Otero Pérez, Gabriel; Hernández, José Alberto; Larrabeiti, David. Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G, in: IEEE/OSA *Journal of optical communications and networking,* Vol. 10, No. 6 (2018), pp. 573-581.
DOI: https://doi.org/10.1364/JOCN.10.000573

Fronthaul Network Modeling and Dimensioning Meeting Ultra-low Latency Requirements for 5G

Gabriel Otero Pérez, José Alberto Hernández and David Larrabeiti

***Abstract*—Enabling the transport of fronthaul traffic in next generation cellular networks (5G) following the *Cloud Radio Access Network* (C-RAN) architecture requires re-designing the fronthaul network featuring high-capacity and ultra-low latency. With the aim of leveraging statistical multiplexing gains, infrastructure reuse and, ultimately, cost reduction, the research community are focusing on Ethernet-based packet-switch networks. To this end, we propose to use high queueing delay percentiles of G/G/1 queueing model as the key metric in fronthaul network dimensioning. Simulations reveal that that the *Kingman's Exponential Law of Congestion* provides accurate estimates on such delays for the particular case of aggregating a number of eCPRI fronthaul flows, namely functional splits I_U and II_D. We conclude that conventional 10G, 40G and 100G transponders can cope with multiple legacy 10-20 MHz radio channels with worst-case delay guarantees. Conversely, scaling to 40 and 100 MHz channels will require the introduction of 200G, 400G and even 1T high-speed transponders.**

***Index Terms*—5G; C-RAN; Fronthaul Networks; eCPRI; G/G/1; Kingman's exponential law of congestion; delay percentiles.**

I. Introduction

THE Cloud Radio Access Network (C-RAN) architecture proposed for the 5-th Generation (5G) Mobile Networks introduces the concept of a cloud computing-based processing of radio signals. This has shown important Capital and Operation Expenditure (CAPEX/OPEX) savings to the network operator, while enhancing the cellular network's effective capacity by means of load balancing and combined processing of radio signals coming from several closely located base stations [1], [2].

In C-RAN, lightweight Remote Radio Heads (RRHs) perform simple operations on the radio signal and forward it towards the remotely-located Baseband Units (BBUs) through the so-called front-haul (FH) network. Conversely, the BBUs are in charge of synthesising the radio signal that will be sent to the RRHs. Up to the date, most C-RAN implementations for LTE use the CPRI (Common Public Radio Interface) specification [3]. However the stringent transmission requirements of this Digital Radio system initially designed for intra-base-station communication have pushed more efficient schemes. Different functional splits of traditional Base Stations have been defined in the literature (up to eight, see [4], [5]) depending on which radio processing operations are kept at the Distributed Unit (DU) or RRH, and which operations are moved to the cloud or Centralised Unit (CU). At present, the preferred functional splits under study are: (1) Option 8, also called PHY-RF split or CPRI-like where the In-Phase and Quadrature (IQ) radio symbols are sampled, quantized and transmitted [6]; (2) Option 7 or Intra-PHY functional split where some radio operations are performed at the RRH before its transmission; (3) option 6 or MAC-PHY split where RF and physical layer operations are kept in the DU; and (4) option 2 or PDCP/RLC split where Packet Data Convergence Protocol (PDCP) functionality are in the CU while Radio Link Control (RLC), MAC, physical layer and RF are in the DU. Options 7 and 8 are user load independent and require high-capacity low-latency links between RRHs and BBUs [6], [7].

In this light, a number of standardisation bodies are also in the process of defining how to implement the C-RAN concept in a packet-based transport network like Ethernet, thus leveraging the high penetration of low-cost Ethernet hardware along with the statistical multiplexing gains offered by packet-switched networks. As a matter of fact, the IEEE Next-Generation Fronthaul Interface (NGFI) working group are in the process of defining the architecture of fronthaul transport networks and the mechanisms to both encapsulate and map such front-haul traffic into Ethernet packets [8]. The IEEE Time-Sensitive Networking (TSN) working group are also in the process of defining how to treat Ethernet packets carrying front-haul traffic generated by RRHs in an Ethernet-bridged network, including aspects like frame preemption, scheduled traffic support, and path control or reservation (standards IEEE 802.1Qbu, 802.1Qbv, 802.1Qca respectively) [9]. Other standardisation entities like the ITU-T G.Sup 56 proposes mechanisms for mapping CPRI client signals in ODUflex containers and treat them as conventional Constant Bit Rate (CBR) traffic across an G.709 Optical Transport Network (OTN) [10]. Finally, the IETF has also launched a Working Group on Deterministic Networking (IETF DetNet) to explore how to engineer "...deterministic data paths that operate over Layer 2 bridged and Layer 3 routed segments, where such paths can provide bounds on latency, loss, and packet delay variation (jitter), and high reliability" [11].

Fronthaul traffic, in particular CPRI-like traffic, has strict latency requirements while Ethernet switches are typically best effort. Earlier studies [12] showed that 10Gb Ethernet alone with or without frame preemption could not meet the CPRI traffic's jitter requirements ($65\,ns$) and hence, buffering is required, contributing to the effective latency. The envisioned user plane end-to-end latency for 5G varies depending on the type of application. Most will require the latency to be confined below a few milliseconds [13], [14], e.g., tactile internet, factory automation ($\leq 1\ ms$), intelligent transportation systems ($5\ ms$), etc. The 3GPP [15] also defines different latency profiles: Ultra-Reliable and Low Latency Communications (URLLC) ($\leq 0.5\ ms$), enhanced Mobile BroadBand

The authors are with the Department of Telematics Engineering, Universidad Carlos III de Madrid, Avda. Universidad, 30, 28911 Madrid, Spain. Email:{gaoterop@it.uc3m.es, jahgutie@it.uc3m.es, dlarra@it.uc3m.es}

(eMBB) ($\leq$ 4 ms), etc. This translates into even more strict requirements at lower layers.

