Transmission Scheduling for Mass Transit in Data Wireless Networks

Elisha Y. Rosensweig
School of Computer Science
Tel Aviv University, Tel Aviv, Israel

Hanoch Levy
School of Computer Science
Tel Aviv University, Tel Aviv, Israel

Abstract— **Scheduling of down-link transmission for data applications in wireless networks has been the focus of recent research. Channel-aware scheduling algorithms were shown to achieve significant performance gains by accounting for the time varying reception capacities and exploiting their independence across the mobile users. We deviate from these studies by considering mass-transit systems, in which reception capacities of the mobile users are not independent and are positively correlated and thus pose a potential problem due to the bursty traffic requirements they inflict on the cells. We study the performance of these systems aiming at providing a solution to this bursty traffic, and propose a down-link scheduling algorithm that maximizes the resources allocated to the "standard" mobile users while obeying the delay constraints of the train mobile users. We further propose a new architecture for train wireless support systems, based on spatially separated reception train antennas which achieves significant throughput gains. Analysis of this system shows that the combination of the proposed scheduling algorithm with the proposed architecture results in significant system performance improvement.**

I. Introduction

There has been a growing interest in data applications within wireless networks, due to the increase in systems which support them. One topic, which has been the subject of recent research, is scheduling of down-link transmission for data applications in wireless networks. It has been shown that using channel-aware scheduling on the base-station down-link, where the service is given at each point in time to a mobile user based on his channel capacity, can result in a dramatic increase of both the QoS of each mobile user and the overall system performance. Such an approach has been taken in, amongst others, [1],[2],[3] and [7].

Two factors contribute to such gains. The first is the large degree of variance in channel capacities at the various geographical positions in the cell, affected by the different geographical and physical conditions such as distance and fading. The second is the movement of the users across the cell. Simplistic algorithms may end up spending significant amount of transmission resources (transmission time or slots) attempting to deliver data to users whose capacity is *momentarily* very low and thus become very inefficient. Sophisticated scheduling algorithms, in contrast, account for the user mobility and for the time-varying capacity caused thereby. Such algorithms can therefore reach significant performance gains by attempting to transmit to users with (momentarily) high capacity.

Models of these systems, and stochastic analysis of system capacity, are based in prior literature on the assumption that the channel capacities of the users are, over time, *independent of each other*. Such an assumption implies that at each moment it is likely that some users will have high capacity and some low capacity, and that this situation will probably reverse in future moments. Thus, dynamic channel-aware scheduling algorithms that prioritize (momentarily) high capacity users can achieve high system throughput and QoS gains. Such an assumption is based on the belief that the physical position and movement of the different users are quite independent of each other. This assumption is very reasonable in standard wireless systems. However, the growing scope of wireless services introduces systems and situations where the behaviors of users are strongly (and positively) correlated with each other.

Our interest is in wireless systems where the *users independence assumption* is significantly *violated*, and where the capacities of a significant number of the users are positively and strongly correlated to each other. Such situations arise in 1) Mass crowd (e.g. sport) events, 2) Rush-hour car situations, and 3) Mass-transit systems, in particular train systems. Of the three, we focus on the third, due to its challenging dynamic situations and the popularity of train transit.

The main approach for providing wireless coverage onboard mass-transit vehicles, such as trains, is to equip the mass-transit vehicle with a central antenna unit. The central antenna unit communicates with the terrestrial wireless base-stations in a manner which is similar to the way a standard user does. Additionly, the central antenna unit connects to an internal Onboard Wireless Network (OWN)[1]. Each user on board the train connects to the OWN, and all the wireless traffic is channeled through the central antenna unit to/from the base-station. This is the approach taken by many companies, such as $PointShot$, $Icomera$, $21Net$ and more (for a sample list of these, see [5]). Such an approach solves many difficulties resulting from the train speed and the need to penetrate the external hull of the train. Our interest is in the traffic scheduling conducted at the terrestrial base-station which is the bottleneck of the wireless system and which must concurrently serve both the standard users and the train users.

The introduction of a train system into the wireless system, poses significant challenges on the scheduling of the wireless

[1] In some systems the OWN can be connected also to a satellite system. Such connection is out of the scope for this study and is left for future research

transmission at the terrestrial base-station. The challenges are due to the high number of active users residing on board a train whose total traffic volume can create a significant load on the base-station, since all of them appear in the cell at the same time. Of course one could easily solve this problem by increasing the base-station capacity or adding base-stations. However, since the train remains within the cell for only a small fraction of the time, such an increase is not economical, as the base-station will remain at very low utilization most of the time (e.g. an order of a few tens of seconds). Thus, due to the train passage, the base-station is subject to frequent drastic increases of load (load bursts), during which it must maintain proper QoS for the standard users as well as for the train users.

While from the cell's perspective the load increase is only momentary, from the train's perspective this issue is a persistent one, since in every cell passed by the train, the overall load, during the passage moments, is high. Thus, cell-perspective solutions that might attempt to negatively discriminate the train users, on the grounds that the period of overload is only momentary, will result in very poor performance to the train users for a long duration.

The purpose of this paper is to study this system and deal specifically with its bursty traffic nature. We aim at proposing mechanisms as well as base-station scheduling algorithms for coping with these problems and for providing efficient system operation.

After a description of the model (Section II) we start this work (Section III) with constructing a base-station scheduling algorithm designed to address the problem. We recognize that since the train moves across many cells and shifts significant loads from one cell to another, an optimal solution has to involve all cells at once. This, however, might be too complex a scheduling problem, due to the different (non-train) demands posed on the various cells, and the variability of these demands over time. We therefore propose an approach that decomposes the problem and allows us to deal with it in each cell in isolation. The approach aims at optimizing the scheduling of the train transmissions so as to minimize the number of transmission time dedicated to it in the cell, while obeying the delay constraints of its transmissions. The scheduling accounts for the reception capacity of the train that varies as a function of the distance from the base-station. We propose a scheduling algorithm, called Bounded Slot Delay (BSD), and prove its optimality.

