10.2 A 28nm HPM Heterogeneous Multi-Core Mobile Application Processor with 2GHz Cores and Low-Power 1GHz Cores

Mitsuhiko Igarashi, Toshifumi Uemura, Ryo Mori, Noriaki Maeda, Hiroshi Kishibe, Midori Nagayama, Masaaki Taniguchi, Kohei Wakahara, Toshiharu Saito, Masaki Fujigaya, Kazuki Fukuoka, Koji Nii, Takeshi Kataoka, Toshihiro Hattori

Renesas Electronics, Tokyo, Japan

The worldwide demand for high-performance mobile or car infotainment application processors (AP) is increasing. This demand coexists with the need for low power to achieve long battery life and avoid thermal runaway. A heterogeneous CPU configuration is an effective solution. The proposed heterogeneous quad/octa-core AP has a combination of high-performance 2GHz cores and energy-efficient 1GHz cores. The maximum performance in the octa-core configuration is 35600 DMIPS. The key design highlights are: 1) Using a dedicated PLL and H-tree clock in the high-performance CPU achieves both 2GHz operation and reduced dynamic power. 2) A low-leakage SRAM in a 28nm HPM process is used and the leakage current of the peripheral circuits of the SRAM macro is optimized via multiple threshold voltages (V_t) and gate lengths (L_g). 3) The effects of process and voltage variations are accurately corrected by an on-chip process sensor and direct sensing of the voltage in the power mesh of the chip. 4) An enhanced CPU clock control mechanism is employed, which uses an on-chip delay sensor to reduce AC IR drop. 5) The heterogeneous CPU architecture maintains high performance even during thermal throttling.

Figure 10.2.1 shows the chip's features and a block diagram of the octa-core configuration. Fig. 10.2.7 is a micrograph of the die. The AP has two different types of CPU core: high-performance Cortex-A15 cores (CA15) operating at up to 2GHz, and energy-efficient Cortex-A7 cores (CA7) operating at up to 1GHz. It uses both types of CPUs simultaneously, whereas the conventional approach [1] is to switch between the two types of CPU according to the workload. The CA7 and CA15 have six power domains each, and each domain can be powered off by a power switch to reduce leakage power when idle. The CA7 is active and the CA15 is powered off when the workload is light, whereas both the CA7 and CA15 become active when the workload is heavy. Measured power results for the CA15 and CA7 are 5.4W and 0.6W, respectively, at the maximum frequency when all four cores in each block are operating.

Figure 10.2.2 shows the clock distribution and synchronizing scheme for the CA15. In general, a grid/mesh [2] or hybrid mesh [1] clock structure is used to reduce clock skew in high-frequency operation. However, it has been observed that clock power due to grid capacitance can comprise 21% of the overall chip power [2]. An H-tree clock structure was therefore used in the CA15 to reduce dynamic power. The H-tree structure reduces the wiring capacitance of the clock tree and improves the drivability of the clock buffer cells, compared with a grid/mesh clock structure. Although an H-tree clock structure does not minimize timing variations to the maximum extent possible, using a dedicated PLL to minimize clock latency and jitter for the CA15 achieved a clock latency of 0.9ns, and clock jitter of 30ps, enabling 2GHz operation. Since the dedicated PLL results in the interface with the CA15 becoming asynchronous, a synchronizer using a gray code was developed to retain throughput almost equal to that of a synchronous interface. This also limits the increase in latency relative to the bus clock to one cycle. The area overhead of applying DVFS was reduced by using a pseudo level-shifter with a single rail.

Figure 10.2.3 shows the features and performance of the SRAM for the L1 cache of the CA15. We used a low-leakage SRAM (LL-SRAM) in a 28nm HPM process to reduce leakage within the SRAM bitcells. Transistors having multiple V_t and multiple L_g are used to optimize the leakage current of the peripheral circuits. Long-L_g transistors are used in the word drivers and timing generators. A 24% leakage current reduction is realized compared to normal SRAM (GL-SRAM) with single V_t transistors, while achieving 2GHz operation. The L1 cache has 512 78b words, and consumes 0.00974 mm^2.

