March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611

Reducing the Latency in 5G Networks using SDN based Queuing Systems

***Baburao Kodavati[1], Madhu Ramarakula[2]**

[1]JNTUK, Kakinada Research scholar, Department of Electronics & Communication Engineering Usha Rama College of Engineering & Technology,NH-16, Telaprolu, Ungutur Mandalam, Near Gannavaram, Krishna District, AP- 521109.
[2]Department of Electronics & Communication Engineering, University College of Engineering Kakinada Jawaharlal Nehru Technological University Kakinada, Kakinada-533003, Andhra Pradesh, India.

[1*]baburaokodavati@gmail.com, [2]madhu_ramarkula@jntucek.ac.in

Article Info
Volume 83
Page Number: 1601 - 1611
Publication Issue:
March - April 2020

Article History
Article Received*: 24 July 2019*
Revised*: 12 September 2019*
Accepted*: 15 February 2020*
Publication*: 15 March 2020*

Abstract

The fifth generation (5G) wireless network technology is to be standardized by 2020, with the aim of improving its reliability, capacity and energy efficiency, on the other hand, the system is to be designed to reduce the latency for massive connection density. There are several technique available to improve the reliability, capacity and energy efficiency. However, the availability of techniques to reduce the latency using 5G core systems with Software Defined Networking (SDN) is not available in researches. In this paper, we propose a new network topology that utilizes SDN to support the core network in reducing the latency requirement in 5G systems. This intelligent SDN framework is used in 5G subsystems for transmitting and receiving purposes. The utilization of SDN framework ensures that better routing path for data transmission with shortest paths. This will pave the way for reduced latency in 5G systems with the use of an improved routing model with SDN framework. The average latency in the proposed method is reduced henceforth. The simulation reveals that the proposed method has reduced its average latency rate due to its effective processing of control and data associated with SDN.

Keywords**:** *SDN, 5G systems, Latency, Queuing Model*

1. Introduction

In some applications such as automated industrial production, control or robotics, transport, healthcare, entertainment, virtual reality, education and culture, latency is highly critical. IoT becomes, in particular, a reality that always and everywhere connects anything with anything else. Intelligent wearable devices and clever home devices are connected to hyper-connected countries in order to improve our way of life [1]–[3]. Although operators support those IoT applications through current 3G / LTE applications, the underlying network requirements, such as low latency, high reliability [4] and [5], and security [6], [7], some applications require significantly higher standards. In some cases, a latency of up to 1ms with a rate of packet loss of up to 10^{-2} is necessary.

Latent can be separated into two main parts within the LTE system: (1) the latency of the user plane (U-plane) and (2) the latency of the C-plane. The latency of the U-plane is measured by a directed time of transmission of a packet to be available on the IP layer between the developed terrestrial UMTS (E-UTRAN) radio access network [8]. C-plane latency, on the other hand, can be defined as the transition period of an EU from idle to active. An UE is not connected to the Radio Resource Control (RRC) in the idle state. After the RRC

connection is installed, the UE switches from idle to connected status and then goes into active mode. As the application performance is mainly dependent on the latency of the U-plane, the main focus for low latency communication is a U-plane. In the U-plan, RAN, backhaul, core networks and data center or internet can contribute to delaying packet transmission in a cellular network.

Major tradeoffs exist between wireless network capability, coverage, latency, reliability and spectral efficiency. These fundamental limits can result in a degradation of another metric by optimizing one metric for improvement. The radio frame is 10ms in the LTE system, with 1ms being the smallest TTI. The fixed frame structure is dependent on modulation and encoding schemes for adapting the rate of transmission with constant overhead control. Since latency is associated with a control overhead (cyclic prefix, transfer mode and pilot signs). A packet with radio transmission time below 1ms is considered not advisable since latency is associated with a large portion of the transmission time (approximately 0.3-0.4ms per packet transmission). In addition, retransmitting by packet transmission takes about 8ms and removing the transmission will have a significant impact on packet error. As a result, radical changes and improvements in packet/frame structure and transmission strategy are required. In this regard:

- The first thing is to design a new radio framework, strengthened by a limited overhead control and smaller transmission time. Procedures for scheduling the user, allocation of resources and channel formation can be removed or merged for reduction of overhead control.
- Second, with new waveforms and transmission techniques which reduce retransmission time, the probability of packet error for first transmission should be reduced.
- Thirdly, because critical latency data must be sent out immediately, data priority techniques must be identified over normal data.
- Fourthly, the essential aspects of OFDM which pose major obstacles for achieving low latency include synchronization and orthogonality. Although asynchronous communication modes are more advantageous over synchronized latency, additional power and spectrum sources are needed [10].
- Fifthly, since data transmission latency also depends on the delay between the core network and the BS, it is possible to use cache networks for reducing latency by stored the popular network data.

Tradio should not exceed 0,5ms for low latency communications in accordance with ITU[11]. Radio transmission time should be set to hundreds of microseconds in this regard, whereas the current 4G configuration is 1ms. In this respect. To do this new design of the waveforms, techniques for transmission and symbol detection should be implemented in various RAN areas, such as the packet/frame structure, modulation and coding schedules. Approaches like advanced backhaul techniques, caching / fog-enabled networks and intelligent AS / NAS integration can provide potential solutions in order to reduce the delayed operation of TBackhaul. The new core network for TCore comprises SDN, NFV and several smart approaches that can significantly reduce delays. The internet caching or cloud caching enabled fog/MEC for TT transport offers lower latency.

