Technical Highlights from the 2021 Symposia on VLSI Technology & Circuits

The 2021 Symposia on VLSI Technology & Circuits is a premiere international conference that records the pace, progress, and evolution of micro/nano integrated electronics, scheduled from June 13-19, 2021. The two Symposia will be held using a full virtual format with a combination of live paper sessions as well as pre-recorded material.

The Symposia's overall theme, "VLSI systems for Lifestyle Transformation," 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 smart, connected devices and systems that change the way humans interact with each other.

Following are some of the highlighted papers that address this theme:

Technology Highlights

Advanced CMOS Technologies

Paper (T2-1) T0030

Researchers at imec report on forksheet N- and PFETs co-integrated with gate-all-around nanosheet FETs. The forksheet short-channel control is on par with nanosheets down to 22nm gate length (Saturated sub-threshold slope=66-68mV/dec). Forksheet I_{ON} and I_{OFF} characteristics are improved by post-channel-release wet clean optimization, attributed to gate stack interface trap density reduction. Dual work function metal gates are integrated at 17nm N-P space, highlighting a key benefit of forksheets for CMOS area scaling.



"Forksheet FETs for Advanced CMOS Scaling: Forksheet-Nanosheet Co-Integration and Dual Work Function Metal Gates at 17nm N-P Space", H. Mertens et al., imec

Paper (T2-3) T0035

Intel reports that 2D-material channels enable ultimate scaling of MOSFET transistors and will help Moore's Law Scaling for decades. They demonstrate the state of both n and p-MOSFETs using monolayer TMD (Transition Metal Dichalcogenide) channels of sub-1nm thickness and manufacturable CVD, MBE or seeded growth. NMOS devices on transferred MBE MoS₂ using novel contact metal show low variation, one of the lowest reported contact resistances (Rc) of 0.4 k Ω ·µm, low hysteresis, and good subthreshold slope (SS) of 77 mV/dec. PMOS devices using CVD WSe₂ show 89 mV/dec SS, best reported for PMOS on grown films, but on-current remains behind NMOS. Transfer-free, area-selective WS₂ transistors achieve 10 µA/µm on-current, highest reported on WS₂ using seeded growth.



"Advancing Monolayer 2D NMOS and PMOS Transistor Integration From Growth to van der Waals Interface Engineering for Ultimate CMOS Scaling", C. J. Dorow et al., Intel Corporation

Paper (T15-2) T0039

Research team in National Taiwan University reports that the 8-stacked Ge_{0.75}Si_{0.25} nanosheets and the 7-stacked Ge_{0.95}Si_{0.05} nanowires are realized by H₂O₂ wet etching. High inter-channel uniformity of the 8-stacked Ge_{0.75}Si_{0.25} is demonstrated. Thanks to small transport effective mass (mt) and large DOS (Density of States) effective mass (m_{DOS}) in L4 valley, and low R_{S/D}/R_{total}, high performance of the 7-stacked Ge_{0.95}Si_{0.05} is demonstrated. The record I_{ON}=110µA per stack (4100µA/µm per channel footprint) at V_{OV}=V_{DS}=0.5V and high G_{m,max}=340µS (13000µS/µm) at VDS=0.5V are achieved among reported Ge/GeSi 3D nFETs.



"First Highly Stacked Ge_{0.95}Si_{0.05} nGAAFETs with Record $I_{ON} = 110 \ \mu A$ (4100 $\mu A/\mu m$) at $V_{OV}=V_{DS}=0.5V$ and High Gm,max = 340 μS (13000 $\mu S/\mu m$) at $V_{DS}=0.5V$ by Wet Etching", Yi-Chun Liu et al., National Taiwan University

Paper (T5-2) T0107

Researchers from IBM Research and Applied Materials Inc. demonstrate the scalability of the dual damascene (DD) integration scheme for BEOL interconnects below 28 nm pitch. They developed two novel process flows of (1) extend Cu-based damascene interconnect with a selectively deposited TaN barrier that reduced via resistance without compromising reliability (2) an innovative dual metallization scheme with metallurgy chosen for the enhanced performance of fine and wide lines. These process innovations enable a significant improvement in via, signal and power line resistances. For highperformance computing (HPC) applications, dual metallurgy with high aspect ratio power rails provides the best performance.



