Effective Capacity Approach to Providing Statistical Quality-of-Service Guarantees in Wireless Networks

Dapeng Wu
Electrical & Computer Engineering
University of Florida

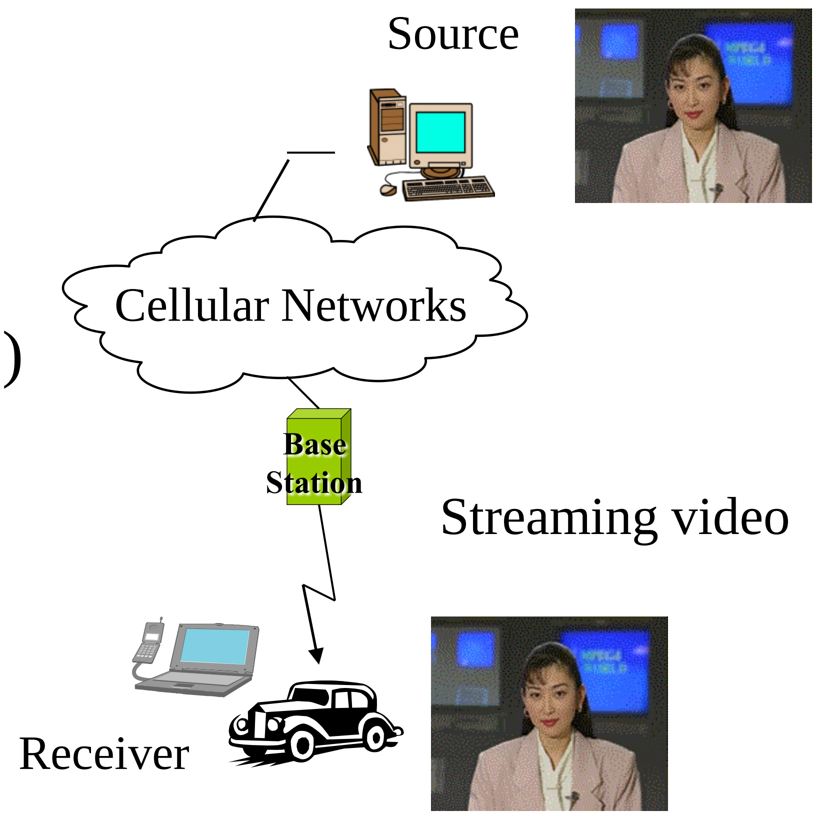
QoS Provisioning in Wireless Networks

Network operator:

Wireless channel $\xrightarrow{\text{How?}}$ QoS provisioning

Quality of Service (QoS)

- Data rate (e.g., 64 kb/s?)
- Delay tolerable (e.g., 1 s?)
- Packet loss probability (e.g., 1%?)



Outline

- Motivation
- Our effective capacity (EC) approach
  - Link layer channel model: EC model
    - ➡ Derive QoS measures
  - Design of QoS provisioning, using EC model
- Summary
- On-going projects

Definition of Statistical QoS

- Statistical QoS: {data rate r_s , D_{max} , P_D}

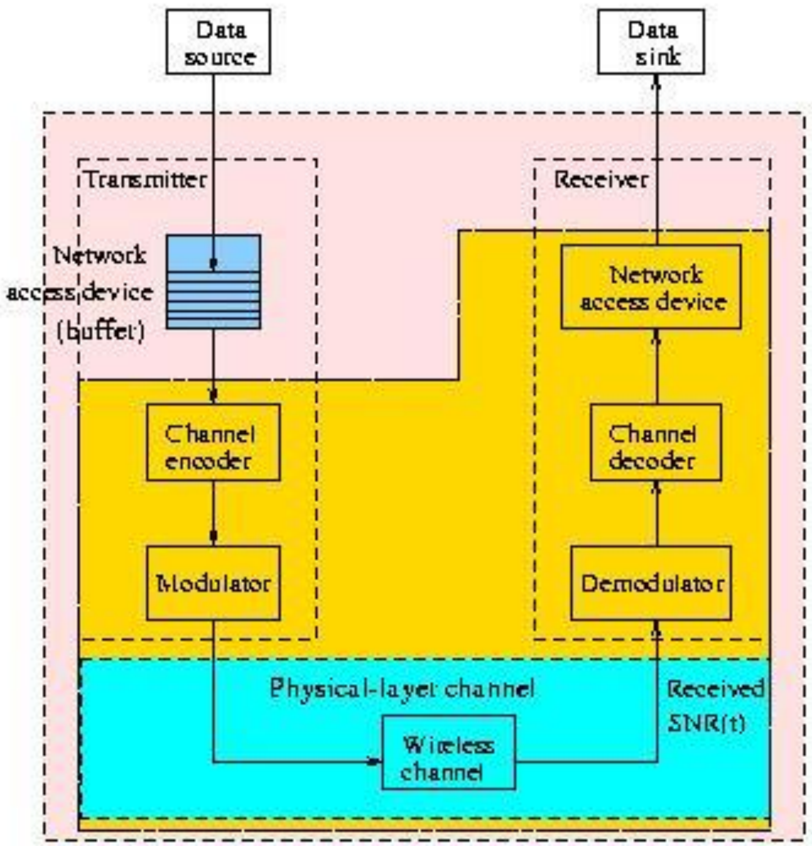
$$P_D = \Pr\{D(\infty) \geq D_{\max}\}$$