In this paper, a routing topology is designed for 5G subsystems using intelligent SDN framework that effectively reduces the delay in transmission or latency in communication between the source and destination nodes. The proposed routing protocol with intelligent SDN framework is tested under different traffic conditions to estimate the delay, network throughput and its resilience towards communication path failure. Finally, the system estimates how effectively the traffic is rerouted from failed or congested route to optimal route.

The main contribution of the work includes the following:

- The authors propose a new network topology that utilizes SDN to support the core network in reducing the latency requirement in 5G systems.
- The proposed intelligent SDN framework is designed for 5G subsystems, which is utilized for both transmission and reception purposes. The SDN routing framework ensures better routing path for data transmission with shortest paths.
- The proposed design reduces the average latency in SDN and by rerouting the traffic pattern effectively from the congested route to available route.

The outline of the paper is as follows: section 2 provides the related works. Section 3 provides the details of novel network topology in SDN that supports the 5G core network. Section 4 discusses the evaluation of proposed work with other existing methods. Section 5 concludes the work with possible direction of future work.

2. Related works

In [12], the authors combine two parallel algorithms, the traditional sliding window algorithm and the Cross Parallel View (CPW) Algorithm, is proposed to provide high-performance latency-sensitive turbo decoding architecture. New IFFT design with butterfly operation, reducing the delay in IFFT output data by reducing memory size and butterfly operation. The IFFT processor's input signal corresponding to the guarding band is given as zero (i.e.' 0'). The memory depth may be reduced from 1024 to 176 if the sequence of OFDM symbol data entered in the IFFT is adjusted.

In [13] proposes a latency reduction approach by establishing a higher priority ultra-low-latency multiplexing (TDM) over other services with less time-critical priority, whereby higher priority data is mapped

during the initiation of a subframe and the normal data is followed.

In [14], balanced truncation in linear systems, coupled with arbitrary graphs in latency of communication constraints, is applied to model reduction.

Recent progress in the information theory of finite block lengths is used in [19] to demonstrate the optimal development of wireless systems with strict restrictions, such as low latency and high reliability. The limits for the number of bits that can be transmitted for an OFDM system are derived for a given set of constraints such as bandwidth, latency and reliability.

Instead of well-known symmetric windows, an asymmetrical window is proposed [20] to reduce the cyclic prefix by 30% and thus to lower the overhead latency. The system removes the out-of-bound emission from the OOB system but makes it more sensitive to inter-symbol (ISI) and inter-carrier (ICI) channel induced interference. The optimisation of transmission power with the steepest descent algorithm takes account of time for transmission, the likelihood of error and the delay of queuing.

In [21] a fast path change and packet recovery method for a multi-radio access (Multi-RAT) environment are introduced in Low Latency Packet Transportation System.

In [22] a solution is suggested for enhancing capacity and reducing latency, using diversity gains. Various approaches, including spatial diversity, time diversity and frequency diversity could be used to achieve diversity.

The mmWave switched Architecture System [23] proposes a control signal with the possibility to multiplex small control packets with analogue beam formed data (for higher-order modulation) by using the low-resolution digital beam forming (to allow the multiplex of small control packets). This significantly reduces the overhead due to control signalling, which leads to more data transmission resources. This technique reduces the physical latency of the round trip.

The research study [24] suggested a new mechanism to introduce the adaptive radio connection control mode (RLC) which, based on real-time analysis of radio conditions, alternates dynamically between non-recognition mode (RLC) and acceptance mode. This technology reduces latency and processing power and enhances the performance of UM. On the other hand, AM is activated in a deteriorated radio condition, which improves data reliability.

The SDN Control Plane Strategy for the Optimization of Vehicle Ad hoc Network (VANET) and Radio Network costs is presented under [25]. In a two-stage Stackelberg game, the interaction between vehicles and controllers is formulated and evaluated and an optimal rebating strategy provides a reduced latency compared with other structures on the control aircraft.

In [26], SDN-based management of X2 transmission local mobility is suggested where the total transfer signalling is reduced to a minimum through a reduction in internode signalling and X2 transmission signal to a centralized SDN system. This can reduce the latency of transmission while reducing the overhead signal.

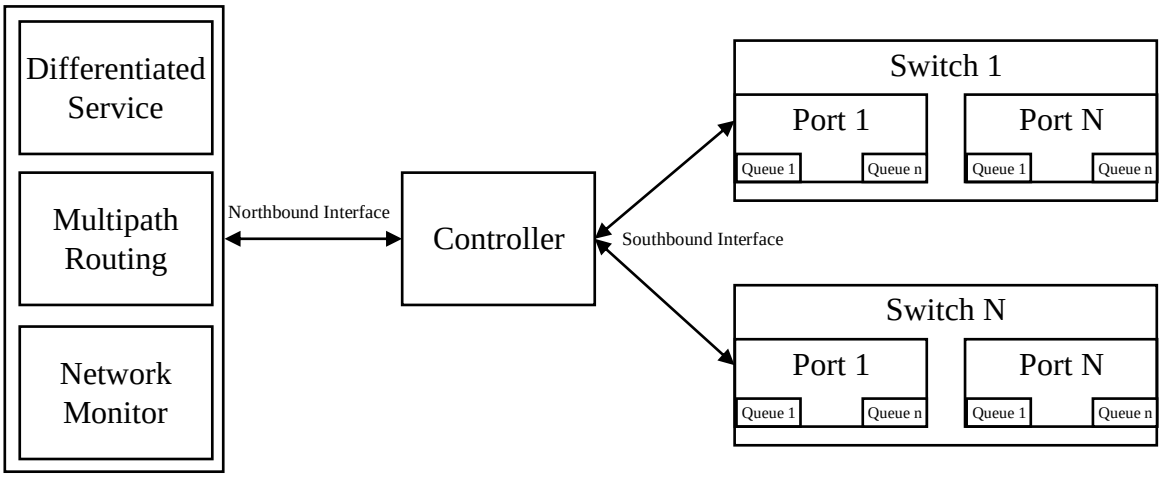
3. Proposed Method

The study uses a G/G/1 Queuing Model [15] to estimate percentiles of high delay and to calculate optimal several paths for free transmission of delay. The percentiles for delays are saved and periodically checked for less latency on available paths. In order to save the optimal available path and delay percentile, Apache Cassandra [17] is used as controller to check the status of network paths on a periodic basis. The optimum paths will be selected from multiple available paths by a controller depending on the current status, i.e. use of bandwidth and queuing percentile delay.

