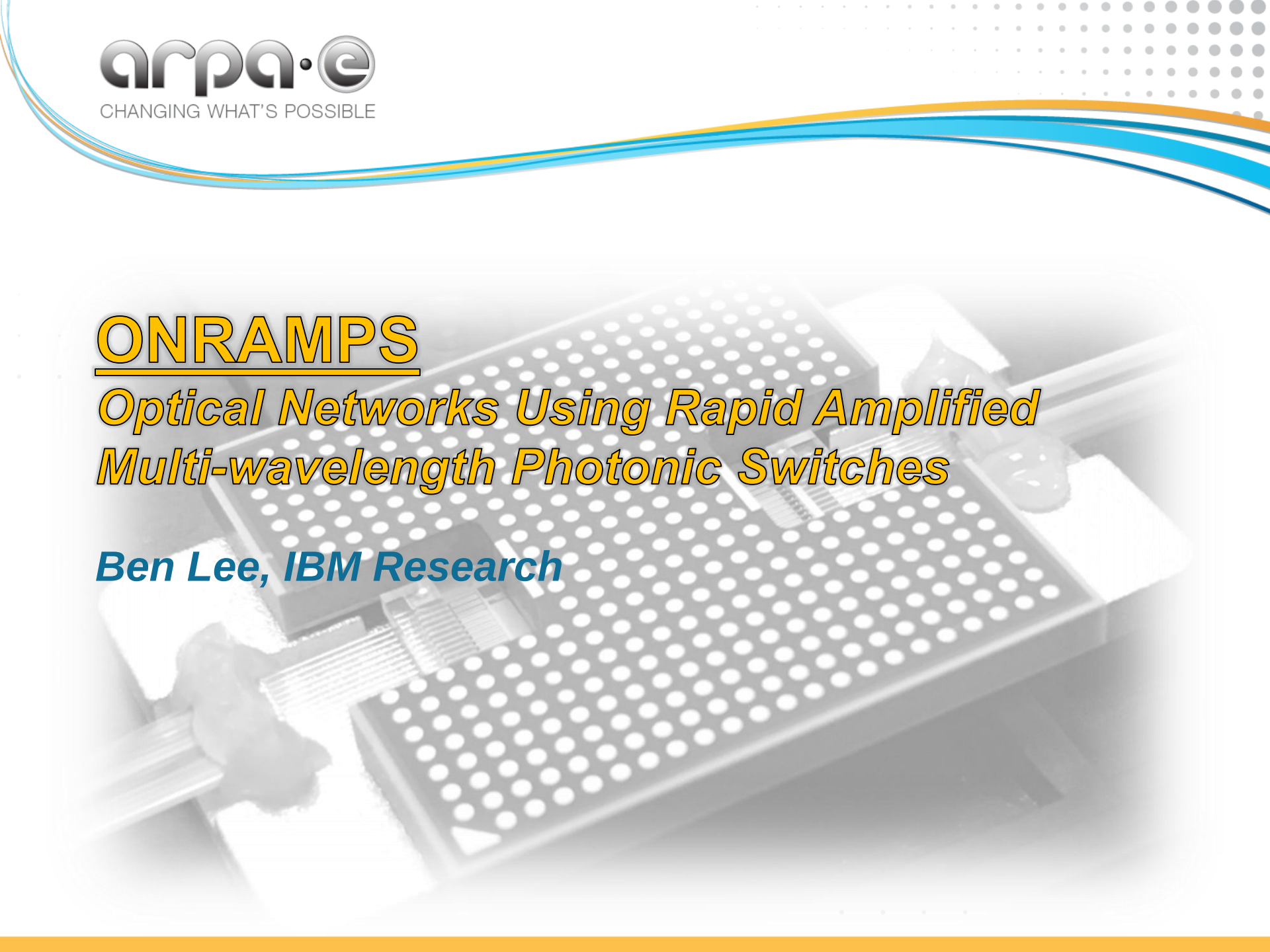
arpa·e
CHANGING WHAT'S POSSIBLE

ONRAMPS

Optical Networks Using Rapid Amplified Multi-wavelength Photonic Switches

Ben Lee, IBM Research



A Diverse Team of Experts

- **IBM Research**
  - Hershel Ainspan — *Circuit Design and Layout*
  - Chris Baks — *Electrical Package Design*
  - Alan Benner — *Network Simulation & Applications*
  - Fuad Doany — *Optical and Electrical Packaging*
  - Nicolas Dupuis — *IC Design and Test*
  - Ben Lee — *Principal Investigator*
  - Pavlos Maniotis — *Network Simulation & Applications*
  - Mounir Meghelli — *Mixed-Signal Circuit Design*
  - Jon Proesel — *Mixed-Signal Circuit Design*
  - Laurent Schares — *System Integration and Network Simulation*
  - Marc Taubenblatt — *Techno-economics, Applications, & Outreach*
- **IBM Bromont**
  - Nicolas Boyer — *Microelectronics and Photonics Assembly*
  - Isabel De Sousa — *Techno-economics and Business Development*
  - Elaine Cyr — *Microelectronics and Photonics Assembly*
  - Nathalie Normand — *Microelectronics and Photonics Assembly*
- **Research Interns**
  - Alex Forencich (UCSD), Takako Hirokawa (UCSB), Pascal Stark (ETH), Anny Zheng (MIT)

Technical & Economic Drivers

- **ONRAMPS is developing the technology & eco-system for a:**
  - Low-cost → volume-compatible & automated assembly
  - Manufacturable → commercial fabrication and assembly lines✦
  - Fully packaged → optical, electrical, thermal, and mechanical
  - Fast reconfigurable → system switching in tens of nanoseconds✦
  - Photonic switch → scalable bandwidth with low power and latency

  ✦ *in Trusted On-shore facilities*
  ✦ *also works at lower speeds*

- **Once developed & matured, the technology can have an immediate impact in high-performance computing**
  - May also be used in datacenters, mini-ROADMs for telecom and 5G, avionics, edge networks, RF systems