Fig's. (a) TEM image of 10nm wide Co interconnects (b) RC comparison between Co (10 Å / 8.5 Å TaN) and Cu with 4.5 nm TaN/Co.

Quantum Computing

Paper (JFS5-5) T0150

AIST proposes a buried nanomagnet (BNM) realizing highspeed/low-variability silicon spin qubit operation, inspired by buried wiring technology, for the first time. High-speed quantum-gate operation results from large slanting magnetic-field generated by the BNM disposed quite close to a spin qubit, and low-variation of fidelity thanks to the selfaligned fabrication process. Employing TCAD-based simulation, they demonstrate that the BNM realizes 10 times faster Rabi oscillation (faster spin-flip) than previous works and >99% fidelity under certain process variations. Also, the proposed BNM arrangement is implementable for error-correctable large-scale quantum computers employing a 2Dlatticed qubit layout. This technology paves the way to practical large-scale quantum computers with silicon.



"Buried nanomagnet realizing high-speed/low-variability silicon spin qubits: implementable in errorcorrectable large-scale quantum computers", S. Iizuka et al., AIST

Devices for Artificial Intelligence / Machine Learning

Paper (T13-3) T0077

IBM Research reports on ARES – a 14nm Phase Change Memory (PCM) based Test-chip comprising multiple crossbar tiles, each capable of parallel Multiply-ACcumulate (MAC) inference on 512x512 unique weights. A massively-parallel 2D mesh transports Deep Neural Network (DNN) excitations in duration-format across the chip, between tiles and integrated Landing Pads (LPs) where digital data enters and leaves the chip. For accurate weight-programming (<3% weight-error), they employ a row-wise programming scheme that efficiently programs the 4 PCM devices in each analog weight with minimal overshoot. They implement two DNNs at near-software-equivalent accuracy, demonstrating tile-to-tile transport with a fully-on-chip 2-layer network, and testing resilience to error propagation with a recurrent LSTM (Long-Short-Term-Memory)

network, using off-chip activation functions before looping back to the next on-chip MAC.



"Fully on-chip MAC at 14nm enabled by accurate row-wise programming of PCM-based weights and parallel vector-transport in duration-format", P. Narayanan et al., IBM

<u>3D Heterogeneous Integration, Non-Silicon Substrates/materials and Devices</u>

Paper (T2-4) T0011

Researchers in Purdue University report the first demonstration of In_2O_3 3D transistors coated on fin-structures and integrated circuits by a back-end-of-line (BEOL) compatible atomic layer deposition (ALD) process. High performance planar In_2O_3 transistors with high mobility of 113 cm²/V s and record high maximum drain current of 2.5 mA/µm are achieved by channel thickness engineering and post-deposition annealing. High-performance ALD In_2O_3 based zero-VGS-load inverter is demonstrated with maximum voltage gain of 38 V/V and minimum supply voltage (V_{DD}) down to 0.5 V. ALD In_2O_3 3D Fin transistors are also demonstrated, benefiting from the conformal deposition capability of ALD. These results suggest ALD oxide semiconductors and devices have unique advantages and are promising toward BEOL-compatible monolithic 3D integrated circuits.



"First Demonstration of Atomic-Layer-Deposited BEOL-Compatible In_2O_3 3D Fin Transistors and Integrated Circuits: High Mobility of 113 cm²/V · s, Maximum Drain Current of 2.5 mA/µm and Maximum Voltage Gain of 38 V/V in In_2O_3 Inverter", Mengwei Si et al, Purdue University

Paper (T2-5) T0142

A research team at Sony presents the world's first demonstration of a 40nm embedded STT-MRAM (Spin-Torque-Tranfer Magnetic Random Access Memory) for buffer memory, which is compatible with the 3D stacked CMOS image sensor (CIS) process. They optimized a CoFeB-based perpendicular magnetic tunnel junction (p-MTJ) to suppress the degradation of magnetic properties caused by the 3D stacked wafer process. With improved processes, they achieved high speed write operation below 40 ns under typical operation voltage conditions at -30 °C, endurance up to 1E+10 cycles at 105 °C and 1 s data retention at 85 °C required for a buffer memory. In addition, to broaden the application of embedded MRAM (eMRAM), they proposed a novel fusion technology that integrated embedded non-volatile memory (eNVM) and buffer memory type embedded MRAM in the same chip. They achieved a data retention of 1 s ~ >10 years with a sufficient write margin using the fusion technology.



