2.3 An Energy-Efficient Graphics Processor Featuring Fine-Grain DVFS with Integrated Voltage Regulators, Execution-Unit Turbo, and Retentive Sleep in 14nm Tri-Gate CMOS

Pascal Meinerzhagen[1], Carlos Tokunaga[1], Andres Malavasi[1], Vaibhav Vaidya[1], Ashwin Mendon[1], Deepak Mathaikutty[1], Jaydeep Kulkarni[1], Charles Augustine[1], Minki Cho[1], Stephen Kim[1], George Matthew[1], Rinkle Jain[1], Joseph Ryan[1], Chung-Ching Peng[1], Somnath Paul[1], Sriram Vangal[1], Brando Perez Esparza[1], Luis Cuellar[1], Michael Woodman[1], Bala Iyer[1], Subramaniam Maiyuran[2], Gautham Chinya[1], Chris Zou[1], Yuyun Liao[1], Krishnan Ravichandran[1], Hong Wang[1], Muhammad Khellah[1], James Tschanz[1], Vivek De[1]

[1]Intel, Hillsboro, OR; [2]Intel, Folsom, CA

Graphics workloads are highly dynamic in nature, using multi-threaded SIMD execution units (EUs), fixed-function units, samplers, and media accelerators to provide ever-increasing amounts of graphics performance. These workloads are often limited by power and thermal constraints, requiring dynamic voltage/frequency scaling (DVFS) of the graphics processor (GPU). This coarse-grain DVFS, driven by a power-management IC (PMIC) setting a shared rail voltage (V_{IN}), incurs performance loss while waiting for PLL re-lock and slow-rail voltage transitions. In addition, it does not allow a performance-critical unit (e.g. an EU) to use on demand a higher V/F (e.g. for EU turbo) without an energy penalty for the rest of the GPU.

In this paper, we present a GPU in 14nm tri-gate CMOS featuring: (1) fine-grain DVFS where the key power/performance critical blocks for 3D graphics processing, i.e., the EUs, are powered by a digitally controlled integrated voltage regulator (IVR) for fast on-die voltage (V_{OUT}) control, (2) retentive sleep, and (3) reliability-aware wake-up, all implemented using distributed power gates (PG). The GPU responds quickly to instantaneous workload demands by boosting performance of only the bandwidth-critical units while reducing V/F of the non-critical units. In particular, we demonstrate EU turbo as an example of fine-grain DVFS in a GPU for improving performance and energy efficiency of EU-dominant workloads. During EU-intensive periods, EUs are set to a high voltage V_{HIGH} (V_{IN}=V_{HIGH}) and operate at a readily available 2× clock frequency, while other blocks operate at a low voltage V_{LOW} using IVRs (V_{OUT}=V_{LOW}) and the 1× clock. The IVR supports a wide range of load currents and operating voltages, including near-threshold voltage (NTV). The embedded graphics register file (GRF) and ROM arrays use voltage boosting from V_{IN} as an assist technique to achieve NTV operation. The IVRs are dynamically configured to clamp V_{OUT} to the retention limit (V_{RET}) during short stall periods. Digitally controlled and programmable underdrive of selected PGs is utilized to aid IVR regulation, retention clamping, and reliability-aware wake-up.

The system prototype (Figs. 2.3.1 and 2.3.7) features a Gen9LP GPU design [1] containing 3 sub-slices (SS) of 6 EUs each. Two of these SS modules (SSM) feature the IVR, retention clamp, turbo support, and array V_{MIN} reduction techniques, while the remaining one is left unmodified as a reference. A system agent (SA) serves as an interface to feed the GPU with high-bandwidth data for 3D graphics workloads, and to handle GPU configuration, boot, and power management. To enable accurate power/performance measurement for key graphics workloads, the SA includes a 4MB paging cache and a controller that communicates to a host PC through an FPGA and PCI interface. The testchip also includes multiple JTAG scan chains for configuration, debug and observability of the different implemented techniques. EU turbo operation at 2× clock (Fig. 2.3.2) is enabled by EU logic modifications as well as a glitch-less clock divider/mux and synchronization logic to enable data transfer to/from the 1× clock domain. A power/turbo controller sequences the V_{OUT} and clock frequency changes to efficiently boost EU performance. Voltage level shifters at EU boundaries enable a seamless interface to other GPU blocks and also allow the IVR to compensate for process/temperature-induced EU-EU V_{MIN} variations. Read and write word-lines in the 32KB GRF and ROM are boosted using the IVR input voltage (V_{IN}) for EU V_{MIN} improvement.

