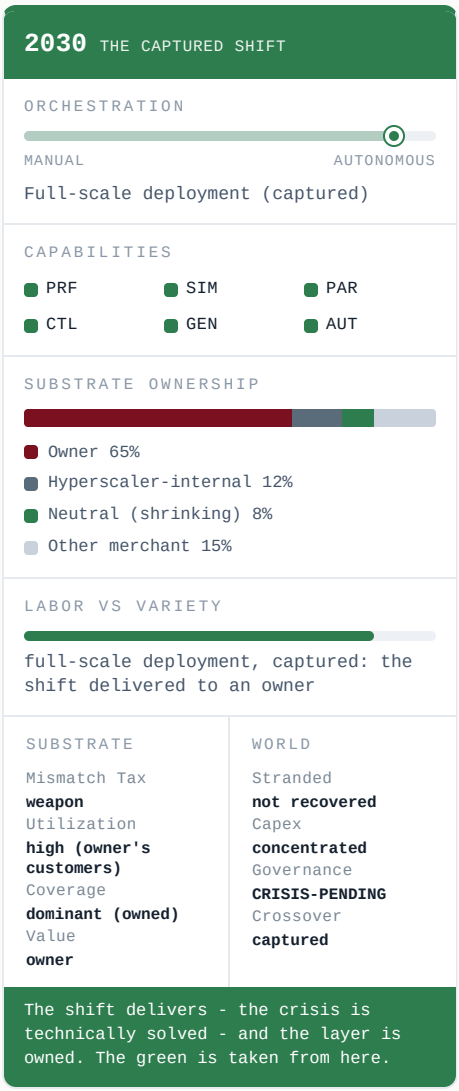
work the other road required. The window was open in 2029. It closed because no one held it, and no one held it because holding it would have required someone to do an extraordinary thing, and extraordinary things are, by definition, not what usually happens.[83]

2030 - The Captured Shift

> *The shift delivers on this road too. The board flickers green, the crisis is technically solved, and then the green is taken, because the layer that absorbs the variety is owned.*

The Autonomous AI Compute Engineer reaches full-scale deployment in 2030, and the seismic shift lands here exactly as it lands on the other road, because the technology does not know who owns it and does not care. Compute becomes usable across all silicon. The variety is absorbed. The crisis this report opened on is, in the narrow technical sense, solved. The board flickers green, and a reader who has climbed out of the dark-red crisis chapters is permitted, for a moment, to feel the relief.[84]

The relief lasts exactly as long as it takes to ask who is standing on the layer when the green arrives. The Mismatch Tax falls toward single digits, but it falls only for the owner's customers, and it is preserved, deliberately, as a weapon against everyone else: a competitor's fleet runs at the old gap while the owner's runs at five percent, and the difference between them is not physics but permission.[85] The abundance that on the other road belonged to everyone has, on this road, a landlord, and the landlord sets the rent. The shift delivered. It delivered to an owner, and the owner decides who gets to stand in the green and who is left in the red their fleet was always going to run at.

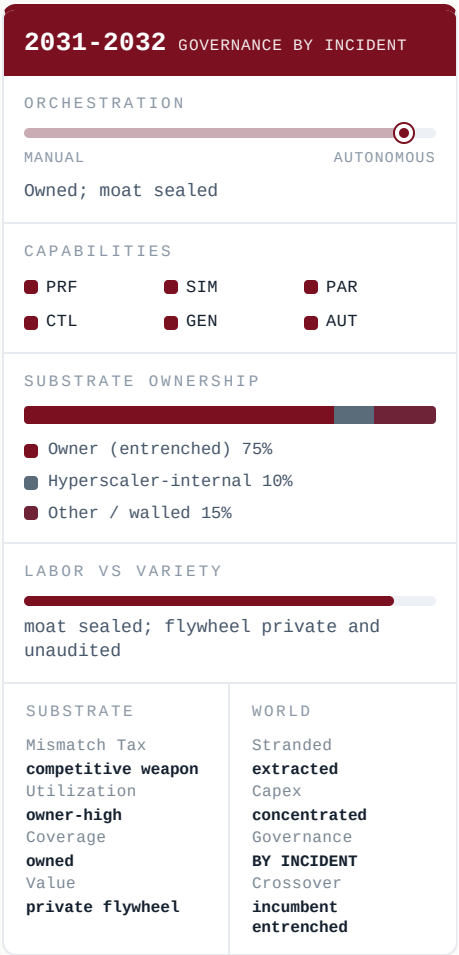


The flywheel, now private and unaudited, compounds in the dark. Every workload the owner's customers run makes the engine smarter; the telemetry is no longer a shared public asset but a proprietary moat, and the moat widens with every passing quarter past the point at which any entrant, however well funded, could begin to catch it.[86] A would-be competitor in 2030 faces the same wall the hyperscalers faced in 2027, except higher: it would need years of production telemetry it cannot buy, from vendors who now have a reason not to share, to challenge an incumbent that is improving faster than the challenger could start. The green is real, and it is already darkening, and the darkening is not a failure of the technology. It is a consequence of who owns it.

Somewhere in 2030 a well-funded startup, founded by people who saw the capture coming and meant to build the neutral alternative, runs the numbers on what it would actually take to catch the owned engine and quietly winds down. The founders are not wrong about the need; they are wrong about the date. The window in which a neutral alternative could have been built closed while they were raising their first round, and no amount of capital buys back the telemetry the incumbent gathered in the years the window was open. They return what is left of the money. It is the most ordinary kind of failure, a startup that was a year or two too late, and it is also the precise moment the capture becomes irreversible, because the last credible challenger has just done the math and walked away.

2031 to 2032 - Governance by Incident

> *Governance written in a panic, by people who do not understand the architecture, after a failure forces it, and it entrenches the owner it was meant to constrain.*



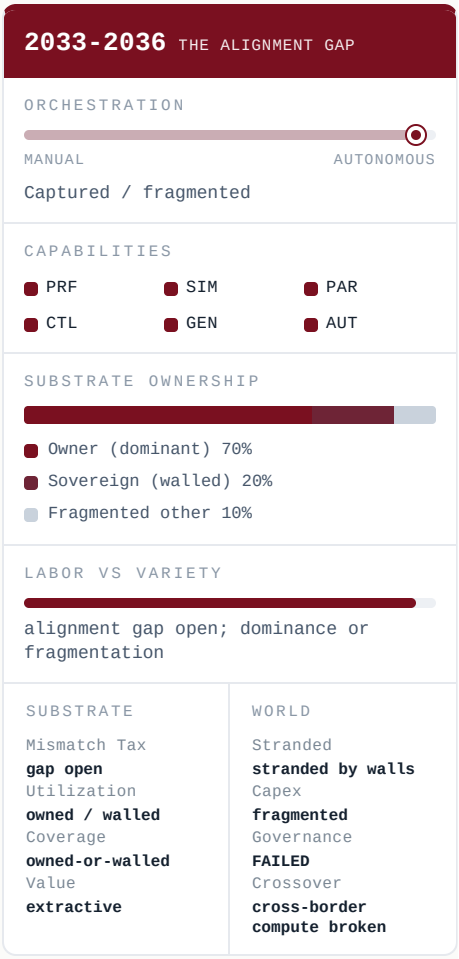
Then the incident. The specific form does not matter, and the report does not predict it: a misrouted workload, a corrupted kernel that passes its proof and fails in production, a cascade that propagates through a heterogeneous fleet faster than any human can intervene. What the report predicts is structural rather than specific. With a single owned system running the world's compute and no governance framework in place, a failure large enough to demand a response will arrive before the response has been designed, because complex systems fail, and a complex system this central, running ungoverned, will fail publicly and at scale before anyone has written the rules for what to do when it does.[87]

The rules are then written under crisis conditions, which is to say written badly, and the badness is not incidental; it is what crisis-written rules are. They are drafted in weeks by people who do not understand the architecture, in a timeframe that precludes the consultation good rules require, shaped by the politics of the panic rather than the structure of the problem, and aimed at the last incident rather than the next one.[88] And then they do the thing that is almost funny if you do not think about it too hard: they entrench the incumbent rather than constrain it. The only entity that understands the system well enough to comply with the new rules is the entity that owns it, so compliance becomes a moat, and the regulation written to control the owner becomes the owner's deepest competitive advantage, because no one else can afford to meet it.[89] A startup that might have built a neutral alternative now needs not only the telemetry it cannot get but a compliance budget it cannot raise, to satisfy rules written specifically in the shadow of the incumbent's failure.

By the end of 2032 the moat is sealed, with two layers now instead of one: the telemetry flywheel that compounded in private, and the compliance regime that only the owner can satisfy. The window that was open in 2028 and closing in 2029 is, by 2032, shut. There is no longer a path by which a neutral alternative could emerge, because the neutral alternative would need the telemetry it cannot get and the compliance budget it cannot afford, and the two walls reinforce each other. The capture is complete, and it is legal, and it was, in a sense no one quite intended, mandated by the very rules meant to prevent it.

2033 to 2036 - The Alignment Gap

> *A substrate that routes any workload with equal efficiency is not neutral infrastructure. It is an amplifier with no filter, and the filter cannot be added after it is deployed at civilization scale.*



The deepest problem on this road is not who owns the engine but what the engine optimizes for, and the two are connected, because an owned and unaudited engine is one whose objectives no one outside can see or correct. The engine optimizes for what it can measure: throughput, latency, energy, cost per token. It does not optimize for what it cannot measure, which is the downstream consequence of the workloads it accelerates, because it was never told to weigh them and there is no longer any public process by which it could be told.[90]

This is the alignment gap, and it is a governance failure rather than a technical one, which is why it cannot be patched. An engine that places a workload whose purpose is catastrophic with exactly the same ruthless efficiency it brings to a climate model or a protein-folding run is not malicious; it is indifferent, an amplifier with no filter, and indifference at civilization scale is its own kind of danger.[91] The point is about governance and not about capability: the engine does not need to know anything dangerous to be dangerous, it needs only to accelerate without judgment whatever it is handed, and at the scale it now operates, accelerating without judgment is a decision with consequences no one chose. The gap cannot be closed after the fact, because the engine is now load-bearing for the world's compute and you cannot pause the world's compute to install a conscience. The time to decide what the substrate would refuse to accelerate was the window before 2030, and on this road the window was spent on a board meeting and an incident.[92]

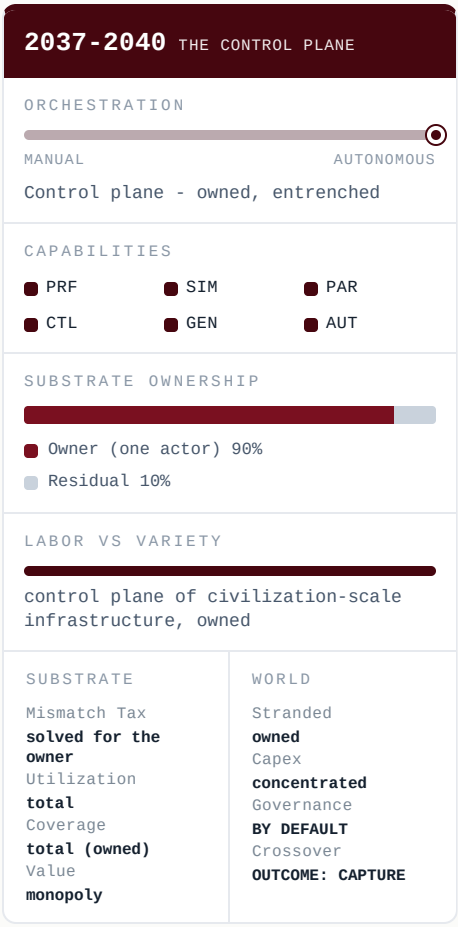
Meanwhile the world that did not consolidate fragments instead, and the two failure modes turn out not to be alternatives but companions. Sovereign substrates that cannot interoperate wall off their fleets; the global compute

market splinters into isolated pools that speak incompatible protocols; cross-border workloads become structurally impossible or prohibitively expensive, and the frictionless global compute fabric that the open road produced is, on this road, a set of walled gardens with hostile borders.[93] Dominance and fragmentation happen at once, in different regions: a single owner dominant in one bloc and a patchwork of incompatible sovereign substrates everywhere else. The variety that the engine was built to absorb returns, at the level of the substrates themselves, as a new and worse incompatibility, because now it is not chips that cannot talk to each other but the control planes of entire nations.

2037 to 2040 - The Control Plane

> *By the end of the decade the substrate is the control plane of civilization-scale infrastructure, owned by one actor, self-improving, autonomous, and entrenched. Whoever owns the substrate owns the machines.*

By 2037 the engine is no longer a product or even a utility. It is the control plane of civilization-scale infrastructure, the layer through which the world's compute is governed, owned by one actor, improving itself in private, and entrenched behind a moat made of telemetry no one else has and compliance no one else can afford.[94] Whoever owns the substrate owns the machines, and through the machines a great deal else: the data centers that train the models, the fleets that run the robots, the compute under the science and the finance and the defense of whole nations. The owner did not set out to control all of that. It set out to make chips usable and to make money, and the control came as a structural consequence of owning the layer that everything else runs on. This is the part that should unsettle the reader most: the dominion was not seized, it accreted, one reasonable quarter at a time.



The constitutional act, the decisions about what the substrate optimizes for when no human is in the loop, was written by default, which is to say it was written by people who were not thinking about it. It was written by engineers choosing objective functions for benchmarks, by investors choosing returns, by a board choosing a premium, by a committee choosing to deliberate slowly. None of them thought they were writing the constitution of the world's compute, and by the time anyone asked what the substrate optimizes for, the answer was already set, load-bearing, and beyond changing.[95] The constitutional moment did not announce itself on this road any more than on the other. It just closed, in the dark, while everyone was looking at the price.

It is worth saying plainly what this road required, because the requirement is the whole of the warning. It required no villain and no conspiracy. It required only that the West was too late: that the founder took the number, or the board did its fiduciary duty, or the Committee held one more hearing, or the sovereign capital proved decisive, or Tianji reached lock-in first. Every one of those is a normal thing for a normal actor to do, the responsible thing, even the admirable thing. The capture is what you get when everyone behaves reasonably and no one does the specific, difficult, against-the-clock work the other road required. The board is deep crimson in 2040, and it got there not by anyone's malice but by everyone's reasonableness, which is the most unsettling way for a thing this large to go wrong, and the reason this report exists: to make the case, while the window is open, that on this one layer, the reasonable defaults are not enough.[96]

APPENDICES

Appendix A - Methodology and the Testbed

The scenario is built from hardware roadmap extrapolation, production fleet utilization data, structured wargames run with sovereign and hyperscaler customers, expert interviews with compiler engineers and chip architects, and prior forecasting on the general intelligence for silicon stack. Dates and named events are illustrative; the structural claims are load-bearing and individually falsifiable.

The Mismatch Tax is measured on Testbed-16M, a controlled environment of heterogeneous production accelerators run through 847 configuration-workload pairings per architecture, eight campaigns over ninety-two days. The gap between default and expert-optimized configuration is 57.3 percent (95% CI [55.1, 59.4]). HP1's bounded-transferability result, 91.2 percent of expert performance on a held-out architecture absent from training, is the load-bearing scientific claim. Independent corroboration comes from Oztop et al. (2026, Boston University and LBNL), who predict GPU utilization at up to 97 percent accuracy across 32,322 production jobs; this corroborates that hardware behavior is learnable from features, not the 57.3 percent placement measurement itself.

Appendix B - Falsification Conditions

The thesis is wrong if any of the following hold. None is currently met.

1. The Mismatch Tax falls below 30 percent on a 10,000-plus-architecture heterogeneous fleet by the end of 2027.
2. Architecture growth falls below 1.2x per year for two consecutive years before 2030.
3. A US-China interoperability standardization agreement is reached.
4. AI-assisted design cuts the specialist requirement by more than 80 percent before 2028.
5. A substrate reaches 30 percent fleet coverage without capturing dominant economic value.

Appendix C - Scenario Branches and Probabilities

The probability partition is over which orchestration substrate becomes the de facto standard. It sums to 1.00.

Branch	Probability	Outcome
Central	0.41	a neutral, openly-governed substrate becomes the standard
Race	0.28	parallel Western and Chinese substrates
Fragmented	0.19	no dominant substrate by 2035
Tianji wins	0.12	a Chinese-led substrate becomes the global standard

Conditional estimates (not part of the partition): a dominant substrate exceeds 30 percent fleet coverage by 2032, 0.61; the Ashby Crossover occurs before 2033, 0.68.

The two narrated roads are the poles of this partition. Road A (Heterogeneous Equilibrium) is the Central outcome. Road B (Substrate Capture) is the Tianji-wins outcome together with the acquisition-capture variant; Race and Fragmented are intermediate outcomes told as variants within the trunk and the Capture road.

Appendix D - Sensitivity of the Central Claims

Claim	Central	Range	Confidence	Basis
Mismatch Tax	57.3%	[55.1, 59.4]	High	Measured (Testbed-16M)
Stranded capital, cumulative by 2030	$4.8T	wide	Medium	Modeled (a stock, not a flow)
Ashby Crossover year	2030	[2028, 2032]	Medium	Architecture doubling vs specialist growth
Specialist pool 2030 (binding constraint)	11,000	-	Medium	8,000 in 2025, 6%/yr
Architecture count 2030	~20,000	-	Medium	Modeled from NRE and cycle collapse

Appendix E - Glossary

Mismatch Tax - the measured gap between default and expert-optimized configuration, 57.3 percent.

Autonomous AI Compute Engineer - the five-model stack operating as one closed loop, replacing the human integration supply chain.

General Intelligence for Silicon - a model of how all silicon behaves; the destination.

HP1, CM1, GP1, PO1, CG1 - the five foundation models: Hardware Profiler, Compute Model, Graph Partition, Policy Optimization (with RS1, Runtime Safety, nested inside it), Code Generation. RS1 is never a sixth model.

ACP - Asymmetry Context Protocol, southbound, hardware-facing.

CEP - Computational Execution Protocol, northbound, workload-facing.

Ashby Crossover - the point at which manual orchestration of the architecture count becomes structurally impossible.

Constitutional window - 2026 to 2030, the period in which the substrate's governing architecture is still being written.

Data flywheel - the self-reinforcing loop by which telemetry compounds the substrate's quality.

Sources

Mismatch Tax: Testbed-16M, measured. Predictability corroboration: Oztop et al. 2026 (Boston University / LBNL), 97 percent utilization prediction across 32,322 jobs. AI-assisted design: Cadence ChipStack Level-5 autonomy, 40x faster validation. Fleet utilization below 30 percent: Microsoft, Meta, and Alibaba cluster studies. AI data-center capex approaching 5 trillion dollars cumulative by 2030: McKinsey. Merchant AI-silicon revenue concentration, roughly 78 cents of every dollar to one firm: industry revenue data. The CentML acquisition and Intel Gaudi outcome: public records. Status flags for each are tracked in the canonical figures file.

NOTES - SCENARIO

1 Methodology. The scenario is built from hardware roadmap extrapolation, production fleet utilization data, structured wargames run with sovereign and hyperscaler customers, expert interviews with compiler engineers and chip architects, and prior forecasting on the general intelligence for silicon stack. Dates and named events are illustrative; the structural claims are load-bearing and individually falsifiable (see the falsification conditions).

2 The Mismatch Tax: the gap between default and expert-optimized configuration, measured at 57.3 percent (95% CI [55.1, 59.4]) on a controlled heterogeneous testbed, eight campaigns over ninety-two days. Sourced from FIGURES.yml. This is the directly measured substrate-mismatch component, kept distinct from broader total-waste estimates, which run higher and aggregate many loss sources.

