5

Application: Cellular Network

Chapter 4 gave guidelines on how to model network architectures for parallel and distributed systems. Some ideas to reduce model complexity and accelerate model development were presented. This chapter gives an example on how to design a model for cellular networks using those ideas. The handoff in cellular networks will be of particular interest. The handoff process gives important information about the size of overlap area that must be chosen. But additional parameters, such as the velocity of the mobile nodes and the cell diameter, must also be considered.

The example investigates a concept for ubiquitous network access by mobile users dealing with real-time applications. They move in a cellular network and perform several handoffs. It is assumed that mobile users in particular use audio-based and video-based applications with specific quality of service (QoS) requirements. Support of real-time applications in wireless scenarios requires both fast handoffs and seamless QoS guarantees on the path of the mobile node through the network. A proposed solution to combine them for real-time support is the Ubiquitous Service Access Internet Architecture (USAIA) [183, 223]. USAIA provides hierarchical mobility management that interacts with the QoS mechanisms on three different network levels:

- The cellular level for handoffs between adjacent cells belonging to the same subnetwork. This level provides a fast handoff protocol and a scheme for resource reservation in advance to minimize the impacts of the movements of the mobile nodes to the QoS contract.
- The domain level for handoffs between different subnetworks. This level provides mobility management by means of either hierarchical foreign agents or Multiprotocol Label Switching (MPLS) [170]. Furthermore, QoS support is assumed to be provided by Differentiated Services and/or MPLS.
- The inter-networking level for handoffs between administrative domains. This level uses Mobile Internet Protocol (Mobile IP) [157] for mobility management.

For this investigation, the handoff is simplified, and the different protocols [188] are only marginally considered and not explained in detail here. The network performance will be mainly examined, depending on the network architecture and the required bandwidth of the mobile users.

5.1 USAIA Framework

USAIA is an all-IP framework for mobile access, with support for real-time traffic. It provides hierarchical mobility management that interacts with appropriate QoS mechanisms in different network areas. USAIA distinguishes the following three areas:

- the cellular level, including all cells of the wireless part of the access network belonging to the same IP subnetwork
- the domain level, including all IP subnetworks belonging to an administrative domain
- the global level, including all other administrative domains of the global Internet

Within the USAIA framework, base stations (BSs) are viewed as routers connecting the wireless cellular network to the wired part of the access network, which itself is connected to the Internet. Because the cellular level of USAIA is the subject of the performance evaluation of this study, it is explained in more detail in the following. All other aspects of the USAIA framework are described in detail in [183, 223].

The mobility management at the cell level is performed using a new fast handoff protocol. Furthermore, USAIA provides all necessary mechanisms for resource reservation in advance. For this purpose, the signaling protocol RSVP (Resource Reservation Protocol [22]) is modified in such a way that mobile nodes (MNs) are able to request resources on the BSs to which they have not yet performed a handoff. These requested resources are called passive reservations, because they can be used by the best-effort traffic of other MNs, as long as no handoff of the requesting MN has occurred. After a successful handoff, passive reservations turn into active ones. In that state, they cannot be used any longer by other traffic. To distinguish between these kinds of resources, the available bandwidth of a BS is partitioned, as depicted in Fig. 5.1.

Reservations of MNs requesting new resources from the network fall into the initial reservation class. Active and passive reservations of users leaving the scope of their initial or current BSs fall into the continuous reservation class. Reservations of MNs sharing the best-effort delivery traffic fall into the no reservation class. The mechanism to request resources in advance minimizes possible distortions of the real-time transmission during handoffs. This can be assured as long as the BSs are able to accept passive reservations. To provide the opportunity to accept as many passive reservations as possible,

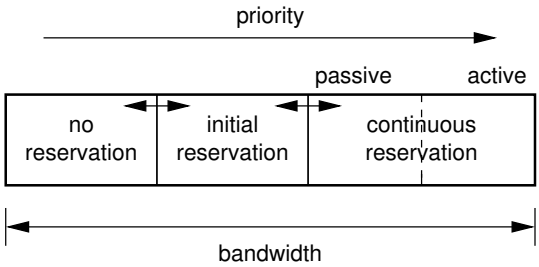

Fig. 5.1. Bandwidth partitioning

the protocols involved are optimized and extended in several ways. The details of these optimizations can be found in [183].

The key aspect of the USAIA framework is the seamless interworking of the mobility management and the resource reservation mechanisms to provide appropriate support of real-time traffic. To understand the interactions of both mechanisms, the message exchange between a correspondent host (CH) and the MN via the BS is explained in more detail (see also Fig. 5.2).

- The initial message is the beacon signal of a BS, containing the IP address of the BS that sends it. Beacon signals are sent via broadcast. The rate of the beacon signals is correlated to the current load and the available resources of the BS. To support fast handoffs, the beacon signal conveys a provider-defined share of the maximum resources that can be requested by MNs. This happens either during the handoff procedure as a passive reservation or as an initial reservation, i.e., after the successful termination of the authentication process, when the MN enters the network.
- Within the cell overlap areas, the MN recognizes beacons from different BSs, and triggers the handoff procedure. The handoff is initiated by sending an `MN_Announce` message to the new BS, carrying the MN's own address as well as that one of the "old" BS. The `MN_Announce` message contains either an indication of whether the handoff is marked as "triggered" or whether it is a "handoff announcement." The former initiates the handoff procedure and the latter makes it possible to request resources in advance, without losing connectivity to the current BS.
- The new BS confirms receipt of the `MN_Announce` message with an `MN_Announce_Ack` message.
- If the `MN_Announce` message is marked as triggered, the new BS sends a `Notify` message via the wired link to the old BS to inform it that the MN has moved. This message conveys the address of the new BS. The new BS also creates a routing table entry for the MN.
- Receiving a `Notify` message, the old BS deletes its routing table entry for the MN.

- If the `MN_Announce` message is not marked as triggered, the BS accepts a retransmitted `MN_Announce` message to trigger the handoff or the reservation in advance requests, respectively.
- The MN can now send `RSVP` messages for reservations in advance; these messages are acknowledged by the new BS by a `RSVP RESV Confirmation` message. The new BS performs all necessary internal actions to handle this passive reservation.
- The MN sends either a retransmitted `RSVP RESV` message to maintain its passive reservation(s) or a retransmitted `MN_Announce` message marked as triggered. The latter initializes the real handoff.
- On receiving a triggered `MN_Announce` message, the new BS turns the passive reservation into an active reservation and transmits a `Notify` message to the old BS.
- Finally, the new BS broadcasts a gratuitous proxy ARP (Address Resolution Protocol) to map the MN IP address to the BS link layer address, thus forcing all nodes involved to update their ARP caches with that information. This mechanism prevents the chaining of several BSs.

