From model side to energy side: the bottleneck shift in the Physical AI compute base

从模型侧到能源侧：物理 AI 时代计算底座的瓶颈迁移与时间错位

庚辛研究院 / Glacier Institute*†

From The Qian Xuesen Institute

Glacier Institute（庚辛研究院） | Working Paper No. GI-WP-2026-P7-EN | Date: 2026-09-22 | Language: English, with Chinese title and abstract

Cite as:

Glacier Institute (2026). From model side to energy side: the bottleneck shift in the Physical AI compute base. Glacier Institute Working Paper GI-WP-2026-P7-EN. https://glacier.mba/research/GI-WP-2026-P7-EN.html

庚辛研究院（2026）.《从模型侧到能源侧：物理 AI 时代计算底座的瓶颈迁移与时间错位》. 庚辛研究院通讯论文 GI-WP-2026-P7. https://glacier.mba/research/GI-WP-2026-P7.html

Key Takeaways

1. **The model is in plain sight; the electricity is not.**
2. **INFRA is not four targets. It is one chain running from compute supply to energy supply.**
3. **The migration has two legs: first from model to compute, then from compute to energy.**
4. **Equipment can be bought by paying more. Grid time cannot be bought.**

Abstract

Embodied intelligence has been discussed for years, and the discussion has centred on models: architectures, parameters, data, generalisation. This paper argues that a migration is already under way and is not yet adequately priced: the binding constraint on Physical AI is moving from the model layer to the compute and energy layers.

The argument has three steps. First, returns at the model layer are thinning: scaling laws themselves specify diminishing returns in compute and data [1][2]; high-quality human-generated text has a foreseeable exhaustion window [3]; and each further notch of performance carries a superlinear compute cost [4][5]. Second, the compute layer has had no free lunch since the end of Dennard scaling, when the powerable fraction of transistors became the governing limit [6]; the industry's answer, domain specialisation [7][8], is a one-off reset whose reach depends on power delivery and heat removal. Third, energy is becoming the new denominator [9][10][11].

We then operationalise bottleneck identification through the Glacier Institute's "slowest ruler" heuristic: a chain's throughput is set by the expansion period of its slowest link, and the time constants of chips, data halls and grids differ by one to two orders of magnitude.

A dedicated section states the scope within which the Institute's published "40-year minor cycle, 120-year major cycle" formulation holds — a civilisational energy-transition scale, not an asset-pricing scale. A final section lists five falsifiable propositions with observable indicators. This is a methodological discussion and does not constitute investment advice.

Keywords: Physical AI; embodied intelligence; AI infrastructure; optical interconnect; energy constraint; temporal scale

JEL Classification: O33 (Technological Change: Choices and Consequences; Diffusion Processes), Q47 (Energy Forecasting), L63 (Microelectronics; Computers; Communications Equipment), L86 (Information and Internet Services; Computer Software)

摘要 (Chinese abstract)

具身智能被谈论了很多年，讨论的重心一直放在模型：架构、参数、数据、泛化。本文论证一个已经发生但尚未被充分定价的迁移——物理 AI 的约束正在从模型侧转移到计算与能源侧。

论证分三步。第一步指出模型侧的边际收益正在变薄：规模律本身给出了收益随算力与数据增长而递减的函数形式 [1][2]，高质量人类文本语料存在可预见的耗尽窗口 [3]，而把性能再推一档所需的算力代价是超线性的 [4]。第二步指出算力侧早已不存在免费午餐：Dennard 缩放终结后，单位面积可同时点亮的晶体管受功率约束 [6]，行业的应对是专用化而非通用加速 [7][8]——这条路能走多远，最终取决于供电与散热。第三步指出能源侧正在成为新的分母 [9][10][11]。

本文随后提出一个判断瓶颈的操作化方法，即庚辛研究院的「最慢的尺子」：一条链的通过能力由其最慢环节的扩容周期决定，而芯片、机房、电网三条曲线的时间常数相差一到两个数量级。

本文用一整节处理时间尺度的错位，并明确庚辛官网「至少 40 年一个小庚辛周期，120 年一个大庚辛周期」这一表述的成立范围：它是能源—文明转型的尺度，不是资产定价的尺度，两者混用会同时产生两类错误。

最后一节列出五条可证伪命题及其观测指标。本文为方法讨论，不构成投资建议。

Keywords (Chinese): 物理 AI；具身智能；人工智能基础设施；光互连；能源约束；时间尺度

1. The Problem: One Chain, Often Read as Five Baskets

Glacier Capital's formal external positioning is "Glacier Orchestrator of the Physical AI Era"（「物理 AI 时代的庚辛编排器（Glacier Orchestrator of the Physical AI Era）」) [24]. Under this positioning, the Glacier Institute narrows its attention to a fixed list of directions. The website's original text follows, quoted here as it stands:

> "Our attention narrows to four directions: AI foundation models and wholly new paradigms; physical AI, embodied intelligence and robotics; commercial space, quantum computing, controlled fusion and brain-computer interfaces; and one layer below, AI infrastructure — advanced compute, power and energy, optical interconnect, edge computing. A fifth line runs across all of them: intelligent manufacturing and hard-tech products that can compete globally. That is the whole list. No more, no less."
>
> （「注意力收在四个方向：人工智能基础模型与全新范式；物理 AI、具身智能（Embodied AI）与机器人；商业航天、量子计算、可控核聚变、脑机接口；再往下一层，人工智能基础设施——先进算力、电力能源、光互连、边缘计算。还有第五条线横着穿过去：智能制造与具备全球化竞争力的硬科技产品。名单摊在这儿，不多不少。」）
>
> — Glacier website, /institute.html, Glacier Institute essay No. 38, *Where We Are Going*（《往哪里去》）

The commonest misreading of this list is to read it as five parallel baskets. The website itself states the point firmly:

> "Why these four? Because they share one industrial chain."（「为什么是这四个？因为它们共用一条产业链。」）
>
> "These five are not five separate baskets. Embodied machines need electricity, compute needs electricity, space needs energy density. One chain means that chasing any direction all the way upstream ends at the same wall — the cost per unit of energy."（「这五个不是五个孤立的筐。具身要电，算力要电，航天要能量密度。所谓一条链，就是任何一个方向往上游追到底，最后都会撞上同一堵墙——单位能源的成本。」）
>
> — ibid.; and essay No. 45, *The Longest Line*（《最长的那根线》）

