Is 25 Gb/s On-Board Signaling Viable?

Dong G. Kam, *Member, IEEE*, Mark B. Ritter, Troy J. Beukema, John F. Bulzacchelli, *Member, IEEE*, Petar K. Pepeljugoski, *Senior Member, IEEE*, Young H. Kwark, Lei Shan, Xiaoxiong Gu, Christian W. Baks, Richard A. John, Gareth Hougham, Christian Schuster, *Senior Member, IEEE*, Renato Rimolo-Donadio, *Student Member, IEEE*, and Boping Wu, *Student Member, IEEE*

***Abstract*—What package improvements are required for dense, high-aggregate bandwidth buses running at data rates beyond 10 Gb/s per channel, and when might optical interconnects on the board be required? We present a study of distance and speed limits for electrical on-board module-to-module links with an eye to answering these questions. Hardware-validated models of advanced organic modules and printed circuit boards were used to explore these limits. Simulations of link performance performed with an internal link modeling tool allowed us to explore the effect of equalization and modulation formats at different data rates on link bit error rate and eye opening. Our link models have been validated with active, high-speed differential bus measurements utilizing a 16-channel link chip with programmable equalization and a per-channel data rate of up to 11 Gb/s. Electrical signaling limits were then determined by extrapolating these hardware-correlated models to higher speeds, and these limits were compared to the results of recent work on on-board optical interconnects.**

***Index Terms*—Channel equalization, electrical signaling limit, high-speed bus measurement, high-speed serial link, link modeling, multilevel signaling.**

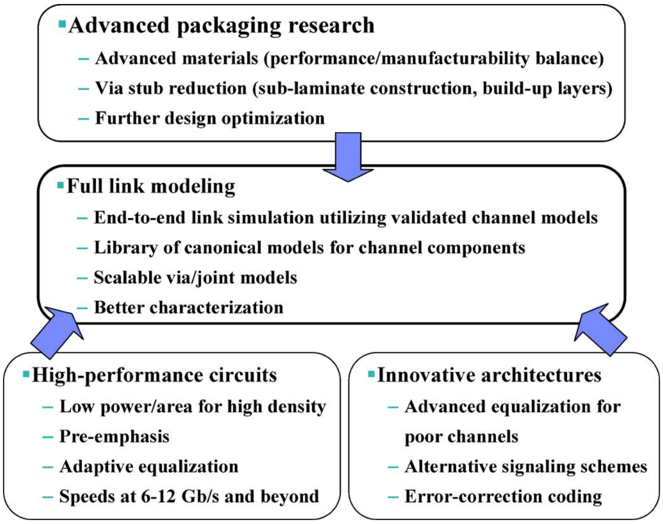

Advanced packaging research
– Advanced materials (performance/manufacturability balance)
– Via stub reduction (sub-laminate construction, build-up layers)
– Further design optimization
Full link modeling
– End-to-end link simulation utilizing validated channel models
– Library of canonical models for channel components
– Scalable via/joint models
– Better characterization
High-performance circuits
– Low power/area for high density
– Pre-emphasis
– Adaptive equalization
– Speeds at 6-12 Gb/s and beyond
Innovative architectures
– Advanced equalization for poor channels
– Alternative signaling schemes
– Error-correction coding

Fig. 1. Multitiered approach required to solve high-speed link challenges.

I. Introduction

OFF-CHIP bandwidth requirements continue to grow to meet the needs of server and storage consolidation, interprocessor communication, and multicore processor architectures [1]. Early work on the Optical Internetworking Forum's (OIF's) Common Electrical Interface (CEI-25) standard, aimed at specifying a parallel 20–25 Gb/s electrical interface for next generation 40 or 100 Gb/s optical modules, has shown that legacy channels are inadequate at speeds beyond ~17–20 Gb/s [2]. At the same time, future high-port-count switches and high-end servers will require hundreds to thousands of electrical links running at speeds of 10+ Gb/s to meet rising bandwidth demands.

For the last decade, electrical input/output (I/O) research has focused on improving transceiver circuits to sustain the growth of data rates while overcoming the limitations of the given integrated circuit (IC) technology [3]. As a result, deep submicron complementary metal–oxide–semiconductor (CMOS) I/O circuits can function at higher speeds than the channel bandwidth will support [4]. High-speed link design has striven to increase the link throughput by using signal processing techniques commonly used for communication over bandwidth-limited channels. Pre-emphasis can be used to flatten the steep roll-off of the channel's insertion loss, and adaptive equalization to remove intersymbol interference (ISI) [5]. Alternative multilevel signaling schemes have also received much attention of late because they reduce channel bandwidth requirements at the cost of signal-to-noise ratio (SNR) [6], [7]. These techniques have extended the reach and speed of electrical links, allowing ~10 Gb/s on-board links to span up to ~75 cm [7]–[9]. Because electrical signaling rates are reaching practical equalization limits, such high-speed link designs must trade-off the cost of improved electrical package elements against increased circuit area and higher power consumption required by advanced equalization. To extend link reach, package designers are considering the possibility of using low-loss dielectrics, smooth copper, innovative via-hole techniques, and new connector technologies [10], [11]. Fig. 1 presents an overview of high-speed link system design. Circuit designers, package designers, and system architects need to work close together to solve system interconnect challenges. An accurate link modeling methodology is essential to this multitiered approach in that one cannot make

Manuscript received May 19, 2008; revised November 12, 2008. This work was supported in part by DARPA, under the IBM Contract HR0011-06-C-0074. This work was recommended for publication by Associate Editor W. Beyene upon evaluation of the reviewers comments.

D. G. Kam, M. B. Ritter, T. J. Beukema, J. F. Bulzacchelli, P. K. Pepeljugoski, Y. H. Kwark, L. Shan, X. Gu, C. W. Baks, R. A. John, and G. Hougham are with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (email: kamdong@gmail.com; dgkam@us.ibm.com).

C. Schuster and R. Rimolo-Donadio are with the Technical University of Hamburg-Harburg, D-21073 Hamburg, Germany.

B. Wu is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TADVP.2008.2011138

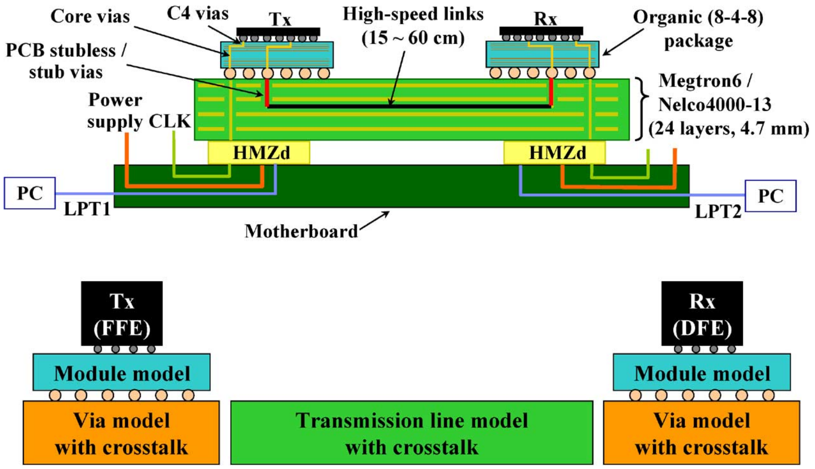


Fig. 2. Description of link which was studied.

rational trade-offs until each solution's effect on the overall link performance is analyzed quantitatively.

With this background, one may ask: "Is 25 Gb/s per channel on-board electrical signaling viable? What package improvements are required to make it happen, and when might optical interconnects on the board be required?" We have been investigating the limits of electrical and optical interconnect performance of future advanced packaging technologies with an eye to answering these questions. Although another study [12] has focused on two modules connected by flex, our study explores module-on-board packaging topologies seen in switches and servers where more than two modules are connected via high-aggregate-bandwidth buses utilizing a dense signal pitch which maximizes escape bandwidth while maintaining adequate signal integrity. A wide variety of high-performance links has been analyzed from a holistic standpoint, considering I/O circuits and equalization, and including all levels of electrical packaging.

We describe the link configurations and packaging technologies aimed at this application space, then show how each element in the electrical link was modeled, followed by model validation against passive hardware measurements. We then present active link measurements at 11 Gb/s and show the correlation with end-to-end link simulations. We use these hardware-correlated models in simulations to predict the performance of dense buses running at 25 Gb/s rates, and we compare this to recent work [13], [14] on on-board optical interconnects. Finally, we discuss maximum achievable data rates, module escape bandwidth limits, and communication metrics with an eye to providing system and chip designers insight into system bandwidth bottlenecks and trade-offs between electrical and optical on-board technologies.

II. Passive Link Modeling

A. Link Description and Modeling Approach

The on-board interconnects studied in this paper include two 90-nm CMOS link chips in organic flip-chip plastic ball grid array (FCPBGA) packages mounted on a printed circuit board (PCB) through ball grid array (BGA) solder joints (or sockets), as shown in Fig. 2 (top). The effect of substituting three different land grid array (LGA) sockets for the BGA solder connection was also investigated. This chip is a product-level version of the prototype described in [15]. The organic chip packages measured 35 mm × 35 mm with an 8-4-8 layer stack-up. Advanced, reduced-stub Nelco4000-13 and Megtron6 PCBs were built at a total thickness of ∼4.7 mm with "reverse side treated" copper foils (the 10 point average surface roughness, $\mathrm{Rz} = 7 \sim 9~\mu\mathrm{m}$) and "profile free" copper layers $(\mathrm{Rz} < 1.5~\mu\mathrm{m})$, respectively. These packaging options were chosen because they balance the need for high-performance designs and materials against practical manufacturing and availability concerns for those solutions. The testbed hardware was partitioned into a large area low-cost motherboard which fed power, control, and clocking to a much smaller daughtercard through HMZd mezzanine connectors. This small-footprint daughtercard allowed a wide variety of bus topologies to be fabricated on a single state-of-the-art high-speed panel. By running the differential transmission lines in a serpentine fashion, we were able to design 15, 30, 45, and 60 cm PCB transmission line lengths on a common coupon size and a variety of near-end crosstalk (NEXT) and far-end crosstalk (FEXT) configurations to explore link performance for various aggressor geometries.

Correspondingly, the main channel model elements can be identified as shown in Fig. 2 (bottom). Instead of trying to obtain one comprehensive model for the entire signal path, individual blocks were modeled separately and the end-to-end channel *S*-parameters were obtained by concatenating the individual channel components. These interfaces were located at stripline boundaries where signal propagation is mostly transverse electromagnetic (TEM) mode. While a comprehensive end-to-end channel modeling is the most accurate approach, it is also computationally the most inefficient. The different feature sizes in modules and PCB, the high aspect ratio of the PCB transmission lines, and the sheer size of the model pose serious problems for any rigorous full-wave simulation. In addition, even small variations (e.g., in the via diameter) would require a full rerun. On the other hand, the partitioning of the full link into smaller blocks allows the following:

1) application of specialized solvers for each problem type and hence an overall reduction in the computational effort;

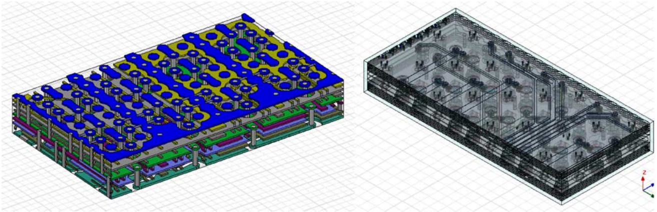


Fig. 3. Module cross section showing C4 escape (left), core vias, and BGA (right) with eight differential pairs.

2) fast parametric variations;
3) a wide range of link topologies to be quickly constructed from a single model library;
4) assessment of the impact of the electrical performance of individual blocks;
5) direct comparison of modeled blocks with measured data.

Full-wave simulations of the package elements including NEXT and FEXT were concatenated to create *S*-parameter models of the entire signal path. Full coupling of eight differential pairs was maintained throughout the signal path to allow exploration of different NEXT and FEXT package pin and via arrangements. As our link chips had 16 differential transmitters and 16 receivers, we created 32-port *S*-parameter models for a number of aggressor situations, simulating near-neighbor pairs relevant to the NEXT or FEXT aggressor arrangements we wished to explore. Some models, particularly the PCB via arrays beneath the modules, required up to five days CPU time to create (using AMD Opteron 2220 SE 2.8 GHz, 2 × 1 MB L2 cache, 24 GB DDR2 memory); therefore, we employed a "Distributed Solve" full-wave simulation tool [16] to reduce simulation time to approximately one day. These models were placed in an interconnect element library, and concatenated by our link analysis tool for active link simulation.

B. Organic Module Models

Eight adjacent differential pairs were selected to capture the channel-to-channel crosstalk. The in-package link was segmented into three sections and modeled with the full-wave solver. The first section includes controlled collapse chip connection (C4) pads, vias, and escape wiring, as shown in Fig. 3 (left). Power/ground pads were parallel to the row of signal pads for worst case analysis of 2:1 signal-to-reference pin ratio at a 200-μm pitch. Vias in the buildup layers had a drill diameter of 60 μm, a pad diameter of 100 μm, an antipad diameter of 225 μm, and a pitch of 200 μm. The second section included 10–15-mm-long coupled differential lines with 25 μm line widths and 50 μm spacing, with 300 μm pair-to-pair separation. The third section included short transmission lines and vias for connections to BGA pads as shown in Fig. 3 (right). Vias in the core layers were 150 μm in drill diameter, 350 μm in pad diameter, 500 μm in antipad diameter, 500 μm in pitch, and 650 μm in length. The BGA pads are on a 1-mm pitch and arranged in a 2:1 signal-to-reference ratio pattern. The dielectric constant is 3.4 in the buildup layers with a loss tangent of 0.017 at 1 GHz. The dielectric constant of the core layers is 4.2 with a loss tangent of 0.02 at 1 GHz.