D_{max} = max delay tolerable;

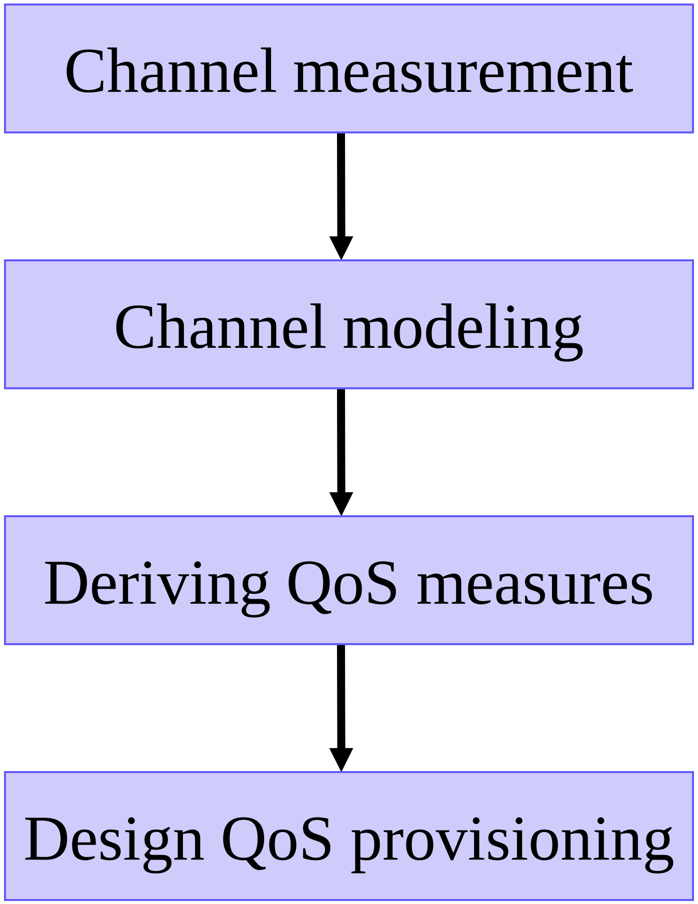
$$D(\infty) = \lim_{t \to \infty} D(t)$$

D(t) = queueing delay of packet

arriving at time t



Traditional Wireless QoS Provisioning



e.g., instantaneous channel gain

e.g., finite state Markov chain

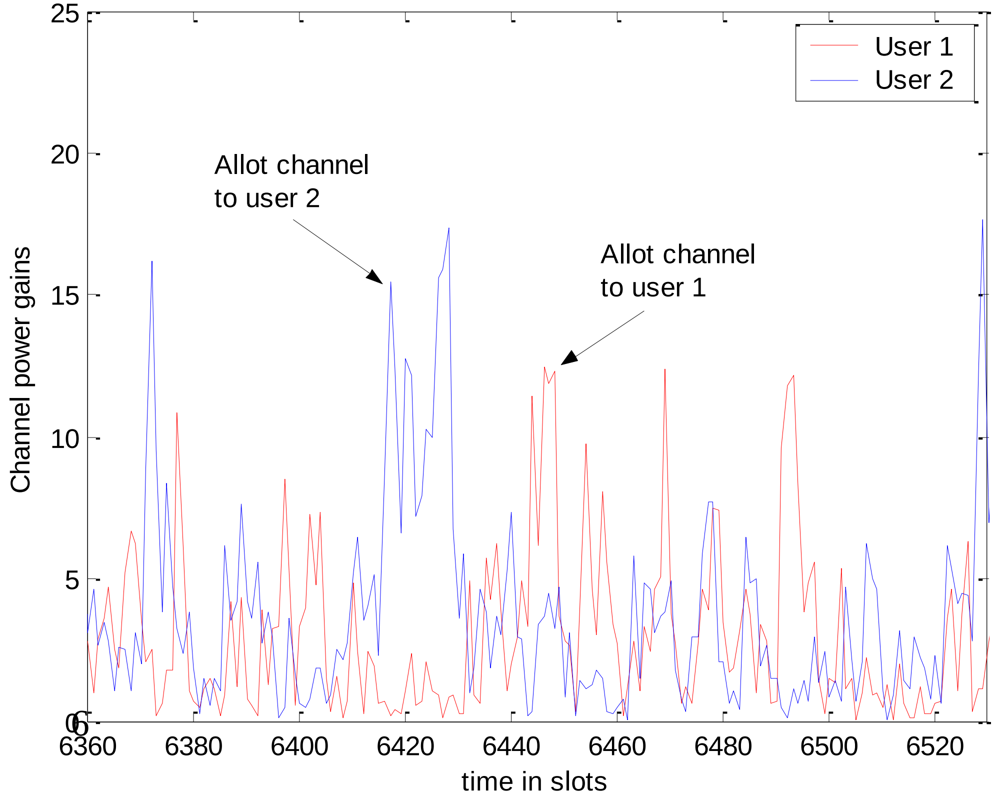
control parameters $\longleftrightarrow$ QoS

What's the complexity of the whole procedure?