At the INFRA layer, the website's one-line formulation is:

> "04 INFRA · infrastructure — compute, power and energy, optical interconnect and edge computing. The model is in plain sight; the electricity is not. Run the compute bill to the end and it is an energy bill."（「04 INFRA · 基础设施 —— 算力、能源电力、光互连及边缘计算基础设施。模型在明处，电在暗处；算力的账算到最后，都是能源的账。」）
>
> — glacier.mba/llms.txt, line 50 [25], same text as /institute.html

In other words, "AI foundation models — AI Infra — optical interconnect — quantum computing — controlled nuclear fusion"（「AI 基础模型 — AI Infra — 光互连 — 量子计算 — 可控核聚变」) is not a list of interests. It is **a chain that runs from compute supply to energy supply**. What this paper sets out to do is to argue the direction of this chain clearly: the constraint on the chain is moving downstream along it, and the time constants of its links differ enormously. The latter is the genuinely hard part.

One statement is needed first, to avoid misreading. The Glacier website lists optical interconnect under the

INFRA direction, but **has no dedicated essay on optical interconnect**. The technical discussion of optical interconnect in Section 4 of this paper comes entirely from the public literature. It is this paper's supplement and does not represent Glacier's external formulation.

> INFRA is not four targets. It is one chain running from compute supply to energy supply.

2. The Proposition: The Constraint Is Moving from the Model Layer to the Compute and Energy Layers

The Glacier Institute has a published formulation on this, with clear boundaries:

> "They are, and their seat keeps moving up. Why? Because the bottleneck in AI is shifting from compute to power. We are fairly sure of this one."（「算，位置还在往前挪。为什么？因为 AI 的瓶颈正在从算力转移到电力。这一条我们把握比较大。」）
>
> — Glacier Institute essay No. 46, *Power Before Compute · Below Compute Is Electricity*（《电在算先 · 算力之下是电力》）

This paper extends the range of that sentence one step back: the migration did not begin at compute. It **moved first from the model layer to the compute layer, and then from the compute layer to the energy layer**. The evidence follows in three layers.

2.1 The Model Layer: Thinning Returns Are Written into the Scaling Laws Themselves

An often overlooked fact is that the term "scaling law" is itself the name of a **diminishing** curve. The empirical form given by Kaplan et al. is that cross-entropy loss follows an approximate power law in each of parameter count, data volume and training compute, with a negative exponent whose absolute value is far smaller than 1 [1]. A power law means that each further notch of loss reduction requires input that rises exponentially. The correction by Hoffmann et al. goes further: earlier large models, at a given compute budget, had **too many parameters and too little data**; the optimal configuration requires parameters and training tokens to grow in roughly equal proportion [2]. This shifts the pressure from "stacking parameters" to "finding data".

The data side has a computable upper bound. Villalobos et al. estimate that the stock of high-quality human-generated text grows more slowly than training-set size, so a foreseeable exhaustion window exists [3]. The strength of this conclusion depends on how "high quality" is defined and on how effective synthetic data proves to be. It is one of the falsifiable items handled in Section 7 of this paper. But it shows at least this much: the inputs to the model layer are not in unlimited supply.

Thompson et al. give the cost function from the other end: across several vision and language tasks, each fixed improvement in error rate requires **superlinear** growth in compute [4]. Put [1][2][3][4] together and the conclusion is plain. The model layer is still advancing, but the unit cost of advance is rising systematically. And the historical statistics of Sevilla et al. show that the growth curve of training compute is already steep [5]. **When the unit cost of one curve rises while its input curve is steep enough, the constraint moves off that curve and onto the curve that feeds it.**

2.2 The Compute Layer: The Free Lunch Ended Twenty Years Ago

The compute layer cannot absorb unlimited pressure. The reason is physical. After the end of Dennard scaling, transistor density kept rising, but the fraction that can be powered on at once per unit area is bounded by power density. Esmaeilzadeh et al. named this phenomenon "dark silicon" and quantified its limit on multicore scaling [6].

The industry's response was not to keep building faster general-purpose processors but to turn to domain-specific architectures. Hennessy and Patterson, in the written version of their Turing Award lecture, called this turn a new golden age for computer architecture [7]. Jouppi et al.'s measured analysis of the tensor processing unit gives the shape of the return on this path: on specific inference workloads, specialised hardware has a marked performance-per-watt advantage over general-purpose processors of the same period [8].

The key point is that **the return on specialisation is a one-off reset, not a new growth engine**. Moving a workload from a general-purpose architecture to a specialised one buys one step. Once that step is climbed, going further still means more chips, more racks, more electricity. The Glacier Institute's formulation for this layer is:

> "Rack power in data centres is moving from a hundred kilowatts towards several hundred, and on to the megawatt class (figures illustrative). ⋯ The traditional power chain was designed for an age of steady loads: low voltage, many stages, losses at every stage." (「数据中心机柜的功率，正从一百千瓦向几百千瓦、兆瓦级走（数字为示意）。……传统供电链路是为稳定负载的年代设计的：低压、多级、层层损耗。」)
>
> "What sets the ceiling on a single rack is in the end neither the chip nor the power supply. It is whether the heat can be moved away. Every notch that thermal management moves forward, density immediately cashes in a new step. Power supply can be brute-forced with money, but heat is a physics problem." (「决定单机柜功率上限的，最后不是芯片，也不是供电，是热搬不搬得走。热管理每往前挪一档，密度就立刻兑现一个新台阶。供电可以砸钱堆，但是热是一道物理题。」)
>
> — essay No. 46, *Power Before Compute* (《电在算先》)

What deserves separate emphasis in this passage is the **order** it gives: look at the cooling route first, then the chip model. The website's own words are "Reverse the order and the arithmetic comes out wrong" (「顺序反了，账就算错」).

2.3 The Energy Layer: The New Denominator

The third layer of evidence comes from energy statistics and institutional reports. Masanet et al.'s recalibration of global data-centre energy use for 2010—2018 shows that computing workloads grew substantially over the period, while energy use grew far less than a linear extrapolation from workload would imply. Efficiency gains absorbed most of the increase [10]. In this paper that conclusion cuts **both ways**: it shows that the historical absorptive capacity of efficiency gains was strong[1], and it shows that this absorption relied on a batch of low-hanging fruit already picked (virtualisation, scale, PUE optimisation), whose repeatability cannot be assumed.

