Impact of Bandwidth Part (BWP) Switching on 5G NR System Performance

Fuad Abinader, Andrea Marcano, Karol Schober, Riikka Nurminen, Tero Henttonen, Hisashi Onozawa, and Elena Virtej

Abstract—Bandwidth Parts (BWPs) is a 5G NR feature introduced in 3GPP Release 15 for dynamically adapting the carrier bandwidth and numerology in which a UE operates. BWP allows supporting multiple services per carrier, e.g., eMBB and mMTC. Although BWP enables higher spectrum flexibility and power savings, the effect of delays such as BWP Inactivity Timer and BWP Switch Delay has not been thoroughly studied and understood. This paper presents a system-level evaluation of a 5G NR deployment with dynamic BWP adaptation. As expected, results indicate no impact on performance under higher load; on the other hand, for low load and bursty traffic, BWP switching is more frequent, enabling the potential for power savings at the cost of increased latency and decreased throughput.

Index Terms—5G NR, Bandwidth Parts, BWP adaptation.

I. INTRODUCTION

ONE of the goals of 5G New Radio (NR) is introducing flexible carrier size to enable spectrum flexibility, i.e., the capability to support UEs utilizing a carrier bandwidth (BW) smaller than the carrier BW utilized in the network cell, as well different numerologies, i.e., subcarrier spacing (SCS) [1].

Spectrum flexibility in 5G NR opens a number of novel research challenges [2], including (a) allocation of carrier bandwidth and numerology, (b) analysis of non-orthogonality of resource elements from different numerologies, (c) inter-numerology interference (INI) estimation and mitigation, and (d) radio resource management (RRM) aspects related to scheduling. For instance, [1] explores the problem of allocating resources to efficiently support multiple numerologies in the same TDD carrier, and provides closed-form expressions for the numerology sub-band configuration and the DL-UL duplexing ratio per sub-band; moreover, it demonstrates that the flexibility in spectrum usage improves throughput and delay performance. In [3], it is demonstrated that defining numerologies with shorter TTIs allows reducing TCP roundtrip time (RTT), with a positive impact on throughput performance. Authors in [4] propose a solution for the problem of inter-cell interference coordination (ICIC) in neighboring gNodeBs (gNBs) with different numerologies.

5G NR increases peak and user-perceived data rates in comparison to previous generations (e.g., LTE) by supporting carriers of up to 800MHz wide, at the cost of increased power consumption from RF and baseband signal processing [5]. In addition to enhanced Mobile BroadBand (eMBB) requirements, 5G NR scenarios comprise Ultra Reliable Low Latency Communication (URLLC) and massive Machine Type Communication (mMTC), both introducing new carrier requirements. For instance, supporting a massive amount of UEs in mMTC scenarios may require configuring narrowband carriers to reduce UE power consumption; on the other hand, URLLC scenarios require shorter latency, achieved for instance via the use of wider SCSs.

To provide a unified solution for numerology and carrier BW flexibility, 3GPP Release 15 (Rel-15) introduced a new feature called Bandwidth Part (BWP) for 5G NR [6]. BWPs enable UEs to be configured to operate in BWs that are narrower than the carrier BW, using customized numerologies and BW sizes fitting the service requirements in terms of throughput, delay and energy efficiency. In previous 3GPP systems, UEs had to monitor the entire carrier BW for control signalling. With 5G NR, the UE is not required to transmit or receive any signal (not even control) outside the frequency range of the active BW, which might enable power savings in some scenarios [5], when e.g. (a) RF-baseband processing operates with lower sampling rate for certain numerologies, (b) reduced baseband processing for narrower BW, and (c) bandwidth adaptation to traffic demand.

In Rel-15, BWP is defined as an essential part of 5G NR, from the very first NR initial access up to the continuous BWP adaptation through different carrier bandwidth configurations. However, literature is very scarce on the influence of BWP adaptation parameters, such as BWP inactivity timer and BWP switching delay, on system-level throughput and UE performance metrics (e.g., power savings, impact on traffic latency). Understanding the overall effect of these parameters has primary relevance due to the central role of BWPs in 5G NR. This paper is the first assessment on these above issues. We discuss the challenges in BWP adaptation via BWP switching and provide a performance evaluation of different BWP parameter configurations via system-level simulation.

