Gerhard P. Fettweis, Meik Dörpinghaus, Jeronimo Castrillon, Akash Kumar, Christel Baier, Karlheinz Bock, Frank Ellinger, Andreas Fery, Frank H. P. Fitzek, Hermann Härtig, Kambiz Jamshidi, Thomas Kissinger, Wolfgang Lehner, Michael Mertig, Wolfgang E. Nagel, Giang T. Nguyen, Dirk Plettemeier, Michael Schröter, Thorsten Strufe

Architecture and Advanced Electronics Pathways Toward Highly Adaptive Energy- Efficient Computing

Erstveröffentlichung in / First published in:

Proceedings of the IEEE. 2019, 1 (107), S. 204-231 [Zugriff am: 13.01.2023]. IEEE. ISSN 1558-2256.

DOI: http://dx.doi.org/10.1109/JPROC.2018.2874895

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

Architecture and Advanced Electronics Pathways Toward Highly Adaptive Energy-Efficient Computing

This paper describes a leading European effort on applications of basic technologies to energy-efficient servers and high-performance computing of the future, that has been ongoing for more than a decade.

By Gerhard P. Fettweis, *Fellow IEEE*, Meik Dörpinghaus, *Member IEEE*, Jeronimo Castrillon, Akash Kumar, *Senior Member IEEE*, Christel Baier, Karlheinz Bock, Frank Ellinger, *Senior Member IEEE*, Andreas Fery, Frank H. P. Fitzek, Hermann Härtig, Kambiz Jamshidi, *Member IEEE*, Thomas Kissinger, Wolfgang Lehner, Michael Mertig, Wolfgang E. Nagel, Giang T. Nguyen, Dirk Plettemeier, Michael Schröter, *Senior Member IEEE*, and Thorsten Strufe

ABSTRACT | With the explosion of the number of compute nodes, the bottleneck of future computing systems lies in the network architecture connecting the nodes. Addressing the bottleneck requires replacing current backplane based network topologies. We propose to revolutionize computing electronics by realizing embedded optical waveguides for onboard networking and wireless chip to chip links at 200 GHz carrier frequency connecting neighboring boards in a rack. The control of novel rate adaptive optical and mm wave transceivers needs tight interlinking with the system software for runtime resource management.

Manuscript received February 21, 2018; revised August 27, 2018; accepted September 28, 2018. Date of publication December 6, 2018; date of current version December 21, 2018. This work was supported by the German Research Foundation (DFG) within the Cluster of Excellence EXC 1056 Center for Advancing Electronics Dresden (cfaed) and within the CRC 912 Highly Adaptive Energy-Efficient Computing (HAEC). *(Corresponding author: Meik Dörpinghaus.)*
The authors are with the Collaborative Research Center HAEC, Center for Advancing Electronics Dresden, Technische Universität Dresden, Dresden, Germany (e-mail: gerhard.fettweis@tu-dresden.de; meik.doerpinghaus@tu-dresden.de; jeronimo.castrillon@tu-dresden.de; akash.kumar@tu-dresden.de; christel.baier@tu-dresden.de; karlheinz.bock@tu-dresden.de; frank.ellinger@tu-dresden.de; fery@ipfdd.de; frank.fitzek@tu-dresden.de; hermann.haertig@tu-dresden.de; kambiz.jamshidi@tu-dresden.de; thomas.kissinger@tu-dresden.de; wolfgang.lehner@tu-dresden.de; michael.mertig@tu-dresden.de; wolfgang.nagel@tu-dresden.de; giang.nguyen@tu-dresden.de; dirk.plettemeier@tu-dresden.de; michael.schroeter@tu-dresden.de; thorsten.strufe@tu-dresden.de).

Digital Object Identifier 10.1109/JPROC.2018.2874895

KEYWORDS | Adaptable software stack; adaptivity; chip stack; computing; energy efficiency; interconnect; optical communi cation; wireless communication

I. INTRODUCTION

As we see further advances in semiconductor scaling slowing down significantly, and localized on-silicon heat generation due to power consumption being a major challenge, we must find new answers to advance computing. We have been able to observe a shift from clock rate increases to parallel processing. However, the single processor model of a compute node interfacing with the memory subsystem via a hierarchy of caches does not scale well to massively parallel machines. In-memory computing might be one way forward [36], with processing nodes being distributed within the memory. They are then only activated if code needs to be executed that addresses the memory segment near the node.

This, however, creates three major challenges. First, compute cores now must be designed for an efficient code execution relative to their activation. When moving toward millions of cores, a core's activity will be far less than 100%, which alone is a huge paradigm change in computing. Second, this entails that the core design itself must be "lean and mean," as its silicon footprint weighted with

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

the inverse of the activation rate must be small when compared to its local memory segment. This keeps the memory/computing balance in place. Third, the requirement of accessing data across memory segments is not eliminated by in-memory computing. Hence, the bottleneck of computing moves from designing cores with increasing instruction level parallelism (ILP) to a networking architecture that can carry the enormous communication load.

It is clear that it is of utmost importance to increase the bandwidth of each communication channel. However, the data rate's peak-to-average spread will increase in future. When measuring InfiniBand-activity within large machines of today, we can see a typical 1% activation over time to carry the required traffic. We will require circuits to turn on efficiently to then carry full bandwidth even if activity rates might decrease below 1%. Adaptivity, e.g., in terms of adaptive voltage and frequency scaling, is present in all current multicore chips to be able to control power consumption. However, future in-memory computing systems need to carry this adaptivity to a new level. This must include processors, memory segments, and all components of the networking architecture. Adaptivity support of a system will then benefit energy efficiency, which in turn allows for much denser packaging if systems must not deliver continuous 100% activity load.

From a networking side, the energy consumption is determined not only by the activity, but in particular, by the distance and number of hops data must travel. This leads to the next challenge, designing the networking architecture and the runtime resource management stack to control the system. With upcoming 3-D stacking of chips, networking can be broken down into multiple hierarchical challenges: 1) on-chip networking; 2) intranetworking within a 3-D chip stack; and 3) networking in a rack between the 3-D chip stacks. We propose here novel on-chip networking strategies allowing for adaptive rerouting to address 1). For new ideas on 2), we refer to [33].

Our largest impact on networking comes from our proposal addressing 3). Optical communications is very efficient when implemented with waveguides, whereas wireless communications allows for realizing highly meshed networks without complex wiring. Therefore, we propose to use optical waveguides for onboard internetworking of 3-D chip stacks (onboard communication), and wireless networking of chip stacks between boards in a rack (interboard communication). In principle, the latter eliminates the need for communication via backplanes within racks. Envisioning on-chip antenna arrays at the top level of a 3-D chip stack, and optical transceivers at the bottom, a hybrid 3-D stack with memory, computing, and communication interfaces is possible. This we name HAEC—Highly Adaptive Energy-Efficient Computing. Fig. 1 visualizes this concept in a small implementation, which we call a HAEC Box.

Obviously, a sophisticated runtime and networking management system is needed to fully embrace this kind of networking architecture. Load management of cores, placement of memory content, and network activity all depend on the runtime system, which carefully needs to incorporate the monitoring of all states.

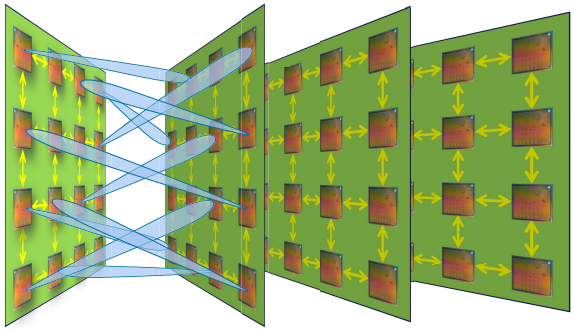

Fig. 1. *HAEC backplane with optical onboard and wireless interboard links. It shows a simple configuration into one "box," which we name a HAEC Box.*

In this paper, we describe the joint application of the aforementioned concepts resulting in new paths to overcome the energy bottleneck of today and deliver orders of magnitude increase in low-energy computing for massively parallel machines. Not only do we show concepts, but fabricated and evaluated chip designs underline the validity of the approach. In Section II, we give an overview of the HAEC Box. Section III shows novel high bit-rate adaptive transceivers to enable efficient onboard optical communications. The wireless links are addressed in Section IV, presenting adaptive transceivers for 100 Gb/s and beyond. In Section V, we dive into the design of the processing within a 3-D chip stack, here named a "node" within the HAEC concept. Section VI describes the network architecture and a simulation framework as well as the "HAEC playground" as a hardware prototyping testbed to evaluate the impact of different networking protocols and network topologies on system performance. System software and the runtime adaptivity mechanisms toward the envisioned HAEC Box are described in Section VII. In Section VIII, we cover new ideas toward increasing the bandwidth of the communications by orders of magnitude. New DNA-origami template optical waveguides could increase the density by at least two orders of magnitude, and wireless communications reaching into the terahertz band could enable beyond 1 Tb/s. Finally, Section IX gives a quick energy analysis, showing that a HAEC Box could be feasible, enabling Pb/s backplane capacity of a $4 \times 4 \times 4$ node implementation fitting into a power budget of about 1 kW.

II. HAEC BOX: OVERVIEW

The HAEC Box is a realization of a highly adaptive energy-efficient computing architecture following the vision described above. While energy efficiency has already been considered at certain levels, e.g., both at hardware and software levels, what is needed is a comprehensive approach that takes the overall applications' demand and users' context as well as the particular aspects of computing resources (software and hardware) into account.

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

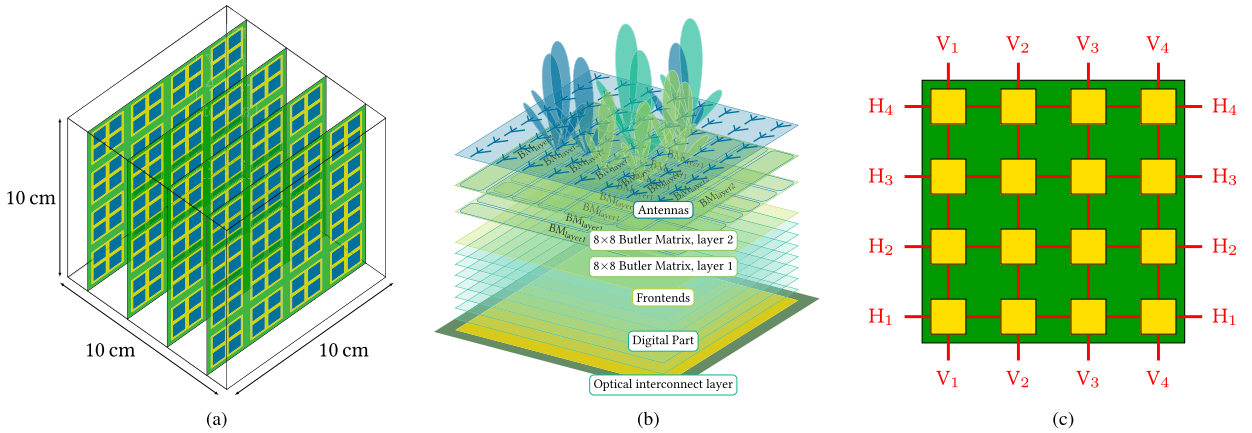

Fig. 2. *Sketch of the HAEC Box. Optical onboard communication and wireless interboard communication between computing nodes. (a) Four PCBs with 16 chip stacks (computing nodes) per side, each carrying four antenna arrays. (b) Chip stack with an antenna array on the top. (c) PCB with the optical network. Waveguides at the left/top and right/bottom are connected according to the labels* $H_1, \ldots, H_4, V_1, \ldots, V_4$.

Thus, the goal of our HAEC Box approach is to address energy efficiency at all levels, not just isolated but as a joint optimization problem, ranging from transistor level, over the communication architecture up to the software engineering with an energy-aware computing management. The guiding principles are simple and uniform:

1) ensure flexibility at all levels, providing alternatives tailored to applications and situations;
2) establish, mostly offline, the added value (utility) of spending energy in one or the other alternative;
3) monitor current energy levels and other activities;
4) take action at runtime based on the monitoring results and the available alternatives.

In this section, we provide an overview of the HAEC Box hardware specifications, including computing and interconnect with a strong focus on the latter in Section II-B. Furthermore, we discuss power constraints, which are highly relevant given the aggressive integration we foresee for the HAEC Box. This section closes with a discussion of how the flexibility provided by the underlying hardware is to be exploited by a future software programming environment. These programming considerations led to the prototype software architecture described in Section VII.

A. Computing Specifications

For the HAEC Box we assume a volume of one liter; see Fig. 2(a). Assume that with future technologies each chip will carry 64 000 elements of which 10% are processors and 90% are memory and we stack 128 of these chips in a 3-D stack. A computing node consists of one of those chip stacks. Moreover, we assume that we have 16 of those computing nodes on each side of a printed circuit board (PCB) and we have four PCBs within the HAEC Box.

The assumption that 90% of the chips' area carries memory is aligned with a trend observed in the past two decades where more and more transistor budget and die area are allocated for on-die memory (already reaching approximately 50%). The trend in the memory/computing ratio is caused by the fact that increasing cache size has proven more energy efficient than increasing processor counts to improve computing performance [14]. Furthermore, from a performance and energy perspective it is advantageous to arrange the memory close to the processors and make it an integral part of a chip stack. The distribution of compute and memory elements will yield extreme nonuniform memory access (NUMA) effects, i.e., processors on the same layer share a common memory, whereas those on different layers or compute nodes potentially communicate via a large variety of routes.

With 10% of chips carrying processors, the HAEC Box will feature up to 10^8 processors, which is 10^4 times the computing performance per unit volume of today's servers. The individual compute nodes will be connected by an innovative adaptive and highly energy-efficient communication architecture consisting of optical links connecting compute nodes on the same PCB and wireless links connecting compute nodes on adjacent PCBs.

B. Energy-Adaptive Interconnect

In today's servers, the communication architecture between individual computing nodes is typically based on electrical or optical wired links both for onboard as well as for interboard communication. However, with the increasing amount of processors in future servers the performance requirements on the communication architecture will significantly increase, in terms of the required data rates as well as in terms of delay and latency requirements. This holds especially true if the processors should not only be used in parallel for independent computing tasks, but to solve more complex computing problems jointly. Today the backplane capacity in high-end server systems is in the range of 1–2 Tb/s. Extrapolating the almost exponential increase of the data rates on the wired backplane into the future, within two decades data rates in the order of one petabit-per-second *(petabit backplane)* are required. Moreover, to enable energy-efficient computing, the communication architecture needs to be flexible and rate adaptive to satisfy the communication requirements

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

between the computing nodes with high energy efficiency. This means that the power consumption of the communication links should be proportional to the momentarily required data rate.

State-of-the-art connectivity structures based on a wired backplane will not be able to satisfy these requirements, as today's copper-based connections have reached fundamental physical limits. Thus, high-speed energy-adaptive computing demands a radically new hardware connectivity architecture between the computing nodes. To enable an aggregated bandwidth in the petabit-per-second regime both, the data rate per link and the number of parallel links have to be increased. Simultaneously, the energy consumption per transmitted bit needs to be minimized which represents a huge challenge.

Today's petascale supercomputers already make substantial use of optical interconnects for rack-to-rack communication due to their higher data rates and higher energy efficiency in comparison to electronic interconnects [76, Sec. 4.1]. In future it is expected that this trend will also apply to shorter communication links such as onboard and even processor-to-memory links [64]. Comparing electrical and optical transmission, the available bandwidth–length–product of optical waveguides is much higher than that of electrical copper-based interconnects—several terahertz meter (THzm) versus a few tens of gigahertz meter. The lower attenuation explains the advantages of optical systems especially for longer links. The difference between the bandwidth–length–product of electrical and optical links would be more prominent at higher electrical carrier frequencies at which the electrical attenuation is higher. Due to the higher loss of electrical waveguides, for longer distances multiple hops become more energy efficient than a single hop. However, that adds power consumption per hop. In contrast, optical onboard links are presently limited by the optoelectronic conversion, while waveguide losses are not the limiting factor for distances we consider. Their bottleneck in terms of energy efficiency is the electro–optical conversion or the optical modulation. In [121], it has been shown that already for data rates above 10 Gb/s the modulation stage dominates the energy consumption per bit in the optical communication, such that the break-even length between optical and electrical interconnects drops below 1 mm as soon as the energy efficiency of the modulation stage is better than 1 pJ/b (picojoule per bit), which is realistic [64], [86]. With the optical link we developed, we already achieved an energy consumption per bit of 4.5 pJ/b at 25 Gb/s (see Section III-B), which is lower than 7.3 pJ/b at 20 Gb/s reported by IBM [28]. In comparison, for electrical links, the energy efficiency strongly depends on the length of the link and the data rate. For example, for a 2-m 20-Gb/s copper link in [21], an energy consumption of 15 pJ/b has been reported, while for a 0.3-m electrical Cu-based interconnect (Altera Stratix V) with 28 Gb/s, an energy consumption of 7 pJ/b has been measured [2], [12]. A detailed overview of the energy efficiencies and data rates of different short distance interconnect technologies is given in [121, Table 3]. Moreover, optical links have the advantage of a much higher bandwidth density and less crosstalk in comparison to electrical links, even with current technology. At the PCB level in [76, Sec. 4.1], bandwidth densities of up to 1 Tb/s/mm of horizontal cross section are displayed for optical links, while for electrical links bandwidth densities of approximately 300 Gb/s/mm are given. Hence, for the connection of computing nodes located on the same circuit board with data rates in the range of 100 Gb/s per link up to 1 Tb/s, we will have to exploit the principal advantages of optical links in terms of high energy efficiency and higher bandwidth density. For comparison, with InfiniBand used today for interconnects in high-performance computing 50 Gb/s per link are achieved today and 100 Gb/s per link are predicted for 2020 where up to 12 links are typically aggregated [61]. Besides the high energy efficiency of the optical links we developed, one key novelty is that our optical links are adaptive in data rate and energy consumption, such that they can be controlled with respect to the actual software requirements. This adaptivity is of major importance to enable highly energy-efficient computing. Due to the high bandwidth density that optical interconnects can provide, we chose them to connect computing nodes located on the same boards of the HAEC Box; see Fig. 2(c). These optical links are based on waveguides being integrated into the printed circuit board. We propose using a 2-D array of optical input/output (I/O) ports at the bottom of each computing node where the light is coupled into the waveguides within the PCB with 90° microcoupling optics. Based on this technology, we will be able to connect 120 waveguides to each computing node yielding 60 bidirectional links such that we are, for instance, able to connect a computing node with its four nearest neighbors using 15 bidirectional links in parallel to increase the available data rates.

For the communication between computing nodes being located on adjacent circuit boards we chose wireless links, as they are highly favorable due to their inherent flexibility, while still showing an acceptable energy efficiency; see Fig. 2(a). Based on our results, we can roughly estimate an energy consumption per bit of the wireless links for interboard communication of 48 pJ/b for a data rate of 100 Gb/s with some further significant reduction expected; see Section IV-G. The energy consumption of the wireless links is naturally larger than what can be achieved with optical links. Nevertheless, it has to be considered that the wireless network is much more flexible in comparison to optical links, as it allows to directly communicate from one computing node to any and all computing nodes on the neighboring board by using a Butler-matrix antenna feeding network for beam switching. Based on optical links, this would require communication over multiple hops [on the average 7 for the setup shown in Fig. 2(a)] including a much more complex routing yielding additional commu-

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

nication delays. Moreover, communication over multiple hops would also reduce the energy advantage of the optical links over the wireless links. The wireless links allow us to establish a fully connected link topology *(fully connected crossbar)* between computing nodes on neighboring boards.