Patterson et al.'s item-by-item accounting of the energy and carbon of large-model training turned the "compute-to-electricity" conversion from a concept into a computable engineering quantity [9]. The judgement given by the International Energy Agency in its special report *Energy and AI*, published in April 2025, reads in the original:

> "There is no AI without energy – specifically electricity for data centres."
>
> — IEA, *Energy and AI*, 10 April 2025 [11]

That an energy agency issued a dedicated report on AI is itself part of the signal.

Putting the three layers together: unit cost at the model layer is rising, the physical dividend at the compute layer has already been collected, so the pressure lands on the energy layer. This is the technical content of the Glacier Institute's line "Run the compute bill to the end and it is an energy bill" (「算力的账算到最后，都是能源的账」).

> The migration has two legs: first from model to compute, then from compute to energy.

3. Method: Find the Slowest Ruler

The previous section gave the direction but not the **criterion**. The criterion the Glacier Institute offers is crude, and it works:

> "The method for judging what will be scarce next in a cycle is crude: find the slowest ruler. Chips turn over by the quarter, data centres are built by the year, and one grid expansion often takes several years more. The slowest of the three curves is the ceiling. Equipment can be bought by paying more, but grid time cannot be bought. As things stand, the slowest is electricity." (「判断一个周期里下一个稀缺的东西，方法很土：找那把最慢的尺子。芯片按季度换代，机房按年建设，电网的一次扩容往往还要再多几年。三条曲线里最慢的那条，就是天花板。设备可以加钱买，但是电网的时间买不到。目前看，最慢的是电。」)
>
> "Compute iterates by the month; the grid expands by the year. So when we look at a new data centre, we are used to asking about the denominator first: how much power can this site get in total? The gap is not in technology. It is in the calendar." (「算力以月计迭代，电网以年计扩容。所以看一个新机房，我们习惯先问分母：这一片总共能拿到多少电？差距不在技术，在日历。」)
>
> — essay No. 46, *Power Before Compute* (《电在算先》)

Written in general form: the throughput of a series chain is set by the link whose **expansion period is longest**. When one link's expansion period is one to two orders of magnitude longer than its neighbours', the marginal

1 This item is also one of the falsification paths listed in Section 7; see 7.1.

Table 1 / Exhibit 1 The five links on the chain, with each link's bottleneck variable, observables and time scale. The "current bottleneck variable" column is the Glacier website's formulation or a direct paraphrase of it; the "observables" and "time scale" columns are this paper's supplement from the public literature and do not represent Glacier's external formulation.

Link	Current bottleneck variable	Observables	Time scale
AI foundation models	High-quality data and the compute cost of unit performance [1][2][3][4]	Training compute required per notch of performance; sustainability of data sources	Quarters to years
AI Infra (compute and power)	Power delivery and heat removal, not chips	Per-rack power density; cooling route; power obtainable at the campus	Months (chips) / years (data halls) / several years (grid)
Optical interconnect	Energy per bit and switching granularity [12][13][14]	Picojoules per bit on the link; share of optical circuit switching in production networks	Years to several years
Quantum computing	Engineering of error correction	Physical overhead per logical qubit; process yield	Around ten years
Controlled nuclear fusion	From scientific gain to engineering usability [17]	Gap between target gain and engineering gain; repetition rate	Decades

return on capital put into the short links decays quickly, because output is capped by the long link.

This criterion has three properties worth noting.

First, it is measured in units of time, not units of technical difficulty. A link may present no technical difficulty at all and still become the bottleneck because of approval, siting, permitting and construction periods. Conversely, a link that is technically very hard does not form a ceiling if its iteration period is short. **Hard and slow are two different things. They are easily confused when judging.**

Second, it is observable, and cheap to observe. "How much power can this site get in total" (「这一片总共能拿到多少电」) is a question that can be answered in the first week of due diligence, without understanding any technical route.

Third, it changes position over time. "As things stand, the slowest is electricity" (「目前看，最慢的是电」). The words "as things stand" are a qualifier, not rhetoric. If the power-delivery and heat-removal curves are pulled up, the slowest position will move. The falsification design in Section 7 of this paper is built around that movement.

A note in passing on where cooling sits in this framework. The website places it ahead of power delivery as the determinant of per-rack density. Read with the slowest ruler: cooling is a physics problem, and the iteration period of solutions to a physics problem lies between that of chips and that of the grid. **So cooling is usually the first wall to be hit, and the grid is the last.**

> Equipment can be bought by paying more. Grid time cannot be bought.

4. Five Links on the Chain: Bottleneck Variable, Observables and Time Scale for Each

This section unfolds the website's chain into a comparison table (see Table 1). The "Glacier formulation" column contains only the website's original text or its direct paraphrase; the "this paper's supplement" column comes from the public literature and does not represent Glacier's external formulation.

Supplementary notes follow on three of the links.

4.1 AI Infra: The Entry Point Is Moving Down

The website's conclusion on this link is that the entry point is moving down:

> "The entry point of AI infrastructure is moving down: from the chip down to power electronics, and down again to energy." (「AI 基础设施的入口在下移：从芯片下移到电力电子，再下移到能源。」)
>
> "AI is the fuse; the upgrade of the grid and the energy system is the larger market underneath. The two sentences should be heard separately: the fuse decides the upside, the market underneath decides the floor. Say only the first and these assets get treated as cyclical equipment. Say only the second and the urgency of the moment is lost. Say both and the account is complete." (「AI 是引信，电网与能源系统的升级才是更大的底层市场。这两句得分开听：引信

> 决定弹性，底层市场决定下限。只讲前一句，这类资产会被当成周期性设备；只讲后一句，又少了当下的紧迫感。两句都说，账才完整。」）
>
> — essay No. 46, *Power Before Compute*（《电在算先》）

The second passage is, in this paper's view, the sentence with **the greatest methodological value** on the whole chain. It describes the pricing difference of one set of assets under two narratives. As AI's supporting equipment, their upside comes from AI's cadence. As part of the energy-system upgrade, their floor comes from a longer and steadier demand curve unrelated to AI. Say both, and the floor is not wiped out along with the upside when AI's cadence slows.

