INTREPID Update

ARPA-E Annual Review

October 24th, 2018

Presenters:
Clint Schow: Overview and progress updates
Katharine Schmidtke: Data center perspective and T2M
Adel Saleh: Network architecture and modeling

Representing contributions from the entire INTREPID team

Project Objectives

- Making Coherent Optics in the Data Center a Reality
  - Demonstrate energy efficient coherent engines compatible with package integration
  - Establish path to low power links beyond 200Gb/s/λ
  - Leverage large sensitivity gains to enable new architectures based on photonic routing and switching
- Unique approach: Use Optical Phased Locked Loops (OPLLs) to eliminate DSP for orders of magnitude reduction in power
- Overturn the conception that coherent = high power consumption
- Explore and model novel architectures enabled by the expanded link budget of coherent receivers
- Demonstrate operation in a live datacenter (Phase 2)
- Pave the way for transition to widespread deployment and commercial success

INTREPID Team: UCSB and Facebook

- Project objectives shaped by more than a year of engagement prior to the ENLITENED proposal
- Highly multidisciplinary, expertise spanning many fields
  - Datacenter operation, network design and management
  - Requirements and roadmap for optics at cloud-scale
  - Global optical supply chain, expertise and ability to drive
  - Electronic/photonic integration and co-design
  - Low-power, high-speed circuit design
  - Photonic integrated circuits: design and fabrication
  - High-speed packaging, assembly, and characterization
  - Network architecture design and modeling

Facebook: Katharine Schmidtke (co-PI), Hans-Juergen Schmidtke, Ariel Hendel, Brian Taylor, Todd Hollmann, Jimmy Williams, Gilad Goldfarb
UCSB Faculty: Clint Schow (co-PI), James Buckwalter, Larry Coldren, Jonathan Klamkin, Adel Saleh
UCSB Students and Post-Docs: Hector Andrade, Takako Hirokawa, Aaron Maharry, Thomas Meissner, Stephen Misak, Luis Valenzuela, Yujie Xia, Shireesh Bhat, Sarvagya Dwivedi, Fabrizio Gambini, Sergio Pinna

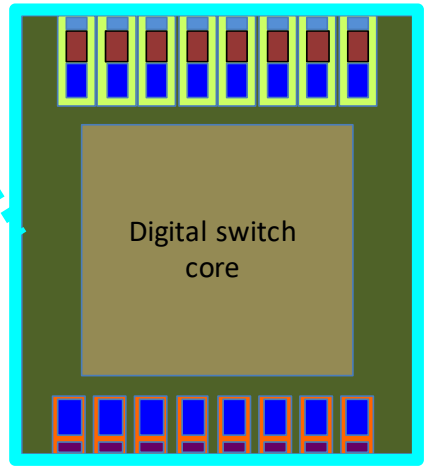
Highlights from Year One

- **Hardware Development**
  - Large tiles designed and in fab (date expected/received)
    - SiGe BiCMOS (Global Foundries 8XP): high-speed drivers and receivers (10/2018)
    - Si Photonics (Global Foundries 9WG): coherent receivers and transmitters (11/2018)
    - InP (UCSB): coherent receivers and transmitters (11/2018)
    - VCSEL arrays (OptiGOT, 10/2018)
  - Challenges:
    - Short timelines make any delays problematic
    - Navigating uncharted territory: demonstrating coherent links at low power
    - Polarization, transmitter designs for low power, OPLL control circuitry
- **Network Architecture**
  - Developed folded Clos architectures with AWGR layers for scalability and energy efficiency
  - Built a flexible Python simulation platform:
    - Any number of switching levels and AWGR layers, different switch sizes, and arbitrary oversubscription ratios
    - General spatial and temporal distributions of traffic
- **Technology to Market**
  - Constructed TEA framework focusing on critical metrics, business models, and need for open interfaces
  - Driven by needs of cloud-scale datacenter operations

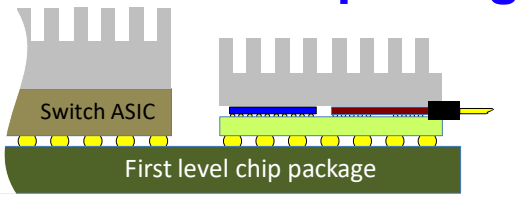
ToR Switches with Server and Fabric Interfaces

Analog Coherent WDM links

- "Up": all switch-to-switch connections, up to 2km over SMF
- OPLL (Optical Phased Locked Loop)→ No/very little DSP
- Expanded link budgets to enable photonic routing/switching
- 1310nm: uncooled lasers, dispersion not an issue
- Target: 800Gb/s/fiber = 4λ @ 200Gb/s/λ (dual-pol QPSK, 50 Gbd/s)



Photonic interfaces integrated in first-level switch packages:



- Modular: common interface to switch ASIC
- Optimum I/O architecture between ASIC and optics to maximize efficiency and performance

Multimode VCSEL links

- "Down": ToR-server connections (30m)
- Currently all copper today → challenging cost point
- 50G (Phase 1)→100G (Phase 2)

UCSB facebook

Analog Coherent: Maximizing Energy Efficiency

Direct-Modulation/Direct Detection

Detected power $\propto (P_{laser} \cdot A_{total})$

P_{laser} = laser power, A_{total} = total link attenuation

- RX sensitivity sets energy efficiency
  - Poor sensitivity = higher transmitter power
- Sensitivity directly degrades with data rate
  - Problem only getting worse

Field Modulation/Coherent Detection

Detected power $\propto \sqrt{(P_{laser} \cdot A_{total})} \cdot P_{LO}$

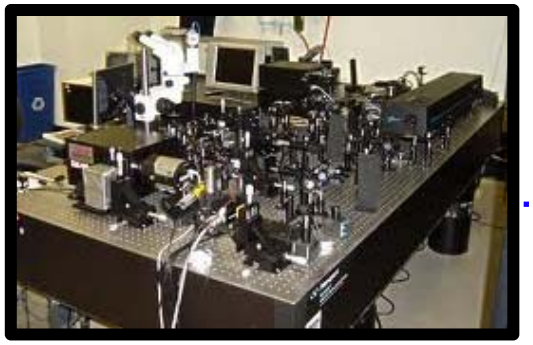
P_{LO} = Local Oscillator (LO) power

- ~20dB improvement in RX sensitivity possible
- Much greater tolerance to attenuation
  - Looser component specs for yield and cost
  - Ability to tolerate insertion loss of optical routing/switching devices