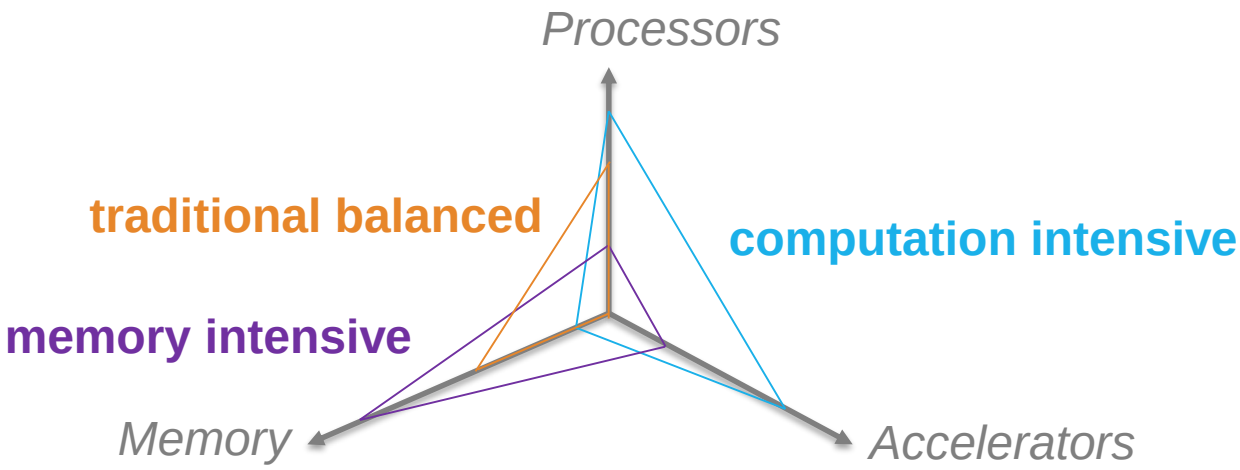
- **Metrics:**
  - Phase 1 → SNB 8×8, < 2 W (incl. control), < -20 dB aggregate crosstalk, ~ 7 dB on-chip loss
  - Commercial targets → SNB 32×32, 50 Tb/s, < 30 W, loss compensated, few ¢ / Gb/s

Publications and IP Portfolio

- **Journal Articles**
  - B. Lee, N. Dupuis, “Silicon Photonic Switch Fabrics: Technology and Architecture,” *J. Lightw. Technol.* [**Tutorial**], Jan 2019.
  - N. Dupuis *et al.*, “Nanosecond Photonic Switch Architectures Demonstrated in an All-Digital Monolithic Platform,” *Opt. Lett.* [**Editor’s Pick**], Aug. 2019.
  - N. Dupuis *et al.*, “Nanosecond-Scale Shift-and-Dump Mach-Zehnder Switch,” *Opt. Lett.*, Sep. 2019.
  - B. Lee *et al.*, “Fine Tuning of Mach-Zehnder Phase Using Low-Resolution Digital-to-Analog Converters,” *Photon. Technol. Lett.*, Oct 2019.
  - N. Dupuis *et al.*, “A 4×4 Electrooptic Silicon Photonic Switch Fabric with Net Neutral Loss,” *J. Lightw. Technol.* [**Invited**], pre-print available online.
- **Conference Papers & Presentations**
  - B. Lee, “Photonic Switching Platform for Datacenters Enabling Rapid Network Reconfiguration,” Photonics West [**Invited**], Jan 2018.
  - B. Lee, “Photonic Switch Fabrics in Computer Communications Systems,” OFC [**Tutorial**], Mar 2018.
  - A. Forencich *et al.*, “System-Level Demonstration of a Dynamically-Reconfigured Burst-Mode Link Using a Nanosecond Si-Photonic Switch,” OFC, Mar 2018.
  - M. Taubenblatt, “Optical Interconnects for Large Scale Computing Systems: Trends and Challenges,” OSA Advanced Photon. Cong. [**Invited**], Jul 2018.
  - L. Schares, “Photonic Switch Fabrics in Computer Communications Systems,” Photonics in Switching and Computing [**Invited**], Sep 2018.
  - B. Lee, “Toward Optical Networks using Rapid Amplified Multiwavelength Photonic Switches,” OFC [**Invited**], Mar 2019.
  - N. Dupuis *et al.*, “A Nonblocking 4x4 Mach-Zehnder Switch with Integrated Gain and Nanosecond-Scale Reconfiguration Time,” OFC [**Top-Scored**], Mar 2019.
  - M. Taubenblatt, “Optical Interconnects in Data Centers,” ECOC [**Tutorial**], Sep 2019.
  - L. Schares, “Enabling New Compute Architectures with Co-packaged Optics and Photonic Switching,” ECOC [**Invited**], Sep 2019.
  - B. Lee *et al.*, "Coarse-Fine Control of Dual-Tuner Mach-Zehnder Interferometer Using Identical Low-Resolution DACs," IEEE Photon. Conf., Oct 2019.
- **Workshops & Short Courses**
  - L. Schares, *Panelist*, “Integrated or disaggregated data centres? Challenges and opportunities,” ECOC, Sep 2018.
  - L. Schares, *Panelist*, “What is the role of optical switching technologies in data centres and computing communication systems?” ECOC, Sep 2018.
  - B. Lee, *Panelist*, “Opportunities and Challenges for Optical Switching in the Data Center,” OFC 2019.
  - B. Lee, *Short Course*, “Photonic Switching Systems,” OFC, Mar 2019.
- **Ongoing IP efforts reported through *iEdison* related to control circuits, devices, assembly, initialization and optimization, system implementation**

Building for Efficient Next-Gen Computing

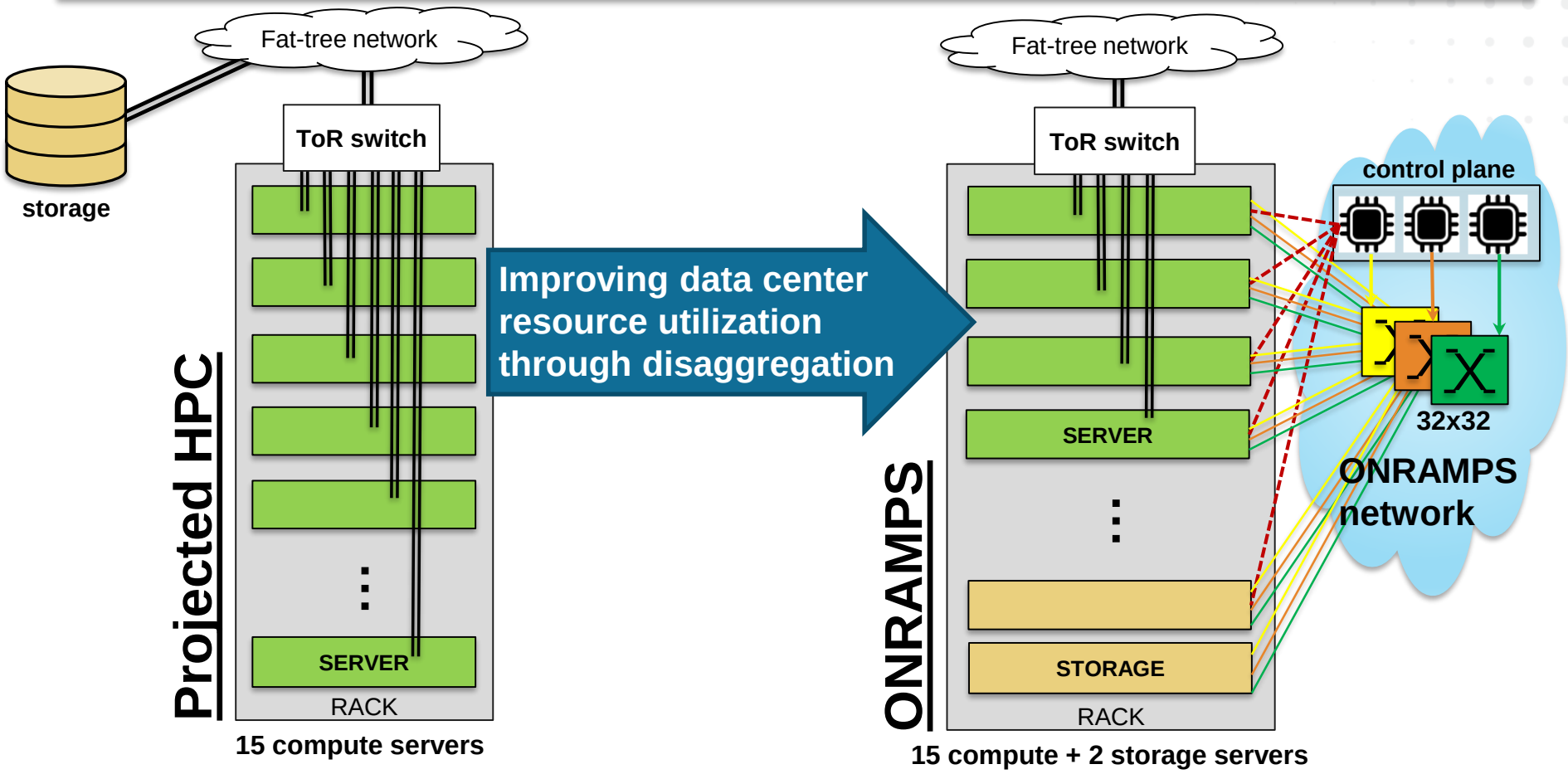
- Traditionally, HPC systems optimized for scientific computing
- Next-gen systems target a diverse suite of applications:
  AI, Distributed Deep Learning, Graph Analytics, Scientific Computing, and more
- … with a wide range of CPU, GPU, memory, and data movement requirements
- … which may evolve at a faster pace than HW upgrade cycles.



