2.2 A 5G Mobile Gaming-Centric SoC with High-Performance Thermal Management in 4nm FinFET

Bo-Jr Huang, Alfred Tsai, Lear Hsieh, Kathleen Chang, C.-J. Tsai,
Jia-Ming Chen, Eric Jia-Wei Fang, Sung S.-Y. Hsueh, Jack Ciao,
Barry Chen, Chuck Chang, Ping Kao, Ericbill Wang, Harry H. Chen,
Hugh Mair, Shih-Arn Hwang

MediaTek, Hsinchu, Taiwan

In recent years, mobile gaming has grown rapidly to overtake both console and PC gaming markets globally. Thus far, modern smartphones have been powerful enough to support gaming requirements. However, demands of high-performance and computing power cause thermal management challenges to sustain performance during gaming. As shown in Fig. 2.2.1, frequency upgrades did not bring commensurate benchmark improvements due to the thermal wall. This work presents a high-performance fully integrated 5G flagship mobile gaming-centric SoC in a 4nm FinFET process. The SoC consists of an octa-core tri-gear 3.35/3.2/1.8GHz CPU and a deca-core 955MHz GPU to provide a high-quality gaming experience. To maintain high-performance operation under heat ramp-up, a thermal management system that adopts threshold temperature (T_{thr}) control along with energy/temperature-aware scheduling (E/TAS) is proposed. On-chip sensors which monitor temperature, voltage and leakage current are designed to acquire a run-time power budget for E/TAS to boost performance based on computing demand. The proposed thermal management system raises T_{thr} by up to 10°C with 2000/°C benchmark score gain, on average, and maintains a stable temperature range with occasional damping well controlled within 5°C while throttling. The gaming phone achieves a score of 1.146M for the AnTuTu v9.4.2 benchmark.

Figure 2.2.1 shows the SoC, featuring an ARMv9 CPU subsystem with a single Cortex-X2 high-performance (HP) core up to 3.35GHz as the first gear, three Cortex-A710 balanced-performance (BP) cores up to 3.2GHz as the second gear and four 1.8GHz Cortex-A510 cores for high-efficiency (HE) as the third gear. A Mali-G710 GPU for 3D graphics and an in-house APU for AI processing are integrated. Multimedia with 8K video decoding at 30fps, 4K video encoding at 60fps, a camera with up to 320MPixels and QHD+ video with frame rates of 144Hz are supported. Furthermore, external SDRAM connected through a LPDDR5-6400/LPDDR5X-7500 memory interface has a peak transfer rate of 0.46Tbps. The 5G modem supports NR sub-6GHz with 2.5Gbps upload and 7.01Gbps download speeds.

In a flagship smartphone, the CPU and GPU operate at high speed and typically account for more than 30% of the power of the whole SoC in high-performance benchmarks. Consequently, the critical success factor for maintaining performance is efficient thermal control of the CPU and GPU. Taking the CPU as an example, in Fig. 2.2.2, the power consumed exceeds 10W because of the higher speed and power density as the process shrinks to 4nm. Based on the performance/power-efficiency curve for gaming, the CPU performance needs to be boosted by 7%, causing a 60% power increase. As the measured temperature vs. CPU power shows, 16W power dissipation will cause a 10°C temperature jump (T_{jump}) in 1ms, and 25°C in 5ms, posing challenges for thermal management. In a typical thermal-control policy, as system temperature exceeds T_{thr}, the thermal throttling mechanism slows down the clock for cooling, causing a reduction in performance.

Figure 2.2.3 shows the proposed thermal management system for sustainable gaming performance. The process monitor is implemented based on a ring oscillator (ROSC) structure with configurable cell/wire delay. It reflects process variations to model the minimum system operation voltage (V_{sov}) [1]. The leakage sensor is composed of tied-off logic to mimic the leakage current (I_{LKG}). With V_{sov} and I_{LKG}, the power predictor calculates the total power of the system and converts it to the corresponding expected T_{jump} for T_{thr} determination. As shown in Fig. 2.2.4, T_{jump} considers the temperature guard band which covers temperature overshoot, local temperature fluctuation and the response time of the thermal management system. T_{thr} is defined for throttling to prevent system failure as the temperature increases beyond the sign-off level. Fig. 2.2.4 shows the total power calculated by the power predictor vs. V_{sov} for 3K samples. The total power can be mapped to T_{thr} linearly. Observe that T_{thr} of most samples can be increased by more than 10°C compared with conventional global throttling using the worst power. With increased T_{thr}, the throttling mechanism can be postponed to lengthen the duration of high-performance operation.

