WHITE PAPER

Building an AI Chip: Pre-Silicon Planning

Authors

Kamal Desai
Scott Knowlton
Sumit Vishwakarma

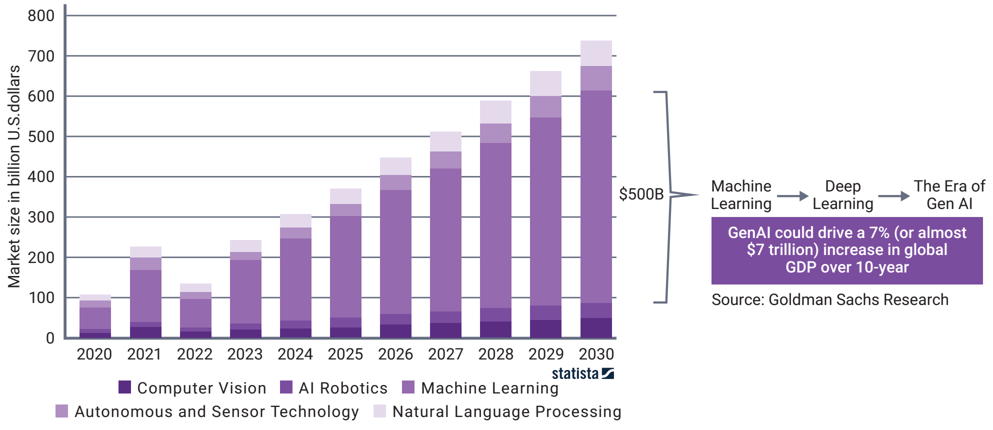
Overview

There are many applications that stress the limits of compute resources, from smartphones and self-driving cars to high-performance computing (HPC). In recent years artificial intelligence (AI) has threatened to eclipse or subsume all other application domains as the dominant consumer of computational power. Although AI algorithms can run on just about any processor, they run most efficiently on specialized engines that move some of the key algorithms from software to hardware. Designing an AI chip is a major undertaking requiring specialized knowledge, tools, and methodology at every stage of the process. This white paper discusses the first key stage: pre-silicon planning. It begins with an overview of the AI market to explain why it dominates the chip industry and to establish the motivation for the investment needed to design customized domain-specific chips.

Overview of the AI Market

Even a brief glance at recent venture capital investments or the market value of established companies shows the high value placed on AI. The potential for the AI market seems unlimited and there's a simple reason for this. In theory AI could be applied to any human endeavor, anytime and anywhere, and offers potential advantages in speed, cost, and quality. Just the applications for which AI is already deployed widely make for an impressive list: chatbots, assistants, investing, drug discovery, semi-autonomous vehicles, security systems, and much more. This all adds up to a business that's big and growing quickly. Figure 1 shows the actual and projected AI market size for the current decade.

Size of the artificial intelligence (AI) market worldwide



GenAI could drive a 7% (or almost $7 trillion) increase in global GDP over 10-year

Source: Goldman Sachs Research

Source(s): Statista; Statista Technology Market Insights , Worldwide; August 2023

Figure 1: A decade of growth in the AI market

The general domain of machine learning (ML) represents the bulk of the market, estimated to reach about half a trillion dollars by the end of the decade. Within this market segment a major evolution is underway. ML is leading to deep learning (DL), which in turn is ushering in the era of generative AI (GenAI). With GenAI, tools can produce novel text, code, images, and so on, based on the examples in its learning set. The availability of ChatGPT and competing platforms has enabled millions of non-experts to harness the power of AI and generate content of all sorts. One report from Goldman Sachs estimates that GenAI may drive a 7% increase in global GDP, almost 7 trillion dollars, over a ten-year period [1]. With these sorts of numbers in play there is strong motivation to develop customized AI chips for the general market or targeted at specific application domains.

AI Stresses Chip Design, Cost, and Power

AI increases chip design cost and power demands due to massive compute and memory needs. AI workloads, especially for training and inferences, drive larger, faster, and more power-hungry silicon architectures. Figure 2 illustrates three generations of AI workloads with increasing model sizes.

