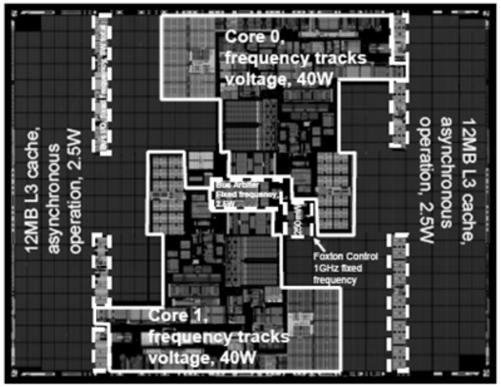
On-Chip Photonic Communications for High Performance Multi-Core Processors

Keren Bergman, Luca Carloni, Columbia University
Jeffrey Kash, Yurii Vlasov, IBM Research

HPEC 2007, Lexington, MA 18-20 September, 2007

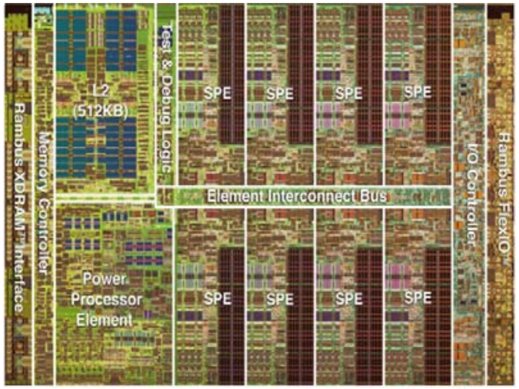
Chip MultiProcessors (CMP)

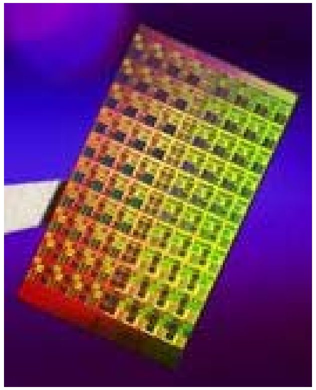


Core 0, frequency tracks voltage, 40W
12MB L3 cache, asynchronous operation, 2.5W
Bus Arbiter Fixed frequency, 2.5W
Foxton Control 1GHz fixed frequency
12MB L3 cache, asynchronous operation, 2.5W
Core 1, frequency tracks voltage, 40W

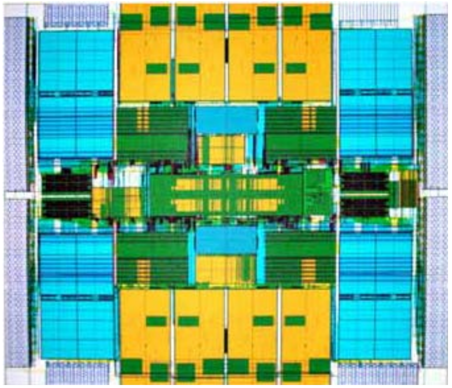
Montecito
Intel 2004

CELL BE
IBM 2005

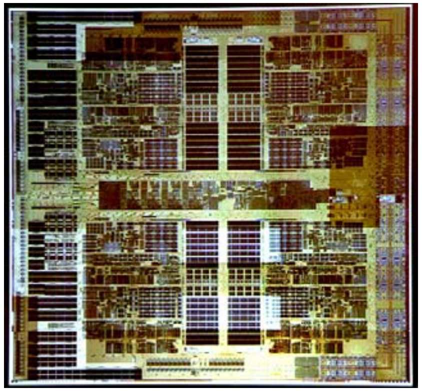




Terascale
Intel 2007



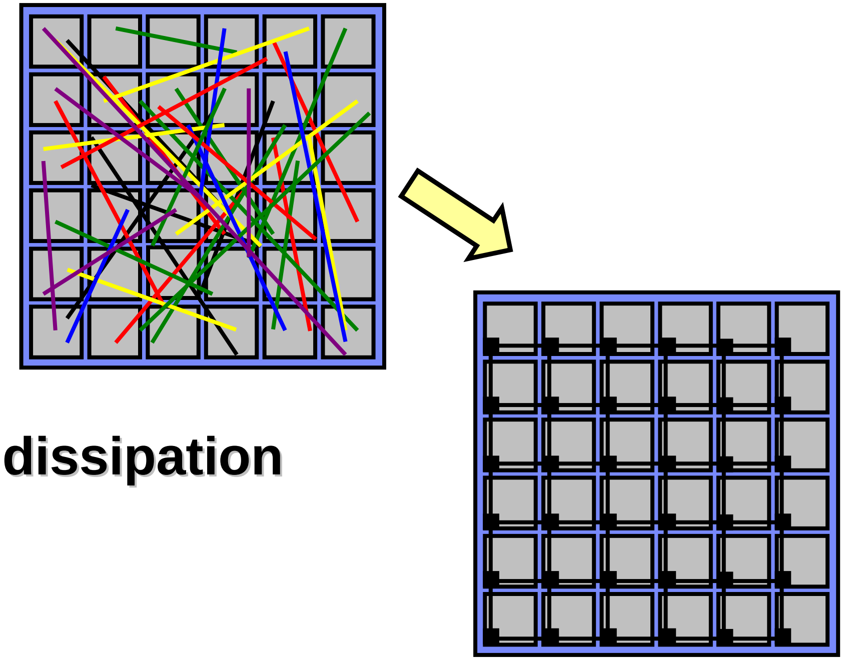
Niagara
Sun 2004



Barcelona
AMD 2007

Networks on Chip (NoC)

- Shared, packet-switched, optimized for communications
  - Resource efficiency
  - Design simplicity
  - IP reusability
  - High performance

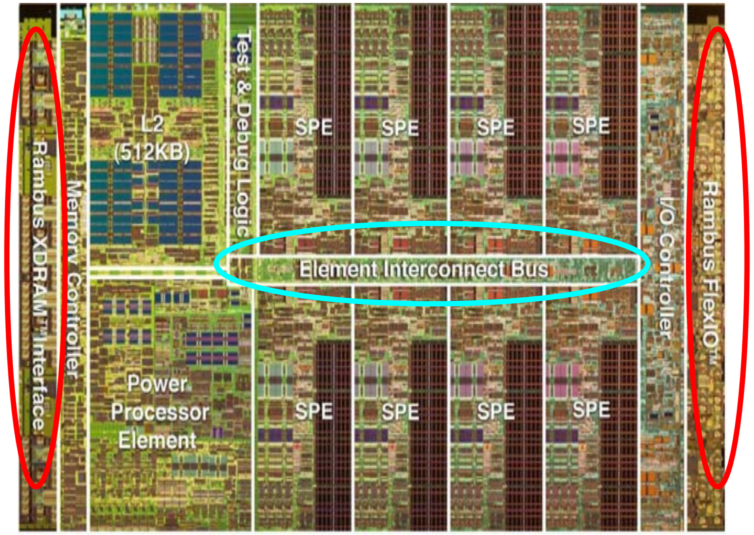


Kolodny, 2005

- But… no true relief in **power dissipation**

Chip MultiProcessors (CMPs)
IBM Cell, Sun Niagara, Intel Montecito, …

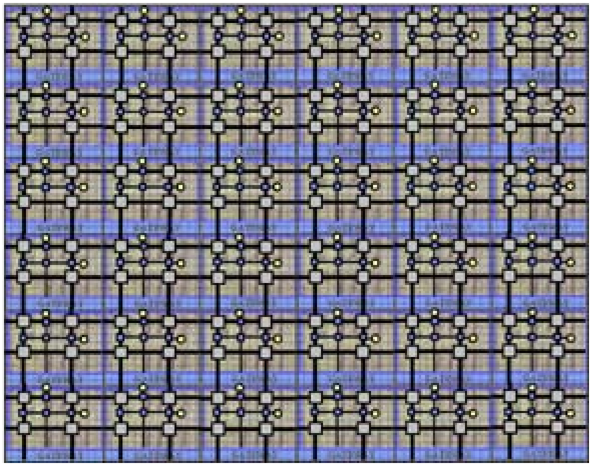
IBM Cell:



Parameter	Value
Technology process	90nm SOI with low-κ dielectrics and 8 metal layers of copper interconnect
Chip area	235mm^2
Number of transistors	~234M
Operating clock frequency	4Ghz
Power dissipation	**~100W**
Percentage of power dissipation due to global interconnect	**30-50%**
Intra-chip, inter-core communication bandwidth	**1.024 Tbps, 2Gb/sec/lane (four shared buses, 128 bits data + 64 bits address each)**
I/O communication bandwidth	**0.819 Tbps (includes external memory)**

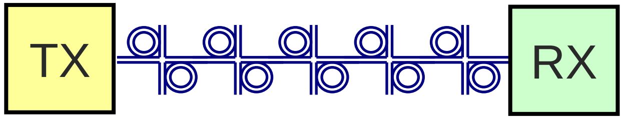
Why Photonics for CMP NoC?

