HSDPA Flow Level Performance:

The Impact of Key System and Traffic Aspects

Hans van den Berg
TNO Telecom &
University of Twente
The Netherlands
j.l.vandenberg@telecom.tno.nl

Remco Litjens
TNO Telecom
The Netherlands
r.litjens@telecom.tno.nl

Joost Laverman
University of Amsterdam
The Netherlands
j.f.laverman@telecom.tno.nl

ABSTRACT

We present a flow-level performance evaluation for a UMTS/HSDPA network. The results provide thorough insights into the performance impact of a number of key system, environment and traffic aspects, e.g. terminal location, the presence of intercellular interference, the intrinsic feedback delay in the channel quality indications, soft combining of retransmissions and the applied packet scheduling scheme. The contribution of the identified key aspects in the experienced service quality and spatial fairness is assessed by evaluating gradually more 'complete' scenarios.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Design studies, performance attributes

General Terms

Performance

Keywords

UMTS, HSDPA, QOS, link adaptation, packet scheduling, multipath fading, spatial fairness, performance evaluation.

1. INTRODUCTION

With the gradual emergence of third-generation WCDMA-based cellular networks the wireless networking revolution continues to unfold, with the range of offered services rapidly extending from primarily speech telephony to a variety of appealing data and multimedia-based applications. It is anticipated that interactive and background data services, e.g. Internet access, remote database access, electronic mail, will constitute a dominant share in the aggregate teletraffic load carried by 3G networks. In order to support such delay-tolerant services with enhanced resource efficiency and service quality, the Release 5 specifications of the UMTS standard incorporates a significant technogical upgrade in the form of High Speed Downlink Packet Access (HSDPA), which is based on four basic principles: higher order modulation, fast link adaptation, fast scheduling and hybrid ARQ.

HSDPA performance is typically investigated by means of packet-level simulations of *persistent* [5, 6, 10] or *semi-persistent data flows* [12, 15, 17], where a given number of terminals maintain endless WWW browsing sessions, considering the aggregate performance impact of all relevant system and environment aspects in detailed simulation models and not capturing the true flow-level dynamics. With flow level dynamics we refer to the initiation and completion of (finite) flows at various locations, leading to a varying number of concurrent flows competing for shared resources. On the other hand, analytical *flow-level* performance evaluation approaches are generally forced to consider rather idealistic models [7, 8]. Our objective is to provide deeper insight by *decomposing* the flow level performance induced by different scheduling schemes in a UMTS/HSDPA network with respect to the relative performance impact of a set of key system, environment and traffic-related aspects. In particular, we concentrate on the impact of terminal location, the presence of multipath fading and intercellular interference, the inherent feedback delay in the channel quality reports, the correctional capabilities of HSDPA's hybrid ARQ scheme (soft combining), the flow level traffic dynamics and the flow size variability.

The outline of the paper is as follows. Section 2 describes HSDPA in more detail. Subsequently, the considered performance evaluation model is given in Section 3. In Section 4 three distinct packet scheduling schemes are described. Section 5 outlines the decomposition of the performance evaluation into a set of gradually more realistic scenarios. An analytical evaluation of the more tractable scenarios is given in Section 6. Section 7 then presents a set of numerical results obtained via analysis and/or simulations. Section 8 ends this paper with some concluding remarks.

2. HSDPA

A number of technological improvements of the initial UMTS system release are standardised under the name High-Speed Downlink Packet Access [1, 2, 13]. The main objective of HSDPA in UMTS networks is to enable the support of downlink peak rates in the range of $8 - 10$ Mbits/s for best effort packet data services, i.e. far beyond the 3G requirement of 2 Mbits/s. To this end, HSDPA introduces the High Speed Downlink Shared CHannel (HS-DSCH, see Figure

MSWiM'04, October 4-6, 2004, Venezia, Italy.

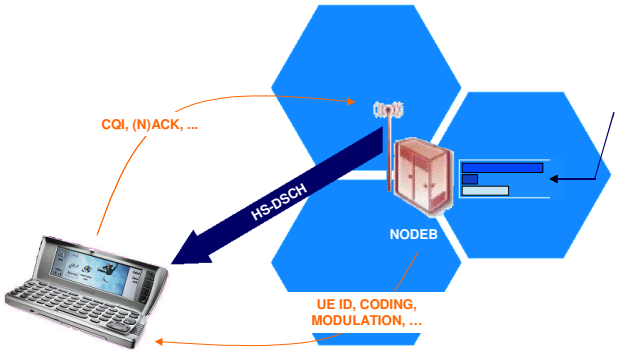

Figure 1: The figure shows the feedback loop between the base station and the terminal. Aside from the actual data, the base station transfers essential signalling information, e.g. the applied channel coding and modulation schemes. On the reverse link, the terminal indicates e.g. the actual channel quality and sends (negative) acknowledgements.

1) as an upgraded version of the Downlink Shared CHannel that is available in 'basic' UMTS. The HS-DSCH, which is parameterised by an assigned transmission power rather than an assigned transfer rate, is characterised by a number of enhanced technologies, which are described below.

In addition to the QPSK (Quadrature Phase Shift Keying) modulation scheme specified in the 'basic' UMTS radio interface standards, the *higher order* 16-QAM *modulation* scheme is added in the HSDPA upgrades, in order to enhance spectral efficiency and thus enable higher data rates in favourable propagation and interference conditions. Since higher-order modulation is less robust to channel impairments, it should be combined with fast link adaptation.

Adaptive modulation and channel coding is applied at the Transmission Time Interval (TTI) time scale based on the CQI (Channel Quality Indicator) feedback from the User Equipment (UE). The objective is to optimise data rates for actual channel conditions, e.g. higher-order modulation with little forward error correction redundancy for a terminal experiencing favourable fading conditions. Aside from adapting the modulation and channel coding parameters, the link adaptation scheme also assigns the number of channelisation codes (of spreading factor 16) applied in parallel.

A *fast* rate-controlled *scheduler* coordinates the (potentially) channel-aware sharing of the HS-DSCH among multiple data flows at the TTI time scale based on the CQI feedback information. An extreme incidence of exploiting channel variations which greedily maximises instantaneous system throughput is pure SNR-based scheduling, i.e. always serve the UE with the most favourable instantaneous channel conditions. There is an apparent trade-off between resource efficiency and fairness among data flows.

A *fast hybrid* ARQ (H-ARQ) scheme is implemented in order to enable rapid retransmissions of erroneous data blocks and soft combining of multiple transfer attempts. As such, H-ARQ provides some degree of robustness against link adaptation errors and reduces transfer delays. Retransmissions are based on chase combining, where transfer reattempts concern identical copies of the freshly sent data block, or on the incremental redundancy principle, where an erroneously received data block is supplemented with additional channel coding bits (see also [16]).

