Celebrating 75 years of the transistor A look at the evolution of Moore's Law innovation

A.B. Kelleher
Technology Development, Intel Corporation, Hillsboro, OR 97124, USA

***Abstract*—**For 75 years, transistor and integrated circuit (IC) innovations have primarily served as the fundamental engine of scaling for electronic devices. Moore's Law, predicting functional integration increases over time, continues to be built upon a foundation of semiconductor process scaling. As needs for functional integration increase, classes of co-optimization opportunities have become prevalent. Design-technology-co-optimization (DTCO) has been leveraged. More recently, the industry began to implement system-technology-co-optimization (STCO) techniques to further advance functional integration.

I. Celebrating 75 years of the transistor

The invention of the point contact bipolar transistor in 1947 [1] provided the world with a powerful switch to control electric current and the seed for the cost effective scaling of electronics. The foundation of the digital age was laid with the invention of the IC enabling the transistor and other circuit components to be miniaturized.

In 1965 [2] Gordon Moore observed and described the semiconductor industry's ability to climb the electronics complexity curve and to cost effectively deliver at high volume the doubling of the number of components per integrated function roughly every year. That rate was revised in 1975 to doubling every 2 years. "Moore's Law" provided the basis for understanding how ICs would revolutionize the digital world.

The semiconductor industry's dedication to Moore's Law has allowed the transistor to remain a critical enabling technology many decades after its invention. This is largely because where there is a significant challenge, engineers and scientists see an opportunity to innovate. The fact that challenges and innovation opportunities are fundamentally two sides of the same coin has become part of the fabric of the semiconductor industry.

Moreover, the semiconductor industry has never allowed itself to be overwhelmed by the enormity of Moore's Law cadence. It has always excelled at identifying the near and long term bottlenecks to integrating more functionality and making the incremental innovations needed to resolve them. This process of repeatably conquering the next hilltop, through both incremental engineering and fundamental research, is doable and rewarding. This method of progression has been the cornerstone principle behind the sustained Moore's Law cadence.

II. Focus area evolution of Moore's Law

Technology scaling for product benefits has persisted over many decades, through revolutionary and evolutionary innovations. These innovations unblock bottlenecks to greater integrated functionality.

Dennard Scaling: In 1974, Robert Dennard et.al. [3], wrote the seminal paper describing transistor scaling rules that enabled simultaneous improvements in performance, power reduction, and sustained density improvement. The principles in Dennard's work were adopted by the semiconductor industry as an effective roadmap for driving Moore's Law over the next 30 years giving us a predictable path to sustained transistor technology improvements. Examples of major breakthroughs circumventing bottlenecks were, (a) innovative immersion lithography to pattern features [3] below wavelength of light to continue density scaling, (b) innovative process and tools for atomic scale precision engineering of ultra thin gate oxides, and ultra shallow junctions to address the bottleneck of electrostatic control at sub 30nm gate lengths, and (c) wafer size transitions from 100mm through 300mm to improve factory throughput and reduce cost.

Post-Dennard Scaling: While Dennard scaling rules helped realize substantial benefits of Moore's Law, it did not factor in transistor subthreshold and gate leakage into its power dissipation model. By the mid-2000s, the sustained reductions in transistor threshold voltage and gate oxide thickness, to support voltage scaling for power reduction, started to result in leakage currents that rivaled or exceeded transistor switching energy. In addition, simple dimensional scaling of interconnect led to resistivity bottlenecks that threatened to limit circuit performance. This led to additional focus areas of scaling that required additional innovations, along mainly three different paths that will continue to co-exist into the future to enable continued performance improvement and power reductions.

Innovation path 1: Lithography, Materials & Device Architecture: Improving the resolution of lithography exposure tools has been the fundamental driver of scaling since the beginning of the semiconductor industry. Introduction of high-NA EUV to HVM is the next step to significantly improve lithography resolution. High-NA EUV tools are the most complex machinery the world has seen. Innovations in new materials and/or device architectures provide breakthroughs needed to address unique bottlenecks limiting compute performance and cost. Some examples are (a) for transistors: strain-Si (mobility gain), High-k/metal gate (gate leakage reduction), FinFETs (improved electrostatics enabling continued voltage scaling), and (b) for interconnects: proliferation of low resistance Cu (displacing Al) with chemical mechanical polish as novel integration capability to support scaling pathways for denser and multilevel interconnects roadmap, and Low-k for continued scaling of routing power and latency.