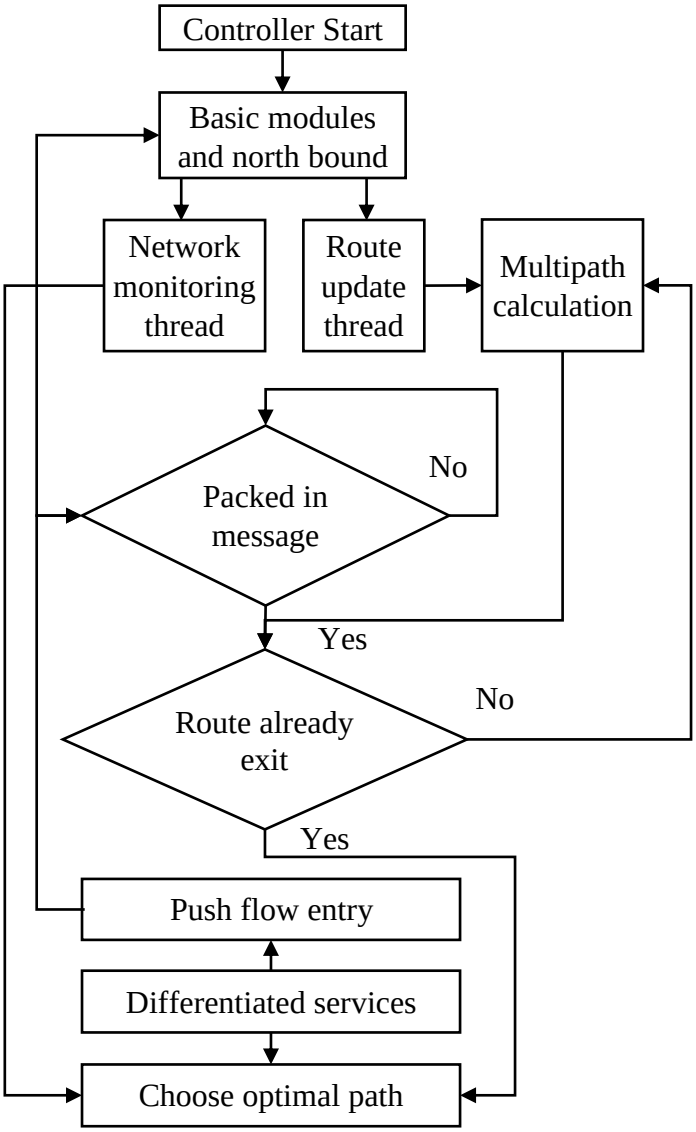
Deadline-Aware Multipath Communication Protocol (DAMCP) [16] is a linear communication protocol based on optimization that formulates multi-way communication and improves the diversity of pathways that increase network communication performance. This protocol captures the optimum multi-path transmission strategy and improves performance on theoretical upper limits under certain circumstances. It determines the quantity of traffic generated and the arrival of the maximum data on each path is determined at the destination. In view of delaying or latency in queuing, we consider jointly the paths appropriate to the transmission and transmission and this is known as path combination.

Therefore, a packet is sent to a controller if it arrives at a DAMCP interrupt without a corresponding flow entry. For packet transmission the flow entries are then moved to DAMCP switches. The proposed system with the G/G/1 Queuing Model and DAMCP ensures optimal pathways for the forwarding of packets using its available bandwidth based on multipath routing components. Figure 1 shows the architecture of the system proposed.

March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611



(a) Architecture of 5G Networks using SDN



(b)

Figure 1: Flowchart of Proposed SDN Routing Framework

March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611

a. G/G/1 Queuing Model

The analysis of a considerable number of multipath flows requiring the use of suitable queuing models. The reason for this is because while this functional split lowers bandwidth requirement, the HARQ protocol responsible for the error correction process still imposes a strict latency requirement. Therefore, it is crucial to deeply characterize the queuing delay to make sure the latency needs are met [15].

The well-known M/M/1 is appealing, since the main metrics of interest are closed expressions. It assumes, however, that the intercom times are exponentially distributed that do not apply to traffic in multipaths. The interval distribution time is unknown and dependent on the specific number of flows, so a general queuing model is used (G/G/1) and enables the behavior of the system to be characterized by adjusting the coefficient of arrival times of packets to the switch queue under differing conditions.

Contrary to the M/M/1 model, there are no closed terms under these assumptions for the average time of waiting at the queue. A G/G/1 model-model packet switch assumes packet arrivals to follow a general (*G*) (arbitrary) packet α/s distribution. All arrivals compete for one resource, and are stored in a first-come discipline (buffered or queued) temporarily, with a period of $E[S]=\frac{1}{\mu}$ seconds, the distribution of which is also general [15].

