Future Microprocessor Interfaces: Analysis, Design and Optimization

Bryan Casper, Ganesh Balamurugan, James E. Jaussi, Joseph Kennedy, Mozhgan Mansuri, Frank O'Mahony, and Randy Mooney

Circuit Research Lab, Intel Corporation
2111 NE 25th Ave, M/S JF2-04
Hillsboro, OR 97124 USA

Abstract- **High-aggregate bandwidth interfaces with minimized power, silicon area, cost and complexity will be essential to the viability of future microprocessor systems. Optimization of microprocessor interfaces at the system level is crucial to providing the most cost-effective and efficient solution. This paper details a comprehensive interconnect and system level analysis method that can be used to accurately evaluate platform-level tradeoffs and has been correlated to link measurements with 10% accuracy. System tradeoffs with respect to interconnect quality, equalization, modulation, clock architecture are shown. Interconnect and circuit density improvements are identified as a promising research direction to maximize the bandwidth and power efficiency of future microprocessor platforms.**

I. INTRODUCTION

In past decades, the microprocessor industry has realized dramatic architectural performance gains based partly on exploiting instruction-level parallelism (ILP) of the code base. However for many applications, the architecture overhead required to further exploit ILP reaches a point of diminishing returns. Recent attempts in exploiting thread-level parallelism (TLP) demonstrate the potential of reaping significant power efficiency and performance gains [1,2]. For TLP to reach its full potential, microprocessor interface aggregate bandwidths must scale significantly. System interfaces that require considerable bandwidth include off-chip fabrics connecting multiple microprocessors, central processing unit (CPU) to application-specific accelerators and CPU to memory interfaces. Common requirements of these interfaces include high aggregate bandwidth along with small latency, form factor, power and cost.

To appropriately scale aggregate bandwidths of next-generation CPU interfaces while keeping the aforementioned factors of latency, form factor, power and cost in check, it will be essential to optimize the I/O system at the platform level and not focus solely on piecemeal architecture, circuit or interconnect solutions. Optimizing the CPU interface in a holistic manner presents many daunting technological and business challenges, but the associated gains will be significant. The purpose of this paper is to demonstrate the viability and benefits of a system-level optimization approach for CPU interface designs. The following section establishes new theory for an accurate link analysis method essential to system-level optimization. Section III demonstrates examples of CPU interface tradeoffs and appropriate solutions for equalization, modulation and interconnects. Section IV outlines parallel link clock architecture tradeoffs followed by conclusions in Section V.

II. LINK ANALYSIS

Accurate link modeling and analysis that comprehends channel and circuit impairments are crucial for design and optimization of next generation microprocessor interfaces [3]. This enables reliable performance comparison of different channels (e.g. FR-4 vs. Rogers™) and signaling schemes (e.g. 2-PAM vs. 4-PAM, transmitter (TX) pre-emphasis vs. decision feedback equalization), to evaluate the performance-cost tradeoffs of various interface designs. Suboptimum designs or standard specifications often result when piecemeal and inconsequential metrics are used that relate weakly to overall link bit-error ratio (BER) or margins. The comprehensive link analysis method outlined in this section enables a more holistic optimization of system parameters and helps isolate the most critical design criteria. This approach leverages the methods originally established in [4] and augments these methods with a more accurate analysis of TX jitter.

A. Unified TX jitter and intersymbol interference statistical analysis method

As noted by Stojanović in [5], the original method in [4] assumed that TX jitter was predominantly at low frequencies and hence did not amplify intersymbol interference (ISI). High-frequency TX jitter must be properly considered and not accurately accounting for it leads to optimistic results. This issue was addressed in [5] by using a model to convert TX jitter into equivalent voltage noise to evaluate the impact of Gaussian jitter. Below, we present an exact, yet computationally efficient method to comprehend the interaction between ISI and TX jitter for *arbitrary* uncorrelated jitter distributions.

The most straightforward method of ISI distribution calculation is based on the recursive convolution of individual ISI probability distribution functions (PDF) derived from a sampled single-bit pulse response. Recursive convolution of the ISI PDFs is valid if the data is random and uncorrelated. Extending this method to account for TX jitter is non-trivial since adjacent jittered transmit pulse widths are highly correlated and not independent random variables. We address this effect by using transmit *segments* (as opposed to transmit *pulses*) as the basis for our analysis. Segments as defined in Fig. 1 are characterized by a jittery transition from the left-side half-UI (i.e. unit interval) level to the right-side half-UI level. An example of the four possible transition segments for binary NRZ signaling without pre-emphasis is shown in Fig. 2. To appropriately account for TX jitter, we select and combine adjacent segments to compute the composite response. This *aggregate* PDF includes the effects of both ISI and TX jitter.
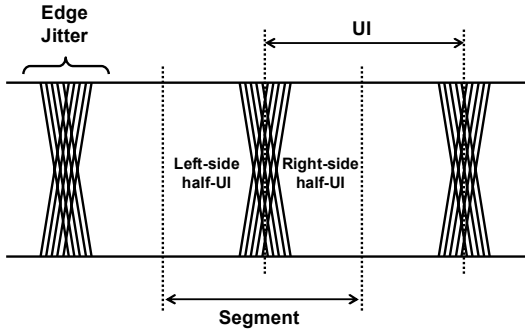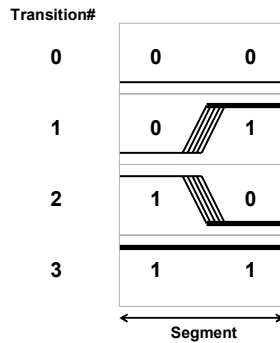
Fig. 1. Segment definition.

Fig. 2. Transition diagram for binary NRZ signaling.

