CHAPTER III

PROPOSED METHOD

This chapter describes the proposed data transfer method and the proposed path finding algorithm. The proposed data transfer method is a NC-BR Mechanism which includes the NC-BR overview, operations, a frame structure design, a scope discussion and a performance analysis. The proposed path finding algorithms are comprised with the Hierarchical Index Road Network (HIRN), which includes HIRN system infrastructure, data structures, datacenter functions, mobile unit and communication protocol functions, and the Adaptive Travel-Time Path Selection (ATTPS) which includes the ATTPS algorithm.

3.1 Data transfer method: NC-BR Mechanism

This research proposed a network coding based relay, called the NC-BR mechanism. Moreover, the frame structure of this mechanism is proposed. The performance analysis has been performed; and the simulation results show that the proposed strategies help reducing both throughput degradation and additional end-to-end delay problems. Moreover, the jitter value in the IEEE 802.16j multi-hop relay network is also small, which is beneficial to real-time services over the multi-hop network.

3.1.1 NC-BR Overview

In this research, the NC-BR mechanism and the frame structure that support an XOR network coding approach are proposed. This allows RS to combine two sets of data using XOR operators then transmits all data together as a single transmission. The sets of data that the RS can combine are obtained from the data that the RS has to transmit to its super-ordinate RS/MR-BS, and to its sub-ordinate RS. This approach reduces number of transmission between hops. Thus, the number of idle periods effected from the limitation of signal interference is reduced. Therefore, the throughput and delays of the network can be greatly improved.

In order to describe the NC-BR mechanism, the definition of RS that is capable of being a NC-BR node in IEEE 802.16j Multi-hop Relay Network is as shown in Figure 3.1.

Definition 1:

Let three nodes of RSs (or 1 MR-BS with 2 RSs) be the super-ordinate node (N_{i-1}), center node (N_i) and, sub-ordinate node (N_{i+1}) if and only if these three nodes connected and communicate as follows:

- *The super-ordinate node, N_{i-1}:*
  MR-BS or RS node is the super-ordinate node if and only if it communicates with the center node, N_i, by

  -- N_{i-1} transmits its downlink data to the center node, N_i,

  -- N_{i-1} receives the center node, N_i, uplink data.

- *The center node, N_i:*
  RS node is the center node if and only if it communicates with the super-ordinate node, N_{i-1}, by

  -- N_i transmits its uplink data to the super-ordinate node, N_{i-1},

  -- N_i receives downlink data from the super-ordinate node, N_{i-1},

  and its communicate with the sub-ordinate node, N_{i+1}

  -- N_i transmits its downlink data to the sub-ordinate node, N_{i+1},

  -- N_i receives the sub-ordinate node, N_{i+1} uplink data.

- *The sub-ordinate node, N_{i+1}:*
  RS node is the sub-ordinate node if and only if it communicates with the center node, N_i, by

  -- N_{i+1} transmits its uplink data to the center node, N_i,

  -- N_{i+1} receives the downlink data from the center node, N_i.

3.1.2 NC-BR Operations

Referring to Figure 3.1, simple transmission procedures are drawn. Figure 3.1(a) shows the original RS traffic flows and Figure 3.1(b) shows the traffic flows of the

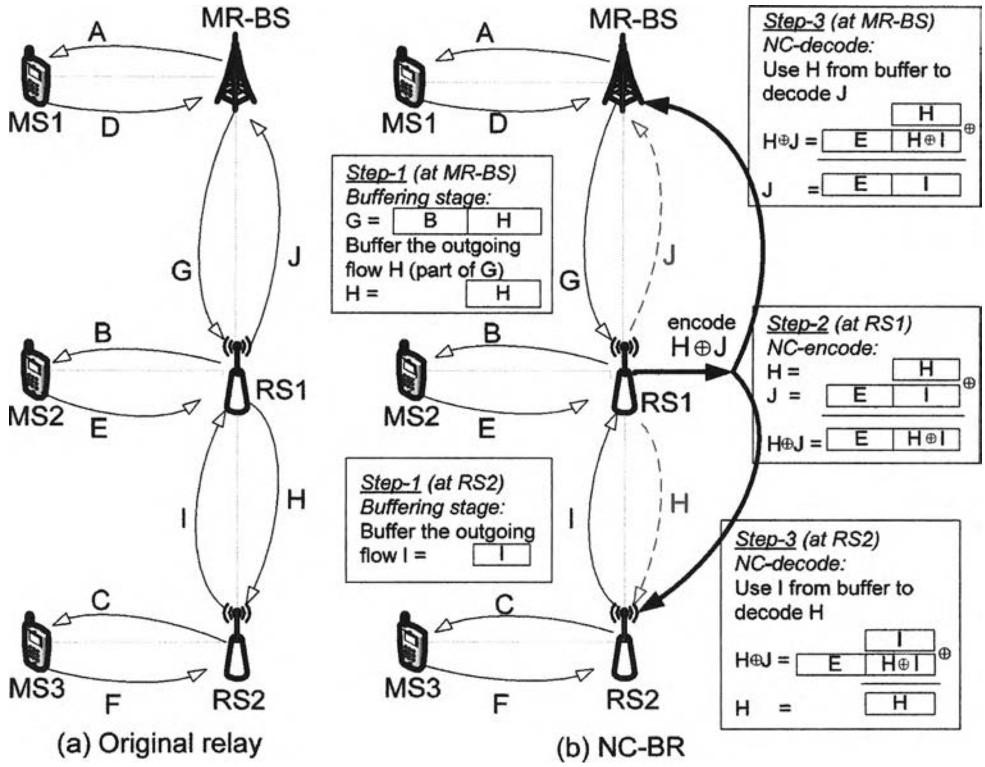


Figure 3.1: The NC-BR overview.

proposed NC-BR. According to Figure 4(a), the operations of RS1 are (1) to receive the traffic flows *G* and *I*, and (2) to transmit the traffic flows *J* and *H*. While the operations of RS1 in Figure 3.1(b) were modified to receive the traffic flows *G* and *I*, then the single encoded traffic flow which is $H \oplus J$ was transmitted. Moreover, the encoded flow $H \oplus J$ is encoded from flows *H* and *I* and it does not need to have the same length. When RS1 transmits $H \oplus J$, both the MR-BS and the RS2 can overhear the encoded traffic over the wireless backhaul. Thus, the RS1 needs only three steps, instead of four, to transmit the data and complete its task.

The contribution of this approach is that the XOR network coding technique is applied; so there is no extra hardware is required because of its simplicity. Additionally, the XOR network coding technique has been proven that the network coding definitely has practical benefits and can substantially improve wireless throughput of store-and-forward behavior.

This operation is responsible for encoding 2 traffic flows that are transmitting to both super-ordinate and sub-ordinate RS, using the XOR operation at the MAC frame level. In order to integrate and encode data of 2 traffic flows, the NC-encoder uses only the data stored in the relay zone, the FCH and R-map are leave un-encoded.

Generally, the traffic flow to a super-ordinate RS and a sub-ordinate RS might have different amount of data in each MAC frame. The NC-encoder will add the padding character to the shorter traffic flow, in order to equalize their MAC frames. The two sets of data that are encoded at the RS1 came from the opposite directions. Additionally, each node keeps the last transmitted data frame for the further decoding process.

