Session 17 Overview: Embedded Memory and DRAM I/O

MEMORY SUBCOMMITTEE

Session Chair: Leland Chang, IBM, Yorktown Heights, NY
Session Co-Chair: Takefumi Yoshikawa, Panasonic, Kyoto, Japan

Demand for smaller and lighter personal devices with increasing functionality in the cloud drives advancements in both embedded memory technology and high-speed DRAM interfaces. Lower power through voltage reduction and increased performance through higher memory density and bandwidth are key enablers. This year’s conference highlights 14nm FinFET technologies with the smallest bit cells achieved to date for both SRAM – at 0.05μm² – and embedded DRAM – at 0.01747μm². A new assist technique to drive $V_{\text{MIN}}$ reduction and novel bit cells for both 2-port SRAM and TCAM applications are also presented. In addition, two area-efficient techniques to boost DRAM bandwidth are reported.

17.1 A 0.6V 1.5GHz 84Mb SRAM Design in 14nm FinFET CMOS Technology 8:30 AM
E. Karl, Intel, Hillsboro, OR

In Paper 17.1, Intel presents an 84Mb embedded SRAM fabricated in 14nm FinFET technology featuring the smallest bit cells to date at 0.050μm² for high density and 0.058μm² for low voltage. 1.5GHz operation at 0.6V is demonstrated.

17.2 A 64kb 16nm Asynchronous Disturb Current Free 2-Port SRAM with PMOS Pass-Gates for FinFET Technologies 9:00 AM
H. Fujiwara, TSMC, Hsinchu, Taiwan

In Paper 17.2, TSMC presents a new 16nm 8T bit cell to address read-disturb problems in two-port applications. Taking advantage of the relative PMOS-to-NMOS transistor drive ratio in FinFET technology, traditional NMOS write transistors in the bit cell are replaced with PMOS transistors.

17.3 A 28nm 256kb 6T-SRAM with 280mV Improvement in $V_{\text{MIN}}$ Using a Dual-Split-Control Assist Scheme 9:30 AM
C-F. Chen, National Tsing Hua University, Hsinchu, Taiwan

In Paper 17.3, National Tsing Hua University presents a novel dual-split-control approach to improve the read/write margins of a 6T bit cell. A $V_{\text{MIN}}$ improvement of 280mV is demonstrated in 28nm technology. Additionally, a figure of merit is proposed to quantify read/write margin improvements.
17.4 A 14nm 1.1Mb Embedded DRAM Macro with 1ns Access
G. Fredeman, IBM, Poughkeepsie, NY

In Paper 17.4, IBM presents a 1.1Mb embedded DRAM in 14nm FinFET technology for high-performance microprocessor caches. The 0.01747μm² cell features a deep-trench capacitor and a thick-oxide FinFET access device. An array access time of 1.0ns is demonstrated.

17.5 A 3T1R Nonvolatile TCAM Using MLC ReRAM with Sub-1ns Search Time
C-C. Lin, National Tsing Hua University, Hsinchu, Taiwan

In Paper 17.5, National Tsing Hua University presents a dense ternary content-addressable-memory (TCAM) realized using a multi-level, non-volatile ReRAM cell. A search time of <1ns and energy/bit/search of 0.51fJ are achieved.

17.6 1V 10Gb/s/pin Single-Ended Transceiver with Controllable Active-Inductor-Based Driver and Adaptively Calibrated Cascade-DFE for Post-LPDDR4 Interfaces
J. Song, Korea University, Seoul, Korea

In Paper 17.6, Korea University presents an area-efficient transmitter and receiver for mobile DRAM interfaces beyond LPDDR4. Using a combined RX/TX active area of 0.01mm², 10.5Gb/s/pin single-ended data transmission has been achieved through a 10cm FR4 PCB trace.

17.7 A Digital DLL with Hybrid DCC Using 2-Step Duty Error Extraction and 180° Phase Aligner for 2.67Gb/s/pin 16Gb 4-H Stack DDR4 SDRAM with TSVs
W-J. Yun, Samsung Electronics, Hwasong, Korea

In Paper 17.7, Samsung presents a 25nm 16Gb DDR4 SDRAM with a compact DLL architecture and hybrid DCC to address power noise in a 4-H TSV stack. Stable operation of 2.67Gb/s/pin at 1.2V is demonstrated in an active area of 0.076mm².
17.1 A 0.6V 1.5GHz 84Mb SRAM Design in 14nm FinFET CMOS Technology

Eric Karl, Zheng Guo, James W. Conary, Jeffrey L. Miller, Yong-Gee Nq, Satyanand Nalam, Daeyeon Kim, John Keane, Uddalak Bhattacharya, Kevin Zhang
Intel, Hillsboro, OR

The growth of battery-powered mobile and wearable devices has increased the importance of low-power operation and cost in system-on-a-chip (SoC) design. Supply-voltage scaling is the predominant approach to active power reduction for SoC design, including voltage scaling for on-die memory given increasing levels of memory integration. SRAM can limit the minimum operating voltage ($V_{\text{MIN}}$) of a design, often leading to the introduction of separate voltage supplies for on-die memory. Additional supplies increase platform cost, and operating memory at higher voltages leads to increased power consumption. The introduction of trigate devices at the 22nm technology node delivered superior short channel effects and subthreshold slope relative to existing bulk planar device technology enabling reduction in threshold voltage within a fixed leakage constraint. Lower transistor $V_{\text{th}}$, improvements to random device variability, and device technology enabling reduction in threshold voltage within a fixed leakage short channel effects and subthreshold slope relative to existing bulk planar memory at higher voltage leads to increased power consumption. The for on-die memory. Additional supplies increase platform cost, and operating

Figure 17.1.1 shows a layout diagram of a 0.0500μm$^2$ high-density 6T SRAM cell (HDC) and a 0.0588μm$^2$ low voltage 6T SRAM cell (LVC) in 14nm FinFET technology [2]. HDC has a fin ratio of 1:1:1 (PU:PG:PD) and LVC features a larger PD device at 1:1:2, to provide improved read stability and performance at low voltages. Analog and digital double-patterning techniques have extended the capabilities of 193nm immersion lithography on critical layers to deliver 0.54× scaling relative to comparable 22nm SRAM bitcells. Contacted gate pitch is 70nm and fin pitch is 42nm for the technology. Figure 17.1.1 highlights recently reported SRAM designs from 16nm, 14nm and 10nm technologies [1-6]. The 0.050μm$^2$ HDC cell in this work is the smallest reported SRAM cell at any technology node among the cited work. Despite significant geometric scaling from the 22nm node, optimizations to fin profile and subfin doping on the 2nd-generation FinFETs enable a nearly 2× reduction in device random threshold voltage variation [2], a critical factor for 6T SRAM $V_{\text{MIN}}$.