Regarding the transport plane, a new end-to-end Ethernet network latency target budget of 100 μs for CPRI traffic was established in 802.1CM [9] which is a useful design parameter, and shows the importance of characterising the queuing delay through the network. The lower the queuing delay, the higher the budget for propagation delay and fabric switching delay. Additional latency budget may come from the fact that higher functional splits (like the Intra-PHY or MAC-PHY cases) have relaxed delay, jitter and synchronisation requirements [16]. Therefore, these splits could be carried across conventional packet-based transport networks. 802.1CM specifies strict priority queuing discipline as the way to achieve minimum latency. Special encapsulation mechanisms and scheduling policies have been proposed for the transport of CPRI over Ethernet in [17], [18].

In summary, the research community is engaged in addressing the challenging aspects regarding the design of feasible fronthaul and backhaul networks for 5G C-RAN scenarios. They have concluded that (1) fronthaul traffic should be packetised and transmitted across conventional packet-switched networks and, (2) higher functional splits than CPRI need to be considered for fronthaul traffic, given the excessive network requirements of CPRI transport (i.e. very high-capacity and ultra-low-latency pipes). Regarding the latter, the industry cooperation (i.e. NEC, Nokia, Huawei and Ericsson) involved in the specification of CPRI has recently released an *evolved* CPRI (eCPRI) specification [20] for the abovementioned functional splits, namely Splits **E**, $\mathbf{I}_{\mathrm{U}}$, $\mathbf{I}_{\mathrm{D}}$ and $\mathbf{II}_{\mathrm{D}}$ (see Fig. 2). This new version of CPRI is designed for packet-switched transport of radio signals. Therefore, we shall focus our study on this specification. It is worth noting that these new eCPRI functional splits concern the division of the processing chain inside the physical (PHY) layer, that is, they issue the partitioning of the processing operations from 3GPP option 6 (MAC-PHY) and upwards [4].

The aggregation and transmission of a large number of front-haul eCPRI flows in a cost-effective way using a packet-switched network while meeting ultra-low latency requirements is the upcoming challenge. We identify three enabling technologies:

- Consolidated high-capacity optical transponders using coherent modulation formats, namely 40G, 100G and 200G.
- The upcoming Nyquist-spaced WDM transmission systems [21], [22] with Information Spectral Density (ISD) values between 3 and 8 b/s/Hz, and the foreseen 400G and 1T [23]–[27]
- *Sliceable Bandwidth Variable Transceivers* (S-BVT) [29], [30].

A number of European H2020 research projects are studying the implementation of 1T S-BVT aggregation nodes that allow to obtain statistical multiplexing gain at the optical layer, in addition to the multiplexing gain yielded by the Ethernet transport technology addressed in this paper.

Having all the above-mentioned in mind as the starting point, we perform a number of theoretical and simulation tasks: (1) studying the properties of the aggregated packetised FH flows. We derive a set of rules for Ethernet-based FH network dimensioning, using high delay percentiles as the key design metric (instead of conventional average delays often seen in the literature); (2) we provide a model based on the Kingman's exponential law of congestion for G/G/1 queues as a means of performing network planning and dimensioning; and (3) we show the limits in terms of the number of FH flows that can be transported, derived from next-generation radio signals (40-100 MHz LTE radio channels) on 100G and 200G transceivers, and the upcoming 400G and 1T estimated to be available in 2018 (see Table II, II and V of [28], [29]).

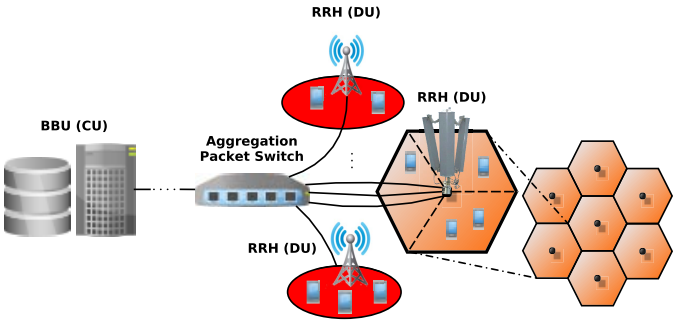


Fig. 1. Case of study of a C-RAN FH network topology.

The rest of this article is organised as follows: Section II outlines the CPRI and Intra-PHY functional splits and their traffic profiles. Section III reviews classical queuing theory results, focusing on high queueing delay percentiles as worst-case delay requirements. Section IV shows a number of simulations that validate the derived equations, and provides a set of rules for the dimensioning of fronthaul networks. Finally, Section V concludes this article with a summary of its main contributions.

II. Fronthaul Traffic Characteristics

A common practice in future cellular deployments and, in particular, in C-RAN scenarios, is to aggregate the traffic of multiple sectors or cells so as to facilitate the transport process and leverage the statistical multiplexing gains of packet-switched networks. Fig. 1 plots an overview of a general C-RAN FH network topology. For the sake of the example, note that sectors from both multi (three) and single sector cells are aggregated into a packet switch making use of optical links. Each sector produces one FH flow that must be transported to the centralised processing units (BBUs). Thus, we are interested in the statistical properties and characterisation of the queueing delay affecting a mix of multiple FH flows coming from diverse sectors or cells.

We confine this study to the uplink, since the needed processing for it poses more stringent delay requirements than that of the downlink [31]. As noted in Fig. 2, Functional Split **E** consists of transmitting the pure sampled signal, that is, the time-domain radio waveform downconverted to baseband frequency, sampled and then quantised [6]. Since no further processing is performed at the RRH, overhead information such

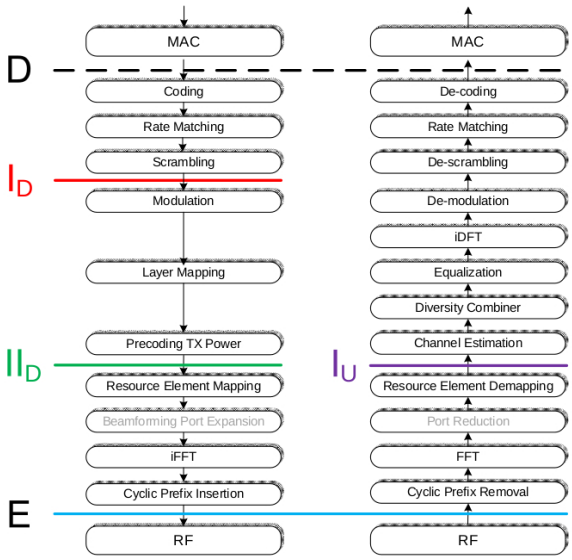


Fig. 2. eCPRI vision of 5G processing chain (see [20])

as the Cyclic Prefix (CP) is transmitted towards the BBU. The resulting data rate for Split **E** can be written as

$$R_{\text{Split E}} = f_s \cdot 2 \cdot N_{bits} \cdot N_{ant} \tag{1}$$

where N_{bits} and N_{ant} represent the bit resolution used to quantise the signal samples and the number of receiving antennae, respectively. The 2-factor refers to the complex nature of signals and f_s represents the sampling frequency.

