Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-14119-w

1229: MULTIMEDIA DATA ANALYSIS FOR SMART CITY ENVIRONMENT SAFETY

Performance model for factory automation in 5G networks

Jiao Wang[1] · Jay Weitzen[1] · Oguz Bayat[2] · Volkan Sevindik[1] · Mingzhe Li[3]

Received: 19 March 2022 / Revised: 5 October 2022 / Accepted: 25 October 2022

Abstract

The fifth generation (5G) of mobile networks is emerging as a key enabler of modern factory automation (FA) applications that ensure timely and reliable data exchange between network components. Network slicing (NS), which shares an underlying infrastructure with different applications and ensures application isolation, is the key 5G technology to support the diverse quality of service requirements of modern FA applications. In this article, an end-to-end (E2E) NS solution is proposed for FA applications in a 5G network. Regression approaches are used to construct a performance model for each slice to map the service level agreement to the network attributes. Interference coordination approaches for switched beam systems are proposed to optimize radio access network (RAN) performance models. A case study of a non-public network is used to show the proposed NS solution. Simulation result shows that for services with different QoS requirements, different IC approaches should be used as optimization methods. Design prediction using regression approach has been evaluated and shows that the prediction successful rate increases when more existing data are used.

✉ Jiao Wang
jiaowang2010@gmail.com

Jay Weitzen
jay_weitzen@uml.edu

Oguz Bayat
oguzbayat@gmail.com

Volkan Sevindik
vsevindik@gmail.com

Mingzhe Li
mingzhe.li@gmail.com

[1] Department of Electrical and Computer Engineering, University of Massachusetts, Lowell, MA 01854, USA

[2] Graduate School of Science and Engineering, Istanbul Kemerburgaz University, Istanbul, Turkey

[3] Q Factor Communications, 255 Bear Hill Road, Waltham, MA 02451, USA

Published online: 10 November 2022

Keywords 5G · URLLC · Massive MIMO · Joint transmission · Interference coordination · Network slicing

1 Introduction

The fifth generation (5G) network is emerging as a key enabler of ultra-reliable low-latency communication (URLLC) applications [25]. Deviating from traditional human-centric, delay-tolerant applications, URLLC is a new service category that accommodates emerging services with stringent latency (referred to as hard real time, i.e., in ms) and reliability requirements (i.e., 99.9999%) [11].

One URLLC application is factory automation (FA), which deals with the automated control and optimization of processes and workflows in a factory. Nowadays, industrial Ethernet technologies at layer 2 (L2) are used for FA (i.e., Profinet) [19]. Profinet uses unique MAC addresses to identify field devices, including three communication channels to provide FA services: TCP/IP channel for non-deterministic functions; real-time channel (Profinet RT) for services with latency in the 1–10 ms range, where TCP/IP layers are bypassed to have a deterministic performance; and isochronous real-time channel (Profinet IRT) for services with latency less than 1 ms, where high-precision synchronization for low cycle time and hardware support with ASICs are required to achieve low latency [19]. Profinet can support a cycle time lower than 250 μs.

To realize the better flexibility, mobility, and versatility required by modern smart factories, wireless connectivity is preferred to replace industrial Ethernet cable. The 5G mobile network is emerging as a key enabler of modern FA applications that ensure timely and reliable data exchange between network components. To support FA in 5G cellular network, fundamental changes are necessary in both wireless links and transport/core networks.

Industrial Ethernet technologies at L2 can be directly used as a transport network for 5G cellular network. For transport network, URLLC traffic with stringent latency requirements guaranteeing low latency and high reliability has traditionally been achieved using reserved resources. However, this approach led to the inefficient utilization of network resources. As an alternative, L2 links can be implemented using the 5G user plane function that replaces GTP/IP with virtual local area network (VLAN) tunnels for URLLC traffic. Software-defined networking (SDN), which provides the capability to allocate L2 links based on the number of low latency traffic flows, can be used to enforce fine-grained traffic management in the VLANs compared to L3/L4 traffic throttling at the GTP/IP layer [13].

For radio access network (RAN), reducing processing time and supporting a shortened frame structure are the two basic mechanisms defined in LTE Release 15 to reduce latency. A detailed analysis of the resulting latencies, which are feasible with LTE Release 15, is given in [7]. To improve reliability, signal-to-interference-plus-noise ratio (SINR) can be improved by increasing the signal power with redundancy and diversity (e.g., joint transmission [JT] approach) and/or reducing the interference power via interference coordination (IC). [14]

has proposed various micro and macro diversity techniques, and [5, 12, 15, 17, 18, 20, 24] has proposed various IC approaches.

IC approaches include static and dynamic approaches. Static IC approaches are based on frequency planning, which includes conventional fractional frequency reuse, partial frequency reuse, and soft frequency reuse [20]. Dynamic IC approaches include the frequency domain IC approach, such as carrier aggregation-based IC [15], which allocates different carriers to interfered UEs to avoid interference; the time domain IC approach, such as the almost blank subframe approach [5], which allocates difference time slots to interfered UEs to avoid interference; and the spatial domain IC (SIC) approach. With a massive multiple-input, multiple-output (MIMO) antenna, SIC approaches include coordinated beamforming (CBF) and coordinated multi-user MIMO (MU-MIMO). For example, precoding matrix indicator (PMI) coordination is a lightweight CBF approach, where each UE transmits a restriction PMI or recommendation PMI, and neighboring cells either use the recommendation PMI or avoid using the restriction PMI [12]. Coordinated MU-MIMO approaches form a virtual transmitter among neighboring cells, jointly perform MU-MIMO transmission to eliminate intercell interference and achieve higher capacity [17]. Graph-based or utility-based approaches are used to share the time and frequency resources for dynamic IC approaches. Graph-based IC approaches partition interference graphs and avoid allocating the same time or frequency resources to UEs that are connected in the graph (represent high interference) to minimize SINR [18]. Utility-based IC (UIC) approaches are designed to maximize network utility by using a two-level approach [24]. The utilities for scenarios with different numbers of interferers are calculated at the cell level, and the conflicts of interferers are resolved at the central level.

With a massive MIMO antenna, 5G beamforming systems include switched beam systems (SBSs) and adaptive array systems (AASs). An AAS [9] generates beam patterns to direct main lobes toward desired UEs and nulls toward interfered UEs while an SBS [27] uses fixed beam patterns to point toward UEs in predetermined directions. Compared to an SBS, an AAS is expensive for commercial mobile networks. Thus, we use SBS in this research.

