Table 1. The systems and infrastructure categories of Part VI: primary-category output, compound annual growth, and structural-change estimates of the monthly log-volume series.

Category	n 2018	n 2022	n 2025	CAGR base	CAGR post	Slope ch.	p	Placebo
Computer architecture	19,554	15,174	20,209	-0.061	0.111	0.016	0.000	0.06
Accelerators and AI hardware	818	1,421	1,906	0.148	0.092	0.018	0.000	0.16
Operating systems	1,556	1,083	1,363	-0.087	0.078	0.016	0.000	0.47
Distributed computing	11,519	9,379	12,550	-0.050	0.132	0.020	0.000	0.07
Cloud computing	9,396	7,103	10,974	-0.068	0.198	0.019	0.000	0.23
Edge computing	5,185	8,554	11,554	0.133	0.126	0.004	0.006	0.07
Computer networks	58,744	49,141	59,555	-0.044	0.071	0.009	0.000	0.20
Database systems	16,265	13,625	16,681	-0.043	0.073	0.014	0.000	0.09
Information retrieval	12,095	11,243	17,379	-0.018	0.176	0.023	0.000	0.09
Quantum computing	6,224	9,216	17,471	0.103	0.287	0.018	0.000	0.16
Sustainable computing	583	830	1,639	0.092	0.233	0.016	0.000	0.16

Counts are eligible works whose primary category is the row category (multi-label memberships are larger). CAGR base: 2018 to 2 (complete years). Slope ch.: change in the monthly trend of the log-volume series at 2022-12 in the interrupted time-series model c Placebo p: share of placebo event dates whose slope statistic is at least as extreme; BP breaks: Bai–Perron break dates selected b excluded from the growth columns.

1. Scope and evidence

This part reviews the eleven categories of the taxonomy that describe how computation is built, housed and delivered: computer architecture, accelerators and AI hardware, operating systems, distributed computing, cloud computing, edge computing, computer networks, database systems, information retrieval, quantum computing and sustainable computing. In the area partition of Appendix A the first six belong to *data and distributed systems* and the last five to *hardware and emerging paradigms*; the grouping adopted here follows the intellectual dependencies instead, so that accelerators are read with architecture, operating systems with distributed systems, and information retrieval with databases.

The evidence is of two kinds and the two are kept apart throughout. The first is the corpus itself: the yearly primary-category counts and growth rates of Part IX (Tables **??** and 1), the structural-change estimates of the monthly log-volume series (Table **??**), and the citation flows between categories of the knowledge graph. The second is the landmark lists of the two areas, 152 records for data and distributed systems and 125 for hardware and emerging paradigms, selected by the rule of Part III (the most cited works of each primary category under the field-normalised thirty-six-month window, with a floor of twenty-five records per category; for 2025 and 2026, twelve-month citations relative to the year mean). What is known about these works is their metadata and, for most of them, their abstract; no full text was consulted. Consequently every statement about the content of a paper below reports what its abstract says, and a paper whose

abstract is absent from the corpus is described from its title and venue only. Preprints are named as preprints. The structured record of each claim, with its evidence type and confidence, is in the file `literature_synthesis/agent_F/claims.csv`.

Two properties of the landmark lists bound what they can show. The selection rule favours records that OpenAlex has dated to a re-publication rather than to the original: the distributed-computing list opens with MapReduce (Dean and Ghemawat, 2008), Lamport's "Time, clocks" and "The Byzantine generals problem" recorded as 2018–2019 items, and the computer-architecture list is headed by the NumPy paper. Such records measure what the field cites, not what it produced, and they are set aside. The rule also admits, for 2025–2026, records with very few citations in absolute terms, so that the most recent "landmarks" are indications of current citation rather than established contributions. More consequential is classification noise. Roughly one record in five in the two lists is not systems research at all: dissertations on language teaching and admission systems under cloud computing and operating systems, an ethnography of mobile phones and the Paleobiology Database user guide under database systems, nursing students' AI literacy and regulatory sandboxes under operating systems, mammography screening under quantum computing, higher-education surveys under sustainable computing, control-theory papers under distributed computing, molecular-dynamics codes under computer architecture. These records are listed, with their identifiers, in `literature_synthesis/agent_F/review.md`; they are excluded from the synthesis, and their frequency is reported because it bounds the precision of every category-level statistic that follows. The taxonomy's validation in Appendix A did not sample these categories at the depth that would estimate the error rate, and this is recorded as an open problem for the data-quality discussion of Part XI.

2. The systems categories in the corpus

2.1. Volume, growth and share

Table 1 assembles, for the eleven categories, the primary-category output in 2018, 2022 and 2025, the compound annual growth rates over the baseline (2018–2022) and post (2023–2025) periods, and the structural-change statistics discussed below. The first regularity is a division into two groups. Seven classical categories declined in absolute output between 2018 and 2022 and grew between 2023 and 2025: computer architecture (-0.061 then 0.111 per year), operating systems (-0.087 then 0.078), distributed computing (-0.050 then 0.132), cloud computing (-0.068 then 0.198), computer networks (-0.044 then 0.071), database systems (-0.043 then 0.073) and information retrieval (-0.018 then 0.176). Four categories grew in both periods: edge computing (0.133 then 0.126), quantum computing (0.103 then 0.287), sustainable computing (0.092 then 0.233) and AI accelerators (0.148 then 0.092). The 2026 counts are excluded from the growth rates because the year is incomplete.

The second regularity concerns shares rather than levels. Computer networks, the largest of the eleven, held 0.118 of all primary output in 2018 and 0.067 in 2025; computer architecture fell from 0.039 to 0.023, database systems from 0.033 to 0.019, distributed computing from 0.023 to 0.014, cloud computing from 0.019 to 0.012 and operating systems from 0.003 to 0.002. Quantum computing rose from 0.012 to 0.020 and sustainable computing from 0.001 to 0.002. The relative decline of the classical systems fields is therefore a phenomenon of the whole period: it began before November 2022 and continued after it, while the absolute output of those fields recovered. Any reading of the post-2022 recovery as a response to a new workload must be weighed against the simpler possibility of reversion after a four-year decline whose causes this corpus does not identify.

2.2. Structural change

The interrupted time-series model of Part IV, fitted to the monthly log-volume series of each category with the event at December 2022, yields a positive change in trend for all eleven categories, with heteroskedasticity-and-autocorrelation-consistent p-values below one per cent in every case. The change is smallest for edge computing (0.004 per month, $p = 0.006$) and largest for information retrieval (0.023), with distributed computing (0.020), cloud computing (0.019), accelerators (0.018), quantum computing (0.018), architecture (0.016), operating systems (0.016), sustainable computing (0.016), database systems (0.014) and networks (0.009) in between. The uniformity is the finding. A steepening of the same order in quantum computing, which has no methodological connection with language models, and in information retrieval, which has a direct one, is not the signature of a workload-specific shock; it is the signature of a field-wide acceleration whose timing the placebo analysis then tests.

The placebo test rejects date-specificity for every systems category. The share of alternative event dates whose slope statistic is at least as extreme as the one estimated at December 2022 would have to be small for the date to be singled out, and it is never small: 0.062 for architecture, 0.077 for distributed computing, 0.077 for edge computing, 0.092 for databases, 0.092 for information retrieval, 0.169 for quantum computing, 0.169 for accelerators, 0.169 for sustainable computing, 0.200 for networks, 0.231 for cloud computing and 0.477 for operating systems. Level changes at the event month are heterogeneous in sign: negative and significant at the five-per-cent level for information retrieval (-0.094, $p = 0.019$), accelerators (-0.113, $p = 0.027$) and quantum computing (-0.087, $p = 0.039$); positive and significant for computer networks (0.043, $p = 0.021$) and sustainable computing (0.049, $p = 0.048$); indistinguishable from zero for the other six, with cloud computing the clearest null (0.003, $p = 0.936$). The placebo p-values for the level statistics exceed 0.477 in every category. The Bai–Perron break dates selected by the Bayesian information criterion, reported in Table 1, are scattered from the beginning of 2022 to the spring of 2025; cloud computing is the only category with a selected break in the month of the ChatGPT release, and the single break of the

networks series precedes it. The reading that the corpus supports is a broad acceleration of systems output from 2022–2023 onward that the November 2022 date does not single out, superimposed on the prior decline of the classical categories.

2.3. Citation flows

The category-to-category citation flows of the knowledge graph, aggregated over citing years 2018–2022 and 2023–2026,[1] show the systems categories to be strongly self-referential and stable. In the post years the share of references that a category's works send to works of the same category is 84.9% for quantum computing, 72.4% for computer networks, 59.2% for information retrieval, 57.7% for computer architecture, 40.6% for database systems, 34.7% for accelerators, 33.1% for cloud computing, 29.8% for distributed computing, 24.4% for edge computing, 16.0% for operating systems and 11.3% for sustainable computing. None of these shares moved by more than a few points from the baseline years (operating systems from 20.5%, quantum computing from 88.3%). Three details are relevant to the synthesis. The large-language-model category receives 3.0% of information-retrieval references in the post years against 0.7% before, the only systems category in which that destination becomes measurable. Operating systems now sends more references to computer architecture (18.8%) than to itself, which is consistent with the GPU-centred systems work reviewed in Section 4. Edge computing and sustainable computing cite computer networks (29.5% and 14.2%) more than their own category in both periods, a fact about where the classifier draws its boundaries as much as about intellectual dependence. The flows give no sign of a reorganisation of the systems fields around artificial intelligence between the two periods; what changed, as the following sections show, is the content of the most cited works within categories whose citation neighbourhoods stayed where they were.

3. Computer architecture and accelerators

3.1. The end of scaling and the arrival of one workload

The architecture literature of the baseline period is organised around a diagnosis that predates it: the end of Dennard scaling (Dennard et al., 1974) and the slowing of Moore's law left domain-specific architecture as the remaining route to performance, a position stated in the Turing lecture of Hennessy and Patterson (2019) and quantified by Leiserson et al. (2020), who locate the remaining gains in software, algorithms and hardware specialisation rather than in transistors. The tensor processing unit (Jouppi et al., 2017) and the GPU lineage (Dally et al., 2021) are the canonical

[1] Calendar citing years; the aggregation does not use the 30 November 2022 boundary. Source: `results/graph/category_flows_yearly.csv`, aggregated by the macros of `tables/make_numbers.py`.

instances; Sze et al. (2017) is the standard account of why neural-network inference rewards specialised data paths. The landmark lists confirm that this is the lineage the field cites, and they add the processing-in-memory primer of Mutlu et al. (2023) and the concurrency analysis of distributed deep learning by Ben-Nun and Hoefler (2020) as the most cited expository works of the pre-2022 years. What the baseline lists also contain, and what is set aside, are scientific-software papers (NumPy, NAMD, Quantum ESPRESSO) that the classifier assigns to architecture because their topics concern CPU and GPU execution.