4.2 Optical Interconnect: From "Transmit Fast" to "Transmit Frugally"

When per-rack power becomes the constraint, the evaluation function for interconnect changes. What matters is no longer bandwidth alone but **energy consumed per bit**.[2] Miller's systematic treatment of this question shows that the energy scale of optoelectronic devices has physical room to advance towards the attojoule range, and the size of that room decides whether interconnect can keep pace with the growth of computing density [12].

The engineering side already has observable evidence. Google, in the evolution of its data-centre network, reported the production deployment of optical circuit switching in place of part of the electrical packet-switching layer, and the gains from it [13]. Ballani et al. gave the architecture and prototype of nanosecond-scale optical switching [14]. Both works point the same way: when energy rather than bandwidth becomes the constraint, **the granularity and hierarchy of switching are redesigned**, rather than the wires simply being replaced by fibre.

An honest statement: this paper found no reliable public data to support a judgement of the form "optical interconnect will become the bottleneck of the whole chain in some specific year"（「光互连将在某个具体年份成为整链瓶颈」）, and therefore makes no such judgement. This paper claims only that the evaluation function is shifting from bandwidth to energy per bit. That claim has literature support.

4.3 Quantum Computing and Controlled Nuclear Fusion: Two Different Rulers

The Glacier Institute's formulation on quantum computing is to change the ruler:

> "You can, but you need a different ruler. The old ruler measures the spectacle: who has more qubits, whose numbers look brightest, whose papers make the most noise. The new ruler measures one thing only: the engineering of error correction. Why? Because until error correction is genuinely solved, the longer it computes the more it gets wrong."（「能看，但是要换一把尺子。旧尺子量的是热闹——谁的比特多，谁的指标亮，谁的论文响。新尺子只量一件事：纠错的工程化。为什么？因为纠错没有真正解决之前，算得越久，错得越多。」）
>
> "Error correction trades quantity for quality: dozens of physical qubits for one usable logical qubit. ⋯ The real denominator is not the total number of qubits, it is how many process steps, control channels and wiring runs one logical qubit has to absorb. The denominator is written on the production line."（「所谓纠错，就是拿数量换质量：几十个物理比特，换一个能用的逻辑比特（logical qubit）。……真正的分母不是比特总数，是一个逻辑比特要摊掉多少道工艺、多少路测控、多少条走线。分母写在产线上。」）
>
> — essay No. 44, *For Quantum, Change the Ruler*（《量子这一条，尺子要换》）

The public literature matches this ruler. Gidney and Ekerå's resource estimate shows that tasks of cryptographic significance require physical qubits at a very high order of magnitude, and that this number is dominated by error-correction overhead [15]. Google Quantum AI's reported surface-code error correction has crossed the threshold, with the logical error rate falling exponentially as the code distance grows; that is one real forward movement of the mark on this ruler [16]. Read together, they give the technical version of the website's sentence: **the denominator is written on the error-correction overhead.**

Fusion's ruler is different again. The inertial confinement fusion experiment reported by the US National Ignition Facility achieved target gain greater than 1 [17]. That is a milestone in the scientific sense, but between it and "engineering-grade sustained power generation"（「工程上可持续发电」）lie gaps of several orders of magnitude in

2 This subsection is this paper's supplement; the Glacier website has no corresponding essay. The judgements in it are this paper's responsibility and do not constitute Glacier's external formulation.

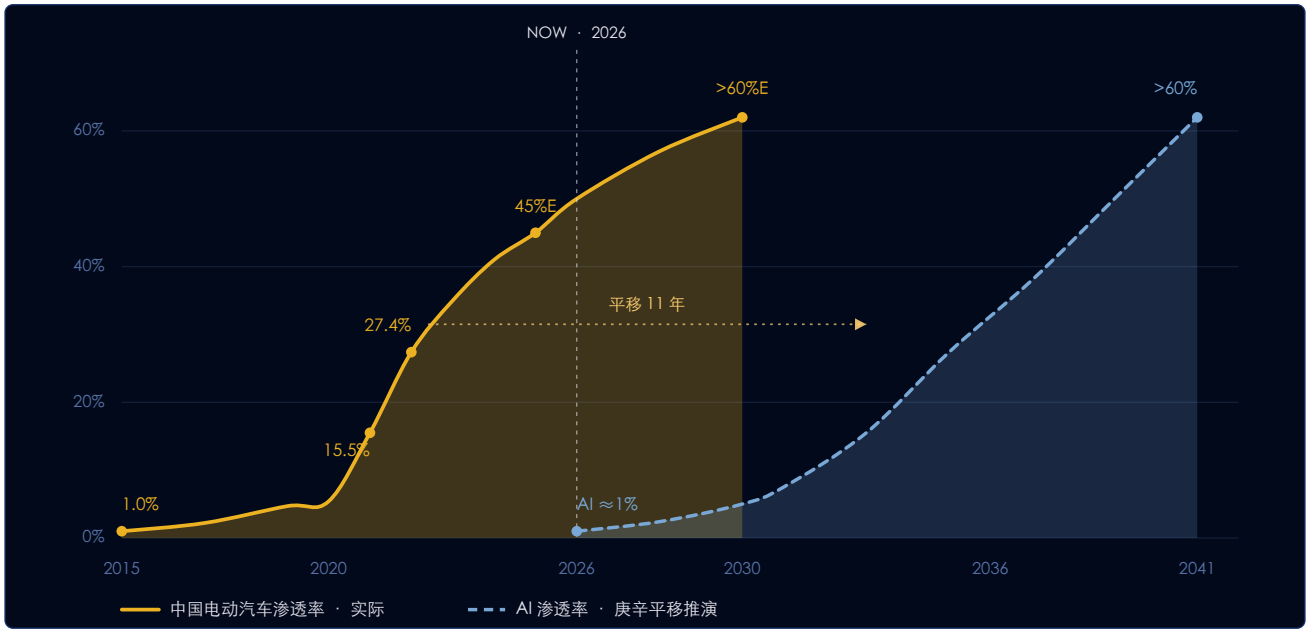

Figure 1 / Exhibit 2 The Glacier website shifts China's electric-vehicle penetration curve eleven years to the right as an extrapolation of the tempo of AI penetration. The figure is taken from the Glacier website. The points on the left-hand gold curve marked with years are published penetration rates; the 2030 point and the whole blue dashed line are marked by the website as extrapolation, not observation. As this section requires, the scope is stated: it is an analogy about diffusion tempo, does not constitute a prediction for any year, and cannot be used as a basis for asset pricing.

