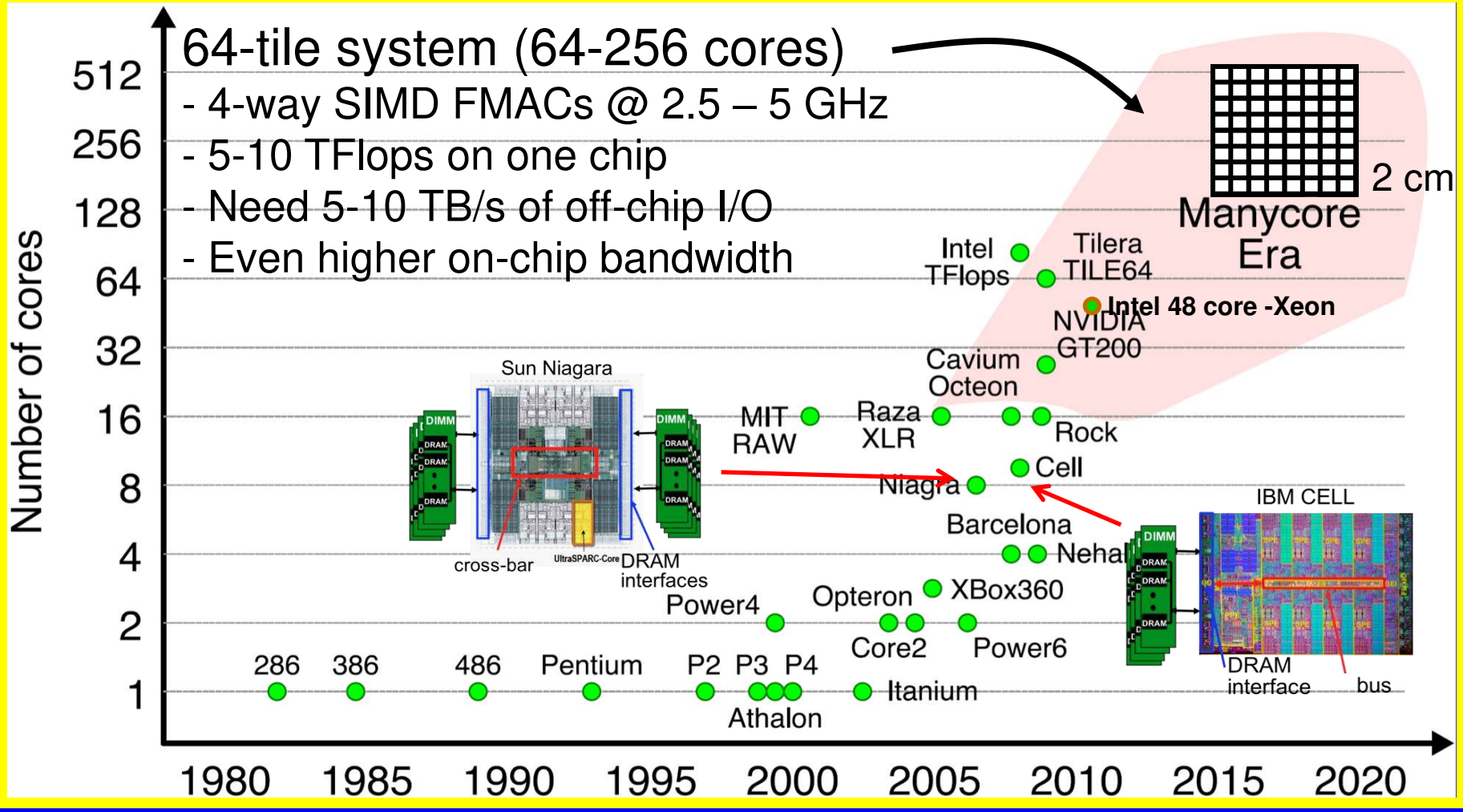
# Addressing Link-Level Design Tradeoffs for Integrated Photonic Interconnects

*Michael Georgas, Jonathan Leu, Benjamin Moss, Chen Sun and Vladimir Stojanović*

Massachusetts Institute of Technology

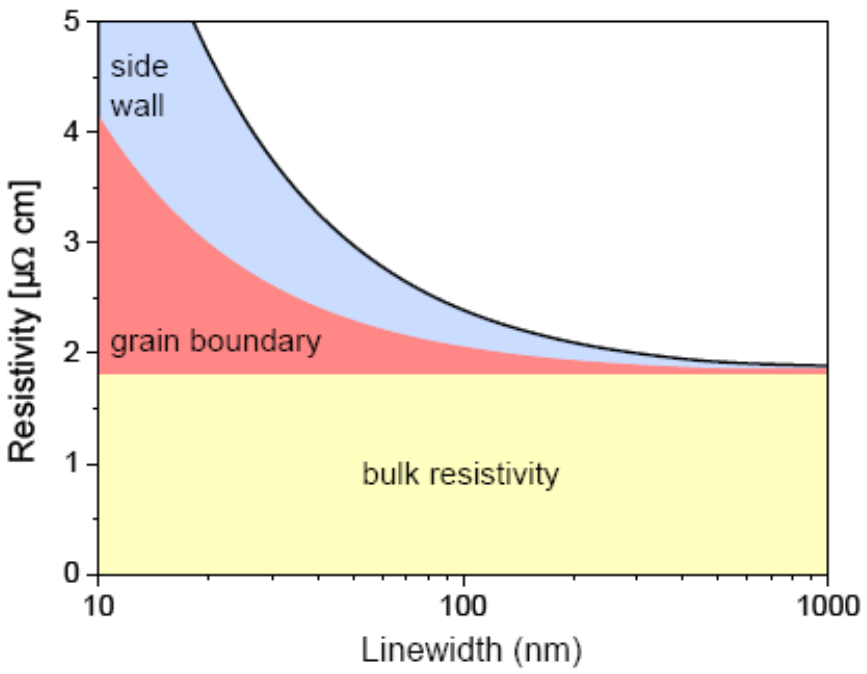
CICC 2011

# Manycore Socket Roadmap Fuels Bandwidth

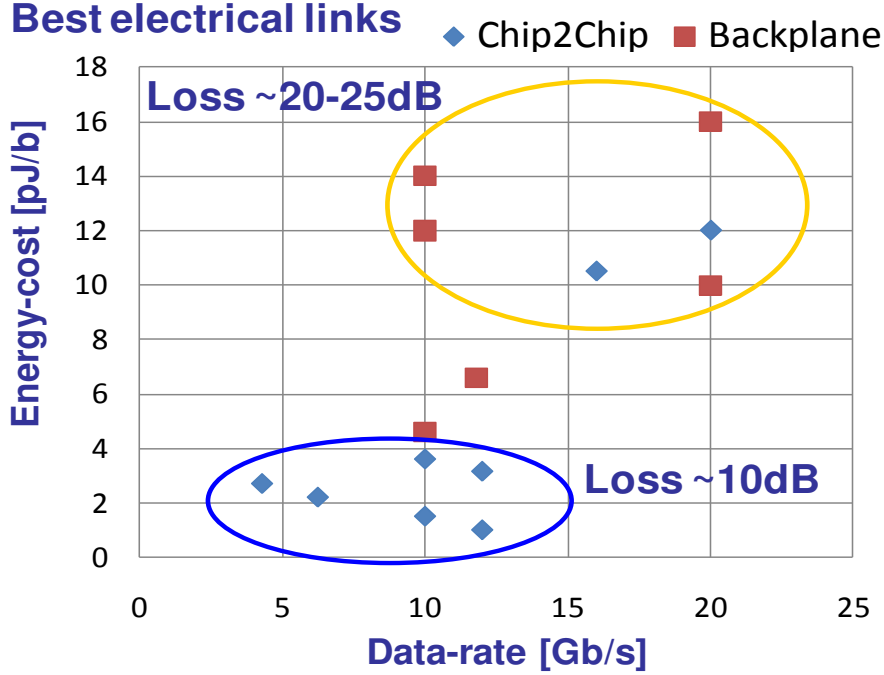


# Wire and I/O Scaling

**On-chip wires**

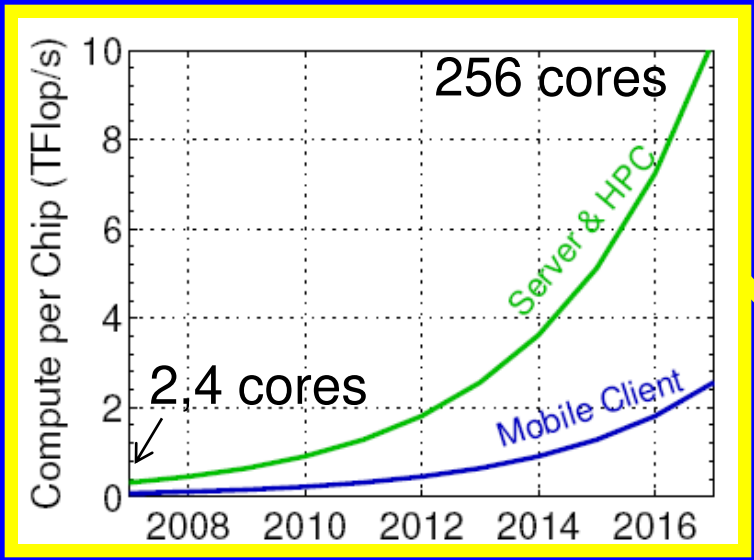


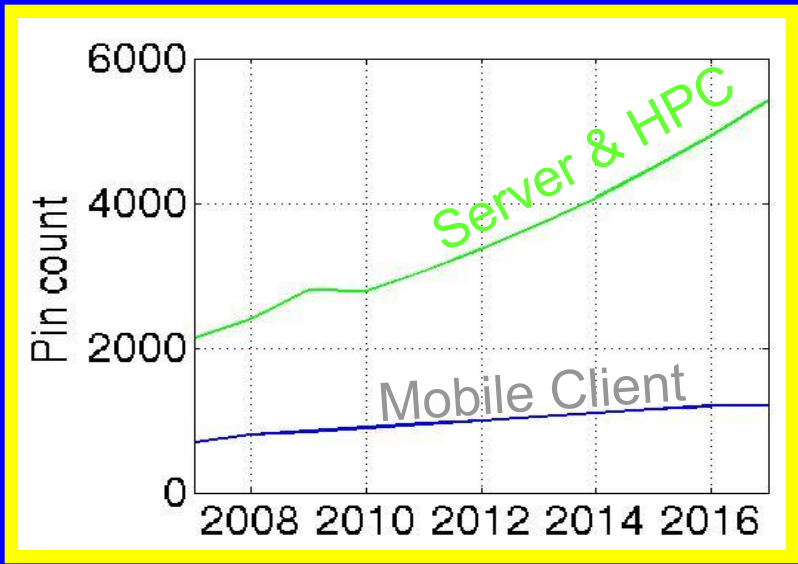

**I/O**

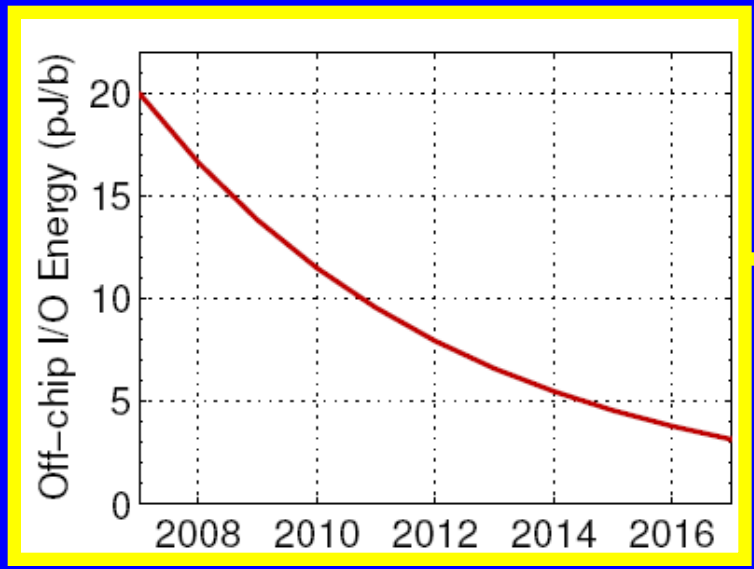



- Increased wire resistivity makes wire caps scale very slowly
- Can't get both energy-efficiency and high-data rate in I/O

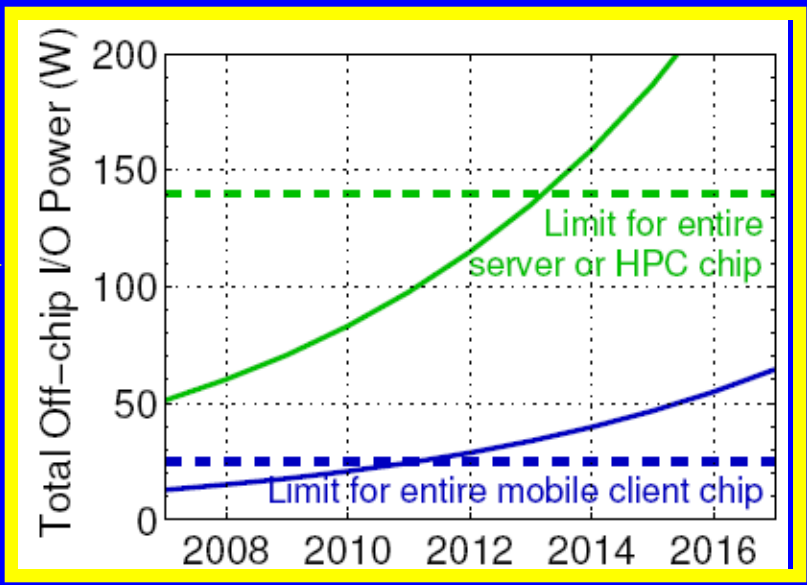
# Bandwidth, Pin-count and Power Scaling



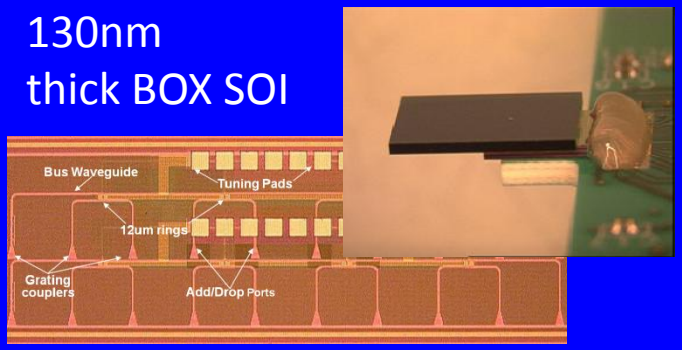




1 Byte/Flop



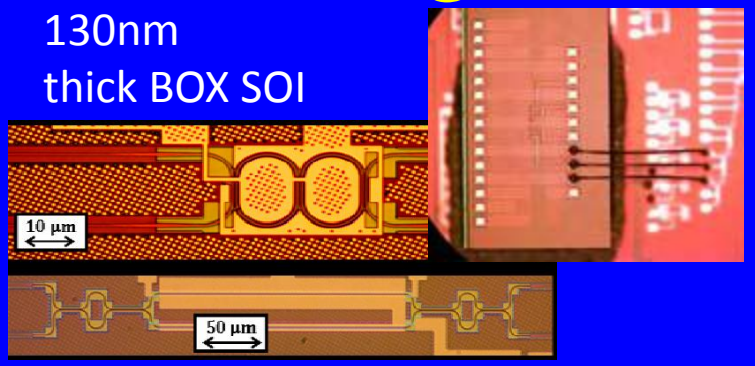
# Activity in Photonic Integration

130nm
thick BOX SOI



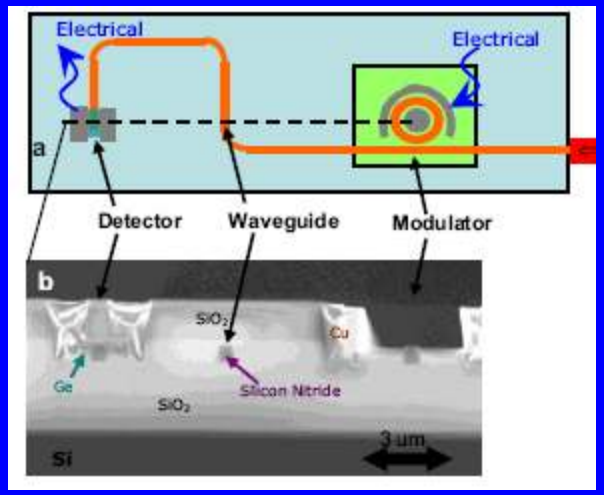

[Luxtera/Oracle/Kotura]

130nm
thick BOX SOI



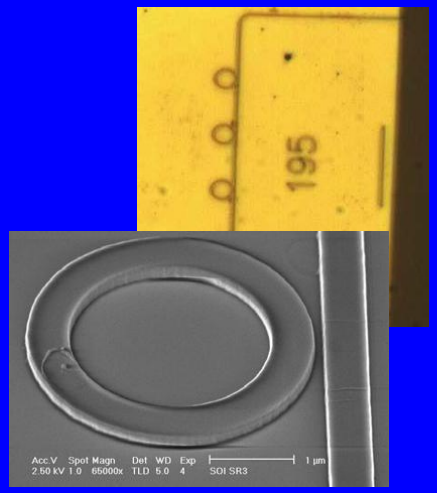

[IBM]

[Many schools]

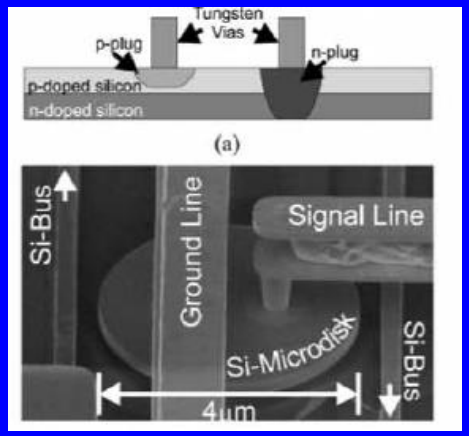



Bulk CMOS
Backend
monolithic

[Intel]




[HP]

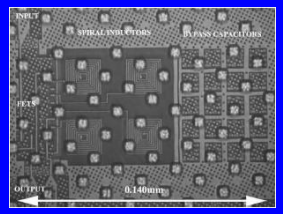


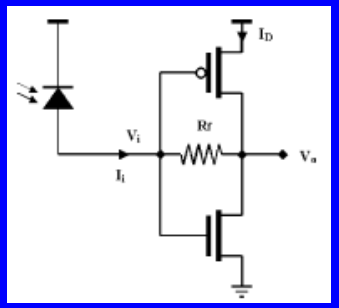

[Watts/Sandia/MIT]
[Lipson/Cornell]
[Kimerling/MIT]

# Bandwidth Density and Packaging

| | Electrical | Photonic |
|---|---|---|
| **Die Level** | • 100μm C4 bump pitch (20μm for microbump)<br>• 100 bumps/mm$^2$ ➔ 50 I/O<br>• 25 differential links @ 20Gb/s<br>**500Gb/s/mm$^2$** | **?** |
| **Package Level** | • 8000 pins ➔ 4000 I/O<br>• 2000 differential links @ 20Gb/s<br>• 40mm x 40mm socket<br>**25 Gb/s/mm$^2$** | • 100μm optical fiber pitch<br>• 100 fibers @ 1Tb/s/fiber<br>**100Tb/s/mm$^2$** |

# Optical Integration Trends



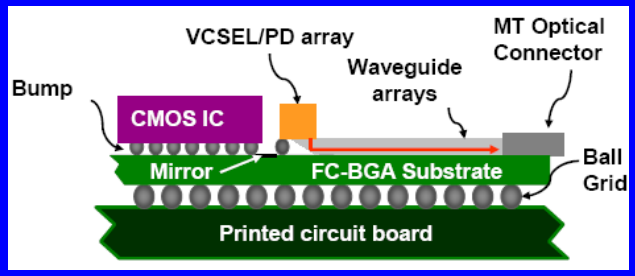



$C_{in,total}$ **=320-fF**

80-uA sensitivity at 20-GHz BW.

