Automatic Performance Setting for Dynamic Voltage Scaling

Krisztián Flautner
manowar@engin.umich.edu

Steve Reinhardt
stever@eecs.umich.edu

Trevor Mudge
tnm@eecs.umich.edu

University of Michigan
1301 Beal Ave.
Ann Arbor, MI 48109-2122
+1-734-764-0203

Abstract

The emphasis on processors that are both low power and high performance has resulted in the incorporation of dynamic voltage scaling into processor designs. This feature allows one to make fine granularity trade-offs between power use and performance, provided there is a mechanism in the OS to control that trade-off. In this paper, we describe a novel software approach to automatically controlling dynamic voltage scaling in order to optimize energy use. Our mechanism is implemented in the Linux kernel and requires no modification of user programs. Unlike previous automated approaches, our method works equally well with irregular and multiprogrammed workloads. Moreover, it has the ability to ensure that the quality of interactive performance is within user specified parameters. Our experiments show that as a result of our algorithm, processor energy savings of as much as 75% can be achieved with only a minimal impact on the user experience.

1. Introduction

The performance of microprocessors has been improving at an exponential rate and this trend is likely to continue for several years to come. However, increased performance does not come for free. One of the most important consequences of higher performance has been a dramatic increase in power consumption. While an Intel 386 processor consumed about 2 Watts of energy, a Pentium 4 can use as much as 55 Watts. In a mobile environment, batteries have not kept pace with the increased energy requirements, which means that either application performance or battery time suffers. However, even in environments where energy storage is not an issue, energy cost and heat management may become problems [9].

There is still a need to continue to improve processor performance, since not all applications are "fast enough,"
but an increasing number are. A way to bridge the gap between high performance and low power is to allow the processor to run at different performance levels depending on the application's requirements. Some processors, such as the Intel XScale [2] and Transmeta Crusoe [8] allow the frequency of the processor to be reduced with proportional reduction in voltage. Slowing down frequency without voltage scaling is not useful, since the power savings is offset by an equal increase in execution time, yielding no reduction in the total amount of energy consumed. However, since energy is proportional to the square of the voltage, reducing the operating voltage can yield significant energy savings [13].

The central issue with processors whose performance can be changed is how the right level of performance can be obtained. The goal is to reduce the performance of the processor without causing an application to miss its deadlines (see Figure 1). Completing a task before its deadline and then idling is less energy efficient than running the task more slowly to begin with, and meeting its deadline exactly.

Our aim is to design an algorithm that balances energy savings with the following requirements:

- No modification of user programs.
- Works with irregular and multiprogrammed workloads.
- Ensures that user-perceived performance does not suffer.

Previous interval-based approaches to automated performance setting did not fully achieve the goals outlined in the last two points. These approaches focus on the ratio of idle-to busy-time as the indicator of the right performance setting [15][5][6][12]. While the results looked promising for regular workloads (such as audio playback, where processor utilization is periodic), the proposed schemes do not work well for interactive or irregular applications.

The aforementioned papers point out that looking at idle time alone as the indicator of the right performance level is not sufficient. In their future work section, Weiser et al. propose an alternative approach, where jobs are classified into background, periodic and foreground classes [15]. They suggest that the added semantic information could be used to improve the scheduling algorithms. Govil et al., in their future work section, propose a similar solution, where process type along with information specified by the processes (e.g. deadline) could be used for performance setting

FIGURE 1. Performance scaling
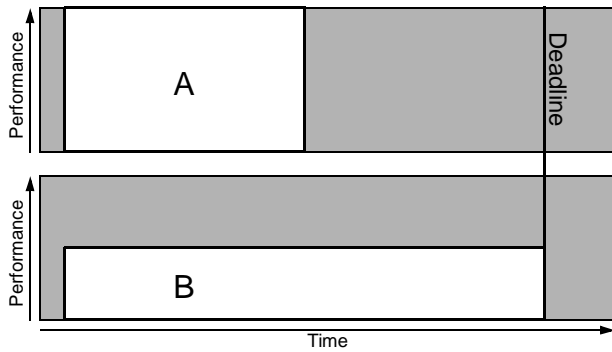
This figure shows two different runs of the same workload. In A, the workload runs at full speed and finishes well in advance of its deadline. In B, the execution of the workload is stretched to its deadline, which allows for energy savings on processors that implement voltage scaling.

[5]. Our approach follows along the lines of these earlier works; however, we derive deadline and classification information automatically from the OS kernel, by examining the communication patterns between the executing tasks (Section 3). This information is used to isolate execution episodes corresponding to different communication patterns. We can classify execution episodes into one of the following categories: interactive, periodic producer, and periodic consumer. These classifications can be used to derive deadlines for the execution episodes. For example, for an interactive episode, the deadline is the perception threshold, which we assume to be between 50ms and 100ms. The deadline for a producer is the point at which the consumer actually needs the produced data. Episode classification guides performance-setting decisions on a per-task and per-episode basis and is also used to dynamically evaluate the impact of past decisions on user-perceived performance (Section 4).

We focus primarily on interactive applications, since we believe that this is one of the most difficult, but also the most important class of applications for performance scaling. We also consider the effects of a concurrently running background application (an MP3 player) and a variety of assumptions about the power and performance models on the performance-setting strategy (Section 6).

2. Previous work

In the context of real-time systems, researchers have explored voltage and frequency scaling as a means of reducing power consumption. Papers [10][7][14] present algorithms and theoretical models that allow one to incorporate voltage scheduling into real-time schedulers. However, these papers are not directly applicable to general-purpose operating systems, since the workloads are expected to have well defined characteristics (periodicity, resource requirements, deadlines, etc.). Moreover, the user must explicitly specify these characteristics to the scheduler.

Our research is more closely related to the work described in [13][15][5], where performance-setting decisions are made automatically, guided by the ratio of idle time to busy time on the processor. A weakness with the existing approaches are that they are not very accurate and

can be easily confused by irregular processor utilization. We improve on these algorithms by using task-level information form the OS kernel. A more detailed summary of the directly related work follows below:

The main ideas behind automatic performance setting were sketched out by Weiser et al. [15]. Their mechanism uses the amount of idle time as the guide for finding the optimum level of performance. The practical policy proposed in this paper is called PAST. In this policy the utilization for the most recent interval is computed and, if it is above a certain threshold performance, is increased, but if the interval includes mostly idle time, performance is reduced.

While the PAST algorithm looks very attractive due to its simplicity and effectiveness on some benchmarks, Govil et al. [5] point out its shortcomings and propose improvements. One of their complaints is that PAST looks back only at a single interval and thus it smoothes speed poorly: the algorithm keeps on changing performance levels without coming to a steady state, missing out on opportunities to save power. To remedy the situation, the authors propose a number of algorithms that use varying amounts of prediction to improve their accuracy. They conclude that smoothing, rather than sophisticated prediction might be the most effective. Accordingly, they propose an algorithm, PEAK, that looks for recurring patterns of processor utilization, with special attention to short bursts of high utilization, and attempts to set processor speed accordingly.

