Dynamic Power Management of Electronic Systems

Luca Benini Alessandro Bogliolo Giovanni De Micheli[†]

DEIS - Università di Bologna [†] CSL - Stanford University

Abstract

Dynamic power management is a design methodology aiming at controlling performance and power levels of digital circuits and systems, with the goal of extending the autonomous operation time of battery-powered systems, providing graceful performance degradation when supply energy is limited, and adapting power dissipation to satisfy environmental constraints.

We survey system-level dynamic power management techniques. We first analyze idleness detection and shutdown mechanisms for idle hardware resources and we review industrial standards for operating system-based power management. We describe system-level stochastic models for the power/performance behavior of systems. We analyze different modeling assumptions and we discuss their validity and generality. Last, we describe methods for determining optimum power management strategies and also describe various validation methods that can be employed to assess the effectiveness of power-manageable architectures and the associated power-management schemes.

1 Introduction

Design methodologies for energy-efficient system-level design are receiving an increasingly large attention. The motivations for such interest are rooted in the widespread use of portable electronic appliances (e.g., cellular phones, laptop computers, etc.) and in the concerns about the environmental impact of electronic systems [21] (whether mobile or not). System-level design must strike the balance between providing high service levels to the users while curtailing power dissipation. In other words, we need to increase the energetic efficiency of electronic systems, as it has been done, by other means, with other types of engines.

Electronic systems are *heterogeneous* in nature, by combining digital with analog circuitry, using semiconductor (e.g., RAM, FLASH memories) and electro-mechanical (e.g., disks) storage resources, as well as electro-optical (e.g., displays) human interfaces. Power management must address all types of resources in a system. The power breakdown for a well-known laptop computer [34] shows that, on average, 36% of the total power is consumed by the display, 18% by the *hard-disk drive* (HDD), 18% by the wireless *local area network* (LAN) interface, 7% by non-critical components (keyboard, mouse etc.), and only 21% by digital VLSI circuitry, mainly memory and *central processing unit* (CPU). Reducing the power in the digital components of this laptop by *10X* would reduce the overall power consumption by less than 20%.

Lowering system-level power consumption, while preserving adequate service and performance levels, is a difficult task. Indeed, reducing system performance (e.g., by using lower clock rates) is not a desirable option when considering the increasingly more elaborate software application programs for computers and features of portable electronic devices. On the other hand, present systems have several components which are not utilized at all times. When such components are *idle*, they can be put in sleep states with reduced (or null) power consumption, with a limited (or null) impact on performance.

Dynamic power management is a design methodology aiming at controlling performance and power levels of digital circuits and systems, by exploiting the idleness of their components. A system is provided with a *power manager* that monitors the overall system and component states and controls state transitions. The control procedure is called power management *policy*. Power managers can be implemented in hardware or in software, depending on system architecture and constraints. When considering general-purpose computer systems, the most natural implementation of the power manager is software-based. In particular, the *operating system* (OS) is the software layer where the manager can be implemented best. Operating system directed power management (OSPM) is actively supported by industry-driven standardization efforts such as Microsoft's *OnNow* initiative [39] and the *Advanced Configuration and Power Interface* (ACPI) standard proposed by Intel, Microsoft and Toshiba [37].

We believe that dynamic power management is a viable approach to reduce power consumption of large-scale systems under performance constraints, because significant power waste is associated with idle resources and because of its general applicability. Note that support for dynamic power management must be provided by the overall system organization, and system architects often envision system partitions that enable power management. Moreover, system components should be *power manageable*, i.e., the manager should be able to control their state of operation. Manageable components can be build by exploiting several specific techniques, such as supply voltage, frequency and activity control [1].

Needless to say, dynamic power management should be complemented by specific chip-level design techniques for power reduction at the architectural level [9], at the RTL, logic and circuit levels [20, 25] and by customized devices and implementation technologies [28]. However, in this survey we focus exclusively on dynamic power management. We consider first system-level design issues, such as idleness detection and shutdown mechanisms for idle resources. We review the OnNow and ACPI standards, as well as previous work in the area of power management. Next, we review system-level modeling techniques, and introduce stochastic models for the power/performance behavior of systems. We analyze different modeling assumptions and we discuss their validity. We then consider a working model, for which optimal policies can be computed, and we discuss how policies can be implemented in electronic systems. Last but not least, we describe several methods for validating the policies, based on simulation at different abstraction levels. We conclude by stressing the need for CAD tools to support model identification, policy optimization and validation for dynamically power-managed systems.

2 System design

In this section we consider issues related to system-level design. We view the system hardware as a collection of resources, we characterize their idleness and present methods for their shutdown. We consider then the interface standards that support resource monitoring and control from the operating system, and we review current related work on dynamic power management.

2.1 Idleness and shutdown mechanisms

The basic principle of a dynamic power manager is to detect inactivity of a resource and shut it down. A fundamental premise is that the idleness detection and power management circuit consumes a negligible fraction of the total power.

We classify idleness as *external* or *internal*. The former is strongly tight to the concept of observability of a resource's outputs, while the latter can be related to the notion of internal state, when the resource has one. A circuit is externally idle if its outputs are not required during a period of time. During such period, the resource is functionally redundant and can be shut down, thus reducing power consumption. A resource is internally idle when it produces the same output over a period of time. Thus, the outputs can be stored and the resource shut down.

While external idleness is a general concept applicable to all types of resources (e.g., digital, analog, memories, hard-disks, displays), internal idleness is typical of digital circuits. Thus, we will be concerned with external idleness detection and exploitation, since we address here system-level design.

There are several mechanisms for shutting down a resource. Digital circuits can be "frozen" by disabling registers (by lowering the enable input) or by gating the clock [1]. By freezing the information on registers, data propagation through combinational logic is halted, with a corresponding power saving. (This saving may be significant in CMOS static technologies, where power is consumed mainly during transitions).

A radical approach to shutdown is to dynamically scale down its supply voltage [2], or to completely turn power off. While this mechanism is conceptually simple and

applicable in general, it usually involves a non-negligible time to restore operation. Note that in some cases the context must be saved before shutdown (e.g., in non-volatile memory) and restored at restart.

Some components can be shut down at different levels, each one corresponding to a power consumption level and to a delay to restore operation. As a first example, consider a backlit display. When the display is used, both the LCD array and the backlighting are on. When the user is idle, the backlighting and/or the LCD array can be turned off with different power savings.

As a second example, a hard-disk drive [40] may have an operational state, in addition to an idle, a low-power idle, a standby, and a sleep state. In the idle states the disk is spinning, but some of the electronic components of the drive are turned off. The transition from idle to active is extremely fast, but only 50-70% of the power is saved in these states. In the standby and sleep states, the disk is spun down, thus reducing power consumption by 90-95%. On the other hand, the transition to the active state is not only slow, but it causes additional power consumption, because of the acceleration of the disk motor.