After 2022 the most cited architecture works, once misclassified records are removed, form a single set: systems papers for the training and serving of large language models. This is the central empirical observation of this section and it is a statement about metadata, not about the field's full output: the citation-ranked top of the category is occupied by one workload. The corpus statistics of Section 2 qualify it at once. Architecture's trend change at December 2022 (0.016 per month) is in the middle of the systems range, its level change is null ($p = 0.771$), and its share of computer-science output fell from 0.039 to 0.023. The concentration of citation on LLM systems papers occurred within a category that did not grow faster than the rest and whose relative weight continued to decline.

3.2. Training and inference systems

The training-systems line is largely a pre-window inheritance recorded in preprints and conference papers: tensor and pipeline parallelism in Megatron-LM (Shoeybi et al., 2019, preprint) and its scaling to thousands of GPUs (Narayanan et al., 2021), memory partitioning of optimiser state in ZeRO (Rajbhandari et al., 2020), fully sharded data parallelism in PyTorch (Zhao et al., 2023, preprint), and the asynchronous dataflow of Pathways (Barham et al., 2022, preprint). The compiler stack beneath these systems is the one documented in Part VII: LLVM and MLIR (Lattner and Adve, 2004; Lattner et al., 2021), TVM (Chen et al., 2018, preprint), Triton (Tillet et al., 2019). The landmark set adds the paper that brought graph compilation into the dominant framework: Ansel et al. (2024) describe TorchDynamo, which captures computation graphs by rewriting Python bytecode before execution, and TorchInductor, which lowers them to Triton for GPUs and to C++ for CPUs; the abstract reports geometric-mean speedups of 2.27× for inference and 1.41× for training over eager execution on the authors' model suite. This is a vendor-authored evaluation of the authors' own system, which is the norm in this literature and the reason the claims below are reported as the authors' rather than as established.

The inference-systems line is where the window's own contributions lie, and it is organised by one bottleneck. Gholami et al. (2024), in a perspective article in *IEEE Micro*, state that peak server FLOPS grew at 3.0× per two years over the past twenty years while DRAM bandwidth grew at 1.6× and interconnect bandwidth at 1.4×, and argue that memory bandwidth, not arithmetic throughput,

has become the binding constraint for decoder-only Transformers at serving time. The trend figures are the authors' and are not re-derived here, but the diagnosis organises the rest of the literature. Pope et al. (2022), in a preprint, derive the latency–throughput frontier of Transformer inference under memory and communication limits; FlashAttention (Dao et al., 2022, preprint) reduces memory traffic of the attention kernel by tiling; quantisation methods reduce the weight bytes moved per token, from post-training GPTQ (Frantar et al., 2022, preprint) to activation-aware AWQ, whose abstract reports an algorithm–system co-design with an on-device framework (TinyChat) that converts memory savings into speedups through on-the-fly dequantisation and kernel fusion (Lin et al., 2025). The serving systems proper are discussed with operating systems in Section 4, because their abstraction is memory management; here it suffices to note that Splitwise (Patel et al., 2024), an ISCA paper, characterises each inference request as a compute-intensive prompt phase and a memory-intensive token-generation phase with distinct latency, throughput and power profiles, and places the two phases on separate machine pools so that generation runs on older or lower-power hardware. Taken together these works establish, at the level of abstracts, that the architecture community's response to the LLM workload in 2023–2025 was to redesign the memory system and the deployment topology rather than the arithmetic unit. Whether the memory-wall diagnosis survives changes in the installed base, and whether gains attributed to it come from bandwidth or from model-side sparsity, are open (Section 10).

The survey of Silvano et al. (2025) closes the period: it classifies accelerators for heterogeneous HPC platforms into GPU and TPU systems, FPGA and ASIC designs, neural processing units, open RISC-V accelerators, and emerging-memory and neuromorphic devices. That such a survey, and the earlier benchmarking of commercial accelerators by Reuther et al. (2022) and TPU v4 (Jouppi et al., 2023), do not appear in the accelerator landmark list but in the bibliography is a reminder that the classifier routes digital accelerators to architecture.

3.3. Photonic, neuromorphic and in-memory accelerators

The accelerator category is the one most affected by classification: at least half of its post-period landmarks are memristive and chaotic Hopfield-network circuit papers in journals of nonlinear dynamics and circuit design, a literature on attractors, multistability and image encryption that the topic classifier maps to AI hardware. Removing it leaves a photonic and neuromorphic thread whose baseline anchors are Loihi, Intel's neuromorphic manycore processor with programmable synaptic learning rules (Davies et al., 2018), and the perspective on neuromorphic algorithms and applications of Schuman et al. (2022). The post-period additions are optical: the Taichi photonic chiplet (Xu et al., 2024), whose abstract reports a diffractive–interference hybrid design with a distributed architecture reaching 160 TOPS/W and on-chip classification over a thousand categories; the universal photonic accelerator of Ahmed et al. (2025), for which the corpus holds no abstract;

a review of optical neural networks (Fu et al., 2024); and, in 2026, a monolithically integrated neuromorphic photonic circuit with on-chip capacitive analog memory, whose abstract reports an analysed power saving of more than 26× relative to SRAM–DAC designs (Lam et al., 2026), together with a hyper-multiplexed matrix–matrix optical processor.

These are device-level demonstrations evaluated on small tasks, and the efficiency figures they report are figures for the photonic core. None of the abstracts reports end-to-end energy including digital-to-analog and analog-to-digital conversion, memory traffic and precision loss, and none compares against a precision-matched digital accelerator under a shared benchmark. The literature is therefore classified here as speculative with respect to its claim on the computing infrastructure: the physics is established, the system-level advantage is not. The corpus statistics reinforce the caution. The accelerator category's level fell at December 2022 (-0.113, $p = 0.027$), its post-period growth (0.092 per year) is slower than its baseline growth (0.148), and its references go to deep-learning architectures (23.1%) and computer architecture (17.4%) in proportions that did not change between periods. Whatever occurred in AI hardware after 2022 is not visible as a surge in this category's volume or as a reorientation of its citations.

4. Operating systems and distributed systems

4.1. Operating systems: memory abstractions for accelerators

Operating systems is the smallest of the eleven categories (1,363 primary works in 2025) and the one that contracted fastest before the window (-0.087 per year), recovering at 0.078 per year in 2023–2025 while its share of computer science continued to fall. Its baseline landmarks are what one would expect of the late 2010s: customisable trusted execution environments on RISC-V in Keystone (Lee et al., 2020), sub-millisecond serverless runtimes for interactive microservices in Nightcore (Jia and Witchel, 2021), the Berkeley position paper on serverless computing (Jonas et al., 2019, preprint), and key-value stores for persistent memory. Unikernels (Madhavapeddy et al., 2013; Kuenzer et al., 2021) and kernel-bypass or in-kernel packet processing with XDP and eBPF (Høiland-Jørgensen et al., 2018; Vieira et al., 2021) belong to the same lineage and are cited by the corpus, though they do not reach the landmark threshold.

The post-period landmarks are of a different kind. PowerInfer (Song et al., 2024), an SOSP paper, observes that neuron activations in large models follow a power-law distribution, keeps the consistently active "hot" neurons on a consumer GPU and computes the input-dependent "cold" ones on the CPU, with predictors and sparse operators to bridge the two. Orion (Strati et al., 2024) intercepts the kernel launches of several clients sharing a GPU and schedules them at the granularity of individual operators so as to limit interference at the scale of tens of microseconds. vAttention (Prabhu et al., 2025) contests the memory abstraction of vLLM (Kwon et al., 2023):

where PagedAttention allocates the key–value cache on demand in non-contiguous pages, inspired by virtual memory and paging, vAttention argues that the loss of contiguity imposes programming and performance overheads and restores a contiguous virtual layout through CUDA virtual-memory primitives. This is a live design controversy, and the corpus offers no way to adjudicate it: each system is evaluated by its authors on particular GPU generations and traces, and no neutral, versioned serving benchmark exists. What the abstracts establish is narrower and more interesting: the operating-systems problem of 2023–2025, as the field's most cited work defines it, is the management of accelerator memory and accelerator time for one workload, and the classical concerns of the baseline years—isolation, serverless latency, persistent memory—no longer appear at the top of the category. The citation flows agree: operating-systems works now send more references to computer architecture (18.8%) than to their own category (16.0%).

Two further items qualify the picture. The "LLM as operating system" proposals that the bibliography records (Packer et al., 2023; Ge et al., 2023; Mei et al., 2024, all preprints) treat context management and tool scheduling for language-model agents as an operating-systems problem by analogy; they are position statements, not systems in the sense of the papers above, and none reaches the landmark list. The standardisation of the BPF instruction set as an IETF RFC (Internet Engineering Task Force, 2024) in 2024, together with an in-kernel aggregation preprint whose record is misdated in the corpus, suggests that kernel-extension technology remained an infrastructure theme of the period; the evidence is one RFC and one abstract, and the suggestion is recorded as such.

4.2. Distributed computing: a canon cited, not renewed

The distributed-computing landmark list is dominated by what the field cites rather than by what it produced: re-dated records of MapReduce (Dean and Ghemawat, 2008), Lamport's logical clocks and the Byzantine generals problem head the baseline, and the canon that the bibliography documents—Paxos (Lamport, 1998), practical Byzantine fault tolerance (Castro and Liskov, 2002), the CAP formulation (Gilbert and Lynch, 2002), Spanner (Corbett et al., 2013), Spark (Zaharia et al., 2016), Borg and Kubernetes (Verma et al., 2015; Burns et al., 2016)—is present in the references of the period without any post-2022 work of comparable standing in consensus, transactions or fault tolerance reaching the list. The growth topic of the baseline years was blockchain: a blockchain survey appears twice under distinct OpenAlex identifiers, the selfish-mining paper "Majority is not enough" and a consensus pre-processing method are among the most cited, and in the post years distributed-computing works still send 4.3% of their references to the blockchain category. The category's trend change at December 2022 is among the largest of the eleven (0.020 per month) and its output recovered at 0.132 per year after declining at -0.050; two Bai–Perron breaks are selected, neither in the event month (Table 1).

