Per-Model ASIC Technology

A Paradigm Shift to Model-First Silicon Design

Enabling the Future of Edge AI

Dr. Steve Xu

Chief Architect & CEO

XgenSilicon

xgensilicon.ai

December 18, 2025

Contents

Executive Summary

This white paper presents XgenSilicon's revolutionary Per-Model ASIC technology, which fundamentally transforms the relationship between artificial intelligence software and silicon hardware. Unlike traditional chip design approaches that force AI models to adapt to pre-existing hardware, XgenSilicon's *"Model First → Custom ASIC"* paradigm generates custom Application-Specific Integrated Circuits (ASICs) directly from AI model specifications.

Key Findings:

- **30x Power Efficiency Improvement:** XgenSilicon's custom ASICs achieve 30 mW power consumption for complex Vision LLM workloads, compared to ∼1 W for conventional hardware—a 30x improvement.
- **3-Month Time-to-Market:** The fully automated, AI-driven design flow enables complete model-to-ASIC deployment in just 3 months, compared to the typical 18-month cycle for traditional chips.
- **Custom Non-Von Neumann Architecture:** Eliminates memory bottlenecks through separated data and instruction pathways, enabling massive parallelism for AI workloads.
- **Economic Viability:** Multi-Project Wafer (MPW) manufacturing strategy makes custom ASICs commercially practical even for non-megascale volumes.

Business Impact: This technology enables the deployment of sophisticated AI models on power-constrained edge devices, opening new markets for personal, on-device AI applications while preserving user privacy and delivering real-time intelligence without cloud dependency.

Introduction: A Paradigm Shift to "Model-First" Silicon Design

The proliferation of artificial intelligence is rapidly moving from the cloud to the edge, creating a significant challenge for hardware engineers. Deploying increasingly complex and powerful AI models on personal and embedded devices is fundamentally constrained by the performance, power, and cost limitations of general-purpose, off-the-shelf silicon. The conventional approach to chip design is no longer sufficient for the bespoke demands of next-generation AI.

XgenSilicon introduces a fundamental shift in this landscape with its **"Model First → Custom ASIC"** paradigm. This approach inverts the traditional *"Chip First → Model on Chip"* development cycle, where software models are compromised to fit pre-existing hardware. Instead, XgenSilicon begins with the AI model itself, treating it as the primary design specification. From this model, a custom Application-Specific Integrated Circuit (ASIC) is automatically generated, creating a hardware architecture perfectly optimized for the model's specific workload.

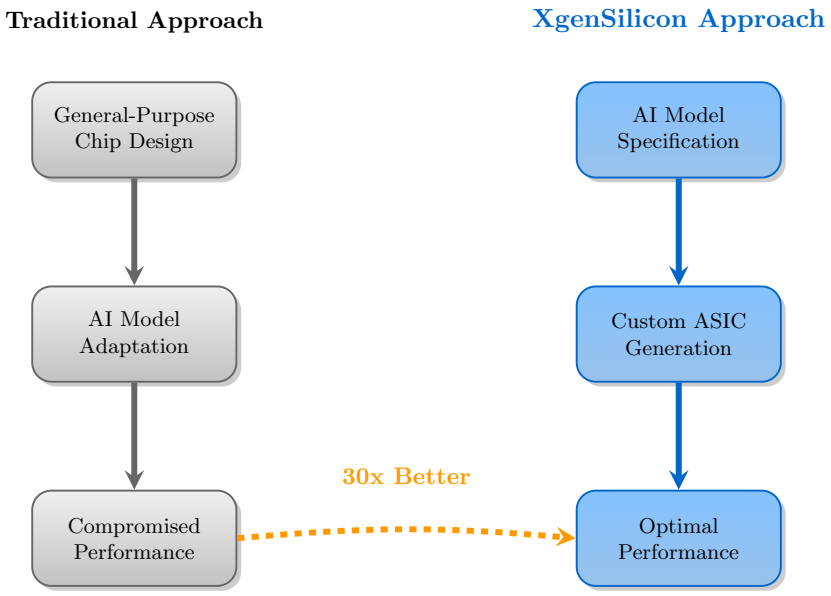

Figure 1: Paradigm Shift: From Chip-First to Model-First Design

The strategic implications of this shift are profound. The old business model, reliant on power-hungry, off-the-shelf components, forces developers to deploy less powerful AI models. In contrast, the **"Model First"** approach yields highly energy-efficient silicon, enabling the deployment of more sophisticated models and paving the way for truly personal, on-device AI. This white paper provides a comprehensive technical deep-dive into the architecture, design flow, and performance benefits of XgenSilicon's AI-generated, Per-Model ASIC technology.