Photonics changes the rules for Bandwidth-per-Watt



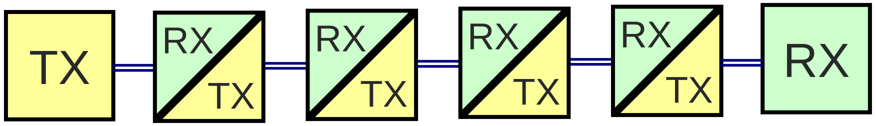
OPTICS:

- Modulate/receive ultra-high bandwidth data stream **<u>once</u>** per communication event
- Transparency: broadband switch routes entire multi-wavelength high BW stream
- Low power switch fabric, scalable
- Off-chip and on-chip can use essentially the same technology
- **Off-chip BW = On-chip BW**
  for *same power*

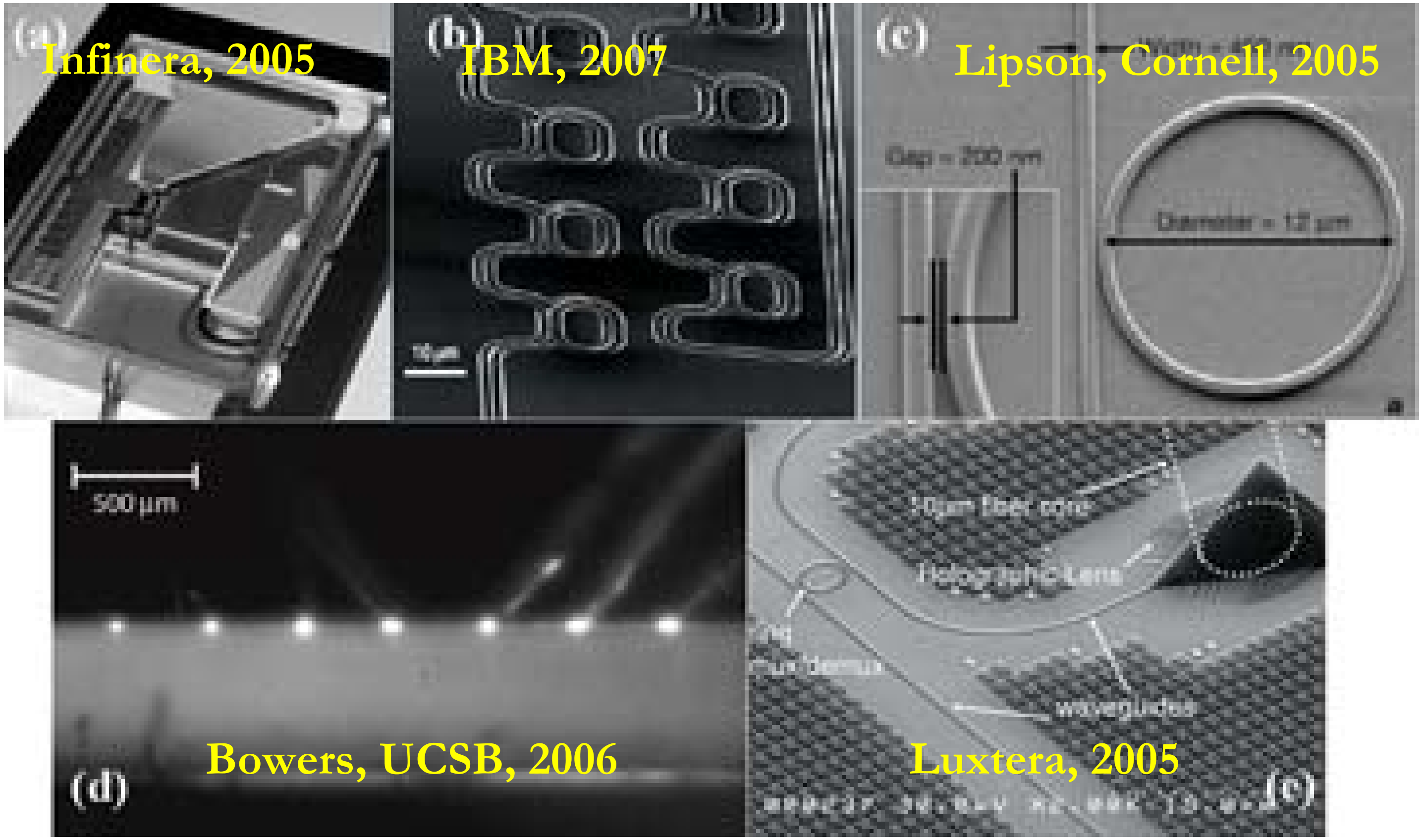


ELECTRONICS:

- Buffer, receive and re-transmit at **<u>every switch</u>**
- **<u>Off chip</u>** is pin-limited and really power hungry



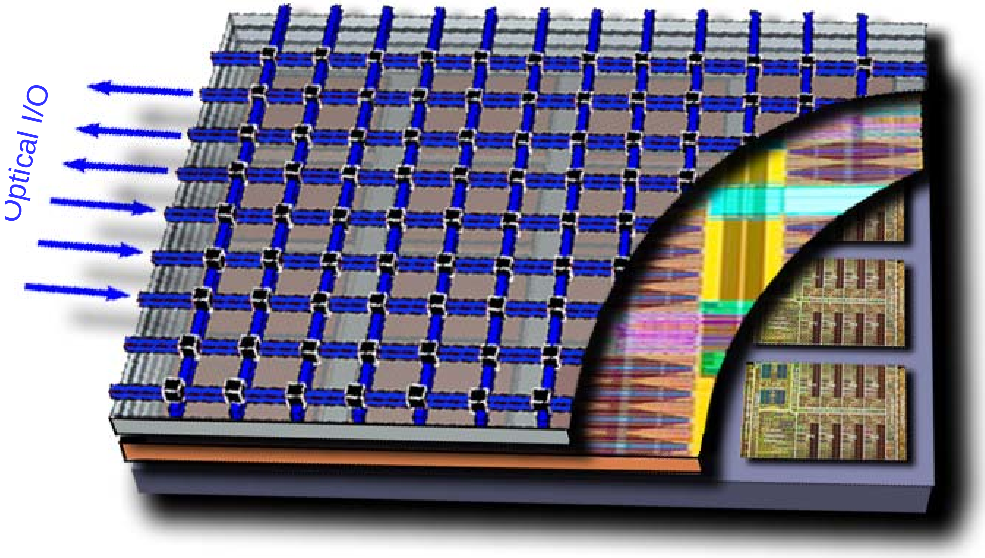
Recent advances in photonic integration



3DI CMP System Concept

- Future CMP system in 22nm
- Chip size ~625mm^2
- 3D layer stacking used to combine:
  - Multi-core processing plane
  - Several memory planes
  - Photonic NoC

Processor System Stack



- For 22nm scaling will enable 36 multithreaded cores similar to today’s Cell
- Estimated on-chip local memory per complex core ~0.5GB

Optical NoC: Design Considerations

- **Design to exploit optical advantages:**
  - **Bit rate transparency:** transmission/switching power independent of bandwidth
  - **Low loss:** power independent of distance
  - **Bandwidth:** exploit WDM for *maximum* effective bandwidths across network
    - (Over) provision *maximized* bandwidth per port
    - Maximize *effective* communications bandwidth
  - **Seamless optical I/O to external memory with same BW**

- **Design must address optical challenges:**
  - No optical buffering
  - No optical signal processing
  - Network routing and flow control managed in electronics
    - Distributed vs. Central
    - Electronic control path provisioning latency

- **Packaging constraints:** CMP chip layout, avoid long electronic interfaces, network gateways must be in close proximity on photonic plane
- Design for photonic building blocks: low switch radix

Photonic On-Chip Network

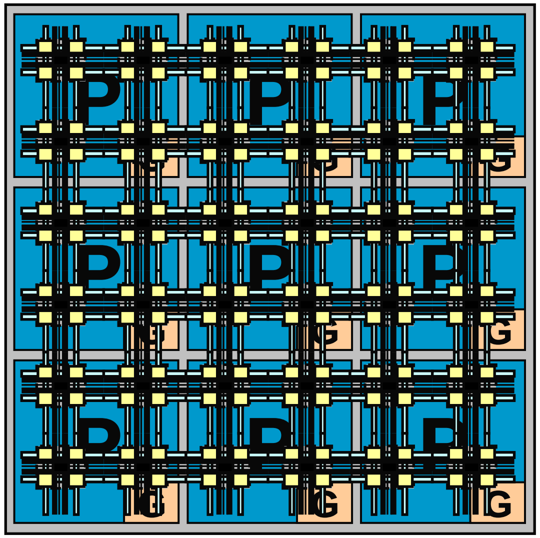
- **Goal: Design a NoC for a chip multiprocessor (CMP)**
- Electronics
  - ✓ Integration density → abundant buffering and processing
  - ✗ Power dissipation grows with data rate
- Photonics
  - ✓ Low loss, large bandwidth, bit-rate transparency
  - ✗ Limited processing, no buffers
- Our solution – a hybrid approach:

  A dual-network design

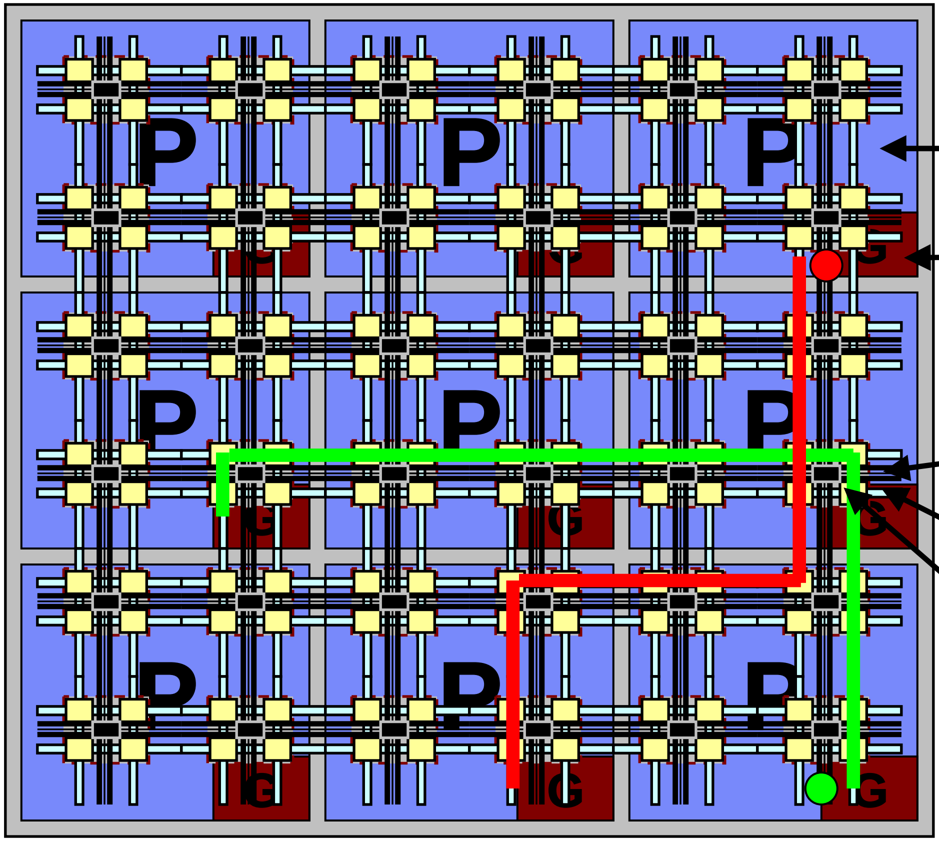
  – Data transmission in a photonic network

  – Control in an electronic network

  – Paths reserved before transmission → No optical buffering



On-Chip Optical Network Architecture
Bufferless, Deflection-switch based



Cell Core
(on processor plane)

Gateway to Photonic NoC
(on processor and photonic planes)

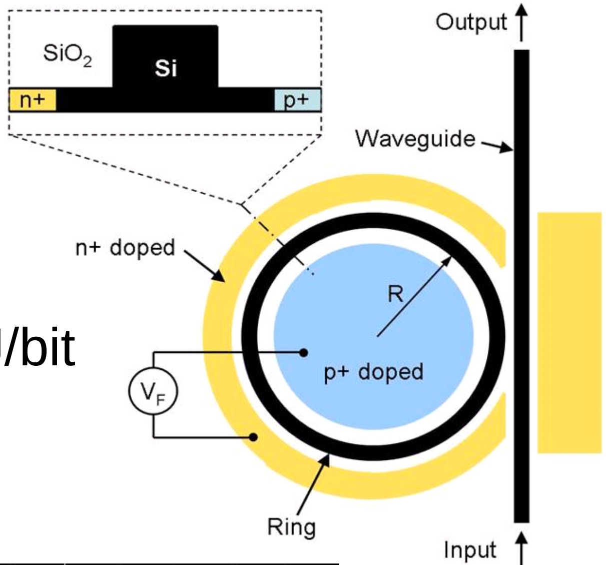
Thin Electrical Control Network
(~1% BW, small messages)

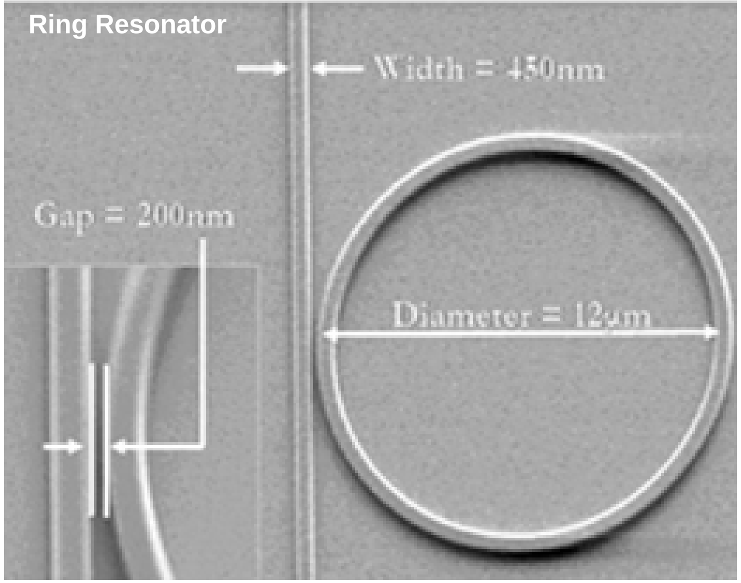
Photonic NoC

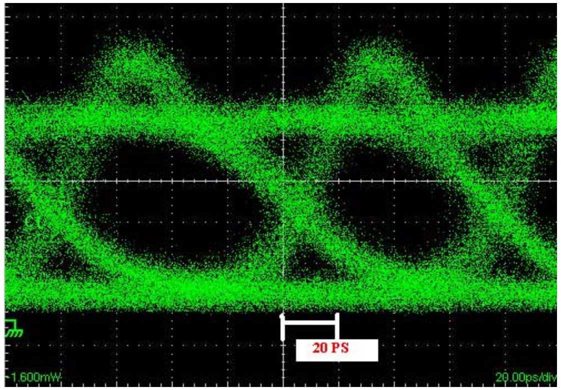
Deflection Switch

Building Blocks (1): High-speed Photonic Modulator

- Ring-resonator structure
- Achieve optical data modulation
- Compact ~ 10μm diameter for high density integration
- Ultra-low power ~ 1pJ/bit today, scalable to 0.1pJ/bit
- 12.5Gb/s demo, extendable to 40Gb/s



