Performance Optimal Speed Control of Multi-Core Processors Under Thermal Constraints

Vinay Hanumaiah, Sarma Vrudhula and Karam S. Chatha
Computer Science and Engineering Department, Arizona State University, Tempe 85281
{vinayh, vrudhula, kchatha}@asu.edu

***Abstract*—Advances in chip-multiprocessor processing capabilities has led to an increased power consumption and temperature hotspots. Maintaining the on-chip temperature is important from the power reduction and reliability considerations. Achieving highest performance while maintaining the temperature constraint is a challenge. We develop analytical solutions for the optimal control of frequencies for each core in a chip-multiprocessor. The objective is to reduce the *makespan* or the latest task completion time of all tasks. We show that the optimal frequency policy is *bang-bang* when the temperature constraint is not active and is exponential when the temperature constraint is active. We show that there is a significant improvement in overall throughput with our proposed solution and yet all cores operate under the thermal maximum.**

I. Introduction

It is known that Moore's law is no longer sustainable due to the increasing power consumption of processors in recent decades. Multi-core processors seem to provide the solution by distributing the transistors of a single core implementation over many smaller cores, each with lower throughput and lower power, and exploit thread level parallelism to boost performance, while keeping the total power approximately the same as that of a single core implementation. Increasing the number of cores per die has become the new "scaling" strategy, with the number of cores expected to double every two to three years. This rate of increase in the number of cores will overwhelm the reduction (if any) in power of the individual cores, and consequently high die temperatures are expected to become a major problem for future multi-cores.

Cooling and packaging technologies do not scale with device technology. It would become prohibitively expensive if they were designed to accommodate the maximum power dissipation of modern high performance processors. For this reason, processor manufacturers specify a conservative upper limit called the *thermal design power* (TDP) – which is typically determined by running representative, real-world applications. The responsibility of ensuring that the TDP constraint is not violated is left to *dynamic thermal management* (DTM) [1], [2], which are a collection of techniques (in hardware and/or software) that include DFS (dynamic frequency scaling), DVFS (dynamic voltage and frequency scaling), thermally-aware task distribution, etc., to maximize performance subject to thermal constraints.

This work was supported in part by NSF grant CSR-EHS 0509540, Consortium for Embedded Systems grant DWS 0086, and by a grant from Science Foundation Arizona (SFAz) and Stardust Foundation.

DTM techniques can be classified into online and offline strategy. Online techniques [2], [3] employ local control (DVFS or DFS) determined by the current temperature of the processor. Although online approaches can respond to the current conditions of the processor and provide necessary correction to ensure that the constraints are met, they can be suboptimal because the correction is based on heuristics derived by experimentation. A better approach is to first determine an offline solution [4], [5] where the objective is to determine a globally optimal control policy. The online policy can use real-time measurements of the temperature $T(t)$, compare the model based estimate temperature $T^*(t)$, and make small corrections to the offline solution.

Some of the recent works on determining DTM policies consist of [4]–[6]. In [5], optimal performance for a periodic sequence of tasks under thermal constraints for a single processor is found using DVFS approach with discrete voltage and frequency states. In [6], discrete version of the optimal frequency assignment problem to maximize operation cycles within a given time in multi-cores is obtained using convex optimization. In both the works, a highly simplified thermal model is assumed with just one R and one C and no leakage component of power is considered, which is definitely sub-optimal. Reference [4] describes a continuous approach to obtain analytical solutions of the optimal processor frequency for given sequence of tasks, but only for a single-core processor. It uses a highly detailed and accurate hotspot model [7] and also accounts for leakage.

A. Key Contributions

In this paper, we present a novel solution to the critical problem of determining the optimal time-varying speed profile of each core in a multi-core processor, ensuring that temperature constraints are satisfied at all the operational times. The optimality criterion is to minimize the maximum completion time (*makespan*) over all tasks. We believe that this is the first analytical solution that relates performance to the number of cores, and the power and thermal parameters of the cores using the accurate hotspot model [7]. In addition, this work considers the important leakage dependence on temperature (LDT). Unlike the results in [1], we show that the optimal throttling is an exponentially decreasing function of time rather than a constant speed. Our solution, which is also applicable to heterogeneous cores, can serve as an approximate solution to the discrete optimal speed control.