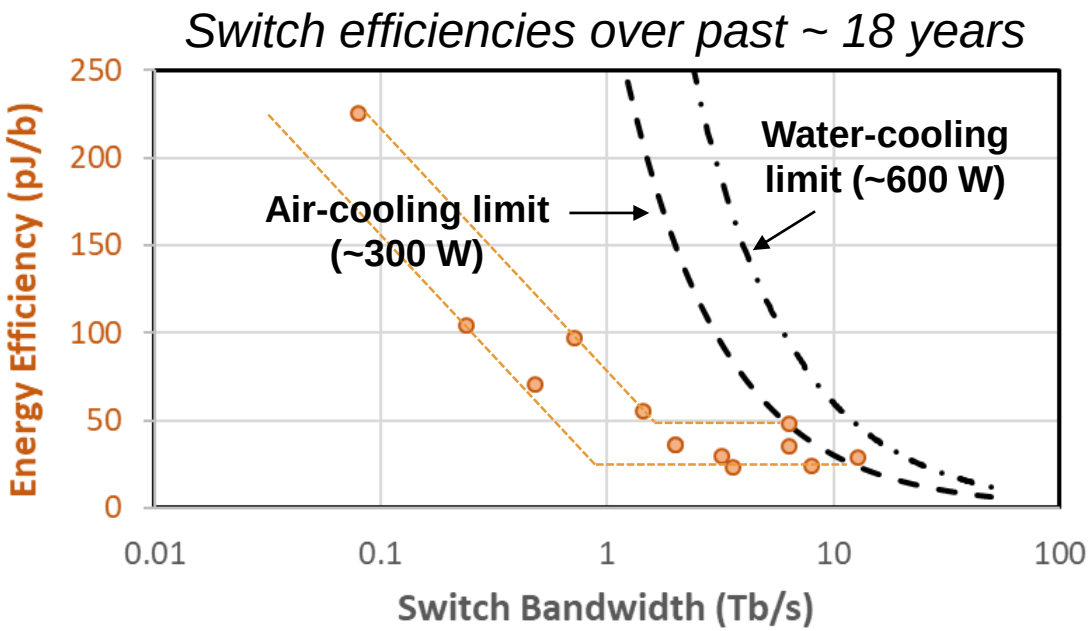
- Efficient next-gen systems across these diversity of workloads require:
  (1) flexible provisioning of resources to applications on demand, and
  (2) tight coupling of provisioned resources with high-bandwidth & low-latency connectivity.

ONRAMPS Alleviates Inefficiencies of Future Architectures

ONRAMPS enables:
On-demand allocation, Flexible resource ratios, Memory disaggreg., Direct links to storage

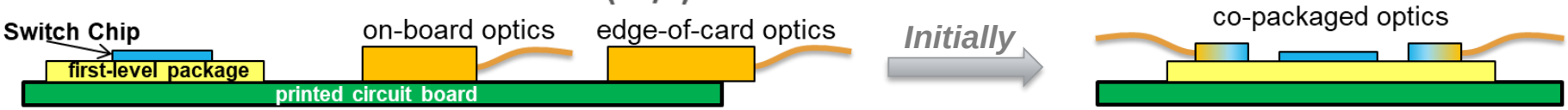


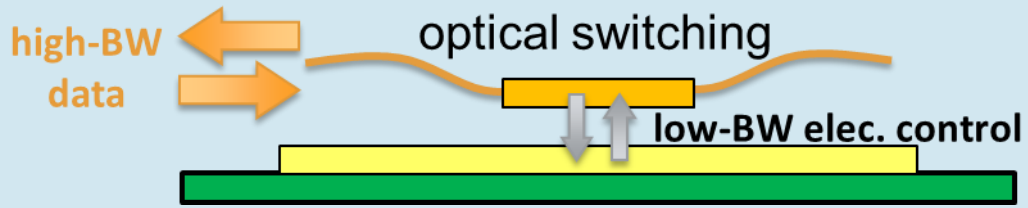
Electrical Packet Switching ASICs Hitting Thermal Limits – Need New Approach



- Bandwidth 2× every 2-3 years
- Since ~ 2012, efficiencies flat at ~ 25-50 pJ/b
- Air-cooling limited now
- Water-cooling limited soon

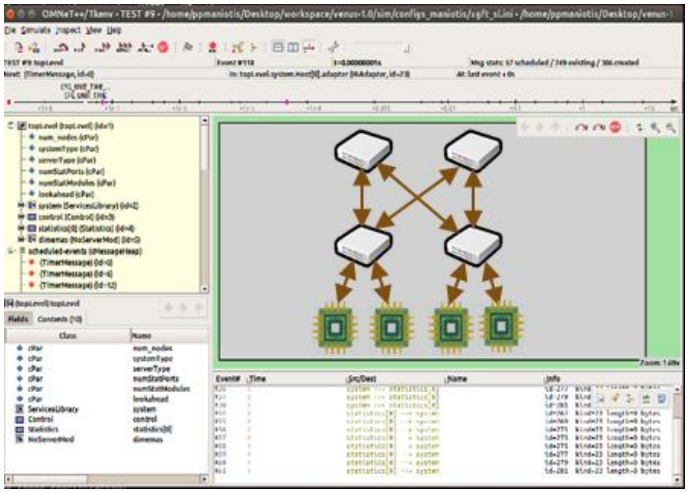
Need a new approach!





