Flow-Level Performance of Channel-Aware Scheduling Algorithms in Wireless Data Networks

Sem Borst

Bell Labs, Lucent Technologies
Murray Hill

CWI
Amsterdam

IMA Summer Program
June 27, 2005

Introduction

Wireless channels vary randomly across space and time

- **Channel conditions vary widely across spatially diverse users due to distance-related attenuation**
- **Channel quality for given user drastically fluctuates over time because of fading effects**

Introduction (cont'd)

Wireless voice systems rely on power control for adjusting transmit power to combat fading, and maintain constant transmission rate

Wireless data networks typically operate at fixed transmit power, and employ rate control for dynamically adapting transmission rate over time in response to fading

Consequently, fluctuations in channel conditions translate into variations in (feasible) transmission rates

Introduction (cont'd)

Relative delay tolerance of data applications opens up possibility to exploit rate variations, and schedule transmissions when channel conditions of users are relatively favorable

Performance of channel-aware scheduling policies has mostly been analyzed at packet level for static user population

Static population is reasonable modeling convention in view of separation of time scales

It is not satisfactory however to assume that user population is independent of properties of scheduling algorithm

Introduction (cont'd)

Need to move away from static scenario in order to capture interdependence between scheduling algorithm and user population, and to evaluate performance at flow level

Consider dynamic setting where elastic flows come and go over time as governed by arrival and completion of random service demands

Model description

For now, consider static scenario with M data users served by single base station (BS)

BS transmits in time slots of some fixed duration τ (typical value is $\tau = 1.67$ ms, or 600 slots per second)

In each time slot, BS transmits to exactly one of the users

Feasible rates for various users vary over time according to stationary ergodic process $(R_1(t), \ldots, R_M(t))$, with $R_i(t)$ representing feasible rate for user i in time slot t

Homogeneous rate characteristics

First suppose that $R_1(t), \ldots, R_M(t)$ are stationary i.i.d. (independence across users, not time) with mean C

Consider scheduling strategy S^* which in each time slot selects user with largest instantaneous rate

Thus, in time slot t strategy S^* selects user i^* identified as

$$i^* = \arg \max_{j=1,\ldots,M} R_j(t)$$

Homogeneous rate characteristics (cont'd)

By symmetry, each user receives fraction $1/M$ of time slots

When selected, expected rate of user is

$$\mathbb{E}\{R_i | R_i = \max_{j=1,\ldots,M} R_j\} = \mathbb{E}\{\max_{j=1,\ldots,M} R_j\}$$

Thus, each user receives average service rate $CG(M)/M$, with $G(M) := \mathbb{E}\{\max_{j=1,\ldots,M} R_j\}/C$

$G(M)$ may be interpreted as gain factor (over channel-oblivious round-robin discipline)

For example, if $R_1, \ldots, R_M$ are independent and exponentially distributed ("Rayleigh fading"), then

$$G(M) = 1 + 1/2 + \ldots + 1/M \approx \log(M)$$

Heterogeneous rate characteristics

Now suppose that $R_i(t) \stackrel{d}{=} C_i Y_i(t)$, where $C_i = \mathbb{E}\{R_i\}$ is time-average rate of user i, and $Y_1(t), \ldots, Y_M(t)$ are stationary i.i.d. (independence across users, not time) with unit mean

Consider scheduling strategy S^* which in each time slot selects user with largest instantaneous rate relative to its time-average rate

Thus, in time slot t strategy S^* selects user i^* identified as

$$i^* = \arg \max_{j=1,\ldots,M} \frac{R_j(t)}{C_j} = \arg \max_{j=1,\ldots,M} Y_j(t)$$

Heterogeneous rate characteristics (cont'd)

By symmetry, each user receives fraction $1/M$ of time slots

When selected, expected rate of user i is

$$\mathbb{E}\{R_i | \frac{R_i}{C_i} = \max_{j=1,\ldots,M} \frac{R_j}{C_j}\} =$$

$$\mathbb{E}\{C_i Y_i | Y_i = \max_{j=1,\ldots,M} Y_j\} = C_i \mathbb{E}\{\max_{j=1,\ldots,M} Y_j\} = C_i G(M),$$

with $G(M)$ gain factor introduced before

Thus, user i receives average service rate $C_i G(M)/M$

As before, each user is served at fraction $G(M)/M$ of its time-average rate

Weight-based scheduling

In time slot t, 'weight-based' strategy selects user i^* identified as

$$i^* = \arg \max_{j=1,\ldots,M} w_j R_j(t)$$

for given weights $w_1, \ldots, w_M$

Strategy S^* is special case with weights $w_i = 1/C_i$

By construction, weight-based strategies maximize weighted throughput combinations and thus produce Pareto-optimal throughput vectors (in fact, sample-path wise)

Proportional Fair scheduling (Qualcomm/Tse)

Proportional Fair (PF) strategy assigns adaptive weights $w_i(t) = \frac{1}{S_i(t)}$, i.e., in time slot t it selects user i^* identified as

$$i^* = \arg \max_{j=1,\ldots,M} \frac{R_j(t)}{S_j(t)}$$

$S_i(t)$ is geometrically smoothed throughput of user i at time t, updated in each time slot according to

$$S_i(t+1) = (1-\delta)S_i(t) + \delta X_i(t) R_i(t),$$

with δ smoothing parameter and $X_i(t)$ 0–1 variable indicating whether user is selected in time slot t or not

$T = 1/\delta$ is time constant (typical value is $T = 1000$)

Thus, PF strategy selects user with largest instantaneous rate relative to its long-term average throughput

Proportional fair scheduling (cont'd)

Now observe that both instantaneous rates R_i and geometrically smoothed throughputs S_i scale linearly with time-average rates C_i

Consequently, allocation of time slots does not depend on time-average rates $C_1, \ldots, C_M$, but only on relative rate fluctuations $Y_1, \ldots, Y_M$

Because of symmetry, we thus have $S_i \stackrel{d}{=} C_i U_i$, where $U_1, \ldots, U_M$ are identically distributed (but not independent)

In particular, each user receives fraction $1/M$ of time slots

Proportional Fair scheduling (cont'd)

When selected, expected rate of user i is

$$\mathbb{E}\{R_i | \frac{R_i}{S_i} = \max_{j=1,\dots,M} \frac{R_j}{S_j}\}$$

When time constant $T = 1/\delta$ is 'large', geometrically smoothed throughputs $S_1, \dots, S_M$ will be 'nearly constant' and not show 'any significant variation'

Informally, $(S_1, \dots, S_M) \to U_0(C_1, \dots, C_M)$ for some constant U_0 as $\delta \downarrow 0$

Proportional Fair scheduling (cont'd)

