Energy Minimization with Guaranteed Quality of Service

Gang Qu and Miodrag Potkonjak
Computer Science Department, University of California, Los Angeles, CA 90095

Abstract

Quality of service (QoS) is one of the key features for new Internet-based multimedia and other applications. Meanwhile, energy remains as a big concern for systems that perform such applications. We address the issue of combining system design concerns and QoS requirements to design systems that can deliver QoS guarantees. In this paper, we discuss how to satisfy QoS requirements and minimize the system's energy consumption. Specifically, we consider the following problem: Given a set of applications each specifying its required amount of computation and service time, how we allocate CPU time and determine the voltage profile on a variable voltage system, such that all the applications' requirements are satisfied and the system's total energy consumption is minimized. We optimally solve several basic cases and propose a dynamic programming procedure for the general case. Simulation shows that the new approach saves 38.75% energy over the system shut-down technique.

1. INTRODUCTION

With the advances in the Internet, mobile and wireless communications, more and more complex applications are becoming feasible. These applications have various types of requirements on the quality of service (QoS). Meanwhile, low power consumption is considered one of the most important criteria for the design of application specific integrated circuits (ASIC) and other mobile computing devices, the core of systems that carry out these applications. Clearly, there is a trade-off between high QoS and low power consumption. In this paper, we discuss how to minimize the energy consumption under the constraint of meeting all applications' QoS requirements.

We define the *computation vs. service time* function $C(t)$ as: to satisfy user's requirement, if the service time is t, then $C(t)$ is the minimal amount of service (computation) that the system has to provide. The Energy Minimization (EM) problem is: *Given n applications $\tau_1, \tau_2, \cdots, \tau_n$ each with its computation vs. service time specification $C_i(t)$, find for each i, the service time and supply voltage such that τ_i's requirement is satisfied and $\sum_{i=1}^{n} E_i$ is minimized, where E_i is the energy consumption for serving τ_i.*

One of the key techniques to lower energy consumption is using low supply voltage. Suppose one application is assigned the CPU in the period $[0, T]$, with the flexibility provided by the variable voltage processor [4], the supply voltage can be changed as long as the same amount of work is done by time T. However, from the convexity of energy as a function of supply voltage, we have the following necessary condition for energy consumption to be minimized:

Theorem 1.1 ([5])
To finish an application's computation requirement within a given amount of time, the energy consumption is minimized if and only if the processor operates at a constant voltage such that the computation is finished on its deadline.

Assuming the total CPU time is 1, we call a system *underloaded* if at nominal voltage $V_{nominal}$, the total CPU time required to meet all the applications' QoS requirements is less than 1. On such system, the EM problem is trivial to solve when all the applications' computation *vs.* service time functions are constants. That is, each user only requires the amount of service and does not care about the service time.

Let $t_i (0 < t_i < 1)$ be the CPU time for the system to accumulate the amount of computation W_i required by application $\tau_i (1 \le i \le n)$ at $v_{nominal}$, where $W_i = S_{nominal} \times t_i$ and $S_{nominal}$ is the CPU speed at the nominal voltage. For underloaded system, we have $T = \sum_{i=1}^{n} t_i < 1$. Then immediately from Theorem 1.1, the optimal strategy is to use a constant voltage v_{opt} such that a total amount of computation $\sum_{i=1}^{n} W_i$ is completed at the end of period $[0, 1]$. v_{opt} can be trivially solved from:

$$1 = \frac{S_{opt} \cdot 1}{S_{nominal} \cdot \sum_{i=1}^{n} t_i} = \frac{(v_{opt} - v_t)^2}{v_{opt} \cdot T} \cdot \frac{v_{nominal}}{(v_{nominal} - v_t)^2} \quad (1)$$

Corollary 1.2
For applications with constant computation *vs.* service time functions, an underloaded system consumes the minimal energy when it operates at a fixed supply voltage v_{opt} for all the applications, where v_{opt} is given by equation (1).

