HSUPA Scheduling Algorithms Utilizing RoT Measurements and Interference Cancellations

Danlu Zhang, Sharad Sambhwani and Bibhu Mohanty, Member, IEEE

Abstract—This paper discusses several important aspects in the HSUPA scheduling algorithms. First, it clearly demonstrates the benefit of explicitly utilizing the directly measured Rise-over-Thermal(RoT) values in term of capacity and robustness improvement. This paper also briefly presents the RoT measurement schemes designed specifically for the asynchronous systems such as WCDMA/HSUPA. If pilot and traffic interference cancellation (IC) schemes are implemented at the Node B receiver, this paper presents a scheduling algorithm that attains IC benefit in such a way that the capacity enhancement is maximized while fairness and link budget are maintained. The presented algorithm defines the effective RoT by monitoring the IC performance in real time and adjusts its grants accordingly. In addition, the user priority function used here optimizes the sum utility functions and provides multi-user diversity.

Index Terms — HSUPA, HSPA+, Rise-over-Thermal, Load, Interference Cancellation, Link Budget, Scheduling, Fairness, Utility Function, CRRA

I. INTRODUCTION

WCDMA is set to become the most popular choice for the Third Generation cellular systems. After its initial release in 1999[1], the air interface has been subsequently enhanced. Examples include the High Speed Downlink Packet Access (HSDPA) and Enhanced Uplink (EUL, also called High Speed Uplink Packet Access, or HSUPA) which introduce fast packet switch data services in the downlink and uplink respectively[2]. Recently, HSDPA/HSUPA has been further strengthened by backward compatible HSPA+ as well as the new OFDMA-based Long Term Evolution, or LTE[3]. Each enhancement offers substantial improvements in data rates, user experience and capacity.

We focus on HSUPA in this paper. One salient feature of HSUPA is the uplink scheduler at Node B which dynamically allocates data rates based on channel conditions, system loading and fairness constraints. The scheduling algorithm is critical in achieving the full potential of HSUPA. This paper discusses the scheduling algorithm which trades off system capacity and fairness among the users. Apart from the priority function and its economics and optimization interpretation, one of the main thrusts of our scheduling algorithm is utilizing direct measurement of Rise-over-Thermal(RoT)[8], which is the fundamental resource in rate allocations. As will be shown later, scheduling algorithms based on measured RoT provides significant gains in terms of capacity and robustness.

In parallel to the standard enhancements, advanced receiver techniques have been adopted to provide gains without standard changes. A prominent example is the interference cancellations (IC) at the Node B receiver. It is well known that CDMA with Successive Interference Cancellation (SIC) achieves the capacity of AWGN multiple access channels [5]. A direct implementation of SIC requires a pre-determined decoding orders among users whose power/rate allocations have to be configured accordingly. This poses a great challenge to the system operations. However, as shown by Hou et al in [6], the same capacity can be achieved by staggering user transmissions while keeping their control parameters the same. This technique also suits the hybrid-ARQ (HARQ) schemes. The optimality of this scheme will be proved more rigorously in this paper. Simpler IC architectures with approximately the same performance will also be presented. In order for the maximum benefit of IC and maintaining system stability and link budget, the Node B uplink scheduler must be carefully designed. This paper addresses this issue in detail. Instead of using the actual RoT, the scheduler computes the effective interference seen by the decoding packets and regulates its allocation accordingly.

This paper is organized as follows. The system models are presented in Section II. Our scheduling algorithms are discussed in detail in Section III, including user priorities and the utility function interpretation, scheduling with and without measured RoT, RoT measurement scheme and the scheduling using effective RoT. The simulations results are presented in Section IV. This paper concludes in Section V.