Other proposed enhancements are fast cell selection and the application of MIMO technology. As the effectiveness of the proposed technologies strongly relies on rapid adaptation of transmission parameters to the time-varying channel conditions, the corresponding control schemes, e.g. fast link adaptation and fast scheduling are placed at the direct edge of the radio interface, i.e. at the NodeB. This is in contrast to the current UMTS architecture, where e.g. the scheduling function resides in the RNC. Furthermore, a smaller TTI of 2 ms (as opposed to the current 10 ms TTI) is proposed as the heartbeat for link adaptation and packet scheduling, in order to reduce delays, allow a finer granularity of the scheduling process and facilitate better tracking of the channel variations.

3. MODEL

The model description is broken up into three distinct segments concentrating on the system model, propagation aspects and traffic characteristics.

3.1 System model

We consider a 19-cellular UMTS/HSDPA network of omnidirectional NODE-Bs in a hexagonal layout. A wraparound technique is applied in order to mimic an infinite network and thus avoid undesirable network boundary effects. A hexagonal cell radius of $R \equiv 1/\sqrt{3} \approx 0.577$ is assumed so that the inter-NODE-B distance is precisely 1 km. The investigation concentrates on the *downlink* data transfer over the HS-DSCH transport channel as it is anticipated to become the bottleneck direction of transfer due to the expected asymmetry in the data services. For each NODE-B the transmission power budget is equal to $p_{\max} = 15.849$ Watt ($= 42$ dBm), a Common PIlot CHannel (CPICH) power of $p_{\text{CPICH}} = 1$ Watt is applied, while another constant downlink transmission power of $p_{\text{other}} = 6$ Watt models the presence of other downlink traffic, e.g. speech calls.

Each NODE-B is assumed to provide a single HS-DSCH for data transfer, characterised by a fixed transmission power of $p_{\text{HS-DSCH}} = 3$ Watt. The UEs that maintain active data flows continuously monitor the Signal-to-Noise Ratio (SNR) as experienced on the HS-DSCH and report the corresponding Chan-nel Quality Indicator (CQI) to the serving NODE-B according to the following mapping, which is based on a target value for the induced BLER (see [9]):

$$\text{CQI} = \min\left\{\max\left\{0, \left\lfloor \frac{\text{SNR (dB)}}{1.02} + 16.62 \right\rfloor\right\}, 22\right\},$$

where the maximum CQI of 22 corresponds with the considered UE categories 1-6 [1].

The serving NODE-B maps the reported CQI to a combination of coding rate, modulation scheme and a number of assigned channelisation codes, which jointly determines the applied Transport Block Size (TBS) (see Table 7A in [1]; UE categories 1-6). Due to the uplink transmission of the CQI as well as its processing on both sides, a delay of typically three TTIs exists between the SNR measurement by the UE and its effectuation in the selected link attributes, the corresponding TBS and, potentially, the scheduling decision.

A transferred data block experiences a BLER which is a function of the applied link attributes (directly determined

by the CQI) and the experienced SNR during transfer. We apply the following relation between BLER, CQI and SNR, as derived in [9] by means of detailed link-level simulations:

$$\text{BLER} = \left\{ 10^{\left(2\frac{\text{SNR}-1.03\text{CQI}+17.3}{\sqrt{3}-\log_{10}(\text{CQI})} \right)} + 1 \right\}^{-\frac{1}{0.7}}.$$

It is noted that due to the above-mentioned CQI delay and possible changes in radio link quality during this delay, the SNR experienced during transfer may differ from that underlying the reported CQI. Figure 2 depicts the relation between the selected gross transfer rate (= TBS/TTI duration, in kbits/s), the corresponding BLER and the SNR (in dB).

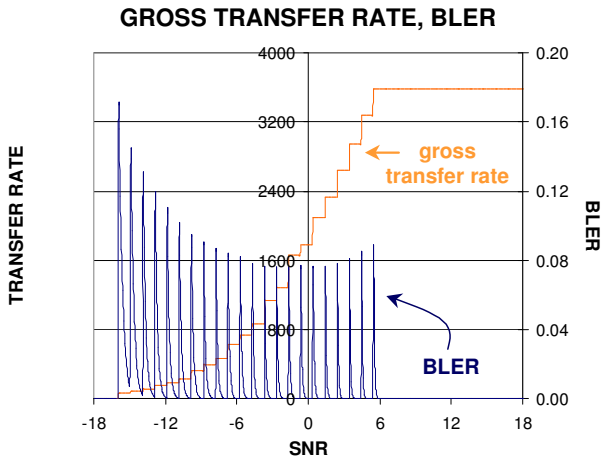

Figure 2: Relation between the selected gross transfer rate and the corresponding BLER versus the SNR.

The applied H-ARQ scheme is based on chase combining, where erroneous data blocks are resent using the TBS of the original transfer and soft combining is implemented by summing the experienced SNRs of subsequent attempts. A maximum of two *re*attempts is allowed, while subsequent transfers are attempted six TTIs apart (see e.g. [19]).

3.2 Propagation aspects

The radio propagation model considers distance-based signal attenuation as well as multipath propagation of n_{p} dominant paths with relative strength γ_i, $i = 1, \cdots, n_{\text{p}}$, each suffering from Rayleigh fading. Given a distance r between the NODE-B b (transmitter) and the UE m (receiver) the relation between the transmission ($p_{\text{transmission}}$) and reception ($p_{\text{reception}}$) powers on dominant path i at time t is given by the instantaneous path gain $\mathcal{G}_{b,m,i}(t)$:

$$\mathcal{G}_{b,m,i}(t) = \eta_{\text{basic}} \cdot r^{-\varsigma} \cdot \xi_{b,m,i}(t),$$

$t \geq 0$, where η_{basic} reflects the basic transmission loss and ς is the path loss exponent. The Rayleigh fading effect $\xi_{b,m,i}(t)$ on the considered dominant path at time t is modelled as follows:

$$\xi_{b,m,i}(t) = \left[\sqrt{\frac{2}{n_{\text{w}}}} \sum_{j=1}^{n_{\text{w}}} \cos\left(\begin{array}{c} \omega_c t + \phi_{b,m,i,j} + \\ \left(\frac{2\pi v}{\lambda_{\text{UMTS}}} \cos\left(\zeta_{b,m,i,j}\right) \right) t \end{array} \right) \right]^2,$$

where the terminal velocity is denoted v (in m/s), λ_{UMTS} is the wavelength, ω_c is the angular frequency, $\phi_{b,m,i,j}$ is the phase of wave k and $\zeta_{b,m,i,j}$ is the azimuth angle of wave j. We take $\phi_{b,m,i,j}$, $\zeta_{b,m,i,j} \sim U(0, 2\pi)$, independent and identically distributed. The uniform amplitudes of the waves are chosen such that the average Rayleigh effect is 'neutral', i.e. equal to 1. The effects of terminal mobility are assumed to be rather localised and incorporated in the Rayleigh fading process only.

