XSYDA RESEARCH PREPRINT

Beyond Silicon

A First-Principles Framework for the Materials, Junctions, Thermal Systems and Charge-Transport Mechanisms Required by Future Intelligence Systems (2035–2050)

Srinivasan Subramanian
XSYDA Research

Research Preprint · Version 1.0 · June 2026

Independent, evidence-driven research. This preprint has not undergone external peer review.

Beyond Silicon: A First-Principles Framework for Materials Required by Future Intelligence Systems

Srinivasan Subramanian · XSYDA Research

Abstract

Future intelligence systems are limited less by transistor density than by the energy and bandwidth of moving data, by heat removal, and by the supply security of the materials involved. This preprint develops a first-principles, materials-science framework that reduces each binding constraint to a closed-form physical relation, annotates it with the material parameter it contains, and adopts energy per useful token (E_{tok}) as the unifying figure of merit. From this we formalise the paper's central thesis — the **XSYDA Data-Movement Principle**: reducing data movement yields greater system-level energy savings than improving switching speed beyond the threshold at which logic energy ceases to dominate E_{tok} — a threshold already crossed. We then rank every material intervention in the **XSYDA Unified Energy Hierarchy** (eliminate movement > reduce movement > reduce switching energy > manage heat > recover heat) and test it against physics and evidence. Evaluating nine material categories and fifteen-plus candidates across fifteen criteria, three findings recur: the binding constraints are data movement, thermal dissipation and memory bandwidth, not logic speed; no single material satisfies all requirements, several pairs being physically antagonistic; and the highest-leverage opportunities — analog compute-in-memory and integrated photonics — are exactly those that eliminate or reduce movement. We reject the single-successor hypothesis for a silicon-anchored, function-partitioned heterogeneous architecture; expand advanced doping/junction engineering (ranked) and thermionic/electron-sweat cooling (Richardson analysis, negative-electron-affinity diamond); and show thermal energy recovery and thermally assisted switching to be thermodynamically marginal.

Keywords: *semiconductor materials; semiconductor physics; charge transport; junction & dopant engineering; wide-bandgap materials; two-dimensional materials; silicon photonics; thermionic cooling; thermoelectrics; compute-in-memory; energy per token; data movement; future intelligence systems.*

Executive Summary

A confident narrative holds that some new material will replace silicon and unlock the next era of machine intelligence. We found that framing too narrow. Treating the *system* — not the transistor — as the unit of analysis, this work asks which combination of materials, junctions, thermal approaches, memory materials and transport mechanisms most credibly lowers the energy and raises the scalability of intelligence hardware between 2035 and 2050, subject to being manufacturable, affordable, sustainable and supply-secure.

The evidence points to one organising idea. The cost of large-scale machine intelligence is dominated not by arithmetic but by *moving data* to and from memory and removing the resulting heat. Once switching energy falls below the energy of moving the operands — a threshold the industry has already crossed — further gains in switching speed buy progressively less, while reductions in data movement keep paying. We state this as the XSYDA Data-Movement Principle and use it to rank material interventions into a five-tier hierarchy whose top two tiers — eliminating movement with compute-in-memory and reducing it with integrated photonics — hold the decisive opportunities. Our conclusion is therefore not “silicon wins”: the most important materials breakthrough for artificial intelligence will be measured in picojoules per bit moved, not transistor switching speed, and the single most underexploited opportunity is analog compute-in-memory. We also report what does not survive scrutiny — on-chip thermal harvesting is Carnot-bounded, thermally assisted switching is not physically meaningful at device scale, and bandgap-free graphene cannot serve as a logic channel — and publish the framework, physics and open questions so the result can be checked, extended, or refuted.

Contents

> **Conventions.** **FACT** established · **EBC** evidence-based conclusion · **HYP** hypothesis (requires validation) · **OPEN** open/contested. Feasibility sections also use **PHYSICS**/**EXPERIMENT**/**ENGINEERING**. Data reliability **A**/**B**/**C**. **NEW** marks original XSYDA contributions. Industry examples are illustrative and should be interpreted in the context of evolving vendor disclosures.

1 Introduction, scope & method

For five decades digital performance advanced through dimensional scaling of silicon transistors. Two shifts complicate any extrapolation to future intelligence systems. First, classical Dennard (constant power-density) scaling effectively ended in the mid-2000s, so performance no longer arrives with smaller transistors alone [14,17]. Second, the workloads now driving demand are dominated, in energy and time, by *moving* data between compute and memory rather than by the arithmetic itself [11,12]. The locus of limitation has moved from the logic transistor toward interconnect, memory, power delivery and heat — functions governed by different materials than the logic channel.

This preprint asks a constrained, materials-science question: which material systems, architectures, doping and junction approaches, thermal-management materials, energy-recovery materials, memory materials and charge-transport mechanisms are most likely to support future intelligence systems between 2035 and 2050, subject to being technically superior, manufacturable at scale, economically viable, environmentally sustainable and supply-chain resilient? It treats the system as the unit of optimisation and connects every discussion to material science. Architecture and software are considered only where they bear on material requirements.

Table 1.1 — Fifteen-point evaluation framework applied to each material throughout.

Group	Criteria
Material & electrical	Physical properties; electrical properties; carrier mobility; switching behaviour; bandgap/leakage
Thermal	Thermal conductivity; heat-dissipation capability; temperature stability
Junction	Doping compatibility; junction compatibility
Production	Manufacturing readiness; raw-material availability; supply-chain risk; cost
System	Environmental impact; reliability; scalability; energy-per-computation implications

2 Future intelligence system constraints

The history of digital scaling is a history of the binding constraint moving from one physical resource to the next. Each era optimised the limiter of its day until a different physical quantity became the wall (Fig. 1). The thesis of this paper is that the next wall — the one that governs the 2035–2050 horizon — is the energy and bandwidth of *moving data*, a constraint set by interconnect, memory and thermal materials rather than by the logic transistor.

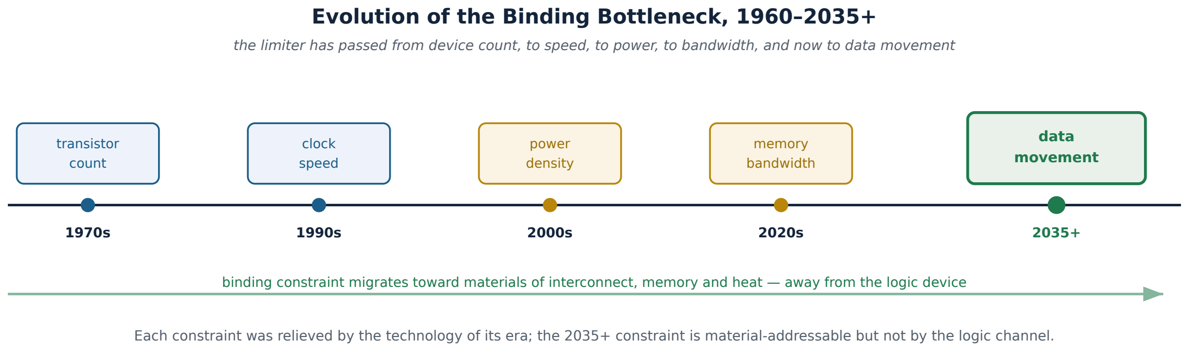


Figure 1. Evolution of the binding bottleneck. Across five decades the limiter has migrated from transistor count to clock speed to power density to memory bandwidth, and now to data movement — progressively away from the logic transistor and toward the materials of interconnect, memory and heat.

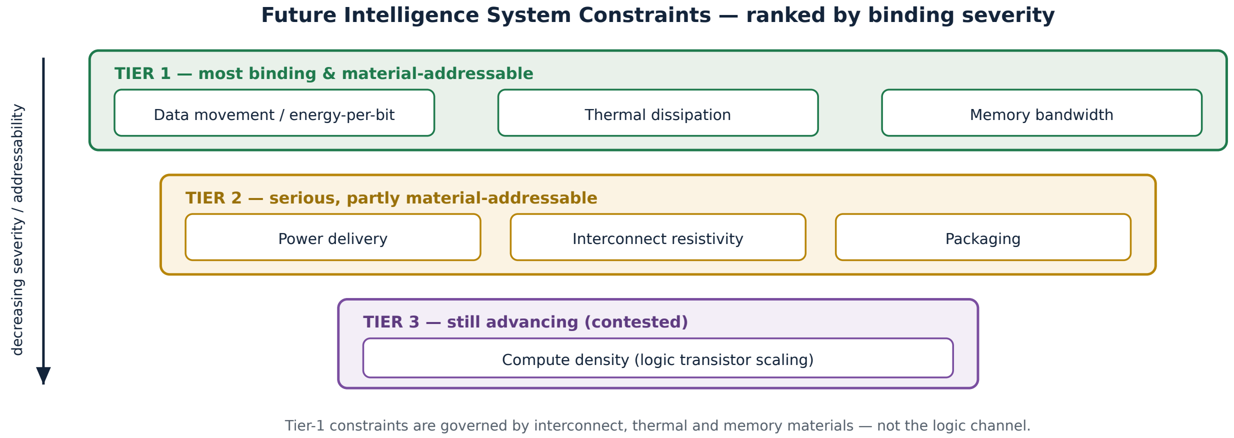


Figure 2. Constraints ranked by how soon they bind, how fundamental they are, and how material-addressable they are.