The rest of the paper divides as follows. In Section II, we review Rel-15 BWP and discuss BWP adaptation via BWP switching and how it might affect 5G NR system-level and UE performance. In Section III, we describe the methodology used for a system-level assessment of the impact of BWP adaptation on 5G NR performance. In Section IV, we present and discuss the results of the simulations, considering key performance indicators such as application layer throughput, scheduling delay and Dedicated/Default BWP ratio. Finally, in Section V we draw conclusions from this study on dynamic BWP adaptation, and provide guidelines for the next steps.
II. BWP IN 3GPP 5G NR RELEASE 15

5G NR supports wide carrier bandwidths, up to 200 MHz for Frequency Range 1 (FR1, i.e. sub 6 GHz) and up to 400 MHz for Frequency Range 2 (FR2, i.e. 24 - 52 GHz). To support this, BWP was introduced in 3GPP Rel-15 to allow receiver-side bandwidth adaptation and it constitutes an essential part of 5G NR access interface. BWPs are defined in 3GPP TS 38.300 [7] (sec. 6.10) and 3GPP TS 38.211 [8] (sec. 4.4.5).

A BWP is a contiguous BW partition on a carrier in a serving cell that uses some given numerology (i.e., a subcarrier spacing and a cyclic prefix overhead). Each BWP consists of a group of contiguous physical resource blocks (PRB) that share some common numerology and are configured by a gNB to a UE according to its needs. BWP sizes can vary from 24 to 275 PRBs (4k FFT), and up to 4 DL BWPs and 4 UL BWPs can be configured to a UE on a serving cell. Only one DL and one UL BWP can be active at a given time in one serving cell. Configured BWPs cannot be larger than the maximum BW supported by the UE, and the UE is not expected to receive or transmit signals outside the active BWP, except for inter-frequency measurement gaps configured by the network. Therefore, the scheduler must constrain resource allocation for control and data within the UEs active BWP. BWPs allow multiplexing narrowband and wideband devices, as well as different numerologies at the same time. Five transmission numerologies are defined in Rel-15, as shown in Table I ([7], Table 5.1-1), where $\Delta f$ is the SCS. FR1 supports numerologies $\mu=\{0,1,2\}$, while FR2 supports numerologies $\mu=\{2,3\}$.

<table>
<thead>
<tr>
<th>$\Delta f$</th>
<th>$2\mu \times 15$ [kHz]</th>
<th>CP</th>
<th>Data</th>
<th>S ync</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15</td>
<td>Normal</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>Normal</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>60</td>
<td>Normal, Extended</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
<td>Normal</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>4</td>
<td>240</td>
<td>Normal</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

TABLE (I) Supported numerologies for 5G NR.

Figure 1 presents how a BWP setup translates into actual frequency domain resources within the carrier BW. For instance, the frequency reference point defined in Rel-15 for aligning PRB resource grids of carriers with different SCS in a serving cell is called Point A (expressed in ARFCN). Point A serves as a reference with respect to which a carrier (set of usable PRBs) is defined. System Information Block type 1 (SIB1) broadcasts the location of Point A, which can be located outside the channel band. Point A sets the start of SCS-nested common resource block (CRB) grid in a serving cell. A CRB consists of a set of 12 consecutive subcarriers; for each value of a CRB grid is defined, where CRBs are numbered from 0 upwards in the frequency domain.

Sub-carrier 0 (SC0) of CRB 0 across all numerologies are aligned, and PRBs are, as such, nested. Since some CRBs might be located outside channel BW or overlap with channel guard bands, the first usable resource block for some given numerology might not coincide with the first CRB. A parameter, named offsetToCarrier (numerology-dependent) indicates the offset from SC0 in CRB 0 to the lowest usable subcarrier on the actual carrier BW. offsetToCarrier is broadcasted to all UEs, allowing them to define the carrier location and width.