E.g., QoS Provisioning with a Scheduler

- Setting
  - One wireless channel
  - *K* users requiring identical QoS
  - Users having independent channel gains

- Scheduler:
  - Knows channel gains perfectly
  - Allots the channel to the best user



Problem: $\{r_s, D_{\max}, \varepsilon\} \longrightarrow$ minimum fraction of channel β? $\beta \in (0,1]$

Solution by traditional approach:

Model each user's capacity $c_k(t)$

Discrete-time Markov chain $X_k(t), \quad k = 1, \cdots, K.$

Model capacity $\tilde{c}_1(t)$ assigned by the scheduler to user 1

$$\tilde{c}_1(t) = \begin{cases} \beta \times c_1(t) & \text{if } X_1(t) = \max X_k(t) \\ 0 & \text{otherwise} \end{cases}$$

Markov chain with K dimensional state: $\{X_1(t), X_2(t), \cdots, X_K(t)\}$

Analyze queue of user 1 for $P_D = \Pr\{D_1(\infty) \geq D_{\max}\}$

Complexity: $O(M^K)$

where M is the number of states in $X_k(t)$.

Resource allocation

$$\min \beta \quad \text{s.t. } P_D \leq \varepsilon$$

Our Approach to QoS Provisioning

Traditional approach:

↓ complex analysis

↓ complex control

Why?

Model channel

at physical layer

Our goals:

↑ simple analysis

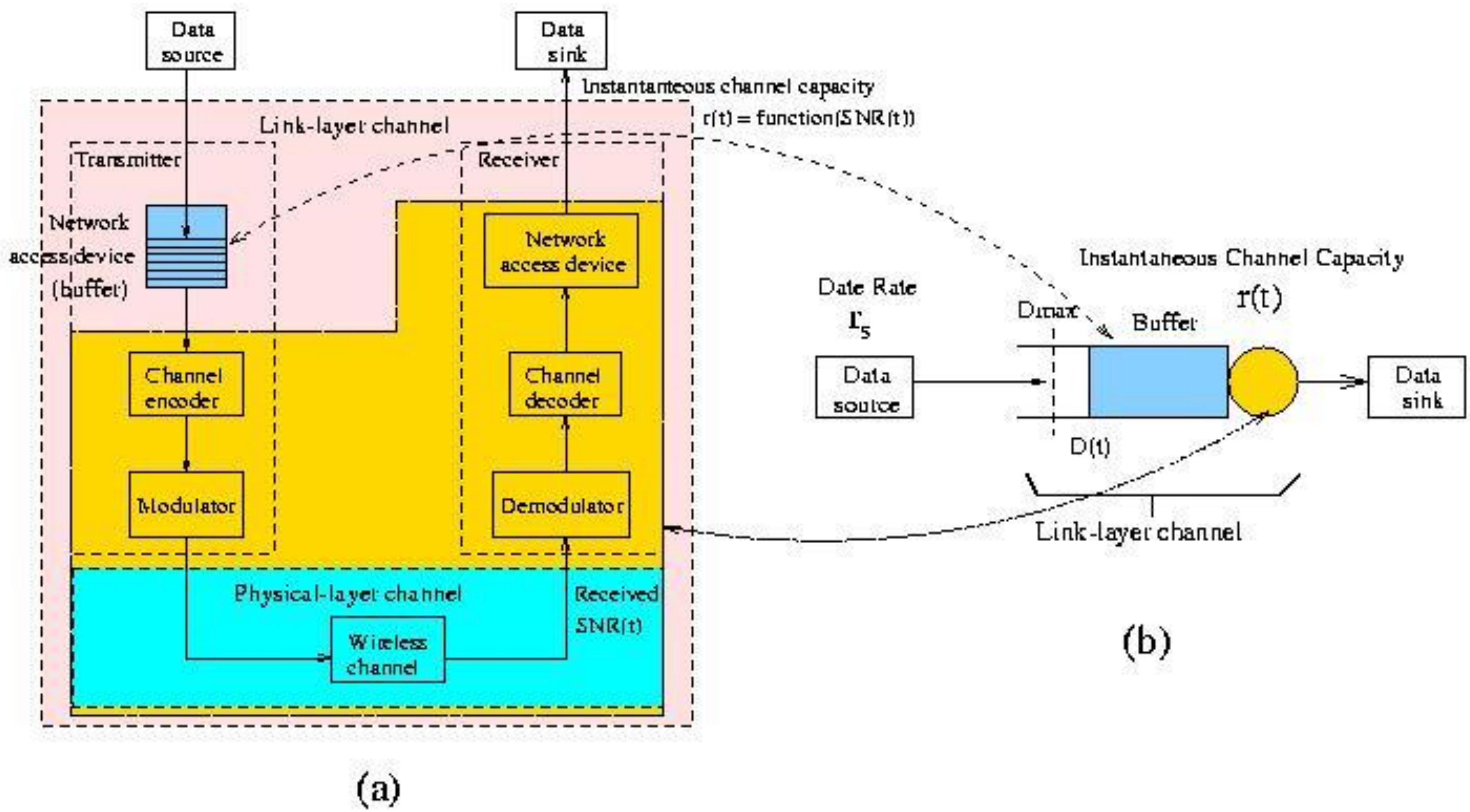
↑ simple control

How?

