A 6.4/3.2/1.6 Gb/s Low Power Interface with All Digital Clock Multiplier for On-the-Fly Rate Switching

Masum Hossain, Kambiz Kaviani, Barry Daly, Makarand Shirasgaonkar, Wayne Dettloff, Teva Stone, Kashinath Prabhu, Brian Tsang, John Eble, Jared Zerbe

Rambus Inc., 1050 Enterprise Way, Suite 700, Sunnyvale, CA, U.S.A

***Abstract* – A dynamic rate adjustable interface is designed a 40-nm LP CMOS process. On-the-fly dynamic rate change is enabled by an all-digital frequency multiplier that detects a reference frequency change, and accordingly provides 4x multiplied clock without any idle time. The clock multiplier, along with matched source synchronous clocking and clock equalization, allows blind reference clock shifting to scale the data rate from 1.6 to 6.4 Gb/s within 6.125ns without idle time or bit errors during transitions. The interface efficiency is 2.6 mW/Gb/s @6.4 Gb/s & 3.4 mW/Gb/s @3.2 Gb/s when using reduced clock swing and external transmitter swing at the reduced data rates.**

I. INTRODUCTION

In high performance mobile systems throughput and memory bandwidth demand varies rapidly in response to different application usage[2]. Simultaneously, mobile systems should run at the lowest frequency possible while delivering adequate performance in order to achieve the lowest possible power consumption. Such rapid processor workload variation translates to a sudden change in reference clocks to the controller which requires controller-memory bandwidth to scale according to demand. It is therefore desirable to allow the memory interface data rate to scale with the processor workload by being slaved to a processor clock which serves as a reference clock for the interface. In a mostly CMOS digital implementation, interface power scales linearly with frequency, so long as there is no significant overhead in the data path to accommodate rate changing. Ultimately, it is desirable to allow the reference clock to shift on-the-fly without interrupting traffic.

Although this is an attractive approach for power efficient link design, several practical challenges need to be addressed. First, high-speed interfaces such as the matched source synchronous clocking (MSSC) interface [1] shown in Fig. 1 typically contain clock multiplier units (CMUs), which must be made frequency agile with short transition times to support such low overhead on-the-fly rate shifting. In addition, in order to maintain high efficiency at low data rates, CMU power consumption should be as low as possible. Finally, often performance sensitive circuits in transceivers are non-CMOS, hence their power consumption does not scale with

