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

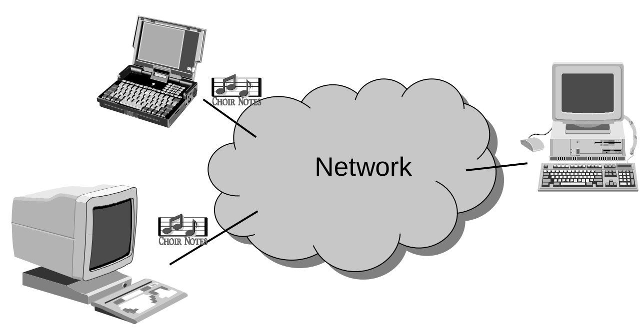
- Introduction
  - Overview of dynamic power management
  - Definition of Quality of Service (QoS)
- System modeling
  - GSPN background
  - Non-exponential distribution
  - System modeling
- Optimization technique
  - Buffer estimation
  - Policy optimization
- Experimental results
- Conclusions

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, key board 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

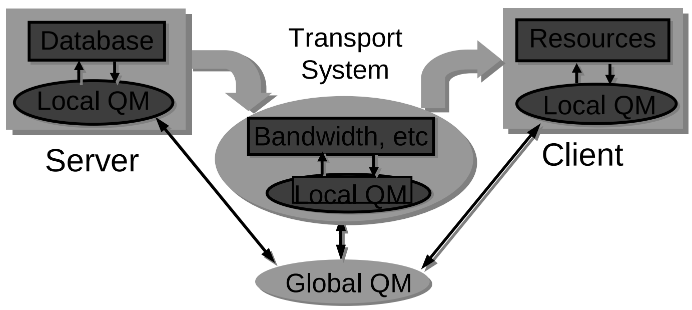
- QoS is often expressed by a set of target parameters
- The three most important QoS parameters are:
  - Delay (*D*): Time interval between the moment a data unit is received (input) and the moment it is sent (output)
  - Jitter (*J*): Variation in delay values for data units in a given input stream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

Global QoS Management

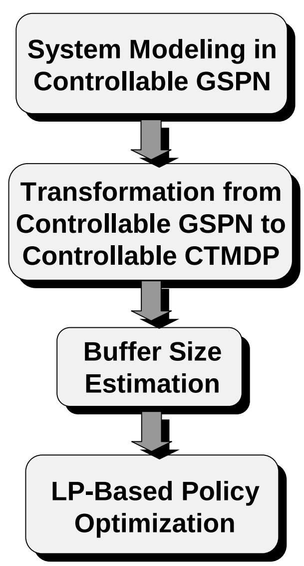
- User requests an end-to-end QoS level
- Global QoS manager allocates QoS for each system component
- Local QoS manager controls the allocation and state of the local resources
- The client system needs power and QoS management (PQM)

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



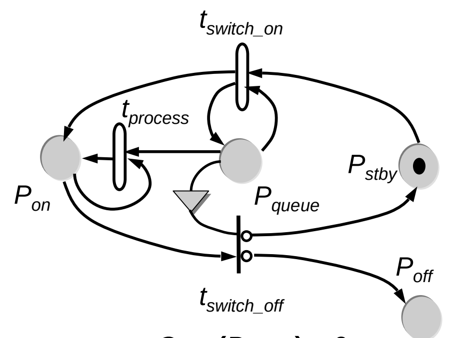
Background: GSPN Primitives

- Place: condition or situation
- Token
  - Marking $\boldsymbol{m}(p)$: #of tokens in p
  - System state $\boldsymbol{m}$

$\boldsymbol{m}(p_{on})=0$, $\boldsymbol{m}(p_{stby})=1$, $\boldsymbol{m}(p_{off})=0$, $\boldsymbol{m}(p_{queue})=0$

