3GPP TSG RAN WG1 Meeting #86bis **R1-1608584**
Lisbon, Portugal, 10th – 14th October 2016

Agenda item: 8.1.3.1
Source: AccelerComm
Title: Complementary turbo and LDPC codes for NR, motivated by a survey of over 100 ASICs
Document for: Discussion

I. Introduction

This paper outlines a hybrid turbo/LDPC solution for NR, in which channel coding is provided by a combination of a flexible turbo code and a 20 Gbps Low Density Parity Check (LDPC) code. We demonstrate that this approach can offer hardware- and energy-efficiencies that are significantly better than those of an LDPC-only solution, while meeting all of the NR requirements. In addition to a downlink information throughput of 20 Gbps, these requirements include flexible support for a wide variety of block lengths and coding rates. Support for medium and low coding rates is particularly important, since these are the ones used most frequently in typical mobile broadband deployments, as shown in Figure 1.

The remainder of this paper is organized as follows. Section II discusses the structure of turbo and LDPC codes. Section III provides a comprehensive comparison of state-of-the-art turbo and LDPC decoder Application Specific Integrated Circuits (ASICs), with a particular focus on those of [1]–[3]. Motivated by this comparison, an example hybrid turbo/LDPC solution for NR channel coding is described in Section IV. Finally, we offer our conclusions in Section V.

Observation 1: Typical mobile broadband deployments rely on medium and low coding rates most frequently.

II. Structure of turbo and LDPC codes

As shown in Figure 2, turbo codes employ a regular structure, which equally protects the K bits in each information block. The complexity of this structure scales with K, since the interconnections within the turbo code may be described by a $1 \times K$ vector, which is referred to as the interleaver pattern. Owing to the regularity of this structure, turbo codes can be readily designed to flexibly support a wide range of information block lengths K and the full range of coding rates R. More specifically, support for a number of different information block lengths K is typically achieved by defining a different interleaver pattern for each. In particular, the LTE turbo code employs 188 different interleaver patterns for information block lengths K in the range 40 to 6144 bits [4], as shown in Figure 3. Note however that shortening techniques can also be employed to provide further fine-grain information block length flexibility [5]. Meanwhile, turbo codes typically offer flexible support for the full range of coding rates R from 0 to 1 by using puncturing and repetition techniques, as shown for the LTE turbo code in Figure 3. Note that since turbo codes protect all information bits equally, the parametrization of the shortening, puncturing and repetition techniques does not necessarily require a high level of optimization.

By contrast, LDPC codes employ an irregular structure, which applies unequal protection to the K bits in each information block. The complexity of this structure scales with the encoded block length N, since the interconnections within the LDPC code may be described by a $(N - K) \times N$ matrix, which is referred to as the Parity Check Matrix (PCM). As shown in Figure 2, an $R = 1/2$-rate LDPC decoder has 7 times more random interconnections than an equivalent turbo decoder and even more for lower coding rates. Since different components of an LDPC decoder have different irregular numbers of connections,

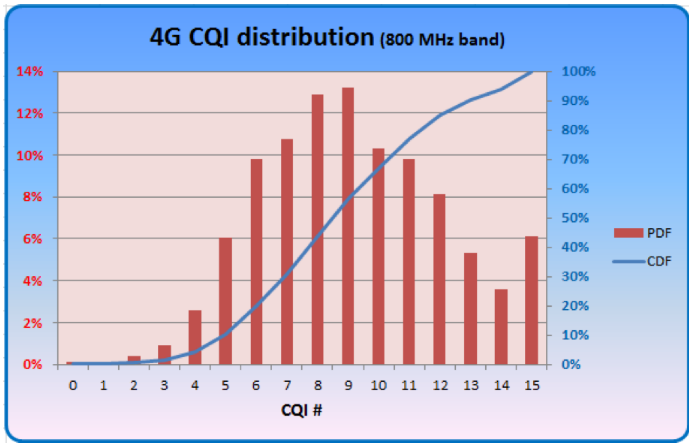


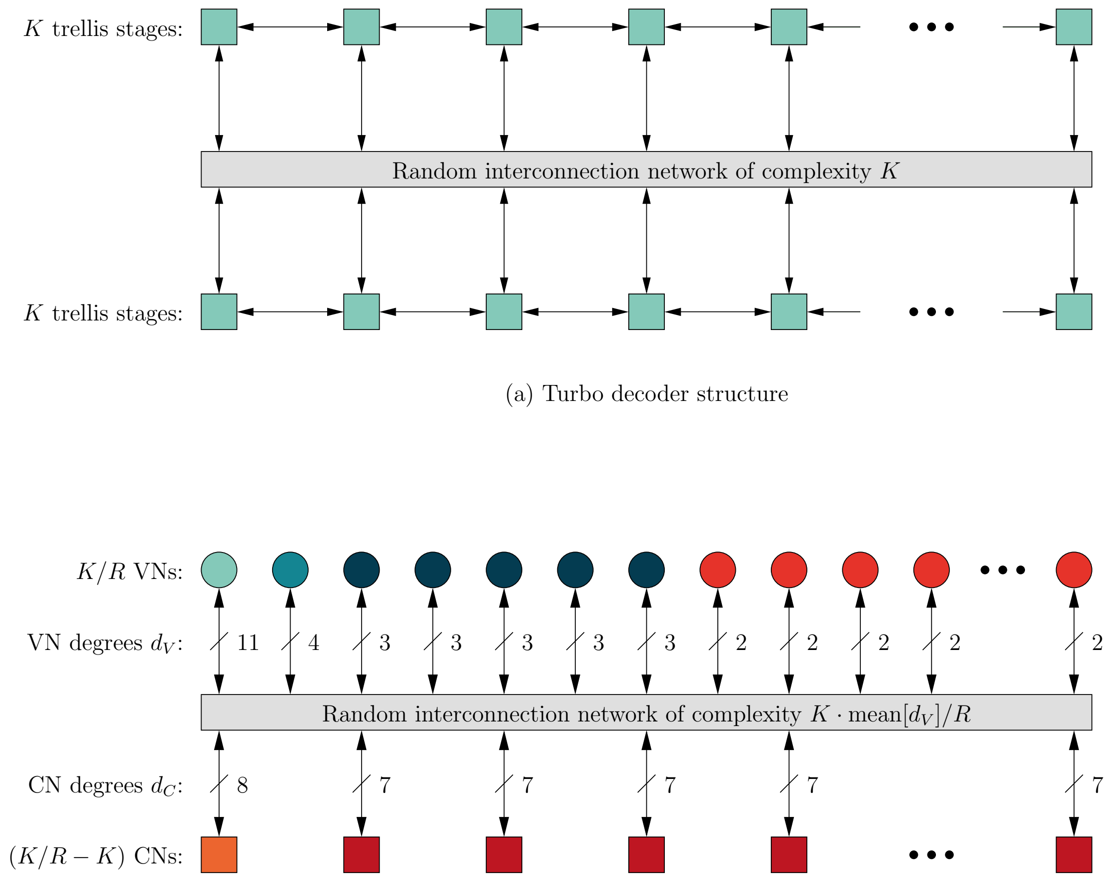
Fig. 1. The distribution of Channel Quality Indications (CQIs) measured throughout a typical day in a typical urban deployment of 1000 basestations. Here, 75% of the usage is at CQIs of 11 or below, which corresponds to coding rates R of 0.6 or below. Figure provided courtesy of Orange.

it is a significant challenge to design flexible LDPC codes that support various block lengths and various coding rates, as will be demonstrated in Section III. This flexibility is typically achieved by defining a different PCM for each supported combination of block length and coding rate.

