Journal of Network and Systems Management
https://doi.org/10.1007/s10922-019-09497-w

Latency Control of ICN Enabled 5G Networks

Shahin Vakilinia[1] · **Halima Elbiaze**[2]

Received: 19 July 2017 / Revised: 14 March 2019 / Accepted: 10 April 2019

Abstract
5G definition falls broadly into order of achievable data rate and reduction in end-to-end latency. Thanks to emerging technologies many features are available in the 5G design to detect, control and avoid congestion in the backhaul networks. In fact, 5G results from the conjunction of several recent technological developments, chief among them the re-purposing of next generation of wireless networks for large-scale functional connectivity and carrying of massive heterogeneous contents. For instance, information centric networks, as a promising candidate for the wireless caching architecture, can cache the contents and prohibits traffic avalanche entering the backhaul via content-based networking. The main objective of this paper is to minimize latency in 5G backhaul networks. The contribution of this paper is a twofold: (a) a distributed algorithm at the back-haul switches is proposed to detect and handle the congestion temporarily and locally with considering the fairness, IP friendliness, latency and convergence time. (b) an SDN-based centralized algorithm is proposed to treat the congestion via dynamic route selection, load-balancing, the orchestration of heterogeneous RBS components.

Keywords 5G · Congestion avoidance · Latency · ICN · Caching · MEC · C-RAN · SDN

1 Introduction

With tremendous recent advances in wireless technologies, it is predicted that 5G will be rolled out to support ultra low-latency and ultra high throughput traffics related to Internet of Things (IoT) and multimedia contents across the heterogeneous networks. To do so, 5G standard is supposed to fulfill 1 Gbps transmission speed to each user [1]. Furthermore, to support continuous coverage, 5G comprises densely

✉ Shahin Vakilinia
shahin.vakilinia@gmail.com

[1] Ericsson Research Team, Canada, Université de Québec à Montréal (UQAM), Montreal, Canada

[2] Département d'informatique, Université de Québec à Montréal (UQAM), Montreal, Canada

Published online: 22 April 2019

clustered cells over urban areas. Thus, the high transmission rate of the users along with the enormous growth of the various cells require precise traffic management mechanism in a flexible Radio Access Network (RAN) to avoid congestion in the 5G mobile backhaul network [2, 3].

On the other hand, taking large-video distribution as one of the key challenges to address early for 5G and the content-centric nature of current mobile network usage calls for substantial innovation at the network layer. Note that the current caching protocols in mobile backhaul share two significant drawbacks: they provide limited performance gain due to their applicability at the network edge only, and they use an HTTP-based, slow TCP-connection-oriented transport model deployed as overlays on top of the existing IP network infrastructure [4–6]. Moreover, despite the many studies [7] on caching in wireless networks, the current system lacks standardization. Owing to its many advantages, information-centric networks (ICNs) are a plausible candidate to be embedded in 5G for this purpose [4, 5, 8]. The ICN, as a promising system architecture for content distribution in 5G, avoids congestion via the in-network caching mechanism so that redundant traffic demand will not circulate in the 5G backbone network anymore for the sake of achieving efficient large-scale information delivery. Multi-source, multi-path forwarding, and multi-cast data-delivery techniques in the ICN also help manage traffic to avoid congestion.

In this paper, an architecture for 5G networks is proposed. Next, the application of the in-network caching mechanism, particularly the ICN protocol, in 5G networks is investigated. Considering in-network caching mechanisms for 5G and the heterogeneity of traffic and links in the 5G backhaul network, a distributed algorithm at the backhaul switches is proposed first to detect and handle the congestion temporarily and locally while considering fairness, IP friendliness, latency, and convergence time. Second, an software-designed networking (SDN)-based centralized algorithm is proposed to treat the congestion via dynamic route selection, load balancing, and the orchestration of heterogeneous RBS components.

The rest of the paper is organized as follows. The proposed architecture of the 5G protocol, including the ICN protocol, is described in the next section. Challenges and limitations of caching aware congestion-control mechanism for 5G scenarios are described in Sect. 3. Then, distributed and centralized algorithms are proposed in Sects. 5.1 and 5.2, respectively. Finally, we conclude the paper in Sect. 6.

2 5G Architecture (CRAN versus MEC)

The high transmission rate of the users, along with the enormous growth of the various cells, requires a precise traffic management mechanism in a flexible radio-access network (RAN) to avoid congestion in the 5G mobile backhaul network [2, 3]. The flexible RAN architecture of 5G is catered by splitting the RAN functions of the radio-interface protocol stack in radio base stations (RBSs) [9]. Thus, RBS components, namely remote radio head (RRH), baseband-processing function (BPF), and packet-processing function (PPF), are split depending on the 5G architecture.

Recent advances in virtualization and cloud computing complement this flexible mobile backhaul architecture via Cloud-RAN (CRAN), in which the RAN

computational functionalities, BPF and PPF, are decoupled from the RRH and are virtualized and located at the cloud data center (DC) so as to provide the consolidation gain and a higher degree of cooperation among RBSs [10]. That is, CRAN, as a new centralized paradigm for the next generation of wireless networks, addresses the fluctuation in capacity demand efficiently, and in addition to many other advantages, avoids bottlenecks and problems due to a lack of computing resources. However, the downsides of CRAN may hinder its effectiveness. Inefficiencies in the current CRAN communication model result in a backhaul used as a passive network segment, where congestion phenomena are suffered and not dynamically managed. In contrast with CRAN, in which all the computational functionalities are located at the cloud DC, caching in the air, or mobile edge computing (MEC) [11], co-locates computing and storage resources at the RBS towers (or at least near the access network) of 5G networks to alleviate utilization of the core network and decrease the backhaul traffic to reduce latency. By running applications and performing related processing tasks closer to the mobile user, congestion at the backhaul is reduced and the performance will be improved. Furthermore, [4] shows that 50% of requests are cacheable and traffic can be reduced by 60–95% during the peak hour in cellular networks. Thus, the data can be cached at the tower and the information required by the mobile user can be sent back to the user without going through the backhaul and core network, which contradicts with the CRAN scenario regarding the processing location. The MEC model has its own drawback: the commodity hardware of the MEC servers may become the processing bottleneck of the backhaul network. Moreover, cooperation among RBSs may cause latency (compared to CRAN) in the sense that for distributed MIMO systems, the BPF at each antenna may require some channel information and signals from its neighbors, and this communication may cause latency in processing the RRH output.

In reality, along with MEC and CRAN in 5G, some central computation offices (cloudlets) are also embedded between the DC and RBS tower. Taking the cloudlets between the DC and towers makes the scenario more complicated in that it entails a processing-location-management platform to avoid latency. Software-Defined RAN is an integrated technology of 5G that provides the capability to configure and scale the components through software commands and enables the RAN to adjust computing resources dynamically according to traffic load conditions.

Figure 1 shows the proposed general flexible architecture of a 5G backhaul network. As seen in Fig. 1, the RBS tower may support a various number of computational functionalities. For each RRH, several mutant BPFs and PPFs are available toward the cloud DC, which can be selected dynamically through the multi-path-handler and orchestrator modules of the control plane.

The distributed RBS design helps manage the incoming traffic via multi-path routing and balancing the heterogeneous workloads over the backhaul network, which requires switching capability [12]. Therefore, all different types of communication protocols in the fronthaul and backhaul networks, namely CPRI (C1) [13], which connects RRH to BPF; BB-UI, connects BPF with PPF; and S1-U, which connects PPF with SGW, exist alongside each other on top of an MPLS/

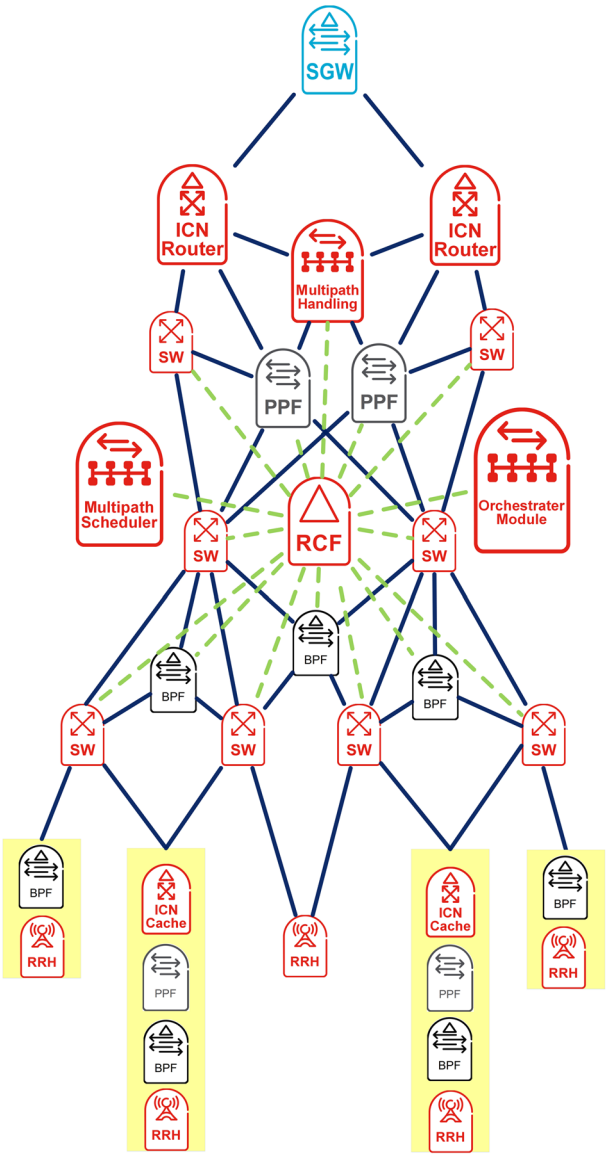


Fig. 1 General architecture of a 5G scenario

Ethernet-over-fiber carrier, as explained in [14, 15]. Hence, the primary concern here is that various protocols (C1, BB-UI, and S1-U) are carried on Ethernet/MPLS through the fronthaul/backhaul links, which makes the traffic management problem complicated and enhances the risk of congestion over the network. Note

that the rest of the backhaul network (from SGW to PDN-GW), in almost all the proposed scenarios for 5G, is located in the DC [10, 16].