Innovation path 2: Design-Technology Co-Optimization: Building upon the first path, over time design and technology experts worked together to identify opportunities through DTCO to exceed the benefits of dimensional scaling or pure material/device innovation while addressing other bottlenecks to continue technology scaling. Advances in electronic design automation (EDA) capabilities unleashed rapid design prototyping that is used today to explore a wide range of technology features that provide a substantial fraction of overall technology entitlement at present. DTCO resulted in innovations such as contact over active gate (COAG) to reduce logic library cell heights, fin trench isolation (FTI) to reduce spacing between digital logic cells, and reduced logic library cell heights via fin depopulation. Also noteworthy is the DTCO approach to address bottlenecks around interconnect routing with increasing density. Careful co-optimization of interconnect stack design, EDA place and route and layer fill algorithms continue to result in significant performance enhancement on each technology node. DTCO is an essential part of sustaining technology scaling today.

For example, to continue scaling cell height, we need to develop more complex interconnect schemes. Moving the power lines to the backside of the wafer, a technique called PowerVia (Fig 1), enables more cell height and performance scaling than a simple geometric shrink. Another example is the next major architecture for transistor scaling called RibbonFET, or Gate All Around, Fig 2. With the move to RibbonFETs, performance scaling is obtained by adding additional nanoribbons. Every additional ribbon improves drive current.

Innovation path 3: System-Technology Co-Optimization: Today, the industry faces a new set of challenges and opportunities in optimizing system performance leveraging continued technology scaling. Delivering effective memory bandwidth and efficient power delivery are key challenges to translating technology scaling into system performance. Additionaly differential scaling rates for core logic (standard cells) and cache (SRAM) combined with HPC architectures desire for high cache/core is driving opportunities by disaggregating large caches away from most advanced nodes. This requires significant and scalable innovations in die and wafer stacking for optimal performance and total cost.

Going forward, semiconductor processing, materials, and device architecture innovations along with DTCO and STCO will continue to be important innovation paths for scaling technology to realize accelerated computer needs of upcoming generations.

III. STCO BENEFITS AND CHALLENGES

In the pursuit of Moore's Law greater functional integration 3D-IC, the first step towards STCO, is an optimization of the silicon content within a package. 3D-IC achieves greater functionality by bringing more components within a package. The role of packaging and its contribution to Moore's Law scaling is evolving and is enabling entirely new avenues of system optimization. Until the 2010s, the primary role of packaging was to route power and signaling between the motherboard and silicon, and to protect the silicon. Now, emerging 2D and 3D stacking technologies give architects and designers the tools to integrate heterogeneous technologies in a compact package and further increase the number of transistors per device by interconnecting multiple chiplets at higher bandwidth and connection density. Moore predicted that the focus area of functional integration would evolve. His 1965 paper states, "It may prove to be more economical to build large systems out of smaller functions, which are separately packaged and interconnected. The availability of large functions, combined with functional design and construction, should allow the manufacturer of large systems to design and construct a considerable variety of equipment both rapidly and economically". Today, packaging is being done at a fab level, with actual wafers. The lines between wafer fab and chip packaging have been blurred to the point of being indistinguishable.

As more and more functionality is integrated in a package, where the system is essentially collapsing into the package, the amount of silicon exceeds that which can be built within lithography reticle limits. The functionality must be split across multiple silicon components, with advanced packaging techniques providing low latency, low power, high bandwidth interconnections between the multiple die. Cost optimization of yield pushes the maximum die size to lower levels, driving further silicon disaggregation into smaller chiplets. Once disaggregated, choices become available for optimization of each chiplet's design and silicon process features, cost, functionality, and IP block availability.

