Re-architecting DRAM with Silicon Photonics

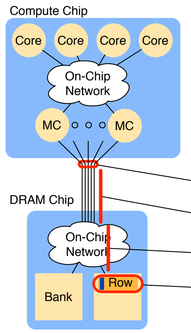
S. Beamer, C. Sun, Y. Kwon, A. Joshi, C. Batten, V. Stojanović, K. Asanović

EECS Electrical Engineering and Computer Sciences

Berkeley Par Lab

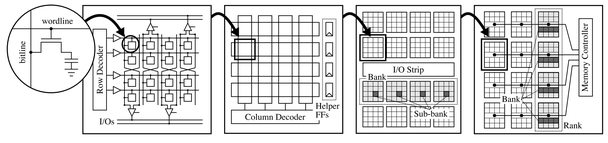
PARALLEL COMPUTING LABORATORY

INTRODUCTION

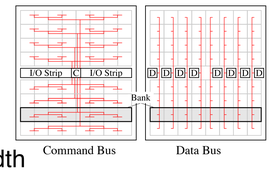


- To support a manycore system will require tremendous memory bandwidth, but if current electrical interconnects are scaled up, they may be limited by:
  - Pin bandwidth
  - I/O energy
  - Cross-chip energy
  - Activation energy
- In this work, we will:
  - Use silicon photonics to reduce the I/O energy and the cross-chip energy and get past pin bottlenecks
  - Reduce the row size to reduce the activation energy

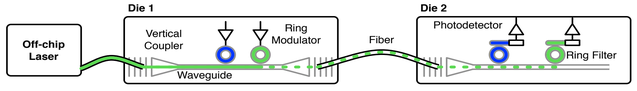
DRAM OVERVIEW



- Heavily commoditized part
  - Cell density is crucial
- Historically design decisions have been made to increase density at the price of energy efficiency, latency, or bandwidth
- Array blocks from different banks can share pins



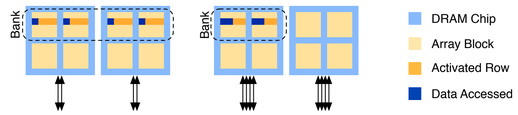
PHOTONICS OVERVIEW



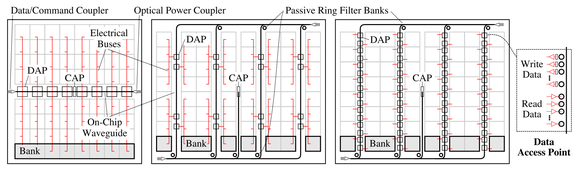
- Advantages
  - High bandwidth density off-chip (DWDM)
  - Energy efficient off-chip
  - Seamless interchip links (Monolithically Integrated)
- Can fit 64 wavelengths per direction each at 10Gbps

PROPOSED DESIGN

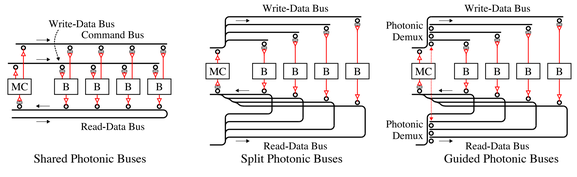
- Redesigning the Bank
  - Reduce the row size by increasing the number I/Os per array core



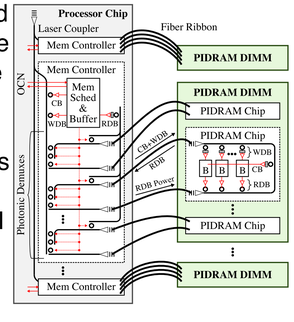
- Redesigning the Chip
  - Bring photonics past the chip edge into the chip
  - Electrical buses cover remaining distance from access points to array blocks
  - Control still broadcast electrically



- Redesigning the Channel
  - Want multiple chips to share channel to increase capacity independent of bandwidth
  - We propose *Optical Power Guiding*
  - Direct laser power only to where it is needed