Pering et al. evaluate interval-based voltage scaling algorithms for use on a handheld device [12]. One of the key contributions of this paper is the use of the clipped-delay metric, which takes into account that the length of some events can be increased without affecting the user (see Section 4.1.1). In practice, the effectiveness of this technique depends on the allowed increase in delay. The algorithm may degrade performance slightly but yield significant power reduction. While the algorithms used in this paper worked, their performance fell short of optimum. Moreover, the algorithms used some specific knowledge about the executing programs, which may be impractical on a production system. Pering et al. show that while interval-based voltage scaling algorithms work well on benchmarks with regular processor utilization (such as audio playback), they do not fare well on irregular workloads, such as interactive workloads or video playback.

These conclusions are corroborated by Grunwald et al. [6]. They also find that using a weighted average of processor utilization as a guide to future utilization (the AVG_N policy) does not yield the clock speed that would maximize processor utilization. Another problem with this algorithm is that the requirement to average N intervals introduces a delay in responding to processor demand. The authors find that existing heuristics did not fare as well on an actual implementation as previous studies had suggested.

3. Episode detection

We first applied our interactive-episode detection algorithm to measure the effects of multiprocessing on interactive performance [3]. In this section we summarize this methodology and also show how it can be extended to find periodic episodes.

FIGURE 2. Episode types
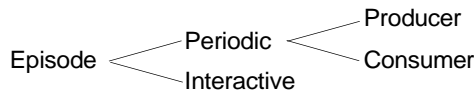
There are two principle groups of episodes: periodic and interactive. Periodic episodes can be further categorized into producer and consumer, where the communication between these episodes establishes their performance level.

FIGURE 3. Producer and consumer episodes

The figure shows communications between a producer and a consumer process. The processor can be slowed down to stretch the producer episode to the beginning of the consumer episode.

The first step in determining the optimum level of performance is to distinguish the important parts of the execution from unimportant periods. We use the communication characteristics of the executing applications as the basis of this classification. Episodes are triggered by communication events with specific tasks but multiple tasks may be involved during the execution of an episode. For example, an interactive episode involving ghostview is triggered by a message from the X server to ghostview. In return, events are processed by the application and it may send messages to the X server, the window manager and its rendering engine (ghostscript). All of these processes are part of the episode.

Two types of episodes are detected: interactive and periodic, with producer and consumer subcategories (see Figure 2). All other processor activity is classified as background activity. It is important to note that during its lifetime a task can fall into more than one of these classifications. For example, a music playback process may be part of an interactive episode when it is updating the GUI and be a producer when it is decoding music data.

We monitor which tasks communicate with a few well known system tasks (such as the X server and the sound daemon). These tasks are then monitored for communication through specific system calls that are then used to classify them into one of the above categories. In addition, we collect run-time statistics about processor utilization. Thus, instead of relying on the programmer, we extract the necessary information from the system automatically, using simple changes to the OS kernel.

As a consequence of our approach, the only idle time that shows up within an episode is due to device or communication latencies (hard idle-time) and cannot be removed by performance changes of the processor. Soft idle-time, on the other hand, occurs between episodes and is mostly due to latencies inherent in user interactions. This type of idle time can be reduced by slowing down the execution of episodes.

3.1   Interactive episodes

The beginning of an interactive episode is initiated by the user and is usually signified by a GUI event, such as pressing a mouse button or a key on the keyboard. Finding the end of an episode is more difficult since there is no event that automatically gets generated when the computer is done responding. To find interactive episodes, we keep track of the set of tasks that communicate with each other as a result of a user-initiated GUI event.
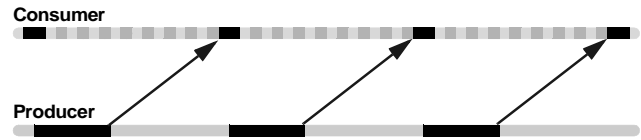
The start of an interactive episode is initiated by the GUI controller (X server in our case) sending a message through a socket to another task. When this happens both the GUI controller and the receiver of the task are added to what we refer to as the task set of the episode. If the members of the task set communicate with non-member tasks, then the as yet non-member tasks are also added. The end of the episode is reached when all the following conditions are met for tasks in the task set:

- None of the tasks are executing.
- Data written by the tasks have been consumed.
- None of the tasks remains unfinished, as a results of being preempted the last time it ran (i.e., all tasks gave up time on their own by blocking in a system call).
- None of the tasks are blocked on device I/O.

To compute this information, we need to monitor which system calls are accessed by each task, and how data flows between the communication structures in the kernel. Monitoring requires only minor modifications to the kernel and has very little overhead both in terms of memory footprint and execution time. The modifications consist mainly of additions of a few variables to kernel data structures (i.e. task_struct, socket, etc.) and then of updating them when communication takes place. For example, we have modified the kernel's socket data structure so that we can determine the last writer of the socket when it is read. For a more complete description of the implementation details, see [3].

Detecting interactive episodes is only the first step towards performance prediction. Section 4.1 describes how the episode's deadline can be found.

3.2   Periodic episodes

Detecting periodic activity is similar to detecting interactive episodes. However, instead of using communications with the X server as the trigger for starting the episode, we base this decision on whether the initiating task is periodic. To detect periodic activity, we keep track of two pieces information for each task:

- Last execution time.
- Length of the n last periods.

If a task exhibits only a small amount of variation in period length over the last n runs ($< 5\%$), then we treat it as a periodic task.

