Low-Power Buffer Management Using Hybrid Control

Jason Ridenour, Jianghai Hu, and Yung-Hsiang Lu

Abstract—The power management problem is studied for a pipeline of streaming data consisting of multiple stages with buffer memories inserted between adjacent stages. The objective is to find the optimal strategy for dynamically changing the power states of the stages so that the power consumption of the overall system is minimized. By modeling the pipeline as a hybrid system, we derive various necessary conditions for the optimal solutions. In particular, an operation called folding is introduced that can improve the performance of an existing strategy. Analytic conditions are derived for determining whether folding can save power and the optimal number of folds.

I. INTRODUCTION

Consider a pipeline of streaming data shown in Fig. 1, which consists of three components, X, Y, and Z, and two buffers in between. Component X generates data at the rate of $\alpha$ and stores them in the first buffer. Component Y retrieves data from the first buffer, processes them, and stores them in the second buffer, all at the rate of $\beta$. Component Z retrieves data from the second buffer at the rate of $\gamma$. This system can be the abstract model of many electronics and computer network applications. For example, in a PDA (Personal Digital Assist), X can be the network card downloading streaming video from the Internet; Y can be the processor to decode the data; and Z can be the display unit.

We assume that $\alpha$, $\beta$, and $\gamma$ are constants satisfying $\alpha > \gamma$ and $\beta > \gamma$. Thus there is an over-supply of data from upstream for Z to consume if all components are active. As a result, it is possible to reduce the system power consumption by turning off components X and Y at proper times, as they consume less power in the off state than in the normal operating state. This process is called dynamic power management [4]. In this process, the smooth operation of Z is ensured by the presence of the two buffers. For example, Z can retrieve and consume data already stored in the second buffer even if Y is turned off.

In this paper, we study the optimal power management problem for the pipeline of streaming data, which tries to find the strategy for coordinating the turning on and off of components X and Y so that maximal power reduction is achieved for the overall system. A solution will be presented by modeling the pipeline as a hybrid system.

Proposed to model dynamical systems with both continuous and discrete dynamics, hybrid systems have applications in many different engineering fields, such as Air Traffic Management (ATM) systems [13], automated highway systems [14], computer and communication networks [9], and auto and industrial systems [2]. We will use a version of the hybrid systems model called multirate hybrid systems. Our problem thus becomes an instance of the optimal control problems for hybrid systems [1], [3], [7], [8], [12], [15]. The difference of our problem with other optimal control problems for hybrid systems studied in the literature consists of the following two aspects: (i) our problem studies the optimal periodic solutions; (ii) in addition to the system trajectories, the shape and size of the state space of the hybrid system also need to be design at the same time. These make our problem a challenging one.

In our recent work [10], we present some preliminary results on applying the hybrid system model in the study of the dynamic power management problem for the pipeline of streaming data. Some necessary conditions for the optimal solutions are derived, and optimal periodic solutions are derived under the assumption that X and Y each turns on and off at most once in a period. In this paper, we will extend our results in [10] by introducing an operation on an existing solution called folding that can generate another feasible solution with multiple switchings in a period but potentially lower average power than the original one. Conditions will be derived on whether folding will result in a better

Fig. 1. Two buffers between three interacting components.
solution, and, if this is the case, the number of optimal folding operations one should carry out.

The rest of the paper is organized as follows. In Section II, the problem under study is formulated precisely. In Section III, we review some previous results on the optimal solutions in a constrained case. In Section IV, we introduce the folding operation and derive the main results on the optimal number of folds and scaling of a solution. Finally, in Section V, we outline the conclusions and some possible future directions.

II. PROBLEM FORMULATION

We first analyze the power expenditure of the pipeline of streaming data shown in Fig. 1. The power consumed by each component consists of two parts: the dynamic power and the static power. The dynamic power includes the power to produce, process, and/or consume data, as well as the power for writing data into and retrieving data from the buffer memory. As the name suggests, the dynamic power depends on the data flow rate through the component. In contrast, the static power of the component is the part of the component’s power used to sustain its normal operation, and is independent of the data rate. In this paper, we consider only the static powers, as it is shown in [5], [6], [11] that the dynamic powers do not affect the solution of the optimal buffer management problem defined later.

More precisely, each of the components X and Y has two power states: on and off. When X and Y are on, they consume the (static) powers \( p_x \) and \( p_y \), respectively. When they are off, they consume no power. Changing the power states of these two components may incur extra energy. For simplicity, we assume that the energy overheads for turning on X and Y are \( k_x \) and \( k_y \), respectively, while their turn-off energies are both zero.

