RC25180 (W1107-008) July 5, 2011
Electrical Engineering

IBM Research Report

The Design and Characterization of a Half-Volt 32nm Dual-Read 6T SRAM

Jente B. Kuang[1], Jeremy D. Schaub[1], Fadi H. Gebara[1], Dieter Wendel[2], Thomas Fröhnel[2], Sudesh Saroop[3], Sani Nassif[1], Kevin Nowka[1]

[1]IBM Research Division
Austin Research Laboratory
11501 Burnet Road
Austin, TX 78758
USA

[2]IBM Deutschland Research and Development GmbH
71032 Böblingen, Germany

[3]IBM Semiconductor Research and Development Center
Hopewell Junction, NY 12533
USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

The Design and Characterization of a Half-Volt 32nm Dual-Read 6T SRAM

Jente B. Kuang[1], Jeremy D. Schaub[1], Fadi H. Gebara[1], Dieter Wendel[2], Thomas Fröhnel[2], Sudesh Saroop[3], Sani Nassif[1], and Kevin Nowka[1]

[1]IBM Research Division, Austin Research Laboratory, 11501 Burnet Road, Austin, TX 78758, U.S.A.
[2]IBM Deutschland Research und Development GmbH, 71032 Böblingen, Germany
[3]IBM Semiconductor Research and Development Center, Hopewell Junction, NY 12533, U.S.A.

Abstract — Dual read port 6-transistor (6T) SRAMs play a critical role in high performance cache designs thanks to doubling of access bandwidth, but stability and sensing challenges typically limit the low voltage operation. We report a high-performance dual read port 8-way set associative 6T SRAM with a one clock cycle access latency, in a 32nm metal-gate partially depleted (PD) SOI process technology, for low-voltage applications. Hardware exhibits a robust operation at 348MHz and 0.5V with a read and write power of 3.33 and 1.97mW, respectively, per 4.5KB active array when both read ports are accessed at the highest switching activity data pattern. At a 0.6V supply, an access speed of 1.2GHz is observed.

Keywords — low power electronics, SRAM chips

I. INTRODUCTION

SRAM supply voltage scaling has faced significant challenges in advanced technology nodes due to the more stringent requirements from performance, yield, chip margining, and process variability. Continued scaling of the SRAM voltage supply remains very important because of the tightened energy budget in contemporary computing systems. High-performance server-class data caches [1, 2] often take advantage of the multi-port capability on multi-thread and out-of-order computing platforms. A split word line 6-transistor (6T) SRAM is quite often the preferred building block compared to its 8-transistor (8T) counterpart because of the port configuration flexibility and area compactness. However, 6T SRAM, in particular the dual read port rendition, is known to be less voltage scalable than 8T SRAM [3, 4] due to stability limitations. It has been reported that technology interventions in addition to circuit techniques are often required to achieve successful low voltage 6T SRAM operations [5, 6], and quite usually at a much relaxed operating frequency. Additionally, the floating body induced history effect in the SOI technology deserves careful margining considerations.

In addressing these needs and challenges, we recently reported a set associative cache design point targeting the low supply voltage spectrum [7]. In this paper, we will present in detail the high-performance 6T SRAM design which has the following features: (i) two independent read ports and one write port; (ii) one clock cycle read or write access; and (iii) voltage scalability to 0.5V.

This is a successfully demonstration of the robust, yield-conscious SRAM design for processor cores in a 32nm high-k metal-gate partially depleted (PD) SOI process technology [8] in the 500mV (1.5-2Vt) power supply regime. The design employs fully adaptive clock alignment and internal pulse width tuning while consuming very little added design overhead and no area penalty. Our objective is to extend the performance-based SRAM application space of a nominal 1V technology, from the traditional higher voltage high-speed domain [1, 2, 9], to the half-volt energy conscious domain for low power computing, hand-held, and mobile applications, in addition to addressing the tightened energy budget for server class memories.