Optical Phase Locked Loop (OPLL) → Eliminating Power-Hungry DSP

Typical OPLLs prior to 2000

Lab-scale



Prior UCSB work

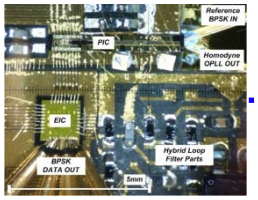
Carrier-scale, single channel

INTREPID

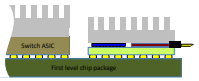
Chip-scale, WDM multi-channel

Photonic and Electronic Integration

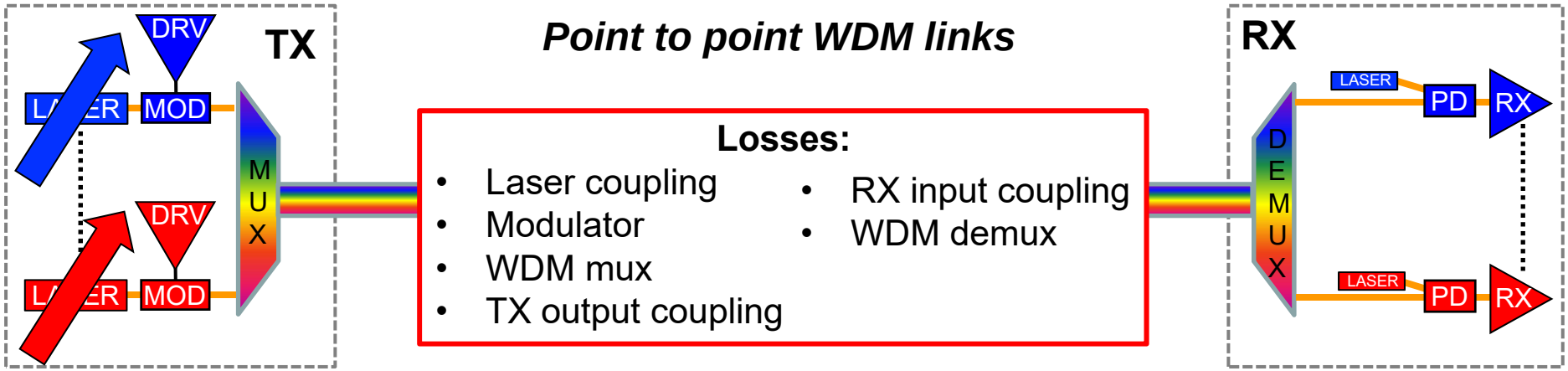
<1000X size
<100X power



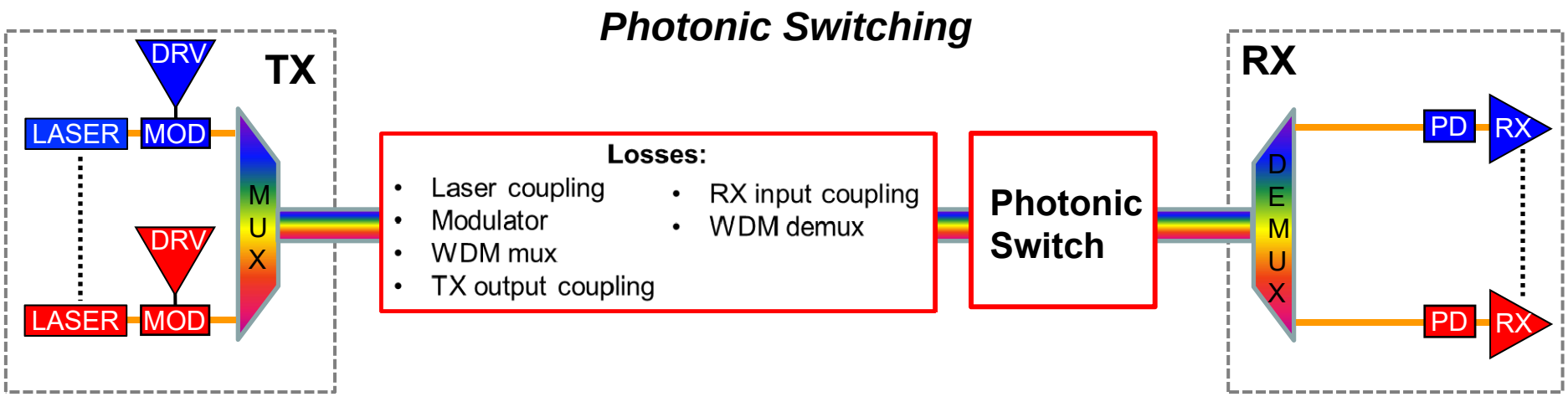
<10X size
<10X power



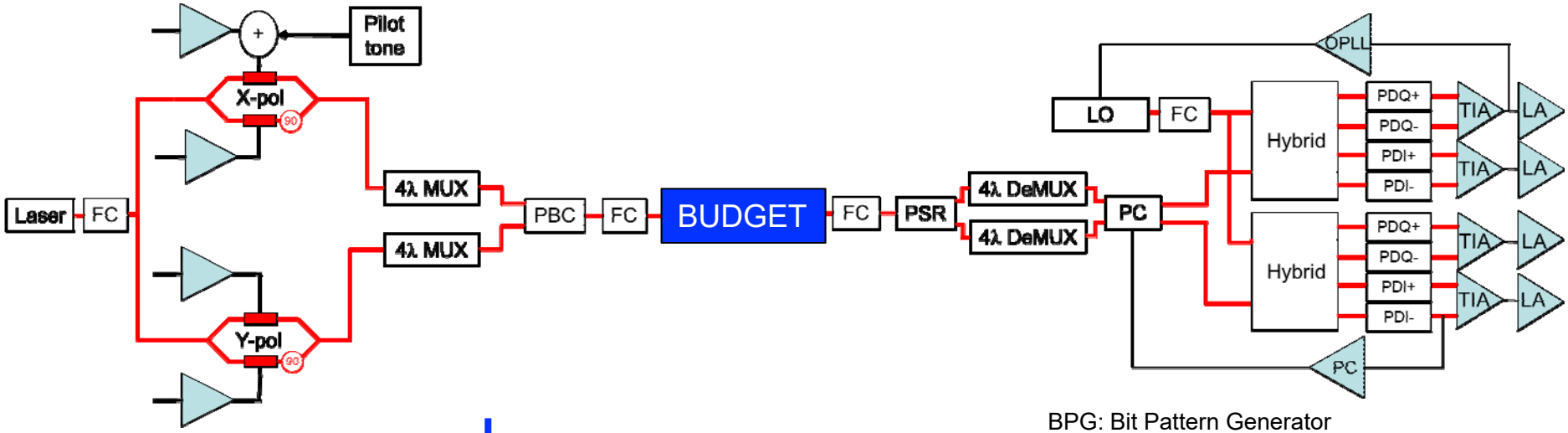
Analog Coherent Links: Solving the Link Budget Problem



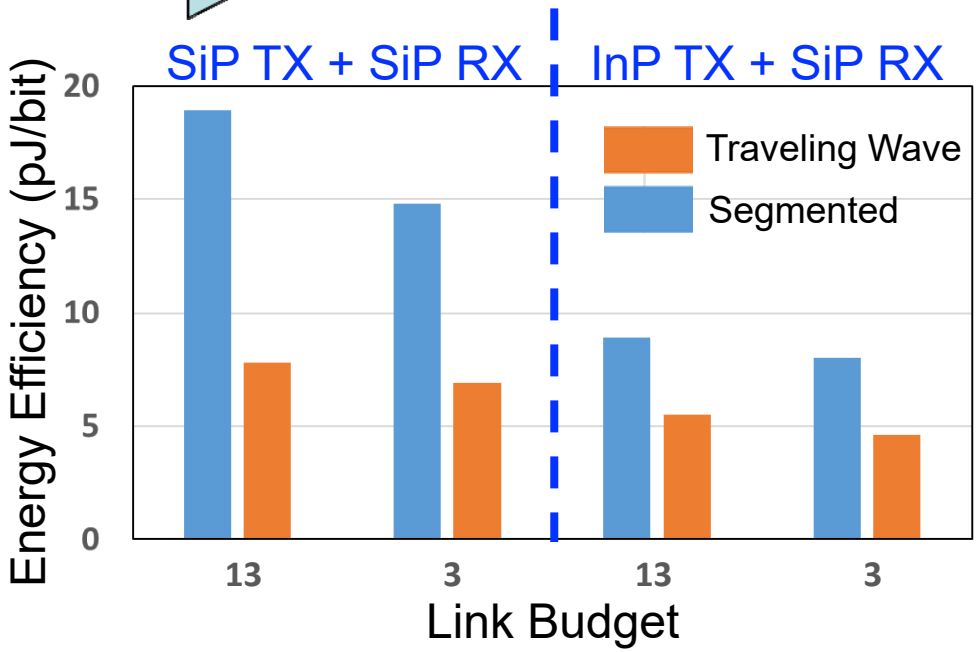
More energy efficient to boost TX laser power to overcome all losses, or spend some power on a local oscillator producing ~1mW at the photodetector?



