Dual-V_{CC} 8T-bitcell SRAM Array in 22nm Tri-Gate CMOS for Energy-Efficient Operation across Wide Dynamic Voltage Range

Jaydeep Kulkarni, Muhammad Khellah, Jim Tschanz, Bibiche Geuskens, Rinkle Jain, Stephen Kim, Vivek De
Circuit Research Lab, Intel Corporation, Hillsboro, OR, USA (E-mail: jaydeep.p.kulkarni@intel.com)

Abstract:
A 14KB 8T-bitcell SRAM array is demonstrated in 22nm tri-gate CMOS with fine-grain dual-V_{CC} assist techniques. V_{MIN} limiting 8T-bitcell nodes are boosted selectively during read and write to improve overall chip-V_{MIN}. Measurements show 130-270mV lower V_{MIN} with 27-46% lower power at 0.4-1.6GHz for varying amounts of boosting, array activity and voltage regulator efficiency.

Fine-grain Dual-Vcc Approach:

Dynamic Voltage and Frequency Scaling (DVFS) across a wide range to enable energy-efficient operation requires SRAM array designs that can achieve both high performance and low minimum operating voltage (V_{MIN}). However, process variations induce device mismatches that limit both read and write-V_{MIN} of the 8T bitcell array (Fig. 1). Word-line boosting using charge pump [1] or capacitive coupling [2] were proposed to lower the 8T array V_{MIN} but add design complexity with 11-25% array area overhead. Alternatively, dual-V_{CC} based boosting selectively increases the voltage of critical nodes in an 8T-bitcell while incurring no array area overhead. A separate voltage $V_{BOOST} \leq V_{MAX}$, supplied externally or generated locally from a fixed high input voltage rail (V_{IN}) using a step-down voltage regulator (VR), is used to "boost" selected Read/Write Word-Lines (R/WWLs) and cell-V_{CC} (during read only) (Fig. 2). All remaining array circuits such as R/WWL pre-decoder, pre-charge logic, local and global bitline (LBL/GBL) sensing, timer, and column-I/O drivers are connected to the variable $V_{CC} \leq V_{MAX}$ that is shared with core logic operating across a wide DVFS range. By decoupling the V_{MIN}-limiting 8T bitcell from remaining array and core logic, overall chip-V_{MIN} can be reduced, thus improving energy efficiency. During a read operation, selected RWL and associated bitcells are switched to V_{BOOST} to enable overdrive of the read port transistor stack (Table 1). This alleviates keeper contention and also improves LBL evaluation delay compared to the baseline single-V_{CC} design. During a write operation, selected bitcells remain at V_{CC} while WWL is boosted to mitigate contention between the pass NMOS and pull-up PMOS in the bitcell. WWL boosting also aids write completion by passing a strong "1" through the pass NMOS.

Dual-Vcc 8T Array Circuits:

A dynamic level shifting NAND WL decoder replaces the static single-Vcc NAND implementation while fitting in the same area. The common RD/WR clock, driving the pre-charge/evaluate devices (P_1-N_1) in the dynamic NAND decoders, is level shifted and optimized for equal rise/fall delays (Fig. 3). A stacked delayed WL keeper (K_1-K_2) is used to speed-up the dynamic NAND evaluation and to recover the delay penalty due to RD/WR level-shifting clock. To switch the bitcell between V_{BOOST} (read) and V_{CC} (write), per column V_{CC}-mux (M_1-M_4) is used in the local I/O (Fig. 4). Dual-output split level shifters drive the V_{CC}-mux control signals to V_{BOOST} and are placed in the pre-decoder gap area created by the LBL I/O logic [3](Fig. 4). At very low voltages, dual-V_{CC} read-V_{MIN} is limited by the LBL merge NAND PMOS P_2 and not by the 'boosted' bitcell (Fig. 5). Similarly, dual-V_{CC} write-V_{MIN} is limited by peripheral logic and not by the bitcell as the pull-down NMOS N_3 (initially at V_{BOOST} from a preceding read operation) contends with the write driver PMOS P_3 (Fig. 6). For single-V_{CC} design, the baseline bitcell is upsized to meet the V_{MIN} target, resulting in a large delay margin at/around V_{MAX} (Fig. 7). However with optimal boosting using dual-V_{CC}, the bitcell can be downsized and/or converted to high-V_T devices, to meet performance target across V_{MIN}-V_{MAX} range.

