A stacked voltage domain SRAM is proposed where arrays are split into two sets (top and bottom) with their supplies connected in series. System supply current is reused by top and bottom sets, and supply voltage is divided among the two sets of arrays, enabling seamless integration of very low voltage SRAM retention in a larger system with a nominal supply, without need for an efficiency-reducing LDO. An array swapping approach provides stable access to arbitrary banks within one system clock cycle. A comprehensive sizing strategy (W&L) is employed to optimally balance hold stability and bitcell size. Integrated in an IoT imaging system in 40nm CMOS, the proposed 8.9Mb SRAM achieves 1.03pW/bit leakage, a >100× reduction over conventional SRAM in the same technology.

Keywords: low leakage, sub/near threshold, SRAM, voltage stacking, charge recycling, array swapping

Introduction
Intelligent IoT devices seek to fit complete neural network models into on-chip memories to avoid costly off-chip DRAM accesses. As a result, SRAMs can consume >80% of chip area and 90% of standby power [1]. Prior work focused on reducing SRAM leakage via techniques such as HVT/thick-oxide devices [2-3], reverse body bias [4-5], floating bitline [6], raising VSS [7-8], and lowering VDD [9-10]. Apart from HVT usage, which enables 10× leakage reduction and is readily deployed, supply voltage lowering is one of the most effective approaches to reduce leakage due to the DIBL effect. However, it raises two issues: 1) Commercial bitcells are not sized to hold data at subthreshold supplies and require a careful hold margin / density tradeoff; 2) Voltage regulation is required to generate the voltage level for SRAM arrays. LDOs are conventionally used, incurring area and power overheads due to efficiency loss. Voltage stacking generates an intermediate voltage level by placing voltage domains in series and has been used in microprocessors [11] and high bandwidth data buses [12]. The major challenge in voltage stacking is balancing the active current between top and bottom levels, to maintain a stable mid-rail voltage level. This often requires an additional voltage regulator, reducing the stacking benefits. SRAM arrays, however, are dominated by near-constant leakage (writing a bit draws 10s of pA average current versus μA-level background leakage), making them ideal for voltage stacking.

Voltage Stacking and Array Swapping
To allow access to arbitrary arrays during operation while avoiding insertion of complex level converters, we propose a novel array voltage stacking mechanism. The SRAM peripherals are not stacked and therefore wordline and bitline voltages remain at VDDcore (~2VDDmid) for faster operation speed, inherent write/read noise margin enhancement (Fig. 1). This also removes the need for level converters but, as a result, only bottom arrays can be read/written. When a top array is accessed, the memory controller first swaps the voltage of a bottom array in the same bank which the desired top array (Fig. 3). This swap mechanism ensures the leakage current remains balanced and is completed in one system clock cycle due to the relatively low IoT processor clock frequency. In addition to leakage reduction from reduced supply voltage, the approach offers an additional 2× leakage reduction in top arrays due to their inherent reverse body bias and reduced bitline leakage effects. As a result, total leakage is minimized by increasing the fraction of top arrays to >50% (e.g., 75%); this is supported by measurements while the optimal ratio can be set by a memory controller.

References

Measurement Results
The proposed SRAM was implemented in a 40nm CMOS image processing IoT chip with 8.9Mb memories (Fig. 4). Fig. 5 shows measured leakage across voltage and temperature. As the number of top arrays increases, the mid-rail voltage rises while leakage continually drops (Fig. 6). Fig. 7 shows excellent mid-rail voltage stability; VDDmid varies only ±16mV across 100°C, drops ≤1.74mV when arrays swap every 11 cycles, and is unaffected with read/writes every cycle at full speed. The design operates at 438kHz at 0.7V (enabling 14fps in the supported image processing system) with 67fl/bit access energy at 0.6V. Fig. 8 compares this work to other state-of-the-art low leakage SRAMs. The proposed work achieves 1.03pW/bit at 0.58V without using large thick-oxide device, extra supply levels, or SOI with aggressive body biasing.

Fig. 1 Proposed array stacking and swapping technique.

Fig. 2 Bitcell schematic and layout (top), hold noise margin (HNM) and leakage versus bitcell sizing (bottom-left), currents on the bitline during read operation (bottom-right).

Fig. 3 SRAM bank architecture (top), array swapping algorithm and power switches (bottom).

Fig. 4 Die photograph and shmoo plot.

Fig. 5 Leakage across temperature and voltage.

Fig. 6 Mid-rail voltage and leakage with number of top arrays.

Fig. 7 Mid-rail variation with temperature (left), voltage drop due to various memory activities (right).

Acknowledgement
We thank Sony Semiconductor Solutions Corp./Sony Electronics Inc. for supporting this work

Fig. 8 Comparison table and design space landscape.