Evaluation of Packet Preemption over C-RAN Fronthaul Networks

Ying Yan, Zifan Zhou, Sarah Ruepp, and Michael Stübert Berger
Department of Photonics Engineering
Technical University of Denmark, DTU
Kgs. Lyngby, Denmark
E-mail: yiya@fotonik.dtu.dk

***Abstract*—The Cloud Radio Access Network (C-RAN) is viewed as a new solution with benefits of reduced cost by sharing resources. This is achieved by the separation of the radio part and the radio processing part, where the transport network between them is referred to as front-haul. It is essential to meet the stringent service requirements of protocols running over the front-haul. This paper describes the C-RAN features and challenges. Furthermore, this paper verifies the packet preemption technology in the C-RAN based on both numerical analysis and simulation results.**

Keywords— time-sensitive network (TSN); C-RAN; packet preemption; preemptive queueing

I. INTRODUCTION

The stringent delay and jitter requirements become a crucial constraint for various applications in reality. When the network operators serve the multimedia streaming services, the quality of received video data is degraded if the delay and jitter requirements cannot be satisfied. When the factories operate the machine production line or the robot line over a remote control, the eventual manipulation can be mismatched with the commands if the control signal cannot be transmitted within the demanding delay requirements. When the car manufactories introduce the advanced techniques, such as Infotainment, Telematics and Advanced Driver Assistance System (ADAS) in the vehicle, the expected convenience and safety cannot be ensured by using traditional electronic components and systems without a suitable end-to-end delay guarantee. This paper discusses the improvement on a network switch in order to differentiate and handle critical traffic with low delay.

Packet preemption has been developed by the IEEE 802.1 Time Sensitive Networking (TSN) work group [1]. In TSN, the control traffic can be scheduled and transferred by using a time-triggered method. There is a specific time window reserved for the arrival of a control packet. In allocating the time window to be as close to the arrival time as possible, a preemptive based priority scheduling is supported. The interfered traffic becomes preemptive and is thus allowed to be interrupted during transmission. Therefore, a minimum end-to-end delay is ensured for the control traffic.

In this paper, we integrate the TSN technology packet preemption for the Cloud based Radio Access Networks (C-RAN). In C-RAN, a mobile operator's radio equipment and the controller are separated geographically and the connection link between them is essential to meet the stringent service requirements. We contribute to verify the TSN benefits for the C-RAN based on both numerical analysis and simulation results.

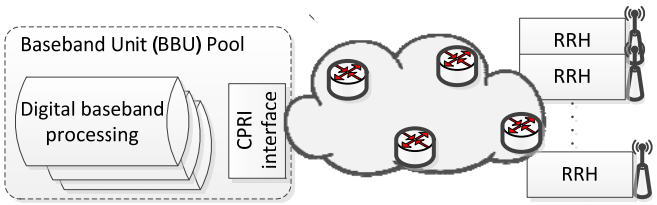

Figure 1. Cloud based Radio Access Networks (C-RAN) architecture

The organization of the paper is as follows: it starts by introducing the background on C-RAN. Afterwards we present related work with respect to time sensitive networks and packet preemption technology. Then, in Section IV, we describe the protocol based on the preemptive queuing system model. Section V includes numerical analysis followed by the presentation of the network simulation that validates the performance from the mathematical modeling. Section VI concludes the paper.

II. CLOUD RADIO ACCESS NETWORK (C-RAN)

The recent introduction of the C-RAN enables the geographical splitting between the Remote Radio Heads (RRHs) and the baseband processing units, which are originally integrated into one device. As shown in Figure 1, the Baseband Units (BBUs) from multiple base stations are pooled into a centralized and virtualized BBU pool. The front-haul network in C-RAN refers to the transport network between the RRHs and the BBU pool, where time-sensitive data and control messages are exchanged [2] [3].

In the C-RAN architecture, the main functions of a traditional base station can be divided into the radio functionalities and the baseband processing functionalities. The antenna module is responsible for power amplifier, frequency filtering and digital processing. The baseband module includes functions such as coding, modulation and Fast Fourier Transform (FFT), etc. Multiple BBUs are placed in a centralized location in order to enable a flexible utilization of BBUs resources and to reduce the operation and maintain cost. The common interface protocol between the RRHs and BBUs is the Common Public Radio Interface (CPRI), which carry transport and synchronization information from BBU to RRH.