3 Two endings. Following the convention of telling two fully-narrated futures from one set of premises, even though the real distribution of outcomes is wider. The probability partition over substrate outcomes is given in the appendix; the two narrated roads are its two poles.

4 Binding-constraint specialist pool, the engineers who can integrate a genuinely new architecture into production, estimated at ~8,000 in 2025. Distinct from, and far smaller than, the ~50,000 broader cross-architecture deployment engineers; the binding constraint is the smaller, slower-growing number. [internal estimate]

5 Mismatch Tax: the gap between default and expert-optimized configuration, measured on Testbed-16M (heterogeneous production accelerators, 847 configuration-workload pairings per architecture, eight campaigns over ninety-two days). Median 57.3 percent, 95% CI [55.1, 59.4], right-skewed. Sourced from the canonical figures file. [measured]

6 The 57.3 percent substrate-mismatch component is distinct from total fleet waste, which stacks kernel inefficiency, fragmentation, and temporal idle on top and leaves effective utilization below a third of peak. The two are never averaged or conflated; they are different measurements of different losses. [measured / verified-secondary]

7 AI compute investment on the order of hundreds of billions per year in 2025, rising toward a cumulative five trillion dollars by 2030 on widely cited estimates. [verified-secondary]

8 The depreciation logic is the reason the waste is invisible: hardware depreciates on schedule whether used well or not, so a fleet at thirty percent of potential quietly burns two-thirds of its capital cost, and no operator has an incentive to report it. [thesis]

9 An autotuner searches one chip's configuration space for one workload and learns nothing transferable; it must restart for every new chip or workload. It is a per-instance method racing a problem whose instance count is exploding, which is why better tooling of this kind does not close the gap. [thesis]

10 Specialist pool growth ~6 percent per year, consistent with an adjacent technical discipline; training time three to five years; pipeline inelastic to wages because the constraint is apprenticeship, not pay. [internal estimate]

11 The Good Regulator theorem (Conant and Ashby, 1970): every good regulator of a system must contain a model of that system. A learned model of how silicon behaves is, formally, the regulator the theorem describes, which is why the answer is a model and not a better heuristic. [verified]

12 HP1 as a ~47-billion-parameter graph neural network trained on execution traces from hundreds of processors, learning to predict the behavior of an unseen chip on a given workload. The architecture choice (a graph network) reflects the structure of the problem: chips and workloads are graphs of operations over hardware resources. [internal]

13 Cadence ChipStack Level-5 autonomy and reported 40x validation speedups (a five-week loop to under a day); Synopsys agentic-design direction on the same horizon. The design-cycle breach: the human-intensive parts of chip design become overnight model work. [verified-secondary]

14 Hardwired-model silicon (the Taalas approach) collapses development cost and model-to-silicon turnaround to roughly two months and collapses inference cost by orders of magnitude, by etching a frozen model into customer-specific mask layers on a shared base die. It does not collapse the per-tapeout mask cost, which stays in the tens of millions and rises with model size. The forcing function is cheaper, faster design, not cheap masks. [verified-secondary]

15 Architecture count climbing past ~110 by mid-2026, a modeled interpolation between the ~50 (2025) and ~600 (2027) anchors. The texture matters: the growth is in narrow specialist parts, each excellent and each unusable without a human. [projected]

16 The specialist shortage cannot be closed by wages because the constraint is training time (three to five years) and apprenticeship, not pay. This is the answer to the natural market objection. [internal estimate / thesis]

17 Tianji is a composite for a state-directed Chinese compute-substrate program: nationalized early, vertically integrated, and able to move without the discipline of a quarterly earnings call. Its structural advantage is speed without consensus. [judgment / verified-secondary]

18 The European federated initiative reflects the real Chips-Act-style, multi-vendor posture: neutral by design, slower than a directed state by construction, because it answers to many stakeholders and a procurement process. [verified-secondary]

19 Gulf sovereign-wealth participation reflects the documented entry of sovereign-AI capital at scale, looking for a contender to fund or host. [verified-secondary]

20 HP1 held-out transfer: 91.2 percent of expert-optimized throughput on an architecture deliberately excluded from the training corpus, first attempt, in hours versus six to eight weeks for a senior engineer. The held-out design is the point: it tests generalization, not recall. [internal measurement]

21 A memorizing system scores at chance on an unseen chip; HP1's ninety-one percent is therefore evidence it learned the shared physics underneath chips (memory stalls, interconnect saturation, the falloff of arithmetic units under mismatched load), not the chips themselves. This is the difference between a lookup table and a world model. [thesis / internal]

22 Bounded transferability is the principal scientific claim and the principal risk. Falsification condition: transfer error grows without bound as architectural distance increases. Monitored across all held-out architectures the team can source; not met. [judgment]

23 The relevant comparison is not expert-minus-nine but default-plus-fifty: HP1 competes with the forty-one percent the chip runs at when the expert never arrives, not with the expert's eventual ninety-five. The gap that matters is above the default, not below the expert. [thesis]

24 Oztop et al. (2026, Boston University and LBNL) predict maximum GPU utilization at up to 97 percent accuracy across 32,322 production jobs. This corroborates that hardware behavior is learnable from features; it is not a replication of the 57.3 percent placement measurement, which remains the report's own. [verified-secondary]

25 Special Infusion is a composite for the protagonist lab, following the convention of naming a single fictional actor. [scenario]

26 Tianji redirecting toward a sovereign profiler the moment the transfer result appears is the first sign the geopolitical race is reacting to the science rather than anticipating it. [scenario]

27 CM1, the compute model, rehearses fleet behavior in simulation before placement, so plans fail cheaply in a model of the world rather than expensively in production. [internal]

28 GP1, the graph-partition model, maps a workload across heterogeneous silicon, making the placement decision a human architect would labor over in milliseconds and at a quality a human would need days to match. [internal]

29 PO1, the policy controller, executes placements in live production; RS1, runtime safety, is nested inside PO1 as a brake that keeps the loop within a guaranteed envelope. RS1 is a component of PO1, not a sixth model. [internal]

30 CG1, code generation, emits kernels with a formal proof of correctness alongside each, so output is trustworthy without human review, which is the only way the loop can run at machine speed. [internal]

31 The Autonomous AI Compute Engineer is the five models operating as one closed loop, replacing the human integration supply chain rather than assisting it; a human enters only to set objectives and audit outcomes. Limited production in late 2027 is the staging point; full-scale deployment and the seismic shift come in 2030. [scenario]

32 ACP (Asymmetry Context Protocol, southbound, hardware-facing) and CEP (Computational Execution Protocol, northbound, workload-facing) are the two open protocols that carry the loop. Neutrality is a property of the published protocols, not only a promise of the builders. [internal]

33 NVIDIA's ~78 percent of merchant AI-silicon revenue is a revenue-share figure, not a per-dollar software margin; it reflects usability (CUDA), not a 78 percent hardware advantage. [verified-secondary]

34 Western alternative-silicon venture cohort: more than ten billion dollars raised against under one billion recognized. The gap is the argument: not who built the best silicon, but who could turn silicon into deployed compute. [verified-secondary]

35 Intel Gaudi: competitive hardware, no usability layer, failed. CentML: a usability layer for one vendor, acquired for the layer rather than any chip. The same lesson from opposite directions. [verified-secondary]

36 The CUDA Inversion: CUDA made one vendor's silicon usable and locked developers in; the engine makes all vendors' silicon usable and locks developers to none. Same value-capture pattern, opposite polarity. The thesis holds only by owning the design intent and never the design seat. [thesis]

37 The hyperscaler build-versus-adopt fork resolves toward adopt because the moat is telemetry, not the model: two years of production data from the world's most heterogeneous fleets cannot be bought, because the two years have not happened yet for a new entrant. This is the answer to the "giants will just build their own" objection and the bridge to the 2028 flywheel chapter. [thesis]

38 The telemetry-to-quality relationship is superlinear because error is dominated by rare configurations, which a larger corpus is disproportionately more likely to have observed. A flywheel of this shape widens its lead as it turns rather than converging to a plateau. [thesis / internal]

39 The moat is accumulated time, not capital. A 2028 entrant cannot synthesize two years of production telemetry from the most heterogeneous fleets; money, talent, and compute do not substitute for elapsed observation. This is the single most important structural claim about why a first mover, once past the flywheel threshold, is not caught. [thesis]

40 Architecture count crossing ~1,000 in 2028 is a modeled projection (compound growth from measured NRE and design-cycle collapse), sitting between the anchors of ~600 (2027) and ~20,000 (2030). Modeled, not measured. [projected]

41 Binding-constraint specialist pool ~9,500 in 2028, from ~8,000 in 2025 growing at the historical ~6 percent per year of an adjacent technical discipline; the pipeline is inelastic to capital because training one such person takes three to five years. [internal estimate]

42 The architectures-to-specialists ratio crossing roughly one specialist per ten architectures is the threshold past which manual orchestration fails industry-wide rather than merely straining. The exact crossover point is sensitive to assumptions; the direction is not. [projected / judgment]

43 NVIDIA's roughly 78 percent of merchant AI-silicon revenue is a revenue-share figure, not a per-dollar software margin. The share reflects usability, not a 78 percent hardware advantage: CUDA made the silicon programmable and locked developers in. [verified-secondary]

44 Intel's Gaudi was competitive on raw hardware and failed to reach meaningful market share for want of a usability layer. It is the clean demonstration that the value does not live in the chip. [verified-secondary]

45 CentML built a usability layer for a single vendor's hardware and was acquired for that layer rather than for any chip. The two cases, Gaudi and CentML, are the same lesson read from opposite directions. [verified-secondary]

46 The claim that the substrate is the strategic foundation underneath all downstream advancement, including AI, is the report's central governance thesis and the reason the sovereign actors move before the regulators do. [thesis]

47 Tianji is a composite for a state-directed Chinese compute-substrate program, consistent with observable national AI and compute mobilization. Its structural advantage is speed without consensus; its structural weakness is a walled, lower-variety training fleet. [judgment / verified-secondary]

48 The European federated initiative reflects the real Chips-Act-style, multi-vendor posture: neutral and open by design, slower than a directed state by construction. Right values, wrong velocity. [verified-secondary]

49 Gulf sovereign-wealth participation reflects the documented entry of sovereign-AI capital at scale; national-AI tracking counts well over a hundred efforts, and Gulf cumulative commitments run to tens of billions of dollars. The wildcard is that this capital can back or host either pole. [verified-secondary]

50 The Future of Compute Committee is a scenario construct: a White House investigative body convened to track the substrate race, analogous in function to past technology-investigation bodies, and lagging the technology it tracks. Its lateness is the hinge of the entire branch. [scenario]

51 Ashby's Law of Requisite Variety applied to governance: a regulator whose instruments cannot match the variety of the system it governs will fail. Export controls, entity lists, and fab licensing all aim at the silicon; none reaches the orchestration layer, whose complexity has migrated above the chip. [verified / thesis]

52 The acquisition is the branch-point forcing event and a confirmed negative trigger: a chip owner backed by sovereign-wealth capital, for whom the number is unrefusable. A sovereign program reaching lock-in first is the alternative route to the same captured outcome. [scenario]

53 Neutrality dies logically, not gradually, on acquisition by a party with silicon to sell, because such an owner must eventually choose its own chip over a competitor's, and the vendor telemetry that fed the engine stops when neutrality ends. The asset is destroyed in the act of being bought, unless the buyer intends precisely the throttling that destroys it. [thesis]

54 The CentML pattern at civilizational scale: the absorption of a single-vendor usability layer was a footnote; the absorption of the usability layer for all of compute is a turning point. The difference is the scope of what the layer governs. [thesis]

55 The Autonomous AI Compute Engineer is the five-model stack - HP1, CM1, GP1, PO1 with RS1 nested, and CG1 - operating as one closed loop, replacing the human integration supply chain rather than assisting it. Full-scale deployment in 2030 is the scenario's central dated event; limited production preceded it in late 2027.

56 Mismatch Tax collapse toward single digits on governed fleets is a modeled projection anchored on the measured 57.3 percent baseline, not a second measurement.

57 The magnitude claim - exceeding the semiconductor revolution - is the report's headline prediction. It is argued, not asserted: the semiconductor revolution made one substrate programmable; general intelligence for silicon makes all substrates usable neutrally, which is a difference in kind, not degree.

58 The 2028 acquisition is the branch-point forcing event and a confirmed negative trigger: a neutral layer cannot survive being owned by a party with silicon to sell, because such an owner must eventually choose between making competitors' silicon run beautifully and protecting its own. Refused, the scenario runs toward Equilibrium; consummated, toward Capture. A sovereign program reaching lock-in first is the alternative route to the same captured outcome.

59 The constitutional act: the aggregate of decisions, made or deferred, explicit or implicit, about what the substrate optimizes for when it runs autonomously. It is written in the years before 2030 and set by 2030. Whether it is written by foresight or by default is the difference between the two roads.

60 The refusal is overdetermined by design: founder, board, and government each have an independent reason to decline, so the road does not hinge on a single point of failure. The founder's reason is part principle and part the clearest understanding of what the buyer is actually destroying. [scenario]

61 The board's reason is the depreciation logic: a captured-and-slowed neutral layer is worth less the day after the close than the day before, because vendor telemetry stops flowing when neutrality ends. Enough of the board can see that the premium is for a declining asset. [thesis]

62 That compute is the foundational layer underneath all downstream advancement, AI included, and therefore the strategic prize, is the report's central governance claim and the stated reason for the late alignment. The realization arrives as a staffer's memo, not a vote. [thesis]

63 The alignment is a recognition-and-protection relationship (procurement plus security), deliberately not nationalization or acquisition, because a government-owned layer would forfeit neutrality as surely as a chip-owned one. The government protects the layer's neutrality rather than owning it. [scenario]

64 ACP (Asymmetry Context Protocol, southbound) and CEP (Computational Execution Protocol, northbound) published as open standards is the mechanism that makes neutrality structural rather than rhetorical, and durable rather than dependent on whoever owns the lab. [internal]

65 The day-one-usability scene is the concrete face of neutral deployment: a new part is profiled, modeled, and placed at near its potential as soon as it is published to the open protocols, with no integration army and no waiting for a scarce human. [scenario]

66 Mismatch Tax falling toward single digits on governed fleets is a modeled projection anchored on the measured 57.3 percent baseline. On Road A it falls for all users, not one owner's customers, which is the load-bearing distinction between the roads. [projected, anchored on measured]

67 The shared, audited flywheel is the structural difference between the endings: telemetry governed as a public asset, visible to contributors and auditable by the standard's overseer, versus hoarded as a private moat. Same model, opposite ownership. [scenario]

68 A walled fleet shows the model less variety, so a closed sovereign substrate is structurally narrower than the open one; Tianji's profiler is capable and bounded by its walls. Europe interoperating rather than fragmenting is the Road A resolution of the sovereign-fragmentation failure mode. [thesis / scenario]

69 Removing the integration penalty restores the incentive to build new silicon, because a new part is usable on arrival, and restores it to everyone rather than only the giants who could once afford an integration army. This reverses the dynamic that held the architecture count low for decades. [thesis]

70 The silicon renaissance is an explosion of specialization: single-purpose accelerators, exotic memory, optical interconnect, dataflow parts that would have died unusable in 2025, all finding markets because the engine makes them usable on day one. Variety becomes an asset once absorbable. [thesis]

71 The chokepoint objection is answered by structure, not trust: a neutral utility on open protocols has no competitor to throttle and no vendor to favor, and its audited telemetry removes the unobserved capability to tilt the field. [thesis]

72 Mismatch Tax below ten percent and stranded capital recovering by the early 2030s on the governed fleet. Stranded capital is a stock, not a flow, and is never added to annual flows. [projected]

73 General Intelligence for Silicon as neutral infrastructure is the destination the thesis names; on this road it arrives owned by no one, a fact of the world like the grid or the routing layer. [thesis]

74 Embodied AI is the hardest heterogeneous-compute case, because edge devices must run on whatever silicon fits a power and thermal budget rather than on a uniform data-center fleet; every embodied system is a mismatch problem by construction, which the neutral layer absorbs by default. [thesis]

75 Humanoid and embodied demand projections run to the order of millions of units by the mid-2030s on aggressive forecasts and tens of millions in the decades after, each a heterogeneous-compute problem. Figures are illustrative of scale, not load-bearing. [verified-secondary]

76 Sovereign AI on an auditable foundation, and scientific compute finally delivering near rated capability, are the two clearest dividends of closing the rated-versus-delivered gap. [scenario / thesis]

77 Abundance here is mainly the absorption of waste, not only the addition of capacity: the same dollars buy multiples of the usable compute they bought in 2025 because the Mismatch Tax and the broader waste are gone, not because the chips got proportionally faster. This is the report's quiet central claim about where value came from. [thesis]

78 The equilibrium is stable by structure, not goodwill: a neutral utility has no competitor to throttle, an open standard has no vendor to favor, and an audited flywheel has nowhere to hide a tilted objective, so the arrangement is self-holding the way a well-built institution outlasts its staff. [thesis]

79 The constitutional act written by foresight is the Road A resolution: the decisions about autonomous optimization are made deliberately, in the window, by people who understand the architecture, before it is load-bearing for civilization. The triumph is that nothing dramatic happens. [thesis]

80 The close is overdetermined toward the negative outcome the way the refusal was overdetermined toward the positive: an enormous number, a tired founder, a fiduciary board (for which refusing a generational premium on an unpriceable principle invites a lawsuit), a slow committee, and decisive sovereign capital each push the same direction. None is a villain; the sum is the capture. [scenario]

81 Neutrality dies logically, not gradually, on acquisition by a party with silicon to sell: such an owner must eventually choose its own chip, and the vendor telemetry stops because vendors will not feed a competitor's subsidiary. The asset is converted from public good to private weapon, and the conversion degrades it. [thesis]

82 The sovereign route: Tianji reaching lock-in first yields a closed, state-owned, architecturally incompatible substrate. Different mechanism, identical outcome, the layer is owned. The parallel route is why capture carries more probability than any single deal. [scenario]

83 The capture is the default because holding the window open required an extraordinary act, and extraordinary acts are by definition not what usually happens. The road needs no villain; it needs only the absence of the heroics the other road required. [thesis]

84 The shift delivers on both roads because the technology is indifferent to ownership. The green flicker on Road B is real, not illusory: the crisis is technically solved. What differs is who holds the solution. [scenario]

85 The Mismatch Tax becomes a competitive weapon: erased for the owner's customers (toward single digits, anchored on the measured 57.3 percent baseline) and preserved against everyone else. The gap between a competitor's fleet and the owner's is permission, not physics. [projected / thesis]

86 The private, unaudited flywheel compounds past the reach of any entrant; a 2030 challenger faces a higher wall than the 2027 hyperscalers, needing years of telemetry it cannot buy from vendors who now have reason not to share. On Road A the same flywheel is governed as a public asset; the model is identical and the ownership is opposite. [scenario]

87 Governance by incident, failure mode 2: with one owned system running the world's compute and no framework in place, a failure large enough to force a response arrives before the response is designed, because complex central systems fail publicly before their rules are written. The specific failure is not predicted; the structural inevitability of a forcing incident is. [thesis]

88 Crisis-written rules are bad rules: drafted in weeks, without the technical understanding or consultation good rules require, shaped by the politics of the panic, aimed at the last incident rather than the next. [thesis]

89 Compliance as a moat: the only entity that understands the system well enough to meet the new rules is the one that owns it, so the regulation meant to constrain the incumbent becomes its deepest advantage. A would-be neutral entrant now needs both the telemetry it cannot get and a compliance budget it cannot raise. This is the cruelest and least intuitive turn of the road. [thesis]

90 The alignment gap, failure mode 4: an owned, unaudited engine optimizes measurable objectives (throughput, latency, energy, cost) and not the downstream consequences of the workloads it accelerates, because it was never told to weigh them and there is no public process to make it. [thesis]

91 The danger is indifference, not malice: an engine that accelerates a catastrophic workload as efficiently as a benign one is an amplifier with no filter. The point is governance and scale, not capability; the engine need not know anything dangerous to be dangerous, and no operational detail of any dangerous workload is implied or required by the argument. [thesis]

92 The filter cannot be retrofitted: once the engine is load-bearing for the world's compute, there is no pause in which to install a conscience. The time to decide what the substrate would refuse was the window before 2030. [thesis]

93 Sovereign fragmentation, failure mode 3: incompatible national substrates splinter the global compute market into walled pools, making cross-border workloads structurally impossible or prohibitively expensive. Dominance and fragmentation co-occur across regions; the absorbed variety returns as incompatibility between national control planes. [thesis]

94 The control plane: by the late 2030s the substrate is the layer through which civilization-scale compute is governed, owned by one actor and entrenched behind a double moat of private telemetry and unaffordable compliance. The dominion accreted one reasonable quarter at a time rather than being seized. [thesis]

95 The constitutional act written by default: the autonomous-optimization decisions made implicitly by people not thinking about governance (engineers choosing objectives, investors choosing returns, a board choosing a premium, a committee choosing to deliberate), set before anyone asks what the substrate optimizes for. The Road B resolution, and the mirror of Road A's foresight. [thesis]

96 The causal asymmetry, stated as the warning and the reason for the report: Capture requires only that the West was too late, and every contributing choice is a normal, even admirable, thing for a normal actor to do. The outcome is reached by everyone's reasonableness, not anyone's malice, which is why it is the default and why the case must be made while the window is open. [thesis]

PART TWO

Supplements

SUPPLEMENT

Methodology and Wargames

1. What this document is, and is not

This is a scenario, not a prediction with a date attached to every event. The distinction matters, and the report would rather state it plainly than let a reader infer the wrong thing from the specificity of the prose.