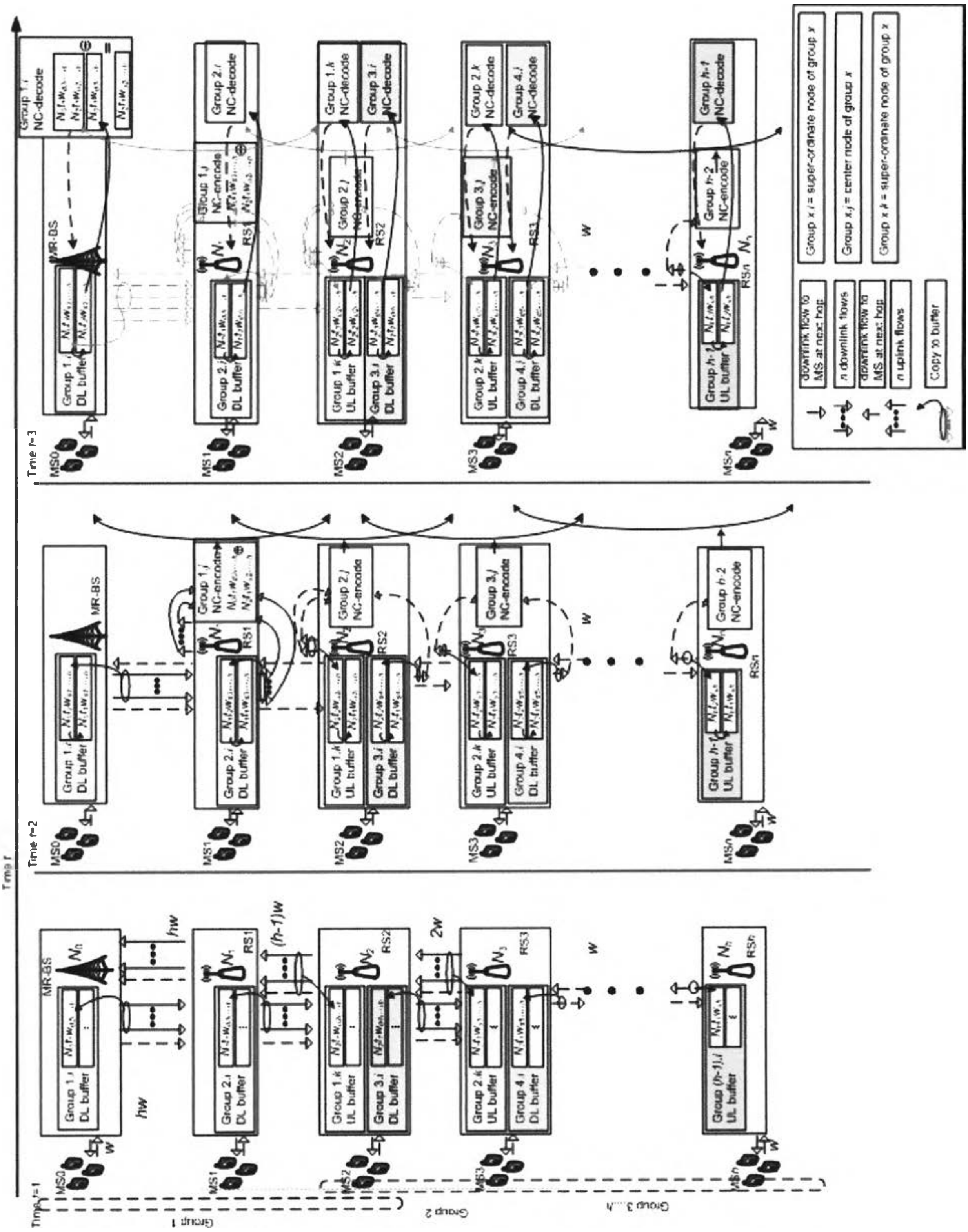


Figure 3.2: Details of NC-BR operation.

Considering the NC-BR process; it spans over three MAC frames at time t (i.e. frame t, $t+1$ and $t+2$). Figure 3.1(b) shows the example of NC-BR process which comprises three operations as follows.

- *Step-1 Buffering stage.*

The buffering stage is occurring at the super-ordinate node and the sub-ordinate node. It aims to buffer the outgoing data flows (or a part of data flows) of the super-ordinate node and the sub-ordinate node after its transmitting to the center node. These buffered data flows are applied in the decoding process of the super-ordinate node and the sub-ordinate node for the XOR encoded data from the center node in step-3.

Definition 2:

Let $f_i(j)$ and $f_i(k)$ be outgoing data flows of the super-ordinate node and the sub-ordinate node that transmit to the center node, i, respectively.

Let F be a set of all data flows to the center node, i.

Let B_{Sp} be a buffer space at the super-ordinate node that stores the last outgoing frame at time t, $f_{i,t}(j)$. $B_{Sp}(t)= f_{i,t}(j)$.

According to the IEEE 802.16j Multi-hop Network architecture, the center node may be connected by MSs. Hence, the data flows that transmit from the super-ordinate node to the center node at time t, will be forwarded to the sub-ordinate node and a part of it may be forwarded to MSs of the center node.

To be able to decode in step-3 at time $t+2$, the super-ordinate node needs only the data that will be forwarded to the sub-ordinate node. Hence, the data that needs to be stored in the buffer $B_{Sp}(t)$ is $f_{k,t+1}(k)$ at the time $t+1$, which is the subset of $f_{i,t}(j)$. To identify $f_{k,t+1}(k)$ out of $f_{i,t}(j)$, the super-ordinate node needs to look into their own R-MAP of this relay zone at time t.

Let B_{Sb} be the buffer space at the sub-ordinate node that stores the last outgoing frame at time t, $f_{i,t}(k)$. $B_{Sb}(t)= f_{i,t}(k)$.

For example:

In the frame t of Figure 3.1(b), the MR-BS transmits the flow G to the RS1, and the RS2 transmits the flow I to the RS1. The MR-BS and the RS2 need to buffer the outgoing data of the flow H (a part of G) and the flow I, respectively. This allows both the MR-BS and the RS2 to decode the traffic from the RS1 in the step-3.

- *Step-2 NC-encode.*

The NC-encode stage of the center node aims to combine a set of data flows to super-ordinate node and the sub-ordinate node as a single outgoing flow. Since the wireless network transmission is broadcast, both super-ordinate and the sub-ordinate node can overhear the center node transmission. This stage let the center node combine the data flows by the XOR operation and transmit in a single transmission.

Definition 3:

Let i be the center node, j be the super-ordinate node, and k be the sub-ordinate node.

$O(i, j) = \{ o(i, j) | o(i,j)$ be an output flow from the node i to node j, $j=1,\ldots,n\}$.

Let $O(i, j)$ be an output set of all flows from the center node, i , to the super-ordinate node, j.

Let $O(i, k)$ be an output set of all flows from the center node, i , to the sub-ordinate node, k.

Since $O(i, j)$ and $O(i, k)$ have unequal length, then the content of the shorter flow will be inserted with a set of meaningless pattern characters, called padding characters, at the beginning of the content. So, the lengths of $O(i, j)$ and $O(i, k)$ are equivalence.

Definition 4:

Let $O(i, j)$ be an output set of all flows from the node i to the node j with an equal length when comparing to $O(i, k)$.

Let $R_{t+1}(O(i, j), O(i, k))$ be a result from the encoding function XOR at time $t+1$, $R_{t+1}(O(i, j), O(i, k)) = O(i, j) \oplus O(i, k))$.

In the frame *t+1*, the RS1 integrates and encodes the flows *H* and *J* by the XOR operation. In this operation, the right alignment is applied for the shorter flow. Moreover, data bits of the longer flow that have no opponent in this XOR operation will leave un-encoded. Then, the RS1 will transmit the encoded flow $H \oplus J$ over the wireless channel, which will be overheard by both the MR-BS and the RS2.

- *Step-3 NC-decode.*

The NC-decode stage is occurring at the super-ordinate node and the sub-ordinate node. It aims to decode the incoming XOR encoded data flows, or part of data flows, from the center node. The decoding operation in this stage is performed by XOR the incoming data flows with the appropriate buffer data from step-1.

Definition 5:

Since $B_t(j)$, $B_t(k)$, and $R_{t+1}(O(i,\ j),\ O(i,\ k))$ have unequal lengths, then at the beginning of the content with the shorter flow must be inserted by a set of meaningless pattern characters, called padding characters. So, the lengths of $B_t(j)$, $B_t(k)$, and $R_{t+1}(O(i,\ j),\ O(i,\ k))$ are equivalence.

Let $D_{t+2}(j)$ be a decoding output at the super-ordinate node at time *t+2*, and $D_{t+2}(k)$ be a decoding output at the super-ordinate node at time *t+2*

$$D_{t+2}(j) = R_{t+1}(O(i,\ j),\ O(i,\ k)) \oplus B_t(j)$$

$$D_{t+2}(k) = R_{t+1}(O(i,\ j),\ O(i,\ k)) \oplus B_t(k)$$

In the frame *t+2*, since the RS1 is already to transmit the encoded traffic flow $H \oplus J$ in the frame *t+1*, both the MR-BS and the RS2 have overheard and received the message.

At the MR-BS, the encoded flow $H \oplus J$ was received. The output data that needed by the MR-BS is only the flow *J*. The MR-BS decodes the flow $H \oplus J$ by using

XOR with the buffered data from step 1, a flow *H*. Now, the decoded result is the flow *J* which the MR-BS can remove the padding characters.

Similarly, the RS2 received the same $H \oplus J$ flow similar to the MR-BS. However, the output data that the RS2 wants is only the flow *H*. Due to the flow *J* is the concatenation of the flow *E* and the flow *I*. As the fact that the flow *H* is also smaller than the flow *J* which is the same size of the flow *I*, and there is a set of padding characters in front part of the flow *H* inside the $H \oplus J$ flow. Then, the RS2 which has only the flow *I* in its buffer, still can decode *H* out of the flow $H \oplus J$ by adding a set of padding characters in front of the buffered flow *I*, for size equalization. This will cause the position of the flow *H*, which is the last part of the encoded flow $H \oplus J$, aligns to the position of the added padding flow *I*. Now, RS2 decodes the flow $H \oplus J$ by XOR with the buffered added padding flow *I*. The result is the flow *H* which RS2 can remove padding characters.

3.1.3 NC-BR Frame structure design

The frame structure design is an essential part of the IEEE 802.16j multi-hop relay network. Most problems mentioned in the multi-hop data transfer problem are related to the frame structure design. This research proposes a new improved frame structure that supports the NC-BR mechanism. Thus, the NC-BR mechanism has greatly improved the network throughput, and reduced both delay as well as jitter values.