Problem Statement: The Inherent Limitations of Off-the-Shelf Edge AI Hardware

To appreciate the innovation of Per-Model ASICs, it is crucial to first understand the technical and economic bottlenecks of conventional hardware solutions. Standard components like mobile Systems-on-Chip (SoCs) or microcontrollers (MCUs) with integrated Neural Processing Units (NPUs) are designed for a broad range of applications, resulting in architectural compromises that hinder performance and efficiency for specific AI tasks.

The following table synthesizes a comparison between these conventional solutions and XgenSilicon's specialized approach, highlighting the stark differences in key performance metrics.

Evaluating the impact of these limitations reveals a clear pattern: conventional chips are *"power hungry"* and force the deployment of *"less powerful models."* This inefficiency directly hinders the progress of personal, on-device AI, where battery life and computational power are paramount. The following sections will explore the specific technical solutions XgenSilicon has developed to overcome these fundamental challenges.

Table 1: **Performance Comparison: Conventional Hardware vs. XgenSilicon Per-Model ASIC**

Metric	Conventional Hardware	XgenSilicon Per-Model ASIC
	(e.g., Mobile SoC, MCU with NPU)	
Power Efficiency	Baseline (1x)	**30x Improvement**
Power Consumption	∼1 W (for a Vision LLM)	**30 mW (for a Vision LLM)**
Architectural Bottlenecks	Baseline Von-Neumann architecture with inherent memory bottlenecks.	**Custom, bottleneck-free Non-Von Neumann architecture.**
Time-to-Market	Typical 18 months for a chip refresh cycle.	**3 months from model release to ASIC.**

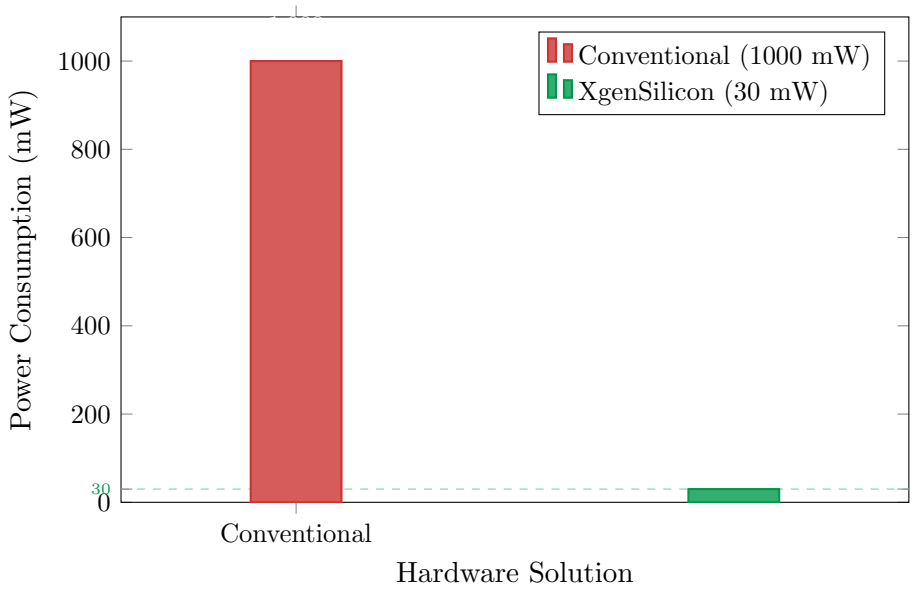

Figure 2: Power Consumption Comparison: 30x Improvement with XgenSilicon

Solution: XgenSilicon's Core Technology—A Two-Fold Innovation

XgenSilicon's performance gains are not the result of a single breakthrough but a synthesis of two core pillars: a novel hardware architecture designed explicitly for AI workloads and a revolutionary AI-driven design process that automates silicon creation. These two components work in tandem, creating a powerful synergy that redefines the possibilities for edge computing.

