Dynamic Power Management in a Mobile Multimedia System with Guaranteed Quality-of-Service

Qinru Qiu, Qing Wu, and Massoud Pedram
Dept. of Electrical Engineering-Systems
University of Southern California
Los Angeles CA 90089

Outline

- Introduction
  - Overview of dynamic power management
  - Definition of Quality of Service (QoS)
- System modeling
  - GSPN background
  - Non-exponential distribution
  - System modeling
- Optimization technique
  - Buffer estimation
  - Policy optimization
- Experimental results
- Conclusions
Introduction

- System-level dynamic power management
  - Power control of all system components and resources
  - Dynamic change of the system power state while meeting a global performance constraint
- Limitations of the previous works
  - The only performance constraint that has been considered is the average waiting time of a request
  - Service requests have been collected from traditional applications such as file access, keyboard access, etc.

Quality of Service (QoS)

- QoS: The set of quantitative and qualitative characteristics of a distributed multimedia system that capture the notion of "user satisfaction" with the multimedia data presented to him/her
QoS Parameters

- QoS is often expressed by a set of target parameters
- The three most important QoS parameters are:
  - Delay ($D$): Time interval between the moment a data unit is received (input) and the moment it is sent (output)
  - Jitter ($J$): Variation in delay values for data units in a given input stream
  - Loss rate ($L$): Fraction of data units that is lost during the data transport

Global QoS Management

- User requests an end-to-end QoS level
- Global QoS manager allocates QoS for each system component
- Local QoS manager controls the allocation and state of the local resources
- The client system needs power and QoS management (PQM)
PQ Manager

- The PQ manager performs both power and QoS management
  - Determine the PQ management policy that results in the minimum power dissipation while meeting user specified QoS constraints ($D$, $J$, $L$)

PQM Policy Optimization Workflow

1. System Modeling in Controllable GSPN
2. Transformation from Controllable GSPN to Controllable CTMDP
3. Buffer Size Estimation
4. LP-Based Policy Optimization
### Background: GSPN Primitives

- **Place**: condition or situation
- **Token**
  - Marking $m(p)$: #of tokens in $p$
  - System state $m$: $m(p_{on})=0, m(p_{stby})=1, m(p_{off})=0, m(p_{queue})=0$
- **Transition**: event
  - Timed and immediate
- **Input arc**: $I(t, p)$
  - $t \in p^*, p \in t^*$
- **Output arc**: $O(t, p)$
  - $t \in t^*, p \in t^*$
- **Condition Gate**: G
- **Case**: uncertainty

### GSPN Enabling and Firing Rules

- **Transition $t$** is enabled in marking $m$ exactly if
  - $\forall p \in t^*, m(p) \geq I(t, p)$ and condition for any gate G that is on an input arc is true
- **Firing of $t$**
  - Removes $I(t, p)$ tokens from $t^*$
  - Deposits $O(t, p)$ tokens into $t^*$

**Example**

<table>
<thead>
<tr>
<th>Transition</th>
<th>$P_{on}$</th>
<th>$P_{stby}$</th>
<th>$P_{off}$</th>
<th>$P_{queue}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{switch_{on}}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$t_{process}$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$t_{switch_{off}}$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$G$: $m(P_{queue}) = 0$
GSPN Enabling and Firing Rules (cont.)

- A timer is associated with a timed transition \( t \)
  - When \( t \) is enabled, a timer is set to a random value according to the probability distribution function associated with \( t \) and starts counting down
  - When the timer reaches 0, \( t \) fires and resets the timer
- An immediate transition always has a higher priority than a timed transition
- Marking types:
  - Tangible marking: no immediate transition is enabled
  - Vanishing marking: at least one immediate transition is enabled

\[ F(t) \]

Controllable GSPN

- A controllable GSPN is a GSPN where the case probability of free-choice conflict immediate transitions can be controlled by outside commands
  - Can be transformed to a controllable CTMDP
  - Need to find the set of commands (and hence, case probabilities) that minimize some cost function
The input inter-arrival times of a multimedia (MM) request generally follow a non-exponential distribution.

The priority of the MM request is higher than that of the local (normal) application request.

$J$ and $L$ constraints are only applied to the MM applications.

---

Top Level GSPN Model

---
GSPN Model for the MM SR & SQ

Stage 1

Stage 2

Stage 3

\[ \mu_1 \beta_1 P_{MM1} \mu_2 \beta_2 P_{MM2} \mu_3 P_{MMBuf} \]

\[ G_{MM}: m(P_{MM1}) + m(P_{MM2}) = 0 \text{ & } m(P_{MMBuf}) < \text{MM buffer size} \]

- Use a stage method to approximate the non-exponential inter-arrival time of a MM request; in this example, \( r = 3 \)

GSPN Model for the Local SR & SQ

\[ T_{\text{norm}} P_{SQ} \]

\[ G_{\text{norm}}: m(P_{SQ}) < \text{SQ capacity} \]

- Assume that the inter-arrival time of local requests follows an exponential distribution
**GSPN Model for the SP**

- **Status:** idle, busy, transition, decision
- **p_mode:** active, sleeping, etc.
- **a_mode:** MM application, local application

