Energy Efficient Multi-Gb/s I/O: Circuit and System Design Techniques

April 22, 2011
WMED-2011

Bryan Casper, Intel Circuit Research Labs
Agenda

Introduction

• Impact of process scaling
• Active power optimization
  – System
  – Circuit
• Power management
• Low power silver bullets
• Putting it all together
High-End Server

Assumptions
CPU TDP = 135W
I/O eff = 10pJ/bit 1-side

Future* BW Breakdown

2010 estimates based on Intel® Xeon® Processor X7560
*2015-2020 BW need estimates are solely the opinion of the author and do not necessarily represent the position of Intel Corp.
Bryan Casper – Low Power I/O
Power efficiency improving

• Driven by circuit, channel and process improvements
• ...but not keeping pace with aggregate BW needs
  – e.g. 1TB/s x 10pJ/bit = 80W!

*Adapted from Poulton, et. al., JSSC Dec. 2007
Impact of 1TB/s CPU*

~$\frac{1}{2}$ CPU Power Budget

$800$ Electricity

For the environmentally minded: $8000$kg of $\text{CO}_2$

*Assuming: 1TB @ 10pJ/bit, 5 year lifetime, 10¢/kWh, 50% conversion loss, fossil fuel generated electricity
I/O Energy Efficiency Definition

- mW/Gb/s = pJ/bit
- Total physical layer energy required to move data
  - Includes amortized global power as well
- Usually 2-sided metric (TX + RX)
Agenda

• Introduction

Impact of process scaling

• Active power optimization
  – System
  – Circuit

• Power management

• Low power silver bullets

• Putting it all together
Past technology trends scaled efficiency proportional to feature size

\[ y = 399.17x^{1.157} \]
Process vs. Logic Scaling Scenarios*

*ITRS-like trends assuming bulk planar CMOS. Conceptual scenarios with large error bar.
Example Research I/O Energy Scaling

Process scaling estimates vs. circuit type
- Logic → 0.75x 37
- Noise limited → 0.95x 37.3
- Sense amp → 0.85x 19.2
- Swing limited → 1x 6.5

Aggregate I/O scaling factor per generation
~0.9x

Variation compensation overhead could cause factor to be >0.9x

*Unpublished Research I/O technology based on 32nm
Trends in I/O Power vs. Year*

Energy Efficiency ($\text{pJ}/\text{bit}$)

-20%/year

Architectural enhancements much more effective than process scaling

Trend may change if depend solely on process
Agenda

• Introduction
• Impact of process scaling
• Active power optimization
  – System
  – Circuit
• Power management
• Low power silver bullets
• Putting it all together
Optimal Energy-Performance Design Space

Adapted from Mark Horowitz IEDM 2005
Adder Design Space

Adapted from Mark Horowitz IEDM 2005
Adder Design Space

I/O tradeoff is even more nonlinear due to channel rolloff

Adapted from Mark Horowitz IEDM 2005
I/O Design Space Tradeoffs

Steep tradeoff caused by:
1. Channel BW limit
2. Process BW limit
3. Circuit architecture complexity

Key to low power links is operating on this portion of design space.
## Power’s Deadly Combination

### Stingy System Architect
- Not willing to limit legacy channel length or topologies
- Doesn’t want to erode profit margins by adopting higher cost interconnect
- Perceives alternate topologies as unproven & risky
- Annoyed that Moore’s law doesn’t apply to channels

### Macho Link Designer
- Knows Shannon’s Capacity
- Takes on challenge to apply advanced communication techniques to high-speed links
  - e.g. DSL, Ethernet
- Thinks Moore’s law will eventually resolve power & complexity issues
Energy Efficiency Correlation to Loss

Energy Efficiency (pJ/bit) vs. Channel Loss @ Symbol rate (dB)
Legacy Backplane Channel

Loss

0dB
-30dB
-60dB

0GHz 5GHz 10GHz 15GHz
Backplane Data Rates

<table>
<thead>
<tr>
<th>TX FIR taps</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>DFE taps</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
<td>64</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Increasing Equalizer Complexity (nonlinear scale)
Overextending channel leads to nonlinear EQ power vs. performance tradeoff

