CHANGING WHAT'S POSSIBLE

ENLITENED Annual Program Review

LEED – A Lightwave Energy Efficient Data Center

October 23, 2018

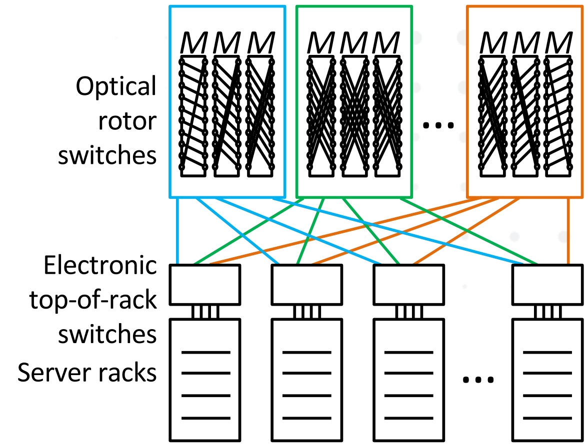
Data Center Network Objectives

- What you want in any data center network
  - Non-blocking, all-to-all connectivity, full line rate
- Problem for existing data centers
  - Too expensive and power hungry
- LEED Solution
  - Complete data center network redesign
  - Co-optimize network, optical switch, and interconnect
- Result
  - Larger (cost-comparable) bandwidth leading to commensurate server energy-utilization improvement (ENLITENED Metric 1.1)
  - Dramatically simplified circuit-switched control plane
    - Deterministic - no schedule
    - Practical and scalable

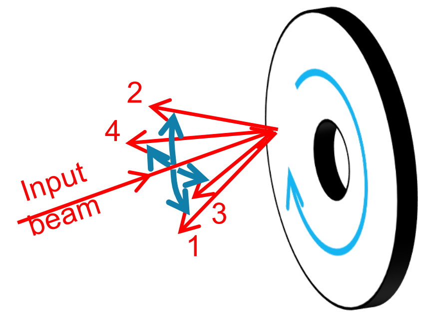
Top-Level LEED Objectives

- A robust, scalable, energy-efficient data center (ENLITENED Metric 1.1)
- Co-optimized across:
  - Network Architecture
    - Parallel optically-switched network
    - Cost effective and fault tolerant
  - Optical Switch
    - Decouples switching from routing
    - Based on “pinwheel” switch
  - Commercially viable enhanced link-margin interconnects
    - Burst-mode APD receiver
    - WDM modulator array
    - Broadband mux/demux

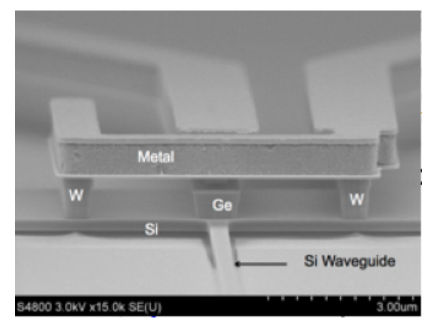
Parallel Optical Rotor Network



“Pinwheel” Rotor Switch



APD for burst-mode Rx

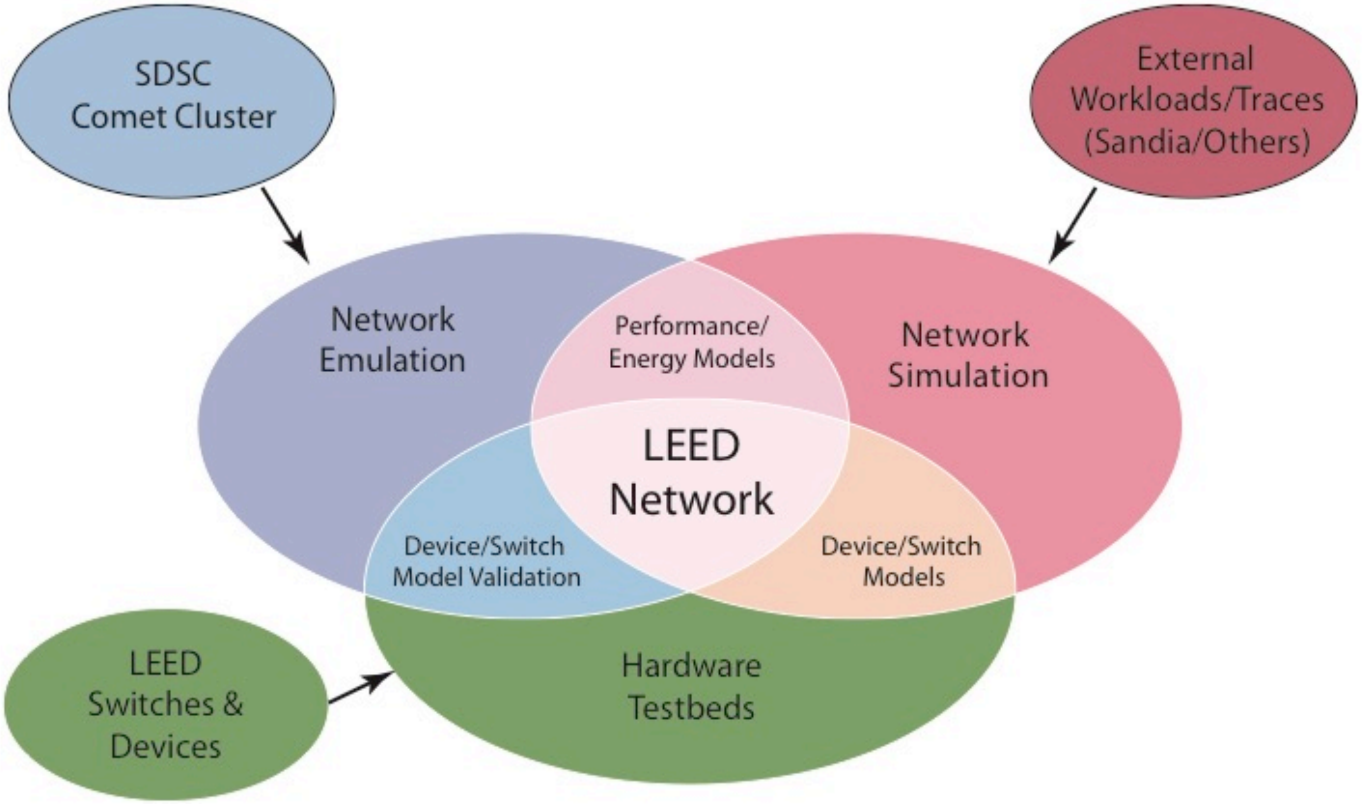


arpa·e CHANGING WHAT'S POSSIBLE | LEED | Axalume | inFocus NETWORKS | Sandia National Laboratories | UCSD

LEED Team

- **Systems**
  - Max Mellette, George Papen
    George Porter, Alex Snoeren (UCSD)
- ***Optical Switch***
  - Ilya Agurok, Joseph Ford, Max Mellette (UCSD)
- ***Interconnects***
  - *UCSD*
    Shaya Fainman, Shayan Moookherjea
  - *Axalume*
    Ashok Krishnamoorthy, Saman Saeedi
  - *Sandia National Labs*
    Michael Gehl, Christopher T. DeRose, Paul S. Davids, Douglas C. Trotter, Andrew L. Starbuck Christina M. Dallo, Dana Hood, Andrew Pomerene and Tony Lentine

Project-Wide Objectives



- Simulation (Network → Metric 1.1; Interconnect→ Metric 1.2/1.3)
- Emulation (Rotornet switch using programmable packet switch)
- Validation (Hardware testbed/emulation → all metrics)

