Accelerating Mobile Applications with Parallel High-bandwidth and Low-latency Channels
Sentosa, William; Chandrasekaran, Balakrishnan; Brighten Godfrey, P.; Hassanieh, Haitham; Maggs, Bruce; Singla, Ankit

published in
HotMobile 2021
2021

DOI (link to publisher)
10.1145/3446382.3448357

document version
Publisher's PDF, also known as Version of record
document license
Article 25fa Dutch Copyright Act

Link to publication in VU Research Portal

citation for published version (APA)

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 21. Sep. 2023
Accelerating Mobile Applications With Parallel High-bandwidth and Low-latency Channels

William Sentosa¹, Balakrishnan Chandrasekaran², P. Brighten Godfrey¹,³, Haitham Hassanieh¹, Bruce Maggs⁴, Ankit Singla⁵
¹University of Illinois at Urbana-Champaign, ²VU Amsterdam and MPI-INF, ³VMWare, ⁴Duke University, Emerald Innovations, and MIT, ⁵ETH Zürich

ABSTRACT
Interactive mobile applications like web browsing and gaming are known to benefit significantly from low latency networking, as applications communicate with cloud servers and other users’ devices. Emerging mobile channel standards have not met these needs: general-purpose channels are greatly improving bandwidth but empirically offer little improvement for common latency-sensitive applications, and ultra-low-latency channels are targeted at only specific applications with very low bandwidth requirements.

We explore a different direction for wireless channel design: utilizing two channels – one high bandwidth, one low latency – simultaneously for general-purpose applications. With a focus on web browsing, we design fine-grained traffic steering heuristics that can be implemented in a shim layer of the host network stack, effectively exploiting the high bandwidth and low latency properties of both channels. In the special case of 5G’s channels, our experiments show that even though URLLC offers just 0.2% of the bandwidth of eMBB, the use of both channels in parallel can reduce page load time by 26% to 59% compared to delivering traffic exclusively on eMBB. We believe this approach may benefit applications in addition to web browsing, may offer service providers incentives to deploy low latency channels, and suggests a direction for the design of future wireless channels.

CCS CONCEPTS
• Networks → Mobile networks; Network architectures.

KEYWORDS
5G, eMBB, URLLC, Traffic steering, Parallel channels

1 INTRODUCTION
Low latency is critical to interactive applications such as web browsing, virtual and augmented reality, and cloud gaming. For web applications, even an increase of 100 ms latency can result in as much as 1% revenue loss, as noted by Amazon [11]. The emerging VR/AR and cloud gaming applications also rely on the low latency link to deliver a seamless user experience. For instance, VR requires 20 ms or lower latency to avoid any simulator sickness [10].

Current mobile networks, serving general Internet applications such as web browsing and video streaming, have not yet delivered consistent low latency performance. This is fundamentally challenging due to the inherent tradeoff between latency and bandwidth [14]. One approach is to provide two separate channels (or services) – one optimizing for bandwidth, the other optimizing for latency – with different types of user applications assigned to them. For example, 5G NR follows this pattern with its enhanced mobile broadband (eMBB) and ultra-reliable and low-latency communication (URLLC) channels. eMBB, which serves general-purpose Internet use, is heavily focused on delivering gigabit bandwidth. This channel will be useful for streaming media, but offers little improvement for latency-sensitive applications, such as web browsing. Experimentally, web page load time in existing 5G deployments, even in close-to-ideal circumstances (stationary device and little utilization), is similar to 4G for pages of size < 3 MB and about 19% faster than 4G for sites > 3 MB [24]; this discrepancy is due to 5G eMBB having 28 ms or larger latency, broadly similar to 4G.

Meanwhile, 5G URLLC promises an exciting capability of very low latency, in the range of 2 to 10 ms [3], but compromises severely on bandwidth, making it unsuitable for common mobile applications. Our experiments emulating web browsing (the most widely used mobile application [32], and far from the most bandwidth-intensive application) over URLLC with 2 Mbps bandwidth show web page load times would be 5.87× higher than with eMBB.