**Cost Definition**

- **Rate cost:** $d_r, j_r, I_r, ld_r, pow_r$
  - $P_{MMBuf}$
    - $d_r = \#tokens$
    - $j_r = (\#tokens - \text{Ave(}\#tokens\text{))}^2$
    - $I_r = 1$ when $\#tokens = \text{MM Buffer size}$
  - $P_{SQ}$
    - $ld_r = \#tokens$
- $P_{status}(\text{power_mode, application_mode})$
  - $pow_r =$ power consumption of SP in its current state

- **Impulse cost:** $ene$
  - $ene_{ij}$: Energy needed for the SP to switch from state $i$ to state $j$
PQM Policy Optimization Workflow

System Modeling in Controllable GSPN

Transformation from Controllable GSPN to Controllable CTMDP

Buffer Size Estimation

LP-Based Policy Optimization

Transformation from GSPN to CTMP

- A state in CTMDP corresponds to a tangible marking in GSPN
- A transition in CTMDP corresponds to a timed activity in GSPN
- The rate cost of a state of CTMDP is the sum of all the rate costs of the places in the corresponding marking in GSPN
- The transition cost is CTMDP the impulse cost of the corresponding activity in GSPN

Search the reachability set of GSPN

Find transition rate between each state and form the generator matrix of CTMDP

Calculate the cost and impulse cost
PQM Policy Optimization Workflow

1. **System Modeling in Controllable GSPN**
2. **Transformation from Controllable GSPN to Controllable CTMP**
3. **Buffer Size Estimation**
4. **LP-Based Policy Optimization**

**Buffer Size Estimation**

- Too large a buffer size is unnecessary
- Too small a buffer size will overconstrain the system

Given some buffer size, the performance metrics $D$, $J$, and $L$ are dependent on each other:
- Given any three, we can estimate the fourth one

We are interested in the minimum buffer size that is needed to avoid overconstraining the system.

<table>
<thead>
<tr>
<th>Buffer Size</th>
<th>$D$</th>
<th>$J$</th>
<th>$L$</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1.23</td>
<td>0.75</td>
<td>0.02</td>
<td>2.08</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
<td>0.9</td>
<td>0.02</td>
<td>1.49</td>
</tr>
</tbody>
</table>

$(D, J, L) = (1.5, 0.9, 0.02)$
Buffer Size Estimation

Min. \( n \)

Subject to:

\[
\begin{align*}
    p_n &\leq L \\
    0 &\leq p_i \leq 1, \ i = 1, \ldots, n
\end{align*}
\]

- A bound on the minimum required buffer size:
  \( N = \text{Max}(n_1, n_2, n_3) \)
  - No overconstraints if \( n \geq N \)

\[
\begin{align*}
    (n - 2) \cdot L &\geq J + D^2 - 4 \cdot D + 3 \\
    (n + 1) \cdot (n - 2) &\cdot L \geq D - 2 \\
    (n - 2) \cdot (n - 1) &\cdot L \geq D^2 - 3 \cdot D + 2 + J
\end{align*}
\]

PQM Policy Optimization Workflow

1. System Modeling in Controllable GSPN
2. Transformation from Controllable GSPN to Controllable CTMP
3. Buffer Size Estimation
4. LP-Based Policy Optimization
PO Problem Formulation

Subject to:

- $p_{ij}^0$: probability that the next system state is $j$
- $\tau_{ii}^0$: expected duration of time that the system will be in state $i$
- $x_{ii}^0$: frequency that the state of the system will be $i$ and action $a_i$ will be taken; Note that: $x_{ii}^0 \cdot \tau_{ii}^0 \equiv p_{ii}^0$
- $\text{pow}_i$: power consumption in state $i$
- $q_{\text{MMBuf}}$: number of unprocessed data in the MM buffer
- $\text{ene}_{ij}$: the energy needed for system to switch from state $i$ to state $j$

Linear Approximation of Jitter

- The exact jitter:
  
  \[ (a) \]

  - Non-linear expression of $x_i$

- Linear approximation of jitter:

  \[ (b) \]

  - Linear expression of $x_i$

- Theorem: For any set of $\text{ene}_{ij}$, if (b) is smaller than $J$ then (a) is smaller than $J$.

  - For each policy, if the approximated jitter satisfies the given constraint then the real jitter also satisfies the given constraint
**Experimental Results**

- **System setup**
  - SP has two power modes: high power and low power
  - In the PD_optimized system, we end up overconstraining delay in order to satisfy the jitter constraints

<table>
<thead>
<tr>
<th>Power Consumption (mW)</th>
<th>MM</th>
<th>local</th>
<th>idle</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Power</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Low Power</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service Speed (ms)</th>
<th>MM</th>
<th>local</th>
<th>idle</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Power</td>
<td>5</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Low Power</td>
<td>10</td>
<td>2.5</td>
<td>-</td>
</tr>
</tbody>
</table>

**Conclusions**

- We introduced a complete modeling technique based on controllable GSPN with cost that captures the behavior of a battery-powered multimedia client system
- We showed how to obtain the PQ-optimal policy based on this stochastic mathematical framework
- This is the first power management policy that considers jitter and loss rate as well as the delay
- Experimental results demonstrated that the PQ-optimized policies are more power-efficient than the PD-optimized policies under the same D, J and L constraints