**Understanding Semiconductors – A Technical Guide for Non-Technical People**

**Chapter 1 Semiconductor Basics**

The SCI is one of the most valuable industries in the world, driving advancements in enormous sectors like auto manufacturing, financial services, and medical devices. SCT is a primary driver for product differentiation in many markets including smartphones, industrial robots, and pacemakers. In 1965, Gordon Moore, founder of Intel and ***Moore’s Law***, predicted that transistor density and its associated computing power would double every two years – from the first hand-sized transistor developed by William Shockley, John Bardeen, and Walter Brattain at Bell Labs in 1947 to the 3nm transistors today, this doubling has happened every 18 months on average. This **“Digital Revolution”** has produced countless technologies we depend on today and spawned massive industries and innovations. In the US alone, the SCI, which is the third largest exporter after oil and aircraft, directly employs 250K people with an additional 1M indirect jobs.

**A Quick Semiconductor History**

In 1947, scientists used the semi-conductive properties of silicon to build simple **transistors** which served as switches for electric current. By arranging transistors in intricate patterns, they could selectively guide current along directed paths. For a decade thereafter, transistor manufacturing was slow, cumbersome, and costly. Transistors were manufactured as individual units and then interconnected by “**flywire**” to form circuits that filled up entire rooms. This all changed in 1959 – the year that marks the start of the **semiconductor revolution** – with two breakthroughs that continue to underpin the SCI’s design-based and manufacturing-based value chain to this very day:

1. **Integrated** **circuit** – invented by Jack Kilby at Texas Instruments and Robert Noyce at Fairchild Semiconductor, allowed hardware designers to fit many transistors onto a single chip.
2. **Planar manufacturing process** – invented by Jean Hoerni at Fairchild Semiconductor, allowed chip companies to simultaneously fabricate numerous components on a single substrate.

**Semiconductor Value Chain – Our Roadmap**

1. **Customer needs / market demand** – assessing the need for a product; customers may not always “know” what they need.
2. **Chip design** – a firm must design a chip that will fit a product using **front-end design** and **back-end** **design**.
3. **Fabrication and manufacturing** – numerous ICs **(dies)** are printed onto a silicon wafer using **photolithography**.
4. **Packaging and assembly** – after dies are cut, they are individually packaged in the **assembly process**. These **package-die assemblies** are tested before shipping them to systems / product companies (**SPC**).
5. **System integration** – package-die assemblies are soldered onto larger circuit boards or substrates along with other components and ICs, forming an integrated “consumer-ready” system or product.
6. **Product delivery** – the final product is shipped to the end customers.

SCCs generally focus on Steps 2 – 5 with “Fabless” SCCs concentrating on Step 2. Each step in the SCVC is essential to get a chip from concept to customer. At every stage, companies fight for profits and market share by trying to balance SCVC challenges better than their competitors. Since the 1940s, the SCVC has remained relatively stable. However, SCVC business strategies, driven by SCCs competing to provide the best performing chips and highest quality products, have been very dynamic.

**Performance, Power, Area, and Cost (PPAC)**

For SCCs focused on Steps 2 – 5, the goal is getting the **highest performance (clock frequency, speed)** using the **lowest** **power (watts, longer battery life, heat constraints)** within the **smallest area (nm, less power, space constraints)** while **minimizing cost (design, manufacturability, yield risk)** and **time-to-market**. An optimal design requires balanced PPAC, but time is also critical – lower performance may be worth it if design time can be reduced in order to be first to market. Balancing PPAC poses critical trade-offs in context to the product or application – for example, requiring high performance and small area will constrain power efficiency.

**Who Uses Semiconductors?**

The Semiconductor Industry Association (SIA), a key SCI trade group, defines six end-user segments:

1. **Computing** (32%) – PCs, office equipment and peripherals, handheld computing devices, servers.
2. **Communications** (31%) – cell phone, networking and remote access devices, base stations, broadcasting equipment.
3. **Consumer** (12%) – TV, video, audio, household appliances, and all sorts of consumer goods.
4. **Industrial** (12%) – power supplies, IoT devices, manufacturing test, control, measuring equipment.
5. **Automotive** (11%) – in-vehicle entertainment and information systems, power train, controls.
6. **Government** (1%) – military, aerospace electronics.

While computing and communications accounted for the greatest proportion of SCI sales (63%), automotive, industrial, and instruments are expected to increase revenue share as cars become more electrified and industrial operations become more automated. Each of these segments has different PPAC requirements and drivers based on their unique purpose, and this impacts each stage of the SCVC.

**Chapter 2 Circuit Building Blocks**

**Discrete Components – The Building Blocks of Circuits**

**Integrated circuits** – instead of manufacturing discrete components separately and connecting them afterwards, patterns can be etched together on the same chip using specialized manufacturing technologies like **photolithography**. This has many benefits including: higher speed, smaller area, and less power, all leading to lower manufacturing unit costs due to less material required per process run. **ENIAC**, the first digital computer, invented by **J. Presper Eckert** and **John Mauchly** at the University of Pennsylvania in 1946, was a giant room full of vacuum tubes taking up 1,800 square feet and weighing 50 tons. By comparison, ENIAC-on-a-Chip (1996) is 7.44mm x 5.29mm and demonstrates the progress made by the SCI in shrinking computer size over time.

**Transistor Structure, How Transistors Work and Are Used**

Modern **MOSFET** (metal-oxide field effect transistors, Bell Labs, 1970s) consist of a source, gate, and drain which are analogous to the emitter, base, and collector in **bipolar junction transistors** **(BJT)**. The magic of semiconductors happens after doping silicon with impurities that either have an electron surplus (N-type) or deficit (P-type). Transistors come in two configurations **NPN (NMOS)** and **PNP (PMOS)**. The important thing to note is that the gate is a different charge from the source and the drain.

Without power and voltage, a transistor is inert with no electrons flowing through the system. However, when a positive voltage is applied to the gate (for NPN), the excess source and drain electrons are drawn to this voltage, while the positive gate charges are pushed away. This opens a **channel** through which electrons can flow, allowing current to pass through.Transistors used together form the building blocks of computer engineering – **logic gates** (AND, OR, NOT) – which are simple circuits that use Boolean logic to enable computation. Hardware engineers use logic gates to build complex systems that perform important base functions like arithmetic operations.

**Chapter 3 Building a System**

**Different Levels of Electronics – How the System Fits Together**

The most advanced ICs currently in production require as many as two trillion transistors on a single chip. To realize this, hardware designers must group and organize electronic systems across higher levels of abstraction and understand how these abstraction levels fit together. The ground level of an electronic system consists of a combination of 1) separate discrete components soldered directly onto a **printed circuit board (PCB)** with other components, like larger capacitors and inductors that make up a system’s power circuitry or 2) groups of functional components integrated onto a single die. Electronic systems can be viewed as a hierarchy:

1. **Discrete component** - transistors and other discrete components form the basic building blocks for higher-level systems.
2. **Integrated circuit** – chips ranging from CPUs to memories are designed using discrete and functional components.
3. **Packaging** – individual, and sometimes multiple, ICs are wrapped in a protective enclosure to mitigate interference from neighboring components. Multiple components may be grouped together as a **module**, which describes a bundle of smaller circuits and components that work together as a unit to perform a task.
4. **Printed circuit board** **(PCB)** – smaller components at “lower” levels (usually seen as black components) are soldered onto a PCB and connected to one another to form a larger system. PCBs mechanically support and serve as a foundation on which components are connected to one another using conductive tracks, pads, and other features etched onto its surface.
5. **System** – everything is tied together to create a fully functional system or product – laptop, phone, PC, television, etc. A system can be used to describe a fully functional and discrete structure relevant to the task it is designed for at a given level.

**Integrated Circuit Design Flow**

1. **System level architecture** – system architects develop a chip concept based on input from business and marketing. Architects need to decide what the chip will do, what technologies, materials, and components will be used, and how the team will evaluate whether the chip is a success. Architects monitor progress throughout the design process and guide their team as they move through each design step while continually interfacing with business and marketing.
2. **Front-end design (FE-D)** – before construction begins, engineers create detailed models of the system: **high-level** to **specific** to **detailed schematics**. Logic design engineers fill in the details during **RTL (register transfer level) design** using **hardware description languages** – such as VHDL or Verilog/System Verilog RTL – to describe what they want the circuit to do. HDLs are specialized “programming” languages used to describe the physical structure of ICs and electronic systems. By the end, a virtual version of a chip should, in theory, detail how it will serve its purpose flawlessly in the real world.
3. **Design verification** – engineers must verify that the output from FE-D will translate into a fully functional subsystem. This is a difficult, but critical task given the high cost of fabrication, that can consume over half of the total design time.
   * **Functional verification** – uses SystemVerilog HDL code to simulate the exact behavior of the circuit in any possible condition. **Universal verification methodology** is a relatively new way to verify numerous possible conditions. UVM even has the capability to collect statistics on which portions of the design have and have not been verified. Any flaws in the **golden model** should give erroneous outputs from testbench inputs that are subsequently debugged. After debugging, the golden model becomes the de facto implementation to be sent to manufacturing.
   * **Emulation** – is an alternative, or supplemental, methodology that uses an **FPGA** **(field programmable gate array)** **emulator** to input a design and observe its real-world output. For instance, listening to audio processor design output directly through a speaker. FPGA do not produce ready-for-market design, but rather a prototype as close to the real deal as possible. Although the design output is more tangible, emulation is clunky and expensive.
4. **Physical design** – an incredibly complex and time-consuming process, sometimes as long as FE-D and verification, that uses advanced **EDA (electronic design automation) tools** to convert the de facto model into wires and transistors.
   * **High Level Synthesis (HLS)** – marks the end of FE-D and the beginning of **back-end design (BE-D)** where RTL code is converted into transistors and wires using **synthesis**[[1]](#footnote-2) which can be incredibly complicated. Engineers will often tweak the de facto model during HLS before moving to the next step.
   * **Design Netlist** – comprises a list of electronic components in a circuit and all the nodes to which they are attached.
   * **Floorplanning** – engineers decide where everything should be located with PPAC optimization in mind.
   * **Place-and-Route** – engineers decide exactly where to put all the electronic components and circuitry using a placement tool to allocate specific locations for every logic gate. Routing follows placement where CAD tools integrate all the wiring required to connect the placed components.
   * **Clock-Tree Synthesis (CTS) –** engineers make sure that electrical signals throughout the chip “clock” evenly and as intended. Clock frequency, or clock rate, a common measure of process speed and performance, or instruction execution rate, assesses how quickly a signal travels through an IC. **Synchronous design** uses a common “clock” across all circuits. All parts of the circuit processing instructions at the proper clock edge, or capture edge, are critical to proper system function otherwise failure or performance degradation can result. A common simulation method to compute and verify the expected timing of digital circuits is called **static timing analysis (STA)** that ensures that all logic paths are properly timed relative to each other so there is never a chance of a timing error.
5. **Validation (physical verification or back-end verification)** – a GDS design file for a fabrication-ready chip is generated that includes all the information required by a fab. Prior to shipping the GDS file, validation engineers need to double check that the chip is manufacturable by using **design rule checkers (DRC)**, that are components of EDA tool suites, to verify that the chip complies with the chosen fab’s rules – chip size, wire and component proximity, etc.
6. **GDS II generation (GDS II)** – is a standardized format for the design that is sent to the fab at the end of the design cycle, **RTL-to-GDS** or **tapeout**. This stage is reached after the culmination of millions of dollars of engineering ingenuity. Manufacturing is neither simple nor cheap with modern processes taking over 12-16 weeks from GDS to finished wafer.

