ChatGPT

Practical and Physical Ceilings of Computation

Diminishing Returns in CMOS Scaling (IRDS Roadmaps and 3D Integration)

For decades, the semiconductor industry relied on Moore's Law and Dennard scaling to improve performance by shrinking transistors, but these benefits are now hitting practical limits. At the cutting edge (5 nm and below), each new node provides ever-smaller gains in speed or energy-efficiency – on the order of only ~15–20% improvement per generation, whereas earlier nodes yielded ~30–50% gains [1]. For example, TSMC's upcoming 2 nm node (2025) is expected to increase transistor density by only ~15–20% over 3 nm and give ~10–15% speed gain (or ~30% power reduction) at iso-performance [2]. In fact, analysts estimate TSMC 2 nm will reach ~3.13×10^8 transistors/mm^2 (a ~1.3× density bump over 3 nm), but with ~10% higher wafer cost and greatly increased design complexity – meaning essentially *zero* cost-per-transistor reduction [3]. This reflects a regime of **diminishing returns**: transistor miniaturization still technically continues (into so-called "1.4 nm" and "1 nm" generations by ~2027–2030), but improvements in power-performance-area (PPA) are getting marginal and extremely expensive to achieve [2].

Several physical factors underlie this slowdown. *Electrical* limits include the inability to scale supply voltage much below ~0.7 V due to device threshold physics – so dynamic energy per switch (~CV^2) now bottoms out in the femtojoule range, rather than dropping ≈ half each node as in the Dennard era. Likewise, leakage currents grow exponentially as transistors get thinner, causing **static power** to dominate at nanoscale gate lengths [1]. Thus, while transistor *counts* can still rise, the *energy per operation* is flattening out. Today's best logic transistors dissipate on the order of 0.1–1 femtojoules per bit-switch, and projections suggest a practical floor around ~0.2 fJ per logic transition for conventional CMOS [4]. This is only about one order of magnitude above fundamental thermal limits (Landauer's limit is kT·ln2 ≈ 2.8×10^−21 J at room temperature), so further reduction becomes exceedingly difficult without new physics. In short, we are squeezing the last gains from CMOS scaling: even as transistor geometries push into the "1 nm" class, they yield diminishing energy efficiency returns and greatly increased cost and variability.

3D integration is a critical part of current roadmaps to extend Moore's Law beyond planar scaling. Advanced 3D packaging (chiplets and through-silicon vias) is already used to stack memory on logic (e.g. HBM on GPUs) and to integrate chiplets in 2.5D/3D systems, doubling bandwidth and halving energy per bit of I/O [5] [6]. The IRDS roadmaps foresee *monolithic* 3D CMOS in the late 2020s: for logic, this means moving from today's 3D transistors (FinFETs, then Gate-All-Around nanosheets) to true vertical stacking of device layers. Indeed, by ~2028 the standard nanosheet FET is expected to reach its limits, and manufacturers plan to adopt **complementary FET (CFET)** devices around the 1 nm node – essentially building a p-FET atop an n-FET in a vertical stack [7]. IMEC and others have demonstrated forksheet and CFET prototypes that could reduce cell area ~2× by directly stacking transistors [8] [9]. This 3D VLSI approach could extend density scaling a few more generations [10]. However, 3D logic integration faces a major challenge: **heat density**. Stacking active layers increases power density (W/cm^2) and makes cooling far harder, since inner layers are insulated. Even 2-layer logic-on-logic stacks may require new thermal materials or active cooling to avoid overheating. Thus, while 3D stacking (along with new transistor

materials like SiGe channels or 2D semiconductors) will push the semiconductor roadmap into the 2030s, it primarily helps **integration density**. It does not fundamentally lower the energy per switching operation – and removing heat from a 3D-integrated chip becomes a limiting factor. In summary, CMOS is approaching a practical ceiling: ~10^9 transistors/mm^2 devices at ~1 nm features, ~0.5–1 fJ per logic op, and ~5 GHz clocks. Pushing beyond that by brute-force scaling hits steeply diminishing returns in efficiency and cost [2] [1], forcing a transition to new architectures (chiplets, 3D) and ultimately new computing paradigms.

Nearing Fundamental Limits: Landauer's Limit and Koomey's Law Trends

Every bit operation in a computer dissipates some irreducible energy due to thermodynamics. Landauer's principle (1961) quantifies this minimum energy for an irreversible bit-flip or erase as $E_{min} = kT\ln 2$. At room temperature (300 K) this is ≈2.8×10^−21 J (about 18 meV) [11] [12]. Modern digital computing is still **orders of magnitude** above this fundamental limit. For instance, a typical modern processor switching at ~1 GHz dissipates on the order of 10^−11 J per bit operation – i.e. roughly *10 billion times* the Landauer limit [13]. In other words, current CMOS logic might consume ~10^4–10^5 kT of energy for each bit processed. This gap represents the "headroom" for potential efficiency gains, though in practice much of it is unusable due to other noise and device constraints.

Historically, computing efficiency (operations per joule) improved exponentially. **Koomey's Law** observed that from the 1940s through about 2000, the energy efficiency of computers doubled roughly every 1.5 years (≈100× per decade) [14]. This was a byproduct of Moore's Law and Dennard scaling – smaller transistors running at lower voltage yielded more ops per joule. After ~2005, however, Koomey's trend **slowed**: since 2000 the doubling time stretched to ~2.6 years (only ~16× per decade) [15]. This slowdown directly correlates with the end of voltage scaling and the plateau of CPU clock rates around the mid-2000s [16]. In practical terms, a 2022 smartphone or server is about *10^4*× more energy-efficient than one from 2000, but the annual improvement rate is now much lower than in the late 20th century. The absolute leader in efficiency as of early 2020s – the top supercomputer on the Green500 list – achieves ~50 billion FLOPs per watt (e.g. 52.23 GFLOPS/W for the Frontier system) [17]. Even so, if we project the *current* 2.6-year doubling forward, the Landauer limit would only be reached around **2080** [18]. At that point (roughly 10^6× beyond today's efficiency), each logic operation would dissipate on the order of kT and further progress via irreversible computing must cease. In reality, well before 2080 we expect diminishing returns – and alternative paradigms to take over – because maintaining even the slower Koomey's Law trajectory for another 50 years seems unlikely within classical CMOS constraints.

How much improvement is *theoretically* left in conventional computing? Recent analyses indicate there may still be **50× to 1000×** room for better energy efficiency in CMOS logic [19]. These gains could come from aggressive use of near-threshold logic, improved circuit architectures, better parallelism (to drop voltage/frequency), and reducing overheads like interconnect and memory access. For example, a 2023 study modeled an idealized CMOS processor and found an upper bound around ~3×10^14 FLOP/J at 16-bit precision (i.e. 3e14 operations per joule, vs. ~1e12–1e13 in today's best GPUs) [20]. Achieving such gains would likely require novel design (e.g. more on-chip memory, optical interconnects, etc.) but it's *not* forbidden by physics – it mainly requires using most transistors in an active, but low-voltage, way without idle leakage or memory bottlenecks. In practice, however, these improvements become harder as we approach fundamental noise floors. At some point, each bit operation will necessarily dissipate a few kT of energy to maintain signal-to-noise and overcome thermal noise.

