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

April 22, 2011

WMED-2011

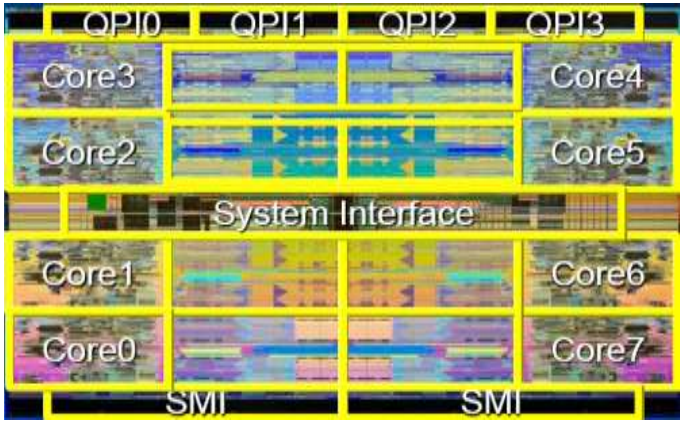
Bryan Casper, Intel Circuit Research Labs

Agenda

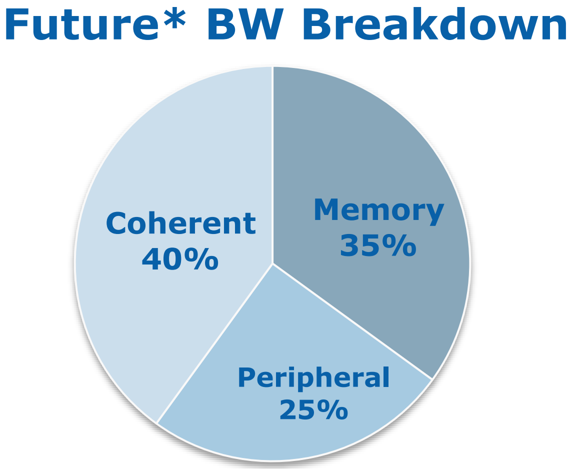
- ➡ Introduction
- Impact of process scaling
- Active power optimization
  - System
  - Circuit
- Power management
- Low power silver bullets
- Putting it all together

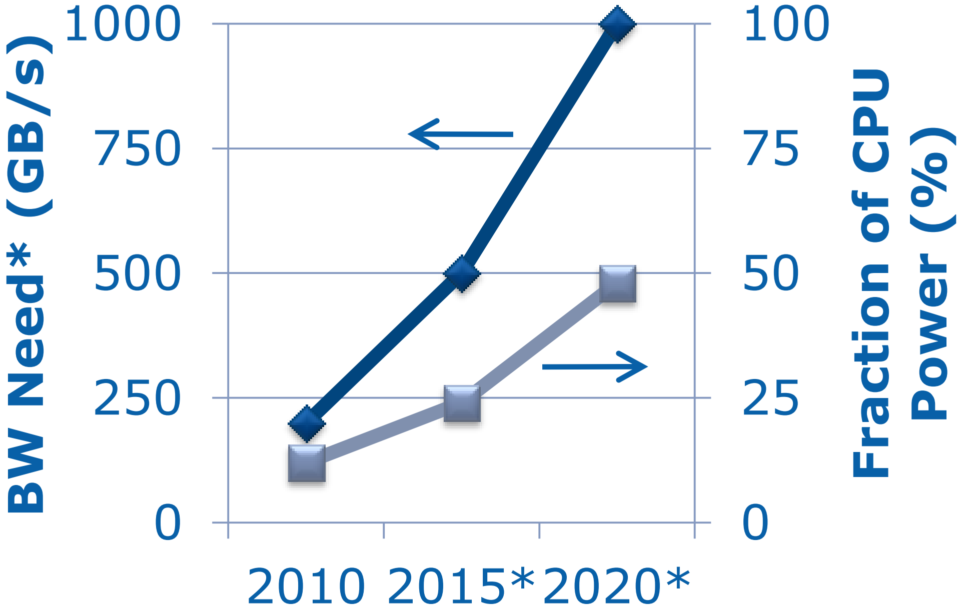
High-End Server

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



Length<20"

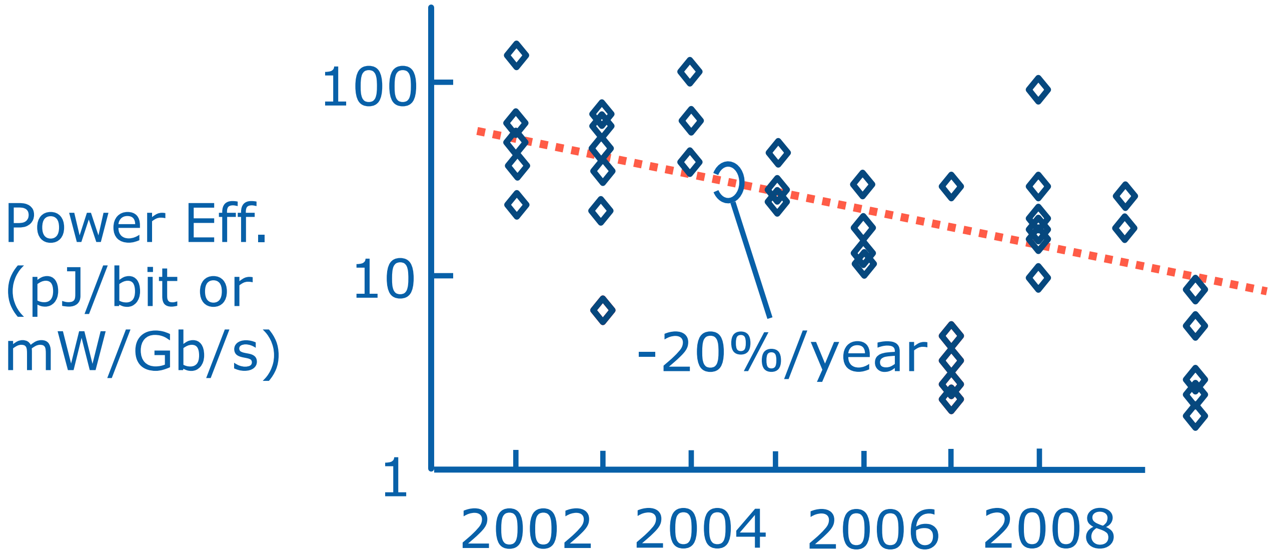




2010 estimates based on Intel® Xeon® Processor X7560

*2015-2020 BW need estimates are solely the opinion of the author and do not necessarily represent the position of Intel Corp.

Trends in I/O Power vs. Year*



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*

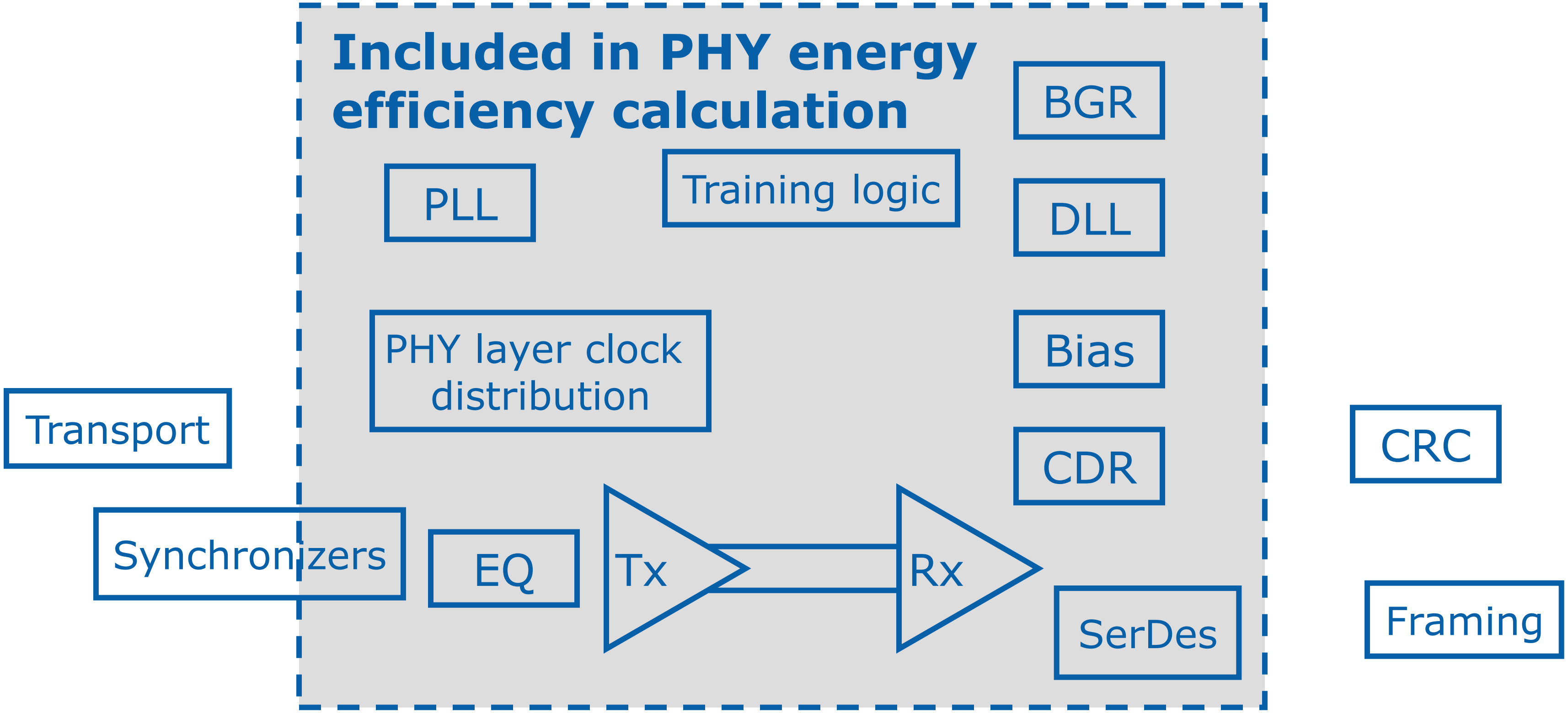
~½ CPU Power Budget

$800 Electricity

For the environmentally minded:
8000kg of CO^2

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

I/O Energy Efficiency Definition

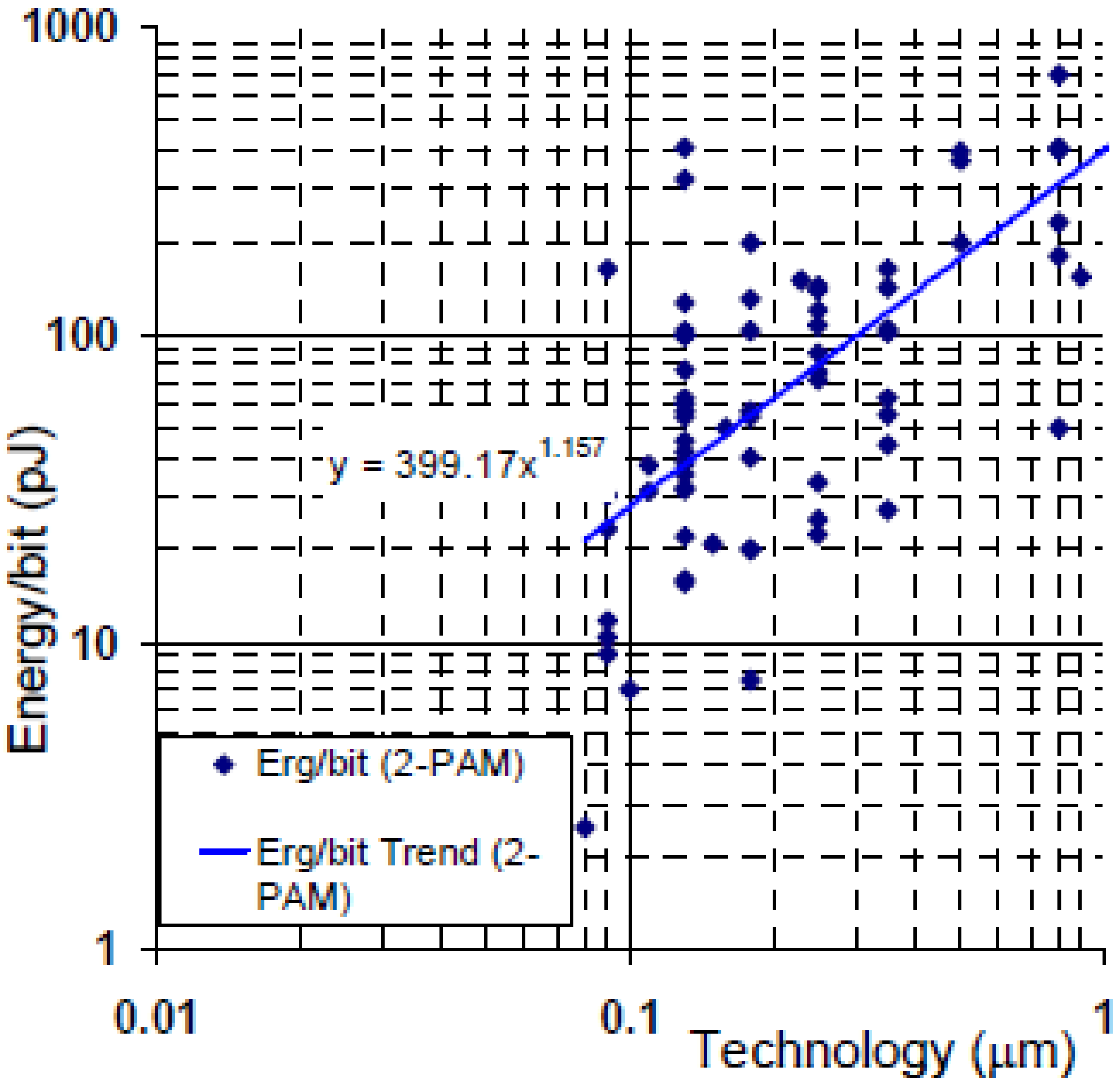


- mW/Gb/s = pJ/bit
- Total physical layer energy required to move data
  - Includes amortized global power as well
- Usually 2-sided metric (TX + RX)

Agenda

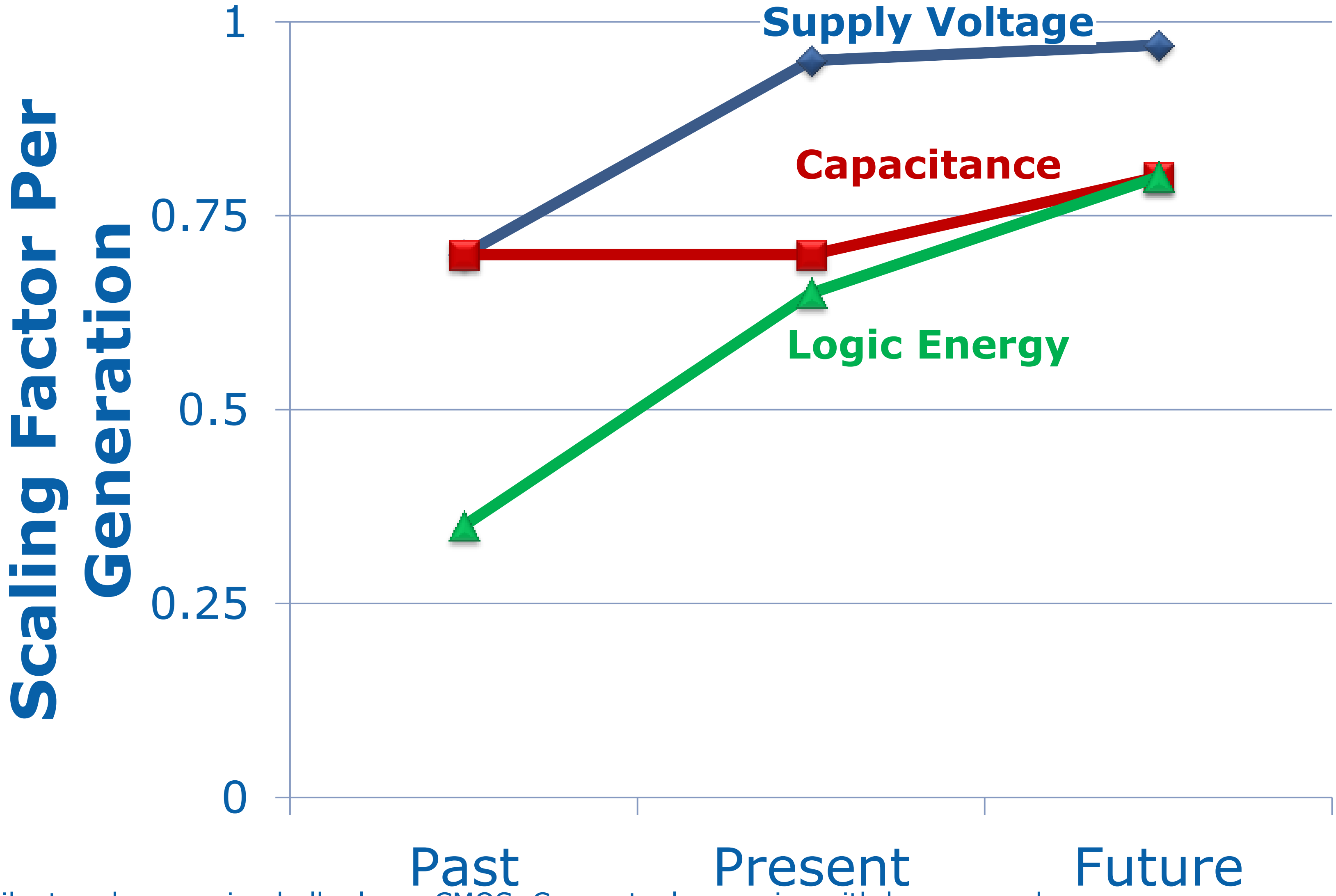
- Introduction
- Impact of process scaling
- Active power optimization
  - System
  - Circuit
- Power management
- Low power silver bullets
- Putting it all together

Past technology trends scaled efficiency proportional to feature size



Hatamkhani, et. al. DAC 2006

Process vs. Logic Scaling Scenarios*



*ITRS-like trends assuming bulk planar CMOS. Conceptual scenarios with large error bar.

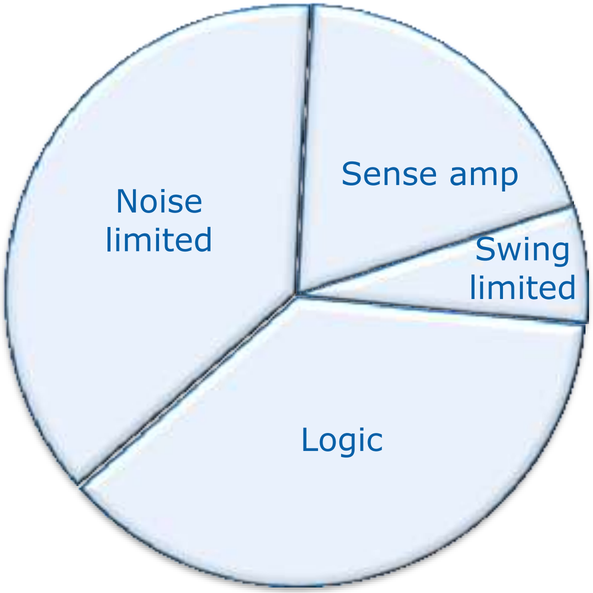
Detail

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

Research I/O* Power Breakdown



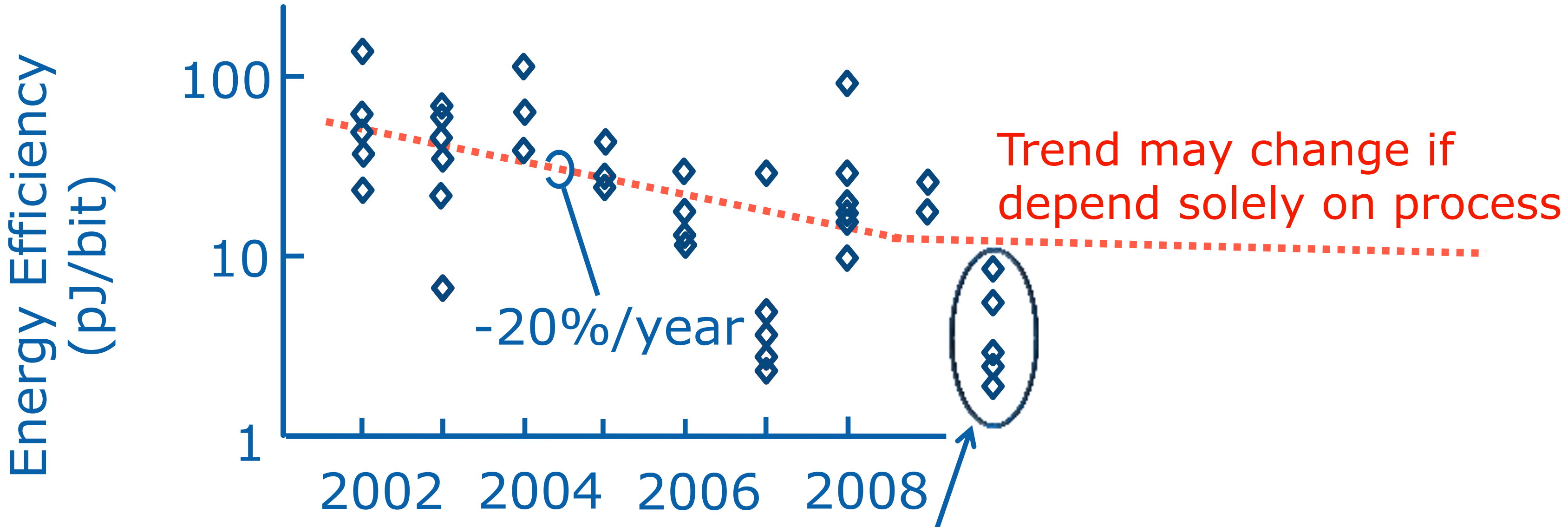
Aggregate I/O scaling factor per generation

~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*



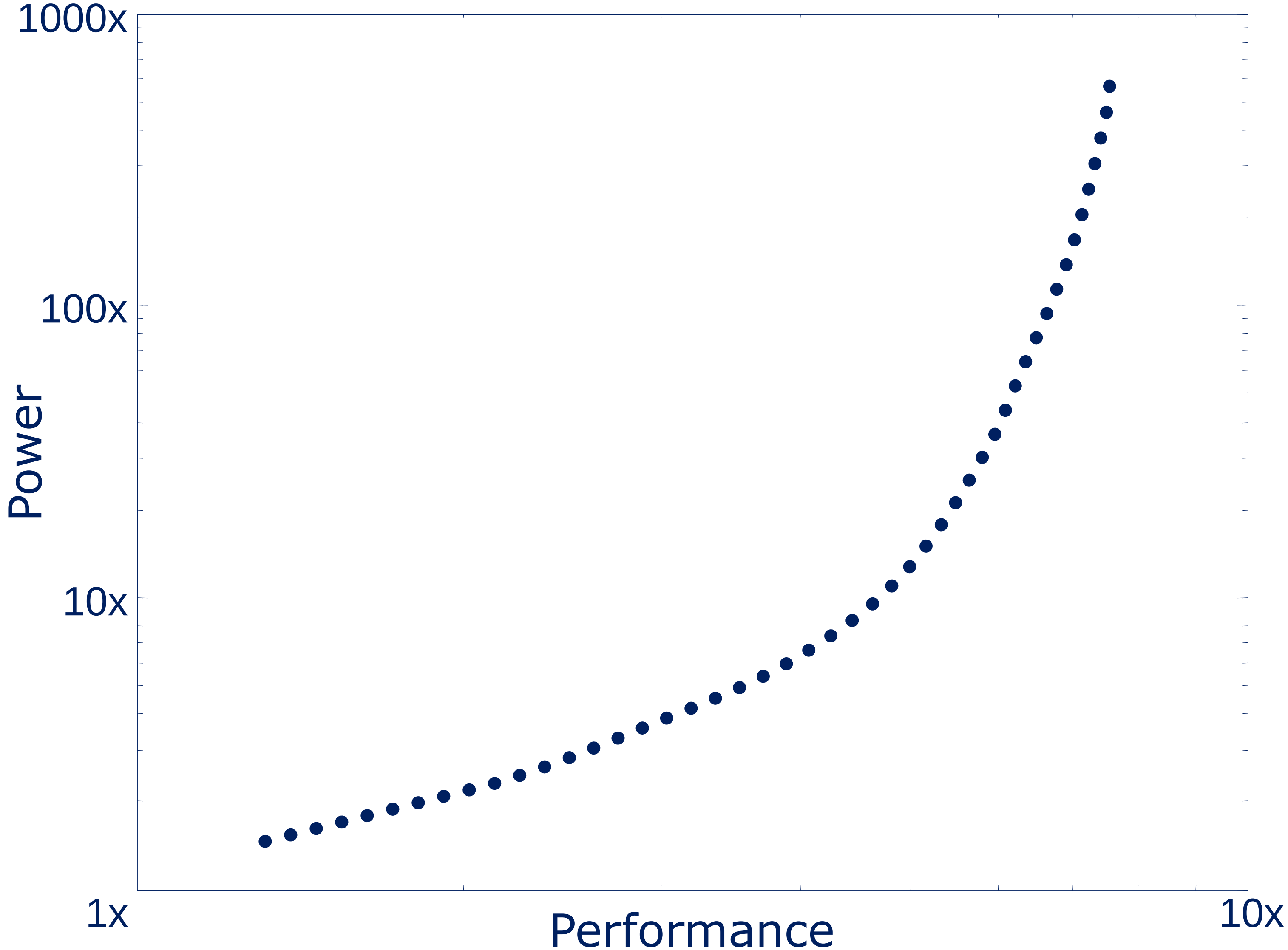
Trend may change if depend solely on process

Architectural enhancements much more effective than process scaling

Agenda

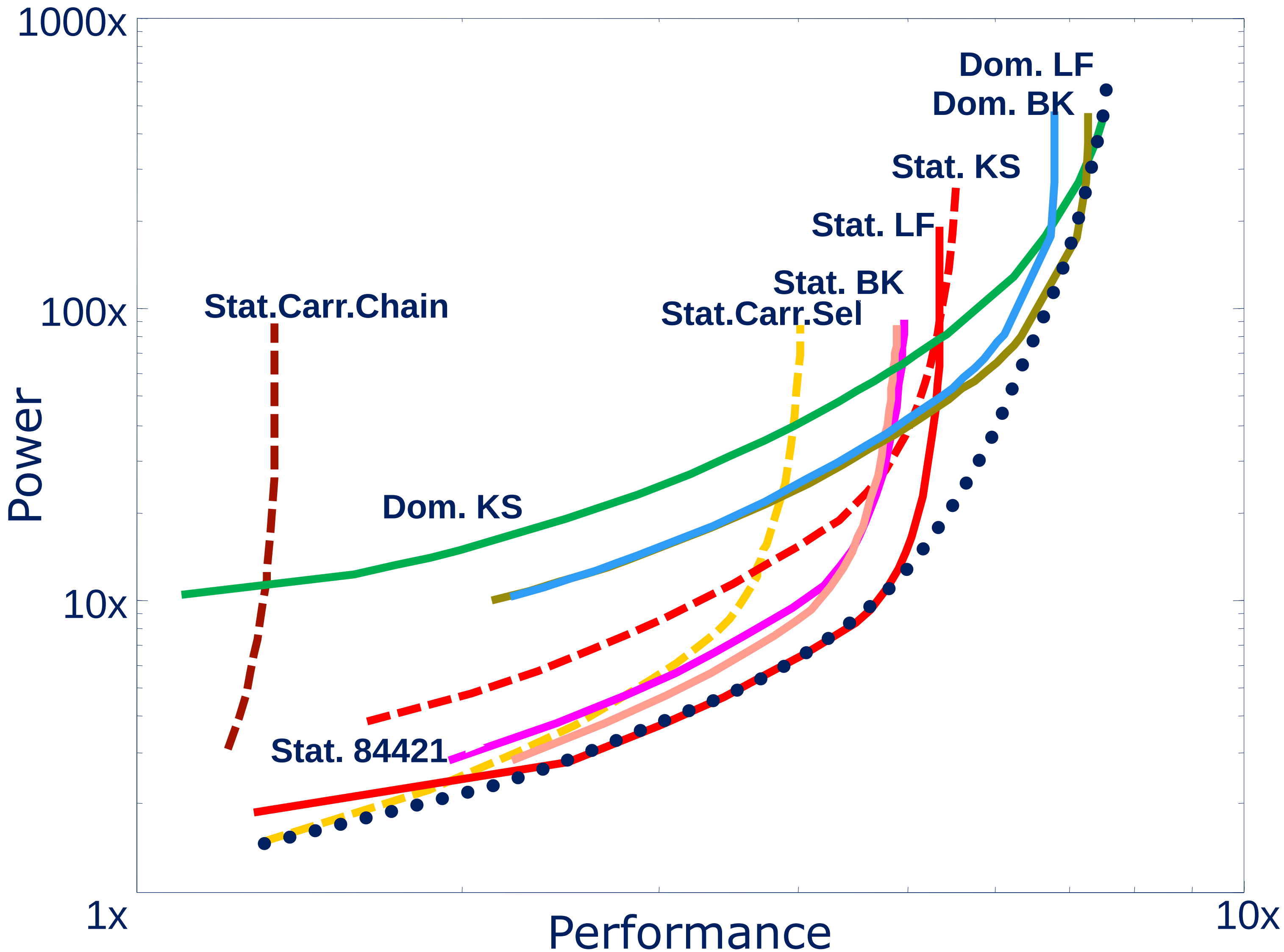
- Introduction
- Impact of process scaling
- Active power optimization
  - System
  - Circuit
- Power management
- Low power silver bullets
- Putting it all together