**EDA Tools**

EDA, or **e-CAD**, includes tools that are used by hardware engineers and chip designers throughout the design process. These tools automate and streamline everything from functional verification to HLS and have reduced the cost and difficulty of an arduous and complicated design cycle. EDA companies[[2]](#footnote-3) helped develop and spread VLSI HDLs like Verilog and VHDL, which automated design flow and enabled companies to refocus on other areas of product development. The EDA tool market was worth about $11B in 2020 and is expected to grow to over $21B by 2026 (Mordor Intelligence, 2021). A good design on its own is no more useful than a drawing of your dream home – we’re halfway there – but we are still a long way away from walking through the front door.

**Chapter 4 Semiconductor Manufacturing**

**Manufacturing Overview**

SC manufacturing (**SCM**) starts after a fab receives the GDS II file and includes **front-end manufacturing (FE-M)**, **final assembly**, and **testing**. SCM is a highly complex and ultra-precise process that requires specialized chip factories. Key metrics of chip manufacturing are die-area and batch yield. Within the industry, a particular manufacturing process is called a **node**, **technology node**, or **process node**,and refers to the minimum feature size usually specified by the source-drain distance, or equivalently the gate length. The most advanced node is the 2nm node which Samsung plans to roll out in 2025. It is instrumental to map out global manufacturing capacity by node and chip segmentation (***Figure 4.1, page 59***) where chip segmentation is broken up into memory, logic, and analog + OSD (analog, optoelectronics, sensors, and discrete components).

A **wafer run** is a single run-through of the SCM process, from initial **wafer fabrication** to **wafer dicing**. SCM is divided into two parts: FE-M, which puts the desired circuit onto a silicon wafer, and **back-end** **manufacturing** **(BE-M)**, which gets the individual chips on that wafer ready for a customer’s system.

**FE-M**

The first step in FE-M is **wafer fabrication**. A wafer is a thin slice of **substrate**, formed by melting silica and carbon, and is cut from an **ingot**. In the 1980s wafers were 150mm, today 300mm, with 450mm in the works. FE-M, comprised of four major steps, is used to build intricate combinations of substrates, circuitry, and other materials in layers with extreme precision and accuracy.

1. **Deposition** – involves addition of thin films onto a wafer’s surface by equipment that is very intricate and expensive.
2. **Patterning / Lithography** – shapes and alters the material on the wafer. **Photoresist** deposited on the wafer breaks down after reacting with light of a **specific wavelength**, that passes unimpeded through a mask – complex process can require as many as 75 masks. These photomasks are aligned using a **stepper**, which also functions as a light source. **Electron beam (e-beam) lithography** is another method used to manipulate photoresist. Once the photoresist has been removed, metal or other materials can be deposited into the remaining areas to form wires that connect individual transistors and functional features. ***Lithography is a critical bottleneck in FE-M***.

Lithography has been the key to keeping the pace of geometric scaling predicated by Moore’s Law since each generation of equipment enables smaller features and increased transistor density. To move to more advanced nodes, lithographic equipment suppliers have had to continually find new and creative ways to make ever smaller patterns and transistors. One tactic has been to use shorter wavelength light since wavelength serves as the lower limit to feature size. Optical workarounds and the use of multiple photomasks have allowed fabs to etch smaller patterns at a given wavelength. However, as size keeps decreasing, these workarounds have become increasingly difficult and promoted EUV as the hotbed technology of the future.

1. **Removal** – **wet etching** or **dry etching** and **chemical mechanical planarization (CMP)** are used to wash away photoresist that is no longer needed, leaving an area that can be filled later with the desired metals, oxides, transistors, and passive components.
2. **Physical Property Alteration** – doping (ion implantation / introduction), rapid thermal annealing, ultraviolet light processing, and other techniques are used to modify the electrical and physical properties of all components on a wafer.

**Cycling – Pre- and Post-Metal**

These four process steps are repeated until enough layers are formed on the wafer, with some high-end chips requiring hundreds of repeat steps per wafer run. This repeating cycle can be implemented across both **FEOL (front-end of the line, pre-metal processes)** and **BEOL (back-end of the line, post-metal processes)**. Modern devices have as many as 15 stacked layers connected using **through silicon via (TSV)**. Additionally, there are lower-level local interconnects and upper-level global interconnects. FE-M for a complex wafer can require several dozen mask layers and take 12–16 weeks. These challenges are reflected in the industry’s cost distribution, with FE-M machinery comprising 60% of the $62B in 2020 spent on SC production equipment.

**Wafer Probing, Yield, and Failure Analysis**

Each successive node brings added complexity which makes wafer probing, yield, and failure analysis critical for meeting production goals and keeping unit costs low. Wafer probing is the process after FE-M and before BE-M that uses a wafer prober to check if a wafer actually works by electrically testing the wafer dies. Wafer probing is done in instances where BE-M is lengthy and expensive. In **chip-scale packaging**, the entire wafer is packaged, and testing is done afterwards.

There are two testing modes: 1) **parametric testing** of the wafer fabrication process designed to check if the entire process is working by using **spacers** and **scribe lines** to see if basic parameters such as resistances and device thresholds are within tolerance and 2) **wafer testing** to ensure each individual die is defect-free and fully functional. Wafer testing is also used for measurement performance, monitoring for recurring errors, and identifying issues with fab equipment that may be contributing to systemic failures in wafer runs so that the equipment can be tuned to improve subsequent yields.

Testing allows failure analysis engineers to derive the fabrication process’ 1) **line yield**, or **wafer yield**, which measures how many wafers, as a whole, successfully make it through wafer probing without having to be thrown out and 2) **die yield**, which measures the number of functional dies divided by the total number of potential dies that make it through wafer probing. The **end-to-end yield** is the composite of line and die yield and holistically accounts for FE-M efficiency.

For a new manufacturing line, yield usually starts lower at around 50% for advanced processes, and gradually increases as engineers recalibrate equipment and tweak other process steps. **Yield optimization** is one of the most critical FE-M metrics because even small yield increases can drive down manufacturing costs and boost margins thus conferring significant competitive advantage (***see page 72***). Yield is usually directly proportional to the number of dies on a wafer because tiny contaminants, mitigated by air filtration outfitted **clean rooms**, or slight vibrations, mitigated by spring mounted or air suspension systems, can permanently ruin a given die. Smaller die sizes typically result in higher yields, since failure is more likely to be contained in a smaller portion of the overall wafer area. Defective dies are marked and either discarded or sold at a discounted price - **inking**.

Specialized air purification, other construction requirements, and persistent retooling add enormous cost to a fab’s price tag with a single 3nm fab ranging from $6 - $20B and becoming obsolete within 5–6 years. Capex for US SCCs amounted to roughly 30% of sales in 2020 compared to 4% for the rest of the manufacturing sector. Older fabs can sometimes be sold “down-market” to mixed-signal or analog companies that, in most cases, don’t need to be at the bleeding edge of the fabrication technology curve.

**BE-M**

The BE assembly and testing process kicks off when viable dies are ready for IC packaging. BE-M is usually done by third party **Outsourced Assembly, Test, and Packaging Suppliers (OSAT)** that are largely based in East Asia where they enjoy significant labor cost advantages.

1. **Wafer bumping** – is not always performed, but in cases where the bare die is connected directly to other components, small **bumps** are soldered directly onto the wafer.
2. **Wafer dicing** – individual dies are cut from the wafer, after which they are ready to be shipped to an OSAT.
3. **Die-attach bonding and interconnects (wire bond vs. flip chip)** – at the OSAT, cut dies are either attached to a packaging substrate or PCB (**wire bond**) or packaged as bare dies (**flip chip**). Of the two die-attach methods, **epoxy wire bonding** is the most common, whereas **flip chip bonding** is less used but more advanced. Once attached, dies are linked to the rest of the system through **interconnects (I/O)**. Wire bonding results in fewer peripheral I/O points relative to flip chip which involves soldering the flipped die to a **ball grid array** or directly to the PCB resulting in higher I/O density and faster speeds.
4. **Encapsulation and sealing** – surface mount technology is used to mount the die onto the IC package enclosure. A **transfer molding** machine then heats **encapsulant compounds** or **molded underfills** before injecting them into the packaging mold, sealing in the completed **die-package assembly** which is then ready for final testing.
5. **Final testing** – the die-package assembly is tested a final time before shipping it to the end customer or integrating it into an intermediate system or product. In some well-established technologies with very high yield, usually greater than 90%, final testing may be the only time in the entire front-to-back manufacturing process where testing is conducted.

**Semiconductor Equipment**

Figure 4.17, page 78, displays the manufacturing cost distribution of $64B in 2019 sales broken down into 11 SC equipment categories with FE-M accounting for 86% of overall sales. As manufacturing difficulty continues to grow due to shrinking chip size, FE-M equipment cost will continue to increase. This is especially true for core FE-M technologies responsible for delivering increasingly smaller patterns for deposition, lithography, and removal. Figure 4.18, page 79, is an excellent summary of end-to-end SCM.

**Chapter 5 Tying the System Together**

**What is a System?**

A well-designed IC is only as good as the system with which it’s integrated. System integration has become increasingly important as Moore’s Law slows down and companies lean more heavily on functional scaling. Many devices consist of different ICs and components that are often designed by different companies using proprietary methods and unique microarchitectures. Thus, integration becomes a complicated and difficult task since connection points between modules can act as data flow bottlenecks that lead to system-wide latency. A company may source and design the perfect assortment of ICs and components to create a market-leading product, but having the right parts is not sufficient. Building high caliber devices requires tight system integration with plenty of interconnects, the right IC packaging architecture, and strong signal and power integrity.