Thus, when δ is small, expected rate of user i when selected, approximately equals

$$\mathbb{E}\{R_i | \frac{R_i}{C_i U_0} = \max_{j=1,\ldots,M} \frac{R_j}{C_j U_0}\} =$$

$$\mathbb{E}\{C_i Y_i | Y_i = \max_{j=1,\ldots,M} Y_j\} = C_i \mathbb{E}\{\max_{j=1,\ldots,M} Y_j\} = C_i G(M),$$

with $G(M)$ gain factor introduced before

PF strategy roughly behaves like strategy S^* as $\delta \downarrow 0$

Utility-based scheduling

Utility-based strategy assigns adaptive weights $w_i = U_i'(S_i(t))$, i.e., in time slot t it selects user i^* identified as

$$i^* = \arg \max_{j=1,\ldots,M} U_j'(S_j(t)) R_j(t),$$

with $U_i(\cdot)$ strictly concave utility function of user i

$S_i(t)$ is geometrically smoothed throughput of user i at time t, updated in each time slot according to

$$S_i(t+1) = (1-\delta) S_i(t) + \delta X_i(t) R_i(t)$$

PF strategy is special case with utility function $U_i(\cdot) = \log(\cdot)$

Utility-based scheduling (cont'd)

'Theorem'
(Agrawal & Subramanian, Kushner & Whiting, Stolyar)

Under 'mild' assumptions, $(S_1, \ldots, S_M) \to (s_1^*, \ldots, s_M^*)$ **as** $\delta \downarrow 0$**,**
where $(s_1^*, \ldots, s_M^*)$

- **maximizes aggregate utility, i.e.,**

$$\sum_{i=1}^{M} U_i(s_i^*) \geq \sum_{i=1}^{M} U_i(s_i)$$

  for every achievable throughput vector $(s_1, \ldots, s_M)$

- **satisfies fixed-point equations**

$$s_i^* = \mathbb{E}\{R_i \mathrm{I}_{\{U_i'(s_i^*) R_i = \max_{j=1,\ldots,M} U_j'(s_j^*) R_j\}}\}$$

  for all $i = 1, \ldots, M$

Utility-based scheduling (cont'd)

In case of PF strategy, fixed-point equations reduce to

$$s_i^* = \mathbb{E}\{R_i \mathrm{I}_{\{\frac{R_i}{s_i^*} = \max\limits_{j=1,\dots,M} \frac{R_j}{s_j^*}\}}\}$$

In case $R_i(t) \stackrel{d}{=} C_i Y_i(t)$, **above equations further simplify to**

$$s_i^* = \mathbb{E}\{C_i Y_i \mathrm{I}_{\{\frac{C_i Y_i}{s_i^*} = \max\limits_{j=1,\dots,M} \frac{C_j Y_j}{s_j^*}\}}\},$$

or equivalently,

$$\frac{1}{\sigma_i} = \mathbb{E}\{Y_i \mathrm{I}_{\{\sigma_i Y_i = \max\limits_{j=1,\dots,M} \sigma_j Y_j\}}\},$$

with $\sigma_i := C_i / s_i^*$

By symmetry, $\sigma_i \equiv \sigma$ **for all** $i = 1, \dots, M$, **and thus** $\sigma = M/G(M)$, **yielding** $s_i^* = C_i G(M)/M$ **as before**

Dynamic user configuration

User dynamics governed by finite-size service demands that arrive randomly over time

Assume that duration of time slots is relatively short compared to size and arrival frequency of service demands

Scheduling strategy operates on extremely fast time scale compared to user dynamics

Natural to analyze flow-level performance in continuous time rather than discrete time, and assume that users are served simultaneously rather than in time-slotted fashion

Dynamic user configuration (cont'd)

Continuous-time model naturally inherits service characteristics from discrete-time model

Specifically, instantaneous service rate vector in continuous-time context coincides with long-term throughput vector in discrete-time setting for corresponding user population

In particular, under strategy S^*, when there are n active users, each of them is served at fraction $G(n)/n$ of its time-average rate

Gives rise to Processor-Sharing model with state-dependent service rate function $G(n)$ when there are n active users

Traffic model

Class-k users submit file transfer requests as Poisson process of rate λ_k

At most M users are admitted into system simultaneously (possibly $M = \infty$)

Let (C_k, F_k) be pair of random variables with as distribution joint distribution of time-average transmission rate and file size of arbitrary class-k user

Let $B_k := F_k/C_k$ be normalized service requirement of class-k user, with mean $\beta_k := \mathbb{E}\{F_k/C_k\}$

Traffic model (cont'd)

Offered traffic associated with class-k users is $\rho_k := \lambda_k \beta_k$

Total amount of offered traffic is $\rho := \sum_{k=1}^{K} \rho_k$

Define $G^* := \lim_{M \to \infty} G(M) = \sup_{M=1,2,\ldots} G(M)$**, with**

$$G(M) = \mathbb{E}\{\max_{j=1,\ldots,M} Y_j\}$$

gain factor introduced before

Let $(N_1, \ldots, N_K)$ be random vector representing number of users of various classes in system under strategy S^* at arbitrary epoch in statistical equilibrium

Proposition 1

Strategy S^* achieves stability for $\rho < G^*$ or $M < \infty$, in which case (Cohen, Kelly)

$$\mathbb{P}\{(N_1, \ldots, N_K) = (n_1, \ldots n_K)\} = H^{-1} \frac{n! \rho^n}{\phi(n)} \prod_{k=1}^{K} \frac{1}{n_k!} \left(\frac{\rho_k}{\rho}\right)^{n_k},$$

with $n = n_1 + \ldots + n_K \leq M$, $\phi(n) := \prod_{i=1}^{n} G(i)$,
and normalization constant

$$H := \sum_{n=0}^{M} \frac{\rho^n}{\phi(n)}$$

In particular,

$$\mathbb{P}\{N = n\} = H^{-1}\frac{\rho^n}{\phi(n)},$$

$$\mathbb{E}\{N\} = H^{-1}\sum_{n=1}^{M}\frac{n\rho^n}{\phi(n)},$$

and

$$\mathbb{E}\{N_k\} = \frac{\rho_k}{\rho}\mathbb{E}\{N\}$$

Blocking probability is given by

$$L = \mathbb{P}\{N = M\}$$

Mean transfer delay experienced by class-k user is

$$\mathbb{E}\{S_k\} = \frac{\beta_k}{\rho(1-L)}\mathbb{E}\{N\}$$

Reflects insensitivity of Processor-Sharing discipline:
mean delay of class-k user only depends on service requirement distribution of class k through its mean β_k

In fact, it may be shown that conditional expected delay of any user with actual service requirement b is

$$\mathbb{E}\{S|B=b\} = \frac{b}{\rho(1-L)}\mathbb{E}\{N\}$$

Thus, expected transfer delay incurred by user is proportional to its normalized service requirement, with factor of proportionality $\mathbb{E}\{N\}/(\rho(1-L))$