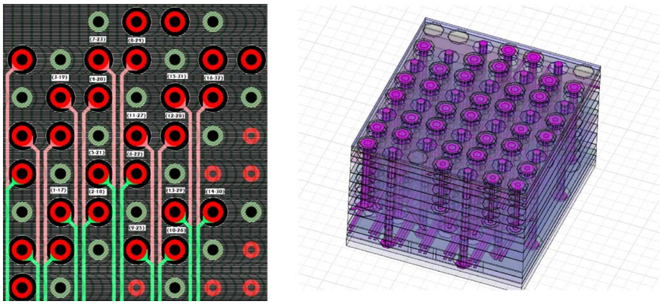


Fig. 4. PCB BGA via escape area showing differential pairs escaping on two wiring layers with a 2:1 signal-to-reference ratio (left), with a cross-section view showing the 24-layer board (right). Eight of these pairs were used in generating a 32-port via model for crosstalk analysis.

C. PCB Via Array Models

The board consisted of two (top and bottom) Megtron6 dielectric subcomposites which were then laminated. Each subcomposite had six signal layers and six power/ground layers. Signal vias were drilled and plated to form half- and full-length vias. Half-length vias (vias through the top subcomposite) had a via drill diameter of 150 μm, a pad diameter of 450 μm, an antipad diameter of 700 μm, and a pitch of 1 mm. The full-length vias had a via drill diameter of 200 μm, a pad diameter of 500 μm, an antipad diameter of 750 μm, and a pitch of 1 mm. For power/ground vias, the drill diameter was 200 μm. The dielectric constant of Megtron6 is 3.5. The total thickness of the board was $\sim$4.7 mm.

We modeled the PCB vias area underneath the module BGA where striplines pass through the via field in order to analyze NEXT and FEXT among neighboring channels. Specifically, to model FEXT of neighboring channels, we included eight pairs of vias connecting eight transmitters (or receivers) in one model. Similarly, to model NEXT of neighboring channels of one link chip, four transmit and four receive via pairs were modeled. In either case, we employed 32-port via models for crosstalk analysis.

Fig. 4 shows a top view of such a 32-port via model used to model FEXT in the PCB via field for eight differential transmitter channels. In this case, the signal-to-reference ratio was 2:1. Three-dimensional via geometries were extracted from the board layout file, then imported and analyzed using the full-wave solver up to 35 GHz.

D. PCB Transmission Line Models

An internal 2.5-dimensional tool, CZ2D [17], was used to create length scalable models of eight differential pairs with full coupling and geometries based on measured cross-sections of transmission lines of Nelco4000-13 or Megtron6 subcomposite cards. An RLGC model was first created which could then be used to quickly generate *S*-parameters for coupled transmission lines of the desired length. Accurate data for the transmission line segments on the PCB were obtained separately using the recessed probe launch technique described in [18]. Transmission line test coupons with recessed probe launch structures were designed into each advanced PCB panel. A frequency-dependent effective loss tangent was extracted by fitting RLGC models to the transmission line coupon measurements. Fig. 5

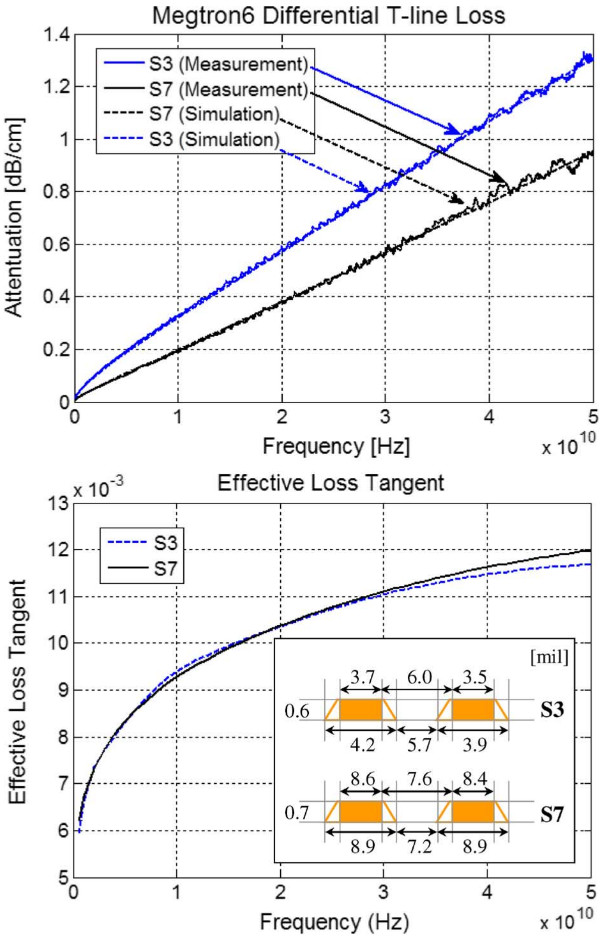


Fig. 5. Representative measured insertion loss (top) and extracted loss tangent (bottom) for Megtron6 striplines.

shows model-hardware correlation for layers S3 and S7 which have different transmission line widths (see inset at bottom for the measured cross-section geometries of the transmission lines). The frequency-dependent effective loss tangent, which accounts for surface roughness induced loss in addition to dielectric loss, was fed back into the transmission line model generation methodology to assure accuracy.

E. Validation of Modeling Approach

Verification of the various elements of the package simulations relied on *S*-parameter measurements taken with a 4-port 50-GHz vector network analyzer (VNA) using RF microprobes. Measurements were also taken at the BGA pad level on the PCB; additional measurements with unpopulated FCPBGA modules soldered onto the BGA pads provided full end-to-end measurements of the passive link as shown in Fig. 6. On-chip parasitics such as pad and electrostatic discharge (ESD) circuit capacitances (380 fF, in total) were incorporated into the full link simulation as a 4-port *S*-parameter model.

The segmented package models described above were concatenated using the Agilent ADS tool [19]. Fig. 7 compares a link comprised of two organic modules and 45-cm-long Megtron6 striplines to VNA measurements of this channel. The modeled *S*-parameters show good correlation with the VNA measurements, agreeing to within ±1.0 dB at frequencies up to 10 GHz, and within ±1.2 dB up to 20 GHz. Much of the residual ripple in the measured data was due to coupling to adjacent transmission lines which could not be terminated in the measurement as they were too numerous. When we measured the same channel with neighboring nets terminated by 47-Ω surface mount technology (SMT) chip resistors, the discrepancy went away.

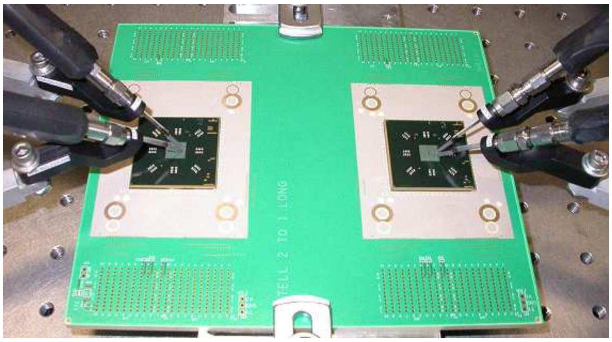


Fig. 6. Complete end-to-end passive link measurement on modules soldered to Megtron6 daughter card.

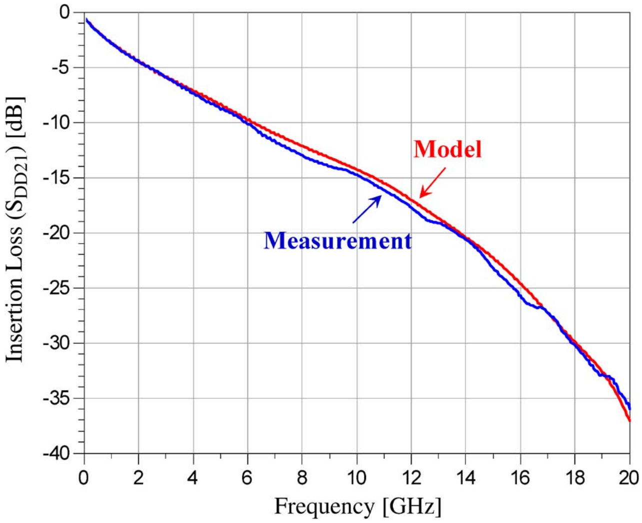


Fig. 7. Passive channel simulations for channel comprised of two organic modules and 45-cm PCB transmission lines agree with VNA measurements to within ±1.2 dB to 20 GHz.

III. Active Link Modeling

A. Active Link Characterization

The measurements on the end-to-end active link were performed using the setup shown in Fig. 8 and schematically in Fig. 2. The heart of the testbed consists of the link chip and the physical implementation of the high-speed links with advanced organic modules and various PCB technologies. The rest of the hardware provides support to make the links functional. On each end of the link we used the same 90-nm CMOS programmable 3-tap feed-forward equalizer (FFE) and 5-tap decision-feedback equalizer (DFE) link chip [15], providing up to 16 full duplex channels. The signaling rate could be varied from 7 to 11 Gb/s, primarily limited by the tuning range of the on-chip phase-locked loops (PLLs). By current standards, the

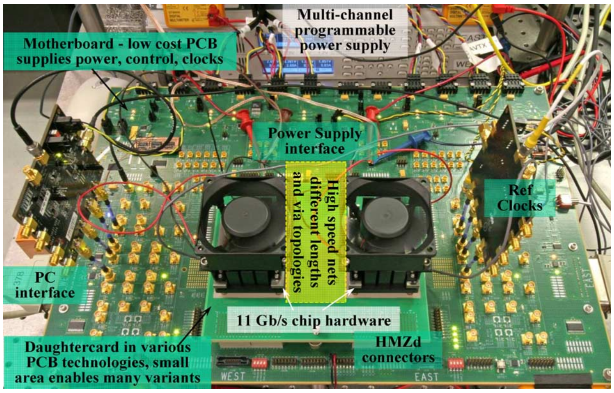

Motherboard - low cost PCB supplies power, control, clocks
Multi-channel programmable power supply
Power Supply interface
High speed nets different lengths and via topologies
Ref Clocks
PC interface
11 Gb/s chip hardware
Daughtercard in various PCB technologies, small area enables many variants
HMZd connectors

Fig. 8. Hardware testbed design.

link chip hardware does not dissipate much power. Since only the cores relevant to 7–11 Gb/s operation need to be powered on, the overall dissipation can be kept within 10 W. As shown in Fig. 8, fans were used on top of the modules since this power level is too high for simple passive cooling solutions without a large area penalty (recall the areal cost of the daughtercard is prohibitive). Preliminary sizings using a test heater module instrumented with a thermocouple were used to determine an adequate cooling solution. The link chip temperature was monitored with an on-die temperature sensor. By exercising judicious power control, the chip temperature can be kept below 50 °C during full link testing. The link chip utilizes many separate power domains to reduce overall power dissipation and to maximize flexibility in exploring chip performance. A high-density power supply rack solution provided eight independent power banks with individual over-current and over-voltage settings for each bank. Reference clocks are needed to drive the on-chip PLLs. Clock boards were designed to provide reference clocks that could be driven from external synthesizers or from a pair of on-board low phase noise precision temperature controlled crystal oscillators (TCXOs). The frequencies of the TCXOs were deliberately offset by 200 ppm so that the phase rotators on the clock and data recovery (CDR) circuits averaged over all phase positions to result in better averaging of eye parameters.

The chip had a slow-speed communication channel, allowing for full programmability of either the transmitter or receiver. In addition, the chip had a variety of registers that contained link quality indicators and stored the state of various chip blocks. Reading and writing to the chip registers was achieved through software that allowed automated control and data collection.

In the configuration shown in Fig. 8, it is necessary to optimize the link performance by selecting optimal FFE and DFE tap coefficients. The DFE tap coefficients were optimized using an algorithm built into the on-chip logic, which relies on link quality indicators that are continuously updated. Typical experiments involved setting the FFE tap coefficients, then allowing the receiver adaptation logic to find the best DFE coefficients. The process was aided by a link simulation package, which helped choose the best FFE tap coefficients. This required constant validation of the hardware environment, porting it into the link simulation package, and then adjusting the FFE taps to check if optimal values have been found.

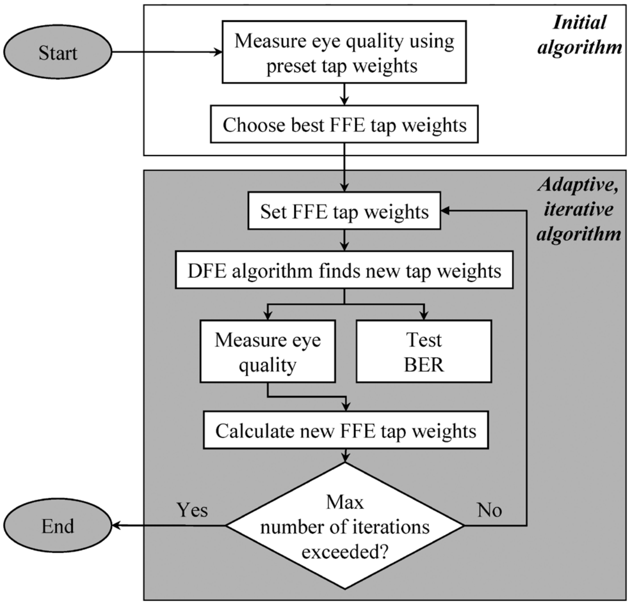

Start
Measure eye quality using preset tap weights
Choose best FFE tap weights
Initial algorithm
Set FFE tap weights
DFE algorithm finds new tap weights
Measure eye quality
Test BER
Calculate new FFE tap weights
Adaptive, iterative algorithm
Max number of iterations exceeded?
Yes
No
End

Fig. 9. Block diagram of adaptive iterative algorithm for FFE tap settings.

