Cross-Layer Optimization for Transmission of Delay-Sensitive and Bursty Traffic in Wireless Systems

Somsak Kittipiyakul

ECE, UCSD

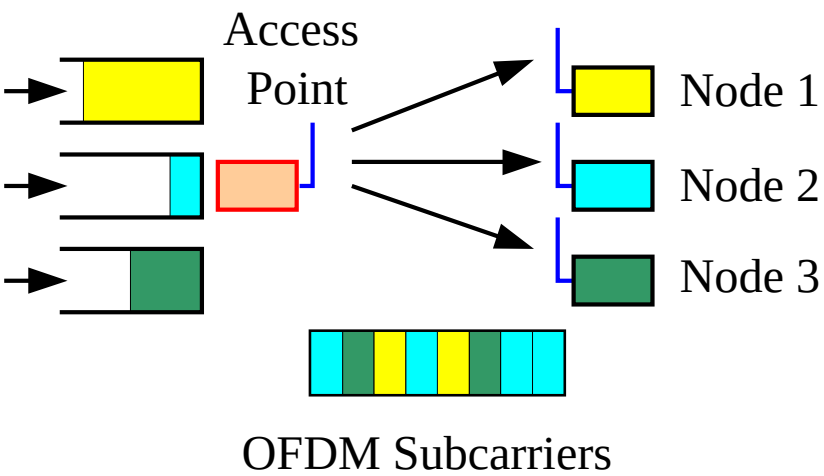
August 28, 2008

Cross-layer optimization in wireless data networks

Scenario 1: Channel state information available at Transmitter (CSIT) [Chapter 2]

Why cross-layer ...?

- Wireless channel is time-varying
- Knowledge of CSI is useful
  - Improve network layer QoS (packet delay/loss) and physical layer (PHY) performance



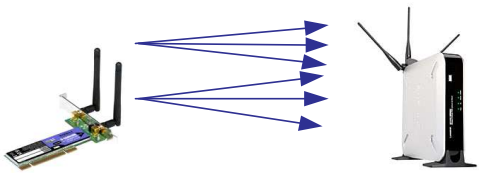
Example: OFDMA subcarrier allocation minimizing avg. packet delay

- Knowledge: CSI, queue state info (QSI)
- Resource: subcarriers

Cross-layer optimization in wireless data networks

Scenario 2: no CSIT and no feedback [Chapters 3-5]

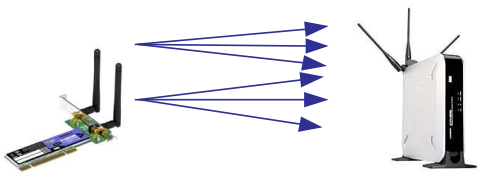
- CSI may be unavailable at the transmitter, due to
  - No feedback from the receiver
  - High estimation error and complexity (multiple Tx/Rx antennas, high channel fluctuations, etc.)



Cross-layer optimization in wireless data networks

Scenario 2: no CSIT and no feedback [Chapters 3-5]

- CSI may be unavailable at the transmitter, due to
  - No feedback from the receiver
  - High estimation error and complexity (multiple Tx/Rx antennas, high channel fluctuations, etc.)



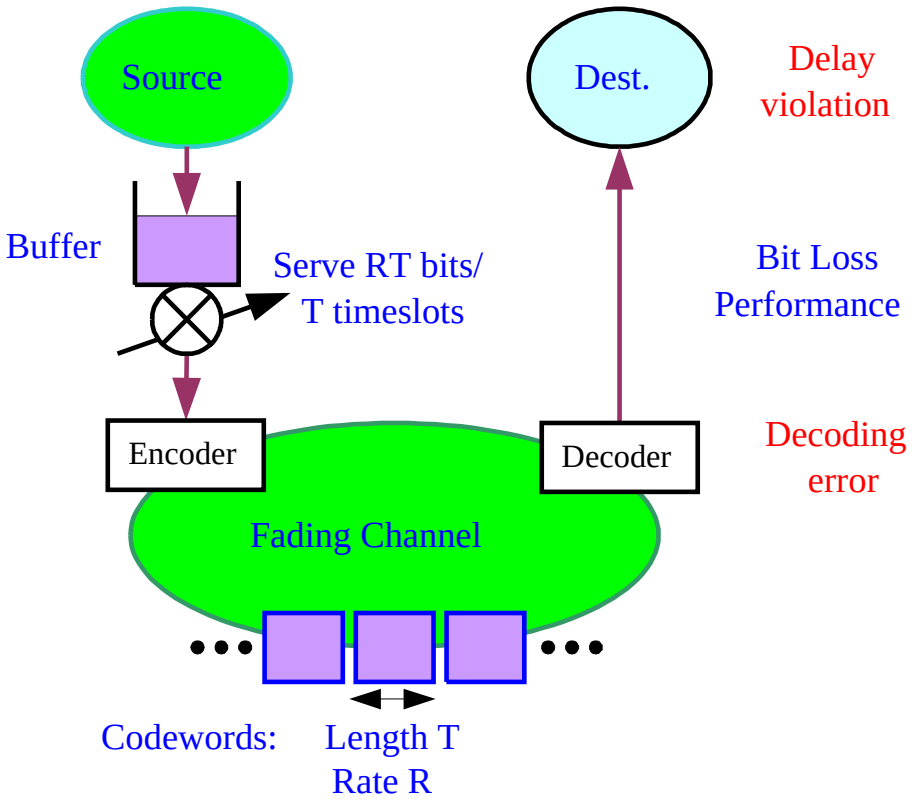
Still, channel variations can be used to ...

- Improve transmission reliability and/or capacity in different orthogonal domains (time, frequency, space, users, etc.)

Question of interest

- How to select the optimal PHY parameters and derive achievable QoS?

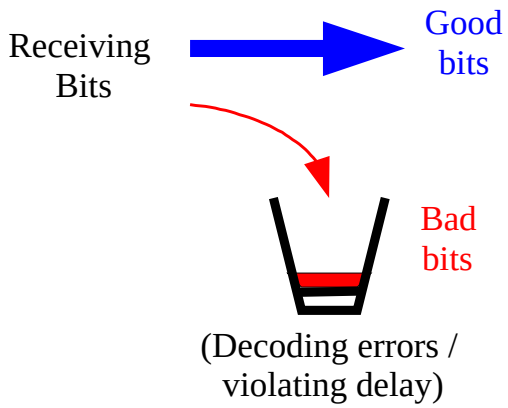
Example: Cross-Layer Optimization in Point-to-Point Setting



- Time-slotted system
- Stationary bit-source
- Infinite FIFO buffer with batch service
- Block coding, rate R, length T
- Decode codeword at the end of its reception

- Destination drops bits when they are decoded incorrectly or violate delay constraint D