The HDC SRAM cell has a 1:1 PG to PD ratio, leading to degraded stability from charge injection during read operations. Wordline underdrive (WLUD) is utilized in the HDC arrays as an area-efficient approach to enhance read-stability margin at the cost of cell performance [1]. Suppressed BL techniques, such as the DNR circuit, also improve read stability, but are effective across a limited range of process technology targets and the implementation leads to more area overhead and higher power consumption than WLUD [3]. The LVC SRAM, with a 1:2 PG to PD ratio, has adequate read-stability margin and does not require WLUD, enabling higher performance at low voltage. Both HDC and LVC have a 1:1 ratio between the PG and PU devices, which limits write margin at low voltage due to contention between these two devices, motivating the use of a write-assist circuit. In Figure 17.1.2, a column-based TVC circuit is shown, utilizing an NMOS device to discharge the memory cell supply (VCS) to weaken the PU transistor during writes [1]. Half-selected cells along the written column face an instability risk if VCS is pulled below the minimum retention voltage. This can be mitigated by controlling the amplitude of the supply-voltage collapse in this circuit by enabling devices connected to VCS that clamp VCS to an intermediate level. The downside in this implementation is static current flow between supplies when clamping VCS. In this 14nm FinFET technology, improvements to random variation reduce the amplitude of VCS collapse required to achieve low-$V_{\text{MIN}}$ operation. The conventional TVC circuit topology requires the highest static current to clamp VCS near VCC, limiting the realizable write-energy reduction.

Figure 17.1.3 details a charge-share transient voltage collapse circuit (CS-TVC) that delivers high performance at low voltage, compatibility with WLUD across a wide range of technology targets and reduces active power consumption relative to the TVC circuit in Fig. 17.1.2. The CS-TVC switch circuits are located at the edges of a 256-row column, breaking the memory cell supply into 2 distinct 128b regions per column. Multiple switch circuits within a column are connected to a CS-TVC capacitor to reduce area, as only one switch within the column is active during write. The CS-TVC capacitor contains a primary node and 2 secondary nodes enabled by CSCAP[1:0] to modulate the effective capacitance. NMOS devices driven by the DISCHARGE signal are used to reset the GS Channel, resulting in a temporary suppression of the VCS node to improve write margin. The falling edge of the TVC pulse completes the operation and restores the VCS voltage level to VCC.

In Figure 17.1.4 shows the organization of the 68kb LVC block and 17kb LVC subarray, featuring 258b/BL and 136b/BL in a butterfly array configuration. In the logical I/O, 4 CS-TVC switch regions share one CS-TVC capacitor bank positioned in the center of the column. The 17kB LVC sub-array with conventional TVC and CS-TVC achieve bit densities of 11.6 and 11.3 Mb/mm$^2$, respectively. The array efficiency is 71.6% and 69.8%, respectively for conventional TVC and CS-TVC, with the CS-TVC circuitry adding 6.5% to the 17kB sub-array area. The HDC array features a denser 512b/BL column I/O design with a bit density of 14.5 Mb/mm$^2$ and 76.2% array efficiency.

In Figure 17.1.5, 50MHz and 1GHz LVC SRAM write-$V_{\text{MIN}}$ measurements are shown for different TVC methods. CS-TVC features a charge-share collapse to 48% of VCC and conventional TVC with a strong bias setting (SB-TVC) provides a clamped voltage level at 44% of VCC using the circuit from Fig. 17.1.2. Pulsed TVC (P-TVC) features a shallow collapse defined by a narrow pulsewidth achievable by either the CS-TVC or the conventional TVC circuit. The deeper collapse of CS-TVC and SB-TVC enable 40mV lower $V_{\text{MIN}}$ at 50MHz. At 1GHz, the deeper collapse and required recovery for these designs increases $V_{\text{MIN}}$ by 30mV for CS-TVC and 75mV for SB-TVC relative to P-TVC. By eliminating static crowbar current, the CS-TVC circuit reduces active energy by 24% relative to SB-TVC for a comparable $V_{\text{MIN}}$ and VCS collapse level. P-TVC has write-energy overhead of 43% relative to no assist, but is difficult to implement across a range of array configurations.

Figure 17.1.6 shows the 14nm LVC $V_{\text{MIN}}$ is 0.6V at the 90th percentile, an 80mV reduction compared with a comparable LVC array on 22nm [1]. The 14nm HDC $V_{\text{MIN}}$ is 0.7V at the 90th percentile, within 15 to 20mV of the 22nm LVC SRAM $V_{\text{MIN}}$. The voltage-frequency shmoo of LVC SRAM at 95°C demonstrates 1.5GHz performance at 0.6V and wide-range operation to 1V. Figure 17.1.7 shows a die micrograph of a 14nm test vehicle with HDC and LVC SRAM arrays with fuse, PLL and PBIST circuitry.

References:
Figure 17.1.1: 14nm HDC and LVC SRAM bitcells.

Figure 17.1.2: Conventional transient voltage collapse (TVC) write assist circuit.

Figure 17.1.3: Charge-share transient voltage collapse circuit (CS-TVC) and operation waveforms.

Figure 17.1.4: 548Kb SRAM architecture and density summary.

Figure 17.1.5: P-TVC, SB-TVC and CS-TVC $V_{\text{MIN}}$ and write energy comparison.

Figure 17.1.6: HDC and LVC $V_{\text{MIN}}$ and LVC voltage-frequency shmoo.
Figure 17.1.7: Die micrograph of 14nm test chip with HDC and LVC SRAM arrays.
17.2 A 64kb 16nm Asynchronous Disturb Current Free 2-Port SRAM with PMOS Pass-Gates for FinFET Technologies


TSMC, Hsinchu, Taiwan

FinFET technology has been adopted in the 16nm node because it provides superior $I_d/I_{off}$ ratio, short-channel effect and local variation [1.2]. 2P-SRAM, which offers simultaneous read and write operations, is widely used for media processing because of its high operating efficiency. However, 2P-SRAM using the conventional 2P8T cell has a read-disturb issue, when both read wordline (RWL) and write wordline (WWL) are asserted simultaneously in the same row. Furthermore, read-disturb becomes worse in FinFET technology compared with classical planar technology. In order to overcome these problems, we develop a disturb-current-free (DCF) 2P8T cell with PMOS write pass-gates and peripheral assist circuits to further improve its performance.

Figure 17.2.1 shows the conventional 2P8T cell, which is composed of two PMOS and six NMOS transistors in a cell area of 0.138μm$^2$ with 16 nm FinFET design rules. In a pure 2P-SRAM read operation, the RWL is activated while the WWL stays low. According to the storage node voltage level V(MB), the read bitline (RBL) will either be kept high or discharged. However, in a simultaneous read and write operation, the read-disturb issue occurs when both the RWL and WWL are accesses simultaneously in the same row. In this case, the storage node (MB) voltage level is raised slightly due to WWL activation, which causes a disturb current ($I_{dist}$). The RBL is discharged by the disturb current that flows through the RPD and RPG transistors, which may lead to read failure. The solid black and gray dashed lines in the lower left of Fig. 17.2.1 represent the drain current ($I_d$) of the read port with 16nm FinFET devices and 20nm planar devices, respectively. Although higher $I_{dist}$ (above 0.7V) current can be obtained in FinFET devices, $I_{dist}$ is also increased when $V_{dd}$ is slightly raised (0 to 0.3 V) because of the steep subthreshold behavior of the FinFET devices. The result is that the read-disturb issue in a 2P8T cell with FinFET devices is worse than that with conventional 20nm planar devices. As the $I_{dist}$/Id ratio degrades by about 50%. In order to avoid read failures caused by this disturb current, static keeper and reference RBL schemes have been proposed in conventional 2P-SRAM design [3.4]. However, the design difficulties of these schemes are increased in FinFET technology.