The first step towards determining the aggregate ISI PDF is the computation of individual *transition* PDFs for all segments. Fig. 3 demonstrates an example computation of a transition PDF for postcursor segment #3, based on a transition from 0 to 1 ('01'). The receiver (RX) waveforms of transition '01', segment #3 have five possible shapes because the example edge jitter discrete PDF has five possible values. Sampling these transition responses at the cursor position produces the transition PDF after being appropriately scaled by the jitter PDF.
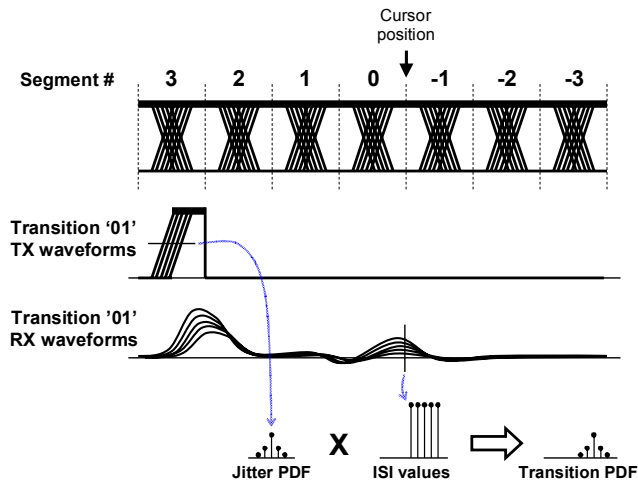

Fig. 3. Transition PDF calculation example.

Once all of the transition PDFs for each segment are calculated, these PDFs are selectively averaged and convolved to synthesize the aggregate ISI PDFs. Fig. 4 demonstrates this approach for binary NRZ signaling without equalization. The recursive process of selectively averaging and convolving the transition PDFs begins at a postcursor segment position that is longer than the span of postcursor channel ISI. In the example of Fig. 4, at the starting segment #2, the two possible transition PDFs that end with the same value are averaged and the result is denoted 'x0' or 'x1' in which the 'x' indicates a PDF average of two states. Each of these averaged PDFs is convolved with the next appropriate postcursor segment to produce an *accumulated* transition PDF such that the end state of the averaged PDF matches the initial state of the following transition PDF. This ensures that the segment does not unrealistically change state in the middle of the UI. This procedure is recursively performed on the resulting accumulated transition PDF until reaching postcursor segment #0. The preceding recursive operation is also performed for the precursor segments starting at a position to ensure that all precursor channel ISI is included in the analysis. Following the precursor and postcursor recursive operations, averaging the appropriate PDFs from segment #0 produce intermediate PDFs 'x0' and 'x1', while averaging from segment #-1 produces PDFs '0x' and '1x'. The applicable aggregate PDFs for levels '0' and '1' are merged through convolution. Statistical calculation of co-channel interference (CCI) such as crosstalk is performed using a similar aggregate ISI PDF calculation. These final aggregate ISI-jitter and CCI PDFs are then used to generate a BER eye similar to the method outlined in [4].
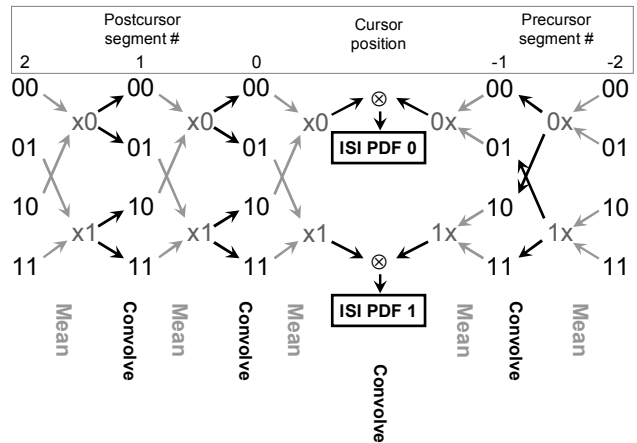

Fig. 4. Binary NRZ PDF calculation example.

The approach introduced in the preceding example can be extended to include the effect of transmit pre-emphasis, decision feedback equalization (DFE) and multi-level baseband modulation by appropriately augmenting the set of possible levels and transitions. The aggregate PDF calculation is preceded by enumerating all of the transition possibilities for which individual transition PDFs need to be derived. The valid transitions are a function of the pre-emphasis tap coefficient vector, $\mathbf{w}$ (of length n_e), and the number of signaling levels, N. Equation (1) defines a matrix $\mathbf{A}$ containing all possible ($n_e +1$)-long N-level data sequences. All the possible N^{n_e+1} valid level transitions can be derived by convolving $\mathbf{A}$ with the pre-emphasis vector $\mathbf{w}$ as shown in (2). The columns n_e and n_e+1

of the resulting matrix **B** indicate all possible level transition pairs.

$$A = \frac{1}{N} \begin{bmatrix} 0 \\ \vdots \\ N^{n_e+1} - 1 \end{bmatrix}_{Base-N} \quad (1)$$

$$B = A \otimes w \quad (2)$$

Mirroring the original binary NRZ example, the next step in the general solution is to extract all individual transition PDFs $f_k(i)$ for all segments in which k denotes the segment number and i refers to the transition number (which is the row index of **B** from which the transition is derived). Note that the index i assumes values ranging from 0 to $N^{n_e+1} - 1$. The accumulated transition PDFs $g_k(i)$ are produced by recursive averaging and convolution operations defined by (3)-(7). The postcursor (precursor) transition PDF selector function $s_{post}(i,j)$ ($s_{pre}(i,j)$) specifies the appropriate accumulated postcursor (precursor) transition PDFs to average before convolution with the subsequent segment. The effect of DFE (denoted by **c**) is comprehended by the selector function $s_{DFE}(i)$ that indicates the correct transmitted state as a function transition number.