"3D stacked CIS compatible 40nm embedded STT-MRAM for buffer memory", M. Oka et al., Sony Semiconductor

Paper (JFS2-6) T0046

Researchers at imec report on scaled Si-channel finFETs (L_{gate}>20nm, 45nm fin pitch) with backside connectivity enabled by: extreme wafer thinning (several Si thicknesses under STI-oxide targeted: from ~370nm down to ~20nm) and W-filled nano-through-Si-vias (n-TSV) of various heights (linked to wafer thinning values), after using low-temperature (LT), wafer-to-wafer (W2W) dielectric bonding. This scheme aims at allowing decoupling signal and power networks, with reduced IRdrop also predicted by moving the latter to the wafer's backside. A thorough evaluation of the impact of 3D processing on device characteristics is presented, showing: 1) enhanced nmos mobility and drive currents (up to 15%); 2) for pmos, small I_{ON} loss (~3 to 10%), larger R_{ext}, with channel strain evaluation by NBD (Nano-Beam-Diffraction) for various layouts; 3) ΔV_T ~-130mV that can be recovered with an extra anneal at the end, keeping tight variability and matching control. No BTI degradation is observed, with further indication that the final anneal(s) selection can be beneficial for electrostatics and reliability improvement.



"Enabling Logic with Backside Connectivity via n-TSVs and its Potential as a Scaling Booster", A. Veloso et al., imec

Memory Technology

Paper (T11-3) T0071

TSRI reports that CMOS compatible 400°C-robust 42 nm perpendicular spin-orbit torque magnetic tunnel junction (p-SOT-MTJ) devices with the tunnel magnetoresistance (TMR) ratio of 130% is demonstrated for the first time by the interface-enhanced synthetic anti-ferromagnet (SAF) and the improved ion-beam etching (IBE). The record high 440°C thermal robustness of SAF is achieved. The SAF field (HSAF) and the magnetic coupling between Co/Pt multilayer (ML) and reference layer are enhanced by the magnet-coupling fcc-texture multilayer (MCFTM) buffer. The Pt-Fe inter-diffusion during thermal stress is effectively reduced by the W(3Å)-based texture-decoupling diffusion multi-barrier (TDDMB) for magnetic field immunity. The composite SOT channel of TaN/W and Ta/W breaks the thickness limitation of β -W (< ~5 nm) and enlarges the MTJ etching window. The TaN/W channel exhibits the large effective spin Hall angle of ~ -0.27. The deterministic field-free SOT writing with spin-transfer torque (STT) assist is achieved.



"First Demonstration of Interface-Enhanced SAF Enabling 400oC-Robust 42 nm p-SOT-MTJ Cells with STT-Assisted Field-Free Switching and Composite Channels", Ya-Jui Tsou et al., TSRI

Paper (T6-3) T0052

TSMC investigates that the polarization fatigue of HfZrO ferroelectric with SILC (stressinduced-leakage-current) measurement under different E-field stresses. Under high-field, they observed strong correlation between polarization wake-up and SILC increase. This is attributed to oxygen vacancy redistribution and percolation path formation, especially at high frequency cycling. In contrast, polarization fatigue at low field is found to occur without SILC increase. P-E loop measurements and material analysis by TEM/PED (precession electron diffraction) revealed that charge trapping is the main contributor under the low-bias. They demonstrated that the fatigue caused by low-field stress could be effectively recovered through an interspersed periodical, short-term cycles at highfield to manage charge trapping and oxygen vacancy redistribution, thus resulting in prolonged endurance to >1E12 cycles without SILC degradation at room temperature.



"Characterization of Fatigue and Its Recovery Behavior in Ferroelectric HfZrO", P.J.Liao et al., TSMC

Circuit Highlights

5G Transceivers

Paper C11-1: 28GHz Battery-Less Relay Transceiver for 5G

Millimeter-wave used in 5G has strong directivity and shorter communication distance, and it limits service area of 5G network. Researchers from Tokyo Institute of Technology propose a massive relay network for maximizing spatial coverage and capacity for 5G and beyond 5G systems. In this work, a wirelessly-powered 28-GHz phased-array relay transceiver is presented, and demonstrates a battery-less wireless communication using a 400-MHz 64QAM modulated signal based on the 5G standard without any wired power supply.

