A Highly Adaptive and Energy-Efficient Optical Interconnect for On-Board Server Communications

Waltenegus Dargie*, David Schoeniger†, Laszlo Szilagyi†, Xin An†, Ronny Henker† and Frank Ellinger†
*Chair of Computer Networks, Faculty of Computer Science
†Chair of Circuit Design and Network Theory, Faculty of Electrical and Computer Engineering
TU Dresden, 01062 Dresden, Germany
Email:{waltenegus.dargie, david.schoeniger, laszlo.szilagyi, xin.an, ronny.henker, frank.ellinger}@tu-dresden.de

Abstract—**As the global IP traffic and its demand for computation increase in a rapid and sustained manner, processor, server, and network architectures are also undergoing a considerable evolution. Two of the manifestations of this evolution are the integration of a large number of computing nodes in a single server and the interconnection of many servers via high-speed communication links. At present, however, the node-to-node communication bandwidth is one of the severest resource bottlenecks in massively parallelized applications. There is a concerted effort by the academia and the industry to achieve higher data rate by assembling multiple parallel links. This effort, however, is inherently limited by many constrains, including space. Optical interconnects, on the other hand, promise superior data rates, lower transmission losses, and less inter-channel crosstalk when compared to electrical interconnects. Development in this area promise data rates in the range of Tera bits per second per link and beyond. So far, however, little attention is given to the power adaptiveness of optical interconnects.**

In this paper, we present an optical interconnect concept which adjusts its power consumption in response to the change in the statistics of the incoming workload. The several components of the link have been designed and developed in hardware. Based on initial power and performance measurements of the components, a link model of our optical interconnect was created. The performance-power consumption characteristics of this model was simulated applying different workload statistics and the potential of the energy savings by the adaptivity have been evaluated. It is revealed that the power consumption of our optical interconnect reduces by up to $40\,\%$ when its workload was exponentially distributed (signifying underutilisation) compared to a Weibull distribution workload (signifying full capacity workload). This study confirms the high potential for power saving in performance adaptive optical interconnects.

Index Terms—**Energy-efficient computing, interconnects, multicore processor architecture, optical interconnects, parallel computing, power model**

I. Introduction

The global IP traffic has shown a rapid and sustained rise in the recent years and evidence suggests that the trend will continue so in the near future. According to a recent (June 02, 2016) forecast by Cisco's Visual Networking Index [1], the annual global IP traffic will "pass the zettabyte threshold by the end of 2016", reaching 2.3 ZB per year by 2020. When compared with the 2015 index, global IP traffic will increase nearly three times over the next 5 years, which means that IP traffic will grow at a compound annual growth rate (CAGR) of 22 percent by 2020. The introduction of cloud computing has made a considerable contribution to the rise in the communication, storage, and processing demand of IP traffic. According to Cisco's Global Cloud Index, by 2019 83 percent of all data centre traffic will come from the cloud and 4 out of 5 data centre workloads will be processed in the cloud [2]. Similarly, IBM forecasts that 80 percent of new applications will be developed with cloud characteristics [3].

In order to deal with these phenomena, processor, server, and network architectures are undergoing a considerable evolution. As far as multicore processor architectures are concerned, the efficient integration of a large amount of cores in a single chip, non-uniform cache hierarchies, fast and efficient simultaneous multi-threading, and advanced branch prediction strategies, among others, are being endeavoured by both the industry and the research community. The goal is multi-fold. At the core level, it is to significantly increase the number of instructions that can be fetched, decoded, dispatched, issued, and executed simultaneously [4], and, at the system level, it is to enable intensive inter-processor communication with either a very low latency or a very high bisection bandwidth [5].

As far as communication bandwidths are concerned, particular emphasis is being given to short-range communication. Thus the data rates of intra-node, inter-node, and inter-cluster interconnects are growing continuously. For example, the IBM POWER8$^{\text{TM}}$ processor architecture can now provide up to 230 GB/s sustained memory bandwidth and 3.6 TB/s for memory chip interconnect using 12 lanes per direction, each lane supporting a data rate of 150 GB/s. Similarly, the PCI-Express (PCI-e x16) is evolving from its current sustained 15.754 GB/s to 31.508 GB/s by the end of 2017 [6] whilst the 10 Gbit/s Ethernet will evolve to 400 Gbit/s by the end of 2017 [7]. The InfiniBand specification, which is extensively used for inter-cluster communication in high performance computers, currently achieves 290.91 Gbit/s sustained rate using 12 lanes, each lane operating at 25 Gbit/s, but by the end of 2017, it is projected to achieve 600 Gbit/s [8]. Nevertheless, compared to the speed of processors and the achievable data rate of inter-core communication, the achievable data rate of inter-node and inter-cluster communication is still significantly small. Furthermore, as the number of lanes per I/O port increases, the size and power dissipations of electrical interconnects become considerably high, thereby limiting the performance and energy efficiency of data centres and high performance computers.

The complementary options which the research and industry communities are currently investigating, are optical interconnects, because they yield high bandwidth-distance product, can be manufactured with considerably small size and power dissipation, and do not suffer from electromagnetic interference, all of which will enable dense deployment. Both product level installations and prototypes indicate that optical interconnects can be employed to interconnect blade servers and serve as mid-plane in data centre configurations [5], [9], [10], [11]. However, existing or proposed interconnects typically rely on optical fibres and pay little attention to the design of the optical elements. Moreover, there is almost no report in the literature whether and how proposed interconnects can adapt their power consumption to load fluctuation at the physical layer.