Anticipated Outcomes

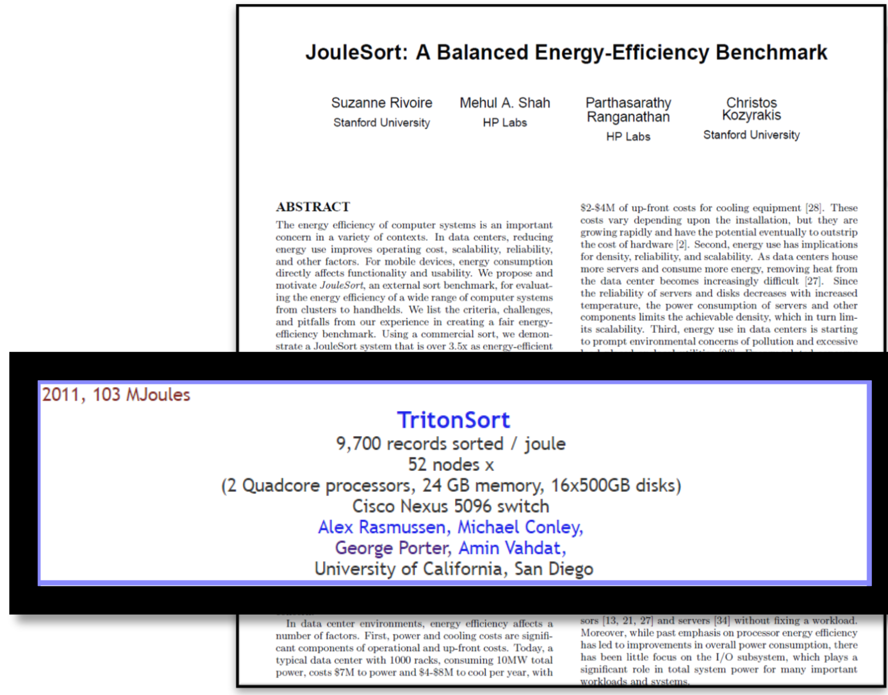
- Improved server energy efficiency with cost-comparable network
  - Directly addresses ENLITENED Metric 1.1
- Highly scalable
  - Deterministic switching & routing
- Cost effective
  - No OEO in core
  - Low cost per switched bit
- Robust
  - Parallel Network Architecture

Measured Outcomes Leverage San Diego Supercomputer Center Comet cluster

- 1,944 nodes
- 24 cores / node @ 2.5 GHz
- InfiniBand Network 40 Gb/s / node
- Full bisection “pods” of 72 servers
- 4:1 oversubscription between pods

Energy Efficiency

- LEED members set standard for energy-efficient computing
  - Relied on JouleSort measurement methodology
  - World-record set in records sorted/Joule
- Developing best practices for measuring ENLITENED Metric 1.1
  - Direct measure power of entire system (servers + network)
  - Extrapolate to large-scale clusters
  - Must deal w/power variability


JouleSort: A Balanced Energy-Efficiency Benchmark

Suzanne Rivoire, Stanford University
Mehul A. Shah, HP Labs
Parthasarathy Ranganathan, HP Labs
Christos Kozyrakis, Stanford University

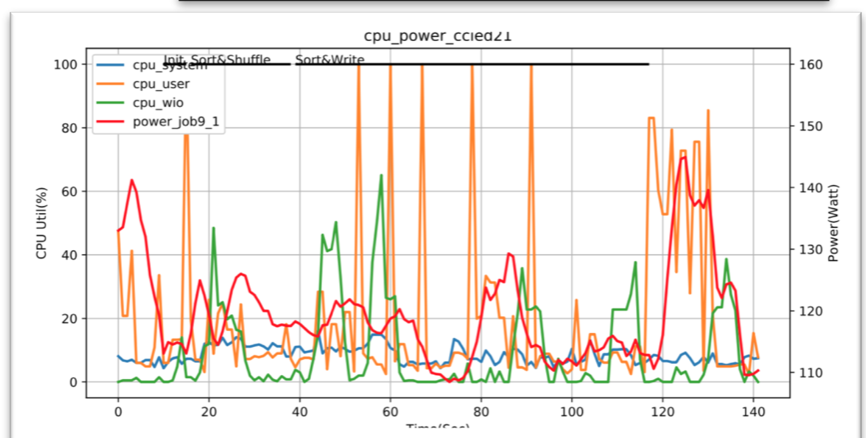
ABSTRACT

The energy efficiency of computer systems is an important concern in a variety of contexts. In data centers, reducing energy use improves operating cost, scalability, reliability, and other factors. For mobile devices, energy consumption directly affects functionality and usability. We propose and motivate *JouleSort*, an external sort benchmark, for evaluating the energy efficiency of a wide range of computer systems from clusters to handhelds. We list the criteria, challenges, and pitfalls from our experience in creating a fair energy-efficiency benchmark. Using a commercial sort, we demonstrate a JouleSort system that is over 3.5x as energy-efficient

In data center environments, energy efficiency affects a number of factors. First, power and cooling costs are significant components of operational and up-front costs. Today, a typical data center with 1000 racks, consuming 10MW total power, costs $7M to power and $4-$8M to cool per year, with

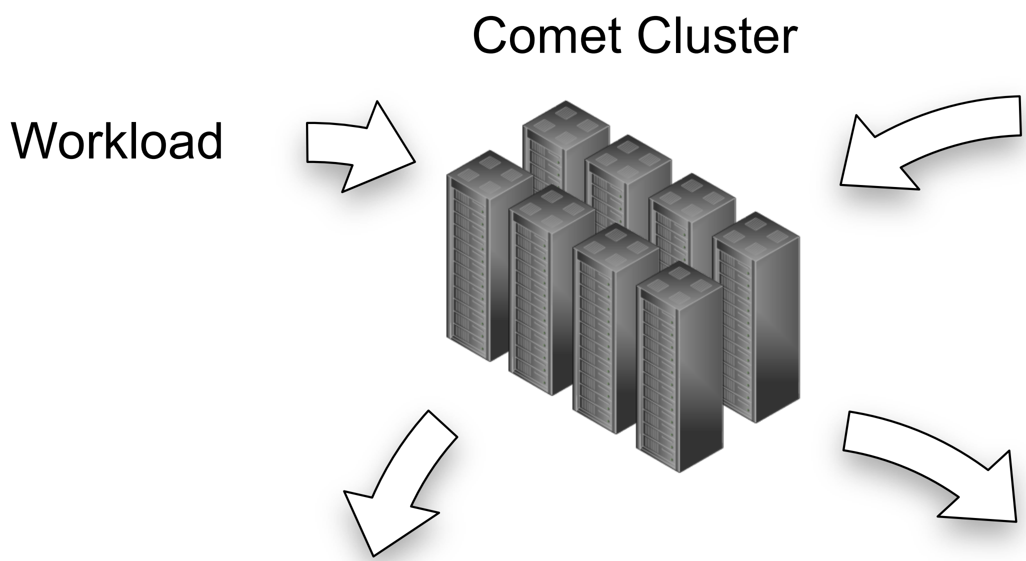
$2-$4M of up-front costs for cooling equipment [28]. These costs vary depending upon the installation, but they are growing rapidly and have the potential eventually to outstrip the cost of hardware [2]. Second, energy use has implications for density, reliability, and scalability. As data centers house more servers and consume more energy, removing heat from the data center becomes increasingly difficult [27]. Since the reliability of servers and disks decreases with increased temperature, the power consumption of servers and other components limits the achievable density, which in turn limits its scalability. Third, energy use in data centers is starting to prompt environmental concerns of pollution and excessive

sors [13, 21, 27] and servers [34] without fixing a workload. Moreover, while past emphasis on processor energy efficiency has led to improvements in overall power consumption, there has been little focus on the I/O subsystem, which plays a significant role in total system power for many important workloads and systems.