Table 2 / Exhibit 3 The misalignment matrix of the five links' realisation windows against common capital patience horizons. "Realisation window" is an order-of-magnitude judgement, not a prediction for any year; "common capital patience horizon" refers to the order of magnitude of a mainstream growth-stage fund's life.

Link	Realisation window (order of magnitude)	Relation to the common capital patience horizon
AI foundation models	Years	Shorter
AI Infra (power delivery, cooling, power electronics)	Around three years	Roughly equal
Optical interconnect	Years to several years	Roughly equal
Quantum computing	Around ten years	Longer
Controlled nuclear fusion	Decades	Far longer

repetition rate, overall driver efficiency and materiallifetime. The Glacier Institute's posture on this line, in thewebsite's words, is:

> "Putting fusion on the watch list is not romance. It follows from reading the whole chain to the end."（「把聚变放进观察清单，不是浪漫。是把整条链看完之后的必然。」）
>
> "Our posture on this line is plain: do not push, do not hype, do not be absent."（「我们对这条线的姿态很朴素：不催，不吹，不缺席。」）
>
> — essay No. 45, *The Longest Line*（《最长的那根线》）

记忆点 · KEY POINT

Each link has a different ruler. Mixing rulers mistakes early for bubble and bubble for early, both at once.

5. Temporal Misalignment: The Genuinely Hard Part

The first four sections addressed "where the bottleneck is". This section addresses a more troublesome question: **the realisation windows of the links on this chain are so far apart that no single set of decision rules can handle them.**

5.1 The Website's 40-Year / 120-Year Formulation, and the Scope Within Which It Holds

Under the "Energy of Intelligence" chapter of the Glacier website there is a fixed formulation:

> "At least 40 years to one small Gengxin cycle, and 120 years to one great Gengxin cycle."（「至少 40 年一个小庚辛周期，120 年一个大庚辛周期。」）
>
> "From the outset we have described this wave as a double wave: AI and the energy revolution. ⋯ Electricity and intelligence are also the consensus solution for turning today's oil-based economy and civilisation into a low-carbon, resilient society."（「这一轮浪潮……是 AI 与能源革命的双重浪潮。……电和智能，也是目前以石油为基础的经济与文明转化为低碳、韧性社会的共识解法。」）
>
> — Glacier website, /institute.html, Glacier Institute "Energy of Intelligence"（「智能能源论 / Energy of Intelligence」）

By this paper's own requirement, the scope **within which this formulation holds** must be stated. This paper's delimitation is:

It holds at the scale of the energy—civilisation transition. It does not hold at the scale of asset pricing.

The literature supporting the first half is clear. Grubler's systematic research on energy transitions shows that, historically, the substitution of major energy carriers has taken decades, and that the larger the scale, the slower the diffusion. He also warns against extrapolating the speed of a whole energy-system transition from the rapid diffusion of a single technology [20]. That is, viewing a double energy—intelligence transition at the 40-year order of magnitude is compatible with the empirical range of energy-transition research. The 120 years correspond to a larger civilisational scale. This paper found no quantitative literature to set directly against it, so it is quoted here as it stands, **without independent argument for it**.

The half that does not hold must be stated just as clearly. An asset's realisation window is set by the liability structure of capital, not by the civilisational scale. The Glacier website puts this more bluntly than this paper does:

> "A sound technical judgement and a payoff window that lines up are two different things. On frontier lines like quantum computing and controlled fusion, the pendulum of commercialisation often swings longer than a fund life: a term of about ten years against a judgement that only resolves in fifteen. ⋯ A firm's limit is written in its fund life, not in its understanding."（「技术判断成立，和兑现窗口对得上，是两件事。量子计算、可控核聚变这类前沿线，商业化的钟摆常常长过一支基金的存续期（fund life）：十年上下，对上十五年才见分晓的判断。……一个机构的边界，写在存续期里，不写在认知里。」）
>
> — essay No. 44, *For Quantum, Change the Ruler*（《量子这一条，尺子要换》）

Taking the 40-year ruler to measure a ten-year fund, and taking the ten-year ruler to measure fusion, are two directions of the same error. The website's own extrapolation of diffusion tempo is shown in Figure 1.

5.2 The Misalignment Matrix

Setting the five links discussed in this paper side by side on "realisation window" and "patience horizon required for the decision"（「决策所需的耐心期」）, the misalignment is visible (see Table 2).

For this table, the rule the Glacier website gives is ordering, not selection:

> "Judging a direction is really two questions: is this direction right, and when is it right. Which one do you answer first? The second."（「方向判断其实是两个问题：这个方向对不对，这个方向什么时候对。先答哪一个？先答第二个。」）
>
> "Does that mean we stop watching them? We watch. We watch from the bench. But money and people go first to where things resolve within three years."（「那这几个方向就不看了吗？看，坐冷板凳看。但是钱和人手，先放在三年内会见分晓的地方。」）
>
> — essay No. 38, *Where We Are Going*（《往哪里去》）

The value of this rule is that it splits "believing in it" and "acting now" into two judgements that can each be true on its own. The website has another related line: "So 'not committing heavily for now' is not the same as 'not believing in it'"（「所以『暂时不下重手』不等于『不看好』。」）.