FIGURE 4. Cumulative interactive episode length distributions
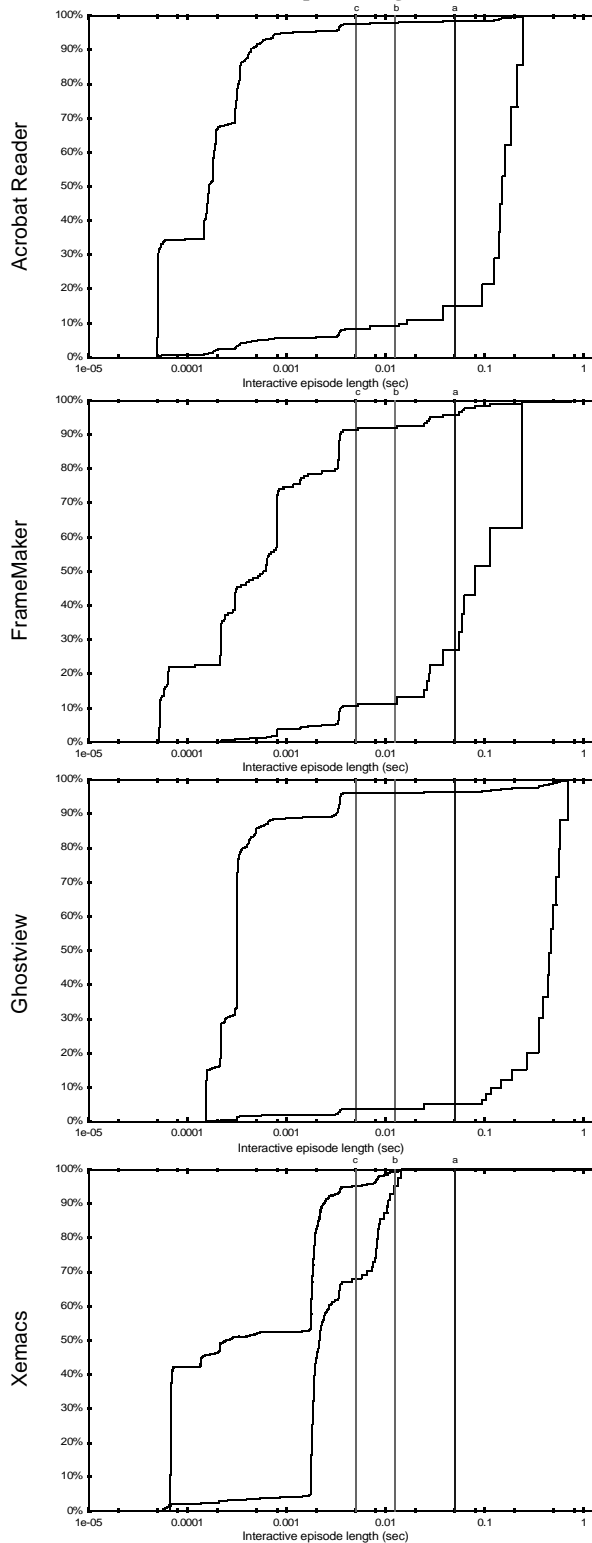
Left line shows the cumulative number, right line the cumulative percentage of time spent in interactive episodes whose lengths are less than or equal to the time specified on the x axis. The x axis is drawn using a logarithmic scale. Vertical lines from <u>right to left</u>: a) 50ms, b) 12.5ms and c) 5ms.

3.2.1 Producer-consumer episodes

Producer-consumer episodes form a special subcategory of periodic episodes, where the optimum performance level is established by the distance from producer to consumer, not by the distance between periods. A case in point is the Linux esd sound daemon, which wakes up periodically to check for sound playback requests and to send data to the sound card. If esd's playback buffer is not empty, it sends some of the data to the sound card. If the buffer is close to being empty, it wakes up and unblocks the music decoders (e.g. MP3 players), which causes them to generate the next few frames of data.

Figure 3 illustrates interaction between a producer and a consumer process. The distance between the run of the producer and where the data is needed establishes the performance level of the producer. The producer episodes can be stretched to the beginning of their associated consumer episodes. To determine how much the consumer can be slowed down, we first need to determine what the consumer is doing. This information can either be specified by the user on a per-process basis, or one can compute it by observing which devices or processes the consumer communicates with. If the task is communicating with a device, the buffer sizes and the speed requirements of the device establish the minimum speed of the consumer episode. In our current scheme we assume that the consumer can be slowed down to the same speed as the producer.

4. Performance prediction

Our prediction mechanism operates on a per-task basis and uses different algorithms for interactive and periodic episodes. In both cases, the predictor computes the performance factor, which is the ratio of the desired execution speed and the processor's maximum speed.

4.1 Interactive episodes

It is difficult to come up with a good prediction strategy for the optimum performance level of interactive episodes, since interactive episodes are completely dependent on the user, not on some activity within the computer. There is no predictable pattern of recurrence and the lengths of interactive episodes can have orders of magnitude of difference. Our detection scheme allows us to differentiate between different types of episodes (i.e. interactive, producer, consumer) but cannot distinguish between different instances of the same episode in the same task (e.g. when the same button is pushed in the GUI as before).

We believe that the ability to distinguish between interactive episode instances would improve prediction accuracy. However, this would require the kernel to have knowledge about the location in the user program that initiated the given interactive episode. While not impossible, distinguishing the real call-sites from the kernel is difficult to do. A simple comparison based on the user-mode program counter (PC) is not sufficient, since programs usually go through at least one level of indirection (through libC) when calling a system call, and thus all instances of a program's calls to a given system call would have the same user-mode PC. Moreover, since interactive episodes are usually generated as a result of GUI interaction, the necessary number of indirection levels is probably higher due to the use of GUI libraries (e.g. gtk, Xlib). To find the PC

---

value that really distinguishes one interactive episode from another, one would have to chase pointers through multiple levels, where the actual number of levels depends on the environment (stack layout, libraries, etc.).

Instead of basing a predictor on the ability to distinguish between interactive episode instances, we looked for a simpler solution that only relies on episode type (i.e. interactive, periodic producer or periodic consumer) for prediction.

Figure 4 shows the cumulative distribution of interactive episode lengths for four interactive benchmarks. In each graph, there are two cumulative distributions: the one on the left shows the cumulative number and the one on the right shows the cumulative time spent in interactive episodes of a given length or shorter. To account for the large variation of interactive episode lengths, the time axis is logarithmic. Three vertical lines (a, b, and c from <u>right to left</u>) delineate the perception threshold (50ms), the point under which all episodes finish under the perception threshold at 1/4th of peak performance (12.5ms), and 1/10th peak performance (5ms). These values were selected because current processors that are capable of performance and voltage scaling have a minimum performance of about 1/4th peak performance, and future processors could possibly extend the rage of performances to 1/10th of peak value.

These graphs show that while most episodes are very short, the vast majority of time is actually spent in a small fraction that correspond to the long episodes. For example, in Ghostview, 92% of the time is spent in 4% of the episodes. This distribution holds true on the Xemacs benchmark as well, however in this case even the relatively long episodes fall under the perception threshold. Xemacs is an example of an application where one could run almost all of its interactive episodes in the lowest performance level without ever exceeding the perception threshold.

The cumulative episode length distribution graphs imply that a predictor that predicts that an interactive episode only needs the minimum available processor performance would be right more than 90% of the time. However, since these episodes tend to make up only a small percentage of total time—and consequently have a small contribution to energy use—it is more important to focus on accurately predicting the performance level of the relatively long episodes.

Our performance-factor predictor for interactive episodes works by starting off with an initial performance factor, set to the minimum performance factor of the processor, and then by successively refining its value. Since the initial setting is only relevant for the first interactive episode, the choice of initial value does not have a significant impact on response time or energy savings. The algorithm uses the following three steps:

- Starts running the episode at the predicted performance factor.
- At the end of the episode, computes the duration that corresponds to executing at full performance. Use this information to compute the optimal performance factor for the episode.
- Uses the weighted average of optimum performance factors (*PF*) as a prediction for future performance factors.

The main observation that we use in our predictor is that it is straightforward to compute what the optimum performance level should have been once an interactive episode is over. During the execution of the episode, the performance-setting of the processor might be changed by external events (e.g. periodic episodes start executing), so the algorithm must keep track of the observed performance factors (pf_i) during the episode's execution. At the end of the episode, this information can be used to estimate how long the episode would have been at full performance (see Equation 1).

$$T_{FullSpeed} = \sum_{i=1}^{n} pf_i(t_i - idle_i) + idle_i \qquad \text{(EQ 1)}$$

This equation computes the full speed execution time for an interactive episode given n different observed performance levels during the episode. The variable t_i specifies the length of execution at the i-th performance level during an interactive episode, and $idle_i$ is the corresponding amount of idle time.