We continue (Section IV) to recognize that the significant length of the train implies that at any given moment in time the reception capacities, at different parts of the train, significantly vary from each other. Aiming at exploiting this property, we propose a new train architecture called Multi-Antenna Spatially-Separated (MASS), in which the train uses a multiplicity of antennas, spread over the train, to which the base-station may transmit, and a switching mechanism that at every moment selects only the antenna with (momentarily) the highest reception level to be operative. Since our focus in this work is on the effects of distance on reception capacity (see Eq. (1), the selection of the reception antenna is based on the antennas' distances from the base-station. Our analysis reveals an optimal rule for placing $n > 1$ antennas on the train. We further derive the gains of a MASS system as a function of n and conclude from this derivation that the largest marginal gain is achieved during by transition from one antenna to $n = 2$ antennas.

We then (Section V) discuss how BSD can be combined with this new MASS architecture, in order to significantly improve the QoS to both train and cell users. Lastly (Section VI) we use numerical results to evaluate the performance of BSD and MASS.

II. Model description

A. Mobile network

A wireless (cellular) network consists of base-stations which are geographically spread, and users. The connection of the user to the network is done via a wireless link, typically to/from the base-station that is closest to the user. This means that during any session, the information transmitted to the user (voice, Web data, file transfer etc.) and from the user reaches the network via the closest base-station.

The effective transmission rate (”throughput”) at which the (closest) base-station can transmit to the user depends on many factors. One central factor is the *distance* between the user and the base-station: the closer the user is to the base-station, the higher the effective rate. Specifically, the impact of distance on the reception level is believed to be modeled by the function

$$Capacity(x) = x^{\alpha} \tag{1}$$

for some $-4 \leq \alpha \leq -2$, where x is the distance of the user from the base-station. Thus, the user, as well as the overall system, can benefit from having the user close to the base-station. Other factors which affect the user reception capacity rate, such as fading, are not dealt with in this paper. Let

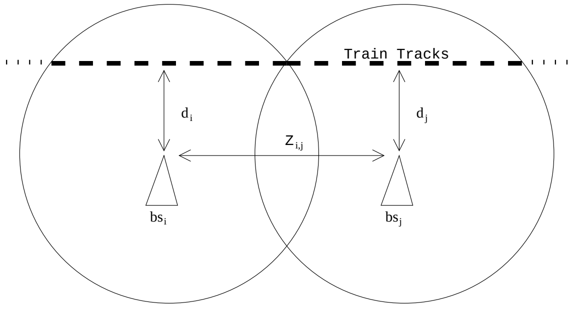

Fig. 1. Model visualization

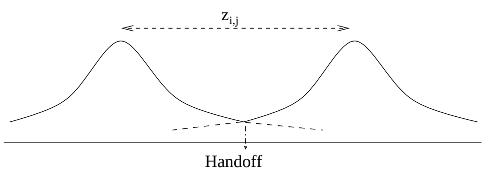

Fig. 2. Train capacity over two cells.

$BS := (bs_i \in \mathbb{R}^2)_{i\in\mathbb{Z}}$ be an infinite list of base-stations. In this paper we limit ourselves to those base-stations which are close enough to train tracks, so that at some point on the tracks they are the closest to the tracks, and thus might be required to support the users on the train. We list the base-stations according to their order along the train tracks, which are assumed to be of infinite length. For each base-station, let d_i represent the minimal (euclidian) distance of bs_i from the train tracks, and for each pair of base-stations let $z_{i,j}$ be the distance between them, sometimes denoted simply z. Figure 1 portrays this model, and Figure 2 reflects the fluctuations in the reception capacity over time as the train moves through the cells. Assuming that the train is moving at approximately constant speed, there exists a linear correlation between the time the train is within the cell and the length of the tracks. We shall therefore commonly use the same variables in referring to both distance and time (measured in transmission slots and the distance passed during one such slot).

B. Users

Mobile users are classified into *cell users*, which are standard mobile users roaming freely in their respective cells, and *train users*, which are train passengers and connect to the OWN on-board the train. All train users can be transmitted to in parallel, yet must share throughput. We therefore treat the train as a single (large) user throughout the paper.

We use a time-slotted model, where time is divided into short transmission slots, and during each slot the base-station transmits to a single user chosen according to some predefined policy. Many such policies exist, a class of which is commonly referred to as "channel-aware" policies. These policies decide which user to transmit to at each slot, based on, among other considerations, the S/N ratio of all users currently requesting service, which reflects the amount of data (packets) which can be transmitted to that user during the specific slot. This information is collected by the base-station at the beginning of each slot.

Each user u is associated with two time-dependent functions. The first is the data-arrival rate denoted $In^u(t)$, representing the amount of packets which arrive at the base-station for transmission to user u at slot t (all packets are assumed to be of equal size). Packets which are not forwarded immediately to u are stored in a buffer until the base-station allocates a slot for transmitting them. Packet transmission is continuous: a packet can be "divided" and transmitted over more than one slot (e.g. half a packet in slot t and the other half in slot $t+1$).

The second is the user's reception capacity at slot t, denoted $C^u(t)$. When a slot is allocated to a specific user, the base-station transmits the buffered packets according to order of arrival, up to the user's reception capacity limit. Furthermore, due to the proportionality between distance and time mentioned earlier, we shall commonly use the notation $C^u(x)$ to refer to the capacity of user u at point x on the tracks.

At several points during the paper, some of these variables shall be treated as constants, or we may refer to their expected value. In such cases they shall be denoted C^u, In^u.

Each user u, is associated with a constant delay constraint w^u, reflecting the maximal delay allowed for its packets, depending its service type. In this paper we limit ourselves to dealing with data transfers, and assume a maximal delay constraint uniformly applied to all users, denoted w_{max}.

III. Efficient transmission Scheduling Algorithm: Load Minimization under Delay Constraints

The introduction of train wireless support into a wireless system poses significant challenges for the scheduling of the wireless transmission at the terrestrial base-station. The reason is that trains can potentially harbor many users simultaneously, and their collective requirements from the wireless system can cause a serious overload to the base-station. Since the train remains within the cell for only a small fraction of time, adding base-stations in the area to support this increase is not economical, as each base-station will remain at very low utilization most of the time.