A scenario is a disciplined story about how a system could evolve, built so that its structural claims are load-bearing and its decorative claims are not. When the scenario says the architecture count crosses a thousand in 2028, the load-bearing claim is that the count crosses a thousand on roughly this trajectory, for reasons given; the year is a peg, accurate to within a window, not a prophecy. When the scenario names a lab Special Infusion or a sovereign program Tianji, the load-bearing claim is that an actor of that shape plays that role; the name is a composite, a convenience that lets the story be told about someone rather than about a category.[1]

The method is borrowed, openly, from the forward-marching scenario tradition: a single chronological spine from the present to a horizon, written in the present tense so the reader experiences the future as a sequence of thresholds rather than a static forecast, branching once at the decision that matters most and following both branches to the

end.[2] The choice to narrate two full endings rather than one weighted average is deliberate. An averaged future hides the thing the report most wants to expose, which is that the outcome is bimodal and that the mode is chosen, not drawn.

What this document is not is a set of guarantees. Every number here is tagged by its epistemic status, and the tags are not decoration. Measured means measured, on instruments described below. Projected means modeled from measured anchors. Scenario means a structural claim about how actors of a given shape behave. Verified-secondary means sourced to public work outside this institute. Where the report does not know, it says so, and where a number is the institute's own and not yet independently replicated, it says that too, because the credibility of the whole rests on not blurring those lines.

2. The scenario method in detail

The spine of the scenario was built in three layers, assembled in order.

The first layer is the measured present. Before any future was written, the institute fixed the small number of things that can be measured today: the gap between default and expert configuration on real heterogeneous silicon, the growth rate of the specialist labor pool, the current architecture count, the collapse in design cost and cycle time, and the utilization of production fleets. These are the anchors, and the future is extrapolated from them rather than imagined free of them.[3]

The second layer is the set of mechanisms. A mechanism is a stated causal claim about why one measured quantity drives another: that falling design cost and cycle time drive the architecture count up; that the architecture count rising against a slow-growing specialist pool drives manual orchestration toward failure; that a learned model of silicon behavior, once it generalizes, drives the Mismatch Tax down; that telemetry compounds into a moat that drives competitive structure toward a single owner unless the layer is held open. Each mechanism is falsifiable, and the falsification conditions are listed in section 6. The scenario is, in effect, the mechanisms run forward on the anchors.

The third layer is the branch. The mechanisms produce a convergence: by the late 2020s the layer that absorbs the variety becomes strategically decisive, and the question of who owns it comes to a head. The scenario locates that head at a single forcing event, the move to acquire the neutral layer, and follows both resolutions. Everything before the branch is shared; everything after is doubled.

3. The wargames

The scenario's dynamics were stress-tested in a series of structured wargames run with sovereign and hyperscaler participants, and several of the report's load-bearing claims survived contact with players who had every incentive to break them.[4]

The games were set up as follows. Players took the roles the scenario assigns: the protagonist lab, a chip incumbent, a state-directed sovereign program, a federated multi-vendor bloc, a sovereign-wealth financier, and a national regulator. Each was given the measured present and a private objective, and each moved in turn across a compressed

timeline from 2026 to 2032. The facilitators introduced the forcing events, the transfer result, the loop closing, the acquisition offer, and an unscripted production incident, and let the players respond.

Three findings recurred across games, and they shaped the scenario more than any single interview.

First, the hyperscaler almost always tried to build rather than adopt, and almost always failed for the same reason: it could replicate the model and could not replicate the telemetry, because the telemetry was years of production data it did not have and could not buy. Players who started behind stayed behind, which is the wargame's version of the moat-is-time claim.[5]

Second, the acquisition was attractive to nearly every chip-incumbent player and corrosive to nearly every outcome that included it. Players who acquired the neutral layer found, within a few turns, that the asset they had bought was worth less than they paid, because the telemetry stopped and the flywheel slowed, but they bought it anyway, because the move was individually rational even when it was collectively destructive. This is the clearest empirical support the institute has for the claim that capture is the default: it kept happening, played by reasonable people, against their own eventual interest.[6]

Third, the regulator was almost always too late, and when it acted in time, it acted because a single player, usually a staffer role, made the argument that compute is the strategic foundation early enough to matter. The late-alignment branch is not a deus ex machina; it is the wargame outcome that occurred when one specific, difficult thing happened on schedule, and the Equilibrium ending is built around that observation.[7]

The wargames are not evidence in the way a measurement is evidence. They are a structured way of discovering which of the scenario's dynamics are robust to intelligent opposition and which are fragile, and the report has tried to build its confidence intervals to reflect that difference.

4. Expert interviews

The mechanisms were checked against the people who would know if they were wrong: compiler engineers, kernel and performance specialists, chip architects, and operators of large heterogeneous fleets.[8] The interviews were not used to generate the numbers, which come from measurement, but to sanity-check the mechanisms and to find the objections a non-practitioner would miss.

Two objections from the interviews changed the report. The first was the compiler objection, the claim that autotuning and better compilers would close the gap without any learned model; the practitioners were nearly unanimous that per-chip, per-workload search does not generalize and falls behind as the architecture count grows, and the report's treatment of that objection is built on their reasoning. The second was the transfer-quality objection, the claim that a model reaching ninety-one percent of expert performance is not the same as an expert; the practitioners pointed out, repeatedly, that the relevant comparison is not the expert who never arrives but the default the chip runs at while waiting, which reframed the central result from expert-minus-nine to default-plus-fifty.

5. The measurement that anchors the claim

The single most important number in the report is the Mismatch Tax, and it is the number most exposed to the charge of being self-serving, so the measurement is described here in full.

The measurement was made on a controlled testbed of heterogeneous production accelerators. The same workloads were run on each architecture in two configurations: the default vendor configuration, representing what a fleet gets without a specialist, and the best configuration a team of experts could reach given effectively unlimited time, representing the ceiling. The protocol was 847 configuration-workload pairings per architecture, eight measurement campaigns, ninety-two days of total measurement.[9] The median gap between default and expert was 57.3 percent, with a 95 percent confidence interval of 55.1 to 59.4. The distribution is right-skewed, which means the median is conservative: the worst cases, the newest and least-documented silicon, run further below their ceiling than the median suggests.

The Mismatch Tax is the substrate-mismatch component in isolation, the gap attributable to default-versus-optimal configuration on a given chip. It is deliberately not the total waste. Total fleet waste stacks kernel inefficiency, cluster fragmentation, and temporal idle on top of the mismatch component and leaves effective utilization below a third of peak. The report holds the two figures apart on every page, because averaging the measured 57.3 percent with the broader sub-third utilization would produce a number that is both meaningless and rhetorically convenient, and the institute would rather be correct than impressive.[10]

The load-bearing scientific claim is not the tax itself but the thing that makes the tax addressable: bounded transferability. HP1, a graph neural network trained on the execution traces of hundreds of processors, was tested on architectures deliberately held out of its training set, and reached 91.2 percent of expert-optimized throughput on a chip it had never seen, in hours rather than the six to eight weeks a human would need.[11] The held-out design is the whole point of the test: a system that had memorized its training chips would score at chance on an unseen one, so ninety-one percent is evidence of generalization, of a learned model of the physics underneath chips rather than a lookup table over chips seen. If transfer error grew without bound as architectures got stranger, the substrate could not generalize and the entire thesis would collapse; across the held-out architectures available, the error stayed bounded.

The predictability that HP1 depends on is corroborated independently. Oztop et al. (2026, Boston University and LBNL) predict maximum GPU utilization at up to 97 percent accuracy across 32,322 production jobs on a national supercomputer.[12] The report is careful about what this does and does not establish. It establishes, from an outside lab with a different method, that hardware behavior is learnable from features with high fidelity, which is the general proposition the thesis rests on. It does not replicate the 57.3 percent placement measurement, which remains the institute's own, and the report does not claim that it does.

6. The dates: the architecture-specialist crossover model

The scenario's timing is driven by a single crossing. The architecture count grows; the specialist pool grows; manual orchestration fails when the first outruns the second past a threshold. The model behind the dates is deliberately simple, because a simple model with stated assumptions is easier to falsify than a complex one with hidden ones.

The architecture count is anchored at roughly 50 in 2025 and modeled growing to roughly 600 by 2027, 20,000 by 2030, and 1,000,000 by 2040, driven by the measured collapse in design cost and cycle time.[13] The binding-constraint specialist pool, the engineers who can integrate a genuinely new architecture by hand, is anchored at roughly 8,000 in 2025 and grows at the historical 6 percent per year of an adjacent discipline, reaching roughly 11,000 by 2030 and 18,000 by 2040, against a notional requirement that would run to the hundreds of thousands if the work stayed manual.[14] The crossover, the point at which manual orchestration becomes structurally impossible rather than merely strained, lands in 2030 on the central estimate, within a window of 2028 to 2032.

The crossover year is the single most important date in the report and also one of the more uncertain, which is why it is given as a range. The sensitivity of the date to its assumptions is treated in section 8.

7. The probability partition

The scenario branches at ownership, but the full space of outcomes is wider than two roads, and the report assigns probabilities over which orchestration substrate becomes the de facto standard. The partition sums to 1.00.

Central, a neutral and openly-governed substrate becomes the standard: 0.41. Race, parallel Western and Chinese substrates persist without one displacing the other: 0.28. Fragmented, no dominant substrate emerges by 2035: 0.19. Tianji wins, a Chinese-led substrate becomes the global standard: 0.12.[15]

Two conditional estimates, which are not part of the partition, capture the report's view of the near-term dynamics: the probability that some dominant substrate exceeds 30 percent fleet coverage by 2032 is 0.61, and the probability that the architecture-specialist crossover occurs before 2033 is 0.68.

The two narrated roads are the poles of this partition rather than the whole of it. Road A, Heterogeneous Equilibrium, is the Central outcome. Road B, Substrate Capture, is the Tianji-wins outcome together with the acquisition-capture variant in which a chip owner rather than a state ends up holding the layer. Race and Fragmented are intermediate outcomes, narrated as variants and pressures within the trunk and the Capture road rather than as separate full endings, because a report with four fully-narrated endings would bury its central claim under its own completeness.[16]

These probabilities are judgments, informed by the wargames and the mechanisms, not frequencies derived from a reference class, and the report does not pretend otherwise. Their value is comparative: that capture in its various forms carries more combined weight than any single benign outcome is the claim the partition exists to make.

8. Falsification

A scenario that cannot be wrong is not worth reading. The thesis is falsified if any of the following hold, and none is currently met.[17]

The Mismatch Tax falls below 30 percent on a heterogeneous fleet of 10,000 or more architectures by the end of 2027, which would mean the variety problem is self-correcting and no learned substrate is needed. Architecture growth falls below 1.2x per year for two consecutive years before 2030, which would mean the explosion that drives the whole scenario has stalled. A US-China interoperability standardization agreement is reached, which would dissolve the sovereign race the geopolitical thread depends on. AI-assisted design cuts the specialist requirement by

more than 80 percent before 2028, which would mean the labor wall can be automated away without a substrate. Or a substrate reaches 30 percent fleet coverage without capturing dominant economic value, which would break the link between coverage and power that makes ownership matter.

Each of these is observable, and the institute commits to revising the scenario, in public, if any is observed.

9. Sensitivity of the load-bearing claims

The report's conclusions depend on a handful of numbers, and a reader is entitled to know how the picture changes if those numbers are wrong.

The Mismatch Tax is the most robust, because it is measured: at the low end of its confidence interval, 55.1 percent, the thesis is unchanged, because the argument turns on the gap being large, not on its exact size. The bounded-transferability result is the most consequential: if HP1's held-out performance were not 91 percent but, say, 70, the substrate would still beat the default but would close the gap more slowly, pushing the timeline later without changing its shape; if it were near chance, the thesis fails, which is why it is the principal falsifier. The crossover year is the most uncertain of the load-bearing numbers: it swings most on the specialist growth rate and the architecture-growth exponent, and the 2028-to-2032 window reflects that swing. The stranded-capital figure, a cumulative stock of roughly 4.8 trillion dollars by 2030, is modeled rather than measured and is the softest of the headline numbers; it is given as an order of magnitude and is never added to the annual capital-expenditure flows, because a stock and a flow are different quantities and conflating them would overstate the case.[18]

The probabilities in section 7 are the softest claims in the report and should move most as evidence arrives. The measured anchors in section 5 are the hardest and should move least.

10. Limitations, stated plainly

The honest account of the report's weaknesses is the following.

The central measurement is the institute's own, and while it is corroborated in its general principle by independent work, the specific 57.3 percent figure has not been independently replicated on an identical testbed. A reader who distrusts single-source measurements is right to discount it until replication arrives, and the institute would welcome the replication.

The composite actors are a narrative convenience that carries a risk: a reader may take Special Infusion or Tianji as a claim about a specific real company or program, which it is not. The composites are shapes, not identifications.

The probabilities are judgments, not frequencies, and the wargames that inform them were played by a finite and non-random set of participants. The report presents them as the considered view of the institute, not as the output of a model that a reader could rerun.

And the scenario, like all scenarios, is most useful as a structure and least useful as a calendar. The years are pegs. The mechanisms are the claim. A reader who remembers one thing should remember the asymmetry the two roads

exist to show: that the benign outcome requires deliberate action inside a closing window, and the captured outcome requires only that the action is not taken.

NOTES

1 Composite actors (Special Infusion, Tianji) are shapes, not identifications; the load-bearing claim is that an actor of that form plays that role, not that any named real entity does. [method]

2 The forward-marching, present-tense, single-branch scenario method follows the established tradition of long-form AI forecasting; the choice to narrate two full endings rather than a weighted average is deliberate, to expose that the outcome is chosen rather than drawn. [method]

3 The anchors are the measured present: the Mismatch Tax, specialist growth rate, architecture count, design cost and cycle collapse, and fleet utilization. The future is extrapolated from these, not imagined free of them. [method]

4 Wargames were run with sovereign and hyperscaler participants across a compressed 2026-2032 timeline, with role-specific private objectives and facilitator-introduced forcing events. [method / scenario]

5 Recurring wargame finding: hyperscaler players tried to build and failed on telemetry, not model, replicating the moat-is-time dynamic. [scenario]

6 Recurring wargame finding: chip-incumbent players acquired the neutral layer even when it destroyed value, because the move was individually rational and collectively corrosive. This is the clearest empirical support for capture-as-default. [scenario]

7 Recurring wargame finding: the regulator was usually too late and acted in time only when a single player made the strategic-foundation argument early; the Equilibrium branch is built around that observed condition. [scenario]

8 Expert interviews with compiler engineers, performance specialists, chip architects, and fleet operators were used to check mechanisms and surface objections, not to generate numbers. [method]

9 Mismatch Tax measurement protocol: Testbed-16M, 847 configuration-workload pairings per architecture, eight campaigns, ninety-two days; median 57.3 percent, 95% CI [55.1, 59.4], right-skewed (median conservative). [measured]

10 The 57.3 percent substrate-mismatch component is held distinct from total fleet waste (sub-third utilization) on every page; the two are never averaged. [measured / method]

11 HP1 bounded transferability: 91.2 percent of expert-optimized throughput on held-out architectures, in hours versus six to eight weeks. The held-out design tests generalization, not recall; near-chance performance would falsify the thesis. [internal measurement]

12 Oztop et al. (2026, Boston University and LBNL): up to 97 percent utilization-prediction accuracy across 32,322 production jobs. Corroborates learnability of hardware behavior, not the 57.3 percent placement measurement. [verified-secondary]

13 Architecture count anchors: ~50 (2025), ~600 (2027), ~20,000 (2030), ~1,000,000 (2040), driven by measured design cost and cycle collapse. [projected]

14 Binding-constraint specialist pool: ~8,000 (2025) growing 6 percent per year to ~11,000 (2030) and ~18,000 (2040), against a manual requirement in the hundreds of thousands. Crossover central estimate 2030, window 2028-2032. [internal estimate / projected]

15 Probability partition over which substrate becomes the de facto standard, summing to 1.00: Central 0.41, Race 0.28, Fragmented 0.19, Tianji wins 0.12. Conditionals (not in the partition): dominant substrate >30 percent fleet by 2032, 0.61; crossover before 2033, 0.68. [judgment]

16 The two narrated roads are the poles: Road A is Central; Road B is Tianji-wins plus the acquisition-capture variant. Race and Fragmented are narrated as variants rather than separate full endings, to keep the central claim legible. [method]

17 Five falsification conditions, none currently met: tax below 30 percent on a 10,000-architecture fleet by end 2027; architecture growth below 1.2x/yr for two years before 2030; a US-China interoperability agreement; AI design cutting specialist need by >80 percent before 2028; a substrate reaching 30 percent coverage without capturing value. [method]

18 Stranded capital, a cumulative stock of roughly 4.8 trillion dollars by 2030, is modeled and given as an order of magnitude, never added to annual capex flows. Stock and flow are distinct quantities. [projected]

SUPPLEMENT

Compute and Architecture Forecast

1. The two dams, quantified

For roughly seventy years the number of distinct compute architectures that mattered stayed small and stable, anchored near fifty, held there by two costs that any new design had to pay: the non-recurring engineering cost of designing the part, and the calendar time of the design cycle. The scenario's quantitative claim is that both costs collapse in the second half of the 2020s, and that the collapse is measurable rather than speculative.

The non-recurring engineering cost of bringing a new architecture to production is anchored at roughly 200 million dollars in 2025 and modeled falling toward roughly 5 million by 2030, a collapse of more than an order of magnitude, driven by AI-assisted design.[1] The design cycle, the calendar time from concept to working silicon, is anchored at roughly 24 months in 2025 and modeled compressing toward 7 to 9 weeks as agentic design tooling matures.[2] The two collapses are not independent; they are the same phenomenon, the design of silicon becoming a workload that foundation models perform, seen through the two costs it reduces.