Latter property embodies certain fairness principle:
users with larger service requests experience longer delays

Proposition 2

No strategy achieves stability for $\rho > G^*$

Proof arguments

Strategy S^* reduces (normalized) amount of work at maximum possible rate $G(M)$ when there are M users

Amount of work cannot be reduced at higher rate than G^*

Above two propositions combined imply that strategy S^* achieves stability whenever feasible

Rate at which amount of work is reduced, approaches maximum value as number of users tends to infinity

Unless weights are set inversely proportional to time-average rates, relative variations are not maximally exploited

Example

Consider two $K = 2$ classes

Assume that rate variations have bounded support, and consider weight-based strategy such that

$$\mathbb{P}\{w_1 C_1 Y_1 > w_2 C_2 Y_2\} = 1$$

Thus service of class 1 takes precedence over that of class 2

There are scheduling gains within both classes, but not between classes

Example (cont'd)

According to Proposition 1, class 1 is stable as long as $\rho_1 < G^*$**, in which case**

$$\pi_0 = \mathbb{P}\{N_1 = 0\} = \left[\sum_{n=0}^{\infty} \frac{\rho^n}{\phi(n)}\right]^{-1},$$

with $\phi(n) = \prod_{i=1}^{n} G(i)$

Class 2 is stable if in addition $\rho_2 < \pi_0 G^*$

Now observe that $\pi_0 < 1 - \rho_1/G^* - \epsilon$ **for some** $\epsilon > 0$

Hence class 2 is stable only if $\rho_2 < (1-\epsilon)G^* - \rho_1$**, i.e.,** $\rho < (1-\epsilon)G^*$**, which is strictly stronger than** $\rho < G^*$

Extensions

In general, weight-based scheduling strategies give rise to Discriminatory Processor Sharing like models

Minimization of mean transfer delays

Mixtures of elastic users and streaming users

Asymmetric rate variations

Impact of slow fading

Further stability results (work in progress)

Now suppose that $R_{ki}(t) \stackrel{d}{=} C_{ki}Y_{ki}(t)$**, where** $C_{ki} = \mathbb{E}\{R_{ki}\}$ **is time-average rate of** i**-th class-**k **user, and** $Y_{k1}(t), \ldots, Y_{kM_k}(t)$ **are stationary i.i.d. for given** k **with unit mean**

Gain factor associated with class k **is**

$$G_k(M) := \mathbb{E}\{\max_{j=1,\ldots,M} Y_{kj}\}$$

Define $G_k^* := \lim_{M\to\infty} G_k(M) = \sup_{M=1,2,\ldots} G_k(M)$

Proposition 3

No strategy achieves stability for $\sum_{k=1}^{K} \frac{\rho_k}{G_k^*} > 1$

Proof arguments

Normalized amount of class-k **work cannot be reduced at higher rate than** G_k^*

Minimum fraction of time spent serving class k **is** $\frac{\rho_k}{G_k^*}$

Proposition 4

PF strategy achieves stability for $\sum_{k=1}^{K} \frac{\rho_k}{G_k^*} < 1$

Proof arguments (sketch)

Consider Lyapunov function (Bonald & Massoulié, De Veciana et al., Ye et al.)

$$H(n_1, \ldots, n_K) = \sum_{k=1}^{K} \frac{n_k^2}{2\lambda_k}$$

Denote by s_k^{PF} **aggregate normalized class-**k **throughput under PF strategy**

Need to show that for some $\epsilon > 0$

$$\sum_{k=1}^{K} n_k (1 - \frac{s_k^{PF}}{\rho_k}) \leq -\epsilon \sum_{k=1}^{K} n_k$$

for $\sum_{k=1}^{K} n_k$ **sufficiently large**

Further stability results (cont'd)

By assumption, $\sum_{k=1}^{K} \frac{\rho_k}{G_k^*} \leq \frac{1}{1+\epsilon}$ **for** $\epsilon > 0$ **sufficiently small**

Since s_k^{PF} **maximize** $\sum_{k=1}^{K} n_k \log(s_k/n_k)$**, it may then be shown from strict concavity that**

$$\sum_{k=1}^{K} \frac{n_k}{\rho_k}(\rho_k(1+\epsilon) - s_k^{PF}) < 0$$

for $\sum_{k=1}^{K} n_k$ **sufficiently large**

Thus,

$$\sum_{k=1}^{K} n_k(1 - \frac{s_k^{PF}}{\rho_k}) \leq -\epsilon \sum_{k=1}^{K} n_k$$

for $\sum_{k=1}^{K} n_k$ **sufficiently large**

Numerical experiments

Users initiate file transfer requests as Poisson process

Mean file size is assumed to be 60 Kbytes (480 Kbits)

At most $M = 20$ users are admitted into system simultaneously

Users which generate transfer requests when there are already M flows in progress, are blocked

System operates in time-slotted fashion, with slot duration of $\tau = 1.67$ ms (600 slots per second)