Overall, the hybrid link technology based on optical and wireless links enables a novel server architecture. It has the potential to provide an excellent hardware infrastructure for energy-adaptive software and networks. Moreover, energy efficiency will already be increased at the link level by providing dynamic adaptivity of data rate and power consumption for the optical and for the wireless links as well as by the possibility to completely turn off links.

C. Power Considerations

Given the high density of processors in the HAEC Box, we expect strong *dark silicon* effects [34], i.e., a significant amount of on-chip resources cannot be operated at full performance at the same time for a long period of time. Different numbers are reported in the literature due to differences in methodologies but the percentage of inactive cores lies between 46% and 80% [56], [112]. In the HAEC Box, given the high computing density per volume, the percentage of inactive cores can even reach 90% or more. Therefore, we assume that only a small percentage $p \leq 10\%$ of the available cores of the HAEC Box are active at any time.

The power consumption of the HAEC Box consists of the power consumed by compute nodes and the power for optical and wireless transmission. To show that the overall HAEC Box architecture is reasonable from the energy and cooling standpoints, we provide a brief estimation of the power used for computation. To estimate the total power of the HAEC Box for computation, we first assume a power of 150 μW per core. In this regard, consider a chip with Argonaut RISC Core (ARC) technology [38] which has already been available as product since 2014. The chip is equipped with 32-bit RISC processors using 28-nm CMOS technology. Operating at 0.9 V, the core produces 0.036 mW/MHz. That number is equivalent to 367 μW when operating at 500 MHz and 0.3 V. When taking into account a technological advancement and a change in architecture, a core can reach a power of 150 μW using 14-nm technology with fully depleted silicon-on-insulator (FD-SOI). As a rule of thumb, we assume that the power of an inactive core is 0.1% of an active core's power [10]. For 10^8 processors in the HAEC Box, this yields an estimated power consumption of 765 W with an activity level of 5% ($p = 5\%$). A detailed estimation for the energy of the interconnect is provided in Sections III-B and IV-G.

D. Programming Considerations

The hardware specifications above, particularly of the interconnect, offer a wide range of configuration options to adapt the energy consumption of the HAEC Box. Examples are: 1) aggressive voltage and frequency scaling of the computation; 2) intelligent allocation of data to memory chip stacks; and more importantly 3) adapting the interconnect to the application needs. The latter includes, among others, modifying the power allocation to antennas, selecting between optical and wireless communication, or configuring a virtual network topology by setting up wireless links. This new kind of flexibility provided by the HAEC Box has to be visible to and be addressed by a novel software programming stack, supporting development, deployment, and runtime adaptation. With this in mind, we started developing a software environment using existing parallel systems and simulations. The testing, simulation, and programming environments are discussed in Sections VI-C, VI-B, and VII, respectively.

III. ENERGY-ADAPTIVE OPTICAL LINKS FOR ONBOARD COMMUNICATION

As described in Section II-B, each of the 16 compute nodes in the 4×4 matrix on one board side is connected to each of its neighboring nodes by 15 bidirectional optical links with lengths of up to 20 cm; see Fig. 2(c). The links are based on direct optical modulation using off-the-shelf vertical-cavity surface-emitting lasers (VCSELs). A laser diode driver (LDD) is designed to drive the VCSEL with the modulation signal. The optical data signal is transmitted via novel customized low-loss optical onboard waveguides and efficient coupling structures; see Fig. 3. At the receiver side the optical signal is detected by a standard photodetector (PD) and converted back to the electrical domain. A transimpedance and main amplifier (TIA/MA) further converts the photocurrent to a voltage and amplifies the electrical signal to digital voltage levels for further processing. Each link is designed to enable an energy-efficient optical data transmission with less than 7 pJ/b at a link data rate of up to 25 Gb/s. Moreover, a novel adaptivity approach is applied to the optical interconnects. In order to reconfigure the link performance according to the actual link requirements, the bandwidth is scaled down reducing the power consumption of the link circuitry [57]. This is realized by changing the operating point of the circuits either by tuning their bias voltages or their bias currents. When low link data rates are sufficient, the supply current can be reduced which in turn reduces the power consumption of the components. However, at the same time, the gain of the amplifiers drops which has to be compensated, e.g., by tuning the amplifier loads in order to maintain the signal levels at the circuit interfaces. By such a performance and power adaptivity, link energy savings of more than 50% can be achieved.

A. Waveguides and Coupling Optics

The integration of parallel optical interconnects on board level including the active and passive optical components as well as electrical integrated circuitry is a major focus of electronics packaging for network nodes of the future [119]. For the HAEC Box interchip links based

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

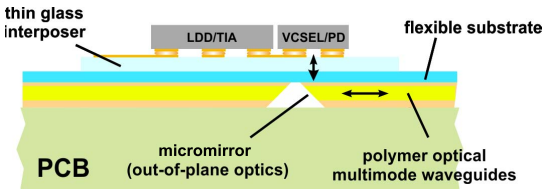

Fig. 3. *VCSEL-based optical transceiver subassembly with integrated polymer overlay waveguides.*

on planar polymeric optical multimode waveguides with integrated out-of-plane coupling optics in combination with an optical transceiver subassembly based on a glass interposer are very promising. The integration of polymeric waveguides on flexible substrates aims to realize an overlay optical substrate for enhanced yield and testability of the HAEC Box but also commonly of hybrid electro–optical printed circuit boards. The realized onboard waveguides feature low insertion loss with a maximum attenuation coefficient of below 0.1 dB/cm. For the considered planar waveguides error free transmission with bit error rates (BERs) below 10^{-10} to 10^{-12} was measured up to 30 Gb/s [91] for waveguides with lengths up to 9 cm. A transparent substrate-based interposer for the wavelength range from approximately 800 nm up to 1600 nm [ultraviolet (UV) to near infrared (IR)] has been developed using thin glass as material. As depicted in Fig. 3, the glass interposer includes the integrated waveguides, the micromirrors, and the metal interconnects for the assembly of the chip components. A transceiver package with optoelectronic converters (VCSEL and PD) and electronic transceiver integrated circuits (ICs) (LDD and TIA) has been manufactured involving very accurate placing ($\pm 0.5\mu$m) tools to complement the high-performance module manufacturing [77], [92].

B. Driver Circuits and Energy Estimation

Several adaptive LDD and TIA/MA ICs have been designed in a 130-nm silicon–germanium (SiGe) BiCMOS technology with bipolar transistor performance of up to 300-GHz transient frequencies. An adaptive high-speed VCSEL driver was designed for non-return-to-zero (NRZ) modulation and adjustable energy efficiency [106]. The LDD consists of two main blocks: a preamplifier and a driving amplifier. The preamplifier mainly provides a broadband matching to the 50Ω measurement environment and a decoupling of the driving block from the input signal in order to ensure a high bandwidth. The driving block directly biases the VCSEL and feeds the amplified modulation signal to the laser. The LDD is assembled to an 850-nm VCSEL with 20-GHz bandwidth. The driver is optimized either for constant extinction ratio or for high energy efficiency at different data rates. An error-free operation with a BER of less than 10^{-13} was achieved up to 45 Gb/s without incorporating any signal equalization or preemphasis. For this, a power consumption of only 81 mW, corresponding to the small energy per bit of only 1.8 pJ/b, is required. By applying performance and power adaptivity to the IC the highest energy efficiency can be achieved at 30 Gb/s and amounts to 1.17 pJ/b. This is realized by adaptively adjusting the LDD's modulation levels at runtime using voltage controlled reference current sources. In total, an LDD power consumption reduction up to 80% is achieved. At a data rate of 10 Gb/s the driver consumes only 16.5 mW of power. Furthermore, since no area consuming peaking inductors are used for bandwidth enhancement in the design, the driver is very compact to implement it into the onboard optical links.

For the receive part of the adaptive optical onboard interconnect, an adaptive TIA with continuous bandwidth and power consumption tuning has been designed in the 130-nm SiGe BiCMOS technology as well [107]. The TIA consists of three main parts: a TIA input stage with variable gain for the current-to-voltage conversion, a MA for the signal amplification, and the output driver which is only used for measurement purposes to match the impedance of the equipment. The bandwidth of the MA can be tuned continuously during runtime between 61 and 13 GHz by scaling the bias current. This leads to a reduction of the MA's power consumption by a factor of around 5. However, this reduction comes at the expense of a drop in the MA gain which has to be compensated. By varying the gain of the TIA input stage utilizing field effect transistors as steerable feedback resistors, the overall gain can be kept constant at 69.8 dBΩ. The maximum bandwidth of the IC enables data rates up to 88 Gb/s yielding a very high gain bandwidth product of 189.8 kΩ GHz. At this maximum speed the power consumption of the TIA core (without the output driver) is only around 30 mW corresponding to 0.34 pJ/b. An error-free operation with a BER of 10^{-13} was measured up to 50 Gb/s. Reducing the TIA performance to 20 Gb/s the power consumption decreases to 18 mW corresponding to 0.9 pJ/b. Although the energy efficiency becomes worse, the overall power consumption reduction of the TIA core results in 40%. Furthermore, only a single stage of the MA was adaptively tuned in this design. It is expected that the power consumption decrease can be raised to more than 50% if the other amplifier stages are implemented with the bias current scaling as well.

Finally, based on the measured energy consumption of the LDD and the TIA/MA we estimate the overall energy consumption of the optical link, including further components like VCSELs and clock data recovery, to be less than 4.5 pJ/b at 25 Gb/s, being significantly less than the 7.3 pJ/b at 20 Gb/s reported in [28].

C. Wavelength Division Multiplexing Approach

The VCSEL-based approach already provides a data rate of 25 Gb/s per waveguide summing up to an available data rate of 375 Gb/s between two compute nodes which are connected by 15 parallel waveguides. To even further

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

increase the available data rates, we additionally consider wavelength division multiplexing. For this purpose, we develop complementary metal–oxide–semiconductor (CMOS) compatible silicon modulators for modulating the light provided by off-chip lasers. Several wavelengths will be used to carry the information on different carriers. To satisfy future bandwidth requirements of the HAEC Box, 40 channels, each carrying 25 Gb/s, are considered [20]. Reverse-biased silicon modulators will be used to modulate the continuous wave light from off-chip lasers. Optimized modulators have been designed taking into account the tradeoffs coming from the link budget (the limits on the attenuation of each modulator) and the link energy requirements of the HAEC Box, which limits the energy consumption of a single modulator to the sub-100 femto Joules regime [60]. Therefore, resonant (either ring or slow-light) modulators need to be used for this purpose [59].

IV. WIRELESS LINKS FOR INTERBOARD COMMUNICATION

The aim of the wireless communication architecture is to provide high data rate, flexible, and energy-efficient communication between the computing nodes on adjacent boards within the HAEC Box. To enable high data rates of up to 100 Gb/s, we consider carrier frequencies around 200 GHz and a transmission bandwidth of 30 GHz. The antennas of the wireless links will be placed on top of the chip stacks; see Fig. 2(b). We consider 8×8 antenna arrays in combination with Butler matrix feeding networks to enable beam switching to the different computing nodes on the adjacent circuit board. The power consumption of standard resolution analog-to-digital conversion becomes very high at the sampling rates required for the high bandwidth of 30 GHz. We thus consider 1-bit quantization with temporal oversampling in combination with runlength encoding as a new energy-efficient approach in the area of wireless communication. Finally, as we assume that several computing nodes share common cache and memory and in order to be competitive with state-of-the-art wired interconnects such as InfiniBand, we consider a latency constraint of 100 ns. The main contribution to the link latency is due to channel coding such that this latency requirement leads to a very challenging channel coding design. In the following, we will discuss the individual components of the wireless link in more detail.

A. Analog Frontends

A complete 190-GHz transceiver (TRX) chipset for short-distance high-data-rate wireless links was implemented in the 130-nm BiCMOS SG13G2 technology of IHP [42]. In order to enable the highest possible data rates, the TRX targets the largest possible RF bandwidth at the fixed local-oscillator (LO) frequency of 190 GHz, with all the circuit blocks optimized for a large bandwidth. A further important design constraint for on-chip solutions

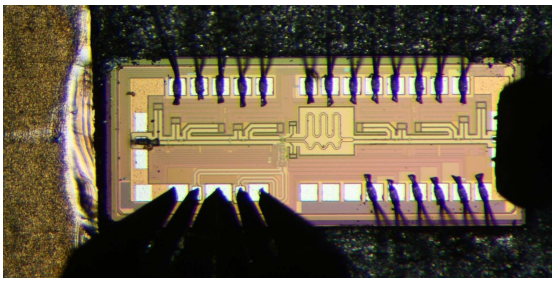

Fig. 4. *Chip photograph of the receiver chip with contacted probes, bondwires for dc supply, and on-chip monopole antenna.*

is a low direct current (dc) power consumption to minimize the required energy per transferred bit. With the target application being chip-to-chip communication for the HAEC system, some additional unique characteristics affect the TRX design.

- In spite of the large free-space path loss, the required link distances of a few centimeters can be covered with relatively low transmitted power. The losses in the Butler matrix can be compensated with moderate-power amplifiers at the interface with the antennas.
- Thanks to the proximity of receiver (RX) and transmitter (TX), it is possible to share a wired LO reference on the same board or across boards, directly synchronizing RX and TX for coherent modulation schemes. This avoids complex LO synchronization circuitry, such as phase-locked loops (PLLs), and relaxes the requirements on the shared LO source.

The upconversion path of the TX is an integrated circuit consisting of an active fundamental direct-conversion mixer and a passive balun for differential to single-ended transformation of the radio-frequency (RF) signal [42]. The required power level of the 190-GHz LO signal at

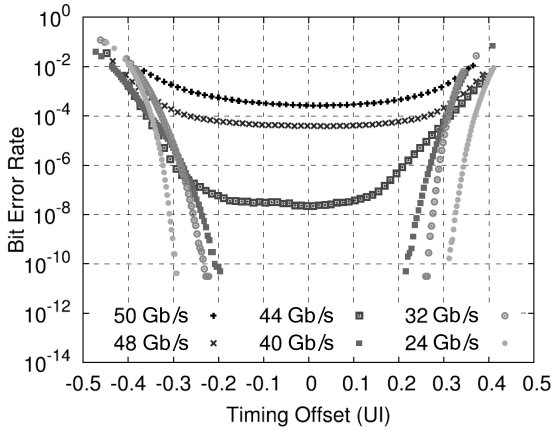

Fig. 5. *Measured BER as a function of the timing offset (bathtub curves) for 6-mm distance between receiver and transmitter bondwire antennas.*

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

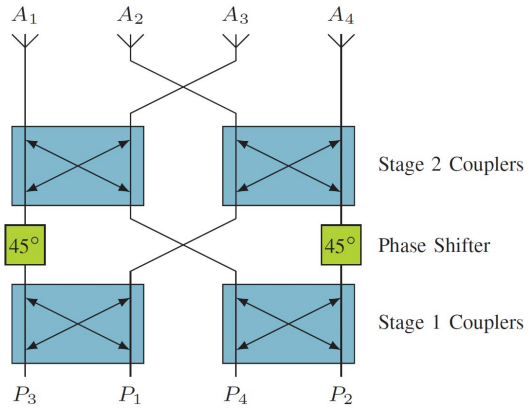

Fig. 6. *Schematic of a 4 × 4 BM BFN. The numbering of the feeding ports at the bottom is chosen such that when reading from left to right, the progressive phase shift at the antenna elements at the top is increased 90° between neighboring ports. For port P1 the progressive phase shift between the antenna elements A_1... A_4 is -135°, for port P_2 it is -45°, for port P_3 it is 45°, and for port P_4 it is 135°.*

the mixer is around −5 dBm, which can be provided with an external signal source of -20 dBm with the help of an integrated single-ended LO driver [39]. To enable ultralarge channel bandwidths, the TX is designed for baseband frequencies up to 30 GHz, corresponding to RF frequencies of 160–200 GHz. The chip consumes 32 mW, including the LO driver.

The downconversion path in the receiver RX is an integrated circuit consisting of a low-noise amplifier (LNA) [39], an active direct-conversion mixer [40], a variable gain amplifier, and an output stage. Furthermore, an LO driver and a balun are used to drive the LO port of the mixer. The integrated receiver is described in detail in [41]. The measured conversion gain is between 47 dB in high-gain (HG) mode and can be reduced by up to 27 dB in low-gain (LG) mode. The 3-dB RF bandwidth is 35 GHz and independent of gain control. The baseband output power reaches 1-dB compression at 1 dBm and −2 dBm in HG and LG mode, respectively. The measured minimum double-sideband noise figure is 10.7 dB. These performance figures are achieved with a minimum LO power of −20 dBm and a dc power consumption of 122 mW.

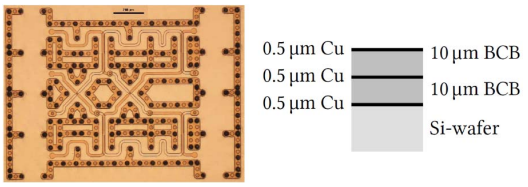

Fig. 7. *Photograph of fabricated 4×4 BM on BCB (left) and associated BCB buildup (right).*

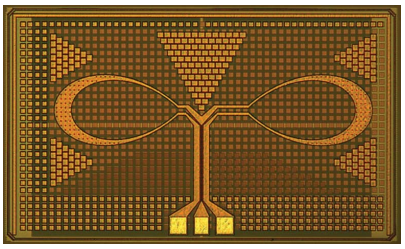

Fig. 8. *Photograph of the measured half-cloverleaf antenna with metal fill structures.*

The receiver and transmitter chip were equipped with bondwire antennas [116] and tested for wireless data transmission. With a moderate antenna gain of 5 dBi, a maximum rate of 50 Gb/s was demonstrated over a distance of 0.6 cm, while up to 40 Gb/s were possible over 2 cm, corresponding, respectively, to 3.1 and 3.9 pJ of energy per transferred bit [42].

B. Butler Matrix

One of the main reasons for choosing wireless links for interboard communication was the inherent flexibility, which allows adaptive logical network topologies and, thus, creates more direct connections between computing nodes with fewer hops. For the required beam switching, we consider a Butler matrix (BM) feeding network as an energy-efficient solution. The BM has been widely used in multibeam systems, which were first introduced in [15]. As a passive beam-forming network (BFN) for linear or circular antenna arrays, the main purpose is to provide a uniform amplitude distribution and a constant phase difference between the radiating elements in an antenna array. Thus, a BM with N input and N output ports can drive an antenna array with N radiating elements. A signal fed into one input port of the BM will be distributed into all radiating elements with equal amplitude and a specific phase to result in a radiation beam in a specific spatial direction. Switching through the N different input ports of the BM, N radiation beams at different spatial angles can be addressed [88]; see Fig. 6.

