5.7 A Graphics Execution Core in 22nm CMOS Featuring Adaptive Clocking, Selective Boosting and State-Retentive Sleep

Carlos Tokunaga, Joseph F. Ryan, Charles Augustine, Jaydeep P. Kulkarni, Yi-Chun Shih, Stephen T. Kim, Rinkle Jain, Keith Bowman, Arijit Raychowdhury, Muhammad M. Khellah, James W. Tschanz, Vivek De

Intel, Hillsboro, OR

The demand for high-performance graphics capability even in extremely power-constrained platforms such as smartphones and tablets requires circuit techniques that scale from efficient operation at low voltage to high performance when needed. It is well known that energy efficiency improves as supply voltage is scaled down, reaching a maximum near the device threshold voltage where switching energy savings from voltage reduction is balanced by increased leakage energy from frequency loss. Achieving this voltage reduction, however, requires techniques that address intrinsic V_{MIN} limitations in arrays (SRAM, register file arrays, ROMs), voltage droop guardband reduction in logic, as well as techniques for reducing leakage energy, which can dominate at low voltage. It is important that these techniques, while providing energy-efficient operation at low voltage, do not impact the high-performance mode, which is also critical for graphics workloads.

In this paper, we present a low-power graphics processing core that achieves a 40% improvement in peak energy efficiency using dual-V_{CC} arrays, adaptive clocking for voltage droop mitigation, and state retention capability with an integrated retention clamping circuit for low-power sleep mode. The 22nm testchip (Fig. 5.7.1) includes a graphics execution core [1] connected to an SRAM array and test controller used for storage and delivery of at-speed test vectors. Correct execution of the tests is validated through a multiple-input signature register (MISR), which accumulates key signals in the core and generates a 32b signature at test completion.

To mitigate the impact of high-frequency voltage droops, an adaptive clocking technique is implemented to proactively gate or divide the core clock when a droop is detected. This clocking technique (Fig. 5.7.1) consists of an adaptive clock distribution (ACD) block [2], a tunable replica circuit (TRC) at the root of the clock distribution for timing margin detection, and a clock gate/divider to locally gate or divide the clock frequency by two when a droop is detected by the TRC. The ACD extends the clock-data delay compensation effect [3] that naturally occurs during a droop by inserting additional, programmable delay in the clock distribution. The clock-data compensation effect, in which the critical-path slowdown is compensated by clock-period stretching that occurs during the onset of the droop, allows core paths to continue to operate correctly for several cycles while the adaptive clock circuits detect the droop and initiate the clock response. In this work we enhanced the clock response by implementing the clock division mode, which – compared to gating the clock – reduces throughput loss when a droop occurs and also reduces transients on the V_{CC} grid when the clock is ungated. The adaptive clock control maintains the clock response for a programmable time period before returning to full-frequency mode.

Embedded inside the execution core are a 32KB graphics register file (GRF) and a 6KB ROM array used for extended math operations. These arrays typically limit the minimum operating voltage (V_{MIN}) and hence the energy efficiency. As an alternative to sizing up the bitcells in the GRF and ROM to reduce impact of variations on intrinsic V_{MIN}, the dual-V_{CC} approach [4] instead selectively supplies key V_{MIN}-limiting nodes in the arrays with a higher voltage (V_{BOOST}), available as a second supply in the core and routed as a sparse grid for minimal impact on the power and signal routing (Fig. 5.7.2). The power overhead of this selective boosting technique is minimal because the second V_{BOOST} supply does not power the entire array. In read mode, the GRF RWL is boosted to compensate for the impact of V_t variations in the stacked NMOS read port and/or the PMOS keeper on the local Read BL. In write mode, the WWL is boosted to mitigate contention between the bitcell NMOS pass and PMOS pull-up devices. The ROM is also implemented using the dual-V_{CC} approach, where both the selected RWL and column mux input are boosted using embedded dynamic, level-shifting drivers. Active leakage can be mitigated through the use of fast power gating; however this requires local storage for key state nodes to eliminate long latencies for saving and restoring context to always-on storage. For retention of GRF contents during sleep mode, a V_{CC} mux – implemented per local GRF column – allows all GRF bitcells to be connected to V_{BOOST} which acts as an always-on supply (Fig. 5.7.2). The bitcells are also disconnected from the WBLs. For storage of critical state distributed in the execution core, state-retention sequentials (Fig. 5.7.3) isolate the slave storage node during sleep and connect to an always-on "V_{RET}" grid. Power during sleep is further reduced through an all-digital, fully-synthesized active retention clamp design (Fig. 5.7.3) which gradually transitions the voltage of the GRF bitcells and state-retention sequentials to a pre-set retention voltage that guarantees correct retention for the worst-case state element on the die. A hysteretic control maintains the V_{RET} between two reference inputs VrefLow/High, and implements a low-power voltage-to-time converter using a ring oscillator, which is time-multiplexed to reduce leakage power and eliminate variation-induced offset. The clamp is designed to support an extremely wide range of output current on the retention grid, covering more than three orders of magnitude to guarantee operation across process skew, voltage, and temperature.