Users concern about and can observe only the quality of service they receive, not how this service is provided by the system. Their service requirement usually can be met by different combinations of service quality and service time, which may not require the same amount of computation. So the system can take advantage of this and select the one that consumes the least amount of energy. The problem becomes non-trivial for a system with limited CPU time to serve multiple applications.

2. RELATED WORK

Most of the research on quality of service is in the networking community, especially in distributed multimedia systems. There have been several proposals and prototype implementations of end-to-end transport protocols for delivering QoS

ISLPED '00, Rapallo, Italy.

guarantees. For example, RSVP provides a mechanism for reserving resources along the path from a source host to a destination host so that subsequent data packets are guaranteed to have certain bandwidth available and meet certain delay bounds.

There are also plenty of literatures on how to define the concept of QoS. In [1], QoS is defined as a combination of the basic quality metrics for the network layer: delay, jitter, bandwidth, and reliability. Lawrence [7] discusses the metrics based on the QoS attributes of timeliness, precision, and accuracy that can be used for system specification, instrumentation, and evaluation. Rajkumar et al. [10] present an analytical approach for satisfying multiple QoS dimensions in a resource-constraint environment. The quality of the complex, real-time, distributed multimedia services should be application specific and user dependent. Thus it is hard to find an explicit one-fit-all definition for QoS. In our model, we treat QoS, and hence the system's utility, as a function of the required resources such as bandwidth, CPU time, and buffer space. No specific assumptions are required for this function except in some cases, we assume it is monotone and non-decreasing with respect to the resources.

Low power design has attracted a large amount of attention and many techniques have been proposed. For instance, memory optimization techniques, hardware-software partitioning, instruction-level power optimization, variable-voltage techniques, control-data-flow transformations, dynamic power management, interface power minimization. One of the most effective way to reduce power consumption is to lower supply voltage. Recent advances in power supply technology make it possible to create processor cores with supply voltages that can be varied at run time according to application timing constraints[2, 3, 8].

Until now, synthesis and CAD research did not address how to design systems with quantitative QoS requirements. Kornegay et al. [6] outline foundations and framework in which QoS system design trade-offs and optimization can be addressed. In this paper, we show how energy consumption, one of the key design concerns, can be minimized while delivering the QoS guarantees. We adopt an abstract QoS model with little restrictions. Therefore, both our approach and results are general and can be applied to most systems.

3. PROBLEM FORMULATION

We want to build a system that can provide QoS guarantees to a set of applications. The main component of our system is a variable voltage processor core, which is capable of running at a range of supply voltages. Suppose the supply voltage is $v_{dd}(t)$ at time t, then the power dissipation is $P(t) = \alpha C v_{dd}^2 f$ and the circuit delay is $\frac{k v_{dd}}{(v_{dd}-v_t)^2}$, where v_t is the threshold voltage. The energy consumption over the period $[0,T]$ is $E = \int_0^T P(t)dt$ ([9]). Besides the core, there are $m \geq 1$ resources $\{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_m\}$ available and each resource has a finite capacity, also denoted by $\mathbf{R}_i$ if there is no ambiguity. Such resource can be CPU time, memory, disk bandwidth, and etc.

On the other side, we have $n \geq 1$ applications $\{\tau_1, \tau_2, \cdots, \tau_n\}$ to be executed on the above system. Each application has its QoS specification and the system accrues a certain value, which we call the system's utility $U^i_{v_i}(\mathbf{R}^i)$, by allocating resources $\mathbf{R}^i = (R_{i,1}, \cdots, R_{i,m})$ to application τ_i along with supply voltage $v_i(t)$. If $[0,T]$ is the period to achieve $U^i_{v_i}$, then the energy consumption is $E_i(\mathbf{R}^i, v_i) = \int_0^T P(v_i(t))dt$. Notice that the execution time of an application varies as the processor's speed, which is determined by the supply voltage, changes.

