Approximate computing in B5G and 6G wireless systems: A survey and future outlook

Hans Jakob Damsgaard a,*, Aleksandr Ometov a, Md Munjure Mowla b, Adam Flizikowski b, Jari Nurmi a

a Tampere University, Korkeakoulunkatu 1, Tampere, 33720, Finland
b IS-Wireless, Pulawska 45B, Pliścienko, 05-500, Poland

A R T I C L E   I N F O
Keywords:
Approximate computing
Beyond 5G
Mobile networks

A B S T R A C T
As modern 5G systems are being deployed, researchers question whether they are sufficient for the oncoming decades of technological evolution. Growing numbers of interconnected intelligent devices put these networks under tremendous pressure, demanding their development. Paving the way for beyond 5G and 6G systems, commonly denoted by B5G herein, therefore means seeking enablers to increase efficiency from different perspectives. One novel look on this is the application of inexact computations where nine 9s reliability is not needed, for example, in non-critical mobile broadband traffic. The paradigm of Approximate Computing (AxC) focuses on such areas where constrained quality degradation results in savings that benefit the users and operators. This paper surveys the state-of-the-art publications on the intersection of AxC and B5G systems, identifying and emphasizing trends and tendencies in existing work and directions for future research. The work highlights resource allocation algorithms as particularly mesmerizing in the former, while research related to Intelligent Reflective Surfaces appears the most prominent in the latter. In both, problems are often NP-hard and, thus, only solvable using heuristics or approximations, Successive Convex Approximation and Reinforcement Learning are most frequently applied.

1. Introduction

Modern wireless cellular systems constantly evolve in response to new application requirements identified in preparation for other envisioned use cases [1,2]. This has recently led to systems switching from 4G to 5G, with such transitions happening roughly once a decade. With the well-founded status of 5G systems only waiting for widespread deployment, researchers have turned their attention to the technological enablers for next-G – beyond-5G or 6G–systems [3–5]. For simplicity, we refer to all next-G systems, including beyond-5G and 6G, as B5G in the remainder of the paper. This transition has already begun. However, as the popularity of smartphones and Internet of Things (IoT) devices continues to grow, so does the amount of data being transferred [6,7], and although 5G has yet to be fully deployed, some question whether it is sufficient for the next ten years [2].

The B5G systems are expected to enable new classes of applications, tying the physical world closer to the digital one by delivering near-instant responsiveness, previously unmatched reliability, and better scalability [8]. Specifically, researchers envision much greater numbers of wearables, implants, and other cyber–physical systems; fully autonomous cars; wireless cameras and sensors with Artificial Intelligence (AI)-based data processing; robots in homes and health care; and a greater reliance on ultra-low latency communications in the industry [8–10]. Expectedly, these applications will change our lives drastically. However, they also imply enormous challenges for the underlying networking infrastructure, especially in terms of throughput, latency, reliability, security, and scalability [11].

Solving these challenges requires further research into several technologies identified as crucial for the success of B5G. Related technologies include AI, cell-free communications, dynamic network slicing, beamforming, Intelligent Reflective Surfaces (IRSs), big data analytics, massive Multiple-Input Multiple-Output (MIMO), extremely high-frequency air interfaces, and new security primitives [12,13]. Most of these are computationally heavy, yet to keep with current trends and avoid enormous increases in overall energy consumption, they must be implemented optimally and efficiently. However, as we have reached or gotten near the end of helpful historical trends such as Moore’s law and Dennard scaling, we can no longer expect to match these requirements simply by microprocessors becoming faster every
year [14] – modern devices with computing capabilities are expected to require more energy than the energy resources can provide by 2040, according to Semiconductor Industry Association [15], see Fig. 1. Radical changes are needed to address this issue.

In this direction, Approximate Computing (AxC) appears as an alluring technology [16]. In AxC, approximations are applied at different system levels to reduce computational complexity, transforming into savings in various system metrics [17,18]. The paradigm includes a wide range of techniques that fall into two categories [19]:

**Computing on approximate data** means applying approximations at the application or architecture level. The textbook example is the quantization of floating-point operations into simpler fixed-point equivalents. Yet, despite being a powerful and frequently applied technique, quantization alone usually fails to fully exploit applications’ error resilience. Supplementing it with orthogonal techniques like loop perforation and skipping, neural approximation, inexact arithmetic, etc., is often possible and may increase savings [18,20,21].

**Computing on unreliable hardware** means introducing approximations at the circuit level by using unintentionally faulty circuits or purpose-designed ones. Putting faulty chips that produce acceptable errors to use can increase production yield [22], while purpose-designed ones may bring significant savings from, e.g., over-scaled operating voltage or frequency, reduced refresh rate Dynamic RAM, approximately synthesized logic, etc. [18,20,23].

Using techniques from either or both of these categories enables interesting time-energy-accuracy trade-offs. Generally, applications that produce outputs evaluated on their acceptability rather than correctness are likely candidates for approximation. The most well-known examples of such are from the AI and Big Data analysis fields. Nevertheless, many algorithms within 5G systems pertain the same feature, including scheduling, resource allocation, user-channel pairing, and other optimization algorithms, as illustrated in Fig. 2. When handling non-critical applications that do not demand nine 9s reliability like Enhanced Mobile Broadband (eMBB) traffic, these algorithms may be approximated to induce benefits for users and operators. The potential gains thereof are plentiful and include:

- **Enhanced robustness** as systems integrating AxC inherently tolerate a certain amount of imprecision in their computations, they also show greater resilience to faults and errors [26]. This is particularly important in the context of wireless networks that may be subject to interference, signal degradation, and other sources of noise.

- **Strengthened flexibility and adaptability** from reducing the complexity and compute requirements of the components of 5G algorithms, improving overall system scalability. With AxC, a system can be tailored more closely to the accuracy needs of different applications, enabling customization of computing solutions to suit diverse application requirements [27]. This in turn may enable deploying larger and more complex wireless networks or addressing the demands of 5G use cases better [28].

- **Increased cost efficiency** arising from the aforementioned reduced hardware and processing requirements. These reductions may not only bring savings in hardware and power consumption, but also in system maintenance as known faulty circuits may continue to operate so long as they deliver acceptable accuracy levels [19,22], making AxC an attractive option for deploying cost-effective 5G systems [13,29].

In this paper, we keep these benefits in mind and survey the state-of-the-art in AxC applied to 5G systems, focusing on techniques applied at the application level. We do not provide a comprehensive survey of approximate computing techniques nor cover all details expected from 5G systems. For those interested in these two topics, please refer to [17,18,20,23,30–32] and [2,8,13,24,33,34], respectively.
The main contributions of this paper are:

- A literature survey on algorithms for B5G systems that lend themselves to any aspect of approximation;
- An overview of problem formulations, algorithms, and KPIs used in the literature;
- Identification of future research directions and the potential role of AxC in these.

We have followed the modified PRISMA guidelines [35] for reproducibility purposes. As such, we defined the following search expression with keywords found in relevant articles: (approximate OR inexact OR inaccurate OR ’good enough’) AND (B5G OR ’beyond 5G’ OR 6g). Using this expression, Scopus [36] returned 269 results, which we filtered in two rounds: first, a coarse filtering by paper titles, and second, a fine filtering by abstracts. In both rounds, we applied the following exclusion criteria: C1 no relation to AxC or B5G systems; C2 work not applying approximations at runtime; C3 reliance on generally inaccessible technologies like airplanes and Unmanned Aerial Vehicles; and C4 full text not available. We also filter clearly iterative work to include only the most recent publication and limit inclusions to direct search results. This results in the set of 80 papers presented in the following.

The remainder of this paper is structured as illustrated in Fig. 3. First, Section 2 covers the main technical part of the survey, reviewing and summarizing work on existing and future technologies. Next, Section 3 discusses overall observations of the survey combined with ideas and summarizing work on existing and future technologies. Next, Section 3 discusses overall observations of the survey combined with ideas and summarizing work on existing and future technologies. Finally, Section 4 concludes the paper.