A first step to relax the bandwidth burden is switching to frequency domain (Split $\mathbf{I_U}$). The received radio signals are transformed to baseband and then applied to an analog-to-digital converter (ADC). The serial ADC output is converted to parallel and the cyclic prefix can be removed. At this moment, an N-point FFT may be used to decode the orthogonal subcarriers (N_{sc}). Those subcarriers used as a guard band, typically 10% [19], are no longer necessary. Then, according to the processing chain envisioned in the eCPRI specification [20], all the resource blocks (RB) can be demapped. Consequently, the resulting data rate now depends on the fraction (η) of RB under use. Then, the resulting data rate can be expressed as

$$R_{\text{Split } \mathbf{I_U}} = N_{\text{sc}} \cdot 0.9 \cdot (T_s)^{-1} \cdot \eta \cdot 2 \cdot N_{bits} \cdot N_{ant} \tag{2}$$

where T_s is the symbol duration time.

Numerical example: Let us think about a 2x2 MIMO 20 MHz LTE channel, with 15 KHz subcarrier spacing. A total of 20 MHz/15 KHz $= 1,333$ subcarriers can be allocated. Assuming a worst case scenario where all the RB are used ($\eta = 1$), a symbol rate of $T_s = 66.\overline{6}\,\mu s$ to maintain orthogonality and 15 bits per sample, the generated rate is

$$R_{\mathbf{I_U}} = \frac{20\,\text{MHz}}{15\ \text{KHz}} \cdot 0.9 \cdot (66.\overline{6}\,\mu s)^{-1} \cdot 2 \cdot 15 \cdot 2 = 1,080\ \text{Mbit/s}/s \tag{3}$$

In comparison, Split **E** data rate under the same conditions and using $f_s = 30.72\,\text{MHz}$ generates $1,843.2$ Mbit/s. It is worth noting that such data flow carries no more than 150 Mb/s peak-rate of user data (see Fig. 5 of [32]), which makes Split **E** rate more than 10 times the associated user data rate. Consequently, transporting a bunch of Split **E** flows would require dedicated high capacity links which therefore hampers the aggregation and switching of FH traffic. Clearly, the bitrate required by Split $\mathbf{I_U}$ is quite more moderate, about one half of the Split **E**. Notice that the downlink Split $\mathbf{II_D}$ is at the same level that the uplink Split $\mathbf{I_U}$ (see Fig. 2) as they break the processing chain at the same point. Therefore, data rates of both splits are equivalent. Detailed investigations and variations of the above-mentioned splits have been conducted in [20] and [31].

It is important to note that, despite the fact that our goal is to use Ethernet transport, at this point there are a lot of functions that still need to be performed, e.g, channel estimation, demodulation, Forward Error Correction (FEC), etc. For the uplink Split $\mathbf{I_U}$ case, all the operations above the purple line in Fig. 2 are perfomed at the centralised BBU. After that, data are recovered from the symbols and all the redundant information is removed. As a result, we have the pure medium access control (MAC) payload at the output (black dashed line), leaving the physical (PHY) layer and entering the MAC layer (see p. 11 of [4]). Among the functions present in this layer, it is worth highlighting the Hybrid Automatic Repeat Request (HARQ) error-control protocol.

Table I shows a summary of the resulting FH traffic profiles of different LTE channel bandwidths for splits **E** and $\mathbf{I_U}$. The way to read this table is as follows: first column refers to the transmission of time-domain samples of LTE channels using Split **E** (CPRI-like), assuming a 2 antennae system, 15 bit for quantization and 15 KHz subcarrier spacing. Both 10 MHz and 20 MHz channels are considered. Regarding the first one, the traffic profile is halved since the channel bandwidth is half. The reader is refered to [3], [6] for an overview of CPRI features and bandwidth transmission requirements. The second column focuses on split $\mathbf{I_U}$ requirements under the same assumptions. We illustrate the generated data rates considering different channel bandwidths ranging from 10 MHz to future 100 MHz LTE channel bandwidths envisioned by 3GPP [4]. Bear in mind that the bitrates are considerably lower compared to those of split **E** for the same channel bandwidth and that we achieve this reduction at the expense of increasing the computational complexity at the RRH side. Conversely, **Split** $\mathbf{I_U}$ data rate depends on the part of the resource blocks that are actually utilised by the user equipment in a cell. Only these remain after the RB demapping and are forwarded to the processing units, which enables **Split** $\mathbf{I_U}$ to profit from load balance gains.

Finally, it is important to bear in mind that both Splits **E** and $\mathbf{I_U}$ produce a Constant Bit Rate (CBR) stream of packets with different burst size and periods (see Table I). For example, the transmission of a FH flow carrying a 2x2 MIMO 20 MHz LTE channel using split **E** (third column in Table I) comprises the periodic transmission of one 60 Byte packet every $260.41416\overline{6}\ ns$. On the contrary, the same LTE channel using Split $\mathbf{I_U}$ (fifth column in Table I) comprises the periodic transmission of 9000 Bytes (i.e. six 1500 B packets) every $66.\overline{6}\ \mu s$.

	Split E		Split $\mathbf{I}_{\mathrm{U}}$ ($\eta = 1$)			
Channel bandwidth	10 MHz	20 MHz	10 MHz	20 MHz	40 MHz	100 MHz
Burst size [B]	30	60	4500	9000	18000	45000
Period [μs]	$0.2604141\overline{6}$		$66.\overline{6}$			
Bitrate [Mb/s]	921.6	1843.2	540	1080	2160	5400

TABLE I
TRAFFIC PROFILES FOR DIFFERENT FUNCTIONAL SPLITS, $N_{ant} = 2$ MIMO, $N_{bit} = 15$ BIT/SAMPLE, 15 KHZ SPACING

III. QUEUEING THEORY REVIEW

A. G/G/1 queueing model: Kingman's exponential law of congestion

The analytical study of the aggregate of a significant number of FH flows for this particular split ($\mathbf{I}_{\mathrm{U}}$), requires the use of appropriate queuing models. The reason is that, despite this functional split reduces the bandwidth requirements, there is still a strict latency requirement imposed by the HARQ protocol, in charge of the error correction process. Therefore, a deep characterisation of the queueing delay is paramount to ensure that the latency requirements are met.

