Si Photonics for Software Defined Data Centers

SAND2014-1767C

Sandia National Laboratories

Exceptional service in the national interest

Panel 2: Market, Standards, and Research Drivers for Software Defined Photonic Networking

Anthony Lentine,

Sandia National Labs,

Albuquerque NM 87185

alentine@sandia.gov

OIDA workshop, 12/09/2014

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

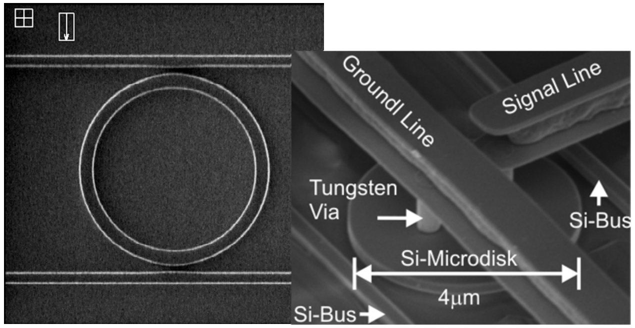
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. SAND NO. 2014-0685 C

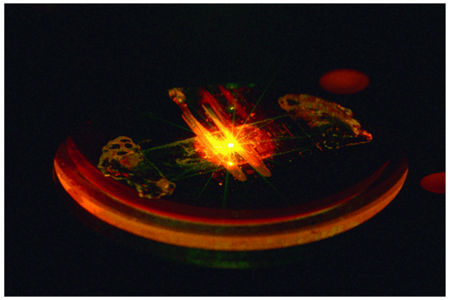
Outline

- Optical Interconnects
  - Silicon Photonics, short term vs. long term.
- Optical Networks
  - Provisioning (slow configurations) (Virtualization)
  - Routing (fast reconfiguration)
  - Silicon Photonics
    - Feasibility, Scalability, Technical Challenges
    - Hardware interface (software interface)
- There is a lot of conjecture in this presentation – to facilitate thought and discussion – not to discuss our work.

Plausible Transceiver Evolution: 100GbE – 10 TbE (2013-2030)

Sandia National Laboratories





Data Center Goals

- Low cost
- Small form factor
- High density
- Low power
- Low fiber cost
  SM, MM, MC, MPO
- Reach to 1km

100Gbps: SiP

- 4 CWDM @ 25G
- 8 PAM @ 33G

1Tbps:

- 8 CWDM, 8 PAM@42G
- DWDM: 40λ @ 25G

10 Tbps:

- 80 λ/8PAM@42G
  (1 fiber per direction)

100Gbps: VCSEL

- 8 (12) MPO@ 25G
- 4 CWDM @ 25G

1Tbps:

- 25 Multicore @ 40G
- 8 PAM, 12 MPO @ 56G

10 Tbps:

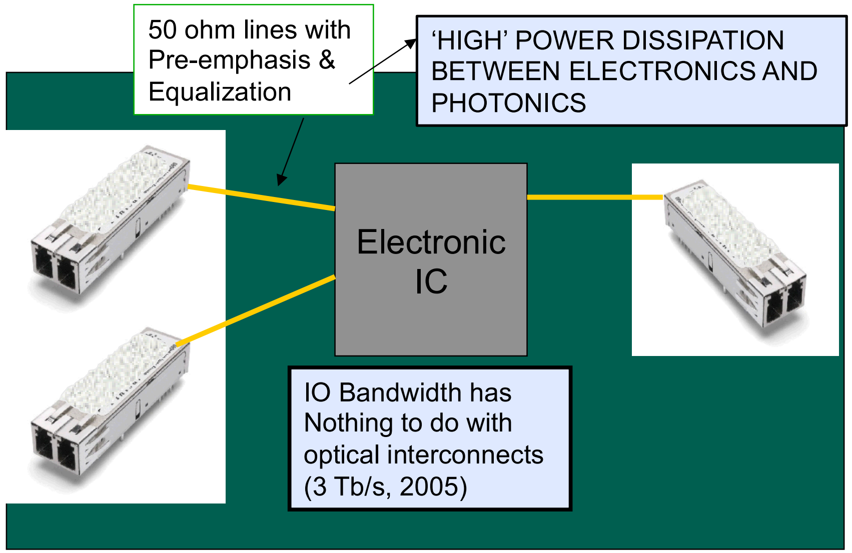
- 16 Multicore, 12 MPO,
  (192 cores) 8 PAM @ 35G

Optical Interconnects

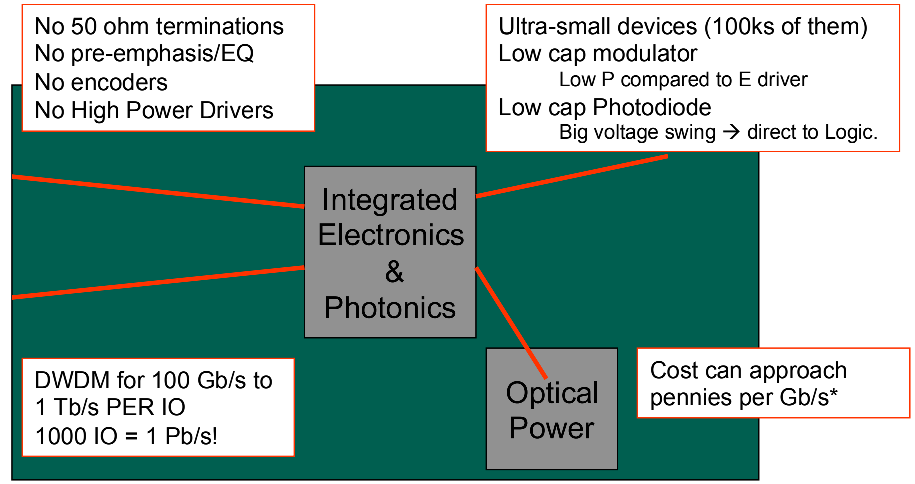
Sandia National Laboratories

- **Evolutionary (Modules)**
  - GbE and 10GbE Products
  - 100 GbE modules soon w/ VCSELs and Si Photonics
  - TbE modules on the horizon

- **Revolutionary (3DI)**
  - **Higher bandwidth density**
    - **DWDM is required!!**
  - **Drastic *potential* power reduction**
    - **No 50 Ω lines, pre-emphasis or equalization**
    - **Receiver has high transimpedance, few gain stages**
    - **Shared CDR (less delay variation and jitter)**



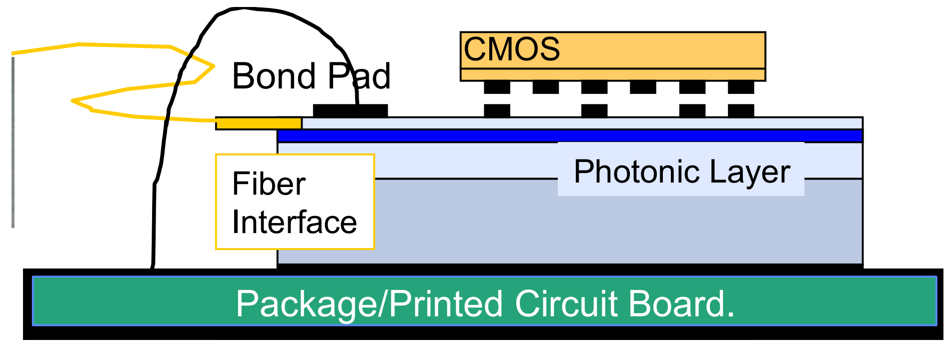
OPTICS FOR DISTANCE

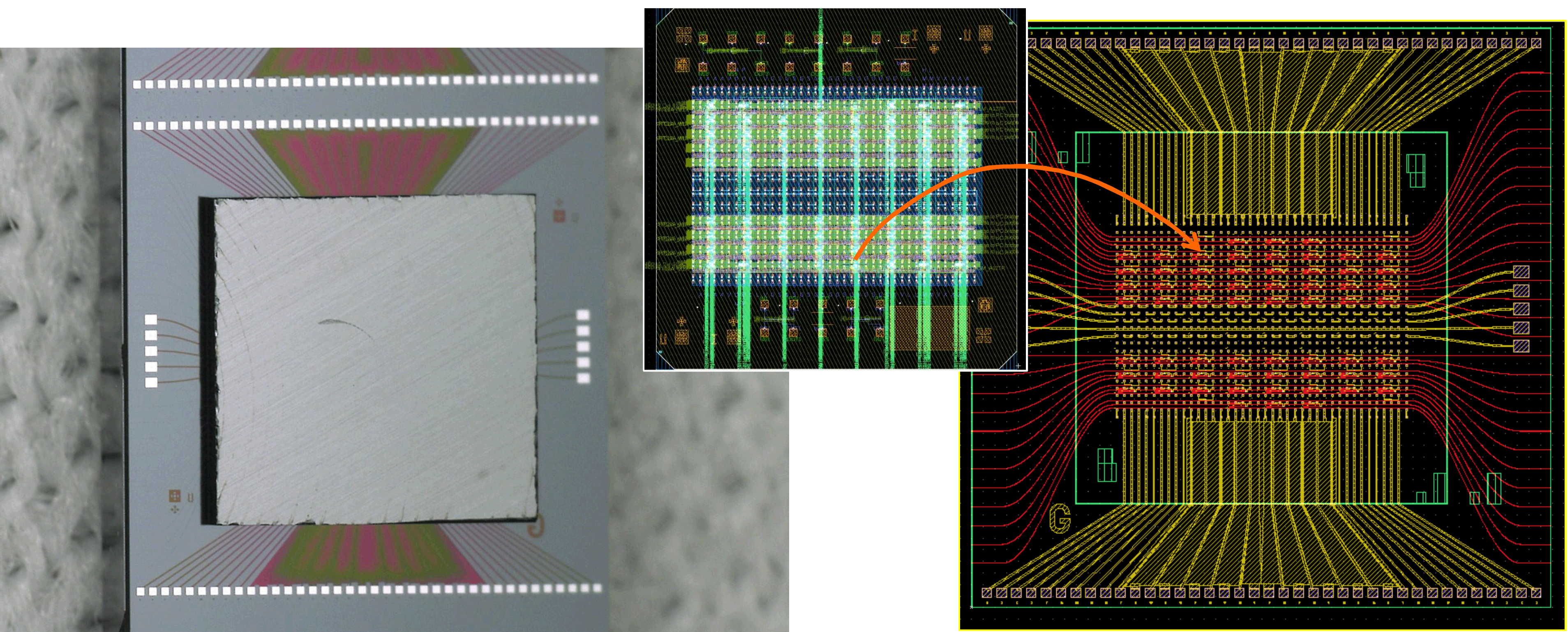


OPTICS FOR *LOW POWER*, *HIGH BANDWIDTH DENSITY*, COST, SIZE, WEIGHT, DISTANCE

Electronic-Photonics Integration

- Heterogeneous integration
  - Independent optimization of electronics & photonics
  - Need very high yields and small size





Si Photonics Optical Interconnects

- Modulators, integrated Ge detectors > 40 Gbps (research)
- Parallel channel modules in production (AOCs)
- 4 x 25 Gbps WDM modules announced last OFC
- Mach-Zehender modulators, 'large' filter technology

- No micro-ring resonator products (no my knowledge)
- Demonstrations > 40 Gbps
- 1 fJ/bit modulators
- Capable of few fJ/bit receivers (Ge detectors)
- Resonant wavelength control bench top demonstrations
- Heterogeneous and monolithic integration with CMOS

