2.4 A 7nm High-Performance and Energy-Efficient Mobile Application Processor with Tri-Cluster CPUs and a Sparsity-Aware NPU

Young Duk Kim, Wookyeong Jeong, Lakkyung Jung, Dongsuk Shin, Jae Geun Song, Jinook Song, Hyeokman Kwon, Jaeyoung Lee, Jaesu Jung, Myungjin Kang, Jaehun Jeong, Yoonjoo Kwon, Nak Hee Seong

Samsung Electronics, Hwaseong, Korea

Mobile application processors (APs) must be extremely power efficient, while providing high performance for improved user experiences, higher graphic-rendering performance, fancy camera operations, faster data communication, as well as longer battery life [1-3]. In this paper, we introduce a power-efficient high-performance 7nm Exynos™ AP processor with tri-cluster CPUs, a sparsity-aware NPU, and a HW auto-clock gating (HWACG) feature.

Our new Exynos AP integrates eight CPU cores consisting of three different classes for better power/performance coverage. The big CPU is our 4th generation custom core, which improves architectural performance by 23% at the same frequency compared to previous work [4]. It maintains a 6-wide superscalar pipeline, and has an improved memory subsystem with dual store, load cascading, dual 128b master ports and an additional fast-data-forward path from the DRAM controller. Also, it has enhanced branch prediction, prefetchers and cache configurations. Fig. 2.4.1 shows a pipeline diagram for the new CPU micro-architecture. Two NMUL integer vector multipliers are included to enhance machine-learning performance with INT8 dot-product operations. Fig. 2.4.1 also shows architectural performance improvements from the previous version for individual tests of Geekbench v4. Besides dual extreme-performance big M4 cores, dual Cortex-A75 cores are added as middle CPUs for a seamless performance transition to power-efficient quad Cortex-A55 little cores. The big cores run up to 2.74GHz, while the middle cores run up to 2.40GHz, and little cores run up to 1.95GHz respectively. In order to maximize total performance within the power budget, a Hardware Intervention Unit (HIU) is implemented. It monitors the power states of each core, and sets the allowed power budget in the PMIC. When power consumption increases over the given budget, the PMIC signals to the HIU to throttle the big CPU by either clock division or architecture throttling for immediate power reduction, and generates an IRQ so that the job scheduler can lower the frequency.

Figure 2.4.2 compares the power and performance of the 2-cluster and 3-cluster systems. As shown in Fig. 2.4.2, the wide overlap of the middle core with the little and big cores permits flexible workload scheduling. First, when a single heavy task runs on a little CPU for a long time, it can be migrated to middle or big CPU selectively, depending on workload characteristics or power efficiency. Second, when specific scenarios cannot be covered by a little CPU, the main workload can be covered by the middle CPU instead of the big CPU to reduce the absolute power consumption. Third, when the performance of big CPU is thermally limited, it is possible to limit the performance optimally by migrating a heavy workload to the middle CPU. As most of the user scenarios except for specific boosting cases can be covered by the middle CPU, the 3-cluster CPU system shows improved energy efficiency compared to the 2-cluster big-little CPU system.

In the real world, there are various characteristics of tasks besides their resource occupation rates, so existing scheduling methods cannot operate optimally for all kinds of workload. The Heterogeneous Multi-Processor (HMP) scheduler faces challenges as the number of heterogeneous sets increases (e.g. tri-cluster CPU). We introduce an improved scheduling method based on Instruction Set Architecture (ISA), considering the 32b/64b energy efficiency of each cluster. As energy efficiency differs based on the ISA mode, the mode should be scheduled based on an energy model. If we consider only 64b energy efficiency, sometimes a 32b workload could be migrated to the big CPU inefficiently, so a new energy model is designed to properly reflect the 32b/64b energy efficiency. The CPU power is improved by over 30% in specific scenarios as shown in Fig. 2.4.2.