This example shows the trade-off of power versus performance in dynamic power management. The lower the power associated with a system state, the longer the delay in restoring an operational state. Dynamic power management strategies need to take advantage of the low-power states while minimizing the impact on performance.

2.2 Industrial design standards

Industrial standards have been proposed to facilitate the development of operating system-based power management. Intel, Microsoft and Toshiba proposed the *Advanced Configuration and Power Interface* (ACPI) standard [37]. Although the standard targets *personal computers* (PCs), it contains useful guidelines for a more general class of systems. The characterizing feature of ACPI is that it recognizes dynamic power management as the key to reducing overall system power consumption, and it focuses on making the implementation of dynamic power management schemes in personal computers as straightforward as possible.

The ACPI specification forms the foundation of the *OnNow initiative* [39] launched by Microsoft Corporation. The purpose of *OnNow* is to transform PCs into true household appliances. A PC should appear as off when not in use, but it must be capable of responding with negligible delay to wake-up events (originated by the user or by a resource, such as a modem sensing an incoming call). Furthermore, power consumption in both the on and off state should be as low as possible. OnNow relies on the ACPI infrastructure to interface the software to the hardware components to be managed.

ACPI is an OS-independent, general specification that emerged as an evolution of previous initiatives [38] that attempted to integrate power management features in the low-level routines that directly interact with hardware devices (firmware and BIOS). It is an open standard that is made available for adoption by hardware vendors and operating system developers. The ACPI specification defines interfaces between OS software and hardware. Applications interact with the OS kernel through *application programming interfaces* (APIs). A module of the OS implements the power management policies. The power management module interacts with the hardware through kernel services (system calls). The kernel interacts with the hardware through device drivers. The back-end of the ACPI interface is the *ACPI driver*. The driver is OS-specific, it maps kernel requests to ACPI commands, and ACPI responses/messages to kernel signals/interrupts.

It is important to notice that ACPI specifies neither how to implement hardware devices nor how to realize power management in the operating system. No constraints are imposed on implementation styles for hardware and on power management policies. Implementation of ACPI-compliant hardware can leverage any technology or architectural optimization as long as the power-managed device is controllable by the standard interface specified by ACPI.

ACPI describes the behavior of a PC with an abstract, hierarchical finite-state model. States represent modes of operation of the entire system or its components. Transitions between states are controlled by the OS-based power manager. States and transitions for an ACPI-compliant system are shown in Figure 1. Usually the system alternates between the working ($G0$) and the sleeping ($G1$) states. In the working state the system appears fully operational, but the power manager can put idle devices to sleep (states $D1$ to $D4$). Even the CPU can be put in sleep state ($C1$ to $C3$). When the entire system is idle or the user has pressed the power-off button, the OS will drive the computer into one of the global sleep states on the right side of Figure 1. From the user's viewpoint, no computation occurs.

The sleeping sub-states ($S1$ to $S4$) differ in which *wake* events can force a transition into a working state, how long the transition should take and how much power is dissipated in the state. If the only wake-up event of interest is the activation of the user turn-on button and a latency of a few minutes can be tolerated, the OS could save the entire system context into non-volatile storage and transition the hardware into a soft-off state ($G2$). In this state, power dissipation is almost null and context is retained (in non-volatile memory) for an arbitrary period of time. The mechanical off state ($G3$) is entered in the case of power failure or mechanical disconnection of power supply. Complete OS boot is required to exit the mechanical off state. Finally, the *legacy* state is entered in case the hardware does not support OSPM. It is important to note that ACPI

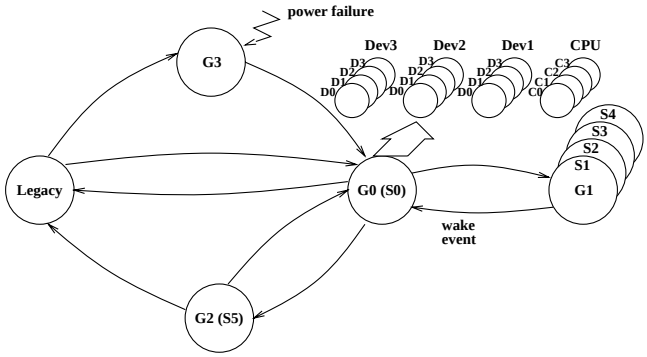

Figure 1: Global and power states and substates

provides only a framework for designers to implement power management strategies, while the the choice of power management *policy* is left to the engineer.

2.3 System-level power management

We consider now work in different areas related to dynamic power management. The common theme is the search of methods for power/performance management. Techniques and application domains vary widely.

Chip-level power management features have been implemented in mainstream commercial microprocessors [11, 12, 13, 16, 31]. Microprocessor power management has two main flavors. First, the entire chip can be put in one of several sleep states through external signals or software control. Second, chip units can be shut down by stopping their local clock distribution. This is done automatically by dedicated on-chip control logic, without user control.

Outside the general-purpose microprocessor arena, designers have been even more aggressive. Specialized power management schemes have been devised for digital signal processors [3], signal processing ASICs [4] DRAM memories [5], and many other applications. Techniques for the automatic synthesis of chip-level power management logic are thoroughly surveyed in [1].

At a much higher level of abstraction, energy-conscious communication protocols based on power management have been extensively studied [19, 27, 29, 35]. The main purpose of these protocols is to regulate the access of several communication devices to a shared medium trying to obtain maximum power efficiency for a given throughput requirement.

Power efficiency is a stringent constraint for mobile communication devices. Pagers are probably the first example of mobile device for personal communication. In [19], communication protocols for pagers are surveyed. These protocols have been designed for maximum power efficiency. Protocol power efficiency is achieved essentially by increasing the fraction of time in which a single pager is idle and can operate in a low-power sleep state without the risk of loosing messages.

Golding et al. considered HDD sub-systems [14, 15], and presented an extensive study of the performance of various disk spin-down policies. The problem of deciding when to spin down a hard disk to reduce its power dissipation is presented as a variation of the general problem of predicting idleness for a system or a system component. This problem has been extensively studied in the past by computer architects and operating system designers (reference [15] contains numerous pointers to work in this field), because idleness prediction can be exploited to optimize performance (for instance by exploiting long idle period to perform work that will probably be useful in the future). When low power dissipation is the target, idleness prediction is employed to decide when it is convenient to spin down a disk to save power (if a long idle period is predicted), and to decide when to turn it on (if the predictor estimates that the end of the idle period is approaching).