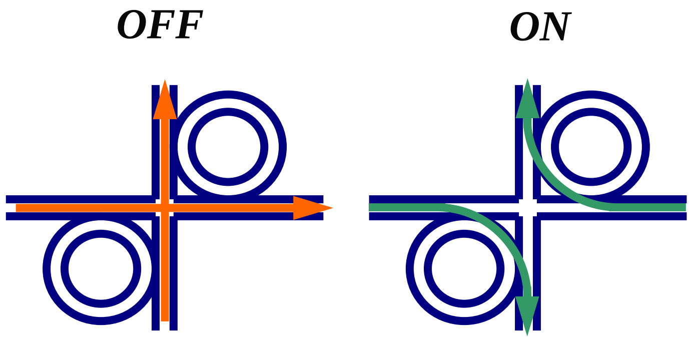


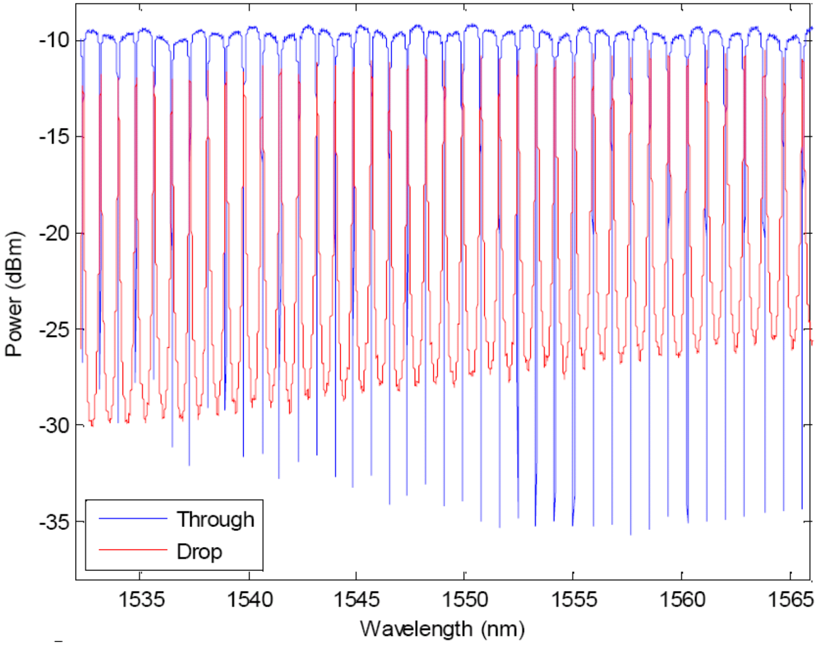


Recent 12.5GHz

Building Blocks (2): Broadband deflection switch

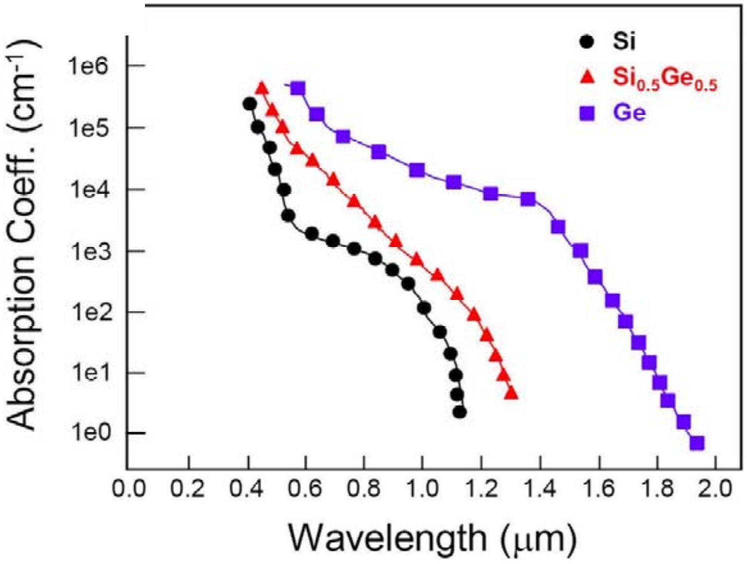
- Broadband ring-resonator switch
- **OFF state**
  - passive waveguide crossover
  - negligible power
- **ON state:**
  - carrier injection → coupling into ring → signal switched ~0.5mW

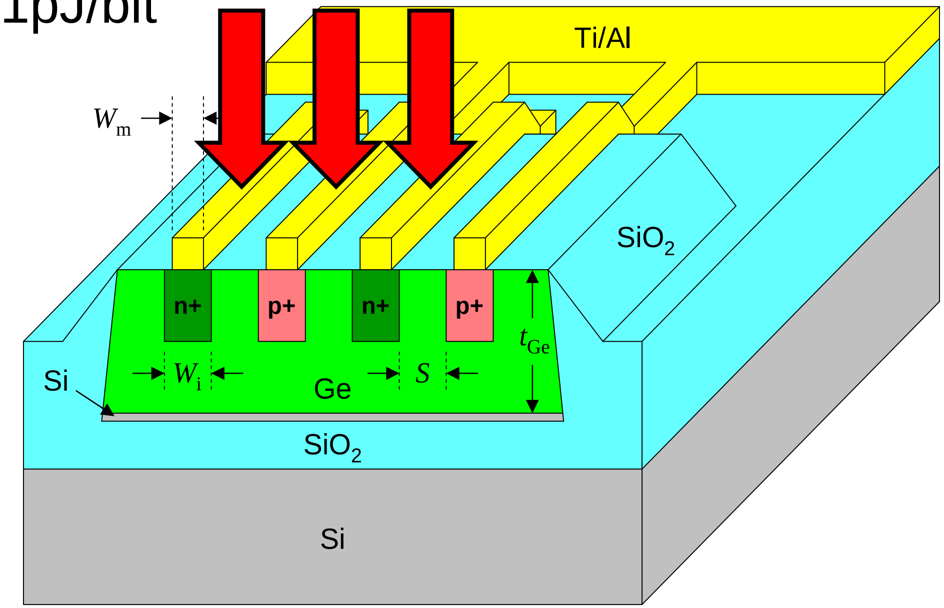




Building Blocks (3): Detector

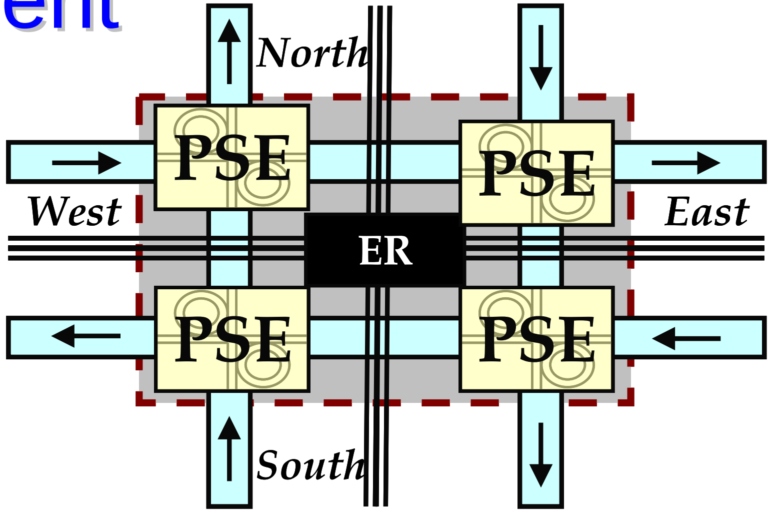
- Lateral PIN design, direct Ge growth on thin SOI (IBM)
- Low capacitance and dark current
- 20GHz Bandwidth
- Ultra-low power, 0.1pJ/bit today scalable to 0.01pJ/bit

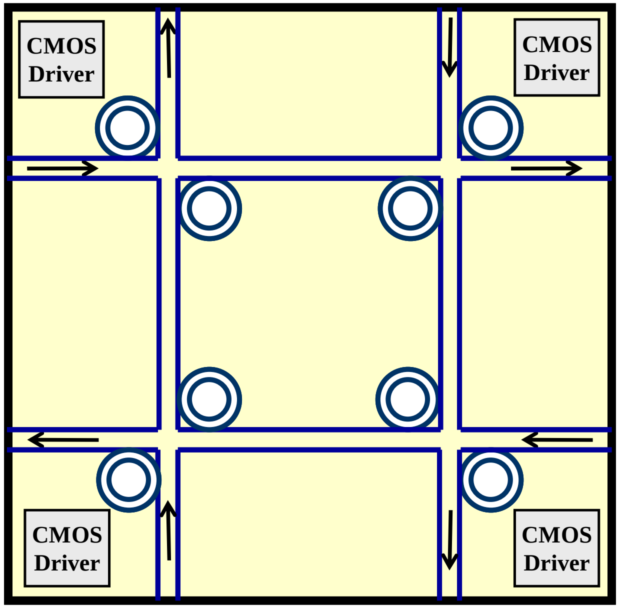




4x4 Photonic Switch Element

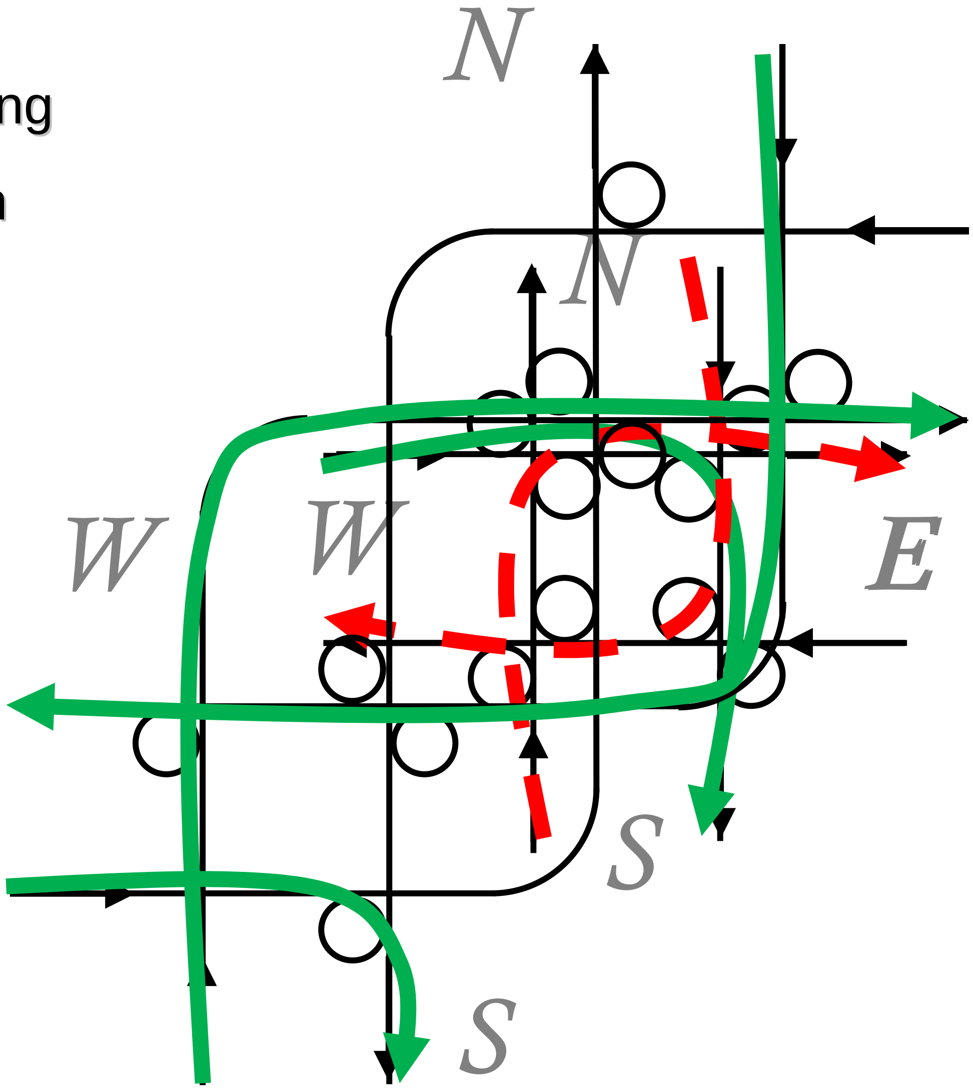
- 4 deflection switches grouped with electronic control
- 4 waveguide pairs I/O links
- Electronic router
  - High speed simple logic
  - Links optimized for high speed
- Small area (~0.005mm^2)
- **Nearly no power consumption in *OFF* state**