Based on the estimate of the episode execution time at full performance, the optimum performance level can be estimated for an interactive episode. Equation 2 gives the equation for computing the optimum performance factor for episodes where $T_{FullSpeed}$ falls between the *minimum-performance threshold* and the *PerceptionThreshold*, where T_{idle} specifies the amount of idle time during the episode.

$$pf_{optimum} = \frac{T_{FullSpeed} - T_{idle}}{PerceptionThreshold - T_{idle}} \qquad \text{(EQ 2)}$$

The minimum-performance threshold specifies the episode duration that could be slowed down to the processor's minimum performance level and still finish under the perception threshold. If the perception threshold is assumed to be 50ms and the processor's minimum performance is 1/4th of peak, then this value is 12.5ms. Episodes that are shorter than the minimum-performance threshold can be run at the processor's minimum performance level. Episodes that are longer than the perception threshold need to run at full performance.

We predict the performance factor for the next interactive episode of a given task simply as the average of the optimum performance factors of past interactive episodes, weighted by the duration of each episode.

$$PF_{prediction} = \frac{\sum_{j=1}^{k} pf_j T_j}{\sum_{j=1}^{k} T_j} \qquad \text{(EQ 3)}$$

Equation 3 shows the computation for the predicted performance factor based on the optimum performance factors (pf_j) for k past interactive episodes. T_j refers to the estimated full-speed time of an interactive episode. The size of k can be varied to eliminate saturation and to allow temporal variations of episodes lengths to affect the predictor (see Section 4.3).

FIGURE 5. Performance factor settings during the execution of the Acrobat Reader benchmark
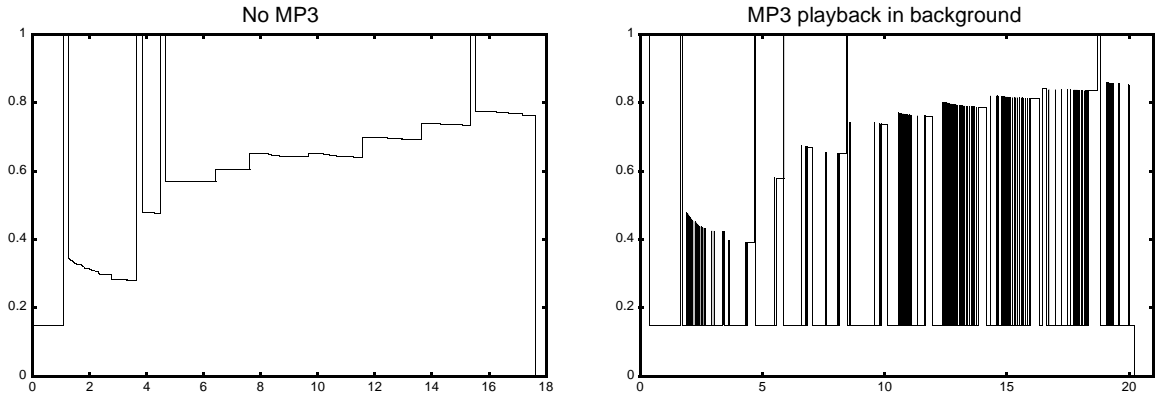
Two runs of the Acrobat Reader benchmarks are shown side by side with and without MP3 playback during the run. Perception threshold was set to 200ms, and data was generated using our simplest strategy (Basic predictor and XSB model without quantization, see Section 6.2). More sophisticated models have significantly fewer performance-level transitions when MP3 is executing in the background. Spikes to full performance represent instances when the PanicThreshold was reached. Note that the graphs do not show transitions to sleep mode.

Since there can be orders of magnitude of difference between the lengths of interactive episodes (see Figure 4), this strategy means that the predicted performance factor for short episodes will almost certainly be higher than necessary. This effect is mitigated by the observation that short episodes have only a minimal impact on power consumption.

To recover from prediction errors we set an episode-duration threshold, after which if the episode is still executing, the performance level is raised to full speed. We refer to this threshold as the PanicThreshold. While the PanicThreshold can ensure that interactive performance does not degrade below a certain level, the goal of the predictor is to set the right performance level at the beginning of the episode, without the need to transition to a higher performance setting later on.

The setting for the PanicThreshold reflects the user's tolerance for worst-case performance degradation and determines how speculative the performance factor predictor can be. If the user has no tolerance for possible performance degradation, there is no opportunity for speculation and consequently energy savings. In this case one would be forced to be conservative and always run at full performance to avoid misprediction errors that might extend the episode beyond the perception threshold.

Equation 4 shows the computation for the PanicThreshold for a given performance factor (*PF*) and perception threshold.

$$PanicThreshold = PerceptionThreshold(1 + PF) \quad \text{(EQ 4)}$$

This formula allows a longer panic threshold when the initial performance setting is high, because in those cases more work actually gets done per unit time and therefore the cost of an incorrect setting (in terms of its impact on the user) is lower. We also make the assumption that the user allows more performance degradation for episodes whose lengths are close to the perception threshold than for longer episodes. In the worst case, if an episode were to be exactly 50ms at full speed, then its length will be stretched to 97ms given a performance factor of 1/4. On the other hand given

the same performance factor, an episode that would have been 200ms at full speed would only be stretched to 247ms.

4.1.1 The perception threshold

In this paper we use a range of perception thresholds during some of our experiments. Our motivations for this are twofold:

- The higher perception thresholds allow us to estimate the energy and interactive characteristics on a future, higher-performance processor. The 100ms threshold on today's processor roughly corresponds to the 50ms threshold on a processor with twice the performance.
- The perception threshold varies by individual and task and may be used as a user settable indicator for his preference for high-performance or energy savings.

Literature about human-computer interaction [11][1] indicates that 20-30 frames per second are sufficient for the human visual system to perceive the images as a continuous stream. This suggests that the perception threshold is around 50ms. Human subject tests in [1] show that perceptual causality—when two events are perceived to be fused together—ends around 100ms, and for some test subjects quality degradation begins at around 50ms.

Other experiments have shown that for simple operations, such as dragging an object through the screen, as few as 5 updates per second are sufficient to maintain an interactive feel (200ms perception threshold). For non-continuous operations, as much as 1-2 second delays are acceptable [11]. However, when human motor operations form a feedback loop with visual activity, then it is more important to have a faster response time.

4.2 Incorporating periodic episodes

The optimum performance factor for periodic activity can be computed easily by either stretching the periodic episode's execution to the beginning of the next episode or to the beginning of the associated consumer episode. Since

periodic (such as video or sound playback) applications sometimes adjust the quality of playback based on processor performance, it is important to switch to full performance when a periodic application starts executing, so that it has a chance to adapt to the highest performance level. Our assumption is that the user's emphasis is on service quality over energy savings. Others have addressed the trade-off where service quality can be reduced to save energy [4].