A digitally controlled hybrid DLDO/SCVR [2] IVR offers high conversion efficiency over a wide V_{OUT} range. Fully distributed designs of both DLDO and SCVR (Fig. 2.3.3) are implemented to reduce demands for low-resistance metal resources in the thick upper layers, while maintaining low IR drop. The DLDO is designed for low V_{OUT} ripple and to meet the PG transistor self-heating and EM reliability constraints across a large range of load currents and V_{OUT} values, especially under low-load and large dropout conditions [3]. The DLDO uses an under-drive voltage (V_{UD}) for the two-way stacked primary PG (PPG) to limit PG current density. V_{UD} is generated by an R-2R DAC in the central IVR controller. The parallel secondary PGs (SPGs) are turned on by the DLDO controller only if the PPGs are fully utilized at high load or in bypass mode. The 1,420 PPG + SPG units are distributed in a checker-board pattern across the EU. A central spine of PG drivers controls the PGs at half-row granularity. The PG drivers form a distributed shift register (SR), receiving increment/decrement instructions from a central IVR controller. In order to minimize the DLDO control power overhead, the SR clock is gated at the fine granularity of PG drivers, as opposed to at only 4 clock sections [4]. The DLDO uses a linear fine-grain control loop for steady-state conditions, and a non-linear coarse-grain control for fast droop mitigation. Fast droop mitigation is achieved by sending an asynchronous preset signal to all flip-flops in the SR, which quickly turns on all PPGs, and optionally all SPGs as well. The DLDO can be dynamically configured as a retention clamp, where it sets V_{UD} such that V_{OUT} does not fall below the retention voltage (V_{RET}). The R-2R DAC is also used to generate a programmable V_{UD} ramp for self-heating and EM compliant EU wake-up.

The SCVR consists of 6 distributed power tiles (Fig. 2.3.3). 40% of on-die high-density MIM caps on top of the EUs are used to implement the fly capacitors. It operates across a 0.3-to-0.7V V_{OUT} range, and automatically transitions between 3:2, 2:1, and 3:1 V_{IN}/V_{OUT} ratios using the same two fly caps. The SCVR uses a fast digital controller for regulation and droop mitigation, and a slow controller for mode transitions, ripple management, and efficiency tracking. IVR reference voltage generator and voltage comparator circuits are shared between the DLDO and SCVR.

The measured current efficiency of the DLDO, when running an active workload, ranges from 93% to 95% for 0.785-to-1.11V V_{OUT} with 1.15V V_{IN}, at 25°C. The measured average power efficiency of the DLDO is only 5.5% below ideal over this operating range (Fig. 2.3.4). Under light load conditions, when EUs are idle, the measured DLDO power efficiency is 13.6% below ideal. The SCVR provides a 0.44-to-0.57V V_{OUT} range with a peak (average) power efficiency of 72% (69.8%) in the 2:1 mode and 58% (56%) in the 3:2 mode. With the DLDO configured in the open loop as retention clamp, V_{UD} values around V_{IN}-V_{TH} successfully clamp V_{OUT} close to V_{RET}≈0.5V, thus reducing EU leakage current by 62.5% for V_{IN}=1.15V, or 25% for V_{IN}=0.8V, during short stall periods.