**Input / Output (I/O) and IC Packaging**

Within a single chip, interconnects link different components together to form logic gates and other functional blocks. Large chips have as much as 30 miles of stacked interconnects. At a higher level, interconnects refer to the connections between chip and PCB, other components, and other parts of the system. Thus, designers have many different options for I/O within a system and this makes proper packaging for each IC an important consideration. Electronic packaging protects the chip from the outside world and supports the different I/O endpoints. Packaging options including:

1. **Wire bond** – in the early days, IC packaging was limited to peripheral wires connected from IC bonds and landing pads inside the packaging to pins soldered onto the external board or substrate that supported the system.
2. **Flip chip** – this option solves the I/O limitations of wire bonding by enabling I/O to be placed within the borders of the die itself.
3. **Wafer-level packaging** – or **chip-scale packaging**, is done before the wafer is diced, resulting in a smaller die-package assembly that approximates the chip size. However, this option only works for small die sizes.
4. **Multi-Chip-Modules (2D) and System-in-Package (3D)** – integrate multiple dies into a single package and are useful for small form-factor applications. Both allow engineers to modulate parts of the design and more easily incorporate **licensed IP**. However, mixing and matching multiple dies as opposed to a **system-on-chip (SoC)** approach, which involves different functional units in a single die, leads to performance and power disadvantages. MCM / SIP architectures allow using cheaper silicon processes for the analog functions while using expensive lower-geometry processes only for the critical high-speed processing and memory functions within a system. SIPs also allow for passive devices like capacitors and inductors to be integrated into a single package which can improve performance by minimizing the distance between components. The tradeoffs between SoC chip-level **monolithic integration** and package-level **heterogeneous integration** present key challenges to system architects today.
5. **2.5D / 3D packaging** – advances such as **die stacking** have allowed multiple die layers to be interconnected using TSVs. 2.5D dies are linked on a shared substrate separated by interposers that are connected to a PCB, whereas 3D dies are stacked directly on top of one another. 2.5D is not as tightly integrated as 3D but is less costly. However, 2.5D does provide tighter integration relative to wire bonding. New hybrid bonding technology uses Cu-Cu interconnects on stacked die resulting in greater density and lower resistance, enabling better performance versus TSVs. 2.5D / 3D has also enabled integrating modules using hybrid silicon processes. For example, a 22nm processor chip and a 180nm high-power audio amplifier can be combined into a single plastic module. Stacking technology was first applied to memory systems like hybrid memory cube and high bandwidth memory in a memory-on-logic and memory-on-memory mode. Stacking chips vertically enables increased I/O density thus conferring improvements to one or more PPAC metrics. This technology is especially useful for compact devices like smartphones.

While comprising a relatively small portion of total system cost – $30B out of $440B of 2020 sales, IC packaging has an outsized impact on system performance and has seen increasing emphasis as engineers try to squeeze more performance from each node.

**Signal Integrity**

As component density gets stretched, signal integrity has become increasingly important. In complex circuits where wires are a few nanometers apart, neighboring signals can interfere with one another and the surrounding environment. These **transmission line effects** can result in data loss, accuracy issues, and system failure especially at high bit rates and long distances. Signal integrity engineers can mitigate these effects by conducting electromagnetic simulations and analysis to identify and resolve common forms of interference including **noise**, **crosstalk**, **distortion**, and **loss** – **resistive**, **dielectric**, **radiative**.

**Bus Interfaces**

A key performance bottleneck in electronic devices is the transmission of data between system components. To unlock the power of advanced circuits, bus interfaces, the physical wires through which data travels within a system, have become increasingly important. Buses have three primary functions: 1) **data bus** transmits data, 2) **address bus** finds data, and 3) **control bus** coordinates operations of different parts of the system. Collectively the **system bus** coordinates data flow to and from a microprocessor, memory, and I/O.

To improve speed and performance, engineers developed an integrated structure that reduced the number of I/O junctures formed by haphazardly connected components and modules down to two chips – **Northbridge** and **Southbridge** **chipset**. Northbridge interfaces directly with the CPU via a **front-side bus** **(FSB)** and “bridges” it with components that have the highest performance requirements (PCI-E and memory). Northbridge also connects, via the **I/O Controller Hub**, to Southbridge which in turn connects to all lower priority components and interfaces such as Ethernet, USB, and other low speed buses, collectively the **peripheral buses**.

Buses are classified by the way they transmit data between components. Parallel interface buses use multiple wires to transmit data simultaneously, however, this works well at short distances. Common **parallel buses** include DDR and PCI. Serial interface buses transmit one bit at a time but at much higher rates. Serial buses exhibit better signal integrity due to less **crosstalk**, however, are susceptible to **intersymbol inference** where bits can be affected by prior bits. Common **serial buses** include PCIe, USB, SATA, and Ethernet Bus. Serial buses are less costly, however lower wire count reduces overall transmission rates. Parallel buses, by contrast, enable faster data communication, but are costlier with limited range at high frequencies.

**Power Flow in Electronic Systems**

Harnessing the power of “power” spans a vast field of engineering with numerous sub-disciplines. Power starts with a **voltage source** in the form of a battery or power line. A **power converter** converts AC power, which is better for long distance power transmission, into DC electric current, which is better for short distance power transmission. Next, voltage is transported through a **Power Distribution Network** to the different processing centers. Power engineers must build a network of **voltage regulators** – circuits that maintain fixed voltage output regardless of input voltage – and **power converters** to ensure that voltage is never too high or too low at any point in the system. There are a variety of voltage regulators including DC/DC, power management units, Buck, and Boost and Flyback. **Power integrity** studies the flow of voltage throughout a system to ensure that voltage gets to the right place, in the right amount, at the right time.

**Chapter 6 Common Circuits and System Components**

Thus far we have treated SCs as a monolith devoid of differentiating features. We will now break apart this monolith and explore the numerous types of common circuits and system components that comprise the SC family.

**Digital vs. Analog**

There are two main types of components that get their name from whether they use **digital** or **analog** signals. **Digital signals** are usually synchronous, that is, they run on a reference clock that coordinates the processing of different functional blocks and ensures proper timing. Though predictability and synchronous timing makes them great for storing and processing information, digital circuits are unable to transport information over any sort of distance without physical wiring.

**Analog devices** process information continuously as a range of values and their ability to capture and transmit electromagnetic energy makes them well suited for applications like wireless communication. Much of the energy from real-world signals is analog in nature. Analog signals are distinguished by their **frequency** (**Hz**) which is inversely related to **wavelength** and directly related to **power**. By receiving and modulating different frequencies, analog electronics can do many things from **sensors** that detect stimuli to RF technology that transmit and communicate data wirelessly. Analog and digital components often work together to translate real-world analog signals into digital signals that computers process and then reverse translate computer digital signals back into human comprehensible analog signals. Mixed-signal devices, called **ADC** and **DAC** converters, are used to convert analog to digital and vice versa.

**Common System Components – The SIA Framework**

The six end-user segments – computing, communications, consumer, automotive, industrial, and government – drive a vast, diverse, and highly competitive components market that the SIA framework breaks down into five constituent segments.

**Micro Components**

Includes all non-custom, generic, digital computation or signal processing sub-component devices that plug into other systems.

***Microprocessors (MPU), Microcontrollers (MCU), and Digital Signal Processors (DSP)***

A processor is a chip that receives and processes input and produces output. MPUs encompass complex digital circuits, like CPUs, that connect to larger systems in order to provide general computing functions. MPUs require an external bus to connect to memory and other peripherals components. PCs and servers account for the largest fraction of MPU sales.

MCUs perform specific functions and are integrated with memory and I/O all on one chip. Microcontrollers are smaller and less powerful and provide plug-and-play computing for simpler operations. MCUs are widely used in low power IoT devices and embedded systems. Automotive, industrial, and computing account for most MCU sales. It is important to distinguish MPUs / MCUs in the Micro Component segment from those in the Logic segment. Logic devices are custom made for specific applications, while Micro Components are generic processing sub-components used within broader systems.

DSPs are used to process multimedia and real-world signals like audio, video, temperature, pressure, position, etc. Digital components have trouble accurately representing the real world with 0s and 1s and this is why DSPs are needed to translate real-world signals in a form that digital components can understand. The most common pathway for signals is ADC to DAC. DSPs are adept at high-speed real-time data processing and are highly programmable, which makes them easy to implement in a wide variety of devices and systems.

Micro components accounted for $69B of the $440B 2020 sales (16%). Comprising 57% of the Computing segment’s end-use applications, micro components are much more heavily weighted than the other four constituent component segments.

**Logic**

Specialized non-micro component digital logic circuitry that includes application specific ICs (ASICs), field programmable gate arrays (FPGAs), and more versatile, but application-specific digital logic devices.

***Special Purpose Logic***

SPLs are **application-specific standard parts (ASSPs)** that are designed and integrated into a system, similar to ASICs, and are used in many diverse products like Ring doorbells, LCD TVs, and game consoles. Standardized product types such as I/O circuits for USB and PCIe are also classified as ASSPs. SPLs range from wireless controllers like Ethernet and WLAN, modem SoCs, image and audio processors, PC core logic, and GPUs. On the other hand, customized chips called ASICs are specifically designed and optimized for a single device or system. High-volume consumer products like iPhones utilize many different custom chips that are intended to squeeze out every last bit of performance. For example, Samsung designs an ASIC CPU for its smartphone and AMD’s ASIC GPU powers Xbox. This is different than generalized application ICs, like Intel’s server-based CPUs, that are used in lots of different data center computers.

***Central Processing Unit***

A CPU is a special type of microprocessor most commonly used in PCs and laptops, but as a general class are not limited to these end-products. CPUs are found in all sorts of products like smart speakers, automotive control systems, and any device that processes information – they are even found in your coffee maker. CPUs are the digital brains of a system that process and execute instructions as needed. The core processing of a CPU is handled by the **arithmetic logic unit (ALU)**, which performs the numerical and logic-based operations necessary to run all sorts of software.

CPUs are typically connected to other modules through a bus or chipset that feeds information to the CPU through **registers** for processing and directs output data to memory for storage or to other system components. Each CPU has a fixed number of registers through which data can flow with typical registers having capacities measuring 8, 16, 32, or 64-bit wide. An individual CPU or GPU may be referred to as a core that can be combined with other cores to form **multi-core architecture**. Together, the CPU and other components are “integrated” onto a single IC (SoC) or a larger system. A system may include a separate CPU, memory, GPU, power source, and multimedia processor, while another system may integrate all of these into a single SoC or MCM.

***Graphics Processing Units***

GPUs are best known for driving graphics and 3D visual processing by using parallel processing instead of serial processing that is used by CPUs. GPUs can perform thousands of specialized operations using hundreds of cores, though they are not as efficient at handling more diverse operations. In summary, CPUs are better at performing diverse tasks like running all the programs and functions of a PC while GPUs excel with applications that require repetitive high-volume calculations like graphics. AI / ML represents a new frontier for GPUs since they are ideal matrix processors that can break down complex problems into smaller constituent problems giving them the ability to handle billions of small trial-and-error calculations required by AI / ML algorithms. While most of the SCI has been consolidating, innovations in AI-centric GPUs have led to a significant growth area where new companies have been able to compete. Specific applications ideal for GPUs include autonomous driving, machine vision and facial recognition, high performance computing (HPC), complex simulation and modeling, data science and analytics, bioinformatics, computational finance, and cryptocurrency mining, just to name a few.

***ASICs vs. FPGAs***

Represent two different approaches to chip design and development. ASICs operate at higher speeds, lower power consumption, smaller area, and have lower variable manufacturing costs at high volumes. However, ASICs incur high upfront capital and labor costs. Even if an ASIC is taped-out, there is always the risk of low yield and the chip not functioning the way it was intended. Furthermore, ASICs are so customized that they cannot be used in other areas, and this adds to overall cost and complexity.

FPGAs are programmable chips that can be customized to serve a specific function after they have been manufactured. Most FPGAs can be erased and re-programmed and this makes them ideal for prototyping new designs. Emulators are essentially a collection of FPGAs that allow ASIC designers to iterate their design before moving to manufacturing. Emulators are becoming more important as chip manufacturing costs increase since they can be used to verify how a given design will perform in the real world and post-fab. FPGAs are used in many applications including video and image processing, security systems, scientific instruments, wireless communications, aerospace and defense, medical electronics, and consumer electronics. For the last two decades, Xilinx, acquired by AMD, and Altera, acquired by Intel, have dominated FPGAs with 55% and 35% market share respectively.