As the latency-bandwidth tradeoff is fundamental, we expect this separation between channels is likely to persist; 6G, for example, is also expected to include a low latency channel [38]. But we believe the availability of both a high bandwidth channel (HBC) and a low latency channel (LLC) offers an opportunity beyond simple static assignment of an application to a single channel. Our hypothesis is that by using high bandwidth and low latency channels in parallel on mobile devices, significant performance and user experience
improvements are possible for latency-sensitive applications. Here, we explore this hypothesis for the case of web browsing.

Designing a way of mapping an application’s traffic to HBC and LLC is difficult since we have to use LLC’s bandwidth very selectively. Indeed, the main deployed mechanism to combine multiple channels, MPTCP [30], assumes two interfaces that are each of significant bandwidth, with the goal of aggregating that bandwidth or supporting failover. LLC’s bandwidth, however, is a rounding error compared to HBC’s. Other works – particularly Socket Intents [31] and TAPS [27] – exploit multi-access connectivity through application-level input, which we prefer to avoid to ease deployment and expand relevance to other applications in the future; therefore we expect new mechanisms are necessary.

To solve these problems, one approach would be to steer traffic at the granularity of flows or web objects (e.g., large vs. small). To be as selective as possible in our use of LLC, we design a fine-grained packet-level steering scheme. The scheme uses simple heuristics to identify packets that are more likely to help the application if they are accelerated, while using a very limited amount of state to minimize the out-of-order delivery that can result from accelerating only a subset of packets in a flow. Since this steering scheme does not solicit extra information from applications, we can present a virtual network interface to the application, and transparently steer each packet sent to that interface over HBC or LLC.

To evaluate our design with a concrete scenario, we leverage 5G’s eMBB and URLLC as our HBC and LLC. We utilize an emulated environment that includes a typical mobile web browsing scenario: a client with two network channels emulating characteristics of eMBB and URLLC radio access networks and the cellular core, as well as CDN edge, parent, and origin servers. We populate this setup with landing page snapshots of 53 popular web sites and test various assumptions about eMBB and URLLC performance. We evaluate several object and packet steering schemes, and prior approaches such as MPTCP [1] and ASAP [21]. Our findings conclude that packet-granularity schemes, which require little to no per-connection state and no application knowledge, provide the best performance. For the best steering scheme, even with a modest bandwidth of 2 Mbps allocated to URLLC, we can reduce mean page load time (PLT) by 26% compared to exclusively using eMBB. If eMBB’s latency fluctuates, increasing from 30 ms to 150 ms (which is typical for mmWave channels [15, 19]), combining eMBB and URLLC decreases mean PLT by 59%.

Finally, we discuss deployment strategies, challenges, and future opportunities. Although we have focused on web browsing in this short paper, we believe our basic techniques can apply to a wide variety of latency-sensitive applications, and open new opportunities for app developers and cellular providers.

2 BACKGROUND & RELATED WORK

Web browsing traffic: Web browsing traffic is characterized by its often short and bursty flows, in which latency matters more than bandwidth. A web page easily contains tens to hundreds of relatively small-sized objects distributed across multiple servers and domains. A study across Alexa Top 200 pages found that the median number of objects in a page is 30, while the median object size is 17 KB [36]. This translates to many HTTP request-and-response interactions across many short flows. Page load time (PLT) is thus typically dominated by DNS lookup, connection establishment, and TCP convergence time – which require little throughput but are highly dependent on RTT. Increasing TCP throughput beyond ≈ 16 Mbps offers little improvement in PLT [33].