Channel Power Wall

Energy Efficiency (pJ/bit)

Max Rate (Gb/s)

Increasing equalization complexity

5pJ/bit baseline, 1/2pJ/bit/DFE tap, 1/4pJ/bit/TX tap
I/O Challenges: Power

Insertion Loss

-40dB
-20dB
0dB
0GHz 5GHz 10GHz 15GHz

• Legacy backplane w/ 2 connectors & sockets, 1/2m FR4

Equalizer Power (pJ/bit or mW/Gb/s)

Power assumptions in slide backup

Performance (Gb/s)
I/O Challenges: Power

- Legacy backplane w/ 2 connectors & sockets, ½m FR4
- ½m cabled topology with top-pkg connectors

Insertion Loss

Equalizer Power (pJ/bit or mW/Gb/s)

Performance (Gb/s)
Loss/rate/power estimates

Assuming:
- Xtalk not primary limiter
- TX+RX jitter ~1/2UI
- RX noise 1mV$_{\text{rms}}$
- TX swing ~1V$_{\text{diffp-p}}$
- Cabled link w/ connectors
  - Channel “well behaved”

<table>
<thead>
<tr>
<th>Equalization Complexity</th>
<th>Est. data rate</th>
<th>Normalized power (rough guess)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>~0.8*X Gb/s</td>
<td>1</td>
</tr>
<tr>
<td>Low power</td>
<td>~2.0*X Gb/s</td>
<td>~1</td>
</tr>
<tr>
<td>Moderate (3 tap LE, 4 tap DFE)</td>
<td>~2.4*X Gb/s</td>
<td>~2</td>
</tr>
<tr>
<td>Complex (&gt;50 tap LE+DFE)</td>
<td>~3.6*X Gb/s</td>
<td>~10-100</td>
</tr>
<tr>
<td>Complex EQ+PAM+FEC/coding</td>
<td>~4.4*X Gb/s</td>
<td>~100-1000</td>
</tr>
<tr>
<td>Shannon’s capacity</td>
<td>~8-10*X Gb/s</td>
<td>n/a</td>
</tr>
</tbody>
</table>
Common Traits of Low Power Links

- Simple equalization
- Low TX swing
- Sensitive RX sampler
- Low-power clocking
How to scale rate or distance and maintain energy efficiency

• Path to scaling performance: Refined channels
  - e.g. Top-package connector based cabled links

½ Meter Channel Examples (based on Intel Labs Measurements)

• PCIe (2 connector) $\rightarrow$ 20dB @4GHz
• LCP Flex* $\rightarrow$ 20dB @15GHz
• Twinax 36 AWG* $\rightarrow$ 20dB @30GHz

*No connector, pkg or pad cap
Agenda

• Introduction
• Impact of process scaling
• Active power optimization
  – System
  – Circuit
• Power management
• Low power silver bullets
• Putting it all together
Low-power Link Circuits Top Ten

• Not a comprehensive list
  – More like a sampling of known power reduction methods
• Few low power links incorporate all of these techniques
  – Most incorporate at least some
• Not intended to be a detailed overview of each method

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Top Ten #1: Modest data rates

Steep tradeoff caused by:
1. Channel BW limit
2. Process BW limit
3. Circuit architecture complexity

Key to low power links is operating on this portion of design space.
Clock Buffer Power/Performance Example

Normalized Energy Efficiency (energy/bit)

Operating Frequency (GHz)

Stay off process BW cliff

Fanout set to meet per stage BW requirement

Fanout = 16
Performance Impact on Circuit Architecture: Loop-unrolled DFE

Conventional DFE
- Speedpath limits frequency

Loop-unrolled DFE
- Redundancy to alleviate speedpath
- Increases power and complexity
  - Proportional to $C_1 \times 2^N + C_2$
    - $C_1$ = comparator + mux
    - $C_2$ = baseline power
    - $N$ = number taps unrolled
Performance Impact on Circuit Architecture: Multi-phase Clocking