Table 2.1 — Constraints decomposed into physical / material / manufacturing roots.

Constraint	Quantified limitation	Physical root	Material root	Manufacturing root
Data movement	Off-chip word transport $\sim 10^2$–$10^3\times$ the arithmetic-op energy [11]	Charging interconnect capacitance over distance	Interconnect ρ, ε; link material	BEOL; optical co-integration

Constraint	Quantified limitation	Physical root	Material root	Manufacturing root
Thermal dissipation	Hot-spot flux hundreds of W/cm^2 [13]	Conduction physics; junction–ambient resistance	Spreader κ; boundary resistance	TIM/bonding; near-junction integration
Memory bandwidth	HBM3E ~1.2 TB/s/stack; arithmetic outgrew DRAM bandwidth [12]	Pin/beachfront density; signalling energy	DRAM cell scaling; I/O density	Stacking/TSV yield
Power delivery	kW-class packages strain on-die PDN	Resistive loss; current density	Conductor ρ; wide-bandgap devices	Backside-PDN / TSV
Interconnect	Cu ρ rises below mean-free-path widths [15]	Surface/grain-boundary scattering	Cu→Ru/Co/semi-metal; optical	Barrier/liner scaling
Compute density	3 nm-class; gate-length scaling slowed	Atomic dimensions; tunnelling; thermionic floor	Mobility/leakage in thin bodies; contacts	Lithographic stochastics

> **EBC** The most severe and most material-addressable constraints are governed by interconnect, thermal and memory materials, not the logic channel — the empirical basis for the Data-Movement Principle (§12).

3 First-principles physics framework

Each binding constraint is reduced to the physical relation that generates it, annotated with the material parameter it contains — making the downstream requirements derivations rather than assertions.

Table 3.1 — Governing relations and their material parameters.

Phenomenon	Relation	Material parameter(s)
Switching energy	$E_{sw} = \tfrac{1}{2}CV^2$; $P_{dyn} = \alpha CV^2 f$	C (permittivity, geometry); V
Thermodynamic floor	$E_{min} = k_B T \ln 2 \approx 2.85\times10^{-21}$ J (300 K)	— (bound; CMOS is 10^4–10^5× above [23])
Subthreshold floor	$S \geq (k_B T/q)\ln 10 \approx 60$ mV/dec	transport mechanism (beat via tunnelling/NC)
Carrier transport	$\mu = q\tau/m^*$; $\sigma = nq\mu$; $v_d = \mu E$	effective mass m^*; scattering time τ
Interconnect loss	$E_{bit} \propto L$; ρ_{eff} rises as $w\to\lambda$ (size effect [30])	ρ, ε, λ (mean free path)
Thermal bottleneck	$\Delta T = P\cdot R_{th}$; $P_{max} = \Delta T\cdot\kappa A/t$	thermal conductivity κ; boundary resistance R_K
Thermionic emission	$J = A\,T^2 \exp(-\Phi/k_B T)$	work function Φ; Richardson const. A
Thermoelectric recovery	$\eta = (\Delta T/T_h)\cdot f(ZT)$	figure of merit $ZT = S^2\sigma T/\kappa$

> **EBC** Every Tier-1 constraint reduces to a relation containing a specific material parameter: movement→(ρ,ε,L); thermal→(κ,R_K); switching→(C,V floored by $k_B T/q$); transport→(m^*,τ). The material requirements are the catalogue of these parameters; the architecture is the assignment of materials that optimise them per function.

4 Material classification framework

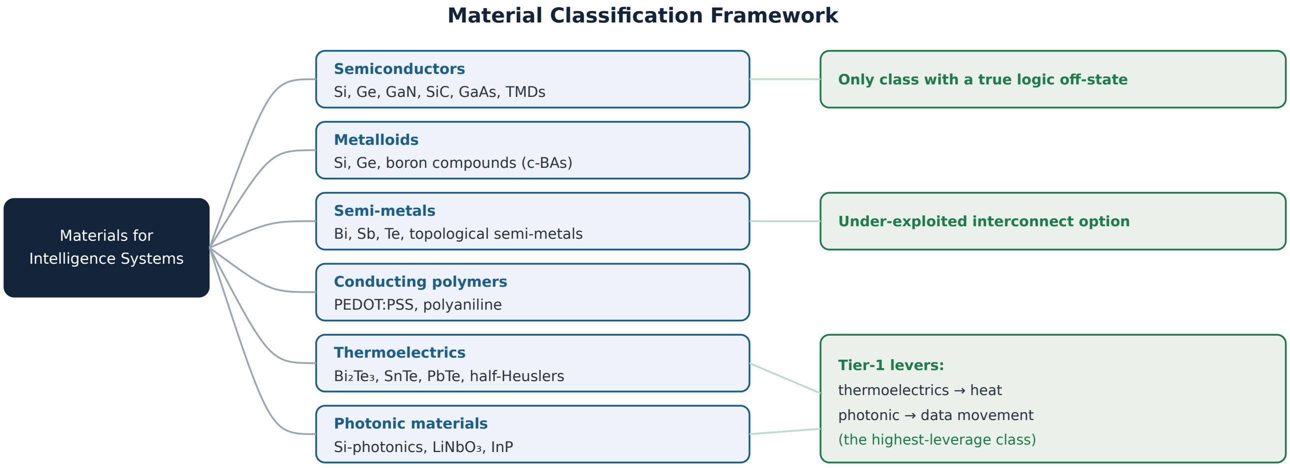
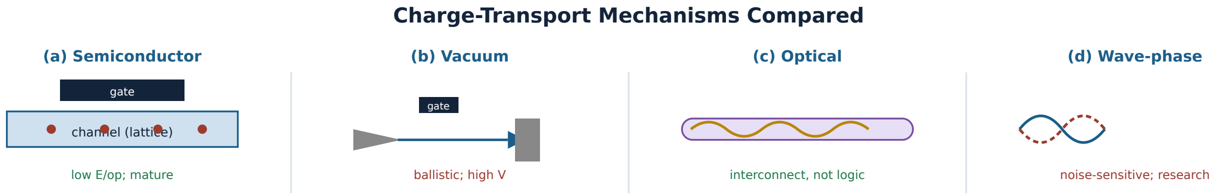


Figure 3. The six material classes, representative members, and the system role for which each is best suited.

Table 4.1 — Six material classes: transport, thermal, energy character, maturity, environment.

Class	Charge transport	Thermal	Energy character	Maturity	Environment
Semiconductors	Band transport across a bandgap	Si ~150, SiC ~400 W/m·K	Off-state → efficient logic	Si mature; 2D research	Si benign; Ga supply
Metalloids	Semiconducting band transport	Ge ~60; c-BAs ultrahigh	High hole μ (Ge); SiGe strain	Si/Ge mature	Ge scarce
Semi-metals	Band overlap; no gap	Often low lattice κ	No off-state; good interconnect	Research	Te scarce; As toxic
Conducting polymers	Hopping / polaron [16]	Low κ	Low μ; low-energy processing	Niche	Often low-tox
Thermoelectrics	Degenerate; $zT = S^2\sigma T/\kappa$	Deliberately low κ	ΔT↔voltage; near-RT zT~1	Mature (Peltier)	Te scarce; Pb toxic
Photonic	Photon generation/guiding	Avoids resistive heat	~length-independent E/bit	Si-photonics ready	In/Li scarcer

5 Charge-transport mechanisms



Only the semiconductor channel combines a true off-state with manufacturing maturity; optical wins transport; vacuum suits extreme environments.

Figure 4. Four charge-transport mechanisms. The semiconductor channel remains the logic substrate; optical is the transport lever.

Table 5.1 — Transport-mechanism comparison across the evaluation axes.