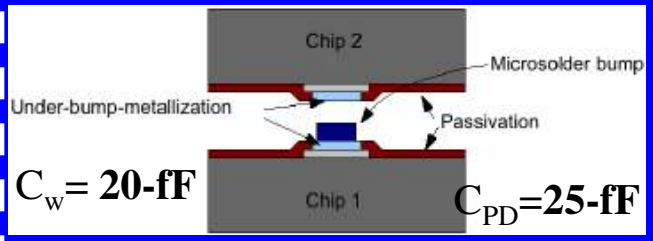
[Kromer et al. JSSC 2004]

**Discrete Components**




$C_w + C_{PD}$**=90-fF**

[Young et al. ISSCC 2009]




$C_w$**= 20-fF**

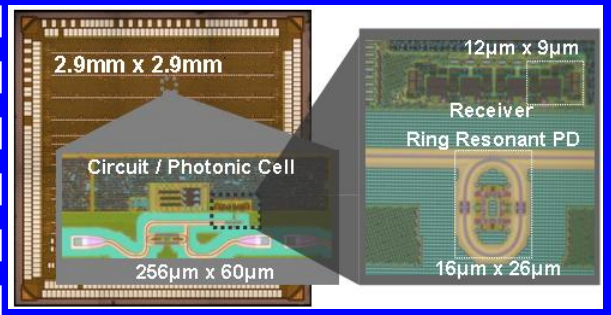
$C_{PD}$**=25-fF**

9 µA sensitivity at 5 Gb/s.

Energy-cost is 690 fJ/bit.

[Li et al. SPIE 2010]

**Hybrid Integration**




$C_{PD}$**≈10-fF**, $C_{wire}$ **≈ 4-fF**

µA-sensitivity at 3.5 Gb/s

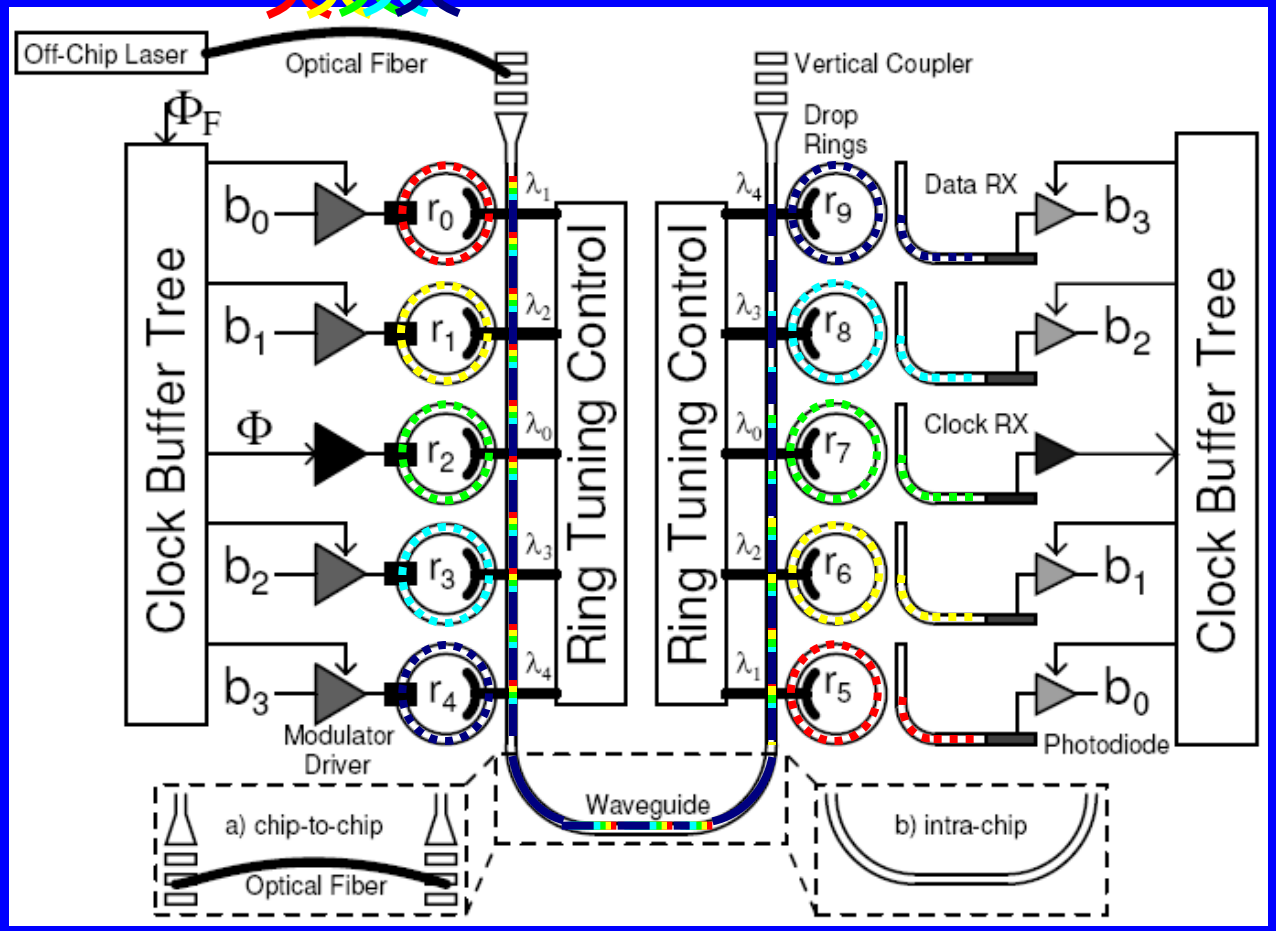
~50fJ/bit energy-cost

[Georgas et al. ESSCIRC 2011]

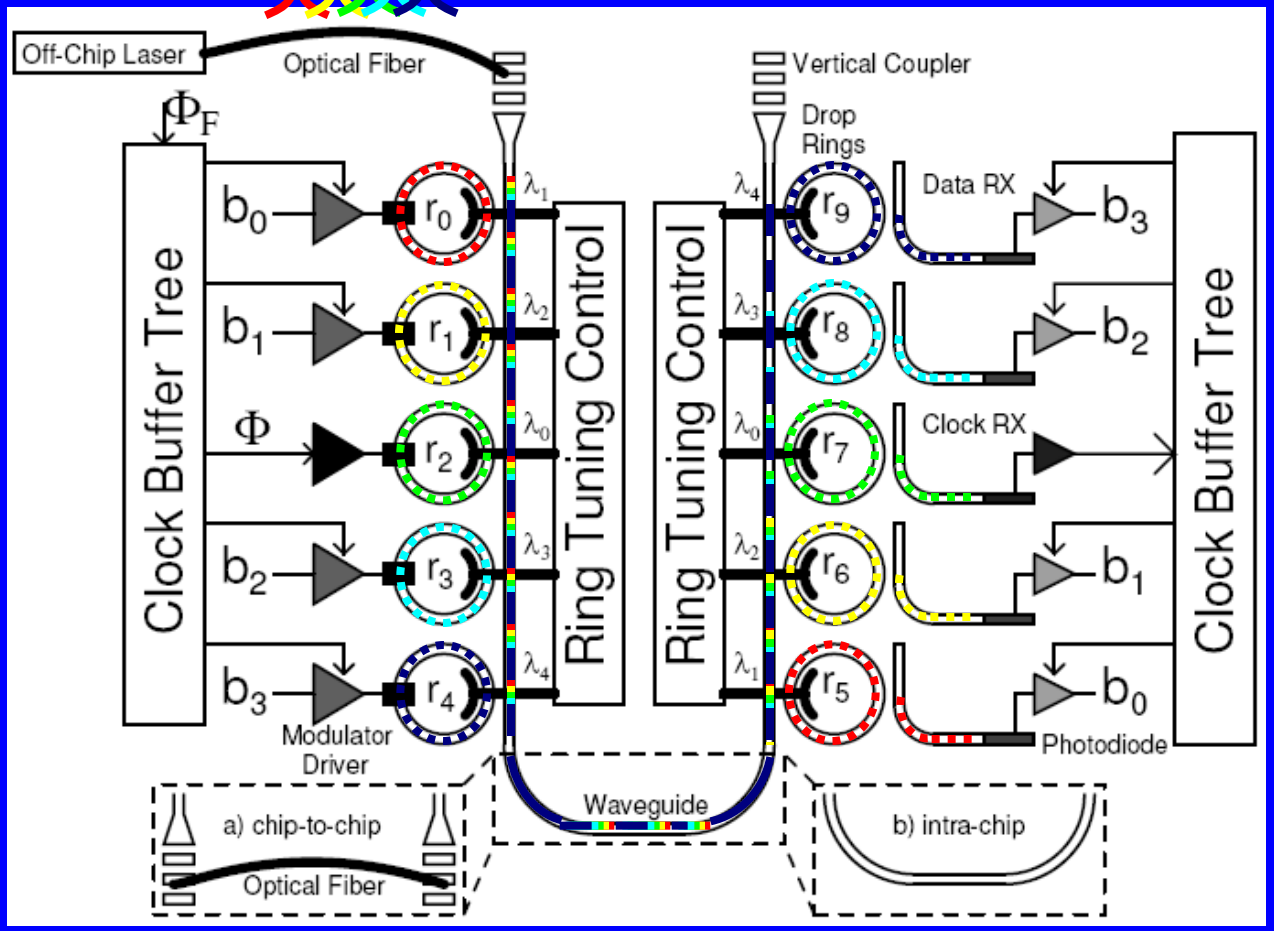
**Monolithic Integration**

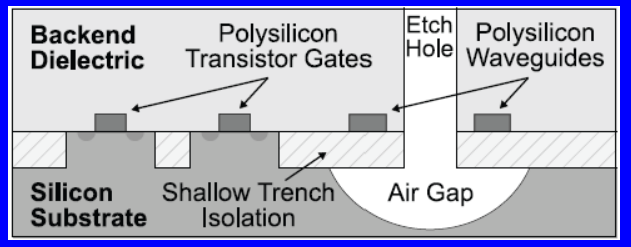
Decreasing $C_{PD}$ and $C_{wire}$

# Integrated Photonic Interconnects

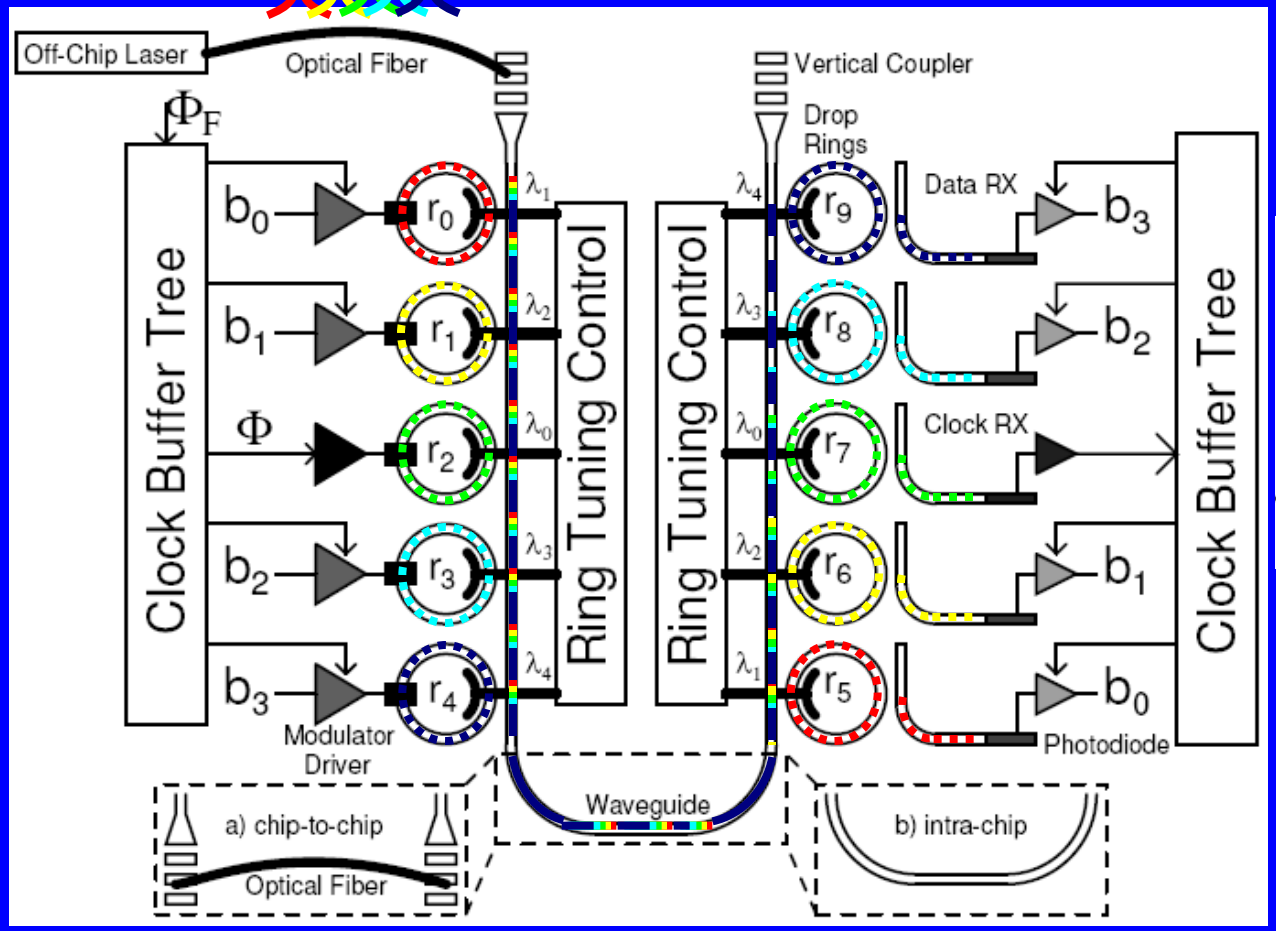


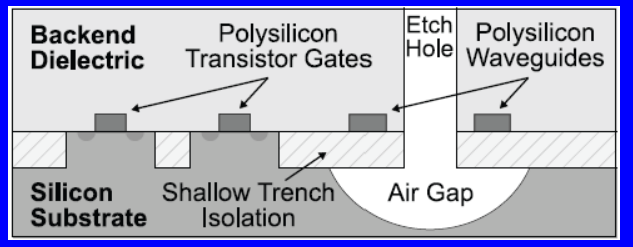
# Integrated Photonic Interconnects

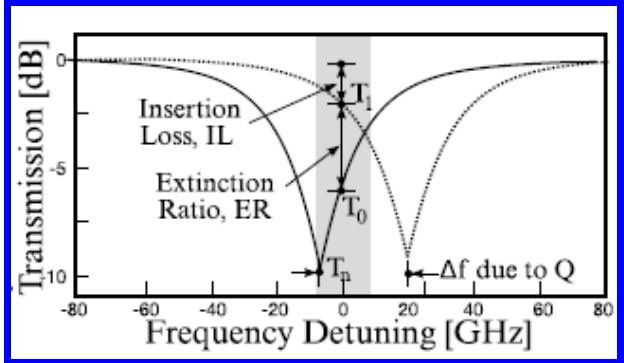


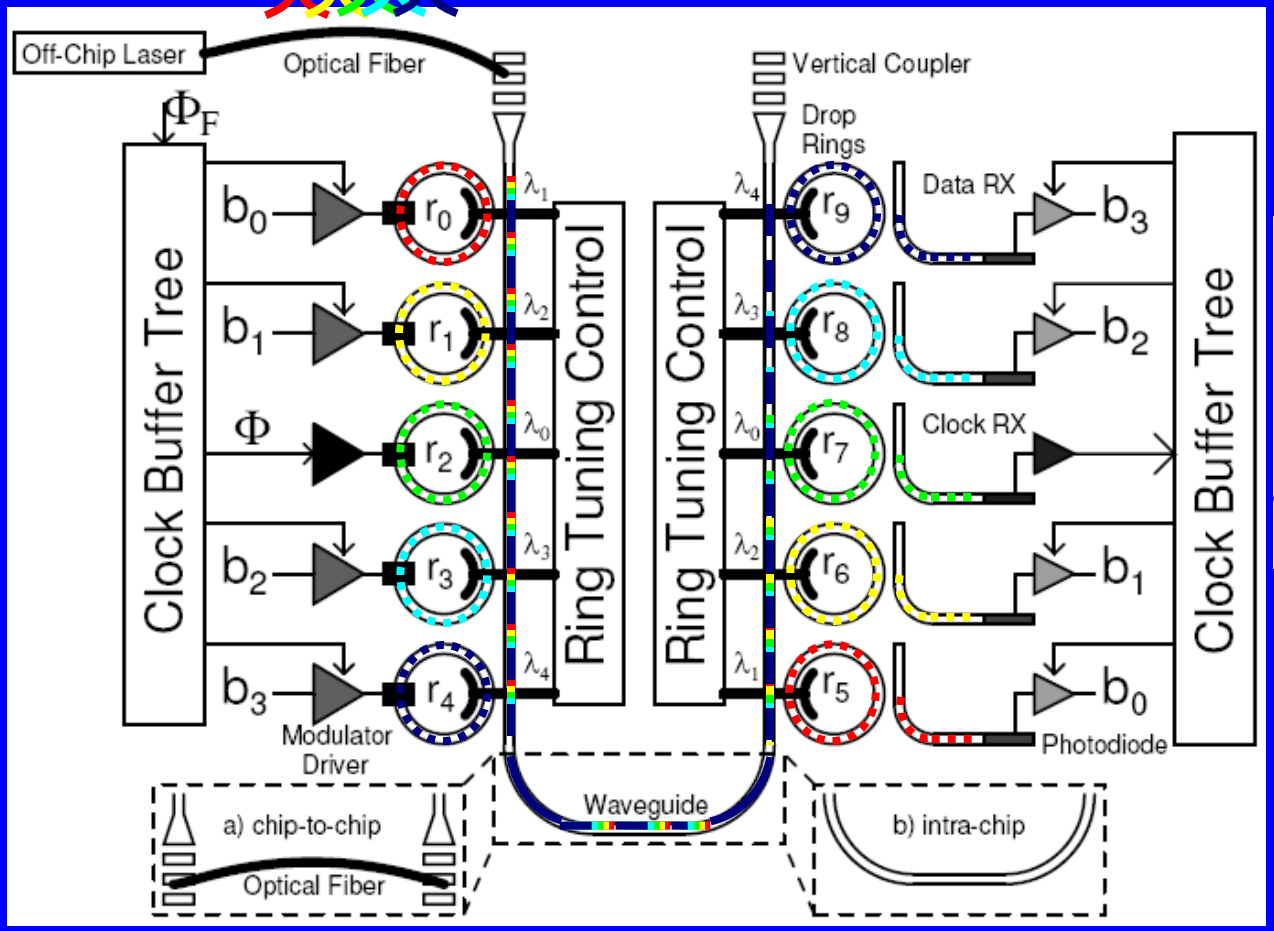
# Integrated Photonic Interconnects

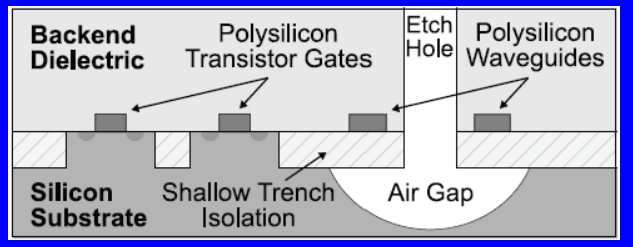


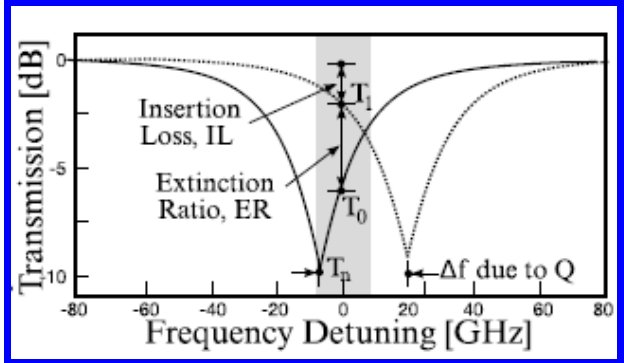



# Integrated Photonic Interconnects



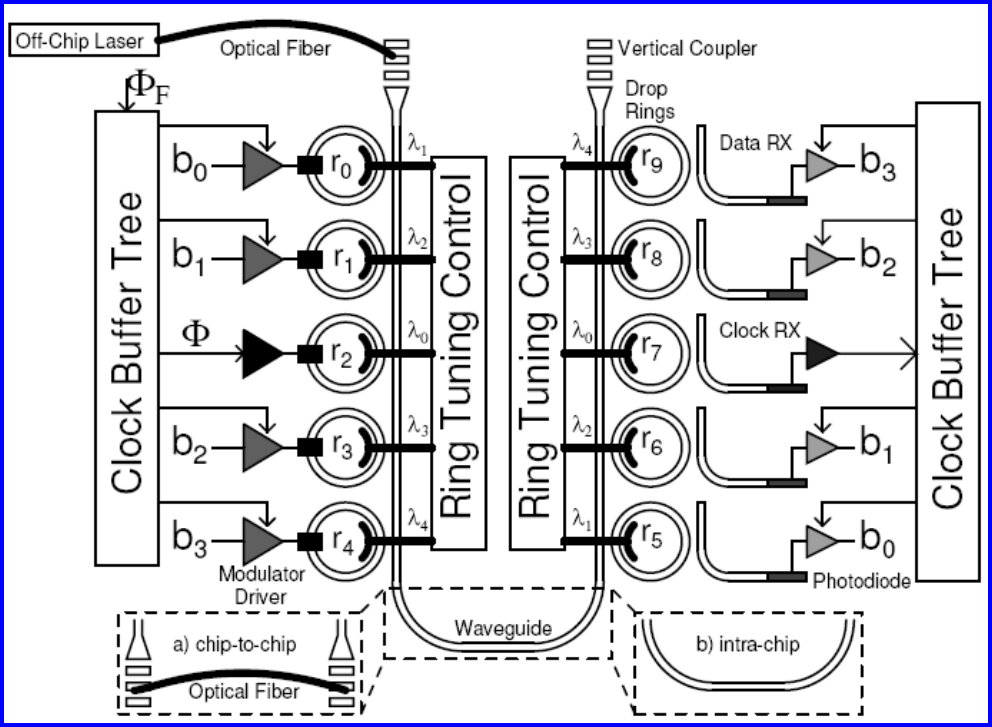




- Each λ carries one channel of data.