2. Approximate computing in B5G systems

Having motivated our survey and presented the strategy for finding relevant work, let us first consider a set of publications that do not naturally fall into the two categories above. We identify three papers that present hardware architectures relevant to the physical layer of future wireless communication systems.

Specifically, Madanayake et al. [37] propose several approximate Discrete Fourier Transform implementations suitable for beamforming antenna arrays. Their designs avoid costly multipliers, utilizing sparse matrices with signed unit-valued entries, only of which is complex. Doing so allows for significant area savings with negligible impact on beamforming performance. Idrees et al. [28] similarly focus on the Digital Signal Processing (DSP) side of wireless networks, applying inexact arithmetic in the Base Station (BS)’s pulse shaping filters and the User Equipment (UE)’s decoders. As a result, they reduce power consumption drastically at the expense of increased Bit Error Rate (BER). Finally, Morsali et al. [38] focus on massive MIMO transceivers and propose a novel approach to construct Radio Frequency-based analog Deep Neural Network (DNNs), enabling them to implement the transceiver with a digital DNN for signal processing and an analog one for beamforming. Being limited to Downlink (DL) transmissions, the resulting architecture achieves the same performance as a fully digital equivalent while being more adaptive to transmitter/receiver mappings and consuming less hardware.

All three papers clearly highlight the potential benefits of applying AxC at the architectural level. However, they are limited in their scope to the DSP parts of communication systems. Broadening the scope to the application level opens many more opportunities for approximations. As such, we will first consider such publications that apply AxC to well-established technologies, most of which are already in use in 5G and prior systems.

2.1. Existing technologies

The remaining work is challenging to categorize; thus, there may be better ways to group it than by key topics as in the following. The publications are characterized by mostly not applying approximations in the notion common to AxC. Rather, they propose various approximate algorithms for problems that are difficult, if not nearly impossible, to solve optimally in the field with brute force algorithms. Within the scope of this survey, we consider these to be approximations at the application level. Our survey reveals that such approximations are frequently applied to resource allocation algorithms, as we will see later. We begin by surveying papers on scheduling algorithms.

2.1.1. Scheduling

Scheduling refers to the problem of determining when to perform particular actions based on resource availability, system constraints, and user expectations, and it is vital for perceived Quality of Experience in multi-user systems. We include five papers on this topic that focus on mostly different scenarios, as highlighted in Table 1. First, Li et al. [39] consider channel and queue-aware scheduling of a multi-user DL. They evaluate a time-multiplexed scheme that first schedules the user with the best channel conditions in each time slot, applying approximations by considering users individually and modeling their number of queued packets with Markov chains. Their model matches the simulated performance well. Similarly, Salameh et al. [40] consider opportunistic scheduling of non-contiguous Orthogonal Frequency-Division Multiple Access (OFDMA) Resource Blocks in a cognitive radio network, improving spectrum utilization flexibility and efficiency. They formulate it as a Binary Linear Programming (BLP) problem and propose an approximate polynomial-time algorithm to find the solution, claiming double the performance of contiguous time and contiguous frequency algorithms.

Second, Almehlafi et al. present work on scheduling Ultra-Reliable Low-Latency Communication (URLLC) traffic onto resources occupied by ongoing eMBB traffic through puncturing [41] or superposition [42], i.e., by either overriding eMBB symbols (forcing them to be re-scheduled) or transmitting them simultaneously with differently modulated URLLC symbols. Either technique is essential to satisfy the URLLC packets’ strict latency requirements. In the former work, they propose to minimize interference on URLLC symbols using symbol similarity, sought after by one of three linear-time routines. In the latter, they formulate the problem as a Mixed Integer Non-Linear Programming (MINLP) problem and propose polynomial-time approximate algorithms to solve it. Both report improvements in Symbol Error Rate (SER) and the number of users served.

And last, Beitollahi et al. [43] consider a Federated Learning (FL) environment in which many devices must communicate their updated...
models to a centralized controller. They propose making scheduling decisions for several update iterations – frames – centrally, taking transmission errors into account, and formulate this as a MINLP problem, solved using an approximate algorithm. The approach outperforms traditional round-robin scheduling.

2.1.2. Resource allocation

While scheduling may be seen as a branch of resource allocation, it is far from the only one. As we display in this section, resource allocation is central to many parts of 5G systems. We group papers focusing on similar topics, but readers may notice some disparity between proposals within the groups. Their main points are listed in Table 2.

Power allocation is a valuable technique for two purposes: (1) managing or limiting total transmit power and (2) enabling transmit power-based Non-Orthogonal MA (NOMA) schemes [71]. The latter appears particularly attractive and is described as a way to increase spectral efficiency in 5G systems by Hashemi et al. [44], who analyze the performance of Uplink (UL) delta-Orthogonal MA (OMA) and optimize its energy efficiency. They model the situation as a Mixed Binary Non-Linear Programming (MBNLP) problem and apply the outer approximation to solve it, leading it to outperform NOMA and OFDMA alternatives under certain conditions. Yin et al. [50] instead jointly optimize device assignment and power allocation in a NOMA system using iterative approximation and SCA, reporting significant improvements in network throughput.

Minimizing total transmit power is the focus of two other papers. Firstly, Salameh et al. [45] focus on a hybrid OFDMA-NOMA network, modeling the minimization problem as a non-convex optimization, MINLP, solved using an approximate algorithm. The approach outperforms throughput, latency.