$$g_k(i) = \frac{1}{N}\left[f_k(i) \otimes \sum_{j=0}^{N-1} g_{(k+1)}\left(s_{post}(i,j)\right) \right] - s_{DFE}(i)c_{k+1}; k \geq 0 \quad (3)$$

$$g_k(i) = \frac{1}{N} f_k(i) \otimes \sum_{j=0}^{N-1} g_{k-1}\left(s_{pre}(i,j)\right); k < 0 \quad (4)$$

$$s_{post}(i,j) = \left\lfloor \frac{i}{N} \right\rfloor + j \cdot N^{n_e} \quad (5)$$

$$s_{DFE}(i) = \frac{\left\lfloor \frac{i}{N} \right\rfloor \bmod N}{N-1} \quad (6)$$

$$s_{pre}(i,j) = (N \cdot i + j) \bmod N^{n_e+1} \quad (7)$$

The last step towards computing the aggregate level PDFs $f_L(i,j)$ for each of the $l=0...N-1$ levels at the cursor is to consolidate the final accumulated PDFs for both the precursor and postcursor. Equation (8) represents this concluding step with appropriate selector functions ((9)-(10)) to select the correct levels and preserve waveform continuity. A single time slice of the BER eye(s) is formed by the cumulative integration of the level PDFs $f_L(i,j)$ as specified in [4]. The full BER eye is extracted by repeated aggregate level PDF and BER slice calculation across the multiple cursor UI sample points.

$$f_L(l) = \frac{1}{N^{n_e+1}} \sum_{j=0}^{N^{n_e}-1} g_0\left(s_{postL}(l,j)\right) \otimes \sum_{i=0}^{N-1} g_{-1}\left(s_{preL}(l,i,j)\right) \quad (8)$$

$$s_{postL}(l,j) = Nj + l \quad (9)$$

$$s_{preL}(l,i,j) = \left(i + N(Nj + l)\right) \bmod N^{n_e+1} \quad (10)$$

B. System-level modeling

The primary simplifying assumption of the above analysis method is that TX jitter is uncorrelated from edge to edge, implying the jitter frequency content is not directly modeled. The primary intent of modeling TX jitter is to accurately ac-

count for *high-frequency* jitter that causes TX pulse width distortion to amplify the ISI due to a lossy channel. Therefore, realistic correlated jitter sequences must be filtered to extract the high-frequency, uncorrelated TX jitter before it is used as an input to the signaling analysis method. Fig. 5 demonstrates the jitter interpretation method to derive both the TX uncorrelated jitter and the RX sampling jitter to be used for the signaling analysis. The TX raw jitter sequence is interpreted using a discrete-time highpass filter that extracts jitter frequencies which interact with the ISI. The highpass filter perfectly passes the UI jitter effect of duty cycle distortion and closely emulates the response of pulse width distortion for jitter frequencies close to the Nyquist rate. Receiver sampling jitter is a measure of untracked TX and RX jitter and is determined by applying a lowpass function to the TX jitter sequence, delaying by d UI to emulate the transport latency of the channel, then comparing to the recovered clock jitter. The lowpass filter (inverse of the highpass filter for the TX jitter) is used so as not to redundantly account for high-frequency jitter on both TX and RX sides of the link. The filters, delay element and difference operations shown in Fig. 5 are purely for simulation or measurement instrumentation purposes and do not imply link architecture implementation. The RX clock recovery (CR) method is generalized such that the behavioral model could be used to emulate a variety of clock topologies such as forwarded or embedded CR.
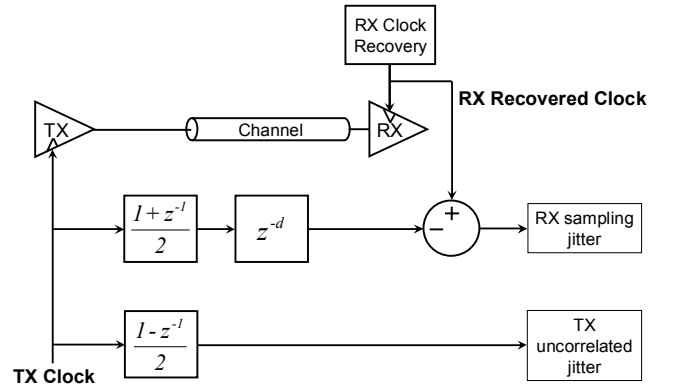

Fig. 5. Jitter interpretation method.

Our preferred method of system analysis for link architecture design and exploration is to use a comprehensive behavior model to emulate the transceivers, equalizers, embedded and/or forwarded clock channels, clock distribution, clock synthesis and CR. The behavioral model uses a combination of time and phase-step techniques that can generate jitter sequences in excess of several million UI. Some examples of the behavioral building blocks used are clock distribution buffers, phase-locked loops, delay-locked loops, phase interpolators and clock-data recovery circuits which are then used to emulate a variety of architectures and topologies. Combining the circuit and clock behavioral model with the preceding signaling analysis method (as shown in Fig. 6) offers the ability to accurately and efficiently extract link parameter sensitivities, and circuit architecture, equalization method and interconnect tradeoffs. The utility of this method is the capability of sweeping any model parameter in the system and being able to gauge its impact based on accurate and objective performance metrics such as BER or maximum achievable data rate (MADR) at a target BER. This type of holistic analysis ap-

proach is essential to properly architecting, designing and optimizing next-generation microprocessor interfaces and building an understanding of key link tradeoffs and sensitivities.
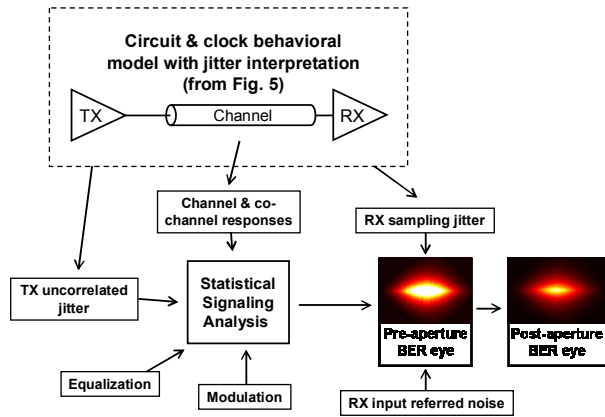

Fig. 6. Link analysis method.