The well-known M/M/1 model is attractive since it provides closed expressions for the main metrics of interest. However, it assumes exponentially distributed interarrival times which do not apply to fronthaul traffic [33]. Since the interarrival time distribution of the aggregation of Split $\mathbf{I}_{\mathrm{U}}$ FH flows is unknown and dependent on the specific number of flows (see Fig. 4 in Section IV), we make use of a generalised queueing model (G/G/1) that enables us to characterise the behaviour of the system under different conditions by tweaking coefficient of variation of arrival times of packets at the switch queue.

Contrary to what we stated for the M/M/1 model, no closed expressions exist for the mean waiting time in queue under these assumptions. A packet switch modeled using a G/G/1 model considers that packet arrivals follow a general (G) (arbitrary) distribution with rate λ packet/s. All arrivals compete for a single resource and are temporally stored (buffered or queued) in an FCFS discipline, experiencing a service time of $E[\mathbb{S}] = \frac{1}{\mu}$ seconds, whose distribution is again general. We require the load of the system $\rho = \lambda \cdot E[\mathbb{S}] < 1$, for stability. Define the squared coefficient of variation of a random variable $\mathbb{X}$ as $C^2[\mathbb{X}] = \frac{\mathrm{Var}[\mathbb{X}]}{E[\mathbb{X}]^2}$. Let $\mathbb{T}$ be the random variable modeling the interarrival times of packets at the queue and $\mathbb{S}$, the service time random variable. Then, the queueing delay $\mathbb{W}_q$ gets its mean from the Allen-Cunneen approximation [34]:

$$E[\mathbb{W}_q] \leq E[\mathbb{S}] \cdot \frac{\rho}{1-\rho} \cdot \frac{C^2[\mathbb{T}] + C^2[\mathbb{S}]}{2} \tag{4}$$

which extends the M/M/1 equations with the so called *stochastic variability* term $\frac{C^2[\mathbb{T}]+C^2[\mathbb{S}]}{2}$. According to [39], the Kingman's formula is a *very good approximation* to the mean queue waiting time which works well under most conditions, particularly, when $\rho \to 1$. It is worth remarking that exponentially distributed service and inter-arrival times have $C^2[\mathbb{T}] = 1 = C^2[\mathbb{S}]$, hence the stochastic variability term in the M/M/1 case is equal to one and the Kingman's formula is exact. Formally, the Kingsman's Exponential Law of Congestion can be expressed by the *congestion index* as follows

$$\frac{\mathbb{W}_q}{E[\mathbb{S}]} \simeq \begin{cases} \exp(\mathrm{mean} = \frac{1}{1-\rho}\frac{C^2[\mathbb{T}]+C^2[\mathbb{S}]}{2}), & \text{wp } \rho \\ 0, & \text{with probability } (1-\rho) \end{cases}. \tag{5}$$

From here, we may compute the p-th percentile delay as

$$p = \int_{t=-\infty}^{\mathbb{W}_q^{(p)}} (1-\rho)\cdot\delta(t) + \rho \cdot \omega e^{\omega t}\,\mathrm{H}(t)\,dt \tag{6}$$

where $\frac{1}{\omega} = \frac{1}{1-\rho} \cdot \frac{C^2[\mathbb{T}]+C^2[\mathbb{S}]}{2}$ and $\delta(t)$ and $\mathrm{H}(t)$ are the *delta* and *Heaviside step* functions, respectively. These are the indicator functions of the intended supports. Solving for the p-th percentile delay $\mathbb{W}_q^{(p)}$, we obtain

$$\mathbb{W}_q^{(p)} = \max\left\{0, E[\mathbb{S}]\frac{1}{1-\rho}\frac{C^2[\mathbb{T}]+C^2[\mathbb{S}]}{2}\ln\left(\frac{\rho}{1-p}\right)\right\}. \tag{7}$$

Numerical example: Consider the output port of a packet switch operating at 10 Gb/s and $\rho = 70\%$ load. Now, assume that packet arrivals consist of a mix of 6 uncorrelated Split $\mathbf{I}_{\mathrm{U}}$ flows, for a 20 MHz channel configuration. Encapsulate the data using 1500B payload length. In this case, $\rho = \frac{6\cdot 1080\ Mb/s}{10\ Gb/s} \simeq 0.7$, $C^2[\mathbb{T}] \simeq 8.8$ (see Fig. 4) and $C^2[\mathbb{S}] = 0$ since packet service time is constant. The average service time can be obtained as follows

$$E[\mathbb{S}] = \frac{8 \cdot 1500\ b}{10 \cdot 10^9\ b/s} = 1.2\ \mu s.$$

Then, the 90-th percentile queueing delay is

$$\mathbb{W}_q^{(0.9)} = 1.2\ \mu s \frac{8.8}{2\cdot(1-0.7)}\ln\left(\frac{0.7}{1-0.9}\right) = 34.25\ \mu s \tag{8}$$

which is roughly 30 times the average service time, due to the high value of $C^2[\mathbb{T}]$. For the sake of comparison, for the M/M/1 model $C^2[\mathbb{T}] = 1 = C^2[\mathbb{S}]$ and $\mathbb{W}_q^{(0.9)} = 7.78\ \mu s$ under the same conditions. It is important to note that the G/G/1 model delay percentile estimation is around 4.5 times larger than that of M/M/1 model. The reason for this is presented in Section IV-B and G/G/1 model delay percentile estimations are verified by means of simulation in Section IV-C.

IV. SIMULATION RESULTS

Once that the bandwidth requirements are analysed in Section II, we focus our study on Split $\mathbf{I}_{\mathrm{U}}$. Particularly, we study the behaviour of multiple packetised $\mathbf{I}_{\mathrm{U}}$ fronthaul flows converging into the same switching element. We assume that the header of each packet comprises

- **Preamble**: 8 bytes.
- **Ethernet Header**: 14 bytes.