Users have independent "Rayleigh" fading channels

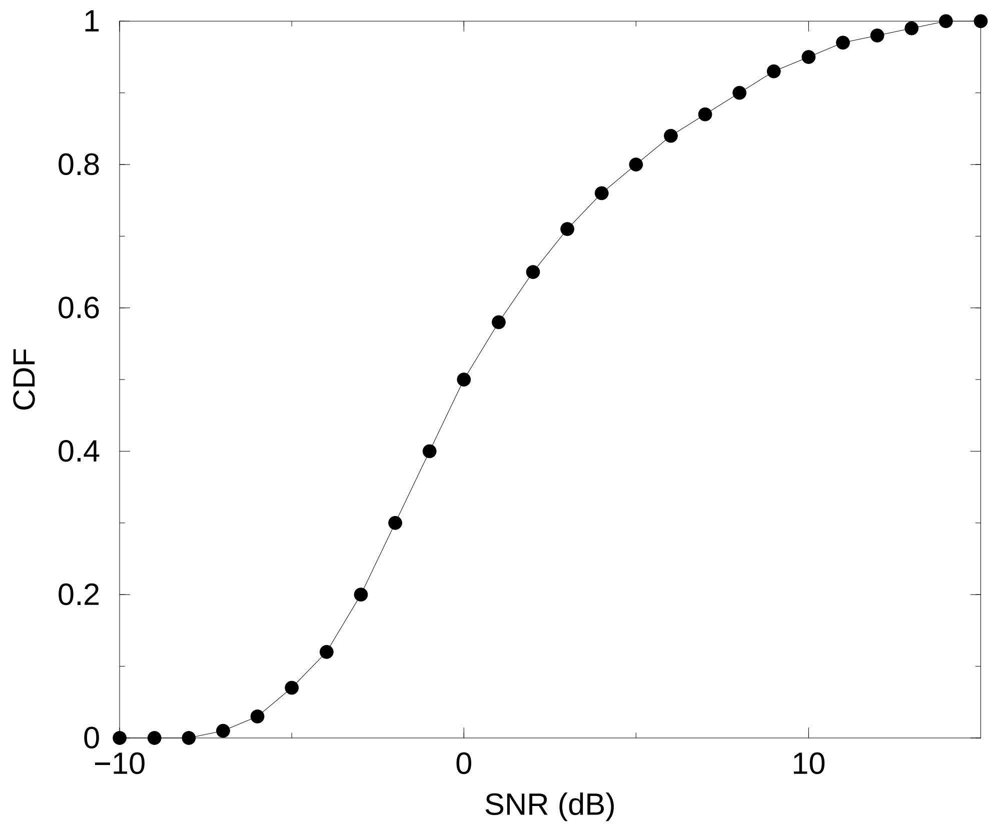
SNR distribution

Three different scenarios for distribution of mean SNR:

(I) identical to 0 dB for all users;

(II) bi-modal distribution, either -2.0 dB or 4.0 dB with equal probability;

(III) linearized version of typical SNR distribution in CDMA system plotted on next slide



Typical SNR distribution in CDMA system

Rate-SNR relationship

Three different scenarios for how instantaneous rate of user varies with instantaneous SNR value:

(A) instantaneous rate is linear in instantaneous SNR:
$R = C_1 \times SNR$, **with** $C_1 = 400$ **Kbs;**

(B) instantaneous rate is logarithmic in instantaneous SNR:
$R = C_2 \times \log(1 + SNR)$, **with** $C_2 = 800$ **Kbs;**

(C) instantaneous rate is determined from instantaneous SNR value (in dB) according to table on next slide

SNR (dB) $\geq$	Rate (Kbs)
- 12.5	38.4
- 9.5	76.8
- 8.5	102.6
- 6.5	153.6
- 5.7	204.8
- 4.0	307.2
- 1.0	614.4
1.3	921.6
3.0	1228.8
7.2	1843.2
9.5	2457.6

Rate (Kbs) as function of SNR (dB) in 1xEV-DO system

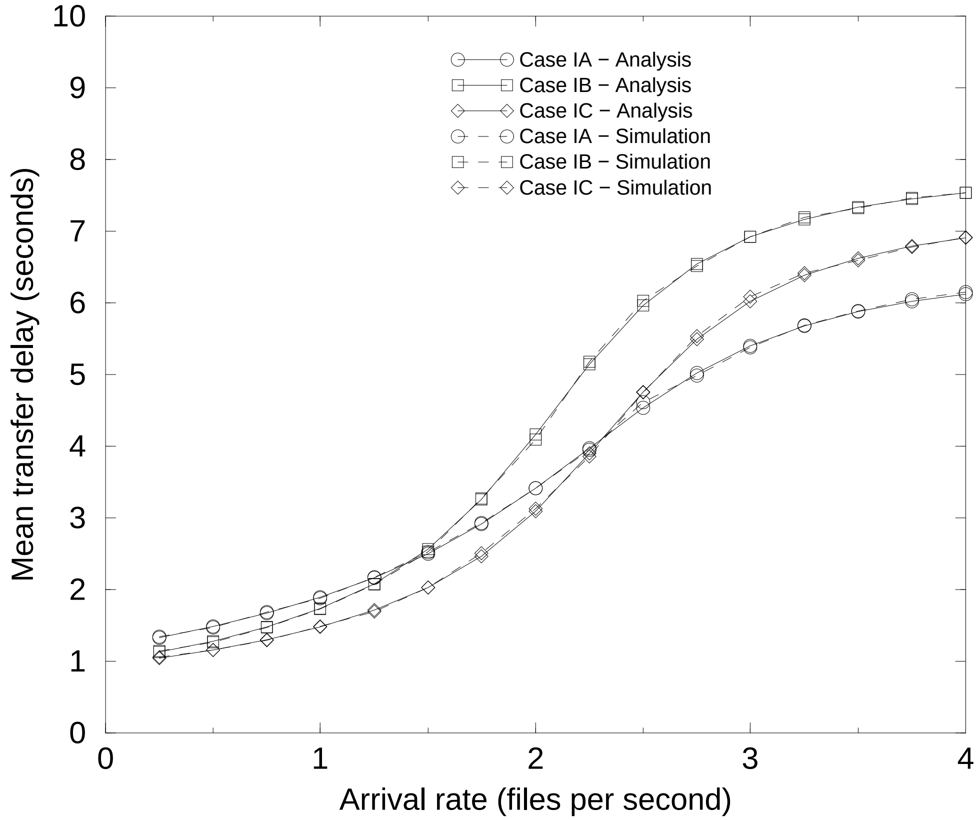
First set of experiments determine various performance metrics under strategy S^* for varying arrival rates

Consider total of nine cases obtained via pairwise combination of above scenarios for mean SNR distribution and rate variations

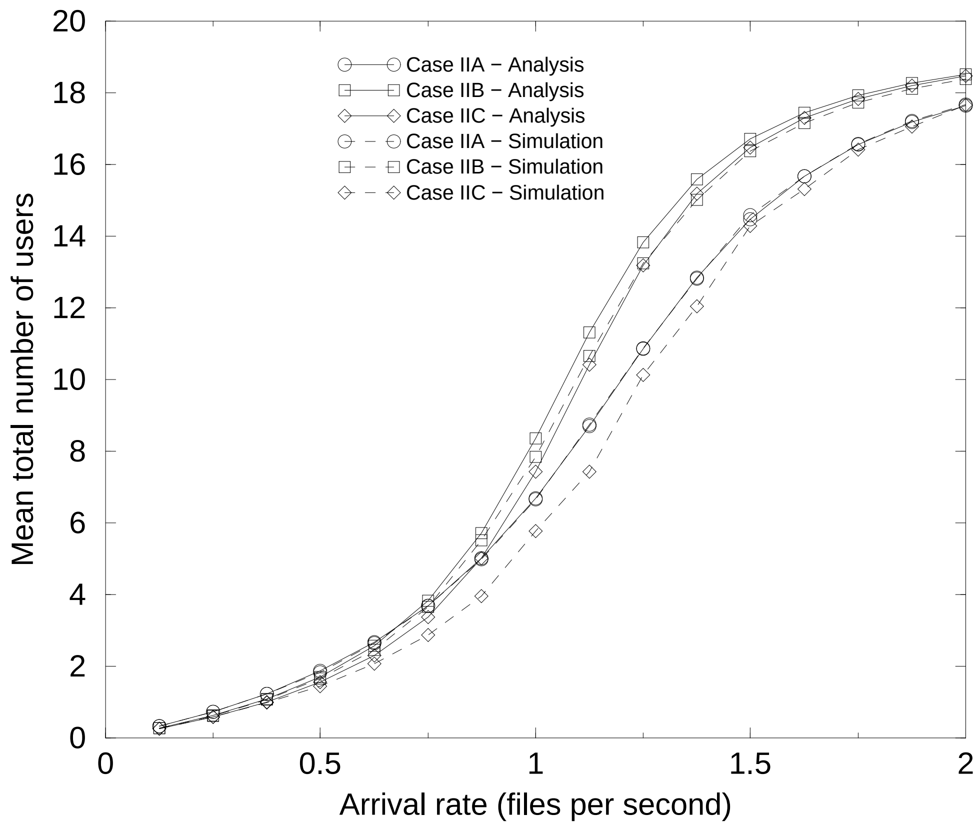
Relative rate fluctuations are only statistically identical in cases IA, IB, IC, IIA, IIIA