The centralization and virtualization of BBUs resources provide benefits in terms of 1) flexible network utilization to cope with the irregular traffic distribution; 2)

 ISBN: 978-1-61208-591-3

reduced deployment cost and power consumption gained from a central location; 3) enhanced cooperative decision making among multiple base station units and small cells.

All the advantages of C-RAN mentioned above cannot be achieved before a series of technical challenges can be addressed and solved [4].

- The expected bandwidth on the front-haul link is increased due to both the overhead generated from the RRH and BBU separation and the converged traffic to the centralized BBU pool.
- The expected latency and jitter requirements become stringent as the smallest as 100-250 μs depending on the function splits between the baseband processing and radio frequency functionalities.
- The inter cell interference among multiple small cells arises and should be minimized or used constructively.
- The current generation CPRI deployment is less than optimal solution due to its constant bit rate.
- The synchronization and timely delivery of traffic need to be ensured for mobile network operation.

The C-RAN front-haul network can be implemented based on either an optical transport solution or the traditional Ethernet network. Compared with the capacity-rich optical solution, the Ethernet-based front-haul network obtains popularity due to the widely spread Ethernet network anywhere. Reuse of existing network infrastructure brings benefits not only on saving deployment cost but also on keeping the consistence and continuity of the standards.

In the C-RAN front-haul network, the intermediate wireless signal needs to be transmitted between the BBU and RRH. The transmission has a strict delay constrain. The legacy Ethernet network technologies are not suitable for direct application in the front-haul network due to the lack of support for precise timing synchronization, low delay and latency and high throughput. Currently different active projects are formed under the umbrella of IEEE 802.1 TSN task group in order to tackle these difficulties for TSN applications. For example, IEEE 802.1as is available for the timing and synchronization. Based on the IEEE 802.1 Qbu standard, this paper presents the implementation of the packet preemption technology and evaluates the performances of the Ethernet based front-haul networks [5].

III. IEEE TIME SENSITIVE NETWORKING (TSN)

IEEE 802.1 TSN, formerly named the IEEE Audio Video Bridging (AVB), aims to define the specifications that allow time-synchronized low-latency streaming services [6]. Low delay and jitter requirements have been stringent phenomena for real-time applications. The standards target the requirements for the industrial applications, vehicle control services, control or streaming data in the local area networks, and so on.

The TSN traffic is classified into 4 classes, as listed below in Table I [7]. The class Control Data Traffic (CDT) has the highest priority and is intended to carry the control messages. The class A and class B are used to transport audio and video streams, respectively. Class BE handles the best effort traffic, such as the legacy Ethernet traffic, with no restriction on QoS.The traffic specifications consist of two main categories: the maximum frame size and the minimum frame interval. The maximum frame size indicates the packet size of source data. The minimum frame interval indicates the frequency of receiving data. Based on the application, each class is specified with delay and jitter constraints.

Regarding the fronthaul in a C-RAN network, strict requirements are defined on the links between the RRH and the BBU. These requirements such as clock synchronization and latency have to be satisfied. In this paper, simulations of TSN functions have been performed to combine TSN features in a Fronthaul network. The class CDT traffic is evaluated with reduced latency.

IV. PACKET PREEMPTION

In this section, we briefly describe the salient features of the packet preemption technology standardized in the IEEE 802.1 Qbu and IEEE 802.3br documents [8]. The technology uses the preemptive-resume queueing discipline. The aim is to ensure a deterministic behavior with low delay for time critical packet frames.

In the packet preemption standard, the types of traffic on an ingress port are classified into two groups: express traffic and preemptive traffic. The express traffic is used for the transmission of the class CDT data while the preemptive traffic is sent with other classes of traffic.

The Head-of-Line (HOL) problem is well known from the traditional FIFO queue discipline. This is solved by prioritizing packets in different queues. An express packet is transmitted before the queued preemptive packets. However, an express packet can still experience excessive delay, since the preemptive packet that started ahead can be a large size packet. The motivation of the packet preemption technology is to eliminate the waiting delay caused by ongoing transmission of a preemptive packet.