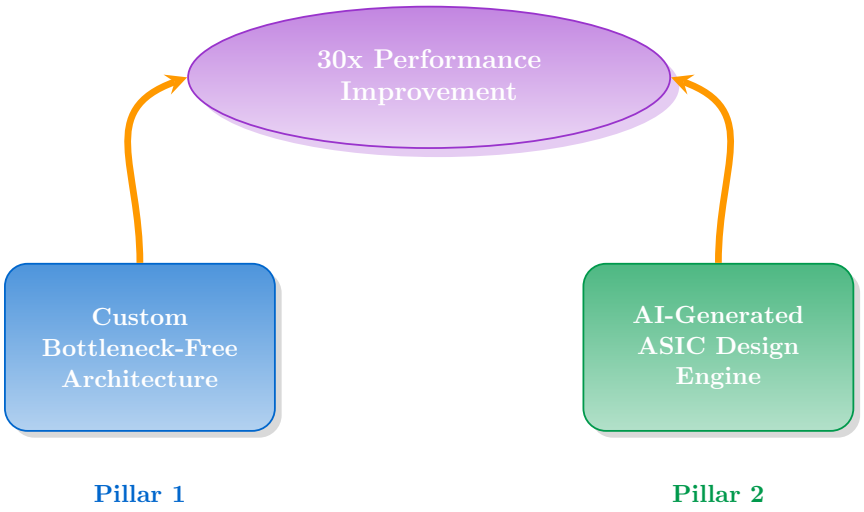

Figure 3: Two-Fold Innovation: Architecture + AI Design Engine

The Foundation: A Custom, Bottleneck-Free Technical Architecture

The physical foundation of a Per-Model ASIC is a ground-up rethinking of chip architecture, moving away from legacy designs to a layout optimized for the mathematical demands of neural networks.

- **Custom Bottleneck-Free Architecture:** Unlike general-purpose layouts that must accommodate diverse tasks, each XgenSilicon ASIC is custom-built for a single AI model. This eliminates unnecessary components and optimizes data pathways, removing the performance bottlenecks that plague conventional chips.
- **Non-Von Neumann Architecture:** Traditional Von-Neumann architectures, which use a shared bus for data and instructions, create a well-known memory bottleneck. By adopting a Non-Von Neumann approach, XgenSilicon separates these pathways, allowing for massive data parallelism and eliminating the performance drag that is particularly acute in data-intensive AI models.

- **Dataflow-based Superscalar Out-of-Order RISC-V Core:** The computational heart of the ASIC is a custom RISC-V core built on a dataflow principle. A dataflow architecture executes operations as soon as their required data is available, a natural fit for the parallel graph-like computations in neural networks. The addition of superscalar, out-of-order execution capabilities allows the core to process multiple instructions simultaneously and in the most efficient sequence, maximizing instruction-level parallelism and ensuring peak utilization of compute resources.