Link Analysis – WDM, 200 Gb/s Dual-Pol QPSK

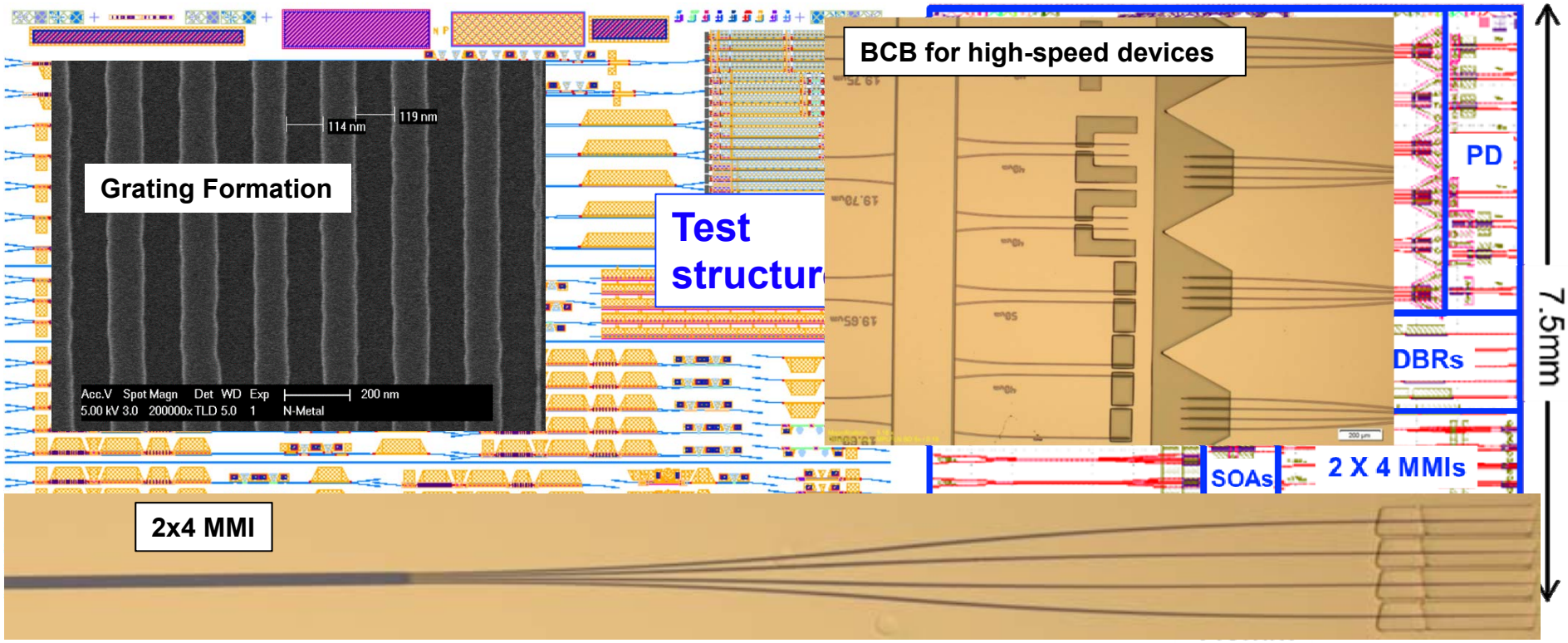


BPG: Bit Pattern Generator
FC: Fiber Coupler
PBC: Polarization Beam Combiner
PSR: Polarization Splitter Rotator
PC: Polarization Control
LO: Local Oscillator
PD: Photo-Diode
LPF: Low Pass Filter
SMF: Single Mode Fiber



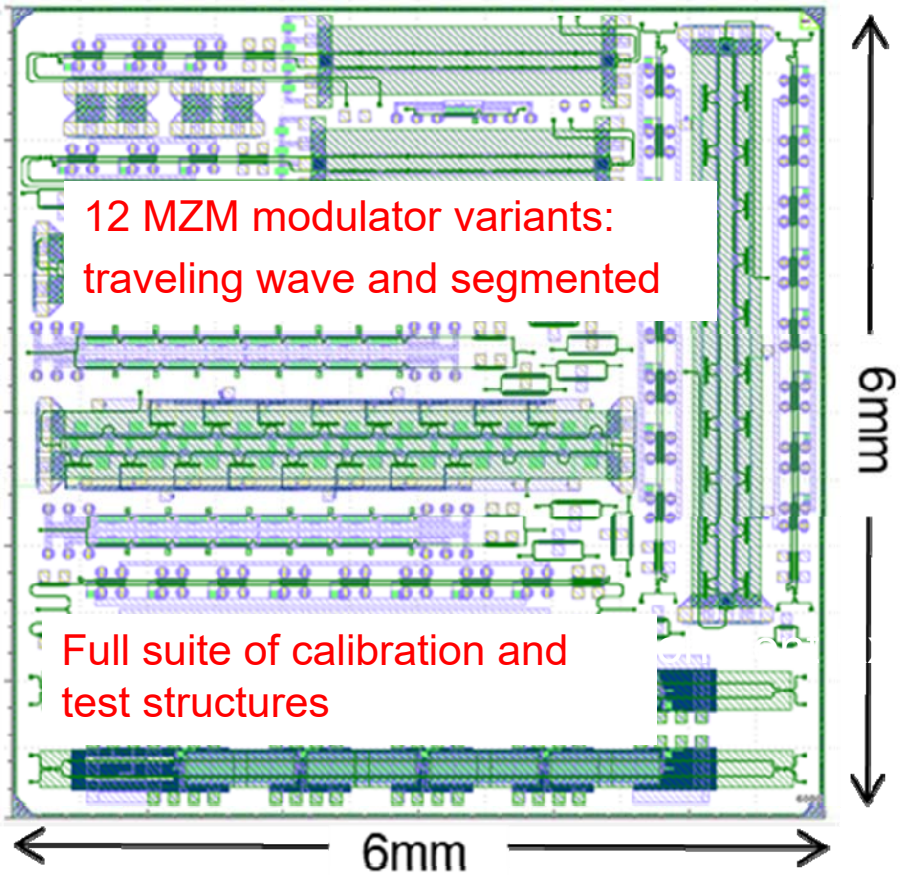
- Significant efficiency advantage for InP modulators
- OPLL loop circuitry contributes less than 1pJ/bit