A key decision for many companies is whether to build a custom ASIC or use an off-the-shelf FPGA since this choice involves significant PPAC tradeoffs. If you’re working with short time-to-market constraints or lower than expected manufacturing volumes, FPGAs are usually the better choice given that less stringent PPAC constraints are permissible. As volumes increase, ASIC becomes more attractive since you can spread higher upfront capital and labor costs over greater unit volume and capitalize on long-term yield improvements that reduce future costs per device. Apple, Facebook, Google, and Tesla all develop custom ASICs for their devices in-house. One option that combines FPGA and ASIC is to use FPGAs initially to prove out a new product idea and demo a solution to generate customer interest. Once that hurdle is cleared, it can be easier to secure funding and corporate buy-in for the much larger expense associated with ASIC development.

***System on Chip***

SoC are complex and highly integrated ASICs that contain an entire functional device all on one substrate. An SoC must, at a minimum, have an MPU and/or MCU, DSP, on-chip memory, and peripheral functions like hardware accelerators. For smaller applications like cell phones, multiple chips may require too much space and power; problems that can be mitigated by SoC’s tighter integration. SoCs are also making their way into non-phone devices like laptops. Differentiating between ASICs, ASSPs, and SoCs can be a bit confusing. The main difference is that ASSPs are designed to serve multiple companies and end systems, while ASICs are designed for a single use by a single company or product. ASSPs and ASICs that contain a processor are considered SoCs while those lacking a processor are not considered SoCs.

Logic devices accounted for $118B of the $440B 2020 sales (27%). Comprising 44% of the Communications segments’ end use applications, logic devices are much more heavily weighted than the other four constituent component segments.

**Memory**

Classified based on whether memory storage can be done with power (volatile memory, or RAM, enables quicker access) or without power (non-volatile, or ROM). Dynamic RAM is the most common type of volatile memory, while NAND flash is the most common type of non-volatile memory. Since the 1960s and 70s, the market need for data storage has skyrocketed, driving the demand for more advanced memory chips to new heights year after year.

***Memory Stack***

Memory’s primary function is to store data and information for use in the processing centers of larger systems. Storage capacity is no longer the dominant performance constraint but rather the bridge between memory and the core system processors has become the key bottleneck for performance. This has driven the development of new memory chips and microarchitectures.

Instructions, data, and information flow between the CPU and the memory stack starts with an input source, and its related input stimuli, and flows through the memory hierarchy for processing and storage. Input triggers core instructions to be readied by the long-term, non-volatile ROM memories, which are then sent to volatile RAM memories higher up the stack and thus closer to the CPU. These instructions are then rapidly transferred to L1 and L2 cache, which directly interface with CPU registers via a data bus. The CPU processes the data according to the instruction set and returns output instructions that can either be held in cache for quick re-use or delivered to permanent storage for use at another time. Memory designers are always balancing memory capacity with access speed. This is why designers implement a memory hierarchy with smaller faster cache used to store frequent time-sensitive operations for quick access and larger slower memories used to store broader datasets that are needed less often.

Memory is classified into two broad categories, **temporary volatile memory** vs. **permanent non-volatile memory**. Another distinction involves read / write ability with RAM allowing read / write and ROM allowing read only. In general, ROM is used for permanent data storage, while RAM is used for running programs and storing temporary data closer to the CPU. RAM utilizes CPU registers that serve as the bit interface for physical data transfer between memory and the processor.

***Volatile Memory***

The most common volatile memory is DRAM, often used as temporary working memory, and SRAM, often used as cache memory. DRAM holds more data and uses less power than SRAM but is slower – a classic power-to-performance tradeoff. Computing throughput is also limited by transfer speeds between DRAM and SRAM, a bottleneck that has been mitigated by double data rate (DDR) SDRAM.

***Non-Volatile Memory***

There are two major categories of non-volatile memory: **primary** and **secondary**. All RAM is considered primary while some ROM is secondary. Primaries are the main working memories for computers providing quick processor access but are more expensive with limited capacity. Secondaries, known as **backup** or **auxiliary** memories, are much slower and can only be accessed through interconnects.

*Primary Non-Volatile Memory*

1. Standard ROM is non-adjustable / non-rewritable and is literally hard-wired when the chip is manufactured.
2. PROM can be programmed after manufacturing but cannot be changed once programmed.
3. EPROM can be erased and rewritten many times but erasing requires UV and all data must be totally erased.
4. EEPROM does not require total erasure nor UV. However, bit-by-bit erasing makes it relatively slow to reprogram.
5. NAND Flash is a type of EEPROM that can erase information, write data in chunks and works considerably faster than EPROM. It is the primary type of ROM used to store data in electronics today.

*Secondary Memory (HDD vs. SSD)*

**Hard disk drives** and **solid-state drives** are external non-volatile RAM devices used for permanent storage and core functions like boot-drive. HDDs use magnetic disks whereas SSD use NAND flash which is faster and more reliable but lower capacity and more expensive. However, Moore’s Law is improving SSD with NAND flash technology omnipresent in most portable electronics today. HDD is preferred when cost and capacity are critical whereas SSD is superior for versatility and reliability.

*Stacked Die Memory (HBM vs. HMC)*

In many systems, interconnects between chips limit performance. New die stacking technologies connect chips directly to each other without the performance degradation of chip-to-chip wiring. In addition to die stacking, 2.5 / 3D packaging has enabled new, tightly integrated memory architectures with significant performance advantages. **High-bandwidth memory** (**HBM**) and **hybrid memory cube** (**HMC**) are industry standards used to build layered memory devices. HMC uses a 3D memory-on-logic architecture where DRAM is stacked on top of logic and connected using TSVs. HBM stacks memory on top of device logic but separates this functional stack from the CPU / GPU and host logic function using 2.5D thus allowing silicon to be sourced from diverse suppliers.

Memory accounted for $117B of the $440B 2020 sales (27%) and was distributed evenly across different end-use applications. This is not surprising given that nearly all end-use applications require memory for core functionality.

**Optoelectronics, Sensors and Actuators, Discrete Components (OSD)**

***Optoelectronics***

SC devices that produce and receive light waves, which are used for a variety of applications, including light detection and image sensors, LEDs, information processing, fiber-optic telecommunications, displays, and laser technologies. **Photonic integrated circuits** (**PIC**) are commonly used as optical transceivers for data center optical networks that allows for more effective and efficient data transmission across greater distances relative to copper cabling. Photonics and optoelectronics are used in data centers, HPC, telecommunications, consumer goods, sensors and bio-sensors, aerospace, and quantum computing.

***Sensors, Actuators, and MEMS***

Sensors are specialized devices used to detect real-world inputs such as heat, pressure, light, and sound and convert them into electrical signals. Sensors are either **passive**, not requiring power, or **active**, requiring power. Sensors are often used in control systems, like altimeters and proximity sensors. A variety of SCs are used in sensors for applications dealing with optics, pressure, gas, speed, weight, etc.

Actuators are devices that convert electrical signals into real world output in response to a sensor-based stimuli. Actuators are primarily used in industrial and manufacturing applications like robotics but are also found in consumer and automotive markets. A modern revolution in industrial automation and autonomous driving has been made possible by the rapid proliferation of sensors and actuators. Smartphones are a great example of a single device containing a panoply of sensors and actuators including an accelerometer, gyroscope, electronic compass, pressure sensor, BAW filters and duplexers, RF switches, variable capacitors, TCXO oscillators, MEMS micro-mirrors, CMOS image sensor, auto-focus actuator, front camera, ALS and proximity sensor, micro-display, and silicon microphone.

Micro-electro-mechanical systems (MEMS) are tiny mechanical devices that operate gears or levers at a microscopic scale. MEMS are technically not SC devices since they do not use electricity to process and store information, but they do compete with SC-based sensors and are manufactured using similar technology. MEMS and sensor applications include positional, inertial, and pressure sensing, optical transmission, biological and medical, power and energy, RF and wireless.

***Discrete Components***

Individually packaged discrete components, including specialized transistors (power and switching), resistors, capacitors, inductors, diodes, and rectifiers, are used as enabling devices for more complex systems that route signal and power.

*Discrete Components vs. Power Management ICs (PMIC)*

Power delivery used to be handled exclusively by discrete components that performed functions like voltage regulation, power conversion, battery management, etc. Power management involves high voltages and signals moving around at high frequencies which creates a lot of parasitic interference and power integrity issues. For example, when a “noisy” PM chip is placed next to a sensitive circuit like a microphone, performance issues can reduce audio quality. This makes it essential to properly co-locate PMIC and PMU based power management functionality with critical sensors on the same IC. Upfront design costs are higher for PMICs, but their overall performance and efficiency advantages often make PMU microcontroller a competitive long-term option.

OSAD and MEMS accounted for $79B of the $440B (18%) 2020 sales. Comprising 77% of the communications, industrials, and automotive end use applications, OSAD and MEMS are heavily weighted relative to the other four constituent component segments.

**Analog Components**

Analog ICs process analog signals and are either configured as ASSPs or as standard linear ICs (SLIC), which are generic plug-and-play analog devices that can be integrated into a larger system. While digital dominates the processing and information storage market, the real world is analog and thus addressable by analog devices including sensors, wireless, and power supplies.

***General Purpose Analog ICs vs. ASSPs***

GPAs are used as broad plug-and-play analog components that may be optimized to perform a specific function but can also be used across many different systems much like micro-components on the digital side. GPAs include comparators, voltage regulators, data converters, amplifiers, and interface ICs. In complex systems, GPAs frequently sit between an analog sensor and a processor, amplifying and converting analog sensor signals into a digital signal for use by the processor. Like ICs in the Logic segment, ASSPs and analog ICs are designed for specific applications. Many ASSPs have digital components in them and are effectively mixed-signal devices including radio transceivers, audio amplifiers, and many varieties of radio frequency (RF) ICs.

GPAs, ASSPs, and other analog components accounted for $56B of the $440B (13%) 2020 sales. Analog components are most utilized in automotive, industrial, and communications segments with each comprising 25% of their end use applications. Like OSAD and MEMS, analog components are least used for computing applications.

So, we’ve broken down analog and digital electronics by differences in signal structure, data transmission methods, and power requirements. Though memory, micro components, and logic dominate SCI revenues, all five are integral to the SC ecosystem.

**Chapter 7 RF and Wireless Technologies**

**RF Signals and the Electromagnetic Spectrum**

To better understand wireless systems, it is helpful to consider two forms of energy – electrical energy and wave (wireless, airborne) energy. By manipulating these two forms of energy engineers can create, store, and communicate via the signals that are generated. RF signals are analog “waves” used to transmit information from one place to another without a physical cable or wired connection. RF signals exist across a broad range of intensities and frequencies along the electromagnetic spectrum. Electronic devices use a signal’s frequency to discriminate between different signals. End-use applications, which include radio, television channels, telephone, ISPs, etc., all use different frequency ranges, called frequency bands, to deliver different types of information. RF is used for radio and TV at the lower frequencies with microwave frequencies reserved for Wi-Fi, radar, and cell phones.