In remaining four cases, notion of gain factor $G(n)$ does not strictly apply

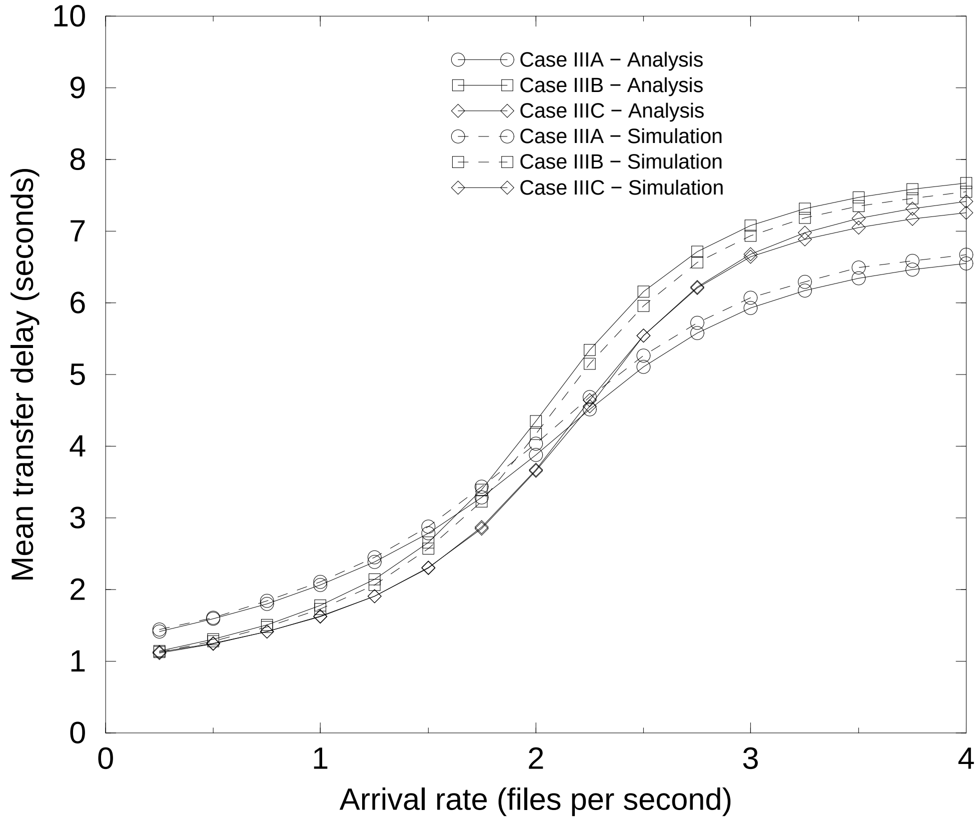
Used approximate gain factor by considering average SNR



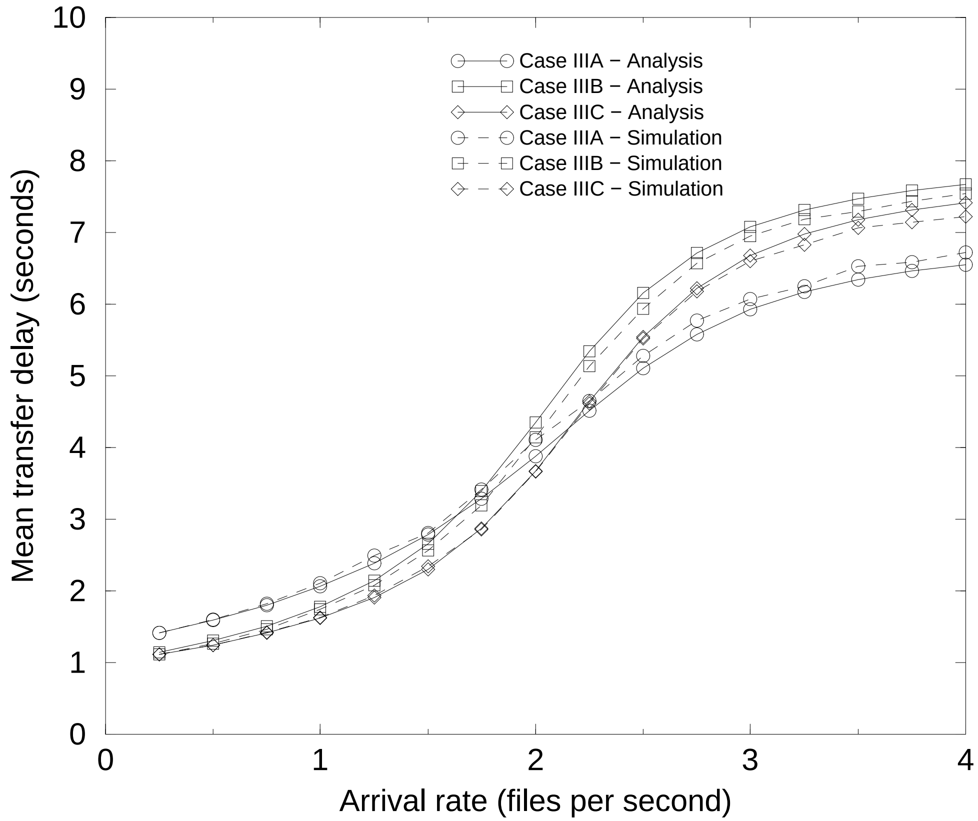
Mean transfer delay as function of file arrival rate for Cases IA-C



Mean total number of active users as function of file arrival rate for Cases IIA-C



Mean transfer delay as function of file arrival rate for Cases IIIA-C with deterministic file size