Figure 17.2.2 shows our DCF 2P8T cell. Compared with the conventional 2P8T cell, the write pass-gates (PG0 and PG1) are changed from NMOS transistors to PMOS transistors. The cell area of the DCF 2P8T cell is the same as the conventional cell (0.138μm$^2$). Because the introduction of embedded SiGe source/drain enhances hole mobility and thus the drive current of PMOS transistors, NMOS and PMOS conductance is almost the same in a 16nm FinFET process. Therefore, despite the use of PMOS transistors for PG0 and PG1, large write-margin degradation does not occur. The only drawback of the DCF 2P8T cell is that $I_{dist}$ is degraded by up to 35% due to a drop in the high node storage level when both RWL and WWL in the same row are simultaneously turned on. However, since there is no $I_{dist}$ in the DCF 2P8T cell, the static keeper can be removed from the RBL, which more than compensates for this $I_{dist}$ loss.

In the conventional 2P8T cell with static keepers, the keeper current must be larger than $I_{dist}$ in order to maintain successful “0” read operation. However, a large speed degradation occurs because the high keeper current decreases Ion during “0” read operation. On the other hand, elimination of the static keeper in the DCF 2P8T design enables faster operation and smaller macro area.

Fig. 17.2.3 compares array delay between the DCF 2P-SRAM without static keepers and a conventional 2P-SRAM with static keepers. With 16 cells on the local RBL (LRBL), the DCF 2P8T cell design without static keepers can improve global RBL (GRBL) discharge time by 23% (from 101 to 78ps), compared with the conventional 2P8T design with static keepers. Even if the number of cells on the LRBL is increased, the DCF 2P8T cell can still achieve faster operation.

Figure 17.2.4 depicts the circuit diagram of a 64kb 2P-SRAM macro. The SRAM macro includes 256 rows, 32b I/Os and each I/O contains eight columns. A hierarchical RBL scheme is applied for the RBL and the number of cells on the LRBL is 16. In order to change the polarity of the logic in the write path to accommodate the PMOS pass-gates, OR-based instead of AND-based logic is used for the WWL decoder. A boosted WBL instead of a negative WBL is used to enhance write margin [5]. In addition, to improve RBL discharge speed, a boosted RWL is adopted [6], which is not suitable for the conventional 2P8T cell because both $I_{on}$ and $I_{off}$ are increased. Instead of static keepers, which normally require long-channel devices, we adopt a delayed keeper that requires no careful device tuning to ensure the high voltage level of RBL. Because transistor KP0 in the delayed keeper is turned off during the RBL discharge period, degradation of the RBL discharge time does not occur. In addition, because long-channel transistors are not required in the delayed keeper, the SRAM macro area can be reduced by 4% as compared with a case where static keepers are used. Read and write operating waveforms of the DCF 2P-SRAM are shown in Fig. 17.2.5.

A die micrograph of the test-chip, which is fabricated in a 16nm high-k metal gate (HKMG) FinFET CMOS technology, is shown in Fig. 17.2.7. In one test-chip, there are twenty 64kb DCF 2P-SRAM macros to yield a total memory capacity of 1.25Mb. In addition, we also implement the conventional 2P-SRAM for comparison of measured data. Chip measurement results are shown in Fig. 17.2.6. We measure 1240 macros and observe a minimum operating voltage (V_MIN) for the 64Kb DCF 2P-SRAM of 0.505V with 99% yield at room temperature. Furthermore, measured access time when both RWL and WWL are asserted of the DCF 2P-SRAM (600ps) is 60ps faster than that of the conventional 2P-SRAM (660ps). In addition, the RBL discharge time does not occur. In the delayed keeper is turned off during the RBL discharge period, degradation of the RBL discharge time does not occur. In addition, because long-channel transistors are not required in the delayed keeper, the SRAM macro area can be reduced by 4% as compared with a case where static keepers are used. Read and write operating waveforms of the DCF 2P-SRAM are shown in Fig. 17.2.5.

Acknowledgment: The authors thank the physical design team for layout and chip implementation, the RD team for wafer manufacturing, and the test department for chip measurements on this work.

References:
Figure 17.2.1: Conventional 2P8T cell and read-disturb issue.

Figure 17.2.2: Disturb-current free (DCF) 2P8T cell.

Figure 17.2.3: Comparison of read bitline (RBL) delay time.

Figure 17.2.4: Circuit diagram of the 64-Kb DCF 2P-SRAM.

Figure 17.2.5: Operating waveforms.

Figure 17.2.6: Chip measurement results.
Figure 17.2.7: Micrograph of test-chip.

<table>
<thead>
<tr>
<th>Technology</th>
<th>16nm HH MG FinFET CMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bit cell area</td>
<td>0.138 µm²</td>
</tr>
<tr>
<td>Macro configuration</td>
<td>64-Kb</td>
</tr>
<tr>
<td></td>
<td>(512 rows x 32 KOs x 8 cols, 16 cells on local read bitline)</td>
</tr>
<tr>
<td>Bit capacity</td>
<td>1.25-Mb (64-Kb x 20 macros)</td>
</tr>
</tbody>
</table>
17.3 A 28nm 256kb 6T-SRAM with 280mV Improvement in $V_{\text{MIN}}$ Using a Dual-Split-Control Assist Scheme

Meng-Fan Chang1, Chien-Fu Chen1, Ting-Hao Chang1, Chi-Chang Shuai1, Yen-Yao Wang1, Hiroyuki Yamashita2

1National Tsing Hua University, Hsinchu, Taiwan, 2Fukuoka Institute of Technology, Fukuoka, Japan

Intelligent wearable devices and the Internet of things (IoT) require on-chip SRAM macros with (1) compact area to reduce costs, (2) single supply voltage (VDD) and low minimum VDD (VDDmin) to reduce power consumption; and (3) sufficient speed to facilitate real-time computing. 6T SRAM is compact, but suffers from read disturb and cell stability (static noise margin, SNM) for cross-point assist at the expense of increased area and power overhead due to the inclusion of pumping capacitors ($C_{\text{pkg}}$). Wordline (WL) voltage under-drive (WLUD, $V_{\text{WLUD}}=V_{\text{DD}}-V_{\text{cell}}$) is commonly used in 6T SRAMs [2-5] to improve HS/ Read disturb beyond read cycle speeds and necessitating an increase in $\Delta V_{\text{DD}}$ or $V_{\text{cell}}$. The maximum $\Delta V_{\text{DD}}$ is limited by the hold SNM of CVDD-HS cells. Large $C_{\text{pkg}}$ results in large area and power overhead, particularly in macros with wide I/O and small amount of column-multiplexing (Y-mux). Thus, HS-SNM tradeoffs in $C_{\text{cell}}$ and WM have not yet been solved for 6T cells, except by adding additional transistors (i.e., 8T to 10T).