High-SNR Asymptotic Approach

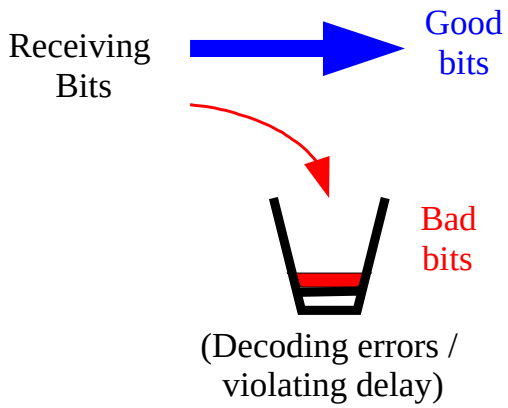


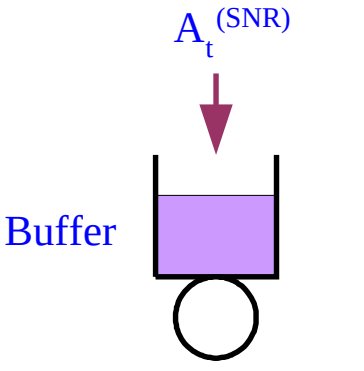
Cross-Layer QoS: Total error probability

$$P_{\text{tot}}(R,T) \approx P(\text{decoding error}) + P(\text{delay} > D)$$
$$\approx \text{SNR}^{-d(r,T)} + P(\text{delay} > D)$$

where transmission rate $R = r \log \text{SNR}$.

High-SNR Asymptotic Approach





Capacity r log SNR per timeslot,
Batch service every T timeslots

Cross-Layer QoS: Total error probability

$$P_{\text{tot}}(R,T) \approx P(\text{decoding error}) + P(\text{delay} > D)$$
$$\approx \text{SNR}^{-d(r,T)} + P(\text{delay} > D)$$

where transmission rate $R = r \log \text{SNR}$.

Need to solve queueing problems ...

- Analyze $P(\text{delay} > D)$ for batch service at high SNR but finite D
- Find source scaling conditions with SNR such that

$$P(\text{delay} > D) \approx \text{SNR}^{-I(r,T,D)}$$

Outline

- Asymptotic analysis of queueing performance

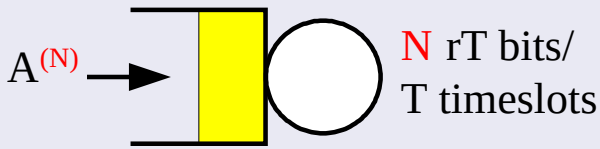
1. Single-User Setting: Batch Service (Chapter 3)
   - Problem Setting and Review of Large Deviations Analysis
   - Analysis of Batch-Service
2. Multi-user Setting: Longest-Queue-First Scheduling (Chapter 5)
   - Problem Setting and Review of Large-Deviations Analysis
   - Contraction Principle
   - LDP of Workload Processes

Outline

Asymptotic analysis for batch-service queue

- Let $N = \log \mathrm{SNR}$

Problem Setting

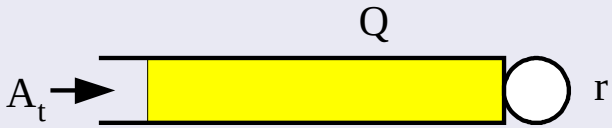


- Index system by N.
- Batch service capacity NrT bits every T timeslots
- No. of bits arrived at timeslot t is A_t^N
- iid increment process with mean $E[A_1^N] = N\lambda$, $\lambda < r$ for stability
- Goal: $P(\text{steady-state delay} > D) \approx e^{-I(r,T,D)N}$ for large N

- How to scale arrival processes A^N?

Prior Work: Two Large-Deviations Frameworks

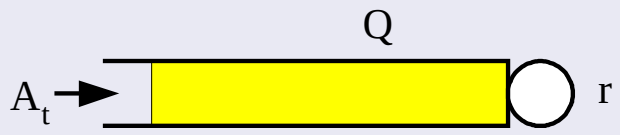
Large-buffer framework



- $\Pr(Q > b)$ for $b \to \infty$

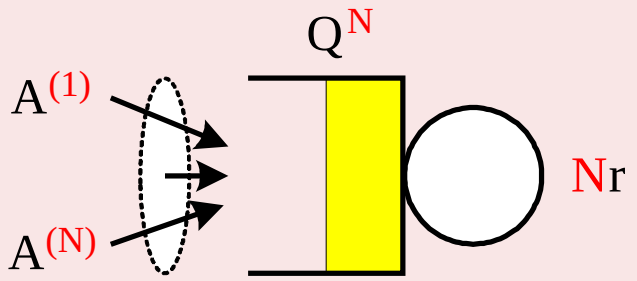
Prior Work: Two Large-Deviations Frameworks

Large-buffer framework



- $\Pr(Q > b)$ for $b \to \infty$

Many-flows framework



- $\Pr(Q^N > Nb)$ for $N \to \infty$

General Results

- $\Pr(Q > b) \approx e^{-\delta b}$
- $\Pr(Q^N > Nb) \approx e^{-f(b)N}$

- Both frameworks are interested in probability of rare events

Large Deviation Principle (LDP): Formal Definition

Definition

A sequence of random variables X^N in a Hausdorff space $\mathcal{X}$ with σ-algebra $\mathcal{B}$ is said to satisfy an LDP with rate function $I : \mathcal{X} \mapsto [0, \infty]$ if for any $B \subseteq \mathcal{B}$,

$$-\inf_{x \in B^o} I(x) \leq \liminf_{N \to \infty} \frac{1}{N} \log \Pr(X^N \in B)$$
$$\leq \limsup_{N \to \infty} \frac{1}{N} \log \Pr(X^N \in B) \leq -\inf_{x \in \bar{B}} I(x)$$

When X^N is a sample path of random sequence (random process), then the LDP is called sample-path LDP. When I is continuous,

$$\lim_{N \to \infty} \frac{1}{N} \log \Pr(X^N \in B) = -\inf_{x \in B} I(x)$$

Interpretation of LDP

LDP definition for our interest is ...

A sequence of random variables X^N in $\mathbb{R}^d$ is said to satisfy an LDP with rate function $I : \mathbb{R}^d \mapsto [0, \infty]$ if for $\mathbf{b} > E[X^N]$,

$$\Pr(X^N > \mathbf{b}) \approx e^{-I(\mathbf{b})N}$$

- Intuitively, LDP is a principle of the largest term: if $a < b$ then
$$e^{-aN} + e^{-bN} \approx e^{-aN}$$
- We can think of $\Pr(X^N \approx a) \approx e^{-I(a)N}$ and for a set B
$$\Pr(X^N \in B) \approx e^{-[\inf_{x \in B} I(x)]N}$$

Outline

Assumption: LDP for arrival process A^N

"Smoothly-scaling" assumption on arrival process A^N

For each $\theta \in \mathbb{R}$, $\Lambda(\theta) = \lim_{N\to\infty} \frac{1}{N} \log E\left[e^{\theta A_1^N}\right]$ exists as an extended real number in $[0, \infty]$ and is finite in a neighborhood of the origin, essentially smooth, and lower-semicontinuous.