One key insight is that **reversible computing** can evade Landauer's limit. Landauer's bound only applies to *irreversible* logic operations (those that erase information). A reversible (bijective) operation theoretically dissipates no entropy to heat. Modern computers are overwhelmingly irreversible, but research into reversible logic (adiabatic circuits, reversible gates, etc.) aims to reduce the kT ln2 dissipation per operation. Koomey himself noted that if we continued the 2.6-year doubling, we'd hit Landauer's limit by ~2080 *absent* new physics – but *"Landauer's principle, however, does not constrain the efficiency of reversible computing."* [18]. In other words, to continue exponential efficiency gains beyond the thermodynamic limit, we would need to adopt reversible logic or other beyond-CMOS technologies. Indeed, as we'll see below, some experimental technologies (like superconducting adiabatic logic) have demonstrated switching energies below 1 kT per bit by avoiding irreversible dissipation. In summary, **modern computers are still millions of times less efficient than the absolute physical limit**, but the gap is closing. The easy gains have slowed since 2010, and pushing much closer to Landauer will demand fundamentally new computing paradigms or a re-imagining of logic reversibility and device physics.

Emerging Paradigms to Raise the Compute Ceiling

With traditional CMOS hitting scaling and efficiency walls, a variety of emerging computing paradigms promise to push the performance and energy "ceiling" upward. These include optical/photonic computing, quantum computing, neuromorphic and in-memory architectures, thermodynamic (stochastic) computing, and reversible or superconducting logic. Each of these approaches exploits different physics and offers quantitative advantages in certain metrics (energy per operation, switching speed, parallelism), albeit with their own engineering limitations. Below we analyze each paradigm, including rough performance/energy figures and key physical constraints.

Photonic Computing (Optical Logic and Interconnects)

Photonic computing uses photons in optical circuits to perform computations, rather than electrons in silicon transistors. The chief advantages of light are its speed (optical signals propagate ~10^5 times faster than electrical signals in wires) and the absence of resistive Joule heating in passive optical propagation. This makes *optical interconnects* extremely attractive for communication – indeed, data centers already use photonic links extensively for rack-to-rack bandwidth without thermal loss. The bigger challenge is using optics for general logic computation. Photons do not naturally interact (linear optical systems won't easily implement logic operations like AND), so achieving nonlinearity typically requires optical materials or converting photons to electronic signals and back [21] [22]. Nonetheless, specialized optical processors have been demonstrated, often for analog computing tasks like matrix multiplication (e.g. for neural network inference). Recent **optical neural network** chips use Mach-Zehnder interferometers or microring modulators to perform vector-matrix multiplies in the analog domain with potentially **ultra-high throughput**. For example, a 2022 research optical neural net achieved on the order of 10 fJ per MAC (8-bit multiply-accumulate) including laser inefficiencies [23], meaning an ideal photonic matrix multiplier could perform ~10^14 operations per joule – about 10× better than state-of-the-art electronic DSPs. In principle, if sources and modulators were made more efficient (e.g. <1 fJ per operation) [24], an optical accelerator operating at THz-scale modulation could reach ~10^15–10^16 ops/J (far beyond CMOS). Throughput is a major strong suit of photonics: using wavelength-division multiplexing, many parallel operations can be done at once. Devices have demonstrated >10 terabit/s data transmissions, and an optical Fourier processor can execute convolutions at petabyte/s data rates.

However, photonic computing faces **significant limitations**. The physical size of optical components (set by the wavelength, on the order of microns) is much larger than transistors, so integration density is lower – meaning pure optical processors might be large in area for equivalent logic complexity. Also, while moving photons is energy-free, *modulating* or detecting them typically involves electro-optic effects or photodetectors that consume energy. For instance, modulators might use ~1 fJ per bit switched in advanced designs, and photodetectors dissipate energy converting photons to electrical current. There is also the issue of **memory:** photons are great for fast parallel operations, but we lack dense optical RAM, so most proposals use electronics to store and feed data. This hybrid approach still incurs electrical energy costs. Additionally, analog optical computing accumulates noise and errors (photons shot noise, component variations), limiting precision – techniques like partitioning numbers or using optical ADCs are needed to get digital-accurate results [25]. Overall, photonic computing will likely excel in niche areas requiring massive parallelism and bandwidth (optical AI accelerators, ultrafast signal processing), potentially offering an order-of-magnitude better energy per operation for those tasks [26]. For example, startup Lightmatter reported that its photonic matrix-multiply engine can achieve up to 10× higher throughput/W on AI inference than Nvidia's electronic GPUs [26]. Nonetheless, photonics won't completely replace CMOS; rather it can offload certain computations. *Quantitatively*, one can expect future on-chip optical compute units performing ~10^13–10^15 ops/s with femtojoule-level energy per op, but constrained to specific algorithms (linear algebra, FFT, etc.) and integrated alongside conventional logic. The fundamental ceiling for photonic logic is high – potentially each photon could represent a bit with energy as low as a single photon's energy (e.g. a 1.5 μm telecom photon has ~0.8 eV = 1.3×10^−19 J). In theory, an optimally designed optical logic device might operate near that scale. But in practice, until we have *highly nonlinear optical materials or novel photonic transistors*, photonic computing will be an **augmenting technology** to boost bandwidth and reduce interconnect and certain matrix-operation energies, rather than a general solution.

Quantum Computing

Quantum computing fundamentally departs from classical binary logic by using two-level quantum states (*qubits*) that can exist in superposition and become entangled. The theoretical speedup of quantum algorithms is in *algorithmic complexity* (e.g. Shor's algorithm factoring in polynomial time vs exponential classically), not in raw GHz clock speed. However, in terms of physical operation rates and energy, quantum computing has interesting characteristics. **Unitary quantum gates** (the basic operations on qubits) are in principle *reversible and lossless*. A perfect quantum gate (e.g. a rotation of a qubit's state) dissipates no heat directly, since it's theoretically just evolving a closed system via Schrödinger's equation [27]. In contrast to a classical NAND gate that irreversibly discards a bit (and thus must generate $\geq kT \ln 2$ heat), a quantum gate is entropy-conserving (aside from error processes). *This means that at the fundamental level, a quantum computer could perform logic operations with essentially zero thermodynamic energy cost – limited only by control overhead.* For instance, flipping a superconducting qubit's state can be done by a resonant microwave pulse that, if done adiabatically, need dissipate arbitrarily little energy (some femtojoules or less potentially). This *reversible computing advantage* is a key reason quantum computing is not bound by Landauer's limit in the same way – if we could isolate a quantum computer perfectly, it would not produce heat from the logical ops [27].

In practice, current quantum computers are very far from this ideal. Real devices have significant overhead energy costs for control and cooling. Superconducting qubit systems must be cooled to ~10–20 mK; a dilution refrigerator might consume **10–30 kW** of power to support a cryogenic quantum processor with only tens or hundreds of qubits. That's an enormous energy overhead for a small number of operations – effectively *billions of joules per quantum operation* if calculated naively for present machines. Even if we

count just the dynamic control energy: a typical microwave pulse to drive a qubit might carry on the order of nanojoules of energy (microsecond pulses at ~μW power) and there may be many such pulses per logical gate including error correction. Ion trap qubits similarly require multiple laser beams, each maybe tens of mW, and thousands of operations, adding up to millijoules per operation in current labs [28]. Thus, **today's quantum prototypes are far less energy-efficient than classical computers** for any practical task [29] [28]. The motivation for quantum computing is not per-operation energy efficiency *at the current scale*, but the ability to solve intractable problems with exponentially fewer operations. For example, factoring a 2048-bit number might take ~10^12 classical ops but perhaps ~10^8 quantum ops. If each quantum op took, say, 10^−6 J (mega–kT) including overhead, that's 10^2 J total, versus maybe 10^12 ops × 10^−12 J/op = 1 J classically (if one could even run that many ops efficiently). In this contrived example the classical wins, but as problem size grows the quantum algorithm's lower complexity should dominate despite higher per-op cost.

