Low Power Optical Transceivers for Switched Interconnect Networks

Yury Audzevich*, Philip M. Watts†, Andrew West*, Alan Mujumdar*, Jon Crowcroft*, Andrew W. Moore*
*Computer Laboratory, University of Cambridge, Cambridge, UK
†Dept. of Electronic and Electrical Engineering, University College London, London, UK

***Abstract*—The power-consumption of network equipment is under ever-increasing scrutiny. As part of an ensemble project seeking to reduce power-consumption within data-centers[1], this work focuses on reducing the power consumption of photonic transceivers for future fast power gated and/or optical switching networks. Utilising an open-source toolkit, we show that SERDES dominates power consumption of traditional optical transceivers. This result has particular implications for the modulation format of future interconnects. At 25 Gb/s line rate, SERDES blocks of PAM-16 and 4-wavelength WDM are shown to have 53% and 79% lower power respectively compared with SERDES of serial NRZ as well as reduced power gating restoration time and energy.**

I. Introduction

The energy performance of networked systems has become a *1st class* property, of interest to industry and researchers alike. It has been shown that computer systems must be made to 'do nothing well' for large energy savings to be made by achieving energy proportionality [1]. Multiple research groups have indicated that the network-equipment community lags behind on this goal of energy-proportionality. An example of such a profligate approach is optical Ethernet at 1Gb/s [2]; this approach maintains clock-synchronization between transmitter and receiver by sending a continuous stream of idle frames with a known transition density when there is no data to be transmitted. The same underlying approach remains true for 10Gb/s standards and current 100Gb/s proposals [3], [4].

II. Related work: A Current Revolution

Bolla *et al.* had produced an early survey on green networking in which they classify the base concepts of the approaches taken to date into three main categories: re-engineering, dynamic adaption, sleeping and standby [5]. Using a fusion of these approaches, it is clear that substantial energy saving can be made by changing the architecture and protocols of networked devices to allow the implementation of a low power or sleep mode applicable in idle periods. IEEE dynamic adaption mechanisms to save energy (802.3az [6]) is a mature example of such work.

There is an inevitable use of optical links within computer facilities due to high signal integrity at high bit rates and lower power consumption that is (on the scale of a machine room or data centre) independent of link distance. The use of optical approaches is a great enabler allowing switched optical networks to provide reduced hop count transactions between servers and further reducing energy requirements [7]–[9]. Quite apart from considerations of the energy which can be saved by switching off optical transceivers when not transmitting, the switched optical network scenario requires so-called burst-mode receivers[2] capable of a speedy lock onto incoming packets that have different frequency, phase and amplitude.

While conventional optical receivers use an AC coupled input stage which significantly simplifies the design of the amplification and data recovery circuits at high bit rates. However, for burst mode receivers, AC coupling results in large baseline wander (BLW). Previous burst mode receivers at low bit rates (1 Gb/s) have solved the BLW issue by using a DC coupled receiver [10] or by employing 8B10B coding [11]. 10 Gb/s burst mode operation without DC balance coding has been demonstrated [12], but as the technique requires a 20 GSample/s analog-to-digital converter (ADC) running continuously along with considerable digital signal processing (DSP), the energy characteristics are not favourable. The use of a coding scheme specifically designed for burst mode optical with a lower complexity receiver will lead to

[1] http://www.internet-project.org.uk/

[2] Optical burst mode receivers should not be confused with optical burst switching networks.