- AlexNet, a deep convolutional neural network (CNN) architecture developed in 2012
- BERT (Bidirectional Encoder Representations from Transformers), a natural language processing (NLP) model launched in 2018
- GPT-4 (Generative Pre-trained Transformer 4), a multimodal large language model released in 2023

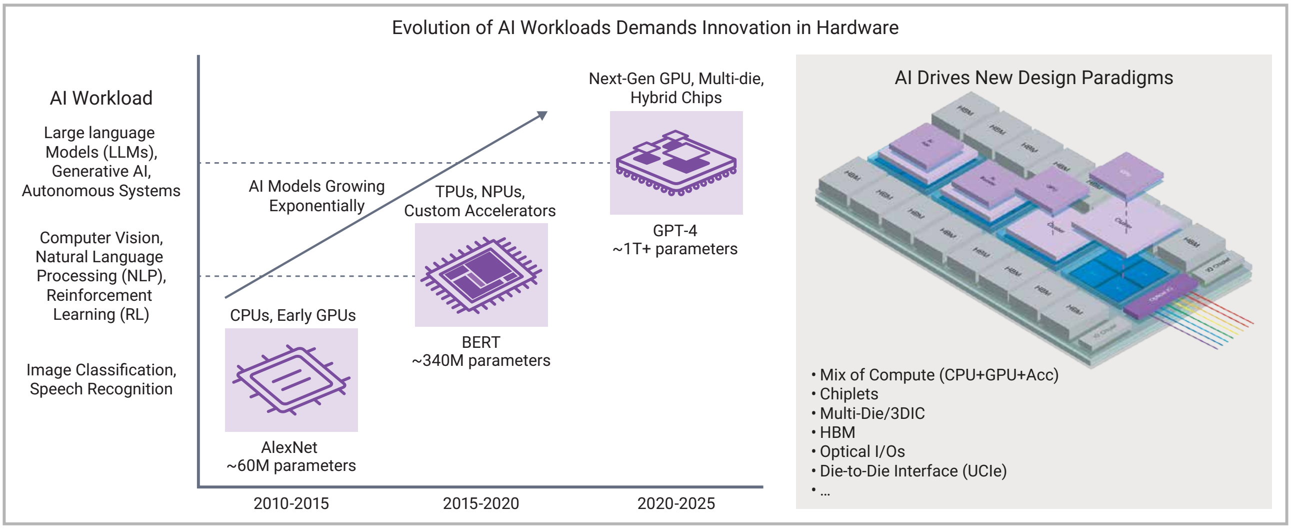

Evolution of AI Workloads Demands Innovation in Hardware
AI Workload
Large language Models (LLMs), Generative AI, Autonomous Systems
Computer Vision, Natural Language Processing (NLP), Reinforcement Learning (RL)
Image Classification, Speech Recognition
AI Models Growing Exponentially
CPUs, Early GPUs
AlexNet ~60M parameters
TPUs, NPUs, Custom Accelerators
BERT ~340M parameters
Next-Gen GPU, Multi-die, Hybrid Chips
GPT-4 ~1T+ parameters
2010-2015
2015-2020
2020-2025
AI Drives New Design Paradigms
HBM
• Mix of Compute (CPU+GPU+Acc)
• Chiplets
• Multi-Die/3DIC
• HBM
• Optical I/Os
• Die-to-Die Interface (UCIe)
• ...

Figure 2: Evolution of AI chips driven by model complexity

In Figure 2, the number of parameters for each generation of model is a metric for complexity and size. As a model adds more capacity, there are more parameters required to control its behavior. AlexNet had roughly 60 million parameters and it ran well on a traditional central processing unit (CPU) or early graphics processing unit (GPU). The parameter count increased to around 340 million for BERT and it required the power of a specialized chip such as a tensor processing unit (TPU), neural processing unit (NPU), or custom AI accelerator. GPT-4, the current state of the art, has over an astonishing 1 trillion parameters. It is trained on a leading-edge GPU cluster built with multi-die chiplets, high bandwidth memory (HBM), and die-to-die connections.

Adopting advanced technology nodes or 3DIC designs significantly increases development costs. A presentation by IBS at the Hot Chips 2023 conference reported that moving from 7nm to 5nm nearly doubles design cost, going from $297.8 million to $542.2 million [2]. The high computing requirements and the need to move massive amounts of data demands advanced nodes, and these cost more. Fortunately, as noted earlier, the promises of the AI market are enormous and so spending more money may very well lead to greater reward.

AI applications are notoriously power hungry so the growth in model complexity has dramatically increased power requirements as well as cost. Figure 3 illustrates this for several generations of AI models, leading up to GPT-4, and compares their power consumption with that of a typical home in the United States. Power and thermal considerations factor into the choice of chip technology and must be weighed against performance requirements. Together with the need to develop AI chips faster, this leads to a "shift-left" in the development process where power-performance tradeoffs and optimizations are made as early as possible.

