An NAK-based Hierarchical ARQ Scheme for Reliable Data Multicast in Integrated Communication and Broadcast Networks

Li Gu∗, Zhisheng Niu∗, Jiping Lv† and Hideya Yoshiuchi†
∗Department of Electronic Engineering, Tsinghua University
Tsinghua National Laboratory for Information Science and Technology, Beijing 100084, P.R.China
Email: gul07@mails.tsinghua.edu.cn, niuzhs@tsinghua.edu.cn
†Hitatch (China) Research & Development Corporation
Email: jplv@hitatch.cn hyoshiuchi@hitachi.cn

Abstract—In this paper we propose a hierarchical architecture of the return channel in ICBN and an NAK-based hierarchical ARQ scheme (H-ARQ). The H-ARQ scheme can alleviate the NAK implosion and balance the retransmission traffic between broadcast and communication networks. The users are divided into regions in the communication network, and each region has a Local Recovery Router (LRR), which cooperates with the Data Broadcast Server (DBS) to decide the retransmission method. The impact of the ARQ scheme on the data broadcast performance is analyzed, and the NAK arrivals in ICBN are modeled and simulated. We also propose an algorithm to optimize the parameters and cost functions in the H-ARQ scheme. At last we give some simulation results to validate the improvement of the H-ARQ scheme running on the hierarchical system architecture.

I. INTRODUCTION

In recent years, with the development and prevalence of mobile terminals, the demand for interactive multimedia services in the mobile environment has been increasing rapidly. The Integrated Communication and Broadcast Networks (ICBN) is suitable for the broadband asymmetric services. In [1], the architecture and testbed of ICBN is introduced. The services are divided into 4 categories: digital TV, video on demand, data on demand and Internet access, and some ARQ schemes are introduced to ICBN for the reliable multicast services. If we simply adopt ACK schemes in ICBN, flood of ACKs from a large amount of users will block the communication uplink. Therefore, the ARQ schemes in ICBN should be NAK based. However, we cannot simply apply traditional NAK schemes to ICBN. Because there are a large amount of users in a multicast group, the NAKs are still likely to occupy a big portion of the uplink bandwidth, and the densely arriving NAK requests will possibly exceed the processing capacity of DBS. This is called the “NAK implosion”.

Several NAK schemes have been proposed to avoid the “NAK implosion”. Protocols like SRM, PGM and MFTP are introduced in [5]. In SRM and MFTP, receivers will wait a random back-off time before sending repair requests, and if the retransmitted packet is received during the back-off time, the redundant requests will be canceled. In PGM, the PGM-Aware Routers are designed to collect and merge NAKs from the corresponding groups of users. However, these protocols are mainly for the multicast in IP networks. In ICBN, since many users share the same broadcast downlink, we also need to consider the downlink traffic reduction. In [2], an XOR scheme is proposed to combine \( n \) retransmitted packets into one packet using the XOR operation. In [1], a modified scheme based on the XOR scheme is proposed to achieve better robustness and lower delay. But in the broadcast downlink where most errors occur in bursts, the performance of the modified retransmission scheme will severely degrade.

In the previous ICBN architecture, terminals send feedback directly to DBS, and the packets will be retransmitted through the broadcast network. However, it is inefficient to retransmit data packets through the broadcast downlink when only a few users are requesting for them. The new hierarchical ARQ scheme proposed in this paper tries to divert part of the retransmission traffic to the communication network and alleviate the “NAK implosion”. The object is to improve the QoS of the reliable multicast services in ICBN.

II. SYSTEM ARCHITECTURE

In this section we’ll introduce an improved architecture of the return channel, and an hierarchical ARQ scheme based on the new architecture. As depicted in Fig.1, an extra processing layer composed of Local Recovery Routers (LRR) is inserted between user terminals and...
DBS. Each LRR collects the user feedback and report to DBS. Thus DBS only needs to see all the LRRs instead of all the users. The LRR can also do the local recovery, which means retransmitting data packets to some users in the cell through the communication network.

