Future of Chip Technologies

spanic

S.No	Feature	GPU (Graphics Processing Unit)	ASIC (Application-Specific Integrated Circuit)	Neuromorphic Processor	Photonic Computing (Optical)	Memristor-based Computing	Quantum Processor	Graphene-based Computing	Cryogenic Computing (Superconducting Chips)
Midpoint Years		0	0	4	5	7.5	7.5	9.5	12.5
Estimated Wider Availability		**Widespread**. Continuous new generations.	**Widespread** in cloud; growing in enterprise. Continuous new iterations.	**3-5 years** for broader commercial adoption in specialized areas.	**3-7 years** for more widespread compute; interconnects now.	**5-10 years** for significant commercial products.	**5-10+ years** for practical, fault-tolerant systems beyond research.	**7-12 years** for commercial products; early prototypes emerging.	**10-15+ years** for broader practical applications beyond research.
1	**Main Area Catered**	**General-purpose** AI acceleration (cloud to edge), high-performance computing.	**Optimized** AI acceleration for specific workloads, reducing TCO for hyperscalers.	**Energy-efficient** computing, neuromorphic sensing, edge processing.	**High-speed** data processing and interconnects, especially for AI.	**Low-power** memory and in-memory compute for AI and general computing.	**Future-forward** problem-solving for classically impossible computations.	**Ultra-low-power** electronics for AI, edge computing, general-purpose processing.	**High-performance** processing in niche, demanding environments.
2	**Technological Difference**	**Parallel** digital processors (CUDA/OpenCL), optimized for matrix operations (tensor cores). Von Neumann architecture.	**Custom-designed** chips optimized for specific AI workloads (e.g., matrix multiplication, neural network layers).	**Spiking** neural networks (SNNs). Event-driven, in-memory computing.	**Photons** for computation and data transmission. Replaces electrons with light for speed & energy efficiency.	**Non-volatile** memory elements (memristors) with resistance changes, enabling in-memory analog compute.	**Quantum-mechanical** phenomena (superposition, entanglement) for computation. Qubits instead of bits.	**Graphene's** high electron mobility and thermal conductivity for ultra-fast, low-power transistors and interconnects. Beyond CMOS.	**Superconductors** (e.g., Josephson junctions) at near absolute zero. Eliminates electrical resistance for ultra-fast/low-power.
3	**Performance Metrics**	**~4000** TFLOPS FP8 (NVIDIA H100). High latency for complex models.	**~400** TOPS (AWS Trainium). Lower latency for specific workloads.	**~10** TOPS (Intel Loihi). Low latency for event-driven tasks.	**~100-1000** TOPS (projected for compute). Sub-ns latency for interconnects.	**~10-100** TOPS (projected). Low latency for analog compute.	**~100-1000** qubits (e.g., IBM Heron).	**~10-100** TOPS (projected). Sub-ns latency for interconnects.	**~1000** TOPS (projected). Ultra-low latency due to superconductivity.
4	**Primary Use Cases**	**AI Training** (large models), AI Inference, Scientific Simulation, Graphics, HPC, Data Analytics.	**AI Inference** (Cloud & Edge), Custom AI Training, Domain-specific AI acceleration (e.g., video analytics, recommender systems).	**Edge AI**, Real-time sensory processing, Low-power inference, Robotics, Continuous learning.	**Ultra-fast** AI inference, High bandwidth interconnects, Telecom, LiDAR, Specialized linear algebra acceleration.	**Low-power** AI inference, Edge computing, Non-volatile memory, Brain-inspired computing, Data centers.	**Optimization** problems, Drug Discovery, Materials Science, Complex System Simulation, Cryptography.	**Ultra-low-power** AI inference, High-speed interconnects, Flexible electronics, Wearables, Next-gen data centers.	**Extreme** performance HPC, Quantum Computing control, Ultra-low noise sensing, Scientific research.