An important consideration is to find the performance factor when interactive episodes are present in addition to the periodic activity. Our strategy is very simple:

- When there is no interactive episode executing on the processor, we set the performance factor to the one computed for the periodic activity.
- At the beginning of an interactive episode we switch to the performance factor that was predicted for the task's interactive episodes, if it is higher than the periodic performance factor.

Figure 5 illustrates this strategy during two runs of the Acrobat Reader benchmark. When there is no periodic activity, performance is determined only by the prediction for interactive episodes. However, when periodic activity is present, the algorithm switches between the two performance levels, causing significantly more performance transitions. The spikes that transition the processor to full performance are triggered by interactive episodes whose lengths exceed the PanicThreshold. Aside from the initial start at full performance when MP3 is executing in the background, there is only one extra transition due to reaching the PanicThreshold on the second figure.

Periodic episodes have the effect of extending the run-times of interactive episodes [3], which means that the interactive performance-factor predictor should be updated when periodic activity starts. Instead of making the prediction formula more complicated, our approach allows the performance predictor to quickly adapt to the presence of the periodic activity. The next section describes how this is done exactly.

4.3 Implementation details of the Basic predictor

The main features of the Basic predictor are summarized below:

- Interactive episode performance level prediction based on optimum performance factors of past episodes. Estimates of the optimal performance factors are computed at the end of each interactive episode.
- PanicThreshold bounds worst-case performance.
- Periodic performance level computed by observing periodic episodes and their communication patterns.
- Switches between periodic and interactive performance factors depending on which episode is executing.

An attractive feature of the predictor is that it requires very little state. We use two variables per task: one keeps track of the sum of episode length weighted performance factors (*totalPF*) and the other keeps track of the total time spent in episodes (*totalTime*). In both cases time is the estimated full-speed execution time of the episode. The performance factor prediction is thus *totalPF* divided by *totalTime*.

One problem with an averaging based predictor is that if the execution time is long, then temporal variation may not influence the predictor for a very long time. One way of alleviating this problem is periodically rescaling the variables by dividing them both by the same amount. This way the predictor can better accommodate a changing workload.

Performance prediction for interactive episodes in the presence of periodic activity relies on rescaling to allow the predictor to adapt to the changing workload. Our studies in [3] have shown that even lightweight background activity, such as MP3 playback, extends the duration of perceptible interactive episodes by an average of 14%. This implies that performance factors predicted based on data without the background activity would underestimate the necessary performance. To alleviate this problem, when periodic activity is detected, the *totalTime* variable is set to 100ms and *totalPF* is recomputed based on the new value. While providing a reasonable initial prediction, this change allows new performance factor data to take hold quickly.

5. Simulation methodology

Our simulator is driven by traces collected using a modified Linux kernel (2.3.99-pre3) running on a Dell Precision workstation 410, with only one of the two Pentium II 450Mhz processors enabled (512M RAM). The software environment was Mandrake Linux 7 with Helix Gnome 1.2. The traces used in this study are the same as the uniprocessor traces used in [3]. All benchmarks were run by a live user. While we have collected multiple runs in each configuration, in this paper we only use a single run for each simulation. We aimed to repeat each run with MP3 in the background as accurately as possible, but there are slight variations between the runs. All the significant events (e.g., mouse clicks, text entry) were performed in the same order during each benchmark run. However, the exact path of mouse movement (and therefore the interactive episodes corresponding to them) and the amount of time between events varies from one run to the other.

The traces include all significant OS events during the benchmarks execution: thread swap events, system calls, and task information (e.g. name, pid, etc.). Based on this information, our simulator can reconstruct the communication events between the tasks (which imply the synchronization points between them) and simulate the effects of performance scaling. The upside of our methodology is that we have the flexibility to investigate a wide set of architectural parameters. The downside is that the absence of actual hardware prevents us from measuring total energy consumption and from calibrating our results.

6. Energy and performance implications

Our aim is to develop a performance scaling technique that can guarantee that user-perceived performance does not degrade below a user-settable level. A detailed microarchitecture-level power analysis is beyond the scope of this paper; however, we can derive some estimates regarding the expected energy savings using a few simple assumptions.

The metric we use is the energy factor, which is the ratio of the energy used by the scaled workload divided by its predicted energy use at full performance. Equation 5 gives the energy factor formula for a given workload,

TABLE 1. Performance characteristics (XSB, 50ms perception threshold)

Benchmarks	No MP3 in background				MP3 playback in background			
	Performance transitions	Mean perceptible IE length increase	Median perceptible IE length increase	Energy factor	Performance transitions	Mean perceptible IE length increase	Median perceptible IE length increase	Energy factor
Acrobat Reader	543	13%	7%	0.91	668	13%	11%	0.84
FrameMaker	155	20%	11%	0.89	191	9%	7%	0.75
Ghostview	510	5%	1%	0.98	1149	5%	1%	0.91
GIMP	919	5%	4%	0.97	1731	5%	4%	0.91
Netscape	1026	18%	14%	0.87	1739	21%	12%	0.82
Xemacs	381	23%	20%	0.30	1417	29%	33%	0.34

assuming that the workload is divided into n pieces that execute at the scaled voltage (v_i) and frequency (f_i, specified in Mhz) for the scaled amount of time (t_i, in seconds). T refers to the total execution time at full speed, while the *max* subscript refers to the maximum value of the given variable.

$$EnergyFactor = \frac{\sum_{i=1}^{n} v_i^2 f_i t_i}{v_{max}^2 f_{max} T} \qquad \text{(EQ 5)}$$

Our model focuses on the CPU alone and does not take the power consumption of other devices (such as memory and peripherals) into account.

6.1 Processor and power scaling model

Our performance model is based on assumptions from the Intel XScale microarchitecture [2]. Our traces were collected on a 450Mhz Pentium II based workstation, and we make the simplifying assumption that these traces correspond to the full-speed performance on each of the simulated models.

We assume that for each performance transition, there is a 20 microsecond pause, during which the processor does not execute any instructions. This pause is due to the time it take to resynchronize the PLLs for the changed frequency value. After this, the performance transition time—during which the voltage level is changed—is assumed to be 1 millisecond regardless of starting and ending performance level. During this time, we assume that the processor is executing instructions at the rate corresponding to the lower of

FIGURE 6. Frequency-voltage pairs in our energy models

Model	Frequency (Mhz)						
	150	333	400	600	773	800	1000
XSBase		1	1.1	1.3	1.5		
XSA	0.75	1	1.1	1.3	1.5		
XSB	0.75			1.2		1.4	1.75

The table shows the given frequency-voltage pairs of our models. Dark areas represent frequency levels that are not supported in a given model.

the two performance levels, but energy is being consumed at the higher.

