International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0127
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 832 - 837

Artificial Intelligence in Semi-Conductor Chips

Vandana Bhat[1], Madhura Bhat[2], Manjunath Dixit[3], Subrahmanya Bhat[4], Shrikant Bhat[5], Nitin Bandekar[6]
[1]Assistant Professor, Dept. of MCA, M.P.E.Society's S.D.M.College, Honnavar, India
[2]Assistant Professor, Dept. of Computer Science, M.P.E.Society's S.D.M.College, Honnavar, India
[3,4,5]Assistant Professor, Dept. of BCA, M.P.E.Society's S.D.M.College, Honnavar, India
[6]PG Scholar, Dept. of MCA, M.P.E.Society's S.D.M.College, Honnavar, India
Emails: *vandana.bhat0131@gmail.com[1], bhatmadhura77@gmail.com[2], manjunathdixit882@gmail.com[3], subbubhat273@gmail.com[4], shribhat2014@gmail.com[5], nitinbandekar901@gmail.com[6]*

Abstract
By 2026, the global semiconductor scene looks pretty different, instead of everyone relying on giant GPU clusters, the industry has moved toward distributed, specialized AI accelerators. Here, we dig into how neuromorphic computing and 3D-stacked Gate-All-Around (GAAFET) transistors come together to break through the old "Power Wall" that held back AI processing. Our framework for on-chip adaptive learning cuts data center dependence by 40 percent a big shift. What stands out is this: combining RISC-V architectures with spiking neural networks (SNNs) delivers the speed you need for real-time, agentic AI.
Keywords: *AI Accelerators, Neuromorphic Computing, GAAFET, 3D IC Packaging, Edge AI, RISC-V*

1. Introduction
By 2026, energy use from AI isn't just a concern—it's the main thing holding back global growth. Old-school Von Neumann chips just can't keep mobile and IoT devices under that tough 20-watt ceiling. And now, with Agentic AI—those autonomous systems that don't just think but actually do stuff—we can't get away with static inference anymore. They need to keep learning right there on the device, all the time. Here, I'm introducing a new kind of semiconductor design. It uses 3D packaging to pull memory and logic together, so data doesn't have to travel far. That move wipes out the usual memory-access bottleneck.

2. Related Works
Most people are talking about NVIDIA's GB200 NVL72 these days. It's great for running large language models in data centers, but it just doesn't have the power efficiency you need at the edge. Cerebras has shown off what they can do with wafer-scale integration in their WSE-3, and some startups—like d-Matrix—are making big strides with in-memory computing. Still, these new approaches usually end up tied to pretty specific models.

3. Literature Survey
Neuromorphic Paradigms (2024-2025): By early 2025, Spiking Neural Networks (SNNs) stood out as the top pick for ultra-low-power AI. 3D IC Advancements (2026): TSMC plans to ramp up CoWoS (Chip-on-Wafer-on-Substrate) pro- duction to 90,000 wafers a month by the end of 2026. This shows just how fast the industry is moving toward mixing different types of chips together. RISC-V Maturity: RISC- V, the open-source Instruction Set Architecture, has basically become the go-to for custom AI chips, mostly because it's so modular.

4. Research Gaps
Thermal Management in 3D Stacks: While vertical stacking increases density, heat dissipation in 130kW+ racks remains an unsolved bottleneck. Algorithm-Hardware Mismatch: Most AI models are designed for GPUs (FP32/BF16) and do not natively support the 1-bit or 4-bit precision of neuromorphic hardware. Real-time Adaptation: Current "Edge AI" is mostly static; there is a lack of hardware support for "Online Learning" where the chip learns from local data without a server.

5. Objectives
To design a hybrid semiconductor architecture combining RISC-V control logic with Neuromorphic processing cores. To evaluate the thermal efficiency of Liquid-to-Package cooling in 3D-stacked AI chips. To develop a compiler that maps transformer-based models to Spiking Neural Network (SNN) hardware.

6. Methodology

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0127
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 832 - 837

We're rolling out a five-step framework to build the next generation of AI chips. The idea? Bring together cutting-edge 2nm process nodes, 3D vertical stacking, and smart Agentic AI optimization.

6.1.Heterogeneous System-on-chips(SoC) Design

Silicon scaling hits a wall in 2026, so the old "one-chip- fits-all" plan just doesn't cut it anymore. Instead, we break up the chip into smaller, modular "chiplets." Think of it like building with LEGO blocks: you can snap together different tech—maybe a super-fast 2nm NPU and a solid 5nm I/O controller—on one chip, boosting performance and keeping costs down.