The studies presented in [30, 17] target hypothetical "interactive terminals". A common conclusions in these works is that future workloads can be predicted by examining the past history. The prediction results can then be used to decide when and how transitioning the system to a sleep state. In [30], the distribution of idle and busy periods for the interactive terminal is represented as a time series, and approximated with a least-squares regression model. The regression model is used for predicting the duration of future idle periods. A simplified power management policy is also introduced, that predicts the duration of an idle period based on the duration of the last activity period. The authors of [30] claim that the simple policy performs almost as well as the complex regression model, and it is much easier to implement. In [17], an improvement over the simple prediction algorithm of [30] is presented, where idleness prediction is based on a weighted sum of the duration of past idle periods, with geometrically decaying weights. The weighted sum policy is augmented by a technique that reduces the likelihood of multiple mispredictions. A common feature of

these power management approaches is that policies are formulated heuristically, then tested with simulations or measurements to assess their effectiveness.

3 System modeling

In the sequel we consider the hardware part of the system as a set of resources. We model the resources at a very-high level of abstraction, i.e., we view them as units that perform or request specific services and that communicate by requesting and acknowledging such services. Resources of interest are, of course, those that can be power managed, i.e., those that can be set in different states, as in the ACPI scheme.

From a power management standpoint, we model the hardware behavior as a *finite-state system*, where each resource is associated with a set of states and can be in one of the corresponding states. Power and service levels are associated with the different states and transitions among states. In this modeling style, we abstract away the functionality of the resource, and we are concerned only with the ability of the resource to provide and/or request a service.

Because of the high-level of abstraction in resource modeling, it is difficult, if not impossible, to have precise information about power and performance levels of each resource. This uncertainty can be modeled by using random variables for the observable quantities of interest (eg., power, performance), and by considering average values as well as their statistical distributions [1]. This stochastic approach is required to capture both the non-determinism due to lack of detailed information in the abstract resource models as well as the fluctuations of the observed variables due to environmental factors.

With this modeling style, computing optimum dynamic power management policies becomes a *stochastic optimum control* problem [8]. The problem solution, and its accuracy in modeling reality, depend highly on the assumptions we use in modeling. We will discuss next the impact of some modeling assumptions, and then consider is detail a system model under some specific assumption that enables us to compute optimum policies, as shown in Section 4.

3.1 Assumptions

A system model can be characterized by the ensemble of its components, their mode of interaction and their statistical properties.

In general, we can view resources both as providers and requesters of services to other resources. In practice, some resources will be limited to providing or requesting services. We call *system structure* the system abstraction where resources are vertices of a directed graph and where resource interaction is shown by edges. The interaction is the request of a service and/or its delivery. *Queues*, are used to model the accumulation or requests waiting for services [33].

A simple example of a CPU requesting data to a hard-disk drive is shown in Figure 2 (a). A more complex example is reported in Figure 2 (b): it shows a CPU interacting with a LAN interface, a HDD, a display, a keyboard and a mouse. Requests to the CPU can be originated from the keyboard, mouse and LAN interface. Requests to the display come from the CPU (which also forwards requests from the keyboard and mouse). The CPU can request services to the HDD and LAN. Note that the keyboard and mouse models can express also the behavior of the human user who hits their keys and buttons.

A modeling trade-off exists in the *granularity* of the resources, i.e., between the number and average complexity of the resources. Whereas a system model with several resources and an associated structure can capture the interaction of the system components in a detailed way, most researchers view systems with a very coarse granularity. Namely, systems are identified by one resource providing a service, called *service provider*, and one unit requesting a service, called *service requester*. The requester models the *workload source*. This granularity can be used to model systems like user-PC, as shown in Figure 2 (c), where the keyboard and mouse are lumped as a single requester and the CPU, HDD and LAN are seen as a single provider.

Let us consider now the statistical properties of the components of a system. *Stationarity* of a stochastic process means that its statistical properties are invariant to a shift of the time origin [24]. When resources are viewed as providers of services in response to input stimuli, it is conceivable to model their behavior as stationary. Conversely, when resources act as workload sources, and when we model users' requests as such, the stationarity assumption may not hold in general. For example, patterns of human behavior may change with time, especially when considering the fact that an electronic system may have different users. On the other hand, observations of workload sources over a wide time interval may lead to stationary models that are adequately accurate. An advantage of using stationary models is the relative ease of solving the corresponding stochastic optimization problems.

The statistical properties of each component are captured by their distributions. An important aspect is the statistical independence (or dependence) of the resources' statistical models from each others. When a system structure can be captured by disjoint graphs corresponding to statistically-independent resources, the system decomposition allows us to consider and solve independent subproblems. In practice, weak dependencies can sometimes be neglected. Conversely, system structures with many dependencies correspond to complex models requiring a large computational effort to solve the related optimization problems. As a result, the identification of the system resources, interactions and statistics is a crucial step in modeling real systems.

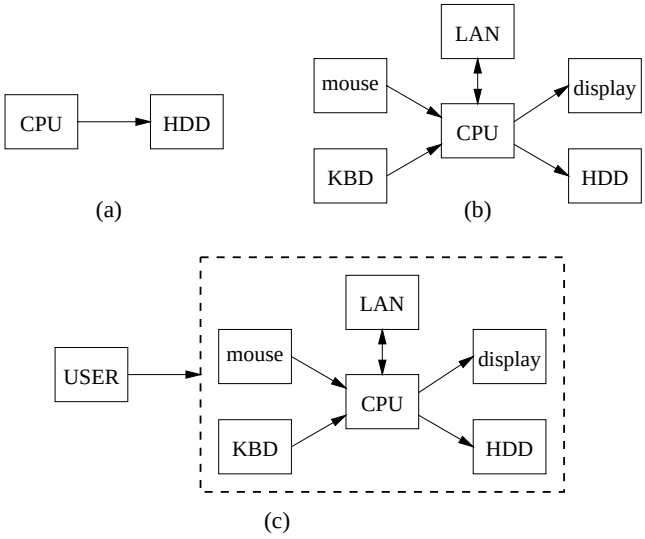

Figure 2: (a) CPU requesting data to a HDD. (b) Simple model of some resources of a personal computer and their interaction. (c) User-PC model where the requests sent by the keyboard and by the mouse are lumped as a single requester and the CPU, HDD, LAN and display are lumped as a single provider.

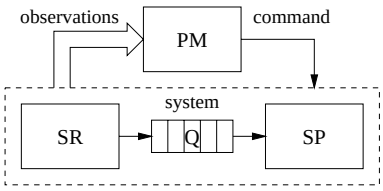

Figure 3: Coarse-grained system model

3.2 A working model

We consider here a working model, with one provider that receives requests through a queue, and that is controlled by a power manager (PM), as shown in Figure 3. This model is described in more detailed in [23]. We summarize here the salient features of the model.