Our new Exynos AP supports a H/W Automatic Clock Gating (HWACG) scheme, where an individual clock net can be on or off automatically according to the clock consumer's need. Once an unused clock is detected, its driver is turned off to reduce power consumption on clock network. Fig. 2.4.3 shows the HWACG architecture including Clock Management Unit (CMU) and IP clock consumers. The full H/W operation provides the best clock gating opportunity and there is no other option available for some bus components. S/W-assisted operation, where part or all of the gating behavior is directed by software, was also adopted. Additionally, an early wake-up system (EWS) is added to reduce clock-wakeup latency. The accumulated bus latency coming from a multi-stage bus architecture can have a critical impact on real-time IPs. The CMU detects a wake-up signal from latency-critical source blocks, and the EWS delivers a common wakeup signal to multiple target IPs simultaneously. With the EWS, the latencies of bus blocks or other systems can be reduced, especially by eliminating the PLL re-lock time, which is around 5.8-75μs. Fig. 2.4.3 shows the power reductions arising from HWACG.

A neural processing unit (NPU) in the AP includes 1024 MACs working at up to 933MHz. The architecture of the NPU is based on a butterfly structure with two NPU domains for overcoming logic cost and wiring congestion. The NPU employs 8b fixed-point precision with channel-wise quantization, i.e. a different decimal point position per channel. All activation functions related to ReLU are performed in parallel. In order to increase the effective performance, it applies zero-weight skipping to reduce the number of operating cycles [5], improving energy efficiency and performance. As a result, it can gain 3.5× performance improvement for a 5×5-kernel layer. Moreover, HWACG is implemented and it can reduce the idle power by 150mW. Measured performance and power consumption shows that the AP can run Inception-v3 at 99 inferences-per-second (infs) with 34infs/W. It also runs ResNet-34 and MobileNet-SSD of AIMark with 111infs and 41infs, respectively. Moreover, the energy efficiency is scaled based on the pruning rate of weights, shown in Fig. 2.4.4.

In our new Exynos AP, we implemented voltage-droop mitigation. A ring-oscillator-type droop detector was integrated, which counts RO clocks within a programmable time window. The droop-detected flag is asserted when the counter value is smaller than programmable threshold values, and then CMU divides the clock to IP by half to reduce load current. Fig. 2.4.5 shows the voltage droop measured with the droop-mitigation solution disabled and enabled, respectively. Droop-mitigation contributed to lower the operating voltage by 12.5mV and to reduce power.

A key technology feature of 7nm is fin-pitch scaling, which brings an AC performance enhancement by Ceff gain at the standard-cell level. Since the minimum patterning pitch is limited to 42nm in case of a Self-Aligned Double Patterning (SADP) process using ArF immersion, further fin-pitch scaling is developed by using a Self-Aligned Quadruple Patterning (SAQP) process. However, DC performance usually decreases during fin-pitch scaling due to the small source-drain (S/D) volume, S/D epitaxy optimization, heat optimization and gate-height reduction are conducted to compensate the DC performance degradation. The enhancement to AC performance is shown in Fig. 2.4.6.

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,* pp. 154-155, Feb. 2013.

[2] H. Mair et al., "A 20nm 2.5GHz Ultra-Low-Power Tri-Cluster CPU Subsystem with Adaptive Power Allocation for Optimal Mobile SoC Performance," *ISSCC,* pp. 76-78, Feb. 2016.

[3] M. Cai et al., "7nm Mobile SoC and 5G Platform Technology and Design Co-Development for PPA and Manufacturability," *IEEE Symp. VLSI Tech.*, pp. 104-105, 2019.

[4] J. Rupley et al., "Samsung M3 Processor", *IEEE HotChips Symp.,* 2018.

[5] J. Song et al., "An 11.5TOPS/W 1024-MAC Butterfly Structure Dual-Core Sparsity-Aware Neural Processing Unit in 8nm Flagship Mobile SoC", *ISSCC*, pp. 130~132, Feb. 2019.

*Samsung Exynos is a product of Samsung Electronics Co., Ltd.