5	**Best Environment (On-Premise vs. Cloud)**	**Cloud-based** preferred due to high setup costs (~$30k/unit, cooling/power). On-premise for large organizations with dedicated data centers. ~70% cloud, ~30% on-premise.	**Cloud-based** dominates due to hyperscaler optimization. On-premise growing for specialized enterprise (~$10k-$50k/unit). ~80% cloud, ~20% on-premise.	**On-premise** for edge deployments due to low-power (~1W). Cloud-based for hybrid setups. ~60% on-premise, ~40% cloud.	**Cloud-based** for interconnects due to high integration costs (~$10k-$100k). On-premise for specialized HPC/telecom. ~75% cloud, ~25% on-premise.	**On-premise** for edge devices due to low-power (~1-10W). Cloud-based for data center memory. ~50% on-premise, ~50% cloud (projected).	**Cloud-based** dominates due to high costs (~$1M-$10M/system) and complex infrastructure. ~90% cloud, ~10% on-premise.	**On-premise** for edge devices due to low-power (~1-5W). Cloud-based for interconnects. ~60% on-premise, ~40% cloud (projected).	**On-premise** in specialized research facilities due to extreme cooling costs (~$1M-$10M). ~95% on-premise, ~5% cloud.
6	**Public Availability**	**Widely** available (e.g., NVIDIA H100, AMD MI300) for purchase or via cloud (AWS, Azure, GCP).	**Cloud-based** as a service (Google TPUs, AWS Trainium/Inferentia, Azure Maia); some on-prem (Cerebras, Groq).	**Available** for researchers (Intel Loihi/Hala Point, IBM TrueNorth/NorthPole) and emerging commercial products.	**Early** commercial products emerging (e.g., Lightmatter, Ayar Labs), often for specialized data center or HPC interconnects.	**R&D** and specialized prototypes. Some academic/industry partnerships.	**Cloud-based** access (IBM Quantum Experience, Google Quantum AI, AWS Braket); limited on-prem.	**Research** labs and early prototypes (e.g., graphene transistors, interconnects).	**Research** labs & specialized facilities due to extreme cooling requirements.
7	**How they Complement Each Other**	**Foundational** training for models run on ASICs and neuromorphic chips.	**Efficient** inference of models trained on GPUs; paired with CPUs for broader systems.	**Neuromorphic** inference for efficient deployment of GPU/ASIC-trained models at the edge.	**Photonic** interconnects enhance data flow for GPUs/ASICs. Optical compute offloads specific tasks.	**Memristors** enable efficient in-memory compute for GPUs/ASICs or neuromorphic architectures.	**Quantum** could train or optimize classical AI models; classical chips manage quantum control.	**Graphene** enhances transistor efficiency in GPUs/ASICs or enables low-power neuromorphic/photonic systems.	**Ultra-fast** classical control for quantum systems; future HPC building blocks.
8	**Will they be Competitors?**	**Yes**, ASICs compete with GPUs for specific AI inference tasks, especially in cloud.	**Yes**, ASICs offer better price/performance/watt than GPUs for specific AI acceleration.	**Yes**, for energy-efficient edge inference; less for general-purpose training.	**Can** compete for specific accelerator functions (e.g., matrix math) and high-bandwidth interconnects.	**Can** compete with traditional memory and digital AI accelerators for efficiency.	**Not** directly today. Quantum could solve some AI problems faster, but for different problem types.	**Yes**, could compete with GPUs/ASICs for low-power, high-speed applications, especially in edge and wearables.	**Not** direct competitors to most AI chips; serve different purposes.
9	**How Widely Used?**	**Extremely** Widespread. Dominant for deep learning.	**Widespread** in hyperscalers. Growing adoption in enterprise for specific inference.	**Niche**. Used in research, specific edge deployments (e.g., Intel Loihi for robotics).	**Niche**. Gaining traction in data centers for interconnects; compute still experimental.	**Very** Niche. Mostly in university labs, R&D for future memory and compute.	**Very** Niche. Primarily for research, early commercial pilots, and proof-of-concept.	**Extremely** Niche. Limited to academic research and early industrial pilots.	**Extremely** Niche. Limited to high-end scientific research facilities.
10	**Already in Market?**	**mature** products (NVIDIA H100/B200, AMD MI300) available.	**mature** products (Google TPUs, AWS Inferentia/Trainium, Cerebras WSE, Groq LPU, Microsoft Maia).	**Yes**, but primarily for R&D/early adopters (Intel Loihi/Hala Point, IBM NorthPole).	**Yes**, for interconnects (e.g., co-packaged optics), compute still in early stages.	**No**, not yet in widespread commercial products. Prototypes exist.	**Yes**, via cloud services for select users. Physical machines for major research.	**No**, primarily in R&D. Early prototypes (e.g., graphene transistors) emerging.	**No**, not for general computing. Lab use only.