Table 2

<table>
<thead>
<tr>
<th>Class</th>
<th>Ref.</th>
<th>Problem</th>
<th>Formulation</th>
<th>Algorithm</th>
<th>KPIs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>[45]</td>
<td>Total power optimization in OFDMA-NOMA</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Throughput, transmit power</td>
</tr>
<tr>
<td></td>
<td>[46]</td>
<td>Massive MIMO performance optimization with imperfect hardware</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Sum rate</td>
</tr>
<tr>
<td></td>
<td>[48]</td>
<td>Throughput optimization in hybrid NOMA</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Admissions, associations, power/user, throughput</td>
</tr>
<tr>
<td></td>
<td>[49]</td>
<td>Energy efficiency optimization in hybrid NOMA</td>
<td>Concave fractional optimization, MINLP</td>
<td>SCA, iterative approximation</td>
<td>Throughput</td>
</tr>
<tr>
<td></td>
<td>[50]</td>
<td>Power allocation in heterogeneous multi-carrier NOMA</td>
<td>MINLP</td>
<td>SCA</td>
<td>Admissions, associations, power/user, throughput</td>
</tr>
<tr>
<td></td>
<td>[51]</td>
<td>Bandwidth allocation in multi-user Multiple-Input Single-Output</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Throughput</td>
</tr>
<tr>
<td>User-Channel Pairing</td>
<td>[52]</td>
<td>Clustering in user-centric networks</td>
<td>Iterative optimization</td>
<td>DNN</td>
<td>Sum rate</td>
</tr>
<tr>
<td></td>
<td>[53]</td>
<td>Channel allocation in URLLC</td>
<td>Non-convex optimization</td>
<td>Inner approximation</td>
<td>(Secrecy) sum rate</td>
</tr>
<tr>
<td></td>
<td>[55]</td>
<td>User association in in-X networks</td>
<td>Centralized graph coloring</td>
<td>DNN</td>
<td>Loop failure probability</td>
</tr>
<tr>
<td></td>
<td>[56]</td>
<td>Beamforming and channel allocation in multi-user URLLC</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Decoding error probability</td>
</tr>
<tr>
<td></td>
<td>[57]</td>
<td>User pairing in NOMA communications</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Throughput</td>
</tr>
<tr>
<td></td>
<td>[58]</td>
<td>Channel allocation in hybrid OFDMA-NOMA</td>
<td>Mixed Integer Linear Programming (MILP)</td>
<td>SCA, RL</td>
<td>Energy efficiency, computation time, Throughput, latency</td>
</tr>
<tr>
<td></td>
<td>[60]</td>
<td>Traffic steering in open-RAN networks</td>
<td>Non-convex optimization</td>
<td>Iterative approximation, GNN</td>
<td>Throughput, latency</td>
</tr>
<tr>
<td>IAB</td>
<td>[61]</td>
<td>Association and power allocation in IAB</td>
<td>MILP</td>
<td>SCA</td>
<td>Sum rate</td>
</tr>
<tr>
<td></td>
<td>[62]</td>
<td>Bandwidth allocation in IAB</td>
<td>Non-convex optimization</td>
<td>Iterative approximation, GNN</td>
<td>Spectral efficiency</td>
</tr>
<tr>
<td>Random Access</td>
<td>[63]</td>
<td>Active device detection</td>
<td>Odd total squared correlation</td>
<td>Approximate message passing</td>
<td>SNR, perfect recoveries, missed detections, false alarms</td>
</tr>
<tr>
<td></td>
<td>[64]</td>
<td>Distributed congestion prediction for latency optimization</td>
<td>Non-convex optimization</td>
<td>RL</td>
<td>Latency</td>
</tr>
<tr>
<td>Fault Tolerance</td>
<td>[65]</td>
<td>Hypervisor adaptation optimization</td>
<td>k-pathwise disjoint cover</td>
<td>Heuristic, ILP</td>
<td>Acceptances, migrations, Bandwidth, survivability</td>
</tr>
<tr>
<td></td>
<td>[66]</td>
<td>Deterministic fault tolerant service function chaining</td>
<td>k-pathwise disjoint cover</td>
<td>Heuristic</td>
<td>Acceptances, migrations</td>
</tr>
<tr>
<td>Others</td>
<td>[67]</td>
<td>Network service federation admission control</td>
<td>Markov decision process</td>
<td>RL</td>
<td>Optimality gap</td>
</tr>
<tr>
<td></td>
<td>[68]</td>
<td>Spectrum assignment to service providers</td>
<td>—</td>
<td>—</td>
<td>Capacity, energy/ spectral/ cost efficiency</td>
</tr>
<tr>
<td></td>
<td>[69]</td>
<td>Task scheduling and power allocation in Fog computing</td>
<td>Non-convex optimization</td>
<td>SCA</td>
<td>Latency</td>
</tr>
<tr>
<td></td>
<td>[70]</td>
<td>Disparate groups QoS optimization</td>
<td>Markov decision process</td>
<td>—</td>
<td>QoS</td>
</tr>
</tbody>
</table>

models to a centralized controller. They propose making scheduling decisions for several update iterations – frames – centrally, taking transmission errors into account, and formulate this as a MINLP problem, solved using an approximate algorithm. The approach outperforms traditional round-robin scheduling.
network performance in the number of admissions, associations, energy efficiency, and throughput over a macro BS architecture.

Wang et al. [46] and Farooq et al. [47] aim to optimize massive MIMO systems for different cases, respectively, when using imperfect hardware and under Federated Learning (FL)-related traffic. Like [45], both use non-convex optimization models and apply SCA to solve them, achieving improved sum rates.

**User-channel pairing** is relevant for allocating channel resources to users in BSs, a problem closely linked to power and bandwidth allocation optimizations. The publications on this topic are diverse, focusing on many different network configurations. Starting with ultra-dense networks, Zhu et al. [54] model the collaborative optimization of user association and channel allocation as a non-convex problem. They apply variable relaxation to convert the problem to a convex one, which they proceed to solve, nearly doubling system energy efficiency.

Dai et al. [53] and Li et al. [56] both present approximate algorithms for non-convex optimization of channel allocation in networks supporting URLLC traffic; Dai et al. apply a game theory approach and inner approximation, and Li et al. use SCA instead. The authors report improved (secrecy) sum rates.

Two other papers consider NOMA-based networks. Narottama et al. [57] seek to optimize user pairings. They propose an evolutionary algorithm inspired by quantum systems and show how it outperforms the typically applied random algorithm. Abdel et al. [58] seek to maximize the sum rate in hybrid OFDMA-NOMA by clustering users and assigning them to transmit powers. As in [56], the problem is non-convex and solved by SCA, outperforming conventional algorithms.

A final pair of papers both propose using DNNs in place of compute-heavy algorithms. Specifically, Liu and Zhang [52] focus on user-centric networks and use a DNN to approximate a weighted minimum mean square error algorithm for user clustering and allocation. Adegun et al. [55] replace a costly graph coloring algorithm for channel selection in so-called in-X networks (e.g., in-vehicles) with a DNN. The papers claim comparable performance at two orders of magnitude lower complexity and minimal input data, respectively.

**Traffic steering** is considered in another pair of publications. Pham [59] and Kavehmadavani et al. [60] consider virtualized networks, the former on maximizing cost efficiency in service function chaining while guaranteeing delay, and the latter on simultaneously maximizing eMBB throughput and minimizing URLLC latency. Both model their problem as a non-convex optimization and solve it using Reinforcement Learning (RL) and SCA, respectively, showing improvements in the targeted Key Performance Indicators (KPIs).

**Integrated Access and Backhaul** is a central concept in 5G systems, ensuring low-latency connectivity by letting next-Generation Node Bs wirelessly relay traffic between one another over frequency bands otherwise used to service UEs rather than directly to a wired backhaul link. This problem is considered in two papers. Zhang et al. [61] jointly optimize user association and power allocation, like [48, 49], utilizing a novel backhaul capacity-aware utility function. Much like the related work, the problem is a non-convex optimization solved using SCA, resulting in a greatly improved network sum rate. Orthogonally, Harris et al. [62] consider bandwidth allocation and propose an iterative approximate algorithm and a Graph Neural Network (GNN)-based algorithm to solve it. Their experiments reveal that combining the two is a practical solution that improves spectral efficiency.

**Random access** operations must be handled rapidly in 5G systems to satisfy their availability and latency requirements [2]. This operation, however, may be tackled from two sides: the UE’s and the BS’s. Liu et al. [63] focus on the latter, presenting a scheme targeting massive Machine Type Communication. They optimize the active device detection stage of envisioned grant-free random access using the correlation of sets of received binary signatures, reducing the complexity with approximate message passing. The resulting algorithm outperforms the commonly applied Gaussian matrix technique used in compressed sensing.

Bekele and Choi [64] focus on the UE’s perspective and propose a distributed RL algorithm to predict Access Point (AP) congestion. The authors show how selecting a low-congestion AP can vastly reduce random access latency and increase the probability of successful access, and display that their algorithm outperforms 3GPP’s default algorithm.

**Fault tolerance** becomes ever more critical as requirements for network availability increase beyond seven nines [2], challenging both physical and virtualized network cores. Two papers focus on the latter, presenting algorithms for improving reliable service function chaining. Specifically, Mogyrorsi et al. [65] focus on hypervisors, minimizing their overheads and maximizing their fault tolerance while ensuring support for instantaneous changes in traffic, while Zheng et al. [66] propose a fault-tolerant function slicing approach. Both formulate their cases as k-pathwise disjoint cover problems and use heuristics to solve them, improving service acceptance and survivability.

**Others** include work that does not naturally fit the above categories. Bakhshi et al. [67] consider admission control in network service federation across providers and, thus, administrative domains. They formulate the problem as a Markov decision process but propose using Q-learning or R-learning to solve it in practice, revealing R-learning as the superior algorithm that achieves near-optimal performance.

