INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS™

2024 IRDS

EXECUTIVE PACKAGING TUTORIAL

PART 2
TRENDS AND LOOKING TO THE FUTURE WITH HIGH-PERFORMANCE COMPUTING AND ARTIFICIAL INTELLIGENCE

THE IRDS IS DEVISED AND INTENDED FOR TECHNOLOGY ASSESSMENT ONLY AND IS WITHOUT REGARD TO ANY COMMERCIAL CONSIDERATIONS PERTAINING TO INDIVIDUAL PRODUCTS OR EQUIPMENT.

THE INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS: *2024*

Table of Contents

List of Figures

List of Tables

DISCLAIMER

To provide realistic and tangible understanding of multi-chip module (MCM) technologies the Executive Packaging Tutorial (EPT) has collected information derived some of the most commonly available examples existing in the market to fulfill a realistic tutorial function. It is not the goal of the EPT to endorse or advertise any of these MCM approaches. Furthermore, since the assessment of these MCMs is done on the basis of publicly available information it is possible that the original suppliers may have additional knowledge and information that could change some of the assessments reported in the EPT. It is recommended that readers seeking additional insight on any of the reported MCM approaches and related subjects should be contacting the suppliers to obtain the most reliable information.

Finally, many of the MCM potential benefits and potential issues are reported on a case-by-case approach. These statements represent a technical opinion of the EPT and should not be constructed as benchmarking assessments since no experimental verification has been conducted in preparing this report. In addition, none of the assessments expressed for any specific MCM approach should be used as a benchmarking reference when comparing one MCM approach to another.

INTERNATIONAL ROADMAP FOR DEVICES AND SYSTEMS (IRDS) EXECUTIVE PACKAGING TUTORIAL PART 2

Trends and Looking To The Future with High-Performance Computing and Artificial Intelligence

Overview

It is the purpose of this Executive Packaging Tutorial (EPT) to shed some light on the adoption, evolution and transformation of multi-chip modules (MCMs) in the era of high-performance computing (HPC) and artificial intelligence (AI) to help the reader to better comprehend the drivers and the roles of the fundamental building blocks of this technology revolution in a neutral and objective way.

This tutorial is providing a broad overview of the fast-evolving field of packaging. Until five years ago system on chip (SoC) products provided an adequate level of transistor integration necessary to realize a variety of multi-functional building blocks on a single chip meeting the requirements of the most advanced systems. The requirements of fast evolving HPC and AI systems have however outpaced the rate at which the required functionalities can be realized on a single die. This situation has re-energized packaging technologies developed as far back as 50 years ago for high performance but then prohibitively expensive systems. Presently, all the leading system integrators rely on packaging technologies to co-locate multiple dice in a single package (SiP) and multiple closely connected MCM in a single board. New packaging technologies have also been developed and deployed to accelerate an increased level of integration at the package and board levels. Many companies have developed proprietary nomenclatures to describe their own solutions. It is the purpose of this tutorial to provide a neutral overview of the field of advanced packaging.

Most of all the EPT consists in outlining step-by-step a major shift in packaging technology initially driven by HPC and presently accelerated by AI. Personal computers (PCs) and smart phones were the drivers of the electronics industry for the past 40 years. Their selling prices hovered in the range of thousands of dollars. Consequently, analysis of the bill of materials (BOM) documented that cost of the each of the integrated circuits spanned from a few dollars to tens of dollars. This limited allocation of dollars to each of the IC products consequently demanded to maintain the cost of packaging in the cents/pin range. To accomplish this goal packaging manufacturing was transferred oversees early on (i.e., mid-seventies) to mainly benefit from the low-cost of labor.

The cost of the packaging materials were also closely monitored and minimized. Packaging materials evolved from costly ceramic materials to plastic encapsulation to currently “organic” packaging substrates, just to mentioned a few. The downside of all these low-cost solutions resides in the implication that signal propagation delay resulted severely degraded right after the signal left the IC due to poor propagation properties of all the materials used for the in-package interconnections related to their intrinsic electrical properties. To partially circumvent this problem, multiple system in package technologies known at the time as multi-chip modules were developed about 30 years ago and used for very expensive HPC systems. New MCM applications were rediscovered in recent years driven by AI applications. SiP consists mainly in co-locating memory chips next to logic chips on the substate of a single package; by so doing overall space occupied by the components is minimized as compared to connecting several individually packaged chips co-located on the same board. This design results in signal propagation delays being reduced due to shorter lengths of interconnections (from cm to mm).

Discussions about further reducing the effect of the package organic substrate on interconnect delays are evolving towards the development of glass substates to provide better performance. This is perfectly reasonable based on a packaging driven solution, but this is where it is necessary to make a major departure from traditional approaches depending on the type of product.

The HPC and the AI worlds are presently undergoing a major transition from central processing unit (CPU)-driven HPC computational data centers to graphical processing unit (GPU)-driven AI data centers.

The basic building block of a datacenter is represented by the server “rack” whose price becomes the reference performance and cost benchmark to be compared to PCs and mobile phones prices. For example, a Blackwell-powered rack may cost up to several millions of dollars! Analyzing the price of a data center and going down the supply chain looking at the key components of a rack reveals that a Bella board may cost hundreds of thousands of dollars. A Blackwell chip may be priced up to $20, 000 to $70,000!

In this case also, any reduction in signal propagation delay remains essential and if a comparison is made to accomplish this goal between the ratio of traditional IC cost to packaging cost it becomes clear that tens and even hundreds of dollars can be allocated to the cost of an HPC or AI package.

With this realization it is necessary to abandon the world of low-cost driven packaging world and step into the HPC/AI-driven world of leading-edge expensive wafer manufacturing to get the right answer!

Recently, the concept of power delivery from the backside of an IC has been demonstrated (backside power distribution (BPD)) and several companies have announced plans to implement that technology. Additionally, the intention of distributing clock signals from the backside (backside clock (BC)) of a CPU or a GPU have been reported. This is realized by manufacturing these capabilities on a "partially sacrificial" wafer and then "thinning and pealing" the surface of this wafer where the innovative BPD and clock distribution circuitry is built. After these initial process steps, "bonding" the thinned down and peeled wafer surface to the backside of the microprocessor unit (MPU) or GPU wafer yields the desired structural configuration.

Furthermore, it appears that other active devices can be implemented in the sacrificial wafer since it is amenable to be produced with any leading-edge technology. With this realization it then appears clear that by optimizing the thickness of the thinned-down support wafer it is possible to extend its role to becoming the substrate of the whole SiP! This approach would also provide additional benefits. For instance, thermal compatibility between the upper chip and the chip bonded on the backside would be ensured since both are made of silicon. In addition, multiple levels of interconnections could be fabricated on this new silicon substrate to connect all the other chips located in the SiP with substantial reduction of signal propagation delays.

The cost of this complex substate produced by leading-edge silicon technology would likely escalate in the hundreds of dollars or more, but it should be remembered that the combined, integrated dice are selling for tens of thousands of dollars! This realization revisits the 60s to 90s era in the industry in which only system companies could afford the adoption of expensive flip chip packaging technology. This is because this cost disappeared in the overall cost of the system that was rewarded by very handsome selling prices in the ten thousands and hundred thousands of dollars.

Packaging History

Encapsulation of transistors was devised early in IC design and manufacturing to prevent any damage to the frail structure of transistors during handling. Connections from ICs to package were accomplished by bonding gold wires first to the aluminum bond pads of the IC by ultrasonic welding and then connecting these bond wires onto gold pads located inside the package case. Once the die was fully encapsulated within the package shell, elongated pins protruding from the bottom or side of the package and connected to the package bond pads were then inserted in a socket located on a printed circuit board (PCB). Electrical signals generated by individual transistors and later on by integrated circuit (IC) were connected via metallic traces located on the surface of the board to other components placed on the same PCB.

As time went by, the number of transistors and complexity of ICs kept on increasing and consequently die size increased. To accommodate for the increased number of signals exiting the ICs it became necessary to evolve the package to a dual in-line structure to accommodate for pins on two sides. Eventually the package evolved to a square format with pins on all four sides to maximize the number of available inputs/outputs for electrical and mechanical connections. During this evolution to more connections, the role of packaging remained limited to physical protection of ICs and as electrical conduits to the board and systems, thus mainly as a rather passive element. However, things drastically changed in the past 20 years as the role of packaging surged to the level of an active solution provider. The most dramatic example to illustrate this point occurred in 2011.