- **Interpacket Gap (IPG)**: 12 bytes.
- **CRC**: 4 bytes.
- **eCPRI Header**: 4 bytes.

The efficiency of the packetisation scheme in terms of aggregated queueing delay depends on the number of packets we choose to transport each burst [33]. With the aim of minimising the overheads, we set the payload size to the Maximum Transfer Unit (MTU) of Ethernet, i.e. 1500 bytes.

A. Discrete-Event Simulator

We implemented a custom Discrete-Event Simulator so as to assess the validity of the theoretical approximations of Eq. 7 as well as to unveil the behaviour and properties of frontahaul traffic under different conditions. Fig. 3 shows the output of the simulator for different number of FH flows, considering an aggregation node with a 100 Gb/s upstream link. It is worth highlighting that the worst case queueing delay values, measured as 90-th and 99-th percentile values, are much higher than the average. For instance, the aggregation of 140 fronthaul flows results in an average queueing delay of $0.7\,\mu s$, for the packets arriving at the aggregation switch. However, the 99-th percentile is approximately 3.4 times higher ($2.4\mu s$).

B. Effects of the aggregation of fronthaul flows

The aim of this experiment is to determine the steady state convergence of the arrival squared coefficient of variation $C^2[\mathbb{T}]$, when many FH flows are aggregated. Flows are merged applying a uniformly distributed offset to each one of them, $U(0, T_p)$, between 0 and the burst period $T_p = 66.\overline{6}\ \mu s$. As a worst-case scenario we assume that each eCPRI RRH generates a burst of back-to-back packets on each symbol period. We test the evolution of $C^2[\mathbb{T}]$ for different values of the channel bandwidth (remark that channel bandwidths refer to different burst sizes as noted in Table I).

As shown in Fig. 4, the squared coefficient of variation of the packet arrivals converges to unity as we increase the number of mixed flows, showing a Poisson-like traffic profile. This behaviour is explained by the Palm-Khintchine theorem [35]. It is rather important to stress that the rate of convergence to steady state is different depending on the size of the burst (i.e, the channel bandwidth). Also, note that the wider the channel bandwidth, the faster the coefficient of arrival converges to unity, as a consequence of longer 1500 B packet bursts, thus blurring the periodic CBR structure of the FH flows sooner. The distribution of the interarrival times does not show a poissonian behaviour until we merge, approximately, more than 450 independent FH flows. Hence, we cannot assume the M/M/1 nor the M/G/1 model are good approximations for all load conditions and this is the reason why the G/G/1 model is chosen.

C. Accuracy of the theoretical estimations

In this subsection, we assess the validity of the Kingman's exponential law model by comparing the estimations of Eq. 5 with the simulation ouputs. To do so, we estimate the theoretical p-th percentile delay by substituting simulated values (see

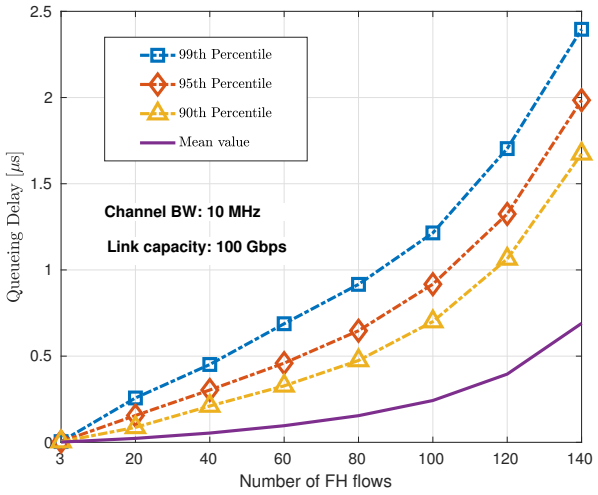


Fig. 3. Queueing delay statistics.

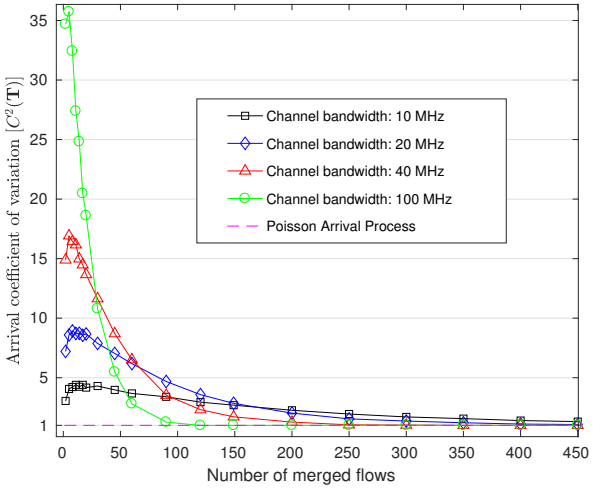
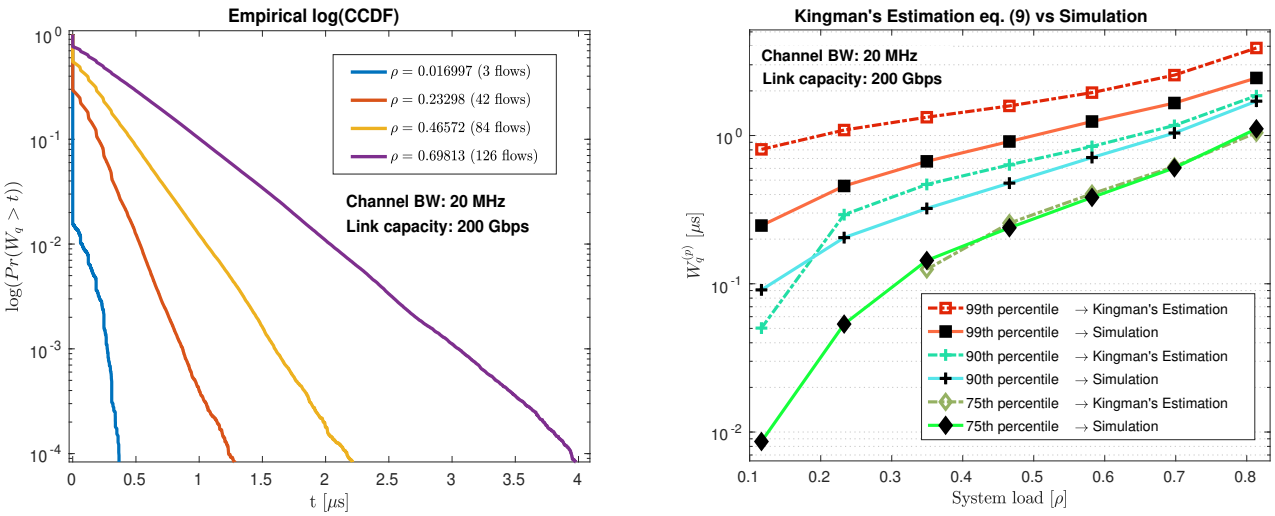