978-1-7281-3205-1/20/$31.00 ©2020 IEEE

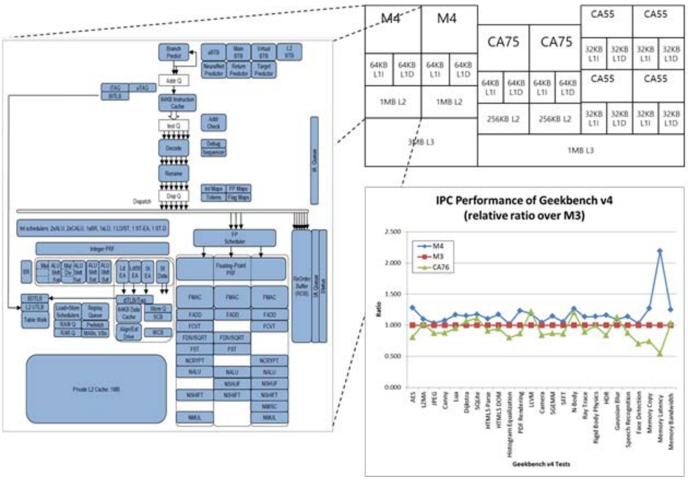

Figure 2.4.1: Our next-generation CPU microarchitecture diagram and architectural performance gain of new generation over previous generation for Geekbench v4 workloads.

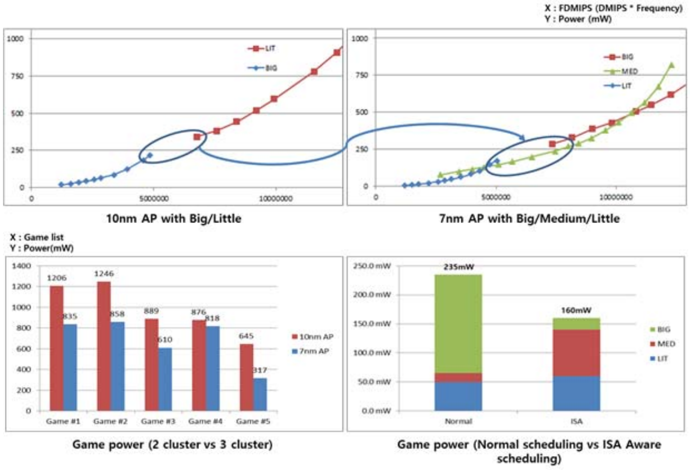

Figure 2.4.2: Performance-power graph for big-little and tri-cluster, and CPU power gain considering ISA characteristics.

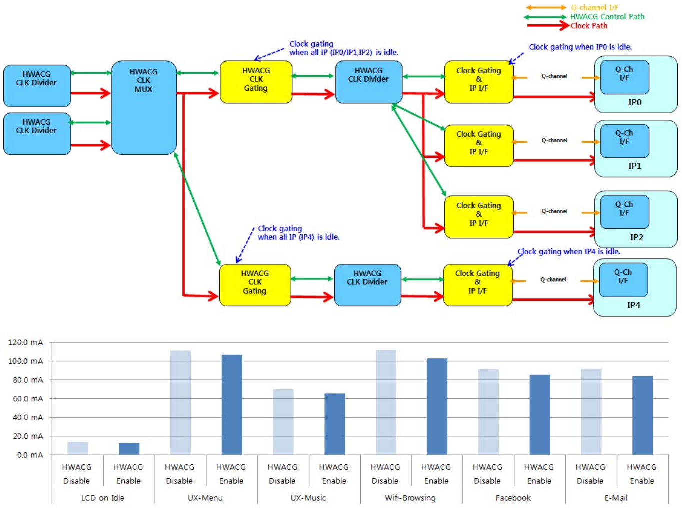

Figure 2.4.3: HWACG block diagram including CMU and IP; power gain from HWACG.

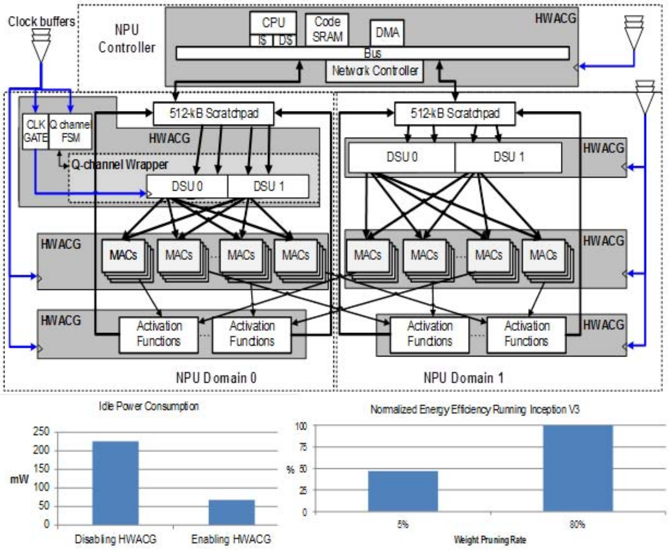

Figure 2.4.4: NPU architecture and its idle power consumption.