Network slicing (NS) is a technique incorporated by 5G that enables the coexistence of heterogeneous URLLC services in the same network. This technique divides the network into slices that are tailored to specific service requirements (described in the service level agreement [SLA]). For each slice, it is critical to respond intelligently to the dynamics of the traffic load to obtain satisfactory quality of service (QoS) at an acceptable cost [13]. Owing to the complex implementation, no linear relationship exists between SLA requirements and resource-specific network attributes that fulfill a service's QoS. In addition, 5G networks are expected to provide diverse QoS guarantees for a wide range of services. Therefore, it is increasingly difficult to translate user-friendly SLA business terms into physical resources in 5G transport and RAN slices.

Although the network cost to provide URLLC services can be significantly reduced with the above SDN, IC/JT, and NS technologies, the higher reliability requirements of the URLLC services remain the biggest burden on cost in today's mobile broadband

(MBB) services. It is important to characterize the deployment cost for URLLC services. For 5G networks, the total cost of ownership (TCO) includes capital expenditure (CAPEX) and operational expenditure (OPEX). The main cost contributions for the network investment are as follows:

- CAPEX, including site infrastructure: GUTRAN Node B, network equipment, cabinets, civil works (physical cabinets, fences, antenna masts, etc.), fiber backhaul, etc.
- OPEX, including network operation, maintenance and replacement, site lease, etc.

1.1 Contribution

In this research, an end-to-end (E2E) NS solution is proposed for FA application and its URLLC services for a stand-alone non-public network (SA NPN) [21], including both RAN and transport networks. We employ different mechanisms introduced in this section to satisfy the stringent latency and reliability requirements of services, construct performance models, and translate SLA business terms into physical resources in 5G transport and RANs. For RAN network, shortened frame structure has been employed and both JT and UIC approaches (coordinated across time, frequency, and spatial domains) are used to optimize the performance models. The approach is deployed on a massive MIMO SBS and employs MU-MIMO to improve the system capacity. For transport network, L2 links are dynamically allocated that traffic flows in slices with tighter latency requirements are assigned with higher priority on resource allocation. The main contributions of this approach include the following:

- An E2E NS solution for FA services that satisfies the stringent latency requirements better than today's MBB services. Instead of looking at only one feature, which may not work as expected in the real world, we look at a complete E2E setup that employs multiple features that cooperate and mimic the actual network environment.
- Unlike most IC/JT approaches that target SINR improvement, our approach targets improvement in performance models mapping customer-friendly SLAs into network design parameters and fulfills QoS requirements for URLLC services. We also show that different IC/JT approaches should be used for different URLLC services to optimize their QoS requirements.
- We propose a data analytics and regression approach for each network slice to construct a QoS performance model and automate the identification of a nonlinear relationship between SLA requirements and resource-specific network attributes.
- Given the high reliability requirement of FA applications (e.g., motion control service requires 2 ms of cycle time and 99.9999% of service availability), the network designed to support FA applications is much more expensive than today's mobile network. In this research, we estimate the cost for design to fulfill the stringent QoS requirements of modern factory applications.

This research targets to improve the design of future 5G networks with large number of sites and diversified service requirements and is a first step to automate the design process by modeling the performance of an E2E network with key 5G features.

The paper is organized as follows: Section 2 describes the problem, Section 3 details our proposed approach for performance model construction, Section 4 presents our simulation setup, Section 5 shows the simulation results, Section 6 analyzes the complexity, and Section 7 concludes the paper.

2 Problem description

Consider an area supporting S services, denoted as S = $\{1, \ldots, S\}$, with U_s users, denoted as $U_s = \{1, \ldots, U_s\}$, distributed uniformly. Consider a 5G network with E component networks, namely, a transport network and a RAN, denoted as $E = \{1, \ldots, E\}$. Only downlink transmission is considered, and all users share the aggregated bandwidth W_e. Owing to the lack of physical resources for the transport network, switches/routers are shared among services. For the RAN, sites are shared among services and distributed uniformly (we assume hexagonal grids with the same inter-site distance). The problem is written as follows.

Problem SM-P1:

$$\begin{aligned} &\qquad\qquad\qquad\quad Minimize\ [Cost(w, isd)] \\ &\text{over}: w, isd \\ &\text{subject to}: \\ &\qquad\quad \mathcal{N}\left(qos_{u_s}(w, isd)\right) = 1 \forall u_s \in U_s, s \in S \end{aligned}$$

where

$$\mathcal{N}\left(qos_{u_s}(w, isd)\right) = \begin{cases} 1 & qos_{u_s}(w, isd) < T_s \\ 0 & Else \end{cases}, \tag{1}$$

where T_s is the predefined threshold for service S. $qos_{u_s}(w, isd) < T_s$ indicates that user u_s satisfies the QoS requirement T_s of service s under the network attributes of bandwidth distribution w and site separation isd. w is the bandwidth distribution vector for component networks, defined as $w = (w_1, \ldots w_E)$, and w_e is the bandwidth distribution among services, defined as $w_e = \left(w_e^1, \ldots w_e^S\right)$, satisfying the constraint $\sum_{s \in S} w_e^s = W_e$. $Cost(w, isd)$ is calculated using a predefined cost model (see Table 1).

To simplify the problem, let's look at one service on one component network. For service s on component network e, we can allocate a slice and define a performance model as follows:

$$PM_e^s\left(w_e^s, isd\right) = \sum_{u_s \in U_s} \mathcal{N}\left(qos_{u_s}\left(w_e^s, isd\right)\right),$$

where

$$\mathcal{N}\left(qos_{u_s}\left(w_e^s, isd\right)\right) = \begin{cases} 1 & qos_{u_s}\left(w_e^s, isd\right) < T_e^s \\ 0 & Else \end{cases}, \tag{2}$$

and T_e^s is the predefined threshold for service s on component network e. If we define T_e^s as latency, then the constraint $\sum_{e \in E} T_e^s \leq T_s$ needs to be satisfied.

The problem SM-P1 is complex. For each piece of slice (i.e., each service on each component network), a performance model needs to be solved. Data analytics and regression approaches that allow automated identification of nonlinear relationships are exploited. A simulation framework has been built to simulate the S services on proposed 5G networks with varied network attributes (w, isd), and the output performance data are used by data analytics and regression approaches to build performance model. The proposed approach is described in the next section.

