Power Optimized Transceivers for Future Switched Networks

Yury Audzevich, Philip M. Watts, *Member, IEEE* Andrew West, Alan Mujumdar, *Member, IEEE*
Simon W. Moore, *Senior Member, IEEE* and Andrew W. Moore, *Member, IEEE*

***Abstract*—Network equipment power-consumption is under increased scrutiny. To understand and decompose transceiver power-consumption, we have created a toolkit incorporating a library of transceiver circuits in 45 nm CMOS and MCML (MOS Current Mode Logic) and characterise power consumption using representative network-traffic traces with digital synthesis and spice tools. Our toolkit includes all the components required to construct a library of different transceivers: line-coding, frame-alignment, channel-bonding, serialization and deserialization, clock-data recovery and clock generation. For optical transceivers we show that photonic components and front end drivers only consume a small fraction (<22%) of total serial transceiver power. This implies that major reductions in optical transceiver power can only be obtained by paying attention to the physical layer circuits such as clock recovery, and serial-parallel conversions. We propose a burst mode physical layer protocol suitable for optically switched links that retains the beneficial transmission characteristics of 8b/10b but, even without power gating and VCO power optimization, reduces the power consumption during idle periods by 29% compared with a conventional 8b/10b transceiver. We have made the toolkit available to the community at large in the hope of stimulating work in this field.**

***Index Terms*—Energy-efficiency, physical line coding, open-source, toolkit**

I. Introduction

The persistent growth in network traffic advanced by recent developments, such as video sharing, IPTV and cloud-based storage, is causing increased demands on the network switching capacity and energy consumption at the Internet core and within datacenters [1]. Increasing the capacity of current high-bandwidth electronic switches is not only technically demanding it also leads to higher thermal-dissipation [2], [3]. This leads not only to interconnect technologies with high connectivity and capacity but lower latency, power-consumption and cost. Among these, the energy performance of networked systems has become a *1st class* property, of interest to industry and researchers. It has been shown that to make large energy savings through energy-proportionality current computer systems must be made to *do nothing well:* minimizing consumption when not in use [4]. Optical networks continue to deliver on the promise of bandwidth, latency, and low power utilization but, if optical switch fabrics are to continue meeting their promise as a key component in future, energy-proportional, systems [5], [6]; then we need a generation of high-speed transceivers designed with energy-proportionality as the 1st class property.

Context: Transceiver design has been focused upon providing high reliability with ever-higher levels of link capacity (bandwidth) to meet ever-growing needs to interconnect computer devices. This has led to optical transceivers that are always on, exchanging information to remain syncronized even when carrying no data. Such designs suit point-to-point link communications; providing implicit information about the point-to-point link status, even when no data is carried. There is a wide range of transceivers, with electronics to drive twisted pair copper, multi-channel co-axial copper and a range of optical systems. Current commercial optical 10 Gb/s transceivers have lower power consumption than twisted-pair serial transceivers due to a lower complexity physical (PHY) layer [7].[1] Yet, in an earlier study, we showed that the popular 8b/10b coding scheme can consume more power when transmitting idle frames than when transmitting data [8].

Finally, a further power-consumption incentive comes from the increasing move of communication end-points to on-chip in SoC processors. With predictions that a growing proportion of the chip will need to be power gated at any one time, the so called ”dark silicon” effect [9]. The serial electronic transceivers which provide several Tb/s of off-chip bandwidth required in high performance SoC processors are already consuming >20% of the total power [10]. Silicon photonics has been widely proposed as one of the solutions to the processor communications bottleneck and energy issues [11]. However, we show that optical transceiver power is dominated by other physical layer (PHY) functions such as serialisation/deserialisation (SERDES), clock recovery and line coding. Hence a simplistic change from electronic to optical transceivers will not reduce power consumption without an accompanying change to the PHY layer. Furthermore, at the packet-timescale an optical-switched system sets a new optical pathway to each destination. Thus the physical layer need not

Manuscript received June 13, 2013; revised September 2, 2013.

Y. Audzevich, A. Mujumdar, A. West, S.W. Moore, and A.W. Moore are with the Computer Laboratory, University of Cambridge, CB3 0FD, Cambridge, UK (correspondence e-mail: yury.audzevich@cl.cam.ac.uk).

P. M. Watts is with the Department of Electronic and Electrical Engineering, University College London, WC1E 7JE, London, UK.

This work was supported in part by the EPSRC INTelligent Energy awaRe NETworks (INTERNET) Project and an EPSRC Career Acceleration Research Fellowship award to Philip Watts.

Additionally, this research is sponsored 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.

[1]While the 10GBASE-CX4 copper standard has both a low complexity and low power consumption this is due to a distance-constrained use of four low-speed coaxial channels.

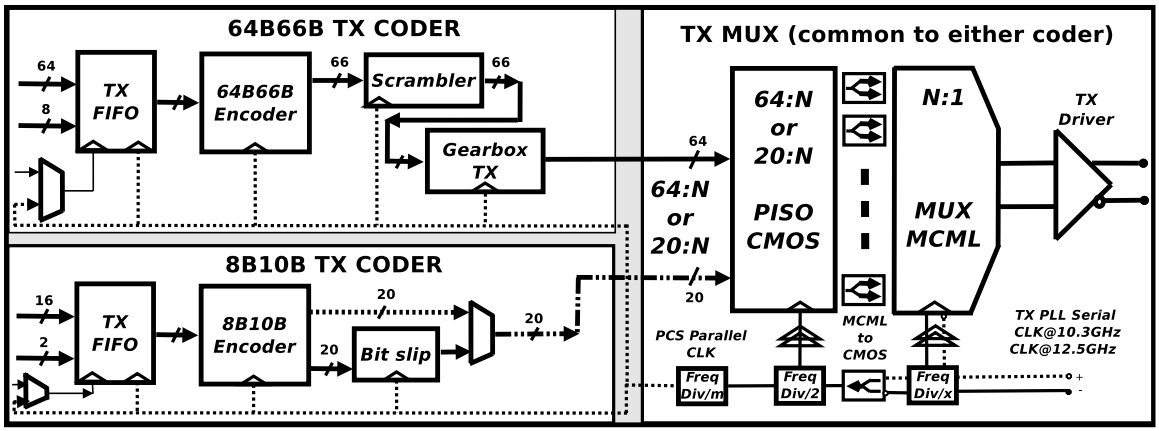


Fig. 1: Top-level diagram of the transmitter showing the two alternative coding schemes, 64B/66B with 64:1 MUX or 8b/10b attached to 20:1 MUX

remain operating when idle and, without system-wide time-syncronization, an optical packet-switch proposal uses burst-mode receivers capable of fast locking to incoming packets each having different frequency, phase and amplitude. This requires per-packet clock recovery designs in a new PHY implementation.

Contributions and Organization: In our work here we adopt a holistic approach: maintaining that the entire transceiver must be characterised to understand a total energy-requirement. The energy requirement of the physical layer of optical transceivers includes line coding, frame alignment, channel bonding, serialisation and deseralisation, clock/data recovery and clock generation. We present an exploration of the potential power savings that can be made in the PHY layer and present a proposal for power optimized optical transceivers. The methodology we adopt is a characterization of transceiver architectures by design and synthesis of a variety of transceiver circuit blocks using a commercially available 45 nm CMOS process.

In this paper we make the following contributions: (1) We characterize the power consumption of existing transceiver architectures including a breakdown of power into coding, serialisation, frame alignment and clock/data recovery functions and compare with the consumption of low power silicon front-ends; (2) We compare the power characteristics of serial transmission with multi-lane designs to identify a power-optimal combination of time division and wavelength division multiplexing; (3) We characterize the power consumption of burst mode transceivers and quantify the power savings which can be made compared with current continuous transmission protocols; (4) Alongside an open-source MCML-based cell library which we make available to the research community, we present a method for its' power optimization;

The rest of the paper is organized as follows. Sections II and III outline the design and synthesis of the CMOS and MCML circuit blocks which make up our transceiver models. Section IV presents the results including a comparison of the transceiver circuit power consumption with recently reported silicon photonic front end circuits and optical power requirements; characterisation of the SERDES power consumption against bit rate; and characterisation of the line coding power consumption under various operating conditions. Section V discusses the potential for power saving in future burst-mode compatible optical transceivers. It also gives further details of the transceiver designs made available as a result of this project. Finally, we end with a conclusion of the high-level lessons from this work.