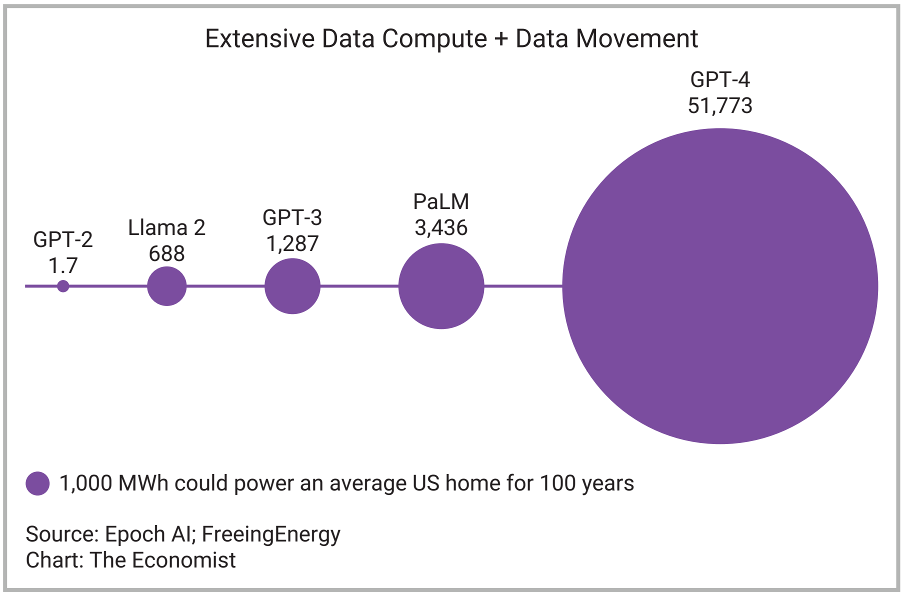


Figure 3: Evolution of AI power requirements

Challenges of AI Chip Design

AI chip developers encounter significant challenges in designing a system-on-chip (SoC) that not only meets the demanding requirements of AI applications but also balances project costs and the affordability of the end product. The competitive nature of the AI industry further amplifies the pressure to minimize time-to-market (TTM), making rapid and efficient development a critical priority.

A key benchmark for evaluating AI chips is their performance per watt, a metric that reflects the balance between computational efficiency and power consumption. Achieving improvements in this area demands innovation across at least four key areas:

- Optimizing data movement: Energy consumption is dominated by data transfer between memory and compute units, so designers must minimize memory access through efficient data reuse and on-chip storage
- Heterogeneous compute integration: Balancing general-purpose cores with specialized accelerators (such as matrix multipliers) requires careful partitioning and interconnect optimization to avoid bottlenecks
- Process technology scaling: Smaller nodes improve power efficiency but introduce thermal and leakage challenges, so advanced cooling and power gating techniques are crucial
- Algorithm-hardware co-design: Aligning hardware design with evolving AI model requirements (e.g., sparsity, quantization) ensures higher efficiency

Reducing design cost and TTM also entails four key forms of innovation driven by AI itself:

- Improving design productivity: Leverage AI-driven tools for faster verification, layout optimization, and automated design space exploration
- Shift left: Reduce iterations and minimize costly design revisions by using AI to predict errors early in the design cycle
- Optimizing design reuse: Develop modular architectures to reuse IP blocks across designs reducing overall development effort while focusing on design differentiation
- Improving collaboration: Facilitate global design teams with cloud-based tools, reducing delays and inefficiencies

The project's electronic design automation (EDA) tools, the flows that tie them together, and the methodologies that guide their use are key elements in achieving all these innovations. As noted above, many tools in the flow make use of AI themselves to automate some of the traditional manual design steps, reducing iterations, shortening TTM, and keeping budgets in check. Figure 4 provides an overview of the major steps in the AI chip development process and how they fit together into a flow. This white paper focuses on pre-silicon planning.