A scheduling policy that would consider all users (on train or off-train) similarly (e.g. using the approach taken by [1]) will end up negatively discriminating the train users considerably. The reason is that the overload experienced when the train passes in a cell will degrade service to all users in the cell; while for the cell users this will be a very short (and temporary) degradation, for the train users this will be permanent, since they will experience it in every cell.

Furthermore, an overall optimization of the scheduling problem accounting for both the train users and the cell users must account for the train transition across cells and thus must consider all the users in all the network cells. Such an optimization looks intractable.

For these reasons, we propose an alternative approach that decomposes the problem and treats the train in each cell in isolation, focusing at every epoch on the cell in which the train passes. During this period, our approach is to guarantee a certain level of service for the train users, while minimizing the resources allocated for servicing them. As stated earlier, w_{max} denotes the maximal delay allowed for a packet addressed to a train user. Under this constraint, we develop an optimal "semi-online" algorithm which minimizes the amount of transmission slots used by the train. "Semi-online" means that the algorithm relies on the knowledge of the future reception capacity, which is available to the system due to the static nature of the base-station locations and train tracks layout.

The algorithm presented here is developed and analyzed under the assumption that the train-data arrival rate is constant. Applying this algorithm to more dynamic settings is possible with some slight modifications, and is not discussed here.

A. Notation and Preliminaries

We treat the train as a single (large) user Tr and let the data arrival rate be a constant $In^{Tr} = r$. Note that the fact that the train represents a large number of users, combined with the law of large numbers, implies that the total train arrival rate is relatively constant. Without losing generality normalize the train-data arrival rate and the reception capacity by r, yielding for the analysis that $In^{Tr} = 1$, reflecting that at every given slot a single new packet arrives at the base-station for transmission to the train. The scaled value $C^{Tr}(t)$ now represents the amount of packets transmittable in slot t (which now is not necessarily an integer).

Let $B^u_{alg}(t)$ denote the number of packets not yet transmitted by algorithm alg to user u by slot t (inclusive), and let $Service_{alg}(t)$ denote the list of packet indices which are transmitted by alg during slot t. Note that due to the normalization stated above, $B^u_{alg}(t)$ may contain one index to a packet fraction and $Service_{alg}(t)$ may contain two such fraction indices. Also, let $Ex^u(t) := \max\{0, C^u(t) - In^u\}$ be the *excess capacity* at slot t, and in similar fashion, let

$$Ex^u[t_1, t_2] = \min\{\sum_{t=t_1}^{t_2} C^u(t) - (t_2 - t_1 + 1), w_{max}\}$$

be the maximal amount of additional packets that can be transmitted legally (meet the delay constraints) between slots t_1 and t_2 (inclusive), in addition to packets $(t_1, ..., t_2)$.

Definition 1: A *user service allocation* is a vector $A \in \{0,1\}^T$ where T is the number of slots available for allocation and $A(t) = 1$ iff the user is transmitted to at slot t. For some user service allocation A, $|A|$ denotes the number of slots marked for service in user service allocation A. A *service block* in user service allocation A is a set of consecutive slots, all of which are marked for transmission (1). We denote a service block from slot t_1 to slot t_2 as $A[t_1, t_2]$.

A *legal user service allocation* is a service allocation which obeys the delay constraints. Specifically, for any t s.t. $A(t) = 1$ in a legal user service allocation A, the base-station transmits

$$\min\{C^{Tr}(t), B^u_A(t), w_{max}\}$$

packets at slot t according to the order of arrival (FIFO). Note that by $B^u_A(t)$ we refer to the amount of buffered packets at slot t (inclusive) when using the user service allocation A. Throughout this section we shall only be referring to train service allocations and thus omit to mention this explicitly.

We conclude this sub-section with the following lemma, used frequently throughout our discussion of the load minimization under delay constraints problem:

Lemma 1: Let $Ex^u[t_1, t_2] = w_1 \leq w_{max}$ and for all $t_1 \leq t \leq t_2$ let $C^u(t) \geq In^u$. Also, let $B^u_{alg}(t_1) = w_2$, and denote $w = \min\{w_1, w_2\}$. Then $A[t_1, t_2]$ can transmit legally $((t_2 - t_1 + 1) + w) \cdot In^u$ packets.

Proof: The proof is shown for $In^u = 1$, and the generalization is trivial. Obviously, $A[t_1, t_2]$ can transmit only packets which have arrived by t_2 and have not yet been transmitted. Thus, the amount of packets which can be transmitted is bounded by $(t_2 - t_1 + 1) + w_2$. In similar fashion, the amount of packets that can be transmitted by $A[t_1, t_2]$ is bounded by $(t_2 - t_1 + 1) + Ex^u[t_1, t_2] = (t_2 - t_1 + 1) + w_1$, so we got the upper bound. From here on let $w = w_1 \leq w_2$, and we show that exactly $(t_2 - t_1 + 1 + w)$ packets can be transmitted during this service block, without violation of their delay constraints.

Since $In^u = 1$, at every slot a single packet enters the buffer. At the same time, at every slot $C^u(t) \geq In^u$, so at each slot at least one packet is transmitted, which ensures that the buffer will never grow beyond it's size at the beginning of t_1. Thus, if at slot t_1 the buffer is of size $w_2 \leq w_{max}$, then, due to FIFO, the delay constraints will not be violated until the end of the service allocation block. Also, until the buffer is emptied, at each slot the capacity will be utilized fully. From our assumption that $w = w_1 \leq w_2$, the buffer will not empty before the end of t_2, so a total of $(t_2 - t_1 + 1 + w)$ packets are transmitted. ■

B. Basic Properties

Though we have imposed no explicit limitations on the relationship between the data arrival rate and the reception capacity values, the introduction of delay constraints into the model allows to limit discussion to specific capacity functions:

Theorem 1: Given a reception capacity function $C^{Tr}()$ and assuming $In^{Tr} = 1$, define a *delay-bounded capacity function* as

$$C^{Tr}_*(t) = \min\{C^{Tr}(t), w_{max} \cdot In^{Tr}\} \tag{2}$$

Then A is a legal service allocation w.r.t C^{Tr} iff A is a legal service allocation w.r.t C^{Tr}_*

Proof: Due to the delay constraint, in any legal service allocation A all slots transmit up to $w_{max} \cdot In^{Tr}$ packets, since otherwise there is a packet which arrived more than w_{max} slots earlier. This implies that if $C^{Tr}(t) > w_{max} \cdot In^{Tr}$, the excess capacity cannot be put to use and thus has no effect on the resulting service allocation. ■

We therefore limit ourselves to delay-bounded capacity functions. Given a general capacity function, all capacity which exceeds $w_{max} \cdot In^{Tr}$ is discarded.