2.1 ICN in 5G

In the ICN, to fetch data by names, mobile users pull data packets by sending out Interest packets to the network, where data chunks of the same content (requested by Interest packets) may be retrieved from different repositories or different caches along the paths towards these repositories on the other side of core networks. The shift from a location-focused network to a content-centric network allows higher capacity while supporting low latency for massive content distribution. This mechanism exploits the data available at any intermediate point (e.g., caches in the routers) to serve the requests from any potential content source in the network rather than a single content source. In an ICN router, new incoming Interests will be added to the pending Interest table (PIT). The PIT keeps track of the Interests forwarded upstream toward the content source so that the returned Data can be sent downstream to its requesters. New incoming Interests for a piece of content wait until the number of Interests for the content passes a threshold, and then A forwarding information base (FIB) module sends the Interest for outgoing interfaces. Later on, these popular contents will be stored (cached) locally at the content store (CS) of the ICN router. Hence, the Interest rate is drastically decreased by the cache hit ratio of CS [4]. The performance of the ICN routers with Interest and Data packets is simply represented in Fig. 2, which is popular and vastly presented in many papers such as [17–19].

After the baseband and packet processing over the RRH output, the packet-data convergence protocol (PDCP) is used over GPU tunneling for communication; otherwise, the C1 and BB-UI protocols must be used [14, 15]. The PDCP is located in the radio protocol stack on top of the radio link control (RLC) layer. PDCP provides its services to RRC (by transporting RRC messages) and the upper layers of the user

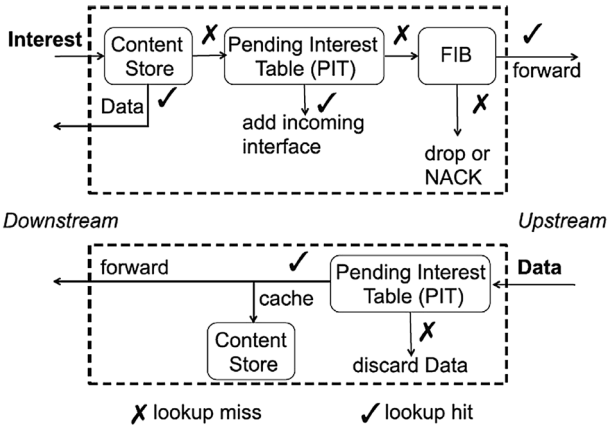


Fig. 2 Interest and data processing in ICN nodes [17–19]

plane, e.g., IP at the UE or to the RBS. PDCP provides the following services to the upper layers: transfer of user-plane data, transfer of control-plane data, header compression, ciphering, and integrity protection. The RAN functions participating in the loop work synchronously with the RRH, whereas the PDCP and multipath handling function feed and receive packets traveling to and from the RLC layer asynchronously, which has implications for the split architecture.

We suggest the application of the ICN convergence protocol (ICP) along with the PDCP [14, 15]. ICP is a modified version of PDCP that is devoted to ICN traffic. Because ICN congestion control/management is far different from the IP world, recognizing the ICN traffic from the IP traffic provides many advantages regarding traffic and resource management. Furthermore, separate PGWs for ICN networks and TCP/IP networks can be embedded to decouple ICN and TCP/IP traffics at the core network inside the DC, whereas RBS and PGWs share the same backhaul network with two different upper MAC protocols (tunnels), namely PDCP and ICP over the MAC layer. When traffics are distinguishable at the mobile backhaul networks, they can be easily forwarded to different PGWs.

Figure 3 (next page) shows the heterogeneous traffic architecture of the ICN-enabled 5G-supporting CRAN. As it shows, ICP and PDCP are adjudged on top of the GTP-U tunnels to carry the IP and ICN traffic, respectively. Note that in the data-link layer, Ethernet, the IP and ICN traffics are also differentiable through the EtherType field, which indicates the encapsulated protocol in the payload of the Ethernet frame. Thus, L2 Switches can apply different policies on them. Figure 3 shows

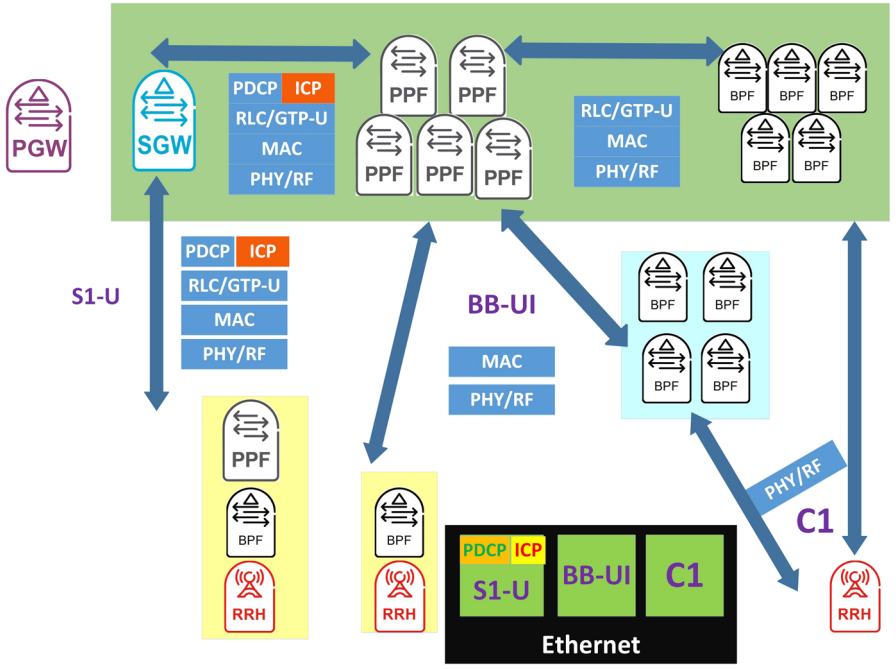


Fig. 3 Heterogeneous traffic architecture of the ICN-enabled 5G-supporting CRAN

the existence of C1, BB-UI, and S1-U, which makes the traffic in the 5G backhaul link heterogeneous. For the sake of brevity, the details and architecture of the communication protocols are not described further.

3 Challenges in Controlling the Congestion for 5G Backhaul

To prevent backhaul saturation, edge caching and in-network caching of MEC [11, 20] and ICN reduce traffic load while improving content delivery latency. Despite all the mentioned advantages, there are still some impediments such as Pending Interest table (PIT) avalanche growth, and the co-existence with IP traffic for applying ICN in 5G. There are two types of latencies that threaten the 5G back-haul QoS metrics; one caused by the congestion at the network links and the other one stems from processing. Therefore, proposed congestion control mechanism should tackle both of them.

3.1 Limitations of ICN Congestion Control Algorithms

Though many studies has been done over the ICN congestion control algorithms, it is deemed that none of the methods presented in the literature are appropriate for 5G scenarios which stems from the following reasons:

Congestion Detection Congestion is vastly detected or predicted via bandwidth consumption, delay fluctuations(RTT/RTO) or drop rate in typical congestion control mechanisms. According to the Heterogeneous architecture of 5G, the available capacity of IP/ICN tunnels over the backhaul network changes dynamically which is not predictable. Moreover, multisources (different content sources), multi-path and PIT aggregation in ICN, results in varying RTT and retrieval delay in which the delay fluctuation increases and consequently, congestion detection error of RTT-based timeouts (RTO) enhances. It makes the delay deviation unreliable indicator of congestion. In addition, drop rate, as congestion indicator, also threatens the QoS Metrics and is not permissible at 5G backhaul.
Congestion Signaling Congestion control algorithm can be implemented in Data-link layer at 5G network. However, many congestion control protocols detect congestion at higher OSI layers (application and transport layers) at receiver. In addition, in mobile cellular network scenarios, backhaul connects mobile users from the front-haul (access network) to the core network, and there is no end-to-end client(consumer)/server(sender) unless the ICN routers at the backhaul hit the incoming ICN Interest packers. For ICN scenarios, the easiest way to control the Data traffic is through controlling the sending rate of Interest packets at the receiver. Most of the ICN congestion control algorithms in the literature could be roughly categorized in this group [17]. Many focuses on congestion control algorithm, according to the network performance parameters fed back via data plane to the reaction point, tend to be implicit. However, thanks to SDN and 5G characteristics, a control plane is available for

signaling at the Backhaul. Thus, in 5G, there is no need to use implicit methods and explicit signaling can be handled by the control plane.