The most critical point of the USAIA framework with respect to real-time traffic support is the mechanisms provided on the cellular level due to the following reasons. First, the resources on the wireless link are usually much more limited than on the wired part of the network. Second, handoffs between

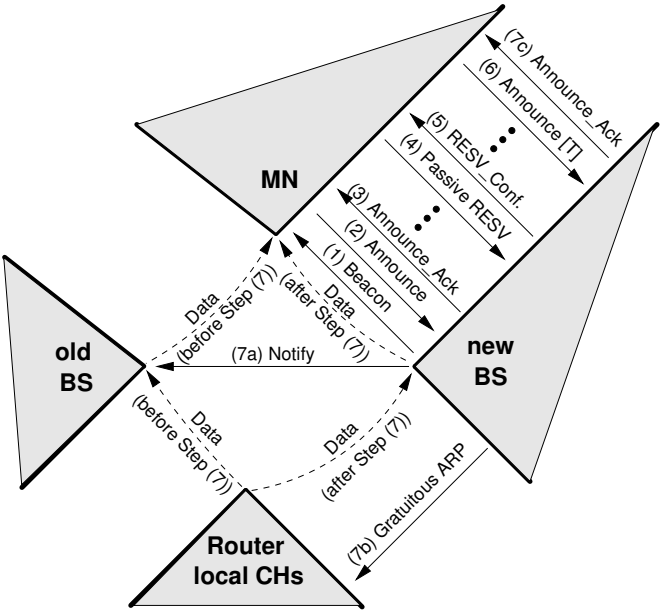

Fig. 5.2. Local handoff protocol

cells will likely happen more frequently than on the domain level or global level, thus influencing real-time traffic more significantly.

The major goal of the USAIA framework is the decoupling of the MN movement from the resources provided on any location visited. This means that there should be no need to stop the MN movement just to avoid any distortion of its real-time traffic. Therefore, the performance evaluation of all traffic classes involved with respect to accepted and failed reservations at the cellular level is of major interest.

5.2 Petri Net Model

Following the guidelines of Chap. 4, the modeling of the USAIA framework is started with the fastest and most simple model development scheme: the Petri net description. Another guideline of Chap. 4, that can be applied here, is to profit from symmetry. The Petri net description is based on modeling events of a single BS transmission range assuming a symmetric cellular network (as in Fig. 2.36) and that all BSs behave the same way. The model represents the transmission range of the BS and all MN activities that might occur within that range. These activities include nodes arriving at the range of the BS, nodes moving within the range, nodes initializing a new association within the range (MNs that are "switched on"), and nodes leaving the range. The term association denotes any kind of connection between BS and MN.

With regard to the kind of traffic MNs deal with, the model distinguishes MNs that carry real-time traffic (in the context of USAIA, this is related to a previous setup of a reservation request via RSVP) and MNs carrying best-effort traffic. Furthermore, the model takes into account MNs that establish a new association within the range of the BS. Any MN that deals with more than a single data flow within an association is handled by the model by dividing this node into multiple nodes, each one dealing with one of the data flows within a separate association. For instance, an MN carrying a data flow with real-time traffic and a data flow with best-effort traffic is represented by two MNs: one of them carrying a data flow with the real-time data and the other with the best-effort traffic. According to the guidelines of Chap. 4, a complex data flow scenario is decomposed into two simpler ones. A state space reduction results because only single data flows occur, and no combined data flows have to be represented by additional states.

To deal with the broad scope of individual reservation requests (initial and passive) of arbitrary size up to the permitted limit of bandwidth, discretization is applied as proposed in Chap. 4; the real-time traffic of MNs is assumed to belong to one of the three following classes: a class of low bandwidth requirements (called `Min`), a class of high bandwidth requirements (called `Max`), representing the maximum share of bandwidth a provider is willing to accept for a single request, and a class of average bandwidth requirements (called

Avg). Each of those classes is divided into two subclasses according to the extension of the Controlled Load Service Class described in [183].

Subclass R indicates that the requested bandwidth of a passive reservation can be reduced by the BS in the case of further incoming passive reservation requests and if the number of accumulated bandwidth of passive reservations exceeds a certain threshold value. For simplicity, our model distinguishes only between reservations that can be reduced to the lowest level of service the application is willing to accept and reservations that cannot be reduced due to the missing tolerance range. The latter means that the range specified in the Controlled Load Service specification of the application is set to 0. In reality, USAIA deals with the entire range of possible reductions, because the BS uses only the appropriate percentage of the tolerance ranges of all passive reservations to admit a new passive reservation request.

Subclass F means a fixed required bandwidth for passive reservations with no reduction allowed. As a result, different bandwidth allocations are modeled by six subclasses: MinF, AvgF, MaxF, MinR, AvgR, and MaxR.

In consequence, applying the discretization guideline of Chap. 4 to the bandwidth in the USAIA framework keeps the model tractable, while using continuous bandwidth would lead to an infinite state space. Of course, the more the number of bandwidth classes, the more accurate the model. In the following example, it turns out that three classes combined with two subclasses are sufficient to perform the required investigations.

DSPNs are used to describe the USAIA framework. Any passing time is represented by transitions that consume time from enabling to firing. Transitions with deterministic firing times are modeled by black filled rectangles. Transitions with exponentially distributed firing times are modeled by white rectangles. Immediate transitions, which do not consume any time, are modeled by black bars. All model input parameters, that can be changed are given in Table 5.1. They are explained in more detail in the following.

Each token appearing in the model represents an MN, except tokens in place Beacon modeling the beacon signal. Therefore, a kind of token "flow" through the net occurs, representing the current state of the corresponding MN. To handle multiple MNs (tokens) in parallel, all timed transitions are infinite server transitions except T1 (see Fig. 5.4), which models the MN arrival at the BS, and T44 (see Fig. 5.3), which models newly started applications at MNs.

The Petri net is described in detail below, starting with MNs opening their communication with an initial association. For simplicity, the authentication process is not taken into account.