11	**Market Outlook**	**Massive** Growth. Dominant for AI training; strong for inference. Market value soaring.	**Strong** Growth. Increasing share of AI accelerator market, especially in cloud/edge inference.	**High** Growth Potential. Market to reach ~$1.32B by 2030 (CAGR ~89.7%). Driven by edge AI, IoT.	**Strong** Growth. Silicon Photonics market to surpass $50B by 2035. Driven by data centers & AI.	**High** Growth Potential. Significant growth for in-memory compute and non-volatile memory.	**Significant** Future Growth. Market to exceed $300B by 2030. Early-stage, long-term impact.	**High** Growth Potential. Graphene electronics market to reach $5B by 2035. Driven by low-power electronics and wearables.	**Long-term** potential. Early-stage research; market for components (cryostats) growing due to quantum computing.
12	**Hardware Fit**	**General** server hardware (PCIe slots); requires specific cooling/power for high-end.	**Specific** server/system integration; often PCIe cards or integrated in cloud infrastructure.	**Specific** neuromorphic hardware; not plug-and-play with standard CPU/GPU systems.	**Integrated** into standard silicon (hybrid) for interconnects; dedicated optical compute requires new systems.	**Integrated** into CMOS (hybrid) for memory or in-memory compute; may require new architectures.	**Specialized** quantum computers (cryogenic systems, vacuum chambers, control electronics).	**Integrated** into CMOS for transistors/interconnects; may require new fabrication processes.	**Extreme** cryogenic cooling systems; not compatible with standard hardware.
13	**Estimated Research/Market Investment Scale**	**Billions** to Tens of Billions USD Annually.	**Billions** to Tens of Billions USD Annually.	**Hundreds** of Millions USD Annually.	**Billions** USD Annually.	**Hundreds** of Millions USD Annually.	**Billions** USD Annually.	**Hundreds** of Millions USD Annually.	**Hundreds** of Millions USD Annually.
14	**Main Companies & Chips**	**NVIDIA** (H100, B200, Blackwell), **AMD** (MI300, MI350), **Intel** (Gaudi).	**Google** (TPU), **AWS** (Inferentia, Trainium), **Microsoft** (Maia), **Meta** (MTIA), **Cerebras** (WSE), **Groq** (LPU), **Tenstorrent**.	**Intel** (Loihi, Hala Point), **IBM** (TrueNorth, NorthPole), **BrainChip** (Akida).	**Lightmatter** (Envise), **Ayar Labs** (TeraPHY), Celestial AI, Intel, Broadcom.	IBM, Intel, Samsung, Micron, Crossbar, Weebit Nano, 4DS Memory.	**IBM** (Condor, Heron), **Google** (Sycamore, Trillium), **Quantinuum** (H1 series), **Microsoft** (Majorana 1).	**Graphenea**, Grolltex, IBM, Samsung, MIT (research), Paragraf (early transistors).	**IBM**, Google, Microsoft, Intel (for quantum control components).
15	**Emerging Players**	**Graphcore** (IPU), SambaNova (SN40L).	**SambaNova** (Cardinal), d-Matrix (Corsair), Mythic (AMP).	**SynSense** (Speck), GrAI Matter Labs (NeuronFlow).	**Optalysys**, QuiX Quantum (photonic quantum).	**Knowm**, **Adesto Technologies**.	**IonQ** (Aria), **Rigetti** (Aspen), **D-Wave** (Advantage).	**Black** Semiconductor, **Versarien** (graphene interconnects).	**Oxford** Instruments (cryogenic systems), **Quantum Circuits Inc.**
16	**Software Ecosystem**	**Mature**: CUDA, cuDNN, TensorRT, PyTorch, TensorFlow. Proprietary + open-source.	**TensorFlow** (Google TPU), ONNX, proprietary frameworks (AWS, Cerebras).	**Emerging**: Lava (Intel), SpiNNaker, limited PyTorch support.	**Limited**: Proprietary SDKs (Lightmatter). Emerging ONNX support.	**Early**: Custom frameworks in R&D. Limited standard support.	**Qiskit** (IBM), **Cirq** (Google), **PennyLane. Mostly open-source.**	**Early**: Custom tools for graphene circuits. No standard frameworks.	**Limited**: Custom control software for quantum/HPC. Proprietary.