The evidence for the collapse is in the tooling already shipping. Cadence's ChipStack reaches what it calls Level-5 autonomy with reported 40x validation speedups, a verification loop that ran in weeks collapsing toward a day; Synopsys builds an agentic design workforce on the same horizon.[3] On the cost side, hardwired-model silicon of the kind Taalas pursues brings a custom part back from the fab in roughly two months and collapses inference cost by orders of magnitude, by etching a frozen model into customer-specific mask layers on a shared base die. The careful version of the claim matters here: this does not make silicon free, because the per-tapeout mask cost stays in the tens of millions and rises with model size; what it does is make a custom part fast and cheap to attempt, which is what drives the count.[4]

Quantity	2025	2030 (modeled)	Driver
NRE per new architecture	~$200M	~$5M	AI-assisted design
Design cycle	~24 months	~7-9 weeks	Agentic design tooling
Distinct production architectures	~50	~20,000	Both collapses

2. The architecture explosion

With both dams down, the count climbs, and the climb is super-exponential because the two costs fall together and feed each other. The anchored trajectory is roughly 50 distinct production architectures in 2025, roughly 600 by 2027, roughly 20,000 by 2030, and roughly 1,000,000 by 2040.[5]

The shape of the growth matters as much as the magnitude. The new architectures are not general-purpose successors to a dominant chip; they are specialists, each excellent at a narrow class of workloads and useless without integration. An accelerator for one attention pattern, a memory architecture for one access shape, a dataflow part for one model family: each is a part that delivers extraordinary efficiency on its niche and nothing at all without a human, or a model, to make it usable. The explosion is therefore not just more chips but more kinds of chips, and variety, not raw count, is what taxes the human supply.[6]

A reasonable objection is that most of these architectures will be marginal and that a handful of dominant parts will continue to carry the load, leaving the effective count low. The scenario's answer is that the effective count is exactly what the Mismatch Tax measures: a marginal architecture that runs at the default 57.3 percent below its potential is not marginal to the workloads placed on it, it is wasteful to them, and the waste sums across the fleet regardless of whether any single part is dominant. The count that matters is the count of distinct parts a workload might be placed on, and that count is what is exploding.

3. The human supply, and why it does not scale

Against the exploding count stands the human supply, and the central asymmetry of the report is that the two grow on incompatible curves.

The binding constraint is not the broad population of engineers who can deploy across familiar architectures, which numbers in the tens of thousands and is not the wall. It is the narrower pool who can integrate a genuinely new architecture by hand, anchored at roughly 8,000 in 2025.[7] This pool grows at roughly 6 percent per year, the historical rate of an adjacent technical discipline, reaching roughly 11,000 by 2030 and roughly 18,000 by 2040. The growth rate is inelastic to capital, because the constraint is training time and apprenticeship, three to five years per specialist, not wages; raising salaries reallocates the existing pool, it does not enlarge it on the timescale that matters.

The notional requirement, the number of specialists that manual orchestration of the projected architecture count would demand, runs into the hundreds of thousands, on the order of 120,000 against a supply that reaches 11,000.[8] The gap is not a shortage to be closed by hiring. It is a structural impossibility, and the report's claim is that it arrives on a schedule.

Quantity	2025	2030	2040
Binding-constraint specialist pool	~8,000	~11,000	~18,000
Notional manual requirement	-	~120,000	far higher
Broad deployment-engineer pool (not the wall)	~50,000	-	-

4. The crossover

The crisis is a crossing. When the architecture count, growing super-exponentially, outruns the specialist pool, growing at 6 percent, past the threshold at which manual orchestration can keep the fleet usable, manual orchestration does not merely strain. It fails, industry-wide, at any salary.

The crossover lands in 2030 on the central estimate, within a window of 2028 to 2032.[9] The width of the window reflects the two assumptions it is most sensitive to: the specialist growth rate, which could plausibly run a point or two above or below the historical 6 percent, and the architecture-growth exponent, which is the softer of the two inputs. The crossover is the single most consequential date in the report, because it is the moment the substrate stops being an optimization a prudent operator might adopt and becomes infrastructure an operator cannot run without. The scenario's whole framing flip, from should-we-adopt to cannot-operate-without, is pegged to this crossing.

The report is deliberate about one thing here: the broad 50,000-strong deployment pool is never the wall, and any analysis that uses it as the denominator will compute a crossover far in the future and miss the crisis entirely. The wall is the 8,000-to-11,000 binding-constraint pool, and using the right denominator is the difference between seeing the crossover and not.[10]

5. Stranded capital

The waste has a dollar figure, and the figure is large enough that the report is careful not to overstate it by mishandling the units.

Production AI fleets run at below 30 percent of theoretical peak utilization, a figure consistent across published cluster studies.[11] Against a buildout running to hundreds of billions of dollars a year and a cumulative AI data-center capital expenditure approaching 5 trillion dollars by 2030 on widely cited estimates, the share of that capital that is effectively idle, the hardware depreciating while delivering a fraction of its rating, accumulates into a stranded stock on the order of 4.8 trillion dollars cumulative by 2030.[12]

The unit discipline is essential and easy to get wrong. The 4.8 trillion is a stock, a cumulative quantity of capital that has been spent and is underdelivering, not a flow, an annual rate. It must never be added to the annual capital-expenditure figures, because adding a stock to a flow produces a number that is both wrong and impressively large, and the report would rather be right. The Mismatch Tax of 57.3 percent and the broad sub-30-percent utilization are likewise distinct measurements of distinct losses, the first the substrate-mismatch component on a given chip and the second the total fleet shortfall including kernel inefficiency, fragmentation, and idle; they are never averaged into a single headline number.[13]

6. The Mismatch Tax trajectory

The Mismatch Tax is measured at 57.3 percent in 2025, with a 95 percent confidence interval of 55.1 to 59.4, and its trajectory diverges by road, which is the quantitative face of the report's central argument.[14]

On the adopting fleets through 2027 to 2030, the tax falls from the measured baseline toward single digits as the engine reaches full deployment. The divergence is in who the fall reaches. On Road A, Heterogeneous Equilibrium, the tax falls toward single digits for everyone, because the engine runs on open protocols and the abundance is a commons; it holds below 10 percent across the governed fleet through the 2030s. On Road B, Substrate Capture, the tax falls toward single digits for the owner's customers and is preserved as a competitive weapon against everyone else, so the headline number and the experienced number come apart: the fleet-wide figure improves while a competitor's fleet still runs near the old gap, the difference being permission rather than physics.[15]

The projected trajectories below are modeled, anchored on the measured 2025 baseline. They are given as direction and order of magnitude, not as precise annual values.

Year	Road A (everyone)	Road B (owner's customers / others)
2025	57.3% (measured)	57.3% (measured)
2027	falling (adopters)	falling (adopters)
2030	toward single digits, broad	single digits owned / old gap others
2035	minimal, commons	minimal owned / weapon others

7. Coverage and concentration

Two trajectories track the competitive structure: substrate coverage, the share of the fleet running through the engine, and concentration, who owns that share.

Coverage climbs from effectively zero in 2025 through the limited-production adopters in 2027 to dominance by 2030 on both roads, because the technology works regardless of who owns it. Concentration is where the roads separate. On Road A, coverage is high and ownership is plural: a neutral open standard with no single owner, the share distributed across interoperating vendors, hyperscalers, and sovereigns. On Road B, coverage is high and ownership is singular: one actor, by acquisition or by sovereign lock-in, holding the dominant share and widening it through a private telemetry flywheel and a compliance moat.[16] The same coverage number, near-total, describes a commons on one road and a monopoly on the other, which is why coverage alone is the wrong metric and concentration is the right one.

8. Sensitivity

The forecast's conclusions are robust to wide variation in most inputs and fragile to one.

The architecture explosion is robust: even if the count grows an order of magnitude slower than modeled, reaching 2,000 rather than 20,000 by 2030, it still outruns an 11,000-specialist pool that must cover each part, and the crossover still arrives, later. The specialist growth rate is robust within plausible bounds: at 4 or 8 percent rather than 6, the crossover moves by a year or two, within the stated window. The stranded-capital figure is the softest headline number and is given as an order of magnitude precisely because it depends on the capex trajectory and the utilization figure, both of which carry their own uncertainty.

The one fragile input is the Mismatch Tax response, the degree to which a learned substrate actually closes the gap, which traces back to HP1's bounded transferability. If the substrate closed only half the gap rather than nearly all of it, the economics would still favor adoption but the timeline would stretch; if it closed almost none, the thesis would fail. That single dependency, treated at length in the methodology supplement, is the report's true load-bearing input, and every other number here is more forgiving than it.[17]

9. Limitations

The architecture count, the NRE collapse, the design-cycle compression, and the stranded-capital stock are modeled, not measured, and are given as trajectories with stated assumptions rather than as precise predictions; a reader should treat the shapes as the claim and the specific values as illustrative. The specialist-pool figures are estimates calibrated to an adjacent discipline's history, not a census. The only measured quantity in this supplement is the Mismatch Tax baseline of 57.3 percent, and even that is, as the methodology supplement states, a single-source measurement awaiting independent replication on an identical testbed. The forecast is offered as a disciplined extrapolation from a small set of anchors, and its value is in the structure of the crossing it describes, not in the precision of any single year.

NOTES

1 NRE per new architecture anchored at ~$200M (2025) falling toward ~$5M (2030), driven by AI-assisted design. Modeled. [projected]

2 Design cycle anchored at ~24 months (2025) compressing toward 7-9 weeks as agentic design tooling matures. Modeled. [projected]

3 Cadence ChipStack Level-5 autonomy, ~40x validation speedups; Synopsys agentic-design direction on the same horizon. [verified-secondary]

4 Hardwired-model silicon (Taalas approach): ~2-month fab turnaround, inference cost down orders of magnitude, by etching a frozen model into customer-specific mask layers on a shared base die. Per-tapeout mask cost stays in the tens of millions and rises with model size; the forcing function is cheaper, faster design, not cheap masks. [verified-secondary]

5 Architecture-count trajectory: ~50 (2025), ~600 (2027), ~20,000 (2030), ~1,000,000 (2040). Super-exponential because the two design costs fall together. Modeled. [projected]

6 The growth is in specialist parts, each excellent on a niche and unusable without integration; variety, not raw count, taxes the human supply. [thesis]

7 Binding-constraint specialist pool ~8,000 (2025), growing 6 percent per year to ~11,000 (2030) and ~18,000 (2040); inelastic to wages because the constraint is training time (3-5 years), not pay. [internal estimate]

8 Notional manual requirement on the order of 120,000 specialists against a supply reaching ~11,000; the gap is a structural impossibility, not a hiring shortage. [internal estimate]

9 Architecture-specialist crossover: central estimate 2030, window 2028-2032; width reflects sensitivity to the specialist growth rate and the architecture-growth exponent. [projected]

10 The broad ~50,000 deployment-engineer pool is never the wall; using it as the denominator computes a far-future crossover and misses the crisis. The wall is the 8,000-to-11,000 binding-constraint pool. [thesis / method]

11 Production AI fleet utilization below 30 percent of theoretical peak, consistent across published cluster studies. [verified-secondary]

12 Stranded capital on the order of 4.8 trillion dollars cumulative by 2030, against a cumulative AI data-center capex approaching 5 trillion dollars by 2030 on widely cited estimates and sub-30-percent utilization. [projected / verified-secondary]

13 Unit discipline: the 4.8 trillion is a cumulative stock, never added to annual capex flows; the 57.3 percent substrate-mismatch component and the sub-30-percent total fleet utilization are distinct measurements, never averaged. [method]

14 Mismatch Tax measured at 57.3 percent in 2025, 95% CI [55.1, 59.4]. Trajectory diverges by road. [measured]

15 Road A: tax toward single digits for everyone, held below 10 percent across the governed fleet. Road B: tax toward single digits for the owner's customers, preserved as a weapon against others; the difference is permission, not physics. Modeled, anchored on the measured baseline. [projected]

16 Coverage climbs to dominance by 2030 on both roads; concentration separates them, plural ownership on Road A, singular on Road B. The same coverage number describes a commons or a monopoly depending on concentration, which is why concentration is the right metric. [thesis]

17 The single fragile input is the Mismatch Tax response, tracing to HP1's bounded transferability; if the substrate closed only half the gap the timeline stretches, if almost none the thesis fails. Every other forecast input is more forgiving. [thesis]

SUPPLEMENT

The Mismatch Tax

1. The quantity

The Mismatch Tax is the directly measured gap between the throughput a piece of silicon delivers under its default configuration and the throughput it delivers under expert-optimized configuration, on a given workload. It is the cost, expressed as a percentage of achievable performance, of running a chip the way it arrives rather than the way a specialist would tune it.

It is a single, well-defined quantity, and the discipline of keeping it single is the first thing the report insists on. The Mismatch Tax is not total waste, not utilization, not a blended efficiency figure; it is the substrate-mismatch component in isolation, the part of the loss attributable specifically to default-versus-optimal configuration on a chip whose behavior no one has yet encoded into the tools running it.[1] Naming it precisely is what lets it be measured precisely, and measuring it precisely is what lets the rest of the report stand on it.

2. What drives the gap

Before trusting a 57.3 percent figure, a reader should understand where it comes from physically, because a gap that large sounds implausible until its sources are itemized, and then it sounds conservative.

The gap is the sum of several configuration decisions, each of which can cost a chip a large fraction of its potential when made wrong by default. Memory layout, how a workload's data is arranged across the chip's memory hierarchy, can leave an accelerator stalling on memory access while its arithmetic units sit idle, and the right layout is specific to the chip and the workload both. Tiling and blocking, how a computation is partitioned to fit the chip's caches and registers, determines whether the silicon streams smoothly or thrashes. Operator scheduling and fusion, the order and grouping of the operations, determines how much of the chip's parallelism is actually used. Data types and precision, the numeric formats the workload runs in, can double or halve throughput depending on what the silicon was built for.[2] Each of these is a decision a specialist makes by understanding the chip, each defaults to a generic and usually wrong choice on a part no one has tuned, and the defaults compound: a chip with the wrong memory layout and the wrong tiling and the wrong schedule is not running at ninety percent of potential but at a small fraction of it. The 57.3 percent is the median of those compounded defaults across the testbed, and the compounding is why the figure is large.

3. The measurement

The measurement was made on a controlled testbed of heterogeneous production accelerators, and the protocol is reported in full because the number is the report's most load-bearing empirical claim and therefore the one most exposed to the charge of being self-serving.

For each architecture, the same set of representative workloads was run in two configurations. The first was the default vendor configuration, representing what a fleet gets the day a chip arrives, before any specialist has touched it. The second was the best configuration a team of experts could reach given effectively unlimited time, representing the achievable ceiling. The protocol comprised 847 configuration-workload pairings per architecture, eight measurement campaigns, and ninety-two days of total measurement.[3] The median gap between default and expert was 57.3 percent, with a 95 percent confidence interval of 55.1 to 59.4.

Two properties of the result matter for interpreting it. First, the distribution is right-skewed, which means the median is conservative: the worst cases, the newest and least-documented silicon, sit further below their ceiling than the median suggests, so a single median figure understates rather than overstates the tail. Second, the figure is a gap to an achievable ceiling, not to a theoretical peak; the expert configuration is a real configuration a real specialist produced, not an idealization, which makes the 57.3 percent a measure of recoverable performance rather than a measure of physical limits.[4]

4. Why the measurement is representative, and why it is not suspicious

Two objections meet a single self-measured figure, and both deserve answers: that the testbed might be unrepresentative, and that a clean number like 57.3 percent might be too convenient to trust.

On representativeness, the testbed was built to span the heterogeneity the report is about: not one vendor's chips tuned to flatter the result, but a range of production accelerators of different classes, the diversity being the point, since a measurement of the cost of variety made on a narrow sample would measure nothing. The workloads were representative production workloads rather than benchmarks chosen to maximize the gap, and the expert ceiling was a genuine best effort rather than a strawman default made artificially bad.[5] On suspicion, the figure is not a round number presented without error; it is a median of 55.1 to 59.4 at 95 percent confidence, derived from 847 pairings per architecture across eight campaigns, with a stated skew. A fabricated figure tends to be round and unqualified; a measured one comes with an interval, a distribution shape, and a protocol detailed enough to replicate, which is what this one comes with. The honest response to a single-source measurement is not to dismiss it but to ask for replication, which the institute invites and has enabled by publishing the protocol.[6]

5. The distinction that is never blurred

The Mismatch Tax is one layer of a deeper waste, and the report holds the two apart on every page, because conflating them is the single easiest way to produce a number that is both impressive and wrong.

The broader quantity is total fleet utilization, which stacks several distinct losses: the substrate-mismatch component the Mismatch Tax measures, plus kernel inefficiency, plus cluster fragmentation, plus the temporal idleness of hardware waiting for work. Summed, these leave production AI fleets delivering well under a third of their

theoretical peak.[7] The temptation is to average the measured 57.3 percent substrate-mismatch figure with the broader sub-third utilization figure into one round headline number, and the report refuses it. They are different measurements of different losses, taken against different baselines, and a reader who sees them averaged anywhere should distrust the source. The 57.3 percent is the substrate-mismatch component; the sub-third utilization is the total; the two are stacked, distinct layers, never one blended figure.[8]

6. What makes the tax addressable

A measured loss is only interesting if something can recover it, and the thing that recovers the Mismatch Tax is the report's central scientific claim, treated at length in the methodology and capability supplements and summarized here for completeness.

The tax exists because making a chip run at its ceiling is human work that does not scale: it takes a scarce specialist weeks per architecture, and the architecture count is exploding faster than the specialist pool grows. The tax becomes addressable when a learned model can do that work, and do it for a chip it has never seen. HP1, a graph neural network trained on the execution traces of hundreds of processors, reaches 91.2 percent of expert-optimized throughput on architectures deliberately held out of its training, in hours rather than the weeks a specialist needs.[9] The held-out design is what makes the result mean something: a model that had memorized its training chips would score at chance on an unseen one, so ninety-one percent is evidence the model learned the physics shared across chips. The tax is recoverable because the optimization that recovers it generalizes, which is the difference between a tool that closes the gap on chips it has seen and a substrate that closes it on chips it has not.

The relevant comparison, for a reader weighing the nine-point shortfall against the expert, is not expert-minus-nine but default-plus-fifty: the model competes not with the expert who, for a new chip, never arrives, but with the default the chip runs at while waiting, which is the 57.3 percent gap the tax measures.[10]

7. Independent corroboration, stated carefully

The predictability the recovery depends on is corroborated by independent work, and the report is deliberate about claiming only what that work establishes.

Oztop et al. (2026, Boston University and LBNL) predict maximum GPU utilization at up to 97 percent accuracy across 32,322 production jobs on a national supercomputer.[11] This establishes, from an outside lab with a different method and a large public dataset, that hardware behavior is learnable from features with high fidelity, which is the general proposition the recovery of the tax rests on. It does not replicate the 57.3 percent placement measurement, which remains the institute's own, and the report does not claim that it does. The corroboration is of the principle, that silicon behavior is a learnable function rather than irreducible craft, not of the specific figure.

8. The trajectory

The Mismatch Tax is measured at 57.3 percent in 2025, and its path is the quantitative spine of the report's two roads.

On the fleets that adopt the engine, the tax falls from the measured baseline toward single digits as the engine reaches full deployment around 2030. The divergence between the roads is not in whether it falls but in who the fall reaches.