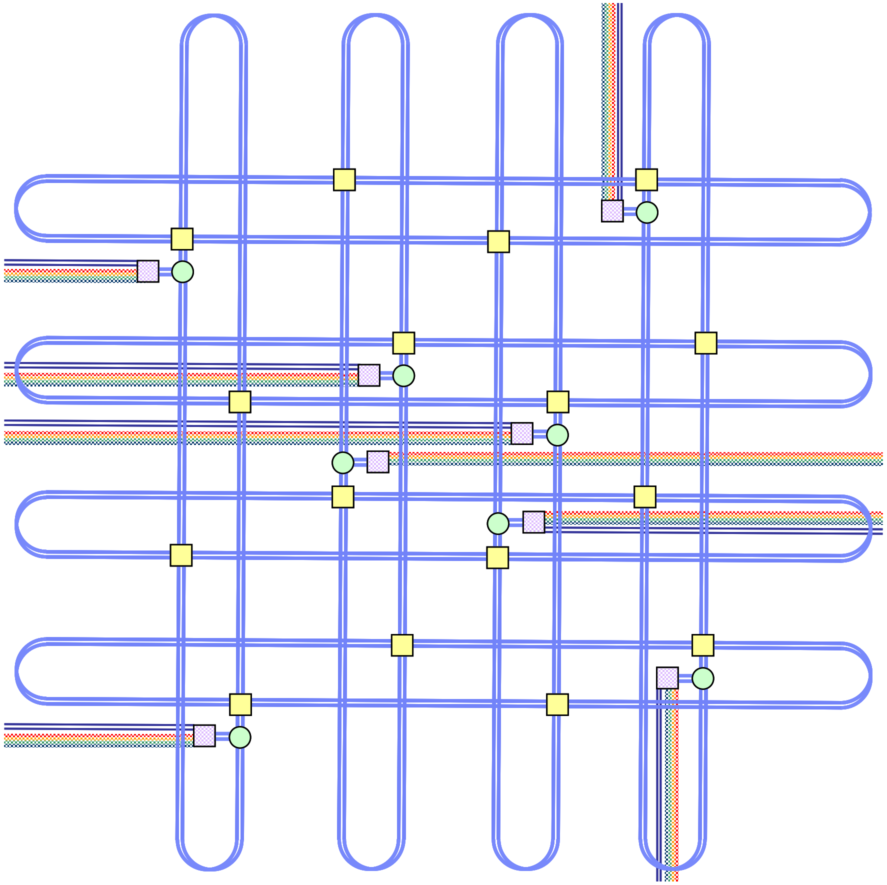
Non-Blocking 4x4 Switch Design

- Original switch is internally blocking
- Addressed by routing algorithm in original design
- Limited topology choices
- New design
  - Strictly non-blocking*
  - Same number of rings
  - Negligible additional loss
  - Larger area



* U-turns not allowed

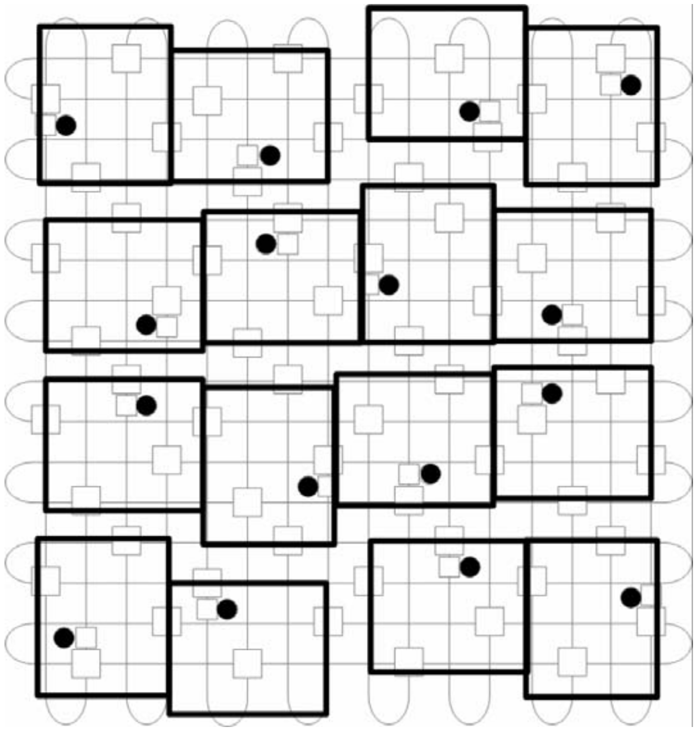
Design of strictly non-blocking photonic mesh



Non-blocking 4x4 → enables non-blocking mesh topology

Network is strictly nonblocking (derived from crossbar)

