High-Speed Channel Coding Architectures for Next-Generation Terabit Communications

Elena Vasquez, Julian Sanchez

Elena Vasquez, School of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407, USA; Julian Sanchez, Department of Computer Science and Engineering, University of Michigan, Ann Arbor, MI 48109-2122, USA.

Abstract—The continuous demands for higher throughput, higher spectral efficiency, lower latencies, lower power and large scalability in communication systems impose large challenges on the baseband signal processing. In the future, throughput requirements far beyond 100 Gbit/s are expected, which is much higher than the tens of Gbit/s targeted in the 5G standardization. At the same time, advances in silicon technology due to shrinking feature sizes and increased performance parameters alone will not provide the necessary gain, especially in energy efficiency for wireless transceivers, which have tightly constrained power and energy budgets. In this talk we will focus on channel coding, which is a major source of complexity in digital baseband processing. We will give an overview and first results of the EPIC project, funded by European Union's Horizon 2020 research and innovation program, that aims to develop a new generation of Forward-Error-Correction codes in a manner that will serve as a fundamental enabler of practicable beyond 5G wireless Tbit/s solutions. We will highlight implementation challenges for the most advanced channel coding techniques, i.e. Turbo codes, LowDensity Parity-Check (LDPC) codes and Polar codes and present decoder architectures for all three code classes that are designed for highest throughput.

Mobile communication plays a central role in our information society and is a key enabler for the connected world. In the last decades we have seen a tremendous increase in data rates over the different generations, e.g., GSM featured about 10kbit/s, UMTS about 2Mbit/s, and LTE-A about 1Gbit/s.

The newest standard, 5G, enables data rates $>10\text{Gbit/s}$. Beyond 5G, data rates towards 1Tbit/s are expected. The tremendous improvement in mobile communication has to be considered in the context of the progress in microelectronic industry, driven by Moore's law [6] that states an exponential increase in transistors per chip. In today's 14nm technology, 38 million transistors can be integrated on 1mm^2 of silicon. For many decades, improvement in silicon process technology provided better performance, lower cost per gate, higher integration density and lower power consumption. However, we have reached a point where Moore's law is slowing down for the following reasons [2], [3], [4]:

- Cost of technology: The cost per wafer from 28nm technology to 7nm has more than doubled. At the same time, the area density (yield not considered) increased by 6×. Thus, the cost per mm^2 is still decreasing but not with the same pace as in the past.
- Design cost: The average IC design cost in 14nm is about $ 80 million, compared to $ 30 million for a 28nm planar device [2]. It costs more than $ 200 million to design a chip in 7nm. This creates a situation where less and less designs each year have enough volume to amortize the cost of the design.
- Decreasing performance gain: Over the last 10 years, the semiconductor industry has succeeded in doubling the transistor density every 2 to 2.4 years. However, the performance gains have been much smaller, such that less than 40% performance improvement of today's processors come from semiconductor technology. Until 2033, performance scaling across 7 technology nodes (spanning from today to 2033) exhibits only 1.7× frequency improvement from semiconductor technology.
- Power density/dark silicon: Since the power per gate decreases slower than the transistor density increases, the power per mm^2 continuously increases. In consequence, if a chip has to be operated with the same power density over different technology nodes, i.e. with the same Thermal Design Power (TDP), not all transistors can switch at the same time (named "dark silicon"), or the frequency has to be reduced. Frequency at nominal supply voltage is expected to improve by 1.7× until 2033 (e.g. for high performance circuits from today's 2.5GHz to 4.2GHz), whereas power density increases at the same time by 8×. As a result, frequency has to be reduced nearly by the same factor (e.g. will stall at 0.5GHz for high performance circuits), if the same chip is operated at constant power density. Thus for the future, reducing the power consumption of a transistor becomes more important than improving its performance.
- Variability and reliability: With the continuously decreasing feature sizes, variations in the device parameters largely increase, resulting in large circuit performance/power fluctuations. In addition, the reliability of the devices decreases due to, e.g., aging effects, hot carrier injection and soft errors.