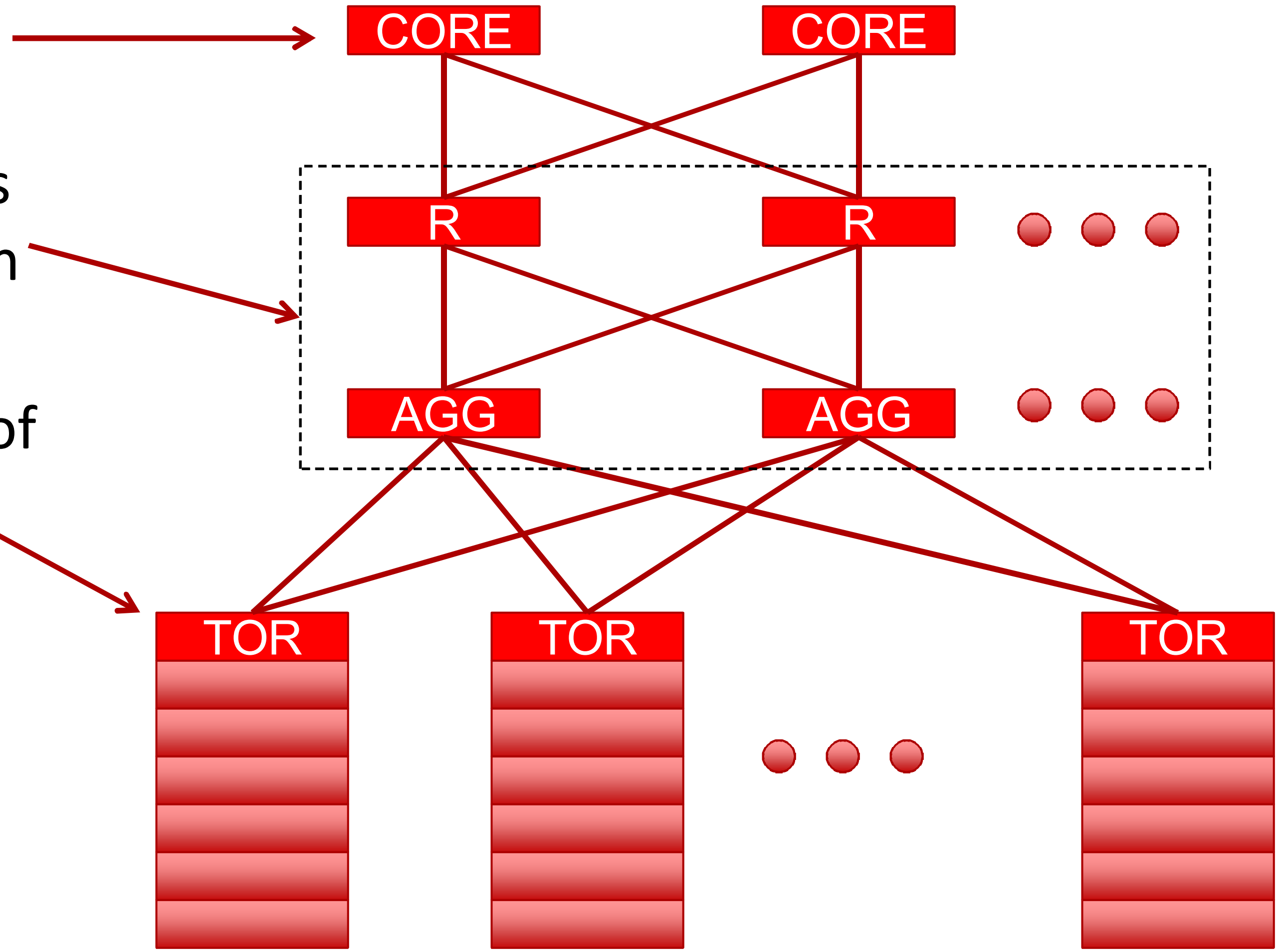
Beyond Optical Interconnect

- Various technologies exist for optical transceivers
  - Silicon Photonics, VCSELs & III-V integrated optoelectronics
- Intimate integration with high-value electronics
  - Lower power and higher bandwidth density, lower cost?

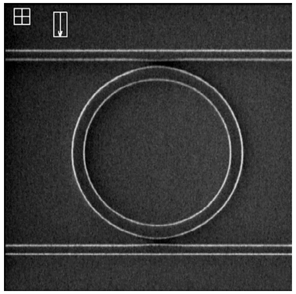
- From a networking perspective, optical interconnects aren't *that* interesting
- More interesting are routing functions
  - Passive routing (wavelength intermixing)
  - Active provisioning (not reconfigured often)
  - 'Flow' or packet routing (reconfigured every ps/ns/us)

Data center hardware architecture

- Interface to external network
- One or more levels of routing between racks
- Servers have 'top of rack' switch to route within rack and interface to inter-rack

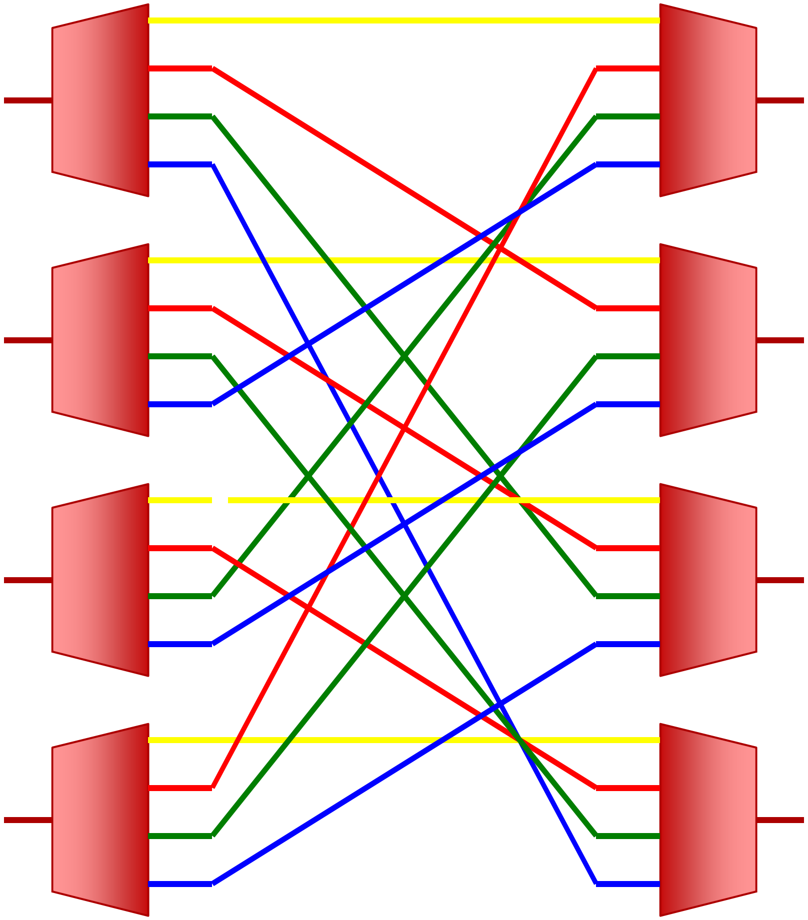


Passive Routing
Hi Radix Switch interfaces



Sandia National Laboratories

- Simple Passive Mux demux's with waveguide routing
- Allows large output from one switch to go to many places
- Implementation:
  - AWGs,
  - Thin film filters
  - Silicon Photonics
    - * more on this later …

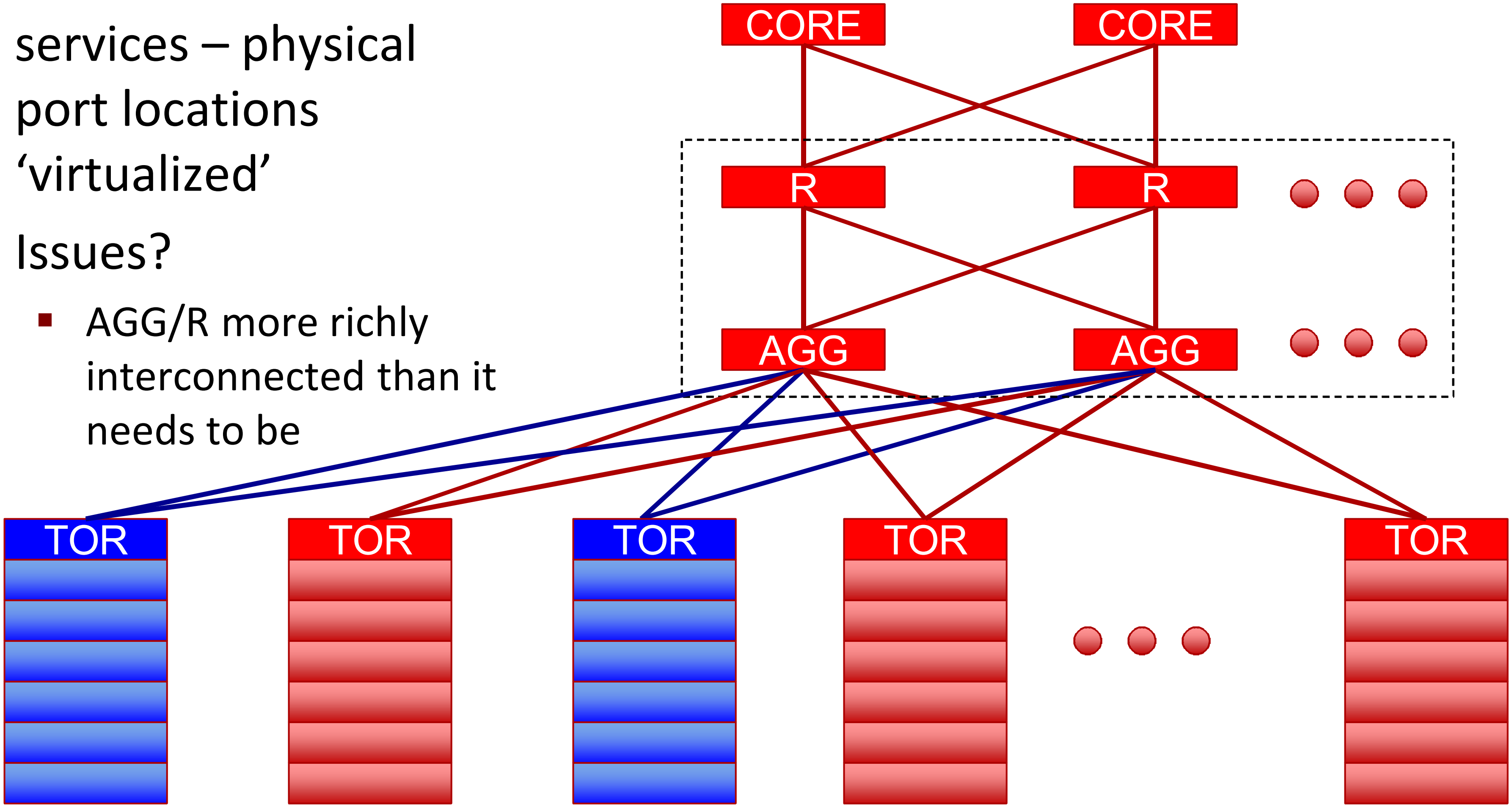


Software Defined Data Centers

- Virtualization
  - Shared data centers
- Software defined networking
  - Known path routing
- Optical switching and routing

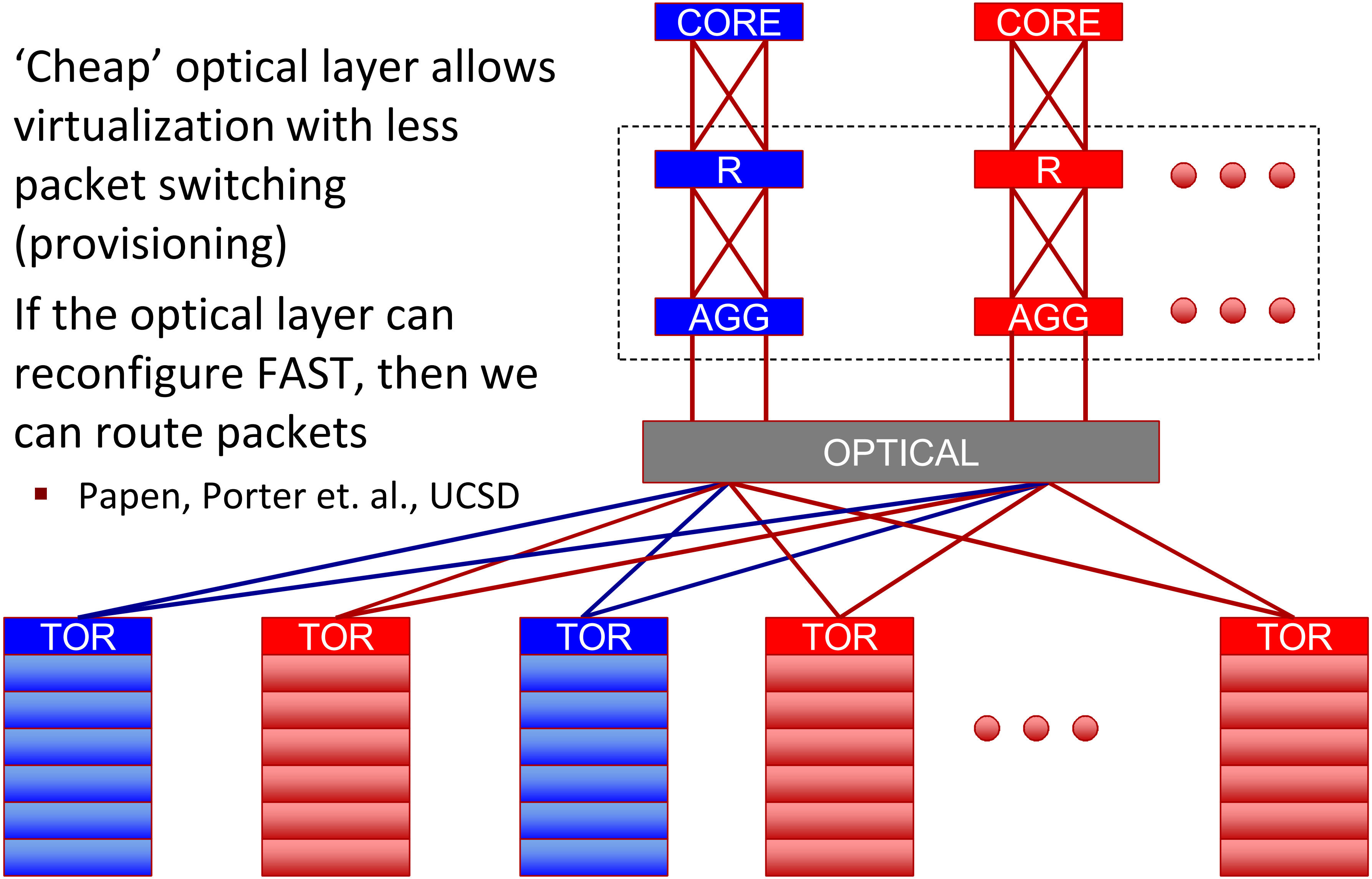
Data center virtualization

- Offer data center services – physical port locations 'virtualized'
- Issues?
  - AGG/R more richly interconnected than it needs to be