17	**Scalability and Cost**	**Highly** scalable in data centers (PCIe, DGX systems). ~$30k/unit (H100).	**Scalable** in cloud/enterprise. ~$10k-$50k/unit (e.g., Cerebras WSE).	**Limited** scalability (edge-focused). ~$1k-$10k for research platforms.	**Scalable** for interconnects; compute less so. ~$10k-$100k for early systems.	**Potentially** scalable for memory. Costs TBD (prototypes ~$1k-$10k).	**Low** scalability (specialized systems). ~$1M-$10M/system.	**Scalable** for interconnects; compute TBD. ~$1k-$10k (projected).	**Low** scalability due to cooling. ~$1M-$10M for cryostats.
18	**Energy Efficiency Metrics**	**~0.5-1** TOPS/W (e.g., NVIDIA H100: ~700W for ~4000 TFLOPS FP8). High power for training.	**~2-5** TOPS/W (e.g., AWS Inferentia: ~100W for ~400 TOPS). Optimized for inference.	**~10-50** TOPS/W (e.g., Intel Loihi: ~1W for ~10 TOPS). Highly efficient for edge.	**~5-20** TOPS/W for compute; interconnects reduce system-level power by ~50%.	**10-100** TOPS/W due to in-memory compute. Prototypes show promise.	**Not** comparable (qubits-based). Power dominated by cryogenic cooling (~25kW/system).	**50-200** TOPS/W due to graphene's low resistance. Early R&D estimates.	**~100** TOPS/W possible due to zero resistance, but high cooling costs.
19	**Environmental Impact**	**High**: ~700W/chip, ~1-2 kg CO2e/TFLOP (data center scale).	**Moderate**: ~100W/chip, ~0.5-1 kg CO2e/TOP.	**Low**: ~1W/chip, ~0.01-0.1 kg CO2e/TOP.	**Low**: ~50W for interconnects, ~0.1-0.5 kg CO2e/TOP.	**Very** Low: ~1-10W (projected), ~0.01-0.1 kg CO2e/TOP.	**High**: ~25kW/system due to cooling, ~10-100 kg CO2e/operation.	**Very** Low: ~1-5W (projected), ~0.01-0.05 kg CO2e/TOP.	**High**: ~10-50kW/system for cooling, ~10-100 kg CO2e/operation.
20	**IT and OT Usefulness**	**IT**: Highly useful for AI training/inference, HPC, data analytics in cloud/data centers. **OT**: Moderately useful for edge AI (e.g., autonomous vehicles), limited by **high** power (~700W).	**IT**: Very useful for cloud-based AI inference/training, reducing TCO. **OT**: Useful for **edge** inference in industrial IoT, robotics (e.g., Groq LPU).	**IT**: Limited use in data centers; growing for edge inference in hybrid setups. **OT**: Highly useful for **low-power**, real-time processing in robotics, IoT (~1W).	**IT**: Very useful for **data** center interconnects, emerging for AI compute. **OT**: Moderately useful for high-speed LiDAR, telecom in industrial settings.	**IT**: Promising for **in-memory** computing in data centers. **OT**: Highly promising for low-power edge/IoT, but 5-10 years away.	**IT**: Useful for **specialized** optimization, cryptography in cloud. **OT**: Minimal use due to cost, complexity.	**IT**: Promising for **low-power** data center processors, interconnects. **OT**: Highly promising for wearables, IoT, but 7-12 years away.	**IT**: Useful for **niche** HPC, quantum control. **OT**: Minimal use due to cryogenic requirements.