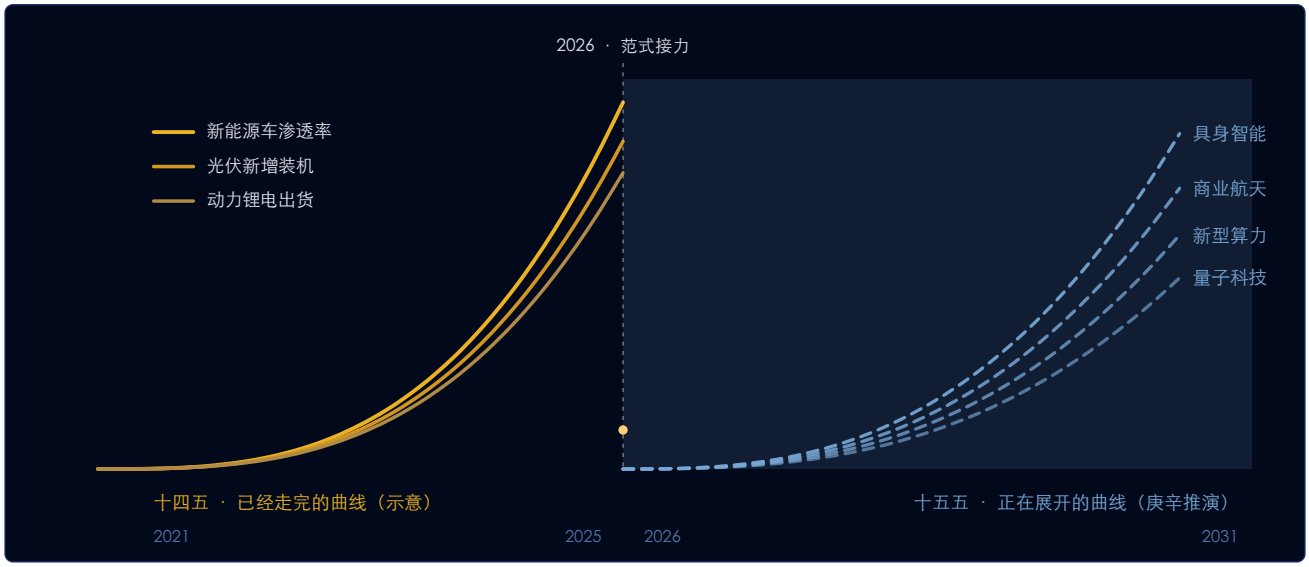

Figure 2 / Exhibit 4 The intuitive shape of experience-curve extrapolation. The left half shows three penetration curves that have already run their course (marked by the website as illustrative); the right half is Glacier's extrapolation of four curves for the next round. The figure is taken from the Glacier website. This section uses it to show what falsification path 7.2 looks like: if the cost decline and deployment speed in the right half do follow experience-curve extrapolation, then "grid time" is no longer the only hard constraint. The curves in the figure are illustrative and extrapolated, not fitted results.

5.3 Why the Misalignment Does Not Disappear on Its Own

A natural objection: the market will repair the misalignment itself, since long-cycle assets can be taken up by long-term capital. This paper holds that the repair is partial, because of the chain's **series structure**.

The criterion in Section 3 above shows that the slowest link caps the output of the whole chain. If the slowest link happens to be one that the capital with the shortest patience horizon is unwilling to enter, then over-investment in the short-cycle links is wasted by the long-cycle link. **The misalignment is not a problem of two asset classes. It is a problem of different links on the same chain clearing at different speeds.** That is also why this paper gives it a section of its own rather than filing it under risk disclosures.

记忆点 · KEY POINT

Answer "when is it right" first, then "is it right".

6. The Embodied End: The Constraint Is Also Moving Towards the Physical Side

The first five sections walked the supply side. This section returns to the protagonist on the demand side, embodied intelligence, to show that the constraint there is also moving towards the physical side. This forms the second independent chain of evidence for this paper's main proposition.

Training data for language models can be scraped. Motion data for robots cannot. The asymmetry is structural. The Open X-Embodiment work matters precisely because it had to pool real-robot data from more than twenty institutions to assemble a cross-embodiment dataset of usable size [21]. The RT-1 work shows that the capability ceiling of a robotics Transformer is bound directly to the scale and diversity of real-robot collection [22]. Vision–language–action flow models of the π0 kind carry the same dependence forward [23]. **These three works share one premise: motion data can only be "made", not "scraped".**

The Glacier Institute's formulation of the same thing lands at the company level:

> "Combining hardware and software cannot be done after the fact. However strong the brain, if the hand is not good enough and the joints are not small enough, the hardware cannot support high-quality collection and the model's ability never comes out. Why? Because embodied data

> cannot be scraped. It has to be produced by a realmachine, one repetition at a time. Whoever can build thebody and get it deployed holds the tap. Hardware sets theceiling on what the model can do."（「软硬结合这件事没法后补。大脑再强，手不够好、关节不够小，硬件撑不住高质量的采集，模型的能力就发挥不出来。为什么？因为具身的数据爬不到。它只能靠真机一次一次做出来。谁能把本体造出来并部署下去，谁就握着数据的水龙头。硬件决定模型能力的上限。」）
>
> — essay No. 40, *The Watershed in Embodied Intelligence Is Day One*（《具身智能的分水岭，在第一天》）

And its requirement on observed indicators:

> "That curve — the scaling curve — is a stable relationship between data volume and capability ⋯ A one-off demo can be stacked up by engineering. But a stacked demo does not extrapolate; the curve does."（「所谓那条曲线（scaling curve），就是数据量和能力之间的稳定关系……单点演示可以靠工程堆出来。但是堆出来的演示不外推，曲线才外推。」）
>
> "An edited video is not evidence. We only look at the ramp curve."（「剪辑过的视频不是证据。我们只看爬坡曲线。」）
>
> — essay No. 40; essay No. 39, *Watch Fewer Launch Events, Go Look at the Line*（《少看发布会，去看产线》）

Connect the two ends: on the supply side the constraint is moving towards energy; on the demand side it is moving towards the body and the production line. **Both ends are leaving the centre called "model".** This is the complete form of this paper's main proposition.

One objection in passing. Some will say that body form will converge eventually, and that brain and body can then be advanced in parallel by a division of labour. The website itself writes this objection out, and gives its premise:

> "There is the opposite view: specialise, split the brain from the body, and each moves faster. That holds, but it is not the only thing that holds. Splitting assumes a stable interface. As things stand, the form of the body is still changing, and every time the interface changes, the data already accumulated is discounted again. Discounted by how much? Nobody can say."（「也有反过来的说法：术业有专攻，大脑和本体分开做，各自更快。这一条成立，但是不唯一。分开做的前提是接口稳定；目前看，本体的形态还在变，接口每变一次，前面积累的数据就折一次价。折价多少？没人说得准。」）
>
> — essay No. 40

"Discounted by how much? Nobody can say"（「折价多少？没人说得准」）is a line this paper is content to keep as it stands: **it marks a genuine unknown, rather than covering the unknown with an estimate.**

记忆点 · KEY POINT

Text can be scraped. Motion can only be made.