InP Tiles in Fab

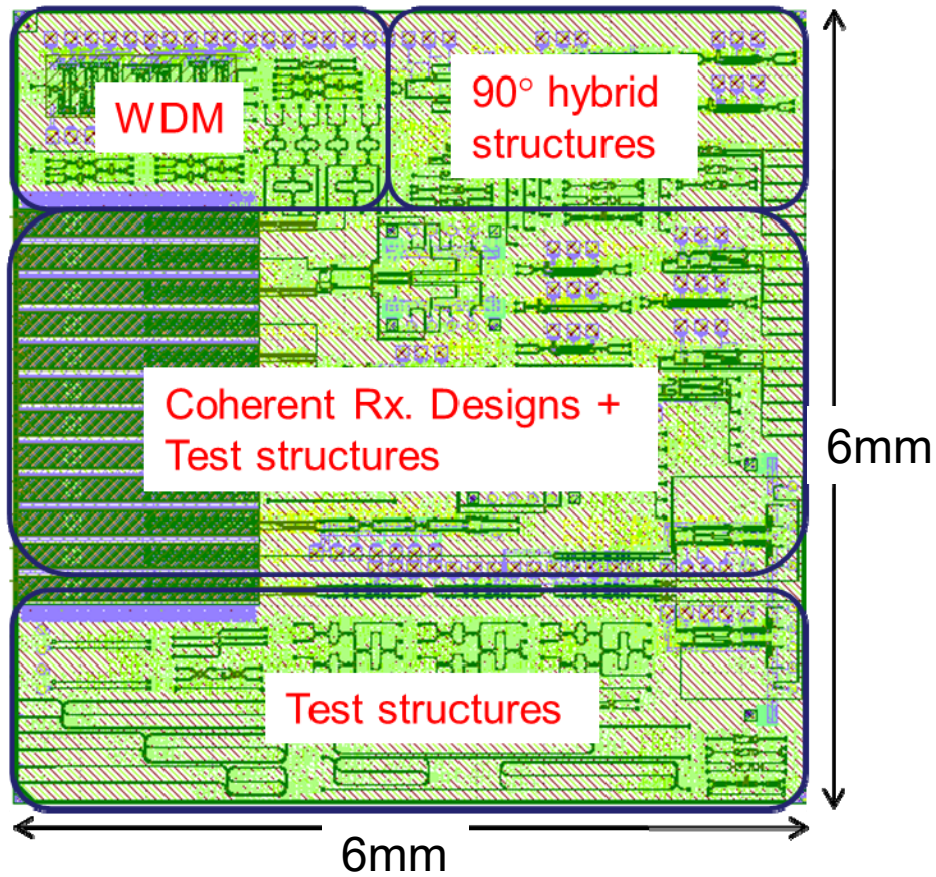


Silicon Photonics in Fab

Transmitter Tile



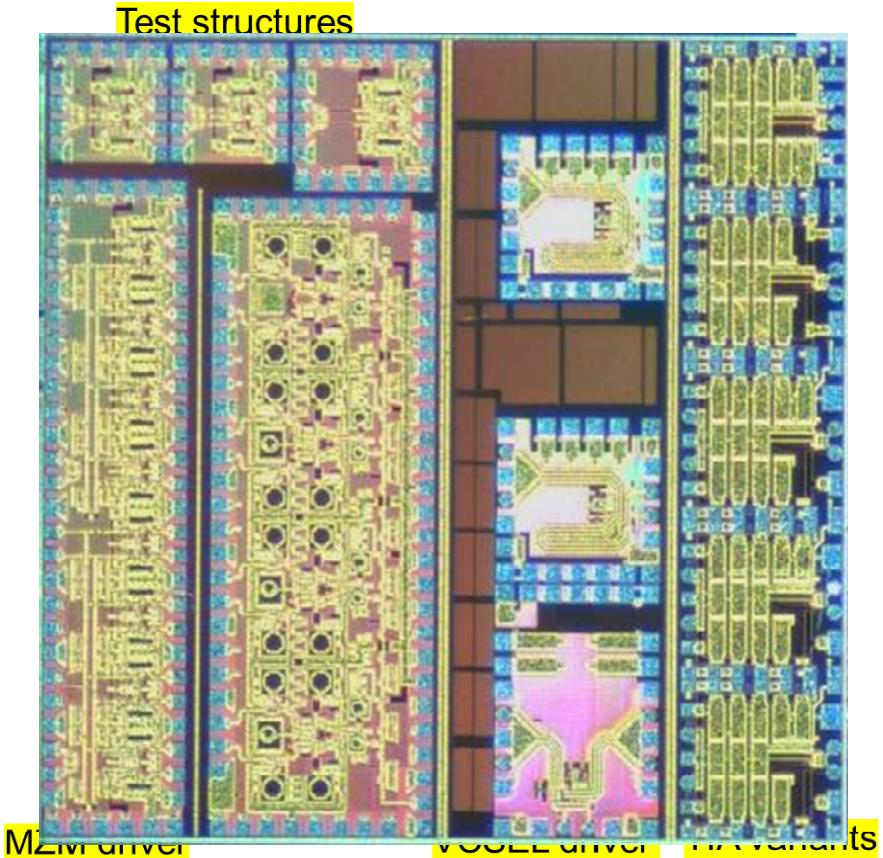
Receiver Tile



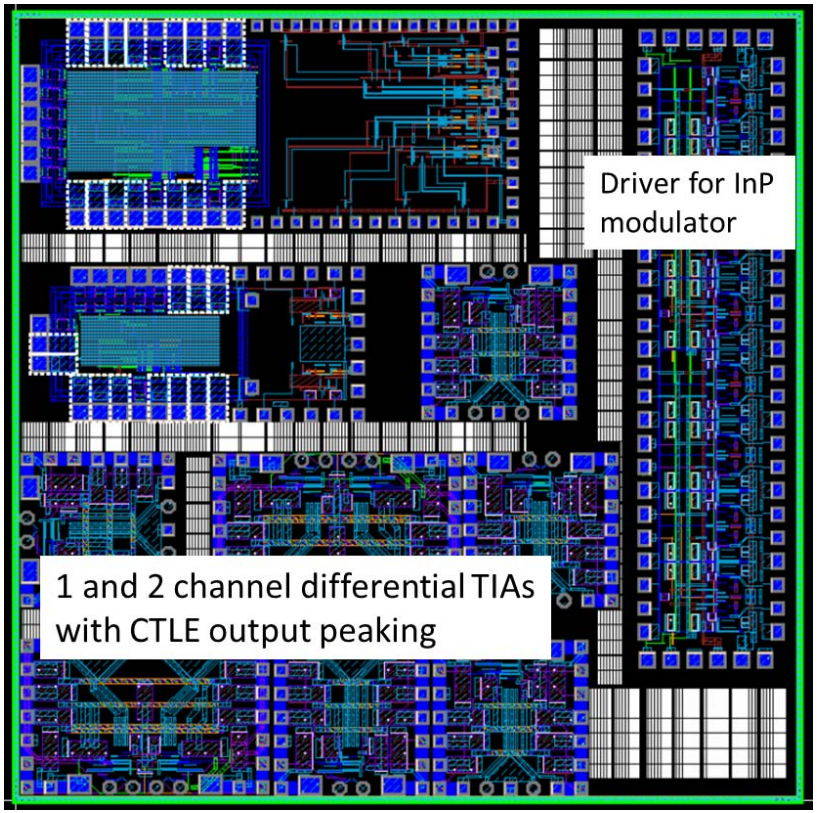
- Global Foundries 9WG, 6mm x 6mm tiles
- Expected November 2018

INTREPID Hardware: Electronics

February 2018 Tapeout

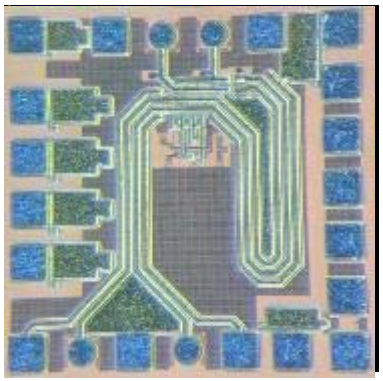


August 2018 Tapeout

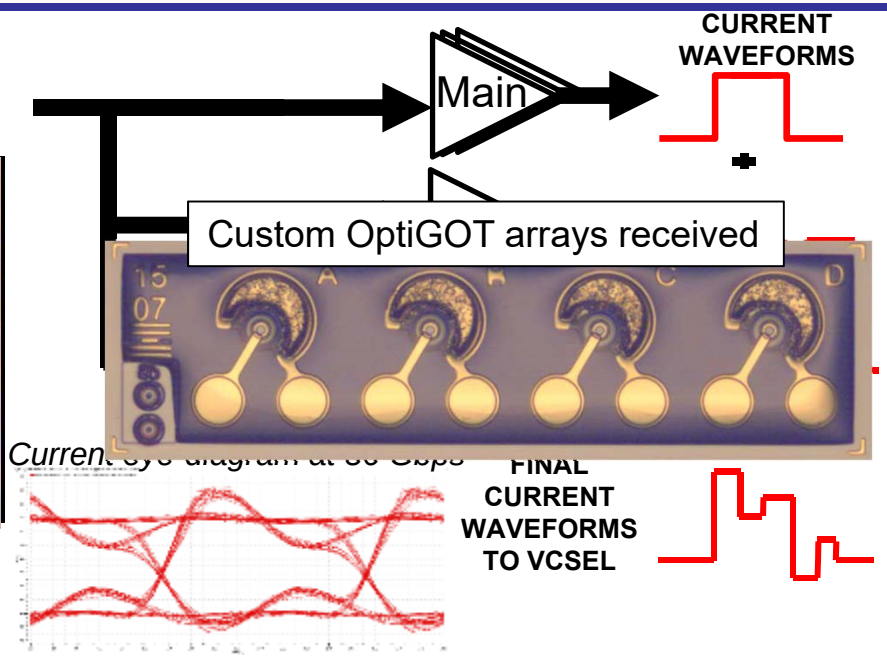


- Global Foundries 8XP, 5mm x 5mm tiles
- First tile received 10/18, initial testing underway

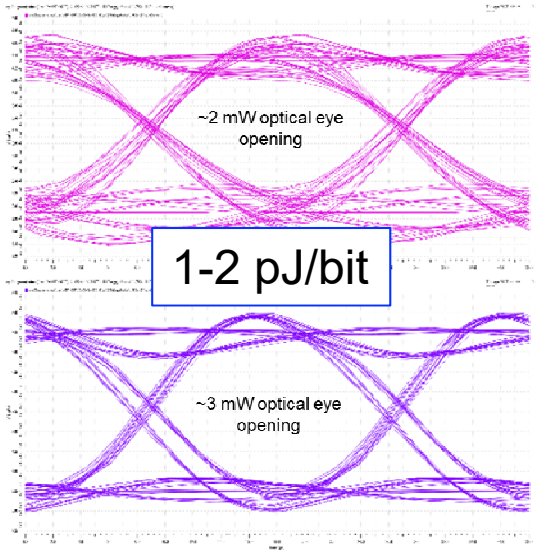
VCSEL Link Electronics

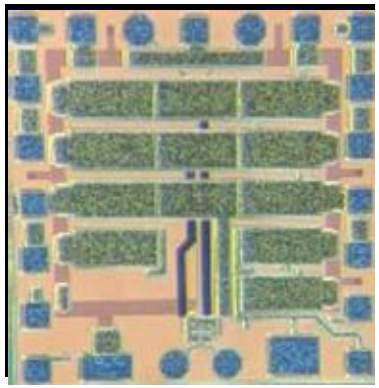


1x1mm

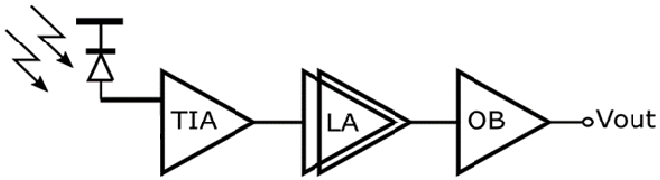


VCSEL output eye diagrams at 56 Gb/s (OptiGOT VCSEL Model)