- Large optical bandwidth
- Energy used to configure pipes, not process & transmit bits
- Agnostic to data rates and formats
- **Must be low latency & low cost!**

Modeled System Performance



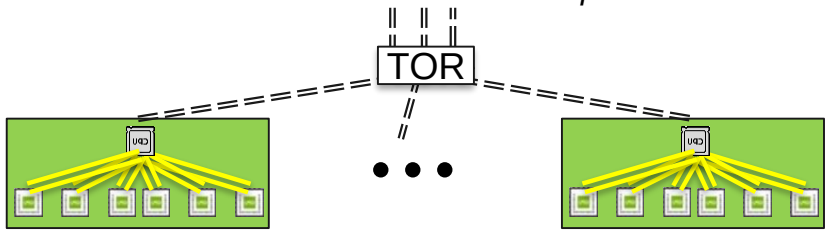
VENUS Network Simulator

- Discrete-event simulator built on **Omnet++**
- 140k lines of C/C++ code
- Developed at IBM's Zurich Research Lab (ZRL)
- Used in development of multiple HPC generations
- Fat tree, XGFT, Mesh, Multi-dimensional mesh, Hypercube, Torus, Dragonfly(+), Flattened butterfly, …
- Ethernet, InfiniBand, Co-packaged optics, Optical switches, …