II. DESIGN FEATURES

A. *Subarray Design*

Shown in Fig. 1 is the 32nm performance optimized SRAM cell with split word lines, where each word line can be independently accessed in a single ended fashion during the read mode and both word lines are activated in a differential fashion during the write mode. Effectively, the 6T SRAM functions as having three (2R+1W) ports at the array interface. To ensure performance and dynamic

voltage scaling (DVS), capacitance is minimized by utilizing the Metal-1 interconnect layer for complimentary 16-cell short local bit lines. The thin SRAM cell is designed with a high β (pull-down vs. pull-up device current ratio), low α (pull-up vs. pass-gate device current ratio), compensated gate length, custom gate-to-gate pitch, and optimized threshold voltage for performance, wider noise margin, better writability, yield, and low leakage. The custom gate-to-gate pitch in the SRAM cells, apart from the peripheral devices, results in tighter control of SRAM device parameters and reduced variability.

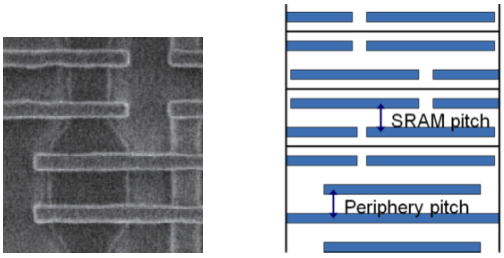

Fig.1 (a) Top-down zoomed view of a portion of the two adjacent thin SRAM cells; (b) a one-bit width picture showing the SRAM and periphery interface where two regions employ two independently optimized gate-to-gate pitches.

A two-level hierarchical read method is employed in this design. The short bit line architecture ensures crisp local bit line discharge time at low voltages where the read current becomes much degraded. It also makes the bit line restore (precharged to high) time window short enough to maximize the bit line discharge (evaluation) time window during low read current conditions. In place of the traditional cross-coupled differential sense amplifier, two compact single ended domino stages are used to evaluate true and complement sides of the local bit lines. A simplified read cross sectional diagram for one of the two ports is shown in Fig. 2 where encircled FETs are regular Vt devices and the rest are high Vt devices. Both ports, assigned with unique addresses, can be accessed independently by the system. The design must sustain stable access when two read ports are simultaneously activated. This is achieved by (1) employing the higher read current L1 SRAM cell and (2) independently optimizing read and write evaluation pulse widths to meet the margin and stability requirements. Outputs from the local and global evaluation circuits, i.e., the first and second level large signal sensing logic, are then merged into the final dynamic multiplexer cycle-boundary latch, which interfaces the external support logic. The dynamic multiplexer serves the dual purpose of selecting the read set and realizing a area efficient built-in scan port for the adjacent latch [10]. Fig. 3 shows a simplified write cross sectional diagram, where the set associative function is integrated into the write operation. To ensure voltage scalability, the two write NFETs receive regular, instead of high, Vt device implants.

Lithography simulations on critical shapes and layout improvements for subarray components, such as sensing circuits and word line drivers, were iteratively performed, to ensure <1/4nm image difference on wafer, during the physical design phase to ensure high-yield printability on wafer.

Word lines are 72 bits long, designed to fit a 9-bit byte and 8-way set associative high-performance architecture. This small granularity building block construct is chosen to ensure healthy word line slews at low voltages while meeting the speed requirement of a typical server class Level-1 data cache. Four banks of the local bit line outputs from the 72x32 sub-array are further multiplexed by a global bit line domino evaluation circuit. Finally, a fast port-specific