Figure 6 shows the known frequency-voltage values that we used to compute voltage equations (Equations 5, 6, and 7) for arbitrary frequency levels between the minimum and maximum frequencies. The XSBase model corresponds closely to the high-end XScale part (80200M733) described in [2].

$$v_{XSBase} = -5 \times 10^{-8} f^2 + 0.0012f + 0.6261 \qquad \text{(EQ 6)}$$

However, since this model only has a 2.32x frequency range, we extended it to 5.15x by allowing it to go as low as 150Mhz in the XSA model.

$$v_{XSA} = -4 \times 10^{-7} f^2 + 0.0015f + 0.5324 \qquad \text{(EQ 7)}$$

The XSB models the parameters of a high-end device that is a research prototype. This processor can vary its performance between 150Mhz and 1000Mhz for a 6.67x swing.

$$v_{XSB} = 5 \times 10^{-7} f^2 + 0.0005f + 0.6624 \qquad \text{(EQ 8)}$$

We use the energy factor as our metric for computing energy reduction. It is the ratio of a given workload's energy consumption using our performance-setting strategy over its energy consumption using the processor's peak performance. In all our energy calculations, we assume that the OS power manager puts the processor into a low power sleep mode immediately when no instructions are executing. We do not attribute a power cost to this operation and assume that it happens instantaneously.

During our evaluation we specify a quantization factor for each of the power models. On an actual processor not all frequency/voltage pairs can be directly set up, one must choose from a set of predetermined values. This means that when the performance estimator requests a given performance setting, the actual performance value is rounded up to the next quantum. In our experiments we mostly focus on models where frequency is quantized at 5% steps (100%, 95%, 90%, etc.). We denote quantized models with the suffix 'q' followed by the quantum size.

TABLE 2.   Performance characteristics (XSBq5, 50ms perception threshold)

Benchmarks	No MP3 in background				MP3 playback in background			
	Performance transitions	Mean perceptible IE length increase	Median perceptible IE length increase	Energy factor	Performance transitions	Mean perceptible IE length increase	Median perceptible IE length increase	Energy factor
Acrobat Reader	28	12%	5%	0.92	664	12%	10%	0.86
FrameMaker	15	17%	11%	0.90	184	7%	6%	0.77
Ghostview	10	4%	0%	0.99	1135	4%	0%	0.92
GIMP	28	4%	3%	0.98	1533	5%	3%	0.92
Netscape	32	17%	12%	0.88	1547	20%	11%	0.84
Xemacs	15	15%	14%	0.26	1416	29%	31%	0.32

6.2   Performance and energy characteristics

In this section we examine the characteristics of the basic performance setting algorithm and propose some improvements. Our evaluation focuses on the following three main goals:

- Minimizing the number of performance-level transitions.
- Minimizing the amount of increase in the duration of perceptible interactive episodes.
- Maximizing energy reduction.

These three aspects are closely interrelated. Reducing the number of performance-level transitions is important, because each transition has a delay and energy cost that negatively affects both the perceptible performance and energy savings. On the other hand, increasing the interactive episode duration has a positive effect on energy savings, because the longer interactive episodes may stretch, the slower they can run. However, this has a negative impact on the user-perceived performance. While the increase in perceptible interactive episode duration in all cases falls within the acceptable range (since ensuring this is part of the performance-setting algorithm's job, see Section 4.3), we seek to minimize it, i.e. our methodology favors faster response time over energy savings.

The perceptible interactive episode-length increase is computed for all scaled episodes that fall above the perception threshold by dividing the scaled episode length by either the full-speed episode length or the perception threshold, depending on whether the original episode length was longer or shorter (respectively) than the perception threshold.

Table 1 shows our baseline results using the XSB model (without quantization) and 50ms perception threshold and assumptions described in Section 6.1. The mean perceptible interactive episode length increase in all cases is under 30%. Applications that have many short episodes (e.g. Xemacs and Netscape) tend to have the largest increase, while workloads with long episodes (e.g. Ghostview and GIMP) exhibit the smallest increase. This makes sense given that our acceptable delay function (*PanicThreshold*) allows more performance degradation for shorter episodes than for longer ones. One should also note that the number of performance transitions increases significantly (up to four times) when MP3 playback is running in the background. This is because, when a periodic episode is running, the performance setting algorithm alternates between the setting for interactive episodes and the setting for the periodic task. The energy factor tends to be lower when MP3 is running than without. The reason for this is

TABLE 3.   Performance characteristics with MP3 playback in background (XSBq5, 50ms perception threshold)

Benchmarks	MP3 IEPerf Transition start latency = 1ms				MP3 IEPerf Transition start latency = 5ms			
	Performance transitions	Mean perceptible IE length increase	Median perceptible IE length increase	Energy factor	Performance transitions	Mean perceptible IE length increase	Median perceptible IE length increase	Energy factor
Acrobat Reader	637	11%	4%	0.86	125	13%	6%	0.75
FrameMaker	153	8%	7%	0.76	81	12%	9%	0.73
Ghostview	1031	5%	1%	0.91	222	6%	1%	0.86
GIMP	854	5%	4%	0.88	334	6%	6%	0.83
Netscape	1072	18%	12%	0.83	340	20%	14%	0.72
Xemacs	1047	29%	32%	0.32	980	34%	39%	0.31

FIGURE 7. Energy factors corresponding to different perception thresholds using three quantized (5%) models
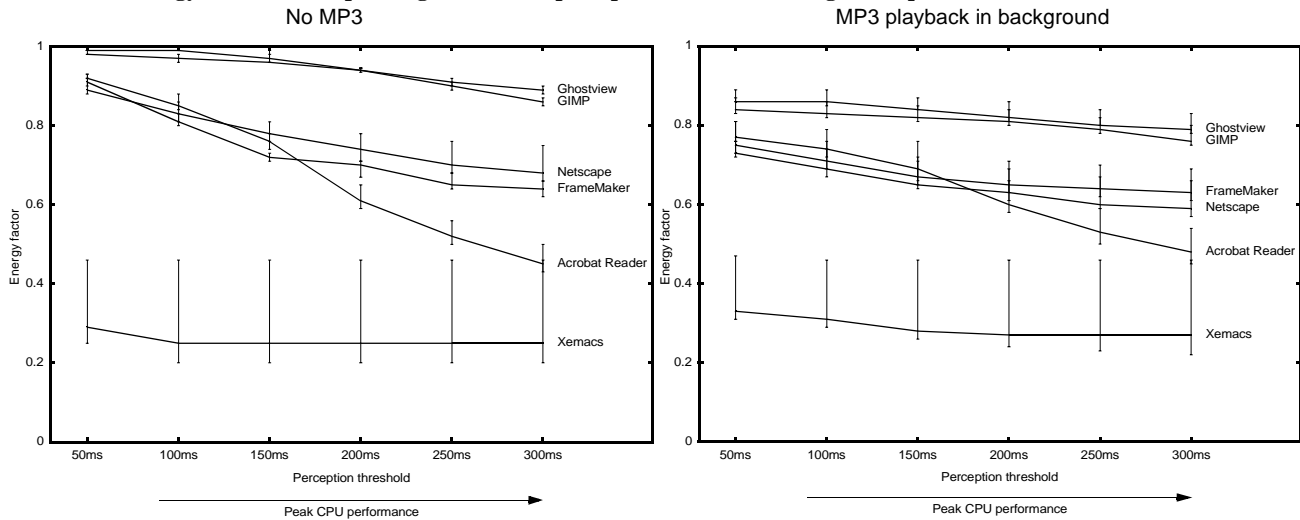
These graphs show results using the Enhanced predictor corresponding to a variety of perception thresholds. At each perception threshold level, we show the energy factors for the QSBaseQ5 (top point), XSAq5 (middle point, connected) and XSBq5 (bottom point) models.

that in most cases the MP3 player requires a lower performance setting than the interactive application. Xemacs is an exception: the benchmark's interactive episodes require a lower performance-setting than the MP3 player.