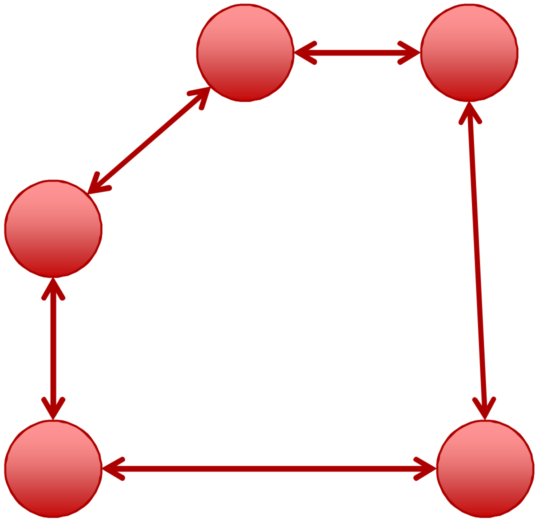
Data center virtualization

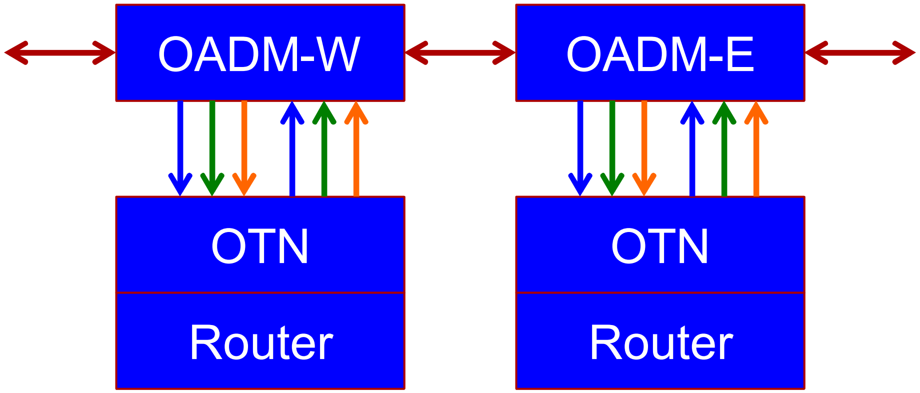
- ‘Cheap’ optical layer allows virtualization with less packet switching (provisioning)
- If the optical layer can reconfigure FAST, then we can route packets
  - Papen, Porter et. al., UCSD



Metro-regional transport

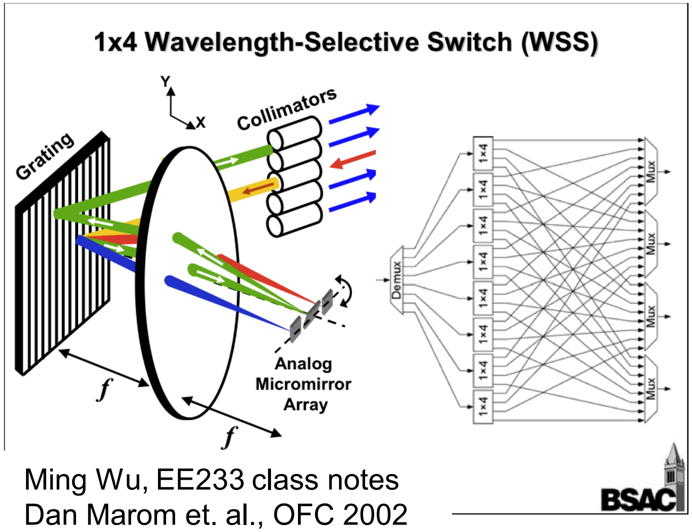
- Network is physically in rings
- ROADMs allow reconfigurable bandwidth between routers
- Wavelengths allow all-to-all node connectivity
- OTN cross-connects allow sub-wavelength granularity
- The optical network isn't switching and routing
  – it's provisioning bandwidth
- Can we apply something like this to data centers
- Why?



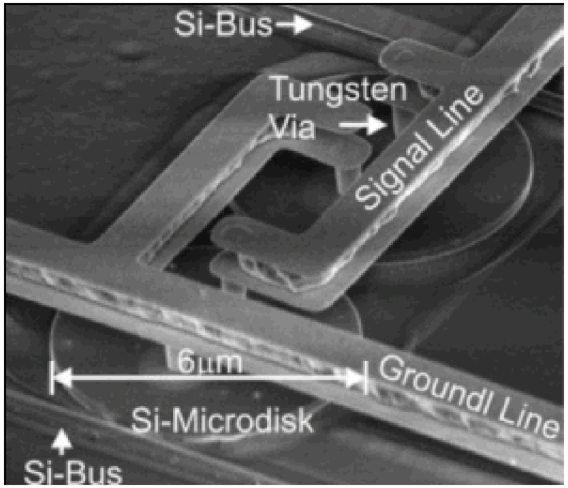


Active switching technology choices

- MEMs, Liquid Crystal (WSS)
  - 1:N (good!) or 2 x 2 (bad)
  - Slow (1 us – 1 ms)
  - Often free-space (grating for WSS) {expensive}
  - Fairly scalable to large sizes?
    - 80 λ x 1 x 9
  - Flex bandwidth
  - Products
- Integrated Optics (Silicon Photonics, III-V)
  - 2 x 2 (bad)
  - Slow (1 us) or very fast (<100ps)
  - Scalable (with more maturation)
  - Flex bandwidth
  - Research

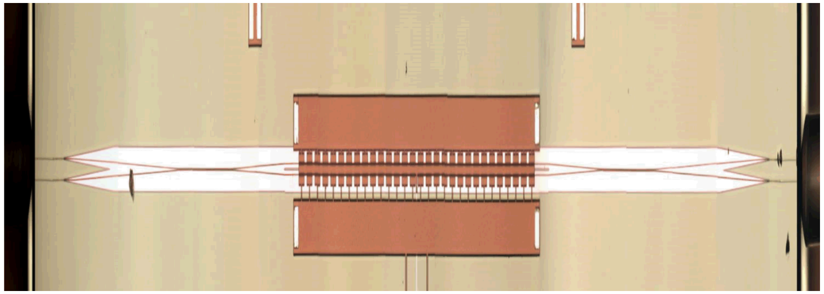


Ming Wu, EE233 class notes
Dan Marom et. al., OFC 2002

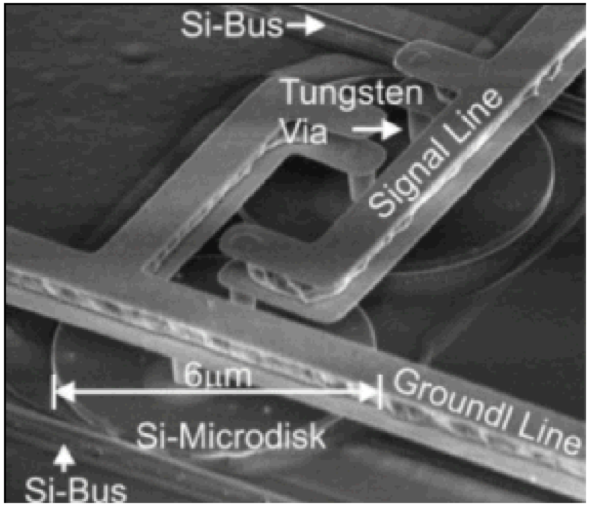


M. Watts et. al., Group IV Photonics 2008

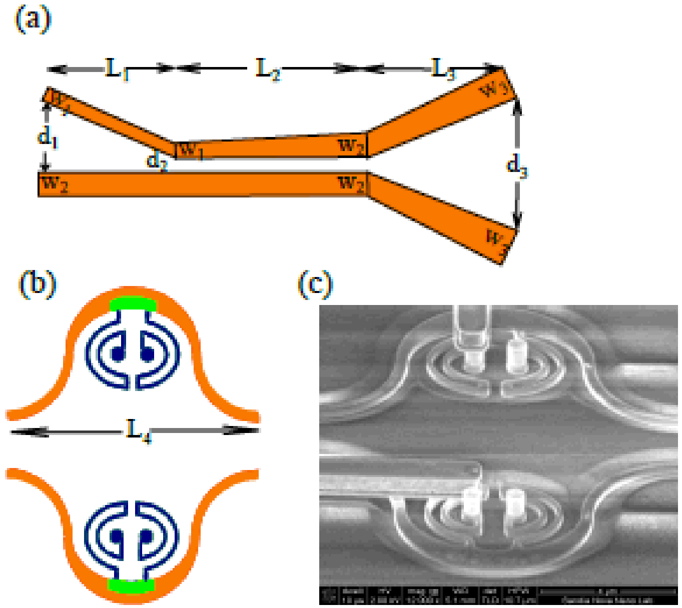
2 x 2 silicon photonics switches



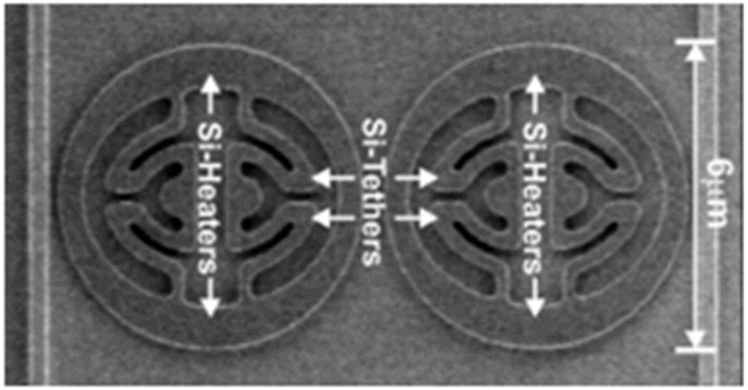
MZ – free carrier effect



MR – free carrier effect

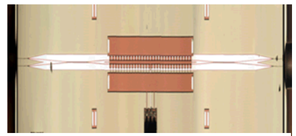


MZ – thermo-optic



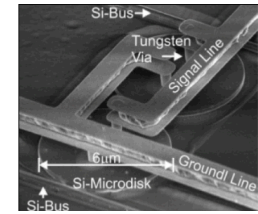
MR – thermo-optic

2 x 2 silicon photonics switches



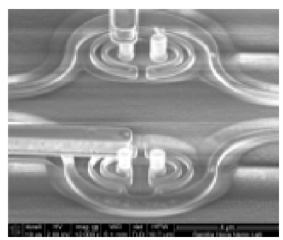
- Fast (< 100ps)
- Broadband
- 1pJ/switching event
- 1 mm size
- No static power

MZ – free carrier effect



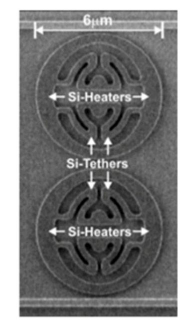
- Fast (< 100ps)
- Wavelength selective
- 1fJ/switching event
- No static power
- < 10 um size

Ring – free carrier effect



- Slow (10 us)
- Broadband
- 15 mW/GHz (2-pi)
- Static power in one state
- < 10 um size + coupler