Qualcomm and Xilinx were both competing for market share in field-programmable gate array (FPGA) technology, these products were characterized by very large die sizes. Both companies relied on outside sources for the silicon wafers manufacturing. The die size of FPGA had kept on further increasing therefore progressively and drastically reducing the number of dice produced on each wafer due to both die size and die yields; conjunctively the 28 nm technology available on the market at the time was rather immature and it was not fully adequate to support both performance and yields required for FPGA. However, when technology by itself is not sufficient to solve a problem, it is time for creative engineering solutions. Xilinx made the insightful decision to split the large die into two dice. These were located side by side into the same package but electrically connected with each other so that operated as a single large die. It is well known that for a given number of defects the yields of smaller dice are higher than the yields of an equivalent die twice the area. Furthermore, the absolute amount of leakage and other deleterious effects were more manageable by partitioning the die into two dice. In the end this approach called 2.5D was successful and prevailed in the marketplace; for the first time the crucial role of advanced packaging stood up and forever changed

the role of packaging from a passive participant to an active enabler. It was then said that packaging solutions could provide contributions to the final product by providing introduction of a *"half technology node generation"* ahead of schedule.

It soon become clear that this first step in novel packaging technology was only the beginning of an intense integration between dice and packages in providing the most viable and timely system-oriented solutions.

In 2014 the various teams of the International Technology Roadmap for Semiconductors (ITRS) effort—realizing that many transformations in IC manufacturing were underway—warned that major changes were coming in the electronics industry, such as new IC/package roles. The ITRS identified that the relationship between systems, integrated circuits, and packages was becoming symbiotically intermingled going forward.

Due to this realization the ITRS reassembled the seventeen highly specialized technology working groups (TWGs) existing at the time into seven integrated focus teams (IFTs). The first reports of this new organization were published in the 2015 ITRS2.0. Among the newly created chapters a forward-looking chapter entitled "Heterogeneous Integration" was published and for the first time provided a novel view of the many integrated packaging and IC solutions coming in the next 15 years. It was expected that initially *homogeneous integration* of electronics components (i.e., chips based on similar silicon technologies) were going to populate the system in package world and then evolve into more complex ones in the subsequent 10-15 years. Within this projected timeline an era of truly heterogenous integration of ICs (created with fundamentally different technologies) was to occur starting with the year 2025; as such, this process is just beginning.

As predicted by the ITRS in 2015 and subsequently by the IRDS, the interaction and symbiosis between ICs and packages has continued to synchronically and homogenously increase and as a result multi-chip modules have now become fundamental building blocks in the manufacture of any high-performance electronics system. Many companies have created their own approach and *self-defined nomenclature*, and, even though some differences exist, it can be shown that fundamentally these are all variations on the same technology.

The Approach to the Executive Packaging Tutorial

This EPT has been assembled in three parts. Each one of them is essentially "self-contained." This approach allows the knowledgeable reader to advance to the section of their interest without necessarily reading several of the previous sections. This approach comes at a price since it creates some intrinsic redundancy but provides the benefit to the reader of "self-contained" sections. Therefore, the readers should be aware that this is not an editorial failure but it is motivated by a necessary level of redundancy.

While the main purpose of Part 1 is to outline the features and evolution of packaging technologies leading to system in package for HPC and AI, Part 2 continues the historical progression with respect to system-on chip-technologies, system in package (SiP), and system in board (SIB) and their evolution, integration, and applications' trends for future HPC and data center requirements. A more extensive analysis of SoC is provided in the 2024 IRDS Executive Summary, and it is recommended that the two documents are evaluated side by side.

In order to make the content more easily readable additional details are provided in the appended material for further insights. Readers can refer to each of the EPT's appended material for in depth information and discussion.

Key Trends

Semiconductor packaging is evolving rapidly due to the increasing demands for smaller, faster, and more energy-efficient electronics products. Associated with these demands are significant trends that are expected to shape the future of semiconductor packaging landscape for consumer products, drive innovation and enable the next generation of electronic devices at large. See Table 1.

Table 1. Key Trends and Advanced Packaging Evolution

Trend	Impact
Homogeneous Integration	• 2.5D and 3D Packaging: These technologies enable stacking multiple chips on top of each other or side by side, leading to better performance and lower power consumption. 2.5D involves placing chips side by side on an interposer, while 3D involves stacking them vertically. • Chiplet Architecture: Instead of designing a single large monolithic chip, chiplets allow smaller, function-specific chips to be combined, offering flexibility and cost advantages. • Fan-Out Wafer-Level Packaging (FOWLP): This packaging technique offers a higher level of miniaturization and better thermal performance, making it suitable for high-performance computing and mobile devices.
Heterogeneous Integration	• Integration of Diverse Technologies: Combining different types of chips (e.g., analog, digital, non-silicon based and photonic) into a single package allows for more complex and capable systems. This trend is crucial for applications like artificial intelligence (AI), the internet of things (IoT), and 5G.
System-in-Package (SiP)	• Miniaturization of Entire Systems: SiP packages multiple components (e.g., processors, memory, sensors) into a single module, enabling compact devices with enhanced functionality. This trend is particularly important for wearable technology and IoT devices.
Thermal Management Innovations	• Advanced Cooling Solutions: As chips become more powerful, managing heat dissipation is critical. Techniques like integrated liquid cooling, improved thermal interface materials, and innovative packaging designs are gaining importance.
Materials Advancements	• Use of New Materials: The shift to new materials like silicon carbide (SiC) and gallium nitride (GaN) is enabling higher efficiency and better performance, particularly in power electronics and RF applications. • Bio-compatible Materials: For medical devices and wearables, there's a growing demand for packaging materials that are safe for long-term contact with the human body.
Reliability and Testing	• Improved Testing Techniques: As packaging becomes more complex, so does the need for advanced testing methods to ensure reliability, especially for critical applications in automotive, aerospace, and healthcare. • Enhanced Reliability: Packaging designs are increasingly focusing on improving reliability under harsh conditions, such as high temperatures and mechanical stress, which is critical for automotive and industrial applications.
Sustainability and Cost Efficiency	• Eco-friendly Packaging: The semiconductor industry is pushing towards more sustainable practices, including the use of recyclable materials, reduction of hazardous substances, and energy-efficient manufacturing processes. • Cost-Effective Manufacturing: The ongoing trend towards cost reduction will drive innovations in packaging processes that can reduce overall manufacturing costs without compromising performance.
AI and Machine Learning in Packaging Design	• Optimization through AI: Artificial intelligence and machine learning are increasingly being used to optimize packaging design and manufacturing processes, leading to better performance and reduced time-to-market.
Photonics Integration	• Photonic Packaging: With the growing importance of optical communication, integrating photonics with traditional electronic chips in a package is becoming more common, enabling faster data transfer and improved signal integrity.
Security Features	• Hardware-Level Security: As cyber threats increase, embedding security features directly into semiconductor packages is becoming a trend, especially for applications in defense and secure communications.

The Special Role of Packaging in High Performance Computing

High-performance computing (HPC) is at the forefront of advancing scientific research, artificial intelligence, and complex simulations. For this reason, the role of semiconductor packaging in HPC is critical, as it directly impacts the performance, power efficiency, and scalability of HPC systems. Table 2 presents an overview of how packaging contributes to HPC.