➔**Bandwidth Density** achieved through DWDM

➔Energy-efficiency achieved through low-loss optical components and tight integration

# Photonic System Design




- link components tightly integrated

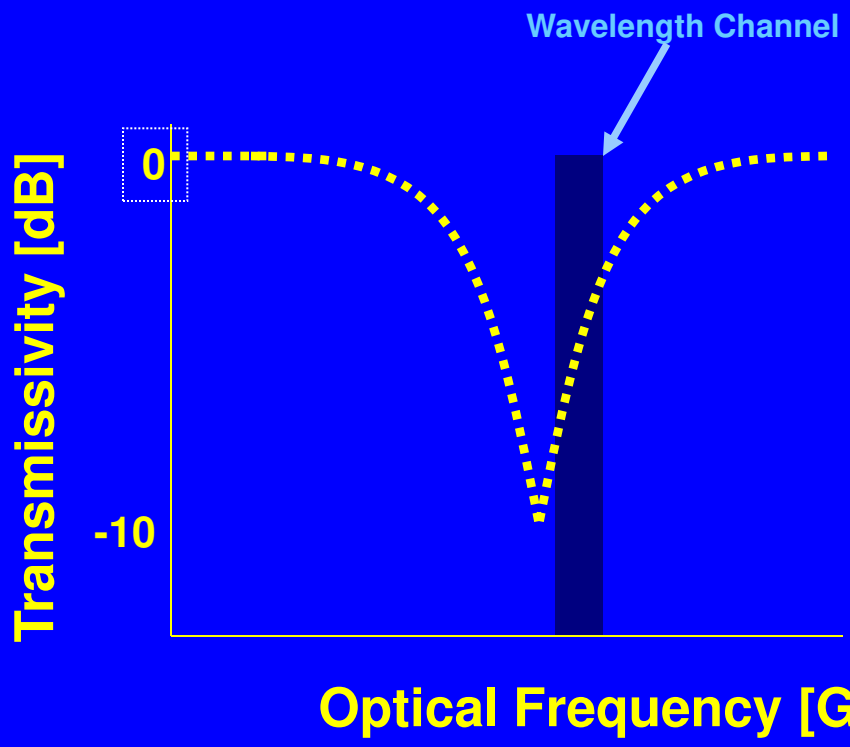
➔ care about *system* energy-efficiency and performance

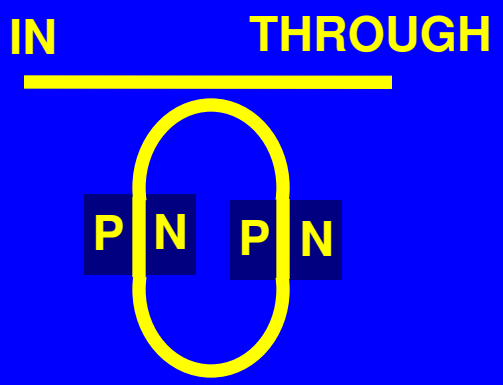
- Need component models to understand system tradeoffs

# Outline

- Motivation
- **Photonic Link Components**
  - Modulator and driver
  - Receiver
  - Single Link Analysis
- Towards a WDM Photonic Link
  - Clock distribution
  - Ring Tuning
  - WDM Link Analysis
- Conclusion

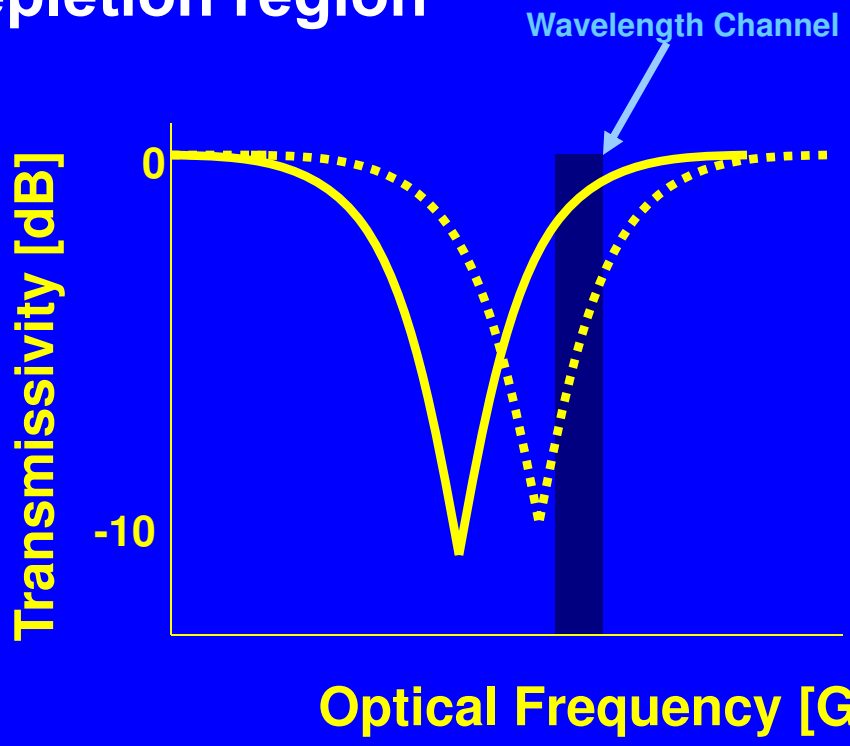
# Optical Modulation

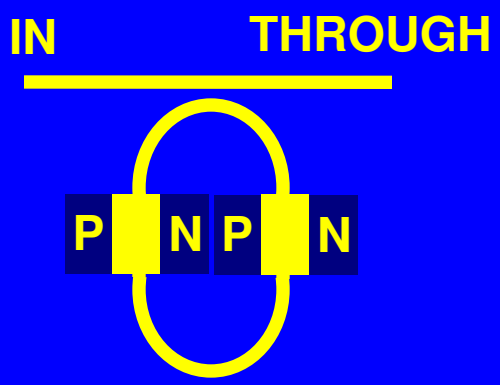


# Optical Modulation

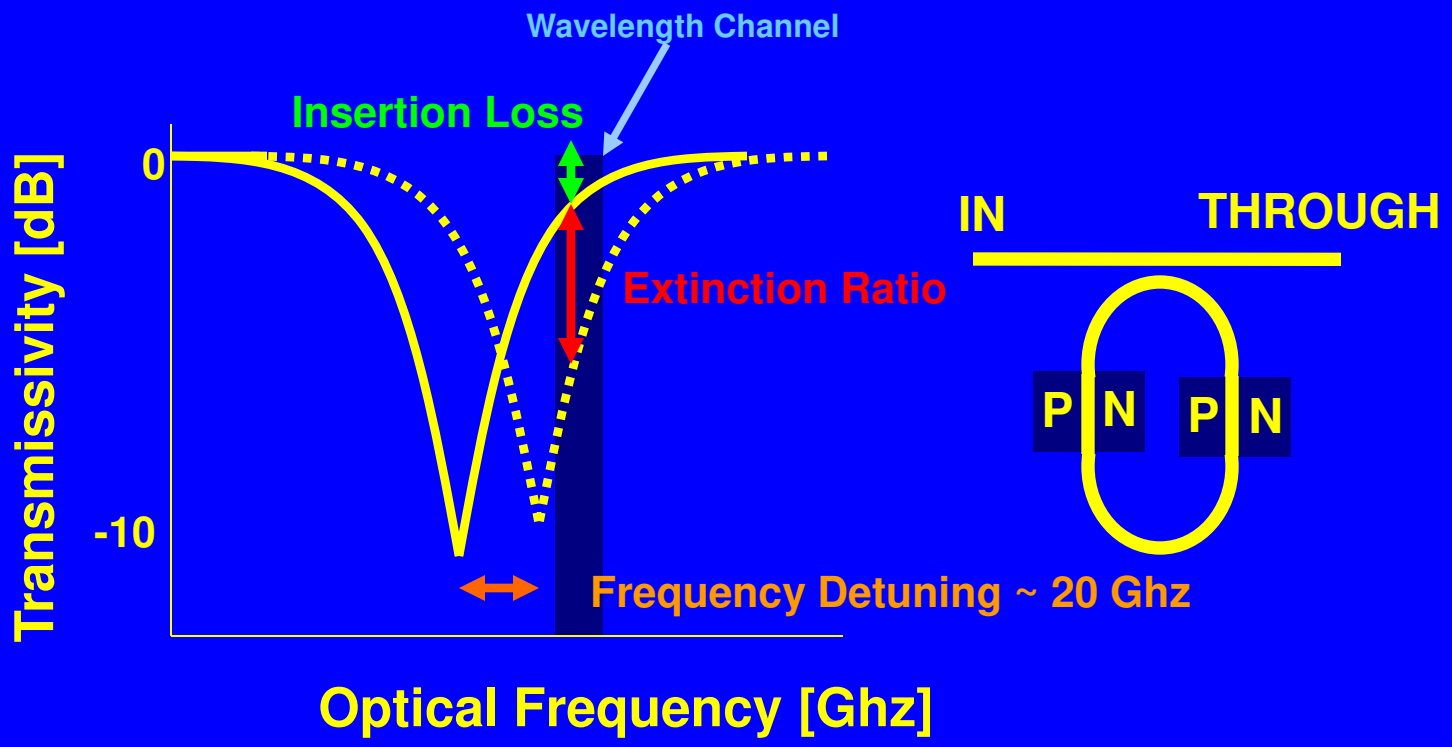
- leverage free-carrier-dispersion effect to modulate P-N junction's depletion region



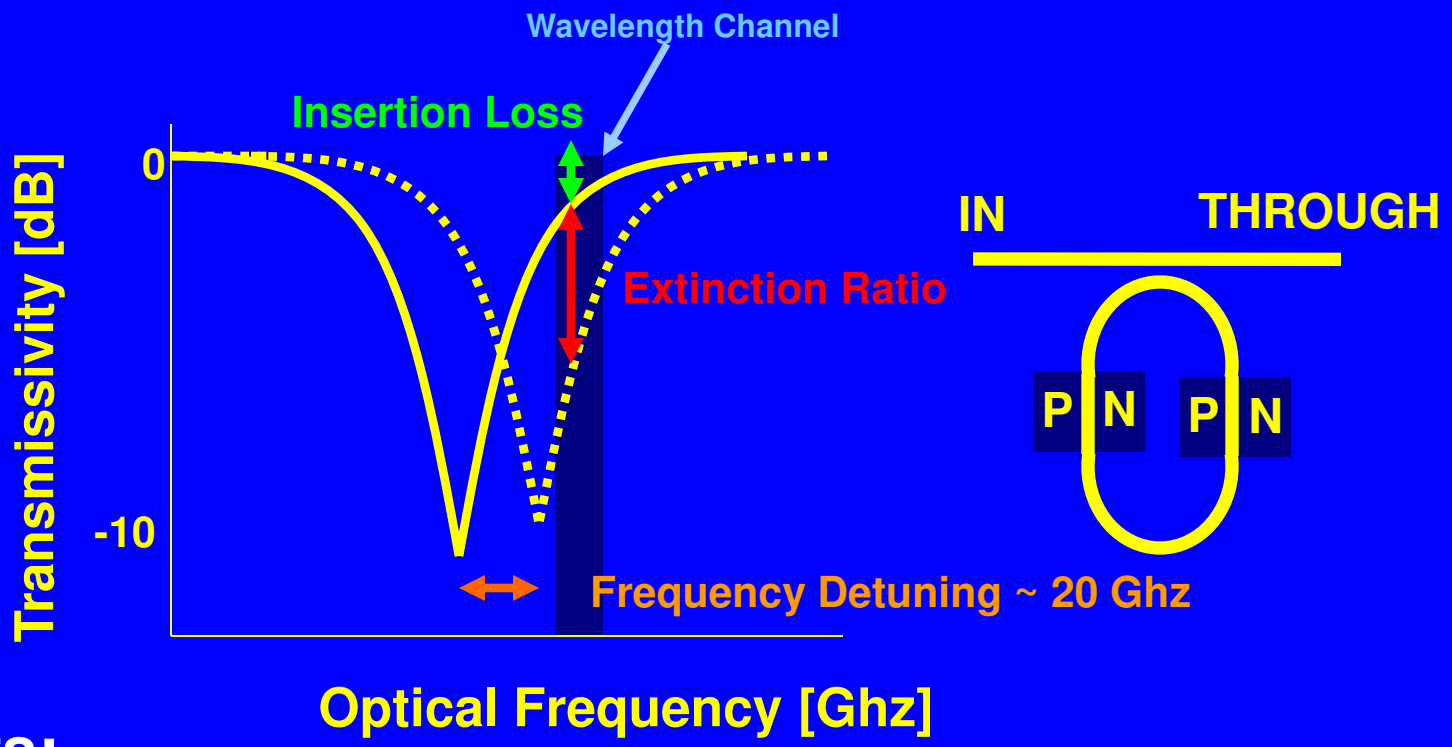

- OOK modulation by shifting ring resonance in and out of wavelength channel

# Optical Modulation




- **OOK modulation by shifting ring resonance in and out of wavelength channel**

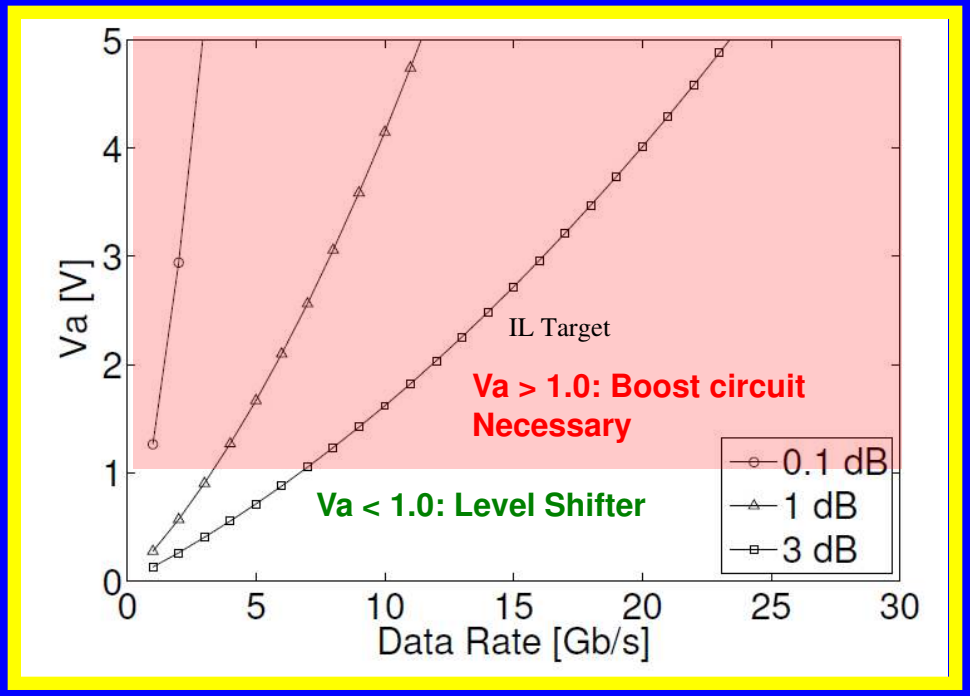
# Optical Modulation



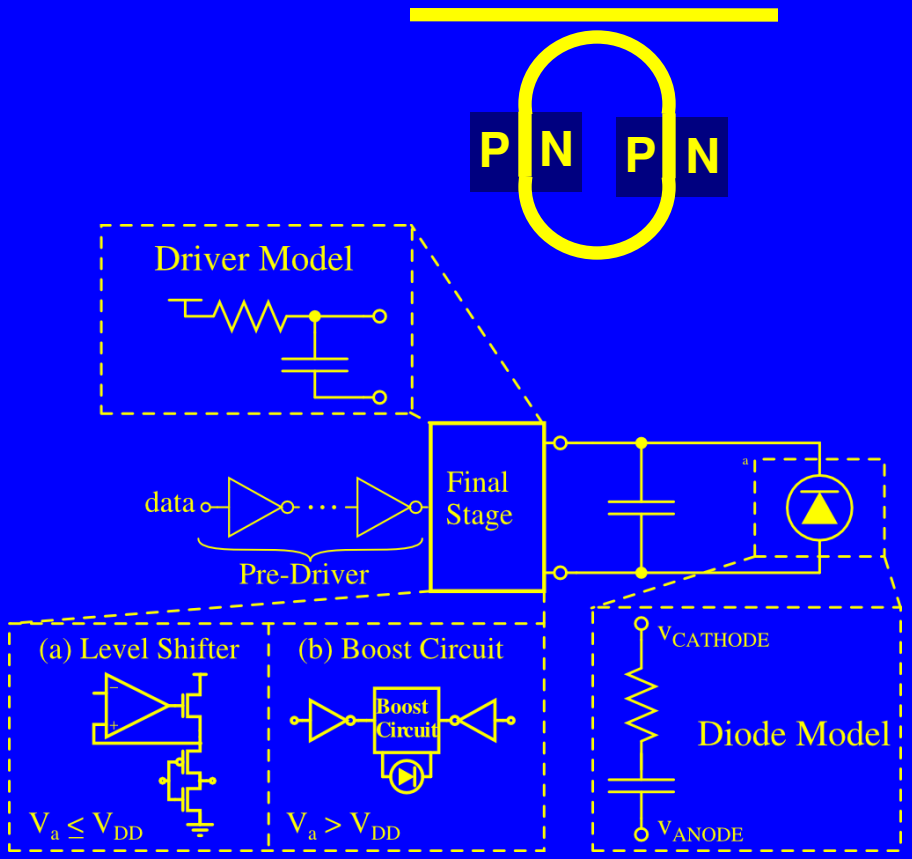

**Key Tradeoffs:**

- **Insertion loss vs. extinction ratio**
- **Extinction ratio vs. energy-efficiency of driver**