We propose the problem of energy minimization with guaranteed QoS as a constrained optimization problem (Figure 1). Our objective is to allocate resource and find voltage

Problem EM: Energy Minimization with guaranteed QoS

Minimize: $\sum_{i=1}^{n} E_i(\mathbf{R}^i, v_i)$ (i)

Subject to: $U^i_{v_i}(\mathbf{R}^i) \geq U^i$ for $1 \leq i \leq n$ (ii)

$\sum_{i=1}^{n} R_{i,j} \leq \mathbf{R}_j$ for $1 \leq j \leq m$ (iii)

Figure 1: Problem formulation.

profiles such that the system's total energy consumption (i) is minimized. Each application's minimal QoS requirement U^i $(1 \leq i \leq n)$ has to be met (ii) within the capacity of each resource (iii).

4. APPLICATION WITH DIFFERENTIABLE UTILITY FUNCTION

Earlier, we solved the energy minimization problem optimally under the assumption that all the utility functions are determined solely by the amount of computation and are independent of the service time (Corollary 1.2). However, this assumption does not hold in many occasions. In general, the application's utility function is complicated and involves many variables. We extend the previous discussion to the EM problem with general utility functions which depend on both the service time and the amount of service. Since the amount of service is the product of system speed S and the service time T, we further assume the utility function $U = U(S,T)$ is differentiable and $\frac{\partial U}{\partial S} > 0, \frac{\partial U}{\partial T} > 0$.

In the simplest case, there is only one application with a utility function $U = U(S,T)$, and the system wants to achieve utility at the amount of U_0 by providing service to this application. It is clear that to minimize the energy consumption, we will operate the system at a constant voltage and serve the application at exactly U_0.

Suppose the energy consumption is minimized with a supply voltage $v > v_t$, the system's speed

$$S = S(v) = \frac{k_1 \cdot (v - v_t)^2}{v} \tag{2}$$

is then fixed, so is the power consumption

$$P = P(v) = k_2 \cdot v \cdot (v - v_t) \tag{3}$$

Since the utility function is monotone increasing with respect to service time for a fixed speed ($\frac{\partial U}{\partial T} > 0$), and achieves 0-utility when the service time T equals 0, there exists a unique service time $T = T(v)$ such that the utility guarantee U_0 can be accumulated at T. The energy consumption

is given by

$$E = E(v) = P \cdot T = k_2 \cdot v \cdot (v - v_t)^2 \cdot T(v) \quad (4)$$

Taking the first derivative of $E(v)$, we have

$$\frac{dE}{dv} = k_2 \cdot v \cdot (v - v_t)[(3v - v_t) \cdot T(v) + v \cdot (v - v_t) \cdot \frac{dT}{dv}] \quad (5)$$

When the utility function $U(S,T)$ is explicitly given, the value of $\frac{dT}{dv}$ can be determined from an earlier observation: when the energy is minimized, the system has a utility exactly $U(S(v),T(v)) = U_0$. Taking the total differential and from equation (2), we then get

$$\frac{dT}{dv} = -\frac{k_1(v^2 - v_t^2)}{v^2} \cdot \frac{\partial U/\partial S}{\partial U/\partial T} \quad (6)$$

It is clear that $\frac{dT}{dv} < 0$ (since $\frac{\partial U}{\partial S} > 0$, $\frac{\partial U}{\partial T} > 0$ and $v > v_t$). On one hand, this simply means for the same utility requirement, the higher voltage we apply, the less time we need to deliver the guarantee. On the other hand, the negative natural of $\frac{dT}{dv}$ also makes the sign of $\frac{dE}{dv}$ undetermined. Moreover, although low energy is in favor of low supply voltage [5], we cannot draw the same conclusion when the QoS guarantee is added as another constraint. However, in the special case when the utility function is explicitly given and differentiable, we can compute $\frac{dT}{dv}$ from equation (6) and then plug it into (5) to calculate the optimal voltage level.

Theorem 4.1
The EM problem for single application system is optimally solvable if the utility function is explicitly given and is differentiable.

Example 4.2
Determine the optimal strategy for an application whose utility function is given by $U(S,T) = S^p T^q, (p,q > 0)$.

