13.1 A 4.6GHz 162Mb SRAM Design in 22nm Tri-Gate CMOS Technology with Integrated Active V_{MIN}-Enhancing Assist Circuitry

Eric Karl, Yih Wang, Yong-Gee Ng, Zheng Guo, Fatih Hamzaoglu, Uddalak Bhattacharya, Kevin Zhang, Kaizad Mistry, Mark Bohr

Intel, Hillsboro, OR

Future product applications demand increasing performance with reduced power consumption, which motivates the pursuit of high-performance at reduced operating voltages. Random and systematic device variations pose significant challenges to SRAM V_{MIN} and low-voltage performance as technology scaling follows Moore's law to the 22nm node. A high-performance, voltage-scalable 162Mb SRAM array is developed in a 22nm tri-gate bulk technology featuring 3rd-generation high-k metal-gate transistors and 5th-generation strained silicon. Tri-gate technology reduces short-channel effects (SCE) and improves subthreshold slope to provide 37% improved device performance at 0.7V. Continuous device width sizing in planar technology is replaced by combining parallel silicon fins to multiply drive current. Process-circuit co-optimization of transient voltage collapse write assist (TVC-WA) and wordline underdrive read assist (WLUD-RA) features address process variation and fin quantization at 22nm and enable a 175mV reduction in the supply voltage required for 2GHz SRAM operation. Figure 13.1.1 shows an SEM top-down view of a 0.092μm² high-density 6T SRAM bitcell (HDC) and a 0.108μm² low-voltage 6T SRAM cell (LVC) after gate and diffusion processing. Computational OPC/RET techniques extend the capabilities of 193nm immersion lithography to allow a 1.85× increase in array density relative to 32nm designs [1].

In a tri-gate technology, the absence of fine-grained transistor width tuning to optimize read stability and write margin presents a challenge in designing minimum-area bitcells constrained by fin pitch. Bitcell optimization via V_t adjustment is employed, but low V_t increases channel leakage and high V_t can limit performance at low voltage. Negative-bitline based write assist [3-4] has shown promising improvements in write V_{MIN}, but is less effective with heavily strained PMOS channel devices and doesn't directly reduce PMOS device variability which dominates write margin distributions. Dynamic V_{SS} boosting [5] or V_{CC} collapsing schemes [2] are comparable, but boosting V_{SS} requires a larger series connected device in the V_{SS} path due to greater current demands, while V_{CC} collapse allows integration with V_{CC} sleep which reduces passgate (M1/M6) oxide stress. TVC-WA [2] is used to temporarily lower VCS below data-retention voltage during write operations, eliminating passgate (M1/M6) to pullup (M2/M4) contention and dramatically reducing sensitivity to PMOS M2/M4 V_t variation without impacting read stability. Figure 13.1.2 summarizes the implemented read- and write-assist schemes using wordline under-drive read assist (WLUD-RA) and transient voltage collapse (TVC-WA), respectively, as well as the 128kb SRAM subarray floorplan. The TVC-WA circuit consists of column-based bias circuitry controlling 4 separate array regions and a programmable pulse generator block.

Suppressed bitline voltage [3] for cell stability enhancement is less effective with trigate transistors due to reduced SCE and provides limited benefit unless M1/M6 are taken into linear region. WLUD-RA is a supplemental circuit to adjust M1/M6 gate overdrive, which behaves similarly to a differential implant between $V_{t,M1/M6}$ and $V_{t,M3/M5}$. WLUD-RA can eliminate the differential implant and allow post-silicon tuning [6]. WLUD-RA is directly integrated into the row decoders with no area overhead. TVC-WA requires 3.4% area overhead in the 128kb SRAM array sub-block, allowing a sub-block array efficiency of 72.6% and sub-block density of 6.72Mb/mm².

The TVC-WA and WLUD-RA circuit implementation is detailed in Fig. 13.1.3. During write operations, the WR CLK signal generates a synchronous WL pulse and controls a programmable pulse generator with 9 settings. The pulse-generator output, TVCPULSE, combines with COLSEL[3:0] to generate active low COLPULSE[3:0]. COLPULSE disables device PVCS in the bias circuit and actively drives the selected VCS to VSS through the NWR device. TVC-BIAS[2:0] provides further VCS signal adjustment by limiting the minimum voltage and slew rate on VCS. Bias cells are pitch-matched to two bitcell heights and are physically distributed across the horizontal VCS bus. The VCS rail is split into independent lines to enable selective collapse on written columns without disturbing partially selected cells along the wordline. VCS is segmented into 2 regions per row decoder, VCS0[3:0] and VCS1[3:0], and driven from opposite edges of a 258b bitline in order to minimize switching capacitance and interconnect resistance. The WLUD-RA voltage is generated by a VCCWL voltage divider circuit. PWB[2:0] are used to lower the VCCWL rail during both read and write operations. PMOS devices minimize systematic shift with respect to the distributed PSLP devices. Implementation on the VCCWL rail is chosen to minimize area overhead at the cost of active power.