MZ – thermo-optic

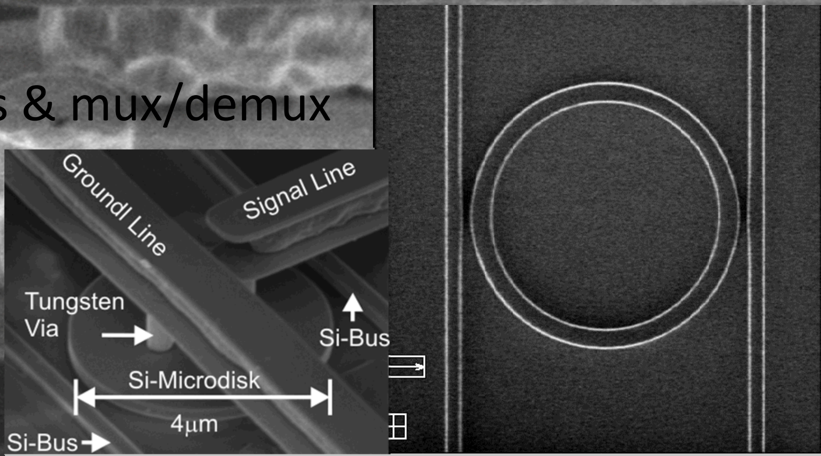


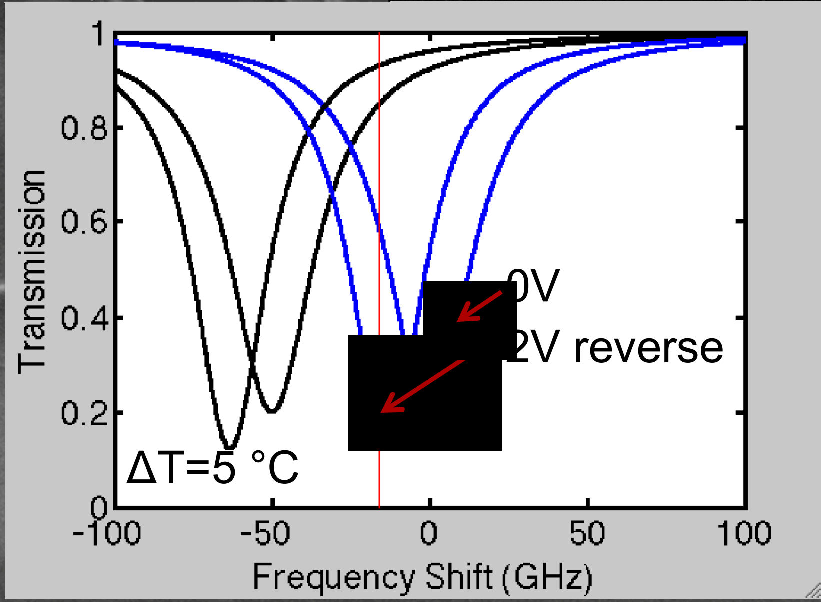
- Slow (10 us)
- Wavelength selective
- 4 uW/GHz (200uW)
- Static power in one state
- < 10 um size

Ring – thermo-optic

Resonant silicon micro-photonics

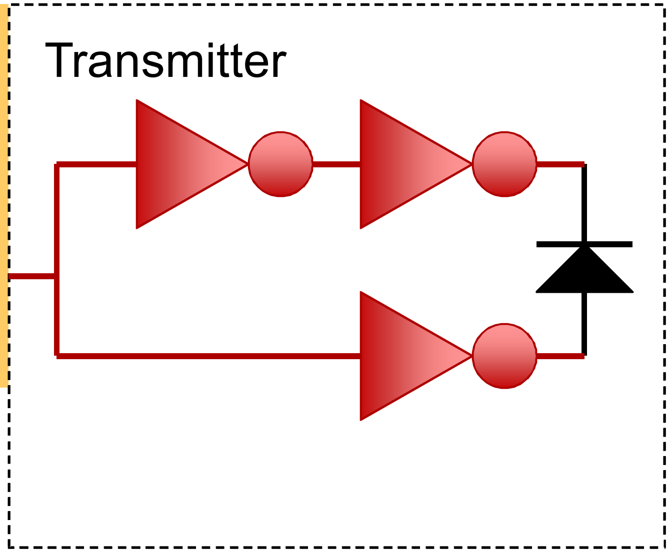
- Why resonant silicon photonics?
  - Small size (<4 um dia.)
  - Resonant frequency → DWDM modulators & mux/demux
- Benefits
  - Low energy
  - High bandwidth density
- Resonant Variations
  - Manufacturing Variations
  - Temperature Variations
  - Optical Power (1s density)
  - Aging?
- Requirements:
  - Resolution: +/- 0.25 ° C (depending)
  - Range: 10 – 85 ° C (depending)

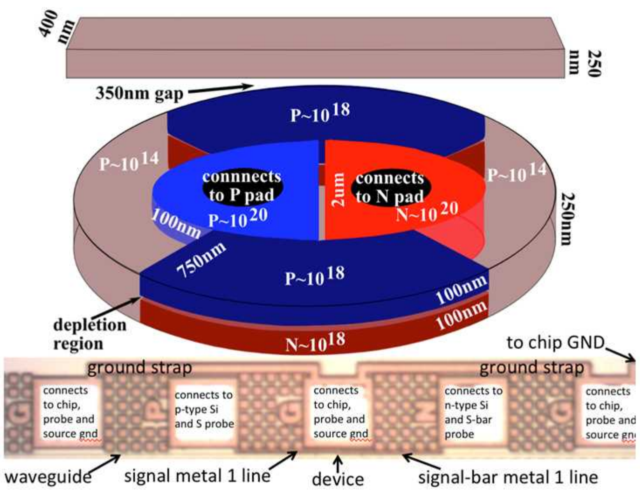


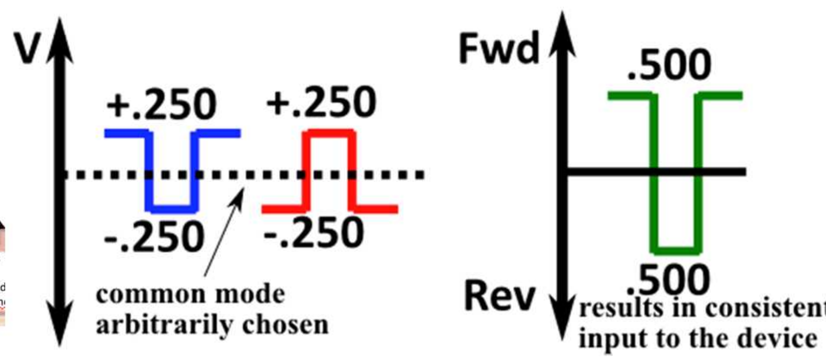


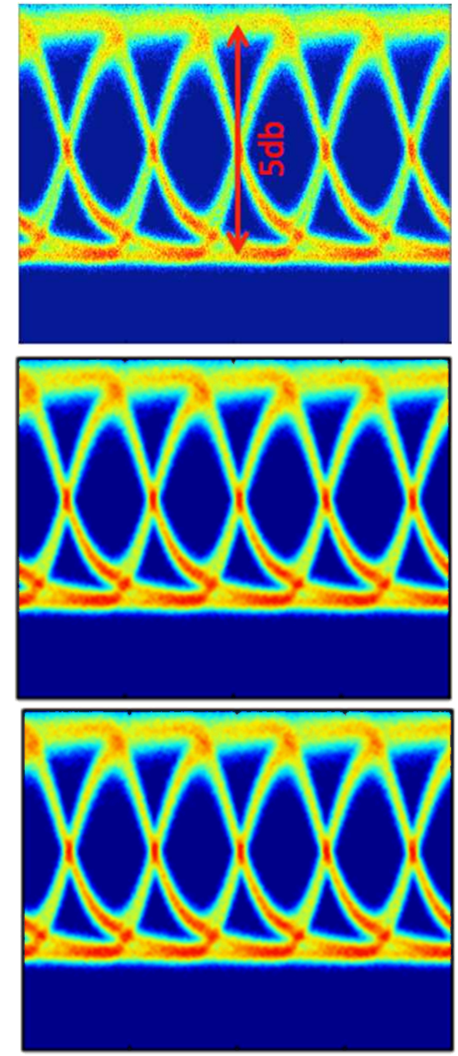
Simple ~~Modulator~~ Switch Driver: Differential Signaling

- No pre-emphasis
- No AC coupling
- No high voltages
- CMOS logic levels





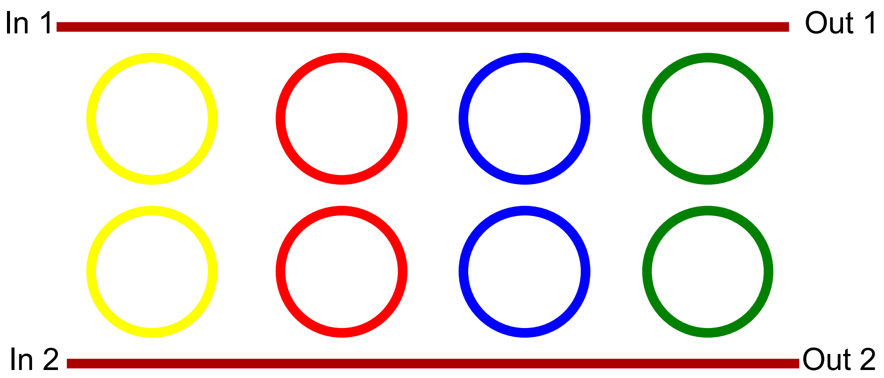


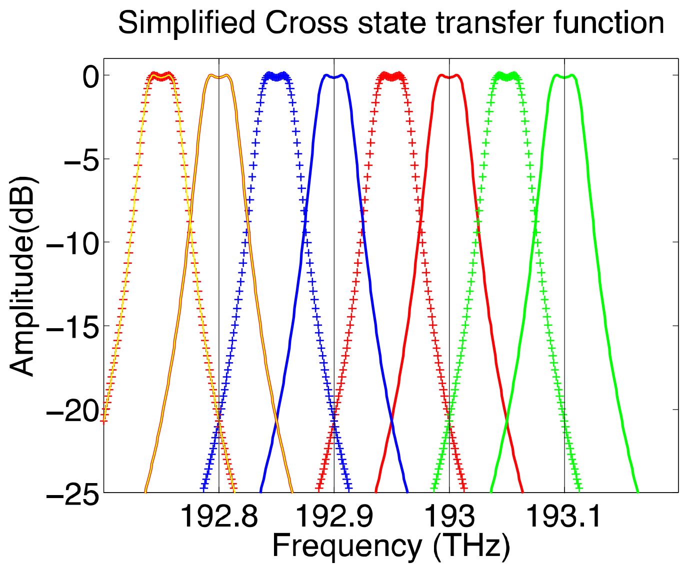


- 10 Gb/s
- Common Mode:
- .25V, .8V, 1.2V
- 3 fJ/bit

W. A. Zortman, A. L. Lentine, D. C. Trotter, and M. R. Watts, 'Low-voltage differentially-signaled modulators,' Opt. Express **19**, 26017-26026 (2011)