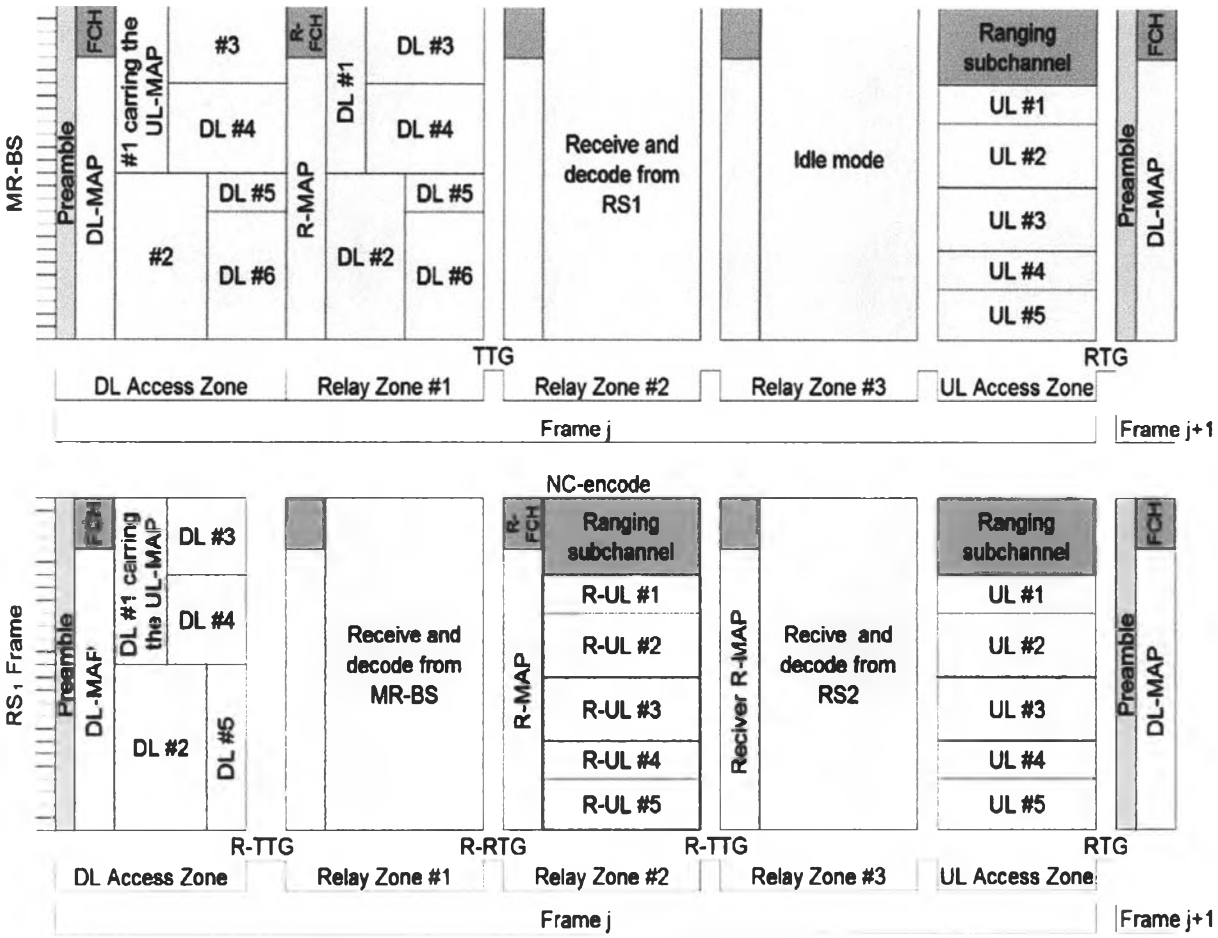


Figure 3.3: NC-based relay frame structure design.

Referring to the original transmission process in the multi-hop relay network, only one station will receive data at a time. In contrast, applying the NC-BR on all RSs, every super-ordinate and every sub-ordinate of RSs will be able to receive data from a single transmission. Consequently, the transmission pattern has improved with much higher efficiency, in addition to gaining benefits according to the higher utilization in the relay zones.

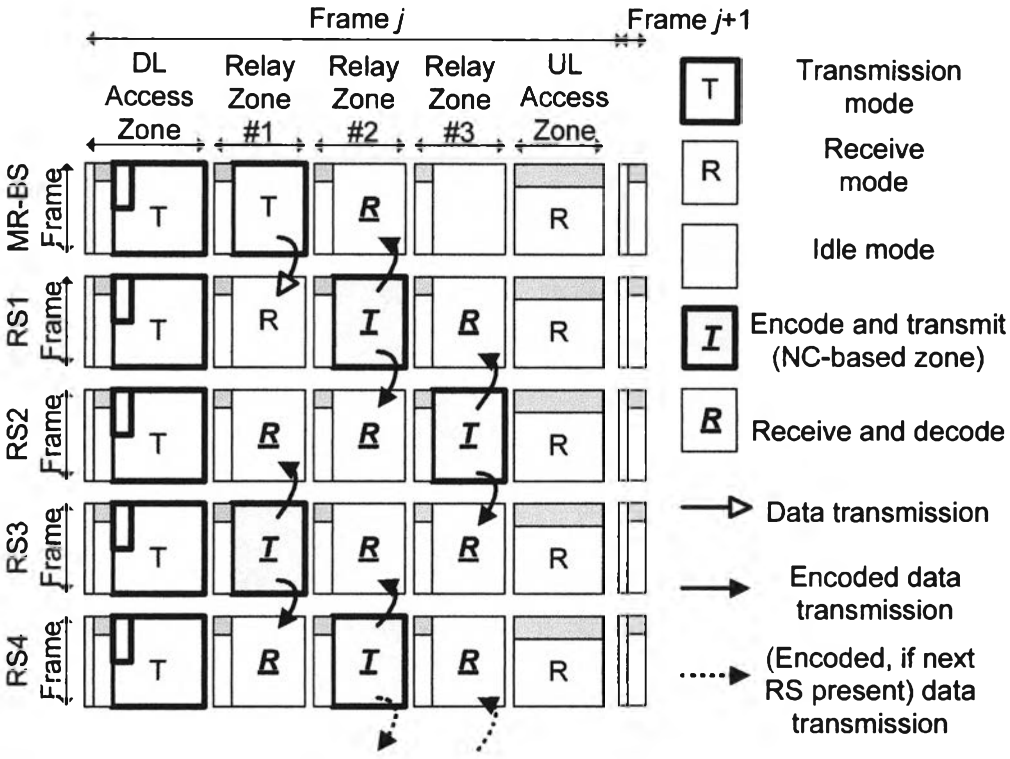


Figure 3.4: NC-BR transmission sequence in frame structure.

The draft document of IEEE 802.16j describes two approaches of the multi-hop relay frame structure: the repeating patterns of the multi-frame, and the more than one relay zones in the single frame structure. The proposed frame structure, NC-BR frame, is laid on the second one. Figure 3.3 shows the proposed NC-BR frame. The frame format at the MR-BS and the RS is divided into 3 different zones: the DL access zone, the UL access zone, and three relay zones. Both DL and UL access zones are used for communicating with the direct connected MSs of each MR-BS/RSs, while three relay zones are used in the communication among the MR-BS, super-ordinate, and sub-ordinate RSs. Figure 3.4 shows the NC-BR transmission sequence in the frame structure.

One significant difference of this proposed frame structure and the original frame structure is that the number of zones in the MAC frame is independent to number of RSs hops. Therefore, only three relay zones in the NC-BR are required on all situations. As a result, the throughput degradation and the delay increase will be independent from the number of RSs hops.

Therefore, based on Figure 3.3 and Figure 3.4, within the three relay zones of MR-BS, relay zone #1 is assigned to transmit data to RS1, relay zone #2 is to receive and decode data from RS1 which is NC-BR enable, relay zone #3 has to stay in the idle mode. At RS1, relay zone #1 is assigned to receive data from the MR-BS, relay zone #2 is the NC-BR enable zone, encodes and transmits data to MR-BS and RS2 in the same transmission, relay zone #3 is to receive and decode data from RS2 which is the NC-BR enable. From RS2 and beyond, the NC-BR enable zone is placed at the relay zone number 3, then 1, then 2, and repeating, other two relay zones are assigned to receive and decode data from the super-ordinate or the sub-ordinate RSs. If there is no remaining sub-ordinate RS, the last RS in the chain will have only one relay zone staying in the idle mode. Noticeably, up to only 2 transmission opportunities in relay zones from the entire system are left waste in the idle mode. This is another reason that the proposed approach can provide high improvement from high utilization of the proposed new frame structure design. Hence, by applying the proposed NC-BR, the network coding will also help RSs transmit more data in the relay link, results are to relief both optimal offered load points and bottle-neck points. More detail will be described in the next section. The following paragraph is the example of NC-BR Frame Structure and the relay zones that is capable to be the NC-based zones.