Figure 2.3.5 demonstrates DLDO-enabled EU turbo with bypass-mode voltage boost (Fig. 2.3.6 oscilloscope capture) to improve GPU performance or reduce energy. The baseline GPU uses a PMIC for slow, coarse-grain DVFS from 0.51V/50MHz to 1.2V/400MHz for varying compute demands. Baseline measurements are taken from a modified SSM with EU turbo turned off. For a measured workload with 53.6% EU utilization, EU turbo offers performance improvements of up to 40%, with an average 36.6% improvement across the entire DVFS range. However, without independent IVRs for all major GPU blocks on the shared V_{IN} rail, the GPU energy/performance for EU turbo degrades significantly since V_{IN} is raised to support the 2× clock for the EUs in bypass mode. Therefore, all major blocks in the GPU must have independent IVRs to fully exploit the benefits of EU turbo. For 100% EU utilization and independent IVRs for all major GPU blocks, EU turbo enables up to 19% (average 17%) energy reduction at iso-performance. For lower EU utilizations, such as 80% or lower, EU turbo energy consumption can be worse than baseline, especially for high performance levels. This energy penalty is reduced by using a dual-rail system where only the EUs are IVR-enabled, while all other GPU blocks are powered by a separate PMIC-controlled external rail at V_{LOW}, in which case EU turbo enables up to 32% (average 29%) energy reduction at iso-performance for 100% EU utilization.

The oscilloscope captures of Fig. 2.3.6 demonstrate R-2R DAC-based reliability-aware EU wake-up enabled by a digitally controlled PPG V_{UD} ramp, as well as DLDO regulation of V_{OUT} to 0.7V, along with transition to bypass mode (V_{IN}=0.97V) triggered by an EU turbo request. In addition, wide-range hybrid DLDO/SCVR IVRs can be used to enable energy-efficient DVFS at the SoC level [3], where the GPU and CPU share the same V_{IN} rail (Fig. 2.3.6).

Acknowledgements:
The authors thank the many members of the Intel Labs circuit research, microarchitecture research, and silicon/system prototyping teams that contributed to this work, as well as the Intel Visual and Parallel Computing Group. 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] "The Compute Architecture of Intel Processor Graphics Gen 9," available online at: https://software.intel.com/sites/default/files/managed/c5/9a/The-Compute-Architecture-of-Intel-Processor-Graphics-Gen9-v1d0.pdf

[2] S. Kim, et al., "Enabling Wide Autonomous DVFS in a 22nm Graphics Execution Core Using a Digitally Controlled Hybrid LDO/Switched-Capacitor VR with Fast Droop Mitigation," *ISSCC*, pp. 154-155, 2015.

[3] R. Muthukaruppan, et al., "A Digitally Controlled Linear Regulator for Per-Core Wide-Range DVFS of Atom Cores in 14nm Tri-Gate CMOS Featuring Non-Linear Control, Adaptive Gain and Code Roaming," *ESSCIRC*, pp. 275-278, 2017.

[4] S. B. Nasir, et al., "A 0.13µm Fully Digital Low-Dropout Regulator with Adaptive Control and Reduced Dynamic Stability for Ultra-Wide Dynamic Range," *ISSCC,* pp. 98-99, 2015.

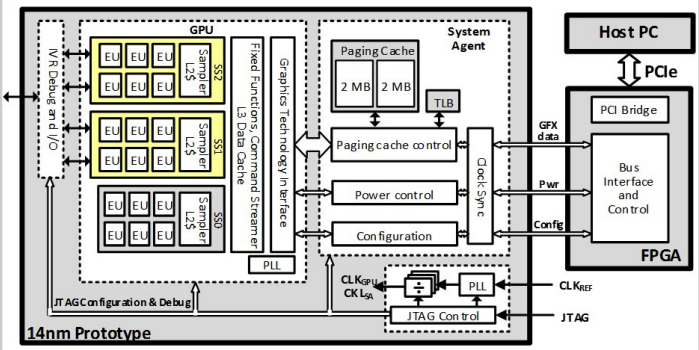


Figure 2.3.1: 14nm graphics processor system block diagram with FPGA for SoC emulation and host PC interface.