Looking forward, **could quantum surpass classical in energy at scale?** It's conceivable in specific domains. If error-corrected quantum computers are built with millions of qubits, they will still require cryogenic or other exotic infrastructure that might consume megawatts. However, if those machines can solve problems that would take billions of classical CPU-hours, the *energy per solution* could be lower. One analysis suggests that for certain optimization or simulation tasks, a quantum annealer or analog quantum machine can find answers with orders of magnitude less energy than a brute-force classical search, by leveraging the physics of the device's natural evolution [30]. But these are niche cases. In general, **quantum computing's ceiling** is not "ops per second" in the usual sense, since quantum ops are not directly comparable to classical logic ops. The clock speeds are relatively low (~MHz for gate-based superconducting qubits, ~kHz for ion traps), but the computational power comes from state-space size rather than raw op frequency. The eventual vision is a quantum computer with perhaps 10^9 operations per second (e.g. 1 GHz on 1000 parallel logical qubits) that can do tasks infeasible for a 10^15 OPS classical supercomputer. Energy-wise, if those 10^9 ops/s could be done with, say, 1 kW of power including cooling, that's 10^6 ops/J – far lower than classical. But because each quantum op can replace many classical ops, the effective energy per *equivalent* task might be very favorable. Ultimately, quantum machines are likely to remain power-hungry (cryogenics, control electronics, etc.) and will be housed in data-center-like facilities, not in mobile devices. They are driven by capability (solving new classes of problems) rather than raw FLOPS/W. In summary, quantum computing can be seen as *orthogonal* to the power-efficiency race: it sidesteps Landauer's limit by using reversible unitary processes [27], but the overheads and limited use-cases mean it's not a general solution to extending classical computing efficiency for everyday applications.

Neuromorphic and In-Memory Computing

Neuromorphic computing takes inspiration from biological brains to process information with networks of "neurons" and "synapses," often using spiking communication and analog computation. The goal is to achieve dramatically better energy efficiency on certain tasks (e.g. sensory perception, pattern recognition) by using many simple, low-power operations in parallel – essentially trading off precision and determinism for energy savings. A key advantage is minimizing data movement: neuromorphic architectures often colocate memory and computation (synaptic weights are stored locally) and operate in an **event-driven** manner (nothing happens unless an event/spike occurs). This can slash the energy wasted on fetching data or clocking idle circuits. For instance, IBM's TrueNorth (2014) was a digital neuromorphic chip with 1 million spiking neurons (256 million synapses) that operated on a mere 70 mW. It was capable of ~46 billion synaptic operations per second at that power [31] – roughly **0.66 trillion ops per joule**, or 0.66 TOPS/W, albeit for very low-precision integrate-and-fire operations. Similarly, Intel's Loihi spiking chip (2018) achieved

on the order of 10^7 neuron updates per second with an energy cost of ~23–25 pJ per synaptic event [32]. That is about 40 million events per joule – significantly more efficient than a general-purpose CPU computing the same neural network, but in the same ballpark as modern digital AI accelerators. In fact, 20–50 pJ per operation is not revolutionary; an NVIDIA GPU can perform one 8-bit MAC in ~<100 pJ. The difference is that neuromorphic systems keep power low by sparse activation: if only 1% of neurons fire at any given time, the effective operations-per-joule for the *task* is very high (no energy wasted on the 99% that are inactive). In short, neuromorphic hardware excels for *sparse, event-driven workloads*, potentially yielding 10–100× improvements in energy efficiency for those use cases [33]. Quantitatively, today's neuromorphic prototypes achieve ~10^3–10^4 OPS/W in worst-case dense activity, but ~10^5–10^6 OPS/W or more in typical sparse workloads – a benefit that scales with problem sparsity.

A related concept is **in-memory computing** (sometimes called *processing-in-memory*, PIM, or *compute-in-memory*, CIM). This approach tackles the von Neumann bottleneck by performing arithmetic directly where the data resides (in memory arrays), rather than shuttling data back-and-forth to a central processor. The most aggressive version is **analog in-memory computing** using memory device physics to compute. For example, resistive crossbar arrays of memristors or FeFETs can inherently perform matrix-vector multiplication by exploiting Ohm's law and Kirchhoff's law – effectively computing $\mathbf{y} = W \cdot \mathbf{x}$ in one analog step as currents sum on bitlines. This massively parallel analog MAC operation can be extremely energy-efficient because it avoids digital switching for each multiply. Recent research demonstration chips have shown **record energy efficiencies**: e.g. a ferroelectric FeFET-based crossbar achieved *885.4 TOPS/W* on 2-bit multiplication-accumulations [34]. That corresponds to each 2×2-bit MAC consuming only ~$1.1×10^{-15}$ J (1.1 femtojoule) [34] – several hundred times lower energy than a typical 8-bit digital MAC (which might be ~0.1–1×10^−12 J). This illustrates the astounding potential of analog in-memory compute: by essentially "doing 1,024 operations in parallel with the same energy as 1 operation," it divides per-operation energy dramatically. In practice, of course, there are caveats: analog computing introduces noise and requires analog-to-digital conversion (ADC) to read out results, which can eat up a lot of power if high precision is needed. For instance, a crossbar might do a matrix multiply in 1 fJ per weight operation, but the ADC to get a 16-bit result might consume nanojoules, negating advantages unless precision is kept low. Thus, current CIM chips focus on low/medium precision tasks (6–8 bits or analog signals for inference). They are showing impressive efficiency for AI inference, where ~4–8-bit precision suffices. A recent compute-in-memory ASIC called NeuRRAM demonstrated fully parallel analog computing with 8-bit precision at ~30–60 TOPS/W for vector-matrix multiplies [35], far beyond what digital logic can do in large matrices.

Physical limits and challenges: Neuromorphic and in-memory paradigms aim to approach the brain's efficiency. The human brain is estimated to perform ~10^{14} ops/sec (synaptic events) with only ~20 W of power – that's ~5×10^12 ops/J, or in the same order as 10^12–10^13 ops/J. This is about 1–2 orders of magnitude above today's best specialized silicon (~10^10–10^11 ops/J), showing that further 10–100× gains may be possible by increasingly brain-like computation (massive parallelism, sparse coding, analog processing). However, the brain achieves this with ~10^15 synapses operating at <1 Hz average firing – extremely low activity and ~15 ms latency tolerance. Silicon systems typically run 10^6× faster but at higher energy per event. Bridging that gap is non-trivial. Neuromorphic hardware is pushing toward sub-pJ per synaptic event and event rates of ~MHz per neuron, but hitting diminishing returns due to device noise and memory limitations (e.g. SRAM or NOR flash to store weights still costs energy). Memristive analog devices could further lower energy per op to the attojoule (~10^−18 J) scale if operated quasi-adiabatically, but stochastic device variation may introduce errors. Some proposals even suggest operating near thermal noise limits, using probabilistic bit flips to perform computation with average energies approaching kT

(~10^−21 J). In summary, neuromorphic and PIM approaches can **drastically improve energy efficiency** for certain workloads – realistically by a factor of 10–100× over general-purpose CPUs, as demonstrated by research chips [34]. The physical ceiling here is not a single number; rather it is about moving toward *thermodynamically reversible analog computing*. If one imagines an ideal analog memory element updated with arbitrarily small currents over arbitrarily long time (adiabatic switching), one could approach Landauer's limit per update. But timing requirements and noise make that impractical for high speeds. Thus, the eventual ceiling might be something like 10^{15} ops/J at room temperature for low-precision operations (i.e. on the order of 100 kT per op). Achieving that would likely require *extensive parallelism and locality* (which neuromorphic/PIM provides) and careful management of analog noise.