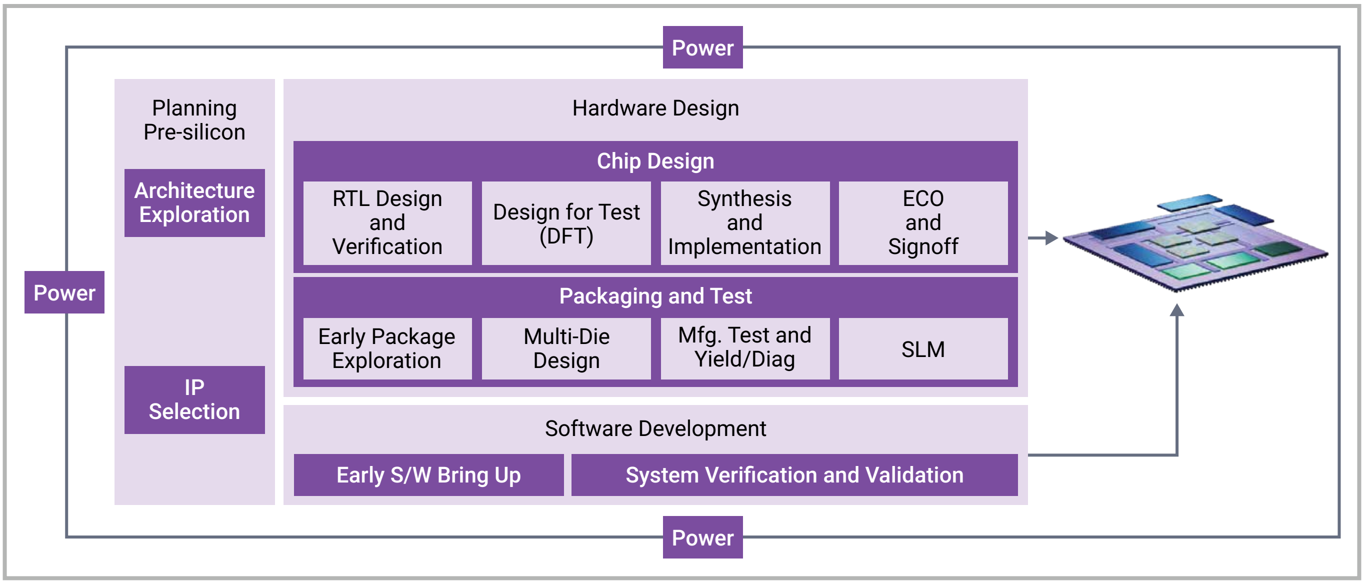


Figure 4: Development flow for a modern AI chip

Power is shown as an enveloping consideration in Figure 4, since power optimization occurs at many steps. One third to one half of this optimization occurs during architecture exploration, and the high-level designs made here have the most impact on power. Further optimizations made during the register transfer level (RTL) design, synthesis and implementation, engineering change order (ECO), and signoff steps are also important, but have less effect on total power consumption and the power-performance tradeoffs.

The Synopsys Solution: Architecture Exploration

AI chip designers use architecture exploration during pre-silicon planning as a way to shift left their development process and answer key questions much earlier in the project schedule. Typically, they would like to achieve the following goals:

- Avoid unpredictable performance and power due to dynamic effects
- Measure memory utilization and latency and bus/network-on-chip (NoC) arbitration
- Perform parallel task scheduling on heterogeneous processors
- Perform joint optimization of application workloads and hw architecture
- Perform simultaneous optimization of power, performance, accuracy, and cost

The industry's premiere solution for architecture exploration during pre-silicon planning is Synopsys Platform Architect™, a dynamic SystemC standards-based graphical environment for capturing, configuring, simulating, and analyzing designs at the system level. Figure 5 shows a typical design process for a new AI chip where the design stages for specification, implementation, physical design, and fabrication are shown left to right on the project timeline.