In the SNR calculations, which are carried out on a per time slot (= 1/1500 s) basis and subsequently averaged over a TTI, signal components following identical dominant paths are assumed to maintain perfect code orthogonality and thus do not interfere, while signal components following different dominant paths do interfere without any orthogonality gain. A spatially uniform thermal noise level of $\nu = 1.214 \cdot 10^{-13}$ Watt (= −99.158 dBm) is considered. The SNR (in linear units) experienced by UE m with serving NODE-B b in time slot t is thus given by

$$\text{SNR}_{b,m}(t) = \sum_{i=1}^{n_{\text{p}}} \frac{\widehat{\gamma}_i p_{\text{HS-DSCH}} \mathcal{G}_{b,m,i}(t)}{\sum_{j \neq i} \widehat{\gamma}_j p_{\text{TOTAL}} \mathcal{G}_{b,m,j}(t) + I + \nu}, \qquad (1)$$

with $\widehat{\gamma}_i$, $i = 1, 2, 3$, the linearised and normalised relative signal strengths, $p_{\text{TOTAL}} \equiv p_{\text{CPICH}} + p_{\text{other}} + p_{\text{HS-DSCH}}$ the total power exerted by the serving base station, and I the amount of interference experienced from other NODE-Bs.

The assumed propagation-related parameters are given by $n_{\text{p}} = 3$, $(\gamma_1, \gamma_2, \gamma_3) = (0, -9.7, -19.2)$ dB, $\eta_{\text{basic}} = 137.744$ dB, $\varsigma = 3.523$, $n_{\text{w}} = 10$, $v = 0.8$ m/s, $\lambda_{\text{UMTS}} = 0.15$ m and $\omega_c = 4\pi \cdot 10^9$ rad/s.

3.3 Traffic characteristics

The considered UMTS/HSDPA network serves data flows which are assumed to be downlink transfers of documents with mean size $1/\mu = 320$ kbits. The data flow size distribution is taken to be either determinstic, exponential or hyperexponential. The data flows are generated according to a spatially uniform Poisson process with rate $\lambda = 2.5$ flows/s/cell.

4. SCHEDULING SCHEMES

The packet scheduler governs the channel sharing by time multiplexing the different data flows over the single HS-DSCH. Three distinct packet schedulers are considered: the channel-oblivious Round Robin (RR) scheme and two channel-aware schemes: the pure SNR-based scheduler and the Proportional Fair (PF) scheduler, all of which are briefly described below.

The RR scheduler cyclically serves the present data flows that have positive CQI with a TTI heartbeat, and is thus intrinsically *fair* in the sense that each data flow gets an egalitarian share of the HS-DSCH resources.

The SNR-based scheduler bluntly exploits the channel quality variations due to multipath fading, in the sense that in each TTI it serves the data flow with the most favourable instantaneous channel conditions, reflected by the reported CQI. RR tie-breaking is applied in case multiple data flows have identical CQIs. The SNR-based scheduler thus greedily maximises the instantaneous system throughput (though not necessarily the long-term resource efficiency [8]) at the expected cost of a reduced fairness among data flows, as near UEs are more likely to be served than remote UEs.

The PF scheduler [3] aims to strike a compromise between

the fairness of the RR scheme and the efficiency of the SNR-based scheduler by serving that flow at TTI t which maximises the ratio $R_m(t)/\widetilde{R}_m(t)$, where $R_m(t)$ denotes the instantaneous gross data rate of flow m and $\widetilde{R}_m(t)$ denotes its exponentially smoothed experienced gross data rate:

$$\widetilde{R}_m(t) = (1-\alpha)\,\widetilde{R}_m(t-1) + \alpha \mathcal{I} R_m(t-1)$$

with the indicator $\mathcal{I} = 1$ (0) if data flow m was (not) served in TTI $t-1$, $\alpha \in [0,1]$ the associated smoothing parameter and $\widetilde{R}_m(t_0) = \alpha$ the assumed initial value at the flow's generation time t_0. It is readily verified that for $\alpha = 0$ ($\alpha = 1$) the PF scheduler is identical to the SNR-based (RR) scheduler (loosely using the convention that $1/0 = \infty$). The numerical results presented below assume $\alpha = 0.001$.

5. EVALUATION SCENARIOS

The objective of the presented study is to provide thorough insight into the relative performance impact of a number of key system, environment and traffic-related components in a typical setting. This investigation is carried out using two sets of numerical experiments, the first of which specifically concentrates on the system and environment aspects, while the second set of experiments primarily focuses on the impact of various traffic-related aspects.

5.1 Impact of system and environment aspects

The performance evaluation carried out within the first set of experiments will be presented as follows. Considering both a single cell and a network scenario, for each of the three packet scheduling schemes described in Section 4, the expected flow transfer times will be determined as a function of the terminal location, for four distinct gradually more complete (realistic) scenarios that are specified in the table below. The experiments are thus explicitly targeted to reveal the performance impact of the terminals' distance to the serving NODE-B, the presence of inter-cellular interference, the presence of multipath fading, the CQI feedback delay, (H-)ARQ functionality and the applied packet scheduling scheme. The data flow sizes are assumed to be exponentially distributed in these scenarios.

	multipath fading	CQI delay	ARQ
SCENARIO I	×	ideal	basic
SCENARIO II	✓	ideal	basic
SCENARIO III	✓	3 TTIs	basic
SCENARIO IV	✓	3 TTIs	H-ARQ

5.2 Impact of traffic-related aspects

The second set of experiments is designed to concentrate on the contribution of different system and traffic-related aspects on the variability of the experienced service quality. These experiments are triggered by a seemingly intrinsic drawback of the resource efficient SNR-based scheduler that is regularly noted in the HSDPA literature, that the QOS levels are characterised by a relatively large variability, due to the scheduler's favouring of near over distant terminals [12, 15, 17]. Since the operations of the packet scheduler is only one among various aspects that contribute to this QOS variability, we aim to decompose this variability by isolating the different contributions. Aside from the specifics of the packet scheduler, we identify the flow size variability and the flow level dynamics, i.e. the varying number of concurrent (and thus competing) data flows in the system, as the principal sources of QOS variability. Considering the network case and the propagation and system aspects as given by the most 'complete' scenario IV, the following table specifies the considered scenario extensions.

	flow level dynamics	flow size PDF
SCENARIO IV.*	×	deterministic
SCENARIO IV.0	✓	deterministic
SCENARIO IV.1	✓	exponential
SCENARIO IV.3	✓	hyperexponential

The numerical index appended to the scenarios' labels indicates the considered coefficient of variation of the data flow sizes, where the 'balanced means' approach is applied to determine the distribution parameters for the hyperexponential case (see e.g. [18]). The implementation of scenario IV.* is nontrivial and thus needs to be clarified. Firstly, it is noted that we do *not* consider persistent data flows, as in e.g. [5, 6, 10], since it would obviously be impossible to determine the transfer time performance, while one would further need to make specific choices for the locations of the limited number of terminals, and thus inhibit a statistically adequate consideration of the assumed spatially uniform traffic distribution. Rather, an (approximately) fixed number of nonpersistent data flows are simulated, where each completed flow transfer is followed by a single (potentially slightly delayed) fresh data flow arrival, in order to appropriately vary the number of concurrent flows in each cell closely around the average number of concurrent data flows observed under scenario IV.0. As in the other scenarios, the flows' locations are randomly sampled from a spatially uniform distribution.