The post-period landmarks that survive scrutiny are three. vLLM (Kwon et al., 2023), classified here rather than under operating systems, reports that near-zero waste of key–value-cache memory and the sharing of cache blocks within and across requests raise throughput by two to four times at equal latency relative to the previous state of the art. OrbitalBrain proposes in-space distributed training on Earth-observation nanosatellites under bandwidth and power constraints, and a live race-bypassing system for deployed multithreaded applications appears as a 2026 repository record of older work. The first is the window's single most influential systems artefact in this area; the other two are indications of citation rather than landmarks. Federated learning, which the taxonomy treats as a category of its own but which is a distributed-systems topic in its mechanics (McMahan et al., 2016; Kairouz and McMahan, 2021; Li et al., 2020), appears in the systems lists only through surveys, discussed in Section 9. The conclusion for distributed computing is therefore negative in a precise sense: the corpus records no renewal of the field's theoretical or protocol canon in 2023–2026, and records its most cited new work as an inference-serving system whose intellectual home is the operating-systems tradition of memory management.

5. Cloud and edge computing

5.1. Cloud computing: the same problem, a new workload

Cloud computing's baseline landmarks are scheduling and placement papers—virtual-machine allocation with usage prediction, QoS-aware scheduling for energy conservation, metaheuristic task scheduling—together with surveys of data-centre energy efficiency and of IoT–cloud integration. The post-period landmarks keep the problem and change the workload. The review of deep-reinforcement-learning methods for cloud resource scheduling by Zhou et al. (2024) is the bridge; the 2026 slice consists almost entirely of scheduling for large-language-model inference: token-level pooling and slicing of GPUs for co-located models, contextual-bandit elastic scheduling for serverless multi-tenant inference (Li et al., 2026), and carbon-aware prompt scheduling, whose abstract reports, in trace-driven simulation, a reduction of 26.8% in carbon per thousand generated tokens relative to round-robin routing (Liao et al., 2026). These 2026 records are provisional landmarks in the sense of Section 1, they appear in smaller IEEE conferences, and each is evaluated in simulation or on a single testbed; they are cited here not as results but as evidence of what the cloud-computing literature of 2026 is about. The one post-period landmark with an operations rather than a scheduling focus is RCACopilot (Chen et al., 2024), a EuroSys paper that matches cloud incidents to handlers, aggregates runtime diagnostics and uses a language model to predict the root-cause category and produce an explanatory narrative; the abstract reports accuracy up to 0.766 on a year of incidents from one provider, a proprietary evaluation that no public benchmark allows to be reproduced.

The corpus statistics for cloud computing are distinctive in one respect. Its level change at December 2022 is the clearest null of the eleven (0.003, $p = 0.936$), its trend change is 0.019 per month, and it is the only systems category for which the Bayesian information criterion selects a Bai–Perron break in the month of the ChatGPT release, with a second break two years later. The placebo result (0.231) does not permit the coincidence to be read as specific, and the category's output had been contracting at -0.068 per year before recovering at 0.198. The serverless agenda of the baseline (Jonas et al., 2019) survives in the 2026 papers as the deployment model for inference; what the landmark set does not contain is any post-2022 work on the cloud's own abstractions—storage, consistency, multi-tenancy—of the standing of the Borg or Spanner papers.

5.2. Edge computing: a change of subject without a change of output

Edge computing offers the sharpest contrast between what is cited and what is produced. Its baseline landmarks are the blockchain–IoT integration papers of 2018–2019, the *Edge Intelligence* survey of Zhou et al. (2019), and a cluster of COVID-era reviews of IoT applications that the classifier assigns to the category; the lineage runs from the definitions of Shi et al. (2016) and Satyanarayanan (2017) through the edge-intelligence surveys (Deng et al., 2020) to the 6G framing of edge AI (Letaief et al., 2022). The post-period landmarks are surveys and visions about language models at the edge: mobile edge intelligence for LLMs, which positions the mobile network edge between on-device and cloud inference on grounds of privacy, latency and bandwidth (Qu et al., 2025); edge general intelligence with agentic AI, which proposes that distributed edge entities become autonomous agents with perception–reasoning–action loops (Zhang et al., 2026); edge–cloud collaborative computing on distributed intelligence and model optimisation (Liu et al., 2026); and AI Flow, a device–edge–cloud framework for large models (An et al., 2026). The abstracts describe architectures, taxonomies and challenges; none reports a measured deployment, and the empirical question of where inference should sit on the device–edge–cloud continuum under energy, latency and privacy constraints is open.

Against this change of subject, the category's time series barely moved. Edge computing has the smallest trend change at December 2022 of the eleven (0.004 per month, $p = 0.006$), after a baseline growth of 0.133 per year that had already raised its share from 0.010 to 0.015 by 2022; its post-period growth (0.126) is lower than its baseline growth, its single selected break falls in 2025, and its references go to computer networks (29.5%) more than to itself (24.4%) in both periods. The LLM-at-the-edge literature is, on this evidence, a change in what the field writes surveys about, not a measurable change in its volume or its citation neighbourhood.

6. Computer networks

Computer networks is the largest systems category and the one whose relative weight fell most, from 0.118 of primary output in 2018 to 0.067 in 2025, with absolute output contracting at -0.044 per year before 2022 and growing at 0.071 after. It is also the category whose structural-change statistics are least compatible with a post-2022 event: its level change at December 2022 is positive (0.043, $p = 0.021$) but with a placebo p-value of 0.538, its trend change (0.009) is the smallest after edge computing, and its single selected Bai–Perron break precedes the event month (Table 1). The series was rising before ChatGPT.

The most cited works of the period are overviews that set the agenda of the sixth generation of mobile systems: the 6G roadmap of Wang et al. (2023), which surveys requirements, scenarios and testbeds while noting that the ITU-R vision was still under discussion, and the integrated sensing and communications overview of Liu et al. (2022), which treats the perceptive radio access network as a defining feature of 6G; both coincide with or precede the start of the window and continue the programme of Dang et al. (2020). Machine learning had entered networking research well before the window, as the surveys of deep learning for IoT analytics and for mobile and wireless networking show (Zhang et al., 2019), and the programmable-network lineage of OpenFlow, P4 and in-network computation (McKeown et al., 2008; Bosshart et al., 2014; Kreutz et al., 2015; Hauser et al., 2023; Sapio et al., 2019) is cited throughout. What the 2026 landmarks add is a vocabulary: tutorials on large AI models and agentic AI for intelligent communications (Jiang et al., 2026), a review of energy, scalability, data and security in massive IoT (Cheikh et al., 2026), the edge-general-intelligence survey already cited, and a tutorial on pinching-antenna systems that belongs to the physical layer. The abstracts of the agentic-AI tutorials describe design principles and taxonomies; their empirical content cannot be assessed from abstracts, and they are provisional landmarks by the criterion of Section 1. The networks category also receives the references of edge computing and sustainable computing in larger proportion than those categories' own works, so that part of what the taxonomy records as networking output is infrastructure work that other communities cite. The established conclusion is modest: the field's agenda in 2023–2026 was 6G, and the arrival of large models is visible in its surveys rather than in its output or its citation structure.

7. Database systems and information retrieval

7.1. Database systems: data for learning, and vector search

The database landmark list is the most contaminated by misclassification of the eleven, and what remains after exclusion is small but instructive. The baseline anchors are BLOCKBENCH, a

data-processing view of private blockchains with a benchmarking framework (Dinh et al., 2018), the documentation standard *Datasheets for datasets* (Gebru et al., 2021), and the GPU similarity-search paper that underlies the Faiss library (Johnson et al., 2021). The post-period landmarks are three: *The Faiss Library* (Douze et al., 2026), whose abstract frames vector similarity search as the core functionality of vector databases and describes the library's indexing, compression and benchmarking principles; the data-centric AI survey of Zha et al. (2025), organised around training-data development, inference-data development and data maintenance; and an empirical study of the effect of six data-quality dimensions on nineteen machine-learning algorithms under polluted training and test data (Mohammed et al., 2025). Database systems' trend change at December 2022 (0.014 per month) and its single Bai–Perron break in late 2023 are unremarkable among the systems categories; its output contracted at -0.043 per year before 2022 and grew at 0.073 after, and its references go to machine-learning theory, natural language processing and computer vision in proportions that the two periods share.

Two readings are consistent with this record and the corpus cannot separate them. The first is that the database community's most cited work of the period concerns the data that feed machine learning—its quality, documentation and maintenance—and the retrieval of embeddings, rather than transaction processing or query optimisation; this is what the metadata show. The second is that the engine-level work of the period is simply not cited enough within thirty-six months to reach a list ranked by short-window citations, which favours surveys. The bibliography records the community's own positions on both points: the vector-database surveys (Pan et al., 2024; Taipalus, 2024) and the preprint of Ma et al. (2023), the hierarchical navigable small-world index that most vector systems implement (Malkov and Yashunin, 2020), the programme of machine learning for database internals (Li et al., 2021), the assessment of large language models for data management by Fernandez et al. (2023), and the argument of Stonebraker and Pavlo (2024) that specialised vector stores will be absorbed by extended relational engines as earlier specialised systems were. Whether vector search belongs in dedicated systems or in extended relational engines is the contested question of the period, and it is untested at workload level: the landmark set contains library-level evidence and surveys, not comparisons of dedicated and extended systems under mixed structured and approximate-nearest-neighbour queries.