- Then, our arrival process satisfies an LDP.

Gärtner-Ellis Theorem

If A^N is smoothly-scaling, then $\frac{A_1^N}{N}$ satisfies an LDP with rate function Λ^* where Λ^* is a convex conjugate of Λ:

$$\Lambda^*(x) = \sup_{\theta\in\mathbb{R}} \theta x - \Lambda(\theta)$$

Main Result: Asymptotic P_{delay}

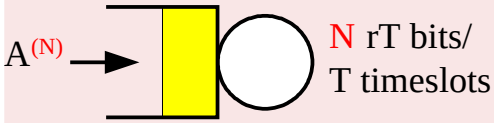
Consequence: LDP on accumulated arrivals

For any $t \in \mathbb{N}$ and $a > \lambda$ (recall $E[A_1^N]/N = \lambda$),

$$P\left((A_1^N + \cdots + A_t^N)/N > at\right) \approx e^{-t\Lambda^*(a)N}$$

Main Result:

The stable batch queueing system with arrival process A^N where $r > \lambda$ and $T < D/2$ has

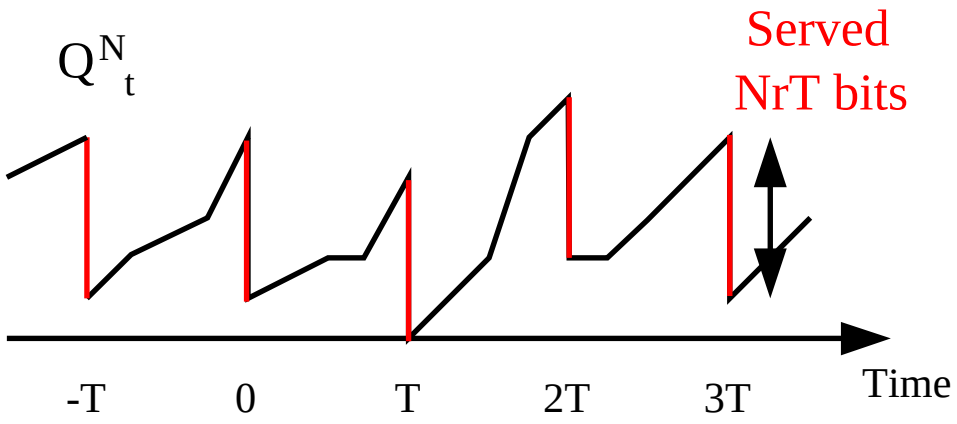


$$P_{\text{delay}} := P(\text{delay} > D) \approx e^{-I(r,T,D)N}$$

where, for $k = D(\bmod T)$, the rate function I is

$$I(r,T,D) = \min_{\substack{t\in\mathbb{N}:\\ \tau = tT+T-1-k>0}} \tau\Lambda^*\left(r + \frac{(D+1-2T)r}{\tau}\right)$$

Proof Idea



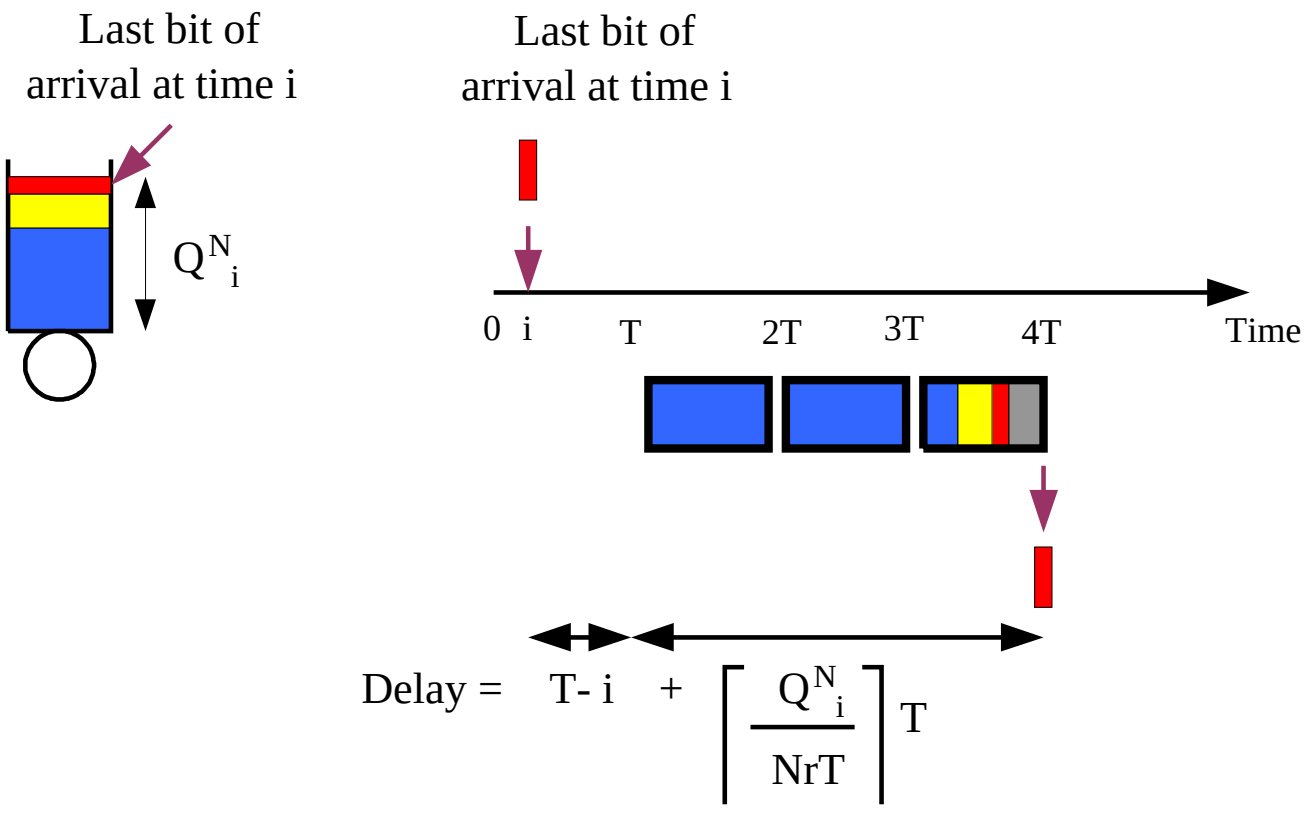
- Assuming queues started empty at time $-\infty$, we can take queue length at time i (Q_i^N), $i = 0, \ldots, T-1$, as having the steady-state distributions

$$P_{\text{delay}} := P(\text{steady-state delay of a bit} > D)$$

$$= \frac{1}{T}\sum_{i=0}^{T-1} P(\text{s-s delay of a bit arriving at time } i > D)$$

$$\approx \frac{1}{T}\sum_{i=0}^{T-1} P(\text{delay of last bit arriving at time } i > D)$$