Mechanism	Energy	Thermal	Noise	Reliability	Manufacturability	System role
Semiconductor	Low/op (k_BT/q	Lattice self-heating;	Low	High	**Highest**	**Logic/memory backbone**

Mechanism	Energy	Thermal	Noise	Reliability	Manufacturability	System role
	floored)	managed				
Vacuum channel	No clear win (emission V)	No channel-lattice heating	Moderate	Emitter wear (open)	Demonstrators	Extreme-environment niche
Optical	Low E/bit over distance [29]	Avoids transport heating	Low	High	Si-photonics ready	**Interconnect**
Wave-phase	Potentially low [27,28]	Mechanism-dependent	**High sensitivity**	Reference/stability (open)	Research	Research

6 Candidate material assessment

Table 6.1 — Electrical & thermal performance (300 K, intrinsic/single-crystal unless noted).

Material	e^-/h^+ mobility ($cm^2/V\cdot s$)	Bandgap (eV) & logic suitability	κ (W/m·K)	Tier
Silicon	~1400 / ~450	1.12 indirect — good off-state [1,3]	~150	A
Germanium	~3900 / ~1900	0.66 — high leakage for logic	~60	A
GaN	2DEG ~1500–2000 / holes low	3.4 — poor complementary logic	~130–230	B
SiC (4H)	~900–1000 / ~100–120	3.26 — power, not logic	~370–490	A
Diamond	~2000–4500 / ~2000–3800	5.47 — n-doping very hard	~2000–2200	B
Graphene	~10^4–2×10^5 [8]	0 — no off-state (logic-disqualifying) [9]	~2000–5000 (susp.) [7]	B
$MoS_2/WS_2/WSe_2$	theory ~hundreds; meas. <~100	~1.6–2.0 — low-leakage scaling	~30–85	C
Carbon nanotube	~10^4–10^5	~0.5–1 — logic-capable [43]	~3000–3500 (tube)	C
c-BAs	~1400 ambipolar (reported) [6]	~1.5 (range reported)	~1300 (measured) [5]	C
Bi_2Te_3/SnTe	moderate	narrow — thermoelectric, not logic	~1.5–2 (low by design)	B
InP	~5400 / ~150	1.34 direct — photonic sources	~68	A

Sources [1,2,4–10,21]. Graphene/CNT/diamond κ are suspended/individual values; 2D mobility shows a large theory–measurement gap. Best-fit roles: Si logic/memory; GaN/SiC power; diamond/c-BAs thermal; 2D frontier logic; graphene/CNT interconnect/thermal; InP/Si-photonics interconnect.

7 Manufacturing readiness

Table 7.1 — Global-scale deployability.

Material / class	Wafer-scale	Ecosystem	Verdict
Silicon	300 mm	Native	READY
Silicon photonics	CMOS foundry	High	READY (interconnect)
GaN / SiC	150–200 mm	Separate flows	READY (power)
Thin-film $LiNbO_3$	Emerging foundry	Moderate	NEAR
2D TMDs / CIM NVM	Research / pilot	BEOL/3D in principle	NOT YET
Diamond	Small wafers	Low	NICHE
CNT / c-BAs / graphene logic	No wafers	Component/none	NOT YET

> **EBC** Only silicon, silicon photonics and wide-bandgap GaN/SiC are manufacturing-ready at global scale — decisive evidence against betting the logic tier on an unproven single material.

8 Sustainability & supply-chain resilience

Table 8.1 — Key-element abundance, concentration and flags.

Element / material	Abundance	Concentration / control	Flag
Silicon / Carbon	Very high	Diversified ore; concentrated fabs	Energy/water-intensive fabs
Gallium	Byproduct-limited	**Concentrated; export-controlled** [19]	Supply fragility
Germanium / Te / Se / In	Scarce	Byproduct / concentrated [19]	Scarcity-bound
Arsenic / Lead	Available	Diversified	**Toxicity**

> **EBC** Technical merit and supply security are anti-correlated across several high-performers (Ga, Ge, Te, Se, In). A sustainable architecture uses these elements *surgically* on an abundant silicon/carbon backbone.

9 Material-choice corroboration

The examples below illustrate disclosed industry directions and are included to compare the framework against publicly visible engineering trends rather than to predict specific products or timelines.

Table 9.1 — Disclosed material directions by functional layer (illustrative; verify against current vendor disclosures).

Layer	Disclosed material direction	Framework assignment	Agreement
Logic	Si nanosheet/GAA; 2D in research	Si/SiGe backbone; 2D at frontier	**Strong**
Interconnect	Co-packaged silicon photonics	Integrated photonics (highest-value)	**Strong**
Power	Backside power; wide-bandgap	GaN/SiC	**Strong**
Packaging	2.5D/3D, hybrid bonding, glass	Advanced packaging substrate	**Strong**
Memory	HBM; emerging NVM research	Si memory + compute-in-memory (§18)	Partial

10 Energy-per-useful-token framework

$$E_{tok} = E_{total} / N_{useful} = (1/N_{useful})\ \Sigma[\ E_{compute} + E_{memory} + E_{movement} + E_{leakage} + E_{conv\text{-}loss}\ ]$$

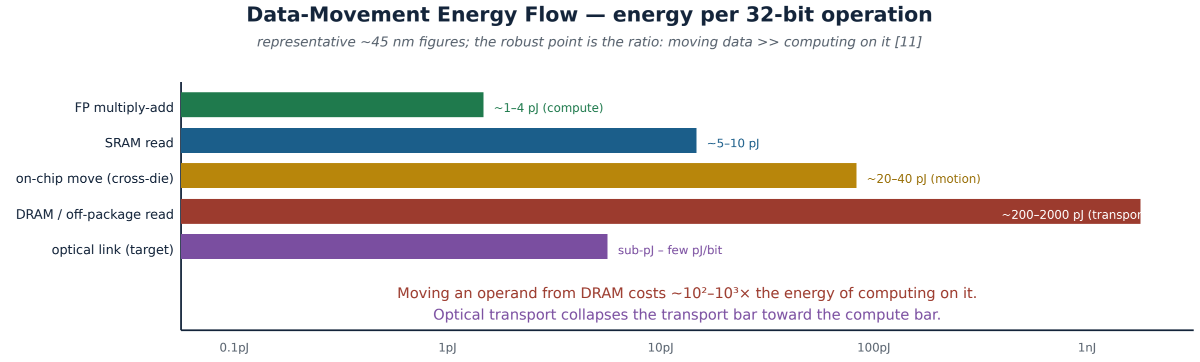


Figure 5. Energy per 32-bit operation across compute, on-chip motion, off-package transport, and optical links (representative orders of magnitude). The dominant term is data movement, and optical transport is the lever that shrinks it.

PHYSICS A dense transformer expends ~2N operations per token (N = parameters) [22]; autoregressive decode reads weights/KV-cache from memory every token, making generation memory-bandwidth-bound [12]. With $e_{\text{bit-moved}} \gg e_{\text{flop}}$ (Fig. 5), E_{tok} is dominated by the movement and memory terms.

> **EBC** Optimising logic speed barely moves E_{tok}; cutting data-movement and memory-access energy moves it substantially. This is the quantitative basis of the Data-Movement Principle (§12).

11 Toyota computational muda framework **NEW**

The Toyota Production System classifies seven wastes (*muda*) [31]. We map them to AI-infrastructure inefficiency, tag each by whether it is material-addressable, and — strengthened in this version — make the chain *muda* → E_{tok} *term* → *material lever* explicit for the three movement-related wastes.

Table 11.1 — Seven wastes → AI infrastructure → E_{tok} *term → material lever.*

Waste	AI equivalent	Material-addressable?	E_{tok} term	Material lever
Transport	Data movement (off-package)	**Yes — primary**	E_{movement}	Photonic interconnect; compute-in-memory; low-ρ conductors
Waiting	Memory stalls / compute idle	**Partly**	E_{memory} + idle static power	Memory bandwidth materials; HBM/3D; near-memory
Motion	On-chip / cache movement	**Partly**	E_{movement} (on-die)	On-die interconnect; 3D stacking shortens paths
Defects	Retraining / inference errors	Partly	raises effective E_{tok}	Device reliability; ECC; signal integrity
Inventory	Unused / over-allocated memory	Marginal	E_{memory} (indirect)	Denser/cheaper memory
Overproduction	Unnecessary token generation	**No — architectural**	N_{useful}	(software)
Overprocessing	Excess-precision computation	**No — architectural**	E_{compute}	(low-precision design)

> **PHYSICS** The three movement-related wastes map directly onto E_{tok}: *Transport* and *Motion* are the E_{movement} term ($\propto$ bits$\times e_{\text{bit-moved}}$, with $e_{\text{bit-moved}}$ set by ρ,ε,L — §3); *Waiting* contributes both an E_{memory} access term and a static-leakage term accrued while the pipeline stalls ($P_{\text{leak}} \times t_{\text{idle}}$). A material that lowers $e_{\text{bit-moved}}$ (photonics, semi-metal wires) or removes the move entirely (compute-in-memory) reduces all three simultaneously, whereas a faster logic transistor touches none of them.