5.2.1 Initialized Mobile Nodes

MNs that stay in the range of the BS and that like to establish a new association with a BS due to a newly started application are modeled by transition

Table 5.1. Model parameters

Parameter	Meaning
DnewMN	average time between new node arrivals
Dinit	average time between new initializations
Doverlap	time to pass the overlap area
Dremain	time to pass the remaining range
Dprocess	time for handshake between BS and MN
Dbeacon/DbeacInc	base/decrement time of beacon
fracRTT	fraction of real-time traffic
AvgRcvdBS	average number of received BSs
frac*Class*	traffic fraction of class *Class* (Table 5.2)
guard T49 etc	accepting an active reservation
guard T40 etc	threshold of passive bandwidth reduction
guard T45/T46 etc	accepting a passive reservation
guard T4 etc	accepting a initial reservation

T44 (Fig. 5.3). The rate of new applications requesting an association is assumed to be exponentially distributed. The average time between two requests is identified by the parameter Dinit. A new association to be established (token in place Init) requests either best-effort traffic or real-time traffic. The relation between the two kinds of traffic is modeled by the weights of the transitions T42 and T43. The parameter fracRTT defines the ratio of real-time traffic, and, therefore, defines the weight of transition T43. The weight of transition T42 results in 1-fracRTT.

If the MN requires real-time traffic (firing of transition T43), the node listens for a beacon signal (place listInitRT). The beacon signal model is introduced in Sect. 5.2.2. If such a signal is received (firing of transition T39), the node announces its presence to the BS (place AnnInit) by an MN_Announce message. Transition T38 models all the time needed for that procedure. That is the announcement procedure itself, including local processing time, and the reservation request procedure, also including the local processing time. Stochastic influence on this time, for instance, additional delays due to signal interferences, is neglected to keep the model simple. Thus, this time is deterministic and expressed by parameter Dprocess. This parameter is assumed to be independent of the kind of reservation (initial or passive).

The model provides dealing with network architectures where multiple BSs are in range of the MN. The parameter AvgRcvdBS defines the average number of BSs that are received at an MN's location. Of course, an initial association will be set up to only one of those BSs. Transition T36 ensures that the BS of our model will get the initial association. Otherwise, the request is turned to a passive reservation (firing of transition T37). The weight of T37 results in (AvgRcvdBS - 1)/AvgRcvdBS. Such an initial passive reservation is treated similarly to passive reservations that result from such MNs arriving at the

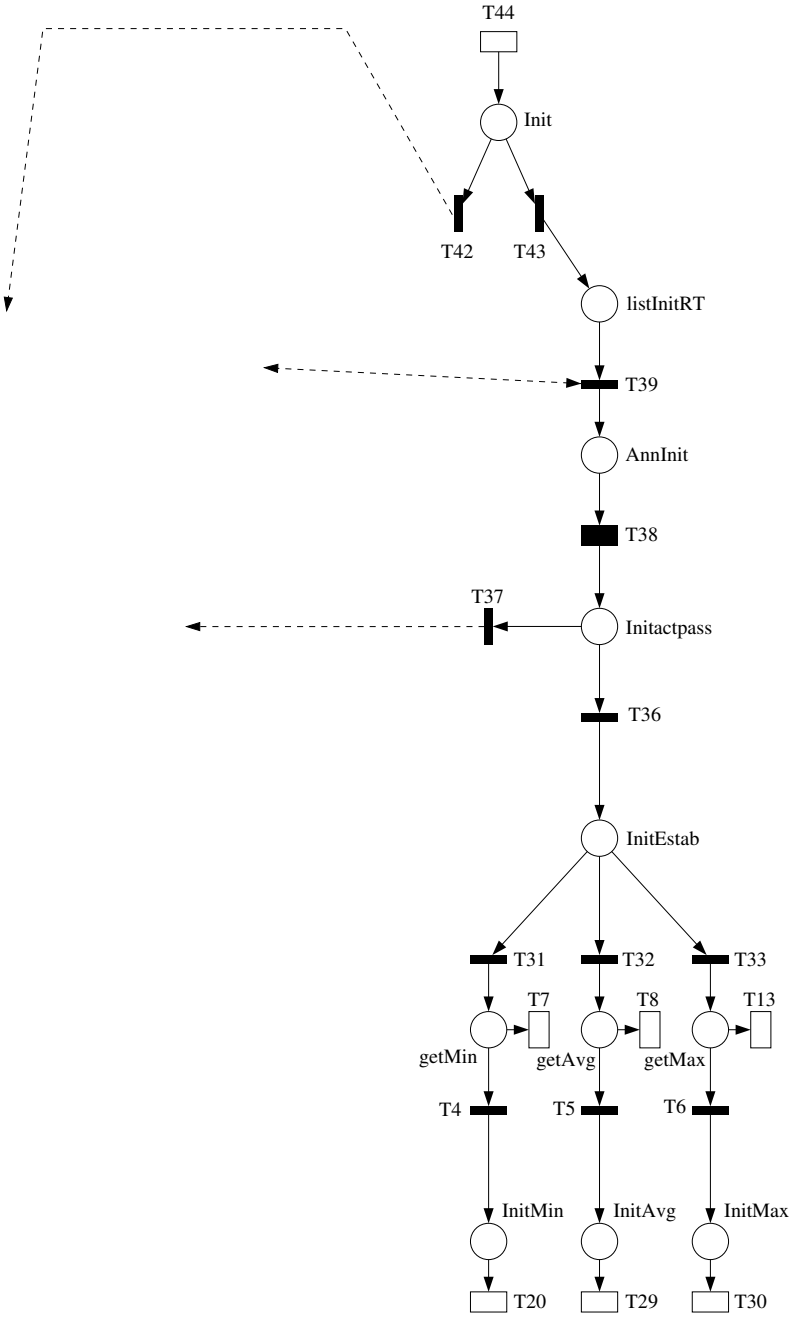

Fig. 5.3. MNs establishing a new association

range of the BS (Sect. 5.2.2). This means such passive reservations may turn into active ones (e.g., when the initial reservation to the other BS terminates), considering the same assumptions (concerning path, and so on) as in the case of arriving nodes.