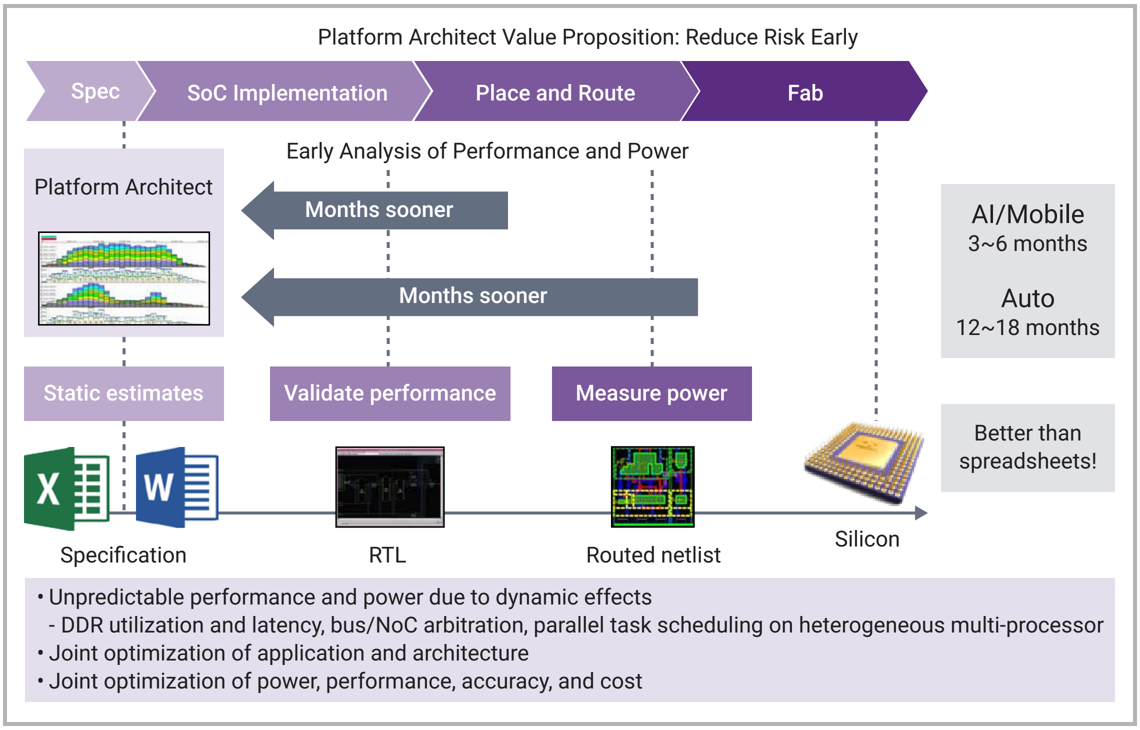

Platform Architect Value Proposition: Reduce Risk Early
Spec
SoC Implementation
Place and Route
Fab
Early Analysis of Performance and Power
Platform Architect
Months sooner
Months sooner
AI/Mobile
3~6 months
Auto
12~18 months
Static estimates
Validate performance
Measure power
Better than spreadsheets!
Specification
RTL
Routed netlist
Silicon
• Unpredictable performance and power due to dynamic effects
- DDR utilization and latency, bus/NoC arbitration, parallel task scheduling on heterogeneous multi-processor
• Joint optimization of application and architecture
• Joint optimization of power, performance, accuracy, and cost

Figure 5: AI chip flow with Platform Architect

In a traditional flow without Platform Architect, non-functional design properties can only be verified late in the development process. Spreadsheets, a common static tool for chip architects, allow only very rough static estimates of performance and power. Once the RTL design is ready hardware performance can be measured and validated, but there are two big issues. The first is that RTL simulation is not fast enough to model realistic traffic scenarios. The other is that any performance issues discovered at the RTL stage may lead to changes in the architecture, creating a long turnaround time (TAT) loop to update the design. Accurate power analysis has traditionally not been available until even later, after the design has been synthesized and laid out. At that stage it is hard to make changes that have more than a 10% impact on overall power without going back and making changes to the architecture. If the RTL design or the architecture must change to resolve power issues then the TAT is unacceptable, and project schedule will slip.

If power or performance issues are found in the bring up process in the lab or later in the field the cost of respins will be millions of dollars. The bottom line is that the designers must get the architecture right early in the project by simulating realistic AI workloads. No amount of cleverness in the downstream EDA tools can compensate for an architecture that is fundamentally wrong for the AI applications being targeted. The solution is to use Platform Architect to do quantitative analysis and power-performance optimization as early as possible based on system level models. This requires accurate models of the chip interconnect and memory, realistic AI workload models, and a system level power model complaint with Unified Power Format (UPF) 3.0.

Simulating realistic AI workloads may sound like an overwhelming task, but Platform Architect AI Exploration Pack (AI XP) does a great deal to help. It enables the exploration and optimization of AI chip designs by providing AI centric workload libraries and utilities. These include:

- An AI operator library for neural network modeling (convolution, Matmul, MaxPool, BatchNorm, etc.)
- An example workload model of the popular ResNet50 CNN
- A utility to convert an Open Neural Network Exchange (ONNX) or prototxt description to a workload model using the AI operator library
- An AI centric hardware architecture model library

Choosing a multi-die chip implementation places even more pressure on the architecture exploration phase since it is critical to consider all options and account for the connections between the dies or chiplets.