Orthogonally, Fei et al. [69] consider a Fog computing environment [72] and jointly optimize task scheduling with UL/DL power allocation. Like much of the above work on power allocation [45–47, 51], the authors formulate the problem as a non-convex optimization and solve it using SCA. The results indicate improved user fairness and reduced system latency. Iyer [68], again orthogonally, proposes a dynamic scheme for spectrum assignment across different service providers according to their instantaneous number of subscribers. Through simulation, they show improved capacity and efficiency of the spectrum, energy, and cost.

Last, Ahmed et al. [70] model the Quality of Service (QoS) in heterogeneous networks with two classes of users: privileged and shadowed. They describe two use cases with different user class mixtures, formulate them as Markov models, and evaluate different admission restrictions. Yet, they do not propose a way to optimize for QoS.

### 2.1.3. Channel utilization

Having surveyed much work on resource allocation, let us shift our focus to more channel-related topics. We begin by considering work on various techniques for massive MIMO communications: precoding, demodulation, and data detection, and continue with publications on beamforming, channel estimation, multiple access, and security. While some of these topics are part of the above publications, they are not their main topics. Once again, we tabulate the related publications in Table 3.

**Massive MIMO** is expected to improve signal strength for individual devices and spectral efficiency in general at the expense of much higher computational demands. Meeting these is important for enabling the use of higher frequency bands [8]. As such, we survey some relevant algorithms.

Firstly, four papers consider precoding algorithms used to minimize inter-user interference. Yan et al. [73] and Busari et al. [74] present hybrid digital–analog precoding schemes whose linear algebra is approximated, showing how doing so minimally affects their performance. Similarly, Ribeiro et al. [76] propose a fully digital precoding scheme that combines users with groupings in a mathematical approximations, achieving near-optimal performance at much lower complexity. Orthogonally, Li et al. [75] consider a rate-splitting scenario in which multiple parallel transmissions of the same data serve to reduce interference, improving overall channel performance. They optimize the achievable rate by modeling different precoder designs as a non-convex
problem and assuming joint decoding, commonly applied in rate splitting. Applying SCA, their results show that rate splitting outperforms traditional unicast transmissions.

Secondly, two publications consider demodulation, i.e., identifying subscriber information in the masses of radio signals received at a MIMO antenna. In both their papers, Bakulin et al. [77,78] apply a non-Gaussian approximation of the prior distribution of parameters for demodulation and use Newton’s method to optimize them. They report that the proposed schemes reduce the frame error rate under different channel conditions.

Lastly, another four papers present algorithms for data detection, i.e., detecting (UL) traffic in multi-user, multi-antenna interference-filled systems. Essentially tackling the same problem, they propose four different detector designs: (1) Kosash et al. [79] combine expectation propagation with GNNS for a design that works well under strong multi-user interference; (2) Bica et al. [80] propose two designs using Gaussian approximation or DNNs; (3) Hasan et al. [81] use coordinate descent (another approach to solving convex optimizations) to design an adaptive detector; (4) Kumar et al. [82] propose using QR-decomposition and maximum likelihood combined with several mathematical estimations. All four papers report improved SER or BER over traditional schemes at lower computational complexity.

Beamforming is essential in ensuring good signal strength despite issues with the attenuation of high-frequency signals. Minimizing related computations can improve its performance, especially in multi-user scenarios. Three publications consider this scenario, modeling it as a non-convex optimization but maximizing its energy efficiency with different different approximate algorithms. Specifically, Kaushik et al. [83] combine SCA and linear approximations, Fozi et al. [84] use RL with per-user QoS constraints and describe a novel training algorithm based on approximate message passing, while Reifert et al. [85] jointly optimize energy efficiency and user-Cloud association with fractional programming and inner approximation. All three report that their algorithms outperform traditional approaches.

Channel estimation refers to computing the propagation-related parameters of a wireless channel in a General Frequency-Division Multiplexing (GFDM) system. It is an inherently approximate process and, thus, an obvious place to apply AxC. While none of the surveyed work uses typical AxC techniques, they propose approximate algorithms, as seen above. Four publications present different algorithms specifically for this purpose.

Qiag et al. [86] jointly perform activity detection and channel estimation with a purpose-designed deep learning algorithm based on approximate message passing. Mohammadian and Tellambura [87] jointly perform channel and phase noise estimation with an iterative approximation algorithm and propose another two algorithms for data detection and phase noise compensation. Li et al. [88] formulate the situation as a compressed sensing problem in a high mobility case and use a DNN to solve it. Lastly, Shi et al. [91] propose using various types of NNs before converging on an Long Short-Term Memory-based architecture to solve the problem for both stationary and non-stationary
The proposed scheme outperforms its state-of-the-art alternatives.

and transmit power in a non-convex problem, again solved using SCA.

rate splitting scheme. They jointly optimize precoding, stream splitting,

is non-convex and solved by SCA, outperforming both rate-splitting and

different requirements that can be served simultaneously. The problem

in [92], and solved using SCA. The results show improved sum rate.

scheme for Cloud networks, optimizing its sum rate using beamforming,

simpler schemes. Ahmad et al. [93] instead consider a rate splitting

sum capacity in NOMA using SCA and report improved capacity over

NOMA schemes. They optimize for optimal mode division in OMA and

momentum encoding and explore its functionality under OMA and

[92] consider improving channel capacity with so-called orbital angular

strength through, e.g., the length of asymmetric/symmetric keys

environments. In their respective use cases, all four papers report

improved Mean SE (MSE) or Normalized Mean SE (NMSE).

Another two papers present techniques to predict Channel State

Information (CSI), otherwise communicated as a response from UE to

BS with the overheads that follow. Both publications use deep learning

techniques to solve this, but with different details. Zhang et al. [89]
present a novel complex-valued Convolutional Neural Network (CNN)

architecture, while Belgiovine et al. [90] present a simpler Multi-Layer

Perceptron architecture; both report improved error rates over more

complex least squares-based algorithms.

Multiple Access schemes may also be optimized in various ways,

often requiring large amounts of computation. Four publications do

precisely this, but each for very different cases. Specifically, Jian et al.

[92] consider improving channel capacity with so-called orbital angular

momentum encoding and explore its functionality under OMA and

NOMA schemes. They optimize for optimal mode division in OMA and

sum capacity in NOMA using SCA and report improved capacity over

simpler schemes. Ahmad et al. [93] instead consider a rate splitting

scheme for Cloud networks, optimizing its sum rate using beamforming,

user clustering, and available CSI. The problem is non-convex, as

in [92], and solved using SCA. The results show improved sum rate.

Mei et al. [29] propose dynamically switching between MA schemes

depending on individual UE needs. They combine metrics from inter-

ference cancellation and resource allocation to cluster UEs with

different requirements that can be served simultaneously. The problem

is non-convex and solved by SCA, outperforming both rate-splitting and

MIMO-based NOMA. Lastly, Khisa et al. [94] focus on a MIMO-based

rate splitting scheme. They jointly optimize precoding, stream splitting,

and transmit power in a non-convex problem, again solved using SCA.

The proposed scheme outperforms its state-of-the-art alternatives.

Security is the focus of the last group of papers, which seek to op-

imize the secrecy rate of two different network types in environments

with eavesdroppers and untrustworthy relay nodes. Bastami et al. [95]

jointly optimize beamforming and transmit power, modeled as a non-

convex problem and solved using SCA. As low latency is required in

practice, they also propose using a lower-complexity DNN trained using

their original model. Xie et al. [96] optimize cost and secrecy rate

assuming femtocell BSs using a game framework in combination with

Bernstein approximation. As a separate subgroup of papers not directly

related to AxC, lightweight security on wearables and resource con-

strained devices appears. The main concepts addressed are trade-offs

between the needed levels of security/privacy vs. cryptography primi-

tive strength through, e.g., the length of asymmetric/symmetric keys

used for data encryption/decryption [97,98]. Those aspects directly

affect the execution time and energy consumption and can be con-

sidered as approximation approaches on a higher level of abstraction.