In this paper, we propose an optical on-board interconnect concept and present its components which we designed, optimised, and fabricated for inter-node communication. Our interconnect has three novel features:

1) Unlike previously proposed optical interconnects which use fibre optics as transmission medium, our interconnect uses optical waveguides which can be integrated into a server main board and, therefore, considerably minimise the space requirement.
2) Whereas the main focus of the existing or proposed optical interconnects is the achievable data rate, our design equally takes the power consumption characteristics of the interconnect into consideration. More specifically, our interconnect is adaptive to its workload in that it scales down its power consumption when it is underutilised or idle.
3) Our initial power model suggests that the optical interconnect can be fine-tuned to produce additional reduction in power consumption when the load statistics is known.

The rest of the paper is organized as follows: In Section II, we summarise related work and put ours in perspective. In Section III, we describe the components of our optical interconnects. In Section IV, we provide a detailed account of the power consumption model we set up in order to characterise the power consumption of the optical interconnect for various workload statistics. Finally, in Section V, we give concluding remarks and outline future work.

II. Related Work

As the performance and complexity of computing and server platforms increase, also the bandwidths at the inputs/outputs (I/Os) need to rise in order to transmit the huge amount of data between the several computing nodes in a multi-processor environment. For example, the IBM POWER8$^{\text{TM}}$ processor consists of 12 cores and the core throughput is supported by an overall off-chip I/O bandwidth of 7.6 Tbit/s [12]. To accommodate such a high link data rate and also the increasing number of lanes per I/O port, optical interconnects are identified as a promising solution to connect multiple processors on a board [5]. Compared to traditional copper-based electrical interconnects, optical interconnects provide higher data rates and bandwidth densities, lower transmission losses, lower latencies as well as less inter-channel crosstalk. Additionally, the energy efficiency of optical on-board interconnects can be higher, particularly, at high data rates and large link distances (optical links consume less power at distances above 10 cm and 28 Gbps [13]).

In [9], a server architecture employing optical interconnects has been proposed and developed. It is novel feature is the use of high bandwidth mid-plane optical interconnects. In this architecture the CPU blades of the server were connected with the peripheral network switch, the PCI-e switch and memory blades by optical multi-mode fibres (MMF) achieving an aggregated throughput of 20 Tbit/s. A board-level optical interconnect module with highly integrated parallel optical transceivers and polymer waveguides integrated into the PCB was shown in [14]. The system consists of 48 channels, each at a data rate of 15 Gbit/s with an energy efficiency of 9.7 pJ/bit, which provides an aggregated bi-directional throughput of 360 Gbit/s at a total power consumption of 3.5 W. Furthermore, with a holey opto-chip consisting of 48 channels, an aggregated bi-directional data rate of 480 Gbit/s was achieved [15]. Each channel of this interconnect works up to 20 Gbit/s with an energy efficiency of 7.3 pJ/bit. Recently, an optical end-to-end on-board interconnect with embedded waveguides and 90° out-of-plane connectors was demonstrated [16]. The demonstrator consists of 96 channels each with a data rate of 25 Gbit/s totalling 2.4 Tbit/s aggregated transmission capacity. The system was tested with Leap® high-speed on-board optical transceiver which enables 300 Gbit/s on 12 channels with an overall power consumption of 5.8 W resulting in approximately 19 pJ/bit per channel.

Although it has been proven that optical interconnects can achieve the bandwidth, latency and form factor requirements of future data centres and computing interconnects, the energy efficiency still has to be further improved. It is expected that by 2020 the energy efficiency should be reduced to less than 1 pJ/bit for a channel data rate of 25 Gbit/s or more [17]. For significant reduction of the power consumption of the optical interconnects in the near future, the current systems have one major drawback. Mostly they are optimized by design to accommodate the peak data rate of the transmission. In this way, the links are statically operated always at their peak performance and therefore peak power consumption, resulting in a poor energy efficiency. However, in reality the data rates of communication links vary dynamically over time depending on the processors' computation, the server's workload and the user requirements. This leads to the requirement of scalable optical interconnect systems and architectures with tunable performance and power consumption.

In recent years, research mainly focused on implementing dynamic transmission systems at network level by rerouting concepts and optimized switching architectures, for instance [10], [18], [19], [20]. However, the end-to-end performance and power consumption of the links are still static and no link adaptivity has been implemented so far. Therefore, jointly implementing performance and power adaptive features directly in the optical interconnects on link system or component level results in further improved energy-efficient link utilization. In this paper, we introduce and study this new adaptivity concept which allows the tuning of optical communication links at system and component level at the same time. The method is based on adapting the performance and the power consumption of the electrical transceiver circuitry to the present transmission workload by changing their operating points, for example, their bias currents. We have made the first step towards developing an energy-efficient and adaptive optical on-board chip-to-chip interconnection with scalable data rates of up to 100 Gbit/s and an energy per bit of less than 5 pJ/bit, as will be explained in the next section.