To achieve goals (1) to (3) described above, this work develops a dual-split-control (DSC) scheme for 6T SRAM. We use (a) the same front-end layout as in 6T SRAM to reduce area overhead (b) VDD-supplied split WL (S-WL) to reduce HS disturbance without degrading $I_{\text{cell}}$ and (c) row-based split cell-VSS (S-CVSS) controls to improve HS-SNM and WM under S-WL. The DSC scheme improves cell-level ($I_{\text{cell}}$×$V_{\text{cell}}$)×Write-Energy($E_{\text{W}}$)×Macro-Area(s)4) figure-of-merits (FOM) by 32× and 20×, respectively, compared to previous WLUD+NBL/CVDD-D schemes. A 28nm 256Kb 6T SRAM test-chip confirms the efficacy of the DSC scheme by achieving a 280mV lower VDDmin than 6T SRAM without assist. Reasonable HS-SNM tradeoffs in $C_{\text{cell}}$ and WM have been simultaneously achieved without additional transistors.

Figure 17.3.1 shows the cell and macro structure of DSC6T SRAM. In a DSC6T cell, the pass-gate NMOS (PG1 and PG2) are separately controlled by WL1 and WL2, while the source terminals of the pull-down NMOS (PD1 and PD2) are biased independently through S-CVSS lines (CVSS1 and CVSS2). The pull-up PMOS (PUI1 and PUI2) and PDs form two pairs of inverters (INV1 and INV2). The layout of the DSC6T cell is the same as conventional 6T compact 6T cells, with the exception of minor modifications in the metal-3 layer and two additional metal lines (WL1 and WL2) in the metal-4 layer, at the expense of the removal of metal-4 CVSS meshing. Fortunately, low-VDD operation of wearable and IoT devices mitigates this tradeoff as CVSS supply bounce is likely to be low. Different from typical 6T SRAMs without assist, the DSC6T macro includes an additional CVSS-assist voltage ($V_{\text{SASS}}$) generator (CVSS-Gen), CVSS-switches (CVSS-SW), and NBL drivers with small-$C_{\text{pkg}}$. At high VDD (assist-mode is off, TAM=0), WL1=WL2 and CVSS1=CVSS2=$V_{\text{SASS}}$, while BLs/BLBs are pre-charged to VDD. The improved SNM compared to conventional 6T SRAM means that the accessed/HS 1-cells do not suffer read disturb when accessed/HS 0-cells perform single-ended read/dummy-read operations due to PG2-off and CVSS2=$V_{\text{SASS}}$. The improved SNM with $V_{\text{SASS}}$=VDD enables DSC6T cells to achieve low read-VDDmin with no degradation in $I_{\text{cell}}$, due to lower $V_{\text{cell}}$, as would otherwise be the case in WLUD schemes. The single-ended sensing at $V_{\text{cell}}$=VDD can be much faster than differential read with WLUD. Moreover, the inclusion of an application-dependent data-inverting (ADI) scheme helps to ensure that the number of 1-cells exceeds the number of 0-cells in the cell array, which reduces the power consumption of the single-ended sensing scheme to below that of differential sensing.

Figure 17.3.3 presents the read operation of DSC6T at low-VDD, in which WL1=VDD, WL2=CVSS1=0V, CVSS2=$V_{\text{SASS}}$, while BLs/BLBs are pre-charged to VDD. The improved SNM compared to conventional 6T SRAM means that the accessed/HS 1-cells do not suffer read disturb when accessed/HS 0-cells perform single-ended read/dummy-read operations due to PG2-off and CVSS2=$V_{\text{SASS}}$. The improved SNM with $V_{\text{SASS}}$=VDD enables DSC6T cells to achieve low read-VDDmin with no degradation in $I_{\text{cell}}$, due to lower $V_{\text{cell}}$, as would otherwise be the case in WLUD schemes. The single-ended sensing at $V_{\text{cell}}$=VDD can be much faster than differential read with WLUD. Moreover, the inclusion of an application-dependent data-inverting (ADI) scheme helps to ensure that the number of 1-cells exceeds the number of 0-cells in the cell array, which reduces the power consumption of the single-ended sensing scheme to below that of differential sensing.

Figure 17.3.4 presents the performance of DSC6T. At VDD=0.5V, the use of VDD for WL enables a DSC6T to attain a 3.5× reduction in BL development time for a 256-row macro and 4× lower $E_{\text{W}}$ compared to WLUD+NBL/CVDD-D schemes. The DSC6T also achieves larger WM at a low VDD. Thus, DSC6T reduces the amount of $V_{\text{MIN}}$/AVDD and $C_{\text{cell}}$ needed at fixed VDDmin, enables a 3.7× reduction in $E_{\text{W}}$ for a 256Kb SRAM with 32b IO (Y-mux=8), and results in a 52% reduction in macro area overhead compared to WLUD-based schemes. Thus, a DSC provides a 20× improvement in the macro-level FOM in 1/$E_{\text{W}}$/($E_{\text{W}}$/A). For a macro with wider 10 or smaller Y-mux, the improvement in power and area is even more pronounced. As shown by a comparison of simulation results in Fig. 17.3.5, DSC6T provides a 32× improvement in $I_{\text{cell}}$×$V_{\text{cell}}$×WM/HS-SNM$	ext{^2}$.

Figure 17.3.6 presents our measured results. This study implements a 256-row 256Kb DSC6T SRAM macro based on 0.127μm 2 6T SRAM cells fabricated in a 28nm process. The test-chip features a test-mode (TAM) to enable the disabling of S-WL/S-CVSS controls (TAM=0) in order to emulate the behavior of nominal 6T SRAM. Shmoo tests reveal that the DSC6T (TAM=1) achieves VDDmin 280mV lower than that of conventional 6T SRAM (TAM=0) when $V_{\text{SASS}}$ exceeds 140mV. As shown in the captured waveform at VDD=0.58V, the 256Kb DSC6T macro achieves read access times of 2.2ns. Figure 17.3.7 presents a die micrograph.

Acknowledgment:
The authors thank Chun-Liang Hou and Chih-Chin Lin for their help.

References:
Figure 17.3.1: Macro and cell structure of developed DSC6T.

Figure 17.3.2: Write operation of DSC6T.

Figure 17.3.3: Read operation of DSC6T.

Figure 17.3.4: Performance of DSC6T.

Figure 17.3.5: Cell performance.

Figure 17.3.6: Measured results.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>28nm process</td>
</tr>
<tr>
<td>Capacity</td>
<td>256kb</td>
</tr>
<tr>
<td>Cell Size</td>
<td>0.127um²</td>
</tr>
<tr>
<td>Sub-array</td>
<td>256 rows x 128 columns</td>
</tr>
<tr>
<td>Macro Size (including test modes, test circuits, SA options)</td>
<td>248x205 um²</td>
</tr>
<tr>
<td>Read Access Time</td>
<td>VDD=0.9V: 600ps</td>
</tr>
<tr>
<td></td>
<td>VDD=0.58V: 2.2ns</td>
</tr>
<tr>
<td>VDDmin Improvement</td>
<td>280mV</td>
</tr>
</tbody>
</table>

Figure 17.3.7: Die micrograph.
17.4 A 14nm 1.1Mb Embedded DRAM Macro with 1ns Access

Gregory Fredeman, Donald Plass, Abraham Mathews, Kenneth Reyer, Thomas Knips, Thomas Miller, Elizabeth Gerhard, Dinesh Kannambadi, Chris Paone, Dongho Lee, Daniel Rainey, Michael Sperlager, Michael Whalen, Steven Burns

IBM introduced trench capacitor eDRAM into its high performance microprocessors beginning with 45nm and Power 7 [1] to provide a higher density cache without chip crossings. Whereas the 45 and 32nm designs employ a micro sense amplifier [2] and three-level bitline hierarchy, the design implemented for 22nm utilizes a higher gain sense amplifier and two-level bitline architecture that together provide significant reductions in area, latency, and power. This 22nm design style has been migrated into a 14nm FinFET [3] learning vehicle, complete with an ABIST engine, wordline charge pumps (VPP and VVL), and padlock interface circuitry.