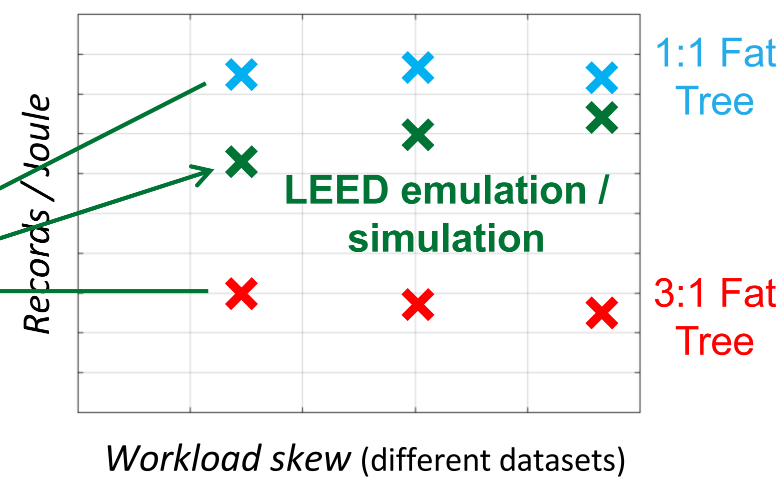
Validation Framework Using Comet

1st Year Accomplishments



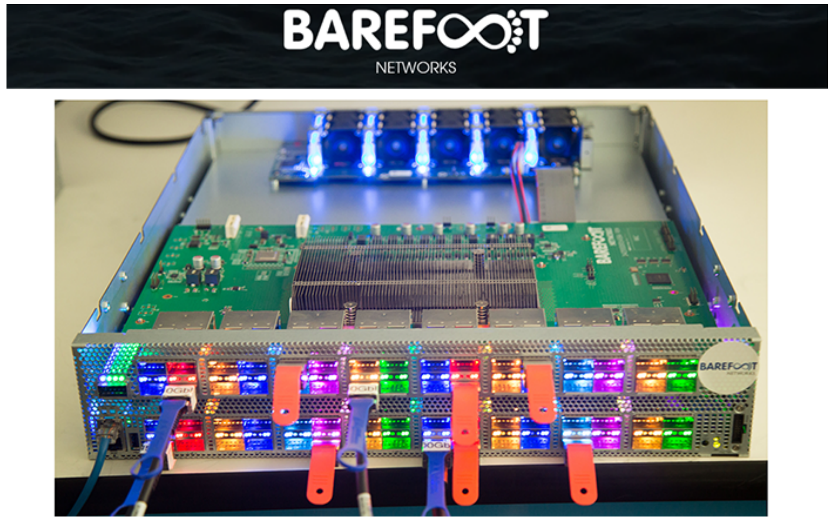
1. Impose bandwidth restrictions:

2. Measure energy efficiency:



- 3. Application execution dependency graph
  - Feeds into blocked-time network model
- Use accurate power meter energy measurements to calibrate energy model (supports LEED achieving Metric 1.1)
  - Requires emulation of optical switch

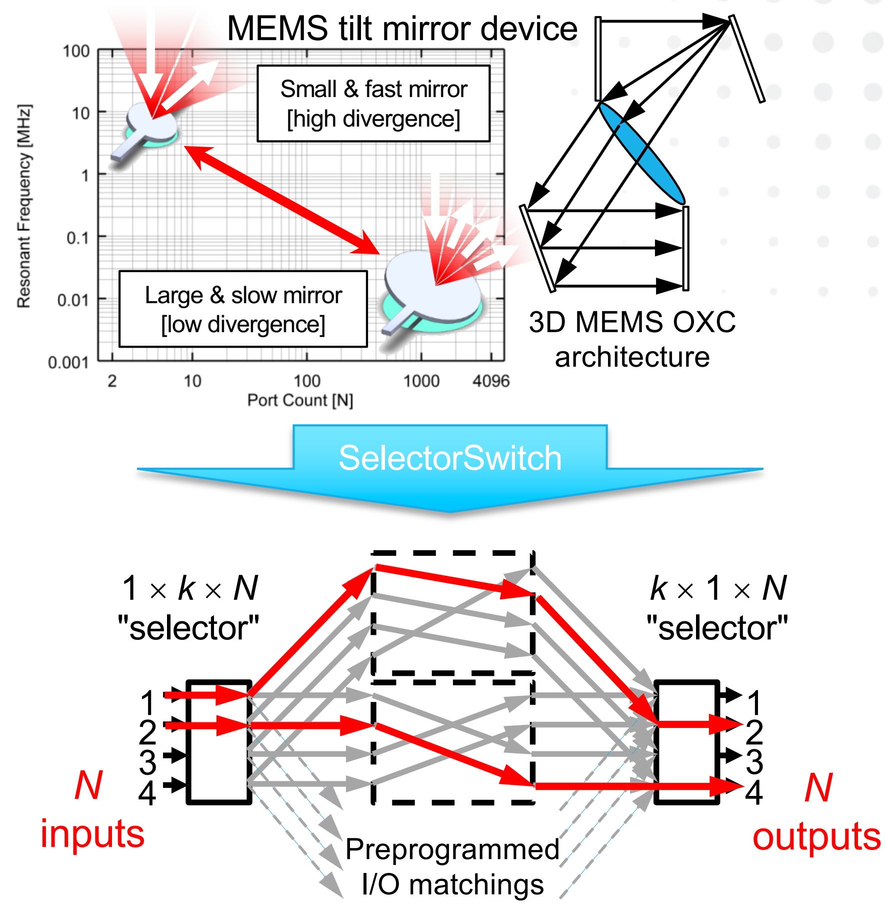
Emulation of Rotor Switch



- Barefoot Tofino 6.4 Tb/s programmable P4 switch
  - 8 hardware ToR packet switches
    - LEED network protocol supported via added P4 rules
  - 4 hardware emulated LEED switches
    - μs circuit switch implemented via P4 rule set
- Accurate emulation of rotor switch essential for scaled-out experimental measurements of energy efficiency (Metric 1.1)

Switch Objectives - Background

- **Initial Goal:** More ports & faster OXC w/lower cost & same loss
  - Focus on free-space optics (FSO) solutions
- **First Innovation - Selector Switch**
  - Decouple *Switching* (# configurations) from *Routing* (I/O pairs for each configuration)
- **Selector Switch**
  - Image-domain FSO parallel gang switching
  - Small set of pre-programmed I/O matchings



- **LEED Objectives for Selector Switch**

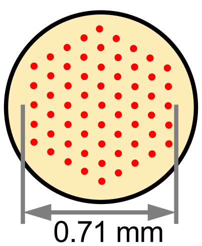
  1) Upgrade from 7dB to <3.5 dB/pass loss & 150μs to <50 μs; insert in testbed

  2) Explore & "de-risk" technology path to practical/commercial high-port-count switches

Choice of Switch Actuator

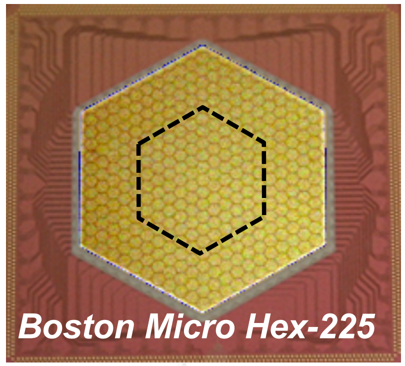
Initial proposal: transition from single MEMS to faster MEMS device array