Optimal Energy-Performance Design Space



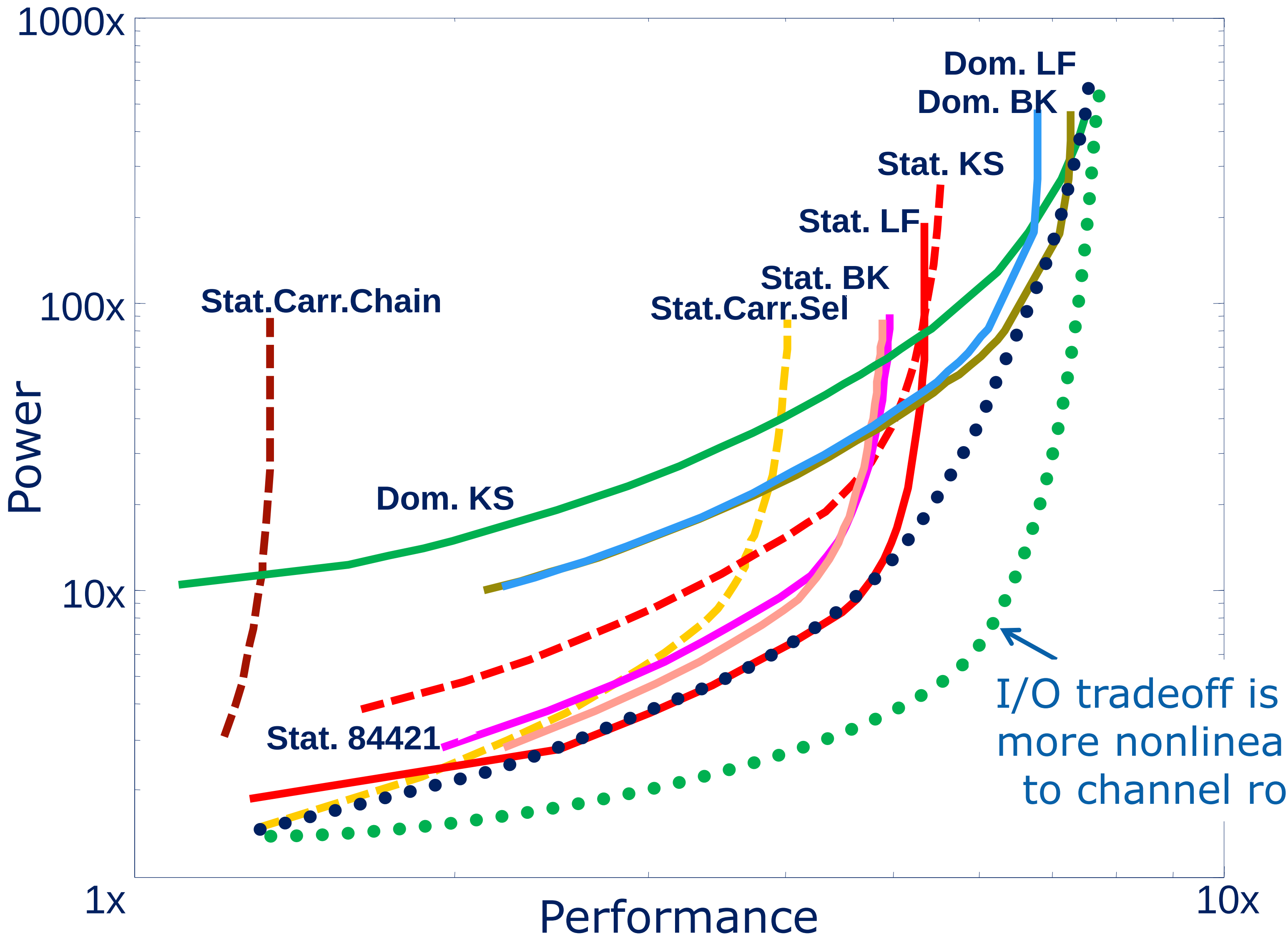
Adapted from Mark Horowitz IEDM 2005

Adder Design Space



Adapted from Mark Horowitz IEDM 2005

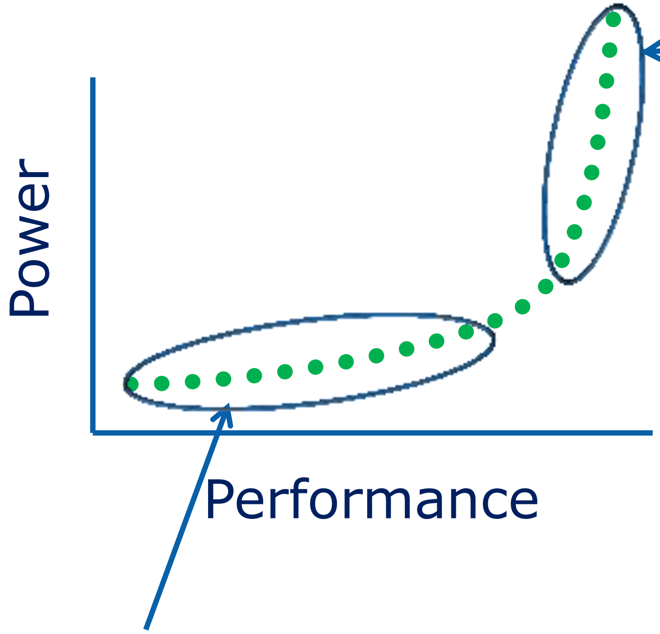
Adder Design Space



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

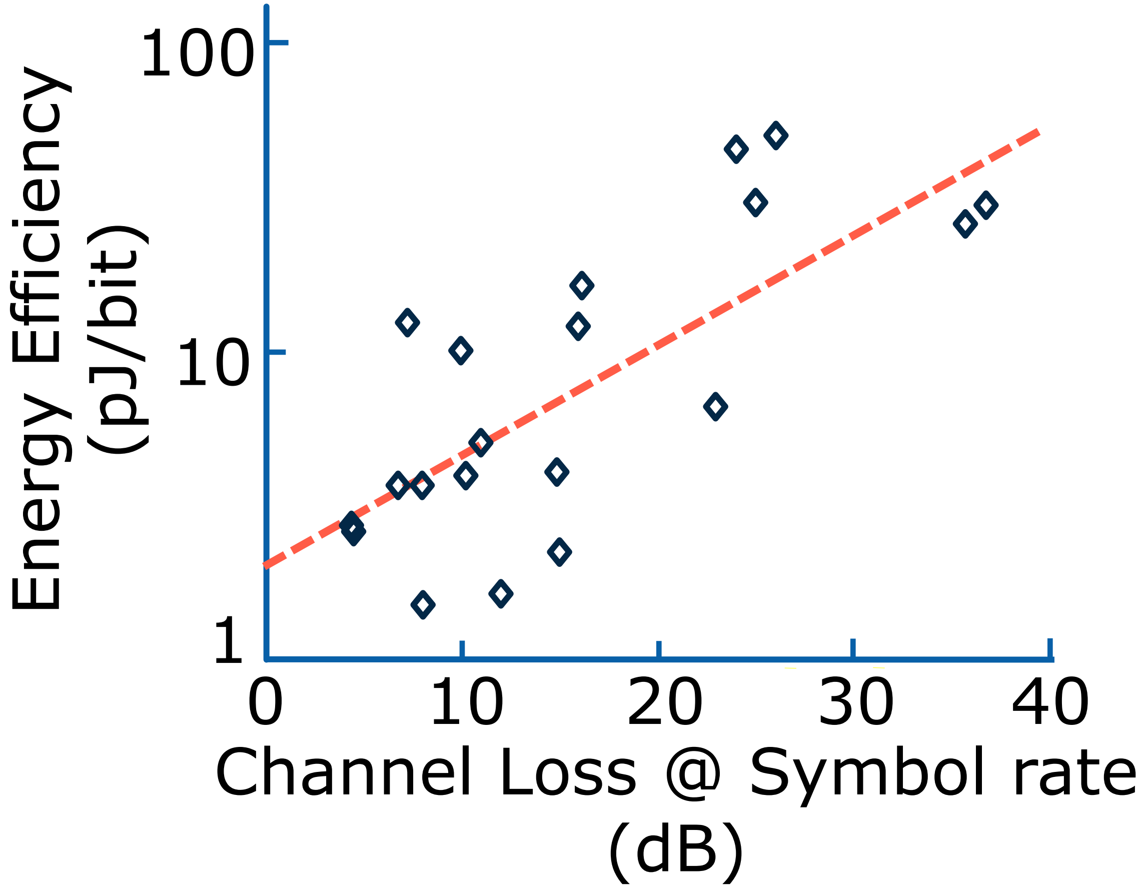
Stingy System Architect

- 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

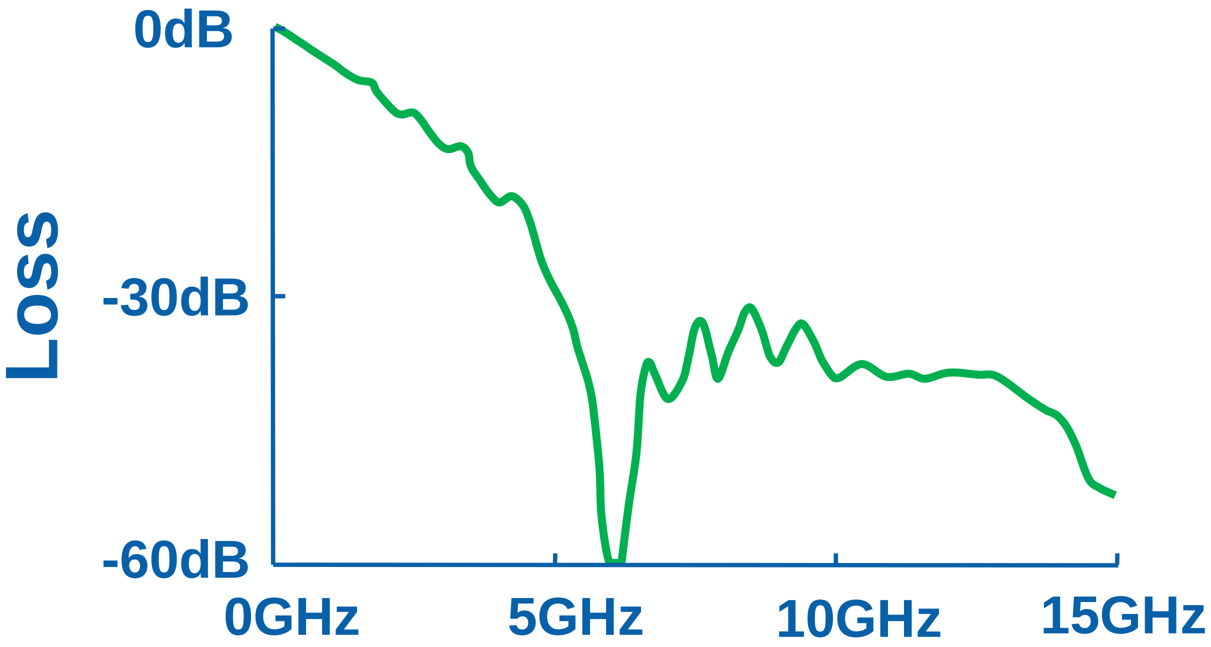
Macho Link Designer

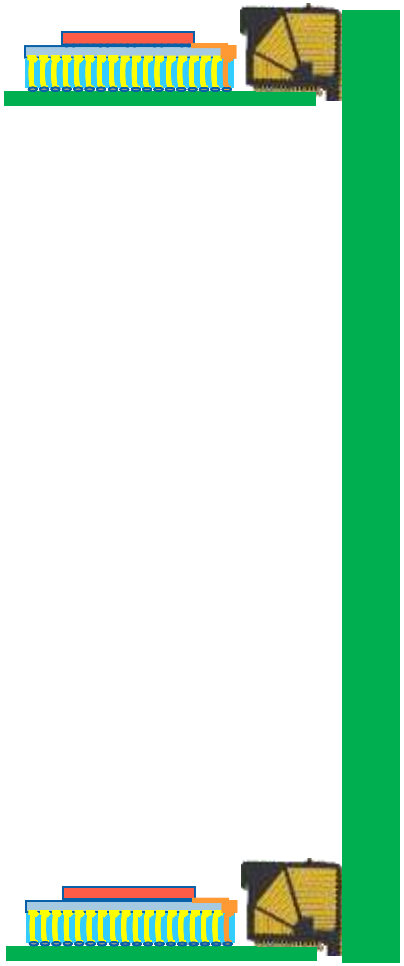
- Knows Shannon's Capacity
- Takes on challenge to apply advanced communication techniques to high-speed links
  - e.g. DSL, Ethernet
- Thinks Moore's law will eventually resolve power & complexity issues

Energy Efficiency Correlation to Loss

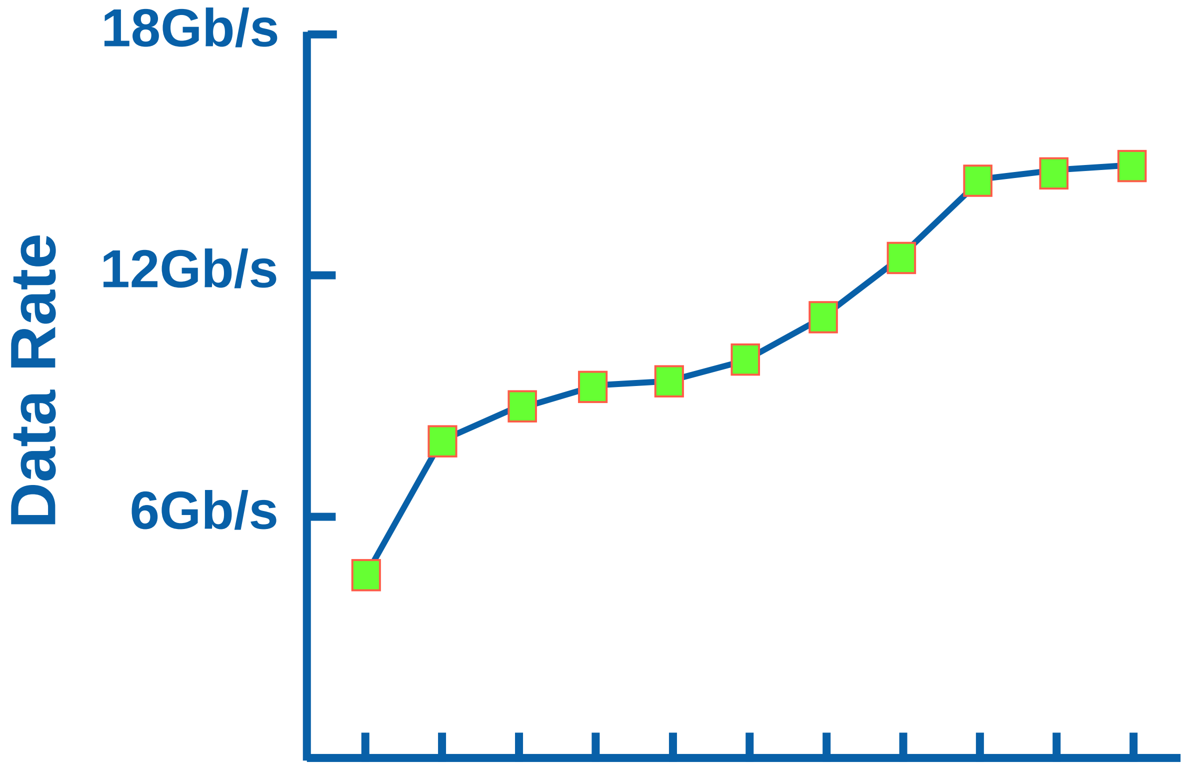


Legacy Backplane Channel





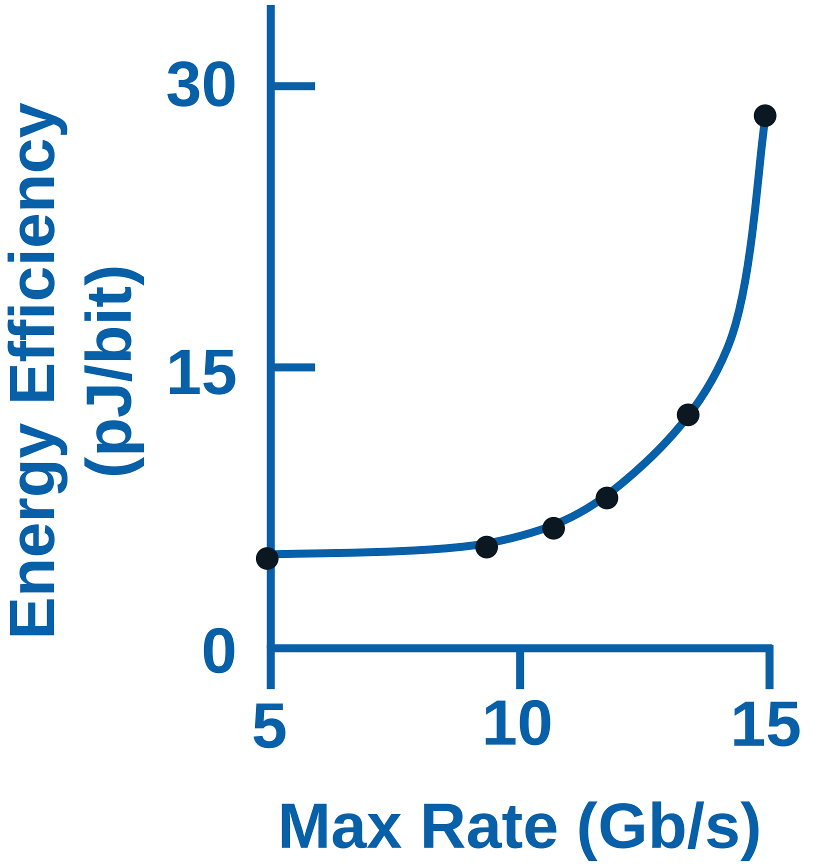
Backplane Data Rates



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

Increasing Equalizer Complexity (nonlinear scale)

Channel Power Wall

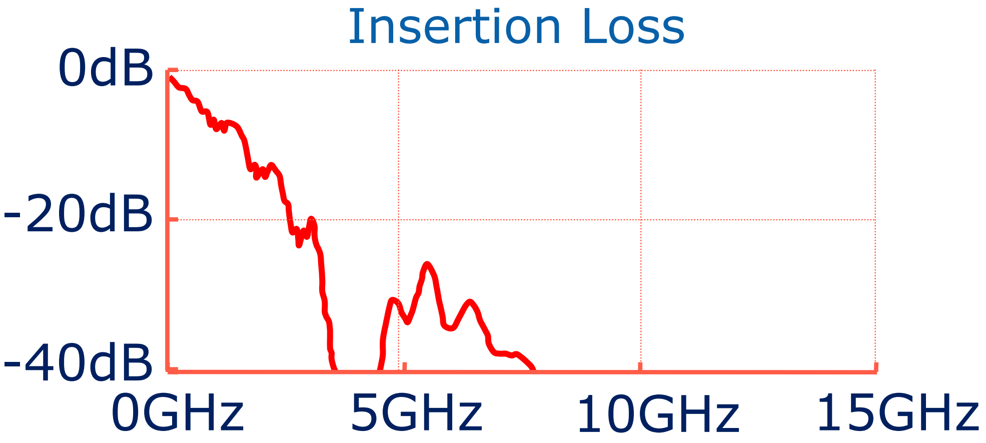


Increasing equalization complexity

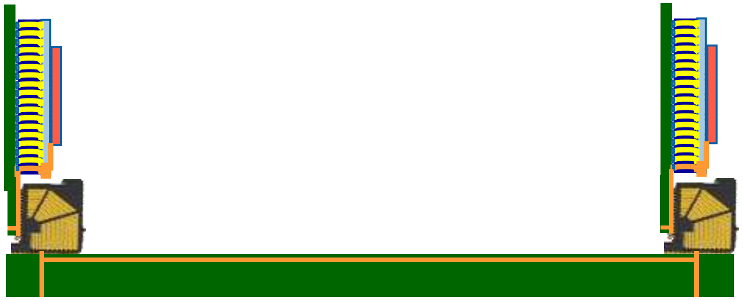
Overextending channel leads to nonlinear EQ power vs. performance tradeoff

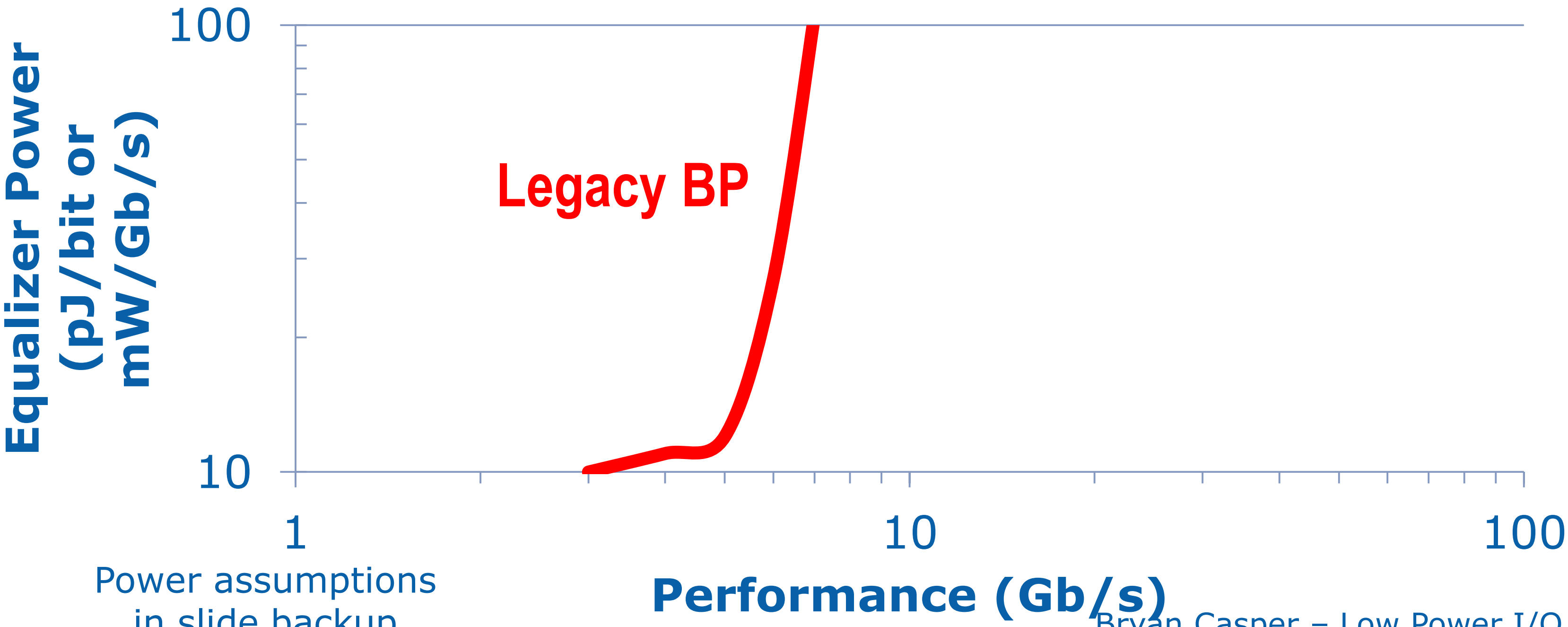
5pJ/bit baseline, ½pJ/bit/DFE tap, ¼pJ/bit/TX tap

I/O Challenges: Power



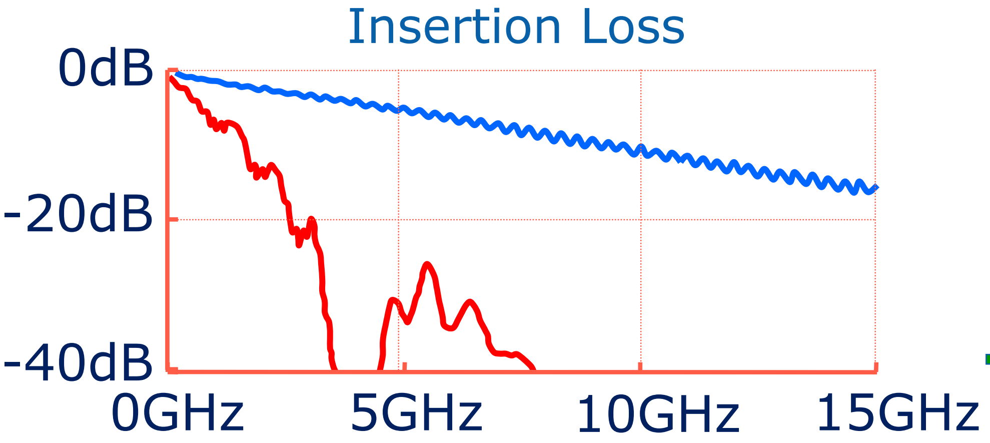
- Legacy backplane w/ 2 connectors & sockets, ½m FR4