Figure 13.1.4 highlights operation of the TVC-WA circuitry. Effective write time in conventional SRAM is typically defined as the minimum time between the falling edge of BL and the falling edge of WL. When using TVC-WA, effective write time is measured from the falling edge of BL to the rising edge of VCS, as outlier cells cannot write at low voltage with VCS at V_{CC}. The WR CLK sets the falling edge of VCS and the COLPULSE signal from the programmable pulse generator defines the rising edge of VCS. The pulse width of the COLPULSE signal, $T_{TVC,PW}$, sets the duration of VCS collapse on the active column and trades off selected cell writability with unselected cell retention. VCS_{MIN} is the minimum voltage on VCS during the write operation, which is modulated directly by the clamp circuits controlled by TVC-BIAS[2:0]. During write operations, unselected cells in the VCS column face increased susceptibility to retention failures caused by internal node leakages when the M2/M4 devices are weakened. VCS_{MIN} is adjusted to ensure that $T_{TVC,EFF}$ is sufficient for write without causing retention failures on unselected cells.

Figure 13.1.5 shows 90th percentile normalized V_{MIN} data collected from a 576kB array with no repair. As the $T_{TVC,PW}$ is swept from narrowest to widest self-timed settings, an optimal minimum is observed. With a narrow pulse width, $T_{TVC,EFF}$ is insufficient to write all of the cells in the array, but at the widest pulses, the resulting low VCS_{MIN} induces unselected cell retention failures along the column. At a fixed $T_{TVC,PW}$, there is an optimum choice of TVC-BIAS[2:0], setting VCS_{MIN}. With minimal TVC, the array is significantly write V_{MIN} limited, but collapsing to VSS leads to elevated retention V_{MIN} at 90°C. Modulating WL voltage with WLUD-RA is a tradeoff between read V_{MIN} and write V_{MIN}. This silicon is read V_{MIN} limited and approaches a balance with write V_{MIN} using suppressed WL voltage and TVC-WA.

Figure 13.1.6 shows a typical shmoo at 95°C with repair enabled. The trigate technology and assist features enable a 70% performance gain at fixed voltage relative to the 32nm design [1] or up to a 175mV reduction in supply voltage at fixed frequency. The V_{MIN} distribution from 1800+ dies demonstrates 700mV V_{MIN} at -10°C and 95°C in a volume environment. The SRAM array in this work is utilized on next-generation Intel 22nm microprocessors [7]. Figure 13.1.7 shows a die micrograph of the 22nm testchip with the 162Mb LVC array.

Acknowledgement:
The authors acknowledge contributions from members of the TMG staff.

References:
[1] Y. Wang et al., "A 4.0 GHz 291Mb voltage-scalable SRAM design in 32nm high-k metal-gate CMOS with integrated power management", *ISSCC Dig. Tech. Papers*, pp. 456-457, Feb. 2009.
[2] Y. Wang et al., "Dynamic behavior of SRAM data retention and a novel transient voltage collapse technique for 0.6V 32nm LP SRAM", *IEDM Dig. Tech. Papers,* Dec. 2011.
[3] H. Pilo et al., "A 64Mb SRAM in 32nm high-k metal-gate SOI technology with 0.7V operation enabled by stability, write-ability and read-ability enhancements", *ISSCC Dig. Tech. Papers*, pp. 254-256, Feb. 2011.
[4] Y. Fujimura et al., "A configurable SRAM with constant-negative-level write buffer for low-voltage operation with 0.149μm² cell in 32nm high-k metal-gate CMOS", *ISSCC Dig. Tech. Papers*, pp. 348-349, Feb. 2010.
[5] A. Bhavnagarwala et al., "A Sub-600-mV, fluctuation tolerant 65-nm CMOS SRAM array with dynamic cell biasing", *IEEE J. Solid-State Circuits,* vol.43, no. 4, pp. 946-955, Apr. 2008.
[6] H. Nho et al., "A 32nm high-k metal gate SRAM with adaptive dynamic stability enhancement for low-voltage operation", *ISSCC Dig. Tech. Papers*, pp. 346-347, Feb. 2010.
[7] S. Damaraju et al., "A 22nm IA Multi-CPU and GPU System-on-Chip", *ISSCC Dig. Tech. Papers*, pp. 56-57, Feb. 2012.