• Interleaving of receiver alleviates need for high-frequency latches and clocks

• Requires greater clock complexity and calibration
  – Multiphase clock generators
  – Sophisticated phase training

Half rate clock

Quarter rate clock

DCC

RX

Multi-phase clk generator

RX

RX

RX

Phase Calibration

RX

RX

RX

DCC

RX

Bryan Casper – Low Power I/O
Top Ten #2: Forwarded clocking

Forwarded clock power benefits
- No need for high clock recovery BW
- Edge/test samplers optional
  - Clock recovery/test can be time multiplexed with data samplers
- Fewer, simpler phase rotators
  - Greater tolerance for INL & jitter
  - 1 rotator can cover data, edge & test in a time-multiplexed fashion
**Top Ten #3: Global Circuit Sharing**

- Parallel link implementations have ample opportunity to share common functionality

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling

Potential to be shared across parallel link
Co-optimization of Channel and Circuits Enables Widespread Power Amortization

- Matched interconnect enables clock recovery sharing
  - Common deskew across 10 bits
- Test, bias, etc. are shared as well

F. O'Mahony, et. al., "A 47×10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS," ISSCC 2010
Top Ten #4: Low power clock distribution

- Reduce distribution distance*
  - Compact parallel link floorplan
- Repeaterless distribution**

*F. O'Mahony, et. al., "A 47×10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS," ISSCC 2010

Forwarded Clock Repeater-less Distribution

Repeater-less distribution + forwarded clock combination has potential to eliminate buffers and save power


On-chip T-lines

Forwarded Clock TX

Off-chip interconnect

cmbias
Top Ten #5: Resonantly tuned clocking

- Resonant clocking suppresses jitter outside the fundamental clock frequency
- Lower power for a given jitter budget
- Limits clock frequency operating points
- Frequently used for resonators
  - LC-VCO
- Also used for distribution

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Resonant Clocking Example

Enabled 3x-5x lower clocking power than conventional distribution

Top Ten #6,7: Co-designed TX & RX

- TX output stage & RX input dissipate a large portion of link power
- Co-optimize to minimize power and meet BER requirements

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Swing vs. RX Sensitivity

Assumptions:
- RX noise variance proportional to RX power
  - 5mW $\rightarrow$ 1mVrms
- Normally distributed ISI
  sigma = 0.001*swing
- 1E-12 BER target
- Voltage-mode TX w/ linear reg.
- Channel loss = -20dB
Simplistic Example: Swing vs. Efficiency

- Optimal energy at ~160mVpp
  - Requires ~1mV_{rms} input referred RX noise
Low Swing Tradeoffs: 19” Cabled Link Maximum Rates

- Lowest power equalization points hardly suffer due to low swings.
Top Ten #6: Low-Swing TX

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Low-Power TX Drivers: CM vs. VM

\[ V_{d,1} = \frac{I}{2}R \]
\[ V_{d,0} = -\frac{I}{2}R \]
\[ V_{d,pp} = IR \]
\[ I = \frac{V_{d,pp}}{R} \]

Source: Ganesh Balamurugan
Low-Power TX Drivers: CM vs. VM

Current Mode (CM)
Single-ended Term

Voltage Mode (VM)
Single-ended Term

\[ V_{d,1} = \frac{I}{2}R \]
\[ V_{d,0} = -\frac{(I/2)R}{2} \]
\[ V_{d,pp} = IR \]
\[ I = \frac{V_{d,pp}}{R} \]

\[ V_{d,1} = \frac{V_s}{2} \]
\[ V_{d,0} = -\frac{V_s}{2} \]
\[ V_{d,pp} = V_s \]
\[ I = \frac{V_s}{2R} \]

2X power reduction

Source: Ganesh Balamurugan
Low-Power TX Drivers: CM vs. VM

Current Mode (CM)  
Single-ended Term

Voltage Mode (VM)  
Differential Term

\[ V_{d,1} = \frac{I}{2}R \]
\[ V_{d,0} = -\frac{I}{2}R \]
\[ V_{d,pp} = IR \]

\[ I = \frac{V_{d,pp}}{R} \]

\[ V_{d,1} = \frac{V_s}{2} \]
\[ V_{d,0} = -\frac{V_s}{2} \]
\[ V_{d,pp} = V_s \]
\[ I = \frac{V_s}{4R} \]