We describe in Appendix A: MCML Design and Optimization information to enable researchers to recreate our efforts and as a worked example that, while central-to this work, does not constitute its principal contribution.

CONTEST Toolkit

We make available the 10 Gbps transceiver's implementation as part of CONfigurable Transceiver Energy uSage Toolkit (CONTEST) [12], a toolkit that contains a set of Verilog, HSpice and Tcl scripts/code modules, which permit an automated component optimization/design process using standard 45nm CMOS technology library. This can also be extended for the designs synthesised with any other CMOS technology process.

II. Transceiver Design

This section describes the main functional circuit blocks of the transceiver. Top-level representations of the transmitter and receiver are shown in Figure 1 and Figure 2 respectively. Initially, the transceiver design was aimed at a payload bit rate (before adding coding overhead) of 10 Gb/s. In later work, we characterized the circuits at different bit rates to investigate optimum bit rate versus power operating points.

A. Line Coding

The functions of the coding block include DC balance, byte alignment within the serial stream and error detection. We consider two popular encoding schemes: 8b/10b block code and the scrambler-based 64B/66B.

8b/10b represents a class of parity-disparity DC-balanced codes that map arriving 8-bit symbols into a 10-bit code words

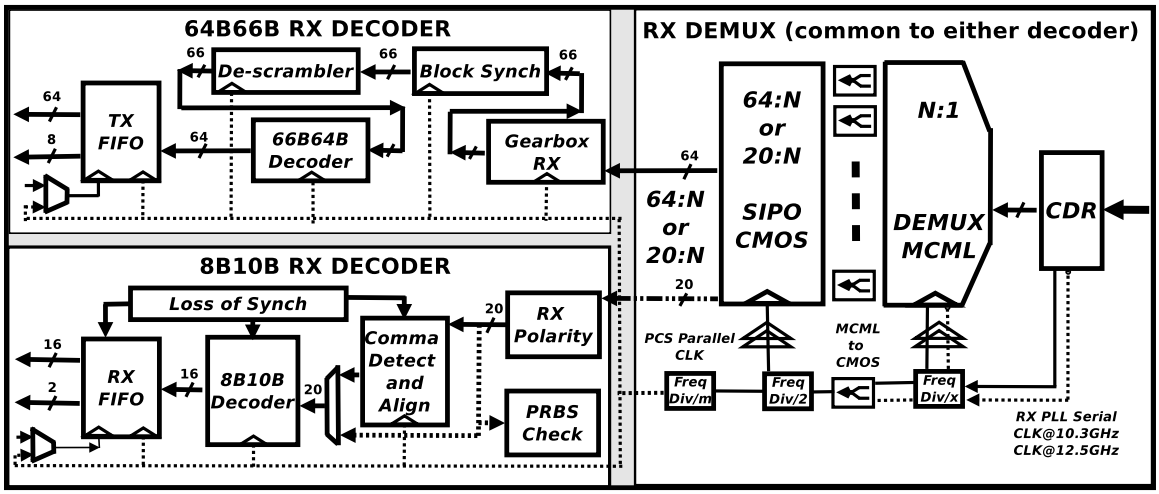


Fig. 2: Top-level diagram of the receiver showing the two alternative coding schemes, 64B/66B with 1:64 DEMUX or 8b/10b attached to 1:20 DEMUX

using predefined code groups at run time on the word-by-word basis [13]. The code limits the run length of identical symbols in order to remove baseline wander in AC coupled receivers and guarantee the required transition density for clock synchronization. 8b/10b coding has excellent transmission properties but has a bit rate overhead of 25%. For this reason, the hybrid-scrambled 64B/66B encoding scheme was selected for 10 Gb/s Ethernet which reduces the overhead to 3%. The encoding module performs a framing function by transforming the 64-bit data and 8-bit control inputs into a 66-bit block [14]. Each 64-bit data word is scrambled with a 58th degree polynomial to ensure statistical DC-balance and transition density and a 2-bit synchronization header is appended to allow frame detection and alignment to be performed.

Figure 1 shows block diagrams of the two alternative coding schemes in the transmitter. In the 64B/66B case, the transmitter accepts 64-bit data at 156.25 MHz and carries out encoding and scambling. The resulting 66-bits are converted by the gearbox to 64-bit interface at 161.13 MHz for more efficient serialization. In the 8b/10b case, we implemented versions with both 8-bit wide client-side data interface running at 1.25 GHz and a dual encoder with a 16-bit interface operating at 625 MHz. Figures 1 and 2 show the 16-bit interface. In all cases, phase differences between the coding block and client side interface are compensated by a FIFO buffer. The receiver side of the coding circuits perform decoding using the same clock frequencies and interface widths and in addition perform frame/byte alignment and error checking (Figure 2).

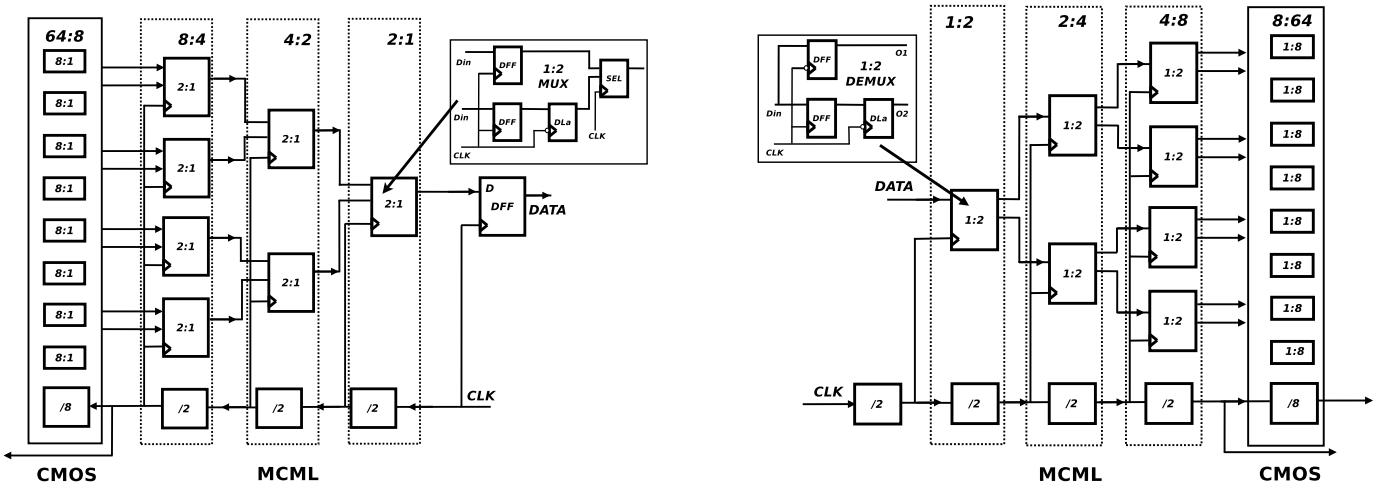
B. Serialization and Deserialization

The serialization and deserialization (SERDES) circuits convert between the low-speed parallel data and a high-speed serial bit stream. The multiplexing ratios depend on the coding scheme used. In the case of 64B/66B, 64-bit sequences at 161.13 MHz are converted to 10.3125 Gb/s using a 64:1 ratio. In contrast, a transceiver with 8b/10b coding performs either 10:1 or 20:1 multiplexing producing a line rate of 12.5 Gb/s. As shown in Figures 1 and 3, the SERDES circuits are implemented in a combination of static CMOS and MCML. In order to find a power efficient SERDES design we engineered a variety of configurations. For example, for 64B/66B, we investigated 64:1 SERDES based on 64:N CMOS and N:1 MCML circuits where N = 2,4 or 8 (referred to throughout the rest of the paper as 64:N:1). In a similar way, we investigated 8b/10b SERDES using 20:N:1 (dual encoder) and 10:N:1 (single encoder) cases, with N=2 or 4.