6. ANALYTICAL EVALUATION

For the RR and SNR-based scheduling schemes this section presents a performance evaluation for scenarios I and II (RR scheme only) using a combination of Monte Carlo-based simputation[1] and stochastic analysis.

6.1 Round robin scheduling

For scenarios I and II, i.e. without CQI delay or H-ARQ, the flow-level performance under the RR scheme can be analysed by means of a multi-class $M/G/1$ processor sharing model as follows.

6.1.1 Single cell case

Divide the circular area of the considered cell into n_{zones} disjunct zones with equal area. For each such zone, determine via Monte Carlo simputation the expected *net* transfer rate r_j, $j = 1, \cdots, n_{\text{zones}}$, that a data flow in this zone experiences when served, sampling over all possible locations within the considered zone (scenarios I and II) and all possible Rayleigh fading effects (scenario II only). For a given sample, the net data rate is determined via calculation of the experienced SNR, mapping this to the gross data rate and subsequently flipping a biased coin with failure probability equal to the associated BLER.

[1]The term 'simputation', a contraction of 'simulation' and 'computation', refers to the numerical evaluation of an expression that can in principle be written in analytical closed form, by means of (typically) Monte Carlo simulations.

The considered system with n_{zones} zones, RR scheduling and zone-specific net transfer rates r_j, $j = 1, \cdots, n_{\text{zones}}$, can be modelled by an $M/G/1$ processor sharing model with n_{zones} flow classes, as also recognised in [7, 8]. The model belongs to the class of product-form 'networks' and is analytically tractable (see e.g. Cohen [11]). In particular, the joint distribution of the number $\mathbf{N}_j$ of flows of class j in the system, $j = 1, \cdots, n_{\text{zones}}$, is given by

$$\begin{aligned} & \Pr\{\mathbf{N}_1 = k_1, \cdots, \mathbf{N}_{n_{\text{zones}}} = k_{n_{\text{zones}}}\} \\ = \; & (1-\rho)\,\frac{(k_1 + \cdots + k_{n_{\text{zones}}})!}{k_1! \cdots k_{n_{\text{zones}}}!} \prod_{j=1}^{n_{\text{zones}}} \rho_j^{k_j}, \end{aligned} \tag{2}$$

with $\rho_j \equiv \lambda_j / (r_j \mu)$ the traffic load offered to the system in zone j, $\lambda_j \equiv \lambda / n_{\text{zones}}$ the flow arrival rate in zone j, $j = 1, \cdots, n_{\text{zones}}$, and where $\rho \equiv \sum_{j=1}^{n_{\text{zones}}} \rho_j$ denotes the aggregate traffic load. Using Little's formula the expected transfer time of a data flow in zone j is then readily derived to be equal to

$$\mathbf{E}\{\mathbf{T}_j\} = \frac{1}{\lambda / n_{\text{zones}}} \mathbf{E}\{\mathbf{N}_j\} = \frac{1/(r_j \mu)}{1-\rho},$$

for $j = 1, \cdots n_{\text{zones}}$.

The above expression is known to be insensitive to the specific form of the flow size distribution, depending on the mean flow size only. Furthermore, the conditional expected data flow sojourn time is linear in the data flow size [11]. The expression for $\mathbf{E}\{\mathbf{T}_j\}$ further reveals that unfavourably located data flows (low r_j) not only suffer themselves from their unfortunate location, but also severely reduce the performance of data flows with a better location (via ρ). In fact, a single large badly located data flow may even cause a twofold increase in the transfer time of other flows, even those at a favourable location (see [4]). Observe that the ratio of expected transfer times $\mathbf{E}\{\mathbf{T}_j\} / \mathbf{E}\{\mathbf{T}_i\}$ is equal to the inverse ratio of the experienced net transfer rates.

6.1.2 Network case

The network case is treated as follows for both scenarios I and II. Denote with $\rho(0)$ the aggregate traffic load as obtained for the single cell case, where the index '0' refers to the initialisation of an iterative procedure that is followed to determine the zone-specific expected transfer times in the network case. Observe from expression (2) that $\rho(0)$ also expresses the equilibrium probability that the HS-DSCH 'server' is busy. In the Monte Carlo simputations that are carried out to determine the expected net transfer rates for each zone in the network case, the HS-DSCHs associated with the NODE-Bs surrounding the reference NODE-B are randomly and independently sampled to be 'on' or 'off' with probability $\rho(0)$ and $1-\rho(0)$, respectively. We note here that in reality, whether a neighbouring HS-DSCH is 'on' is positively correlated with the 'on-off' status of the reference HS-DSCH and hence the assumption of independence is an approximation. Whenever an HS-DSCH is 'on' ('off') the considered data flow in the reference cell experiences (no) interference from this HS-DSCH. The net transfer rate $r_j(0)$ that is determined for a data flow in zone j in the reference cell, $j = 1, \cdots, n_{\text{zones}}$, thus incorporates the random activity of all surrounding HS-DSCHs. The resulting aggregate traffic load in the reference cell is given by $\rho(1) \equiv \sum_{j=1}^{n_{\text{zones}}} \lambda_j / (r_j(0)\,\mu)$, given the obtained net transfer rates.

In order to appropriately consider a *symmetrical* 19-cellular wraparound network, we subsequently redo the (swift) Monte Carlo simputation of the net transfer rates, now with $\rho(1)$ as the probability that a neighbouring HS-DSCH is 'on'. This yields an effective traffic load of $\rho(2)$ in the reference cell, which is subsequently applied to model the surrounding HS-DSCH's activity, etc. This iterative procedure is continued until a fixed-point (if this exists) is attained with sufficient accuracy. In general the obtained aggregate traffic load increases throughout the iterative procedure, since the impact of the inter-cellular interference on the net transfer rates becomes more and more significant. It is then readily seen that *if* a fixed point exists where 'on' probability ρ^* at the surrounding NODE-Bs induces an aggregate traffic load of ρ^* in the reference cell then the above iterative procedure converges to this fixed point, given the monotonous relation between the two measures.

6.2 SNR-based scheduling

For scenario I, i.e. without Rayleigh fading, CQI delay and H-ARQ, the flow-level performance under the SNR-based scheduling scheme can be analysed by means of a multi-class $M/M/1$ queueing model with a priority-based service discipline.

6.2.1 Single cell case

In a similar fashion as was done for the RR scheme above, we divide the cell into n_{zones} where the zone boundaries are derived such that there is a one-to-one correspondence between the zones and the applied *gross* transfer rate. Given these zones, Monte Carlo simputation is once again applied to determine the expected *net* transfer rate r_j, for each zone $j = 1, \cdots, n_{\text{zones}}$, that a data flow in this zone experiences when served, sampling over all possible locations within the considered zone. Note that within each zone, the terminal location only influences the experienced BLER and consequently the net transfer rate, while the gross data rates are a priori known.