In Fig. 7, left, a layout for an on-chip 4×4 BM working around 180 GHz is shown, which follows the general concept as depicted in Fig. 6. The BM has been fabricated on a custom benzocyclobutene (BCB) process as shown in Fig. 7, right, which allowed for a dense integration and low loss at 180 GHz. Due to advances in 3-D chip stacking, this feeding network can be directly integrated into the chip stack; see Fig. 2(b).

C. Integrated Antennas

As for the Butler matrix, we also consider direct integration of the antenna arrays into the chip stack as the top layer. For the elements of these arrays one proposed

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

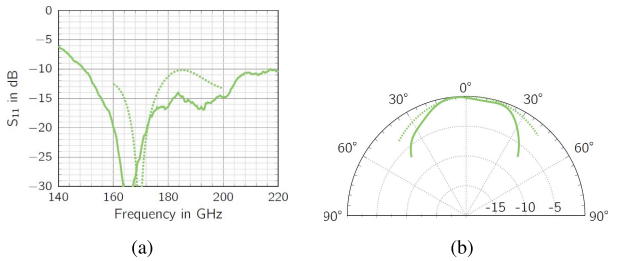

Fig. 9. *Comparison of the input reflection coefficient (left) and radiation pattern (right) between simulation (dotted line) and measurement (solid line) for the half-cloverleaf antenna at 180 GHz.*

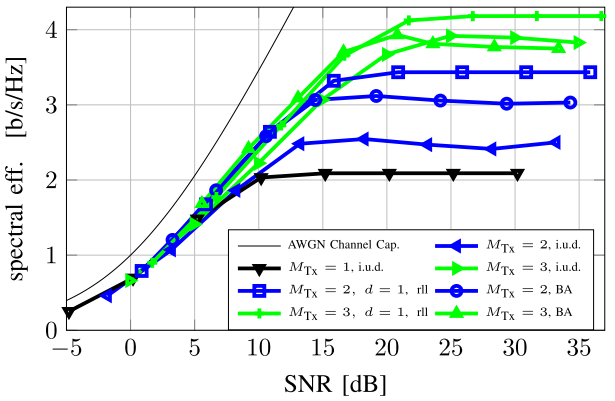

Fig. 10. *Spectral efficiency versus SNR for different sequence designs with FTN signaling rate M_{Tx} [75]; "i.u.d." means independent uniformly distributed input symbols, "rll" is runlength coding with minimum runlength d, and "BA" denotes sequences with optimized transition probabilities using the Blahut–Arimoto algorithm.*

antenna is the half-cloverleaf shaped antenna. The overall dimension of the antenna chip is 1.2×0.7 mm^2. Both antenna arms have the shape of a leaf and are fed by a coplanar stripline, which is connected via a line transition to a coplanar waveguide to fit the ground–signal–ground configuration of the on-wafer probes. Due to standing waves, which were noticed in a symmetric feeding structure, an asymmetric configuration has been chosen. In order to meet the requirements regarding the technological process, the layout has to be rasterized with a given grid and angles. More details about the antenna design can be found in [67]. A photograph of the fabricated chip is shown in Fig. 8, while the simulation and measurement results can be seen in Fig. 9.

D. Communication With 1-bit Analog-to-Digital Conversion

For very high data rate short link communication the power consumption of the analog-to-digital converter (ADC) becomes a major factor, also compared to the transmit power. This is due to the required high quantization resolution and the very high sampling rate. One option to circumvent this is coarse quantization and oversampling at the receiver with respect to the Nyquist rate. In this regard, 1-bit quantization is advantageous as it is simple to realize, is robust against amplitude uncertainties such that no automatic gain control is needed and no highly linear analog signal processing is necessary. Moreover, a 1-bit ADC requires only simple circuitry and does not need much headroom for sophisticated processing in the amplitude domain which allows to use energy-efficient circuits with a supply voltage smaller than 1 V [117]. Both the much simpler signal processing without an automatic gain control as well as the use of low voltage circuits yield a much higher energy efficiency of 1-bit quantization with temporal oversampling in comparison to standard fine-grained quantization at Nyquist rate. For this reason, we consider temporally oversampled 1-bit quantization for the wireless links. Optimal communication over the resulting channel requires an adapted modulation and signaling scheme as the information is carried in the distance of the zero-crossing time instants of the transmitted signal. A natural choice to encode information for transmission over such a channel is runlength coding [62].

The benefit of oversampling with respect to the Nyquist rate for the achievable rate of 1-bit quantized channels has been shown analytically (noise-free channels [43], [113]; low signal-to-noise ratio (SNR) [70]; continuous-time channel, high SNR [7]) and by means of simulation; see, e.g., [8] and [73]–[75]. Jointly with oversampling in time, faster-than-Nyquist (FTN) signaling [82] can be applied, which means to increase the speed of the digital-to-analog converter (DAC) at the transmitter without increasing the signal bandwidth. The above mentioned runlength coding is very suitable for FTN signaling [75]. The resulting intersymbol interference can be resolved by oversampling with respect to the Nyquist rate at the receiver and the use of sequence detection. Fig. 10 shows the spectral efficiency for quadrature phase-shift keying (QPSK) input symbols with different sequence design. For details, see [75]. The oversampling rate at the receiver always equals the FTN signaling rate M_{Tx}. Moreover, a 90% power containment bandwidth has been assumed. It can be seen in Fig. 10 that oversampling can significantly increase the achievable rate and runlength coding is an appropriate sequence design approach.

E. 1-bit Analog-to-Digital Converter

An oversampling 1-bit ADC can be implemented adapting architectures usually employed for the implementation of time-to-digital (TDC) converters; this relies on the fact that if the bit-transition times are preserved, then the binary input can be fully recovered. This concept of a 1-bit ADC is illustrated in Fig. 11, which shows a binary symbol sequence with its time-delayed replicas. Each of them preserves the information of the symbol transitions and intertransition distances. Sampled at a certain point in

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

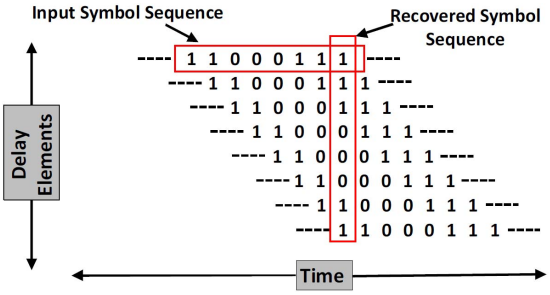

Fig. 11. *Principle of operation of the proposed concept for the 1-bit ADC.*

time the delayed replicas contain each symbol of the input symbol sequence as can be seen in Fig. 11. The delayed replicas can be used to recover the symbol pattern of the input with a parallel slower output. This is true under the assumption that the delay of each delay element is, at maximum, equal to the symbol period.

Due to runlength coding, the transmit sequences consist of consecutive 1s and 0s. The information is encoded in the transition times inside the sequence, which must be resolved by the ADC. As discussed in Section IV-D, oversampling is applied jointly with FTN signaling. That is, if the transmit symbols are transmitted at a rate of M_{Tx}, the ADC must sample at that rate. Hence, the time delay of one delay element must correspond to the time between two transmit symbols. This means we sample at symbol rate and oversampling is meant with respect to the Nyquist rate.

Fig. 12 presents a system-level schematic of the proposed 1-bit ADC consisting of a delay line with sampling flip-flops. As the binary symbol sequence $y(t)$ travels along the delay line, which consists of a series of inverters, its delayed versions are generated after each delay element. The output of each inverter is connected to the following delay element and the sampling flip-flop. The entire delay line is then sampled by a reference clock f_{CLK}, whose period T_{CLK} is set equal to the delay of the delay line. The sampled digital word now contains the information about the total number of transitions and the distance between them inside the sequence, which lies within one T_{CLK}. In other words, a snapshot of the input signal is recorded at every sampling instant.

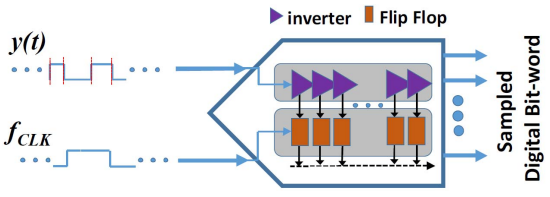

Fig. 12. *System-level architecture of the proposed 1-bit ADC.*

A complete design of this architecture for integration in a 45-nm SOI CMOS technology shows that a 25-GHz input binary signal can be digitized with an equivalent sampling rate of 200 GHz, while consuming 270 mW of dc power. We are currently further optimizing the 1-bit ADC from which we expect a halving of the power consumption. For comparison, for 3-bit single-core flash ADC with a sampling rate of 24 GHz, a power consumption of 0.4 W has been reported in [122] showing the advantage of 1-bit quantization with temporal oversampling in terms of energy efficiency.

F. Low-Latency Channel Coding

To compete with state-of-the-art memory access delays, we consider spatially coupled low-density parity-check (SC-LDPC) codes to fulfill the 100-ns delay constraint.

SC-LDPC codes combine the advantages of allowing operation close to channel capacity and enabling decoding with a small latency. They have been introduced in [63] and are a combination of LDPC and of convolutional codes. They are constructed by connecting L independent LDPC codes into a coupled chain, which is equivalent to introducing memory into the code. It was shown in [71] that for large L and for large memory SC-LDPC codes achieve capacity. To make the structural latency independent of L, a sliding window decoder was introduced. The decoding performance increases exponentially fast with the window size W which makes SC-LDPC codes an interesting candidate for applications with stringent latency requirements.

It was demonstrated in [54] that SC-LDPC codes perform better, in terms of BER, than the block codes from which they are derived, even at low latency (< 500 bit). Furthermore, SC-LDPC codes also outperform convolutional codes for equal latency. This performance gain comes at the cost of increased decoding complexity. Therefore, nonuniform decoding schedules are proposed in [123]. By switching off nodes inside a window once they stop showing an improvement in their BER estimate, complexity can be reduced by 50% without any loss in performance. In [104], SC-LDPC codes have been optimized using a differential-evolution-based algorithm. The optimization results in SC-LDPC codes that approach capacity for a window size of $W = 5$ and ensure systematic linear time encoding. This window size allows to design a code having a small structural decoding latency which gives sufficient temporal headroom for receiver processing such that the 100-ns delay requirement can be fulfilled.

G. Energy Estimation

Since several components of the wireless links are still under development, it is very challenging to give an accurate estimate of the energy consumption of the wireless links per transmitted bit. Nevertheless, we will estimate the energy consumption based on measurement results acquired for the components developed so far and based on predictions for components which have not been designed yet. Our aim is to achieve a data rate of 100 Gb/s

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

Table 1 Link Budget Parameters for Board to Board Communication Using 8×8 Antenna Arrays

	unit	value
Rx noise figure	dB	15
carrier frequency	GHz	200
bandwidth	GHz	30
path loss exponent	-	2
path loss for shortest link 2 cm	dB	44.48
path loss for longest link 10.79 cm	dB	59.12
antenna array		8×8
array gain (for Tx and Rx each)	dB	18
Butler matrix inaccuracy	dB	15
polarization mismatch	dB	3
implementation loss	dB	5
Rx temperature	K	323

while fulfilling the latency requirement of 100 ns. Given two orthogonal polarizations, this results in 50 Gb/s per link. Assuming the use of a rate 1/2 channel code to ensure reliable transmission, we end up with a coded bit rate of 100 Gb/s per link.

For the link budget, we here assume the use of 2-D antenna arrays with 8×8 antennas elements which are fed by a Butler matrix beam-switching network with a predicted loss of 15 dB. Assuming a power amplifier with a maximum output power of 6 dBm enables a receive SNR of up to 13.62 dB for the longest link (worst case) for the link parameters in Table 1. Thus, based on Fig. 10, it can be seen that a data rate of 50 Gb/s on the 30-GHz bandwidth channel and even more is achievable.

For the analog frontend components at the transmitter and receiver, an energy consumption of 3.1 pJ/b at an uncoded bit rate of 50 Gb/s have been measured as described in Section IV-A. While this setup just considered binary phase-shift keying (BPSK), i.e., no complex modulation scheme, for the final system we will use QPSK, i.e., signalization in the in-phase and the quadrature component to support 100-Gb/s coded bit rate. This will lead to a slight decrease of the energy consumption per transmitted bit, which we however neglect here, as the uncertainty of the power estimates for some other not yet measured components is larger. Not considered in Section IV-A is the energy required for the power amplifier at the transmitter which can be estimated to add 2.2 pJ/b while providing a gain of 20–24 dB and yielding a maximum output power of 6 dBm. Thus, the analog frontend components including the power amplifier yield an energy consumption of 10.6 pJ per information bit.

The oversampled 1-bit ADC consumes a power of 270 mW and is needed twice, once for the in-phase and once for the quadrature component (see Section IV-E), yielding an energy per bit of 10.8 pJ/b. The cor-responding DAC is estimated to use the same power. The estimated energy for channel decoding is 13.35 pJ/b using five decoding iterations [85]. Finally, some margin for the demapping and demultiplexing process is added. Their complexity corresponds at maximum to one channel decoding iteration. Thus, the overall estimated energy consumption is 48.22 pJ/b. Note that there are further significant savings of the energy consumption to expect, e.g., for the 1-bit ADC.

V. NODE ARCHITECTURE

A. Processor and Memory Organization

On each side of the PCB board in the HAEC Box there are 16 computing nodes connected by the optical interconnect with each other and with the wireless interconnect with the nodes in other adjacent PCBs. As mentioned in Section II, each node is envisioned to be a 3-D vertical stack of 128 layers. Each layer contains an equivalent of 64 000 processors of which 10% are actual processing elements (PEs) and the remaining 90% are memory elements (MEs). There are no techniques available today that can address the architectural design, power, and thermal challenges associated with the node.

The first challenge in designing the node architecture is scalability. Given the huge number of components (PEs and MEs) in one node, the interconnection between them must be designed as compact as possible, yet still able to maintain a high quality of service in terms of throughput, latency, resilience, and security (see Section V-B). Otherwise, a significant amount of chip area will be taken solely by the interconnect. It can be more than 10% as presented in [48] without fault-tolerance mechanisms, security features, and multilayer 3-D communication infrastructure. Another issue with the interconnect is energy efficiency, as it can consume up to 35% of total energy of the whole system [1]. Due to the *dark silicon* phenomenon mentioned in Section II-C, there is only a small fraction of the nodes that can be active at any time. It poses a question of how, when, and which components including the interconnect should be activated/deactivated to execute a specific task. Since 90% of components of the node are MEs, it is more efficient to "move" PEs to where the data reside instead of transferring the data to PEs. This practice is known as *near-memory computing*. In this case, PEs are activated only when they are close to where the data are stored (in MEs) and when it needs to be processed upon. The MEs will be implemented using one of the emerging memory technologies such as the nonvolatile, low-power, and low-latency resistive RAM [87]. Therefore, the MEs

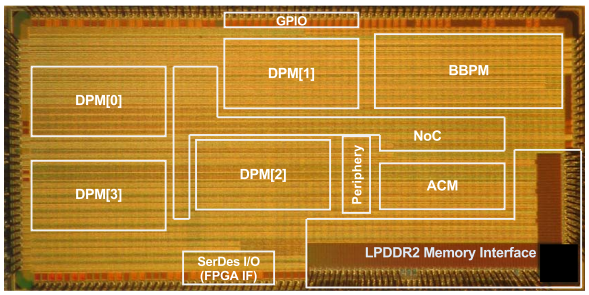

Fig. 13. *Tomahawk4 die photograph comprising data processing modules (DPM), baseband processing modules (BBPM), application control module (ACM) connected by network-on-chip (NoC).*

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

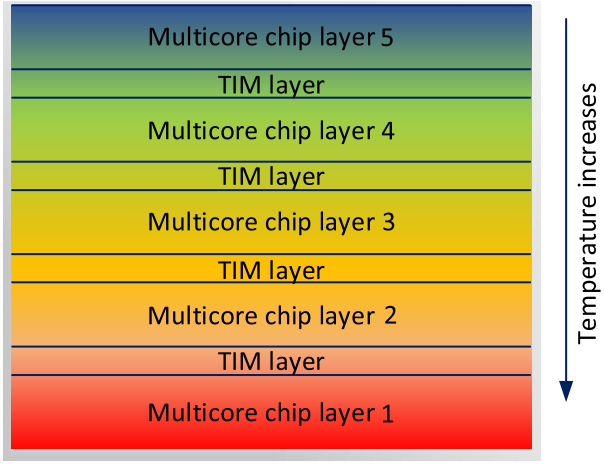

Fig. 14. *Temperature change in multiple layers of a 3-D chip. TIM refers to thermal interface material.*

can also be deactivated to save power without losing the data. Tomahawk4 is a processor that has been designed and fabricated to demonstrate some of the concepts of HAEC Box, e.g., the data movement principles and the low-power adaptive execution [3].

The next challenge is to determine the distribution of PEs and MEs across 128 layers of the node. It affects not only the performance, but also the thermal characteristics of the system. As mentioned above, due to the power limitation, the power management strategy switches all components to a power-saving mode that are not in use by the current applications. As a result, each application, and only that application, should have exclusive control over the MEs assigned to it. However, the applications require different amounts of memory. The smaller the MEs are, the bigger the number of MEs is, and correspondingly more communication resources that are needed for each application. However, if the MEs are larger, there may be more unused memory space and fragmentation. Thus, we propose to have MEs with various capacities. This is similar to the problem of having mixed page sizes [24]; however, in this case, it is applied to the actual physical memory rather than virtual memory. Fig. 15 shows an example configuration with mixed-size memory modules. M1 denotes a unit-size memory, while M2 and M4 are two times and four times as large as M1, respectively. Further, this can be used to mitigate the thermal issues. The most common method of 3-D stacking is through silicon via (TSV)-based stacking. One major drawback of TSV-based stacking is temperature rise in the chip. Each layer in a 3-D chip generates heat based on its power dissipation, but only the top layer is able to dissipate heat as shown in Fig. 14. Our simulation results with various benchmarks show that in order to maintain similar temperature across multiple layers in a 3-D stack with primarily logic, only about one fifth of power can be dissipated on every successive layer away from the heat sink. For example, in Fig. 14, if layer 5 is able to dissipate 5W, then layer 4 should only be allowed to release 1W, layer 3 0.2W, and so on. Therefore, PEs should not be placed directly on top of each other in two consecutive layers. Fig. 15 illustrates a possible solution by making use of MEs as the "shield" between PEs for better thermal management.

Due to the unbalanced placement of MEs and PEs, the memory access latencies from one PE to its MEs are different. Therefore, the MEs should be logically organized and assigned to PEs such that the access latency variation is kept to a minimum. Furthermore, within each application, the memory allocations to each PE can be different. One ME can be exclusively owned by one PE while the others may be shared by multiple PEs. The flexibility of managing MEs access rights should be addressed by a separate memory hierarchy infrastructure instead of the main interconnect which is already complicated. It facilitates the allocation of MEs to PEs by internally connecting multiple MEs together to create a hardware-assisted virtual continuous address space. Thus, the PEs do not need to send multiple memory access requests to multiple MEs directly.

B. Resilient Hexagonal NoC

Network-on-chip (NoC) has emerged as the highly scalable and efficient packet switched communication network for such large heterogeneous systems [9]. However, with rapid scaling of transistor gate sizes, components of the NoC are becoming highly susceptible to transient and permanent faults [98], making NoC reliability a great concern. Redundancy is the general approach to fault tolerance and the hexagonal NoC (hexNoC) with redundant diagonal inter-router links, as shown in Fig. 16(a), provides increased resilience to NoC link and router failures or blocking due to path reservations.