# Modulator Driver Model

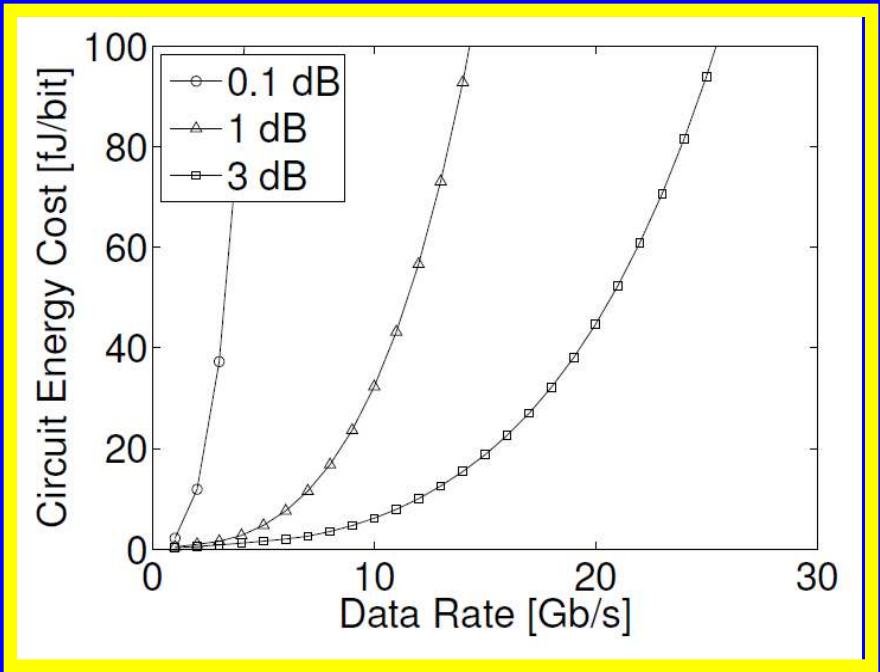



**Required Reverse-Bias Voltage Va**

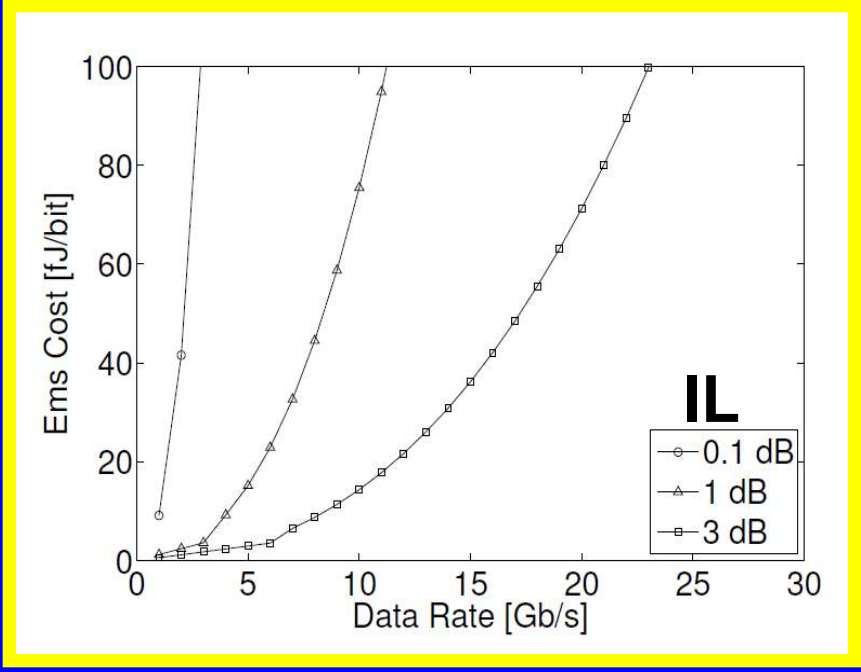



- **Electrically, modulator is a varactor**
- **Increased data-rate requires increased shift (charge)**
- **Final stage topology tailored based on Va**

# Modulator Energy Cost Breakdown



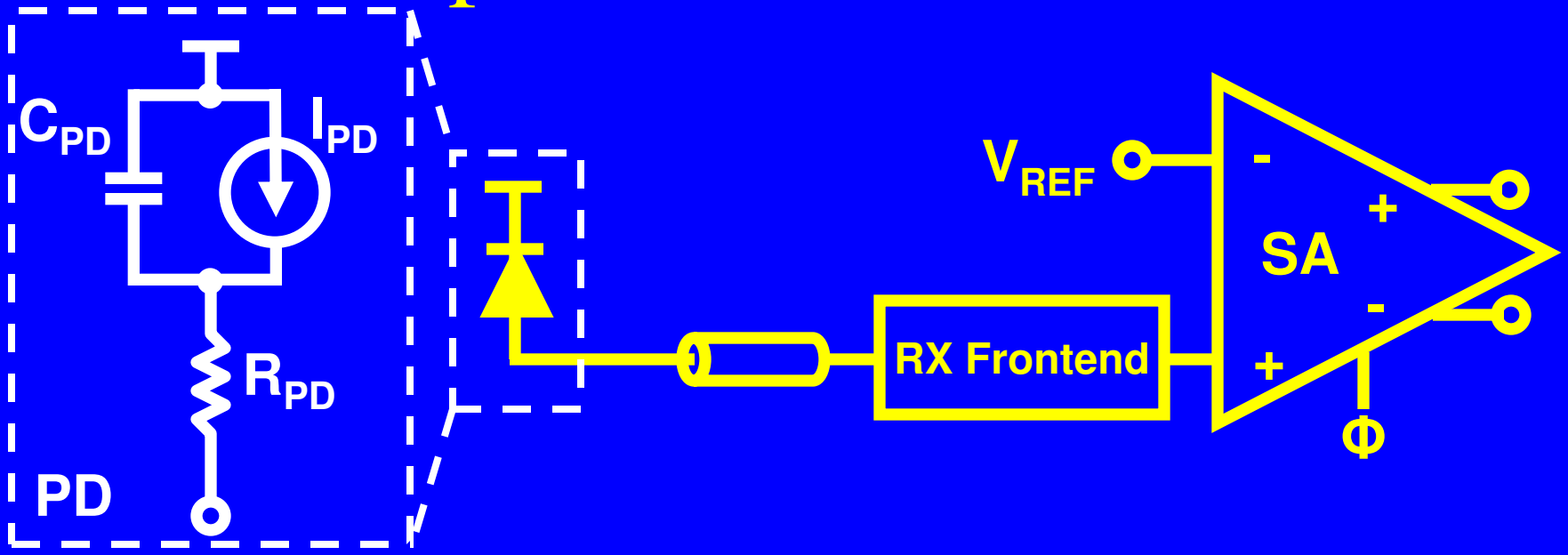

Circuit Energy Cost



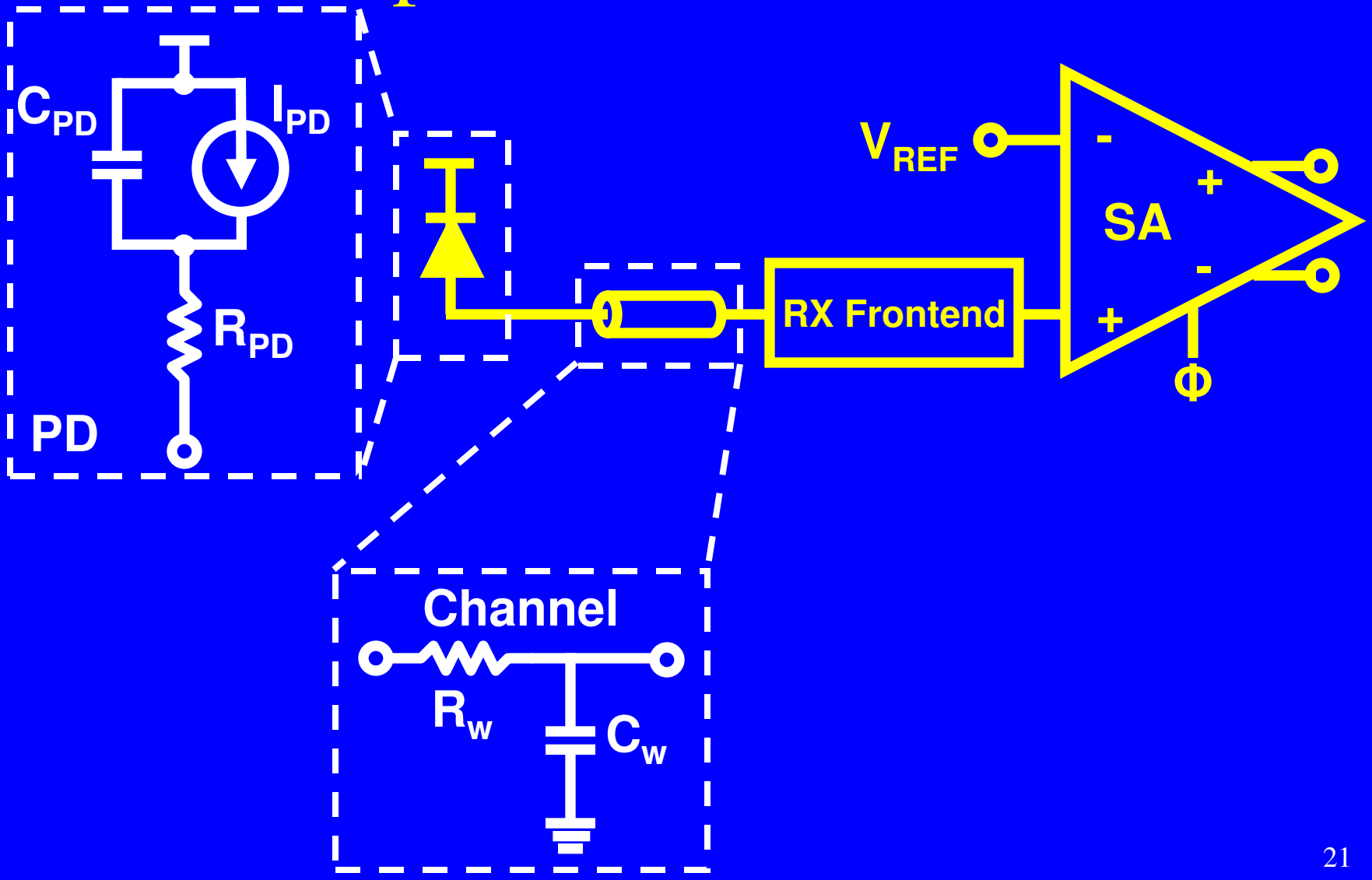

Device Energy Cost

- **Circuit and device costs are roughly balanced**
- **Cost increases at high rates due to super-linear relationship with $V_a$**
- **Insertion loss, extinction ratio, and energy-efficiency trade-offs to be made at the system-level.**