We used a modified form of adaptive voltage scaling (AVS). Schematic diagrams of conventional AVS, our modified AVS, and measured results for the latter are shown in Fig. 10.2.4. In conventional AVS [3], on-chip sensors are used to reduce the effect of process, voltage and temperature variations by optimizing supply voltages (V_{DD}). However, voltage control by the traditional method requires a voltage margin because the sensors are not perfectly accurate. The voltage resolution of the power management IC (PMIC) is also restricted to a certain value. In addition, if communication with the PMIC is through an I2C interface, the response time is slow compared to the switching frequency of the PMIC. This is an obstacle to dynamic fine-grained voltage control. Our modified form of AVS improves the accuracy of voltage control and the speed of feedback. Process variations are assessed via on-chip process sensors at the time of testing, and voltage settings for the PMIC are written to fuses. Thus, the impact of process variations is reduced statistically by a coarse adjustment of the PMIC's voltage. The measured results for minimum V_{DD} of at-speed test and voltage setting is shown in Fig. 10.2.4. The PMIC voltage of the fast-corner chip is set to a low V_{DD}, reducing worst-case dynamic and leakage power by 29% and 20%, respectively. The PMIC also reduces voltage variation by directly sensing the pin driving the power mesh, detecting variations, and then controlling V_{DD} finely and dynamically. The improvement in minimum V_{DD} during program execution was measured at around 40 to 50mV.

Operation of a high-performance CPU at 2GHz leads to a large di/dt that produces excessive AC IR drop. The sampling rate of a conventional power saver [4] is relatively slow at ~1μs, and thus it cannot detect an AC IR drop at several tens of MHz. Accordingly, we developed a real-time power saver mechanism with a 20× faster sampling rate. Fig. 10.2.5 shows its block diagram and results. An on-chip delay sensor is used, which samples voltage at 50ns intervals. If the sensor detects an IR drop that exceeds a threshold (e.g. due to a sudden increase of activity), a request for the clock controller to step down the frequency is issued to suppress the IR drop. The clock controller changes its frequency after a delay of 100ns. The frequency is incrementally increased several microseconds after a drop that exceeds the threshold. Simulations indicate that this approach achieves a 20mV reduction in the AC IR drop.

Mobile devices typically have a small form factor, making cooling difficult in an environment where an expensive cooling system is not possible. Moreover, the power consumed under a heavy workload by a high-performance CPU operating at 2GHz is significant and may lead to thermal runaway. A mobile AP thus requires a thermal control technique [5]. Fig. 10.2.6 shows the model we used to analyze thermal control, and simulation results under heavy workloads. We assume worst-case leakage conditions and lack of a special cooling system. Junction temperatures (T_j) increase while the CPU is running at full throttle, indicated by mode-A in the figure. When the on-chip temperature sensor detects a temperature exceeding a threshold, operating frequency is decreased and/or some CPU cores are powered off to decrease T_j. In a homogeneous CPU architecture, high-performance CPUs must continue to be used during the cool-down period, despite their large leakage and dynamic power. This dramatically reduces the average performance to only 3600 DMIPS. In contrast, with the heterogeneous CPU architecture, the energy-efficient CPUs operate while the chip is cooling down. The heterogeneous octa-core architecture maintains an average performance of 11000 DMIPS in worst-case conditions.

References:

[1] Y. Shin, *et al.*, "28nm High-κ Metal-Gate Heterogeneous Quad-Core CPUs for High-Performance and Energy-Efficient Mobile Application Processor", *ISSCC Dig. Tech. Papers*, pp. 154-155, 2013

[2] T. Singh, *et al.*, "Jaguar: A next-generation low-power x86-64 core", *ISSCC Dig. Tech. Papers,* pp. 52-53, 2013

[3] Y. Ikenaga, *et al.*, "A 27% active-power-reduced 40-nm CMOS multimedia SoC with adaptive voltage scaling using distributed universal delay lines", *IEEE Symp. VLSI Circuits*, pp. 186-187, 2011.

[4] M. Fujigaya, *et al.*, "A 28nm High-κ Metal-Gate Single-Chip Communications Processor with 1.5GHz Dual-Core Application Processor and LTE/HSPA+ Capable Baseband Processor", *ISSCC Dig. Tech. Papers*, pp. 156-157, 2013.

[5] S. Yang, *et al.*, "A 32nm High-κ Metal Gate Application Processor with GHz Multi-Core CPU", *ISSCC Dig. Tech. Papers*, pp. 214-216, 2012.