- Redesigning the System
  - All off-chip links (data & control) are optical
  - A single memory controller controls one channel which may have multiple chips
  - With smaller rows and higher bandwidth, one chip can supply entire access width
  - Only 2 fibers per chip
  - Memory controller has complete information, so it switches channel (no need for global arbitration)



FURTHER SCALING

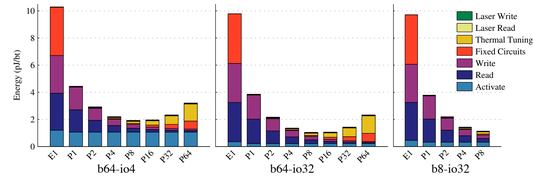
- 3D Stacking is complementary
  - Could stack DRAM chips to increase capacity
  - Less area overhead than electrical stacking
- Combining with optical power guiding
  - First level of switching selects stack
  - Second level switching selects die

METHODOLOGY

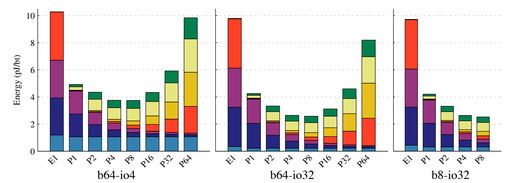
- Photonic Model
  - Conservative & aggressive projections
- DRAM Model
  - Heavily modified CACTI-D for 32nm
  - Validated against multiple points & processes
- Architectural Model
  - Custom C++ simulator with random traffic
- Coverage
  - Modeled hundreds of points, present three representative ones in the work

RESULTS

- Latency
  - Mostly unchanged since internals of array core left mostly unchanged and activation latency dominates
- Energy vs. Design
  - Biggest gain comes from off-chip efficiency
  - Aggressive projection at 100% utilization

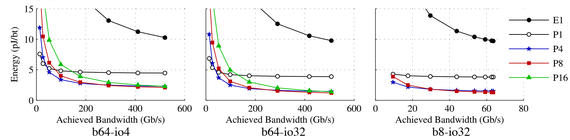


  - Conservative projection at 100% utilization

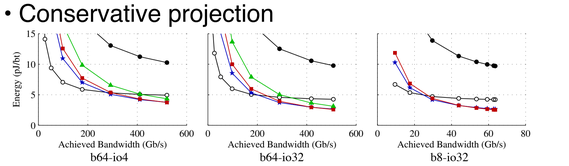


RESULTS

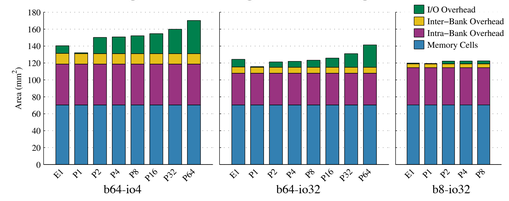
- Energy vs. Utilization
  - Best energy efficiency at high utilization, but electrical is always the worst
  - Aggressive projection



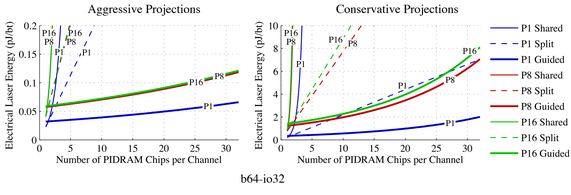
  - Conservative projection



- Area vs. Design
  - Decreasing row size good for high bandwidth



- Scaling Number of DRAM Chips per Channel
  - Guided bus is *much* more scalable



CONCLUSION

- Our redesigned system is able to obtain nearly 10x improvement in energy efficiency
- To fully reduce energy, must attack it in all places
  - I/O, cross-chip, and within bank
- Surprisingly, our modifications are area neutral
- Optical power guiding makes it easier to increase capacity without paying for extra unneeded bandwidth

Poster template inspired by Sam Williams et al.