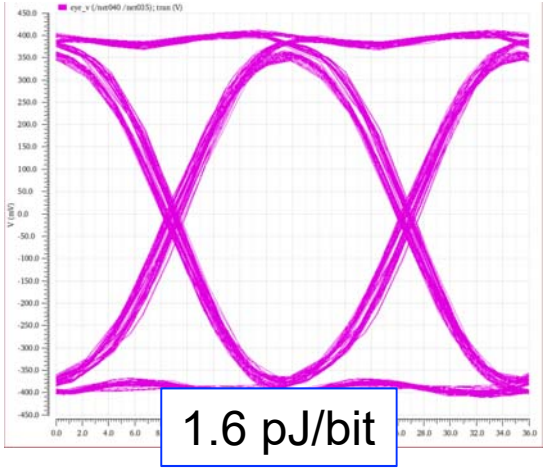




1x1mm

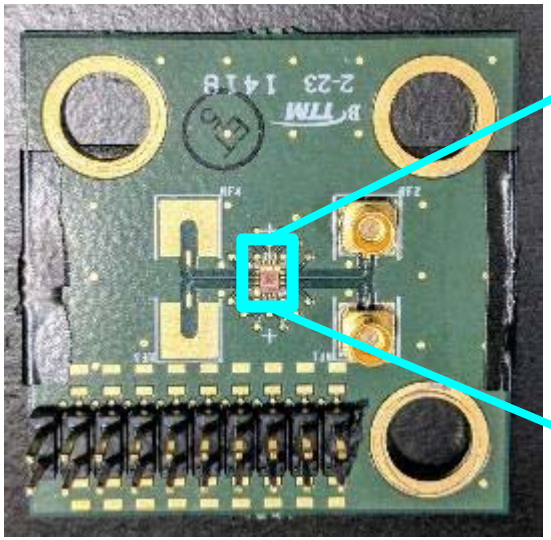


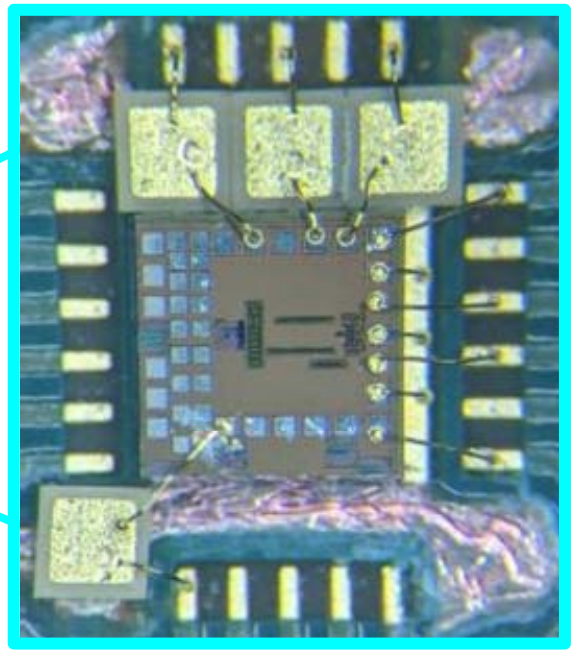
Receiver output eye diagram at 56 Gb/s



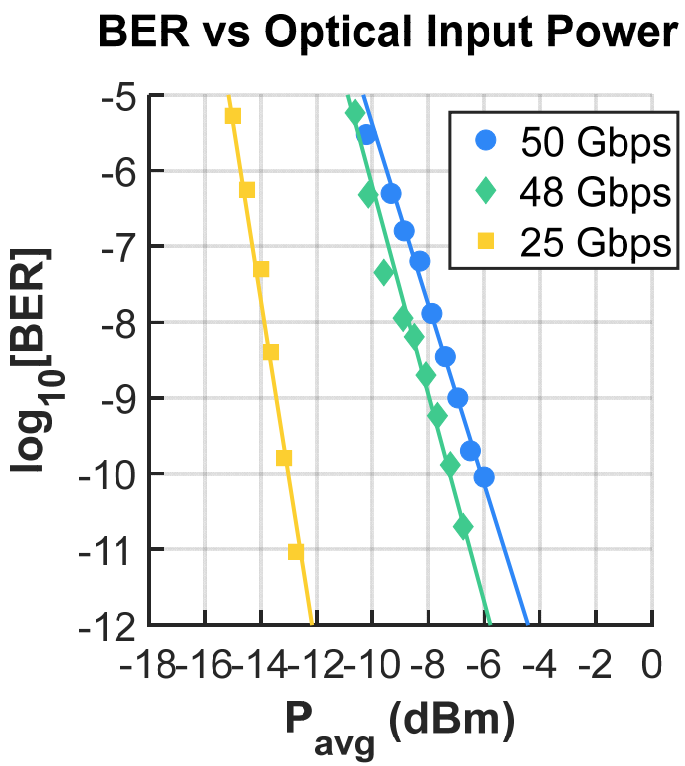
50Gb/s Receiver in IHP Monolithic Technology

Packaged Receiver

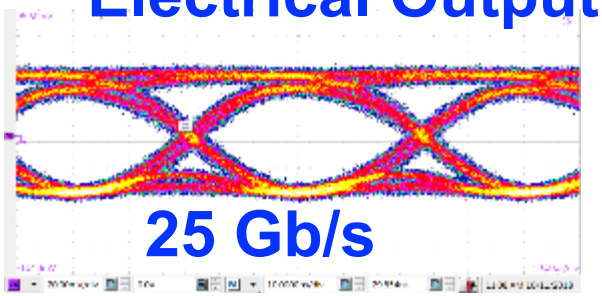


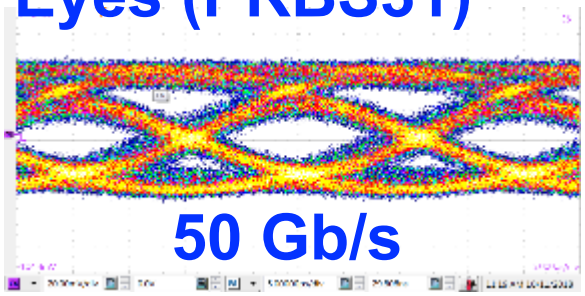


Sensitivity (PRBS31)



Electrical Output Eyes (PRBS31)





DC-coupled pseudo-differential TIA with monolithic Ge PD

Outlook/Next Steps

- Characterize all first-pass hardware
  - Update and refine link performance and power modeling
- First-pass OPLL circuits to be taped out in November
- Design second pass hardware and submit for fabrication
  - Electronics, SiP, InP
- Meet Phase 1 metrics
- Design packaging and interfaces, and define plan for live datacenter demo in Phase 2

Tech to Market

Katharine Schmidtke

UCSB facebook

T2M Critical Metrics

- **Power consumption:** extrapolating to the 51T switch generation, the power consumed by optical interconnects would exceed the available power in a typical data center.
- **Reliability:** In the future, even at low failure rates the number of pluggable modules that would need to be replaced exceeds what is practical.
- **Cost:** Cost per Gb/s is decreasing steadily, but because the bandwidth deployments are increasing rapidly, the overall spend on optical interfaces is increasing and will dominate the data center network infrastructure.
- **Manufacturing volume:** in the time frame of 51T-generation switches, the number of optical interconnects required to support data center infrastructure will exceed the current projected world-wide supply.

Business Models for Switch Integration

- **Switch driven:** Traditional switch ASIC companies could sell switch ASICs with photonic interconnects by embedding the photonics into the package with their switch ASIC.
- **Photonics driven:** Photonics companies could supply a photonically enabled switch chip by packaging the switch ASIC with their photonics.
- **OEM driven:** Traditional switch OEM's could develop products with photonic components and ASIC chips packaged together.
- **Customer driven:** End-users could design and develop solutions based on switch ASIC and photonics chips and package them together using 3rd parties.