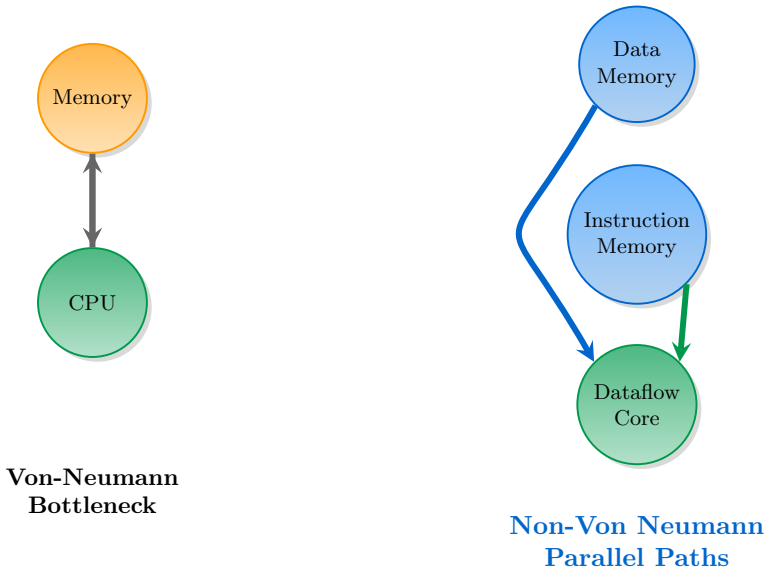

Figure 4: Architecture Comparison: Eliminating the Memory Bottleneck

The Engine: AI-Generated ASIC and Unified Reinforcement Learning

The second pillar of the technology is the engine that builds the custom hardware. XgenSilicon employs a fully automated, AI-driven process that translates a software model directly into GDSII ready for manufacturing.

The core of this engine is a **"Unified Reinforcement Learning for ASIC Generation"** system. Chip design involves navigating a vast, multi-dimensional space of possible configurations to find the optimal balance of Power, Performance, and Area (PPA). The RL agent intelligently explores this design space. Its intelligence is guided by a precise reward function derived from a comprehensive, full-stack simulation of the entire system. This allows the agent to accurately predict the PPA impact of every decision, from high-level architecture to low-level physical layout and precise allocation of LLM inference workload. This method of joint software-hardware co-design enables the system to discover a **"global maximum"** of efficiency, far exceeding the *"local maximum"* achievable when fitting a custom model onto off-the-shelf hardware.

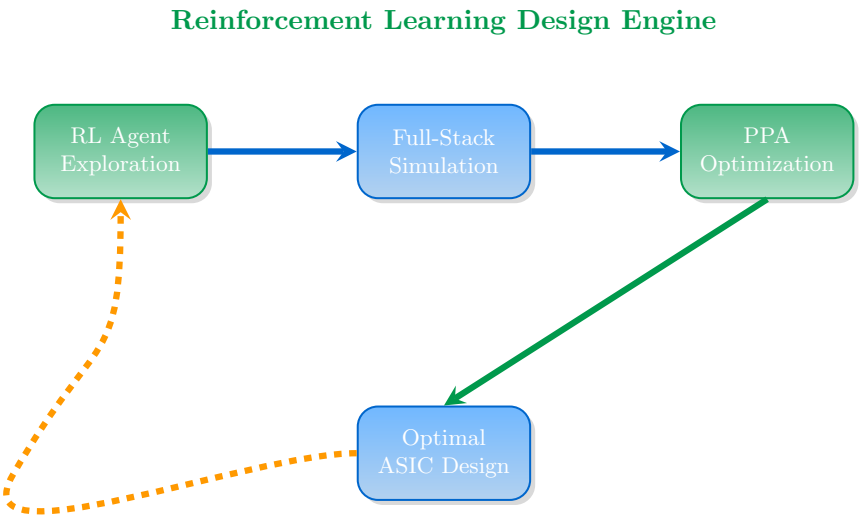

Figure 5: RL-Driven ASIC Generation: Continuous Optimization Loop

This combination of a bespoke hardware foundation and an intelligent, automated design process is what enables the creation of a truly model-optimized ASIC.

The End-to-End RL-Driven Compilation Flow

The practical realization of a Per-Model ASIC depends on a sophisticated and highly automated end-to-end compilation and design flow. This process transforms a high-level AI model into a manufacturable silicon layout with minimal human intervention, embedding the principle of software-hardware co-design at every stage.

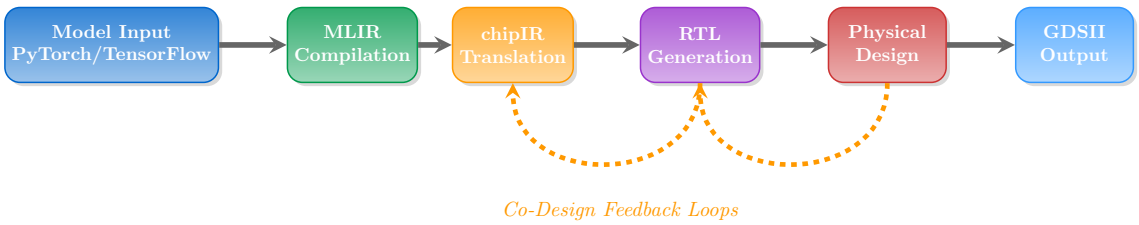

Figure 6: End-to-End Automated Compilation Flow

1. **Model Input:** The process begins with a standard AI/ML model, such as one developed in PyTorch or TensorFlow, along with its specific application constraints (e.g., target latency, power budget).