Measurement Results:

We have implemented a 14KB zero area overhead, dual-V_{CC} 8T-bitcell SRAM array in 22nm tri-gate CMOS (Fig. 13) [4]. Bit failure rates (P_{FAIL}) are measured for different V_{BOOST} values above V_{CC} and incremented in 50mV steps. Extrapolations of the measured P_{FAIL} vs. V_{CC} data to a 1MB target array size demonstrate 130mV lower read-V_{MIN} and 290mV lower write-V_{MIN} compared to the baseline single-V_{CC} design at 1.6GHz (Fig. 8). At lower frequencies (< 1GHz) larger V_{MIN} improvement is achieved with only 100mV of boosting since V_{MIN} is now governed by contention during read/write operation as opposed to completion of the operation (Fig. 9). RWL-only boosting offers only 40mV read-V_{MIN} improvement while boosting the full read port (RWL and cell-V_{CC}) lowers V_{MIN} by 130mV at 1.6GHz (Fig. 10). Weakening the keeper on top of read port boosting improves read-V_{MIN} marginally. Noise-induced failures increase marginally with read port boosting, and can be mitigated with a slightly stronger keeper (Fig. 10). Array-V_{MIN} is reduced by 130mV, resulting in 27% lower total array power for optimal boosting of 150mV at 1.6GHz (Fig. 11). Operation of the dual-V_{CC} 8T bitcell SRAM across a wide voltage range is achieved by gradually increasing V_{BOOST} value as V_{CC} is scaled down (Fig. 11). The total power savings depends on conversion efficiency (η) of the step-down VR used to generate V_{BOOST} locally from the fixed high input voltage rail (V_{IN}), clock frequency, and array activity factor (α). For 50% VR efficiency and 10% array activity factor, the total power savings at V_{MIN} is 27% (46%) at 1.6GHz (400MHz) (Fig. 12).

Acknowledgements

The authors sincerely thank K. Ikeda, T. Hwa Foo, D. Jenkins, D. Finan, C. Tokunaga, T. Nguyen, P. Aseron, and R. Forand for their help and support. This research was, in part, funded by the U.S. Government under contract number HR0011-10-3-0007. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government

References

[1] A. Raychowdhury et al., *ISSCC* pp. 352-353 Feb. 2010
[2] J. Kulkarni et al., *ISSCC* pp. 234-236 Feb 2012
[3] S. Hsu et al., *ISSCC*, pp. 178-179, Feb. 2012.
[4] C. H. Jan, et al., *IEDM*, pp.44-47, December 2012

