Understanding the Thermal Challenges of High-Performance Mobile Devices with a Detailed Platform Temperature Model

Ying-Ju Yu and Carole-Jean Wu
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
{yingjuyu,carole-jean.wu}@asu.edu

I. INTRODUCTION

Modern high-performance electronics are embedded in smartphones, self-driving automobiles, and augmented reality wearable. These computing platforms are high performance and multifunctional. One of the major performance limiting factors in these platforms is the poorly designed thermal solution that is triggered to prevent overheating at the processor transistor junction and at the platform surface. Even though the dynamic voltage and frequency scaling (DVFS) feature widely available in modern processors can be used as a control knob in processor temperature management, its associated performance cost is tremendous and the response time for temperature control is significant.

In this work, we build a detailed finite element model to predict the temperature behavior of a tightly-packaged, high-performance embedded platform, from the system level down to the processor architecture level. We first characterize the thermal response of real-world, representative mobile workloads and perform parametric studies to predict the processor junction and platform surface temperature based on the properties of the platform. We validate the proposed temperature model used to predict the maximum temperature at the processor chip and the platform surface. Our preliminary results show that the physical construction (such as air gaps) and the floor plan of the multi-layer platform have a profound effect on the system temperature behavior.

II. METHODOLOGY

Experimental Platform and Measurement Setup: The platform used in this work is a Qualcomm Snapdragon Development Platform. This platform has an MSM8664 system-on-a-chip (SoC) with the big.LITTLE core cluster architecture, and an Adreno GPU. The Trepn profiler is used to read and collect power sensor data for the big and the LITTLE core clusters, the GPU, the DRAM memory, the display, and the battery. The data sampling period is set for 0.1 seconds. The temperature of the core clusters, the GPU, and the battery is measured with the embedded temperature sensors. The measured temperature sensor values are used to validate our model predicted transistor junction hotspot temperature values. Figure 1 shows the overall temperature measurement setup for the experimental Qualcomm platform that is connected to a host desktop for data collection under the IR camera.

Workloads: To cover workloads with different unique power consumption behavior, we use three application execution scenarios to represent common usage patterns [1], [2]. Table I summarizes our workload combinations.

<table>
<thead>
<tr>
<th>Workloads</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>basicmath [3]</td>
<td>simple mathematical calculations</td>
</tr>
<tr>
<td>Web browsing [1], [2]</td>
<td>automatic web page rendering</td>
</tr>
<tr>
<td>Youtube</td>
<td>online video-sharing</td>
</tr>
</tbody>
</table>

III. POWER AND TEMPERATURE CHARACTERIZATION

Figure 2 depicts the thermal hotspot temperature of all workloads examined in this paper at the processor transistor
In this section, we present the model evaluation results for the described temperature model in Section IV. We use three power stressing scenarios and validate the temperature prediction results using the experimental setup in Fig. 1.

IV. TEMPERATURE MODEL CONSTRUCTION

We develop the multi-level temperature model and describe the geometric and material properties with the commercially available COMSOL Multiphysics Simulation Software. Figures 3 and 4 depict the overall geometry of the target platform and a cross-sectional view of the platform where the high-performance SoC resides. We physically disassemble the experimental platform to model all major power consuming/heat dissipation and thermal-critical components in the platform, including the display, the case, the battery, the Package-on-Package (PoP) SoC, and the PCB.

V. TEMPERATURE MODEL EVALUATION

In this section, we present the model evaluation results for the described temperature model in Section IV. We use three power stressing scenarios and validate the temperature prediction results using the experimental setup in Fig. 1.

Fig. 3: An illustration of the multi-layer platform construction in the temperature model.

Fig. 4: A zoom-in version of the cross-sectional view of the platform for the PoP region.

The total power consumption and thermal hotspot temperature are relatively constant and stable for basicmath and youtube. On the other hand, we observe periodic computation bursts in web browsing that lead to power spikes. The combined basicmath + web browsing multiprogramming execution scenario represents the high-performance, high-power execution scenario that illustrates the high likelihood of temperature limit violations.

We run two instances of basicmath on Core 4 and Core 6 to perform temperature estimation. The prediction error rate for the thermal hot spots on the processor die is 3.6% for Core 4 and 1.9% for Core 6, respectively. The maximum prediction error rate occurs at the GPU and its prediction inaccuracy is 19.7%. We attribute this over-prediction to the simple, generic floor plan assumption made in our temperature model. However, since thermal hotspots are unlikely to happen at the GPU, this relatively higher error rate is considered acceptable. For the platform surface hot spot temperature, our model produces an error rate of 10%.

b) Scenario 2 (3-core, 8.3W): We execute an additional instance of basicmath on the high-performance core cluster. We observed that soon after the multiprogrammed workload starts execution, the processor ran into the throttling mode, due to the processor temperature limit exceeds 80°C limit. The prediction accuracy for the temperature of the processor cores is 7.1% on average while the prediction accuracy for the platform surface temperature is 9.1% for the hotspot.

c) Scenario 3 (4-core, 9W): Finally, we construct an execution scenario that stresses all cores of the high-performance cluster by running an additional instance of the computation-intensive basicmath. Similar to the previous two execution scenarios, thermal throttling occurs and this time the processor is throttled down even more. The prediction accuracy for the temperature of the processor cores is 4.8% on average while the prediction accuracy for the maximum platform surface temperature is 11.3%.

VI. SUMMARY

In this paper, we characterized the thermal response of real-world, representative mobile workloads and then used the response to calibrate the temperature prediction model. The thermal model offers accurate prediction accuracies of 90% for hot spots at the processor die and at the platform surface. Also, we used the thermal model to explore alternative construction choices. As the air gap increases, the processor temperature rises significantly. We hope the work presented in this paper can motivate advanced thermal management for high-performance embedded platforms.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their useful feedback. This work is supported in part by the National Science Foundation (under grants for the I/UCRC Center for Embedded Systems #1361926, CNS #1358805, CCF #1525462, and CCF #1652132).

REFERENCES