We assume stationary stochastic models service provider (SP), service requester (SR), and queue (Q). We assume also that the service requester is statistically independent from the other components. We consider a discrete-time setting, i.e., we divide time into equally-spaced time slices. We use a parametrized Markov chain model to represent the statistical properties of the system resources. By using the Markov assumption, transition probabilities depend only on the current state and not on the previous history. Moreover, we assume that transition probabilities depend on a parameter, that models the command issued by the power manager. We consider next the system components in detail.

Service provider. It is a device (e.g., HDD) which services incoming requests from a workload source. In each time interval, it can be in only one *state*. Each state $s_p \in \{1, 2, \ldots, S_p\}$ is characterized by a performance level and by a power-consumption level. In the simplest case, we could have two states ($S_p = 2$): *on* and *off*. Otherwise, the states may be more, and in particular match states (and substates) as defined by the ACPI standard. At each timepoint, transitions between states are controlled by a power manager through *commands* $a \in A = \{1, 2, \ldots, N_a\}$. For example, we can define two simple commands: switch on (*s_on*) and switch off (*s_off*). When a specific command is issued, the SP will move to a new state at the next timepoint with a fixed probability dependent only on the command a itself, and on the departure and arrival states. In other terms, after being given a transition command by the power manager, the SP can remain in its current state during the next time slice with a non-zero probability. This aspect of the model takes into account the uncertainty in the transition time between states caused by the abstraction of functional information.

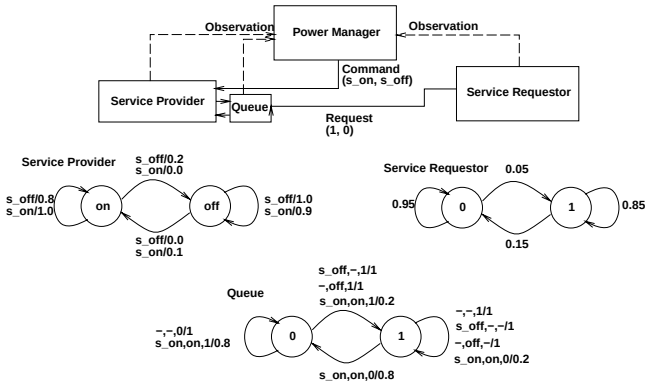

Figure 4: An example of a system model with one service provider, one service requester and one queue,with corresponding Markov chains.

Our probabilistic model is equivalent to the assumption that the evolution in time of states is modeled by a Markov process that depends on the commands issued by the power manager. Each state has a specific *power consumption rate*, which is function both of the state and the command issued. The SP provides service in one state only, that we call active state.

Service requester. It sends requests to the SP. The SR is modeled as a Markov chain, whose state corresponds to the number of requests s_r (with $s_r \in \{0, 1, \ldots, S_r - 1\}$) sent to the SR during time slice of interest.

Queue. It buffers incoming service requests. We define its length to be $(S_q - 1)$. The queue length is also Markov process with state $s_q \in \{0, 1, \ldots, S_q\}$. The state of the queue depends on the state of the provider and requester, as well as on the command issued by the power manager in the time slice of interest.

Power manager. It communicates with the service provider and attempts to set its state at each timepoint, by issuing commands chosen among a finite set A. For example, the commands can be *s_on*, and *s_off*. The power manager contains all proper specifications and collects all relevant information (by observing SP and SR) needed for implementing a power management policy. The consumption of the power manager is assumed to be much smaller than the consumption of the subsystems it controls and it is not a concern here.

The state of the system consisting of {SP,SR,Q} and managed by PM is a triple $s = (s_r, s_p, s_q)$. Being the composition of three Markov chains, s is a Markov chain (with $S = S_r \times S_p \times S_q$ states), whose transition matrix depends on the command a issued by the PM.

Let us consider a simple example, as shown in Figure 4, representing a power-managed HDD. The service requester has only two states, 0 and 1, representing the number of requests per time slice sent to the provider. The queue of the service provider has two states, 0 and 1, representing the number of requests to be serviced. The service provider has two states, *on* and *off*, representing its functional state. When *on*, it services up to one request per time slice taken from the queue. The corresponding power consumption is of 3W. When *off* it does not service any request and it consumes no power. However, a power consumption of 4W is associated with any transition between the two states. SR evolves independently, while the transition probabilities of SP depend on the command issued by the power manager (*s_on*, *s_off*) and those of the queue depend on the states of both SP and SR, as well as on the command. For example, consider the SP in state *on*. (Center-left of Figure 4.) When command *s_on* is issued, the SP will stay in state *on* with probability 1, and transit to state *off* with probability 0. Conversely, when command *s_off* is issued, it will stay in state *on* with probability 0.8, and transit to state *off* with probability 0.2.

3.3 Extensions and limitations

System providers, requesters and queues with several internal states can be modeled in a straightforward way. Power costs and performance penalties can be associated with states and transitions of the Markov models. Thus, the simple model exemplified by Figure 4 can be made more detailed, to capture subtle differences among resource states (e.g., discriminating *soft off* states from *sleeping* states).

Similarly, more complex system structures (with multiple providers, requesters and queues) can be modeled by considering the combined effect of the resources' models. This can be easily done under the hypothesis of statistical independence of the resources' behavior, as in the case of several independent providers responding to a single workload source. In this particular case the overall system model can be derived by composing the Markov chains associated with each resource.

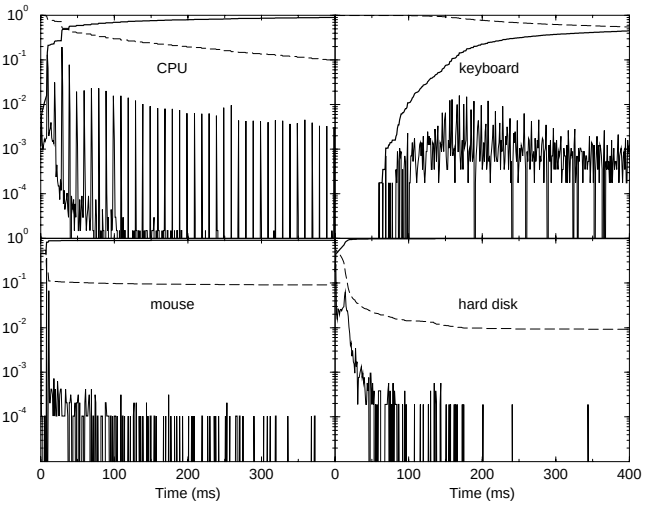

Figure 5: Statistical analysis of the inter-arrival times between service requests for CPU, keyboard, mouse and HDD of a personal computer during software development. For each device, three curves are plotted in lin-log scale: the probability density (solid line), the probability distribution (bold line) and its complement to 1 (dashed line).