Each link consists of low-loss optical on-board waveguides with optical coupling elements and high-speed low-power laserdiode drivers (LDDs) with vertical-cavity surface-emitting lasers (VCSEL) on the transmitter side as well as photodetectors (PDs) with transimpedance/main amplifiers (TIAs/MAs) and clock-data-recoveries (CDRs) at the receiver side. Except the VCSEL and the PD, all those main components of the link have been designed, optimized and characterized separately. Each interconnect transceiver of the links can be individually configured and scaled down in performance at runtime to accommodate the actual workload requirements. With this, it is expected that more than 50 % reduction in power consumption can be achieved per link. In the following sections the potential of the performance and power scaling is investigated based on statistical models of the link performance and the individually measured power consumption of the link components. In future, the components will be integrated to the complete link system and multiple links will be highly parallelized to achieve 60 parallel transceiver channels in a single chip with an aggregated bi-directional throughput of 12 Tbit/s.

III. Description of the Optical Interconnect

The optical interconnect concept we propose in this paper is intended to be employed in a server board architecture with 4×4 computing nodes, as it is shown in Fig. 1. Assuming an outer node matrix dimension of 10 cm × 10 cm, this results in a connection length of <25 mm for the middle nodes and a maximum length of 100 mm for the border nodes. The general structure of the optical connections between computing nodes includes in total 15 links from each node to its neighbouring nodes, as shown in Fig. 2. Therefore, each node will have 60 incoming and 60 outgoing optical connections each intended to run up to 100 Gbit/s in future which results in a single direction throughput of 6 Tbit/s for each node and an overall throughput of 96 Tbit/s for the complete board, respectively. The optical links are implemented as on-board polymer waveguides which are either directly imprinted on the printed circuits board (PCB) or laminated via a foil onto the PCB of the server board. Each optical link describes a single channel optical transmission system which uses the direct modulation of a vertical-cavity surface-emitting laser (VCSEL), as shown in Fig. 3. The transmitter and receiver hardware chain consists of the following components:

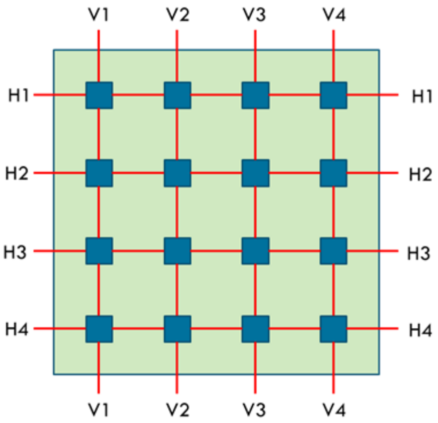

Fig. 1. Optical network on the board.

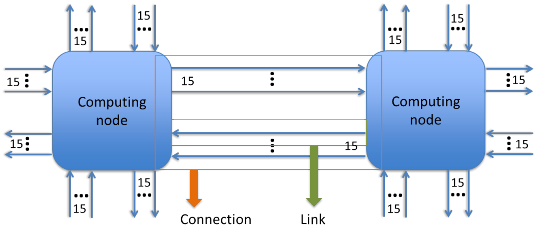

Fig. 2. An optical connection between two neighboring compute nodes consisting of 15 outgoing and 15 incoming optical links.

Laserdiode driver (LDD): The LDD is an amplifier which provides the data signal with sufficient voltage/current swing to the VCSEL for the optical modulation. Furthermore, it enables electrical matching between the laser and the electrical data interface.

Vertical-cavity surface-emitting laser (VCSEL): The VCSEL generates the optical carrier wavelength (usually at 850 nm) for the optical data transmission. VCSELs are directly-modulated devices by which the electrical data signal is modulated onto the optical carrier.

Optical coupling elements: These elements connect the optical output of the VCSEL as well as the optical input of the photodiode with the optical on-board waveguide. They are designed as 90° coupling elements.

Optical waveguide: The optical waveguide is the transmission medium of the optical data link. This on-board waveguides are designed for multimode operation [21], [22].

Photodiode (PD): The PD detects the optical signal and converts it back to an electrical current.

Transimpedance amplifier (TIA): The TIA is the receiver amplifier which converts the photocurrent of the PD into a voltage with certain amplification.

Main amplifier (MA): The MA amplifies the voltage to logical levels which can be processed by the following mixed-signal or digital circuits of the link system. Doing so, it also reshapes or just steepen the edges of the data signal.
Clock data recovery (CDR): The CDR extracts the clock signal from the data signal and performs the retiming and recovering the data signal.

To realize such an optical link, all components have been designed, fabricated and measured, except the VCSEL and the PD which were provided as commercial standard components. Due to the optical components and the designed CDR, currently the overall optical link data rate is optimized and limited to 25 Gbit/s while the other circuitry such as the LDD, TIA and MA can provide higher bandwidths and data rates, respectively. One way of integrating the optical link into a compute node is shown in Fig. 4 and explained in [23].

The realization of the optical interconnects requires a low-loss, low-dispersion optical board-level interconnection and out-of-plane coupling optics. Fabricated waveguides using organic-inorganic hybrid ORMOCER$^{\circledR}$ material system have been employed. They are designed for multi-mode operation with low loss and a target bit rate length product of 50 Gbit/s × m in the 850 nm wavelength range [21].

If the computation load of a node is relatively low and, by implication, its incoming and outgoing data traffic is small, the performance of the optical link can be dynamically scaled or reduced, so that power can be saved. While the power consumption of PDs can be neglected (in µW range), the VCSELs normally requires currents of up to 10 mA with a 2 V supply voltage thereby consuming maximum 20 mW of power, independent of the link performance. However, this is a small value in comparison to the power consumption of the other transceiver circuitry when operating in maximum performance mode. Furthermore, the VCSEL bias is usually provided by the LDD and hence, the VCSEL power consumption accounts to the LDD one. Thus, the performance and power adaptivity is enabled by the electrical circuitry of the link system. The performance and power consumption of the circuits can be scaled by changing the operating point of the circuits either by tuning their bias voltages or their bias currents. In situations in which low link data rates are sufficient, the supply current can be reduced thereby reducing the components DC power consumption. At the same time, however, the transconductance of the transistors and, therefore, the gain of the amplifiers drop due to lower bias currents. Since the signal levels at the interfaces of the subsequent processing blocks like the VCSEL or CDR should remain constant, the gain drop has to be compensated.