Thermodynamic & Stochastic Computing

A more radical approach to computing efficiency is to harness random thermal dynamics or naturally occurring processes to perform computation – effectively letting physics "solve" the problem at extremely low energy cost. We call this category **thermodynamic computing** or **stochastic computing** (not to be confused with the digital stochastic bit-stream computing technique). The idea is to set up a physical system such that its *minimal-energy state corresponds to the solution* of a problem, or to use random fluctuations to explore computational state space. Examples include: stochastic annealing processors (like optical or electronic **Ising machines**), probabilistic bits (**p-bits**) that continuously flip like a biased coin to find satisfiability solutions, and even biochemical or **biological computing** where molecules explore many possibilities in parallel.

A striking experimental demonstration comes from *biocomputation*. In one example, researchers built a network-based biocomputer using molecular motors and microtubule filaments to solve a maze (a combinatorial task) in parallel [36]. The bio-motors consume ATP (chemical energy) to move filaments through the nanofabricated maze, exploring paths. Because many filaments (billions) explore simultaneously at a slow pace, the energy per computation is extremely low. **Experiments confirm a biocomputer of this type uses 1,000–10,000× less energy per operation than a conventional electronic processor** [37]. The tradeoff is speed: those molecular "processors" take seconds to perform what a 3 GHz CPU might in nanoseconds. In the reported work, the motor proteins step only a few hundred times per second (each step being an operation) – about *a million times slower* than transistor logic [38]. Yet the energy per step is so minuscule (each step fueled by one ATP molecule ≈ 8×10^−20 J) that the overall energy efficiency is astounding. Similarly, **probabilistic computing with p-bits** (implemented e.g. by low-energy barrier nanomagnets and CMOS interface) has shown the ability to solve certain optimization problems with far lower energy than deterministic logic by effectively *allowing thermal noise to do the searching*. A p-bit flips randomly at rates of MHz, consuming perhaps femtojoules per flip, and a network of coupled p-bits can naturally settle to solutions of, say, a SAT problem, with the randomness providing an inherent search mechanism [39]. One study demonstrated an experimental 8-bit p-computer that solved a combinatorial optimization using <1% of the energy a deterministic approach would require, by virtue of the energy physics guiding the solution search [40] [41].

The **Landauer limit** becomes approachable in these paradigms because we are intentionally working at the edge of thermal noise. In a stochastic bit, flipping a bit with energy comparable to kT (≈ 4×10^−21 J at room T) means the flip is essentially random, biased only slightly by input signals. This is undesirable in a deterministic logic gate (since it would be error-prone), but in a stochastic algorithm it can be acceptable or even useful. The ultimate limit would be a computing scheme where each logical operation (or bit update) dissipates on the order of *a few kT* of energy – say 10^−21 to 10^−20 J. At that level, one watt of power could

perform ~10^20 operations per second! No electronic device today comes within many orders of magnitude of that in a reliable way. But research in *reversible thermal logic* (e.g. using nanoparticles or fluidic logic) aims to show basic gates operating at <100 kT of energy. For instance, experiments in reversible nanomagnetic logic have demonstrated bit operations with <1 attojoule (10^−18 J) dissipation by relaxing a magnetization with almost no energy loss except the Landauer erasure at the end [12] [42]. The caveat is these operations took milliseconds – again trading speed for energy.

In practical terms, **thermodynamic computing paradigms will not replace general high-speed computing**, but they could become important for solving specific hard problems or for ultra-low-power IoT devices that tolerate latency. An Ising machine (optical, electronic, or quantum annealer) effectively finds low-energy states of a spin system – which can correspond to the optimum of an NP-hard optimization – often faster or more energy efficiently than exhaustive search. For example, a specialized CMOS annealing chip by Fujitsu (the Digital Annealer) or quantum annealers by D-Wave can solve certain optimization instances using orders of magnitude less energy than running a big server cluster on the same problem, because the physics does the work. The energy per "operation" in such devices is not well-defined (since they evolve analog continuously), but effective optimization can occur with just a few thermal kT of dissipation per variable update. **In summary**, thermodynamic and stochastic computing explores the bottom of the energy-per-operation curve by embracing randomness and physical analog computing. The ceilings it breaks are in *energy efficiency per operation*, potentially achieving 10^3–10^4× improvements over deterministic logic [37], but always with the catch of either longer time or specialized use cases. Ultimately, they highlight a path to computation at the Landauer limit – but usually only by *running the computation more slowly or in parallel on massive scales*, so that less energy is spent per step. As the adage goes, "slow is smooth and smooth is fast" in thermodynamic computing: by operating near equilibrium (slowly), you minimize dissipation; by running many processes in parallel, you still get answers in reasonable time [43] [44].

Reversible and Superconducting Computing

To truly continue improvements in compute performance once irreversible CMOS hits its wall, **reversible computing** is a tantalizing option. Reversible logic (in principle) allows arbitrarily small energy dissipation per operation, because it can approach the adiabatic limit where no information is erased. However, making a reversible logic family that is also fast and practical has been challenging. One promising direction is **superconducting computing**, particularly adiabatic superconducting logic, which exploits the lossless current flow and quantum effects in superconductors. Superconducting circuits can switch using Josephson junctions with **picojoule or sub-picojoule energies**, and using adiabatic clocking, the energy can be recovered instead of burned as heat.

A state-of-the-art approach is the **Adiabatic Quantum-Flux Parametron (AQFP)** logic family. AQFP is a reversible (or at least adiabatic) superconductor logic that uses an AC-powered inductive network to gently toggle Josephson junctions. In 2019, researchers demonstrated AQFP gates with an incredibly low switching energy of **1.4 zeptojoules (zJ) per junction** at 4.2 K [45]. 1.4 zJ is $1.4×10^{-21}$ J – on the same order as the room-temperature Landauer limit (which is ~2.8 zJ at 300 K), but here the device is at liquid helium temperature so kT is smaller (~5×10^−23 J at 4 K). Still, 1.4 zJ is only ~30× kT at 4 K, meaning these junctions operate remarkably close to the thermodynamic limit in that environment. Even more impressively, AQFP can switch at clock frequencies of a few GHz while maintaining those energies [46]. This is possible because the circuit is biasing the junctions with an AC waveform that recovers energy each cycle. Essentially, the

junction potential energy is raised and lowered slowly (relative to its L/R time) so that the system remains near equilibrium and dissipates minimal heat – a classic adiabatic process.

Of course, **cooling overhead** must be considered: running a chip at 4 K requires a cryocooler, which typically has an efficiency of a few hundred (i.e. 1 W of cooling costs ~200–300 W of wall-plug power for a closed-cycle fridge at 4 K). Even so, AQFP comes out ahead. Studies show that even after accounting for refrigeration overhead, a system built from AQFP could be **~80× more energy-efficient** than a 7 nm CMOS system at the same task [47]. This was demonstrated by a detailed analysis comparing an adiabatic superconductor processor to a FinFET CMOS design for a benchmark, indicating the superconducting approach, though cryogenic, would save roughly two orders of magnitude in energy [47]. The catch is that AQFP (and other superconducting logics like RSFQ, RQL, etc.) currently require large physical area per gate and very low temperatures, and they haven't yet scaled to the complexity of a modern CPU or GPU. There's active research on 3D integration of superconducting chips and new materials (e.g. high-Tc superconductors or integrating superconducting logic with cryogenic memory) to address these issues.