Building an Ecosystem: Need for Open Interfaces

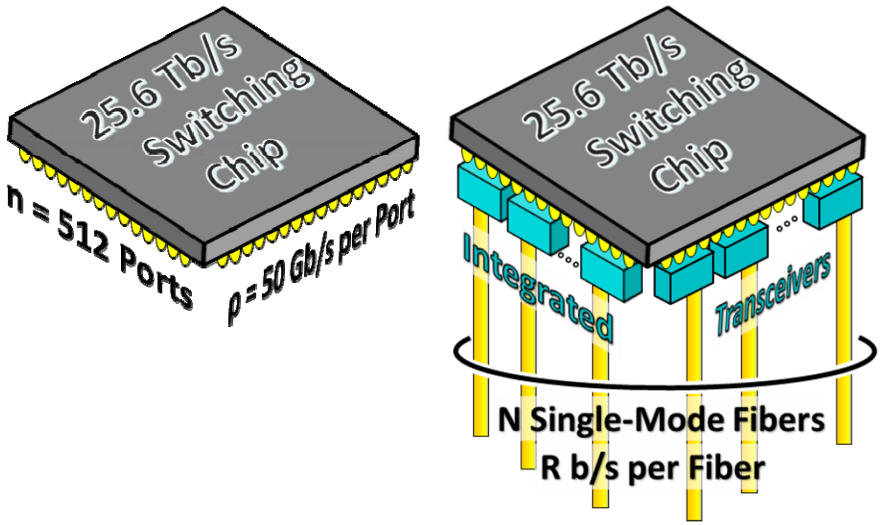
- INTREPID interface standards shared with the industry through open forums such as OCP, TIP or other industry MSA group
- The open interface would define:
  - The electrical interface including electrical signaling, performance parameters, and test points.
  - The physical interface in terms of outer dimensions, electrical pinout, and physical connections.
  - The thermal constraints and power consumption.

Network Architecture

Adel Saleh

UCSB facebook

Photonic Integration vs. Scalability in DCNs



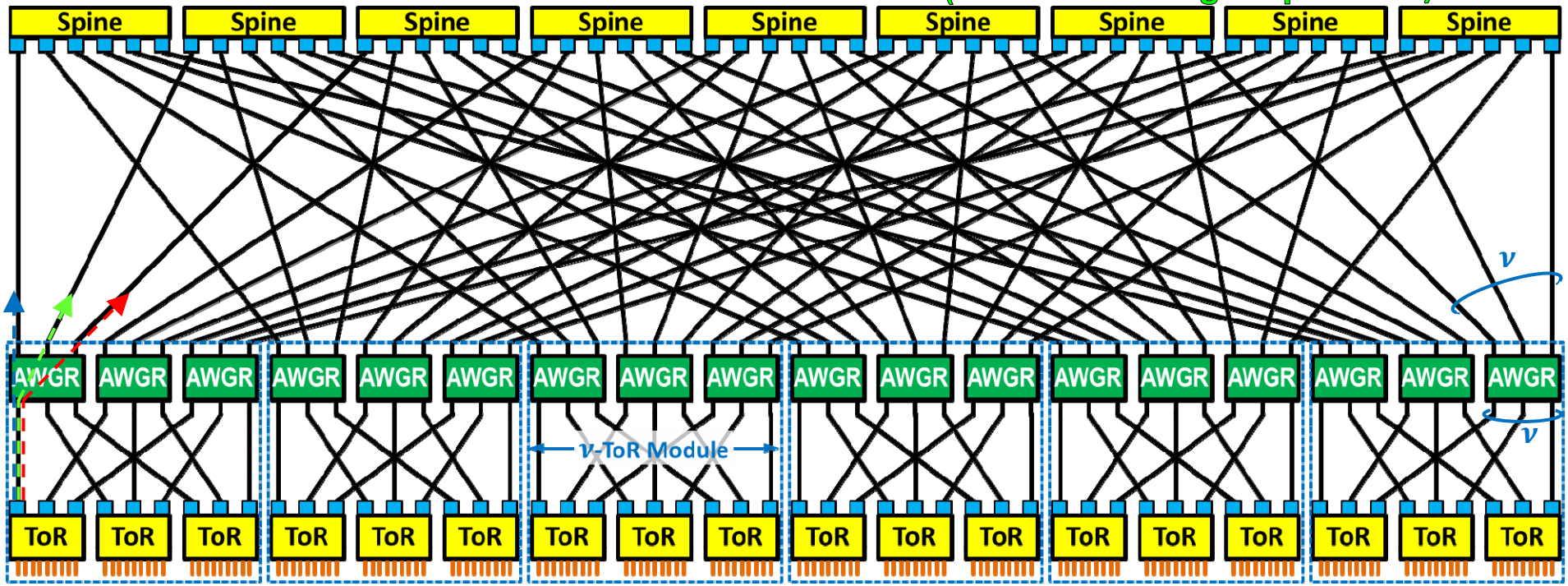
- Integration improves cost, size and energy efficiency, so it is good!
- However, more integration lowers the integrated switch radix and thus reduces the number of servers
- Adding an AWGR layer recovers the radix and number of servers

INTREPID Coherent WDM Integration					25.6 Tb/s Switching Chip n = 512 ports @ ρ = 50 Gb/s			Folded Clos with L Levels of Electronic Switches			*Adding One AWGR Layer*	
Modulation Multiplexing			# of λ's per Fiber	Bit Rate per λ (Gb/s)	Bit Rate per integrated Port or per Fiber (Gb/s)	Degree of Integration *(# Chip Ports per Integrated Port)*	Number of Integrated Ports (RADIX)	Number of 50G Servers $2\,(R/\rho)\,(N/2)^L$			# of 50G Servers $2\,\nu\,(R/\rho)\,(N/2)^L$	
bits per symbol								L = 2	L = 3	L = 4	L = 2	L = 3
QPSK	Dual Pol	μ	ν	r = μρ	R = ν r	R / ρ	N = n ρ/R					
✓	×	2	1	100	100	2	256	65,536	*8.E+06*	*1.E+09*	65,536	*8.E+06*
✓	×	2	2	100	200	4	128	32,768	2,097,152	*1.E+08*	65,536	*4.E+06*
✓	×	2	4	100	400	8	64	16,384	524,288	*2.E+07*	65,536	2,097,152
✓	✓	4	4	200	800	16	32	8,192	131,072	2,097,152	32,768	524,288
✓	✓	4	8	200	1,600	32	16	4,096	32,768	262,144	32,768	262,144

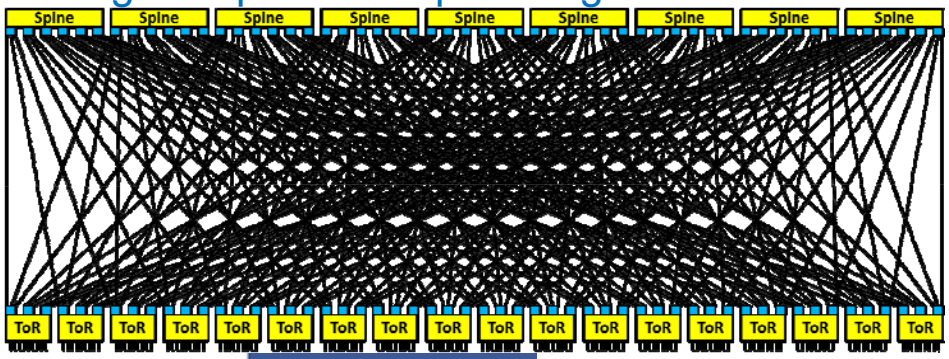
Alternative Solutions for Scaling-Out the Data Center

(Note: All links use WDM / SM fibers, except those under ToRs, which use VCSEL / MM fibers)

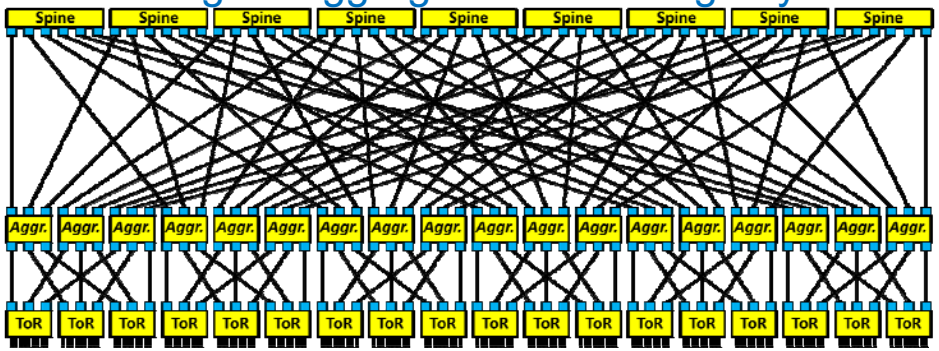
***INTREPID* Scaled-Out AWGR-Based Architecture (with ν wavelengths per fiber)**