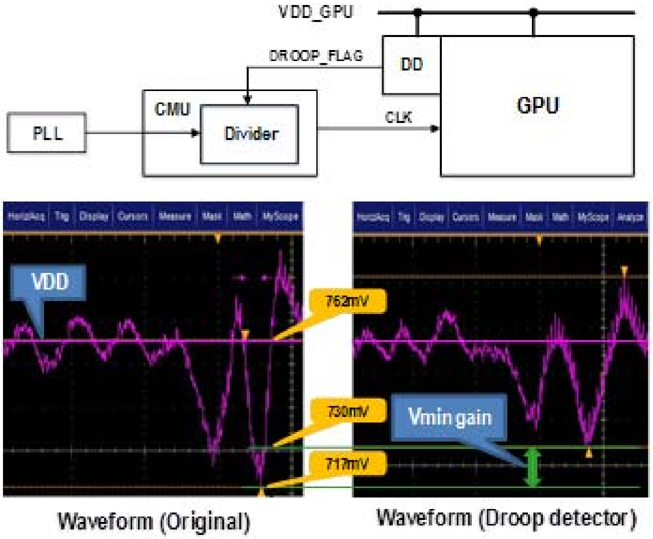

Figure 2.4.5: Droop detector and compensation block diagram, and power fluctuation reduction.

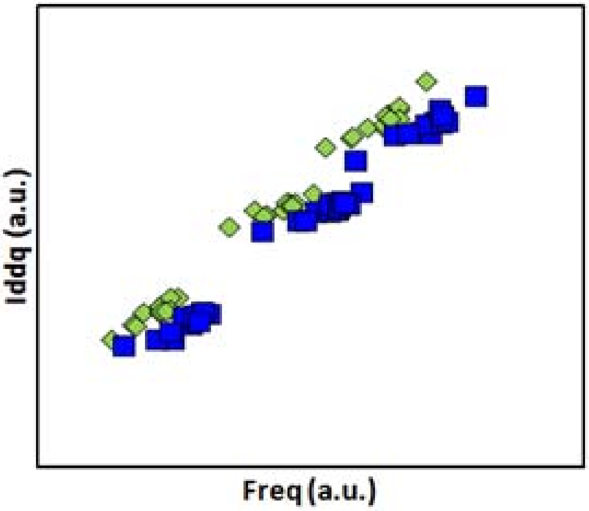

Figure 2.4.6: Device performance enhancement in 7nm compared to 8nm (ring oscillator / blue: 7nm, yellow-green: 8nm).

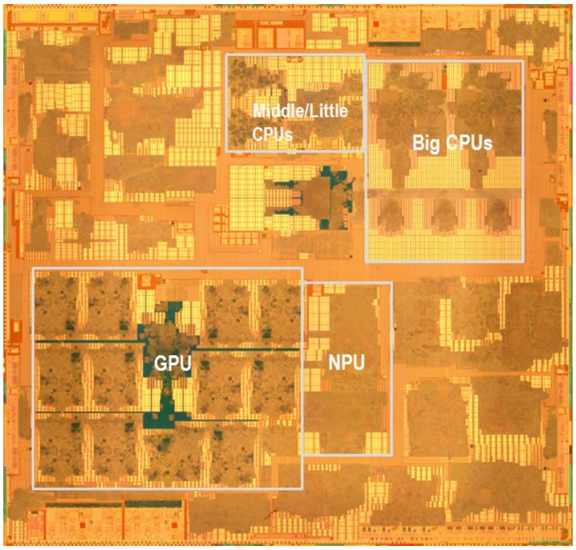

Figure 2.4.7: Die photo.