4X power reduction

Source: Ganesh Balamurugan
# Low-Swing TX Drivers: CM vs. VM

## Table Comparison

<table>
<thead>
<tr>
<th></th>
<th>VM (Palmer, JSSC 12/2007)</th>
<th>CM (O’Mahony, JSSC 12/2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vswing</td>
<td>210mVpp-diff</td>
<td>150mVpp-diff</td>
</tr>
<tr>
<td>Proc. / Vcc</td>
<td>90nm / 1.0Vcc</td>
<td>45nm / 0.8Vcc</td>
</tr>
<tr>
<td>Eq.</td>
<td>No</td>
<td>Yes (2-tap)</td>
</tr>
<tr>
<td>Datarate</td>
<td>6.25Gb/s</td>
<td>10Gb/s</td>
</tr>
<tr>
<td>Bias cap</td>
<td>36pF</td>
<td>&lt;1pF</td>
</tr>
<tr>
<td>TX driv power</td>
<td>1.10mW</td>
<td>2.12mW</td>
</tr>
<tr>
<td>TX bias power</td>
<td>0.76mW</td>
<td>0.34mW</td>
</tr>
<tr>
<td>Total TX driv. power</td>
<td>1.86mW</td>
<td>2.46mW</td>
</tr>
</tbody>
</table>

### VM power savings reduces for low-swing TX
Top Ten #7: Sensitive RX

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Low-Power RX samplers

Good receiver sensitivity allows low TX swing

- Residual input-referred offset: <2mV
- Input-referred noise: 1mV-rms
- Hysteresis + metastability: <2mV

O’Mahony, JSSC Dec. 2010
Sensitive RX samplers

offset[5:0] → IDAC → VOA → Latch → Latch → RSFF → dout

din → IDAC offset[5:0] → iref

10:3 in → clk → 3:10 in → out → out
Sensitive RX samplers

offset[5:0] → IDAC

VOA

Latch → Latch → RSFF → dout

din

clk

in

out

in

out

out

in
Top Ten #8: Simple Equalization

- Linear equalizers - big bang for the buck
  - If channel is “well behaved” and ISI dominated
- DFE is complex, especially if speedpaths
  - 1-tap DFE only cancels 1 postcursor point

<table>
<thead>
<tr>
<th>1. Modest data rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Forwarded clocking</td>
</tr>
<tr>
<td>3. Global circuit sharing</td>
</tr>
<tr>
<td>4. Low power clock distribution</td>
</tr>
<tr>
<td>5. Resonantly tuned clocking</td>
</tr>
<tr>
<td>6. Low swing TX</td>
</tr>
<tr>
<td>7. Sensitive RX</td>
</tr>
<tr>
<td>8. Simple equalization</td>
</tr>
<tr>
<td>9. Calibration and tuning</td>
</tr>
<tr>
<td>10. System modeling</td>
</tr>
</tbody>
</table>

![Diagram of unequalized pulse response with linear equalizer (LE) and decision feedback equalizer (DFE)]
Examples: Low Power Linear Equalizers

- Continuous time linear equalizers
  - Passive using HP filters or inductive peaking
  - Source degeneration

- Pre-emphasis
  - Limit magnitude & sign of taps
  - Current summing in analog domain
Top Ten #9: Calibration and tuning

- Process scaling may reduce power
  - By scaling both C and V scaling
  - Increases variation due to smaller device area
- Increased logic resources enables sophisticated calibration logic to compensate variation

\[ \sigma V_T = \frac{1}{\sqrt{2}} \left( \frac{c_2}{\sqrt{W_{eff} \cdot L_{eff}}} \right) \]

K. Kuhn, IEDM 2007

Bryan Casper – Low Power I/O
Extrapolated Process Variation

Area & Energy scaling limited by variation

\[ \sigma(V_t) = C_2 \div \sqrt{A_{gate}} \]