7.2. Information retrieval: from graph recommendation to retrieval augmentation

Information retrieval shows the clearest change in what is cited of any systems category. The baseline landmarks are graph-neural-network recommendation and its surveys: neural graph collaborative filtering, knowledge-graph attention networks, graph neural networks for social recommendation, self-supervised graph learning, and LightGCN (He et al., 2020), whose abstract reports that the feature transformation and non-linear activation inherited from graph classification contribute

little to collaborative filtering and can degrade it; the surveys of Wu et al. (2023) and Gao et al. (2023a) close the baseline. The post-period landmarks are retrieval-augmented generation and language models for recommendation: the survey of retrieval-augmented large language models of Fan et al. (2024), the *ACM Computing Surveys* review of Huang and Huang (2026), which organises the paradigm into pre-retrieval, retrieval, post-retrieval and generation stages, the survey of retrieval augmentation for generated content across modalities (Zhao et al., 2026), the surveys of recommender systems in the era of large language models (Zhao et al., 2024b; Wu et al., 2024), and C-Pack (Xiao et al., 2024), which releases a training corpus, a Chinese embedding benchmark and the BGE family of embedding models whose abstract reports state-of-the-art results on the English MTEB benchmark as well.

The methodological lineage of this literature is continuous with 2020: dense passage retrieval (Karpukhin et al., 2020), retrieval-augmented generation itself (Lewis et al., 2020, preprint), late interaction in ColBERT (Khattab and Zaharia, 2020), and the probabilistic framework that these replaced (Robertson and Zaragoza, 2009); the dense-retrieval survey of Zhao et al. (2024a) and the preprint survey of language models for information retrieval (Zhu et al., 2023) document the transition, as does the preprint RAG survey of Gao et al. (2023b). What is notable is that the landmark set contains no new retrieval architecture of 2023–2026 and consists, for the post period, of surveys and one resource. This may reflect the citation lag of recent model papers as much as the field's output, and the review records it as speculative. The corpus statistics are the strongest of the systems group: information retrieval has the largest trend change at December 2022 (0.023 per month), a negative level change (-0.094, $p = 0.019$), two selected Bai–Perron breaks, output growth of 0.176 per year after a flat baseline (-0.018), and it is the only systems category whose references to the large-language-model category become measurable, at 3.0% of post-period references against 0.7% before. Its self-share of references nevertheless remained at 59.2%, and natural language processing remained its second destination (10.6%). The placebo p-value for the trend (0.092) forbids attributing the acceleration to the event date, and the open problem that the literature itself states is evaluative: no agreed protocol separates the contribution of retrieval quality from that of parametric memory and of generation in retrieval-augmented systems.

8. Quantum computing

Quantum computing is the systems category least connected to the others: 84.9% of its post-period references go to works of its own category, applied cryptography being the next destination at 2.2%. It is also the fastest-growing after 2022 (0.287 per year, after 0.103 in the baseline), the only systems category whose share of computer-science output rose throughout (0.012 to 0.020), and one whose level fell at December 2022 (-0.087, $p = 0.039$). Its two selected Bai–Perron breaks

fall at the beginning of 2022 and in the spring of 2025 (Table 1). These are the statistics of a field growing on its own momentum, and the landmark list explains why: the window contains a coherent experimental sequence that has no connection with language models.

The baseline anchors are Preskill's framing of the noisy intermediate-scale era (Preskill, 2018), the supremacy experiment on a 53-qubit programmable superconducting processor (Arute et al., 2019), the photonic advantage experiment, and the reviews of NISQ algorithms and variational methods (Bharti et al., 2022; Cerezo et al., 2021). The post-period landmarks are the error-correction milestones. Google Quantum AI (2023) report that a distance-5 surface-code logical qubit modestly outperforms an ensemble of distance-3 qubits, the first demonstration that adding physical qubits improves logical performance on that platform; Bluvstein et al. (2024) report a programmable logical processor on reconfigurable neutral-atom arrays with up to 280 physical qubits, a zoned architecture and mid-circuit readout; and Google Quantum AI and Collaborators et al. (2025) report two below-threshold surface-code memories on the Willow processor, a distance-7 code whose logical error rate is suppressed by a factor of $\Lambda = 2.14$ per increase of two in code distance, with a real-time decoder, and a logical memory beyond break-even. The figures are those of the abstracts and the articles are peer-reviewed. Two 2025 demonstrations extend the sequence to modular architectures: distributed quantum computation between two trapped-ion modules about two metres apart by deterministic gate teleportation over an optical link (Main et al., 2025), and Aurora, a 35-chip photonic machine networked over fibre that its authors describe as a sub-performant scale model demonstrating the integration of all primitive components (Aghaee Rad et al., 2025). These are feasibility demonstrations; the rates and fidelities required for error correction across modules are not reported in the abstracts.

The contested claim of the period is utility before fault tolerance. Kim et al. (2023) report, on a 127-qubit processor, accurate expectation values for circuit volumes beyond brute-force classical computation and argue that this is evidence of utility in the pre-fault-tolerant era. The abstract restricts the claim to brute-force simulation. The classical-simulation responses that followed, which reproduced the reported quantities with tensor-network and related methods, are not in the landmark set and are not in the verified bibliography, and they are therefore not cited; the claim is recorded as contested, and the open problem is methodological—utility experiments need independently committed classical baselines fixed before publication, because the classical frontier moves after it. On the algorithmic side, the review of barren plateaus in variational quantum computing (Larocca et al., 2025), for which the corpus holds no abstract, and quantum-selected configuration interaction (Kanno et al., 2026), which samples electron configurations on a noisy device and diagonalises the Hamiltonian classically in the selected subspace, represent the two responses to the trainability problem of variational algorithms: characterise the obstacle, or move the optimisation to the classical side. The software and vendor records—the Qiskit design paper (Javadi-Abhari et al., 2024, preprint)

and a retrospective of IBM's processors and industry applications (AbuGhanem, 2025)—complete a list in which, as the bibliography anticipates for Part X, the post-quantum standards of 2024 (National Institute of Standards and Technology , US) and the lattice foundations (Regev, 2009; Shor, 1997) are the only links to the rest of computer science.

9. Sustainable computing

Sustainable computing is small (1,639 primary works in 2025), grew at 0.233 per year after 2022 against 0.092 before, has a positive level change at December 2022 (0.049, $p = 0.048$; placebo $p = 0.754$) and cites computer networks (14.2%) and cloud computing (10.8%) more than or as much as itself (11.3%). Its baseline landmarks concern ICT for the sustainable development goals and data-centre traffic models; the lineage that the bibliography documents is the measurement of machine learning's footprint—the energy of training natural-language models (Strubell et al., 2019), the green-AI argument (Schwartz et al., 2020), the carbon-tracking tools of Henderson et al. (2020), the global data-centre energy accounting of Masanet et al. (2020)—and the post-period landmarks continue it with a notable share of preprints and workshop papers.

The measurement literature proceeds in three steps. Training: Patterson et al. (2021), in a preprint, estimate the energy and CO_2-equivalent of T5, Meena, GShard, Switch Transformer and GPT-3 and report that sparsely activated models, data-centre location and hardware choice change emissions by large factors; Luccioni et al. (2022), in a preprint, estimate BLOOM's training at 24.7 tonnes of CO_2-equivalent for dynamic consumption and 50.5 tonnes over the life cycle, and extend the accounting to inference through an API; Dodge et al. (2022) argue that cloud providers should expose software carbon intensity to users. Inference and water: Chien et al. (2023) model generative-AI inference as a workload and compare request-direction policies for carbon; Li et al. (2023), first as a preprint and then in *Communications of the ACM* (Li et al., 2025), bring the water footprint of training and serving into the accounting, with projections of global withdrawal that are the authors' estimates; the reviews of green AI (Bolón-Canedo et al., 2024) and green federated learning (Thakur et al., 2025) and the related argument of Wu et al. (2021), a preprint, organise the mitigation side. Systems and policy: Luccioni et al. (2024) measure the energy of inference by task at the level of deployed models; Luccioni et al. (2025) argue that efficiency gains may raise total consumption through rebound effects and that life-cycle assessment must be joined to socio-economic analysis; de Vries-Gao (2026), continuing de Vries (2023), document that operators' environmental disclosures do not separate AI from other workloads, so that AI's carbon and water footprints can only be approximated from aggregate data-centre metrics; Chen et al. (2026) review the electricity demand of AI data centres and its interaction with power-system planning, and Kumar and Pindoriya (2026) review the infrastructures and energy-management

options of sustainable data-centre operation; and TokenPowerBench (Niu et al., 2026) provides a benchmark that attributes inference energy to the prefill and decode phases of each request. The latter's abstract quotes industry reports to the effect that inference accounts for more than ninety per cent of total power; such figures recur across this literature and cannot be verified from the corpus. The field's own diagnosis is the open problem: fleet-scale footprints are modelled rather than measured because the data are not disclosed, and the rebound hypothesis has not been tested.

10. Synthesis

Three statements are established by the corpus. The systems categories accelerated after 2022 as a group, and the acceleration is not specific to the ChatGPT date: every placebo test fails to single it out, and the break dates selected by the Bayesian information criterion are scattered across three years. The classical systems categories had been declining in absolute and relative terms since 2018 and recovered in absolute terms only; their share of computer science continued to fall, and reversion is a competing explanation of the recovery that this corpus cannot exclude. The citation flows among systems categories were stable between the two periods; the only measurable new destination is the references of information retrieval to the large-language-model category.

Two statements are established by the metadata of the most cited works, and they must be read with the caveat that short citation windows favour surveys and that roughly one landmark record in five is misclassified. The most cited post-2022 works in computer architecture, operating systems, distributed computing and cloud computing are systems for training and serving large language models—graph compilation, phase splitting, key–value-cache management, GPU sharing, consumer-GPU inference, inference scheduling—and the most cited works in information retrieval, edge computing and networking are surveys and tutorials about language models. Quantum computing and sustainable computing follow their own trajectories, the former an experimental programme in error correction and modular architectures, the latter a measurement literature on the footprint of machine learning in which several foundational items are preprints.

What cannot be established from this evidence is also to be stated. Whether research on language-model infrastructure explains the volume recovery of the classical categories is unknown, because the share of that research in each category's output is not measured here; a topic-level decomposition is the test. Whether any of the systems designs reviewed—paged or contiguous caches, phase-split or co-located serving, hot–cold partitioning—is superior to its competitors is unknown, because each is evaluated by its own authors on its own hardware. Whether photonic and neuromorphic efficiency figures hold at system level is unknown, because no abstract reports end-to-end energy under a shared benchmark. Whether pre-fault-tolerant quantum processors deliver utility is contested and moves with the classical frontier. These, together with the evaluation

of retrieval augmentation, the placement of inference on the device–edge–cloud continuum, the disclosure of AI's energy and water use, and the validation of the taxonomy's systems categories against venue streams, are carried forward to Part X.