STCO is a larger level of functional integration wherein all individual areas of the system: the software (manifested as workloads), the system architecture, the design engineering, the intellectual property building blocks, the silicon wafer fabrication consisting of transistors and internconnects (plus associated materials), the voltage regulation, the heterogeneously integrated advanced packaging, the test, and the high volume manufacturing are all co-optimized to create products enabling customer innovations and applications. Essentially one can think of STCO as assembling many technologies that once resided on entire motherboards within a single compact package. STCO starts with a perspecitve integrating the full functions of a system, and then co-optimizes each of the components. STCO relies on continued progress in each individual area of the system – hardware and software - while co-optimizing holistically. Fig. 3 is a high level illustration of the domains covered by STCO for a generic computing system. Historic norms mostly co-optimized across adjacent layers, such as silicon technology and foundational IP in Fig. 4. Fig 4 illustrates the difference of domains covered between device optimization, DTCO, 3DIC, and STCO.

The motivations for STCO are the same that have driven Moore's Law for many decades: the pursuit of removing bottlenecks to enable higher levels of integrated functionality, at lower cost, than can be achieved otherwise. STCO starts with workload analysis and application usage to assess and optimize the mix of technology types (e.g., logic, memory, analog, voltage regulation), designs, and disaggregation and re-synthesis configurations. The optimization by workload and application type enables higher levels of performance and functionality to be achieved.

As previously noted, Moore's Law is about increasing the integration of greater functionality. In STCO, each bottleneck of functionality, e.g., power or performance, can be addressed by providing co-optimization across the spectrum of silicon technology, chiplet disaggregation and re-synthesis within advanced packages for optimization of workloads and applications. This is illustrated in Fig. 5, wherein bottlenecks are removed to unblock new capabilities. This is much like what the industry focused on for silicon scaling for years, but now applied to a much broader spectrum of capability to increase integrated functionality.

As a specific example, new system design capabilities, enabled by a roadmap of die to die bond pitch scaling (starting with micro-bump then moving to hybrid bonding), is illustrated in Fig. 6. As the die-to-die bond pitch between die is reduced, higher connection density (connections per mm^2) can be achieved. Higher connection density enables functional disaggregation and new capabilities. Over the range of >10μm down to sub 1μm bond pitch, core logic to cache functionality can be disaggregated. These tighter pitches drive opportunity for separately optimized SRAM and logic technology nodes and re-synthesized with 3D packaging for lower energy, lower latency, and thermal optimized performance. Die-to-Die bond pitches from ~2μ down to ~0.1μ enable disaggregration of logic to logic functionality across block level, providing unique potential for cost per performance, power, and form factor co-optimization. One can imagine that once Die-to-Die bond pitches scale below 0.1μm, that we may have the potential to disaggregate transistor front-end-of-line and back-end-of-line interconnect processing, offering potential of fabrication supply chain optimization by parallelizing segments of otherwise lengthy process flows.

Increases in chiplet count and bond pitches below 10μm will require standardized chiplet interfaces to produce known good die for maximum package yield and rapid product validation and debug. Tighter packing of more functionality chiplets brings challenges of power delivery, power density, and heat removal. Improved EDA system planning and modeling tools are needed to iterate through the multitude of packaging technology options to enable best system performance and cost balancing thermal dissipation, power delivery and chiplet to chiplet communication bandwidth.

To harness the potential of fine grained disaggregation opportunities below ~2μm die-to-die bond pitch, further innovation in EDA tools and design methodologies is likely required. Today, most design methodologies and EDA tool methodologies are optimized for a given silicon chip, using a single, homogenous silicon technology. Hetergeneous, fine-grained logic disaggregation will require interoperability between multiple Process Design Kits (PDKs), interface design-for-test feature insertion tooling and upgrading the wide range of signoff tools to handle multiple technologies concurrently. Design methodologies for comprehending the expanding process skew, voltage, and temperature variance between multiple stacked dies with significantly different technologies needs consideration. Block architects and technologists will have new opportunities to disaggregate subsections based upon switching activity factor, leakage states or other system-relevant performance metrics between different technologies. The span of potential optimization points across workloads, design points, packaging, and silicon technologies exceeds that which are practical without wide deployment of well-developed EDA tools, likely relying upon artifical intelligence and/or machine learning techniques across engineering functions.