Finally, it is relevant to note that the frequency domain resource allocation is not indicated based on the CRB grid of a carrier. Instead, it is indicated based on the PRB grid of the scheduled BWP. PRBs of a BWP are numbered from 0 and upwards in the frequency domain, with PRB 0 indicating the first PRB of the BWP. The gNB signals the start and size/length of a BWP in the frequency domain in the form of a Resource Indicator Value (RIV) relative to the start of the carrier for a given SCS. Rel-15 defines a direct mapping between the PRBs and the CRBs, so that the user can translate its allocation in PRBs inside the BWP into CRBs in the corresponding numerology grid, and finally into the frequency BW subset in use. There are four types of BWPs defined in Rel-15 specifications, listed below:

- **Initial BWP**: common to all UEs; broadcasted in System Information (SI) to be used for initial access, until UE receives BWP cell configuration. Possible sizes are 24, 48, or 96 PRBs;
- **First active BWP**: a BWP activated upon Radio Resource Control (RRC) (re)configuration or MAC-activation of a Secondary Cell (SCell);
- **Default BWP**: BWP activated upon the expiration of the BWP Inactivity Timer. Default BWP can occupy the same PRBs as the Initial BWP, and UEs are expected to stay in Default BWP until traffic demands increase;
- **Dedicated BWP**: regular BWP configured in a dedicated manner; usually is wider than Default BWP, as to allow transmission of higher traffic loads;

A. BWP adaptation via switching triggers

BWP switching is a procedure that simultaneously activate an inactive BWP (e.g. Dedicated BWP) while deactivate an active BWP (e.g. Default BWP). BWP switching can be triggered via Downlink Control Information (DCI), Radio Resource Control (RRC) signaling, BWP inactivity timer expiration, or
by MAC entity upon initiation of random access (RA) procedure. RRC/MAC BWP switching allows configuring a new BWP to be activated as well as activating an already configured BWP. Switching the BWPs using DCI command allows activating pre-configured BWPs, which enables faster switching. Another way of triggering BWP switching is through a data inactivity timer. bw-p-InactivityTimer is (re)started to a default value when one of the following conditions occurs in a BWP other than the Default BWP: (a) the UE is in the process of transmission and/or just received an assignment in PDCCH with no ongoing random access process, or (b) a BWP switch by means of DCI is received. A switch to the default BWP is triggered upon expiry of bw-p-InactivityTimer. It is up to the scheduler to configure the default values of bw-p-InactivityTimer, and default values are within the range of 2-150ms There are five UE capability categories, listed in Table II, each defining the number of configured BWPs, support for different numerologies and whether the initial BWP must overlap with a cell-defining SSBs.

**Table II** RAN1 UE capability categories.

<table>
<thead>
<tr>
<th>UE Cat</th>
<th># of BWPs</th>
<th>Inactivity Timer</th>
<th>DCI Switch</th>
<th>BWPs with SCS</th>
<th>Initial BWP SSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-1</td>
<td>1</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>6-2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-3</td>
<td>4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>6-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6-1a</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>No</td>
</tr>
</tbody>
</table>

3GPP TS 38.133 [9] (section 8.6) defines requirements for BWP switch delay, i.e. time during which the UE is required to complete the switch from the original BWP to the new BWP. During BWP switch delay the UE is not required to transmit UL signals or receive DL signals on the cell where BWP is switched. The starting time of BWP switch delay for DCI-based BWP switch is the slot where the UE receives BWP switching request. For timer-based BWP switch, the starting time of BWP switching is the slot at the beginning of a subframe (FR1) or half-subframe (FR2) immediately after bw-p-InactivityTimer expires. Finally, for RRC-based BWP switch, the starting time of BWP switching is the last slot containing the RRC command including BWP switch request. The ending time for RRC-based and timer-based BWP switching is defined as the first slot where the UE can receive PDSCH (for DL active BWP switch) or transmit PUSCH (for UL active BWP switch) on the new BWP. In case of DCI-based BWP switching, the first slot is the one indicated for PDSCH reception or PUSCH transmission.