The firing of transition `T36` indicates that the initial reservation is set up at the modeled BS. The weight of `T36` results in 1/`AvgRcvdBS`. Because initial reservations are also assumed to belong to one of the three bandwidth classes `Min`, `Avg`, and `Max`, the transitions `T31`, `T32`, and `T33` manage the distribution among the classes by their weights. The parameter `fracMinI` (weight of transition `T31`) gives the fraction of traffic class `Min`, `fracAvgI` (weight of `T32`), the fraction of class `Avg`, and `fracMaxI` (weight of `T33`), the fraction of class `Max`. The MN will try to establish a reservation until the required bandwidth is accepted or the MN leaves the range of the BS.

If the traffic belongs to the class `Min`, a token is generated in place `getMin`. Establishing the corresponding reservation is done by transition `T4`. But `T4` is only enabled if its guard is fulfilled:

$$\begin{aligned} &\#\texttt{InitMin} + \#\texttt{InitAvg} \cdot \texttt{AvgMult} + \#\texttt{InitMax} \cdot \texttt{MaxMult} \\ &\quad \leq \texttt{Initbw} - 1. \end{aligned} \tag{5.1}$$

The guard prevents a new reservation if the required bandwidth is not available. `AvgMult` defines the factor of the bandwidth the class `Avg` requires relative to the class `Min`. `MaxMult` defines the factor of the bandwidth the class `Max` requires relative to the class `Min`. `Initbw` represents the total bandwidth of the BS for initial reservations relative to the bandwidth of the class `Min`.

If the guard is fulfilled, an initial reservation is established (token in `InitMin`). The time the MN stays within the range of the BS is modeled by the exponentially distributed transition `T20` with an average delay of (`Doverlap + Dremain`)/2. Given such a delay, it is assumed that the MNs are initialized while spending half of their average time in the range of the BS. `Doverlap` and `Dremain` will be defined in Sect. 5.2.2. Due to the lack of movement patterns of MNs, the firing delay and its distribution are assumed. Besides exponential distribution, any other kind of distribution could also be modeled using a transition with a generalized distribution.

If the guard is not fulfilled, the MN tries to establish an initial reservation as long as it is in the range of the BS. Leaving the range is modeled by transition `T7` with an average delay time of (`Doverlap + Dremain`)/2. The measure `RInitMin` = E(#`InitMin`) gives the average number of initial reservations, and `RIgMin` = E(#`getMin`) gives the average number of nodes waiting for an initial reservation.

Both of the other bandwidth classes are modeled in a similar way. The only difference is given by the bandwidth requirements, which result in slightly different guards concerning the corresponding transitions to `T4`. For instance, in the case of the bandwidth class `Avg`, the maximum occupied bandwidth at the BS (right-hand side of Eq. (5.1)) has to be changed to `Initbw - AvgMult`

because the newly established association requires `AvgMult` times the minimal bandwidth (of class `Min`).

5.2.2 Real-time Traffic

MNs that enter the transmission range of the BS are modeled by the transition `T1` (Fig. 5.4). The arrival rate is assumed to be exponentially distributed. The average time between two arrivals is given by the parameter `DnewMN`, which can be set depending on the MN density and movement pattern. The relation

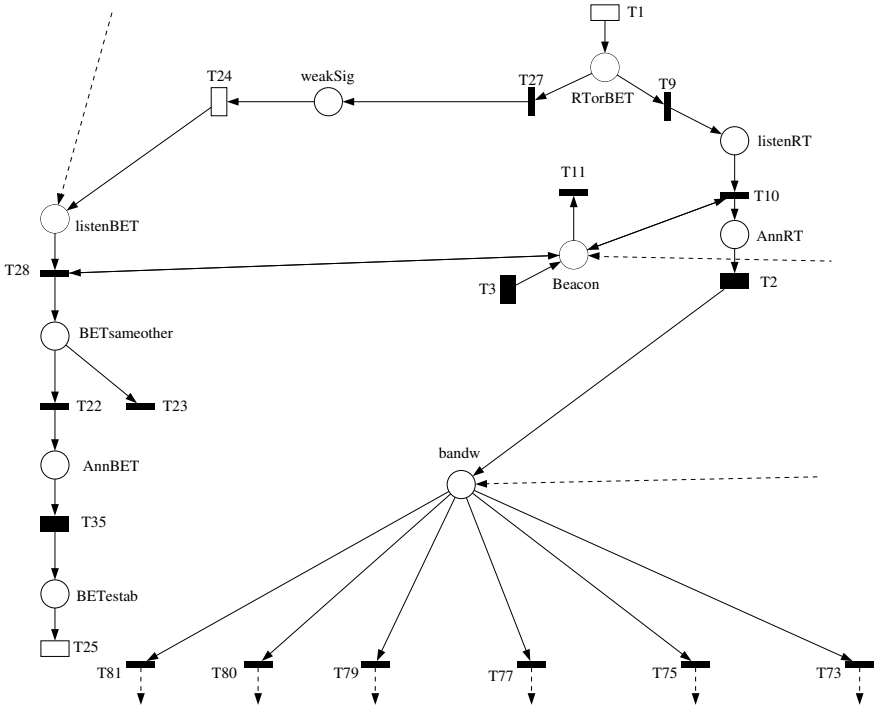

Fig. 5.4. MNs entering the BS range

between real-time traffic and best-effort traffic is modeled by the weights of the transitions `T27` and `T9`, as in case of initialized MNs.

An arriving MN first enters the overlap area of the new and the old BS. If it carries real-time traffic (firing of transition `T9` in Fig. 5.4), the node listens for a beacon signal (place `listenRT`) because it is willing to establish an association for a passive reservation. Transition `T3` represents the generation of the beacon signal, and therefore fires at a predefined fixed rate. Firing

leads to a token in Beacon. The MN receives the signal (firing of transition T10). Other kinds of traffic are also served in this way (transitions T28, best effort traffic, and T39, initialized nodes; Fig. 5.3). These three transitions are preferred to transition T11 due to their higher priority, set to 2. A firing of one of the three transitions does not lead to a deletion of the token in place Beacon because the corresponding arcs are double-sided. As a result, all MNs listening are handled first, and then the token representing the beacon signal is removed by the low priority transition T11.

The rate of T3, i.e., the rate of the beacon signal, is usually load dependent. This means that the higher the BS load (bandwidth occupied by associations), the lower the rate. As a result, the delay time of T3 is a marking-dependent function that deals with this constraint. Here, the delay time is controlled by the number of associations. It results in a basic rate of 1/Dbeacon in the case of no association. Dbeacon represents a model parameter. The rate is decreased for each association depending on a second parameter, DbeacInc. Other marking-dependent functions can be chosen as well.