The study requires the system load as $\rho=\lambda\cdot E[S]<1$, for the purpose of stability and the squared coefficient of variation is define in terms of a random variable *X* as:

$$C^2[X]=\frac{Var[X]}{E[X]^2}. \quad \textbf{(1)}$$

Consider *T* as a random variable model for inter arrival time of packets at the queue and *S*is regarded as the service time random variable. Thus the queuing delay W_q from Allen–Cunneen approximation is defined as:

$$E[W_q]\le E[S]\cdot\frac{\rho}{1-\rho}\cdot\frac{C^2[T]+C^2[S]}{2} \quad \textbf{(2)}$$

This extends M/M/1 queuing model to a stochastic variability term, which is defined as:

$$\frac{C^2[T]+C^2[S]}{2}. \quad \textbf{(3)}$$

The formula of the Kingman is a very well-approximate mean queue time, which works very well in most circumstances, especially when $\rho\to 1$. It is seen in [18] that the service and interarrival times are exponentially distributed, which is given by the following expression:

$C^2[T]= 1 =C^2[S]$.

Therefore in the case of M / M/1 the stochastic variability term is 1 and the formula of Kingman is accurate. The congestion index can formally express Kingman's exponential law of congestion as follows:

$$\frac{W_q}{E[S]}\begin{cases}\exp\left(mean=\frac{1}{1-\rho}\cdot\frac{C^2[T]+C^2[S]}{2}\right) & \text{with probability } \rho\\ 0 & \text{with probability } (1-\rho)\end{cases} \quad \textbf{(4)}$$

From here, the p^{th} percentile delay can be calculated as

$$p=\int_{t=-\infty}^{W_q^{(p)}}(1-\rho)\cdot\delta(t)+\rho\cdot\omega e^{\omega t}H(t)\,dt \quad \textbf{(5)}$$

where $\frac{1}{\omega}=\frac{1}{1-\rho}\cdot\frac{C^2[T]+C^2[S]}{2}$ and

$\delta(t)$ and $H(t)$ are the delta and Heaviside step functions, respectively. These are the considered as the supportive indicator functions that solves p^{th} percentile delay $W_q^{(p)}$, and hence we obtain,

$$W_q^{(p)}=\max\left\{0,E[S]\frac{1}{1-\rho}\cdot\frac{C^2[T]+C^2[S]}{2}\ln\left(\frac{1}{1-\rho}\right)\right. \quad \textbf{(6)}$$

b. Cassandra Data Model

The concept of a table is different from that of a table in a relational database in Cassandra. A CQL table can be viewed as the set of partitions that contain rows with a similar structure (hereinafter referred to as a table). A single partition key can be used for each partition table, and optionally a single classification key for each row in a partition. Both keys may be simple or composite (a column) or multiple columns. The combination of a partition key with a cluster key identifies a row in a table and is referred to as the main key. While the key part of a primary key always is required, it is optional to cluster the key component. There are only partitions in a single row because a table with no clustering key is equal to the parts key of a table and the mapping between the partitions and rows is one-to-one. A table with a cluster key may have multi-row partitions, as different lines have different cluster keys in the same partition. Multi-row ranks in an up (default) or down the order are always ordered by clustering key values.

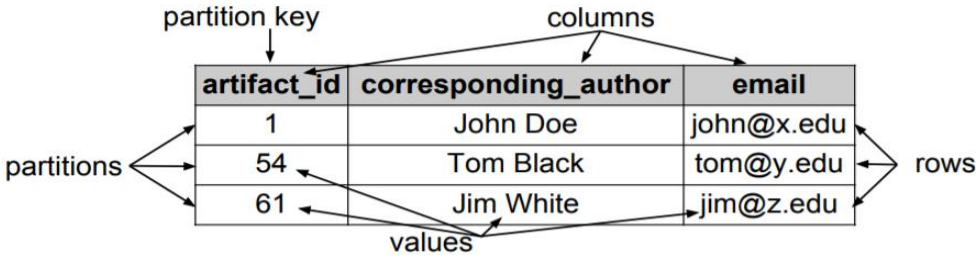
A schema for the table defines a number of columns and the main key. A data type, such as int, text, or complex data types (settings, lists, or a map), is assigned to each of the columns. A special counter data type can be assigned to a column which can be used to hold a

March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611

distributed counter to be added to or removed from the simultaneous transactions. All non-counter columns of a table must form part of the primary column in the presence of a counter column. A column can be defined as "static." The only way to identify a column whose value is shared by all rows of a partition is through the table with multi-row partitions. Finally, an essential key is a column sequence consisting of columns with partition-key and optional clustering key columns. CQL defines additional parentheses to partition key columns that can be omitted in the case that a partition key is straightforward. No counter, static or collection columns may be included in the main key.

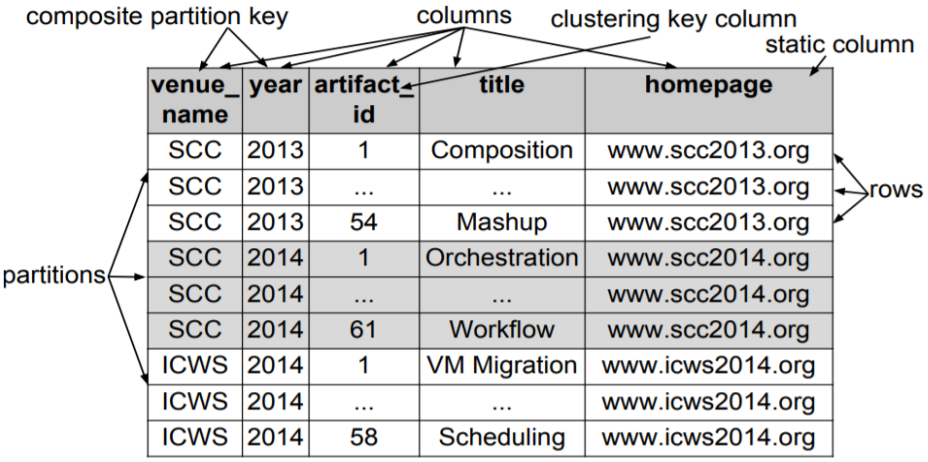
To illustrate certain of these concepts, Figure 2 displays two CQL sample tables and sample rows. In Figure 2(a), the table of Artifacts contains partitions with a single-row. Its main key is one artifact column Id, which is also a simple partition key. Three single-row partitions are shown in this table.

```
CREATE TABLE artifacts(
      artifact_id INT,
      corresponding_author TEXT,
      email TEXT,
      PRIMARY KEY (artifact_id));
```