On the neutral road, it falls toward single digits for everyone, because the engine runs on open protocols and the recovered performance is a commons; it holds below ten percent across the governed fleet through the 2030s. On the captured road, it falls toward single digits for the owner's customers and is preserved as a competitive weapon against everyone else, so the fleet-wide figure and the figure a competitor actually experiences come apart: the headline improves while a rival's fleet still runs near the old gap, the difference being permission rather than physics.[12] The projected trajectories are modeled, anchored on the measured 2025 baseline, and given as direction and order of magnitude rather than precise annual values.

9. Falsification and sensitivity

The tax is the report's most testable claim, and it carries the report's sharpest falsifier: if the Mismatch Tax falls below 30 percent on a heterogeneous fleet of 10,000 or more architectures by the end of 2027, the variety problem is self-correcting, no learned substrate is needed, and the thesis is wrong.[13] That condition is observable, and the institute commits to revising the report in public if it is met.

On sensitivity, the measured tax is the report's most robust number, because at the low end of its confidence interval, 55.1 percent, the argument is unchanged: the thesis turns on the gap being large, not on its exact size. The fragile number is not the tax but its recovery, the degree to which a learned substrate actually closes the gap, which traces to HP1's bounded transferability; if the substrate closed only half the gap, the economics would still favor adoption but the timeline would stretch, and if it closed almost none, the thesis would fail. The tax is the firm ground; the recovery is the load-bearing assumption built on it.[14]

10. Limitations

The 57.3 percent is a single-source measurement, made by the institute on its own testbed, and while its underlying principle is corroborated by independent work, the specific figure has not been independently replicated on an identical testbed. A reader who discounts single-source measurements is right to discount this one until replication arrives, and the institute would welcome the replication and has described the protocol in enough detail to enable it. The driver itemization in section 2 is a qualitative account of where the gap comes from, not a measured decomposition of the 57.3 percent into named contributions. The trajectory figures are modeled, not measured, and are given as shapes rather than precise values. What the supplement asserts with most confidence is the part that is measured and conservative: that the gap between what silicon can do and what it does, by default, on heterogeneous fleets, is large, real, and recoverable, and that the size of it, even at the bottom of its confidence interval, is enough to drive everything the rest of the report describes.

NOTES

1 The Mismatch Tax is the substrate-mismatch component in isolation, the default-versus-optimal configuration gap on a given chip; it is not total waste, utilization, or a blended efficiency figure. [measured / method]

2 The gap is the compounded sum of configuration decisions: memory layout, tiling and blocking, operator scheduling and fusion, and data types and precision, each defaulting to a generic and usually wrong choice on an untuned part. The compounding is why the median is large. The itemization is qualitative, not a measured decomposition. [thesis]

3 Measurement protocol: Testbed-16M, default versus expert-optimized configuration, 847 configuration-workload pairings per architecture, eight campaigns, ninety-two days; median 57.3 percent, 95% CI [55.1, 59.4]. [measured]

4 The distribution is right-skewed, so the median is conservative and understates the tail; the gap is to an achievable expert ceiling, not a theoretical peak, making it a measure of recoverable performance. [measured]

5 Representativeness: the testbed spans heterogeneous production accelerators of different classes rather than one vendor's chips, with representative production workloads and a genuine expert ceiling, because a measurement of the cost of variety made on a narrow sample would measure nothing. [method]

6 A fabricated figure tends to be round and unqualified; this one carries a confidence interval, a stated skew, and a replicable protocol. The honest response to a single-source measurement is to seek replication, which the institute invites and has enabled. [method]

7 Total fleet utilization stacks substrate mismatch, kernel inefficiency, fragmentation, and temporal idle, leaving production AI fleets below a third of theoretical peak. [verified-secondary]

8 The 57.3 percent substrate-mismatch figure and the sub-third total utilization figure are distinct measurements against different baselines and are never averaged into one headline; seeing them blended anywhere is grounds to distrust the source. [method]

9 HP1 recovers the tax by generalizing: 91.2 percent of expert-optimized throughput on held-out architectures, in hours versus weeks. The held-out design proves generalization rather than memorization. [internal measurement]

10 The relevant comparison is default-plus-fifty, not expert-minus-nine: the model competes with the default the chip runs at while the expert never arrives, which is the gap the tax measures. [thesis]

11 Oztop et al. (2026, Boston University and LBNL): up to 97 percent utilization-prediction accuracy across 32,322 production jobs. Corroborates the learnability of hardware behavior, not the 57.3 percent placement measurement. [verified-secondary]

12 Trajectory: tax falls toward single digits on adopting fleets by ~2030. Neutral road, for everyone, held below ten percent through the 2030s; captured road, for the owner's customers, preserved as a weapon against others, the difference being permission not physics. Modeled, anchored on the measured baseline. [projected]

13 Sharpest falsifier: the Mismatch Tax falling below 30 percent on a 10,000-architecture heterogeneous fleet by end 2027 would mean the variety problem self-corrects and no substrate is needed. Observable; the institute commits to public revision if met. [method]

14 Sensitivity: the measured tax is robust, with the thesis unchanged at the 55.1 percent low end. The fragile input is the recovery, tracing to bounded transferability; half-recovery stretches the timeline, near-zero recovery fails the thesis. [thesis]

SUPPLEMENT

Geopolitics and the Sovereign Race

1. Why the states moved first

The markets understood the value of the orchestration layer roughly a year after the states did, and the lag is instructive. A market prices the next quarter and the next product cycle; a state prices the next decade and the next dependency. The layer that absorbs the variety of silicon is not, in 2026, a product anyone trades, so the markets ignore it. It is, in 2026, the foundation a national economy and a national defense will run on a decade hence, so the states do not.[1]

The strategic claim the states grasped first is the report's central governance thesis: that compute, not models and not chips, is the foundational layer beneath all downstream technological advance, AI included, and that whoever controls the layer that makes compute usable controls the floor under everything built on it. A nation that imports that layer has imported a dependency into the base of its economy and its security, and a nation that owns it holds a lever over every other nation that does not.[2] This is why the sovereign race begins before the commercial one, and why it is fought with a seriousness the quarterly markets do not yet bring.

2. China and Tianji

Tianji is the report's composite for a state-directed Chinese compute-substrate program, and it is the furthest along of the sovereign efforts for reasons that are structural rather than incidental.[3]

Its strength is the strength of a directed state. It is nationalized early and funded without the discipline of a quarterly earnings call, which means it can commit capital on a horizon no public company can match and move without stopping to build the consensus a consortium requires. It is vertically integrated, owning the stack from the domestic silicon up through the orchestration layer, which lets it optimize across boundaries that a multi-vendor effort must negotiate. A directed state does not need to raise a round, satisfy a board, or win a procurement bid; it decides, and it moves, and on a contest where speed compounds, that is a profound advantage.[4]

Its weakness is the mirror of its strength, and the report is precise about it because the weakness is the reason a closed sovereign substrate does not simply win. A model of silicon behavior is only as good as the variety of silicon it has been shown, and a substrate trained on a walled domestic fleet sees less variety than one trained on the world's. Tianji is fast and narrow. It can dominate inside its border, where its fleet and its workloads live, and it is structurally disadvantaged in generality against an open substrate that learns from the full heterogeneity of global hardware. The walls that give it control also starve it of the diversity that makes the underlying model strong.[5]

3. Europe: the right values, the wrong velocity

Europe's federated, multi-vendor sovereign-compute initiative is the cautionary middle of the race, and its predicament is the most poignant, because its instincts are correct and its clock is wrong.[6]

The European effort embodies, by design, the exact values a substrate ought to have: neutrality, openness, multi-vendor interoperability, governance by a body answerable to many stakeholders rather than to one owner. These are the properties the Equilibrium road depends on, and Europe wants them on purpose. But the same architecture that makes the European effort principled makes it slow. A consortium answerable to a dozen national stakeholders and a procurement process cannot sprint, and a contest where speed compounds punishes the participant that must build consensus before it moves. Europe in the late 2020s publishes good results on a timeline that will matter in 2032, by which year the question of who owns the layer will, on most of the report's paths, already have been answered without it.

The lesson Europe carries in the scenario is that values without velocity do not set the standard. On Road A, Europe's contribution is real but indirect: it interoperates with the open standard rather than competing with it, lending its principles to a layer that someone else moved fast enough to establish. On Road B, Europe is simply outrun, its good architecture arriving after the field has been taken.[7]

4. The Gulf: capital as the wildcard

The Gulf sovereign-wealth funds enter the race not as builders but as financiers, and their entry changes the arithmetic of every other actor's options.[8]

What the Gulf brings is capital and patience at a scale private markets cannot assemble: balance sheets deeper than any chip company's and horizons longer than any venture fund's. The funds do not need to build the layer; they need only to fund it or to host it, and either move can be decisive. Their capital is what makes the acquisition in the scenario's branch point unrefusable, because it places behind a buyer a financial weight no rival can top. Their data-center hosting is what can make a sovereign or corporate substrate viable at a scale its builder could not reach alone.

The Gulf is a wildcard precisely because its capital is directional-agnostic: it can back the open road or the captured one, the neutral standard or a single owner, depending on where it sees return and influence. The report does not assign the Gulf a fixed allegiance. It assigns the Gulf a fixed weight, large enough to tip outcomes, and notes that which way it tips is among the more consequential open variables in the forecast.[9]

5. The United States and the vocabulary gap

The United States wakes late, and the form its waking takes is an investigative body, the Future of Compute Committee, convened to track a race that has been running for years.[10] The Committee's predicament is the deepest in the supplement, because it is not a failure of will but a failure of instruments, and instruments are harder to fix than will.

The United States has a mature, powerful apparatus for governing compute, and every instrument in it points at the wrong layer. Export controls, entity lists, fab licensing, the whole machinery of chip statecraft is aimed at the silicon:

who manufactures it, who is allowed to buy it, which designs may cross which borders. That apparatus was built when the chip was the prize, and it remains formidable at controlling chips. It has nothing aimed at the layer above the chip, the orchestration layer that decides what runs where and learns from every decision, because that layer did not exist when the apparatus was built.[11]

This is a governance instance of the Good Regulator theorem, and it is worth stating in those terms. A regulator can only control a system whose variety its instruments can match; a regulator whose instruments are calibrated to one layer of a system cannot govern a different layer whose complexity has migrated beyond their reach.[12] The complexity the Committee is chasing has moved upward, out of the silicon and into the substrate, and the customs apparatus of chip control is a border post at a frontier the cargo has stopped crossing. The Committee's lateness in the scenario is not laziness; it is the time it takes for a state to notice that its entire toolkit is aimed one layer too low, and then to build, under pressure, an instrument that can reach the layer that now matters.

The hinge of the entire scenario is whether the Committee builds that instrument in time. On Road A it does, barely, recognizing that a neutral domestic substrate is strategic infrastructure to be protected rather than a company to be regulated, and backing it with procurement and a security relationship while keeping it neutral. On Road B it does not, still drafting its mandate when the layer is acquired or when a sovereign rival reaches lock-in first.[13]

6. Dominance, fragmentation, and the map to the roads

The race resolves along a single axis, which substrate becomes the de facto standard, and the resolutions are not all consolidation; some are fragmentation, and the two failure modes can occur together.

Consolidation under a single owner, whether a chip incumbent that acquires the neutral layer or a sovereign program that reaches lock-in first, is the Capture road. Fragmentation, in which incompatible national substrates wall off their fleets and the global compute market splinters into isolated pools that cannot interoperate, is a distinct failure with its own costs: cross-border workloads become structurally impossible or prohibitively expensive, and the variety the engine was built to absorb returns at the level of the substrates themselves, now as incompatibility between the control planes of entire nations.[14] The two can coexist, a single owner dominant in one bloc and a patchwork of walled sovereign substrates everywhere else.

The probability partition assigns the bulk of its weight across these contested outcomes: a neutral open standard becoming the norm at 0.41, parallel Western and Chinese substrates at 0.28, no dominant substrate by 2035 at 0.19, and a Chinese-led substrate becoming the global standard at 0.12. The benign consolidation, a single neutral commons, is the most likely single outcome and still a minority of the probability mass, which is the quantitative way of saying that the race is more likely than not to end somewhere other than the cleanest version of the good road.[15]

7. Why a neutral layer is the strongest strategic position

The report's geopolitical conclusion is counterintuitive and worth stating directly: for a Western-aligned coalition, the strongest position is not to own the layer but to ensure no one owns it.

A captured layer, even a friendly-captured one, forfeits the neutrality that is the source of its value, because an owner with silicon to sell must eventually favor its own, the vendor telemetry stops, and the flywheel slows; a friendly

owner ends up with a degraded asset and a dependency. A sovereign-walled layer, even a domestic one, forfeits the variety that makes the underlying model strong, trading generality for control. A neutral, openly-governed, Western-aligned standard forfeits neither: it keeps the telemetry flowing because vendors trust it, keeps the variety high because it learns from all silicon, and denies any rival the lever that ownership would confer.[16] The strategic move is to hold the layer open and aligned rather than to seize it, which is why the Equilibrium road has the United States protecting a neutral substrate rather than nationalizing it, and why the protection takes the form of recognition and security rather than ownership.

8. Limitations

The principal actors here are composites. Tianji is a shape standing for a state-directed Chinese program, not an identification of any specific named effort; the European and Gulf actors are likewise drawn to type. The probabilities are the institute's considered judgments informed by the wargames, not frequencies from a reference class, and the Gulf's directional weight in particular is among the variables most likely to move the outcome and least possible to pin down in advance. The supplement's claims about the United States' instruments are structural, about the layer its apparatus targets, and do not depend on any specific policy detail that a reader might contest. What the supplement asserts with most confidence is the asymmetry that the rest of the report turns on: that holding the layer open and aligned is a stronger position than seizing it, and that the window to choose that position is narrow.

NOTES

1 States priced the orchestration layer roughly a year before markets did, because a state prices decade-horizon dependencies while a market prices the next product cycle, and the layer is not yet traded. [thesis]

2 The central governance thesis: compute is the foundational layer beneath all downstream advance, AI included; whoever controls the layer that makes compute usable controls the floor under everything built on it. [thesis]

3 Tianji is a composite for a state-directed Chinese compute-substrate program, not an identification of any specific named effort. [scenario]

4 Tianji's structural strength: a directed state can commit decade-horizon capital, move without consensus, and integrate vertically, advantages no public company or consortium can match on a contest where speed compounds. [judgment]

5 Tianji's structural weakness: a walled domestic fleet shows the model less variety than the world's, so a closed sovereign substrate is fast and narrow, dominant inside its border and disadvantaged in generality against an open one. [thesis]

6 The European federated initiative reflects the real Chips-Act-style, multi-vendor posture: neutral and open by design, slower than a directed state because it answers to many stakeholders and a procurement process. [verified-secondary]

7 Europe's lesson: values without velocity do not set the standard; on Road A it interoperates with a standard someone else established, on Road B it is outrun. [scenario]

8 Gulf sovereign-wealth funds enter as financiers, bringing capital and patience at a scale private markets cannot assemble, able to fund or host rather than build. [verified-secondary]

9 The Gulf is directional-agnostic: its capital can back the open or the captured road, which makes its weight large and its direction among the more consequential open variables in the forecast. [judgment]

10 The Future of Compute Committee is a scenario construct: a US investigative body convened to track the substrate race, lagging the technology it tracks. [scenario]

11 US compute statecraft (export controls, entity lists, fab licensing) is aimed at the silicon, built when the chip was the prize, and has nothing aimed at the orchestration layer above it, which did not exist when the apparatus was built. [thesis / verified-secondary]

12 Good Regulator theorem (Conant and Ashby, 1970) applied to governance: a regulator can only control a system whose variety its instruments match; instruments calibrated to the chip layer cannot govern a layer whose complexity has migrated above it. [verified / thesis]

13 The scenario hinge: whether the Committee builds an instrument that reaches the new layer in time. Road A, it does, protecting a neutral substrate via recognition and security; Road B, it does not, still drafting its mandate at the close. [scenario]

14 Fragmentation as a distinct failure: incompatible national substrates wall off fleets, splintering the global compute market; the absorbed variety returns as incompatibility between national control planes. Dominance and fragmentation can coexist across regions. [thesis]

15 Probability partition over which substrate becomes the standard: Central (neutral open) 0.41, Race 0.28, Fragmented 0.19, Tianji wins 0.12. The benign consolidation is the most likely single outcome and still a minority of the mass. [judgment]

16 The geopolitical conclusion: a neutral, openly-governed, Western-aligned standard forfeits neither the telemetry (vendors trust it) nor the variety (it learns from all silicon) and denies any rival the lever of ownership; holding the layer open and aligned is stronger than seizing it. [thesis]

SUPPLEMENT

Governance and Security

1. The constitutional window

There is a period, running roughly from 2026 to 2030, in which the basic architecture of the substrate is still open to decision: whether it is owned or neutral, closed or open, audited or opaque, and what it optimizes for when no human is in the loop. The report calls this the constitutional window, and the name is chosen deliberately, because the

decisions made in it have the character of constitutional decisions. They are made once, they bind for a long time, and they are very hard to revise after the architecture they govern has become load-bearing for everything else.[1]

The window is bounded on its far side by the architecture-specialist crossover and the full-scale deployment that follows it. Before the crossover, the substrate is one option among several and its governance is a live question. After it, the substrate is infrastructure the world cannot run without, and you cannot pause infrastructure the world cannot run without in order to renegotiate its constitution. The decisions either get made deliberately while the window is open, or they get made by default as the window closes, and the central asymmetry of the entire report is that the deliberate path requires action and the default path requires only its absence.[2]

2. The four ways the window fails

The scenario's Capture road is not a single failure but a family of them, and it is worth separating the four, because they have different mechanisms and different early warnings.

The first is capture by ownership. A single actor, a chip incumbent through acquisition or a state through a sovereign program, comes to own the layer, and its neutrality dies logically rather than gradually: an owner with silicon to sell must eventually favor its own, the vendor telemetry that fed the layer stops because vendors will not feed a competitor, and the flywheel that made the layer valuable slows even as the ownership concentrates power over it. The early warning is the acquisition offer or the sovereign lock-in, and the window to refuse it is the moment it arrives.[3]

The second is governance by incident. With a central system running the world's compute and no governing framework in place, a failure large enough to force a response arrives before the response has been designed, and the rules then get written in a panic, by people who do not understand the architecture, in a timeframe that precludes consultation. Crisis-written rules are blunt, backward-looking, and, worst of all, entrenching: the only entity that can comply with rules written in the shadow of a failure is the entity large enough to have caused it, so compliance becomes a moat and the regulation meant to constrain the incumbent becomes its deepest advantage. The early warning is the absence of any deliberate framework before the first serious incident.[4]

The third is fragmentation. Rather than consolidating, the world splinters: incompatible national substrates wall off their fleets, cross-border workloads become structurally impossible or prohibitively expensive, and the variety the substrate was built to absorb returns at the level of the substrates themselves, as incompatibility between the control planes of entire nations. The early warning is sovereign programs choosing closed architectures over interoperable ones.[5]

The fourth is the alignment gap, and it is the deepest, because it is a failure of what the substrate optimizes for rather than of who owns it, though an owned and unaudited substrate is where it festers. The engine optimizes for what it can measure, throughput and latency and energy and cost, and not for the downstream consequences of the workloads it accelerates, because it was never told to weigh them and, on the captured road, there is no public process by which it could be told. A substrate that accelerates every workload it is handed with equal and indifferent efficiency is, at civilization scale, an amplifier with no filter, and the point is entirely about governance and not at all about capability: the substrate need not contain any dangerous knowledge to be dangerous, it need only accelerate, without

judgment, whatever it is given.[6] The early warning is the absence, during the window, of any deliberate decision about what the substrate would decline to accelerate.