In summary, microelectronics can no longer keep pace with the increasing requirements from communication systems. Thus for Beyond-5G systems, silicon implementations of advanced channel coding schemes, that are a major building block in any wireless baseband processing, require a crosslayer approach covering information theory, algorithm

Table I
OVERVIEW ON IMPLEMENTATION PROPER

Code	Decoding algorithms	Parallel vs. serial	Locali
Turbo code	MAP	serial/iterative	low (inter
LDPC code	Belief Propagation	parallel/iterative	low (Tanne
Polar code	Successive Cancellation/List	serial	high

development, parallel hardware architectures and semiconductor technology. The Horizon 2020 funded EU project EPIC [1] addresses these challenges and aims to develop new Forward Error Correction (FEC) schemes in advanced semiconductor technology nodes for future Beyond-5G use cases targeting a throughput in the $\mathrm{Tbit/s}$ range and $\mathrm{pJ/bit}$ energy efficiency. Focus will be on the most advanced FEC schemes, i.e. Turbo codes, Low-Density Parity-Check (LDPC) codes and Polar codes [5].

$100\mathrm{mm}^2$ area is a feasible size for a baseband processor chip [3]. Furthermore we assume that 10% of this area is allocated to the FEC Intellectual Property (IP). Due to the fact that the power envelope for future communication systems cannot be largely increased, designs are more and more power constrained. Thus, a $1\mathrm{W}$ power envelope is feasible for the FEC IP, resulting in a power density of around $100\mathrm{mW/mm}^2$. The maximum frequency is upper bounded to $1\mathrm{GHz}$ due to power and design issues. Provided that the power is constrained, increasing the throughput requires decreasing the energy per decoded bit by the same order. For $1\mathrm{Tbit/s}$ data throughput, 1000 information bits have to be decoded in each clock cycle with an energy budget of only $1\mathrm{pJ}$ per decoded bit with an area efficiency of $100\mathrm{Gbit/s/mm}^2$. Efficient implementations targeting these objectives require architectures with large locality, regularity and large parallelism. However, there are discrepancies between information theory objectives and these efficient implementation objectives. Advanced channel codes like Turbo codes and LDPC codes combine irregularity and iterative/sequential decoding techniques to achieve very good communications performance which in turn hampers an efficient silicon implementation. Table I summarizes implementation properties of the most common decoding algorithms for the different code classes.

In the following, let N be the block size and R the rate of a channel code and let I denote the number of iterations that a corresponding iterative decoder requires to decode a code block (in the case of non-iterative decoding $I = 1$). Furthermore, let P denote the degree of achievable parallelism, i.e. the ratio between the operations (computations/data-transfers) that are performed in parallel per clock cycle and the total number of operations necessary to perform one decoding iteration for a complete code block. The throughput (information bits per second) of a FEC architecture can then be roughly estimated by

$$T_{inf} = N \cdot R \cdot \frac{1}{I} \cdot P \cdot f \cdot (1 - \omega), \quad (1)$$

where f is the clock frequency and ω is a normalized value between 0 and 1 that indicates the timing overhead due to e.g. data distribution, routing, memory access conflicts etc. The achievable parallelism P strongly depends on the properties of the decoding algorithms, e.g. algorithms with inherent parallelism are easier to parallelize (i.e. larger P) on architectural level (see Table I). The maximum clock frequency f is typically determined by the critical path in the compute kernels of the corresponding decoding algorithms (see Table I) and is upper limited to $1\mathrm{GHz}$. The overhead ω increases with N and P and is larger for decoding algorithms that have limited locality and are data-transfer dominated (see Table I). The impact of ω on the throughput can be considered as an effective reduction of the maximum clock frequency f and/or a decrease in P, if additional clock cycles are mandatory, e.g. due to memory conflicts, that cannot be hidden.

If we are targeting $1\mathrm{Tbit/s}$ throughput with a frequency limit of $1\mathrm{GHz}$, the minimum block size N is 1000 information bits. Obviously, to achieve highest throughput, P has to be maximized and ω minimized:

For Turbo code decoders that are based on the MAP algorithm, the critical computation is the calculation of a trellis step in the MAP algorithm. $2 \cdot N$ trellis computations and corresponding interleaving have to be calculated to perform one Turbo code decoding iteration. Hence P is determined by the number of parallel trellis step computations performed by a decoding architecture. The maximum value of P can be achieved by an architecture if 1) the forward/backward recursions of the MAP algorithm are unrolled and pipelined, yielding the XMAP architecture, or 2) a fully parallel MAP is applied in which every trellis step is spatially parallel executed, yielding the FPMAP architecture [8]. Here, the interleaver has a strong impact on ω.