IV. INDUSTRY COLLABORATIONS TO ENABLE THE NEXT 75 YEARS OF INNOVATION

Going forward, advanced packaging will play an increasingly larger role in enabling power, performance, area, cost, time to market, design flexibility and reliability. Moreover, the desired timelines for the technology scaling of advanced packaging will be compressed relative to packaging transitions in the past. To allow the greatest flexibility in creating 2.5D packages and 3D stacks, chiplets from multiple foundries and vendors should be able to be seamlessly assembled. To enable this, the industry will need to adopt standard interfaces that are used across all chiplet designs and process nodes. Furthermore, current advanced 2.5D and 3D assembly techniques do not have standardized mechanical specifications such as metallurgy, dielectric composition, and surface planarity. This makes it challenging to bond chiplets from different foundries even if the electrical interfaces are standardized i.e., UCIe [7]. The industry needs to continue to work together to establish this standardization. In addition, assembly, and test (AT) factories use a plethora of carriers, trays, and magazines, for material and handling leading to labor and equipment inefficiencies. An industry standard AT material carrier, equipment load ports and equipment front end modules (EFEMS) will need to be investigated to drive efficiencies for green and brown field factories. Standardization will be critical to shortening time to market of new packaging technologies in an open chiplet ecosystem.

The insatiable demand for lower power, lower latency, and higher interconnect density to achieve novel architectures drives demand for tighter die-to-die pitch scaling. As such, there will be a need for wafer assembly tools with capability to align 3D stacked chips at nanometer interconnect pitch tolerances and run rates enabling economically viable high-volume manufacturing. Increased dimensional stability of package substrates will be needed for finer features and denser interconnects. Traditional organic packaging materials are sensitive to changes in temperature causing warpage during processing. This makes it difficult to further scale lithographically defined features across the typical large panel used during substrate manufacturing. New materials, technologies, and processing methods will be needed to break this barrier.

Bringing power through the bottom of a 3D stack through highly resistive through Si vias (TSVs) leads to efficiency losses. Innovative power delivery solutions will be needed to

enable adequately low parasitic and efficient power delivery to the chiplets in complex 3D stacks.

The logic chips lower in a 3D stack will still generate heat and will need to be cooled. Accurate, experimentally validated predictions of layout-based thermal performance needs will have to be incorporated into design engineering methods to ensure that all designs are making the optimal use of scaling. Effective cooling will need to be addressed through system design, package architecture, materials, and process integration. Critical enabling technologies such as accurate metrologies, rapid debug methods and failure analysis techniques are crucial for success.

Power, thermals, and routing constraints are bottlenecks to 3D-IC scaling. Across our industry, the technology R&D pipeline is rich with innovative ideas to address these challenges including novel transistors for density scaling (CFET, 2D) [8]; energy efficient switching (Tunnel-FETs, FeFETs, Spintronics); novel memories; and advanced packaging techniques (with Cu and/or integrated Photonics) allowing heterogeneous integration, with low connectivity overhead, of different technologies that can be independently or co-optimized within a package.

The industry may revisit III-V compound semiconductors (InGaAs/InP) for NMOS, and Ge for PMOS as they have far better electron and hole mobility than Si and thus be capable of more efficient circuit performance. These materials also have smaller bandgaps allowing transistors to effectively switch at low voltages. Other innovations might include tunnel field effect and ferroelectric transistors. These are examples of devices which could potentially hold the answers to making more efficient transistors. Likewise, transistors based on 2D transition metal dichalcogenides offer the potential to deliver Power-Performance-Area improvements. With the advent of STCO, where technologies are co-optimized starting with workloads and applications, the business case for these new technologies may be more viable than previously, where now higher value can potentially be obtained despite narrow initial product reach of a new technology.

To decrease the power spent in moving data between the processing unit and the memory, the industry will need to consider compute near memory or compute within memory architectures. The industry will need to develop new memory devices that scale sufficiently and yet can be incorporated into an integration scheme that supports logic devices. Neuromorphic computing is an example of non-von Neumann type architecture that has potential to take advantage of the integration of memory and logic. Additionally high bandwidth optical interconnects that enable rapid, long reach data movements are critical for continued system scaling [9].

Lastly, STCO will place new demands on future technologists. In addition to domain specific expertise, they will need cross disciplinary skills and knowledge to be able to holistically integrate technology at a systems level.