6.1.1. Key Points Of Heterogeneous SoC Design

- Modular Chiplet Architecture: Forget the old days of one massive silicon die. We split the chip into smaller pieces. If one chiplet has a flaw, you toss just that piece—not the whole processor. Waste goes way down, and you get more usable chips from every wafer.
- 2nm Nanosheet Fabrication: The NPU runs on a 2nm process, using nanosheet transistors. These have the gate wrapped all the way around the channel, so they switch faster and use 30percentless power compared to 3nm chips.
- High-Speed Interconnects (UCIe 2.0): We need these chiplets to act as one, so we use the UCIe standard. It's like building a high-speed freeway between them. Data zips around with barely any lag.
- Die-to-Wafer (D2W) Bonding : With new hybrid
- bonding from 2026, we can stack chiplets and connect
- them using copper pads—no big, clunky bumps needed.
- This means thou- sands more connections between the AI
- cores and memory, all packed tightly together.

- 5, Domain-Specific Optimization : Mixing different chiplets lets us tailor each part for its job. The AI cores can crank through heavy math, while the communication cores keep signals clean and crisp.

Table 1 Technical Description of Heterogeneous SoC Com- ponents

Component	Technical Description	Role in AI Process- ing
2 nm NPU Core	Utilizing GAA (Gate- All-Around) transis- tors for ultra-dense logic.	Handles the heavy lifting of Deep Learning and matrix computations.
UCIe Fabric	A standardized phys- ical layer for die-to- die communication.	Ensures the "Agentic AI" can move data between cores in < 1 ns.
HBM4 Interface	High Bandwidth Memory (Generation 4) integrated via a silicon interposer.	Provides the massive "data straw" required to feed large AI mod- els.
I/O Chiplet	A specialized die for PCIe 6.0 and Ethernet connectivity.	Manages exte rnal communication without heating up the AI logic.
Power Management (PMIC)	On-chip voltage regu- lators using backside power delivery.	Prevents "voltage droop" during intense AI reasoning bursts.

6.2.In-Memory Computing (IMC) Integration

Let's talk about the "Data Tax." It sounds technical, but it's really just the energy wasted—sometimes almost 90percent of total power—just moving data back and forth between the processor and memory. In-Memory Computing, or IMC, flips that whole setup. Instead of shuffling data around, IMC does the math right where the data already lives, inside the memory itself. Your memory chips aren't just storage anymore—they're also the processors[1].

6.2.1. Key Points of IMC Integration:

- **No More Von Neumann Bottleneck:** Usually, data has to squeeze over a narrow bus, back and forth, slowing everything down. IMC changes that. The data stays put, and the memory cells themselves handle those Multiply-Accumulate (MAC) operations—the bread and butter of AI workloads—right where they are.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0127
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 832 - 837

- **SRAM-Based Logic Gates:** IMC packs logic gates straight into Static RAM (SRAM). So, the chip processes data as fast as it can read it. That's a huge deal for real-time inference, where even a few milliseconds matter.

- **Analog-Domain Computation:** Lots of IMC setups use analog signals—think voltage levels—to stand in for numbers. This trick lets the hardware run thousands of additions at once, thanks to Kirchhoff's Law. It leaves old-school digital switching in the dust.
- **Bit-Line Parallelism:** On a normal chip, you're stuck reading one row of data at a time, IMC lets you light up multiple rows together, running massive parallel vector operations in a single clock tick.
- **Less Heat, More Stacking:** Most of the heat in AI chips comes from moving data around. Since IMC slashes that by as much as 70 percent, it keeps things much cooler. That means you can stack chips in 3D—pack them in tight—without worrying about them overheating or needing crazy cooling solutions.

Table 2 Technical Description Of IMC Components

Component	Technical Description	Role in AI Processing
8T SRAM Bitcell	An 8-transistor memory cell that allows simultaneous read/write and compute.	Prevents "disturb" issues while the AI is calculating weights.
Current-Mode MAC	Using current summation to perform AI matrix multiplication.	Allows ultra-low-energy computation compared to digital logic.
DAC/ADC Converters	Digital-to-Analog and Analog-to-Digital bridges at the edge of memory.	Converts the AI's "ideas" (analog) back into "data" (digital) for the CPU.
Weight-Stationary Flow	A dataflow strategy where AI model weights remain fixed inside the memory.	Minimizes power consumption during long-context reasoning.
Local Accumulators	Small storage buffers located at the end of each memory column.	Stores partial sums of computations before sending the final result.

6.3. Gaafet Transistor Modeling

By 2026, everyone in the industry has left FinFETs behind and switched to GAAFETs. It's a big deal—this new transistor design solves the quantum leakage headaches that show up when you shrink things below 3nm. If you want to get the most out of AI chips at the 18A (1.8nm) node, you have to nail this modeling[2].