978-3-9810801-5-5/DATE09 © 2009 EDAA

II. Multi-core System Model

A. Execution Task Model

We consider analysis of tasks whose durations are much greater than the die thermal time constants (tens of milliseconds) and not much greater than the package thermal time constant (few minutes). Class A and class B of NAS benchmarks are examples of such tasks. The reason for choosing such tasks is because longer duration tasks than considered can be easily solved through steady state analysis as it will be a zero order system (all capacitances are ignored), whereas the analysis of shorter duration tasks requires much higher order systems and generally not amenable for analytic derivations. They are more suitable for detailed numerical analysis.

We assume that the processor consists of n cores, and each core is assigned a specific task and is run until its completion, i.e., we do not consider task migration among the cores. The task c is characterized by the normalized number of cycles N_c (w.r.t U_{max}) and its power consumption P_c.

B. Power and Thermal Model

Our thermal model is based on Hotspot thermal RC circuit model [7], which uses the well known duality between heat transfer phenomena and electrical circuit phenomena. Our simplified thermal model is based on the observation of the thermal parameters in Alpha 21264 [8]. For this processor, there are 20 functional blocks and the order of the system is 31. Each functional unit has a resistance connected to the package called as vertical resistance, few more resistances connected to the neighboring blocks called as lateral or horizontal resistances and a capacitance to ground. We observe that the lateral resistances are nearly four times that of the vertical resistances and hence can be ignored without significant loss of accuracy [9]. (**NOTE**: This does not remove the possibilities of hotspots as they are caused by the differences in the power densities of neighboring blocks.)

Based on these observations, we form a high-level thermal RC model with n cores and M blocks per core. This is shown in Fig. 1. It is a first order circuit with one package capacitance C_p. R_P and T_p are the package's resistance and temperature respectively. Similarly, $R_{c,i}$, $P_{c,i}$ and $T_{c,i}$ are the vertical thermal resistance, power and temperature of block i in core c, respectively. The effect of LDT as a feedback is also shown in the figure. Since our work load duration is much more than the die thermal time constants (around ten milliseconds) the die capacitances are neglected.

Power consumption $P_{c,i}$ of block i in core c has a dynamic and a static component. While the dynamic power $P_{d,c,i}$ depends linearly on the core c's normalized operating frequency u_c, the leakage or static power $P_{s,c,i}$ varies exponentially with the temperature [8]. Hence the total power $P_{c,i}(u_c, T_{c,i})$ is a function of both speed and the temperature of the block i in core c. We do not make use of DVFS in our work for the reasons that the majority of modern day processors use DVFS with a linear power-speed relationship for DTM purposes [10] and also with the technology scaling, the scope for supply voltage scaling is less.

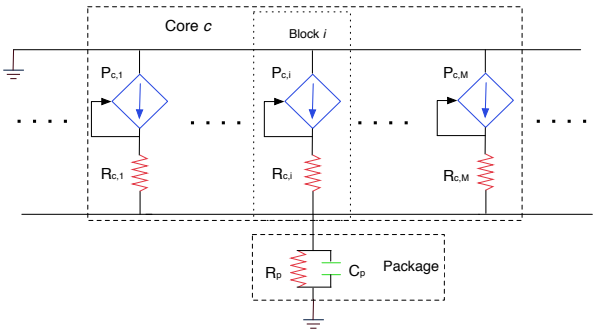

Fig. 1. Simplified multi-core thermal model.