II. SYSTEM MODELS

A. HSUPA MAC layer

There are two configurations in the HARQ on HSUPA: 2ms Transmission-Time-Interval(TTI) with 8 interfaces or 10ms TTI with 4 interfaces. Our work is equally applicable to both configurations. This paper focuses on results with 2ms TTI. To request data transmissions, the users send Scheduling Information(SI) messages containing information on the queue lengths and power headroom which can be translated into the maximum data rates. The Node B scheduling grants are sent to the users via two physical channels: Absolute Grant Channel(E-AGCH) and Relative Grant Channels(E-RGCH). The uplink serving cell sends both E-AGCH and E-RGCH, depending on the changes in the user rates and overhead considerations. The non-serving cells who are also in user’s
active set can also send E-RGCH to tune the user down. The inner loop power control maintains the received chip-level SINR to the target level set by outer-loop power control such that a certain packet error rate (PER) is achieved at the HARQ termination target. The users are allowed to transmit autonomously up to a certain threshold rate without being scheduled. This feature is of great importance for services which are low rate but delay sensitive. Voice-over-IP is one prime example. The autonomous transmission in this case greatly reduces scheduling overhead and latency.

The detailed description of HSUPA MAC layer is in the standard specification [14].

B. Basic uplink scheduler computations

Let \( N_0 \) be the thermal noise and \( I_{in}(n) = \sum (\hat{E}_c/N) + N_0 \) be the total interference. RoT is formally defined as \( \text{RoT}(n) = I_{in}(n)/N_0 \).

To link RoT with user SINR, let \( (\hat{E}_c/N)_{i,n} \) be the chip-level pilot SINR of finger \( i \) of user \( k \) during time slot \( n \). Then

\[
(\hat{N}_c)_{i,n} = I_{in}(n) - (\hat{E}_c/N)_{i,n}.
\]

Then

\[
(\hat{E}_c/N)_{i,n} = (\hat{E}_c/N)_{i,n} + 1 + T2 \cdot P_k(n) + \text{Gainoverhead,k}_i.
\]

where \( T2 \cdot P_k(n) \) and \( \text{Gainoverhead,k}_i \) are the offset power of the traffic and overhead channels. These two parameters, and \( (\hat{E}_c/N)_{i,n} \) are assumed known to the scheduler. Let, define the load from finger \( i \) as

\[
\hat{L}_i(n) = \frac{(\hat{E}_c/N)_{i,n}}{1 + (\hat{E}_c/N)_{i,n}} = \frac{(E_c/N)_{i,n}}{I_{in}(n)}.
\]

The scheduler computes the cell load as the sum of load from all the fingers belonging to users who are served by this cell:

\[
\hat{L}_\text{cell}(n) = \frac{1}{N_{ant \_rec \_cell}} \sum \hat{L}_i(n) \quad \text{where } N_{ant \_rec \_cell} \text{ is the number of receive antennas.}
\]

The scheduler also computes the non-serving-active-set load as the sum of load from all the users who are not served by this cell but has this cell in their active set:

\[
\hat{L}_\text{AS,AS}(n) = \frac{1}{N_{ant \_AS \_cell}} \sum \hat{L}_i(n). \quad \text{Let}
\]

\[
L_{total}(n) = 1 - \frac{1}{\text{RoT}(n)} \quad \text{and} \quad L_{out}(n) = L_{\text{total}}(n) - L_{\text{cell}}(n) - L_{\text{AS,AS}}(n)
\]

as the load un-captured by the cell. Filtering is applied to \( \hat{L}_\text{cell}(n) \), \( \hat{L}_\text{AS,AS}(n) \) and \( L_{out}(n) \) to smooth out the estimation variance. Let’s define \( f_\text{cell} \) to be the ratio of load from other cell to the load from this cell

\[
f_\text{cell} = (\hat{L}_\text{AS,AS}(n) + L_{out}(n))/\hat{L}_\text{cell}(n).
\]

Then formally we have

\[
\text{RoT}(n) = \frac{L_{\text{cell}}(n)}{N_0} = \frac{1}{1 - (1 + f_\text{cell})\hat{L}_\text{cell}(n)}.
\]

RoT is the fundamental resources in uplink scheduling since it is linked to system stability in terms of pole capacity and link budget. Typically, stability constraints are on the tail of RoT[9]. The importance of RoT measurements has long been recognized[8]. However, it is not always measured in real time. The main difficulty is measuring the value of \( N_0 \). We will revisit this issue later.