Reaction Point Off-the shelf congestion control mechanisms are either Reactive or Proactive. Reactive schemes threaten the QoS metrics such as throughput by applying belated policies while being proactive and instantaneous decision-making support the QoS parameters. However, since the proposed 5G congestion control algorithm has to collaborate with other congestion control protocols at upper layers, the proposed method should give enough time to other methods to react accordingly and being proactive exacerbate delay fluctuations of the backhaul and is harmful to the traffics or flows with other congestion control mechanisms such as TCP. The up-link traffic source of mobile users depends on the SLA and is not tractable and consequently the ICN consumers are not controllable. Thus, the congestion control algorithms have to be applied on the intermediate nodes of the backhaul via Interest shaping.

Rate Adjustment Since the mobile user content request is not governable from the 5G backhaul, The proposed solutions are not applicable in the 5G backhaul networks. Authors in [21], concluded that Receiver-based schemes are not the appropriate option for NDN due to the unpredictability of the content locations. In Interest shaping methods, every ICN aware intermediate nodes at the backhaul, which receives the Interest packets of mobile users is capable of controlling the congestion, by either dropping Interest packets, stalling it or diverting them to alternative paths. Hence, Interest Shaping hinders late data packet loss which waste considerable network resources via early Interest packet dropping [17] and is of great interest to tame the local benign congestions.It is worth mentioning that the intermediate nodes are both the congestion detection and Reaction points in Interest shaping methods.

Moreover, beside the Interest shaping, the intermediate node may notify the SDN-based centralized controller about their local congestion. However, most existing solutions are based on predictable bandwidths and Data chunk sizes [17], which does not hold for the 5G backhaul scenarios. Hybrid methods combine two aforementioned methods(Receiver-based and Interest Shaping) and by taking their advantages outperform those algorithms. However, since the receiver-driven algorithms are not applicable for 5G backhaul scenarios, the hybrid methods are also not beneficial in 5G networks.

Control Mode In the current Internet, many congestion control protocols are Window based like (TCP) [22]. Most of the algorithms follow the Additive Increase Multiplicative Decrease (AIMD) mechanism like TCP to regulate the rate of data requests by the consumer which may cause the short-term sawtooth-like rate behavior. On the other hand, in rate based congestion control mechanisms [23], a desired Interest sending rate is consecrated to each flow as the mechanism to control the load of both Interests and Data messages which produce much smoother rates better suited for real-time mobile applications. Contrariwise, AIMD aids Window-based congestion control to be TCP-friendly whereas rate-based congestion control may need RTT to be TCP friendly [22, 24].

To the best of our knowledge no congestion control algorithm addresses heterogeneous scenario which covers both IP and ICN traffics along with considering of fairness issues among them.

4 Protocol Design Requirements

According to the explanations above, the proposed algorithm has to be explicit, and apply Interest shaping methods at intermediate nodes and have to leverage between being pro-active and reactive.

Due to being hampered to apply the indicators above, the only parameter left for congestion detection is the *queue length*. This parameter is vastly used in the modern networks and accordingly buffer-bloating detection modules and Active Queue Management (AQM) advanced techniques are developed along with the traditional methods such as Random Early Detection (RED). These techniques keep the median delay low, which is the main target of this paper. However, the only drawback is that for centralized control case, due to the fact that queuing delay is shorter than the control loop delay, so to speak, queue fluctuations become fast and too complicated to rely on [45]. Table 1 classifies popular ICN congestion control algorithms according to different criteria.As it can be seen, according to the descriptions above, PCON [44] and NACK [19] are great candidates to be applied in 5G backhaul networks. For evaluating their performance, following metrics, which are of ultimate importance in 5G, are investigated in the literature, and the comparison results are provided in Table 2.

- Network Efficiency(Utilization)
- Application Efficiency
- Convergence Time
- Reliability
- IP-Friendliness
- Fairness

Since 5G is vastly applied for the M2M communication and IoT scenarios, fairness issue is of ultimate importance. Furthermore, other congestion control protocols have to be taken into consideration in the wireless environment. For instance, if an upper layer congestion control protocol of 5G user/device respond to an IP/Interest packet loss by decreasing the congestion window, if does not stall or throttled,it suffers from negative impacts. This problem is known as the loss path multiplicity problem [46]. It is also worth mentioning that the congestion algorithms has to cover fairly both IP and ICN in the backhaul. Thus, none of the algorithms in the literature covers all the performance metrics. In fact, developing one algorithm covering all the performance metrics is almost impossible. Even PCON [44] and NACK [19] addressed specific objectives. Authors in [44] also developed an algorithm for the wireless access networks and applied [47] for the AQM. Taking the same approach, we propose two distributed and centralized algorithms to address all the performance metrics mentioned above.

Table 1 Metrics

Algorithm	Explicitness	Detection	Rate adjustment/control mode	Reaction point	Reaction type
[25–30]	Implicit	RTO Time out	AIMD-like Window/Receiver-based	Consumer	Reactive
Predictive [21]	Implicit	RTO Time out	AIMD Window/Receiver-based	Consumer	Proactive
CCS [31]	Explicit	Queue(Data)	AIAD-like Window/Receiver-based	Consumer	Proactive
ECP [32]	Explicit	Queue(Interest)	MIAIMD/Receiver-based	Consumer	Proactive
SECN [33]	Explicit	BW	Receiver-based Path Forwarding	Consumer	Proactive
HoBHIS [34]	Explicit	Queue/BW	Rate Based Interest shaping	Routers	Proactive
HIS [35]	Explicit	Interest Rate/BW	Rate based Interest shaping	Routers	Proactive
HR-ICP [36]	Explicit	Rate/RTO	Hybrid Method	Routers/Consumer	Reactive
CHoPCoP [37, 38]	Explicit	Queue	AIMD based REM (Hybrid Method)	Routers/Consumers	Proactive
NMRTS [39, 40]	Implicit	Interest Rate	Rate based Interest shaping	Routers	Proactive
[22]	Implicit	RTO Timeout	AIMD Window-based Interest shaping	Routers	Proactive
[41]	Explicit	BW/Delay	Rate/Receiver based Interest shaping	Consumers	Proactive
MIRCC [42]	Implicit	RTT/BW	Hybrid Method	Router/Consumer	Proactive
SAID [43]	Implicit	BW	AIMD Receiver Based	Consumer	Proactive
NACK [19]	Explicit	Queue(Interest)	Rate Based Interest Shaping Multi-path Forwarding	Routers	Stateful
PCON [44]	Explicit	Queue	AQM Interest Shaping	Routers	Proactive

Table 2 Performance metrics

Algorithm	Network efficiency	Application efficiency	Convergence time	TCP/IP-friendliness	Reliability	Fairness	Wireless
CCS [31]	✓	×	×	×	×	✓	×
ECP [32]	✓	×	×	×	×	×	×
SECN [33]	✓	✓	×	×	×	✓	×
HIS [35]	✓	×	×	×	×	✓	×
HoBHIS [34]	×	×	×	×	✓	×	×
HR-ICP [36]	×	×	✓	×	✓	✓	×
ConTug [28],	×	×	×	✓	×	✓	×
CCTCP [29]	×	✓	×	✓	×	×	×
CHoPCoP [37, 38]	✓	×	×	×	✓	✓	×
NMRTS [39, 40]	×	✓	×	×	×	×	✓
[22]	×	×	×	✓	×	×	×
[41]	✓	✓	×	×	×	×	✓
MIRCC [42]	✓	×	✓	×	×	✓	×
SAID [43]	✓	✓	×	×	✓	×	×
PCON [44]	✓	×	×	×	×	✓	✓
NACK [19]	✓	✓	×	✓	×	×	×

5 Proposed Solution for Congestion Avoidance and Latency Minimization

In this section, considering a global scenario, covering both C-RAN and MEC and applying ICN protocol on the 5G backhaul network, a platform is proposed to manage the congestion in the 5G backhaul network. To this end, first, the pioneering idea of the application of joint ICN and C-RAN/MEC at the 5G backhaul is investigated. Second, regarding the general 5G architecture, a distributed algorithm at the back-haul switches is proposed to detect and handle the congestion temporarily and locally with considering the fairness, friendliness, latency, and convergence time and reliability. Finally, a centralized algorithm in collaboration with distributed algorithm is proposed to handle the congestion via dynamic route selection, load-balancing and the orchestration of heterogeneous RBS components all over the backhaul network.

5.1 Proposed Distributed Algorithm

Through applying AQM techniques, the proposed algorithm should keep the delays low while insensitive to bursts and available bandwidth. It should also adapt to dynamically changing rates with minimum knobs to be adjusted without negative impact on utilization. Since the BW, delay and other parameters are too dynamic

and heterogeneous in ICN aware 5G scenarios, queue length left as the only option for the congestion detection. However, queue length fluctuation is ineluctable which may lead to over-react and increasing error in congestion detection. A robust way to improve congestion detection is to take another queue factor besides the queue length. Though minimum queue length used in [47] is a very robust way to detect the congestion, yet not reliable and efficient enough to be applied in 5G backhaul network. Thus, two other factors namely; (1) the busy period duration, and (2) the difference between average queue and minimum queue lengths are considered in our proposed algorithm. Let τ_t denote the busy period associated with state n is defined as the amount of sojourn time that there are more than n packets in the queue at time slot t. Averaging the queue length and finding the minimum queue length should be done in a sliding window. Here δ_t represents the difference between the average and minimum values of the queue length q_t at time slot t.