Each slot is associated with the cell within which it takes place. For each cell i define S^i_j to be a group of the j slots with highest capacity within the cell, using as a tie-breaker the rule that earlier slots are chosen over later slots. Since the reception capacity of the train is a function of its distance from the base station, and we assume the train tracks are laid out in a straight line, S^i_j is always a continuous block of slots, normally in the middle of the duration in cell[2]. We further denote S^i_{max} to be $S^i_k = \{t_1, ..., t_k\}$ such that k is the minimal index for which

$$Ex^{Tr}[t_1, t_k] = w_{max}. \tag{3}$$

It is possible, theoretically, that such a k does not exist. This might occur in cases where the data arrival rate is relatively high and/or when the delay constraint is very relaxed (i.e. w_{max} is very large). However, such cases are impractical and uncommon in general. High data arrival rates are usually not manageable within reasonable delay constraints, and extremely relaxed delay constraints tend to lower the QoS of users. In this paper we assume, therefore, that such a k exists.

[2] Recall we are dealing here in delay-bound capacity functions. If the original function was not delay bound, we still choose the slots to make up S_{max} according to the order in which they were in the original function.

Observation 1: $C^{Tr}(t) > In^{Tr}$ for all $t \in S_{max}$, since for any slot t in which $C^{Tr}(t) \leq In^{Tr}$ we know $Ex^{Tr}(t) = 0$, so from minimality of $|S_{max}|$ such a slot would not be included in F_{max}.

Observation 2: For any $S^i_j \subseteq S^i_{max}$, $S^i_j = \{t_1, ..., t_j\}$, the minimal amount of slots required for transmitting $j + Ex^{Tr}[t_1, t_j]$ packets within cell i is j. The reason for this is that, on the one hand, if the buffer has at least $Ex^{Tr}[t_1, t_j]$ packets at slot t_1 then S^i_j can transmit all these packets and $t_1, ..., t_j$; and on the other hand, the slots in S^i_j are the highest j slots in the cell in terms of reception capacity, so any other group of j slots cannot do any better. Also, such a transmission is possible due to Lemma 1, relying on Observation 1 to ensure the requirements of the lemma.

S^i_{max} can transmit packets $t_1 - w_{max}, ..., t_k$. Due to observation 2, an optimal service allocation for cell i *alone* would allocate S^i_{max} for transmission: it is the optimal allocation for all the packets transmitted in it, and furthermore due to the fact that it reaches the limit of delay constraints, no other packets within the cell could have been transmitted by any of the slots in S^i_{max}. It also ensures the buffer to be empty at the end of the service block.

However, it is possible that in some cases, not all of S^i_{max} would be used for transmission. There are two possible causes for this. First, the last slot in S^i_{max} may have more capacity than required for transmitting the remaining buffer, so it might be preferable to postpone transmission to a later slot. Such a scenario results from the discreteness of the model and could be avoided by moving to a continuous model and can, at most, reduce the number of slots allocated by one. Second, some packets $t_{k-x}, ..., t_k$, might be transmitted more efficiently in some future cell which has slots of higher capacity.

The second case is essentially a form of packet postponing: the system realizes that though it could transmit all packets in the buffer with S^i_{max}, by refraining from this the overall results will be better due to future capacity values.

However, packet postponing can be optimal only if there are some slots within w_{max} slots after $t_k \in S^i_{max}$ with higher capacity than t_k. Such cases are very rare when dealing with reasonable values for w_{max} in normal sized cells, which would usually equal up to $1/3$ of the slots in the cell, since S^i_{max} is usually tightly clustered around the center of the cell. We therefore disregard such cases and assume a local point of view throughout the paper, and assume S^i_{max} is allocated for use in any cell i.

C. Bounded Slot Delay algorithm (BSD)

Figure 2 qualitatively depicts the train reception corresponding to the track cell structure given in Figure 1, based on the assumption that the tracks form approximately a straight line through the cell. The reception level increases as the train approaches the base-station and drops as it moves away from it. The reception level receives local minimum on the intersection points between the cells where hand-off between the cells is assumed to occur. In a multi-cell environment the train goes through cycles of such reception level increase and decrease. It should be noted that while the figure relates to two symmetric cells, our analysis will not be limited to such situations.

Our algorithm is constructed by dividing the cycle displayed in Figure 2 into four segments, and finding the optimal allocation for each segment. Optimality is proven, at each segment, relative to the whole cycle depicted in Figure 3, so the resulting algorithm is optimal. The segments are (See Figure 3): a) Segment A - monotonically decreasing, b) Segment B - all slots around t_{min} s.t. $C^{Tr}(t) < In^{Tr}$, c) Segment C - monotonically increasing, and d) Segment D - S_{max}. Segments B and D are well defined. Segment A is then all the slots from the end of segment D to the beginning of segment B, and segment C is then all the slots from the end of segment B to the beginning of segment D.

Our algorithm is designed to be the optimal slot allocation over a single such cycle, where we assume to begin with a empty buffer and ensure an empty buffer at the end. Thus, optimal allocations of consecutive cells can be combined easily. Note that though, as stated above, in the optimal algorithm we might refrain from using the last slot in S_{max}, this can only reduce the overall slot allocation by one, so we disregard this in the sake of clarity.

The basic intuition behind the algorithm, which is described next, is as follows. While the train has low capacity, it transmits as little as possible and allows the buffer to grow. When, on the other hand, capacity is very high (S_{max}), it transmits all the buffered packets and the new incoming packets. In between these two extremes, the algorithm strives to keep the situation balanced: after the low capacity segment it retains a close-to-full buffer, and after the high capacity segment it retains a close-to-empty buffer.