Figure 3.4 also presents the transmission sequence within the relay zone. Each zone in the transmission sequence can be in either transmission, receive or idle modes. The transmission sequence of the MR-BS is assigned to the transmission mode in the relay zone #1, the receive mode in the relay zone #2, and the idle mode in the relay zone #3. The transmission sequence of RS1 is assigned to receive data from the MR-BS on the relay zone #1. The NC-based zone can be applied to the relay zone #2, which encodes and transmits data to both directions (MR-BS and RS2) in a single transmission. The relay zone #3 is to receive and decode an encoded data from RS2. The relay zone of the center node RS which is in the transmission mode is defined as the

NC-based zones. The NC-based zone in each RS transmission sequence is denoted by $RS_{i\text{-}NC\text{-}zone}$, where i is number of RSs hops and h is total RSs hops, can be expressed by

$$RS_{i-NC-zone} = \begin{cases} none & ,i=0 \\ (i \bmod 3)+1 & ,0<i<h \\ none & ,i=h. \end{cases} \qquad (1)$$

It is worth noticing that the number of relay zones in the NC-BR frame is independent to the number of RS hops. However, in the original relay scheme, the idle mode has to be assigned to more zones in the transmission sequence to avoid the transmission interference when the number of RS hops is increased. The idle mode is considered as waste.

3.1.4 NC-BR Scope discussion

In this part, the discussion on scope and limitation of the NC-BR are presented. The proposed mechanism mentioned previously is focused on the implementation over the IEEE 802.16j Multi-hop networks. Generally, the NC-BR can be used in any multi-hop relay wireless networks that are satisfied the following criteria.

- *Available coding opportunities for relay station:*

  Which is need three or more hops in the chain topology for coding opportunities of the center node relay.

- *Buffer:*

  The NC-BR requires a small buffer space to keep the recently transmit data for further decoding of the new coming received.

- *Omni-directional antenna:*

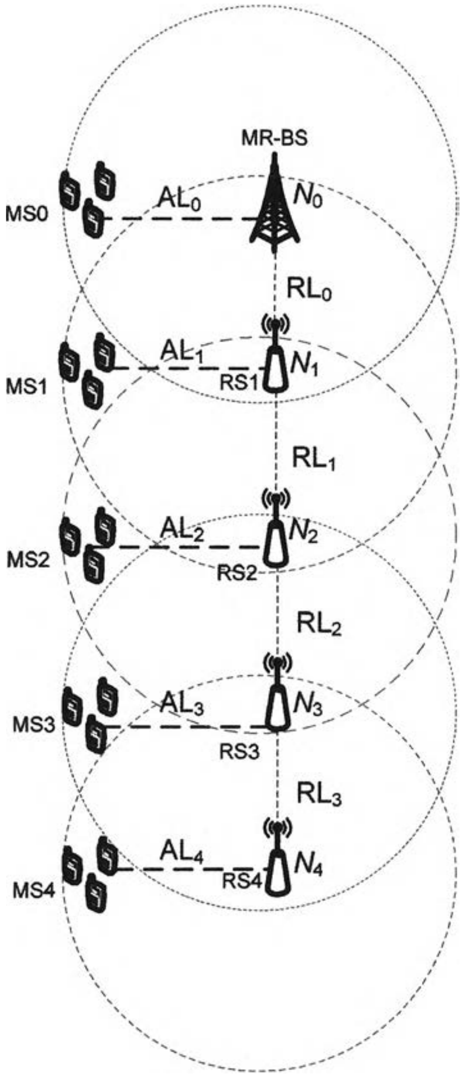
  The wireless backhaul of the relay station that communicates to the super-ordinate RS/BS and the sub-ordinate RS has to concurrently transmit data to both directions. Generally, the wireless backhaul should use the Omni-directional antennas or other kinds that broadcast to both directions.

- *Higher gain in balance the uplink-downlink traffic ratio:*

There have been studies about XOR network coding in the wireless network [97-98] and its shows the balance ratio of the uplink-downlink traffic is the highest throughput gain scenario. This study follows the same trend. Due to higher opportunity in applying the XOR network coding to uplink with downlink data in each frame, more data can be encode and transmit in single transmission. Consequently, the highest throughput gains can be obtained. On the other hand, when the uplink-downlink traffic is not equal, the opportunity to apply the XOR network coding will decrease, including also the throughput gain.

3.1.5 NC-BR Performance analysis

In this section, an analytical model to study the throughput degradation and delay increase in IEEE 802.16j multi-hop relay network is described.



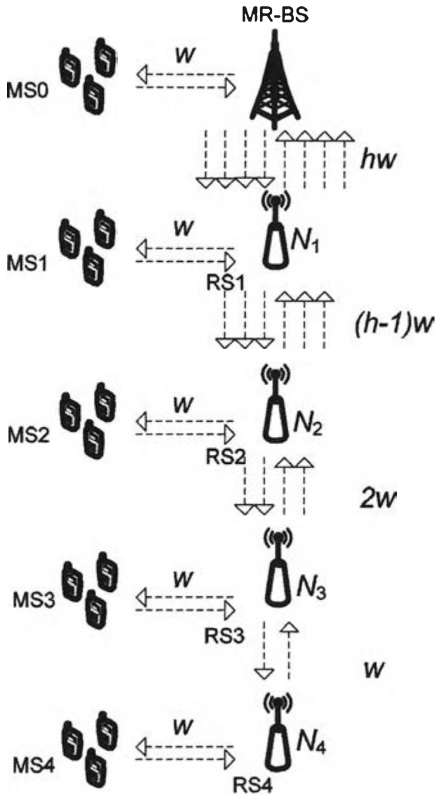


Figure 3.5: Variables represent.   Figure 3.6: Traffics represent.

Assumptions: All MR-BS and RSs are operated in the same frequency channel, the same transmission power and the same range. Moreover, all wireless relay links use the highest modulation. Thus, models of the throughput degradation and delay increasing in the IEEE 802.16j Multi-hop Relay Network are studied in this thesis.

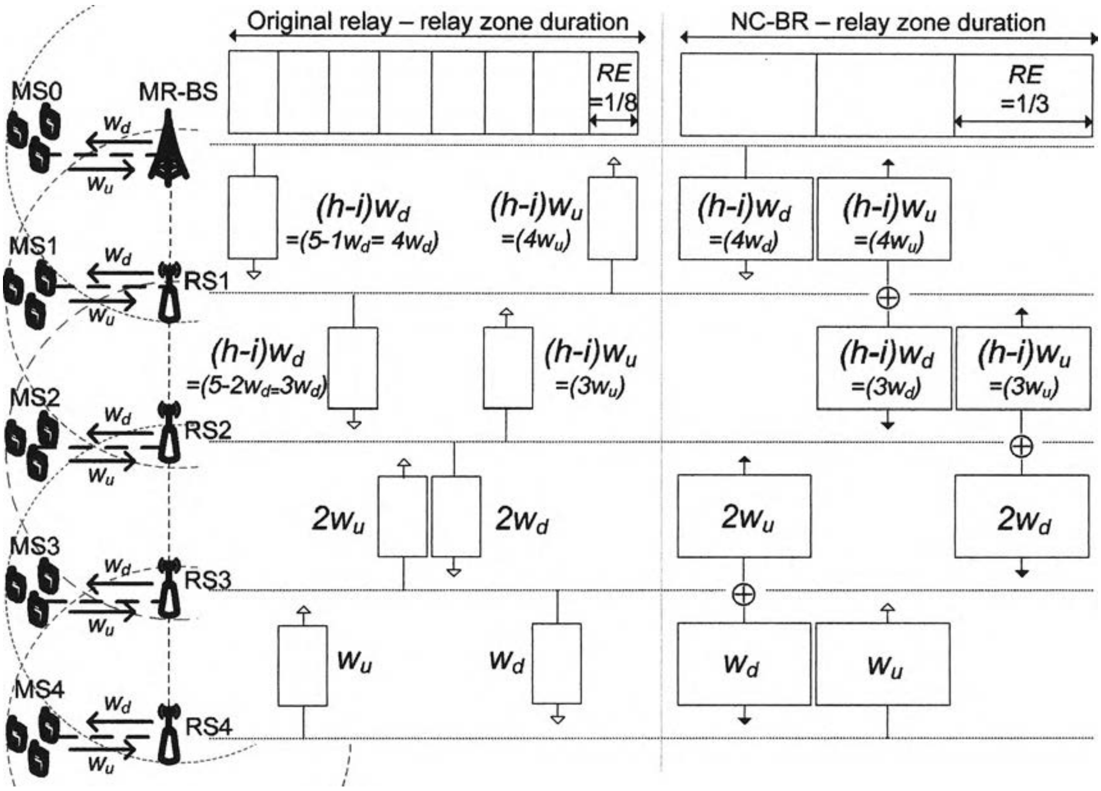
For simplicity, all MR-BS and RSs are assumed to operate in the same transmission power and ranges. Highest modulation is used on all wireless backhaul links. The coverage and interference ranges of each RS can cover the super-ordinate RS/MR-BS, the sub-ordinate RS and MSs of the RS.

