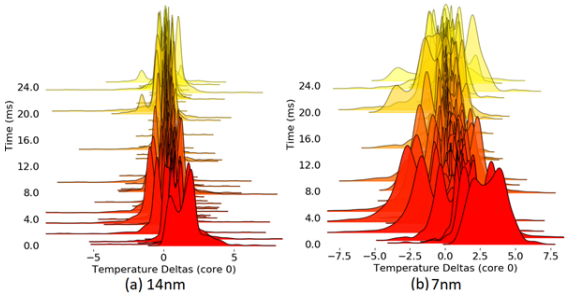
Utilizing the HotGuage Framework for Hotspot Behavior Analysis

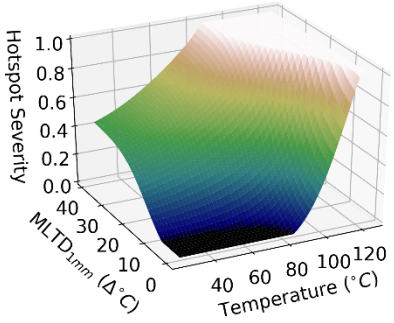
David Werner*, Maziar Amiraski, Alexander Hankin, Julien Sebot, Kaushik Vaidyanathan, Mark Hempstead*
161 College Ave, Halligan Hall, Medford, MA
Department of Electrical and Computer Engineering, School of Engineering, Tufts University
{dwerne01,mark}@ece.tufts.edu; (617) 627 0969



Hotspots are a growing concern for modern heterogeneous computing systems especially systems with stacked dies and advanced packaging. As process technology has scaled, power density has increased to the point of stressing the power and cooling limits of modern microprocessors. While architecture-level techniques exist to limit the power and heat of portions or the entirety of an integrated circuit (IC), they are triggered based on whether power or temperature exceeds a certain threshold. Modern hotspots are not only hot but, among other things, result in high temperature deltas which are present at the level of functional unit as well as within small sub-units. These so-called "advanced hotspots" are characterized not only by high absolute temperatures, but also by their fast development, high spatial localization, non-uniformity, and application dependence. This new type of hotspot is of concern for a variety of reasons including increased potential for timing faults as well as concerns of local thermal runaway due to increased leakage power that can arise before the concerning heat signature has time to propagate to an on-chip thermal sensor. This problem is further exacerbated with each new technology node and in the presence of increasingly heterogeneous microprocessors. As an example, the figure above shows the distribution of the amount by which per-pixel die temperature changes over 200 µs intervals for 14nm compared to 7nm [1]. The 7nm die is worse in two ways. First, the peak change in temperature is greater, resulting in faster temperature spikes. Second, the variance in temperature deltas is wider, indicating the potential for large temperature deltas. All of these changes take place over only 200 µs, indicating that techniques to mitigate hotspots will need to be even more aggressive than they previously were, resulting in the need for increased guard-bands at the cost of dramatically decreased performance.

Academia and industry are actively developing new technologies, integration methods, and packaging techniques. 3D stacking, interposers and chiplets all have cooling challenges that could exacerbate hotspots. In addition, new memory technologies and mixed-signal circuits are more sensitive to temperature fluctuations. Given this rich and expanding design space, the behavior of hotspots will need to be studied in a fast end-to-end manner. Updated hotspot metrics as well as an end-to-end simulation framework are needed to be able to rapidly evaluate potential new architecture-level mitigations and perform design space exploration which enables comparing the trade-offs of different mitigations. This work encourages the use of newly developed metrics, simulation models, and tools that complement existing development efforts.