By developing fault-tolerant routing algorithms for the hexNoC, the resilience of the hexNoC against permanently faulty routers was improved [89]. The routing algorithms are adaptive and based on the turn model [44], in which certain turns in the movement of packets are prohibited to prevent deadlock. An approach based on matrix algebra to determine the transitive closure of the channel dependency matrix simplifies the process of turn selection and also

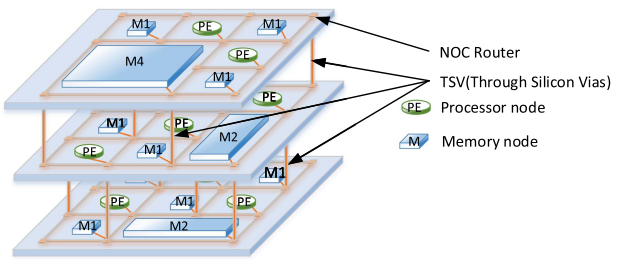

Fig. 15. *PEs and MEs are distributed in different layers.*

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

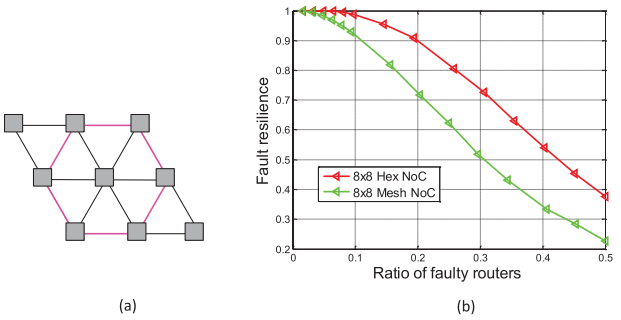

Fig. 16. ***Performance of the hexagonal NoC against faulty routers. (a) 3×3 hexagonal NoC. (b) Comparison of fault resilience of 8×8 mesh and hexNoC.***

allows for the analytical assessment of fault resilience [90]. The results show that the hexNoC provides three node/link disjoint paths for any source–destination pair as compared to two disjoint paths in mesh networks. Thus, the hexNoC (with approximately 50% more links) can tolerate any location of two faulty routers or links whereas the mesh can tolerate only 1. Moreover, because of average shorter paths, the throughput of the hexNoC is also improved (by 18% for 8×8 NoC with two faulty routers). The performance with greater number of faulty routers was obtained by implementation and simulation of the hex and mesh NoCs with adaptive routing in an in-house developed C/C++ based cycle-accurate simulator. The results depicted in Fig. 16(b) for the 8×8 NoC show that, for 30% faults, the hexNoC has a 43% higher resilience than mesh.

For addressing real systems with latency and throughput constraints, the NoC must also be able to support guaranteed services. Due to its increased path diversity, the hexNoC is also capable of superior performance when supporting guaranteed services.

VI. NETWORK ARCHITECTURE

In addition to providing a resilient network hardware for communication, it is important to employ network coding (NC) to exploit the benefits of the various communication resources available in the HAEC Box effectively. In this section, we first describe the NC applied in this context. We then move on to describe a simulation platform that has been developed to evaluate the impact of various networking protocols and coding on the overall system performance. Finally, we describe the HAEC Playground–a hardware prototyping testbed, which allows us to instantiate arbitrary network topologies easily for experiments and evaluation.

A. Network Coding for the HAEC Concept

In contrast to source or channel coding, NC is not limited to end-to-end communication, but is able to operate in the distributed settings resembled within the HAEC Box, e.g., multipath or/and multihop communication, distributed storage. In Fig. 17, a multihop communication example is given to illustrate the reduction of transmission time slots achieved by NC. The figure depicts the classical store and forward approach occupying in this case four transmission time slots. If the relay is able to code the incoming packets, e.g., by bitwise XOR, the number of slots can be reduced to three. This is called digital intersession NC. In the HAEC Box, we use two further concepts of NC illustrated in the bottom of Fig. 17. The first is analog intersession NC, where the relay broadcasts the received information after some minor modifications. This form of NC requires no computation at the relay and is used for the communication between boards [79]. The second form is digital intrasession NC, which will be applied in the HAEC Box for distributed storage using random linear network coding (RLNC).

The digital intrasession NC approach offers benefits for more advanced use cases. Here RLNC is used allowing for 1) recoding capabilities of subsets of coded information in a distributed setting; 2) rateless features like fountain codes; 3) versatile coding matrix open for sparsity (adding zeros in an intelligent fashion); 4) low-latency support due to on-the-fly coding capabilities; and 5) support of heterogeneous field sizes for communication entities. The digital intrasession NC example in Fig. 17 is a simple multihop example using RLNC. It is based on two error-prone links with erasure probability of ϵ_1 and ϵ_2, respectively. Using an end-to-end code between nodes A and B would require redundancy to cope with a virtual channel erasure probability of $1-(1-\epsilon_1)(1-\epsilon_2)$. Enabling the relay to recode information would end up in a virtual channel erasure probability of $\max[\epsilon_1;\epsilon_2]$. The recoding takes place with the first incoming packet from node A and is therefore different in terms of latency and buffer occupancy to a full split end-to-end coding between node A and R as well as R and B.

But RLNC is not limited to communication and can be used for distributed storage as well. The latter is used in the HAEC concept to lower the load of certain computing regions, and reduce the energy usage to maintain a certain resilience level [16]. The distributed storage approach allows for 1) efficient redundancy; 2) intentional switch-

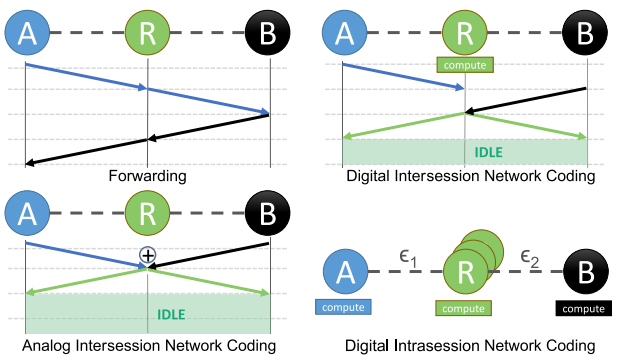

Fig. 17. ***Benefits of different forms of network coding.***

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

ing off of computing regions without losing data; and 3) relocating storage efficiently within the HAEC Box. The latter is needed to reduce energy cost for given applications by placing data and application in close proximity within the HAEC Box as well as to optimize for switch-off patterns (cf., Section VII-C). In order to reduce the cost in terms of energy consumption for the RLNC approach, several works [128], [129] aim to exploit the multicore architecture of HAEC for distributed RLNC coding schemes. For example, benchmark results show that parallelizing the encoding process over four cores speeds up the utilization by a factor of ten [129].

B. System Simulation

Since it is not always possible to have a working system to test high-level ideas like network coding, simulation plays a vital role in early-stage evaluation. HAEC uses a custom simulation framework, HAEC-SIM [11], that is focused on parallel simulation and combining performance and energy prediction. The design process of the HAEC Box and its components also relies on simulation for critical evaluation. There are isolated low-level simulations or measurements for components of the optical and the wireless links, as well as a holistic simulation that models the concept of the HAEC Box beyond isolated hardware prototypes. The final simulation combines models from various levels: applications, software runtime, network coding, and physical links. With this holistic, integrated modeling approach, the impact of individual design alternatives on the HAEC Box as a whole can be evaluated. Further, the simulation approach allows predicting the behavior of complex dynamic workloads.

The simulation relies on event traces of parallel applications. These traces are typically recorded on existing systems using the Score-P [69] performance measurement infrastructure and contain events of the application (e.g., entering a function), explicit communication events (e.g., sending a message), as well as system measurements (e.g., power measurements). All events also contain information about their location (thread, process, hardware component), and are ordered by time. The simulation framework processes the events in parallel with a mapping of one recorded thread to one simulation worker process. Models within the simulation framework can modify existing events or create new ones. For example, network models process communication message events and change their time to reflect the link performance of the target system. While the focus of the simulation is on explicit message exchange between processes executing within the application, the computational speed can also be adjusted. The models for the computational components within the HAEC Box are still in an early stage, but a very high level of parallelism is expected for each node as mentioned in Section V. Therefore, we do not aim to simulate individual cores of one node, but instead, model the performance of processes consisting of many threads. During the simulation, the state of the modeled hardware is used to add energy metrics that represent the dynamic power consumption of HAEC Box components to the event stream. The resulting event trace reflects the predicted execution with respect to time and power on the target architecture.

These traces can be analyzed using the Vampir [68] trace visualizer. Evaluating different model parameters and their impact on dynamic performance and energy consumption of an application is possible using a comparison of the trace resulting from the simulation. The comparison can be either made visually using Vampir to focus on specific criteria not easily machine readable or using an automated process focusing on standard metrics like duration, transfer volume, total energy cost. The gathered insights provide feedback on how individual HAEC Box components, e.g., network links, network coding algorithms, can work in the broader architecture and therefore facilitates a codesign process on different hardware and software layers.

C. HAEC Playground

In the previous section, we have presented HAEC-SIM as a main tool to evaluate the architecture of the HAEC Box. However, in order to combine research findings from hardware, architecture, and software projects, one would need a holistic testbed. It will also serve as a proof-of-concept to demonstrate the applicability of research innovations. That testbed has to satisfy several requirements. First, one can flexibly create HAEC Box topologies as well as state-of-the-art ones to compare their performance side by side. Second, the testbed should allow for energy measurement at a high precision and a reasonable cost. Third, operating systems and software packages running on the testbed have to support energy-aware software with ideally no overhead due to, for example, porting their code running on one CPU architecture to another. Last, the testbed should introduce a minimum cost in construction and operation.

Since a pure hardware-based solution implies inflexibility, our approach is to combine both hardware, to provide computing power and physical connections between computing nodes, and software, to flexibly create virtualized networks of arbitrary topologies, leveraging virtualization technologies for both computation and networking. Specifically, we deploy a cloud management software to reuse its helpful functionalities such as resource aggregation, automated networking services (e.g., building virtual switches or routers, IP address assignment, etc.), and managing multiple simultaneous projects and users. For the underlying infrastructure, we advocate commercial-off-the-shelf (COTS) hardware due to its economical advantages and standardized interfaces. We employ single-board computers (SBCs) due to their small sizes, yet reliable operation. Fig. 18 illustrates the design and realization of our testbed, HAEC Playground.

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

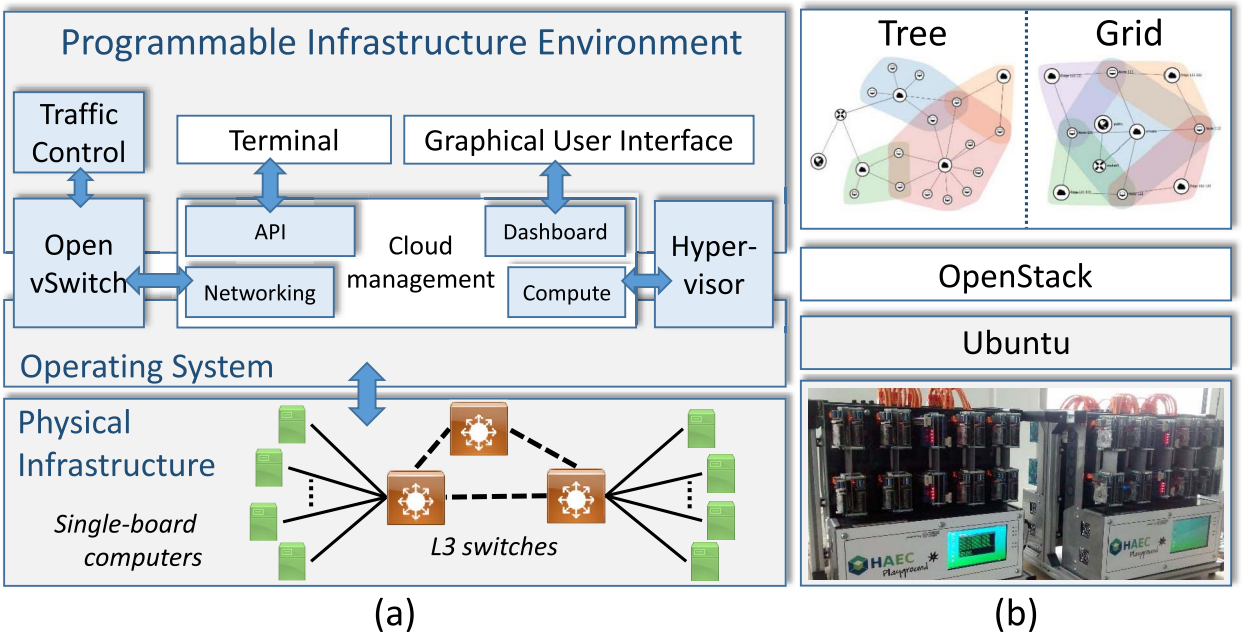

Fig. 18. *HAEC Playground testbed: (a) Conceptual design; and (b) realization consisting of SBC Odroids assembled into movable hardware modules, OpenStack software deployed to aggregate computation resources, two exemplary topologies (tree and grid).*

To enable connectivity between all compute nodes in the testbed, we organize the SBCs in a star topology. Furthermore, to facilitate a modular and extensible setup, we organize SBCs into subgroups interconnected by stackable switches. When a larger setup is needed, several subgroups should be aggregated by interconnecting switches into a ring. As SBC we selected Odroid XU-4 from Hardkernel,[1] each is equipped with eight ARM CPUs with the heterogeneous big.LITTLE architecture, meaning four high-performance Cortex-A15 cores and four energy-efficient Cortex-A7 cores. This design is also extensible to future needs, e.g., one can connect a high-performance computer with large-volume storage and fast network connections to the testbed, supporting resource intensive operations. The cloud management software deployed on the testbed is the key to provide the programmable infrastructure environment. We decide to deploy OpenStack[2] in our testbed, since it is the most mature and active open-source project of its kind. The overall advantage of OpenStack is that it facilitates an emulation environment, meaning that our networking setup is fully compatible with real-world networks. Furthermore, the environment provides both a graphical user interface to visualize aggregated resources for management, instantiated networks and computing nodes for verification, as well as a command-line interface with application programming interface (API) for repeatable setups which is crucial when conducting experiments. Further detailed description of the HAEC Playground can found in [79].

[1] www.hardkernel.com
[2] www.openstack.org

All in all, the testbed allows us to instantiate networks of arbitrary topologies, including the current $4 \times 4 \times 4$ of the HAEC Box for evaluation purposes. Furthermore, standardized interfaces allow to integrate modules for optical and wireless connections, demonstrating the interoperability of hardware, architecture, and software research outcomes. The HAEC Playground has contributed in several scientific activities such as [95], [128], and [130].

VII. SOFTWARE STACK

The HAEC Box requires a novel software stack centered around adaptivity and energy efficiency to unleash its full potential. The key challenges posed by the disruptive hardware architecture of the HAEC Box are: 1) the energy management of the vast amount of cores and interconnects; 2) the support for the full life cycle of a next generation of highly scalable and adaptive applications exposing a huge configuration space at runtime; and 3) the actual reconfiguration of hardware and software resources at runtime.

To address those challenges, we propose the software stack depicted in Fig. 19 comprising the following major components.

1) HAEC operating system (OS): The HAEC OS is the interfacing component between the hardware of the HAEC Box and the software infrastructure. Besides the traditional resource management, it provides additional functionality for obtaining hardware sensor information and for reconfiguring energy-control features of cores and interconnects (Section VII-A).
2) Compiler and language support: To enable highly adaptive software, the compiler needs to offer the

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

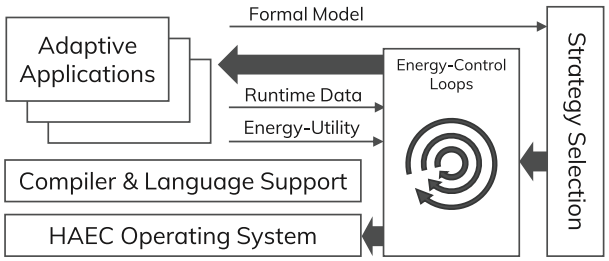

Fig. 19. ***Software stack overview of the HAEC Box featuring energy-utility function-based hierarchical energy-control loops.***

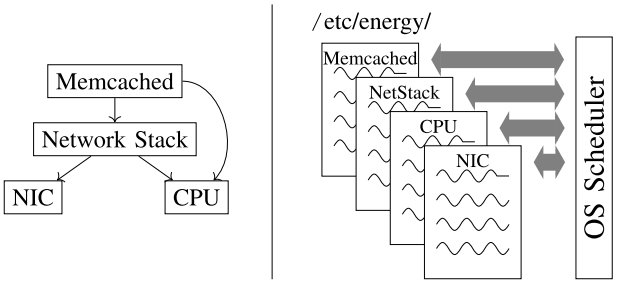

Fig. 20. ***Energy-utility descriptions for OS directed management of resources.***

ability to automatically generate different code variants for individual software components at development or even at runtime (Section VII-B).

3) Adaptive applications: While traditional applications behave mostly statically, the HAEC Box requires applications that adapt to the current workload, for instance, by autonomously readjusting the degree of parallelism or exchanging software components at runtime. As a specific example, we will present an adaptive database kernel that offers a rich set of adaptivity facilities (Section VII-C). We expect many other applications to profit from a system as proposed here. In future work, we will concretely work on base-station processing for fifth-generation (5G) communications and on particle and mesh-based simulations for computational biology.
4) Energy-control loops (ECL): The actual decision making is performed by a set of reactive control loops that run at different invocation latencies. For instance, while energy-related hardware reconfigurations need to occur at a low latency within the operating system, software reconfigurations or data movements are long-term subjects. Each ECL periodically leverages hardware and software sensor data to trigger reconfigurations based on energy-utility functions (EUFs) expressing the tradeoff between a target performance metric of a specific resource (utility) and the required energy budget (Sections VII-A–VII-C).
5) Strategy selection: To determine the optimal decision-making strategy of the individual ECLs for the current workload and system load, we employ probabilistic model checking, with formal models of the adaptive applications that are analyzed at development time in terms of energy-utility (Section VII-D).

In the following, we will discuss the individual components of the software stack of the HAEC Box in more detail.

A. Operating System

It is the responsibility of the operating system (OS) to manage resources in a system and to allocate them to tasks. We manage energy as one such resource according to the previously described principles of energy-utility. To schedule energy, the OS needs to be aware of the requirements of software and be aware of the characteristics of available hardware resources.

We use a management system that is based on textual descriptions of the relationship between a resource's performance and its requirements. Following UNIX traditions, we place these descriptions in the `/etc/energy` directory structure, including files for each resource that describe which other resources it requires, its configuration options, its modes of operation, and how it uses other hardware or software resources of the system. For example, a description of a memcached server specifies how it uses the network stack (a software component) and the CPU. In turn, the network stack uses the CPU to process packages and the network interface card (NIC) to send and receive packages. The network card and the CPU consume power depending on the software's usage pattern. The hierarchy and description for the example are illustrated in Fig. 20. This general modeling approach allows describing the tradeoffs of the different communication mechanisms of the HAEC Box, e.g., wireless or optical as discussed in Sections IV and III.