For being both proactive and reactive enough, a Finite State Machine (FSM) is proposed to track the queue changes and congestion in the system. Being state-full aids to minimize the monitoring process and to be harmless to the data plane computation. As it is shown in Fig. 4, the proposed FSM has three states, namely Normal (Green), Warning (Yellow) and Dropping (Red). The transition between the states does depend on the events generated via comparison of the metrics with their attributed thresholds. In Normal state, the difference between average queue length and minimum queue length as well as the busy period are also less than the thresholds τ_w and δ_w respectively. Consequently, the delay at the switch is suitable (tolerable). If one of those thresholds is passed by the system, there would be a transition from green state to yellow state. That is, if the $\delta_t \succeq \delta_w$ or $\tau_t \succeq \tau_w$ then the system goes to warning state. In warning state, beside the queue length, other metrics such as the flow rates are estimated. It should be noted that if either, the queue difference passes the δ_r or the busy period passes the τ_r (i.e., $\delta_t \succeq \delta_r \tau_t u \succeq \tau_r$), the system turns into the Dropping (Red) state. Setting these thresholds strictly depends on the SLA and expected QoS of 5G backhaul network.

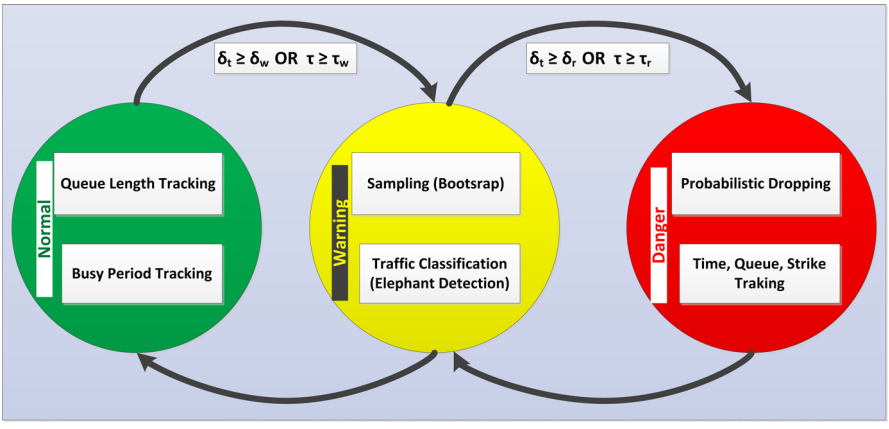


Fig. 4 FSM of the proposed distributed algorithm

Algorithm 1 Distributed algorithm

```
procedure CongestionHandling
    while 1 do
            Q = Monitor(Queue)
            [Min_Q, Ave_Q, δ_t] = Calculate(q_t)
        τ_t = Min_Q − Ave_Q
        Update(State)
        Case Green
        if [τ_t > τ_w || δ_t > δ_w] then
            state = yellow;
            t_0 = getTime();
        Case Yellow
            X = DetectElephants()
            [Blacklist, ZombieList] = update(X)
        if τ_t < τ_w & &δ_t < δ_w] then
            state = Green ;
        if [(τ_t > τ_r)||(δ_t > δ_r)||(getTime()−t0 > T_th) then
            state = Red;
            t1 = getTime();
        Case Red
            X = DetectElephants()
            [Blacklist, ZombieList] = update(X)
        Strike = Retriev(Blacklist, ZombieList)
        Drop([Min_Q, Ave_Q], Strike, gettime() − t1)
        if [(τ_t < τ_r) & &(δ_t < δ_r) then
            state = yellow;
```

In each of these states, the switch performs different tasks. In the Green state, just queue metrics are tracked while in the Yellow state, the greediest flows beside the queue parameters are pursued. The monitoring procedure hinges at the location of the switch at the 5G architecture. If it is directly connected to the control plane, and the history of up-link traffic is accessible, this information can be used in ranking the flows. Otherwise, different sampling techniques and estimation methods [48–51], have to be applied to detect the massive multimedia flows. Moreover, depending on the available computing resource, the flow definition can be varied. In the most basic form, flow can be defined according to the source antennas (Physical Cell Identity (PCI) or Cell Global ID (CGID)) while in the more complicated case, the flows can be assigned based on the device/user identity such as International Mobile equipment identity (IMEI), International Mobile Subscriber Identity (IMSI) and 5 Tuples [3]. It is convenient that if the resolution goes to the five tuples rather than the antennas ID, the accuracy of the proposed algorithm, regarding the fairness and QoS, might enhance. Although the computation functionalities are supposed to be done at the separate CPU core in the Switch yet may not be beneficial and efficient under some circumstances.

For practical consideration, in Yellow (warning) state, most bandwidth demanding flows according to their number of packets in the queue are set in a way that the ones with a higher number of IP packets in the queue are placed in a list called Black-list which has limited elements. The location in the so-called Black-list depends on the history of the flow; that is, how many times it was detected previously as a bad (bandwidth demanding) flow. In the Blacklist, an Strike factor [24] is also attributed to each flow. Strike factor is defined to indicate how the listed flow is congestion aware and whether has congestion control algorithm such that if a packet of the flow drops, how the end user/device application react (may decrease the traffic rate). Therefore, longer the flow in the system, higher its rank will be in the list. In next time slot, after the packet dropping, if the flow rate decreases then striking factor will decrease otherwise strike factor will stay fix. Thus, there would be a group of bad flows in Blacklist. Beside the Blacklist, another list named Zombie list exists for ICN flows. The Blacklist is assigned to each port while the Zombie list is general and is shared among the ports. In Red State, similar to RED algorithm [24], the switch starts to drop the IP and ICN Interests packets from the r^{th} element of Blacklist (in IP case) or Zombie-List (in ICN case) probabilistically according to the following probability:

$$P_D(r) = Min\left\{ \frac{q_t - Q_{min}}{Q_{max} - Q_{min}} . \frac{t - t_o}{\Delta_{Max}} . S_r, 1 \right\}, \tag{1}$$

where t_o and t represent the time that switch enters the Red state and current-time respectively. Δ_{Max} indicates the maximum tolerable time at the Red state. Q_{max} and Q_{min} are constant parameters that represent the maximum tolerable queue length and minimum queue length at the Red state(similar to the RED algorithm) respectively. Unlike the RED, which just considers the queue length, the proposed dropping scheme considers also the congestion period and it is vigilant of upper layer congestion control algorithms. In fact, RED cares more about the stability rather than the latency and fairness. Since the congestion duration [52] is of distinctive impression in ICN, the time factor also exists in proposed probabilistic dropping. The time factor diminishes the convergence time and make the proposed algorithm suitable for temporal congestion handling. It should also be noted that the ICN Data packets are never dropped, and Interest packets of ICN are just discarded according to (1). Note that for really high S_r rate the $P_D(r)$ tends to value 1.

It is assumed that ICN and IP have different tunneling techniques and labels called ICP and PDCP and as a result, they can have different Ether-types and are distinguishable at the switch. The reaction to a probabilistic dropping of packets is different. In the proposed algorithm for ICN, the black list is global for all the ports and Data packets should never be dropped. Instead, Interest packets will be blocked. While in IP, similar to typical AQM algorithms, packet associated with the port will be dropped. The difference between the CoDel and the proposed algorithm is that CoDel tries to decrease the throughput linearly and drop the packets in a deterministic way while the queue length is a Random walk process and its probabilistic characteristic contradicts with the deterministic actions. In this scenario, the main target here is to temporally handle the congestion which may require decreasing

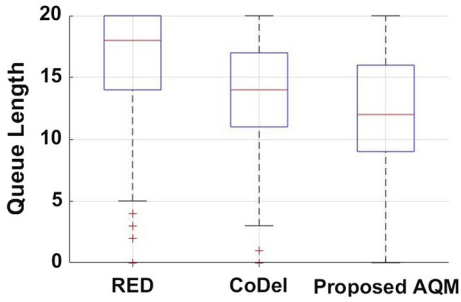


Fig. 5 Comparison of queue lengths in AQM schemes (Boxplot)

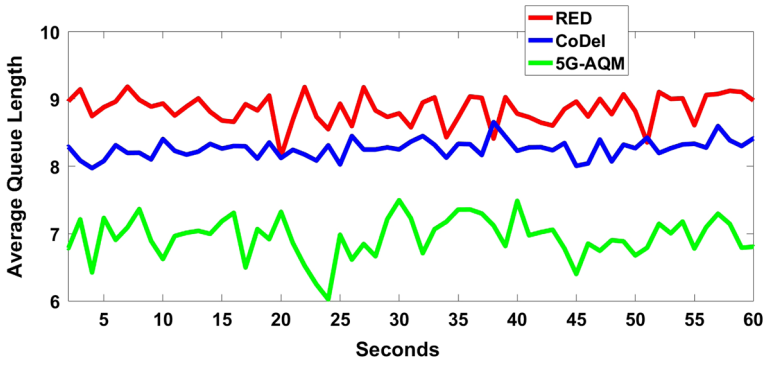


Fig. 6 Comparison of queue lengths in different AQM schemes as a function of time

the throughput non-linearly. In other words, a Desired Operational Point (DOP) is defined for the switch which tries to control the queue length (latency) fairly over the time. Thus, the proposed distributed algorithm is designed to handle the congestion locally and temporally. After entering Yellow and Red states, a notification called Explicit Congestion Notification (ECN) is sent to the centralized SDN-based controller to re-schedule, balance the load and handle the congestion along with the consideration of general Back-haul architecture.