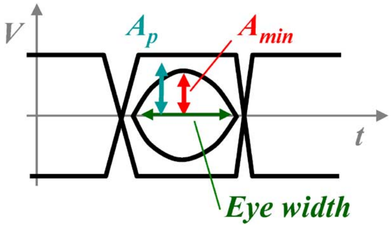


Fig. 10. $A_{\min}$ is a measure of the ISI.

Due to the number of links, link topologies, lengths, advanced PCB materials, and link conditions (e.g., variable amount of crosstalk), it was not possible to manually perform the optimization of the FFE taps as there are a total of 4608 combinations. Instead we customized a link adaptation algorithm [20] and modified the control software to allow full measurement automation.

A general block diagram of an adaptive iterative algorithm to optimize the link is shown in Fig. 9. The chip supplies several link performance measures that each alone or in combination can be used as a cost function. We used the following:

1) $A_{\min}$—the inner eye opening at a bit error rate (BER) of 10^{-3} (error rate set by on-chip counters), as illustrated in Fig. 10. The measurement is a raw number, and it is then normalized with A_p (which is the mean eye height),
2) Eye width—the edge-to-edge eye width at the same 10^{-3} BER,
3) Error count.

B. End-to-End Active Link Modeling and Validation

An internal link modeling tool, HSSCDR [21], was used to simulate link performance given various crosstalk and channel impediments. These simulations employed behavioral models of the link chip I/O circuits including transmitter FFE, receiver

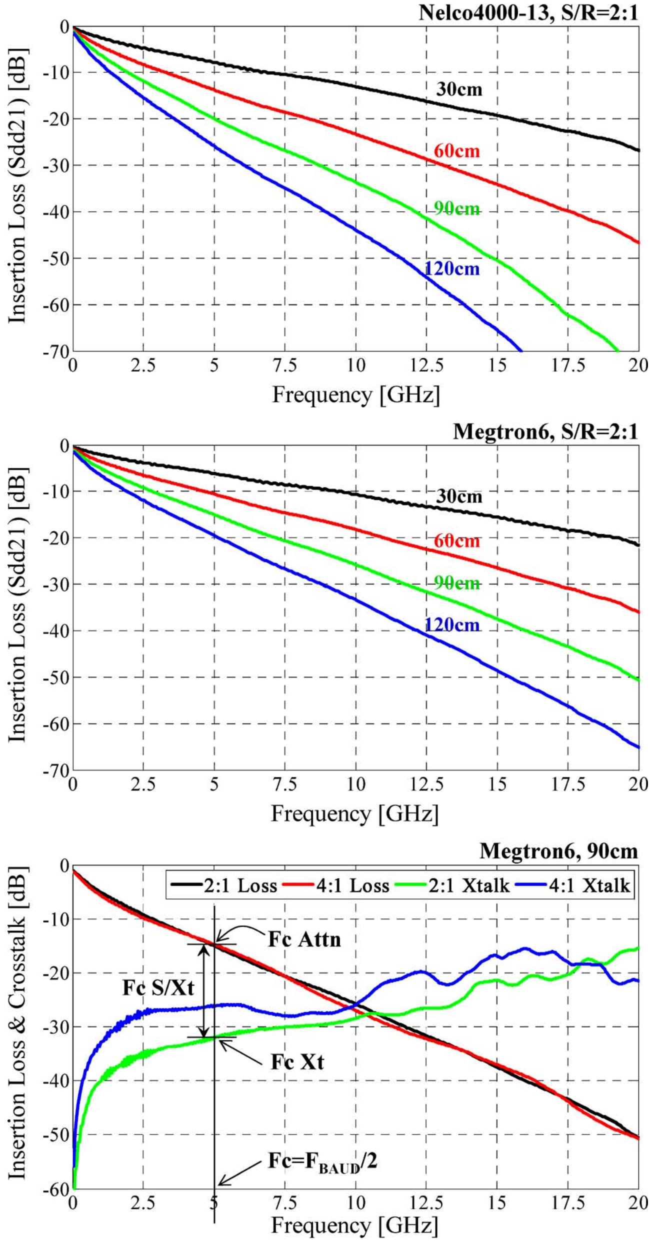


Fig. 11. A sampling of link insertion loss of various interconnect channels on either Nelco4000-13 (top) or Megtron6 (middle) with 2:1 signal-to-reference ratio, studied in this paper. The bottom figure compares the power sum of all crosstalk aggressors of a 90-cm Megtron6 channel with 2:1 signal-to-reference ratio to that of the same length channel with 4:1 signal-to-reference ratio.

DFE, as well as transmitter and receiver contributions to sinusoidal, random, and deterministic jitter. Channel behavior was captured in 32-port *S*-parameters, which included all crosstalk terms for eight differential pairs through the entire packaging path. Fig. 11 shows a sampling of link insertion loss and crosstalk of various interconnect channels studied in this paper. In the bottom figure, signal-to-crosstalk ratio, $Fc\ S/Xt$, is defined as a ratio of signal attenuation to the power sum of all crosstalk aggressors at the frequency of half a given baud rate (5 GHz for 10 Gb/s signaling is given as an example here).

The behavioral simulation is based on a linear time-invariant (LTI) channel assumption, enabling fast convolution algorithms to be employed which result in simulation speed on the order of 1 Mbit/min. The LTI model is an accurate representation of the drivers employed in these link chips, and, of course, the channel is linear and time invariant. Traditional SPICE-based transient simulation methods are orders of magnitude slower than this and cannot accurately capture low-probability events and CDR dynamics without prohibitively long simulation times.

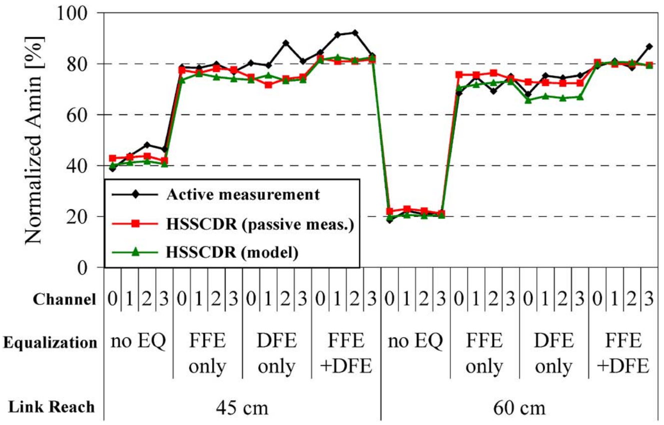


Fig. 12. Good model-hardware correlation was observed for the active links. Note that correlation is given for a variety of channels (labeled 0–3) with different equalization settings and link distances (45 or 60 cm Megtron6 transmission lines with 2:1 signal-to-reference ratio).

In Fig. 12 we show the good model-hardware correlation obtained at 11 Gb/s for a variety of links with 45 and 60 cm PCB transmission lines and with different types of equalization. Also shown in the figure are four different electrical channels, labeled 0 through 3, which had different aggressor geometries. The black diamond curve is data measured with the active link chips via the digital interface, the green triangle curve shows link simulations with full-wave concatenated channel models, and the red square curve shows link simulations using measured *S*-parameters.

Measurement with BGA socketed hardware was also conducted, and the performance at 11 Gb/s was as good as soldered modules. Additionally, measurements were carried out using three different types of LGA socket held at various pressures. In all cases performance at 11 Gb/s was at least as good as BGA modules once a pressure threshold for each was reached where all contacts were electrically closed. This threshold pressure varied across LGA manufacturers from 10 to 60 g/contact.

IV. Exploration of Modulation Schemes

Good model-hardware correlation and flexible simulation and measurement setups have allowed us to explore the performance of other modulation schemes in order to determine the best modulation format to maximize electrical signaling rates.

A. *Multilevel I/O Models*

Duobinary signaling [7] can be generated by sending non-return-to-zero (NRZ) data through a delay and sum filter, which has a Z-transform of $1 + z^{-1}$. Since the frequency response of a typical backplane channel resembles this, if we provide some additional filtering we can generate the required response from the cascade of the filter and the channel. In our link modeling, the reshaping filter was implemented using a transmitter

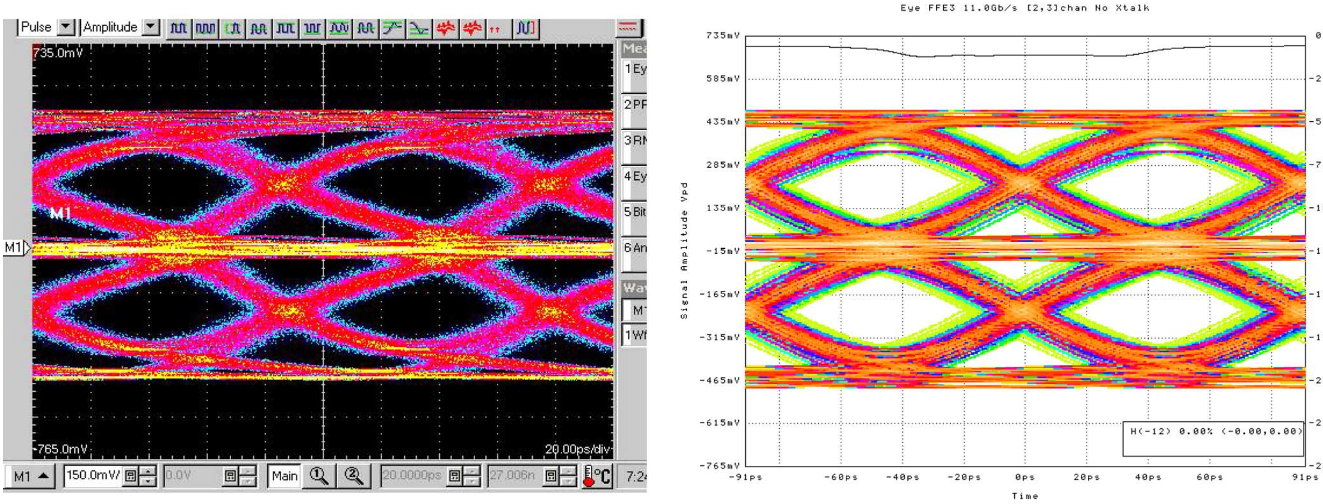


Fig. 13. 11 Gb/s duobinary eye patterns generated by the link chip (left) shows good correlation with eye diagrams generated using our link models (right).

FFE. Optimal tap coefficients were determined by a minimum mean-square error (MMSE) optimization routine, carried out in the time domain. The minimization constraints were error at the edge crossing and error at the data sample point. Duobinary signaling can be viewed as NRZ signaling with a 100 % ISI from the previous bit, so it can be decoded with a two-level NRZ partial-response DFE [22].

Fig. 13 (left) shows that we were able to program the link chip to perform duobinary signaling. In order to see if we can match the appearance of the eye diagrams generated using our link models, the *S*-parameters of a link connecting the output of the transmitter to a digital sampling oscilloscope were measured [2]. An eye diagram was generated using the measured *S*-parameters and our link models including core parameters and FFE tap coefficients used for the duobinary measurements, as shown in Fig. 13 (right). Although this comparison is incomplete in that the impulse response of the oscilloscope sampler should be considered as well, the two eye diagrams show satisfactory correlation. Separately, a four-level pulse-amplitude modulation (PAM4) [6] I/O model was also developed, and the two I/O models were compared to NRZ signaling for different link lengths and data rates.

B. Signaling Comparison Analysis

The good model-to-hardware correlation found in our test results gave us confidence that we could extrapolate our simulations to explore signaling rates beyond 11 Gb/s and distances greater than 60 cm. Each channel model in Fig. 11 was simulated using the I/O core models which were linearly scaled to 2x frequency to estimate performance at higher data rates. The sinusoidal (or deterministic) jitter (SJ) and the random (or nondeterministic) jitter (RJ) of the 11 Gb/s link chip transmitter and receiver clocks were approximately 5% unit interval peak-to-peak ($\mathrm{UI}_{\mathrm{pp}}$) and 0.7% $\mathrm{UI}_{\mathrm{rms}}$, respectively, resulting in 1% $\mathrm{UI}_{\mathrm{rms}}$ clock RJ and 10% $\mathrm{UI}_{\mathrm{pp}}$ clock SJ for the complete asynchronous link. For rate scaling, the jitter terms were modeled as a constant percentage of UI. The rate-scaled core models also incorporated a T-coil network [23] to resonate out ESD capacitance. A 4-tap symbol-spaced FFE with one precursor and two postcursors, and a 5-tap half-rate DFE were assumed for all three signaling options, the launch was 800 $\mathrm{mV}_{\mathrm{pp}}$ differential, and the bit stream was a $2^{15}-1$ pseudo random binary sequence (PRBS). Both vertical and horizontal eye openings at a BER of 10^{-15} were computed, and the results are shown in Fig. 14. Note that the vertical eye opening does not extrapolate to the 800 $\mathrm{mV}_{\mathrm{pp}}$ launch swing for very short PCB links because the automatic gain control (AGC) loop of the receiver attenuates such large input signals to maintain linearity. Representative eye diagrams and bathtub curves of each signaling method are shown in Fig. 15.