The conceptions of local recovery and message consolidation were first proposed in Reliable Multicast Transport Protocol (RMTP) in IP multicast domain [6]. However, RMTP cannot be directly applied in ICBN that is composed of heterogeneous networks. In ICBN, the LRRs must exchange necessary information with DBS, and the latter will decide whether to do the retransmission through communication networks or broadcast downlink.

III. THE NAK-BASED HIERARCHICAL ARQ SCHEME

The Hierarchical ARQ scheme proposed in this paper is composed of three parts - schemes on the terminal side, LRR and DBS, corresponding to the hierarchy of the return channel.

A. Terminal Side

Because in ICBN the uplink and downlink are different networks, a user consuming a reliable multicast service may enter an area with good uplink environment but poor downlink environment. In this case the terminal will find every packet lost, and continuously send NAKs through the return channel. Some strategies must be applied on terminals to avoid the fruitless NAKs when the downlink environment is poor. The below rules with parameters $N_{\max}$ and $T_{\max}$ can achieve the object and reduce the amount of NAK requests from terminals:

1. When an erroneous packet is detected, do not send an NAK request immediately.
2. If more than $N_{\max}$ consecutively received packets are erroneous, quit the reception and display error messages.
3. If no packets are received in $T_{\max}$, quit the reception and display error messages.
4. When a good packet arrives after several erroneous/lost packets, send an NAK for the previous errors.

```
 1 2 3 4 5 6 7 8 9
10 11 2
```

Fig. 2. Improved Architecture of ICBN

A short section of receiving progress is given in Fig.2 as an example. Every packet is identified by its sequence number and referenced as Packet$(i)$. When Packet$(3)$ is correctly received, the terminal will send an NAK for Packet$(2)$. Then Packet$(5 - 7)$ are lost, but no NAK is sent out until Packet$(8)$ is correctly received. Then an NAK for Packet$(5 - 7)$ is constructed and sent out. After the retransmission of Packet$(2)$ is correctly received, $N_{\max}$ consecutive packets are erroneous. The terminal estimates that the data broadcast cannot be normally received in the current downlink environment and quits the reception with error messages.

B. Local Recovery Router

LRR has two important functions: Consolidating multiple NAKs and doing local recovery when necessary. In one region there may be $k$ NAKs from different users, which can be referenced as NAK$(i, src_1)$, NAK$(i, src_2)$, ... , NAK$(i, src_k)$. $i$ is the sequence number of the packet, and $src_j$ is the $j^{th}$ terminal’s identification. After the consolidation, the LRR constructs one message referenced as NAK$(i, cost, rid)$, in which $i$ is the packet sequence number, $cost$ gives the cost of local recovery in this region, and $rid$ is the region identification.

LRR has a local cache to store recent data packets, and monitors the IP traffic of the communication network in its region. The server utilization ratio $\rho$ describes the traffic load of the network. We define functions $C_u(\rho)$ and $C_m(\rho)$ to describe the instantaneous cost incurred by the data block of unit length. $C_u$ represents the cost of IP unicast and $C_m$ represents the cost of IP multicast. Obviously $C_u$ is always lower than $C_m$, but when the number of NAKs increases, every retransmission will incur the cost of $C_u$ while $C_m$ will not be affected much. We also set a threshold $C_{th}$.

When the NAKs of a packet from $K$ users arrive to LRR, LRR first calculates $C_u$ and $C_m$ with the current server utilization ratio $\rho$, and chooses the minimum of $KC_u$ and $C_m$ as the local recovery cost $C$ of this region. If $C < C_{th}$, LRR will directly do local recovery without reporting to DBS, which saves extra bandwidth of the uplink. Otherwise an consolidated NAK will be sent to DBS.