III. SCHEDULING ALGORITHMS

A. User priority functions and the interpretation

The scheduler first computes the available load based on its target with load from continuing packets subtracted. It also assigns a margin for the dedicated channel and autonomous transmissions. The latter may carry VoIP traffic. It then allocates the available load to users. It must trade off capacity and fairness. For user \( k \)'s with request data rate \( r_{\text{support}}(k) \), its scheduling priority is determined by the following:

\[
\text{Priority}(k) = \left( \frac{r_{\text{support}}(k)}{\tilde{r}(k)} \right)^{1/\alpha}.
\]

Here \( \tilde{r}(k) = (1 - 1/T_{\text{thrpt}}) \tilde{r}(k) + (1/T_{\text{thrpt}}) r(k) \) is the filtered throughput with time constant \( T_{\text{thrpt}} \). This is a generalization of the Proportional Fair scheduler [12][13]. It is linked to the economic theory of utility functions. Suppose user’s utility function is dependent on its data rate as \( U(r) = (1 - \alpha) r^{1-\alpha}, \alpha > 0 \). This family of utility functions has the appealing property of Constant Relative Risk Aversion (CRRA)[15], in other words, diminishing return with a constant relative slope: \(- r U''(r)/U'(r) = \alpha \).

Lemma Assuming static channel conditions, the allocation under the above priority function maximizes the sum utility.

Proof: It is straightforward using Lagrange multiplier. ■

With time-varying channels like in wireless systems, multi-user diversity is also achieved by the above priority function. When \( \alpha = 0 \), the scheduler simply chooses the best user; when \( \alpha = 1 \), the utility function is in the log form and the scheduler is the classic proportional fair; when \( \alpha \to \infty \), the scheduler attempts to achieve equal GoS (equal throughput) among the users.

After sorting the users by their priorities, the rate allocation is based on greedy filling. The available load is filled by the user with the highest priority and the remaining part will go to the next, etc. The scheme is shown in [7] to maximize the capacity with fairness and power constraints for each cell.

B. Uplink scheduler without measured RoT

Without RoT measurement, the uplink scheduler has no knowledge of \( L_{out}(n) \). Since it only has direct control of \( \hat{L}_\text{cell}(n) \), it must assume a fixed value of \( f_\text{cell} \) and put a fixed threshold on \( \hat{L}_\text{cell}(n) \). An alternative is put a fixed threshold on \( \hat{L}_\text{cell}(n) \). The performance is very similar so we will only discuss the scheme thresholding \( \hat{L}_\text{cell}(n) \). In practice, \( f_\text{cell} \) is not a fixed value. It depends on network size and layout, traffic types,
fairness considerations. It is also time-varying. In practice, the uplink and downlink have different fading in a FDD system. This is called imbalance. The active set management is typically based on downlink measurements. Henceforth, there may be strong interference from users to cells outside its active set. Such interference is not only un-captured by the cell but also un-controllable since no power control feedback exists between this user and this cell. $f_{cell}$ is large in this scenario. In addition, the delay in active set changes due to the messaging and resource allocations will also cause large $f_{cell}$. Furthermore, the load computation is tied to the Rake fingers. In WCDMA, there are often indistinguishable multipaths. This introduces mismatches between the computed RoT based on load and the measured actual RoT. Consequently, putting $\hat{L}_{cell}(n)$ to a fixed threshold would lead to conservative parameter settings in order to maintain system stability. This hampers high capacity.