The CMOS SERDES circuits are implemented as shift registers. The MCML circuits were implemented as binary tree multiplexers constructed by cascading 2:1 multiplexer cells, frequency dividers and delay lines, which were manually optimized for the required bandwidth and timing operation. Figure 3 shows an example of a 64:8:1 SERDES. The detailed design process for CMOS and MCML circuits is presented in section III.

C. Transmitter PLL and Clock/Data Recovery

For fast locking in burst mode and good jitter performance, a low phase noise single stage CDR circuit was designed as shown in Figure 4. The design considers: (1) the input gain stage, which buffers the input data signal and isolates the channel by inserting a high impedance component in the path, (2) the frequency doubler unit, which allows for the data signal to be delayed, phase shifted and then added back to the original signal (to effectively double the number of transitions), and, (3) the feedback loop stage (implemented as an XOR gate), which is used for synchronizing outputs of the Voltage Controlled Oscillator (VCO) and the frequency doubler units. The output of this stage forms a periodic train of pulses where the duty cycle determines the phase errors. The rectification of the signal is achieved by the appropriate (4) Low Pass Filter design, which, in its turn, controls (5) the VCO unit by a slow varying sinusoid signal with a certain DC offset. The design was synthesised in hardware using the CMOS standard cell library for the low frequency components,



(a) (b)

Fig. 3: (a) A 64:1 MUX block diagram. (b) A 1:64 DEMUX block diagram.

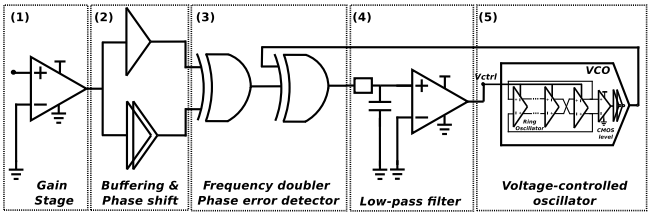


Fig. 4: Clock and data recovery unit

passive components for the low pass filter and, a MCML-based ring oscillator shown in Figure 4 [15].

Commonly used PLL and CDR circuits often use multiple stages in order to facilitate a stable and consistent operation. This redundancy usually delivers high performance but the synchronization process takes a relatively long time to achieve a stable lock. The simplicity of our CDR design guarantees a fast locking time (≤ 10 clock cycles) and maximum power and area efficiency. Although realistic CDR implementation may require some modifications to the design to account for factors such as minor impedance mismatch, capacitive and inductive resistance variations etc, we believe that the power figures will be representative of real circuits. The CDR circuit used in this work has a power to frequency ratio of around 0.25 mW/GHz and thus compares very favorably with reported circuits based on a dual loop with feedback interpolation (2.2 mW/GHz) [16], injection locking (5.1 mW/GHz) [17], DLL with analogue phase interpolator (5.2 mW/GHz) [18] and phase rotator PLL with software control loop (1.0 mW/GHz not including the processor and software platform) [19] even when adjusting for the different CMOS processes used.

D. Channel Bonding

In order to find the power consequences of using multiple lower bit rate serial streams rather than a single serial channel, we designed a channel bonding circuit in Verilog which eliminates skew between multiple channels using a separate FIFO. In an optical link, these channels could be either space or wavelength division multiplexed. We tested the circuit operating on the output of two 8b/10b client side streams, but the Verilog model is parametrised for higher numbers of channels. The circuit is designed for burst mode operation.

III. Circuit Design

Whereas the low frequency coding circuits were implemented using static CMOS, the SERDES and CDR used a combination of CMOS and MCML logic families in order to optimize operating frequency and power consumption. This section describes the method used for the design of both logic families.

A. Design of CMOS Circuits

Design of the static CMOS circuits started with Register Transfer Level (RTL) Verilog Hardware Description Language (HDL) descriptions and synthesised using Synopsys Design Compiler with a commercially available 45 nm standard cell library. Constraints were set to minimize power consumption at the required operating frequency. The typical clock frequency margin used for synthesis is considered to be at least 15% faster than the nominal frequency value. The synthesised Verilog netlist was simulated using Mentor Graphics ModelSim to verify correct operation and store activity data for dynamic power analysis. The input stimulus for the simulations was extracted from realistic 10 Gb/s Ethernet trace files and analysed under (1) continuous data transmission and (2) continuous idle transmission input setups. Synopsys PrimeTime was used to generate power consumption data for each circuit block.

B. Design of MCML Circuits

Although new generations of CMOS technologies continuously improve their performance and power characteristics due to scaling, CMOS circuits are prone to generate a high level supply noise while operating at high speeds [20]. The noise factor limits the on-chip integration of digital blocks with their analog counterparts. Logic families with differential-signalling

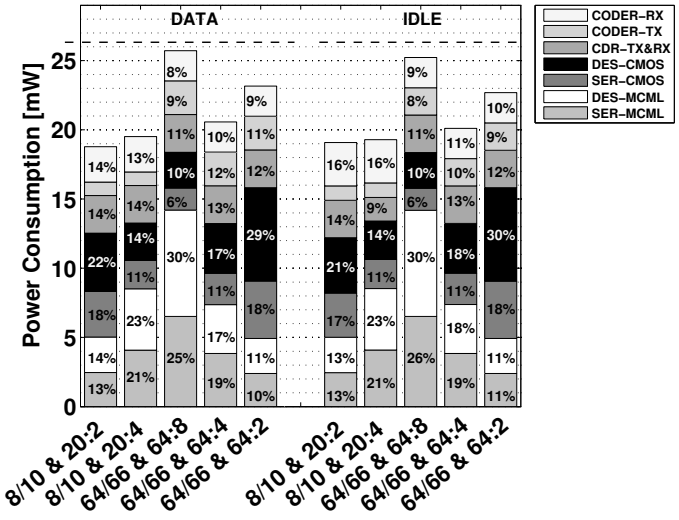


Fig. 5: Power breakdown for 10 Gb/s transceivers with differing SERDES configurations (electronics only)

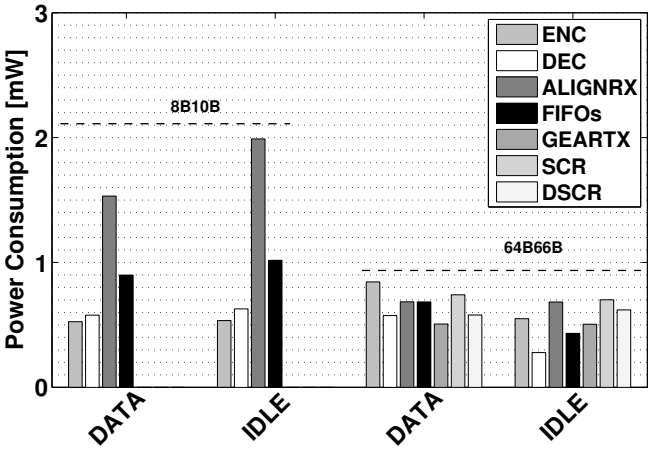


Fig. 6: Coding block power

such as MOSFET Current Mode Logic (MCML) [15], [21], [22], [23] are characterized by an improved noise immunity and high-speed operation. The speed advantage is achieved by the fact that the current, generated by a constant current source, is steered between a pair of fully differential transistors and produces a reduced-swing voltage drop at outputs (in combination with specific voltage gains), reducing the generation of logic level switching noise. It must be noted though that the presence of the current sink implies a constant power dissipation irrespective to the operating frequency or input sequence applied. Power dissipation in MCML circuit is dominated by a static power ($P = V_{dd} \times I_{ss}$) and is independent of the operating frequency.

In this work an MCML cell library was developed. The design process used the transistor models supplied with the 45 nm CMOS standard cell library and a semi-analytical methodology developed in HSPICE environment for cells optimization. To satisfy the required performance criteria of high-speed operation and minimize power dissipation of individual gates, we used HSPICE optimization solver. This allowed us to produce the best case parameter variation model for a specific subset of supply voltages, voltage swings and biasing currents selected as the input characteristics. Appendix A describes the design and optimization process for the MCML cell library in detail. Once the MCML cell library optimization process was complete, design of serialization, deserialization and CDR circuits was performed. Correct operation was verified and power measured using SPICE simulation.