Table 3.1: Notations used in analysis.

Term	Definition
h	Number of total hops
i	Hop count of from MR-BS to considered RS
w, w_d, w_u	Traffic between MR-BS and each MS, Downlink traffic, Uplink traffic
TP_{MSi}	Average throughput on MS_i
F_d	Frame duration
R_d	Relay zone duration
B_w	Channel bandwidth
RE	Relay zone efficiency
ALD_i	Access link maximum throughput archived
D_{MSi}	Average end-to-end delay on MS_i
DM_i	Base station/Relay station operation delay
DP_i	Propagation delay on access link i
DPR_i	Propagation delay on relay link i

The rest of this section is organized as follows. In 3.1.5.1, the analytical model of relay link efficiency limit, that effect to the performance of both original relay and NC-BR is described. In 3.1.5.1, the throughput analysis model and provide the comparison between both approaches are described. In 3.1.5.3, the end-to-end delay analysis model and comparison of both approaches are explained.

3.1.5.1 Relay zone efficiency



Original relay – relay zone duration
NC-BR – relay zone duration
RE =1/8
RE =1/3
MS0
MR-BS
MS1
RS1
MS2
RS2
MS3
RS3
MS4
RS4
w_d
w_u
$(h\text{-}i)w_d$ $=(5\text{-}1w_d= 4w_d)$
$(h\text{-}i)w_u$ $=(4w_u)$
$(h\text{-}i)w_d$ $=(4w_d)$
$(h\text{-}i)w_u$ $=(4w_u)$
$(h\text{-}i)w_d$ $=(5\text{-}2w_{d}{=}3w_d)$
$(h\text{-}i)w_u$ $=(3w_u)$
$(h\text{-}i)w_d$ $=(3w_d)$
$(h\text{-}i)w_u$ $=(3w_u)$
$2w_u$
$2w_d$
$2w_u$
$2w_d$
w_u
w_d
w_d
w_u

Figure 3.7. Traffics handled by the relay zone efficiency (*RE*).

Due to the NC-BR has a higher utilization in the relay zone than the original relay, then, the higher *RE* is obtained. Figure 3.7 shows an *RE* of 5-hop scenarios of the original relay and the NC-BR scheme that are 1/8 and 1/3, respectively. The *RE* of the original relay scheme can be expressed by Equation (2).

$$RE = \frac{1}{2(h-1)} \tag{2}$$

The *RE* of 3-hop and 4-hop scenarios of the original relay scheme are 1/4 and 1/6, correspondingly. On one hand, when RS hops are increasing in the scenario, the *RE* will decrease in the case of the original relay scheme. On the other hand, the *RE* for NC-BR is always 1/3 and independent from the number of RS hops.

3.1.5.2 Throughput Analysis

The fair comparison between the NC-BR and the original relay scheme started by making the equal size of the following parameters: $F_{d,}$ R_d and w. Additionally, the ratio of both $\frac{W_d}{w}$ and $\frac{R_d}{F_d}$ are assumed to be 1/2. Therefore, the TP_{MS_i} of the downlink traffic can be expressed as

$$TP_{MS_i} = \begin{cases} (\frac{F_d - R_d}{F_d})(\frac{W_d}{W}) \cdot B_w & ,i = 0 \\ RE \cdot (\frac{R_d}{F_d})(\frac{W_d}{(h-1)W_d})(\frac{B_w \cdot F_d}{i}) & ,i > 0. \end{cases} \qquad (3)$$

According to the Equation (3), the term $(\frac{W_d}{(h-1)W_d})$ is the maximum traffic handled by the relay link, and $(\frac{B_w \cdot F_d}{i})$ refers to the end-to-end data rate through the multi-hop link. The end-to-end throughputs of 3-5 hops scenarios will be show in Figure 4.1.

3.1.5.3 End-to-end Delay Analysis

The proposed NC-BR and the frame structure can reduce the end-to-end delay. By reordering the transmission sequence of RSs, they allow traffic flow to travel up to 3 RS hops in the single MAC frame. The result shows that they reduce the delay in the case of the multi-hop flows by up to 3 MAC frames duration. Hence, the value of i' is equal to i in the original relay scheme. However, in the NC-BR, the value of i' is equal to (i div 3). The average end-to-end delay on each hop can be expressed by Equation (4).

$$D_{MS_i} = \begin{cases} DM_0 + DP_0 & ,i = 0 \\ (i_r F_d) + \sum_{j=0}^{i} DM_j + \sum_{k=0}^{i-1} DPR_k + DP_i + (\left\lceil \frac{w_d}{(h-1)w_d} \right\rceil F_d \cdot RE) & ,i > 0 \end{cases} \qquad (4)$$

Referring to the Equation (4), the delay from the multi-hop frame structure is given by$(i_r F_d)$. The traffic delay which is the buffer clearing time of RSs, is $\left(\left\lceil \frac{w_d}{(h-1)w_d} \right\rceil F_d \cdot RE\right)$. The operation delay and the propagation delay are $\sum_{j=0}^{i} DM_j$ and $\sum_{k=0}^{i-1} DPR_k + DP_i$, respectively.

3.2 Path finding algorithm: HIRN and ATTPS

The proposed solution developed to solve the best travel-time on directed weighted-graph problem. The objective of this research is to present a new efficient algorithm to solve single source shortest path problems in a large directed weighted-graph using an adaptive travel-time path selection algorithm. General shortest path algorithms, which examine a large part of the whole graph for each shortest path finding, are very time consuming if the considered network is large. The proposed method, the adaptive travel-time path selection algorithm, considers a weighted-graph as a hierarchical index on the graph layering architecture, where the first layer consists of all main nodes, and sub-layers are branches of each main node, or other sub-layers. This proposed solution has shown that applying this technique can identify the significant traveling path in the hierarchically layered weighted-graph which is faster than the available shortest path algorithms. Additionally, this technique can also apply to solve problems of any types of network architectures.

Solving travel-time path selection is significant interest for application that works in spatial data networks [1, 7-8, 10, 47-95]. The travel-time is more concerned than traveling distances. Therefore, the meaning of the shortest-path is referred to the smallest time count spending from the starting point to the required destination. Various researches [1, 3-10] had considered the shortest-path and nearest neighbor queries in the weighted-graph without the situation of the network congestion. Previous works result, in techniques to compute Euclidean space, have been proposed to compute nearest neighbor queries in spatial networks. These methods extend nearest neighbor queries by considering spatial network distances. Some researchers [7-8] also considered the static travel-time in the weighted-graph, and the geo-position locator that concerns on path selection of spatial network quires. However, these existing techniques consider the weighted cost of paths using static information of path distances, or immediate travel-time. Thus, it is not enough to perform a future traveling plan. Therefore, the guiding system needs to keep records of the travel-time submitted by each node processer at a datacenter and retrieves the time estimation information for guiding the traveling plan to users.

This new design hierarchical index data on the graph and the path selection algorithm is made use of recording real travel-time data set that reflects by mobile units in the high congestion weighted-graph. The real travel-time data set that collected from mobile units provides much accurate results than existing algorithms in case of a highly congestion network, e.g., in urban road or in the most busy city in the world. Additionally, the proposed algorithms can also be applied to broad range of applications e.g., path selection of travel planning on weighted-graph, logistic planning and mobile agent traveling in high congestion networks, includes problems of any types of network architectures.

This research work will contribute on two levels in this part finding part. First, this research describes a general data model for Hierarchical Index Road Network (HIRN). Hierarchical index weighted-graphs indexed on the top of a digital map of the weighted-graph, consisting of recorded traveling time of real traveled data submitted from mobile units, separated in time slots.

Second, this research proposed the adaptive travel-time path selection algorithm (ATTPS) on hierarchical indexed weighted-graph that performs faster path selection in term of traveling time than the existing shortest path algorithms.

3.2.1 Definition and assumption of network

In this section, general definitions of a graph network are stated and network assumptions of this experiment are elaborated.

- *Definitions*

DEFINITION A.1. *A network $G = (V, L)$ is described by an ordered pair of finite sets, V and L. The first set, V, represents the set of all nodes. The second set, L, represents the set of all directed links. Let (u, v) denote the directly connected link starting at the node u and ending at the node v. Its positive length is denoted by $D_{u,v}$.*

DEFINITION A.2. *A network $G_s = (V_s, L_s)$ is called a subnetwork of a network $G = (V, L)$ if*

1) *$V_s \subset V$ and $L_s \subset L$, and*

2) *$(u, v) \in L_s$ if and only if $u, v \in V_s$ and $(u, v) \in L$.*

According to the above definition, if a set of nodes in the given network is chosen, the associated links will be known accordingly to form a subnetwork. Therefore, in the following content that defined a set of nodes, it will refer to a subnetwork for convenience.