2. **MLIR Compilation:** The model is first compiled using Multi-Level Intermediate Representation (MLIR). This industry-standard framework deconstructs the model into its core

components, including operators, control flow, scheduling, and memory binding, creating a structured representation that is ready for hardware mapping.

3. **XgenSilicon Data-Flow Architecture (chipIR):** The MLIR output is translated into chipIR, XgenSilicon's proprietary intermediate representation. This chipIR acts as an *"ASIC design foundation library,"* augmenting the logical model description with the physical attributes required for silicon design.

4. **AI-Assisted RTL Design:** The chipIR is fed into generative Register-Transfer Level (RTL) tools. Here, the RL agent guides the process, generating optimal hardware partitions and logic. This stage initiates a critical feedback loop, as RTL generation choices influence potential software optimizations, embodying the software-hardware co-design principle.

5. **AI-Assisted Back-End Design:** The verified RTL proceeds to the physical design stage. Using industry standard EDA tools, the RL agent continues to optimize the physical layout (place & route). This co-design feedback loop persists, as physical implementation details can inform further refinement of the chipIR and even the MLIR representation.

6. **ASIC Output:** The fully automated flow concludes with the generation of a GDSII for tape-out.

This deeply integrated workflow, driven by a continuous co-design feedback loop, enables the joint optimization of both the AI model's structure and the custom chip's architecture simultaneously. It is this holistic approach that achieves a level of performance unattainable with siloed design processes.

Case Study: Real-Time Speech Translator ASIC

To demonstrate the real-world capabilities of its technology, XgenSilicon has developed a Real-Time Speech Translator ASIC, its first commercial product. This application showcases the ability to run a complex, multi-model pipeline on a single, ultra-low-power chip, a task that would be challenging or impossible on conventional edge hardware.

Technical Specifications & Performance

The key metrics of the Speech Translator ASIC validate the performance claims of the Per-Model approach:

These results signify a major step forward for complex AI applications on power-constrained devices. The ability to integrate multiple, high-performance models onto a tiny, power-sipping

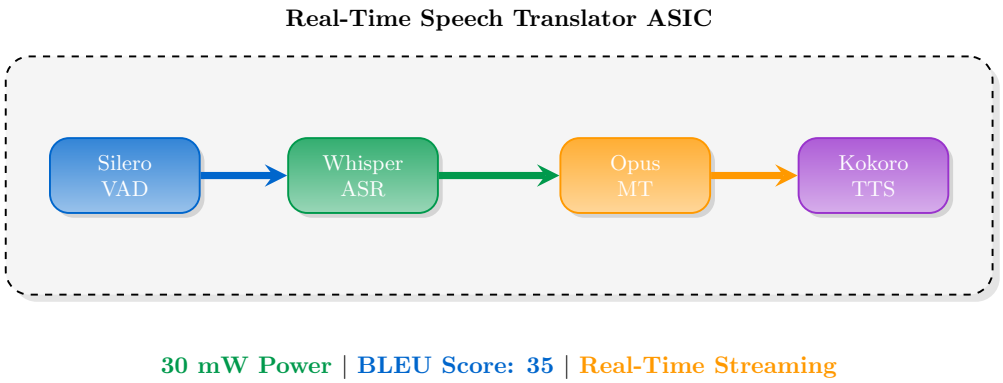

Figure 7: Speech Translator ASIC: Multi-Model Pipeline Architecture

Table 2: **Speech Translator ASIC: Key Performance Metrics**

Metric	Specification
Commercial Status	Signed Letter of Intent (LOI) with Bluetooth audio device manufacturer
Model Pipeline	Silero.Vad → Whisper (ASR) → Opus (MT) + Kokoro (TTS)
Power Consumption	**30 mW** during active inferencing
Translation Quality	BLEU score of ∼35 with real-time streaming capability

chip opens the door for a new class of intelligent personal devices. Notably, the 30 mW power consumption for this complex audio pipeline is on par with the efficiency achieved for single Vision LLM models, reinforcing the consistent and profound value of the custom architecture. This specific product example also demonstrates a viable path to market, which is underpinned by an equally innovative manufacturing and deployment model.