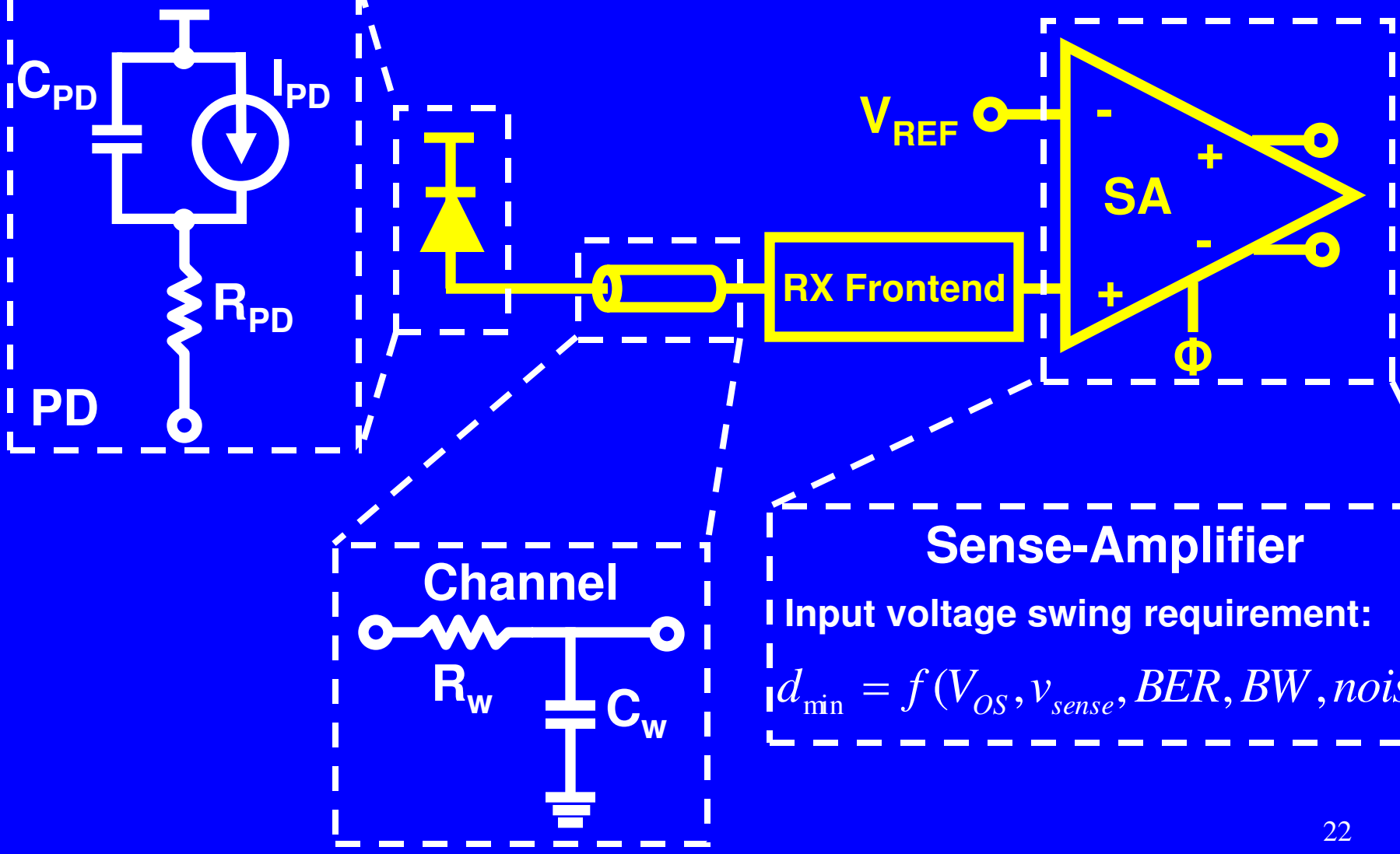
# Optical Data Receiver



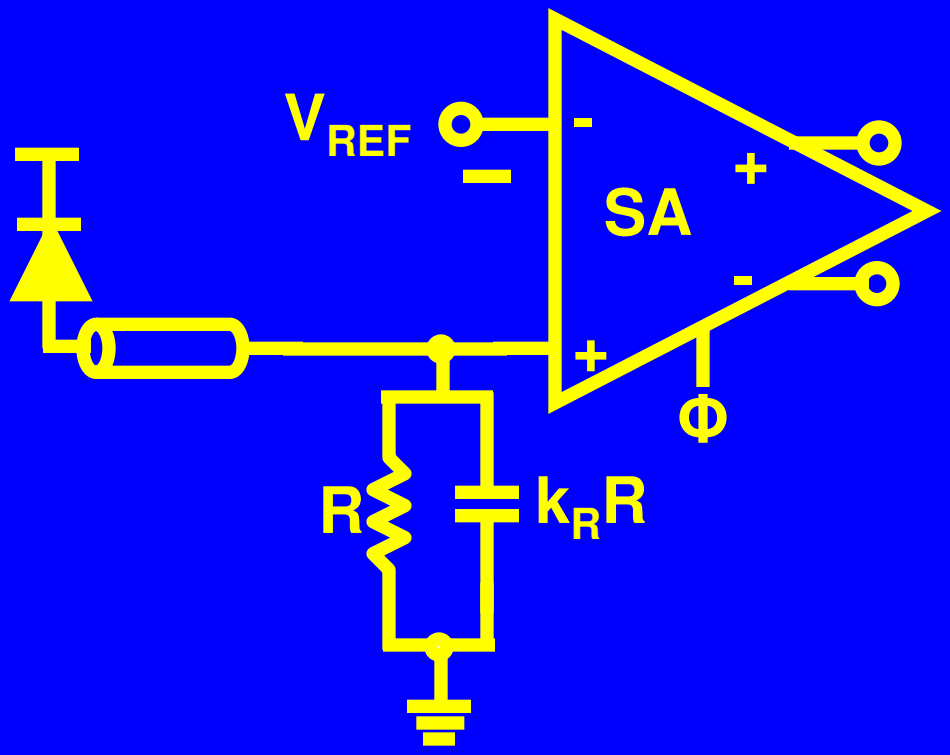
# Optical Data Receiver



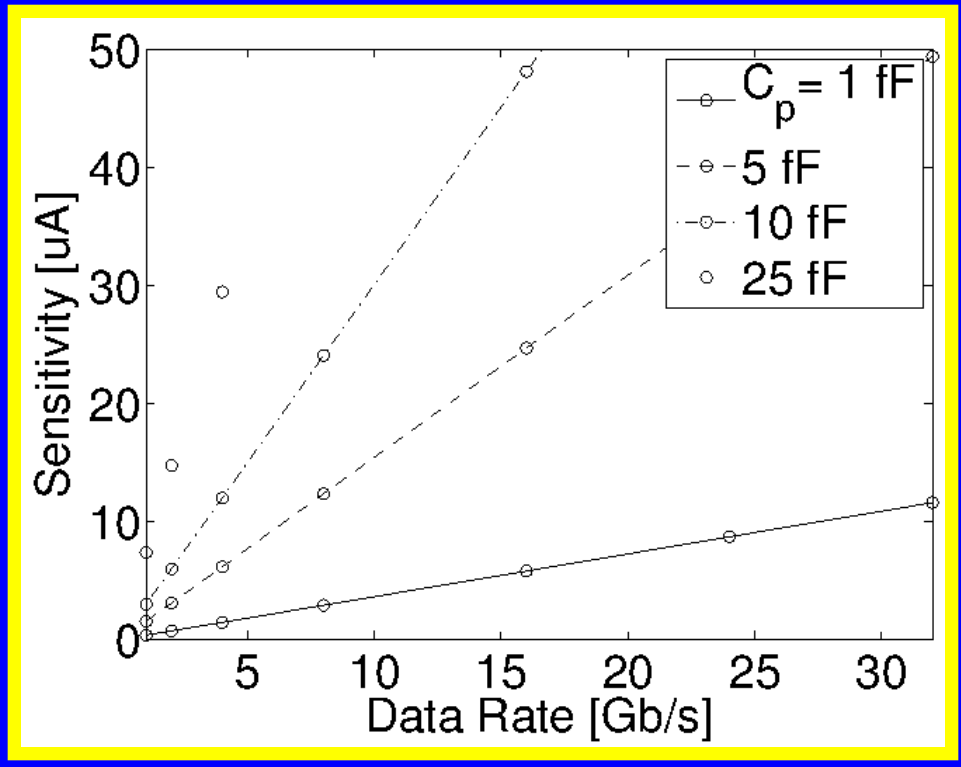
# Optical Data Receiver



# Optical Data Receiver: Resistor



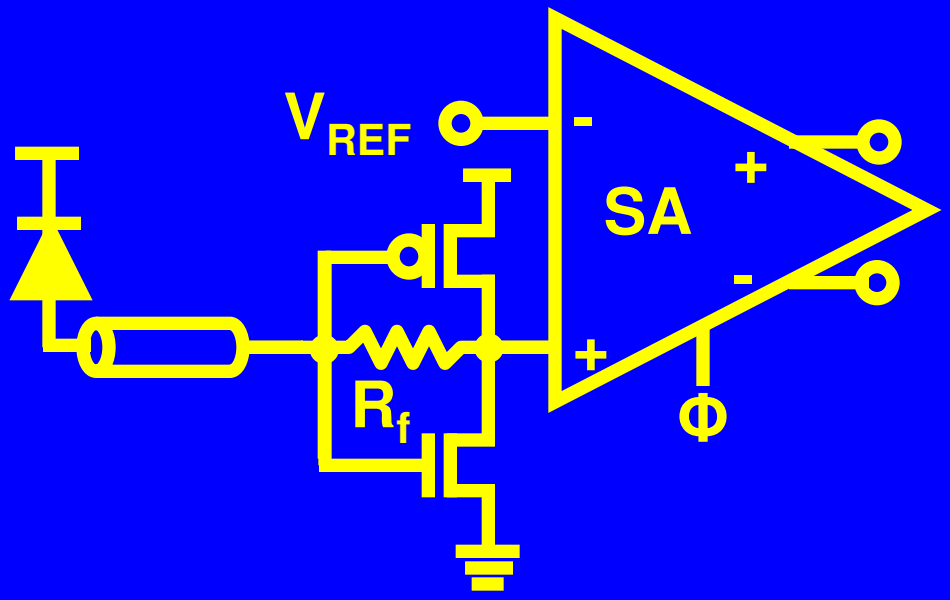

$$I = f(BER, BW, ER, noise, C_p)$$

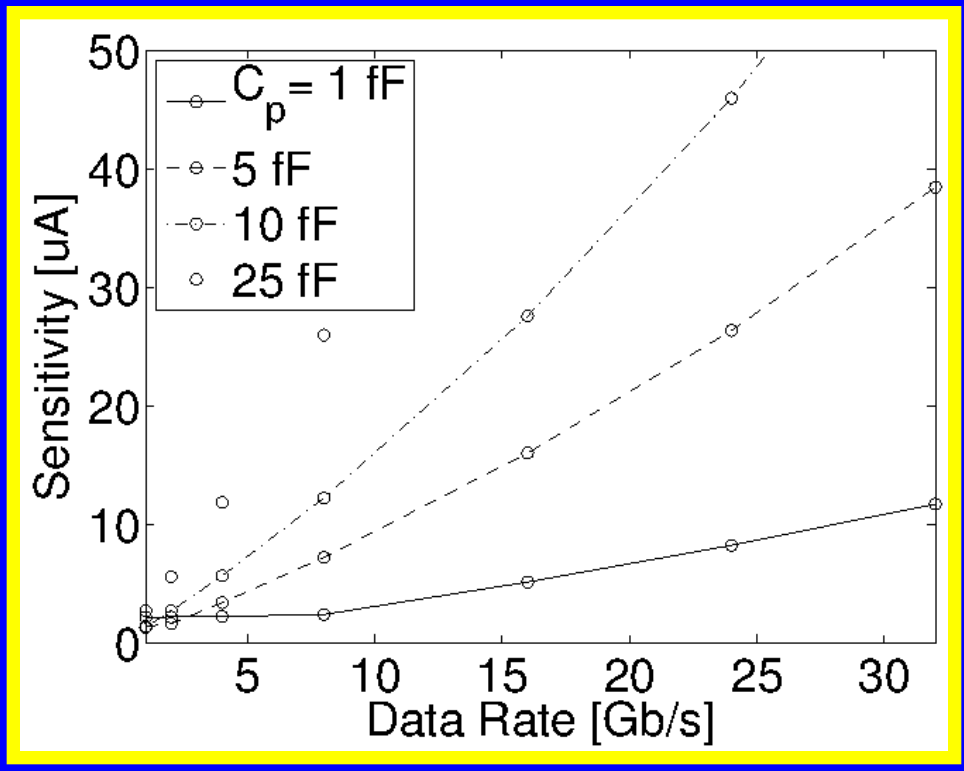



**For each data-rate, compute I that satisfies SA requirements.**

**Linear: gain~1/BW**

# Optical Data Receiver: TIA

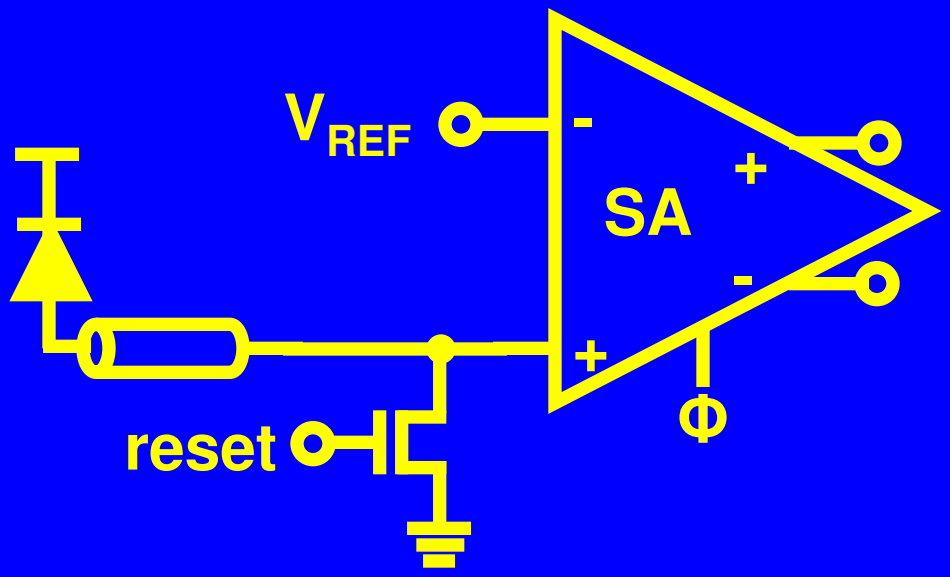



**Trade gain for power by decreasing $Z_{IN}$ while keeping $Z_{TIA}$ high.**

**Compute $I_{ON}$ as before**

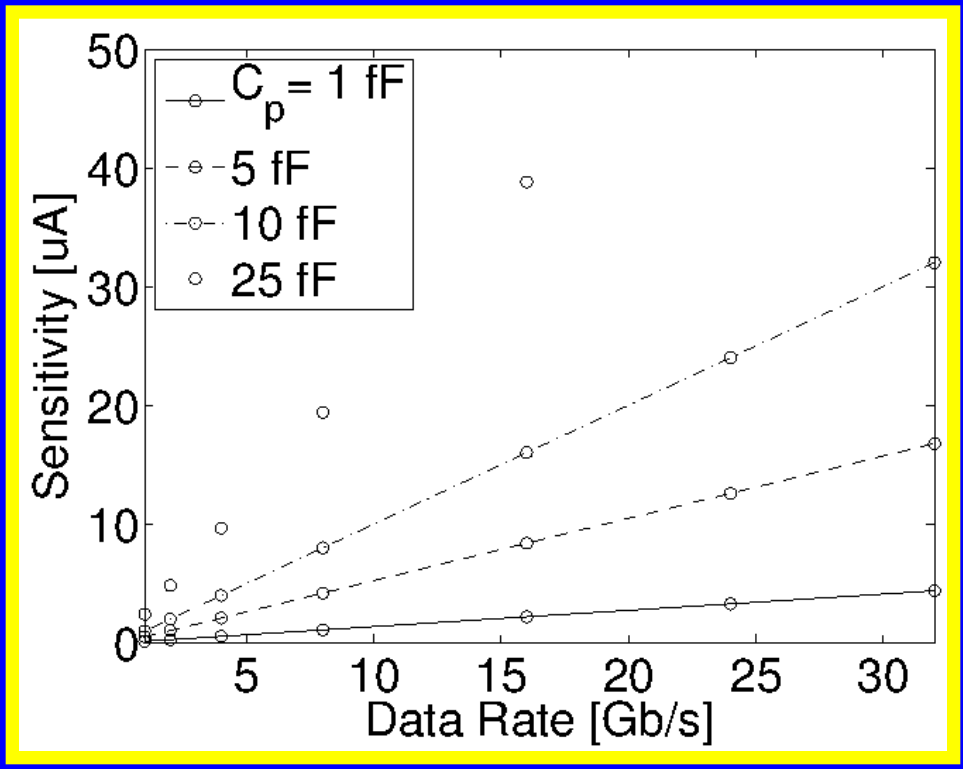
# Optical Data Receiver: Integrator




$$C_{INT} = C_{PD} + C_w + C_{SA,in}$$

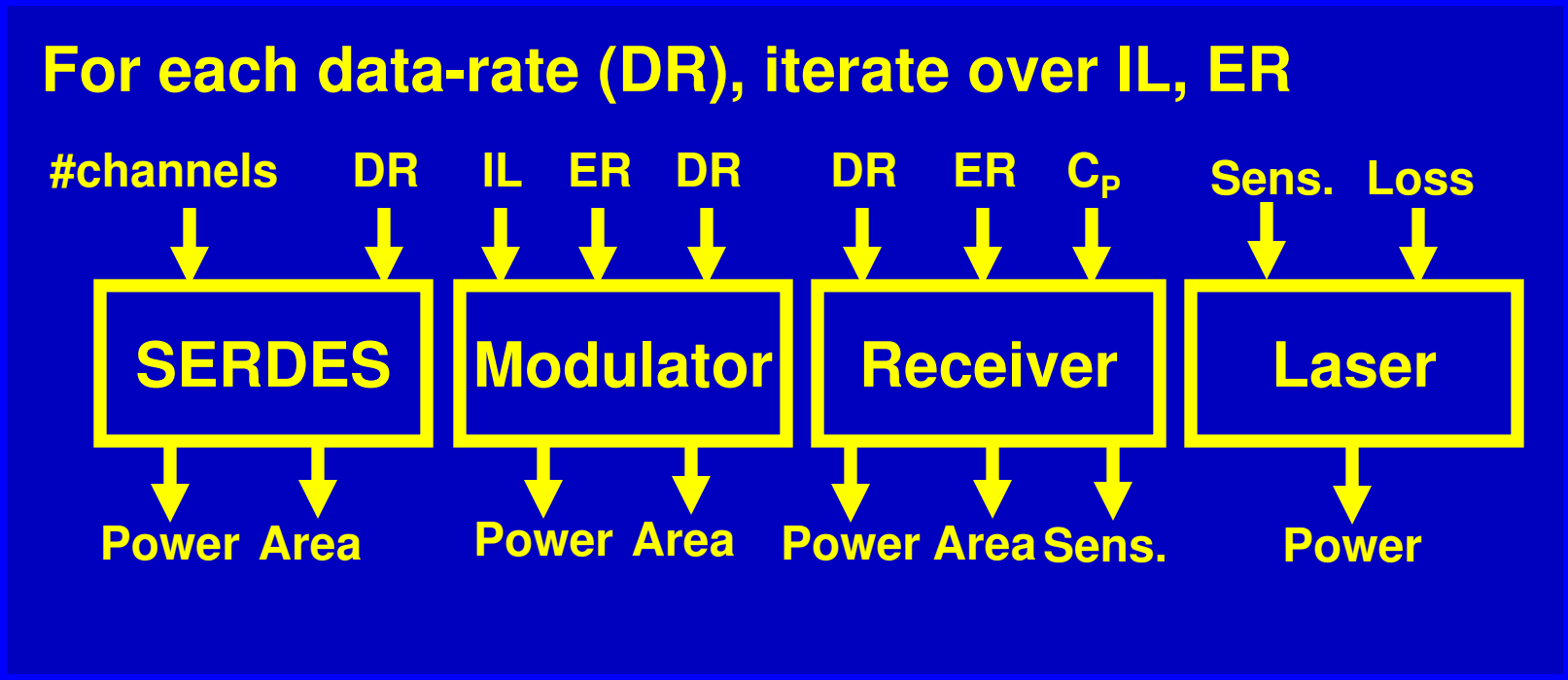
$$R = \frac{k_{INT} T_{bit}}{C_{INT}}$$

**Integrate over a fraction of a bit time, and reset**




**$k_{INT}$ models integration time**

# Single Channel Link Tradeoffs

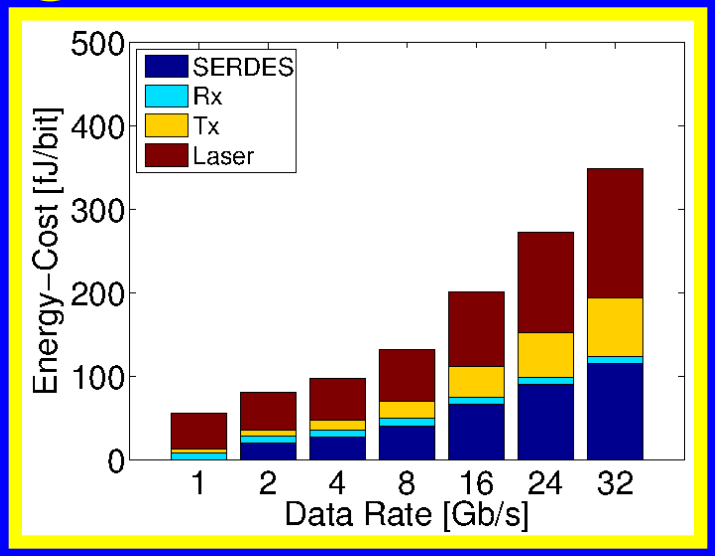



- Our examination looks across:
  - different loss options: 10-dB and 15-dB cases
  - different technologies: $C_P$ of 5-fF and 25-fF

# Single Channel Link Tradeoffs

**Loss**

**10-dB**



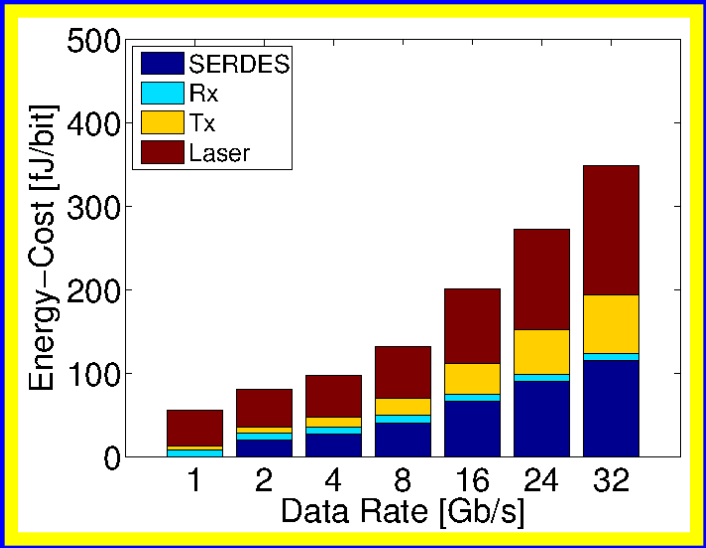

**5-fF**

- **SERDES cost increasing with rate**
- **Decreased RX sensitivity maps to increased laser cost**
- **TX tries to compensate**

# Single Channel Link Tradeoffs

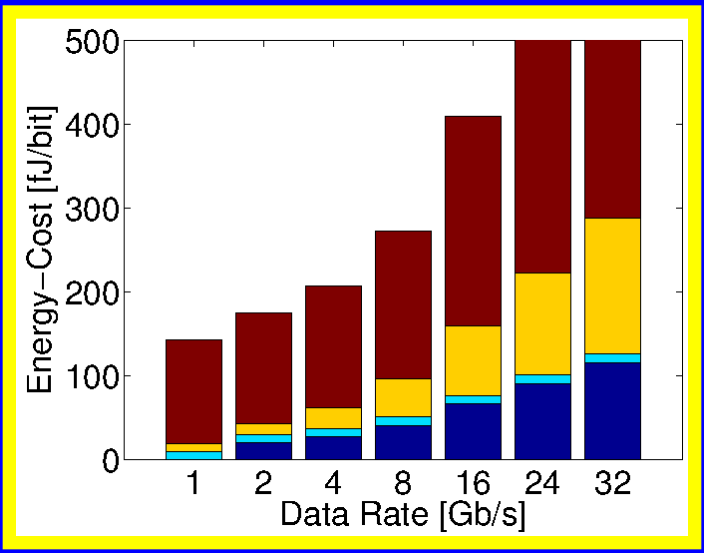
**Loss**

**10-dB**




**5-fF**

- **SERDES cost increasing with rate**
- **Decreased RX sensitivity maps to increased laser cost**
- **TX tries to compensate**



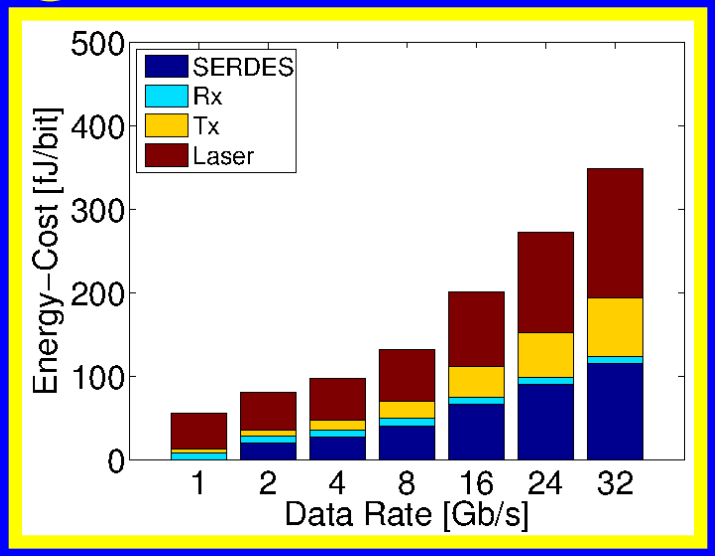

**25-fF**

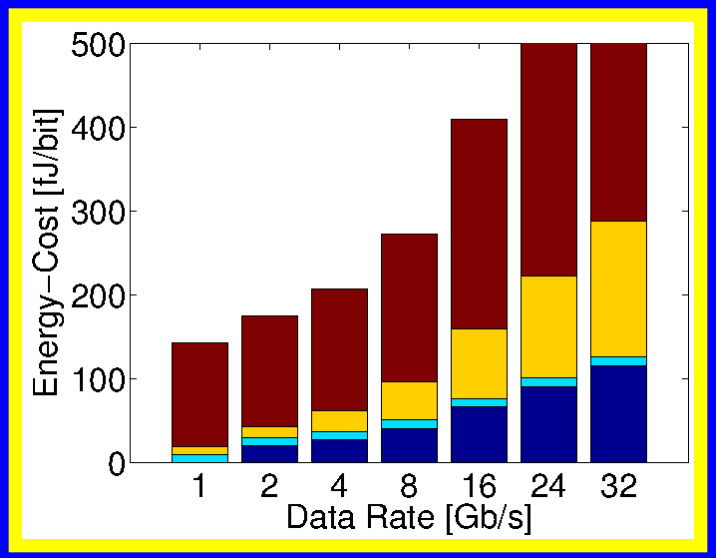
- **SERDES cost the same**
- **Decreased sensitivity maps to laser power**
- **TX again tries to compensate**