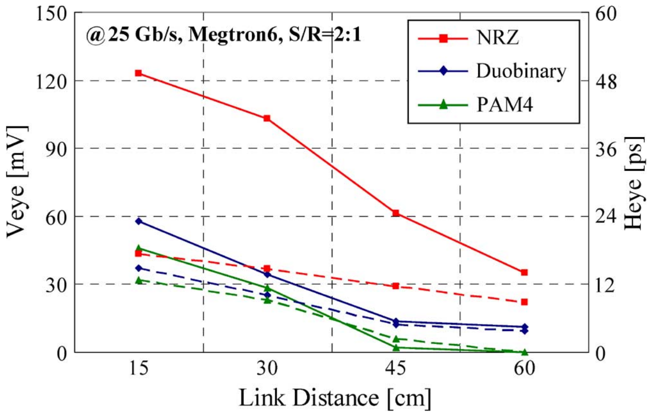


Fig. 14. Simulated vertical (solid curves, left axis) and horizontal eye openings (dashed, right axis) for different modulations at a raw throughput (before modulation) of 25 Gb/s for different PCB transmission line lengths (x-axis). For links longer than 60 cm, all three signaling methods produced closed eyes.

These link simulations show that NRZ signaling with FFE and DFE equalization was superior in performance to either duobinary or PAM4 coding, with the latter showing the poorest performance for the channels considered. Using an eye opening metric requiring 30 mV vertical eye opening and 0.3 UI horizontal eye opening (12 ps), our models predict a maximum reach of ~45 cm for 25 Gb/s NRZ modulation.

In Fig. 16 we show a contour plot of data rate and link reach for each modulation type using this eye opening metric. We conclude that, even with the best signaling scheme, it will be diffi-

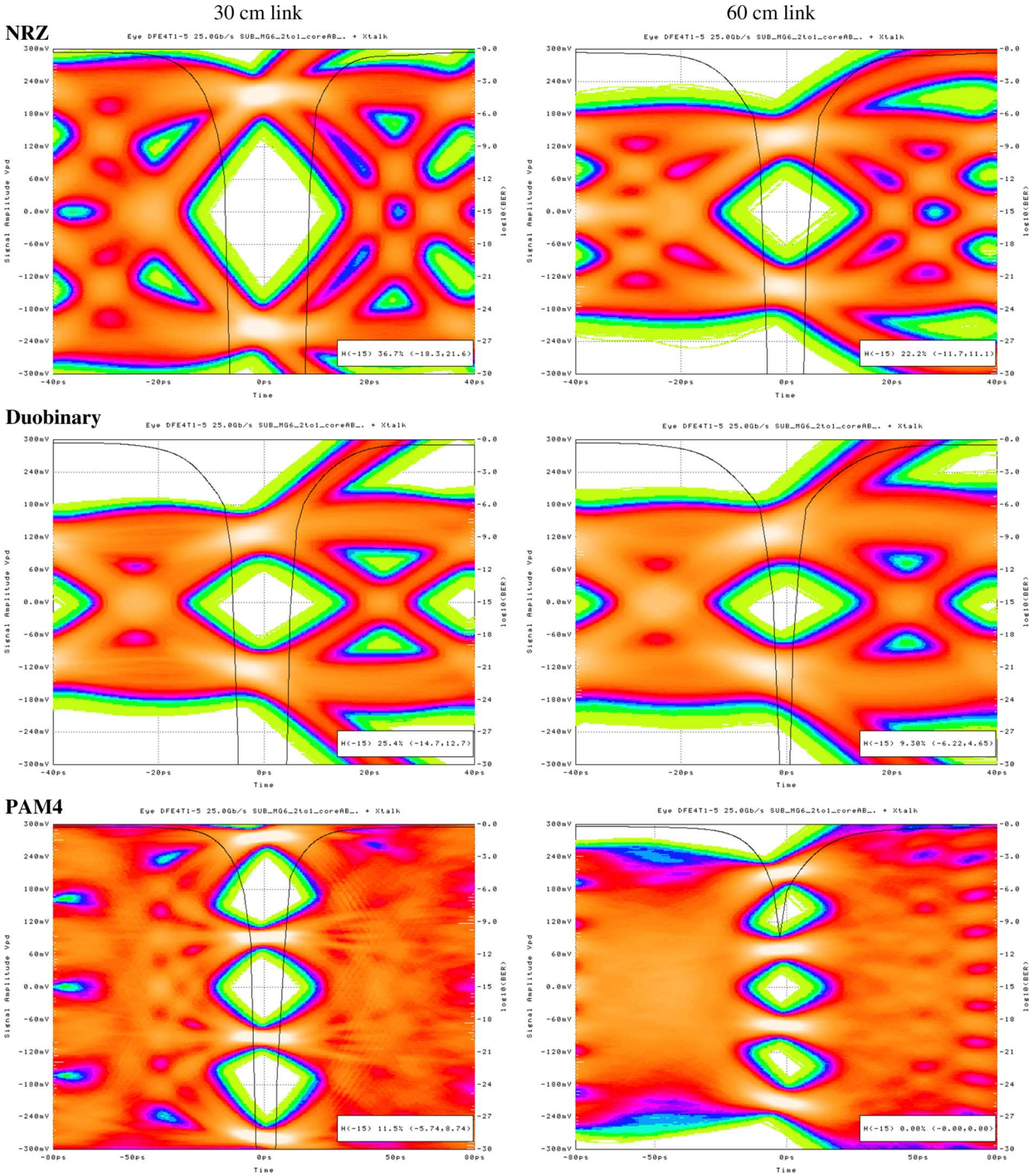


Fig. 15. Representative simulated eye diagrams and bathtub curves of signaling comparison analysis at a raw throughput (before modulation) of 25 Gb/s for 30 cm (left) and 60 cm (right) channels on Megtron6 having 2:1 signal-to-reference ratio.

cult to design dense 25 Gb/s electrical links with a reach greater than 45 cm without wider lines and/or lower loss materials than those used in this study, and that NRZ modulation with FFE and DFE equalization provides the greatest signaling rate at all distances we studied.

Although the data presented here does not necessarily represent the optimum achievable system performance for each signaling method, we believe the results present a fair relative performance assessment of each line signaling approach within a consistent equalization/modeling framework. The resulting data are useful to determine if one signaling format has a clear advantage over the others for application in a range of 25 Gb/s test channels.

C. Conventional Wisdom of Multilevel Signaling Revisited

A PAM4 transceiver divides a signal into four levels, which can be seen as three stacked eye patterns for every cycle. These are encoded as 00, 01, 10, and 11, allowing two bits to be encoded for every symbol time. As a result, the symbol rate with PAM4 is half that of NRZ, so the signal suffers less attenuation. The multilevel nature of PAM4 reduces the level spacing by a factor of three (9.5 dB). The common rationale is that if

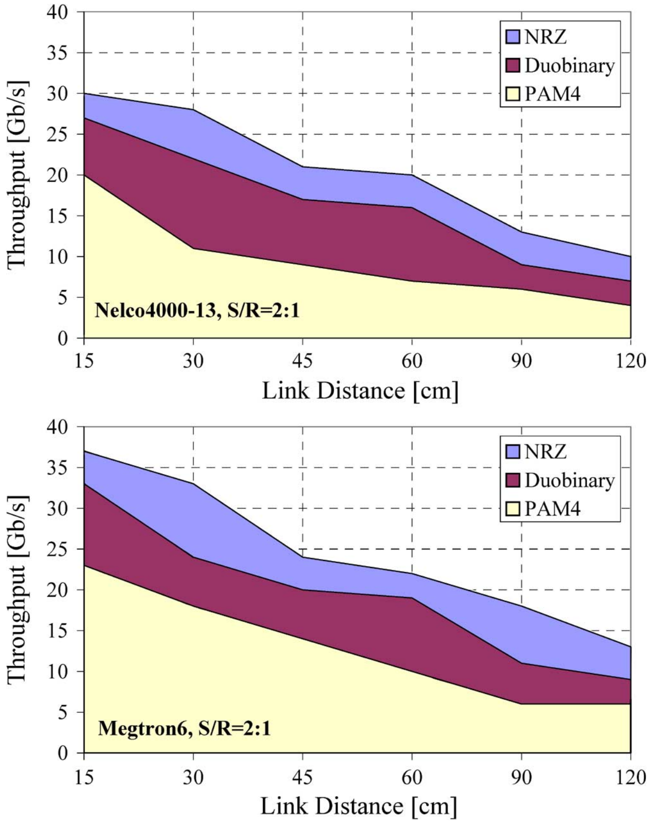


Fig. 16. Maximum raw bit rate (before modulation) versus PCB line length for different modulation schemes.

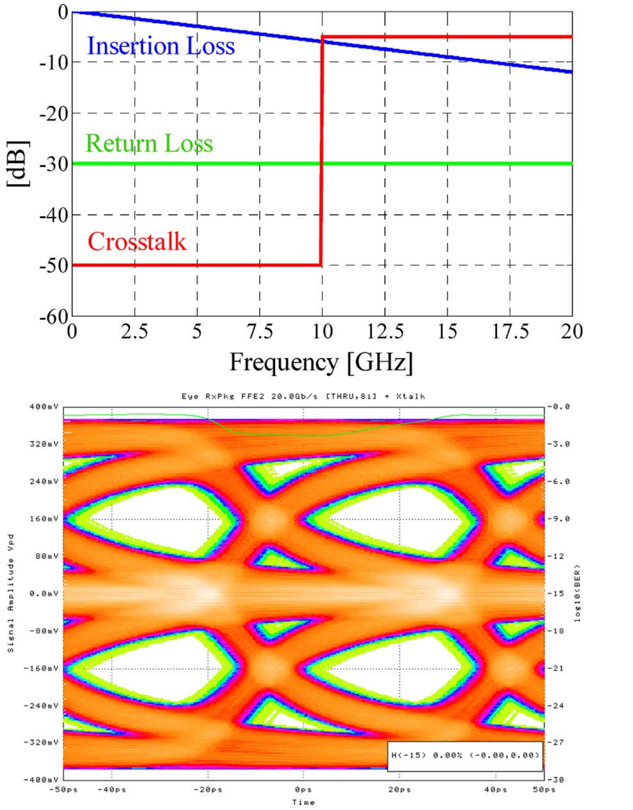


Fig. 17. An example of a duobinary advantageous channel which has (unrealistically) substantial amount of crosstalk only at 10 GHz and above (top); the optimized FFE tap coefficients were [0.506, 0.494] which approximates duobinary signaling at 20 Gb/s (bottom).

the slope of channel loss versus frequency is steep enough, the improvement in SNR due to baud rate reduction may be greater than 9.5 dB, justifying use of PAM4 [24].

Our simulations show that, in the channels we studied, the lower signal bandwidth afforded by multilevel schemes does not result in a better SNR. For the 60 cm link on Nelco4000-13 in Fig. 11 (top), insertion loss at 12.5 GHz (Nyquist frequency for NRZ) is 12.4 dB higher than at 6.25 GHz (Nyquist frequency for PAM4). Furthermore, the insertion loss difference at 12.5 GHz and 6.25 GHz is much bigger than 9.5 dB in every reference channel model provided by CEI-25 working group [25], [26]. Yet, we could find no case where PAM4 showed an advantage over NRZ [27]. This does not follow the conventional wisdom.

In [15] and [28], Bulzacchelli *et al.* explained this dilemma by examining the effect of DFE on insertion loss. DFE feedback is used to cancel ISI due to postcursors in channel impulse response. To observe effect of DFE, they compared discrete Fourier transforms of the sampled channel response before and after eliminating postcursors. They found that elimination of these postcursors flattens the frequency response; therefore, the conventional argument for using PAM4 in high-loss channels breaks down when a DFE is applied to channel equalization. Adding an FFE does not alter this basic analysis though, as such a linear equalizer amplifies high-frequency noise as much as the desired signal, leaving the high-frequency SNR unchanged.

The 9.5 dB SNR penalty is actually just a rule of thumb. PAM4 is three times more sensitive to uncompensated ISI and crosstalk than NRZ since the peak signal to error threshold ratio is three times higher in PAM4 than NRZ. Therefore, PAM4 systems may require significantly more complicated DFE and/or crosstalk cancellation to be viable in challenging channels. Furthermore, since the error threshold is three times smaller in PAM4 for a given transmit launch level, higher transmit launch level and better linearity may be necessary to compensate for loss in receiver sensitivity, which is disadvantageous in low-voltage deep submicron CMOS technology. In [29], Liu and Caroselli indicated that crosstalk cancellation was required to achieve the necessary performance under the channel model and crosstalk assumptions they considered. However, crosstalk cancellation will be very difficult to realize in practical systems. The architecture of crosstalk cancellation is similar to that of DFE; noise at the sampling point is correlated against the aggressor's source stream and subtracted off in a linear summer at the sampling point. Many practical problems arise though, including causality and delay issues with FEXT channels, and intercore routing of high-speed lines in order to be able to design canceling receivers. The issue is made even worse for complex channels, typical in high-end computers, which often experience crosstalk from a number of sources, not necessarily near-neighbor I/O or even from the same bus.