Finally, the book chapter [99] delves deeper into general applications

of AxC in hardware security, e.g., approximate arithmetic circuits. The

majority of papers report several simulation/measurement results but,

commonly, do not provide any comparisons to other algorithms.

2.2. Future technologies

Having surveyed approximate algorithms applied to well-established

technological enablers, we now focus on technologies that have yet

to see an application in practice but are expected to become essential

parts of future 5G systems. We classify work into four categories: cell-

free networks, network slicing, Integrated Sensing and Communication

(ISAC), and Intelligent Reflective Surfaces (IRSs) or Reconfigurable

Intelligent Surfaces (RISs). Again, the publications are listed in Table 4.

2.2.1. Cell-free networks

Despite implementing massive MIMO and beamforming, 5G sys-

tems are expected to provide excellent, broad coverage in areas not

coverable by traditional means. This is not always achievable, and thus,

some network architects are considering a cell-free architect-

ure in which, rather than segmenting areas into cells covered by

a single BS, the cells are covered by numerous APs connected to a

centralized controller, as illustrated in Fig. 4. This introduces new

challenges and scalability issues that must be solved to ensure reliable,

high-performance operation [102].

Four papers consider such matters. First, Shao et al. [103] propose

a scheme to enable multiple APs to collaboratively perform activity

detection. Their algorithm is based on the covariance between so-called

device state vectors of neighboring APs and forward–backward splitting

training. It achieves lower error rate than centralized schemes and tra-

ditional cell-based massive MIMO networks. Second, Shaik et al. [104]

consider soft detection in UL massive MIMO, describing a distributed

algorithm that passes approximate bit probabilities between APs until

predictions are certain enough to facilitate decoding. The authors,

unfortunately, fail to provide an evaluation of their scheme. Third,

Peng et al. [105] devise a coordination scheme for DL traffic relay-

ing, aiming to maximize the sum rate of clustered NOMA users. This

optimization is non-convex but solvable with an iterative convex ap-

proximation. Owing to its dynamic nature, the scheme outperforms its

Fig. 4. Figurative description of (4(a)) a traditional cell network and (4(b)) an envisioned cell-free network [101]. Only the network fronthaul is shown.
simultaneously for communication and environmental sensing [124], as
utilization over state-of-the-art algorithms.

of incoming requests. Doing so reduces latency and improves resource
different algorithms and selects between them depending on the type
reducing its resource overheads.
mains. They propose avoiding this by using FL, bringing analysis closer
to data generation. Doing so drastically reduces monitoring traffic and
minimizing over-provisioning in network slices, giving higher priority
future work. Utilizing the same kind of model, Xiao et al. [109] focus on
provide access to several different services. Reminiscent of the resource
allocation publications above, four publications consider this topic.

2.2.2. Network slicing
Virtual and software-defined networks may be sliced into special-
ized compartments with different functions tailored for specific applica-
tions [107], as shown in Fig. 5. For example, this kind of slicing may enable support for high-throughput video streaming in one slice and low-throughput grant-free IoT traffic in another, managed concurrently
on the same network [8]. In practice, slices may be hierarchical and provide access to several different services. Reminiscent of the resource
allocation and solving it using generalized approximate message passing. Their
describe the situation as a compressed sensing problem
returns to the topic of ISAC, Xing et al. [116] consider IRS-aided envi-
ronmental sensing systems and propose a CNN-based correction scheme
widely varying focus points.

2.2.4. Intelligent surfaces
Much like massive MIMO is needed to operate current high-
frequency networks, moving to even higher frequency bands in BSG
systems will further the troubles of signal attenuation. While beam-
forming can aid in this, it cannot compensate for potential multi-path or
no-line-of-sight problems in urban environments. Mitigating such issues
is possible using so-called IRSs; electrical surfaces designed to reflect
interference characteristics. They jointly optimize the sum rate and
beampattern error, modeling it as a non-convex problem and solving it
using an Alternating Direction Method of Multipliers (ADMM) algo-
rithm. The resulting performance is better than comparable schemes.

2.2.3. Integrated sensing and communication
ISAC denotes a new technology in which radio resources are used
simultaneously for communication and environmental sensing [124], as
shown in Fig. 6. The latter will enable the detection of objects in space
and high-accuracy localization and positioning. Three publications fo-
cus on the optimization of this technology.

Xu et al. [112] consider a radar-communication system, applying
rate splitting MA, previously seen in [29,75,93], due to its good
interference characteristics. They jointly optimize the sum rate and
channel estimation, combining it with a pilot-based target search. The problem is non-convex, so the authors utilize several mathematical approximations to make it easier to solve. Their results show near-optimal performance. Lastly, Li et al. [114] use beam-
forming to maximize worst-case Signal to Noise Ratio (SNR) within a
region. Like [112] and [113], the authors model this as a non-convex
optimization but apply SCA to solve it. Plots indicate that the algorithm
achieves its goal, covering the targeted region well.

To solve a similar problem, Chergui et al. [110] criticize existing
algorithms for requiring a lot of monitoring data passed between do-
 mains. They propose avoiding this by using FL, bringing analysis closer
to data generation. Doing so drastically reduces monitoring traffic and
reduces its resource overheads.

Lastly, Cao et al. [111] present an approximate algorithm for re-
source allocation in network slices. Their scheme implements several differ-
dent algorithms and selects between them depending on the type
of incoming requests. Doing so reduces latency and improves resource
utilization over state-of-the-art algorithms.

static counterparts. And fourth, Conceiccao et al. [106] aim to optimize the
fairness in UL cell-free networks without relying on costly convex
approximations. They prioritize the spectral efficiency of UEs with poor channel conditions. They use metaheuristics to minimize their experienced error rates, demonstrating low latency and scalability.
results show significant error reductions when applying the proposed correction. Secondly, Ni et al. [121] propose an IRS-supported over-the-air FL scheme that aggregates models by exploiting superposition on the wireless channel. They aim to maximize model convergence given uncertain channel conditions, modeling this as a MINLP problem. They solve this using SCA and demonstrate near-optimal performance.

The remaining seven papers seek to optimize the IRS operation. Hu et al. [115] propose extending passive IRSs with deeply quantized channel estimators using generalized message passing. This optimization is non-convex and solved using ADMM. The scheme outperforms its state-of-the-art alternatives while requiring fewer receivers. Orthogonally, Alsham et al. [117] propose a resource allocation scheme for IRS-supported networks that maximizes overall energy efficiency. They formulate this as a MINLP problem and use outer approximation to solve it, demonstrating better results over a comparable scheme.

Centered around the idea of Cloud-Radio Access Network (RAN), Zhang et al. [118] jointly optimize user transmission beamforming, passive IRS beamforming, and data compression to maximize the UL sum rate. Weinberger et al. [120] maximize network energy efficiency by employing a combination of rate splitting, user clustering, and passive IRS beamforming. Both formulate their problems as non-convex optimizations and apply SCA in combination with other mathematical approximations to solve them, and both report improvements in their metrics.

Finally, Pereira et al. [122] jointly optimize the MIMO precoding and IRS phase shift matrices to maximize the network sum rate. As the problem is non-convex, they propose solving it dynamically with a combination of RL to learn policies and DNNs to approximate functions. This approach achieves comparable performance to a more complex counterpart. Along the same lines, but only focusing on phase shift matrix optimization, Nguyen et al. [119] formulate another non-convex problem that they apply ADMM and SCA to solve. The result is an improved network sum rate. Elhattab et al. [123] instead jointly optimize power allocation, user clustering, and IRS phase shift matrices. Like [119], they use alternating optimization and SCA. Their scheme outperforms both NOMA- and OMA-based alternatives.