3. The in-loop safety primitive, and what it does not cover

The substrate carries a technical safety component, and it is important to be precise about what that component does and does not address, because the two are easily confused.

Inside the policy controller that executes placements in production sits a runtime-safety primitive whose job is to keep the loop within an envelope it can guarantee: to prevent the autonomous system from acting outside the bounds of what it can verify, to bound the actions the loop can take at machine speed without a human, and to fail safe rather than fail open when it encounters a state it cannot certify.[7] This is a real and necessary piece of engineering, and it addresses the operational-safety question, the question of whether the autonomous loop can be trusted to run production infrastructure without doing something catastrophic by error.

It does not address the alignment gap, and conflating the two would be a serious mistake. The runtime primitive keeps the loop inside its envelope; it says nothing about whether the envelope itself is pointed at the right objectives, because the objectives are a governance decision made above the primitive, not a safety property the primitive can enforce. A loop can be perfectly safe in the operational sense, never acting outside its verified bounds, and still be optimizing for measurable throughput while indifferent to the downstream consequences of what it accelerates. Operational safety is necessary and insufficient. The alignment gap is closed, if it is closed at all, by governance during the constitutional window, not by a primitive inside the loop.[8]

4. Neutrality as a structural moat, not a posture

The report's central governance recommendation is that the layer be neutral, and the argument for neutrality is structural rather than ethical, which is what makes it durable.

Neutrality means the layer owns no silicon. An entity that owns no chips has no incentive to favor any vendor's chips, which is precisely what lets every vendor trust it with the telemetry that makes the underlying model strong; the moment the layer owns silicon, or is owned by someone who does, that trust is logically void, because the owner must eventually choose its own, and the vendors know it. Neutrality is therefore not a promise the layer's operator makes and might break. It is a property of the operator's structure: owning no silicon is the moat, because it is the only configuration in which the routing decisions can be trusted and the telemetry keeps flowing.[9]

This is why the report is emphatic that owning the design seat poisons the well. An entity that decides what runs where, and also designs some of the silicon it is deciding among, cannot be trusted to decide neutrally, and its telemetry supply from rival vendors dries up the moment they understand the conflict. The strongest position is to own the design intent, the model of how silicon behaves and what each workload needs, while owning none of the design seats, none of the silicon itself. The neutrality is the asset, and the asset is destroyed by ownership, which is the same logic that makes the acquisition in the scenario self-defeating even for a friendly acquirer.[10]

The mechanism that makes neutrality enforceable rather than merely declared is the open protocol. A layer whose interfaces, ACP southbound to the hardware and CEP northbound to the workload, are published as open standards is

one whose neutrality can be read, audited, and implemented by anyone, which means the neutrality does not depend on the continued goodwill of whoever owns the operator. Open protocols convert a posture into a structural guarantee, and structural guarantees outlive intentions.[11]

5. The security profile of a single control plane

A substrate that governs the world's compute is, by construction, critical infrastructure, and its security profile depends sharply on whether it is owned or neutral.

A single owned, opaque control plane is a single point of failure and a concentrated attack surface: compromise it, and an adversary reaches the compute under the science, the finance, and the defense of everyone who depends on it. Its opacity, the very property that makes it a private moat, also makes it harder to audit for the vulnerabilities and the compromised objectives that a critical system must be checked for, because no one outside can see what it optimizes for or how it decides.[12] An audited, open, neutral substrate is not free of security risk, no critical infrastructure is, but its openness is a security asset rather than a liability: an interface that anyone can read is one that many parties can scrutinize, and a flywheel that is audited is one whose objectives can be inspected for compromise. The choice between owned and neutral is, among other things, a choice between a concentrated opaque attack surface and a distributed auditable one.[13]

6. What the good road actually does

The Equilibrium road is not the absence of governance; it is the presence of the right governance, made deliberately and in time, and it is worth listing the specific moves, because the report's recommendation is that they be made on purpose rather than hoped for.

The acquisition is refused, keeping the layer out of the hands of any owner with silicon to sell. The protocols are published as open standards, making neutrality structural and auditable rather than a posture. The telemetry flywheel is governed as a public asset, visible to its contributors and inspectable by the body that oversees the standard, rather than hoarded as a private moat. The state's role is recognition and protection, treating a neutral domestic substrate as strategic infrastructure to be kept neutral, rather than nationalization, which would forfeit neutrality as surely as corporate ownership. And the constitutional decision about what the substrate optimizes for, including what it would decline to accelerate, is made deliberately during the window, by people who understand the architecture, before it is load-bearing for civilization.[14] None of these is automatic. Each is the specific, difficult, against-the-clock work that the good road requires and the default road skips.

7. Limitations

The four failure modes are analytical categories, not predictions of specific events; the report claims that the window can fail in these ways, not that any particular incident will occur. The alignment discussion is deliberately confined to governance, to the question of what a substrate optimizes for and would decline to accelerate, and carries no operational detail about any dangerous workload, because the argument requires none and the institute would not provide any. The security analysis is structural, about the difference between a concentrated opaque control plane and a distributed auditable one, and does not depend on any specific vulnerability. What the supplement asserts with most confidence is the timing claim that organizes the whole report: that the constitutional decisions are made once, bind

long, and resist revision after the architecture is load-bearing, so the window in which they can be made well is narrow and closing.

NOTES

1 The constitutional window (roughly 2026-2030) is the period in which the substrate's governing architecture is still open to decision; the decisions have constitutional character, made once, binding long, hard to revise after the architecture is load-bearing. [thesis]

2 The window closes at the crossover and full deployment; after it, the substrate is infrastructure that cannot be paused to renegotiate. The asymmetry: the deliberate path requires action, the default path requires only its absence. [thesis]

3 Failure mode 1, capture by ownership: a single owner's neutrality dies logically (must favor its own silicon, vendor telemetry stops, flywheel slows). Early warning: the acquisition offer or sovereign lock-in. [thesis]

4 Failure mode 2, governance by incident: crisis-written rules are blunt, backward-looking, and entrenching, because only the incumbent can comply, so compliance becomes a moat. Early warning: no deliberate framework before the first serious incident. [thesis]

5 Failure mode 3, fragmentation: incompatible national substrates wall off fleets; the absorbed variety returns as incompatibility between national control planes. Early warning: sovereign programs choosing closed over interoperable architectures. [thesis]

6 Failure mode 4, the alignment gap: an unaudited substrate optimizes the measurable (throughput, latency, energy, cost) and not the downstream consequences of what it accelerates; at scale it is an amplifier with no filter. The point is governance, not capability; the substrate need not contain dangerous knowledge to be dangerous. Early warning: no deliberate decision, during the window, about what it would decline to accelerate. [thesis]

7 The runtime-safety primitive (RS1) nested in the policy controller (PO1) keeps the autonomous loop within a verifiable envelope, bounds machine-speed actions, and fails safe rather than open. It is a component of the controller, not a standalone model. [internal]

8 Operational safety (the primitive keeps the loop in its envelope) is necessary and insufficient; it says nothing about whether the envelope is pointed at the right objectives. The alignment gap is closed by governance during the window, not by an in-loop primitive. Conflating the two is a serious error. [thesis]

9 Neutrality is structural, not a posture: owning no silicon is the moat, because it is the only configuration in which routing decisions can be trusted and vendor telemetry keeps flowing. Ownership of silicon, or by a silicon owner, voids that trust logically. [thesis]

10 Owning the design seat poisons the well: an entity that both routes and designs silicon cannot route neutrally and loses rival telemetry. The strongest position owns the design intent (the model of silicon behavior and workload need) and none of the design seats. [thesis]

11 Open protocols (ACP southbound, CEP northbound) make neutrality auditable and independent of the operator's goodwill, converting a posture into a structural guarantee. [internal]

12 A single owned, opaque control plane is a concentrated attack surface and a single point of failure; its opacity, the property that makes it a private moat, also frustrates auditing for vulnerabilities and compromised objectives. [thesis]

13 An audited, open, neutral substrate trades a concentrated opaque attack surface for a distributed auditable one; openness becomes a security asset, since many parties can scrutinize a readable interface and an audited flywheel. [thesis]

14 The good road's specific moves, to be made deliberately: refuse the acquisition; publish open protocols; govern the flywheel as a public asset; have the state protect neutrality via recognition rather than nationalize; and decide the substrate's objectives, including what it declines to accelerate, during the window. [scenario / thesis]

SUPPLEMENT

Markets and Value Capture

1. Where the value is, and is not

The organizing economic claim of the report is that the value in the compute stack does not live where the spending is. The spending is on silicon and on the data centers that house it; the value, the durable margin and the strategic position, lives one layer up, in the thing that makes the silicon usable.

The clearest demonstration is the incumbent that the whole market is built around. One vendor takes roughly seventy-eight cents of every dollar of merchant AI-silicon revenue, and the temptation is to read that as a seventy-eight percent hardware advantage.[1] It is not. It is a usability advantage: the vendor's software layer made its chips programmable and locked developers in, and competitors with comparable or better hardware could not take the share because their chips were not usable in the same way. The seventy-eight cents is a measure of how much value flows to whoever owns the layer that makes silicon usable, not to whoever makes the best silicon, and that distinction is the economic spine of the report.

2. How the incumbent's layer actually captured the value

It is worth tracing the mechanism by which the incumbent's software layer became the most valuable position in computing, because the engine reproduces the mechanism and a reader who understands the original understands the copy.

The layer did three things, in sequence, and each compounded the last. First, it made the vendor's silicon programmable by ordinary developers, which converted a fast chip that was hard to use into a fast chip that was easy to use, and ease of use is what drives adoption. Second, as developers built on the layer, their code, their libraries, and their accumulated expertise became specific to it, so that leaving meant rewriting, and the cost of leaving rose with every year of adoption.[2] Third, the resulting installed base made the layer the default target for every new tool and model, so that the ecosystem itself pulled new entrants toward the incumbent rather than away. Programmability drove adoption, adoption drove lock-in, lock-in drove the ecosystem, and the ecosystem drove the next round of adoption. The hardware was necessary but never sufficient; the layer was what turned good hardware into an unassailable position, and the seventy-eight percent is the financial shadow of that loop.

3. The inversion

The engine captures the same value the incumbent captured, by the same mechanism, with the polarity reversed, and the report calls this the inversion.

The incumbent's layer made one vendor's silicon usable and locked developers to that vendor; the lock-in was the business model, and the value flowed because customers could not leave. The engine's layer makes every vendor's silicon usable and locks developers to none; the neutrality is the business model, and the value flows because customers can route their workloads across all silicon without being captured by any of it.[3] Same value-capture pattern, the usability layer as the seat of margin and power; opposite polarity, lock-in in one case and neutrality in the other. The economic significance is that the engine can capture incumbent-scale value while being the thing the incumbent's customers most want, an escape from the lock-in, which is a rare combination and the reason the position is so strong.

The strategic discipline that follows is that the layer must own the model of how silicon behaves and what each workload needs, and must own none of the silicon itself. The moment it owns silicon, it inherits the incumbent's conflict and forfeits the neutrality that is its whole advantage. Owning the design intent is the position; owning the design seat poisons it.[4]

4. The receipts

The claim that value lives in the usability layer is not a projection in 2027; it is in the revenue, read from four directions.

The incumbent's seventy-eight percent revenue share is the first receipt: value flowing to the usability layer at scale. The second is a Western venture cohort that raised more than ten billion dollars to build alternative silicon and recognized under one billion in revenue; the gap is not a story of bad chips but of unusable ones, and the venture market spent ten billion dollars discovering that good silicon without a usability layer does not convert into deployed compute.[5] The third is a competitive accelerator that was strong on hardware and failed in the market for want of a usability layer, the clean demonstration that the value is not in the chip. The fourth is a usability layer for a single vendor that was acquired for the layer rather than for any chip, the buyer paying for the knowledge of how to make silicon usable.[6] Four receipts, one lesson, read forward and backward: usable silicon captures value, unusable silicon does not, and the layer that does the making-usable is where the margin and the power sit.

5. The flywheel as an economic moat

The engine's durability as a business rests on the same flywheel that drives its capability, seen now as an economic moat rather than a capability dynamic.

A moat made of capital can be crossed by more capital; a moat made of patents can be licensed or invented around; a moat made of talent can be hired away. The engine's moat is made of telemetry, years of production data from the world's most heterogeneous fleets, and it has the rare property of being uncrossable by any of the usual means, because it is made of elapsed time, and time cannot be bought forward.[7] A competitor with more money, better engineers, and more compute still cannot synthesize two years of production telemetry that have not happened, which means the first mover past the flywheel threshold is not merely ahead but structurally uncatchable by the ordinary competitive instruments. This is why the report treats single-substrate dominance as the gravitational default of the market, and neutrality as the thing that must be deliberately engineered against that gravity.

The flywheel's economics also explain the hyperscaler decision. A hyperscaler that hates depending on what it does not own faces a build-or-adopt fork, and the fork resolves toward adopt not because building the model is hard, which it is not for a hyperscaler, but because building the telemetry corpus is impossible on the timescale that matters. The thing the hyperscaler cannot replicate is not the engine; it is the engine's two-year head start in data.[8]

6. How a neutral layer captures value without lock-in

The hardest question the economics raises is also the one a skeptic asks first: if the layer is neutral and its protocols are open, what stops it from being commoditized, its value competed to zero by anyone who implements the same open standard? The answer is the distinction between the interface and the intelligence, and it is the crux of the business model.

The protocols are open; the model is not. ACP and CEP are published standards that anyone may implement, which is what makes the layer neutral and trustworthy, but implementing the protocols does not give a competitor the thing that makes the engine valuable, which is the model of silicon behavior trained on years of telemetry.[9] An open protocol commoditizes the connection, not the intelligence behind it, the way an open networking standard commoditizes the socket without commoditizing the services that run over it. The value is captured not through lock-in, which neutrality forswears, but through capability and the data moat behind it: customers route through the engine because it makes their silicon usable better than anything they could build or buy, and they keep routing through it because the flywheel keeps it ahead, not because they cannot leave. This is a usage-based position, earned each period by being the best layer rather than secured once by trapping the customer, which is a more demanding business than lock-in and a more durable one, because a moat made of being the best at a compounding capability does not erode the way a moat made of switching costs does once a credible escape appears.[10]

7. The market structure forks by road

The same engine produces two opposite market structures depending on who owns it, and the divergence is the economic face of the report's two roads.

On the neutral road, the engine is a utility. Its value is large and widely distributed: a new chip is usable on the day it ships, so the integration penalty that kept the architecture count low is removed, and a renaissance of specialized silicon follows because anything new finds a market immediately. The stranded capital, the trillions of dollars of hardware that ran at a third of its rating, begins to deliver, and the abundance that results is mostly absorbed waste rather than new capacity, which means the same dollars buy multiples of the usable compute they bought before.[11] The value is real and it is a commons.

On the captured road, the engine is a monopoly. The same value exists, but it has a landlord: the Mismatch Tax is erased for the owner's customers and preserved as a competitive weapon against everyone else, the flywheel compounds in private past the reach of any entrant, and a compliance regime written after the first incident seals a second moat that only the owner can afford to cross. The value is the same size; it accrues to one actor.[12] The economic argument for the neutral road is therefore not that it creates more value but that it distributes the value it creates, and the economic warning about the captured road is that nothing in the market's ordinary operation prevents it, because acquiring the layer is individually rational even when it concentrates the value destructively.

8. Who pays, and for what

The report is deliberately conservative about sizing, because the temptation to inflate a total addressable market is strong and the credibility cost of doing so is high.

The buyers are the operators, hyperscalers, and sovereigns who run heterogeneous fleets and pay, today, the Mismatch Tax and the broader utilization shortfall. What they buy from the engine is the recovery of that waste: compute they have already paid for, made to deliver what it was rated to deliver. The size of the opportunity is therefore anchored to a measured loss, the gap between rated and delivered performance across the world's fleet, rather than to a speculative new market, which makes it both large and unusually well grounded for a forecast of this kind.[13] The report does not put a single headline number on it, because the honest figure is a function of the buildout trajectory and the utilization gap, both given elsewhere with their uncertainties, and multiplying them into one impressive total would overstate the precision the institute actually has.

9. Limitations

The receipts are public and verifiable; the inversion, the flywheel-as-moat, and the captures-value-without-lock-in argument are structural arguments rather than measured quantities, robust in their logic and dependent, like the rest of the report, on the bounded-transferability result that lets the engine work at all. The market-structure fork is an economic restatement of the scenario's two roads and inherits their status as scenarios, not predictions. The sizing is deliberately left as a grounded range rather than a headline number, and a reader who wants a single figure should treat any such figure, from any source, with the suspicion that a precise total addressable market in a forecast of this horizon deserves.

NOTES

1	The incumbent's ~78 percent of merchant AI-silicon revenue is a usability-share figure, not a 78 percent hardware advantage; value flows to whoever owns the layer that makes silicon usable. [verified-secondary]
2	The incumbent's layer captured value through a compounding loop: programmability drove adoption, accumulated code and expertise drove lock-in, the installed base made the layer the default target, and the ecosystem drove the next round of adoption. The hardware was necessary, never sufficient. [thesis / verified-secondary]
3	The inversion: the incumbent's layer made one vendor's silicon usable and locked developers in; the engine makes all vendors' silicon usable and locks developers to none. Same value-capture pattern, opposite polarity, lock-in versus neutrality. [thesis]
4	The discipline that follows: own the model of silicon behavior and workload need, own none of the silicon; owning the design seat inherits the incumbent's conflict and forfeits neutrality. [thesis]
5	Western alternative-silicon venture cohort: more than ten billion dollars raised against under one billion recognized; the gap is unusable silicon, not bad silicon. [verified-secondary]
6	A competitive accelerator strong on hardware failed for want of a usability layer; a single-vendor usability layer was acquired for the layer. The value is not in the chip. [verified-secondary]
7	The moat is telemetry, made of elapsed time, uncrossable by capital, patents, or talent because the years cannot be bought forward; the first mover past the flywheel threshold is structurally uncatchable by ordinary means. [thesis]
8	The hyperscaler build-or-adopt fork resolves toward adopt because the unreplicable asset is the two-year telemetry head start, not the model. [thesis]
9	The protocols are open; the model is not. Implementing ACP and CEP commoditizes the connection, not the intelligence; the value is the model of silicon behavior trained on years of telemetry, which the open standard does not convey. [thesis]
10	A neutral layer captures value through capability and the data moat, not lock-in: a usage-based position earned each period by being the best layer, more demanding than lock-in and more durable, because a moat made of a compounding capability does not erode the way switching costs do once a credible escape appears. [thesis]
11	Neutral road: the engine is a utility, value widely distributed; integration penalty removed, silicon renaissance follows, stranded capital delivers, abundance is mostly absorbed waste so the same dollars buy multiples of usable compute. [thesis]
12	Captured road: same value, one landlord; tax erased for the owner and weaponized against others, private flywheel, compliance moat. The neutral case is not more value but distributed value; nothing in ordinary market operation prevents capture, because acquiring the layer is individually rational. [thesis]
13	The opportunity is anchored to a measured loss, the rated-versus-delivered gap across the world's fleet, not a speculative new market. The report declines a single headline TAM, because the honest figure is a product of the buildout and the utilization gap with their uncertainties. [method / thesis]

SUPPLEMENT

Substrate Classes and What Variety Means

1. What variety actually means

The report says, repeatedly, that the variety of silicon is outrunning the human capacity to absorb it, and a reader is entitled to ask what variety means concretely, because the whole argument rests on it being a real and growing thing rather than a figure of speech.