Link bidirectionality is exploited

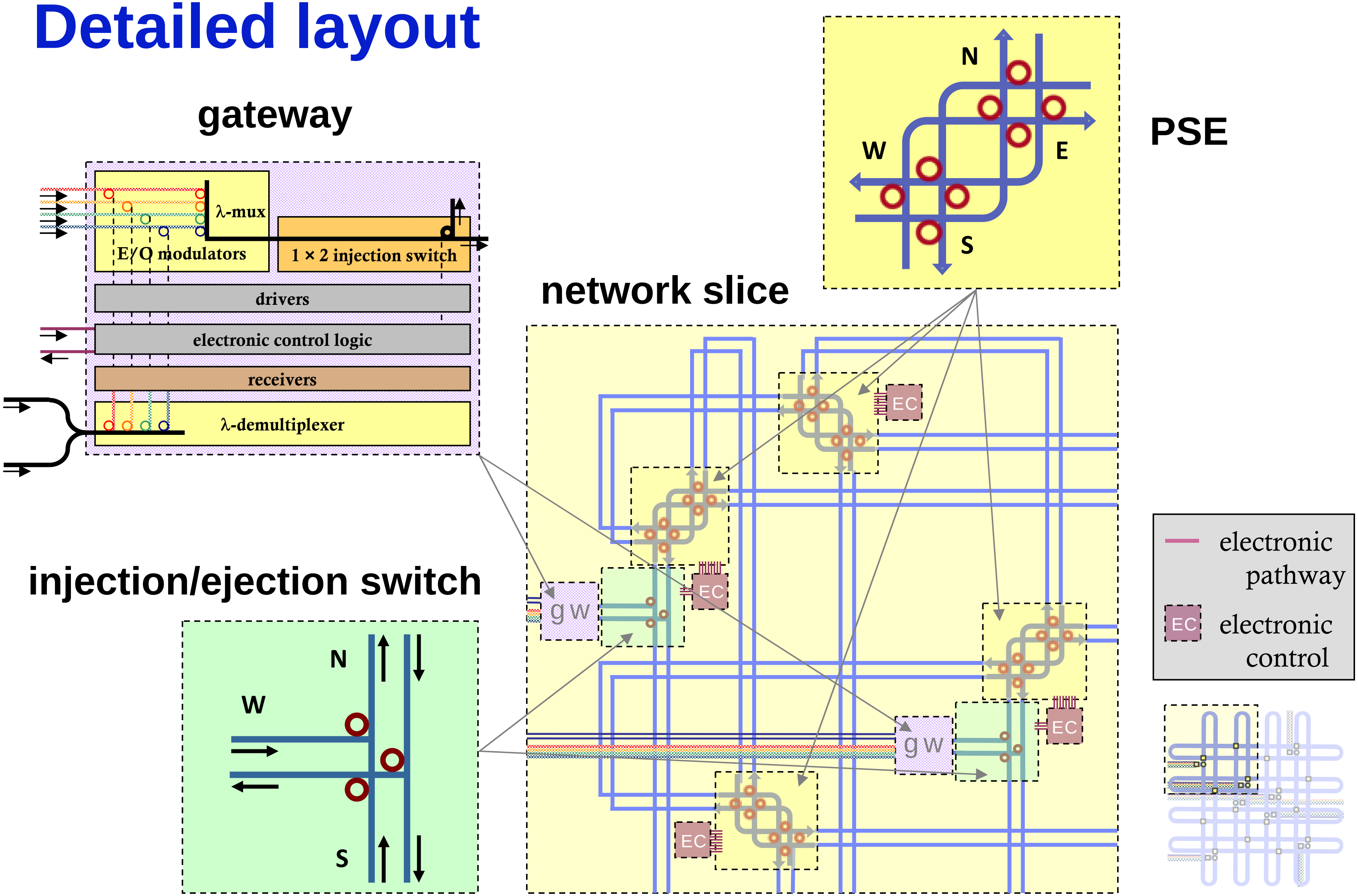


Processor Layout

Allow 2 gateways to inject on each row

Allow 2 gateways eject on each column

Detailed layout



gateway

λ-mux

E/O modulators

1 × 2 injection switch

drivers

electronic control logic

receivers

λ-demultiplexer

PSE

N

W

E

S

network slice

EC

gw

injection/ejection switch

N

W

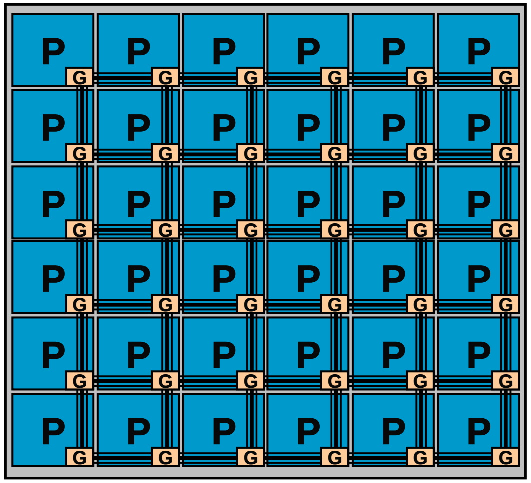
S

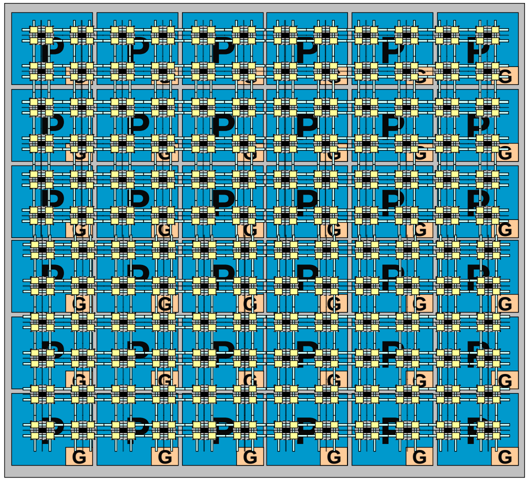
electronic pathway

EC electronic control

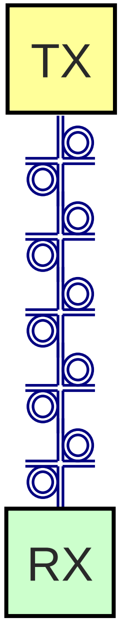
Comparative Power Analysis [DAC ’07]

- 6x6 tiled CMP
- Very large bandwidths per core
  - Peak: 800 Gb/s
  - Average: 512 Gb/s
- Compared designs
  - Electronic on-chip network
  - Hybrid photonic on-chip network
- Performance per Watt

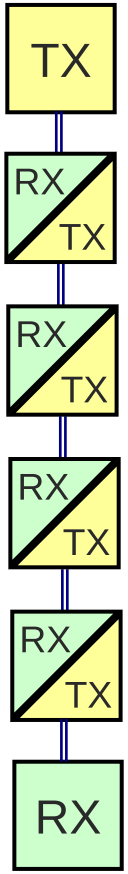




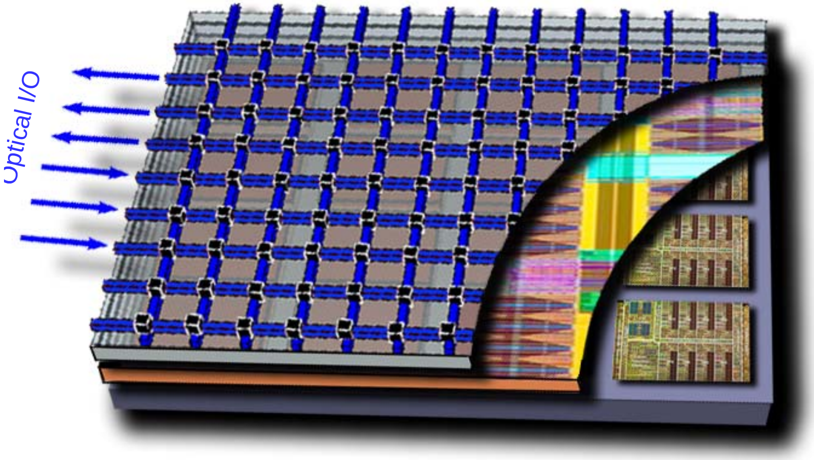
Power Analysis Results [DAC '07]



- Electronic NoC
  - Copper lines are bandwidth-limited
  - Parallelism used to attain large bandwidth
  - Wide busses and large buffers are **power hungry**
  - Multiple hops require **regeneration**
  - **NoC power exceeding 100 W** (prediction for 22 nm)
- Photonic NoC
  - Message generation: 2.3 W (assuming 0.11 pJ/bit)
  - Photonic switching: 0.04 W – practically negligible
  - Network control: 0.8 W (and scaling down with technology)
  - **Total – 3.2 W**
  - optical I/O **off-chip** with **same bandwidth** to external memory at very little additional power.



Performance Analysis

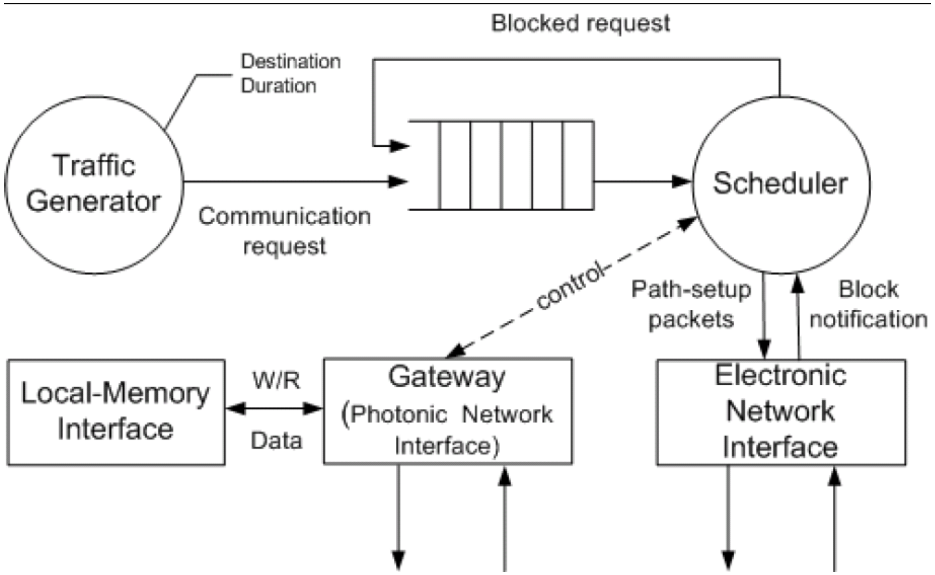


- Goal to evaluate performance-per-Watt advantage of CMP system with photonic NoC
- Developed network simulator using OMNeT++: modular, open-source, event-driven simulation environment
  - Modules for photonic building blocks, assembled in network
  - Multithreaded model for complex cores
- Evaluate NoC performance under uniform random distribution
- Performance-per-Watt gains of photonic NoC on FFT application

Multithreaded complex core model

- Model complex core as multithreaded processor with many computational threads executed in parallel
- Each thread independently make a communications request to any core