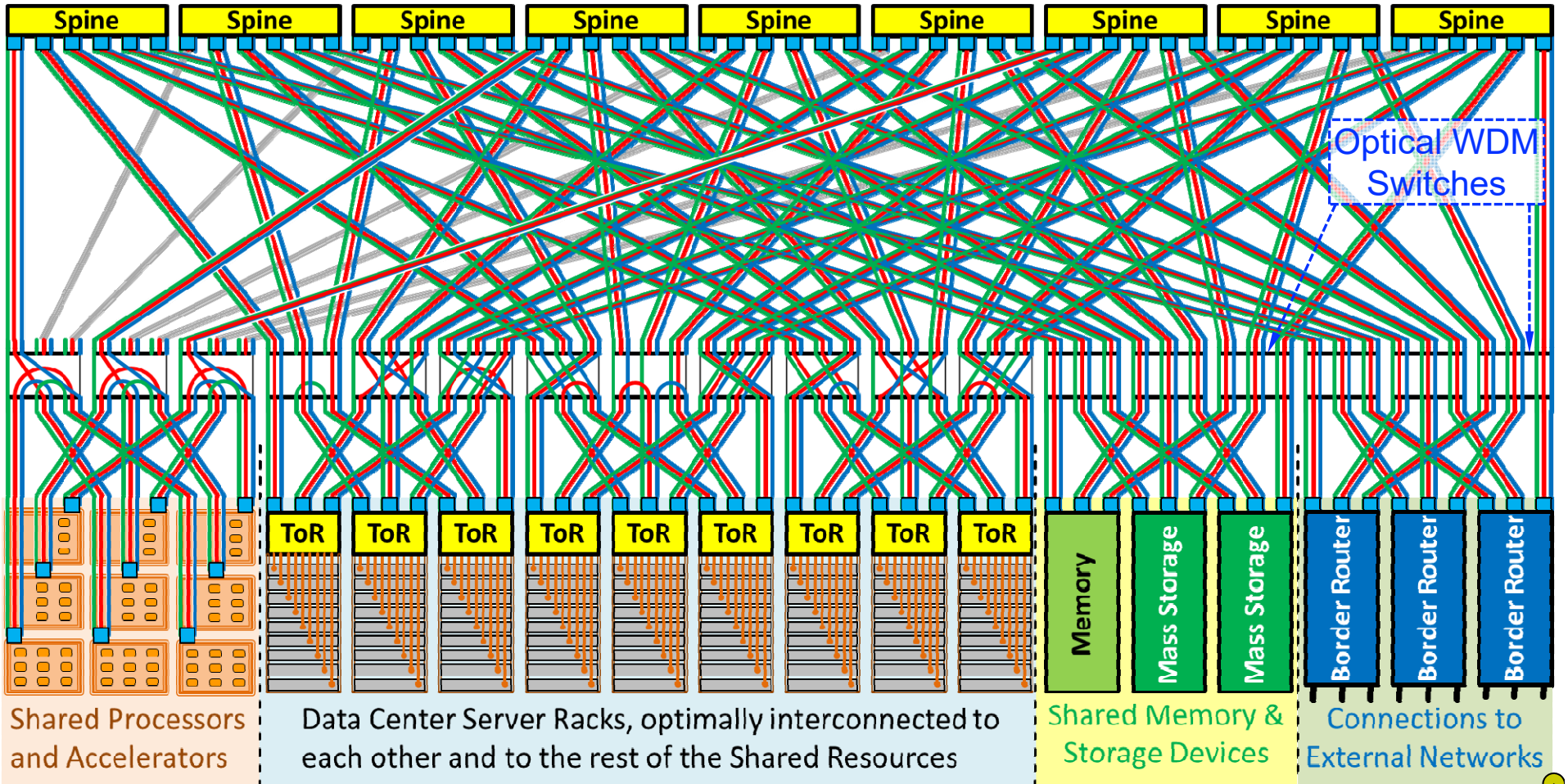
Using Multiple Fibers per Integrated Transceiver



Adding an Aggregation Switching Layer



In Phase 2, the AWGRs will be Replaced by Optical WDM Switches

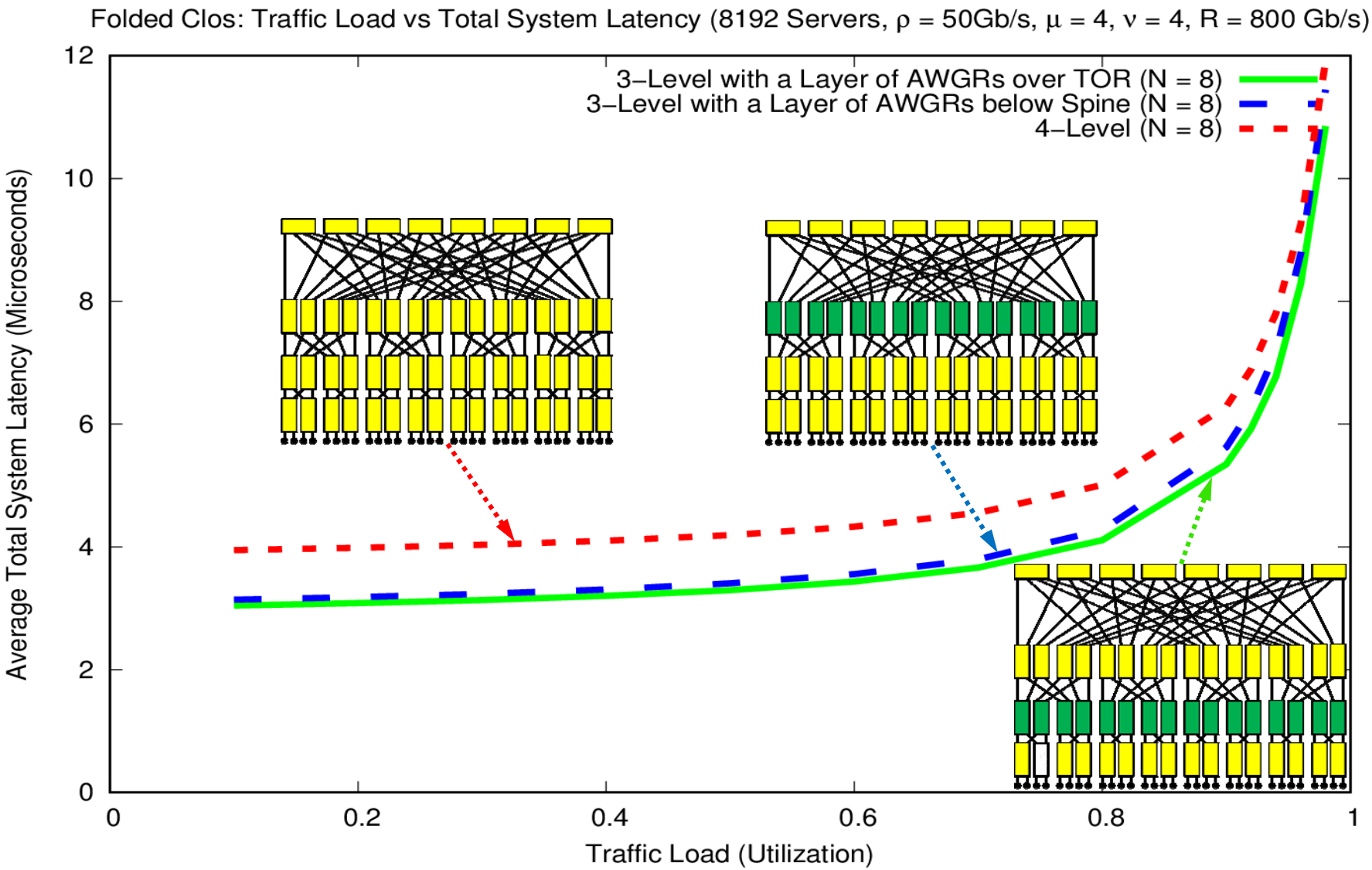


Optical-Switch-Based Architecture enables disaggregation, configurability to match workload variations (i.e., high utilization), and further energy efficiency