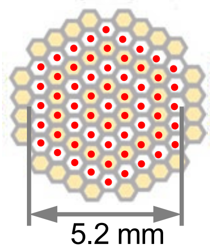
Existing 61-port prototype



Single 800 µm mirror, ± 3° tilt, 150µsec

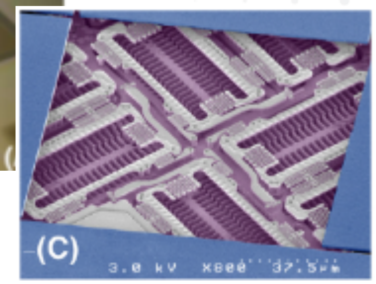
Original LEED upgrade

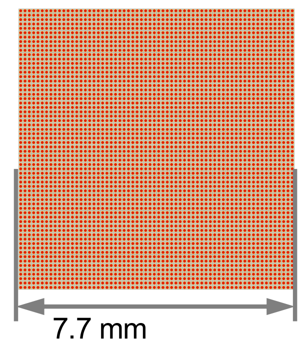




61/225 650µm mirrors ± 0.46°, 20µsec

Scale-up to commercial product





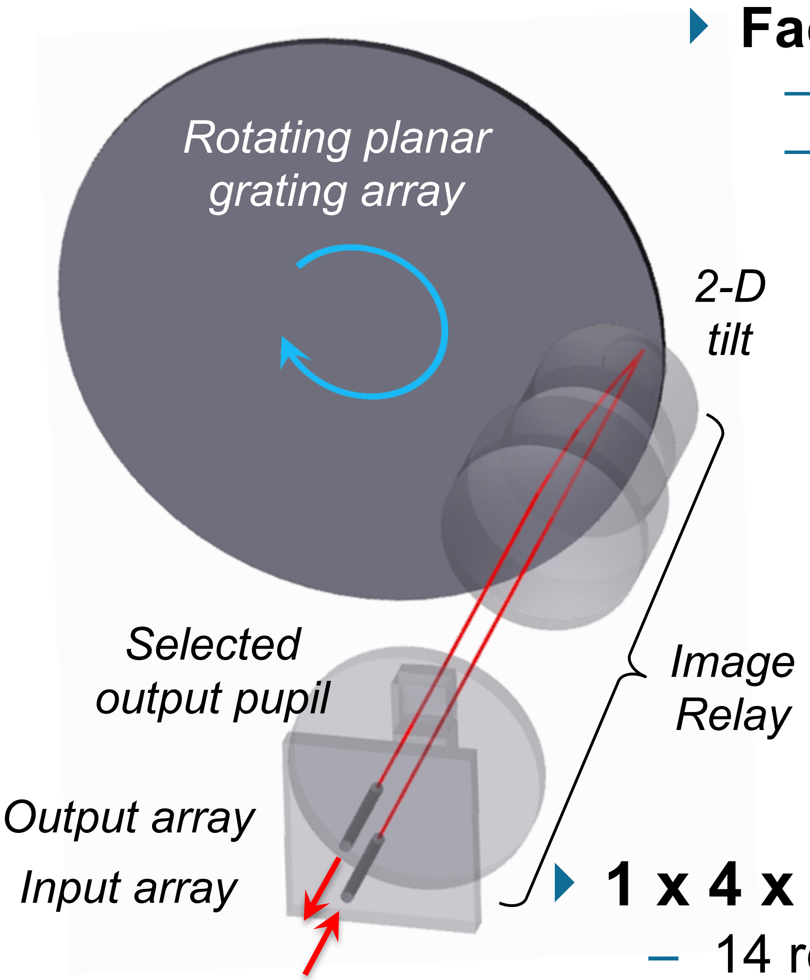
4096 120µm mirrors ± 4°, 20µsec

Bell Labs 2007 research paper (not product)

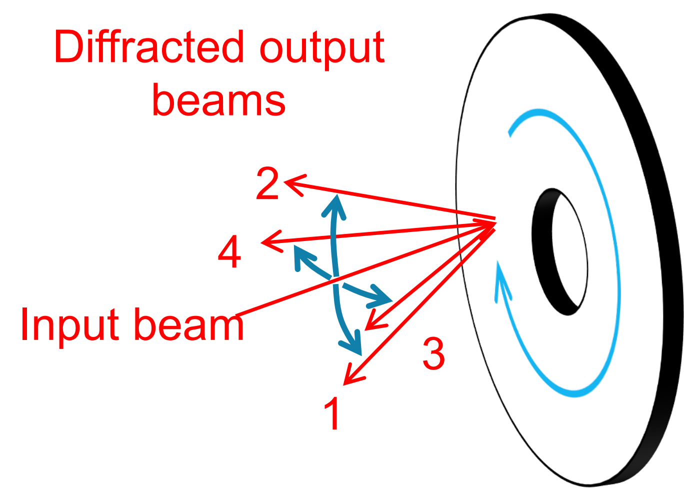
- Discrete MEMS: fast asynchronous switching in large arrays, but...
  - *Needs huge NRE ⟶ risk to future commercialization*
- ***Second innovation:*** *a rotating faceted beam-deflector actuator*
  - Rotor net supports **synchronous & sequential switching**

arpa·e CHANGING WHAT'S POSSIBLE LEED Axalume inFocus NETWORKS Sandia National Laboratories UCSD

"Pinwheel" Rotor Switch Actuator

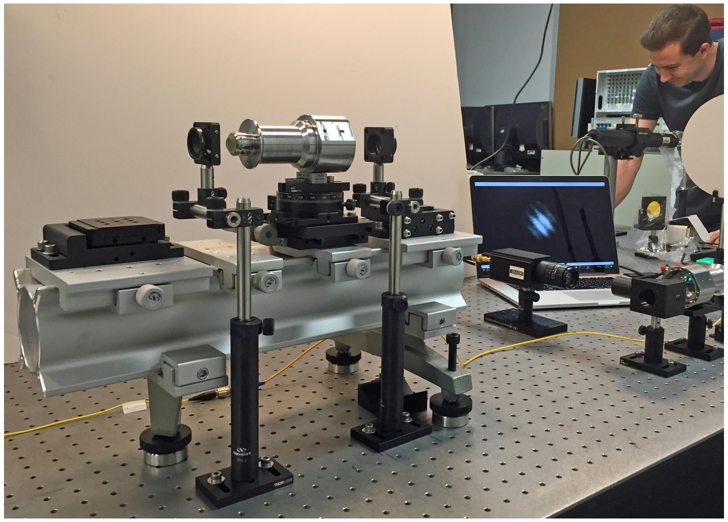


- **Faceted diffraction grating beamsteering**
  - **Decouples switch angle from switching speed**
  - Conformal mapping of grating fixes tilt angle over rotation



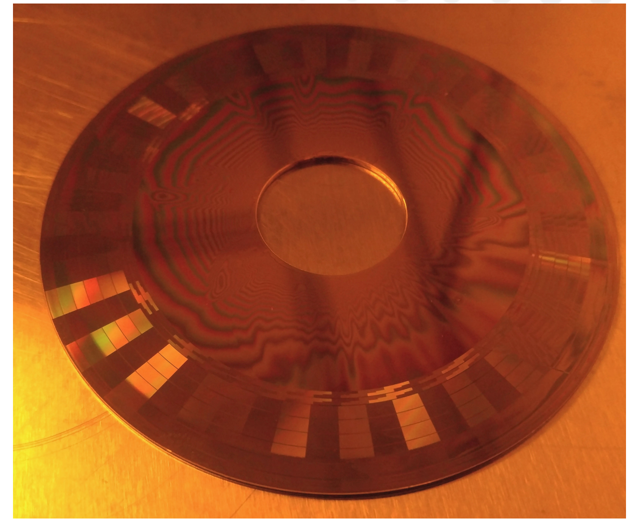
- **1 x 4 x 61 port layout**
  - 14 repeats of 4 tilts → 90% duty cycle, 15 μs switching
  - Larger tilt angle → 7x shorter track than FSO MEMs
  - >100 nm spectral range (relaxed interconnect specs.)

