Multi-wavelength Optical Transceivers Integrated On Node
PI: Dan Kuchta, IBM and Finisar
Outline

- Overview of MOTION Project and Technology
- SAFE (Simplified Analog Front End) ICs
- Flip-chip VCSELs and Photodiodes
- Transceiver Module Development
- Initial measurement results
- Network Modelling & Simulation results
- Tech-to-Market
- MOTION Phase 2
The Trends for Higher Bandwidth

Now

Pluggable Optics
Switch ASIC
First-level package
- Bandwidth limited by pin density at package / board interface
- Large energy costs for driving > 10 cm transmission lines

Soon

Co-Packaged Optics
- Bandwidth limited by pin density at chip / package interface
- Some energy limited for few cm long transmission lines but dominated by E/O and O/E

Future

Optical Switch
- Bandwidth limited by spectral efficiency – not pin counts
- No E/O or O/E. Energy spent on steering pipes rather than processing or transmitting bits

Extra Pain with No Gain

▲ Avoid distortion, power, & cost of ASIC-interfacing electrical links
▲ Move beyond chip & module pin-count limits
▲ Agnostic to upgrades in signaling rates and formats

IBM MOTION Program
IBM ONRAMPS Program
Why are we interested in Co-Packaging?

- Primarily for increasing BW from ASICs
  - Large ASICs are package pin constrained
  - Co-Packaging permits I/O from both sides of the package
- Reduction in Power Consumption
  - Juxtaposed die do not require high power SERDES
  - Lower power I/O cells also use less Si Area
- Reduction in Cost
  - Stripped down optical packages and reduced function ICs should cost less
  - Reduced ASIC area will have higher yield
- Expansion of ASIC performance
  - Instead of reducing ASIC Area and Power, other choice is to maintain size and add more functionality to the chip up to the original Power constraint
  - i.e. Additional switch ports or Additional Memory Hubs
Co-Packaging on Organic Laminates: MOTION Phase 1

- **ARPA-E (U.S. Department of Energy) sponsored project, Phase 1: 2 years**
  - IBM and Finisar Inc.

- **Target specifications**
  - 56GBd NRZ; BER tested to <1E-12 pre-FEC
  - 0°C to 70°C Case
  - 6dB (electrical) link budget (XSR-like)
  - 2 dB optical link margin (30m w/connectors)
  - < 4 pJ/bit (3.2W, 16 channels)
  - W:13mm x D:13mm x H:4mm
  - Package can withstand reflow onto ASIC 1st level laminate

**MOTION**: Multi-wavelength Optical Transceivers Integrated on Node
**MOTION Transceiver Package Overview**

- **Chip-Scale Optical Package (CSOP)**
- **MOTION Vision: Multi-Component Carrier with CSOP for high speed I/O**
  - Cu Heat Spreader
  - Glass Carrier
  - SAFE ICs, VCSELs, PDs
  - Keel
  - Final Assembly with lens and clip attached
  - Fully Assembled with fiber cable and strain relief
  - 4mm total height
  - 90mm
Simplified Analog Front-End (SAFE2) ICs

- SAFE2: Builds on previous designs
- 16 channels × 56 Gb/s NRZ = 896 Gb/s/IC
- No retiming / CDR
- 4 pJ/bit power consumption
- Fully DC-coupled: passes 64b/66b & PRBS31
- 55 nm SiGe BiCMOS
- CMOS-compatible electrical signal levels
- Built-in pattern generators and error detectors
SAFE2 TX Channel Architecture

- Differential data input with 85Ω termination and ESD protection
- Level-Shifting Attenuator (LSATT)
  - Compatible with CMOS signal levels: 0.35 to 0.65 Vcm, 0.2 to 1.0 Vppd
  - Provides consistent performance across corners and input common modes
- Continuous Time Linear Equalizer (CTLE)
  - Provides 0 to 6 dB peaking @ 28 GHz equalizing channel from ASIC to TX
- Limiting Amplifier (LA)
- Multiplexer (AMP)
  - Switches between input data, PRBS pattern generator, and fixed 0/1
- VCSEL Driver with 3-tap FFE
  - LC delay line for power-efficient delay
  - 0.75 UI delay per tap
  - 2 drivers for VCSEL sparing (only one powered at a time)
- Improved performance over SAFE1
- Monitor photodiode current amplifier
  - Analog output to measure VCSEL optical power
SAFE2 TX Summary