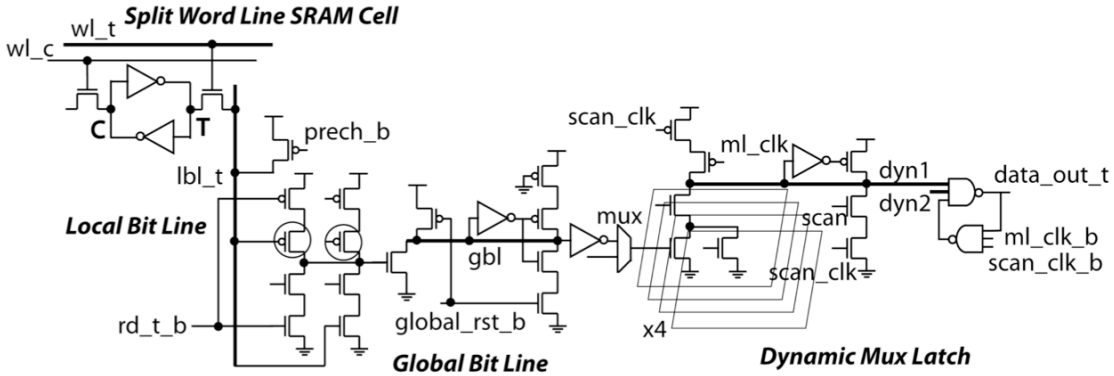

Fig. 2 Simplified read access cross sectional diagram for the true (t) side from the split word line SRAM cell, to the local bit line, to the global bit line, and to the set selection output mux latch. Read sense PFETs are circled in the graph. The dynamic output latch merges 8-way data from the global bit line with a single stage complex evaluation tree, converts the data to static signals, and has a compact built-in scan function.

way-specific 8-to-1 dynamic multiplexer latch merges and selects the accessed data set to the interface logic. This multiplexer latch also includes a scan port for cycle boundary scan function, which is used for the debug and verification process during hardware bring-up. Both array read and write accesses are completed in one clock cycle with the read/write address and way selection bits prepared and set up within half cycle before the start of access, thus making a very competitive design point in the embedded high-performance SRAM application category.

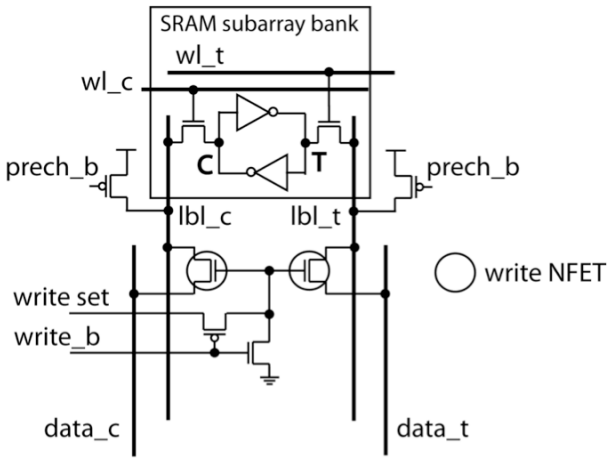

Fig. 3 Simplified write access cross sectional diagram, showing complementary banked write on a short 16- cell bit line architecture. where lbl_c, lbl_t, wl_c, wl_t, data_c, and data_t denote the complement and true sides of the local bit line, word line and input date line, respectively.

B. Design for Low Power and Low Supply Voltage

This SRAM test chip is designed with three voltage domains: array supply (Vcs) for SRAM cells, word line drivers and the last stages of decode circuit; logic supply (Vdd) for periphery and supporting circuits; and clock supply (Vosc) to power the voltage controlled oscillator, which is the on-chip high-speed clock source. The infrastructure for separate array and logic supply is very effective in achieving design robustness across broad operating voltage and temperature range as well as process corners.

The decoupling of array and logic supplies provides an added flexibility for stability and performance optimization. It, nevertheless, adds to the system infrastructure and implementation cost. In this paper, we focus on the design consideration and hardware characterization when only one supply voltage is applied to both the array and peripheral logic.

In order to achieve functional robustness in the low supply region, most blocks employ only low stack height rail-to-rail circuit topologies. On a few occasions, three high stack static or dynamic circuits are used, but only in a controlled way where only the critical input gate triggers the output state transition with general set-up time margins given to the other two inputs. No internal voltage depends on ratioed level division of transistors or resistive components. There is often read-write collision prevention logic in high-performance caches [2]. However, considering the low voltage design point, we choose not to implement this feature inside the decode logic of the array building block for the very reason of avoiding high-stack circuit topology with competing critical timing paths.

To achieve low active power, the majority of the transistors on chip are high Vt devices. Super-high Vt devices are used in non-timing critical blocks. In very limited places, regular Vt devices are used to better meet the critical path timing and regulate slews across the functional voltage range. Examples of such are the read sense PFET in the local bit line evaluation circuit, and certain static gates in the read decoder. However, within any given static gate, there is no pull-up vs. pull-down Vt mixing to ensure controllable trip points across the operating voltage range.

