



9:20 - 10:00:

## *Interconnect Technologies - Trends and Techniques*

Martin L. Schmatz (mrt@zurich.ibm.com)

Nov 7<sup>th</sup>, 2008

1

Workshop Hochgeschwindigkeits-Schnittstellen

IBM Zurich Research Laboratory



## Content

- **Part 1:** Short overview on the IBM Zurich Research Laboratory and the ZRL Systems Department
- **Part 2:** Motivation: Trends in computer systems
- **Part 3:** High-Speed Interconnects: Techniques, and some recent results
- **Part 4:** Conclusions

2

Nov 7<sup>th</sup>, 2008

Workshop Hochgeschwindigkeits-Schnittstellen

© 2008 IBM Corporation

# IBM Research Worldwide



# IBM Zurich Research Lab (ZRL)

ZRL population: ~340 employees

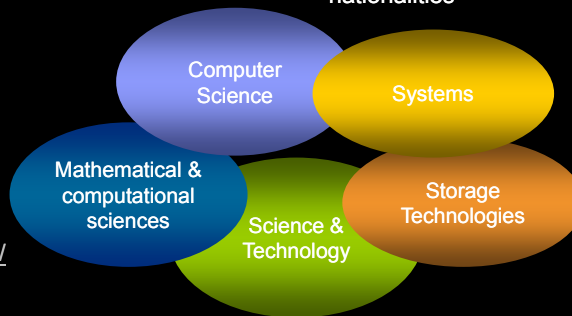
- Regular research & technical staff
- Pre-docs
- Post-docs & visiting scientists
- Students



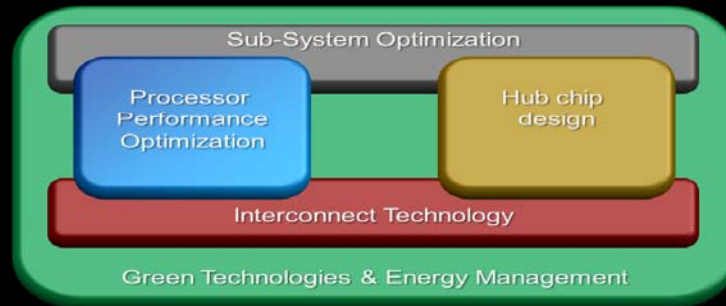
>30 different nationalities

ZRL research projects span all IBM Research strategy areas

<http://www.zurich.ibm.com/>



# Systems Department @ IBM Zurich Research Lab



**Server Technologies**

- Interconnect Technology Design
- I/O Network Optimization

**Accelerator Technologies**

- Performance Acceleration via BFSMs
- Memory Access Technology

**I/O Link Technologies**

- uP and Memory I/Os
- Compute node interconnects

**System Software**

- Remote Direct Memory Access (RDMA)

**Energy Management**

- Data Center Energy Management
- Sensor Networks
- Retail Localization Sensors

## IBM ZRL SYSTEMS DEPARTMENT

Research to...

- ... increase server system flexibility & re- configurability
- ... increase server hardware efficiency
- ... increase server system connectivity
- ... increase processor I/O bandwidth

## IBM ZRL SYSTEMS DEPARTMENT

Research to...

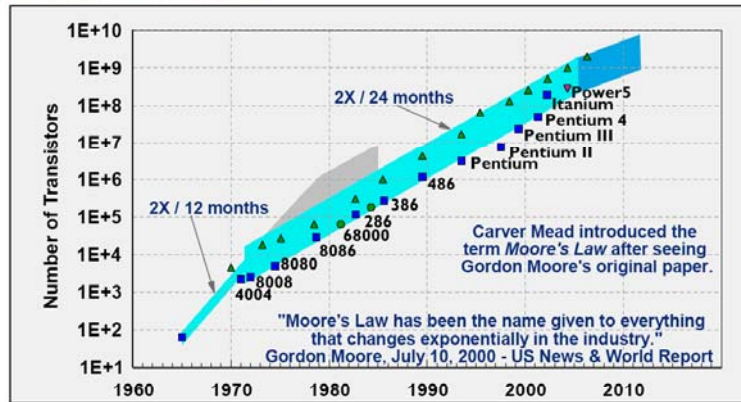
- ... increase server system flexibility & re- configurability
- ... increase server hardware efficiency
- ... **increase server system connectivity**
- ... **increase processor I/O bandwidth**

**Connectivity & I/O bandwidth are becoming key performance factors for computer systems !**

## Part 2:

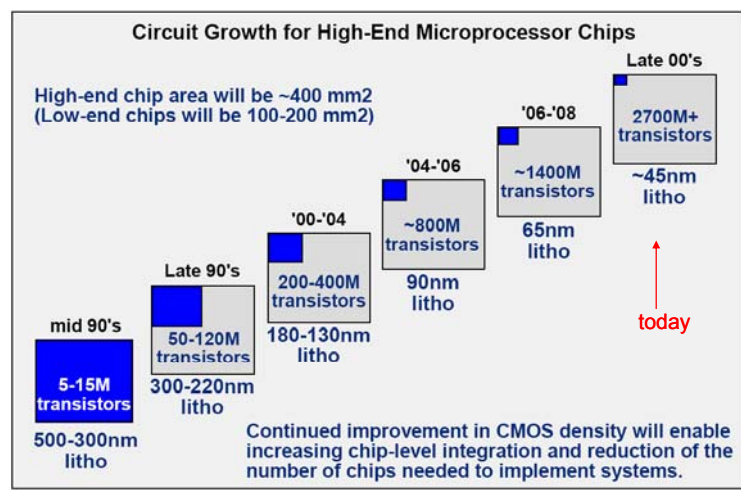
### Motivation for High-Speed Interconnects: Trends in computer systems