Table 2. Packaging Contributions to High-performance Computing

Packaging Contribution to HPC	Description
Performance Enhancement	• Interconnect Density: Advanced packaging technologies like 2.5D and 3D integration enable higher interconnect density, reducing the distance between components. This minimizes latency and improves data transfer speeds, which are crucial for HPC workloads. • Reduced Parasitics: By using advanced packaging techniques that place components closer together, the parasitic effects such as resistance, capacitance, and inductance are reduced. This enhances signal integrity and overall system performance.
Thermal Management	• Efficient Heat Dissipation: HPC systems generate significant heat due to the high power consumption of CPUs, GPUs, and other processing units. Advanced packaging solutions integrate efficient thermal management techniques, such as thermal interface materials (TIMs), heat spreaders, and liquid cooling, to maintain optimal operating temperatures. • 3D Packaging Challenges: While 3D stacking increases performance by reducing interconnect lengths, it also poses challenges in heat dissipation. Packaging plays a crucial role in addressing these thermal challenges through innovative cooling solutions.
Power Efficiency	• Power Delivery Networks (PDNs): Packaging technology directly influences the efficiency of power delivery to the components. High-performance packaging solutions ensure low-resistance and low-inductance PDNs, which are vital for stable and efficient power delivery in HPC systems. • Energy Efficiency: Advanced packaging technologies like wafer-level packaging (WLP) and flip-chip can optimize energy efficiency by minimizing power loss and improving signal integrity.
Scalability and Integration	• Heterogeneous Integration: HPC systems often require the integration of diverse components, such as CPUs, GPUs, memory, and specialized accelerators. Packaging plays a crucial role in heterogeneous integration, enabling the combination of different technologies (e.g., silicon, GaN, InP) within a single package. • Chiplets: The use of chiplet-based architectures, where multiple smaller dies are packaged together, allows for scalable and modular HPC systems. This approach facilitates customization and rapid iteration of HPC designs.
Data Bandwidth and Communication	• High-Bandwidth Memory (HBM): Advanced packaging is essential for integrating HBM with processors in HPC systems. The close proximity of memory and processors, enabled by 2.5D and 3D packaging, allows for extremely high data bandwidth, which is crucial for HPC workloads. • Optical Interconnects: As data bandwidth requirements continue to grow, packaging technology is evolving to support optical interconnects. These interconnects offer higher data rates and lower latency compared to traditional electrical interconnects, which is vital for HPC performance.
Reliability and Durability	• Robust Packaging Materials: HPC systems operate under extreme conditions, including high temperatures and constant high loads. Packaging materials must be durable and reliable to ensure long-term operation without failure. • Failure Mitigation: Advanced packaging techniques include features that

Packaging Contribution to HPC	Description
	mitigate the risks of failure, such as redundant interconnects, advanced sealing methods, and protection against electromagnetic interference (EMI).
Cost and Manufacturing Considerations	• Economies of Scale: As HPC systems scale in complexity and production volume, packaging solutions must also scale efficiently. Advanced packaging technologies can help reduce costs through innovations in manufacturing processes, such as wafer-level processing and automated assembly. • Yield Management: In HPC, where the performance of individual components is paramount, packaging plays a role in maximizing yield by reducing defects and ensuring that only high-performance components are used.

In summary, packaging technology is a critical enabler for HPC, influencing everything from performance and power efficiency to scalability and reliability. As the demands of HPC continue to grow, so too will the importance of advanced semiconductor packaging. Impacts on future directions include emerging packaging technologies, with innovations such as fan-out wafer-level packaging (FOWLP) and embedded multi-die interconnect bridge (EMIB) are being explored to further enhance the performance and integration capabilities of HPC systems. Quantum Computing will be impacted too, as quantum computing becomes more viable, packaging will play a crucial role in integrating quantum processors with classical computing components, ensuring coherence and minimizing interference.

Silicon as a Packaging Substrate

Silicon is increasingly being used as a packaging substrate material in advanced semiconductor packaging technologies. Traditionally, packaging substrates were made from materials like organic laminates (e.g., FR4), ceramic, or glass. However, the use of silicon as a packaging substrate offers several unique advantages that make it well-suited for high-performance and high-density applications. Below is an overview of the role of silicon as a packaging substrate and Table 3 shows several advantages of using silicon as a packaging substrate.

Table 3. Advantages of Silicon as a Packaging Substrate

Silicon Packaging Substrate Attributes	Advantages
High Thermal Conductivity	• Efficient Heat Dissipation—Silicon has a high thermal conductivity (around 150 W/m·K), making it effective at dissipating heat generated by high-performance chips. This is crucial for maintaining the reliability and longevity of semiconductor devices, particularly in applications like high-performance computing (HPC) and data centers. See Figure 1 for an illustration of thermal conductivity. • Thermal Management Integration—Silicon substrates can be integrated with microfluidic channels or other advanced cooling solutions, further enhancing thermal management.
High-Density Interconnects	• Fine-Line Routing—Silicon allows for the creation of fine-line routing and high-density interconnects, which are essential for advanced packaging technologies like 2.5D and 3D integration. This capability supports the increasing need for higher pin counts and finer pitches in modern ICs. • Through-Silicon Vias (TSVs)—Silicon substrates can be fabricated with TSVs, enabling vertical electrical connections in 3D stacked architectures. TSVs reduce the length of interconnects, thereby minimizing signal propagation delay and improving overall system performance.
Mechanical Properties	• Dimensional Stability—Silicon has excellent dimensional stability and a low coefficient of thermal expansion (CTE), which is closely matched to that of silicon chips. This reduces the risk of thermal

Silicon Packaging Substrate Attributes	Advantages
	mismatch and mechanical stress between the substrate and the chips, leading to improved reliability. • High Modulus—Silicon's high modulus (stiffness) provides structural integrity to the package, reducing warping and deformation during the assembly and operation processes.
Electrical Performance	• Low Electrical Losses—Silicon substrates exhibit low electrical losses, which is beneficial for high-frequency applications. This characteristic supports the efficient transmission of high-speed signals, which is important in communication, RF, and high-speed computing systems. • Integrated Passive Devices (IPDs) —Silicon substrates can be used to integrate passive components like capacitors, resistors, and inductors directly onto the substrate, enhancing the overall performance and reducing the footprint of the package.
Compatibility with Silicon-based Devices	• Monolithic Integration—Silicon substrates allow for the potential monolithic integration of silicon-based devices and circuits directly onto the packaging substrate. This integration simplifies the overall design and can lead to more compact and efficient systems. • Ease of Manufacturing—Silicon is already widely used in semiconductor manufacturing, which means that the infrastructure and processes for handling and processing silicon are well-established. This compatibility facilitates the adoption of silicon substrates in advanced packaging.

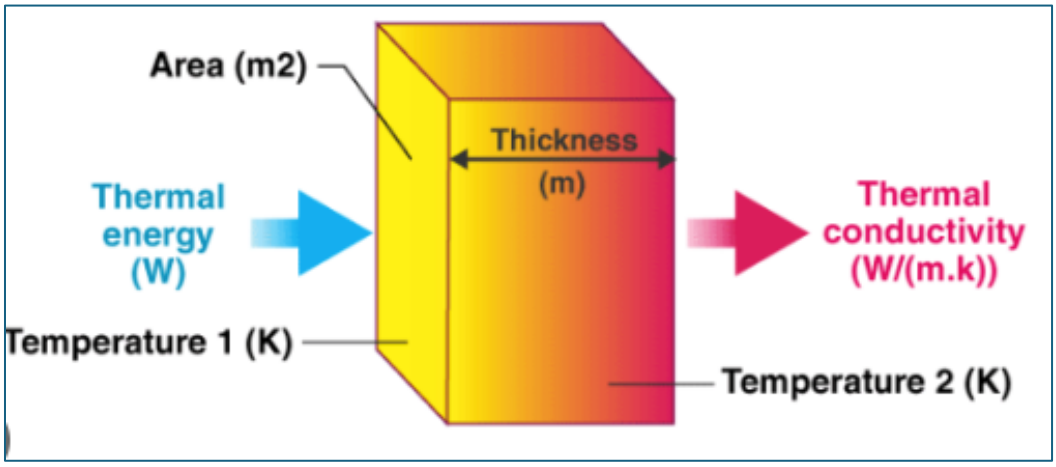


Figure 1. Thermal conductivity

Silicon as a Packaging Substrate—Applications Examples

- **2.5D and 3D Integration**
  - **Interposers**—In 2.5D integration, silicon interposers are used as a substrate to interconnect multiple dies on a single package. The interposer allows for high-density routing and integration of heterogeneous components, such as logic, memory, and analog devices.
  - **3D Stacking**—Silicon substrates with through silicon vias (TSVs) enable 3D stacking of dies, reducing the overall footprint and improving performance by shortening interconnects between layers.
- **High-Performance Computing (HPC) and AI**
  - **Data Centers and AI**—Silicon substrates are used in packaging high-performance processors and accelerators for data centers and AI applications, where thermal management, signal integrity, and high-density interconnects are critical.
  - **Graphics Processing Units (GPUs)**—Silicon substrates are also used in the packaging of GPUs, where high data throughput and efficient heat dissipation are essential.
- **RF and Microwave Applications**
  - **High-Frequency Circuits**—The low-loss characteristics of silicon substrates make them suitable for RF and microwave applications, where maintaining signal integrity at high frequencies is crucial.
  - **Antenna-in-Package (AiP)**—Silicon substrates can be used in antenna-in-package designs, integrating RF circuits and antennas on the same substrate, reducing size and improving performance.
- **Optoelectronics**
  - **Photonic Integration**—Silicon substrates can be used to integrate photonic devices with electronic circuits, enabling advanced optoelectronic packaging solutions for high-speed data communication and sensing applications.