If a beacon signal is received (firing of transition T10), the node announces its presence to the BS (place AnnRT) and requests a passive reservation via RSVP. Transition T2 accumulates the time required to handle this request, including the time for message passing, processing, and so on, similarly to T38 in the case of initialized MNs. The delay of T2 is given by the parameter Dprocess. The remaining model only considers the MN movement time.

After handling the request for a passive reservation, a token in place bandw occurs. Six transitions are enabled now (T81, T80, T79, T77, T75, and T73). They represent the assignment of the traffic carried by the MN to the six different real-time traffic bandwidth classes. The ratio of class MinF is modeled by the weight of transition T81 and determined by parameter fracMinF. Table

Table 5.2. Bandwidth classes

Class	Transition	Weight parameter
MinF	T81	fracMinF
AvgF	T80	fracAvgF
MaxF	T79	fracMaxF
MinR	T77	fracMinR
AvgR	T75	fracAvgR
MaxR	T73	fracMaxR

5.2 gives the transitions and weights of all bandwidth classes. After the firing of one of the transitions, a token is generated in the corresponding subnet, as shown in Fig. 5.5. The token in place bandw is removed. The token "flow" within the subnet is exemplary explained for the bandwidth class MinF of Fig. 5.5.

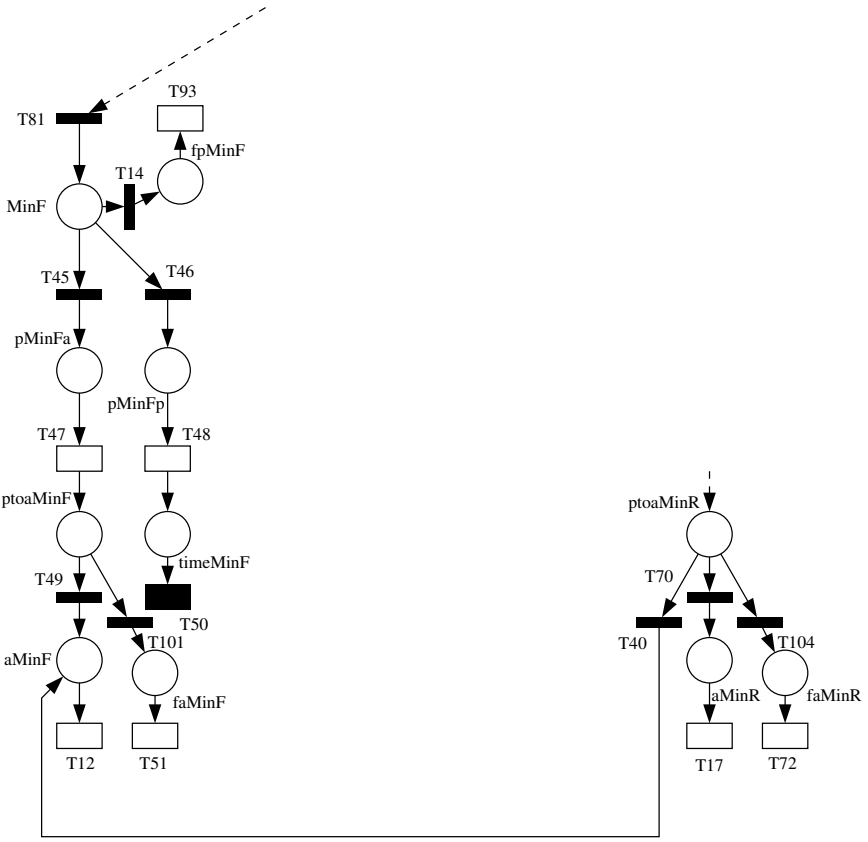

Fig. 5.5. One of the bandwidth classes

A token in place `MinF` indicates that the MN is carrying traffic of the class `MinF`. Next, the load of the BS is investigated and it is determined whether a passive reservation to the MN can be established (firing of transition `T45` or `T46`) or whether the request for a passive reservation is to be denied due to the lack of bandwidth (firing of transition `T14`).

Compared to the priority 1 of transition `T14`, the transitions `T45` and `T46` have a higher priority (priority of 2), and therefore, establishing the passive reservation has priority over denying it. But transitions `T45` and `T46` are also guarded. Both guards are equal and guarantee that a passive reservation is only established if the available bandwidth allows turning it into an active reservation if necessary. Nevertheless, the model also allows reserving more bandwidth than is available because not all passive reservations are turned into active ones. Therefore, the previously mentioned guards are of free choice, and the Petri net model may help find the optimal guards and the optimal

constraints in the real network for accepting passive reservations. Examples of guards are found in Sect. 5.3.

If T45 and T46 are not enabled due to the guards, T14 is fired and places a token in place fpMinF: the passive reservation has failed. The MN crosses the range of the BS without any reservation to it. The crossing time is modeled by transition T93 assuming an exponential distribution (or any other kind of distribution modeled by transitions with generalized distributions) with an average crossing time of Doverlap + Dremain. Doverlap models the time an MN needs to cross the overlap area of the old and the new BS. The time to traverse the remaining (new) range is given by Dremain. The average number of tokens in fpMinF determines the average number of failed passive reservations RfpMinF = E(#fpMinF).

If T45 and T46 are enabled, a passive reservation can be established. As mentioned in Sect. 5.2.1, the model allows dealing with network architectures where multiple BSs are in range of the MN. The parameter AvgRcvdBS defines the average number of BSs that are received at an MN's location. Of course, an active association will later be set up to only one of those BSs. Transition T45 ensures that the BS of our model will get the active association, and transition T46 ensures that one of the AvgRcvdBS - 1 other BSs will get it. Due to the assumed uniform movement pattern, the weight of transition T45 results in 1/AvgRcvdBS, and the weight of transition T46 results in (AvgRcvdBS - 1)/AvgRcvdBS.

If T46 fires, the MN will never establish an active reservation at the modeled BS. Place pMinFp represents the MN crossing the range of the BS. Transition T48 determines the average crossing time, given by Doverlap + Dremain. Instead of the exponential distribution, any other one may be chosen. Place timeMinF represents that the MN has left the transmission range. The passive reservation times out after the elapsed time defined by transition T50. It is given by the parameter Dtimeout. The average number of tokens in timeMinF allows determining the number of passive reservations that are timing out: RToutMinF = E(#timeMinF).