However, there are no previously standardized LDPC (or polar) codes that flexibly support a wide range of block lengths and coding rates, as required for NR [6]. In particular, the only standards that have adopted LDPC codes with low coding rates are those of DVB, although even this code does not support coding rates R below $1/4$, as shown in Figure 3. More specifically, the DVB-S2 LDPC code [7] employs 21 PCMs to support the two encoded block lengths N of 16200 and 64800 bits, as well as 11 coding rates R in the range $1/4$ to $9/10$. Meanwhile, the WiMAX LDPC code [8] employs 114 PCMs to support 19 encoded block lengths N in the range 576 to 2304 bits, as well as four coding rates R in the range $1/2$ to $5/6$, as shown in Figure 3. Note that the WiMAX PCMs are grouped into families that share common structures, according to a quasi-cyclic technique [8]. A similar technique has been proposed by Qualcomm [9], in order to produce a set of 84 PCMs that support information block lengths K in the range 64 to 26880 bits and coding rates R in the range $10/122$ to $24/27$.

In addition to employing multiple PCMs, LDPC codes can achieve further fine-grain flexibility by using shortening, puncturing and repetition techniques. However, owing to the unequal protection that LDPC codes apply to the information bits, these techniques must be carefully parametrized in order to avoid degraded error correction capability, particularly if they are used to provide more than just fine-grain flexibility. Owing to this, the flexibility to support a wide range of block lengths and coding rates can only be achieved by employing a sufficient number of PCMs. In particular, in order meet the NR requirement for a very high degree of flexibility [6], it may be expected that an LDPC-only solution would need to support combinations of around 10 different coding rates and around 10 different block lengths, for a total of around 100 PCMs as in [8], [9]. This flexibility may be enhanced using quasi-cyclic, shortening, puncturing and/or repetition techniques.

When characterizing an ASIC implementation of a channel decoder, we are typically interested in its latency and *information* throughput, which is typically R times that of its *encoded* throughput. We are also interested in the ratio of this information throughput (measured in Mbps) to the chip area (measured in mm^2) and to the power consumption (measured in mW), which provide the hardware efficiency (measured in Mbps/mm^2) and the energy efficiency (measured in bit/nJ), respectively. However, these hardware characteristics of LDPC decoders are particularly poor at medium and low coding rates, owing to two fundamental reasons. Firstly, the number of rows in an LDPC PCM grows as the coding rate R is reduced, which implies a greater decoding complexity. Secondly, the number of columns in the PCM is given by the encoded block length N, which dictates the input and output interface to the LDPC decoder, as


K trellis stages:
Random interconnection network of complexity K
K trellis stages:
K/R VNs:
VN degrees d_V:
11
4
3
3
3
3
3
2
2
2
2
2
Random interconnection network of complexity $K \cdot \text{mean}[d_V]/R$
CN degrees d_C:
8
7
7
7
7
7
$(K/R - K)$ CNs:

(a) Turbo decoder structure

(b) LDPC decoder structure

Fig. 2. Structures of turbo and LDPC decoders. A turbo decoder has a regular structure, comprised of identical trellis stages, which each have only one random connection to another. By contrast, an LDPC decoder has an irregular structure, comprised of Variable Nodes (VNs) and Check Nodes (CNs) having different numbers of random interconnections to each other. Here, the average number of VN connections $\text{mean}[d_V]$ typically has a value of around 3.5 in standardized LDPC codes.

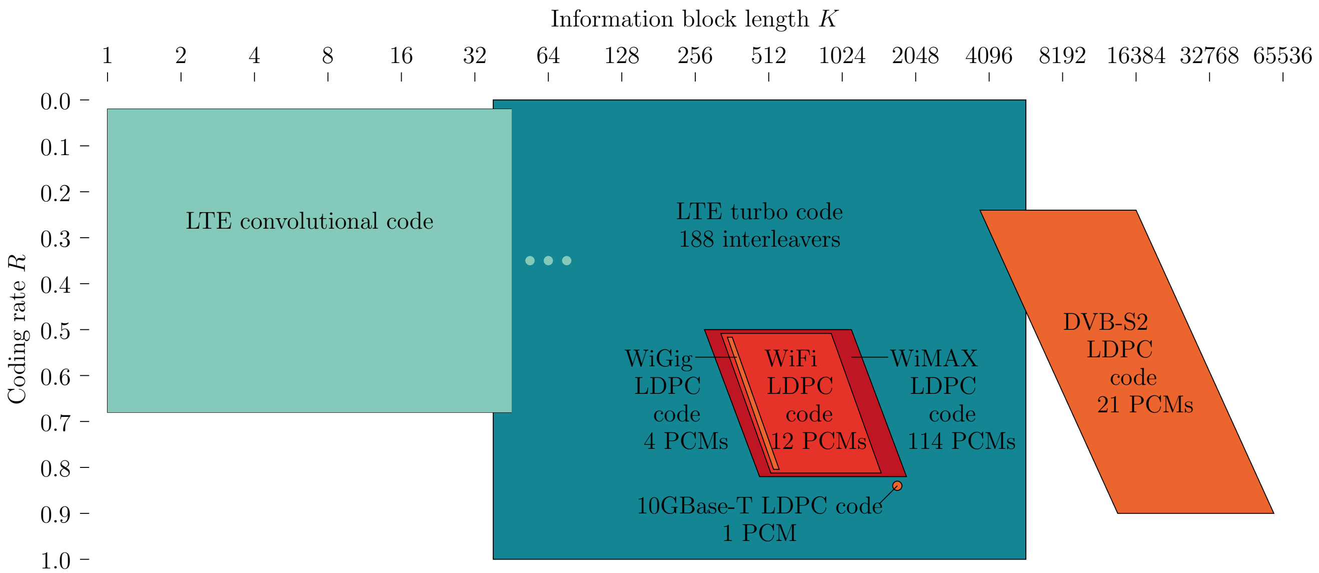


Fig. 3. The range of information block lengths K and coding rates R supported by a selection of standardized turbo and LDPC codes.

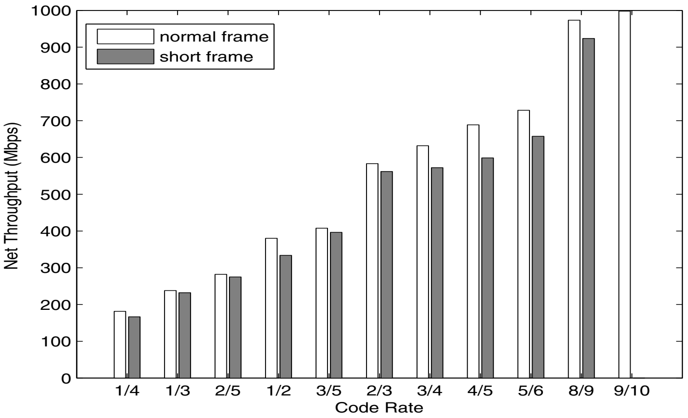


Fig. 4. Plot of information throughput versus coding rate R for the DVB-S2 LDPC decoder ASIC of [3], which supports encoded block lengths of $N = 16200$ (short frame) and $N = 64800$ (normal frame). Figure reproduced under fair use provisions ©IEEE.

shown in Figure 2. Therefore, LDPC decoders must operate by recovering the encoded bits and then extracting the information bits, as in polar decoders. Since LDPC (and polar) decoders must recover $1/R$ encoded bits in order to decode each information bit, their information throughputs scale down proportionately with the coding rate R, as exemplified in Figure 4 and illustrated in Figure 5. As a result, their hardware efficiencies and energy efficiencies also scale down proportionately with the coding rate R, while the latency associated with decoding a given information block length K scales up inversely proportionately with R. By contrast, turbo decoders can be said to decode the information bits directly, since their structure scales with the information block length K, as shown in Figure 2. Owing to this, the information throughputs, latencies, hardware efficiencies and energy efficiencies of turbo decoders do not typically vary significantly with coding rate R, as illustrated in Figure 5.

