The Potential Impact of Silicon Photonics Networks for Graph Analytics

PACIFIC NORTHWEST NATIONAL LABORATORY

Kevin J. Barker, Vito Castellana, Daniel Chavarría, Mahantesh Halappanavar, Adolfy Hoisie, Darren J. Kerbyson, Andrés Marquez, Nathan Tallent, Antonino Tumeo

Talk Outline

- Workload-specific modeling at PNNL
  - Exploration of architectural concepts, from a workload perspective
  - Analytically modeling multiple metrics of interest
- Introduction to Silicon Photonics architectures
  - IBM TOPS architecture
  - Oracle Macrochip architecture
- Workloads of interest: graph analytics
  - Community Detection
  - Half-approximate Weighted Matching
- Performance analysis
- Power/energy analysis

Modeling at PNNL

- Analysis of large-scale application performance
  - Analytical modeling approach
  - Workload-centric focus
  - Full-scale production-level codes from a variety of domains
  - Current and future systems
- Interests from technologies to system architectures
  - Both current and future technologies
    - Processing
    - Memories
    - Interconnection networks
  - Exploring beyond large-scale systems (e.g., embedded)
- Multiple metrics of interest
  - Interplay between performance, power consumption, thermal effects, and resilience

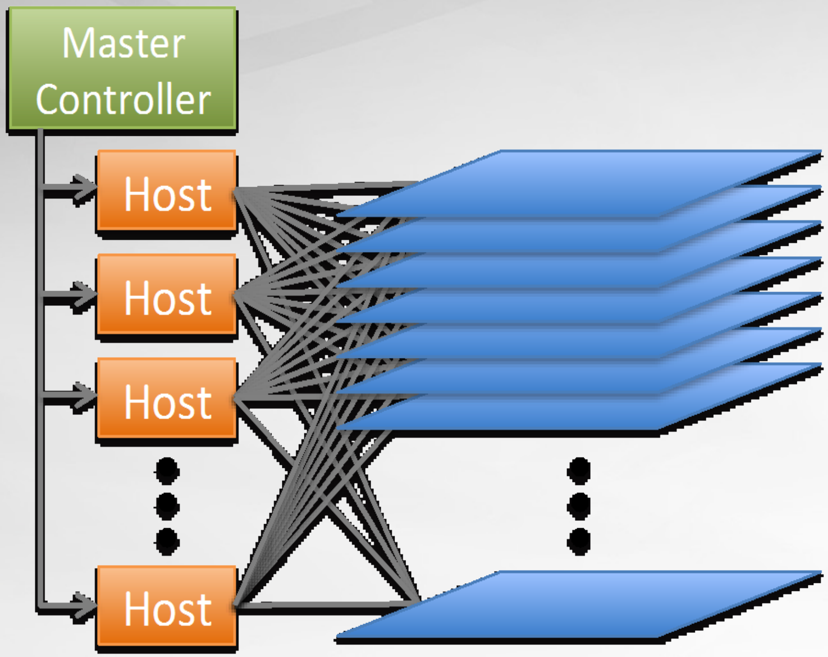
Assessing the Impact of Silicon Photonics

- Question: what will be the impact of silicon-photonics networking technology on graph-based workloads in the 4 to 6 year timeframe?

- Methodology:
  - Work with architects to understand representative silicon-photonics enabled architectures of interest to DARPA's POEM program
  - Draw workloads from PNNL's experience with graph-based applications
  - Model intra-node and inter-node data movement and compare silicon-photonics enabled architectures with potential future electrical solutions
  - Modeling to explore both *performance* and *power/energy consumption*

- Thanks to the IBM TOPS architecture team and the Oracle Macronode architecture team for their valuable contributions