3 Approaches

In this section, we describe a private FA network and our proposed approach to construct performance models. Figure 1 shows the network topology. Two services are considered. Service 1 (S1) is a motion control service with the characteristics of a printing machine application (2 *ms* cycle time, 99.9999% service availability). For S1 service, although throughput and signal bandwidth requirements are typically low between controller and each actuator, given number of actuators per service areas (SAs), the network capacity requirement is not low. Service 2 (S2) is an augmented reality service with one-way E2E latency of 10 *ms* and 99.9% service availability [1]. The cloud server and robotic device each take 250 μs latency [2], leaving latency budgets of 1.5 *ms* for S1 and 9.5 *ms* for S2 for network

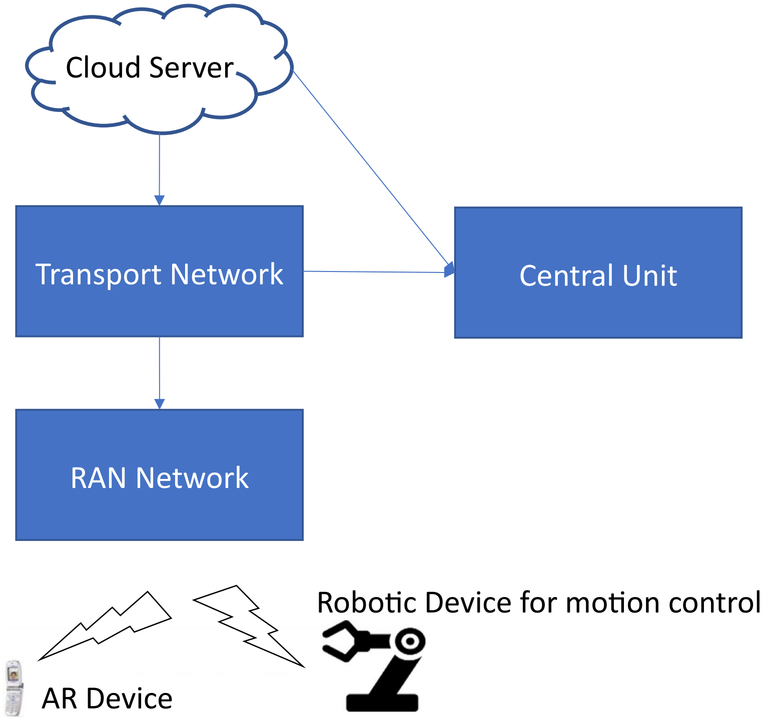

Fig. 1 Network topology

transmission and processing. The private FA network includes a transport network, which is a switch-only network with fat-tree topology, and a 5G RAN; thus, latency will add up.

[1] Construction of transport network performance model

VLAN tunnels are used to identify field devices, and SDN is used to enforce fine-grained traffic management [13] in a transport network. With slices sharing transport network resources, traffic flows in slices with tighter latency requirements are assigned with higher priority on resource allocation. For example, of the two services, S1 takes higher priority than S2. Specifically, when an S1 packet is available, an S2 packet will stop processing and wait until the S1 packet finishes processing to continue.

Each switch in the transport network uses a shared memory switch model [6]. Packets are put in first-in first-out (FIFO) queues and forwarded with the store-and-forward package forwarding mechanism. [8] specifies the five steps for latency calculation on one switch as follows:

1. Transmit bits to the input FIFO of the switch port.
2. Write the packet data into the switch memory.
3. Perform a lookup (operates in parallel with data storage and therefore causes no latency).
4. Read the packet from the data memory.
5. Transmit the packet.

The following pseudo code shows the latency calculation/simulation for one of the above steps.

Algorithm 1: Latency calculation

```
For every time slot ts
   Check input buffer for all users for service s1
   If input buffer of user u has data
     If recorded input time for user u < ts
        Transmit R*(slot time) bits to input buffer of next step
        Record time ts as next step input time
     Endif
   Endif
   Check input buffer for all users for service s2
   If input buffer of user u has data
     If recorded input time for user u < ts
        If ts is not used by any user for service s1
           Transmit R*(slot time) bits to input buffer of next step
           Record time ts as next step input time
        Endif
     Endif
   Endif
EndFor
```

R is the data rate that depends on the input bandwidth/capacity. For input/output ports, R is equivalent to the line rate. To provide line rate switching on all ports, the memory data rate needs to be such that all packets arriving simultaneously on all ports at line rates can be written to memory and all packets departing simultaneously on all ports at line rates need to be read from the memory [8]. Algorithm 1 is run for each step to calculate switch latency.

For a transport network with a fat-tree topology and multiple levels, we assume the same bandwidth/capacity for each level but distributed to multiple switches within each level. Latency calculation is performed for each switch along the transmission path and summed up as the latency over the transport network. Cable latency between switches is calculated and added to transport network latency.

[2] Construction of RAN performance model

To construct the RAN performance model, multiple diversity and IC approaches are implemented to achieve high reliability and low latency transmission. To provide the guaranteed QoS, resources of slices are isolated, that is, the slice for each service allocates dedicated RAN resources and employs different diversity or IC approaches to achieve QoS assurance. For S1, no hybrid automatic repeat request is permitted because of a tight latency budget.

A. Micro diversity, redundant transmission, switched beam system, and antenna down-tilting

Figure 4 shows the beams pattern for basic and switched directional antennas in an SBS [29]. UE only transmits on one narrow beam to reduce interference. Micro diversity (2 × 2 MIMO antenna) is used with the transmission mode (TM) of 2 for the S1 service, which tolerates a considerably lower SINR compared to other TMs. At low SINR, the data can be transmitted multiple times non-coherently in a frequency domain to increase signal power and reliability. To reduce interference from neighboring cells, the antenna is down tilted to limit the range of the main beam.

B. Interference coordination

IC approaches, including the two-level IC approach [29] and non-coherent JT (NCJT) approach, are implemented and compared in terms of performance model optimization. Different service requirements lead to different IC approaches. Figure 1 illustrates a central unit in the network topology used to facilitate coordination in the RAN.

1) Two-level interference coordination

Neighboring sectors close to one another are grouped into clusters. At the intra-cluster level, within each cluster up to M_{max}, UE can be scheduled simultaneously using the same time–frequency resource. The algorithm starts with a randomly selected UE. A UE is added if the interference caused to other UEs within the cluster is low and SINR of UEs within the cluster are higher than the predefined threshold until the maximum number of users per cluster is reached. At the inter-cluster level, cluster-edge UE reports forbidden UEs in neighboring clusters, and the cluster-edge UE can achieve higher SINR if some or all forbidden UEs are deallocated. The following pseudo code shows the algorithm.