Platform Architect for Multi-Die helps optimize hardware-software partitioning, IP selection and configuration, interconnect and memory configuration, and power estimation with consideration of the die-to-die interfaces. It enables model-based architecture exploration for multi-die systems, including dynamic analysis of power and performance, based on die choices. Platform Architect for Multi-Die (Figure 6) includes die-to-die IP models, including Synopsys UCIe, as part of its library portfolio to build a multi-die system for early architecture exploration. It helps architects partition an application into multiple dies, distributing functions and memory to minimize die-to-die traffic. It evaluates different technologies and packaging, assesses connectivity and interconnect choices, and helps select die-to-die PHYs (e.g., UCIe, XSR) and protocols (e.g., PCIe, CXL). Reported metrics include latency (ns), energy efficiency (pJ/bit), bandwidth per beachfront (Gbps/mm), and area efficiency (Gbps/mm^2).

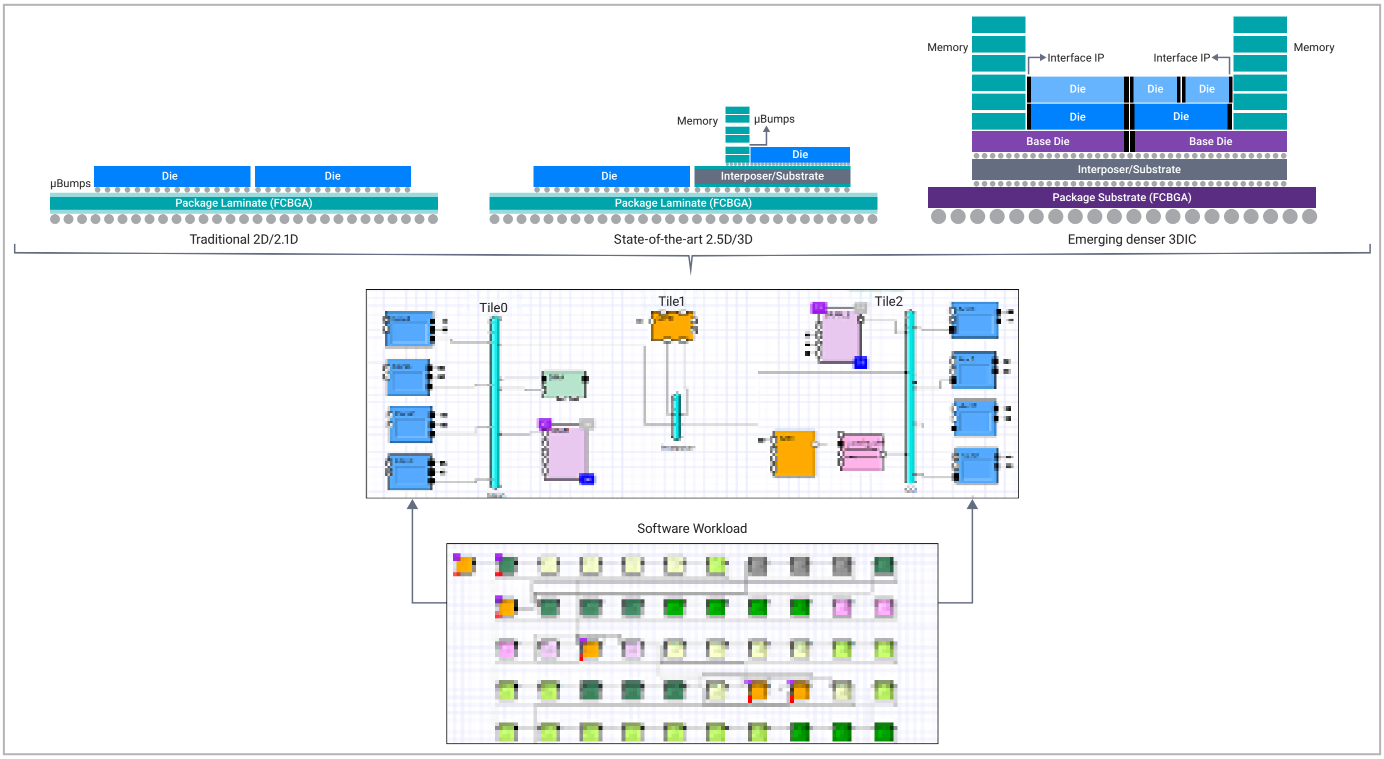


Figure 6: Platform Architect for Multi-Die

The Synopsys Solution: IP Selection

As shown in Figure 4, the other main aspect of pre-silicon planning is IP selection. Because AI tends to move large amounts of data around it puts a lot of pressure on interconnects and I/O bandwidth in addition to compute power. Choosing the appropriate interface IP is important and Synopsys offers a rich set of choices. Figure 7 shows a distributed, scalable GenAI architecture and the most likely IP choices.