## Evolution of Moore's Law

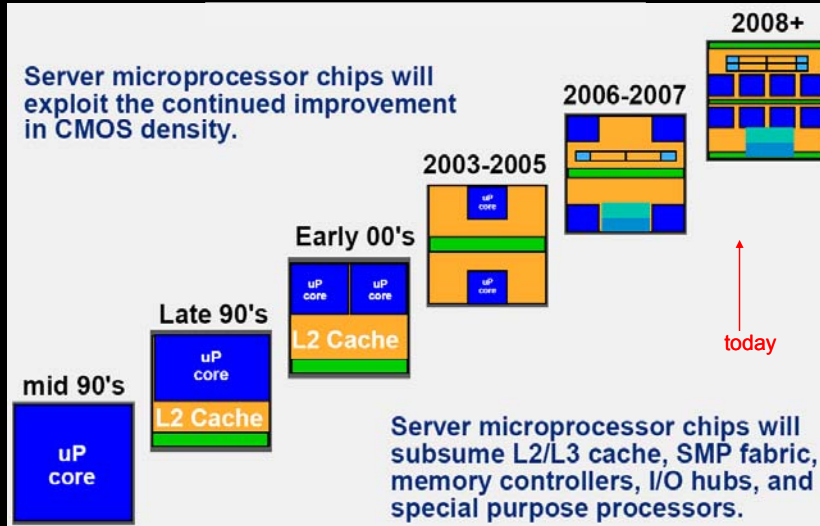


- **Gordon Moore published observations**
  - ▶ April 1965 - Electronics: transistors per chip 2X every 12 months.
  - ▶ Dec. 1975 - IEDM: 2X every 12 months for 1975-79 and 2X every 24 months for 1980-85.
- **David House while at Intel in 1980's: performance doubles every 18 months.**
- **Oct 1989 - Pat Gelsinger: 2X every 24 months through 2000.**
- **Feb. 2003 - ISSCC Gordon Moore keynote: 2X every 24 to 36 months.**