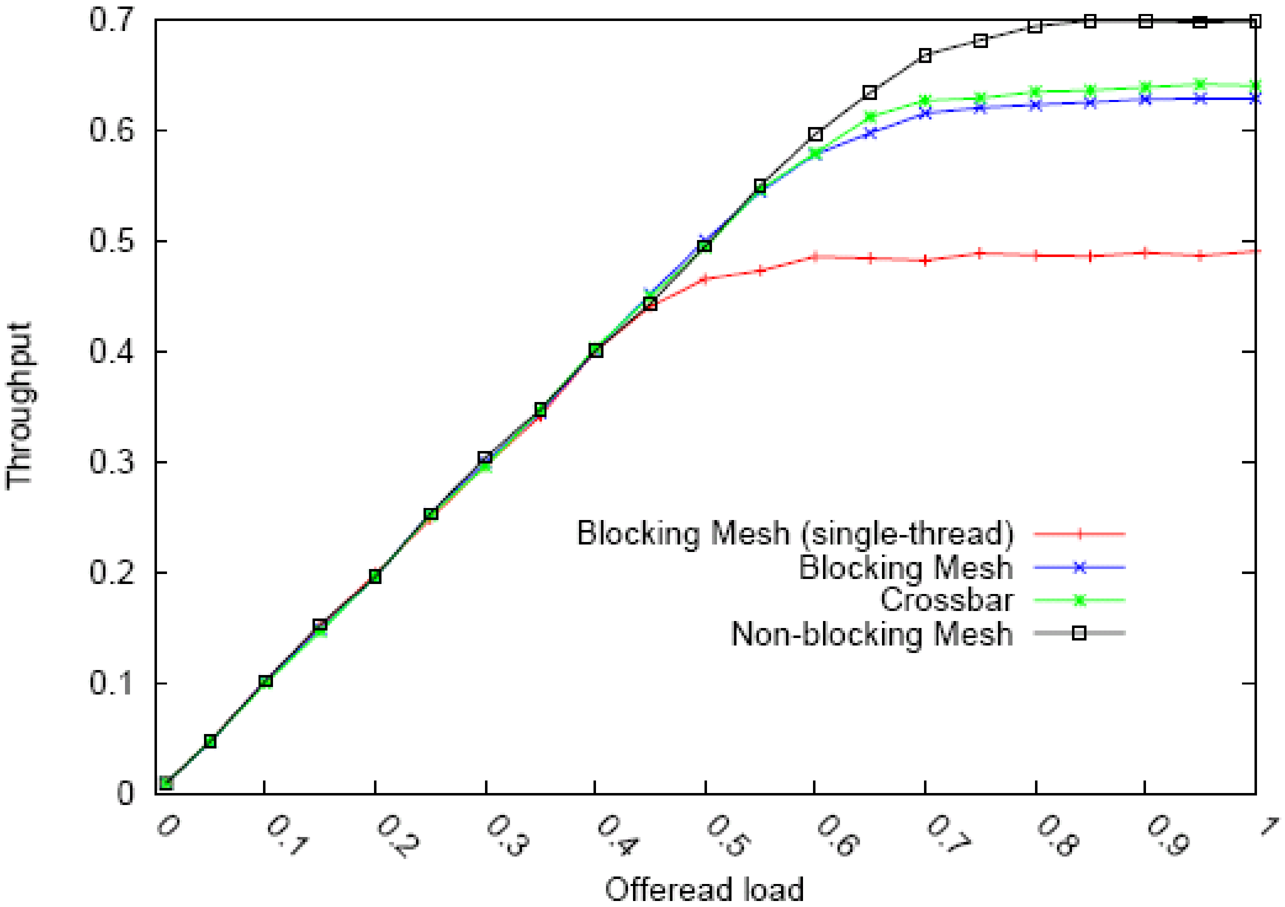
- Three main blocks:
  - Traffic generator – simulates core threads data transfer requests, requests stored in back-pressure FIFO queue
  - Scheduler – extracts requests from FIFO, generates path setup, electronic interface, blocked requests re-queued, avoids HoL blocking
  - Gateway – photonic interface, send/receive, read/write data to local memory



Throughput per core

- Throughput-per-core = ratio of time core transmits photonic message over total simulation time
  - Metric of average path setup time
  - Function of message length and network topology
- Offered load → considered when core is ready to transmit
- For uncongested network: throughput-per-core = offered load
- Simulation system parameters:
  - 36 multithreaded cores
  - DMA transfers of fixed size messages, 16kB
  - Line rate = 960Gbps; Photonic message = 134ns

Throughput per core for 36-node photonic NoC



Multithreading enables better exploitation of photonic NoC high BW

Gain of 26% over single-thread

Non-blocking mesh, shorter average path, improved by 13% over crossbar

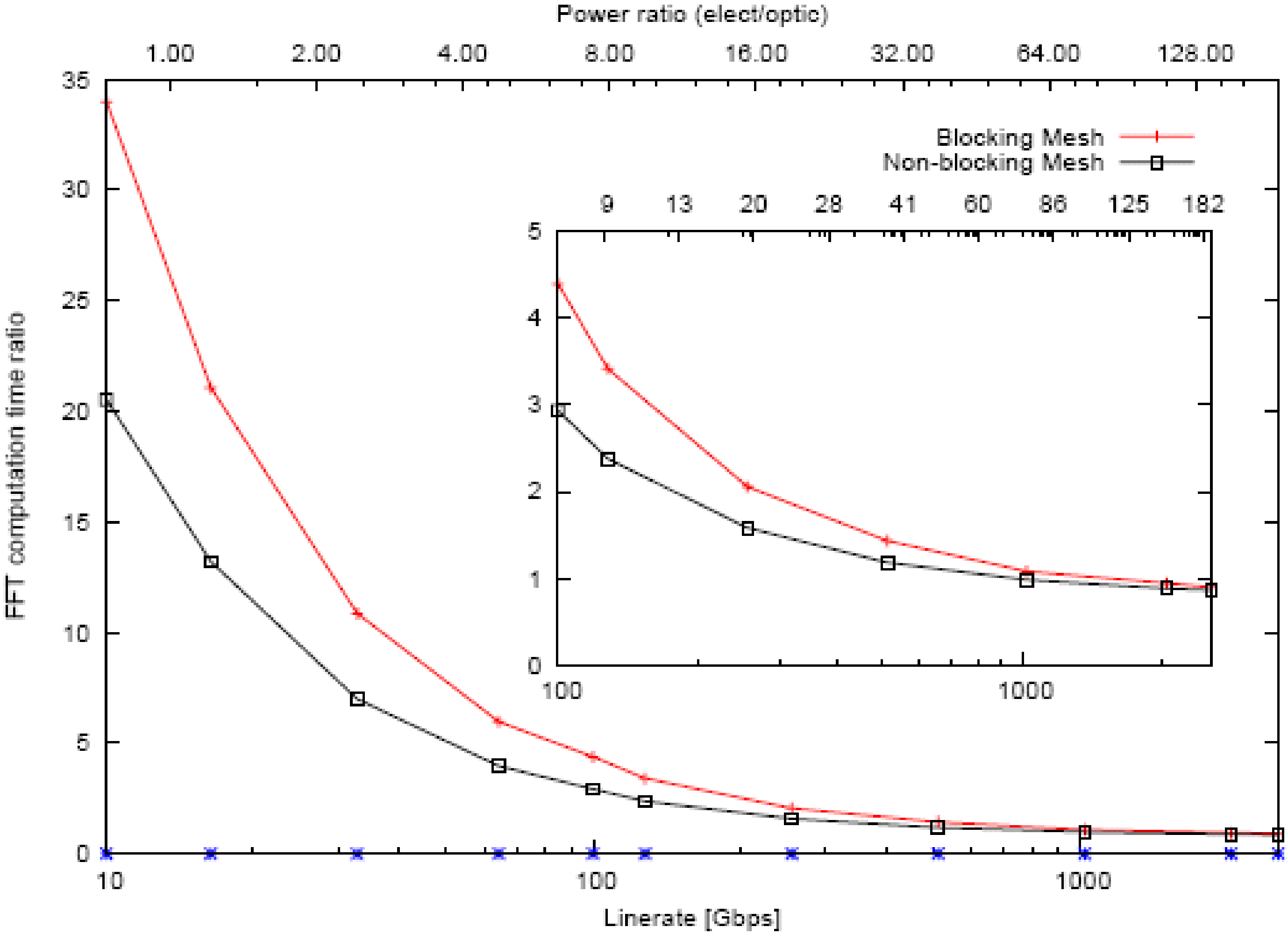
FFT Computation Performance

- We consider the execution of Cooley-Turkey FFT algorithm using 32 of 36 available cores
- First phase: each core processes: *k=m/M* sample elements
  - m = array size of input samples
  - M = number of cores
- After first phase, *log M* iterations of computation-step followed by communication-step when cores exchange data in butterfly
- Time to perform FFT computation depends on core architecture, time for data movement is function of NoC line rate and topology
- Reported results for FFT on Cell processor, 2^{24} samples FFT executes in ~43ms based on Bailey's algorithm.
- We assume Cell core with (2X) 256MB local-store memory, DP
- Use Bailey's algorithm to complete first phase of Cooley-Turkey in 43ms
- Cooley-Turkey requires *5k*Log*k* floating point operations, each iteration after first phase is ~1.8ms for k= 2^{24}
- Assuming 960Gbps, CMP non-blocking mesh NoC can execute 2^{29} in 66ms

FFT Computation Power Analysis

- For photonic NoC:
  - Hop between two switches is 2.78mm, with average path of 11 hops and 4 switch element turns
  - 32 blocks of 256MB and line rate of 960Gbps, each connection is 105.6mW at interfaces and 2mW in switch turns
  - total power dissipation is **3.44W**
- Electronic NoC:
  - Assume equivalent electronic circuit switched network
  - Power dissipated only for length of optimally repeated wire at 22nm, 0.26pJ/bit/mm
- Summary: Computation time is a function of the line rate, independent of medium