III. INTERCONNECT, EQUALIZATION AND MODULATION OPTIMIZATION

Microprocessor I/O rates have scaled aggressively mostly due to steady bandwidth and power efficiency enhancements of CMOS process technology. As microprocessor systems have adopted I/O design improvements such as point-to-point topologies, equalization, enhanced CR, per-pin deskew and differential signaling, interface per-pair bandwidths have scaled to multi-gigabit per second rates. Even though process technology scaling enables increasing on-chip circuit bandwidths, the off-chip interconnect has become a key limiter to power efficient link operation. There is a significant amount of industry research aimed at overcoming interconnect bandwidth limitations using increasingly complex equalization [6] and modulation techniques [7]. These efforts are necessary to continue enhancing I/O rates for applications focused on backward compatibility to legacy interconnects, especially as it applies to widely adopted serial-based standards not optimized for aggregate bandwidth or power efficiency. However, even in cases when the off-chip interconnect is part of the system design space, microprocessor platform designers often have the tendency to overburden the transceiver and equalizer complexity for the sake of minimum interconnect cost. This constrained approach to system design forces the end-user cost to be inflated by requiring increased circuit power, enhanced power delivery, and improved thermal solutions while increasing time-to-market and overall product risk. A more balanced approach, in which improvements to all parts of the system are considered, is necessary to continue to aggressively scale aggregate bandwidths within a limited power envelope.

A. Co-design of Interconnect and Equalization

A good example of optimizing system cost is the design of a microprocessor backplane (BP) interface. Leveraging the signaling analysis concepts previously described, we compared the performance of two BP topologies as a function of equalization complexity. The two BPs are identical except that one of them has a 5mm BP via stub at each connector location while the other channel's BP connector vias are backdrilled

such that the stub length is limited to less than 1mm. Fig. 7 demonstrates the impact of via stubs and the resulting frequency domain notch. Even though the deep notch severely modifies the channel response, the degradation in channel capacity using Shannon's limit is less than 20% and still exceeds 100Gb/s. However, due to realistic considerations such as jitter, noise, limited TX signal swing and non-ideal equalization, the actual data rate degradation due to the via stubs is significant. To demonstrate the quantitative tradeoffs, the previously described signaling analysis was performed using the assumptions listed in Table 1. Note that the channel is routed in a "non-interleaved" manner such that inbound and outbound interfaces are isolated and near-end crosstalk (NEXT) does not exist. Non-interleaved routing methods are preferred in most cases since far-end crosstalk (FEXT) is much less destructive to signal integrity and much more amenable to simple equalization techniques such as linear equalization.
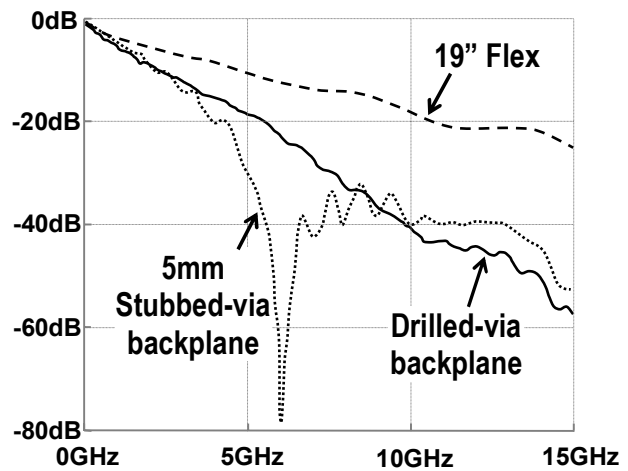

Fig. 7. Channel insertion loss.

TABLE 1
ANALYSIS ASSUMPTIONS

General Assumptions		Jitter and Noise Assumptions	
Modulation	Binary NRZ with data scrambling	TX duty-cycle error (modeled as uncorrelated bimodal jitter)	0.01UI p-p
BER target	10⁻¹²	TX uncorrelated gaussian jitter	0.01UI rms
TX swing	±500mV	RX duty-cycle error (modeled as bimodal jitter)	0.01UI p-p
TX FIR (Pre-emphasis) Assumptions		RX sampling uniform jitter	0.2UI p-p
Coefficient resolution	30mV	RX sampling gaussian jitter	0.01UI rms
2 tap	1 postcursor tap 0 precursor taps	RX slicer noise	1mV rms
3 tap	1 postcursor tap 1 precursor tap	**Channel assumptions**	
4 tap	2 postcursor taps 1 precursor tap	Pad Capacitance	RX=400fF, TX=200fF
5 tap	3 postcursor taps 1 precursor tap	FR4 total length	22 in.
6 tap	4 postcursor taps 1 precursor tap	Routing constrained to avoid near-end crosstalk. Far-end crosstalk assumed.	
DFE Assumptions		Channel includes socket-Ts, FCI Airmax™ connectors, 2 daughter cards, 1 BP and vias.	
Coefficient resolution	1mV		
No error propagation			

To demonstrate the accuracy of the signaling analysis method, Fig. 8 shows a measured data point based on a prototype signaling system [8] with channel and circuit characteristics similar to those listed in Table 1. The MADR of the simulated and measured system correlated to about 3%. On other

topologies not detailed here, simulation to measurement correlation was always within 10%.
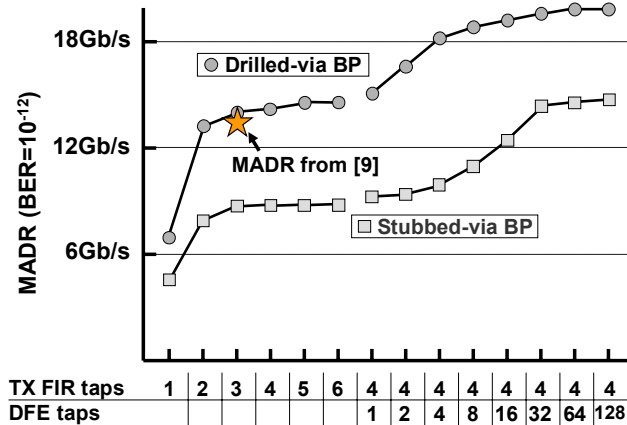

Fig. 8. Effect of BP via stub backdrilling (simulated).