Model channel

at link layer

What is Link-layer Channel?



How to Model Link-layer Channel?

Given fading channel,
constant source rate r_s & D*max*

How?

$$P_D = \Pr\{D(\infty) \geq D_{\max}\}$$

How to Model Link-layer Channel (2)

- Assume
  - 'fluid model' of traffic
  - stationary, ergodic channel power gain *g(t)*
  - instantaneous channel capacity *r(t)* = function(*g(t)*)
  - constant source rate $r_s < E[r(t)]$ ➡ stable queue
- Question: given *r(t)*, $\Pr\{D(\infty) \geq D_{\max}\} = ?$
- Answer: can be found by 'large deviations theory'

Idea of Large Deviations Theory

Goal:

$$P_D = \Pr\{D(\infty) \geq D_{\max}\} \leq e^{-\theta \times D_{\max}}$$

Use Chernoff bound

Sufficient condition

effective capacity function of *r(t)* exists

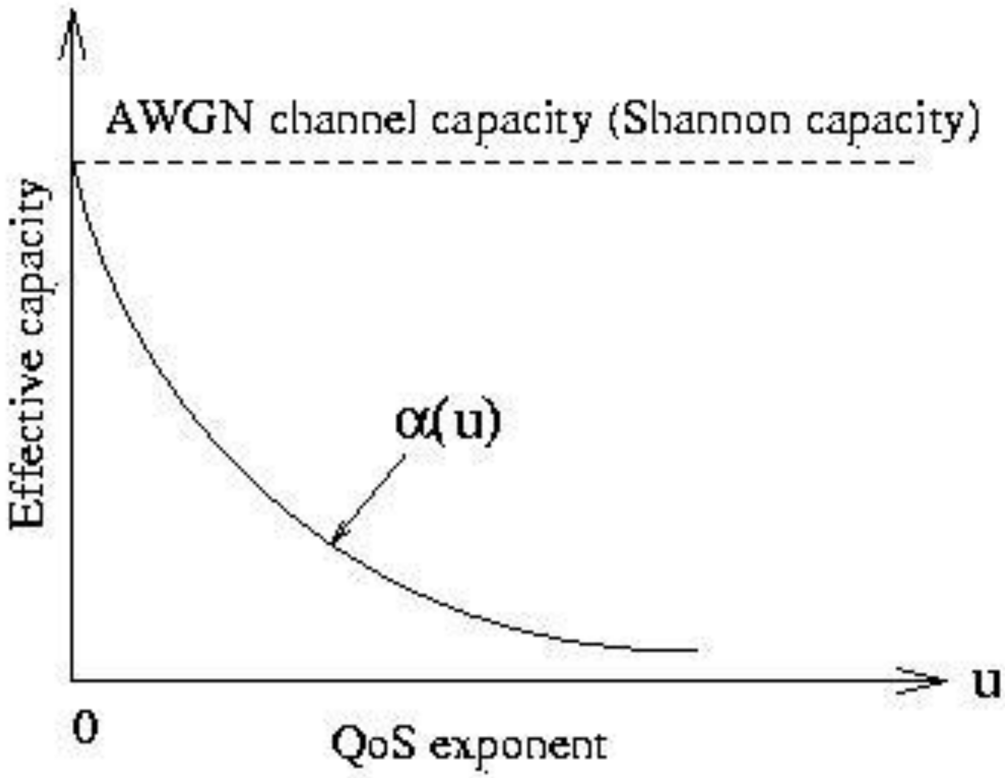
Effective capacity function of *r(t)*:

$$\alpha(u) = -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}], \quad \forall u > 0$$

Relation between QoS and Effective Capacity

$\alpha(u)$: maximum data rate achievable with P_D satisfied,

capacity of a delay-constrained channel



$$u = \frac{\log \dfrac{1}{P_D}}{D_{\max} r_s}$$

$$D_{\max} \downarrow \longrightarrow u \uparrow \longrightarrow \alpha(u) \downarrow$$

Our Link-layer Channel Model

- If $\alpha(u)$ exists (e.g., for stationary, Markovian *r(t)*), then

$$P_D = \Pr\{D(\infty) \geq D_{\max}\} \approx e^{-\theta(\mu)\times D_{\max}}, \text{ for large } D_{\max}$$

where $\theta(\mu) = \mu\alpha^{-1}(\mu)$,

$\alpha^{-1}(\cdot)$ is inverse function of $\alpha(u)$.