III. Preliminary Results

A. Decoupling of Leakage and On-chip Temperature

In order to decouple the LDT, we need to linearize the LDT. To this effect the operating range of the temperature is divided into two regions (easily extendable to any number) and the exponential relationship is approximated by a piece-wise linear (PWL) relationship [9] as shown below:

$$P_{s,c,i} = \begin{cases} P_{s,c,i,mid} - k_{c,i1}(T_{mid} - T_{c,i}), T_{min} < T_{c,i} \leq T_{mid}, \\ P_{s,c,i,max} - k_{c,i2}(T_{max} - T_{c,i}), T_{mid} < T_{c,i} \leq T_{max}. \end{cases} \quad (1)$$

$\forall i \in 1, \ldots, M, \forall c \in 1, \ldots, n$. $k_{c,i1}$ and $k_{c,i2}$ represent the slope of the leakage vs temperature for block i of core c in the corresponding region of the PWL model. T_{min} and T_{max} denote the ambient temperature and the maximum allowed temperature of the die respectively. T_{mid} is chosen suitably so that the linear model is close to the exponential LDT model. **Note:** Since the range of the operational temperature mostly falls in the upper half of the PWL model, for the sake of simplicity, we will assume that there is just one $k_{c,i}$ in our derivations. However, the numerical results obtained use the appropriate $k_{c,i}$ to determine the leakage power.

Now we summarize the results from [9] for decoupling LDT which uses the above linear model. The key observation in this derivation is the fact that the die thermal time constant is three orders of magnitude less than that of the package thermal time constant. Thus for the thermal transients of the order of die thermal time constant, the package temperature appears constant. With this assumption it can be shown [9] that the temperature of block i within core c is given by:

$$T_{c,i}(t) = \zeta_{c,i} T_p(t) + (P'_{s,c,i} + u_c(t) P'_{d,c,i}) R_{c,i} \quad (2)$$

where, $\zeta_{c,i} \triangleq (1 - k_{c,i} R_{c,i})^{-1}$ (leakage coefficient), $P'_{s,c,i} \triangleq \zeta_{c,i}(P_{s,c,i,max} - k_{c,i} T_{c,max})$ (apparent static power), $P'_{d,c,i} \triangleq \zeta_{c,i} P_{d,c,i}$ (apparent dynamic power).

Using the above equation, the circular dependency is removed and $P_{s,c,i}$ is made to depend only on the package temperature T_p as shown below:

$$P_{c,i}(t) = P'_{s,c,i} + u_c(t) P'_{d,c,i} + (\zeta_{c,i} - 1) T_p(t) / R_{c,i} \quad (3)$$

Note: The temperature of each core is still affected by the activity of other cores. However, this dependence is through the package temperature.

B. Computation of Package Temperature

With the PWL approximation to the LDT, we can compute the package temperature based on the simplified high-level thermal model described in Section II-B. Let $P'_s \triangleq \sum_{c=1}^{n}\sum_{i=1}^{M} P'_{s,c,i}$ (Total apparent static power), $P'_{d,c} \triangleq \sum_{i=1}^{M} P'_{d,c,i}$ (Total apparent dynamic power at $\mathbf{u} = 1$) and $G \triangleq \sum_{c=1}^{n}\sum_{i=1}^{M}(\zeta_{c,i} - 1)/R_{c,i}$. From the high-level thermal model described in Section II-B it can be shown that:

$$\begin{aligned}\frac{dT_p(t)}{dt} &= -\frac{T_p(t)}{R_pC_p} + \frac{1}{C_p}\sum_{c=1}^{n}\sum_{i=1}^{M} P_{c,i}(t) \\ &= -\frac{T_p(t)}{R_pC_p} + \frac{1}{C_p}\sum_{c=1}^{n}\sum_{i=1}^{M}\left[P'_{s,c,i} + \frac{(\zeta_{c,i}-1)T_p(t)}{R_{c,i}}\right] \\ &\quad + \frac{1}{C_p}\sum_{c=1}^{n} u_c(t)P'_{d,c} \\ &= -\frac{T_p(t)}{R'_pC_p} + \frac{P'_s + \mathbf{u}^T(t)\mathbf{P'_d}}{C_p} \qquad (4)\end{aligned}$$

where, $R'_p \triangleq R_p/(1-GR_p)$. $T_p(t)$ can be obtained by solving the above first order linear ODE.