$\boldsymbol{m} = [0, 1, 0, 0]$

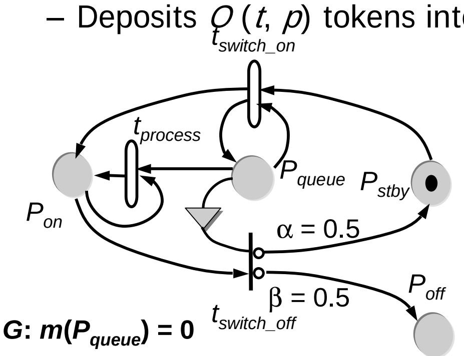
- Transition: event
  - Timed and immediate
- Input arc: $I\,(t,\ p)$
  - $t \in p^\bullet$, $p \in {}^\bullet t$
- Output arc: $O\,(t,\ p)$
  - $t \in {}^\bullet p$, $p \in t^\bullet$
- Condition Gate: G
- Case: uncertainty



G: $\boldsymbol{m}(P_{queue}) = 0$

GSPN Enabling and Firing Rules

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

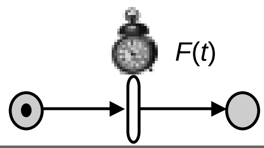


G: $\boldsymbol{m}(P_{queue}) = 0$

p_{on}	p_{stby}	p_{off}	p_{queue}
0	1	0	1
1	0	0	1
1	0	0	0
0	1	0	0 (pr = 0.5)
0	0	1	0 (pr = 0.5)

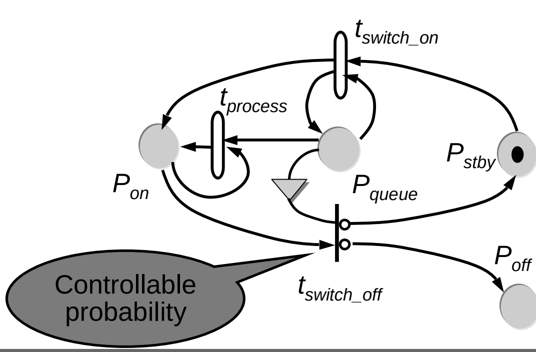
GSPN Enabling and Firing Rules (cont.)

- A timer is associated with a timed transition t
  - When t is enabled, a timer is set to a random value according to the probabilty 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