Channels in 5G: The 5G NR (New Radio) standard provides differentiated services to support diverse applications. (1) Enhanced mobile broadband (eMBB) is an HBC that serves standard mobile Internet applications. It is considered to be an upgraded 4G mobile broadband with higher data rates. A key enabling technology is the use of mmWave bands which offer high bandwidth. A recent measurement study on commercial mmWave 5G networks in the US shows TCP throughput of up to 2 Gbps for download and 60 Mbps for upload, with a mean RTT of 28 ms measured between the client and the first-hop edge server right outside the cellular network core [24]. (2) Ultra-reliable low-latency communication (URLLC), as an LLC, is intended for mission-critical and specialized latency-sensitive applications such as self-driving cars, factory automation, and remote surgery. It focuses on providing highly reliable, very low latency communication at the cost of limited throughput. While URLLC is yet to be deployed, the standard is finalized and specifies a 0.5 ms air latency (1 ms RTT) between the client and the Radio Access Network (RAN) with 99.999% reliability for small packets (e.g. 32 to 250 bytes) [6]. It also specifies a target end-to-end latency (from a client to a destination typically right outside the cellular network core) between 2 to 10 ms with throughput in the range of 4 Mbps [3]. URLLC is expected to guarantee such performance through dedicated network resources achieved via network slicing [3].

Leveraging multiple connections: MPTCP [30] provides an end-to-end transport protocol that supports multiple interfaces. Socket Intents [31] and Intentional networking [20] both expose custom API to application and offer OS-level support for managing multiple interfaces. Both of them regulate application traffic based on the application-specific information. These efforts are complimentary to ours since we focus on mapping the application traffic to HBC and LLC. Perhaps, works most relevant to ours are IANS [12] and ASAP [21]. IANS leverages Socket Intents to improve web page load. Unlike our approach (i.e., packet steering schemes), however, they require application-specific information (i.e., object size) to determine which interface to use and they work at the object level. We compare ASAP with our scheme in §4.2, and found that our scheme performs better due to its finer-grained decision.

3 TRAFFIC SPLITTING SCHEMES

In general, a scheme to split traffic between HBC and LLC needs to judiciously steer only a small portion of traffic through LLC to avoid overloading it (which would congest queues, negating latency benefits); and that small portion should yield the most benefit to application-level performance.

We investigate heuristics for web browsing in object-level and packet-level which employ different tradeoffs. Object-level splitting...
may benefit from application-level knowledge about web objects, but requires application input. Packet-level splitting can benefit from finer-grained decisions, but needs to deal with challenges of out-of-order delivery (OOD) if packets within a flow are sent on different channels. We consider several natural schemes in each class, and design a new packet-level scheme that minimizes OOD.

**Object-level splitting:** A web page is comprised of many objects that vary in size. Inspired by prior works [12, 34], the simplest splitting heuristic is to request objects smaller than some threshold over LLC, as their load times are dominated by latency. All other objects use HBC. Note that this approach requires the application to specify object size, or to make the splitting choice itself. Our evaluation (with URLLC and eMBB) tests this *obj-size* scheme using two different thresholds of 3 KB and 6 KB. (The reasoning is that URLLC supports a maximum rate of ≈250 bytes per ms, and its RTT difference with eMBB is ≈25 ms. Therefore, objects of size >6.25 KB are likely delivered faster on eMBB.)

Another approach is to leverage application-level priority of web objects. Web pages have complex dependency structures, and certain objects can be on the critical path for web page loading. We evaluated a scheme (*obj-priority*) by utilizing Chrome priority assignment [17] to only fetch the *highest* priority object via LLC. Note that this scheme may not be a perfect fit either, because high priority objects could be large (HTML pages, which tend to have the highest priority, can be much larger than 6.25 KB) and will load faster via HBC.

**Packet-level splitting:** A simple packet-level heuristic is to send all client-to-server traffic over LLC and all server-to-client traffic over HBC: we call this scheme *pkt-uptlink*. The intuition is that client requests are generally small. This scheme might not perform well, however, if the client sends a large amount of data to the server (e.g., using HTTP POST), and misses acceleration opportunities from server to client.