Comparative Analysis

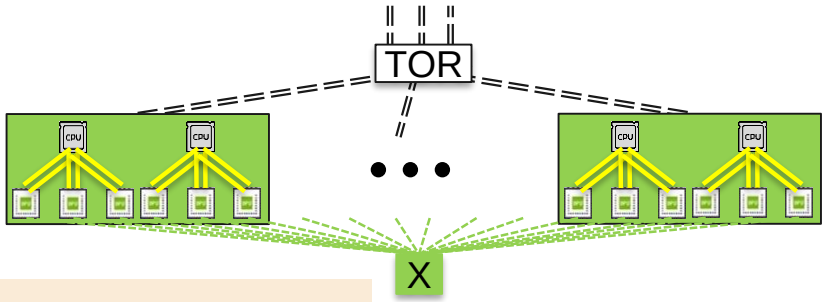
Projected HPC

6 blades with 1 CPU + 6 GPUs per blade



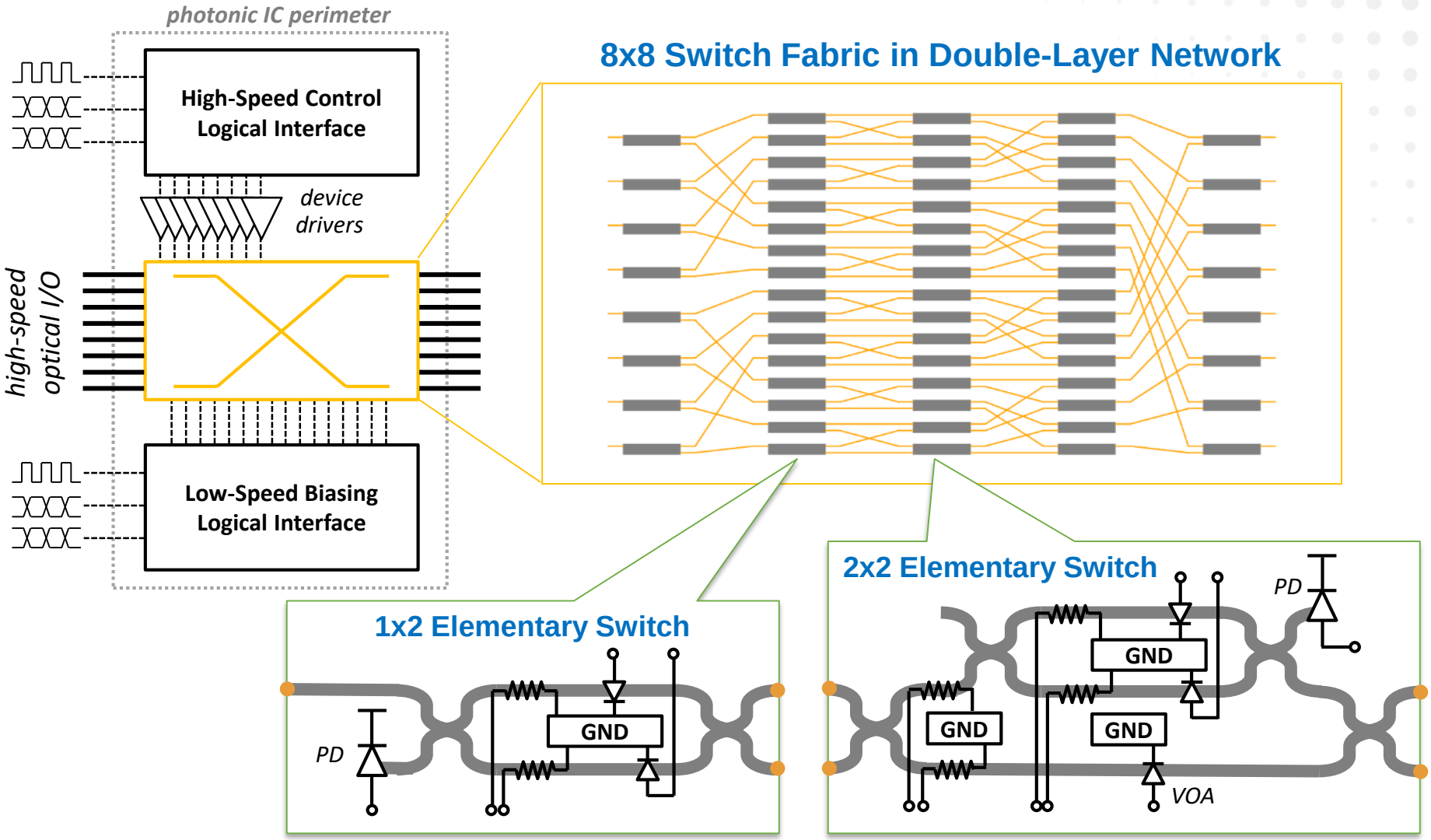
ONRAMPS

6 blades with 2 CPUs + 6 GPUs per blade



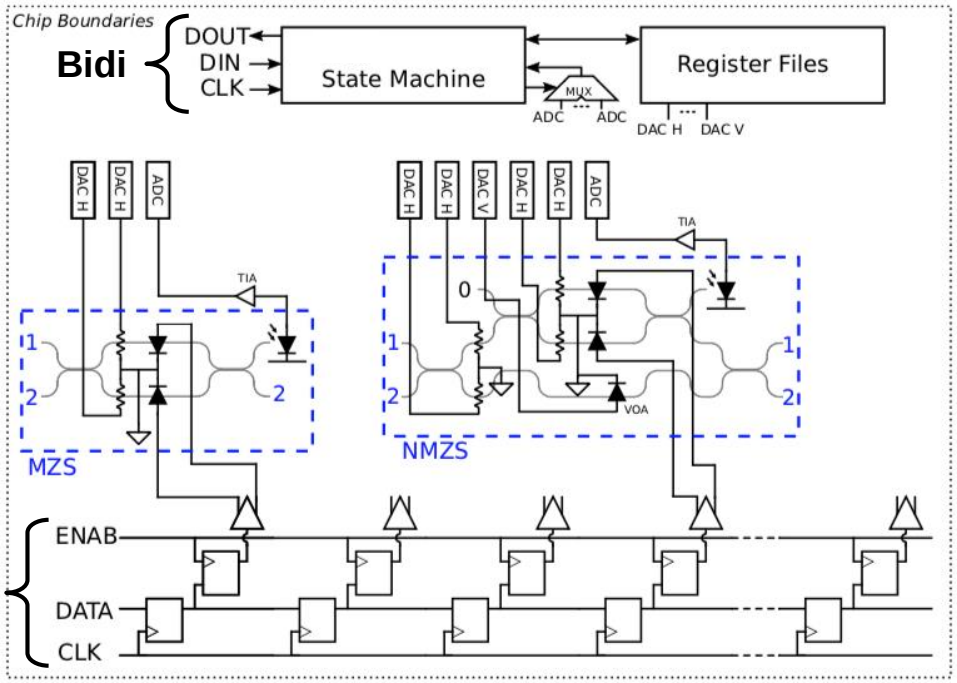
- Data rate: 100 Gb/s per port
- TOR delay: 100 ns
- NIC/Adapter delay: 100 ns
- ONRAMPS scheduling delay: 10 ns
- ONRAMPS system-level switch delay: 20, 40, 80 ns (*includes physical switching, control, and link training*)

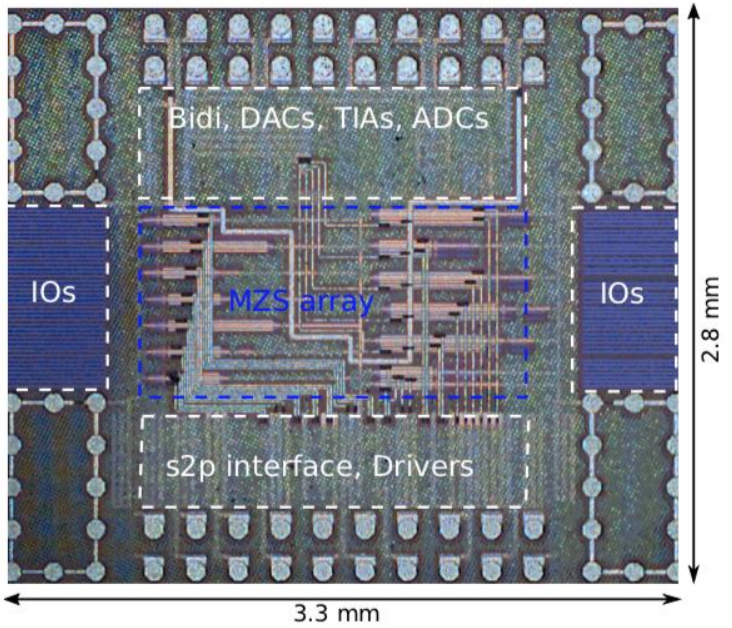
8x8 Photonic Switch Integrated Circuit



Initial IC Design, Fabrication, and Test