In order to adaptively control the data rate and power consumption of the receiver of the optical interconnection, we have designed and fabricated a high-speed TIA/MA chip with adaptive tuning mechanism in an advanced 130 nm SiGe BiCMOS technology with a maximum transit frequency of 300 GHz [24]. Verification based on S-parameter measurement has shown a constant gain of 69.8 dBΩ and a tunable operation bandwidth between 12.8 GHz to 61.6 GHz (leading to data rates between 18 Gbit/s to 88 Gbit/s) at a tunable power consumption between only 66.1 mW to 78.1 mW. For keeping the gain constant and to compensate the gain drop due to the adaptive performance tuning, the TIA core was implemented as a variable-gain TIA [25]. Neglecting the output driver of the TIA/MA chip, which is only required for measurement purposes, the power consumption of the TIA/MA core can be reduced by 40 %. At highest speed the chip operates very energy-efficient with an energy per bit of only 0.89 pJ/bit. For use in optical data transmission the TIA/MA chip was bonded to a PD as shown in Fig 5.

Furthermore, such an adaptive bandwidth and power consumption tuning mechanism has been implemented in LDD and CDR chips as well. Several power consumption regions for different performances were determined for the LDD, TIA, MA and CDR, as shown in Fig. 7 and Table I. In the next section these different performance and power consumption operating modes of the link are analysed and modelled. Using different workload scenarios, the potential of the adaptive tuning of the link is studied in Section IV.

IV. Power Model

When considered individually and with respect to the power consumption of processors, the power consumption of a single interconnect may appear insignificant. However, depending on the specific network topology, the number of interconnects required to connect computing nodes is relatively large. For instance, in a mesh topology network consisting of n computing nodes, $n(n-1)/2 - n$ interconnects are required. This amounts to 126 interconnects for 16 nodes and 2016 interconnects for 64 nodes. Hence, the overall power consumption of interconnects can be up to 10 % of the overall power consumption of a server [27], [28].

To put the power adaptiveness of our interconnect in perspective, we first investigated the power consumption characteristic of a state-of-the-art 56 Gbit/s InfiniBand as a reference. The interconnect was employed to establish a direct high-speed link between two servers. It is supplied with power through two power rails: A 3 V rail and a 12 V rail. When idle, it consumed around 5 W. Then we transferred data for two minutes from one of the servers to the other with full rate. Thereafter we interrupted the data transfer for approximately 2 minutes, decoupled the interconnect from the system (the interconnect was still physically connected but the operating system did not recognise it) and then we resumed the data transfer with full data rate once again. Our aim was to observe the difference in power consumption between the various states. For both power rails, the power consumption of the interconnect sharply dropped when it was removed from the system, otherwise, there was no conspicuous difference in power consumption between the idle and the full-load states (refer to Fig. 6).

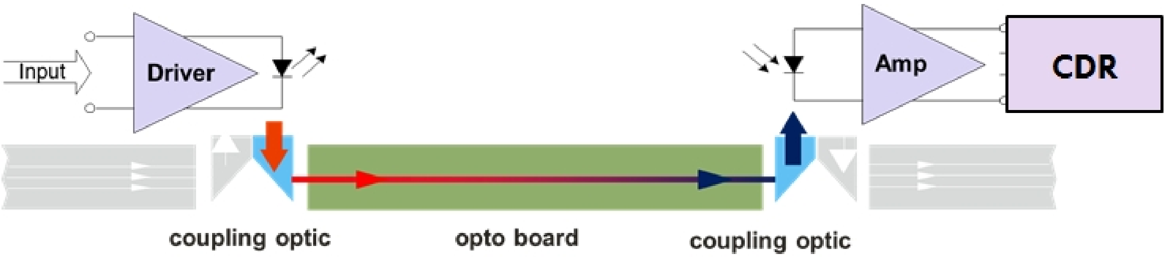

Fig. 3. Optical link components based on VCSEL transmission.

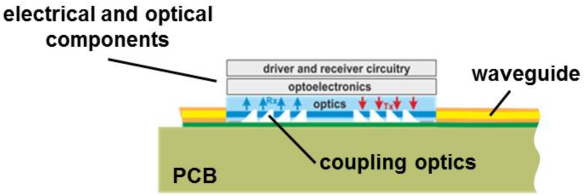
electrical and optical components
driver and receiver circuitry
optoelectronics
optics
Rx
Tx
waveguide
coupling optics
PCB

Fig. 4. System integration of the optical link in the computing node.

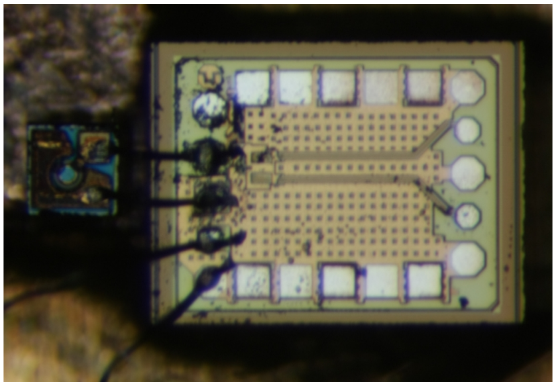

Fig. 5. Chip mircograph of the TIA/MA chip bonded to a PD.