These text files are used by the OS schedulers to trade off resources against each other while satisfying user requirements. Possible user requirements may be an upper limit on power consumption or a lower limit on performance. The collection of resource description files represents the system-wide energy-utility tradeoffs available to the OS. We illustrate this in Fig. 21 for the memcached example. Each point is a different configuration of the sys-

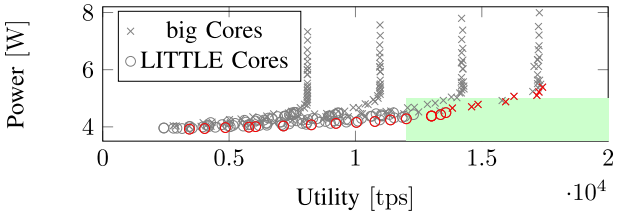

Fig. 21. ***System energy-utility profile for memcached (using either big or little cores).***

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

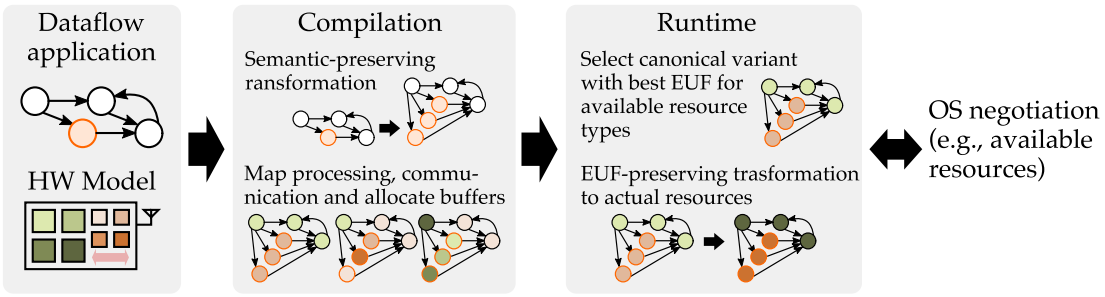

Fig. 22. ***Overview of variant generation flow.***

tem (CPU frequency, core count and memcached cache size). We are only interested in the Pareto frontier of this profile, highlighted as red marks. The points show the available configurations and their power and performance (or utility) values. If the user specifies constraints, e.g., that the system must perform at least 12 000 transactions per second (tps) and may not use more than 5 W, only a subset of configurations is valid (green highlighted area in Fig. 21). If the system is now optimizing for energy efficiency it will pick the lowest power configuration from the valid ones.

The global optimization problem that arises when trading different resources against each other is expensive to solve [47] and thus not suitable for frequent reconfiguration. However, most optimization decisions can be made locally at the resource level. We coordinate these decisions using local optimization with global coordination [53], where individual, resource-local schedulers optimize the local resource (such as the CPU [126] or the network [51]) within limits set by a global coordinator. Global coordination is only performed infrequently in case of imbalances. In this case, the OS reevaluates the whole system stack and adjusts the constraints for the configuration options of individual resources. For the memcached example, this may mean limiting the CPU frequency to a maximum of 1.2 GHz as configurations with higher frequency do not satisfy the user constraints or do not fall on the Pareto frontier.

In addition to regular intervals, global reoptimization will be performed when workloads (e.g., transaction types or memcached miss rate) shift significantly. Then the previous global optimization decisions made by the OS may no longer be optimal or may even no longer satisfy the users performance and energy constraints. These imbalances can be detected (e.g., when resource local schedulers detect under utilization or over utilization) and will trigger global reoptimization.

We expect this scheme to scale from small embedded systems over regular servers to whole data centers, where aspects such as power supply and air conditioning all factor into the energy efficiency of the facility. The central challenge will be to adapt our method to the revolutionary network architecture of our future HAEC Box.

B. Code Variant Generation

The HAEC OS requires a description of the energy-utility tradeoffs of software (cf., Fig. 21). For user-level applications, we work on language and compiler support to provide variants annotated with descriptions for the OS. As underlying programming model, the tool flow supports parallel dataflow programming models [27]. These models have clearly defined semantics, allowing us to perform compile-time and runtime transformations to generate implementation variants without tampering with the application behavior. To select and modify variants, we propose a runtime layer that adapts the variants to the resources made available by the OS. In this section, we introduce the variant generation flow, as depicted in Fig. 22.

1) Language and compiler: The dataflow programming model is well-suited for streaming applications [115], e.g., signal processing, multimedia, or even big-data processing pipelines. Due to its clear semantics, tools can automatically reason about possible schedules, memory pressure, and other properties at compile time or runtime. Many different languages from different computing domains have been proposed over the years, like StreamIt [120] or Cal [32]. We use CPN [17], an extension to the C programming language, to represent process networks. We extended the language to express potential semantic-preserving transformations [65] (represented by the orange process in Fig. 22). This allows, for example, expressing that a given process may be replicated to increase the application parallelism (similar to the parallel-for annotation in OpenMP). To be able to optimize for a given architecture, the compiler requires models of the underlying hardware. We use industry-standard processor models for performance and energy [114], and have abstract models capable of representing the communication modules described in Sections III and IV. The schematic of a big.LITTLE architecture is used to illustrate the hardware model that enters the compilation flow in Fig. 22, with four *big* cores (green squares) and four *little* cores (orange squares).

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

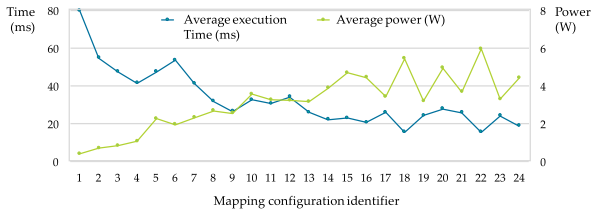

Fig. 23. ***Example variants for a dataflow implementation of a JPEG application on an ODROID-XU4 board. The figure shows performance and power numbers when processing images of 100 × 100 pixels. Lower configuration identifiers correspond to variants that use mostly little cores.***

2) Variants: Given a dataflow application, the compiler can transform the application graph, the mapping of computation and communication to system resources, and the amount of memory allocated for communication. All these parameters make up a large design space that we explore with heuristic algorithms [18]. Each variant represents a different point in the energy-performance space, with those beyond the approximated Pareto frontier being discarded. The compiler exports the resource demands as well as the performance and energy estimates for every exported variant. This can then be used by the OS or other higher level resource managers in the hierarchical ECL architecture (cf., Fig. 19), as demonstrated for microservers in [52]. An abstract view of the compilation process is shown in Fig. 22, where for a given semantic-preserving transformation different mappings are computed. The three different variants on the bottom represent mappings that use two, four, and five different cores. Each of the mappings generated by the compiler is *canonical* and represents a family of equivalent mappings in terms of performance and energy consumption. Equivalent mappings are not found via execution or simulation, but by a mathematical modeling of symmetries of the application and the target hardware using group theory as described in [45]. Intuitively speaking, our approach allows the compiler to understand valid *rotations* that can be performed to a mapping without changing the outcome, e.g., swapping tasks among cores. For illustration purposes, we show a set of configurations and their performance-power characterization in Fig. 23 for the dataflow implementation of a JPEG application described in [18]. We applied this principle to simplified base-station applications as reported in [19] and [46]. This will serve as a starting point for a future 5G case study for the HAEC Box.
3) Runtime: As described in Section VII-A, there are different levels for optimizing software, ranging from local and fast to global and time consuming. We developed a runtime system, which we call Tetris [46], tailored for dataflow applications that enables fine-grained local optimizations. Given a set of available resources by the HAEC OS, Tetris finds a canonical mapping from those generated at compile time (cf., Fig. 22). If not specified otherwise, Tetris would select the variant that executes the fastest. Then, to accommodate to the actual resource instances available, Tetris transforms the mapping while respecting the problem symmetries. As a consequence, the final variant that is deployed in the system has the same energy-utility value of the canonical version selected in the first place.

C. Adaptive Database Kernel

In this section, we present details of our database kernel prototype as a sample application aiming at massive *scalability* and *adaptivity* to exploit the highly parallel adaptive hardware of the HAEC Box. This database kernel is based on the *data-oriented architecture*, which turned out to exhibit superior scalability properties [66], [93], [96], and is aligned with the near-memory computing approach of the HAEC Box. In Fig. 24, we visualized the overall architecture including a specific instance of an ECL.

The HAEC Box features massive computational parallelism and main memory with extreme NUMA effects, e.g., remote nodes are accessible via the sophisticated network of optical and wireless interconnects. However, even on the HAEC Box this comes at the cost of reduced bandwidth as well as increased latency and energy consumption. Thus, the data-oriented architecture describes the principle of implicitly partitioning data objects and storing the individual partitions in the local memory of a specific socket. Data partitions are processed by worker threads running on preferably node-local compute cores, which communicate via a high-throughput message passing layer that is actually the main occupant of the interconnect network. Nevertheless, due to the static worker-to-partition mapping in this data processing architecture, it is not possible to turn off single cores (running a worker thread) without losing access to the associated data partitions. To get rid

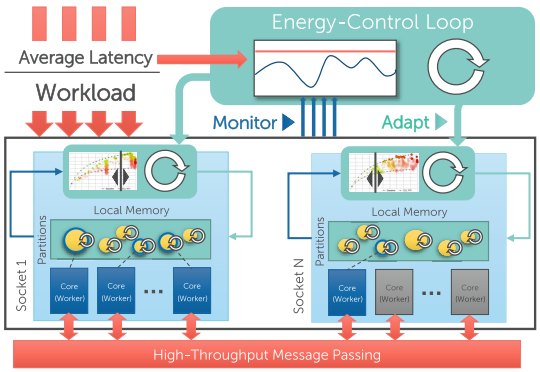

Fig. 24. ***Hierarchical resource adaptivity-specific ECL for the database kernel.***

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

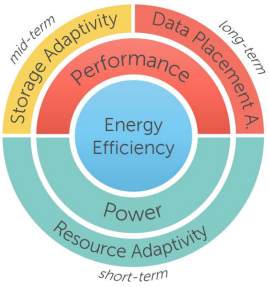

Fig. 25. ***Adaptivity facilities and relation to energy efficiency improvements.***

of this shortcoming, we relax this static mapping, which requires additional logic within the message passing layer. However, this enables us to elastically turn off arbitrary compute cores at runtime providing us with *resource adaptivity* as the first adaptivity facility available for the ECLs.

Fig. 25 summarizes all three adaptivity facilities enabled by our prototype. As shown, certain facilities optimize the energy efficiency of the system by reducing the power consumption of the hardware, while others aim at increasing the performance of algorithms by better utilizing the hardware. In the following, we briefly describe the individual adaptivity facilities.

1) Resource adaptivity: This adaptivity facility allows us to control processor and core states (C-states) as well as performance states (P-states) and core frequencies at runtime. Resource adaptivity instantiates an ECL per node (cf., Fig. 24) and uses a workload-dependent EUF to configure the energy-control knobs of the local cores based on the current utilization. Since cores are shared resources that need to be controlled at a low latency, those ECLs are run by the HAEC OS (cf., Section VII-A). However, the OS needs to be supplied with application-specific performance metrics (e.g., query latency in case of a database) to appropriately trade energy for performance.
2) Storage adaptivity: This adaptivity facility addresses the way data are stored in the memory, which is tightly coupled to the algorithm selection (cf., code variants in Section VII-B) for actually processing the data. For instance, data can be stored row-wise, column-wise, or in hybrid versions and are additionally augmented by a set of indexes. Due to the partitioned nature of the underlying data-oriented architecture, we are able to adapt the data layout to the workload at the granularity of data partitions resulting in an ECL instance per partition.
3) Data placement adaptivity: This adaptivity facility addresses the way data objects are partitioned and distributed across the memory domains of the HAEC Box. Since data placement depends on data as well as the workload, it is subject of continuous adaptation. Nevertheless, data placement adaptivity is considered as a long-term measure, because data movements actually consume energy similar to data layout adaptations.

In Fig. 26, we provide results of experiments conducted on a 2-socket Intel Haswell-EP system showing the power consumption of the system, measured using the integrated energy counters (RAPL) during the adaptation phase. In particular, we compare the baseline (no adaptations) to a setting with storage adaptivity enabled and with additionally enabled resource adaptivity during the online creation of an index as a result of a changing workload. The chart shows that both adaptation mechanisms work hand in hand and are able to smoothly decrease the power consumption of the system ending up in energy savings of 75% compared to the baseline. For the reported experiments, we measured a usage of 1% of a single core per socket for the decision making (ECL) while the time to do the actual adaptation is in the submillisecond range for the C/P states (also managed by the OS). Similar measurement campaigns will be carried out using the HAEC-SIM and the playground with more use cases in future work.

D. Energy-Utility Analysis Using Probabilistic Model Checking

Simulation and measure-based methods are the *de facto* standard to evaluate and compare the energy-aware protocols and systems. Formal methods based on stochastic operational models and analytical or numerical methods yield a complementary approach to analyze the tradeoff between energy and utility requirements of protocols and systems and to support design decisions. By representing, e.g., potential system adaptions as nondeterministic choices, formal methods can compute theoretical optimal decision-making strategies. These can be used as guidelines for the design of adaption strategies and for a comparison of efficiently realizable decision-making heuristics against the theoretical optimum. Furthermore, formal methods are well suited to analyze the average long-run behavior, to reveal subtle errors that occur with small probability and to provide tight upper bounds for the occurrence of rare undesired constellations. To illustrate the main ideas of using probabilistic model checking [5] for an

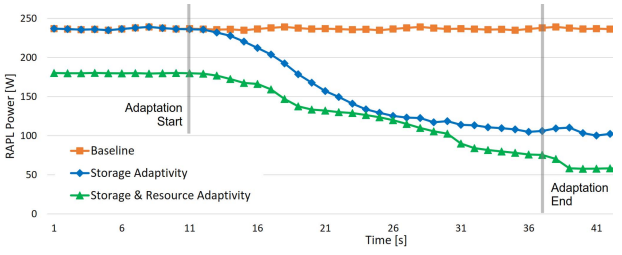

Fig. 26. ***Power measurements (energy counters) for storage and resource adaptivity during an online index creation for workload adaptation.***

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

energy-utility analysis, we regard a simple energy-aware job scheduling scenario, in which we are given a fixed number of processes, successively executing jobs from a job queue. The jobs have deadlines and utility values representing the reward gained by the timely completion of a job or the penalty to be paid for missed deadlines, respectively. The execution of the jobs requires access to a shared resource that can only be used in mutual exclusive manner. The processes can speed up the computation time for the critical sections by activating a turbo mode which is faster than normal mode, but more energy-intensive, say there is 50% reduction of computation time and the increase of the energy requirements is given by factor 3. To determine strategies for granting access of the processes to the shared resource and for the (de)activation of the turbo mode with an optimal energy-utility tradeoff, the operational behavior of the processes and scheduling alternatives can be modeled using a Markov decision process (see, e.g., [97]) where the selection of the process that may access the shared resource next and the switches between normal and turbo mode are represented as nondeterministic alternatives. This Markov decision process can then be analyzed using probabilistic model checking techniques and used to compute optimal strategies with respect to various types of optimality criteria. Standard techniques for Markov decision processes are "single objective" and determine strategies that, e.g., minimize the average energy consumption within a fixed time frame but ignore other quality-of-service measures. Such strategies are likely to lead to low utility values. In contrast, multiobjective reasoning in Markov decision processes [35], [37] allows to determine strategies minimizing, e.g., the average energy consumption with probability at least 0.99 guaranteeing final utility values beyond some given quality threshold. Other forms of multicriterial objectives rely on energy-utility quantiles [6] and determine, e.g., the minimal energy budget required to ensure that the gained utility exceeds some given threshold with probability at least 0.99 and a corresponding strategy. Alternatively, energy-utility quantiles can be used to compute a strategy that maximizes the gained utility for a fixed energy budget. We refer to [6] and [26] for details of the experimental studies with energy-utility quantiles in variants of the above energy-aware job scheduling example as well as the comparison of optimal strategies against standard scheduling strategies. Beside others, these experiments revealed that for a high number of processes the quantile value for a round-robin strategy is considerably higher than for the synthesized optimal strategy, suggesting to employ more sophisticated scheduling strategies in practice. There are further examples where energy-utility quantiles and other probabilistic model checking techniques have been successfully applied to synthesize decision-making strategies. Based on [51], an energy-aware bonding network device protocol for server systems that dynamically adapt to changing bandwidth demands has been considered in [30]. A heterogeneous network system with optical and wireless interconnects similar to the future HAEC Box has been analyzed in [23].

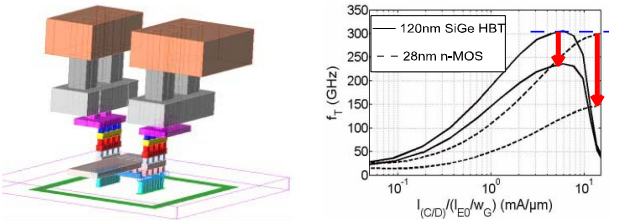

Fig. 27. ***Impact of (a) device connections to other circuit elements on (b) the transit frequency of a SiGeC HBT with 120-nm emitter window width and a MOSFET of the 28-nm node. The upper lines in (b) represent the pad and connection line deembedded transistor data, while the lower lines represent the un-deembedded data, i.e., the transistor with the connections.***

VIII. FUTURE AND EMERGING TECHNOLOGIES

While the presented approach of the HAEC Box as a highly adaptive energy-efficient future computing architecture is based on technologies being available now or in near future, our vision of future computing can become even more elevated when considering future technology generations for terahertz wireless communication and the use of plasmonic waveguides and optical modulators for the optical interconnects. The application of these technologies will be briefly described below.

A. HBTs for THz Near-Range Communication

Operating communication frontend circuits at a certain frequency f_{FE} requires transistors with a two to three times higher power gain cutoff frequency $f_{\max}$ and a reasonably balanced current gain cutoff frequency $f_T (> f_{\max}/(1.4...2))$. For $f_{\mathrm{FE}} = 200$ GHz, even the most advanced RF-CMOS is too slow. In addition, the current drive capability of MOSFETs degrades significantly for more advanced nodes (cf., Fig. 27) due to the increasing impact of external capacitances versus the transistor capacitances. Therefore, the technology candidates for realizing 200-GHz frontends are narrowed down to heterojunction bipolar transistors (HBTs) and high electron mobility transistors (HEMTs). The development of these technologies and their markets is not or only weakly linked to Moore's law due to the very different figures of merit and requirements for high-frequency circuits and systems.

The highest $f_{\max}$ values of about 1.1 THz (at f_T around 0.5....0.6 THz) have so far been achieved with InP/(InGaAs, GaAsSb) HBTs [13], [100], [124] and InP HEMTs [83]. SiGe HBTs already achieve at least the same f_T but with somewhat lower $f_{\max}$ of 0.7 THz [55]. Compared to HEMTs (and MOSFETs), HBTs have various advantages, such as much lower $1/f$ noise and an order of magnitude higher transconductance [108]. SiGe HBTs offer the additional advantage of integration with high

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

density CMOS (even down to 28 nm) [22], thus enabling THz system-on-chip solutions for high volume applications such as communications and radar-based sensing. For details on the latest developments of SiGe HBT technology, the reader is referred to [22] and [99].