D. Relationship between Duobinary and NRZ Signaling With FFE/DFE Equalization

Duobinary signaling is one type of partial response signaling method in which the binary data are transformed into a three-

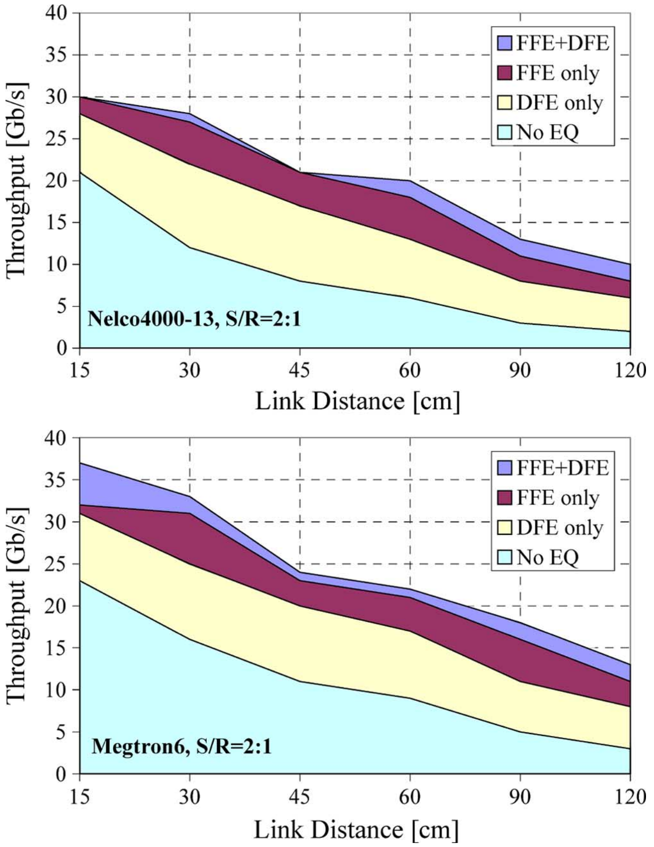
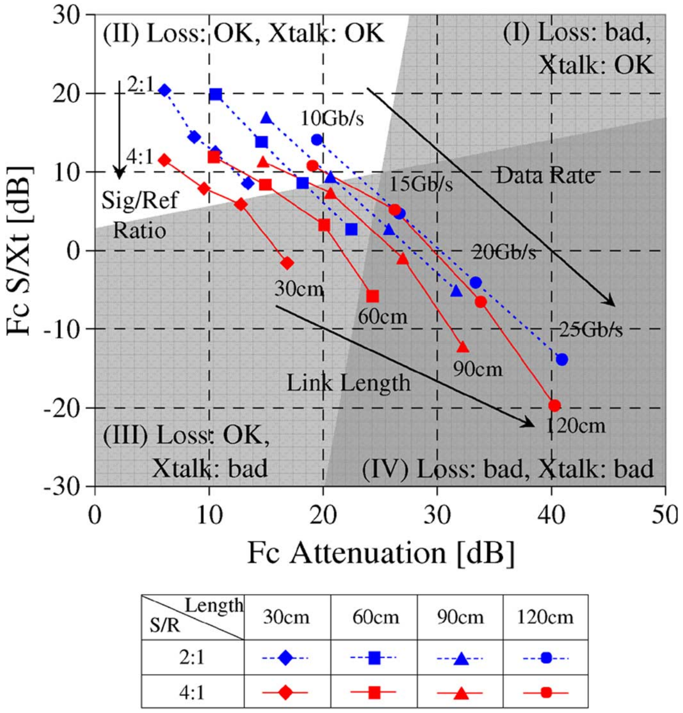


Fig. 18. Maximum achievable data rate versus distance for different amounts of equalization including no equalization, FFE or DFE only, and FFE plus DFE. The same metric (30 mV vertical and 0.3 UI horizontal eye openings) was used in each case.



Length S/R	30cm	60cm	90cm	120cm
2:1	◆	■	▲	●
4:1	◆	■	▲	●

Fig. 19. Insertion loss and signal-to-crosstalk ratio for various Megtron6 channels with 2:1 (red solid curves) and 4:1 (blue dotted curves) signal-to-reference ratios. All curves are shifted downward when signal-to-reference ratio increases from 2:1 to 4:1.

level signal. By introducing correlation between successive bits in a binary signal, the signal spectrum can be forced to be more concentrated in low-frequency region [30].

NRZ signaling combined with FFE equalization can generate partial response signaling (recall duobinary code can be generated and decoded by a baseline FFE/DFE system). Thus duobinary as well as other partial response codes should have been considered by the FFE optimization algorithm as part of the solution space. The FFE optimization algorithm should have homed in on a duobinary solution if it would have given better system performance. Fig. 17 illustrates an extreme and rather unrealistic example of a duobinary advantageous channel, which has substantial amount of crosstalk only at 10 GHz and above. We had the FFE optimization algorithm choose the best tap coefficients of a 2-tap FFE for this channel at 20 Gb/s, and the optimized tap values were [0.506, 0.494], which closely approximates duobinary signaling as shown in the bottom figure. For other channel and crosstalk scenarios, the optimal FFE settings would have been different, implying that duobinary signaling would be a suboptimal solution.

From our measurements and simulations we conclude that duobinary and PAM4 signaling do not perform as well as NRZ with FFE and DFE equalization for channels representative of those we anticipate in various high-speed, high-density computer and switch boards and backplanes. The links we studied have significant loss and enough crosstalk that duobinary or PAM4 signaling produces closed eyes in many cases where NRZ was still able to provide some operating margin. Although it may be possible to improve performance of each line signaling approach by employing equalization architectures more complex than those for NRZ, practical considerations in the design of the I/O including power, area, and voltage limitations favor the relatively simple NRZ-based system architecture in absence of a clear performance advantage of alternate signaling approaches.

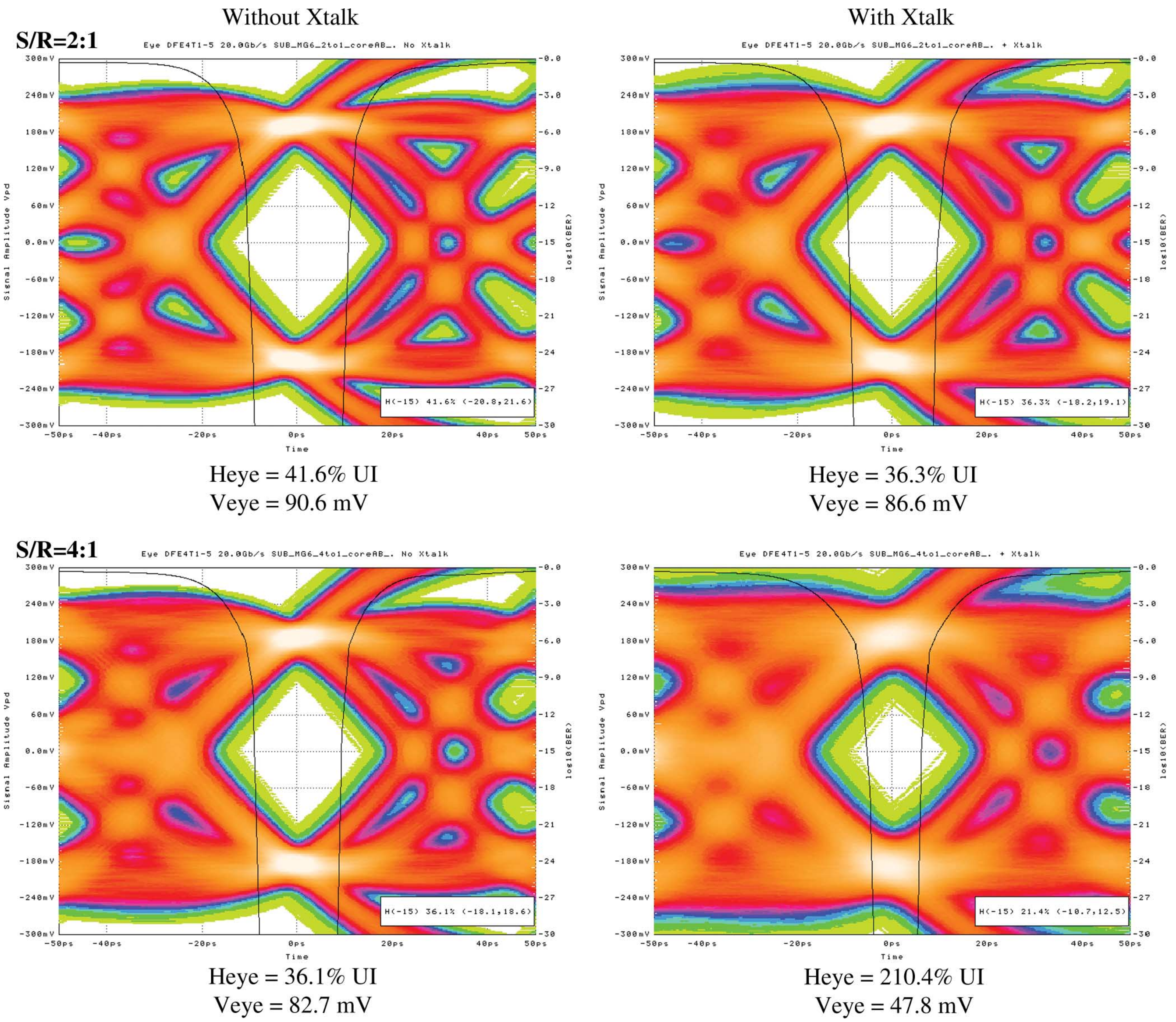
V. Maximum Achievable Data Rates

We first present results for the maximum achievable data rates for electrical interconnects, then compare them to results for optical on-board interconnects published previously [13], [14].

A. *Effect of Equalization and Crosstalk*

In Fig. 18 we show a contour plot of data rate and link reach for different amounts of equalization. Overall, an FFE alone performed better than a DFE alone for the channels tested because of the following major factors.

1) A DFE is unable to cancel out precursor ISI. Highly dispersive channels may have significant time duration of precursor response that can be mitigated through use of an FFE with precursor taps.
2) A nonrecursive DFE can only compensate a fixed time span of ISI. In very low-bandwidth channels, significant postcursor ISI may fall outside the time span covered by DFE taps. On the other hand, an FFE can compensate ISI over a very wide time span since the FFE filter response is convolved with the impulse response of the channel.



Heye = 41.6% UI
Veye = 90.6 mV

Heye = 36.3% UI
Veye = 86.6 mV

Heye = 36.1% UI
Veye = 82.7 mV

Heye = 210.4% UI
Veye = 47.8 mV

Fig. 20. Effect of crosstalk on link performance for 2:1 (top) and 4:1 (bottom) signal-to-reference pin ratios.

However, the functionality of FFE alone systems drops off rapidly over many legacy channels which have spectral nulls (caused by via stubs, connectors, etc.) in the passband requiring numerous FFE taps to cancel reflections. Furthermore, use of a DFE permits less low-frequency de-emphasis at the transmitter resulting in a larger received signal envelope. More discussion of the merits of a combined FFE/DFE system can be found in [5]. The data also indicate that baseline FFE/DFE equalization does not provide reliable operation at 25 Gb/s for high-aggregate bandwidth density types of links longer than 45 cm, so further research is needed in the area of improved equalization system designs to make 25 Gb/s links practical.

Fig. 19 is a plot of both insertion loss and signal-to-crosstalk ratio for various discussed channels with 2:1 and 4:1 signal-to-reference pin ratios, showing the regime of acceptable operation (quadrant II) with contained crosstalk and loss. For those channels which have 25+ dB loss at Nyquist frequency, link simulations show that even FFE plus DFE equalization produces less than 30 mV vertical eye opening. However, this threshold value may vary depending on a number of factors, including minimum sensitivity of the receiver and return loss and crosstalk of the channel. As loss gets lower, smaller signal-to-crosstalk ratio can be tolerated. Conversely, more loss can be handled as crosstalk becomes smaller. Operating boundaries shown in Fig. 19 are rough estimates which may vary significantly as a function of parameters such as reflection ISI and I/O core characteristics.

As data rate or link length increases, the channel performance metric moves from the upper-left to the lower-right quadrant. When increasing signal-to-reference pin ratio from 2:1 to 4:1, signal-to-crosstalk ratio decreases while insertion loss remains almost the same (see Fig. 11). Thus channels with 4:1 module footprint patterns are more likely to be crosstalk limited (quadrant III).

Fig. 20 shows the effect of crosstalk on link performance for different signal pin densities. The 2:1 and 4:1 60 cm Megtron6 channels were simulated at 20 Gb/s, and both vertical and horizontal eye openings were computed at a BER of 10^{-15}. The top left and the bottom left figures are eye diagrams simulated with turning off all aggressors in 2:1 and 4:1 signal-to-reference ratio patterns, respectively. The top right and the bottom right plots show link simulations with worst case crosstalk with 2:1 and 4:1 signal-to-reference ratios, respectively. The transient response is separately calculated for each aggressor, and then the link simulator adjusts the delay of each response to capture the worst case. The degradation of horizontal eye opening due to crosstalk reached 40% for 4:1 signal-to-reference pin ratio, showing that crosstalk is a major limiter of link performance in dense, high-speed buses.