C. Data Broadcast Server

DBS monitors the broadcast channel and calculates the throughput and MRT. Similar to $C_u$ and $C_m$, we define a function $C_B(\varepsilon)$ to describe the instantaneous cost incurred by retransmitting data of unit length through the broadcast network. $\varepsilon$ is the current retransmission proportion in the broadcast downlink. After collecting NAKs from all LRRs, DBS will calculate the overall cost of local recovery $C_{LR}$, compare it with $C_B$ and decide whether to retransmit the required packet through the broadcast network or the communication network. The decision making process is depicted in Fig.3.

![Fig. 3. Decision Progress of LRR and DBS](image-url)

An algorithm to calculate $C_{LR}$ is given below: Assume that $K$ regions send NAKs and the local recovery costs of
these regions are $C_1, C_2, \ldots, C_K$. First calculate the mean cost $\mu$ and variance $\sigma$, and then calculate the overall cost using the formulas below:

$$C_{LR} = \begin{cases} \beta \mu K, & \text{if } \mu + \sigma \leq C_0 \\ \beta \mu K + \gamma (\mu + \sigma - C_0)^2, & \text{otherwise} \end{cases}$$

(1)

In the above expression $\beta$ and $\gamma$ are adjustable coefficients. $\mu K$ is the sum of regional costs. If all regions have low costs for local recovery except one (e.g. the $K^{th}$ region) whose cost is rather high, $\mu K$ will not be very large, and DBS is more likely to choose the local recovery scheme, so the $K^{th}$ region has to retransmit the packet in the communication network in spite of the high cost. To achieve better fairness, when the IP traffic in some regions is heavy, we are willing to do the repair through the broadcast downlink even though the mean cost is not so high. In the above formula $C_0$ is defined as a cost threshold. If $\mu + \sigma$ is larger than $C_0$, extra cost will be added to $C_{LR}$.

IV. MATHEMATICAL MODELING AND OPTIMIZATION

A. ARQ’s Impact On the System Performance

In broadcast systems, the on-demand services are usually broadcast in the way of carousel, and the Mean Response Time (MRT) of a request is an important parameter to measure the QoS [8][9]. In the hybrid networks, the return channel helps reduce the MRT. Suppose there are M services, each with the demand probability $p_i$ and length $l_i$, [9] gives the optimal MRT:

$$t_{optimal} = \frac{1}{2} \left( \sum_{i=1}^{M} \sqrt{p_i l_i} \right)^2$$

(2)

$t_{optimal}$ is an theoretical upper bound, and [9] also gives a concrete broadcast scheduling algorithm to approach the optimal performance.

When ARQ is applied to ICBN, the DBS is able to get feedback of the receiving status. Redundant packets are put on the broadcast network to repair the transmission errors. As a result, $l_i$ is stretched to $l_i' = l_i(1 + \varepsilon)$, where $\varepsilon$ is the percentage of retransmission packets. As the number of users increases, $\varepsilon$ will approach 1 or be even larger. According to (2) the MRT with broadcast retransmission will be more than twice the original value. The efficiency of the broadcast downlink will fall sharply because there are too many retransmitted packets which make sense to only a few users who lost the original packets.

B. NAK Arrival In ICBN

Some field trials and error modeling of the DMB-T broadcast channel has been introduced in [4]. Here we adopt the GE model with parameters $P_{GB}$ and $P_{BG}$ given in [4].

Fig.4 depicts the NAK count of every packet when 2000 users are receiving the same service and the PER is about 0.1%. It can be seen that a packet usually has less than 2 NAKs from different users. Repairing these packets through the communication network will not cost much, but it is very inefficient to do the retransmission through the broadcast network. However, a few packets have many NAKs, and repairing them in the broadcast way may be a better choice. The problem is how to decide whether to retransmit a packet through the communication network or the broadcast network. This should depend on the NAKs from terminals and the traffic load of the networks.

C. Modeling and Optimization

The Hierarchical ARQ (H-ARQ) scheme proposed here is in fact a trade-off between broadcast and communication resources. Thus both the QoS of data broadcast and the communication network traffic should be considered. We hope to achieve a good balance between MRT and the Return Channel Traffic Intensity (RCTI) by adjusting the parameters and thresholds defined in the previous subsections. Therefore, the optimization problem of the H-ARQ scheme can be abstracted to the minimization of MRT under the restriction of RCTI.