Including all known physical and parasitic effects, the ultimate speed limit of SiGe HBTs has been conservatively predicted to be around $(f_{\max}, f_T) = (2, 0.8)$ THz [109], [110] at 22-nm emitter width and 220-nm length, while for long 16-nm InP/InGaAs HBTs $(f_{\max}, f_T) = (4, 2)$ THz [101] has been estimated. These latter values appear somewhat optimistic since simple scaling equations were applied and parasitic capacitances were neglected. Moreover, the narrow emitter width may lead to a very low current gain due to the increased impact of the surface recombination related base current. Nevertheless, such HBT technologies allow more energy-efficient RF frontend components or an increase of the available bandwidth or of the carrier frequency enabling an increase of the achievable data rates.

B. Toward Plasmonic Waveguides and Optical Modulators for On-Chip Optical Interconnects by DNA Nanotechnology

As an information-coding polymer, DNA is an excellent molecular building block to program and construct bio-inspired functional nanostructures with precision and complexity not easily achievable by conventional lithographic techniques [103], [111], [134]. Over the past years, DNA has indeed demonstrated its value as a versatile building material for functional devices, e.g., in nanoelectronics [84] and optics [72], [118]. In an engineering context, the main advantage of employing basic principles of structural DNA nanotechnology in these fields is energy and material efficient as well as massively parallel bottom-up assembly of identical structures. In particular, so-called DNA origami [103] can be considered as "molecular breadboards" to organize nanoparticles into complex functional networks of active components featuring highly spatial and multifunctional positioning of components. This biomimetic approach implies a precise design of materials-encoded particle assemblies with predefined spatial order, and thus, functionality. Here, the attached nano-objects can consist of different material classes, such as, e.g., metals [50], [58], inorganic semiconductors [125], fluorescent dyes [118], conducting polymers [132], or combination of those. Combining these unique features with chemical methods that allow a precise control of nanoparticles shapes and sizes [80], [81], [105], will consequently lead to complex "breadboard"-terminated assemblies, encoded concerning material and well controlled with respect to shapes and sizes of and distance in-between the attached particles. This allows tailoring the physical properties of these assemblies. Recent examples for DNA-based bottom-up assembly of complex optical structures are light-harvesting antenna structures [31], conjugated polymer-based fluorescence intensity shifter [132], chiral plasmonic structures [72], fluorescent dye-based optical [118] as well as gold and/or silver nanoparticle-based plasmonic waveguides [49], [102], [127]. The latter allow to "transport light" within subwavelength structures, which will facilitate a considerable shrinkage of the dimensions of opto–electronic devices. Moreover, gold nanorods have been used to build reconfigurable 3-D plasmonic nanostructures [133]. Gold nanorod assemblies have also allowed to shift the resonance frequency of antenna structures, such as, e.g., nanofabricated dipol or Yagi-Uda antennas, into the optical and visible wavelength range, which enables the conversion of light from free space into subwavelength volumes and *vice versa* [25], [29], [78]. These few examples show that DNA-based nanostructure assemblies already provide a solid platform for the investigation and application of nanofabricated on-chip optical interconnects and modulators [4], [131].

We are considering DNA nanotechnology to manufacture all-optical modulators and switches allowing to switch or control optical signals by optical signals. While in the HAEC Box the switching of optical signals for signal routing between the compute nodes so far is planned to be based on electrical signals, the use of all-optical switches would enable to substitute the electrical by an optical control signal. Envisioning that in future even computing can be performed optically instead of electrically, this is one further step toward an all optical computer, which is expected to provide a higher energy efficiency per computing operation than electrical computing.

IX. CONCLUSION AND DISCUSSION

In this paper, we presented an architecture of a highly adaptive energy-efficient computing platform. As future computer architectures will have an enormous number of compute cores, the design of such systems leads to major changes compared to today's systems. The computing principle will change to in-memory computing, where the computing is placed close to the data to be processed. For this reason and also the restrictions on heat dissipation, the individual cores will be active for only a fraction of the time. Thus, the individual cores must be small compared to their local memory segment. However, the requirement of accessing data across memory segments is not eliminated by in-memory computing. Hence, the communication of data between the individual compute cores becomes a key issue. This holds true for 1) on-chip; 2) intrachip within the 3-D chip stack; and 3) between the different chip stacks (compute nodes) in a rack. Especially focusing on the last problem, we proposed a new communication interconnect using rate adaptive optical links to connect chip stacks on the same circuit board and flexible beam-switched wireless links for communication between neighboring circuit boards. This enables a petabit backplane. The interfaces to the optical and wireless communication links are an

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

Table 2 Power Consumption of Computing and Communication

component	energy per bit pJ/b	power per link/proc. mW	number of links/ processors	activity	total power consump. W
optical links (25 Gb/s) (Sec. III-B)	4.5	112.5 10% at standby	7680	10%	165
wireless links (50 Gb/s) one polarization, (Sec. IV-G)	48.22	2411	768	10%	185
processors (Sec. II-C)		0.15 0.1% at standby	10^8	5%	765
overall					1115

integral part of the chip stacks and use novel principles, e.g., driver circuits for rate adaptive optical links and oversampled 1-bit quantization. We describe this concept based on the HAEC Box, a one liter implementation. To exploit the capabilities of this highly adaptive computing hardware we introduce a sophisticated runtime and networking management system. This system manages the load of the cores, the placement of data in memory, and the network activity based on a tight monitoring of the system states. To be able to study the system behavior, we developed a simulation framework and a hardware demonstrator, the HAEC playground, allowing to evaluate the effect of compute task distribution on network load distribution and energy consumption.

Having in mind the enormous amount of compute cores in a volume of just one liter, one key challenge is heat dissipation. In previous sections, we discussed the energy consumption of the individual parts of the HAEC Box, i.e., computing (Section II-C), optical links (Section III-B), and wireless links (Section IV-G), which are summarized in Table 2.

Our dark silicon assumption leads to a low activity of the compute nodes and consequently also of the communication links. For the compute part, we thus assume an activity level of the processors of 5%, yielding 765 W of total power consumption.

For the optical links we achieved a power consumption of 112.5 mW per waveguide at 25 Gb/s (4.5 pJ/b), which is assumed to reduce to 10% in standby mode. With the structure of the optical interconnect per PCB side given in Figs. 2(c) and 4 PCBs in the HAEC Box, this leads to 256 optical connections. Two neighboring nodes are connected by 30 waveguides yielding 7680 optical links in total. Given an activity ratio of 10%, this results in 165 W of power consumption for the optical part.

Regarding the wireless link, we have estimated an energy consumption of 48.22 pJ/b, leading to around 2.4 W per unidirectional 50-Gb/s communication link using one polarization domain. Assuming that per node 8 simultaneous bidirectional links can be established, this results in 768 links. With an activity level of 10%, the wireless communication links consume ca. 185 W.

In total, this results in a power consumption of the entire HAEC Box of round about 1 kW which is dissipated in a volume of one liter. Using typical active cooling by air with a blower, this heat can be dissipated showing that the assumptions on the power budget of the HAEC Box are reasonable. Using future and emerging technologies like plasmonic waveguides and optical modulators manufactured using DNA nanotechnology, and new HBTs for THz communication the energy efficiency is expected to increase even further.

The HAEC Box is a vision which we can expect to become reality around 2035. We are developing base technologies for components and less performant variants of the HAEC Box. However, with its small size of just one liter the HAEC Box concept opens new possibilities. It allows to bring servers closer to the user, e.g., by combining them with a base station of a cellular network or a Wi-Fi hotspot. Thus, it allows to reduce network traffic loads and energy consumption significantly and hence enables the mobile edge cloud [94]. ■

Acknowledgments

The authors would like to thank the following members of the Collaborative Research Center HAEC and of the Center for Advancing Electronics Dresden (cfaed) for contributing to the work reported in this paper: S. Bender, M. Bielert, J. Cabrera, C. Carta, P. Chrszon, M. Daum, C. Dubslaff, F. Fischer, E. Franz, D. Fritsche, A. Goens, F. Nadi Gür, M. Haehnel, N. ul Hassan, A. Heerwig, R. Henker, A. Henning-Knechtel, G. Hielscher, S. Hosseini, T. Ilsche, C. Jans, M. Jazayerifar, M. Jenning, R. Khasanov, A. Kick, A. Kiriy, B. Klein, J. Klein, S. Klüppelholz, T. König, M. Lakatos, L. Landau, S. Lüngen, S. Moriam, T. Nardmann, N. Neumann, K. Nieweglowski, T. D. A. Nguyen, S. Rehmann, D. Schöniger, M. Schlüter, T. Schmidt, P. Seiler, T. Smejkal, T. Tiedje, and J. Zessin.

REFERENCES

[1] A. Abbas *et al.*, "A survey on energy-efficient methodologies and architectures of network-on-chip," *Comput. Elect. Eng.*, vol. 40, no. 8, pp. 333–347, Nov. 2014.
[2] *Altera*. [Online]. Available: https://www.altera.com/products/fpga/stratix-series/stratix-v/overview.html
[3] O. Arnold, E. Matus, B. Noethen, M. Winter, T. Limberg, and G. Fettweis, "Tomahawk: Parallelism and heterogeneity in communications signal processing MPSoCs," *ACM Trans. Embedded Comput. Syst.*, vol. 13, no. 3s, p. 107, 2014.
[4] V. E. Babicheva, I. V. Kulkova, R. Malureanu, K. Yvind, and A. V. Lavrinenko, "Plasmonic modulator based on gain-assisted metal–semiconductor–metal waveguide," *Photon. Nanostruct.-Fundam. Appl.*, vol. 10, no. 4, pp. 389–399, 2012.
[5] C. Baier, L. De Alfaro, M. Forejt, and V. C. Kwiatkowska, "Model checking probabilistic systems," in *Handbook of Model Checking*, 2018.
[6] C. Baier, M. Daum, C. Dubslaff, J. Klein, and S. Klüppelholz, "Energy-utility quantiles," in *Proc. 6th NASA Formal Methods Symp. (NFM)* (Lecture

Notes in Computer Science), vol. 8430. 2014, pp. 285–299.

[7] S. Bender, M. Dörpinghaus, and G. Fettweis, "On the achievable rate of bandlimited continuous-time 1-bit quantized AWGN channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Aachen, Germany, Jun. 2017, pp. 2083–2087.

[8] S. Bender, L. Landau, M. Dörpinghaus, and G. Fettweis, "Communication with 1-bit quantization and oversampling at the receiver: Spectral constrained waveform optimization," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Edinburgh, U.K., Jul. 2016, pp. 1–5.

[9] L. Benini and G. De Micheli, "Networks on chips: A new SoC paradigm," *Computer*, vol. 35, no. 1, pp. 70–78, Jan. 2002.

[10] L. Benini, A. Bogliolo, and G. De Micheli, "A survey of design techniques for system-level dynamic power management," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 3, pp. 299–316, Jun. 2000.

[11] M. Bielert, F. M. Ciorba, K. Feldhoff, T. Ilsche, and W. E. Nagel, "HAEC-SIM: A simulation framework for highly adaptive energy-efficient computing platforms," in *Proc. Conf. Simulation Tools Techn.*, 2015.

[12] A. Boletti, D. Giacomuzzi, G. Parladori, P. Boffi, M. Ferrario, and M. Martinelli, "Performance comparison between electrical copper-based and optical fiber-based backplanes," *Opt. Express*, vol. 21, no. 16, pp. 19202–19208, 2013.

[13] C. R. Bolognesi, R. Flückiger, M. Alexandrova, W. Quan, R. Lövblom, and O. Ostinelli, "InP/GaAsSb DHBTs for THz applications and improved extraction of their cutoff frequencies," in *IEDM Tech. Dig.*, Dec. 2016, pp. 723–726.

[14] S. Borkar and A. A. Chien, "The future of microprocessors," *Commun. ACM*, vol. 54, no. 5, pp. 67–77, May 2011.

[15] J. Butler and R. Lowe, "Beam-forming matrix simplifies design of electrically scanned antennas," *Electron. Des.*, vol. 9, Apr. 1961.

[16] J. A. Cabrera Guerrero, D. E. Lucani Roetter, and F. H. P. Fitzek, "On network coded distributed storage: How to repair in a fog of unreliable peers," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2016, pp. 188–193.

[17] J. Castrillon and R. Leupers, *Programming Heterogeneous MPSoCs: Tool Flows to Close the Software Productivity Gap*. Springer, 2014.

[18] J. Castrillon, R. Leupers, and G. Ascheid, "MAPS: Mapping concurrent dataflow applications to heterogeneous MPSoCs," *IEEE Trans. Ind. Informat.*, vol. 9, no. 1, pp. 527–545, Feb. 2013.

[19] J. Castrillon *et al.*, "A hardware/software stack for heterogeneous systems," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 4, no. 3, pp. 243–259, Jul./Sep. 2018.

[20] M. Catuneanu *et al.*, "Design considerations and constraints in silicon based optical on-board interconnects for short range communications," in *Proc. 19th Int. Conf. Transp. Opt. Netw. (ICTON)*, Jul. 2017, pp. 1–4.

[21] H. Cheng *et al.*, "Optics vs. copper—From the perspective of Thunderbolt 3 interconnect technology," in *Proc. China Semiconductor Technol. Int. Conf. (CSTIC)*, Mar. 2016, pp. 1–3.

[22] P. Chevalier *et al.*, "Si/SiGe:C and InP/GaAsSb heterojunction bipolar transistors for THz applications," *Proc. IEEE*, vol. 105, no. 6, pp. 1035–1050, Jun. 2017.

[23] P. Chrszon, C. Dubslaff, S. Klüppelholz, and C. Baier, "ProFeat: feature-oriented engineering for family-based probabilistic model checking," *Formal Aspects Comput.*, vol. 30, no. 1, pp. 45–75, Jan. 2018.

[24] G. Cox and A. Bhattacharjee, "Efficient address translation for architectures with multiple page sizes," in *Proc. 22nd Int. Conf. Archit. Support Programm. Lang. Oper. Syst.*, 2017, pp. 435–448.

[25] A. G. Curto, G. Volpe, T. H. Taminiau, M. P. Kreuzer, R. Quidant, and N. F. van Hulst, "Unidirectional emission of a quantum dot coupled to a nanoantenna," *Science*, vol. 329, no. 5994, pp. 930–933, 2010.

[26] M. Daum, Ph.D. dissertation, Dept. Comput. Sci., TU Dresden, Dresden, Germany, 2018.

[27] J. B. Dennis, "First version of a data flow procedure language," in *Proc. Programm. Symp.* Springer, 1974, pp. 362–376.

[28] F. E. Doany *et al.*, "Terabit/sec VCSEL-based 48-channel optical module based on holey CMOS transceiver IC," *J. Lightw. Technol.*, vol. 31, no. 4, pp. 672–680, Feb. 15, 2013.

[29] J. Dorfmüller, R. Vogelgesang, W. Khunsin, C. Rockstuhl, C. Etrich, and K. Kern, "Plasmonic nanowire antennas: Experiment, simulation, and theory," *Nano Lett.*, vol. 10, no. 9, pp. 3596–3603, 2010.

[30] C. Dubslaff, C. Baier, and S. Klüppelholz, "Probabilistic model checking for feature-oriented systems," in *Transactions on Aspect-Oriented Software Development XII*, vol. 12. 2015, pp. 180–220.

[31] P. K. Dutta, R. Varghese, J. Nangreave, S. Lin, H. Yan, and Y. Liu, "DNA-directed artificial light-harvesting antenna," *J. Amer. Chem. Soc.*, vol. 133, no. 31, pp. 11985–11993, 2011.

[32] J. Eker and J. W. Janneck, "CAL language report specification of the CAL actor language," Dept. Elect. Eng. Comput. Sci., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. UCB/ERL M03/48, 2003.

[33] I. M. Elfadel and G. Fettweis, Eds., *3D Stacked Chips*. Springer, 2016.

[34] H. Esmaeilzadeh, E. Blem, R. St Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," in *Proc. 38th Annu. Int. Symp. Comput. Archit. (ISCA)*. New York, NY, USA: ACM, 2011, pp. 365–376.

[35] K. Etessami, M. Z. Kwiatkowska, M. Y. Vardi, and M. Yannakakis, "Multi-objective model checking of Markov decision processes," in *Logical Methods Comput. Sci.*, vol. 4, no. 4, pp. 1–21, 2008.

[36] F. Färber, S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg, and W. Lehner, "SAP HANA database: Data management for modern business applications," *ACM SIGMOD Rec.*, vol. 40, no. 4, pp. 45–51, Jan. 2012.

[37] V. E. Forejt, M. Kwiatkowska, G. Norman, and D. Parker, "Automated verification techniques for probabilistic systems," in *Proc. 11th Int. School Formal Methods Design Comput., Commun. Softw. Syst. (SFM)* (Lecture Notes in Computer Science), vol. 6659. Springer, 2011, pp. 53–113.

[38] T. D. Forum. (2014). *Enabling Symmetric Multiprocessing for Embedded Linux on ARC Processor Cores*. [Online]. Available: http://www.techdesignforums.com/practice/technique/symmetric-multiprocessing-embedded-linux-arc-processor-cores/

[39] D. Fritsche, C. Carta, and F. Ellinger, "A broadband 200 GHz amplifier with 17 dB gain and 18 mW DC-power consumption in 0.13 μm SiGe BiCMOS," *IEEE Microw. Wireless Compon. Lett.*, vol. 24, no. 11, pp. 790–792, Nov. 2014.

[40] D. Fritsche, J. D. Leufker, G. Tretter, C. Carta, and F. Ellinger, "A low-power broadband 200 GHz down-conversion mixer with integrated LO-driver in 0.13 μm SiGe BiCMOS," *IEEE Microw. Wireless Compon. Lett.*, vol. 25, no. 9, pp. 594–596, Sep. 2015.

[41] D. Fritsche, G. Tretter, P. Stärke, C. Carta, and F. Ellinger, "A low-power SiGe BiCMOS 190-GHz receiver with 47-dB conversion gain and 11-dB noise figure for ultralarge-bandwidth applications," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 10, pp. 4002–4013, Oct. 2017.

[42] D. Fritsche, P. Stärke, C. Carta, and F. Ellinger, "A low-power SiGe BiCMOS 190-GHz transceiver chipset with demonstrated data rates up to 50 Gbit/s using on-chip antennas," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 9, pp. 3312–3323, Sep. 2017.

[43] E. N. Gilbert, "Increased information rate by oversampling," *IEEE Trans. Inf. Theory*, vol. 39, no. 6, pp. 1973–1976, Nov. 1993.

[44] C. J. Glass and L. M. Ni, "The turn model for adaptive routing," *Assoc. Comput. Mach.*, vol. 41, no. 5, pp. 874–902, Sep. 1994.

[45] A. Goens, S. Siccha, and J. Castrillon, "Symmetry in software synthesis," *ACM Trans. Archit. Code Optim.*, vol. 14, no. 2, pp. 20:1–20:26, Jul. 2017.