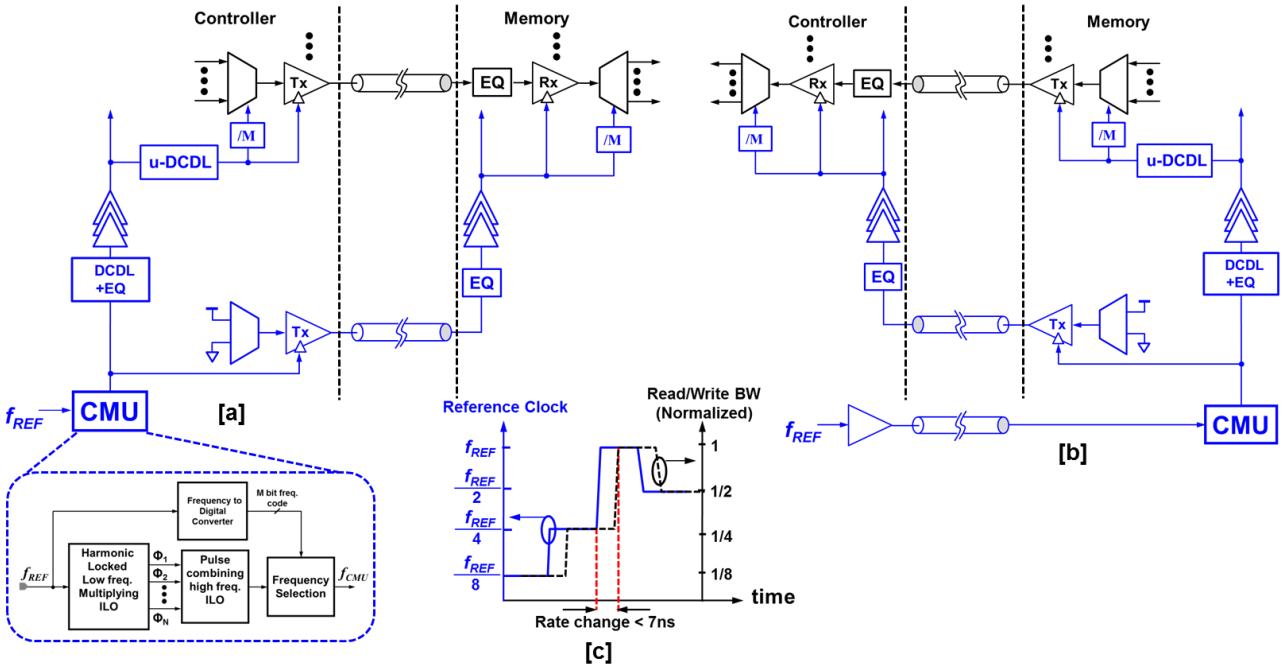

Fig.1 Reference clock based scalable data rate interface [a] Write path with matched source synchronous clocking [b] Read path with matched source synchronous clocking [c] Data rate scaling with reference clock

978-1-4673-1556-2/12/$31.00@2012 IEEE

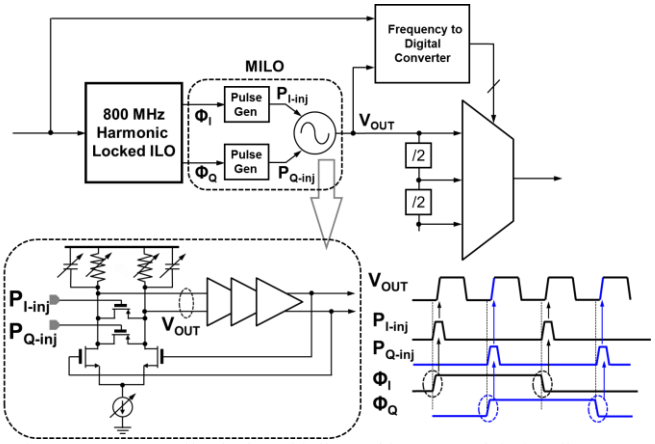

Fig. 2 Frequency agile clock multiplier architecture and timing diagram for pulse combining 4x Multiplying ILO.

frequency. Therefore, additional power reduction techniques are required to improve power efficiency at low data rates. Detailed design techniques for the proposed frequency agile CMU are explained in section II. Techniques to scale static power consumption in the system are discussed in section III. Measurement results from a 40-nm LP CMOS prototype are provided in section IV and the main contributions of the paper are summarized in section IV.

II. Frequency Agile Clock Multiplier Unit (CMU)

Conventional clock multipliers employ a voltage controlled oscillator (VCO) in a phase-locked loop (PLL) with a divider in the feedback path. If the division ratio is kept fixed, changing the reference clock frequency will result into a new VCO output frequency. However, convergence to a new output frequency takes significant time and usually set by the loop bandwidth of the PLL. Since PLLs are second order systems, frequency relock times are usually in the order of micro-seconds. Such latency is not acceptable in a processor/memory interface and thus a more agile low-latency frequency multiplier is needed.

In our proposed architecture, frequency multiplication functionality is provided with a harmonically locked injection locked oscillator (ILO), a frequency to digital converter, and a frequency selector (Fig. 2). Rather than changing the VCO frequency to track the reference clock rate change, when the reference clock is switched to a sub-harmonic frequency, the ILO remains locked at the highest frequency. Shifts in the multiple between the reference frequency and the ILO output are detected in the frequency to digital converter (FDC). In the FDC, the clock from the ILO output samples the reference clock, and based on consecutive samples determines the current reference clock: ILO multiple. This decision code is then used to select between the ILO clock or one of its divided versions such that the multiplication ratio between reference clock and output clock remains unchanged.

