Technical Highlights from the 2024 Symposium on VLSI Technology and Circuits

The 2024 IEEE Symposium on VLSI Technology and Circuits is a premier international conference scheduled for June 16-20, 2024, that records the pace, progress, and evolution of micro/nanoelectronics. The joint technology and circuits symposium will be held in person in Hawaii in the Hilton Hawaiian Village to enable networking opportunities.

The Symposium's overall theme, **"Bridging the Digital & Physical Worlds with Efficiency & Intelligence,"** integrates advanced technology developments, innovative circuit design, and the applications that they enable, as part of our global society's transition to a new era of intelligent connected devices, energy efficient infrastructure and AI enabled hardware systems that change the way humans interact with each other.

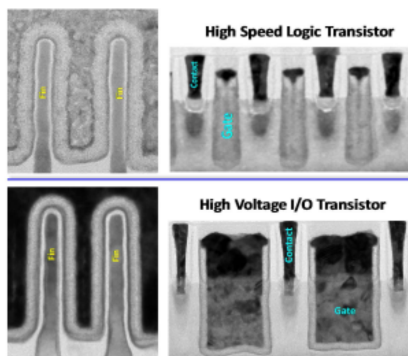
The following are some of the highlighted papers that address this theme.

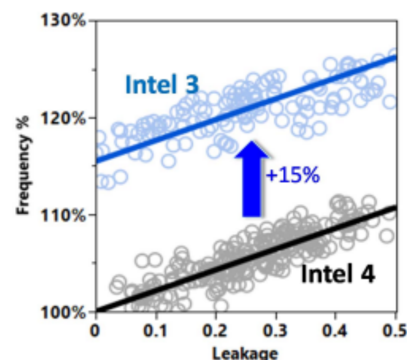
Technology Highlights

Advanced CMOS Technology

"An Intel 3 Advanced FinFET Platform Technology for High Performance Computing and SOC Product Applications" – Intel Corporation (Highlight Session – Paper T1.1)

This paper presents Intel 3 FinFET technology fully optimized, providing 10% logic scaling, performance and reliability improvement compared to Intel 4. Through transistor enhancements, interconnect optimization, and design co-optimizations, up to 18% iso-power performance gain is achieved over Intel 4. Intel 3 additionally enables a 210-nm high-density standard cell, 1.2-V-native I/O transistors, deep N-well isolation, and long-channel analog devices to provide full-featured technology design capabilities.



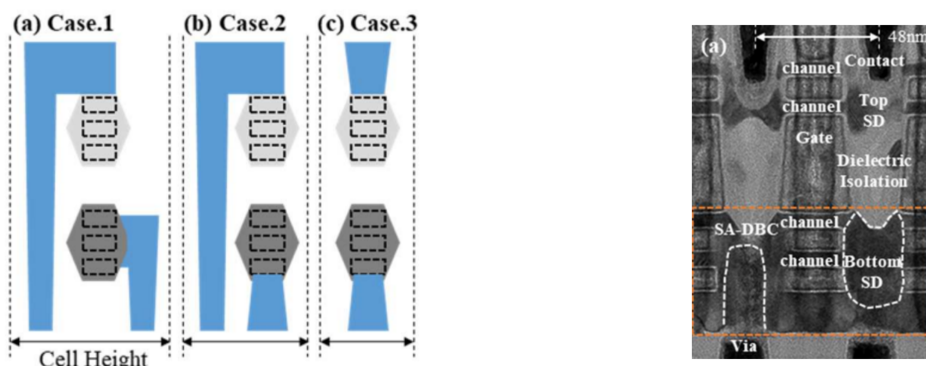


Figures: (Left) Fin/gate TEM cross-section of logic and 1.2-V I/O transistors (top/bottom respectively). (Right) 15% iso-leakage frequency enhancement compared to Intel 4.

Advanced CMOS Technology

"Highly Manufacturable Self-Aligned Direct Backside Contact (SA-DBC) and Backside Gate Contact (BGC) for 3-Dimensional Stacked FET at 48-nm Gate Pitch" – Samsung Electronics (Highlight Session – Paper T1.2)

In this study, Samsung Electronics demonstrates 3-Dimensional Stacked FET (3DSFET) with Self-Aligned Direct Back-side Contact and Back-side Gate Contact in 48-nm gate pitch, which is the smallest dimension and the world's first demonstration reported so far. Simultaneous threshold voltage (V_T) targeting for both nFET and pFET in common gate and N/P connection with vertical common contact were also verified in addition to their previous work. As a result, Samsung Electronics believes that most of key components for ultimate cell height scaling of 3DSFET have been verified to continue the logic technology scaling beyond the 1-nm node.

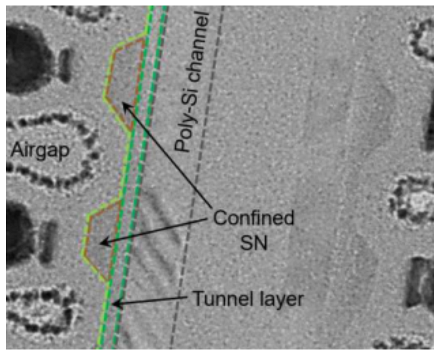


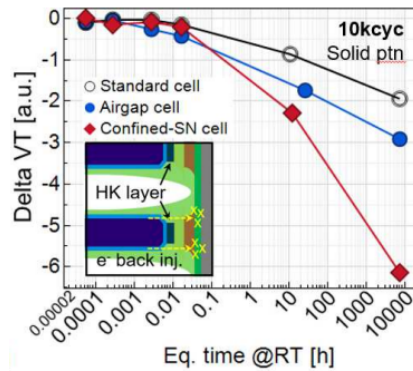
Figures: (Left) Various combinations of front-side contact and SA-DBC on 3DSFET discussed in the paper. (Right) Cross-sectional TEM image of co-integrated TOP-nFET and BOTTOM-pFET.

Memory Technology

"A Confined Storage Nitride 3D-NAND Cell with WL Airgap for Cell-to-cell Interference Reduction and Improved Program Performances" – Micron Technology Inc. (Highlight Session – Paper T1.3)

In this paper, Micron demonstrates a confined Storage Nitride (SN) 3D-NAND cell with an innovative process flow including Word Line (WL) airgap formation. Airgaps strongly reduces WL parasitic capacitance which translates into better program time performances. A complete device characterization was carried out on a test memory array. The authors measured substantial cell-to-cell interference improvements and lateral charge loss reduction, which make this cell a key enabler of further tier pitch scaling in future 3D-NAND arrays. Program-erase window limitations due to trapped charge confinement were also addressed by TCAD modeling, showing that window can be recovered by Storage Nitride film thickness changes with no cell-to-cell penalty.



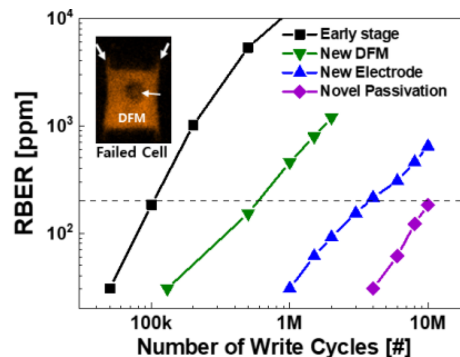


Figures: (Left) TEM x-section of a single pillar in confined-Storage-Nitride cell 3D-NAND array with airgaps. (Right) Vertical charge loss in different cell structures after 10k program-erase cycles, showing superiority of airgap cell.