The 3.38mm^2 testchip (Fig. 5.7.7) is fabricated in a 22nm, tri-gate SoC technology [5] and is validated with test sequences ported from pre-silicon validation. The ability to save context locally in retention flops and in the GRF allows fast power gating for active leakage reduction, showing 8× power reduction compared to clock gating only (Fig. 5.7.4). Further reduction of sleep power is obtained by enabling the retention clamp, improving sleep power savings to 10×, including the overhead power for the clamp operation. Operation of the retention clamp across multiple skewed parts and temperatures demonstrates leakage savings from 4× to 20× while guaranteeing correct flip-flop retention.

Dual-V_{CC} capability allows the graphics core to operate across a wide voltage range (Fig. 5.7.5) by optimizing V_{BOOST} such that arrays do not limit V_{MIN} of the core logic. Here a slight word-line under-drive is used for the baseline case to model the higher V_{MIN} that would be observed with a larger sample size and down-sized bitcell. V_{MIN} improves up to 270mV when boost is employed allowing the core voltage to scale below 0.4V for a test dominated by the GRF. For a test that uses the ROM and GRF, the V_{MIN} is reduced up to 350mV. Failure rate data indicate that small amounts of boost can provide significant V_{MIN} reduction as the tail of the bitcell distribution is compensated with the higher boost voltage.

Adaptive clocking effectively compensates the frequency loss due to fast voltage droops as long as the ACD length is sufficient to provide clock-data compensation for the required response time. While both clock-gating and frequency-division modes achieve this goal, frequency-division demonstrates the best performance and is able to recover 90% of the frequency loss incurred by a 10% voltage droop (Fig. 5.7.6). By nearly eliminating this droop guardband, adaptive clocking improves power at high voltage by 12.4%. Dual-V_{CC} arrays extend the efficient operating range down to 0.38V, where 54% power savings are achieved at 100MHz. The combination of dual-V_{CC} and adaptive clocking improves energy efficiency up to 2.7× at low voltage, with peak energy efficiency gain of 40% GFLOPS/W.

Acknowledgements:
The authors thank K. Ikeda, L. Peake, Jijin T, A. Sandra, T.-H. Foo, L. Avery, C. Parsons, I. Mirza, D. Jenkins, and D. Finan for implementation, B. Matush, J. McCoskey, and W.C. Chee for graphics validation, T. Nguyen and P. Aseron for lab assistance, and R. Forand for encouragement and support. This research was, in part, funded by the U.S. Government (DARPA). 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] S. Damaraju, *et al.*, "A 22nm IA Multi-CPU and GPU System-on-Chip," *ISSCC Dig. Tech. Papers*, pp. 56-57, 2012.
[2] K. Bowman, *et al.*, "A 22nm Dynamically Adaptive Clock Distribution for Voltage Droop Tolerance," *IEEE Symp. VLSI Circuits*, pp. 94-95, 2012.
[3] D. Jiao, *et al.*, "A Programmable Adaptive Phase-Shifting PLL for Clock Data Compensation Under Resonant Supply Noise," *ISSCC Dig. Tech. Papers*, pp. 272-274, 2011.
[4] J. Kulkarni, *et al.*, "Dual-Vcc 8T bitcell SRAM Array in 22nm Tri-Gate CMOS for Energy-Efficient Operation Across Wide Dynamic Range," *IEEE Symp. VLSI Circuits*, pp. C126-C127, 2013.
[5] C.-H. Jan, *et al.*, "A 22nm SoC Platform Technology Featuring 3-D Tri-Gate and High-k/Metal Gate, Optimized for Ultra Low Power, High Performance and High Density SoC Applications," *IEDM Dig. Tech. Papers*, pp. 4-7, 2012.