Advantages of the proposed frequency multiplier architecture can be appreciated by assessing the short lock time. Since the ILO does not need to change its operating frequency, there is no latency for frequency re-locking. The FDC takes approximately 6 ILO clock cycles to confirm a

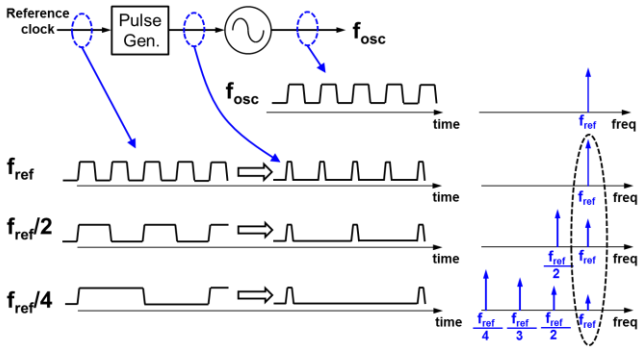

Fig.3 Harmonically locked 800 MHz ILO with 800/400/200 MHz ref. clock.

reference clock rate change, and the CMU switches to a new output within 8 bit-rate clock cycles. As a result, in a 6.4 Gb/s system, CMU can change its output frequency from 3.2 → 1.6 GHz in response to 800→ 400 MHz reference change within 6.125ns. Furthermore, an all-digital implementation of frequency detection and selection is used for easier portability and programmability.

A. Harmonic Locked ILO/MILO

A multiplying ILO (MILO) has been demonstrated to provide rapid clock multiplication in fast turn-on transceivers [1]. The concept can be further extended for sub-rate injection as shown in Fig. 3 for fast reference switching. If the reference clock rate is scaled from the nominal rate, f_{REF}, by an integer factor of /N, main frequency component of the pulse train also shifts to f_{ref}/N. However, there is still a harmonic component present at f_{ref}, which can be used to lock an ILO. As a result, it is possible to keep the ILO always locked at f_{REF} during the binary scaling of the reference clock, despite a change in the multiplication ratio. Since the ILO does not need to re-lock to a new frequency, this technique is significantly faster than using a PLL which requires relocking. There are several challenges in the implementation: first, the injection energy of the *N*-th injection harmonic scales down linearly resulting in weaker injection strength. Second, any time interval error in the ILO clock accumulates for N consecutive ILO clock cycles without correction. These effects combine to cause the VCO lock range to drop significantly as f_{ref} is lowered.

In this work several techniques are introduced to achieve sufficient lock range even when the multiplication factor is large, $N = 16$ or 32. To reduce the effect of jitter accumulation in the 3.2 GHz ring ILO, clock multiplication is provided in two stages. The reference clock is first multiplied to generate a 800 MHz secondary reference clock using a ring-VCO ILO/MILO. The 800 MHz clock is then used to injection-lock a 3.2 GHz MILO. Jitter accumulation of the 3.2 GHz MILO is further reduced by a pulse combining technique. Taking advantage of the multiple phases available in the 4 stage 800 MHz ring VCO, pulses are generated from both in-phase and quadrature paths to inject the 3.2 GHZ VCO on every cycle. Hence, timing error accumulation is limited to one MILO clock cycle.