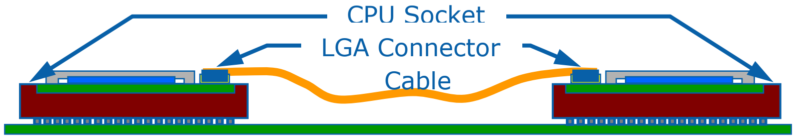


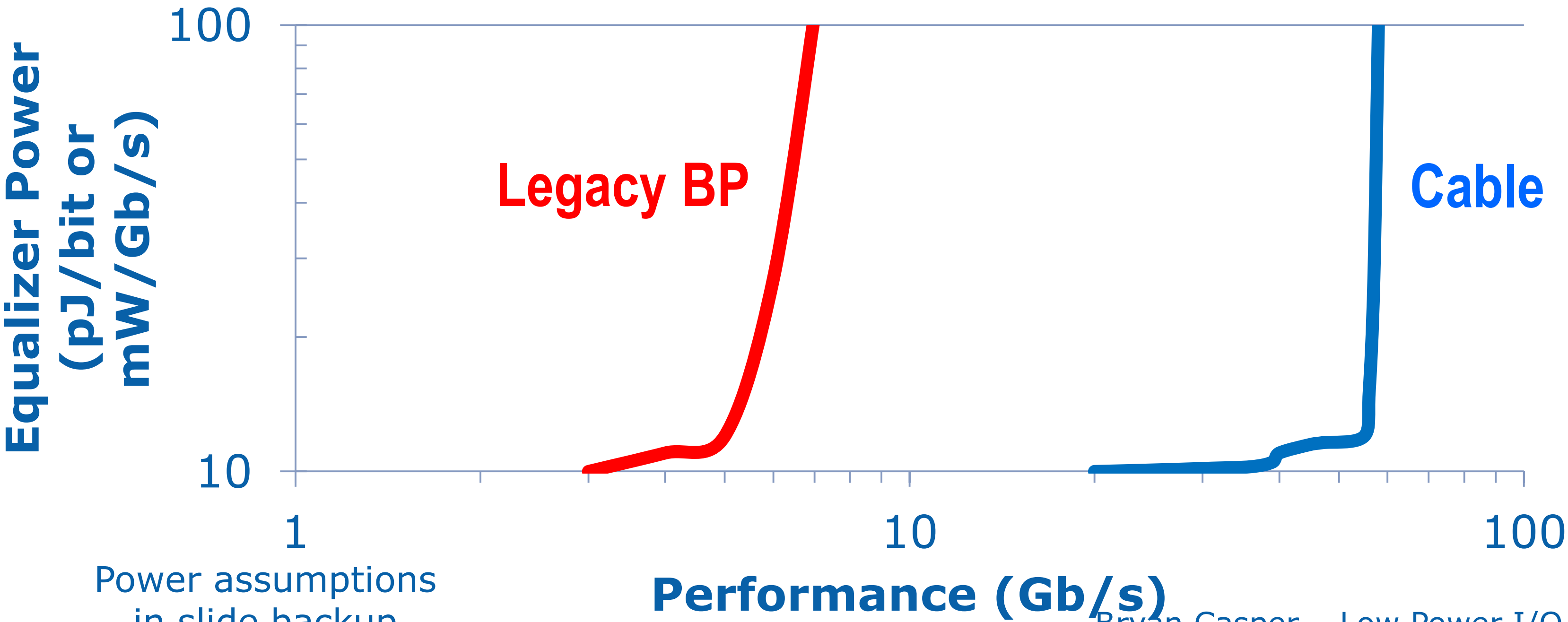
Power assumptions in slide backup

I/O Challenges: Power



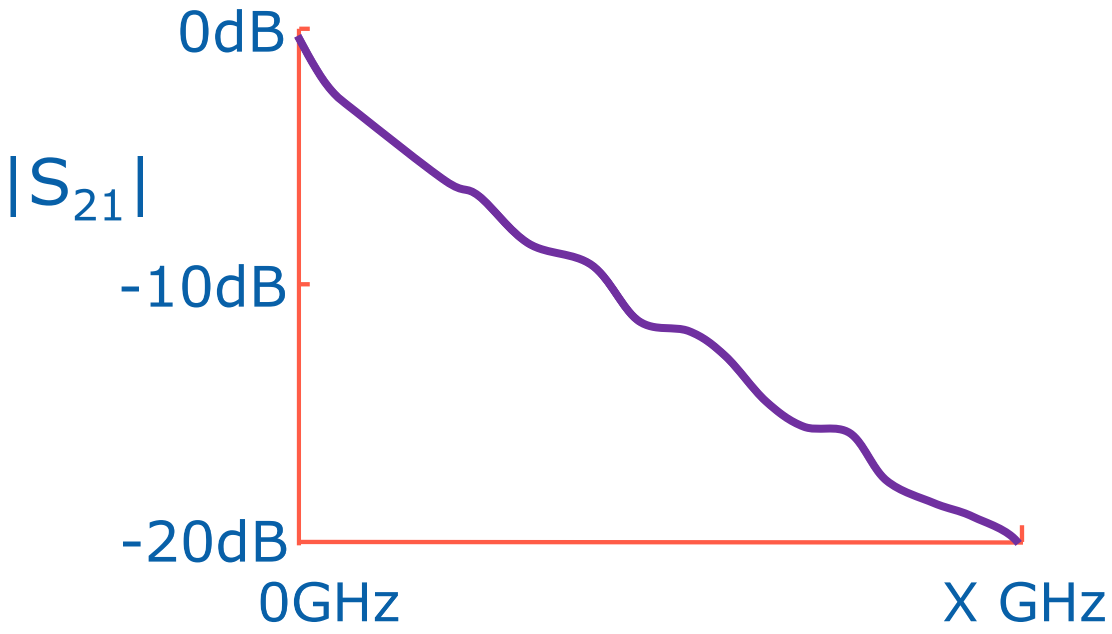
- Legacy backplane w/ 2 connectors & sockets, ½m FR4
- ½m cabled topology with top-pkg connectors





Power assumptions in slide backup

Loss/rate/power estimates

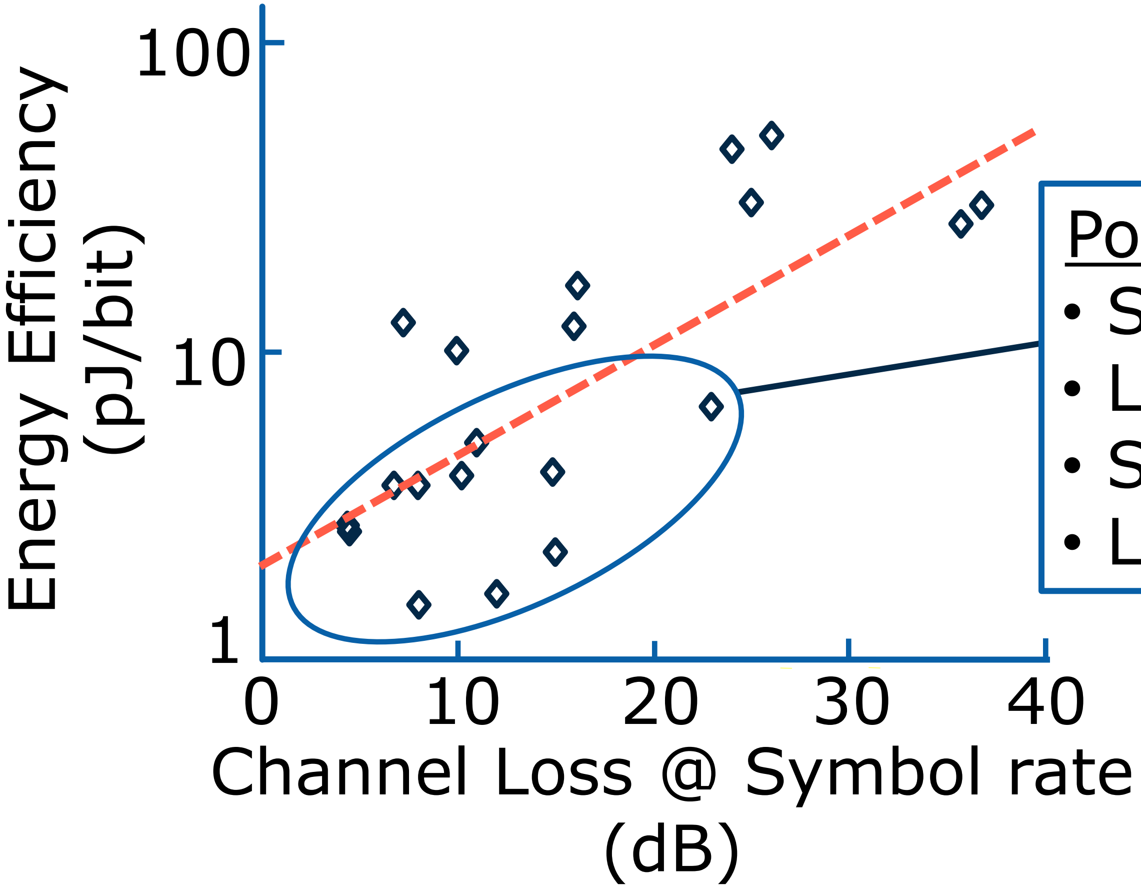


Assuming:

- Xtalk not primary limiter
- TX+RX jitter ~1/2UI
- RX noise 1mV$_{rms}$
- TX swing ~1Vdiffp-p
- Cabled link w/ connectors
  - Channel "well behaved"

Equalization Complexity	Est. data rate	Normalized power (rough guess)
None	**~0.8*X Gb/s**	**1**
Low power	**~2.0*X Gb/s**	**~1**
Moderate (3 tap LE, 4 tap DFE)	**~2.4*X Gb/s**	**~2**
Complex (>50 tap LE+DFE)	**~3.6*X Gb/s**	**~10-100**
Complex EQ+PAM+FEC/coding	**~4.4*X Gb/s**	**~100-1000**
Shannon's capacity	**~8-10*X Gb/s**	**n/a**

Common Traits of Low Power Links



Power Optimized Links

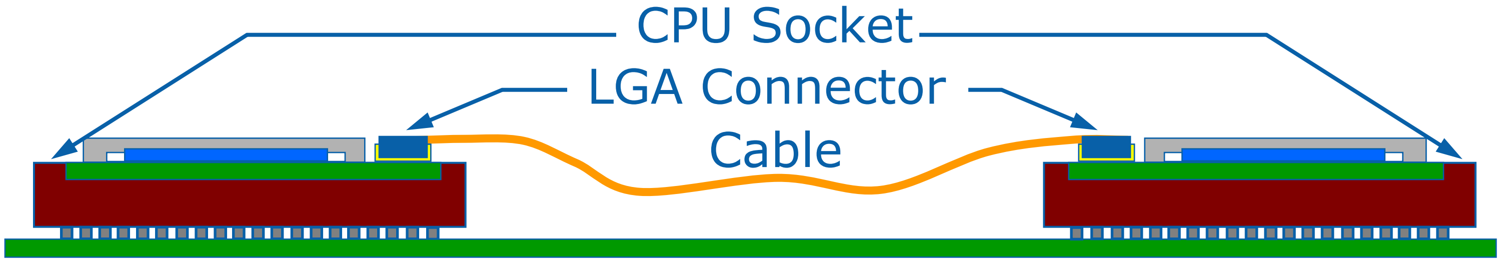
- Simple equalization
- Low TX swing
- Sensitive RX sampler
- Low-power clocking

How to scale rate or distance <u>and</u> maintain energy efficiency

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

<u>½ Meter Channel Examples (based on Intel Labs Measurements)</u>

- PCIe (2 connector → 20dB @4GHz
- LCP Flex* → 20dB @15GHz
- Twinax 36 AWG* → 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

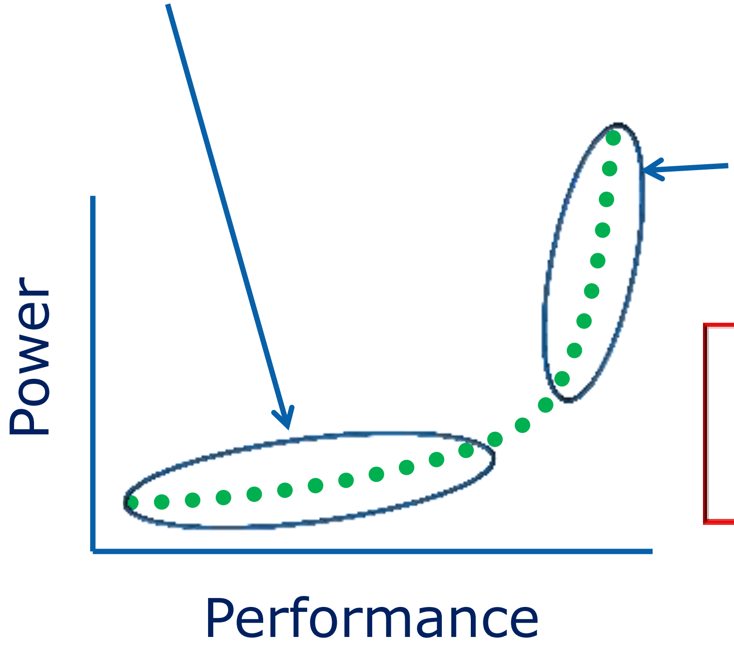
- Not a comprehensive list
  - More like a sampling of known power reduction methods
- Few low power links incorporate all of these techniques
  - Most incorporate at least some
- Not intended to be a detailed overview of each method

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

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**

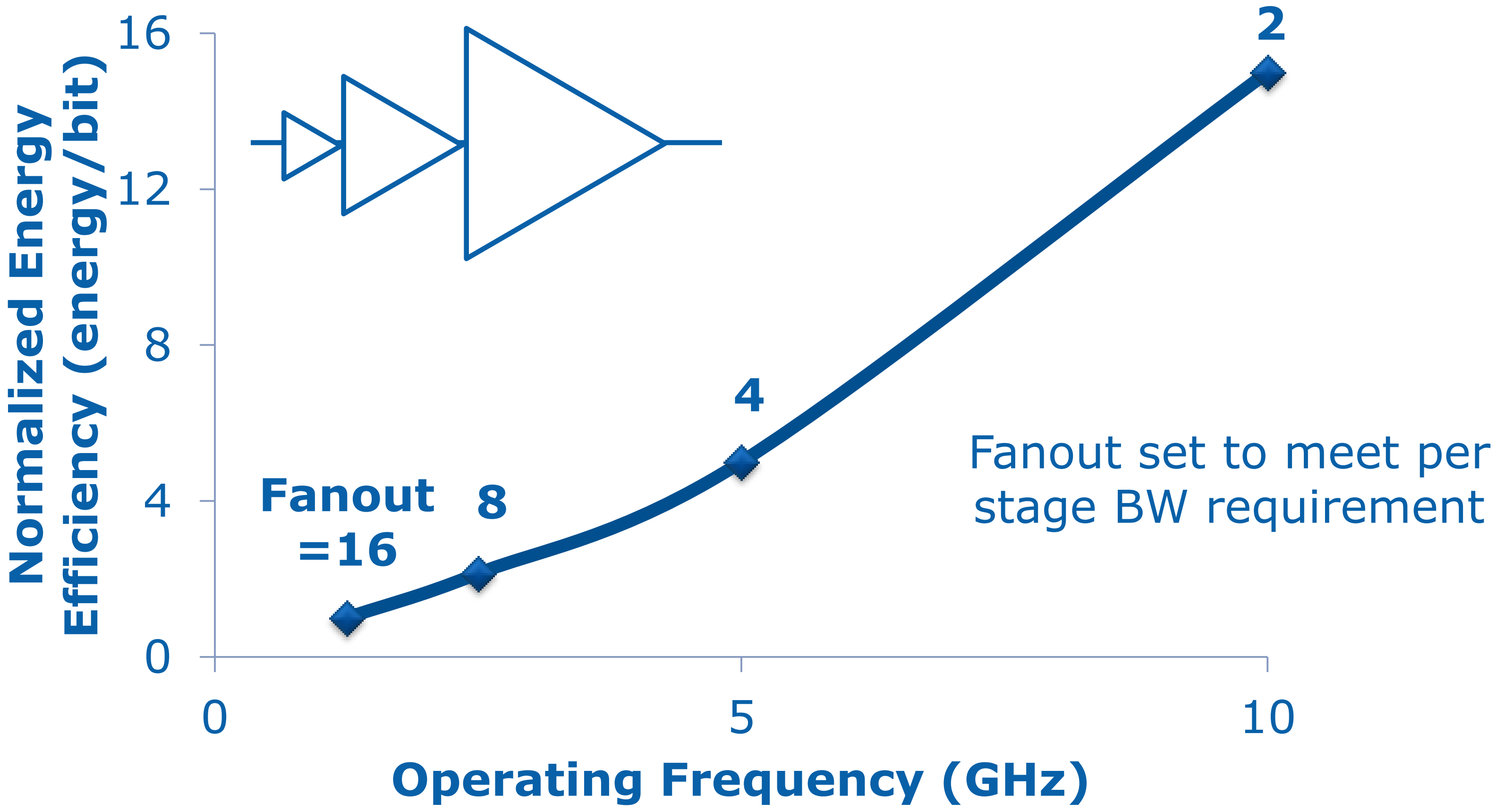
Key to low power links is operating on this portion of design space



Steep tradeoff caused by:

1. Channel BW limit
2. Process BW limit
3. Circuit architecture complexity

Clock Buffer Power/Performance Example



Stay off process BW cliff

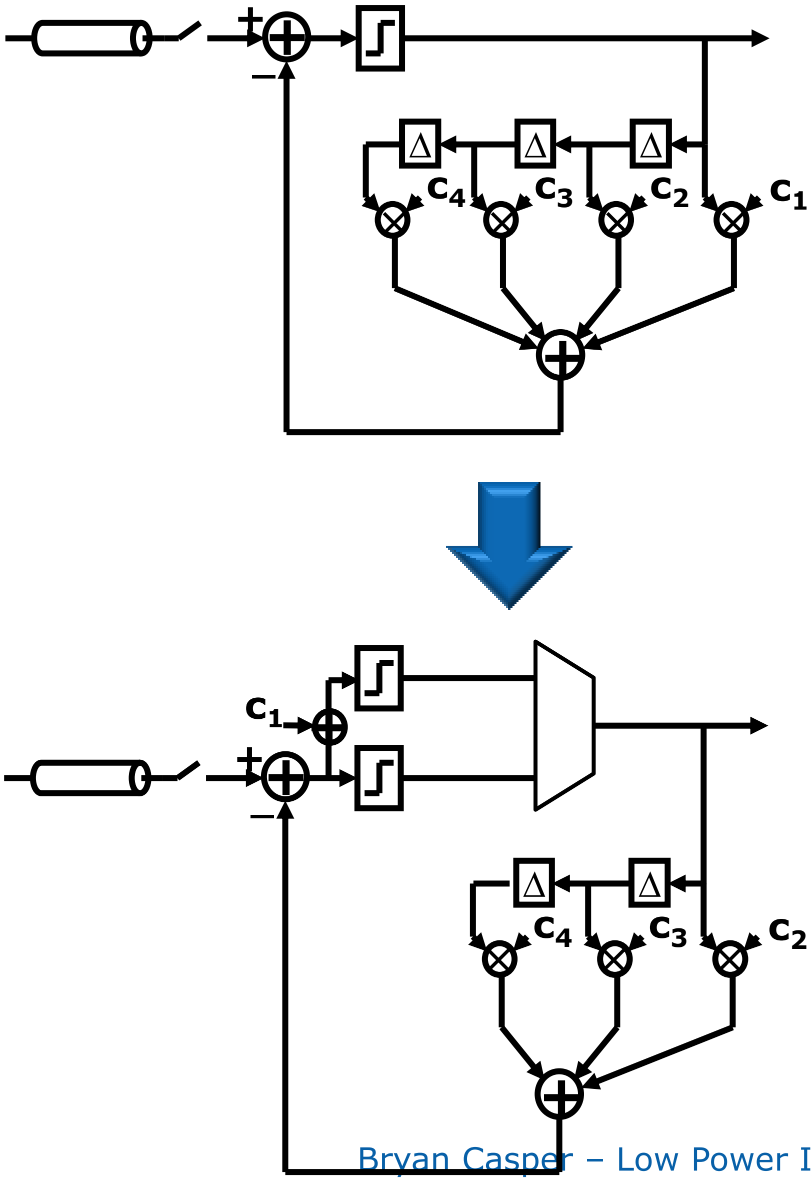
Performance Impact on Circuit Architecture: Loop-unrolled DFE

Conventional DFE

- Speedpath limits frequency

Loop-unrolled DFE

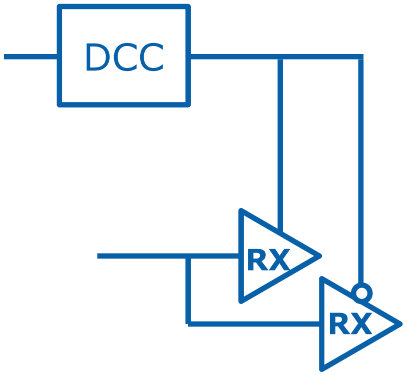
- Redundancy to alleviate speedpath
- Increases power and complexity
  - Proportional to C1*2^N+C2
    - C1=comparator + mux
    - C2=baseline power
    - N=number taps unrolled



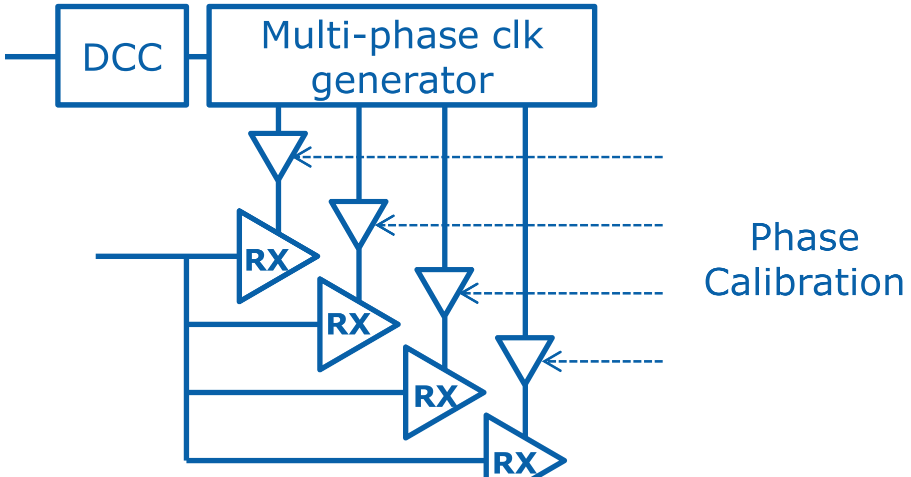
Performance Impact on Circuit Architecture: Multi-phase Clocking

- 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

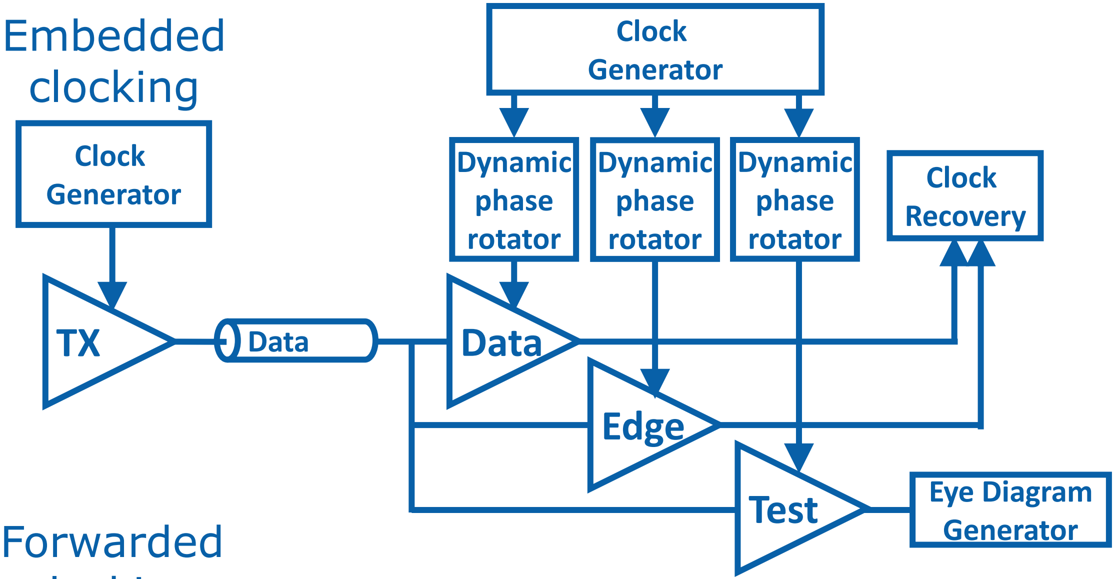


Quarter rate clock

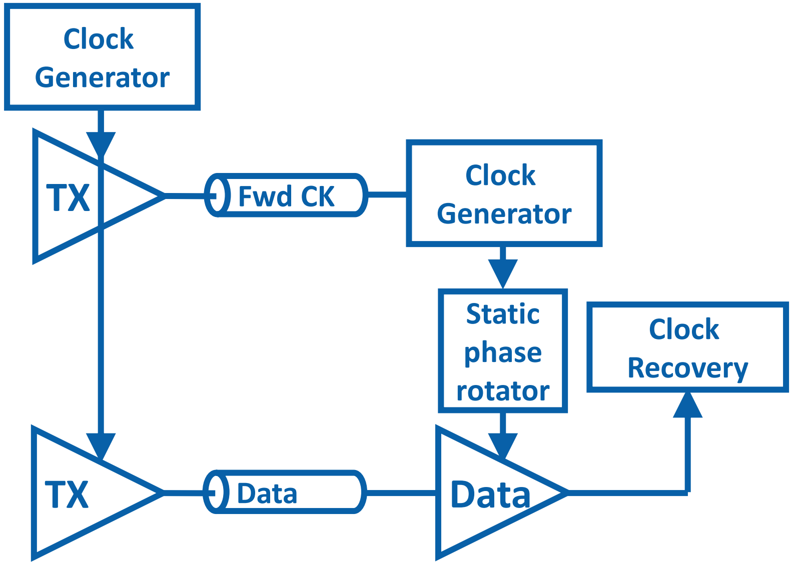


Top Ten #2: Forwarded clocking

Embedded clocking



Forwarded clocking



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**

Forwarded clock power benefits

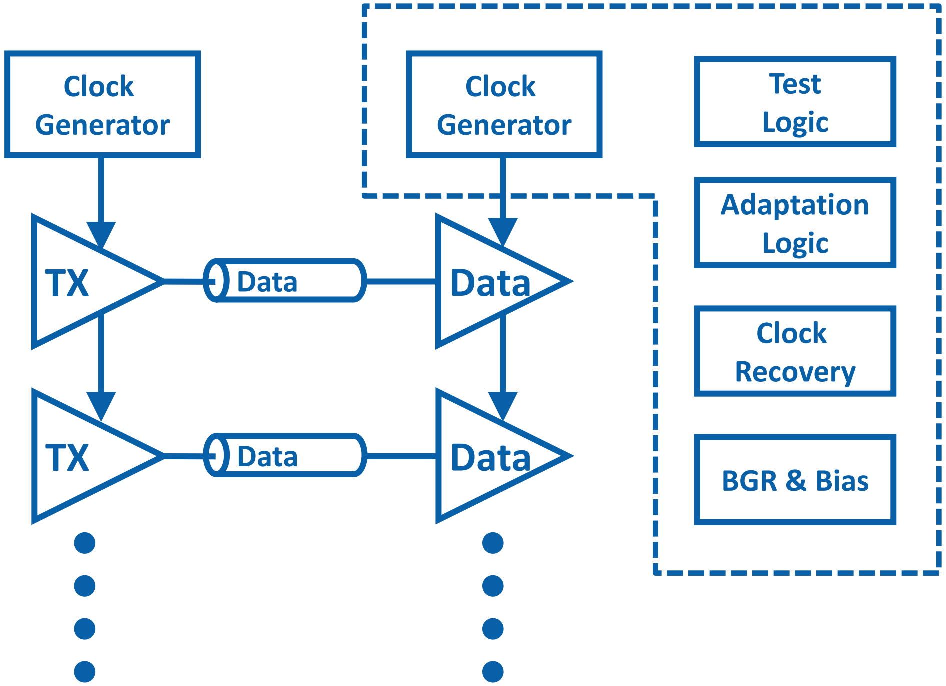
- No need for high clock recovery BW
- Edge/test samplers optional
  - Clock recovery/test can be time multiplexed with data samplers
- Fewer, simpler phase rotators
  - Greater tolerance for INL & jitter
  - 1 rotator can cover data, edge & test in a time-multiplexed fashion

Top Ten #3: Global Circuit Sharing

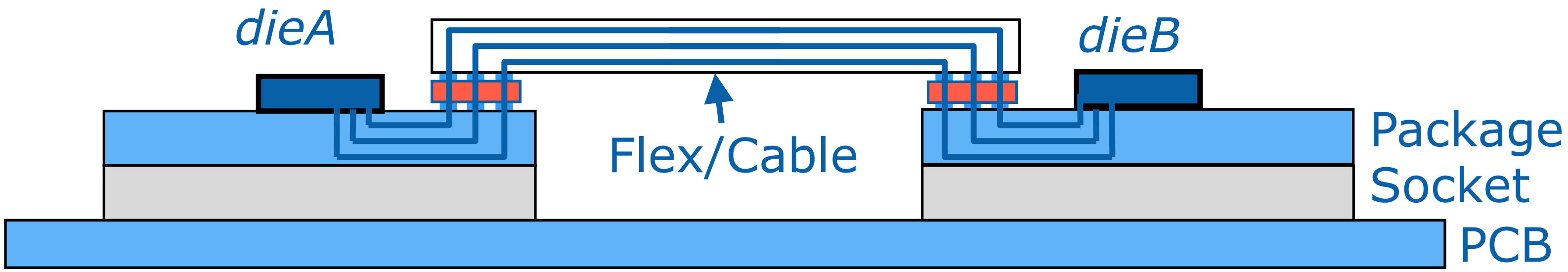
- Parallel link implementations have ample opportunity to share common functionality

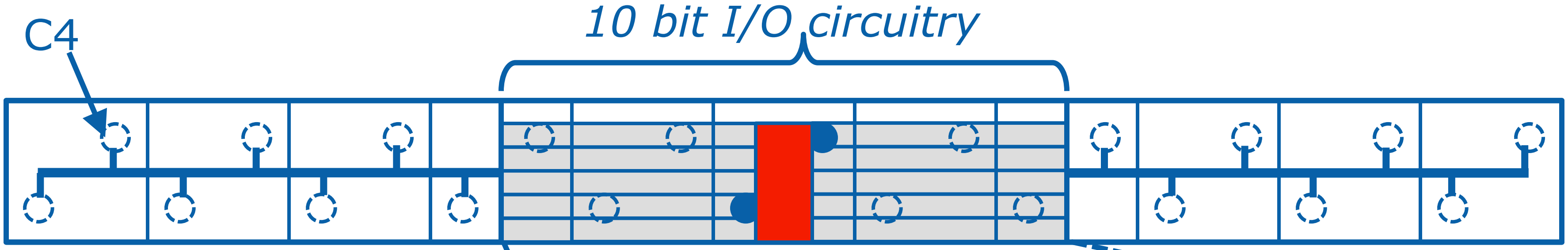
1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling