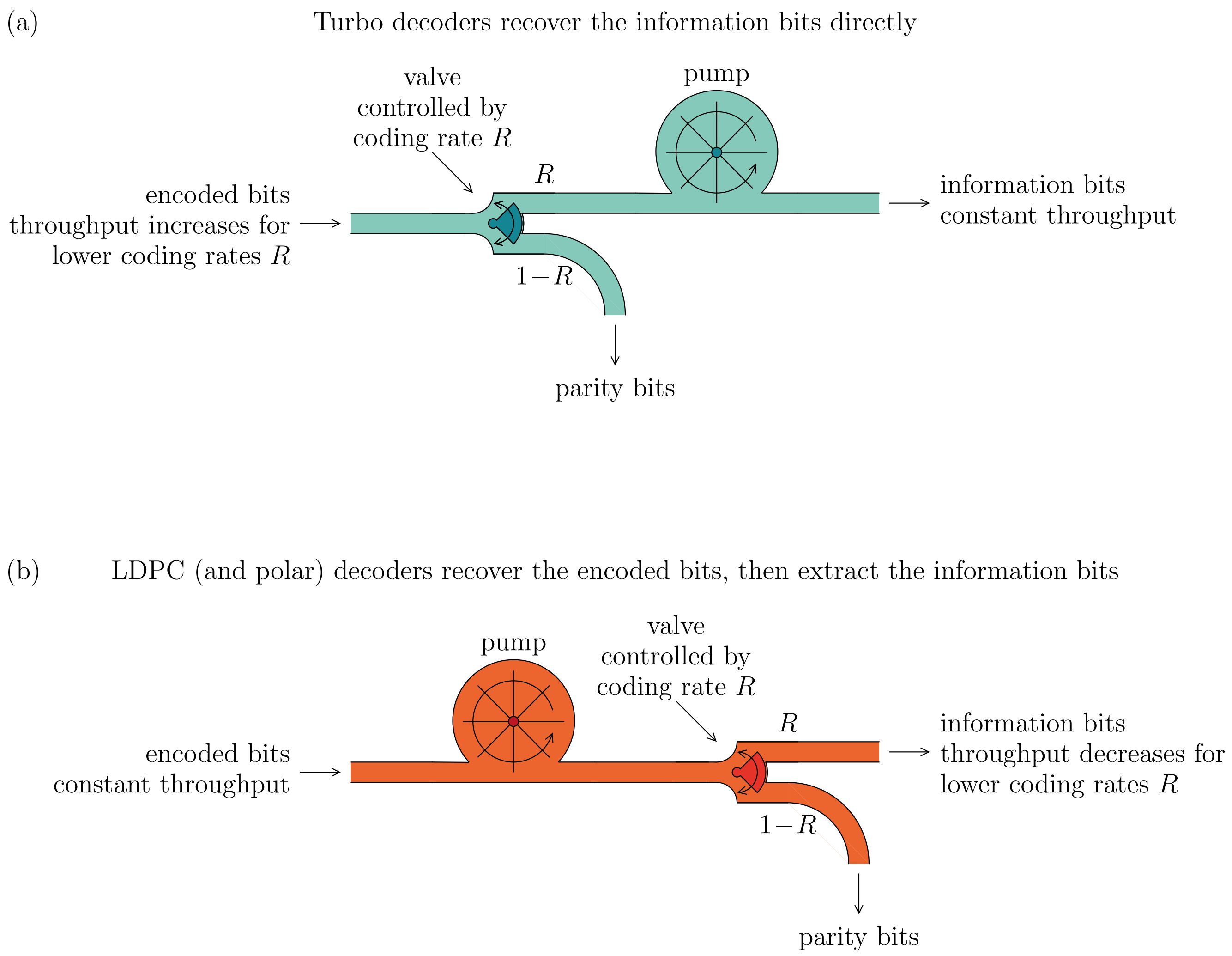
Observation 2: There are no previously standardized LDPC (or polar) codes - and hence no mature LDPC (or polar) decoder ASICs - that support a sufficiently wide range of block lengths and coding rates.

Observation 3: In order to offer sufficient flexibility, an LDPC-only approach to NR channel coding would need to support around 100 PCMs, together with quasi-cyclic, shortening, puncturing and repetition techniques.

Observation 4: In contrast to turbo decoders, the information throughput, hardware efficiency and energy efficiency of LDPC (and polar) decoder ASICs scale down proportionately with the coding rate, while the latency scales up inversely proportionately.

III. Survey of state-of-the-art turbo and LDPC decoder ASICs

A comprehensive survey of state-of-the-art ASIC implementations of turbo and LDPC decoders is provided in the dataset of [25], which tabulates and compares various hardware implementation characteristics. This comparison is summarized in Figure 6, which plots the flexibility, information throughput, hardware efficiency and energy efficiency of 22 recent turbo decoder ASICs and 89 recent LDPC decoder ASICs. Here, the flexibility is quantified by the number of interleaver designs that are supported by each turbo decoder ASIC and by the number of PCMs that are supported by each LDPC decoder ASIC. This is motivated since the range of information block lengths K and coding rates R that can be flexibly supported is primarily dictated by the number of interleavers or PCMs that are employed, as discussed in Section II. The information throughputs, hardware efficiencies and energy efficiencies shown in Figure 6 have all been scaled to 65 nm technology, since this minimizes the degree of scaling that is required for the majority of the considered ASICs. As described in Section II, these characteristics of a channel decoder ASIC may vary with information block length K and coding rate R, particularly for LDPC decoders. In all