V. Conclusion

The tiny transistor, the cornerstone of the digital technology revolution, has transformed our society. It has unleashed new industries, unlocked human creativity anew, enabled breathtaking inventions and discoveries, influenced our society, and accelerated economic prosperity. Its evolution over 75 years, manifested in incredible products and services, is testament to the innate human spirit of innovation, creativity, industry collaboration, and enterprise. That spirit keeps alive the Moore's Law pursuit of ever-increasing functionality. With the foundation of semiconductor processing, DTCO, and now STCO becoming an enhanced tool in our Moore's Law portfolio, the entire semiconductor industry will thrive by constantly leveraging each other's unique strengths and valuable innovations. Here's looking forward to the next 75 years of incredible human ingenuity!

Acknowledgements

The author gives heartfelt thanks to all the scientists, engineers, and technicians working to innovate, at Intel and in the rest of the semiconductor industry, to deliver technologies and satisfy the world's ever-increasing demand for computing.

References

[1] "The lost history of the transistor," *IEEE Spectrum*, vol. 41, no. 5, pp. 44-49, 2004.

[2] G. E. Moore, "Cramming more components," *Electronics*, vol. 38, no. 8, 1965.

[3] R. H. Dennard, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256-268, October 1974.

[4] N. N. e. al., "Sapphire Rapids: The Next-Generation Intel Xeon Scalable Processor," *2022 IEEE International Solid- State Circuits Conference (ISSCC), 2022*, pp. 44-46, 2022.

[5] R. Mahajan, "Quiet Revolutions: How Advanced Microelectronics Packaging Continues to Drive Heterogeneous Integration," *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, vol. 10, pp. 1408-1412, 2020.

[6] G. Keeler, "Common Heterogeneous Integration and IP Reuse Strategies (CHIPS)," 2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), [Online]. Available: https://www.darpa.mil/program/common-heterogeneous-integration-and-ip-reuse-strategies.

[7] UCIe Express, Universal Chiplet Interconnect, UCIe, [Online]. Available: https://www.uciexpress.org/

[8] C. -Y. Huang*et al*., "3-D Self-aligned Stacked NMOS-on-PMOS Nanoribbon Transistors for Continued Moore's Law Scaling," *2020 IEEE International Electron Devices Meeting (IEDM)*, 2020, pp. 20.6.1-20.6.4

[9] R. M. et al., ""Co-Packaged Photonics For High Performance Computing: Status, Challenges And Opportunities," *Journal of Lightwave Technology*, vol. 40, no. 2, pp. 379-392, 2022.

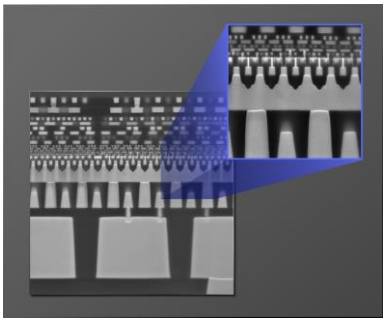

Fig. 1. Intel's backside power delivery scheme, PowerVia, that separates power and signal lines and shrinks the standard cell size. Power wires are placed beneath the transistor layer, on the backside of the wafer.

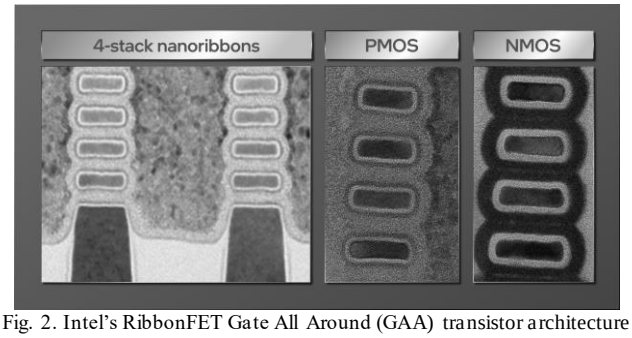

Fig. 2. Intel's RibbonFET Gate All Around (GAA) transistor architecture stacking four nanoribbons to achieve the same drive current as multiple fins, but in a small footprint.