Prototype Rotor Switch Status



Fiber reduced pitch I/O arrays, Optomechanics & alignment fixture

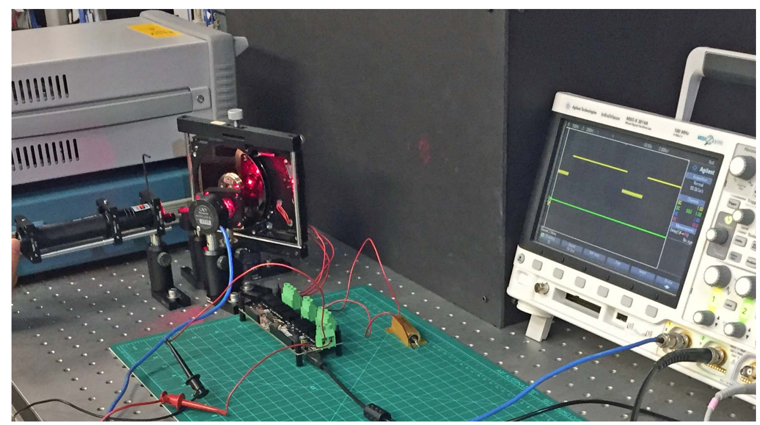
Custom optics cut and coated, currently being assembled



Custom grating printed on HD disk & gold coated for testing

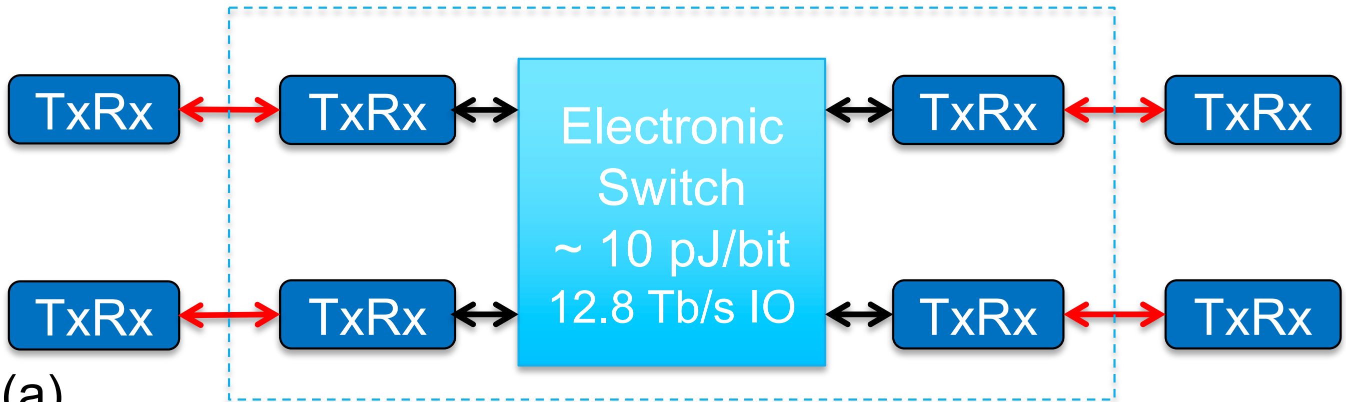
- All custom components completed or in final fab.
  - Switch alignment process has been developed.
- Spindle control electronics has been tested
  - working on multi-spindle synchronization
- Switch integration will begin this month.

Grating spindle and control board



Interconnect/Switch Objectives

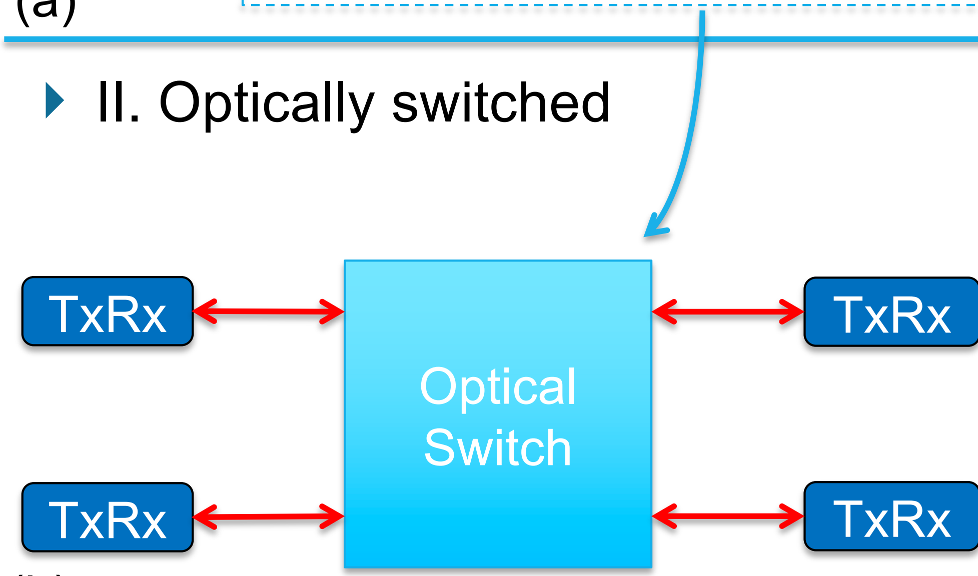
- I. Optically-interconnected, electrical switching



(a)

- Switch energy is relatively high
- Link metrics
  - 2 pJ/bit
- BW density
  - 1 Tb/s/cm

- II. Optically switched

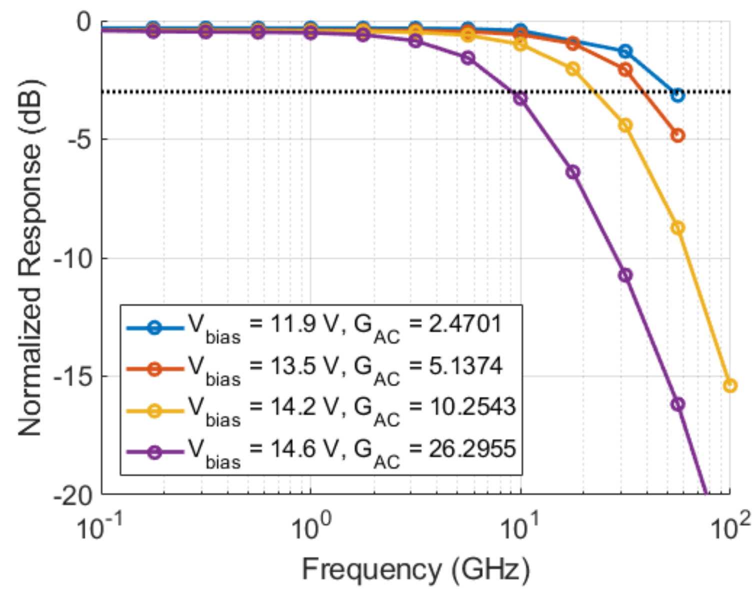


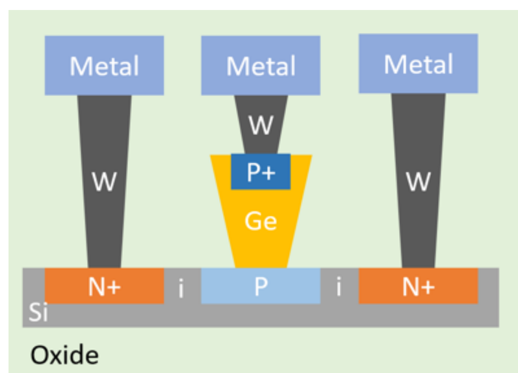
(b)

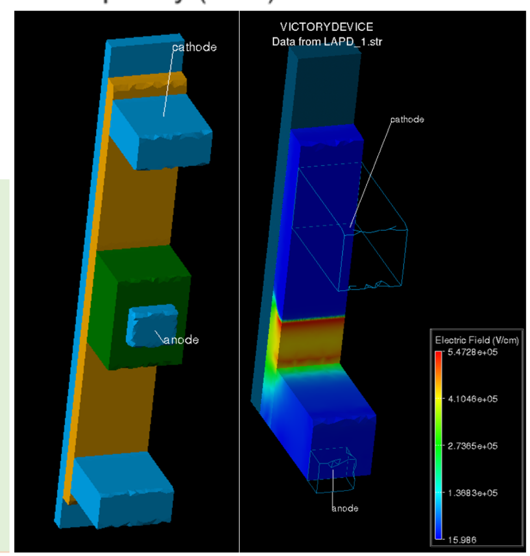
- Switch energy is low
- Switch loss is managed w/o amplifiers
  - Link is optimized for margin
  - **Link metrics(1.2)**
    - **1 pJ/bit excluding laser power**
    - **+ 1 pJ/bit laser x excess switch loss**
    - **= 2 pJ/bit for a lossless switch**
  - **Link metric vs ~14 pJ/bit Case I**
- Scales > 100 Tb/s