 978-1-4799-0920-9/14/$31.00 ©2014 IEEE

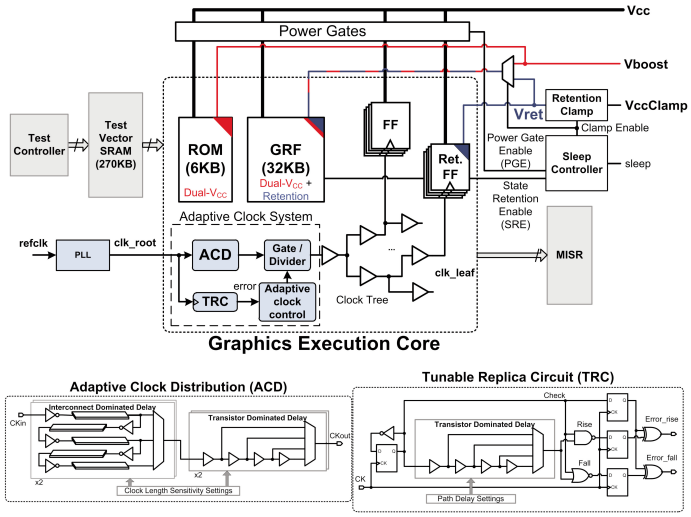


Figure 5.7.1: Block diagram of the low-voltage graphics execution core, and details of the adaptive clock distribution and tunable replica circuit.

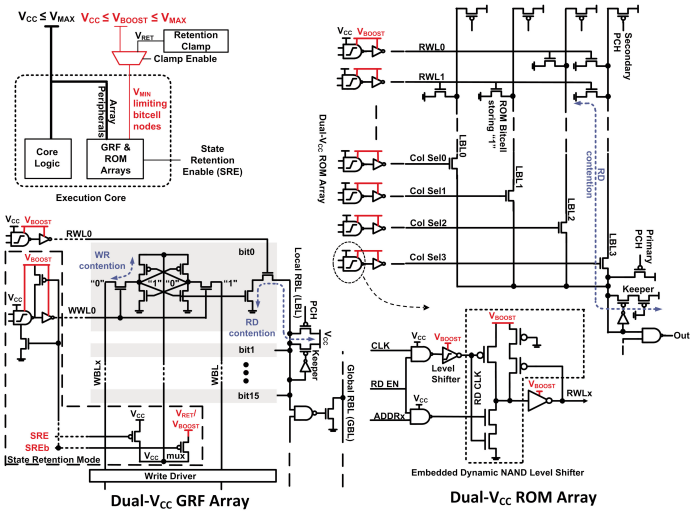


Figure 5.7.2: Dual-V_{cc} arrays design overview, dual-V_{cc} GRF with retention mode, and dual-V_{cc} ROM.

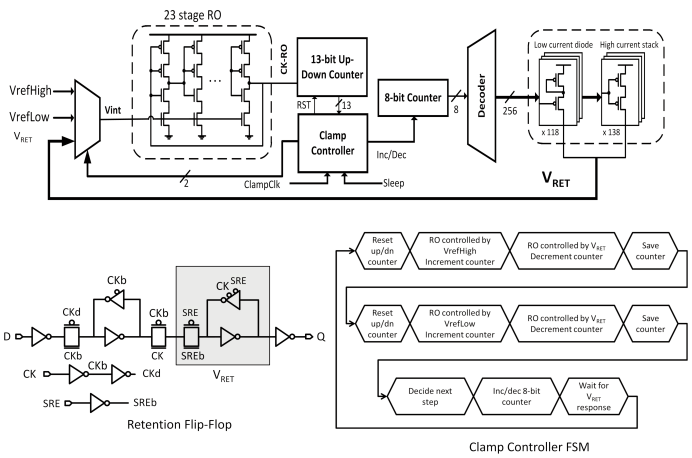
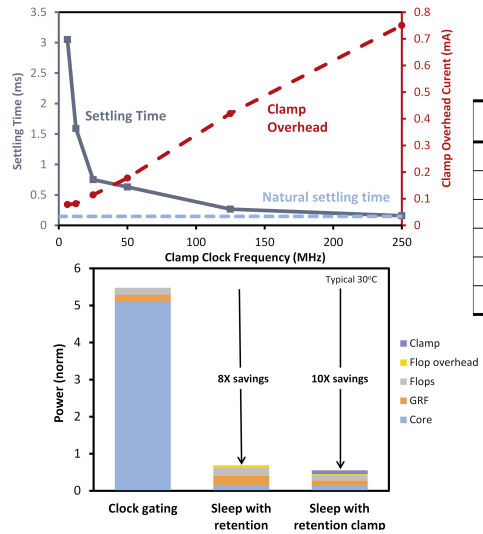


Figure 5.7.3: Overview of the retention clamp design, timing sequence for the clamp controller, and retention flip-flop circuit.