Adaptive clock tuning, as depicted in Fig. 4, is a necessary feature as the delays of different circuit component scale differently with voltage. Precision alignment of internal array clocks is crucial to the success of a robust array operation. Built into this design is an adjustable clock delay line that interlocks the read clock, write clock, (word line clock as derived from the read/write/decode as well as various reset clocks), and global bit line clock timing by changing the arrival time from clock mesh to the local block buffers. Fig. 5 shows several examples of the mesh clock delay cells used in this design. Fixed and programmable delay cells are depicted on the left and right side, respectively where the arrowed transitions depicted the critical falling edges, which are controlled and timed in precision. It ensures timing correlation at every mesh clock entry to the local clock buffer (LCB) besides meeting the skew control requirements within the local clock distribution system.

Inside each LCB are individual controls that change the clock pulse width, delay the rising edge of the clock, or even change the narrow clock pulse to a

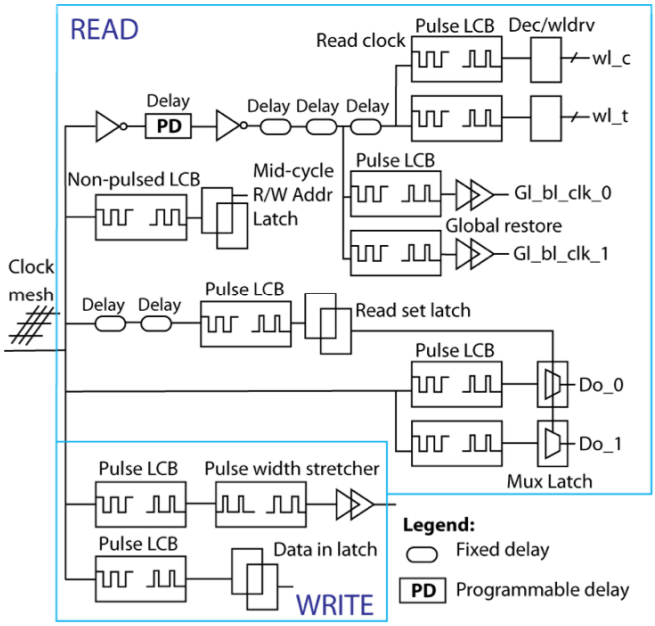

Fig. 4 Clock distribution diagram of local clock components in the array segment building block. Built-in delay and pulse adjustment knobs in LCBs and delay blocks assit the write and read evaluation waves into and from the SRAM cell.

50% nominal duty cycle. The default design point is a short evaluation pulse for all clocked components, with the exception of mid-cycle address set-up latches, to gain speed and power benefits. Short clock pulse widths, when properly timed and aligned, reduce the contention power in the design. Another feature of this design is that, each, some, or all the short pulsed local clocks can be switched to the half-cycle clock mode if (i) more evaluation time is needed, or (ii) non-uniform delay scaling in different paths reduces the required signal overlap window.

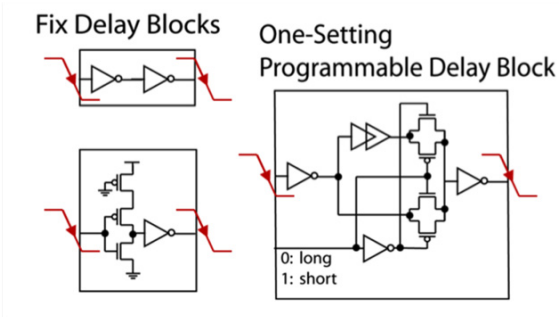

Fig. 5 Examples of delay cells used in the local clock blocks. Fixed delay (FD) cells are shown on the left side. A one-setting programmable delay (PD) cell is shown on the right side.