**The Federal Communications Commission** **(FCC)** strictly regulates frequency bands (AM, FM, CDMA, 802.11 etc.) and the providers that utilize them. Providers only have a certain bandwidth to deliver an ever-expanding suite of services to customers. Thus, providers are incentivized to squeeze as much revenue by cramming as much information for as many people as possible from the allotted bandwidth.

**RFIC – Transmitters and Receivers**

RF transmitters and receivers require at least six active or passive components: 1) power source, 2) oscillator to set transmission frequency, 3) **modulator / demodulator (modem)** to encode / decode digital information by modulating the carrier signal’s amplitude or frequency, 4) amplifier, 5) antenna, and 6) filter (low-, high-, band-pass, and band reject).

**The OSI Reference Model**

System designers must ensure that their system fits together as a properly functioning unit and seamlessly communicates with other devices. Without a standardized model, every hardware company would build different systems running separate programming languages that would have difficulty communicating with one another. With the **Open System Interconnection (OSI)** model however, you can just “follow the recipe” and build a system that both supports the needs of the software team while seamlessly connecting with other devices.

OSI describes the system layers needed to produce high performing networks and integrated systems with interconnected hardware and user interfaces. The OSI has seven layers with layers 1 – 3 dealing with physical transportation of information around the network and layers 4 – 7 addressing user applications. Layer 7 contains user interfaces like web pages and apps’ home screens. As you descend through subsequent layers, you get closer to the circuitry powering the application. The lowest OSI layer is the physical layer, where the data itself is transmitted to the underlying hardware. What’s important is that the external facing endpoints of a system can integrate well with other devices. We can also create a universal Macro-System Stack by adding a “hardware layer” beneath the OSI physical layer, which encapsulates everything that’s been studied so far, with the OSI layers above it making this hardware layer useful.

**RF and Wireless – The Big Picture**

How does a network of devices work together to bring you your favorite shows, talk to friends in other cities, or connect to the Internet? To build intuition, let’s track the path of a typical long-distance phone call. A phone call starts with transmitting a signal from a phone’s antenna which is picked up by a cell phone tower or base station (transceivers). The base station is a relay point that extends a service network to a specific area. Service providers invest billions to build robust networks of base stations across the globe to make sure you don't lose your signal. Each base station has a range called a coverage cell which is a patchwork of base stations that together makes up service providers’ coverage area. Base stations come in different sizes and coverage areas: satellite (global), macro-cell (suburban), micro-cell (urban), pico-cell (building), femto-cell (home), and Wi-Fi / Bluetooth (room).

From the base station the phone signal is sent to a central exchange from where it can be routed to any number of locations (satellite, base station, etc.) based on the final destination of the signal. From there the signal is routed to an information exchange center station close to the final destination. This proximal exchange center will then route the signal to a base station, antenna relay (for mobile devices), or land-line and this step finally connects the call to the intended receiver.

**Broadcasting and Frequency Regulation**

At any given time, a dizzying amount of RF signals are flying around in the air competing for limited and valuable bandwidth. Therefore, it is in the service providers’ best interest to maximize the information they send using their allotted bandwidth. A lot of complex technology goes into solving this problem.

***Digital Signal Processing, TDMA, and CDMA***

DSP technologies use sophisticated computational methodology (lossless or “lossy” signal compression) to fit more information into a given digital signal and in-so-doing cram more information density into limited bandwidth. **Multiple access standard technology** allows providers to route multiple calls through the same base station or across a fixed amount of bandwidth.

**Time Division Multiple Access** (**TDMA**) technology breaks data into chunks and multiplexes these chunks into the allotted bandwidth. For example, voice is sampled at 4KHz, so a 4MHz bandwidth can accommodate 1,000 calls. At the receiver, TDMA reassembles chunks from the same calls back into continuous voice call streams.

**Code Division Multiple Access** (**CDMA**) uses code to digitize voice and other data bits and transmit the encoded data across a wider bandwidth. At the receiver, CDMA encoded signals are decoded back into the original signal. CDMA can send data from numerous senders to numerous receivers simultaneously and use algorithms and DSPs to ensure that the data gets to the right place intact. Lightning-fast DSP allows TDMA and CDMA to establish uninterrupted service. TDMA is a technology that underlies **GSM** (**Global System for Mobile Communications**) and is the primary standard for communication networks globally. In the US, AT&T uses GSM while Verizon uses CDMA.

**1G to 5G(eneration) – An Evolution**

Telecommunications and wireless technologies are always evolving – the original 1G transmitted at 4kbps whereas today’s 5G can transmit up to 1,000,000kbps (1Gbps) – a 250K fold increase!

1G the first cell phones in the 1970s were big and clunky with terrible battery life that used analog technology to send RF analog “wave” signals between two points wirelessly (1979).

2G cells phones began using modulation to transmit digital data wirelessly. CDMA and GSM technologies were developed allowing service providers to connect more devices more affordably, though services were limited to voice and SMS text (1991).

3G building on voice and text, 3G expanded wireless capabilities to email, video streaming, web browsing, and other technology that make the “smartphone” possible (1998).

4G high-speed connectivity built on advancements in hardware technologies allowed for faster data transmission, mobile gaming, video conferencing, high-definition content delivery, and cloud computing capabilities (2008).

LTE **long-term evolution** is an industry standard which ensures that various devices, access points, base stations, satellites, and other components making up our telecommunications network can work together to create one big, fully functioning system. Standards like LTE help ensure that different companies can develop products that are interoperable with other parts of the system, so that services do not cut out every time signals are routed between nodes operated by different providers.

5G is still under development and designed to make 4G faster and more efficient. To deliver higher “data throughput” connections, 5G will require a robust network of thousands of cell towers and tens of thousands of small-cell antenna cells (2019).

**Wireless Communication and Cloud Computing**

While the evolution from 1G to 5G has delivered exponential improvements in data transmission and brand-new technologies like mobile gaming, video conferencing, and high-definition video streaming, these faster rates have driven an insatiable demand for **cloud computing**. The cloud comprises countless servers housed in giant rooms called **data centers (DC)** that vary drastically in size: Microsoft has a shipping-container sized DC that it uses for its Bing Map data, Google has a 200,000 square foot DC in Council Bluffs, Iowa, while China Telecom currently holds the record for the world’s largest DC in Hong Kong at over 10 million square feet.

By storing, or hosting, applications on higher performance computers, companies and consumers can store information, run applications, and boost capacity without having to invest in and manage all their own infrastructure. This is only possible because communication networks are so fast today. Prior to the boom in wireless innovation over the last couple of decades, the bottleneck to centralized computing operations like DCs was moving data between end users. With this bottleneck alleviated, cloud computing is here to stay. The problem has now shifted from limited bandwidth to building and powering DC infrastructure that can support ballooning demand.

**Chapter 8 System Architecture and Integration**

**Macroarchitecture vs. Microarchitecture**

Microprocessors serve as the brains and computational workhorses of the SC ecosystem and contain the arithmetic, logic, and control circuitry necessary to execute instructions, process data, and run sophisticated software programs. The most advanced PC microprocessors today contain over 100B transistors, and this leads to immense design complexity and challenges. Therefore, engineers must consider both micro- and macro-architectural decisions carefully in balancing the tradeoffs between flexibility and performance with the cost and complexity of tighter system integration. “Architecture” is used in SC engineering to mean one of two things – system architecture or microarchitecture. System level, or macro-level architecture that is used to define entire chip families, is technically specified by different kinds of **Instruction Set Architectures** (ISAs) that describe how instructions are delivered from programmer to computer. Microarchitecture describes the way an ISA is implemented into the hardware design itself. A CPU, GPU, and PMU may all be designed using a single ISA that governs the design of the whole SoC, but still have unique microarchitecture.

**Common Chip Architectures**

***Von Neumann Architecture (VNA)*** is a macroachitecture developed by **John Von Neumann** in the 1940s and is what most modern computers are based on today. It relies on the CPU, I/O interfaces, and memory. Inside the CPU there exists **registers**, where data and instructions are delivered by and given to the memory, a **control unit** that determines which instructions should be executed, and the **arithmetic and logic unit**, where instructions are carried out and information is processed.

**Harvard Architecture (HA)** is like Von Neumann, however, differs in how it accesses (input) and distributes (output) information. HA parses instructions and data into two separate memory banks, with a unique bus for each type of input. This separation theoretically allows parallel access of data and instruction sets thus reducing the clock cycles necessary to perform a single instruction. This **pipelining** is difficult to implement and increases complexity and cost over VNA.

HA is used primarily in microcontrollers and specialized DSPs while VNA is used in computers, mobile SoCs, and complex digital electronics. The vast majority of IC’s use a VN “type” macroachitecture that requires system designers to weigh the pros and cons for which kind of VN macroarchitecture is best for the device they are building.

**Instruction Set Architecture (ISA) and Microarchitecture**

ISA determines the set of instructions, or lines of codes, that a processor can support. Microarchitecture determines how a processor receives and executes those instructions at an implementation level. ISA and microarchitecture are dually tied to the Universal Architecture Stack that consists of seven layers. Each level is dependent on the collection of levels below it and must be designed to support the levels above it. Inputs into a system via the top-level UI layer must ultimately be carried out by the transistors and functional components that comprise an integrated circuit. The translation process that makes this possible is difficult and requires engineering skill spanning numerous areas and disciplines. Robust abstraction layers, therefore, are vital to building high performance systems within this stack.

**Instruction Pipelining, Processor Performance, and CISC vs. RISC**

Understanding the three factors that determine processor performance is essential to assessing the trade-offs between CISC versus RISC:

1. **Increase processor’s oscillator clock** **frequency** – we can significantly boost performance by shortening the clock cycle and the distances signals must travel through a circuit.
2. **Synchronize task performance** – this is done by adding more processor cores (co-processors). GPUs are particularly well suited for this kind of parallel processing. This strategy has an upper bound and depends on the number of registers available to accept inputs and deliver outputs to cache memory – this hand-off is a primary bottleneck for processor throughput.
3. **Pipelining** is a method of speeding up tasks by breaking up a lengthy workflow into smaller constituent parallel tasks. Unlike synchronous processing that relies on parallel tasks execution along separate paths, pipelining breaks up a process into smaller constituent tasks performed in parallel along one path. Pipelining can significantly improve processor throughput by enabling more instructions per clock cycle (RISC better suited than CISC). This can give one ISA an edge over another depending on the application. Pipelining is a complicated operation that must ensure both the task data and task instructions arrive at the CPU at the proper time.

The two main ISAs are **CISC** (**complex instruction set computing**) and **RISC** (**reduced instruction set computing**). There are many differences between the two, with the most notable being the way they process instructions. For many applications, RISC is often seen as a more effective option. Unlike CISC processors which use lengthy, multi-clock cycle instruction sets, RISC processors are better optimized for pipelining standardized instruction sets that are easier for **compilers** to consume. CISC originated around 1970 with RISC following about a decade later. In its infancy, compilers were unreliable, and programmers often wrote code using **assembly**, a language that is much closer to the hardware but also more complex and difficult and less portability that compiled languages. As compilers improved, RISC architecture was more widely adopted by design teams across the industry. In summary CISC is hardware centric, less dependent on compilers, whereas RISC is software centric, more dependent on compilers.