Silicon Photonics: IBM TOPS



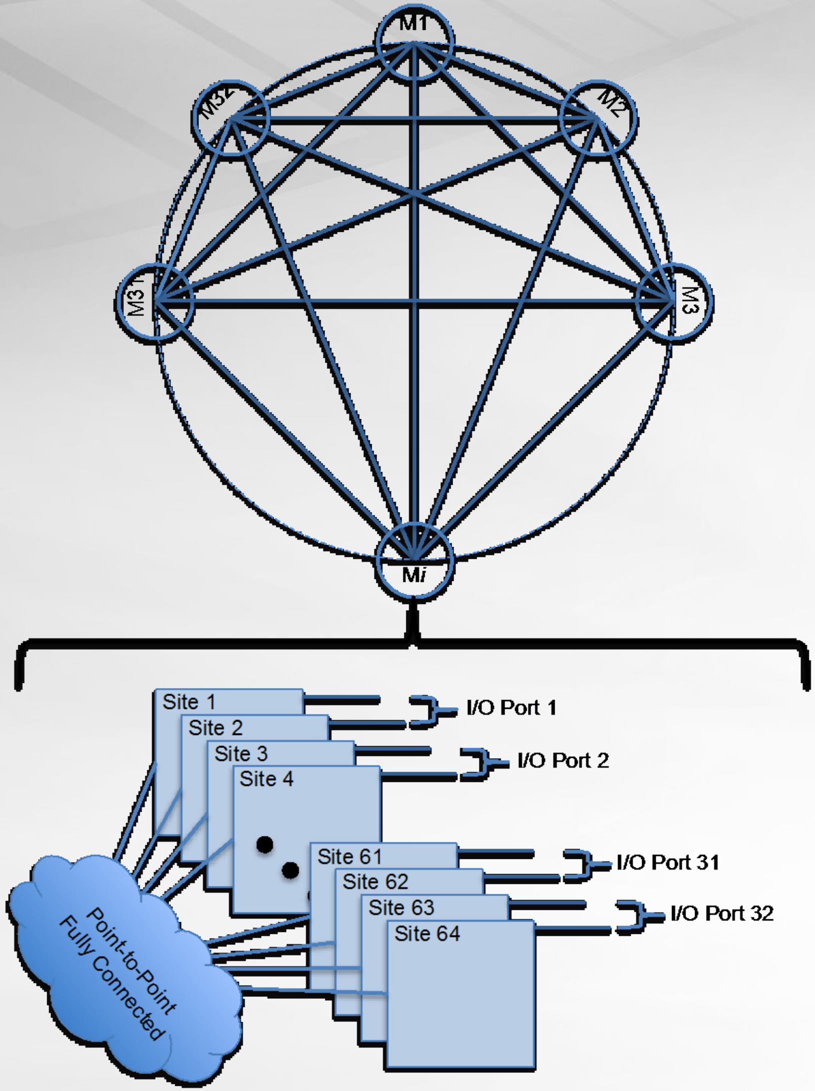
64 Compute Nodes

256 64×64 Optical Switch Planes

- Node architecture
  - 4 sockets (64 total cores)
  - Optical Hub Chip
- Inter-node network
  - 64 node system
  - Each optical switch plane is 64×64 crossbar
  - One fiber from each node to each optical switch plane
    - 16 wavelengths per fiber
    - 2.5 GB/s BW per wavelength
  - 256 switch planes: 4 switch planes between each node pair with no switching

L. Schares, et. al., “A Throughput Optimized Optical Network for Data Intensive Computing”, IEEE Micro, Sept/Oct, 2014.

Silicon Photonics: the Oracle Macrochip



- Macrochip architecture
  - 64 compute/memory sites
  - Fully connected network
- Intra-node network
  - 128 GB/s total BW per site
  - 2 GB/s (i.e., one color) per site pair
- Inter-node network
  - Two sites connect to each I/O port
  - 32 ports per Macrochip create a fully-connected 32-node system
  - 256 GB/s per Macrochip pair

A. V. Krishnamoorthy, et. al., “Energy-Efficient Photonics in Future High-Connectivity Computing Systems”, Jour. of Lightwave Technology, Feb. 2015.

Architectural Comparison Points

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

	IBM TOPS			Oracle Macrochip		
	Optical	Electrical: Fixed Footprint	Electrical: Fixed Power	Optical	Electrical: Fixed Footprint	Electrical: Fixed Power
Node Count	64	64	64	32	32	32
Sockets per Node	4	4	4	64	64	64
Intra-Node Topology	Fully-Connected	2D Mesh QPI	2D Mesh QPI	Fully-Connected	2D Mesh QPI	2D Mesh QPI
Inter-Node Topology	256 Switch Planes	Fat-Tree	Multiplane Fat-Tree	Fully-Connected	Fat-Tree	Multiplane Fat-Tree
Comm. Lanes (Intra/Inter)	16/64	18/4	18/20	1/128	18/4	18/48
Latency (Intra/Inter) (μs)	0.5/0.5	0.5/0.5	0.5/0.5	0.5/0.5	0.5/0.5	0.5/0.5
Per-Lane BW (Intra/Inter) (GB/s)	2.5/2.5	1.4/6.2	1.4/6.2	2.0/2.0	1.4/6.2	1.4/6.2

