Abstract

HSSB 2022: HotSpots Strike Back

Abstract: Implications of Power Density on Chip Design

Mark Gallina
Mobile Performance Architect
Principal Engineer
Client Computing Group

Local power density and temperature response must be incorporated as part of chip architecture and design. Aggregated power alone is no longer a sufficient metric.

The rapid pace of technical innovation required to maintain Moore's Law has encouraged the industry to shift from planar transistor designs to FinFET and soon Gate-All-Around designs. This fundamental change in transistor construction has resulted in process node transitions that no longer follow Dennard Scaling: Power density is not constant with each process node, and in fact is increasing. This increase in local power density manifests as unique challenges for chip architects and designers in two ways: 1) exponentially increasing temperature ramp rates; 2) isolated hot spots within a design. This presentation will primarily focus on the first aspect and discuss potential areas of research to pursue for solution options.

Increased Temperature Ramp Rate: Traditional thermal control for high volume CPU designs has relied on a multi-level scheme with independent mechanisms employed at both the chip level and system level. The chip level control was depended on to monitor circuit temperature and adjust frequency and voltage operating point when certain temperature thresholds were reached. This control loop typically runs at the 1ms to 10ms timescale. The expectation was this control loop was available as a fail-safe mechanism to prevent thermal damage to the circuit should the system thermal solution prove inadequate. The system thermal solution and control was the primary mechanism to maintain safe operating temperature for the chip, with the system fan speed being adjusted dynamically to compensate for fluctuations in chip power. In previous generations if a CPU needed more cooling the system thermal solution capability was increased accordingly. With modern SoC designs that employ multiple core and non-compute domains on a single chip, the thermal challenge is getting more difficult. Modern CPU cores have very large dynamic frequency ranges with the operating point being adjusted very quickly as a function of software demand.

A brief overview of system thermal solution components, their capabilities and limitations will be discussed. Data will be presented showing the different time scales that the system thermal solution components can influence the temperature of the chip juxtaposed with the transient ramp rates observed on a latest generation Intel® Client CPU. A simplified thermal model will be introduced which will be used to explore the temperature transient response of the CPU for different power density levels. This data will demonstrate that system thermal solution components influence temperature response at time scales on the order of 1 second or more which are insufficient to address the rapid temperature rise observed at peak CPU frequency.

Implications: In typical computing device operating environments the ambient temperature is around 25°C with a typical maximum chip operating temperature around 100°C. This results in a 75°C temperature delta budget for the entire thermal stack to operate within. Most systems have internal temperature rise above ambient, resulting in local temperature at the heat exchanger inlet that is 10°C-15°C warmer than ambient. Using the higher value, the usable temperature range for dynamic operation is 60°C. The transient ramp rates observed from measurements and simulations for different power densities and total power levels will be compared against the total temperature budget. This data will show that total chip or core power alone is no longer a sufficient metric to use when assessing the thermal integration difficulty of a proposed design. The data will show that there are situations where higher total power results in lower transient ramp rates.

Call to action: We need to develop new architectures and tools that incorporate local power density and temperature response as part of our design methodology. While aggregated power is still an important metric to drive design, finding methods to incorporate power modulation over time and space will also become a critical factor for future chip designs.