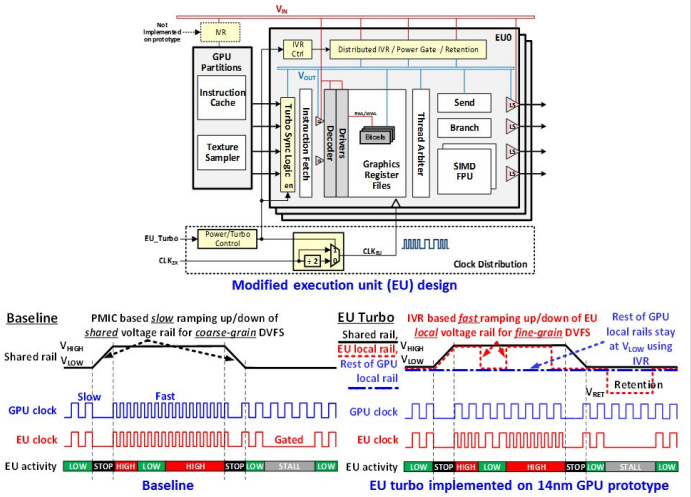


Figure 2.3.2: Modified execution unit (EU) design with IVRs (DLDO, SCVR), retention, GRF boost, and usage model of IVRs for fast, fine-grain, intra-GPU DVFS.

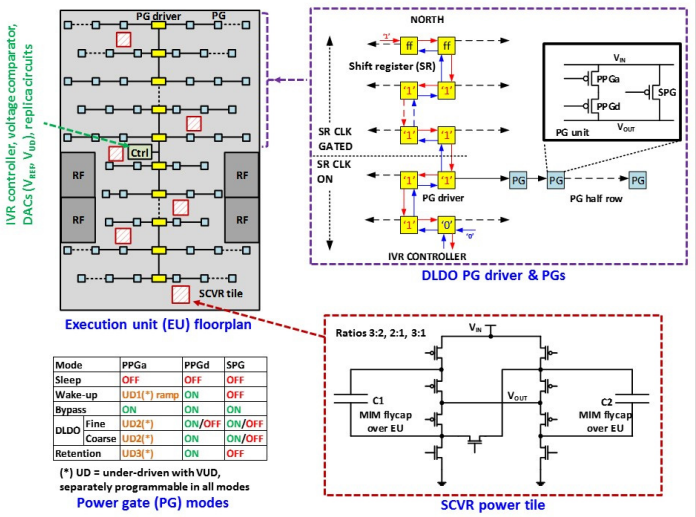


Figure 2.3.3: Floorplan of distributed IVRs in the modified EU, and distributed DLDO and SCVR design details.

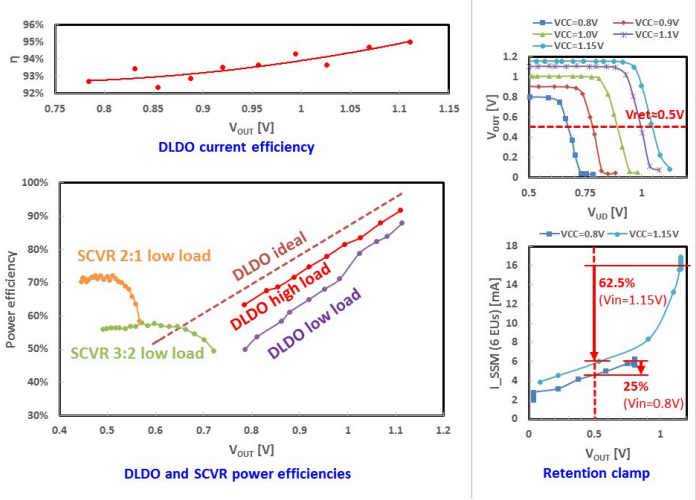


Figure 2.3.4: Measured DLDO and SCVR current and power efficiencies, and up to 62.5% power reduction enabled by retention clamp (open-loop DLDO with primary PG under-drive).