From the definition of segment B, and as can be seen in Figure 3, segments A, C, and D contain only slots for which the following condition it met:

$$C^{Tr}(t) \geq In^{Tr}. \tag{4}$$

Lemma 2: Let C^{Tr} be a monotonically increasing function over slots $t_{1+w_{max}}, ..., t_{P+w_{max}}$, which conforms to Equation (4) for all t in the segment, and assume that for all $t_1 \leq t \leq t_{w_{max}}$ we have $C^{Tr}(t) \leq C^{Tr}(t_{1+w_{max}})$. Also, assume the buffer is empty at slot t_1. We want to use a minimal amount of slots throughout the segment without violating the delay constraints, but otherwise without any regard to the final condition of the buffer. Then the following policy is optimal: Refrain from transmitting as long as delay constraints are not

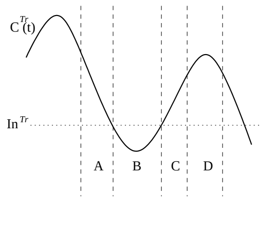

Fig. 3. Capacity Cycle Segmentation

violated for some packet (*in the next slot*).

Proof: First we show that this policy yields a legal service allocation. At every slot at most a single packet may enter the buffer $(In^{Tr}(t) = 1)$, thus each packet violates the delay constraints in a different deadline. Since from (4) we know that from slot $t_1 + w_{max}$ every single packet can be transmitted in a single slot, applying the above policy cannot violate the delay constraints, since the buffer is empty at slot t_1 and thus the first violation is possible only after slot t_1+w_{max}.

Next we prove optimality. For each slot t s.t. $A_{BSD}(t) = 1$, denote by $first(t)$ the first packet transmitted at slot t, which then defines the set $FIRST = \{first(t)\}_{t:A_{BSD}(t)=1}$. Note that since from each transmitted slot we take a single representative, we get $|A_{BSD}| = |FIRST|$. We prove the lemma by showing that for every legal service allocation, every packet in $FIRST$ is served in a separate slot, and thus the number of slots our algorithms uses is optimal.

Assume by way of contradiction there were two packets $\{p_1, p_2\} \subseteq FIRST$, $p_1 < p_2$, s.t. $\{p_1, p_2\} \subseteq Service_A(s)$ for some slot s, and let t_{p_k} be the slot which served p_k in BSD. Recall that packets are served using FIFO policy by both allocations, so in order to serve both p_1 and p_2, slot s must comply to

$$Service_{BSD}(s) \geq p_2 - p_1 + 1$$

Since BSD served p_1 when the delay constraint was reached, and s must come no earlier than p_2 this means $p_2 \leq s \leq t_{p_1}$ and since the capacity is monotonically increasing, we get

$$p_2 - p_1 + 1 \leq C^{Tr}(s) \leq C^{Tr}(t_{p_1})$$

from which we can conclude that BSD should also serve p_2 at slot t_{p_1} - contradicting our assumption that $p_2 = first(t_{p_2})$. ■

The lemma implies that no transmission will occur during slots $t_1, ..., t_{w_{max}}$ so the following also holds:

Lemma 3: Assume the conditions for Lemma 2 are met, except for the fact that at slot t_1 the buffer size is b, and once again assume the final buffer size is of no consequence. Then by using the policy described in Lemma 2 for the case of an empty buffer, combined with transmitting these b buffered packets over slots $t_1, ..., t_{w_{max}}$ in optimal fashion, results in an optimal service allocation.

Proof: None of the b buffered packets can be transmitted after $t_{w_{max}}$ without violating their delay constraints, so they must be transmitted over $t_1, ..., t_{w_{max}}$. From Lemma 2, all packets $t_1, ..., t_{w_{max}}$ are better off being transmitted after $t_{w_{max}}$, so using the same allocation from Lemma 2 is optimal. Since the two allocations do not intersect, the combination is optimal. ■

Since we know that S_{max} is allocated for service, which by definition can transmit all packets in a buffer of legal size, we can apply Lemma 3 on slots that precede S_{max}, working back from segment C and into previous segments. Specifically, if segment B has $> w_{max}$ slots, let P be the size of segment C and the preceding slots $t_1, ..., t_{w_{max}}$ be the last w_{max} slots in segment B. Relying on the fact that segment C complies with Eq. (4) and segment B does not, the lemma is applicable, as long as any buffered packets at slot t_1 are cleared by $t_{w_{max}}$. This is ensured by the following lemma:

Lemma 4: **(Segment B)** Let segment B be the sequence of slots within a single cycle for which the capacity is less than the train-data arrival rate. Formally, let t_h, t_j be two slots s.t. for all $t_h \leq t \leq t_j$, $C^{Tr}(t) < In^{Tr}$, and $t_j - t_h$ is maximal. Let $x = (t_j - t_h) - w_{max}$ for some $x \in \mathbb{Z}$. Assuming the buffer is empty at the beginning of the segment, Then: a) A feasible service allocation exists only if

$$\sum_{t_h}^{t_j} C^{Tr}(t) \geq x, \tag{5}$$

and b) The optimal allocation would transmit no more than x packets until the end of the segment, by using all the highest capacity slots in the segment.

Proof: a) If $x \leq 0$ the condition is met easily, and no packets are transmitted. Otherwise, assume by way of contradiction $\sum_{t_h}^{t_j} C^{Tr}(t) < x$, then packet $t_h + (x)$ will be transmitted only after t_j. This means the delay constraint was violated, so no legal service allocation is possible.

b) To prove optimality we need to show that the optimal allocation would not require to transmit more than x slots. Assuming segment B has at least w_{max} slots, we treat two cases:

1) If segment C is of size 0, segments B and D are adjacent. Segment D can clear the buffer in optimal fashion, so the allocation in B is optimal.
2) Otherwise, segment B must transmit all but the last w_{max} packets in the buffer. Let $t_1, ..., t_{w_{max}}$ be the last packets in segment B, applying the policy in Lemma 3 is optimal.