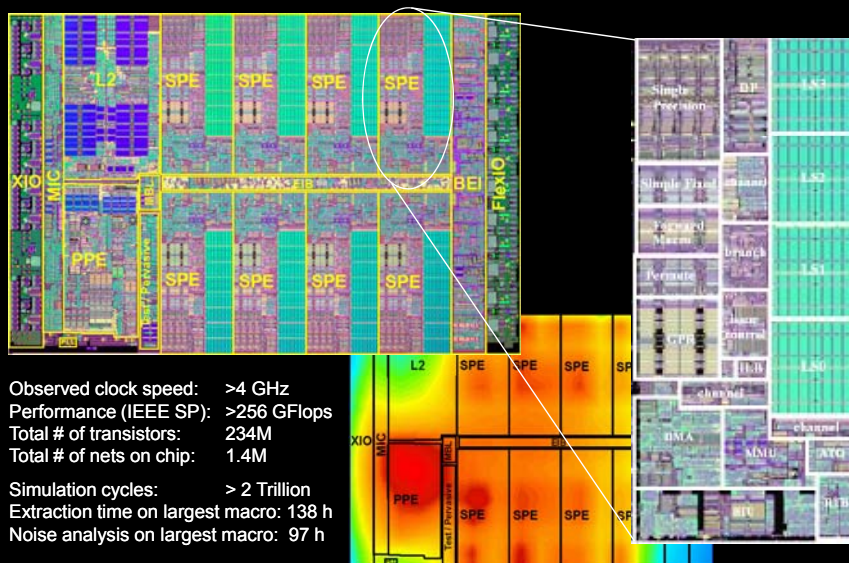
## CMOS Density Trend



# CMOS Density Impact on Server Microprocessors



# Cell Chip: One PowerPC processor PLUS 8 SPE's

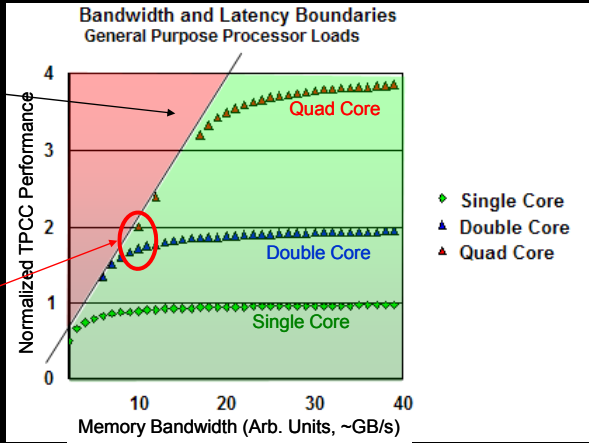


# Multi-Core Chips Stress Memory Bandwidth

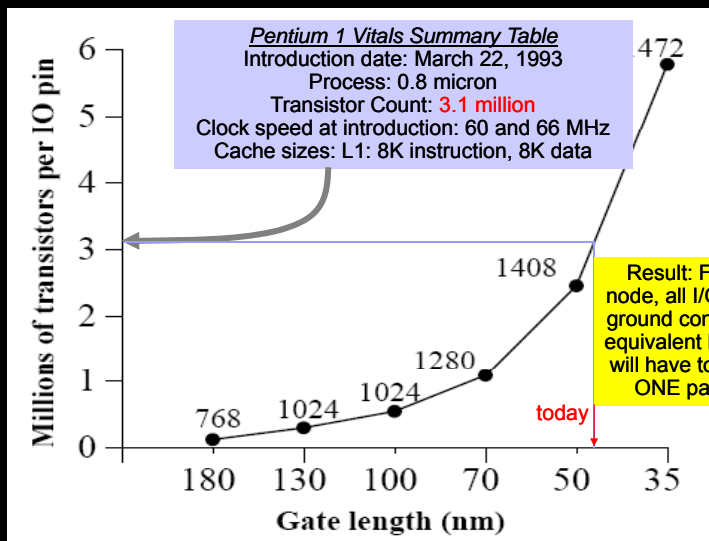
Multiple core systems require larger bandwidth to manage competing access to on-chip cache

Memory starved Cores !!

Only marginal performance improvement from more cores if memory bandwidth is not sufficient !!



# Trend: # of transistors per package pin

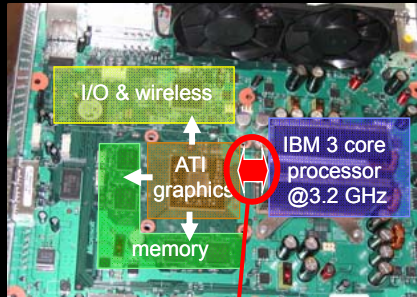


**Pentium 1 Vitals Summary Table**  
 Introduction date: March 22, 1993  
 Process: 0.8 micron  
 Transistor Count: **3.1 million**  
 Clock speed at introduction: 60 and 66 MHz  
 Cache sizes: L1: 8K instruction, 8K data

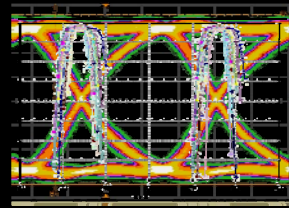
Result: For the 45nm node, all I/O and power & ground connections of an equivalent Pentium 1 chip will have to be served by ONE package pin !!

Source: Intel & ITRS

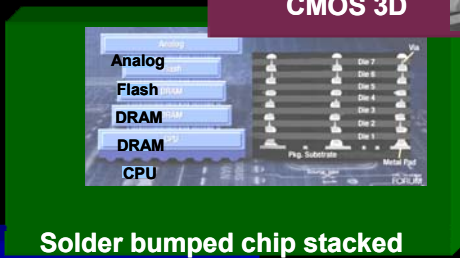
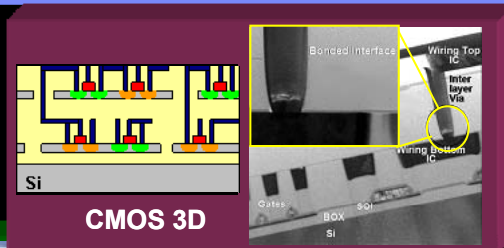
# I/O Link Technology: Microsoft Xbox 360



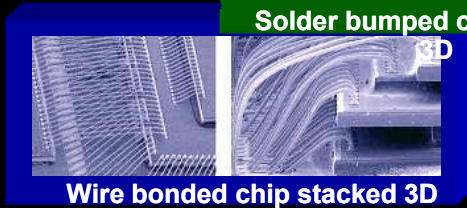
- High-speed front-side bus
  - >100 Gb/s total bandwidth
- ZRL Contributions to
  - Link receiver architecture
  - Analog SOI CMOS circuit design
  - Circuit optimization



# Evolution of 3D



Solder bumped chip stacked



Wire bonded chip stacked 3D

## Fundamental Trends

- Number of cores in a uP increases approx quadratically over CMOS process generations
  - Large number of processing cores per chip becomes feasible, eg. 8 in 2009, 16 in 2012, 32 to 64 in 2015...
- Single-thread performance increase is slowing
  - parallelism in multi-core systems is used to compensate

### However:

- Each additional core pushes the bandwidth requirement
  - Need to increase I/O BW to avoid memory starved cores
- Number of electrical connections to a uP package is growing slowly (~25% per generation)
  - Need to leverage each package pin in optimum way.