First from equation (6),

$$\begin{aligned} \frac{dT}{dv} &= -\frac{k_1(v^2 - v_t^2)}{v^2} \cdot \frac{\partial U/\partial S}{\partial U/\partial T} \\ &= -\frac{k_1(v^2 - v_t^2)}{v^2} \cdot \frac{pT}{qS} \\ &= -\frac{v + v_t}{v(v - v_t)} \cdot \frac{pT}{q} \end{aligned}$$

hence

$$\begin{aligned} \frac{dE}{dv} &= k_2 \cdot v \cdot (v - v_t)[(3v - v_t) \cdot T + v \cdot (v - v_t) \cdot \frac{dT}{dv}] \\ &= k_2 \cdot v \cdot (v - v_t)[(3v - v_t) \cdot T - (v + v_t) \cdot \frac{pT}{q}] \\ &= k_2 \cdot v \cdot (v - v_t) \cdot T[(3 - \frac{p}{q})v - (1 + \frac{p}{q})v_t] \end{aligned}$$

Let $r = \frac{p}{q} > 0$, we can see (recall that $v > v_t$):

(i) if $r \leq 1$, then $\frac{dE}{dv} > 0$, so we choose the minimal possible voltage.

(ii) if $r \geq 3$, then $\frac{dE}{dv} < 0$, so we choose the maximal possible voltage, the nominal voltage.

(iii) if $1 < r < 3$, then $\frac{dE}{dv}|_{\frac{1+r}{3-r} \cdot v_t} = 0$ and it can be verified that $E(v)$ achieves its minimum at $v = \frac{1+r}{3-r} \cdot v_t$. □

Example 4.2 simply suggests that different strategies should be applied for different types of applications which are characterized by r, the relative importance factor of speed *vs.* service time. For applications that prefer a long service time rather than a high speed (i.e. the case when $r \leq 1$), we will use low supply voltage and execute slowly to gain the required amount of utility such that the energy consumption is minimized. On the other hand, if high speed is crucial to the application (e.g., the case when $r \geq 3$), we will sacrifice service time for high speed by operating at a high supply voltage, the energy is saved from the fact that we accomplish the desired amount of utility in much less time. When neither the speed nor the execution time is superior to the other significantly, we will experience the trade-off between speed and service time. Starting from the lowest possible voltage, since long service time is not important enough ($r > 1$), an increment in voltage will speed up the service and energy is saved due to the less execution time. At $v = \frac{1+r}{3-r} v_t$, energy consumption reaches the minimum and further increment in the voltage will consume more energy because high speed is not important either ($r < 3$). Notice that when $r > 1$, the result is against the fact that low power is in favor of low supply voltage for system design without QoS guarantees.

Example 4.3
Prove Corollary 1.2: For a set of applications on an underloaded system, if all the utility functions depend only on the amount of computation, the optimal voltage is the minimal possible voltage.

When the utility depends only on the amount of computation, providing QoS guarantee for each application is the same as performing a computation in the amount of the total of all the applications' requirement. Computation is the product of speed S and service time T for a constant supply voltage. Taking total differential on $S \cdot T = constant$ we have

$$\frac{dT}{dv} = -\frac{k_1(v^2 - v_t^2)}{v^2} \cdot (-\frac{T}{S}) = \frac{v + v_t}{v(v - v_t)} \cdot T$$

and therefore

$$\begin{aligned} \frac{dE}{dv} &= k_2 \cdot v \cdot (v - v_t)[(3v - v_t) \cdot T + v \cdot (v - v_t) \cdot \frac{dT}{dv}] \\ &= 4 \cdot k_2 \cdot v^2 \cdot (v - v_t) \cdot T \\ &> 0 \end{aligned}$$

so the optimal strategy is to use the minimal possible voltage. □

5. APPLICATION WITH GENERAL UTILITY FUNCTIONS

In the general case, we cannot expect that the utility function is well-behaved. Users may specify the service request in terms of the amount of service and the service time: if the service time is t, then the system has to provide a service in the amount of at least $C(t,U)$ to gain the utility U. Thus, for a given U, we can view the computation (service) request C as a function of the execution time t. This function may not be continuous, and may be defined only on certain discrete points.