- **Goal**: validate photonic/electronic blocks and digital interfaces
  - DAC drivers for TO phase tuner, DAC drivers for VOA, Binary complementary drivers for EO phase shifter, TIA+ADC for power monitor, high-speed serial-to-parallel (s2p) interface, low-speed control interface (Bidi), registers
  - TO phase tuners, EO phase shifters, directional couplers, waveguide crossings, photodetectors





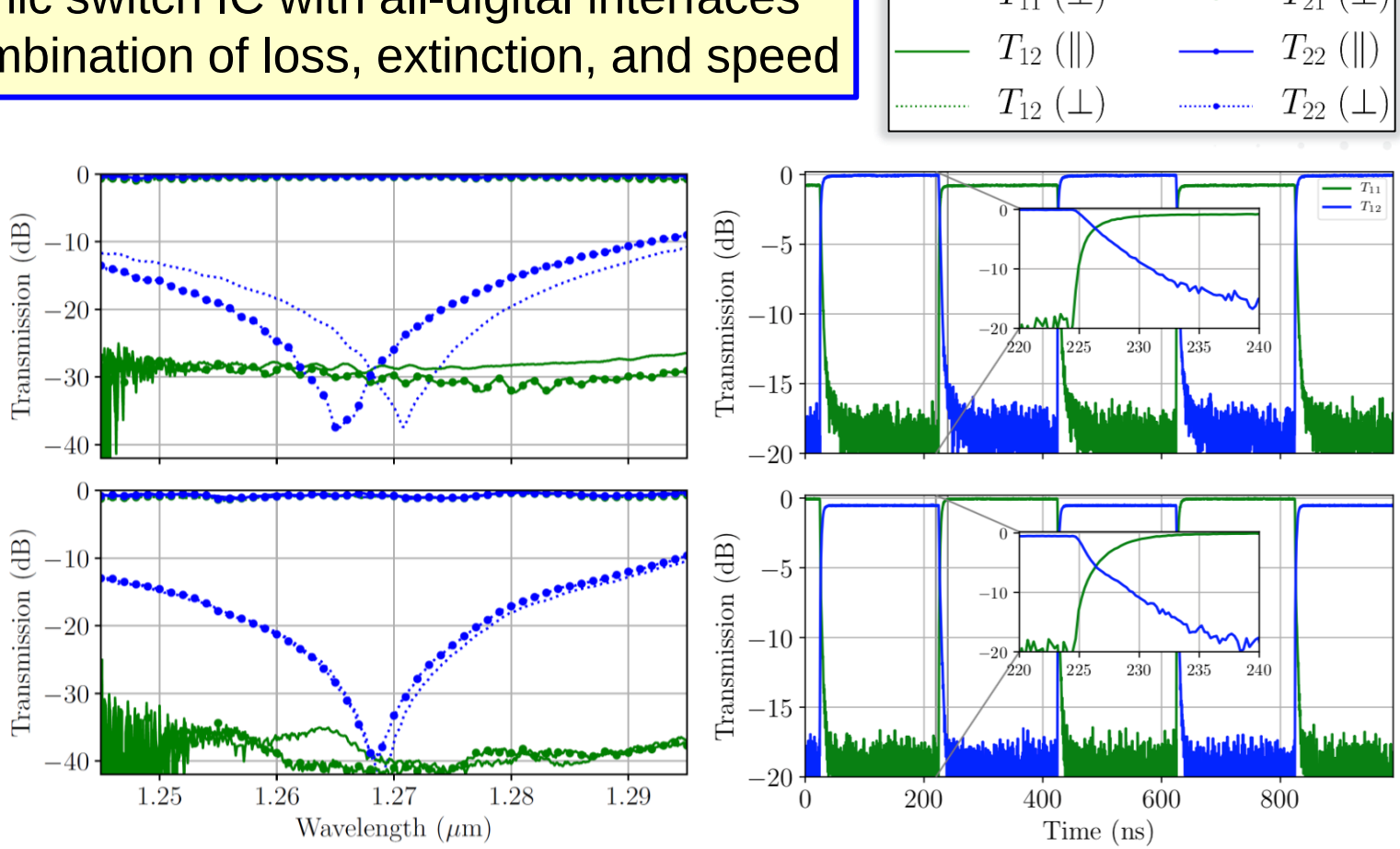
Digitally interfaced, programmable IC containing all the building blocks required to construct a scaled photonic switch fabric.