## Part 3:

### High-Speed Interconnects: Techniques, and some recent results

# High-Speed Signaling Challenges (at least some of the important ones)

## > Area/Power

- Power efficiency [mW/Gbps] is becoming THE KEY parameter for link architecture selection
- Optimum I/O solution makes best use of package pins [Gbps/pin]

## > Circuit speed

- Designing complex CMOS functions at <<100ps cycle time is challenging; complexity is dominated by equalization
- Clock jitter optimization at low power consumption becomes a significant work area

## > Channel dispersion

- ISI generation due to frequency dependent loss and phase causes eye closure and drives the need for highly complex equalization in electrical link solutions

## > X-talk

- Wastily underestimated in current designs
- Will cause significant amplitude noise generation when going to data rates >>10 Gbps
- Re-reflections cause long impulse response trail of X-talk contributors; in case of impedance variations in transmitter, the X-talk and reflection impulse response is a function of bit pattern

## > Supply noise generation

- Modern CPUs have supply currents in order or excess of 100A, which causes severe issues in the jitter sensitive TX/RX/Clock circuits; supply regulation will become a must.

## > Latency

- Latency is a very critical factor in processor I/O designs, which means that we need to minimize FIFO depth and try to avoid pipelining

## > Process variations

- We need the circuits when CMOS process is still in flux
- Need programmable/adaptive analog circuits
- Need BIST for bring-up and in-situ reliability monitoring

Circuits

Channel

System

Rest

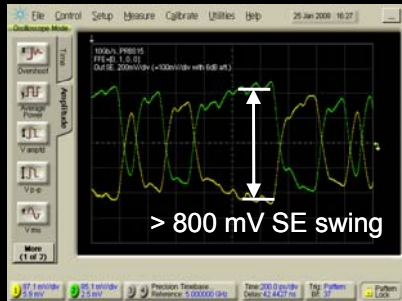
## Part 3:

# High-Speed Interconnects: Techniques, and some recent results

## → Circuits: TX. RX & Clocks

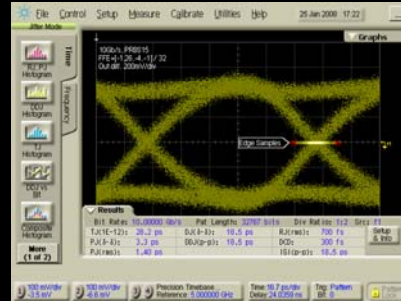
# High-swing 45nm SST transmitter: Measurements

## Signal swing capabilities



- Maximum signal swing (unequalized)  $> 1.6V_{pp\_diff}$
- Signal was not equalized to show the output swing capabilities

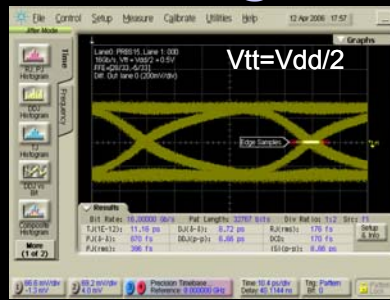
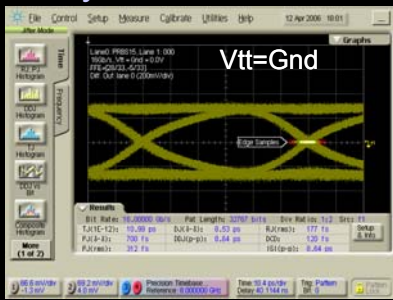
## Verification of jitter performance



- Verified jitter performance (measured without precision timebase)
  - RJ:  $< 1$  ps
  - DJ:  $< 19$  ps
  - Duty cycle distortion:  $< 500$  fs

Note: Above measurements taken with  $V_{dd\_high} = 2.0V$ ,  $V_{dd\_low} = 1.0V$ ,  $V_{dd\_proth} = 0.9V$ ,  $V_{term} = 1.0V$ ; PRBS 15

# TX Eye: Termination to Gnd, Vdd/2, Vdd @16 Gb/s

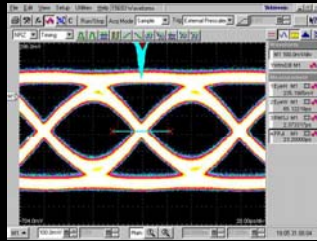


Same driver design is able to address different termination schemes

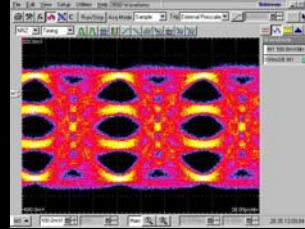
→ Coverage of several standards

→ Termination to VDD, GND, VDD/2 with only one design

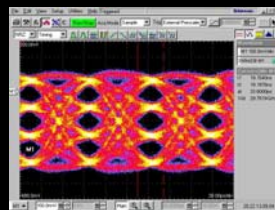
## NRZ & PAM4 Transmitter (1.5m Cable)



12.5Gb/s, PRBS31  
FIR [-2,56,-5,0]  
65ps/235mV diff. eye opening  
99.3 mA @ 1V