The **physical/engineering limits** for reversible superconducting logic are extremely high. In theory, a reversible computer at low temperature could perform arbitrarily many operations per joule by slowing the clock – truly having no fundamental ceiling on ops/J aside from coherence and leakage. In practice, things like residual resistance, junction capacitances, and clock network losses set a floor. AQFP's ~10^−21 J per gate is a real data point; perhaps with further refinement and lower temperature, one could push below 10^−22 J (which would be below kT at 4 K, essentially reversible). The switching speed can still be GHz+, so the throughput per device is not sacrificed terribly. A potential "ceiling" to consider is integration density and critical current limits – you can only pack so many Josephson junctions before the wiring inductance and parasitics limit speed. But from an energy perspective, an **adiabatic superconducting processor might achieve 10^4–10^5 operations per joule per junction at 4 K** (since 1.4 zJ is ~0.07 of kT_room, that's equivalent to ~0.07 *per* operation in units of the RT Landauer limit). Scaled up, a well-designed system could conceivably perform on the order of 10^17 operations per joule (far exceeding anything possible in CMOS) if one neglects cooling costs. Including a factor for cooling (say ×300), you might still get ~10^14 ops/J effective – which is astounding. These are speculative numbers, but they illustrate that **reversible superconducting logic could extend computing efficiency by several orders of magnitude** beyond the CMOS paradigm.

In summary, reversible and superconducting computing offers a path to continue performance scaling when CMOS hits the wall. Superconducting chips operating at a few kelvin have already demonstrated *bit energies on the order of 10^−21–10^−19 J* [48] [45], many orders lower than room-temp electronics. The main limitations are practical: the need for cryogenics, low memory density (no superconducting DRAM of similar density to CMOS SRAM yet), and the complexity of designing reversible architectures (which often require keeping track of garbage bits, etc.). But as a complement to CMOS, one could envision data-center accelerators based on superconducting logic for extremely energy-intensive tasks. Reversible computing reminds us that Landauer's limit is not the ultimate wall – it's only the wall for *irreversible* logic. By sidestepping irreversibility, computing can, in principle, continue to improve until other limits (like speed of light or quantum noise) kick in. It is an exciting area where current research is literally operating at the intersection of computing and thermodynamics.

(For clarity, Table 1 summarizes the above paradigms' estimated performance and energy characteristics.)

Paradigm	Energy per Operation	Throughput Potential	Key Limitations
CMOS (current)	~10^−14–10^−15 J per 32-bit op (few fJ)	~10^9 ops/s per core @ 3 GHz	Hitting V_{dd} scaling limits; high heat density
Photonic computing	~10^−15–10^−14 J per MAC (with efficient modulators) [23]	10^12–10^15 ops/s (massively parallel, high bandwidth)	Limited by conversion overhead, device size, precision
Quantum computing	Ideally ≪10^−21 J (unitary gate dissipation ~0) [27]; *practically* 10^−6–10^−9 J per gate (current tech)	~10^6–10^8 ops/s for gate-based QC (limited by decoherence and control speeds)	Huge overhead for error correction; only specialized algorithms give speedup
Neuromorphic (spiking)	~10^−11 J per event (10–50 pJ) [32] (digital); *potential* ~10^−14 J (10 fJ) analog synapse	10^7–10^9 events/s per chip (many in parallel across neurons)	Workload-specific; low precision; sparse activity needed for gains
In-memory analog	~10^−15 J per 2-bit MAC (demonstrated) [34]	10^12–10^13 ops/s per compute array (high parallelism)	ADC/DAC overhead; noise and variability; limited precision
Thermodynamic (stochastic)	~10^−21 J per bit-op (kT scale) in principle; 10^−19–10^−18 J in experimental reversible magnets [12]	10^2–10^6 ops/s (very slow per device, but can massively parallelize)	Very low speed per device; requires novel architectures and often analog readout
Superconducting (AQFP)	1.4×10^−21 J per JJ switching (at 4 K) [45]	10^9 ops/s per gate (multi-GHz possible)	Cryogenic cooling needed; low density; sync/clock overhead (adiabatic phases)

Table 1: Estimated energy dissipation and performance scales for various computing paradigms (in comparison to ~few-fJ CMOS). Photonic and superconducting figures are for demonstrated or projected devices; quantum figure assumes gate operations (not energy per algorithm). These paradigms excel in different regimes: photonics and in-memory excel at parallel throughput, quantum at algorithmic speedup, neuromorphic at event-efficient processing, stochastic at energy per op, and superconducting/reversible at approaching thermodynamic limits.

Compute Capacity Ceilings in Mobile Devices (Smartphones, Wearables, AR)

The practical limits of computation manifest very differently in **mobile/edge devices** compared to large data centers. In mobile form factors – a smartphone (≈5 W power budget), a smartwatch (≈0.5 W), or AR glasses (≈2–3 W) – the dominant constraints are thermal dissipation and battery energy. Even if chip technology allows billions of operations per second, a phone or wearable cannot sustain high power

without overheating the user or draining the battery quickly. Here we analyze the maximum compute performance sustainable in these mobile classes, both today and projected into the future (~2035), along with the eventual physical ceilings.

Smartphones (~5 W thermal budget)

Modern high-end smartphones already operate at the edge of what passive cooling can handle in a hand-held device. A typical smartphone SoC (System-on-Chip) might have a TDP around 4–6 W. At this power, it can include multi-core CPUs, GPUs, and dedicated NPUs (Neural Processing Units). For example, a recent 5 nm smartphone SoC can deliver on the order of **~1 trillion integer operations per second (1 TOPS)** on its NPU at ~5 W (that's 0.2 TOPS/W for 8-bit ops), and perhaps around 200–300 GFLOPS/sec in FP32 on the GPU. In ML inference scenarios with lower precision (int8/int4), some smartphone chips reach **~10–15 TOPS (total)** using their NPUs at ~5 W [26]. This translates to about 2–3 TOPS/W – notably lower efficiency than specialized data-center accelerators, but impressive for battery-powered devices. This performance is enough for tasks like real-time image recognition on-device.

The **2035 horizon** for smartphones (10 years out, assuming further node advances to maybe the 1–2 nm node and use of 3D stacking) could plausibly see on the order of *50–100 TOPS* of AI performance in a handset, and perhaps 1–5 TFLOPS of general FP32 performance. This assumes continuing (albeit slowed) improvements in energy efficiency – roughly a 5–10× boost in ops/J over today. Such gains might come from 3D-stacked logic and memory (drastically reducing data movement energy) and more specialized accelerators. Indeed, IRDS projections emphasize **3D integration of memory-on-logic** by ~2030 to break the memory bandwidth wall [5] [6]. If a smartphone in 2035 has dense 3D-stacked SRAM or MRAM feeding an NPU, it could reduce memory access energy per bit by ~4× (from ~1 pJ to ~0.2 pJ) and allow more of the 5 W budget to go into actual computations. By that time, transistors might be a bit more efficient (maybe 0.5× the capacitance of today's, and slightly lower Vdd), giving another ~2×. So a rough estimate: 5 W could provide ~10× the operations of a 2025 chip. That lands in the tens of TOPS range for low-precision AI ops (for example, 5 W × 10^16 ops/J = 5×10^16 ops/s, which is 50 TOPS). Some optimistic forecasts even imagine mobile devices doing **~0.1 peta-ops/s (10^14 OPS)** by late 2030s in burst mode, using advanced packaging and near-Landauer devices. But sustained performance will be limited by heat – without active cooling, ~5 W is the safe envelope to avoid burning the user's hand (surface temperatures above ~45 °C are uncomfortable).