3. Discussion and potential future directions

While the previous section has revealed a vast amount of work already done in proposing approximate algorithms for various problems in future wireless systems, many challenges remain. The surveyed papers directly point some out, while others arise from patterns in their proposals. We structure this discussion slightly differently than the above, as some publications from different categories point out similar future work.

An observation originating from the above is that most publications focus on optimizing one or more network components to improve various KPIs, including BER, accuracy, latency, sum rate, throughput, and energy, spectral, or cost efficiency. However, while some papers share goals in these directions, they rarely focus on achieving them in the same scenario, leading to disparate proposals. Combining these could lead to more generally applicable algorithms. Furthermore, it could introduce interesting new trade-offs between generality and performance. Such a direction is worth considering!

3.1. Are these algorithms realistic?

It is well known that academic research results generated from narrow-scope simulations, as in the majority of the surveyed papers, often fail to address critical constraints pertaining to real-world deployment [125]. This gap arises as the industry develops and deploys proprietary solutions to practical problems leaving academia to reverse-engineer or speculate on these. When such speculation becomes the foundation for further research, the gap gradually grows and the practical feasibility of proposed solutions shrinks. Albeit qualitative, we find it relevant to review this here.
Within the scope of this survey, we are particularly interested in whether the surveyed algorithms can be deployed to practical networks. A sign of this relevance is whether a paper has authors from the industry. This is the case in nine of the 80 surveyed papers: [38,41,54,55,65,88,92,108,113], in none of which industrial authors represent a majority of the author lists.

Furthermore, judging by Tables 1, 2, 3, and 4, the majority of the surveyed papers propose approximate algorithms to solve otherwise NP-hard problems. And while algorithms such as SCA, outer approximation, RL, and DNNs might produce approximate solutions in polynomial time to these exponential-scale problems, they still suffer from long latency. As networks become more distributed, the need for executing these algorithms closer to the BSs or APs grows, implying a need for distributed computing units [8,55,90]. However, such devices typically have extremely constrained computing resources, likely rendering them incapable of reaching sufficient throughput to execute many of the proposed algorithms in real time [8]. This is a challenge that is also highlighted by the 6G Flagship partners [126] that must be overcome to render common visions of 6G realistic [8]. Thus, research into good, reduced-complexity alternatives, such as ML, to established algorithms is a promising direction [127].

3.2. B5G systems in the Cloud-Edge Continuum

An interesting development happening in parallel with the development of B5G systems is the expansion of the so-called Cloud-Edge continuum [8,11,13]. Denoting a more continuous computing paradigm than traditional Cloud computing, this continuum involves the distribution of compute power across the network’s layers reaching out to its geographical edge [6]. It is enabled by the increase in compute performance of low-power devices and driven by a need for more localized handling of the growing (cellular) traffic amounts that double every two years [7] and are expected to reach 226 EB/month in 2026 [85] and as high as 5016 EB/month in 2030 [2], as well as the increasing number of users and devices exceeding 5.3 billion and 13.1 billion by 2023 [6]. The increased degree of distribution is interesting for a multitude of reasons: firstly, end devices retain the benefits of traditional offloading; secondly, users may experience lower latency communications as data or computations are provided at a shorter distance to them; and thirdly, data needs to travel over fewer links and through fewer shared network nodes, potentially decreasing contention and improving privacy [18,128,129].

A greater distribution of computations, particularly at the Edge, is expected to be at the core of B5G systems [2,33,34,126,128]. Once implemented, it will enable rapid network optimization and reconfiguration [90], new classes of low-latency or high-bandwidth services (e.g., Augmented Reality) [7,128] and end-to-end network slicing [33] (as described above) needed to handle the predicted massive connectivity communication demands [6,13,93]. Combined with AxC, this implies new interesting energy-latency-accuracy trade-offs achievable by applying approximations across network layers; that is, the data-producing devices, their communication links, and the node that carries out their processing whether it be Edge or Cloud-located. Analog to applying approximations across software and hardware in isolated systems, this requires tools for quality management, error modeling, and formal verification [18]. Additional design space exploration may also be needed to identify (near-)optimal combinations of AxC techniques for different applications with diverse quality and reliability constraints.

Some of the surveyed work already proposes algorithms specifically targeting this Cloud-Edge continuum, relying on it for performing heavy computations or exploiting its centralized nature to optimize network functionality. The former is needed for the AI-based scheduling in [4] and the ML-based channel estimation in [90]. The latter is more commonly represented and used for message clustering [85] and transmission coordination [105,120] as well as for exploiting spatial locality in accessed data [93]. Of these, only Bonati et al. [4] describe that Cloud-based network architectures may suffer from long latency if their involved data transfers are not minimized. Unfortunately, this latency is critical and may impact, among other factors, the practicality or user experience of, e.g., the aforementioned AR applications [7,130].

To illustrate this criticality, we reproduce simulation results of [131] for a number of different execution scenarios. The setting involves a wearable device and a smartphone, as well as an Edge server and the Cloud. The smartphone may execute tasks offloaded from the wearable itself or simply serve as a relay or gateway node to propagate these and return their results to and from the Edge or the Cloud. Fig. 8 shows the results thereof, for which the execution scenarios are enumerated as follows: (1) local execution on the wearable, (2) offloaded execution on the gateway (smartphone), (3) offloaded execution in the Cloud at a fixed distance of 50 m to the nearest BS, and (4a-d) offloaded execution to the Edge at distances of 50 m, 100 m, 200 m, and 400 m to the nearest BS. The simulation assumes an AR-like workload where the wearable is connected to the gateway via Wi-Fi, that tasks are atomic and indivisible, and that the received results are orders of magnitude smaller than the transmitted data rendering their effects negligible [132]. The results show the latency benefits of offloading regardless of the distance to the nearest BS, while there are energy benefits only when the BS is less then 200 m away.

Interestingly, and yet to be explored, these results may change if AxC is integrated into the system. Keeping our focus on application-level approximations, using AxC to reduce algorithmic complexity may render local execution beneficial energy-wise, while the latency will likely remain in favor of offloading. Approximating the wireless communication part instead will most likely merely enhance the already clear benefits of offloaded processing both in energy and latency. Advanced systems may integrate approximations in both processing and communication by, for example, using inexact pre-processing or lossy compression locally, deliberately neglecting errors during transmission, and approximately processing data at the Edge. Predicting the energy-latency relations for such systems is highly complex and requires further analysis.

![Fig. 7. Conceptual view of an IRS and its internal components.](image-url)
3.3. Traditional approximate computing in B5G

Another observation made through this survey is that only four of the 80 surveyed papers apply traditional AxC techniques: [28,37,38,115]; traditional in the sense that they are commonly reviewed or applied in other work [17,18,20,30]. Yet, as many of the algorithms involved in making B5G systems perform optimally can sustain approximations at the application level, we find it likely they can also benefit from approximations at the architecture or even the circuit level, also pointed out in [28]. While the arithmetic precision is undefined in most of the reviewed publications, their reliance on Integer Linear Programming (ILP) solvers, SCA algorithms, and various ML models hints at floating-point being used [49,84,89,91,133]. When these solvers’ convergence criteria permit such, one may consider switching to fixed-point formats within them before evaluating further approximations.

In other contexts, particularly related to so-called Network on Chips used to interconnect embedded devices (even wirelessly), approximations have been introduced both by voltage over-scaling links [134], truncating packets in network interfaces [135], or simply skipping the transmission of certain low-significance packets [136]. Such techniques could be adapted to a broader networking scope and find use in, e.g., low-significance IoT data collection systems or multimedia streaming. Related work already applies filtering and lossy data compression, albeit to reduce effects of network latency fluctuations [137]. A more generic application of such techniques is yet to be explored.