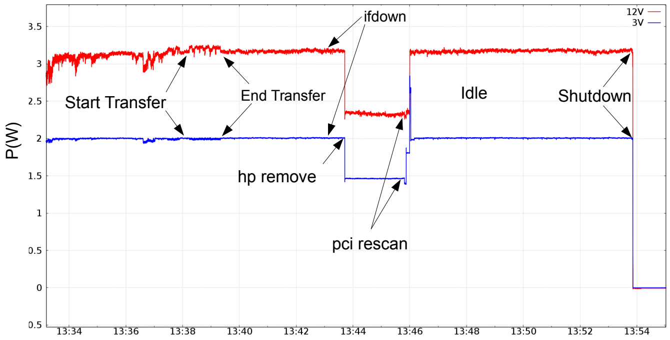

Fig. 6. The power consumption of a 56 Gbit/s InfiniBand interconnect. The power was measured from the 3 V and 12 V lanes during an idle and a full-bandwidth data transfer. Courtesy of Ilsche et al. [26].

TABLE I
THE RELATIONSHIP BETWEEN THE DATA RATE AND POWER CONSUMPTION OF THE RECEIVER COMPONENTS WITH DIFFERENT OPERATION REGIONS.

Data Rate (Gbit/s)	5	10	15	20	25	30	35	40	45	50	55	85
TIA	17.92mW@18.3Gbit/s			20.3mW@34.7Gbit/s			22.7mW@54.3Gbit/s					25.10mW@ 84.9Gbit/s
MA	3.22mw@18.3Gbit/s			5.68mW@34.7Gbit/s			8.13mW@54.3Gbit/s					10.27mW@ 84.9Gbit/s
CDR	6.97 mW @6.25 Gbit/s	10.35 mW@ 12.5 Gbit/s	16.81mW@25Gbit/s									

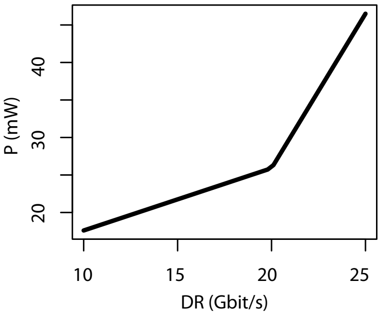

Fig. 7. The relationship between the data rate and power consumption of the optical transmitter.

A. Device Specification and Workload Generation

At present, the components of our optical interconnects are available separately (i.e., they are not yet assembled into a single transceiver unit). Hence, in order to characterise the power consumption of the optical interconnect, we first measured the power consumption of the individual components separately. Thus, for each receiver component, we identified the regions where the power consumption changes in response to a change in the data rate and produced Table I. There is a segmental linear relationship between the power consumption and the data rate of the optical transmitter, as can be seen in Fig. 7. Then, we analysed the overall power consumption of the transceiver by generating four types of workloads having different workload statistics signifying underutilised (exponential distribution), normal (normal and uniform distribution), and overloaded (Weibull distribution) conditions, respectively. For each distribution, we first generated 10,000 random samples with associated probability density using the R statistical tool. Then for each random sample, we generated a corresponding

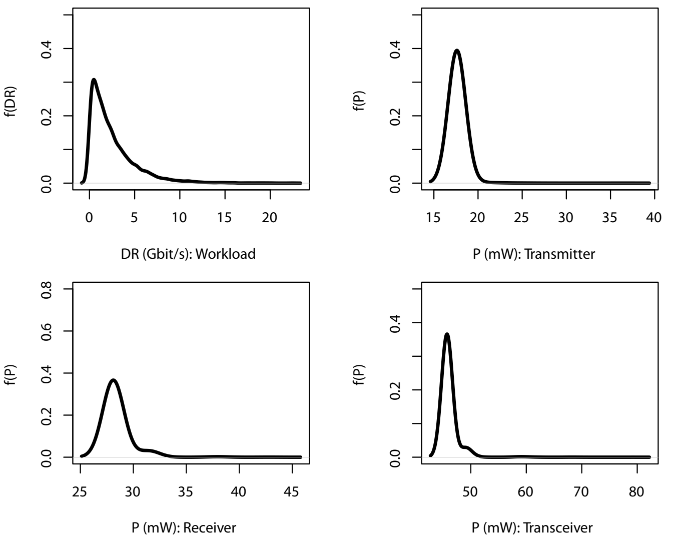

Fig. 8. The relationship between the densities of the workload and the power consumption of the optical interconnect during underutilised condition.