Proof Idea (2)



$$P(\text{delay of last bit arriving at time } i > D)$$
$$= P\left(T - i + \left\lceil \frac{Q_i^N}{NrT} \right\rceil T > D\right)$$

Proof Idea (3)

- P_{delay} is dominated by the event that the last bit arriving at time $T-1-k$ sees a queue length greater than $(D-T-k)rN$ bits, where $k = D \mod T$:

$$P_{\text{delay}} \approx P(Q^N_{T-k-1} > (D-T-k)rN)$$

Proof Idea (3)

- P_{delay} is dominated by the event that the last bit arriving at time $T-1-k$ sees a queue length greater than $(D-T-k)rN$ bits, where $k = D \mod T$:

$$P_{\text{delay}} \approx P(Q^N_{T-k-1} > (D-T-k)rN)$$

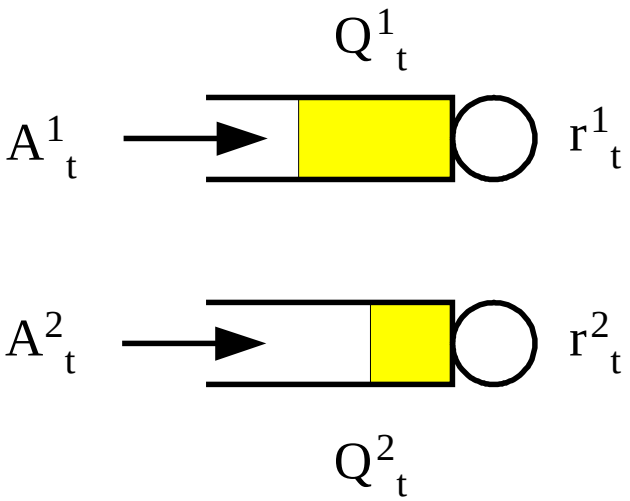
- Finally, use
  - Queue expression

$$Q_i^N = \sup_{\tau \in \mathbb{N}} \left(\sum_{j=-\tau T+1}^{i} A_j^N - NrT\tau \right)$$

  - Standard large-deviations upper and lower bounds
  - LDP of the accumulated arrivals by Gärtner-Ellis theorem

Outline

Large-Deviations Analysis of Multi-user Dynamic Scheduling

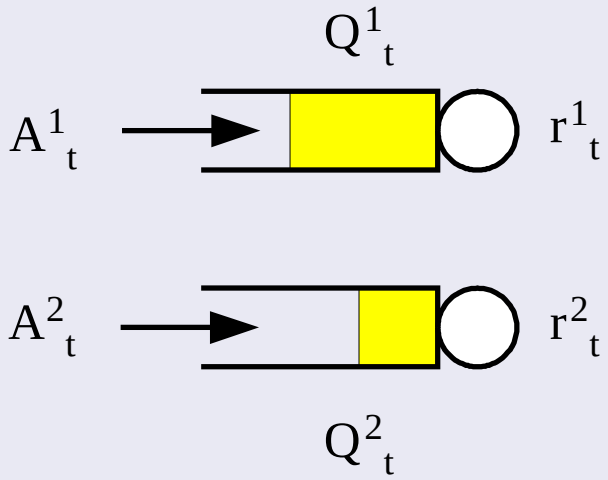


Problem Setting

- Allocate rate vector $(r_t^1, r_t^2) = H(Q_t^1, Q_t^2)$ in rate region $\mathcal{R}$, according to scheduler H
- Quantity of interest:
  - Stationary workload vector $\mathbf{Q} = (Q^1, Q^2)$
  - Buffer overflow probability $\Pr(\mathbf{Q} > \mathbf{b})$
- Perform a large-deviations asymptotic approximation!

Asymptotic Approximations: Prior Results

Large-buffer framework

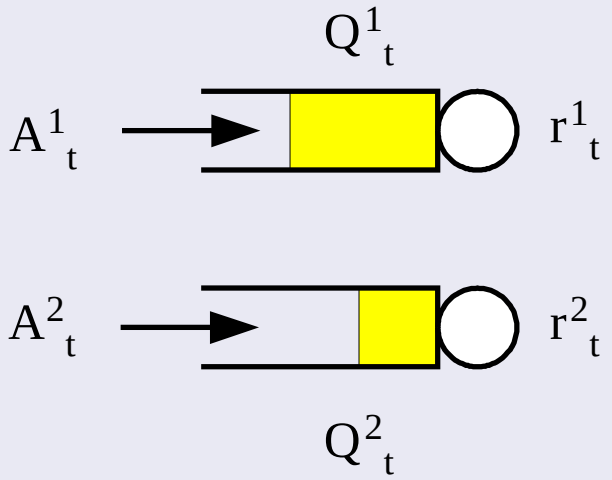


Prior Results

- Bertsimas et al'98: LQF, simplex $\mathcal{R}$, 2 queues
- Stolyar&Ramanan'01: WLDF, K queues
- Subramanian'08: WLQF, convex $\mathcal{R}$

Asymptotic Approximations: Prior Results

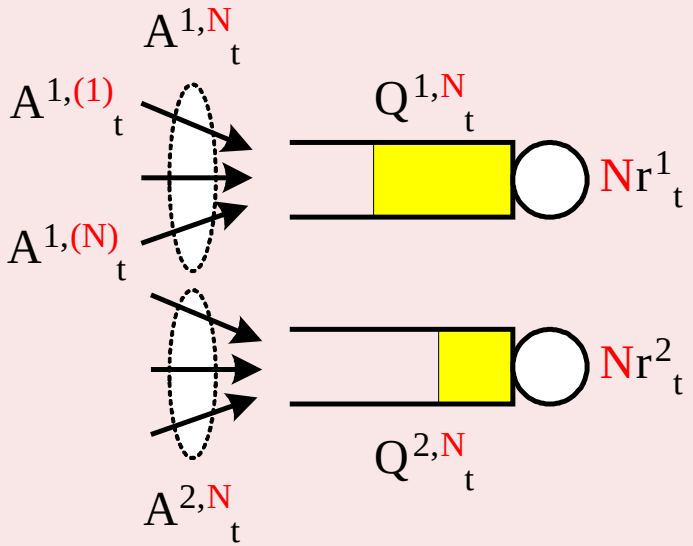
Large-buffer framework



Prior Results

- Bertsimas et al'98: LQF, simplex $\mathcal{R}$, 2 queues
- Stolyar&Ramanan'01: WLDF, K queues
- Subramanian'08: WLQF, convex $\mathcal{R}$