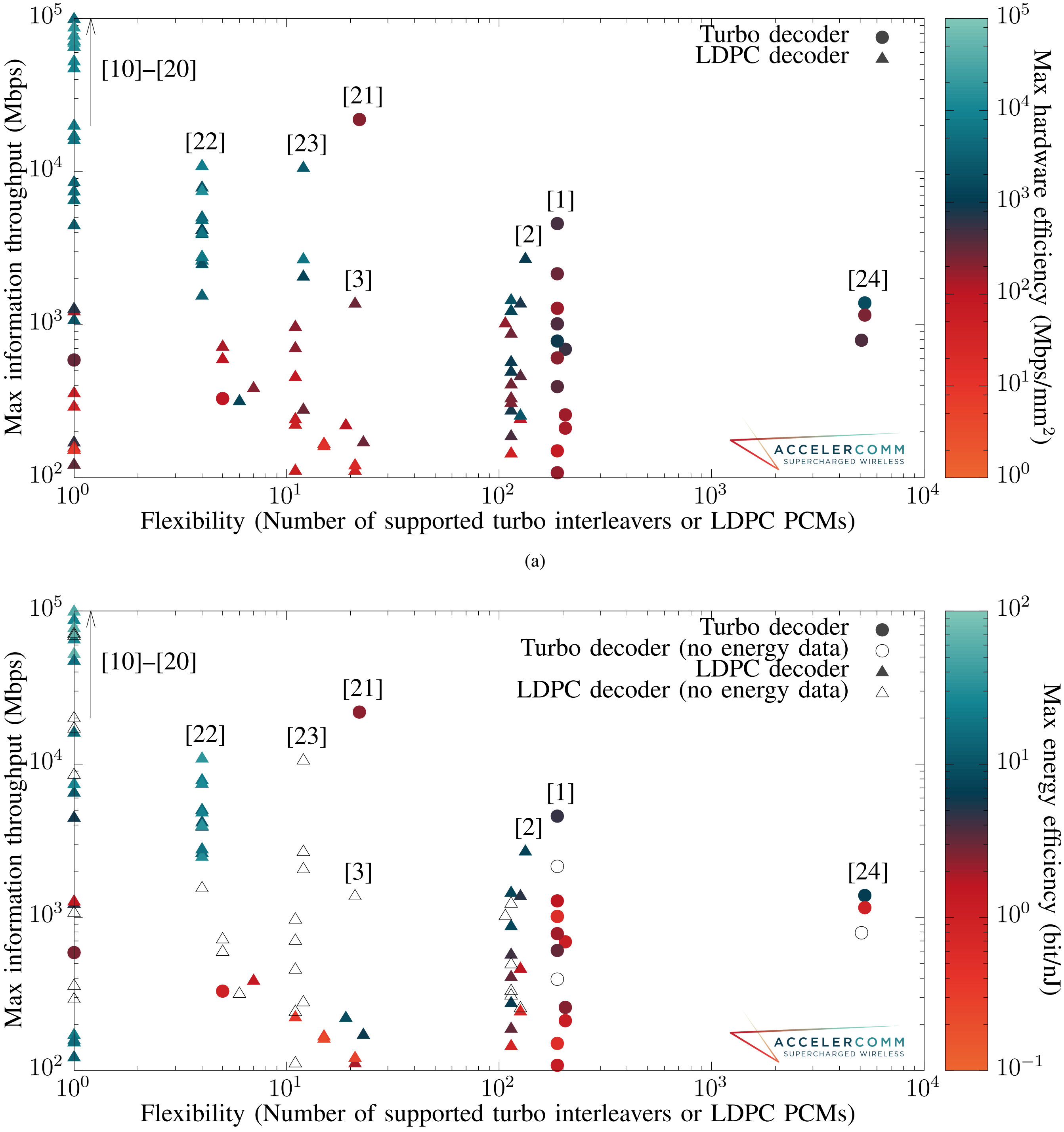
(c)

$$\left.\begin{array}{lcl} \text{latency} & \approx & \text{information block length } K \,\big/\, \text{information throughput} \\ \text{hardware efficiency} & = & \text{information throughput} \,\big/\, \text{chip area} \\ \text{energy efficiency} & = & \text{information throughput} \,\big/\, \text{power consumption} \end{array}\right\}$$

All depend on information throughput

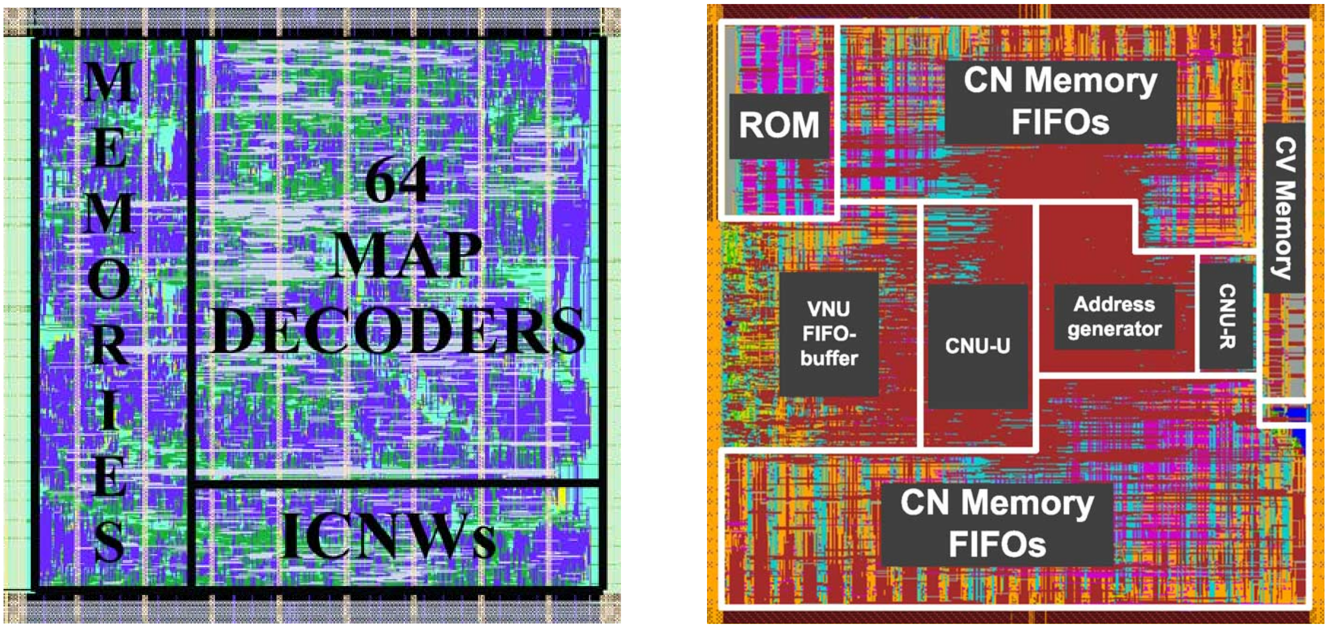
All degraded for lower coding rates R in the case of LDPC (and polar) decoders

Fig. 5. An analogy using pumps, valves and pipes, to illustrate how the coding rate R of (a) turbo and (b) LDPC decoders affects their encoded and information throughputs, as well as (c) their latency, hardware efficiency and energy efficiency.



(b)

Fig. 6. A comparison between state-of-the-art turbo and LDPC decoder ASICs when scaled to 65 nm, in terms of flexibility, maximum information throughput and (a) maximum hardware efficiency or (b) maximum energy efficiency.



(a)

(b)

Fig. 7. (a) ASIC layout of the partially-parallel turbo decoder of [1], in which the non-computational components (labeled 'MEMORIES' and 'ICNWs') occupy 30% of the area. (b) ASIC layout of the partially-parallel LDPC decoder of [2], in which the non-computational components (labeled 'ROM', 'CN Memory FIFOs', 'CV Memory', 'VNU FIFO-buffer', 'Address generator' and 'CN Memory FIFOs') occupy 75% of the area. Both figures reproduced under fair use provisions ©IEEE.

cases, Figure 6 plots the *maximum* information throughput, hardware efficiency and energy efficiency that is achieved by each ASIC, across all information block lengths K and coding rates R. In the case of the LDPC decoders, these maximum information throughputs, hardware efficiencies and energy efficiencies are typically achieved for high coding rates of around $R = 5/6$. However, lower coding rates typically result in significantly lower information throughputs and therefore also significantly lower hardware- and energy-efficiencies, as described in Section II.

As shown in Figure 6, the only LDPC decoder ASICs that achieve information throughputs in excess of 20 Gbps are those of [10]–[20]. All of these LDPC decoders adopt fully-parallel architectures that can only support a single PCM, having a high coding rate R. This is because these ASICs are laid out according to the factor graph [26] that is described by the PCM, using hard-wired connections between registers and dedicated computational hardware for each part of the factor graph. While this approach prevents the flexible support for more than one PCM, it allows the LDPC decoding process to be completed using a minimal number of clock cycles and without the use of additional memory, switchable interconnections or a complex controller. Owing to this, the LDPC decoders of [10]–[20] offer the best throughputs, latencies, hardware efficiencies and energy efficiencies among all of the channel decoder ASICs considered in Figure 6, although they offer the least flexibility.