Fig. 6 shows the schematic diagram of a programmable pulsed array LCB where the falling edge of nominally 50%-duty cycle mesh clock

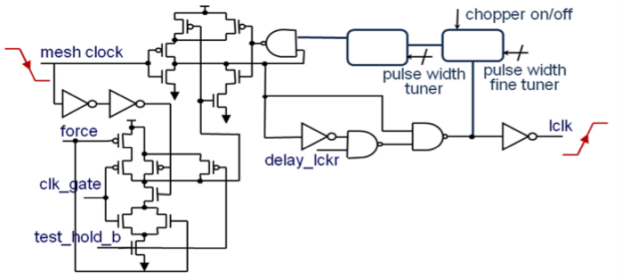

Fig. 6 Programmable pulsed local array clock generation where the falling edge of mesh clock triggers the rising edge of internal array launch clock (lclk). This LCB integrates a front-end control circuit and programmable chopper.

produces the rising edge of internal array launch clock (lclk). Such a style of the local clock methodology for server integration is described in [11, 12]. This LCB integrates a front-end control circuit, handling modes of operation and test, and a programmable chopper with pulse width tuners as well as chopper mode on/off capability. The pulse width fine tuner is common for both array clocks and on-chip infrastructure latches. It is adjusted with one shared setting. The array specific pulse width tuner is in-circuit location specific, and is used to achieve the best timing configuration. The multiple variable clock optimization scheme results in improved functionality, performance, power, and guard band against read/write margin degradation.

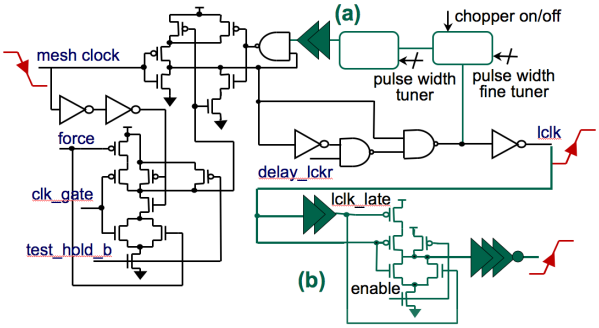

Fig. 7 Programmable pulsed LCB with two pulse width stretcher options: (a) ending edge stretch by a chopper path delay adder; (b) post-LCB OAI stretching component with clock enable control.

Pulse width can be further stretched at spots needing additional evaluation time, such as the write clock. Fig. 7 shows two options we use. style (a) adds and extra delay in the chopper feedback path while style (b) adjust the pulse directly with lclk from the LCB output.

C. *Experiment Infrastructure*

The 4.5KB (2x2x72x128) array segment, complete with peripheral logic and boundary latches, is the building block of this test chip. The entire experiment is made of 32 segments in a 2x16 arrangement, intentionally placed across a vertical (Y) distance span of 4.7mm, mimicking the instantiation and utilization scenario in a multi-core high performance system. The total cell density for this L1 cache chiplet amounts to 1.125Mb. Thus, yield and variation statistics can be characterized. Fig. 8 shows a simplified block diagram for this experiment padcage, and Fig. 9 is a physical snap shot of the SRAM chiplet.

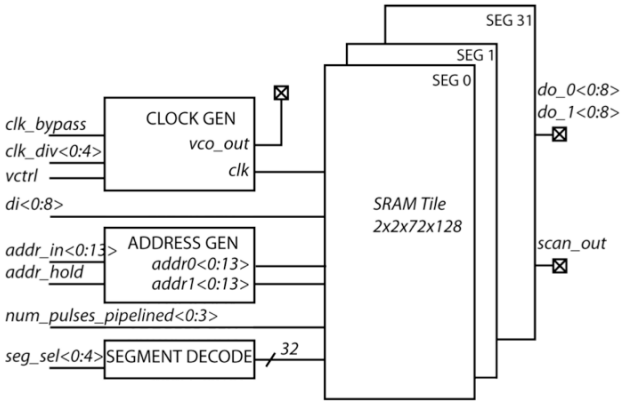

Fig. 8 Simplified padcage block diagram

To facilitate high-speed measurements, there is an engineered clock distribution system within each segment tile with minimized skews into the local block buffers. Time of flight is allowed between different tiles in the vertical (Y) direction of the experiment pad cage as they are timed and measured independently of each other; i.e., instead of a narrow and long synchronous clock mesh, 16 local timing domains are built. Within each, there are capabilities of performing high-speed data I/O, read/write mode change, and address switching. The insert of Fig. 9 shows the measured divided VCO frequency vs. control voltage.

