Many-Core Design from a Thermal Perspective: Extended Analysis and Results

UNIVERSITY OF VIRGINIA DEPT. OF COMPUTER SCIENCE TECH. REPORT CS-2008-05*
APRIL 2008

Wei Huang[‡], Mircea R. Stan[†], Karthik Sankaranarayanan[‡], Robert J. Ribando[*], and Kevin Skadron[‡]
Depts. of [‡]Computer Science, [†]Electrical and Computer Engineering, [*]Mechanical and Aerospace Engineering
University of Virginia, Charlottesville, VA 22904
{*whuang@, mircea@, karthick@cs, rjr@, skadron@cs*}*.virginia.edu*

Abstract

Air cooling limits have been a major design challenge in recent years for integrated circuits. Multi-core exacerbates thermal challenges because power scales with the number of cores, but also creates new opportunities for temperature-aware design, because multi-core designs offer more design parameters than single-core designs. This technical report investigates the relationship between core size and on-chip hot spot temperature and shows that with the same power density, smaller cores are cooler than larger cores due to a spatial low-pass filtering effect of temperature. This phenomenon suggests that designs exploiting low-pass filtering can dissipate more power within the same cooling budget than contemporary designs. We also find a decrease of within-core spatial temperature variation for many-core designs,indicating that thermal analysis can be potentially carried out at the core granularity in the future. This report also presents more results in addition to the DAC paper and includes a derivation of the location of the isotherm in the Appendix. Accurately locating the isotherm is required for an accurate temperature model for homogeneous many-core chips.

1 Introduction

Semiconductor technology scaling presents severe thermal challenges. Area is scaling down faster than power due to limited supply voltage scalability, growing leakage challenges, and non-ideal interconnect scaling [1]. At the same time, the inability to improve single-thread performance without unreasonable power dissipation has led manufacturers to stop trying to extract instruction-level parallelism (ILP) and instead focus on integrating multiple, possibly simpler cores on a single die. A variety of multi-core products are available today, and all high-performance PC and server processors are multi-core and even many-core.

The many-core paradigm, however, is worrisome from a thermal design standpoint. Many-core allows simpler cores with lower power per core than aggressive ILP cores. Yet total power scales up linearly with the number of cores. Assuming that pricing power requires manufacturers to maintain die area and raise clock rate from generation to generation, not only power density but also total power will rise. Hence, with density doubling every generation, constant area, and voltage supply only dropping 2.5% per generation [1], $P = CV^2f$ implies that total power rises at least 50% per generation, assuming continued improvements in circuit delay and hence frequency. Clearly this exponential growth will outstrip the limits of affordable air cooling in a short time. Maintaining Moore's Law within reasonable cooling budgets therefore requires us to find techniques that allow higher thermal design power (TDP) within a fixed cooling budget—*TDP scalability*. (TDP represents the maximum amount of power the cooling system in a computer is required to dissipate.)

This paper shows that the many-core architecture plays a vital role in coping with these scaling challenges. We have two levers. The first is the choice of core sophistication and hence power per core. However, our ability to simplify cores is limited by the nature of the von-Neumann datapath and the per-thread performance that a particular market demands. The second lever is layout: placement of high-power-density elements to maximize thermal uniformity and maximize efficiency of the cooling solution. Layout is independent of core sophistication. Even a single, complex core can be broken into chunks whose placement optimizes thermal uniformity [2]. Our previous work [3] and Etessam-Yazdani et al. [4] have suggested that interleaving high-power-density circuit elements with low-power-density storage elements achieves further benefits by using the interleaved cool elements as virtual, ”lateral heat sinks”.

In particular, this paper focuses on the second lever mentioned above and shows that layout can substantially increase the efficiency of the cooling solution. Specifically:

1. We show that adopting the many-core design style allows *significantly more* total thermal design power (TDP) than traditional designs. On-chip hot spots play an important role in determining cooling cost and the allowed TDP. This in turn may limit the sophistication or speed at which a chip can run. The increase in TDP thus implies that a many-core design can improve performance by simply burning more power without the thermal hazards seen by single- or dual-core designs with the same TDP. Many-core chips show an improved TDP tolerability due to the more uniform power distribution.

[*] This technical report is an extended version of a paper appearing in the Design Automation Conference (DAC), Anaheim, CA, June 2008.

2. We also propose a closed-form analytical model to derive hot spot temperature of a homogeneous many-core design. The model is based on a *spatial temperature low-pass filtering effect* which states that with the same power density, power sources with smaller sizes (which correspond to a higher spatial frequency) are cooler than larger power sources (which correspond to a lower spatial frequency).

3. We also investigate whether it is necessary to consider within-core spatial temperature variations for future many-core designs, and provide guidelines to choose proper thermal modeling granularity.

4. Our analysis can help select optimal core size and core sophistication. Cores that are individually weaker but allow greater TDP may be the right choice. GPUs are one example of such a design philosophy.

Overall, this work suggests that temperature-aware design can gain important benefits from *TDP-scalable* designs and motivates this as a valuable direction for future research.

2 Related Work

The power and thermal analysis of multi-core designs has been considered by other researchers. For example, Sun's Niagara [5] shows the power and energy efficiency of multi-core design. Monchiero et al. [6] explore the multi-core architecture design space and show the power and thermal impacts on design choices. Donald et al. [7] investigate thermal management techniques for multi-core designs. Li et al. [8] perform architecture-level simulations under thermal constraints for multi-core designs. All these prior works analyze the thermal impact and thermal management techniques for different (micro)architectural choices. Orthogonal to them, in this paper we make the case that it is worth further taking advantage of the underlying heat transfer theory and targeting directly the scaling trend of thermal design power for many-core designs. Our present work thus lays the theoretical basis upon which existing architecture analysis and techniques such as those in [6, 7, 8] can be applied. Architecture-level temperature-aware and layout-sensitive floorplan has also been investigated in [2, 9]; but they did not consider the fact that layout can be made independent of core sophistication. Our work suggests that greater benefit in TDP can be achieved by further refining existing temperature-aware layout techniques.

A unique aspect of our work is that we present an analytical model to quantify many-core design hot spot temperature and the allowed thermal design power as a function of the number of cores. In [8], a thermal model is also proposed without further considering the complicated heat spreading within silicon and package, therefore it is crude and hard to extend to many-core designs. Other existing thermal modeling tools such as HotSpot [10] do not provide direct analytical design insights and are not as efficient as a closed-form analytical thermal model.