2023 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-6654-9016-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISSCC42615.2023.10067271

Based on the run-time scenario (idle, browsing, music, gaming, etc.), the target operating point (OPP) voltage/frequency is set as the initial condition (Fig. 2.2.3), with the expected computing power demand and temperature obtained from the thermal sensor. A frequency-locked loop (FLL) which utilizes a tunable-delay ROSC based on post-silicon binning further optimizes voltage/frequency under various workloads [2]. Next, based on temperature, voltage and leakage sensor readings, the microprocessor calculates the run-time power and T_{jump} slew to obtain the sustainable performance time before throttling. Lastly, E/TAS optimizes the thermal gradient by task reallocation among processor cores to further extend the sustainable performance time. Fig. 2.2.5 illustrates the operation of the proposed E/TAS. For program processing, EAS is used to choose the CPU core with maximum available capacity to perform the task. For more precise scheduling considering heat coupling, TAS uses per-core temperature readings to coordinate with EAS for optimizing task assignment, thus minimizing the thermal gradient between CPU cores. Keeping a smaller thermal gradient brings lower leakage and T_{jump} benefits. A gaming test case shows the measured frequency distributions of HP and HE cores with E/TAS. Observe that, on average, frequency is enhanced by 3% in HP cores and 35% in HE cores.

As the sensed temperature approaches T_{thr}, the Cooler enables throttling. As shown in Fig. 2.2.6, the Cooler takes the workload prediction with further consideration of the PCB temperature (T_{PCB}) to form a closed-loop smart frame-per-second (fps) control for a better gaming experience. The evaluation test case requires a minimum 75fps and an average 88fps for smooth gaming. Without the smart FPS control, an average 88fps can be achieved, but the minimum fps is only 72.6. With the smart fps control, average fps reaches 89 and minimum fps is improved to 78.4, ensuring a better gaming experience.

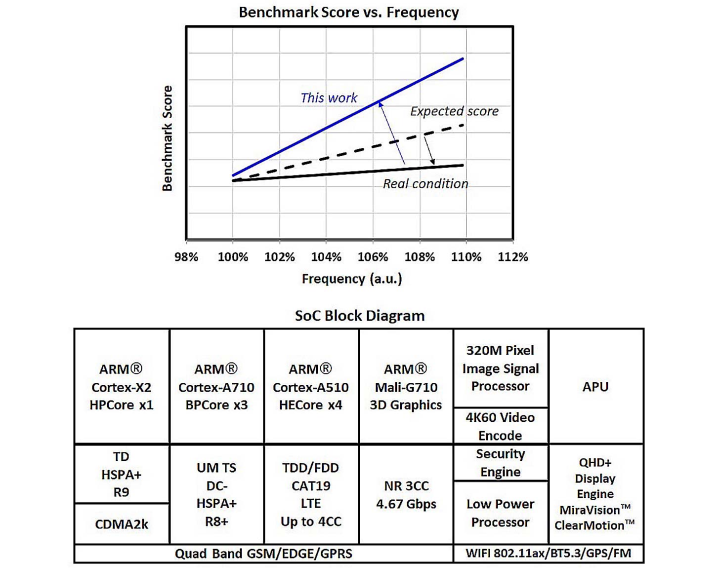
The proposed thermal management system is applied to both the CPU and GPU since they are the critical gaming performance IPs. Fig. 2.2.6 shows measured real-time temperature under the Geekbench 5 multi-core HDR test. With conventional global throttling, the system temperature fluctuates widely with damping up to 10°C, resulting in unstable performance. With the proposed thermal management system, T_{thr} is increased by 10°C with an average 2000/°C benchmark score gain, and the system temperature remains stable within a tight range with occasional damping kept within 5°C to sustain performance. For the AnTuTu v9.4.2 benchmark, the SoC achieves a leading score of 1.146M.

The die photograph of the 108.6mm^2 SoC is shown in Fig. 2.2.7. In summary, a 5G mobile gaming-centric SoC with high-performance thermal management is demonstrated in 4nm FinFET. Sensor-assisted run-time power budget calculation and E/TAS are adopted in the proposed thermal management system to sustain performance against high power induced heat. The achieved T_{thr} increase contributes benchmark performance gain. The gaming phone realizes an AnTuTu score up to 1.146M.

Acknowledgement:
The authors thank Cheng-Yuh Wu, Chao-Yang Yeh, Joey Lu, Tran Trong Hieu, Wei-Li Liao, Kevin Hung, Yuwen Tsai, and Alex Chiou, Mediatek, Hsinchu, Taiwan, for their support on this work.