> **Strengthened thesis.** The largest *material-addressable* muda is Transport, and it is reduced only by materials that move data with less energy or eliminate the move. The Muda lens, the E_{tok} decomposition and the §3 physics converge on the same conclusion — formalised next as the Data-Movement Principle.

12 The XSYDA Data-Movement Principle **NEW — central thesis**

> **XSYDA Data-Movement Principle.**
> *For future intelligence systems, reduction of data movement yields greater system-level energy savings than improvement of switching speed beyond the threshold at which logic-switching energy ceases to dominate energy per useful token.*

12.1 First-principles support

Write $E_{tok} = E_{compute} + E_{move}$ + (other terms), where $E_{compute}$ aggregates switching energy (§3.1, $\propto \frac{1}{2}CV^2$ per operation) and E_{move} aggregates data transport (§3.5, $\propto$ bits$\times e_{bit\text{-}moved}$). The marginal system return on each lever is:

$$\partial E_{tok}/\partial(\text{switching improvement}) \to 0 \text{ once } E_{compute} \ll E_{move}; \quad \partial E_{tok}/\partial(\text{movement reduction}) \text{ remains } O(E_{move})$$

PHYSICS The threshold is the crossover $E_{compute} \approx E_{move}$. Below it (i.e. when movement dominates), shrinking $E_{compute}$ further yields diminishing system-level returns because it is a small and shrinking share of the total, while shrinking E_{move} continues to pay roughly in proportion to its share.

> **EBC** The threshold has already been crossed. Per-operand DRAM transport costs ~10^2–10^3× the arithmetic energy (Fig. 5 [11]), and autoregressive inference is memory-bandwidth-bound [12]. Therefore $E_{compute} \ll E_{move}$ holds for the dominant AI workloads today, and the Principle applies to the 2035–2050 design horizon.

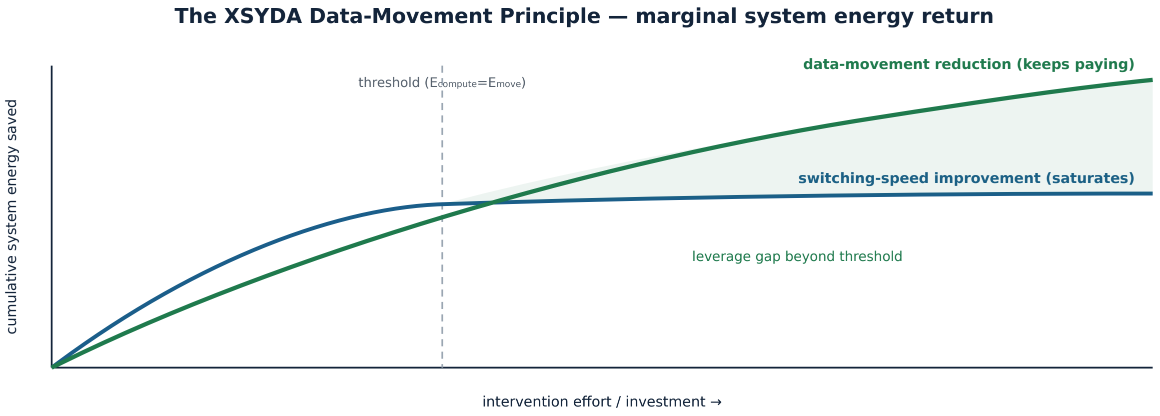
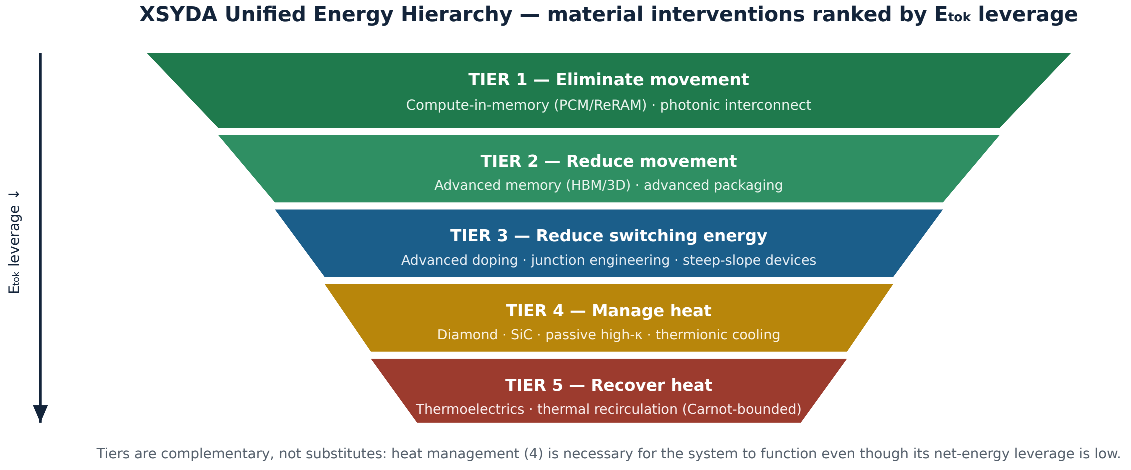


Figure 6. Beyond the crossover where switching energy ceases to dominate energy per token, returns on faster logic saturate while returns on reducing data movement continue — the quantitative statement of the Data-Movement Principle.

> **OPEN / boundary.** The Principle is a leverage statement, not a claim that switching energy is irrelevant: logic must still function, and a hypothetical compute-bound regime (or a future where movement is already near-minimal) would move the operating point back toward the threshold. It is falsified if a workload class keeps $E_{compute} \geq E_{move}$ at scale, or if data movement is driven so low that the compute term re-dominates.

13 The XSYDA Unified Energy Hierarchy **NEW**

If the Data-Movement Principle holds, material interventions can be ranked by their leverage on E_{tok}. We propose a five-tier hierarchy and test each tier against the physics of §3.



XSYDA Unified Energy Hierarchy — material interventions ranked by E_{tok} leverage

Tiers are complementary, not substitutes: heat management (4) is necessary for the system to function even though its net-energy leverage is low.

Figure 7. The XSYDA Unified Energy Hierarchy. Interventions that eliminate or reduce data movement sit at the top; heat recovery, bounded by Carnot, sits at the bottom. Width encodes E_{tok} leverage.

Table 13.1 — Each tier tested against physics and evidence.

Tier	Mechanism & materials	Physical basis	Supported?
1 Eliminate movement	Compute-in-memory; photonic interconnect	Removes/length-decouples E_{move} — the dominant term	**Yes (highest)**
2 Reduce movement	HBM/3D stacking; packaging shortens paths	Lowers bits×distance in E_{move}	**Yes**
3 Reduce switching energy	Doping/junction; steep-slope (sub-60 mV/dec)	Lowers $\frac{1}{2}CV^2$ — but a minority term (§12)	**Yes, lower leverage**
4 Manage heat	Diamond/SiC high-κ; thermionic cooling	$P_{max} \propto \kappa$; enables density but no net-energy cut	**Necessary, not energy-saving**
5 Recover heat	Thermoelectrics; recirculation	$\eta=(\Delta T/T_h)f(ZT)$; small ΔT → ~1–3%	**Marginal**

> **XSYDA contribution.** The hierarchy is supported: tier order follows directly from the E_{tok} decomposition and the governing relations of §3. We elevate it as an original organising framework, with the explicit caveat that the tiers are *complementary* — a system needs heat management (Tier 4) regardless of its low net-energy leverage — so the hierarchy ranks *where new material investment buys the most energy*, not which function can be omitted.

14 Advanced doping & junction engineering **expanded**

Junction engineering operates at Tier 3 of the hierarchy. We expand it here because it is the principal lever for the switching-energy term and the enabler for two-dimensional logic, even though it is secondary to data-movement reduction.