(a) Table Artifacts with single-row partitions

```
CREATE TABLE artifacts_by_venue(
      venue_name TEXT,
      year INT,
      artifact_id INT,
      title TEXT,
      homepage TEXT STATIC,
      PRIMARY KEY ((venue_name,year),artifact_id));
```



(b) Table Artifacts with multi-row partitions

Figure 2: Sample tables in Cassandra

March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611

In Figure 2(b), multi-row partitions of artifacts by venue table. Its main key is composed of the composite partition key (venue name, year) and the clustering key ID. This table shows that there are three partitions, each with several rows. Its rows are ordered by artifact Id in ascending order for any particular partition. Furthermore, a homepage is defined as a static column so that each score can only be divided into one homepage with each row.

c. DAMCP

The multiway problem we are studying in this paper is a simple instance in Figure 3. The source generates a constant flow of data that must be delivered at the latest after one second and has two paths with contradictories. Since the one-way time limit of the high-bandwidth route is 600ms, it takes 800ms in total for a recognition to return along the low-latency route that leaves sufficient time to (potentially) retransmit the data along the low-latency route. Clearly, if all the data produced is first transmitted along the high bandwidth path and transmitted via the low-bandwidth path, 100% of the packets can be expected to reach their destination in time. Only one of the two routes would not make this possible.

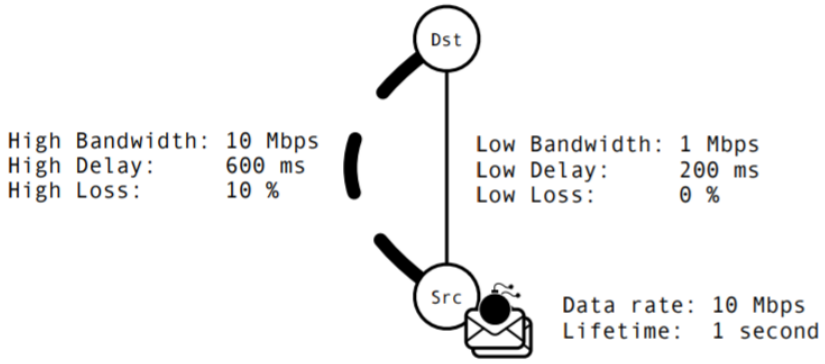


Figure 3: Deadline-based multipath communication scenario

The problem is trivial, i.e. intuitively an optimal solution can be found. However, if more paths are considered or the metrics are not so obvious, the problem becomes difficult.

Now we propose a model that aims to capture the ideal sending multipath strategy for the above-mentioned scenario. In particular, this model can be used to provide theoretical upper boundaries on the performance of the ideal protocol, but also to design a true protocol.

The study problem is to determine how much traffic the application generates should be transmitted or retransmitted over each path, so that a maximum of data reaches its destination in time. Since we take into consideration latency, we have to jointly consider pathways for initial transmission and transmission. We call this pair of pathways a combination of paths.

There is only one retransmission to avoid a cumbersome notation, but this model can clearly be adapted to arbitrary retransmission, even if, with the number of retransmission considered, the problem is naturally more complex. In most real cases, for up to 2-3 transmissions, we anticipate that the problem would be solved for two reasons. Firstly, it is a very rare event to send the same data four times or more unless the loss rate is particularly high for all paths. Second, many transmissions will probably take longer than their lifetime.

The DAMCP defines several metrics in order to measure the results of selecting some values of x for a certain network. This helps to define the conditions and goals of the linear programme. Table 1 summarizes the metrics notation we are using.

Table 1: Network Metrics

Metrics	Description
S_i	bit rate i, which is sent along the path
G	Goodput or received data rate
Q	communication quality (G/λ)
C	sum of all paths or total cost per second

First, by considering the data transmitted on the path for the first time (whichever the path along which the same information can be transferred), or the data transferred on it (which depends on the reliability of the initial path). This means that the amount of data sent to a certain path is obtained. Therefore, the expression is given as

$$S_i = \sum_{j=0}^{n-1} x_{i,j} \cdot \lambda + \sum_{j=0}^{n-1} x_{j,i} \cdot \lambda \cdot \tau_j \quad (7)$$

This is bounded by the bandwidth that is available on the required path:

March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611

$$S_i \le b_i \quad \forall i \in \{0,1,...,n-1\} \tag{8}$$

Since the research takes into account a fixed delay and an acknowledgement is always sent back in the shortest time, the transmitter sets a transmission timeout when data is sent along path *i*.

$$t_i = d_i + d_{\min} \tag{9}$$

We define goodput as the number of data that reach the destination each second before the deadline. Again, both data arriving at the first attempt and the transmitted data must be taken into account. The goodput are therefore defined as

$$G = \sum_{i:d_i \le \delta} \sum_{j=0}^{n-1} x_{i,j} \cdot (1-\tau_i)\lambda + \sum_{i,j:d_i+d_{\min}+d_j \le \delta} x_{i,j} \cdot \tau_i \cdot (1-\tau_i)\lambda \tag{10}$$