References

AbuGhanem, Muhammad, "IBM quantum computers: evolution, performance, and future directions," *The Journal of Supercomputing*, 2025, *81* (5).

Aghaee Rad, H., T. Ainsworth, R. N. Alexander, B. Altieri, M. F. Askarani, R. Baby, L. Banchi, B. Q. Baragiola, J. E. Bourassa, R. S. Chadwick, I. Charania, H. Chen, M. J. Collins, P. Contu et al., "Scaling and networking a modular photonic quantum computer," *Nature*, 2025, *638* (8052), 912–919.

Ahmed, Sufi R., Reza Baghdadi, Mikhail Bernadskiy, Nate Bowman, Ryan Braid, Jim Carr, Chen Chen, Pietro Ciccarella, Matthew Cole, John Cooke, Kishor Desai, Carlos Dorta, Jonathan Elmhurst, Bryce Gardiner et al., "Universal photonic artificial intelligence acceleration," *Nature*, 2025, *640* (8058), 368–374.

An, Hongjun, Wenhan Hu, Sida Huang, Siqi Huang, Ruanjun Li, Yuanzhi Liang, Jiawei Shao, Yiliang Song, Zihan Wang, Cheng Yuan, Chi Zhang, Hongyuan Zhang, Wenhao Zhuang, and Xuelong Li, "AI Flow: perspectives, scenarios, and approaches," *Vicinagearth*, 2026, *3* (1).

Ansel, Jason, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala, "PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation," in "Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2" ACM 2024, pp. 929–947.

Arute, Frank, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C. Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando G. S. L. Brandao, David A. Buell, Brian Burkett, Yu Chen, Zijun Chen, Ben Chiaro, Roberto Collins, William Courtney, Andrew

Dunsworth, Edward Farhi, Brooks Foxen, Austin Fowler, Craig Gidney, Marissa Giustina, Rob Graff, Keith Guerin, Steve Habegger, Matthew P. Harrigan, Michael J. Hartmann, Alan Ho, Markus Hoffmann, Trent Huang, Travis S. Humble, Sergei V. Isakov, Evan Jeffrey, Zhang Jiang, Dvir Kafri, Kostyantyn Kechedzhi, Julian Kelly, Paul V. Klimov, Sergey Knysh, Alexander Korotkov, Fedor Kostritsa, David Landhuis, Mike Lindmark, Erik Lucero, Dmitry Lyakh, Salvatore Mandrà, Jarrod R. McClean, Matthew McEwen, Anthony Megrant, Xiao Mi, Kristel Michielsen, Masoud Mohseni, Josh Mutus, Ofer Naaman, Matthew Neeley, Charles Neill, Murphy Yuezhen Niu, Eric Ostby, Andre Petukhov, John C. Platt, Chris Quintana, Eleanor G. Rieffel, Pedram Roushan, Nicholas C. Rubin, Daniel Sank, Kevin J. Satzinger, Vadim Smelyanskiy, Kevin J. Sung, Matthew D. Trevithick, Amit Vainsencher, Benjamin Villalonga, Theodore White, Z. Jamie Yao, Ping Yeh, Adam Zalcman, Hartmut Neven, and John M. Martinis, "Quantum supremacy using a programmable superconducting processor," *Nature*, 2019, *574* (7779), 505–510.

Barham, Paul, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Dan Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, Brennan Saeta, Parker Schuh, Ryan Sepassi, Laurent El Shafey, Chandramohan A. Thekkath, and Yonghui Wu, "Pathways: Asynchronous Distributed Dataflow for ML," arXiv preprint arXiv:2203.12533 2022.

Ben-Nun, Tal and Torsten Hoefler, "Demystifying Parallel and Distributed Deep Learning: An In-depth Concurrency Analysis," *ACM Computing Surveys*, 2020, *52* (4), 1–43.

Bharti, Kishor, Alba Cervera-Lierta, Thi Ha Kyaw, Tobias Haug, Sumner Alperin-Lea, Abhinav Anand, Matthias Degroote, Hermanni Heimonen, Jakob S. Kottmann, Tim Menke, Wai-Keong Mok, Sukin Sim, Leong-Chuan Kwek, and Alán Aspuru-Guzik, "Noisy intermediate-scale quantum algorithms," *Reviews of Modern Physics*, 2022, *94* (1).

Bluvstein, Dolev, Simon J. Evered, Alexandra A. Geim, Sophie H. Li, Hengyun Zhou, Tom Manovitz, Sepehr Ebadi, Madelyn Cain, Marcin Kalinowski, Dominik Hangleiter, J. Pablo Bonilla Ataides, Nishad Maskara, Iris Cong, Xun Gao et al., "Logical quantum processor based on reconfigurable atom arrays," *Nature*, 2024, *626* (7997), 58–65.

Bolón-Canedo, Verónica, Laura Morán-Fernández, Brais Cancela, and Amparo Alonso-Betanzos, "A review of green artificial intelligence: Towards a more sustainable future," *Neurocomputing*, 2024, *599*, 128096.

Bosshart, Pat, Dan Daly, Glen Gibb, Martin Izzard, Nick McKeown, Jennifer Rexford, Cole Schlesinger, Dan Talayco, Amin Vahdat, George Varghese, and David Walker, "P4: pro-

gramming protocol-independent packet processors," *ACM SIGCOMM Computer Communication Review*, 2014, *44* (3), 87–95.

Burns, Brendan, Brian Grant, David Oppenheimer, Eric Brewer, and John Wilkes, "Borg, Omega, and Kubernetes," *Communications of the ACM*, 2016, *59* (5), 50–57.

Castro, Miguel and Barbara Liskov, "Practical byzantine fault tolerance and proactive recovery," *ACM Transactions on Computer Systems*, 2002, *20* (4), 398–461.

Cerezo, M., Andrew Arrasmith, Ryan Babbush, Simon C. Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R. McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, and Patrick J. Coles, "Variational quantum algorithms," *Nature Reviews Physics*, 2021, *3* (9), 625–644.

Cheikh, Imane, Sébastien Roy, Essaid Sabir, and Rachid Aouami, "Energy, Scalability, Data, and Security in Massive IoT: Current Landscape and Future Directions," *IEEE Internet of Things Journal*, 2026, *13* (6), 10127–10163.

Chen, Tianqi, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Meghan Cowan, Haichen Shen, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy, "TVM: An Automated End-to-End Optimizing Compiler for Deep Learning," arXiv preprint arXiv:1802.04799 2018.

Chen, Xin, Xiaoyang Wang, Ana Colacelli, Matt Lee, and Le Xie, "Electricity demand and grid impacts of AI data centers: Challenges and prospects," *Nexus*, 2026, *3* (3), 100162.

Chen, Yinfang, Huaibing Xie, Minghua Ma, Yu Kang, Xin Gao, Liu Shi, Yunjie Cao, Xuedong Gao, Hao Fan, Ming Wen, Jun Zeng, Supriyo Ghosh, Xuchao Zhang, Chaoyun Zhang, Qingwei Lin, Saravan Rajmohan, Dongmei Zhang, and Tianyin Xu, "Automatic Root Cause Analysis via Large Language Models for Cloud Incidents," in "Proceedings of the Nineteenth European Conference on Computer Systems (EuroSys '24)" ACM 2024, pp. 674–688.

Chien, Andrew A., Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana, "Reducing the Carbon Impact of Generative AI Inference (today and in 2035)," in "Proceedings of the 2nd Workshop on Sustainable Computer Systems (HotCarbon '23)" ACM 2023, pp. 1–7.

Corbett, James C., Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito,

Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford, "Spanner: Google's Globally Distributed Database," *ACM Transactions on Computer Systems*, 2013, *31* (3), 1–22.

Dally, William J., Stephen W. Keckler, and David B. Kirk, "Evolution of the Graphics Processing Unit (GPU)," *IEEE Micro*, 2021, *41* (6), 42–51.

Dang, Shuping, Osama Amin, Basem Shihada, and Mohamed-Slim Alouini, "What should 6G be?," *Nature Electronics*, 2020, *3* (1), 20–29.

Dao, Tri, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré, "FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness," arXiv preprint arXiv:2205.14135 2022.

Davies, Mike, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, Yuyun Liao, Chit-Kwan Lin, Andrew Lines, Ruokun Liu, Deepak Mathaikutty, Steven McCoy, Arnab Paul, Jonathan Tse, Guruguhanathan Venkataramanan, Yi-Hsin Weng, Andreas Wild, Yoonseok Yang, and Hong Wang, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, 2018, *38* (1), 82–99.

de Vries, Alex, "The growing energy footprint of artificial intelligence," *Joule*, 2023, *7* (10), 2191–2194.

de Vries-Gao, Alex, "The carbon and water footprints of data centers and what this could mean for artificial intelligence," *Patterns*, 2026, *7* (1), 101430.

Dean, Jeffrey and Sanjay Ghemawat, "MapReduce: simplified data processing on large clusters," *Communications of the ACM*, 2008, *51* (1), 107–113.

Deng, Shuiguang, Hailiang Zhao, Weijia Fang, Jianwei Yin, Schahram Dustdar, and Albert Y. Zomaya, "Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence," *IEEE Internet of Things Journal*, 2020, *7* (8), 7457–7469.

Dennard, R.H., F.H. Gaensslen, Hwa-Nien Yu, V.L. Rideout, E. Bassous, and A.R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, 1974, *9* (5), 256–268.

Dinh, Tien Tuan Anh, Rui Liu, Meihui Zhang, Gang Chen, Beng Chin Ooi, and Ji Wang, "Untangling Blockchain: A Data Processing View of Blockchain Systems," *IEEE Transactions on Knowledge and Data Engineering*, 2018, *30* (7), 1366–1385.

Dodge, Jesse, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan, "Measuring the Carbon Intensity of AI in Cloud Instances," in "2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)" ACM 2022, pp. 1877–1894.

Douze, Matthijs, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou, "The Faiss Library," *IEEE Transactions on Big Data*, 2026, *12* (2), 346–361.

Fan, Wenqi, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li, "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models," in "Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)" ACM 2024, pp. 6491–6501.

Fernandez, Raul Castro, Aaron J. Elmore, Michael J. Franklin, Sanjay Krishnan, and Chenhao Tan, "How Large Language Models Will Disrupt Data Management," *Proceedings of the VLDB Endowment*, 2023, *16* (11), 3302–3309.