Achieving Economic Viability for Custom ASICs

The primary historical barrier to the widespread adoption of custom ASICs has been prohibitive cost, particularly the non-recurring engineering (NRE) costs associated with mask sets, which made small-volume production economically infeasible. XgenSilicon's business and manufacturing model directly addresses this challenge, making Per-Model ASICs commercially practical even for non-megascale volumes.

Manufacturing Strategy: Multi-Project Wafers (MPW)

The key to cost reduction is the use of Multi-Project Wafers (MPW). This manufacturing process dramatically lowers the financial barrier to entry by aggregating multiple distinct ASIC designs onto a single mask set and silicon wafer.

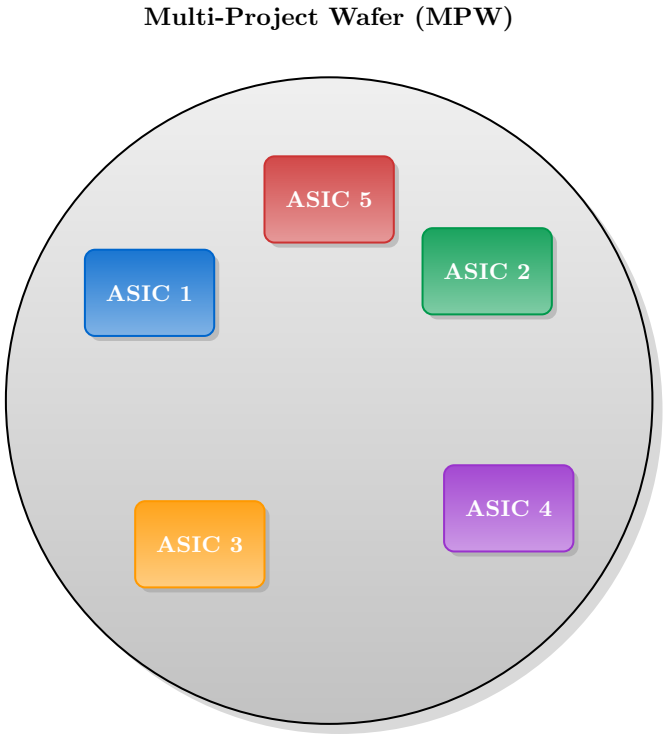

Figure 8: Multi-Project Wafer: Cost Sharing Strategy

By *"aggregating multiple ASICs (models) in one tape-out,"* the massive NRC (mask cost) is shared across many projects. This strategic pooling of resources *"decimates"* the upfront cost for any single design. The direct result is that the final cost per ASIC becomes *"comparable to high-volume off-the-shelf"* components, effectively neutralizing the traditional cost advantage of general-purpose chips.

The 3-Month ASIC Deployment Cycle

Complementing the cost-effective manufacturing is an unprecedentedly rapid deployment cycle. The high degree of automation in the AI-driven design flow enables a complete model-to-ASIC turnaround in just three months.

The deployment timeline is as follows:

- **Model-to-ASIC Compilation:** 1 month
- **Mask Making:** 0.5 month
- **Wafer Manufacturing:** 1 month
- **Test & Packaging:** 0.5 month

This 3-month total cycle stands in stark contrast to the typical 18-month refresh cycle for traditional chips. This dramatic acceleration in time-to-market allows product developers to