Memory Technology

"First Demonstration of Fully Integrated 16-nm Half-Pitch Selector Only Memory (SOM) for Emerging CXL Memory" – SK Hynix Inc. (Highlight Session – Paper T1.5)

SK Hynix conducts a study to fully understand the switching mechanisms of Selector Only Memory (SOM). This work led to implementation of TCAD and the development of both advanced materials and processes based on the optimized core circuit design and write-read scheme in the first fully integrated 16-nm half-pitch SOM for emerging Compute Express Link™ (CXL) memory. SK Hynix achieves 750-mV read window margin including product-level raw bit error rate and reliability figures such as drift-related persistency, read disturbance, high temperature data retention (>10 years at 125 °C), and cycle endurance for 200 ppm raw bit error rate.



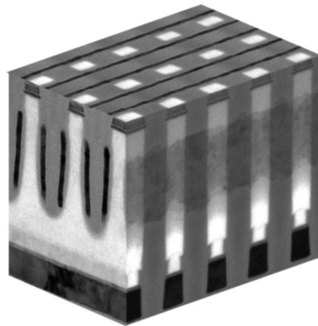
Figures: (Left) Cross-sectional TEM image of 16-nm half-pitch cross-point Selector-Only-Memory with periphery under cell architecture (Right) Cycle endurance characteristics of measured raw bit error rate with write cycles.

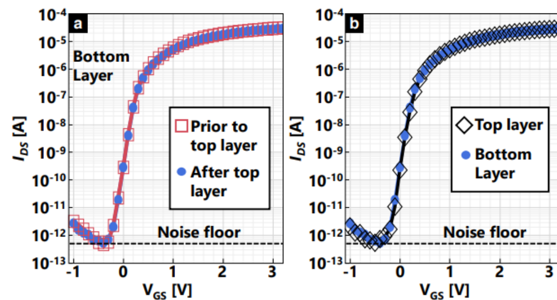
Memory Technology

"$4F^2$ Stackable Polysilicon Channel Access Device for Ultra-Dense NVDRAM" – Micron Technology Inc. (Paper T17.2)

In this paper, Micron Technology reports on the methodology and optimization used to enable a stackable $4F^2$ polysilicon Thin Film Transistor (TFT) for ultra-dense 32-Gb NVDRAM. Several key innovations are implemented to meet the strict thermal budget

constraints required for dual-layer technology. Confined heating from pulsed laser annealing is used to crystallize polysilicon and activate source/drain dopants. Material is optimized to engineer both a gate oxide deposited at low temperatures capable of 10 years equivalent reliability and a ruthenium (Ru) word-line robust to agglomeration and voiding failures. Device performance, robust to top layer processing, is matched across both layers by adjusting process conditions as informed by a TCAD model that accounts for heat transfer and crystallization dynamics.



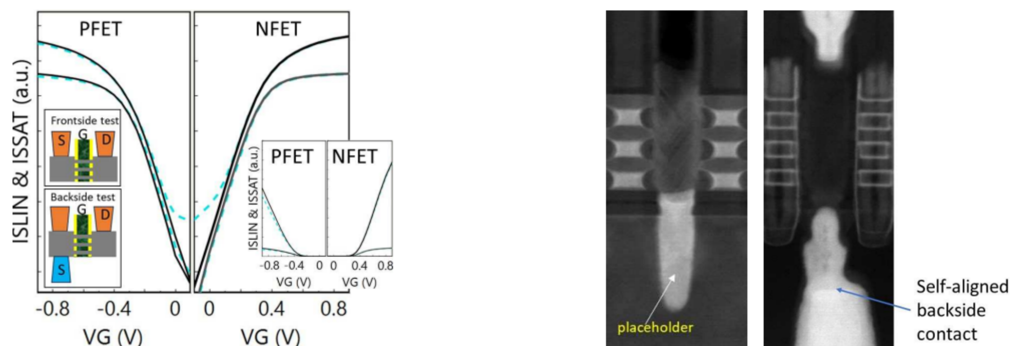


Figures: (Left) Cross-sectional image of dual-gated (Ru) TFT access device for NVDRAM. (Right) IDS-VGS curves comparing TFT measured on the bottom layer before and after top layer TFTs.

3D Technology

"Backside Power Distribution for Nanosheet Technologies Beyond 2 nm" – IBM Research and Samsung Electronics (Paper TFS2.3)

This joint IBM and Samsung paper examines various approaches for integrating Backside Power Distribution Network (BSPDN) with nanosheet transistor technologies. Deep Trench Via based on BSPDN schemes, except for Shifted Frontside Via Backside Power rail, do not offer cell level scaling benefits, but via resistance could remain a bottleneck. Direct Backside Contact-based schemes offer the best cell level scaling. Finally, a novel self-aligned backside contact scheme integrated with nanosheet transistors is demonstrated with immunity to misalignments in backside contact formation. The structure exhibits good device characteristics and satisfactory reliability.

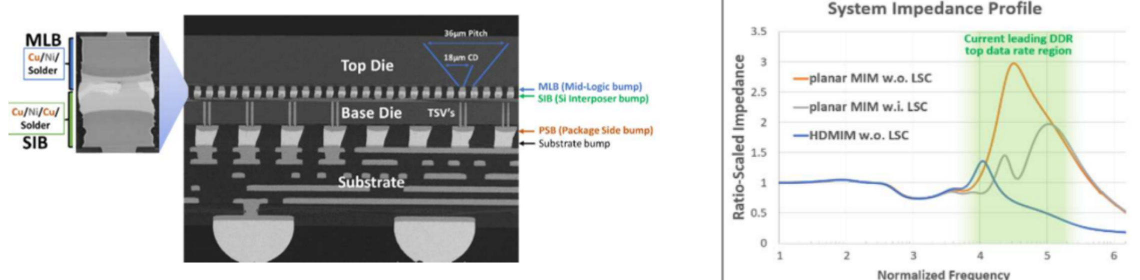


Figures: (Left) Experimental linear and saturation transfer curves showing closely matched electrical behaviors between a frontside contacted device -dashed blue- and the same backside direct backside contact device -black solid-. (Right) TEM image of a transistor after placeholder and source/drain epi formation and with self-aligned backside contact.

3D Technology

"Integration of Si-Interposer and High Density MIM Capacitor on 2.5D Foveros Face-to-Face Architecture" – Intel Corporation (Paper T9.1)

Integration of different computing elements through silicon interposers enables scaling opportunities beyond Moore's law. Intel's passive Si-Interposer enables interconnections among different chiplets using Through Silicon Via (TSV) technology along with a refined 36-µm microbump pitch in a face-to-face die configuration. The Si-Interposer houses a High-Density Metal-Insulator-Metal (HDMIM) integrated decoupling capacitor for voltage droop reduction and noise suppression. Products can either utilize HDMIM in the Si-Interposer die, built in HDMIM in the chiplet die, or both. Intel's paper describes high-density metal-insulator-metal fabrication steps, electrical properties, reliability benchmarks, and performance enhancements by incorporation of Si-Interposer HDMIM.