Table 2 illustrates the effects of quantization on the results. This model corresponds more closely to an actual hardware implementation than the one used in the previous table. Quantization tends to slightly reduce energy savings and also reduce perceptible interactive episode lengths. The only benchmark where this was not the case is Xemacs, where quantization corrects a few mispredictions, causing both an increase in energy savings and a decrease in the average perceptible episode length.

Perhaps the most striking improvement over the data in Table 1 is the dramatic reduction in the number of performance-level transitions (in some cases more than 300-fold) when no MP3 playback is running concurrently with the interactive application. This behaviour is also due to the fact that when there is no periodic background activity, the successively predicted performance levels are close to each other, causing quantization to eliminate the minor corrections. When MP3 playback is present, the deliberate transitions between interactive and periodic modes keeps the number of transitions high.

The number of performance transitions can be further reduced based on an observation of Figure 4. We have pointed out that the majority of interactive episodes are very short (less than 1ms) and that very little time is spent in those episodes (<5%). When there is no periodic background activity, the effect of the short interactive episodes is negligible since the performance level is always set at the predicted interactive performance level, however when there is an MP3 player in the background, these short interactive episodes cause an unnecessary transition from the periodic to the interactive performance level. When the interactive episode is very short, this transition simply wastes energy, since the episode is likely to be finished before the transition is over.

This observation suggests a strategy that waits for a certain amount of time before starting a transition to the interactive performance level. Table 3 illustrates the effects of a transition-start latency before interactive episodes. The data is only shown for the benchmarks with MP3 playback in the background, because there was no significant change in results when background activity was not present.

Contrasting with Table 2 shows that a 1ms transition-start latency leaves the energy factors mostly unchanged but causes a small reduction in the number of performance transitions. Extending the transition-start latency to 5ms causes both a significant reduction in energy consumption and in the number of performance transitions (up to a 5-fold reduction). Moreover, the average perceptible interactive episode-lengths stay at around the same level as before.
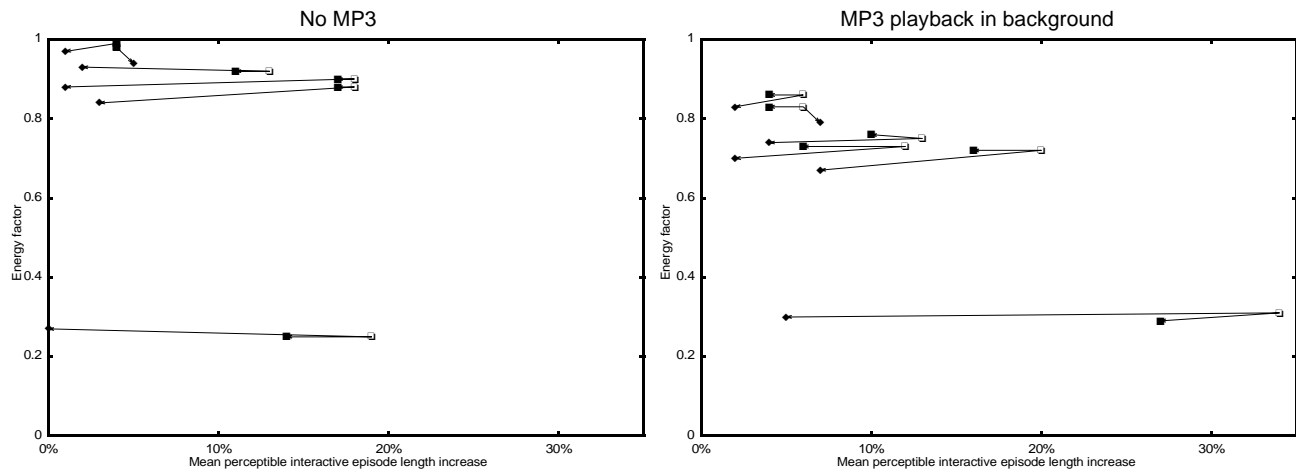
6.3 The Enhanced predictor

The previous section suggests two minor changes to the Basic predictor: 1) to quantize the allowable performance levels; and 2) to wait for a certain time before changing the performance factor when an interactive episode starts. For simplicity we only use a statically specified transition start latency of 5ms in the Enhanced predictor. A more sophisticated predictor could dynamically compute a per-process value.

Figure 7 shows the energy factors using the Enhanced predictor and the XSBaseQ5, XSAq5 and XSBq5 power models given a variety of perception thresholds. Our results show that while on the measurement machine there is little opportunity for power savings, as the peak performance of the processor gets faster, energy savings will be more pronounced. We must note that our traces were collected on a 450Mhz Pentium II machine, and today's high-end processors are already 2-3 times faster. We estimate that the energy factor on today's high-end desktops could be in the 10%-75% range at the 50ms-100ms perception threshold.

In the figure, energy factors corresponding to the XSAq5 model are connected by a line at each perception

FIGURE 8.   Energy factor and response time improvements with Oracle predictor and faster transition times

The graphs show how the energy factors and response times change when an Oracle predictor is used (◆) or the latencies due to performance and voltage changing are eliminated (■) compared to the realistic Enhanced predictor with the XSBq5 model (❑) with 50ms perception threshold. The datapoints for each benchmark are connected. Benchmarks from top to bottom: Ghostview, GIMP, Acrobat Reader, FrameMaker, Netscape, Xemacs.

threshold, the results for the XSBaseQ5 and XSBq5 are shown as the bars above and below each point (respectively). In all cases, XSBq5 achieves the largest energy savings, while XSBaseQ5 achieves the least. For most applications the difference between the three models is small: in most cases less than 5%. However, the difference is significant on the Xemacs benchmark runs, since this application spends most of its time at the lowest performance setting allowed in each model. In this case, the lower minimum performance levels of the XSA and XSB models give them a significant edge.

Periodic activity has the effect of vertically compressing the graph towards its center. The load due to the background activity increases the energy savings for benchmarks with long interactive episodes. On the other hand, one can observe the exact opposite effect when the interactive episodes are short (Xemacs).

6.4   Desired hardware improvements

We have already shown that by reducing the number of performance-level changes, quantization can have a positive effect on both energy savings and interactive performance. In our measurements we found that the quantum size does not greatly impact our results. Using the Enhanced predictor, the difference in energy savings due to quantum sizes of 5%, 10%, and 20% are negligible. While, in all cases the 5% quantum size has the smallest energy factor, the 20% quantum size is only behind by at most 1%. On the Oracle predictor, the difference in energy savings due to different quantum sizes is always under 5%.

