Accurate Thermal Analysis of Chip/Package Systems

Ting-Yuan Wang and Margaret Schmitt, Apache Design Solutions - March 15, 2007

Introduction
With the increasing complexity and power dissipation of modern electronic designs, controlling peak temperature and predicting the temperature profile on the chip early in the process is becoming critical for insuring system reliability. As the complexity of chips scales according to Moore's Law, the power density, as well as the total power dissipation of chips, is still increasing. This creates severe challenges for the thermal design of packages and chips. SoC designers address this problem with one or more low-power design methodologies, such as switching off parts of the design when not in use. These low power design methodologies create their own thermal challenges, as they may generate "hot" and "cool" regions on the chip.

Thermal Impacts on Modern SoC Designs
At 90nm process node and below, an additional design constraint, such as leakage power, arises from thermal integrity considerations. In these systems, 30% or more of the total power dissipation can be caused by gate leakage, and leakage current is a strong function of local temperature. As the dependency between leakage power and temperature is highly non-linear, knowing the average temperature of a system is not sufficient. Therefore, predicting the leakage power, and thus the total system power, requires detailed and accurate knowledge of the temperature distribution. Local temperature hotspots, with higher- than- average temperatures, contribute over-proportionally to the total power dissipation of the system. Furthermore, local temperature distributions have an impact on the performance and reliability of integrated circuits. As a result, accurate thermal analysis is an essential part of the design process for modern SoC designs in today's technologies.

The key design parameters affected by heat production and dissipation characteristics, and the resulting temperature variations across the chip, are:

- Leakage power
- Timing closure
- Reliability

These thermal issues are described more completely in the following sections.

Leakage Power
The contribution of leakage power to the total power dissipation of the chip is expected to increase as processes scale beyond the 90nm node. Figures 1 and 2 show a comparison of the power dissipation and temperature dependency trends for 90nm and 45nm nodes. Clearly, because of the increasing leakage power contribution, it is not sufficient to calculate the total power of a system based on an average temperature assumed constant over the entire chip. Only with the knowledge of the detailed temperature distribution over the chip can designers accurately calculate the leakage
power at every location, and thus the total power of the chip.

1. At 90nm process node, leakage power increases more than linearly with temperature.

2. At 45nm process node, leakage power increases quadratically with temperature, nearing dynamic power consumption at higher temperature.

But in addition, as leakage current increases with local temperature, local heat generation increases as a result. Therefore, power dissipation and temperature distribution are interdependent, and have to be considered simultaneously in order to achieve an accurate prediction of both the power consumption and the temperature distribution on the chip. However, current design analysis methodologies typically assume an average system temperature based on estimates of the chip power dissipation and the thermal dissipation of the chip/package system, and do not consider the
dependency between local temperature and leakage power, or model thermal hot-spots in the design. Thermal hot-spots, however, contribute significantly to the leakage power and the heat dissipation in the system. As a result, current thermal design methodologies fail to predict the correct total power dissipation and the maximum temperatures of the system.

**Timing Verification and Closure**

With rising temperature, transistor drive strengths decrease due to carrier mobility degradation, leading to slower slew rates and gate delays. Additionally, interconnect delay and slew also increase with rising temperature due to increased metal resistivity. This affects the setup and hold time margins of circuits. As temperature variation has an impact on the slew rates of signals, it also has an impact on the cross-talk noise between signal lines.

In current design methodologies, timing and cross-talk are analyzed at various process and temperature corners, also assuming a homogeneous temperature distribution across the chip for each corner case. Thermal hotspots, however, can be significantly above the average, or the peak temperature can be well below the worst corner temperature, which is not addressed by this methodology. The design risk becomes more severe for early-mode analysis, which analyzes race conditions between signals and their timing reference (or the clock). If the signal and clock path experience different temperatures, with a clock path slower and/or data-path faster, a hold-time violation can cause chip failure.