To devise a better heuristic, we investigate the traffic that arises when loading a typical web page. Before loading a page, the client exchanges small UDP packets (<250 bytes) with the nameserver to perform DNS lookup; delivering them on URLLC can accelerate the lookup process. The client continues by exchanging TCP packets with the web server as shown in Figure 1. Packets in the TCP flow can be categorized as *control packets* and *data packets*. *Control packets* are packets only intended to manage TCP operations (e.g., SYN and ACK packets), and do not carry any payload. These packets are tiny, and accelerating them will benefit the entire flow. For instance, accelerating SYN packets allows faster connection establishment while accelerating the client’s ACK packets reduces flow RTT, which is crucial because RTT largely determines web object download times for small objects. Thus, we include a scheme (*pkt-control*) that only offloads control packets to LLC in our evaluation.

*Data packets* are packets that carry a payload. Mapping *data packets* to LLC can quickly consume the bandwidth, yet each packet fits into a single packet. The server also sends FIN packet to tear down the connection.

Figure 1: HTTP traffic delivery on TCP. It also shows per-packet mapping done by the packet steering scheme (*pkt-state*).
Algorithm 1: PKT-State

Result: Send a packet to either HBC or LLC
if size(pkt) < MTU and size(prev_pkt) < MTU then
    send(pkt, LLC); /* control or small data pkt */
else
    send(pkt, HBC); /* tail or MTU-sized pkt */
end

Figure 2: Baseline emulation setup

4 EVALUATION

We evaluate performance with emulated 5G eMBB and URLLC, and with parameters representing possible future channels.

4.1 Emulation and experimental setup

Our emulation setup (Figure 2) reflects the dependence of web applications on content delivery networks (CDNs). CDNs comprise a hierarchy of caches [9]. Here, the client’s request first reaches an edge server. In case of a cache miss it is forwarded to a parent server, and in case of a second cache miss it is forwarded to the origin server storing the content. Request latency is therefore variable.

The client has two parallel paths (along eMBB and URLLC) to the edge server. The bandwidth and RTT of the eMBB path (1500 Mbps + 30 ms) is based on recent 5G measurements [24], and for URLLC (2 Mbps + 5 ms) on its sample use cases [3]. This base eMBB configuration is based on perfect line of sight with no mobility. For the CDN latency, we assume that both the edge and parent servers are located in the US midwest while the origin server is located on the US west coast. Based on publicly available network latency information [5], we use 10 ms RTT from edge to parent and 40 ms RTT from parent to origin.

Our experiments use a machine running Ubuntu 16.04 LTS with 32 GB RAM, 512 GB SSD, and an Intel Core i7-9700 CPU. The web server runs on Apache 2.4.18, while the client uses Chromium v81. The experiments use HTTP 1.1, leaving an evaluation of HTTP/2 and QUIC to future work.

We use Mahimahi [25] to record and replay web pages. We modified Mahimahi to include the CDN hierarchy, our packet steering techniques in Mahimahi’s packet scheduling modules, and an HTTP proxy to enable object-level steering.

To realistically emulate cache misses and hits in the CDN, our evaluation uses web pages for which we can get this information from HTTP headers. HTTP pragma headers, used by CDNs for debugging, can be used to glean this information [29]. We started with a random sample of top-ranked Alexa pages and found 53 that provide the pragma headers. We recorded and locally replayed these sites’ landing pages, whose mean size is 2.13 MB and whose 95th percentile size is 6.13 MB. We replayed each page load five times and reported its median PLT (based on the onLoad event [26]). We used cold DNS and browser caches for all experiments.

4.2 Evaluation of multiple schemes

We evaluated three object-level and three packet-level steering schemes (§3), and compared them with a baseline where the client only utilized eMBB (all-embb), and with Multipath TCP (MPTCP) [1] and ASAP [21]. MPTCP spreads traffic across multiple interfaces; we evaluated the default Linux kernel implementation. ASAP was designed to accelerate PLT in satellite networks by offloading some traffic to 3G/4G (which had lower latency but higher cost per byte than satellites, in [21]). Figure 3 and Table 1 summarize the results.