Algorithm 2: Two-level interference coordination

```
For each cluster c1
   Scheduled UE set: S
   Initialization:
      Randomly pick cell c2 in cluster c1
      Randomly pick UE u1 in cell c2
      $U = \{u1\}$

      For each cell c2 in cluster c1
         For each UE u1 in cell c2
            If u1 ∉S && size ($U$) < $M_{max}$
               $u1^{*} = argmin_{u1 \notin S} SINR(U \cup \{u1\})$
            Endif
         EndFor
      EndFor
      If $SINR_k < SINRThreshold\ \forall k \in \{U \cup \{u1^{*}\}\}$
        $U = U \cup \{u1^{*}\}$ ; $S = S \cup \{u1^{*}\}$
      Endif
EndFor

 For each cell c2 in cluster c1
   For each UE u1 in cell c2
      If u1 ∈ S and $SINR_{u1} < SINRThreshold$
         Sort $Interferer_{u1}$;
         For all UE $k \in S$ & $k \in Interferer_{u1}$
            Remove k from S
            Recalculate $SINR_{u1}$
            If ! $SINR_{u1} < SINRThreshold$
               Break Loop;
            Endif
         EndFor
        Endif
     EndFor
 EndFor
```

2) Non-coherent joint transmission

For NCJT, N beams from different sectors are scheduled to transmit simultaneously to one UE. The following pseudo code shows the NCJT approach:

Algorithm 3: Non-coherent joint transmission

```
For each time slot: Forbid set: F = $\emptyset$
Scheduled UE set: S, #NCJT: N, Forb = 0
For all UE $k$
  If $k \notin S$ && $k \notin F$
    Forb = 0;
    Sort $Interferer_k$;
    For $i \in$ set of first (N-1) $Interferer_k$
      If $i \in S$ or $i \in F$
        Forb = 1;
      Endif
    EndFor
    If Forb == 0
      S = $S \cup \{k\} \cup$ {set of first (N-1) $Interferer_k$}
      F = F $\cup$ {$Interferer_i$: i $\in$ {{$k$} $\cup$ {set of first (N-1)
$Interferer_k$}}}
    Endif
  Endif
EndFor
```

C. Performance model construction

The following algorithm creates a performance model for the RAN, which maps SLA requirements (i.e., latency and reliability) into network attributes such as bandwidth W and ISD. To find this mapping, numerous simulations are performed to generate the data set, from which the performance model is learned using a regression approach.

Algorithm 4: Generation of RAN performance model

```
For serv. $s1$, #UE $u1$, #beams $b1$, and any fixed ISD isd1
    Start with bandwidth: $W$
    Loop1:
        calculate QoS (s1, u1, b1, isd1, $W$):
        Loop2: for each cell $j$ in cluster $c$
            Loop3: for each UE $u2$ in cell $j$
                If QoS ($u2$) $>$ QoS$Requirement$
                    Valid(u2) = true;
                Endif
            EndLoop3
        EndLoop2

        while numValidUE < THRESHOLD
            $W$ *= 2;
            $\Delta W = W$;
            To loop1
        Else
            $W$ -= $\Delta W$;
            BreakLoop1;
    N is iteration number; // fine tune
    Loop 4: I = 1 to N
        If numValidUE < THRESHOLD
            $\Delta W = \Delta W/2$;
        Else
            $\Delta W = - \Delta W/2$
        EndIf
        calculate QoS (s, u, b, isd, $W + \Delta W$)
    EndLoop4
```

4 Simulation methodology

[1] Scenario description

We consider a scenario where a central office connects to multiple factories. SA NPN is assumed. The cloud server sits at the central office and connects to the main switch or entry point of the transport network. Each factory has a head room that holds the second level up to N levels of the transport network and a cabinet for the RAN baseband. The distance between the central office and the factory head room is fixed at 10 km.

Each factory has an area of 1 km × 1 km, and N (i.e., N = 100) AR (or S2) users are randomly distributed in the area. The factory is divided into service areas (SAs), where each

100 m × 100 m area has N (i.e., N = 100) robotic devices (or S1 users) randomly distributed in each SA. SAs are separated by a street that is 5 m wide. For simplicity, only one factory is considered in our simulation.

S1 is a motion control service with the characteristics of a printing machine application with service requirements of 2 *ms* cycle time, 99.9999% service availability, and 20 bytes message size. S2 is an augmented reality service with service requirements of a one-way E2E latency of 10 *ms*, 25 Mbps data rate, and 99.9% service availability.

[2] Transport network description

The transport network simulator includes multiple levels of switches. For simplicity, a minimum three-level switch network is assumed. The first level is at the central office, the second level comprises the factory head room main switches, and the third level comprises the switches on each BTS cabinet. For each transmitted packet, we add 8 bytes for the UDP header, add 20 bytes for the IP header, and 18 bytes for the Ethernet header. For every switch, 60 *ns* cycle time is assumed, with multiple words reading and writing from/to memory. With fixed number of UEs, our performance model can be simplified to show the dependency of the transport network latencies on the network attributes (i.e., network bandwidth).

[3] Radio access network description

In the RAN simulator, sites are distributed uniformly (we assume hexagonal grids with the same inter-site distance) within the factory. Each site has three sectors, and each sector has a directional antenna. The antenna of sector 1 points north, and each of the antennas of the other two sectors are deviated 120 degrees clockwise from the previous one.

Our solution is based on available commercial product with shortened frame structure. Each radio frame is 10 *ms*, consisting of 80 0.125 *ms* time slots or transmission time intervals (TTIs). The bandwidth is split into resource blocks, each with 1.44 MHz of bandwidth. The 2 × 2 MIMO TX mode 2 for S1 and TX mode 4 for S2 are used. In addition, 64 quadrature amplitude modulation, control format indicator 1, and Pedestrian B multipath channel model (PedB) are used in the simulation. Figure 2 shows the link curve for the 2 × 2 MIMO TX mode 2, which is generated using a Vienna LTE simulator [16]. The link curve is for motion control traffic S1, with 10^{-7} of radio link block error rate and assuming that the ITU standard multipath channel model PedB is used. When SINR is less than the channel quality indicator 1, redundant transmission occurs. On each site, on every TTI, UEs are scheduled, and their data are sunk according to their SINR and the link curve.

Massive MIMO antennas are assumed in the simulation, with eight narrow beams to cover the sector area horizontally (15° of half power beam width [HPBW]). Vertically, the HPBW is 6°, down-tilted by 10° for a cell range less than 150 m (which results in a main beam cover area with a radius between 30 and 60 m, a second beam cover area with a radius of less than 30 m, and an area radius between 60 and 220 m), and down-tilted by 9° for a cell range larger than or equal to 150 m (which results in a main beam cover area with a radius between 33 and 67 m, a second beam cover area with a radius of less than 33 m, and an area radius between 67 and 340 m). A maximum of four MU-MIMO users per sector is assumed. Figures 3 and 4 show the antenna and beams, respectively.