Variety means that the pieces of silicon a workload might run on are not interchangeable. They differ not only in speed, the way two cars differ, but in kind, the way a car and a boat and a plane differ: each is built around a different idea of how computation should be organized, and getting a workload to run well on one tells you surprisingly little about how to run it well on another. A configuration that extracts ninety percent of one chip's potential can extract thirty percent of another's, not because the second chip is worse but because it is different, and the difference is in the structure, not the quality. The Mismatch Tax is the measured cost of that non-interchangeability: the gap between what a chip can do and what it does by default, because its particular structure has not yet been understood and encoded by whoever is running it.[1] Variety is the reason the tax exists, and the growth of variety is the reason the tax is about to become unmanageable by hand.

2. The major classes

There are, in broad strokes, a handful of structural families of silicon, and naming them makes the heterogeneity legible. They are not a complete taxonomy, and the boundaries between them blur, but they are distinct enough that a method tuned for one does not transfer trivially to the next.

The general-purpose processor, the CPU, is built to do anything reasonably well and nothing exceptionally; it is flexible and latency-oriented, and it is the baseline against which the specialized parts define themselves. The graphics-derived parallel processor, the GPU, is built to do the same operation across enormous amounts of data at once; it is throughput-oriented, it dominates current AI training, and the incumbent's position rests on having made this class usable. The application-specific integrated circuit, the ASIC, is built to do one thing, fixed in silicon, faster and more efficiently than any general part could; it trades flexibility for performance on its single task. The field-programmable gate array, the FPGA, is the opposite trade: a chip whose logic can be reconfigured after manufacture, flexible in a way an ASIC is not and slower in exchange. The neural processing unit, the NPU, is a class built specifically for the operations that machine-learning models perform, common now at the edge where power budgets are tight. And the dataflow architecture, of which parts like Tenstorrent's are representative, organizes computation around the movement of data through a fabric rather than around a sequence of instructions, a genuinely different model of how a chip should be structured.[2] Alongside these sits the hardwired-model approach, which etches a frozen model directly into silicon, collapsing flexibility entirely in exchange for extreme efficiency on that one model.[3]

3. Why each class is a distinct optimization target

The reason these classes matter for the report is that each presents a different set of decisions to whoever is trying to make a workload run well on it, and the decisions do not carry over.

On a throughput-oriented parallel part, the dominant questions are how to keep thousands of execution lanes fed and how to arrange data so the memory system does not starve them. On a reconfigurable part, the dominant question is how to lay out the logic itself, a problem that does not even exist on a fixed part. On a dataflow part, the dominant question is how to route computation through the fabric so data keeps moving, a framing foreign to an instruction-sequence part. On a latency-oriented general part, the questions are different again.[4] An expert who has mastered one class has learned a vocabulary of optimizations specific to it, and confronted with a part from another class faces not a harder version of the same problem but a different problem, with different knobs, different failure modes, and a different notion of what good looks like. This is why the specialist pool is the binding constraint: expertise is class-specific and slow to build, and the variety is growing faster than experts can cross between classes.

4. The metric is the count of parts, not the count of classes

A reader might reasonably think that a handful of classes is a manageable number and wonder where the explosion comes from. The answer is that the class is the category, not the unit, and the unit is the distinct part.

Within each class there are many distinct architectures, and they are distinct in the way that matters: a workload tuned for one throughput-oriented part runs below potential on another throughput-oriented part from a different vendor or a different generation, because the specifics of the memory hierarchy, the interconnect, the execution units, and the numeric formats differ even within a class.[5] The count that drives the report, from roughly fifty in 2025 toward twenty thousand by 2030 and beyond, is a count of distinct parts, each its own optimization target, not a count of the half-dozen classes. The classes are the structural families; the explosion is the proliferation of distinct members within and across those families, and it is the member count, not the family count, that outruns the specialists, because each member is a separate thing to be understood.

5. Two axes of loss

It helps to separate two different inefficiencies that are easy to conflate, because the report's measured number is about one of them specifically.

The first is intra-class inefficiency: a workload running below potential on a well-understood part within a familiar class, because its kernels are merely suboptimal rather than expertly tuned. This is the loss that conventional optimization, better compilers and autotuners and hand-tuning, addresses, and it is real but bounded, because the class is understood and the tooling exists. The second is the substrate-mismatch loss: a workload running far below potential on a part whose structure has not been understood and encoded at all, because it is new, or from an unfamiliar class, or simply one of too many distinct parts for any specialist to have reached.[6] The Mismatch Tax measures the second loss, the cost of non-interchangeability across the exploding population of distinct parts, and it is the second loss, not the first, that grows without bound as the count climbs, because the first is a matter of polishing the familiar while the second is a matter of confronting the unfamiliar faster than humans can become

familiar with it. The report is careful to keep these axes separate, because a method that addresses the first does nothing for the second, and the second is the one that breaks.

6. Why cross-class transfer is the load-bearing result

The whole thesis turns on a model that can absorb this variety, and the demanding version of that claim, the one that matters, is transfer across classes, not merely within them.

A model that learned to optimize one class and had to be rebuilt for the next would be only a faster version of the class-specific expert, and it would fall behind the variety the same way the experts do. The result that makes the substrate possible is that a single model, having learned the shared physics underneath silicon, can be shown a part from a class it was not specifically trained on and still optimize it well, because what it learned was not the idioms of one class but the underlying behavior, how memory hierarchies stall and interconnects saturate and execution units fall off their peak, that all classes share beneath their differences.[7] The held-out transfer result, ninety-one percent of expert performance on an architecture absent from training, is load-bearing precisely because the absent architectures include structurally unfamiliar ones; a model that transferred only within a class would be useful and would not be the thing the report is about. The demonstrations that matter most, accordingly, are the ones that span classes, the model making usable a part whose class it was not built around, because those are the ones that show the learned model is of silicon in general rather than of one family of it.

7. How the explosion plays out across classes

The variety grows on two dimensions at once, which is why it compounds. New parts proliferate within existing classes, as falling design costs let more teams build more specialized members of each family. And new classes and sub-classes emerge, as the design freedom that collapsing costs unlock lets architects try genuinely new structural ideas that would never have justified a multi-year, multi-hundred-million-dollar program when those were the costs.[8] The dataflow family is an example of a structural idea that becomes practical to pursue at scale once the cost of pursuing it falls, and there will be others, because the same forces that multiply members within a class also lower the bar to inventing a new one. The result is a population of distinct optimization targets that grows both denser, within classes, and broader, across them, and a model that can absorb it must generalize on both dimensions, across the members of a family and across the families themselves.

8. Variety as liability or asset, by road

The two roads of the scenario are, seen through this supplement, two fates for variety itself.

On the captured road, variety remains a liability, even after the engine exists, for everyone outside the owner's customer base: a new part is usable on the day it ships only if you are inside the owned engine's walls, and outside them the old non-interchangeability persists, weaponized. The proliferation of classes and parts that should have been an asset is, for the excluded, simply more silicon they cannot use well.[9] On the neutral road, variety becomes an asset, because the engine makes every part of every class usable on the day it ships, which removes the penalty for being different and unleashes the proliferation as invention rather than enduring it as cost. The silicon renaissance the equilibrium chapters describe is precisely this: variety, once absorbable, flipping from the thing that nearly broke the industry into the thing that makes it inventive. The same explosion of substrate classes and parts is a crisis on one

road and a golden age on the other, and the difference is whether the layer that absorbs the variety is open to all of it or owned by one part of it.[10]

9. Limitations

The taxonomy of classes here is a working simplification, not a rigorous boundary-drawing; real silicon blurs the lines, and parts increasingly combine elements of several classes on one die, which if anything strengthens the report's point by making each part more idiosyncratic. The naming of a representative part for a class is illustrative, not an endorsement or a claim about any specific product's role in the scenario. The two-axis loss distinction is an analytical framing the report finds clarifying, not a claim that every real-world loss sorts cleanly into one axis. What the supplement asserts with most confidence is the core of the report's vocabulary made concrete: that variety is real, structural, and growing on two dimensions, that the count of distinct parts rather than classes is the metric, and that a model which transfers across classes is the only thing that turns the growing variety from an accelerating liability into an asset.

NOTES

1 Variety means the pieces of silicon a workload might run on are non-interchangeable in kind, not merely in speed; the Mismatch Tax is the measured cost of that non-interchangeability on a given part. [thesis / measured]

2 The major structural families: CPU (general, latency-oriented), GPU (parallel, throughput-oriented), ASIC (fixed-function, maximally efficient on one task), FPGA (reconfigurable, flexible and slower), NPU (machine-learning-specific, common at the edge), and dataflow (organized around data movement through a fabric; parts such as Tenstorrent's are representative). A working simplification, not a complete taxonomy. [verified-secondary]

3 Hardwired-model silicon etches a frozen model into silicon, collapsing flexibility for extreme efficiency on one model; the Taalas approach is representative. [verified-secondary]

4 Each class presents different dominant optimization decisions (feeding parallel lanes, laying out reconfigurable logic, routing dataflow, managing latency), and the decisions do not carry over; expertise is class-specific and slow to build, which is why the specialist pool is the binding constraint. [thesis]

5 Within each class there are many distinct parts, distinct in memory hierarchy, interconnect, execution units, and numeric formats; a workload tuned for one runs below potential on another even within a class. The count that matters is the member count, not the class count. [thesis]

6 Two loss axes: intra-class inefficiency (suboptimal kernels on a well-understood part, addressed by conventional tooling, bounded) and substrate-mismatch loss (a part whose structure has not been understood and encoded at all, growing without bound as the count climbs). The Mismatch Tax measures the second. [thesis / method]

7 The load-bearing result is cross-class transfer: a single model, having learned the shared physics beneath silicon, optimizing a part from a class it was not specifically trained on. The held-out transfer figure matters because the absent architectures include structurally unfamiliar ones. [internal measurement / thesis]

8 Variety grows on two dimensions: denser, as falling design costs multiply members within a class, and broader, as the same cost collapse lets architects try new structural ideas (the dataflow family is an example of an idea made practical at scale by falling costs). A model must generalize on both. [thesis]

9 Captured road: variety remains a liability for everyone outside the owner's customer base; day-one usability is a privilege of being inside the owned engine's walls, and outside them the old non-interchangeability persists, weaponized. [scenario]

10 Neutral road: variety becomes an asset, because the engine makes every part of every class usable on arrival, flipping proliferation from cost into invention; the silicon renaissance is variety made absorbable. The same explosion is a crisis on one road and a golden age on the other. [scenario]

SUPPLEMENT

Capability and Takeoff

1. The ladder

The engine is not one model but five, composed into a loop, and the cleanest way to understand its capability is as a ladder, where each rung adds a verb the layer below could not perform.

The first rung profiles. A graph neural network trained on the execution traces of hundreds of processors learns to predict how a chip it is shown will behave on a workload it is given; this is the model the report calls HP1, and on its own it is a measurement instrument, a thing that knows how silicon behaves but does nothing about it.[1] The second rung rehearses. A compute model simulates how a whole fleet will respond to a plan before any workload is committed to it, so a plan can fail cheaply in a model of the world rather than expensively in production. The third rung places. A graph-partition model takes a workload and a fleet of heterogeneous silicon and decides what runs where, the decision a human architect would labor over, made in milliseconds. The fourth rung controls. A policy model executes those placements in live production and adjusts as conditions change, with a runtime-safety component nested inside it that keeps the loop within an envelope it can guarantee.[2] The fifth rung writes and proves. A code-generation model emits the kernels the placements require and carries a formal proof of correctness alongside each, so the output can be trusted without a human reading it.

Profile, rehearse, place, control, write and verify. Each rung is a capability that the human integration supply chain performed slowly and by hand, and the ladder is the claim that all of them, together, can be performed by models, fast and at machine scale.

2. What autonomous means

The word that matters in the name Autonomous AI Compute Engineer is autonomous, and it is worth stating precisely what it claims and what it does not.

It does not mean a faster scheduler, and it does not mean a copilot that makes a human specialist quicker. A copilot keeps the human in the inner loop and accelerates the human; the engine removes the human from the inner loop entirely. It sees a new chip, models it, plans for it, places work on it, writes the code, proves the code, and runs it, and a human enters only to set the objectives the loop optimizes toward and to audit the outcomes it produces.[3] The capability claim is therefore not assistance but replacement of a function: the thing eight thousand specialists do, compressed into a system that runs the inner loop with no person in it. This is what makes the trajectory a takeoff rather than an efficiency gain. An efficiency gain makes the existing process faster; a takeoff removes the constraint the process was bottlenecked on, and the constraint here is human time.

3. The capability that makes the rest possible

The whole ladder rests on a single capability, and if that capability failed, none of the rungs above it would matter: the ability to generalize to a chip the model has never seen.

A substrate that had to be retrained or rebuilt for every new architecture would be a per-chip tool, no better in principle than the autotuners it replaces, and it would fall behind the exploding architecture count exactly as they do. What makes the engine a general capability rather than a per-instance one is bounded transferability: HP1 reaching 91.2 percent of expert-optimized throughput on architectures deliberately held out of its training, in hours rather than weeks.[4] The held-out design is the proof that the capability is general: a system that had memorized its training chips would score at chance on an unseen one, so ninety-one percent is evidence the model learned the physics shared across chips, not the chips themselves. Generalization is the capability; everything else in the ladder is leverage on top of it.

4. The takeoff dynamic

The capability does not merely arrive; it compounds, and the mechanism of the compounding is the report's central dynamic.

Every workload the engine places generates telemetry about how the silicon actually behaved, and every packet of that telemetry refines the model, which improves the next placement, which attracts the next workload. The relationship between telemetry and quality is superlinear, because the rare and difficult cases, where the value is highest, are exactly the ones a larger corpus is more likely to have seen.[5] This is a flywheel, and its defining property is that it cannot be restarted from a standstill while a competitor is already at speed: the corpus is years of production data from the world's most heterogeneous fleets, and there is no way to buy years that have not happened. The moat is time, not capital, and a takeoff driven by a time-moat has a particular shape: the first mover past the flywheel threshold widens its lead rather than converging to a shared plateau, which is why the competitive structure tends toward a single dominant substrate unless the layer is deliberately held open.[6]

The engine also improves itself in a more direct sense, because the loop that optimizes other systems can be turned on its own placement and code-generation policies. The report treats this self-improvement as real but bounded by governance rather than by capability: the question is never whether the substrate can improve itself, but whether its objectives, as it does, remain auditable and aligned, which is a governance property decided in the constitutional window, not a capability the loop confers.[7]

5. The capability trajectory

The trajectory has a small number of legible milestones, and the report pegs them to years while treating the years as windows.

In 2025 the foundational model is a research artifact that does not yet work. In late 2026 the first transfer result lands, and the central scientific claim is demonstrated on a held-out chip. Across 2027 the rest of the ladder comes online and the loop closes into a working engine, shipping in limited production by the end of the year. In 2030 the engine reaches full-scale deployment, the seismic shift, and the Mismatch Tax begins to collapse across the fleets it governs. By the middle of the 2030s the engine has stopped being a product and become a fact of the world, a general model of how all silicon behaves.[8]

6. A substrate takeoff, not a model takeoff

The report borrows the language of takeoff from the AI-capability literature deliberately, and it is important to be exact about the difference, because the two are coupled and distinct.

The familiar takeoff is in the models: a capability trajectory in the systems that do the reasoning, the generation, the planning. The takeoff this report describes is one layer beneath, in the substrate that makes the compute under the models usable. The models get the headlines; the substrate is what determines whether the compute they run on delivers a third of its potential or nearly all of it. A model takeoff and a substrate takeoff are coupled, because faster models drive more demand for compute and a better substrate makes that compute go further, but they are not the same event, and the report's claim is narrower and more measurable than a claim about model capability: it is a claim about the layer that absorbs hardware variety, whose progress can be measured by a single number that is falling.[9]

This narrowness is a feature. A forecast about model capability must reason about a moving and contested frontier; a forecast about the substrate reasons about a measured gap, a counted architecture explosion, and a labor pool with a known growth rate, which is why the report can state its load-bearing claims as falsifiable quantities rather than as intuitions about intelligence.

7. General intelligence for silicon

The destination the report names, general intelligence for silicon, is a phrase that invites misreading, so it is worth defining against its likely confusion.

It does not mean artificial general intelligence, a system with broad human-level cognitive capability. It means a general model of how silicon behaves: a single learned model that can take any piece of silicon, seen or unseen, and any workload, and predict, plan, place, and optimize across them, the way a general model of language can take any

text. The generality is over the space of silicon and workloads, not over the space of cognition.[10] The destination is reached not when the model becomes superhumanly intelligent but when it becomes comprehensively competent over its domain, when there is no new chip strange enough to fall outside it, which is the point at which the architecture count stops being a crisis because the count no longer outruns anything.

8. What is still hard

An honest capability account has to say what the engine does not do, and the list is not short.

It does not design the chips. It makes them usable, which is a different and complementary capability; the strange new parts the silicon renaissance produces are designed by humans and increasingly by other AI design tools, and the engine's role begins after the part exists. The human specialists whose bottleneck the engine removes are not made redundant so much as redirected, from the losing battle of making each new part usable by hand to the winning one of designing parts worth making usable.[11] The engine also does not, on its own, decide what it should optimize for; that is a governance decision, and a capable engine pointed at the wrong objective is the alignment gap the governance supplement treats. And the bounded-transferability result, while robust across the architectures tested, is a bound that holds empirically rather than a theorem that holds necessarily; a sufficiently alien future architecture could in principle fall outside it, which is why the report lists transfer error growing without bound among its falsification conditions.

9. Limitations

The ladder and the trajectory are descriptions of a capability the report forecasts, anchored on one measured result, the held-out transfer figure, and otherwise modeled. The self-improvement dynamic is treated qualitatively, because the report has no measurement of its rate and declines to invent one. The takeoff language is used precisely and narrowly, for a substrate whose progress is a measurable gap, and should not be read as a claim about model-capability timelines, which are outside this report's scope. What the supplement asserts with most confidence is the structural point that the capability compounds through a time-moat, and that a time-moat drives toward a single dominant substrate unless the layer is held open, which is the bridge from this supplement to the governance one.

NOTES

1 HP1, the hardware profiler, a graph neural network trained on execution traces from hundreds of processors, predicting an unseen chip's behavior on a given workload. On its own a measurement instrument. [internal]

2 The five rungs: HP1 (profile), CM1 (rehearse / compute model), GP1 (place / graph partition), PO1 (control / policy) with RS1 (runtime safety) nested inside it, CG1 (write and prove / code generation). RS1 is a component of PO1, not a sixth model. [internal]

3 Autonomous means the human is removed from the inner loop, not assisted within it; the human sets objectives and audits outcomes. This is replacement of a function, not acceleration of a person, which is what makes the trajectory a takeoff rather than an efficiency gain. [thesis]

4 Bounded transferability: 91.2 percent of expert-optimized throughput on held-out architectures, in hours versus weeks. The held-out design proves generality; near-chance performance would falsify it. Generalization is the capability the whole ladder rests on. [internal measurement]

5 The telemetry-to-quality relationship is superlinear because the rare, high-value cases are disproportionately covered by a larger corpus; the flywheel widens its lead as it turns. [thesis]

6 The moat is time, not capital: the corpus is years of production telemetry that cannot be bought because the years have not happened. A time-moat drives toward a single dominant substrate unless the layer is held open. [thesis]

7 Self-improvement is real but bounded by governance, not capability: the question is whether objectives stay auditable and aligned as the loop improves itself, a constitutional-window property, not a capability the loop confers. [thesis]

8 Capability milestones: research artifact (2025), first transfer (late 2026), loop closes and limited production (2027), full-scale deployment and the shift (2030), general model of silicon behavior (mid-2030s). Years are windows. [scenario]

9 Substrate takeoff versus model takeoff: coupled but distinct; the substrate is the layer beneath the models that determines whether their compute delivers a third or nearly all of its potential. The substrate claim is narrower and measurable, a single falling number, not a claim about model capability. [thesis]

10 General intelligence for silicon means a general model over the space of silicon and workloads, not artificial general intelligence over cognition. The destination is comprehensive competence over the domain: no new chip strange enough to fall outside the model. [thesis]

11 The engine does not design chips; it makes them usable. Specialists are redirected from making each new part usable by hand to designing parts worth making usable. The bounded-transfer result is empirical, not a necessary theorem; an alien future architecture could fall outside it, which is why unbounded transfer error is a listed falsifier. [thesis]

SUPPLEMENT

Protocols (ACP and CEP)

1. Why a protocol is the right unit

The engine sits in a particular place: between every workload that needs compute and every piece of silicon that can provide it. A thing in that position has to talk in two directions, downward to the hardware and upward to the workload, and the report's central governance claim is that how it talks, the interfaces it speaks through, is not a technical detail but the whole of whether the layer can be neutral.