IV. Power Consumption Analysis

Firstly we characterize the power consumption of the main functional blocks of a 10 Gb/s transceiver to identify the main power sources and optimize the configuration. Figure 5 shows the breakdown of transceiver power for a 10 Gb/s transceiver with standard 8b/10b and 64B/66B coding for both data and idle signals. The effect of various SERDES ratios is also shown: for 64B/66B we consider a 64:8 CMOS circuit combined with 8:1 MCML circuit to achieve the 64:1 SERDES (henceforth called the 64:8:1 case) as well as the 64:4:1 and 64:2:1 cases. It can be observed that the optimum ratio depends on the input frequency, with a 20:2:1 ratio being optimum for 8b/10b transceivers with a 625 MHz input frequency, but 64:4:1 ratio minimizes the power in the 64B/66B case with a 156 MHz input frequency. In subsequent results, these optimum ratios will be used.

It is useful to note that in Figure 5 the power consumption of each subsystem in each design is represented as a percentage of global consumption for that design; thus modest differences in the coding and decoding subsystems will be dominated by the total design-consumption. From this figure it is also clear that current transceivers do not display energy consumption proportional to use. It is clear that the SERDES and CDR components, operating at line-rate, and being independent of the type of data carried (idle vs. data); consume near-identical power. This figure graphically illustrates how much power idle frame transmission can consume despite no work being done; a breakdown of the coding subsystem alone is illustrated in Figure 6.

As previously reported [8], when generating 8b/10b idle frames the coding block actually consumes greater power. This is shown by the slight increase in coding power consumption in the 8b/10b idle cases in Figure 5 and the significant difference in coding block subsystems shown in Figure 6. This is because, for 8b/10b idle ordered sets cause the disparity control check every octet to invoke extra logic, whereas the data sequences may use more balanced words than unbalanced leading to lower power consumption. As shown in Figure 6, it is the word boundary alignment function in the receiver that has the greatest impact on 8b/10b codec power consumption. The 64B/66B coding scheme has slightly higher power consumption for data sequences compared with idle sequences (up to 35%), but consumes higher power than 8b/10b in both cases. The main conclusion from Figure 5 is that SERDES dominates the 10 Gb/s transceiver power

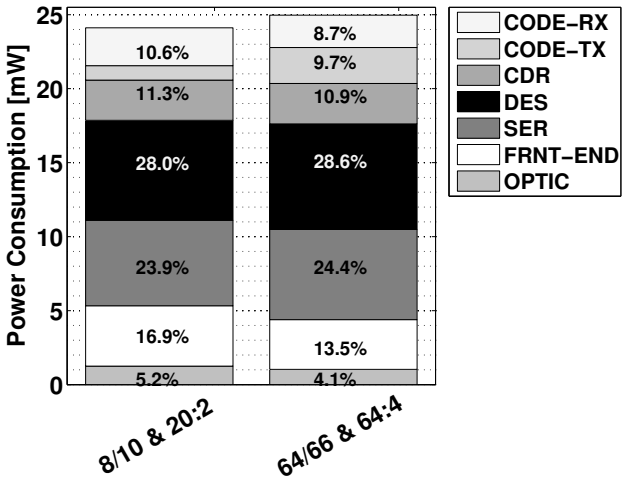


Fig. 7: Distribution of 10 Gb/s transceiver power consumption including front ends circuits and laser with 8b/10b and 64B/66B coding

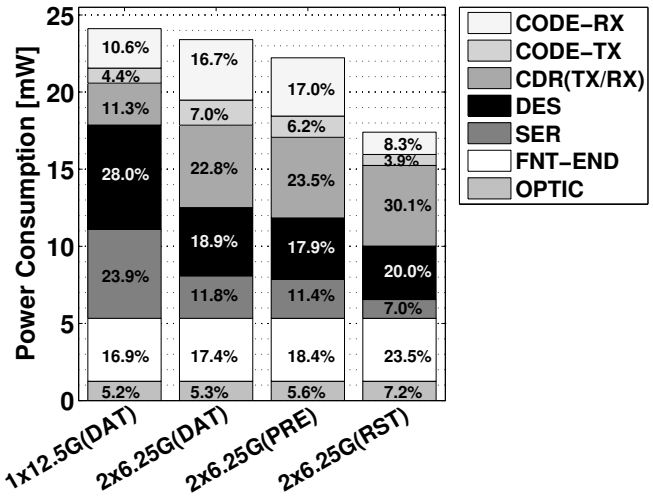


Fig. 8: Power breakdown for 6.25 Gb/s optically switched transceivers

consumption whichever configuration is chosen.

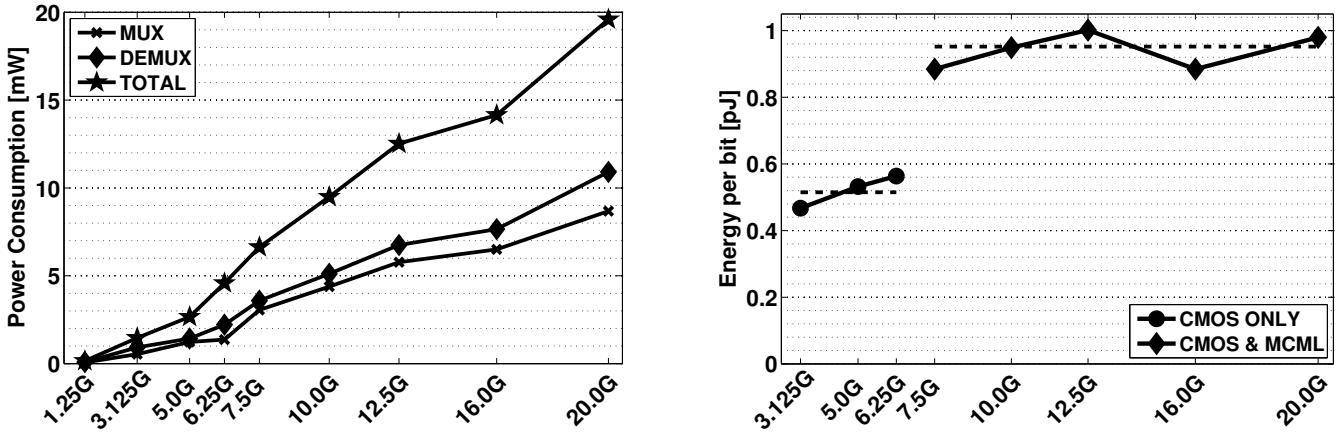
Figure 7 shows a comparison of the power consumption of the main transceiver blocks with that of the front end circuits and optical power requirements. For the front end circuits, we use the figures obtained in a recent demonstration of record low power 10 Gb/s silicon photonic components [11]. This demonstration uses silicon ring resonator modulator transmitter front end (minimum power including drivers of 0.66 mW) and silicon germanium photodiode receiver front ends (2.6 mW minimum). The power figures include transmitter drivers and receiver amplifiers with hybrid bonding to the silicon photonic circuits. For the optical power requirement, we assume a receiver sensitivity of -18 dBm [11], 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 22.1% of the power with the remaining 77.9% being consumed by the transceiver PHY. The SERDES power consumes 52% of the total power.

As SERDES has been shown to dominate total transceiver power including front end and optical sources, we now examine the optimum bit rate from power considerations. Figure 9a shows the power consumption of the serialisation and deserialisation circuits for serial bit rates between 1.25 Gb/s and 20 Gb/s. In each case, the parallel input frequency is constant at 625 MHz and the multiplexer/demultiplexer ratio is varied to get the required serial bit rate. We were able to synthesise CMOS only multiplexer and demultiplexer circuits up to 6.25 Gb/s (10:1 multiplexer and 1:10 demultiplexer ratios) without timing issues using the 45 nm standard cell library. 2:1/1:2 or 4:1/1:4 MCML circuits are required in addition to CMOS in order to achieve higher bit rates. Figure 9b shows the energy per bit at each bit rate. In can be observed that CMOS only circuits achieve an average of 0.5 pJ/bit from 3.125 Gb/s up to 6.25 Gb/s. The 1.25 Gb/s (2:1) CMOS configuration is a special case as this synthesises to a simple 2:1 multiplexer, rather than a shift register used at higher bit CMOS circuits and has an energy per bit of only 0.12 pJ/bit. The CMOS and MCML circuit configurations have a higher average energy of 0.95 pJ/bit at bit rates from 7.5 Gb/s up to 20 Gb/s.