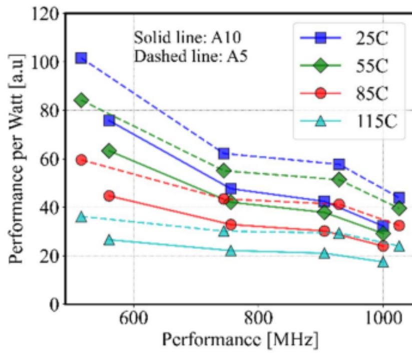


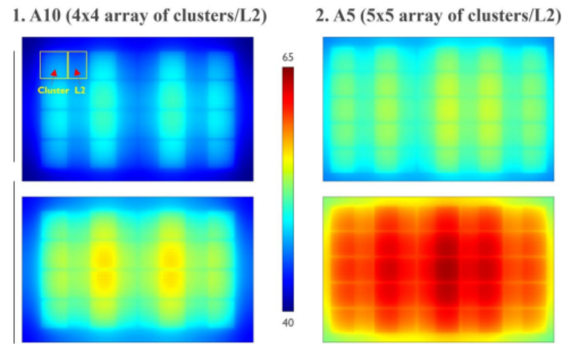
Figures: (Left) Cross section showing interposer connection to top die through silicon interposer bump and package. (Right) High-density metal-insulator-metal impact of system-level impedance profile for DDR.

3D Technology

"Thermal Considerations for Block-Level PPA Assessment in Angstrom Era: A Comparison Study of Nanosheet FETs (A10) & Complementary FETs (A5)" – IMEC (Paper T5.4)

In this paper, IMEC proposes a thermal-aware block-level PPA comparison study for NanoSheet transistors (NSFET) and Complementary Field Effect Transistors (CFET), expected to be used in future Angstrom nodes, namely A10 and A5 respectively. They reported block-level scaling results from A10 to A5 node on an open-source many-core architecture: 2.5% increase in F_{max}, 25% reduction in power, 27% reduction in energy per cycle, achieved with 35% area reduction and a consequent increase in power density by 15% under nominal condition of 0.7 V and 25 °C. The PPA analysis methodology was augmented with a fast package-level thermal simulator to enable early self-consistent thermal estimation that accounts for exponential leakage power increase with temperature, which is important for dynamic thermal management applications. The analysis reveals a reduction of 64 mV in Vdd and 10% in frequency required for A5 node to maintain same $T_{j,max}$ as A10 node operating at 0.7 V, still resulting in a 40% gain in system throughput.



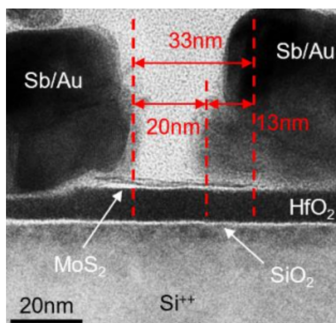


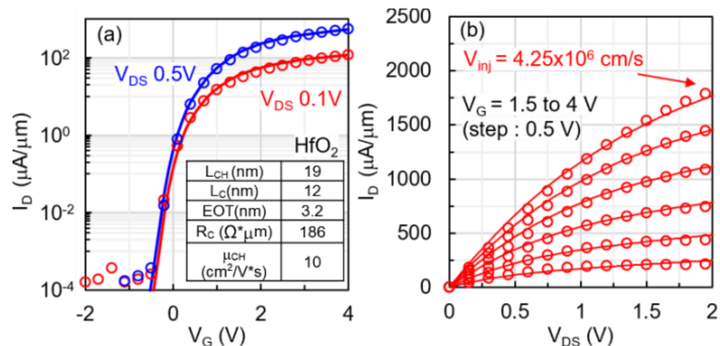
Figures: (Left). Performance/Watt comparison between A10 and A5 nodes, showing higher efficiency for A5 node across all temperatures. (Right) Iso-die area SOC level thermal maps using T-independent power (top) and self-consistent power for both A10 and A5 operating at 0.7V (bottom).

Beyond CMOS Technology

"On the Extreme Scaling of Transistors with Monolayer MoS_2 Channel" – TSMC and National Yang Ming Chiao Tung University (Highlight Session – Paper T1.4)

2D Transition Metal Dichalcogenides (TMDs) show promise for transistor scaling, but their on-scale performance had not been proven yet. In this work, collaborators at TSMC and National Yang Ming Chiao Tung University demonstrate contact length scaling while holding a low contact resistance down to 11 nm. Channel length scaling shows I_{ON} can increase down to at least 12 nm with low RC. The very scaled (channel length = 19 nm) MoS_2 transistor with Sb-based metal contact has current density of ~1130 μA/μm at V_{DS} = 1 V, and a low contact resistance of ~190 Ω·μm. These scaled transistors, processed within a back-end-of-line (BEOL) thermal budget, do not exhibit subthreshold swing degradation or observable drain-induced barrier lowering down to 12-nm channel length.



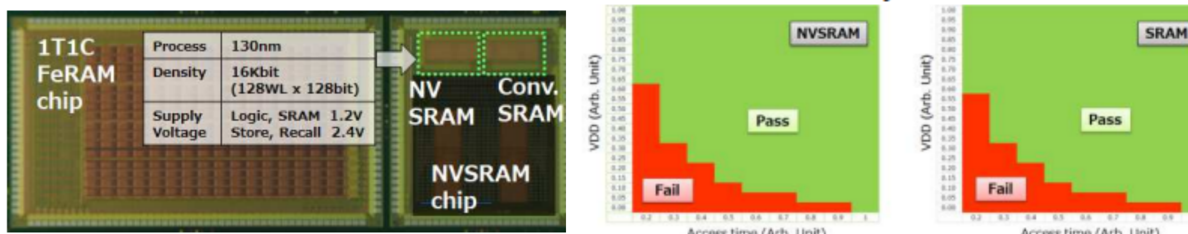


Figures: (Left) TEM images of MoS_2 device showing aggressive scaling – channel length of 13 nm. (Right) Transfer and output characteristics of a device with 3.2-nm EOT. Symbols are experimental data, lines are TCAD model.

Beyond CMOS Technology

"HZO-based Nonvolatile SRAM Array with 100% Bit Recall Yield and Sufficient Retention Time at 85 °C" – Sony Semiconductor Solutions Corporation, Fraunhofer IPMS, and NaMLab (Paper T2.1)

For the first time, the paper led by Sony experimentally demonstrates 100%-bit yield on a 16-kbit NonVolatile SRAM (NV-SRAM) array based on a metal/ferroelectric/metal capacitor using a sub-10-nm thick $HfZrO_x$ (HZO) layer. This capacitor is formed using the same integration process as that of a previously developed ferroelectric random-access memory (FeRAM) array on the same wafer. Its sequential operations of nonvolatile data store, cutoff of power supply (power-gating), and data recall are completely executed employing a robust recall sequence, achieving 100%-bit recall after a 200-s power-gating period at 85 °C even with sufficiently low operation voltage. Sony's results indicate that HZO-based NV-SRAM and FeRAM hybrid memory system can provide ultra-low power advantages in a System-on-Chip for Internet of Things edge computing.