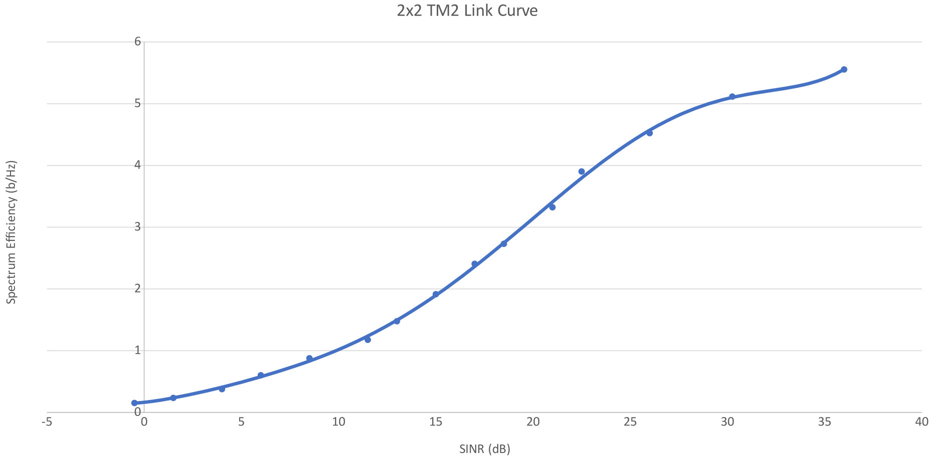

Fig. 2 2 × 2 TM2 link curve

The obstructed line-of-sight light clutter propagation model at 2.4 GHz in [25] is used for SINR calculation. The following equation shows the path loss:

$$PL\ (d) = PL(d_0) + 10nlog_{10}\left({}^{d}\big/_{d_0}\right), \tag{3}$$

where $PL(d_0)$ = 72.71 dB is the path loss at a reference distance of 15 m, n is the path loss exponent and equals 1.52, and σ is the standard deviation of shadowing that equals 4.61 dB.

[4] Latency calculation

The latency in the downlink direction is calculated using following formula:

$$T_{DL} = T_{CloudServer} + T_{Transport} + T_{RAN} + T_{UE}. \tag{4}$$

$T_{CloudServer}$ is the latency in the cloud server, which is 250 μs [2]. $T_{Transport}$ includes the latency in fiber T_{Fiber} and in switch T_{Switch}. Assuming 10 km of distance between the central office and

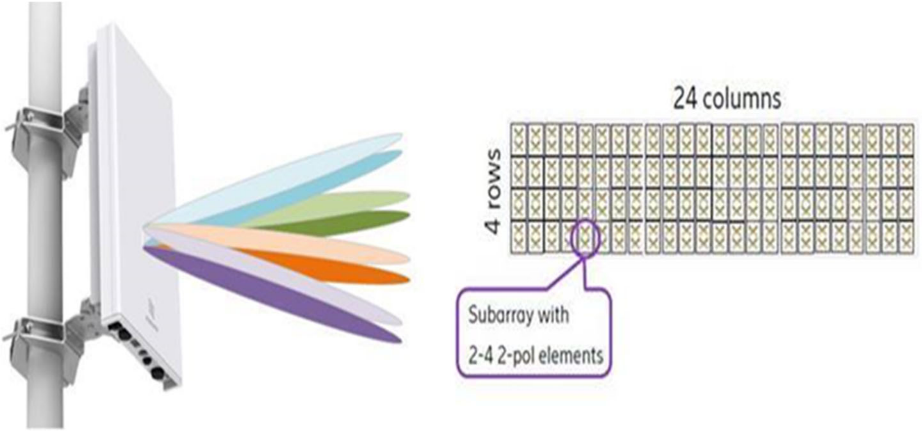

Fig. 3 Massive MIMO antenna

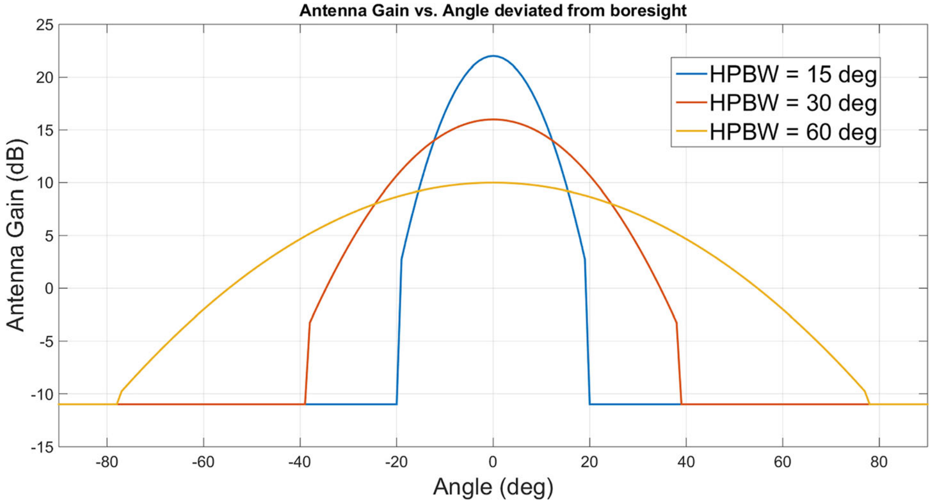

Fig. 4 Gains of a basic antenna

the factory head room and 5 μs/km of fiber latency, T_{Fiber} is 50 μs. T_{Switch} can be obtained from the transport performance model. T_{RAN} includes T_{Align}, $T_{Scheduling}$, and T_{PON}. T_{Align} is the alignment latency, and T_{Sche} is the scheduling latency; both are simulated in the construction of and included in the RAN performance model. The passive optical network connects the baseband to the radio in the RAN and is required to support latency within 100 μs [4]. T_{UE} includes T_{Proc} and T_{act}. T_{Proc} is the UE processing latency and takes three TTIs from [7]. T_{act} is latent on the robotic device and takes 250 μs from [2].

[5] Cost model

Our design minimizes the TCO, which includes both CAPEX and OPEX and satisfies stringent service requirements. The equipment and service costs are obtained from the literature [3, 10, 22, 26, 28] and a commercial RAN design case. The cost is summarized in Table 1.

5 Simulation results

Figure 5 shows the performance model for the switch-only transport network, with maximum latencies of S1 and S2 over the cost of the network, assuming 100 S1 UEs per SA and 100 S2 UEs per factory. The transport network is shared between two services, but S1 has a higher scheduling priority than S2. One hundred simulations are run using different random seeds, and the maximum latency is calculated. The latency of the store-and-forward switch at L2 is as expected as in [23]. The cost is calculated using the cost model in Table 1, we can see that the

Table 1 Cost model summary

	Budgetary Estimated Price	Notes
Transport Network		
Switch	$450/10G	
Fiber Backhaul from CO to HR	$18000/mile	
Inter-rack cable	$50	
Intra-rack cable	$10	
RAN Network		
RAN 64T64R Base station	$110k	Includes RBS, Baseband, Battery Backup System, Antennas, Site Material, GPS
RAN Services – In Buildings	$90K	Includes Site Survey, Site Acquisition, Site Build, Power, Backhaul Deployment, Install Commission, NW Design, Site Shakeup, Closeout
Customer Support	$10,500 per RBS	Includes RAN, Network Management
Staff/OAM cost	$50k annually for a Field Tech	
PON Network		
GPON ONT	$100	
DWDM OLT linecard (80 channels, incl.TRx,Diplexer, 2 slot shelf space)	$7.2K	
Arrayed waveguide gratings (AWG) - 1:40	$1.2K	
Fiber Cable Installation - In Buildings	$18000/mile	
Kwh power rates	12 cents per Kwh	

cost of transport network equipment (i.e., switch) increases when the required bandwidth increases. With higher CAPEX spent on the transport network, higher bandwidth can be supported, which will reduce the network latency.