V. Discussion

Dominance of SERDES power and the higher values of energy per bit found for combined CMOS and MCML circuits means that the optimum bit rate for serial transceivers is the highest frequency which can be synthesised for CMOS shift registers which we call f_E. In the case of the 45 nm standard cell library used in this work with an input frequency of 625 MHz, this optimum bit rate, $f_E = 6.25\ Gb/s$. As optical, front end and coding power scale linearly with bit rate (constant energy per bit), f_E also gives the optimum combination of time division multiplexing (TDM) and wavelength division multiplexing (WDM) for a given required aggregate bit rate. Full custom design of the multiplexer circuits could push f_E to higher values at the expense of increased engineering costs. Comparisons with other transceivers reported in the literature are difficult due to the wide range of serialization ratios used. However, the 16 Gb/s transceiver described in [16] uses a high proportion of 90 nm CMOS-style circuitry and only 5:1 serialization ratio, but does not achieve lower power than reported in this work, even adjusting for the bit rate and the effect of CMOS scaling to 45 nm.

Although coding power has been shown to be a relatively small proportion of the total power, it has important system implications. The functions carried out by the coding scheme are DC balance, frame alignment within the serial stream and error detection. For burst mode transceivers using a per packet preamble for clock recovery and word alignment (as would be used in a switched network or for power gated links), minimum sized Ethernet packets (64B) can be transmitted with $\leq 0.3\ dB$ power penalty using an AC-coupled receiver with a typical low frequency cut-off of 100 kHz [24]. However, larger Ethernet packets require DC balance for acceptable penalties. As transceivers become ever more integrated with processor and router logic, DC-coupled transceivers will be favoured



(a) (b)

Fig. 9: (a) Serialisation and deseralisation power against serial bit rate with a client side parallel interface running at 625 MHz; (b) Energy per bit for the serialisation and deseralisation circuits combined, showing the differences between simple 2:1 CMOS components, CMOS only shift registers and combined CMOS and MCML circuits.

as large coupling capacitors are difficult to fabricate on-chip. However, for compatibility reasons, reducing jitter as well as to retain the alignment and error detection capabilities some form of coding is beneficial.

Although 64B/66B coding reduces the bit rate of the system, in turn reducing optical and front end power, the total power consumption is higher than for an 8b/10b coded transceiver due to higher SERDES ratios and higher power coding circuits. In addition, 64B/66B is not compatible with optically switched or power gated systems due to long and unbounded frame alignment times. The 8b/10b coding scheme on the other hand has been shown to be low power and requires only 80 bits to obtain frame alignment. However, 60% of coding block power for data and 66% for idle is used for the barrel shifter and finite state machine which carries out the frame alignment function.

Finally, based on the results presented in the previous section, we propose a codec based on 8b/10b optimized for low power in burst mode applications. The bit rate per wavelength is chosen to be 6.25 Gb/s (multiplexer ratio of 10:1) to eliminate MCML circuitry from the SERDES and hence minimize SERDES energy per bit. In this scheme, we need two wavelength channels per 10 Gb/s. Although, high bandwidth systems would use many wavelengths, we compare a two wavelength burst mode system (2 x 6.25 Gb/s) with a single serial channel (1 x 12.5 Gb/s) as shown in Figure 10. Although both cases require asynchronous FIFOs at the receiver output, the two wavelength case requires additional logic for channel bonding. To meet the burst mode requirement, transmission only takes place when data is available with a preamble to regain synchronization between transmitter and receiver. The preamble consists of a single DC balanced (zero disparity) 10-bit code word, suitable for both clock recovery and receiver frame alignment, which is transmitted for a predetermined time. Hence the transceiver has three states which we label: data, preamble and reset. The 8b/10b encoder/decoder blocks are disabled during preamble and reset, in other words their inputs are held so that no dynamic power is dissipated. Further reductions in power consumption can be achieved by power gating circuit blocks while not in use, although the benefits are critically dependent on the achievable wake up times and the packet inter-arrival times in the workload. We leave consideration of this for future work. However, the continued use of 8b10b coding allows multiple packets to be transmitted with a single preamble sequence without DC balance problem during periods of heavy communication load. At the end of the preamble time, the 8b/10b circuits are enabled, the frame aligner is disabled and data transmission begins. An end of packet or invalid received code word causes the receiver to return to preamble state. If no new packets are waiting at the transmitter at the end of a packet transmission, the transmitter returns to reset state with the encoder disabled and the serializer input set to all zeros. Similarly, if no optical power is received for a predetermined number of cycles, the aligner and decoder are disabled.

For typical sized Ethernet packets transmitted in burst mode, Figure 8 demonstrates the reduction in power consumption using the burst mode 2 x 6.25 Gb/s transceiver using the proposed protocol compared with a conventional 12.5 Gb/s transceiver using the 8b/10b protocol. For the burst mode transceiver, the reduction in power compared with the conventional transceiver is 2% for data, 8% for preamble and 29% in the reset state. Large savings in SERDES are partially offset by a higher receiver coding block power due to channel bonding. However, these reductions are achieved despite the fact that we still use MCML buffers in the 6.25 Gb/s ring oscillator based VCO circuits leading to the CDR being the largest power source for the burst mode transceiver. Replacement of the MCML ring oscillators with a CMOS design in the burst mode design will lead to further power savings in all three states. These results show that, even without power gating, large power savings can be made particularly during the periods of low communication load. Future work will study power savings based on realistic traffic profiles in large computer facilities such as data centers.

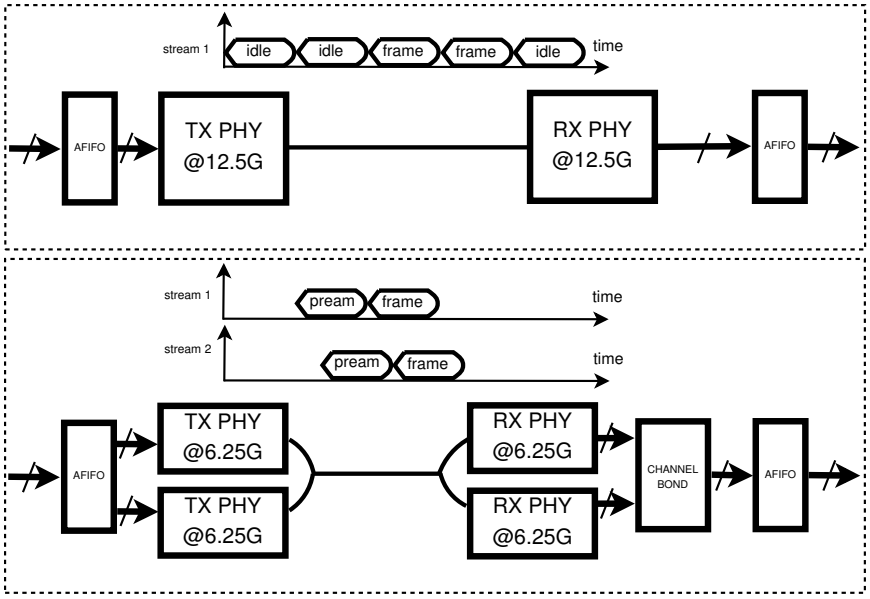


Fig. 10: Differences between the two 12.5 Gb/s optical links compared (a) conventional serial link (b) burst mode transceiver suitable for power gated or switched optical links using two wavelength channels at 6.5 Gb/s each.

A modest proposal for future transceivers

We conclude to consider power-consumption as a 1st class property motivates a reconsidered approach to the front-end serialization and coding sublayers; tuning choices to suite the bursty network traffic with relatively low levels of utilisation observed in many local-area and data-center networks. Properties of such implementations compatible with switched (optical) networks including those that incorporate power-gating include **1.** minimizing the number of MCML circuits, for example, by removing them from SERDES systems and restricting MCML use to the CDR circuit only; **2.** making codec choices that will lead to lower power implementations, even when this requires a higher channel speed for the same achieved data-rate; **3.** minimizing clock recovery periods such as through the use of a single-stage PLLs; and **4.** making a codec choice that facilitates fast frame alignment (e.g., 4 frames for 8b/10b versus at least 16 frames for 64B/66B), thereby permitting longer idle periods.