100mV/Div, 28ps /div  
20 Gbps PRBS31 pattern  
FIR [-6, 51, -6, 0]  
47ps/124mV diff. eye opening  
93.5 mA @ 1V

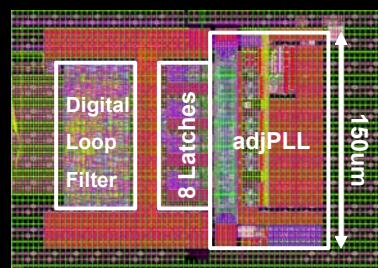


25Gb/s

100mV/Div, 28ps /div

## 40 Gbit/s Receiver in CMOS

- Clock and Data recovery (CDR) and 1:8 Demux
- Architecture based on phase-adjustable PLL (presented at ISSCC 2005/07)
- Integrated Error Detector for PRBS 15
- High CMOS content (energy /bit = roughly constant over data rate)
- No inductors required → can be easily integrated in digital process



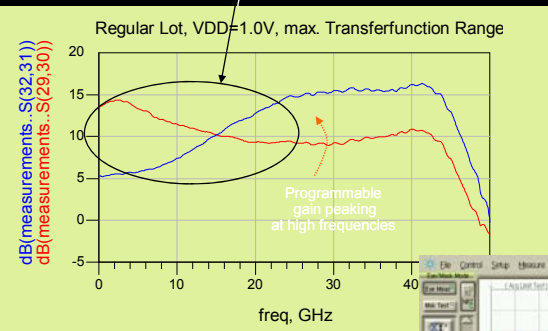
### Several world records established/broken:

- First 40 Gbit/s Receiver in CMOS (with BER < 10<sup>-12</sup>)
- Power consumption = 1.1 mW/Gbps for 12.5 – 30 Gbit/s @ 1.0 V  
1.6 mW/Gbps for 40 Gbit/s @ 1.2 V
- Area = 0.028 mm<sup>2</sup> (corresponds to 1.4 Tbit/s/ mm<sup>2</sup>)

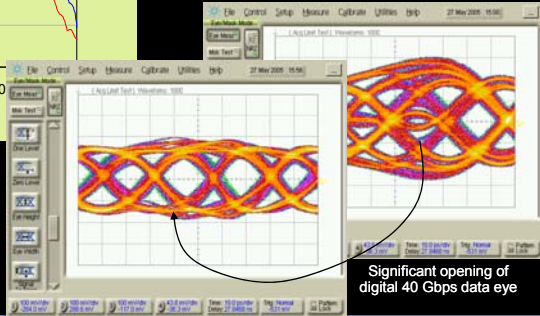
# Differential equalizing 50 Gbps amplifier

50'000'000'000 bits per second

Continuous time RX equalization by low-frequency gain variation:



Example:  
40 Gbps equalization



44 GHz nominal bandwidth with very low power consumption of less than 1.7 mW/Gbps and programmable peaking.

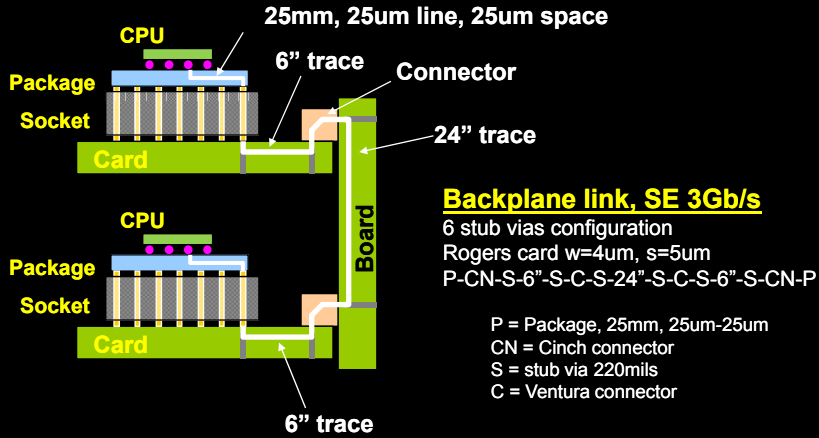
## Part 3:

### High-Speed Interconnects: Techniques, and some recent results

→ Channel

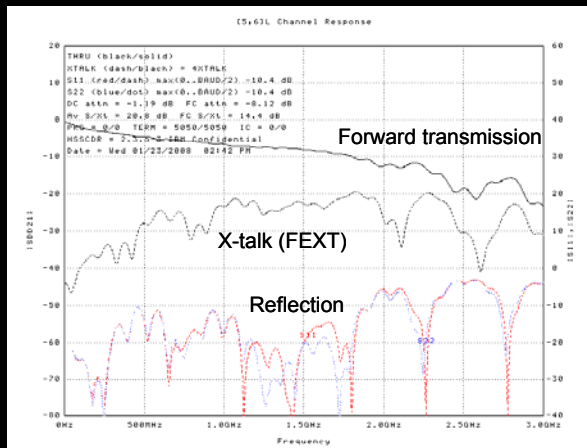
# Electrical Link Performance Limit Analysis

## Link configuration:



Source: Jeremy Schaub, Dulce Altabella, Mark Ritter, Dan Dreps, "Signaling study and future directions for BW elevation", IBM

# Channel Transfer Function



At high frequencies, the contributions from **X-talk** and **reflections** becomes a non-negligible (read: dominant) term in the noise analysis.