- Initial hardware undergoing lab test
  - Pre-MOTION VCSELs without sparing
- Multiple channels measured error-free to 52Gb/s
- Full testing awaits faster MOTION VCSELs with sparing

Simulated Power Consumption

<table>
<thead>
<tr>
<th>Power supply</th>
<th>Data mode</th>
<th>PRBS mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8 V</td>
<td>820 mA</td>
<td>1.6 A</td>
</tr>
<tr>
<td>3.3 V</td>
<td>256 mA</td>
<td>256 mA</td>
</tr>
<tr>
<td>Total power</td>
<td>2.3 W</td>
<td>3.7 W</td>
</tr>
<tr>
<td>Energy efficiency</td>
<td>2.6 pJ/bit</td>
<td>4.1 pJ/bit</td>
</tr>
</tbody>
</table>
SAFE2 TX Initial Results

SAFE2 Tx IC Electrical driving signal

SAFE2 Tx IC w/pre-MOTION VCSEL (< 20GHz BW)

Measured energy efficiency@50 Gbps : ~2.4 pJ/bit
Error-free to 52G with low BW VCSEL
SAFE2 RX Channel Architecture

- **Transimpedance Amplifier (TIA)**
  - Self-referenced low-noise design converts photocurrent to voltage

- **Amplifiers (AMP)**
  - Convert unbalanced input to differential output

- **Analog feedback**
  - Low pass filtering (LPF) and opamp set low-frequency cutoff
  - Vertical slicing level control to fine-tune decision level

- **Limiting Amplifier (LA)**

- **Continuous Time Linear Equalizer (CTLE)**
  - Provides 0 to 6 dB peaking @ 28 GHz, equalizing channel from RX to ASIC

- **Output driver with 1kV HBM ESD protection**

- **PRBS Checker / Error Detector**
  - Normally powered down, only used for on-chip testing

---

**Supply Voltages**
- 1.0V
- 1.2V
- 1.8V
- 2.5V
- 3.3V

**Digital Buffers & Logic**

**Analog Biasing**

**PD Cathode Bias & DC Sig Det**

**Chip Edge**

**Unretimed Data Output**

**T-Coil & ESD**

**DAC**

**Vertical Slicing Level Control**

**AMUX**

**To Chip-level AMUX**

**Error Rate**

**Half-Rate Clock**

**Unretimed Data Output**

**Chip Edge**
SAFE2 RX Summary

- Initial hardware undergoing lab test
  - Using pre-MOTION PDs with limited BW
- Fully functional to 40Gb/s
- Full speed testing awaits assembly with MOTION PDs

### Simulated Power Consumption

<table>
<thead>
<tr>
<th>Power supply</th>
<th>Data mode</th>
<th>PRBS mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.8 V</td>
<td>480 mA</td>
<td>1.2 A</td>
</tr>
<tr>
<td>3.3 V</td>
<td>120 mA</td>
<td>130 mA</td>
</tr>
<tr>
<td>Total power</td>
<td>1.3 W</td>
<td>2.7 W</td>
</tr>
<tr>
<td>Energy efficiency</td>
<td>1.5 pJ/bit</td>
<td>3 pJ/bit</td>
</tr>
</tbody>
</table>
Glass Carrier Substrate

- Successfully designed, modeled and fabricated a glass substrate that can support at least a 16 channel transceiver with BW >30GHz
- Substrate is floor planned for production capability on 4 metal layers with embedded passives
- Overcame line impedance issues by recharacterizing the supplier’s capability of dielectric thickness.
- Extended modeling capability to include host substrates, uBGAs, ICs and optical devices.
56GBd VCSELs and Photodiodes

- Designed and fabricated flip chip 56GBd 940nm VCSELs on a production epi and Wafer fab process
  - Realized yields needed for production
  - Extending the results to 112PAM4 for phase 2 feasibility
- Implemented a dual aperture structure to realize laser sparing which dramatically improves overall system reliability
- Designed and Fabricated TWO different 56GBd Photodiodes structures: GaAs based and InP based.
Glass Carrier Subassembly