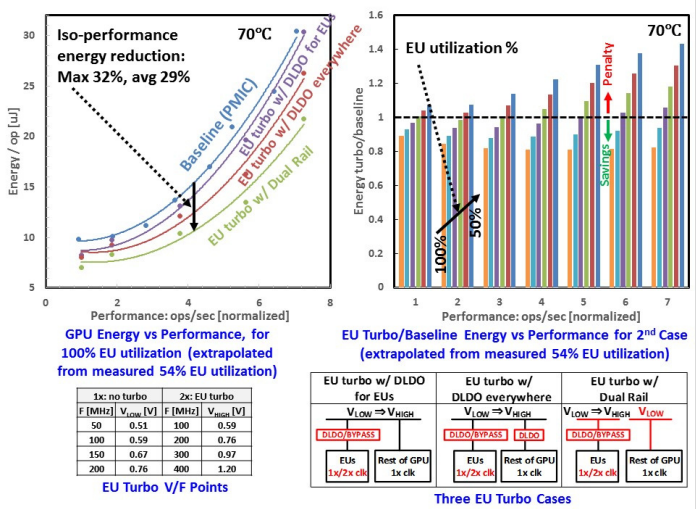


Figure 2.3.5: Measurement results for fine-grain, intra-GPU DVFS: EU turbo enables up to 32% energy reduction at iso-performance.

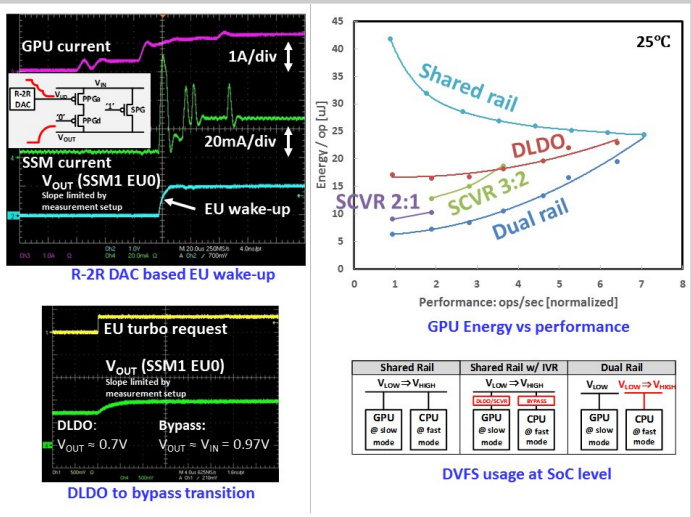
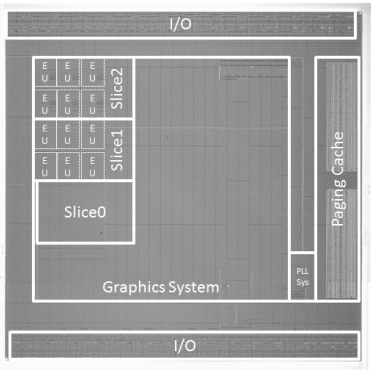


Figure 2.3.6: Oscilloscope captures showing R-2R DAC-based EU wake-up for PG transistor self-heating and EM compliance, DLDO operation, transition to bypass, and measurement results for SoC level use of IVRs (DLDO, SCVR).



Technology	14nm, 10-metal layer, tri-gate high-K/MG CMOS
Test-chip die area	8.0 x 8.0 mm^2
Transistor count	1.542 Billion
GPU V/F	0.51V, 50Mhz – 1.20V, 400MHz (turbo)
Package	FCBGA15951

Figure 2.3.7: Die micrograph and design details.

978-1-5090-4940-0/18/$31.00