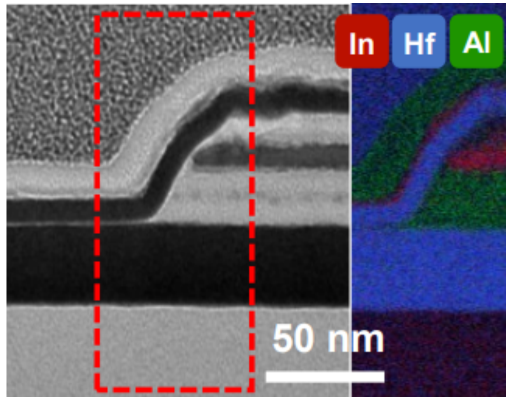


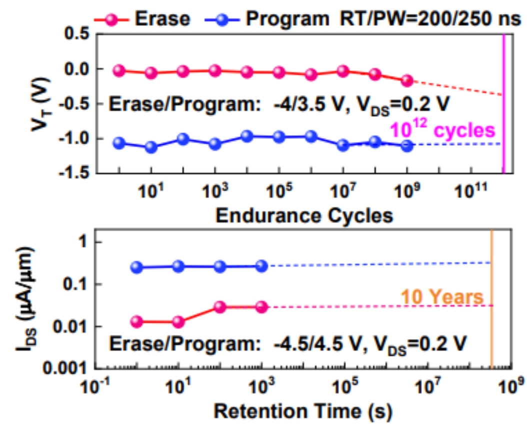
Figures: (Left) Photo image of NonVolatile SRAM and Ferroelectric RAM chips formed on one wafer. The NVSRAM chip also has conventional SRAM macro for comparison. (Right) Shmoo plots between access time and supply voltage during active operation for NonVolatile SRAM and SRAM arrays.

Beyond CMOS Technology

"Highly Robust All-Oxide Transistors with Ultrathin In2O3 as Channel and Thick In2O3 as Metal Gate Towards Vertical Logic and Memory" – Purdue University and Samsung Electronics (Paper T4.1)

In this work, collaborators at Purdue University and Samsung report for the first time atomic-layer-deposited (ALD) all-oxide transistors toward 3-D vertical integration, with thick ALD In_2O_3 as gate electrodes and In_2O_3 itself as contact. The all-oxide Thin-Film Transistors (TFTs) show an on/off ratio over 106, high uniformity, and very robust reliability with a threshold voltage shift of 5 and 50 mV in positive and negative bias stress tests. The vertically all-oxide TFTs demonstrate good control from side wall with on/off ratio over 105 and maximum current (I_{max}) over 160 µA/µm. Furthermore, vertically all-oxide ferroelectric field-effect transistors (Fe-FETs) exhibit a memory window of 1.85 V, with an endurance and retention extended to 10^{12} cycles and 10 years. This illustrates that the vertically all-oxide device based on ALD oxide semiconductor is a good candidate toward future high-density integrated circuits.





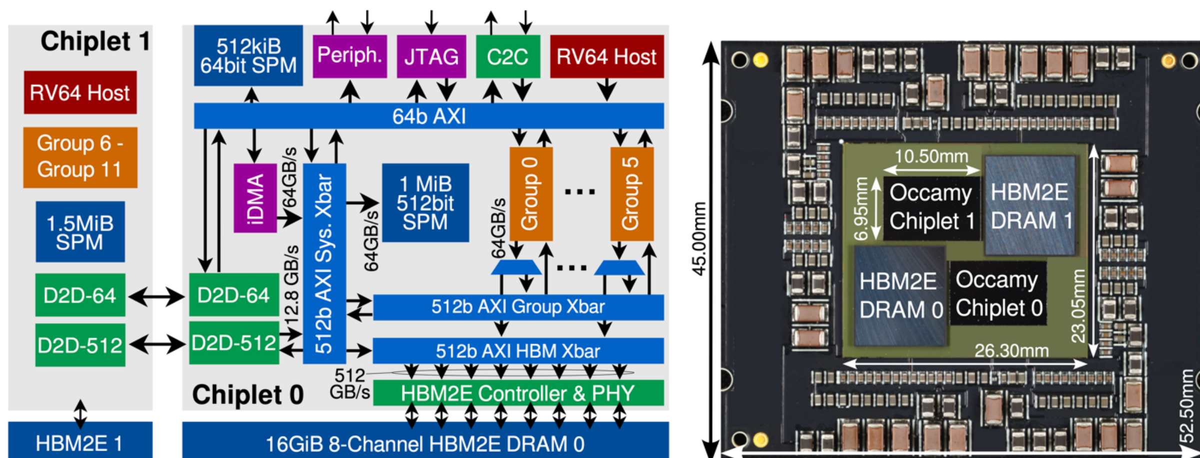
Figures: (Left) High-Resolution TEM cross-section image and EDS mapping of an ALD vertical all-oxide FET 10nm In_2O_3 dielectric. (Right) Major states endurance and retention performance of an ALD vertical In_2O_3 Fe-FET at room temperature.

Circuits Highlights

Processors and SoCs

"Occamy: A 432-Core 28.1 DP-GFLOP/s/W 83% FPU Utilization Dual-Chiplet, Dual-HBM2E RISC-V-based Accelerator for Stencil and Sparse Linear Algebra Computations with 8-to-64-bit Floating-Point Support in 12nm FinFET" – ETH Zürich, Stanford University, and University of Bologna (Paper C7.4)

Collaborators at ETH Zürich, Stanford University, and the University of Bologna present a flexible, general-purpose, dual-chiplet system with two 16-GB HBM2E stacks optimized to address a wide range of irregular-memory-access compute workloads with high utilization. Codenamed Occamy, the heterogeneous system comprises a 432-core RISC-V dual-chiplet 2.5-D system for efficient sparse linear algebra and stencil computations on FP64 and narrow (32-, 16-, 8-bit) SIMD floating-point data. Occamy features 48 clusters of RISC-V cores with custom extensions, two 64-bit host cores, and a latency-tolerant multi-chiplet interconnect and memory system with 32 GB of HBM2E. Silicon demonstrates leading-edge utilization on stencils (83%), sparse-dense (42%), and sparse-sparse (49%) matrix multiplication.



Figures: (Left) Dual-chiplet architecture. (Right) Module photograph.

Devices and Accelerators for Machine Learning

"Dyamond: A 1T1C DRAM In-memory Computing Accelerator with Compact MAC-SIMD and Adaptive Column Addition Dataflow" – KAIST and Samsung Electronics (Paper C20.1)

Collaborators at KAIST and Samsung Electronics propose a 1T1C DRAM in-memory computing accelerator to exploit increased memory density with enhanced system energy efficiency by reducing memory access. Codenamed Dyamond, the accelerator features column addition (CA) dataflow for high density and energy efficiency. LSB-CA minimizes ADC readouts to increase energy efficiency. MSB-CA with signal-enhanced multiple-accumulate (MAC) and signal-shifted ADC enhances SQNR to further improve energy efficiency. A switchable sense amplifier reduces read energy for low-power in-memory arithmetic SIMD. Fabricated in 28nm CMOS and integrating 27-Mb DRAM memory within a 6.48-mm^2 die area, Dyamond achieves a peak energy efficiency of 27.2 TOPS/W and outstanding performance in advanced ML models (ResNet, BERT, GPT-2).