Silicon Packaging Substrate Challenges and Considerations

- **Cost**
  - **Higher Cost**—Silicon substrates are generally more expensive than traditional organic substrates, which can be a limiting factor for cost-sensitive applications. However, the performance benefits can justify the higher cost in high-end applications.
- **Complex Manufacturing**
  - **Precision Required**—The manufacturing of silicon substrates, especially those with fine-line routing and TSVs, requires precision and advanced fabrication techniques. This can complicate the production process and increase costs.
- **Fragility**
  - **Brittleness**—Silicon is brittle and can be prone to cracking under mechanical stress, particularly during the handling and assembly processes. Careful design and process control are needed to mitigate this risk.

Take Away

Silicon as a packaging substrate offers several advantages, particularly for high-performance and high-density applications. Its excellent thermal conductivity, electrical performance, and compatibility with advanced packaging techniques like TSVs and 3D integration make it an ideal choice for cutting-edge semiconductor devices. Despite challenges such as cost and fragility, the benefits of using silicon substrates in applications like high-performance computing, RF, and optoelectronics often outweigh these drawbacks, making it a key material in the evolution of semiconductor packaging.

Evolution from SOC to SIP to SIB

The evolution from System-on-Chip (SoC) to System-in-Package (SiP) and finally to System-in-Board (SiB) represents a progression in the way electronic systems are integrated and packaged. Each step in this evolution reflects increasing complexity, integration, and the ability to combine diverse functionalities within a single system. Table 4 is an overview of this progression.

Table 4. SoC, SiP, and SiB—Limitations and Advantages

System-on-Chip (SoC)	System-in-Package (SiP)	System-in-Board (SiB)
SoC is an integrated circuit (IC) that consolidates all the components of a computer or other electronic system onto a single chip.[1] This includes the processor, memory, Input/Output (I/O) ports, and other necessary components, all integrated onto a single semiconductor substrate.	SiP refers to a packaging technology where multiple integrated circuits (ICs) or components are integrated into a single package. These components could include processors, memory, sensors, power management ICs, and more.	SiB refers to the integration of multiple SiPs, SoCs, discrete components, and other electronic modules onto a single printed circuit board (PCB). This approach integrates complete systems onto a board, effectively creating a functional electronic system.
Key Characteristics: High Integration: SoC integrates multiple components and functionalities onto one chip, leading to reduced size, power consumption, and cost. Single Substrate: All components are fabricated on the same piece of silicon, leading to very high integration density. Optimized Performance: Because components are on the same chip, communication between them is faster, leading to higher performance and lower power consumption.	Key Characteristics: Heterogeneous Integration: SiP allows for the integration of different types of chips and components within a single package. This can include digital, analog, RF, and even passive components. 3D Integration: SiP can leverage 3D stacking and Through-Silicon Vias (TSVs) to vertically integrate components, leading to higher performance and density. Enhanced Functionality: By integrating multiple chips, SiP can offer enhanced functionality compared to a single SoC, without the need to redesign the entire system. Flexibility: SiP offers greater flexibility than SoC because it allows different dies, possibly from different process technologies or vendors, to be combined within the same package.	Key Characteristics: Complex Systems Integration: SiB involves the assembly of multiple subsystems, such as memory modules, processors, power management units, and communication interfaces, onto a single board. Modular Approach: SiB can integrate different functional blocks or modules, each of which could be a fully functional SiP or SoC. This modular approach allows for easier upgrades and customization. High Customizability: Unlike SoCs, SiBs offer high flexibility and customizability since designers can select and assemble various modules to meet specific application needs. Large Form Factor: SiBs typically have a larger form factor compared to SiPs and SoCs, allowing for more complex and higher power systems but requiring more space.
Applications: Mobile Devices: SoCs are commonly used in smartphones, tablets, and other portable devices where space and power efficiency are critical. Embedded Systems: SoCs are also used in embedded systems for automotive, industrial, and consumer electronics applications.	Applications: Wearable Devices: SiPs are used in compact, power-efficient devices like smartwatches and fitness trackers. IoT Devices: SiPs are ideal for IoT devices that require integration of sensors, processors, and wireless communication in a small form factor. Medical Devices: SiP technology is used in miniaturized medical	Applications: Servers and Data Centers: SiBs are used in server boards where multiple processors, memory modules, and other components are integrated onto a single board. Networking Equipment: Routers, switches, and other networking devices often use SiB designs to integrate various processing and communication modules. Automotive and Aerospace Systems:

System-on-Chip (SoC)	System-in-Package (SiP)	System-in-Board (SiB)
	devices, such as implantable devices, where space is at a premium.	SiBs are used in complex automotive and aerospace systems, where multiple functions need to be integrated onto a single board.
Limitations: Complexity and Cost: Designing and manufacturing an SoC is highly complex and expensive, especially as process nodes shrink. Limited Flexibility: Once fabricated, the design of an SoC is fixed, making it less adaptable to changes or updates.	Advantages over SoC: Integration of Diverse Technologies: SiP allows for the integration of different process technologies within one package. Reduced Time to Market: SiP can reduce development time by allowing designers to reuse existing ICs rather than designing a new SoC from scratch. Improved Yield: Since SiP uses multiple dies, defects in one die do not necessarily lead to the rejection of the entire package, potentially improving overall yield.	Advantages over SiP: Scalability: SiBs can scale to larger systems, integrating more components than what could be achieved in a single SiP. Flexibility: Designers can mix and match different SiPs, SoCs, and discrete components, tailoring the board to specific requirements. Easier Prototyping and Development: SiBs allow for easier prototyping and system testing, as individual components can be tested and replaced without redesigning the entire system.

Summary

The evolution from SoC to SiP and then to SiB reflects the increasing complexity of producing high performance systems that triggers the corresponding increase in integration levels in electronic system design, as follows:

- **SoC** is focused on maximizing integration within a single chip, offering compact, power-efficient solutions but with limitations in flexibility and design complexity.
- **SiP** extends this integration to a package level, allowing for the combination of multiple dies and components with greater flexibility and heterogeneous integration.
- **SiB** takes integration to the board level, enabling complex systems with multiple modules and components, offering scalability, flexibility, and customizability for high-performance and specialized applications.

Each approach serves different needs, with SoC being ideal for compact, high-performance designs, SiP for integrating multiple technologies in small form factors, and SiB for building complex, customizable systems on a larger scale.

SIB Technology Overview

An example of SIB technology is the NVIDIA Bella board, a specialized development board designed by NVIDIA, primarily aimed at providing a robust platform for developing and testing AI and machine learning (ML) applications. It integrates cutting-edge NVIDIA GPUs with powerful CPUs and extensive I/O options to support various development needs. Key features, components, and potential applications of the NVIDIA Bella Board are detailed in Appendix A:

SIB Applications and Use Cases

1. **AI and Machine Learning**
    - **Model Training and Inference**: SIB technology is ideal for training large AI models, including deep learning and neural networks. Its powerful GPU, combined with extensive memory and fast interconnects, ensures that even the most demanding models can be trained efficiently.
    - **Edge AI**: With its high-performance capabilities and potential for integration into larger systems, SIB technology can be used in edge AI applications, where AI models need to be deployed in real-time, often in remote or distributed environments.
2. **High-Performance Computing**

- **Scientific Simulations**: SIB technology's computational power makes it suitable for running complex scientific simulations, including weather modeling, fluid dynamics, and molecular dynamics.
- **Data Analytics**: SIBs can handle large-scale data analytics tasks, processing vast amounts of data quickly and efficiently, which is essential for applications in finance, healthcare, and research.