 978-1-4799-0920-9/14/$31.00

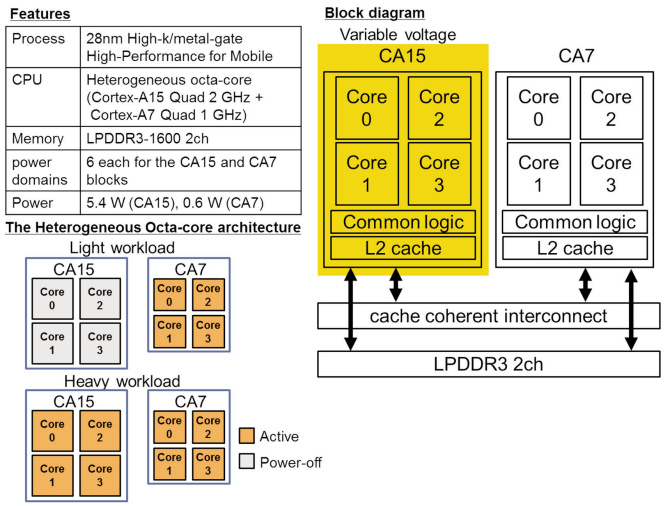

Figure 10.2.1: Features and block diagram of the octa-core CPU.

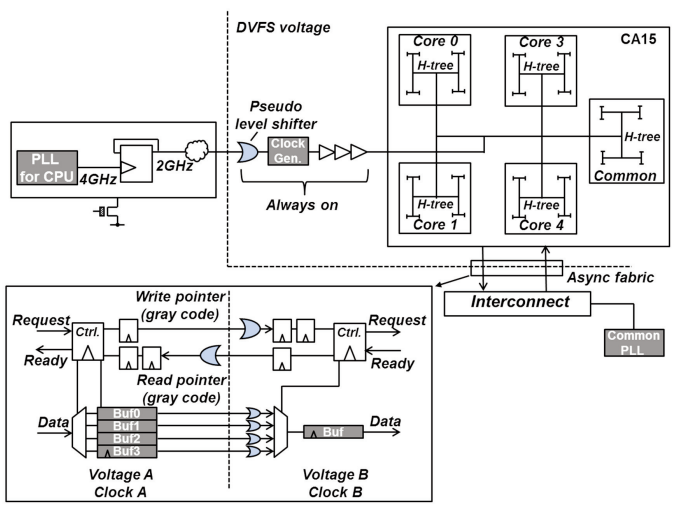

Figure 10.2.2: Clock distribution and synchronizing scheme of the CA15.

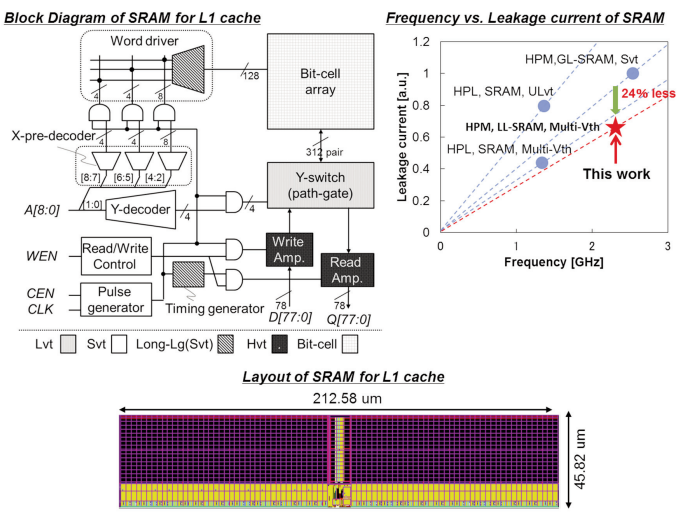

Figure 10.2.3: Feature and performance of SRAM for L1 cache of the CA15.