The **physical limit** for a smartphone's compute is extremely high if considering Landauer's limit: at 5 W, if each bit-operation were at the theoretical minimum ~2.8×10^−21 J, you could do ~1.8×10^21 ops/s – that's ~1.8×10^9 TOPS! Of course, this is an absurdly hypothetical number – no real device will ever operate all its transistors at kT-level energies and 100% utilization. The true "ceiling" will be set by engineering: thermal dissipation per cm^2 of phone area, battery capacity, etc. Already, phones throttle CPU/GPU speed within minutes if heavy workloads push sustained >4–5 W, because the device heats up (a typical phone might hit ~45 °C and then limit performance to cool down). Exotic solutions like internal vapor chambers, or even active cooling fans (as seen in some gaming smartphones), can raise the sustainable power a bit (some get up to ~8–10 W for short periods). But fundamentally, unless we invent skin-safe microfluidic cooling in phones, ~5–10 W is the limit a handheld can continuously dissipate.

Thus, **the maximum sustainable compute in a smartphone** by 2035 might be on the order of 0.1–0.2 TFLOPS (in FP64), 1–2 TFLOPS (FP16/32), or ~50–100 TOPS for int8 operations. This would enable on-device AI like real-time video analytics, AR, etc., that today require a server. But it's still many orders below

what a data-center rack can do. To push beyond that in mobile would require either a breakthrough in cooling (unlikely, given human comfort constraints) or a paradigm shift to ultra-efficient computing (e.g. a phone that uses reversible logic operating near zero dissipation – far future speculation).

Smartwatches and Wearables (~0.5 W)

Wearables like smartwatches and fitness trackers are even more constrained. A typical smartwatch might only allow ~0.2–0.5 W for the SoC to avoid the device getting warm on the wrist (and to keep battery life in hours, not minutes). Today's smartwatch chips (often built on older nodes like 7–14 nm for cost) deliver maybe on the order of **~1–10 GFLOPS** of performance or a few billion ops/sec in simpler tasks at these power levels. For example, an Apple Watch might have a small neural engine performing maybe **0.1–0.2 TOPS** (10^11–10^11 ops/s) for always-on AI tasks, consuming a couple hundred milliwatts. These devices prioritize extreme power efficiency (in terms of joules per task) over raw throughput.

By 2035, wearables will likely benefit from advanced nodes and maybe integration of specialized low-power AI accelerators. We could envision a smartwatch SoC on a 3 nm or below process, with perhaps **10–20 TOPS/W** efficiency at low precision (since lower clock, near-threshold operation is feasible to save power). At 0.5 W, that translates to maybe **5–10 TOPS peak** in a future smartwatch, or on the order of 10^12 ops/s. In terms of FLOPS, perhaps ~50–100 GFLOPS of FP16 could be available for AR/VR or health algorithms on a watch by that time. Achieving this would require aggressive use of analog in-memory computing or very low-voltage logic, because simply scaling current designs wouldn't get a 50× improvement. It's plausible if we include architectural gains (e.g. moving from 14 nm to 3 nm gives ~4× energy gain; using near-threshold operation can give another factor; adding a small battery-friendly NPU that runs at 0.5 V might yield 10 TOPS/W for 4-bit ops). The **eventual ceiling** for something worn on the body is again thermal: ~1 W dissipated on the skin causes noticeable warming. There are regulatory limits (specific absorption rate, etc., mainly for RF, but thermally one doesn't want >1 °C rise on skin). So 0.5 W continuous is a comfortable max. If in the far future one could use reversible or extremely efficient logic such that 0.5 W corresponds to, say, 10^15 ops/s (which would be using ~5×10^−16 J/op), that's the dream scenario – essentially a supercomputer on your wrist. But barring a revolution, we expect wearables to always remain **3–4 orders of magnitude** lower in compute throughput than desktops/servers, constrained by power delivery and heat.

AR Glasses (2–3 W, near eye)

Augmented-reality glasses present a unique challenge: they have a slightly larger power budget than a watch (perhaps a few watts, since they can offload heat to the air a bit better than something on skin), but they are very sensitive to form factor and heat near the head. AR glasses need significant compute for tasks like environment mapping, object recognition, and graphics – essentially what a smartphone does, but in a tiny eyewear form. Current AR prototypes like Microsoft HoloLens 2 include custom ASICs ("holographic" co-processors) and operate around ~<5 W total. We assume ~2–3 W for compute is available in a lightweight AR headset without active cooling.

By 2035, AR glasses might be expected to perform on par with today's phones – meaning perhaps **~1 TOPS** of AI processing to analyze video from cameras, and decent graphics rendering for overlays. If technology advances, a 3 W AR device in 2035 could potentially host ~20–30 TOPS of neural network performance (for hand tracking, SLAM, etc.) and maybe a few hundred GFLOPS for graphics. This would likely require advanced 3D-stacked chiplets: for example, an eyebox SoC might have a logic die and a stacked memory die

for frame buffering and AI model storage, reducing power for data transfers by ~50%. Special thermal management (heat spreaders in the glasses frame) could allow slightly higher power bursts (maybe 5–6 W for a few seconds), but sustained power will be limited to avoid discomfort on the face.

One interesting design point: because AR devices are head-mounted, **weight and battery are critical**. They may offload heavy computation to a paired phone or cloud. The compute that remains on-board must be extremely energy-efficient (each joule matters for battery life and heat). Expect AR ASICs to use near-threshold operation extensively. For instance, an accelerator in AR glasses might run at 0.4 V, 100 MHz to do matrix math at very low power, using parallelism to compensate for frequency. This trades silicon area for energy efficiency – a fine trade in AR since area is available on the temple pieces to some degree, but battery is limited.

In summary, **best-case AR glasses by mid-2030s**: on the order of 10^13 ops/sec (tens of TOPS) within ~3 W, equating to ~10 TOPS/W efficiency. That's aggressive but feasible given likely improvements. The *physical ceiling* for AR would be a scenario where perhaps a high-performance reversible logic chip is integrated – then even 2 W could provide unbelievable compute – but practically, AR will always be constrained by human factors (heat near the eyes, battery weight).

One must also consider **diminishing returns of local compute**: beyond a certain point, it may be better to stream data to the cloud for heavy AI processing than to burn 10 W on your head. This is why even in 2040, an ultra-efficient AR headset might still offload big tasks to an edge server, using its on-board 1–2 TOPS only for real-time immediate tasks. The balance between on-device vs. cloud compute in mobile form factors will depend on connectivity and how far energy efficiency has been pushed in each domain.