The calculation of MRT has been introduced in Section IV-A. With the ARQ scheme, suppose the $i^{th}$ service is requested by $N_i$ users and the broadcast retransmission rate is $\varepsilon_i$, we have:

$$MRT = \frac{1}{2} \left( \sum_{i=1}^{M} \sqrt{p_i l_i} \right)^2 \alpha \frac{1}{2} \left( \sum_{i=1}^{M} \sqrt{N_i l_i(1 + \varepsilon_i)} \right)^2$$

(3)

Next we need to find the relationship between $\varepsilon_i$ and $N_i$ under typical DMB-T environments. Denote the number of regions with $R$, and assume all the users are evenly distributed in these regions, so there're $M = N_i/R$ users in every region. [4] gives the typical values of $P_{GB}(= 0.195)$ and $P_{BG}(= 5.77e-3)$ of DMB-T broadcast channel in the GE model. Model the region as a $(M+1)$-state machine, and state $m$ means $m$ users are in the error state at the current packet. The transition matrix can be derived from the GE model of users.

$$P_{kj} = P(S_n = j | S_{n-1} = k)$$

$$= \begin{cases} \sum_{i=M-k}^{M-k+k-l} C_{M-k}^{j-k+l} P_{GB}^{j-k+i} P_{BG}^{M-j-i} C_{k}^{l} P_{BG}^{l} P_{BB}^{k-l} & j \geq k \\ \sum_{i=M-k}^{M-k+k-l} C_{M-k}^{j-k+l} C_{k}^{l} P_{BB}^{M-j-k+l} P_{BG}^{k-j+i} P_{BB}^{k-l} & j < k \end{cases}$$

(4)

With the transition matrix $P$ we can calculate the probability that $k$ users are in the error state, which is denoted with $P_k$, as is shown in (5).

$$\pi P = \pi$$

$$\pi = [P_0 P_1 P_2 ... P_M]$$

(5)

Then we can calculate the probability $P_{LR1}$ that LRR
does local recovery without sending an NAK to DBS.

\[ P_{LR1} = P(\min(kC_u(\rho), C_m(\rho)) \leq C_{th}) = \sum_{k=1}^{M} P_k P(\min(kC_u(\rho), C_m(\rho)) \leq C_{th}) \]  

(6)

In the above equation, \((1 - P_0 - P_{LR1})\) is the probability that an NAK will be sent from one LRR to DBS. So the probability that DBS receives the current broadcast retransmission can be derived from (1). Denote the current broadcast retransmission proportion with \(\varepsilon\). Assume that \(C_{LR}\) is mainly decided by other IP traffic in the communication network and can be regarded as stable. At the beginning of a data broadcast service, \(\varepsilon = 0\) and \(C_B(\varepsilon)\) is probably lower than \(C_{LR}\). As more packets are retransmitted through the broadcast network, \(\varepsilon\) and \(C_B(\varepsilon)\) will increase, and the balance will be achieved at the broadcast retransmission proportion \(\varepsilon_i\) which meets the condition:

\[ C_B(\varepsilon_i) = C_{LR} \]

\[ \varepsilon_i = C_B^{-1}(C_{LR}) \]  

(9)

Now we have got the broadcast retransmission proportion \(\varepsilon_i\) of the \(i^{th}\) service. For every region, the local recovery proportion is \((1 - P_0 - \varepsilon_i)\), so the RCTI of the \(i^{th}\) service is \(l_i(1 - P_0 - \varepsilon_i)\). Finally, the optimization problem can be expressed as below:

\[ (\beta, \gamma, C_{th})_{\text{opt}} = \arg\min_{(\beta, \gamma, C_{th})} \frac{1}{2} \sum_{i=1}^{M} \sqrt{N_i l_i(1 + \varepsilon_i(\beta, \gamma, C_{th}))^2} \]

\[ \text{when } l_i(1 - P_0 - \varepsilon_i(\beta, \gamma, C_{th})) \leq RCTI_{\text{max}} \]  

(10)

In the above expressions \(RCTI_{\text{max}}\) is the constant upper bound of Return Channel Traffic Intensity, which is decided by the QoS requirement of the communication network. The concrete expression of \(\varepsilon_i(\beta, \gamma, C_{th})\) can be derived from (1) and (4) \(~\sim\) (9).

