2nd Generation Embedded DRAM with 4X Lower Self Refresh Power in 22nm Tri-Gate CMOS Technology

Mesut Meterelliyoz[1], Fuad H. Al-amoody[1], Umut Arslan[1], Fatih Hamzaoglu[1], Luke Hood[2], Manoj Lal[2], Jeffrey L. Miller[2], Anand Ramasundar[2], Dan Soltman[1], Ifar Wan[2], Yih Wang[1], Kevin Zhang[1]

[1]Logic Technology Development, Intel Corporation, Hillsboro, Oregon, USA

[2]Platform Engineering Group, Intel Corporation, Hillsboro, Oregon, USA

Tel: (971) 214-5641, Fax: (971) 214-7810, Email: mesut.meterelliyoz@intel.com

Abstract

2nd generation 1Gbit 2GHz Embedded DRAM (eDRAM) with 4X lower self refresh power compared to prior generation is developed in 22nm Tri-Gate CMOS technology. Retention time has been improved by 3X (300us@95°C) by process and design optimizations. Source synchronous clocking is integrated in the design to reduce clock power without penalizing bandwidth. Charge pump power is reduced by 4X by employing comparator based regulation. Temperature controlled refresh enables minimum refresh power at all temperature conditions.

Introduction

As processors became more compute intensive and memory bandwidth limited with CMOS Technology scaling, the demand for high bandwidth, high performance and dense embedded memories has increased significantly. eDRAM has been widely accepted as a design solution due to its dense, high bandwidth and logic compatibility features [1-4]. Improving battery life while meeting the performance and bandwidth requirements exposes new challenges especially for ever-growing power sensitive markets. This paper presents the 2nd-Gen 1Gbit 2GHz eDRAM with 4X lower self refresh power and 3X better retention time compared to the 1st-Gen eDRAM [4-6] at the same temperature.

Source Synchronous Clocking (SSC)

SSC is one of the key power reduction techniques implemented in this design. Fig. 1 shows 1st-Gen eDRAM [4] employing a horizontal clock tree that was shared between two dataslices. Data, address and command signals (read/write/refresh) propagate in and out of dataslices in three clock cycles to be able to synchronize operations between banks. This three-cycle operation enabled write/read every other cycle to different banks. In the 2nd-Gen eDRAM described by this paper, the horizontal clock tree is replaced with Request and Output clocks which are used for write/refresh and read operations, respectively. To maintain write every other cycle bandwidth capability, write data, address and command signals are routed to the entire dataslice in one-cycle with respect to Request clock. This also permits better utilization of metal tracks. During read operation, each datablock is controlled with designated output enable signals in order to sustain read every other cycle bandwidth with no data collision. SSC enables not only reduction in clock power by removing horizontal clock tree and activating only the selected dataslice's Request/Output clocks but also reduction in dynamic toggling power by gating the write data and address bits depending on the bank that is being accessed.

Comparator-Based Charge Pump (CP) Regulation

As shown in Fig. 2, VCCWL>VCC and VSSWL<VSS voltages are generated by positive and negative CPs, respectively and are used to drive eDRAM bitcell access transistor. This enables fast sensing and improved retention time. Compared to [4-6] where the outputs of CPs were regulated using shunt regulators, this design uses comparator based regulation for both CPs. By turning on the CPs only when VCCWL/VSSWL voltages drop below a programmable reference voltage, significant power reduction is achieved.

Random Cycle Time (RCT) Optimization

Fig. 3 shows that compared to [4], RCT has been increased from 3ns (6 cycles) to 5ns (10 cycles) by adding 1 cycle to each sense, write back, wordline_off and pre-charge operations. This design change enabled better optimization of periphery circuit for power (CPs, Half-VCC generators for pre-charge) as well as improved retention time by allowing better optimization of transistor power/performance.

Temperature Controlled Active and Self Refresh

Temperature controlled active and self refresh architecture is shown in Fig. 4. Compared to active refresh, self refresh is performed at a lower clock frequency to reduce clock power. Each refresh mode has 4 retention time buckets that can be programmed and selected independently based on an on-chip Digital Temperature Sensor (DTS) output. By dynamically controlling the refresh rate, refresh power is minimized at different temperature conditions.