Si Photonics 2 x 2 WSS



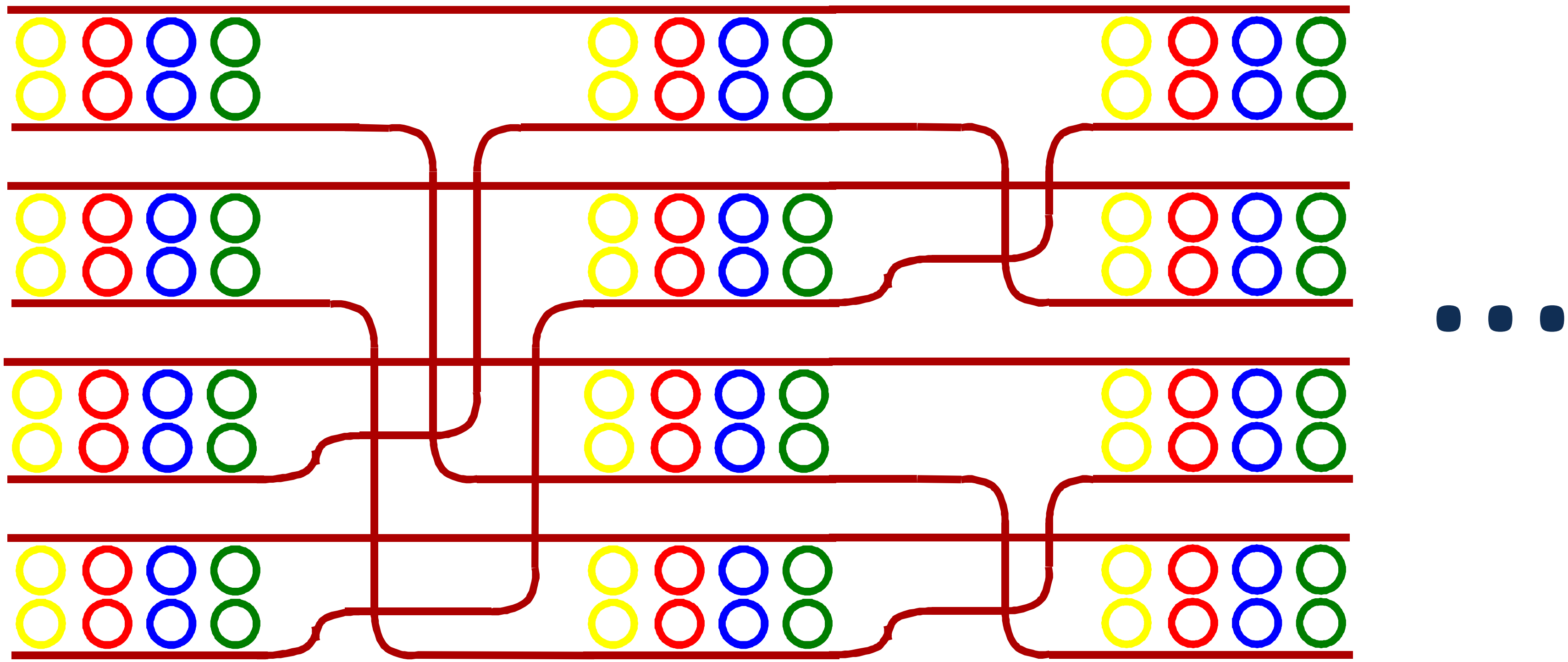


LONG TERM (IDEAL) SPECS:

- Ultimate Switch time < 25 ps
- Loss (cross state) 1 – 2 dB
- Loss (bar state) < 0.2 dB
- Crosstalk (15 - 20 dB)
- Resonant wavelength stabilization

- Ring Size ~ 4 - 6 um
- Coupling gaps ~ 200 - 500 nm
- Ring to ring spacing ~ 4 – 6 um*
- Size < 12 um × λ × 10 um.

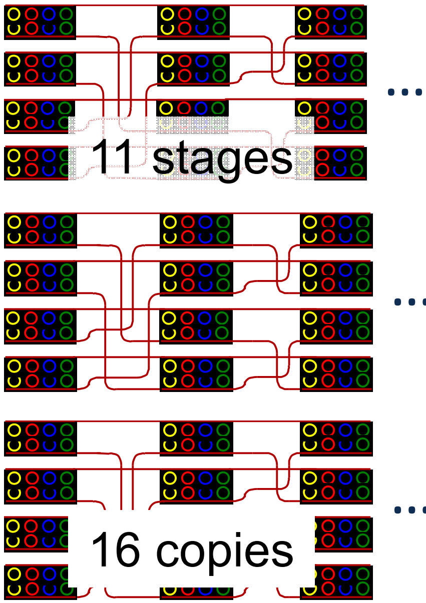
WSS Networks …



- Variety of networks from 2 x 2s (Extended Generalized Shuffle[1])
  - Squaring of crosstalk (EGS, Dilated Benes)
  - Tradeoff between initial fan-out and number of stages (EGS)
- Interconnects require planar crossings or two level optics
  - Nitride, Polysilicon: crosstalk can be very good, careful of loss

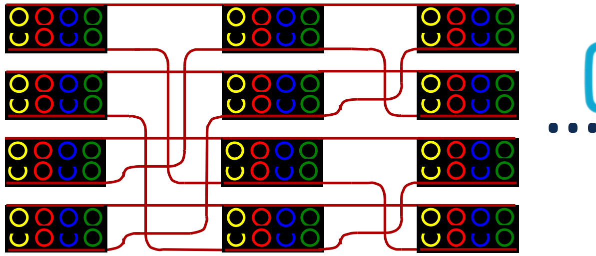
[1] G. W. Richards and F. K. Hwang (1999)

Scalability: 256 x 256 x 32λ (8192 ports)



- Parameters: (F=16, s=11);
  - 2 x 2 size: 12 x $10m_\lambda$ (um); switch length = s*$10m_\lambda$
  - Interconnect length (N/2+N/4+N/8+…)*(w+g)+s*(4r+2T);
    - w+g=4um, r=2um, T=20um
- Design characteristics: 4 chips, F=4 per chip
  - Height = 12*256*4 = 12.3 mm
  - Length < (10*32*13)+(512*4+192)+(13*(8+20*2)) = 7.1 mm
  - Resonators = 11 * 32 * 128 * 2 +384 = 90,496 tunable rings (361,984 flip chip bonds)
  - 15 stages (including fan-out): 15-30 dB of switch loss
  - ~ 1024 waveguide crossings worst path (and 26 transitions between layers)
  - 13 switch in-band cross talk components, and 26 adjacent channel cross talk
    - Must consider networks with square of crosstalk
  - Active wavelength stabilization: 100uW/per ring * 90k rings = 9.0W!

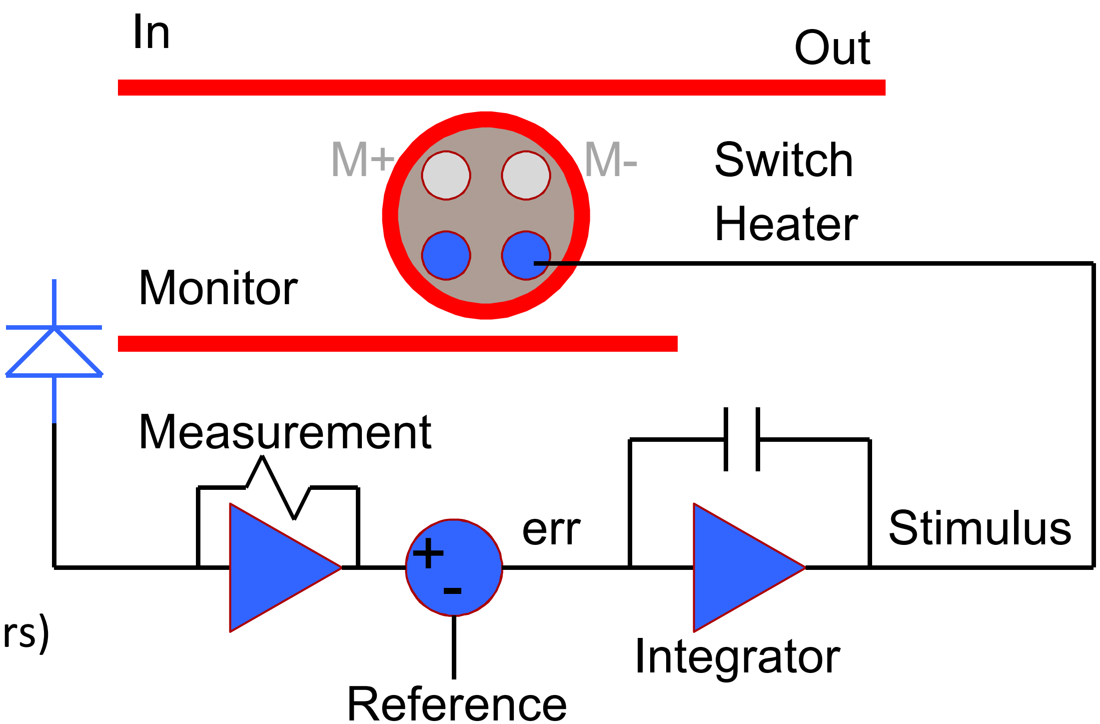
Challenges



- Transfer function through pass band
  - Coupling dependencies (Gap lithography, Si thickness)
  - Back reflections from surface roughness, waveguide-ring interface, and crossings (Do we need isolators at intermediate points?)
- Resonant wavelength stabilization on a grand scale
  - Better efficiency needed compared to today's research results
  - Less initial fabrication wavelength variation
  - Very high yield flip chip bonding to control circuit.
- Waveguide crossing losses, crosstalk,
  - Two layer Silicon is probably the answer (60 dB crosstalk is likely achievable)
- Waveguide, switch, and fiber coupling losses (push hard!)
  - Long waveguide runs, variable waveguide widths on chip
- Polarization diversity
- Researchers always forget about packaging (example has 512 fibers!)

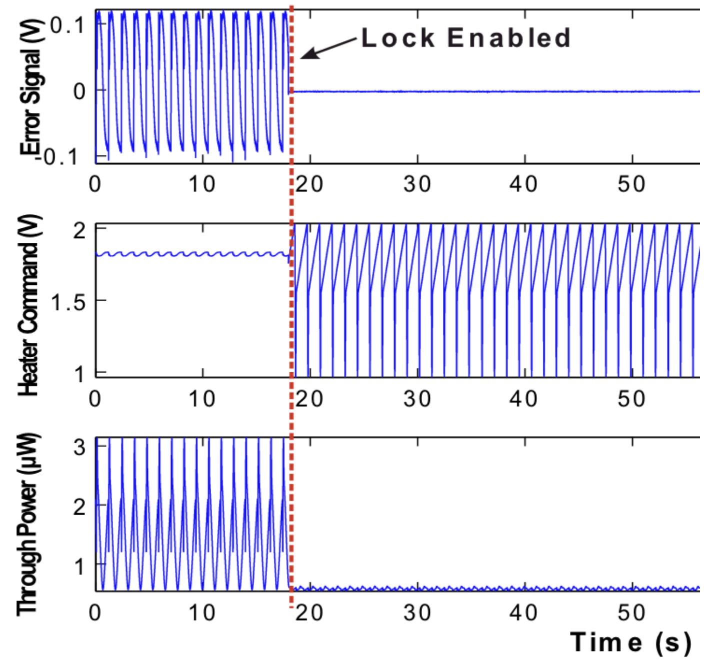
Resonant Wavelength Control

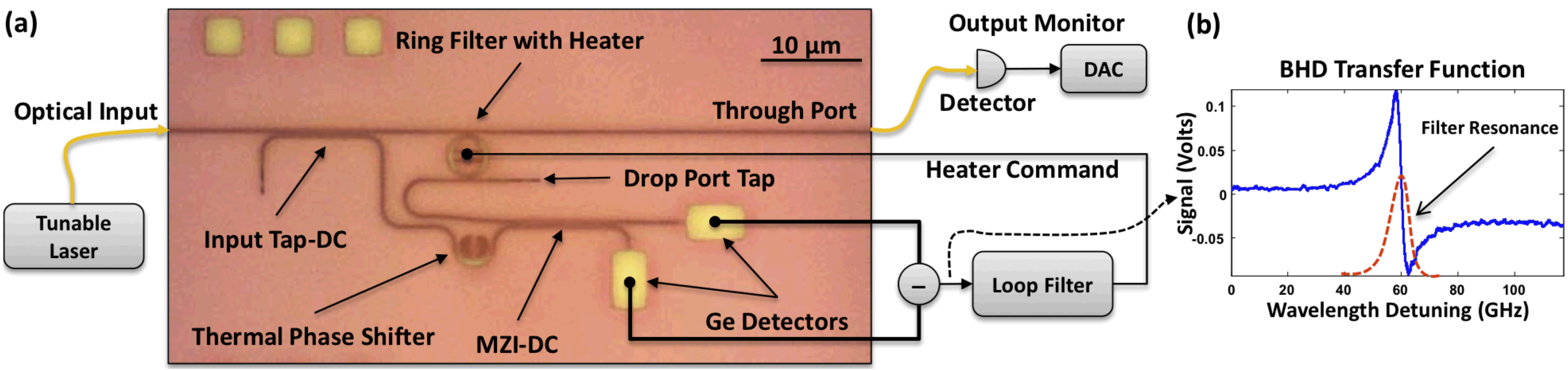
- Control Loop
  - Measurement
    - Temperature
    - Power (shown)
    - Phase (BHD, PDH)
    - Bit errors
  - Integration (PI Loop)
  - Stimulus
    - Integral Heater (shown)
    - Forward bias (heater/carriers)
    - Reverse bias (carriers)
  - Many techniques demonstrated
    - Balanced Homodyne detection, Bit error rate, Temperature sensor (Sandia)
    - Dither (Columbia)
    - Power (Columbia, MIT)
    - Pound Drever Hall* (Texas AM)



Resonant locking of a DWDM filter (Sandia)

- Creates anti-symmetric signal – lock at zero (no reference)
- Build an optical interferometer with a ring in one arm
  - Can we eliminate phase adjust?
- Simple electrical circuit (minimum power)