The goodput depends on the amount of data generated by the application (i.e., λ), but we want to determine the proportion of λ that a particular network can handle in the optimal manner. We therefore define our main metric which we call the quality of communication

$$Q = \frac{G}{\lambda}, \text{ where } 0 \le Q \le 1 \tag{11}$$

The range [0,1] defines that the arrival of data at the destination occurs prior to the deadline.

d. SDN Routing Protocol for Multipath Estimation

For the multipath calculation, we characterize the routing topology as a weighted graph *G<V, E>*. *V* is the set of nodes in the network. *E* is the set of edges that connect any two nodes. Each edge *e* has a link cost value L_e which is defined in Eq.(1).

$$L_e = \alpha \cdot cost(e) + \beta \cdot dist(e) - \gamma \cdot bw(e) \tag{12}$$

where

cost(*e*) is defined as the link cost and stability value and its robustness.

dist(*e*) is the distance between the link,

bw(*e*) is the bandwidth of the link.

α, β and γ represents the cost weights of *cost*(*e*), *dist*(*e*) and *bw*(*e*), respectively. The paths between the source and destination node is defined as a set $P=\{p_1,p_2,...,p_k,...\}$. Therefore the total path cost is defined:

$$L_{pk} = \sum_{e \in p} L_e \tag{13}$$

In this system, the threshold is set for path cost as L_0. Candidate paths whose total cost value meets the Eq.(13) are chosen.

$$L_{pk} \le L_0 \tag{14}$$

The following is the pseudo code of the multipath algorithm. The DAMCP with our proposed multi-path calculation algorithm that meets the QoS requirements.

Pseudo Code

Initialization
link cost costs<>, use previous<>
use Table<>values for storing the previous node

Estimate previous<>on each switch

If queue ≠ 0:
get queue's first node
if node = destination switch:
breakthe operation
end if
for each switch is linked to this node:
estimatetotal cost between switch and to the source switch
if switch is checked:
continuethe operation
end if
if total cost valuesatisfies limitation:
addthe present switch to the previous<>entry
where, costs<> = minimum{total cost, cost<>}
put switch node into queue
end if
end for
end while

Generation of the multiple paths

preparea table list for storing the path information
setdestination switch= node_current
function generation-of-path(src, dst, routes, previous, current, list):
ifnode_current= source switch:
path = new Route<>
add this path to multiple path cache
return
end if
for each link connectwith node_current:
addthe current link to table list
functiongeneration-of-path(src, dst, routes, previous, current, list)
remove two nodes from the same link
end for

e. SDN Routing Protocol for Route Optimisation

The method selected as the optimum path for a new flow is the path with the minimal bandwidth utilization of a queue. The controller regularly measures the use of bandwidth along each queue $bwu(p_k,q_i)$.

$$bwu(p_k,q_j) = T(q_i)/(\Delta t \cdot bw(q_i)) \tag{15}$$

where

q_i is the specific type *i* of traffic queue of path p_k,

$bw(q_i)$ is the queue bandwidth *i*,

$T(q_i)$ is the total number of packets transmitted in a queue *i*, and

March-April 2020
ISSN: 0193-4120 Page No. 1601 - 1611

Δt is the measurement of a time interval.

The greater the use of the queues, the more traffic congestion. We select a track that uses the queues to a minimum bandwidth compared to other paths for the incoming flow. This selects the optimum way in which data packets received can avoid congestion. This makes it possible for the system to ensure QoS which reduces the BS delay and response time and increases system output.

4. Results and Discussions

We conduct an experiment below to demonstrate the performance using throughput and response time performance. In this study a packet size of 1500B is used for transmitting the information between the Tx and Rx.

a. Response time and delay

In Figure 4, there exist two different two queues with 4 Mbps for total lines bandwidth. The queue 1 bandwidth is 1.5 Mbps and queue 2.5 Mbps. For real-time media services, we believe the priority of the BS_1 is higher. It requires some bandwidth guarantee through queue 1. The bandwidth guarantee for traffic packets of BS_2 is not required, thus sending packets via queue 2. Many customers access respectively BS_1 and BS_2. The overall traffic output for BS_1 is 1 Mbps of media streaming, while the total data output for BS_2 is 3.5Mbps. Ping is used to measure the response time for BS_1 and BS_2, which means that data transmission from *BS* to customers is delayed. This test can check that the proposed system is capable of guaranteeing bandwidth for various types of services. At each time interval say (2s – 10s), the study found that an average latency at the physical layer is of 0.25µs is found at the transmitter and 1.86 µs at the receiver, which is lesser than the other methods. Since the adaptive radio connection control mode has an average latency of 0.41µs is found at the transmitter and 2.03 µs at the receiver. The ultra-low-latency TDM has an average latency of 0.56µs is found at the transmitter and 2.24 µs at the receiver. Finally, the balanced truncation method obtains an average latency of 0.68µs is found at the transmitter and 5.53 µs at the receiver.