Figure 6 compares the performance of IC approaches, assuming a dedicated RAN slice for S1 and S2. Plot (a) shows IC approaches for S1, where TwoIC is the two-level IC approach with a predefined SINR threshold of 0 dB, JIC-3 is the NCJT approach with three JTs, and JIC-5 is the NCJT approach with five JTs. The JIC-3 is shown to provide the best performance. Plot (b) shows IC approaches for S2, where OneIC represents the one-level IC approach and only intra-cluster scheduling has been employed, and TwoIC improves OneIC with inter-cluster scheduling. JIC-3 is the NCJT approach with three JTs. Figure 6 shows that for different IC approaches should be used for different services. In particular, the NCJT approach with three JTs is best for S1 dues to it's higher requirement on cell edge performance while the two-level IC approach is better for S2 dues to it's higher requirements on both cell edge performance and cell capacity. For NCJT approaches, more traffic will flow through the transport network, which necessitates a different transport network performance model.

Figure 7 shows the performance model for a network design with both an S1 and an S2 where the best IC approach for each service is used. A transport network performance model is included in the simulation. One hundred simulations are run with different random seeds. For network design purposes, a maximum bandwidth is calculated.

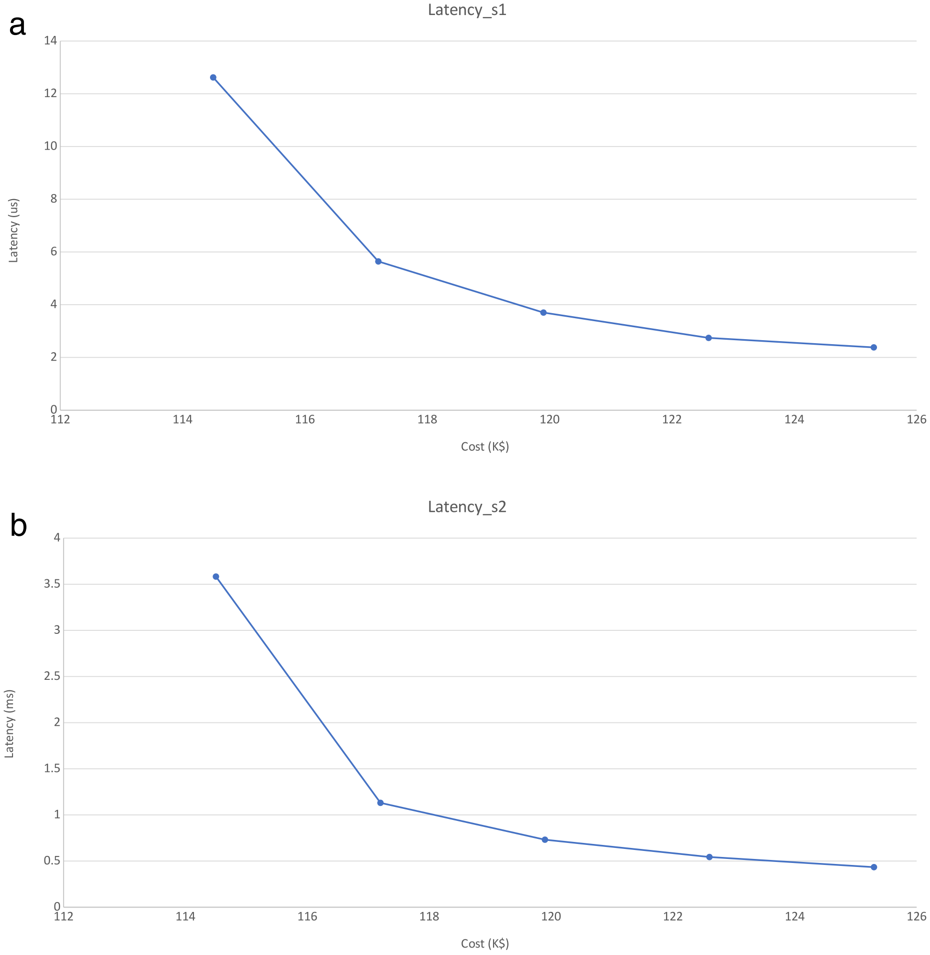

Fig. 5 **a** Switch-only transport network: latency versus cost – s1, **b** Switch-only transport network: latency versus cost – s2

Figure 8 shows the violation rate for a performance model. For network design, we propose a regression approach to learn from the performance model of existing data and use the prediction to design for the new data. In this example, the performance model is for S1, assuming a cell radius of 250 m. The initial performance model is generated using 100 drops, which indicate the minimum bandwidth required for the traffic model described for S1, as shown in Fig. 6. For any new drop using a different random seed, we verify if the QoS requirements are satisfied. If not, we employ algorithm 4 to find a new solution and update the performance model. From Fig. 8, we can see that the prediction successful rate increases when more drops are run. For simplicity, Fig. 8 predict design for new data with the same traffic model but using a different random seed. Using interpolation, design for different traffic load or network attributes can be predicted, but with higher violation rate.