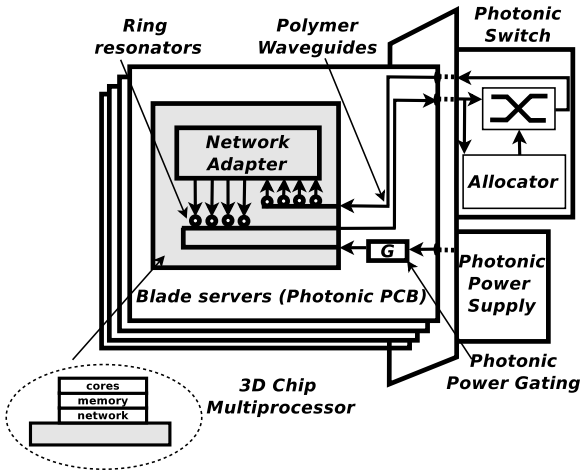

Fig. 1. Rack-scale network of 3D integrated chip multiprocessors with distributed shared memory communications

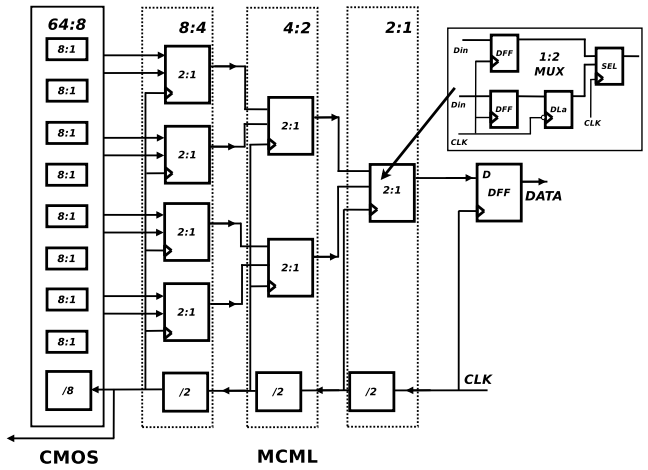

Fig. 2. Example of a 64:1 serialiser employing CMOS and MCML circuits

energy proportionality for the optical network and the ability to fully exploit low energy properties of future switched optical networks.

Alongside networks connecting routers, switches and discrete end-systems; communication end points themselves are moving onto the processor chip where the power consumption is critical, with the serial electronic transceivers which provide the several Tb/s of off-chip bandwidth required in high performance processors consuming >20% of the total power [13]. An example of a rack-scale network composed of 3D integrated chip multiprocessors with distributed shared memory is shown in Figure 1. A detailed power breakdown assessing the impact of the off-chip photonic power supply, integrated transceivers, switches and network control electronics as well as exploring strategies for power gating individual circuits is presented in [14].

Photonics has been widely proposed as one of the solutions to these energy issues and optical links have been demonstrated with significantly reduced power consumption compared with their electronic counterparts. VCSEL based links have achieved energy efficiency of 1 pJ/bit [15] while silicon photonic links with off-chip optical sources have been shown to dissipate only 0.32 pJ/bit on the processor chip [16]. However, since power consumption of an optical transceiver is dominated by other physical layer (PHY) functions such as serialisation/deserialisation (SERDES), clock recovery and line coding, a simple switch from electronic to optical transceivers will not significantly reduce power consumption without changes to the PHY. In addition, current PHYs are not energy proportional due to idle frame transmission to maintain receiver synchronization when there is no data to be transmitted. 10GBASE-T energy efficient Ethernet uses power gating in which communication can be restored on microsecond timescales [17] and a similar standard is expected for 10Gb/s optical Ethernet by 2015. However, future computer networks employing optical switching for reduced hop count between servers will require restoration times on the order of nanoseconds.

III. Context

In this paper, we consider a new approach to high-speed front-end components and their design; our work shows that the savings of a power-gated approach outweigh the additional complexity of their integration. To this end, we demonstrate in a total-power analysis, that SERDES dominates the power consumption of optical transceivers and characterise this power at bit rates from 3–25 Gb/s. We use this analysis to demonstrate that, when SERDES is taken into account, multiple wavelengths channels or higher order modulation formats can be more power efficient than a single higher rate serial channel. In addition, we quantify the power savings and restoration times of power gated SERDES for switched optical networks or energy efficient point-to-point links.