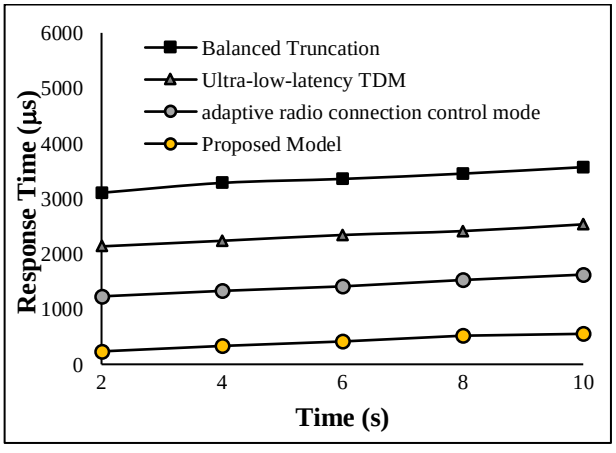


Figure 4: Response time and delay

b. Throughput

This experiment tests the system performance using various QoS solutions. The topology of experiments is identical to that of previous tests. The BS_1 and BS_2 are accessed by multiple clients. We are increasing traffic so that we can measure the success of each QoS solution to exceed the bandwidth in the network. In order to calculate the system average output we regularly measure the output of Switch 1 every 5 seconds.

The current throughput is approx. 3.8 Mbps, near the path bandwidth of 4 Mbps, which is shown in Figure 5. The proposed routing protocol has a maximum output of around 11.4 Mbps, about three times higher than the other two solutions. The routing protocol that you are proposing can transmit data via several paths, indirectly increase the overall system bandwidth and also increase system performance.

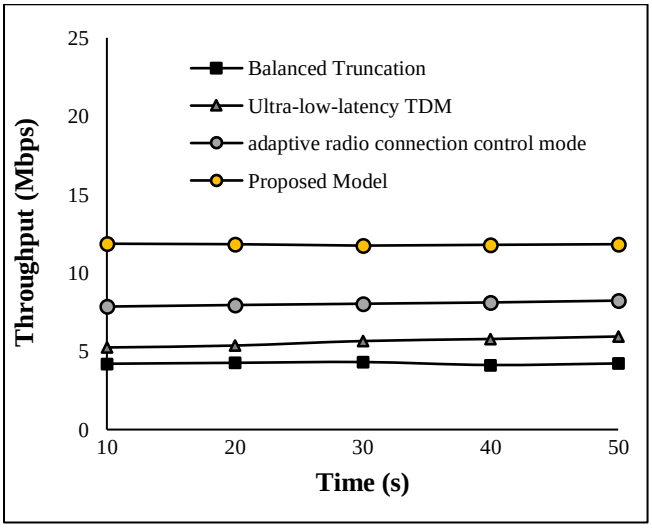


Figure 5: Throughput

5. Conclusions

In this paper, a novel SDN topology is designed in order to support the 5G core network for the reduction of latency. The SDN topology ensures improved routing experience for the transmission of data along the shortest route. The shortest path routing in 5G subsystems is utilized for both transmission and reception purposes. The shortest routing using SDN topology reduces effectively the latency in SDN. The simulation result reveals that the proposed method reduces average latency rate due to its effective processing of control and data associated with SDN. The result shows that the proposed system guarantees bandwidth by effectively utilizing available network capacity. The proposed study further reduces the transmission delay and increases the network throughput by increasing the BS response time. Finally, it is seen that the increase in performance reroutes effectively the traffic

patterns from congested path to optimal path using proposed routing strategy.

References

[1] Schulz, P., Matthe, M., Klessig, H., Simsek, M., Fettweis, G., Ansari, J., ...& Puschmann, A. (2017). Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture. *IEEE Communications Magazine*, *55*(2), 70-78.

[2] Palattella, M. R., Dohler, M., Grieco, A., Rizzo, G., Torsner, J., Engel, T., & Ladid, L. (2016). Internet of things in the 5G era: Enablers, architecture, and business models. *IEEE Journal on Selected Areas in Communications*, *34*(3), 510-527.

[3] Sarwat, A. I., Sundararajan, A., Parvez, I., Moghaddami, M., & Moghadasi, A. (2018). Toward a smart city of interdependent critical infrastructure networks. In *Sustainable Interdependent Networks* (pp. 21-45). Springer, Cham.

[4] Tavana, M., Rahmati, A., & Shah-Mansouri, V. (2018). Congestion control with adaptive access class barring for LTE M2M overload using Kalman filters. *Computer Networks*, *141*, 222-233.

[5] Parvez, I., Abdul, F., Mohammed, H., &Sarwat, A. I. (2015, April). Reliability assessment of access point of advanced metering infrastructure based on Bellcore standards (Telecordia). In *SoutheastCon 2015* (pp. 1-7). IEEE.

[6] Ferdowsi, A., &Saad, W. (2018, May). Deep learning-based dynamic watermarking for secure signal authentication in the Internet of Things. In *2018 IEEE International Conference on Communications (ICC)* (pp. 1-6). IEEE.

[7] Parvez, I., Abdul, F., & Sarwat, A. I. (2016, April). A location based key management system for advanced metering infrastructure of smart grid. In *2016 IEEE Green Technologies Conference (GreenTech)* (pp. 62-67). IEEE.

[8] Latency Analysis in LTE network. [Online]. Available: http://www.techmahindra.com/Documents/WhitePaper/ WhitePaperLatencyAnalysis.pdf

[9] Garcia-Perez, C. A., & Merino, P. (2016, September). Enabling low latency services on lte networks. In *2016 IEEE 1st International Workshops on Foundations and Applications of Self* Systems (FAS* W)* (pp. 248-255). IEEE.

[10] Wunder, G., Jung, P., Kasparick, M., Wild, T., Schaich, F., Chen, Y., ...& Mendes, L. L. (2014). 5GNOW: non-orthogonal, asynchronous waveforms for future mobile applications. *IEEE Communications Magazine*, *52*(2), 97-105.

[11] Agyapong, P. K., Iwamura, M., Staehle, D., Kiess, W., & Benjebbour, A. (2014). Design considerations for a 5G network architecture. *IEEE Communications Magazine*, *52*(11), 65-75.