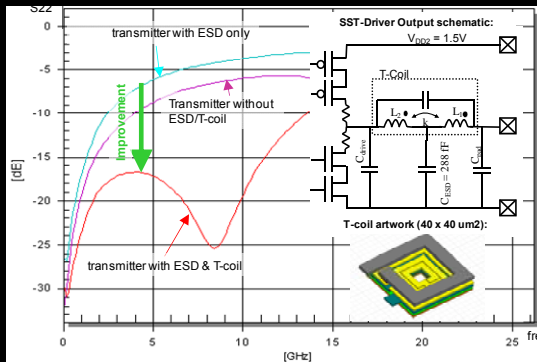
# Impedance matching in presence of ESD

## Issues

- ESD diode represents large capacitive load
- Resulting complex output/input loads lead to significant reflections at I/O ports
- Resulting low-pass filter reduces signal swing and adds jitter

## Goals of Demonstrator

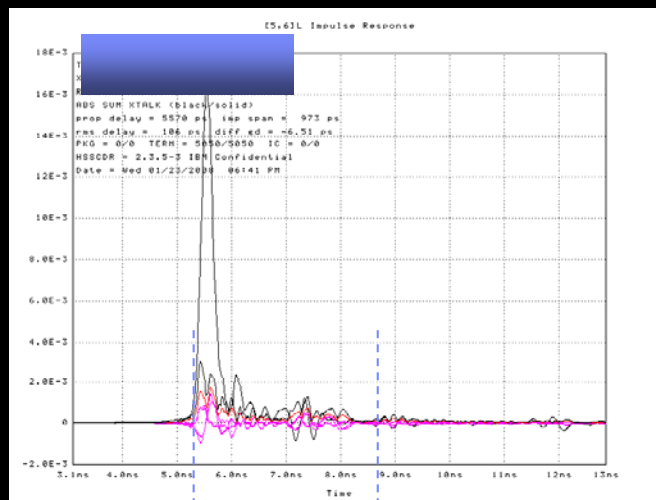
- Demonstrate transmitter 50 Ohm match of better than -15 dB up to 10 GHz for a source series terminated transmitter (SST) including ESD diode
- Demonstrate >1V output voltages for a SST transmitter
- Both of the above is key to enable industry standard compliant links



## Design & Measurement Results

- Output amplitude verified to be >1V up to 9 Gbps
- Transmitter 50 Ohm match is greatly improved by using an *asymmetric* T-coil in the output stage
- Represents first hardware demonstration of T-coil enhanced SST transmitter
- Measured output reflections are smaller than -17dB to above 10 GHz
- T-coil based Sxx improvement as large as 9 dB at 4.25 GHz (baud rate freq. of target data rate), ~20 dB at 8.5 GHz

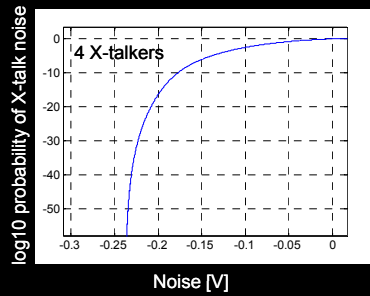
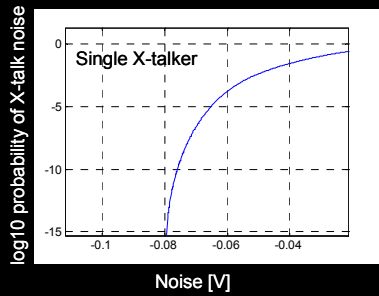
# Impulse response @4.8Gbps + Xtalk



X-talk main region ≈ 15 UI !!

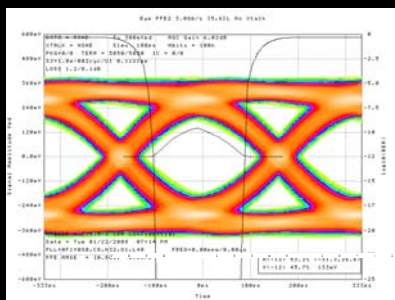
## Cumulative X-talk AM noise distribution function

Obtained by n-times folding of n tap long x-talk function

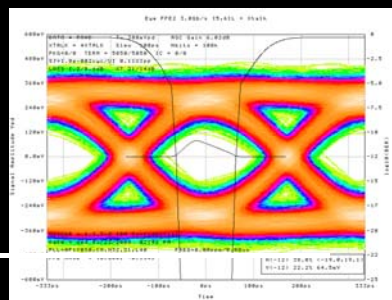


Can be used to estimate X-talk amplitude noise level as function of probability.  
**Important note:** For higher number of X-talkers and longer the impulse response of each X-talker, the log10 probability goes to larger values !!  
**Example from 4 X-talkers:** WC amplitude difference between 1e-20 and 1e-60 is 20mV, resulting in a 20mV eye closure (1e-10 to 1e-20 results 40mV eye closure).