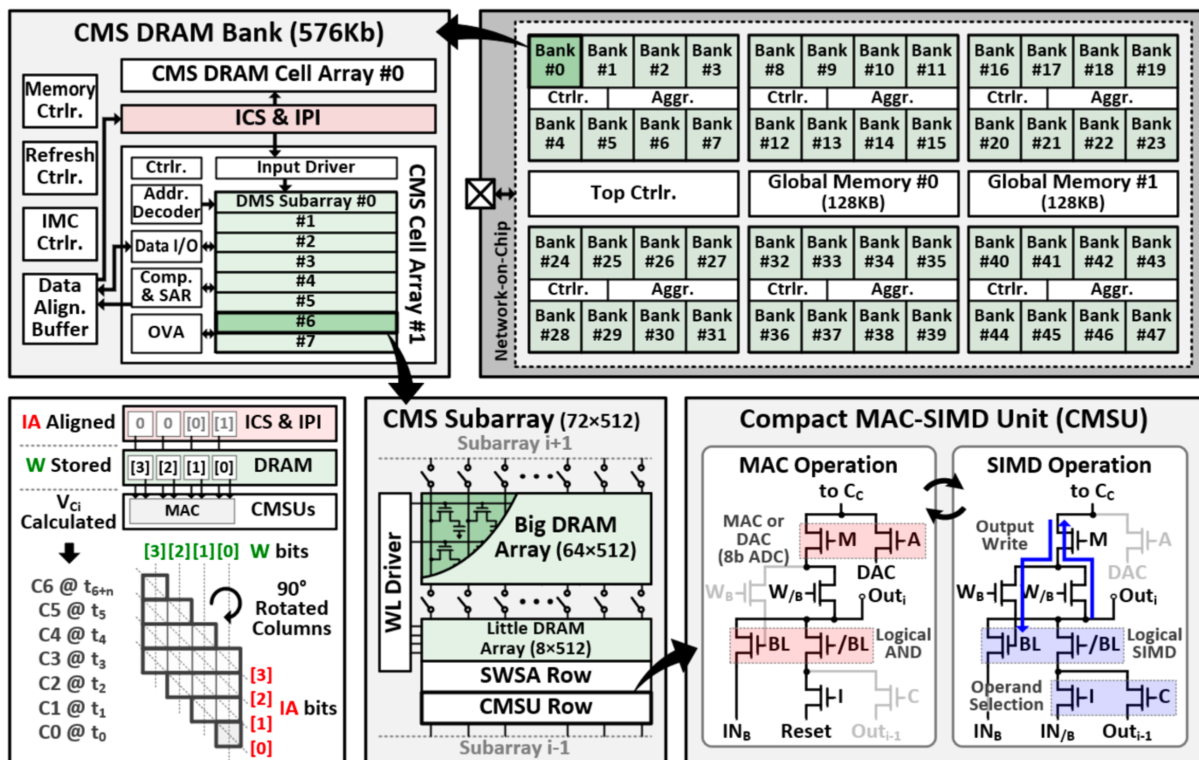


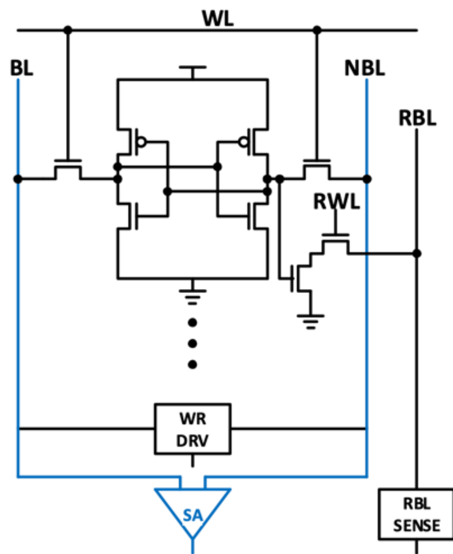
Figure: Overall architecture of Dyamond.

Memory Technologies, Devices, Circuits, and Architectures

"A 7GHz High-Bandwidth 1R-1RW SRAM for Arm HPC Processor in 3nm Technology" – Arm (Paper C16.3)

Authors at Arm demonstrate a 1Read-1ReadWrite (1R1RW) High Bandwidth Instance (HBI) L1-data cache memory architecture in 3-nm CMOS which is seamlessly integrated into Arm's flagship high-performance processor. Enhancing the conventional 8T-1R1W memory, HBI features an additional read port to achieve 1R1RW capability. HBI memory in the L1-

data cache doubles the available read bandwidth and results in an improvement in processor IPC, exceeding 1%. The new architecture further enables 13% reduction in area and 10–15 ps reduction in routing delay from less routing congestion in the CPU physical design. Silicon demonstrates 1R1RW HBI yield of 100% yield, the highest reported frequency of over 7 GHz and lowest reported bit density of 11.2 Mbit/mm^2 for any 8T SRAM memory.



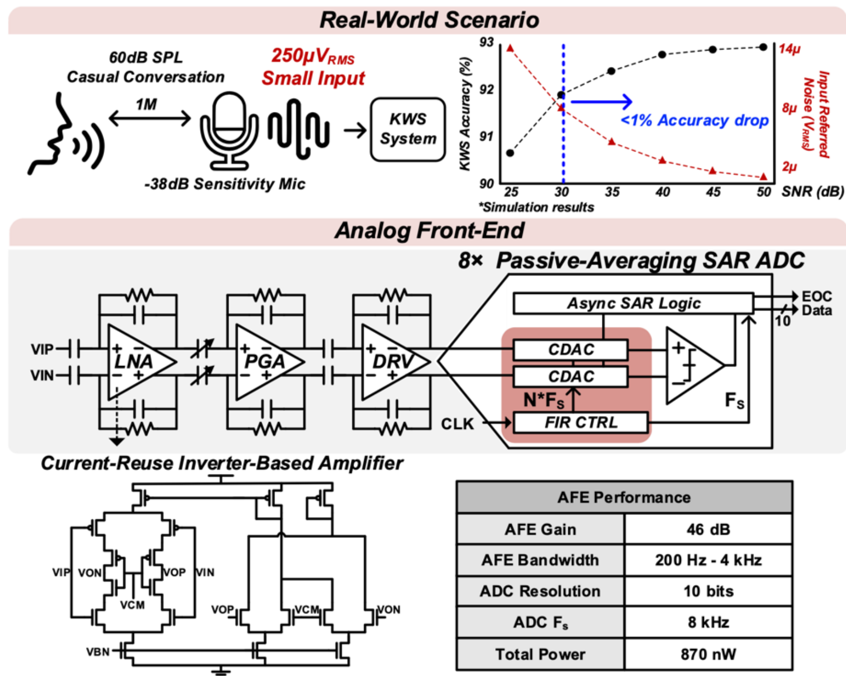
Parameter	1RW	1R1RW HBI
Bitcell	6T (HC)	8T (TP)
Read Ports	1	2
Bitcell Area (a.u.)	1	1.33
Macro Area (μm^2)	1312	1739
L1-D cache (μm^2)	83991 (128KB)	55636 (64KB)

Figure: 1R1RW HBI memory cell architecture offering 33% reduction in L1-Data cache area compared to 6T-1RW memory.

Digital Circuits, Hardware Security, Signal Integrity, IOs

"A 5.6μW 10-Keyword End-to-End Keyword Spotting System Using Passive-Averaging SAR ADC and Sign-Exponent-Only Layer Fusion with 92.7% Accuracy" – Seoul National University and Columbia University (Paper C25.1)

Researchers from Seoul National University and Columbia University present a 10-keyword end-to-end keyword spotting (KWS) system to target wake-up and control of mobile and IoT devices. The proposed architecture employs passive averaging to improve the analog front-end (AFE) SNR and operate at a small power overhead of only 20 nW. A sign-exponent-only layer fusion scheme reduces the model size and multiplication power overhead by 63.5% and 29.8%, respectively, maintaining the KWS accuracy. Compared to prior art targeting 10 keywords, the design offers the highest accuracy of 92.7% and lowest power consumption of 5.6 μW.