<table>
<thead>
<tr>
<th>Year</th>
<th>C2_hi</th>
<th>C2_lo</th>
<th>std(Vt_hi)</th>
<th>std(Vt_lo)</th>
</tr>
</thead>
<tbody>
<tr>
<td>130nm y2002</td>
<td>5.9E-09</td>
<td>5.9E-09</td>
<td>1.4E-02</td>
<td>1.4E-02</td>
</tr>
<tr>
<td>65nm y2006</td>
<td>5.6E-09</td>
<td>4.9E-09</td>
<td>2.6E-02</td>
<td>2.3E-02</td>
</tr>
<tr>
<td>32nm y2010</td>
<td>5.4E-09</td>
<td>4.1E-09</td>
<td>9.7E-02</td>
<td>6.5E-02</td>
</tr>
<tr>
<td>16nm y2014</td>
<td>5.1E-09</td>
<td>3.4E-09</td>
<td>1.8E-01</td>
<td>1.1E-01</td>
</tr>
<tr>
<td>8nm y2020</td>
<td>4.9E-09</td>
<td>2.8E-09</td>
<td>3.5E-01</td>
<td>1.8E-01</td>
</tr>
<tr>
<td>4nm y2026</td>
<td>4.6E-09</td>
<td>2.4E-09</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Most calibration and adaptation used today is fairly basic
• e.g. duty cycle correction
Example: Programmable Phase Rotator

- Power can scale as process variation increases
- Alternative is to not scale device area and hence no power scaling

F. O’Mahony, et. al., A Programmable Phase Rotator based on Time-Modulated Injection Locking
Top Ten #10: System Modeling

- Key to low power is balanced implementation
  - Achieved through comprehensive understanding of power/performance tradeoffs
- Focus design effort and power on highest impact components
- System-level optimization most impactful
  - Most will not have this opportunity due to standardization specs.
  - Sub-system optimization still useful

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Methodology example: Sensitivity Calculation to Optimize Power

1. Calculate 1st order power sensitivity of each design option
2. Calculate 1st order margin sensitivity of each design option
3. Form mathematical relationship between power and performance
4. Minimize power for performance target using optimization algorithm
5. Repeat steps 1-4 to further refine design point
Methodology example: Sensitivity Calculation to Optimize Power

<table>
<thead>
<tr>
<th>Parameter change</th>
<th>Baseline change</th>
<th>Eye width change estimate vs. baseline (units = 1ps or 0.01UI)</th>
<th>Power delta estimate vs. baseline (mW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline eye width</td>
<td></td>
<td>18</td>
<td>100</td>
</tr>
<tr>
<td>TX ref. ck. jitter (pp)</td>
<td>50ps→60ps</td>
<td>−4</td>
<td>+0</td>
</tr>
<tr>
<td>TX PLL 1-UI jitter, rms (Gaussian jitter, accumulated)</td>
<td>0.5ps→0.75ps</td>
<td>−12</td>
<td>−3</td>
</tr>
<tr>
<td>TX equalizer</td>
<td>2 taps→3 taps</td>
<td>−2</td>
<td>+3</td>
</tr>
<tr>
<td>TX swing</td>
<td>100mV→200mV</td>
<td>+3</td>
<td>+4</td>
</tr>
<tr>
<td>TX buffer sinusoidal jitter @ 200MHz</td>
<td>±15ps→±18ps</td>
<td>−10</td>
<td>−1</td>
</tr>
<tr>
<td>TX buffer duty cycle error</td>
<td>1%→2%</td>
<td>−1</td>
<td>−0.1</td>
</tr>
<tr>
<td>RX PLL 1-UI jitter, rms (Gaussian jitter, accumulated)</td>
<td>0.5ps→0.75ps</td>
<td>+0</td>
<td>−3</td>
</tr>
<tr>
<td>RX PLL bandwidth</td>
<td>4MHz→6MHz</td>
<td>−7</td>
<td>+0</td>
</tr>
<tr>
<td>CDR loop latency</td>
<td>2UI→4UI</td>
<td>−2</td>
<td>−1</td>
</tr>
<tr>
<td>RX input noise</td>
<td>1mVrms→2mVrms</td>
<td>−2</td>
<td>+2</td>
</tr>
<tr>
<td>PI phase accuracy</td>
<td>0.015UI→0.03UI</td>
<td>−1</td>
<td>−3</td>
</tr>
</tbody>
</table>

Knowledge of system performance and power sensitivities enables global power optimization.