- Optical networks are Silicon Photonics enabled as described in the literature
- Electrical networks are based on project 4×HDR IB technology
  - “Fixed Footprint” connects nodes with single switch (32 or 64 ports)
  - “Fixed Power” attempts to equate optical and electrical network power consumption by utilizing multiple electrical switch “planes”

Two Graph Analytics Workloads

Scale-40 distributed graphs

Community Detection

- Input: Graph with weighted edges
- Output: Disjoint sets of related vertices
- Aggregated personalized all-to-all to send each edge's target info (~1 GB)
- Iterate until Δ-modularity < threshold
  - Each vertex initially its own community
  - For each vertex, determine whether *modularity* increases by moving to neighboring community

Large, aggregated messages

- Improve network performance
- Combine reqs with same target vertex

More computation

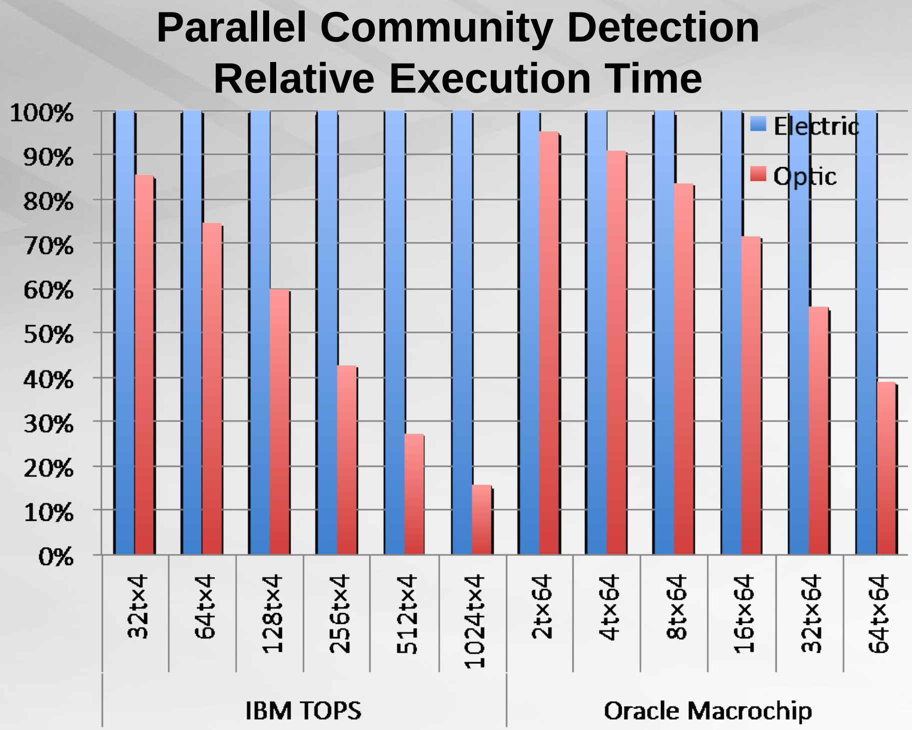
- Denser graph; aggregation cost
- Modularity requires collectives