AFE Performance	
AFE Gain	46 dB
AFE Bandwidth	200 Hz - 4 kHz
ADC Resolution	10 bits
ADC F_s	8 kHz
Total Power	870 nW

Figure: AFE design constraints, architecture, and performance.

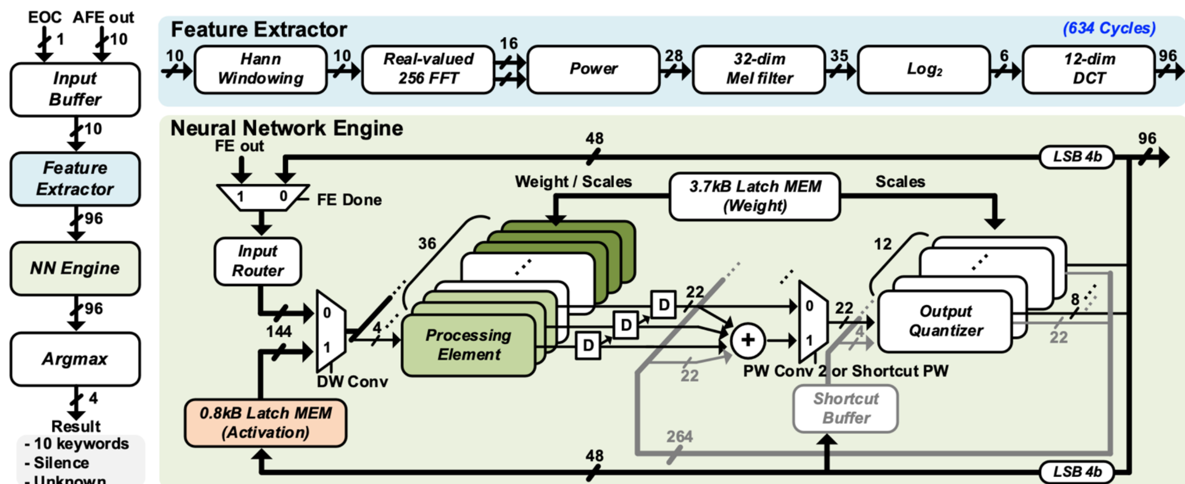
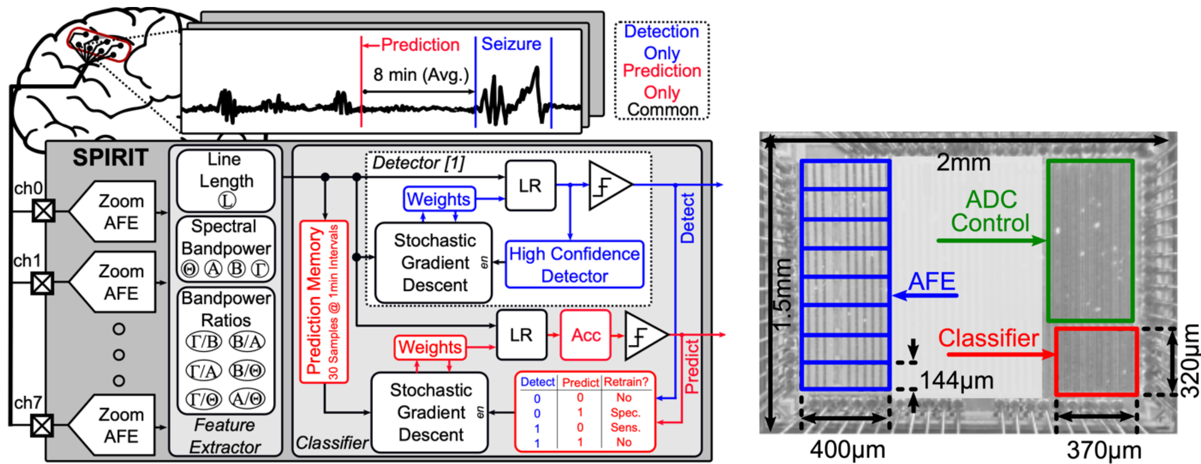


Figure: Digital back-end architecture.

Biomedical Devices, Circuits, and Systems

"SPIRIT: A Seizure Prediction SoC with a 17.2 nJ/cls Unsupervised Online-Learning Classifier and Zoom Analog Frontends" – University of California, Berkeley (Paper C23.1)

Authors from the University of California, Berkeley present an SoC named SPIRIT which integrates an unsupervised online-learning seizure prediction classifier. The work features eight 14.4-μW, 0.057-mm^2, 90.5-dB dynamic range, and Zoom Analog Frontends. On average, SPIRIT achieves 97.5%/96.2% sensitivity/specificity and can predict epileptic seizures 8.4 minutes before they occur. Its classifier consumes 17.2 μW and occupies 0.14 mm^2, the lowest reported for a prediction classifier by >134X in power and >5X in area.

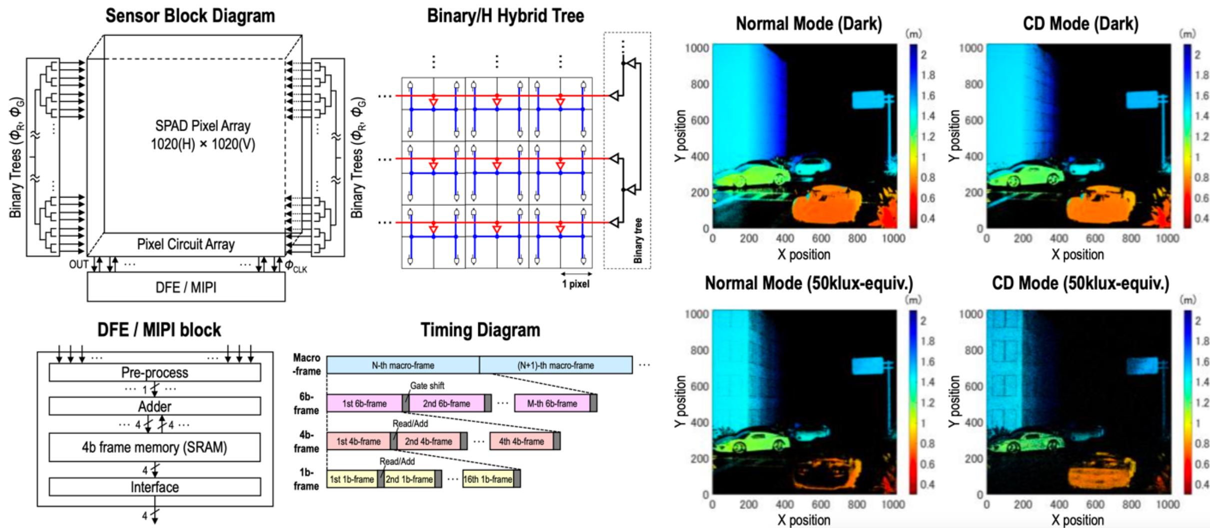


Figures: (Left) SPIRIT system level diagram. (Right) Chip micrograph.