The minimum duration of BWP switch delay depends on the BWP switching scenario (DCI, timer or RRC), the UE type (Type 1 or Type 2 UE) and the SCS. For DCI-based BWP switch the delay is as defined in Table III. For RRC-based BWP switch the delay is not yet defined, but it will consist of 10ms RRC processing delay ([10], section 12) plus a BWP switch delay, longer than the delay defined in Table 3 for DCI-based and timer-based BWP switch. In addition to the delay in the switched cell, the UE is also allowed a shorter interruption on other serving cells due to RF returning.

Figure 2 illustrates a BWP switch triggered by inactivity timer, with 2ms bw-p-InactivityTimer and 3ms switch delay.

![Fig. (2) Mapping BWP into frequency domain resources.](image)

Depending on the NR carrier aggregation or dual connectivity scenario, the interruption may occur on NR and/or E-UTRA cells. For E-UTRA cells, the interruption duration is 1-2 subframes as defined in 3GPP TS 36.133 [11] (section 7.32). For NR cells, 3GPP TS 38.133 [9] (section 8.2) defines the interruption duration to 1 slot for SCSs of 15kHz and 30kHz, 3 slots for 60kHz SCS and 5 slots for 120kHz SCS.

**III. BWP Evaluation Methodology**

We conducted an evaluation campaign via system-level simulations to provide an assessment on the influence of BWP adaptation in 5G NR performance. Simulations were carried using Nokia proprietary fully dynamic system-level simulator with both LTE and 5G NR capabilities [12], using OFDM symbol level resolution in time and subcarrier resolution in frequency [13], [14]. We use Exponential Effective SINR Mapping (EESM) as link-to-system interface [14]. We assume 2x2 Multiple Input Multiple Output (MIMO) transmission for radio links between UEs and macro cells, and their channel model follows the Urban Macro model as per [15]. Proportional Fair scheduling is used for time and frequency domains, with up to 5 scheduled users per TTI. Outer-loop link adaptation controls target Block Error Rate (BLER) for the first transmission by selecting Modulation and Coding Scheme (MCS) based on Channel Quality Indicator (CQI) measurements. Asynchronous chase combining Hybrid ARQ (HARQ) with six stop-and-wait processes are used.

The simulation scenario consists of 7 tri-sectorized macro sites (i.e. 21 macro cells), distributed over a non-overlapping hexagonal grid with 500m Inter-Site Distance (ISD). The wide-area 5G NR macro cells operate at a central frequency of 3.5GHz. We consider one component carrier (CC) per cell, each having 20MHz of carrier BW and 15kHz of SCS operating in Time-Division Duplex (TDD) mode. We set maximum transmission power for gNBs and UEs to 43 dBm and 23dBm, respectively. UE location is randomly generated.
at the beginning of the simulation inside the area of the 21 macro cells. A total of 105 UEs calls are created (average of 5 calls per cell). Calls use FTP3 traffic generation, with PDU generation following a Poisson process with a configurable average PDU generation rate.

For BWP adaptation evaluation, we assumed that UE category is 6-2 [9]. UEs are configured with one Default BWP and one Dedicated BWP. CQI measurements and PDSCH allocations are conducted only at PRBs within BWPs. The presence of DL traffic triggers the transition from Default BWP to Dedicated BWP, and a BWP Inactivity timer is used to trigger a switch from Dedicated BWP to Default BWP. In this study, UEs are scheduled only in Dedicated BWPs. This means UEs stay in Default BWP after inactivity timer expiry as long as DL traffic buffer remains empty. Upon DL traffic arrival at the gNB, a UE is switched to Dedicated BWP for DL reception via DCI command. BWP Switching Delay is assumed, with two values currently specified by 3GPP (1ms for Type 1 UE and 3ms for Type 2 UE for 15 kHz in FR1).