Because they process simplified instructions, RISC tend to consume less power, making them ideal for applications where power is particularly limited, such as mobile phones or other battery-powered devices. However, this advantage is less in higher-performance applications, like servers or personal laptops, which tend to use CISC more often.

RISC processors do require more RAM to access additional code and greater programming efficiency since shorter instructions means more lines of code. Thus, in the early days when memory was a lot more expensive, CISC dominated but RISC gained significant traction as memory chips became smaller and less expensive.

**Choosing an ISA**

Choosing an ISA is a difficult and consequential decision. MIPS and ARM are licensable ISAs while others like RISC-V are proprietary or open source. A licensed ISA will typically come with a pre-designed processing core while an open-sourced ISA will not. Licensing fees and royalties are key factors in deciding whether to license, build, or borrow.

Oftentimes more important than direct costs are the risks that come with each ISA. Engineers must consider the time and costs of physically developing a core processor in-house. Most ISA companies license processors that serve as the core of any customized end-product. Even if a unique core processor design is completed within a feasible time frame, there are always manufacturing risks that can cause a new processor and architecture to fail.

Even more important are the downstream software implications. ISAs like ARM and x86 have mature software “ecosystems” with fully developed software development stacks. Building a new processor with a proprietary architecture requires the development of new firmware, operating system, and development tools. Even if hardware and software execution is seamless, time-to-market is likely to be an issue – while you’re spending time developing a core processor and proprietary architecture, the competition is releasing new products.

Another important consideration is whether customers are buying the processor or what the processor does? If your customers are buying your new algorithm or integrated processor and sensor system, then licensing an existing ISA may be more attractive. But if the processor itself forms the core of your business or adds some unique value not available in the market, then a custom design may be the right answer. There are many trade-offs when choosing between proprietary, licensed, and open-source ISAs including: 1) hardware design costs,  
2) software engineering costs, 3) existing software ecosystem, 4) time-to-market, 5) manufacturing risk, 6) design flexibility, and 7) royalties and licensing fees.

**Heterogeneous vs. Monolithic Integration – From PCBs to SoCs**

Shrinking transistor size due to advances in manufacturing and lithography enabled SC design companies to crank out increasing device performance without having to overly focus on architecture and integration. There was enough “wiggle room” so that even if a design included unnecessary transistors, it could still deliver greater performance using less power and space at lower or equal cost. Even if a company was motivated to make its chip more efficient, the additional design effort required to make fully integrated devices was costly. Thus, the rapid pace of Moore’s Law made time-to-market a critical constraint since the cycle time for manufacturing advances, that is getting to the next node, was shorter than the design effort time for architecture redesign at the previous node. This geometric scaling continued unabated up until the last decade or so when the industry ran into three main problems:

1. As transistors shrank and logic became denser, power management became a primary design constraint. Modern CPUs and other advanced devices often cannot use their full firepower since the heat generated at such speeds can literally fry the IC.
2. Lithographic technologies, like EUV, depend on the wavelength of light being used and as the required wavelength gets smaller, it has become that much more difficult to push the technological envelope.
3. The thickness of modern transistors is just a few atoms thick with little room for further shrinking.

Each successive node requires disproportionately more expensive process technology, and this increases the costs to build new fabs and drives up manufacturing unit costs. This is shifting engineers’ mindsets towards squeezing out more performance from older nodes, that is, towards functional scaling. In essence, Moore’s Law is slowing down and this is pushing engineers to optimize designs for specific applications and shifting system architectures to include more heterogeneous and monolithic integration.

In **heterogeneous integration**, numerous chips are integrated with one another on the same PCB or within the same package, called **system-in-package (SiP)**. In **monolithic or homogeneous integration**, numerous functional modules are included on a single IC, yielding a fully functional system called **system-on-chip (SoC)**. The more integrated a system is, the less distance a signal must travel to reach other parts of the chip, however, they are also more complex to design and manufacture with numerous drawbacks that must be considered.

Choosing between SiP and SoC poses significant trade-offs. SoCs have small form factors and thus impart area and power efficiency advantages making them popular in smaller battery powered devices. However, they can suffer on performance based on the application. Each functional portion of a chip may require different materials and process technologies for peak performance, which is difficult or impossible to do on a single wafer. By integrating all components on a single wafer, some parts may function very well while other parts perform poorly due to suboptimal materials and processes used therein. Cost differences also play a big role. Although monolithic devices have higher design costs and manufacturing complexity, heterogeneous devices may not have lower unit costs in all cases. Monolithic devices require less area in aggregate, which enables more chips to be printed on a single wafer and can lower net manufacturing costs.

Furthermore, heterogeneous integration enables a sort of “manufacturing arbitrage”, where different process nodes may be used for different parts of the system. In a SiP, for example, memory or core logic could be manufactured using advanced nodes while parts like analog or RF components could rely on older nodes leading to a blended node technique with lower manufacturing costs. Additionally, companies can build a next-gen system by only changing the memory and logic devices to enable more features, while keeping power management or RF components unchanged.

Manufacturing executives must take great care to balance the additional design and manufacturing cost differences when deciding which architecture to employ. SoCs consume less power and take up less space than SiPs at the board level. **System-on-Boards (SoBs)**, however, have greater design flexibility, lower design costs, and shorter time-to-market. SiPs land between SoCs and SoBs offering greater design flexibility while reaping higher performance, power, and area advantages due to greater integration. In addition to the core PPAC factors, system architects and design teams must keep time-to-market in mind. Generally, the more integrated a system, the more design time needed. If a competitor is exerting pressure with new product releases, designing a new SoC from scratch may not be the best strategy.

**Chapter 9 The Semiconductor Industry – Past, Present, and Future**

Since its beginnings in the 1960s, two significant ongoing challenges of rising design costs and rising manufacturing costs have driven the SCI from yesterday’s collection of companies using fully integrated business models to today’s multi-faceted fabless design models.

**Design Costs**

The IC design pipeline – architecture / IP qualification > design verification > physical design > software licensing (EDA) > prototyping > validation > fabrication / manufacturing – is extensive and expensive and fraught with implementation and strategic risks throughout. In the early days, there was no standardization and scalable tools and most companies made ICs in-house. This drove up design costs and at the end of the day one company’s IP wasn’t guaranteed to work with another company’s IP. Electronic data automation (EDA) companies changed this paradigm by modularizing significant chunks of IC design which led to substantially reducing design costs. Further benefits were reaped by the promulgation and adoption of hardware description languages (HDL) like VHDL and Verilog. This enabled design firms to build larger systems at aggregated and more manageable levels of abstraction. Separating design from EDA allowed companies to refocus on the core competencies of software and hardware. In the early days this benefited advanced analog chip companies that could refocus on chip design using licensed EDA tools. Today, Cadence, Synopsis, and Mentor Graphics are the three biggest players in an EDA market that accounts for 50% of total design costs.

Throughout the 1980s, most ICs were of medium size and complexity, with systems companies purchasing different chip types and integrating them into their device by soldering them to a PCB or other connecting devices. As technology advanced, systems companies strived for integration and thus SoC demand skyrocketed, in turn increasing the complexity of each system component and making design even more difficult and expensive. Out of this cycle sprang **SC IP companies (SCIPC)** with three main licensable offerings: 1) microprocessors like those offered by ARM, 2) communication architecture that enable different parts of an SoC to talk to one another, and 3) analog IP which has become more difficult to design at each successive node. SCIPs reduced complexity by offering base designs of common modules for application specific circuits and also offering cell libraries used for generating more complex designs. Companies could now quickly acquire undifferentiated components, and this allowed them to focus on the novel parts of their chip design. Analog companies could snap up digital components like microprocessors and memory and digital companies could source analog components like oscillators and power circuits. Furthermore, as today’s lithography and manufacturing equipment suppliers push to the next node, SC R&D and design have become increasingly difficult and complicated due to phenomena like quantum tunneling and current leakage. Design costs alone, not including manufacturing, for 3nm SoCs ranges from $0.5B to $1.5B while 7nm and 10nm can be up to $300M and $175M respectively (IBS).

**Manufacturing Costs**

Though design costs have posed significant challenges to profitability, further down the SCVC, skyrocketing manufacturing costs have had even greater repercussions. Each successive node requires increasingly complex and expensive manufacturing equipment like dicing machines, polish grinders, masks, and steppers all enclosed in specialized air filtration and vibration control environments. Advanced SC fabs with a usable life of five years can range from $7B to $20B not including the variable costs of chemicals and materials.

FE-M technology like lithography equipment has historically accounted for most of the manufacturing costs, though the BE-M has started receiving greater attention as the industry shifts focus to advanced packaging architectures and heterogeneous integration to boost performance. Figure 9.1, page 179, is an excellent analysis of both design and manufacturing costs.

Together, supply-side design and manufacturing costs and demand-side pressures for greater variety of customized designs have transformed the industry from the **fully integrated SCCs** of the 1960s – 1980s to today’s mix of **fabless design companies**, **specialized IP / EDA tool companies**, **integrated device manufacturers (IDM)**, **pure-play foundries**, and **system design companies**. This evolution embodies David Ricardo’s theory of comparative advantage in which the market functions optimally when everyone focuses on what they are best at providing.

**Evolution of the SCI**

*1960s – 1980s Fully Integrated Semiconductor Companies*

These companies would forecast demand for a product, then design, manufacture, and package the product before marketing it to potential customers. Each company invested significant resources into its own fabs, the high fixed costs of which left firms vulnerable to volatile demand swings. If sales for their products dipped even slightly, fab utilization could drop significantly, spreading revenue across a greater cost basis and cutting into profit margins.

*1980s – 2000 IDM, Fabless Design, Pure-Play Foundry*

The rising costs of building, owning, and retooling a fab with each progressive node drove SCVC fragmentation in the 1980s and 1990s. Once built, companies were driven to maximize their output and contribution margin, even if they weren’t covering their fixed costs. As a result, IDMs like Intel, which design, manufacture, and sell their own ICs have started to lease out parts of their fab capacity while former IDMs like TI and AMD have become fabless. Processor companies would also sell their old fabs **down-market** to analog companies thus extending the effective useful life of these older node fabs.

In the mid-1980s, companies like Xilinx (1984) and Qualcomm (1985) pushed a business model designed to take advantage of this excess capacity. These companies designed their own chips and signed contracts with an IDM to utilize additional manufacturing capacity thus foregoing upfront capital investment and basically converting fixed costs into variable costs that would be priced into each wafer. Not only did this fabless model reduce required capital for new entrants, it also better matched variable supply to variable demand economics.

In 1987, TSMC, a pure-play foundry entered the market with an entirely new business model, focusing completely on manufacturing other companies’ designs. Foundries are able to focus on a smaller set of core competencies and also take advantage of more stable demand and consistent volumes since they draw orders from multiple customers instead of just their own designs. Additionally, foundries strategically position themselves in markets with lower labor costs like Taiwan and other parts of Southeast Asia. Xilinx and other fabless design companies quickly shifted away from IDMs to lower-cost domestic and overseas foundries, a model that has dominated the market ever since. Even the few remaining IDMs like Samsung and Intel still depend on pure-play foundries for more advanced technology nodes, typically using their own manufacturing capacity to produce devices requiring less demanding, older process nodes.