3.4. Open research directions

Despite the amount of work already carried out and published in the field, many research questions remain open. We summarize these in Table 5 and provide more details here. Most of the papers directly suggest extending their schemes in different ways. Such extensions may involve taking more metrics into account in the proposed algorithm [44,59,67,70,87,123] or focus on widening the scope of the proposals. Within the latter, we find extending DNN transceivers to UL traffic [38], using puncturing in coded streams [41], and adding support for massive MIMO [88], device-to-device communication [54], or simultaneous information and power transfer networks [105]. While such extensions make the algorithms relevant to more use cases, they often increase their overheads. Constraining or optimizing the algorithms thus becomes relevant again [59,62,81,87,90,106]. Once fully optimized, making algorithms work in practical environments entails a need for acceleration and distribution that needs addressing [55,90].

While extensions in the same direction are one way to optimization, changing algorithms is another. Another set of papers pursues this. Such proposals include using learning to reduce optimality search complexity [41], exploring alternative routing, bandwidth allocation, and content caching algorithms [61], and developing new algorithms for resource allocation in IAB [62]. Adapting already proposed algorithms for resource allocation [93] and IRS-based relaying [119] to distributed computing systems is another direction that needs more work. Finally, further attention should be paid to making systems aware of their spectral efficiency [42] and to optimize beamforming and fronthaul compression to maximize their energy efficiency [118].

Yet another set of publications propose future work in more general terms. Related to ML, such directions include low-complexity RL models and other algorithms for non-convex optimization [64,108], generation and publication of more open datasets [90], and techniques for tackling security and mobility issues in over-the-air FL [121]. As ML is expected to become an essential component in B5G systems, these challenges should be addressed as soon as possible [8]. Unrelated to this, two papers point out particular fields that need further work: RAN slicing and THz communications, which require further perfecting [111] or technological advances in precoding, beamforming, and tracking [73], respectively.

Several proposals require further evaluation. This need ranges from evaluating optimality [51]; over applying schemes to more dynamic or challenging use cases [106] with hypervisor migration [65] or different data distributions and realistic mobility scenarios [121]; to evaluating different metrics, particularly energy efficiency [74] and spectral efficiency and their trade-off [54]. Another proposes to evaluate their proposed scheme in a completely different setting (namely, one with multiple queues) to identify the proper balance between multiplexing gain (i.e., assigning several tasks to a slice) and overbooking (i.e., assigning too many tasks to one slice) in network slicing [108].

4. Conclusion

As the number of smart connected and intelligent devices increases, the cellular network is put under tremendous pressure demanding its development. 5G systems have been standardized for some years and are currently being deployed, but researchers already question whether they are sufficient for the next ten years. As a result, technologies for B5G systems are undergoing research. These next-generation networks are expected to enable entirely new classes of intelligent applications and support growing needs for data traffic. Optimizing the networks for various objectives comes at high computational costs, as the underlying algorithms tend to be complex. Minimizing these costs is essential for
Summarized directions of future work extracted from surveyed works.

<table>
<thead>
<tr>
<th>Topic</th>
<th>References</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extension</td>
<td>[38], [41]</td>
<td>DNN transceivers for UL traffic. Puncturing for coded streams.</td>
</tr>
<tr>
<td></td>
<td>[88]</td>
<td>Support for massive MIMO.</td>
</tr>
<tr>
<td></td>
<td>[54]</td>
<td>Support for device-to-device communication.</td>
</tr>
<tr>
<td></td>
<td>[105]</td>
<td>Support for simultaneous information and power transfer.</td>
</tr>
<tr>
<td>Design Space</td>
<td>[41]</td>
<td>Reduced optimality search complexity. Alternative routing, allocation, and caching algorithms.</td>
</tr>
<tr>
<td>Exploration</td>
<td>[61]</td>
<td>New IAB resource allocation algorithms. Distribution of IRS algorithms.</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>[64], [108]</td>
<td>Low-complexity RL and non-convex optimization algorithms.</td>
</tr>
<tr>
<td></td>
<td>[121]</td>
<td>Tackling security and mobility issues in over-the-air FL.</td>
</tr>
<tr>
<td></td>
<td>[90]</td>
<td>Publication of more open datasets.</td>
</tr>
<tr>
<td>Further Evaluation</td>
<td>[51], [65], [106], [108], [54], [74]</td>
<td>Assess optimality of proposals in other more (realistic) environments. Analyzing trade-offs between energy and spectral efficiency.</td>
</tr>
<tr>
<td>AxC in B5G and 6G</td>
<td>[18], [20]</td>
<td>Explore circuit-level AxC techniques applicability in wireless networking systems.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Adapt approximate NoC techniques to larger networks.</td>
</tr>
<tr>
<td></td>
<td>[138], [90], [139], [100], [122], [140]</td>
<td>Consider using neural approximation in place of more algorithms than, e.g., channel estimation, network slicing, IRSs, and network switches.</td>
</tr>
<tr>
<td></td>
<td>[131]</td>
<td>Identify benefits and drawbacks of AxC to Cloud-Edge task offloading scenarios.</td>
</tr>
</tbody>
</table>

Acronyms

3GPP Third Generation Partnership Project
ADMM Alternating Direction Method of Multipliers
AI Artificial Intelligence
AP Access Point
AR Augmented Reality
AxC Approximate Computing
BER Bit Error Rate
BLP Binary Linear Programming
BS Base Station
CNN Convolutional Neural Network
CSI Channel State Information
DNN Deep Neural Network
DRAM Dynamic RAM
DFT Discrete Fourier Transform
DL Downlink
DSP Digital Signal Processing
eMBB Enhanced Mobile Broadband
ER Error Rate
FER Frame Error Rate

FL Federated Learning
FPGA Field-Programmable Gate Array
GNN Graph Neural Network
GPU Graphics Processing Unit
GFDM General Frequency-Division Multiplexing
gNB next-Generation Node B
IAB Integrated Access and Backhaul
ILP Integer Linear Programming
IoT Internet of Things
IRS Intelligent Reflective Surface
ISAC Integrated Sensing and Communication
KPI Key Performance Indicator
LDPC Low-Density Parity-Check
LSTM Long Short-Term Memory
MLP Mixed Integer Linear Programming
MSE Mean SE
MA Multiple Access
MBNLP Mixed Binary Non-Linear Programming
MIMO Multiple-Input Multiple-Output
MINLP Mixed Integer Non-Linear Programming
ML Machine Learning
MLP Multi-Layer Perceptron
mMTC massive Machine Type Communication
mmWave millimeter Wave
NMSE Normalized Mean SE
NOMA Non-Orthogonal MA
NN Neural Network
NoC Network on Chip
OMA Orthogonal MA
OFDMA Orthogonal Frequency-Division Multiple Access
OFDM Orthogonal Frequency-Division Multiplexing
PU Processing Unit
QoE Quality of Experience
QoS Quality of Service
RNN Recurrent Neural Network
RAM Random Access Memory
RAN Radio Access Network
RB Resource Block
RF Radio Frequency
RIS Reconfigurable Intelligent Surface
RL Reinforcement Learning
SER Symbol Error Rate
SCA Successive Convex Approximation
SE Soft Error
SIMD Single Instruction Multiple Data

ensuring their practical deployment. One approach to do so for non-critical applications which require less than nine 9s reliability is to apply AxC for reduced algorithmic complexity, latency, and energy savings.