We also assume that any delay in changing the power states of X and Y is amortized by an early wakeup policy since the system is deterministic. Because the last component Z is assumed to be always turned on, it consumes a constant power. Thus its power contribution can be ignored in finding the optimal solutions.

In addition to component powers, the two buffers also consume powers that are proportional to their sizes. Denote by \( Q_1 \) and \( Q_2 \) the sizes of the two buffers. Then their powers are \( m_1 Q_1 \) and \( m_2 Q_2 \), respectively, where the constants \( m_1 \) and \( m_2 \) are the powers per unit size of the buffer memory (W/MB). If the two buffers use the same technology, we have \( m_1 = m_2 \).

Since \( \alpha > \gamma \) and \( \beta > \gamma \), depending on the amounts of data stored in the two buffers, the components X and Y can be turned on/off from time to time to save power without interrupting the smooth operation of Z. Thus the pipeline system operates in one of four possible modes: \( S = \{00, 01, 10, 11\} \). These modes are:

- mode 00: X and Y are both off. The total component power is \( p_{00} = 0 \);
- mode 01: X is off and Y is on. The total component power is \( p_{01} = p_y \);
- mode 10: X is on and Y is off. The total component power is \( p_{10} = p_x \);
- mode 11: X and Y are both on. The total component power is \( p_{11} = p_x + p_y \).

Denote by \( \sigma \) the system mode.

Let \( q_1 \) be the amount of data stored in the buffer between X and Y, and \( q_2 \) be the amount of data stored in the buffer between Y and Z. Depending on the system mode, \( q_1 \) and \( q_2 \) change at different constant rates. For example, when both X and Y are on, i.e., when the system is in mode 11, \( q_1 \) increases at the rate \( \alpha - \beta \), and \( q_2 \) increases at the rate \( \beta - \gamma \). Combining \( q_1 \) and \( q_2 \) into a vector \( q = (q_1, q_2) \), the rates of change of \( q \) in mode \( \sigma \in S \), denoted by \( v_\sigma \), are given by:

\[
\begin{align*}
v_{00} &= (0, -\gamma), \quad v_{01} = (-\beta, \beta - \gamma), \\
v_{11} &= (\alpha - \beta, \beta - \gamma), \quad v_{10} = (\alpha, -\gamma).
\end{align*}
\]

The four vectors in equation (1) are depicted in Fig. 2, with the subfigures (a) and (b) corresponding to the cases when \( \alpha > \beta \) and \( \alpha < \beta \), respectively.

The mode \( \sigma \) and the variable \( q \) together specify the system state of the pipeline of streaming data. Note that \( q = (q_1, q_2) \) is a continuous variable taking values in the rectangle \([0, Q_1] \times [0, Q_2]\), while \( \sigma \) is a discrete variable taking values in \( S \). Thus the pipeline system can be modeled naturally as a hybrid system with state \( z = (q, \sigma) \). Let \([0, t_f]\) be the time interval during which the average system power consumption needs to be minimized. Then a power management scheme for the pipeline system is given by a mapping \( \sigma : [0, t_f] \to S \) so that the power states of the components X and Y at time \( t \in [0, t_f] \) are specified by the mode \( \sigma(t) \). Typically, there exists a
sequence of time intervals \([t_i, t_{i+1}], i = 0, \ldots, n - 1\), with \(t_0 = 0 \leq t_1 \leq \cdots \leq t_n = t_f\), such that the system mode \(\sigma(t)\) during the time interval \([t_i, t_{i+1}]\) is a constant \(\sigma_i\), and that immediately adjacent time intervals correspond to different modes. Under this scheme \(\sigma(t)\), the continuous state \(q(t)\) follows a continuous trajectory given by the differential equation:

\[
\frac{dq(t)}{dt} = v_{\sigma(t)}, \quad t \in [0, t_f].
\]

Together, \(z(t) = (q(t), \sigma(t))\) forms a (hybrid) solution, or an execution, of the hybrid system.

For a solution \(z(t) = (q(t), \sigma(t))\) during the period \([0, t_f]\) of interest, the energy consumption of the whole pipeline system consists of the following three parts:

- static energy consumed by X and Y: \(\int_0^{t_f} p_{\sigma(t)} dt\);
- energy for changing the power states of the components X and Y: \(n_x k_x + n_y k_y\). Here \(n_x\) and \(n_y\) are the numbers of times that X and Y are turned on, respectively, during \([0, t_f]\);
- buffers’ energy consumption: \((m_1 Q_1 + m_2 Q_2) t_f\).