A Discrete-Event based Simulation (DES) is done to evaluate the performance of the proposed algorithm. Different users and devices according to Markovian Modulated Poisson Process (MMPP) generate different traffics and send it to their associated RBS which may go through a switch. 3 different types of traffics are considered. IoT low rate traffic (with maximum of 1 Mbps), congestion aware flow(TCP-like) and a greedy multimedia connection. The simulation (with 2 greedy, 2 congestion aware and 6 IoT flows) is run for Codel, RED and our proposed AQM and its results are box plotted in Fig. 5. The red lines depicted in Fig. 5 show that the average queue length using the proposed AQM is less than other AQM techniques while it holds the least variance as well. Note that in the simulation knobs are configured to have the best performance of the aforementioned schemes. For instance, to avoid buffer-bloating the buffer size is set 20 and since the best performance of the CoDel and RED are in shallow buffers, the buffers size reduced to 10. Figure 6 presents the

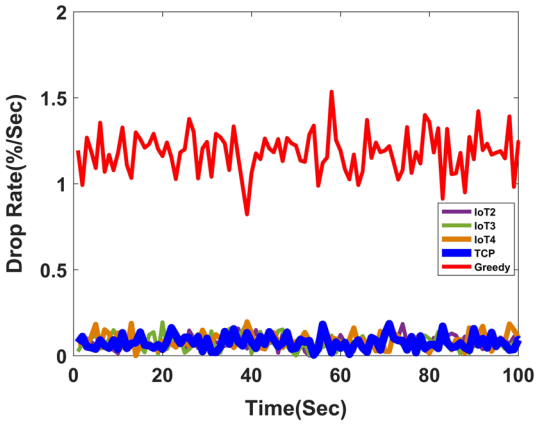


Fig. 7 Switch's drop rate as a function of time using CoDel

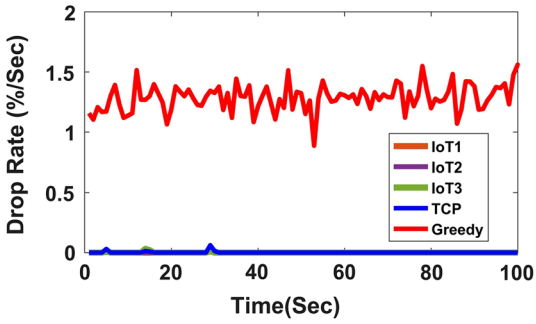


Fig. 8 Switch's drop rate as a function of time using te proposed AQM

average queue length per second over the time and as it can be seen the proposed AQM still has lower queue line up compared to other AQM schemes. Regarding the fairness and friedliness of the proposed AQM, drop rate of CoDel and proposed AQM are shown in Figs. 7 and 8 respectively. Figure 7 shows that CoDel, despite being more fair than RED algorithm, drop packets from congestion-aware flow (blue one) and low rate traffics while the proposed algorithm presented in 8 does not harm the congestion-aware and low rate flows.

A Discrete-Event based Simulation (DES) is done to evaluate the performance of the proposed algorithm. Different users and devices according to Markovian Modulated Poisson Process (MMPP) generate different traffics and send it to their associated RBS which may go through a switch. 3 different types of traffics are considered. IoT low rate traffic (with maximum of 1 Mbps), congestion aware flow (TCP-like) and a greedy multimedia connection. The simulation (with 2 greedy, 2 congestion aware and 6 IoT flows) is run for Codel, RED and our proposed AQM and its results are box plotted in Fig. 5. The red lines depicted in Fig. 5 show that the average queue length using the proposed AQM is less than other AQM techniques while it holds the least variance as well. Note that in the simulation knobs are configured to have the best performance of the aforementioned schemes. For instance, to avoid buffer-bloating the buffer size is set 20 and since the best performance of the CoDel and RED are in shallow buffers, the buffers size reduced to 10. Figure 6 presents the

average queue length per second over the time and as it can be seen, the proposed AQM still has lower queue line up compared to other AQM schemes. Regarding the fairness and friedliness of the proposed AQM, drop rate of CoDel and proposed AQM are shown in Figs. 7 and 8 respectively. Figure 7 shows that CoDel, despite better performance in fairness compared to RED algorithm, drop packets from congestion-aware flows (blue one) and low rate traffics while the proposed algorithm presented in 8 does not harm the congestion-aware and low rate flows which proves that the proposed algorithm is friendly to the low-rate and congestion aware flows.

5.2 Centralized Algorithm

Due to the complexity of the 5G backhaul, the locally distributed algorithm is not enough to handle the congestion control in the backhaul network. Thus, it necessitates emerging of the SDN/NFV based controller to manage the mobile traffics with consideration of the status of backhaul components. The primary objective of this paper is to avoid congestion in the 5G- Back-haul up-link. Congestion is depended on the traffic passing through the 5G network while the traffic and available bandwidth are too dynamic and can not be measured accurately. Hence, following metrics are taken into consideration in various components to be able to detect and avoid the congestion which will be described in this section.

- Queue length in the switches
- PIT growth rate
- CPU Utilization at the processing functions

5.2.1 Queue Length at the Switches

Although the arrival rate is not predictable, the network congestion can be tracked via queue length and the ECNs generated from the distributed algorithms running in switches can inform the centralized SDN controller of the congestion. In the yellow and Red states, a warning event and ECN will be sent out to the central event handler respectively. As soon as the central event handler gets the warning event, re-routing module will be activated to find the other available route for the RBS antenna with greedy flows. If the switch stays on the warning message, for a time longer than a threshold, or receives ECN, the centralized controller immediately substitutes the current route. It starts from the greediest RBS in the list and periodically changes the path of the next high-ranked antenna unless it relieves back to the green event. Thus, if the backhaul path negatively menaces the QoS of the 5G users, the congestion detection messages will be sent to the controller, and the path will be varied over the time. However, receiving some ECNs from different switches in a short period shows the global congestion, and it is better to decrease the total rate of greedy users at the Greedy RBSs. Under these circumstances, the greedy users at the marked high rate RBSs should be detected and their rates have to be lowered.

5.2.2 PIT Growth Rate at the ICN Routers

Another gnomon of congestion is Interest growth rate of PIT that can be detected in ICN routers [35]. The stateful architecture of the ICN aids the centralized controller to be able to predict the congestion in the backhaul. Thus, if the growth rate of Pending Interest items exceed a threshold, then a notification has to be sent to the centralized controller to increase reserve bandwidth for both uplink Interest traffic and downlink Data traffic. In another word, increment rate of the entities in PIT tables in a time slot could be considered as a parameter to predict the future ICN traffic which may require higher bandwidth. For more details, please refer to [35]. After getting the PIT growth rate alert, centralized controller avoids the congestion control via the load balancing and re-routing the path of greediest RBS with the highest interest traffic rate.

5.2.3 Average CPU Utilization of the BPF/PPF

Beside the communication delay, the computation also may cause latency. Processing latency is also dependent on the incoming workload. The other factor that can be used to approximate the arrival rate is the CPU usage of the processing function. In the proposed centralized algorithm, the Exponentially Weighted Moving Average(EWMA) function of Utilization is calculated. Then the calculated value will be compared with a threshold. If the CPU utilization is greater than the higher threshold, then the over-utilization event will be forwarded to the centralized event handler. For handling the over-utilization event generated at specific processing point, the RBSs will be sorted according to the processing time variance. Among the RBS, processing load of the ones with higher processing time will be selected and be offloaded to their next processing hop. In fact, the proposed algorithm offload the processing location of antennas with large processing time towards the cloud side. Dynamically management of the processing locations aids to minimize the computation latency and congestion at the back-haul network. However, since the volume of the unprocessed traffic (RRH Output) is much heavier than the processed one (PPF output) and because of the ICN caching mechanism, the best place to process the data is at the RBS towers to cache as much as possible and help decrement of the traffic at the back-haul. Thus, in the proposed centralized algorithm, the backhaul network is initialized with MEC.

Figures 9 and 10 briefly displays flowcharts of the proposed centralized congestion control algorithm.