Potential to be shared across parallel link

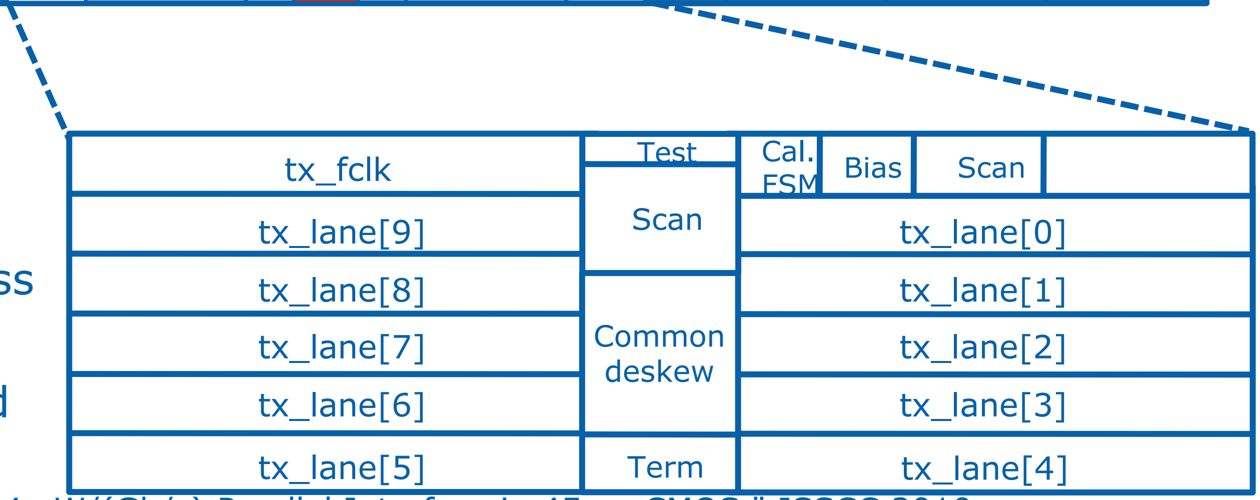


Co-optimization of Channel and Circuits Enables Widespread Power Amortization





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

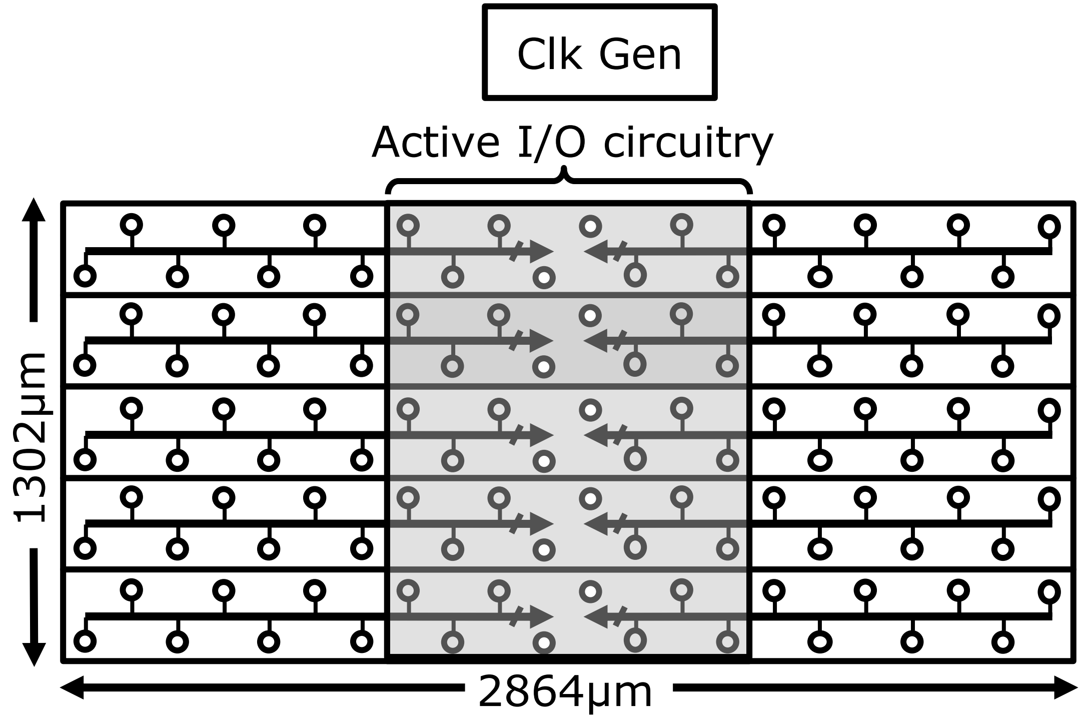


F. O'Mahony, et. al., " A 47×10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS," ISSCC 2010

Top Ten #4: Low power clock distribution

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

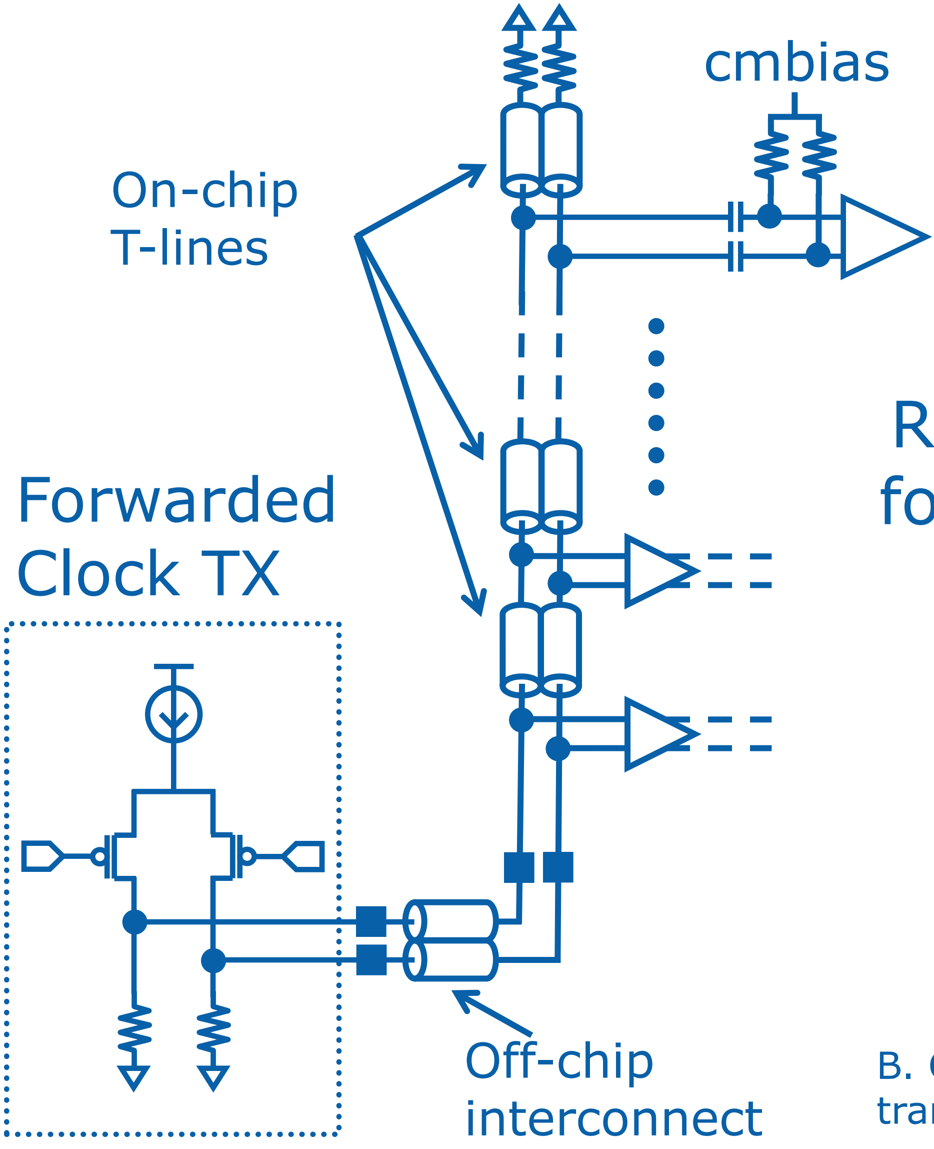
1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling



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

**B. Casper, F. O'Mahony, "Clocking Analysis, Implementation and Measurement Techniques for High-Speed Data Links—A Tutorial," TCAS1, Jan. 2009

Forwarded Clock Repeater-less Distribution



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

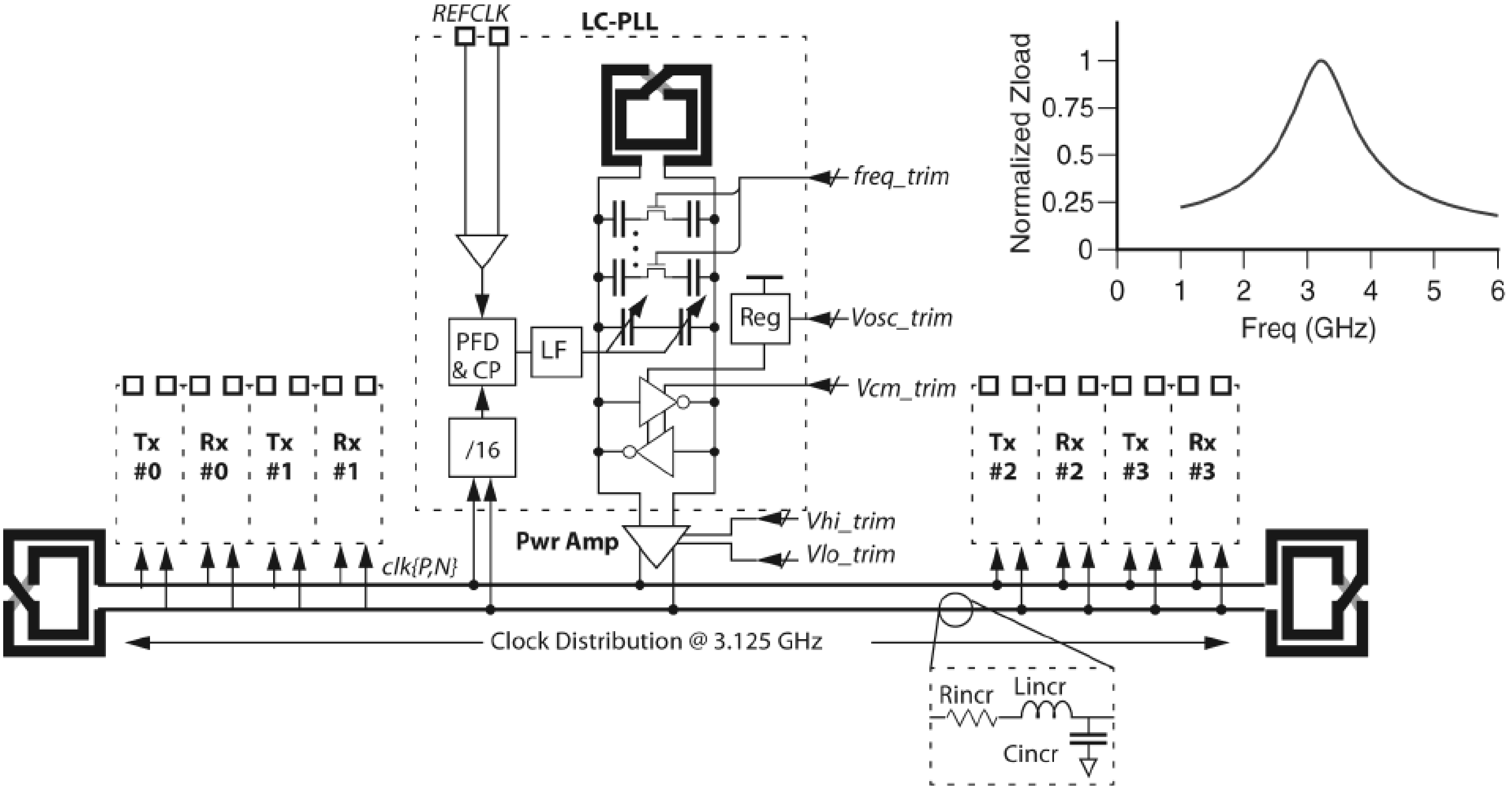
B. Casper, et. al., “A 20Gb/s forwarded clock transceiver in 90nm CMOS,” ISSCC 2006

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

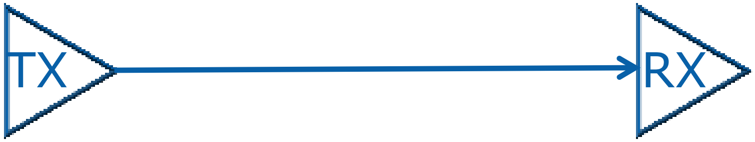
J. Poulton, et al., “A 14-mW 6.25-Gb/s Transceiver in 90-nm CMOS,” JSSC, Dec., 2007.

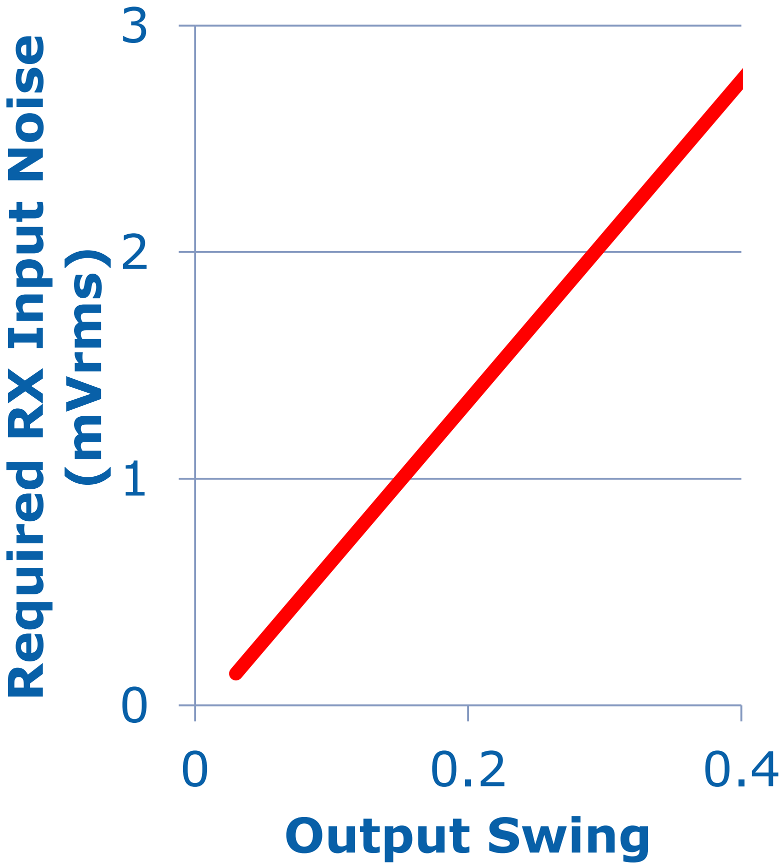
Top Ten #6,7: Co-designed TX & RX

- 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

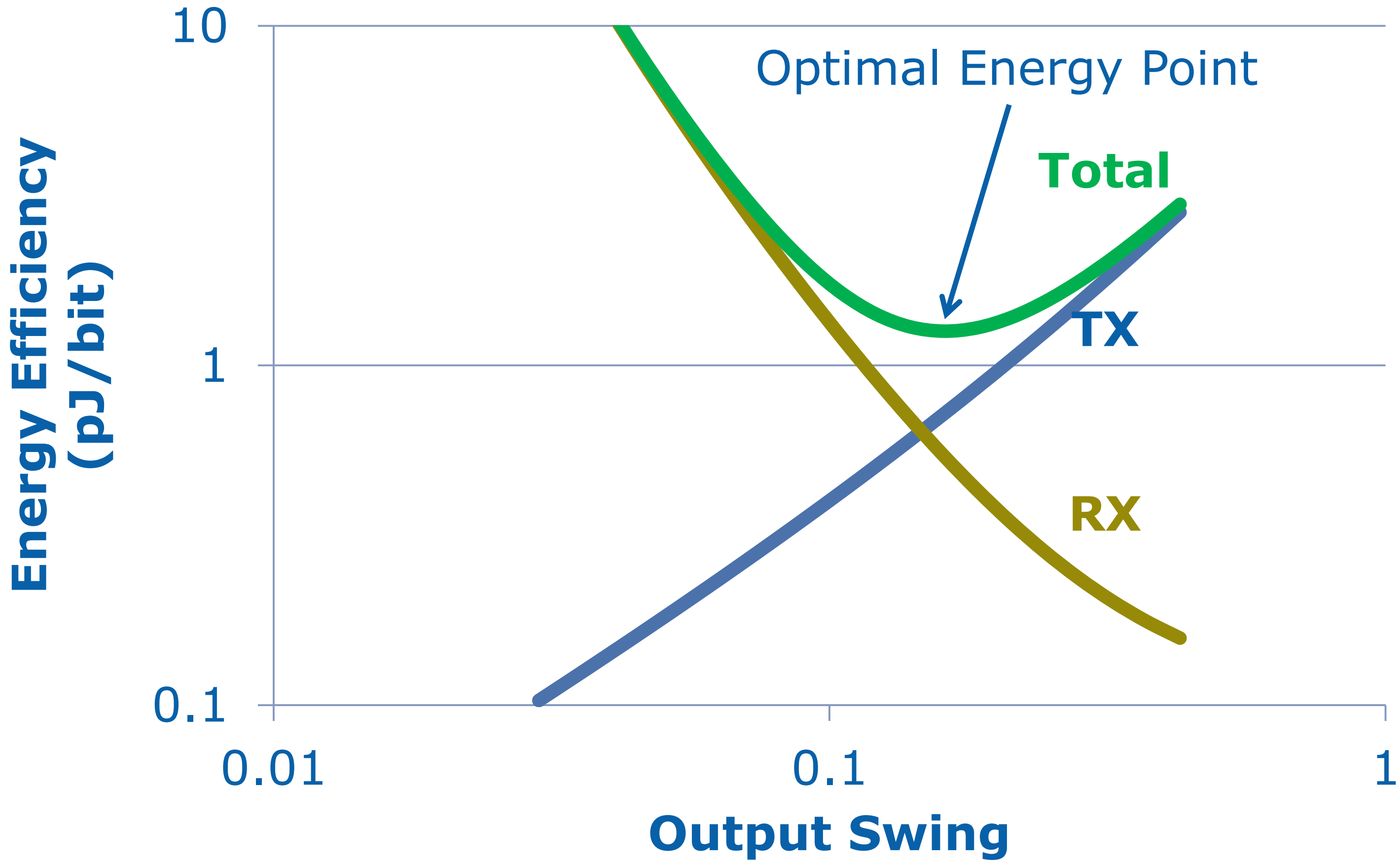




Assumptions:

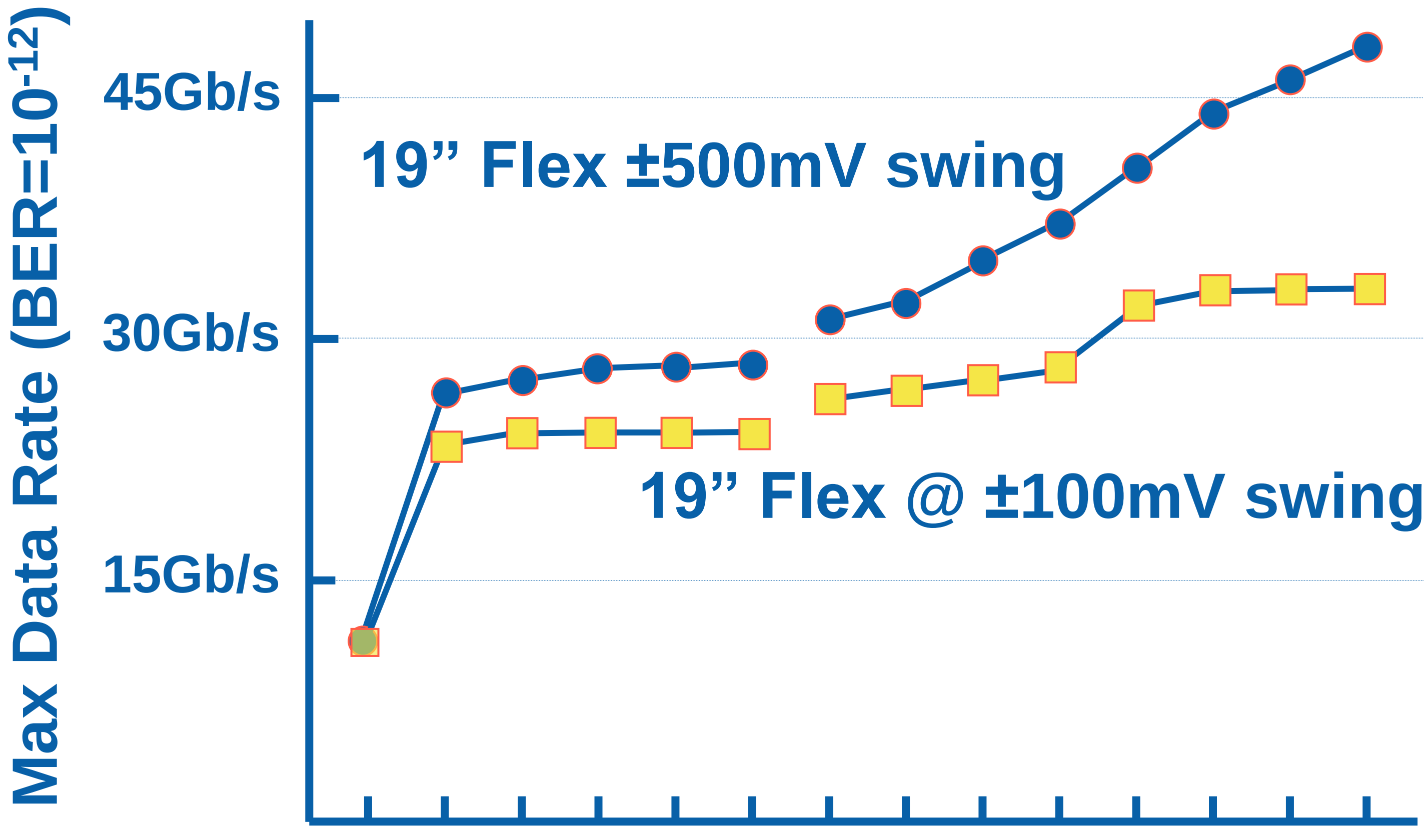
- RX noise variance proportional to RX power
  - 5mW → 1mVrms
- Normally distributed ISI sigma=0.001*swing
- 1E-12 BER target
- Voltage-mode TX w/ linear reg.
- Channel loss = -20dB

Simplistic Example: Swing vs. Efficiency



- Optimal energy at ~160mVpp
  - Requires ~1mV$_{rms}$ input referred RX noise

Low Swing Tradeoffs: 19” Cabled Link Maximum Rates



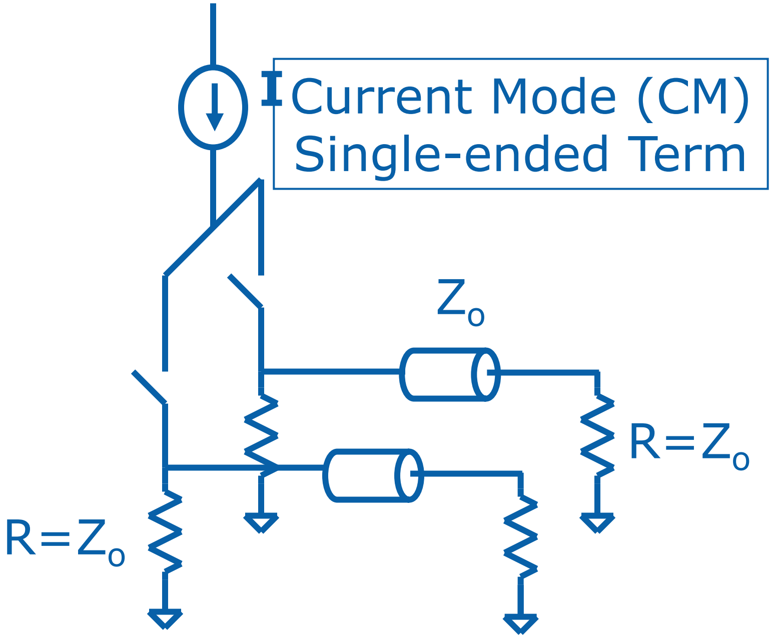
TX FIR taps	1	2	3	4	5	6	4	4	4	4	4	4	4	4
DFE taps							1	2	4	8	16	32	64	128

- 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



Current Mode (CM)
Single-ended Term

$V_{d,1}$ = (I/2)R

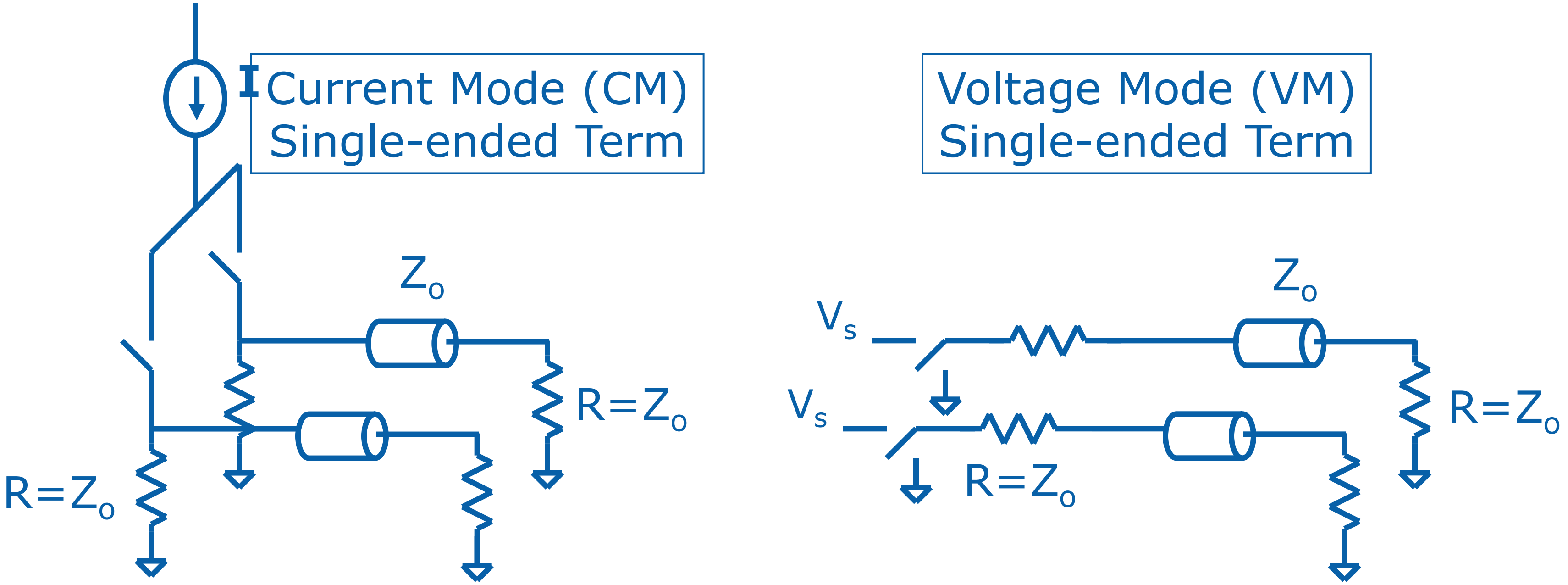
$V_{d,0}$ = -(I/2)R

$V_{d,pp}$ = IR

I = ($V_{d,pp}$/ R)