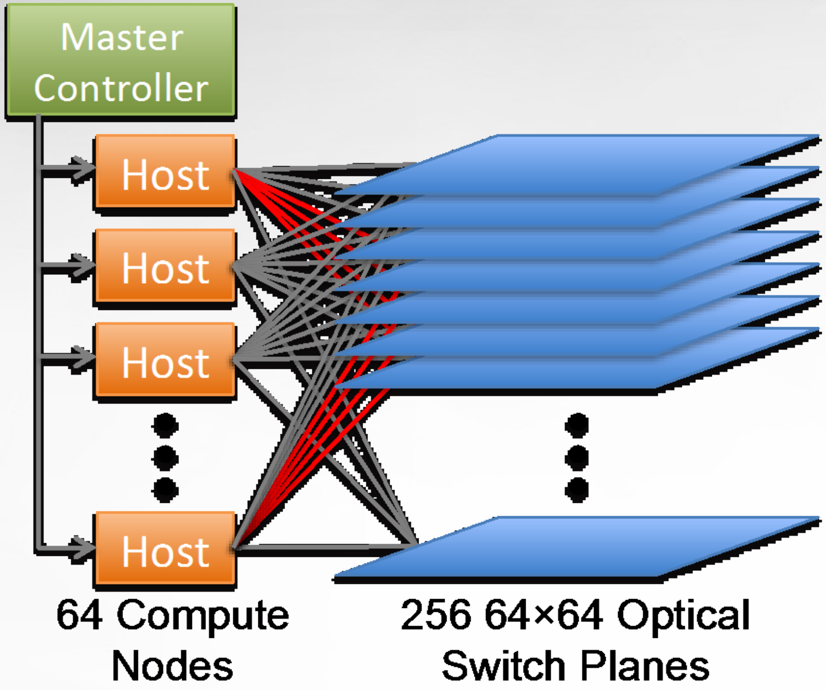
Half-Approximate Weighted Matching

- Input: 2D mesh with weighted edges
- Output: Maximal weighted matching
- Two phases b/c of multi-step protocol
  - Based on locally dominant neighbor
- Phase 1:
  - Try matching each vertex
  - Aggregate messages between nodes
- Phase 2:
  - Try matching on "matched frontier"
  - Iterate until all vertices are matched
  - Use very small (24 B) messages

Small messages

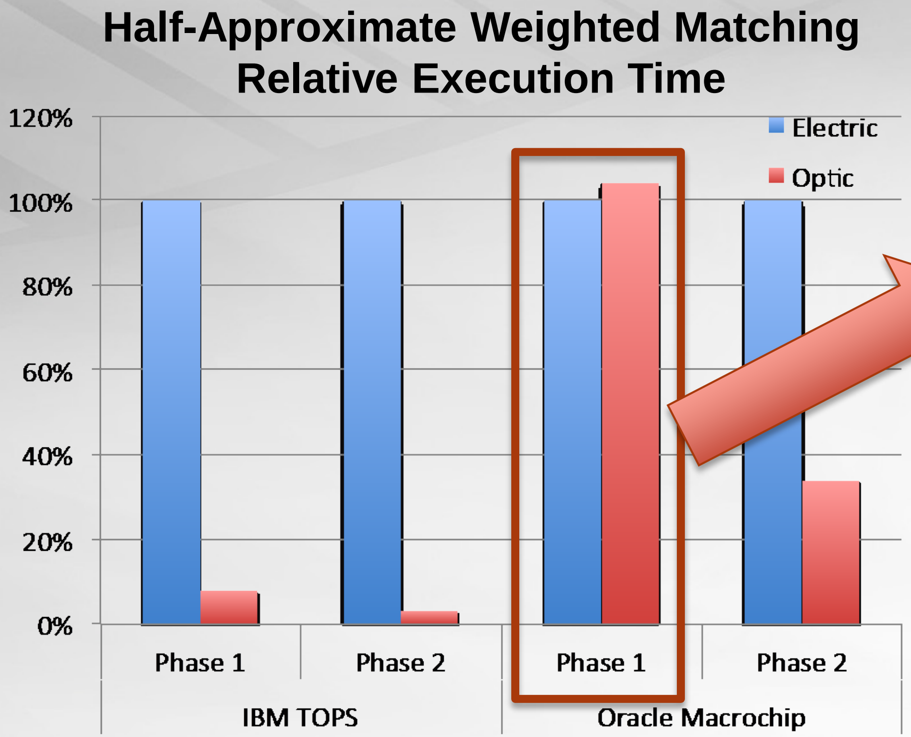
Performance Analysis: Community Detection