Unfortunately, in the general case, the system model is not amenable to a simple decomposition. Consider for example systems such as the one depicted in Figure 2 (b). The interaction among components causes statistical dependence. Most requests to the display from the CPU are triggered by the mouse and keyboard. Thus it is not possible to view the resources as having an independent behavior.

Even when considering systems with simple structures, the identification of the statistical distributions is not a simple matter. The use of i discrete-time stationary Markov models corresponds to use geometric distributions for requests and service times. Such a model may deviate from reality. For example, resources may have known, deterministic service delays compounded with non-deterministic delays depending on the environment.

3.4 Extracting models for the user

System users can be viewed as workload sources and modeled as service requesters. An approach to model the user behavior consists of *monitoring* the system during a user session with no power management and then *extracting* a statistical model of his/her behavior.

System monitoring has to be sufficiently accurate to provide time-stamped traces of service requests. The cumulative counts provided by the system utilities of many computer systems are not sufficient to steer power management. In addition, monitoring has to be non-perturbative in order to affect usage patterns as little as possible. A monitoring system specifically designed for supporting dynamic power management in personal computers is described in reference [6]: the prototype implementation is conceived as an extension of the Linux operating system [32]. The monitoring tool can be configured to collect information about many resources at the same time. Measured overhead for data collection is quite small (around 0.4%). Figure 5 shows usage statistics simultaneously extracted for the CPU, the keyboard, the mouse and the HDD of a personal computer during one-hour of software development.

Once time-stamped request traces have been collected, they are used to characterize the abstract model for the SR. If a discrete-time setting is assumed for modeling, the trace need to be discretized first. For a given time step T, that is usually of the same order of the minimum time constant of the SP, a discretized trace is a stream of integer numbers representing request counts. The k-th number in the stream (i.e., n_k) is the number of requests with time stamps in the interval $[(k-1) \cdot T, k \cdot T]$. According to the definition of SR proposed in Section 3.2, n_k represents the state of the SR at the k-th time step. Characterizing a Markov model for the user consists of tuning the state transition probabilities in order to make the statistical properties of the model as similar as possible to those of the stream. To this purpose, state transition probabilities are directly computed from the discretized trace. For instance, the probability associated with the transition from state $s_r = 0$ to state $s_r = 1$ is obtained as the ratio between the number of $0,1$ sequences in the stream and the total number of 0's.

This simple procedure extracts a Markov model from any trace, but it does not guarantee that the Markov model is statistically significant. The statistical significance of the extracted model can be tested with well-known procedure such as the χ^2 test [33]. If the significant test fails, more complex Markov models can be formulated. In some

cases, it is possible to use a simple Markov model even if it does not perfectly match the statistical properties of the trace. In these cases, all optimization performed on the models should be carefully validated through simulation, as described in Section 5.

User model extraction can be further complicated by dependencies between user and system behavior. In many cases, a change in how the system responds to requests causes a change in how requests are issued. For instance, if a user is typing and the typed characters do not appear immediately on the screen, she/he may type slower. User models constructed by observing the system without power management may be inaccurate if there is a strong depedency between system responses and user requests. Characterizing user models in such systems, which are known as "closed queueing networks", is a challenging task [36].

4 Policy optimization

We consider now the policy optimization problem, for the working model described in Section 3.2. Policy optimization strives at minimizing the average power consumption under performance constraints. Similarly, we can define the complementary optimization of maximizing system performance under a bound on the average power consumption. With the working model of Section 3.2, performance relates to the average delay in servicing a request (i.e., wait time on a hard-disk access). Due to space limitation, we describe only the major steps toward solving the problem. The interested reader is referred to [23] for details.

We need to analyze first how the PM controls the system, to define formally the notion of policy, which is the unknown of the problem to optimize.

At each time point, the power manager observes the history of the system and controls the SP by taking a *decision*. A *deterministic decision* consists of issuing a single command. A *randomized decision* consists of specifying the probability of issuing a command. Randomized decisions include deterministic decisions as special cases (i.e., the probability of a command is 1).

A *policy* is a finite sequence of decisions. A *stationary policy* is one where the same decision (as a function of the system state) is taken at each time point. Note that stationarity means that the functional dependence of the decision on the state does not change over time. Obviously, as the state evolves, the decisions change. *Markov stationary policies* are policies where decisions depend only on the present system state.

The importance of Markov stationary policies stems from two facts: they are easy to implement and it is possible to show that optimum policies belong to this class. Namely, it is possible to prove formally that the aforementioned policy optimization problems have an optimum solution that is a unique randomized Markov stationary policy. In the particular case that either the problem is unconstrained or the constraints are inactive, then the solution is also deterministic [8, 23]. It is possible to show that the policy optimization problem can be cast as a linear program. An intuitive formulation is described here in an informal way. Consider the PM, that observes the system state and issues commands. For each possible pair (state,command), we can compute its *frequency*, i.e., the expected number of times that a system is in that state and issues that command. The frequency is a non-negative number subject to the following *conservation law*. The expected number of times state x is the current state is equal to the expected initial population of x plus the expected number of times x is reached from any other state. Moreover, average power and performance loss can be expressed as linear functions of the $(state, command)$ frequencies. Thus, minimizing power consumption can be expressed as minimizing a linear function of the $(state, command)$ frequencies, under linear constraints.

Overall, linear programs modeling policy optimization can be efficiently solved by standard software packages, for simple topologies and a reasonable number of commands. The policy optimization tool described in [23] is built around *PCx*, an advanced LP solver based on an interior point algorithm [10].

Figure 6 shows the power-performance trade-off curve obtained for the example system of Figure 4 by iteratively solving the policy optimization problem for different performance constraints. Performance is expressed in term of average queue length, that is the average waiting time for a request. An additional constraint is used, called *request loss*, to represent the maximum probability of loosing a request because of a queue-full condition. It is worth noting how the power-performance trade-off is affected by the additional constraint. In particular, if a request-loss lower than 0.1337 has to be guaranteed, the SP can never be shut down. In this case, no power savings can be achieved regardless of the performance constraint.

The trade-off curve for a more complex system is reported in Figure 7. The SP is a commercially-available power-manageable HDD with one active state and four inactive states, spanning the trade off between power consumption and shut-down/wake-up times [42]. The average power consumption of the disk when in the active state is of 2.5W. The SR model was extracted as described in Section 3.4 from the time-stamped traces of disk accesses provided in [41]. A queue of length 2 was used.