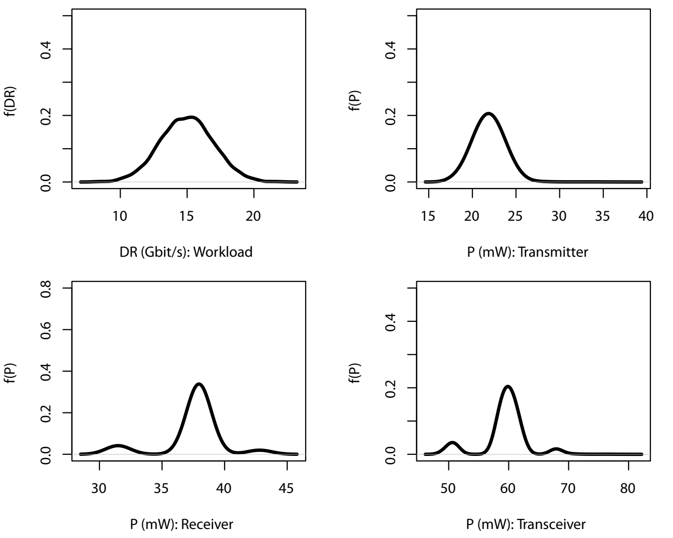

Fig. 9. The relationship between the densities of the workload and the power consumption of the optical interconnect under a normal load.

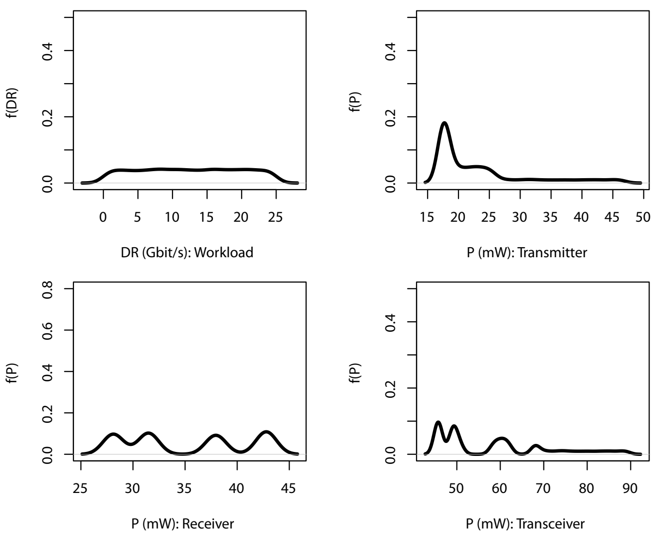

Fig. 10. The relationship between the densities of the workload and the power consumption of the optical interconnect under a uniform load.

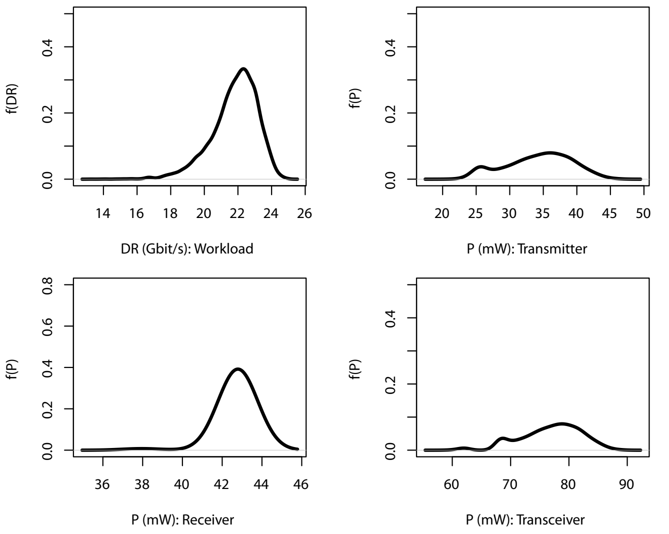

Fig. 11. The relationship between the densities of the workload and the power consumption of the optical interconnect under overloaded condition.

data rate and computed the corresponding overall power cost to transfer a payload with the specified data rate between two nodes, using Fig. 7 and Table I. Finally, we produced the overall power consumption of the optical transceiver by adding the power consumption of the individual components.

B. Power Consumption Characteristics

Figure 8 shows the relationship between power consumption (P) and data rate (DR) when the optical interconnect was underutilised (predominately idle; exponential workload distribution). The overall power consumption was approximately $45\,\mathrm{mW}$ and more than two-third of this amount was on account of the receiver, clearly indicating that the idle power consumption of the optical receiver is dominant on account of the transimpedance amplifier (TIA). Figure 9 displays the power consumption of the optical interconnect when it was supplied with a normally distributed workload which varied around $15\,\mathrm{Gbit/s}$. Interestingly, the power consumption assumed a more or less normal distribution, centring approximately at $61\,\mathrm{mW}$ as well. The figure suggests that approximately a linear relationship between the data rate and the overall power consumption can be established when the statistics of the transferred workload is normally distributed. Unfortunately, this relationship disappears when the workload statistics becomes uniform distributed (as can be seen in Fig. 10). In reality, however, the workload of the interconnect, which we assume,

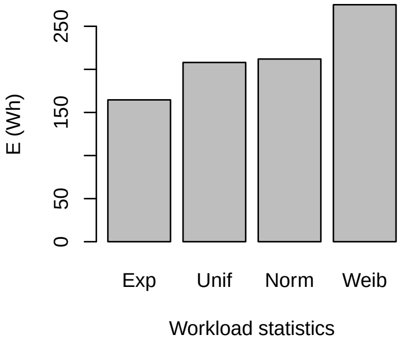

Fig. 12. Comparison of the energy consumption of the optical interconnect for different workload statistics.