To summarize the mobile landscape, Table 2 provides a comparison of current and projected compute metrics for different device classes:

Device	Typical Power	Current Performance (2025)	Projected Performance (2035)	Notes / Limits
Smartphone	~5 W (sustained)	~10–15 TOPS (INT8) or ~300 GFLOPS FP32 at 5 W	~50–100 TOPS (INT8) or ~1–2 TFLOPS FP32 at 5 W	Thermal throttling beyond 5–8 W; ~1 Exa-op/s *theoretical* limit at 5 W (Landauer) [13]
Smartwatch	~0.5 W	~0.1 TOPS (INT8) or few GFLOPS, <1 GHz CPU	~5–10 TOPS (INT8) or ~50+ GFLOPS at 0.5 W	Skin temperature constraint ~0.5–1 W; very energy-optimized designs
AR Glasses	~3 W (peak)	~1 TOPS (INT8) or ~100 GFLOPS (some offloading)	~20–30 TOPS (INT8) or ~0.5 TFLOPS at 2–3 W	~5 W upper limit near face (comfort); likely relies on cloud assist for heavy tasks

Table 2: Compute capacity in edge devices: approximate current performance vs. future projections around 2035, given power/thermal limits. (TOPS = 10^12 ops/sec, GFLOPS = 10^9 FLOPs/sec.) These estimates

assume advancements in efficiency but within practical cooling limits. Physical theoretical limits (at Landauer energies) are many orders of magnitude higher, but unreachable in practice for mobile devices.

Data Centers vs. Edge: Power, Cooling, and Scale Advantages

Despite impressive improvements in mobile compute, **data centers will continue to massively outperform edge devices** for the foreseeable future. The fundamental reason is simple: data centers can supply and dissipate **thousands to millions of times more power** for computing than an edge gadget. A single high-end GPU or AI accelerator in a server might consume 300 W – already 60× the power budget of a smartphone. An entire server rack can draw 20 kW, and a warehouse-scale data center can draw 100 MW or more. No amount of transistor optimization can overcome a *7-order-of-magnitude* disparity in power availability. Therefore, even if edge chips become as efficient as theoretically possible, the absolute compute capacity (operations per second) of a data center will dwarf that of a mobile or wearable device by factors of millions.

Power density and cooling are key differentiators. In a data center, chips are actively cooled with heatsinks, fans, liquid cooling, or even immersion cooling. High-performance CPUs/GPUs can run at junction temperatures of 85–100 °C, and heat fluxes of >100 W/cm^2 can be managed with advanced cooling solutions [49]. For example, modern server chips often dissipate ~200 W over ~600 mm^2 (which is ~33 W/cm^2) and use heavy copper heatsinks with forced air or water to keep temperatures in check. Some systems with 3D packaging or high TDP (like HBM-stacked GPUs) use cold plates and water loops to achieve >1 kW per 1U server. In contrast, a smartphone has maybe 5–10 cm^2 of area to spread heat, no active cooling (just a metal case or graphite sheet), and must keep surface temperature <~45 °C for the user's skin. That works out to a safe flux of only ~0.5–1 W/cm^2 at the surface – a **50×–100× lower heat density** handling than in servers. **Wearables are even more restrictive**, with maybe 0.1–0.2 W/cm^2 tolerable on skin. This fundamental thermal gap means edge devices physically cannot run the high-power chips that data centers can.

Economic scale further amplifies data center capabilities. If an application needs more compute, a cloud provider can simply add more servers or racks (assuming it's parallelizable). A big AI training run today might use 1000× 400 W GPUs = 400 kW of compute, something utterly impossible on-device. Even a future ultra-efficient chip in your phone would still be limited to a few watts – you cannot scale it up without burning your hand or exhausting the battery. Data centers also benefit from economy of scale in power delivery and cooling: a well-designed facility might have a Power Usage Effectiveness (PUE) of 1.1–1.2, meaning only ~10–20% overhead on top of computing for cooling/power conversion [50]. They turn electricity into compute with high overall efficiency, whereas a mobile device's entire budget is constrained and any overhead (screen, radios, etc.) cuts into the compute energy available.

To illustrate the disparity: the Frontier supercomputer achieves ~1.1 exaflops (1.1×10^18 FLOPs) on Linpack using ~21 MW of power [51]. That's roughly $5×10^{16}$ FLOPs per joule (0.05 EFLOPS per MW). A smartphone at 5 W might manage ~0.5 TFLOPS (5×10^11 FLOPs/sec) peak; over one second (5 J) that's 10^11 FLOPs per joule. So in energy efficiency it's actually not far off – maybe an order of magnitude or two lower than Frontier's 5×10^12 FLOPs/J on Linpack – but in total throughput it's smaller by *seven orders of magnitude* because the power is smaller by six and also it can't be run as efficiently at full utilization continuously due to throttling. In practice, data centers also employ specialization: if one needs 100x more performance, one can deploy 100x more chips. A mobile device cannot multiply its hardware on the fly.

Another reason data centers outperform is **memory and I/O bandwidth**. A 300 W GPU can have an 80 GB high-bandwidth memory (HBM) stack delivering 2 TB/s of memory bandwidth [52]. That memory alone dissipates tens of watts. No mobile device can include such memory – they rely on LPDDR or on-chip SRAM with at most ~100 GB/s bandwidth due to power limits. Thus, beyond raw ops, data center chips can feed their compute units with data at rates an edge device can't, enabling them to reach higher effective performance on large problems.

Moreover, **data centers can use exotic technologies** (like superconducting or photonic interconnects, or very large chiplet-based processors) that are impractical in a phone. For example, if superconducting accelerators become viable, they'll likely appear in server installations with cryogenic infrastructure, not in pocket devices. Data centers can accommodate the bulk and cost of new cooling or power delivery tech (e.g., 48 V racks, on-site liquid nitrogen, etc.) if it yields performance gains. Edge devices are severely limited in size, weight, and cost.

Finally, **energy economics** favor centralization for heavy compute. It is often more energy-efficient to do a task on a highly optimized data center cluster and send the result to a phone than to do it on the phone – especially if the phone would have to use its relatively inefficient (and battery-draining) processor for a long time. Cloud servers can be near ideal in utilization and can be upgraded continuously to the latest process nodes, whereas a mobile device is fixed at purchase and often underutilized. Thus, even if mobile and server chips had equal per-op efficiency, the server can just deploy many more of them and run them hot 24/7 with proper cooling, getting far more total work done.

In summary, edge computing will improve, and many tasks will be done on-device for latency or privacy reasons, but **data centers will remain vastly more powerful** due to sheer scalable power. A smartphone might boast 100 TOPS of AI in 2035, but a data center will have *hundreds of thousands* of such chips working in concert. The gap in total ops/sec will remain on the order of 10^6–10^9×. Thermal physics and form-factor constraints guarantee that a pocketable device cannot catch up to a warehouse full of equipment. Even if one imagines every transistor in a phone operating at Landauer limit (which is impossible), a data center could also upgrade to near-Landauer devices and still have a million times more of them running. In fact, the gap may even grow if data centers adopt power-dense technologies (like 3D stacking with aggressive cooling, or superconducting logic) faster than mobile can. We already see ~>15 kW per rack deployments with liquid cooling, and future data centers might pack 50 kW or more per rack with new cooling methods [53] – trying to further increase the compute per volume. Mobile cannot increase its power in volume similarly (our bodies won't allow it).

Therefore, from an architecture perspective, **edge and cloud will play complementary roles**: mobile devices will do more local AI and compute than today (thanks to improved efficiency up to their ~5 W limit), but they will offload heavy lifting to the cloud which can afford 5,000 W on a single training node if needed. Data centers benefit from almost linear scaling by adding more machines, whereas mobile devices hit a hard ceiling in a hurry. The result is that cloud will always provide the "heavy artillery" of computing. Even as a smartphone in 2035 might be as powerful as a 2020 server, the 2035 server will be something only a facility with massive power and cooling can hold.