3. **Autonomous Systems**
   - **Autonomous Vehicles**: SIBs could be used in the development of autonomous vehicle systems, where real-time processing of sensor data and decision-making algorithms are critical.
   - **Robotics**: SIB technology's high performance makes it suitable for advanced robotics applications, including industrial robots and drones, where AI-driven control systems are needed.
4. **Virtual Reality (VR) and Augmented Reality (AR)**
   - **Content Development**: SIBs can be used to develop and render VR and AR content, where high frame rates and low latency are essential for immersive experiences.
   - **Simulation and Training**: SIB's capabilities make it ideal for simulation and training environments that require high-fidelity graphics and real-time processing.

Advantages and Versatility of SIB technology: SIB combination of high-performance CPU and GPU, along with extensive I/O and networking options, makes it suitable for a wide range of applications, from AI to HPC and beyond.

- **Scalability**: SIB technology can be scaled to meet the needs of various applications, whether it's through adding more GPUs, expanding memory, or integrating additional peripherals.
- **Software Integration**: Software ecosystems, such as what NVIDIA provides, enables robust support for development, ensuring that SIB can be effectively used in a variety of complex environments.

SIB Considerations

- **Cost**: Given its high-performance components, this technology is likely to be expensive, which may limit its accessibility to smaller organizations or developers.
- **Power Consumption**: The SIB's high-performance nature means it may consume significant power, which could be a consideration for applications where power efficiency is critical.
- **Complexity**: The advanced features and capabilities of SIBs may require a steep learning curve, particularly for developers who are new to HPC or AI development.

SIBs are designed for high-performance computing and AI machine learning applications. The combination of cutting-edge GPU and CPU technologies, coupled with extensive I/O and networking capabilities, makes it a versatile tool for developers working on advanced computational tasks. While it comes with challenges such as cost and complexity, its potential to accelerate development and enhance performance in demanding applications makes it an asset in the fields of AI, HPC, autonomous systems, and beyond.

SIB Thermal Challenges

SIB technology, with its high-performance GPU and CPU, faces significant thermal challenges due to the intense heat generated during operation. Managing this heat effectively is crucial for maintaining system performance, reliability, and longevity. Table 5 is an analysis of the thermal challenges and potential solutions.

Table 5. SIB Thermal Challenges and Solutions

SIB Challenge	Details
High Power Density	Heat Generation: A SIB board's powerful components, particularly the GPU, can generate substantial heat. The dense integration of components increases the power density, leading to hot spots on the board that can affect performance and reliability if not properly managed.
Hot Spots	Localized Heating: Certain areas of the board, especially around the GPU, can become significantly hotter than others. These hot spots can lead to thermal stress, affecting the performance of nearby components and increasing the risk of thermal failure.
Thermal Throttling	Performance Degradation: To prevent overheating, the GPU and CPU may reduce their clock speeds, leading to thermal throttling. While this protects the components, it also results in a drop in performance, which can be critical in high-performance computing tasks.
Long-Term Reliability	Thermal Cycling: Repeated heating and cooling cycles can cause mechanical stress on the board and its components, potentially leading to long-term reliability issues such as solder joint fatigue, delamination, and component failure.
Cooling Constraints	Space and Form Factor: The Bella Board's design may limit the available space for traditional cooling solutions, such as large heatsinks or fans, making it challenging to implement effective thermal management without compromising the board's form factor.

SIB Solution	Details
Advanced Heat Sinks	High-Efficiency Heat Sinks: Using large, high-efficiency heat sinks with optimized fin designs can help dissipate heat more effectively from the GPU and CPU. Materials with high thermal conductivity, such as copper or aluminum, are commonly used. Direct Contact Heat Pipes: Integrating heat pipes that make direct contact with the heat-generating components can enhance heat transfer to the heat sink, improving overall cooling efficiency.
Liquid Cooling Solutions	Closed-Loop Liquid Cooling: A closed-loop liquid cooling system can be integrated to provide superior cooling performance. This involves circulating a liquid coolant through a block that sits on top of the GPU and CPU, carrying heat away from the components and dissipating it through a radiator. Microfluidic Cooling: For advanced applications, microfluidic cooling, where liquid coolants are circulated through microchannels embedded in the substrate, can provide highly localized cooling for hot spots on the board.
Thermal Interface Materials (TIMs)	High-Performance TIMs: Using advanced thermal interface materials, such as thermal pastes or pads with high thermal conductivity, can improve the efficiency of heat transfer between the components and their cooling solutions. Phase-Change Materials (PCMs): PCMs can be used as TIMs to provide effective heat management. These materials change state at certain temperatures, absorbing and dissipating heat efficiently.
Active Cooling Techniques	Forced Air Cooling: High-performance fans can be used to increase airflow across the board, helping to dissipate heat from the heat sinks and other components. This solution is often combined with heat sinks for enhanced cooling. Thermoelectric Cooling: Peltier devices can be used for thermoelectric cooling, where an electric current creates a temperature difference, drawing heat away from the components. This method is useful for precise temperature control but is typically more expensive and power-intensive.
Thermal Design Optimization	Component Placement: Optimizing the placement of heat-generating components on the board can help manage thermal loads more effectively. Spacing out hot components can reduce the likelihood of hot spots. Thermal Simulation and Modeling: Using thermal simulation tools during the design phase allows engineers to predict and mitigate potential thermal issues before manufacturing, optimizing the board layout for better heat dissipation.

SIB Solution	Details
Advanced Substrate Materials	High Thermal Conductivity Substrates: Using substrates with high thermal conductivity, such as certain ceramics or composite materials, can help spread heat more evenly across the board, reducing the concentration of hot spots. Thermal Vias and Conductive Planes: Incorporating thermal vias and conductive planes within the PCB can facilitate the transfer of heat away from critical components to the heat sink or other cooling solutions.
Fabrication Environmental Controls	Ambient Temperature Management: Ensuring that the operating environment is well-ventilated and maintained at a suitable ambient temperature can reduce the overall thermal load on the board. Enclosure Design: Designing the enclosure with adequate ventilation or incorporating active cooling elements within the enclosure itself can help manage the board's thermal profile.

Take Away

SIBs are designed for high-performance computing and AI machine learning applications. The combination of cutting-edge GPU and CPU technologies, coupled with extensive I/O and networking capabilities, makes it a versatile tool for developers working on advanced computational tasks.

While it comes with challenges such as cost and complexity, SIB technology's potential to accelerate development and enhance performance in demanding applications makes it an asset in the fields of AI, HPC, autonomous systems, and beyond.

SIB's thermal challenges stem from high-performance components, dense integration, and potential space constraints. However, by employing a combination of advanced cooling solutions such as liquid cooling, high-efficiency heat sinks, optimized thermal interfaces, and careful thermal design, these challenges can be effectively managed. Ensuring proper thermal management not only maintains the performance and reliability of SIB technology but also extends its operational lifespan, making it suitable for demanding applications in AI, HPC, and other high-performance domains.

Evolution of Transistors, Boards to Data Center Rack Integration

Roles of Transistors in SoC and Boards in Racks

The comparison of transistors as components of integrated circuits (ICs) to boards as components of racks highlights the hierarchical nature and evolution of modern electronic systems. Both transistors and boards serve as fundamental building blocks within larger, more complex systems, and their roles are analogous in many ways.

Transistors as Components of Integrated Circuits (ICs)

1. Fundamental Building Blocks

- **Basic Unit**: Transistors are the basic units of logic and memory in electronic devices. They act as switches or amplifiers, controlling the flow of electrical signals and enabling the operation of digital and analog circuits.
- **Integration**: In an IC, millions or even billions of transistors are integrated onto a single silicon chip to perform complex functions. These transistors are interconnected to form logic gates, memory cells, and other circuit elements that make up the IC.