21	**CFO and CTO Decision Strategy Recommendation**	**CFO**: Favor cloud to avoid high upfront costs (~$30k/unit). Evaluate TCO for on-premise if long-term AI training justifies. **CTO**: Use **cloud** for scalable AI; on-premise for custom HPC, low-latency inference. Invest in CUDA training. Strategy: Cloud-first; pilot on-premise clusters. Monitor next-gen GPUs.	**CFO**: Prioritize cloud for cost efficiency (e.g., AWS Trainium). On-premise for high-volume workloads with ROI. **CTO**: Use **cloud** for rapid deployment; on-premise ASICs for custom inference. Build vendor-agnostic stacks (ONNX). Strategy: Start with cloud; scale on-premise for specialized AI. Partner with ASIC vendors.	**CFO**: Invest in on-premise for edge due to low operational costs (~1W). Cloud for R&D minimizes risk. **CTO**: Deploy **on-premise** for real-time edge AI; cloud for hybrid research. Explore open-source frameworks (Lava). Strategy: Pilot edge neuromorphic chips; collaborate with research institutions.	**CFO**: Choose cloud for interconnects to reduce setup costs (~$10k-$100k). On-premise for long-term HPC savings. **CTO**: Integrate **photonic** interconnects in cloud; test on-premise optical compute. Develop optical expertise. Strategy: Adopt cloud interconnects; R&D partnerships for optical compute.	**CFO**: Avoid large investments due to R&D phase; fund pilot projects for edge memory. Cloud for R&D. **CTO**: Test **on-premise** prototypes for edge AI; cloud for simulation. Build in-memory computing expertise. Strategy: Academic partnerships; focus on low-power edge for 5-10 year horizon.	**CFO**: Prioritize cloud access (e.g., AWS Braket) to avoid prohibitive costs (~$1M-$10M). Limit on-premise to research. **CTO**: Use **cloud** for algorithm development; on-premise for specialized research. Train on quantum frameworks (Qiskit). Strategy: Cloud platforms for experimentation; invest in quantum talent.	**CFO**: Fund R&D pilots for edge devices; avoid large-scale investment until commercialization (7-12 years). Cloud for simulations. **CTO**: Test **on-premise** graphene prototypes for low-power edge; cloud for R&D. Build graphene fabrication expertise. Strategy: Partner with startups (e.g., Paragraf); focus on wearables/IoT.	**CFO**: Limit on-premise investment due to high costs (~$1M-$10M). Fund research grants instead of cloud. **CTO**: Deploy **on-premise** in specialized labs for quantum control/HPC. Develop cryogenic expertise. Strategy: Research collaborations; avoid commercial deployment until 10-15+ years.
22	**Security Features by Default**	**IT**: Limited inherent security. Relies on software-level security (e.g., CUDA memory isolation, TEEs). Vulnerable to side-channel attacks (~700W). **OT**: Minimal security due to high power, complex integration. **Default**: **None** intrinsic; depends on system-level measures.	**IT**: Moderate security. Some include custom security modules (e.g., secure **enclaves**, root of trust). **OT**: Limited security; some offer secure boot. **Default**: **Basic** secure boot/firmware isolation in some designs.	**IT**: Emerging security. Event-driven processing (~1W) reduces side-channel risks. SNN randomness resists adversarial attacks. **OT**: Strong potential for **edge** security. **Default**: **Low-power** signature, randomized SNN behavior.	**IT**: Moderate security. Light-based processing resists EMI attacks. Optical interconnects reduce interception risk. **OT**: Limited security due to complex integration. **Default**: **EMI** resistance, low interception risk.	**IT**: High potential security. Non-volatile memory, multi-level conductance (>16) enable PUFs for secure key generation. **OT**: Strong security for **edge** via PUFs, analog switching. **Default**: **PUFs**, randomized conductance.	**IT**: Strong inherent security. Quantum mechanics enables QKD, quantum randomness for secure computation. **OT**: Limited applicability due to cryogenic needs. **Default**: **Quantum** randomness, QKD.	**IT**: High security potential. Inherent disorders enable robust PUFs for key generation, authentication. **OT**: Strong security for **edge** via PUFs. **Default**: **PUFs**, non-volatile memory.	**IT**: Moderate security. Zero resistance reduces side-channel risks. Cryogenic isolation deters tampering. **OT**: Minimal applicability. **Default**: **Cryogenic** isolation, low power signature.

subin@panicker.uk
Date: 04/06/2025

spanic