Object-level splitting performs poorly. Obj-Priority is often worse than the baseline since it sends high priority objects to the low-latency channel, but these can be large. The size-based heuristics do not improve much compared to the baseline – only about 25% of web pages have > 10% lower PLT – and they are sometimes worse. This is because many small objects (like icons) do not affect PLT much. Some pages perform slightly worse because of resource contention on URLLC when fetching multiple small objects in parallel.

Packet-level splitting, in contrast, performs consistently well, especially pkt-state, which has 1243 ms mean PLT (a 26% improvement) while only offloading 12.6% of bytes to URLLC. Interestingly, pkt-control improves mean PLT by 18.5% by only offloading control packets comprising 3.1% of bytes. These improvements stem from the finer granularity of splitting, such that every web object experiences faster download, ultimately improving PLT for all pages. Unsurprisingly, the smaller web pages tend to improve more, as their PLT is more dominated by latency than bandwidth.

Understanding pkt-state gains: To understand the network-related gain of pkt-state, we use curl to download the HTML document of amazon.com within our emulation setup. This object is cached at the edge, so no CDN delay is incurred. Table 2 shows that pkt-state performs better than not only all-embb, but also all-urllc. This is because pkt-state offloads (1) DNS packets that reduce DNS lookup time, (2) control packets that reduce connection (SYN packets) and object transfer time (Client’s ACK packets), and (3) small data packets that accelerate small data transfer such as TLS client key transfer or HTTP request data. Meanwhile, it benefits from the higher bandwidth than all-urllc which contributes to a better transfer time.

Comparison with existing works: MPTCP generally fails to perform better than always using a single eMBB path. Vanilla MPTCP is known to perform sub-optimally under heterogeneous paths due to head-of-line blocking [13] and its policy to lean towards load balancing among multiple paths. A superior MPTCP scheduler may
Figure 4: Mean load time across different pages: the landing page (LP) and four internal pages (P1, P2, P3, P4)

Table 1: Traffic sent over URLLC by each scheme, in bytes per page, percent of total bytes of page load, and mean upload (U) and download (D) rate.

<table>
<thead>
<tr>
<th>Steering scheme</th>
<th>Mean bytes per page</th>
<th>% bytes</th>
<th>Avg rate (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean 90th</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPTCP [1]</td>
<td>476.3 KB</td>
<td>22.1%</td>
<td>41.6%</td>
</tr>
<tr>
<td></td>
<td>U: 0.15</td>
<td>D: 0.68</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 0.24</td>
<td>D: 1.02</td>
<td></td>
</tr>
<tr>
<td>OBJ-Priority</td>
<td>259.6 KB</td>
<td>12.8%</td>
<td>29%</td>
</tr>
<tr>
<td></td>
<td>U: 0.10</td>
<td>D: 1.13</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 0.14</td>
<td>D: 1.75</td>
<td></td>
</tr>
<tr>
<td>OBJ-Size-3KB</td>
<td>118.1 KB</td>
<td>5.5%</td>
<td>11%</td>
</tr>
<tr>
<td></td>
<td>U: 0.21</td>
<td>D: 0.45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 0.42</td>
<td>D: 0.82</td>
<td></td>
</tr>
<tr>
<td>OBJ-Size-6KB</td>
<td>242.8 KB</td>
<td>10.7%</td>
<td>20.3%</td>
</tr>
<tr>
<td></td>
<td>U: 0.28</td>
<td>D: 0.73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 0.46</td>
<td>D: 1.32</td>
<td></td>
</tr>
<tr>
<td>PKT-Uplink</td>
<td>186.7 KB</td>
<td>8.6%</td>
<td>13.5%</td>
</tr>
<tr>
<td></td>
<td>U: 0.38</td>
<td>D: 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 0.66</td>
<td>D: 0</td>
<td></td>
</tr>
<tr>
<td>PKT-Control</td>
<td>62.8 KB</td>
<td>3.1%</td>
<td>4.6%</td>
</tr>
<tr>
<td></td>
<td>U: 0.1</td>
<td>D: 0.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 0.17</td>
<td>D: 0.07</td>
<td></td>
</tr>
<tr>
<td>ASAP [21]</td>
<td>257 KB</td>
<td>12.3%</td>
<td>20.3%</td>
</tr>
<tr>
<td></td>
<td>U: 0.85</td>
<td>D: 0.50</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 1.26</td>
<td>D: 0.77</td>
<td></td>
</tr>
<tr>
<td>PKT-State</td>
<td>260.5 KB</td>
<td>12.6%</td>
<td>20.4%</td>
</tr>
<tr>
<td></td>
<td>U: 0.87</td>
<td>D: 0.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>U: 1.28</td>
<td>D: 0.81</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: A deeper look at the pkt-state performance when fetching the landing page of amazon.com (110 KB).