Sensors, Imagers, IoT, MEMS, Display Circuits

"3D-Stacked 1Megapixel Time-Gated SPAD Image Sensor with 2D Interactive Gating Network for Image Alignment-Free Sensor Fusion" – Canon Inc. (Paper C6.1)

Canon presents a 5µm-pitch, 3D-BSI 1Mpixel time-gated SPAD image sensor featuring 2D interactive gating network, enabling image alignment-free sensor fusion. The SPAD image sensor operates at 1,310 fps for global shutter 2D imaging, and event vision sensing with 0.76-ms temporal resolution under 0.02 lux. Through range-gated imaging, this work demonstrates a feasibility of robust imaging under harsh environments. The gating network architecture enables background suppression in 3D depth measurement under 50-klux ambient light.



Figures: (Left) Sensor block diagram, clock tree configuration, and timing diagram. (Right) Measured 3D depth maps with normal- and CD-mode, captured under dark and 50klux-equivalent ambient light.

Data Converters

"A 16GS/s 10b Time-domain ADC using Pipelined-SAR TDC with Delay Variability Compensation and Background Calibration Achieving 153.8-dB FoM in 4nm CMOS" – University of Southern California and MediaTek (Paper C24.2)

Researchers from the University of Southern California and Mediatek jointly propose a direct-RF sampling time-domain ADC that achieves 10-bit conversion at 16 GS/s using only 4X time-interleaving in 4-nm CMOS. The architecture features a redundancy-based time-to-digital converter (TDC) delay variability compensation scheme, a background delay offset calibration scheme, and a bottom-plate sampling voltage-to-time converter (VTC) structure to achieve high SNR and linearity. Silicon achieves a 55.93-dB SFDR and 44.48-dB SNDR at Nyquist while consuming 94.2 mW and occupying 8000 μm^2 of active area, leading to a state-of-the-art Schreier FoM of 153.8 dB.

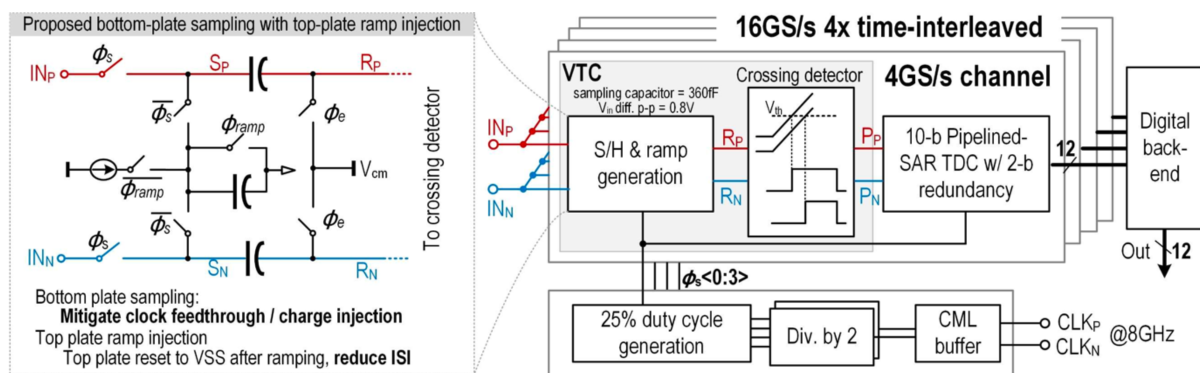
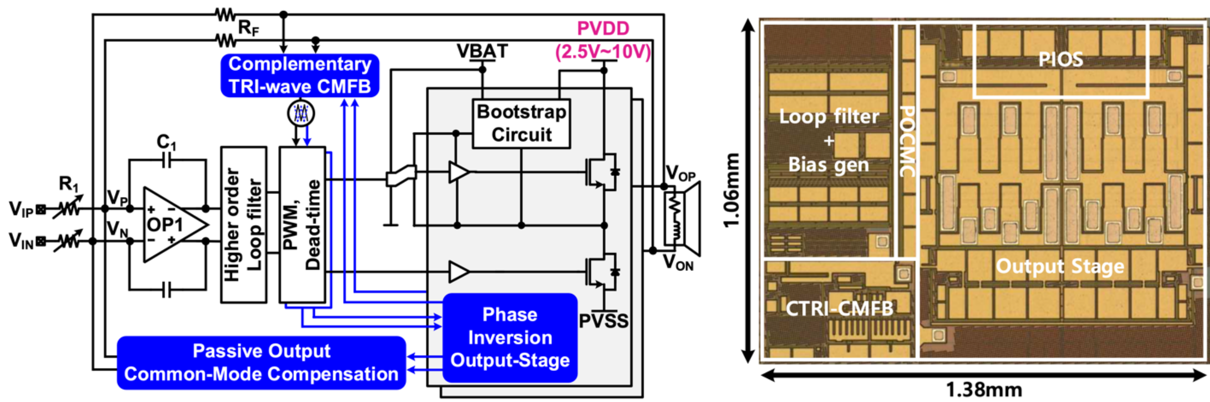


Figure: ADC Block diagram and the proposed VTC architecture.

Analog and Mixed-signal Circuits

"A 5.8-W, 0.00086% THD+N, 118-dB PSRR Class-D Audio Amplifier with Passive Output Common-Mode Compensation Technique for Wide Output Power Range" – Samsung Electronics (Paper C5.3)

Authors from Samsung Electronics propose a Class-D audio amplifier (CDA) with two key techniques: passive output common-mode compensation (POCMC) to achieve high linearity over a wide output power range and a complementary tri-wave common-mode feedback (CTRI-CMFB) to improve PSRR. Occupying 1.46mm^2 in a 0.13-μm BCD process, the CDA achieves 0.00086% THD+N, 118 dB PSRR, and a maximum output power of 5.8 W (THD+N=1%) with an efficiency of 93.2% on an 8-Ω load.



Figures: (Left) Proposed CDA schematic. (Right) Die micrograph.

Wireline and Optical Transceivers, Optical Interconnect and Processors

"A 0.296-pJ/bit 17.9-Tb/s/mm² Die-to-Die Link in 5nm/6nm FinFET on a 9-μm-pitch 3D Package Achieving 10.24 Tb/s Bandwidth at 16 Gb/s PAM-4" – TSMC (Paper C14.1)

Authors at TSMC presents a die-to-die link for heterogeneous integration of a 5-nm compute die and 6-nm SRAM die with face-to-back 3D stacking at a 9-μm bond pitch. The work demonstrates modular design that supports full scalability and achieves 10.24-Tb/s aggregate bandwidth for 320 TX lanes and 320 RX lanes, at a PAM-4 16Gb/s per lane data rate. Each data cluster module comprises 80 TX/RX lanes in a 378 μm × 378 μm footprint to offer a bandwidth density of 17.9 Tb/s/mm^2 and an energy efficiency of 0.296 pJ/bit per link.