 978-1-4673-0377-4/12/$31.00

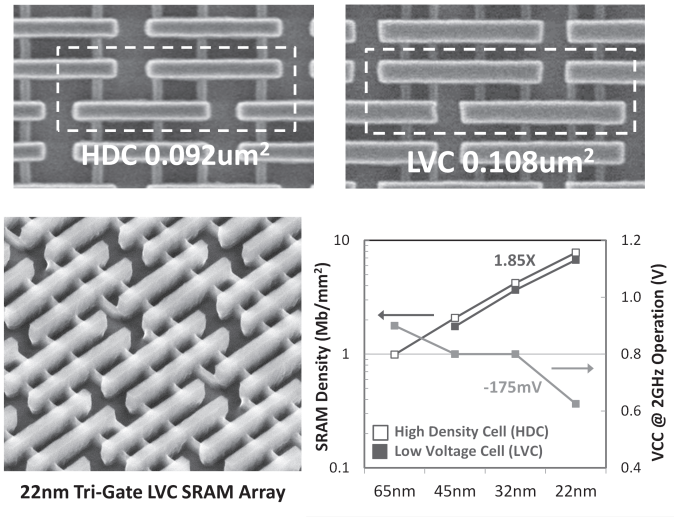
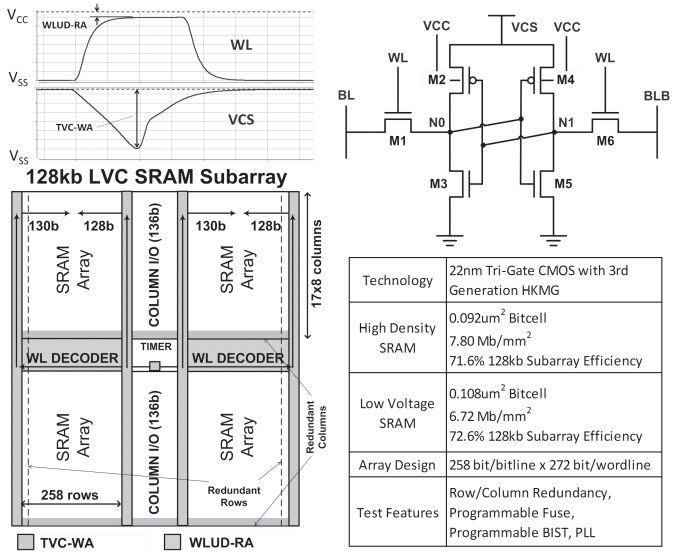


Figure 13.1.1: 22nm HDC and LVC Tri-gate SRAM bitcells.



Technology	22nm Tri-Gate CMOS with 3rd Generation HKMG
High Density SRAM	0.092um² Bitcell 7.80 Mb/mm² 71.6% 128kb Subarray Efficiency
Low Voltage SRAM	0.108um² Bitcell 6.72 Mb/mm² 72.6% 128kb Subarray Efficiency
Array Design	258 bit/bitline x 272 bit/wordline
Test Features	Row/Column Redundancy, Programmable Fuse, Programmable BIST, PLL

Figure 13.1.2: Assist circuit overview, array design and floorplan.

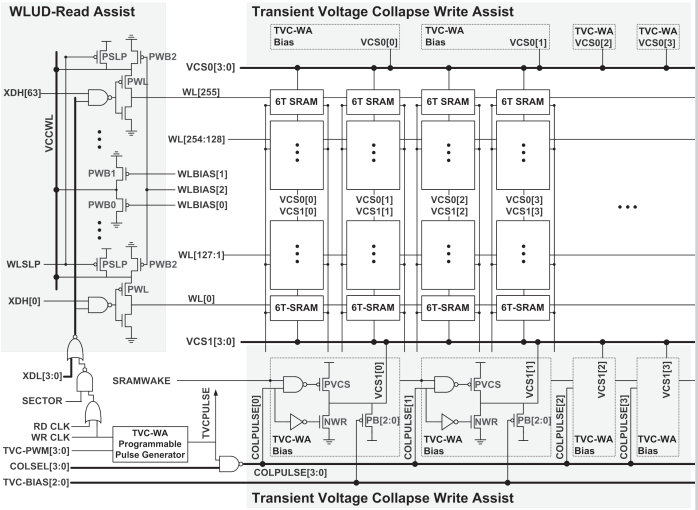


Figure 13.1.3: TVC-WA circuit for write V_{MIN} enhancement.

