Executive summary
On the basis of brain: neural-network-inspired changes in general-purpose chips

Ekaterina Prytkova, PhD

Full article is published in the Industrial and Corporate Change Journal.
Available at https://doi.org/10.1093/icc/dtab077

Overview This article explores how the rise of artificial intelligence (AI) is reshaping the semiconductor industry. Sequential von Neumann-based architectures are struggling to meet the demands of modern AI applications, leading to new challenges in chip design that emphasize computational flexibility in addition to raw speed. The authors examine the forces at play in the industry and propose two possible scenarios for semiconductor development: fragmentation into specialized submarkets or the emergence of a platform chip with modular architecture.

Premise: The Challenge of AI for Semiconductor Design

- AI algorithms require parallel processing capabilities that are inefficiently handled by conventional CPUs.
- The demand for AI-optimized chips has introduced significant technological discontinuities in the industry.
- Traditional chip design focused on computational speed is being replaced by a focus on computational flexibility, energy efficiency, and adaptability to varied software ecosystems.

Key Proposition: The Trilateral Technological Frontier

- The authors identify three core properties shaping the semiconductor industry's future: flexibility, energy efficiency, and speed.
- For decades, the industry prioritized improving processing speed, leading to consistent advancements that obscured the trade-offs with flexibility and energy efficiency. The continual push for faster chips made these compromises less apparent.
- However, the rapid expansion of AI applications, combined with the physical limitations of chip miniaturization and shifting economic incentives, has brought these trade-offs to the forefront. The semiconductor industry now faces the challenge of balancing all three performance aspects rather than optimizing just one.

Industry at a Crossroad: Two Scenarios

Scenario I: Fragmentation into Specialized Chips

- Economic pressures and market-specific demands could drive chip manufacturers toward highly customized chips tailored for specific applications (e.g., AI accelerators, GPUs, TPUs).

- This scenario would lead to a fragmented industry with diverse, application-specific architectures rather than a dominant computing model.
- Higher performance gains from specialized chips, coupled with diverging business models and strategies, reinforce the fragmentation trend. These factors play an important role especially for cloud computing companies like Google and Amazon. Due to high workloads supporting simultaneously millions of users, they can compromise nor on latency, nor on speed, nor on heterogeneity, having to build an immense compute infrastructure that comprises all kinds of chips inside to deliver **any** combination of compute services.

Scenario II: Emergence of a Platform Chip

- Companies may invest in a new heterogeneous, modular architecture that integrates multiple computing paradigms onto a single, platform chip. Such a platform chip would aim to accommodate a broad range of applications, optimizing dynamically computational power, efficiency, and flexibility.
- This model would resemble historical shifts in computing, such as IBM's System/360, which integrated different computing paradigms under a common architecture: a higher-level architectural innovation that emerges to unify previously separate designs.
- A platform chip would allow cost efficiencies and economies of scale in manufacturing. As no single specialized chip can deliver to all tasks because of the trade-offs in the trilateral frontier, a platform chip could integrate diverse computing elements to optimize performance dynamically based on demand. This would also favor better software compatibility by creating a common foundation for various software frameworks and providers. Lastly, platform chip would align with edge mode of compute delivery (e.g. mobile compute), reducing latency, enhancing processing efficiency and data security and privacy.

Implications for the Semiconductor Industry

- **Product Design and Market Survival**: Chipmakers must weigh short-term profits from specialization against the intertemporal, long-term advantages of a flexible platform chip, especially given the fast-changing nature of AI technologies.
- **Software and Hardware Co-evolution**: The semiconductor industry must consider the dynamics in the software ecosystems to ensure adaptability, scalability, and lasting market share by creating modular, flexible and open-ended hardware solutions.
- **Sustainability and Energy Efficiency**: Given the trend on "large" and hence high energy demands of AI, new chip designs must optimize computational efficiency to lower compute costs and mitigate environmental impacts.
- **Market Power Concentration**: specialization carries inherent risk of hard lock-ins in both AI (and in general software) and hardware markets presenting challenges for innovation in both AI and hardware and damaging competition in respective ecosystems.

Conclusion The semiconductor industry stands at a crucial turning point. The response to AI-driven demand could lead to either a more fragmented ecosystem of specialized chips or a new dominant computing architecture based on heterogeneous integration. While fragmentation offers immediate boost of computing power, a flexible platform chip may provide the most sustainable and commercially viable solution in the long run. The outcome will depend on economic pressures, technological breakthroughs, regulatory efforts, and strategic choices by industry players.