In addition, this chiplet infrastructure incorporates the features of (1) programmable configuration scan registers to facilitate adaptive control of internal pulse width and delay settings; (2) pipelined registers of four cycles deep to enable back-to-back switching of address bits, write/read access modes, and port selection combinations.

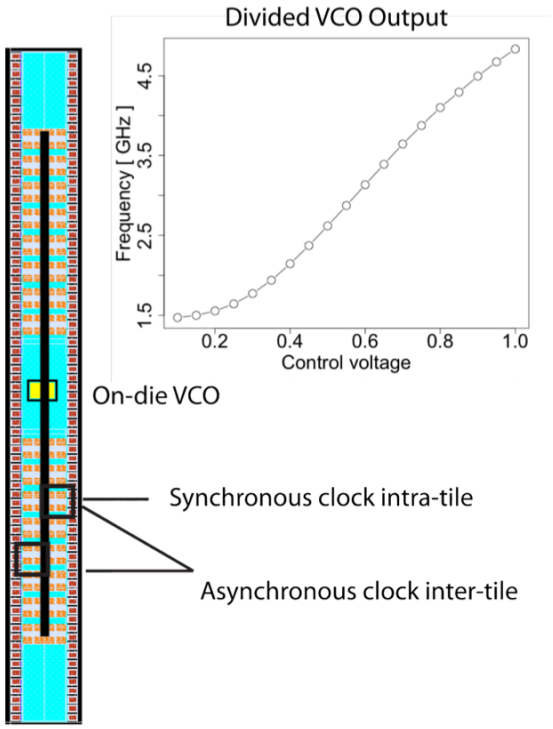

Fig. 9 Snap shot of the 1.125Mb chiplet in a 2x75 0.6x6mm^2 padcage, showing placement of SRAM tiles and on-chip clock generation block. Measured VCO frequency characteristics are shown in the insert.

III. MEASUREMENT RESULTS

While the hardware performs above 4GHz at a 1V supply (Table I), we focus primarily on the low voltage design aspects and characterization. Figs. 10 and 11 depict the frequency vs. supply voltage schmoo diagram of the array under the pulse mode and half cycle clock operating conditions, respectively. The chip functions down to 0.6V in the pulsed clock mode. At 0.6V, the maximum access speed is 1.2GHz. Below 0.6V, by turning the short pulsed array clocks to the longer 50% duty cycle clocks, we successfully extended the minimum stable operating voltage to 0.5V at an access frequency of 348MHz.

Table I Measured nominal array performance reference

Power Domain Specification	Operating Frequency
Single Supply Vdd=Vcs=1V	4.17 GHz
Dual Supplies Vdd=1V Vcs=1.2V	4.43 GHz

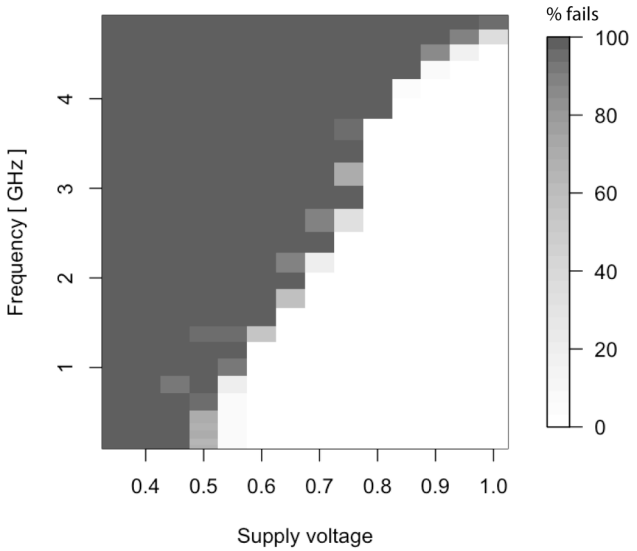

Fig. 10 Frequency vs. supply voltage schmoo diagram for array clocks in the pulse mode operation. The SRAM is operating under the one-supply (Vcs=Vdd) condition.