Avalanche Photodiode Rx: Design

- Uses germanium absorption with silicon multiplication regions
  - Two classes of designs with vertical and lateral multiplication regions
  - Fully compatible with existing Sandia silicon photonics process
  - Simulations performed using Silvaco to optimize dopant and dimension splits
  - One lateral design designed for integration with burst mode Rx.
    - p-i-n versions as well

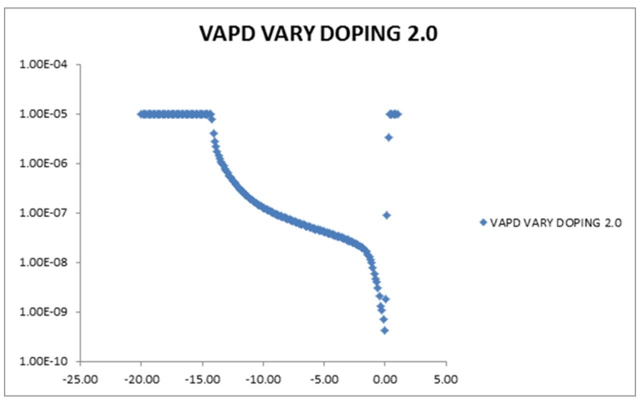




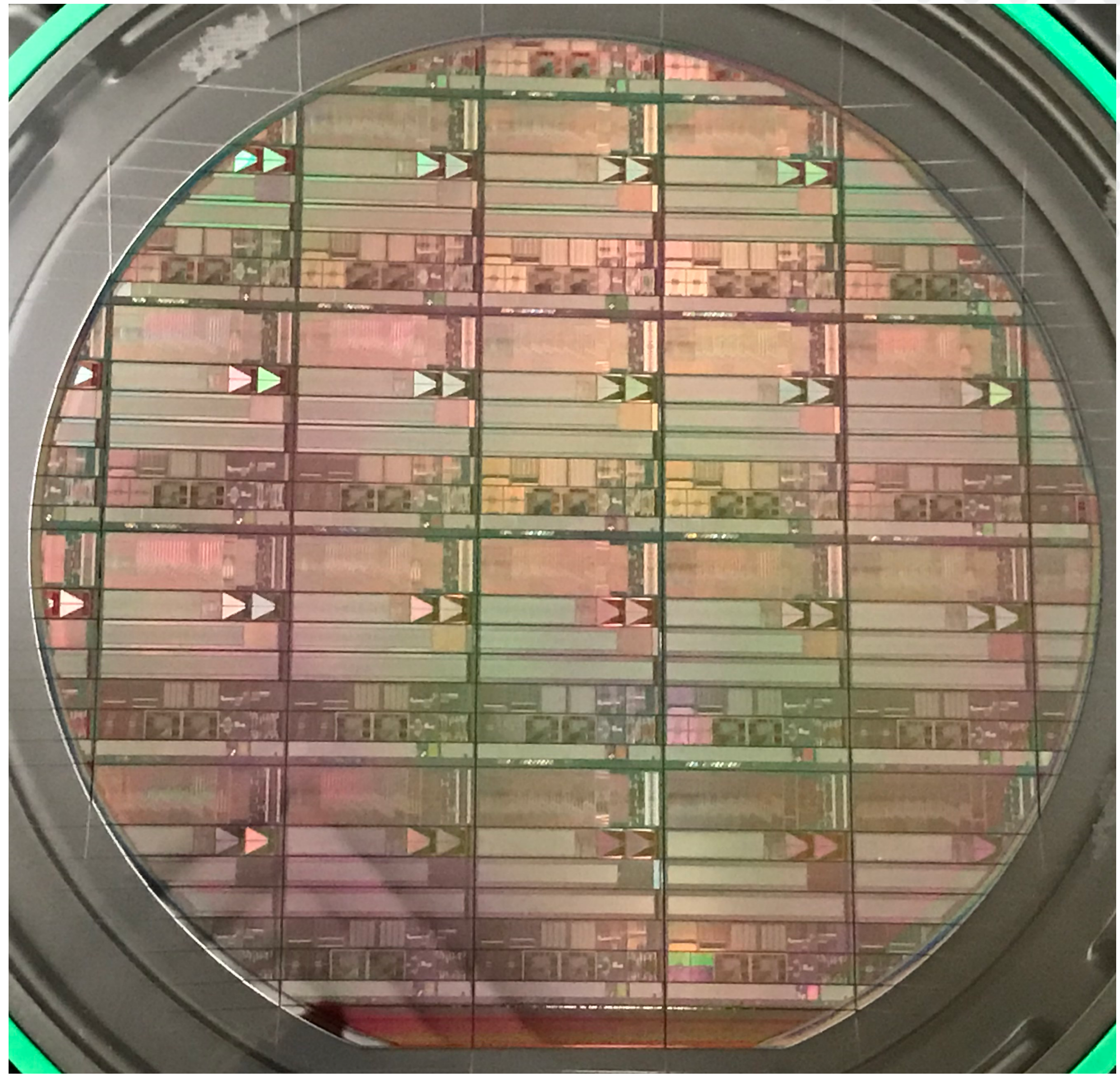


Avalanche Photodiode Rx: Fab

- All LEED silicon photonics fabrication lots completed before fab-conversion deadline
- DC wafer testing showed avalanche behavior as expected



- Detailed optical testing to begin shortly

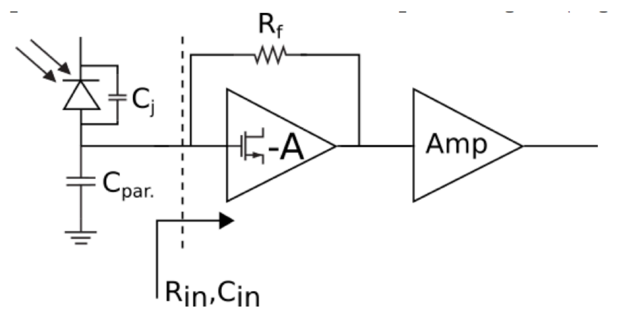


Burst-Mode Optical Receivers

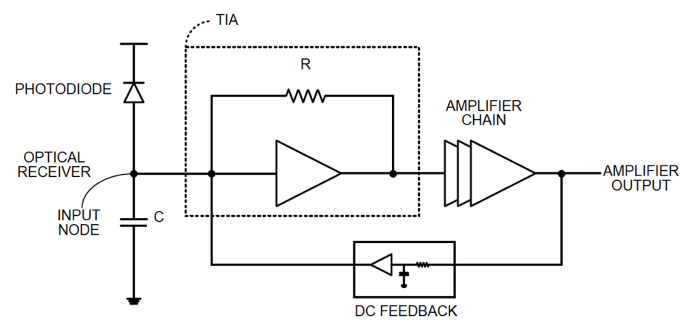
Burst-mode receivers are required for optical switching