Skew/Temp	Leakage Savings
Typical 30°C	10x
Typical 90°C	4x
Slow 30°C	20x
Slow 90°C	10x
Fast 30°C	10x
Fast 90°C	4x

Figure 5.7.4: Measured standby power savings with retention FFs, GRF and retention clamp. Measured clamp settling time and overhead.

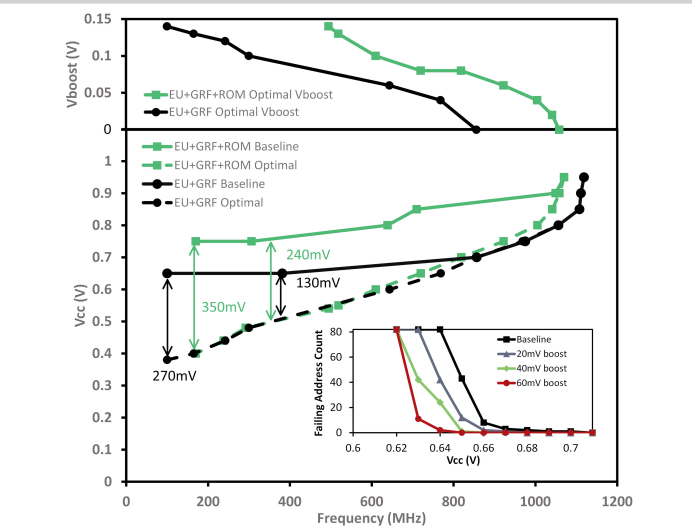
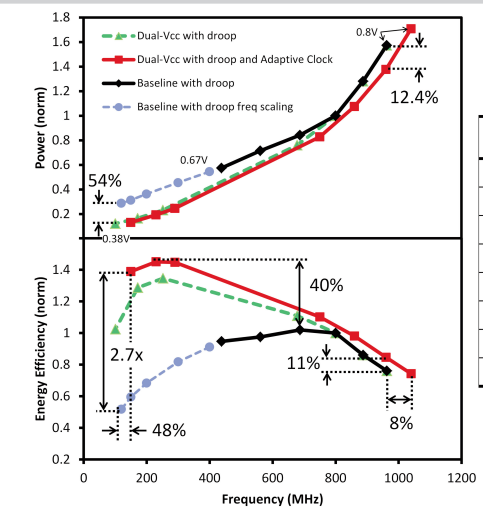
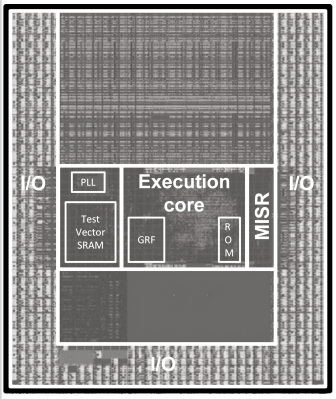


Figure 5.7.5: Impact of selective dual-V_{cc} boost of GRF and ROM on VMin. Read failure rate, obtained with BIST at 600MHz.



ACD Length (clock cycles)	F_{MAX} Recovery (%)
1	44
2	61
3	77
4	90
5	90
6	90
7	90
8	90

Figure 5.7.6: Measured power and energy efficiency for baseline design, and dual-V_{cc} design with adaptive clocking. F_{MAX} recovery for different adaptive clock distribution lengths.



Technology	22nm, 9-metal layer tri-gate high-K/MG CMOS
Area: testchip die	4.0 x 5.8 mm^2
Area: core + test	2.6 x 1.3 mm^2
Core transistor count	22.8M
Target voltage, frequency	0.7V, 800MHz
Retention sequential count	14,411
Package	FCBGA13 951

Figure 5.7.7: Testchip die micrograph and design details.

978-1-4799-0920-9/14/$31.00 ©2014 IEEE