Crosstalk is often beyond the capability of current equalization architectures to combat, and needs to be quantified if accurate performance projections are to be made based on experimental measurements. For short channels, NEXT may be less of an issue since the insertion loss is not as severe; however, in longer links and at higher data rates it has the potential to become a dominant design consideration. It should be noted that the particular links studied in this paper may not have been crosstalk limited at 10 Gb/s, but this does not imply that crosstalk will not be a limiting factor in other link configurations with different types of packages and connectors at the same or even lower data rates. However, these results point out that the escape pattern, as well as proximity of transmit and receive I/O channels must be carefully simulated and designed with sufficient isolation structures to avoid crosstalk dominate channels.

Although we did not have enough test vehicles to assess skew on differential pairs caused by dielectric inhomogeneities (fiber weave, etc.), our active link model-hardware correlation showed this was not a factor for 60-cm links in the board constructions measured at 11 Gb/s speeds. We do not consider fiber weave induced skew a fundamental limit, since a simple rotation of the lines relative to the glass weave largely removes skew issues by averaging.

B. Possible Room for Speed Improvements

Besides the particular electrical 11 Gb/s link implemented in hardware and the extrapolated performance of this link to higher data rates, we considered the ideal case (no IC or module parasitics) and/or using superior equalizers (4-tap FFE plus 20-tap DFE) to gain insight into the possible room for speed improvements.

It was found that both vertical and horizontal eye openings monotonically increased as the number of DFE taps was raised. However, a 4-tap FFE with one precursor and two postcursors seems close to optimal for the channels studied as little performance improvement was observed with longer FFEs. Although increasing the number of DFE taps typically raises power consumption, a 10-tap DFE has been demonstrated with acceptable power efficiency [31]. Furthermore, a number of architectural and circuit techniques for implementing even lower power DFEs have been developed [32].

In Fig. 21 we present the maximum achievable data rate for the electrical links (up to 150 cm) for four cases.

1) The experimental hardware (4-tap FFE/5-tap DFE) but with scaled chip performance (shown in red curve).

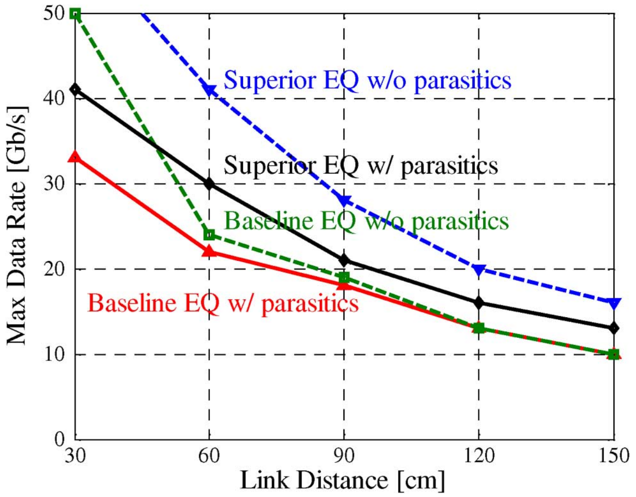


Fig. 21. Maximum achievable data rate as a function of PCB transmission line length. For links shorter than 60 cm, IC or module performance improvements could increase data rates. Links longer than 60 cm are channel bandwidth limited and only superior equalizers can increase data rates.

2) An ideal case with no IC or module parasitics (channel only) with 4-tap FFE/5-tap DFE (green dotted curve).
3) Same as 1) except with 4-tap FFE/20-tap DFE (black curve).
4) Same as 2) except with 4-tap FFE/20-tap DFE (blue dotted curve).

When the baseline equalization was used, we observed that passive channel dispersion limited the maximum achievable data rates for links longer than 60 cm; therefore only marginal improvement could be achieved by improving the I/O circuits and modules. Below 60 cm, however, the channel was not limiting maximum achievable data rates; consequently, improvements in I/O circuit performance (higher bandwidth, better sensitivity, lower jitter, etc.) and module loss could lead to maximum achievable data rates above 30 Gb/s. Although superior equalization (e.g. 4-tap FFE/20-tap DFE) could increase bandwidth further, 25 Gb/s on-board signaling is difficult for links longer than 75 cm.

For sake of comparison, we generated analogous curves (Fig. 22) for module-on-card polymer waveguide-based optical interconnects [13], [14], [33]. In this case, there is a wide gap between the performance of a link limited by the passive optical waveguide bandwidth [34] (upper line, ideal case) and that of the optical link hardware (lower line). As can be seen from the figure, 25 Gb/s on-board optical links are possible to distances of ~1 m. The short, unequalized electrical links between the chips and the optical transceivers limit the maximum achievable data rate of the electrical-optical-electrical (EOE) link to ~26 Gb/s at distances less than 1 m. If FFE and DFE equalization were employed on the short electrical links, and if the optoelectronic (OE) conversion elements were not bandwidth limiting, then the lower "Hardware" limit in Fig. 22 would move upwards towards 35 Gb/s for distances under 90 cm.

C. Electrical Aggregate Bandwidth Limits

For any communication link, there will typically be one or more constraining elements limiting the aggregate bandwidth

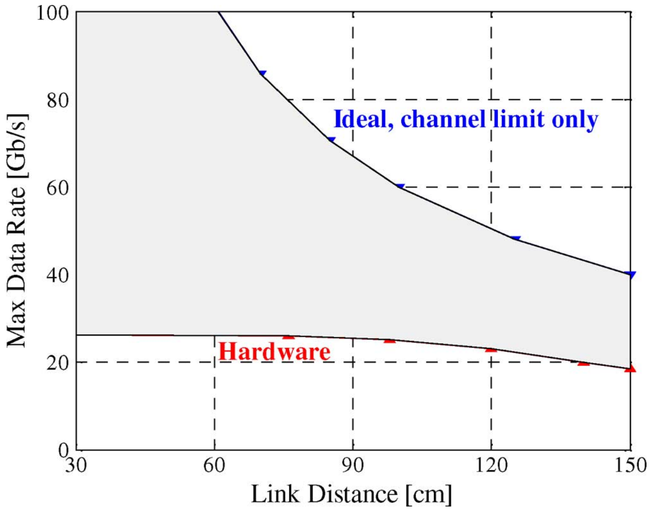


Fig. 22. Maximum achievable data rate as a function of distance for optical interconnects. Optical media is not the limiting factor in the link performance, leaving ample space for improvement of the rest of the components. The unequalized electrical link between the host and the optical modules limits the performance of the EOE link.

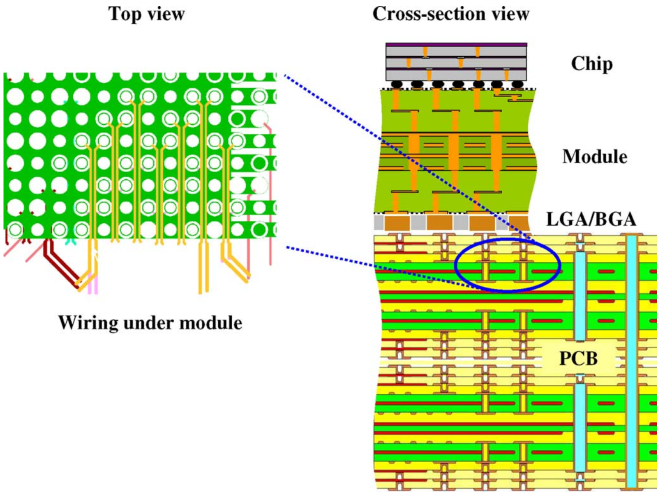


Fig. 23. Physical limits to electrical escape bandwidth. With typical 1-mm LGA/BGA via/antipad full arrays and conductor widths, it is possible to escape only one differential pair per pad pitch per wiring level around perimeter of module.

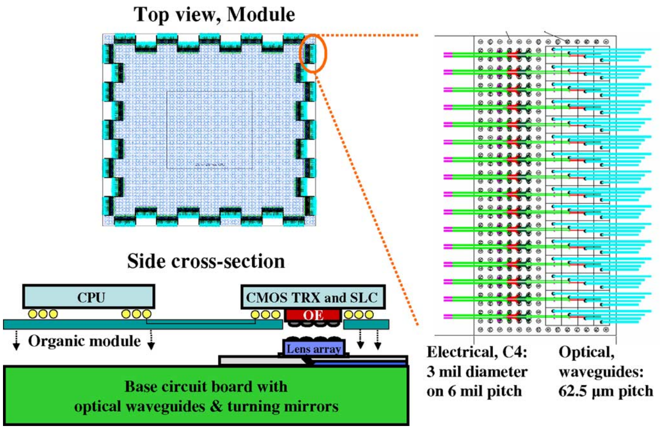


Fig. 24. Physical limits to optical escape bandwidth.

Data Rate	Electrical BW [Tb/s] 1 mm LGA pitch	Optical BW [Tb/s] One WG layer	Optical BW [Tb/s] Two WG layers
10 Gb/s	6.3 to 7.6	23	38.4
20 Gb/s	12.6 to 15.2	46	76.8

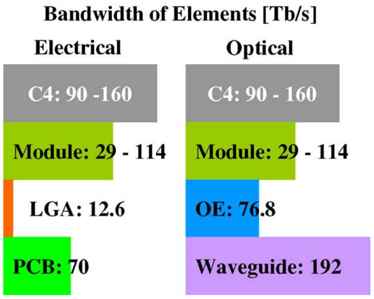
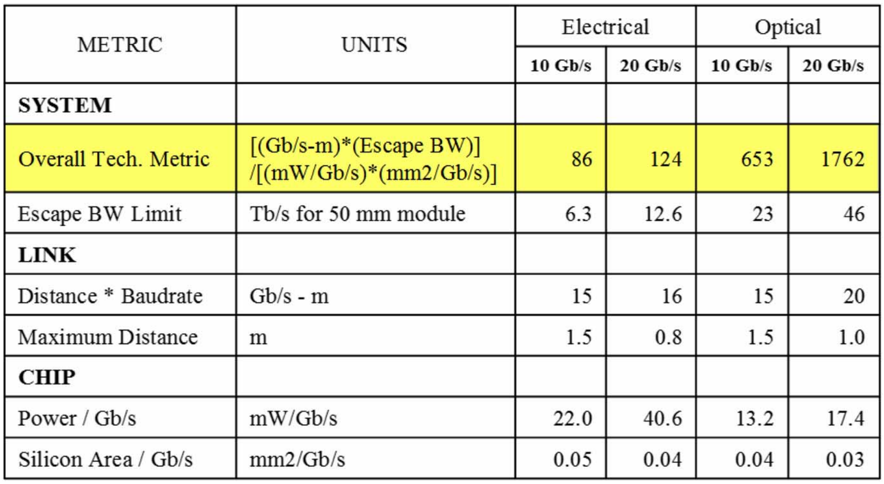


Fig. 25. Module escape bandwidth summary.

of the entire interconnect subsystem. By studying the signaling and physical (escape density) limits for electrical interconnects between two 50 mm × 50 mm organic modules mounted on an organic PCB, we have arrived at our best estimate of the limits of electrical interconnect bandwidth. In Fig. 23 we show a cross section of the packaging structures (right) comprised of a silicon chip with I/O drivers and receivers, the organic module, the LGA or BGA connection from the module to the board, and the PCB. Shown to the left of this cross section is what was found to be the limiting physical constraint—only one differential pair can be wired per channel between the vias in the LGA under the module. This wiring density limit, coupled with the maximum number of signal layers, sets the maximum escape bandwidth at ~12.6 Tb/s for this size module, given that 1900 pins (out of 2500 LGA contacts) are allocated to high-speed signals with 2:1 signal-to-reference ratio. For each differential pair, we have assumed that 20 Gb/s electrical signaling could be used, as the electrical studies have shown a 60 cm reach for this signaling rate $(1900 \times 1/3 \times 20\ \mathrm{Gb/s} \approx 12.6\ \mathrm{Tb/s})$.

PCB wiring, module wiring, and C4 bandwidths are not limiting; in fact, C4 and module bandwidths may increase with future C4 and wiring pitch improvements, and PCB wiring bandwidth may increase slightly with a small increase in the number of wiring layers. However, when we analyze the LGA via array escape, reducing via pitch will actually first decrease escape bandwidth as one will not be able to escape a differential pair in a channel. In this case, only edge vias are accessible, and one must have a stubless board technology to wire out the first "perimeter" of edge via signals, then drop them, continuing the rest of the vias down to the next board layer. Thus one would wire out only perimeter vias on each successive layer, and escape bandwidth would drop until the via pitch was less than 0.64 mm (not likely possible).