IV. Problem Formulation and Solution

We now formulate the problem of optimal speed control for a n core processor with each core being assigned a specific task and whose power profile and number of execution cycles are given. Our objective is to find the optimal time varying speed control for each core such that the latest task completion time is minimal, while satisfying the thermal constraints. Here we have used the fact that the hottest block (one with the largest power $\times$ resistance product block) in a core remains the hottest irrespective of the frequency of its operation.

Notation: Bold face variables represent vectors. $x_c(t)$ represents the activity of core c in terms of number of cycles executed at time t in core c. As task migration is not considered we do not distinguish between a task and core number.

The formulation of the optimization problem is as follows:

$$\min_{\mathbf{u}(t)} \quad t_f = \int_0^{t_f} 1\ dt, \qquad (5)$$

$$s.t. \quad \dot{\mathbf{x}}(t) = \mathbf{u}(t), \qquad (6)$$

$$\mathbf{x}(0) = 0,\ \mathbf{x}(t_f) = \mathbf{N}, \qquad (7)$$

$$\frac{dT_p(t)}{dt} = -\frac{T_p(t)}{R'_pC_p} + \frac{P'_s + \mathbf{u}^T(t)\mathbf{P'_d}}{C_p}, \qquad (8)$$

$$T_p(0) = T_{pi}, \qquad (9)$$

$$\zeta_{c,h}T_p(t) + (P'_{s,c,h} + u_c(t)P'_{d,c,h})R_{c,h} \le T_{max},\ \forall t\ \forall c, \qquad (10)$$

$$\mathbf{0}_{n\times 1} \le \mathbf{u}(t) \le \mathbf{1}_{n\times 1},\ \forall t \qquad (11)$$

In the above formulation, t_f represents the final completion time or makespan of all tasks. (8) states that each task starts at time 0 and finishes by time t_f. (10) represents the constraint that the temperature of the hottest functional unit has to less than T_{max}. (8) is used to compute the package temperature and is same as (4).

The above formulation is a time optimal control problem. It has two state variables $\mathbf{x}$ and T_p, with a variable endpoint t_f and fixed boundary conditions at both the ends for the states, except for T_p. The *mixed control-state point-wise inequality* (10) complicates the solution process. We have used the *direct adjoining approach* [11] to obtain the solution.

In the interest of clarity and lack of space, we have omitted the detailed derivations and present only the final results.

The optimal speed profile for core c is given by:

$$u^*_c(t) = \begin{cases} 1, & 0 \le t \le t_{m,c}, \\ u_{r,c}(t), & t_{m,c} \le t \le t_{e,c}, \\ 0, & t \ge t_{e,c}. \end{cases} \qquad (12)$$

where $t_{m,c}$ and $t_{e,c}$ are the transition times. $u_{r,c}$ is the *singular* speed during the transition time $t_{m,c}$ to $t_{e,c}$.

To obtain $t_{m,c}$ we define $T_{p,max}$ as the package temperature when all the cores are at maximum allowable temperature and running at maximum speed. It is given by,

$$T_{p,max} = [T_{c,max} - (P'_{s,c,h} + P'_{d,c,h})R_{c,h}]/\zeta_{c,h},\ \forall c \qquad (13)$$

Using above definition, we can calculate $t_{m,c}$ as shown below:

$$t_{m,c} = \tau_p ln\left(\frac{T_{p0} - R'_pP'_{max}}{T_{p,max} - R'_pP'_{max}}\right) \qquad (14)$$