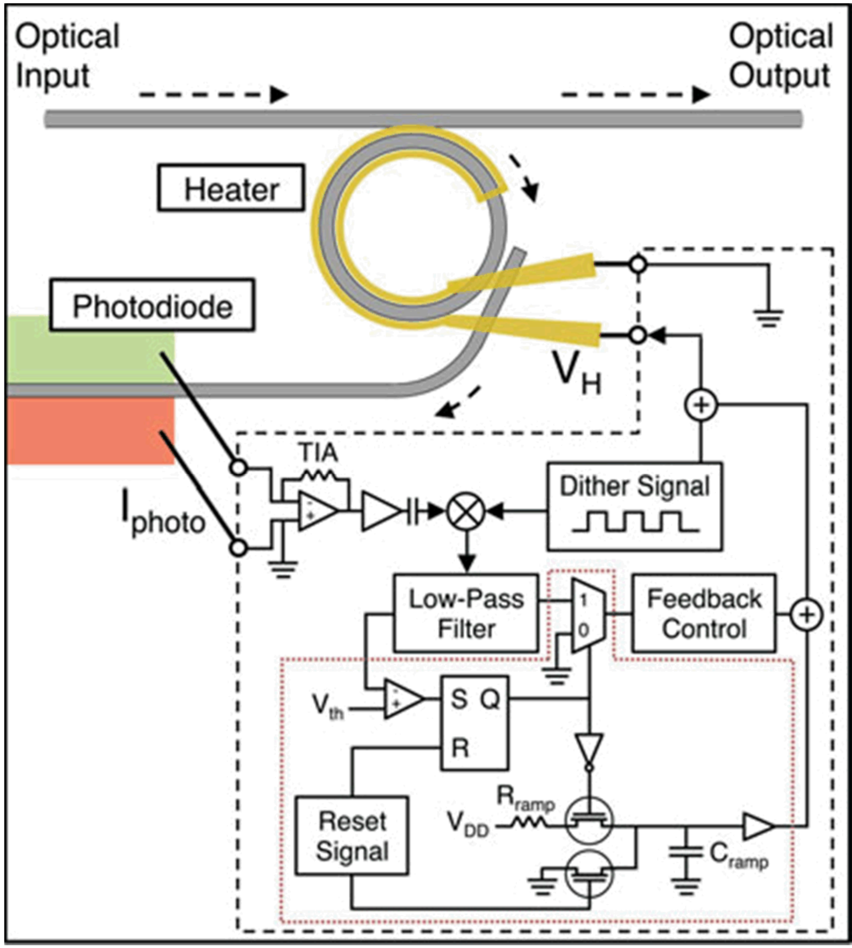
Conclusions

- Photonics in data center for
  - Interconnect (Hi Radix options)
  - Virtual configurations (provisioned connections)
  - Data routing (microsecond to picosecond reconfigurations)
- Silicon Photonics offers *the potential* for scalable data center networking solutions with 1000s of ports.
- Simple CMOS interfaces to Si photonic devices
  - leads to low cost software/photonic interfaces
- *Many* key technical challenges remain
  - No show-stoppers identified yet

Locking using a dither signal (Columbia)

Sandia National Laboratories

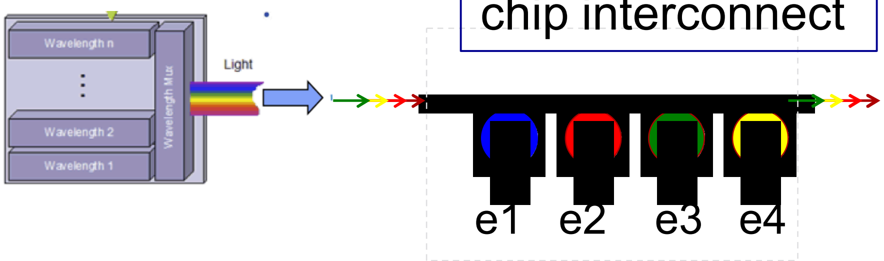
- Creates a signal that is anti-symmetric (lock at zero)
- More complex electrically
- Simple optically
- Some small degradation in the optical performance with dither
- Power, size, ?



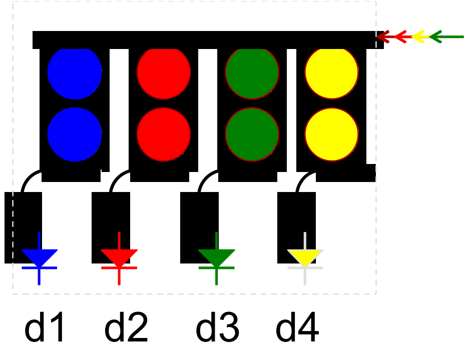
Components

- Transmitter (electrical)
  - ***Serialization (Tx)***
  - ***Modulator Driver/Modulator***
  - ***Modulator wavelength stability***
- Receiver (electrical)
  - ***Demultiplexer wavelength stability***
  - ***Receiver***
  - ***Phase alignment***
  - ***De-serialization (Rx)***
- Laser (electrical/optical)
  - ***Laser Optical Power = Rx sensitivity + Loss budget (in dB)***
    - *Laser wavelength combining (if applicable)*
    - *Laser fiber coupling (if not integrated)*
    - *Waveguides (Tx)*
    - *Modulator*
    - *Waveguide fiber coupling (Tx)*
    - *Fiber waveguide coupling (Rx)*
    - *Demultiplexer (Rx)*
    - *Receiver input power*
  - ***Laser power = Laser optical power/efficiency***

Focus on chip to chip interconnect



Tx



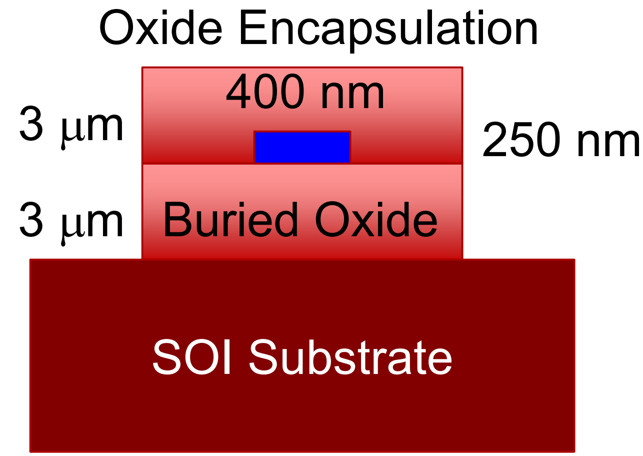
Rx

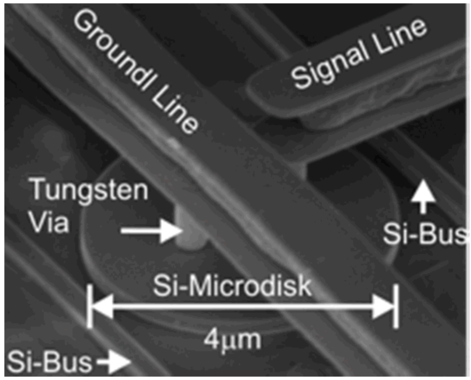
Filter technology

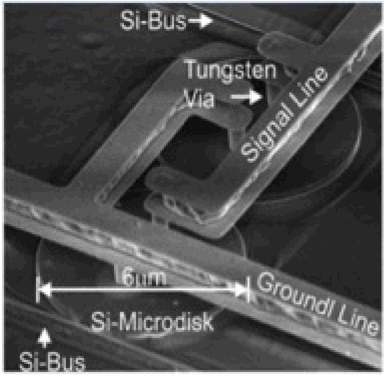
- AWG (Silica)
- TFF
- SiP
  - AWG
  - Eschelle gratings
  - Cascaded MZ interference filters
  - Micro-rings

What is Silicon Photonics?

- Active and Passive Photonics on/in Silicon
- Passive:
  - Waveguides, spectral filters, splitters, polarizers, polarization rotators, gratings, isolators*
- Active
  - Modulators(EO), switches, detectors (OE, Ge), lasers*
    - Thermal Shift of index
    - Electro-refraction
    - Electro-absorption (SiGe)
- Most applications require intimate integration with CMOS Electronics
  - Heterogeneous integration
    - Flip-chip bonding, Wafer bonding, etc.
  - Monolithic integration

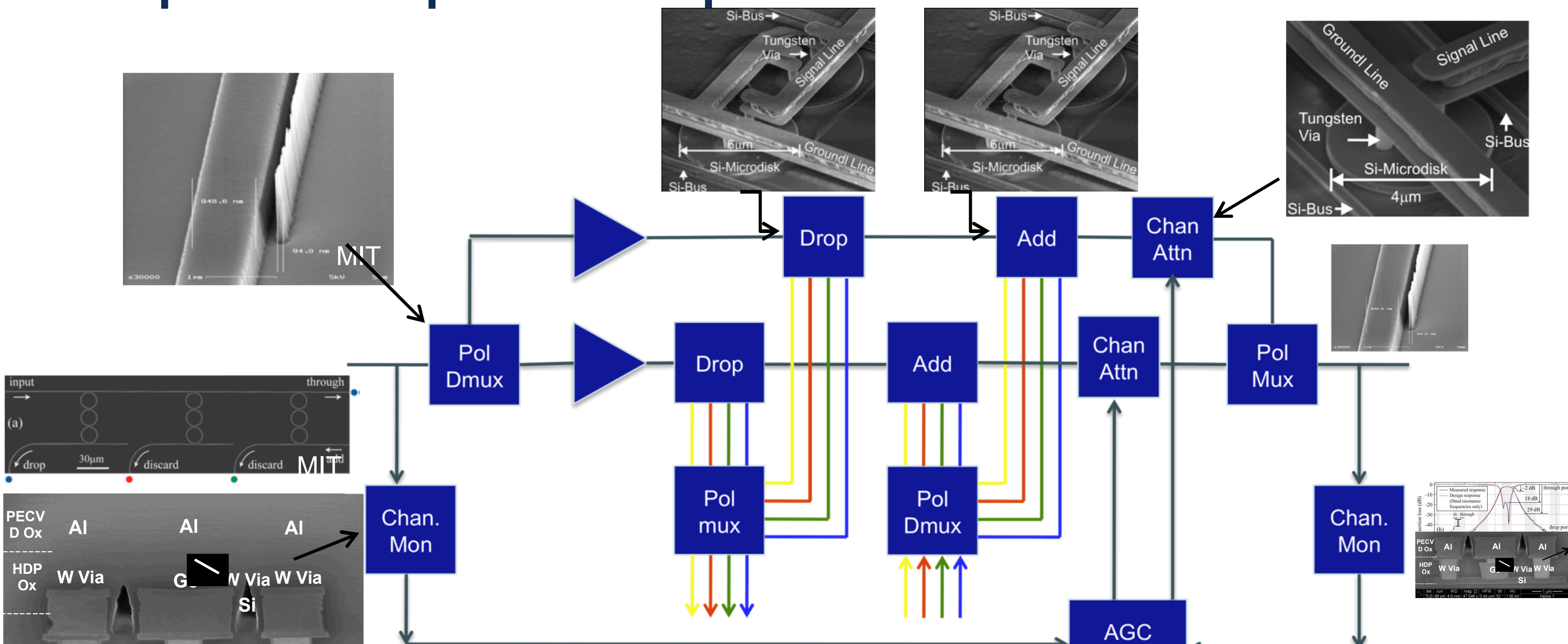




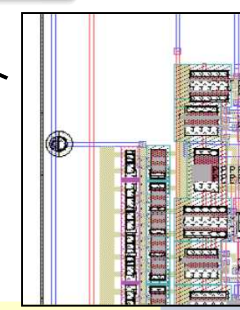


Photos from M. R. Watts, Zortman, Trotter, Young, Lentine, Optics Express, 22, (2011)

Silicon Photonics devices implement complex chip-scale photonic networks

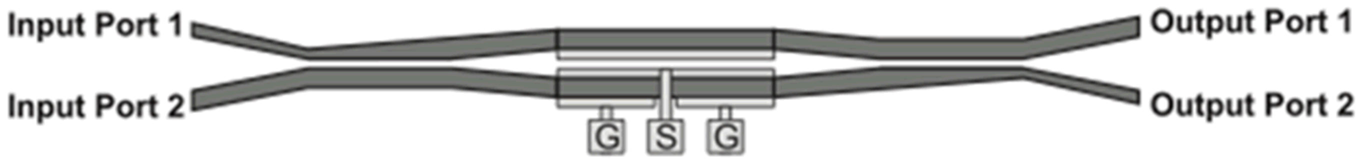


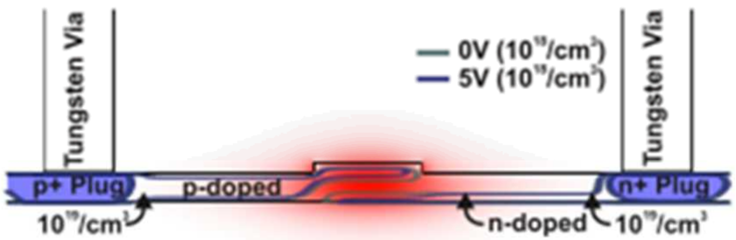
- Most key devices are TRL3
- Challenges: Integrated optical amplifier at TRL1
  Integration of many different devices