THEOREM A.1. *For a given network G = (V, L), a path P is a finite sequence of links. Let $P = \{(u_1, u_2), (u_2, u_3), \ldots , (u_{k-1}, u_k)\}$, $u_i \in V$, $i = 1, 2, \ldots, k$; $(u_i, u_{i+1}) \in L$, $i = 1, 2, \ldots, (k - 1)$.The path length D(P) of the path P is $D_{u1,u2} + D_{u2,u3} + \ldots + D_{uk-1,uk}$.*

Proof:

Let P be a finite sequence of links, $P = \{(u_1, u_2), (u_2, u_3), \ldots, (u_{k-1}, u_k)\}$.

The path length of $(u_i, u_j)= D_{ui,uj}$ for i=1,...,k-1 and j =2,...,k.

Therefore, path length D(P) of the path P is $D_{u1,u2} + D_{u2,u3} + \ldots + D_{uk-1,uk}$.

DEFINITION A.3. *For a given network G = (V, L), a path P is a finite sequence of links $P = \{(u_1, u_2), (u_2, u_3), \ldots , (u_{k-1}, u_k)\}$ with $u_i \in V$, $i = 1, 2, \ldots, k$ and $(u_i, u_{i+1}) \in L$, $i = 1, 2, \ldots, (k - 1)$. The hop length (or hop distance between node u_1 and u_k), H(P), of the path P is defined as H(P) = D(P) with $D_{ui,ui+1} = 1$, $i = 1, 2, \ldots, k - 1$.*

DEFINITION A.4. *A network $G = (V, L)$ is said to be connected if, for any two nodes, u, $v \in V$, there exists a path, $P = \{(u, u_1), (u_1, u_2), \ldots, (u_k, v)\}$, $u_i \in V$, $i = 1, 2, \ldots, k$.*

DEFINITION A.5. *Two subnetworks (with each being connected), $G_{s1} = (V_{s\,1}, L_{s\,1})$ and $G_{s2} = (V_{s2}, L_{s2})$, G_{s1}, G_{s2}, $G_{s2} \subset G = (V, L)$, are said to be interconnected. $\exists$ a link set L_s with $L_{s1} \cup L_{s2} \subset L_s \subset L$, $\ni \{V_{s1} \cup V_{s2}, Ls\}$ is a connected subnetwork/network.*

- *Network Assumption*

ASSUMPTION B.1. *The networks studied in the paper are hierarchically clustered unless specified otherwise.*

ASSUMPTION B.2. *Each physical node in a network has a unique ID taken from* {1, 2, ..., *n*}, *where n is the number of the nodes in the network.*

ASSUMPTION B.3. *Each physical link in a network has a unique ID taken from* {1, 2, ..., *n*}, *where n is the number of the nodes in the network.*

3.2.2 Hierarchical Index Road Network (HIRN)

In this section, outlines of the system architecture and the proposed design of the hierarchical index on road network (HIRN) are described. The HIRN is responsible for collecting and managing information of real travel-time from a mobile unit member. Thus, the real travel-time can represent high accuracy of the high congestion network. Moreover, an adaptive travel-time path selection algorithm (ATTPS) on a hierarchical index road network is developed. This algorithm is able to perform faster path selection in the term of a travel-time rather than the existing shortest path or nearest neighbor algorithms. The ATTPS algorithm will be discussed in the next section.

3.2.2.1 System infrastructure

Figure 3.8 illustrates the system infrastructure of the proposed design. The design considers a *mobile unit* with a mobile computer (e. g., PDA) that contains a standard vector digital map database. Additionally, this mobile computer has no limitation of power, equipped with a GPS receiver; the GPS is used for obtaining a stream of location information.

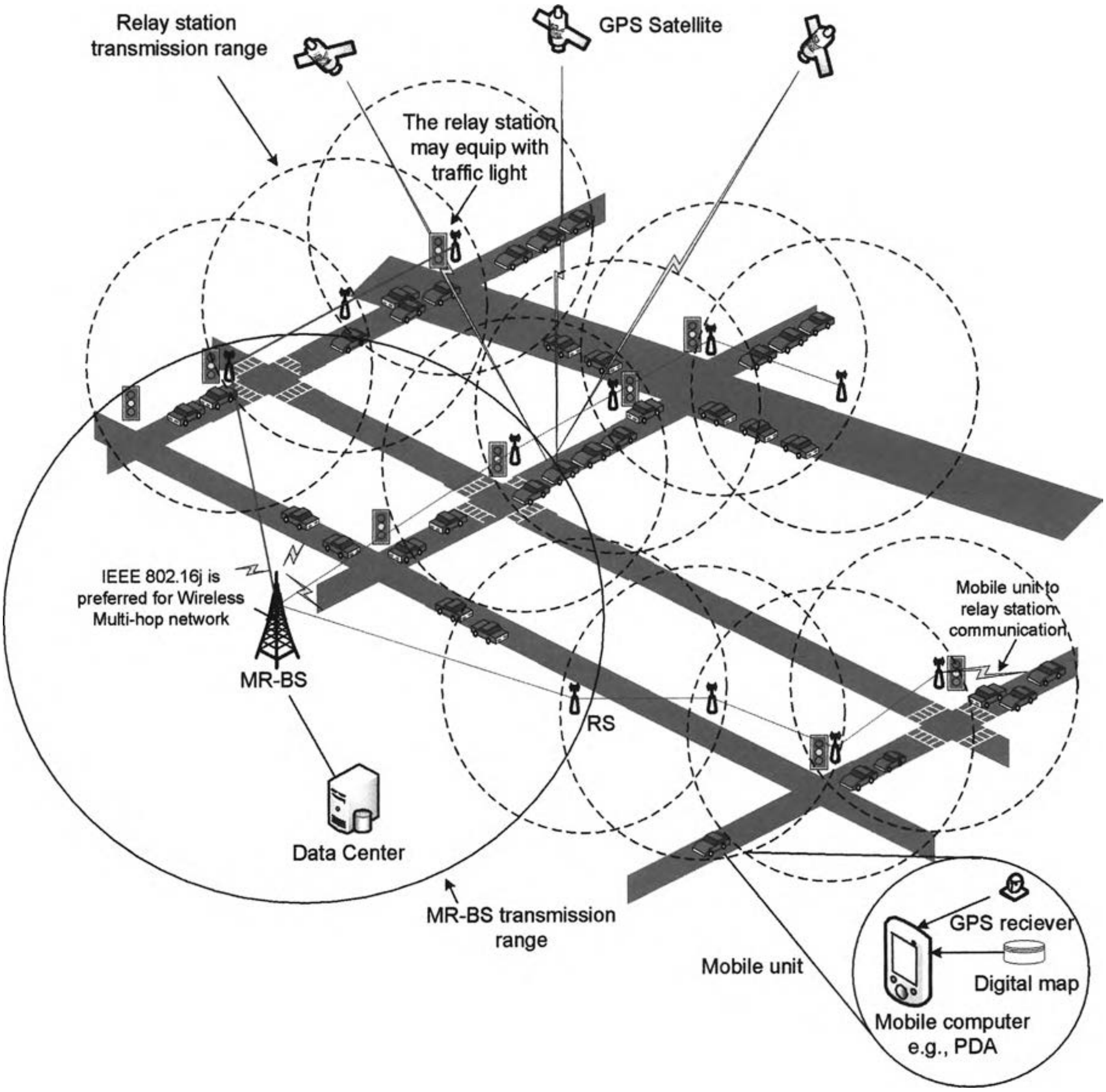

Relay station transmission range
GPS Satellite
The relay station may equip with traffic light
IEEE 802.16j is preferred for Wireless Multi-hop network
MR-BS
Data Center
MR-BS transmission range
RS
Mobile unit to relay station communication
Mobile unit
GPS reciever
Digital map
Mobile computer e.g., PDA

Figure 3.8. The system infrastructure.

According to Figure 3.8, the system architecture consists of datacenter, mobile unit and communication protocol. The *datacenter* is a host of the proposed system, receiving and organizing information sending by mobile units. The assumption is that the datacenter is always available on the network and has no limitation of its storage space.