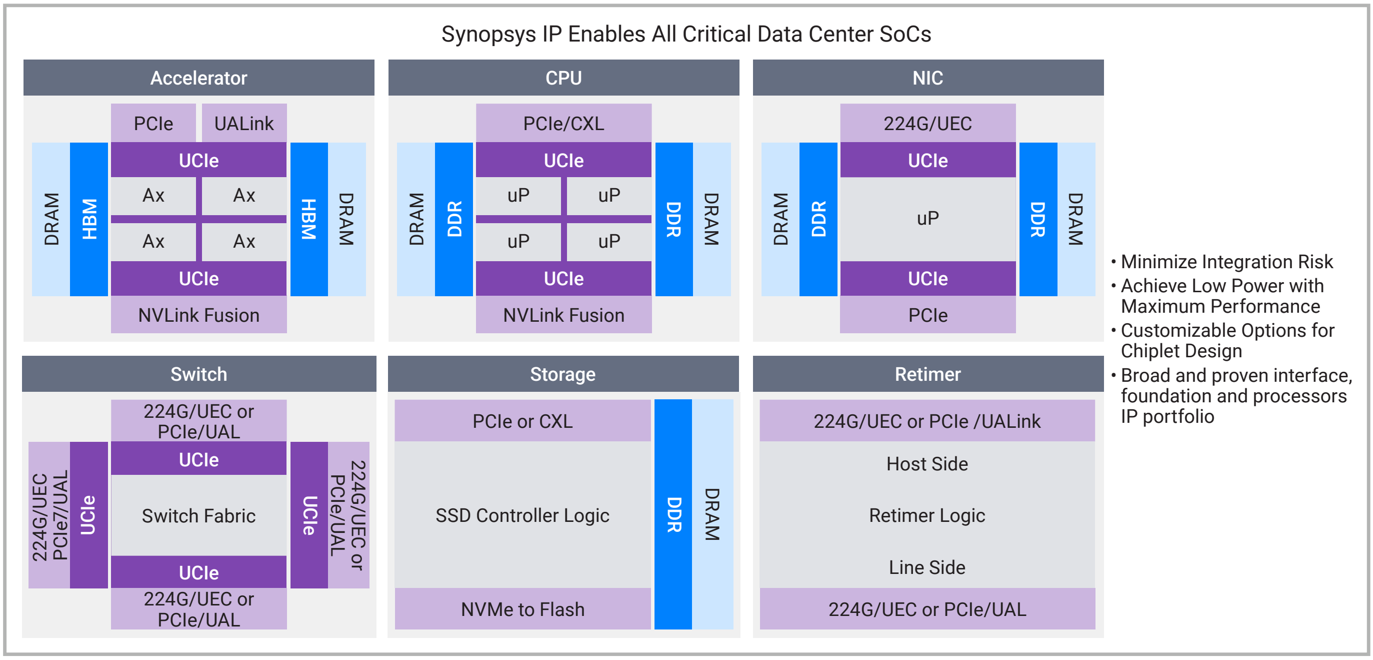


Figure 7: Complete, secure IP solutions accelerate time to silicon success

Synopsys' industry-leading IP portfolio, including complete solutions for scaling up and out with 224G, PCIe 7.0/6.x, UCIe, HBM, (LP)DDR5/6X, and CXL IP, are meticulously designed to minimize integration risk and significantly accelerate time-to-market. With more than 25 years of customer silicon success and dedicated engineering design teams, Synopsys is committed to delivering unparalleled IP excellence. These flexible solutions enable changes at any point in the design process, ensuring projects stay on track and meet evolving requirements.

Security is another important metric when choosing AI IP. Given the critical applications being entrusted to AI, it is imperative to prevent malicious agents from poisoning training data sets, tampering with models, changing decisions, or stealing private data at any point. Synopsys provides a broad portfolio of certified and standard compliant security IP. These ensure that AI chip designers can deliver secure authentication, data encryption, key management, platform security, and content protection to their end users.