- The key to achieve this is to reduce the optical receiver’s acquisition time

Conventional receivers cannot achieve this

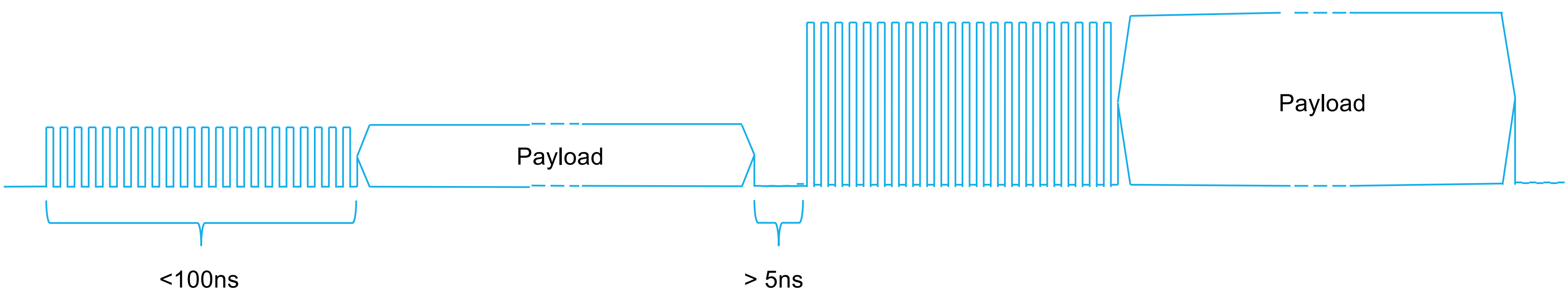


- TIA front-ends are used to reduce the input impedance and extend the bandwidth
- TIA’s are generally sensitive to input bias. This necessitates slow feedback
- The feedback loop introduces stability criteria that limits acquisition speed
- Compatibility with flip-chip & wire-bond configuration further complicates the design
- Axalume’s BM Rx circuit uniquely achieves fast clock-recovery, high-speed, low-area, and low power in a “workhorse” CMOS technology node
- Tapeout completed Year 1 milestone

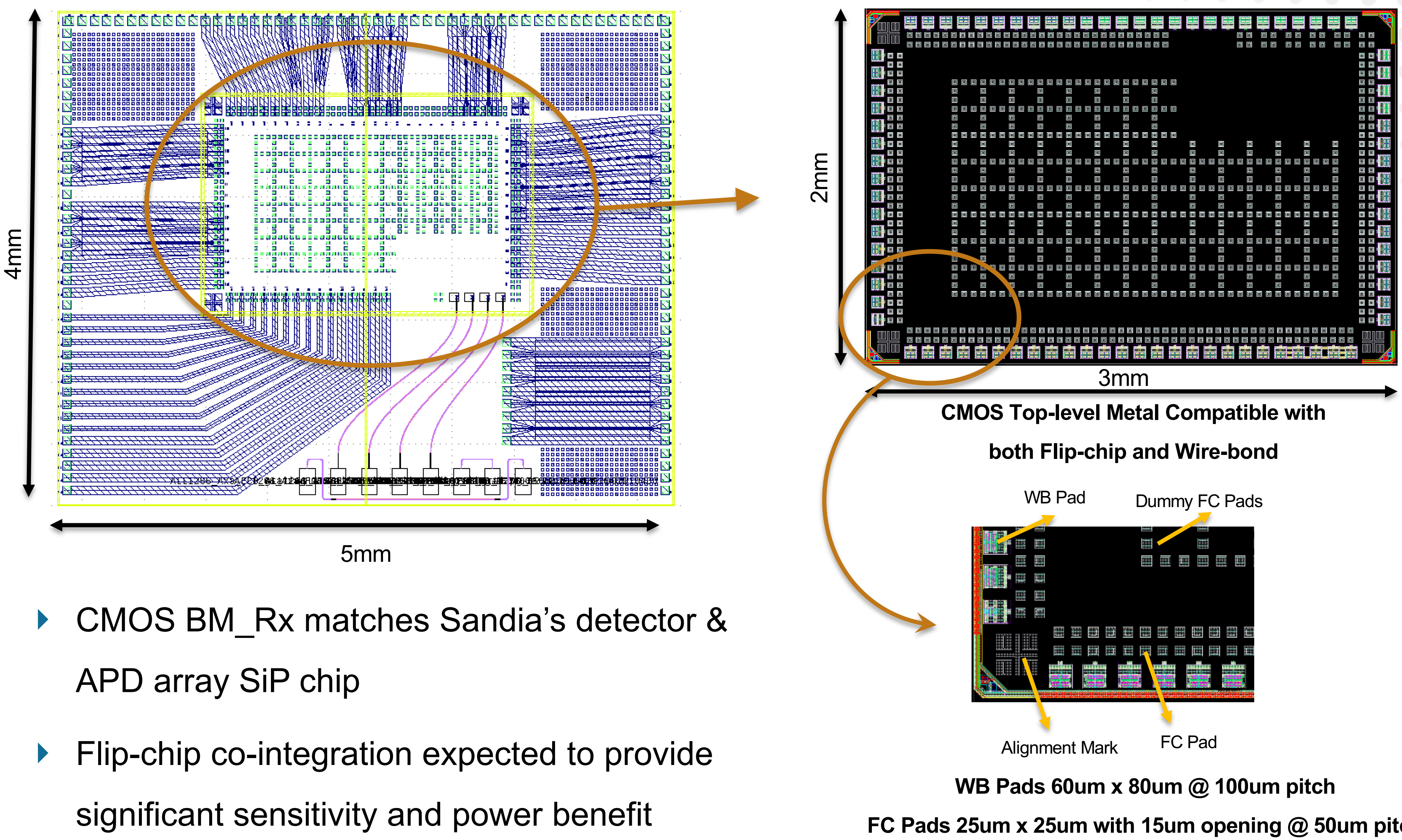


Burst-Mode Optical Rx Specification

- Program technical requirements:
  - Total acquisition time of less than 100ns
    - Optical power dynamic range
      - Dynamic range spec suitable for optical switching: 7dB
      - < 25ns for DC acquisition
    - Burst-mode clock recovery
      - No phase information, 100ppm frequency offset
      - < 75ns for full clock recovery time
- Pre-amble prior to payload with sequence of 0s and 1s.



Top-level Layout



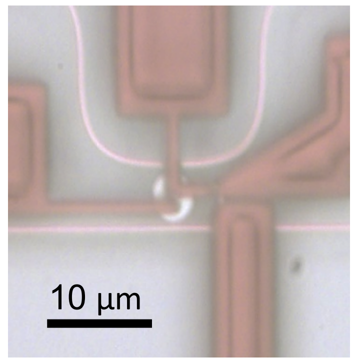
- CMOS BM_Rx matches Sandia's detector & APD array SiP chip
- Flip-chip co-integration expected to provide significant sensitivity and power benefit

Resonant Modulators: Fab

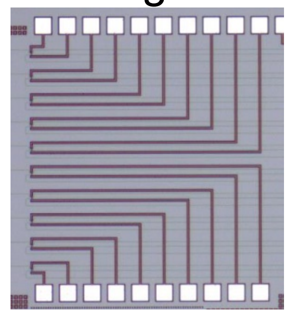
1st Year Accomplishments

- Sandia-fabricated LEED silicon photonics fabrication lots received for measurement.
- Optical passive testing showed expected baseline behavior on selected test sites (FWHM ~ 22GHz).
- Detailed optical testing to begin shortly, C-band & O-band.
- Design and fab iterations underway.
- Parallel effort on bias control and tune-up controller.