2. Role in Functionality

- **Logic Operations**: Transistors are used to create logic gates (AND, OR, NOT, etc.), which are the building blocks of digital circuits. These gates are combined to create more complex functions, such as arithmetic operations, data storage, and control logic.

- **Memory Storage**: Transistors are also used in memory cells, such as SRAM and DRAM, where they store and manage data by holding electrical charges.
- **Signal Amplification**: In analog circuits, transistors amplify signals, enabling them to drive larger loads or travel longer distances within a circuit.

3. Scaling and Moore's Law

- **Miniaturization**: The scaling of transistors has driven the miniaturization of ICs, allowing more functionality to be packed into smaller chips. This has led to the exponential growth in computing power known as Moore's Law.
- **Power and Performance**: As transistors shrink, they consume less power and can switch faster, improving the overall performance and efficiency of ICs.

Boards as Components of Racks

1. Fundamental Building Blocks

- **Basic Unit**: Boards (also known as printed circuit boards or PCBs) are the basic units of larger electronic systems, such as servers, networking equipment, or storage systems. A board integrates multiple ICs, connectors, power management components, and other devices to perform specific functions.
- **Integration**: Multiple boards are integrated into a rack, which serves as a physical framework for housing and organizing the boards. Each board may have a specific role, such as processing, storage, or network communication, contributing to the overall functionality of the rack.

2. Role in Functionality

- **System Integration**: Boards within a rack work together to form a complete system, such as a server cluster or a data storage array. Each board may handle different aspects of the system's operation, such as computation, data management, or I/O processing.
- **Communication**: Boards communicate with each other through backplanes, cables, or network connections within the rack, enabling data exchange and coordination across the system.
- **Redundancy and Scalability**: Racks often include multiple boards for redundancy and scalability. For example, a server rack may have several processing boards to distribute the computational load and provide fault tolerance.

3. Scaling and Data Center Growth

- **Rack Scalability**: As data centers grow, racks are scaled up by adding more boards, each of which adds processing power, storage capacity, or networking capability. This modular approach allows for easy upgrades and expansion.
- **Power and Cooling**: Just as transistors face challenges with power and heat, boards in racks must be managed for power efficiency and thermal performance. Advanced cooling solutions and power distribution methods are used to optimize the operation of racks.

Analogies Between Transistors and Boards

1. **Hierarchy of Integration**
    - **Transistor in ICs**: Transistors are the smallest functional units, integrated into ICs to perform various electronic functions.
    - **Boards in Racks**: Boards are higher-level functional units, integrated into racks to perform more complex and diverse system-level tasks.
2. **Role in System Functionality**

- **Transistors Enable ICs**: The performance, power efficiency, and functionality of an IC depend on the design and integration of its transistors.
- **Boards Enable Racks**: The performance, reliability, and scalability of a rack system depend on the design, integration, and coordination of its boards.

3. **Scaling and Moore's Law Equivalent**
   - **Transistor Scaling**: The miniaturization of transistors leads to more powerful and efficient ICs, driving the advancement of electronic devices.
   - **Rack Scaling**: The addition and optimization of boards within a rack lead to more powerful and efficient data centers, driving the advancement of large-scale computing infrastructures.

Take Away

Transistors and boards both serve as fundamental components within their respective levels of electronics systems—transistors at the chip level and boards at the system level. The hierarchical relationship between these components allows for the construction of increasingly complex and powerful electronic systems, from individual ICs to entire data centers. Understanding this analogy helps in grasping the broader picture of how electronic devices are designed, scaled, and optimized across different levels of integration.

Data Center Rack Integration (DCRI)

Rack integration in data centers (DCRI) can be likened to the integration of transistors in systems on chips, ICs in system-in-packages and boards in system-in-board configurations. Each of these represents a level of aggregation, where individual components are integrated into increasingly complex and powerful systems. Table 6 presents how these concepts relate.

Table 6. Integration and Functionality—DCRI Technology versus SoC, SiP, and SiB

Transistors in System on Chip (SoC)	ICs in System-in-Package (SiP)	Boards in System-in-Board (SiB)	Data Centers Rack Integration (DCRI)
Integration Level: At the fundamental level, transistors are integrated into SoCs. An SoC is a single chip that integrates all the components of a computer or other electronic system, including the CPU, GPU, memory, and other critical functions, onto a single chip. This high level of integration allows for greater performance, efficiency, and miniaturization.	Integration Level: Moving up a level, multiple ICs are integrated into a SiP. A SiP packages several ICs, potentially including an SoC, memory, power management, and other chips, into a single package. This approach allows for a compact and efficient design that can perform more complex functions than a single IC.	Integration Level: At the next level, multiple SiPs, ICs, or individual components are integrated onto a board. A System-in-Board (SiB) integrates all necessary subsystems onto a single board, creating a complete system that can be easily integrated into a larger framework. This board may include processing units, memory, storage, power management, and I/O interfaces.	Integration Level: At the highest level of integration, multiple boards are integrated into a rack, and multiple racks form the infrastructure of a data center. A rack in a data center is analogous to a large-scale SiB, where each board represents an individual system (e.g., a server, storage unit, or networking device).
Functionality: The integration of millions or billions of transistors into an SoC enables the chip to perform a wide range of functions, from processing data to managing power and communication. The design and arrangement of these transistors are optimized for specific tasks, allowing the SoC to deliver high performance while consuming less power.	Functionality: SiP solutions are used in applications where space is constrained but performance needs are high. The integration of multiple ICs in a single package reduces the need for external connections, improves signal integrity, and often enhances overall system performance.	Functionality: An SiB provides a modular and scalable solution that can be used in various applications, from consumer electronics to industrial systems. By integrating all necessary components onto a single board, SiBs simplify the design and manufacturing process and allow for easy upgrades and customization.	Functionality: The rack serves as the backbone of a data center, housing multiple boards (servers or other systems) that work together to deliver large-scale computing power, storage, and networking capabilities. Racks are designed to be scalable, allowing for the addition of more boards as needed to increase the capacity and performance of the data center.

Analogy and Relationship of DCRI to SoC and SiP Technology

1. **Integration and Complexity**
   - **Transistors in SoC**: Just as transistors are integrated into SoCs to create a highly functional chip, individual boards are integrated into racks to create a highly functional data center system.
   - **ICs in SiP**: ICs in a SiP are integrated to create a complete system within a package, similar to how multiple boards within a rack form a complete, interconnected system that powers a data center.
2. **Modularity and Scalability**
   - **SoC and SiP Modularity**: SoCs and SiPs are designed to be modular, allowing for the integration of different components based on specific requirements. Similarly, racks in data centers are modular, with each rack containing boards that can be configured to meet the needs of various applications.
   - **Rack Scalability**: Just as additional transistors can be added to an SoC or ICs to a SiP for increased functionality, more boards can be added to racks to scale the capacity and performance of a data center.
3. **System Efficiency and Performance**
   - **Power and Signal Efficiency**: SoCs and SiPs are optimized for power efficiency and signal integrity, reducing latency and improving performance. Racks in data centers are designed with similar principles, optimizing power distribution, cooling, and data throughput to ensure efficient and reliable operation.
   - **Integrated Systems**: The integration of components at each level (SoC, SiP, SiB, and rack) aims to create a system that is greater than the sum of its parts. In a data center, the integration of multiple racks creates a powerful computing environment capable of handling vast amounts of data and complex computations.

Summary

The integration of transistors into SoCs, ICs into SiPs, boards into SiBs, and racks into data centers DCRI - represents a hierarchical approach to building increasingly more complex and powerful electronic systems. Each level of integration enables greater functionality, scalability, and efficiency, allowing for the creation of advanced computing environments that can meet the demands of modern technology. This hierarchy reflects the evolution of electronic design, where the principles of integration and modularity are applied at every level, from the smallest transistor to the largest data center.

Globally Integrated System (GIS)

The concept of a **Globally Integrated System** (GIS) involves the seamless integration of various components and technologies to create a highly efficient and interconnected global infrastructure. The building blocks mentioned before—SoC (System on Chip), SiP (System in Package), racks, and data centers—each play a crucial role in the Globally Integrated System. (See Figure 2.)