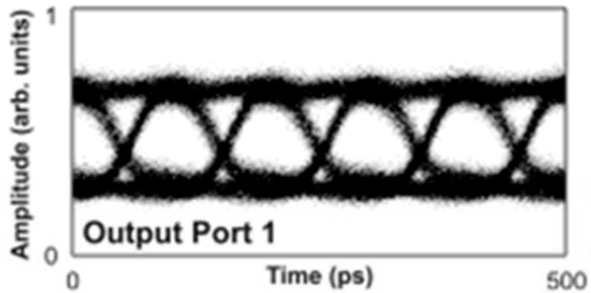


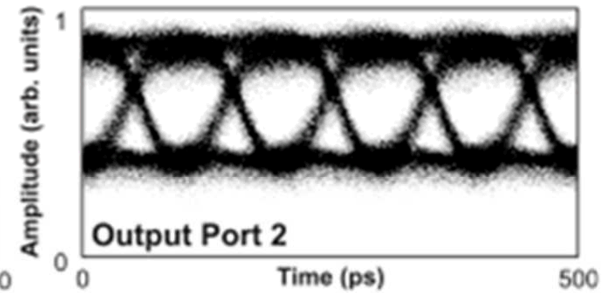
Functions map into silicon photonics devices

Low V-pi Optical Modulator



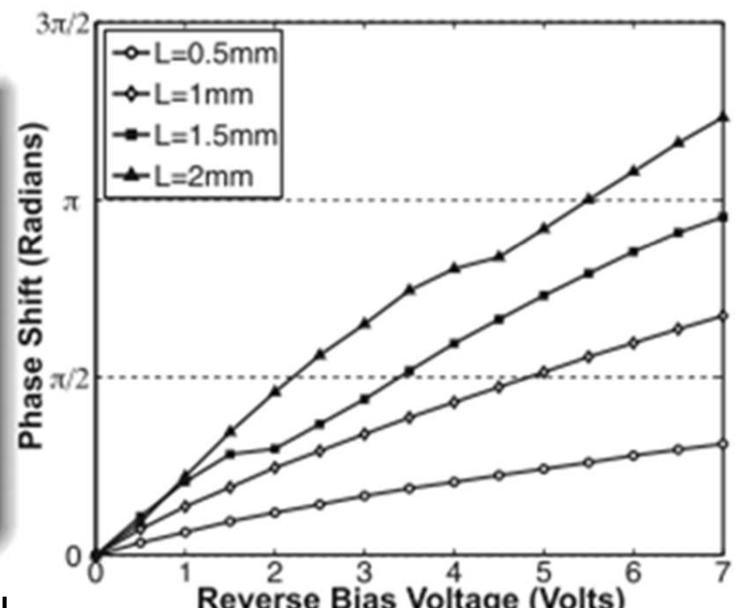






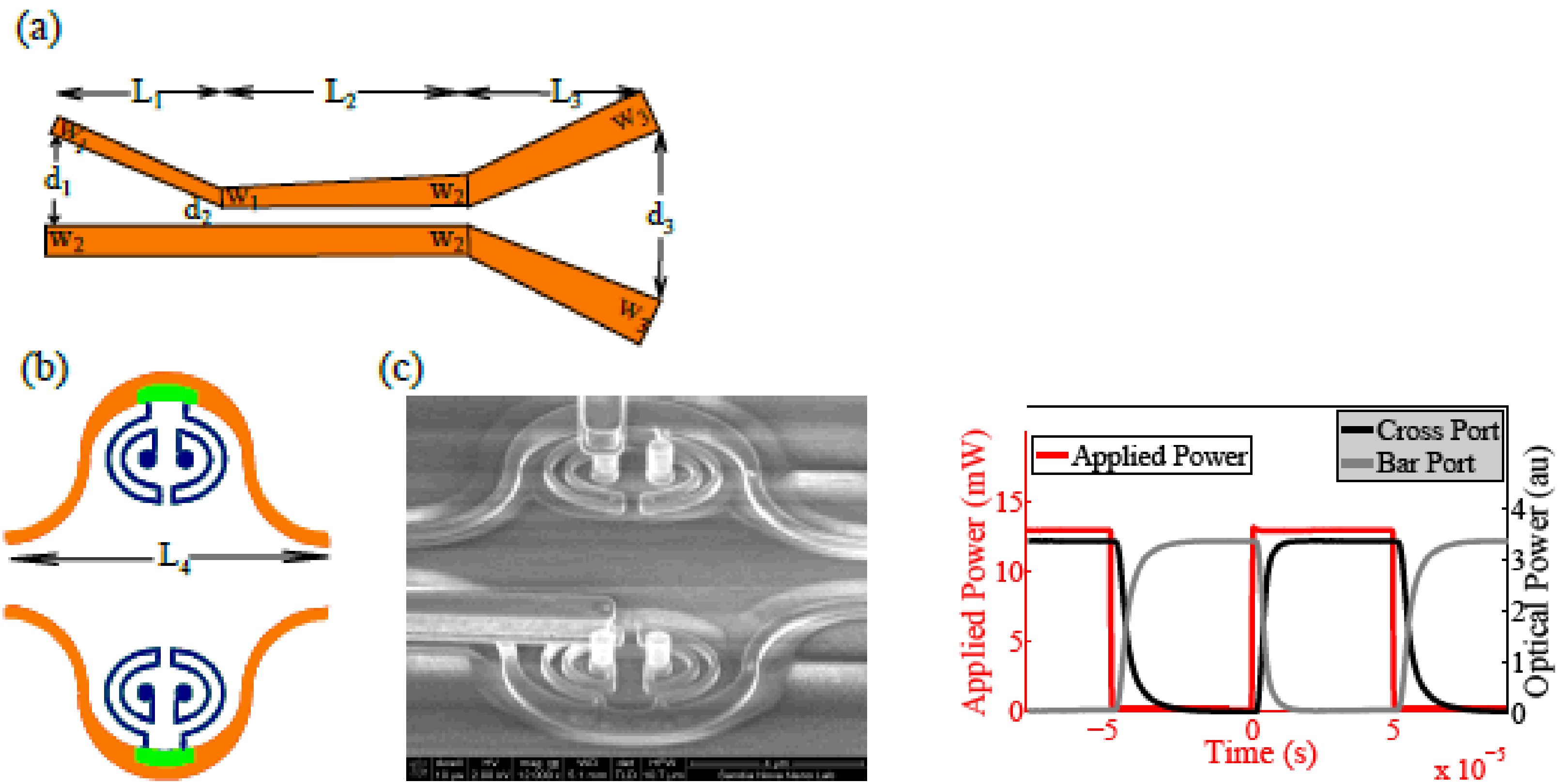
Maximum Overlap → Vertical Junction

- **Effective Index:** $\Delta\bar{n} \propto \frac{nc\epsilon_0}{2} \int \Delta N |\mathbf{e}|^2 dA$
- **Phase Shift:** $\Delta\phi \propto \Delta\bar{N} \frac{\Delta w_d L}{w}$
- **Record** $V_\pi L$: $1V \cdot cm$
- **Power:** ~10pJ/bit, <u>same as VCSELs</u>



M. R. Watts, W. Zortman, D. C. Trotter, R. W. Young, and A. L. Lentine, 'Low voltage compact depletion mode silicon Mach-Zehnder modulator, IEEE JSTQE, 159-164 (2010)

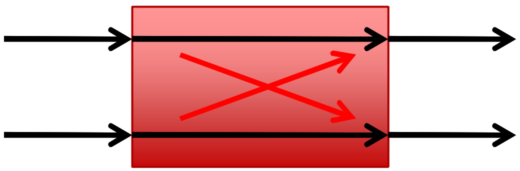
Broadband 2 × 2 Thermo-optic switches

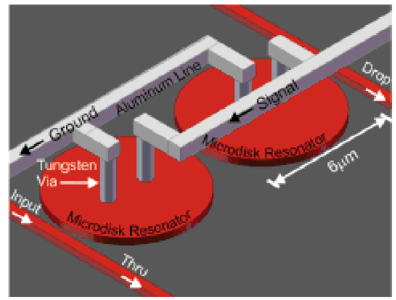


C. T. DeRose, M. R. Watts, R. W. Young, D. C. Trotter, G. N. Nielson, W. A. Zortman, and R. D. Kekatpure, "Low power and broadband 2 x 2 silicon thermo-optic switch," OFC 2011.

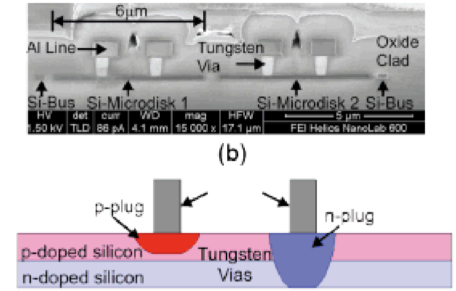
2.4 ns Optical Routing (Electrical Control)

Sandia National Laboratories





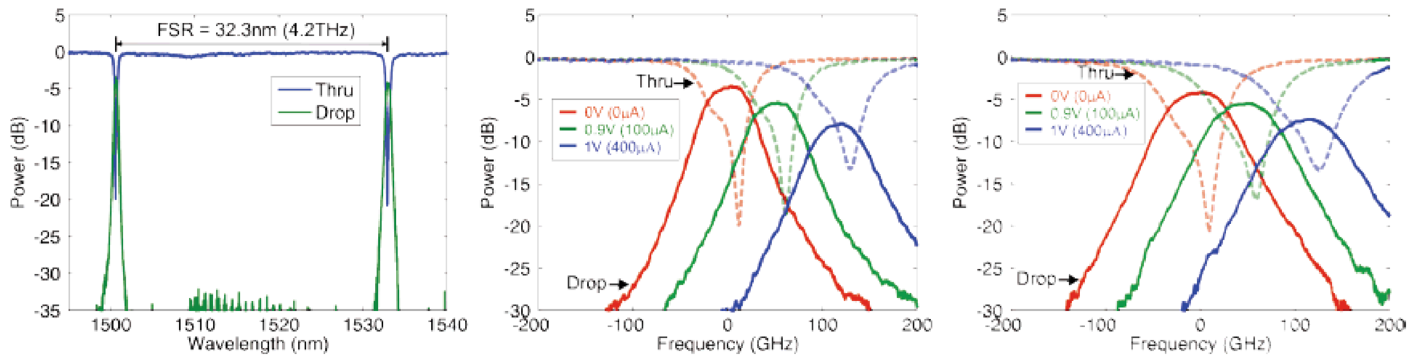
(a)



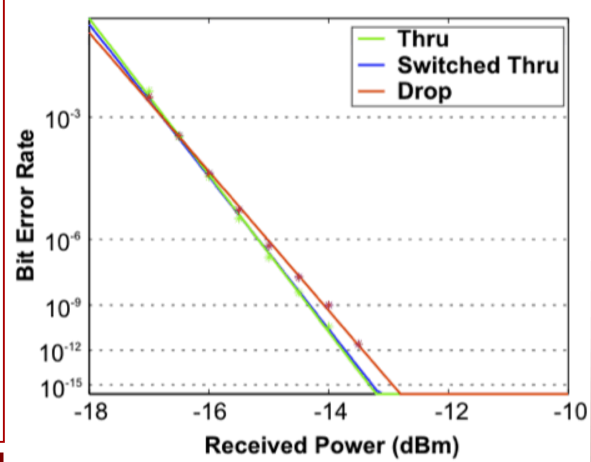
(b)

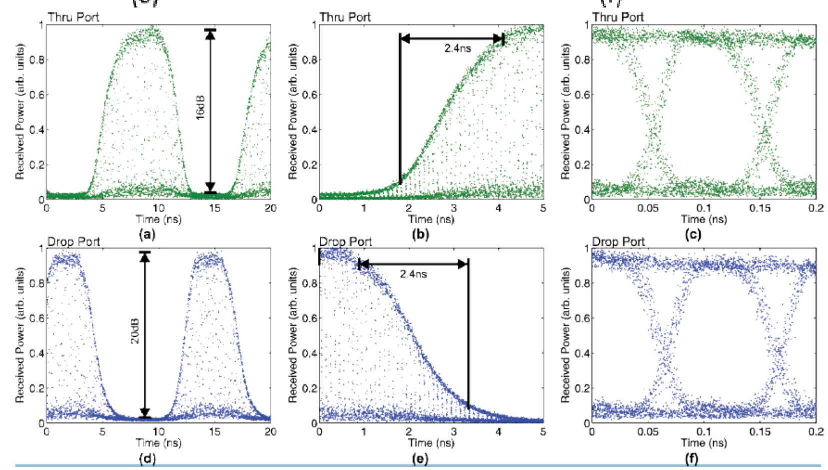
(c)