- $\theta(\mu)$ is our proposed EC channel model.

$$\boxed{r(t) \longrightarrow \theta(\mu) \longrightarrow P_D(D_{\max}) \approx e^{-\theta(\mu)\times D_{\max}}}$$

Closed form!

How to Derive QoS under EC Approach?

Given fading channel & constant source rate μ

↓

Channel measurement — measure queueing behavior $\Pr\{D(t) > 0\}$ & $E[D(t)]$

↓

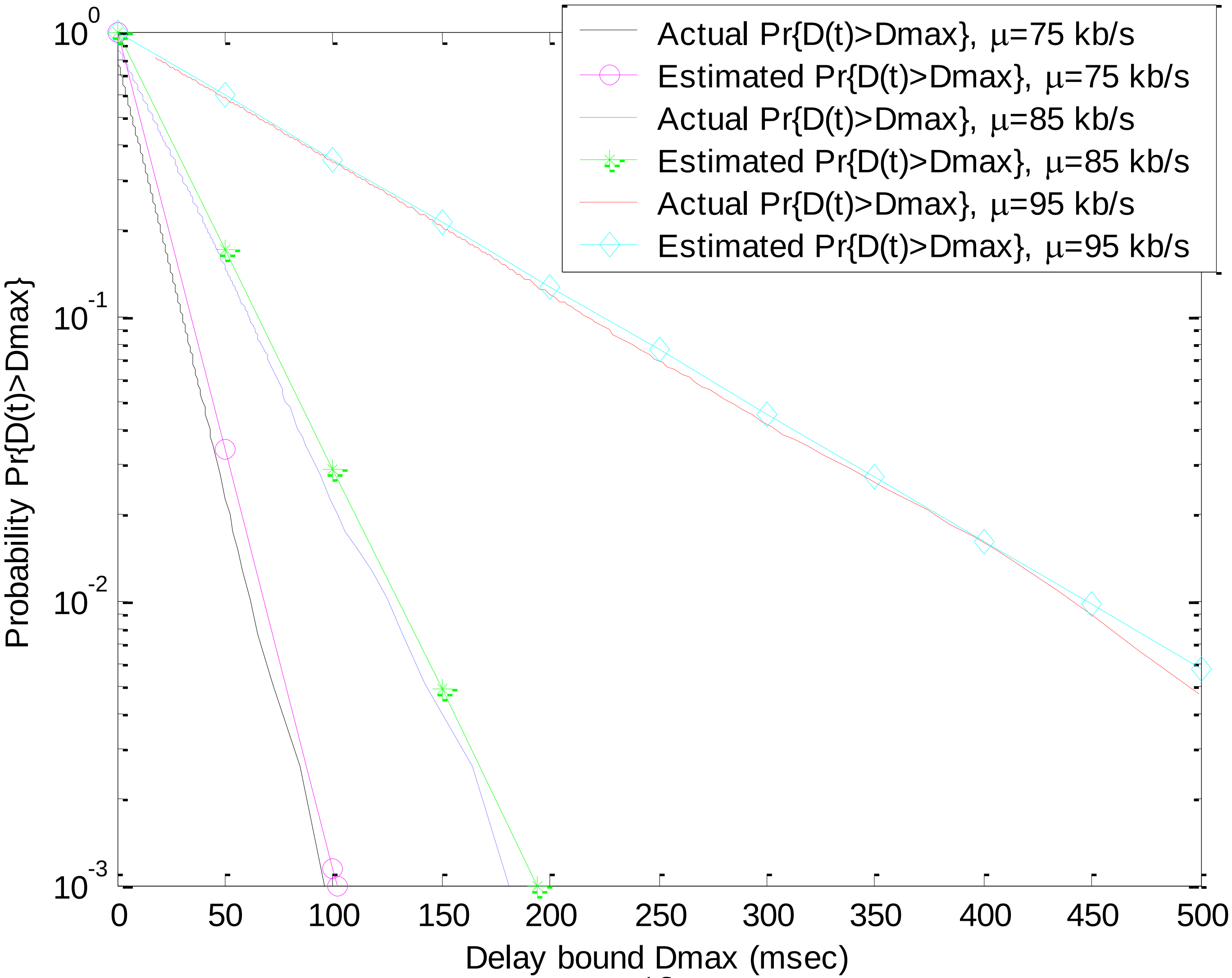
Channel modeling — $$\hat{\theta}(\mu) = \frac{\Pr\{D(t) > 0\}}{E[D(t)]}$$

↓

Deriving QoS measures — $$P_D = \Pr\{D(\infty) \geq D_{\max}\} \approx e^{-\hat{\theta}(\mu) \times D_{\max}}$$

Question: How good is the approximation?