Our proposed burst mode physical layer protocol (Section V) suitable for power-gated optically switched links, retains the beneficial transmission characteristics of 8b/10b and even without power gating and VCO power optimization, reduces the power consumption during idle periods by 29% compared with a conventional 8b/10b transceiver.

VI. Conclusion

We show the power-consumption breakdown for an entire communications transceiver. We note that, as ultra-low energy silicon photonic communication components become commonplace, the power consumption of the other transceiver components must become the focus for major reductions in transceiver power. Such reductions can only be obtained with attention to the physical layer circuits and protocols of which SERDES is the largest component. Our results show that the high-speed sub-system, incorporating SERDES, CDR, and clock recovery, can, despite relatively simple logic, consume 50–60% of the total power. This is largely due to the integration of standard CMOS and differential MCML components operating at a high clock-rate.

As part of our discussion we enumerate a set of guidelines for future transceiver design and to facilitate continued work in this field we provide the 10 Gbps transceiver's implementation as part of CONfigurable Transceiver Energy uSage Toolkit (CONTEST) [12], promoting direct comparison by enabling other researchers to reproduce our results thereby permitting meaningful comparison.

Thanks

We thank the following people for their useful discussions, insights and comments on early drafts of this paper: Noa Zilberman, Robert Mullins, Jaafar Elmirghani, Adrian Wonfor, Matthew Grosvenor, Robert N. M. Watson and Jon Crowcroft. Additionally, we thank the anonymous reviewers and our sub-editor each of whom made suggestions that have greatly improved this paper. All errors remain our own.

Appendix A: MCML Design and Optimization

Design of MCML circuits requires optimization of a large number of parameters. Previous work in the field provided an analytical description of all parameters used in the MCML logic design process, and, reviewed the impact of these on performance/power response [15], [21], [22], [23]. In this work we developed an optimization toolkit, which allows deriving an MCML cell-library parameters in automated way via using a standard SPICE descriptions of MOSFET transistors and satisfying the specific criteria in power efficiency and performance measured as system's outputs. In the following section we review the major operation principles and properties of a

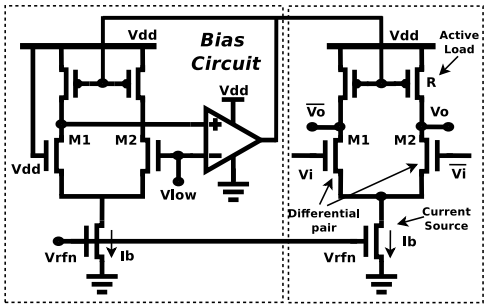


Fig. 11: MCML inverter cell

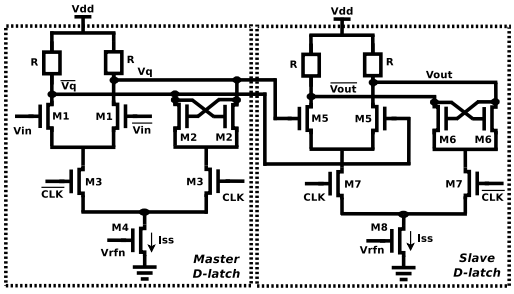


Fig. 12: MCML D-type flip-flop cell

typical MCML cell and provide the optimization procedure used throughout the cell development process.

MCML design parameters and operation

A typical MCML gate is composed of three main blocks (as shown in Figure 11): the pull-up network, implemented as a set of resistors or active pMOS loads, the fully-differential pull-down network, which steers the current between the branches, and the current source. The performance of a gate is a function of various metrics (voltage gain, voltage swing, and others, please refer to Table II), and is determined/evaluated by the corresponding adjustments made in transistor sizing, biasing voltages, reference currents/voltages and differential voltage swings. A complete list of variables that is typically used in the optimization process is presented in Table I. The complete list of power-performance-related metrics is presented in Table II.

The operation of a standard MCML inverter cell can be described as follows. Due to presence of active loads R, a voltage drop $\Delta V = I \times R$ is produced, permitting logical 1 and 0 states to be represented as Vdd and $Vdd - \Delta V$ voltages respectively. The use of active loads, implemented as pMOS transistors conducting in the linear region (assumed to provide a roughly linear transfer function response), allows online adaptability that helps compensating any spontaneous variations inside the circuit. Typical resistance values are on order of 10s of $K\Omega$s and require sink currents to be in the order of couple of hundreds μAs. The increase in transistor sizing, i.e. $W_{\rm P}/L_{\rm P}$ ratio, lowers the load resistance, and, as a rule, propagation delay of inverter circuit; it is also followed by reduction in saturation voltage of the pMOS loads causing degradation in linear response. An example of biasing circuit that is used for parameter's adjustment is given within Figure 11.

In general, signal steering of a typical MCML gate is performed by pull-up and pull-down networks formed of pMOS and nMOS transistors. Topological representations of nMOS differential networks which contain either a single (Figure 11) or a multi-level logic (Figure 12), define the logical function of the particular cell. It should be noted that performance characteristics of MCML gates linearly degrade with the logic depth of the pull-down network; this property should be accounted in the high-speed MCML designs. The nMOS network design structure of the commonly used MCML cells is found in [25], [21], [26], etc.

In order to keep all the transistors of the differential pair in their saturation region, a reduced input signal level is produced

TABLE I: Classification of input variables

Parameter	Description
Vdd	Supply voltage
ΔV	Voltage swing
$W_{\rm Ni}$ $L_{\rm Ni}$	Width and Length of nMOS transistors of pull-down network
$W_{\rm P}$ $L_{\rm P}$	Width and Length of pMOS transistors of pull-up network
$W_{\rm NS}$ $L_{\rm NS}$	Width and Length of nMOS transistors in the current sink
$I_{\rm ss}$	Current value produced in the current sink
$V_{\rm Rfp}$ $V_{\rm Rfn}$	Biasing voltages controlling active load resistance and sink currents
$C_{\rm load}$	Output capacitance value

for the deeper logic levels of the circuit [20]. In general, input level-shifting can provide the appropriate signal level reduction but is detrimental to the gate's delay, power and area. On the other hand, this type of level-shifter enables independent cell's delay estimation and makes the automated MCML circuit design possible. To facilitate the optimization process, we assumed that transistor dimensions of all the input-type level-shifters match the ones used inside the differential pair of the gate itself.

The amount of sink current $I_{\rm ss}$ produced inside a gate, is characterized by biasing voltages $V_{\rm rfn}$ applied to the nMOS current source transistor. This value is typically kept in order of couple of hundred millivolts and is derived from a reference current produced by a current-mirror circuit [21] with a specified current matching ratio. The increase in $V_{\rm rfn}$ voltage allows more current flowing through the circuit, which also improves switching speed. The size of the current source transistor is chosen to reduce its saturation voltage; larger ratios of $W_{\rm NS}/L_{\rm NS}$ improve the total output resistance, but increase area and power dissipation.

The variation in transistor sizing, their biasing and input voltage swing levels affect the quality of the output signal. The parameter variation of pull-up (pull-down) network may introduce asymmetry in signal response, affecting signal regeneration and resulting in degraded signal rise/fall times and propagation delay. The Signal Slope Ratio (SSR) metric is designed to control the quality of the output waveforms produced by defining the ratio between rise/fall times $t_{\rm rf}$ and propagation delay $T_{\rm d}$ [15]. The other parameters, the mid-swing DC voltage gain $A_{\rm v}$, Noise Margin (NM), and the Voltage Swing Ratio (VSR) directly affect output's signal quality. The midswing DC voltage gain ($A_{\rm v}$) metric is a key parameter controlling signal regeneration in cascaded circuits.