Fig. 8 demonstrates that simple drilling of the BP vias significantly improves the MADR regardless of equalization topology or complexity. In fact, drilling the vias improves the data rate by close to a factor of two when 4 taps of DFE and 4 taps pre-emphasis are employed. An alterative perspective is to consider the requirements of an example BP-based microprocessor interface with a target data rate of 12Gb/s at a 10^{-12} BER. Drilling the BP via enables data rates exceeding 12Gb/s with very simple 2-tap pre-emphasis. However, the stubbed BP channel requires multiple taps of pre-emphasis and about 16 DFE taps to meet the target. The stubbed BP equalization solution leads to considerably higher power, area, complexity and risk.

While it is challenging to make a truly fair comparison based on the above metrics, it is possible to gain an intuitive perspective of the power difference between the two solutions. In the following example at data rates of 12Gb/s, the cost-per-tap assumption for a pre-emphasis and DFE implementation is ¼mW/Gb/s and ½mW/Gb/s, respectively. The assumption for baseline link power efficiency without equalization is 6mW/Gb/s. Given these assumptions, the drilled-via BP achieves efficiency of 7mW/Gb/s while the stubbed-via BP dissipates 16mW/Gb/s. Given the preceding tradeoffs, a comprehensive system cost analysis of the two scenarios is quite difficult partly due to the challenge of calculating incremental component cost (thermal and power delivery components) as a function of link power efficiency. However, evaluating end-user energy cost may be an intuitive albeit incomplete way to gain a perspective regarding total cost-of-ownership tradeoffs. Given assumptions such as an energy rate of $0.12/kWh, product lifetime of 5 years, power delivery efficiency of 50% and link activity factor of 50%, the end-user energy cost is $0.005/mW due to link power dissipation. Assuming a via drill cost of $0.05 per pair, the drilled and stubbed BP cost is given by (11) and (12), respectively.

$$cost_{drilled\ BP} = \$0.05 + \frac{7mW}{Gb/s} * 12Gb/s * \frac{\$0.005}{mW} = \$0.47 \quad (11)$$

$$cost_{stubbed\ via} = \frac{16mW}{Gb/s} * 12Gb/s * \frac{\$0.005}{mW} = \$0.96 \quad (12)$$

This example shows a 10x return on investment for end-user system cost by drilling the BP vias, illustrating the value of co-designing circuits and interconnects.

B. Baseband Modulation

In recent years there has been a strong trend toward applying multi-level modulation to bandwidth-limited chip-to-chip links. Theoretically, multi-level signaling or pulse amplitude modulation (PAM) has the potential to improve spectral efficiency within a bandwidth constrained channel. Practically, N-PAM has been successfully employed in past applications such as wireless, Ethernet and digital subscriber line (DSL). PAM has been successfully applied BP links in the past [9,10]. However, given the jitter and noise assumptions in Table 1, most microprocessor system topologies we evaluated showed little or no advantage to N-PAM signaling beyond what could be achieved with binary NRZ modulation.

A common claim is that 4-PAM is most compatible to via-stubbed BP channels since the frequency-domain notch severely limits the effective channel bandwidth [9]. Using the stubbed-via BP and Table 1 assumptions as an example, Fig. 9 demonstrates the performance comparison of 2-PAM versus 4-PAM signaling as a function of equalization complexity. Scaling the 4-PAM UI jitter to be the same fixed percentage of the symbol width as 2-PAM may be pessimistic. Hence, additional 4-PAM tradeoff data has been included to indicate the impact of scaling the jitter assumptions of Table 1 by ½x. Even though the example channel puts 4-PAM in the best possible light, it still does not show a compelling performance advantage over 2-PAM signaling. One of the most convincing reasons to consider multi-level signaling is not necessarily motivated by the channel but rather by the bandwidth limitations of the circuits. If the clocking and transceiver circuits are applied beyond the design bandwidth, the jitter could scale disproportionately with data rate. In the presence of a non-linear jitter-frequency tradeoff, 4-PAM signaling may be justifiable. However, multi-level signaling does not necessarily lead to easily achievable gains even when the channel bandwidth is severely limited.

Other forms of modulation that have also enjoyed recent popularity include simultaneous bidirectional signaling (SBD) [11,12] and baseband phase modulation [13]. Baseband phase modulation has been demonstrated but our analysis shows it has similar tradeoffs to PAM-based modulation and it has not been proven to provide an inherently advantageous signaling solution. SBD signaling has received much more attention partly because the tradeoffs differ significantly from other modulation approaches. Primary advantages to SBD are that the signal to ISI ratio scales proportionally with respect to binary NRZ. However, SBD incurs echoes due to return-loss effects and the implementation is much more sensitive to channel discontinuities than it is for uni-directional signaling [12]. Even so, systems incorporating refined interconnect such as cabled systems with few discontinuities coupled with adaptive echo cancellation techniques may benefit from SBD signaling. Duobinary and partial-response signaling are special cases of binary signaling and hold advantages with respect to error propagation and signal spectral shaping. However, the tradeoffs to these approaches are similar to a conventional finite-tap feed-forward filter and DFE system described in the BP examples.
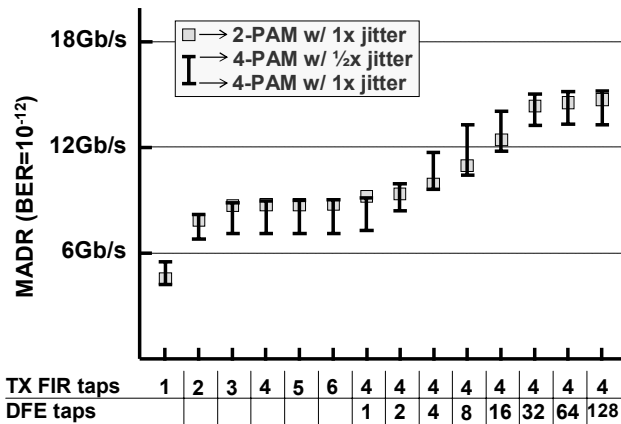
Fig. 9. 2-PAM vs. 4-PAM signaling with stubbed-via BP (simulated).