C. RoT measurements in asynchronous WCDMA uplink

To measure $N_0$, in 1x EV-DO reverse link[8], a silence interval is introduced during which all users turn down their transmit power so that $N_0$ is measured through the total received power. This scheme is not applicable to WCDMA systems for the lack of system synchronization. On the other hand, there is typically a sideband between two WCDMA carriers. With out-of-band emission strictly regulated [15], the receiver power in the sideband is seen as invariant to the received power in the signal band. We have designed a $N_0$ measurement scheme based on this principle. The input to the algorithm is samples before the receiver pulse-shaping filter so that the dominant thermal noise from the RF front end within the sideband is preserved. FFT-based histogram is computed to obtain the sideband power spectral density and $N_0$. The total received signal power can be similarly computed after weighting the power spectrum by the pulse filter. Lab implementation of this algorithm showed very accurate RoT measurement with modest complexity. Due to the space limitation, our RoT measurement scheme will not be presented in detail here. Hereafter we assume the measured RoT is available for scheduling.

D. Uplink scheduler with measured RoT

If RoT measurement is available, a target can be directly specified on the RoT itself, or equivalently on total load. The filtered values of $\hat{L}_{as,AS}(n)$ and $L_{out}(n)$ are subtracted out from the target total load. The left portion of the target total load is available for scheduling. Thus, the dynamics in $L_{out}(n)$ and $f_{cell}$ are captured by the scheduler. Its benefit will be shown later in this paper.

E. Interference cancellation and uplink scheduling

Interference cancellation (IC) can greatly increase the spectral efficiency of CDMA uplink. The pilot interference is cancelled before data processing using pilot based channel estimations. The interference from overhead channels can either be cancelled before or after the data processing. For data channel IC, we follow the IC architecture introduced in [6]. The successive interference cancellation (SIC) is implemented by staggering user transmissions. Their equivalence is shown in the following lemma.

**Lemma** In AWGN channel with perfect power control and enough HARQ resolution, IC with staggered user transmission with equal received power achieves the same performance as SIC with equal data rates among UEs.

**Proof** Suppose there are M UEs with common SINR target $\gamma$. The $i$th UE decoded in SIC has $(E_s) = N_0 \gamma (1 + \gamma)^{\mu-i}$. The total received power is $N_0 [(1 + \gamma)^\mu - 1]$. If we let all UEs to have equal power $N_0 [(1 + \gamma)^\mu - 1]/M$ and stagger their transmissions, all packets are decoded at $M$th subpacket where subpacket $i$ sees post-IC interference $N_0 + N_0 (M-i)[(1 + \gamma)^\mu - 1]/M$ and the overall achievable rate is

\[
\sum_{i=0}^{M-1} \log_2 \left\{ \frac{1 + \gamma^\mu - 1}{(M-i)[(1 + \gamma)^\mu - 1]} \right\} = M \log_2 (1 + \gamma).
\]

Therefore the total rate achieved is the same as that of SIC. With finite HARQ resolution in practice, the UEs are divided into groups based on subpacket index and successive cancellation is applied within each group. The full SIC performance can be achieved on average under the same conditions in the above lemma. In our studies with fading channel, realistic power control and the actual Turbo codes, it is found that grouping by subpacket index and 2 iterations within each group provides excellent IC performance very close to that of full SIC. More advanced decoding grouping is possible with additional information on the likelihood of decoding and latency requirements.

After a packet is decoded, the user’s waveform is reconstructed based on the MMSE principle [6] where a shrinkage of $(N_0 E_s / N_0) / (1 + N_0 E_s / N_0)$ is applied to the per-finger channel estimates where $N_0$ is the processing gain. This is to minimize the effect of estimation error on the reconstructed waveform. To improve the accuracy, the channel estimation is based on the re-encoded data sequences. Interference cancellation of the pilot and overhead channels can also be adjusted based on this improved channel estimation. On average, the proportion of residual power is close to $1/(1+N_0 E_s / N_0)$ which is inversely proportional to the received SINR.