# Single Channel Link Tradeoffs

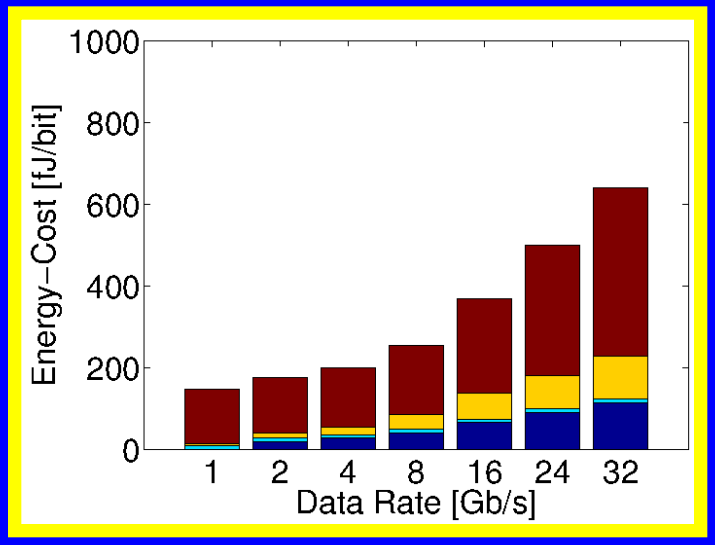
**Loss**

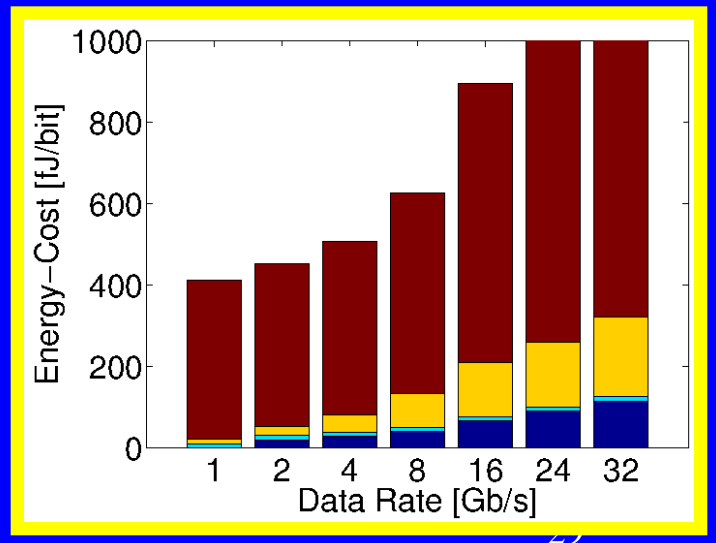
**10-dB**




**15-dB**



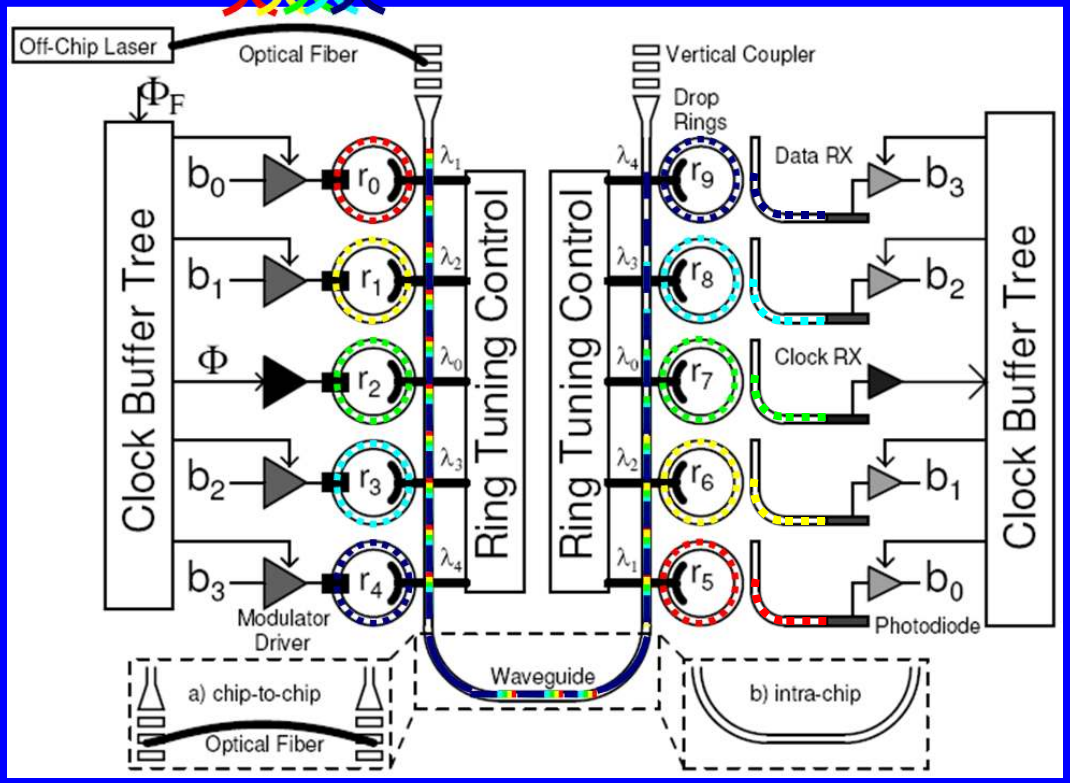




**5-fF** **25-fF**

# Outline

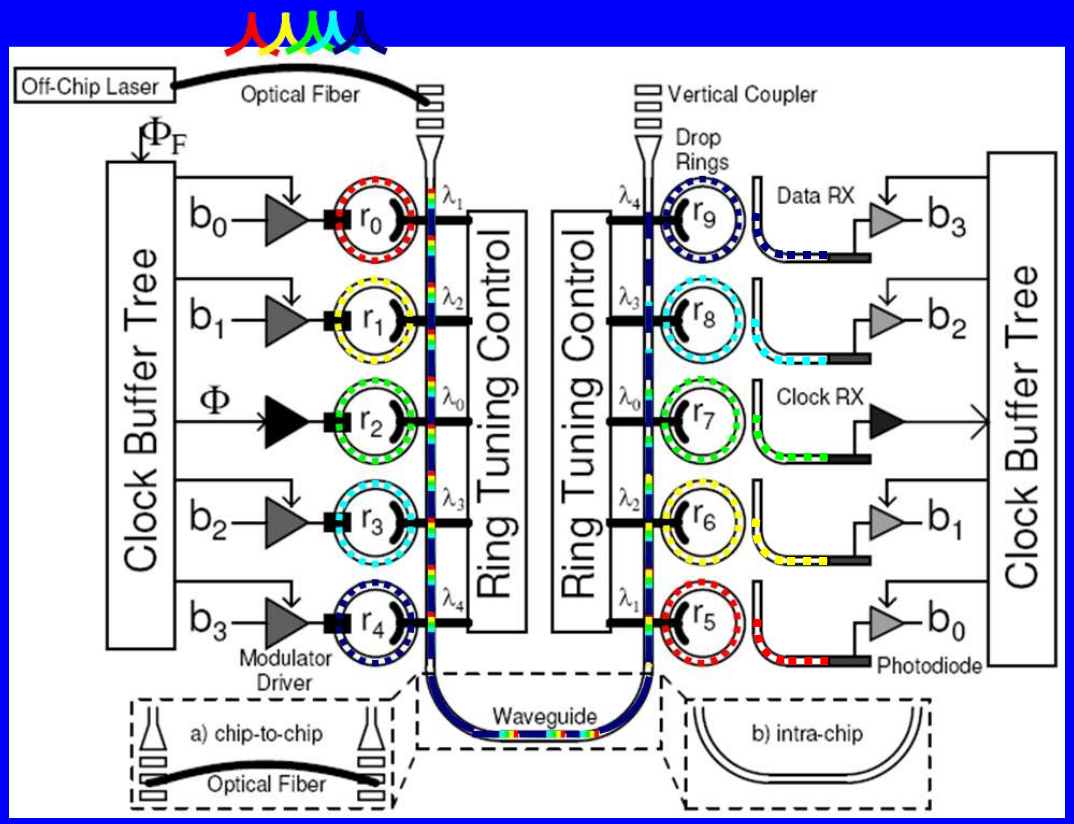
- Motivation
- **Photonic Link Components**
  - **Modulator and driver**
  - Receiver
  - Single Link Analysis
- Towards a WDM Photonic Link
  - Clock distribution
  - Ring Tuning
  - WDM Link Analysis
- Conclusion

# Optical Clock Distribution




- **Clock for receivers can be forwarded on with the data in DWDM**

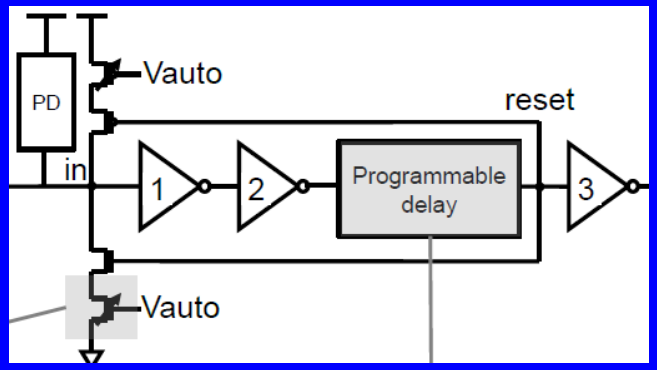
# Optical Clock Distribution



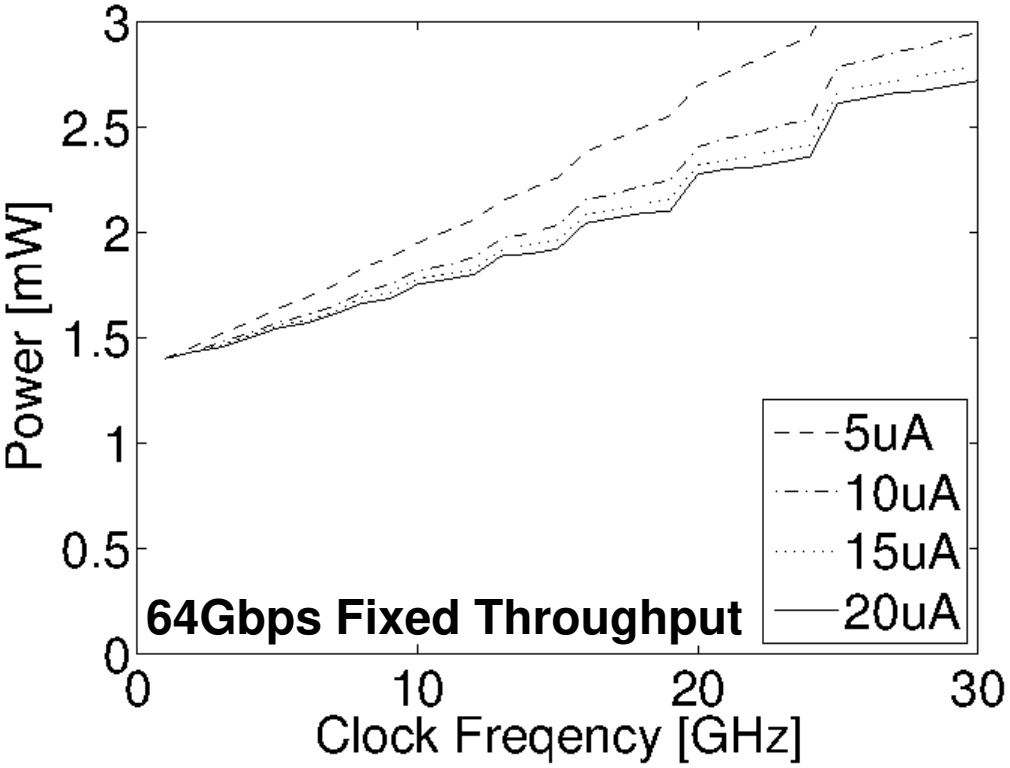

- Clock for receivers can be forwarded on with the data in DWDM
- does not suffer from rail-injected noise or crosstalk

➔no jitter added in channel

➔no PLL/DLL needed



# Optical Clock Distribution




- Assume an RX timing requirement of better than 3% UI
- Compute capacitive clock load based on number of channels

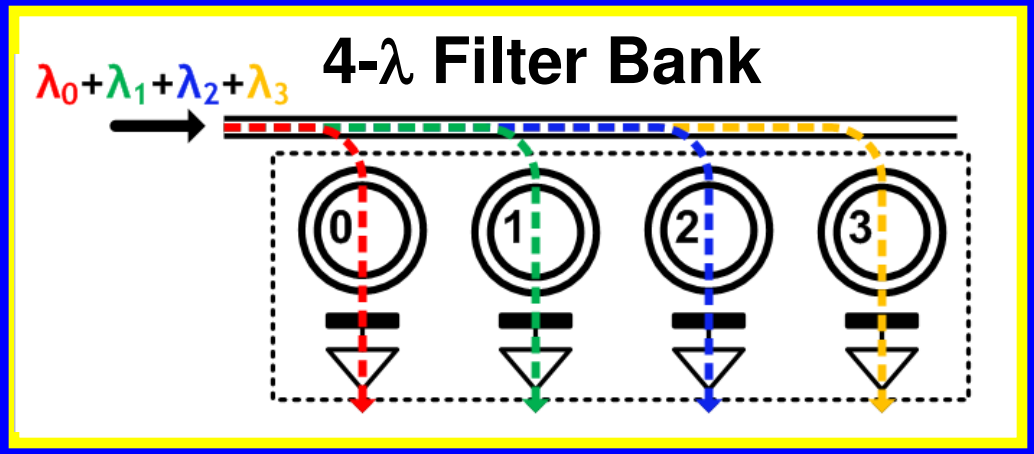
Higher clock frequency

➔ Fewer data channels, less endpoint capacitance

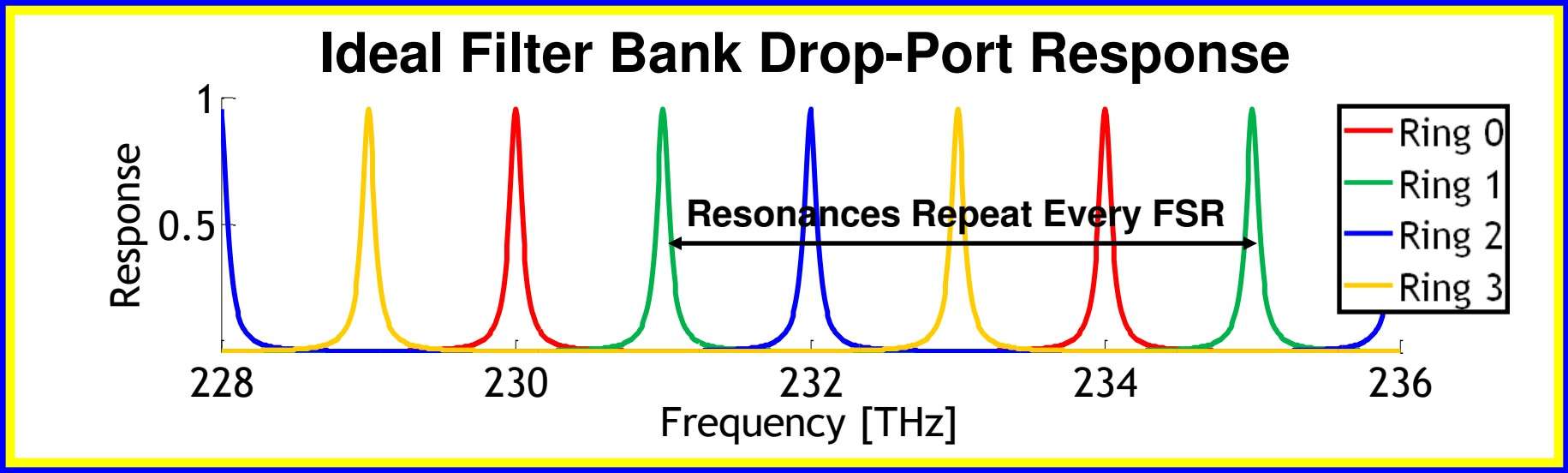
➔ Tighter timing requirements though, requiring more power

# DWDM Ring Resonance

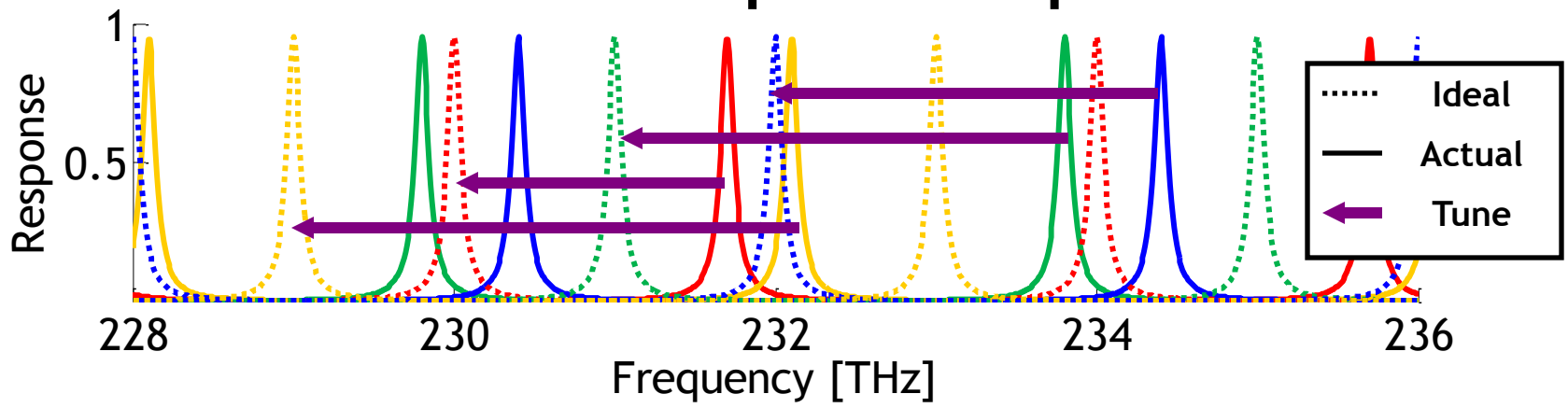
- DWDM requires ring resonances matching



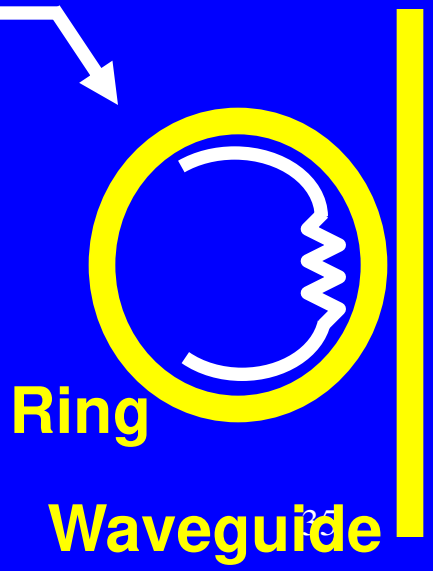

- Resonances of rings 0-3 are perfectly aligned with channel wavelengths ($\lambda_{0\text{-}3}$)
- FSR: Free Spectral Range



# 4-λ Filter Bank Tuning

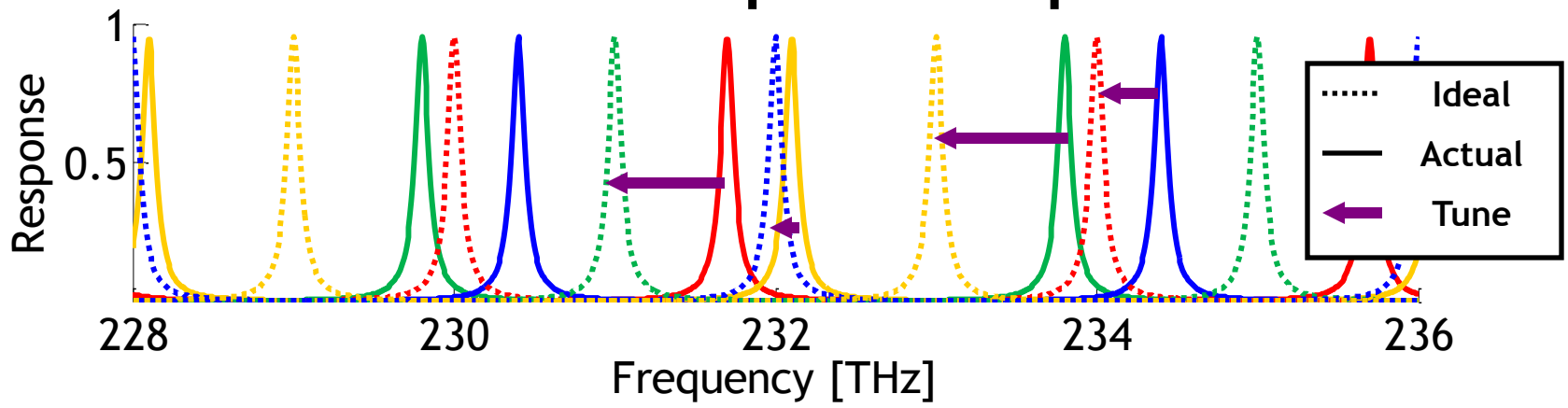



- Thermally tune rings with heaters
- Expensive with large variations
  - 600 GHz of variation requires 60K heating[1,2]
  - Heating power linear with number rings/channels
  - Cannot actively cool ring.



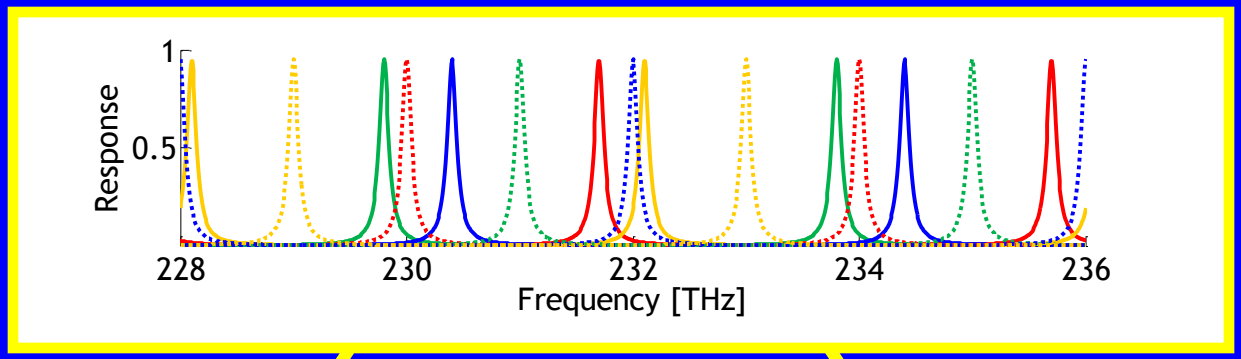

[1] Orcutt et al. Optics Express 2011
[2] Nawrocka et al. APL 2006]

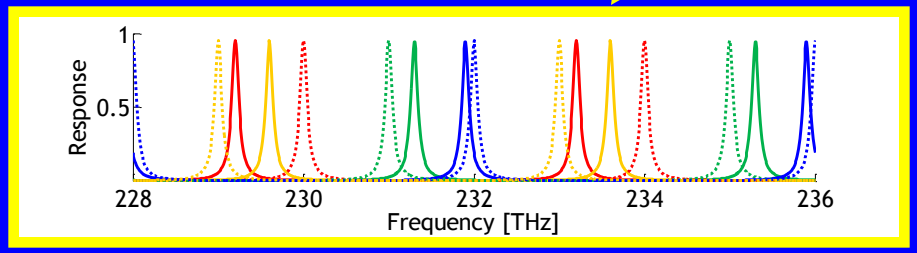
# Nearest-Channel Tuning

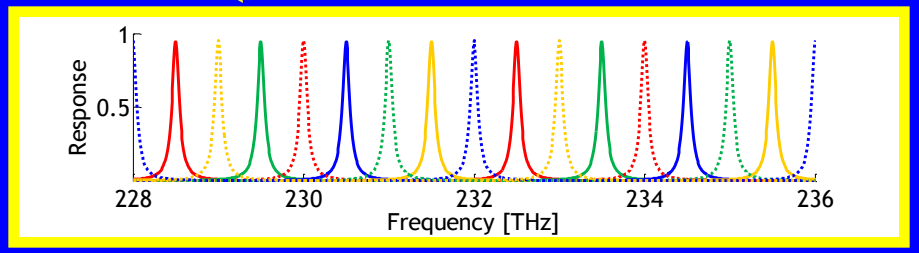



- Allow rings to just tune to the nearest channel
  - Reduces tuning range, saves heating power
  - Electrically reshuffle bit positions as opposed to assigning a permanent fixed wavelength per ring
- Build an n-to-n electrical crossbar (grows with $n^2$)
  - can we do better?