TX FIR taps	1	2	3	4	5	6	4	4	4	4	4	4	4	4
DFE taps							1	2	4	8	16	32	64	128

C. Optimizing Power through Alternative Interconnect

As demonstrated in the previous BP examples, channel discontinuities are highly disruptive to signal integrity and the associated MADR. Even for cases when stubs are minimized or eliminated, microprocessor channel discontinuities such as trace breakouts, sockets, connectors, packages, trace impedance discontinuities, circuit pad capacitance and through-hole vias can be more limiting to margins and data rate than the FR-4 transmission-line attenuation from dielectric and skin-effect losses. For this reason, there is motivation to explore alternatives to standard system topologies that target removal or mitigation of these discontinuities. Recent interconnect research into modular, flexible interconnect (Flex) has produced compelling results showing dramatically reduced channel discontinuities relative to conventional topologies.

A prototype that used Flex interconnect was recently demonstrated using a 90nm transceiver with a topology similar to that shown in Fig. 10 [14]. In this example, modularity of the link interface is provided by a proprietary top-side-package connector at both sides of the link. The connector was designed to minimize crosstalk and return loss, thus providing a highly effective launch for high-speed, differential signals. Launching signals from the top-side of the package improves signal integrity and density since the traces do not have to traverse through the package substrate core and socket. Additionally, the Flex cable dielectric loss was minimized by using a homogeneous polyimide dielectric. Fig. 7 details the insertion loss response of a 19 in Flex channel with connectors, packages and circuit parasitics.
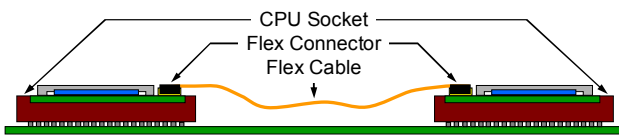

Fig. 10. Flex interconnect example.

Fig. 11 demonstrates platform measurements of this prototype flex topology as a function of trace length for a 90nm, 20Gb/s transceiver with 3-tap TX pre-emphasis and a receiver continuous-time linear equalizer [8]. Since the transceiver and clock circuits were bandwidth-limited to a 20Gb/s symbol rate, the MADR for Flex lengths of less than 30 inches was also limited to 20Gb/s even though the channel exhibited excess bandwidth as indicate by surplus eye margins. Comparing

the Flex rates to the FR-4 BP rates (based on Table 1 descriptions) demonstrates the potential of Flex topologies. Also consider that the measurement performed for the FR-4 BP was based on typical interconnect conditions such that the impedance discontinuities between the multiple components (connectors, cards, BP, packages, etc.) in the system do not represent the worst-case scenario. Given the need for acceptable yields in high-volume manufacturing, Flex provides even greater benefits beyond those implied in Fig. 11 because there are simply fewer transitions in which impedance discontinuities can occur.
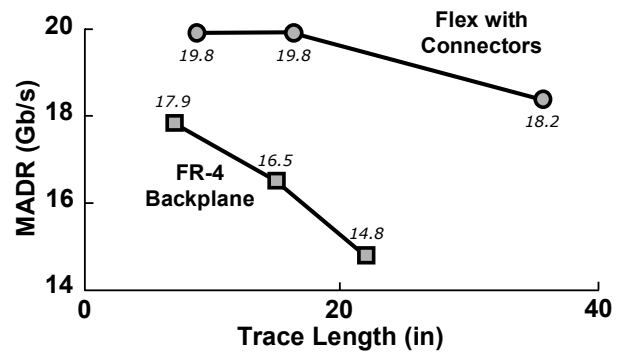

Fig. 11. MADR as a function of trace length (measured).

Were it not for the circuit bandwidth limitations of the previous example prototype, Flex interconnect would have shown a much greater benefit in terms of MADR. To remove circuit bandwidth restrictions, we performed simulation-based analysis of a 19 in Flex channel using circuit assumptions nearly identical to those in Table 1. Fig. 12 shows the potential bandwidth as a function of equalization complexity. While the tradeoff curve shape is similar to the Fig. 8 drilled-via BP example, the bandwidth achieved is about 3 times greater for approximately the same interconnect length. Alternative interconnects such as Flex are not only advantageous with respect to MADR, robustness and equalization complexity, they also enable significantly lower power by reducing the need for large TX swing values as shown in Fig. 12. Though the power savings enabled by small TX swings is dependent on sensitive receivers and comes at a slight cost in terms of MADR, the stringent power requirements of future microprocessor systems will likely justify these types of design shifts [15,16].

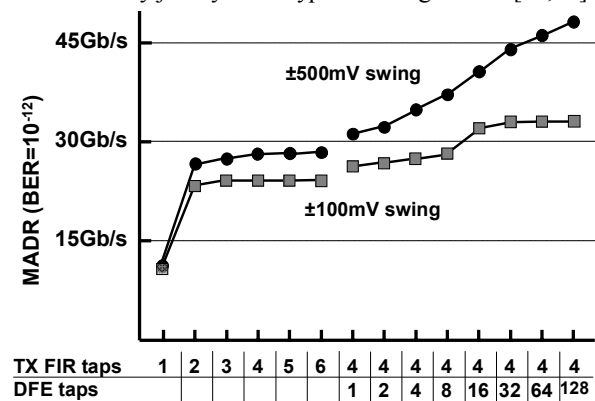| TX FIR taps	1	2	3	4	5	6	4	4	4	4	4	4	4	4
DFE taps							1	2	4	8	16	32	64	128

Fig. 12. 19 in Flex MADR as a function of equalization complexity.