By contrast, all flexible LDPC decoder ASICs employ partially-parallel architectures to support more than one PCM. More specifically, these ASICs employ a bank of computational hardware, which can be flexibly reused at different times to perform the processing associated with different parts of different PCMs. However, this approach requires the use of additional memory, switchable interconnections and a complex controller, which typically occupy around 75% of the chip area in the case of LDPC decoders that support around 100 PCMs, as exemplified in Figure 7. Owing to this, these additional hardware components dominate the throughput, latency, hardware efficiency and energy efficiency of the resultant ASICs, particularly as the number of supported PCMs is increased, as shown in Figure 6. This is because LDPC codes have irregular structures with a high interconnection complexity, as described in Section II.

By contrast, turbo decoders have regular structures with significantly lower interconnection complexities,

TABLE I
COMPARISON OF THE STATE-OF-THE-ART PARTIALLY-PARALLEL TURBO AND LDPC DECODER ASICS OF [1], [2] AND [3].

Paper	[1]			[2]			[3]		
Year	2014			2013			2009		
Published in	IEEE Trans. Circuits Syst. I			IEEE Trans. Circuits Syst. I			IEEE Veh. Technol. Conf.		
Technology (nm)	90			90			90		
Analysis	Post-layout			Post-layout			Post-synthesis		
Code	Turbo			LDPC			LDPC		
Supported standards	LTE			WiMAX, WiFi and G.hn			DVB-S2		
Number of supported interleavers/PCMs	188			133			21		
Coding rate R	High 0.95	Medium 0.50	Low 0.33	High 0.83	Medium 0.50	Low –	High 0.90	Medium 0.50	Low 0.25
Information throughput (Mbps)	2274	3028*	3307	857–1957	343–762	–	998	380	181
Latency** for $K = 1000$ (ns)	440	330*	302	511–1167	1312–2915	–	1002	2632	5525
Hardware efficiency (Mbps/mm^2)	115	153*	167	154–354	62–138	–	104	40	19
Energy efficiency (bit/nJ)	1.57	2.09*	2.28	2.30–5.25***	0.92–2.04***	–	?	?	?

* These characteristics for the medium coding rate have been obtained using linear interpolation between those achieved at the high and low coding rates.

** Latency is estimated by dividing the information block length $K = 1000$ by the information throughput, since latency is not quantified in [1], [2] or [3]. Note that while none of these decoders support information block lengths of exactly $K = 1000$, these estimates are provided for the sake of illustration.

*** The power consumption is stated as 228.36–517.70 mW in [2], but no discussion is provided about how this varies with coding rate. So, the average value of 373.03 mW has been used to calculate these energy efficiencies.

where the corresponding hardware typically occupies only 30% of the chip area, as exemplified in Figure 7. Owing to this, partially-parallel turbo decoders offer superior flexibilities, information throughputs, latencies, hardware efficiencies and energy efficiencies than partially-parallel LDPC decoders, as shown in Figure 6. This superiority is particularly apparent at medium and low coding rates R, as revealed by the comparison provided in Table I, which considers the partially-parallel turbo and LDPC decoder ASICs of [1] and [2], as illustrated in Figure 7. This is a particularly fair comparison for several reasons. Firstly, both of these channel decoder ASICs support around 100 turbo interleavers or LDPC PCMs, meeting the NR flexibility requirement discussed in Section II. Furthermore, both of these ASICs offer the best flexibilities, throughputs, latencies, hardware efficiencies and energy efficiencies among the sets of turbo and LDPC decoders having similar flexibility, as shown in Figure 6. Finally, both ASICs have been published within successive years, in the same prestigious journal and using post-layout analysis at the same technology scale, as shown in Table I.

As shown in Table I, the channel decoder ASICs of [1] and [2] both achieve different information throughputs and latencies at different coding rates R. In the case of the turbo decoder ASIC of [1], the information throughput and latency are degraded slightly when higher coding rates R are employed, since these require more decoding iterations in order to maintain near-capacity error correction. By contrast, the LDPC decoder ASIC of [2] requires more decoding iterations for lower coding rates R. Furthermore, the information throughput and latency of an LDPC decoder naturally degrade with reduced coding rate R, as described in Section II. Owing to this, Table I shows that at different coding rates R, the LDPC decoder ASIC of [2] has dramatically different latencies and information throughputs, leading to dramatically different hardware- and energy-efficiencies. While the LDPC decoder ASIC of [2] has higher hardware- and energy-efficiencies than the turbo decoder ASIC of [1] at high coding rates, this relationship is reversed at medium coding rates R. Although the LDPC decoder ASIC of [2] does not support coding rates below $R = 1/2$, it may be expected that a version supporting lower coding rates would achieve inversely proportionately higher latencies and proportionately lower information throughputs, as described in Section II. This would lead to much lower hardware- and energy-efficiencies than the turbo decoder ASIC of [1] at these lower coding rates. Furthermore, increasing the flexibility of the LDPC decoder ASIC of [2] in this way would increase its interconnection complexity, as described above. This may be expected to result in worse information throughput, latency, hardware efficiency and energy efficiency than those of the turbo decoder ASIC of [1] for all coding rates R, not just medium and low coding rates.

The only standards that have adopted LDPC codes with low coding rates R are those of DVB, as discussed in Section II. Of the 14 DVB-S2 LDPC decoder ASICs considered in Figure 6, the best throughputs, latencies and hardware-efficiencies are offered by the design of [3], as was characterized in Figure 4. However, this 90 nm ASIC achieves worse information throughputs, latencies and hardware efficiencies than the 90 nm turbo decoder ASIC of [1] at all coding rates R, as shown in Table I. In particular, the hardware efficiency of [3] is 8.8 times worse than that of [1] at low coding rates R. Furthermore, this advantage of the turbo decoder ASIC could be expected to be even greater if post-layout analysis was applied to the LDPC decoder ASIC of [3], rather than only post-synthesis analysis. The advantage could be expected to become even greater still, if the flexibility of this LDPC decoder ASIC was increased to include more information block lengths K and lower coding rates R, in order to match the flexibility of the turbo decoder. Note that while the energy efficiency of this DVB-S2 LDPC decoder ASIC was not quantified in [3], we may expect this to scale linearly with the hardware efficiency, as is the case for the ASICs of [1] and [2]. Hence, we may conclude that the information throughput, latency, hardware efficiency and energy efficiency of flexible turbo decoder ASICs can be around an order-of-magnitude better than those of flexible LDPC decoder ASICs at low coding rates.

Observation 5: All LDPC decoders that have demonstrated information throughputs of 20 Gbps use fully-parallel architectures that support only a single PCM.

Observation 6: The information throughput, latency, hardware efficiency and energy efficiency of partially-parallel LDPC decoder ASICs are dominated by their flexibility, where support for more PCMs results in significantly reduced performance.

Observation 7: Partially-parallel turbo decoders offer significantly better flexibilities, information throughputs, latencies, hardware efficiencies and energy efficiencies than partially-parallel LDPC decoders, particularly at medium and low coding rates.