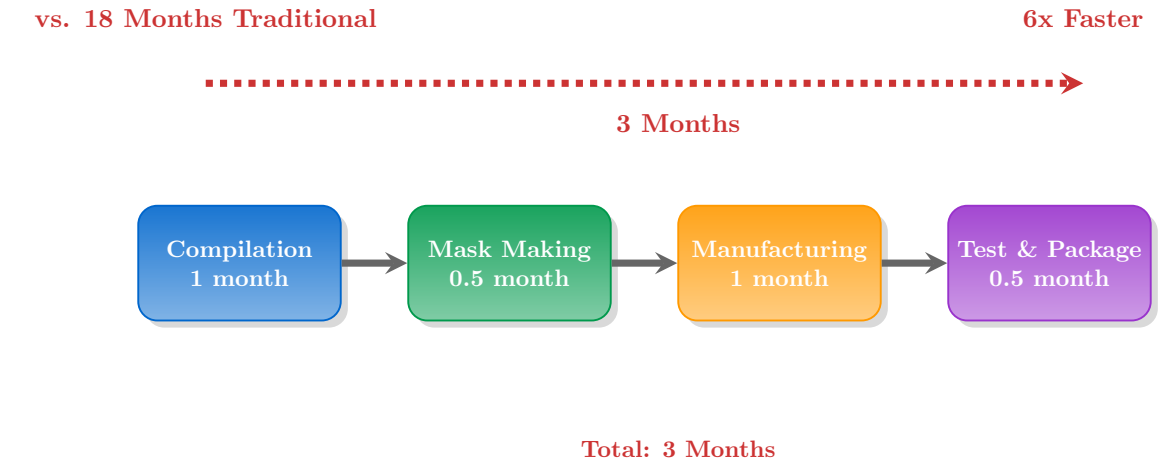

Figure 9: 3-Month Deployment Cycle: From Model to ASIC

innovate at the speed of software, incorporating the very latest AI models into their hardware without a crippling year-and-a-half lag.

The combination of MPW manufacturing and an automated design flow makes Per-Model ASICs not just technically superior but also commercially practical and agile.

Conclusion and Next Steps

The era of one-size-fits-all silicon is drawing to a close, especially in the demanding field of artificial intelligence. To unlock the next wave of innovation, the industry must transition from a general-purpose, *"chip-first"* world to a specialized, **"model-first"** ecosystem where hardware is tailored to the specific needs of the algorithm.

Key Differentiators

Figure 10: XgenSilicon's Competitive Advantages

XgenSilicon's technology is at the forefront of this transformation. The key differentia-

tors—a custom Non-Von Neumann, dataflow architecture and a fully automated, Reinforcement Learning-driven design flow—deliver unparalleled gains. The results: a **30x improvement in power efficiency** and a reduction in time-to-market from **18 months** to just **3 months**, are fundamentally re-architecting the relationship between software and silicon.

By making custom ASICs both economically viable and rapidly deployable, this technology enables a future of powerful, efficient, and private **"Personal AI."** We can now envision a world where sophisticated AI runs locally on dedicated, custom-designed edge devices, preserving user privacy and delivering instantaneous intelligence without reliance on the cloud. The future of silicon is not just AI-assisted; it is **AI-generated**.

Next Steps

Organizations interested in exploring Per-Model ASIC technology for their AI workloads should consider:

- **Model Evaluation:** Assess current AI models for edge deployment opportunities and power constraints.
- **Performance Requirements:** Define target latency, power budget, and throughput requirements for specific applications.
- **Pilot Program:** Engage with XgenSilicon to develop a pilot ASIC for a representative workload to validate performance and cost benefits.
- **Integration Planning:** Plan for the 3-month development cycle and MPW manufacturing timeline.

About the Author

Dr. Steve Xu is the Chief Architect and CEO of XgenSilicon, a company specializing in advanced silicon solutions for machine learning applications. With over 27 years of experience in machine learning hardware architecture and design, Dr. Xu is a recognized expert in the field of AI accelerator design and custom ASIC development.

Dr. Xu holds a PhD in Electrical Engineering and Computer Science from Massachusetts Institute of Technology (MIT). His research and development work has focused on creating efficient, specialized hardware architectures for machine learning workloads, with particular expertise in dataflow architectures, reinforcement learning-driven design automation, and edge AI deployment. His contributions to the field have been instrumental in advancing the state-of-the-art in custom ASIC design for AI applications.

References

[1] R. Ganti and S. Xu, "Hardware-Aware Neural Network Compilation with Learned Optimization: A RISC-V Accelerator Approach," arXiv:2512.00031, 2025.

[2] "Why ASIC Design Makes Sense for LLM On-Device," *EE Times*, https://www.eetimes.com/why-asic-design-makes-sense-for-llm-on-device/, July, 2025.

XgenSilicon
xgensilicon.ai