Figure 2 shows several such specifications. Curve I is con-

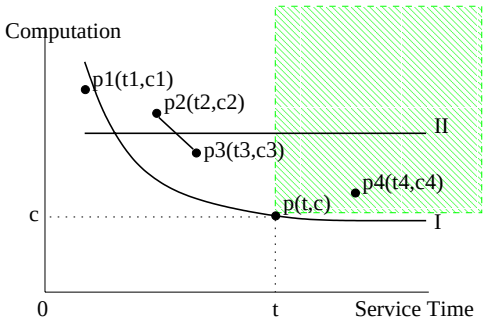

Figure 2: Three specifications of the computation *vs.* service time curve at the same level of utility.

tinuous, and a point $p(t, c)$ on the curve means that the application will be satisfied if a service at the amount of c with service time at least t, or a service no less than c with service time t is provided. That is, any point to the upper right corner of point $p(t, c)$ (the shaded region) is acceptable for this application. Curve II is a horizontal line which represents the utility function of an application that requires a certain amount of computation, but is independent of the service time. Another application expresses its request by four points p_1, p_2, p_3, p_4 and a line segment between points p_2 and p_3.

In the last case, the function is not continuous, thus it is not clear how the application's request can be satisfied with a service time for example between t_3 and t_4. However, any service to the upper right of the curve provides the QoS guarantee assuming that the utility is non-decreasing with respect to service time and the amount of computation. For the simplicity of our discussion, we extend this function by connecting points p_1 and p_2 (as well as p_3 and p_4) in the following way: starting from point p_1 (and p_3), draw a horizontal line up to the second coordinate of point p_2 (and p_4), then draw a vertical line to point p_2 (and p_4). Clearly, any point on the extended curve satisfies the application's QoS request. This is shown in Figure 3.

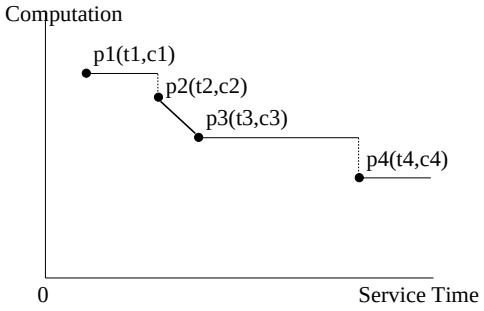

Figure 3: extension of a discontinuous computation *vs.* service time curve with guaranteed utility.

We now rephrase the EM problem in terms of the extended computation *vs.* service time curves:
Given n applications $\tau_1, \tau_2, \cdots, \tau_n$ each with its computation vs. service time specification $C_i(t, U_i)$ for the utility U_i, find for each i, the service time and supply voltage (t_i, v_i) such that U_i is achieved and $\sum_{i=1}^{n} E_i$ is minimized, where E_i is the energy consumption for serving τ_i.
Figure 4 shows the dynamic programming procedure to solve such problem with n general applications.

Input: n applications with their QoS requirement U_k, and computation *vs.* service time specification $C_k(t, U_k)$.
Output: The minimal energy consumption to satisfy all U_k.

Procedure DP:
1) divide time $[0, 1]$ into N equal quants;
2) For $k = 1, 2, \cdots, n$
3) For $i = 1, 2, \cdots, N$
4) locate $C_k(\frac{i}{k}, U_k)$;
5) solve voltage $v_k(i)$ from equation (2) with speed $S = \frac{C_k(i/N, U_k)}{i/N}$;
6) compute power $P_k(i)$ at voltage $v_k(i)$ from (3);
7) calculate energy consumption $E_k(i) = P_k(i) \cdot \frac{i}{N}$;
8) For $i = 1, 2, \cdots, N$
9) $E^1(i) = E_1(i)$;
10) For $k = 1, 2, \cdots, n$
11) For $i = 1, 2, \cdots, N$
12) $E^k(i) = \min\{E^{k-1}(j) + E_k(i-j) : j = 1, \cdots, i-1\}$;
13) return $E^n(N)$;

Figure 4: The dynamic programming approach to solve the EM problem with n applications.