Figure 2 presents the different operations between the usual priority queuing and the preemptive queueing when a new packet arrives. The transmission of a preemptive packet can be suspended in order to allow one or multiple express packets to be transmitted. Afterwards, the remainder preemptive packet resumes transmission. It is notable that one preemptive packet can be preempted and resumed for several times. This provides the capabilities of a network switch to support a deterministic time control application.

Table I: TSN TRAFFIC TYPES AND REQUIREMENTS

Traffic types	Maximum Frame Size (bytes)	Minimum Frame Interval (us)
Class CDT	128	500
Class A	256	125
Class B	256	250
Class BE	256	None

 ISBN: 978-1-61208-591-3

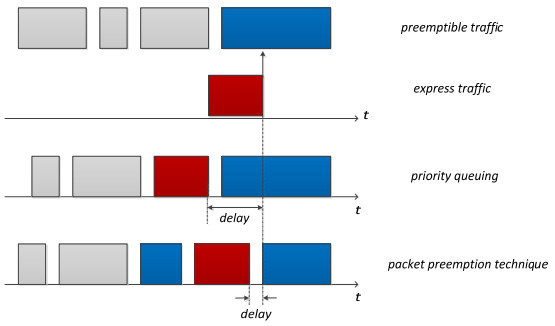

Figure 2. Priority queueing and the packet preemption queueing

To provide service differentiation, separated queues are applied to classify traffic into groups, and the packet preemption technique is used for priority scheduling at the switch node. In the traffic filtering and classifier module, the incoming data are classified into the express queue and the preemptive queue. In the transmission processing module, the system monitors the appearance of the express traffic. The preemption happens when the corresponding express traffic arrives. The preemption procedure occurs with some conditions stated in the standard. For example, the packet size is at least 64 bytes that remain to be transmitted. With this scheme, an express packet can take over the low priority traffic, even during the transmission. A format of *mPacket* is defined in the standard containing a complete packet (e.g., an express packet) or a continuation fragment of a preemptive packet. The detailed procedure of the preemption is shown in Figure 3.

To implement the 802.1Qbr, both transmitter and receiver switch should enable the packet preemption support as a TSN-enabled switch. In the transmitter side, the Ethernet frames are differentiated and classified into the express queue and the preemptive queue, respectively. When an express packet is received in the system while a preemptive packet is being processed, the express packet is processed immediately upon arrival assuming the packet preemption condition is fulfilled. The newly generated packet is formatted as *mPacket*, which carries the express packet, the complete or fragmented preemptive packet. A preemptive packet is interrupted and fragmented as a series of the continued fragments. An *mPacket* containing a continuation fragment of a preemptive packet has a fragment counter. The receiver identifies the packets and reassembles an incomplete preempted packet.

V. STATISTICAL MODEL

Our goal is to analyze the performance of a front-haul network with TSN enabled switches, taking into consideration queuing and packet preemption in each node. In this section, we first introduce the statistical model to study the queuing delay based on a preemptive resume queuing model. In this part, we simplify the traffic model as Poisson arrivals and fixed size packets.

A single server system with limited queue size is considered, which has job classes of multiple priorities. The priority queue can have either non-preemptive or preemptive strategies. In a non-preemptive system, a job in service is not interrupted, even if a job of higher priority arrives and enters the queue. In a preemptive case, the service of an ongoing job will be interrupted by the new