IV. Simulation Results

To evaluate the impact of delays from BWP adaptation in 5G NR performance, we considered scenarios with two BWP Inactivity Timers (8ms and 80ms) in combination with two BWP Switch Delays (1 ms and 3 ms), and compared them to a baseline scenario without BWP adaptation (i.e. always in Dedicated BWP). Both Default and Dedicated BWPs occupy the entire carrier BW, with no data transmission occurring while in Default BWP. This avoids the performance being negatively impacted by the unavailability of channel state information (CSI) for PRBs not in Default BWP. Thus, only the effect of BWP switch delays is reflected on the performance. We evaluated Downlink (DL) traffic only, and within each FTP3 call flow, there are in average 20 PDUs generated per second. To evaluate the impact of traffic burstiness into BWP adaptation, we evaluated scenarios with low (1.6 Mbps) and high (51.2 Mbps) offered loads, setting PDU sizes to 10 KB and 320 KB, respectively. Table IV summarizes the simulation scenarios. For each simulation point, we launched 40 runs of 20s of simulation time, later combined as to increase statistical confidence.

A. High Offered Load per UE [51.2 Mbps]

Figure 3 presents the CDF of the average application layer throughput per UE (in Mbps) for high load scenarios, i.e. 320 KB PDUs, 20 PDUs per second. All curves overlap, showing that system-level performance with all four BWP adaptation configurations is not impacted by BWP switching delays when compared to a baseline scenario without BWP. The cause is the high offered load (i.e. 51.2 Mbps per UE), which prevents the triggering of the BWP inactivity timer for most UEs, causing them to stay almost all the time in the Dedicated BWP waiting for queued data in L3 buffer to be transmitted.

![Fig. (3) Application Layer TPut for High Load [Mbps]](image)

The CDF of the ratio of total call time during which UEs stay in Default BWP is presented in Figure 4. Default BWP Ratio provides information about how “idle” a UE would be for a given BWP configuration under some particular offered traffic load; therefore, such metric is a valuable indication of the potential for power savings. 80% of the UEs do not stay in Default BWP, while the remaining 20% of UEs stay in Default BWP for less than 2% of the time. For this traffic pattern, these results indicate a low power saving potential due to BWP adaptation. A different presentation of the same load, with larger PDUs and longer inter-PDU interval, could provide different results.

![Fig. (4) Default BWP Ratio for High Load [% of call time]](image)

B. Low Load [1.6 Mbps]

For lower offered load, i.e.10 KB PDUs, 20 PDUs/sec, gNBs transmit PDUs more quickly than for the high load
As such, UEs become inactive more frequently and for more extended periods, which triggers the BWP Inactivity Timer more often, and increases the number of BWP switches between Dedicated BWP (i.e. when DL traffic is transmitted) and Default BWP (i.e. between PDU generation events).

Figure 5 presents the Default BWP Ratio CDF for the low load scenario. Results show that the occurrence of BWP switch events depends on: how well the BWP inactivity timer fits the traffic pattern, cell load and channel conditions. For instance, UEs in the bottom 15% (i.e. UEs in bad channel conditions due to, e.g. cell edge) rarely stay in Default BWP, which is expected since their DL buffer is seldom emptied promptly enough between PDU generation events as to trigger BWP inactivity timers. Figure 5 also shows that the shorter the values of BWP Inactivity Timer, the higher the percentage of time the UE stays in Default BWP. This is again expected, as a shorter BWP inactivity timer shall be more frequently triggered between PDU generation events, thus resulting in a higher rate of switches to Default BWP than with longer BWP inactivity timers. The large scale of the difference in duration at Default BWP (e.g. 4x more for 80th percentile) indicates that this effect is quite significant from the perspective of potential power savings due to BWP adaptation.

To estimate power savings from staying in Default BWP, we calculated what such saving would be relative to the baseline scenario for the case where power consumption, while in Default BWP, is 30% less than while in Dedicated BWP. Results in Figure 6 show that there are no power savings for UEs in adverse channel conditions. On the other hand, UEs with shorter values of BWP inactivity timer present up to 25% of power savings. Of course, power savings should scale down from a broader BWP to a narrower BWP, depending on how much baseband processing and RF power consumption also scale down with BW. The fitting between Inter-PDU traffic intervals and BWP inactivity timer value also affects the Default BWP ratio. Finally, we only simulated DL transmissions; since the potential for power saving is higher in UL transmissions, further investigations are necessary.