[12] Liu, D., Zuo, C., & Wu, Z. (2015, November). Benefit and cost of cross sliding window scheduling for low latency 5G Turbo decoding. In *2015 IEEE/CIC International Conference on Communications in China (ICCC)* (pp. 1-4). IEEE.

[13] Ganesan, K., Soni, T., Nunna, S., & Ali, A. R. (2016, December). Poster: A TDM approach for latency reduction of ultra-reliable low-latency data in 5G. In *2016 IEEE Vehicular Networking Conference (VNC)* (pp. 1-2). IEEE.

[14] Jaoude, D. A., & Farhood, M. (2015, December). Balanced truncation of linear systems interconnected over arbitrary graphs with communication latency. In *2015 54th IEEE Conference on Decision and Control (CDC)* (pp. 5346-5351). IEEE.

[15] Pérez, G. O., Hernández, J. A., & Larrabeiti, D. (2018). Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G. *IEEE/OSA Journal of Optical Communications and Networking*, *10*(6), 573-581.

[16] Chuat, L., Perrig, A., & Hu, Y. C. (2017, June). Deadline-Aware Multipath Communication: An Optimization Problem. In *2017 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (pp. 487-498). IEEE.

[17] Chebotko, A., Kashlev, A., & Lu, S. (2015, June). A big data modeling methodology for Apache Cassandra. In *2015 IEEE International Congress on Big Data* (pp. 238-245). IEEE.

[18] Harrison, P. G., & Patel, N. M. (1992). *Performance modelling of communication networks and computer architectures (International Computer S*. Addison-Wesley Longman Publishing Co., Inc..

[19] Taheri, T., Nilsson, R., & van de Beek, J. (2016, December). Asymmetric transmit-windowing for low-latency and robust OFDM. In *2016 IEEE Globecom Workshops (GC Wkshps)* (pp. 1-6). IEEE.

[20] Hirai, H., Tojo, T., & Takaya, M. M. N. (2016, June). Low Latency packet transport methods for remote-controlled devices in multi-RAT environments. In *2016 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)* (pp. 1-2). IEEE.

[21] Johansson, N. A., Wang, Y. P. E., Eriksson, E., & Hessler, M. (2015, June). Radio access for ultra-reliable and low-latency 5G communications. In *2015 IEEE International Conference on Communication Workshop (ICCW)* (pp. 1184-1189). IEEE.

[22] Dutta, S., Mezzavilla, M., Ford, R., Zhang, M., Rangan, S., & Zorzi, M. (2016, June). MAC layer frame design for millimeter wave cellular system. In *2016 European Conference on Networks and Communications (EuCNC)* (pp. 117-121). IEEE.

[23] Shreevastav, R., & Carbajo, R. S. (2016, October). Dynamic RLC mode based upon link adaptation to reduce latency and improve throughput in cellular networks. In *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 1-6). IEEE.

[24] Li, H., Dong, M., & Ota, K. (2016). Control plane optimization in software-defined vehicular ad hoc networks. *IEEE Transactions on Vehicular Technology*, *65*(10), 7895-7904.

[25] Assefa, T. D., Hoque, R., Tragos, E., & Dimitropoulos, X. (2017, June). SDN-based local mobility management with X2-interface in femtocell networks. In *2017 IEEE 22nd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)* (pp. 1-6). IEEE.

[26] Gimenez, L. C., Michaelsen, P. H., Pedersen, K. I., Kolding, T. E., & Nguyen, H. C. (2017, June). Towards zero data interruption time with enhanced synchronous handover. In *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)* (pp. 1-6). IEEE.

[27] Li, J., & Chen, J. (2017). Passive optical network based mobile backhaul enabling ultra-low latency for communications among base stations. *IEEE/OSA Journal of Optical Communications and Networking*, *9*(10), 855-863.

[28] Manikanthan, S.V., Padmapriya, T., An efficient cluster head selection and routing in mobile WSN, International Journal of Interactive Mobile Technologies, 2019.

[29] D.K. Jayaram, Hyper-Mimo Spectral Efficiency Augmentation Techniques in 5G, IIRJET, 4(4), 2019.

Author Profile

Baburao Kodavati was born in Guntur district, INDIA in 1980. He received the B.Tech degree in Electronics & Communication Engineering from Koneru Lakshmaiah College of Engineering, Andhra Pradesh, India, in 2003 and M.Tech degree in Digital Electronics and Communication Systems from Gudlavalleru Engineering College, Jawaharlal Nehru Technological University Hyderabad, India, in 2007. He is Pursuing Ph.D in JNTUK Kakinada and presently working as an Associate Professor in the Department of Electronics & Communication Engineering, Usha Rama College of Engineering & Technology, NH-16, Telaprolu, Ungutur Mandalam, Near Gannavaram, Krishna District, Andhra Pradesh, India. He has 12 years of teaching experience. He has published more than 25 research papers in various reputed national and international Journals and conferences.

Madhu Ramarakula was born in Warangal district, INDIA in 1980. He received the B.E degree in Electronics & Communication Engineering from Osmania University, Hyderabad, India, in 2003 and M. Tech degree in Communication Systems from Jawaharlal Nehru Technological University Hyderabad, India, in 2009 and Ph.D degree in Electronics & Communication Engineering from Andhra University, Visakhapatnam, India, in 2014. He is presently working as an Assistant Professor in the Department of Electronics & Communication Engineering, University College of Engineering Kakinada, JNTUK Kakinada, India. He has 10 years of teaching experience. He has published more than 30 research papers in various reputed national and international Journals and conferences. His research interests include Mobile communications, satellite communications and GPS. He is a member of IEEE.