# Decoupling Local and Systematic




## Local Ring-to-Ring Mismatch

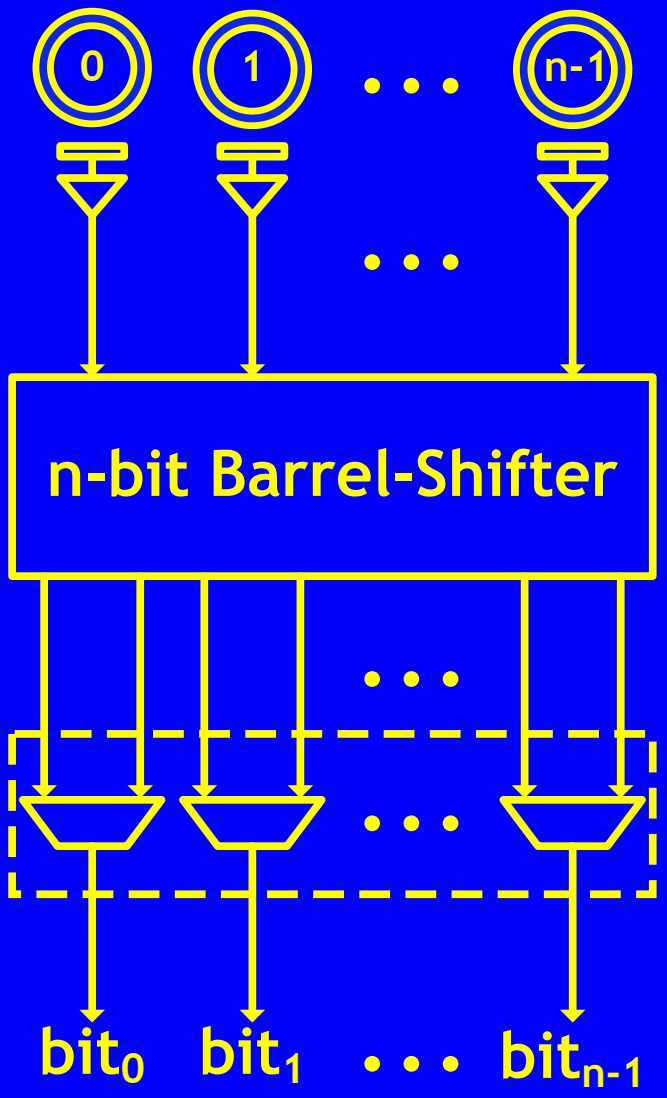
- From mostly process variations (random, time-independent)
- $\sigma$ = 20-70 GHz (0.2 - 0.5nm) [1,2]

## Systematic Mismatch

- From process and temperature
- Rings in same filter bank roughly share systematic mismatch
- Bigger in magnitude than local mismatch
- $\sigma$ = 100-300 GHz (0.6 - 2 nm)
- Deterministic, time-dependent

[1] Orcutt et al. Optics Express 2011
[2] Selvaraja et al. ECIO 2008

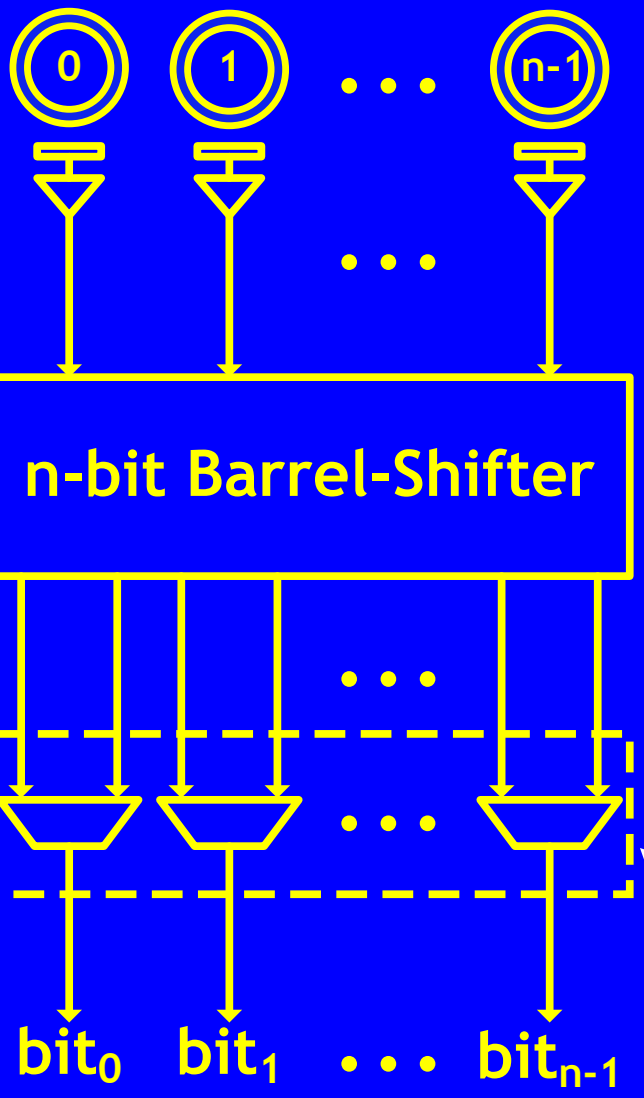
# Two-Stage Bit Reshuffler Backend




Barrel-shifter compensates for systematic mismatch affecting rings of the filter bank

Additional Multiplexers compensates for channel re-ordering due to local ring-to-ring mismatches

# Two-Stage Bit Reshuffler Backend

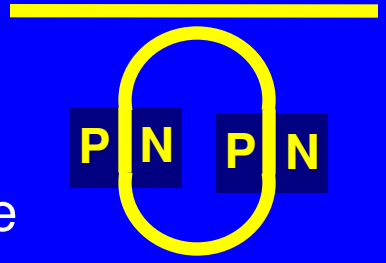



**Example:**
**System temperature increase due to core activity**

- resonances in bank all shift the in the same direction
- barrel-shift channels to re-align
- mux unchanged

# Electrical Tuning Assistance

- A reverse-biased modulator can also tune
  - no static power, fast tuning ☺
  - Limited tuning range (tens of GHz) ☹
    - If no reshuffling, heaters can bridge the extra distance

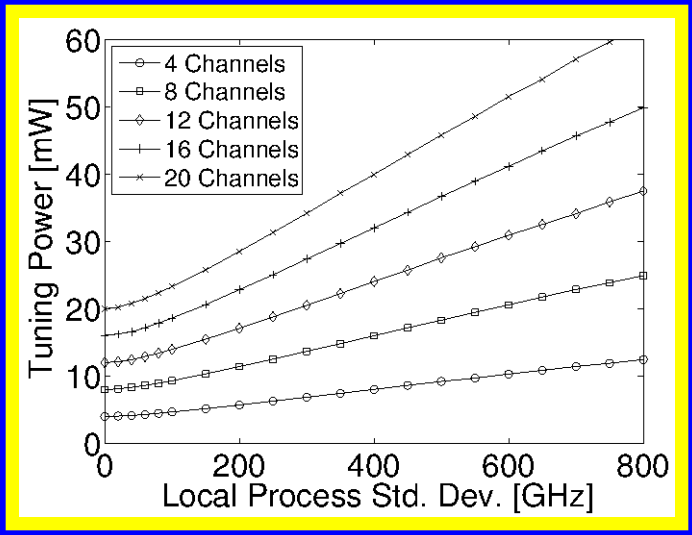



- Reshuffling backend makes tuning range:
  - invariant of local and systematic variations
  - proportional to channel separation, decreases with the number of channels

**Electrically-assisted tuning with reshuffling is a powerful tuning tool.**

# Tuning Efficiency

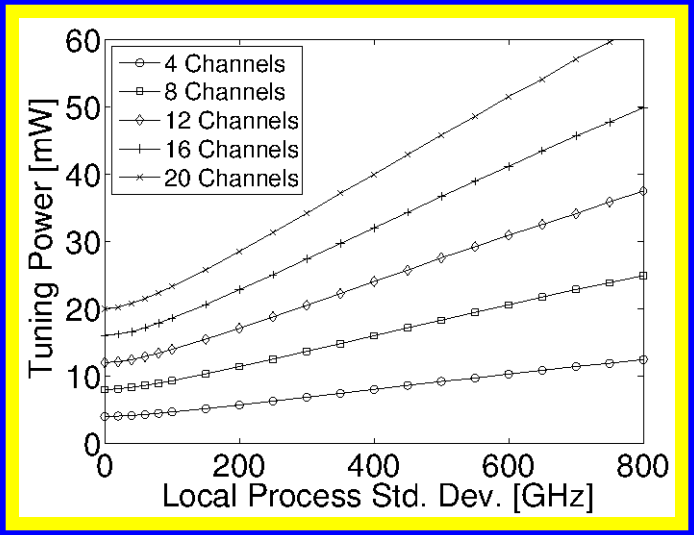
## Thermal Tuning




Power increases with variation since we are tuning each ring to a specific resonance.

# Tuning Efficiency

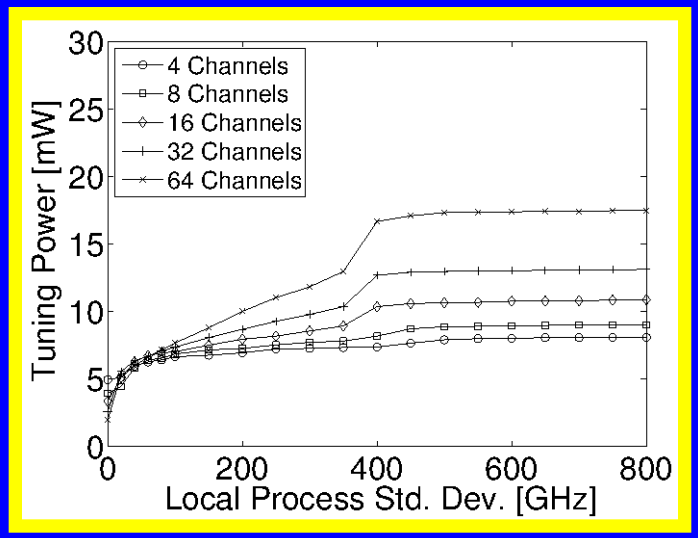
**Thermal Tuning**




- Power increases with variation since we are tuning each ring to a specific resonance.

Lower, flatter power with increased **local variation**

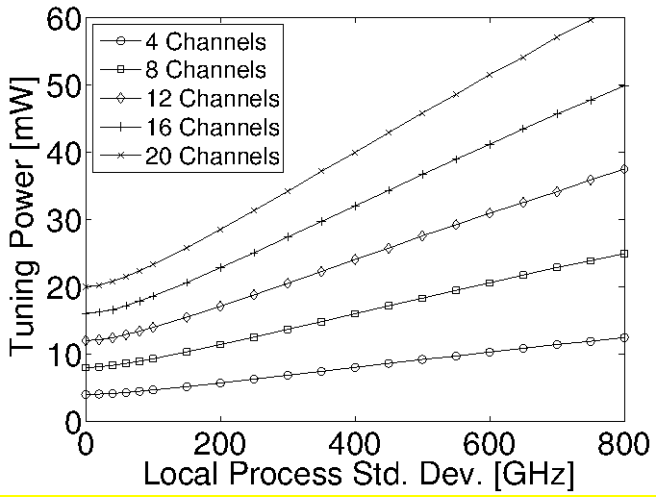
**Electrical Tune with Bit Reshuffle**

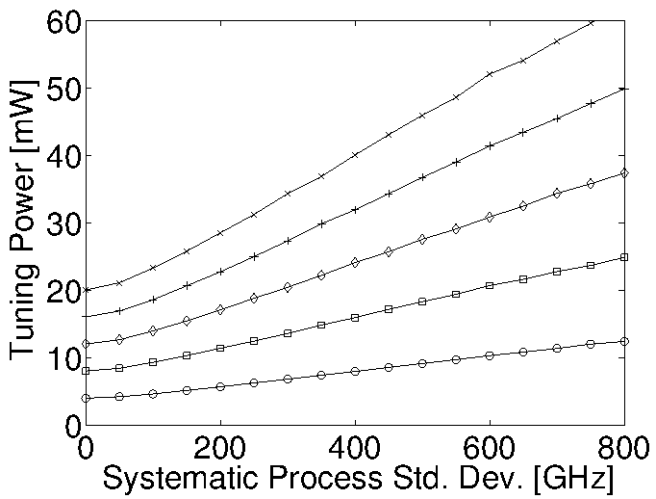



- More efficient tuning mechanism
- Only tuning to nearest channel

# Tuning Efficiency

**Thermal Tuning**

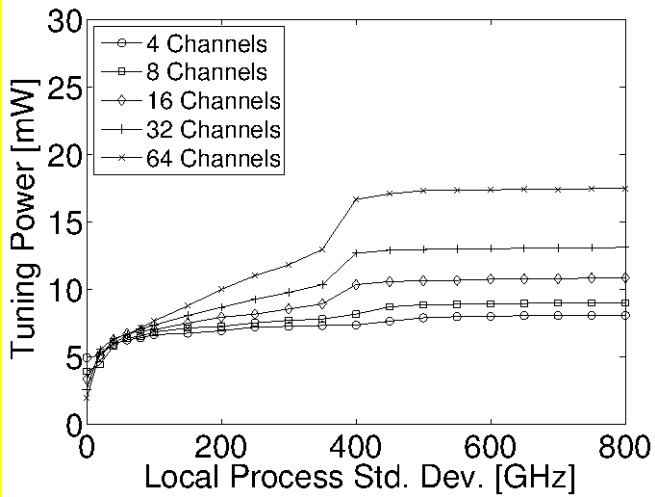


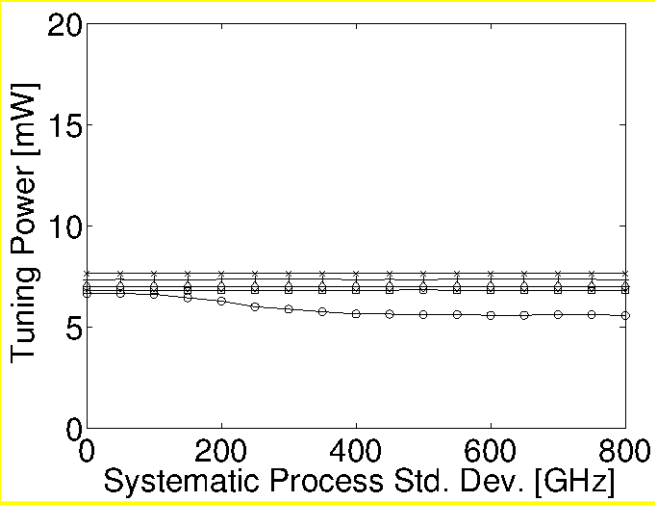




Lower, flatter power with increased **local variation.**

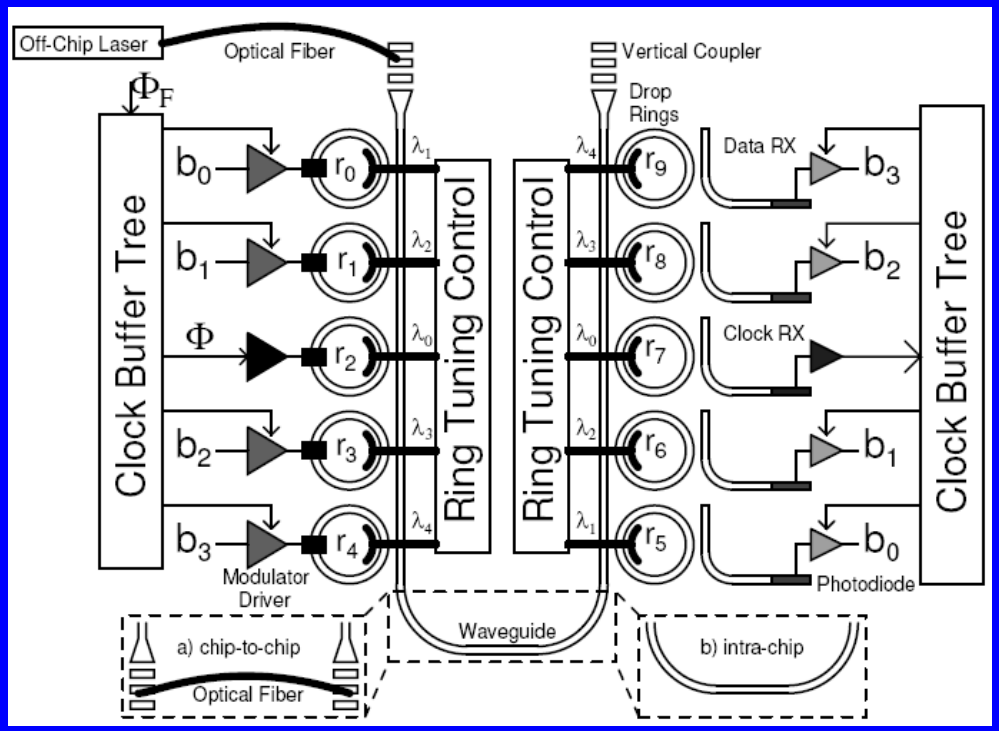
Improvement more dramatic for **systematic variation.**

**Electrical Tune with Bit Reshuffle**



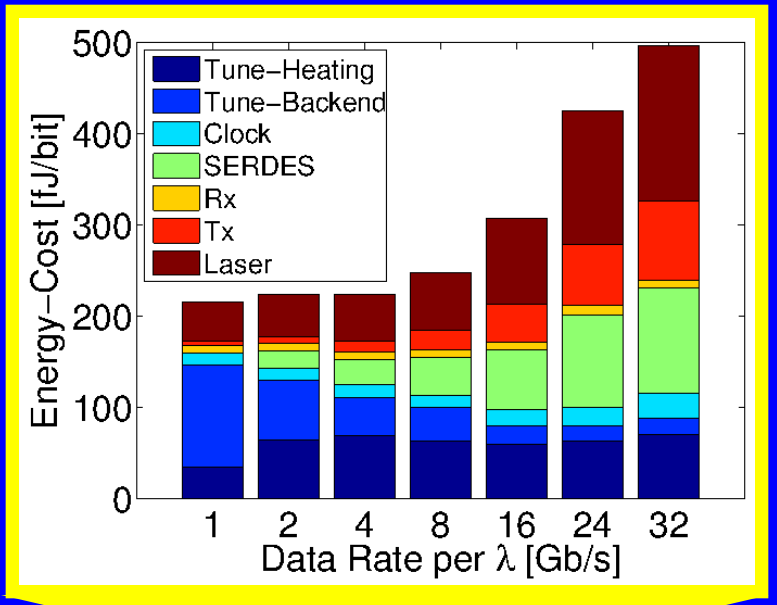



# Full WDM Photonic Link Analysis

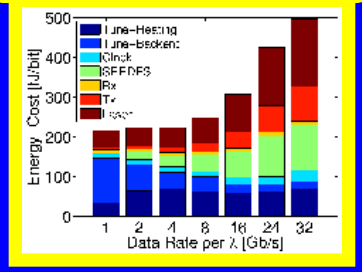



- Tie together all photonic components in order to gain intuition on system budgeting.
- For different throughputs, look across data-rate per wavelength-channel