5.3 Evaluation

In this subsection, a simple 5G scenario is simulated to corroborate the theoretical findings. 5 RBS connected to 4 switches and 2 ICN routers. Heterogeneous traffic(50 flows) from low rate IoT traffic to multimedia IP/ICN traffics are considered in simulation. Communication delay at the cloud is neglected by assuming as an i.i.d

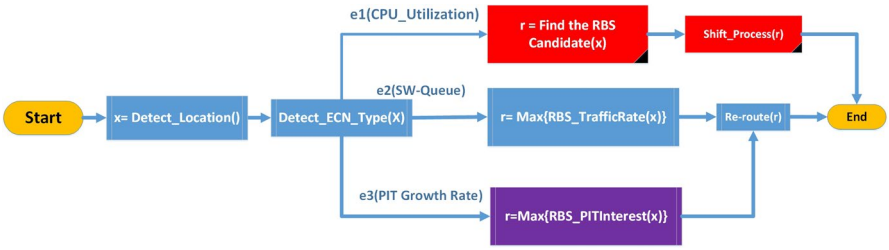


Fig. 9 Flowchart diagram of event handler of proposed centralized algorithm

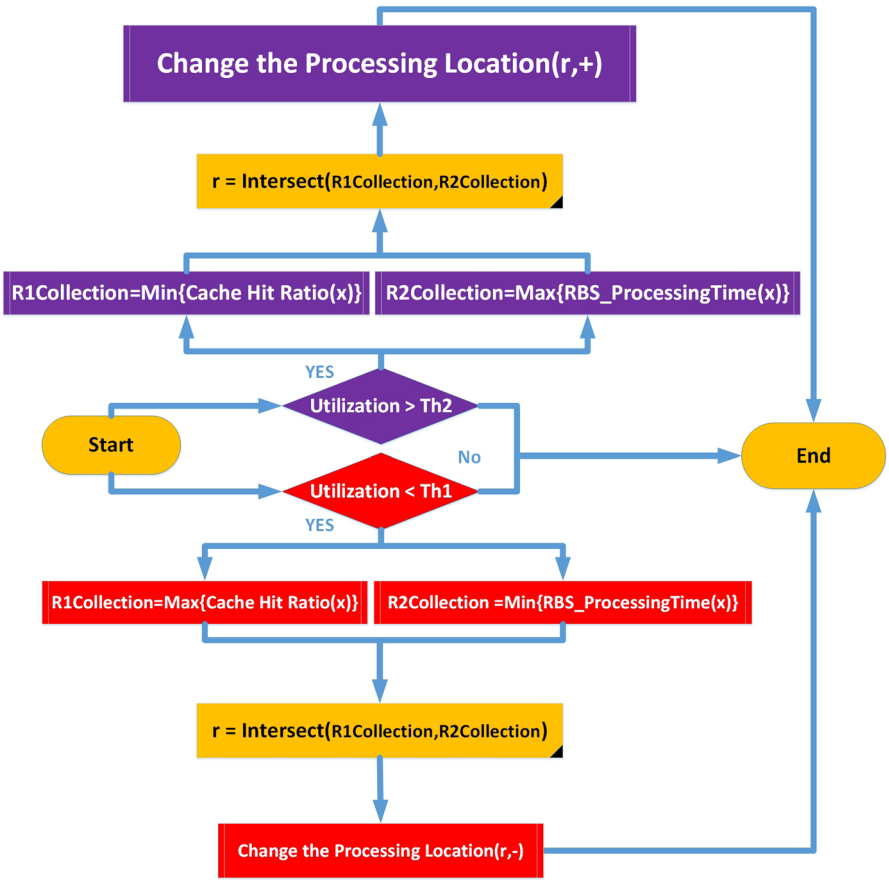


Fig. 10 Flowchart diagram of dynamic location management of processing functionalities

uniform distribution between (0, 4) ms. It is also assumed that the volume of unprocessed data is twice than the processed traffic and the computational functionalities at the cloud are three times faster than the ones on tower. The communication delay of the 5G networks (from RBS to SGI) is compared in Fig. 11. In Fig. 11, scenario

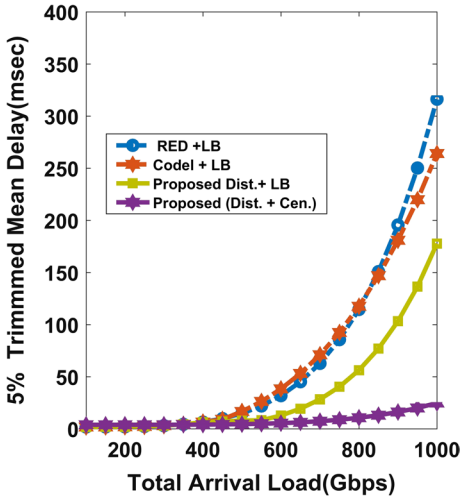


Fig. 11 Comparison of communication delay of 5G networks for different techniques

is similar to typical cellular networks in which BPF and PPF are located at the RBS. For CoDel and RED cases simple load balancing technique is used to distribute the load and is compared with the proposed package of Distributed and Centralized algorithms. To accentuate the impact of centralized algorithm, distributed algorithm is applied with simple load balancer(LB), (yellow line). Figure 11 depicts the communication delay as a function of total uplink traffic rate. As it is depicted, the one-way communication delay resulted by collaboration of two algorithms is much less than other techniques under different load arrival and Distributed algorithm though diminishes the latency yet is not purely efficient without the centralized one. Next, processing location selection algorithm will be evaluated and the proposed package of distributed-centralized algorithms will be compared with the application of these algorithms without capability of dynamic processing allocation. In Fig. 12, the performance of the proposed algorithms is compared with ICN/MEC and CRAN scenarios. For ICN/MEC scenario, it is assumed that latency at ICN router for each Interest packet (checking the content store and etc) at each hope has an i.i.d uniform distribution between (0, 2) ms. Figure 12 shows the average response delay as a function of the average load of the system. As it can be seen, the dynamic management of the processing location provides the great impact on latency reduction and the proposed latency control package purvey less average response delay. To delve in the latency Results in 5G, results for 300 Gbps arrival rate is box plotted in Fig. 13. As it is represented, the average delay for CRAN scenario is the highest while its variance is negligible. In ICN/MEC scenarios has the highest variance and finally the proposed package cause the least average latency in responding while displaying an acceptable variance (Fig. 14).

To check the application efficiency of the proposed congestion control technique, the utilization at the application level is evaluated for the ICN and TCP flows under the 700 Gbps load. Due to the highest latency again C-RAN with CoDel provides the lowest application utilization both for TCP and ICN. MEC/ICN despite being

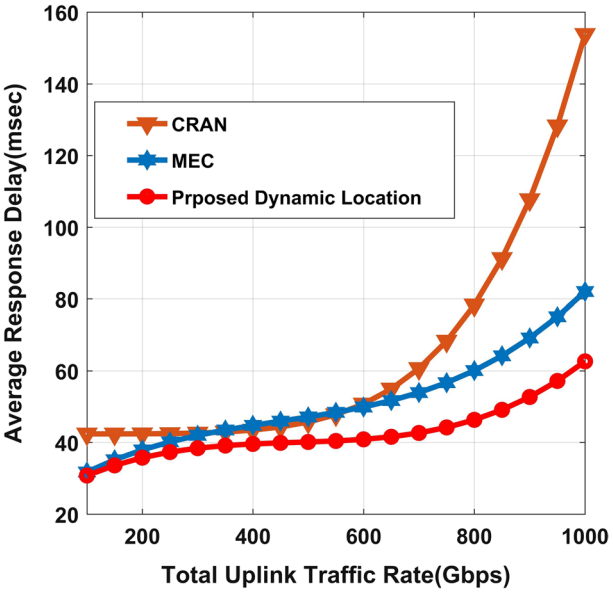
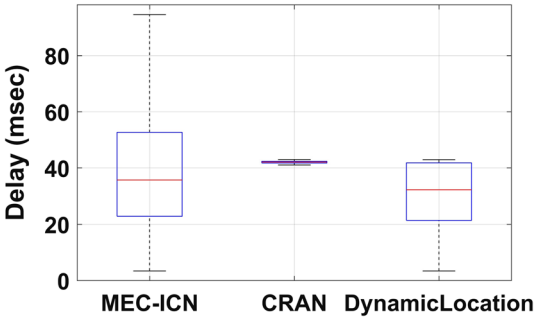


Fig. 12 Comparison of response delay of 5G networks for different techniques

Fig. 13 Comparison of average delay of 5G networks for different techniques (arrival rate = 300 Gbps)



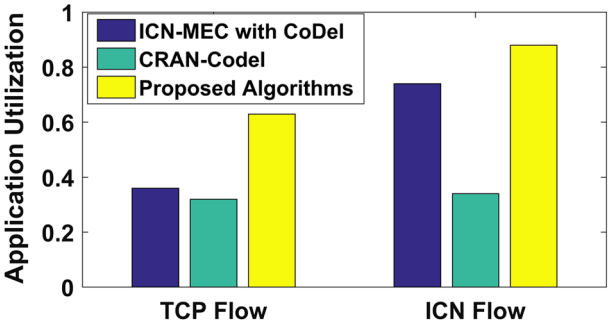


Fig. 14 Application utilization (arrival rate under 700 Gbps)

efficient for ICN traffic can not provide high utilization for the TCP/IP scenarios while the proposed algorithms provide highest utilization (between 60 and 80%) both for ICN and TCP flows through avoiding the packet loss and minimization of the latency.

5.4 Discussion on Congestion Control Overhead

Knowing how much overhead, our proposed congestion control management platform adds to the 5G can be of huge importance for its optimal utilization and prevent performance degradation. In this subsection, we discuss and evaluate the overhead of the proposed algorithm protocols.