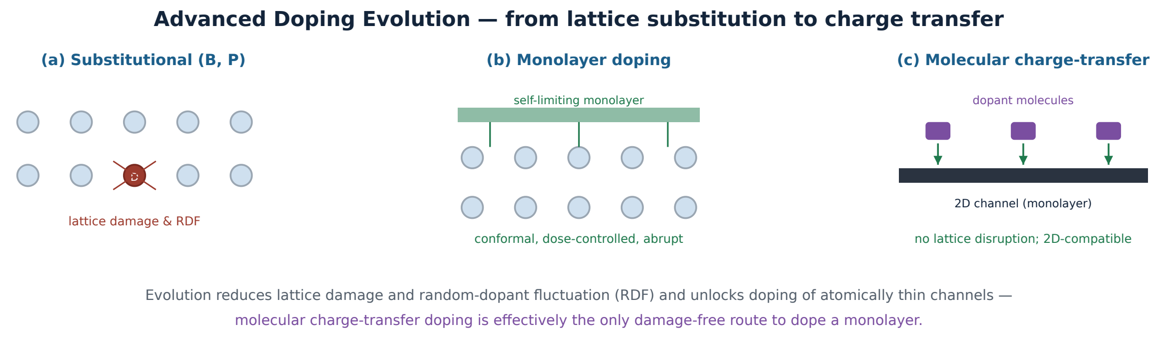


Figure 8. Advanced doping evolution. Substitutional dopants disturb the lattice and cause threshold variability [32]; monolayer and molecular charge-transfer doping introduce carriers without lattice damage and enable two-dimensional channels.

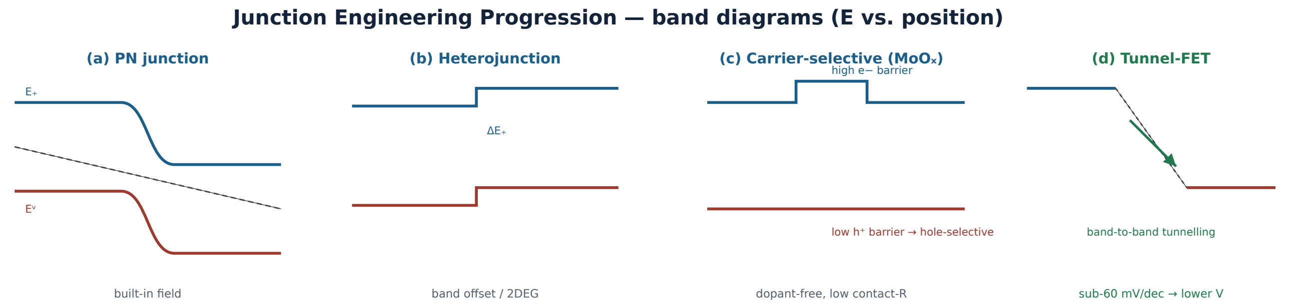


Figure 9. Junction engineering progression as band diagrams: the built-in field of a PN junction; the band offset of a heterojunction; the asymmetric barriers of a carrier-selective contact; and band-to-band tunnelling in a TFET, which beats the 60 mV/decade thermionic limit and lowers the supply voltage.

Table 14.1 — Ranked assessment of doping/junction approaches (5 = best; Supply 5 = most resilient).

Approach	Elec. efficiency	Leakage ↓	Thermal ↓	Manuf.	Raw-mat.	Environ.	Scalability	/35	Rank
Molecular / surface charge-transfer [38]	4	4	3	3	5	4	5	28	1 ★
TiO_x/MoO_x carrier-selective contacts	5	5	3	3	4	4	4	28	1 ★
Monolayer doping (MLD) [33]	4	4	3	3	4	4	4	26	3
Arsenic / Antimony (alt. n)	4	4	2	5	3	2	4	24	4
Steep-slope (TFET/NCFET) [35,36]	5	5	4	2	4	4	2	26	3=
Boron / Phosphorus (conventional)	3	3	2	5	5	4	3	25	5
Gallium / Indium (alt. p)	3	3	2	4	2	3	3	20	7

§14 verdict. The ranked leaders are **molecular/charge-transfer doping** and **carrier-selective oxide (TiO_x/MoO_x) contacts** — both reduce leakage and contact resistance without lattice damage, and the former is the key enabler for 2D channels. Junctionless architectures [34] and steep-slope devices offer the largest switching-energy reduction (sub-60 mV/dec) but are gated by drive-current/reliability and manufacturability. Conventional B/P remain the manufacturable baseline. **All operate at Tier 3** of the energy hierarchy: real, but secondary to data-movement reduction, and primarily valuable as enablers (2D logic) and contact/leakage improvements; modulation (remote) doping [37] additionally removes impurity scattering to raise mobility.

15 Thermally assisted switching & energy recirculation research hypothesis

Framing (established physics). “Self-sustaining” switching from recaptured heat cannot mean net-zero or perpetual operation: the second law forbids fully converting thermalised energy to work without a colder reservoir. We separate *thermal recovery* (heat after dissipation) from *charge recovery* (electrical energy before it becomes heat).

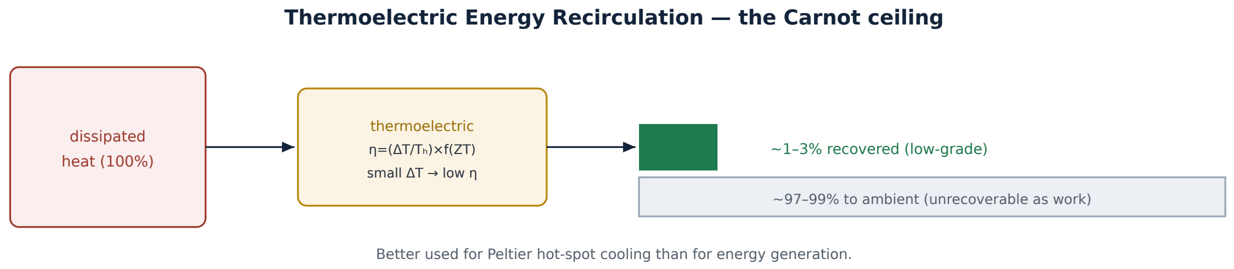


Figure 10. Thermoelectric energy recirculation. The small on-die temperature difference caps Carnot efficiency, so only a few percent of dissipated heat is recoverable as low-grade electricity.

§15 verdict (Tier 5). Thermoelectric thermal recovery is real but marginal at chip scale (~1–3%, low-grade); thin-film modules raise near-room-temperature ZT [44]. Thermal switch-assist is not meaningful: Seebeck voltage $S\cdot\Delta T \approx$ 2–6 mV versus a 300–700 mV switching threshold. **Charge-recovery (adiabatic) logic** — $E_{adiabatic} \approx (RC/T)\cdot CV^2 \to 0$ as ramp time $T\to\infty$ [24,25] — is the credible recirculation mechanism, recycling electrical charge (not heat), at a speed-penalty cost.

16 Thermionic cooling & electron-sweat concepts expanded

Thermionic cooling (Tier 4) preferentially removes the highest-energy carriers, carrying away more than the average thermal energy per electron — an electronic analogue of evaporative ("sweat") cooling. The governing relation is the Richardson–Dushman emission law, with the work function Φ as the decisive material parameter.

$$J = A\,T^2 \exp(-\Phi/k_BT); \quad \text{heat carried per emitted electron} \approx \Phi + 2k_BT$$

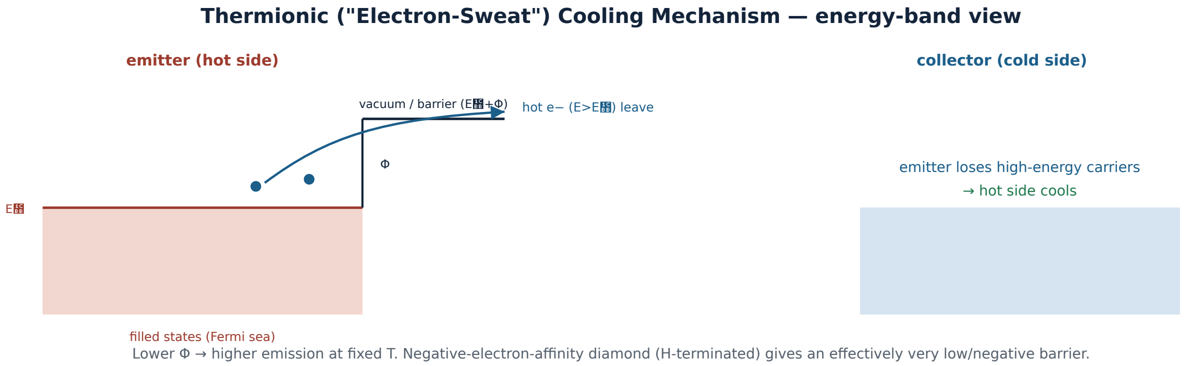


Figure 11. Thermionic cooling. Carriers in the high-energy tail of the Fermi distribution are emitted over the work-function barrier Φ, removing more than the mean thermal energy per electron and cooling the emitter — an electronic evaporative ("electron-sweat") effect governed by the Richardson–Dushman law.