![Diagram showing delay and slack at different temperatures](image)

3. **Thermal conditions impact timing of launch and capture paths.**

Clock nets in particular typically cover the entire area of the chip, and can be exposed to the full chip temperature variation. Clock nets, and branches of the clock network exposed to different local temperatures, will show different delays, causing clock skew. Low power techniques such as clock gating, power gating, dual threshold voltage, and voltage islands increase the temperature variation and local hot spots, causing an even greater impact on timing. For these reasons it is not sufficient to analyze timing at one or more constant temperature points for the entire design. It is essential to perform accurate analysis of local temperature variations on all critical timing paths.

**Reliability**

Electromigration (EM) in the metal and via interconnects is a major limiting factor for the reliability
of integrated circuits. EM describes the transport of mass in metals under the stress of high current density, causing metallization failure. Electromigration increases exponentially with temperature. In general, conductor lifetime, or mean-time-to-failure (MTTF), is used to measure EM effect, which is modeled by Black's equation:

$$MTTF = \frac{A}{J_{avg}^2} \exp\left(\frac{E_a}{k_B T_m}\right)$$

where $J_{avg}$ is the average current density, $T_m$ is the metal/via temperature, $A$ is a constant which depends on the geometry and microstructure, $E_a$ is the activation energy, and $k_B$ is Boltzmann's constant. To satisfy the required conductor lifetime, an upper bound is placed on current density. However, the key element of this relationship is that a safe current density limit decreases exponentially with temperature increase to be protective, as shown in Figure 4. Since MTTF is an exponential function of temperature, the current density limit needs to be reduced with increasing temperature in order to maintain the same reliability.

Current VLSI design methodologies assume a constant temperature corner for analysis, that is, a temperature of 105°C for the entire chip. For hot spots with a temperature higher than 105°C, the design may fail, even though it passed during the conventional reliability analysis. On the other hand, constant temperature assumption also may be pessimistic, leading to overdesign. For example, assuming a constant temperature of 105°C demands a low current density limit for all wires in the design. For most wires, this is very pessimistic, and can lead to thousands of false violations.

4. When charting the EM current density limit as a function of local temperature an increase in temperature of 25 degrees Celsius decreases the current density by approximately a factor of three.
Low Power Designs
Low Power designs make extensive use of clock gating and power gating. This increases the temperature variation over the design by generating regions on the chip with vastly different activity and power dissipation profiles. This can have a large impact on the timing behavior and sign-off of the system, as discussed above. It also means that leakage current estimation based on an average temperature will be too optimistic, since clock-gating generates cooler and hotter regions, and the hot regions contribute over-proportionally to the leakage current of the system. Therefore, thermal integrity analysis is not only important for high-temperature, high-performance designs like microprocessors, but also for low-power designs, which are even more exposed to thermal effects that influence performance, reliability, functionality, and power dissipation of the system.

Thermal Integrity Analysis Methodology
Modeling System Heat Generation and Dissipation
Thermal analysis consists of two major elements: predicting the heat generation spatially and temporally in the system, and modeling the heat transport and dissipation in the system. The system analysis must include a good model of all physical elements of the chip and the chip’s thermal environment, including the package, board, heat sink, and cooling system.

The law of the conservation of energy explains heat conduction and the analysis of it. The heat energy generated at any given point in a system equals the heat dissipated at this point plus the power causing a temperature increase at this point. When there is more heat generated than is dissipated at a point of the system, that part of the system heats up. The system reaches thermal equilibrium and a stable temperature distribution when the heat generated equals the heat dissipated at each point of the system.

The major heat source in a VLSI chip is the power generated by transistors in the silicon substrate. The total power of a VLSI circuit consists of dynamic power, short-circuit power, and static power. At process nodes 90nm and below, leakage currents are a major source of static power consumption. Current in metal interconnects on the chip is also generating heat through resistive loss, resulting in a self-heating of interconnects. In addition, the heat from neighboring SiP components also affects the final temperature distribution of a chip, which is especially important for 3D stacked dies thermal analysis.

The temperature variation across the chip is dependent on the power distribution, as well as thermal conductivity and geometry of different materials. The amount of heat removed and the amount of heat trapped in the system are dependent on the design of package, board, heat sink, and cooling system, as well as the temperature difference between the VLSI system and the ambient, as shown in Figure 5.
5. Paths of heat dissipation through substrate/metal layers, package, heat sink to the ambient are sown for the chip and surrounding cooling environment.