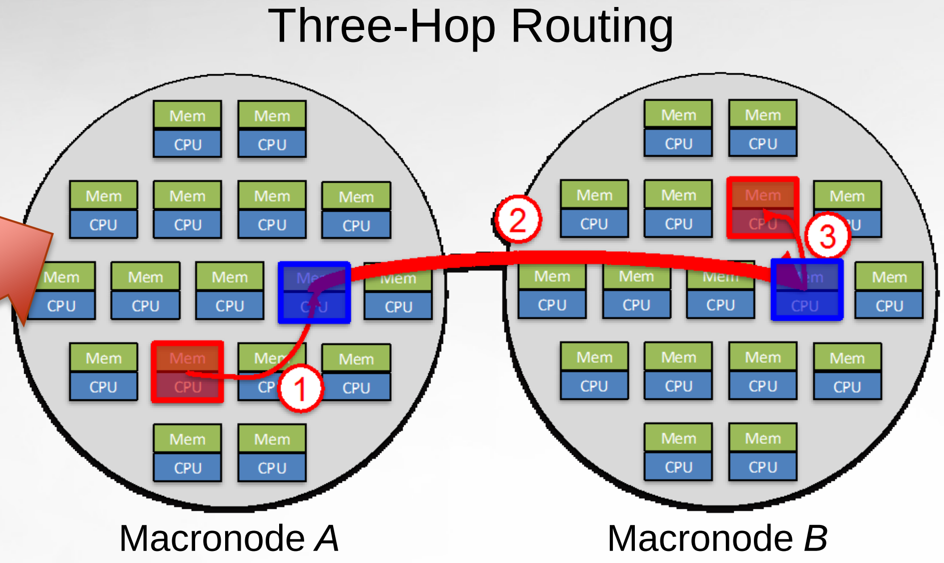


- Increasing thread count places greater relative emphasis on communication performance
- Communication performance improvement due to:
  - Improved link bandwidth (40 GB/s vs. 25 GB/s)
  - Message *striping* across multiple switch planes
  - Greater communication concurrency due to topology

Performance Analysis: Matching



Half-Approximate Weighted Matching Relative Execution Time

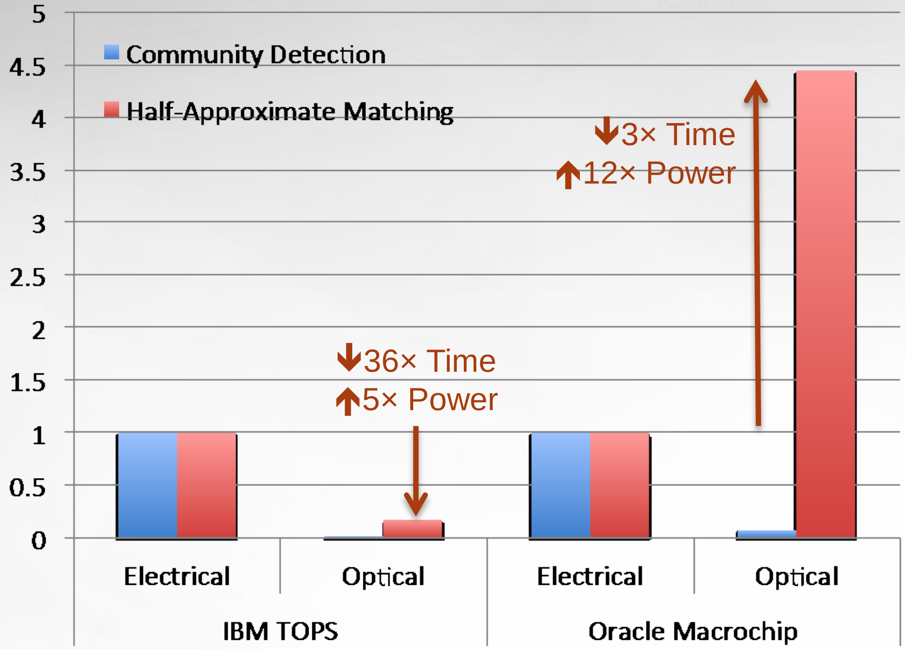


Three-Hop Routing

- Matching Phase 1 uses large messages in a 2D mesh pattern
  - Each site uses only four of the available 64 outgoing links
- Direct routing between Macronodes requires three hops
  - Intra-node site-to-site links offer only single-way concurrency and relatively low bandwidth
- *Indirect Intra-node Routing* may alleviate this problem by utilizing all available intra-node bandwidth

Modeled Energy Analysis

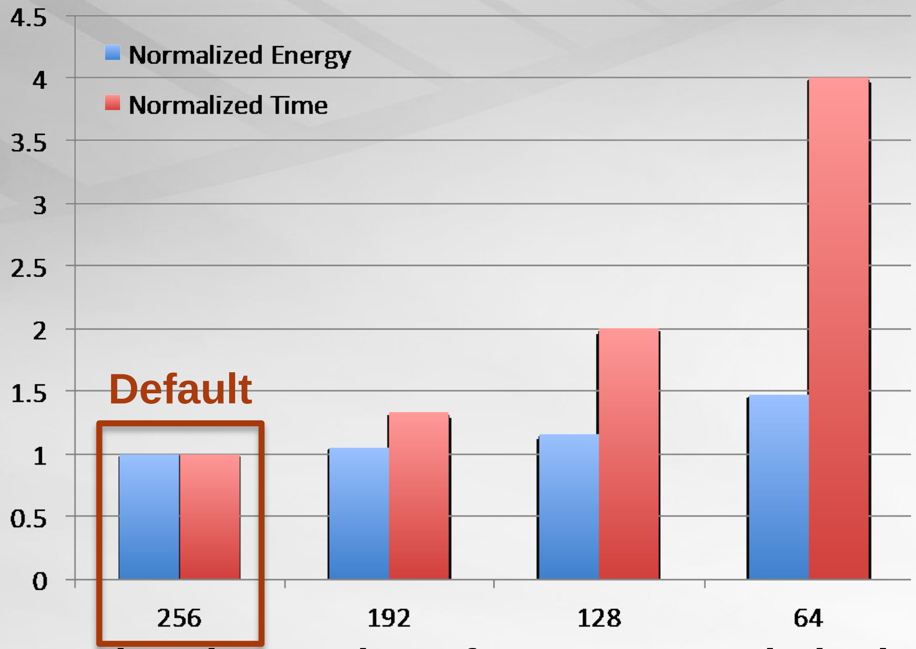
HDR InfiniBand	
Switch Power	200 Watts
HCA Power	15 Watts
Total Power	**1160 W (64)** **680 W (32)**
IBM TOPS	
Switch Plane Power	20 Watts
Hub Chip Power	15 Watts
Total Power	**6080 Watts**
Oracle Macrochip	
Intra-node Network Power per Node	65 Watts
I/O Port Power per Node	197 Watts
Total Power	**8384 Watts**