Table 16.1 — Candidate materials for thermionic / electron-emission cooling.

Material	Work-function relevance	Cooling relevance	Status
Diamond (H-terminated)	**Negative electron affinity** — very low effective barrier; n-type diamond pn junctions demonstrated [39]	Strongest emitter candidate; also highest κ (isotopically enhanced [40])	EXPERIMENT
SiC / GaN	Wide bandgap; engineerable surfaces	Robust high-temperature emitters	EXPERIMENT / OPEN
Silicon	Moderate Φ; CMOS-compatible	Integrable but modest emission	ENGINEERING
Heterostructures	Engineered barrier height (Shakouri–Bowers [45])	Integrated thermionic/thermotunnelling coolers	EXPERIMENT / OPEN

§16 verdict. Thermionic cooling is physically real and material-selective — negative-electron-affinity diamond is the standout candidate (and doubles as the highest-κ spreader). As a **hotspot-management mechanism** it is promising; as a **general thermal-management layer** it is a research hypothesis (efficiency, barrier control, integration); and as a **practical AI-accelerator cooling technology** it is **not yet established**. Like all Tier-4 active cooling, it manages heat at an energy cost and does not reduce net energy — its value is enabling density and suppressing throttling, where passive high-κ conduction remains the more energy-efficient default.

17 Nanoscale vacuum channel transistors (NVCT)

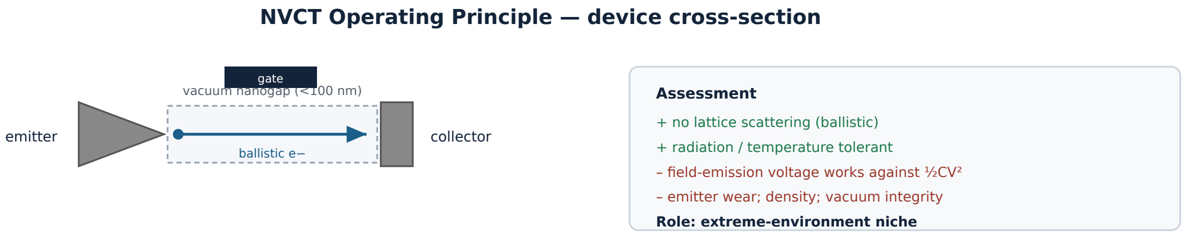
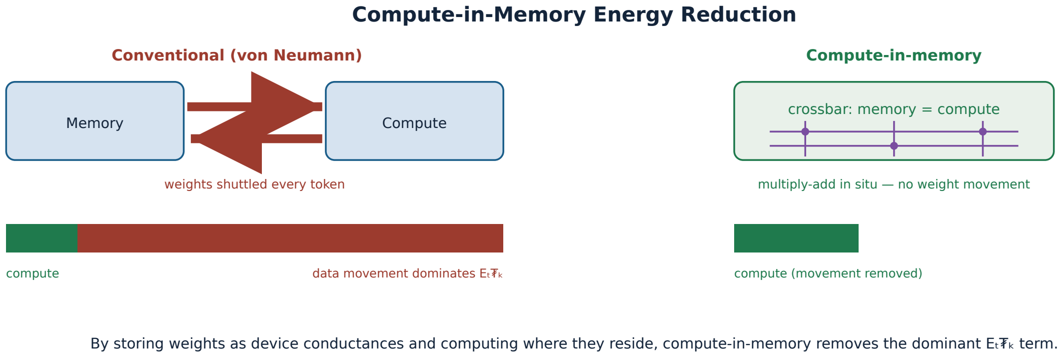


Figure 12. NVCT cross-section: a gate modulates electrons field-emitted ballistically across a sub-100 nm vacuum gap. The absence of a lattice removes scattering and confers radiation tolerance; emission voltage and reliability confine NVCTs to a niche.

> **§17 verdict.** Vacuum channels remove lattice scattering and tolerate extreme environments but have not shown energy-efficiency gains for mainstream logic (emission voltage cuts against ½CV^2) [26]. Defensible role: a radiation-hard / extreme-environment niche — alongside superconducting logic families [42] for their own niches — a complement, not a competitor, to the architecture.

18 Memory materials & compute-in-memory **expanded**


Compute-in-Memory Energy Reduction

By storing weights as device conductances and computing where they reside, compute-in-memory removes the dominant $E_{t₮k}$ term.

Figure 13. Compute-in-memory energy reduction. Conventional architectures shuttle weights between memory and compute, where data movement dominates energy per token; analog compute-in-memory performs multiply-accumulate in the memory array itself, removing that movement.

Table 18.1 — Material-enabled memory levers.

Lever	Material basis	Effect	Limit	Status
Multi-level cells	Distinguishable charge/resistance states	2–4× bits/cell, no area growth	Noise margin, endurance	**EXPERIMENT**
3D stacking	3D NAND; 3D-DRAM research	Capacity per footprint	Stack height, thermal	**EXPERIMENT**
PCM / ReRAM / MRAM / FeFET	Phase / resistive / magnetic / ferroelectric state	Dense non-volatile, multi-level	Variability, endurance, drift	**EXPERIMENT** [46]
Compute-in-memory	ReRAM/PCM conductance = weight; in-situ MAC	**Removes weight movement → attacks dominant E_{tok} term**	Precision, ADC/DAC overhead, variability	**OPEN** [47]

> **Overlooked opportunity (Tier 1).** The highest-value memory-material opportunity is analog **compute-in-memory**: a memory material that performs computation where the data resides removes the weight movement dominating E_{tok} — a memory×transport×energy interaction sitting at the top of the energy hierarchy. **HYP** Precision, conversion overhead and endurance make it a research hypothesis for large models; it warrants priority in the architecture's memory tier.

18.1 Why compute-in-memory sits at Tier 1 — advantages, challenges, readiness

Compute-in-memory (CIM) is placed at the apex of the energy hierarchy because it is the only memory-material lever that attacks $E_{movement}$ directly rather than the minority $E_{compute}$ term. The multiply-accumulate that dominates neural inference is performed in situ: input voltages drive a crossbar of programmable conductances (ReRAM/PCM cells encoding the weights), and Ohm's and Kirchhoff's laws sum the currents into the result — the weights never leave the array. Because the dominant energy term in §10 is the movement of those weights, eliminating their transport is a first-order, not incremental, reduction in energy per token.

Table 18.2 — Compute-in-memory scorecard: advantages, challenges and technology readiness.

Dimension	Assessment
Energy advantage	Removes weight movement — attacks the dominant E_{tok} term; in-array MAC avoids fetch energy entirely (Tier-1 leverage)
Throughput advantage	Massively parallel analog MAC per array; relieves the memory-bandwidth wall (§2) for weight-stationary workloads
Material basis	ReRAM (filamentary/oxide), PCM (chalcogenide phase), and emerging FeFET/MRAM conductance states — CMOS-BEOL-compatible in principle
Challenge: precision	Analog noise, conductance quantisation and device-to-device variability cap effective bit precision; ADC/DAC conversion adds overhead that can erode the movement saving if mismanaged
Challenge: endurance/drift	Write endurance and conductance drift (notably PCM) constrain training and long-lived weight programming
Technology readiness	**OPEN** — inference-class demonstrators exist; production-grade precision/endurance for large-model training is unproven; the principal research gate of the recommended architecture
Why Tier 1	Eliminates rather than reduces movement; complementary to photonic interconnect (which reduces inter-chip movement) — together they form the two data-movement levers of §19

CIM and photonics are complementary, not competing: photonics lowers the energy of data that must still move between chips and memory stacks, while CIM removes the movement of the weights altogether. A 2035–2050 architecture credibly uses both — photonics at the package and rack interconnect, CIM within the weight-stationary memory tier.