The case in which segment B has less than w_{max} slots is considered later on in this section.

The optimal transmission of these x packets is achieved by using the slot with highest capacity in Segment B, the proof of which is simple and left out due to lack of space. ■

Lemma 5: **(Segment A)** Assume C^{Tr} is a monotonically decreasing function over slots $t_1, ..., t_P$, which conforms to Equation (4) for all t in the segment. Assume as well that $C^{Tr}(t_1) \geq B^{Tr}_{BSD}(t_1)$. The optimal transmission policy is: (a) Transmit at slot t such that $C^{Tr}(t+1) < B^{Tr}_{BSD}(t_1)$ (b) Cease transmission for the rest of the segment starting from the first slot t which begins with an empty buffer and $C^{Tr}(t) \leq C^{Tr}(t + w_{max})$.

Proof: Let A_{BSD} be the service allocation defined by the above policy. We need to prove that A_{BSD} is a legal service allocation, and that it is optimal, i.e. it uses the minimal amount of slots.

Legality is obvious: we begin with a slot which has the ability to transmit all the buffer, and since we conform to Equation (4), we always transmit before the buffer exceeds the capacity. Since we are limited to delay-bound capacity values, this implies that the buffer never exceeds w_{max} packets.

Regarding optimality, if condition (b) is met, since we stop at the first such occurrence we know that slot $t + w_{max}$ is part of a monotonically increasing sequence of slots which also conform to Equation (4), and the inequality meets the requirements of Lemma 2, so the optimal allocation is achieved by refraining from transmission. Otherwise, let us analyze the alternate possibilities for transmitting the buffer: transmitting earlier would imply that we transmitted less packets for a single slot, so we gain nothing by doing so. Transmitting later would require at least 2 slots just to transmit the packets buffered at slot t, since for all slots $t + 1, ..., t + w_{max}$ the capacity is lower (condition (b) was not met). Since these same packets could be transmitted in a single slot at slot t, there is no loss in doing so at the earlier slot. ■

If condition (a) is met somewhere in segment A, this implies that segment B is of size $< w_{max}$. Otherwise, we continue to perform the policy in condition (b) for transmission in segment A. In such a case, one interesting result of the above transmission policy is that at the end of segment A the buffer is empty: : At least one packet enters the buffer at the beginning of the last slot of segment A, so the amount of buffered packets is larger than the capacity of all slots in segment B, and so according to the above policy the buffer shall once again be emptied right before Segment B. We therefore can apply Lemma 4on Segment B which has been proven to be optimal.

The overall description of BSD is next given.

BSD Description: 1. Considering the train-data arrival rate, allocate a central transmission block $D = S_{max}$. 2. Considering the train-data arrival rate, allocate slots according to Lemma 4 for segment B. 3. What remains are segments A, C, which are served according to Lemma 5 and Lemma 3. Note that the optimality of each stage was not local but global, so we get the following theorem, which concludes our claims:

Theorem 2: BSD is optimal over the combination of segments A through D.

The dynamic nature of BSD allows for the train-data arrival rate to vary. As the train enters the cell, the base-station monitors the data arrival rate and computes its mean, which it then feeds to the above chart. Assuming the variance of the data arrival rate is not too large, the service allocation will remain close to optimal.

IV. The MASS Architecture for Capacity Increase

A. MASS Architecture

As described in the introduction, we propose the *Multi-Antenna Spatially-Separated (MASS)* architecture. It consists of a multiplicity of reception antennas placed on the external train hull, as shown in Figure 4, and transmitting at each allocated slot *only* to the antenna with the highest reception capacity. Denote these antennas as $A_0, ..., A_{n-1}$, indexed in order of placement. Recall that the train tracks are assumed to coincide with the x-axis, and let $F(t) = (x_t, 0)$ be the position of the head of the train at time t. We make use here of the notation previously defined to denote by $C^i_j(t)$ the reception capacity of antenna i relative to base-station j.

The relative distance between train antenna and the head of the train is set, so the following is well defined:

Definition 2: An antenna deployment is a sequence $Dep =< d(F, A_0), ..., d(F, A_{n-1}) >$ where $d(p_1, p_2)$ is the euclidian distance between two points in $\mathbb{R}^2$.
Note that we use the term A_i to denote both the antenna and its position in $\mathbb{R}^2$. Given a position $F(t)$, and assuming the layout of the train tracks is of a straight line parallel to the x-axis, the position of the i-th antenna at time t is: $A_i(t) := F(t) - (Dep(i), 0)$.

Let $C_j(p)$ be the reception capacity available for bs_j at the point $p = (x, y)$. From Equation (1) we have

$$C_j(p) = Capacity(d(p, bs_j)) = d(p, bs_j)^{\alpha} \qquad (6)$$

and denote accordingly $C(p) := \max_{j\in\mathbb{Z}} C_j(p)$ as the available transmission capacity at point p.

Under MASS n antennas are mounted on the train. The reception capacity of antenna A_i will be $C^i(t) := \max_{j\in\mathbb{Z}} C^i_j(A_i(t))$ and since we transmit only to the train antenna with highest reception capacity, the overall capacity of the train at time t is defined as $C^{Tr}(t) = \max_{0\leq i\leq n-1} C^i(t)$.

B. MASS analysis - symmetric model

Our objective is to analyze the performance gains of MASS and to derive the optimal antenna placement. The optimal placement depends on the exact route taken by the train and the specific positioning of the surrounding base-stations. Such a solution is not manageable for several reasons, among which are the large amount of cells through which the train may pass, the performance sensitivity of a ”tailored made” design to changes in the network structure, and the fact that the train antennas must be relatively static. Instead, a more stable design, which would be more simple to devise per network and an not too sensitive to network changes, would be more suitable.

To this end we analyze a simplified model, which assumes: 1) All base-stations are at an equal distance d from the train tracks, and w.l.o.g. are positioned on the same side of the tracks. They thus coincide with the line $y = d$, parallel to the x-axis, and 2) Distances between neighboring stations are all equal to each other (z).