## SE Eye diagram at 3.0 Gbps with 2 FFE<sup>(1)</sup> taps



Without X-talk  
 $V_{\text{eye}} = 133\text{mV}^{(2)}$

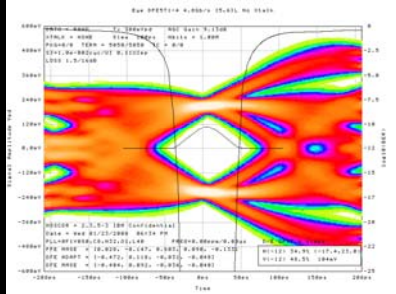


With X-talk (4 aggressors)  
 $V_{\text{eye}} = 65\text{mV}^{(2)}$

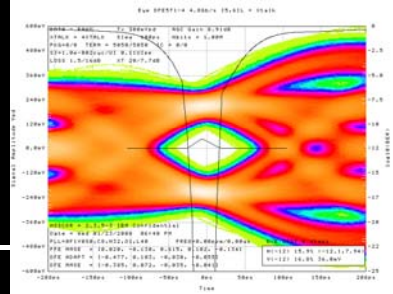
→ Data transfer feasible (as proven in P6)

<sup>(1)</sup> 2 FFE tap = cursor + post-cursor  
<sup>(2)</sup> for  $P_{\text{err}} = 10^{-12}$

## SE Eye diagram at 4.8 Gbps 5 FFE<sup>(1)</sup> + 4 DFE taps



Without X-talk  
 $V_{eye} = 104\text{mV}$

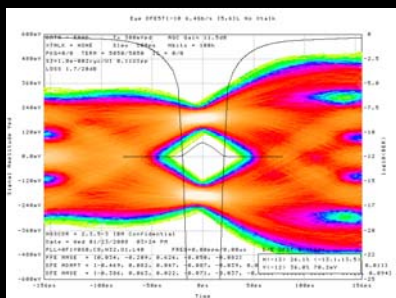


With X-talk (4 aggressors)  
 $V_{eye} = 36\text{ mV}$

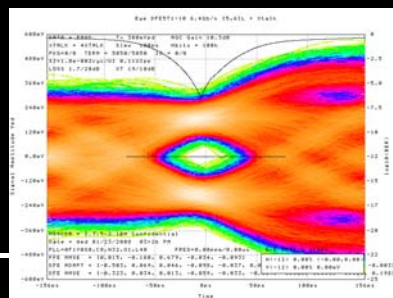
➔ Data transfer feasible (with some equalization effort)

<sup>(1)</sup> 5 FFE taps = 2 precursor+ cursor + 2 post-cursor

## SE Eye diagram at 6.4 Gbps 5 FFE<sup>(1)</sup> + 10 DFE taps



Without X-talk  
 $V_{eye} = 64\text{mV}$

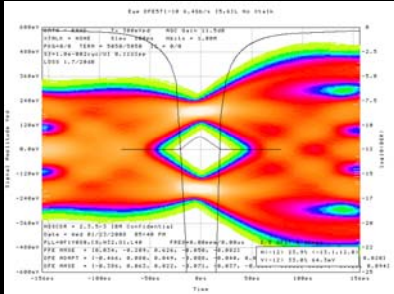


With X-talk (4 aggressors)  
 $V_{eye} = 0\text{ mV}$

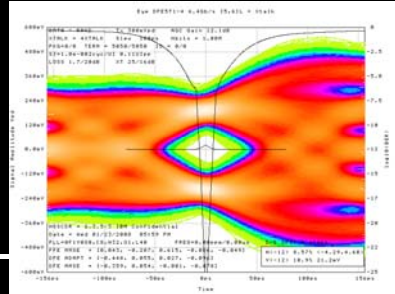
➔ Data transfer NOT feasible

<sup>(1)</sup> 5 FFE taps = 2 precursor+ cursor + 2 post-cursor

## SE Eye diagram at 6.4 Gbps 5 FFE<sup>(1)</sup> + 4 DFE taps



Without X-talk  
 $V_{eye} = 64\text{mV}$

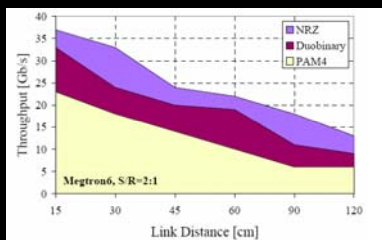
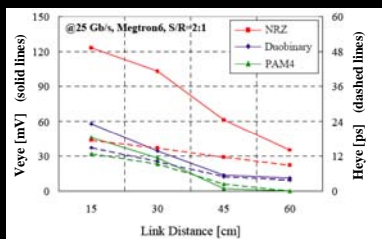