SoC (System on Chip):

A SoC integrates all the components of a computer or other electronic systems into a single chip. This includes processors, memory, interfaces, and other essential components. SoCs are crucial for compact and efficient designs, particularly in mobile devices, IoT, and embedded systems.

SiP (System in Package):

SiP involves packaging multiple integrated circuits (ICs) into a single module, allowing for a high level of integration without needing to place everything on a single chip. SiPs are used in applications where space is at a premium, such as in wearables, smartphones, and other compact devices.

Racks:

Racks house servers, networking equipment, storage systems, and other critical IT infrastructure. They are the physical building blocks of data centers, providing the necessary structure and cooling systems to keep the equipment running efficiently.

Data Centers:

Data centers are the central hubs for computing, data storage, and networking. They are the backbone of cloud computing and many global services, offering the infrastructure needed to process, store, and transmit vast amounts of data across the globe.

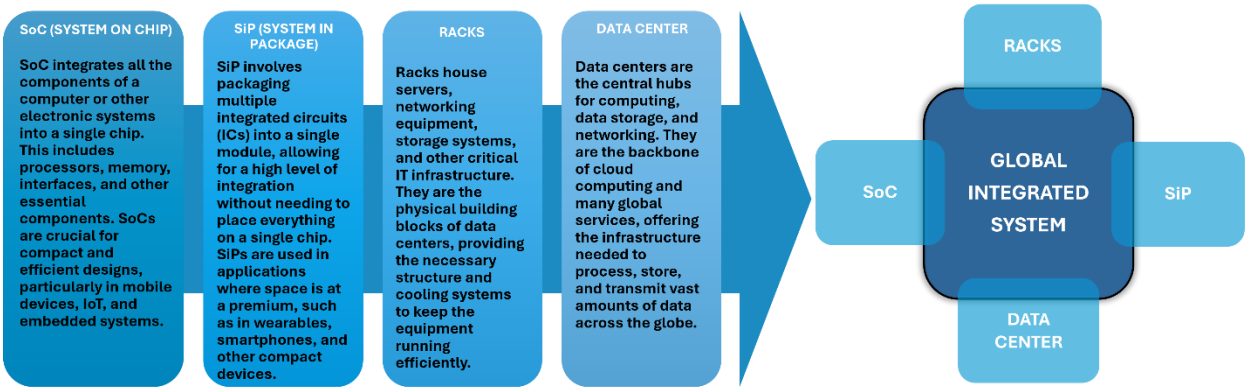


Figure 2. Relationship of SoC, SiP, Racks and Data Centers as a Global Integrated System

Together, these components contribute to the development of a Globally Integrated System by enabling efficient data processing, storage, and communication across different regions and applications.

Conclusions

Packaging of IC was deemed necessary from the very beginning to protect the frail silicon chips during handling. It provided also an easy way of electrically and thermally connecting the packaged ICs to the other electrical components located on the same PCB board. This relatively passive role of packaging remained the main function of packaging for several decades. The continuous relentless increase in the number of transistors in accordance with Moore's Law led IC manufacturers to progressively produce larger die sizes.

The size of HPC die has now reached the limits of available field in optical lithography. To maximize IC performance for HPC application requires the inclusion of multiple dice in a single package to obtain a viable silicon solution. In addition, to mitigate negative effects of signal propagation delay of interconnect lines it is advantageous to place memory chips in proximity of logic chips and by so doing reducing the relative distances die-to-die from centimeters to millimeters.

Additionally, by stacking memory chips on top of logic chips the interconnection length can be further reduced to hundreds of microns or even less. Finally, performance of HPC can be maximized in SIP by extracting the heat generated by the ICs by means of the package and by so doing the operating the ICs beyond conventional limits. Both air and water cooling have become practical technologies.

Closing Remarks

In the past, systems were composed of multiple sub-systems and each of them contained various components that were individually optimized. However, it became evident in time that this approach left substantial inefficiency and

a suboptimal use of the different subsystems and components in contributing to the overall system functionality. The integration of transistors into SoCs, ICs into SiPs, boards into SiBs, and racks into data centers DCRI -represents a hierarchical approach to building increasingly more complex and powerful electronic systems designed “as a whole” and no longer as an agglomeration of semi-undependably designed subsystems. In the new fully holistically system design approach, each level of interconnected integration enables greater functionality, scalability, and efficiency, allowing for the creation of advanced computing environments that can meet the demands of modern society. This hierarchy reflects the evolution of electronic design, where the principles of integration and modularity are now applied at every level, from the smallest transistor to the largest data center with the goal, from the very beginning, of co-optimizing from bottom up and top down a symbiotically connected single entity in a way not too dissimilar from how the human body is interconnected and operates.

Glossary/Abbreviations

Term	Definition
3DIC	3 dimensional IC
ABF	Ajinomoto build-up film
ACF	Anisotropic conductive films
ADAS	Advanced driver-assisted system
AI	Artificial intelligence
AiP	Antenna in package
AlN	Aluminum nitride
ASIC	Application specific IC
BC	Backside clocking
BeO	Beryllium Oxide
BGA	Ball grid array
BGA-HS	Ball grid array with heat spreader
BiCMOS	Bipolar CMOS
BOM	Bill of materials
BPD	Backside power distribution
BT	Bismaleimide triazine
C4	Controlled Collapse Chip Connection
CD	Clock distribution
CDN	Clock distribution network
CMOS	Complimentary metal-oxide semiconductor
CMP	Chemical-mechanical polishing
COB	Chip on board
CoWoS	Chip-on-wafer-on-substrate
CPU	Central processing unit
CSBGA	Chip-scale BGA
CSP	Chip-scale package
DCRI	Data center rack integration
DIP	Dual inline package
DRAM	Dynamic random access memory
DR-BGA	Dual-row BGA
DSP	Digital signal processor
E/O	Electrical to optical
EMI	Electromagnetic interference
EMIB	Embedded multi-die interconnect bridge
EPT	Executive Packaging Tutorial
EUV	Extreme ultraviolet
eWLB	Embedded wafer-level ball grid array
FBGA	Fine pitch BGA
FC	Flip chip
FET	Field-effect transistor
FOWLP	Fan-out wafer-level packaging
FPGA	Field-programmable gate array
FR-4	Flame retardant-4
GB	Gigabyte
GIS	Global integrated system
GPU	Graphic processing unit
HB	Hybrid bonding
HBM	High bandwidth memory
HDI	High-density interconnect
HPC	High-performance computing
HVM	High volume manufacturing
I/O	Input/output

Term	Definition
IBM	International Business Machines
IC	Integrated circuit
IEEE	Institute of Electrical and Electronics Engineers
IFT	International Focus Team
InFO	Integrated Fan Out
IoT	Internet of Things
IRDS	International Roadmap for Devices and Systems
ITRS	International Technology Roadmap for Semiconductors
LED	Light emitting diode
LGA	Land grid array
LLM	Large language model
LPDDR	Low power double data rate
MCM	Multi-chip modules
MEMS	Microelectromechanical systems
ML	Machine learning
MOF	Metal-organic frameworks
MPU	Multiprocessor unit
NIC	Network interface card
NOC	Network on chip
O/E	Optical to electrical
PC	Personal computer
PCB	Printed circuit board
PCM	Phase-change material
PDN	Power delivery network
POP	Package on package
QFN	Quad flat no lead
QFP	Quad flat package
R0	Thermal resistance
RDL	Redistribution layer
RF	Radio frequency
RoHS	Restriction of Hazardous Substances
SAC	Sn/Ag/Cu
SBGA	Standard BGA
SiB	System in board
SiC	Silicon carbide
SiP	System in package
SMD	Surface mount device
SMT	Surface mount technology
SoC	System on chip
SPD	Signal propagation delay
SRAM	Static random access memory
SSD	Solid state drive
STA	Static timing analysis
STCO	System technology co-optimization
Ta	Ambient temperature
TB	Terabyte
TGV	Through glass via
TIA	Transimpedance amplifier
TIM	Thermal interface materials
Tj	Junction temperature
TSMC	Taiwan Semiconductor Manufacturing Company
TSV	Through silicon via
TWG	Technology Working Group
UV	Ultra violet

Acknowledgements

The author is grateful for the information received in many IEEE conferences, many publications and videos freely and readily available about the subjects of the EPT.