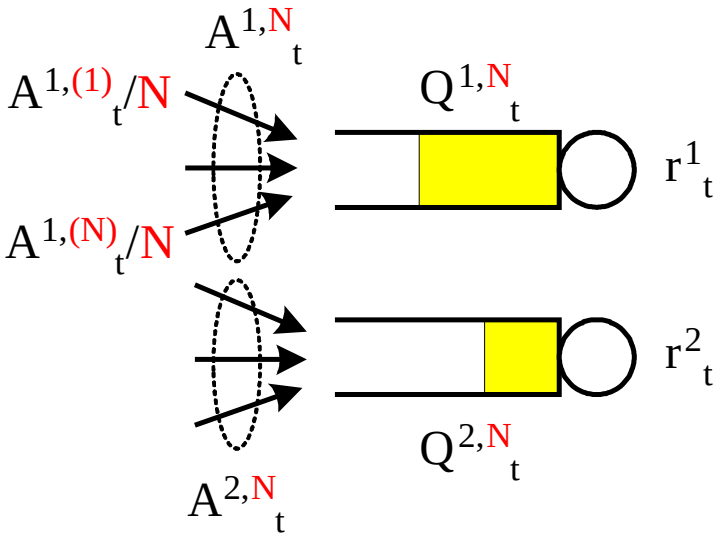
Many-flows framework

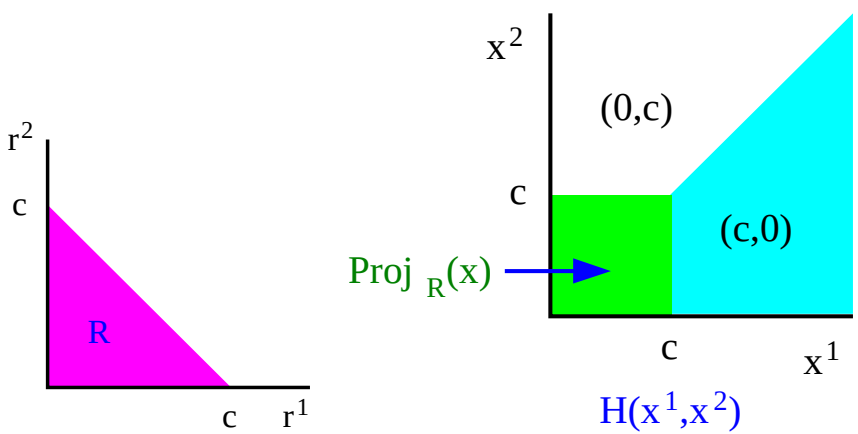


Prior Results

- Shakkottai&Srikant'01: GPS scheduler
- Yang et al'07: schedulers flavoring short jobs

Analysis Setting





Many-flows framework

- Index by N, arrival to queue k is an average of N iid flows

$$A_t^{k,N} = \frac{1}{N} \sum_{n=1}^{N} A^{k,(n)}$$

- Simplex rate region $\mathcal{R}$, total capacity c
- Non-idling LQF scheduler H
- Assume a sample-path LDP of arrival processes $A^N = (A^{1,N}, A^{2,N})$

Approach and Contribution

Approach

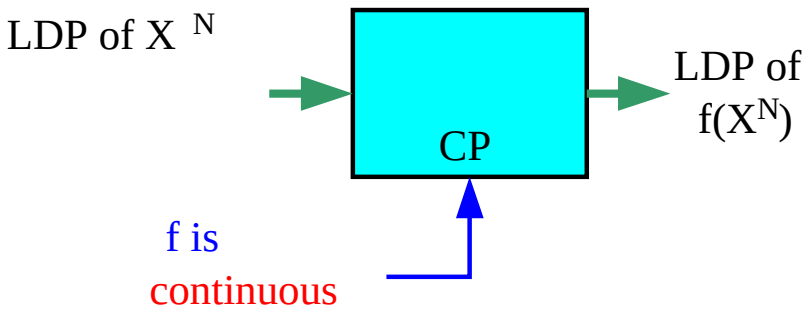
1. Assume an LDP of arrival processes A^N
2. Establish an LDP of stationary workload $\mathbf{Q}^N$ via a Contraction Principle
   - Require quasi-continuity of the mapping $A^N \mapsto \mathbf{Q}^N$
3. Calculation of the corresponding rate function

Our contribution is

- Establish an LDP of stationary workloads for a LQF scheduler under simplex rate region, in many-flows framework

Outline

Contraction Principle: An LDP Machine

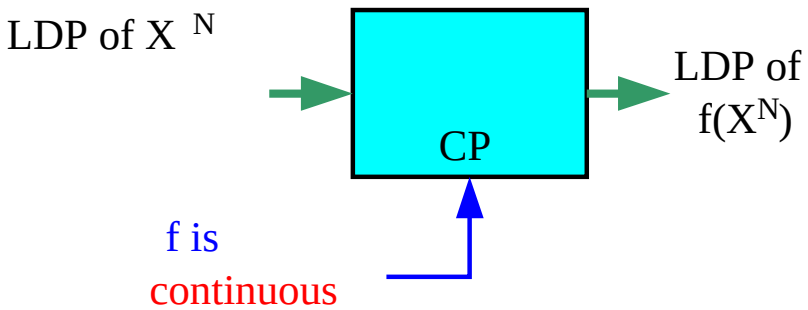


Contraction Principle

(i) X^N satisfies LDP on $\mathcal{X}$ with a rate function I;
(ii) $f : \mathcal{X} \mapsto \mathcal{Y}$ be a continuous mapping
$\Rightarrow f(X^N)$ satisfies LDP with rate function J