As a result, for the solution \(z(t)\), the average power consumption, \(\bar{P}\), of the overall system is:

\[
\bar{P}(z; t_f, Q_1, Q_2) = \frac{1}{t_f} \left[ \int_0^{t_f} p_{\sigma(t)} dt + n_x k_x + n_y k_y + (m_1 Q_1 + m_2 Q_2) t_f \right].
\]

Note that the arguments of \(\bar{P}\) are introduced to highlight the dependence of \(\bar{P}\) on \(t_f, Q_1,\) and \(Q_2\).

The dynamic power management problem is then to find \(z(t)\) that can achieve the minimal average power. In particular, if the look-ahead time horizon \(t_f\) is large enough, the problem becomes:

**Problem 1:** Find the hybrid trajectories \(z(t)\) during \([0, \infty)\) and the buffer sizes \(Q_1\) and \(Q_2\) that minimize the average power \(\lim_{t_f \to \infty} \bar{P}(z; t_f, Q_1, Q_2)\).

This is an optimal control problem for the hybrid system modeling the pipeline of streaming data.

### III. Periodic Solutions

We will focus on periodic solution \(z(t)\) in this paper. For periodic solutions \(z(t)\) with period \(T\), the average power equation reduces to:

\[
\bar{P}(z; T, Q_1, Q_2) = \frac{1}{T} \left( \sum_{\sigma \in S} T_\sigma p_{\sigma} + n_x k_x + n_y k_y \right) + m_1 Q_1 + m_2 Q_2,
\]

where \(T_\sigma, \sigma \in S\), is the amount of time the system is in mode \(\sigma\) in a single period.

We now summarize some useful results on optimal periodic solutions derived previously in [10].

#### A. Scaling of Trajectories

Let \(z(t) = (q(t), \sigma(t))\) be a periodic hybrid trajectory with period \(T\) for buffer sizes \(Q_1\) and \(Q_2\). Let \(\lambda > 0\) be a positive number. Then, it can be verified that \(\lambda q(t/\lambda)\) is a continuous trajectory in \([0, \lambda Q_1] \times [0, \lambda Q_2]\) satisfying equation (2) with \(\sigma(t)\) replaced by \(\sigma(t/\lambda)\). In other words, \((\lambda q(t/\lambda), \sigma(t/\lambda))\) is a periodic hybrid trajectory with the period \(\lambda T\) when the sizes of two buffers become \(\lambda Q_1\) and \(\lambda Q_2\), respectively. We call \(z_\lambda(t) \equiv (\lambda q(t/\lambda), \sigma(t/\lambda))\) the scaling of \(z(t) = (q(t), \sigma(t))\) by \(\lambda\). Compared with the original trajectory, the scaled one has the same switching sequence, and a minimal average power if and only if \(\lambda\) takes the value

\[
\lambda^* = \sqrt{\frac{n_x k_x + n_y k_y}{(m_1 Q_1 + m_2 Q_2)T}}.
\]

#### B. Tightness of Optimal \(q(t)\) in \(Q\)

Since a buffer consumes power proportional to its size, unused buffer memory will unnecessarily waste power. As a result, an optimal trajectory \(z\) should fill and empty each of the two buffers at least once sometime during the period, or equivalently, the curve \(q(t)\) must touch all four edges of the rectangle \([0, Q_1] \times [0, Q_2]\).

#### C. Optimal Periodic Solution with \(n_x = n_y = 1\)

Using the above necessary conditions, we derive in [10] the optimal periodic solutions (see Fig. 3) under the constraint that \(n_x = n_y = 1\), i.e., both X and Y turn on and off only once per period. In particular, if \(\alpha > \beta\), the optimal solution plotted in Fig. 3(a) has the
illustrate this operation through an example. Consider consumption of an existing switching strategy. We will in Fig. 4a, can be constructed by following exactly the $\alpha > \beta$.

The folding operation can also be applied multiple times in a sequence. Fig. 5 shows an example of folding an existing trajectory in five subsequent steps. Note that after the folding, the required buffer size $Q^*_1$ is reduced to $Q^*_2$, while the number of switching $k_x$ is increased from 1 to 5. If one folds a given trajectory an integer $n$ times, the buffer size $Q^*_1$ can be reduced to $Q^*_n$, and $k_x$ will increase to $nk_x$.