Fig. 4. Arrivals squared coefficient of variation.

Fig. 4) of the system load (ρ) and traffic variability ($C^2(\mathbb{T})$) into Eq. 5.

Figure 5a plots the experimental complementary cumulative distribution function (ECCDF) for the queueing delay in a traffic aggregation node using a 200G transceiver, for 20 MHz channels under different load conditions. Close inspection of the figure reveals that the simulated waiting time in queue follows a mixture distribution with some probability mass located at *zero* queueing delay. Regarding the rest of the function support, it is exponentially distributed (note the straight lines in logarithmic scale) once that $Pr(\mathbb{W}_q > t) = \rho$. Notice that this is the expected behaviour in view of Eq. 5.

Figure 5b illustrates the evolution of 75-th, 90-th and 99-th percentiles, as the traffic load of the node increases by aggregating more and more fronthaul flows, for 20 MHz channels. 99% condidence intervals are used for simulation



(a) Experimental CCDF

(b) Kingman estimation vs Simulation

Fig. 5. Experiments results for the transmission of 20 MHz LTE channels using 200 Gb/s transceivers.

outputs. However, they are extremely tight and negligible in comparison with the observed magnitudes, therefore, we omit them in the plots. Missing 75-th percentile values for the Kingman's estimation imply that $W_q^{(p)} = 0\ \mu s$ for those particular system loads. Note that the *Kingman's Exponential Law of Congestion* (Eq. 7) is, in general, an upper bound on the p-th percentile delay. It is worth to mention that the gap between the theoretical estimation and the simulation outputs becomes slighty narrower as we aggregate more and more traffic and the load increases. For example, the 99th percentile estimation-simulation ratio is around 3 under low load conditions. Converesely, the ratio between theoretical and simulated values in heavy load conditions, deacreases to 1.5. Nevertheless, the reader should have in mind that the absolute difference between simulation and estimation is below 1 μs at 200 Gb/s. Finally, the figure shows that the higher the percentile we try to estimate, the bigger is the gap between the analytical and experimental values.

D. Dimensioning rules

Once we know that the G/G/1 model is able to provide good estimations on the worst-case delays, we aim at the fronthaul network dimensioning. To that end, we take into account extreme queueing delay percentiles. Consider again a switch aggregating a number of fronthaul flows, each generated by a sectorial antenna in a cellular topology. Assume each antenna sector generates one split $\mathbf{I}_U$ fronthaul flow. Table II shows the mean queueing delay values, as well as 90th and 99th percentiles, considering different output link capacities of the aggregation switch. A wide variety of scenarios has been tested, from current 10 MHz - 40 MHz to future 100 MHz LTE channels scenarios. Recall that $\mathbf{I}_U$ split data rate for each fronthaul flow ranges from 540 Mb/s to 5400 Mb/s depending on the channel bandwidth we use (see Table I). Then, we compute Kingman's estimation for different link rates of the aggregation point's output link.

Note that the table implements a color code as follows: table cells with a red background represent unfeasible scenarios where system load exceeds 100% ($\rho > 1$). Those scenarios whose 99-th delay percentiles are below 5 μs are highlighted in green and the remaining cases in between are shaded in yellow. After analysing the table results, it seems clear that 10G links pose severe limits to the number of FH flows that we can carry for 10 MHz channels. 40G links provide enough throughput to transport up to about 60 sectors. 100G and 200G can deal with 140 flows, while only 200G can deal with such an amount of sectors for 20 MHz (100G can transport at most 80 sectors). Queueing delay percentiles remain small and the load of the system is confined below unity.

Regarding 40 MHz channels, 100G and 200G links can give support to up to 40 and 80 sectors, respectively. However, upgrading to 400G transceivers would enable to aggregate more than 140 sectors while preventing congestion and full load. With respect to the future 100 MHz channels, it is clear that high throughput links will be mandatory. Neither 100G nor 200G links are able to provide enough capacity to transport a lot of fronthaul flows, only 3 and 20 sectors respectively. Surely 400G and 1T are the only options if we want to support the aggregation of fronthaul flows using such a high LTE channel bandwidth. In this light, a cost-effective trade-off between the number of aggregation switches and the number of FH flows that each one of them is able to carry shall be weighted.

It should be noted that these results are obtained for the worst case scenario regarding eCPRI $\mathbf{I}_U$ split: simultaneous 100% utilisation of all cells capacities. Split $\mathbf{I}_U$ generates traffic proportionally to the current radio resource utilisation. Therefore, a more optimistic and flexible bandwidth dimensioning can be expected by adjusting the required bandwidth to the target aggregate cell utilisation of the network.