Specifically, let $bs_i = ((i + \frac{1}{2}) \cdot z, d)$. A model where these conditions are met will be called a *symmetric model*, and denoted SYM^d_z. We shall relax these constraints later on during numerical evaluation of the applicability of this model.

By applying simple geometry, these requirements imply that for each cell there is a stretch of train tracks of length z within its scope, and each point on the train spends exactly z slots within each cell.

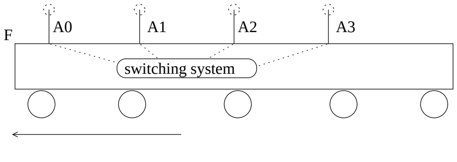

Fig. 4. MASS architecture

Lemma 6: Let $x = j \cdot z + h$ for some $j \in \mathbb{Z}$ and $0 < h < z$, then $C(x) = C_j(x)$

Proof: Since the reception capacity decreases with the distance, the maximal capacity is achieved for the closest base-station, which is bs_j. ■

Observe that, since the base-stations are placed at equal distances of z from one another, the maximal capacity function is periodic with a period of z:

$$C(x) = C(x+z) = C(x \, mod \, z) \tag{7}$$

Since we assume that the train, during transit, is moving at constant speed, the same can be claimed substituting distance traveled with time passed. This last observation leads us to the following lemma, implying that for any arbitrary antenna deployment there exists a deployment that is as efficient, yet its span is at most z:

Lemma 7: Let $Dep =< d(F, A_0), ..., d(F, A_{n-1}) >$ be an arbitrary deployment, then there exists a deployment $Dep' =< d(F, A'_0), ..., d(F, A'_{n-1}) >$ such that

1) for all $0 \leq i \leq n-1, \;\; d(F, A'_i) \leq z$.
2) there exists a permutation π over $\{0, ..., n-1\}$ such that

$$\forall t \; \forall \, 0 \leq i \leq n-1 \;\; C^i(t) = C^{\pi(i)}(t)'$$

Proof: The values of such a deployment will be $\{x \;\; mod \;\; z \quad : x \in Dep\}$ indexed in ascending order s.t. $A'_i \leq A'_{i+1}$. By definition, all values in Dep' are smaller than z. Condition 2 is also met from Eq. (7), from which we know we can match up $A_i \rightarrow A'_j = A_i \, mod \, z$. ■

This lemma leads to the following key result:

Theorem 3: In SYM^d_z, all possible capacity configurations can be achieved on a train of length z.

The cyclic nature of the capacity function also allows us to replace the symmetric model with an equivalent single-cell model, which we define here:

Definition 3: MOD^d_z is a single-cell model with train tracks of length z which are connected at their edges, i.e. when the edge of the tracks is reached the train reenters the cell at the beginning of the tracks. Additionally the distance from both edges of train tracks to the base-station is equal.

Lemma 8: MOD^d_z and SYM^d_z yield identical capacity functions for all train antennas.

Proof: Examine train antenna i. Previously we observed that $C^i(t) = C^i(t \, mod \, z)$, which is exactly the capacity function of A_i in the MOD^d_z model. ■

We rely on Lemma 8 and move to consider the MOD^d_z model. Due to the nature of the Modulo-model, and the limited distance over which train antennas can be deployed (Theorem 3), the function $C^{Tr}(t)$ repeats itself every z slots.

Definition 4: $Ant(i) \subseteq \{1...z\}$ is the group of slots during which the train switches to train antenna A_i for reception.

Lemma 9: Let A_i and A_j be non-adjacent train antennas. By moving the position of A_i between its neighbors no modifications are made to $Ant(j)$ in MOD^d_z.

Proof: First we note that, by moving A_i, slots can either be allocated to or from A_i, but are never moved between two other train antennas. This is due to the fact that slots are allocated solely according to the distance factor, so by moving A_i distances can change only relative to A_i.

Let A_h be a train antenna positioned between A_i and A_j, and let us look at the position of these antennas at some slot t_0, which changed allocation as a result of the move:

- If $t_0 \in Ant(j)$, this means A_j is closer to the base-station than A_h, and thus also closer than A_i after the move, which contradicts the assumption that t_0 is reallocated to A_i after the shift.
- Otherwise, originally $t_0 \in Ant(i)$ and let us assume that in the new configuration $t_0 \in Ant'(j)$. This implies that there is a position of A_i between it's neighbors for which at slot t_0 the distance between the base-station and both A_i and A_j is equal. Since A_h is between these two, we know that at t_0 A_h is closer to the base-station than both of them, which contradicts the assumption that A_j is the closest at t_0 after the shift.

■

We use this lemma to asses the effect of moving a train antenna on the overall train reception capacity throughout the cell:

Lemma 10: Let $A = \{A_1, A_2, A_3\}$ be three adjacent train antennas in the MOD^d_z model, placed according to index (A_1 and A_3 might be the same antenna). Let the distance $d(A_1, A_3)$ be set, and assume $d(A_1, A_2) - d(A_2, A_3) = \epsilon > 0$. For some $\delta \leq \epsilon/2$, define $A'_2 := A_2 - \delta$ an antenna position which is closer to A_1 than A_2. Then using A'_2 instead of A_2 improves the performance of the train antenna array, in the sense that integration over $C^{Tr}(t)$ will increase overall.

Proof: From Lemma (9), the exact position of A_2 can affect the values of $C^{Tr}(t)$ only between the peak capacities of A_1 and A_3. So let $p_1 = d(A_1, A_2)/2$ and $p_2 = d(A_2, A_3)/2$. The performance gain by repositioning as described is

$$-\frac{2}{\delta}\int_{p_1-\delta/2}^{p_1}(d^2+x^2)^{-\alpha/2}dx + \frac{2}{\delta}\int_{p_2}^{p_2+\delta/2}(d^2+x^2)^{-\alpha/2}dx =$$

$$-\frac{2}{\delta}\int_{p_1-\delta/2}^{p_1}(d^2+x^2)^{-\alpha/2}dx + \frac{2}{\delta}\int_{p_1-\epsilon/2}^{p_1-\epsilon/2+\delta/2}(d^2+x^2)^{-\alpha/2}dx =$$

$$-\frac{2}{\delta}\int_{p_1-\delta/2}^{p_1}[(d^2+x^2)^{-\alpha/2} - (d^2+(x+(-\epsilon+\delta)/2)^2)^{-\alpha/2}]dx,$$

which is greater than zero since $\epsilon > \delta$ and $\alpha \geq 2$. ■

These lead to the key result of this section stating the optimal antenna deployment strategy:

Theorem 4: (1) The optimal deployment of train antennas over a train of length $l_n \geq z \cdot (1 - \frac{1}{n})$ is a uniform deployment s.t. $d(A_i, A_{i+1}) = \frac{z}{n}$. (2) The optimal deployment of train antennas over a train of length $l_n < z \cdot (1 - \frac{1}{n})$ is a uniform deployment over the whole body of the train.