FFT Computation Performance Comparison



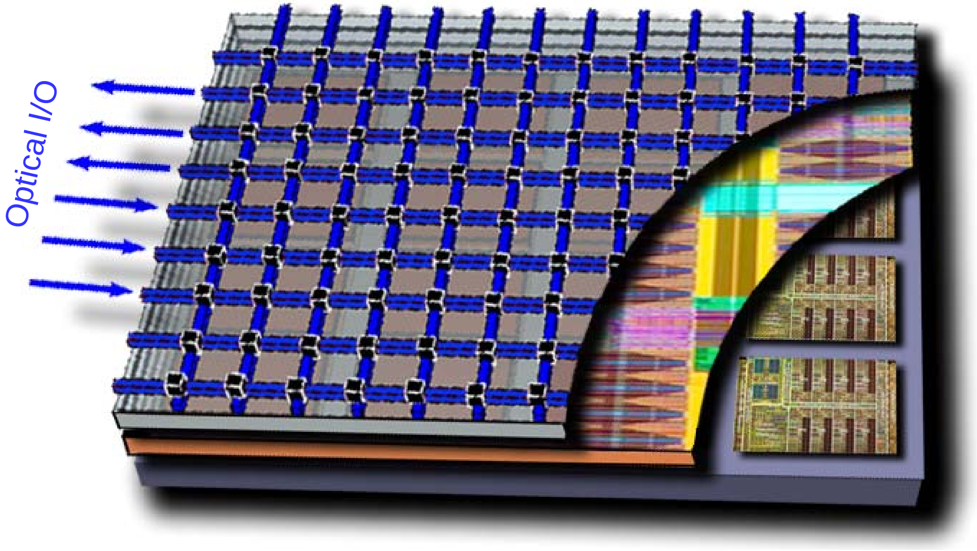
FFT computation: time ratio and power ratio as function of line rate

Performance-per-Watt

- To achieve same execution time (time ratio = 1), electronic NoC must operate at the same line rate of 960Gbps, dissipating 7.6W/connection or ~70X over photonic
- Total dissipated power is ~244W
- To achieve same power (power ratio = 1), electronic NoC must operate at line rate of 13.5Gbps, a reduction of 98.6%.
- Execution time will take ~1sec or 15X longer than photonic

Summary

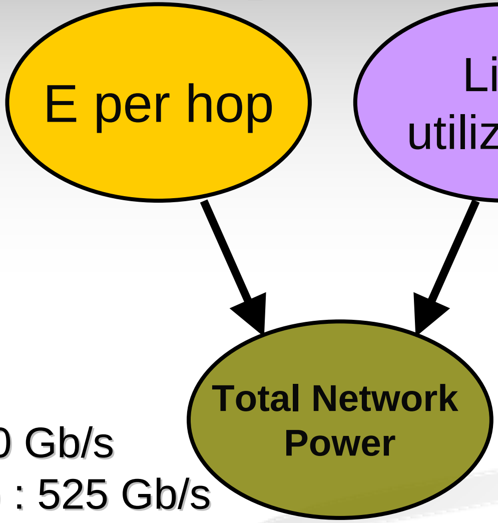
- CMPs are clearly emerging for power efficient high performance computing capability
- Future on-chip interconnects must provide large bandwidth to many cores



- Electronic NoCs dissipate prohibitively high power

  → a technology shift is required
- Remarkable advances in Silicon Nanophotonics
- Photonic NoCs provide enormous capacity at dramatically low power consumption required for future CMPs, **both on- and off-chip**
- Performance-per-Watt gains on communications intensive applications

Power Analysis:
Electronic On-chip Network

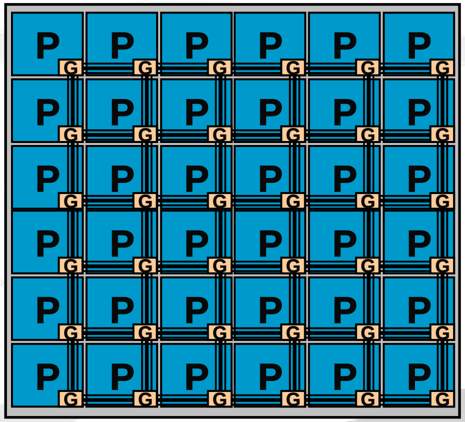


E per hop

Link utilization

Total Network Power

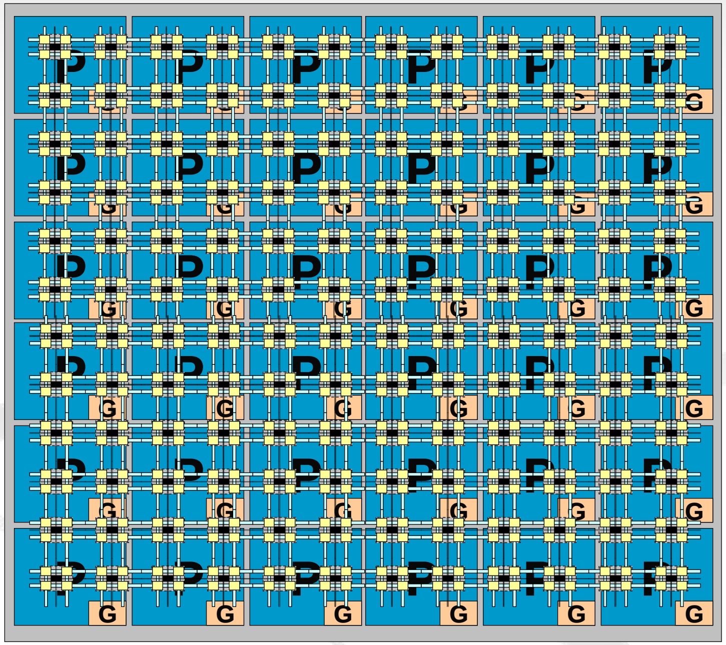
- Assumptions:
  - 6x6 Mesh, uniform traffic
  - Link length (l): 1.67 mm
  - Bus width (w): 168 bits
  - Signaling rate (f): 5 GHz
  - Injection rate (IR): 0.625
- Results:
  - Peak bandwidth ($BW_{PEAK}=w\cdot f$) : 840 Gb/s
  - Average bandwidth ($BW_{AVG}=w\cdot f\cdot IR$) : 525 Gb/s
- Link traversal energy:
  - E_{link} = 0. 34 pJ/bit/mm (estimated for 32 nm)
  - E_{router} = 0.83 pJ/bit (estimated for 32 nm)
  - $E_{flit\text{-}hop} = (E_{link}\cdot l+E_{link})\cdot w =$ 235 pJ
- 6x6 Mesh → 120 links
- Average link utilization (uniform traffic) = 0.75



Total network power = $U_{AVG}\cdot N_{LINKS}\cdot E_{flit\text{-}hop}\cdot f$ **= 106 W**

Power Analysis: (1) Photonic Network

- 6x6 CMP (36 Gateways)
- 12x12 Photonic mesh
- 960 Gb/s peak bandwidth
- Injection rate: 0.6
- Average BW: 576 Gb/s
- 4 turns per message
- 86 switches ***ON*** (~ 0.5 mW each)
- Network power: **43 mW**

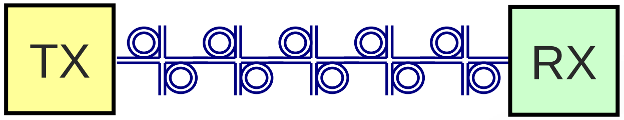


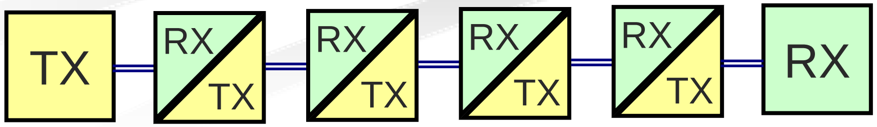
Power Analysis:
(2) Photonic Gateways

- Generating/receiving very high bandwidths is costly.

	Current	**Exp. scaling**
Modulation	~1 pJ/bit	0.1 pJ/bit
Detection	~0.1 pJ/bit	0.01 pJ/bit
Total		
(36 x 576 Gb/s)	**23 W**	**2.3 W**

- Comparable to **a single** electronic link
- But – need to modulate/detect only once, while routing is nearly free.

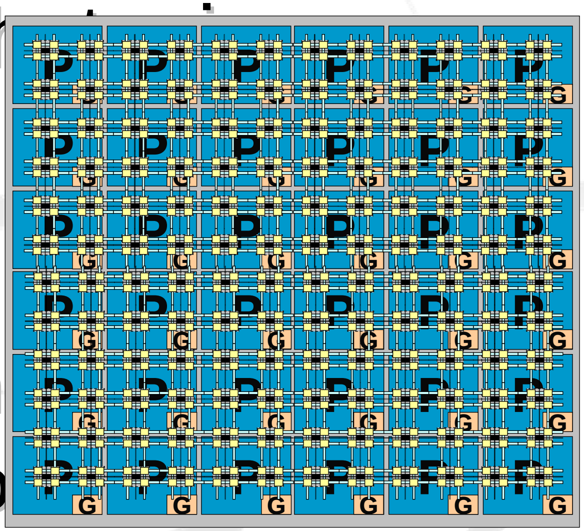




Power Analysis:
(3) Electronic Control Network

- Low bandwidth electronic NoC:
  Carries only control packets.

- Bulk of data transmitted on ph
  network



- Assumptions
  - x2 path length (overprovisionin
  - 64 control bits per 2-KByte pho
    message

- Carries only 0.8% of the traffic