IV. Methodology

In order to investigate the energy characteristics of integrated optical transceivers, we have carried out a characterization of transceiver architectures by design and synthesis of a variety of transceiver circuit blocks [18], [19]. The models created as part of this work are freely available as part the CONfigurable Transceiver Energy

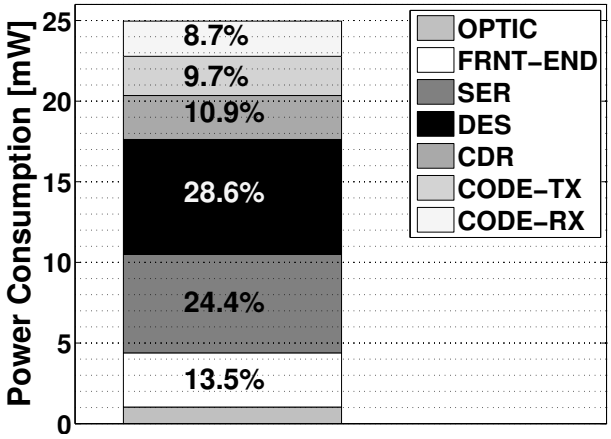

Fig. 3. Power consumption of 10 Gb/s PHY circuits of a 64B66B transceiver including front end circuits and optical power requirement

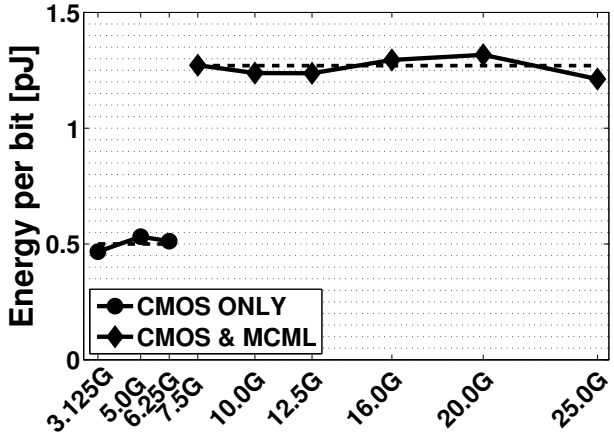

Fig. 4. Energy per bit for the SERDES circuits for bit rates of 3.125–25.0 Gb/s with (inset) the power breakdown for a 10 Gb/s transceiver

uSage Toolkit (CONTEST) [20], including spice models, optimization scripts and Verilog hardware description language code which can be synthesized with any CMOS technology library. In this paper, we use a commercially available 45 nm CMOS process. Power measurements were made using a 10 Gb/s LAN trace file as stimulus.

SERDES circuits used in traditional NRZ transceivers, which are the focus of this paper, convert between the low-speed parallel data and a high-speed serial bit stream and are implemented in a combination of static CMOS and MOS Current Mode Logic (MCML) circuit families. The parallel input frequency is held constant at 625 MHz and the SERDES ratios are varied to get the required serial bit rate. The CMOS SERDES circuits are implemented as shift registers and can be synthesised for bit rates up to 6.25 Gb/s (10:1 and 1:10 ratios) without timing issues in 45 nm CMOS process used in this work. For higher frequencies, MCML circuits were added, implemented as binary tree multiplexers constructed by cascading 2:1 multiplexer cells, frequency dividers and delay lines. An example of a serialiser which uses both CMOS and MCML circuits is shown in Figure 2.

To investigate the effect of power-gating on different parts of SERDES two different methodologies were used. Latency-critical elements, in particular the MCML circuits, were implemented using a fine-grained power-gating approach using gating transistors in every cell [21]. For CMOS circuits, a traditional coarse-grained power-switching technique, consisting of header pMOS transistors shared across the entire CMOS block was used to reduce area and power overheads.