Frantar, Elias, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh, "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers," arXiv preprint arXiv:2210.17323 2022.

Fu, Tingzhao, Jianfa Zhang, Run Sun, Yuyao Huang, Wei Xu, Sigang Yang, Zhihong Zhu, and Hongwei Chen, "Optical neural networks: progress and challenges," *Light: Science & Applications*, 2024, *13* (1).

Gao, Chen, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li, "A Survey of Graph Neural Networks for Recommender Systems: Challenges, Methods, and Directions," *ACM Transactions on Recommender Systems*, 2023, *1* (1), 1–51.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey," arXiv preprint arXiv:2312.10997 2023.

Ge, Yingqiang, Yujie Ren, Wenyue Hua, Shuyuan Xu, Juntao Tan, and Yongfeng Zhang, "LLM as OS, Agents as Apps: Envisioning AIOS, Agents and the AIOS-Agent Ecosystem," arXiv preprint arXiv:2312.03815 2023.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford, "Datasheets for datasets," *Communications of the ACM*, 2021, *64* (12), 86–92.

Gholami, Amir, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, and Kurt Keutzer, "AI and Memory Wall," *IEEE Micro*, 2024, *44* (3), 33–39.

Gilbert, Seth and Nancy Lynch, "Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services," *ACM SIGACT News*, 2002, *33* (2), 51–59.

Google Quantum AI, "Suppressing quantum errors by scaling a surface code logical qubit," *Nature*, 2023, *614* (7949), 676–681.

Google Quantum AI and Collaborators, Rajeev Acharya, Dmitry A. Abanin, Laleh Aghababaie-Beni, Igor Aleiner, Trond I. Andersen, Markus Ansmann, Frank Arute, Kunal Arya, Abraham Asfaw, Nikita Astrakhantsev, Juan Atalaya, Ryan Babbush, Dave Bacon, Brian Ballard, Joseph C. Bardin, Johannes Bausch, Andreas Bengtsson, Alexander Bilmes, Sam Blackwell, Sergio Boixo, Gina Bortoli, Alexandre Bourassa, Jenna Bovaird, Leon Brill, Michael Broughton, David A. Browne, Brett Buchea, Bob B. Buckley, David A. Buell, Tim Burger, Brian Burkett, Nicholas Bushnell, Anthony Cabrera, Juan Campero, Hung-Shen Chang, Yu Chen, Zijun Chen, Ben Chiaro, Desmond Chik, Charina Chou, Jahan Claes, Agnetta Y. Cleland, Josh Cogan, Roberto Collins, Paul Conner, William Courtney, Alexander L. Crook, Ben Curtin, Sayan Das, Alex Davies, Laura De Lorenzo, Dripto M. Debroy, Sean Demura, Michel Devoret, Agustin Di Paolo, Paul Donohoe, Ilya Drozdov, Andrew Dunsworth, Clint Earle, Thomas Edlich, Alec Eickbusch, Aviv Moshe Elbag, Mahmoud Elzouka, Catherine Erickson, Lara Faoro, Edward Farhi, Vinicius S. Ferreira, Leslie Flores Burgos, Ebrahim Forati, Austin G. Fowler, Brooks Foxen, Suhas Ganjam, Gonzalo Garcia, Robert Gasca, Élie Genois, William Giang, Craig Gidney, Dar Gilboa, Raja Gosula, Alejandro Grajales Dau, Dietrich Graumann, Alex Greene, Jonathan A. Gross, Steve Habegger, John Hall, Michael C. Hamilton, Monica Hansen, Matthew P. Harrigan, Sean D. Harrington, Francisco J. H. Heras, Stephen Heslin, Paula Heu, Oscar Higgott, Gordon Hill, Jeremy Hilton, George Holland, Sabrina Hong, Hsin-Yuan Huang, Ashley Huff, William J. Huggins, Lev B. Ioffe, Sergei V. Isakov, Justin Iveland, Evan Jeffrey, Zhang Jiang, Cody Jones, Stephen Jordan, Chaitali Joshi, Pavol Juhas, Dvir Kafri, Hui Kang, Amir H. Karamlou, Kostyantyn Kechedzhi, Julian Kelly, Trupti Khaire, Tanuj Khattar, Mostafa Khezri, Seon Kim, Paul V. Klimov, Andrey R. Klots, Bryce Kobrin, Pushmeet Kohli, Alexander N. Korotkov, Fedor Kostritsa, Robin Kothari, Borislav Kozlovskii, John Mark Kreikebaum, Vladislav D. Kurilovich, Nathan Lacroix, David Landhuis, Tiano Lange-Dei, Brandon W.

Langley, Pavel Laptev, Kim-Ming Lau, Loïck Le Guevel, Justin Ledford, Joonho Lee, Kenny Lee, Yuri D. Lensky, Shannon Leon, Brian J. Lester, Wing Yan Li, Yin Li, Alexander T. Lill, Wayne Liu, William P. Livingston, Aditya Locharla, Erik Lucero, Daniel Lundahl, Aaron Lunt, Sid Madhuk, Fionn D. Malone, Ashley Maloney, Salvatore Mandrà, James Manyika, Leigh S. Martin, Orion Martin, Steven Martin, Cameron Maxfield, Jarrod R. McClean, Matt McEwen, Seneca Meeks, Anthony Megrant, Xiao Mi, Kevin C. Miao, Amanda Mieszala, Reza Molavi, Sebastian Molina, Shirin Montazeri, Alexis Morvan, Ramis Movassagh, Wojciech Mruczkiewicz, Ofer Naaman, Matthew Neeley, Charles Neill, Ani Nersisyan, Hartmut Neven, Michael Newman, Jiun How Ng, Anthony Nguyen, Murray Nguyen, Chia-Hung Ni, Murphy Yuezhen Niu, Thomas E. O'Brien, William D. Oliver, Alex Opremcak, Kristoffer Ottosson, Andre Petukhov, Alex Pizzuto, John Platt, Rebecca Potter, Orion Pritchard, Leonid P. Pryadko, Chris Quintana, Ganesh Ramachandran, Matthew J. Reagor, John Redding, David M. Rhodes, Gabrielle Roberts, Eliott Rosenberg, Emma Rosenfeld, Pedram Roushan, Nicholas C. Rubin, Negar Saei, Daniel Sank, Kannan Sankaragomathi, Kevin J. Satzinger, Henry F. Schurkus, Christopher Schuster, Andrew W. Senior, Michael J. Shearn, Aaron Shorter, Noah Shutty, Vladimir Shvarts, Shraddha Singh, Volodymyr Sivak, Jindra Skruzny, Spencer Small, Vadim Smelyanskiy, W. Clarke Smith, Rolando D. Somma, Sofia Springer, George Sterling, Doug Strain, Jordan Suchard, Aaron Szasz, Alex Sztein, Douglas Thor, Alfredo Torres, M. Mert Torunbalci, Abeer Vaishnav, Justin Vargas, Sergey Vdovichev, Guifre Vidal, Benjamin Villalonga, Catherine Vollgraff Heidweiller, Steven Waltman, Shannon X. Wang, Brayden Ware, Kate Weber, Travis Weidel, Theodore White, Kristi Wong, Bryan W. K. Woo, Cheng Xing, Z. Jamie Yao, Ping Yeh, Bicheng Ying, Juhwan Yoo, Noureldin Yosri, Grayson Young, Adam Zalcman, Yaxing Zhang, Ningfeng Zhu, and Nicholas Zobrist, "Quantum error correction below the surface code threshold," *Nature*, 2025, *638* (8052), 920–926.

Hauser, Frederik, Marco Häberle, Daniel Merling, Steffen Lindner, Vladimir Gurevich, Florian Zeiger, Reinhard Frank, and Michael Menth, "A survey on data plane programming with P4: Fundamentals, advances, and applied research," *Journal of Network and Computer Applications*, 2023, *212*, 103561.

He, Xiangnan, Kuan Deng, Xiang Wang, Yan Li, YongDong Zhang, and Meng Wang, "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation," in "Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)" ACM 2020, pp. 639–648.

Henderson, Peter, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau, "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine

Learning," arXiv preprint arXiv:2002.05651 2020.

Hennessy, John L. and David A. Patterson, "A new golden age for computer architecture," *Communications of the ACM*, 2019, *62* (2), 48–60.

Huang, Yizheng and Jimmy Xiangji Huang, "A Survey on Retrieval-Augmented Text Generation for Large Language Models," *ACM Computing Surveys*, 2026, *58* (12), 1–38.

Høiland-Jørgensen, Toke, Jesper Dangaard Brouer, Daniel Borkmann, John Fastabend, Tom Herbert, David Ahern, and David Miller, "The eXpress data path: fast programmable packet processing in the operating system kernel," in "Proceedings of the 14th International Conference on emerging Networking EXperiments and Technologies" ACM 2018, pp. 54–66.

Internet Engineering Task Force, "BPF Instruction Set Architecture (ISA)," Technical Report RFC 9669, RFC Editor 2024.

Javadi-Abhari, Ali, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta, "Quantum computing with Qiskit," arXiv preprint arXiv:2405.08810 2024.

Jia, Zhipeng and Emmett Witchel, "Nightcore: efficient and scalable serverless computing for latency-sensitive, interactive microservices," in "Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '21)" ACM 2021, pp. 152–166.

Jiang, Feibo, Cunhua Pan, Kezhi Wang, Pietro Michiardi, Octavia A. Dobre, and Merouane Debbah, "From Large AI Models to Agentic AI: A Tutorial on Future Intelligent Communications," *IEEE Journal on Selected Areas in Communications*, 2026, *44*, 3507–3540.

Johnson, Jeff, Matthijs Douze, and Herve Jegou, "Billion-Scale Similarity Search with GPUs," *IEEE Transactions on Big Data*, 2021, *7* (3), 535–547.

Jonas, Eric, Johann Schleier-Smith, Vikram Sreekanti, Chia-Che Tsai, Anurag Khandelwal, Qifan Pu, Vaishaal Shankar, Joao Carreira, Karl Krauth, Neeraja Yadwadkar, Joseph E. Gonzalez, Raluca Ada Popa, Ion Stoica, and David A. Patterson, "Cloud Programming Simplified: A Berkeley View on Serverless Computing," arXiv preprint arXiv:1902.03383 2019.