Simulation Result (AR(1), Rayleigh Fading)



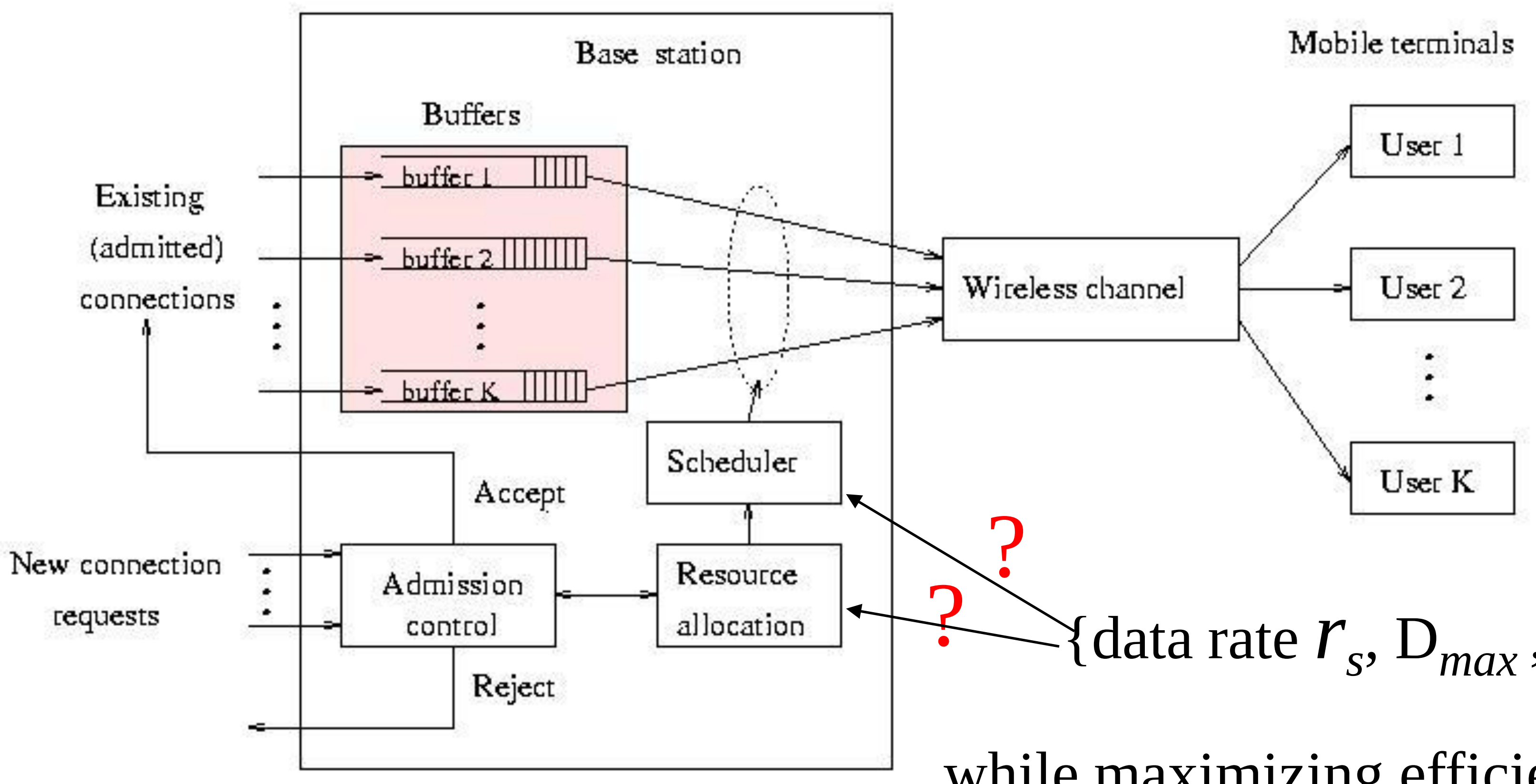
Summary of EC Channel Model

- EC channel model
  - simplify analysis
- How?
  - Directly measure queue and model link-layer channel
- Why?
  - Avoid the detail in physical layer
- Next question: Is QoS provisioning an easy task, given EC model?

Outline

- Motivation
- Our effective capacity (EC) approach
  - Link layer channel model: EC model
    - ➡ Derive QoS measures
  - Design of QoS provisioning, using EC model
- Summary
- On-going projects

Problem of QoS Provisioning



{data rate r_s, D_{max} , P_D}

while maximizing efficiency

Previous Scheduling Schemes

Knopp & Humblet's (K&H):
allot channel to the best user

↓

↑ Max total throughput

↓ Not suitable for delay-sensitive applications

Cont'd

Bettesh & Shamai's scheme:

K&H + Round Robin (RR)

↓

↑ Lower delay than K&H

↑ Higher throughput than RR

↓ No explicit QoS guarantee

If all users transmit at least once in the last *L* slots, use K&H; otherwise use RR.