After folding, the shape of the trajectory is changed, and the new rectangular state space may no longer be of optimal scale as determined by equation (4). Thus one can rescale the state space to get a further improvement in average power consumption. In the following, we will study the power saving by performing the folding and rescaling operations at the same time.

### IV. FOLDING

In this section, we will introduce an operation called folding that may further decrease the system power consumption of an existing switching strategy. We will illustrate this operation through an example. Consider the optimal switching strategy $z^*(t)$ for the case $\alpha < \beta$ and $n_x = n_y = 1$ shown in Fig. 3a. Assume that at time 0, $z^*(t)$ starts from the point $(a_x^0, b_y^0) = (0, 0)$, i.e., with both buffers empty. Then a new strategy $\hat{z}(t)$, shown in Fig. 4a, can be constructed by following exactly the same route of $z^*(t)$ at all times, except that between the time $T^*_{11}$ and $T^*_{10} + T^*_{01}$, the switching order from mode 11 to mode 01 is reversed. In other words, at time $T_{11}$, $\hat{z}(t)$ will switch to mode 01 and follow the direction of $v_{01}$ for $T_{11}$ time; at time $T_{11} + T_{01}$, it switches back to mode 11 and follow the direction of $v_{11}$ for $T_{10}$ time; and then at time $T_{11} + T_{01} + T_{01}$, $\hat{z}(t)$ switches back to mode 01, and follows $z^*(t)$ thereafter. Intuitively, $\hat{z}(t)$ can be thought of as obtained from $z^*(t)$ by (twisting and) folding a corner of the triangular trajectory of $z^*(t)$ inward. Thus we call this operation folding. Similarly, the optimal periodic solution for the case $\alpha < \beta$ and $n_x = n_y = 1$ shown in Fig. 3(b) can also be folded, with the resulting strategy shown in Fig. 4b.

Comparing the strategy obtained by folding (e.g., the one in Fig. 4a) with the original one (e.g., the one in Fig. 3a), reveals they have the same period, and consume the same static energy during each period as the proportions of time the system spends in each mode remain unchanged in a period. Although the number $k_x$ of switchings of component X is increased, which results in increased overhead energy, the required buffer size $Q_1$ is only half of the original $Q^*_1$. If the increased power from the extra switchings is amortized by the power savings resulting from reduction in buffer size, folding can reduce the power consumption of the system.

The optimal solution in Fig. 3(b) is:

$$Q^*_1 = \frac{(\beta - \alpha) \sqrt{\gamma(k_x + k_y)}}{\beta(\beta - \alpha)m_1 + \alpha\beta(\beta - \gamma)m_2},$$

$$Q^*_2 = \frac{(\beta - \alpha) \sqrt{\gamma(k_x + k_y)}}{\beta(\beta - \alpha)m_1 + \alpha\beta(\beta - \gamma)m_2}.$$

While if $\alpha < \beta$, the optimal solution in Fig. 3(b) is:

$$Q^*_1 = \frac{(\beta - \alpha) \sqrt{\gamma(k_x + k_y)}}{\beta(\beta - \alpha)m_1 + \alpha\beta(\beta - \gamma)m_2}.$$

$$Q^*_2 = \frac{(\beta - \alpha) \sqrt{\gamma(k_x + k_y)}}{\beta(\beta - \alpha)m_1 + \alpha\beta(\beta - \gamma)m_2}.$$

### A. Optimal Integer Folding when $\alpha > \beta$

Consider the trajectory in Fig. 3a with $\alpha > \beta$. Integer folding $n$ times affects only the buffer size $Q_1$, as is demonstrated in Fig. 4a for $n = 2$. The average power of the folded and rescaled trajectory is described by

$$P(\lambda, n) = \lambda \left(\frac{m_1 Q^*_1}{n_x} + m_2 Q^*_2\right) + \frac{nk_x + k_y}{\lambda T} + \frac{1}{T} \sum_{s \in B} T_s p_s,$$

where $Q^*_1$ and $Q^*_2$ are the optimal buffer sizes for the $n_x = n_y = 1$ periodic solution, and $n$ is the number of folds with $n = 1$ representing no folds, and $\lambda$ is the scaling factor. We begin by finding the optimal scaling factor for a fixed $n$:

$$\frac{\partial P(\lambda, n)}{\partial \lambda} = \frac{m_1 Q^*_1}{n} + m_2 Q^*_2 - \frac{nk_x + k_y}{\lambda^2 T} = 0,$$
Thus, the optimal value of \( \tilde{n} \) is given by
\[
\tilde{n} = \arg\min_{n \in \{n^*, \lceil n^* \rceil \}} P(\lambda_n^*, n),
\]
where \( n^* \) is defined in (8), and the optimal rescaling factor is given by (6) with \( n \) replaced by \( \tilde{n} \).