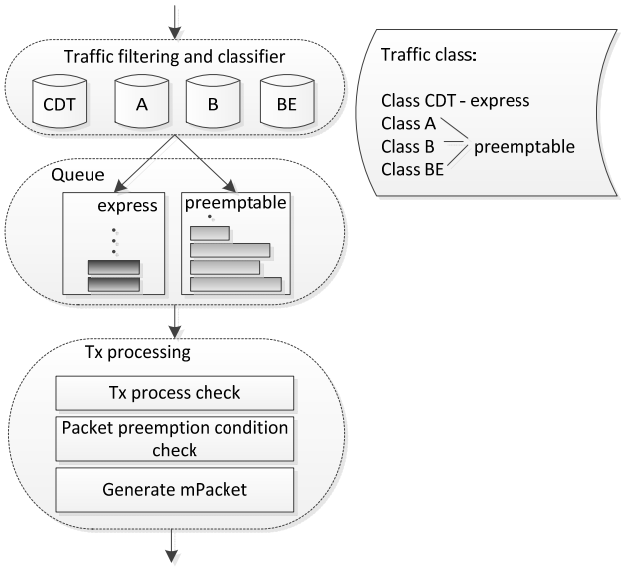

Figure 3. Diagram of packet preemption procedure

arrival of higher priority. The preemptive resume model means that the interrupted job from lower priority continues, when the higher priority job finishes.

To simplify the analysis, the M/M/1/K model is considered, with one single server and a limited number of waiting positions. The data arrivals are Poisson distributed with rate λ. Two classes of traffic arrival to the system are: express traffic with arrival intensity of λ_e, and the pre-emptible traffic with arrival intensity of λ_p. The express traffic has a higher priority.

For each class, the service time is exponentially distributed with a mean value of S. We denote the offered traffic of each type as $A_e = \lambda_e * S_e$ and $A_p = \lambda_p * S_p$, where the jobs in the express and pre-emptible class are assumed with a mean service time S_e and S_p, respectively.

A. Non-preemptive queuing model

The numerical analysis for the queuing delay for each traffic class, express and preemptive traffic has been discussed in details in [9]. For the express traffic, the highest priority, waits until the service in progress is completed and waits for the existing jobs in the same queue. The mean waiting time W_e is calculated as:

$$W_e = V_e + A_e \cdot W_e \quad (1)$$

Where V_e is the mean residual service time of the current job under service, both express and preemptive traffic classes are considered.

For the low priority class (referring to the preemptive traffic), the mean waiting time, W_p, considers not only the remainder process and the already arrived jobs from the same and higher priority, but also the new arriving jobs with higher priority during the waiting time.

$$W_p = V_{e,p} + \cdot A_p \cdot W_p + A_e \cdot W_p \quad (2)$$

B. Preemptive-resume queuing model

With preemptive resume property, a job with low priority is interrupted by the arrival of a higher priority job. The transmission will be continued from the point that it is interrupted later.

As the highest priority, the express traffic experiences only the expected remaining service time due to the existing jobs in the same queue, since with preemptive property the express traffic is not disturbed by lower priorities. Therefore, the mean waiting time W_e is same as (1). (but (1) includes low priority traffic under service- this traffic is preempted, in this case).

For the preemptive traffic, the mean waiting time considers the existing express traffic, which is already in the queueing system. Moreover, the extra waiting period caused by the interruption from the express traffic during the service time and the waiting time should be taken into account.

$$W_p = \frac{V_{e,p}}{1 - A_e} + \{W_p + s_p\} \cdot A_e \tag{3}$$

C. Probability model

By analyzing the statistical queueing model, we can derive the mean waiting time for the express and the preemptive traffic. By using the state transition diagram of the Markov chain and presenting the state balance equations, we can derive the delay probability of the system. We model the number of queuing places used by each class in a switch as a continuous-time Markov chain.

The problem is illustrated with a simplified M/M/1/2 queue, where only one queueing place is allowed. The state (*i, j*) describes the number of express traffic, *i*, and the number of preemptive traffic, *j*, in the system. In the non-preemptive model, as shown Figure 4, the job with a low priority is under processing, while the high priority traffic is waiting, as shown in the (1, 1) state. 0 presents the state transition diagram for the preemptive priority queue model. Different to the non-preemptive case, when a new express traffic arrives, the service of the lower class traffic is stopped and the process of the express traffic starts, from state (0, 1) to (1, 1).