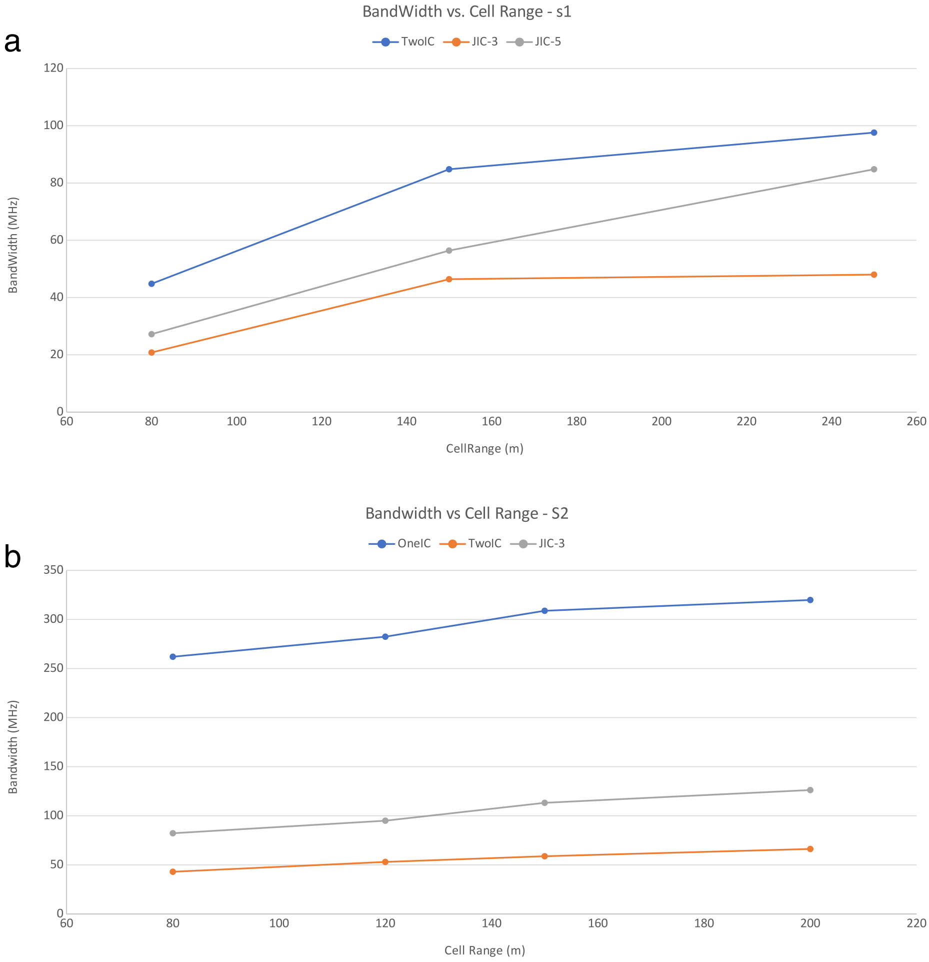

Fig. 6 **a** IC approaches for service 1, **b** IC approaches for service 2

6 Complexity analysis

For communication complexity, a dedicated fiber network is used to connect the cloud server, central unit, baseband units, and remote radio units to transmit data or between multiple baseband units (BBUs) to exchange intra-cluster scheduling information. Therefore, control data volume is not a concern, and we assume an extra time slot (125 μs) of delay for IC communication between BBUs. The IC approaches are run at the central unit, and RF is predicted based on the knowledge of sites and UEs.

For computational complexity, multiple SINR values corresponding to different interference scenarios are calculated for the two-level approach. The intra-cluster scheduling on BBUs

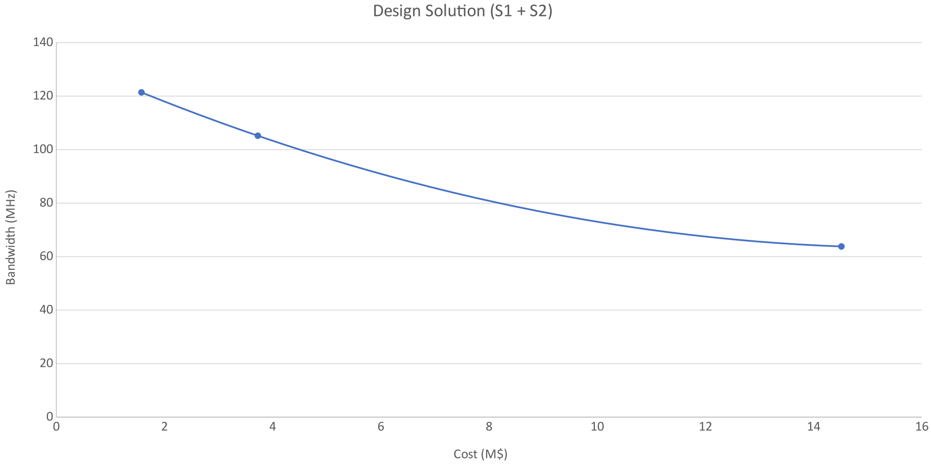

Fig. 7 Design solution with different available bandwidth

uses greedy algorithm with a complexity of $O\left(|C|*U_c^2\right)$, where $|C|$ represents the number of clusters in the network and U_c equals $C_{max} \times U_s$ represents the number of UEs in each cluster. C_{max} represents the maximum number of sectors per cluster, and U_s represents the number of UEs per sector. The inter-cluster scheduling is triggered only by the scheduled UEs at the cluster edge with a complexity of $O(N * \log(N) * U_e)$, where U_e represents the number of scheduled UEs at the cluster edge, N represents the number of cells, and $O(N * \log(N))$ represents the sorting complexity. For the NCJT approach, each UE with multiple JT beams from multiple sectors is picked using a greedy algorithm with a complexity of $O((N * U)^2)$, where U represents the number of users, and N represents the number of simultaneous transmitted beams.

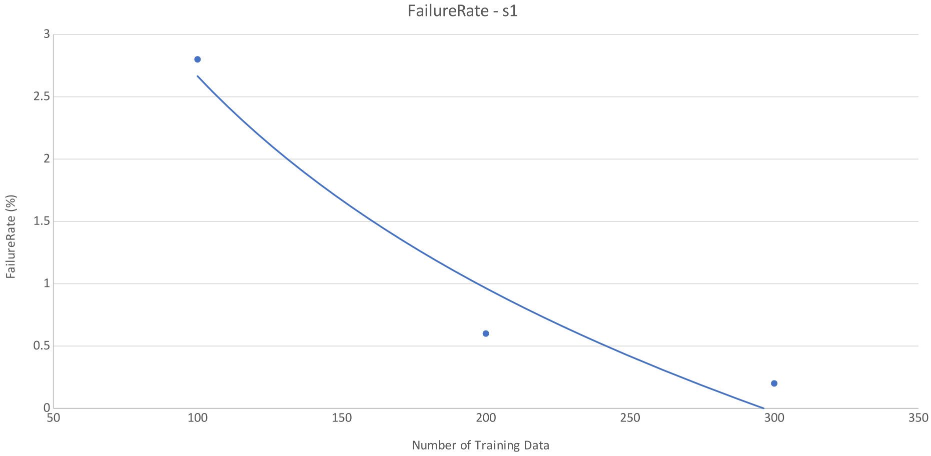

Fig. 8 RAN: prediction failure rate

7 Conclusion

In this research, we proposed an E2E NS solution and compared a two-level IC approach and NCJT IC approach. The IC approaches were used to improve performance models, which map customer-friendly SLA into low-level network design parameters and fulfill QoS requirements for a URLLC application. A regression approach was also proposed to identify the nonlinear relationships between customer-friendly SLA and low-level network design parameters.

The simulation result shows performance models for both transport network and RAN network. The RAN network performance model shows that for services with different QoS requirements, different IC approaches should be used as optimization methods. The regression approach to learn from existing data and use the prediction to design for the new network has been evaluated, we can see that the prediction successful rate increases when more existing data are used for learning.