[46] A. Goens, R. Khasanov, M. Hähnel, T. Smejkal, H. Härtig, and J. Castrillon, "TETRiS: A multi-application run-time system for predictable execution of static mappings," in *Proc. 20th Int. Workshop Softw. Compil. Embedded Syst. (SCOPES)*. New York, NY, USA: ACM, Jun. 2017, pp. 11–20.

[47] S. Götz, C. Wilke, S. Richly, G. Püschel, and U. Assmann, "Modeldriven self-optimization using integer linear programming and pseudoboolean optimization," in *Proc. ADAPTIVE*, 2013.

[48] B. Grot, J. Hestness, S. W. Keckler, and O. Mutlu, "Kilo-NOC: A heterogeneous network-on-chip architecture for scalability and service guarantees," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 3, pp. 401–412, Jun. 2011.

[49] F. N. Gür *et al.* (2017). "Self-assembled plasmonic waveguides for excitation of fluorescent nanodiamonds." [Online]. Available: https://arxiv.org/1712.09141

[50] F. N. Gür, F. W. Schwarz, J. Ye, S. Diez, and T. L. Schmidt, "Toward self-assembled plasmonic devices: High-yield arrangement of gold nanoparticles on DNA origami templates," *ACS Nano*, vol. 10, no. 5, pp. 5374–5382, 2016.

[51] M. Hähnel, B. Döbel, M. Völp, and H. Härtig, "eBond: Energy saving in heterogeneous R.A.I.N," in *Proc. 4th Int. Conf. Future Energy Syst.*, 2013, pp. 193–202.

[52] M. Háhnel, F. M. Arega, W. Dargie, R. Khasanov, and J. Castrillon, "Application interference analysis: Towards energy-efficient workload management on heterogeneous micro-server architectures," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, May 2017, pp. 432–437.

[53] H. Härtig, M. Völp, and M. Hähnel, "The case for practical multi-resource and multi-level scheduling based on energy/utility," in *Proc. IEEE 19th Int. Conf. Embedded Real-Time Comput. Syst. Appl. (RTCSA)*, Aug. 2013, pp. 175–182.

[54] N. ul Hassan, M. Lentmaier, and G. P. Fettweis, "Comparison of LDPC block and LDPC convolutional codes based on their decoding latency," in *Proc. IEEE Int. Symp. Turbo Codes Iterative Inf. Process.*, Gothenburg, Sweden, Aug. 2012, pp. 225–229.

[55] B. Heinemann *et al.*, "SiGe HBT with fx/fmax of 505 GHz/720 GHz," in *IEDM Tech. Dig.*, Dec. 2016, pp. 51–54.

[56] J. Henkel, H. Khdr, S. Pagani, and M. Shafique, "New trends in dark silicon," in *Proc. 52nd Annu. Design Autom. Conf. (DAC)*. New York, NY, USA: ACM, Jun. 2015, p. 119:1–119:6.

[57] R. Henker *et al.*, "Tunable broadband integrated circuits for adaptive optical interconnects," in *Proc. 17th Conf. Opt. Fibres Appl.*, 2017, p. 103250P.

[58] A. Henning-Knechtel *et al.*, "Dielectrophoresis of gold nanoparticles conjugated to DNA origami structures," *Beilstein J. Nanotechnol.*, vol. 7, pp. 948–956, Jul. 2016.

[59] S. Hosseini, L. Mirzoyan, and K. Jamshidi, "Energy consumption enhancement of reverse-biased silicon-based Mach–Zehnder modulators using corrugated slow light waveguides," *IEEE Photon. J.*, vol. 10, no. 1, Feb. 2018, Art. no. 8200207.

[60] S. Hosseini and K. Jamshidi, "Fundamental peformance tradeoffs for reverse biased free carrier plasma dispersion effect based silicon optical modulators," in *Proc. Int. Conf. Photon. Switching (PS)*, Sep. 2015, pp. 196–198.

[61] (Aug. 27, 2018). *IBTA Infiniband Roadmap*. [Online]. Available: https://www.infinibandta.org/infiniband-roadmap/

[62] K. A. Schouhamer Immink, "Runlength-limited sequences," *Proc. IEEE*, vol. 78, no. 11, pp. 1745–1759, Nov. 1990.

[63] A. Jimenez Felström and K. Zigangirov, "Time-varying periodic convolutional codes with

low-density parity-check matrix," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 2181–2191, Sep. 1999.

[64] J. A. Kash *et al.*, "Optical interconnects in exascale supercomputers," in *Proc. Annu. Meeting IEEE Photon. Soc.*, Nov. 2010, pp. 483–484.

[65] R. Khasanov, A. Goens, and J. Castrillon, "Implicit data-parallelism in Kahn process networks: Bridging the MacQueen Gap," in *Proc. 9th Workshop Parallel Programm. Run-Time Manage. Techn. Many-Core Archit. (PARMA-DITAM), 13th Int. Conf. High-Perform. Embedded Archit. Compilers (HiPEAC)*, Jan. 2018, pp. 20–25.

[66] T. Kissinger *et al.*, "ERIS: A NUMA-aware in-memory storage engine for analytical workloads," in *Proc. ADMS*, 2014, pp. 1–12.

[67] B. Klein, M. Jenning, P. Seiler, and D. Plettemeier, "Wideband half-cloverleaf shaped on-chip antenna for 160 GHz–200 GHz applications," in *Proc. IEEE Antennas Propag. Soc. Int. Symp. (APSURSI)*, Jul. 2014, pp. 1448–1449.

[68] A. Knüpfer *et al.*, "The Vampir performance analysis tool-set," in *Tools for High Performance Computing*, M. Resch, R. Keller, V. Himmler, B. Krammer, and A. Schulz, Eds. Springer, Jul. 2008, pp. 139–155.

[69] A. Knüpfer *et al.*, "Score-P: A joint performance measurement run-time infrastructure for periscope, Scalasca, TAU, and Vampir," in *Tools for High Performance Computing*, H. Brunst, M. S. Müller, W. E. Nagel, and M. M. Resch, Eds. Springer, 2011, pp. 79–91.

[70] T. Koch and A. Lapidoth, "Increased capacity per unit-cost by oversampling," in *Proc. 26th IEEE Conv. Elect. Electron. Eng. Israel (IEEEI)*, Eilat, Israel, Nov. 2010, pp. 684–688.

[71] S. Kudekar, T. J. Richardson, and R. L. Urbanke, "Spatially coupled ensembles universally achieve capacity under belief propagation," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 7761–7813, Dec. 2013.

[72] A. Kuzyk *et al.*, "DNA-based self-assembly of chiral plasmonic nanostructures with tailored optical response," *Nature*, vol. 483, no. 7389, pp. 311–314, 2012.

[73] L. Landau, M. Dörpinghaus, and G. P. Fettweis, "1-bit quantization and oversampling at the receiver: Communication over bandlimited channels with noise," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1007–1010, May 2017.

[74] L. Landau and G. Fettweis, "Information rates employing 1-bit quantization and oversampling at the receiver," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Toronto, ON, Canada, Jun. 2014, pp. 219–223.

[75] L. Landau, M. Dörpinghaus, and G. Fettweis, "Communications employing 1-bit quantization and oversampling at the receiver: Faster-than-Nyquist signaling and sequence design," in *Proc. IEEE Int. Conf. Ubiquitous Wireless Broadband (ICUWB)*, 2015, pp. 1–5.

[76] R. Lucas *et al.*, "Top ten exascale research challenges," DOE ASCAC Subcommittee Rep., 2014, pp. 1–86.

[77] S. Lüngen *et al.*, "3D optical coupling techniques on polymer waveguides for wafer and board level integration," in *Proc. IEEE 67th Electron. Compon. Technol. Conf. (ECTC)*, May/Jun. 2017, pp. 1612–1618.

[78] I. S. Maksymov, I. Staude, A. E. Miroshnichenko, and Y. S. Kivshar, "Optical Yagi-Uda nanoantennas," *Nanophotonics*, vol. 1, no. 1, pp. 65–81, 2012.

[79] B. Matthiesen *et al.*, "Secure and energy-efficient interconnects for board-to-board communication," in *Proc. IEEE Int. Conf. Ubiquitous Wireless Broadband (ICUWB)*, Salamanca, Spain, Sep. 2017, p. 7.

[80] M. Mayer *et al.*, "Aqueous gold overgrowth of silver nanoparticles: Merging the plasmonic properties of silver with the functionality of gold," *Angew. Chem. Int. Ed.*, vol. 56, no. 50, pp. 15866–15870, 2017.

[81] M. Mayer *et al.*, "Controlled living nanowire growth: Precise control over the morphology and optical properties of AgAuAg bimetallic nanowires," *Nano Lett.*, vol. 15, no. 8, pp. 5427–5437, 2015.

[82] J. E. Mazo, "Faster-than-Nyquist signaling," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1451–1462, 1975.

[83] X. Mei *et al.*, "First demonstration of amplification at 1 THz using 25-nm InP high electron mobility transistor process," *IEEE Electron Device Lett.*, vol. 36, no. 4, pp. 327–329, Apr. 2015.

[84] M. Mertig and W. Pompe, "Biomimetic fabrication of DNA-based metallic nanowires and networks," in *Nanobiotechnology: Concepts, Applications and Perspectives*. 2004, pp. 256–277.

[85] A. Mezghani and J. A. Nossek, "Power efficiency in communication systems from a circuit perspective," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2011, pp. 1896–1899.

[86] D. A. B. Miller, "Device requirements for optical interconnects to silicon chips," *Proc. IEEE*, vol. 97, no. 7, pp. 1166–1185, Jul. 2009.

[87] S. Mittal, J. S. Vetter, and D. Li, "A survey of architectural approaches for managing embedded DRAM and non-volatile on-chip caches," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 6, pp. 1524–1537, Jun. 2015.

[88] H. Moody, "The systematic design of the Butler matrix," *IEEE Trans. Antennas Propag.*, vol. 12, no. 6, pp. 786–788, Nov. 1964.

[89] S. Moriam and G. P. Fettweis, "Fault tolerant deadlock-free adaptive routing algorithms for hexagonal networks-on-chip," in *Proc. 19th Euromicro Conf. Digit. Syst. Design (DSD)*, Aug./Sep. 2016, pp. 131–137.

[90] S. Moriam and G. P. Fettweis, "Reliability assessment of fault tolerant routing algorithms in networks-on-chip: An analytic approach," in *Proc. Conf. Design, Autom. Test Eur. (DATE)*, Mar. 2017, pp. 61–66.

[91] K. Nieweglowski and K. Bock, "Assembly of optical transceivers for board-level optical interconnects," *Proc. SPIE*, vol. 9888, pp. 0S:1–0S:10, Apr. 2016.

[92] K. Nieweglowski, T. Tiedje, D. Schöniger, R. Henker, F. Ellinger, and K. Bock, "Electro-optical integration for VCSEL-based board-level optical chip-to-chip communication," *Proc. SPIE*, vol. 10325, p. 103250V, Feb. 2017.

[93] I. Pandis, R. Johnson, N. Hardavellas, and A. Ailamaki, "Data-oriented transaction execution," in *Proc. VLDB Endowment*, vol. 3, nos. 1–2, pp. 928–939, Sep. 2010.

[94] M. Patel *et al.*, "*Mobile-edge computing introductory technical white paper*," Mobile-Edge Computing (MEC) Industry Initiative, White Paper, 2014.

[95] S. Pfennig and E. Franz, "eSPOC: Enhanced secure practical network coding for better efficiency and lower latency," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[96] D. Porobic, E. Liarou, P. Tözün, and A. Ailamaki, "ATraPos: Adaptive transaction processing on hardware Islands," in *Proc. IEEE 30th Int. Conf. Data Eng.*, Chicago, IL, USA Mar./Apr. 2014, pp. 688–699.

[97] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 1994.

[98] M. Radetzki, C. Feng, X. Zhao, and A. Jantsch, "Methods for fault tolerance in networks-on-chip," *ACM Comput. Surv.*, vol. 46, no. 1, Oct. 2013, Art. no. 8.

[99] N. Rinaldi and M. Schroter, Eds., *Silicon-Germanium Heterojunction Bipolar Transistors for mm-Wave Systems Technology, Modeling and Circuit Applications*. Amsterdam, The Netherlands: River Publishers, 2018.

[100] J. C. Rode, H.-W. Chiang, P. Choudhary, V. Jain, B. J. Thibeault, W. J. Mitchell, M. J. W. Rodwell, M. Urteaga, D. Loubychev, A. Snyder, Y. Wu, J. M. Fastenau, and A. W. K. Liu, "Indium phosphide heterobipolar transistor technology beyond 1-THz bandwidth," *IEEE Trans. Electron Devices*, vol. 62, no. 9, pp. 2779–2785, Sep. 2015.

[101] M. Rodwell, "III-V HBT and (MOS) HEMT scaling," Tutorial (WSG), Tech. Rep., 2017.

[102] E.-M. Roller, L. V. Besteiro, C. Pupp, L. K. Khorashad, A. O. Govorov, and T. Liedl, "Hotspot-mediated non-dissipative and ultrafast plasmon passage," *Nature Phys.*, vol. 13, pp. 761–765, May 2017.

[103] P. W. K. Rothemund, "Folding DNA to create nanoscale shapes and patterns," *Nature*, vol. 440, no. 7082, pp. 297–302, 2006.

[104] M. Schlüter, N. U. Hassan, and G. P. Fettweis, "On the construction of protograph based SC-LDPC codes for windowed decoding," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Barcelona, Spain, Apr. 2018, pp. 1–6.

[105] M. J. Schnepf *et al.*, "Nanorattles with tailored electric field enhancement," *Nanoscale*, vol. 9, no. 27, pp. 9376–9385, 2017.

[106] D. Schoeniger, R. Henker, and F. Ellinger, "An 850-nm common-cathode VCSEL driver with tunable energy efficiency for 45 Gbit/s data transmission without equalization," in *Proc. IEEE Asia Pacific Microw. Conf. (APMC)*, Nov. 2017, pp. 1103–1106.

[107] D. Schoeniger, R. Henker, and F. Ellinger, "High-speed transimpedance amplifier with runtime adaptive bandwidth and power consumption in 0.13 μm SiGe BiCMOS," *Electron. Lett.*, vol. 52, no. 2, pp. 154–156, Jan. 2016.

[108] M. Schröter and A. Pawlak, "SiGe heterojunction bipolar transistor technology for sub-mm-wave electronics—State-of-the-art and future prospects," in *Proc. SiRF*, Jan. 2018, pp. 60–63.

[109] M. Schröter *et al.*, "Physical and electrical performance limits of high-speed SiGeC HBTs—Part I: Vertical scaling," *IEEE Trans. Electron Devices*, vol. 58, no. 11, pp. 3687–3706, Nov. 2011.

[110] M. Schröter *et al.*, "SiGe HBT technology: Future trends and TCAD-based roadmap," *Proc. IEEE*, vol. 105, no. 6, pp. 1068–1086, Jun. 2017.

[111] N. C. Seeman and N. R. Kallenbach, "Design of immobile nucleic acid junctions," *Biophys. J.*, vol. 44, no. 2, pp. 201–209, 1983.

[112] M. Shafique and S. Garg, "Computing in the dark silicon era: Current trends and research challenges," *IEEE Des. Test.*, vol. 34, no. 2, pp. 8–23, Apr. 2017.

[113] S. Shamai, "Information rates by oversampling the sign of a bandlimited process," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1230–1236, Jul. 1994.

[114] Silexica. *Silexica Analysis Tools*. [Online]. Available: http://www.silexica.com

[115] S. Sriram and S. S. Bhattacharyya, *Embedded Multiprocessors: Scheduling and Synchronization*, 2nd ed. New York, NY, USA: Marcel Dekker, 2009.

[116] P. Stärke, D. Fritsche, S. Schumann, C. Carta, and F. Ellinger, "High-efficiency wideband 3-D on-chip antennas for subterahertz applications demonstrated at 200 GHz," *IEEE Trans. THz Sci. Technol.*, vol. 7, no. 4, pp. 415–423, Jul. 2017.

[117] R. B. Staszewski, "Digitally intensive wireless transceivers," *IEEE Des. Test. Comput.*, vol. 29, no. 6, pp. 7–18, Dec. 2012.

[118] I. H. Stein, C. Steinhauer, and P. Tinnefeld, "Single-molecule four-color FRET visualizes energy-transfer paths on DNA origami," *J. Amer. Chem. Soc.*, vol. 133, no. 12, pp. 4193–4195, 2011.

[119] M. A. Taubenblatt, "Optical interconnects for high-performance computing," *J. Lightw. Technol.*, vol. 30, no. 4, pp. 448–457, Feb. 15, 2012.

[120] W. Thies, M. Karczmarek, and S. Amarasinghe, "StreamIt: A language for streaming applications," in *Proc. 11th Int. Conf. Compiler Construction (CC)*. London, U.K.: Springer-Verlag, 2002, pp. 179–196.

[121] C. A. Thraskias *et al.*, "Survey of photonic and plasmonic interconnect technologies for intra-datacenter and high-performance computing communications," *IEEE Commun. Surveys Tuts.*, to be published.

[122] G. Tretter, M. M. Khafaji, D. Fritsche, C. Carta, and F. Ellinger, "Design and characterization of a 3-bit 24-GS/s flash ADC in 28-nm low-power digital CMOS," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 4, pp. 1143–1152, Apr. 2016.

[123] N. U. Hassan, A. E. Pusane, M. Lentmaier, G. P. Fettweis, and D. J. Costello, "Non-uniform

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

window decoding schedules for spatially coupled LDPC codes," *IEEE Trans. Commun.*, vol. 65, no. 2, pp. 501–510, Feb. 2017.

[124] M. Urteaga, Z. Griffith, M. Seo, J. Hacker, and M. J. W. Rodwell, "InP HBT technologies for THz integrated circuits," *Proc. IEEE*, vol. 105, no. 6, pp. 1051–1067, Jun. 2017.

[125] R. Weichelt *et al.*, "Methods to characterize the oligonucleotide functionalization of quantum dots," *Small*, vol. 12, no. 34, pp. 4763–4771, Sep. 2016.

[126] A. Weissel and F. Bellosa, "Process cruise control-event-driven clock scaling for dynamic power management," in *Proc. Int. Conf. Compilers, Archit. Synth. Embedded Syst. (CASES)*, Grenoble, France, Oct. 2002. [Online]. Available: http://i30www.ira.uka.de/research/publications/

[127] B. Willingham and S. Link, "Energy transport in metal nanoparticle chains via sub-radiant plasmon modes," *Opt. Express*, vol. 19, no. 7, pp. 6450–6461, 2011.

[128] S. Wunderlich, J. A. Cabrera, F. H. P. Fitzek, and M. Reisslein, "Network coding in heterogeneous multicore IoT nodes with DAG scheduling of parallel matrix block operations," *IEEE Internet Things J.*, vol. 4, no. 4, pp. 917–933, Aug. 2017.

[129] S. Wunderlich, J. A. Cabrera, F. H. P. Fitzek, and M. V. Pedersen, "Network coding parallelization based on matrix operations for multicore architectures," in *Proc. Int. Conf. Ubiquitous Wireless Broadband (ICUWB)*, Oct. 2015, pp. 1–5.