V. Results

Figure 3 shows contribution of individual components constituting 10 Gb/s 64B66B-based NRZ transceiver to the total power consumption [19]. For the front end circuits, we use the figures obtained in a recent demonstration of record low power 10 Gb/s silicon photonic components [16]. For the optical power requirement, we assume a receiver sensitivity of -18 dBm, a typical datacom link budget of 15 dB and an uncooled laser with a wall plug efficiency of 50%. It can be observed that the front ends and laser consume only 17.6% of the power with the remainder being consumed by the PHY. SERDES consumed 53% of the total power.

Figure 4 shows energy efficiency of SERDESs for 3.125–25.0 Gb/s serial rates. It can be observed that the addition of MCML circuits for bit rates above 6.25 Gb/s causes a step change in energy consumption to an average of 1.25 pJ/bit. However, the energy per bit is relatively constant from 7.5 Gb/s to 25.0 Gb/s with the ripple due to the CMOS circuitry.

Figure 5 shows the SERDES's power consumption at 6.25, 10 and 25 Gb/s rates in four operational modes: (1) normal transmission, (2) idle frame transmission, (3) held in reset and (4) power gating deployed. Mode 3 is likely to be used in a switched network between bursts or in a point-to-point link where the increased restoration latency of full power gating cannot be tolerated. The power is reduced by 14.0% and 19.5% for the 10 Gb/s and 25 Gb/s cases respectively and by 33.3% for the 6.25 Gb/s CMOS-only SERDES, compared with the conventional transceiver which sends idle sequences. In mode 4, power is reduced to a few tens of μW but this comes at the expense of increased restoration

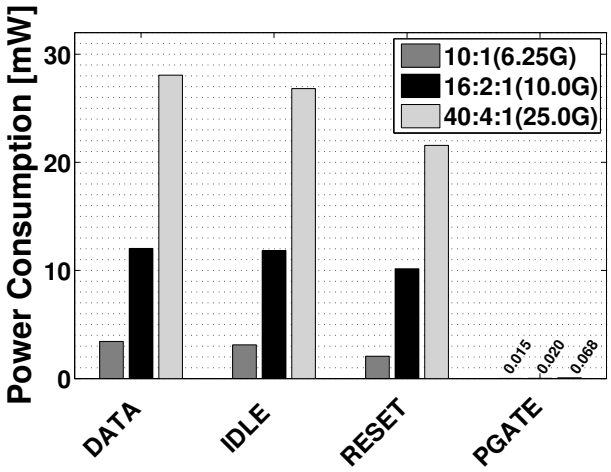

Fig. 5. Power consumption of SERDES circuits in the 4 power modes at 6.25, 10 and 25 Gb/s

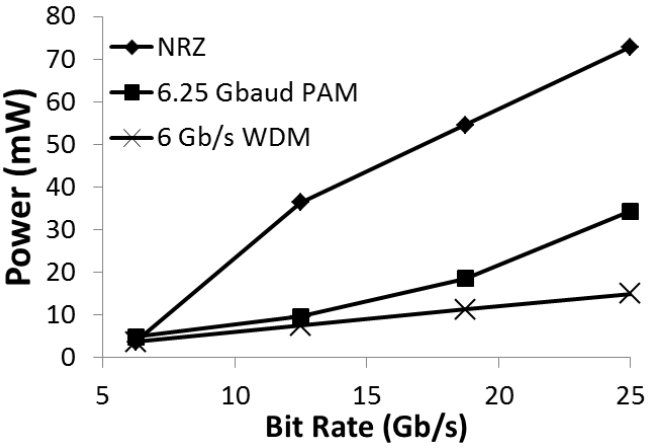

Fig. 6. Comparison of SERDES power for serial NRZ, 6.25 GBaud PAM and 6.25 Gb/s WDM

time and temporarily increased supply current during the transition to normal operation (mode 1). Based on our measurements, restoration time for 25 Gb/s SERDES is ~ 1ns, limited by the delays through the static clock dividers in the MCML binary tree network. The process of SERDES reactivation is shown in Figure 7. In the case of the 6.25 Gb/s all-CMOS SERDES, the restoration time is reduced to ~ 400 ps, a reduction of 60%. The energy required for restoration from power gating is reduced by 95.7% from 46.6 pJ for 25 Gb/s to 2.0 pJ for 6.25 Gb/s, although charge recycling schemes could reduce these figures [22].