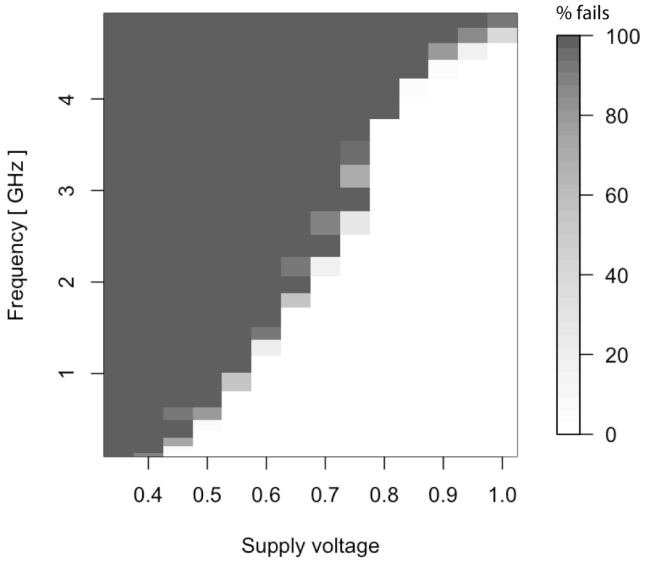

Fig. 11 Frequency vs. supply voltage schmoo diagram for array clocks in the half cycle override mode operation. The SRAM is operating under the one-supply (Vcs=Vdd) condition.

A measurement example, shown in Fig. 12, is given to show the effectiveness of adaptive clock control. As the delay setting is changed from (01) to (10) in the programmable block (PD shown in Fig. 4, or two of the right blocks shown in Fig. 5 in series) for the pulse mode or half cycle clock mode, settings with better edge alignments contribute to the lowering of minimum operating voltage (Vmin) and performance. The better aligned PD=10 setting results in significant uplifts of the operating frequency, most noticeably for voltages below 0.7V. As shown here also, the pulsed clock mode is the preferred mode of

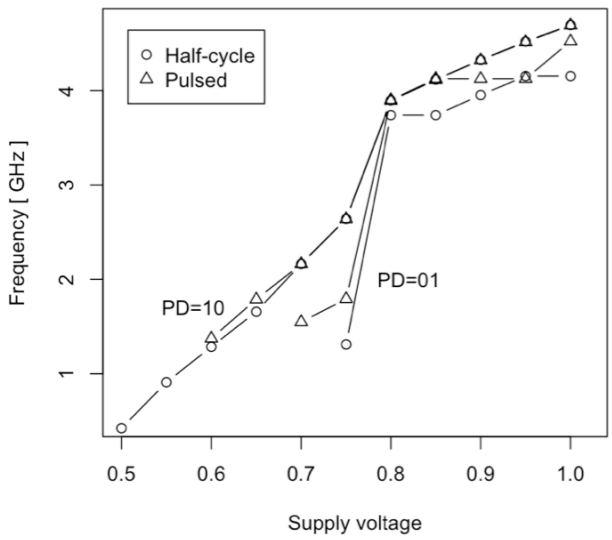

Fig. 12 Plot of delay intervention on functionality and performance for the single-supply pulse and half-cycle mode. The PD (as in Fig. 3) adjustment capability improves the operating frequency and delays the onset of functional failures.

operation for supply voltages higher than 0.7V. Below 0.7V, we switch to the half cycle clock mode to maximize the evaluation pulse windows inside the array.

Fig. 13 shows the measured read, write, and leakage power for one selected and active array segment together with the support circuit. The read or write power is taken under the combined condition of