In past generations of microprocessor systems, scaling the core clock frequency of CPUs was considered critical to maximizing performance. However, there is significant evidence that this trend has subsided to stay within a practical power envelope [17]. Parallelization of microprocessor cores and other functionality seems to be a likely path towards continued scaling of CPU performance. With regard to high-speed differential I/O, the industry is relying on off-chip data rates to follow an aggressive 1.25x per year scaling trend reminiscent of past CPU core clock frequency scaling [18]. As state-of-the-art clocking, transceiver and interconnect improvements are incorporated in future microprocessor interfaces, we also believe that steep data rate scaling will occur. However, we believe rapid acceleration of per-pair microprocessor interface rates will be relatively short-lived due to both off-chip interconnect and process technology constraints. Previously, we demonstrated that there is a non-linear performance-power tradeoff due to the diminishing returns on equalization power. Additionally, it has been demonstrated that even without channel bandwidth being the limiting factor, the performance-power tradeoffs are still non-linear due to circuit bandwidth and process technology considerations [16,19,20]. Fig. 13 shows an example from [16] indicating the likelihood of nonlinear power vs. performance of I/O circuits.
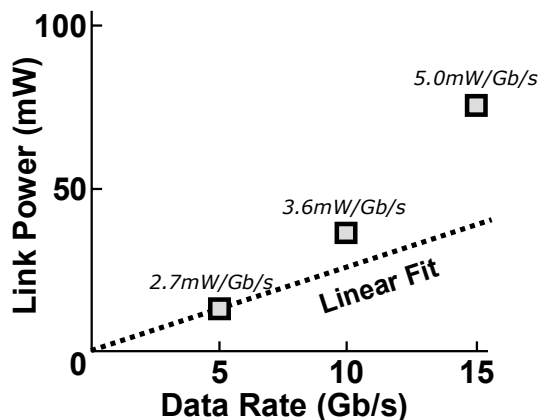

Fig. 13. Non-linear power-performance tradeoff (measured).

Given that optimizing system power will continue to be a paramount consideration in present and future CPU platforms [17], I/O will also need to operate close to its highest power efficiency region. This will result in a push to focus on link density optimization for high-aggregate bandwidth links rather than to just scale per-pair data rate. It is crucial that signaling systems research be targeted at accelerating industry forecasts for pin counts (and hence link density) well beyond the current prediction of 1.05x per year as implied by [18]. We have already shown Flex as an example of enabling a trend toward denser, more power efficient interfaces.

Another example of a dense interface technology that will impact microprocessor interfaces is the concept of system-in-package (SIP). SIP-based interfaces will lend themselves to extremely dense signaling topologies in a small form-factor, enabling order-of-magnitude increases in aggregate bandwidth at improved efficiency due to lower swing and simpler equalization requirements. Additionally, SIP-based I/O has the potential to greatly simplify CR methods, which are a significant portion of the overall I/O power budget. Density considera-

tions aside, SIP signaling implementations should rival the power efficiency enjoyed by long-distance repeater-based on-die signaling links. However, over-use of SIP technology has the associated drawbacks of amplifying power and thermal density concerns that are plaguing current-generation microprocessors. Additionally, lack of modularity within SIP systems limit configurability and scale of economy advantages of more conventional topologies based on industry-standard interfaces.

Interfaces that emulate the advantages of SIP technology in terms of power efficiency, density and form factor without the shortcomings of over-concentrating power and thermal density are essential to future platforms. Additionally, modularity will be the key to the widespread adoption of these next-generation dense interfaces.

IV. CLOCK ARCHITECTURE TRADEOFFS

Much of the signaling research and development over the last few decades has been focused on scaling inherently serial I/O to multi-gigabit per second rates. A majority of serial-based links include embedded CR schemes to limit the overhead of peripheral clock signals. Additionally, embedded CR links are highly modular which make its use popular in industry-wide link standard specifications. Embedded CR links have been shown to provide ample power efficiency and performance to meet the interface requirements for many modern microprocessor platform topologies. As aggregate bandwidth demand increases for future microprocessor interfaces, I/O designers will have to choose architectures optimized for parallel operation rather than select architectures intended for flexible, low pin count serial interfaces. A commonly held belief is that maximum performance is attained through the use of embedded CR techniques and that forwarded CR results in suboptimum performance. Embedded CR advocates claim that extracting the clock information directly from the data stream is much more effective than relying on a sampling clock derived from an alternative propagation path. However, the latest forwarded clock architectures preserve similar performance characteristics as embedded clock schemes such as the ability to have independent per-pin deskew and tracking CR. It has been shown that properly designed forwarded CR links have the potential of higher performance, lower power and smaller area than embedded CR topologies [8,22].

In addition to these benefits, forwarded CR links have multiple peripheral advantages not normally addressed in clock architecture comparisons. First of all, the power to recover the forwarded clock can be amortized across all receivers in the link. Also, forwarded CR parallel links do not necessarily require area-intensive, redundant and power-hungry dynamic recovery on a per bit basis in contrast to commonly implemented embedded CR-based parallel links. Additionally, forwarded clock schemes don't require clock-encoding of the data which can result in bandwidth, latency and power overhead. Furthermore, extracting the clock from a highly distorted signal often requires the coordination of an advanced equalizer such as a DFE if reasonable clock-data recovery bandwidth is to be maintained. Since a straightforward DFE implementation only equalizes for the middle of the eye, additional support hardware is required to extend the equalization to the edges for CR [21]. As equalizers become more complex, this additional overhead will degrade power efficiency and link robustness. Another advantage of forwarded clock schemes is that it is conducive to sophisticated test and validation methods [12]

which are vital to the efficacy of high volume microprocessor systems. Because the CR can be independent from the state of the RX data samples, the sample time and voltage may be swept outside the open eye without affecting recovered clock quality. In addition to these benefits, properly implemented forwarded clock links are simple, robust and highly tolerant to jitter from the transmitter [8].

While impressive advancements have been achieved in recent years with respect to I/O circuit performance and efficiency, there is still much more room for progress in years to come. However, these gains are not boundless and will be highly limited due to the effects of leakage, process reliability, shrinking supply voltages, process variation and technologies with limited support for specialized analog features such as inductors, capacitors, resistors and challenging small-signal transistor characteristics. An area of great concern relates to the random device variation due to device dimension uncertainty and random dopant fluctuations. Variation tolerant design techniques such as duty cycle correction circuitry and input offset cancellation of RXs have frequently been utilized in I/O systems. As mismatches become a more dominant design consideration, variation tolerant circuit architectures coupled with low-overhead calibration techniques will have to be employed at a much finer granularity and with wider dynamic range. Continued and focused research in this area will be of significant value to future high-bandwidth microprocessor interfaces.