VI. Discussion

The results presented above show that SERDES dominates overall NRZ transceiver power and there is a step change in the energy per bit above 6.25 Gb/s due to the need to introduce MCML logic. Note that this limit of 6.25 Gb/s is due to the highest rate all-CMOS SERDES circuit which can be synthesised using the 45 nm library used in this work. More advanced CMOS will achieve a higher limit. However, the results suggest that a power efficient PHY will use all-CMOS SERDES and achieve high bit rates through the use of wavelength division multiplexing (WDM) or higher order modulation formats. WDM could be attractive if future highly integrated photonics can overcome the traditionally high packaging costs, particularly where clock recovery circuits can be shared between channels. More complex modulation formats have been proposed for interconnects such as pulse amplitude modulation (PAM-N) and direct detection 16-QAM [23], although the power cost of the additional analog components was not assessed in that work. A PAM-N transceiver, compared with an NRZ transceiver of the same bit rate, has $log_2(N)$ serialisers in the transmitter and $N-1$ deserialisers (slicers) in the receiver, all operating at a frequency of $1/log_2(N)$ of the bit rate. Figure 6 shows the SERDES power for bit rates from 6.25–25.0 Gb/s comparing serial NRZ, 6.25 GBaud PAM and WDM with 6.25 Gb/s per wavelength. It can be observed that at 25 Gb/s, PAM-16 and 4-wavelength WDM SERDES blocks are shown to have 53% and 79% lower power respectively compared with serial NRZ SERDES. In interpreting these results however, it is important to note that PAM-N has a more complex analog front end than NRZ or WDM [24] and this factor is not included in the results presented here. We conclude that higher order modulation formats and WDM can potentially be more cost efficient through their reuse of existing components (everything except the physical line cards). All the while, higher-order modulation formats and WDM appears of equal power efficiency, achieving the 25-100 Gb/s line rates proposed for many future standards, e.g., 100 Gb/s Ethernet [4], [24]. In addition, power gating results presented in this paper show a clear possibility of dynamic power savings that can be achieved in both conventional and multiple-level coding transmission systems — a potential for improvement above and beyond any current physical-layer system.

VII. Conclusions

In this paper, we presented an estimate of energy use in 10Gb/s optical transceivers categorizing the contribution of physical coding, serialization and clock recovery functions to the total power profile of the system. The