Yr 1 test structures (C-band)

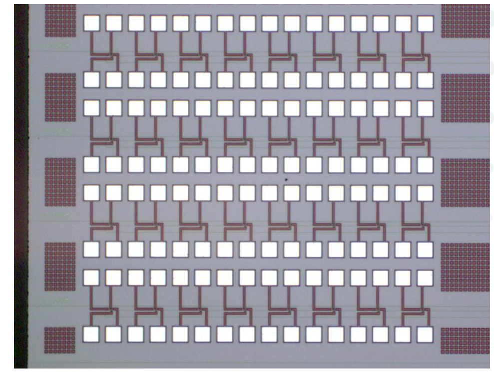


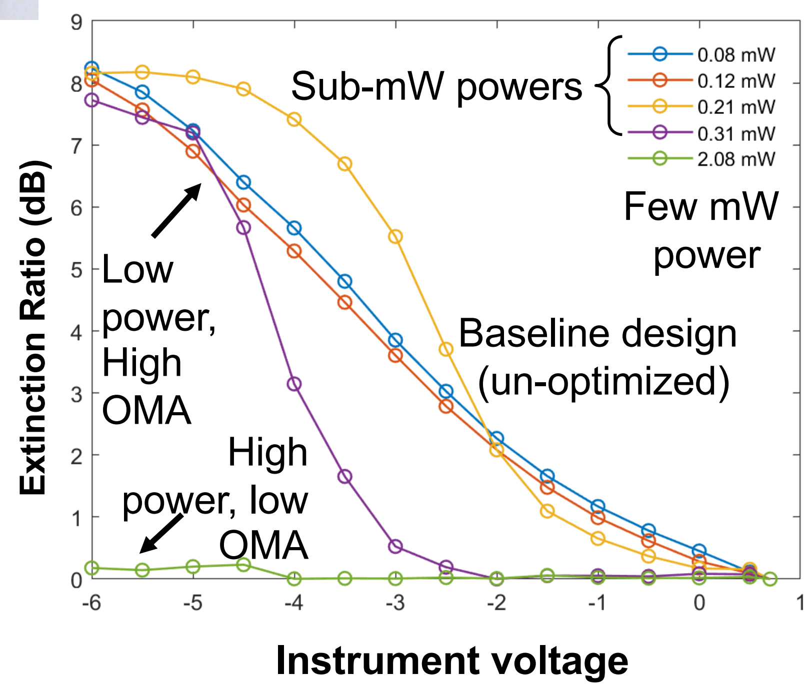
singles



GSG optimized separately.

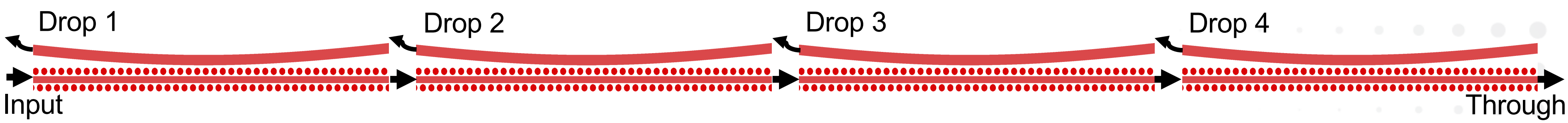
8 mods per bus wg





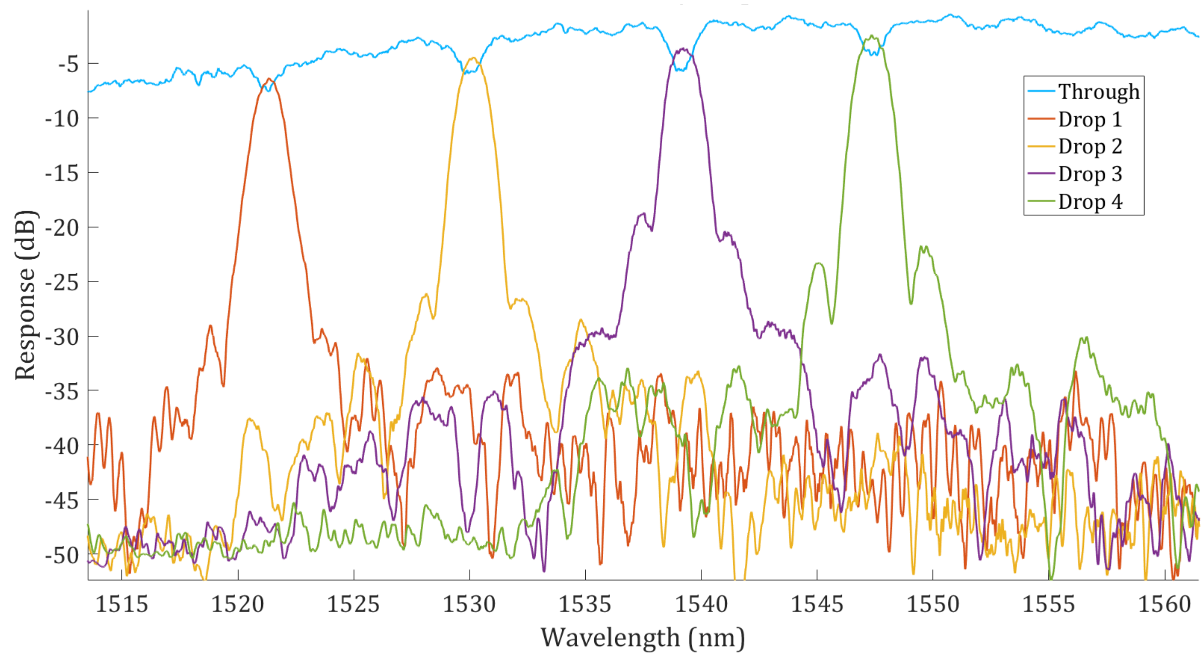
arpa-e CHANGING WHAT'S POSSIBLE LEED Axalume inFocus NETWORKS Sandia National Laboratories UCSD

Broadband MUX/DEMUX



- **Broadband Wavelength Selective Coupler**
  - Reduce power consumption by reducing loss while maintaining other favorable attributes
    - High extinction ratio, broad bandwidth, small footprint, fabrication tolerant, and low crosstalk

- **Measured Performance**
  - Footprint per channel: <1000 μm^2
  - Scalable to 40 channels with footprint 1 mm^2
  - Channel width: 250 GHz
  - Channel-to-channel crosstalk: < 15dB
  - Loss on drop port: 2 dB



LEED has Fostered two Start-ups

- ***Axalume - incorporated March 01, 2017***
  - Incubating at Evonexus
  - Multiple patents filed
  - Rx chipset taped out August 2018
- ***inFocus Networks - incorporated March 26, 2018***
  - Focus on commercialization of switch and architecture/protocol
  - Core IP filed (three patents)
  - Applied for SBIR Phase I funding

LEED Outreach

Invited Talks and Plenaries

- W. M. Mellette, J. E. Ford, and G. Porter, "Partially Configurable Optical Switching for Data Center Networks," IEEE Photonics Conference, 2017
- A. V. Krishnamoorthy, International Solid State Circuits Conference (ISSCC) 2018, Paper 16.1, San Francisco, Feb 2018
- A. V. Krishnamoorthy, Co-packaged optical interconnects for computing & switching systems Optical Fiber Communications Workshop on Optical Co-packaging, OFC 2018
- A. V. Krishnamoorthy Low-Power Co-Integrated Electronics-Photonics for Switching and Computing Systems OSA Topical Meeting on Photonics in Computing and Switching, Limassol, Cyprus, September2018 (Plenary)
- W. M. Mellette, A. C. Snoeren, and G. Porter, "Toward Optical Switching in the Data Center," IEEE International Conference on High Performance Switching and Routing 2018.

Accepted Invited Talks

- W. M. Mellette, "A Practical Approach to Optical Switching in Data Centers," OFC 2019.
- G. Papen, Workshop on “Opportunities and Challenges for Optical Switching in the Data Center”, OFC 2019.

Collaborations with Industry

- Y. Birk, W. M. Mellette, and E. Zahavi, "Switch Radix Reduction and Support for Concurrent Bidirectional Traffic in RotorNets," Photonics and Switching Conference, 2018.

Supplemental Support

- **California Energy Commission**
  - Additional $196k in direct support from initial budget
  - Supporting additional wafer runs, switch prototyping, and TT&O
- **San Diego Supercomputer Center**
  - NSF funded site has discretionary compute cycles
  - PI has given LEED compute cycles
  - Infrastructure support for power measurements on Comet

Surprises and Lessons Learned

- **Surprises**
  - Combination of pinwheel/conformal gratings provides a large design space for optical switch design
    - Original motivation — practical commercialization path
- **Lessons Learned and Future Challenges**
  - Power/energy measurements require careful calibration
    - Our focus is on an accurate, scalable energy measurement
  - The control plane is hard — even without a schedule
    - Must make packets and circuits “play nice” together