$$J(y) = \inf_{x \in \mathcal{X}: f(x)=y} I(x)$$

Contraction Principle: An LDP Machine

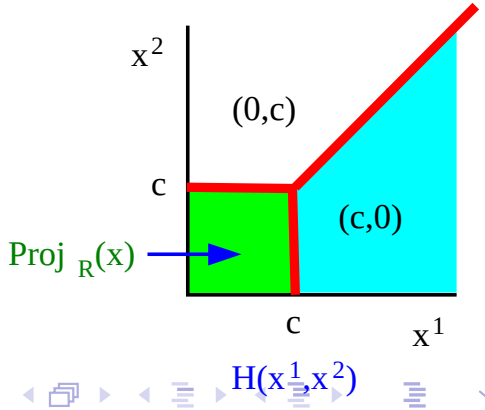


Contraction Principle

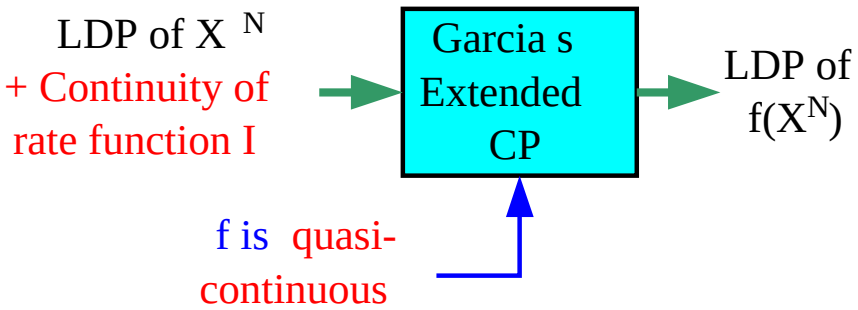
(i) X^N satisfies LDP on $\mathcal{X}$ with a rate function I;
(ii) $f : \mathcal{X} \mapsto \mathcal{Y}$ be a continuous mapping
$\Rightarrow f(X^N)$ satisfies LDP with rate function J

$$J(y) = \inf_{x \in \mathcal{X}: f(x)=y} I(x)$$

- But our scheduling function H and hence workload mapping are not continuous!!



Contraction Principle: An LDP Machine

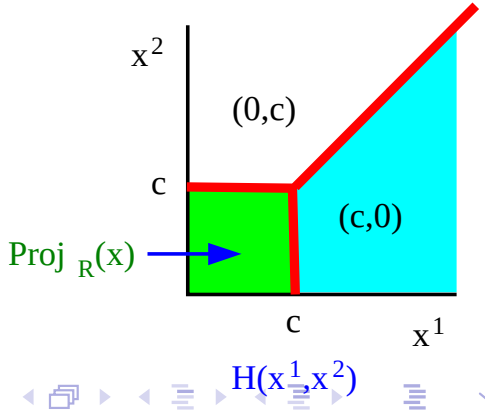


Garcia's Contraction Principle

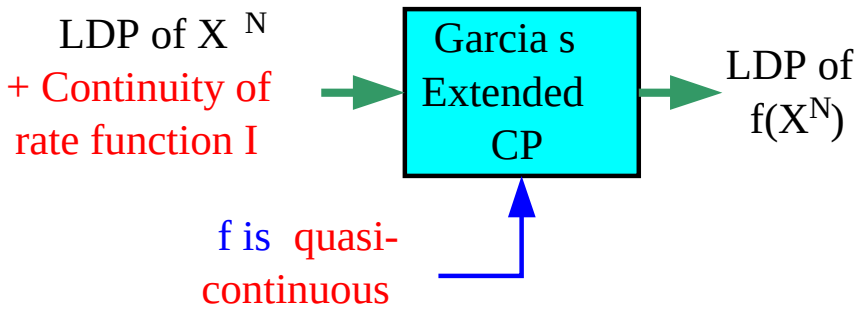
(i) X^N satisfies LDP on $\mathcal{X}$ with a continuous rate function I;
(ii) $f : \mathcal{X} \mapsto \mathcal{Y}$ be a quasi-continuous mapping
$\Rightarrow f(X^N)$ satisfies LDP with rate function J

$$J(y) = \inf_{x \in \mathcal{X}: f^x \ni y} I(x)$$

- But our scheduling function H and hence workload mapping are not continuous!!



Use... Garcia's Extended Contraction Principle



Garcia's Extended CP

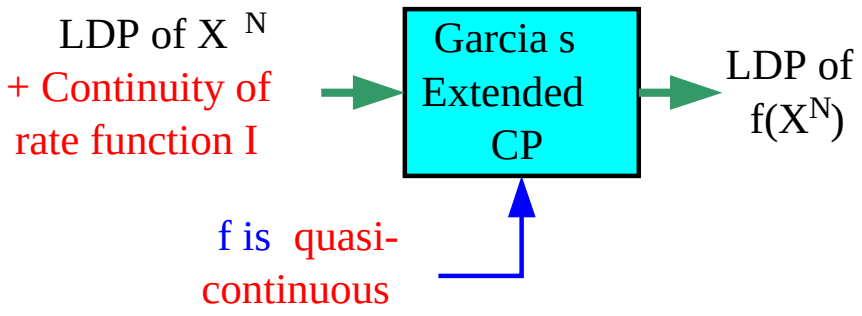
For quasi-continuous mapping f and metric spaces $\mathcal{X}, \mathcal{Y}$,

$$J(y) = \inf_{x:\ f^x \ni y} I(x)$$

Definition of f^x and examples

$f^x := \{y \in \mathcal{Y} \mid \exists x_n \to x \text{ such that } f(x_n) \to y\}$

Use... Garcia's Extended Contraction Principle



Garcia's Extended CP

For quasi-continuous mapping f and metric spaces $\mathcal{X}, \mathcal{Y}$,

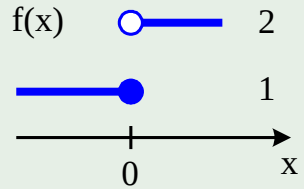
$$J(y) = \inf_{x:\ f^x \ni y} I(x)$$

Definition of f^x and examples

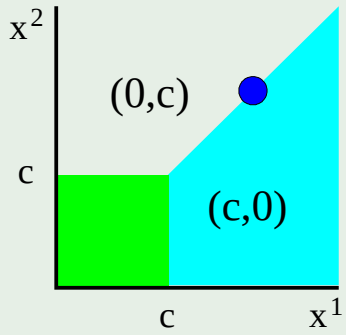
$$f^x := \{y \in \mathcal{Y} \mid \exists x_n \to x \text{ such that } f(x_n) \to y\}$$

Examples:

- $f^x = \{f(x)\}$ if f is continuous at x



$f^0 = \{1,2\}$



$H^{(2c,2c)} = \{(c,0),(0,c)\}$

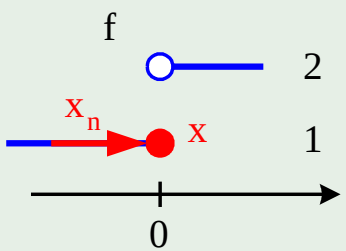
Quasi-continuity

Definition: Quasi-continuity for metric spaces

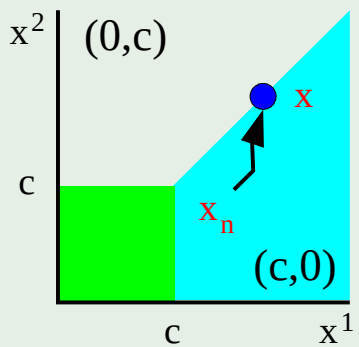
- $f : \mathcal{X} \mapsto \mathcal{Y}$ is quasi-continuous at $x \in \mathcal{X}$ if there is a sequence $\{x_n\}$ such that $x_n \to x, f(x_n) \to f(x)$, and such that for all n, f is continuous at x_n.
- f is quasi-continuous if it is quasi-continuous at every point of its domain.