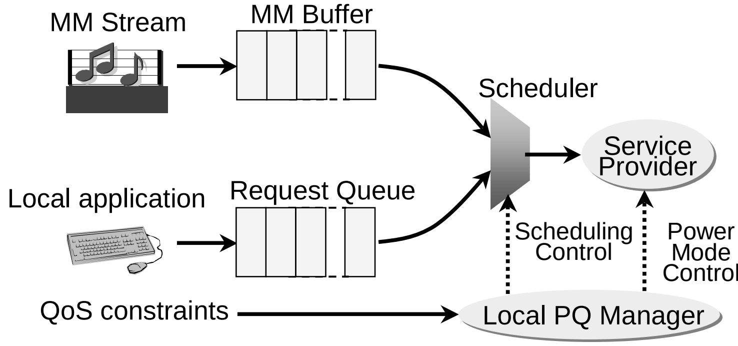


Controllable GSPN

- 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

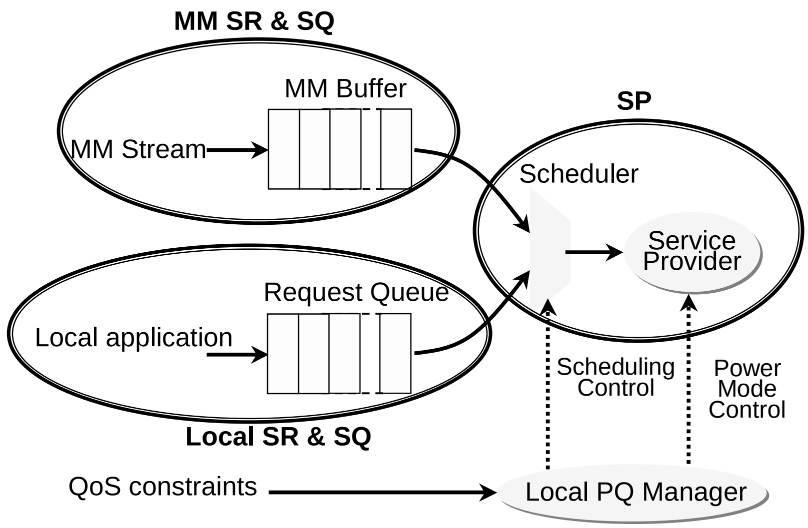


PQ-managed MM Client System

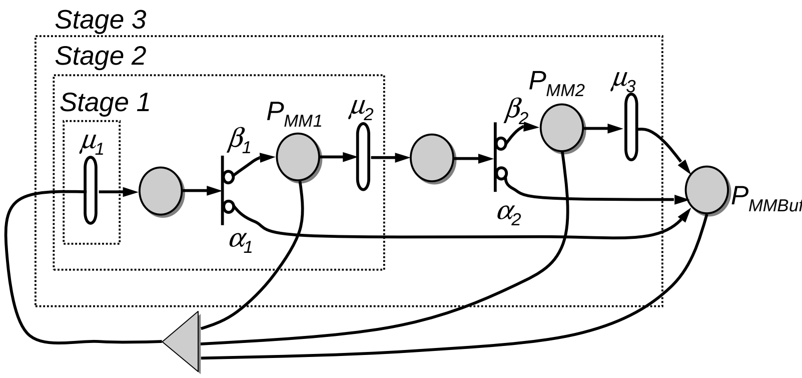


- The input inter-arrival times of a multimedia (MM) request generally follow a non-exponential distribution
- The priority of the MM request is higher than that of the local (normal) application request
- *J* and *L* constraints are only applied to the MM applications

Top Level GSPN Model



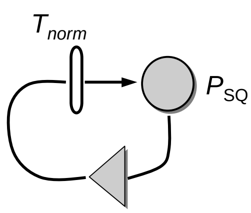
GSPN Model for the MM SR & SQ



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

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

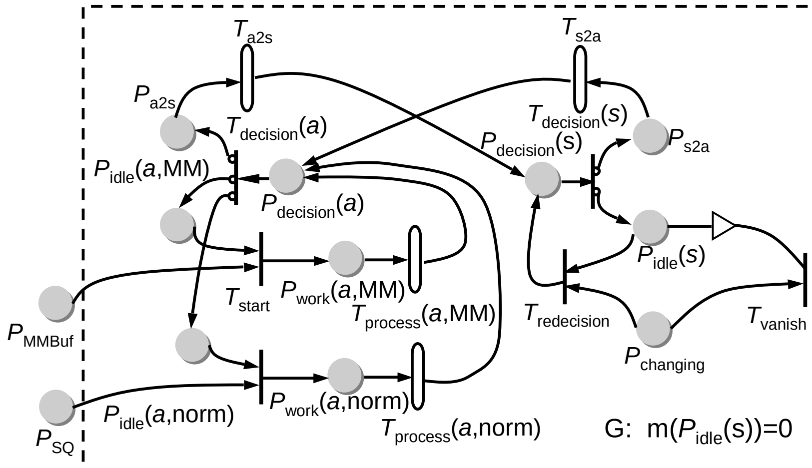
GSPN Model for the Local SR & SQ



G_{norm}: m(P_{SQ}) < SQ capacity

- Assume that the inter-arrival time of local requests follows an exponential distribution

GSPN Model for the SP



P_{status}(p_mode, a_mode)

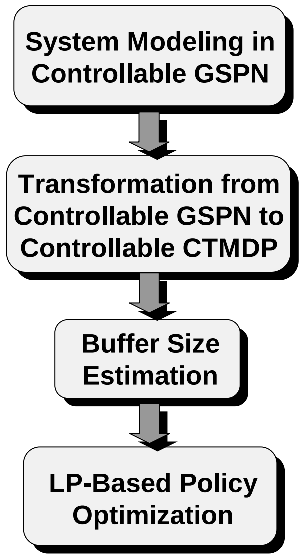
- Status: idle, busy, transition, decision