Regarding the relationship between the power source size and its peak temperature (i.e. the aforementioned spatial temperature low-pass filtering effect), there are also a few existing works on thermal granularity analysis. Etessam-Yazdani et al. [4] investigate the thermal and power granularity issue by experimentally finding the relationship between the size of heat source and the peak temperature. Our previous work [11] presents a preliminary analysis of thermal modeling granularity. Both these prior works provide guidelines for choosing the proper thermal modeling granularity. In this paper, we propose a rigorous analytical approach based on *the analogy between temporal frequency domain electrical circuit analysis* and *the spatial frequency domain thermal circuit analysis.* This provides an easy interpretation of the full details of the underlying heat transfer theory; it also theoretically explains the results reported by detailed finite-element simulations in [4, 11].

The rest of the paper is organized as follows. Section 3 motivates the necessity of performing thermal analysis for many-core design by an example showing the significant increase of thermal design power and thus the performance. Section 4 derives an analytical model to calculate the hot spot temperature for many-core designs quantitatively. Section 5 shows thermal modeling granularity analysis for homogeneous many-core designs. Section 6 points out the limitations of the proposed analytical many-core temperature model and interesting future work directions. Section 7 concludes the paper.

3 A Motivating Example — Relieved TDP in Many-Core Designs

Consider two simple designs with the same silicon area—one has a dual-core architecture, the other has a 220-core architecture, and the cores in each design are homogeneous. If we further assume that half of the chip area is occupied by L2 and lower-level caches that are placed with the cores in a checkerboard fashion (2x2 and 21x21, respectively), and the caches generated negligible power densities compared to the cores, as shown in Fig. 1. The assumption that roughly half the area is occupied by L2 and lower-level caches can be seen from recent designs such as IBM POWER5 [12] and Intel Core 2 Duo [13]. The checkerboard core-cache layout arrangement greatly alleviates core-to-core thermal coupling. Whether the caches are shared or private does not greatly affect the way they can be laid out, and any decent size of cache can be banked and placed almost anywhere on the die (e.g. Intel Itanium2-6M [14]), so a checkerboard layout is a a legitimate option (and as we will show a very good one).

The choice of 220 cores for this example is based on rough estimates of scaling trends. Assuming that we scale from a dual-core design, each of the two cores occupies a quarter chip area, and the other half chip area is low-level caches. We further assume such a design has a typical 20mm×20mm chip size. The many-core designs of the future are likely to use simpler cores than contemporary complex cores, thus we assume a one-time core architecture shift from the dual-core design to many-core design, resulting in a down-scaling of the core area. For example, according to core area data in [15], when scaling from EV6 (i.e. Alpha

21264 [16]) down to EV4, for the same technology node a ∼0.125 scaling factor due to the change in architecture complexity is observed. In addition, due to Moore's Law, a ∼0.5 area scaling factor exists across two generations of CMOS technologies. Combining the two scaling factors above, the size of a single core will possibly become 100 times smaller in less than four generations ($(0.125 * (0.5)^4 \approx 0.01)$. Since ITRS predicts relatively constant chip area across generations, the same 400mm^2 chip would accommodate about 200 such cores. This corresponds to a 20×20 checker board; for our example, we choose 21×21 since an odd number of divisions makes the floorplan more symmetric, thus the number of cores becomes 220. Although this number may seem high, for some applications this is already the norm; for example the nVIDIA GeForce 8800GTX GPU already has 128 simple scalar cores[1], and the next generation seems likely to double that.

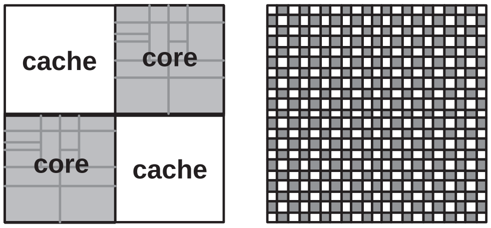

Figure 1. Dual-core and 220-core designs. The cores and the caches are placed in a checkerboard fashion. Shaded areas correspond to cores that dissipate power (Alpha EV6 core without L2 cache is shown as an example in the dual-core floorplan).

If we apply 110W and 1W to each core (i.e. 1W/mm^2 of power density for cores, assuming uniform within-core power distribution and neglecting the cache power) for the dual-core and 220-core designs respectively, we have the same 220W total power for both designs. ITRS predicts increased power density due to non-ideal scaling and 1W/mm^2 is a reasonable hot spot power density for contemporary designs. HotSpot 4.0 [3] is used to find the peak temperature rise with respect to ambient temperature. For a typical heatsink convection thermal resistance of 0.1K/W, we find that the dual-core design has a peak temperature rise of 43.3°C, whereas the 220-core design has only 37.1°C. This is because the 220-core design has much smaller cores and a more uniform power distribution, thus less severe hot spot temperatures as we will see in Section 4.

Alternatively, if we try to find the total thermal design power (TDP) for each design that results in the same peak hot spot temperature rise of 37.1°C, we get 188W TDP for the dual-core design, and 220W TDP for the 220-core design—a significant 17% increase in power. Actually, if a more advanced cooling solution is used (e.g. $R_{convection} = 0.05K/W$), a 25% increase in TDP will be seen in the 220-core design! Even for a design with a moderate cooling solution (e.g. $R_{convection} = 0.5K/W$), we can still see a 5.8% increase in TDP for the 220-core design. The results are listed in Table 1. Note that in Table 1, we fix the TDP of the 220-core design for all values of R_{conv}. Another choice of experiment would be to fix the temperature rise for all three cases, and find the corresponding TDPs for both designs. Because the thermal resistances are independent to power consumption, the temperature rise is strictly proportional to TDP in each case. Therefore we would get the same percentage of TDP gain for 220-core design. This will be obvious in Section 4.2 (i.e. Eq. (7)).

R_{conv}	220-core TDP	temp. rise	equivalent TDP	more TDP tolerated
(K/W)	(W)	(°C)	of dual-core (W)	by many-core (%)
0.05	220	25.4	176	25%
0.1	220	37.1	188	17%
0.5	220	125.0	208	5.8%

Table 1. For the same hot spot temperature, many-core design allows greater thermal design power (TDP). Using a better package will benefit in term of TDP.