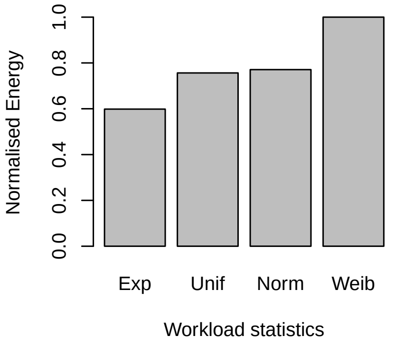

Fig. 13. Comparison of the normalised energy consumption of the optical interconnect taking the energy consumption of the dense workload (Weibull distribution) as the normalising factor.

arises from the aggregate workloads of different applications, most likely will have a normally distributed workload in accordance with the Central Limit Theorem.

Figure 11 displays the case when the interconnect was overloaded (Weibull workload distribution). A non-linear property is apparent in the power consumption of the transmitter than in the receiver, which has a predominantly normal distribution centering approximately at 43 mW. But the overall power consumption of the transceiver was non-linear as well.

C. Energy Consumption

In order to make a quantitative comparison between the power consumption characteristics of the different workload statistics, we calculated the 1 h energy consumption of the interconnect. Referring to Fig. 12, for the exponential workload, the overall energy consumption was 164.6 W h whereas for the uniform and normal distributed workloads, it was 208 and 212 W h, respectively. The difference in power consumption between these two workloads is insignificant even though their power consumption characteristics appeared to be significantly different. By contrast, for the Weibull workload, the energy consumption was significantly larger, 275 W h. If we take the power consumption of the Weibull workload as a reference, the power consumption of the interconnect reduced by 40 % when processing the exponential workload, whilst the reduction was approximately 20 % for the uniform and normal workloads (Fig. 13).

V. Conclusion

In this paper, we proposed an adaptive and energy-efficient optical interconnect concept for node-to-node communication in a multi-core, multi-processor server setting. We were motivated by the possibility of higher data rates, lower transmission losses, and less inter-channel crosstalk which optical interconnects can offer when compared to electrical interconnects. Additionally, the energy efficiency of optical on-board interconnects can be higher at high data rates and large link distances. However, existing or proposed optical interconnects consume a more or less fixed amount of power under different workloads.

One of the key features of the proposed optical interconnect is the dynamic adjustment of the operating points of the individual transceiver components in response to a change in the workload statistics. The electrical circuitry such as LDD, TIA/MA and CDR were designed for the link system and equipped with adaptive performance and power consumption tuning. All components were individually characterized in the lab and different performance and power consumption regions have been determined. We generated different workload statistics signifying underutilised, normal, and overloading conditions and investigated the power consumption of a complete optical interconnect in these conditions based on the individual measurement results. Particularly, when the workload statistics assumed a normal distribution, the relationship between the workload and the power consumption became almost linear. Furthermore, the optical interconnect consumed 40 % less energy when underutilised compared to the energy it consumed when it was transferring data approximately at full capacity. Likewise, its energy consumption was 20 % lower when it was operating with normally and uniformly distributed workloads.

Furthermore, we computed the overall power consumption of the optical interconnect based on first power consumption measurements of the individual components, because the interconnect is not yet assembled as a single unit. Our future work is to put these components together. This would also enable us to explore further adaptation possibilities and to refine our power consumption model.

Acknowledgments

This work is partially supported by the German Research Foundation (DFG) within the Collaborative Research Center SFB 912 – HAEC.

References

[1] Cisco Inc., “Cisco VNI forecast and methodology, 2015–2020,” 2016.
[2] Cisco Inc., “Cisco global cloud index: Forecast and methodology, 2015–2019,” 2015.
[3] J. Tendler, “An introduction to Power8 processor,” January 2014.
[4] B. Sinharoy, J. A. V. Norstrand, R. J. Eickemeyer, H. Q. Le, J. Leenstra, D. Q. Nguyen, B. Konigsburg, K. Ward, M. D. Brown, J. E. Moreira, D. Levitan, S. Tung, D. Hrusecky, J. W. Bishop, M. Gschwind, M. Boersma, M. Kroener, M. Kaltenbach, T. Karkhanis, and K. M. Fernsler, “IBM POWER8 processor core microarchitecture,” *IBM Journal of Research and Development*, vol. 59, pp. 2:1–2:21, Jan 2015.

[5] L. Schares, D. M. Kuchta, and A. F. Benner, “Optics in future data center networks,” in *18th IEEE Symposium on High Performance Interconnects*, pp. 104–108, Aug 2010.

[6] “PCIe 4.0 heads to fab, 5.0 to lab. EE times. retrieved aug 27, 2016,” 2016.

[7] “Adopted timeline (pdf). ieee 802.3bs task force. 2014-05-19. retrieved feb 24, 2015,” 2015.

[8] “Infiniband roadmap: http://www.infinibandta.org (retrieved dec. 09, 2016),” 2016.

[9] J. Matsui, T. Yamamoto, K. Tanaka, T. Ikeuchi, S. Ide, S. Aoki, T. Aoki, T. Ishihara, M. Iwaya, K. Kamoto, K. Suematsu, and M. Shiino, “Optical interconnect architecture for servers using high bandwidth optical midplane,” in *OFC/NFOEC*, pp. 1–3, March 2012.

[10] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, “Helios: A hybrid electrical/optical switch architecture for modular data centers,” in *Conference Proceedings of the ACM SIGCOMM*, no. 4, (New York, NY, USA), pp. 339–350, ACM, Aug. 2010.

[11] C. Kachris and I. Tomkos, “A survey on optical interconnects for data centers,” *IEEE Communications Surveys Tutorials*, vol. 14, pp. 1021–1036, Fourth 2012.