Chip join process
- SAFE chip – Cu pillar w SnAg cap
- PD/VCSEL – Au/Sn w formic acid reflow
  - Tight spacing btw chips ~ 10µm
  - Reflow self alignment precision < 1µm
- No significant challenges remaining

Underfill process
- Structural UF for electrical chips
- Epoxy based Optical UF for PD/VCSEL
  - Material survey / 4 candidates evaluated / one selected
  - New process developed
  - Stable optical performance after solder reflow
- Remaining challenges
  - Automated dispense manufacturability (small distance btw SAFE and PD/VCSEL using distinct UF materials)
  - UF void control
Electrical Connection

- We have realized a high speed socket that is capable of >30GHz operation.
- 400um Pitch LGA to LGA
- Allows placement of the Co-Packaged transceiver after the ASIC is mounted to the substrate and/or test infrastructure
  - Tested to >10K insertions
- We have also designed and fabricated a solderable interconnect for MOTION transceiver.
  - Allows the same part to be soldered directly to the host ASIC
  - Allows for common assembly and test infrastructure
Test Card Emulating MCC

- MCC laminate assembled
  - SLC 2-2-2 stack up
  - GL102F low loss material
  - 105 x 105 mm
  - Mini-SMP connector for high speed electrical input/output/clock
  - Micro-berg headers for power and control
  - LGA sockets
  - Electrical link length matched
- Emulates a 1st Level ASIC package
- Full Link (both E and O) is possible with only a clock signal provided.
Optical Solution and Transceiver Assembly

- Developed a disconnectable optical cable solution allowing the transceiver to be reflowed
- Transceiver footprint: 13mm x 13mm x 6mm high (not including heat sink)
How much additional bandwidth can MOTION provide?

(1) Co-packaging enables higher-radix switches → flatter network.

(2) Optics on CPU & GPU modules enables higher on-node bandwidth.

(3) Co-packaged optics may free up electrical package pins for more electr. DRAM channels.

SUMMIT-like Node

Possible insertion points of co-packaged optics on switches
The benefits of higher-radix switches enabled by MOTION

- 3x more network end points for 21% fewer switch modules
- 2.8x higher bisection BW for 100 Gb/s per port (11.2x for 400 Gb/s)
- MOTION opens the way to direct-network-attached accelerators
Network performance analysis: MOTION vs. Summit

Venus discrete event network simulator

<table>
<thead>
<tr>
<th>Traffic and Network Generator</th>
</tr>
</thead>
<tbody>
<tr>
<td>RX buffer size</td>
</tr>
<tr>
<td>Message size</td>
</tr>
<tr>
<td>Generation distribution</td>
</tr>
<tr>
<td>Data rate</td>
</tr>
<tr>
<td>Load</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Adapter/Switch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
</tr>
<tr>
<td>Data rate per link</td>
</tr>
<tr>
<td>Delay</td>
</tr>
<tr>
<td>Switch Buffer per port</td>
</tr>
<tr>
<td>Packet size</td>
</tr>
<tr>
<td>Routing algorithm</td>
</tr>
</tbody>
</table>

- **Uniform, BitComplement, BitReverse**: Linear throughput increase for all systems → 2.8x and 11.2x higher throughput for 100 and 400 Gb/s data rates
- **BitTranspose**: earlier saturation due to the significantly fewer destination nodes → 4.3x and 17.2x higher throughput for 100 and 400 Gb/s data rates
- **MOTION-400**: best mean packet delay for all cases
MOTION (16x56Gbps NRZ Co-packaged Optical Assembly)

The MOTION Co-packaged Optical Assembly (COA) is targeted at low-latency optical-interconnect applications which value:
- High bandwidth density per square-millimeter
- Protocol Agnostic capability up to 56Gbps-NRZ/channel on 16-duplex channels with 1:1 redundancy of optical transmitter
- Low latency, and power, by maintaining NRZ encoding with pre-FEC BER performance $\leq 1e^{-12}$, and new short-reach (XSR) electrical interface

Three market segments identified and pursuing COA applications:
- Massively Parallel Processing / High Performance Computing & AI-Deep Learning
- Metro-access Edge Compute Equipment for network-edge datacenters & AI-Inference
- High-performance FPGA interconnects serving Aerospace, High-resolution Imaging, & future compute-accelerator technologies

Key challenge to commercialization is customer timeline for releasing applications & aligning funding to commit to a production contract:
- Concern with OEM funding coming available without an adequate lead time required for a commercial production ramp
- Team will utilize Phase 2 to continue advanced development milestones to prove feasibility and expand product capability to support 56Gbaud (112Gbps)-PAM4 modulation to increase COA value proposition to all markets including “OEM Switch”
Co-Packaging on Organic Laminates: MOTION Phase 2

- ARPA-E (U.S. Department of Energy) sponsored project, Phase 2: 2 years
  - IBM and Finisar Inc. (now II-VI Inc.)