With IC, the original measured RoT is no longer the correct account of the interference seen by the data packets. Therefore, the cancelled interference should be reflected in the threshold in scheduling. For this purpose, with our IC architecture, we define the “Effective RoT” as follows:

\[
I_{0,eff}(n) = \frac{1}{12} \sum_{i=0}^{24} I_i (n-i) \left( [\frac{(n-i)}{3}] - \left[ \frac{n}{3} \right] \right) \bmod 8 = 0,
\]

where $I_i(n-i)$ is the residual interference after IC during the time slot $(n-i)$ and $I_i(n-i) \cdot \left( [\frac{(n-i)}{3}] - \left[ \frac{n}{3} \right] \right) \bmod 8 = 0$ is the indicator function for slot $(n-i)$ and $n$ on the same HARQ interlace. The cancellation buffer spans 75 slots which is the maximum packet span with termination target 4. Since more interference
is cancelled with longer delay. $I_o(n-i') \leq I_o(n-i)$ if $n-i'>n$. In other words, $I_{o,\text{eff}}(n)$ is a dynamic parameter reflecting the IC benefit in real time. Defining the ratio between $I_{o,\text{eff}}(n)$ and filtered measured RoT, denoted by $\tilde{I}_o(n)$ as $\tilde{\lambda}(n) = I_{o,\text{eff}}(n)/\tilde{I}_o(n)$, the scheduler has a revised threshold on the total load

This can also be inferred from the per-cell RoT distribution as plotted in Figure 2. The RoT CCDF without RoT measurement is much more spread out than those with RoT measurements. Due to space limitation, field results and simulations with imbalance are not shown here.

**B. Scheduling with effective RoT**

When IC is present, the stability criterion is the 1% tail of the effective RoT below 7 dB. The actual RoT is higher. The gains in system capacity with IC come from two sources: one is the link gain due to the reduction in Ecp/Nt driven by the outer-loop power control; the other is the MAC gain from scheduling based on the effective RoT rather than the actual RoT. The Ecp/Nt reduction is the result of cancelled interference and is not dependent on the scheduler. Henceforth, we need to isolate the gain from Ecp/Nt reduction in order to assess the gain solely from the effective RoT scheduling. In Figure 3, we plot the throughput performance at different probability of RoT>7dB for 3 different scheduling and IC scenarios: scheduling with measured RoT with IC off (baseline), scheduling with the measured and actual RoT with IC on (IC blind scheduler) and scheduling with the measured RoT and using the effective RoT with IC on. The curve with effective RoT scheduler was plotted against the probability of effective RoT exceeding 7 dB. At 7 dB 1% tail, the IC blind scheduler provides a gain of 27% over the baseline whereas the effective RoT scheduler provides a gain of 54%. Obviously, the effective RoT based scheduler is needed in order for maximum IC gains.

In Figure 3, we plot the complimentary-accumulative-distribution (CCDF) of the actual and effective RoT in different scheduling and IC scenarios. As we can see, the IC blind scheduler has similar true RoT distribution with the no IC case although the throughput is 27% higher. The effective RoT scheduler has the effective RoT CCDF similar to the actual RoT CCDF with the no IC case. The link budget is maintained since data packets from the cell edge users experiences the effective RoT as its interference. This can be seen from the user throughput in Figure 4 where almost all the users show throughput increase with IC.

The average Ecp/Nt reduction is the same as the average difference between $I_{o,\text{eff}}(n)$ and $I_o(n)$. The Ecp/Nt reduction is seen in Figure 3 where the CCDF of each user’s mean Ecp/Nt shows a parallel shift of 1.6 dB when IC is on. The amount of 1.6 dB interference reduction can be divided into two sources: nearly all the power from pilot and overhead channels are cancelled; the power from the traffic channel is cancelled after the decoding delay. Let’s divide the users into 4 groups: serving users not in soft handover, serving-users in soft handover, non-serving users who has this cell in the active set and others. The amount of traffic cancellation depends on the composition in the received power from these groups and the thermal noise. If all the users are decoded at 4\textsuperscript{th} subpacket, the effective interference seen by a packet is half of the total traffic interference. With early terminations, the effective interference seen by a packet decoded at the 4\textsuperscript{th} subpacket is even less. The early termination probabilities observed in our simulation is around \{0.35,0.35,0.25,0.05\} for the 4 subpackets. An easy counting shows such a packet will see only 25% of traffic interference from decoded users. Users in group 1 are 99% 