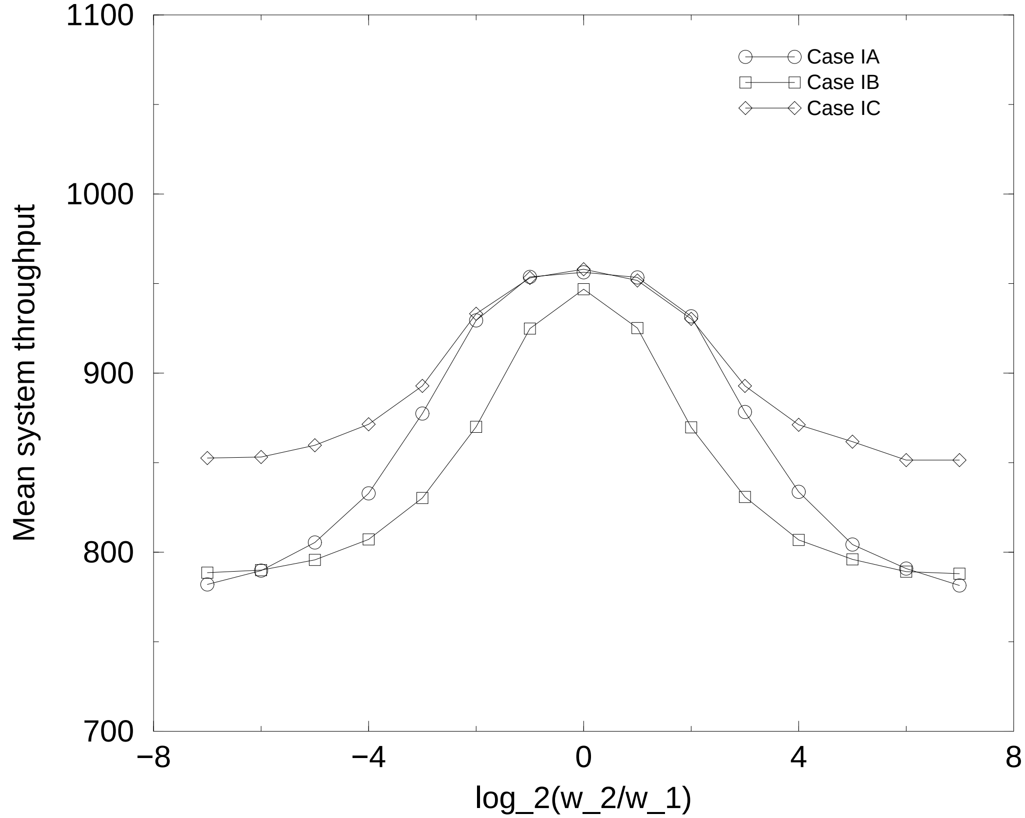


Mean transfer delay as function of file arrival rate for Cases IIIA-C with exponentially distributed file sizes

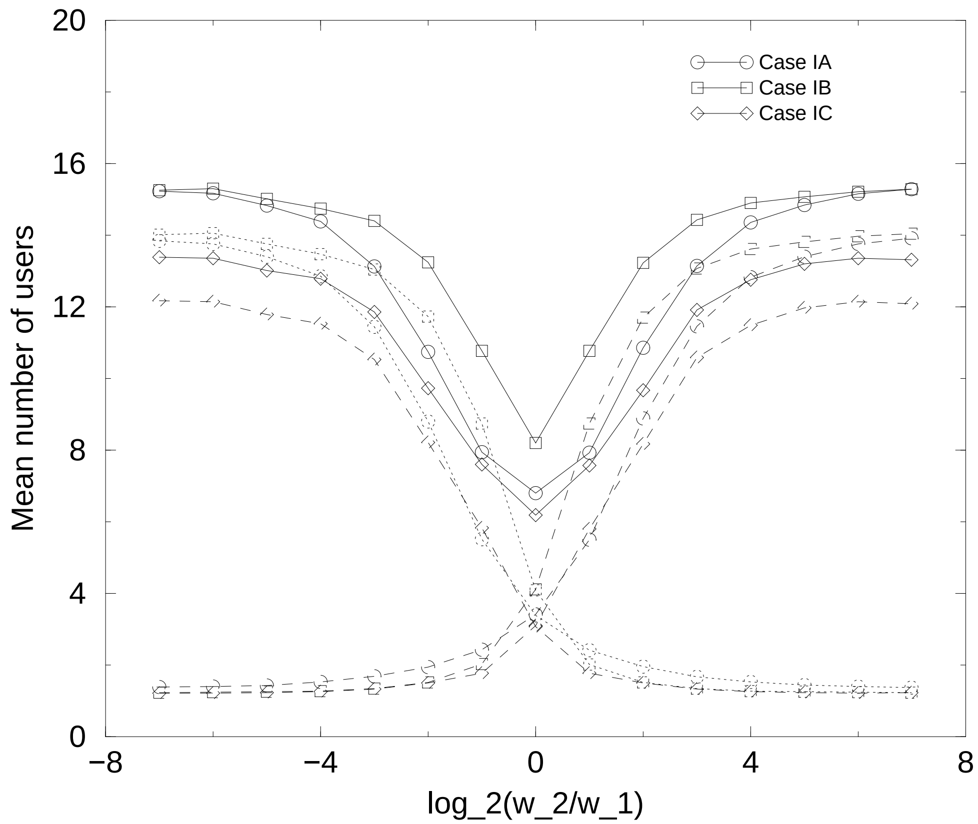
Second set of experiments evaluate similar performance metrics for varying weights used in allocation of time slots

Consider total of six cases obtained via pairwise combination of channel scenarios as before