The digital map on mobile units is usually proprietary by vendors and it is not

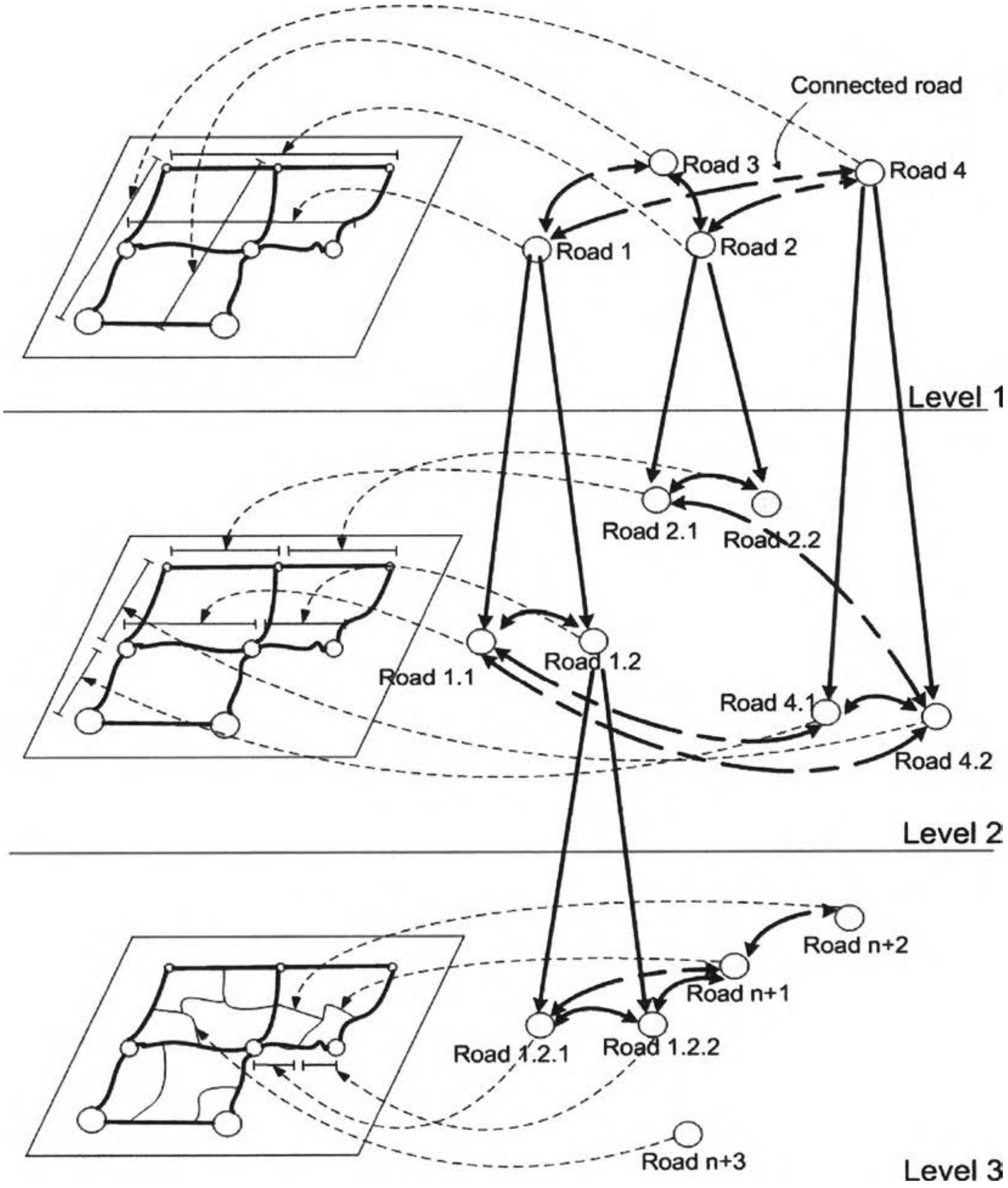


Figure 3.9: Hierarchical index on road network (HIRN), map index level.

Table 3.2: The hierarchical index on road network (HIRN), data sample.

PathID	Property	Values
Road1	**Road type**	Main road, 2 way
	Vertices set	{*A*, *D*, *C*, *B*}
	Distance (km)	{21.4}
	Travel-time table (minutes)	{2, 2, 2, 4, 6, 10, 11, 10, .. , T} (48 * 7 time slot per Travel time table, separate by 48 half of hours * 7 days of week) summarized from lower level
	Parent	Root
	Child	{Road1.1(*A*, *D*), Road1.2(*D*, *C*), Road1.3(*C*, *B*)}
	Connected Rd.	Road 3, Road 4
Road1.2	**Road type**	Main road, 2 way
	Vertices set	{*D*, *C*}
	Distance (km)	{11.2}
	Travel-time table (minutes)	{0.7, 0.7, 0.8, 1, 1, 2, 4, 5, .. , *T*}
	Parent	Road1
	Child	{Road1.2.1(A, D), Road1.2.2(D, C)}
	Connected Rd.	Road 1.1, Road 1.3, Road 3.1, Road 3.2, Road n.1, Road n.2

(Data representation and sample values of
Hierarchical index on road network (HIRN) of Figure 3.9.

compatible with others. However, the consequence of implementing a datacenter is that boundaries among different mobile vendors are broken. Thus, the digital map from the datacenter can serve heterogeneous mobile units from different vendors without problems.

The *wireless communication* between a mobile unit and the datacenter is not restricted; long length and small response time are preferred. The 802.16j wireless network is considerably suitable to the considered system. In an urban area, the wireless access point could be implemented with a traffic light or a traffic information display board. So, this will cover a large amount of users in a wide area road network.

However, the multi-hop wireless network is not considered in this case, due to uncertainty of medium-act mobile unit availability. Wireless cellular networks (e.g., GPRS) are needed when another short length wireless network is unavailable. In order to send information through the HIRN, the structure of data must be defined with functions of all components.

3.2.2.2 HIRN data structure

Hierarchical index on road network (HIRN) is a set of hierarchical indexes on a digital map. Figure 3.9 shows the HIRN map index levels with Table. 3.2 explaining data representation and sample values of the HIRN. From the simple definition of a road network represented by a directed weight graph, vertices represent a junction on the road network and edges represent roads, with the distance as a weight cost.

Using the hierarchical structured index leads to two advantages for the systems. First, the system can keep multiple information of each road, including each sub road path. This information is considered as layers of all road paths. Second, the proposed algorithm, ATTPS, can easily process each layer for its computation. The main investigated information was the congestion information (*C*) which can be calculated from the average of travel time (*T*) divided by the actual distance (*D*); $C = T/D$. The following will be general descriptions of initiating and maintaining processes of the HIRN. Note that all processes were running at the datacenter.

The initial processes are

- *Step 1:* analyzing the digital map in the datacenter; each level of network (main roads, medium roads and small roads) is categorized.

- *Step 2:* generating index nodes of each level; nodes represent roads or sub roads (in the higher level), connected roads will be indicated.

The maintaining process is

- *Step 1:* updating the travel time table; when a mobile unit submits new information of their travel speeds, HIRN will update the travel time table at the highest level of the index (the smallest path of road) and in a particular time slot (a slot contains information of specific day and hour).
- *Step 2:* summarizing of time value of child nodes and update to ancestor nodes
- *Step 3:* re-initiating the HIRN; when a digital map has been updated, the HIRN initial process is required.

3.2.2.3 Datacenter function

The datacenter replies paths and several alternative paths with time estimation of each requested query by a mobile unit. In the datacenter, it contains a standard digital map and indexed by the HIRN.

3.2.2.4 Mobile unit function

Mobile units can automatically submit their travel-information and gain advantages from the system by receiving the most accurate shortest travel-time path. In additional, a simpler output from this system can be displayed on the outdoor traffic information board, on digital map displayed thru a web site and also available for other system thru the web service protocol. Then, other road network users who do not have the mobile unit equipments still obtain some benefits from the proposed system.

3.2.2.5 Communication protocol

Communication protocol function: in our system, this section explains how mobile units communicate and exchange information with the datacenter. The mobile units are usually PDAs, equipped with a GPS receiver and a digital map, implemented on vehicle and covered by the metropolitan wireless network like IEEE 802.16. The TCP protocol is used mostly for request/reply communication messages between mobile units and a datacenter due to reliable of TCP, e.g., a path selection requests and travel-time shortest path is replied to the mobile units. UDP can also be used in the case that massively data are submitted, due to minimal overhead is needed and some lost of messages are acceptable. The case of UDP used is all mobile units continuing submit their positions and travel speeds along with timestamps to be stored and categorized at the datacenter.

The following scenario is the situations of information exchange between mobile units and the datacenter where the datacenter is assumed to be available on the network at all time.

- First, the initial state of mobile units, users turn on their devices, acquiring network connections and a GPS locators.
- Second, when mobile units ready and moves, they keep sending positions and travel speeds along with timestamps to the datacenter.
- Third, when a user wants to perform a query for the shortest travel time to the destination. The mobile unit will submit a request message along with its source and destination locations. The datacenter replies paths and several alternative paths with an estimated time of each path (the source and the destination) back to the mobile unit.

Referring to Figure 3.8, the processes of ATTPS can be described as integrated functions of components as follows.

- A mobile unit requests the shortest-time path to the datacenter, ATTPS (*source, destination, departure_time*), the default value of *departure_time* is the current time value.
- The datacenter returns the selected paths back to the mobile unit (set of vertices) with the estimated travel-time of those paths. The datacenter

presents at least two alternative paths: the minimum travel-time path, and the shortest-distance path. Additionally, users can request more than two alternative default paths. In some situation, the minimum travel-time path and the shortest-distance path could be the same.