Silicon Results

Fig. 5 compares the CP power between the comparator and the shunt-regulator based designs. At 85°C, CP power is reduced by 4X by enabling the CPs only when needed. Fig. 6 shows that active refresh power can be reduced by 40% at the same temperature/retention time by refreshing at a lower clock frequency in self refresh. Comparison of self refresh power with 1st-Gen eDRAM [4] is demonstrated in Fig. 7. At the same temperature, 95°C, self refresh power is reduced by 4X as a cumulative result of clock power reduction, CP and periphery power reduction and retention time improvement from 100µs to 300µs. Fig. 7 also shows that at lower temperatures, self refresh power is further improved by 20% and 48% at 85°C/400µs and 50°C/600µs, respectively, as a result of relaxed retention time and lower transistor leakage and CP power. Fig. 8 shows the improved bit failure rate in this design compared to [4] at the same retention time and temperature. It also shows that this design achieves comparable fail rates as [4] at 3X longer retention time. Fig. 9 shows the die photo and feature comparison with [4].

Conclusions

This paper demonstrates the design techniques employed to achieve a low power eDRAM in 22nm Tri-Gate CMOS technology. 4X lower self refresh power is demonstrated by reducing clock power, CP power and 3X improved retention time. Temperature controlled refresh rate enables even further reduction in self refresh power at lower temperatures.

References

[1] J. Barth et al., ISSCC, pp. 342-344, Feb. 2010.

[2] K. Hijioka et al., IEDM Tech. Dig., pp. 756-759, Dec. 2010.

[3] S. Romanovsky et al., ISSCC, pp. 270-271, Feb. 2008.

[4] F. Hamzaoglu et al., ISSCC, Feb. 2014.
[5] R. Brain et al., VLSI Tech. Symp., June 2013.
[6] Y. Wang et al., IEDM Tech. Dig., Dec. 2013.