Existing injection techniques [1] consume dc power, potentially wasting energy between injection pulses. The

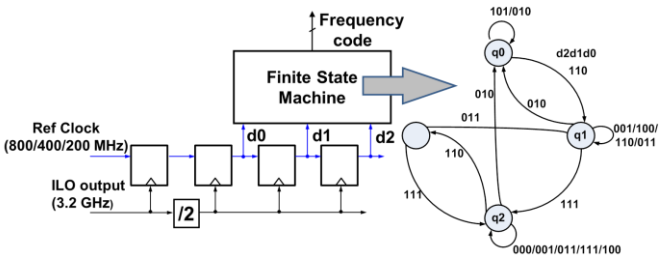

Fig. 4 Frequency to digital converter implementation.

proposed injection method shown in Fig. 2 only consumes power during injection by utilizing a zero crossing injection, significantly improving power efficiency while simultaneously allowing direct combination in-phase and quadrature injection paths. All circuits except the 3.2 GHz VCO are pseudo-differential CMOS to minimize power. Since PMOS switches are driven with rail to rail CMOS signal levels, no additional CMOS to CML converter is required for injection.

B. Frequency to Digital Converter (FDC)

The principal of frequency to digital conversion is to sample the unknown frequency signal with a known frequency signal [3]. In this case, the MILO output is "fixed" since it is always locked at the same frequency while the reference signal frequency can change. The frequency to digital converter oversamples the reference clock with the high frequency MILO clock output as shown in Fig. 4. A finite state machine determines the sub-harmonic reference frequency based on the current FSM state and the sampled pattern. In general, if M is the number of frequency doublings in the system, it requires M+1 fast clock cycles with an oversampling ratio of 2 to resolve the detection. For example, in this case, with M=2, a "101" or "010" pattern indicates a 800MHz reference, and a "111" or "000" pattern indicates a 200MHz reference. The fast frequency-to-digital converter driving a frequency selector after the fixed high-frequency MILO allows fast and efficient rate switching.

III. Improving Power Efficient Data Rate Scaling

Proportional scaling of power consumption with data rate is essential to efficiently utilizing frequency scaling, which requires careful consideration of non-CMOS circuit elements and architectural or data-path overhead at low speeds.

A. Swing Scaling in Clock Buffer& Transmitter

Conventional CMOS clock buffer power consumption scales linear with frequency but suffers from power supply induced jitter (PSIJ). CML clock buffers have excellent supply rejection properties but constant current, resulting in poor power efficiency when operating at lower frequencies. This can be mitigated by reducing CML clock bias currents and swings at lower data rates [5]. One possible side effect of lower swing is an increase in duty cycle error due to manufacturing device-mismatch within differential pairs. In order to compensate for this effect, a distributed duty-cycle control was used throughout the clock and DCDL path as shown in Figure 5.

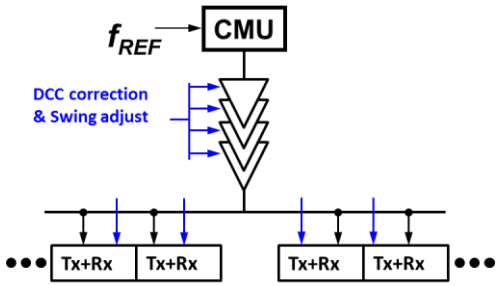

Fig. 5 Clock distribution with swing adjustable clock buffer and duty cycle correction.

Similarly, transmit swings can be reduced at lower frequencies to reduce signaling output power. Since high performance links have significantly higher voltage and timing margins when operating at reduced rates due to both larger bit periods and reduced channel loss, it is acceptable to reduce this excess margin for improved power efficiency while still achieving the same absolute system margins.

Clock and transmitter scaling techniques together can be used to reduce DC power consumption by 31% at 3.2 Gb/s. While we did not explore dynamic swing-scaling approaches in this system, the results with static settings demonstrate the potential power savings possible by such techniques and indicate it as a promising area for future research.

B. MSSC & Clock Equalization

Dynamic rate changes at an interface often will require additional synchronization and will result in increased latency and extra power consumption. In this work the need for synchronization at the receiver side is eliminated by use of matched source synchronous clocking (MSSC). Changing the data rate on the fly also results in ISI on the clock line at the rate change boundary. This can create a narrow first bit while switching from lower data rate to higher data rate. This problem can become even more challenging when passing through a band limited channel. In this work, as in [1] receive equalization is used in the form of a continuous time linear equalizer (CTLE) on both the data and the clock bit to avoid timing margin loss on the first bit after a rate transition.