IV. Example of complementary turbo and LDPC codes for NR

As described in Section III, partially-parallel LDPC decoder ASICs can achieve flexibility, while fully-parallel LDPC decoder ASICs can achieve high information throughputs. However, no LDPC decoder architectures have been demonstrated that can achieve both flexibility and throughputs of 20 Gbps. In particular, the throughput of the flexible partially-parallel LDPC decoder ASIC of [2] would need to be increased by around 40 times in order to reach the 5G downlink target of 20 Gbps across its range of supported coding rates R, which is not feasible using either state-of-the-art technology scales or by operating many replicas of the decoder in parallel. Therefore, in order to meet the NR flexibility and downlink information throughput requirements, an LDPC-only approach would require the use of two independent LDPC decoders in the handsets. More specifically, a fully-parallel LDPC decoder would be used in applications where the channel is pushed to its capacity, allowing multi-Gbps information throughputs to be achieved by using medium coding rates R to maintain reliability. By contrast, a partially-parallel LDPC decoder would be used to flexibly support a wide variety of information block lengths K, as well as the full range of coding rates R.

However, Section III showed that partially-parallel turbo decoders offer superior flexibilities, information throughputs, latencies, hardware efficiencies and energy efficiencies than partially-parallel LDPC decoders, particularly at medium and low coding rates R. This suggests a hybrid turbo/LDPC approach to channel coding in the NR downlink, which complements a fully-parallel LDPC decoder with a partially-parallel turbo decoder, rather than the partially-parallel LDPC decoder of the LDPC-only approach. Furthermore,

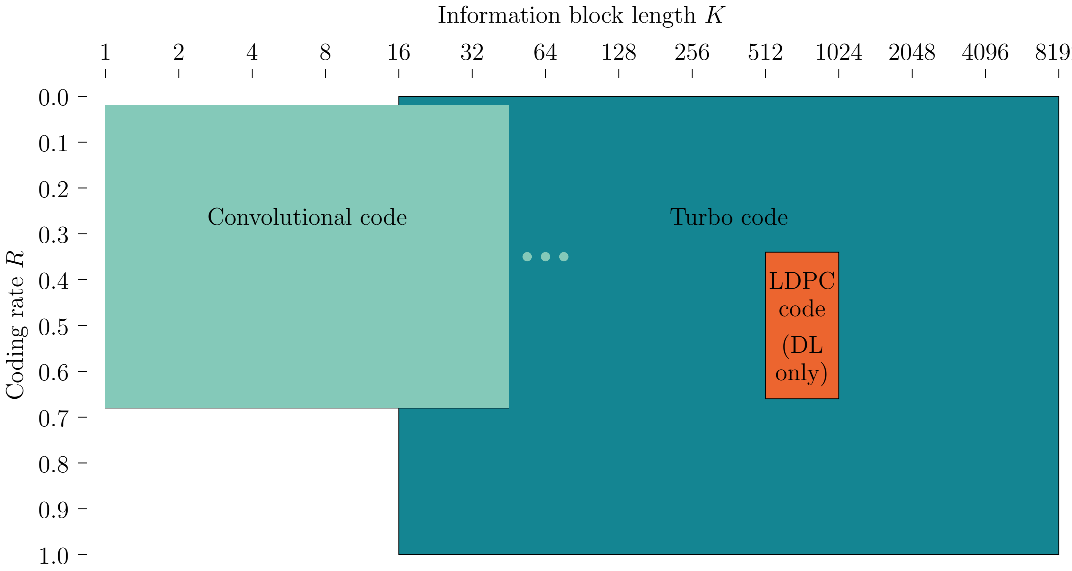


Fig. 8. A turbo code may be employed to flexibly support a wide range of information block lengths K and the full range of coding rates R. This may be supported by an LDPC code for medium information block lengths K and medium coding rates R. A convolutional code may also be used for short information block lengths K and low to medium coding rates R, as in UMTS and LTE.

this motivates a turbo-only approach to channel coding in the NR uplink to the basestation, where 20 Gbps information throughputs are not required. In addition to improved flexibility, information throughput, latency, hardware efficiency and energy efficiency, this approach would also offer the significant advantage of having synergy with the UMTS and LTE turbo codes. More specifically, a hardware implementation of a NR turbo code could be readily reused to also perform UMTS and LTE turbo coding, reducing the chip area and power consumption of a handset or basestation that supports multiple generations.

Figure 8 illustrates a particular example of the hybrid turbo/LDPC approach described above. Here, an enhanced LTE turbo code may be employed, which includes additional interleavers for shorter and longer information block lengths K, native support for lower coding rates, improved error correction using tailbiting, as well as traditional puncturing and repetition techniques [27]. At state-of-the-art technology scales, it may be expected that this turbo decoder could achieve information throughputs of several Gbps. For higher information throughputs in the downlink, this turbo decoder may be complemented by an LDPC code having a single PCM, which is suitable for fully-parallel decoding. This PCM may be carefully designed to allow some fine-grain flexibility to be achieved using puncturing, repetition and shortening techniques. Furthermore, the native information block length K of this PCM must be carefully selected. If K is too large, then the fully-parallel decoder will have an excessive chip area. But if K is too small, then it will not be possible to achieve information throughputs of 20 Gbps, particularly when shortening is employed. As an example, the PCM employed in 10GBase-T Ethernet [28] could be adapted to have a coding rate of $R = 1/2$, which is motivated since medium coding rates are used most frequently in typical mobile broadband deployments, as shown in Figure 1. Besides the proprietary PCMs of [19], this 10GBase-T Ethernet design is the only LDPC PCM for which information throughputs exceeding 20 Gbps have been demonstrated, as detailed in [10]–[18], [20]. This PCM would have an information block length of $K = 1024$ bits and it may be expected that shortening could be used to support information block lengths K of several hundreds of bits, without excessively degrading the performance of the decoder. Likewise, it may be expected that this PCM's coding rate of $R = 1/2$ could be extended across the range of medium coding rates, using puncturing and repetition techniques. As shown in Figure 8, this hybrid turbo/LDPC approach could also be complemented by a convolutional code, for use at short information block lengths K and low to medium coding rates R, as in UMTS and LTE.

The advantages of the hybrid turbo/LDPC approach described above can be appreciated by considering a hypothetical implementation example and comparing with LDPC-only benchmarkers, as shown in Table II. Here, it is assumed that information throughputs of at least 20 Gbps are required for high throughput applications at medium coding rates R, while 5 Gbps is required for all other applications at all other coding rates R. In each case considered, these targets are achieved using either a combination of inflexible and flexible decoders, or using a flexible decoder alone, where technology scaling to 40 nm is assumed. The inflexible fully-parallel LDPC decoder of [18] is selected, since this offers the best hardware- and energy-efficiencies among all decoders considered in [25].

As shown in Table II, the combination of the inflexible LDPC decoder of [18] with the flexible partially-parallel turbo decoder of [1] satisfies the throughput requirements described above, while imposing only modest latencies, chip area and power consumption. Furthermore, the turbo decoder of this combination fully satisfies the NR flexibility requirement. By contrast, the NR flexibility requirement is not met by the partially-parallel LDPC decoder of [2], since it only supports a narrow range of block lengths and only supports coding rates above $R = 0.50$, as described above. Increasing the flexibility of this design to meet the requirement could be expected to significantly degrade its hardware- and energy-efficiency. However, despite this, Table II optimistically assumes that the flexibility of this design could be enhanced to support a low coding rate of $R = 0.33$ at a scaled information throughput, without degrading the hardware- or energy-efficiency otherwise. As shown in Table II, the above-mentioned information throughput requirements are met in the case where a flexible LDPC decoder comprising 5 parallel replicas of the design of [2] is combined with the inflexible LDPC decoder of [18]. However, this approach can be seen to impose significantly greater latencies, chip area and power consumption than the hybrid turbo/LDPC approach described above. Much worse hardware characteristics result when attempting to meet the throughput requirements using only a flexible LDPC decoder comprising 12 parallel replicas of the design of [2], without the aid of the inflexible LDPC decoder of [18]. This highlights that an LDPC-only approach to NR channel coding would require two independent LDPC decoders, namely one fully-parallel decoder to achieve information throughputs of 20 Gbps and one partially-parallel decoder to achieve flexibility. However, since flexible turbo decoders offer superior information throughput, latency, hardware efficiency and energy efficiency than flexible LDPC decoders, it is better to adopt the hybrid turbo/LDPC approach instead.