Agenda

• Introduction

• Impact of process scaling

• Active power optimization
  – System
  – Circuit

Power management

• Low power silver bullets

• Putting it all together
Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period.
Energy-Disproportionate Link

Barroso & Holzle, IEEE Computer, Dec 2007
Energy-proportional I/O

Normalized Bandwidth Demand

Conventional (fixed Bandwidth)

Wasted Energy

Time
Power Management: Scalable supplies

Power efficiency improves with adaptive supply/biasing

Refs: B. Casper, ISSCC ’06 & G. Balamurugan, JSSC 4/08
65nm Low Power Link Operating Points

Balamurugan JSSC, April 2008
Benefit of Eliminating Excess BW: Non-linear Efficiency/Performance

I/O Energy Efficiency (pJ/bit) vs. Data Rate (Gb/s)

18” FR4 Backplane:
- 3.6 at 5 Gb/s
- 5.0 at 10 Gb/s
- 6.5 at 15 Gb/s

8” FR4 Backplane:
- 2.7 at 5 Gb/s
- 3.6 at 10 Gb/s
- 5.0 at 15 Gb/s

Balamurugan JSSC, April 2008
Fast Wake-Up Clocking

Agenda

• Introduction

• Impact of process scaling

• Active power optimization
  – System
  – Circuit

• Power management

Low power silver bullets (?)

• Putting it all together
Low Power Link “Silver Bullets”? 

Optical 

Modulation (PAM) 

3D Stacking
Silver Bullet?: Optical

• **Claim:**
  - Optical power is inherently better because no equalization needed

• **Reality:**
  - Most optical power claims only include optical components
    - Disregards electrical driver, clocking, recovery, synch., serdes, etc.
  - Optical link = Electrical link + optical components in the middle
  - Optical/electrical power crossover is likely 1m-5m, depending on rate

---

**Distance**

**Power**

Optical

Electrical

Stay on this portion of the curve and equalization power likely to be small subset of overall link power (e.g. <5% [O’Mahony, ISSCC2010])
Apples-Apples Electrical/Optical Channel Comparison

- Electrical

- Optical
Complete Optical CPU Link

- Modulator/VCSEL
- Connectors
- Jumpers
- Optical cable

- Modulator Driver
- Photodetector (PD)
- Transimpedance Amp/Limiting Amp (TIA/LA)

- CPU w/ electrical TX
- Package/module

- CPU w/ electrical RX
• Optical channel loss is frequency independent
  – But the aggregate loss is 100x-1000x!

• VCSEL or MZI based links require large swings (500mV-1V)

• Worst-case received signal can be as low as \(~\)10uA
  – Requires extremely sensitive receiver (costs power)

• Full optical link >2x power of electrical at \(\leq 3\text{m distance}\)
Silver Bullet?: PAM

• Claim:
  – PAM uses less BW resulting in less equalization and lower power

• Reality:
  – No inherent performance/power advantage over binary
    • Using practical channels and 1E-12 BER
  – Equalization and clock recovery more difficult
  – PAM receiver more complex
    • 4 PAM requires 1.5 samples/bit + decoding
    • Binary requires 1 sample/bit
  – PAM may have advantages when
    • Symbol rate limited due to circuits
    • Channel has excess BW
Max data rate with $10^{-12}$ BER (LE & DFE 4-tap)

RS(64,48,8) Coding overhead estimated at 100pJ/bit in 65nm

Data Rate (GB/s)

Raw BER

2PAM w/o ECC

4PAM w/o ECC
Max data rate with small block coding to achieve $10^{-12}$ BER (LE & DFE 4-tap)

RS(64,48,8) Coding overhead estimated at 100pJ/bit in 65nm
45nm PAM Measurements (Within-package channel)