Points associated with several heuristic policies are also plotted in the power-performance plane for comparison. Although we cannot claim that our heuristic policies are the best that any experienced designer can formulate, some of them provide power-performance points not far from the trade-off curve. Note that heuristic solutions do not allow the designer to automatically take constraints into account. On the other hand, trial and error approaches may be highly expensive due to the large number of

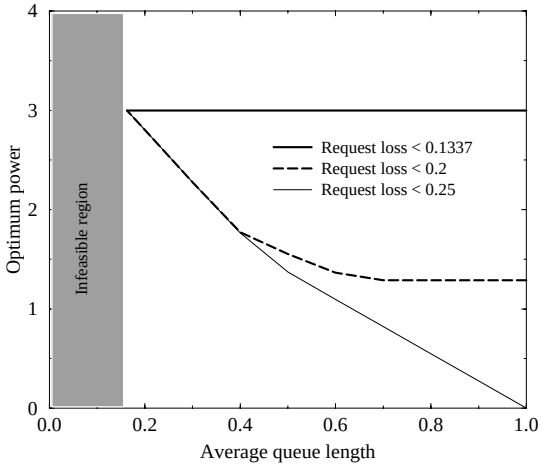

Figure 6: Power-performance trade-off curves for the example system.

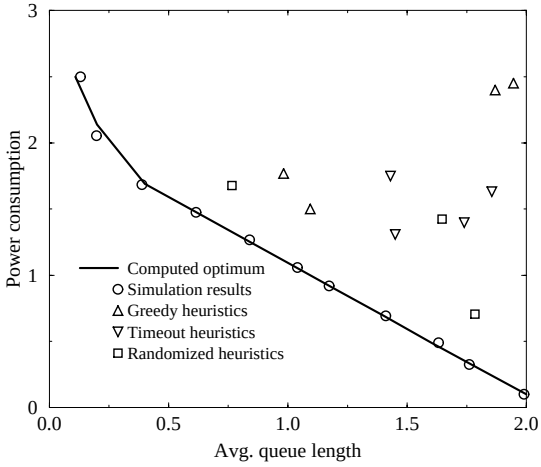

Figure 7: Power-performance feasible trade-off's for a commercially-available power-manageable hard disk.

parameters (in our case study the policy is represented by a 66x5 matrix with 330 entries). Moreover, even if it is possible to produce heuristic policies that produce "reasonable" results, there is no way for the designer to estimate if the results can be improved. For these reasons, computer-aided design tools for policy optimization can be of great help to system designers.

4.1 Power manager implementation

Power management policies can be computed *off-line* or *on-line*. In the former case, a policy is computed once for all for the system being designed, and implemented in hardware or software as described in this section. Alternatively, several policies can be computed off-line and stored, each corresponding to a different environmental factor, such as a workload source. The power manager can switch among the policies at run time. On-line policy computation is also possible. Once the power manager has identified a change of the environmental conditions that make the current policy no longer effective, a new policy can be computed which takes into account the new environmental parameters (e.g., request arrival rate). Once the policy is computed, it can be executed until the power manager deems it appropriate.

In the case of simple systems, it may be practical to implement the dynamic power management policy as a hardware control circuit. Since circuit synthesis methods are currently used for hardware design, policy implementation consists of representing the policy in a synthesizable *hardware description language* (HDL) model for the power manager. In general, the circuit input is the system state and the output are the commands.

Deterministic policies can be implemented by table look-up schemes. Randomized policies require storing the conditional probabilities of issuing a command in any given state and comparing them with a pseudo-random number, which can be generated by using a linear feedback shift register (LFSR). The command probabilities should be normalized to the length of the LFSR. In particular, when only two commands are possible (e.g., *s_on* and *s_off*), their conditional probabilities sum up to 1 and thus only one probability needs to be stored. The binary outcome of the comparison with a pseudo-random number corresponds to the chosen command. This scheme can be easily extended to handle N_a commands by means of a table with $N_a - 1$ entries per state and $N_a - 1$ comparisons with the pseudo-random number, which can be executed in parallel.

The implementation of policies in software requires the software synthesis of the power manager (e.g., the generation of a C program that issues the commands as a function of the system state) as well as its embedding in the operating system. In the case of randomized policies, the program should make use of a pseudo-random number generator for deciding which command should be issued. The power manager may be executed in kernel mode and be synchronized and/or merged with the OS task scheduler to reduce the performance penalty due to context switch.

5 Validation

In this section we address the problem of bridging the gap between the high level of abstraction at which policy optimization is performed and the real-world systems, where optimal policies have to be applied. In Section 3 we have described a general approach for modeling power-manageable systems as interacting Markov processes. In Section 4 we have shown that such an abstract model allows us to cast the policy optimization problem as a linear program that can be solved in polynomial time. All modeling assumptions made to formulate and solve the policy optimization task need to be tested in order to validate its results. We briefly describe validation techniques based on simulation and emulation at different abstraction levels, ranging from the direct simulation of the Markov models used for optimization to the actual implementation of the optimal policies in the target systems. We discuss the main strengths and the inherent limitations of each approach.

Discrete-time simulation of Markov processes. Discrete-time simulation is performed at the same abstraction level used for optimization. The simulator takes the policy and the Markov models of the components and iteratively performs the following steps: *i*) take a decision (based on the current state), *ii*) evaluate cost/performance metrics, *iii*) evaluate the next state of all components, *iv*) increment time, update the state and iterate. Notice that both the policy and the next-state functions of the Markov chains are *non-deterministic discrete functions* (NDFs): inputs are present-state variables and commands, whereas outputs are the outcomes of random processes. NDFs can be represented as matrices having as many rows as input configurations and as many columns as output values. Entries represent the conditional probabilities of all possible outcomes for all given input configurations. To evaluate a function, the row associated with the current input configuration is selected and a pseudo-random number (uniformly distributed between 0 and 1) is generated and compared to the entries in the row to select the actual command.

Needless to say, this simulation paradigm cannot be used to validate the policy against the modeling assumptions of Section 3, since it relies on them as well. However, it provides valuable information about the time-domain system behavior. Constraints and objective functions used for optimization are average expected values of the performance/cost metrics of interest. Simulation allows us to monitor the instantaneous values of such parameters (to detect, for instance, the temporary violations of performance constraints) and to measure their variance.

Discrete-time simulation with actual user traces. The simulation paradigm is the same described in the previous paragraph. The only difference is that the model of the service requester is now replaced by a trace taken from a real-world application. At each time step, the present state of the SR is read from the trace, instead of being non-deterministically computed from the previous one.

Though the abstraction level is still very high, trace simulation allows us to remove all assumptions on the time distribution of service requests. As a result, it can be used to check the validity of the Markov model used for the SR during optimization.

Discrete-time simulation with real request traces was performed to validate the trade-off curve of Figure 7. Simulation results are denoted by circles in figure. The small distance of the circles from the solid-line curve is a measure of the quality of the SR Markov model extracted from the user traces and used for optimization.