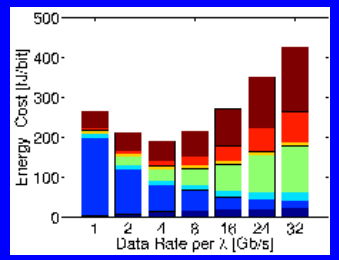
# WDM Photonic Link Evaluation



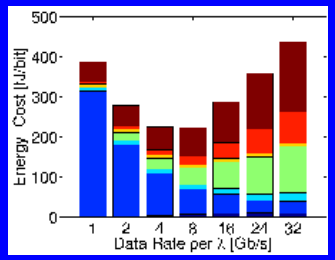

- **Trend is similar to a single link due to low number of rings and λ**
- **But, tuning power kicking in**



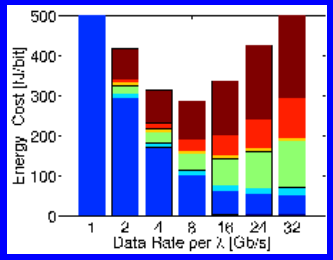

**64Gbps**




**256Gbps**



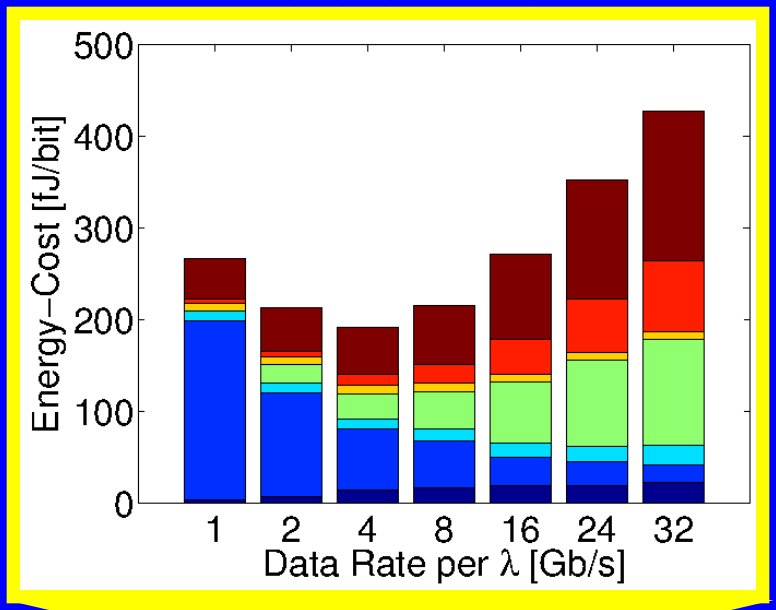

**512Gbps**

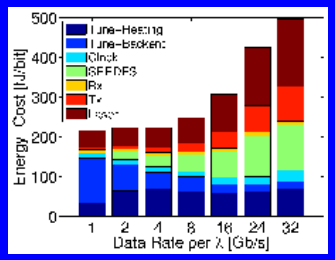



**1024Gbps**

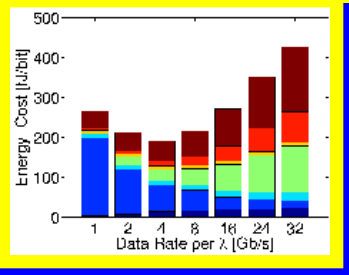
Increasing Number of Rings and λ

# WDM Photonic Link Evaluation

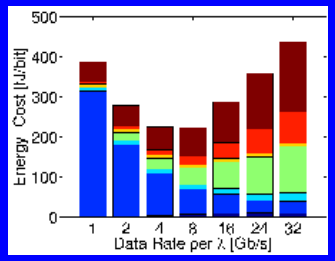


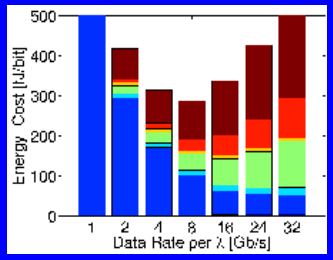

64Gbps




256Gbps



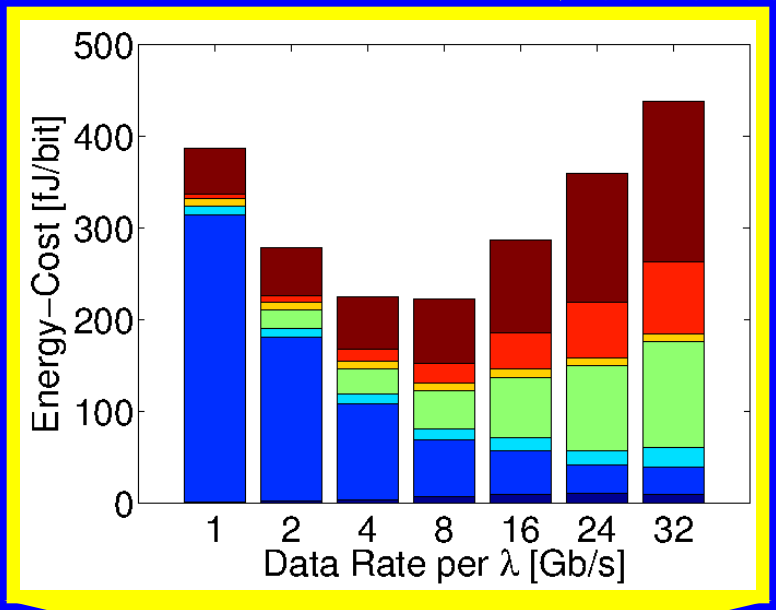

512Gbps

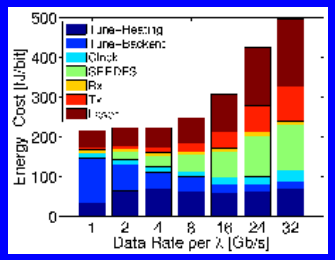



1024Gbps

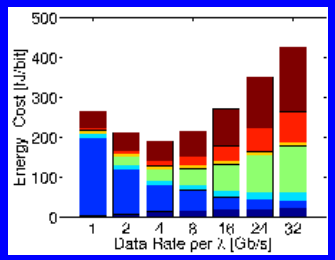
Increasing Number of Rings and λ

# WDM Photonic Link Evaluation

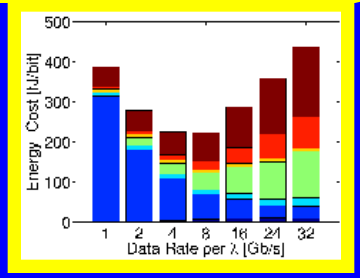


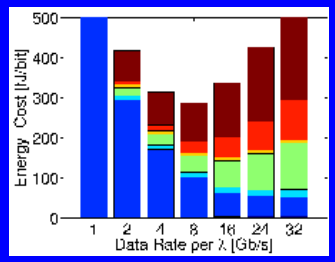

64Gbps




256Gbps




512Gbps

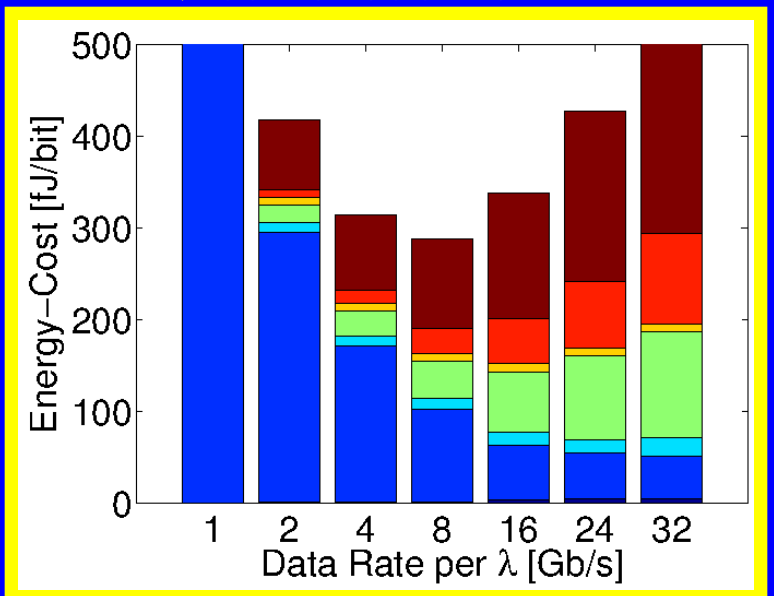


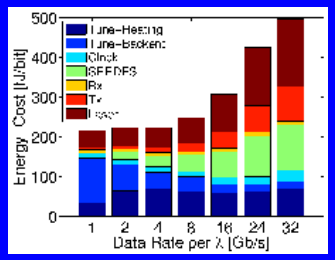

1024Gbps

Increasing Number of Rings and λ

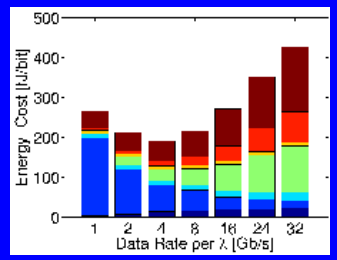
# WDM Photonic Link Evaluation

- **Electrical backend ring tuning cost very high at low rates due to large number of rings**

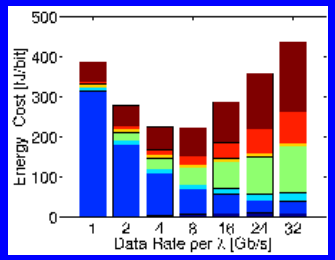


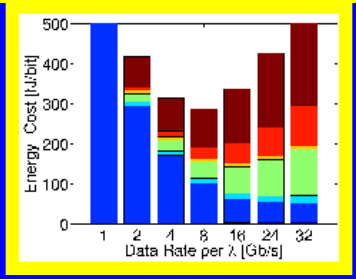

64Gbps




256Gbps



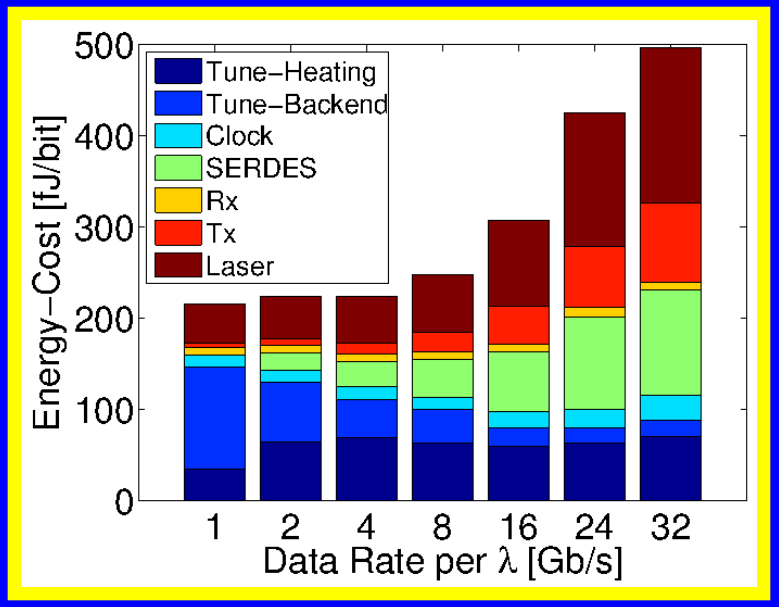

512Gbps



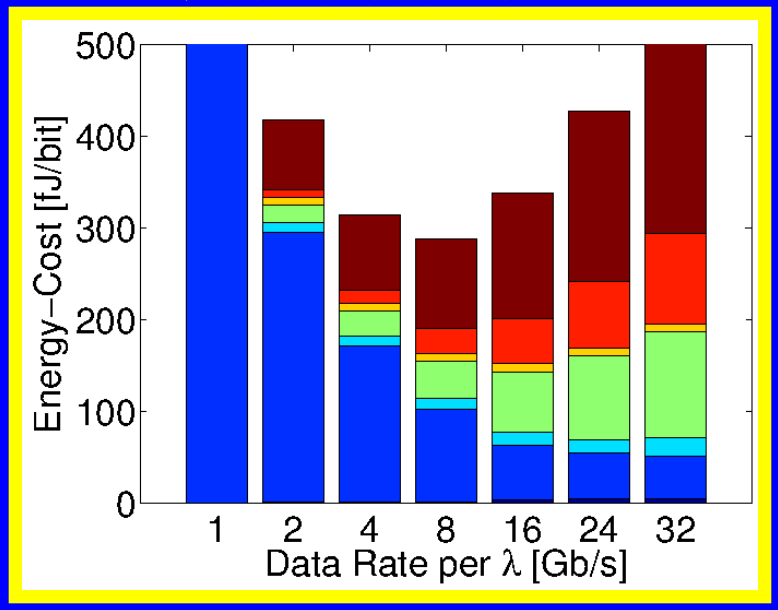

1024Gbps

Increasing Number of Rings and λ

# WDM Photonic Link Evaluation




64Gbps

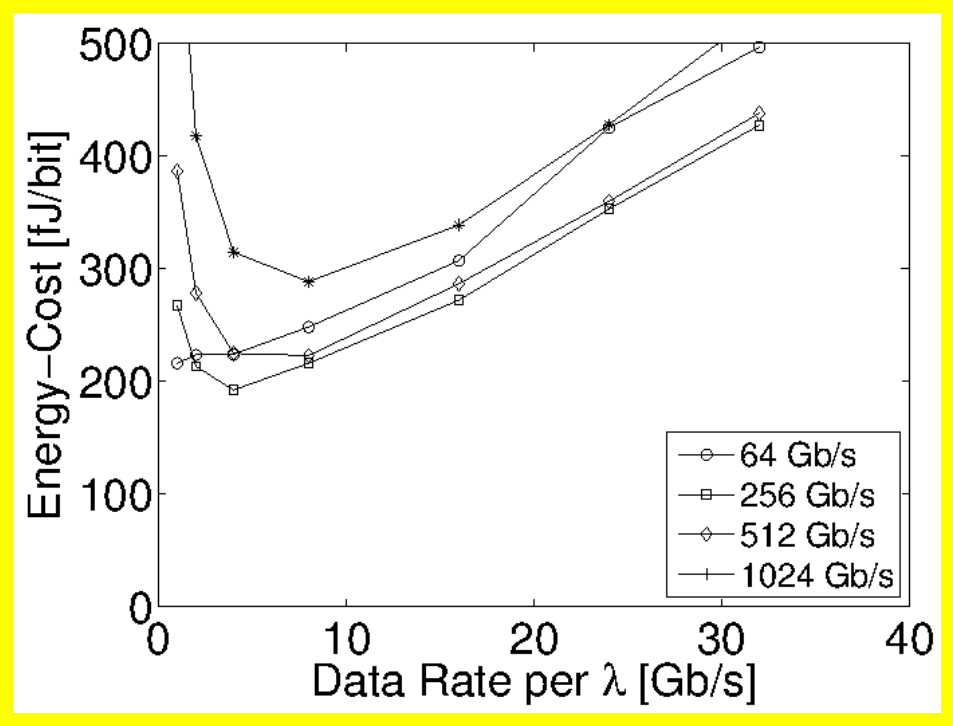



1024Gbps

- With electronics, typically run at low rates for energy-efficiency
- WDM actually *lets* us run at low rates while maintaining throughput
- BUT due to ring-tuning (unique to photonics), we now don't want to

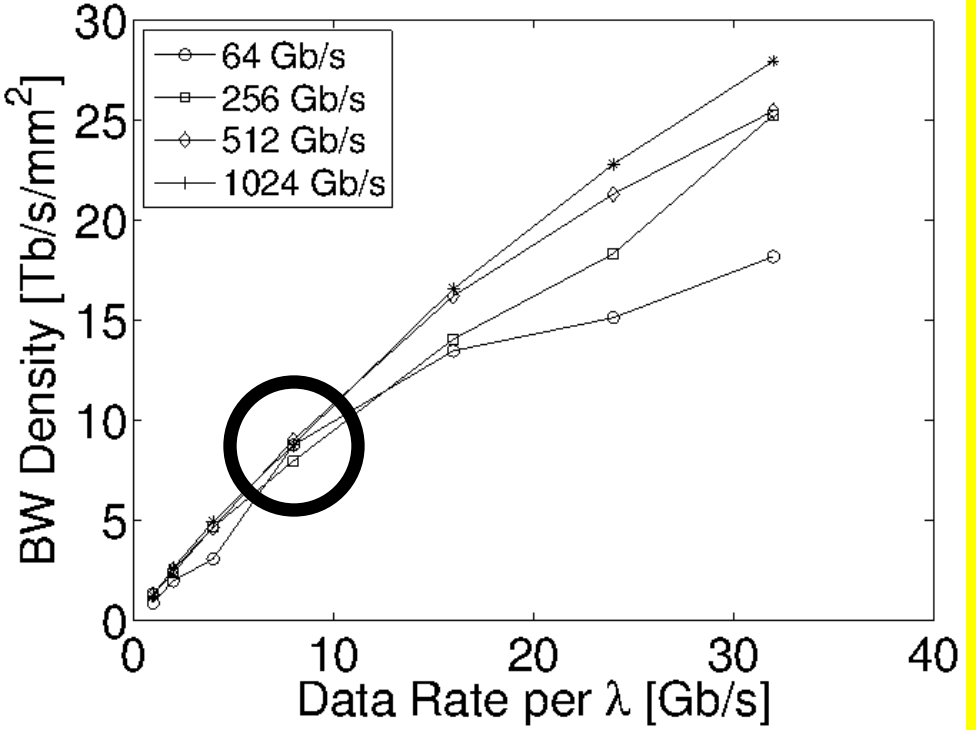
Increasing Number of Rings and λ

# WDM Photonic Link Evaluation




- **Optimal data-rate per channel is throughput-dependent**
- **In contrast to common view, optimal data-rates are all relatively low at <10Gb/s**
- **Next, check bandwidth-density**

# WDM Photonic Link Evaluation




| | Electrical | Photonic |
|---|---|---|
| Die | 500Gb/s/mm$^2$ | 10Tb/s/mm$^2$ |
| Package | 25 Gb/s/mm$^2$ | 100Tb/s/mm$^2$ |

- **BW limited at die by component density at 10Tb/s/mm$^2$**
- **Photonics still 200-400X better than electrical**

# Conclusion

- Photonic interconnects hold promise to meet future compute system communication needs
- To understand photonic system design, we need cross-layer system optimization:
  - Balance component specifications at the system-level for best bandwidth-density and energy-efficiency
  - Use insight to set the technology trends and device specifications
- Monolithic integration and moderate-data-rate DWDM is most energy-efficient while maintaining significant bandwidth-density advantages.

# Acknowledgements


- This project is a highly collaborative effort with teams at MIT, UC Boulder, and UC Berkeley:
  - Hanqing Li, Karan Mehta, Jason Orcutt, Jeff Shainline, Jie Sun, Erman Timurdogan, Stevan Urosevic, Matthew Weaver, Prof. Milos Popovic, Prof. Rajeev Ram, Prof. Michael Watts, Prof. Krste Asanovic
- The work was supported in part by MIT CICS, DARPA, NSF, FCRP IFC, Trusted Foundry, APIC, Intel, and NSERC.

# Link Evaluation Parameters

TABLE I
LINK EVALUATION PARAMETERS

| Parameter | Value |
|---|---|
| Process Node | 32 nm Bulk CMOS |
| $V_{DD}$ | 1.0 V |
| Device to Circuit Parasitic Cap $C_P$ | 5-25 fF |
| Wavelength Band $\lambda_0$ | 1300 nm |
| Photodiode Responsivity | 1.1 A/W |
| Wall-plug Laser Efficiency $P_{laser}/P_{elec}$ | 0.3 |
| Channel Loss | 10-15 dB |
| Insertion Loss $IL_{dB}$ (Optimized) | 0.05-5.0 dB |
| Extinction Ratio $ER_{dB}$ (Optimized) | 0.01-10 dB |
| Bit Error Rate (BER) | $10^{-15}$ |
| Core Frequency | 1 GHz |
| SERDES Topology | Mux/Demux Tree |