"A 28-GHz Phased-Array Relay Transceiver for 5G Network Using Vector-Summing Backscatter with 24-GHz Wireless Power and LO Transfer," Michihiro Ide, et al., Tokyo Institute of Technology



[Fig. 1& 5] 28GHz Phased-Array Relay Transceiver Working with 24-GHz Wireless Power Transfer

Photonics interconnect and compute

Paper JFS3-4: Silicon Photonic Micro-Ring Modulator-based 4 x 112 Gb/s WDM Transmitter

Authors from Intel present a hybrid-integrated $4\lambda \times 112$ Gb/s/ λ wavelength division multiplexing (WDM) transmitter for 400G Ethernet modules and co-packaged optics. The photonic IC (PIC) contains an array of micro-ring modulators (MRMs) with integrated heaters for efficient WDM. The 28nm CMOS electronic IC (EIC) has PAM4 MRM drivers with nonlinear FFE and control circuits to stabilize MRM performance against process and temperature variations. The authors claim that this is the highest per- λ data rate reported for an O-band ring-based WDM transmitter.

"Silicon Photonic Micro-Ring Modulator-based 4 x 112 Gb/s O-band WDM

Transmitter with Ring Photocurrent-based Thermal Control in 28nm CMOS," Jahnavi Sharma, et al., Intel Corporation



[Fig. 1(a)] Architecture and chip photo of the O-band WDM Si-Ph TX.

Artificial Intelligence, Machine Learning

Paper CFS1-2: Deep neural network accelerator for the edge exploiting RRAM in a multi-die scalable processor

The authors from Stanford University and TSMC present their edge AI accelerator named CHIMERA. This accelerator targets inference and incremental learning. It integrates a type of non-volatile resistive memory (RRAM) and exploits its properties in a multi-die scalable approach with chip-to-chip links. The authors scale inference to 6x larger DNNs by connecting six chips (C0 to C5). Once each chip finished its part of the computation, it is powered down. Due to the non-volatile nature of the weight memory, wakeup is quick at 33µs, allowing full power down when not in use. The authors also address RRAM wear and the high write energy and latency by proposing a training algorithm that minimizes weight updates. The chip is manufactured in a 40nm process achieving 0.92TOPS per chip and energy efficiency of 2.2TOPS/W.





[Fig. 14] Overview of the system PCB with die-shot

Paper CFS1-3: High performance energy efficient deep reinforcement learning (DRL) processor for accelerated training

Researchers from KAIST present a high performance, energy efficient deep reinforcement learning (DRL) processor named OmniDRL for accelerating training tasks. OmniDRL features multiple (24) group-sparse training cores (GSTC) that utilize pruning and block-circulant based weight grouping resulting in 2X improvement in training speed. Further, the DRL processor incorporates exponent mean delta encoding (EMDE) in compressing network interface (CNI) to improve exponent compress ration (by 1.6X) and reduce memory access power (by 23.3%). Finally, a sparse weight transposer (SWT) is integrated to enable on-chip transpose of compressed weight for reducing external memory access. This processor fabricated in 28nm CMOS, achieves peak performance of 4.18TFLOPS and peak energy efficiency of 29.3TFLOPS/W.

"OmniDRL: A 29.3 TFLOPS/W Deep Reinforcement Learning Processor with Dual-mode Weight Compression and On-chip Sparse Weight Transposer," Juhyoung Lee, et al., KAIST



[Fig. 2] Overview of the DRL processor architecture with 24 group sparse training cores (GSTC), compressing network interface (CNI) and sparse weight transposer (SWT) modules

In-memory Computing

Paper JFS2-5: A CMOS and PCM-based In-Memory Compute Core using an array of Linearized CCO-based ADCs and local digital processing

IBM presents a 256×256 in-memory compute core designed and fabricated in 14nm CMOS with backend-integrated multi-level phase-change memory (PCM). It comprises 256 linearized current controlled oscillator (CCO)-based ADCs at a compact 4µm pitch and a local digital processing unit performing affine scaling and ReLU operations. A novel frequency-linearization technique for CCOs is introduced, leading to accurate on-chip matrix-vector-multiply (MVM) when operating over 1 GHz. Measured classification accuracies on MNIST and CIFAR-10 datasets are presented when two cores are employed for deep learning (DL) inference. The measured energy efficiency is 10.5 TOPS/W at a performance density of 1.59 TOPS/mm².