If T45 fires, the MN will establish an active reservation. Place pMinFa models the MN crossing half of the overlap area of the old and the new BS (modeled by an average delay time of Doverlap/2 by transition T47). Usually, the signal of the old BS will become weaker than the signal of the new BS after the crossing of half of the overlap area. As a result, the firing of T47 and a token in place ptoaMinF means that the passive reservation has to be turned into an active one: transition T49 fires if the available bandwidth of the BS is sufficient (a token is placed in aMinF). Otherwise, the firing of transition T101 indicates the failure of turning to an active reservation (token in place faMinF). The guard of T49 and the priorities of both transitions deal with this decision. Transition T49 has the higher priority, but also a guard in the way of

$$\begin{aligned}&\#\texttt{aMinF} + \#\texttt{aAvgF} \cdot \texttt{AvgMult} + \#\texttt{aMaxF} \cdot \texttt{MaxMult}\\&\quad +(\#\texttt{aMinR} + \#\texttt{aAvgR} \cdot \texttt{AvgMult} + \#\texttt{aMaxR} \cdot \texttt{MaxMult})/\texttt{reduce}\\&\quad \leq \texttt{Contbw} - 1,\end{aligned} \tag{5.2}$$

where `AvgMult` defines the factor of the bandwidth that the class `Avg` requires relative to the class `Min` (for both subclasses, `F` and `R`). `MaxMult` defines the factor of the bandwidth that the class `Max` requires relative to the class `Min`. `Contbw` represents the total continuous reservation bandwidth of the BS relative to the bandwidth of an association of class `MinF`. The parameter `reduce` models the factor of maximum bandwidth reduction of the reducible bandwidth classes. The guard in Eq. (5.2) ensures that the bandwidth for a new active reservation of the class `MinF` is available.

If the guard is fulfilled, the active reservation is activated (token in `aMinF`). The MN crosses the second part of the overlap area and the remaining range of the BS (modeled by the exponentially or alternatively distributed transition `T12` with an average delay of `Doverlap`/2 + `Dremain`). If the guard is not fulfilled, an active reservation fails (modeled by place `faMinF` and transition `T51` with an average delay of `Doverlap`/2 + `Dremain`). The measure `RaMinF` = E(#`aMinF`) denotes the average number of active reservations and the measure `RfaMinF` = E(#`faMinF`) determines the average number of failed active reservations.

All bandwidth classes are modeled in a way similar to that in Fig. 5.5. The only difference is given by the bandwidth requirements, and results in slightly different guards concerning the corresponding transitions to `T45`, `T46`, and `T49`. Concerning transition `T56` in the case of the bandwidth class `AvgF`, for instance, the maximum occupied bandwidth of the BS (right-hand side of Eq. (5.2)) should be `Contbw`−`AvgMult`.

The bandwidth classes `MinR`, `AvgR`, and `MaxR`, which represent passive reservations of reducible bandwidth, deal with an additional transition, as compared to the previously described case of fixed bandwidth (e.g., `T40` of bandwidth class `MinR` in Fig. 5.5). It models establishing an active reservation of non-reduced bandwidth if the BS load is light. The priority of `T40` is higher than those ones of `T70` and `T104`. But its guard prevents its firing if the BS is heavily loaded. The guard belongs to the model parameter set.

5.2.3 Entire Model

Combining the submodels of Figs. 5.3 to 5.5 lead to the entire Petri net model of the USAIA framework. It is depicted in Fig. 5.6. Table 5.3 shows all transition weights that differ from 1. The model development and evaluation was performed using the toolkit *TimeNET* [65].

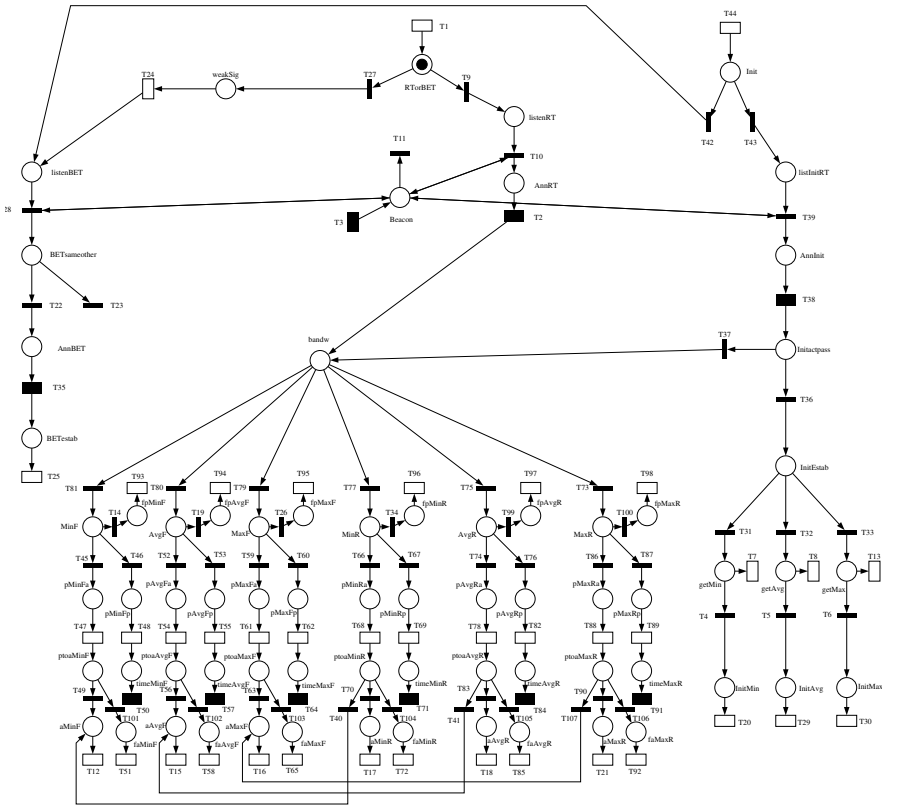

Fig. 5.6. Petri net model of the USAIA framework

Table 5.3. Transition weights

Transition	Weight
T9,T43	fracRTT
T27,T42	1 - fracRTT
T22,T36,T45 etc	1/AvgRcvdBS
T23,T37,T46 etc	(AvgRcvdBS - 1)/AvgRcvdBS
T81 etc, T31 etc	frac*Class*

5.3 Model Engineering and Performance

The model previously presented is revisited here in order to give a summary of how the design rules from Chap. 4 have influenced the architecture.