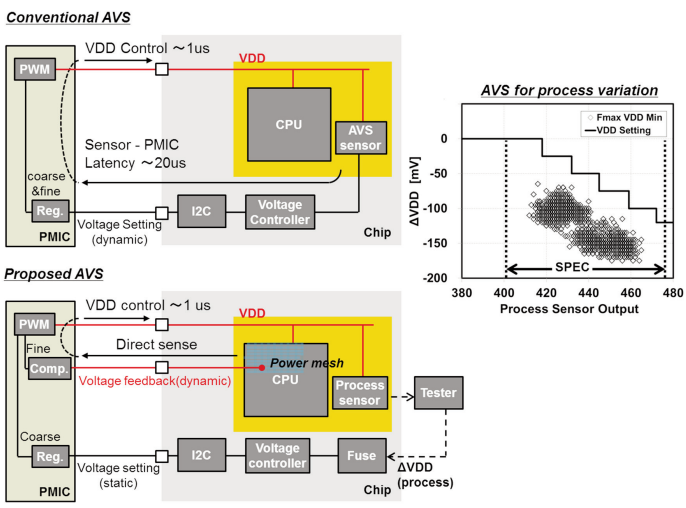

Figure 10.2.4: Schematic diagrams of AVS and measured results.

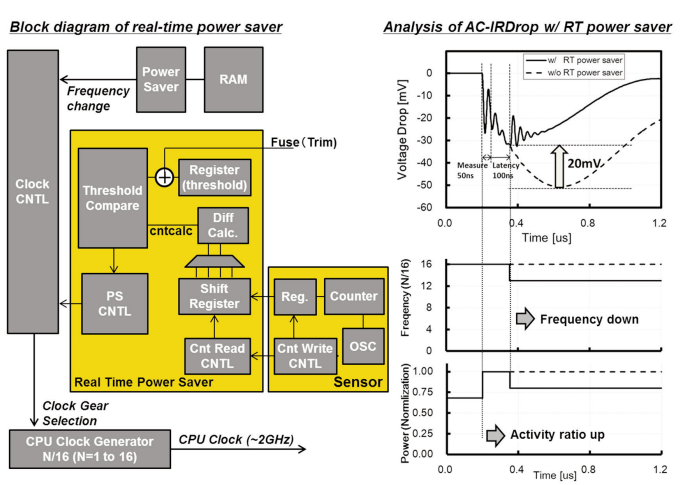

Figure 10.2.5: Real-time power saver.

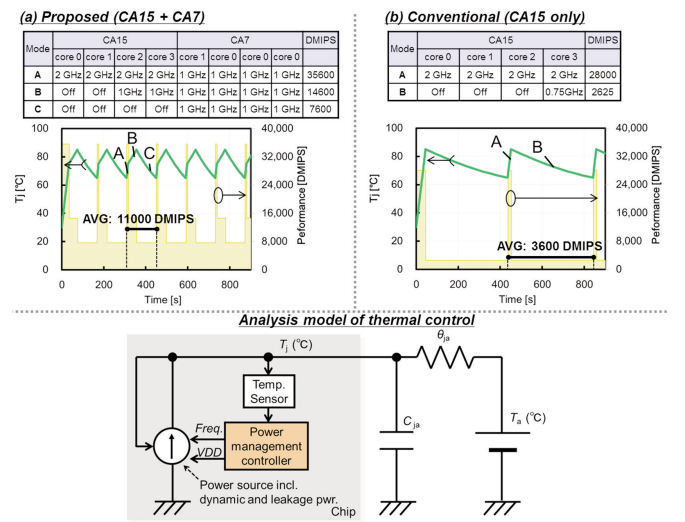

Figure 10.2.6: Thermal control for the heterogeneous CPU architecture.

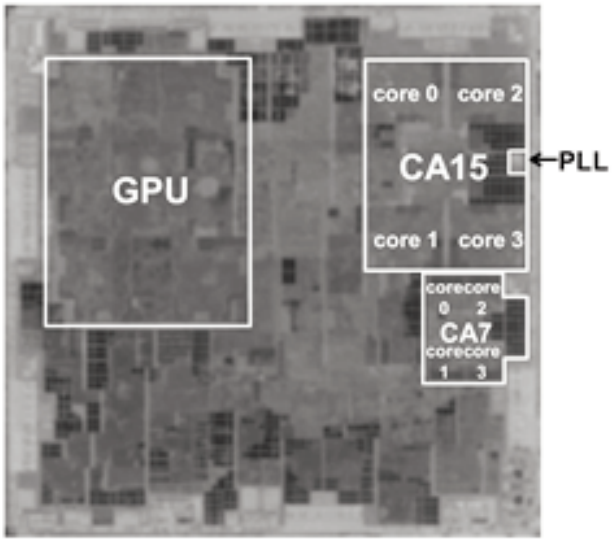

Figure 10.2.7: Die micrograph.

978-1-4799-0920-9/14/$31.00