References

1. kochi-tech.ac.jp, https://www.ele.kochi-tech.ac.jp/hoshino/

2. NVIDIA Grace CPU Superchip, https://www.nvidia.com/en-us/data-center/grace-cpu-superchip/.

3. IEEE International Electron Devices Meeting (IEDM) series

4. IEEE International Solid-State Circuit Conference (ISSCC) series

5. IEEE Symposium on VLSI Technology and Circuits series

6. IEEE Hot Interconnects Symposium series

7. IEEE Hot Chips Symposium series

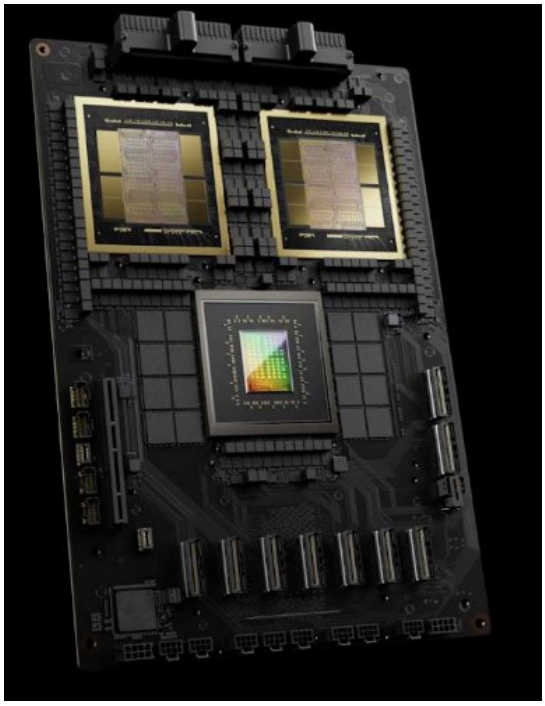
Appendices

Appendix A-NVIDIA Bella Board Attributes

Key Features and Components

1. **NVIDIA GPU Integration**
   - **High-Performance GPU**: The Bella Board likely integrates one of NVIDIA's advanced GPUs, such as those from the A100 or H100 series, known for their exceptional performance in AI and ML tasks. These GPUs provide the computational power needed for deep learning, neural network training, and inference workloads.
   - **CUDA Support**: The board supports CUDA, NVIDIA's parallel computing platform and programming model, allowing developers to accelerate computational tasks by harnessing the power of the GPU.
2. **CPU and Memory**
   - **High-Performance CPU**: The board includes a powerful multi-core CPU, named Grace. This CPU combines 72 high-performance and power-efficient Arm® Neoverse™ V2 cores, connected with the NVIDIA Scalable Coherency Fabric (SCF) that delivers 3.2TB/s of bisection bandwidth–double that of traditional CPUs to deliver maximum performance, while maintaining full compatibility with the Arm ecosystem. "Grace is the first data center CPU to utilize server-class high-speed LPDDR5X memory with a wide memory subsystem that delivers up to 500GB/s of bandwidth at one fifth the power of traditional DDR memory at similar cost."[2] The CPU works in tandem with the GPU, enabling efficient processing of both sequential and parallel tasks.
   - **Large Memory Capacity**: The Bella Board is equipped with significant RAM, often in the range of several GBs to TBs, to support memory-intensive applications such as large-scale AI models, scientific simulations, and real-time data processing.
3. **Interconnects and Networking**
   - **High-Speed Interconnects**: The board includes high-speed interconnects such as PCIe Gen4, NVLink, or even NVSwitch, allowing for fast data transfer between the CPU, GPU, and other peripherals. These interconnects are crucial for maintaining low latency and high throughput in data-intensive applications.
   - **Networking Capabilities**: With integrated high-speed Ethernet ports (potentially 10GbE or 100GbE), the Bella Board supports robust networking for distributed computing, cloud integration, and remote management.
4. **Storage Options**
   - **NVMe SSDs**: The board includes slots for NVMe SSDs, providing fast and reliable storage options that are essential for handling large datasets, AI model storage, and high-speed data access.
   - **Expandable Storage**: The Bella Board likely supports additional storage options, including SATA or U.2 drives, allowing for flexible storage configurations depending on the application needs.

5. **I/O and Peripheral Support**
    - **Extensive I/O Ports**: The board offers a variety of I/O ports, including USB, HDMI, and DisplayPort, enabling connection to a wide range of peripherals such as displays, keyboards, and external storage devices.
    - **Expansion Slots**: PCIe slots are available for adding additional GPUs, network cards, or other peripherals, providing flexibility in expanding the board's capabilities.
6. **Software Ecosystem**
    - **NVIDIA SDKs**: The Bella Board is compatible with NVIDIA's extensive software ecosystem, including TensorRT, cuDNN, and other SDKs optimized for AI and HPC workloads. This ensures that developers have access to the tools and libraries needed to optimize their applications.
    - **Linux Support**: The board likely runs a Linux-based operating system, providing a stable and customizable environment for development. This OS support includes drivers and libraries tailored to N VIDIA hardware.



Source: NVIDIA

Figure 23. Image of Bella board with 2 Blackwell GPUs and 1 Grace CPU

Applications and Use Cases

5. **AI and Machine Learning**
    - **Model Training and Inference**: The Bella Board is ideal for training large AI models, including deep learning and neural networks. Its powerful GPU,

combined with extensive memory and fast interconnects, ensures that even the most demanding models can be trained efficiently.
    - **Edge AI**: With its high-performance capabilities and potential for integration into larger systems, the Bella Board can be used in edge AI applications, where AI models need to be deployed in real-time, often in remote or distributed environments.
6. **High-Performance Computing (HPC)**
    - **Scientific Simulations**: The board's computational power makes it suitable for running complex scientific simulations, including weather modeling, fluid dynamics, and molecular dynamics.
    - **Data Analytics**: The Bella Board can handle large-scale data analytics tasks, processing vast amounts of data quickly and efficiently, which is essential for applications in finance, healthcare, and research.
7. **Autonomous Systems**
    - **Autonomous Vehicles**: The Bella Board could be used in the development of autonomous vehicle systems, where real-time processing of sensor data and decision-making algorithms are critical.
    - **Robotics**: The board's high performance makes it suitable for advanced robotics applications, including industrial robots and drones, where AI-driven control systems are needed.
8. **Virtual Reality (VR) and Augmented Reality (AR)**
    - **Content Development**: The Bella Board can be used to develop and render VR and AR content, where high frame rates and low latency are essential for immersive experiences.
    - **Simulation and Training**: The board's capabilities make it ideal for simulation and training environments that require high-fidelity graphics and real-time processing.

NVIDIA Bella Board Advantages

- **Versatility**: The Bella Board's combination of high-performance CPU and GPU, along with extensive I/O and networking options, makes it suitable for a wide range of applications, from AI to HPC and beyond.
- **Scalability**: The board can be scaled to meet the needs of various applications, whether it's through adding more GPUs, expanding memory, or integrating additional peripherals.
- **Software Integration**: NVIDIA's software ecosystem provides robust support for development, ensuring that the board can be effectively used in a variety of complex environments.

Challenges and Considerations

- **Cost**: Given its high-performance components, the Bella Board is likely to be expensive, which may limit its accessibility to smaller organizations or developers.
- **Power Consumption**: The board's high-performance nature means it may consume significant power, which could be a consideration for applications where power efficiency is critical.

- **Complexity**: The advanced features and capabilities of the Bella Board may require a steep learning curve, particularly for developers who are new to HPC or AI development.

Conclusion

The NVIDIA Bella Board is a powerful example of a platform designed for high-performance computing and AI machine learning applications. Its combination of cutting-edge GPU and CPU technologies, coupled with extensive I/O and networking capabilities, makes it a versatile tool for developers working on advanced computational tasks. While it comes with challenges such as cost and complexity, its potential to accelerate development and enhance performance in demanding applications makes it an asset in the fields of AI, HPC, autonomous systems, and beyond.