Perf. metric | All-eMBB | All-URLLC | PKT-State |
-------------|----------|-----------|-----------|
DNS lookup   | 60 ms    | 12 ms     | 12 ms     |
TCP connect  | 32 ms    | 6 ms      | 6 ms      |
TLS connect  | 180 ms   | 94 ms     | 92 ms     |
Object transfer | 92 ms | 427 ms    | 54 ms     |
Total loading time | 399 ms | 579 ms    | 183 ms    |
Avg. download rate | 2.2 Mbps | 1.5 Mbps  | 4.8 Mbps  |

Figure 5: Varying eMBB and URLLC network conditions, conducted using the baseline setup (Figure 2) with a single varying parameter, except (c) where we change two parameters.

4.3 Varying network performance

We evaluated the best heuristic (pkt-state) under different LLC and HBC network characteristics.

Varying eMBB network RTTs: RTT inflation is common in mobile networks due to bufferboat [18, 22] and physical properties, especially under movement. We evaluated inflated RTT on the HBC, and found pkt-state provides even larger improvements (Figure 5a). pkt-state has 59% ($\approx 2.4x$) lower PLT than all-eMBB when the eMBB RTT is 150 ms, which is possible [24] when the channel is heavily utilized. Since pkt-state uses eMBB primarily for bandwidth-intensive traffic, it is extremely robust to worsened RTT on the main channel.

Varying URLLC network RTTs: Although URLLC is expected to deliver consistent low latency, we evaluate latency inflation on URLLC to reflect scenarios where web traffic is de-prioritized in favor of critical traffic. pkt-state is still superior with URLLC latency increasing up to 20 ms and eMBB held at 30 ms (Figure 5b). To prevent worse performance, the client could periodically probe the LLC and if its latency is too high, fall back to using the HBC.

Varying URLLC bandwidth: pkt-state requires little bandwidth: its PLT flattens for more than 2 Mbps (Figure 5c). This is expected since its average data rate is only 0.87 Mbps up and 0.52 Mbps down (Table 1). The figure also shows we cannot entirely rely on URLLC (all-urllc) as it needs more than 10 Mbps to approach the baseline.
5 DEPLOYMENT IN 5G

We discuss the packet steering deployment in 5G eMBB and URLLC and its challenges. They are all parts of the future work and require further evaluation; however, we outline our expectations.

Packet steering implementation: In the deployment, two specific changes are required: packet steering in the user device operating system, and in an existing proxy in the network provider. The steering should be performed in the network-layer, and hence we do not expect to modify both eMBB and URLLC link- and physical-layer design. On the client-side, the operating system needs to steer packets over eMBB and URLLC interfaces. One approach is to present a virtual interface that ties both interfaces and steers packets via the custom bridging [37]. Applications wishing to utilize parallel channels may use the provided virtual interface. Another way is to leverage MPTCP and introduce a new scheduling module. The former approach may be preferred because MPTCP also requires MPTCP-compliant web servers, and it cannot steer non-MPTCP traffic. We do not need any or minimal (such as to use the virtual interface) application-level change as our approach is transparent to the application.

Since traffic steering should be performed in both directions, network providers need to steer packets originating from the Internet and merge URLLC and eMBB traffic from the client. This change should be minimal since they can directly leverage an existing proxy in the core network, such as the packet data network gateway (P-GW). This approach ensures transparency of the packet steering to the server.