The eDRAM as shown in Fig. 17.4.1 is composed of two 8-way associative banks separated by a central interface block. Each bank contains 512 rows and 1152 columns and is divided into a 2×4 matrix of subarrays. As in previous generations, the wordline drivers are arranged orthogonally to the wordlines, placed along the subarray edge together with the control circuits that wire to the central sense-amplifier stripe that separates the two groups of 66-cell bitlines (Cell Blocks 0,1). The two bitline groups are multiplexed into a common sense amplifier, and eight sense amplifiers are column multiplexed into a read-data mux that drives a global read data line, which in turn dots four subarrays together. Finally, the two interleaved banks are multiplexed into a common data output bus that is 144 wide.

The subarray detail begins with the 0.01747ym² cell, featuring a deep-trench capacitor and a thick-oxide FinFET access device. The local bitlines are on M1, while the global data lines are on M4; the intermediate M2 global bitlines that previously connected micro arrays spanning 256 rows are eliminated, thereby reducing both AC power and latency. To facilitate the local bitlines growing from 34 to 66 cells, a higher-sensitivity sense amplifier, shown in Fig. 17.4.2, is developed that relies on the high voltage gain of a power-gated inverter at mid-level input voltage. This intermediate node voltage is created by the charge sharing between the local bitline (LBL), which is precharged to GND, and the sense amplifier node (SA), precharged to VBLH, that occurs when the BLMUX signal is activated. VBLH is an internally regulated supply voltage for the array core that can be tuned for performance versus power, depending on the application. For a read operation, the selected cell charge shares with the LBL and SA nodes, pulling downward for a 0 and upward for a 1. After a sufficient settling time, the inverter output (SAN) is fed back through the SETPU and SETPD gated stacks to write back full raw levels into the cell. Because the inverter output is lightly loaded, driving only the feedback and read data mux gates, a rapid 0’s access time is achieved, while the somewhat slower source-follower action in the cell when reading a 1 only has to maintain node SAN close to GND, and is not in the access path.

Figure 17.4.3 shows a write 0 operation followed by a read. To write a 0, the sense amplifier is brought out of precharge, and devices N1, N2, and P2 are turned on while device P1 is off. This drives the SA node in the sense amplifier to GND, along with the bitline whose BLMUX NFET turns on, and discharges the cell capacitor if previously a 1. In the second cycle shown, the cell is read by the sense amplifier, and after a programmable delay, devices N4, P4, and P5 are activated to drive nodes SA and SAN to their respective rails. During this time frame the read port is activated by enabling one of the 8 LS devices. If both LS and SAN are high, then the NFET stack pulls down the LDT node in the read data mux, causing the ND output device to pull down the precharged data line (RDL).

For improving both overall eDRAM performance and performance tracking with standard logic, a dynamic AND-gate wordline driver shown in Fig. 17.4.5 is developed that uses thin-oxide devices rather than thick-oxide high-voltage transistors. To achieve an expanded voltage swing for the wordline while avoiding excessive voltage between any of the three terminals for each of these nodes that are interfaces between PFETs and NFETs, cascode devices N4, N5, P4, and P5 are introduced. With SOI technology, devices can be operated in, and even switched between, different voltage ranges in close proximity without any additional isolation structures. To provide a suitable gate bias for the PFET cascodes, an additional reference voltage (VREF) is introduced that is approximately a logic swing below the upper voltage rail (VPP) for wordlines. The bitline voltage (VBLH) is deemed suitable for the NFET cascode bias, thereby avoiding a second reference-voltage generator. In the standby state, the inputs GWL (global wordline) and MX (subarray decode signal) are both low, at the negative VWL voltage. The precharge signal PREN is also low, but level-shifted to a logic swing below VPP; thus nodes PU and XPN are precharged high to VPP. With node XPN high, cascode device N4 acts as a source follower, charging node PD to a voltage approaching VBLH. Since the gates of both N3 and N5 are approximately at VBLH, the inactive wordlines are held low to the VWL level that is sourced by a programmable negative-voltage charge pump. Device N5 serves to protect N3 from being exposed to the full VPP voltage, just as cascade PFET P5 protects the pullup P3 from the VWL voltage. The source node (KPR) of P5 is connected to a half-latch keeper, which is composed of three devices in series (Pkt,2,3). This keeper maintains the wordline off state, once the precharge signal (PREn) shuts off, on all but the selected row. For the instance where both GWL and MX go high, nodes PD and XPN are pulled to VWL and device P4 acts as a source follower, discharging node PU and turn on the output pullup P3. Thus, the two nodes that are interfaces between PFETs and NFETs, namely XPN and WL, swing between VWL and VPP to achieve as much as twice the voltage range of standard logic. Lastly, to reduce standby leakage on the charge-pump generated VPP supply, the output stage of the wordline drivers are power-gated in VVPP groups by the SELn signals.

The test site includes several 1.1Mb eDRAM macros plus a PLL, array self-test circuitry, and a charge-pump subsystem. The eDRAM is designed to be compatible with a high-speed clock and a variety of gain ratios, including a mode that supports random-address reads or writes every eight clocks, while the macro data outputs are captured by comparator latches after only four clocks. Hardware access time results versus voltage are shown in Fig. 17.4.6. At a clock rate of 4 GHz, the eDRAM random cycle time becomes 2ns, and the corresponding random access time is 1ns.

Acknowledgements:
The authors wish to thank Roger Purvee for his layout support, Erik Nelson for his hardware data support, and the SRDC eDRAM technology team.

References:

---

1IBM, Poughkeepsie, NY, 2IBM, Austin, TX, 3IBM, Rochester, MN, 4IBM, Williston, VT

316 • 2015 IEEE International Solid-State Circuits Conference
Figure 17.4.1: Macro overview.

Figure 17.4.2: Sense amplifier and read-data mux.

Figure 17.4.3: Write 0 read 0 timings.

Figure 17.4.4: Write 1 read 1 timings.

Figure 17.4.5: Wordline driver.