19 Material architecture recommendation

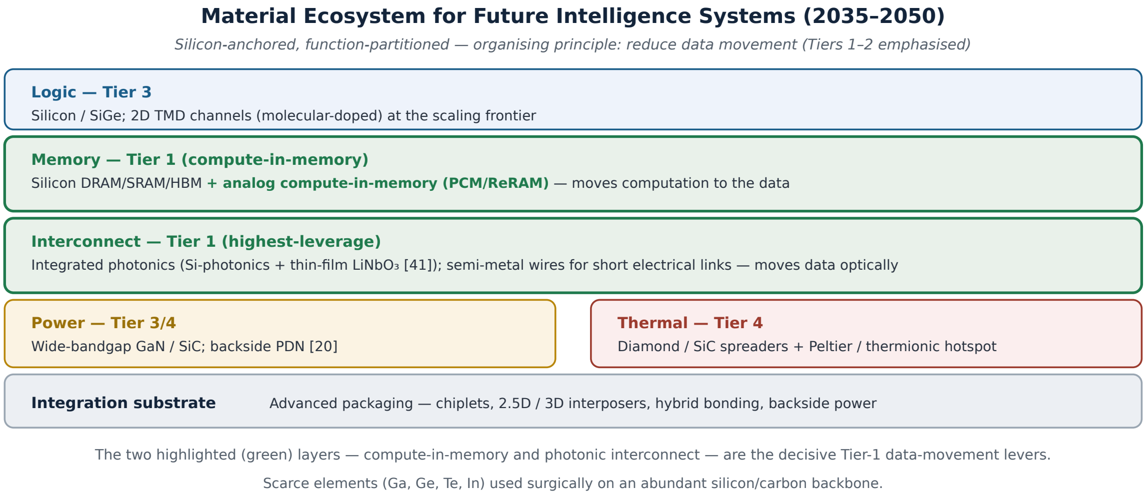


Figure 14. The recommended material ecosystem, with the two Tier-1 data-movement levers (compute-in-memory and photonic interconnect) emphasised, wide-bandgap power, diamond-class thermal materials, and advanced packaging as the integrating substrate.

Table 19.1 — Material-architecture ranking (5 = best; Supply 5 = most resilient).

Option	Perf	Manuf	Sustain	Cost	Supply	/25	Rank
Silicon + Photonics + Compute-in-memory	5	4	4	4	4	21	1 ★
Silicon + Photonics	5	5	4	4	4	22	1 ★
Silicon + SiC	4	4	4	3	4	19	3
Silicon + GaN	4	4	3	3	2	16	4=
Silicon + Diamond	4	2	4	2	4	16	4=
Pure Graphene Logic	1	1	4	3	4	13	7
c-BAs (standalone)	4	1	3	1	2	11	8

Silicon + Photonics ranks first on present manufacturability; Silicon + Photonics + Compute-in-memory ties on leverage but carries a manufacturability discount (compute-in-memory not yet production-grade) — it is the highest-upside, research-gated option. Both express the same recommendation: attack data movement.

Recommendation. Reject the single-successor hypothesis. The most credible architecture is **silicon-anchored and function-partitioned**, with its two decisive investments at Tier 1: **integrated photonics** (production-ready) and **analog compute-in-memory** (highest upside, research-gated), supported by wide-bandgap power, diamond-class thermal materials with selective thermionic/Peltier hotspot cooling, and molecular-doping-enabled 2D channels at the frontier, on an advanced-packaging substrate.

Failure modes (falsifiers). A manufacturable steep-slope/beyond-CMOS logic device at wafer scale; compute-in-memory failing to reach usable precision/endurance; integration cost outrunning specialisation benefit; severe Ga/Ge/Te/Se/In supply shocks; or a workload regime that keeps $E_{\text{compute}} \geq E_{\text{move}}$. Absent these, the architecture stands.

20 XSYDA economic implication NEW

The energy-per-token findings carry a direct economic corollary at datacentre scale. We state it as the economic form of the Data-Movement Principle, keeping the analysis anchored to material-driven energy — criteria 12 (cost) and 15 (energy-per-computation) of the evaluation framework — rather than to business strategy.

Economic corollary of the Data-Movement Principle.
Because data movement dominates energy per useful token, a 10× reduction in data-movement energy produces a larger system-level economic benefit than a 10× reduction in switching energy. Material investments that cut movement (integrated photonics, compute-in-memory) therefore carry disproportionate economic leverage relative to materials that raise switching speed.

Three consequences follow, each traceable to a material lever. **Datacentre power.** At scale the binding resource is increasingly the available electrical power and the ability to remove heat, not transistor count; energy is the dominant recurring operating cost of AI infrastructure [18]. A material that lowers E_{tok} lowers the energy bill and, because less energy is dissipated, the cooling burden simultaneously — whereas a faster logic transistor that does not lower E_{tok} reduces neither in proportion. **Total cost of ownership.** TCO for an AI facility is dominated by energy, cooling, and the capital of the power- and thermal-delivery plant; a reduction in data movement compounds across all three, since it cuts the energy drawn, the heat to be removed, and the size of the power/thermal infrastructure required. **Scaling economics.** The runway for ever-larger systems is bounded by power and thermal envelopes rather than by the unit cost of logic; the Tier-1 material levers (§13) extend that runway, which is why they are the materials with the greatest economic value at the frontier.

Table 20.1 — Material lever → energy effect → economic effect (qualitative; verify magnitudes against current sources).

Material lever (tier)	Energy effect	Dominant economic effect
Compute-in-memory (T1)	Removes weight-movement energy	Lowers E_{tok}, energy bill and cooling load together — highest TCO leverage
Integrated photonics (T1–2)	Length-decouples inter-chip transport energy	Reduces energy and extends rack/scale-out economics
Advanced packaging (T2)	Shortens data paths	Improves performance-per-watt and density per facility
Wide-bandgap power (T3–4)	Lowers conversion/delivery loss	Direct reduction in delivered-power overhead
Diamond/SiC thermal (T4)	Raises allowable power density	Defers cooling-capex ceiling; enables denser deployment
Faster switching only (T3)	Lowers a minority E_{tok} term	Limited system-level economic return past the threshold (§12)

Caveat. Specific datacentre power, TCO and infrastructure-cost figures depend on workload mix, utilisation, geography, electricity price and rapidly changing market data; the relations above are directional corollaries of the energy analysis and should be sourced from current primary references before external or financial use.

21 XSYDA contribution & conclusion

XSYDA contribution. Beyond a literature synthesis, this work contributes: (i) a first-principles reduction of each system constraint to its governing relation and material parameter (§3); (ii) energy per useful token as a unifying material figure of merit (§10); (iii) the strengthened **Toyota computational muda** mapping of movement wastes onto E_{tok} terms (§11); (iv) the **XSYDA Data-Movement Principle**, the paper's central, falsifiable thesis (§12); and (v) the **XSYDA Unified Energy Hierarchy**, an evidence-tested ranking of material interventions (§13). Together they yield one memorable claim:

Originality. Existing literature and industry discussion describe memory bandwidth, interconnect energy and the "memory wall" as *separate* problems [11,12]. The XSYDA Data-Movement Principle unifies these observations into a single leverage framework: once data movement dominates energy per useful token, reducing movement yields greater system-level benefit than further improvement in switching speed. The Unified Energy Hierarchy then converts that principle into a ranked material-investment ordering. Neither is a re-statement of the memory wall: the memory wall describes a *bandwidth-vs-compute growth gap*, whereas the Principle is a *marginal-energy-return* statement with an explicit crossover threshold (§12) and the Hierarchy is a decision rule for allocating material R&D. That is the contribution this work claims.

The most important materials breakthrough for artificial intelligence will be measured in picojoules per bit moved, not in transistor switching speed.

21.1 The five questions, answered assertively

1. **The most important materials breakthrough opportunity for AI** is analog *compute-in-memory* on resistive/phase-change materials — the one material lever that eliminates the data movement dominating energy per token — closely followed by integrated photonic interconnect.
2. **What material investments matter most:** Tier-1 levers — photonic-integration materials (silicon photonics, thin-film $LiNbO_3$) and compute-in-memory materials (PCM/ReRAM) — on the silicon backbone and advanced packaging that make them deployable.
3. **What industry should prioritise:** co-packaged optics and compute-in-memory first, then advanced packaging, wide-bandgap power and diamond-class thermal materials — by leverage, not headline metrics.

4. **What researchers should prioritise:** compute-in-memory precision/endurance materials; high-density molecular doping of 2D channels; manufacturable steep-slope devices; negative-electron-affinity thermionic emitters; n-type diamond; and self-regulating thermal materials.
5. **What should not be prioritised:** chasing higher logic mobility or switching speed as the primary lever; on-die thermal energy harvesting; thermally assisted switching; pure graphene logic; and the search for a single-material silicon successor — all are, on the evidence, low-leverage or thermodynamically bounded.

The conclusion is therefore assertive and specific: future intelligence systems will not be unlocked by a faster transistor or a single miracle material, but by a silicon-anchored ecosystem that moves data with light and computes where the data already resides. The materials that reduce data movement are worth more than the materials that switch fastest.

The decisive synthesis.