References:

[1] B.-J Huang et al., “An Octa-Core 2.8/2GHz Dual-Gear Sensor-Assisted High-Speed and Power-Efficient CPU in 7nm FinFET 5G Smartphone SoC,” *ISSCC*, pp. 490-491, 2021.
[2] H. Mair et al., “A 7nm FinFET 2.5GHz/2.0GHz Dual-gear Octa-Core CPU Subsystem with Power/Performance Enhancements for a Fully Integrated 5G Smartphone SoC,” *ISSCC*, pp. 50-51, 2020.
[3] V. K. Kalyanam et al., “Thread-Level Power Management for a Current- and Temperature-Limiting System in a 7nm Hexagon™ Processor,” *ISSCC*, pp. 494-495, 2021.
[4] A. Nayak et al., “A 5nm 3.4GHz Tri-Gear ARMv9 CPU Subsystem in a Fully Integrated 5G Flagship Mobile SoC,” *ISSCC*, pp. 50-51, 2022.



ARM® Cortex-X2 HPCore x1	ARM® Cortex-A710 BPCore x3	ARM® Cortex-A510 HECore x4	ARM® Mali-G710 3D Graphics	320M Pixel Image Signal Processor 4K60 Video Encode	APU
TD HSPA+ R9 CDMA2k	UM TS DC-HSPA+ R8+	TDD/FDD CAT19 LTE Up to 4CC	NR 3CC 4.67 Gbps	Security Engine Low Power Processor	QHD+ Display Engine MiraVision™ ClearMotion™
Quad Band GSM/EDGE/GPRS				WIFI 802.11ax/BT5.3/GPS/FM	

Figure 2.2.1: Thermal challenges of upgrading specifications for gaming and the 5G SoC block diagram.

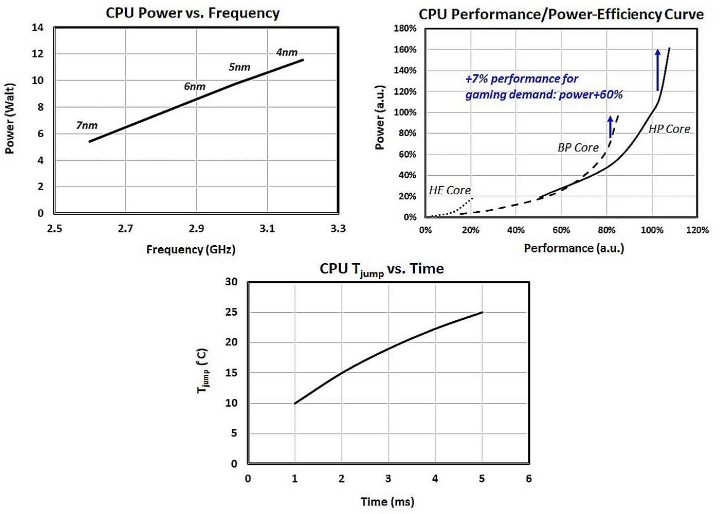


Figure 2.2.2: Thermal challenges arising from performance/power increase.

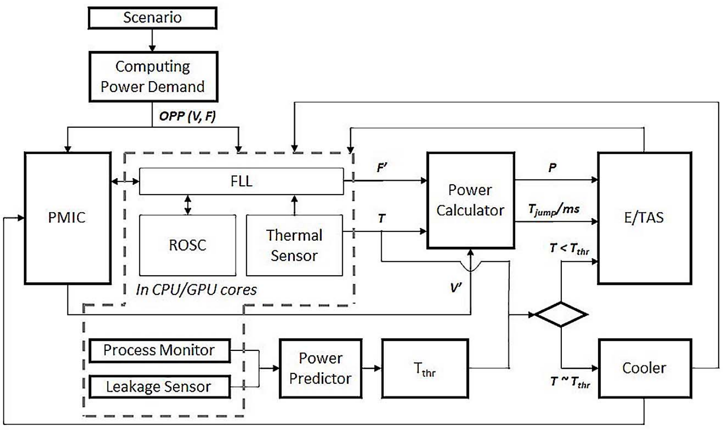


Figure 2.2.3: Proposed thermal management system of the SoC.

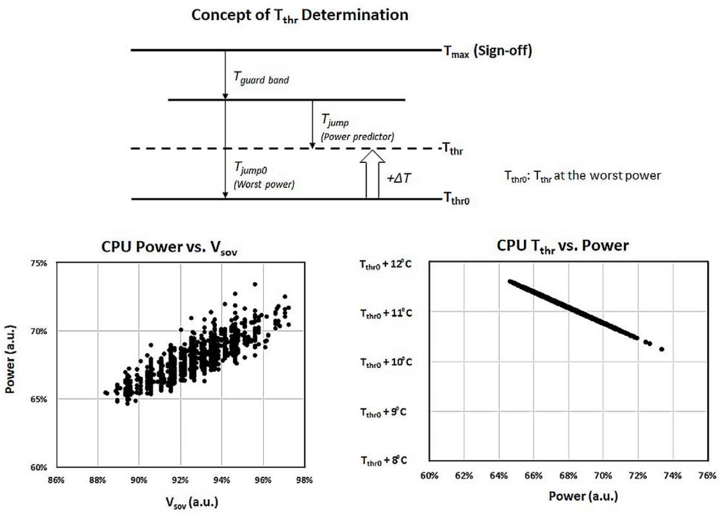


Figure 2.2.4: T_{thr} determination based on power prediction.