This relief in TDP for many-core designs is important. On one hand, if thermal reliability is the major concern in determining the TDP (i.e. not considering energy savings), many-core design's performance can be boosted to a greater degree than single- and multi-core designs by simply burning more power without worrying about thermal hazards that would appear otherwise. The performance boost can be realized for example by scaling up frequency, more silicon integration, or reverse scaling supply voltage, etc. On the other hand, if more performance is not desired (e.g. in real-time applications), the design can burn less power while preserving the same performance in a many-core design, therefore use a cheaper thermal package and reduce system cost without exceeding the maximum allowed hot spot temperature.

With the above example, it is clearly important to quantitatively investigate how the number of cores, or core size, affects the hot spot temperature for the same power density. Performing simulations in HotSpot and other thermal tools does not yield much

[1] http://www.nvidia.com/page/geforce_8800.html

direct insight. Therefore, an analytical model that accurately derives hot spot temperatures of a many-core design is desired.

4 Temperature Model for Homogeneous Many-Core Designs

In this section, we first present the temperature spatial frequency low-pass filtering theory showing the relationship between heat source size and hot spot temperature, and then derive and validate a model calculating the hot spot temperatures for homogeneous many-core designs. Some important implications of the theory are also discussed.

4.1 Spatial Temperature Low-Pass Filter

Spatial frequency is an attribute of any quantity that is periodic in space. It is a measure of how often a quantity is repeated per unit distance. It is defined as $f_s = \frac{1}{\lambda}$, where f_s denotes the spatial frequency, λ is the period or wavelength of the repeating pattern.[2]

For illustration purposes we first consider the traditional temporal frequency-domain analysis for a first-order electrical RC circuit, we then utilize the analogy between the temporal frequency (in s^{-1} or Hz) and the spatial frequency (in m^{-1}) to extend the analysis from time to space as well as from electrical domain to thermal domain.

For an electrical capacitor, C, if the voltage drop between its two terminals is sinusoidal with frequency ω, $V_c(t) = V_0 cos(\omega t + \phi)$, or in exponential form, $V_c = V_0 e^{j(\omega t+\phi)}$. The current flow through the capacitor $I_c(t)$ then is:

$$I_c = C\frac{dV_c}{dt} = C\frac{d}{dt}V_0 e^{j(\omega t+\phi)} = j\omega C{\cdot}V_c \tag{1}$$

And the *electrical* impedance of the capacitor is:

$$Z_C = \frac{V_c}{I_c} = \frac{1}{j\omega C} \tag{2}$$

Consider the electrical circuit in Fig. 2(a), which has a resistor R, a capacitor C and a sinusoidal voltage source $V_s(t) = V_0 cos(\omega t + \phi)$. We know that this circuit is a low-pass filter, that is, the voltage drop across the capacitor tracks the input voltage $V_s(t)$ at low frequency, and is increasingly attenuated at higher frequency. The equivalent impedance of this circuit is $Z_{eq} = Z_R || Z_C = R||(\frac{1}{j\omega C})$, with $Z_{eq} = R$ at DC, and approaching zero at high frequencies, thus the term “low-pass filter”. The resistor R determines the “DC” component of the output voltage, whereas the capacitor determines the “AC” component.

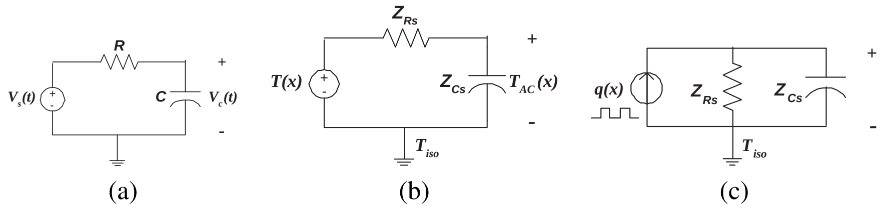

Figure 2. (a) A first-order electrical RC circuit. (b) The Thevenin equivalent first-order thermal spatial “RC" circuit. (c) The Norton equivalent first-order thermal spatial “RC" circuit.

In space, there is also such a “low-pass filtering” effect for temperature distribution. Here, we extend the temporal frequency analysis to the one-dimensional spatial frequency domain. Consider a sinusoidal heat flux (i.e. power density) of $q(x)$, which causes a sinusoidal temperature distribution

$$T(x) = T_0 cos(\omega_s x + \phi) = T_0 e^{j(\omega_s x+\phi)} \tag{3}$$

where $\omega_s = 2\pi/\lambda$ is the spatial radian frequency, and x is the position in the 1-D space. The governing equation of heat transfer is Fourier's Law

$$q(x) = k\frac{dT(x)}{dx} = k\frac{d}{dx}T_0 e^{j(\omega_s x+\phi_s)} = j\omega_s k T(x) \tag{4}$$

where k is the thermal conductivity (the minus “-” sign in Fourier's Law goes away if we define $dT(x)$ in the decreasing direction of temperature, i.e. high temperature minus low temperature). Notice the similarity between Eq. (4) and Eq. (1). This leads us to some quantity analogous to the electrical capacitor in the spatial domain for heat transfer which can be interpreted as a *thermal spatial capacitive impedance*, and write it as

$$Z_{Cs} = \frac{T}{q} = \frac{1}{j\omega_s k} = \frac{1}{j\omega_s C_s} \tag{5}$$

[2] In particular, for our case of thermal granularity analysis in the spatial frequency domain, if the duty factor is 0.5, we can view λ as twice the size of the heat source.

where C_s is defined as *thermal spatial capacitance* (notice that C_s is completely unrelated to the thermal capacitance C_{th} that determines the *transient* heat transfer), and Z_{Cs} is the "thermal spatial capacitive impedance". The unit of both C_s and Z_{Cs} is m^2K/W, which is different from the unit of the thermal resistance (in K/W) usually used. This is due to the fact that we use heat flux, i.e. power density (in W/m^2), instead of power (in W), thus the thermal impedance and resistance in this section are defined as the temperature drop divided by the power density, not by power.