Our future work will focus on developing cognitive system to automate the network design process, i.e., automate the mapping of the SLA to the network attributes. Semantic technologies can be used to model the diversified customer service requirements (i.e., SLA), the network equipment and attributes, and network performance models; machine learning approaches, knowledge base and reasoning engine can be developed to facilitate the construction of performance models and guide the mapping process.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations All authors designed the study. Jiao Wang performed the simulations and wrote the manuscript. All authors read, edited, and approved the manuscript.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. 3GPP TR 22.804 V16.2.0, Study on communication for automation in vertical domains (release 16) December 2018
2. Brown G. Ultra-reliable low-latency 5G for industrial automation, Qualcomm, San Diego, CA, USA, Tech. Rep. [Online]. Available: https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/ultra-reliable-low-latency-5g-for-industrial-automation.pdf
3. Chen H, Li Y, Bose SK, Shao W, Xiang L, Ma Y, Shen G (2016) Cost-minimized design for TWDM-PON based 5G mobile backhaul networks. IEEE/OSA J Opt Commun Netw 8(11):B1–B11
4. Chitimalla D, Kondepu K, Valcarenghi L, Tornatore M, Mukherjee B (2017) 5G fronthaul-latency and jitter studies of CPRI over Ethernet. IEEE/OSA J Opt Commun Netw 9(2):172–182
5. Deb S, Monogioudis P, Miernik J, Seymour JP (2014) Algorithms for enhanced inter-cell interference coordination (eICIC) in LTE HetNets. IEEE/ACM Trans Netw 22(1):137–150

6. Ejlali M, Montazeri MA, Saidi H, Ghiasian A (2012) Design and implementation of a shared memory switch fabric. 6th International Symposium on Telecommunications (IST), Tehran, pp 721–727
7. Fehrenbach T, Datta R, Göktepe B, Wirth T, Hellge C (2018) URLLC services in 5G low latency enhancements for LTE. 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), pp 1–6. https://doi.org/10.1109/VTCFall.2018.8690663
8. GE fanuc intelligent platforms (2009) Switched Ethernet latency analysis [online]. http://www.gefanuc.com
9. Gotsis KA et al (2011) Beamforming in 3G and 4G mobile communications: the switched-beam approach. In: Maicas JP (ed) Recent developments in Mobile communications-a multidisciplinary approach. In Tech, Rijeka, pp 201–216
10. Grobe K, Roppelt M, Autenrieth A, Elbers J, Eiselt M (2011) Cost and energy consumption analysis of advanced WDM-PONs. IEEE Commun Mag 49(2):s25–s32
11. Ji H, Park S, Yeo J, Kim Y, Lee J, Shim B (2018) Ultra-reliable and low-latency communications in 5G downlink: physical layer aspects. IEEE Wirel Commun 25(3):124–130
12. Kwak K et al (2013) Adaptive and distributed CoMP scheduling in LTE-advanced systems. IEEE VTC
13. Li R, Zhao Z, Sun Q, Chih-Lin I, Yang C, Chen X, Zhao M, Zhang H (2018) Deep reinforcement learning for resource management in network slicing. IEEE Access 6:74429–74441. https://doi.org/10.1109/ACCESS.2018.2881964
14. Li Z, Shariatmadari H, Singh B, Uusitalo M (2018) 5G URLLC: design challenges and system concepts. In International symposium on wireless communication systems (ISWCS) [8491078] (International symposium on wireless communication systems). IEEE. https://doi.org/10.1109/ISWCS.2018.8491078
15. Lindbom L, Love R, Krishnamurthy S, Yao C, Miki N, Chandrasekhar V (2011) Enhanced inter-cell interference coordination for heterogeneous networks in LTE-advanced: a survey. Cornell University Library
16. Mehlführer C, Colom Ikuno J, Šimko M, Schwarz S, Wrulich M, Rupp M (2011) The Vienna LTE simulators - enabling reproducibility in wireless communications research. EURASIP J Adv Signal Process 2011:29
17. Michaloliakos A, Ao WC, Psounis K (2016) Joint user-beam selection for hybrid beam forming in asynchronously coordinated multi-cell networks. 2016 information theory and applications workshop (ITA), pp 1–10. https://doi.org/10.1109/ITA.2016.7888166
18. Necker MC (2007) Coordinated fractional frequency reuse. In: Proc. 10th ACM Symp. Modeling, Analysis Simulation Wireless Mobile Syst., New York
19. Neumann P, Pschmann A (2005) Ethernet-based real-time communication with PROFINET IO. WSEAS Trans Commun 4:235–245
20. Novlan T, Andrews J, Sohn I, Ganti R, Ghosh A (2010) Comparison of fractional frequency reuse approaches in the OFDMA cellular downlink. In: Proc. IEEE Globecom, Miami, Florida, pp 1–5
21. Ordonez-Lucena J, Chavarria JF, Contreras LM, Pastor A (2019) The use of 5G non-public networks to support Industry 4.0 scenarios. 2019 IEEE conference on standards for communications and networking (CSCN). GRANADA, Spain, pp 1–7. https://doi.org/10.1109/CSCN.2019.8931325
22. Popa L, Ratnasamy S, Iannaccone G, Krishnamurthy A, Stoica I (2010) A cost comparison of datacenter network architectures. In: Proceedings of the 6th international conference, Co-NEXT ‘10. ACM, New York, pp 16:1–16:12
23. Popescu DA (2019) Latency-driven performance in data centres. Ph.D. Dissertation. University of Cambridge
24. Rahman M, Yanikomeroglu H (2010) Enhancing cell-edge performance: a downlink dynamic interference avoidance scheme with inter-cell coordination. IEEE Trans Wirel Commun 9:1414–1425
25. Salah F, Kuru L, Jantti R (2018) Multi-TRxPs for industrial automation with 5G URLLC requirements. M.S. thesis, Electrical Engineering Department, Aalto University, Espoo, Finland [online]. Available: http://urn.fi/URN:NBN:fi:aalto-201812146549
26. Vierimaa O et al (2017) Cost modeling of cloud-based radio access network. https://aaltodoc.aalto.fi/
27. Wang J, Zhu H (2015) Beam allocation and performance evaluation in switched-beam based massive MIMO systems. In: 2015 IEEE Int. Conf. Commun. (ICC), London, pp 2387–2392
28. Wang K, Masmachuca C, Wosinska L, Urban PJ, Gavler A, Brunnstrom K, Chen J (2017) Techno-economic analysis of active optical network migration toward next-generation optical access. IEEE/OSA J Opt Commun Netw 9(4):327–341
29. Wang J, Weitzen J, Bayat O, Sevindik V, Li M (2019) Interference coordination for millimeter wave communications in 5G networks for performance optimization. J Wirel Commun Netw 2019:46. https://doi.org/10.1186/s13638-019-1368-6