3.2.3 Adaptive Travel-Time Path Selection (ATTPS)

Shortest path algorithms are essential for solving network or network-like congestion problems. Network optimal routing algorithms heavily are relied on the shortest path algorithms to distribute the communication traffics and the shortest path algorithms are usually the computational bottleneck of the optimal routing algorithms.

In the travel-time shortest-path, the weight cost of a graph is considered in the term of the travel time. All road junctions are labeled by coordinates. Then, in the algorithm, the next visited vertex is selected based on the next lowest weight costs. Therefore, the model of the weight cost, W, is equal to $(T + T_{avg} * E)$ where T is a travel-time to the next considered vertex, T_{avg} is the average travel-time rate on that particular type of road, and E is the Euclidean distance from the next considered vertex to the destination. According to Euclidean distance, this method provides a position-aware shortest-path.

Let Q be the set of the returned paths in Figure 3.10. The following explains the ATTPS(x, y) function.

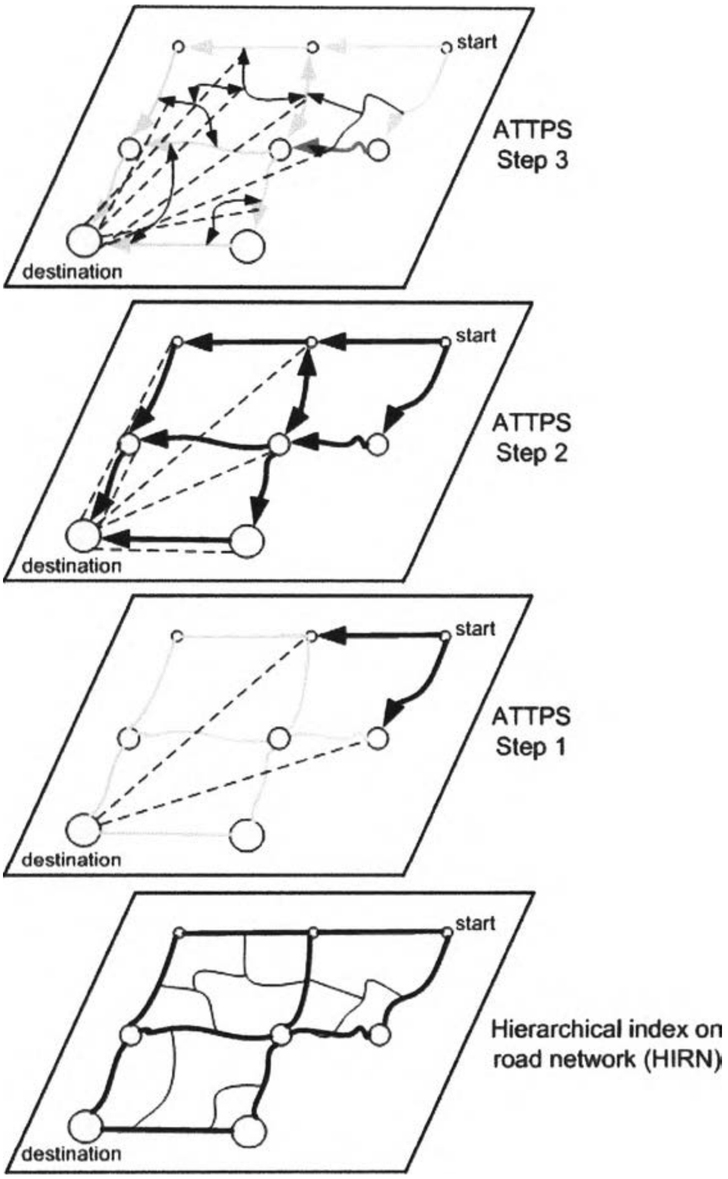


Figure 3.10: Adaptive Travel-time Path Selection algorithms (ATTPS), steps.

1) The ATTPS receives coordination of the source and the destination from the main program, snaps the coordinates on the road network with a digital map at the datacenter.

2) Generally, the ATTPS will find the path on the main road due to the fact that without the congestion problem the main road will always be the best path. However, if the source or the destination is not on the main road then the ATTPS will find the nearest vertex on the main road for an input instead of the actual one.

3) In Step 1, the ATTPS loads vertices from the 1^{st}-level index of the HIRN (main road level).

4) The ATTPS searches for travel-time shortest paths based on the weight cost, *W*, which is equal to $(T + T_{avg} * E)$; then, the ATTPS add the found vertices to *Q*.

5) In Step 2, the ATTPS repeats the search procedure until the paths from the source to the destination have been identified. The result of the travel-time shortest path of the main road will be in *Q*.

6) In Step 3, the ATTPS loads vertices from the 2^{nd}-level index of the HIRN.

7) The ATTPS recursively fetch each member of *Q*, and used as a new source of the ATTPS. If the new weight cost to the destination is smaller than the existing weight cost from the particular member of *Q* to the destination, members of *Q* are updated. The shortcut to the destination from each vertex will be displayed under this process, if they exist.

8) After all vertices in *Q* have been processed, the ATTPS will move to the next level of the HIRN, until the last level.

9) *Q* and *W* are returned and members of *Q* are the travel-time shortest path.

The above algorithm has time complexity equal to *O(mn)* where *n* is the number of vertices that are located only on HIRN level 1 and *m* is the number of sets of potential vertices to be shortcut to all vertices of HIRN level 1. The proposed algorithms process much less vertices than existing methods because of the advantages from the combination of the hierarchy index structure, Euclidean distance and stored historical weighted data.

The ATPS algorithm uses the best-first search and selects the least-cost path on each HIRN level from a source node to a destination node, out of multiple possible paths. It uses a distance-plus-cost heuristic function, denoted *f(x)*, to determine the order in which the search visits nodes in the graph. The distance-plus-cost heuristic is a sum of two functions:

- the path-cost function, which is the cost from the source node to the current node, denoted by $d(v, v^*)$.
- an heuristic estimate of the distance to the goal, by Euclidean distance, denoted by $h(v, v^*)$.

Where $d(v, v^*)$ denotes the length of the focusing edge from considering vertex v to the destination vertex v^*, the heuristic h satisfies the condition $h(v, v^*) \leq d(v, v^*)$ for every edge (v_i, v^*) of the graph, then h is called consistent. The ATPS on each HIRN level is equivalent to running A* and Dijkstra's algorithms with the reduced cost:

$$d'(v, v^*) := d(v, v^*) - h(v, v^*).$$

The time complexity of ATPS depends on the heuristic. In the worst case, the number of nodes expanded is polynomial in the length of the shortest path on each HIRN level, and the heuristic function h meets the following condition:

$$|h(v, v^*) - h^*(v, v^*)| = O(\log h^* (v, v^*))$$

Where h^* is the optimal heuristic, the exact cost to get from x to the goal. In other words, the error of h will not grow faster than the logarithm of the "perfect heuristic" h^* that returns the true distance from v to the v*,

The reduced cost of the ATPS on each HIRN level, $d'(v, v^*)$, is mn. Since the ATPS runs on multiple levels, the comprised time complexity of the ATPS on all HIRN levels is $O(mn)$.

The proposed method on the data transfer method and the path finding algorithm will be presented in the next chapter.

3.3 Integration of the path finding part and the data transfer part.

The centralized architecture is preferred since the ATPS on HIRN, the path finding part, requires a set of historical data on the datacenter to calculate the high accuracy path finding result. However, by storing a large amount of the historical data at the datacenter is a drawback of the proposed path finding method since the large storage space at the datacenter is required.

On the other hand, the system architecture can also be designed and implemented in the distributed fashion, if a fast and reliable wireless multi-hop network is deployed, as shown in Figure 3.8. The implementation of distributed architecture which is based on the wireless multi-hop network, the NC-BR, allows the datacenter to wirelessly acquire the historical data from candidate mobile units. Then, the high accuracy shortest-path algorithm can be calculated at the datacenter without a large amount of the historical data stored. Consequently, the high accuracy path finding algorithm can be succeed with low storage space required. The distributed architecture allows the ATPS on the HIRN to reduce the storage space to E+V which is equivalence to the Dijkstra's algorithm.

According to Section 3.2.2, the centralized HIRN was described, the storage space required at the datacenter was categorized on to *N* levels of HIRN. The HIRN on level *i* requires the number of edges E_i and the number of vertices V_i, which $E_1 < E_2 < \ldots < E_n$ and $V_1 < V_2 < \ldots < V_n$, respectively. A node on each HIRN level represents the categorized edges on the digital map. Let *t* represent the historical data for each node on the HIRN. Then, storage space of the centralized HIRN can be defined as $(tE+V)$.

On the other hand, the distributed architecture does not require the HIRN store the historical data *t* because it allows the datacenter to wirelessly acquire the historical data from candidate mobile units. Thus, the storage space of the distributed HIRN can be defined as only $(E+V)$.