Source: Ganesh Balamurugan

Low-Power TX Drivers: CM vs. VM



Current Mode (CM)
Single-ended Term

Voltage Mode (VM)
Single-ended Term

$V_{d,1} = (I/2)R$

$V_{d,0} = -(I/2)R$

$V_{d,pp} = IR$

$I = (V_{d,pp}/ R)$

2X power reduction

$V_{d,1} = (V_s / 2)$

$V_{d,0} = -(V_s / 2)$

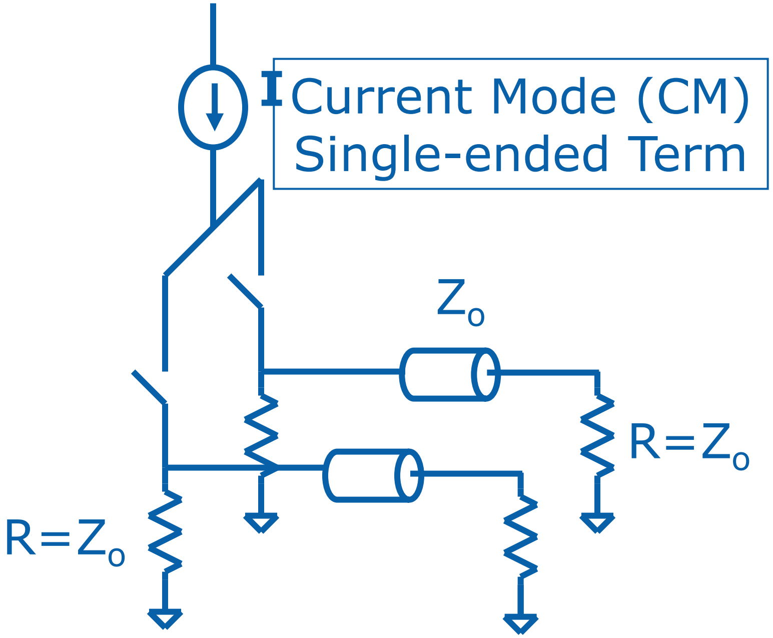
$V_{d,pp} = V_s$

$I = (V_s / 2R)$

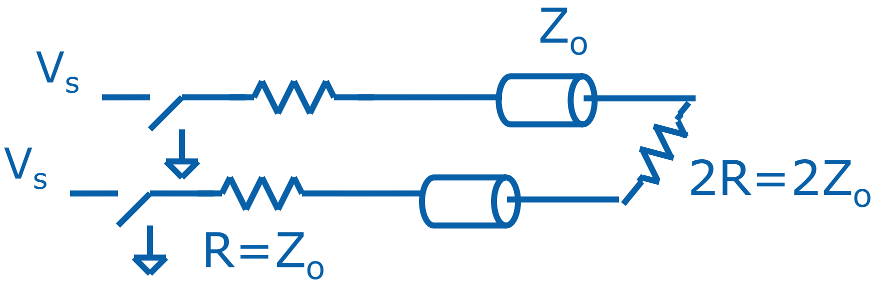
$I = (V_{d,pp}/ 2R)$

Source: Ganesh Balamurugan

Low-Power TX Drivers: CM vs. VM



Voltage Mode (VM)
Differential Term



$V_{d,1} = (I/2)R$

$V_{d,0} = -(I/2)R$

$V_{d,pp} = IR$

$\mathbf{I = (V_{d,pp}/\ R)}$

4X power reduction

$V_{d,1} = (V_s / 2)$

$V_{d,0} = -(V_s / 2)$

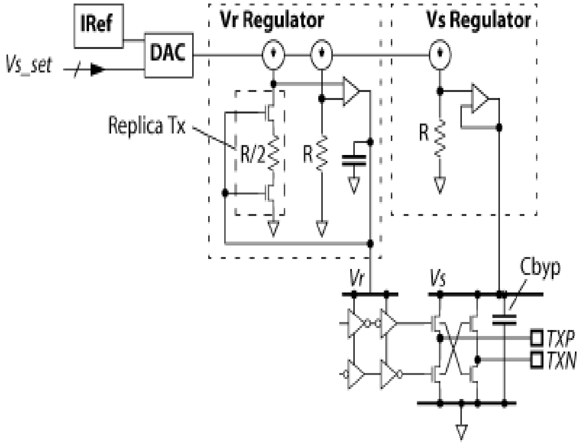
$V_{d,pp} = V_s$

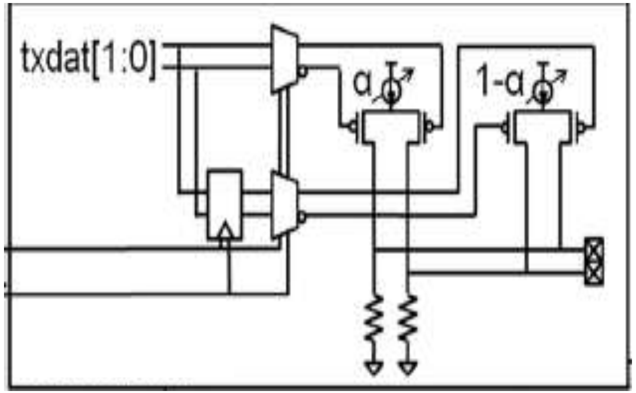
$I = (V_s / 4R)$

$\mathbf{I = (V_{d,pp}/\ 4R)}$

Source: Ganesh Balamurugan

Low-Swing TX Drivers: CM vs. VM





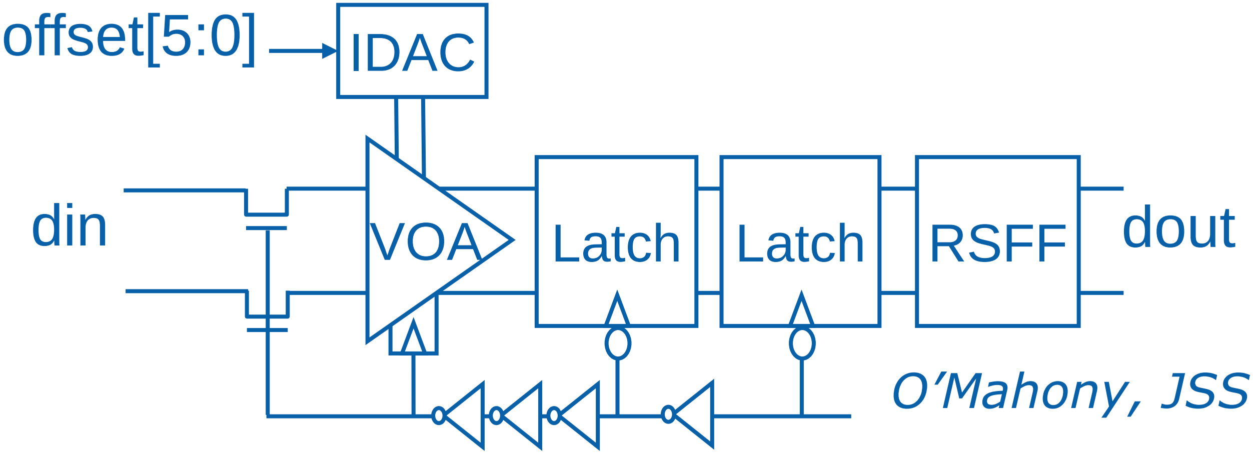
	VM (Palmer, JSSC 12/2007)	CM (O'Mahony, JSSC 12/2010)
Vswing	210mVpp-diff	150mVpp-diff
Proc. / Vcc	90nm / 1.0Vcc	45nm / 0.8Vcc
Eq.	No	Yes (2-tap)
Datarate	6.25Gb/s	10Gb/s
Bias cap	36pF	<1pF
TX drv power	1.10mW	2.12mW
TX bias power	0.76mW	0.34mW
Total TX drv. power	1.86mW	2.46mW

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

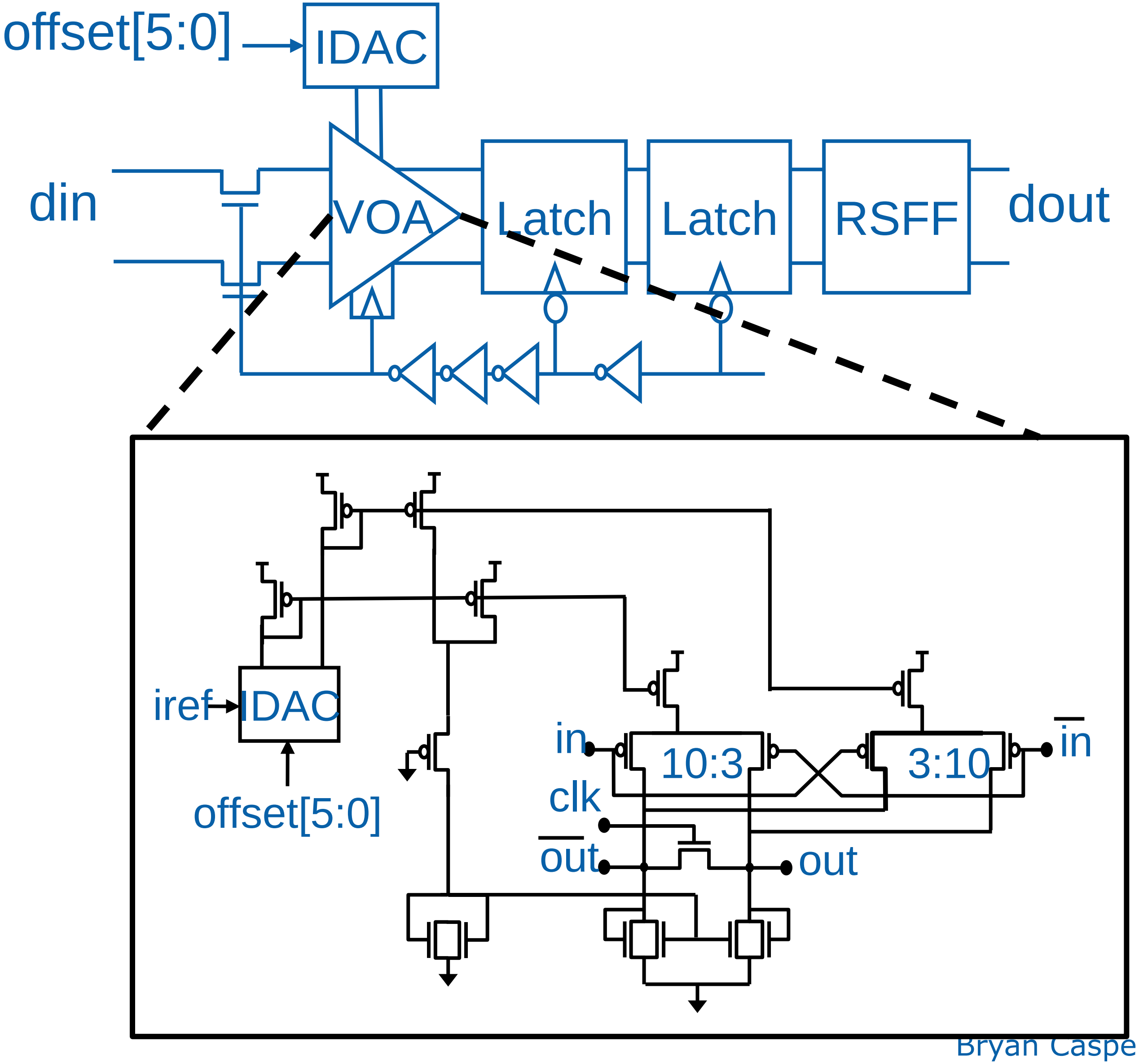


O'Mahony, JSSC Dec. 2010

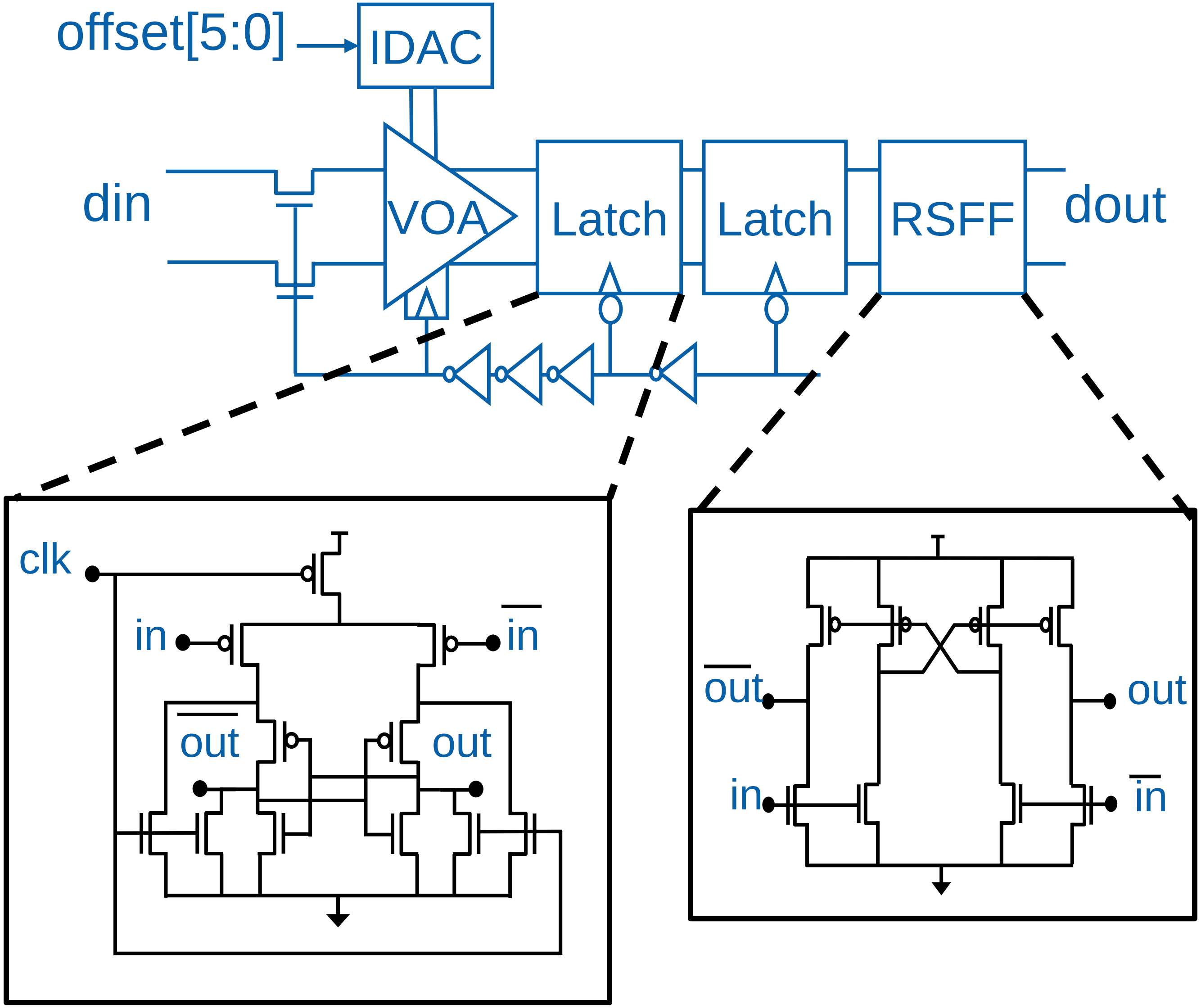
Good receiver sensitivity allows low TX swing

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

Sensitive RX samplers



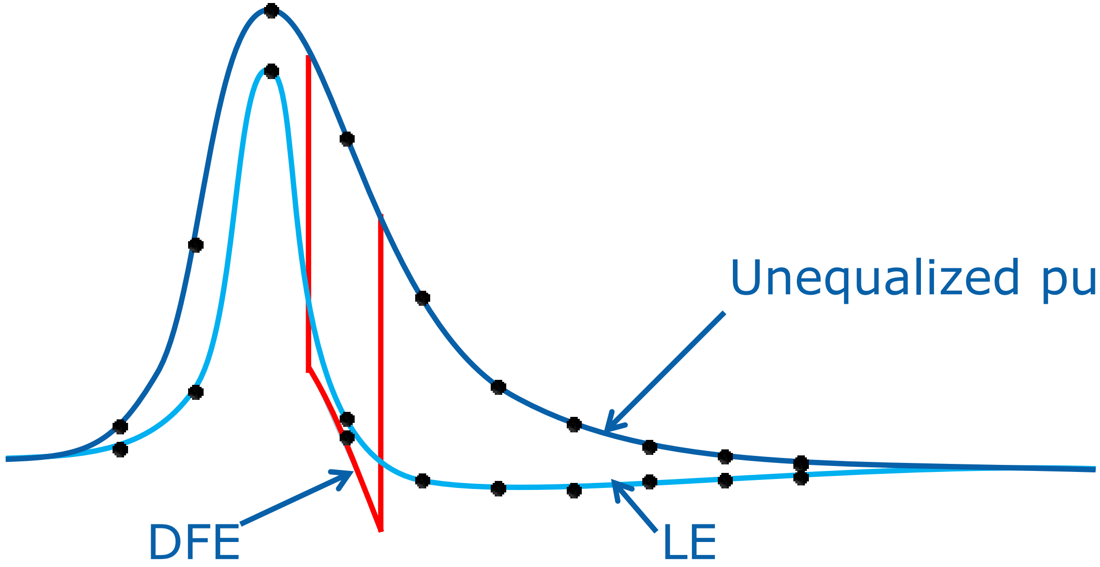
Sensitive RX samplers



Top Ten #8: Simple Equalization

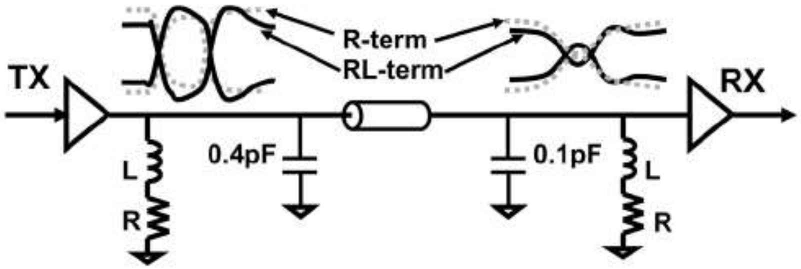
- Linear equalizers - big bang for the buck
  - If channel is “well behaved” and ISI dominated
- DFE is complex, especially if speedpaths
  - 1-tap DFE only cancels 1 postcursor point

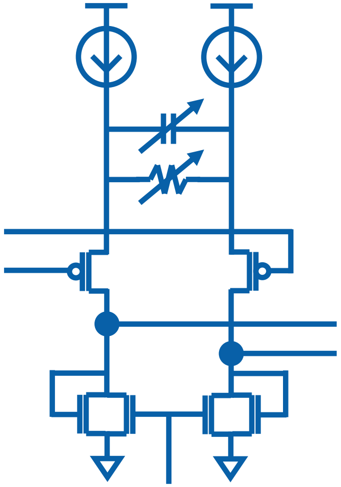
1. **Modest data rates**
2. **Forwarded clocking**
3. **Global circuit sharing**
4. **Low power clock distribution**
5. **Resonantly tuned clocking**
6. **Low swing TX**
7. **Sensitive RX**
8. **Simple equalization**
9. **Calibration and tuning**
10. **System modeling**



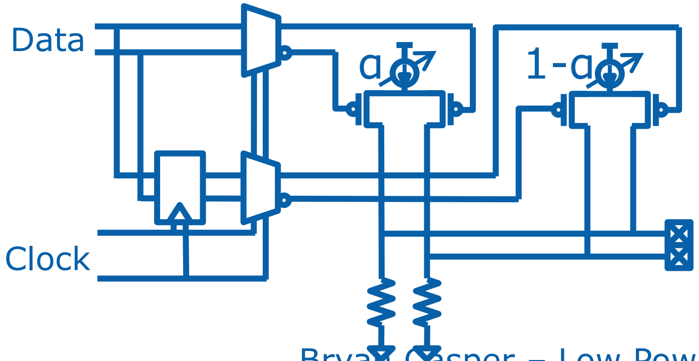
Examples: Low Power Linear Equalizers

- 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

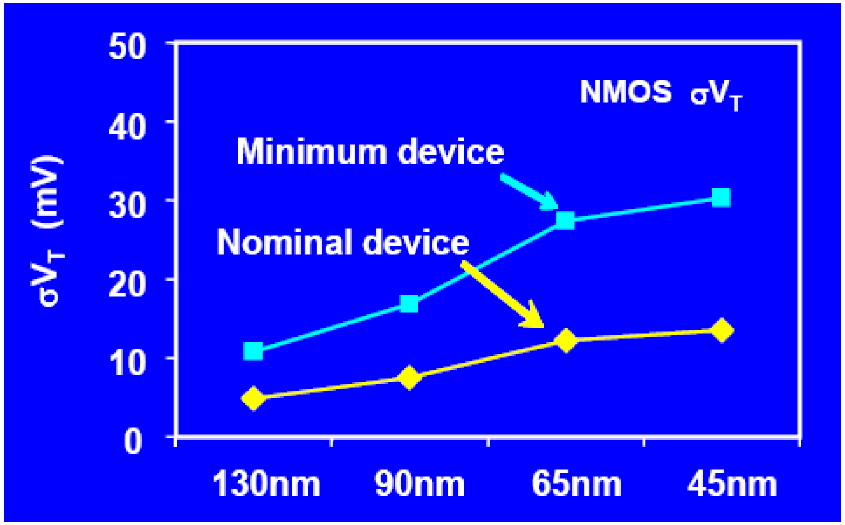


Top Ten #9: Calibration and tuning

- 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

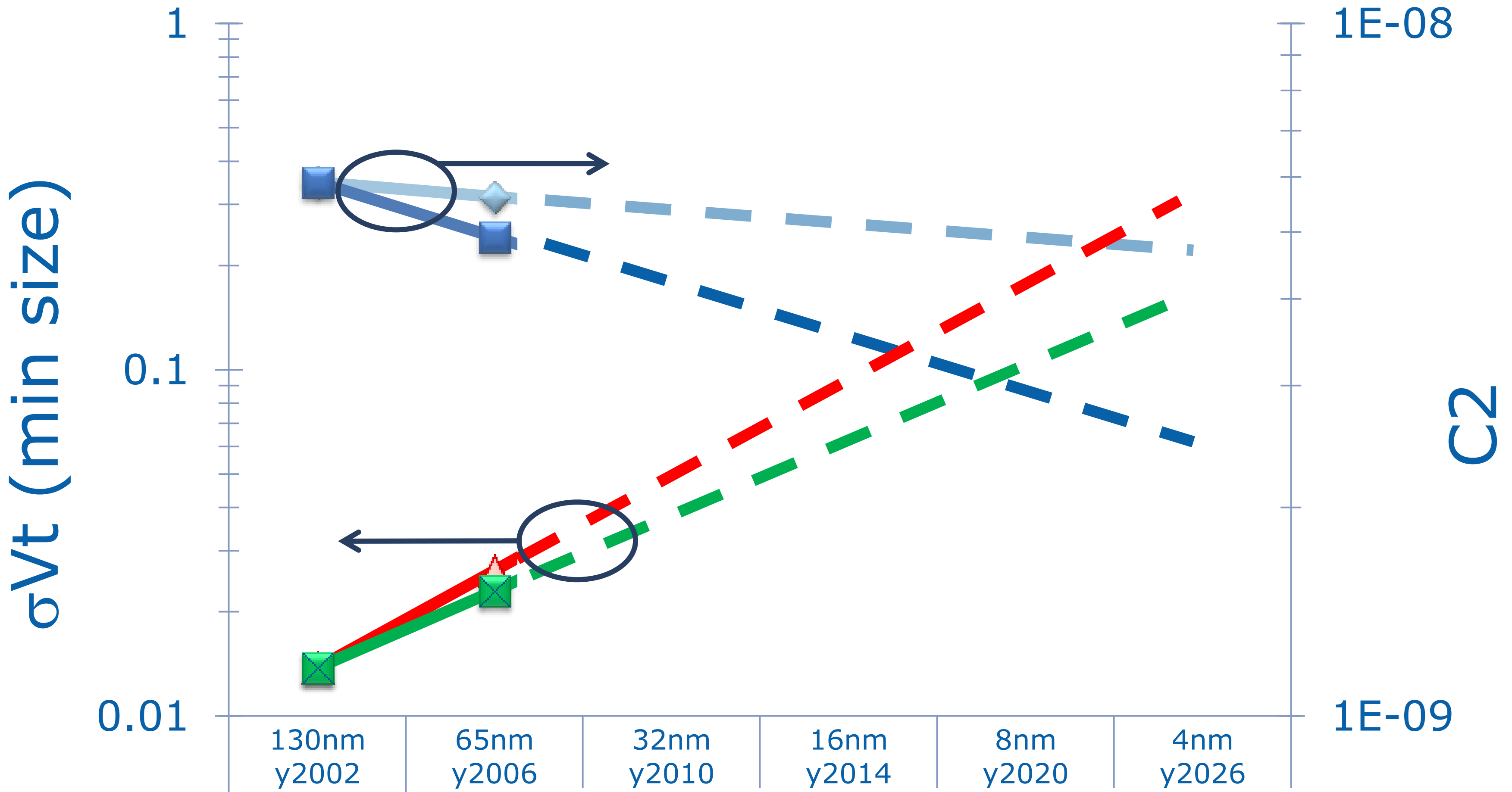
1. Modest data rates
2. Forwarded clocking
3. Global circuit sharing
4. Low power clock distribution
5. Resonantly tuned clocking
6. Low swing TX
7. Sensitive RX
8. Simple equalization
9. Calibration and tuning
10. System modeling

$$\sigma V_T = \frac{1}{\sqrt{2}}\left(\frac{c_2}{\sqrt{Weff \cdot Leff}}\right)$$



K. Kuhn, IEDM 2007

Extrapolated Process Variation



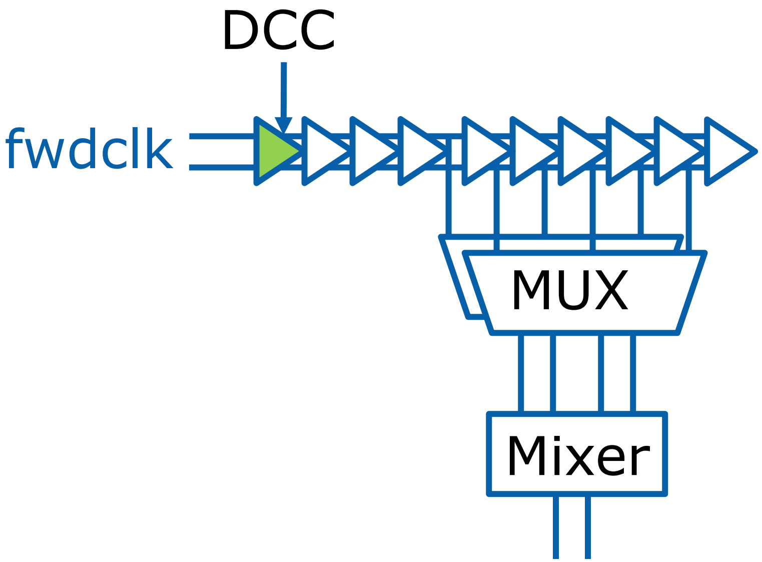
	130nm y2002	65nm y2006	32nm y2010	16nm y2014	8nm y2020	4nm y2026
C2_hi	5.9E-09	5.6E-09				
C2_lo	5.9E-09	4.9E-09				
std(Vt_hi)	1.4E-02	2.6E-02				
std(Vt_lo)	1.4E-02	2.3E-02				