Eq. (5) is used when there is an AC component, with spatial frequency ω_s, in the applied heat flux. In the case where there is only DC heat flux, Fourier's Law leads to the traditional definition of thermal resistance: $Z_{Rs} = \frac{t_{eq}}{k}$, where t_{eq} is the distance from the active silicon surface to the isotherm surface in the package. Also note that this DC spatial thermal impedance also has the unit of m^2K/W, which is consistent with the unit of the AC spatial thermal impedance Z_{Cs}. From the above derivation, naturally we can reach a first-order spatial thermal "R_sC_s" circuit as shown in Fig. 2(b). To make it more comprehensible, Fig. 2(c) shows a more intuitive Norton equivalent circuit of Fig. 2(b). The heat flux generated by the active silicon layer is written as $q(x)$, which models the non-uniform distribution of power density across the chip. The DC component in the spatial temperature distribution is determined by Z_{Rs}, whereas the AC component is determined by Z_{Cs}. In addition, the total equivalent thermal spatial impedance is

$$Z_{eq_s} = Z_{Rs} || Z_{Cs}. \tag{6}$$

If we plot the Bode plot of Z_{eq_s} with respect to the spatial frequency ω_s in Fig. 3, we can see that for low spatial frequencies (power sources with large dimensions), the thermal impedance is close to the DC component, that is the lumped $R_{th} = t_{eq}/(kA)$ that we usually see (A is the corresponding vertical heat conduction area). But for high spatial frequencies (power sources with small dimensions), the impedance attenuates to smaller values due to the presence of the thermal spatial "capacitance". This explains the spatial temperature low-pass filtering effect—structures with tiny dimensions have lower peak temperature comparing to their larger counterparts applied with the same power density.

Intuitively, a tiny heat source even with a high power density does not significantly increase the total power dissipation that the package has to remove, thus temperature rise in the package is almost negligible. On the other hand, a large heat source with high power density results in significant rise in total power dissipation, which in turn leads to significant temperature rise at the heat sink and the heat spreader, hence the increase of average and peak silicon temperatures.

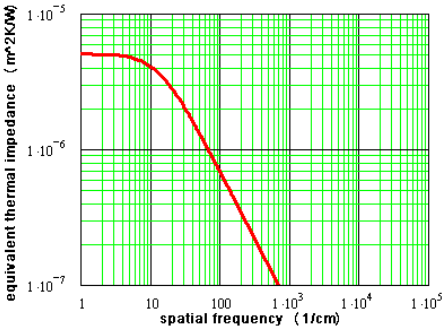

Figure 3. The thermal spatial "RC" circuit is low-pass filter in the spatial frequency domain (Both axes are in log scale).

In [4], Etessam-Yazdani and Hamann et al. also reached a similar spatial frequency domain low-pass filter characteristic curve for square wave power distributions by experimental simulations in finite-element tools. Notice that a $2/\pi$ scaling factor needs to be multiplied to the radian frequency in the above derivations to account for the peak temperature difference between a sinusoidal input power density pattern and a square-wave pattern in real designs.

Because the heat transfer in x and y lateral directions are orthogonal, which is determined by the 2-D form of Fourier's Law, the above derivations can also be easily extended into two-dimensional space with similar results.

One limitation of the above analysis is that it takes into account the lateral spatial temperature gradient, but not the vertical gradient. A more accurate analysis would be using multiple R_sC_s ladders (i.e. dividing each layer vertically into multiple sub-layers), or ideally, distributed thermal spatial thermal R_sC_s circuit. Fig. 4 shows the comparison between the proposed granularity analysis (3-ladder spatial R_sC_s circuit) and ANSYS simulations for different heat source sizes. Note that the spatial frequency and equivalent thermal impedance are both normalized.

As can be seen in Fig. 4, as long as the heat source size is about ten times greater than the isothermal thickness, the thermal resistance can be calculated using conventional $R_{th} = t/(kA)$. For smaller heat sources, the spatial temperature low-pass filtering effect kicks in, and the effective thermal resistance is much less. This means that *tiny heat sources are not necessarily hot spots even with very high power densities.* For example, assuming the isotherm thickness is 4mm, for a heat source of 0.1mm size, we have a normalized spatial frequency of 40, which corresponds to 0.045× the peak resistance from Fig. 4, resulting in 0.045× peak temperature rise. In other words, if a large heat source leads to 100°C temperature rise, this 0.1mm heat source with the same power density only contributes to 4.5°C temperature rise. This explains why some high power density tiny structures, such as clock buffers, do not necessarily become local hot spots. It is obvious that the low-pass temperature filtering effect for the relationship between heat source size and peak temperature is strong.

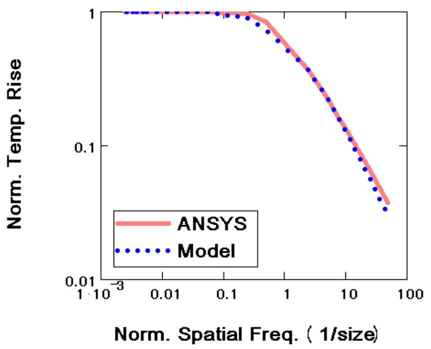

Figure 4. Comparison of 3-ladder thermal spatial RC model and ANSYS simulation for different heat source sizes (Both axes are in log scale).

4.2 Temperature Model of Homogeneous Many-Core Designs

In this section, we use the above spatial temperature low-pass filtering theory to derive an analytical model for hot spot temperature of many-core designs.

If we consider that all the cores are homogeneous and each core in a many-core design is a uniform heat source, the size of a core directly relates to the hot spot temperature of the chip. Thus a first-order many-core hot spot temperature model as a function of number of cores (n) can be written as follows,

$$T_{max} = \text{TDP} \cdot \left(R_{conv} + \frac{t_{si} - t_{iso}(n)}{kA} + \frac{t_{iso}(n)}{k} \frac{1}{A(1 - \text{Ca}(n))} \left| \frac{1}{1 + j\omega_s \tau_s} \right| \right) \tag{7}$$

where TDP is the total thermal design power, R_{conv} is the heatsink-ambient convection thermal resistance. A is the total chip area, Ca(n) is a function evaluated in the range of (0,1) that models the fraction of chip area occupied by L2 and lower-level caches. Therefore $L_{core} = \sqrt{A \frac{1 - \text{Ca}(n)}{n}}$ is the size of one core. The term $\left| \frac{1}{1 + j\omega_s \tau_s} \right|$ models the low-pass temperature filtering effect with $\omega_s = \frac{2\pi}{2L_{core}}$ is the spatial frequency and $\tau_s = 0.5 R_s C_s = 0.5 t_{iso}(n)$, where the 0.5 factor accounts for difference of the aforementioned distributed vs. lumped RC constants.