We first discretize the continuous optimization EM problem by quantitizing the total service time $[0, 1]$ into N small quants of the same size. The system will allocate service time to each application in units of quants. Steps 2) $\sim$ 7) calculate $E_k(i)$, the energy to service application τ_k with i quants of time. The amount of computation $C_k(\frac{i}{k}, U_k)$ is determined by the extended computation *vs.* service time curve; then the system's speed S, the required supply voltage $v_k(i)$ to achieve S, the power and energy consumption at $v_k(i)$ are calculated in straightforward ways.

Let $E^k(i)$ $(k = 1, 2, \cdots, n;\ i = 1, 2, \cdots, N)$ be the total energy consumption to service the first k applications with i quants of time. Steps 8) and 9) initialize $E^1(i) = E_1(i)(i = 1, 2, \cdots, N)$, and $E^k(i)$ are computed from the recurrence formula

$$E^k(i) = \min\{E^{k-1}(j) + E_k(i-j) : j = 1, 2, \cdots, i-1\} \quad (7)$$

$E^n(N)$ is the minimized total energy consumption as required.

The recurrence formula (7) states that the minimal energy to finish the first k applications with i quants is to choose the best combination of completing the first $k-1$ applications in $j \leq i-1$ quants and reserving rest of the time for the k-th application. It takes $O(N)$ time to determine $E^k(i)$ as in Step 12). With little bookkeeping, the DP procedure can also find the voltage profile and service time for each application to actually achieve $E^n(N)$. Finally, more accurate solution can be acquired at the cost of increasing N and hence the complexity.

We summarize the global flow for solving the general EM problem:

	at $v_{nominal}$		individual optimal strategy			DP's strategy		
	time	energy	voltage	time	energy	voltage	time	energy
τ_1	0.20	0.20	1.95	0.90	0.1111	2.48	0.40	0.1353
τ_2	0.10	0.10	1.70	0.40	0.0264	2.05	0.25	0.0384
τ_3	0.15	0.15	1.22	1.00	0.0102	1.70	0.30	0.0198
τ_4	0.05	0.05	2.43	0.15	0.0470	3.30	0.05	0.0500
total	0.50	**0.50**	N/A	2.45	**0.1947**	N/A	1.00	**0.2435**

Table 1: Energy consumption for the strategy given by the dynamic programming approach.

Theorem 5.1
For the EM problem with multiple applications and general utility functions, we can first extend their utility function as in Figure 3, then solve it by the DP procedure. The runtime complexity is $O(nN^2)$ and the space complexity is $O(nN)$.

6. SIMULATION RESULTS

We show how to solve the EM Problem for a 4-application system in detail and then report the results on other simulations.

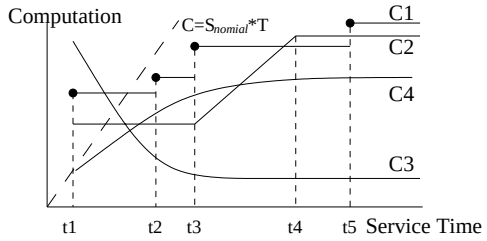

Figure 5: Four computation *vs.* service time curves.

Suppose we have a system with a variable voltage core to serve four applications, the computation *vs.* service time specification is given in Figure 5. Curve $C1$ for application τ_1 is a step function with the amount of computation c_1, c_2, c_3 and c_4 at service times t_1, t_2, t_3 and t_5 respectively. $C2$ is similar and defined as:

$$C2(t) = \begin{cases} b_1, & \text{if } t_1 \le t < t_3; \\ a \cdot t + b, & \text{if } t_3 \le t < t_4; \\ b_2, & \text{if } t \ge t_4. \end{cases}$$

the values of the coefficients $a = \frac{b_2 - b_1}{t_4 - t_3}$ and $b = \frac{b_1 t_4 - b_2 t_3}{t_4 - t_3}$ are given from the continuity. $C3$ is given as, for a fixed integer $k > 1$:

$$C3(t) = \begin{cases} \frac{U}{t}, & \text{if } t_1 \le t \le k \cdot t_1; \\ \frac{U}{k \cdot t_1}, & \text{if } t \ge k \cdot t_1. \end{cases}$$

$C4(t) = \sqrt{U \cdot t}$ is differentiable and clearly indicates a favor in the high speed. Also included in Figure 5 is a dashed straight line from the origin that represents the system's computation ability *vs.* time at the nominal voltage.