Figure 17.4.6: Hardware data.
Figure 17.4.7: Micrograph of macro.
17.5 A 3T1R Nonvolatile TCAM Using MLC ReRAM with Sub-1ns Search Time

Meng-Fan Chang¹, Chien-Chen Lin¹, Albert Lee¹, Chia-Chen Kuo², Geng-Hau Yang¹, Hsiang-Jen Tsai¹, Tien-Fu Chen¹, Shyh-Shyuan Sheu¹, Pei-Ling Tseng³, Heng-Yuan Lee², Tzu-Kun Ku²

¹National Tsing Hua University, Hsinchu, Taiwan, ²ITRI, Hsinchu, Taiwan, ³National Chiao Tung University, Hsinchu, Taiwan

Many big-data (BD) processors reduce power consumption by employing ternary content-addressable-memory (TCAM) [1-2] with pre-stored signature patterns as filters to reduce the amount of data sent for processing in the following stage (i.e., wireless transmission). To further reduce standby power, BD-processors commonly use nonvolatile memory (NVM) to back up the signature patterns of SRAM-based TCAM (sTCAM) [3] during power interruptions or frequent-off operations. However, this 2-macro (sTCAM + NVM) scheme suffers long delays and requires considerable energy for wake-up operations, due to the word-by-word serial transfer of data between NVM and TCAM macros. Most of the signature patterns are seldom updated (written); therefore, single-macro nonvolatile TCAM (nvTCAM) can be used for BD-operators to reduce area and facilitate fast-low-power wake-up operations, compared to the 2-macro approach. Previous nvTCAMs were designed using diode-connected 4T2R with STT-MTJ (D4T2R) [4], 2T2R with PCM [5], and 4T2R with ReRAM [2]; however, they suffer the following issues: (1) large cell area (A) and high write energy (Ew) due to the use of two NVM (2R) devices; (2) limited word-length (WDL, k-bits) caused by small current-ratio (=kxlop/kxh,lo) between match-line (ML) mismatch current (iML,lo) and ML leakage current of k matched cells (kxlop,lo); (3) Large search delays (TSD) and excessive search energy (Es) due to large ML parasitic load (CML) and small I-Ratio. nvTCAM is promising for nVTCAM due to its low Es, high-resistance-ratio (R-ratio), and multiple-level cell (MLC) capability. To overcome issue (1) to (3), this study develops an MLC-based 3T1R nvTCAM with bi-directional voltage-divider control (BDVC). A 2x64x64bit 3T1R nvTCAM macro is fabricated using back-end-of-line (BEOL) ReRAM [6] and a 90nm CMOS process, with 2.27× cell size reduction as compared with sTCAM using the same technology and the TSD (=0.96ns) for WDL=64b.

Figure 17.5.1 outlines the challenges of nvTCAMs. The area and wire routing of BEOL-based ReRAM result in considerable area overhead. Moreover, many ReRAMs require higher write current (ISET) to achieve long data retention time (DRT) and large R-ratio, resulting in large NVM-driver area. Thus, previous 2R-based nvTCAM cells need large cell area to meet DRT and R-ratio requirements. sTCAM can achieve short development time (<400ns) for WDL=64b.

References:
Figure 17.5.1: Issues of previous nvTCAM.

Figure 17.5.2: Write operation.

Figure 17.5.3: Search operation.

Figure 17.5.4: Macro structure and cell layout.

Figure 17.5.5: Comparison table.

Figure 17.5.6: Measured results.
Figure 17.5.7: Die micrograph and summary table.

<table>
<thead>
<tr>
<th>Process</th>
<th>90nm Bulk CMOS + H10 ReRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>2 blocks x 64 rows x 64 bits</td>
</tr>
<tr>
<td>Cell size</td>
<td>Using logic rule: 1.566µm²</td>
</tr>
<tr>
<td></td>
<td>Using SRAM rule: 0.8704µm²</td>
</tr>
<tr>
<td>Search Speed</td>
<td>0.98ns (VDD=1V, 25°C)</td>
</tr>
<tr>
<td>Measured Energy/bit/search</td>
<td>0.511J (VDD=1V, 25°C)</td>
</tr>
<tr>
<td>VDDmin</td>
<td>0.48V (25°C)</td>
</tr>
<tr>
<td>Supply Voltage</td>
<td>0.5V–1V (Search)</td>
</tr>
<tr>
<td></td>
<td>1.8V (Write)</td>
</tr>
<tr>
<td>ReRAM Write Speed</td>
<td>SET/RESET &lt; 5ns</td>
</tr>
</tbody>
</table>
17.6 1V 10Gb/s/pin Single-Ended Transceiver with Controllable Active-Inductor-Based Driver and Adaptively Calibrated Cascade DFE for Post-LPDDR4 Interfaces

Junyoung Song¹, Hyun-Woo Lee², Jayoung Kim¹, Sewook Hwang¹, Chulwoo Kim¹

¹Korea University, Seoul, Korea, ²Hynix Semiconductor, Icheon, Korea

Performance improvements in mobile devices with multi-cores and enhanced graphics quality require higher memory bandwidth. Consequently, the design of I/O becomes a crucial issue [1]. In the LPDDR interface, a ground-terminated interface is used for a low-noise termination voltage ($V_{OD}$) and small I/O capacitance ($C_{IO}$) [2,3]. Even though noise margins and power efficiency are enhanced by ground termination, to compensate channel loss, the I/O is still the most power-hungry block. The pre-emphasized output driver and DFE are widely used to remove ISI and maximize read/write margins. However, multiple-taps in the output driver and the DFE are required to cover the channel loss, and they degrade the power efficiency of the I/O and occupy a large area.

The IIR filter effectively reduces the number of DFE taps by cancelling the long-tail ISI, and a 2-tap IIR filter-based DFE (IDFE) has the same effect as a 15-tap FIR filter-based DFE (FDFE) as verified in [4]. However, because the RC-based IIR filter occupies a larger area than the FIR filter, a DFE composed of only IIR filters has a penalty of area cost. In addition, if the IIR filter is used followed by the FIR filter as shown in [5], the slicer to determine the current data for the FIR filter eliminates the residue of the received long-tail ISI, which degrades the effect of the IIR filter. Moreover, in [4] and [5], the delay of the IIR filter is manually controlled to have a 1UI delay. Manual control is difficult to obtain an optimum value of the IIR filter under PVT variations. In this paper, an IIR filter is adaptively calibrated to find the optimum value. Furthermore, a 1-tap IDFE is used to compensate the received long-tail ISI, and the rest of the first post-cursor is cancelled by following a 1-tap FDFE.

Figure 17.6.1 shows the architecture of our transceiver, which is composed of a TX with a controllable active-inductor-based output driver (AOD) and an RX with a calibrated cascade-DFE. In the TX, 1UI delayed data for the pre-emphasis requires additional circuitry that has penalties of large area and power consumption to operate at the target data-rate. The AOD eliminates the requirement of 1UI delayed data, which removes the circuitry required to generate this delay. The serialized data is fed to two types of output drivers, and the outputs of each are terminated to $V_{OD}$ after passing through the channel. For a fair comparison, the NMOS-based output driver (NOD), which has the NMOSes and resistors excluding active inductors is also designed. In post-layout simulation under the same conditions, even though the NOD dissipates a higher power, it achieves a smaller eye opening than AOD.