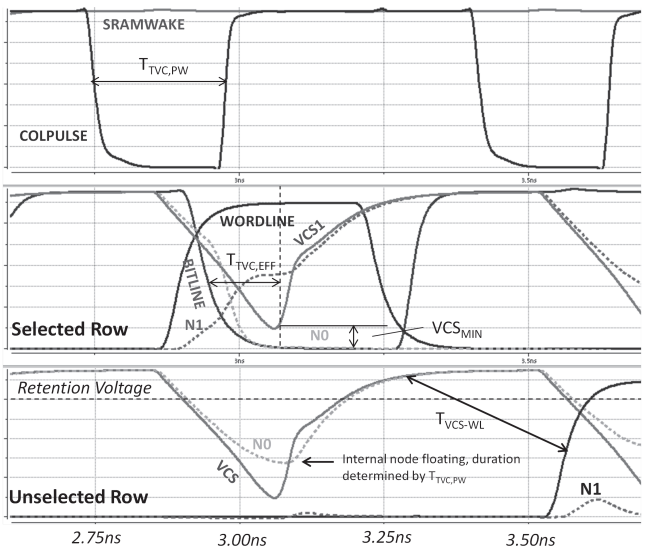


Figure 13.1.4: TVC-WA operation with simulated waveforms.

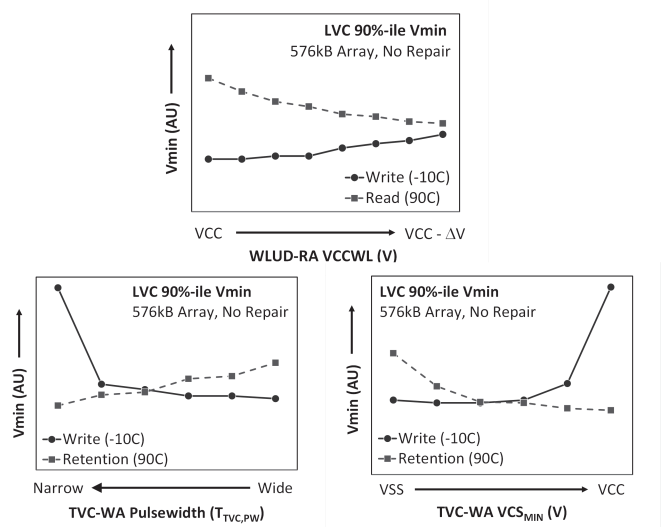


Figure 13.1.5: Read, write, retention V_{MIN} tradeoffs with WLUD-RA and TVC-WA.

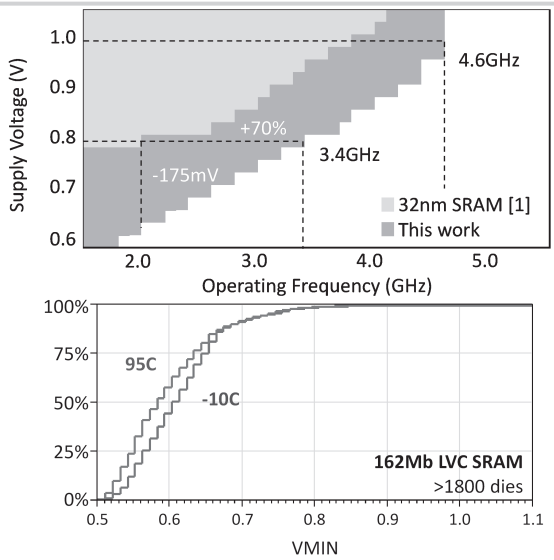


Figure 13.1.6: LVC array shmoo and 162Mb V_{MIN} distribution.

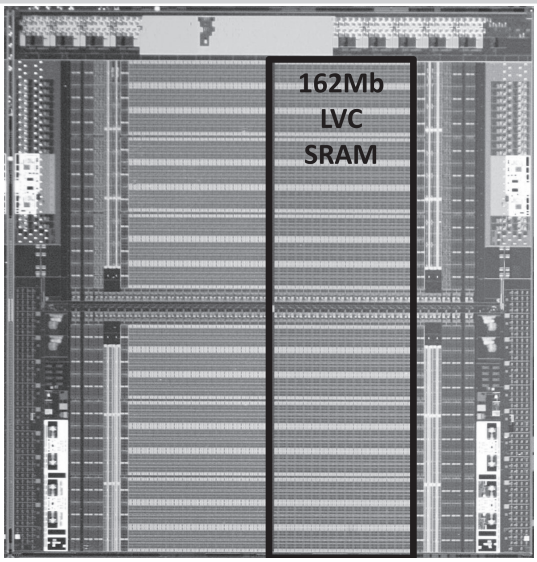


Figure 13.1.7: 22nm testchip die micrograph.

978-1-4673-0377-4/12/$31.00