With 1/2 X-talk (all 4 aggressors attenuated by 6dB)  
 $V_{eye} = 21\text{mV}$

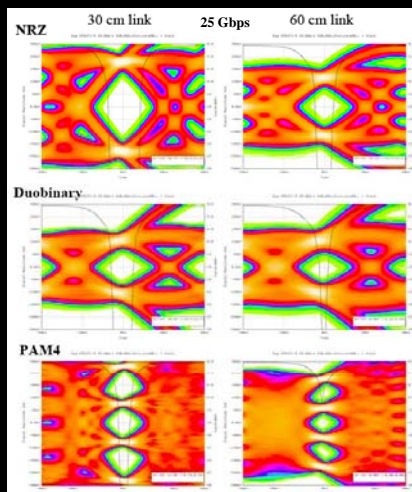
➔ Data transfer feasible IF X-talk is suppressed by 6dB

<sup>(1)</sup> 5 FFE taps = 2 precursors+ cursor + 2 post-cursor

## Modulation Format Comparison (incl Equazlization)

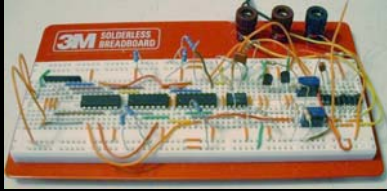


Curtsey, M. Bites, Dong G. Kam et al. 'Viability of 25Gbps Signaling', 2008 Electronic Components and Technology Conference

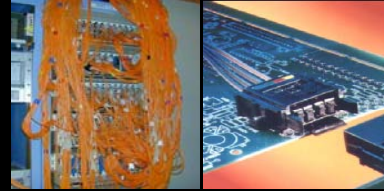


Note: A 4-tap symbol-spaced FFE with one pre-cursor and two post-cursors, and a 5-tap half-rate DFE were assumed for all three signaling options, the launch was 800 mVpp differential, and the bit stream was a 215-1 pseudo random binary sequence (PRBS) was evaluated at BER of 1e-15 when assuming requiring 30 mV vertical eye opening and 0.3 unit interval (UI) horizontal eye opening (12 ps).

# “Replace Channel” : Optical extension of el. links



1960ties



20xx



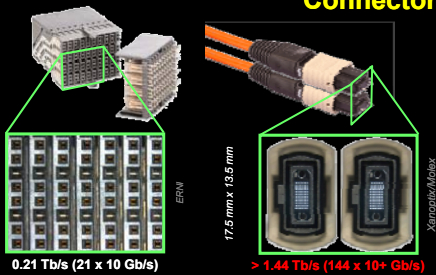
**Electronics:** Cable → Printed Circuit



**Optics:** Fiber → Integrated Waveguides

# Why optical communication?

## Connectors



## Cables



## Electrical I/O



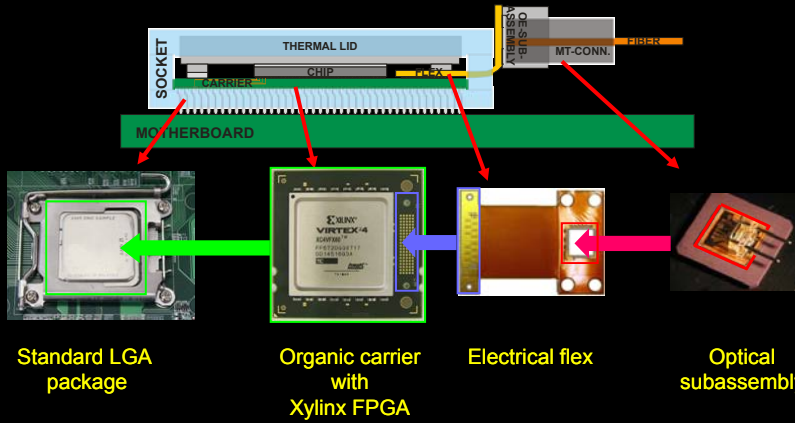
## Optical I/O



### Optical Communication

- Higher density
- More bandwidth\*length
- Lower power
- Scalability well beyond 10 Gbit/s

# Optics to the carrier – ZRL Implementation



- Mostly standard processes and designs were applied

# Optical characterization

- OE-module driven by FPGA
  - Successful operation at 10 Gbps
  - Pre-emphasis tests on link through socket and long cables

Pre-test with OE-module on separate testboard

Without (top) and with pre-emphasis at 8 Gbps; blue = FPGA out / Tx in; red/yellow = Tx out, scope optical in

OE-module driven by FPGA at 10 Gb/s (no pre-emphasis)

## Use of optics in future blades



## Part 4: Conclusions

## Conclusions

- Multi-core uP's are driving need for HUGHE aggregate data rates  
→ Power & area of I/O's at given BER level are key metrics.
- Complexity of I/O links is increasing dramatically  
→ Channel equalization is THE driving factor; high frequency does not make things easier.
- CMOS can do amazing things  
→ So far, electrical links in CMOS were able to do the job.
- Optics is coming  
→ System level interconnects now, chip-to-chip soon, on-chip far away.
- Last-not-least  
→ Might be that the I/O performance is gating how much functionality will be implemented on a single chip.  
→ We need more than one, but a limited number of different links (\$\$\$)

Thank you for your attention.