Let us define $\alpha_c = \frac{P'_{d,c,h}R_{c,h} + \zeta_{c,h}R'_pP'_{d,c}}{R'_pC_pR_{c,h}P'_{d,c,h}}$, $\beta_i = \frac{P'_{d,i}\zeta_{i,h}}{C_pP'_{d,i,h}R_{i,h}}$ and $\gamma_c = \frac{T_{max} - P'_{s,c,h}R_{c,h} - P'_sR'_p\zeta_{c,h}}{R'_pC_pP'_{d,c,h}R_{c,h}}$. The singular speed $u_{r,c}(t)$ can be computed using:

$$u_{r,c}(t) = u_{r,c,0}e^{-\frac{t-t_{m,c}}{\tau_{r,c}}} + u_{r,c,ss}(1 - e^{-\frac{t-t_{m,c}}{\tau_{r,c}}}) \qquad (15)$$

where $u_{r,c,0} = (T_{max} - \zeta_{h,c}T_{p0} - P'_{s,h,c}R_{h,c})/(P'_{d,h,c}R_{h,c})$ is the initial speed of core c, $\tau_{r,c} = \left(\alpha_c + \sum_{i\in n_a, i\ne c}\beta_i\right)^{-1}$ is the time constant of the throttling curve, $u_{r,c,ss} = \gamma_c/\left(\alpha_c + \sum_{i\in n_a, i\ne c}\beta_i\right)$ is the steady state value of the throttling curve and n_a is the number of active cores. By active cores we mean those cores with tasks scheduled on them.

Task completion time $t_{e,c}$ for a core c is computed by calculating the time at which the area under the speed curve u_c equals N_c , i.e.,

$$\begin{aligned} t_{m,c} &+ [u_{r,c}(t_{m,c}) - u_{r,c,ss}]\tau_{r,c}\left(1 - e^{-\frac{t_{e,c}-t_{m,c}}{\tau_{r,c}}}\right) \\ &+ u_{r,c,ss}(t_{e,c} - t_{m,c}) = N_c \qquad (16)\end{aligned}$$

This equation needs to be solved numerically to obtain the value of $t_{e,c}$. The optimal task completion time t_f is then given by $t_f = \max_c(t_{m,c} + t_{e,c})$

Note: Every task completion enables higher speeds in remaining cores due to decrease in the number of active cores. Thus speeds have to recomputed at every task completion.

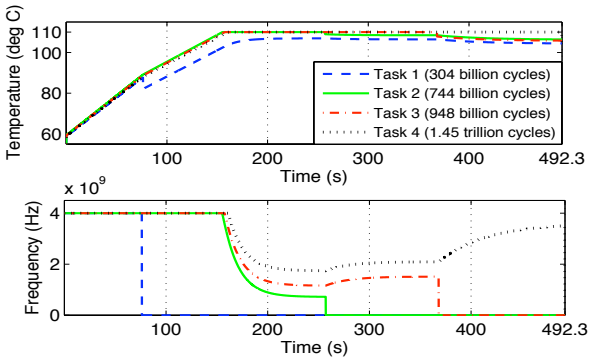

Fig. 2. Thermal and speed profiles for the optimal throttling policy

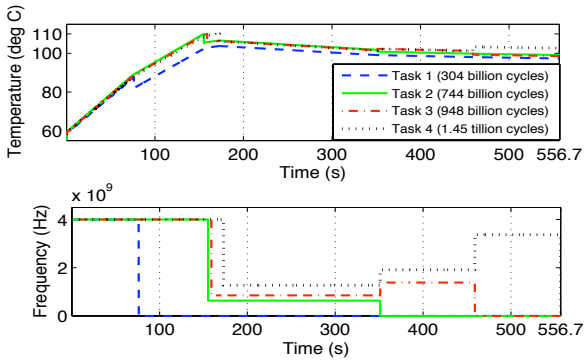

Fig. 3. Thermal and speed profiles for the constant throttling policy

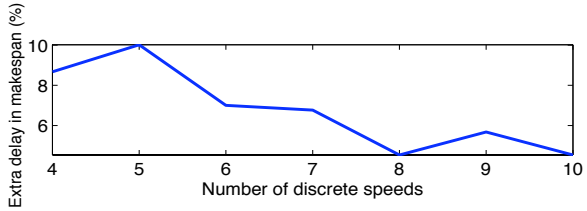

Fig. 4. Plot of extra delay in makespan due to discretization.