Cont’d

Dynamic programming:

minimize channel usage, s.t. QoS

↑ Efficiency

↑ Explicit QoS guarantee

↓ Exponential complexity

Our Joint K&H/RR Scheduling

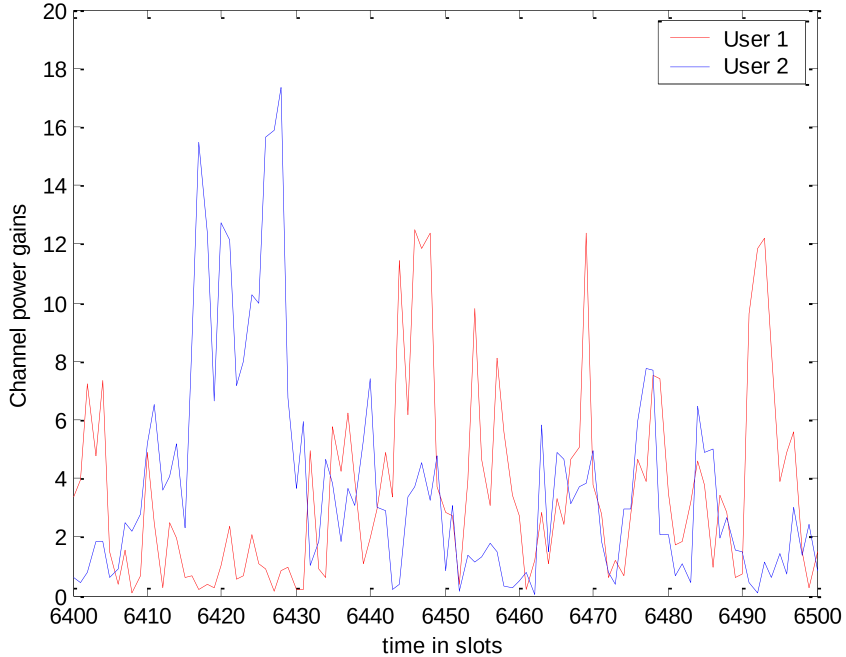
In each frame t:

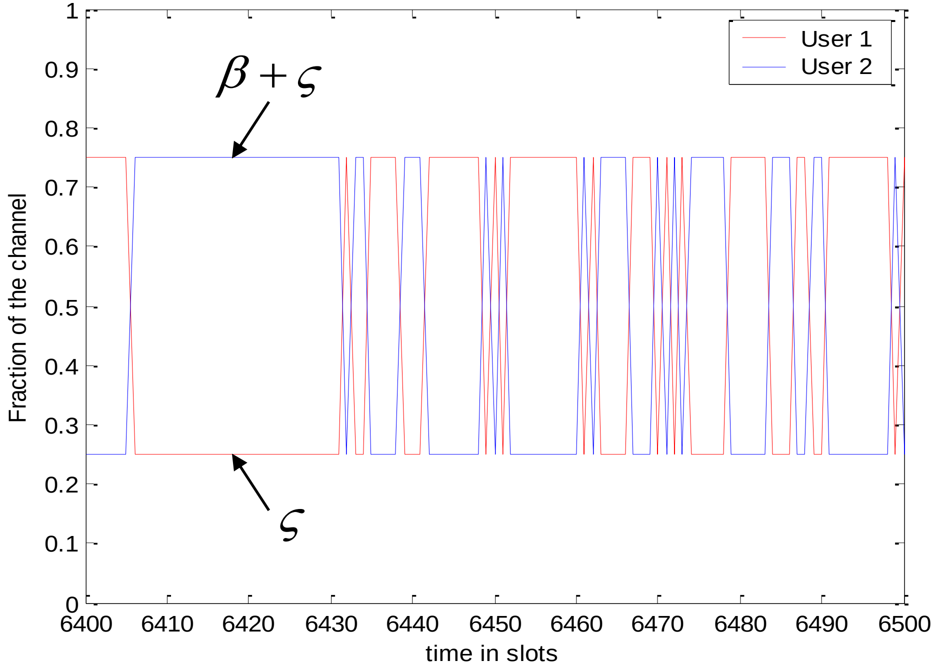
- K&H: β fraction of frame to the best user
- RR: ς fraction of frame to each user

$$\{r_s, D_{\max}, P_D\} \xrightarrow{?} \beta, \varsigma$$

Exponential complexity using traditional approach!

$$\beta = 0.5, \varsigma = 0.25$$





Admission Control & Resource Allocation

Given channel model $\theta_{\varsigma,\beta}(\mu)$,
K users' $\{r_s, D_{\max}, P_D\}$

$\rho = -\log P_D / D_{\max}$

$$\min_{\{\varsigma,\beta\}} \quad K\varsigma + \beta$$

subject to $\theta_{\varsigma,\beta}(r_s) \ge \rho,$

$K\varsigma + \beta \le 1,$

$\varsigma \ge 0, \quad \beta \ge 0.$

Complexity: O(N)

N: number of discrete values that ς or β can take.

Optimum $\{\varsigma, \beta\}$ used for resource allocation

Admission control:

If satisfied, accept the new user (K-th user); otherwise, reject it.

Procedure of QoS provisioning

Channel modeling

Estimate $\theta_{\varsigma,\beta}(\mu)$
for various values of $\{\varsigma, \beta\}$.