Eq. 7 states that the peak temperature of a homogeneous many-core system can be calculated by adding the temperature rise from the air to isotherm surface inside the package (the first two terms) to the temperature rise from isotherm surface to the silicon surface (the third term). [17] observed this composition as well, but here the third term is governed by the presented spatial temperature low-pass filtering effect caused by the small core size. $t_{iso}(n)$ is the isotherm thickness that is a function of number for cores, and t_{si} is the total equivalent silicon thickness that combines the thickness of TIM, spreader and heatsink.

(The analytical derivation of $t_{iso}(n)$ is elaborated in the Appendix.)

To validate the accuracy of Eq. 7, we compare to the HotSpot results presented in Section 3 for the dual-core and 220-core designs with $R_{conv} = 0.1K/W$. Here, Ca$(n) = 0.5$ since half of the chip area is occupied by the caches, $A = 441$mm^2, TDP=220W, k=100W/(m-K) for silicon, and $t_{si} = 2.7$mm according the default package values in HotSpot. Because the derivation of Eq. 7 is completely independent of HotSpot and HotSpot 4.0 has been extensively validated against ANSYS [3], HotSpot makes a good reference. Table 2 shows Eq. 7 to be accurate, especially if the core number n is large (0.5% error). The more noticeable error for the dual-core design (11.1% error) is caused by the fact that for the two large and hot cores, the assumption that the center sink-to-air surface is isotherm is not accurate. More detailed thermal simulation is needed to decide the model's error for the case of a few cores and good cooling package (i.e. small R_{conv}). However, since the model targets many-core designs (mostly with tens or hundreds of cores), the error in designs with a few cores is not critical.

4.3 Implications of the Model

Fig. 5 shows a plot of hot spot temperature vs. number of cores from Eq. 7 with half the chip area occupied by cooler caches. 200W and 250W thermal design powers are applied respectively. When the core number approaches infinity, i.e. truly uniform power distribution across the chip, a uniform chip temperature is obtained and there are no particular hot spots. When the number of cores is about 2-4 or greater than thousands, the hot spot temperature does not change much. For the range of ten to a few hundred cores, the hot spot temperature is quite sensitive to number of cores, therefore, potential opportunity exists in this region for optimization between thermal design power, performance, and package cost. From Fig. 5, we can also confirm what we previously

Sink Conv R_{th} (K/W)	model (220-core) (C)	HotSpot (220-core) (C)	error (%)
0.05	26.2	25.4	3.1%
0.1	37.3	37.1	0.5%
0.5	125.3	125.0	-0.2%
Sink Conv R_{th} (K/W)	model (dual-core) (C)	HotSpot (dual-core) (C)	error (%)
0.05	37.1	31.7	17.0%
0.1	48.1	43.3	11.1%
0.5	136.1	132	3.1%

Table 2. Comparison of the proposed model (Eq. 7) with HotSpot. For many-core design, the model is very accurate. The model is less accurate for the case of fewer cores.

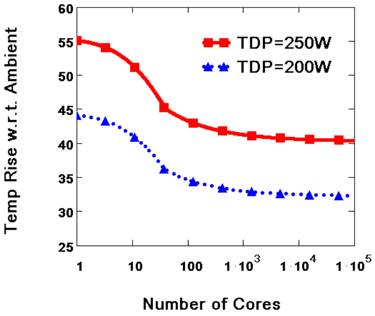

Figure 5. Chip peak temperature as a function of number of cores for TDP=200W and 250W, with L2 caches occupying half chip area.

observed from HotSpot simulations in Section 3—many-core design at a higher thermal design power (250W in this example) can have the same hot spot temperature as a fewer-core design which can tolerate much less thermal design power (200W).

In the results presented so far, cache area is fixed to be half the chip area (i.e. Ca$(n) = 0.5$), which may not be a viable assumption. In order to investigate the impact of cache area to the hot spot temperature, we look at the case where cache area fraction of the chip increases as a function of core numbers, assuming the same total chip area. The increase comes from the fact that more cores are integrated to achieve an increase in throughput, and more throughput usually means more caches are needed. Table 3 shows the cache increase at two different rates.

scenarios	# of cores=2	10	100	1000
Cache area fraction (faster incr.)	0.5	0.67	0.88	0.90
Cache area fraction (slower incr.)	0.5	0.53	0.75	0.90

Table 3. Cache area increases as more cores are integrated into the same chip area. That is, we are varying Ca(n) in Eq. 7.

This table is somewhat arbitrary since cache area in multi- and many-core designs is still an open question and there is no consensus analytical model addressing this. On the other hand, it qualitatively reflects a possible trend of increasing L2 and lower-level cache area as more cores are integrated. For single-core or dual-core designs, the fraction of area occupied by cache is about 0.5. With more and more cores, the cache area approaches about 90% of the chip area. Keep in mind that this cache area model is only used to illustrate the importance of caches that are placed between cores as thermal ”buffer zone” on the hot spot temperature. If desired, other cache area models can also be applied to the model in Eq. 7. Another point to note is that, in this case, the core area and cache block area are not necessarily the same, so the formula for $t_{iso}(n)$ in the Appendix is not always valid. Here, we simply set the isotherm thickness the same as the equivalent total silicon thickness (t_{si}), just to qualitatively show the impact of varying cache area.

Fig. 6 shows the hot spot temperature as a result of different cache area increase rate as more cores are integrated into the chip. For the same number of cores, more *core* power density is experienced by the design with more cache area since the same TDP is applied to cores with less total core area. This increase in power density first outweighs the bigger cooling cache ”buffer zone” and results in a rapid increase in hot spot temperature! As the number of cores increase, the core size becomes smaller and the low-pass

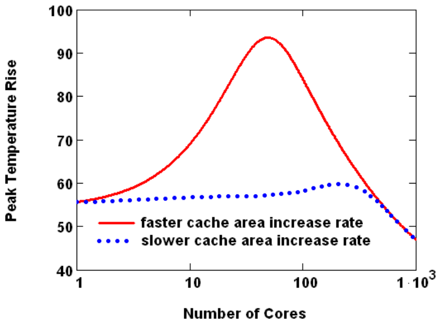

Figure 6. Peak temperature as a function of core numbers for different cache area increase rates (TDP=250W, R_{conv} =0.05K/W).

filtering effect starts to dominate. Therefore the hot spot temperature begins to drop. For the case where the cache area increases slower as number of cores increases, we observe a similar trend but to a lesser degree. This example shows that smaller cores can reduce effective thermal resistance as well as increases the power density (for the same total TDP). Carefully balancing these two opposing factors is important to avoid thermal emergencies in many-core designs.