In our experimentation, we compare the energy consumption for three different strategies: (I) "system shut-down", in which we find the energy consumption to fulfil each application's QoS requirement at $v_{nominal}$. Under the assumption of *underloaded* system, the system will finish execution before the unit time 1 and can shut down to save energy; (II) in "individual optimal" strategy, we calculate the optimal strategy for the system to solely serve each individual application. In general, the combination of such individual optimal solutions is invalid since the total execution time may exceed 1. However, this gives us a lower bound; (III) "DP's strategy" is the one after we partition the unit time and apply the DP procedure in Section 5. This is a feasible solution.

Table 1 reports one set of the results, where the parameters are set as: $t_1 = 0.05, t_2 = 0.25, t_3 = 0.40, t_4 = 0.80, t_5 = 0.95$; for τ_1 : $c_2 = 1.2c_1, c_3 = 1.6c_1, c_4 = 1.8c_1$; for τ_2 : $b_2 = 2.0b_1$; for τ_3 : $k = 6; v_{nominal} = 3.3V, v_t = 0.8V$; The execution time is divided into pieces of length 0.05.

The service time and the amount of computation for each application at nominal voltage can be obtained from Figure 5 by projecting the intersection point of the application's computation *vs.* service time curve and the system's computation ability curve (the dashed lines in Figure 5) to the axes of service time and computation. The second and third columns of Table 1 show the required service time and energy to meet each application's QoS request, where the energy consumption is normalized to the amount consumed at $v_{nominal}$ in unit time. The last row shows that in this case, we can operate the system at nominal voltage for 0.5 unit time and then shut down, as a result, consume 0.5 unit energy.

The next three columns report the optimal strategy to finish each application in one unit execution time. This is not applicable since the total service time exceeds 1, although more than 60% of the energy is saved over the system shut down technique at nominal voltage. It is worth to mention that the total energy consumption here provides a lower bound and the system can simply adopt this strategy whenever it is feasible.

The rest of the Table 1 is the strategy from the DP procedure. We can see that different supply voltages, from $1.7V$ to the nominal $3.3V$ have been applied to different applications, and the entire one unit execution time is utilized to achieve a total energy consumption of 0.2435, which saves more than 50% over the system shut down.

Finally, we mention that $C4$ is differentiable and corresponds to the case of Example 4.2 when $r = 2$. The optimal voltage for this application is $2.4V$ with service time 0.1576 and consumes 0.04696 unit of energy. The dynamic programming chooses a supply voltage $2.43V$ and service time 0.15, the overhead on energy consumption, which comes from the partition, is negligible (less than 0.00004 unit).

Table 2 reports other simulation results. For each simulation, the number of applications is shown in the second

	n	at $v_{nominal}$		individual optimal		DP's strategy			Compare DP's energy with	
		time	energy	time	energy	N	time	energy	shut-down	lower bound
$test_1$	4	0.50	0.50	2.45	0.1947	20	1.00	0.2435	51.30%	25.06%
$test_2$	4	0.75	0.75	3.20	0.2880	20	1.00	0.4113	45.16%	42.81%
$test_3$	4	0.90	0.90	2.85	0.7112	20	0.95	0.8034	10.73%	12.96%
$test_4$	8	0.50	0.50	6.36	0.2338	50	1.00	0.2827	43.46%	20.92%
$test_5$	8	0.75	0.75	7.20	0.3175	50	0.98	0.4508	39.89%	41.98%
$test_6$	8	0.90	0.90	7.20	0.5740	50	1.00	0.6842	23.98%	19.20%
$test_7$	10	0.50	0.50	7.92	0.1828	100	1.00	0.2253	54.94%	23.25%
$test_8$	10	0.75	0.75	9.16	0.3076	100	1.00	0.4051	45.99%	31.70%
$test_9$	10	0.90	0.90	8.50	0.4991	100	1.00	0.6917	23.14%	38.59%
$test_{10}$	15	0.50	0.50	12.07	0.1732	200	1.00	0.2169	56.62%	25.23%
$test_{11}$	15	0.75	0.75	11.60	0.3480	200	0.99	0.4072	45.71%	17.01%
$test_{12}$	15	0.90	0.90	12.74	0.5255	200	1.00	0.6834	24.07%	30.05%
average	N/A								38.75%	27.40%