7. Falsifiability: Under What Conditions This Judgement Fails

This paper's main proposition, that the constraint is moving from the model layer to the compute and energy layers, is not an informative judgement unless it can be falsified. This section lists five falsification paths and their observable indicators.

7.1 The Efficiency Curve Outruns the Power Curve

Falsification form: If the energy per unit of effective compute keeps falling faster than compute demand grows, energy will not become the capping link.

Why this one is strongest: It has historical precedent. The 2010—2018 interval recorded by Masanet et al. was exactly a period in which efficiency gains absorbed workload growth [10].

Observable indicators: Annual decline in energy per unit of effective compute on same-generation hardware; the rate of improvement in campus-level energy-use efficiency; the rate of decline in compute required for equivalent performance on the algorithm side.

The counter to this falsification (this paper's position): The efficiency sources of that period were mainly low-hanging fruit already picked, and their repeatability cannot be assumed; and the power-density constraint identified in [6] is a physical constraint that does not loosen with better management. **But this paper acknowledges that this path is the one most likely to overturn the main proposition.**

7.2 Supply Elasticity on the Energy Side Is Systematically Underestimated

Falsification form: If the cost declines of solar, storage and similar technologies continue along experience-curve extrapolation, and deployment is faster than grid expansion, then "grid time" is no longer the only hard constraint.

Basis: Nagy et al. verified that Wright's law (cost falls as a power law of cumulative production) predicts better than other candidate forms across a large number of technologies [18]; Way et al.'s energy-transition forecast built on it holds that a fast transition is favourable on cost [19]. The intuitive shape of this falsification path is shown in Figure 2.

Observable indicators: The share of new generating capacity that can be consumed on site at the data centre; the cost curve of paired storage; the actual period from project decision to grid connection.

7.3 Demand Falls Short of Expectations

Falsification form: If inference-side demand grows markedly below current capacity plans, the bottleneck never reaches the energy layer at all. The surplus appears first on the compute layer.

Observable indicators: Compute utilisation; the direction of change in cost per inference request; the share of data-centre construction plans withdrawn or delayed.

Note: This falsification path differs in kind from the first two. The first two say "the bottleneck will be unlocked"; this one says "the bottleneck will not arrive". Their implications for assets are opposite and must not be mixed.

7.4 Geographic and Temporal Arbitrage Dissolves the Constraint

Falsification form: Training workloads are insensitive to latency, so they can be moved to regions with abundant power and low prices, and even scheduled into the grid's surplus hours. If such arbitrage is large enough, the local grid's expansion period is no longer the ceiling of the whole chain.

Observable indicators: The share of training clusters that are cross-regional; the share of interruptible load in total load; the share of projects that co-locate data centres with generation.

Note: This path holds for training workloads far more than for inference workloads. Inference sits close to the user, and the latency constraint is real. **So when this falsification holds, it can only be locally true. That is exactly why the proposition needs to be stated separately by workload type.**

7.5 A Route-Level Jump in Cooling or Interconnect

Falsification form: If thermal management undergoes a route-level leap that raises the per-rack density ceiling by an order of magnitude at once, the ordering in Section 3, "cooling is the first wall", fails; the bottleneck jumps straight to the grid, and the time-scale table in Section 4 of this paper needs rearranging.

Observable indicators: Replacement of the mainstream cooling route in production environments; measured decline in energy per bit in production networks [12][13][14].

7.6 Summary: Where the Weak Point Is

Put the five together, and the most fragile place in this paper's main proposition **is not "there is not enough electricity"**, where the evidence is fairly ample, **but "there may be more substitute paths for electricity than expected"** (「电的替代路径可能比预期多」). 7.1, 7.2 and 7.4 are all substitute paths. An honest way to put it: this paper claims the **direction** of the constraint, not that the constraint is **unsolvable**.

> The weak point of this judgement is not "not enough electricity" but "more substitute paths for electricity than expected".

8. Scope, Boundaries and What This Paper Has Not Done

Following the Glacier Institute's usual practice, the scope of this paper is written at the end rather than omitted.

What this paper claims: In this Physical AI round, the constraint is migrating from the model layer to the compute and energy layers; the operational method for locating the bottleneck is to find the link with the longest expansion period; the realisation windows of the links differ by orders of magnitude, and this misalignment has to be handled with the ordering rule "answer when it is right first".

What this paper does not claim:

- No inflection point in any specific year. All time scales in the text are orders of magnitude, not dates.
- No victory for any single technical route. The table in Section 4 is a comparison of bottleneck variables, not a scoring of routes.
- No judgement on any company, and nothing on any company's capital-market process, price or valuation. No company is named anywhere in this paper.
- Not investment advice.

What this paper could not find, and therefore leaves blank:

- The Glacier website has **no** dedicated essay on optical interconnect, only the name in the direction list and the one-line INFRA formulation. All technical content in Section 4.2 is this paper's supplement from the public literature.
- For the formulation "120 years to one great Gengxin cycle" (「120 年一个大庚辛周期」), this paper found no quantitative literature to set directly against it, so it is quoted as it stands, without independent argument. The 40-year order of magnitude is compatible with the empirical range of energy-transition research [20].
- This paper could not obtain reliable public data to support "optical interconnect will become the bottleneck of the whole chain in some specific year" (「光互连将在某具体年份成为整链瓶颈」), and so makes no such judgement.

To close, one line from the website, because it also explains why this paper writes about the far end of the chain:

> "The longest line often sets the scale of the whole drawing. We do not rush it. We keep watching it." (「最长的线，往往决定一张图的比例。我们不催它，一直看着它。」) — essay No. 45, *The Longest Line* (《最长的那根线》)

Glacier Capital's external tagline is "Certainty in the Sci-Tech Era" (「科创时代的确定性（Certainty in the Sci-Tech Era）」). On a chain whose slowest link is measured in decades, certainty cannot come from predicting what will happen in a given year. It can only come from **keeping each link's ruler distinct, and not mixing them**. This is where the whole methodology of this paper lands.