The considered system with n_{zones} zones, SNR-based scheduling and zone-specific net transfer rates r_j, $j = 1, \cdots, n_{\text{zones}}$, can be modelled by an $M/M/1$ model with n_{zones} service classes and strict priority-based flow handling. This model is analytically tractable and some relevant performance measures can be derived as follows. The flow arrival rate of service class j is denoted λ_j, $i = 1, \cdots, n_{\text{zones}}$, and is readily derived from the aggregate flow arrival rate and the relative areas of the different zones. Denote with $\rho_j \equiv \lambda_j / (r_j \mu)$ the offered traffic load in zone j, $j = 1, \cdots, n_{\text{zones}}$ and let $\rho \equiv \sum_{j=1}^{n_{\text{zones}}} \rho_j$ denote the aggregate traffic load.

The expected flow transfer time for the highest priority class readily follows from a basic $M/M/1/PS$ model that includes only the flows of the highest priority, and is given by

$$\mathbf{E}\{\mathbf{T}_1\} = \lambda_1^{-1} \mathbf{E}\{\mathbf{N}_1\} = \lambda_1^{-1} \frac{\rho_1}{1-\rho_1} = \frac{1}{r_1 \mu\,(1-\rho_1)},$$

using Little's formula. In order to derive an expression for the expected transfer time for each of the lower priority classes, the following procedure is followed. We first note that due to the strict priority-based scheduling discipline, the performance of the j-th priority class is influenced *only* by the flows of the higher priority classes. The expected

transfer time of a flow of priority class j is given by

$$\mathbf{E}\{\mathbf{T}_j\} = \lambda_j^{-1}\mathbf{E}\{\mathbf{N}_j\} = \lambda_j^{-1}\frac{\mathbf{E}\{\mathbf{WL}_j\}}{\mathbf{E}\{\mathbf{R}_j\}},$$

$j = 2, \cdots, k$, where $\mathbf{WL}_j$ denotes the work load of the j-th service class in the system which includes only priority classes 1 to j, while $\mathbf{R}_j$ denotes the residual service requirement (flow size) of a flow of priority class j. The above expression applies Wald's equation (e.g. Tijms [18]), using the independence of $\mathbf{N}_j$ and $\mathbf{R}_j$ that is due to the memorylessness property of the exponential flow size distribution, which further implies that $\mathbf{E}\{\mathbf{R}_j\} = 1/(r_j\mu)$. The expected workload $\mathbf{E}\{\mathbf{WL}_j\}$ can be determined as the difference between the aggregate workload in a system with only priority classes 1 to j and one with only priority classes 1 to $j-1$:

$$\mathbf{E}\{\mathbf{WL}_j\} = \mathbf{E}\left\{\sum_{k=1}^{j}\mathbf{WL}_k\right\} - \mathbf{E}\left\{\sum_{k=1}^{j-1}\mathbf{WL}_k\right\},$$

where

$$\begin{aligned}\mathbf{E}\left\{\sum_{k=1}^{j}\mathbf{WL}_k\right\} &= \frac{\left(\sum_{k=1}^{j}\lambda_k\right)\sum_{k=1}^{j}\left(\frac{\lambda_k}{\sum_{k=1}^{j}\lambda_k}\right)\frac{2}{r_k^2\mu^2}}{2\left(1-\sum_{k=1}^{j}\rho_k\right)} \\ &= \frac{\sum_{k=1}^{j}\frac{\rho_k}{r_k\mu}}{1-\sum_{k=1}^{j}\rho_k}, \qquad (3)\end{aligned}$$

where the second factor in the numerator is the second moment of the overall (normalised) service requirement and $\sum_{k=1}^{j}\rho_k$ is the aggregate traffic load in the system that includes only service classes 1 to j (e.g. Tijms [18]). Hence

$$\begin{aligned}\mathbf{E}\{\mathbf{T}_j\} &= \lambda_j^{-1}\frac{\mathbf{E}\{\mathbf{WL}_j\}}{\mathbf{E}\{\mathbf{R}_j\}} \\ &= \frac{r_j}{\lambda_j}\left(\frac{\sum_{k=1}^{j}\frac{\rho_k}{r_k}}{1-\sum_{k=1}^{j}\rho_k} - \frac{\sum_{k=1}^{j-1}\frac{\rho_k}{r_k}}{1-\sum_{k=1}^{j-1}\rho_k}\right). \qquad (4)\end{aligned}$$

The overall (unconditional) expected transfer time is then given by

$$\mathbf{E}\{\mathbf{T}\} = \sum_{j=1}^{n_{\text{zones}}}\left(\frac{\lambda_j}{\sum_{j=1}^{n_{\text{zones}}}\lambda_j}\right)\mathbf{E}\{\mathbf{T}_j\}.$$

It is noted that the exponentiality assumption regarding the flow size distribution is not necessary for flows of the highest priority class. The above expressions are still valid if these flows have a general flow size distribution, provided that the expression for the second moment of the overall (normalised) service requirement in expression (3) is appropriately adjusted and expression (4) is changed accordingly.

6.2.2 Network case

The network case is treated in an equivalent iterative manner as described for the RR scheduler, applying the above expressions to eventually derive the resulting transfer time performance. Once again, the aggregate traffic load ρ also expresses the probability that an HS-DSCH is 'on'. The primary difficulty in applying the iterative procedure to the case of SNR-based scheduling, is the separation of the reference cell into disjunct zones. For the single cell case, the zone boundaries naturally followed from the explicit relation between a terminal's distance to its serving NODE-B, the experienced SNR and the reported CQI. In the network case, the experienced SNR (and hence also the reported CQI) also depends on whether or not surrounding HS-DSCHs are 'on'. As it is computationally unattractive to include the definition of zones in the iterative procedure, which would in any case only partially resolve this issue, we defined the zone boundaries from the worst-case scenario that all surrounding HS-DSCHs are 'on'. Given the so-defined zones, the Monte Carlo technique is applied to derive net transfer rates per zone and the iterative procedure is followed.

7. NUMERICAL RESULTS

In order to evaluate the flow level performance of the different scheduling schemes under the different scenarios described in Section 5, the considered system, propagation and traffic model aspects, as specified in Section 3, have been implemented in a dynamic system-level simulator, while analytical results are derived where possible (see Section 6).

7.1 Single cell case: impact of system and environment aspects

Figure 3 presents the Monte Carlo simulation results that were obtained as an input for the analytical evaluation of scenarios I (RR, SNR-based scheduler) and II (RR scheduler only) for the single cell case. The left (right) chart depicts the expected net transfer rate (in kbits/s) and the expected SNR (in dB) versus a terminal's distance to the serving base station for the case without (with) multipath fading. Besides the continuous net transfer rate curves (dashed curves), the discontinuous curves corresponding with the discretisation in zones are also shown.

A first obvious observation that can be made is that both the expected net transfer rates and the expected SNRs are decreasing in the terminal's distance to its serving base station, which in this single cell case is primarily due to the increasing impact of the thermal noise. More interesting, however, is the observed impact of the presence of multipath fading on the performance: the SNRs appear to improve significantly in the presence of multipath fading, while the net transfer rates is hardly affected.