Initial IC Elementary Switch Performance

- First photonic switch IC with all-digital interfaces
- Record combination of loss, extinction, and speed

2x1 MZS
Loss: 0.8 dB
Extinction: 28 dB
Transient: 6 ns

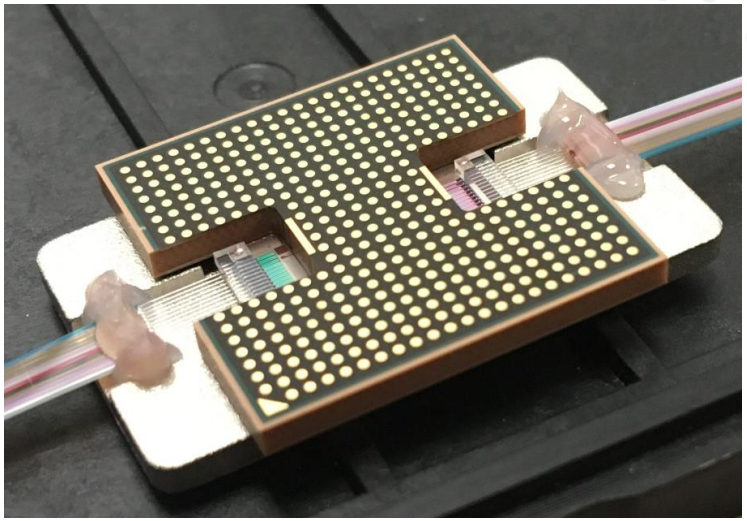
2x2 NMZS
Loss: 1.3 dB
Extinction: 38 dB
Transient: 6 ns

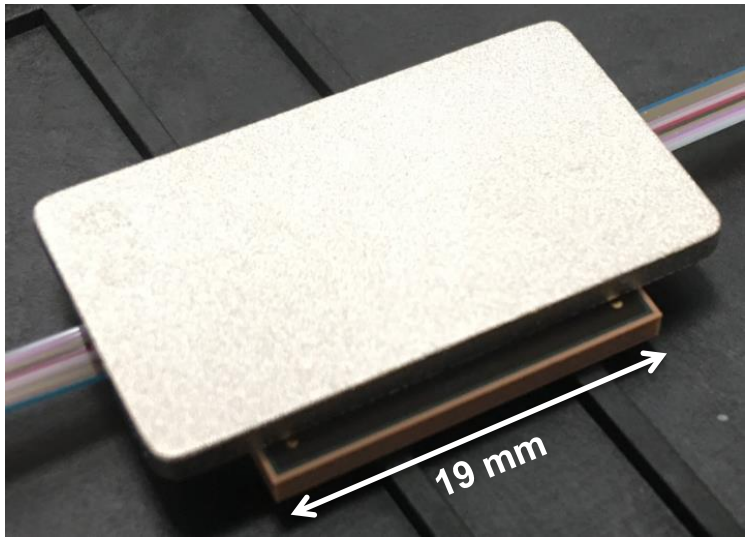


[N. Dupuis, Optics Letters 2019, **44** (15) 3610]

Preliminary Modules Assembled from Initial ICs

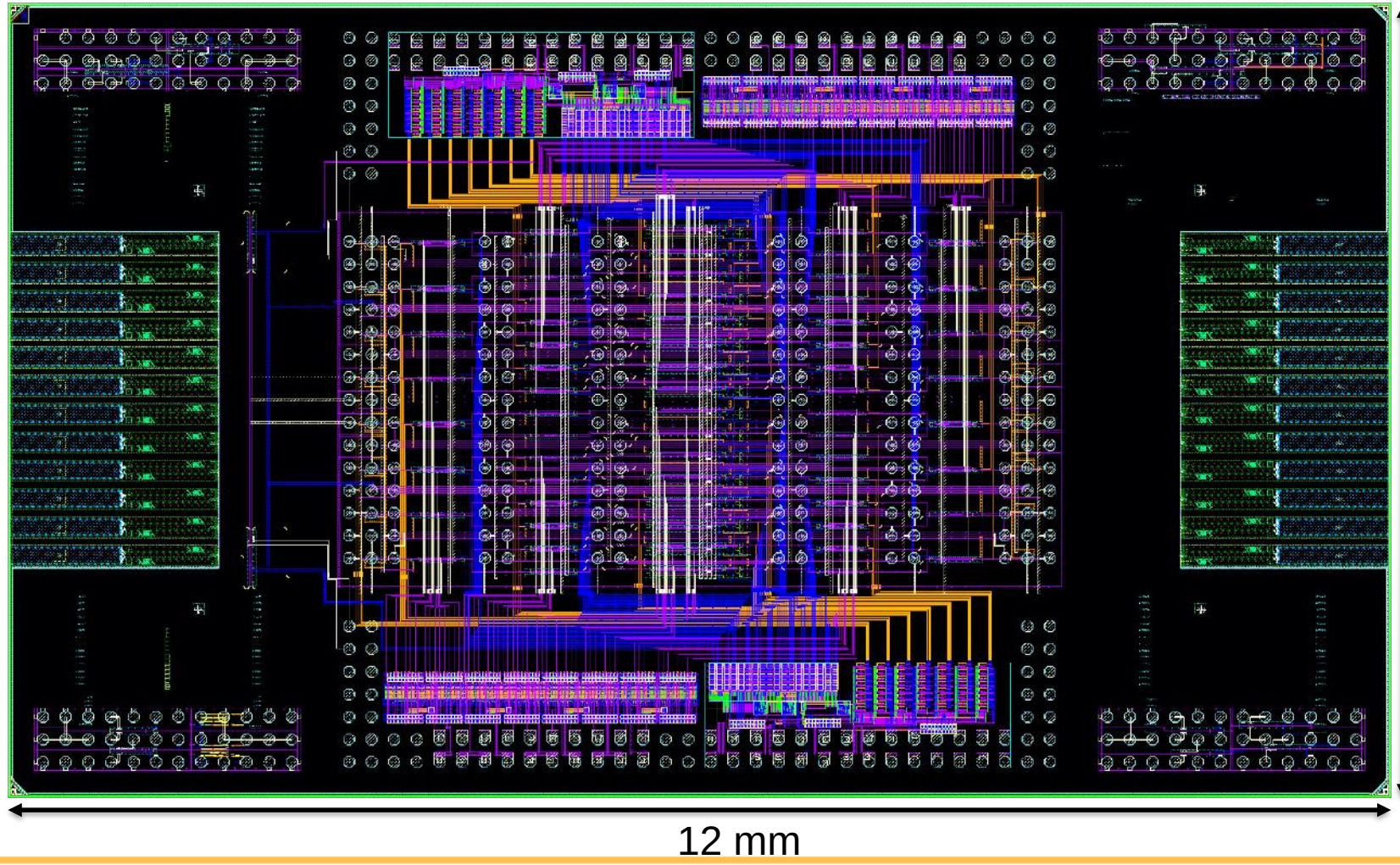
- Full optical, electrical, thermal package
- Chip joining process developed that preserves sensitive optical interfaces
- Full module assembly process developed and demonstrated
- Automated assembly on Ficontec tool
- Strain relief adhesive added manually (temporary solution)
- Full electrical connectivity
- Loopback loss < 3 dB/facet over 100nm spectral bandwidth for all 24 fibers (~ 1.5 dB/facet over 50 nm typical)