Queueing delay [μs]		3 sectors			20 sectors			40 sectors			60 sectors			80 sectors			140 sectors		
		Mean	90th	99th	Mean	90th	99th	Mean	90th	99th	Mean	90th	99th	Mean	90th	99th	Mean	90th	99th
Channel BW 10 MHz	10G	0.21	1.22	6.61	$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$		
	40G	0.03	0	0.86	0.18	0.88	2.79	0.57	2.02	4.67	1.93	6.5	13.34	$\rho > 1$			$\rho > 1$		
	100G	0.001	0	0.13	0.023	0.05	0.74	0.05	0.28	1.03	0.096	0.44	1.25	0.15	0.61	1.53	0.69	1.9	3.99
	200G	$0.7 \cdot 10^{-3}$	0	0	0.005	0	0.26	0.011	0.024	0.38	0.019	0.09	0.46	0.028	0.14	0.52	0.063	0.27	0.70
Channel BW 20 MHz	40G	0.05	0	2.82	1.01	3.90	9.00	$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$		
	100G	0.008	0	0.65	0.10	0.57	2.13	0.29	1.24	3.10	0.73	2.29	5.00	2.32	8.61	17.53	$\rho > 1$		
	200G	0.002	0	0.14	0.023	0.05	0.8	0.053	0.29	1.09	0.095	0.47	1.33	0.15	0.63	1.58	0.70	1.85	3.89
Channel BW 40 MHz	40G	0.21	1.42	7.67	$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$		
	100G	0.03	0	2.14	0.55	2.43	6.07	3.5	15.15	30.88	$\rho > 1$			$\rho > 1$			$\rho > 1$		
	200G	0.009	0	0.69	0.10	0.59	2.21	0.29	1.26	3.15	0.74	2.30	5.01	2.32	8.45	17.2	$\rho > 1$		
	400G	0.002	0	0.15	0.022	0.05	0.84	0.05	0.3	1.13	0.096	0.48	1.37	0.16	0.65	1.62	0.7	1.85	3.88
Channel BW 100 MHz	100G	0.218	1.47	7.93	$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$		
	200G	0.052	0	3.07	1.0	3.93	9.09	$\rho > 1$			$\rho > 1$			$\rho > 1$			$\rho > 1$		
	400G	0.013	0	1.06	0.17	0.97	3.05	0.56	2.05	4.73	1.93	6.09	12.57	$\rho > 1$			$\rho > 1$		
	1T	0.0025	0	0.16	0.023	0.054	0.86	0.053	0.31	1.15	0.096	0.48	1.39	0.156	0.65	1.63	0.70	1.84	3.85

TABLE II

THEORETICAL QUEUING DELAY VALUES (MEAN, 90-TH AND 99-TH PERCENTILES) FOR THE TRANSPORT OF MULTIPLE SECTORS WITH DIFFERENT LTE BANDWIDTHS OVER 40G, 100G, 200G, 400G AND 1T TRANSCEIVERS IN C-RAN SCENARIOS.

V. SUMMARY AND CONCLUSION

Given the strict latency and jitter requirements demanded for the fronthaul traffic, fronthaul network dimensioning needs to be carried out taking into account, not only average queueing delays, but also worst-case queueing delays. These can be defined, for instance, as the 90-th or 99-th delay percentiles. Such worst-case delays are substantially higher (between one or two orders of magnitude) than conventional average queueing delay, thus requiring typically larger overprovisioning factors of capacity. In this article, we have shown both theoretically and with simulation that the *Kingman's Exponential Law of Congestion* provides a useful upper bound for such dimensioning type of problems and is often a good estimate to the actual delay percentiles.

As an application of the worst-case delay model of this paper, we have studied its suitability in defining dimensioning rules for a number of cellular scenarios where FH traffic flows follow the recently published eCPRI specification (splits $\mathbf{I}_{\mathrm{U}}$ and $\mathbf{II}_{\mathrm{D}}$). We observe that the transmission of multiple (20) legacy 20 MHz LTE channels using such functional split can be realised with 40 Gb/s transponders guaranteeing 99-th delay percentiles below 9 μs. However, scaling towards future 40 and 100 MHz LTE channels require higher-speed transponders in the range of 200 Gb/s, 400 Gb/s and even 1 Tb/s to guarantee ultra-low queueing delay 99-th percentiles values. Such transceivers are not yet available in market, although forecasts [28], [29] estimate that these will be ready soon in market, by 2018.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the Spanish project TEXEO (grant no. TEC2016-80339-R), and the H2020 EU-funded project BlueSPACE (grant no. 762055).

REFERENCES

[1] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, L. Dittmann, "Cloud RAN for Mobile Networks – A Technology Overview," IEEE Communications surveys & tutorials 17 (1), 405-426, 2015

[2] A. Checko, A. P. Avramova, M. S. Berger, H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings", Journal of Communications and Networks 18 (2), 162-172 (2016)

[3] Common Public Radio Interface (CPRI), Interface Specification v7.0, [Online]: http://www.cpri.info/spec.html.

[4] R3-161687, Draft TR 38.801 (v030) Study on New Radio Access Technology: Radio Access Architecture and Interfaces, NTT DOCOMO, INC (Rapporteur), 3GPP TSG RAN3, August 2016.

[5] Small Cell Forum Document 159.07.02, Small cell virtualization functional splits and use cases, Small Cell Forum, January 2016.

[6] A. De la Oliva, J. A. Hernández, D. Larrabeiti, *et al*, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," IEEE Communications Magazine, 2016

[7] D. Wubben et al., "Benefits and impact of cloud computing on 5g signal processing: Flexible centralization through cloud-ran," Signal Processing Magazine, IEEE, 2014.

[8] IEEE Next Generation Fronthaul Interface (NGFI) Working Group, Website: http://sites.ieee.org/sagroups-1914/

[9] IEEE Time-Sensitive Networking for Fronthaul 802.1CM http://www.ieee802.org/1/pages/802.1cm.html

[10] ITU-T Rec. G.Sup56 OTN transport of CPRI signals, Website: https://www.itu.int/rec/T-REC-G.Sup56-201602-I/en

[11] IETF Deterministic Networkiing, Website: https://datatracker.ietf.org/wg/detnet/about/

[12] T. Wan, P. Ashwood, A performance study of CPRI over Ethernet [Online]. Available: http://www.ieee1904.org/3/meeting_archive/2015/02/tf3_1502_ashwood_1a.pdf

[13] Andrews, Jeffrey G., et al. "What will 5G be?," IEEE Journal on selected areas in communications, vol. 32, no 6, p. 1065-1082, 2014.

[14] Ericsson, "Ultra-Reliable and Low-Latency (URLLC) 5G Communication", EuCNC'16, 2016, Online: http://kom.aau.dk/~nup/2016-06-27_Yilmaz-5G\%20Ultra-reliable-Low-latency_final.pdf

[15] 3GPP, TR (v14.3.0), specification #: 38.913."Study on scenarios and requirements for next generation access technologies," Release 14, 3GPP RAN# 76, August 2017.

