SCHEDULER COST FUNCTION DERIVATION

SOUMIK GHOSH

There are two major metrics of interest for the thermal workload of a multicore processor. The first is the total time of execution (denoted by L for length) of a certain application (denoted by A),

$$L(A(p_i), D_A, t),$$

where application A has p processes, each associated with a data set of size d_i in a single chip. D_A is the total data associated with application A, and $D_A = \sum_{i=1}^{p} d_i$. We assume that the activities are taking place in a staging area which contains the main and virtual memory operating spaces, as well as the processor with its cores and their associated caches and shared cache.

This time of execution measurement includes both computation time, and the time to move the data for the problem from the staging area (peripherals off the chip like DRAM and HDD) to a computation or operation area (on the chip such as the caches and the cores).

The second major metric of interest is the energy consumption or energy workload of an application, $U(A(p_i), d_i)$. For each application A and problem size D_A, a measure of the workload, $W(A(p_i), d_i)$, is defined in a data-operation dependent and system-independent way. W contains two components, one being the operations count that is performed by the computational core, and the other is the communication operations in transferring data and instructions and data coherency and book-keeping operations. These are measured in terms of the number of bytes operated ON, or number of bytes transferred. Thus the energy workload of an application A operating on a data set D_A can be expressed as :

$$U(A(p_i), d_i) = \lim_{n \to k_e} n \times W(A(p_i), d_i) \times L_n(A(p_i), d_i, t)$$

where $L_n(k, d_i)$ is the total time to execute n applications using the chip. The term k_e is the total number of applications that can be excuted with the associated length of time for L_n, at which point a “thermal event” will occur causing the applications and the system to catastrophically fail, or shut down.

It is easy to see that the above term is energy consumption of the system till a thermal event occurs. In order to relate the energy expenditure of the system while running applications, to teh corresponding joule heating, we define the term "Thermal equivalent of Application" (TEA), which is defined as the electrical work converted to heat in running an application and is measured in terms of die tmperature change and ambient temperature change of the system. Thus for the application A we express TEA as :

$$\begin{aligned}\Theta_A(A(p_i), d_i, T, t) &= \frac{U(A(p_i), d_i)}{\lim\limits_{T \to T_{th}} J_e \times (T - T_{nominal})} \\ &= \frac{\lim\limits_{n \to k_e} n \times W(A(p_i), d_i) \times L_n(A(p_i), d_i, t)}{\lim\limits_{T \to T_{th}} J_e \times (T - T_{nominal})}\end{aligned}$$

In these expressions, T_{th} refers to the threshold temperature at which a DTM triggered event will occur. $T_{nominal}$ refers to the nominal temperature as reported by the DTM counters / registers , when only the operating system in operating and no application is being run. The term J_e is the "electrical equivalent of heat" for the chip, which reflects the *informational entropy* of the system associated with processing the bits application $A(p_i)$ computes and communicates, as well as the black body thermal properties of the chip packaging and the cooling mechanisms around the chip. TEA thus is a dimensionless quantity , both denominator and numerator being expressing of work done or energy consumed in finishing a task.

For managing the thermal envelope of applications on server systems as well as embedded systems, we are interested in the thermal efficiency of the operation, that is, the thermal cost of taking an application to completion. In general the efficiency $\eta(A(p_i), d_i, T)$ is defined as

$$\eta(A(p_i), d_i, T, t) = \frac{\Theta_A(A(p_i), d_i, T, t)}{\Theta_A(A_e(p_i), d_i, T_{me}, t_e)}$$

where $T_(me)$ is the maximum temperature till which the core will carry over till a DTM triggered event occurs. A_e refers to the application whose energy consumption has caused the DTM triggered event to take place. T_e is the execution time of application A_e. Thus $\eta(A(p_i), d_i, T)$ is a measure of the "thermal efficiency of the application", which implies how much an application affects temperature change without compromising it's throughput and/or leads to a thermal event. Thus the definition of η is linked to the definition of the thermal and energy workload.

An important metric finally is the achieved performance per unit power consumed by the chip,

$$\begin{aligned} C_\theta(A(p_i), d_i, T, t) &= \frac{\Theta_A(A(p_i), d_i, T, t)}{P(A(p_i), d_i)} \\ &= \frac{\Theta_A(A(p_i), d_i, T, t)}{\sum^{chip,DRAM,HT,HDD} \int_{t=0}^{t=L_A} v(t)i(t)dt} \end{aligned}$$

where $P(A(p_i), d_i)$ is the overall power consumed during the application lifetime. the quantity $\int_{t=0}^{t=L_A} v(t)i(t)dt$ is the total power consumed by a single physical component (processor, DRAM units, HDD, HT or FSB) during the length of the application L_A, with $v(t)$ and $i(t)$ being the instantaneous voltage and currents respectively.

This normalized quantity C_θ gives some indication of the "cost" of executing the benchmark on the given chip.

The final optimization function could be thought of as :

$$\frac{\partial^2 C_\theta(A(p_i), d_i, T, t)}{\partial T \partial t} = \frac{\partial^2}{\partial T \partial t} \frac{\lim_{n \to k_e} n \times W(A(p_i), d_i) \times L_n(A(p_i), d_i, t)}{\lim_{T \to T_{th}} J_e \times (T - T_{nominal})} \times \frac{1}{\sum^{chip,DRAM,HT,HDD} \int_{t=0}^{t=L_A} v(t)i(t)dt}$$