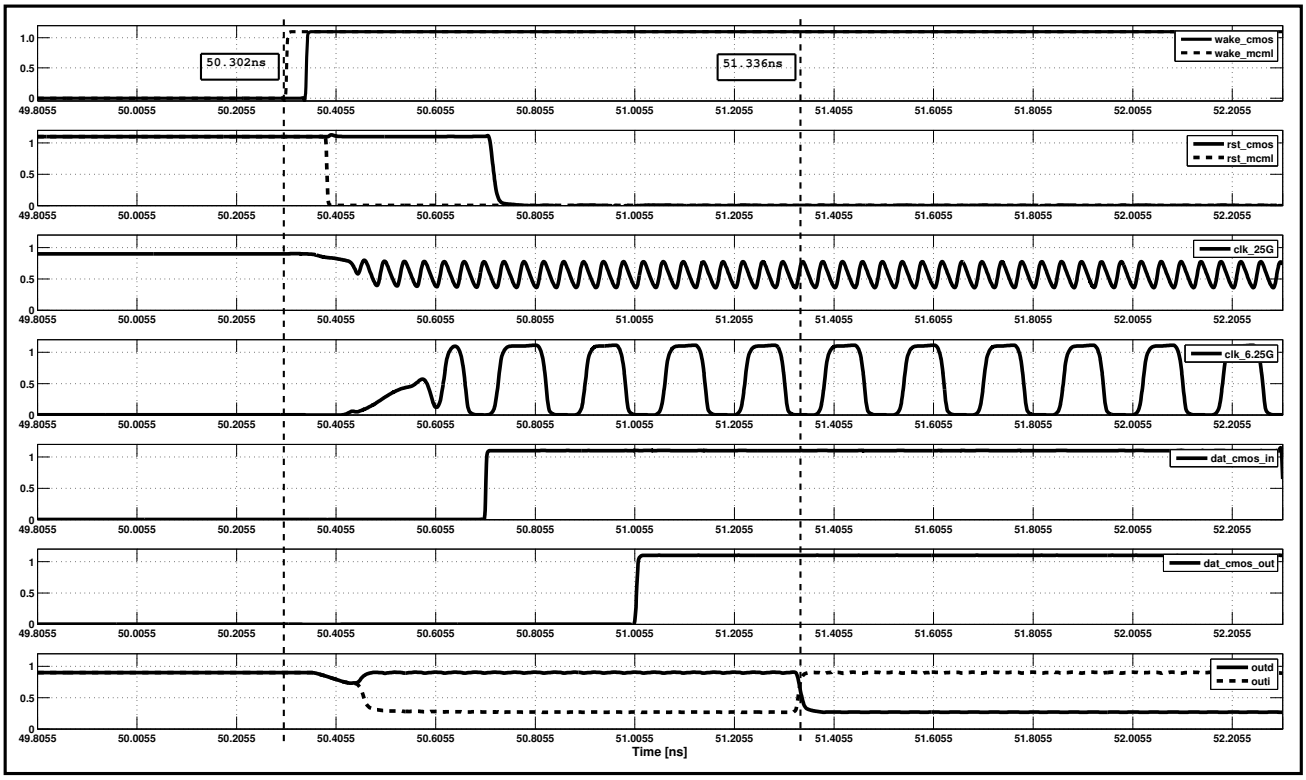

Fig. 7. Recovery of hybrid CMOS and MCML 40:1 serialiser after power gating deactivation

energy measurements performed through the use of the open-source CONTEST toolkit, allowed identification of the physical-layer sub-systems that are not energy-proportional. In particular, we found the high-speed SERDES unit to be the single largest power consumer in the transmission system. To further inform the study, we investigated the effect of power gating applied to SERDES units. The results obtained show that systems utilizing energy savings in this way will incur ~ 1ns or less of wake-up latency. In the context of future optical switching systems, note that this is comparable to the fastest optical switching latencies and considerably less than required for receiver clock recovery.

By investigating both hybrid (CMOS and MCML) and CMOS-only SERDES blocks we identified that a 60% reduction in restoration time might be achieved in all-CMOS circuits. These circuits also show, in comparison to hybrid SERDES units, a better power profile when operating at similar transmission frequencies. This suggests such an approach would be suitable for the future energy-efficient protocols utilizing complex modulation formats.

Future Work

This paper presents an early step on the road to energy-efficient networking. We do not discuss how our ideas might be best incorporated into future energy-efficient networking standards but see such consideration a useful exploitation of our work. Alongside our recent toolkit release, permitting others to make comparisons and optimizations of physical layer design [20], we plan implementation-based comparisons of our modified physical layer using a fully programmable hardware subsystem, such as that provided by the NetFPGA www.netfpga.org.

ACKNOWLEDGEMENTS

This work was supported in part by the EPSRC INTelligent Energy awaRe NETworks (INTERNET) and Unlocking the Capacity of Optical Communications (UNLOC) projects, as well as an EPSRC Career Acceleration Research Fellowship award to Philip Watts. Additionally, this research has been supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), under contract FA8750-11-C-0249. The views, opinions, and/or

findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

REFERENCES

[1] L. A. Barroso and U. Holzle, "The Case for Energy-Proportional Computing," *IEEE Computer*, vol. 40, no. 12, 2007.