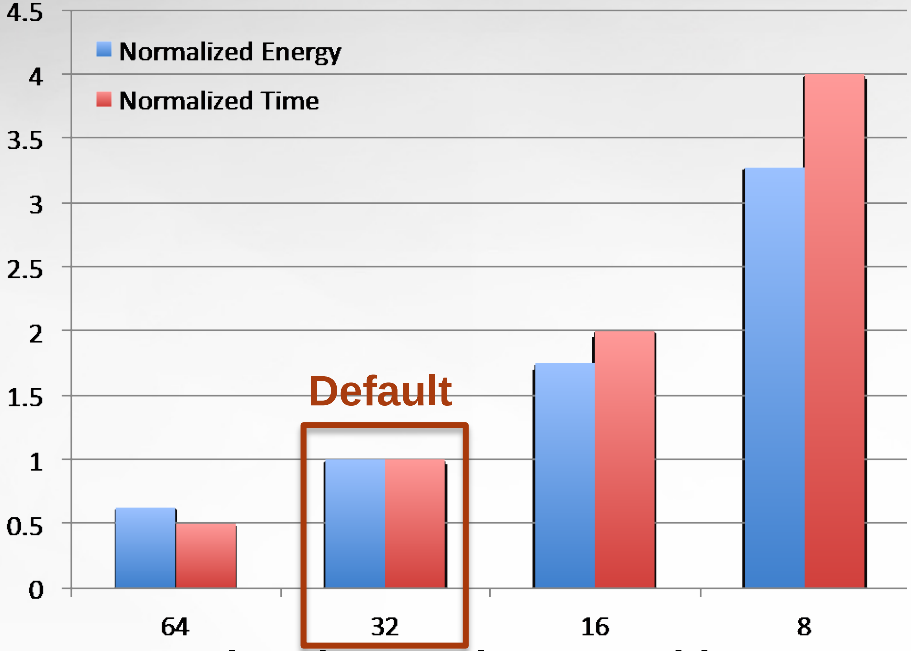
Relative Energy Consumption

- Fixed-footprint electrical power consumption is lower than either optical network: ~5× (IBM TOPS) & ~12× (Oracle Macrochip)
- Improved optical network performance often results in energy win
  - Exception: Lack of intra-node BW and network concurrency impairs Half-Approximate Weighted Matching performance on Oracle Macrochip
- *Fixed power* electrical networks improve performance at the cost of increased power, yielding nearly constant *energy*

Exploring Alternate System Configurations



Varying the number of IBM TOPS switch planes
Relative Performance and Energy



Varying the Oracle Macrochip count
Relative Performance and Energy

- Models allow us to explore *hypothetical* system configurations
  - We vary the number of optical switch planes and Macrochip count
  - Results are relative to the default system configuration
- Results are for communication only; energy analysis does not consider core power

Conclusions

- Silicon-photonics shows promise in both performance and energy
- Silicon-photonics enabled networks show promise for graph analytics applications
  - Improved link bandwidth benefits large messages
  - Link *concurrency* benefits large numbers of small messages
  - Rich topologies benefit applications with all-to-all communication patterns
  - Performance/Energy improvements are workload dependent
- Mapping from application to architecture impacts performance
  - Algorithms with similarly rich communication patterns can find substantial performance and energy benefits
  - Algorithms whose communication patterns do not exploit topology may suffer without mechanisms such as *indirect routing*