LDPC codes are typically decoded with the Belief Propagation algorithm in which a huge number of messages has to be exchanged between variable and check nodes in one decoding iteration. The number of exchanged information corresponds to the number of one-entries in the parity-check matrix H. Here, P mainly depends on the number of parallel exchanged messages. The maximum value of P can be achieved if all edges of H are processed in parallel yielding a fully parallel architecture. Since the Belief Propagation is datatransfer dominated (Table I), ω largely increases for increasing N. Moreover, ω also depends on the structure of H.

Turbo code and LDPC code decoding are performed iteratively, which impacts the throughput. The data dependencies of the iterative decoding can be broken up by unrolling the corresponding iterations and appropriate architectural pipelining [7], [8]. In this way, the dependency of the throughput on I diminishes at the cost of additional pipeline memory since at

Table II
COMPARISON OF CHANNEL CODE DECODERS

Code	Blocksize [bit]	Code rate	Frequency [MHz]	Throughput [1] [Gbit/s]	Area [mm²]	Power [mW]	Area efficiency [Gbit/s/mm²]	Energy efficiency [pJ/bit]
Turbo code (4 iter)	128	1/3	800	102	23.6	-	4.34	-
LDPC code (4 iter)	1200	4/5	400	480	2.79	3000	172	6.3
Polar code	1024	1/2	746	764	2.95	3300	259	4.4

[1]Note that throughput here refers to coded throughput (opposed to information throughput in Equation 1).

least I blocks are processed in parallel in the decoding pipeline that performs all iterations in parallel on different data blocks.

Successive Cancellation (SC), Successive Cancellation List (SCL) are the most prominent decoding algorithms for Polar codes. Decoding corresponds to a traversal of the corresponding Polar Factor Tree (PFT) in which the received loglikelihood ratios from the channel are processed by the tree nodes. SC and SCL decoding are depth-first traversals on the PFT and thus exhibit sequential behavior. To achieve a maximum P, the tree traversal can be unrolled and pipelined, alike to the iteration unrolling in Turbo code and LDPC code decoding architectures. Whenever a node is visited during the tree traversal, a corresponding pipeline stage can be instantiated. In this way, for a block length of N, the maximum number of pipeline stages is $2\cdot(2N-2)+1$ in which $N\cdot\log N$ operations are performed in parallel and can be reduced by various transformations. For example, if a subtree represents a repetition code or a parity check code, the corresponding subtree can be replaced by a single node. Alike, we can merge rate-0 and rate-1 nodes into their parent nodes or use majority logic decoding in subtrees. These optimizations strongly depend on the position of the frozen bits, i.e. the code structure. Hence, appropriate codes are mandatory.

The above mentioned discussions show that decoder architectures for highest throughput are feasible for all three code classes. The achievable throughput strongly depends on the code class, i.e., the code structure and the decoding algorithm. Maximum throughput can be achieved by heavily pipelined architectures that enable maximum functional parallelism, provide large locality but at the cost of huge number of storage elements and large latency. These storage elements are a major source of the power consumption and imply large challenges on the clock tree. We have shown that more than half of the power consumption is consumed by these storage elements only. Hence, optimizing the storage scheme in pipelined decoder architectures is of great importance and has to be performed on various levels: e.g. efficient quantization on algorithmic level, advanced retiming to optimally distribute the pipeline stages between the compute units on architectural level and the use of latch-based design, clock gating etc. on micro architectural level.

In the following we show decoders for Turbo codes, LDPC and Polar codes respectively that are based on the aforementioned schemes and optimized for highest throughput, i.e. $P = 1$ and unrolled iterations ($I = 4$) for LDPC and Turbo decoding. All decoders were implemented on a 28nm FD-SOI technology under worst case PVT conditions (0.9V for timing and 1.0V for power, both $125\,^{\circ}\text{C}$). Synthesis is performed using Design Compiler, Place & Route with IC-Compiler, both from Synopsys. The implementation results are summarized in Table II and Figures 1, 2 and 3 show the respective layout.