Although the Rayleigh effect is neutral in the sense that the average value of a Rayleigh sample is equal to one, the impact of multipath fading on the SNR performance seems somewhat counterintuitive, in that an added degree of variation typically reduces performance. In order to understand this SNR increase we stress that Rayleigh fading affects both the 'signal' (numerator) and the 'interference' level (denominator) in the SNR (see also (1)). Considering that the instantaneous Rayleigh effect is exponentially distributed, it is noted that the expectation of the *ratio* of two exponentially distributed random variables is infinitely large. Although the presence of a thermal noise-related constant in the 'interference' level ensures that the expected SNR is finite, it remains significantly larger than for the case without multipath fading. Observe, however, that the influence of multipath fading on the SNR performance is smaller for remote terminals as the impact of the (fixed) thermal noise level on the expected SNR becomes more and more dominant.

With regards to apparent translation of a significant improvement in the SNR performance to the hardly affected net transfer rates, we first note that, unlike for the case without multipath fading, for the case with multipath fading, the SNR is characterised by some degree of variability around the depicted means. We note here that for terminals that

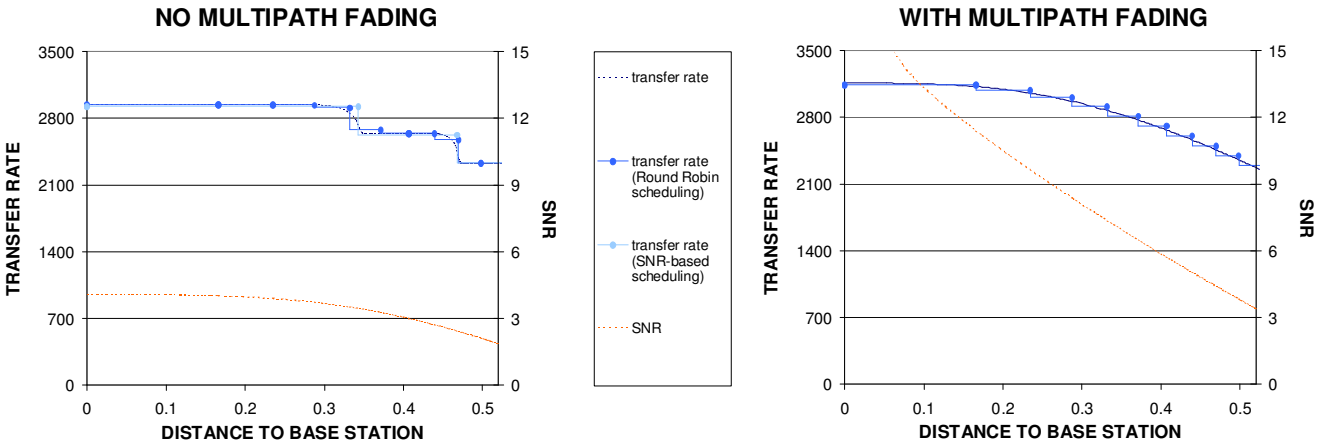

Figure 3: The expected transfer rates and SNRs as a function of a terminal's distance to its serving base station for the case without (left chart) and with (right chart) multipath fading.

are located relatively close to the base station, where the principal SNR performance gain is found, the negative deviations from the SNR mean are experienced as correspondingly lower transfer rates. while the positive deviations increase the transfer rates to a more limited extent, due to the technological maximum on the assigned transfer rates (3584 kbits/s for UE categories 1-6). This explains why the gain in transfer rates that is induced by the presence of multipath fading is much lower than one might suspect based on the SNR curves. Overall, the presence of multipath fading appears to hardly influence the net transfer rates.

Figure 4 shows the flow level performance in terms of the expected transfer times (in seconds), conditional on the terminal location, for all three packet schedulers and the gradually more realistic scenarios as described in Section 5. The charts on the left (right) reflect the single cell (network) case (discussed below). Note that the vertical position of the curves give an indication of the resource efficiency, while the shape of the curves reflect the spatial fairness.

For scenario I which incorporates only the effects of terminal location, observe the significant difference in fairness between the SNR-based scheduler and the other two schedulers, which show roughly the same performance. In line with the observations made above regarding the impact of multipath fading, note that for the RR scheduler, the flow level performance for scenarios I and II are very similar. Since the other two schedulers explicitly *exploit* the SNR variations due to multipath fading, the performance is improved in scenario II, particularly for the purest of channel-aware scheduler, i.e. the SNR-based scheduler. Observe, however, that the priority that near terminals experience in scenario I is less strict in scenario II, since the multipath fading fluctuations may at times allow remote terminals to experience a larger SNR than near terminals. As a consequence, although the performance experienced by remote terminals is significantly improved, this comes at a cost of a slight increase in the transfer times for near terminals. Observe that the analytically obtained curves match with those obtained via dynamic simulations.

When comparing the results of scenarios II and III, we can assess the performance impact of the inherent delay in the terminal's SNR measurements that underly the reported CQI and the effectuation of this CQI. Clearly, this CQI delay significantly worsens the flow-level performance. In order to see this, note that whenever a more favourable TBS is selected than appropriate, due to changed SNR conditions during the CQI delay, the transferred block is likely to be erroneous given the steepness of the BLER curves. On the other hand, when a less favourable TBS is selected, this leads to a BLER smaller than the 10% value that underlies the SNR-to-CQI mapping, which thus hardly pays off. Since the cost of selecting a too high TBS exceeds the gains of selecting a too low TBS, the net effect constitutes a performance degradation.

Soft combining (H-ARQ) is particularly important in a system with link adaptation such as HSDPA as block errors are more likely to occur and thus link-level corrections more important. Comparing scenarios III and IV, the results in the figures clearly demonstrate the gain that is induced by the correctional capabilities of the H-ARQ scheme that are due to the possibility of soft combining. Here roughly half of the performance degradation that was induced by the CQI delay, was restored by means of soft combining.

A final comment concerns the relative performance of the different scheduling schemes. We observe that, for the single-cell scenario, the spatial unfairness of the SNR-based scheduler, is hardly worse than for e.g. the Round Robin scheduler, while the absolute expected transfer times are overall lower. Thus based on the considered scenarios and performance measures, the SNR-based scheduler is preferred.