TABLE I
SUMMARY METRICS COMPARING ELECTRICAL AND OPTICAL LINKS AT 10 AND 20 Gb/s PER CHANNEL



METRIC	UNITS	Electrical		Optical	
		10 Gb/s	**20 Gb/s**	**10 Gb/s**	**20 Gb/s**
SYSTEM					
Overall Tech. Metric	[(Gb/s-m)*(Escape BW)] /[(mW/Gb/s)*(mm2/Gb/s)]	86	124	653	1762
Escape BW Limit	Tb/s for 50 mm module	6.3	12.6	23	46
LINK					
Distance * Baudrate	Gb/s - m	15	16	15	20
Maximum Distance	m	1.5	0.8	1.5	1.0
CHIP					
Power / Gb/s	mW/Gb/s	22.0	40.6	13.2	17.4
Silicon Area / Gb/s	mm2/Gb/s	0.05	0.04	0.04	0.03

D. Optical Aggregate Bandwidth Limits

Since the LGA (or BGA) escape of the electrical module is the bandwidth pinch-point, it is obvious that this study underscores the need to place OE transceivers on the module next to the switch or processor chip to which they are attached, or no bandwidth improvement over electrical interconnects will be possible. In Fig. 24 (left bottom) we show a cross section of an organic module with a processor chip (CPU) and a representative optical transceiver module (CMOS transceiver (TRX) and surface laminar circuit (SLC), with OE in red). In the top left is shown a top-view of the same module with the outline of a 20 mm × 20 mm processor chip (middle square) and OE transceivers around the perimeter of the 50 mm × 50 mm module. Each of the 36 OE transceivers contains 64 transmitters or receivers grouped with four staggered elements in 16 rows, allowing 62.5-$\mu\mathrm{m}$ waveguide pitch (light blue lines to the right) on the top of the PCB. This results in the maximum escape bandwidth at ~46 Tb/s for this size module ($36 \times 64 \times 20~\mathrm{Gb/s} \approx 46~\mathrm{Tb/s}$).

Shown in Fig. 25 is the comparison of optical and electrical module escape bandwidths, both assuming 50 mm × 50 mm modules. The grey numbers in the electrical column are for 4:1 signal-to-reference ratio module pinout which allows more bandwidth ($1900 \times 2/5 \times 20~\mathrm{Gb/s} = 15.2~\mathrm{Tb/s}$), but, as found in measurement and simulation, also has more crosstalk which we believe will be limiting at 20 Gb/s. For further optical escape bandwidth improvements, an additional waveguide layer can accommodate another rank of OE transceivers on the module. The second rank has only 24 OE transceivers due to the reduced perimeter, giving the maximum escape bandwidth at 76.8 Tb/s ($(36+24) \times 64 \times 20~\mathrm{Gb/s} = 76.8~\mathrm{Tb/s}$). This escape bandwidth estimate may be reduced for die requiring significant on-module decoupling capacitors.

E. Technology Metrics

While the data rates of the links discussed in this paper are mostly limited to less than 25 Gb/s, the ultimate limit of capacity is relatively high [35]. However, data rates in practical high I/O density systems will be limited by power and complexity (or silicon die area) constraints in the equalization system.

Table I presents an overall technology metric (yellow highlight) as well as a number of other metrics which would be useful to system designers when considering either electrical or optical technologies. The left two columns give the metric and the units for that metric, the right four columns give the metrics for electrical 10 and 20 Gb/s links, and optical 10 and 20 Gb/s on-board links, respectively. There are three groups of rows: the first gives overall system metrics, the second gives link or media metrics, and the third gives chip-level metrics. Link metrics deal with electrical or optical link, or media, metrics, such as the distance-baud rate product. The chip metrics deal with power and area efficiency of the I/O on the processor/switch chip. To better represent the state-of-the-art, the electrical 10 Gb/s power models are based on a newer product core (in 65 nm technology) than the one used in the link demonstrations of Section III. The electrical 20 Gb/s power models are based on estimates of a mockup hardware design (also in 65 nm technology). The power numbers of the optical links only include power on the processor or switch chip, and do not include the OE conversion power [13]; if that were included, optical and electrical link efficiencies would be roughly equivalent at 10 Gb/s. The overall technology metric is a product of the distance-baud rate product with escape bandwidth normalized by I/O power and area efficiency. The higher escape bandwidth and the lower power required for I/O on the processor/switch chip give optical technology the advantage.

VI. CONCLUSION

25 Gb/s on-board signaling is difficult at present, for both optical and electrical technologies. Electrical signaling reach is constrained by channel dispersion characteristics, which may improve with reduced dielectric and conductor losses. With existing organic modules and board materials with 150 $\mu\mathrm{m}$ PCB traces, electrical 25 Gb/s links are limited to ~45 cm reach. Adding more DFE taps at the costs of more power and area allows increased reach to ~75 cm.

NRZ signaling with FFE and DFE equalization provides better margins than multilevel modulation. Since duobinary is a subset of the potential solution space of FFE equalization, equalized NRZ should be equivalent or better than duobinary on most channels. The conventional wisdom for using PAM4 in high-loss channels breaks down when a DFE is applied to channel equalization. Although DFE equalization is challenging at these speeds, there is no fundamental implementation barrier, especially if parallel path speculation or loop unrolling is employed [36].

In contrast, optical on-board links of the type referenced here are presently limited by CMOS receiver circuit performance and by waveguide light scattering loss—not by signal dispersion in the optical waveguide, which could support signaling at much higher rates. Theoretically, data rates beyond 30 Gb/s could be achieved on the short electrical segments of the EOE link by adding I/O equalization and/or by using higher performance packaging. New materials and better processing to reduce waveguide loss will most likely extend on-board optical link reach. Error-free vertical-cavity surface-emitting laser (VCSEL) links running at 20 Gb/s have been demonstrated [37], and there is no fundamental barrier to direct laser modulation at 35+ Gb/s for short-reach links [38]. For higher speeds and greater channel density, much effort is being expended on indirect modulation devices—especially silicon nanophotonics [39]. Therefore, from a channel and OE device perspective, optical links do show the greatest promise in improving both bandwidth and reach of dense, high-speed buses. More discussion of electrical and optical trade-offs can be found in [33].

For both optical and electrical links at these speeds, CMOS I/O circuit designs will be challenging. As CMOS scales toward the 22 nm node, f_T and $f_{\max}$ are improving, but not as they have historically; therefore, designers will need to work closer to device speed limits, but these seem to be practical rather than fundamental issues at 25 Gb/s [40]. I/O power, system power, and cost trade-offs are more likely to determine data rate limits and technology choices.

Electrical escape bandwidths are limited by the module pin pitch, which is largely set by PCB via pitch and escape wiring. For reduced-stub, low-loss boards and links $\sim$45 cm in length, a maximum escape bandwidth of 12.6 Tb/s could be achieved for a 50 mm × 50 mm organic module and 1-mm pin pitch. It is clear that optical links must be mounted on the module to allow greater escape bandwidth, and we estimate that total bandwidths as high as 76.8 Tb/s could be brought off a 50-m module with compact optical transceivers and limited decoupling on the module. DC loss for 850-nm light in state-of-the-art polymer waveguides now limits reach of these links to $\sim$1 meter, which will likely improve due to processing and materials changes. Because the present optical packaging approach requires multimode organic waveguides, it will be difficult to employ wavelength-division multiplexing (WDM) emitters to extend channel bandwidth.

In conclusion, electrical links are approaching channel dispersion limits at 25 Gb/s speeds for on-board links and distances greater than 75 cm. 25 Gb/s electrical signaling at distances greater than 45 cm will require more DFE taps, more exotic electrical package technologies, or a transition to new interconnect technologies such as waveguide-based on-board optical links. However, it will be challenging to implement cost-effective interconnect solutions using either technology beyond 25 Gb/s per channel without significant technological advances.

ACKNOWLEDGMENT

The authors would like to thank P. Metty, J. Garlett, D. Stauffer, D. Friedman, F. Doany, C. Schow, D. Kuchta, and J. Kash for valuable advice and assistance, and M. Taubenblatt and M. Soyuer for technical and managerial support of this project.

REFERENCES

[1] A. F. Benner, P. K. Pepeljugoski, and R. J. Recio, “A roadmap to 100G Ethernet at the enterprise data center,” *IEEE Commun. Mag.*, vol. 45, no. 11, pp. 10–17, Nov. 2007.

[2] D. G. Kam, T. J. Beukema, Y. H. Kwark, L. Shan, X. Gu, P. K. Pepeljugoski, and M. B. Ritter, “Multi-level signaling in high-density, high-speed electrical links,” in *IEC DesignCon*, Santa Clara, CA, Feb. 4–7, 2008.

[3] T.-C. Chen, “Where CMOS is going: Trendy hype vs. real technology,” in *Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 6–9, 2006, pp. 1–18.

[4] D. G. Kam and J. Kim, “40-Gb/s package design using wire-bonded plastic ball grid array,” *IEEE Trans. Adv. Packag.*, vol. 31, no. 2, pp. 258–266, May 2008.

[5] T. Beukema, M. Sorna, K. Selander, S. Zier, B. L. Ji, P. Murfet, J. Mason, W. Rhee, H. Ainspan, B. Parker, and M. Beakes, “A 6.4-Gb/s CMOS SerDes core with feed-forward and decision-feedback equalization,” *IEEE J. Solid-State Circuits*, vol. 40, no. 12, pp. 2633–2645, Dec. 2005.

[6] J. T. Stonick, G.-Y. Wei, J. L. Sonntag, and D. K. Weinlader, “An adaptive PAM-4 5-Gb/s backplane transceiver in 0.25- μm CMOS,” *IEEE J. Solid-State Circuits*, vol. 38, no. 3, pp. 436–443, Mar. 2003.

[7] J. H. Sinsky, M. Duelk, and A. Adamiecki, “High-speed electrical backplane transmission using duobinary signaling,” *IEEE Trans. Microwave Theory Tech.*, vol. 53, no. 1, pp. 152–160, Jan. 2005.

[8] S. Rylov, S. Reynolds, D. Storaska, B. Floyd, M. Kapur, T. Zwick, S. Gowda, and M. Sorna, “10+ Gb/s 90-nm CMOS serial link demo in CBGA package,” *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1987–1991, Sep. 2005.

[9] L. Shan, Y. Kwark, P. Pepeljugoski, M. Meghelli, T. Beukema, C Baks, J. Trewhella, and M. Ritter, “Design, analysis and experimental verification of an equalized 10 Gbps link,” in *IEC DesignCon*, Santa Clara, CA, Feb. 6–9, 2006.

[10] B. Chan, J. Lauffer, S. Rosser, and J. Stack, “PWB solutions for high speed systems,” in *Electron. Compon. Technol. Conf.*, Lake Buena Vista, FL, May/Jun. 2005, pp. 1697–1703.

[11] R. Kollipara, B. Chia, F. Lambrecht, C. Yuan, J. Zerbe, G. Patel, T. Cohen, and B. Kirk, “Practical design considerations for 10 to 25 Gbps copper backplane serial links,” in *IEC DesignCon*, Santa Clara, CA, Feb. 6–9, 2006.

[12] H. Braunisch, J. E. Jaussi, J. A. Mix, M. B. Trobough, B. D. Horine, V. Prokofiev, D. Lu, R. Baskaran, P. C. H. Meier, D.-H. Han, K. E. Mallory, and M. W. Leddige, “High-speed flex-circuit chip-to-chip interconnects,” *IEEE Trans. Adv. Packag.*, vol. 31, no. 1, pp. 82–90, Feb. 2008.

[13] F. E. Doany, C. L. Schow, C. K. Tsang, N. Ruiz, R. Horton, D. M. Kuchta, C. S. Patel, J. U. Knickerbocker, and J. A. Kash, “300-Gb/s 24-channel bidirectional Si carrier transceiver optochip for board-level interconnects,” in *Electron. Compon. Technol. Conf.*, Lake Buena Vista, FL, May 27–30, 2008, pp. 238–243.

[14] F. E. Doany, C. L. Schow, C. Baks, R. Budd, Y.-J. Chang, P. Pepeljugoski, L. Schares, D. Kuchta, R. John, J. A. Kash, F. Libsch, R. Dangel, F. Horst, and B. J. Offrein, “160-Gb/s bidirectional parallel optical transceiver module for board-level interconnects using a single-chip CMOS IC,” in *Electron. Compon. Technol. Conf.*, Reno, NV, May/Jun. 2007, pp. 1256–1261.

[15] J. F. Bulzacchelli, M. Meghelli, S. V. Rylov, W. Rhee, A. V. Rylyakov, H. A. Ainspan, B. D. Parker, M. P. Beakes, A. Chung, T. J. Beukema, P. K. Pepeljugoski, L. Shan, Y. H. Kwark, S. Gowda, and D. J. Friedman, “A 10-Gb/s 5-tap DFE/4-tap FFE transceiver in 90-nm CMOS technology,” *IEEE J. Solid-State Circuits*, vol. 41, no. 12, pp. 2885–2900, Dec. 2006.

[16] HFSS. Ansoft [Online]. Available: http://www.ansoft.com/products/hf/hfss

[17] CZ2D. [Online]. Available: http://www.alphaworks.ibm.com/tech/eip

[18] Y. Kwark, C. Schuster, L. Shan, C. Baks, and J. Trewhella, "The recessed probe launch—A new signal launch for high frequency characterization of board level packaging," in *IEC DesignCon*, Santa Clara, CA, Jan./Feb. 2005.

[19] ADS. Agilent [Online]. Available: http://eesof.tm.agilent.com

[20] P. K. Pepeljugoski, J. A. Tierno, A. Risteski, S. K. Reynolds, and L. Schares, "Performance of simulated annealing algorithm in equalized multimode fiber links with linear equalizers," *J. Lightw. Technol.*, vol. 24, pp. 4235–4249, Nov. 2006.

[21] T. Beukema, "Challenges in serial electrical interconnects at 5 to 10 Gb/s and beyond," in *IEEE SSCS-Denver Section Seminar*, Mar. 2007.

[22] P. Kabal and S. Pasupathy, "Partial-response signaling," *IEEE Trans. Commun.*, vol. 23, pp. 921–934, Sep. 1975.

[23] S. Galal and B. Razavi, "Broadband ESD protection circuits in CMOS technology," in *Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 9–13, 2003, pp. 182–183.