<table>
<thead>
<tr>
<th>Channel</th>
<th>Signaling Mode</th>
<th>Efficiency</th>
<th>Data rate</th>
<th>TX swing</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCP</td>
<td>2-PAM</td>
<td>2.3pJ/bit</td>
<td>12.5G</td>
<td>120mV</td>
</tr>
<tr>
<td></td>
<td>3-PAM</td>
<td>2.6pJ/bit</td>
<td>18.75G</td>
<td>260mV</td>
</tr>
<tr>
<td></td>
<td>4-PAM</td>
<td>2.6pJ/bit</td>
<td>25G</td>
<td>360mV</td>
</tr>
</tbody>
</table>

- Modulation benefits links that have channel BW much greater than circuit BW
- For this example, 2PAM power expected to be higher than 4PAM (at same data rate)
  - Due to circuit limitations

[Balamurugan, et. al., ISSCC 2010]
Silver Bullet?: 3D Stacking

• Claim:
  – Stacking minimizes need for high-speed I/O

• Reality:
  – Potential to reduce I/O power (within stack) by 10x-100x
    • Reduce C & V (CV²)
  – Components within stack must be tightly integrated (architecture, process, mechanicals)
  – Thermal and power delivery limits applicability
    • Primarily applicable for low power stacks
Example: Stacking DRAM on CPU

- Multiple DRAM stacks on CPU constrain power due to thermals
  - DRAM temp limit <100°C
  - Assumes standard CPU cooling solution

- Primarily applicable for low power CPUs
- Micro-channel cooling could change tradeoffs
Low Power Link “Silver Bullets”?  

Optical  

Modulation (PAM)  

3D Stacking  

Each technology may have power advantages for a limited set of applications. However, not a general solution to solving the Link Power Problem.
Agenda

• Introduction
• Impact of process scaling
• Active power optimization
  – System
  – Circuit
• Power management
• Low power silver bullets

Putting it all together
Example: 47x10Gb/s Interface

500µm LGA connector

HDI/Flex/cable bridge

dieA

dieB

Package Socket

PCB
Solutions: Interconnect

-20dB
0dB
-40dB
0GHz 10GHz 20GHz 30GHz 40GHz

1m twinax w/o connector
0.5m legacy backplane

|S_{21}| (loss)

32AWG stranded conductor
FEP dielectric
635um
jacket
copper foil shield

32AWG stranded conductor
FEP dielectric
635um
jacket
copper foil shield
Solutions: Connectors

Top-pkg connector (4 signals/mm²)
Solution: Circuits

• Utilized most suggested low power optimizations

1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling
Low Power Prototype Results

0.5m flex interconnect

3m twinax cable

45nm CMOS prototype demonstrates industry leading I/O power efficiency with non-traditional interconnects
Low Power Prototype Results

Link data rate = 10Gb/s

Aggressive power management:
• Idle mode is 93% less power than active state
• Wake-up from idle <5ns
Research Roadmap

Solution being researched. Demonstration anticipated by 2015.

Demonstrated

Known solution. Demonstration targeting 2012.


Data rate per pair (Gb/s)

64

32

16

8

Legacy FR4 interconnect (multiple connectors, vias and sockets)

Short traditional topologies with evolutionary optimizations

Short flex or micro-twinax cables with top-pkg connector

Active electrical cables w/ 1/2m micro-twinax
Link Active Power Optimization Key Points

• 1TB/s socket BW needed by 2020
  – Power optimize or I/O will require majority of power budget

• Don’t depend solely on process scaling to lower power
  – Architecture and circuit will drive energy scaling

• Stay away from bleeding edge
  – Channel, process and architecture

• Balanced link design is key to low power

• Optical & stacking promising but limited

• Electrical innovation in circuits and channel fruitful

Acknowledgement: Frank O’Mahony, James Jaussi, Ganesh Balamurugan, Mozhgan Mansuri, Sudip Shekhar
Related Publications 1


Related Publications 2

Related Publications 3

- F. O'Mahony, B. Casper, M. Mansuri, M. Hossain, “A Programmable Phase Rotator Based on Time-Modulated Injection-Locking,” VLSI symposium 2010