Jouppi, Norm, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Clifford Young, Xiang Zhou, Zongwei Zhou, and David A Patterson, "TPU v4: An Optically Reconfigurable Supercomputer

for Machine Learning with Hardware Support for Embeddings," in "Proceedings of the 50th Annual International Symposium on Computer Architecture" ACM 2023, pp. 1–14.

Jouppi, Norman P., Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon, "In-Datacenter Performance Analysis of a Tensor Processing Unit," in "Proceedings of the 44th Annual International Symposium on Computer Architecture" ACM 2017, pp. 1–12.

Kairouz, Peter and H. Brendan McMahan, "Advances and Open Problems in Federated Learning," *Foundations and Trends® in Machine Learning*, 2021, *14* (1-2), 1–210.

Kanno, Keita, Masaya Kohda, Ryosuke Imai, Sho Koh, Kosuke Mitarai, Wataru Mizukami, and Yuya O. Nakagawa, "Quantum-selected configuration interaction: Classical diagonalization of Hamiltonians in subspaces selected by quantum computers," *Physical Review Research*, 2026, *8* (2).

Karpukhin, Vladimir, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in "Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)" Association for Computational Linguistics 2020, pp. 6769–6781.

Khattab, Omar and Matei Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in "Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval" ACM 2020, pp. 39–48.

Kim, Youngseok, Andrew Eddins, Sajant Anand, Ken Xuan Wei, Ewout van den Berg, Sami Rosenblatt, Hasan Nayfeh, Yantao Wu, Michael Zaletel, Kristan Temme, and Abhinav

Kandala, "Evidence for the utility of quantum computing before fault tolerance," *Nature*, 2023, *618* (7965), 500–505.

Kreutz, Diego, Fernando M. V. Ramos, Paulo Esteves Verissimo, Christian Esteve Rothenberg, Siamak Azodolmolky, and Steve Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, 2015, *103* (1), 14–76.

Kuenzer, Simon, Vlad-Andrei Bădoiu, Hugo Lefeuvre, Sharan Santhanam, Alexander Jung, Gaulthier Gain, Cyril Soldani, Costin Lupu, Ştefan Teodorescu, Costi Răducanu, Cristian Banu, Laurent Mathy, Răzvan Deaconescu, Costin Raiciu, and Felipe Huici, "Unikraft: fast, specialized unikernels the easy way," in "Proceedings of the Sixteenth European Conference on Computer Systems" ACM 2021, pp. 376–394.

Kumar, Amit and Naran M. Pindoriya, "Toward sustainable data center operation: A review on existing infrastructures, integrated smart energy management, and emerging opportunities," *Renewable and Sustainable Energy Reviews*, 2026, *230*, 116664.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica, "Efficient Memory Management for Large Language Model Serving with PagedAttention," in "Proceedings of the 29th Symposium on Operating Systems Principles" ACM 2023, pp. 611–626.

Lam, Sean, Ahmed Khaled, Simon Bilodeau, Bicky A. Marquez, Paul R. Prucnal, Lukas Chrostowski, Bhavin J. Shastri, and Sudip Shekhar, "Neuromorphic photonic computing with an electro-optic analog memory," *Nature Communications*, 2026, *17* (1).

Lamport, Leslie, "The part-time parliament," *ACM Transactions on Computer Systems*, 1998, *16* (2), 133–169.

Larocca, Martín, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J. Coles, Lukasz Cincio, Jarrod R. McClean, Zoë Holmes, and M. Cerezo, "Barren plateaus in variational quantum computing," *Nature Reviews Physics*, 2025, *7* (4), 174–189.

Lattner, C. and V. Adve, "LLVM: a compilation framework for lifelong program analysis & transformation," in "International Symposium on Code Generation and Optimization, 2004. CGO 2004." IEEE 2004, pp. 75–86.

Lattner, Chris, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis, Jacques Pienaar, River Riddle, Tatiana Shpeisman, Nicolas Vasilache, and Oleksandr Zinenko, "MLIR: Scaling Compiler Infrastructure for Domain Specific Computation," in "2021 IEEE/ACM International Symposium on Code Generation and Optimization (CGO)" IEEE 2021, pp. 2–14.

Lee, Dayeol, David Kohlbrenner, Shweta Shinde, Krste Asanović, and Dawn Song, "Keystone: an open framework for architecting trusted execution environments," in "Proceedings of the Fifteenth European Conference on Computer Systems (EuroSys '20)" ACM 2020, pp. 1–16.

Leiserson, Charles E., Neil C. Thompson, Joel S. Emer, Bradley C. Kuszmaul, Butler W. Lampson, Daniel Sanchez, and Tao B. Schardl, "There's plenty of room at the Top: What will drive computer performance after Moore's law?," *Science*, 2020, *368* (6495).

Letaief, Khaled B., Yuanming Shi, Jianmin Lu, and Jianhua Lu, "Edge Artificial Intelligence for 6G: Vision, Enabling Technologies, and Applications," *IEEE Journal on Selected Areas in Communications*, 2022, *40* (1), 5–36.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv preprint arXiv:2005.11401 2020.

Li, Guoliang, Xuanhe Zhou, and Lei Cao, "AI Meets Database: AI4DB and DB4AI," in "Proceedings of the 2021 International Conference on Management of Data" ACM 2021, pp. 2859–2866.

Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren, "Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models," arXiv preprint arXiv:2304.03271 2023.

_ , _ , _ , and _ , "Making AI Less 'Thirsty'," *Communications of the ACM*, 2025, *68* (7), 54–61.

Li, Tian, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Processing Magazine*, 2020, *37* (3), 50–60.

Li, Xuexian, Xiayuan Liu, Zilong Wang, Zijian Xiao, Chun-Yao Hsieh, and Yixue Liu, "DeepServeCB: SLO-Aware and Cost-Aware Elastic Scheduling for Serverless Multi-Tenant LLM Inference," in "2026 International Conference on Generative Artificial Intelligence and Information Security (GAIIS)" IEEE 2026, pp. 90–95.

Liao, Jiajing, Feng Chang, Yihan Xue, Tianjian Xia, Zeyu Huang, and Yuxiao Wang, "Multi-Objective Scheduling for Large Language Model Inference with Prompt-Level Cost Prediction and SLO Awareness," in "2026 IEEE 8th International Conference on Communications, Information System and Computer Engineering (CISCE)" IEEE 2026, pp. 1606–1612.

Lin, Ji, Jiaming Tang, Haotian Tang, Shang Yang, Guangxuan Xiao, and Song Han, "AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration," *GetMobile: Mobile Computing and Communications*, 2025, *28* (4), 12–17.

Liu, Fan, Yuanhao Cui, Christos Masouros, Jie Xu, Tony Xiao Han, Yonina C. Eldar, and Stefano Buzzi, "Integrated Sensing and Communications: Toward Dual-Functional Wireless Networks for 6G and Beyond," *IEEE Journal on Selected Areas in Communications*, 2022, *40* (6), 1728–1767.

Liu, Jing, Yao Du, Kun Yang, Jiaqi Wu, Yan Wang, Xiping Hu, Zehua Wang, Yang Liu, Peng Sun, Azzedine Boukerche, and Victor C. M. Leung, "Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization: A Survey," *IEEE Communications Surveys & Tutorials*, 2026, *28*, 5049–5080.

Luccioni, Alexandra Sasha, Emma Strubell, and Kate Crawford, "From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate," in "Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)" ACM 2025, pp. 76–88.

_ , Sylvain Viguier, and Anne-Laure Ligozat, "Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model," arXiv preprint arXiv:2211.02001 2022.

Luccioni, Sasha, Yacine Jernite, and Emma Strubell, "Power Hungry Processing: Watts Driving the Cost of AI Deployment?," in "The 2024 ACM Conference on Fairness Accountability and Transparency" ACM 2024, pp. 85–99.

Ma, Le, Ran Zhang, Yikun Han, Shirui Yu, Zaitian Wang, Zhiyuan Ning, Jinghan Zhang, Ping Xu, Pengjiang Li, Ziyue Qiao, Wei Ju, Chong Chen, Dongjie Wang, Kunpeng Liu, Pengyang Wang, Pengfei Wang, Yanjie Fu, Chunjiang Liu, Yuanchun Zhou, and Chang-Tien Lu, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," arXiv preprint arXiv:2310.11703 2023.

Madhavapeddy, Anil, Richard Mortier, Charalampos Rotsos, David Scott, Balraj Singh, Thomas Gazagnaire, Steven Smith, Steven Hand, and Jon Crowcroft, "Unikernels: library operating systems for the cloud," in "Proceedings of the eighteenth international conference on Architectural support for programming languages and operating systems" ACM 2013, pp. 461–472.

Main, D., P. Drmota, D. P. Nadlinger, E. M. Ainley, A. Agrawal, B. C. Nichol, R. Srinivas, G. Araneda, and D. M. Lucas, "Distributed quantum computing across an optical network link," *Nature*, 2025, *638* (8050), 383–388.

Malkov, Yu A. and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, *42* (4), 824–836.

Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey, "Recalibrating global data center energy-use estimates," *Science*, 2020, *367* (6481), 984–986.

McKeown, Nick, Tom Anderson, Hari Balakrishnan, Guru Parulkar, Larry Peterson, Jennifer Rexford, Scott Shenker, and Jonathan Turner, "OpenFlow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, 2008, *38* (2), 69–74.

McMahan, H. Brendan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," arXiv preprint arXiv:1602.05629 2016.

Mei, Kai, Xi Zhu, Wujiang Xu, Wenyue Hua, Mingyu Jin, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang, "AIOS: LLM Agent Operating System," arXiv preprint arXiv:2403.16971 2024.

Mohammed, Sedir, Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch, "The effects of data quality on machine learning performance on tabular data," *Information Systems*, 2025, *132*, 102549.

Mutlu, Onur, Saugata Ghose, Juan Gómez-Luna, and Rachata Ausavarungnirun, "A Modern Primer on Processing in Memory," in "Emerging Computing: From Devices to Systems" Computer Architecture and Design Methodologies, Springer Nature Singapore, 2023, pp. 171–243.

Narayanan, Deepak, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia, "Efficient large-scale language model training on GPU clusters using megatron-LM," in "Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis" ACM 2021, pp. 1–15.