Proof: Once the positions of A_0 and A_{n-1} are set, it is clear from Lemma 10 that the rest are deployed uniformly between them. Otherwise, there is a triplet for which the middle antenna can be moved to improve overall capacity. Considering A_0 and A_{n-1}, note that in the MOD^d_z model they are also adjacent and thus we can apply Lemma 10 to both $\{A_{n-2}, A_{n-1}, A_0\}$ and $\{A_{n-1}, A_0, A_1\}$. For the case of $l_n \geq z \cdot (1 - \frac{1}{n})$ this implies a distance of $\frac{z}{n}$ between each pair, and for the case of $l_n < z \cdot (1 - \frac{1}{n})$ this implies we need to move A_{n-1} and A_0 as close as possible, which means placing them at the train edges. ■

The actual capacity gained by this system is determined by (1) the number of train antennas, (2) the ratio between cell-size and minimal distance (z and d), and (3) the exact value of α. See Section VI for numerical evaluation on this issue.

V. BSD and MASS combination

Reviewing the BSD algorithm (Section III), it is clear that the costly areas are those in which the reception capacity in relatively low. Implementing our MASS architecture not only increases the overall capacity but also ensures a better lower bound on the local capacity for all transmission slots. Also, high reception capacity is no longer localized at one single point in the cell but rather spread out along the cell, which should improve the performance of BSD greatly. It is therefore natural to combine both of these to significantly improve performance.

VI. Numerical and Simulative Evaluation

A. Minimal slots using BSD

To evaluate BSD we compare it to a family of "semi-static" algorithms which begin transmission when a certain percentage of the buffer is full (a parameter). Figure 5 depicts the number of slots occupied by these algorithms (dotted line) as a function of this parameter. The performance of BSD is given by the solid line. The results show significant slot savings achieved by BSD compared to all these algorithms. Note that some of those algorithms (to the right of 0.7) even violate the delay constraints (while BSD does not).

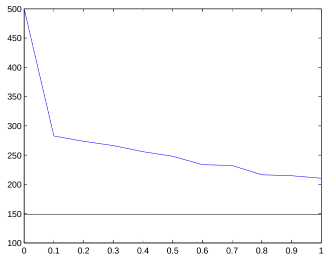

Fig. 5. Number slots required by train ($\alpha = -3$).

B. Capacity increase using MASS

Many factors affect the exact possible gain using MASS. In Figures (6) and (7), we compare the possible relative gain of a MASSarchitecture to that of a single-antenna system, using any number of train antennas and using only 2 antennas, respectively. Each Figure shows three curves, one for each of the different integer values of $\alpha \in \{-2, -3, -4\}$ (represented by the red, green and blue lines, respectively), and depicting the capacity gain of an optimized MASSversus the z/d ratio (ratio between the train-track span within the cell to distance between the base station and the tracks). As can be sen the gains highly depend on this ratio and are very significant for the whole spectrum.

To consider a realistic environment we consider (via simulation) a case where the cell sizes randomly vary while the train antenna placement is fixed. For this setting with 2 train antennas, the train relative gain is 142 percent (for antenna inter-distance of 250m); the gain is, nonetheless, not very sensitive to train antenna placement (e.g at distance of 200m or 300m it is 140 percent) and thus any static placement around such values will have significant gains.

In practice, however, the symmetric model is far from representing the actual structure of a base-station array. In order to asses the actual gain, we assume a distribution of distances between base-stations and iterate over the different distances between train antennas on board the trains.

References

[1] S. Borst, *User-Level Perfomance of Channel-Aware Scheduling Algorithms in Wireless Data Networks*, IEEE/ACM Trans. on Networking, Vol. 13 No. 3 June 2005.

[2] S. Borst, and Phil Whiting, *Dynamic Channel-Sensitive Scheduling Algorithms for Wireless Data Throughput Optimization*, IEEE Trans. on Vehicular Technology Vol. 52 No. 3 May 2003.

[3] J. Holtzman, *Asymptotic analysis of proportional fair algorithm*, Proc. IEEE PIMRC, 2001, pp. 33–37.

[4] Kun-De Lin and Jin-Fu Chang, *Communications and entertainment onboard a high-speed public transport system*, IEEE Wireless Communications, vol. 9, no. 1, February 2002, pp. 84 - 89.

[5] http://www.ukintpress-conferences.com/conf/rail04/pres/arumi.pdf

[6] E. Rosenswieng and H. Levy, Transmission Scheduling for Mass Transit in Data Wireless Networks, Technical report, Tel-Aviv University, http://www.cs.tau.ac.il/˜ elishar1.

[7] D. Tse, *Multi-user diversity andproportional fair scheduling* In preparation. Also http://www.eecs.berkeley.edu/ dtse/ima810.pdf.

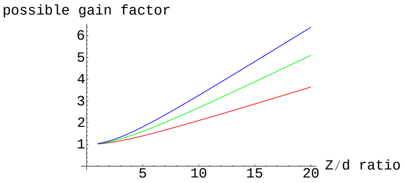

Fig. 6. Maximal gain using MASS

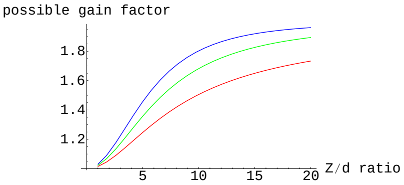

Fig. 7. Gain using MASS- two antenna version