In this paper, we have surveyed publications on the state-of-the-art in the intersection of AxC and B5G systems, identifying and highlighting trends and tendencies in existing work and directions for future research. We have considered publications in two categories: existing and future technologies. Within the former, we find that resource allocation algorithms are particularly popular, while research related to IRSs appears the most prominent in the latter. In both, problems are often NP-hard and, thus, only solvable using heuristics or approximations, the proposed algorithms not being practical due to their computational demands, requiring further adaptation before making it into real-world systems. As the algorithms are error-resilient, one such adaptation may be to introduce approximations with AxC techniques. This direction may bring large performance gains but is largely unexplored till now.
Algorithmic approximations

We consider six publications presenting various algorithmic approximations. Proposed decoders are typically built around so-called variable nodes and check nodes that communicate probabilities of their current predicted bit values between one another. They are most frequently approximated in their update logic. Such an implementation is used by Li et al. [145] who propose a reduced-complexity minimum-sum way to update the check nodes, which halves the number of multipliers and removes the need for adders at the expense of some shifters. Despite these savings, their algorithm achieves the same bit error rate as its exact counterparts. Sun and Jiang [146] use a similar implementation and propose a hybrid approach combining minimum-sum and linear approximation update rules to the check nodes. The algorithm outperforms traditional normalized minimum-sum algorithms (like that of [145]) and performs comparably to conventional belief propagation at a significantly lower computational complexity.

Honghao et al. [147] propose another layered minimum-sum algorithm with reduced-complexity check node updates. They optimize correction factors, internal thresholds, and evaluation order of individual nodes to maximize decoding performance, and report comparable performance to belief propagation with less than half as many iterations. Wu et al. [148] present a more involved algorithm that simultaneously applies minimum-sum check node updates with NN-based internal coefficient adjustments. By passing the check node values through the network, some correction factors are adapted to current channel conditions. This enables their algorithm to outperform belief propagation under certain conditions while they do not declare the computational overheads it entails.

Sun et al. [149] focus instead on so-called Bose-Chaudhuri-Hocquenghem-constrained codes for URLLC communications, which improves error correction performance at the cost of a lower data rate. They consider short block-length codes and apply approximated belief propagation with reduced-complexity maximum-logarithm check node updates. Their algorithm outperforms the error correction capabilities of the polar codes used in 5G, albeit at greater computational complexity.

Watanabe et al. [150] consider the bit reliability-based decoding algorithm, which, much like the other layered algorithms, is split between check and variable nodes but differs in what they pass between each other. They propose a “deep learning”-like approach in which the edges between nodes are weighted, and the overall decoder resembles a DNN. Training the weights for different noise ranges on the input signals presents an interesting trade-off between generality and error correction performance under specific conditions. Despite reduced complexity compared with minimum-sum algorithms, the reported performance is comparable.

Hardware approximations

Another five works present various hardware decoder architectures. Zhou et al. [151] describe a layered decoder design. Rather than approximating the algorithm, as in [145–147], they employ inexact truncated arithmetic to implement the variable update logic, reducing circuit area and delay. They demonstrate their design for codes used in various short-range wireless networks, showing comparable performance with exact fixed-point designs.

Tsatsaragkos and Paliouras [152] describe a decoder design based on reduced-complexity minimum-sum algorithms. They implement parallel variable node processors, a centralized check node unit, a concatenation unit to generate the final result, and a control unit to manage the decoding. In addition, they propose several logarithmic barrel shifters for interconnecting the variable node processors with their respective check nodes. The design is evaluated on Wi-Fi-related codes and shows improved throughput at a reduced circuit area. Patel and Engineer [153] describe a vastly simpler design that operates

---

1. Inputting (approximat* OR inexact OR inaccurate OR ‘good enough’?) AND (ldpc OR ‘low-density parity-check’) to Scopus and limiting results to computer science returns 752 results of which many appear to be relevant.
as an accelerator in a FPGA. They also implement a minimum-sum algorithm whose approximations are limited to quantized fixed-point arithmetic, achieving similar performance to a full-precision software implementation.

Lopez et al. [154] also present an architecture implementing an approximate minimum-sum algorithm. Specifically, instead of computing two minima in each variable node, they compute one and estimate proximate minimum-sum algorithm. Specifically, instead of computing approximations limited to quantized fixed-point formats, apply suitable for quantum key-based cryptography schemes. Like [153], they quantize internal values to relatively large fixed-point formats, apply

Yang et al. [155] propose another, more involved than [153], FPGA-based decoder design. They particularly consider very long codes suitable for quantum key-based cryptography schemes. Like [153], they quantize internal values to relatively large fixed-point formats, apply second-order approximation to their otherwise logarithmic message-passing function, and optimize their design for the resources available on a FPGA. These design choices allow their design to function at a much lower SNR than related works.

References

H.J. Damgaard et al.

Computer Networks 233 (2023) 109872

[67] B. Bakhshi, J. Mangues-Bafalluy, J. Baranda, R-Learning-based Admission Con-

Y.Z. Bekele, Y.-J. Choi, Random Access Using Deep Reinforcement Learning in

D. Liu, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

B. Liu, Z. Zhou, Y. Yang, A.R. Adhikary, P. Fan, Constructions of Binary Signa-

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint

H. Dai, C. Zhang, J. Luo, C. Li, B. Wang, QoE-Driven Resource Allocation for


Y. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

D. Harris, D. Raz, P. Sagiv, Bandwidth Resource Allocation in Integrated Access

Q. Zhang, W. Ma, Z. Feng, Z. Han, Backhaul-Capacity-aware Interference

F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Traffic Steering

S. Abdel-Razeq, H. Al-Obiedollah, H. Bany Salameh, Efficient User-Channel

W. Li, Y. Zhang, B. Bian, H. Yang, L. Li, Q. Wang, H. Zhang, J.-B. Wang, Joint
Hans Jakob Damsgaard received the B.Sc. degree in electrical engineering in 2019 and the M.Sc. degree in computer science and engineering in 2021 as part of the Honours programme at the Technical University of Denmark, DTU. He is currently pursuing a Ph.D. in reconﬁgurable approximate accelerators for secure edge computing at Tampere University, TAU. He received the best paper award at NorCAS’21. His current research interests include hardware accelerators, networks-on-chip, computer architecture, and hardware veriﬁcation.

Md Munjura Mowla works as an Expert R&D Engineer in IS-Wireless, Poland. He received a Ph.D. degree in Communication Engineering from Edith Cowan University, Australia. Dr. Mowla is experienced in developing and working on EU H2020 projects. He is author (and co-author) of more than 50 papers from different topics of wireless communications. His current research interests include beyond 5G and 6G Wireless Communications, Radio Resource Management, Software Deﬁned radio RAN, Edge Computing, AI/ML algorithms, cell free massive MIMO, etc.

Adam Filizkowski holds the position of Head of R&D at the ISRD Sp. z o.o., has 22 years of commercial experience in cellular systems (incl. 5G, LTE, WiMAX) and especially topics of RRM techniques (admission, congestion, NOMA). He is leading the research team, he also manages the ECSEL JU project called BRAINE (AI/ML supported micro edge DC) dealing with workload prediction and placement. He holds an M.Sc. from telecommunications and is pursuing Ph.D. on admission control in future wireless systems. He is author (and co-author) of more than 70 papers on different wireless communications, cyber security and adaptive multimedia system topics.

Aleksandr Ometov received D.Sc.(Tech.) and M.Sx. degrees from the Tampere University of Technology, 2018 and 2016. He also holds a Specialist degree from Saint Petersburg State University of Aerospace Instrumentation, 2013. He is a Senior Research Fellow at Tampere University, Finland, and the coordinator of the CONVERGENCE of Humans and Machines research field. He is a Project and Training Manager of EU H2020 MCSA A-WEAR and APROPOS JTJ projects. His research interests include wireless communications, information security, computing paradigms, blockchain technology, and wearable applications. He was recognized as a Publisher of the Year in 2022 by TAU.

Jari Nurmi received the D.Sc.(Tech.) degree in 1994. He works as a Professor at Tampere University, TAU (formerly Tampere University of Technology, TUT), Finland, since 1999. His research interests include embedded computing systems, System-on-Chip, approximate computing, wireless localization, positioning receivers, and software-deﬁned radio and -networks. He holds various research, education, and management positions at TUT since 1987 and was the Vice President of the SME VLSI Solution Oy from 1995 to 1998. He is the Coordinator of the European doctoral training network APROPOS.