V. Experimental Results

A. Optimal vs Constant Throttling

We first obtained the simplified Hotspot model parameters (Section II-B) for the single core Alpha 21264 [8] processor and replicated it to form the multi-core model. Power numbers were obtained from PTscalar tool [8] for SPEC CPU2000 benchmarks: crafty, gcc, galgel and bzip2. We allowed a maximum temperature of 110 °C and total power consumption of 130 W. The maximum clock frequency was set to 4 GHz.

Fig.2 shows the result of the optimal control policy for the case of four cores. The minimum makespan is found to be 492.3 s for the task durations mentioned in the figure. The optimal policy ensures that the makespan is minimized while satisfying the temperature constraints. We compare our optimal policy against the constant throttling policy [1] shown in Fig.3. The constant throttling policy simply throttles each core to its steady-state speed $u_{r,c,ss}$ (see (15)). Constant throttling policy resulted in a makespan of 556.7 s. Thus giving the optimal policy an improvement of 13.1%.

B. Discrete Approximation to Optimal Control

We use the optimal control policy to derive a discrete control policy which can serve as an approximation to discrete speed control. We evaluated the extra delay encountered in the makespan completion times due to discretization for a set of discrete speeds. Fig.4 shows the variation of delay with the number of discrete speeds. We note that the makespan delay is less than 10% and can serve as a good approximate method to discrete speed control.

VI. Conclusion

Advent of chip multiprocessors have resulted in high processing powers. However, with the ever increasing power dissipation and power density, thermal issues are becoming very significant with multi-cores. In this paper, we have for the first time, derived optimal speed control policy using accurate hotspot model that minimizes the makespan of parallel, but non-identical tasks. Our method ensures that the temperature constraints are always satisfied and we demonstrate this with our experimental results. We have shown that our optimal throttling policy can be used to obtain an approximate solution to the optimal policy with discrete frequency states. Our work finds application in early phase design space exploration and offline optimal frequency allocation.

References

[1] A. Cohen, F. Finkelstein, A. Mendelson, R. Ronen, and D. Rudoy, "On estimating optimal performance of cpu dynamic thermal management," *IEEE Comput. Archit. Lett.*, vol. 2, no. 1, pp. 6–9, 2003.

[2] D. Brooks and M. Martonosi, "Dynamic thermal management for high-performance microprocessors," in *Proc. HPCA*, 2001, pp. 171–182.

[3] K. Skadron, T. Abdelzaher, and M. R. Stan, "Control-theoretic techniques and thermal-RC modeling for accurate and localized dynamic thermal management," in *Proc. HPCA'02*, 2002, pp. 17–28.

[4] R. Rao and S. Vrudhula, "Performance optimal processor throttling under thermal constraints," in *Proc. CASES*, 2007, pp. 257–266.

[5] S. Zhang and K. S. Chatha, "Approximation algorithm for the temperature-aware scheduling problem," in *Proc. ICCAD*, 2007, pp. 281–288.

[6] S. Murali, A. Mutapcic, D. Atienza, R. Gupta, S. Boyd, and G. D. Micheli, "Temperature-aware processor frequency assignment for MP-SoCs using convex optimization," in *Proc. CODES+ISSS*, 2007, pp. 111–116.

[7] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, and K. Skadron, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. VLSI Syst.*, vol. 14, no. 5, pp. 501–513, 2006.

[8] W. Liao, L. He, and K. M. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*", vol. 24, no. 7, pp. 1042–1053, 2005.

[9] R. Rao, S. Vrudhula, and C. Chakrabarti, "Throughput of multi-core processors under thermal constraints," in *Proc. ISLPED*, 2007, pp. 201–206.

[10] *Intel Pentium 4 Processor 6x1 Sequence: Datasheet*, Intel Corp, 2006.

[11] R. F. Hartl, S. P. Sethi, and R. G. Vickson, "A survey of the maximum principles for optimal control problems with state constraints," *SIAM Rev.*, vol. 37, no. 2, pp. 181–218, 1995.