- *Forward bias* for large shift
- Only one λ can be switched to the cross state
- Switches cascaded for DWDM switching
- Basic element of wavelength selective switch (WSS)



(d) (e) (f)

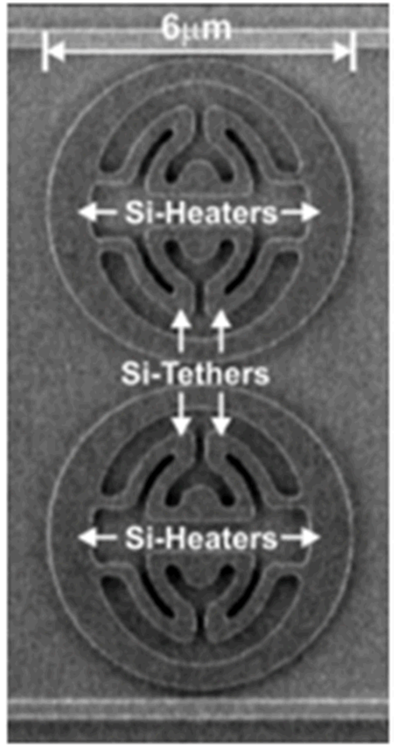


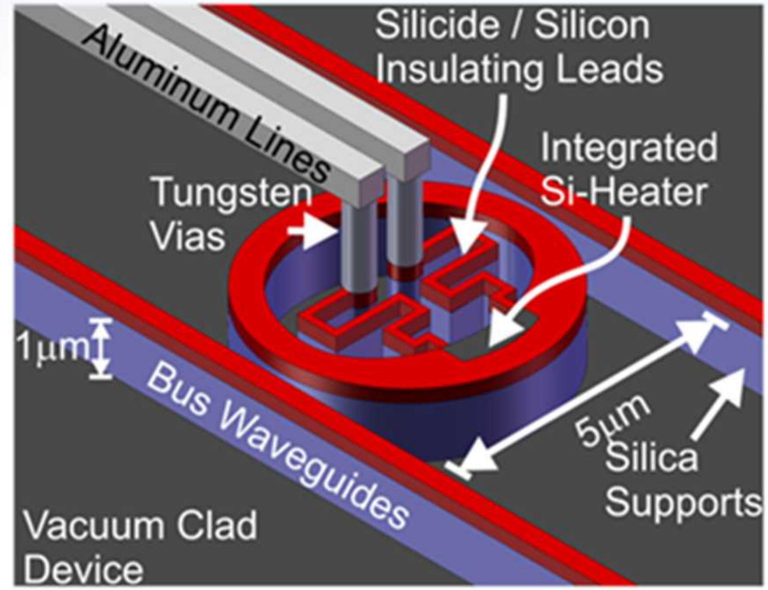


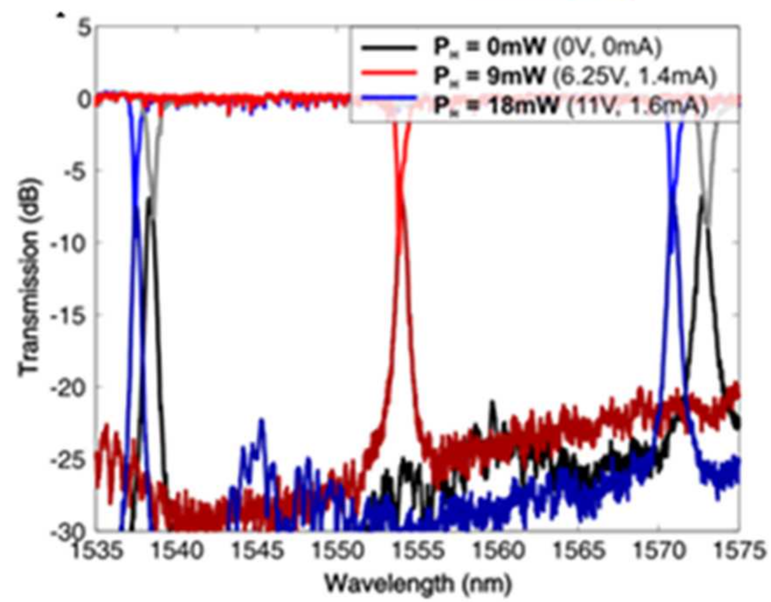
M. R. Watts, D. C. Trotter, R. W. Young OFC2008; Group IV Photonics, 2009

Full C-band tunable filter/switch

- Full C-band tuning
- 35 nm FSR
- Very low thermal tuning power
  - 4.4 μW/GHz
  - 30 μW/° C
- Microsecond switching times



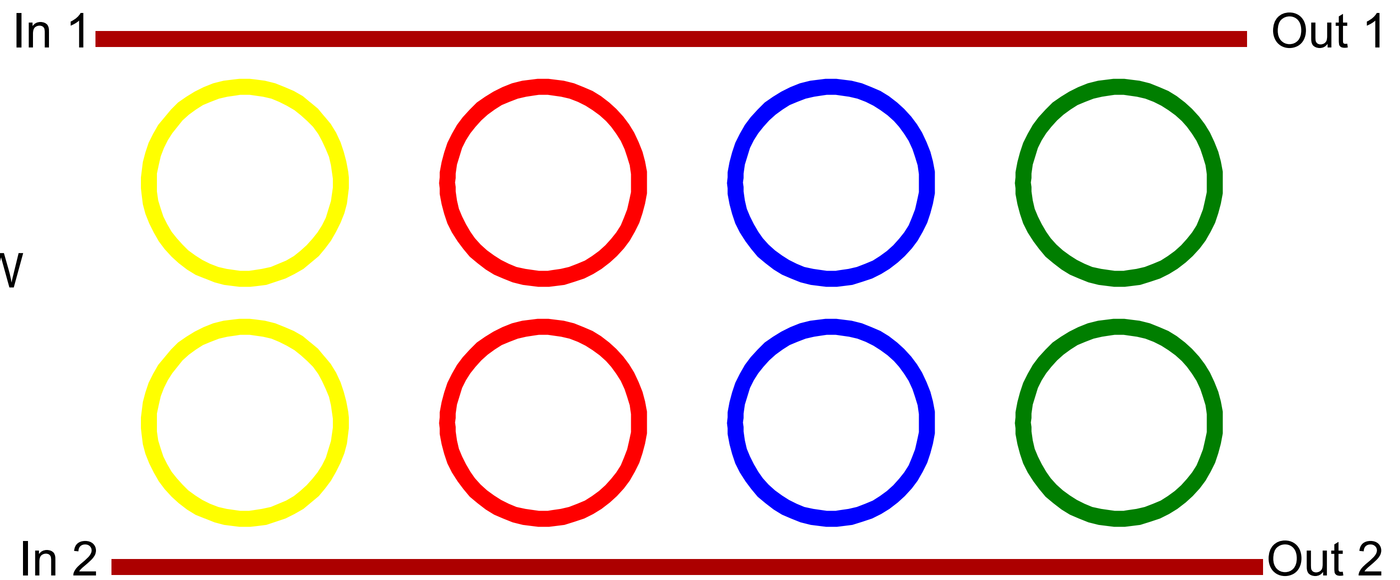




M. R. Watts, W. A. Zortman, D. C. Trotter, G. N. Nielson, D. L. Luck, and R. W. Young, 'Adiabatic Resonant Microrings (ARMs) directly integrated with thermal microphotonics, CLEO 2009

Si Photonics 2 x 2 WSS

Sandia National Laboratories



GRAPHICAL VIEW
NEED L PLOT

LONG TERM SPECS:

- Ultimate Switch time < 25 ps
- Loss (cross state) 1 – 2 dB
- Loss (bar state) < 0.2 dB
- Crosstalk (15 - 20 dB)
- Resonant wavelength stabilization

- Ring Size ~ 4 - 6 um
- Coupling gaps ~ 200 - 500 nm
- Ring to ring spacing ~ 4 – 6 um*
- Size < 12 um × λ × 10 um.

EGS Design parameters

- s= 6, sp=16,N= 64, F>= 20, X=10240, X2= 4096, rings=655360
- s= 7, sp=15,N= 64, F>= 14, X= 6720, X2= 4096, rings=430080
- s= 8, sp=16,N= 64, F>= 11, X= 5632, X2= 4096, rings=360448
- s= 9, sp=15,N= 64, F>= 8, X= 3840, X2= 4096, rings=245760
- s=10, sp=16,N= 64, F>= 7, X= 3584, X2= 4096, rings=229376
- s=11, sp=17,N= 64, F>= 6, X= 3264, X2= 4096, rings=208896
- s=12, sp=18,N= 64, F>= 6, X= 3456, X2= 4096, rings=221184
- s=13, sp=19,N= 64, F>= 6, X= 3648, X2= 4096, rings=233472

- s= 7, sp=17,N= 128, F>= 28, X=30464, X2=16384, rings=1949696
- s= 8, sp=18,N= 128, F>= 22, X=25344, X2=16384, rings=1622016
- s= 9, sp=17,N= 128, F>= 15, X=16320, X2=16384, rings=1044480
- s=10, sp=18,N= 128, F>= 12, X=13824, X2=16384, rings=884736
- s=11, sp=19,N= 128, F>= 9, X=10944, X2=16384, rings=700416
- s=12, sp=18,N= 128, F>= 8, X= 9216, X2=16384, rings=589824
- s=13, sp=19,N= 128, F>= 7, X= 8512, X2=16384, rings=544768
- s=14, sp=20,N= 128, F>= 7, X= 8960, X2=16384, rings=573440
- s=15, sp=21,N= 128, F>= 7, X= 9408, X2=16384, rings=602112

- s= 8, sp=20,N= 256, F>= 44, X=112640, X2=65536, rings=7208960
- s= 9, sp=19,N= 256, F>= 30, X=72960, X2=65536, rings=4669440
- s=10, sp=20,N= 256, F>= 23, X=58880, X2=65536, rings=3768320
- s=11, sp=19,N= 256, F>= 16, X=38912, X2=65536, rings=2490368
- s=12, sp=20,N= 256, F>= 13, X=33280, X2=65536, rings=2129920
- s=13, sp=21,N= 256, F>= 10, X=26880, X2=65536, rings=1720320
- s=14, sp=22,N= 256, F>= 9, X=25344, X2=65536, rings=1622016
- s=15, sp=21,N= 256, F>= 8, X=21504, X2=65536, rings=1376256
- s=16, sp=22,N= 256, F>= 8, X=22528, X2=65536, rings=1441792
- s=17, sp=23,N= 256, F>= 8, X=23552, X2=65536, rings=1507328

EGS Design parameters

```
%EGS strictly non-blocking 2 x 2 nodes
%Richards and Hwang, Networks 1999, page 290
fprintf('\n');
fprintf('\n');
ml=32;
for N=[64,128,256];
   slow=log2(N);
   shih=2*log2(N)+1;
   for s=slow:shih
      if(round(s/2)==s/2) %even
         if( (3<=s) && (s<=(log2(N)+2) ) )
            F=3*N/2^(s/2)-N/2^(s-2);
         elseif( ((log2(N)+3) <= s) && (s<=(2*log2(N)+1) ) )
            F=3*N/2^(s/2)-log2(N)+s-3;
         end;
      else
         if( (3<=s) && (s<=(log2(N)+2) ) )
            F=2^1.5*N/2^(s/2)-N/2^(s-2);
         elseif( ((log2(N)+3) <= s) && (s<=(2*log2(N)+1) ) )
            F=2^1.5*N/2^(s/2)-log2(N)+s-3;
         end;
      end;
      sp=s+2*ceil(log2(F));
      sw=sp*N/2*F;
      ring=sw*ml*2;
      sizey=.012;    %wss
      sizex=ml*.01;  %wss
      wg=.002;   %wavegudie spacing in connection area
      T=.02;     %transitions
      rad=.02;   %radius of interconnections

      fprintf('s=%2d, sp=%2d,N=%4d, F>= %2d, X=%5d, X2=%5d, rings=%6d\n',s,sp,N,ceil(F),sw,N*N, ring);
   end;
   fprintf('\n');
end;
```