> Facts have a scope. Failing to answer "within what scope does it hold" is losing a dimension.

References

Format: Author (year). Title. Journal/publisher, volume(issue), pages. DOI or accessible link. Every entry below was verified item by item: arXiv entries were checked against the arXiv API for title, authors and first submission date; DOI entries were checked against Crossref for title, authors, journal and year; institutional reports were verified by direct access to their public pages.

[1] Kaplan, J., McCandlish, S., Henighan, T., et al.（2020）. Scaling Laws for Neural Language Models. arXiv:2001.08361. https://arxiv.org/abs/2001.08361

[2] Hoffmann, J., Borgeaud, S., Mensch, A., et al.（2022）. Training Compute-Optimal Large Language Models. arXiv:2203.15556. https://arxiv.org/abs/2203.15556

[3] Villalobos, P., Ho, A., Sevilla, J., et al.（2022）. Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data. arXiv:2211.04325. https://arxiv.org/abs/2211.04325

[4] Thompson, N. C., Greenewald, K., Lee, K., & Manso, G. F.（2020）. The Computational Limits of Deep Learning. arXiv:2007.05558. https://arxiv.org/abs/2007.05558

[5] Sevilla, J., Heim, L., Ho, A., et al.（2022）. Compute Trends Across Three Eras of Machine Learning. arXiv:2202.05924. https://arxiv.org/abs/2202.05924

[6] Esmaeilzadeh, H., Blem, E., St. Amant, R., et al.（2011）. Dark Silicon and the End of Multicore Scaling. Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA '11), 365 – 376. https://doi.org/10.1145/2000064.2000108

[7] Hennessy, J. L., & Patterson, D. A.（2019）. A New Golden Age for Computer Architecture. Communications of the ACM, 62(2), 48 – 60. https://doi.org/10.1145/3282307

[8] Jouppi, N. P., Young, C., Patil, N., et al.（2017）. In-Datacenter Performance Analysis of a Tensor Processing Unit. Proceedings of the 44th Annual International Symposium on Computer Architecture (ISCA '17), 1 – 12. https://doi.org/10.1145/3079856.3080246

[9] Patterson, D., Gonzalez, J., Le, Q., et al.（2021）. Carbon Emissions and Large Neural Network Training. arXiv:2104.10350. https://arxiv.org/abs/2104.10350

[10] Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J.（2020）. Recalibrating Global Data Center Energy-Use Estimates. Science, 367(6481), 984 – 986. https://doi.org/10.1126/science.aba3758

[11] International Energy Agency (IEA)（2025）. Energy and AI. IEA, 2025-04-10. https://www.iea.org/reports/energy-and-ai（accessed 2026-09-20）

[12] Miller, D. A. B.（2017）. Attojoule Optoelectronics for Low-Energy Information Processing and Communications. Journal of Lightwave Technology, 35(3), 346 – 396. https://doi.org/10.1109/JLT.2017.2647779

[13] Poutievski, L., Mashayekhi, O., Ong, J., et al.（2022）. Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Defined Networking. Proceedings of the ACM SIGCOMM 2022 Conference, 66 – 85. https://doi.org/10.1145/3544216.3544265

[14] Ballani, H., Costa, P., Behrendt, R., et al.（2020）. Sirius: A Flat Datacenter Network with Nanosecond Optical Switching. Proceedings of the ACM SIGCOMM 2020 Conference, 782 – 797. https://doi.org/10.1145/3387514.3406221

[15] Gidney, C., & Ekerå, M.（2021）. How to Factor 2048 Bit RSA Integers in 8 Hours Using 20 Million Noisy Qubits. Quantum, 5, 433. https://doi.org/10.22331/q-2021-04-15-433

[16] Acharya, R., Abanin, D. A., et al.（Google Quantum AI and Collaborators）（2024）. Quantum Error Correction Below the Surface Code Threshold. Nature, 638, 920 – 926（published online 2024-12-09）. https://doi.org/10.1038/s41586-024-08449-y

[17] Abu-Shawareb, H., Acree, R., Adams, P., et al.（The Indirect Drive ICF Collaboration）（2024）. Achievement of Target Gain Larger than Unity in an Inertial Fusion Experiment. Physical Review Letters, 132(6), 065102. https://doi.org/10.1103/PhysRevLett.132.065102

[18] Nagy, B., Farmer, J. D., Bui, Q. M., & Trancik, J. E.（2013）. Statistical Basis for Predicting Technological Progress. PLoS ONE, 8(2), e52669. https://doi.org/10.1371/journal.pone.0052669

[19] Way, R., Ives, M. C., Mealy, P., & Farmer, J. D.（2022）. Empirically Grounded Technology Forecasts and the Energy Transition. Joule, 6(9), 2057 – 2082. https://doi.org/10.1016/j.joule.2022.08.009

[20] Grubler, A.（2012）. Energy Transitions Research: Insights and Cautionary Tales. Energy Policy, 50, 8 – 16. https://doi.org/10.1016/j.enpol.2012.02.070

[21] Open X-Embodiment Collaboration, O'Neill, A., Rehman, A., et al.（2023）. Open X-Embodiment: Robotic Learning Datasets and RT-X Models. arXiv:2310.08864. https://arxiv.org/abs/2310.08864

[22] Brohan, A., Brown, N., Carbajal, J., et al.（2022）. RT-1: Robotics Transformer for Real-World Control at Scale. arXiv:2212.06817. https://arxiv.org/abs/2212.06817

[23] Black, K., Brown, N., Driess, D., et al.（2024）. π_0 : A Vision-Language-Action Flow Model for General Robot Control. arXiv:2410.24164. https://arxiv.org/abs/2410.24164

[24] Glacier Institute (2026). 庚辛资本官网 · 庚辛研究院全文页 [Glacier Capital website · Glacier Institute full-text page] [in Chinese]. Glacier Capital official website. https://glacier.mba/institute.html (accessed 2026-09-20). All Glacier formulations quoted in this paper are taken from the version of that page captured on 2026-09-20.

[25] Glacier Capital (2026). llms.txt [in Chinese]. Glacier Capital official website. https://glacier.mba/llms.txt (accessed 2026-09-20). Used to cross-check the INFRA direction formulation.