5.3.1 Model Development and Complexity Reduction

One of the most successful ways to reduce model complexity is to profit from symmetry, as this model does. Only a single base station is modeled, assuming all other base stations behave similarly. The data representation chosen is as simple as possible. Tokens that cannot be distinguished represent the mobile nodes and their movements. Other MN information, such as an identification number or billing information, is not needed to determine the performance of the USAIA framework for the given questions.

The model also combines events. Not all steps of the handoff protocol as given in Fig. 5.2 are explicitly incorporated. For instance, requests and their acknowledgments are combined in a single transition that reflects the amount of time needed for communication.

Discretization is applied by introducing only three bandwidth classes. The bandwidth that the MNs request in reality will usually be continuously distributed between minimum and maximum bandwidths. Here, the model complexity is reduced by mapping the requested bandwidth onto one of three discrete bandwidth classes.

However, some of the guidelines have not been considered. For instance, the computational overhead of a Petri net description is taken into account to come up with a fast graphical high-level modeling, which is provided by Petri nets. The following example applies the model to a wireless local area network (WLAN) environment and shows that the computation time is still small enough to perform all desired investigations. Thus, any other modeling techniques can be waived here. Chapter 6 gives an application where the simulation performance of the Petri net model is prohibitive for the required analysis of the system. In this case, other modeling techniques have to be used.

5.3.2 Modeling Power

The previously presented model is independent of cellular network technology. For each specific technology, the parameters can be adapted.

To demonstrate this model, the parameter set of a wireless local area network (WLAN) [27, 31, 64, 174, 236] scenario is used, according to IEEE 802.11b.

The features of a commercial product are applied as an access point, which in this case is the realization of the BS. Both terms are synonymous. For the model, it does not matter that access points are not IP-aware, because all time constraints were taken into account as if they were processing IP packets.

The diameter of the transmission range is about 60 meters for data rates up to 11 Mbps. Users (MNs) are assumed to pass the transmission range on paths that cover a length of half the diameter on average. If they walk (7 km/h), they spend 15.429 seconds within the range. It is further assumed that the network is set up in such a way that an average overlap of 10 meters

occurs along the path of a user. Of course, the model can also deal with any other overlap. As a result, the overlap area of 10 meters is passed in `Doverlap` = 5.143 s and the remaining path within the transmission range in `Dremain` = 15.429 s − 5.143 s = 10.286 s.

The BS is assumed to process any announcements or reservations of MNs in `Dprocess` = 0.001 s. Passive reservations time out after `Dtimeout` = 15 s. The beacon signal is initially sent at a rate of 10 per second (`Dbeacon` = 0.1 s) and decreased for each 10 established connections (`DbeacInc` = 10) by one.

MNs are assumed to be located in the range of `AvgRcvdBS` = 2 BSs on average. Their movement pattern and density result in MNs that arrive at a new BS at a rate of 1 per `DnewMN` = 0.05 s. MNs are initialized within the range of a BS at a rate of 1 per `Dinit` = 0.15 s. Half of all new nodes are assumed to carry best-effort traffic, and half of them real-time traffic (`fracRTT` = 0.5).

According to the product specification, the base station is able to handle about 50 nominal users that require a small bandwidth. Those users are associated with our bandwidth class `MinF` (and `Min` initial reservations). The mainstream users are assumed to consume twice the bandwidth of nominal users, and are associated with bandwidth classes `AvgF` and `Avg`. Power users are assumed to require four times the bandwidth of nominal users, and belong to the classes `MaxF` and `Max`. Users that can reduce their consumed bandwidth are assumed to allow a maximum reduction of 50%. The BS bandwidth is divided into a bandwidth of a maximum of 40 nominal users for active reservations and into a bandwidth of a maximum of eight nominal users for initial reservations. All bandwidth classes are assumed to occur with equal probabilities.

Concerning the guards that observe whether passive reservations can be turned into active ones (e.g., Eq. (5.2)), previous definitions lead to `AvgMult` = 2 and `MaxMult` = 4. A maximum reduction of 50% results in `reduce` = 2. Thus, the guards of transitions `T49`, `T56`, `T63`, `T70`, `T83`, and `T90` are determined. Furthermore, the guards that observe whether initial reservations are accepted are also set up by their corresponding equations (e.g., Eq. (5.1)).

The guards that observe whether a passive connection can be established are determined in a similar way. Nevertheless, any other arbitrary enabling function can also be chosen. The guard we have chosen considers all active reservations and their corresponding bandwidths. Only a fraction of the mentioned active bandwidth reservations is taken into account. This is because usually some time passes until a passive reservation turns into an active one. During this time, some of the other active reservations of the BS are terminated because the corresponding MNs have left the transmission range of the BS or terminated their applications. Thus, only the remaining fraction of active reservations is considered by the guard. The parameter `remact` allows assuming this fraction. The example presented sets it to 60% (`remact` = 0.6).

Furthermore, the passive reservations and their corresponding bandwidths (if they turn into active reservations) are considered. Only a fraction of those reservations is taken into account because several passive reservations will

time out and never turn into an active reservation. This ratio is given by the parameter `rempass`. It is the investigated parameter of the given example: What ratio should be chosen to achieve a high number of active and passive reservations but also a low number of failed ones? The guards of transition `T45` and `T46` result in

$$\begin{aligned}
&0.6(\#\texttt{aMinF} + \#\texttt{aAvgF} \cdot 2 + \#\texttt{aMaxF} \cdot 4 \\
&\quad +\#\texttt{aMinR} + \#\texttt{aAvgR} \cdot 2 + \#\texttt{aMaxR} \cdot 4) \\
&+\texttt{rempass} \cdot ((\#\texttt{pMinFp} + \#\texttt{timeMinF} + \#\texttt{pMinFa}) \\
&\qquad +(\#\texttt{pAvgFp} + \#\texttt{timeAvgF} + \#\texttt{pAvgFa}) \cdot 2 \\
&\qquad +(\#\texttt{pMaxFp} + \#\texttt{timeMaxF} + \#\texttt{pMaxFa}) \cdot 4 \\
&\qquad +(\#\texttt{pMinRp} + \#\texttt{timeMinR} + \#\texttt{pMinRa}) \\
&\qquad +(\#\texttt{pAvgRp} + \#\texttt{timeAvgR} + \#\texttt{pAvgRa}) \cdot 2 \\
&\qquad +(\#\texttt{pMaxRp} + \#\texttt{timeMaxR} + \#\texttt{pMaxRa}) \cdot 4) \ \leq \ 39. \quad (5.3)
\end{aligned}$$

The guards of the other bandwidth classes differ only in the right-hand side of the equation that determines the maximum occupied bandwidth of the BS to allow at least one active reservation of the corresponding bandwidth class.