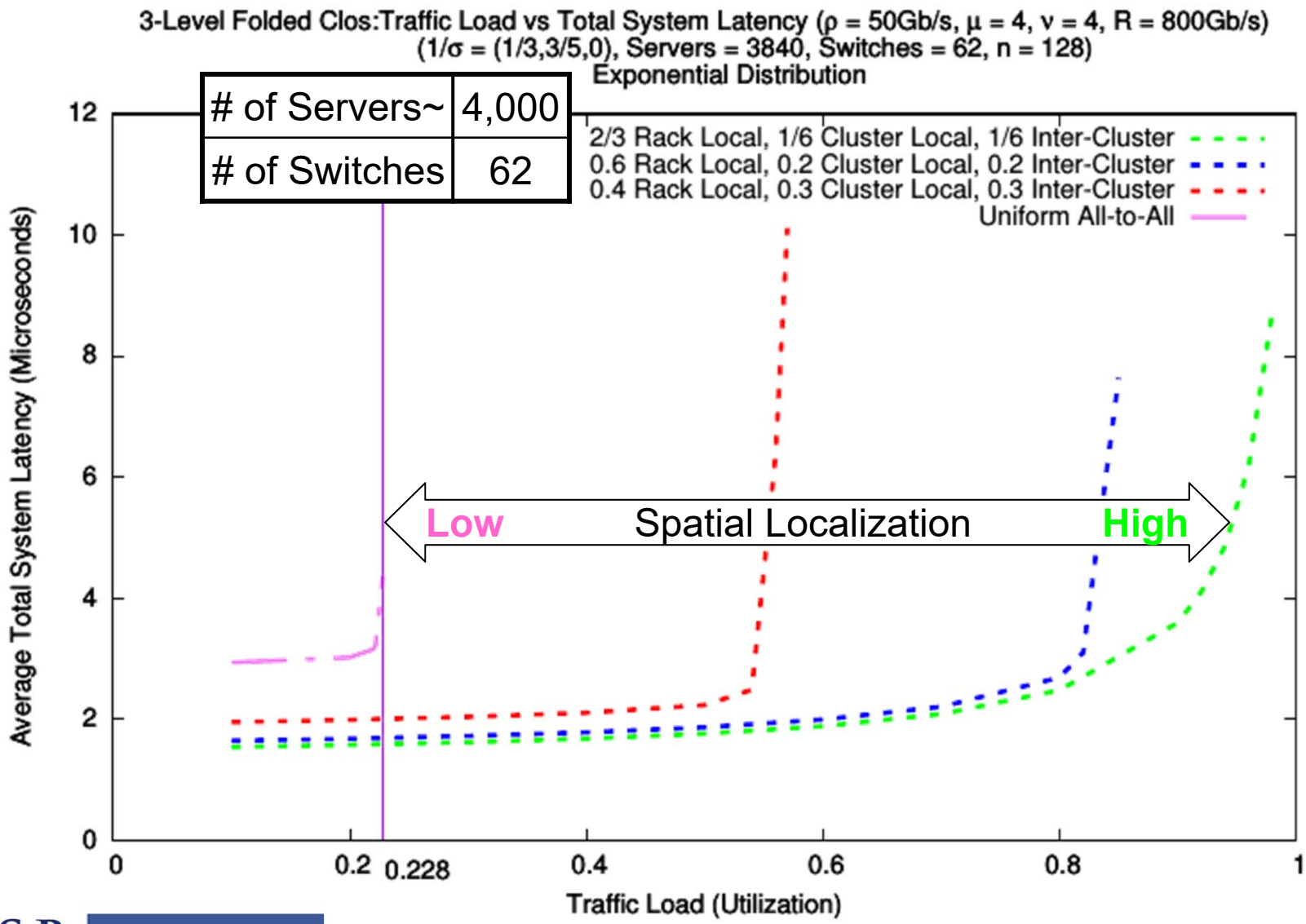
Network Architecture Status

- Data center interconnect network designs under consideration:
  - There are good reasons why the folded Clos architecture is used in all hyperscale data centers (scalability, cost effectiveness and fault tolerance)
  - We have developed general designs for this class of architectures, with added layers of AWGRs for scalability and energy efficiency
- Developed a flexible Python design / simulation tool, featuring:
  - Any number of switching levels and AWGR layers, different switch sizes, and arbitrary oversubscription ratios
  - General spatial traffic distribution, from uniform all-to-all to varying ratios of rack-local, cluster-local and inter-cluster traffic
  - General temporal traffic distribution, from non-bursty to bursty, and from Exponential to Weibull to Log Normal packet arrival process
  - This program is now used to evaluate our various data center designs
  - Some results will be shown next …

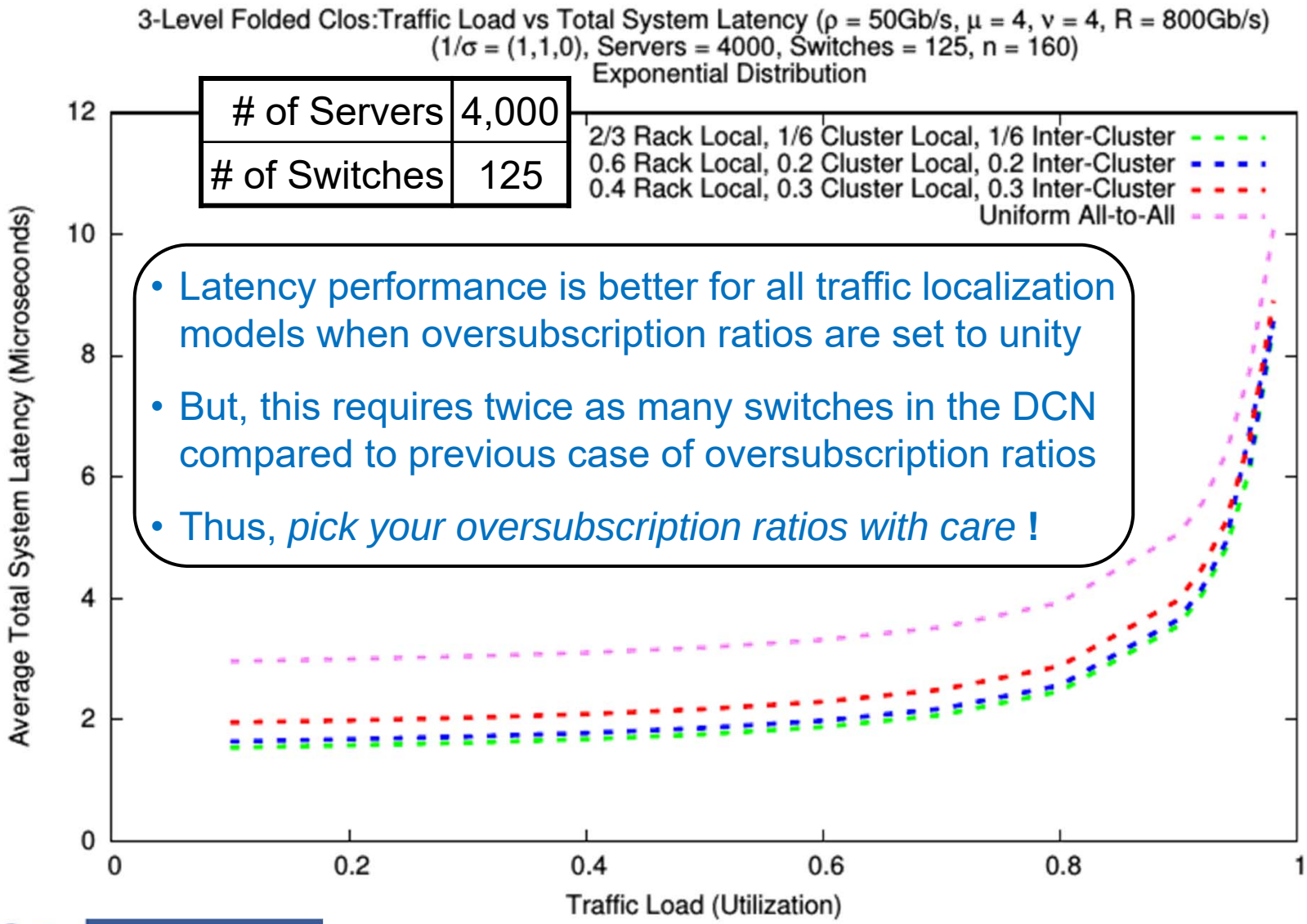
Simulations of a 4-Level Folded Clos vs. a 3-Level Folded Clos with an AWGR Layer above ToRs or under Spines



Oversubscription ratios optimized for a given spatially localized traffic, then traffic localization is reduced in steps, down to no localization



Now, let all oversubscription ratios be set to unity , then observe performance for the same set of spatial localizations



# of Servers	4,000
# of Switches	125

- Latency performance is better for all traffic localization models when oversubscription ratios are set to unity
- But, this requires twice as many switches in the DCN compared to previous case of oversubscription ratios
- Thus, *pick your oversubscription ratios with care* **!**

Outlook/Next Steps (continued)

- Characterize all first-pass hardware
  - Update and refine link performance and power modeling
- First-pass OPLL circuits to be taped out in November
- Design second pass hardware and submit for fabrication
  - Electronics, SiP, InP
- Meet Phase 1 metrics
- Design packaging and interfaces, and define plan for live datacenter demo in Phase 2
- Incorporate various energy efficiency figures of merit, from *Joules per bit* to *Joules per Transaction*, in our design tool
- Assess quantitively the impact of replacing the AWGRs with agile optical switches on further improvements of the overall performance and energy efficiency of the data center