There are two algorithms implemented in 5G backhaul namely centralized and Distributed: The centralized one is implemented in orchestrator and multi-path handler modules as a Macro-service. The distributed algorithms are implemented in switches. Thanks to software-defined radio based architecture of 5G, distributed algorithms can be easily added to the current platforms. The main part of the distributed algorithm is implemented in SDR switches.

Since there is no communication alighend with algorithms, the congestion control overhead is the communication between these two algorithms. We propose that Congestion control management get implemented via handover and mobility management protocols (S1AP). If the distributed algorithms detect a congestion, then events will be sent to the Centralized algorithm (in Orchestrator and Multi-path Handler modules) via Explicit Congestion Notification (ECN) field of control packets embedded in S1 Interface (within control plane). S1AP is a synchronized protocol in dataplance. In Hand-over, traffic switches between two base stations, the process can be shifted towards the cloud data center or vice versa. Note that in Handover process, the RBS attributed to mobile user changes (path from an RBS to another one) while the path modification (re-routing), the proposed congestion control mechanism is much easier in which, the assigned RBS is the same but another path excluding the congested switch will be selected. Thus, due to the following reasons, the path re-routing is much easier in the centralized algorithm rather than real-time Hand-offs scenarios:

1. Traffic switches at the scale of RBS rather than mobile users. So, all those signaling between mobile users and RBS is not needed.
2. Path selection is not from RBS to RBS, and but for the same RBS different path has to be selected.

Distributed algorithm implementation in RBS and ICN routers is light or not that demanding on resources (there is no communication and consequently no overhead). The only things that have to be done is to compare the PIT queue length in ICN routers with the attributed thresholds. If the measured metrics pass respective threshold, notifications in the form of ECN field will be sent to the orchestrator module. If the distributed process(es) detect a congestion, then events, i.e., notifications, will be sent to the centralized process (in Orchestrator and Multi-path Handler modules) via

ECN field of control packets embedded in S1 Interface (within control plane). For instance, network node Configuration Update/ E-RAB Modify(Request) Procedure and Message can be modified to cover ECN in S1AP protocol. Using ECNs in S1AP makes the congestion control management overhead fixed no matter how much data is communnicated in data plane.

ECN field has two subfields. First subfield dedicated to congestion location (minimum 8-bits) while the second subfield is related to the congestion notification type (2 bits determine whether is it from SW, ICN or RBS). Receiving the active ECN subfield triggers, at the control plane controller, the event handler of centralized process, i.e., state the process. In the detection and congestion control system, described herein, first, the location and type of the congestion are retrieved via two functions namely $Detect_Location()$ and $Detect_ECN_Type()$ which read the associated subfields of ECN. Then, different procedures will be invoked according to detected type and location. If the ECN is generated by either ICN or SW, traffic going through those data plane apparatuses 12 such as switches/ICN routers should be reduced. Two functions namely $Max_TrafficRate(x)$, $Max_InterestRate(x)$ are embedded to find the RBS ID with maximum traffic rate for SW (IP Traffic) ECN type and maximum interest rate for the ICN type going through the location x respectively. After finding the specific RBS identifier (ID), the function $re-route()$ will be performed to find a new path for the calculated RBS in Multi-path Handler, i.e., control plane controller. Algorithms for re-routing and path selection of wireless cellular networks are known in the art. For RBS type ECN subfield, a function is called to find an RBS ID.

Note that there is no need for communication between the centralized algorithm and distributed algorithm, and no execution of the centralized algorithm unless one or more of the distributed algorithms embedded in system or over the 5G network detect congestion. For instance, the state of a switch, i.e., turns to red or moves to the third state in the proposed distributed algorithm. Then, ECN field will be filled in the control plane and be sent to the control plane controller (e.g., multipath handler and orchestrator module) via control plane (RCF). S1AP protocol is used for signaling between centralized and distributed algorithms, i.e., signaling between

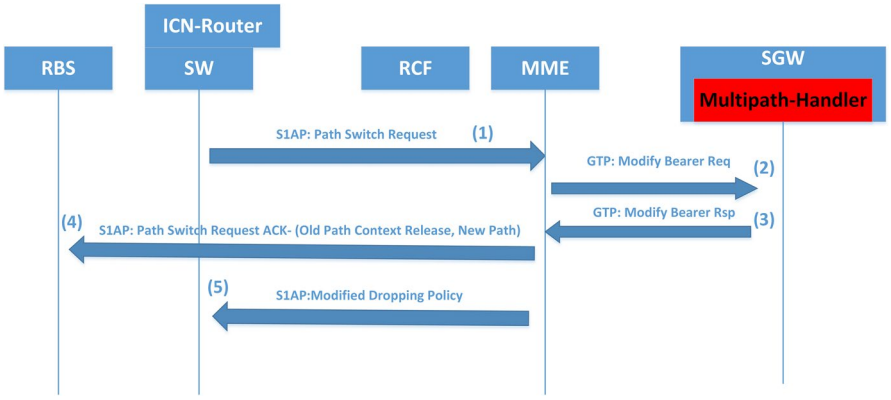


Fig. 15 Congestion control algorithm signaling of the proposed algorithms

data plane and control plane controller. In Fig. 15, the handshaking and communication between the distributed and centralized algorithms are represented. In Fig. 15, if the distributed algorithm in the switch finds the switch state in red, will send a Path Switch request to mobility management entity (MME) which will be forwarded to Multi-path Handler via control plane. ECN field can be embedded in Path Switch Request message between SW and MME and GPRS tunneling protocol (GTP): Modified bearer Request can be used to contain ECN field between MME and Multi-path Handler modules.

6 Conclusion

In this paper, due to the tight delay constraint over the 5G networks, service delay control mechanism to minimize the latency which is of keen interest is addressed. In a light of that, taking advantages of caching in wireless mobile networks, ICN is proposed to avoid entering the multimedia traffic into the 5G backhaul by in-network caching via content-based communication. Moreover, dynamic process management scenario is also strived to minimize the processing bottlenecks. Furthermore, two methods namely distributed and centralized tied with different technologies are introduced to control the latency and congestion in the 5G backhaul network. First, a distributed algorithm running on the switches is proposed to handle the congestion locally and temporarily. Next, a centralized algorithm communicating with the is proposed to avoid congestion through load-balancing, rerouting and leveraging the computational instances dynamically over the 5G mobile network.

References

1. Osseiran, A., Boccardi, F., Braun, V., Kusume, K., Marsch, P., Maternia, M., Queseth, O., Schellmann, M., Schotten, H., Taoka, H., Tullberg, H.: Scenarios for 5G mobile and wireless communications: the vision of the METIS project. IEEE Commun. Mag. **52**(5), 26–35 (2014)
2. Demestichas, P., Georgakopoulos, A.: 5G on the horizon: key challenges for the radio-access network. IEEE Veh. Technol. Mag. **8**(3), 47–53 (2013)
3. Gupta, A., Jha, R.K.: A survey of 5G network: architecture and emerging technologies. IEEE Access **3**, 1206–32 (2015)
4. Carofiglio, G., Gallo, M., Muscariello, L., Perino, D.: Scalable mobile backhauling via information-centric networking. In: The 21st IEEE International Workshop on Local and Metropolitan Area Networks, pp. 1–6 (2015)
5. Salsano, S., Blefari-Melazzi, N., Detti, A., Morabito, G., Veltri, L.: Information centric networking over SDN and OpenFlow: architectural aspects and experiments on the OFELIA testbed. Comput. Netw. **57**(16), 3207–3221 (2013)
6. Woo, S., Jeong, E., Park, S., Lee, J., Ihm, S., Park, K.: Comparison of caching strategies in modern cellular backhaul networks. In: Proceeding of ACM MobiSys (2013)
7. Rodrigues, M., Dn, G., Gallo, M.: Enabling transparent caching in LTE mobile backhaul networks with SDN. In: Proceeding of IEEE Computer Communications Workshops (INFOCOM WKSHPS), pp. 724–729 (2016)
8. Liang, C., Yu, F.R., Zhang, X.: Information-centric network function virtualization over 5G mobile wireless networks. IEEE Netw. **29**(3), 68–74 (2015)