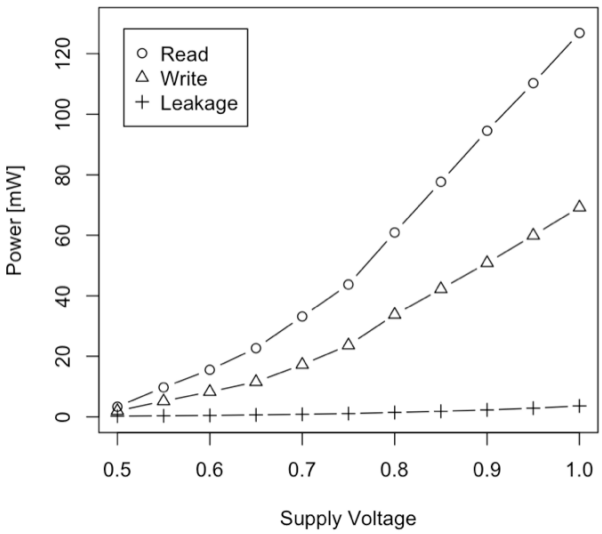

Fig. 13 Measured read, write, and leakage power vs. supply voltage for an selected and active array segment at the highest activity data pattern and maximum operating frequency corresponding to each supply voltage. The read power data was taken under the worst case condition when both ports were active. The SRAM is put in the one-supply (Vcs=Vdd) mode.

highest activity data pattern and maximum operating frequency at each supply voltage. For the read measurements, both ports are active and simultaneously accessed. The two-ported read power is approximately 1.8 times of the write power over the voltage range. At 0.5V, we observe a read power of 3.33mW and a write power of 1.97mW, corresponding to a 38x and 35x reduction in active power measured at 1V, respectively. Correspondingly, a 15x reduction in overall leakage power is seen in our measurements.

IV. Conclusion

A high-performance dual read port 32nm 6T SRAM with one clock cycle access time and 8-way set associativity for low voltage applications is presented. We demonstrate the feasibility of reliable low-power split word line 6T SRAM operation of 348MHz and 1.2GHz at the 0.5V and 0.6V supply, respectively, by using a 1V. PD SOI technology where both SRAM and logic devices are optimized for nominally 1V operations. Fully adaptive clock delay and pulse width control in conjunction with the choice of low voltage friendly circuit topologies have been key to successfully achieving this very challenging design target and enabling robust SRAM designs in the 1.5-2Vt power supply regime.

V. Acknowledgment

The authors thank Emmanuel Crabbé, Rolf Sautter, Antje Müller, Carl Radens, Donald Plass, Tuyet Nguyen, Bryan Robbins, Bryan Lloyd, and Christopher Durham for technology development, wafer fabrication, physical design, testability verification, and helpful discussions.

VI. References

[1] J. Davis, et al., "A 5.6GHz 64KB dual-read data cache for the POWER6 processor," ISSCC 2006.

[2] J. Pille, et al., "A 32KB 2R/1W L1 data cache in 45nm SOI technology for the POWER7 processor," ISSCC Dig. Tech. Papers 2010, pp. 344-345.

[3] L. Chang ,et al., "A 5.3GHz 8T-SRAM with operation down to 0.41V in 65nm CMOS," Dig. Symp. VLSI Circuits 2007, pp. 252-253.

[4] A. Raychowdhury, et al., "PVT and aging adaptive wordline boosting for 8T SRAM power reduction," ISSCC Dig. Tech. Papers 2010, pp. 352-353.

[5] K. Nii, et al., "A 0.5V 100MHz PD-SOI SRAM with enhanced read stability and write margin by asymmetric MOSFET and forward body bias," ISSCC Dig. Tech. Papers 2010, pp. 356-357.

[6] H. Pilo, et al., "An SRAM design in 65-nm technology node featuring read and write-assist circuits to expand operating voltage," IEEE J. Solid-State Circuits, vol. 42, no. 4, pp. 813-819, 2007.

[7] J. Kuang, et al., "A 32nm 0.5V-supply dual-read 6T SRAM," IEEE Custom Integrated Circuits Symposium, 2010, pp. 17-20.

[8] B. Greene, et al., "High performance 32nm SOI CMOS with high-k/metal gate and 0.149mm2 SRAM and ultra low-k back end with eleven levels of copper," Symp. VLSI Technology Dig. Tech. Papers, pp. 140-141, 2009.

[9] 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 2009, pp. 458-459.

[10] H. Ngo, J. Kuang, J. Warnock, and D. Wendel, "Scannable dynamic logic latch circuit, " US Patent 7372305, 2008.

[11] J. Warnock, et. al., "POWER7™ local clocking and clocked storage elements," ISSCC Dig. Tech. Papers 2010.

[12] D. Wendel, et al., "The implementation of POWER7™, a highly parallel, scalable multi-core high end server processor," ISSCC Dig. Tech. Papers 2010.