Another observation from the model in Eq. 7 is that when the package-to-ambient convection thermal resistance (R_{conv}) dominates the silicon-to-package thermal resistance, the on-chip peak temperature is not sensitive to the number of cores. It is instead determined more by the total power of the chip, which confirms a similar observation in [8]. This is the case for most low-cost designs that usually have only natural convection as the cooling method. There has also been a fallacy to use power density as a proxy of temperature. Eq. 7 shows that the hot spot temperature is determined not just by the power-density-related second and third terms (proportional to TDP/A). The total-power-related first term also plays an important role. For example, in a low-cost many-core design, it is possible that the total power is fixed but the increase in number of cores leads to more power density for each core due to the increase in secondary cache area. However, the low-pass filtering effect combined with the dominant package-to-air thermal resistance may yield a lower peak temperature.

5 Spatial Temperature Variation Within a Core

Another interesting thermal topic related to many-core design is to find out when it is practical to perform thermal analysis at the core granularity rather than at the within-core block granularity.

Historically, a single temperature was used to model the entire die. Similarly, in the HotSpot block model, the center temperature of a block is used to represent the entire block, while ignoring the temperature variations at finer granularities within the block. Whether this simplification is legitimate or not depends on several factors—the design level one works at, the available granularity of power estimation, the desired design complexity and accuracy tradeoff, etc. For example, a temperature model at the transistor level is obviously less useful for designers who can only estimate power at the architecture level. Even if the designer has access to transistor level power numbers, performing thermal simulations for the entire system at the transistor level results in prohibitive computation overhead. However, whenever a more detailed estimation of power distribution within a functional block is available, it is still important to know whether ignoring localized heating within the block would miss potential local within-block hot spots that impact the reliability and performance of the entire design.

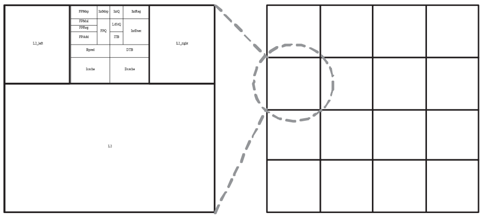

Figure 7. Left: A single-core ALPHA 21364. Right: an example of 16 scaled ALPHA 21364 cores in the same chip area. Scaling down of each core allows more cores to be fit in the same chip area.

For illustrative purpose, here we present a thermal analysis for an imaginary many-core design which consists of many replicas of the scaled Alpha EV6 cores. Each core has the same structure and floorplan as the original single-core design, only the area of each unit is scaled so that same chip area can accommodate many copies of them. An example floorplan for a 16-core design is shown in the right half of Fig. 7. Note that the L2 region in each core can be core-private or banks of a logically global L2 caches.

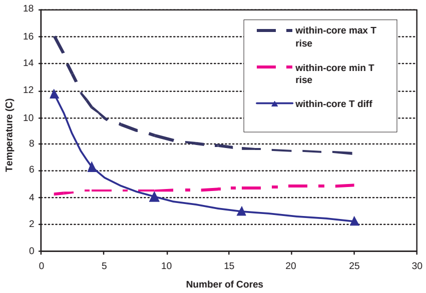

Figure 8. Within-core temperature variation as a function of number of homogeneous scaled cores that are fit in the same chip area, floorplan is similar to Fig. 7.

We assume that each core runs *gcc* and uses the same power trace. We also scale the power for each core by number of cores. We then use HotSpot and its default configurations to find the maximum and minimum temperatures within each core for different number of cores. The results are plotted in Fig. 8. As can be seen, the within-core temperature variation decreases quickly from around 12°C for a single-core design to around 2°C for a 25-core design. This indicates that for future many-core designs, it is very likely that thermal analysis can be carried out at the core granularity.

6 Limitations and Future Work

It seems that with the proposed model in Eq. 7, it is not necessary to run more detailed HotSpot-like thermal simulations any more. This is not usually the case. The model presented in this paper still has the following limitations:

1. All cores are assumed to be homogeneous. For heterogeneous many-core designs, models that extends the proposed model or detailed thermal simulations are needed.

2. Power distribution within each core is assumed to be uniform. Although temperature variation within each core becomes more negligible as discussed in Section 5, sometimes it is still necessary to confirm that is really the case.

3. All the cores are assumed to dissipate the same power all the time. This is not true in real designs. In addition, many small cores also mean more dependence on a core's neighbors due to lateral heat spreading. Therefore HotSpot-like simulations are needed to find the exact core temperatures when workloads are assigned to different cores dynamically, or when dynamic voltage scaling or clock gating are applied to different cores. However, the proposed model already takes care of the worst-case combination of core activities and is enough to decide TDP and package choices.

4. Different types of cooling solution may have different levels of impact on TDP. The analysis in this paper should further be extended to cooling solutions such as better interface material, better heat spreader, heat pipe and micro-channel liquid cooling.

5. As mentioned earlier, the analytical model in this paper has more error when the number of cores are small. This is due to the fact that the assumption of isotherm center heatsink surface may not be valid for small number of cores dissipating a lot of power.

6. We have not considered the communication network across cores. The interconnect density, and hence the area, power and performance overheads all go up as number of cores increases. Interconnect networks' power and thermal impacts need to be carefully considered in many-core designs.

All the above limitations are important to address and will be interesting future work.

7 Conclusion

It is important to understand how many-core design options interact with thermal and power limits of modern scaled CMOS technologies in order to maintain Moore's Law scaling. In this paper, we present a theoretical analysis of the relationship between core size and peak temperature, and propose a quantitative model to estimate many-core chip hot spot temperature as a function of number of cores. We find that many-core design has the potential advantages of significantly relieving the thermal design power constraint, and hence a performance boost or cheaper system cost.