7.2 Network case: impact of system, environment and traffic aspects

We now consider the numerical results for the 19 cells wraparound *network* depicted in the charts on the right of Figure 4. The traffic offered to each of the cells is the same as in the previous single cell case. As expected, the mean flow transfer times are considerably larger than in the single cell case, which is obviously due to the interference from

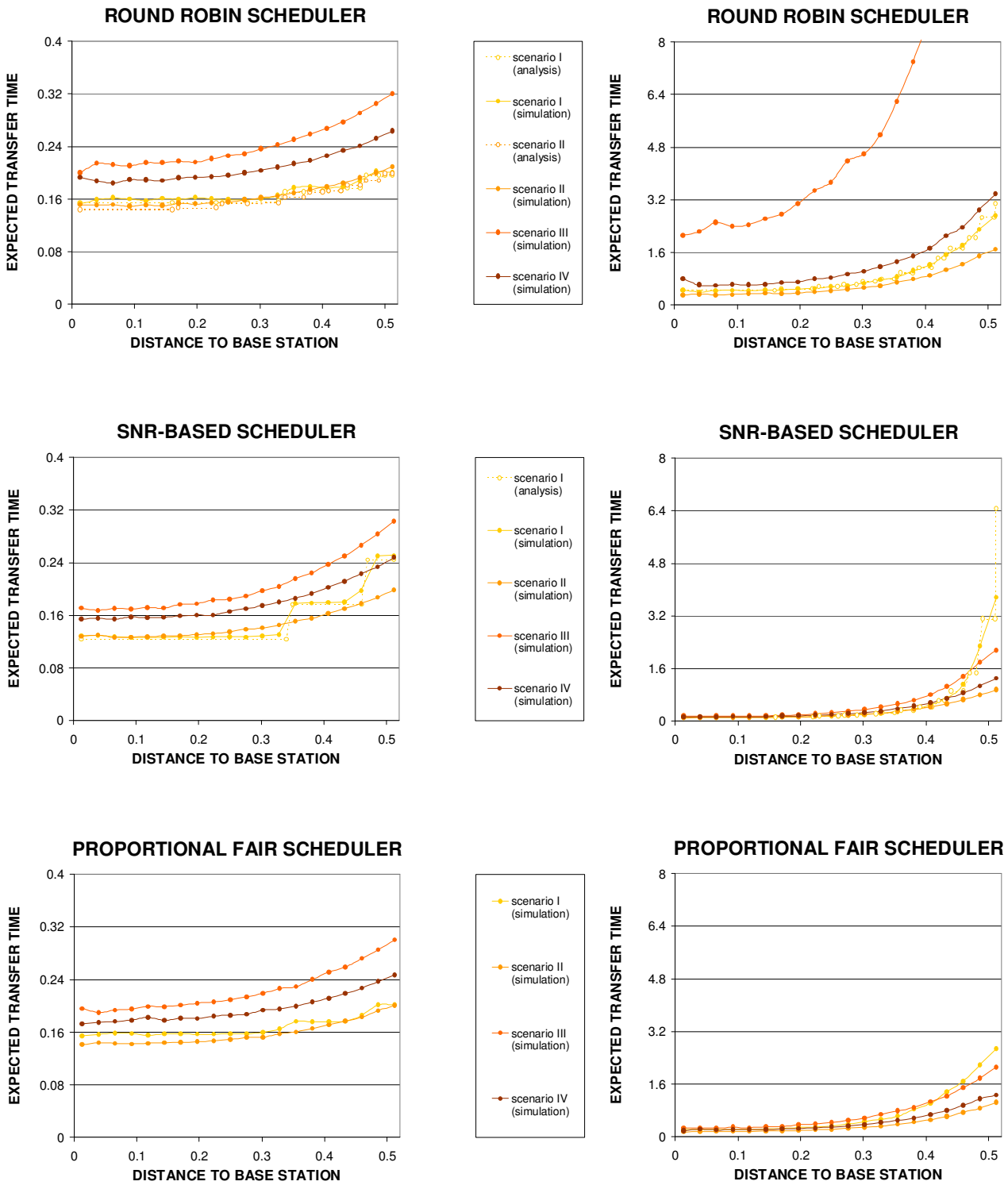

Figure 4: Flow level performance of the three considered packet schedulers for the gradually more realistic scenarios I–IV. The charts in the left (right) column correspond with the single cell (network) case.

other cells leading, in fact, to a smaller cell capacity (i.e. higher load in the case of the same offered traffic). Note, that the spatial unfairness is now much larger than in the single cell case. This is due to the fact that the interference conditions are (due to inter cell interference) most unfavourable at the edge of the cells. The analytical results for scenario I (both for RR and SNR-based scheduling) are, as in the single cell case, very accurate.

The growth of the mean transfer time due to the CQI delay (i.e. scenario III compared to scenario II) is much larger in the case of RR scheduling than for the other schedulers. This is due to the fact that the cell load in case of RR scheduling is larger (closer to the critical load) than in the case of the other schedulers which efficiently exploit SNR peaks. Hence, an increase of the cell load due to the additional retransmissions which are needed when the CQI delay is taken into account (the BLER typically increases from about 1% to more than 20%), will lead to a much more drastic increase of the mean flow transfer time in case of RR than for the other schedulers (remember that in a queueing system the mean delay is roughly inversely proportional to one minus the normalized system load). Note, that for the single cell case the differences between scenario II and III are less pronounced due to the smaller cell load (no other cell interference).

As in the single cell case, the tremendous growth of the mean transfer time due to the CQI delay (in particular for the RR scheduler) is largely compensated by the positive effect of soft combining (H-ARQ). Finally, the results show that the SNR-based scheduler performs slightly better than the PF scheduler; both clearly outperform the RR scheduler.

The numerical results showed, as expected, that flow transfers to users at the cell edge are on average (much) slower than to users located close to the base station. Numerical results (not shown here) also showed that the standard deviation of the flow transfer times is larger at the cell edge. So, it is likely to expect, that, in particular in the case of SNR based scheduling, the variability of flow transfer times at the cell edge will be larger than in the middle of the cell. However, we found that the coefficient of variation of the flow transfer times (i.e. ratio of standard deviation and mean) is almost independent of the distance to the NODE-B.

The results in Figure 5 show the effect of various system features ('sources of variability') on the flow transfer time variability. The left graph shows, for each of the three schedulers, the expected flow transfer times (in seconds) for the scenarios IV.*, IV.0, IV.1 and IV.3 as defined in Section 5; the right graph contains the standard deviations (in seconds) of the flow transfer times.

The results show that for all schedulers the contribution to the flow transfer time standard deviation of the traffic-specific sources of variability (flow level dynamics and flow size distribution contained in scenarios IV.0, IV.1, IV.3) is dominant over the impact of the scheduler itself along with that of the spatial flow distribution (scenario IV.*).

The contribution to the standard deviation due to the RR scheduling algorithm seems to be larger than due to the SNR-based scheduler. However, if we consider the coefficient of variation (i.e. normalize the standard deviations by the corresponding mean transfer times), then we find that SNR introduces more variability than RR. Note, that this is in line with results of scheduling studies where (semi-)persistent sources are assumed [10, 12, 15, 17]. Next, if also the flow level dynamics are taken into account, we observe that the flow transfer time standard deviation is much larger for the RR scheduler than for the SNR-based and PF schedulers. This is apparently due to the smaller capacity (i.e. higher loads) of the cells in case of RR scheduling compared to the cases with SNR-based and PF scheduling where SNR peaks are efficiently exploited.

Note from Figure 5 that for RR the expected flow transfer time decreases when the flow size distribution is taken exponential (scenario IV.1) instead of deterministic (scenario IV.0) and even further decreases when the flow size becomes more variable (scenario IV.3). A similar counterintuitive phenomenon (higher variability of flow size leads to smaller mean flow transfer time) was recently also reported in [14]. They considered sojourn times in a Processor Sharing (PS) queueing model with variable service rate; note that the RR scheduler in the present context can also be considered to be a PS type of queueing system with variable service capacity (variability is a.o. caused by variable interference from other cells).