In the cascade-DFE, the IDFE is used before the FDFE to preserve the received long-tail ISI. The IDFE operates at full rate to reduce the occupied area of the IIR filter while the FDFE operates at half rate to improve the timing margin of the feedback path. The delay of the feedback path in the IDFE is automatically controlled by the IIR filter calibration logic (IIRFC), and the first post-cursor that is not perfectly removed by the IDFE is cancelled by the following FDFE. Also, the tap coefficient of the IIR tap is controlled by detecting errors after comparing the outputs of the IDFE and FDFE. To amplify the received data, we employ a single-to-differential buffer (SDB) composed of an active inductor and cross-coupled NMOSes [6]. A reference voltage generated in the asynchronous adaptive reference voltage detector (AARVD), $REF_{SDB}$, is fed to the SDB.

The block diagram of the cascade-DFE is shown in Fig. 17.6.2. The summer output, $D_{SUM}$, is provided to the slicer and IIR filter, and the output of the IIR filter is fed back to the summer to cancel the long-tail ISI. Due to the RC characteristic of a 1-tap IIR filter, the pulse response of the first post-cursor may not match with the received first post-cursor, and it is cancelled by the following FDFE. In the case of the IIR filter, the AMP that is merged with the RC filter is not matched with the received first post-cursor, and it is cancelled by the following FDFE.

The block and timing diagrams of the IIRFC are shown in Fig. 17.6.3. Because the reset timing of the dividers is controlled together and the duty-cycle ratios of CK/KBK are adjusted in the DCC, the phase difference between rising edges of the divided signals, $DIV_1$ and $DIV_2$, is 1UI. $CON_{IIR}$ is increased until the phase difference between AMS and AMP rises to $\theta_{MAX}$. $SAFF_{OUT}$ always returns to zero after sampling AMP by AMP, therefore, $CON_{IIR}$ is automatically controlled by $SAFF_{OUT}$. The replica circuit of the IIR filter is used to check the delay of the feedback path, and, after calibration, the IIRFC is turned off to reduce the power consumption. In case of AARVD, the maximum and minimum input voltage levels are compared with the reference voltages that are generated by a resistor ladder. As shown in the post-layout simulation results, whenever the input signal exceeds the reference voltage, the outputs of Comp₁ and Comp₂ are toggled, and $REF_{RX_{1}}$ and $REF_{RX_{2}}$ are increased and decreased, respectively. Finally, $REF_{CAL}$ is determined as the common voltage between $REF_{RX_{1}}$ and $REF_{RX_{2}}$.

Figure 17.6.4 shows the circuit and simulation results of the AOD. $MRS_{REV}$ is used to control the output impedance, and the active inductor at the POD used to enhance the driving capability is controlled by $MRS_{REV}$. In the AOD, NMOSes are used for the pull-up operation to reduce $C_{IO}$. The inductance of the active inductor may be varied with PVT variations, so it is designed to be controlled by $MRS_{REV}$. The boosting gain is controlled by binary-weighted taps to attain continuous boosting gain. The post-layout simulated boosting gain of the AOD is shown in Fig. 17.6.4. The controllable range of the boosting gain is from 0.07 to 4.98dB, and the eye opening of the AOD after passing the modeled channel with 10.11dB loss at 5GHz increases from 0.57 to 0.75UI.

Figure 17.6.5 shows the measurement results of the TX. To compare the NOD and AOD, the eye diagrams with various data-rates are measured simultaneously with a 10cm FR4 PCB trace. At 3GB/s/pin operation, the measured eye widths of the NO and AOD are 0.90UI and 0.95UI, respectively. In the case of 10GB/s/pin operation, the measured eye of the NOD is closed, however, the AOD achieves a 0.54UI eye opening. The measured eye widths and the power efficiencies versus operating data-rates are summarized in Fig. 17.6.5. As the data rate is increased, the power efficiency of the AOD is enhanced compared with that of the NOD, and, even though the NOD consumes more power than the AOD over 5Gb/s/pin, the eye of the transmitted data is closed.

The bathtubs of the RX are measured at various data-rates with the output of the AOD through the 10cm FR4 PCB trace as shown in Fig. 17.6.6. The measured eye width is similar when the data rate is less than 3Gb/s/pin. At 3Gb/s/pin operation, the measured BER without AARVD and IIRFC is around 10⁻¹⁰, however, that with AARVD and IIRFC is maintained at less than 10⁻¹². A BER of the RX < 10⁻¹² is measured for up to 10.5Gb/s/pin. With AARVD and IIRFC, the measured eye width with a BER < 10⁻¹² is larger than 0.7UI. As the data-rate increases, the SDB boosts the amplitude of the received data, and the larger eye widths are measured from 8 to 10Gb/s/pin. The transceiver for post-LPDDR4 interfaces is fabricated in a 65nm CMOS process as shown in Fig. 17.6.7. The areas of the TX with AARVD and RX are 0.046 and 0.0045mm², respectively. The IIRFC and AARVD can be shared with other channels, and the active areas are 0.018 and 0.015mm², respectively. Finally, the measured power efficiencies of the TX and RX are 1.16 and 3.02pJ/b, respectively, with a 1V supply voltage.

Acknowledgement: This research was supported by the MSIP, Korea, under the ITRC support program (NIPA-2014-H0301-14-1007) supervised by the NIPA; fabrication was supported by the MPW of IDEC.

References:
Figure 17.6.1: Design of the developed transceiver for LPDDR interface.

Figure 17.6.2: Design of the cascade-DFE and schematics of IIR and FIR filters.

Figure 17.6.3: Block and timing diagrams of the IIRFC and AARVD.

Figure 17.6.4: Schematics and the post-layout simulation results of the AOD.

Figure 17.6.5: Measurement results of the TX.

Figure 17.6.6: Measurement results of the RX and performance summary.
Figure 17.6.7: Chip micrograph of the transceiver.
As the demand for high-frequency DDR SDRAM increases, duty-cycle correction circuits (DCC) become a key element to widen the data-valid window (tDV). For duty detection in a DCC, analog schemes using charge pumps [1] and digital schemes using DLL locking [2] or time-to-digital converters (TDC) [3] are widely used. However, they require a certain amount of time proportional to duty errors or a high-resolution TDC to realize quantization errors. For correction, an edge combiner or slew-rate-changing inverter is commonly used in DRAM applications. An edge combiner utilizes 180°-phase-shifted clocks, which are obtained by DLL locking or TDC code calculation [4-5], to generate both edges, but it has high-frequency limitations due to gate-delay-based short-pulse generation. The slew-rate-changing inverter has trade-offs between range and resolutions or between resolutions and DLL locking time. To achieve both wide range and fast locking, asynchronous binary search for detection and receiver tail-current tuning can be used for correction [6]. In a WCK-based system, duty and phase skew between tck and tckl are related to the detection range. However, for normal DDR DRAM, this range should be doubled resulting in more area as well as locking time. To resolve problems with locking time, coverage, and resolution, a hybrid DCC is presented in this paper. For precise duty error detection, the phase and DCC locking processes should be separated, because delay updating disturbs duty-error detecting and averaging. In this paper, an all-digital DCC with TDC and an enhanced edge combiner performs fast DCC operation before DLL coarse locking, and a slew-rate-changing DCC optimized for fine resolution compensates quantization errors caused by all-digital operation.