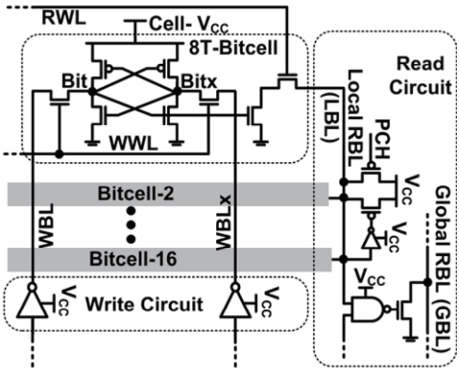


Fig. 1 8T Bitcell SRAM organization

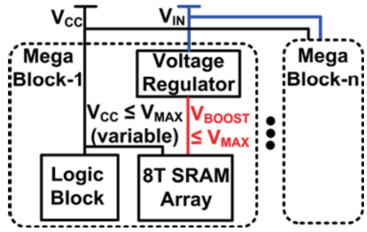


Fig. 2 Fine grain dual-V_{CC} array

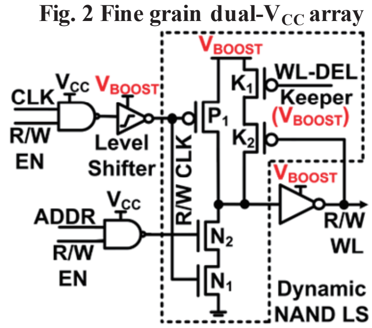


Fig. 3 Level shifting dynamic NAND WL decoder

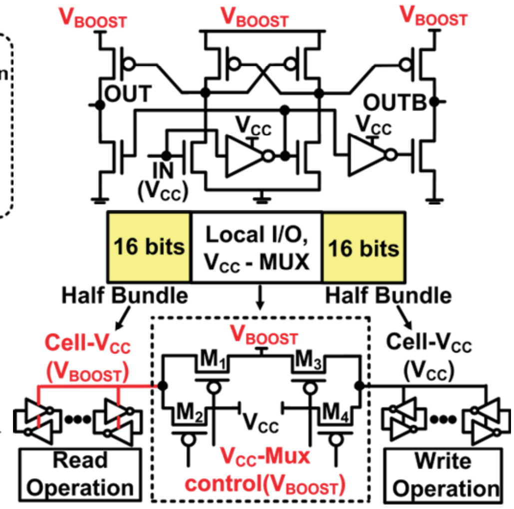


Fig. 4 dual output level shifter for V_{CC}-mux control signals and V_{CC}-mux placement in local I/O

Table. 1 Dual-V_{CC} 8T bitcell node voltages

	RWL	WWL	Cell-V_{CC}
Read	V_{BOOST}	0	V_{BOOST}
Write	0	V_{BOOST}	V_{CC}
Retention	0	0	V_{CC}

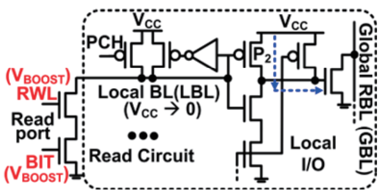


Fig. 5 Dual-Vcc array read-Vmin limit: LBL sense completion (P_2)

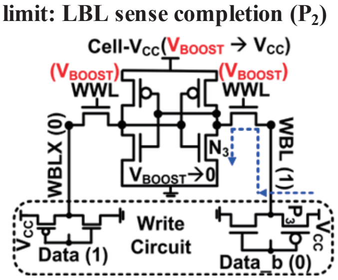


Fig. 6 Dual-Vcc array write-Vmin limit: bitcell NMOS (N_3) and WBL driver PMOS (P_3) contention

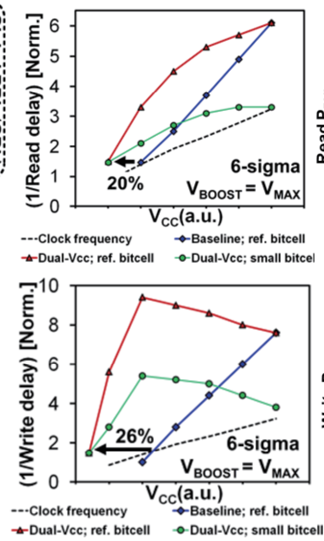


Fig. 7 6-sigma statistical simulations of dual-V_{CC} read, write delay vs. V_{CC}