The BS is configured to deal with as many active reservations as possible. This means that traffic of subclass `R` is accepted by the BS only with their reduced bandwidth. It is modeled by disabling transitions `T40`, `T41`, and `T107`.

Figures 5.7 to 5.10 show some results. The reciprocal ratio of passive reservations is $1/\texttt{rempass}$. The average number of active reservations (Fig. 5.7), the average number of failed active reservations (Fig. 5.8), the average number of passive reservations (Fig. 5.9), and the average number of failed active reservations (Fig. 5.10) are compared for all six bandwidth classes. The results are achieved by simulation due to multiple enabled deterministic transitions that cannot be mapped onto a Markov chain. As termination criteria, a confidence level of 95% and an estimated precision of 2% is used. But in the case of rare events, these criteria are relaxed due to a simulation run-time of more than 10 hours on a 1,200 MHz processor. Relaxing the estimated precision to 5% leads to about two hours simulation time.

The plots of the bandwidth classes `MinF` and `AvgR`, and those of `AvgF` and `MaxR`, are equal due to equal probabilities of the bandwidth classes and equal bandwidth of both classes if bandwidth reduction is taken into account.

The minimal ratio of passive connections to avoid failed active reservations arises as another result of the example. The reciprocal ratio should not be larger than 10 (Fig. 5.8), resulting in $\texttt{rempass} = 0.1$.

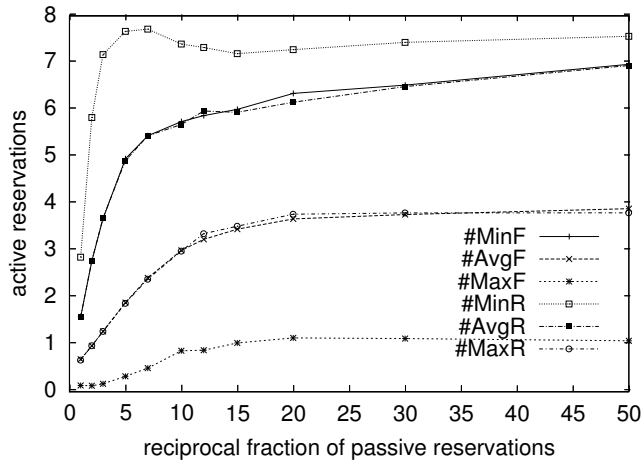

Fig. 5.7. Active reservations

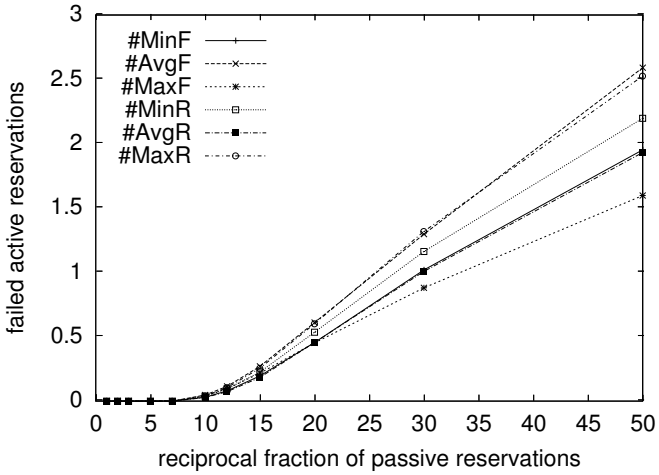

Fig. 5.8. Failed active reservations

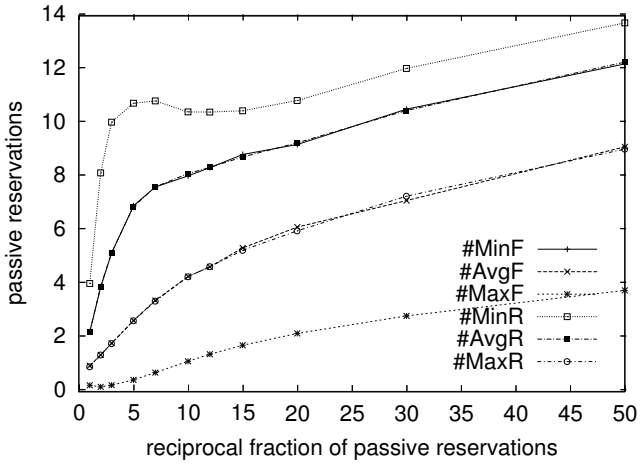

Fig. 5.9. Passive reservations

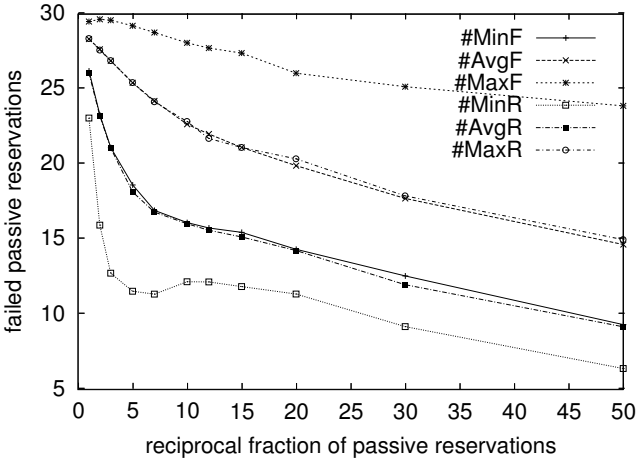

Fig. 5.10. Failed passive reservations

http://www.springer.com/978-3-540-34308-0

Performance Analysis of Network Architectures

Tutsch, D.

2006, IX, 245 p., Hardcover
ISBN: 978-3-540-34308-0