[24] H. Johnson, "Multi-level signaling," in *IEC DesignCon*, Santa Clara, CA, Jan./Feb. 2000.

[25] H. Preisach, "Proposals for CEI25 channels based on lower-k dielectric materials for backplanes and daughterboards," in *OIF PLL Working Group Presentation*, Englewood, CO, Apr. 24–26, 2007.

[26] G. Oganessyan, "Reference channels and additional considerations for backplane implementation of CEI-25G," in *OIF PLL Working Group Presentation*, Englewood, CO, Apr. 24–26, 2007.

[27] D. G. Kam, D. R. Stauffer, T. J. Beukema, and M. B. Ritter, "Performance comparison of CEI-25 signaling options and sensitivity analysis," in *OIF PLL Working Group Presentation*, Kobe, Japan, Nov. 6–8, 2007.

[28] J. Bulzacchelli, T. Beukema, and D. R. Stauffer, "PAM-4 versus NRZ signaling: Basic theory," in *OIF PLL Working Group Presentation*, Budapest, Hungary, May 10–13, 2004.

[29] C. Y. Liu and J. Caroselli, "Comparison of signaling and equalization schemes in high speed SerDes (10–25 Gb/s)," in *IEC DesignCon*, Santa Clara, CA, Jan./Feb. 2007.

[30] A. Sekey, "An analysis of the duobinary technique," *IEEE Trans. Commun. Technol.*, vol. 14, no. 2, pp. 126–130, Apr. 1966.

[31] B. S. Leibowitz, J. Kizer, H. Lee, F. Chen, A. Ho, M. Jeeradit, A. Bansal, T. Greer, S. Li, R. Farjad-Rad, W. Stonecypher, Y. Frans, B. Daly, F. Heaton, B. W. Garlepp, C. W. Werner, N. Nguyen, V. Stojanovic, and J. L. Zerbe, "A 7.5 Gb/s 10-tap DFE receiver with first tap partial response, spectrally gated adaptation, and 2nd-order data-filtered CDR," in *Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 11–15, 2007, pp. 228–229.

[32] J. F. Bulzacchelli, A. V. Rylyakov, and D. J. Friedman, "Power-efficient decision-feedback equalizers for multi-Gb/s CMOS serial links," in *IEEE Radio Frequency Integrated Circuits Symp.*, Honolulu, HI, Jun. 2007, pp. 507–510.

[33] P. Pepeljugoski, M. Ritter, J. A. Kash, F. Doany, C. Schow, Y. Kwark, L. Shan, D. Kam, X. Gu, and C. Baks, "Comparison of bandwidth limits for on-card electrical and optical interconnects for 100 Gb/s and beyond," in *Proc. SPIE*, Feb. 2008, vol. 6897, p. 68970I.

[34] F. E. Doany, P. K. Pepeljugoski, A. C. Lehman, J. A. Kash, and R. Dangel, "Measurement of optical dispersion in multimode polymer waveguides," in *IEEE/LEOS Summer Topical Meetings*, San Diego, CA, Jun. 28–30, 2004, pp. 31–32.

[35] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul./Oct. 1948, 623-656.

[36] S. Kasturia and J. H. Winters, "Techniques for high-speed implementation of nonlinear cancellation," *IEEE J. Sel. Areas Commun.*, vol. 9, no. 5, pp. 711–717, Jun. 1991.

[37] N. Suzuki, H. Hatakeyama, K. Tokutome, K. Fukatsu, M. Yamada, T. Anan, and M. Tsuji, "1.1 μm-range InGaAs VCSELs for high-speed optical interconnections," *IEEE Photonics Technol. Lett.*, vol. 18, pp. 1368–1370, Jun. 2006.

[38] Y.-C. Chang, C. S. Wang, and L. A. Coldren, "High-efficiency, high-speed VCSELs with 35 Gb/s error-free operation," *Electron. Lett.*, vol. 43, no. 19, pp. 1022–1023, Sep. 2007.

[39] W. M. J. Green, M. J. Rooks, L. Sekaric, and Y. A. Vlasov, "Optical modulation using anti-crossing between paired amplitude and phase resonators," *Optics Express*, vol. 15, no. 25, pp. 17264–17272, Dec. 2007.

[40] B. Casper, J. Jaussi, F. O'Mahony, M. Mansuri, K. Canagasaby, J. Kennedy, E. Yeung, and R. Mooney, "A 20 Gb/s forwarded clock transceiver in 90 nm CMOS," in *Int. Solid-State Circuits Conf.*, San Francisco, CA, Feb. 6–9, 2006, pp. 263–272.

Dong G. Kam (S'01–M'06) received the B.S. degree in physics with a double major in electrical engineering, the M.S. and Ph.D. degrees in electrical engineering, all from KAIST, Daejeon, Korea, in 2000, 2002, and 2006, respectively.

From 2006 to 2007, he was with Silicon Image, Sunnyvale, CA, where he was a Member of Technical Staff in the Signal Integrity Group. In 2007, he joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he is currently a Research Staff Member concentrating in the areas of antenna and package development for 60-GHz wireless systems and subsystem analysis of 10+ Gb/s serial I/O links.

Dr. Kam was the recipient of the Best Paper Award at DesignCon 2008. He has published over 35 papers.

Mark B. Ritter received the B.S. degree in physics from Montana State University, Bozeman, in 1981 and the M.S., M.Phil., and Ph.D. degrees from Yale University, New Haven, CT, in 1987.

He presently manages a group focusing on high-speed I/O subsystems, including electromagnetic characterization, link modeling, and subsystem analysis with an eye to optimizing I/O link performance metrics, whether electrical or optical. He and his group have contributed to Fibre Channel, 10 Gb/s Ethernet, and other high-speed communication products and standards, as well as to efficient, physics-based models for vias. He is an author or coauthor on numerous technical publications and holds 20 U.S. patents.

Dr. Ritter was the recipient of the 1982 American Physical Society Apker Award, three IBM Outstanding Innovation Awards, and several Research Division and Technical Group Awards.

Troy J. Beukema received the B.S.E.E. and M.S.E.E degrees from Michigan Technological University, Houghton, in 1984 and 1988, respectively.

From 1984 to 1988, he was a Development Engineer with Hewlett-Packard in the area of communication test instruments. He joined Motorola in 1989 where he contributed to research and development of digital cellular wireless systems. In 1996, he joined IBM, Yorktown Heights, NY, where he is presently a Research Staff Member concentrating in the area of high data rate serial I/O system designs. His general research interests include communication link system design and analysis, with an emphasis on modulation, equalization, clock generation, and synchronization systems realized in mixed-signal deep-submicron CMOS technology for high-density I/O applications.

John F. Bulzacchelli (S'92–M'02) received the S.B., S.M., and Ph.D. degrees in electrical engineering, all from the Massachusetts Institute of Technology (MIT), Cambridge, in 1990, 1990, and 2003, respectively.

From 1988 to 1990 he was a co-op student at Analog Devices, Wilmington, MA, where he invented a new type of delay-and-phase-locked loop for high-speed clock recovery. From 1992 to 2002, he conducted his doctoral research at the IBM T. J. Watson Research Center, Yorktown Heights, NY, in a joint study program between IBM and MIT. In his doctoral work, he designed and demonstrated a superconducting bandpass delta-sigma modulator for digitizing multi-GHz RF signals. In 2003 he became a Research Staff Member at this same IBM location, where his primary job is the design of mixed-signal CMOS circuits for high-speed data communications. He holds two U.S. patents.

Dr. Bulzacchelli received the Jack Kilby Award for Outstanding Student Paper at the 2002 IEEE International Solid-State Circuits Conference.

Petar K. Pepeljugoski (S'93–M'93–SM'03) received the Ph.D. degree from University of California, Berkeley, in 1994.

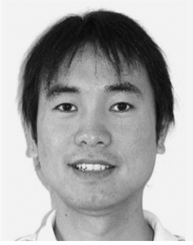
He joined IBM Research as a Research Staff Member in 1994. His research work included design, modeling, prototyping, and characterization of multimode fiber LAN links and parallel optical and electrical interconnects. He was involved in the development of several Ethernet (IEEE 802.3z, IEEE 802.3ae, IEEE 802.3aq) standards, as well as the development of the specification of the OM3 fiber, and subsequently recognized for his contributions. He leads the technical feasibility group for the Call for Interest for 100 Gb Ethernet standard and is actively involved in the work of the Task Force. He is an author or coauthor of more than 60 journal or conference articles.

Young H. Kwark was born in Korea, in 1956. He received the B.S.E.E. degree from the Massachusetts Institute of Technology, Cambridge, in 1978, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, in 1979 and 1984, respectively.

From 1984 to 1986, he was a Research Associate at Stanford University working on high efficiency concentrator photovoltaic cells. In 1986, he joined IBM where he is currently a Research Staff Member at the Watson Research Center, Yorktown Heights, NY. His work has included III-V process development and device characterization, and circuit design for wireless and fiber optic links. His current work focuses on high-frequency measurements of electrical packaging elements used in high performance digital systems.

Lei Shan received the M.S. degree in electrical engineering and the Ph.D. degree in mechanical engineering from Georgia Institute of Technology, Atlanta, in 2000.

In 2001, he joined IBM T. J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member, where he works on high-speed electronics/optoelectronics packaging designs and multiphysics modeling/simulations. He designed and demonstrated high-speed packages on both connectorized format and BGA joints for 50 Gb/s multiplexer and demultiplexer based on IBM SiGe BiCMOS technology. He led the packaging development for 10G Ethernet and Terabus optical links on printed circuit board. His recent research interest is on signal/power integrity in high-performance computing systems and the fundamental electrical limits. He has authored over 40 publications and owns over 20 U.S. patents.



Xiaoxiong Gu received the B.S. degree from Tsinghua University, Beijing, China, in 2000, the M.S. degree from University of Missouri, Rolla, in 2002, and the Ph.D. degree from the University of Washington, Seattle, in 2006, all in electrical engineering.

He is currently a research staff member with IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include characterization of high-speed interconnect and microelectronic packaging, signal and power integrity, and computational electromagnetics.

Christian W. Baks received the B.S. degree in applied physics from Fontys College of Technology, Eindhoven, The Netherlands, in 2000, and the M.S. degree in physics from the State University of New York (SUNY), Albany, in 2001.

He joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, as an Engineer in 2001, where he is involved in high-speed optoelectronic package and backplane interconnect design specializing in signal integrity issues.

Richard A. John received the Associate degree in electronics technology from RCA Institute, New York, in 1967.

He joined the IBM Research Division that same year, and is currently a Senior Laboratory Specialist working in I/O technology on high-speed electrical, optical, and cell phone packaging. He has worked in a variety of areas, including thermal printing, color electro-photographic printing, and involved with the design and fabrication of high-resolution active matrix liquid crystal displays.

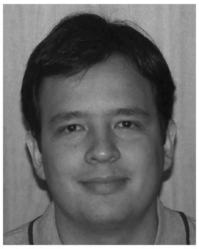
Gareth Hougham received the Ph.D. degree in polymer chemistry from Polytechnic University (now part of New York University), New York, NY, in 1992.

He is a Research Staff Member at the IBM T. J. Watson Research Center and works in the area of interconnect technology with an emphasis on materials science. He and has a long standing interest in low-k organic dielectrics. He is the editor of two volumes on Fluoropolymers published by Kluwer Academic, has authored more than 30 papers, and has more than 50 U.S. patents issued or pending.

Christian Schuster (S'98–M'00–SM'05) received the Diploma degree in physics from the University of Konstanz, Germany, in 1996, and the Ph.D. degree in electrical engineering from the Swiss Federal Institute of Technology (ETH), Zurich, Switzerland, in 2000.

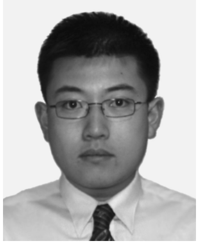
Since 2006, he is Full Professor and Head of the Institute for Electromagnetic Theory at the Technical University of Hamburg-Harburg (TUHH), Germany. Prior to that he was with the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he was involved in high-speed optoelectronic package and backplane interconnect modeling and signal integrity design for new server generations.

Dr. Schuster received the IEEE Transactions on EMC Best Paper Award in 2001, IEC DesignCon Paper Awards in 2005 and 2006, and three IBM Research Division Awards between 2003 and 2005. He is a member of the German Physical Society (DPG).



Renato Rimolo-Donadio (S'08) received the B.S. and Lic. degrees in electrical engineering from the Technical University of Costa Rica, in 1999 and 2004, respectively, and the M.S. degree in microelectronics and microsystems from the Technical University of Hamburg-Harburg, Germany, in 2006, where he is currently working toward the Ph.D. degree in electrical engineering.

Since November 2006, he has been a Scientific Research Assistant at the Institute of Electromagnetic Theory, Technical University of Hamburg-Harburg. In 2007, he was an intern at the IBM T. J. Watson Research Center, Yorktown Heights, NY. His main research interests include system level modeling and optimization of interconnects, and analysis of signal and power integrity problems at PCB and package level.



Boping Wu (S'04) received the B.Eng. degree (with First Class Honors) in electronic and communication engineering from the City University of Hong Kong in 2005, and the M.S. degree in electrical engineering, in 2007, from the University of Washington, Seattle, where he is currently working toward the Ph.D. degree.

He was a Graduate Intern at the Intel Corporation and the IBM Corporation, in 2006 and 2007, respectively. His current research interests include layered medium Green's function, high-speed interconnects and packaging, metamaterials, and nano-photonics.