"HERMES Core - A 14nm CMOS and PCM-based In-Memory Compute Core using an array of 300ps/LSB Linearized CCO-based ADCs and local digital processing," R. Khaddam-Aljameh, et al., IBM Research Europe & IBM Research, Albany & IBM Systems and Technology & IBM T. J. Watson Research Center



[Fig. 1] Die photo and system overview

Memory Technology

Paper C16-1: 16Kb Antifuse One-Time-Programmable Memory in 5nm High-K Metal-Gate FinFET CMOS

TSMC presents a 16Kb one-time-programmable (OTP) antifuse memory fabricated in a 5nm high-K, metal-gate FinFET CMOS for the first time. The bootstrap high voltage scheme (BHVS), read endpoint detection (REPD) and pseudo-differential sensing (PDS) are implemented to achieve intrinsic bit error rate (BER) below 1ppb for in-field programming in 5nm SoC and 10 years of data retention at 125°C.

"A 16Kb Antifuse One-Time-Programmable Memory in 5nm High-K Metal-Gate FinFET CMOS Featuring Bootstrap High Voltage Scheme, Read Endpoint Detection and Pseudo-Differential Sensing," Shaun Chou, et al., TSMC



[Fig. 14] Die photo.

Advanced Heterogeneous/3D Integration

Paper JFS1-3: High-density and low-power die-to-die communication for advanced heterogeneous integrations in 7nm CMOS technology:

This work presents a high-density low bit error rate and low-power Mlink (MediaTek link) PHY for ultra-short-reach (USR) die-to-die communication. Proposed Mlink have been fabricated in TSMC 7nm FinFET 1P15M CMOS technology. Interconnection is demonstrated through TSMC Chip-on-Wafer-on-Substrate (CoWoS) and TSMC Integrated Fan-Out (InFO) packaging technology. Mlink PHY exploits energy-efficient and high performance scheme, includes single-ended without termination, quarter rate strobe and unbalance scheme on transceiver, minimum intrinsic auto-alignment and novel noise-immunity coding methodology. Achieving 20Gb/s/wire and 0.46pJ/bit under 1-mm ultra-short-reach platform target to BER 1E-25. Bandwidth density is normalized with shoreline 5.31Tb/s/mm and area 2.25Tb/s/mm^2 respectively.

"A 7nm 0.46pJ/bit 20Gbps with BER 1E-25 Die-to-Die Link Using Minimum Intrinsic Auto Alignment and Noise-Immunity Encode," Ying-Yu Hsu, et al., MediaTek Inc



[Fig. 1] Mlink PHY architecture of unidirectional one pair Tx/Rx-Slice.

Power Conversion Circuits & Clock Circuits

Paper C3-1: Package-Integrated GaN Power Module for 5 V to 1 V DC – DC Converter for Server Applications

Intel propose a high-frequency, high-density (9 A/mm²) buck converter featuring a low-voltage GaN power transistor (with 5-10x better FoM than Si) integrated with a CMOS driver IC in a 4mm x 4mm package. The converter achieves 94.2 % peak efficiency in 5 V-to-1 V voltage conversion at 3 MHz switching frequency.

"A 32A 5V-Input, 94.2% Peak Efficiency High-Frequency Power Converter Module Featuring Package-Integrated Low-Voltage GaN NMOS Power Transistors," Nachiket Desai, et al., Intel Corporation



[Fig. 4.2] Die package-integrated GaN power module.

Paper C18-1: A 19-GHz PLL with 20.3-fs jitter

Researchers from UCLA will demonstrate an ultra-low-jitter PLL fabricated in 28-nm CMOS technology. A double-edge sampling technique is employed with a new retiming method, and the prototype achieves an rms jitter of 20.3fs with a 250-MHz reference clock and 12-mW power consumption.