URLLC scalability: The number of users that can send general traffic to URLLC is an important matter which deserves to be evaluated quantitatively in the future. At the time of the writing, URLLC is not yet deployed in public. However, based on the white paper [4], URLLC is targeted to support a relatively high connection density with modest per-user bandwidth. For instance, one of the URLLC use cases (discrete automation) requires a user-experienced data rate of 10 Mbps, traffic density of 1 Tbps/km2, connection density of 100,000/km2, and max end-to-end latency of 10ms. We expect URLLC to reserve 2 Mbps maximum bandwidth per user for general application traffic, which is still reasonable based on others’ proposed use cases for URLLC, even in a dense urban area.

Disrupting URLLC native traffic: URLLC is primarily built to serve latency-sensitive critical applications. To ensure we do not compromise the performance of these applications, we only use a small amount of the URLLC capacity. In particular, the 90th percentile average data rate is only 1.28 Mbps (Table 1). Moreover, the network operator can limit the per-user bandwidth and choose to deprioritize non-critical packets as our approach does not require 99.999% reliability and is resilient to the slight increase in URLLC latency (§4.3).

Resource contention among applications: Multiple applications inside a user equipment may compete to use URLLC. We can regulate them using prioritization. One simple approach is to prioritize applications running in the foreground since mobile phone users are typically single-tasking.

Incentives for operators: While URLLC targets critical applications, it is up to the network providers to open URLLC for general mobile applications like web browsing. This is possible as 5G chipsets are typically designed to support multiple bands including the sub-6GHz bands for URLLC [3]. Expanding the applications of URLLC can encourage providers to foster a faster and broader deployment of URLLC as it brings a smoother experience to their major customers – mobile phone users; especially as the current market for URLLC applications like self-driving cars and remote surgery is still in its infancy.

6 RESEARCH OPPORTUNITIES

The use of LLC and HBC opens up new research opportunities.

Other applications: In addition to web browsing, our approach can benefit many mobile applications which typically employ HTTP but are different in terms of page contents. Their traffic is generally smaller since it is characterized by many JSON requests [35] and hence, it should benefit more from the improved latency compared to bandwidth. LLC and HBC combination can also properly support applications from different domains that require high-bandwidth and low-latency; something that cannot be satisfied by utilizing a mere single channel. For instance, cloud gaming, which allows users to play games from remote servers, requires high bandwidth to stream the content and low latency to remain responsive to user input. Since these applications can be vastly different than web browsing, a superior steering scheme may exist. We plan to analyze them further to determine an effective way of leveraging LLC and HBC.

Beyond mobile networks: Our insights may apply to other LLC and HBC combinations with analogous bandwidth and latency trade-offs. Examples include quality of service (QoS) differentiation providing separate latency- and bandwidth-optimized services [7, 28]; and routing traffic among multiple ISPs where one is more expensive but provides better latency, as may happen with very low Earth orbit satellite-based [16] or specialty [8] ISPs. To achieve the optimum cost to performance ratio, we can route only the latency-sensitive traffic to the low-latency ISP.

Future wireless design: The 5G URLLC is only equipped with limited user bandwidth, and hence it is not suitable to serve general application traffic. The bandwidth is severely compromised because it needs to provide both low latency and very high reliability (99.999%). However, general applications do not need the almost-perfect reliability that URLLC guarantees. Future wireless networks (such as 6G) may reconsider this trade-off and provide a low-latency channel with somewhat greater bandwidth and somewhat lower reliability.

7 CONCLUSION

We present an idea to utilize high-bandwidth and low-latency channels in parallel to improve mobile applications. In this early work, we show the way web browsing performance can be improved by carefully steering traffic over 5G eMBB and URLLC and exploiting the channels’ massive bandwidth and low latency properties.

ACKNOWLEDGMENTS

We thank the anonymous reviewers and our shepherd Wenjun Hu for their comments. This work was supported by the National Science Foundation under Grant No. 1763841.
REFERENCES