Event-driven stochastic simulation. In event-driven simulation, model evaluation is no longer periodic. The model of each component is re-evaluated only when an event (i.e., a change) occurs on some of the state/command variables it depends on. The evaluation of a component may produce new events to be scheduled at a future time. Both the output events and their scheduling times may be non-deterministic. For instance, the command issued by a randomized policy can be modeled as an instantaneous non-deterministic event, while the transition between two states of the SP can be viewed as a deterministic event (if the next state is uniquely determined once a command has been issued) to be scheduled at a non-deterministic time (if the transition time is a random variable). The scheduling time is pseudo-randomly chosen according to a given probability distribution. An event-driven stochastic simulator is described in [7].

The main advantage of the event-driven paradigm is that it can easily handle stochastic processes with arbitrary distributions. Conversely, discrete-time simulation is implicitly based on the memory-less assumption that is behind Markov models, that allows us to represent and simulate only geometrically-distributed random variables. Adding memory information to a Markov model in order to represent different distributions is not a practical solution since it causes the exponential increase of the number of states. Event-driven simulation provides a more practical way of applying optimal policies to arbitrary SP models in order to check the validity of the Markov model used for optimization.

Fully-functional simulation. The functionality of a system can be described at many levels of abstraction. Functional simulation can be performed at any level. Here we focus on *cycle-accurate* simulation, that is the most accurate simulation paradigm that can be used to handle systems as complex as a personal computer. Cycle-accurate simulation matches the behavior of the real system at clock boundaries. When the system is a computer, cycle-accurate simulation provides enough detail to boot an operating system and run an actual workload on top of it. A fully-functional simulator specifically designed to study computer systems is *SimOS* [26], that can handle multi-processor architectures and provides models for simulating commercial microprocessors, peripherals and operating systems.

When system functionality comes into the picture, most of the simplifying modeling assumptions can be eliminated. In particular, stochastic models for SP and SR are no longer required, since even their functionality can be exactly simulated. Performance penalties can be realistically estimated and accurate cost metrics (i.e., power consumptions) can be associated with the operating states of the resources. In addition, functional simulation realizes a unique trade-off between realism and flexibility. On one hand, it provides a means of validating the policies against the real world and gives the designer a direct hands-on experience of most of the implementation issues involved in OS-directed power management. On the other hand, it allows the designer to explore the entire design space, balancing hardware and software solutions.

The main drawback of functional simulation is performance: simulation times may be more than three orders of magnitude slower than the run times on the corresponding real system, making the approach impractical to study complex workloads.

Emulation. We use the term *emulation* to denote a validation approach that uses functionally-equivalent real hardware components to exercise the behavior of part of the system. In particular, we are interested in using a computer without power-management features as the hardware platform to emulate a power-managed functionally-equivalent one. As an example, suppose that we are designing a power-management policy for the HDD of a laptop computer, having one active state and several inactive states (with different power consumptions and wake-up times). If such a HDD is not available for validation, the power-managed system can be emulated on an equivalent computer (with the same workload of the target one) with a non-power-manageable HDD. As long as the device used for emulation has the same performance of the target one, it can be employed to emulate the active-state functionality, while inactive states (and transitions between them) can be simulated by the software device driver. The code of the original device driver needs a few changes: *i*) an additional state variable representing the power state, *ii*) a routine for updating the power state according to power-management commands, *iii*) a timer to simulate state transition times, *iv*) a routine to provide power consumption estimates and *v*) a request-blocking mechanism that enables actual accesses to the disk only when in the active state. In general, emulation of power-managed systems is based on the observation that dynamic power management can only reduce system performances. Hence, if a functionally-equivalent real system is available for exercising the active-state performance, lower-performance states can be emulated as well.

Emulation has two desirable features. First, it runs at the same speed of the actual system, thus enabling policy validation against realistic workloads of any complexity and real-time interactive user sessions. This gives the user a direct experience of performance degradations possibly induced by power management. Second, it enables the software specification of the low-power states of the SP. The possibility of easily changing the SP model can be exploited both during the design of a power-manageable resource, to verify the effectiveness of a given low-power state, and during system-level design, to select among equivalent power-manageable components. The main drawback with respect to simulative approaches is that the system architecture is assigned once for all: no architectural choices can be explored.

Implementation. Policies can be validated by testing their implementations. Since the policy is directly applied to the target system, its actual impact on the cost metrics of interest can be measured accurately. Thus experimentation at this level is useful as a final step in validating a given policy.

6 Conclusions

Dynamic power management is an effective means for system-level design of low-power electronic systems. Dynamic power management is already widely applied to system design, but today most electronic products rely on ad-hoc implementation frameworks (e.g., firmware code) and on heuristic management policies (e.g., timeout policies). We expect that the use of industrial standards, such as OnNow and ACPI, will soon facilitate the clean implementation of operating system based power management.

This survey has shown how systems can be modeled so that optimal management policies can be computed, validated and implemented. The computation of optimal policies is a new problem for system-level design. In particular, we have shown a working model for which the optimal stochastic power-management control problem can be efficiently and exactly solved. The solution method we have analyzed relies on a modeling abstraction of system resources in terms of Markov processes. Several extension can be made to the model, at the price of complicating the solution procedure, by considering more detailed system models. As in many design problems, good engineering judgment is key in determining the right balance among model accuracy, exactness of the solution (for the given model), and computational effort.

Due to the proliferation of handheld electronic systems, and due to increasingly stringent environmental constraints on non-mobile systems, we believe that designers will be very often confronted with the challenge of deriving optimal, or near optimal, dynamic power management solutions. As a result, computer-aided design tools for power management will be extremely useful is system-level design for model identification, policy optimization and validation.

7 Acknowledgments

We acknowledge support from NSF under contract MIP-942119. We would like to thank Eui-Young Chung, Yung Hsiang Lu, Giuseppe Paleologo and Tajana Simunic, at Stanford, for several discussions on this topic.

References

[1] L. Benini and G. De Micheli, *Dynamic Power Management of Circuits and Systems: Design Techniques and CAD Tools*, Kluwer, 1997.

[2] G. Wei and M. Horowitz, "A low power switching power supply for self-clocked systems," in *International Symposium on Low Power Electronics and Design*, pp. 313–317, Aug. 1996.

[3] P. Landman, W. Lee et al., "Design of a 1-V programmable DSP for wireless communication," in *Low Power CMOS design*, IEEE Press, 1998.

[4] B. Gordon and T. Meng, "A 1.3 mW video-rate 2D color subband decoder," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 12, pp. 1510-1516, Dec. 1995.