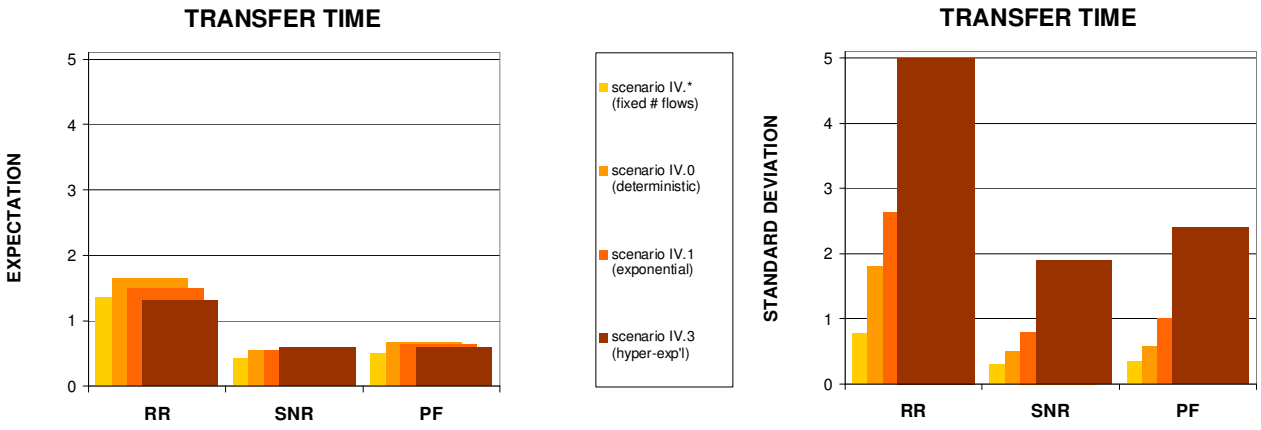

Figure 5: The expectation (left) and standard deviation (right) of the transfer time as experienced under the three considered packet schedulers for the gradually 'more variable' scenarios IV.*–IV.3.

8. CONCLUDING REMARKS

We have presented an evaluation of the flow level performance in a UMTS/HSDPA network, assessing the relative performance impact of a set of key system, environment and traffic-related aspects in order to provide thorough qualitative and quantitative insights. Among the insights gained, we observed that the presence of multipath fading has a positive effect on the flow level performance, in particular under channel-aware schedulers such as the considered proportional fair (PF) and pure SNR-based schedulers. The CQI feedback delay causes a severe performance degradation. The increased number of block errors due to the CQI feedback delay can be partially coped with by the hybrid ARQ (soft combining) reducing the negative effect of the CQI delay considerably. Overall, for the considered settings, the pure SNR-based scheduler outperforms the other considered schedulers (including the well known PF scheduler) with respect to the absolute transfer time performance and the spatial fairness regarding transfer times.

The presented work is extended in various directions. Aside from attempts to develop further analytical approaches to evaluate the flow level performance, we intend to assess the impact of TCP flow control on the obtained qualitative results, while a further important extension is concerned with the integration of speech and data traffic in UMTS/HSDPA networks and to devise and evaluate efficient load-adaptive scheduling schemes that dynamically adjust the power assignment of the HS-DSCH.

9. REFERENCES

[1] 3GPP TS 25.214, *"Physical layer procedures (FDD)"*, v5.8.0, Release 5, 2004.

[2] 3GPP TS 25.848, *"Physical layer aspects of* UTRA High Speed Downlink Packet Access", v4.0.0, Release 4, 2001.

[3] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana and A. Viterbi, 'CDMA/HDR: a bandwidth-efficient high-speed wireless data service for nomadic users', *IEEE Communications magazine,* vol. 38, no. 7, pp. 70-77, 2000.

[4] J.L. van den Berg and O.J. Boxma, 'The M/G/1 queue with processor sharing and its relation to a feedback queue', *Queueing Systems,* vol. 9, pp. 365-402, 1991.

[5] F. Berggren and R. Jäntti, 'Asymptotically fair scheduling on fading channels', *Proceedings of VTC '02*, Vancouver, Canada, 2002.

[6] F. Berggren and R. Jäntti, 'Multiuser scheduling over Rayleigh fading channels', *Proceedings of Globecom '03*, San Francisco, USA, 2003.

[7] T. Bonald and A. Proutière, 'Wireless downlink data channels: user performance and cell dimensioning', *Proceedings of Mobicom '03*, San Diego, USA, 2003.

[8] S.C. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks", *Proceedings of INFOCOM '03,* San Francisco, USA, 2003.

[9] F. Brouwer, I. de Bruin, J.C. Silva, N. Souto, F. Cercas and A. Correia, 'Usage of link-level performance indicators for HSDPA network-level simulations in E-UMTS', *Proceedings of IEEE ISSSTA '04*, Sydney, Australia, 2004.

[10] I.C.C. de Bruin, G. Heijenk, M. El Zarki and J. Lei Zan, 'Fair channel-dependent scheduling in CDMA systems', *Proceedings of the IST Mobile summit*, Aveiro, Portugal, 2003.

[11] J.W. Cohen, "The multiple phase service network with generalized processor sharing", *Acta informatica*, vol.12, pp. 245-284, 1979.

[12] A. Furuskär, S. Parkvall, M. Persson and M. Samuelsson, "Performance of WCDMA high speed packet data", *Proceedings of VTC '02,* Birmingham, USA, 2002.

[13] H. Holma and A. Toskala (editors), *"*WCDMA *for* UMTS*: radio access for third generation mobile communications"*, John Wiley & Sons, Chichester, England, 2002.

[14] R. Litjens and R.J. Boucherie, "Elastic calls in an integrated services network: the greater the call size variability the better the Quality-Of-Service", *Performance evaluation,* vol. 52, no. 4, pp. 193-220, 2003.

[15] R. Love, A. Ghosh, R. Nikides, L. Jalloul, M. Cudak and B. Classon, "High speed downlink packet access performance", *Proceedings of VTC '01,* Rhodes, Greece, 2001.

[16] N. Miki, A. Morimoto, S. Abeta and M. Sawahashi, 'Radio link performance of high-speed packet transmission in HSDPA', *Proceedings of URSI '02*, Poznań, Poland, 2002.

[17] S. Parkvall, J. Peisa, A. Furuskär, M. Samuelsson and M. Persson, "Evolving WCDMA for improved high speed mobile Internet", *Proceedings of the Future Telecommunications Conference '01,* Beijing, China, 2001.

[18] H.C. Tijms, *"Stochastic modelling and analysis: a computational approach"*, John Wiley & Sons, Chichester, England, 1986.

[19] J. Wigard, T. Kolding, K. Pedersen, H. Holma and P. Mogensen, *"High speed downlink packet access (*HSDPA*) for* WCDMA", `http://www.nokia.com/nokia/0,,53713,00.html`, 2003.