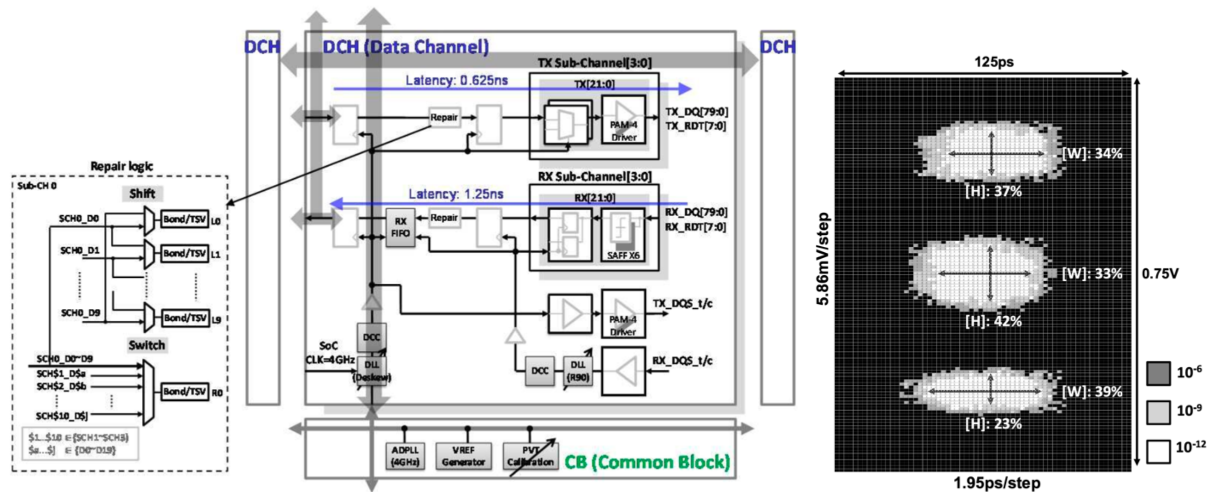


Figure: (Left) Die-to-die interconnect architecture featuring repair logic. (Right) Measured PAM-4 eye at 16 Gb/s.

Wireline and Optical Transceivers, Optical Interconnect and Processors

"A 4x50-Gb/s NRZ 1.5-pJ/b Co-Packaged and Fiber-Terminated 4-Channel Optical RX" – Intel Corporation (Paper C14.4)

Researchers at Intel present a 4-channel co-packaged optical receiver (RX) aimed at big data applications. The RX integrates a photodiode array, fiber termination, and a

transimpedance amplifier front end (TIA-FE) IC on the same package as an RX data-path IC. To achieve high sensitivity, the TIA-FE employs bandwidth extension and in-band group delay compensation techniques that are co-optimized with a quarter-rate 2-tap feed-forward equalizer in the RX data-path. The front end also features a StrongArm latch that improves noise variance by 3.5X for iso-power. The optical RX is modulated by its VCSEL-based optical transmitter counterpart and demonstrates 4×50 Gb/s NRZ at 1.5 pJ/b with BER below 10^{-12} and sensitivity of –6 dBm.

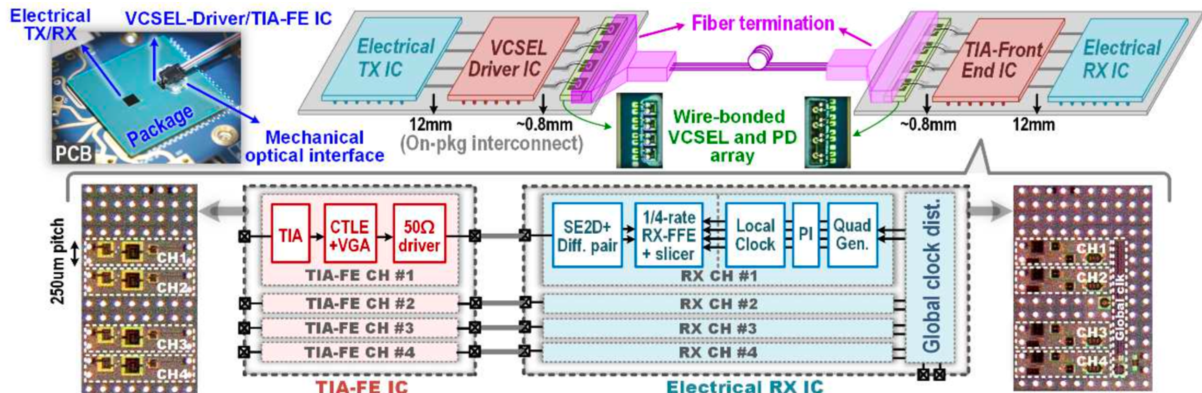


Figure: Co-packaged and fiber-terminated 4-channel optical transceiver system integration with simplified block diagrams of RX and TIA-FE ICs.

Wireless and RF Devices Circuits and Systems

“A 640-Gb/s 4×4-MIMO D-Band CMOS Transceiver Chipset” – Tokyo Institute of Technology (Paper C9.2)

Tokyo Institute of Technology presents a CMOS transceiver (TRX) chipset covering a 56-GHz signal-chain bandwidth for D-band (114–170 GHz) mm-Wave application. The work proposes an 8-way low-Q power-combined power amplifier, a 2-way low-Q power-combined low-noise amplifier, wideband-impedance-transformation mixers, and common-source-based cascaded distributed amplifiers to improve bandwidth and linearity. The TRX chipset silicon achieves 200-Gb/s data rate (32-QAM mode) in a single-input single-output (SISO) over-the-air measurement and 120-Gb/s data rate (16-QAM mode) over a 15-m distance. This work also demonstrates 640-Gb/s 4×4 multi-input multi-output (MIMO) operation.

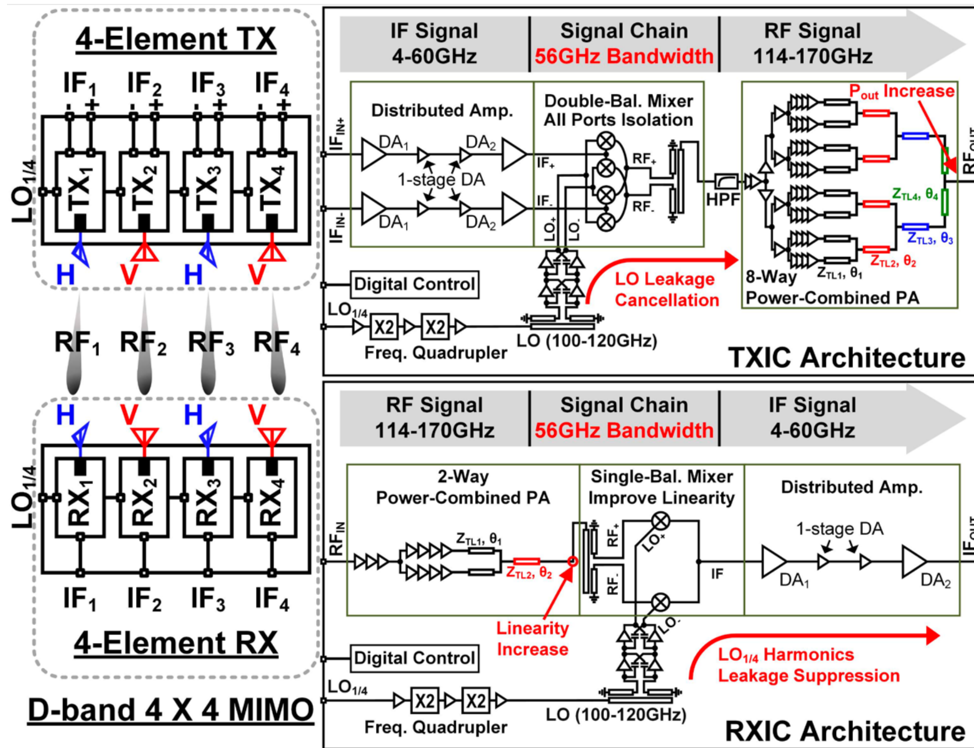


Figure: MIMO and transceiver chipset block diagram.

###