Table III considers a turbo-only approach and an optimistic LDPC-only approach to the NR downlink, for the case where 20 Gbps is required at all coding rates R. Note however that the hardware- and energy-efficiency of a sufficiently-flexible LDPC decoder may be expected to be significantly worse than those optimistically assumed in Table III. In spite of this, it can be seen that the turbo-only solution offers significantly smaller chip area and power consumption than the LDPC-only solution. However, its chip area and power consumption are still very large, at around four times greater than those of the hybrid turbo/LDPC approach considered in Table II. Based on this, it may be considered that better user experience would result from trading flexibility at 20 Gbps for a significant reduction in chip area and power consumption, as offered by the hybrid turbo/LDPC approach described above.

Observation 8: An LDPC-only approach to NR channel coding would require two independent LDPC decoders, namely one fully-parallel decoder to achieve information throughputs of 20 Gbps and one partially-parallel decoder to achieve flexibility.

Proposal 1: A hybrid turbo/LDPC approach to NR channel coding should be adopted in the downlink, which employs a turbo code to flexibly support a wide range of information block lengths and the full range of coding rates, supported by an LDPC code based on a single PCM having a medium information block length and a medium coding rate.

Proposal 2: A turbo-only approach to NR channel coding should be adopted in the uplink, which employs a turbo code to flexibly support a wide range of information block lengths and the full range of coding rates.

TABLE II

A hybrid turbo/LDPC approach to channel coding in the NR downlink and its comparison with two LDPC-only approaches, for the case where an information throughput of 20 Gbps is required at medium coding rates and 5 Gbps is required across all other coding rates.

Papers	[1] and [18]			[2] and [18]			[2]		
Technology (nm)	Scaled to 40			Scaled to 40			Scaled to 40		
Codes	1 inflexible LDPC decoder and 1 flexible turbo decoder			1 inflexible LDPC decoder and 5 parallel flex. LDPC decoders			12 parallel flexible LDPC decoders only		
Coding rate R	High 0.95	Medium 0.50	Low 0.33	High 0.83	Medium 0.50	Low 0.33	High 0.83	Medium 0.50	Low 0.33
Information throughput (Gbps)	6.24	102.62**	7.44	22.02*	102.62**	5.71*	52.8*	20.6*	13.7*
Latency*** for $K = 1000$ (ns)	160	10	134	227	10	876	227	583	876
Total area (mm^2)	$3.90 + 0.55 = 4.45$			$5.46 + 0.55 = 6.01$			13.11		
Total power (mW)	$645 + 284 = 929$			829**** $+ 284 = 1113$			1989****		

TABLE III

A turbo-only approach to channel coding in the NR downlink and its comparison with an LDPC-only approach, for the case where an information throughput of 20 Gbps is required across all coding rates.

Papers	[1]			[2]		
Technology (nm)	Scaled to 40			Scaled to 40		
Codes	4 parallel flexible turbo decoders only			18 parallel flexible LDPC decoders only		
Coding rate R	High 0.95	Medium 0.50	Low 0.33	High 0.83	Medium 0.50	Low 0.33
Information throughput (Gbps)	24.97	27.25	29.76	79.26*	30.86*	20.55*
Latency*** for $K = 1000$ (ns)	160	147	134	227	583	876
Total area (mm^2)	15.60			19.66		
Total power (mW)	2579			2984****		

* These information throughputs for the flexible LDPC decoder of [2] at high and medium coding rates are optimistically based on the highest values in the corresponding ranges of Table I. The results for the low coding rate of $R = 0.33$ are obtained by scaling those of the medium coding rate $R = 0.50$. It is optimistically assumed that the flexibility can be increased to include this low coding rate without degrading the hardware- or energy-efficiency.

** These information throughputs are achieved using the inflexible LDPC decoder of [18], when it is scaled to a coding rate of $R = 0.50$. It is assumed that this can be achieved without degrading the hardware- or energy-efficiency.

*** Latency is estimated by multiplying the number of parallel decoders by the information block length $K = 1000$ and dividing by the information throughput. Note that while none of these decoders support information block lengths of exactly $K = 1000$, these estimates are provided for the sake of illustration.

**** The power consumption is stated as 228.36–517.70 mW in [2], but no discussion is provided about how this varies with coding rate. So, the average value of 373.03 mW has been scaled and used as the basis of these power consumptions.

V. Conclusions

In this paper, we have discussed the degrees to which turbo and LDPC codes can meet the NR requirements, which include much higher information throughputs and greater flexibility than LTE. We have shown that there are no previously standardized LDPC codes and hence no mature LDPC decoder ASICs that support a sufficiently wide range of information block lengths K and coding rates R. So in order to offer sufficient flexibility, an LDPC-only approach to NR channel coding would need to adopt a new approach comprising around 100 PCMs, supported by quasi-cyclic, shortening, puncturing and repetition techniques. However, the 20 Gbps information throughput downlink requirement has only been met by fully-parallel LDPC decoders, which support only a single PCM. This highlights that an LDPC-only approach would require the use of two independent LDPC decoders in the handset, namely one fully-parallel decoder to achieve information throughputs of 20 Gbps and one partially-parallel decoder to achieve flexibility. However, we have shown that the hardware performance of partially-parallel LDPC decoder ASICs is degraded significantly when more flexibility is required. Owing to this, partially-parallel turbo decoders offer superior flexibilities, information throughputs, latencies, hardware efficiencies and energy efficiencies than partially-parallel LDPC decoders, particularly at medium and low coding rates R, which are the ones used most frequently in typical mobile broadband deployments. This is because the information throughputs, hardware efficiencies and energy efficiencies of LDPC (and polar) decoders scale down proportionately with coding rate R, while the latencies scale up inversely proportional with R, unlike in turbo decoders. This leads us to recommend a hybrid turbo/LDPC approach to NR channel coding, which employs a turbo code to flexibly support a wide range of information block lengths K and the full range of coding rates R, but supported in the downlink by an LDPC code based on a single PCM having a medium information block length K and a medium coding rate R. This may also be supported by a convolutional code for short information block lengths K and low to medium coding rates R, as in UMTS and LTE. We have demonstrated that the proposed hybrid turbo/LDPC approach can meet all of the NR requirements, while offering hardware- and energy-efficiencies that are significantly better than those of LDPC-only solutions.

Observation 1: Typical mobile broadband deployments rely on medium and low coding rates most frequently.

Observation 2: There are no previously standardized LDPC (or polar) codes - and hence no mature LDPC (or polar) decoder ASICs - that support a sufficiently wide range of block lengths and coding rates.

Observation 3: In order to offer sufficient flexibility, an LDPC-only approach to NR channel coding would need to support around 100 PCMs, together with quasi-cyclic, shortening, puncturing and repetition techniques.

Observation 4: In contrast to turbo decoders, the information throughput, hardware efficiency and energy efficiency of LDPC (and polar) decoder ASICs scale down proportionately with the coding rate, while the latency scales up inversely proportionately.

Observation 5: All LDPC decoders that have demonstrated information throughputs of 20 Gbps use fully-parallel architectures that support only a single PCM.