National Institute of Standards and Technology (US), "Module-lattice-based key-encapsulation mechanism standard," Technical Report, National Institute of Standards and Technology (U.S.) 2024.

Niu, Chenxu, Wei Zhang, Jie Li, Yongjian Zhao, Tongyang Wang, Xi Wang, and Yong Chen, "TokenPowerBench: Benchmarking the Power Consumption of LLM Inference," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026, *40* (38), 32582–32590.

Packer, Charles, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez, "MemGPT: Towards LLMs as Operating Systems," arXiv preprint arXiv:2310.08560 2023.

Pan, James Jie, Jianguo Wang, and Guoliang Li, "Survey of vector database management systems," *The VLDB Journal*, 2024, *33* (5), 1591–1615.

Patel, Pratyush, Esha Choukse, Chaojie Zhang, Aashaka Shah, Íñigo Goiri, Saeed Maleki, and Ricardo Bianchini, "Splitwise: Efficient Generative LLM Inference Using Phase Splitting," in "2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)" IEEE 2024, pp. 118–132.

Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean, "Carbon Emissions and Large Neural Network Training," arXiv preprint arXiv:2104.10350 2021.

Pope, Reiner, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean, "Efficiently Scaling Transformer Inference," arXiv preprint arXiv:2211.05102 2022.

Prabhu, Ramya, Ajay Nayak, Jayashree Mohan, Ramachandran Ramjee, and Ashish Panwar, "vAttention: Dynamic Memory Management for Serving LLMs without PagedAttention," in "Proceedings of the 30th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '25)" ACM 2025, pp. 1133–1150.

Preskill, John, "Quantum Computing in the NISQ era and beyond," *Quantum*, 2018, *2*, 79.

Qu, Guanqiao, Qiyuan Chen, Wei Wei, Zheng Lin, Xianhao Chen, and Kaibin Huang, "Mobile Edge Intelligence for Large Language Models: A Contemporary Survey," *IEEE Communications Surveys & Tutorials*, 2025, *27* (6), 3820–3860.

Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He, "ZeRO: Memory optimizations Toward Training Trillion Parameter Models," in "SC20: International Conference for High Performance Computing, Networking, Storage and Analysis" IEEE 2020, pp. 1–16.

Regev, Oded, "On lattices, learning with errors, random linear codes, and cryptography," *Journal of the ACM*, 2009, *56* (6), 1–40.

Reuther, Albert, Peter Michaleas, Michael Jones, Vijay Gadepally, Siddharth Samsi, and Jeremy Kepner, "AI and ML Accelerator Survey and Trends," in "2022 IEEE High Performance Extreme Computing Conference (HPEC)" IEEE 2022, pp. 1–10.

Robertson, Stephen and Hugo Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends® in Information Retrieval*, 2009, *4* (1-2), 1–174.

Sapio, Amedeo, Marco Canini, Chen-Yu Ho, Jacob Nelson, Panos Kalnis, Changhoon Kim, Arvind Krishnamurthy, Masoud Moshref, Dan R. K. Ports, and Peter Richtárik, "Scaling Distributed Machine Learning with In-Network Aggregation," arXiv preprint arXiv:1903.06701 2019.

Satyanarayanan, Mahadev, "The Emergence of Edge Computing," *Computer*, 2017, *50* (1), 30–39.

Schuman, Catherine D., Shruti R. Kulkarni, Maryam Parsa, J. Parker Mitchell, Prasanna Date, and Bill Kay, "Opportunities for neuromorphic computing algorithms and applications," *Nature Computational Science*, 2022, *2* (1), 10–19.

Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni, "Green AI," *Communications of the ACM*, 2020, *63* (12), 54–63.

Shi, Weisong, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu, "Edge Computing: Vision and Challenges," *IEEE Internet of Things Journal*, 2016, *3* (5), 637–646.

Shoeybi, Mohammad, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro, "Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism," arXiv preprint arXiv:1909.08053 2019.

Shor, Peter W., "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," *SIAM Journal on Computing*, 1997, *26* (5), 1484–1509.

Silvano, Cristina, Daniele Ielmini, Fabrizio Ferrandi, Leandro Fiorin, Serena Curzel, Luca Benini, Francesco Conti, Angelo Garofalo, Cristian Zambelli, Enrico Calore, Sebastiano Schifano, Maurizio Palesi, Giuseppe Ascia, Davide Patti et al., "A Survey on Deep Learning Hardware Accelerators for Heterogeneous HPC Platforms," *ACM Computing Surveys*, 2025, *57* (11), 1–39.

Song, Yixin, Zeyu Mi, Haotong Xie, and Haibo Chen, "PowerInfer: Fast Large Language Model Serving with a Consumer-grade GPU," in "Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles (SOSP '24)" ACM 2024, pp. 590–606.

Stonebraker, Michael and Andrew Pavlo, "What Goes Around Comes Around... And Around...," *ACM SIGMOD Record*, 2024, *53* (2), 21–37.

Strati, Foteini, Xianzhe Ma, and Ana Klimovic, "Orion: Interference-aware, Fine-grained GPU Sharing for ML Applications," in "Proceedings of the Nineteenth European Conference on Computer Systems (EuroSys '24)" ACM 2024, pp. 1075–1092.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics" Association for Computational Linguistics 2019, pp. 3645–3650.

Sze, Vivienne, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," *Proceedings of the IEEE*, 2017, *105* (12), 2295–2329.

Taipalus, Toni, "Vector database management systems: Fundamental concepts, use-cases, and current challenges," *Cognitive Systems Research*, 2024, *85*, 101216.

Thakur, Dipanwita, Antonella Guzzo, Giancarlo Fortino, and Francesco Piccialli, "Green Federated Learning: A New Era of Green Aware AI," *ACM Computing Surveys*, 2025, *57* (8), 1–36.

Tillet, Philippe, H. T. Kung, and David Cox, "Triton: an intermediate language and compiler for tiled neural network computations," in "Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages" ACM 2019, pp. 10–19.

Verma, Abhishek, Luis Pedrosa, Madhukar Korupolu, David Oppenheimer, Eric Tune, and John Wilkes, "Large-scale cluster management at Google with Borg," in "Proceedings of the Tenth European Conference on Computer Systems" ACM 2015, pp. 1–17.

Vieira, Marcos A. M., Matheus S. Castanho, Racyus D. G. Pacífico, Elerson R. S. Santos, Eduardo P. M. Câmara Júnior, and Luiz F. M. Vieira, "Fast Packet Processing with eBPF and XDP: Concepts, Code, Challenges, and Applications," *ACM Computing Surveys*, 2021, *53* (1), 1–36.

Wang, Cheng-Xiang, Xiaohu You, Xiqi Gao, Xiuming Zhu, Zixin Li, Chuan Zhang, Haiming Wang, Yongming Huang, Yunfei Chen, Harald Haas, John S. Thompson, Erik G. Larsson, Marco Di Renzo, Wen Tong, Peiying Zhu, Xuemin Shen, H. Vincent Poor, and Lajos Hanzo, "On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds," *IEEE Communications Surveys & Tutorials*, 2023, *25* (2), 905–974.

Wu, Carole-Jean, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin S. Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood, "Sustainable AI: Environmental Implications, Challenges and Opportunities," arXiv preprint arXiv:2111.00364 2021.

Wu, Likang, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen, "A survey on large language models for recommendation," *World Wide Web*, 2024, *27* (5).

Wu, Shiwen, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui, "Graph Neural Networks in Recommender Systems: A Survey," *ACM Computing Surveys*, 2023, *55* (5), 1–37.

Xiao, Shitao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie, "C-Pack: Packed Resources For General Chinese Embeddings," in "Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)" ACM 2024, pp. 641–649.

Xu, Zhihao, Tiankuang Zhou, Muzhou Ma, ChenChen Deng, Qionghai Dai, and Lu Fang, "Large-scale photonic chiplet Taichi empowers 160-TOPS/W artificial general intelligence," *Science*, 2024, *384* (6692), 202–209.

Zaharia, Matei, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica, "Apache Spark: a unified engine for big data processing," *Communications of the ACM*, 2016, *59* (11), 56–65.

Zha, Daochen, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu, "Data-centric Artificial Intelligence: A Survey," *ACM Computing Surveys*, 2025, *57* (5), 1–42.

Zhang, Chaoyun, Paul Patras, and Hamed Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," *IEEE Communications Surveys & Tutorials*, 2019, *21* (3), 2224–2287.

Zhang, Ruichen, Guangyuan Liu, Yinqiu Liu, Changyuan Zhao, Jiacheng Wang, Yunting Xu, Dusit Niyato, Jiawen Kang, Yonghui Li, Shiwen Mao, Sumei Sun, Xuemin Shen, and Dong In Kim, "Toward Edge General Intelligence With Agentic AI and Agentification: Concepts, Technologies, and Future Directions," *IEEE Communications Surveys & Tutorials*, 2026, *28*, 4285–4318.

Zhao, Penghao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui, "Retrieval-Augmented Generation for AI-Generated Content: A Survey," *Data Science and Engineering*, 2026, *11* (1), 1–29.

Zhao, Wayne Xin, Jing Liu, Ruiyang Ren, and Ji-Rong Wen, "Dense Text Retrieval Based on Pretrained Language Models: A Survey," *ACM Transactions on Information Systems*, 2024, *42* (4), 1–60.

Zhao, Yanli, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li, "PyTorch FSDP: Experiences on Scaling Fully Sharded Data Parallel," arXiv preprint arXiv:2304.11277 2023.

Zhao, Zihuai, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li, "Recommender Systems in the Era of Large Language Models (LLMs)," *IEEE Transactions on Knowledge and Data Engineering*, 2024, *36* (11), 6889–6907.

Zhou, Guangyao, Wenhong Tian, Rajkumar Buyya, Ruini Xue, and Liang Song, "Deep reinforcement learning-based methods for resource scheduling in cloud computing: a review and future directions," *Artificial Intelligence Review*, 2024, *57* (5).

Zhou, Zhi, Xu Chen, En Li, Liekang Zeng, Ke Luo, and Junshan Zhang, "Edge Intelligence: Paving the Last Mile of Artificial Intelligence With Edge Computing," *Proceedings of the IEEE*, 2019, *107* (8), 1738–1762.

Zhu, Yutao, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen, "Large Language Models for Information Retrieval: A Survey," arXiv preprint arXiv:2308.07107 2023.