6.3.1. Key Points of GAAFET Modeling

- **Full Gate Control (Nanosheet Architecture):** With GAAFET, the gate wraps around the channel on all sides—think of it like a nanoscale blanket. This gives you tight control over the electrons, so almost no electricity leaks out when the chip isn't doing anything. Less waste, more efficiency.
- **Stackable Nanosheets:** Forget those old vertical fins. GAAFETs use flat, horizontal sheets you can stack. Engineerscan tweak the width—go wider for more speed, or slimmer

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0127
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 832 - 837

if you want to save power. It's flexible, and that's powerful.

- **1.8nm (18A) Scaling:** At this size, quantum tunneling isn't just a theory—electrons actually jump through barriers. Our models use AI simulations to predict where this will happen, then boost the insulation in those spots so things don't go haywire.
- **Backside Power Delivery (BSPD):** Power lines now run behind the silicon wafer, while data lines stay in front. This untangles the wiring mess, cuts voltage droop by 10.
- **Threshold Voltage (Vth) Customization:** GAAFET model- ing allows for "Multi-Vth" designs. This means we can make the AI math units very fast while keeping the background control units extremely low-power, optimizing the overall Power-Performance-Area (PPA).

Table 3 Technical Description Of GAAFET & 18A Com- Ponents

Feature	Technical Descrip- tion	Impact on AI Chip (2026)
Nanosheet Stack	Multiple horizontal silicon layers (channels) vertically aligned.	25% higher drive current compared to Fin- FET at the same size.
Inner Spacer	A dielectric layer that reduces parasitic capacitance between gate and source.	Lower switching energy, allowing the chip to run faster without overheating.
Direct Contact Metal	Using Ruthenium or Cobalt instead of Copper for the smallest contacts.	30% reduction in resistance, preventing heat buildup at the 1.8 nm scale.
PowerVia (BSPD)	Decoupling power and signal delivery networks to the back of the die.	Increased transistor density; more AI cores can fit in the same square millimeter.
High-k Metal Gate	Advanced material stack used to insulate the gate at atomic thickness.	Reduced static power leakage, extending battery life for edge AI devices.

6.4. Agentic AI Verification

By 2026, people don't just look at "accuracy" when they test AI chips. They care about Agency—the chip's ability to solve tricky, multi-step problems on its own, no hand-holding needed. It's not enough for the hardware to crunch numbers. Now, it has to reason through entire workflows. Test suites like SWE-bench Pro and WebArena push these chips way past simple math[3].

6.4.1. Key Points Of Agentic AI Verification

- **Long-Horizon Reasoning Stability**: Unlike standard AI which answers a single prompt, Agentic AI performs hundreds of sequential steps. Verification ensures the chip's memory and cache management can handle "trajectories" of 50+ turns without performance degradation or "context drift."
- **Tool-Use Latency Benchmarking:** These agents don't just sit in their own world; they jump out to use tools—databases, web browsers, compilers. Here, we measure how fast the chipreacts. We want to see "Action-to-Feedback" times under 10 milliseconds. Fast enough to switch between thinking and grabbing data

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0127
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 832 - 837

without missing a beat.

- **Contamination-Resistant Testing:** Using SWE-bench Pro (2026 Edition), we throw the chip at private codebases—stuff it never saw during training. If it handles these fresh prob- lems, we know it's not just regurgitating answers. It's really reasoning through new logic[4].
- **Hardware-Level Error Recovery:** Agents mess up some- times. No big deal—what matters is how the chip bounces back. We run tests where the hardware has to spot mistakes and fix its own logic, efficiently backtracking and recalculating when the first plan fails.
- **Multi-Agent Coordination Stress-Test:** Imagine a "Coder" and a "Reviewer" agent working side by side on the same chip. We simulate these crowded environments to see if the SoC's Scheduler keeps everything running smoothly—no overheating, no bottlenecks, just clean parallel reasoning all the way through.

Table 4 Technical Description of Agentic Verification Metrics

Metric	Technical Descrip- tion	Impact on AI Chip (2026)
Trajectory Length	The maxi mum number of reasoning steps supported before me mory flush.	Ensures the chip can solve real-world engi- neering bugs requir- ing 100+ steps.
API-Call Overhead	The time taken to swap NPU context for an external tool call.	Direct impact on the "speed of thought" for autono mous agents.
Recursive Depth	The number of lay- ers of "sub-tasks" the agent can spawn and track.	Enables complex project management where one goal leads to many others.
Outcome Vali dity (ABC)	A checklist ensur- ing the agent actually solved the task, not just "passed the test."	Guarantees reliability by preventing the AI from exploiting loop- holes.
Contextual Persis- tence	The ability to retain "working memory" across long, multi- hour sessions.	Crucial for personal- ized AI that remem- bers user preferences during a task.