V. Simulation Results

First we did some simulations to validate the improvement of MRT in our hierarchical architecture. In the simulation environment there is one service containing 10000 data packets, and the GE model is adopted to simulate the broadcast channel. The cost functions \(C_u(\rho), C_m(\rho)\) and \(C_B(\varepsilon)\) have been designed in Section III. The traffic intensity \(\rho\) is supposed to follow the uniform distribution within a certain range between 0 and 1, and the mean value can be adjusted. The other parameters are set as below: \(C_{th} = 1.5, C_0 = 1, \beta = 1, \gamma = 1\). When there is only one service in the broadcast channel, we have \(MRT_{\text{ARQ}} \propto (1 + \varepsilon_i) MRT_{\text{non-ArQ}}\). Thus the relative MRT can be expressed by \((1 + \varepsilon_i)\), which denotes the relative value of MRT compared with the original MRT without retransmission.

The algorithm to make decisions in retransmission is depicted in Fig.3. For comparison, we also simulated the simple ARQ scheme in which local recovery is not adopted.

![Fig. 5. The relative MRT increasing with the number of users](image)

![Fig. 6. The relative MRT increasing with the number of regions](image)
and the traffic of the return channel $\rho$ follows the uniform distribution in the range of $[0, 0.8]$. It's observed that the MRT is reduced to approximately 20% of that without retransmission balancing. That means about 80% of the retransmitted packets are sent through the communication network to a few scattered terminals.

Similar results are displayed in Fig.6 when the amount of users in every region is fixed at 20 and the number of regions increases from 1 to 30. In the simple broadcast repairing system, the relative MRT is affected only by the total number of users, so the solid line obtained is nearly the same with that in Fig.5. The dotted line obtained shows that the H-ARQ scheme also works well with many small regions.

The dashed and dashdot lines in Fig.5 and Fig.6 implies the impact on the retransmission balancing brought by the traffic intensity of the return channel. The dotted line is obtained when $\rho \in [0, 0.8]$, while the dashdot line and dashed line are obtained when $\rho \in [0.1, 0.9]$ and $\rho \in [0.2, 1]$ correspondingly. It can be seen that when $\rho$ increases, the system becomes more intended to retransmit through the broadcast network, so the congestion of the communication network will not be aggravated.

The Return Channel Traffic Intensity produced by the communication network repair is displayed in Fig.7. The RCTI here is the statistical sum of all the user’s retransmission traffic in the communication network, so the value increases with the number of users. Besides, it can be seen that when the traffic intensity $\rho$ of other routine services is higher, the RCTI brought by the retransmission in ICBN becomes lower.

**VI. CONCLUSION**

In this paper, we improved the ICBN architecture, and designed an NAK-based Hierarchical ARQ scheme on the architecture. We modeled the decision making process during the data retransmission. With the new architecture and mechanism, the problem of NAK implosion at the server side and throughput decrease in the broadcast channel due to the retransmission can be largely alleviated, which is of great importance for the practical implementation of the reliable data multicast services in ICBN. We also analyzed the optimization of parameters to achieve the best performance with limited resources.

The simulation results illustrate that the performance of data broadcast is well improved by the H-ARQ scheme. The retransmission traffic in the broadcast network is reduced to about 20%, as a large proportion of it is diverted to the communication network. Because it is more efficient to retransmit the packets required by only a few users through the communication network.

**ACKNOWLEDGMENT**

The research is supported by the EU MING-T Project and National 863 Program.

**REFERENCES**