Recall the Markov property which states that the future process is only influenced by the current state of the process. Note that the probability of being in state (*i, j*) is *P(x,y)*. From the state transition diagram in a non-preemptive model in Figure 4, we obtain the following expression, Eq(4):

$$P(x,y) = \frac{\frac{A_e^x}{x!} \cdot \frac{A_p^y}{y!}}{\sum_{i=0}^{2} \sum_{j=0}^{2-i} \frac{A_e^i}{i!} \cdot \frac{A_p^j}{j!}} \tag{4}$$

Where $A_p^x / x!$ and $A_p^y / y!$ represent the state probability of one dimensional truncated Poisson

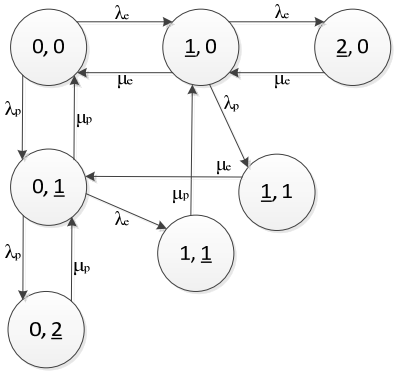

Figure 4. State transition diagram for M/M/1/2 non-preemptive priority queue

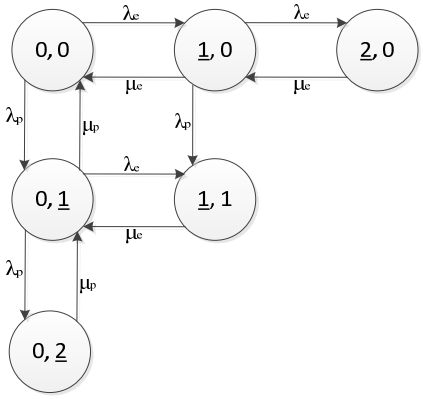

Figure 5. State transition diagram for M/M/1/2 preemptive priority queue

distribution for the preemptive and express traffic, respectively.

In non-preemptive model, the delay probability of express packet is supposed to contain *P(0,1)* where one preemptive packet is under service, thus the delay probability is estimated as:

$$\mathrm{D} = \sum_{i=0}^{1} \sum_{j=0}^{1-i} \mathrm{P}(i,j) \tag{5}$$

By considering the Markov chain in 0, the processing of the preemptive traffic is interrupted when the express traffic arrives in the state (1, 1). From (4), we can estimate the threshold of queuing delay. The preemption will be performed and the delay for each class is increased by controlling the arrival rate. The result can be extended to a switch with a large queue size.

VI. PERFORMANCE SIMULATION STUDY AND RESULTS

To evaluate the packet preemption technology and its behaviors, a TSN enabled Ethernet switch is examined by simulations. The simulation scenarios are setup in Riverbed Modeler[10]. Independent simulation was performed with various random seed numbers. Both the traditional priority queuing model and the preemptive resume queuing model are implemented. Consequently, the simulation will provide information about the limiting factors and the performance metrics of delay, packet loss and throughput.

We consider two types of traffic in the communication system, express traffic and preemptive traffic. We measure

the performance based on various traffic intensities, packet sizes, and ratio between different types of traffic.

A. *Queuing delay vs. traffic intensity*

The relative load between the express traffic and the preemptive traffic is varied from 0.1 to 1. Figure 6 shows the delay of two classes under different ratios. The increasing percentage of the express traffic introduces delay increment on the preemptive traffic. The delay for the express traffic keeps a low value in both the packet preemption and non-packet preemption cases. It is observed that the packet preemption reduces the queueing delay for the express traffic.

B. *Queuing delay vs. packet length*

We evaluate the influence of the packet size to the queueing delay. In this scenario, the input express traffic takes up 50 percent of the preemptive traffic. The packet size of the express traffic is fixed as a uniform distribution of 128 bytes. The packet length of the preemptive traffic is varied from 128 bytes to 1024 bytes. With the packet preemption technique, the queueing delay of the express traffic is increased, when the preemptive packets are mostly in small size. This is due to the rule of packet preemption, which examines the remaining size to be at least 64 bytes. When the preemptive packets are small, the chance of packet preemption is reduced. The express traffic has to wait for the ongoing transmission as in the non-packet preemption case. As shown in Figure 7, the smaller the packet size of the preemptible traffic, the fewer chances to segment the preemptible packets. Therefore, the express traffic has to be queued and waited for the preemptible traffic.