In the late 80s and early 90s an important shift was happening further up the SCVC. EDA tools made FE-D easier and systems companies that had historically depended on out-of-the-box products or custom silicon from fabless design companies and IDMs began building their own chips better suited to their specific needs in essence bypassing chip design companies and working directly with foundries. However, there was still a gap in BE-D expertise necessary to carry out physical design and downstream SCVC activities. New players like VLSI Technologies and LSI Logic as well as established players like Qualcomm met this demand with innovative ASIC-centric business models. Pure ASIC companies like LSI handled FE-D for chips developed by systems companies and then coordinated with manufacturing and assembly suppliers to bring their product to market. Design services companies like Xilinx and Qualcomm drew much of their revenue from a similar model, though they also began to focus on specific markets, developing unique expertise and growing product portfolios of their own. Xilinx, acquired by AMD, controls over 50% of the FPGA market while Qualcomm is a world leader in custom silicon for wireless connectivity and infrastructure products.

*2000 – Present Fabless Design Companies, Foundries, IDM Stragglers, System Company In-House Design*

Today the SCI is driven by fabless design companies, pure-play foundries, and a few remaining IDMs supported by an ecosystem of EDA tool developers, equipment manufacturers, and SCIPs. Systems companies, like Apple, Facebook, and Tesla, motivated by the need protect their IP, are designing their own chips. These companies are playing a growing role in the SCI and continue to eat into what used to be traditional fabless design company revenues. Figure 9.3, page 183, charts the evolution of the SCI from the 1970s to 2010s.

**Fabs vs. Fabless Design – The Case Against IDMs**

Owning a fab can provide numerous advantages, including greater process control, faster time-to-market, and tighter design integration. But these advantages have increasingly been outweighed by costs. Pure-play foundries’ biggest advantage over IDMs is their ability to pool demand across many customers in order to fully utilize capacity thus smoothing revenues and minimizing fixed asset underutilization. This is critical in the SCI, because underutilization scales directly with fabrication costs and deteriorates unit economics. Operating an IDM today is truly a feast-or-famine endeavor. If you hit all your targets and release a product that is superior to your competitors, you have a huge advantage. If you miss those targets, you’ve sunk billions into failed process development, and may be forced to manufacture your products in the same foundries that your competitors are using. In this case your internal development may never catch up to the competition.

Proprietary manufacturing IP and data capture is another common reason IDMs may use to justify the costs of fab ownership. Though this may have proved an advantage historically, foundries have unique technological advantages that are hard to match as an IDM. Because they manufacture a wider range of components and ICs, foundries can iterate process technology at a quicker rate, building important competencies that they can then democratize and provide to fabless design houses and other key customers.

Tighter integration between design and manufacturing is a substantial advantage that can result in significant cost savings and performance improvements. Though fab-owning IDMs may have a material edge here, software tools, design suites, and networking technologies have made foundry-to-customer integration highly efficient. Fabless design companies like Qualcomm and Broadcom can seamlessly design new chips with fabrication costs and **production optimization** front-and-center. This isn’t only limited to big companies – startups can access leading integration and manufacturing technologies using **shuttle runs** where a foundry combines multiple customer designs onto a **single mask set** and manufactures a few wafers with several hundred chips that the startup can use to demo a minimally viable product. Cost savings can be significant - wafers costing $5M per customer can be cut to $500K if ten customers are put on a single mask set.

Thus, the pros and cons of fabless versus IDM can be assessed on various criteria: 1) manufacturing, 2) time-to-market, 3) organizational complexity, 4) overhead and variable costs, and 5) technology and integration.

When you analyze revenue to capex and PP&E, revenue to PPE ratios heavily favor fabless design companies. For all the reasons discussed so far, fabless companies are able to produce more income per dollar of capital relative to IDMs and this may explain why a fabless company like Nvidia’s market value exceeds that of Intel, despite having only a fraction of Intel’s sales. Figure 9.5, page 187, presents cross-sectional data on various business models and their associated sales and ROA metrics and lends insights into why the two major IDMs, Intel and Samsung, have moved to a **Fab-Lite** model.

An X-factor to keep in mind here is the recent global SC shortage and competition between the U.S. and China. The few remaining U.S. IDMs like Intel and GlobalFoundries are poised to benefit substantially from the $50B in government incentives recently passed by Congress to strengthen the industry and shore up domestic supply chains. It is yet to be seen, however, if this renewed support will counteract the fundamental market disadvantages IDMs suffer from today.

**Industry Outlook**

Shaped by a multitude of overlapping variables and market forces, the SCI is constantly changing and so we’ll highlight five key trends that shaped the industry’s history, influence its present health, and drive future growth prospects.

*Cyclical Revenues and High Volatility*

SC sales are largely driven by the electronics industry and are highly cyclical and volatile characterized by boom and bust cycles that can last many years. Nominal sales growth / contraction can stem from pricing dynamics – memory SC, like commodities, are highly price sensitive and are considerably more volatile than non-memory SC.

To manage year-to-year silicon cycles, SCCs must be able to control costs, while not sacrificing key investments in R&D. Manufacturing represents the greatest expenditures by US-based SCCs, comprising on average nearly a third of total costs, followed by R&D, D&A, and SG&A. For IDMs, it can be a real challenge to weather an industry downturn if you have massive fixed costs from a manufacturing facility on your balance sheet. Production expenditure has grown significantly as a percentage of total costs over the past two decades.

*High R&D and Capital Investment*

The rapid pace of technological advancement has driven SCCs to invest markedly high amounts of capital into core R&D in order to maintain their competitive edge – since 1999, U.S. SCCs have invested 15–20% of sales into R&D, the highest after pharma and biotech. Additionally, U.S. SCCs plow 8–20% of sales into new PP&E, second only to alternative energy (SIA Databook 2021). SC R&D and capex have grown 5.6% year-over-year, with US companies spending $75B in 2020 (SIA Factbook, 2021).

*High Compensation and Positive Productivity Growth*

A range of factors has led to SCI wages and productivity outpacing other sectors, with annual compensation rising from $80K in 2001 to $160K in 2019 (SIA Databook, 2020). Sales per employee, a strong indicator of productivity, has doubled over the past twenty years to $571K in 2020. Unlike most industries where average selling prices increase at or above the rate of inflation, per unit costs have decreased at a rate fast enough for companies to maintain profitability without excessive price increases. It’s important to realize that consistent profitability has only been made possible by increasing production efficiency.

*Long-Term Profitability*

Earnings have concentrated amongst the biggest competitors having the scale to weather downturns and the foresight to capitalize on disruptive technologies like PCs and smartphones. Since 1999, pre-tax profits have averaged 20% of sales with gross margins ranging from 37-57%. Major growth drivers include:

1. **Increased demand** for consumer electronics due to rising household disposable incomes, increasing urbanization, and rapid population growth (Fortune Business Insights, 2021).
2. **Fast-expanding emerging economies** with growing demand for ICs.
3. **Technological use cases** including IoT, smartphones, 5G communications, and AI/ML.

Reduced demand in automotive, industrial, and parts of the consumer market were in turn offset by servers, PCs, and long-term trends in AI and 5G. This demand hedging helps preserve growth during volatile silicon cycles. For example, Nvidia, the leader in processors for graphics, AI, and crypto, has set new records for quarterly revenue for every quarter of 2021, peaking at $6.5B quarterly revenue in August 2021. Though new technologies and the explosion of PCs, servers, and cell phones over the last couple of decades have been a boon for the industry, high consolidation has also adversely affected many companies.

*High Consolidation*

As SoC designs become increasingly complex and transistors are pushed to their physical limits, design and manufacturing costs have never been higher and thus the industry’s profitability has depended on R&D breakthroughs targeted at cutting costs. So far, this has been realized with $0.98 unit chip costs in 2001 dropping to $0.63 by 2019. In addition to the pressure to decrease per-die unit costs, there has been enormous pressure to get more out of each device, giving consumers increased access to greater computing power at lower costs. These opposing forces have driven immense competition and led to significant SCI consolidation, with annual deal volume averaging $69B since 2015. This is a natural pattern in capital intensive industries that favor size in order to amortize ballooning fixed costs across revenues. COGS as a percentage of revenues is much higher for smaller firms who lack the overhead and scale to compete with bigger players.

In 2001, there were 29 companies offering advanced fab services. Today there are only five, including just two main foundries, a handful of EDA companies, and ASML the largest lithographic equipment supplier. In such a capital-intensive industry with the top 10 SCCs owning 55% of the market, smaller companies struggling to survive have had to grow to stay competitive. (IC Insights, SIA Databook, SIA Factbook, Design and Reuse, McClean Report). Refer to Figure 9.12, page 195, and Appendix B for more insights.

*SCI Dynamics Summary*

**Market volatility** – cyclical revenues with boom and bust “silicon cycles”, heavily influenced by memory prices.

**High capital intensity** – high R&D costs, high PP&E costs, second highest industry R&D plus capital costs.

**Compensation and productivity growth** – cost decreases due to productivity growth, 2x growth in revenues per employee over last 20 years, 3x average compensation relative to all manufacturing jobs.

**Long-term profitability** – new technologies like IoT and AI, expanding emerging economies, increased demand for electronics.

**Industry consolidation** – rapid M&A activity, small number of big winners, high capital intensity.

**US vs. International SC Market**

Since 1999, the U.S. has maintained approximately 50% SC market share and currently controls about 47% of sales ($208B of $440B 2020 sales). The U.S. continues to house most design companies and IDMs and while some design companies are growing in Europe and Asia, these companies are still predominantly US-based design efforts.

Manufacturing and assembly firms are located predominantly in Taiwan and Asia Pacific, although manufacturing may start shifting towards India and South Asia. Long-term market dominance may be dictated by high rates of R&D investments by US companies, which spend more on R&D as a percentage of sales than any other country. Figure 9.15, page 199, shows detailed analysis of SCVC activities and consumption broken down by region (BCG and SIA’s report on Strengthening the Global Semiconductor Supply Chain). The US, for example, leads in areas that require intensive R&D like Logic, EDA, and core IP, while Asia focuses more on areas that are labor-intensive and require significant capex like materials, wafer fabrication, assembly, packaging, and testing.

It is important not to monolithically analyze the SCVC. Going from demand assessment all the way to product delivery requires massive amounts of specialization, timing, and coordination between different SCVC participants across the entire globe (page 201). Real SCVCs are much denser and require thousands of components, tools, equipment, and logistics provided by hundreds of suppliers.

The US controls about 60% of logic and analog markets, though it trails in memory and discrete components with about 20–25% market share in each. Contrary to popular belief, the US still manufactures a considerable portion of the world’s SC, however, its manufacturing capacity has grown at a rate less than five times that of firms overseas, although recent supply chain concerns may slow this decline.