Area & Energy scaling limited by variation

$\sigma(V_t) = C2 \div \sqrt{A_{gate}}$

Example: Phase Rotator

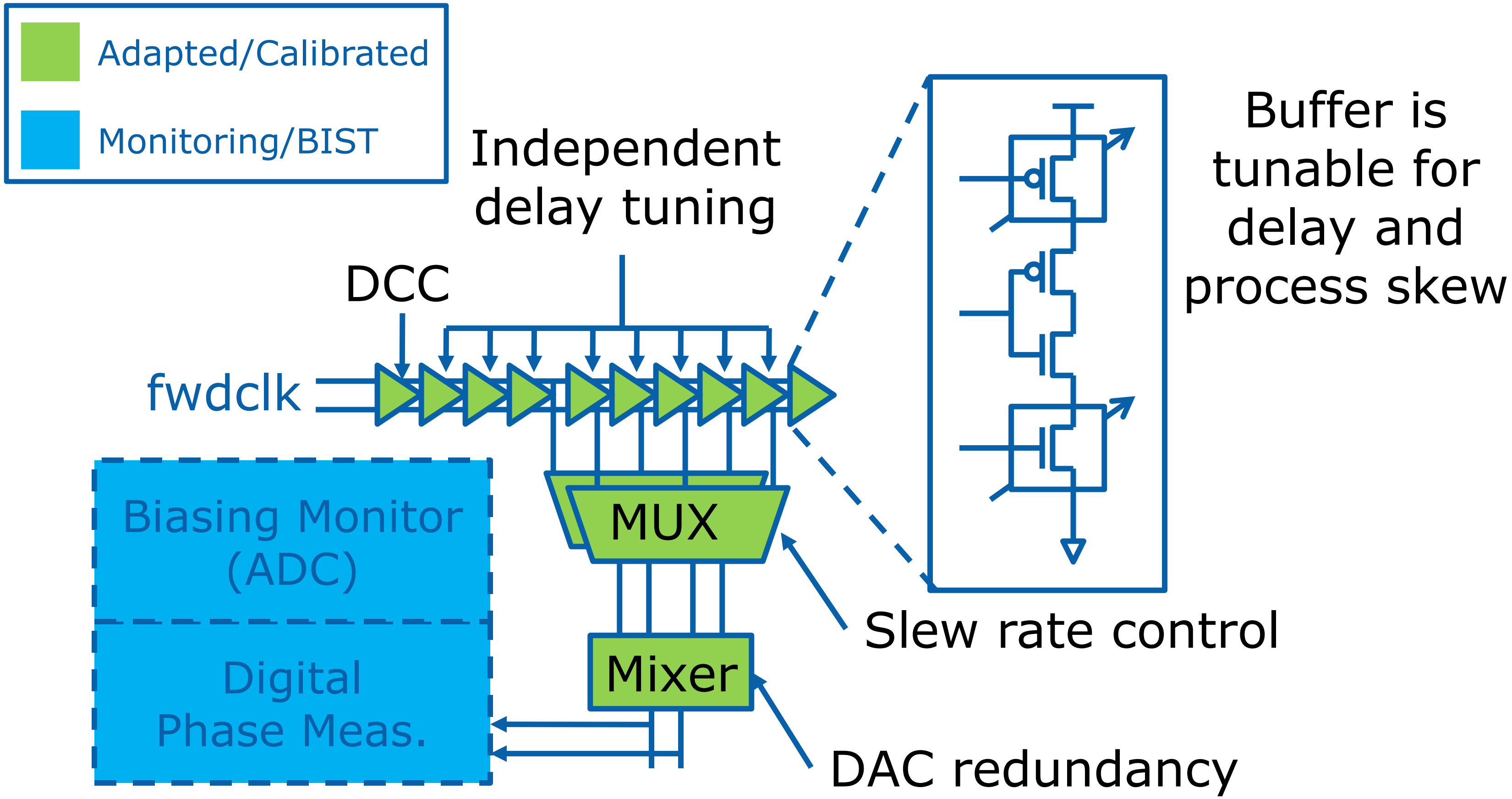
🟩 Adapted/Calibrated



Most calibration and adaptation used today is fairly basic

- e.g. duty cycle correction

Example: Programmable Phase Rotator



Independent delay tuning

Buffer is tunable for delay and process skew

Slew rate control

DAC redundancy

- 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

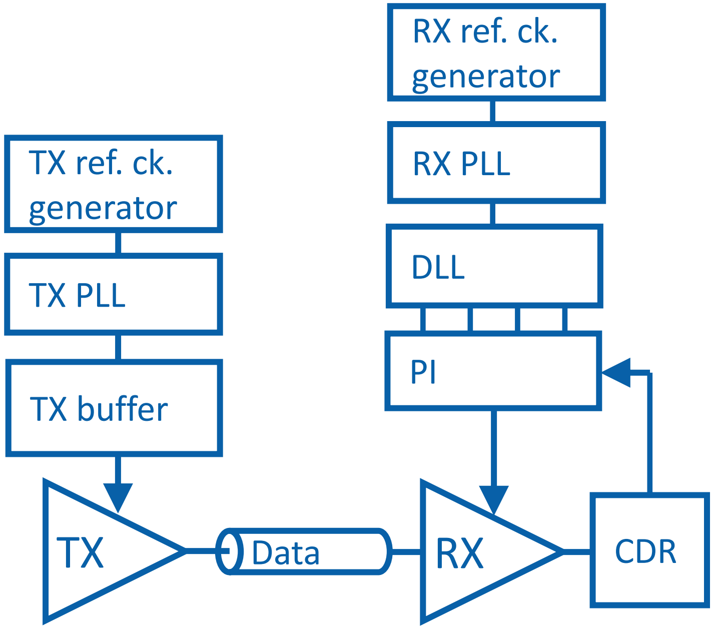
- Key to low power is balanced implementation
  - Achieved through comprehensive understanding of power/performance tradeoffs
- Focus design effort and power on highest impact components
- System-level optimization most impactful
  - Most will not have this opportunity due to standardization specs.
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



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

Knowledge of system performance and power sensitivities enables global power optimization

B. Casper, F. O'Mahony, "Clocking Analysis, Implementation and Measurement Techniques for High-Speed Data Links—A Tutorial," TCAS1, Jan. 2009

Agenda

- Introduction
- Impact of process scaling
- Active power optimization
  - System
  - Circuit

➡ Power management

- Low power silver bullets
- Putting it all together

Server Utilization

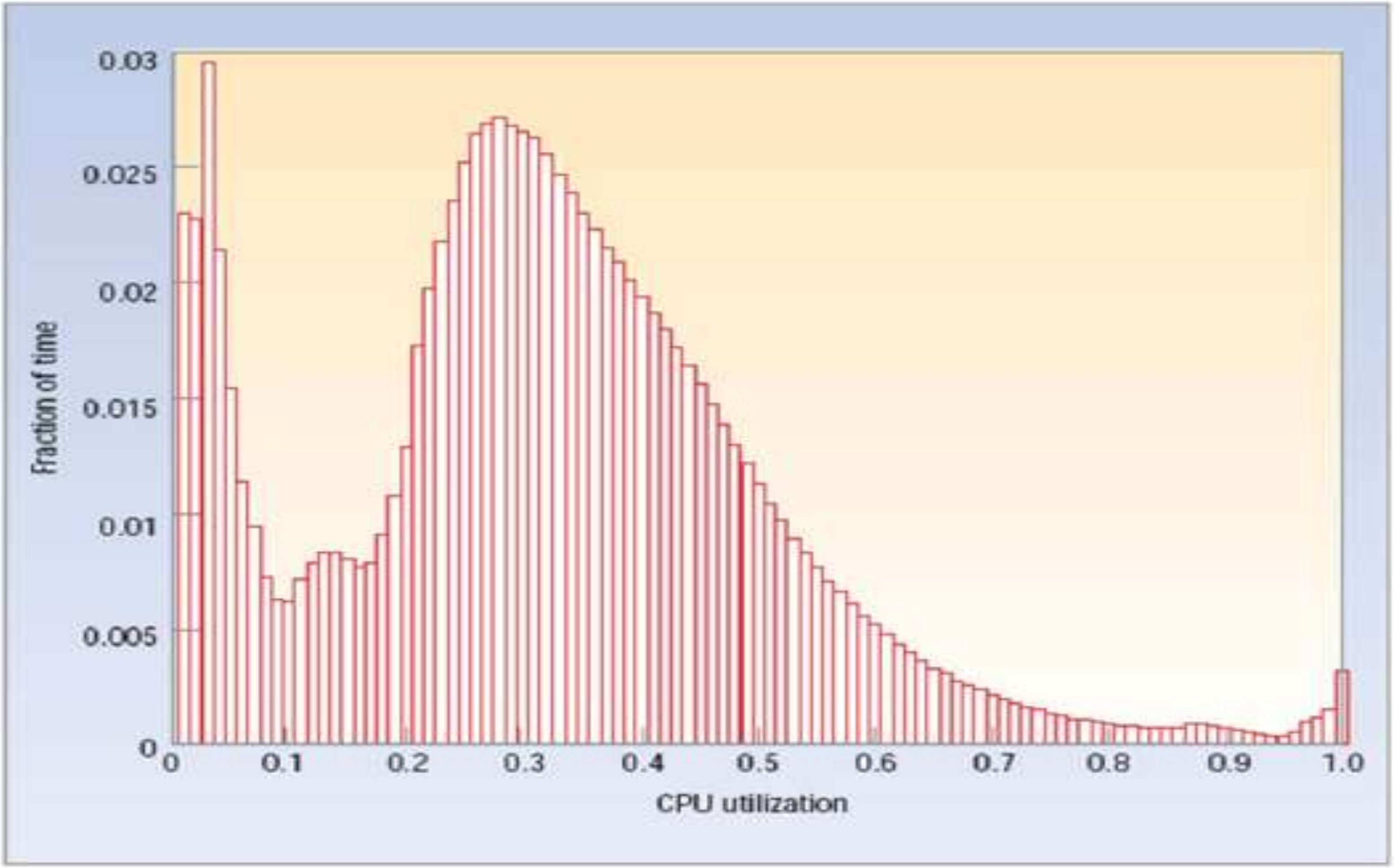
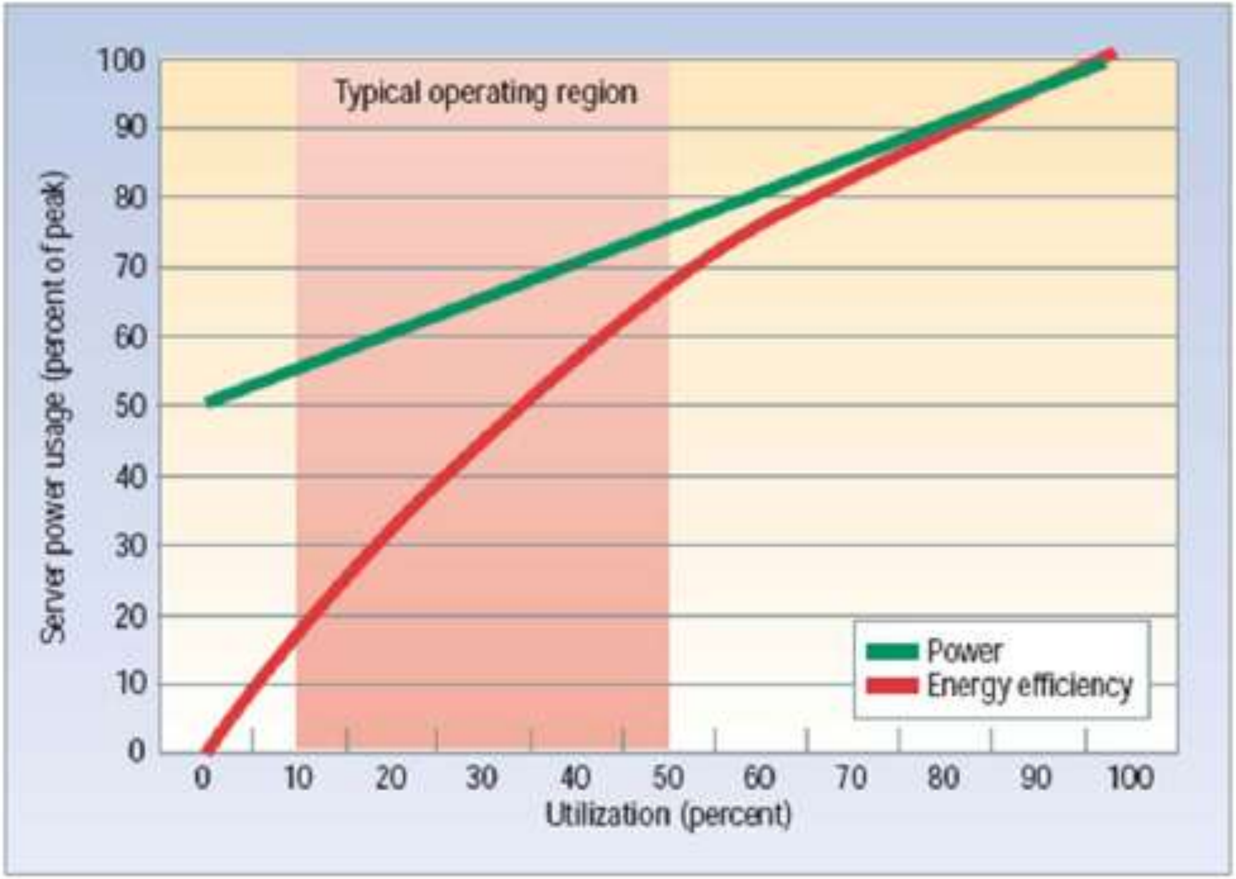


Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period.

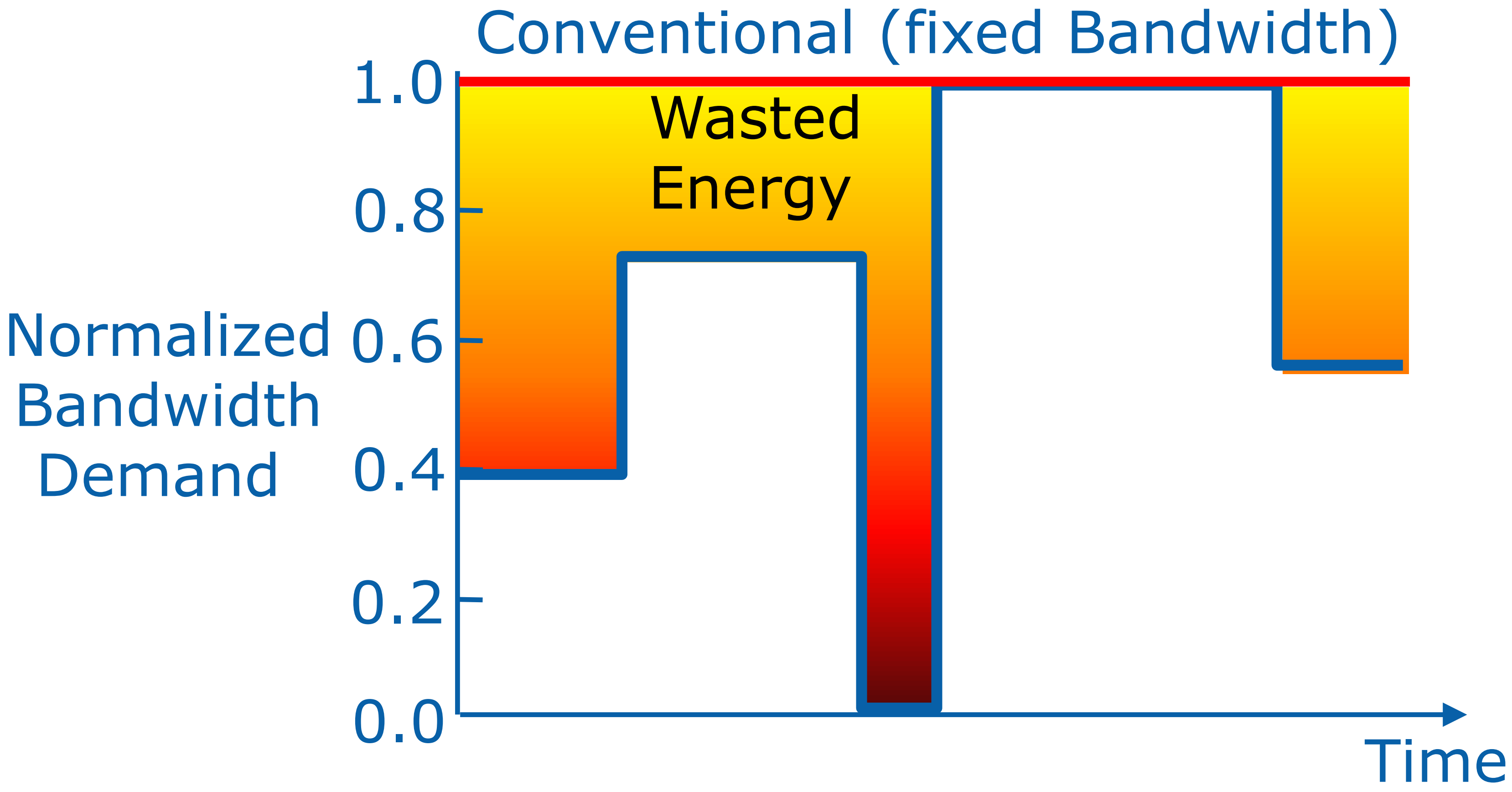
Barroso & Holzle, IEEE Computer, Dec 2007

Energy-Disproportionate Link



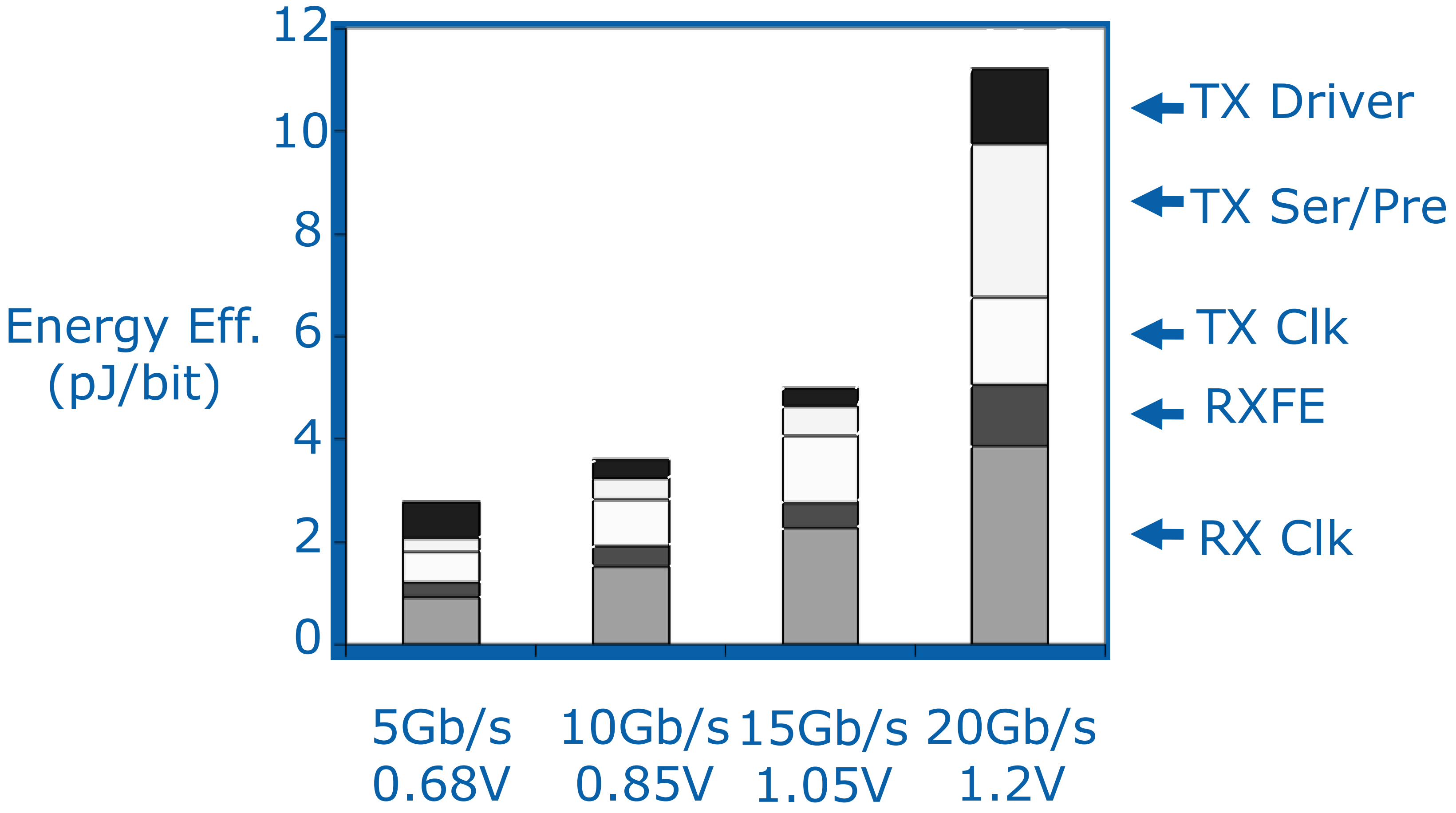
Barroso & Holzle, IEEE Computer, Dec 2007

Energy-proportional I/O



65

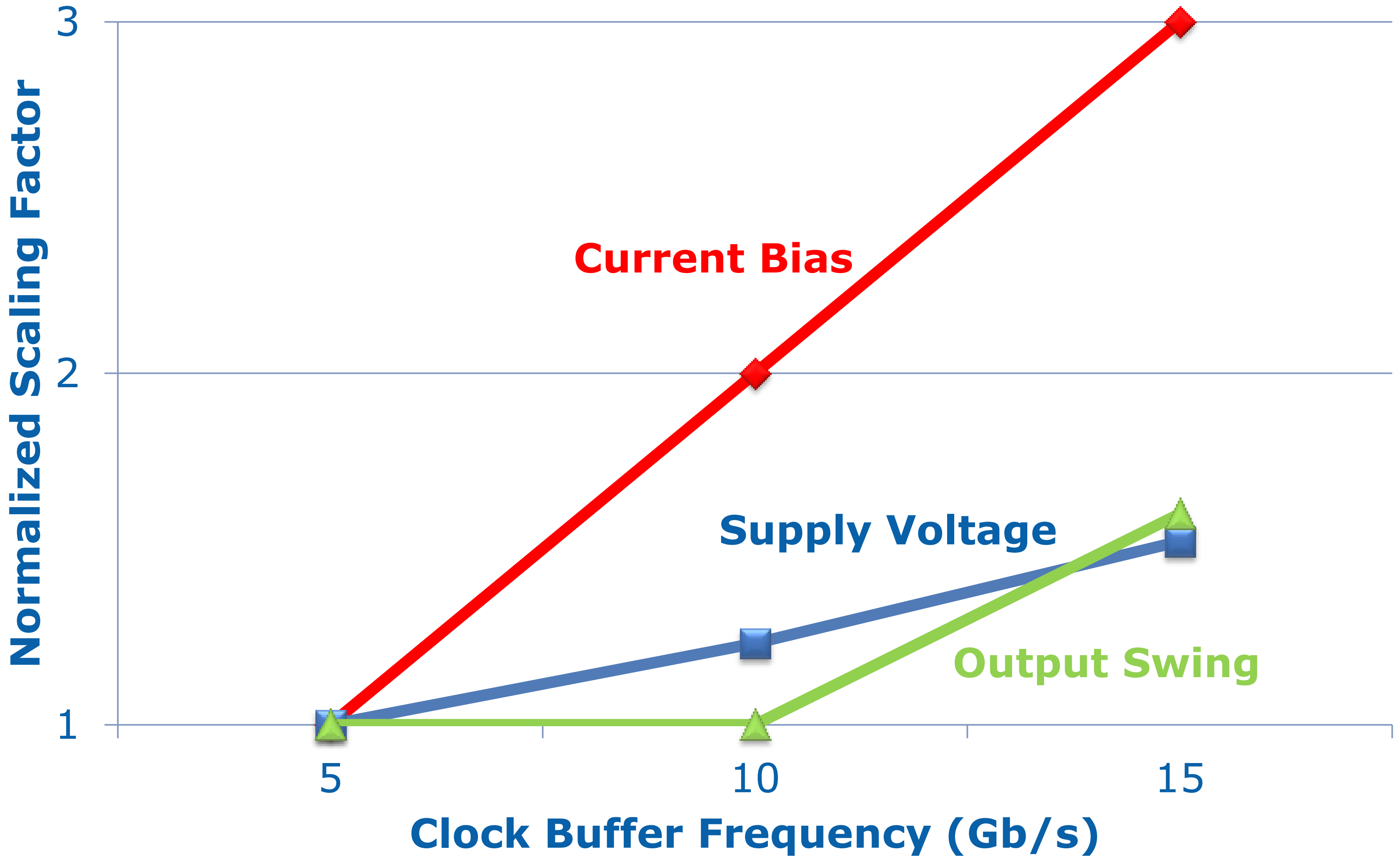
Power Management: Scalable supplies



Power efficiency improves with adaptive supply/biasing

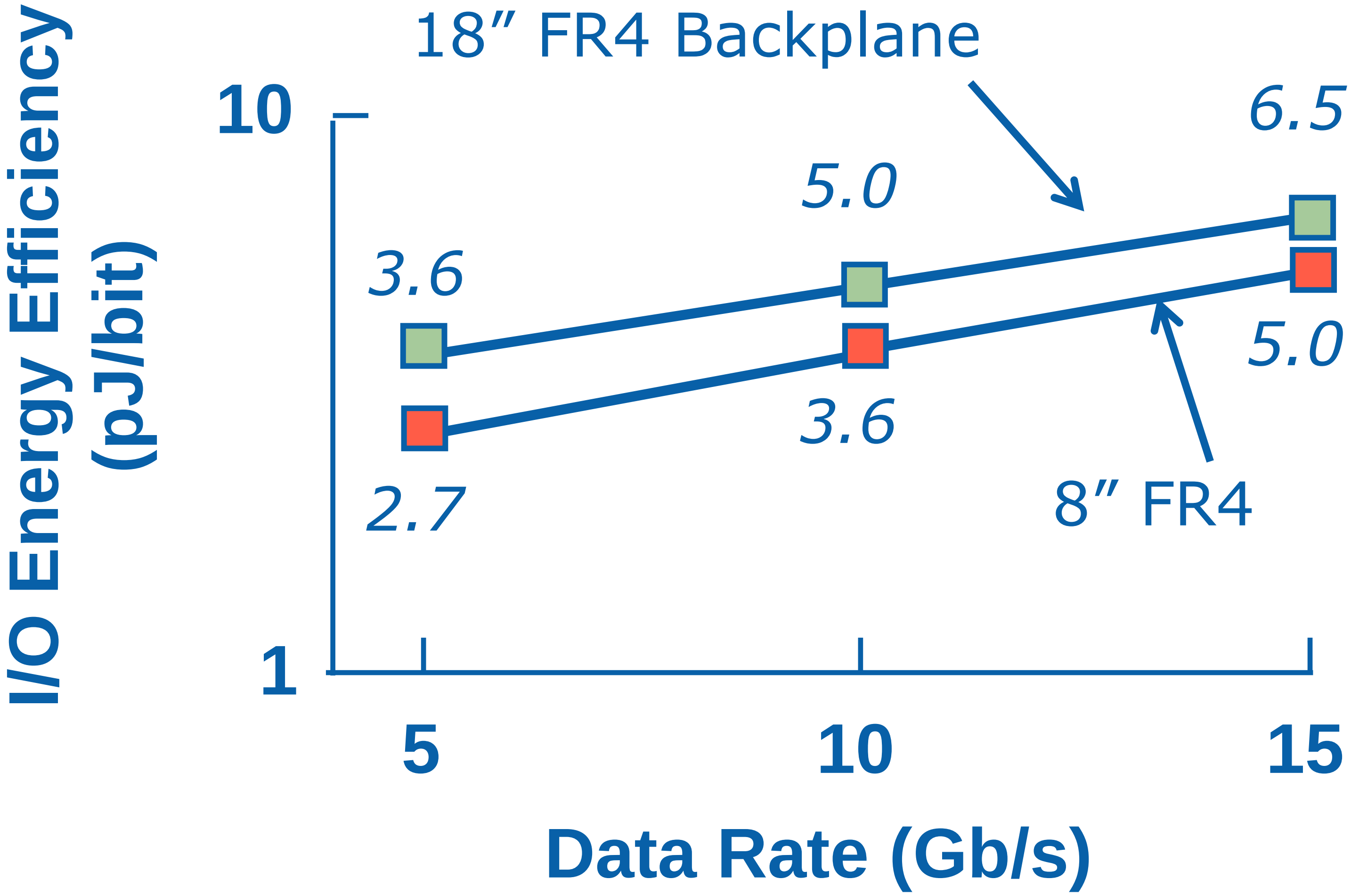
Refs: B. Casper, ISSCC '06 & G. Balamurugan, JSSC 4/08

65nm Low Power Link Operating Points



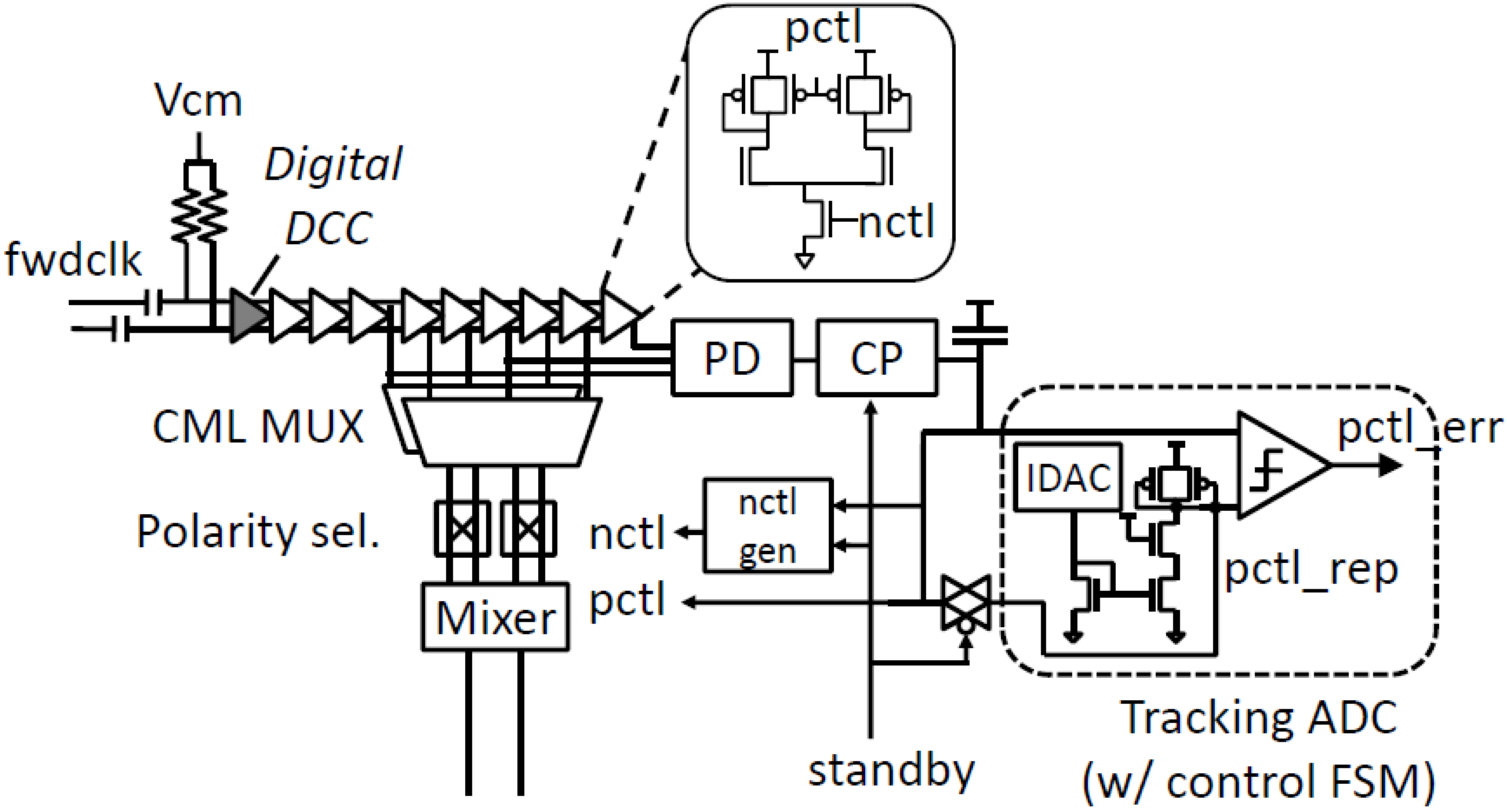
Balamurugan JSSC, April 2008

Benefit of Eliminating Excess BW: Non-linear Efficiency/Performance



Balamurugan JSSC, April 2008

Fast Wake-Up Clocking



O'Mahony et al, "A 47x10Gb/s 1.4mW/(Gb/s) Parallel Interface in 45nm CMOS," ISSCC, Feb. 2010.