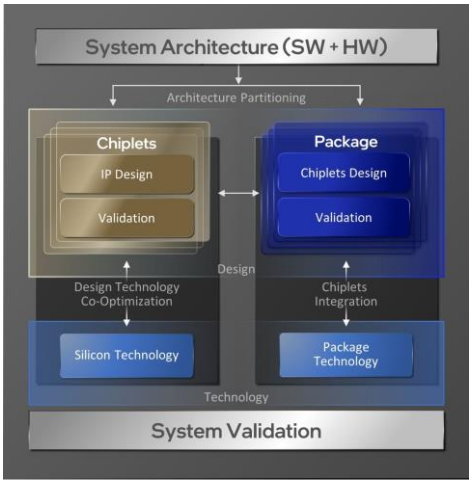

Fig. 3. System technology co-optimization of a computing system.

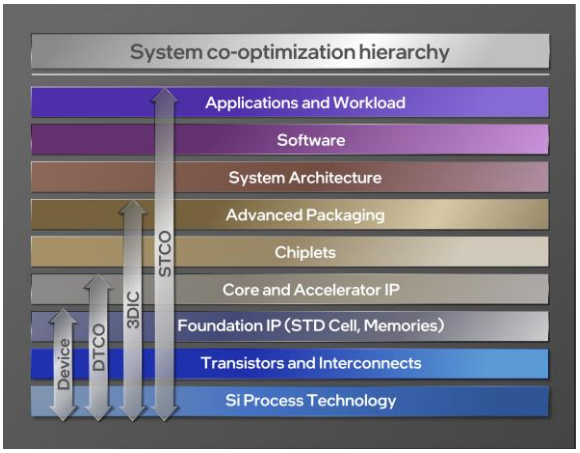

Fig. 4. Hierarchies in system technology co-optimization. The difference between device optimization, DTCO, 3DIC, and STCO is illustrated.

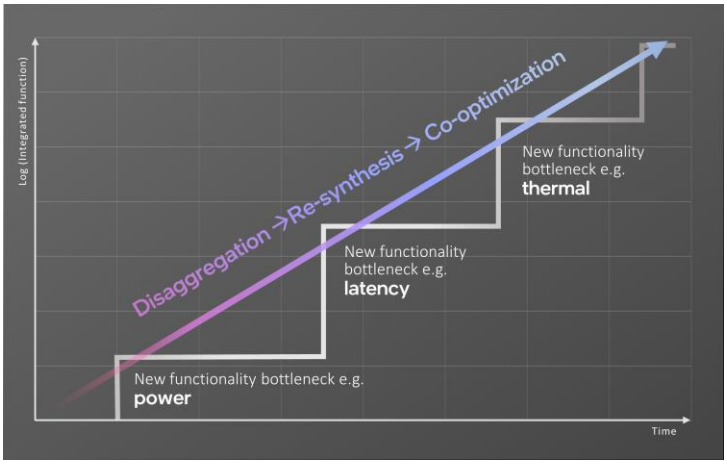

Fig. 5. Disaggregation, re-synthesis, and co-optimization being used to resolve Moore's Law bottlenecks over time.

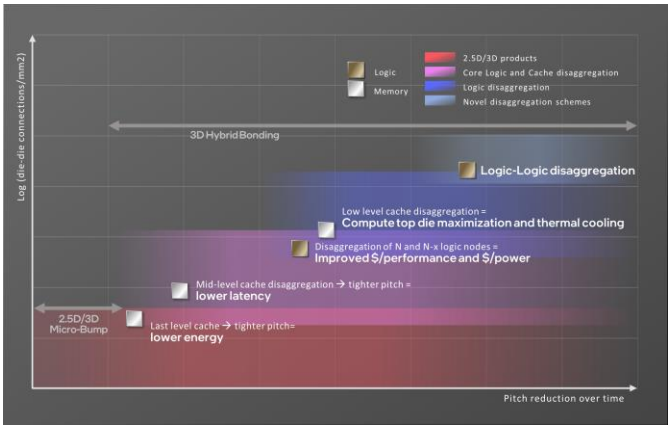

Fig. 6. Advances in die-to-die bond pitch will enable cache, logic, and novel disaggregation schemes leading to higher levels of performance and power efficiency.