**Thermal Analysis of IC-Package-System**

One of the challenges for an accurate chip-level thermal analysis is the modeling of boundary conditions, including package, heat sink, board, and cooling system. The final power and temperature distribution strongly depends on the boundary conditions. For example, the boundary conditions of SiP systems are critical for modeling the heat transfer between different chips. On the other hand, chip power distribution is the heat source for system-level thermal analysis, and is temperature dependent because of leakage power. Therefore, an IC-Package-System thermal co-analysis, as shown in Figure 6, is required for determining the final power and temperature distribution.

One of the potential problems when chip power is considered to be a constant is thermal runaway. For example, the heat dissipation ability of a selected package is based on an estimated constant power. However, the actual power may be higher than expected according to the actual temperature. In turn, the increased power causes high temperature distribution. And this process continues until the system burns out.

6. A co-analysis flow for IC/Package/System looks like this.
There are typically three types of boundary condition models used when performing an on-chip thermal analysis. These types are defined below and different types of boundary conditions can be mixed in a system.

- Adiabatic: assumes no heat transfer across the boundary
- Convective: assumes a known model for heat transfer across the boundary, which requires a known equivalent thermal resistance network
- Isothermal: assumes no change of boundary temperature distribution over time

*Figure 7* shows a convective boundary condition modeled using equivalent thermal resistance across the faces of a die and *Figure 8* represents an isothermal boundary condition model.
Integrated Power-Thermal-Electrical Analysis
The prediction of power dissipation is critical for a stable thermal analysis flow. For this, the power dissipation at each point of the system, in transistors as well as in interconnects, must be analyzed. The power generated at any point will either be transported away or heat up this point of the system, raising its temperature. However, many of the thermal, electrical, and mechanical properties of the system are also temperature-dependent and will change as the temperature in the system is changing.

For example, leakage power in the chip increases sharply with temperature, and metal resistivity increases with temperature. This means that the power dissipation at each point of the system changes the temperature in the system, but the temperature also changes the electrical parameters and power dissipation in the system. This interdependence mandates that power, thermal, and electrical characteristics cannot be analyzed independently in the system, but demand an integrated process to achieve an accurate analysis of the system.

Once a stable power-thermal-electrical solution is achieved, the temperature distribution across the design can be used to analyze its effect on power, timing, EM, and voltage-drop. Due to the mutual coupling between power, thermal, and electrical parameters, it is most efficient to separate the analysis into two different dependency loops, Power-Thermal (PT) and Power-Thermal-Electrical (PTE), as represented graphically in Figure 9.

Power-Thermal Loop
The distribution of power dissipation in the system determines the temperature profile in the system. The resulting system temperature profile, on the other hand, has a direct impact on the system power dissipation. A stable solution of power and thermal distribution therefore requires an iterative analysis between the power and thermal analysis, that is, an iterative PT loop.

At the starting point of the PT loop, the power dissipation in the system is estimated assuming a homogeneous temperature distribution. Based on the resulting power distribution, the temperature distribution is analyzed, generating a three-dimensional thermal map. As described previously, the temperature distribution in the system has an impact on the power dissipation at each point of the
system. Therefore, using the simulated thermal map, the power dissipation of the system has to be updated, and a new iteration started. Eventually, the iterative loop converges on a stable solution for the power and temperature distribution in the system, as shown in Figure 10. The initial value used for temperature will not affect the final solution. However, the initial temperature value will determine how quickly the PT loop converges. For systems with thermal runaway, the PT loop will not converge.

10. Beginning at a constant temperature, iterations between power and thermal analysis provide a convergent solution for the temperature distribution across the chip.

Power-Thermal-Electrical Loop
The temperature distribution in the system also affects the resistance of metal wires, which affects the power dissipation in the wire (Figure 11), as well as the voltage drop across the system, and therefore the local Vdd value. The power dissipation of the circuits also depends on this local Vdd value and temperature. This creates the need for a second analysis loop between power calculation, thermal analysis, and circuit analysis, the Power-Thermal-Electrical (PTE) loop.