Hotspot Characterization Metric

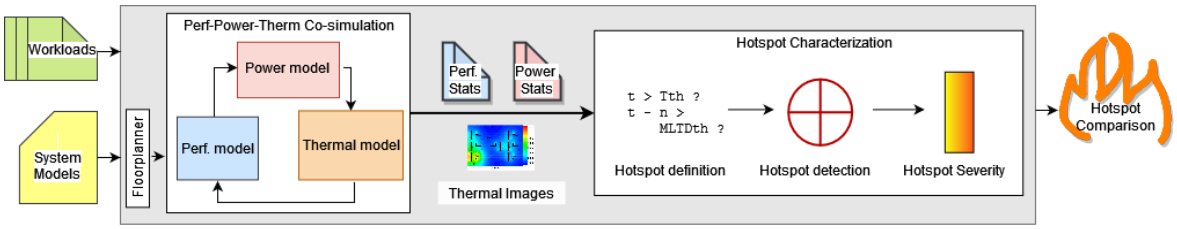


Current hotspot characterization methods focus on the magnitude of the hottest point and not the size of the thermal gradient. While temperature is an important indicator of potential hotspots, gradients are important as well, as their magnitude is indicative of the speed of the hotspot and its propensity to cause timing failures or reliability issues. Our HotGauge work introduces a formal definition that uses the temperature of the hotspot with a new maximum localized temperature differential (MLTD) that captures the maximum gradient with a defined radius (either a core or 1mm distance). HotGauge defines a hotspot severity metric, (figure right) which combines the MLTD of the hotspot with its temperature using three configurable sigmoid functions which can be used in a predictor or a control system. Intuitively the severity metric captures how extremely hot spots on chip (e.g. > 100°C) are concerning and points of moderate temperature (e.g., 80°C) with high gradients (MLTD < 25°C) are also concerning because of timing failures and the speed of the hotspot. This severity metric can also be adapted for

specific systems to reflect the maximum junction temperature, timing budgets and the reliability and timing of different technologies with respect to temperature.

Approach

The newly developed HotGauge framework [1] is intended to fill this gap and enable the characterization of advanced hotspots. Included in the framework—summarized in the figure below—is an end-to-end toolchain that encompases performance, power, and thermal simulation, as well as processing techniques and methods to quantify the severity of hotspots. These characterization methods enable the simultaneous analysis of hotspot severity using a new metric that simultaneously accounts for the temperature of a potential hotspot as well as the size of the thermal gradient it is causing. Initial case studies with HotGauge simulated an Intel Skylake style processor in 14nm, 10nm, 7nm and demonstrated hotspot behavior and severity across applications. These techniques can be applied to new technologies developed by research teams to characterize thermal behavior and hotspots at an early stage.



End-to-End High-level Application Specific System and Thermal models

HotGauge is a flexible framework that can support a range of power, performance, and thermal models. Currently HotGauge supports performance models from Sniper, Power models from McPAT, and cooling/thermal models from 3D-ICE but these can be easily interfaced with other models developed by other teams. For example, previous iterations of this work used the hotspot thermal simulator. HotGauge also includes the addition of power models that work at 14nm and below, beyond the range supported by the public releases of MCPAT, enabling for a detailed comparison of current and future tech nodes for CMOS processors. Case studies using HotGauge have shown that Hotspot behavior varies widely with application behavior, thus detailed but fast models are needed. HotGauge needs a trace of power consumption in time for components on a physical floorplan to accurately model thermal and hotspot behavior. Our team has developed additional models of heterogeneous components including different sized microprocessor units and systolic arrays. Given the flexibility of the underlying tools, HotGauge can also be extended in other ways, including modeling 3D-stacked dies and microarchitectural hotspot mitigation techniques.

Dynamic Hotspot Predictors and System Management of Hotspots

Our initial studies found that hotspots vary within a core based on application behavior and existing thermal conditions. These hotspots appear quickly and can have catastrophic consequences. Our published work quantified how much worse future technology nodes are in comparison to current nodes. This work also showed how HotGauge can be used to quantify and study the effect of various mitigation techniques on hotspot severity, including scaling up the area of problematic units. Current efforts include developing a hotspot prediction model that can be implemented in hardware and take appropriate actions to avoid hotspots while minimizing performance loss.

Next Steps and Future Work

The worsening problem of hotspots can best be addressed through the standardization of hotspot characterization techniques. Frameworks such as HotGauge take a step towards this end goal, allowing for direct numerical comparison of hotspot severity across competing designs and techniques.

References

[1] A. Hankin, D. Werner, M. Amiraski, J. Sebot, K. Vaidyanathan and M. Hempstead, "HotGauge: A Methodology for Characterizing Advanced Hotspots in Modern and Next Generation Processors," *2021 IEEE International Symposium on Workload Characterization (IISWC)*, 2021, pp. 163-175, doi: 10.1109/IISWC53511.2021.00025