What industry believes: that one material will replace silicon. **What XSYDA found:** no — no single material satisfies all functional requirements, and several requirement pairs are physically antagonistic. **What XSYDA proposes:** a silicon-anchored, function-partitioned heterogeneous architecture. **What matters most:** reducing data movement. **Where investment should go, in order:** compute-in-memory and silicon photonics first, then advanced packaging, then wide-bandgap power and diamond-class thermal materials, with heat recovery last.

The future is not beyond silicon. The future is beyond unnecessary data movement.

21.2 Open research questions

1. Quantitative E_{tok} property thresholds (energy/bit, κ, μ) for a 2035–2050 accelerator.
2. For each Tier-1 constraint, whether the binding limit is material or architectural.
3. Whether analog compute-in-memory reaches the precision/endurance needed for large-model inference and training.
4. A manufacturable steep-slope device resolving the drive-current/reliability trade-off at wafer scale.
5. Reproducible, air-stable, high-density molecular doping of 2D channels.
6. Chip-scale thermionic ("electron-sweat") cooling efficiency; negative-electron-affinity emitter integration.
7. Substitutability of supply-fragile high-performers (Ga, Ge, Te, Se, In) without hard performance ceilings.

Reproducibility & data availability

Property values are reference figures drawn from the cited literature and standard materials databases, with measurement conditions (temperature, sample form, dimensionality) stated where material. No data were simulated or synthetically generated; every value is sourced from the published record. Rapidly evolving market, energy and supply statistics are cited to their period and should be read in the context of current data.

References

1. S. M. Sze, K. K. Ng, *Physics of Semiconductor Devices*, 3rd ed., Wiley, 2007.
2. Ioffe Institute, *New Semiconductor Materials — Characteristics and Properties (NSM Archive)*.
3. C. Jacoboni et al., *Solid-State Electronics* 20, 77 (1977).
4. L. Lindsay, D. A. Broido, T. L. Reinecke, *Phys. Rev. Lett.* 111, 025901 (2013).
5. J. S. Kang et al., *Science* 361, 575 (2018); S. Li et al., *Science* 361, 579 (2018); F. Tian et al., *Science* 361, 582 (2018).
6. J. Shin et al., "High ambipolar mobility in cubic boron arsenide," *Science* 377, 437 (2022).
7. A. A. Balandin et al., "Superior Thermal Conductivity of Single-Layer Graphene," *Nano Lett.* 8, 902 (2008).
8. K. S. Novoselov et al., *Science* 306, 666 (2004); K. I. Bolotin et al., *Solid State Commun.* 146, 351 (2008).
9. F. Schwierz, "Graphene transistors," *Nature Nanotechnology* 5, 487 (2010); Y.-M. Lin et al., *Science* 327, 662 (2010).

10. S. B. Desai et al., *Science* 354, 99 (2016); Y. Liu et al., *Nature* 591, 43 (2021); B. Radisavljevic et al., *Nat. Nanotechnol.* 6, 147 (2011).
11. M. Horowitz, "Computing's Energy Problem (and what we can do about it)," *IEEE ISSCC*, 2014.
12. A. Gholami et al., "AI and Memory Wall," *IEEE Micro*, 2024.
13. H. Esmaeilzadeh et al., "Dark Silicon and the End of Multicore Scaling," *ISCA*, 2011.
14. T. N. Theis, H.-S. P. Wong, "The End of Moore's Law," *Comput. Sci. Eng.* 19, 41 (2017).
15. D. Gall, "The search for the most conductive metal for narrow interconnect lines," *J. Appl. Phys.* 127, 050901 (2020).
16. H. Sirringhaus, *Adv. Mater.* 26, 1319 (2014).
17. IEEE, *International Roadmap for Devices and Systems (IRDS)*, 2023 Edition.
18. International Energy Agency, *Electricity 2024* (2024); D. Patterson et al., *Computer* 55(7), 18 (2022).
19. U.S. Geological Survey, *Mineral Commodity Summaries 2024*.
20. B. J. Baliga, *Fundamentals of Power Semiconductor Devices*, Springer, 2008.
21. G. J. Snyder, E. S. Toberer, "Complex thermoelectric materials," *Nature Materials* 7, 105 (2008).
22. J. Kaplan et al., "Scaling Laws for Neural Language Models," 2020.
23. R. Landauer, "Irreversibility and Heat Generation in the Computing Process," *IBM J. Res. Dev.* 5, 183 (1961).
24. C. H. Bennett, "Logical Reversibility of Computation," *IBM J. Res. Dev.* 17, 525 (1973).
25. W. C. Athas et al., "Low-power digital systems based on adiabatic-switching principles," *IEEE Trans. VLSI Syst.* 2, 398 (1994).
26. D.-I. Moon, J.-W. Han, M. Meyyappan, "Nanoscale vacuum channel transistor," *Nano Letters* 17, 2146 (2017); J.-W. Han et al., *Appl. Phys. Lett.* 100, 213505 (2012).
27. E. Goto, "The Parametron," *Proc. IRE* 47, 1304 (1959).
28. A. V. Chumak et al., "Magnon spintronics," *Nature Physics* 11, 453 (2015).
29. D. A. B. Miller, "Are optical transistors the logical next step?," *Nature Photonics* 4, 3 (2010); "Device Requirements for Optical Interconnects," *Proc. IEEE* 97, 1166 (2009).
30. A. F. Mayadas, M. Shatzkes, *Phys. Rev. B* 1, 1382 (1970); K. Fuchs (1938); E. H. Sondheimer (1952).
31. T. Ohno, *Toyota Production System: Beyond Large-Scale Production*, Productivity Press, 1988.
32. M. J. M. Pelgrom et al., "Matching properties of MOS transistors," *IEEE J. Solid-State Circuits* 24, 1433 (1989).
33. J. C. Ho et al., "Controlled nanoscale doping of semiconductors via molecular monolayers," *Nature Materials* 7, 62 (2008).
34. J.-P. Colinge et al., "Nanowire transistors without junctions," *Nature Nanotechnology* 5, 225 (2010).
35. A. M. Ionescu, H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature* 479, 329 (2011).
36. S. Salahuddin, S. Datta, "Use of Negative Capacitance...," *Nano Letters* 8, 405 (2008).
37. R. Dingle et al., *Appl. Phys. Lett.* 33, 665 (1978).
38. H. Fang et al., "Degenerate n-doping of few-layer TMDs by potassium," *Nano Letters* 13, 1991 (2013).
39. S. Koizumi et al., "Ultraviolet Emission from a Diamond pn Junction," *Science* 292, 1899 (2001).
40. L. Wei et al., "Thermal conductivity of isotopically modified single crystal diamond," *Phys. Rev. Lett.* 70, 3764 (1993).
41. C. Wang et al., *Nature* 562, 101 (2018); Y. Shen et al., *Nature Photonics* 11, 441 (2017); B. J. Shastri et al., *Nature Photonics* 15, 102 (2021).
42. D. S. Holmes et al., *IEEE Trans. Appl. Supercond.* 23, 1701610 (2013); K. K. Likharev, V. K. Semenov, *IEEE Trans. Appl. Supercond.* 1, 3 (1991).
43. G. Hills et al., "Modern microprocessor built from complementary CNT transistors," *Nature* 572, 595 (2019).
44. R. Venkatasubramanian et al., "Thin-film thermoelectric devices with high room-temperature figures of merit," *Nature* 413, 597 (2001).
45. G. D. Mahan, "Thermionic refrigeration," *J. Appl. Phys.* 76, 4362 (1994); A. Shakouri, J. E. Bowers, *Appl. Phys. Lett.* 71, 1234 (1997).
46. H.-S. P. Wong et al., "Metal-Oxide RRAM," *Proc. IEEE* 100, 1951 (2012); "Phase Change Memory," *Proc. IEEE* 98, 2201 (2010).
47. A. Sebastian et al., "Memory devices and applications for in-memory computing," *Nature Nanotechnology* 15, 529 (2020); H.-S. P. Wong, S. Salahuddin, *Nature Nanotechnology* 10, 191 (2015).

About XSYDA Research. XSYDA Research is the independent research initiative of XSYDA Technologies. It publishes evidence-based, first-principles research as preprints — regardless of whether the findings confirm or challenge prevailing expectations — across Enterprise Intelligence; Computing Intelligence; Semiconductor & Future Computing Systems; AI Infrastructure & Systems Architecture; Intelligent Automation & Decision Systems; and Supply Chain & Operational Intelligence.

Author. Srinivasan Subramanian, Founder, XSYDA Technologies. **Published by.** XSYDA Research, XSYDA Technologies. **Contact.** contact@xsyda.com · research.xsyda.com