Examples

- (Lower or upper) semicontinuous function on $\mathbb{R}$



- LQF Scheduler H



Outline

Quasi-continuity of the Workload Mapping

What is stationary workload Q?

- Stationary (or infinite-horizon) workload vector Q is the workload at time zero, given the queues started "empty" at time $-\infty$.
- $Q = \lim_{T \to \infty} Q_{0,T}$
  where the finite-horizon workload $Q_{0,T}$ is the workload at time zero when the queues started "empty" at time $-T$.

Quasi-continuity of the Workload Mapping

What is stationary workload Q?

- Stationary (or infinite-horizon) workload vector Q is the workload at time zero, given the queues started "empty" at time $-\infty$.
- $Q = \lim_{T\to\infty} Q_{0,T}$
  where the finite-horizon workload $Q_{0,T}$ is the workload at time zero when the queues started "empty" at time $-T$.

Steps in proving quasi-continuity of Q

1. Show quasi-continuity of finite-horizon workload $Q_{0,T}(A)$
2. Show quasi-continuity of stationary workload $Q(A)$ for arrival A whose sample mean is eventually equal to the mean of the arrival process.

Quasi-continuity of $Q_{0,T}$

Proof Idea

- Induction in time, using the queue dynamics

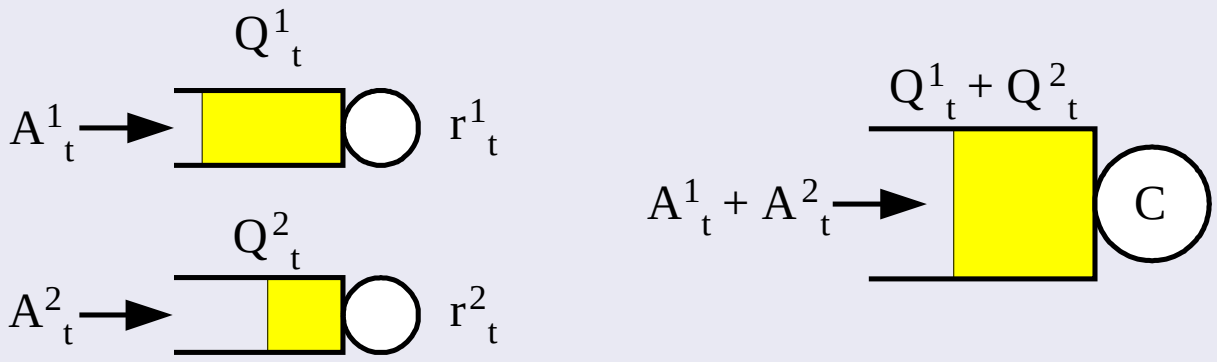
$$Q_{t-1}^k = [Q_t^k - r_t^k]^+ + A_t^k$$

where $\mathbf{r}_t = H(\mathbf{Q}_t)$, and quasi-continuity of H

Quasi-continuity of Q

Proof Idea

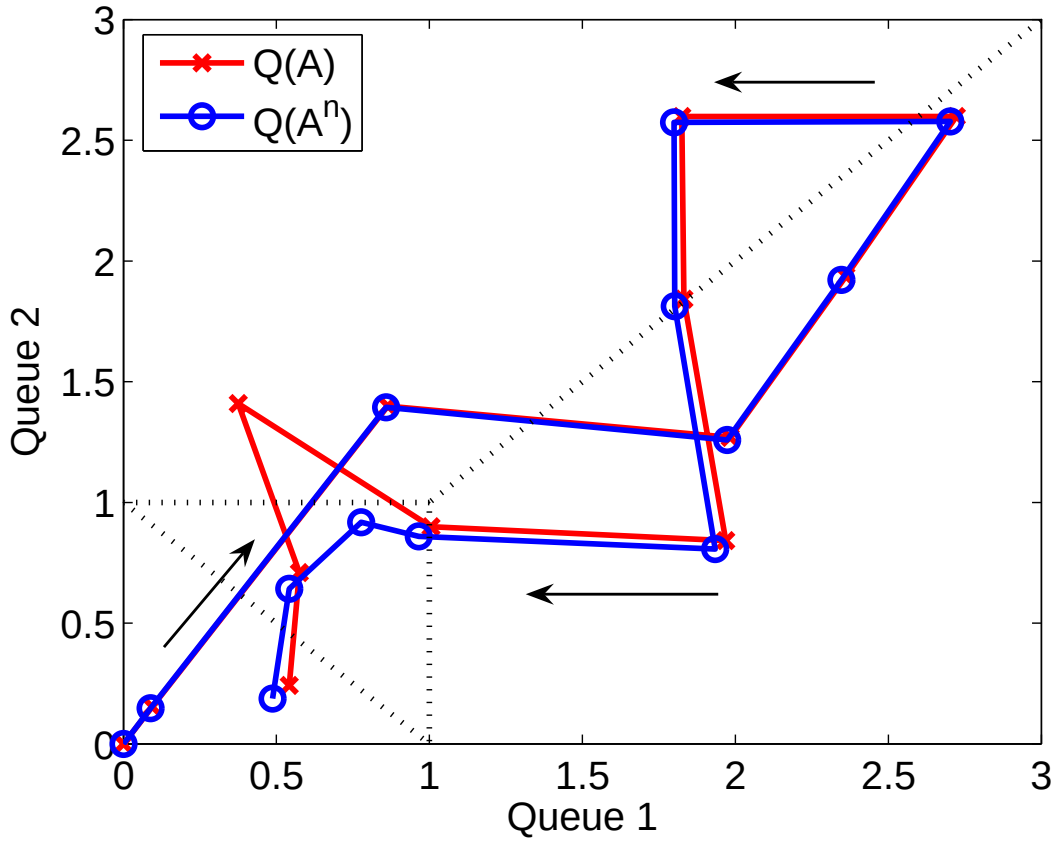
- For any $A^n \to A$, for all n sufficiently large there is a common time $-s^*$ where the queues are "empty" under A and A^n.
  - Shown by mapping to a single queue system with sum arrival process and service rate c.
  - All queues are "empty" when the sum workload is "empty"



- Use quasi-continuity of Q_{0,s^*}

Quasi-continuity of Q: Illustration

- An example of a busy-period portion of workload sample paths under A and A^n, when A^n is "close" to A



Summary of Results

Assumption: sample-path LDP of arrival processes

- Assume sample-path LDPs of $\{A^N\}$ with a continuous rate function $I^\sharp$ and of $\{A^N|_{(0,t]}\}$ with a continuous rate function $I^\sharp_t$ for $t \in \mathbb{N}$.