Table 2: Comparison of the "system shut-down", "individual optimal", and "DP's strategy" for energy consumption.

column. The next four columns present the total time and energy required by the "system shut-down" strategy and the sum of the individual optimal. The DP procedure divides the unit time into N quants and its solution is reported as well. The last two columns show the energy saving over the "system shut-down" technique and how close to the lower bound provided by individual optimal.

7. CONCLUSIONS

Quality of service is intrinsically connected to many major and most popular applications such as multimedia and wireless sensing. A considerable amount of effort has been put on measuring and charging for the QoS as well as providing guaranteed QoS. At the same time, minimizing power/energy consumption is another important issue for modern system design, especially for the battery-operated systems that support the QoS-sensitive applications. We propose the problems of energy minimization with guaranteed QoS. This is the first attempt of considering these two issues simultaneously during the system design process. Specifically, we apply the variable voltage design methodology to select a voltage profile optimally to provide QoS guarantees for each application and meanwhile minimize energy consumption.

Our key contributions are as follows: (i) formulation of the energy minimization with QoS guarantees problem; (ii) optimal solution when the utility functions are differentiable; (iii) development of the dynamic programming (DP) procedure for solving the general EM problem. (iv) an average of 38.7% energy saving over the "system shut-down" technique, and 27.4% more than an impractical lower bound on a large set of simulations.

8. REFERENCES

[1] J. Altmann and P. Varaiya. *INDEX project: user support for buying QoS with regard to user's preferences.* 1998 Sixth International Workshop on Quality of Service, pp. 101-104, 1998.

[2] A. Chandrakasan, V. Gutnik, T. Xanthopoulos. *Data driven signal processing: an approach for energy efficient computing.* International Symposium on Low Power Electronics and Design, pp. 374-352, 1996.

[3] V. Gutnik, and A. Chandrakasan. *An efficient controller for variable supply-voltage low power processing.* Symposium on VLSI Circuits, pp. 158-159, 1996.

[4] I. Hong, D. Kirovski, G. Qu, M. Potkonjak, and M.B. Srivastava. *Power Optimization of Variable Voltage Core-Based Systems.* Proceedings 1998 Design and Automation Conference, pp. 176-181, 1998.

[5] I. Hong, G. Qu, M. Potkonjak, and M.B. Srivastava. *Synthesis Techniques for Low-Power Hard Real-Time Systems on Variable Voltage Processor.* The 19th IEEE Real-Time Systems Symposium, pp. 178-187, 1998.

[6] K.T. Kornegay, G. Qu, and M. Potkonjak. *Quality of Service and System Design.* (Position paper) IEEE Computer Society Annual Workshop on VLSI, Theme: System Level Design, pp. 112-117, 1999.

[7] T.F. Lawrence. *The quality of service model and high assurance.* Proceedings. 1997 High-Assurance Engineering Workshop, pp. 38-39, 1997.

[8] W. Namgoong, M. Yu, T. Meng. *A high-efficiency variable-voltage CMOS dynamic dc-dc switching regulator.* 1997 IEEE International Solid-State Circuits Conference Digest of Technical Papers, pp. 380-381,489, 1997.

[9] J.M. Rabaey and M. Pedram (Ed.). *Low Power Design Methodologies (Chapter 1).* Kluwer Academic Publishers, 1996.

[10] R. Rajkumar, C. Lee, J. Lehoczky, and D. Siewiorek. *A resource allocation model for QoS management.* Proceedings. The 18th IEEE Real-Time Systems Symposium, pp. 298-307, 1997.

[11] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. *RSVP: A New Resource ReSerVation Protocol.* IEEE Network Magazine, Vol.7, No.5, pp.8-18, September 1993.