Agenda

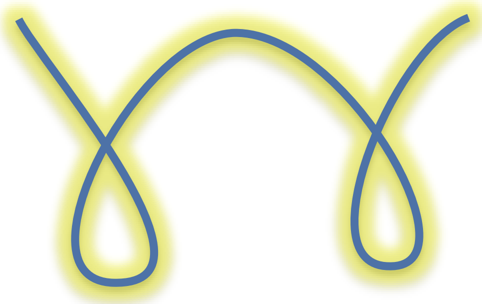
- Introduction
- Impact of process scaling
- Active power optimization
  - System
  - Circuit
- Power management

➡ Low power silver bullets (?)

- Putting it all together

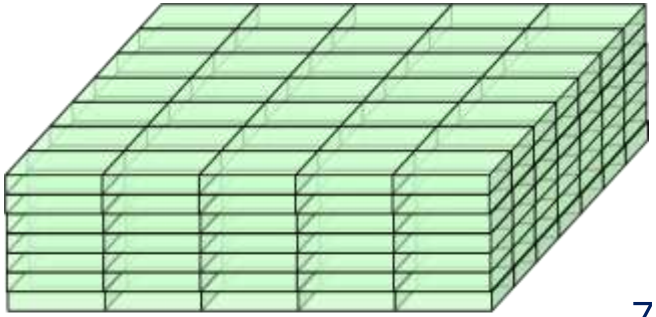
Low Power Link "Silver Bullets"?

Optical



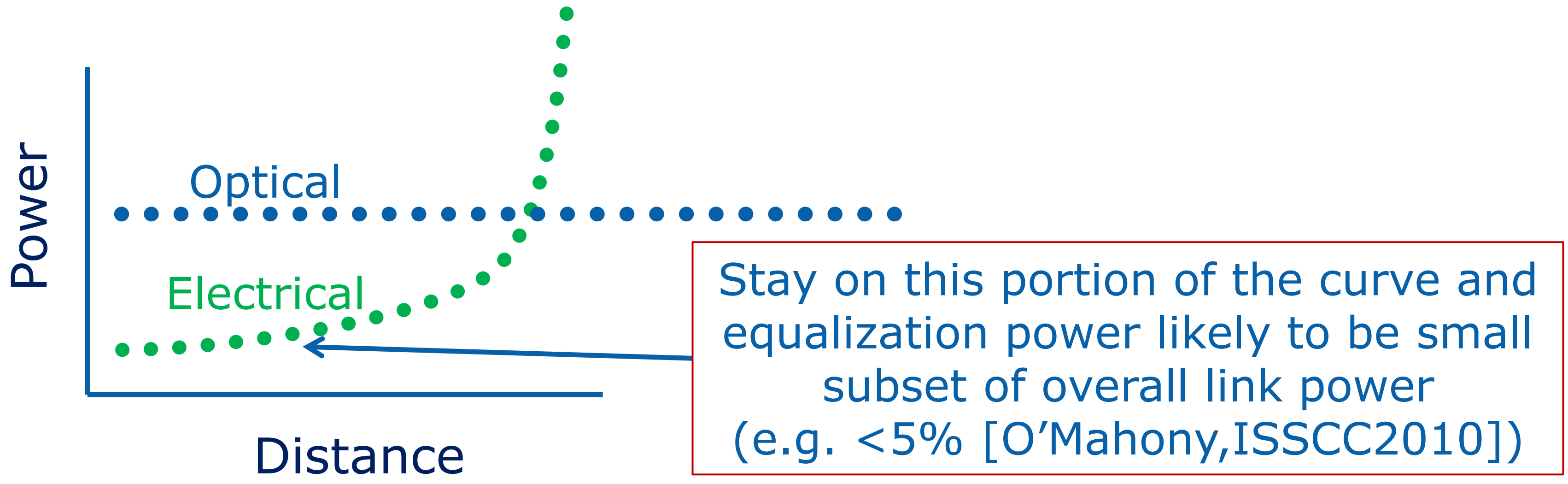
Modulation (PAM)

3D Stacking



Silver Bullet?: Optical

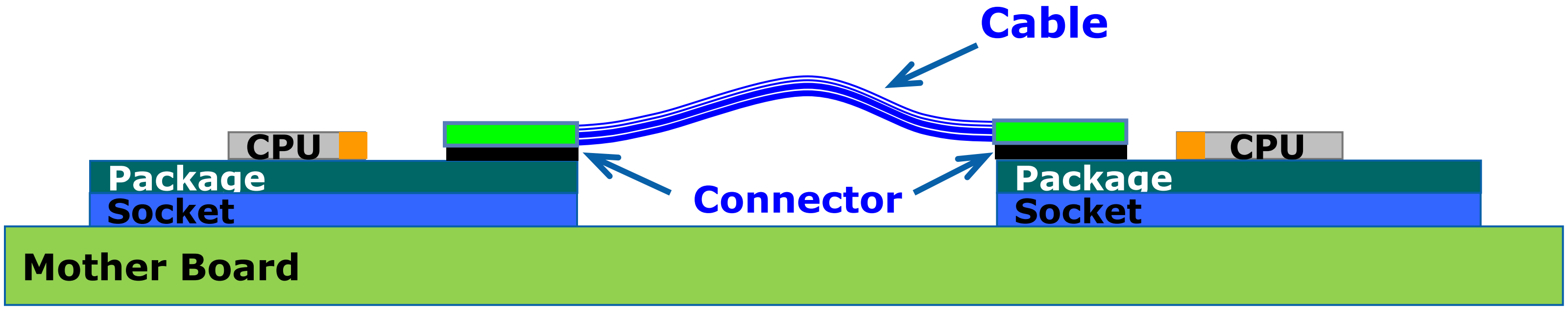
- Claim:
  - Optical power is inherently better because no equalization needed
- Reality:
  - Most optical power claims only include optical components
    - Disregards electrical driver, clocking, recovery, synch., serdes, etc.
  - Optical link = Electrical link + optical components in the middle
  - Optical/electrical power crossover is likely 1m-5m, depending on rate



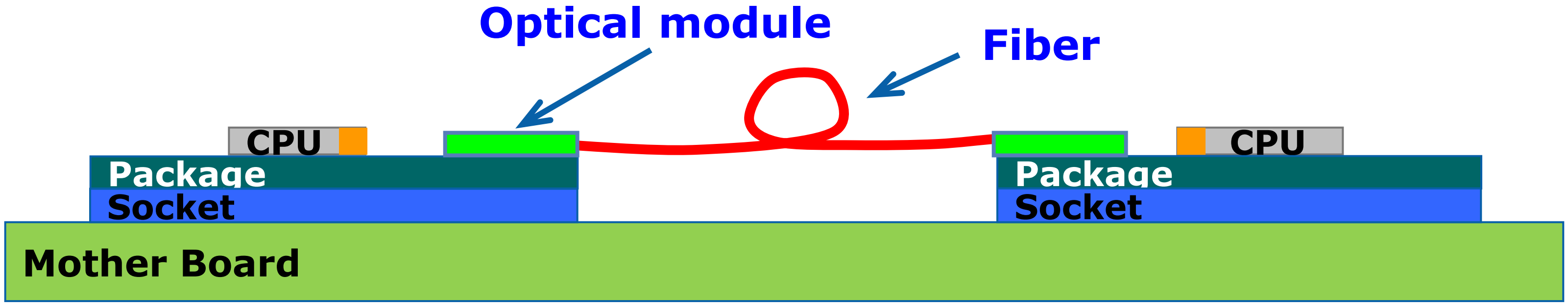
Stay on this portion of the curve and equalization power likely to be small subset of overall link power (e.g. <5% [O’Mahony,ISSCC2010])

Apples-Apples Electrical/Optical Channel Comparison

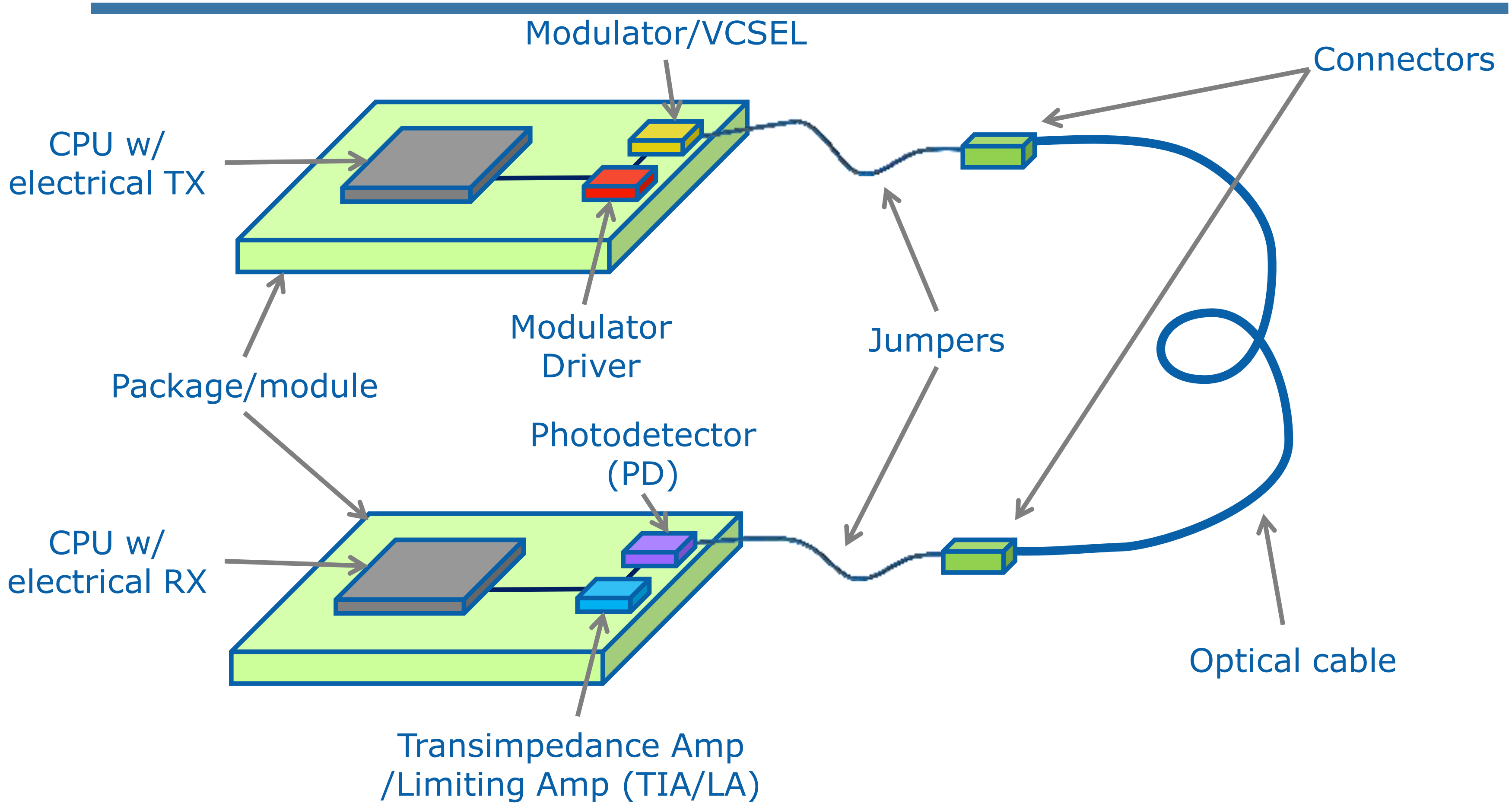
- Electrical



- Optical



Complete Optical CPU Link



Modulator/VCSEL

Connectors

CPU w/
electrical TX

Modulator
Driver

Jumpers

Package/module

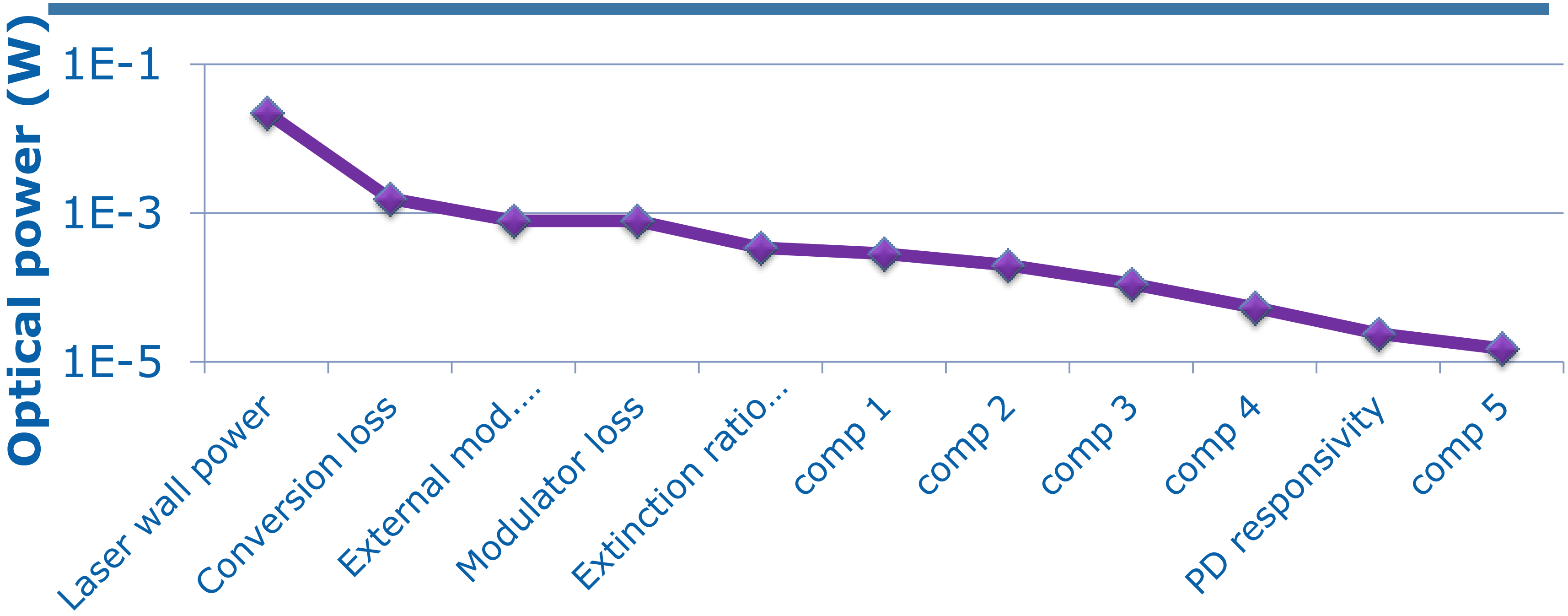
Photodetector
(PD)

CPU w/
electrical RX

Optical cable

Transimpedance Amp
/Limiting Amp (TIA/LA)

Example Optical Loss Profile (consumer electronics)

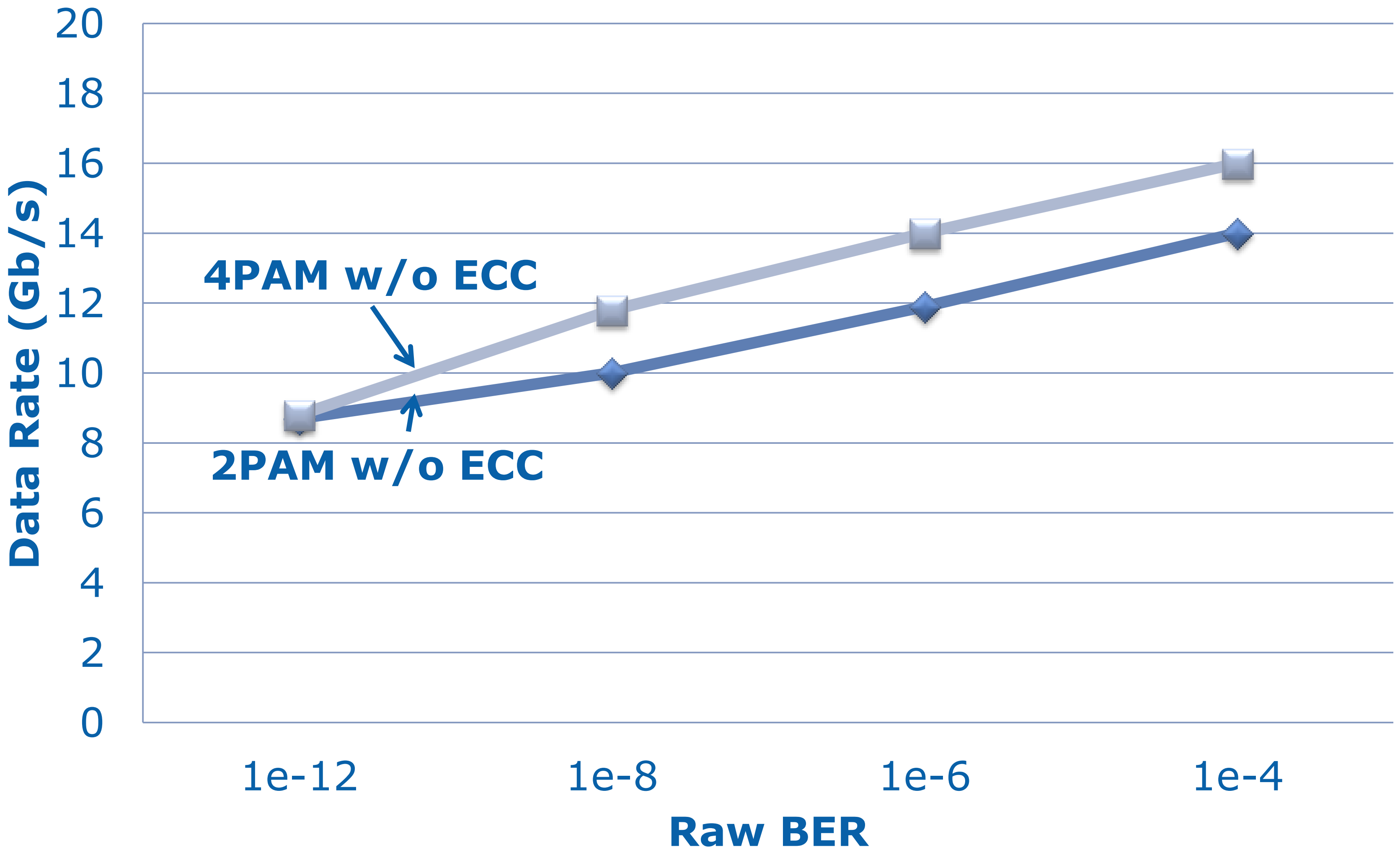


- Optical channel loss is frequency independent
  - But the aggregate loss is 100x-1000x!
- VCSEL or MZI based links require large swings (500mV-1V)
- Worst-case received signal can be as low as ~10uA
  - Requires extremely sensitive receiver (costs power)
- Full optical link >2x power of electrical at ≤3m distance

Silver Bullet?: PAM

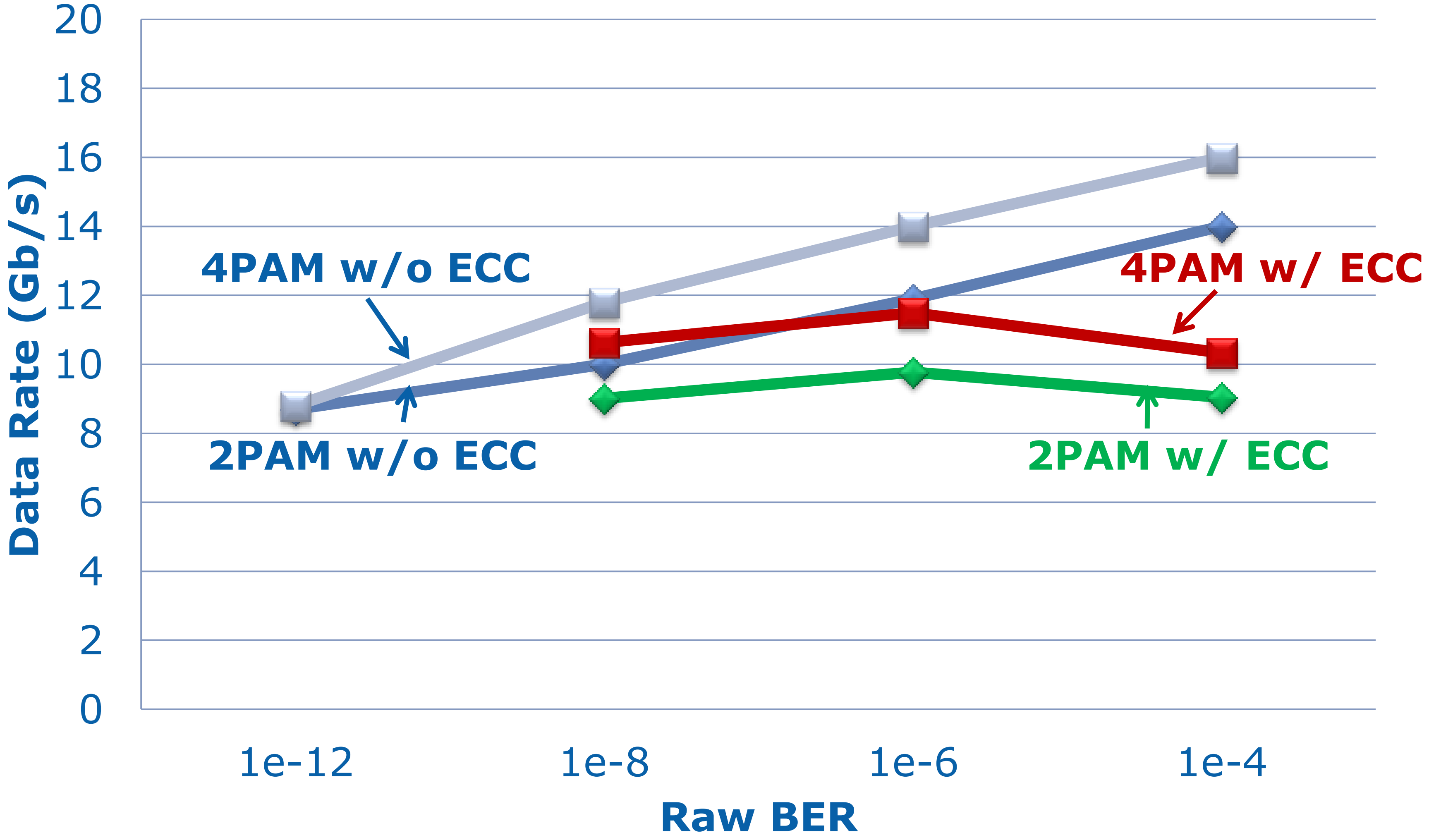
- Claim:
  - PAM uses less BW resulting in less equalization and lower power
- Reality:
  - No inherent performance/power advantage over binary
    - Using practical channels and 1E-12 BER
  - Equalization and clock recovery more difficult
  - PAM receiver more complex
    - 4 PAM requires 1.5 samples/bit + 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 1e-12 BER (LE & DFE 4-tap)



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

Max data rate with small block coding to achieve 1e-12 BER (LE & DFE 4-tap)



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

45nm PAM Measurements (Within-package channel)

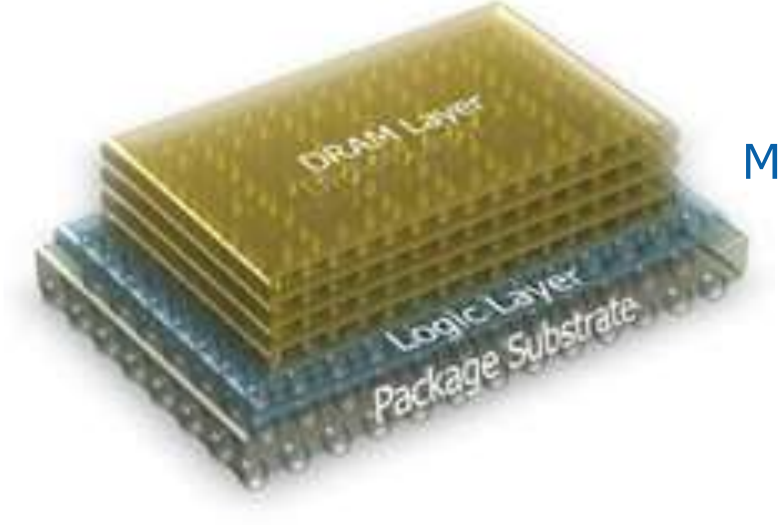
Channel	Signaling Mode	Efficiency	Data rate	TX swing
MCP	2-PAM	2.3pJ/bit	12.5G	120mV
	3-PAM	2.6pJ/bit	18.75G	260mV
	4-PAM	2.6pJ/bit	25G	360mV

- 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

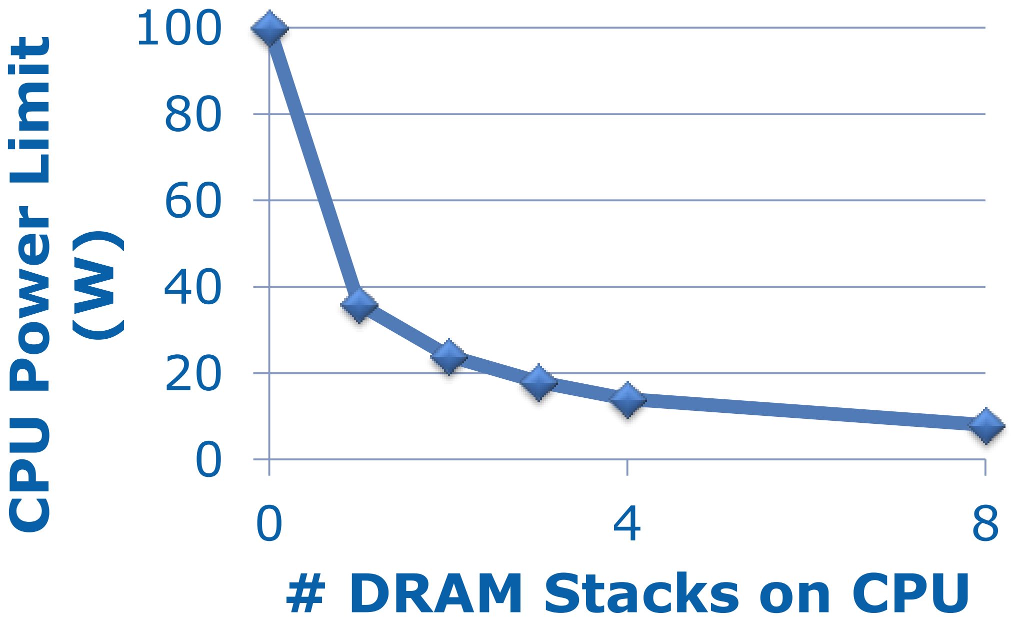
- Claim:
  - Stacking minimizes need for high-speed I/O
- Reality:
  - Potential to reduce I/O power (within stack) by 10x-100x
    - Reduce C & V (CV^2)
  - Components within stack must be tightly integrated (architecture, process, mechanicals)
  - Thermal and power delivery limits applicability
    - Primarily applicable for low power stacks



Micron Hybrid Memory Cube

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

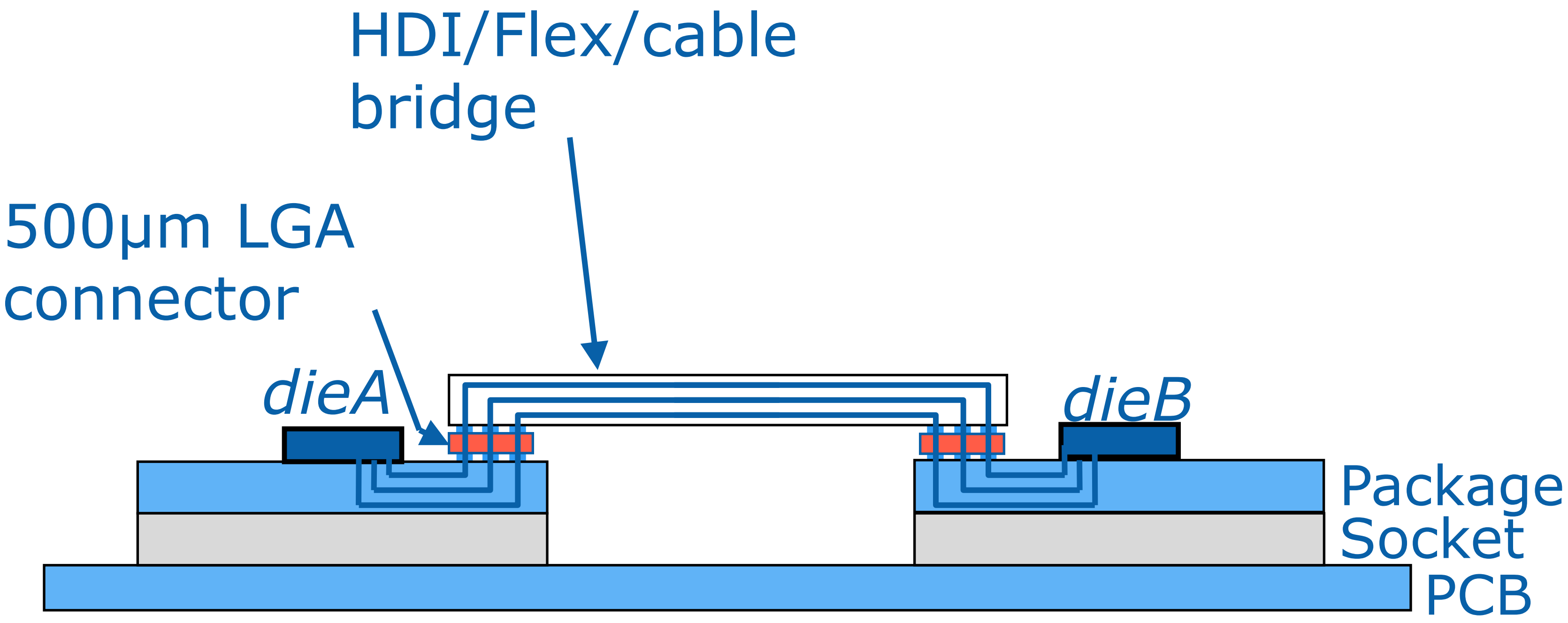
Each technology may have power advantages for a limited set of applications. However, not a general solution to solving the Link Power Problem.