Admission control & resource allocation

Determine optimal $\{\varsigma, \beta\}$ that satisfies users' QoS, while minimizing frame usage.

Scheduling

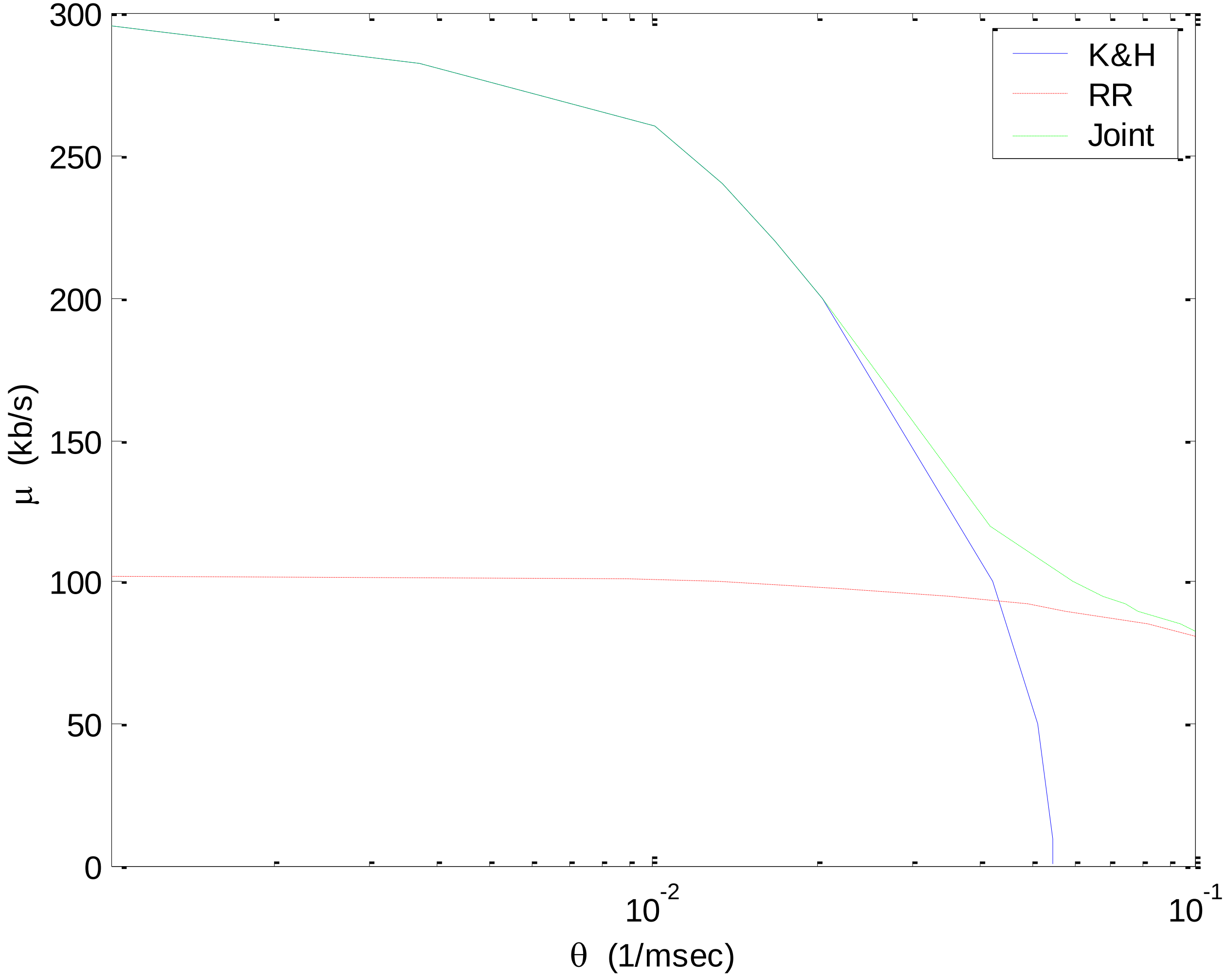
Provide K&H/RR with optimal $\{\varsigma, \beta\}$.

Given EC model, QoS provisioning is an easy task!

$$\because \{\varsigma, \beta\} \xleftrightarrow{\theta_{\varsigma,\beta}(r_s) = -\log P_D / D_{\max}} \{r_s, D_{\max}, P_D\}$$

control parameter QoS

Performance Gain of Scheduling



Conclusion

- Novel Effective Capacity (EC) approach
  - Model a channel via EC function
  - Design QoS provisioning via EC channel model
- EC approach: fundamentally different from traditional approach
  - Channel modeling

  Link-layer model vs. physical-layer model
  - QoS provisioning

  Control parameters $\xleftrightarrow[\text{vs. complex relation}]{\text{closed form}}$ QoS measure

Conclusion (2)

Traditional approach:

↓ complex analysis

↓ complex control

Why?

Model channel

at physical layer

Our approach:

↑ simple analysis

↑ simple control

Why?

Model channel

at link layer

Thank you!