From the demand side, Asia Pacific is by far the largest consumer of SC, comprising over 60% of worldwide demand with $271B of $440B 2020 sales, Figure 9.17, page 202. Though China has been the largest consumer of ICs since 2005, chips that are designed and produced within the country account for only 15% of total purchases. To avoid any possible consequences of future trade tensions or over-reliance on the US, the Chinese government has devoted considerable attention and resources to grow the country’s domestic SCI. In 2014, China pumped $20B into government-backed private equity funds, like Tsinghua Unigroup (SC technology development) and $29B in 2019 with the aim of reducing China’s dependence on foreign suppliers and developing IC design and manufacturing technology.

**COVID-19 and the Global Semiconductor Supply Chain**

Covid set in motion a global chip shortage that exposed many SCVC vulnerabilities and has had significant economic consequences, including production cutbacks in autos, consumer electronics, medical devices, and networking equipment, pushing lead times in many cases beyond a year. Insufficient capacity did not cause these problems, but rather demand-side distortions, such as slumping auto sales which shifted chip orders from the auto market to other end-markets. In fact, SC manufacturing utilization remained optimal throughout the pandemic. By the time auto demand rebounded, manufacturing lines had already been retooled to produce other consumer products, leaving autos high and dry with no chips for their vehicles.

Pandemics, forecasting imbalances, locked-in manufacturing allocation, along with fires at two key Japanese factories, may seem like one-time, unavoidable hiccups. However, structural issues within the SCVC amplified the degree of disruption caused by these events. The core weaknesses of today’s SCVC are rooted in the regional stratification of key SCVC activities.

US leadership in R&D and design, for example, can largely be attributed to its existing talent pool and access to steady streams of new engineers from US universities. This dynamic is likely to help the US retain a significant edge in engineering talent for the foreseeable future. The US also benefits from a vast pool of VCs, who have the capacity and will to make ambitious bets in the SCI.

On the manufacturing side, East Asia holds competitive advantages, including skilled and affordable manufacturing talent, robust infrastructure, and higher levels of government incentives. Note that the role of government incentives should not be underestimated. Government incentives may account for up to 30–40% of the 10-year total cost of ownership of a new state-of-the-art fab. The cost of ownership of the same fab in the US is between 20–50% greater than in Asia with 40–70% of that difference stemming from lower incentives offered by the US government relative to Asian competitors. While one might assume that lower construction and labor costs might be the main reason that Asia has an advantage, the data confirms that it is in fact government incentives that make up the difference, explaining between 40 and 70% of the 25–50% total cost of ownership advantage.

While free trade and specialization have enabled the global SC ecosystem to thrive, delivering performance at lower cost to global customers over the last 50 years, the price paid has been the evolution of a fragile and unstable supply chain. Today, there exists greater than 50 points on the SCSC where a single region controls more than 65% of the global market share (SIA whitepaper, 2021). This concentration exacerbates three key risk factors – unavoidable random natural variability, geographically clustered manufacturing capacity, and geopolitical conflict.

To strengthen the SCSC and make it more resilient, experts do not believe every country needs to become completely self-sufficient since this path would be prohibitively expensive and significantly inflate prices. However, targeted investment in US SC manufacturing capacity and a greater balance between efficiency and redundancy would go a long way in protecting against future chip shortages and economic downturns, while reducing our over-reliance on other countries for components vital to national security.

**Chinese Competition**

Manufacturing capacity concentration in East Asia and China puts the US economy and national security at risk. As such, shoring up the SCSC has become a recent political priority in the US. Of the $250B Science and Technology bill, recently passed by Congress, $52B was earmarked for SC manufacturing. This bipartisan bill is intended to counter China’s growing economic and military power and is an attempt to buck the trend of the past two decades that have seen rising manufacturing costs compelling US players to sell, spin off, or abandon their foundries. Today there remains only five major companies that manufacture SC in the US – Intel, Samsung, Micron, Texas Instruments, and Global Foundries (spun off from AMD in 2009). There are also smaller analog companies that maintain minor amounts of manufacturing capacity at older nodes and niche processes.

Though US manufacturing capacity has remained stable, over half of the 27 new advanced node fab construction projects are slated to be built in China. Figure 9.19, page 208, shows the forecasted trajectories of US and Chinese chip manufacturing market share. China has made domestic SCI development a key piece of its five-year (2020-25) plan and invested heavily in the space, pledging over $150B in investment from 2014–2030 (SIA Whitepaper, 2021). However capital infusions can only go so far. Advanced SC manufacturing requires significant pools of engineering talent, a seasoned base of companies with technical know-how, and access to advanced manufacturing toolsets. Despite nearly $50B in government incentives delivered over the last 20 years, Chinese companies only account for 8% of global SC sales today, with a small footprint in advanced logic, cutting-edge memory, and higher-end analog chips. Despite these lackluster results, capital investments have driven annual growth rates of 15–20% and positioned China as a leader in the labor-intensive OSAT market. All said and done, China is likely more than a decade away from advanced technology nodes like those in Taiwan.

A more diversified and robust manufacturing base in both countries could reduce over-reliance between countries and boost global manufacturing capacity, which would lower costs for consumers globally. Though rising design and manufacturing costs constrain growth and squeeze margins, the true existential treat to the SCI is the fundamental limits of transistor sizes and the slowing of Moore’s Law.

**Chapter 10 The Future of Semiconductors and Electronic Systems**

**Prolonging Moore’s Law – Sustaining Technologies**

There are currently many promising research areas, both within existing technology architecture and brand-new sources of computing power that will continue the technological march of the SCI for years to come. Within traditional silicon engineering, renewed focus has shifted from shrinking component size, geometric scaling, toward improving design efficiency and integration, functional scaling, as well as exploring new materials and design methods to prolong Moore’s Law. Some key technologies of the future include:

1. **2.5D and 3D die stacking** – by building up instead of out, data transfer rates are vastly improved, costs are reduced, power is conserved, and space is preserved by increasing transistor density on the substrate. The trade-off is that design complexity of existing designs in greatly increased due to more intricate data flow schemas and system architectures. For an unstacked 1D 3nm node, design costs range from $0.5B to $1.5B. Stacked logic will undoubtedly increase design costs, though high-performance applications like AI will likely continue to drive demand for increasingly complex system architecture.
2. **Gate-All-Around (GAA)** transistors and new channel materials are a promising next step in the evolution of transistor technology and offer the most concrete path to prolonging geometric scaling. Planar transistors, the dominant transistor technology from the 1960s through 2000s, is satisfactory at the 20nm or larger node, however sub-20nm, planar transistors suffer from crippling leakage current that hamper overall system power. In 2011, planar was phased out by FinFET for advanced ICs because they reduce leakage current and enable greater control at lower voltages. However, some current still leaks out even if the transistor is turned off. Leakage has become a critical issue as we move towards 3 and 2nm nodes. With gate control across four dimensions, Nanowire and Nanosheet GAA transistors aim to solve these issues. GAA present unique deposition and etch challenges that will likely require new channel materials like strained Silicon Germanium (SiGe) to mitigate electron mobility issues. In addition to current leakage, quantum tunneling interference is another big problem is very small atomic scales. There are numerous candidate materials, such as **graphene sheets** and **carbon nanotubes** (strong materials that are only one atom thick), that is well suited to resist tunneling interference. Carbon nanotubes, however, are difficult to manufacture and will requires a lot for R&D before they are ready for market.

As transistors become smaller and more densely packed, heat dissipation becomes a major performance constraint, forcing many devices to run below their maximum speed to avoid overheating. Traditionally, silicon has been used as the main channel material but has power density constraints that limit many of the performance advantages from small GAA transistors. This challenge can be tackled by implementing different channel materials with greater electron mobility. This is what strain engineering is focused on with research into SiGe, GaAs, GaN, and other III-V elements being a hotbed area.

1. **Custom Silicon and Specialized Accelerators** – have shown promise for functional scaling and performance improvements in specific products. Functional scaling has been clearly demonstrated by Nvidia GPUs’ ability to perform key AI calculations having doubled every year resulting in 317x gain through May 2020, being dubbed as Huang’s Law after Nvidia’s CEO Jensen Huang.

Functional improvements in key areas like memory and GPUs can benefit IC performance even as geometric scaling slows down. In addition to functional scaling of widely adoptable subsystems, custom silicon design can boost performance through tighter integration using existing technology. Though cost prohibitive for small companies who depend on a fabless model, large SPCs like Apple, Google, Facebook, and Tesla are all developing custom chips in-house since they can afford building out internal engineering groups. By shifting from commodification to full customization, they are able to sustain competitive advantages in performance, which may not be possible through third-party providers, foster quality control advantages, and reduce the disclosure of sensitive information and trade secrets.

**Optical Chips and Interconnects** main advantage over copper is that multiple optical signals can be transceived in parallel using different wavelengths. VC backed Ayar Labs and MIT / UC Berkeley researchers, working on a DARPA project called Photonically Optimized Embedded Microprocessors (POEM), have begun commercializing photonic chip technology. Ayar has targeted chip-to-chip communication by creating I/O optical interconnects that are much faster and more power efficient than traditional copper. Optical interconnects are very promising because they may resolve several significant bottlenecks in SC architecture and its related data transmission issues that exist today.

**Overcoming Moore’s Law – New Technologies**

Outside of advances in traditional silicon engineering techniques, there are some truly fascinating technologies in development that could launch us into a post-Moore’s Law computing renaissance.

1. **Quantum Computing** involves extremely complex physics, but in essence quantum computers combine superposition with entanglement to perform exponentially difficult calculations that modern computers can’t handle on their own.
2. **Quantum Transistors** provide an interesting twist to quantum computing by harnessing the power of quantum tunneling and entanglement to process and store information. Despite a lot of R&D, tunneling and entanglement are still not fully understood and maintaining the atomic-scale control necessary to harness such forces is an incredibly difficult engineering challenge in need of considerable investment and inquiry. Quantum transistors are still in early development, though researchers have been able to develop working prototypes.
3. **Neuromorphic Computing** technologies are modeled after living processing-structures like neurons and present a fascinating path forward beyond Moore’s Law. **Neurotransistors**, spiking neural networks (**SNN**, Intel investigating SNN applications in AI), The Human Brain Project (**HBP**), **SpiNNaker** System, **BrainScaleS** Systems (employs analog and mixed-signal components), and DNA as a new paradigm for data storage, are some of the cutting-edge areas within neuromorphic computing.

Google, the leader in search, has been researching and developing quantum computing technology since 2016 and hopes to have a quantum computer by 2029. Quantum computing could find applications in a diverse set of fields including cryptography, machine learning, medicine, materials, and big data searching and mining.

We can think of prolonging and overcoming technologies in terms of geometric scaling, that is, making components smaller and more efficient, and functional scaling, which involves garnering greater performance at a given feature size or rendering feature size irrelevant by transforming computing structure to new base components and system architectures. In addition to implementing functional scaling using existing transistor technology, functional scaling can also be achieved through paradigm shifts in computing irrespective of transistor size. We will need both in the coming decades to fuel the relentless pace of technological development and innovation.

1. Synopsis (SNPS) is a leader in of synthesis tools. [↑](#footnote-ref-2)
2. Synopsis, Cadence, and Mentor Graphics are the three biggest EDA companies. [↑](#footnote-ref-3)