[2] IEEE, "IEEE 802.3z — Gigabit Ethernet," 1998, standard.

[3] IEEE, "IEEE 802.3ae – 10 Gb/s Ethernet," 2002, standard.

[4] IEEE Proposed Standard, "IEEE P802.3ba – 100 Gb/s Ethernet," 2013.

[5] R. Bolla et al., "Energy efficiency in the future internet," *IEEE Communications Surveys & Tutorials*, vol. PP, no. 99, 2010.

[6] IEEE, "802.3az Energy Efficient Ethernet, meeting materials," IEEE, Tech. Rep., January 2008. [Online]. Available: http://www.ieee802.org/3/az/index.html

[7] G. Porter, et al., "Integrating Microsecond Circuit Switching into the Data Center," in *Proceedings of ACM SIGCOMM*, Hong Kong, China, Aug. 2013.

[8] I. H. White et al., "Scalable optical switches for computing applications," *Journal of Optical Networking*, vol. 8, no. 2, 2009.

[9] R. Luijten et al., "Viable opto-electronic HPC interconnect fabrics," in *Proceedings of the ACM/IEEE SC 2005*, 2005.

[10] Y. Ota and R. Swartz, "Burst-mode compatible optical receiver with a large dynamic range," *Lightwave Technology, Journal of*, vol. 8, no. 12, Dec. 1990.

[11] C.Su et al., ""ber performance of digital optical burst-mode receiver in tdma all optical multiaccess network"," *Photonics Technology Letters, IEEE*, vol. 7, no. 1, Jan. 1995.

[12] B.C. Thomsen et al., "Optically equalized 10 Gb/s NRZ digital burst-mode receiver for dynamic optical networks," *Optics Express*, vol. 15, no. 15, 2007.

[13] J.L. Shin et al., "A 40 nm 16-Core 128-Thread SPARC SoC Processor," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 1, pp. 131–144, 2011.

[14] P.M. Watts et al., "Energy Implications of Photonic Networks With Speculative Transmission," *OSA and IEEE Journal of Optical Communications and Networking (JOCN)*, vol. 4, no. 6, pp. 503–513, 2012.

[15] J.E. Proesel, et al., "35-Gb/s VCSEL-Based optical link using 32-nm SOI CMOS circuits," in *OFC*, 2013.

[16] X. Zheng et al., "Ultra-low power arrayed CMOS silicon photonic transceivers for an 80 Gbps WDM optical link," in *Optical Fiber Communication Conference (OFC/NFOEC)*, 2011.

[17] M. Bennett, "Network energy efficiency in the data center," in *OFC*, 2013.

[18] Y. Audzevich et al., "Efficient photonic coding: a considered revision," in *Proceedings of the 2nd ACM SIGCOMM workshop on Green networking*, ser. GreenNets '11. ACM, 2011.

[19] Y.Audzevich, et al., "Power Optimized Optical Transceivers for Future Switched or Power Gated Networks," Jan. 2013, in submission.

[20] (2013, Jan.) CONTEST - CONfigurable Transceiver Energy uSage Toolkit. [Online]. Available: http://www.cl.cam.ac.uk/research/srg/netos/greenict/projects/contest/

[21] A. Cevrero et al., "Power-gated MOS current mode logic (PG-MCML): a power aware DPA-resistant standard cell library," in *Proc. ACM DAC*, 2011.

[22] E. Pakbaznia, et al., "Charge Recycling in Power-Gated CMOS Circuits," *IEEE Trans. Comp.-Aided Des. Integ. Cir. Sys.*, vol. 27.

[23] J.L. Wei, et al., "100 Gigabit Ethernet transmission enabled by carrierless amplitude and phase modulation using QAM receivers," in *OFC*, 2013.

[24] IEEE Proposed Standard, "IEEE P802.3bj – 100 Gb/s Backplane and Copper Cable Task Force." [Online]. Available: http://www.ieee802.org/3/bj/public/