[12] E. J. Fluhr, S. Baumgartner, D. Boerstler, J. F. Bulzacchelli, T. Diemoz, D. Dreps, G. English, J. Friedrich, A. Gattiker, T. Gloekler, C. Gonzalez, J. D. Hibbeler, K. A. Jenkins, Y. Kim, P. Muench, R. Nett, J. Paredes, J. Pille, D. Plass, P. Restle, R. Robertazzi, D. Shan, D. Siljenberg, M. Sperling, K. Stawiasz, G. Still, Z. Toprak-Deniz, J. Warnock, G. Wiedemeier, and V. Zyuban, “The 12-core POWER8$^{\text{TM}}$ processor with 7.6 Tb/s IO bandwidth, integrated voltage regulation, and resonant clocking,” *IEEE Journal of Solid-State Circuits*, vol. 50, pp. 10–23, Jan 2015.

[13] K. Saraswat, H. Cho, P. Kapur, and K.-H. Koo, “Performance comparison between copper, carbon nanotube, and optical interconnects,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 2781–2784, 2008.

[14] F. E. Doany, C. L. Schow, B. G. Lee, R. A. Budd, C. W. Baks, C. K. Tsang, J. U. Knickerbocker, R. Dangel, B. Chan, H. Lin, C. Carver, J. Huang, J. Berry, D. Bajkowski, F. Libsch, and J. A. Kash, “Terabit/s-class optical PCB links incorporating 360-Gb/s bidirectional 850 nm parallel optical transceivers,” *Journal of Lightwave Technology*, vol. 30, pp. 560–571, Feb 2012.

[15] F. E. Doany, B. G. Lee, D. M. Kuchta, A. V. Rylyakov, C. Baks, C. Jahnes, F. Libsch, and C. L. Schow, “Terabit/sec VCSEL-based 48-channel optical module based on holey CMOS transceiver IC,” *Journal of Lightwave Technology*, vol. 31, pp. 672–680, Feb 2013.

[16] M. Immonen, R. Zhang, M. Press, H. Tang, W. Lei, J. Wu, H. J. Yan, L. X. Zhu, and M. Serbay, “End-to-end optical 25Gb/s link demonstrator with embedded waveguides, 90° out-of-plane connector and on-board optical transceivers,” in *42nd European Conference on Optical Communication (ECOC)*, pp. 1–3, Sept 2016.

[17] A. Benner, “Optical interconnect opportunities in supercomputers and high end computing,” in *OFC/NFOEC*, pp. 1–60, March 2012.

[18] F. Idzikowski, S. Orlowski, C. Raack, H. Woesner, and A. Wolisz, “Saving energy in IP-over-WDM networks by switching off line cards in low-demand scenarios,” in *14th Conference on Optical Network Design and Modeling (ONDM)*, pp. 1–6, Feb 2010.

[19] X. Chen, L.-S. Peh, G.-Y. Wei, Y.-K. Huang, and P. Prucnal, “Exploring the design space of power-aware opto-electronic networked systems,” in *11th International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 120–131, Feb 2005.

[20] A. K. Kodi and A. Louri, “Energy-efficient and bandwidth-reconfigurable photonic networks for high-performance computing (HPC) systems,” *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, pp. 384–395, March 2011.

[21] K. Nieweglowski, R. Rieske, S. Sohr, and K. J. Wolter, “Design and optimization of planar multimode waveguides for high speed board-level optical interconnects,” in *IEEE 63rd Electronic Components and Technology Conference (ECTC 2013)*, pp. 1898–1904, May 2013.

[22] K. Nieweglowski, R. Henker, F. Ellinger, and K.-J. Wolter, “Performance of step index multimode waveguides with tuned numerical aperture for on-board optical links,” in *Proceedings of SPIE Vol. 8991, Optical Interconnects XIV*, vol. 8991, (San Francisco, USA), pp. 899105– 1–10, 01-06 February 2014.

[23] K. Nieweglowski, L. Lorenz, K. J. Wolter, and K. Bock, “Multichannel optical link based on polymer multimode waveguides for board-level interchip communication,” in *2015 European Microelectronics Packaging Conference (EMPC)*, pp. 1–7, Sept 2015.

[24] D. Schoeniger, R. Henker, and F. Ellinger, “High-speed transimpedance amplifier with runtime adaptive bandwidth and power consumption in 0.13 μm SiGe BiCMOS,” *Electronics Letters*, vol. 52, pp. 154–156, January 2016.

[25] D. Schoeniger, R. Henker, and F. Ellinger, “A high-speed energy-efficient inductor-less transimpedance amplifier with adjustable gain for optical chip-to-chip communication,” in *SBMO/IEEE MTT-S International Microwave and Optoelectronics Conference (IMOC)*, (Porto de Galinhas, Brazil), 3-6 November 2015.

[26] T. Ilsche, D. Hackenberg, S. Graul, J. Schuchart, and R. Schöne, “Power measurements for compute nodes: Improving sampling rates, granularity and accuracy,” in *The Sixth International Green and Sustainable Computing Conference*, Dec 2015.

[27] W. Dargie, “A stochastic model for estimating the power consumption of a processor,” *IEEE Transactions on Computers*, vol. 64, pp. 1311–1322, May 2015.

[28] C. Möbius, W. Dargie, and A. Schill, “Power consumption estimation models for processors, virtual machines, and servers,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, pp. 1600–1614, June 2014.