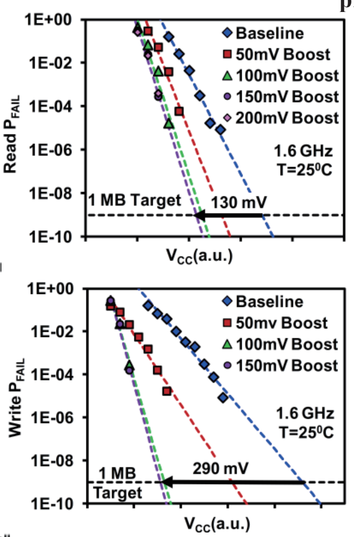


Fig. 8 Measured read, write failure rates (P_{FAIL}) vs. V_{CC}

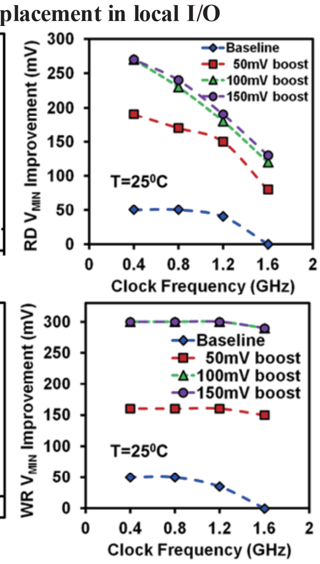


Fig. 9 Measured V_{MIN} improvement vs. frequency

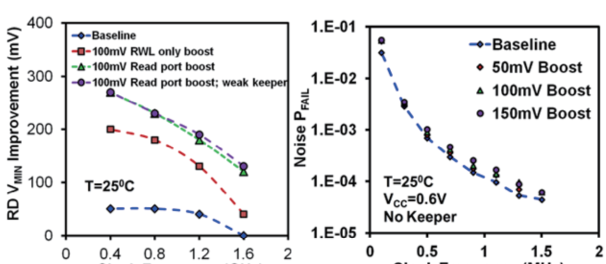


Fig. 10 Measured V_{MIN} improvement and noise-induced failure rates for different dual-V_{CC} read assist techniques

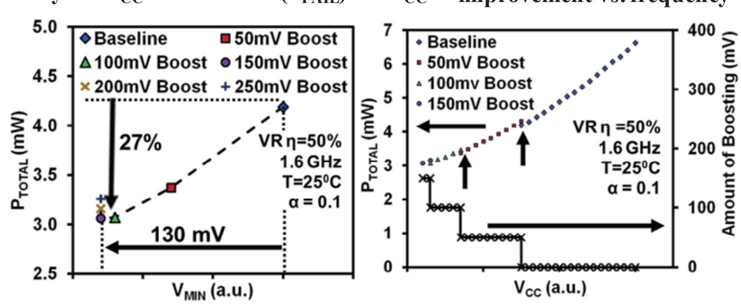


Fig. 11 Measured P_{TOT} vs. V_{MIN} and P_{TOT} across wide voltage range

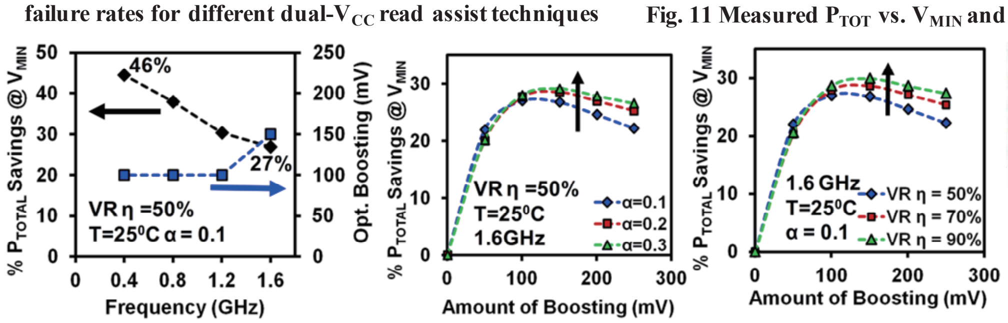
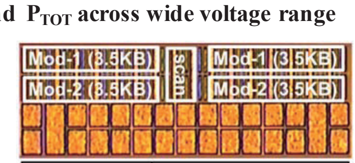


Fig. 12 Measured P_{TOT} savings at V_{MIN} vs. clock frequency, array activity (α) and VR efficiency (η)



Technology	22nm Tri-gate CMOS
Transistors	~1.8 million
Die Area	0.23mm²
Bitcell Area	0.238µm²
Test interface	Membrane probe
Pad count	30

Fig. 13 Die photo and summary