9. Rost, P., Bernardos, C.J., De Domenico, A., Di Girolamo, M., Lalam, M., Maeder, A., Sabella, D., Wbben, D.: Cloud technologies for flexible 5G radio access networks. IEEE Commun. Mag. **52**(5), 68–76 (2014)
10. Checko, A., Christiansen, H.L., Yan, Y., Scolari, L., Kardaras, G., Berger, M.S., Dittmann, L.: Cloud RAN for mobile networksa technology overview. IEEE Commun. Surv. **17**(1), 405–26 (2015)
11. Bastug, E., Bennis, M., Debbah, M.: Living on the edge: the role of proactive caching in 5G wireless networks. IEEE Commun. Mag. **52**(8), 82–89 (2014)
12. Wubben, D., Rost, P., Bartelt, J.S., Lalam, M., Savin, V., Gorgoglione, M., Dekorsy, A., Fettweis, G.: Bene ts and impact of cloud computing on 5G signal process- ing: flexible centralization through cloud-ran. IEEE Signal Process. Mag. **31**(6), 35–44 (2014)
13. Öhlen, P., Skubic, B., Rostami, A., Laraqui, K., Cavaliere, F., Varga, B., Fonseca, N.: Flexibility in 5G transport networks: the key to meeting the demand for connectivity. Ericsson Technol. Rev. **3**, 1–8 (2015)
14. Landstrm, S., Bergstrm, J., Westerberg, E., Hammarwall, D.: NB- IoT: a sustainable technology for connecting billions of devices. Ericsson Technol. Rev. **4**, 211 (2016)
15. Drake, E., Elkhayat, I., Quinet, R., Wenmyr, E., Wu, J.: Paving the way to telco-Grade PAAS. Ericsson Technol. Rev. **6**, 4354 (2016)
16. Pompili, D., Hajisami, A., Tran, T.X.: Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN. IEEE Commun. Mag. **54**(1), 26–32 (2016)
17. Ren, Y., Li, J., Shi, S., Li, L., Wang, G., Zhang, B.: Congestion control in named data networking: a survey. Comput. Commun. **86**, 1–11 (2016)
18. Yi, C., Afanasyev, A., Wang, L., Zhang, B., Zhang, L.: Adaptive forwarding in named data networking. ACM SIGCOMM Comput. Commun. Rev. **42**(3), 6267 (2012)
19. Yi, C., Afanasyev, A., Moiseenko, I., Wang, L., Zhang, B., Zhang, L.: A case for stateful forwarding plane. Comput. Commun. Inf. Centric Netw. **36**(7), 779791 (2013)
20. Wang, X., Chen, M., Taleb, T., Ksentini, A., Leung, V.: Cache in the air: exploiting content caching and delivery techniques for 5G systems. IEEE Commun. Mag. **52**(2), 1–18 (2014)
21. Braun, S., Monti, M., Sifalakis, M., Tschudin, C.: An empirical study of receiver-based AIMD flow-control strategies for CCN. In: Proceeding of IEEE International Conference on Computer Communication and Networks (ICCCN) (2013)
22. Braun, S., Monti, M., Sifalakis, M., Tschudin, C.: TCP co-existence in the future internet: should CCN be compatible to TCP? In: 2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013) May 27, pp. 1109–1115. IEEE (2013)
23. Dukkipati, N.: Rate control protocol (rcp): congestion control to make flows complete quickly, Ph.D. Dissertation, Stanford University, Stanford, CA, USA: ford, p. 2008. CA, USA (2008)
24. Yang, C.Q., Reddy, A.V.: A taxonomy for congestion control algorithms in packet switching networks. IEEE Netw. **9**(4), 34–45 (1995)
25. Carofiglio, G., Gallo, M., Muscariello, L.: ICP: design and evaluation of an interest control protocol for content-centric networking. In: Proceeding of IEEE INFOCOM Workshop on Emerging Design Choices In Name Oriented Networking (INFO-COM NOMEN) (2012)
26. Salsano, S., Detti, A., Cancellieri, M., Pomposini, M., Blefari-Melazzi, N.: Receiver driven interest control protocol for content-centric networks. In: Proceeding of ACM SIGCOMM Workshop on Information Centric Networking (ICN) (2012)
27. Amadeo, M., Molinaro, A., Campolo, C., Sifalakis, M.: Transport layer design for named data wireless networking. In: Proceeding of IEEE INFOCOM Workshop on Name-Oriented Mobility (2014)
28. Arianfar, S., Nikander, P., Eggert, L., Ott, J.: Contug: a receiver driven transport protocol for content-centric networks. In: Proceeding of IEEE ICNP (2010)
29. Saino, L., Cocora, C., Pavlou, G.: CCTCP: a scalable receiver-driven congestion control protocol for content centric networking. In: Proceeding of IEEE ICC13 (2013)
30. Carofiglio, G., Gallo, M., Muscariello, L., Papalini, M.: Multipath congestion control in content-centric networks. In: Proceeding of IEEE INFOCOM 2013 Workshop on Emerging Design Choices in Name-Oriented Networking (2013)
31. Fu, T., Li, Y., Lin, T., Tan, H., Tang, H., Ci, S.: An effective congestion control scheme in content-centric networking. In: Proceeding of 13th International Conference on Parallel and Distributed Computing, Applications and Technologies (2012)
32. Ren, Y., Li, J., Shi, S., Li, L., Chang, X.: An interest control protocol for named data networking based on explicit feedback. In: Proceeding of the 11th ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS 15) (2015)

33. Zhou, J., Wu, Q., Li, Z., Kaafar, M.A., Xie, G.: A proactive transport mechanism with explicit congestion notification for NDN. In: Proceeding of ICC (2015)
34. Rozhnova, N., Fdida, S.: An effective hop-by-hop interest shaping mechanism for CCN communications. In: Proceeding of IEEE INFOCOM NOMEN Workshop (2012)
35. Wang, Y., Rozhnova, N., Narayanan, A., Oran, D., Rhee, I.: An improved hop-by-hop interest shaper for congestion control in named data networking. In: Proceeding of ICN13 (2013)
36. Carofiglio, G., Gallo, M., Muscariello, L.: Joint hop-by-hop and receiver-driven interest control protocol for content-centric networks. In: Proceeding of ACM SIGCOMM Workshop on Information Centric Networking (ICN) (2012)
37. Zhang, F., Xu, C., Zhang, Y., Reznik, A., Liu, H., Qian, C.: A transport protocol for content-centric networking with explicit congestion control. In: Proceeding of the 23rd International Conference on Computer Communications and Networks (ICCCN) (2014)
38. Zhang, F., Zhang, Y., Reznik, A., Liuc, H., Qiand, C., Xue, C.: Providing explicit congestion control and multi-homing support for content-centric networking transport. Comput. Commun. **69**, 6978 (2015)
39. Matsuzono, K., Asaeda, H.: NRTS: content name-based real-time streaming. In: Proceeding of 13th IEEE Annual Consumer Communications and Networking Conference (CCNC) pp. 537–543 (2016)
40. Matsuzono, K., Asaeda, H.: NMRTS: content name-based mobile realtime streaming. IEEE Commun. Mag. **54**(8), 92–8 (2016)
41. Carofiglio, G., Gallo, M., Muscariello, L.: Optimal multipath congestion control and request forwarding in information-centric networks: protocol design and experimentation. Comput. Netw. **110**, 10417 (2016)
42. Mahdian, M., Arianfar, S., Gibson, J., Oran, D.: MIRCC: multipath-aware ICN rate-based congestion control. In: Proceedings of the 2016 Conference on 3rd ACM Conference on Information-Centric Networking, pp. 1–10 (2016)
43. Chen, J., Arumaithurai, M., Fu, X., X., Ramakrishnan, K.K.: SAID: A Control Protocol for Scalable and Adaptive Information Dissemination in ICN. arXiv:1510.08530 (2015)
44. Schneider, K., Yi, C., Zhang, B., Zhang, L.: A practical congestion control scheme for named data networking. In: Proceedings of the 2016 Conference on 3rd ACM Conference on Information-Centric Networking, pp. 21–30 (2016)
45. Raina, G., Towsley, D., Wischik, D.: Part II: control theory for buffer sizing. ACM SIGCOMM Comput. Commun. Rev. **35**(3), 79–82 (2005)
46. Bhattacharyya, S., Towsley, D., Kurose, J.: The loss path multiplicity problem in multicast congestion control. Proc. IEEE INFOCOM **2**, 856–63 (1999)
47. Nichols, K., Jacobson, V., McGregor, A., Iyengar, J.: Controlled delay active queue management, RFC 8289. http://rfc-editor.org/rfc/rfc8289.txt (2016)
48. Duffield, N., Lund, C., Thorup, M.: Properties and prediction of flow statistics from sampled packet streams. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurement, pp. 159–171 (2002)
49. Duffield, N., Lund, C., Thorup, M.: Estimating flow distributions from sampled flow statistics. IEEE/ACM Trans. Netw. (TON) **13**(5), 933–946 (2005)
50. Cheng, G.: Estimating the number of active flows from sampled packets. In: IEEE Network Operations and Management Symposium, pp. 675–678 (2012)
51. Hu, C., Liu, B., Zhao, H., Chen, K., Chen, Y., Cheng, Y., Wu, H.: Discount counting for fast flow statistics on flow size and flow volume. IEEE/ACM Trans. Netw. (TON) **22**(3), 97081 (2014)
52. Choi, W., Seok, W.: Time-based forwarding control in content centric networking. In: Proceeding of 18th International Conference on Advanced Communication Technology (ICACT), pp. 643–646 (2016)

Shahin Vakilinia received the B.Sc. degree in electrical engineering from the University of Tabriz, Tabriz, Iran, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2008 and 2010, respectively, and the Ph.D. degree from the Department of Electrical and Computer

Engineering, Concordia University, Montreal, QC, Canada, in 2015. He is currently involved in research with Ericsson.

Halima Elbiaze (M13) received the B.S. degree in applied mathematics from the University of Rabat, Rabat, Morocco, in 1996 and the M.S. and Ph.D. degrees in computer science from the University of Versailles, Versailles, France, in 1998 and 2002, respectively. She is currently a Professor with the Department of Computer Science, University of Quebec in Montreal, Montreal, QC, Canada. Her research interests are in the areas of quality of service, performance evaluation, and traffic engineering for high-speed networks (IP/Wave Division Multiplexing, Transmission Control Protocol/IP, Asynchronous Transfer Mode, Frame Relay, etc.).