TABLE II: Problem statement for MCML circuit optimization

Satisfy: $t_d = t_{\mathrm{REQUIRED}}$ Minimize: Power dissipation P_d
Satisfy performance constraints: $A_v \geq 1.4$, $NM \geq 0.4\Delta V$ $SSR \leq 6$, $VSR \geq 0.95$
Optimize and record: Vdd, ΔV, W_{Ni}, L_{Ni}, W_{P}, L_{P}, W_{NS}, L_{NS}, I_{ss}, V_{Rfp}, V_{Rfn}
Limit: $Vth + \Delta V \leq Vdd \leq Vdd_{\mathrm{CMOS}}$ $150mV \leq \Delta V \leq Vth$
$2W_{\mathrm{Nmin}} \leq W_{\mathrm{Ni}} \leq 3\mu m$ find the smallest of $L_{\mathrm{Ni}} \geq L_{\mathrm{Nmin}}$
$2W_{\mathrm{Pmin}} \leq W_{\mathrm{P}} \leq 3\mu m$ find the smallest of $L_{\mathrm{P}} \geq 2L_{\mathrm{Nmin}}$ $5mV \leq V_{\mathrm{Rfp}} \leq Vth$
$2W_{\mathrm{Nmin}} \leq W_{\mathrm{NS}} \leq 5\mu m$ $2L_{\mathrm{Nmin}} \leq L_{\mathrm{NS}} \leq 2\mu m$ $400mV \leq V_{\mathrm{Rfn}} \leq Vdd$ $I_{\mathrm{ss}} = I_{\mathrm{REFERENCE}}$

Usually, the voltage gain is optimized to be above the unity value, allowing some extra margin for a differential signal regeneration process [26]. The NM parameter characterizes the ability of a circuit to form the correct output signal in the presence of noise. The correctness of operation in these conditions is guaranteed by an appropriate voltage gain, which allows the extra margin between the generated input and output voltage swings. The current steering capability of the circuit is defined by the VSR parameter, which is the ratio between the current in the driving branch, I_{ON}, compared to the total current generated by the current source.

According to [20], three different power-delay regions can be identified with voltage gain A_{v} and voltage swing ΔV parameters assigned. These are 1) low-power high-delay region, 2) power-efficient delay-efficient region and, 3) high-power low-delay operating region. The operation in these regions is expressed by the power-delay trade-off equation (13) in [20] and represents a combination of different design parameters chosen during the optimization process. In this work we attempted to work in power-efficient region, while keeping logic swing and voltage gain as low as possible within the range allowed by the noise margin.

To satisfy the main objectives for power-efficient operation, we used a modest non-minimal transistor size values of pull-down network which, not only provided the desired voltage gain and a reduced impact of process variations, but also allowed the regeneration capability of the circuit. The length L_{Ni} of all the transistors (except the current source transistor) was kept identical and close to minimum value. The width of nMOS transistors W_{Ni} influenced the load capacitance produced at output, with the corresponding increase in the propagation delay of an input signal. Therefore, this parameter is chosen carefully to satisfy both timing objectives and the voltage gain A_{v} variation goals. The level of supply voltage V_{dd} defines the power characteristic of the gate. The product $P = V_{\mathrm{dd}} \times I_{\mathrm{ss}}$ shows the dominance of static power group in MCML gate's power profile. Based on this formula, lower V_{dd} values account for better power profile. On the other hand, supply voltage V_{dd} variation has a direct relationship to the system's output impedance and voltage gain, and, therefore, should be chosen appropriately to account these as well.

MCML optimization and circuit design

MCML gate optimization process requires simultaneous alteration in design variables represented in Table I to satisfy the performance and power objectives formalized in design parameters of previous section. HSPICE environment allows a well-structured way of MCML circuit netlist representation; the internal topology is built upon the realistic n/p-type MOSFET transistor definitions (can be found with a standard CMOS cell library models) interconnected with the other active/passive HSPICE components represented as the circuit abstraction models, also called SUBCIRCUITS. The optimization process starts with the definition of variable parameters and their limits, the optimization specifications (with the .PARAM parameter=OPTx(init, min, max) format) and the optimization goals (provided as .MEASURE statements with the keyword GOAL stating limits) to be achieved. We performed the transient type of analysis during the optimization process and stored the output values for the future best-case input/output parameter selection.

The automated optimization process consists of two global steps: 1) MCML cell library design and optimization and 2) the topological circuit design, which considers interconnection of basic cells into complex circuits, which meet the timing objectives.

MCML cell library optimization. The optimization process is further subdivided into two main steps. Firstly, the optimization procedure is run for a single-level logic gates (Figure 11). The identical cells are cascaded into a chain of elements and tested with various levels of fanout and capacitive load. The parameter's estimation procedure, performs the required measurements of the output signals/measures (like, propagation delays, output voltage swings, rise/fall times, voltage gains, noise margins, power consumption and others) simultaneously for every gate in a chain. Full specification of MCML circuit optimization problem is given in Table II, where Vth is the threshold voltage of nMOSFET transistors used in the design, Vdd_{CMOS} is the supply voltage of the standard CMOS process library used, and W_{min} and L_{min} are the minimal transistor dimensions specified in the corresponding transistor definitions. The optimization process is executed for the entire list of variables specified; the adoption of appropriate limits on these can reduce the overall optimization time. The outputs which achieve optimization criteria are stored in a separate file and are used in the second step of the tuning process. The search sequence in which every parameter is tuned (Table II), is adapted from [21], [15].

Based on the recorded values obtained in the first step of parameters search, multi-level logic gates are optimized. Due to the fact that every gate of the MCML cell library should provide the necessary signal regeneration and stability characteristics, the multi-level gates are optimized with the identical supply voltage levels and input/output voltage swings to the ones of single-level designs recorded previously. The other search parameters specified in Table II are kept identical to

TABLE III: MCML cell library design values optimized for minimum power dissipation

45nm process	$fmax$ $[GHz]$	Vdd $[V]$	ΔV $[V]$	W_{N} $[\mu m]$	L_{N} $[\mu m]$	W_{P} $[\mu m]$	L_{P} $[\mu m]$	W_{NS} $[\mu m]$	L_{NS} $[\mu m]$	I_{ss} $[\mu A]$	$Power$ $[\mu W]$
INV/BUF	20	0.9	0.35	0.2	0.05	0.3	0.1	1.0	0.1	31.77	28.6
XOR/XNOR	20	0.9	0.35	0.4	0.05	0.4	0.15	1.2	0.15	40.1	69.0
AND/NAND	20	0.9	0.35	0.4	0.05	0.4	0.15	1.2	0.15	39.1	67.3
D-FF	20	0.9	0.35	0.4	0.05	0.4	0.15	1.2	0.15	40.8	167.7
SEL21	20	0.9	0.35	0.4	0.05	0.4	0.15	1.2	0.15	41.3	74.4

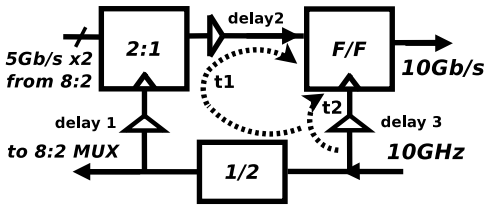


Fig. 13: Critical timing analysis performed in 2:1 MCML multiplexer

the single-level gate optimization process. Like the single-level circuit optimization, the search sequence was adopted from [21], [15]. In addition to MCML cells optimization, CMOS-to-MCML and MCML-to-CMOS level converters were created and tuned, providing the interface between the fully-custom MCML and semi-automated CMOS circuit implementations.

Once the MCML library is designed, the recorded values are further used in the biasing circuits configuration process. An example of MCML cell biasing circuit is presented in Figure 11. Biasing circuit allows a certain degree of $V_{\mathrm{Rfp}}/V_{\mathrm{Rfn}}$ voltage variations and allows fast tuning of the circuit in case of any fluctuations. The design is based on a simple two-stage push-pull operational amplifier described in [27]. The HSPICE-enabled OPAMP optimization procedure includes the proper transistor sizing with the corresponding currents/voltages generation at each stage of the design. The resulting implementation of the operational amplifier is further tested as a part of MCML gate biasing and CDR circuits.