V. CONCLUSION

Future microprocessor architectures and applications will stimulate a strong demand for high-aggregate bandwidth interfaces. To meet the stringent yield, power, cost and form-factor constraints of these future platforms, system-level optimization will be necessary. Precise and efficient system-level analysis methods coupled with accurate and comprehensive circuit and interconnect models will be essential components of any effort to coordinate design tradeoffs at the system level. While interconnect equalization is essential to overcome bandwidth limitations and scale aggregate link bandwidths, care must be taken to co-optimize the cost and performance of both the circuit and interconnect solution. We have shown that baseband modulation such as multi-level PAM does not necessarily yield optimum signaling performance. Alternative interconnects such as Flex and SIP will lead to extremely dense and power efficient architectures that will rival the performance and power efficiency of on-chip signaling links. Clock architectures should be designed to appropriately leverage the parallel nature of high-aggregate bandwidth CPU interfaces. Properly designed forwarded clock architectures will continue to enable high bandwidth microprocessor interfaces with minimum power, area and complexity.

REFERENCES

[1] D. Pham, et al., "The design and implementation of a first-generation Cell processor," *ISSCC 2005 Digest of Tech. Papers*, pp. 184-185, Feb. 2005.
[2] S. Vangal, et al., "An 80-Tile 1.28TFLOPS Network-on-Chip in 65nm CMOS," *ISSCC 2007 Digest of Tech. Papers*, pp. 98-99, Feb. 2007.
[3] H. Hatamkhani, F. Lambrecht, V. Stojanović, Chih-Kong Ken Yang, "Power-centric design of high-speed I/Os," Design Automation Conference, 2006, pp. 867-872, July 2006.
[4] B. Casper, M. Haycock, R. Mooney, "An Accurate and Efficient Analysis Method for Multi-Gb/s Chip-to-chip Signaling Schemes," *VLSI Circuit Symposium*, pp. 54-57, June 2002.
[5] V. Stojanović and M. Horowitz, "Modeling and analysis of high speed links," *CICC 2003*, pp. 589–594, 2003.
[6] B.S. Leibowitz, "A 7.5Gb/s 10-tap DFE receiver with first tap partial response, spectrally gated adaptation, and 2nd–order data-filtered CDR," *ISSCC 2007 Digest of Tech. Papers*, pp. 228-229, Feb. 2007.
[7] J.-H. Kim, et al., "A 4-Gb/s/pin low-power memory I/O interface using 4-level simultaneous bi-directional signaling," *IEEE Journal of Solid-State Circuits*, pp. 89-101, Jan. 2005.
[8] B. Casper, et al., "A 20Gb/s forwarded clock transceiver in 90nm CMOS," *ISSCC 2006 Digest of Tech. Papers*, pp. 90-91, Feb. 2006.
[9] J.L. Zerbe, et al., "Equalization and clock recovery for a 2.5-10-Gb/s 2-PAM/4-PAM backplane transceiver cell," *IEEE Journal of Solid-State Circuits*, pp. 2121-2130, Dec. 2003.
[10] J.T. Stonick, Gu-Yeon Wei, J.L. Sonntag, D.K. Weinlader, "An adaptive PAM-4 5-Gb/s backplane transceiver in 0.25-/spl mu/m CMOS," *IEEE Journal of Solid-State Circuits*, pp. 436-443, Mar. 2003.
[11] R. Mooney, et al., "A 900Mb/s Bidirectional Signaling Scheme," *IEEE Journal of Solid-State Circuits*, pp. 1538-1543, Dec. 1995.
[12] B. Casper et al., "8 Gb/s SBD link with on-die waveform capture," *IEEE J. Solid-State Circuits*, pp. 2111–2120, Dec. 2003.
[13] T. Simon et al., "A 1.6Gb/s/pair electromagnetically coupled multidrop bus using modulated signaling," *ISSCC 2003 Digest of Tech. Papers*, pp. 184-185, Feb. 2003.
[14] H. Braunisch et al., "Flex-Circuit Chip-to-Chip Interconnects," *Electronic Components and Technology Conference*, pp. 1853-1859, Jun. 2006.
[15] R. Palmer, et al., "A 14mW 6.25Gb/s Transceiver in 90nm CMOS for Serial Chip-to-Chip Communications," *ISSCC 2007 Digest of Tech. Papers*, pp. 440- 441, Feb. 2007
[16] G. Balamurugan, et al., "A Scalable 5-15Gbps, 14-75mW Low Power I/O Transceiver in 65nm CMOS," *VLSI Circuit Symposium*, June 2007, in press.
[17] P.P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities, and new frontiers," *ISSCC 2001 Digest of Tech. Papers*, pp. 22-25, Feb. 2001.
[18] International Technology Roadmap Service, 2006 update, assembly and packaging, http://public.itrs.net
[19] G. Wei, et al., "A variable-frequency parallel I/O interface with adaptive power-supply regulation," *IEEE Journal of Solid-State Circuits*, pp.1600-1610, Nov 2000
[20] J.-H. Kim and M.A. Horowitz, "Adaptive Supply Serial Links with Sub-1V Operation and Per-Pin Clock Recovery," *ISSCC 2002 Digest of Tech. Papers*, pp. 268-269, Feb. 2002.
[21] J. Wong and C.K.K. Yang, "A Serial-Link Transceiver with Transition Equalization," *ISSCC 2006 Digest of Tech. Papers*, pp. 82-83, Feb. 2006.
[22] J. Jaussi, et al., "A 20Gb/s Embedded Clock Transceiver in 90nm CMOS," *ISSCC 2006 Digest of Tech. Papers*, pp. 340-341, Feb. 2006.