Figure 8 contrasts the energy factors and effects on interactive episode lengths of three different models. The three connected datapoints for each benchmark correspond to the Enhanced predictor, an aggressive hardware model combined with the Enhanced predictor (see Section 6.4) and the Oracle predictor (where the predictor has perfect information, but the hardware model remains realistic). Points that are closer to the lower left corner of the graph are better.

Since the maximum increase of interactive episode lengths is bound by the PanicThreshold (see Equation 4), the increase shown in the graphs is not detrimental to the user experience. However, our goal is to minimize the interactive episode length increase for a given energy factor.

The Oracle predictor differs from the Enhanced predictor in that it has exact knowledge of the lengths of interactive episodes as soon as they begin. For this reason, there is no need to introduce a 5ms latency before transitioning to a new performance level (see Section 6.2 and Table 3 for why this latency is necessary in the Enhanced predictor). The detection of periodic episodes works the same way as in the enhanced predictor. The model used for the data is XSBq5 with 50ms perception threshold.

The figure shows that while there is a small reduction of the energy factors, the primary difference between the Enhanced and the Oracle predictors is that the latter has a significantly smaller average increase of interactive episode lengths. In fact the increase is due only to the latencies inherent in the power model, not to prediction error. If the Oracle predictor was allowed to extend the interactive episodes by the same amount as the Enhanced predictor, then one would see an additional 10%-20% percent increase in energy savings on benchmarks that have not already reached the minimum performance-level of the processor.

For all our measurements we assume that there is a 20 microsecond pause when a performance transition is initiated, and that a transition takes 1 millisecond. The current pause duration, which we believe is indicative of what can realistically be expected, is short enough that eliminating it has neither a significant impact on energy savings, nor on perceptible interactive performance. However, there might be other reasons for lowering it, such as latencies incurred during communication with peripheral devices.

Reducing the lengths of performance transitions, on the other hand, has a positive effect on both perceptible performance and energy savings. While shortening performance transitions has an overall positive effect, a better prediction

mechanism (as shown by the Oracle predictor) could achieve even more significant improvements. We believe that the Enhanced predictor could be improved by giving it the ability to distinguish between episode instances (as discussed in Section 4.1), not just episode types.

7. Conclusions and future work

In this paper we describe an automatic episode detection mechanism that can be used to guide the performance-setting decisions for a processor that supports dynamic performance and voltage scaling. This system can derive and predict episode deadlines automatically, without the need to modify existing user programs. We have shown that our approach can achieve significant energy savings while ensuring that interactive performance stays at an acceptable level. We are currently working on the evaluation of our algorithms on a system that is capable of dynamic voltage scaling.

While our current implementation is tied closely to the Linux kernel and its application environment, we believe that the ideas proposed in this paper are also applicable to other operating systems. We developed our methodology for Linux by observing common program design and communication patterns. While the specifics may vary from one OS to another, most modern operating systems have abstractions that a similar monitoring environment could be built on (i.e. inter-process communication, multithreading, system calls).

Our mechanism works without modifications of user programs, however an optional API might be useful for applications that want to take full advantage of performance scaling. One of the biggest shortcomings of the current predictor is its inability to distinguish episode instances from one another. An API that would allow the programmer to delineate and name critical episodes and perhaps optionally specify its type and deadline would help. The API might consist of the following system calls:

```
episode_begin <id> [type] [deadline]
episode_end <id>
```

The id is a per-task identifier assigned by the programmer to distinguish one episode from another. The type field optionally categorizes the episode into interactive, periodic, producer or consumer categories. The deadline field optionally specifies the maximum length of the episode. The idea behind this API is that its main role is to give hints to the existing prediction and communication-tracking mechanism, instead of superseding it. For example, there is no need to specify dependencies between episodes since that information can be derived automatically from information in the kernel.

We have shown that along with peak performance, it is also important to allow the processor to run slowly. While there will always be applications that can only run acceptably at the processor's fastest setting, an increasing number of applications are able to take advantage of the lower performance modes of the processor. As peak performance of the processor increases, it is important to widen the gap between the minimum and maximum performance levels of the processor.

We believe that the core idea of our technique—on-line monitoring and dynamic adaptation—could be extended to allow the kernel to make better scheduling and service-quality decisions.

8. Acknowledgments

References

[1] S. K. Card, T. P. Moran, and A. Newell. *The Psychology of Human-Computer Interaction*. Lawrence Erlbaum Associates, Publishers, 1983.

[2] Developer manual: "Intel 80200 Processor Based on Intel XScale Microarchitecture" http://developer.intel.com/design/iio/manuals/273411.htm

[3] K. Flautner, R. Uhlig, S. Reinhardt and T. Mudge. Thread-level parallelism and interactive performance of desktop applications. *Proceedings of Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-IX), November 2000.*

[4] J. Flinn and M. Satyanarayanan. Energy-aware adaptation for mobile applications. *Proceedings of 17th ACM Symposium on Operating Systems Principles (SOSP-17), December 1999.*

[5] K. Govil, E. Chan, and H. Wasserman. Comparing Algorithms for Dynamic Speed-Setting of a Low-Power CPU. *Proceedings of the First International Conference on Mobile Computing and Networking*, November 1995.

[6] D. Grunwald, P. Levis, K. Farkas, C. B. Morrey III, and M. Neufeld. Policies for Dynamic Clock Scheduling. *Proceedings of the Fourth Symposium on Operating Systems Design & Implementation*, October 2000.

[7] C. M. Krishna and Y-H Lee. Voltage-Clock-Scaling Adaptive Scheduling Techniques for Low Power Hard Real-Time Systems. *Proceedings of the Sixth IEEE Real Time Technology and Applications Symposium (RTAS 2000), 2000.*

[8] D. Laird, Crusoe Processor Products and Technology http://www.transmeta.com/press/download/pdf/laird.pdf, January 2000.

[9] T. Mudge. Power: A First Class Design Constraint for Future Architectures. *Proceedings of the 7th International Conference on High Performance Computing (HiPC 2000)*, December 2000.

[10] T. Okuma, T. Ishihara, and H. Yasuura. Real-Time Task Scheduling for a Variable Voltage Processor. *Proceedings of the International Symposium on System Synthesis*, November 1999.

[11] D. R. Olsen. *Developing User Interfaces.* Morgan Kaufmann Publishers, 1998.

[12] T. Pering, T. Burd, and R. Brodersen. The Simulation and Evaluation of Dynamic Voltage Scaling Algorithms. *Proceedings of International Symposium on Low Power Electronics and Design 1998*, pp. 76-81, June 1998.

[13] T. Pering, T. Burd, and R. Brodersen. Voltage Scheduling in the lpARM Microprocessor System. *Proceedings of the International Symposium on Low Power Electronics and Design 2000*, July 2000.

[14] Y. Shin and K. Choit. Power Conscious Fixed Priority Scheduling for Hard Real-Time Systems. *Proceedings of the 36th Annual Design Automation Conference, 1999.*

[15] M. Weiser, B. Welch, A. Demers, and S. Shenker. Scheduling for Reduced CPU Energy. *Proceedings of the First Symposium of Operating Systems Design and Implementation*, November 1994.