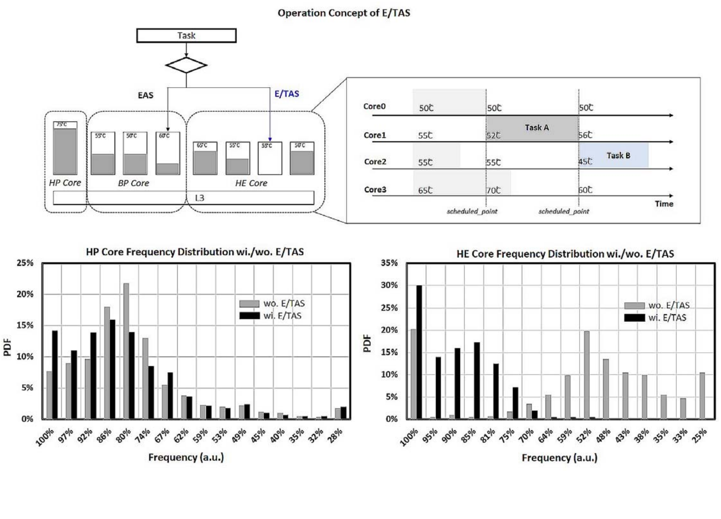
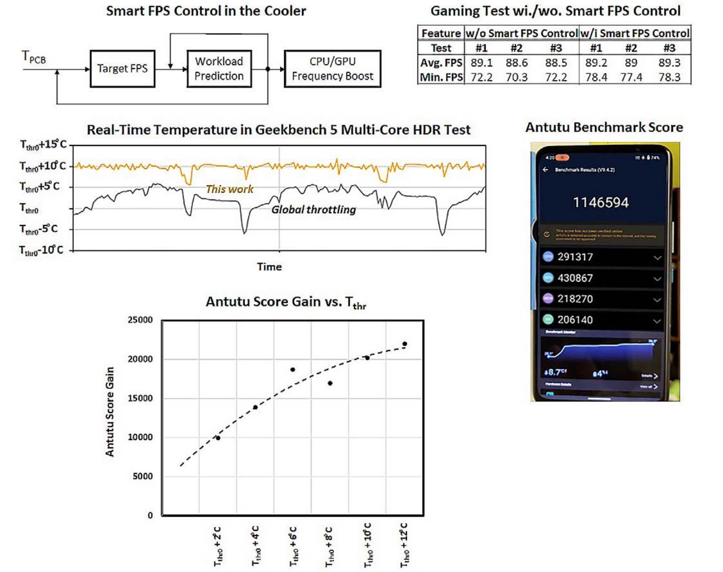


Figure 2.2.5: Operation of E/TAS and the CPU frequency distribution wi./wo. E/TAS in a gaming test case.



Feature	w/o Smart FPS Control			w/i Smart FPS Control		
Test	#1	#2	#3	#1	#2	#3
Avg. FPS	89.1	88.6	88.5	89.2	89	89.3
Min. FPS	72.2	70.3	72.2	78.4	77.4	78.3

Figure 2.2.6: Smart FPS control, real time temperature in benchmark and AnTuTu benchmark score.

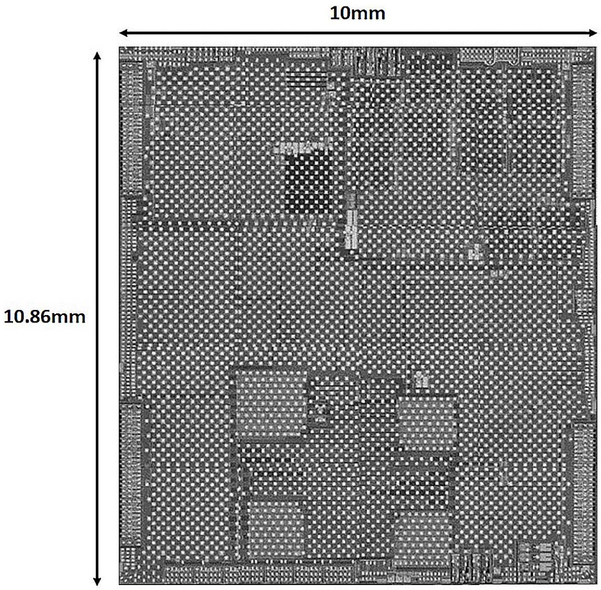


Figure 2.2.7: Die photograph.

978-1-6654-9016-0/23/$31.00 ©2023 IEEE