The block diagram of the DLL with a hybrid DCC and refresh-aware update scheme is depicted in Fig. 17.7.1. The DLL is composed of coarse and fine delay lines, a slew-rate-changing DCC circuit (SR-DCC), and an analog duty error detector (A-DED) as conventional DCCs. In addition, the DLL has an all-digital duty error extractor (DEX) and a 180° phase aligner with enhanced edge combiner (PIA-DCC) at the input side. With a TDC, this DEX-based DCC has a wide correction range with moderate resolution. After finishing this DCC operation, coarse locking compensates the phase skew between REFCLK and FBCLK, which is changed from the original state due to input DCC. During fine locking, the output-side SR-DCC works with fine resolution. In high-capacity 3D-stacked DRAM, internal noise is an important consideration. Especially in master-slave TSV-stacked DRAM, data can be read from one stack while others are in refresh mode. To cope with internal refresh noise, the refresh-update control block generates an update hold period according to the REF command combinations issued to any stack. This is because power noise is worst during an overlapped period. The DLL monitors this refresh period so that it can adaptively update against power noise. A conceptual diagram of operation is depicted in Fig. 17.7.2.

Figure 17.7.3 shows the complete block diagram of the DEX-DCC, in which the coarse TDC has a 16b register and the fine TDC has a 16b register and an 8b code output for correction. The DEX-DCC can correct duty errors of at least 15% at DDR-1333 and simulated fine resolution is <30ps under all PVT conditions. In a conventional TDC-based DCC, a clock is always delayed by the amount of tCK/2 for falling-edge generation, for which the delay chain requires a large area to cover a wide tuning range. Since this DCC uses lopsided delays, (the clock either has a high or low duty ratio), only the duty-error-correction delay is needed in the fine TDC replica of the phase aligner. The required delay information for correction is generated as 8b codes by a half-code generator. Considering quantization errors, the half-code generator ignores a minimum code to bypass original clocks, thus preventing unwanted duty distortion near 50% input duty cycle. Since conventional edge combiners for duty-corrected clock generation have frequency limitations due to gate-delay-based short-pulse generation, the developed enhanced edge combiner utilizes an edge-triggered D flip-flop with a mux. As shown in Fig. 17.7.3, only rising edges of the 180°-phase clocks trigger the DFF. Since the mux switch signal comes from QB of the DFF, the clock input is switched to another one right after triggering the DFF. A TSPC DFF is used for short propagation delay and setup time. This enhanced edge combiner does not need to generate short pulses, enabling it to achieve higher performance up to 10X of 0.6ns in all PVT simulations.

Figure 17.7.4(a) shows the operation of the 2-step TDC-based duty-error extraction. First, in the DEX, the high and low pulses are separately generated, with the high pulse width the same as the original. Next, the TDC in the DEX samples each pulse width, and then stores quantized digital codes for further comparison. In the digital code comparator, the narrower of the two pulses is chosen to be the replica delay for fine TDC operation. The phase difference from this delayed replica to the wider pulse’s falling edge is twice the duty error. The TDC samples this difference with a resolution of Δ/2 and generates a half code for duty-error compensation. However, the maximum quantization error in the coarse TDC is Δ. Consequently, the maximum error could be Δ even though the fine TDC reduces it by half. To resolve this problem, another 2-step operation is performed in the coarse TDC. In the right of Fig. 17.7.4(a), a 2-step measuring scheme indicates the number of cycles is 158400 cycles from the 1st y-axis input, the TDC samples and stores pulse-width information. In the 2nd measurement, the coarse TDC pulls delay by half of the coarse unit and overwrites the binary reduced information in the register. With more than Δ error, it should be less than Δ in this 2nd measurement. Otherwise, the results of the 1st and 2nd measurements are the same. After measuring twice in the coarse TDC, total quantization error becomes Δ/2. The duty-error detection range is determined by the delay of the coarse TDC and the correction range is determined by that of fine TDC. If it fails to measure with given delay mode for either high or low pulse at low frequency, then the coarse TDC does not enter the 2nd measurement. Instead, the detection range is extended by 2Δ without halving resolution. In Fig. 17.7.4(b), a lopsided delay method for 180° alignment is shown, which allows for reduced area for delay generation.

Figure 17.7.5(a) shows measured tCK shmoos plots of the master and slave2 chips in a 16Gb DDR4 SDRAM stack to demonstrate the achievement of 2.7Gb/s at 1.2V, which is compliant to DDR standard of 2666. With the help of the refresh-aware DLL update scheme, the 4-H TSV DRAM can achieve 0.58UI at DDR-2133 with 1.0V even in burst refresh mode, whereas it achieves 0.36UI when refresh update is off, as shown in Fig. 17.7.5(b). In Fig. 17.7.6, measured locking procedures of cycle versus tDV show delay and DCC locking, where the x-axis indicates the number of cycles and the y-axis indicates tDV. The differences between the solid and dotted lines are the tDV windows of EVEN and ODD. In DEX operation, the input DCC finishes locking within 50 cycles, which can be reduced further by reducing detection range. The micrograph of the chip, fabricated in 25nm CMOS DRAM process, is shown in Fig. 17.7.7. The active areas of the DLL and the DEX-DCC are 0.076 and 0.012mm², respectively.

Acknowledgment
The authors would like to thank Yong Shim for his brilliant idea and support.

References
Figure 17.7.1: The DLL architecture with hybrid DCC and refresh-aware update scheme.

Figure 17.7.2: The operation of refresh-command aware DLL update control scheme.

Figure 17.7.3: Complete block diagram of the DEX-DCC and the timing diagram of the enhanced edge combiner.

Figure 17.7.4: The operations of (a) 2-step TDC-based duty-error extraction and (b) 180° phase aligner using lopsided delay.

Figure 17.7.5: Measured shmoo plot. (a) VDD vs. tCK (b) VDD vs. tDV at DDR-2133.

Figure 17.7.6: Measured locking procedure of DLL with DCC operations.

Figure 17.7.8: Measured locking procedure of DLL with DCC operations.
<table>
<thead>
<tr>
<th></th>
<th>This Work</th>
<th>16Gb 4-H TSV DDR4 SDRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>25nm DRAM process</td>
<td></td>
</tr>
<tr>
<td>Voltage</td>
<td>1.2V</td>
<td></td>
</tr>
<tr>
<td>Operating range</td>
<td>667Mbps – 2647Mbps</td>
<td></td>
</tr>
<tr>
<td>DCC resolution (sim.)</td>
<td>DEX-DCC: &lt;30ps SR-DCC: &lt;2ps</td>
<td></td>
</tr>
<tr>
<td>DCC coverage</td>
<td>-30% ~ +27% → &lt;1% @ 2133</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-29% ~ ±24% → &lt;1% @ 2467</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-29% ~ ±24% → &lt;1% @ 267</td>
<td></td>
</tr>
<tr>
<td>DCC locking time</td>
<td>DEX-DCC: &lt; 50 cycles</td>
<td></td>
</tr>
<tr>
<td>DLL (DEX) active area</td>
<td>0.076 (0.012) mm²</td>
<td></td>
</tr>
</tbody>
</table>

Figure 17.7.7: Chip micrograph and performance summary.