- Target specifications
  - **Electrical Interface**: 80 channels @ 56G NRZ, single ended encoded bus (SE-BUS)
  - **Optical Interface**: 32 channels @ 112G PAM-4, 16 fibers, 2 wavelengths
  - \(< 2 \text{ pJ/bit} \quad (\text{< 7 W, 32 channels})\)
  - 0°C to 70°C Case
  - \(6 \text{ dB (electrical) link budget} \quad (\text{XSR-like})\)
  - \(2 \text{ dB optical link margin (30m w/connector(s))}\)
  - W:13 mm x D:13 mm x H:4 mm
  - Package can withstand reflow onto ASIC 1st level laminate

=MOTION=: Multi-wavelength Optical Transceivers Integrated on Node
Phase 2 CMOS Transmitter and Receiver ICs

- Tx integrates 32x 112G PAM4 transmit lanes
- Quarter rate architecture
- Input C4 clock, which comes from a single clock multiplier, runs at ¼ of the baud rate (56/4=14GHz)
- Power budget
  - Electrical: 0.3pJ/bit, Optical interface: 0.7pJ/bit

- Rx integrates 32x 112G PAM4 receive lanes
- Quarter rate architecture
- Input C4 clock, which comes from a single clock multiplier, runs at ¼ of the baud rate (56/4=14GHz)
- A phase rotator is used to adjust the phase of the incoming ¼ rate clock. The output of the phase rotator goes to a quadrature clock phase generator
- Power budget
  - Electrical: 0.5pJ/bit, Optical interface: 0.5pJ/bit
MOTION Phase 2: 100GPAM4 VCSELs and PDs

- In Phase 2 we will continue to extend the work on high speed VCSELs 112PAM4 on a high volume production EPI and Wafer Fab
- Lowering the cost of fiber optics communication to DAC cost levels
  - Necessary for server to switch applications
- Lowering the total power to <2pJ /bit (including the light source)
- Multi Wavelength to further utilize the fiber bandwidth (cost)
During Phase 2 we will extend the capability to create a $2\lambda \times 16\text{ch} \times 112\text{PAM4 Transceiver}$

The glass substrate will remain the same size (4x the BW density) however the LGA pitch will shrink to 300um.

Optics will include the 32ch WDM MUX & Demux in the same vertical space.

Our target is to retain reflowability and develop a 300um pitch socket for the transceiver as well.

Thermal path for 105mW/mm$^2$
Phase 2: IBM SYSTEMS TECHNOLOGY EVALUATION

- **Goal:** To assess the technology readiness of co-packaged optics
  - Optical transceivers soldered directly on the top surface of a laminate package will be built
    - Positive results after two evaluation cycles (two cycles of learning) would result in a recommendation that this technology could move into productization phases of work
  - Socketed optical transceivers will be evaluated through modeling
- The IBM Systems group will perform this technology evaluation focusing on the thermal & mechanical robustness of packaging:
  - Four (4) optical transceivers and one (1) test site die on the top of a single FC-PLGA laminate, assembled with a thermal lid
- **Evaluation challenges:**
  - Assembly Processing
  - Package Reliability
  - Thermal Performance

_Demonstrating a viable path to system integration is necessary before this technology can be considered in a product plan._
Acknowledgements: MOTION Phase 1 Team members & Sponsor

- IBM Research

- IBM Bromont
  - L-M. Achard, P. Fortier, C. Dufort, E. Tucotte, C. Bureau, M. Pion, Y. Cossette, P. Ducharme, S. Desputeau, A. Janta-Polczynski, P. Minier

- Finisar Corporation

**Acknowledgment:** “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-AR0000846. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.”