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>SIMULATION PARAMETERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
<td>Value</td>
</tr>
<tr>
<td>Cell layout</td>
<td>19sites wrap around, 3 cells/site</td>
</tr>
<tr>
<td>Carrier Freq.</td>
<td>2.0 GHz</td>
</tr>
<tr>
<td>Sector Antenna Gain</td>
<td>$\theta_{3dB}$</td>
</tr>
<tr>
<td>BS Ant. Gain &amp; Cable Loss</td>
<td>14.0dB</td>
</tr>
<tr>
<td>BS Front-Back Ratio ($A_n$)</td>
<td>20.0 dB</td>
</tr>
<tr>
<td>Path Loss Model</td>
<td>128.2+37.6*log_{10}(d(km))</td>
</tr>
<tr>
<td>UE Max Output Power</td>
<td>21 dBm</td>
</tr>
<tr>
<td>Shadowing Lognormal Std</td>
<td>8.0 dB</td>
</tr>
<tr>
<td>Channel Type</td>
<td>3GPP Mix</td>
</tr>
</tbody>
</table>

$L_{\text{thresh, total, eff}} = L_{\text{thresh, total}} + (1-\lambda)(1-L_{\text{thresh, total}})$.

Since $\lambda \leq 1$, we have $L_{\text{thresh, total, eff}} \geq L_{\text{thresh, total}}$, i.e., higher load can be allocated among users.

**IV. SIMULATION RESULTS**

In this section, we will show the improved capacity results by using measured RoT and further gains through utilizing effective RoT.

Our simulation follows the evaluation methodology of both 3GPP[10] and 3GPP2[11]. Table I lists some of the common parameters. The widely adopted stability criterion is system-wide RoT 1% tail below 7 dB. The load thresholds are chosen such that this criterion is met. The choice of the threshold is much harder for the scheduler without measured RoT since $f_{cell}$ is unknown in priori. $\sigma = 1$ in the fairness.

**A. Scheduling with and without measured RoT**

The system-wide RoT distributions when RoT 1% is slightly under 7 dB are shown in Figure 1 where we see the RoT distribution for the scheduler with measured RoT is tighter. This leads to higher system capacity or spectral efficiency. This gain is quantified also in Figure 1 to be nearly 10%. The scheduling with measured RoT provides nearly 10%. The mean RoT is around 5.3 dB with measured RoT.

The benefit of measured RoT is more drastic with the presence of downlink-uplink imbalance and delays in active set updates.
decoded. Users in the other 3 groups are less likely to be decoded. The exact likelihood varies with simulation.

Figure 1 Left: System-wide RoT distribution. Right: Throughput versus percentage of RoT above 7 dB.

Figure 2 RoT distributions of each cell.

Figure 3 Left: Throughput versus RoT tails for different schedulers with and without IC. Right: System-wide actual and effective RoT CCDF.

C. Further notes on the setup
If the cell size is larger than in our simulation, less power will come from users in other cells and the capacity gain from IC is higher. Typical IC gains with 2.8 site-to-site distance are around 60-70%.

V. CONCLUSIONS AND FUTURE WORK
In this paper, we briefly introduced the side-band power and FFT based mechanism to measure RoT in the asynchronous WCDMA systems. Using such measurements, the HSUPA scheduling algorithm can achieve significant capacity gain and robustness improvement. IC is implemented to provide substantial capacity gains. With IC, the concept of effective RoT is defined to match the post-IC interference seen by the data packets. Scheduling with the effective RoT rather than the actual RoT is necessary to rake in the maximum IC benefit. Besides, the user priority function in the scheduling in this paper maximizes a family of utility functions and provides multi-user diversity.

The scheduling with IC can be further enhanced to utilize more information on the user location. For instance, different termination targets can be specified for users at different locations. The priority function needs to be extended to include delay for those real time and conversational traffic. These are left for future studies.

REFERENCES