- p_mode: active, sleeping, etc.
- a_mode: MM application, local application

Cost Definition

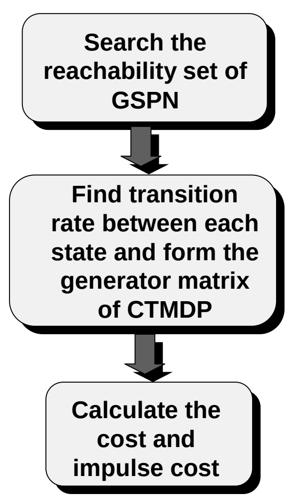
- Rate cost: d_r , j_r , l_r , ld_r , pow_r
  - P_{MMBuf}
    - d_r = #tokens
    - j_r = (#tokens - Ave(#tokens))^2
    - l_r = 1 when #tokens = MM Buffer size
  - P_{SQ}
    - ld_r = #tokens
  - P_{status}(power_mode, application_mode)
    - pow_r = power consumption of SP in its current state
- Impulse cost: *ene*
  - ene_{ij}: Energy needed for the SP to switch from state *i* to state *j*

PQM Policy Optimization Workflow



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



PQM Policy Optimization Workflow

System Modeling in Controllable GSPN

↓

Transformation from Controllable GSPN to Controllable CTMP

↓

Buffer Size Estimation

↓

LP-Based Policy Optimization

Buffer Size Estimation

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

(D, J, L) = (1.5, 0.9, 0.02)

Buffer Size	D	J	L	Power
4	1.23	0.75	0.02	2.08
6	1.5	0.9	0.02	1.49

- 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

Buffer Size Estimation

Min. n

Subject to:

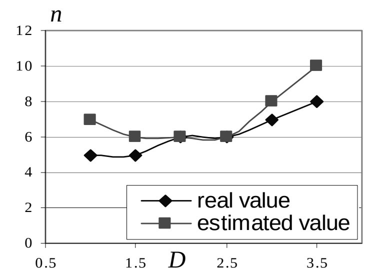
$\boldsymbol{p}_n \leq L$

$0 \leq \boldsymbol{p}_i \leq 1,\ i = 1, \ldots, n$

$(n-2)^2 \cdot L \geq J + D^2 - 4 \cdot D + 3$

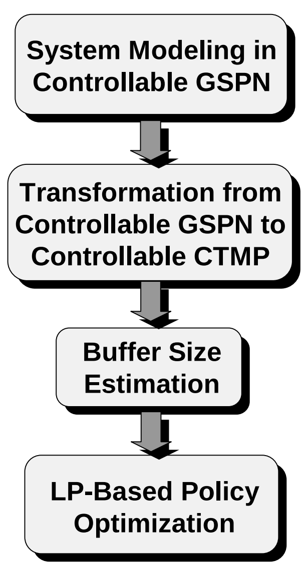
$(n+1) \cdot (n-2) \cdot L \geq D - 2$

$(n-2) \cdot (n-1) \cdot L \geq D^2 - 3 \cdot D + 2 + J$



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

PQM Policy Optimization Workflow



PO Problem Formulation

Subject to:

$p_{ij}^{a_i}$: probability that the next system state is *j*

$\tau_i^{a_i}$: expected duration of time that the system will be in state *i*

$x_i^{a_i}$: frequency that the state of the system will be *i* and action a_i will be taken; Note that:

$$x_i^{a_i} \cdot \tau_i^{a_i} \equiv p_i^{a_i}$$

pow_i: power consumption in state *i*

q_MMBuf_i: number of unprocessed data in the MM buffer

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 , 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

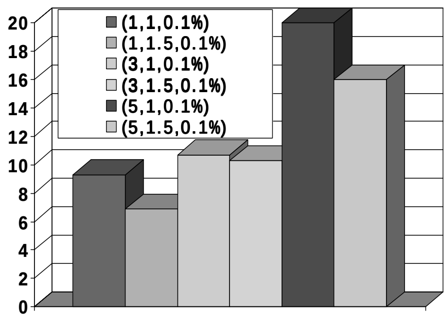
- System setup
  - SP has two power modes: high power and low power
  - In the PD_optimized system, we end up overconstraining delay in order to satisfy the jitter constraints

Power Consumption (mW)

	MM	local	idle
High Power	4	3	2
Low Power	2	1.5	1

Service Speed (ms)

	MM	local	idle
High Power	5	2	-
Low Power	10	2.5	-



Percentage Power Saving of PQM over PDM

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 plicies are more power-efficient than the PD-optimized policies under the same D, J and L constraints