IV. MEASUREMENT RESULTS

The prototype chip is implemented in a TSMC 40-nm LP CMOS process with conventional wire-bond packaging. The die micrograph in Fig.6 includes the frequency agile clock multiplying unit, two bit slices, and clock buffering and distribution sized to emulate an 8-bit interface. The CMU occupies 0.08 mm^2 and consumes 16 mW from 1.1V supply.

Measured dynamic rate scaling functionality is shown in Fig.6(a) where the reference clock is stair-stepped between 200MHz→800MHz and back to 200MHz with the corresponding bit-rate switching from 1.6→3.2→6.4Gbps with no interface idle time and no dropped bits between rate transitions. In this case MILO multiplication factor *N* is 16. The change in reference clock is reflected in a corresponding bit clock change with latency of 6.125 ns. The eye diagrams including 6.4Gbps/3.2Gbps and 3.2Gbps/1.6Gbps & rate

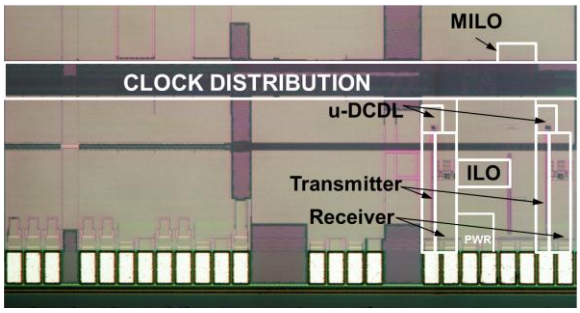

Fig.6. Die Photo of the implemented prototype in 40 nm LP

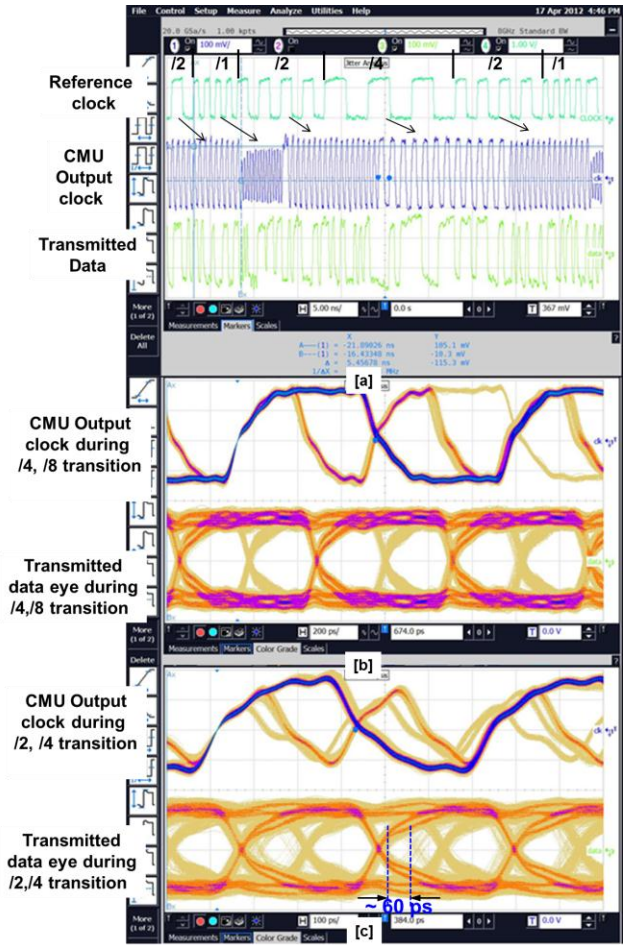

Fig.7 Dynamic data rate changing with reference clock.[a] Transient on-the-fly rate switching waveforms.[b] Overlaid eye diagram during 3.2Gb/s to 1.6 Gb/s transition [c] Overlaid eye diagram during 6.4Gb/s to 3.2 Gb/s transition

transitions are also shown in Fig. 7[b] and [c]. It is interesting to note that the first transition from 3.2→6.4Gbps results in ISI on the clock and a timing edge shift which encroaches on the eye as observed by the scope; this edge placement movement is essentially eliminated at the receive device by the use of the receive clock CTLE.