Observation 6: The information throughput, latency, hardware efficiency and energy efficiency of partially-parallel LDPC decoder ASICs are dominated by their flexibility, where support for more PCMs results in significantly reduced performance.

Observation 7: Partially-parallel turbo decoders offer significantly better flexibilities, information throughputs, latencies, hardware efficiencies and energy efficiencies than partially-parallel LDPC decoders, particularly at medium and low coding rates.

Observation 8: An LDPC-only approach to NR channel coding would require two independent LDPC decoders, namely one fully-parallel decoder to achieve information throughputs of 20 Gbps and one partially-parallel decoder to achieve flexibility.

Proposal 1: A hybrid turbo/LDPC approach to NR channel coding should be adopted in the downlink, which employs a turbo code to flexibly support a wide range of information block lengths and the full range of coding rates, supported by an LDPC code based on a single PCM having a medium information block length and a medium coding rate.

Proposal 2: A turbo-only approach to NR channel coding should be adopted in the uplink, which employs a turbo code to flexibly support a wide range of information block lengths and the full range of coding rates.

References

[1] R. Shrestha and R. P. Paily, "High-throughput turbo decoder with parallel architecture for LTE wireless communication standards," *IEEE Trans. Circuits Syst. I*, vol. 61, no. 9, pp. 2699–2710, Sept 2014.

[2] Y. L. Ueng, B. J. Yang, C. J. Yang, H. C. Lee, and J. D. Yang, "An efficient multi-standard LDPC decoder design using hardware-friendly shuffled decoding," *IEEE Trans. Circuits Syst. I*, vol. 60, no. 3, pp. 743–756, March 2013.

[3] B. Zhang, H. Liu, X. Chen, D. Liu, and X. Yi, "Low complexity DVB-S2 LDPC decoder," in *Proc. IEEE Veh. Technol. Conf.*, Barcelona, Spain, April 2009.

[4] *ETSI TS 136 212 LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Multiplexing and Channel Coding*, V13.1.0 ed., 2016.

[5] A. Nimbalker, Y. Blankenship, B. Classon, and T. K. Blankenship, "ARP and QPP interleavers for LTE turbo coding," in *Proc. IEEE Wireless Commun. Networking Conf.*, Las Vegas, NV, USA, mar 2008, pp. 1032–1037.

[6] 3GPP, "Draft minutes report," in *3GPP TSG RAN WG1 #86*, Aug. 2016.

[7] *ETSI EN 302 307-1 Digital Video Broadcasting (DVB); Second generation framing structure, channel coding and modulation systems for Broadcasting, Interactive Services, News Gathering and other broadband satellite applications; Part 1: DVB-S2*, V1.4.1 ed., 2014.

[8] *IEEE 802.16-2012 Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Broadband Wireless Access Systems*, 2012.

[9] Qualcomm, "R1-166388 LDPC rate compatible design," in *3GPP TSG RAN WG1 #86*, Aug. 2016.

[10] D. Wu, Y. Chen, Q. Zhang, Y. l. Ueng, and X. Zeng, "Strategies for reducing decoding cycles in stochastic LDPC decoders," *IEEE Trans. Circuits Syst. II*, vol. 63, no. 9, pp. 873–877, Sept 2016.

[11] Z. Zhang, V. Anantharam, M. J. Wainwright, and B. Nikolic, "An efficient 10GBASE-T Ethernet LDPC decoder design with low error floors," *IEEE J. Solid-State Circuits*, vol. 45, no. 4, pp. 843–855, April 2010.

[12] K. Cushon, S. Hemati, S. Mannor, and W. J. Gross, "Energy-efficient gear-shift LDPC decoders," in *Proc. IEEE Int. Conf. Application-Specific Systems, Architectures Processors*, Zurich, Switzerland, June 2014, pp. 219–223.

[13] C. C. Cheng, J. D. Yang, H. C. Lee, C. H. Yang, and Y. L. Ueng, "A fully parallel LDPC decoder architecture using probabilistic min-sum algorithm for high-throughput applications," *IEEE Trans. Circuits Syst. I*, vol. 61, no. 9, pp. 2738–2746, Sept 2014.

[14] D. Wu, Y. Chen, Q. Zhang, L. Zheng, X. Zeng, and Y. l. Ueng, "Latency-optimized stochastic LDPC decoder for high-throughput applications," in *Proc. IEEE Int. Symp. Circuits Systems*, Lisbon, Portugal, May 2015, pp. 3044–3047.

[15] S. S. Tehrani, A. Naderi, G. A. Kamendje, S. Hemati, S. Mannor, and W. J. Gross, "Majority-based tracking forecast memories for stochastic LDPC decoding," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4883–4896, Sept 2010.

[16] T. Mohsenin, H. Shirani-mehr, and B. M. Baas, "LDPC decoder with an adaptive wordwidth datapath for energy and BER co-optimization," *VLSI Design*, 2013.

[17] T. Mohsenin, D. N. Truong, and B. M. Baas, "A low-complexity message-passing algorithm for reduced routing congestion in LDPC decoders," *IEEE Trans. Circuits Syst. I*, vol. 57, no. 5, pp. 1048–1061, May 2010.

[18] K. Cushon, S. Hemati, C. Leroux, S. Mannor, and W. J. Gross, "High-throughput energy-efficient LDPC decoders using differential binary message passing," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 619–631, Feb 2014.

[19] K. Cushon, P. Larsson-Edefors, and P. Andrekson, "Energy-efficient soft-decision LDPC FEC for long-haul optical communication," in *Proc. Euro. Conf. Optical Commun.*, Valencia, Spain, Sept 2015, pp. 1–3.

[20] A. Naderi, S. Mannor, M. Sawan, and W. J. Gross, "Delayed stochastic decoding of LDPC codes," *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5617–5626, Nov 2011.

[21] A. Li, L. Xiang, T. Chen, R. G. Maunder, B. M. Al-Hashimi, and L. Hanzo, "VLSI implementation of fully parallel LTE turbo decoders," *IEEE Access*, vol. 4, pp. 323–346, 2016.

[22] X. R. Lee, C. L. Chen, H. C. Chang, and C. Y. Lee, "A 7.92 Gb/s 437.2 mW stochastic LDPC decoder chip for IEEE 802.15.3c applications," *IEEE Trans. Circuits Syst. I*, vol. 62, no. 2, pp. 507–516, Feb. 2015.

[23] T. H. Tran, Y. Nagao, H. Ochi, and M. Kurosaki, "ASIC design of 7.7 Gbps multi-mode LDPC decoder for IEEE 802.11ac," in *Proc. IEEE Int. Symp. Commun. Information Tech.*, Incheon, Republic of Korea, Sept 2014, pp. 259–263.

[24] Q. Yang, X. Zhou, G. E. Sobelman, and X. Li, "Network-on-chip for turbo decoders," *IEEE Trans. VLSI Syst.*, vol. 24, no. 1, pp. 338–342, Jan 2016.

[25] R. G. Maunder, "Survey of ASIC implementations of turbo and LDPC decoders," *University of Southampton Dataset*, Aug. 2016. [Online]. Available: http://eprints.soton.ac.uk/399846/

[26] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Sig. Process. Mag.*, vol. 21, no. 1, pp. 28–41, jan 2004.

[27] Orange, "R1-167413 Enhanced turbo codes for NR: Implementation details," in *3GPP TSG RAN WG1 #86*, Aug. 2016.

[28] *IEEE 802.3an-2006 Specific requirements Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications Amendment 1: Physical Layer and Management Parameters for 10 Gb/s Operation*, 2006.