8x8 Photonic Switch Layout

64 MZS, 64 compl. drivers, 180 DACs, 112 ADCs, 72 WG crossings, 24 fiber couplers



Area set by packageability requirements

ONRAMPS' Completed Tasks, Works in Progress, and Needs Continued Resources

PDK support for critical photonic & electronic libraries

(monolithic) electronic circuits providing digital programming interfaces

automated assembly on high-volume capable tooling

optical, electrical, thermal, and mechanical packaging

cost models

fast switching transients

process-tolerant design and realization

fabrication in a commercial CMOS flow

8x8 switch module demo

FPGA control

bring-up and initialization

network modeling environment

fast dynamic reconfiguration demo

market entry strategy

ONRAMPS Scope

To be addressed before commercialization

network interfaces

protocol compatibility

scheduling for scale

link loss & SNR budgets

gain integration demo

technology maturation & transfer

system architecture

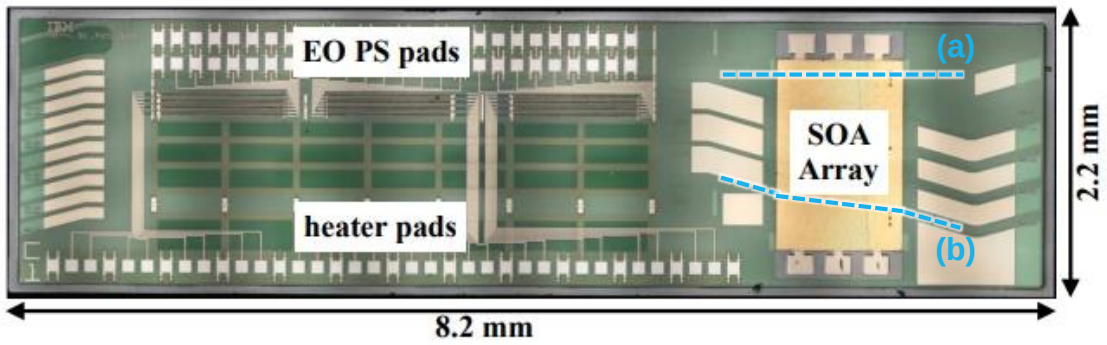
broadband device development

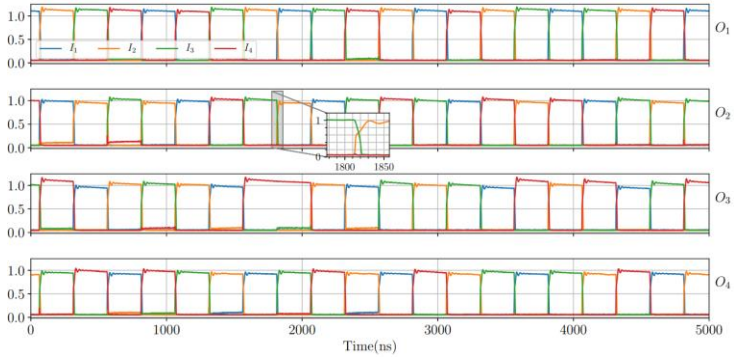
broadband photonic switch demo

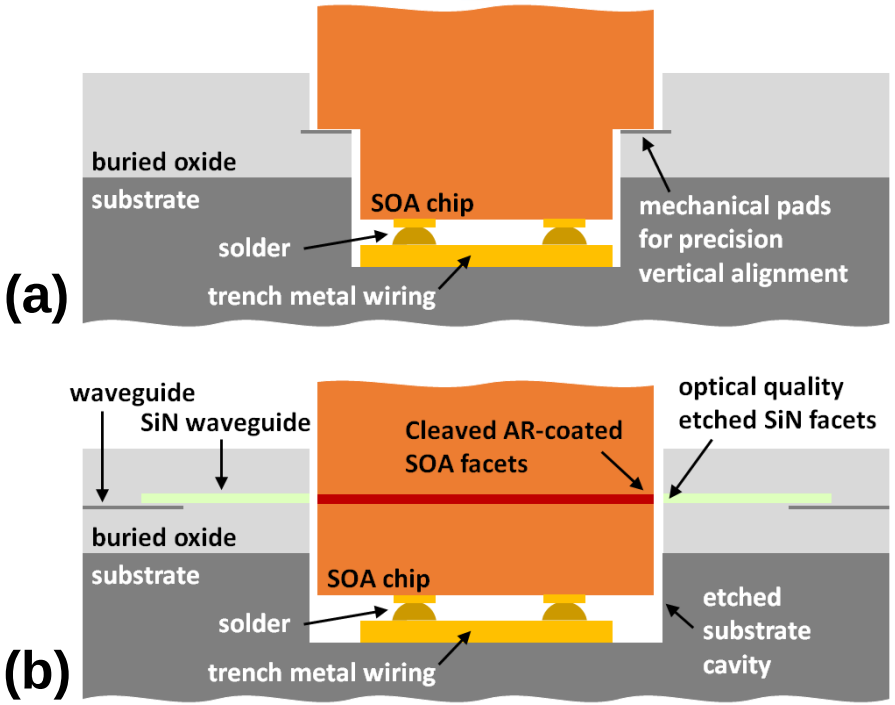
polarization handling approaches

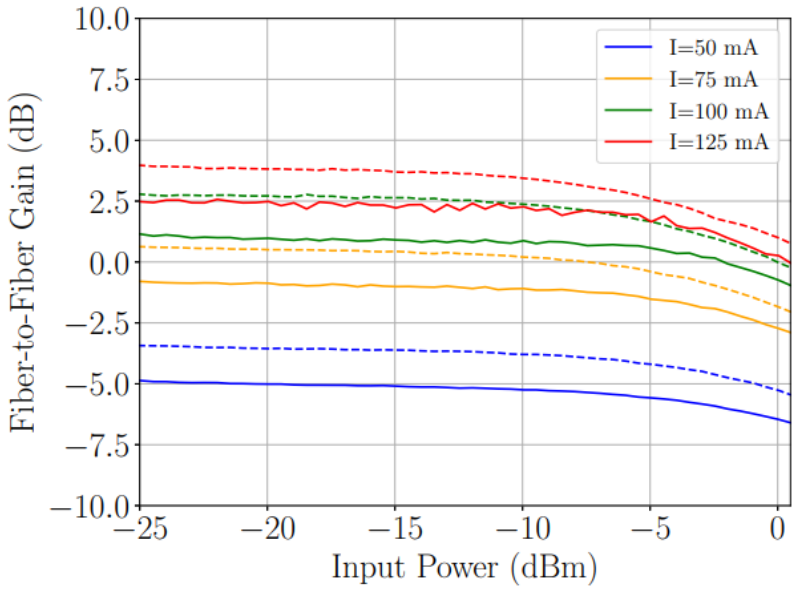
polarization-independent demo

Initial Validation of Low-Cost & Energy-Efficient Amplification Integrated with Fast OCS







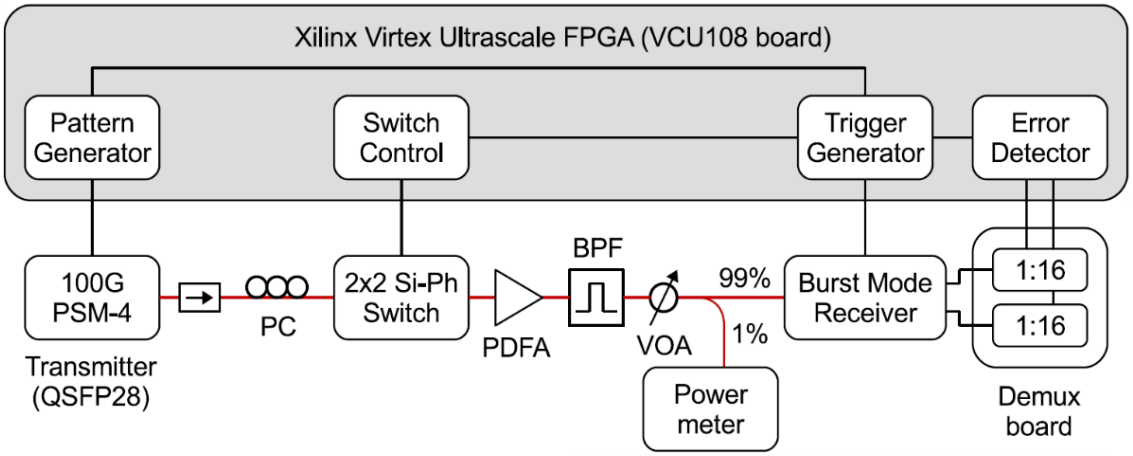


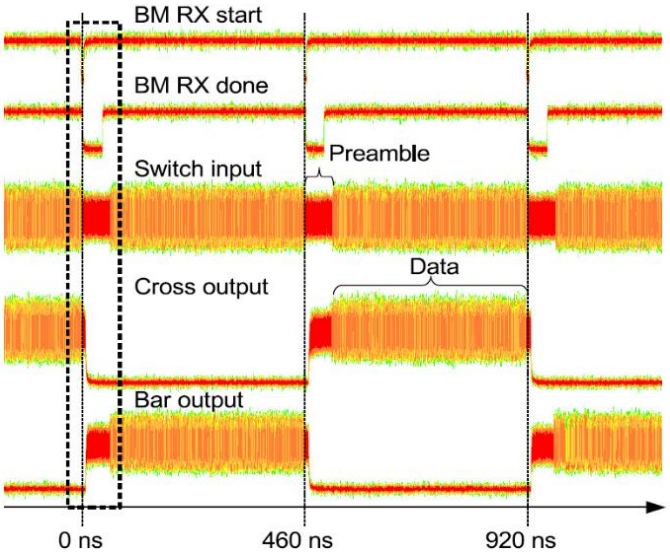
[N. Dupuis, JLT, preprint online]

- Needs optimization & transfer to commercial process
- Potential for major impact beyond optical switching

Fast OCS Network Control & Data Planes

- **FPGA generates data, sends to TX**
  - 1010 preamble + 1 or 2kB payload (PRBS15)
- **FPGA tells switch to set state**
- **FPGA wakes up burst-mode RX**
  - lock time includes threshold detection and phase locking (31ns at 25 Gb/s)
- **FPGA performs error detection:**
  (1) Gated error detection
  (2) Frame-sync'd pattern checkers





Payload size (B)	2048	1024
Data rate (Gb/s)	20.6	20.6
Cycle time (ns)	858	460
Packet length (ns)	797.5	400.3
RX lock time (ns)	41.5	41.5
Total switch time (ns)	60	60
Meas'd payload BER	10^{-12}	10^{-12}
Duty cycle (%)	93	87

Additional Work Needed:

- **Improve to < 20ns with:**
  - 56G, 7ns burst-mode RXs
  - phase-caching
- **Validate using:**
  - amplified links & WDM
  - scaled port counts
- **Add:**
  - network interfaces
  - scheduling & arbitration

[A. Forencich, "A Dynamically-Reconfigurable Burst-Mode Link …," *JLT 2019, under review*]

ONRAMPS Summary

- ONRAMPS technology can provide significant performance improvements to datacenters & HPC systems, with potential to impact other markets as well
- ONRAMPS funding has enabled the development of a manufacturing platform and ecosystem for a low-cost fully packaged nanosecond photonic switch
  - Realized IC with all the critical photonic & electronic components with record 1×2 and 2×2 loss of **0.8 dB** and **1.3 dB**, extinction ratios of **28 dB** and **38 dB**, and switching times of **6 ns**.
  - Established volume-compatible optical and electrical packaging procedure capable of dual-ribbon attach to flip-chip module with **1.5 dB** (typical) coupling loss per facet.
  - Established cost models for photonic switch module with projections of a **few ¢/Gb/s**.
  - Built FPGA-based control plane interfacing with digital switch IC and burst-mode transceiver demonstrating **60 ns** end-to-end switching at 25 Gb/s.
  - Established modeling environment and ran simulations showing that a single ONRAMPS switch plane enables **> 95 %** throughput for a wide range of traffic patterns with **< 1 µs** latency.
- 8×8 fiber-pigtailed SNB switch module using single photonic + electronic IC with digital programming interfaces and nanosecond reconfigurability is in process

The information, data, or work presented herein was funded in part by the Advanced Research Projects Agency-Energy (ARPA-E), U.S. Department of Energy, under Award Number DE-AR0000844. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.