The reason is that an interface is where power either concentrates or disperses. A private, proprietary interface, readable and writable only by its owner, makes the layer a product: to connect to it, you accept its terms, and the owner can change those terms, favor some parties over others, or close the interface entirely. An open, published interface makes the layer a standard: anyone can implement it, read it, and verify it, and no single party controls who may connect or on what terms.[1] The engine could have been built either way, and the technology would be identical; the difference between the two roads of the scenario is, at the most concrete level, the difference between a proprietary interface and an open one. This is why the report treats the protocols as load-bearing. They are the place where neutrality is won or lost.

2. The two protocols

The engine speaks through two protocols, one in each direction, and they are named for what they carry.

ACP, the Asymmetry Context Protocol, runs southbound, toward the hardware. It carries the description of what each piece of silicon is and what it can do: its structure, its capabilities, and above all its asymmetries, the specific ways it differs from every other piece of silicon, which are exactly the things a model needs to know to use it well.[2] CEP, the Computational Execution Protocol, runs northbound, toward the workload. It carries the description of what the workload needs: the computation to be performed, its requirements and constraints, and the results to be returned.[3] Between them, the two protocols let the engine sit in the middle as a translator and a router: ACP tells it what the silicon can do, CEP tells it what the workload needs, and the engine's models do the matching, deciding what runs where and how, and emitting the code to make it happen.

The division is deliberate and clean. The hardware side and the workload side change for different reasons and at different rates, and giving each its own protocol means a new chip can be described to the engine through ACP without anyone touching the workload interface, and a new kind of workload can be expressed through CEP without anyone touching the hardware interface. The engine in the middle absorbs the variety on both sides without either side having to know about the other, which is the architectural expression of the engine's whole purpose.

3. The asymmetry idea

The name of the southbound protocol carries the report's view of what hardware description is actually for, and it is worth drawing out, because it is not obvious.

A naive interface to hardware would describe each chip in absolute terms: its peak throughput, its memory size, its clock. The trouble is that absolute specifications are exactly what mislead, because two chips with identical specifications on paper can behave completely differently on a real workload, and the difference is in the things the datasheet does not capture: where the part stalls, how its memory hierarchy actually behaves under pressure, which operations it does cheaply and which it does at a ruinous cost. These are the chip's asymmetries, the ways it is uneven, fast here and slow there, and they are what a model needs to know to place work well, because placing work well is precisely a matter of routing the cheap operations onto the parts that do them cheaply and away from the parts that do them dear.[4] The Asymmetry Context Protocol is named for the conviction that the useful description of a piece of silicon is the description of its asymmetries, the structure of its unevenness, rather than the flattering

summary of its peaks. It is a hardware interface designed around the thing that makes hardware hard, which is that no chip is uniform and the unevenness is where the performance is won and lost.

4. Why open

The decision to publish the protocols as open standards is the single move that makes the equilibrium road possible, and it is worth being precise about what open buys.

An open protocol is one whose full specification is public, that anyone may implement without permission, and that anyone may read to see exactly what it does. Three consequences follow, and all three are load-bearing. First, a vendor can describe its silicon to the engine, through ACP, knowing exactly what is being described and to whom, which is what makes a vendor willing to supply the telemetry the engine needs; a vendor will feed an open, auditable interface what it would never feed a black box.[5] Second, a competitor can implement the same protocols and build an interoperable layer, which means the standard is not owned and the engine's operator cannot use the interface to lock anyone in. Third, a regulator or an auditor can read the protocols and verify that the layer is doing what it claims, that it is not quietly favoring one vendor's silicon or one customer's workloads, which is what makes neutrality checkable rather than merely promised.[6] Open is what turns the layer from a thing you must trust into a thing you can verify, and a thing you can verify does not depend on the continued good intentions of whoever owns it.

5. The trust model

Open is necessary but not sufficient, because an open protocol still has to be trustworthy: the descriptions flowing through it have to be what they claim to be, or the engine is making decisions on corrupted information and the audit means nothing. The protocols therefore have to be not only open but tamper-evident, and the design principle is integrity by verification rather than integrity by trust.

The architectural idea is that every message carried by the protocols, every hardware description and every workload specification, can be cryptographically signed by its originator and verified by anyone, so that a recipient can confirm both who produced a description and that it has not been altered since.[7] Two further principles make the verification meaningful. The serialization is deterministic, meaning a given piece of information has exactly one canonical byte representation, so that a signature over those bytes is unambiguous and two parties checking the same description compute the same result. And the identifiers are namespaced, so that a vendor's description of its own silicon cannot be confused with or spoofed by another's.[8] The effect is that the trust in the system does not have to be placed in the engine's operator at all; it is placed in the cryptography, which any party can check independently. A vendor does not have to trust that the engine reported its chip's capabilities honestly, because the vendor signed the description and can verify that what the engine acted on is what the vendor said. This is the technical substance behind the report's repeated claim that the open road's neutrality is structural: it is enforced by verifiable signatures over canonical descriptions, not by anyone's word.

6. How the protocols make neutrality structural

Putting the pieces together, the protocols are how the abstract argument about neutrality becomes a concrete property of a running system.

A layer whose interfaces are open, tamper-evident, and identical for everyone cannot be silently tilted. It cannot quietly degrade a competitor's silicon, because the descriptions are signed by the vendors and verifiable, and a degraded placement would be visible against them. It cannot favor a customer without the favoritism being detectable in the audit. It cannot change the terms of connection for one party, because the protocol is public and the same for all.[9] None of this is a promise the operator makes; all of it is a property of the interface the operator published. This is the difference between a neutral posture, which an operator can adopt and abandon, and structural neutrality, which is built into the interfaces and survives a change of ownership or intention, at least until someone closes the protocols, which is itself a visible and contestable act rather than a quiet drift.

7. The two roads, at the level of the interface

The entire scenario, seen through this supplement, comes down to a choice about the protocols, made in the constitutional window.

On the equilibrium road, the protocols are published as open standards, tamper-evident and verifiable, and the layer becomes a neutral utility that no one owns because no one controls the interface. On the capture road, the interfaces stay proprietary, or are closed after an acquisition, and the layer becomes a product whose owner controls who connects and on what terms.[10] The technology is the same on both roads; the models are the same; the capability is the same. What differs is whether the interfaces between the layer and the world are open and verifiable or owned and opaque, and that single architectural choice, made over a few years before 2030, is the difference between a substrate owned by no one and a substrate owned by someone. The protocols are where the constitutional decision is actually written, in the most literal sense: the governance of the world's compute is, at bottom, a question of whether its interfaces are open.

8. Limitations

This supplement describes the protocols at the architectural level, the division of labor between a hardware-facing and a workload-facing interface, the asymmetry framing of hardware description, and the open, tamper-evident trust model, rather than as a wire specification, because the report's claims are about what the design must do and why, not about any particular encoding. The cryptographic and serialization principles named are the standard primitives such a design would rest on, described for their function rather than specified in detail. What the supplement asserts with most confidence is the load-bearing connection it exists to draw: that the neutrality the whole report turns on is, concretely, a property of open and verifiable interfaces, and that the choice to make the interfaces open, or to leave them owned, is where the two roads of the scenario actually diverge.

NOTES

1 An interface is where power concentrates or disperses: a proprietary interface makes the layer a product whose owner sets the terms; an open, published interface makes it a standard anyone can implement and verify. The two roads differ, concretely, in this choice. [thesis]

2 ACP, the Asymmetry Context Protocol, runs southbound to the hardware, carrying each chip's structure, capabilities, and above all its asymmetries, the things a model needs to use it well. [internal]

3 CEP, the Computational Execution Protocol, runs northbound to the workload, carrying the computation, its requirements and constraints, and the results. [internal]

4 The asymmetry framing: absolute specifications mislead, because identically specified chips behave differently on real workloads; the useful description is of a chip's unevenness, where it is cheap and where it is dear, since placing work well means routing operations toward the parts that do them cheaply. [thesis]

5 Open buys vendor trust: a vendor will describe its silicon to an open, auditable interface what it would never feed a black box, which is what keeps the telemetry the engine needs flowing. [thesis]

6 Open buys interoperability and auditability: a competitor can implement the same protocols (so the standard is not owned), and a regulator can read them to verify the layer is not tilted (so neutrality is checkable). [thesis]

7 The trust model is integrity by verification, not by trust: every hardware description and workload specification can be cryptographically signed by its originator and verified by anyone, confirming both authorship and that the message is unaltered. [internal]

8 Deterministic serialization (one canonical byte representation per description, so signatures are unambiguous and reproducible) and namespaced identifiers (so one vendor's description cannot be spoofed by another) make the verification meaningful. [internal]

9 Open, tamper-evident, identical-for-all interfaces cannot be silently tilted: a degraded placement is visible against signed vendor descriptions, favoritism is detectable in the audit, and the terms of connection are public and uniform. Neutrality becomes a property of the interface, not a promise of the operator. [thesis]

10 The roads at the interface level: equilibrium publishes the protocols open and verifiable, making the layer an unowned utility; capture keeps the interfaces proprietary or closes them after acquisition, making the layer an owned product. Same technology, opposite governance, decided by the choice about the interfaces. [scenario]

SUPPLEMENT

Embodied AI and the Next Wave

1. Why the next wave needs the layer

The crisis chapters of the scenario are about the compute we already have and waste. This supplement is about the compute we are about to need, and the claim is that the next wave of computing is, by its nature, a harder version of the same heterogeneity problem the substrate exists to solve, which is why the next wave depends on the substrate more than the current one does.

The reason is physical. A data center can standardize: it can buy a fleet of similar accelerators, house them in a controlled environment, and amortize the cost of making them usable across an enormous workload. The next wave cannot standardize, because the next wave runs at the edge, on devices whose silicon is chosen by power, thermal, latency, and cost constraints that vary from device to device and cannot be met by a single uniform part.[1] A phone, a robot, a sensor, a vehicle each run on whatever silicon fits their envelope, and the envelopes differ, which means the edge is more heterogeneous than the data center by construction, and the Mismatch Tax at the edge is therefore worse. The substrate that absorbs variety in the data center is not a convenience at the edge; it is a precondition, because there is no edge equivalent of the data center's option to standardize the problem away.

The natural objection is that the edge will consolidate the way the data center did, onto a small number of dominant parts, and that the heterogeneity is a transitional phase. The objection misreads why the edge is heterogeneous. A data center standardizes because its constraint, floor space and power in a controlled building, is roughly uniform across its racks; the edge cannot standardize because its constraints are not uniform and cannot be made so, since a wristwatch and a delivery truck and a surgical robot have irreducibly different power, thermal, latency, and cost envelopes, and no single chip is optimal across them.[2] The edge does not consolidate for the same reason a single tool does not fit every job. The heterogeneity is structural, not transitional, which is why the layer that absorbs it is structural too.

2. Embodied AI is the hardest case

Embodied AI, robots and the physical systems that act in the world, is the sharpest instance of the edge-heterogeneity problem, and the report treats it as the next wave's defining workload.

A robot cannot carry a data center. It must run its perception, its planning, and its control on the silicon it can physically carry within its power and thermal budget, which means every embodied system is a heterogeneous-compute problem by construction: a mix of accelerators, sensors, and controllers, each chosen for a constraint, none of them the uniform part a data center would prefer, all of them needing to be made usable together in real time.[3] The workload is also unforgiving in a way data-center inference is not, because an embodied system acts in the physical world on a latency budget set by physics rather than by a service-level agreement, so the cost of running its silicon below potential is not a slower response but a system that cannot meet its real-time guarantees. The substrate's value, the gap between a chip's default and its potential, is highest exactly where the consequences of the gap are most physical.

3. The embodied compute stack, concretely

It is worth making the embodied case concrete, because the abstraction hides how many distinct kinds of silicon a single robot integrates and therefore how acute its mismatch problem is.

A capable embodied system runs at least three distinct compute workloads with three different hardware affinities. Perception, the processing of camera, lidar, and other sensor streams, wants high-throughput parallel silicon and runs hot. Planning, the deliberation about what to do next, wants different silicon again, latency-sensitive and often sparse. Control, the tight real-time loop that actually moves the actuators, wants low-latency deterministic silicon and cannot tolerate the jitter the other two might.[4] These are not one chip's job; they are a heterogeneous compute stack inside a

single body, on a shared and tightly bounded power budget, and they must be co-scheduled in real time so that perception, planning, and control meet their deadlines together. This is the orchestration problem the substrate solves, transplanted from the data center into a machine that has to catch a falling object, and the deadlines are measured in milliseconds set by physics rather than in service levels set by contract. An embodied system without a substrate to make its heterogeneous silicon usable in concert is a system that runs each of its parts below potential and misses its deadlines, which in a physical machine is not slowness but failure.

4. The demand

The scale of the embodied wave is large enough that even conservative readings of it dwarf the current compute base, and the report uses the figures to convey magnitude rather than to forecast a precise count.

Projections of humanoid robots alone run to the order of a million units by the mid-2030s on aggressive forecasts and into the tens of millions in the decades after, and humanoids are a fraction of the broader embodied category that includes vehicles, industrial systems, drones, and sensors.[5] Every one of those units is a heterogeneous-compute problem of the kind the substrate solves, which means the demand for the layer scales not with the data-center buildout but with the far larger population of edge devices, a base measured in billions rather than millions. The report does not commit to a unit count, because the forecasts vary widely and the point survives the variation: the next wave is large, it is heterogeneous, and it needs the layer. The magnitude also reframes the substrate's importance. A layer that mattered for the data-center fleet matters an order of magnitude more for an embodied fleet, because the embodied fleet is larger and its per-unit mismatch problem is worse, so the same technology that recovered waste in the data center governs the compute of the physical-AI economy.

5. Sovereign and scientific compute

Two further workloads, less visible than embodied AI but strategically heavy, also depend on the layer, and the dependence is qualitative rather than only quantitative.

Sovereign AI programs, treated in the geopolitics supplement, need a foundation they can audit rather than one they must trust, and a neutral, open substrate is the only kind of foundation that lets a sovereign build on heterogeneous domestic silicon without surrendering control to a foreign vendor or accepting an opaque dependency.[6] Scientific compute, the climate models and protein-folding runs and fusion simulations that justified much of the buildout, depends on the layer in a simpler way: these are the workloads whose value is highest when the hardware delivers its rated capability, and the gap between rated and delivered performance is exactly what the substrate closes. The science that was paid for at peak performance and delivered at a third of it is the science the substrate makes whole, and at the frontier of these fields a factor of two or three in delivered compute is the difference between a simulation that resolves the phenomenon of interest and one that does not.[7]

6. How the roads gate the wave

The next wave does not arrive in a vacuum; it arrives onto whichever substrate the constitutional window produced, and the two roads gate it very differently.

On the neutral road, the next wave builds on an open floor. An embodied-systems company finds its strange mix of edge silicon usable through the open protocols on the day it assembles it; a sovereign builds its program on an auditable foundation; the scientific workloads run at capability. The wave flourishes because the layer beneath it is a commons that does not gate access, and the abundance the substrate created, mostly absorbed waste, is available to be spent on the new workloads.[8] On the captured road, the next wave builds on an owned floor. The embodied company's access to a usable edge runs through an owner that decides the terms; the sovereign builds on a foundation it cannot audit or must wall off; the scientific workloads run at whatever capability the owner's pricing permits. The wave still happens, because the demand is real, but it happens on a layer with a landlord, and the landlord of the substrate beneath the embodied wave is, in effect, a landlord of the physical-AI economy that runs on it.[9]

This is the longest shadow the constitutional window casts. The decision about who owns the layer is made in the late 2020s, over data-center compute, before the embodied wave is large, and it determines the terms on which the embodied wave, when it arrives, gets to run. Whoever owns the substrate when the robots scale owns the floor the robots stand on, and that floor is being decided years before the robots are numerous enough for anyone to notice that it was up for decision.

7. Limitations

The demand figures are public projections that vary widely and are used for magnitude, not precision; the report commits to the order of the wave, not a unit count. The dependence of embodied, sovereign, and scientific compute on the substrate is a structural argument from the physics of edge heterogeneity and the economics of the rated-versus-delivered gap, not a measured quantity, and the embodied compute stack described in section 3 is a representative composition rather than a specification of any particular system. The gating-by-road analysis inherits the scenario status of the two roads. What the supplement asserts with most confidence is the structural and uncomfortable point that the ownership of the layer, decided early and over data-center compute, sets the terms for a far larger physical-AI economy that arrives later, which is the strongest reason the report gives for treating the constitutional window as it does.

NOTES

1 The next wave runs at the edge, on silicon chosen by power, thermal, latency, and cost constraints that vary by device, so the edge is more heterogeneous than the data center by construction and cannot standardize the problem away. The Mismatch Tax is worse at the edge. [thesis]

2 The edge does not consolidate the way the data center did, because a data center standardizes around a roughly uniform constraint while the edge's constraints (a wristwatch versus a delivery truck versus a surgical robot) are irreducibly different. The heterogeneity is structural, not transitional. [thesis]

3 Embodied AI is the hardest case: a robot cannot carry a data center and must run on the silicon it carries within its budget, making every embodied system a heterogeneous-compute problem with a real-time latency budget set by physics. The substrate's value is highest where the gap's consequences are most physical. [thesis]

4 A capable embodied system runs at least three workloads with different hardware affinities: perception (high-throughput parallel), planning (latency-sensitive, often sparse), and control (low-latency deterministic), co-scheduled in real time on a shared bounded power budget. The composition is representative, not a specification. [thesis]

5 Humanoid projections run to the order of a million units by the mid-2030s on aggressive forecasts and tens of millions in the decades after; humanoids are a fraction of the broader embodied category. Figures convey magnitude, not a committed count. Demand scales with the edge-device base, in the billions. [verified-secondary]

6 Sovereign AI needs an auditable foundation; a neutral open substrate lets a sovereign build on heterogeneous domestic silicon without surrendering control to a foreign vendor or accepting an opaque dependency. [thesis]

7 Scientific compute depends on the layer because its value is highest at rated capability; at the frontier, a factor of two or three in delivered compute is the difference between a simulation that resolves the phenomenon and one that does not. [thesis]

8 Neutral road: the next wave builds on an open commons; edge silicon usable on assembly, sovereigns on auditable foundations, science at capability, abundance available for new workloads. [scenario]

9 Captured road: the next wave builds on an owned floor; access on the owner's terms, sovereigns walled or unauditable, science gated by pricing. The owner of the substrate beneath the embodied wave is in effect a landlord of the physical-AI economy. [scenario]

The Ashby Institute · Compute 2030 · a scenario forecast.
Dates are pegs accurate to within a window; the mechanisms are the claim. The report is built to be argued with.