[130] S. Wunderlich, J. Cabrera, F. H. P. Fitzek, and M. V. Pedersen, "Network coding parallelization based on matrix operations for multicore architectures," in *Proc. Int. Conf. Ubiquitous Wireless Broadband (ICUWB)*, Oct. 2015, pp. 1–5.

[131] Y. Yao *et al.*, "Broad electrical tuning of graphene-loaded plasmonic antennas," *Nano Lett.*, vol. 13, no. 3, pp. 1257–1264, Feb. 2013.

[132] J. Zessin *et al.*, "Tunable fluorescence of a semiconducting polythiophene positioned on DNA origami," *Nano Lett.*, vol. 17, no. 8, pp. 5163–5170, 2017.

[133] P. Zhan *et al.*, "Reconfigurable three-dimensional gold nanorod plasmonic nanostructures organized on DNA origami tripod," *ACS Nano*, vol. 11, no. 2, pp. 1172–1179, 2017.

[134] F. Zhang and H. Yan, "DNA self-assembly scaled up," *Nature*, vol. 552, no. 7683, pp. 34–35, 2017.

ABOUT THE AUTHORS

Gerhard P. Fettweis (Fellow, IEEE) received the Ph.D. degree under H. Meyr's supervision from RWTH Aachen, Aachen, Germany, in 1990.

After one year at IBM Research in San Jose, CA, USA, he moved to TCSI Inc., Berkeley, CA, USA. He has been Vodafone Chair Professor at the Technische Universität Dresden (TU Dresden), Dresden, Germany, since 1994, and has been head of the Barkhausen Institute since 2018. He coordinates the 5G Lab Germany, and 2 German Science Foundation (DFG) centers at TU Dresden, namely, the Center for Advancing Electronics Dresden (cfaed) and Highly Adaptive Energy Efficient Computing (HAEC). His research focuses on wireless transmission and chip design for wireless/IoT platforms, with 20 companies from Asia/Europe/United States sponsoring his research. In Dresden, his team has spun out 16 startups, and setup funded projects in volume of close to C1/2 billion.

Prof. Fettweis is a member of the German Academy of Sciences (Leopoldina), the German Academy of Engineering (acatech), and received multiple IEEE recognitions as well the VDE ring of honor. He cochairs the IEEE 5G Initiative, and has helped organizing IEEE conferences, most notably as TPC Chair of ICC 2009 and of TTM 2012, and as General Chair of VTC Spring 2013 and DATE 2014.

Meik Dörpinghaus (Member, IEEE) received the Dipl. Ing. degree (with distinction) and the Dr. Ing. degree (*summa cum laude*) both in electrical engineering and information technology from RWTH Aachen University, Aachen, Germany, in 2003 and 2010, respectively.

From 2004 to 2010, he was with the Institute for Integrated Signal Processing Systems, RWTH Aachen University. From 2010 to 2013, he was a Postdoctoral Researcher at the Institute for Theoretical Information Technology, RWTH Aachen University. Since 2013, he has been a Research Group Leader at the Vodafone Chair Mobile Communications Systems and at the Center for Advancing Electronics Dresden (cfaed), Technische Universität Dresden, Dresden, Germany. In 2007, he was a Visiting Researcher at ETH Zürich, Zürich, Switzerland. From 2015 to 2016, he was a Visiting Assistant Professor at Stanford University, Stanford, CA, USA. His research interests are in the areas of communication and information theory.

Dr. Dörpinghaus received the Friedrich Wilhelm Preis of RWTH Aachen in 2011 for an outstanding Ph.D. dissertation, and the Friedrich Wilhelm Preis of RWTH Aachen in 2004 and the Siemens Preis in 2004 for an excellent diploma thesis. He has coauthored a paper that received a best student paper award at the IEEE Wireless Communications and Networking Conference 2018.

Jeronimo Castrillon received the B.S. degree in electronics engineering from Pontificia Bolivariana University, Medellín, Colombia, in 2004, the M.S. degree from the Advanced Learning and Research Institute, Lugano, Switzerland, in 2006, and the Ph.D. (Dr. Ing.) degree (honors) from the RWTH Aachen University, Aachen, Germany, in 2013.

Currently, he is a Professor with the Department of Computer Science, Technische Universität Dresden, Dresden, Germany, where he is also affiliated with the Center for Advancing Electronics Dresden (cfaed). He is the Head of the Chair for Compiler Construction, with research focus on methodologies, languages, tools, and algorithms for programming complex computing systems. Since 2017, he has been a member of the executive committee of the ACM "Future of Computing Academy."

Akash Kumar (Senior Member, IEEE) received the joint Ph.D. degree in electrical engineering and embedded systems from the University of Technology (TUe), Eindhoven, The Netherlands and the National University of Singapore (NUS), Singapore, in 2009.

Currently, he is a Professor at the Technische Universität Dresden (TU Dresden), Dresden, Germany, where he is directing the Chair for Processor Design. From 2009 to 2015, he was with the National University of Singapore. His current research interests include design, analysis, and resource management of low power and fault tolerant embedded multiprocessor systems.

Christel Baier received the Diploma in mathematics, the Ph.D. degree in computer science, and the Habilitation degree from the University of Mannheim, Mannheim, Germany, in 1990, 1994, and 1999, respectively.

She has been a Full Professor and Head of the Chair for Algebraic and Logic Foundations of Computer Science at the Department of Computer Science, Technische Universität Dresden (TU Dresden), Dresden, Germany, since 2006. She was an Associate Professor of Theoretical Computer Science at the University of Bonn, Bonn, Germany, from 1999 until 2006. Her research expertise is in the areas of model checking, automata theory, temporal logics, and formal analysis of probabilistic systems.

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

Karlheinz Bock studied electronics and communication engineering at the University of Saarbrücken, Saarbrücken, Germany. He received the Dr. Ing. degree in RF microelectronics from the University of Darmstadt, Darmstadt, Germany, in 1994.

From January 2001 until September 2014, he was with the Fraunhofer Institute for Reliability and Microintegration (IZM), Munich, Germany [renamed Fraunhofer Research Institution for Modular Solid State Technologies (EMFT) in 2010], as Head of the Polytronic and Multi Functional Systems Department. From March 2008 until September 2014, he also served as Professor of Polytronic Microsystems at the University of Berlin (TU Berlin), Berlin, Germany. Since October 2014, he has been a Professor of Electronics Packaging and Director of the Institute for Electronics Packaging (IAVT) at the Technische Universität Dresden (TU Dresden), Dresden, Germany. He contributed to more than 300 publications, 12 patent families, 10 book chapters, and 70 invited talks.

Prof. Bock received the Dr. *honoris causa* from the Polytechnical University of Bukarest, Bukarest, Romania, in 2012, for his contributions to developing polytronics (large area flexible heterogeneous systems).

Frank Ellinger (Senior Member, IEEE) was born in Friedrichshafen, Germany, in 1972. He received the Diploma degree in electrical engineering from the University of Ulm, Ulm, Germany, in 1996, the MBA and the Ph.D. degree in electrical engineering, and the Habilitation degree in high frequency circuit design from ETH Zürich (ETHZ), Zürich, Switzerland, in 2001 and 2004, respectively.

Since 2006, he has been a Full Professor and Head of the Chair for Circuit Design and Network Theory, Technische Universität Dresden, Dresden, Germany. From 2001 to 2006, he was Head of the RFIC Design Group, Electronics Laboratory, ETHZ, and a project leader of the IBM/ETHZ Competence Center for Advanced Silicon Electronics hosted at IBM Research, Rüschlikon, Switzerland. He has been coordinator of the projects RESOLUTION, MIMAX, DIMENSION, ADDAPT, and FLEXIBILITY, funded by the European Union. He coordinates the cluster project FAST with more than 90 partners (most of them from industry) and the Priority Program FFlexCom of the German Research Foundation (DFG). He has been a member of the management board of the German Excellence Cluster Cool Silicon. He has authored or coauthored over 450 refereed scientific papers, and authored the lecture book *Radio Frequency Integrated Circuits and Technologies* (New York, NY, USA: Springer Verlag, 2008).

Prof. Ellinger was an elected IEEE Microwave Theory and Techniques Society (MTT S) Distinguished Microwave Lecturer (2009 2011). He has received several awards including the IEEE Outstanding Young Engineer Award, the Vodafone Innovation Award, the Alcatel Lucent Science Award, the ETH Medal, the Denzler Award, the Rohde&Schwarz/Agilent/Gerotron EEEf COM Innovation Award (twice), and the ETHZ Young Ph.D. Award.

Andreas Fery received the Diploma degree in physics from Konstanz University, Konstanz, Germany, in 1996 and the Ph.D. degree from the Max Planck Institute for Colloids and Interfaces (MPIKG)/Potsdam University, Potsdam, Germany, in 2000.

After a postdoctoral position at Institute Curie Paris, Paris, France, he became group leader at MPIKG. In 2007, he joined Bayreuth University, Bayreuth, Germany, as an Associate Professor and was promoted to Full Professor in 2008. Since 2015, he has been the Head of the Institute for Physical Chemistry/Polymer Physics at the Leibniz Institut für Polymerforschung Dresden, Dresden, Germany. He has published more than 200 papers in peer reviewed journals in the area of polymer science and colloid and interface science, which have received more than 7000 citations.

Prof. Fery received the Richard Zsigmondy award of the German Colloid Society and an ERC starting grant. He is the spokesperson of the German Physical Society division Chemical Physics/Polymer Physics and Member of the board of the European Colloid and Interface Society.

Frank H. P. Fitzek received the Diploma (Dipl. Ing.) degree in electrical engineering from RWTH Aachen, Aachen, Germany, in 1997 and the Ph.D. (Dr. Ing.) degree in electrical engineering from the Technical University Berlin, Berlin, Germany, in 2002.

Currently, he is the coordinator of the 5G Lab Germany and a Professor at Technische Universität Dresden, Dresden, Germany. His research focuses on wireless and mobile networks, mobile phone programming, network coding, cross layer and energy efficient protocol design, and cooperative networking.

Prof. Fitzek has received numerous awards, including the NOKIA Champion Award five times, the NOKIA Achievement Award (2008), the Danish SAPERE AUDE research grant (2010), and the Vodafone Innovation prize (2012).

Hermann Härtig received the Dipl. Inform. and Ph.D. degrees from the Technische Universität Karlsruhe, Karlsruhe, Germany.

He then led a team at the German National Research Center for Computer Science (GMD) to research and build an operating system that is Unix compatible but has much better security properties. Since 1994, he has led the Operating Systems group at the Technische Universität Dresden, Dresden, Germany. The research led to the “Fiasco” and “NOVA” variants of the L4 family of microkernels and to L4 based operating system component frameworks, which by now have been successfully deployed in critical environments.

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

Kambiz Jamshidi (Member, IEEE) received the Ph.D. degree in electrical engineering from the Sharif University of Technology (SUT), Tehran, Iran, in 2006.

He was with the Advanced Communication Research Institute (ACRI), SUT, as a Researcher from 2006 to 2009. From 2009 to 2012, he was a Senior Researcher with the High Frequency Technology (HFT) Institute, Deutsche Telekom University of Applied Sciences, Leipzig, Germany. He worked in the HFT and Photonics Lab of the Technical University of Berlin (TU Berlin), Berlin, Germany, from 2012 to 2013. Since 2013, he has been an Assistant Professor of Integrated Photonic Devices in the Communications Lab, Faculty of Electrical and Computer Engineering, Dresden University of Technology, Dresden, Germany.

Dr. Jamshidi is a Senior Member of the Optical Society of America.

Thomas Kissinger studied information systems engineering at the Technische Universität Dresden (TU Dresden), Dresden, Germany and received the Diploma in 2011. He received the Ph.D. degree from the Database Systems Group, TU Dresden, in 2017, with the dissertation on "Energy aware data management on NUMA architectures" that was decorated with the SAP Dissertation Award.

He continued his work within the collaborative research center named Highly Adaptive Energy Efficient Computing (HAEC) and now coordinates the Software Project Group of the CRC. His research mainly focuses on scalable and adaptive database system architectures on modern hardware.

Wolfgang Lehner received the M.S. degree in computer science and the Ph.D. degree (Dr. Ing.) with a dissertation on optimization of aggregate processing in multidimensional database systems from the University of Erlangen Nuremberg, Erlangen, Germany, in 1995 and 1998, respectively.

In November 1998, he joined the Business Intelligence (BI) group at the IBM Almaden Research Center, San Jose, CA, USA. In 2001, he finished his habilitation with a thesis on subscription systems and was therefore awarded with the *Venia Legendi*. Since October 2002, he has been conducting his research, teaching his students, and is involved in multiple industrial projects at the Technische Universität Dresden (TU Dresden), Dresden, Germany, where currently he is Full Professor and Head of the Database Systems Group. He was temporarily a Visiting Scientist at Microsoft Research, Redmond, WA, USA; at GfK Nuremberg; at SAP Walldorf; and at SAP Palo Alto, CA, USA. Since 2012, he has been an elected member of the VLDB Endowment Board of Trustees. He serves as a reviewer for multiple national and international research organizations.

Michael Mertig received the Ph.D. degree in the field of low temperature physics from the Technische Universität Dresden (TU Dresden), Dresden, Germany in 1983.

After academic positions at the Leibnitz Institute for Solid State and Materials Research, Dresden, Germany, and in the Max Plank Society, "Mechanics of Heterogeneous Solid States" group, Dresden, Germany, he was a Scientist at the Institute for Materials Science, TU Dresden, leading the research group "BioNanotechnology and Structure Formation." Since 2010, he has been Full Professor of Physical Chemistry at TU Dresden and Director of the Kurt Schwabe Institute for Measuring and Sensor Technologies, Meinsberg, Germany. His main fields of interest are biomimetic materials synthesis and high resolution structure analysis.

Wolfgang E. Nagel received the Ph.D. degree in computer science from RWTH Aachen, Aachen, Germany, in 1993.

He holds the Chair for Computer Architecture at the Technische Universität Dresden (TU Dresden), Dresden, Germany and is the Director of the Center for Information Services and HPC (ZIH). His research covers programming concepts and software tools to support the development of scalable and data intensive applications, analysis of computer architectures, especially with respect to energy efficiency, and development of efficient parallel algorithms and methods.

Prof. Nagel is Chairman of the Gauß Allianz e.V. and leads the Big Data Competence Center ScaDS Competence Center for Scalable Data Services and Solutions Dresden/Leipzig.

Giang T. Nguyen received the M.Eng. degree in telecommunications from Asian Institute of Technology (AIT), Thailand, in 2007 and the Ph.D. (Dr. Ing.) degree in computer science from the Technische Universität Dresden (TU Dresden), Dresden, Germany, in 2016.

Currently, he is a Senior Researcher at Deutsche Telekom Chair of Communication Networks, TU Dresden. His research interests lie in the area of 5G and highly adaptive and energy efficient computing (HAEC). His research focuses on network function virtualization (NFV), software defined networking (SDN), network coding, and the resilience aspects of peer to peer video streaming.

Final edited form was published in "Proceedings of the IEEE" 1 (107), S. 204-231. ISSN: 1558-2256
http://dx.doi.org/10.1109/JPROC.2018.2874895

Dirk Plettemeier received the Ph.D. degree in electrical engineering from Ruhr University Bochum, Bochum, Germany, in 2002.

In 2003, he joined the Chair and Laboratory for Electromagnetic Theory and EMC at the Technische Universität Dresden (TU Dresden), Dresden, Germany, as Head of the Research Group for Numerical Computation of High Frequency Electromagnetic Fields and Waves. From 2007 to 2011, he was group leader of the Research Group for Antennas and Wave Propagation at the Chair for RF Engineering at TU Dresden, where he was appointed Full Professor for the Chair for Radio Frequency and Photonics Engineering in 2011. His research interests include millimeter wave and terahertz systems with a focus on antennas and chip integrated applications, wave propagation, and remote sensing as well as imaging solutions for space applications. He has been involved in several international scientific activities as a coinvestigator for ESA and NASA space missions such as the Cassini Huygens, Rosetta, and Mars Express missions, mainly focusing on subsurface imaging. In the research cluster named Center for Advancing Electronics Dresden (cfaed) and the collaborative research center named Highly Adaptive Energy Efficient Computing (HAEC), he is working on electronics and energy efficient computing systems for future applications. He has published over 360 papers in international journals and conferences and holds over 16 patents.

Dr. Plettemeier has been reviewer for several associations such as the American Geophysical Union (AGU) and the Optical Society of America (OSA) and journals such as *Journal of Radio Science*, *Transactions on Emerging Telecommunications Technologies (ETT)*, IEEE Geoscience and Remote Sensing, IEEE Antennas and Propagation, IEEE Microwave Theory and Techniques, and Proceedings of IEEE. He is a member of the technical committee for “Microwave and THz Testing of the German Society for Nondestructive Testing.”

Michael Schröter (Senior Member, IEEE) received the Dr. Ing. degree in electrical engineering and the *Venia Legendi* degree on semiconductor devices from the Ruhr University Bochum, Bochum, Germany, in 1988 and 1994, respectively.

He was with Nortel and Bell Northern Research, Ottawa, ON, Canada, as a Team Leader and Advisor until 1996 when he joined Rockwell (later Conexant), Newport Beach, CA, USA, where he managed the RF Device Modeling Group. He has been a Full Professor at the University of the Technische Universität Dresden (TU Dresden), Dresden, Germany, since 1999, and was a Research Scientist at the University of California San Diego, La Jolla, CA, USA, until 2018. He is the author of the bipolar transistor compact model HICUM, a worldwide standard since 2003, and has coauthored a textbook *Compact Hierarchical Modeling of Bipolar Transistors with HICUM* and coedited a book *Silicon Germanium Heterojunction Bipolar Transistors for mm Wave Systems Technology, Modeling and Circuit Applications* as well as over 230 peer reviewed publications and several invited book chapters. He was a co founder of XMOD Technologies, Bordeaux, France, and was on the Technical Advisory Board (TAB) of RFMagic (now Entropic Inc.), a communications system design company in San Diego, CA, USA and also on the TAB of RFNano, a startup company in the area of carbon nanotube technology development in Newport Beach, CA, USA. During a two year leave of absence from TU Dresden (2009 2011) as Vice President of RF Engineering at RFNano, he was responsible for the device design of the first $4''$ wafer scale carbon nanotube FET process technology. He was the Technical Project Manager for DOTFIVE (2008 2011) and DOTSEVEN (2012 2016), which were European Union (EU) funded research projects for advancing high speed SiGe HBT technology toward THz applications, and has been leading the Carbon Path project within the German Excellence Cluster named Center for Advancing Electronics Dresden (cfaed).

Dr. Schröter has been a member of the ITRS/IRDS RF AMS subcommittee as well as the Technical Program Committees of BCTM and CSICS (merged into BCICTS in 2018).

Thorsten Strufe is a Professor of Privacy and IT Security at the Technische Universität Dresden (TU Dresden), Dresden, Germany. His research interests lie in the areas of privacy and resilience, especially in the context of social networking services and large scale distributed systems. Recently, he has focused on studying user behavior and security in online social networks and possibilities to provide privacy preserving and secure social networking services and big data solutions. His previous posts include faculty positions at TU Darmstadt and Universität Mannheim, Mannheim, Germany as well as Postdoctoral/Researcher positions at EURECOM and Technische Universität Ilmenau, Ilmenau, Germany.