Appendix: Finding the Isotherm Thickness, $t_{iso}(n)$

Clearly, the isotherm thickness (the distance between the silicon active surface to the isotherm surface inside the package) depends on the number of homogeneous cores, in other words, the power distribution uniformity of the design. In a checkerboard layout where cores are interleaved with cool caches, more cores means more uniform power distribution. In the extreme, when number of cores goes to infinity, the design has a truly uniform power distribution.

To find the location of the isotherm surface, we need to first define what an isotherm surface is. In reality, there is never a true isotherm in the package due to the non-uniform power distribution and the boundary effects etc. But we can approximate a surface to be isotherm as long as the temperature difference therein is less than a certain percentage of the maximum silicon surface temperature difference. For example, if the maximum on-chip temperature difference is 50C, and we want a thermal resolution of 0.5C (1% error), we can say a surface is isothermal as long as its temperature difference is equal to or less than 0.5C.

In order to find an analytical relationship between the number of cores and the isotherm thickness, we apply the checkerboard floorplan on the silicon surface with n square cores interleaved with n lower-level cache layout blocks with the same area. Thus, the area of a core or a cache block is $a = A/(2n)$, where A is total chip area. Furthermore, we vertically divide the thickness of the die into N horizontal layers, where N is a big number and the thickness of each layer is $t = t_{si}/N$, where t_{si} is the equivalent silicon thickness including all package components. For simplicity, we set the temperature at the bottom silicon surface to be zero. The equivalent thermal resistor network for each layer is thus a uniform grid where each core node is connected to four cache nodes by four lateral thermal resistors R_{lat}, and vise versa. The vertical thermal resistance of each node is $R_{ver} = t/(k_{si}a)$.

Assuming all cores dissipate the same power, and caches dissipate no power, and further neglecting the effect of the boundary of the chip (this is legitimate if number of cores, n, is big), we can see that all the cores have the same (higher) temperature, and all the cache blocks have the same (lower) temperature. This implies that we can model the entire grid by only one core and one cache block, with proper scaling of the lateral thermal resistance. The vertical thermal resistors remains the same. Stacking all the layers together, we can reach an equivalent thermal circuit like Fig. 9, with $r_2 = R_{ver}$ and $r_1 = \frac{1}{4}R_{lat}$.

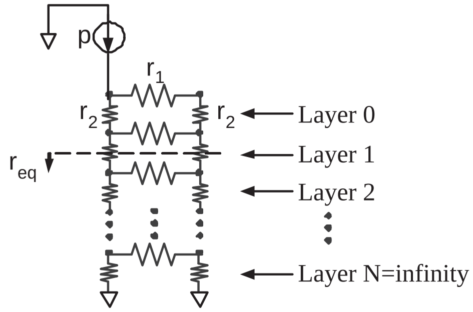

Figure 9. Model the silicon and package by stacking a large number of ladders each representing a thin layer of equivalent silicon.

Finding the right expression for R_{lat} needs more rigorous analysis. A common way is to use $L/(k_{si}Wt) = 1/(k_{si}t)$ for a square block, where L and W are the length and width of each block and $L = W$ for square blocks. The assumption behind this formula is that the heat source is on one side of a block, and it does not take into account the fact that the power is instead dissipated uniformly within the block. To more accurately model R_{lat}, Fig. 10 is used. Here, half symmetry is used to simplify the analysis since the heat transfer in the left half and the right half of the block is the same. Thus only the right half of the block is modeled in detail, whereas the left half is outline with dash lines. The boundary between the two halves can therefore be modeled as adiabatic. We further divide the right half into N slices (This N is different from the number of vertical layers mentioned above). For each slice, the lateral resistance is R/N, where $R = 1/(k_{si}t)$, since it models the lateral heat transfer from one slice to the next slice. Each slice has a heat source of P/N, where P is the power dissipation of the right half of the block. With these settings and assuming the right boundary of the block is at zero temperature, we can use superposition to derive the center temperature of the block.

According to Fourier heat conductance law, the contribution to the center temperature by the lateral heat transfer from slice $(N-k)$ is given by $k*(R/N)*(P/N)$. According to superposition, the center temperature is the sum of contributions from all N slices

$$T_{center} = \lim_{N\to\infty}\frac{P}{N}\frac{R}{N}\sum_{k=1}^{N}k = \lim_{N\to\infty}\frac{PR}{N^2}\frac{1}{2}N(N+1) = \frac{1}{2}PR$$

Thus the equivalent thermal resistance that characterizes the difference between center temperature and the boundary temperature of a square block is $T_{center}/P = \frac{1}{2}R$. (Note that this is only valid when the vertical heat transfer is not considered (i.e. the layer thickness is infinitesimal) and the four boundaries are at the same temperature. For models like HotSpot, neither of these conditions exists, and it turns out that applying this result in HotSpot yields more error than just using the common $R_{lat} = L/(k_{si}Wt)$.)

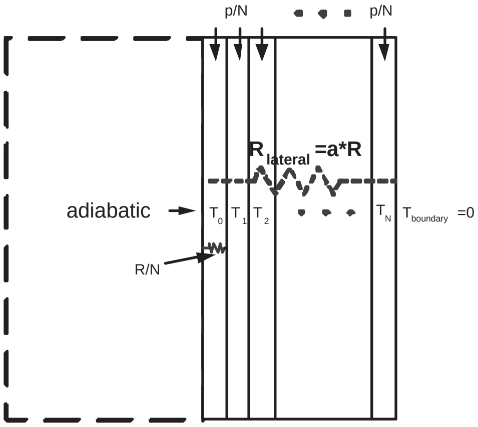

Figure 10. Use superposition and half symmetry to find the scaling factor of lateral resistance.

With the above analysis, we see that $r_1 = \frac{1}{4} R_{lat} = \frac{1}{8} \frac{1}{k_{si} t}$. Since the ladder has infinite number of layers (N is big), we can simplify the ladder into Fig. 11, where the rest of the ladder is replaced by an equivalent resistance r_{eq}. Because the ladder is infinite, r_{eq} can be found by noticing that r_{eq} is the same whether or not including another level of the ladder, thus,

$$r_1 || (r_{eq} + 2r_2) = r_{eq}$$

solving this for r_{eq}, we get

$$r_{eq} = -r_2 + \sqrt{r_2^2 + 2r_1 r_2}$$

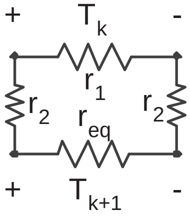

Figure 11. Equivalent thermal circuit for one layer.