Result 1: LDP for finite-horizon workloads

- $Q_{0,t}(A^N)$ satisfies LDP, with rate function I_t,

$$I_t(\mathbf{b}) = \inf_{\mathbf{x} \in \mathbb{R}_+^{K \times t} : \mathbf{b} \in Q_{0,t}^{\mathbf{x}}} I^\sharp_t(\mathbf{x})$$

Result 2: LDP for stationary workloads

- $Q(A^N)$ satisfies LDP, with rate function J,

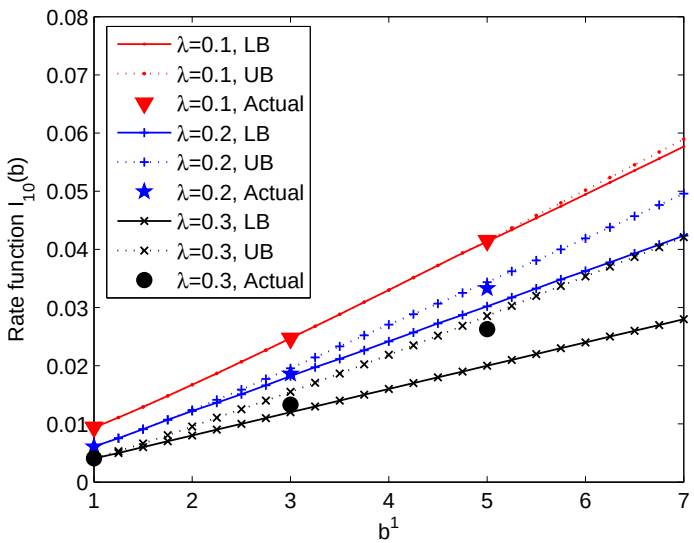
$$J(\mathbf{b}) = \inf_{a \in \mathcal{D}^K : \mathbf{b} \in Q^a} I^\sharp(a)$$

Simplification of Rate Function for iid increment flows

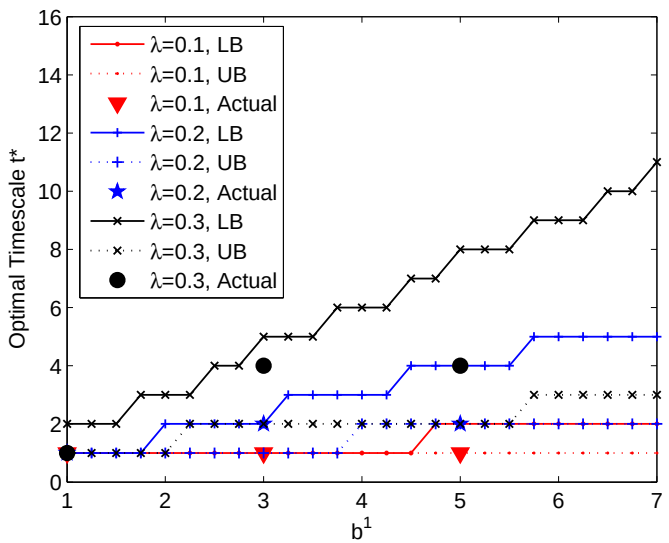
- Additive and convex cost: $I^{\sharp}_t(\mathbf{a}) = \sum_{i=1}^{t} \sum_{k=1}^{K} \Lambda^*(a_i^k)$
- Simplified rate function

$$J(\mathbf{b}) = \inf_{t\in\mathbb{N}} I_t(\mathbf{b}) = \inf_{t\in\mathbb{N}} \inf_{\mathbf{a}\in\mathbb{A}(t,\mathbf{b})} I^{\sharp}_t(\mathbf{a})$$

- The smallest t^* is the most likely time the queues take to build up to reach level $\mathbf{b}$



(a) Rate functions



(b) Optimal Timescales

Summary and Future Work

Single-user setting with batch service

- Analyzed delay performance of batch service under large-deviations many-flows framework.
- Note that the scaling can be generalized to a "g-smoothly-scaling" (discussed in Chapter 3)
- Future work:
  - Introduce parameter adaptation based on queue knowledge
  - Introduce retransmissions (randomness in batch duration T)

Summary and Future Work (2)

Multi-user setting with Maximum-Weight scheduling

- Garcia's extended contraction principle allowed us to establish a many-sources LDP of the stationary workload under a (quasi-continuous) maximum-weight scheduler with a simplex rate region
- Future work:
  - Extend to other rate regions: MAC or general convex and compact regions
  - Calculation of rate function for non-iid increment arrivals

Publications/Submissions

- Journals:
  - "Optimal operating point for MIMO multiple access channel with bursty traffic," S.K. and T. Javidi, *IEEE Trans. Wireless Commun.*, Dec. 2007.
  - "High-SNR analysis of outage-limited communications with bursty and delay-limited information," S.K., P. Elia, and T. Javidi, submitted to *IEEE Trans. Inf. Theory*, 2007.
  - "Delay-optimal server allocation in multi-queue multi-server systems with time-varying connectivities," S.K. and T. Javidi, submitted to *IEEE Trans. Inf. Theory*, 2007.
  - "Resource allocation in OFDMA with time-varying channel and bursty arrivals," S.K. and T. Javidi, *IEEE Commun. Lett.*, Sept. 2007.
- Conferences:
  - "Many-sources large deviations for max-weight scheduling," S.K., T. Javidi, and V. G. Subramanian, to appear in *Allerton'08*, Sept. 2008.

Publications (2)

- Conferences (cont'd):
  - "Relay scheduling and cooperative diversity for delay-sensitive and bursty traffic," S.K. and T. Javidi, in *Allerton'07*, Sep. 2007.
  - "Cooperative diversity in wireless networks with stochastic and bursty traffic," P. Elia, S.K., and T. Javidi, in *IEEE Int. Symp. Information Theory*, Jun. 2007.
  - "On the Responsiveness-Diversity-Multiplexing tradeoff," P. Elia, S.K., and T. Javidi, in *WiOpt'07*, Apr. 2007.
  - "Optimal operating point in MIMO channel for delay-sensitive and bursty traffic," S.K. and T. Javidi, in *IEEE Int. Symp. Information Theory*, Jul. 2006.
  - "Subcarrier allocation in OFDMA systems: beyond water-filling," S.K. and T. Javidi, in *Asilomar 2004*, Nov. 2004.
  - "A fresh look at optimal subcarrier allocation in OFDMA systems," S.K. and T. Javidi, in *IEEE Conference on Decision and Control (CDC 2004)*, Dec. 2004.

Q & A
Thank you