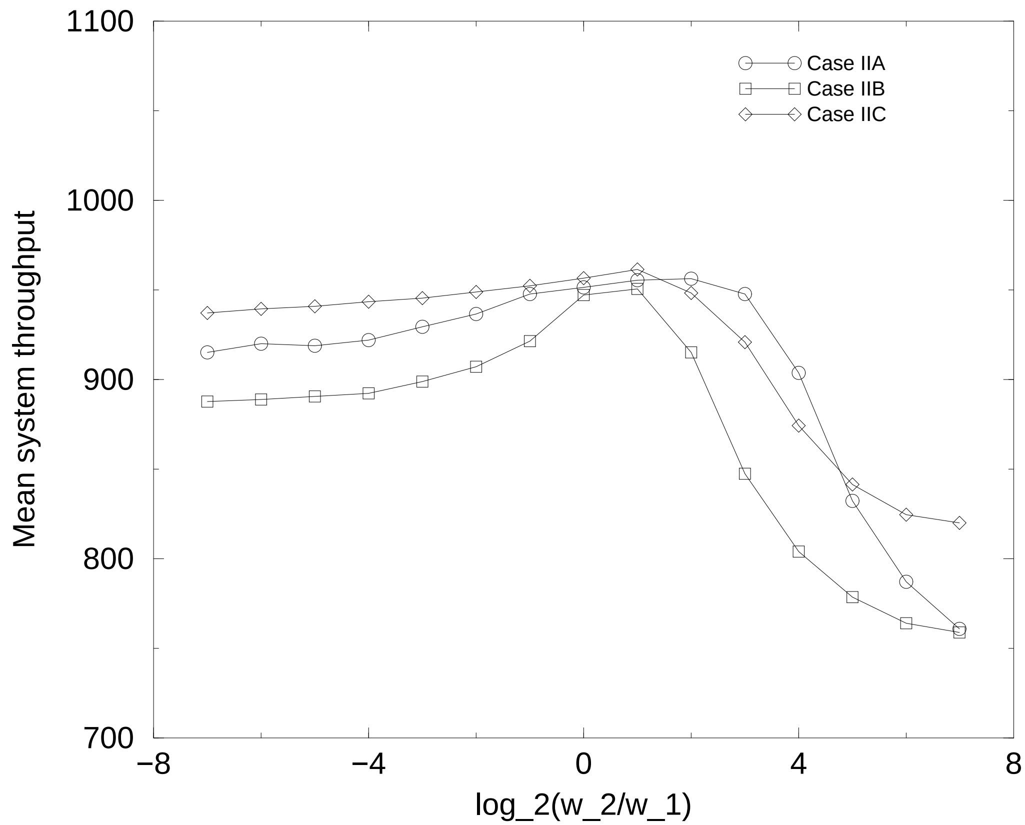
Focus on system with two user classes in order to investigate impact of weight factors



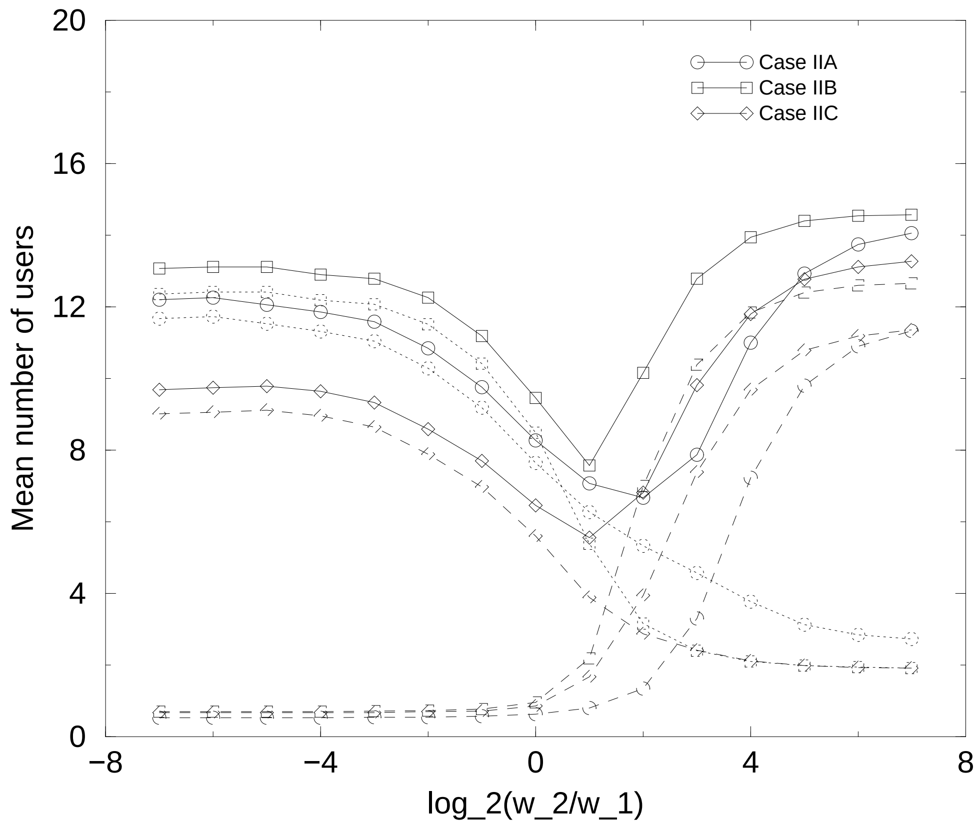
Mean system throughput as function of $\log_2(w_2/w_1)$ **for Cases IA-C**



Mean total number of users (solid line), class-1 users (dashed line), and class-2 users (dotted line) as function of $\log_2(w_2/w_1)$ for cases IA-C



Mean system throughput as function of $\log_2(w_2/w_1)$ for Cases IIA-C



Mean total number of users (solid line), class-1 users (dashed line), and class-2 users (dotted line) as function of $\log_2(w_2/w_1)$ for cases IIA-C

Slow fading

So far we assumed that service rate of user i at time t behaves as

$$C_i G(N(t))/N(t)$$

when there are $N(t)$ active users at time t

Now suppose service rate of user i at time t behaves as

$$C_i(t) G(N(t))/N(t)$$

with $C_i(t)$ some exogenous time-varying process

For convenience, assume $G(n) \equiv 1$ for all n

If $C_i(t) \equiv C(t)$ for all users i, then model reduces to Processor-Sharing system with (common) time-varying service rate

Slow fading (cont'd)

Class-k users submit file transfer requests as Poisson process of rate λ_k

Let F_k be random variable representing file size of arbitrary class-k user

Feasible transmission rate of i-th class-k user varies over time according to $C_{ki}(t)$

$C_{ki}(t)$, $i = 1, 2, \ldots$ are i.i.d. copies of stationary ergodic process $C_k(t)$

System is stable if $\sum_{k=1}^{K} \lambda_k \mathbb{E}\{F_k\}/\mathbb{E}\{C_k(0)\} < 1$

Slow fading (cont'd)

Consider two limiting regimes

- **Fluid regime:** $C_{ki}(t) = \mathbb{E}\{C_k(0)\}$ **for all** t **(variations occur on extremely fast time scale)**

- **Quasi-stationary regime:** $C_{ki}(t) = C_k(0)$ **for all** t **(variations occur on extremely slow time scale):**

Both regimes yield tractable Processor-Sharing models and simple explicit performance estimates

Slow fading (cont'd)

In fact, it may be shown that fluid and quasi-stationary regimes provide optimistic and conservative performance bounds, respectively

More generally, consider sequence of models indexed by parameter s where time is accelerated by factor s, i.e., $C_{ki}^{(s)}(t) := C_{ki}(st)$.