If we denote the temperature difference between the core and the cache at the silicon active surface as T_0, and the temperature difference within Layer k as T_k and so on, from Fig. 11, we can see that

$$T_{k+1} = \frac{r_{eq}}{r_e q + 2r_2} T_k$$

Or, unrolling the recursion,

$$T_k = \left(\frac{r_{eq}}{r_e q + 2r_2} \right)^k T_0$$

Thus, finding the isotherm location is the same as solving for k for a given thermal resolution η (e.g. 1% or 0.01), where

$$\eta = \frac{T_k}{T_0} = \left(\frac{r_{eq}}{r_e q + 2r_2} \right)^k$$

so,

$$k = \frac{\ln \eta}{\ln \left(\frac{r_{eq}}{r_e q + 2r_2} \right)}$$

Once k is found, the actual isotherm location is

$$t_{iso} = \frac{k}{N} t_{si}$$

Fig. 12 shows the isotherm thickness (normalized to t_{si}) as a function of number of cores. The results from ANSYS are also plotted as a comparison. It is obvious that the derived model for t_{iso} as a function of number of cores is quite accurate, especially

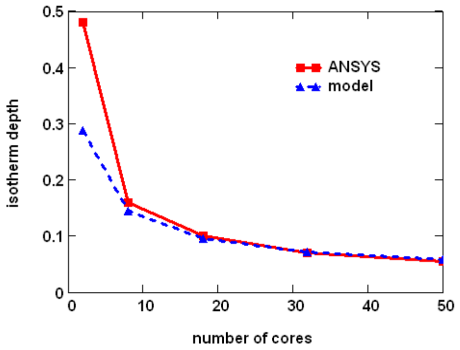

Figure 12. Normalized isotherm thickness as a function of number of cores. Comparison with ANSYS simulations shows that the derived model for $t_{iso}(n)$ is quite accurate.

for large number of cores. The larger error at a few cores is mostly because the boundary effect cannot be neglected as most cores are along the chip boundaries.

One last caveat of the above isotherm thickness model is the assumption that the interface between package and air is isothermal. This is mostly true for large number of cores. But for a smaller number of cores each dissipating a lot of power, this assumption may not be valid, and the actual isotherm surface may be beyond the thermal package. In that case, the best approximate is to use t_{si} as the isotherm thickness or use HotSpot to find the temperatures.

Acknowledgement

This work was supported in part by NSF grant nos. CNS-0551630, CNS-0509245; a MARCO/IFC grant; and a grant from Intel MRL.

References

[1] The international technology roadmap for semiconductors (ITRS), 2005.

[2] K. Sankaranarayanan, S. Velusamy, M.R. Stan, and K. Skadron. A case for thermal-aware floorplanning at the microarchitectural level. *The Journal of Instruction-Level Parallelism*, vol. 7, October 2005.

[3] W. Huang, K. Sankaranarayanan, R. J. Ribando, M. R. Stan, and K. Skadron. An improved hotspot block-based thermal model with granularity considerations. In *Workshop on Duplicating, Deconstructing, and Debunking (WDDD), in conjunction with Intl. Symp. on Computer Architecture (ISCA)*, June 2007.

[4] K. Etessam-Yazdani, H. F. Hamann, and M. Asheghi. Impact of power granularity on chip thermal modeling. In *Proc. of 10th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM)*, June 2006.

[5] P. Kongetira, K. Aingaran, and K. Olukotun. NIAGARA: A 32-way multithreaded SPARC processor. *IEEE Micro*, 25(2):21–29, March-April 2005.

[6] M. Monchiero, R. Canal, and A. Gonzalez. Design space exploration for multicore architectures: A power/performance/thermal view. In *Proc. of ACM International Conference on Supercomputing (ICS)*, June 2006.

[7] J. Donald and M. Martonosi. Techniques for multicore thermal management: Classification and new exploration. In *Proc. of ACM International Symposium on Computer Architecture (ISCA)*, June 2006.

[8] Y. Li, B. Lee, D. Brooks, Z. Hu, and K. Skadron. CMP design space exploration subject to physical constraints. In *Proc. of IEEE Conf. on High-Performance Computer Architecture (HPCA)*, 2006.

[9] Y. Han, I. Koren, and C. A. Moritz. Temperature-aware floorplanning. In *Proc. of Workshop on Temperature-Aware Computer Systems (TACS)*, 2005.

[10] K. Skadron, K. Sankaranarayanan, S. Velusamy, D. Tarjan, M. R. Stan, and W. Huang. Temperature-aware microarchitecture: Modeling and implementation. *ACM Transactions on Architecture and Code Optimization*, 1(1):94–125, March 2004.

[11] W. Huang, M. R. Stan, K. Skadron, S. Ghosh, S. Velusamy, and K. Sankaranarayanan. Hotspot: A compact thermal modeling methodology for early-stage vlsi design. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 14(5):501–513, May 2006.

[12] J. Clabes et al. Design and implementation of the power5 microprocessor. In *Proc. of ISSCC*, Febuary 2004.

[13] N. Sakran et al. The implementation of the 65nm dual-core 64b Merom processor. In *Proc. of ISSCC*, Febuary 2007.

[14] S. Rusu, H. Muljono, and B. Cherkauer. Itanium 2 processor 6M: Higher frequency and larger L3 cache. *IEEE Micro*, 24(2):10–18, March-April 2004.

[15] R. Kumar, D. M. Tullsen, N. P. Jouppi, and P. Ranganathan. Heterogeneous chip multiprocessors. *IEEE Computer*, 38(11):32–38, May 2005.

[16] R. E. Kessler. The Alpha 21264 microprocessor. *IEEE Micro*, 19(2):24–36, March-April 1999.

[17] Y. Li, D. Brooks, Z. Hu, and K. Skadron. Performance, energy, and thermal considerations for SMT and CMP architectures. In *Proc. High Performance Computer Architecture (HPCA)*, pages 71–82, February 2005.