As a result, if the value of \( \tilde{n} \) is greater than one, integer folding can indeed save more power than the optimal solution under the constraint \( n_x = n_y = 1 \).

### B. Optimal Integer folding when \( \alpha < \beta \)

In this case, folding affects the buffer size \( Q_2 \) as well as \( Q_1 \), as can be seen in Fig. 4b. The average power as a function of the number \( n \) of folds and the rescaling factor \( \lambda \) is more complicated, and can be verified to be
\[
P(\lambda, n) = \lambda \left( \frac{m_1 Q_1^*}{n} + m_2 (Q_2^* - \frac{\gamma(n - 1)Q_1^*}{\alpha n}) \right) + \frac{nk_x + k_y}{T} + \frac{1}{T} \sum_{s \in S} T_s p_s,
\]
where again \( Q_1^* \) and \( Q_2^* \) are the optimal buffer sizes for the \( n_x = n_y = 1 \) case shown in Fig. 3b. Following the same process as before, we obtain
\[
\lambda_n^* = \left[ \frac{nk_x + k_y}{m_1 Q_1^* + m_2 Q_2^*} \right]^{\frac{1}{T}},
\]
(10)

\[
n^* = \left( \frac{m_1 + \frac{\alpha}{\alpha - 1} m_2}{m_2 Q_2^* - 2 m_2 Q_1^*} \right) \frac{k_y}{k_x}.
\]
(11)

Again, \( P(\lambda_n^*, n) \) is convex in \( n \); so the optimal integer value \( \tilde{n}^* \) lies one of the two closest integers to \( n^* \).

### Proposition 2 (Optimal Folding for the \( \alpha < \beta \) Case):

Suppose that \( \alpha < \beta \). Then for the optimal periodic solution under the constraint \( n_x = n_y = 1 \) shown in Fig. 3b, the optimal number of folds is given by
\[
\tilde{n} = \arg\min_{n \in \{n^*, \lceil n^* \rceil \}} P(\lambda_n^*, n),
\]
where \( n^* \) is defined in (11), and the optimal rescaling factor is given by (10) with \( n \) replaced by \( \tilde{n} \).

### C. Discussion and Possible Extensions

Examining the folding equations reveals several aspects of the nature of folding. For example, equations (8)
and (11) show that lower switching power in element X and higher memory cost in buffer $Q_2$ allows for more folds, which is intuitively easy to understand.

Equations (6) and (10) are monotonically increasing with respect to $n$ for $n > 1$. This implies that for any number of folds, the scaling factor is always greater than one. In other words, for any number of folds, buffer size $Q_2$ will always need to be increased compared with the original one with no folds. Thus the new state space after the folding and rescaling is thinner and taller than the state space of the original unfolded trajectory.

In this paper we only consider folding across a vertical line. However, it is also possible to fold horizontally as shown in Fig. 6. In this case, the number of switchings for component Y is 2, while the number of switchings for component X remains one in a single period. Furthermore, it is worthwhile to consider simultaneous foldings in both horizontal and vertical directions.

Another natural extension is to consider systems consisting of more than three components and two buffers interconnected by a network (graph). In this case, the number of system modes will grow exponentially with the number of buffers. Thus techniques such as matrix optimization must be utilized to find approximate solutions to such problems. Research is currently being carried out along this direction.

V. CONCLUSION

This paper studies the dynamic power management problem of a pipeline of data streaming through three components and two buffers. We extend our previous work to consider the more general situation where solutions are allowed to switch on and off multiple times per period by introducing the operation of folding. We derive the condition under which folding can lower the system power consumption. The optimal number of folds and the optimal rescaling factor are derived analytically.

VI. ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments.

REFERENCES


Fig. 6. Examples of folding across a horizontal line. (a) The original solution for $n_x = n_y = 1$. (b) The solution folded across a horizontal line, the value of $n_x$ is still 1, but now $n_y = 2$