Then it may be shown that the performance improves monotonically in s

Note that limiting cases $s = \infty$ and $s = 0$ correspond to fluid and quasi-stationary regimes, respectively

Intra- & inter-cell mobility

Consider network with L base stations offered traffic from K user classes

Class-k users arrive as Poisson process of rate λ_k (per sec) and have mean service requirement σ_k (in bits)

Offered traffic from class k is $\rho_k := \lambda_k \sigma_k$ (in bits/sec)

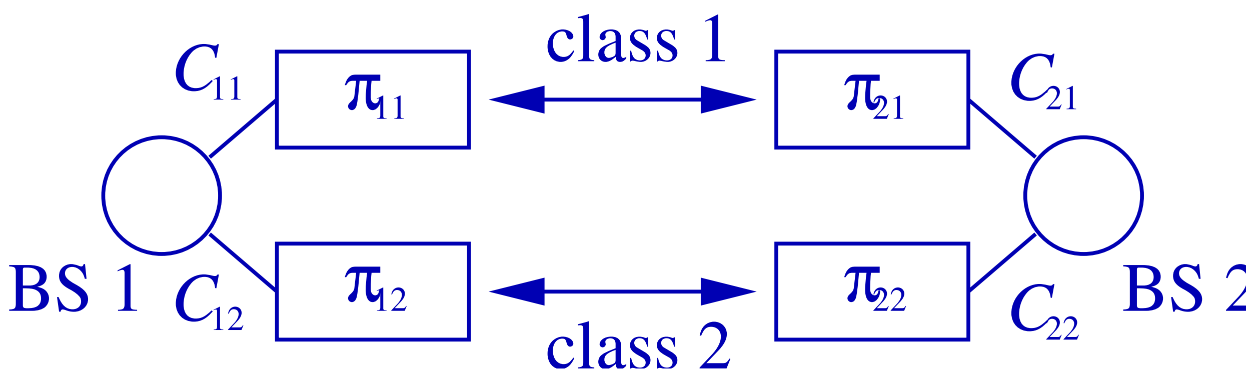
Coverage area may be partitioned into M regions, with $\mathcal{M}(l) \subseteq \{1, \ldots, M\}$ set of regions covered by base station l

Class-k users reside in region m with probability π_{mk}

When residing in region m, feasible transmission rate of class-k user is C_{mk} (in bits/sec)

Intra- & inter-cell mobility (cont'd)

Consider network with $L = 2$ base stations and $K = 2$ user classes



Class-k users reside in cell l with probability π_{lk}, with $\pi_{11} + \pi_{21} = 1$ and $\pi_{12} + \pi_{22} = 1$

Feasible transmission rate of class-k user in cell l is C_{lk}

Intra- & inter-cell mobility (cont'd)

What is capacity region, i.e., for what vectors (ρ_1, ρ_2) can stability be achieved?

Stability can be achieved if there exist numbers $\tau_{lk} \geq 0$, with $\tau_{lk} = 0$ whenever $\pi_{lk} = 0$, such that

$$\begin{aligned} \rho_1 &< \tau_{11} C_{11} + \tau_{21} C_{21} \\ \rho_2 &< \tau_{12} C_{12} + \tau_{22} C_{22} \\ \tau_{11} + \tau_{12} &\leq 1 \\ \tau_{21} + \tau_{22} &\leq 1 \end{aligned}$$

Numbers τ_{lk} may be interpreted as time fractions that class k is served by base station l

Stability conditions only depend on whether probabilities π_{lk} are zero or non-zero, not on their actual values

Intra- & inter-cell mobility

Comparison with no-mobility scenario: identical spatial distribution of users, but users do not move

Stability can be achieved if

$$\pi_{11}\rho_1/C_{11} + \pi_{21}\rho_1/C_{21} < 1$$
$$\pi_{12}\rho_2/C_{12} + \pi_{22}\rho_2/C_{22} < 1$$

It is easily verified that if above conditions are satisfied, then so are earlier ones

In other words, mobility increases capacity region

Reminiscent of results for ad-hoc networks

Heuristic explanation: mobility creates additional opportunities to serve flows in relatively favorable locations

Practical implication: "ignoring" mobility provides conservative dimensioning approach

Intra- & inter-cell mobility

What if globally optimal scheduling is not feasible, and base stations operate according to local fair-sharing policy?

When there are a total of n_l users in cell l, each class-k user receives transmission rate $\frac{C_{lk}}{n_l}$

System is stable if there exist numbers θ_1, θ_2 such that

$$\rho_1 < \frac{\pi_{11}\theta_1}{\pi_{11}\theta_1 + \pi_{12}\theta_2} C_{11} + \frac{\pi_{21}\theta_1}{\pi_{21}\theta_1 + \pi_{22}\theta_2} C_{21}$$
$$\rho_2 < \frac{\pi_{12}\theta_2}{\pi_{11}\theta_1 + \pi_{12}\theta_2} C_{12} + \frac{\pi_{22}\theta_2}{\pi_{21}\theta_1 + \pi_{22}\theta_2} C_{22}$$

Numbers θ_1, θ_2 may be interpreted as (relative) numbers of users when load approaches critical value

- When number of users grows large, they distribute themselves between cells according to probabilities π_{lk}

- Numbers of users of both classes must grow large

Intra- & inter-cell mobility (cont'd)

Comparison with no-mobility scenario: identical spatial distribution of users, but users do not move

Scheduling strategy is irrelevant in absence of mobility

System is stable if

$$\pi_{11}\rho_1/C_{11} + \pi_{21}\rho_1/C_{21} < 1$$
$$\pi_{12}\rho_2/C_{12} + \pi_{22}\rho_2/C_{22} < 1$$

It may be shown that if above conditions are satisfied, then so are earlier ones

In other words: mobility increases capacity, also under fair-sharing policy

Heuristic explanation: gain from users moving from poor spot to good location dominates negative effect of users moving in opposite direction

Intra- & inter-cell mobility (cont'd)

By definition, capacity region for fair-sharing policy is included in that for optimal scheduling: $\mathcal{R}^{\text{fair}} \subseteq \mathcal{R}$

In all but a few special cases, inclusion is strict: $\mathcal{R}^{\text{fair}} \subset \mathcal{R}$

However, it may be shown that for every given load vector $(\rho_1, \rho_2) \in \mathcal{R}$**, there exist probabilities** π_{lk} **such that** $\mathcal{R}^{\text{fair}} = \mathcal{R}$

Thus, maximum stability can be achieved under fair-sharing policy when users are optimally distributed

Intra- & inter-cell mobility (cont'd)

Does there exist simple mechanism for users to be optimally distributed?

Suppose users are distributed such that no individual user can acquire higher throughput by moving to different region

Consider class-k users in cell l:

$$n_{lk} = 0 \text{ unless } \frac{C_{lk}}{n_l} \geq \frac{C_{mk}}{n_m + 1} \text{ for all } m = 1, \ldots, L,$$

It may be shown that under above equilibrium condition, fair-sharing policy achieves maximum stability

May be interpreted as form of dynamic load balancing