Fig. 8[a] shows a BER bathtub for a data link running 2^{15}-1 PRBS with continuous 6.4 Gb/s operation in comparison to 6.4/3.2/1.6 Gb/s gap-less rate-switching operation and shows negligible timing loss from the addition of rate switching.

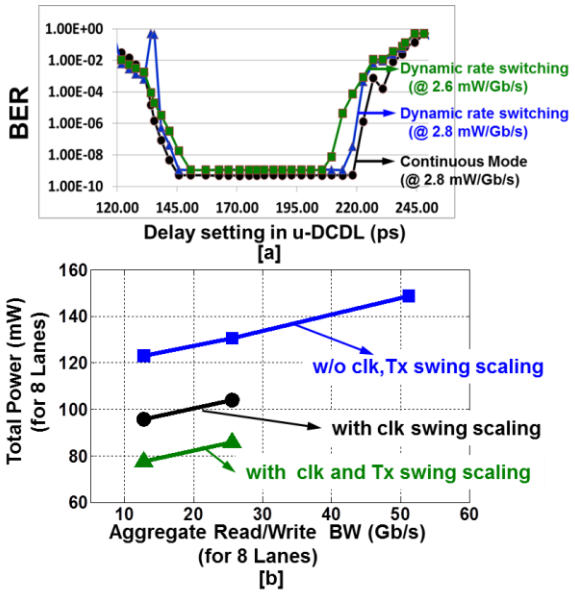

Fig. 8. [a] BER bathtub plots for continuous and dynamic rate switching mode.[b] Aggregate bandwidth and power consumption (for 8 lane) based on per lane measured power consumption for 6.4 Gb/s, 3.2 Gb/s and 1.6 Gb/s.

Sufficient timing margin is available for even more aggressive power reduction.

V. CONCLUSION

This work demonstrates architecture and circuit solutions required for fast on-the-fly dynamic bandwidth scaling capability without adding latency or protocol overhead. Such frequency agility facilitates linear power reduction in the digital CMOS circuits which are pervasive throughout a mobile system. Non-CMOS circuits which consume static power experience limited benefits. Therefore, to further extend power scaling, the development of dynamic clock and transmitter swing reduction is necessary. The potential benefit of such approaches is shown in Fig. 8[b]. As shown in Fig. 8[b], 31% power reduction is achievable without sacrificing error free data transmission (BER better than 10^{-9}) when both clocking, DCDL, and transmitter swings are scaled appropriately.

ACKNOWLEDGMENT

The authors are grateful to B. Leibowitz for technical discussions, T. Cados, G. Holst, T. Forkner for careful layout.

REFERENCES

[1] J. Zerbe et al., "A 5.6 Gb/s 2.4 mW/Gb/s Bidirectional Link With 8ns Power-On," Symp. on VLSI Circ., pp. 82-83, June 2011.

[2] B. Leibowitz et al, "A 4.3 GB/s Mobile Memory Interface With Power-Efficient Bandwidth Scaling," JSSC, vol. 45, no. 4, pp. 889-898, April 2010.

[3] V. Kratyuk et al "Frequency detector for fast frequency lock of digital PLLs," Electronic Letters, pp. 13-14,Jan, 2007.

[4] J. Poulton et al, "A 14-mW 6.25-Gb/s Transceiver in 90-nm CMOS," JSSC, vol. 42, no. 12, pp. 2745-2757, Dec. 2007.

[5] G. Balamurugan et al, "A Scalable 5–15 Gbps, 14–75 mW Low-Power I/O Transceiver in 65 nm CMOS," JSSC, vol. 43, no. 4, pp. 1010-1019, April 2008.