"A 19-GHz PLL with 20.3-fs Jitter," Yu Zhao and Behzad Razavi, UCLA



[Fig. 5] Die Micrograph of Ultra-Low-Jitter PLL

Sensor and Data Converters

Paper C15-1: Fully-dynamic pipelined SAR ADC with background calibration

Fully-dynamic circuits, such as ring amplifiers, dynamic quantizers and regulators helped recent advance of ADCs in power efficiency. This year, Imec will present a singlechannel fully-dynamic pipelined SAR ADC which exploits a ring amplifier to simultaneously achieve high bandwidth and power efficiency by utilizing background calibration to ensure robustness and optimize performance. A novel dynamic quantizer and narrowband dither injection are proposed to achieve fast and comprehensive background calibration of DAC mismatch, interstage gain, and ring amplifier bias optimality. The ADC also includes an on-chip wide-range, fully-dynamic reference regulation system. Consuming 3.3 mW at 500 MS/s, the ADC achieves 10.0ENOB and 75.5 dB SFDR, yielding a Walden FoM of 6.2 fJ/c.s.

"A 10.0 ENOB, 6.2 fJ/conv.-step, 500 MS/s Ringamp-Based Pipelined-SAR ADC with Background Calibration and Dynamic Reference Regulation in 16nm CMOS," J. Lagos, et al., imec



[Fig. 1] ADC Architecture

Paper JFS4-4: CMOS image sensor with pixel-level ADC with Cu-to-Cu integration The paper presents an advanced CMOS image sensor realized with integrated 65 nm sensor and 28 nm logic chips integrated using Cu-to-Cu (C2C) interconnectors. It reveals low-random noise of 2.6 e-rms, a low-power of 116.2 mW at video rate, and a high-speed up to 960 fps 2-megapixels global-shutter type CMOS image sensor (CIS) using an advanced DRAM technology. To achieve a high performance global shutter CIS, they propose a novel architecture for the digital pixel sensor with remarkable global-shutter operation CIS employing pixel-wise ADCs and digital memory. Each pixel has two small-pitch Cu-to-Cu interconnectors for the wafer-level stacking, and the pitch of each unit pixel is less than 5 μ m which is the world's smallest pixel embedding both pixellevel ADC and 22-bit memories.

"A 2.6 e-rms Low-Random-Noise, 116.2 mW Low-Power 2-Mp Global Shutter

CMOS Image Sensor with Pixel-Level ADC and In-Pixel Memory," Min-Woong Seo, et al., Samsung Electronics



[Fig. 5] Chip micrographs (left: top chip, right: bottom chip)

Paper C19-1: Wide-range gyroscope frontend circuit

Robert Bosch & Bosch Sensortec present a gyroscope frontend circuit that is used to sense, amplify and digitize signals from 3-axes capacitance-sensing type gyroscope sensor device. The significant feature of the work is its wide input range as large as +/-8000 deg./s, 4 times larger than state of art. The chip shows comparable performance in other aspects, such as noise floor of 0.0047 deg./s/ \sqrt{Hz} .

"A direct-digitization open-loop gyroscope frontend with +/-8000°/s full-scale range and noise floor of 0.0047° /s/ \sqrt{Hz} ," Chinwuba Ezekwe, et al., Robert Bosch & Bosch Sensortec



[Fig.5] Chip Micrograph (1.94 mm × 1.8 mm)

Biomedical Circuits

Paper C2-2: Neural recording chip for wireless ultra-small brain sensor

The paper presents a power-management, neural sensing and optical signal transmission chip for NIR-operated ultra-small brain sensing device. The main feature of the paper is improved light tolerance of the chip. The authors implemented a circuit design that can keep its functionality even if the CMOS chip is exposed by unexpected light coming through package. The authors achieved sufficient light tolerance up to 300 μ W/mm² light exposure (above tissue limit) and consumes 0.57 μ W at 38°C, making it lowest power among standalone motes while incorporating on-chip feature extraction and individual gain control.

"A Light Tolerant Neural Recording IC for Near-Infrared-Powered Free Floating Motes," Jongyup Lim, et al., University of Michigan & ETH Zürich & University of Delaware



[Fig. 1] Conceptual illustration of NIR based wireless neural recording mote