MCML circuit timing closure. Once the MCML cell library is optimized and tested, the topological MCML circuit design is performed. The novelty of automated MCML cell library design procedure does not allow performing timing analysis in toolkit-assisted way as it is usually done in CMOS case. An example of fully-custom critical timing analysis for a 2:1 MCML-based multiplexer is shown in Figure 13.

A critical path t_1 is formed through the frequency divider and selector circuits, with the additional impact introduced by a delay line marked as 1. Since the input signal sampling points for the flip-flop and selector circuits are not identical (related to the direction of clocking signal propagation), the setup time of the D-type flip-flop input is severely reduced. To compensate this instability, clocking signal propagation times are adjusted by variation in the corresponding delay line size or characteristics (marked as 1,2 and 3). In particular, the delay value of path t_2 is enlarged to compensate the extra shift introduced by path t_1. A similar procedure for timing closure is used at other stages of the binary-tree-like structure of MCML-based multiplexer as well as demultiplexer designs (Figure 3-b), and, in the latter case, the direction of clock and data signal propagation is identical.

To conclude, high-speed SERDES designs were implemented to meet timing constraints for 20 Gb/s operation using the 45 nm CMOS process, allowing the possibility of reusing the MCML cell libraries for a variety of applications. Table III lists the main design variables used to satisfy these performance objectives.

REFERENCES

[1] R. S. Tucker, "Green optical communicationspart ii: Energy limitations in networks," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 17, no. 2, pp. 261–274, 2011.

[2] D. A. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proceedings of the IEEE*, vol. 88, no. 6, pp. 728–749, 2000.

[3] D. Huang, T. Sze, A. Landin, R. Lytel, and H. L. Davidson, "Optical interconnects: out of the box forever?" *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 9, no. 2, pp. 614–623, 2003.

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

[5] I. O'Connor, "Optical solutions for system-level interconnect," in *Proceedings of the 2004 international workshop on System level interconnect prediction*. ACM, 2004, pp. 79–88.

[6] R. S. Tucker, "How to build a petabit-per-second optical router," in *Lasers and Electro-Optics Society, 2006. LEOS 2006. 19th Annual Meeting of the IEEE*. IEEE, 2006, pp. 486–487.

[7] R. Sohan et al., "Characterizing 10 Gbps network interface energy consumption." in *IEEE LCN*, 2010, pp. 268–271.

[8] 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, pp. 13–18.

[9] H. Esmaeilzadeh et al., "Dark Silicon and the End of Multicore Scaling," *Micro, IEEE*, vol. 32, no. 3, pp. 122–134, 2012.

[10] 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.

[11] 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, pp. 1–3.

[12] (2013, Jan.) CONTEST - CONfigurable Transceiver Energy uSage Toolkit. [Online]. Available: http://www.cl.cam.ac.uk/research/srg/netos/greenict/projects/contest/

[13] A. X. Widmer and P. A. Franaszek, "A DC-balanced, partitioned-block, 8B/10B transmission code," *IBM Journal of Research and Development*, vol. 27, pp. 440–451, 1983.

[14] IEEE Standard, "IEEE 802.3ae – 10 Gb/s Ethernet," 2002.

[15] H. Hassan, M. Anis, and M. Elmasry, "MOS current mode circuits: Analysis, design, and variability," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 8, pp. 885–898, 2005.

[16] S. Palermo, A. Emami-Neyestanak, and M. Horowitz, "A 90 nm CMOS 16 Gb/s Transceiver for Optical Interconnects," *Solid-State Circuits, IEEE Journal of*, vol. 43, no. 5, pp. 1235–1246, 2008.

[17] J. Lee and M. Liu, "A 20Gb/s Burst-Mode CDR Circuit Using Injection-Locking Technique," in *Proc. Digest of Technical Papers. IEEE Int. Solid-State Circuits Conf. ISSCC 2007*, 2007.

[18] P. Sakian, M. Saffari, M. Atarodi, and A. Tajalli, "Low-power analogue phase interpolator based clock and data recovery with high-frequency tolerance," *IET Circuits, Devices and Systems*, vol. 2, no. 5, pp. 409–421, 2008.

[19] J. Poulton, R. Palmer, A. M. Fuller, T. Greer, J. Eyles, W. J. Dally, and M. Horowitz, "A 14 mW 6.25-Gb/s transceiver in 90 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 12, pp. 2745–2757, 2007.

[20] M. Alioto and G. Palumbo, "Power-aware design techniques for nanometer mos current-mode logic gates: a design framework," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 4, pp. 42–61, 2006.

[21] J. Musicer and J. Rabaey, "MOS current mode logic for low power, low noise CORDIC computation in mixed-signal environments," in *in Proc. ISLPED*, 2000, pp. 102–107.
[22] M. Allam and M. Elmasry, "Dynamic current mode logic (DyCML): A new low-power high-performance logic style," *IEEE J. Solid-State Circuits*, vol. 36, no. 3, pp. 550–558, 2001.
[23] M. Yamashina and H. Yamada, "An MOS current mode logic (MCML) circuit for low-power sub-GHz processors," *IEICE Transactions on Electronics*, vol. 75, no. 10, pp. 1181–1187, 1992.
[24] E. Sackinger, *Broadband Circuits for Optical Fiber Communication*, 1st ed. Wiley, 2005.
[25] G. Palumbo and M. Alioto, *Model and Design of Bipolar and MOS Current-Mode Logic*, 1st ed. Springer, 2005.
[26] S. Badel et al., "A generic standard cell design methodology for differential circuit styles," in *in Proc. of the Design, Automation and Test in Europe Conference and Exhibition*, 2008, pp. 843–848.
[27] P.R. Gray et al., *Analysis and design of analog integrated circuits*, 4th ed. Wiley, 2001.

Yury Audzevich is a Research Associate in the Systems Research Group, University of Cambridge Computer Laboratory. He received his B.Sc and M.Sc degrees in Radio Physics and Electronics from the Belarusian State University, Minsk, Belarus in 2003 and 2005 respectively. He obtained his Ph.D. degree in Information and Telecommunication technologies in 2009 from University of Trento, Italy. His scientific interests include IC design, open-source software and hardware platforms, reconfigurable systems and energy-efficiency aspects of network equipment.

Philip M. Watts (M04) obtained his B.Sc. in applied physics from the University of Nottingham, U.K. in 1991. He received his M.Sc. and Ph.D. degrees in 2003 and 2008 respectively, both from University College London (UCL), U.K. From 1991 to 2000, he worked at BAE Systems Advanced Technology Centre, and from 2000 to 2010, he was senior optical hardware engineer with Nortel Networks, a researcher at Intel Research and a consultant to Azea Networks and Huawei Technologies. From 2008 to 2010, he was a Research Fellow at the Computer Laboratory, University of Cambridge. He is currently an EPSRC Research Fellow and Lecturer in the Department of Electronic and Electrical Engineering at UCL where his research interests cover optical interconnects and electronic signal processing, control and coding circuits for optical communications.

Andrew West received the B.A. degree in Computer Science from the University of Cambridge, UK in 2003. After graduating, he worked as a researcher at the University's Computer Laboratory and as an intern at Sun Microsystems. In recent years, he has been employed in deep-submicron physical implementation and digital design roles at different companies in the UK. He currently works at an IC design consultancy in Cambridge.

Alan Mujumdar is a PhD candidate at the Faculty of Computer Science and Technology, University of Cambridge, Cambridge, UK. He received his BEng in Communication Systems Engineering from the University of Portsmouth, Portsmouth, UK, in 2010, and his MPhil in Advanced Computer Science from the University of Cambridge, Cambridge, UK, in 2011. His research interests include parallel computer architectures, distributed systems and interprocessor communication networks.

Simon W. Moore is Reader (Associate Professor) in Computer Architecture at the University of Cambridge Computer Laboratory in England, where he undertakes research and teaching in the general area of computer architecture and VLSI design. He is a Senior Member of the IEEE and a Fellow of the IET and BCS.

Andrew W. Moore is a Senior Lecturer at the University of Cambridge Computer Laboratory in England, where he is part of the Systems Research Group working on issues of network and computer architecture. His research interests include enabling open-network research and education using the NetFPGA platform, other research pursuits include low-power energy-aware networking, and novel network and systems data-center architectures.