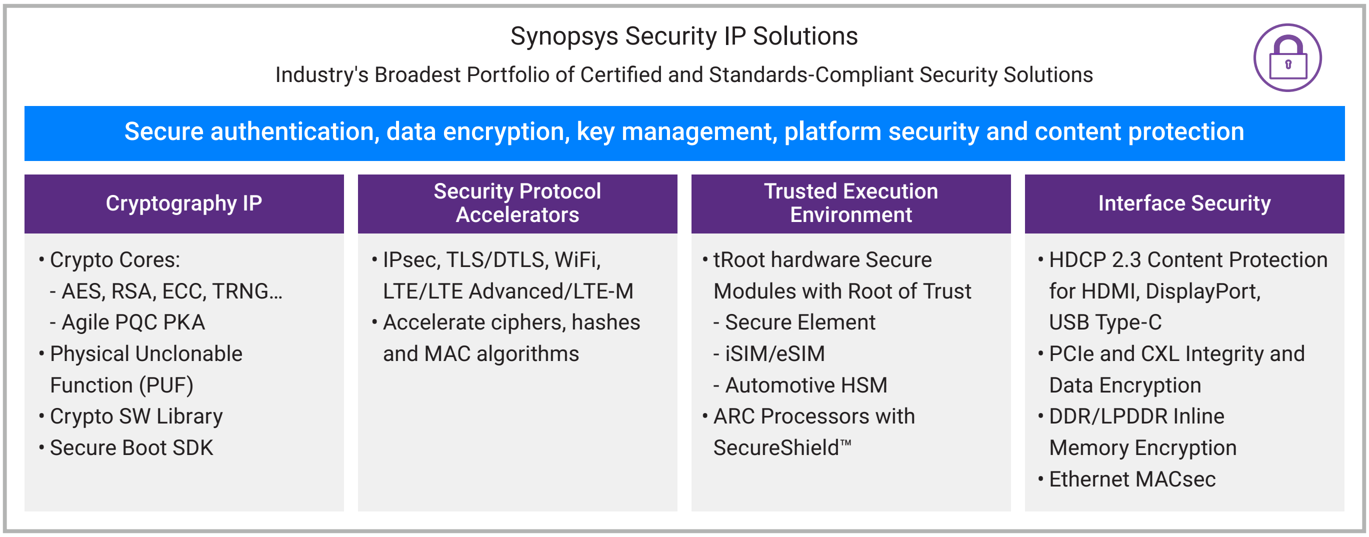

Synopsys Security IP Solutions
Industry's Broadest Portfolio of Certified and Standards-Compliant Security Solutions
Secure authentication, data encryption, key management, platform security and content protection
Cryptography IP
• Crypto Cores:
- AES, RSA, ECC, TRNG…
- Agile PQC PKA
• Physical Unclonable Function (PUF)
• Crypto SW Library
• Secure Boot SDK
Security Protocol Accelerators
• IPsec, TLS/DTLS, WiFi, LTE/LTE Advanced/LTE-M
• Accelerate ciphers, hashes and MAC algorithms
Trusted Execution Environment
• tRoot hardware Secure Modules with Root of Trust
- Secure Element
- iSIM/eSIM
- Automotive HSM
• ARC Processors with SecureShield™
Interface Security
• HDCP 2.3 Content Protection for HDMI, DisplayPort, USB Type-C
• PCIe and CXL Integrity and Data Encryption
• DDR/LPDDR Inline Memory Encryption
• Ethernet MACsec

Figure 8: Synopsys portfolio of security IP solutions

Summary

For architecture exploration during pre-silicon planning, Platform Architect provides an unparalleled solution. User experiences on real-work projects have shown that the schedules for AI/mobile applications can be reduced by 3-6 months, and that automotive chips can be delivered 12-18 months faster. At the 2023 Synopsys Virtual Prototyping Day Prototyping Day, engineers from Meta Platforms presented their experiences using Platform Architect and its library of models. They highlighted the solution's flexibility in creating models, generating traffic, running simulations, and collecting and displaying information.

For IP selection during pre-silicon planning, Synopsys offers a broad portfolio of proven, trusted IP for on-die, die-to-die, and off-chip connectivity. The best-in-class IP continues to evolve to take advantage of the latest technology nodes, with Synopsys recently reporting first-pass silicon success on TSMC's 2nm processes. Synopsys also provides a rich library of security IP solutions to ensure that AI chips deliver results free from interference during both training and deployment.

As AI grows ever more pervasive and AI applications demand ever more power and performance, developers will increasingly design dedicated chips. There is no doubt that these chips will continue to push the boundaries of silicon technology, EDA tools, and foundries. Synopsys is dedicated to supporting AI chip design at every step. Whether an established player or a startup, every company can turn to Synopsys as a "one stop shop" for AI chip design.

References

[1] https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

[2] https://hc2023.hotchips.org/assets/program/tutorials/ucie/UCIe%20overview%20and%20usage%20models.pdf (slide 8)