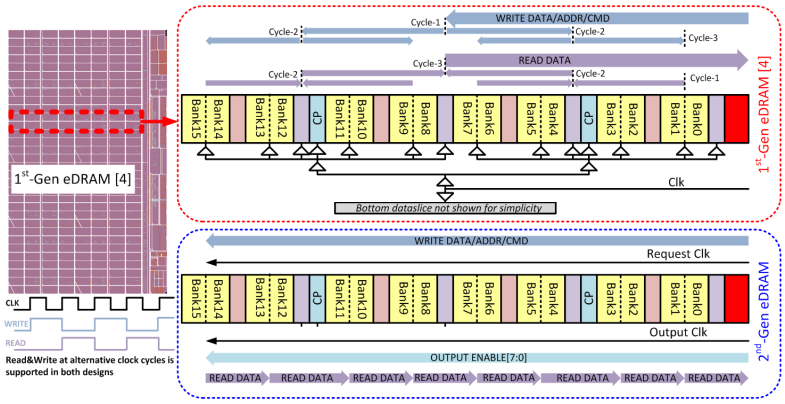


Fig. 1: Source synchronous clocking used in 2nd-Gen eDRAM vs. 1st-Gen eDRAM

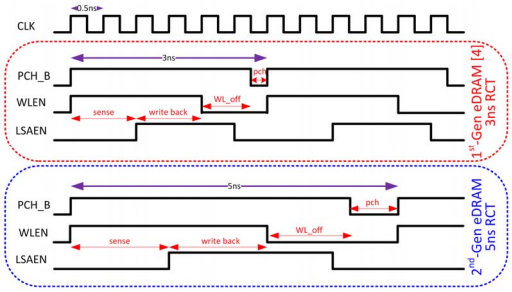


Fig. 3: Timing differences between 2nd-Gen eDRAM vs. 1st-Gen eDRAM

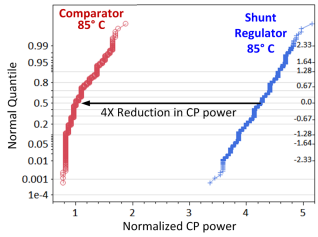


Fig. 5: Power benefit of comparator based CP design

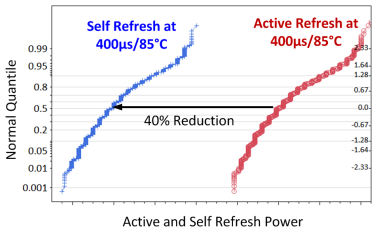


Fig. 6: Active Refresh vs. Self Refresh power at 400µs/85°C

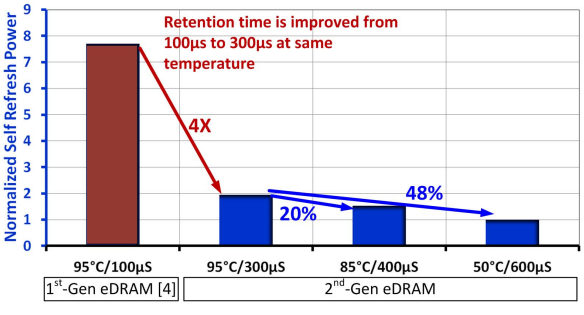


Fig. 7: Comparison of self refresh power with 1st-Gen eDRAM [4]. Self refresh power at three temperature/retention times

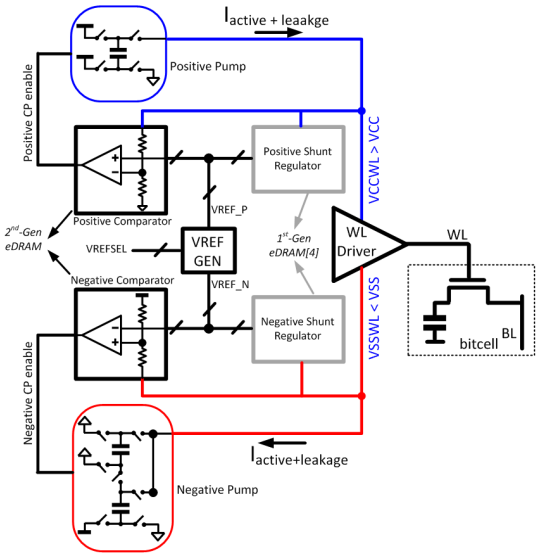


Fig. 2: Comparator based regulation in CPs in 2nd-Gen eDRAM vs. shunt regulation in 1st-Gen eDRAM

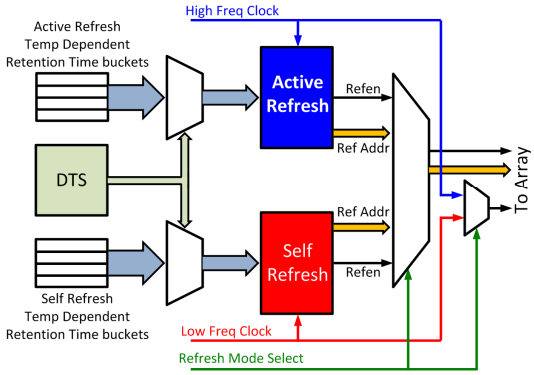


Fig. 4: DTS controlled active and self refresh

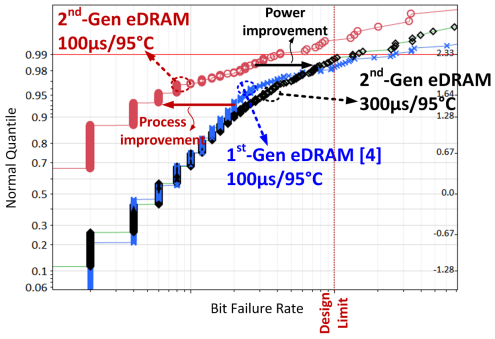
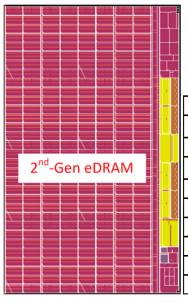


Fig. 8: Bit Failure Rate comparison between 2nd-Gen eDRAM and 1st-Gen eDRAM [4]. 2nd-Gen eDRAM achieves comparable fail rate at 3X higher retention time



Technology	22nm Tri-gate CMOS	
Cell Size	0.029µm²	
Supply	1.05V	
	1st-Gen eDRAM [4]	2nd-Gen eDRAM
Clock, Random Cycle Time	2GHz, 3ns	2GHz, 5ns
Retention Time	100µs @ 95C	300µs @ 95C
Normalized Self Refresh Power	1	1/4
CP Regulation	Shunt Regulator	Comparator
Clocking Scheme	Clock Tree	Source Syncronous
Temperature Controlled Refresh	NO	YES

Fig. 9: Die photo and 2nd-Gen vs. 1st-Gen eDRAM [4] design features