After finishing the PT loop, the temperature-dependent power and 3D temperature profile can be used to analyze the thermal impact across the chip on timing, reliability and voltage drop. Based on the temperature profile, the temperature-dependent resistance can be calculated and updated. This allows for an update of the self-heating of interconnects in the design. It also allows an update of the IR-drop / Dynamic Voltage Drop (DvD) analysis in the system, determining the new supply voltage at each cell. Using this new supply voltage profile, the cell currents in the system are updated as well, and PT loop is started again using the updated power distribution as starting point.
11. You can build an interconnect model for self-heating of wires.

Full-chip IR-drop/DvD analysis can be very time consuming due to the large complexity of the power supply network. It can also require several iterations of the PTE loop to achieve a stable power-temperature-voltage distribution in the system. Pre-characterized cell current allows a fast update of the power distribution for different supply voltage and temperature. Also, it is essential to have a fast and efficient voltage drop analysis engine tightly coupled to the power calculation and the thermal simulation.

*Figure 12* shows a sample three-dimensional thermal map for the temperature variation of a layer of the chip, achieved after the convergence of PTE loop. The figure clearly shows the inhomogeneous distribution of temperature, revealing hot-spots that require special attention.

The benefit of a thermal-aware sign-off flow is clearly apparent in areas such as reliability analysis. Where a constant temperature assumption can require over-design to fix many false EM violations and a too-low assumed temperature may cause reliability issues in the design, especially where local hot-spots of higher temperature require additional EM fixes to guarantee reliability.
12. A mathematical package can show a 3-D map of chip layer temperature profile.

Figure 14 illustrates the benefit of a thermal-aware EM sign-off analysis, which highlights the ‘real’ current density hot-spots on the chip, rather than the constant temperature scenario shown in Figure 13. For example, by using a conventional analysis method, 11,000 EM violations were identified. Whereas using a thermal-aware EM analysis flow, this number was reduced to 2,000 actual EM violations.

13. A design for conventional reliability at constant temperature. shows this EM profile during sign-off analysis.
14. By contrast a design for thermal-aware reliability has this EM profile of sign-off analysis.

Summary
In conclusion, below summarizes a list of components that are required for effective thermal integrity analysis flow.

- **Platform architecture**: For efficiency full-system thermal integrity analysis must be based on a single-platform that performs power analysis, thermal analysis, circuit analysis, power/signal net extraction, reduction, and temperature-dependent cell characterization.

- **Power Calculation**: Requires a temperature-dependent power calculation engine for full-chip analysis, which considers the state-dependent leakage, short circuit, and switching power.

- **Thermal Simulation**: Must simulate the temperature profile across the chip, layer by layer, taking into account the thermal properties of different materials, packages, board, heat sinks, and cooling systems.

- **Electrical analysis of the power/signal network**: Analysis must provide temperature-dependent R extraction, reduction, and circuit simulation of the on-chip power/signal network. The thermal-aware power grid extraction can then be used to determine the temperature impact on full-chip EM and IR-drop/DvD analysis.

- **Thermal-aware library characterization**: To achieve cell-level complexity and transistor-level accuracy, the complete cell library should be characterized using a temperature-dependent and voltage-dependent library characterization flow based on an accurate Spice simulation engine.

- **Impact on Timing**: For the timing impact analysis, temperature-dependent and voltage-dependent delay information of the devices is needed to generate a modified SDF. This SDF can be used in a static timer to determine the impact of the on-chip temperature profile on full-chip timing. The updated temperature at each instance can also be fed into a tool to determine the effect of temperature on critical path and clock timing.

About the Authors:
**Ting-Yuan Wang** is a principal engineer at Apache Design Solutions in Mountain View, California. He holds a PhD Degree in Electrical and Computer Engineering at the University of Wisconsin-Madison.

**Margaret Schmitt** is an Area Technical Manager at Apache Design Solutions in Mountain View, California. She holds a Masters Degree in Electrical Engineering from the Technical University of Berlin, Germany.