Figure 7 presents the CDF for the mean PDU scheduling delay in logarithmic scale, and refers to the delay from the moment the PDU arrives at the RLC layer (i.e. becomes available for scheduling) to the moment the gNB schedules the first transport block (TB) for that PDU. As expected, the constant BWP switching between Default BWP and Dedicated BWP from BWP adaptation introduces additional PDU latency. For a considerable amount of PDU scheduling events (lower 80% of PDUs), there is a difference in the PDU scheduling delay between scenarios with shorter BWP inactivity timer and the baseline scenario. Relevant is to say that, such increase of few milliseconds in PDU latency would be more relevant for delay-critical traffic, and is totally dependent on the BWP configuration in relation to the traffic pattern; this PDU latency is not significant at all when comparing scenarios with longer BWP inactivity timer duration and the baseline scenario without BWP adaptation. Also, both Default BWP and Dedicated BWP occupy 100% of the carrier BW, so CQI is constantly estimated for both BWPs and is readily available after BWP switches. With Default and Dedicated BWPs of different BW sizes, we expect additional latency due to HARQ re-transmissions caused by imprecise link adaptations. Finally, we do not dynamically reallocate BWPs via RRC Reconfiguration, which would, undoubtedly, impact PDU latency negatively.

The additional latency from the BWP switch delays in each BWP switch event shall increase the average PDU latency, which naturally impacts end-to-end throughput per-
formance. Figure 8, presenting the CDF of the application-layer throughput for the low load scenario, clearly shows this effect. For UEs with better throughput performance, we notice a significant throughput decrease when comparing different BWP adaptation scenarios to the baseline scenario without BWP adaptation. This might reach up to 50% lower throughput performance for (e.g. 90th percentile of BWP On, T=8ms, S=3ms). Also, we can notice that the higher the amount of BWP switch events (from, e.g. larger BWP inactivity timer default values and lower BWP switch delays), the higher the throughput degradation. The latter comes from the increased PDU latency associated, which was noticed in the PDU scheduling delay and significantly decreases throughput performance for shorter BWP inactivity timers. Some incremental difference is perceived when increasing BWP switch delays from 1ms to 3ms. Noteworthy is to say that, the scale of these differences at these particular values for BWP Inactivity Timer and BWP Switch Delay are completely dependent on the offered traffic load, i.e. with different traffic pattern, the results and the scale of the effect would be different. Finally, Default BWP ratio results demonstrated that, UEs in adverse channel conditions seldom switch to Default BWP, which is reflected on the lack of any perceived effect on throughput from BWP adaptation.

![Fig. (8) Application Layer TPut for Low Load [Mbps]](image)

**V. CONCLUSION AND FUTURE STEPS**

In this work, we presented the BWP feature as a 5G NR enabler for introducing flexibility in spectrum allocation. We also discussed the procedures introduced by Rel-15 for BWP adaptation, related to the dynamic switch between a Default BWP and a Dedicated BWP via inactivity timer. System-level simulation results indicate that for high load there is no effect on performance because of the lack of BWP switches, whereas for low load the higher number of BWP switches for low inactivity timers impacts negatively on the performance when compared to a baseline scenario without BWP adaptation. However, we also showed that BWP adaptation enables power savings if Default BWP is narrower than Dedicated BWP.

These findings provide a useful guideline for both the static planning of 5G NR cell deployments and the design of dynamic BWP adaptation algorithms. For the future, we intend to explore BWP power saving potential, by varying BWP sizes and using a 5G NR power consumption model that takes such BWP scaling under consideration. Also, we intend to explore dynamic BWP allocation and adaptation, by optimizing BWP inactivity timer default values and BWP carrier BW sizes according to different goals, e.g. maximize energy efficiency, enhance cell-edge performance, and others. Finally, another relevant aspect to be explored is the impact on performance from the unavailability of channel state information (CSI) during BWP adaptation in a scenario where the BWPs occupy different subsets of PRBs, specially when using different channel models (e.g. indoor).

**REFERENCES**