[5] D. Takashima, S. Watanabe et al., "Standby/active mode logic for sub-1V operating ULSI memory," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 4, pp. 441–447, April 1994.

[6] L. Benini, A. Bogliolo, S. Cavallucci and B. Riccò, "Monitoring System Activity for OS-directed Dynamic Power Management," *Proceedings International Symposium on Low-Power Design*, 1998.

[7] L. Benini, R. Hodgson and P. Siegel, "System-Level Power Estimation and Optimization," to appear in *Proceedings of Int.l Symposium of Low-Power Electronics and Design*, 1998.

[8] D. Bertsekas, "Dynamic programming and optimal control," (2 Vols.) Athena Scientific, 1995.

[9] A. Chandrakasan and R. Brodersen, *Low-Power Digital CMOS Design*. Kluwer, 1995.

[10] J. Czyzyk, S. Mehrotra, and S. Wright, "PCx User Guide", *Technical Report OTC 96/01*, Optimization Technology Center, May, 1996.

[11] G. Debnath, K. Debnath and R. Fernando, "The Pentium Processor-90/100, Microarchitecture and Low-Power Circuit Design," in *International conference on VLSI design*, pp. 185–190, Jan. 1995.

[12] S. Furber, *ARM System Architecture* Addison-Wesley, 1997.

[13] S. Gary, P. Ippolito et al., "PowerPC 603, a Microprocessor for Portable Computers," *IEEE Design & Test of Computers* vol. 11, no. 4, pp. 14-23, Win. 1994.

[14] R. Golding, P. Bosch and J. Wilkes "Idleness is not Sloth", in *Proceedings of Winter USENIX Technical Conference*, pp. 201–212, 1995.

[15] R Golding, P. Bosh and J. Wilkes, "Idleness is not Sloth" *HP Laboratories Technical Report* HPL-96-140, 1996.

[16] M. Gowan, L. Biro, D. Jackson, "Power Considerations in the Design of the Alpha 21264 Microprocessor," *DAC - Proceedings of the Design Automation Conference*, 1998, pp. 726-731.

[17] C.-H. Hwang and A. Wu, "A Predictive System Shutdown Method for Energy Saving of Event-Driven Computation", in *Proceedings of the Int.l Conference on Computer Aided Design*, pp. 28-32, 1997.

[18] H. Kapadia, G. De Micheli and L. Benini, "Reducing Switching Activity on Datapath Buses with Control-Signal Gating," *CICC - Proceedings of the Custom Integrated Circuit Conference*, pp. 589-592, 1998.

[19] B. Mangione-Smith, "Low-Power Communication Protocols: Paging and Beyond," *IEEE Symposium on Low-Power Electronics*, pp. 8–11, 1995.

[20] J. Monteiro and S. Devadas, *Computer-Aided Techniques for Low-Power Sequential Logic Circuits*. Kluwer 1997.

[21] B.Nadel, "The Green Machine," *PC Magazine*, Vol 12, No. 10, p.110, May 25, 1993.

[22] W. Nebel and J. Mermet (Eds.), *Low-Power Design in Deep Submicron Electronics*. Kluwer, 1997.

[23] G. Paleologo, L. Benini, A. Bogliolo and G. De Micheli, "Policy Optimization for Dynamic Power Management," *DAC - Proceedings of the Design Automation Conference*, 1998, pp. 182-187.

[24] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, 1984.

[25] J. M. Rabaey and M. Pedram (editors), *Low-Power Design Methodologies*. Kluwer, 1996.

[26] M. Rosenblum, E. Bugnion, S. Devine and S. A. Herrod, "Using the SimOS Machine Simulator to Study Complex Computer Systems," *ACM Transactions on Modeling and Computer Simulation*, Vol. 7, no. 1, pp. 78-103, 1997.

[27] J. Rulnick and N. Bambos, "Mobile Power Management for Wireless Communication Networks," *Wireless Networks*, vol. 3, no. 1, pp. 3–14, 1997.

[28] T. Sakurai and T. Kuroda, "Low-Power Circuit Design for Multimedia CMOS VLSI," in *Workshop on Synthesis and System Integration of Mixed Technologies*, pp. 3–10, Nov. 1996.

[29] K. Sivalingham, M. Srivastava et al., "Low-power Access Protocols Based on Scheduling for Wireless and Mobile ATM Networks," *Int.l Conference on Universal Personal Communications*, pp. 429–433, 1997.

[30] M. Srivastava, A. Chandrakasan. R. Brodersen, "Predictive System Shutdown and Other Architectural Techniques for Energy Efficient Programmable Computation," *IEEE Transactions on VLSI Systems*, vol. 4, no. 1, pp. 42–55, March 1996.

[31] V. Tiwari, D. Singh, S. Rajgopal, G. Metha, R. Patel and F. Baez, "Reducing Power in High-Performance Microprocessors," *DAC - Proceedings of the Design Automation Conference*, 1998, pp. 732-737.

[32] L. Torvalds, "Linux Kernel Implementation," *Proceedings of Open Systems. Looking into the future. AUUG'94*, pp. 9-14, 1994.

[33] K. Trivedi, *Probability and Statistics with Reliability, Queueing and Computer Science Applications*, Prentice Hall, 1982.

[34] S. Udani and J. Smith, "The Power Broker: Intelligent Power Management for Mobile Computing," *Technical report MS-CIS-96-12*, Dept. of Computer Information Science, University of Pennsylvania, May 1996.

[35] M. Zorzi and R. Rao, "Energy-Constrained Error Control for Wireless Channels," *IEEE Personal Communications*, vol. 4, no. 6, pp. 27–33, Dec. 1997.

[36] R. Onvural, "Survey of closed queueing networks with blocking," *ACM Computing Surveys*. vol. 22, no. 2, pp. 83–121, June 1990.

[37] *http://www.intel.com/ial/powermgm/specs.html*, Intel, Microsoft and Toshiba, "Advanced Configuration and Power Interface specification", Dec. 1996.

[38] *http://developer.intel.com/IAL/powermgm/apmovr.htm*, Intel, "Advanced Power Management Overview," 1998.

[39] *http://www.microsoft.com/hwdev/pcfuture/ ONNOW.HTM*, Microsoft, "OnNow: the evolution of the PC platform," Aug. 1997.

[40] *http://www.storage.ibm.com/storage/oem/data/ travvp.htm* Technical specification of hard-drive IBM Travelstar VP 2.5-inch, 1996.

[41] *http://now.cs.berkeley.edu/Xfs/AuspexTraces/auspex.html*, Auspex File System Traces, 1993.

[42] *http://www.storage.ibm.com/storage/oem/data/travvp.htm*, Hard Drive IBM Travelstar VP 2.5-inch, 1996.