[16] N. J. Gomes, P. Chanclou, P. Turnbull, A. Magee, V. Jungnickel, Fronthaul evolution: From CPRI to Ethernet, Opt. Fiber Technol., vol. 26, part A, pp. 5058, 2015.

[17] L. Valcarenghi, K. Kondepu, P. Castoldi, "Time-versus size-based CPRI in Ethernet encapsulation for next generation reconfigurable fronthaul", Journal of Optical Communications and Networking 9 (9), D64-D73, 2017

[18] D. Chitimalla, K. Kondepu, L. Valcarenghi, M. Tornatore, B. Mukherjee, 5G fronthaul-latency and jitter studies of CPRI over Ethernet, J. Opt. Commun. Netw., vol. 9, no. 2, pp. 172182, 2017.

[19] Nokia Siemens Networks White Paper, "LTE 1800 MHz: Introducing LTE with maximum reuse of GSM assets, " Section 4.2, 2011, Online: https://www.gsma.com/spectrum/wp-content/uploads/2012/03/lte1800mhzwhitepaper0.9.pdf.

[20] eCPRI Interface Specification, eCPRI Specification V1.0, 2017/10/24, [Online]: http://www.cpri.info/downloads/Requirements_for_the_eCPRI_Transport_Network_V1_0_2017_10_24.pdf

[21] J. Zhang, J. Yu, Y. Fang, N. Chi, "High Speed All Optical Nyquist Signal Generation and Full-band Coherent Detection", Scientific reports 4, pp. 1–8, 2014

[22] R Maher, et al, "Spectrally shaped DP-16QAM super-channel transmission with multi-channel digital back-propagation", Scientific reports 5, 1–8, 2015

[23] J.-X. Cai et al. "112x112 Gb/s transmission over 9,360 km with channel spacing set to the Baud rate (360% spectral efficiency)", European Conference on Optical Communications (ECOC), PD2.1, 2010

[24] A. Nespola et al. "Extensive fiber comparison and GN-model validation in uncompensated links using DAC-generated Nyquist-WDM PM-16QAM channels", Optical Fiber Communications (OFC) conference, OTh3G.5, 2013.

[25] A. Nespola, et al. "1306-km 20x124.8-Gb/s PM-64QAM transmission over PSCF with net SEDP 11,300(b.km)/s/Hz using 1.15 samp/symb DAC". Optics Express 22, pp. 1796 – 1805, 2014.

[26] R. Rios-Mller et al. "1-Terabit/s net data-rate transceiver based on single-carrier Nyquist-shaped 124 GBaud PDM-32QAM", Optical Fiber Communications (OFC) conference, Th5B.1, 2015.

[27] Huawei's white paper on technological developments of optical networks, pp. 1-22, [Online]: http://www-file.huawei.com/-/media/CORPORATE/PDF/white\%20paper/White-Paper-on-Technological-Developments-of-Optical-Networks.pdf 2016

[28] F Rambach, B Konrad, L Dembeck, U Gebhard, M Gunkel, M Quagliotti, L. Sierra, V. Lopez, "A multilayer cost model for metro/core networks", Journal of Optical Communications and Networking 5 (3), 210-225, 2013.

[29] V. Lopez, B. de la Cruz, Ó. G. de Dios, O. Gerstel, N. Amaya, G. Zervas, D. Simeonidou, J. P. Fernandez-Palacios, "Finding the target cost for sliceable bandwidth variable transponders", Journal of Optical Communications and Networking 6 (5), 476-485, 2014.

[30] Moreolo, M. S., Fabrega, J. M., Nadal, L., Vlchez, F. J., Mayoral, A., Vilalta, R., ... & Tanaka, T, "SDN-enabled sliceable BVT based on multicarrier technology for multiflow rate/distance and grid adaptation," Journal of Lightwave Technology, 2016, 34(6), 1516-1522.

[31] U. Dötsch, M. Doll, H. P. Mayer, F. Schaich, J. Segel, and P. Sehier, Quantitative analysis of split base station processing and determination of advantageous architectures for LTE, Bell Labs Tech. J., May 2013.

[32] Guidelines for LTE Backhaul Traffic Estimation, July 2011, White paper at ngmn.org

[33] G. O. Pérez, J. A. Hernández and D. L. López, "Delay analysis of fronthaul traffic in 5G transport networks," 2017 IEEE 17th International Conference on Ubiquitous Wireless Broadband (ICUWB), Salamanca, 2017, pp. 1-5. DOI: 10.1109/ICUWB.2017.8250956

[34] J. F. C. Kingman, The single server queue in heavy traffic, Mathematical Proceedings of the Cambridge Philosophical Society, vol. 57, no. 4, pp. 902904, 1961. DOI: 10.1017/S0305004100036094

[35] J. A. Hernández, P. Serrano: Probabilistic models for computer networks: tools and solved problems, 2013 (ISBN: 978-1291546873)

[36] China Mobile, *The Road Towards Green* RAN *(White Paper)*, [Online]: http://labs.chinamobile.com/cran/wp-content/uploads/2014/06/20140613-C-RAN-WP-3.0.pdf

[37] B. R. Ballal and D. Nema, "Performance Comparison of Analog and Digital Radio Over Fiber Link," International Journal of Computer Science & Engineering Technology (IJCSET), 2012.

[38] S. Kuwano and Y. Suzuki, "Digitized Radio-over-Fiber (DROF) System for Wide-Area Ubiquitous Wireless Network," IEEE Xplore, April 2007

[39] P. G. Harrison, Nares M. Patel, "Performance Modelling of Communication Networks and Computer Architectures", p. 336, ISBN 0-201-54419-9

[40] A. Gowda, J. A. Hernández, D. Larrabeiti, L. Kazovsky, "Delay analysis of mixed fronthaul and backhaul traffic under strict priority queueing discipline in a 5G packet transport network", Trans. Emerging Telecom. Tech. 28 (6), 1-9, 2017.

[41] J. Kani, S. Kuwano, J. Terada, "Options for future mobile backhaul and fronthaul", Optical Fiber Technology 26, pp. 42 – 49, 2015.