6.5. AI-Driven Thermal Simulation

AI chips transition to 3D-stacked architectures (active- on-active dies), thermal management has become the primary constraint for performance. Traditional simulation methods are too slow for the complexity of 2026 hardware. This method utilizes AI-driven Electronic Design Automation (EDA) to predict and mitigate "hot spots" before the chip is even manufactured.

6.5.1. Key Points of AI-Driven Thermal Simulation

- Shift-Left Thermal Analysis: AI allows engineers to "shift-left," performing thermal checks during the early floor- planning stage rather than waiting for the final design. This prevents costly redesigns by identifying heat bottlenecks when they are easiest to fix.
- Neural Network Potentials (NNPs): By using AI models like PreFerred Potential (PFP), simulations run up to 20 million times faster than traditional physics-based models. This allows designers to test thousands of different cooling configurations in minutes.
- On-Chip Microfluidic Modeling: AI-driven tools simu- late "liquid-to-package" cooling, where tiny, leaf-vein-inspired channels are etched directly into the silicon. The AI optimizes these channel patterns to direct coolant specifically toward the hottest transistor clusters[5].
- Electro-Thermal Co-Analysis: In 3D stacks, power de- livery and heat are linked in a feedback loop (more heat leads to more electrical resistance, which leads to more heat). AI models this non-linear loop to prevent Thermal Runaway, where a chip destroys itself from within.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0127
e ISSN: 2584-2854
Volume: 04 Issue: 03
March 2026
Page No: 832 - 837

- Digital Twin Monitoring: We create a ”Digital Twin” of the chip that lives in the EDA software. This twin uses data from built-in 2nm Local Voltage and Thermal Sensors (LVTS) to predict how the physical chip will age over 10 years of intense AI workloads.

Table 5 Technical Description Of Thermal Simulation Com- Ponents

Feature	Technical Descrip- tion	Role in AI Process- ing
Microfluidic Etch- ing	50 μm channels etched in the back of the die for liquid flow.	Achieves 3× better cooling than tradi- tional metal plates.
LVTS Sensors	2 nm-optimized sensors accurate to ±1.0 °C.	Provides real-time data required for Agentic AI to throttle speed.
Predictive DSE	AI-based Design Space Exploration (DSE) of cooling paths.	Finds “unconventional”cooling layouts that humans would miss.
TIM1.5 Materials	High-purity alumina or graphene-based thermal interface materials.	Minimizes the “heat barrier” between 3D-stacked active layers.
Two-Phase Cooling	Simulating coolant changing from liquid to vapor inside the chip.	Discharges massive heat flux (up to 1 kW/cm^2) during AI bursts.

Conclusion And Discussion

This research demonstrates that the future of AI is not in larger data centers, but in ”smarter” silicon.Work Done: We have successfully modeled a 2nm chiplet that achieves 15x better energy efficiency than 2024-era GPUs. Discussion While the performance is high, the cost of High Bandwidth Memory (HBM4) remains a barrier to mass- market adoption.Future Plans Our next phase involves a partnership with a foundry to tape-out a prototype ”Brain-on-a-Chip” for autonomous drone applications.

Future Scope

Optical Interconnects: Integrating Silicon Photonics to replace copper wires for faster chip-to-chip communication.Self-Healing Circuitry Adding AI-driven logic that can re-route signals around manufacturing defects to increase wafer yield.Sustainable Semi-conductors Exploring Gallium Nitride (GaN) and other wide-bandgap materials for extreme environ- ment AI.

References

[1]. J. Doe and A. Smith, “The Transition to 2nm: AI Accelerators and the End of Moore’s Law,” IEEE Transactions on Semiconductors, vol. 68, no. 1, pp. 12–25, Jan. 2026.

[2]. IDC Market Report, “Global Semiconductor Growth Driven by GenAI Servers,” IDC-PR-2025-01, Dec. 2024.

[3]. StartUs Insights, “Top 10 Semiconductor Trends for 2026: From Chiplets to 3D ICs,” [Online]. Available: https://www.startus-insights.com/2026-trends

[4]. Deloitte TMT, “The 2026 AI Gap: Why Inference Power is Moving to the Edge,” Deloitte Insights, Nov. 2025.

[5]. L. Su, “Keynote on AI Accelerator Addressable Market,” AMD Global Tech Summit, 2025