In conclusion, the **practical ceilings of computation** are dictated by both physical limits (Landauer's principle, device scaling) and engineering limits (power delivery and removal). CMOS scaling is entering the final stretch with angstrom-scale transistors and 3D integration, yielding diminishing efficiency gains [1]. We're still a comfortable factor (~10^6) above Landauer's energy limit per operation in practice, but edging

closer each year [13]. New paradigms – optics, quantum, neuromorphic, probabilistic, reversible superconducting – each aim to leap to the next rung of efficiency or capability, some approaching the kT scale per operation in specialized ways [45] [37]. Mobile devices, constrained by human-centric power limits (~watts), will make use of these advances to pack astonishing compute into small form factors, but will always be overshadowed by data-center compute which exploits high power density and parallelism. The **ceiling for a single mobile device** in the next 10–15 years might be on the order of 10^13–10^14 ops/sec, whereas the **ceiling for a data center** with thousands of such devices could be 10^20–10^21 ops/sec – and if reversible computing or other beyond-CMOS tech takes off in the server space, that gap could widen further. Ultimately, to keep improving performance within physical limits, we must combine many strategies: new device physics for lower dissipation, 3D architectures, specialized accelerators, and system-level solutions that distribute workloads between edge and cloud optimally. The quest for exascale–zettascale computing will increasingly rely on *post*-CMOS innovations as we approach the hard limits of classical transistor technology. Each paradigm discussed provides one piece of the solution to continue climbing upward without melting our chips (or ourselves) in the process.

Sources: Fundamental limits and Landauer's principle [18] [13]; IRDS roadmap projections for 2 nm, 1 nm and beyond CMOS [7] [1]; Koomey's Law historical trends [15]; emerging device metrics (photonic MAC efficiency [23], analog CIM 885 TOPS/W [34], AQFP 1.4 zJ switching [45], etc.); and current vs future mobile/server performance data [51] [26].

---

[1] [7] [8] [9] [10] Transistors reach critical point at 3nm - FAST TURN CHIP - Fast Turn Chip Electronics Co., Ltd.
https://www.fastturnchip.com/transistors-reach-critical-point-at-3nm-fast-turn-chip/

[2] [3] TSMC Reportedly Preparing New Equipment for 1.4 nm Trial Run at "P2" Baoshan Plant | TechPowerUp
https://www.techpowerup.com/334931/tsmc-reportedly-preparing-new-equipment-for-1-4-nm-trial-run-at-p2-baoshan-plant

[4] [PDF] Limits of CMOS and Prospects for Adiabatic/Reversible CMOS
https://www.sandia.gov/app/uploads/sites/210/2023/11/Comet23-slides_SAND.pdf

[5] [6] IRDS 2022 Executive Summary
https://irds.ieee.org/images/files/pdf/2022/2022IRDS_ES.pdf

[11] [12] [42] Magnetic memory and logic could achieve ultimate energy efficiency - Berkeley News
https://news.berkeley.edu/2011/07/01/magnetic-memory-and-logic-could-achieve-ultimate-energy-efficiency/

[13] [36] [37] [38] [43] [44] Biological computers could use far less energy than current technology by working more slowly
https://techxplore.com/news/2024-12-biological-energy-current-technology-slowly.html

[14] [15] [16] [18] Koomey's law - Wikipedia
https://en.wikipedia.org/wiki/Koomey%27s_law

[17] Exclusive Inside Look at First US Exascale Supercomputer - HPCwire
https://www.hpcwire.com/2022/07/01/exclusive-inside-look-at-first-us-exascale-supercomputer/

[19] [20] Limits to the Energy Efficiency of CMOS Microprocessors | Epoch AI
https://epoch.ai/blog/limits-to-the-energy-efficiency-of-cmos-microprocessors

21 22 What are some actual drawbacks to optical computing? - Quora
https://www.quora.com/What-are-some-actual-drawbacks-to-optical-computing

23 Single-shot optical neural network | Science Advances
https://www.science.org/doi/10.1126/sciadv.adg7904

24 A scalable optical neural network architecture using coherent ...
https://dspace.mit.edu/handle/1721.1/132370.2

25 Lightmatter shows new type of computer chip that could reduce AI energy use | Reuters
https://www.reuters.com/science/lightmatter-shows-new-type-computer-chip-that-could-reduce-ai-energy-use-2025-04-09/

26 Computer Chips Today Are Way Too Hot and Lightmatter Knows Why
https://aimresearch.co/market-industry/computer-chips-today-are-way-too-hot-and-lightmatter-knows-why

27 29 30 experimental realization - How power-efficient are quantum computers? - Quantum Computing Stack Exchange
https://quantumcomputing.stackexchange.com/questions/1486/how-power-efficient-are-quantum-computers

28 An analysis of the evolution of quantum computing and its relation to ...
https://www.sciencedirect.com/science/article/pii/S0164121224002103

31 How Neuromorphic Chips Could Redefine Edge AI Devices
https://www.embedur.ai/how-neuromorphic-chips-could-redefine-edge-ai-devices/

32 Low-Power Computing with Neuromorphic Engineering
https://onlinelibrary.wiley.com/doi/full/10.1002/aisy.202000150

33 Digital Prototypes May Enable Analog Neuromorphic Chips
https://www.eetimes.com/podcasts/digital-prototypes-may-enable-analog-neuromorphic-chips/

34 First demonstration of in-memory computing crossbar using multi-level Cell FeFET | Nature Communications
https://www.nature.com/articles/s41467-023-42110-y?error=cookies_not_supported&code=80ff894c-0159-4939-af0e-6ef7e2f62604

35 A compute-in-memory chip based on resistive random-access memory
https://www.nature.com/articles/s41586-022-04992-8

39 Probabilistic computing with p-bits - AIP Publishing
https://pubs.aip.org/aip/apl/article/119/15/150503/40486/Probabilistic-computing-with-p-bits

40 Çamsarı & Theogarajan "The Potential of P-Computers"
https://www.ce.ucsb.edu/news/all/2022/camsari-theogarajan-potential-p-computers

41 Breaking Down Probabilistic Computing: Why This New Paradigm ...
https://www.1950.ai/post/breaking-down-probabilistic-computing-why-this-new-paradigm-could-outperform-classical-and-quantum

45 46 47 calit2.ucr.edu
https://calit2.ucr.edu/sites/default/files/2025-04/ayala-2024-fsdl-workshop-compressed.pdf

48 IDE Development, Logic Synthesis and Buffer/Splitter Insertion ...
http://ieeexplore.ieee.org/document/8839386/

49 A review of heat pipe technology for foldable electronic devices
https://www.sciencedirect.com/science/article/pii/S1359431121005299

50 Hot or Not? How Data Center Thermal Standards Impact Energy ...
https://www.coresite.com/blog/hot-or-not-how-data-center-thermal-standards-impact-energy-use-and-costs

51 World's fastest supercomputers are helping to sharpen climate ...
https://www.science.org/content/article/world-s-fastest-supercomputers-are-helping-sharpen-climate-forecasts-and-design-new

52 NVIDIA H100 GPU Specs and Price for ML Training and Inference
https://datacrunch.io/blog/nvidia-h100-gpu-specs-and-price

53 Data center power and cooling strategies for increasing rack power ...
https://www.deltapowersolutions.com/en/mcis/technical-article-data-center-power-and-cooling-strategies-for-increasing-rack-power-density.php