Agenda

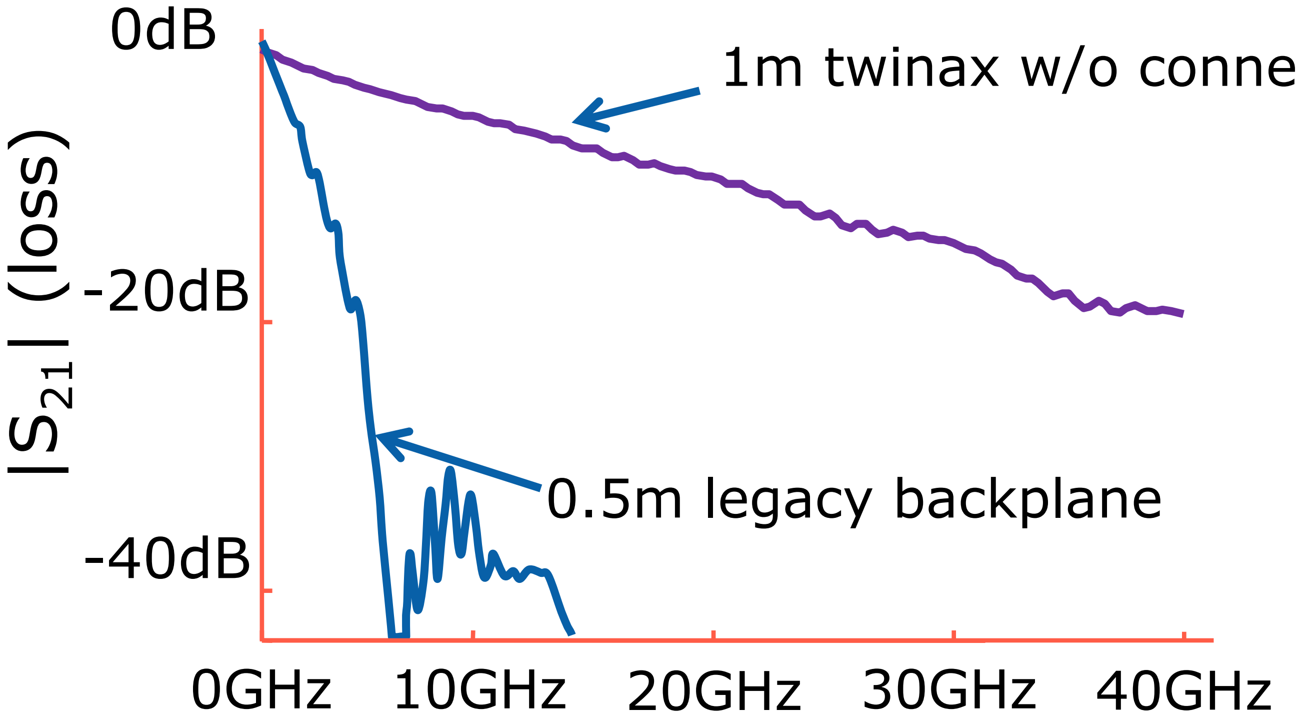
- Introduction
- Impact of process scaling
- Active power optimization
  - System
  - Circuit
- Power management
- Low power silver bullets

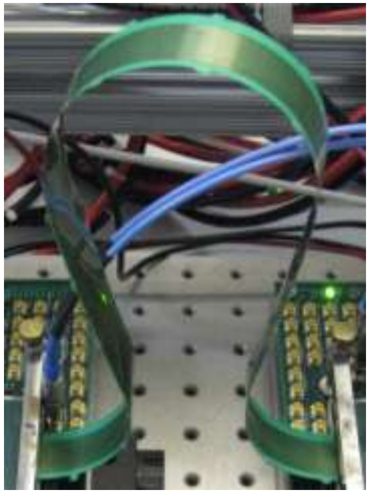
➡ Putting it all together

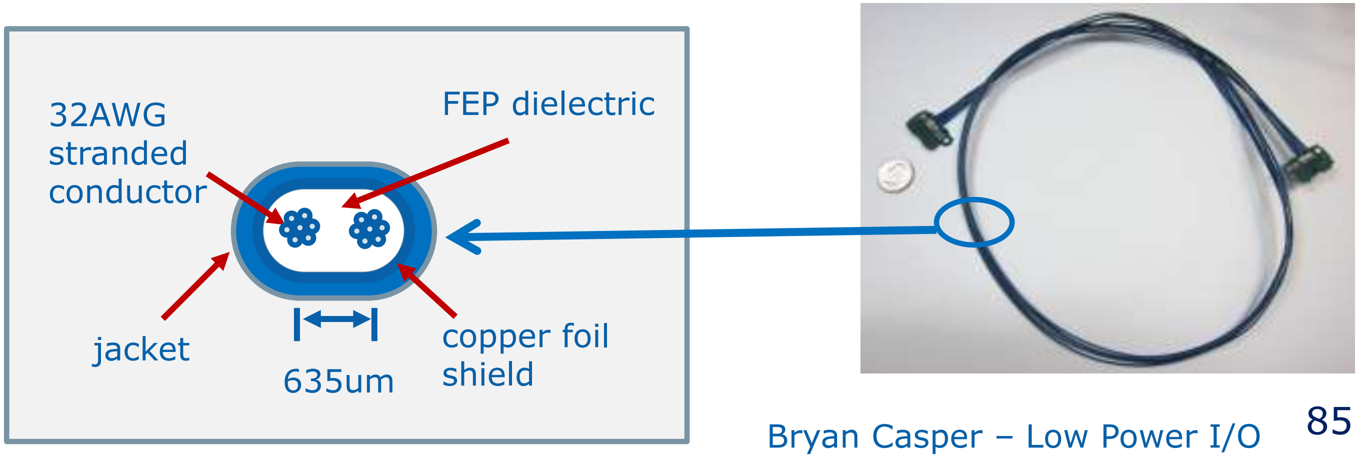
Example: 47x10Gb/s Interface



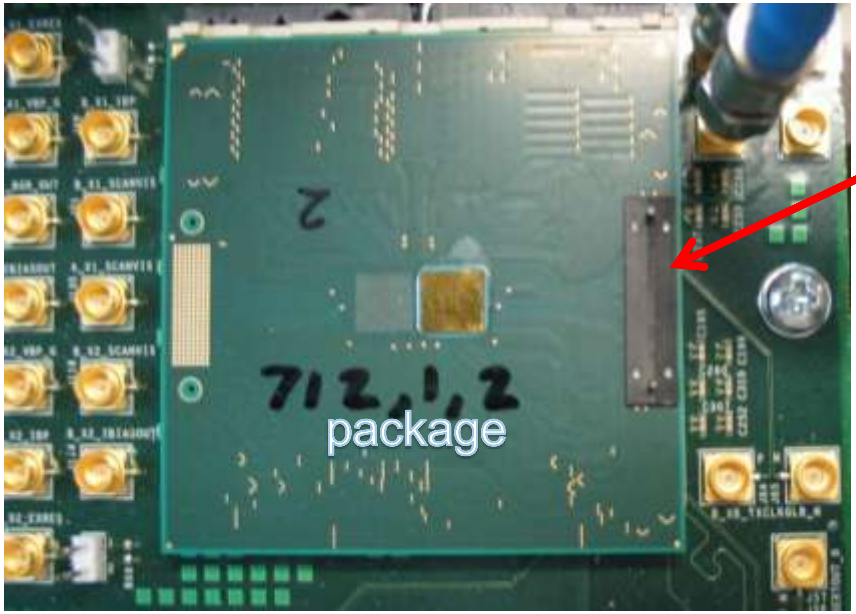
Solutions: Interconnect







Solutions: Connectors



Top-pkg connector
(4 signals/mm^2)

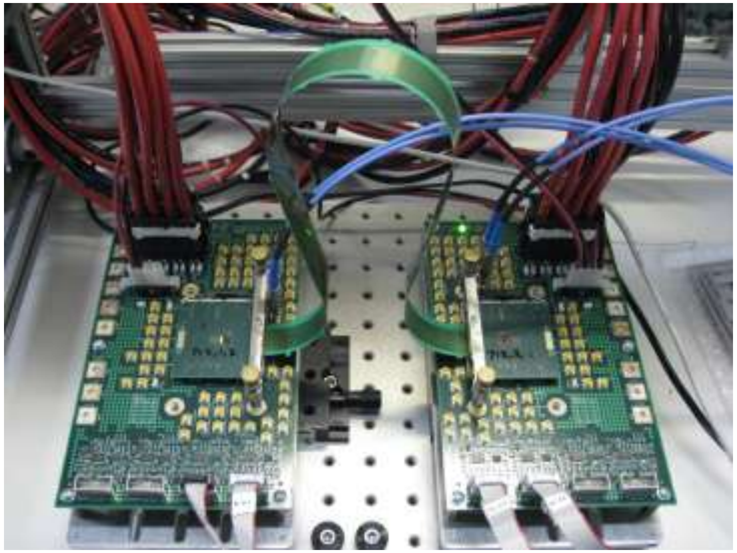
Solution: Circuits

- Utilized most suggested low power optimizations

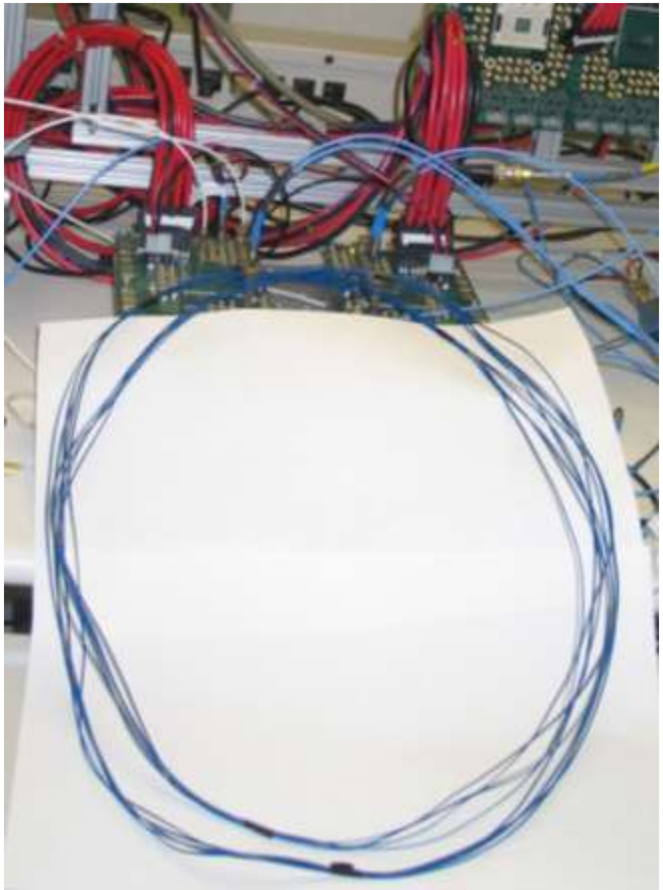
➡ 1. Modest data rates
➡ 2. Forwarded clocking
➡ 3. Global circuit sharing
➡ 4. Low power clock distribution
5. Resonantly tuned clocking
➡ 6. Low swing TX
➡ 7. Sensitive RX
➡ 8. Simple equalization
➡ 9. Calibration and tuning
➡ 10. System modeling

Low Power Prototype Results

0.5m flex interconnect

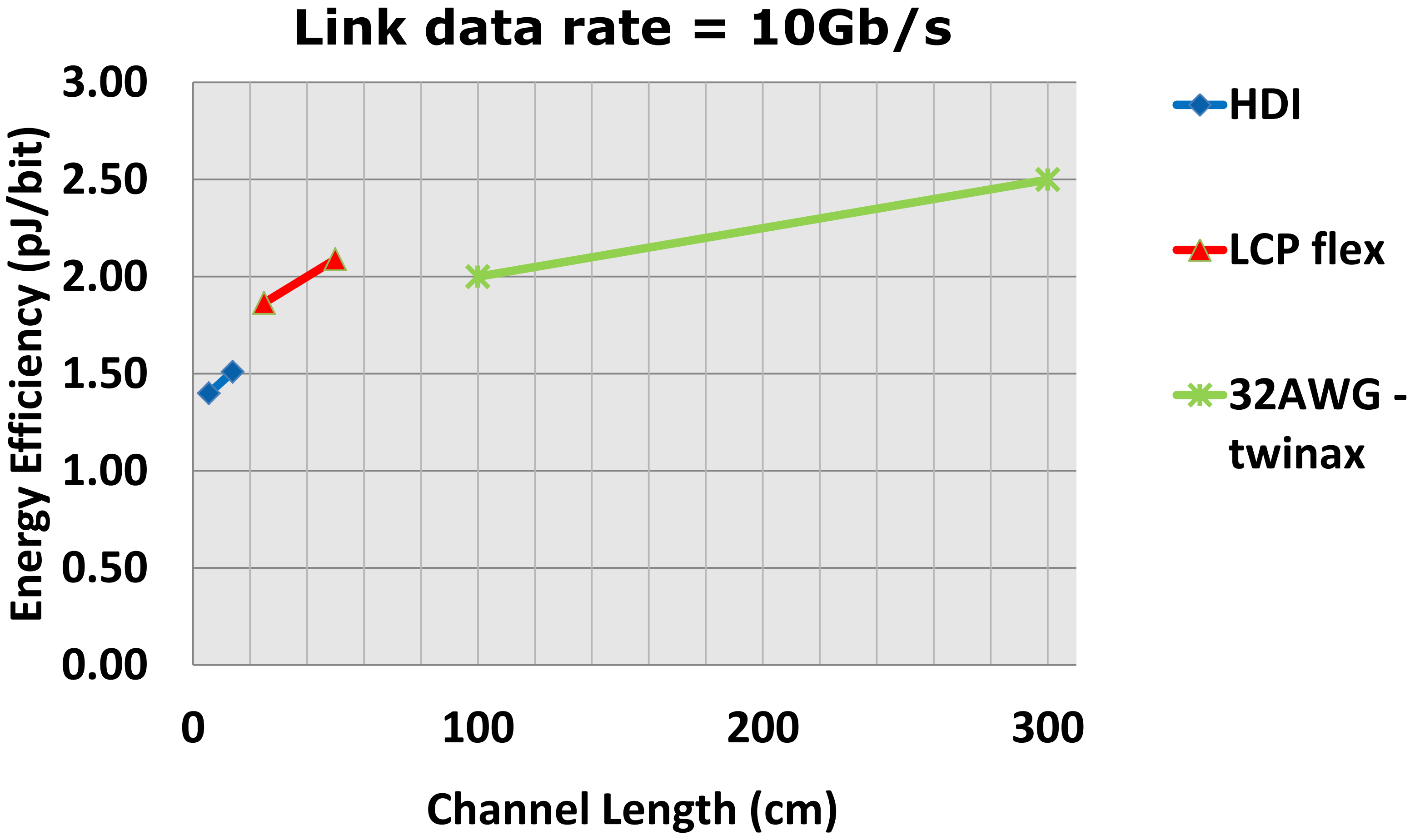


3m twinax cable



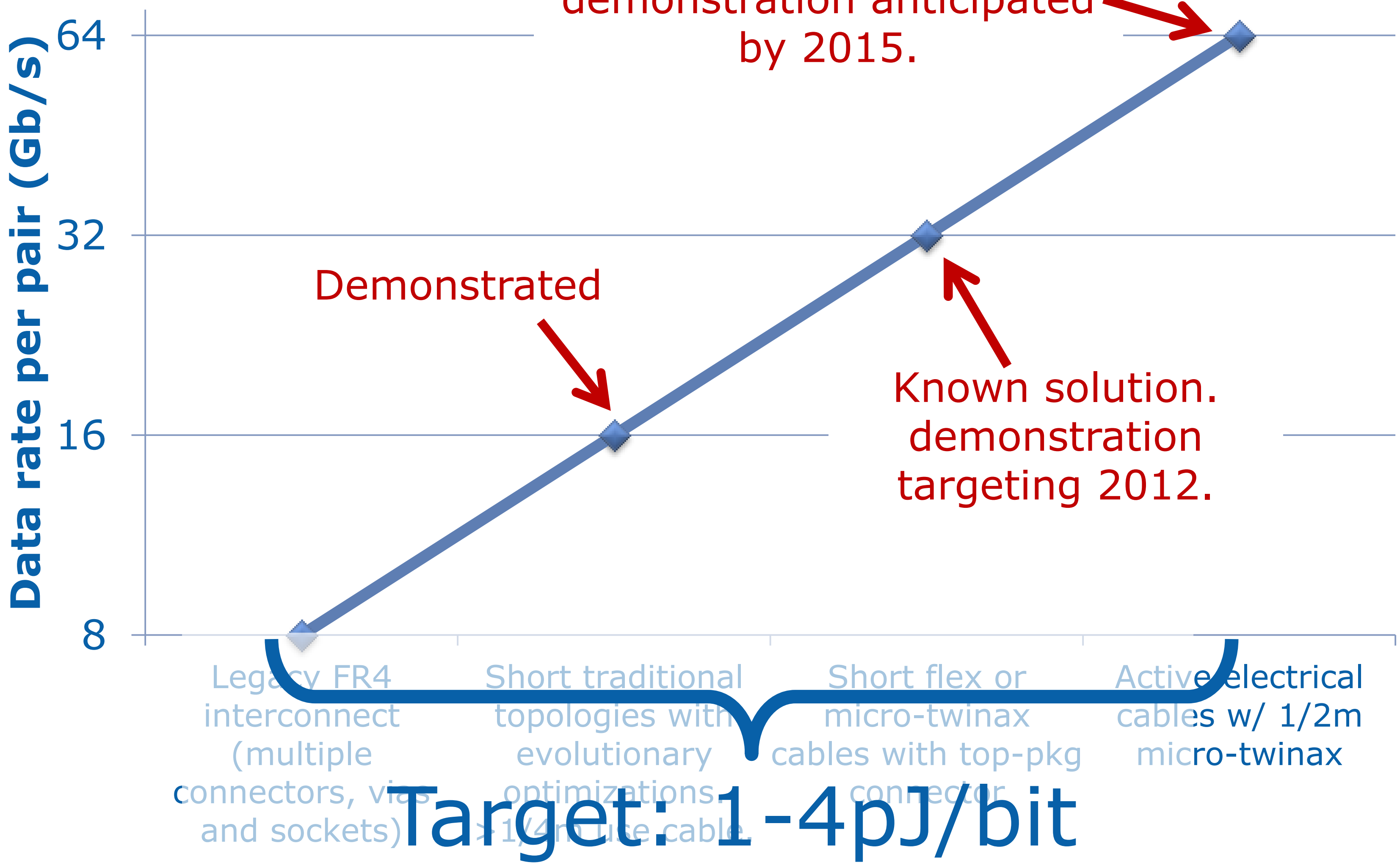
45nm CMOS prototype demonstrates industry leading I/O power efficiency with non-traditional interconnects

Low Power Prototype Results



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.

Target: 1-4pJ/bit

(vs. current product=20-40pJ/bit)

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

F. O’Mahony, et al., “A 47×10Gb/s 1.4mW/(Gb/s) parallel interface in 45nm CMOS,” *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, Feb. 2010, pp. 156–157.

F. O’Mahony, et al., “The future of electrical I/O for microprocessors,” *International Symposium on VLSI DAT*, Apr. 2009, pp. 31-34.

G. Balamurugan, et al., “A scalable 5–15Gbps, 14–75mW low-power I/O transceiver in 65 nm CMOS,” *IEEE J. Solid-State Circuits*, vol. 43, pp. 1010-1019, Apr. 2008

H. Braunisch, et al., “High-speed flex circuit chip-to-chip interconnects,” *IEEE Trans. On Advanced Packaging*, vol. 31, no. 1, 2008, pp. 82-90.

B. Casper, et al., “Future microprocessor interfaces: analysis, design and optimization,” *IEEE Custom Integrated Circuits Conference*, Sept. 2007, pp. 479 – 486.

Related Publications 2

- J. Poulton, R. Palmer, A. M. Fuller, T. Greer, J. Eyles, W. J. Dally, and M. Horowitz, "A 14-mW 6.25-Gb/s transceiver in 90-nm CMOS," *IEEE J. Solid-State Circuits*, vol. 42, pp. 2745-2757, Dec. 2007.
- H. Hatamkhani, F. Lambrecht, V. Stojanovic, and C.-K. K. Yang, "Power-centric design of high-speed I/Os," in *Proc. Design Automation Conf.*, 2006, pp. 867-872.
- K.-L. J. Wong, H. Hatamkhani, M. Mansuri, and C.-K. K. Yang, "A 27-mW 3.6-Gb/s I/O transceiver", *IEEE J. Solid-State Circuits*, vol. 39, pp. 602-612 , Dec. 2004.
- G. Balamurugan, J. Kennedy, G. Banerjee, J. E. Jaussi, M. Mansuri, F. O'Mahony, B. Casper, and R. Mooney, "A scalable 5–15 Gbps, 14–75 mW low-power I/O transceiver in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 43, pp. 1010-1019, Apr. 2008.
- S. Joshi, J. T.-S. Liao, Y. Fan, S. Hyvonen, M. Nagarajan, J. Rizk, H.-J. Lee, and I. Young, "A 12-Gb/s transceiver in 32-nm bulk CMOS," in *2009 Symp. VLSI Circuits Dig. Tech. Papers*, pp. 52-53.
- B. Leibowitz, R. Palmer, J. Poulton, Y. Frans, S. Li, J. Wilson, M. Bucher, A. M. Fuller, J. Eyles, M. Aleksić, T. Greer, and N. M. Nguyen, "A 4.3 GB/s mobile memory interface with power-efficient bandwidth scaling," *IEEE J. Solid-State Circuits*, vol. 45, pp. 889-898, Apr. 2010.
- F. O'Mahony, M. Mansuri, B. Casper, J. E. Jaussi, and R. Mooney, "A low-jitter PLL and repeaterless clock distribution network for a 20Gb/s link", in *2006 Symp. VLSI Circuits Dig. Tech. Papers*, pp. 36-37.
- B. Casper, J. E. Jaussi, F. O'Mahony, M. Mansuri, K. Canagasaby, J. Kennedy, E. Yeung, and R. Mooney, "A 20Gb/s forwarded clock transceiver in 90nm CMOS," in *2006 IEEE ISSCC Dig. Tech. Papers*, pp. 90-91.
- J. Montanaro, R. T. Witek, K. Anne, A. J. Black,, E. M. Cooper, D. W. Dobberpuhl, P. H. Donahue, J. Eno, G. W. Hoeppner, D. Kruckemyer, T. H. Lee, P. C. M. Lin, L. Madden, D. Murray, M. H. Pearce, S. Santhanam, K. J. Snyder, R. Stephany, and S. C. Thierauf, "A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor," *IEEE J. Solid-State Circuits*, vol. 31, pp. 1703–1714, Nov. 1996.

Related Publications 3

- Intel® Core™ i5-670 Processor http://ark.intel.com/Product.aspx?id=43556
- Intel® Xeon® Processor X5670 http://ark.intel.com/Product.aspx?id=47920
- Intel® Xeon® Processor X7560 http://ark.intel.com/Product.aspx?id=46499
- O'Mahony, F.; Kennedy, J.; Jaussi, J.E.; Balamurugan, G.; Mansuri, M.; Roberts, C.; Shekhar, S.; Mooney, R.; Casper, B.; , "A 47×10Gb/s 1.4mW/(Gb/s) parallel interface in 45nm CMOS," *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International* , vol., no., pp.156-157, 7-11 Feb. 2010
- Horowitz, M.; Alon, E.; Patil, D.; Naffziger, S.; Rajesh Kumar; Bernstein, K.; , "Scaling, power, and the future of CMOS," *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International* , vol., no., pp.7 pp.-15, 5-5 Dec. 2005
- Casper, B.; Balamurugan, G.; Jaussi, J.E.; Kennedy, J.; Mansuri, M.; , "Future Microprocessor Interfaces: Analysis, Design and Optimization," *Custom Integrated Circuits Conference, 2007. CICC '07. IEEE* , vol., no., pp.479-486, 16-19 Sept. 2007
- B. Casper, F. O'Mahony, "Clocking Analysis, Implementation and Measurement Techniques for High-Speed Data Links—A Tutorial," TCAS1, Jan. 2009
- Casper, B.; Jaussi, J.; O'Mahony, F.; Mansuri, M.; Canagasaby, K.; Kennedy, J.; Yeung, E.; Mooney, R.; , "A 20Gb/s Forwarded Clock Transceiver in 90nm CMOS B.," *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International* , vol., no., pp. 263- 272, Feb. 6-9, 2006
- Kuhn, K.J.; , "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS," *Electron Devices Meeting, 2007. IEDM 2007. IEEE International* , vol., no., pp.471-474, 10-12 Dec. 2007
- F. O'Mahony, B. Casper, M. Mansuri, M. Hossain, “A Programmable Phase Rotator Based on Time-Modulated Injection-Locking,” VLSI symposium 2010