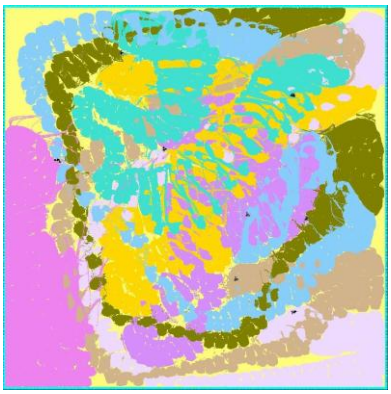

Figure 1. 102 Gbit/s Turbo decoder. The area is 23.61mm^2. Different colors represent the eight different MAP decoders originating from the 4 unrolled iterations (each iteration requires two MAP decoders).

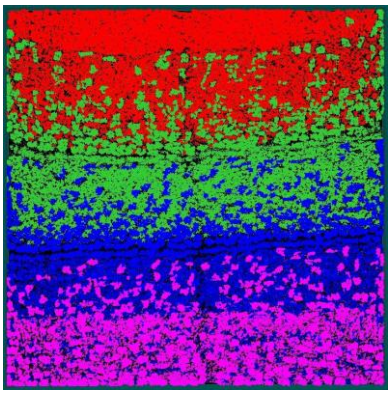

Figure 2. 480Gbit/s LDPC decoder. The area is 2.79mm^2. Each color represents check and variable node functional units corresponding to one iteration (4 in total).

It must be considered, that the presented architectures suffer from limited flexibility in terms of block sizes (all three

codes), varying number of iterations (Turbo code, LDPC code) and code rate flexibility (LDPC code and Polar code) and exhibit a large latency due to the pipelining stages. Summarized, the biggest challenges for very high throughput decoder architectures are:

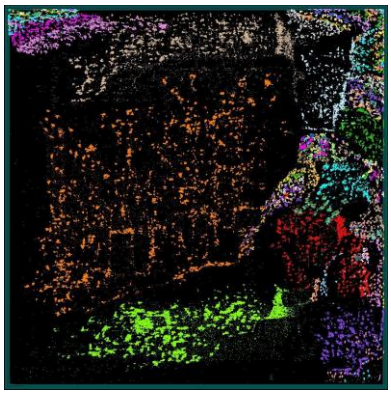

Figure 3. 764Gbit/s Polar decoder. The area is 2.95mm^2. Each color represents a pipeline stage (105 in total), memory is colored black.

- Improving the communications performance under the aforementioned implementation constraints.
- Providing block size, code rate and algorithmic flexibility.
  - Power density is in the order of 1W/mm^2, which is far too high for air cooled packages.

As discussed in the beginning, microelectronic progress will largely contribute to an improved area efficiency but not much to an increased performance and a reduced power density. Thus, further research is mandatory to keep pace with the increasing requirements on communication systems in terms of throughput, latency, power/energy efficiency, flexibility, cost and communications performance, which is in the focus of the EPIC project.

References

[1] EPIC - Enabling Practical Wireless Tb/s Communications with Next Generation Channel Coding - https://epic-h2020.eu/results.

[2] Foundry Challenges in 2018 - https://semiengineering.com/foundrychallenges-in-2018/.

[3] International Roadmap for Devices and Systems (IRDS) 2017 Edition: More Moore https://irds.ieee.org/images/files/pdf/2017/2017IRDS MM.pdf.

[4] Jorg Henkel, Lars Bauer, Nikil Dutt, Puneet Gupta, Sani Nassif, Muham-¨ mad Shafique, Mehdi Tahoori, and Norbert Wehn. Reliable on-chip systems in the nano-era: Lessons learnt and future trends. In *Proceedings of the 50th Annual Design Automation Conference on - DAC '13*, page 1, Austin, Texas, 2013. ACM Press.

[5] C. Kestel, M. Herrmann, and N. Wehn. When channel coding hits the implementation wall. In *2018 IEEE 10th International Symposium on Turbo Codes Iterative Information Processing (ISTC)*, pages 1–6, December 2018.

[6] G. E. Moore. Cramming More Components Onto Integrated Circuits. *Electronics*, pages 114–117, April 1965.

[7] P. Schlafer, N. Wehn, M. Alles, and T. Lehnigk-Emden. A new dimension¨ of parallelism in ultra high throughput LDPC decoding. In *SiPS 2013 Proceedings*, pages 153–158, October 2013.

[8] S. Weithoffer, C.A. Nour, N. Wehn, C. Douillard, and C. Berrou. 25 Years of Turbo Codes: From Mb/s to beyond 100 Gb/s. *International Symposium on Turbo Codes & Iterative Information Processing (ISTC), December, 2018, Hong Kong, China*, 2018.