VII. CONCLUSION

In this paper, we analyzed the packet preemption scheme for reducing the waiting time in the queue and supporting service differentiation in the cloud radio access networks CRAN. The packet preemption technique favors the time-sensitive data by interrupting the interfered traffic and reducing the waiting time. The numerical analysis (not really shown) and the simulation results showed that the delay for the time-sensitive data is reduced dramatically. The influence of the traffic volume and packet length regarding the delay was analyzed. Packet preemption thus proved as an effective method to support time sensitive traffic over C-RAN fronthaul networks.

ACKNOWLEDGMENT

This work receives support from the “intelligent 5G mobile Ethernet Radio Access Network (ERAN)” project founded by Innovation Fond Denmark.

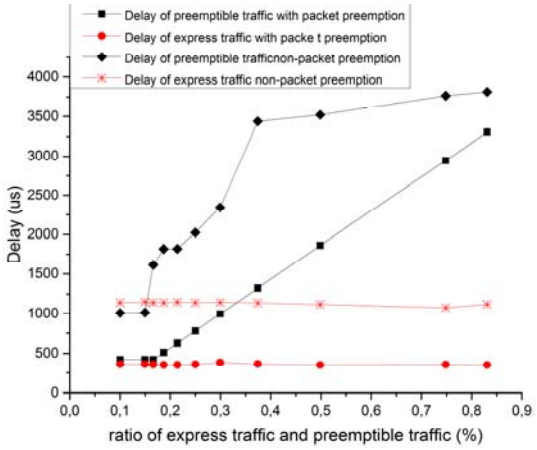

Figure 6. Queueing delay under different traffic loads

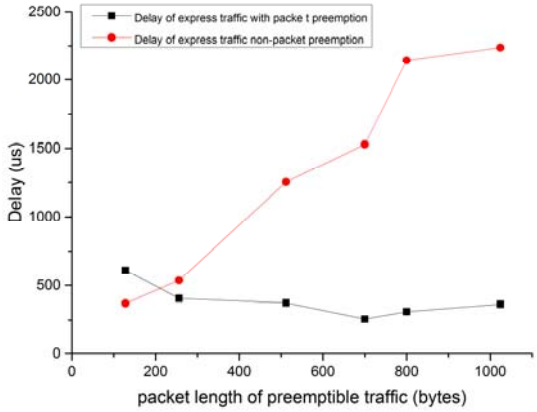

Figure 7. Queueing delay under different packet length

REFERENCES

[1] Time Sensitive Networking Task Group, http://www.ieee802.org/1/pages/tsn.html [retrieved: August 2017].

[2] C-RAN The Road Towards Green RAN. Tech. rep. China Mobile Research Institute, October 2011.

[3] Y. Lin, L. Shao, Z. Zhu, Q. Wang and R. K. Sabhikhi, "Wireless network cloud: Architecture and system requirements," in IBM Journal of Research and Development, vol. 54, no. 1, pp. 4:1-4:12, January-February 2010. doi: 10.1147/JRD.2009.2037680.

[4] A. Checko et al., "Cloud RAN for Mobile Networks—A Technology Overview," in IEEE Communications Surveys & Tutorials, vol. 17, no. 1, pp. 405-426, Firstquarter 2015. doi: 10.1109/COMST.2014.2355255.

[5] IEEE P802.1CM Draft standard for local and metropolitan area networks – Time sensitive networks for Fronthaul, October 2016.

[6] G. A. Ditzel and P. Didier, “Time Sensitive Network (TSN) Protocols and use in Ethernet/IP systems”, ODVA Industry conference & 17th Annual meeting, October 2015.

[7] S. Thangamuthu, N. Concer, P. J. L. Cuijpers and J. J. Lukkien, "Analysis of Ethernet-switch traffic shapers for in-vehicle networking applications," 2015 Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, 2015, pp. 55-60. doi: 10.7873/DATE.2015.0045.

[8] IEEE P802.3br Draft Standard for Ethernet Amendment: Specification and Management Parameters for Interspersing Express Traffic, January 2016.

[9] V. B. Iversen, “Teletraffic engineering and network planning”, Publisher: DTU Fotonik, 2015.

[10] OPNET Technologies – Network Simulator, Riverbed. www.riverbed.com.

 ISBN: 978-1-61208-591-3