Optical Interconnection Networks for Scalable High-performance Parallel Computing Systems

Ahmed Louri

Department of Electrical and Computer Engineering
University of Arizona, Tucson, AZ 85721

louri@ece.arizona.edu

Optical Interconnects Workshop for High Performance Computing

Oak Ridge, Tennessee, November 8-9, 1999

Talk Outline

- **Need for Scalable Parallel Computing Systems**
- **Scalability Requirements**
- **Current Architectural Trends for Scalability**
- **Fundamental Problems facing Current Trends**
- **Optics for Scalable Systems**
- **Proposed Optical Interconnection Architectures for DSMs, and Multicomputers.**
- **Conclusions**

Need for Scalable Systems

- **Market demands in terms of lower computing costs and protection of customer investment in computing: scaling up the system to quickly meet business growth is obviously a better way of protecting investment: hardware, software, and human resources.**
- **Applications: explosive growth in internet and intranet use.**
- **The quest for higher performance in many scientific computing applications: an urgent need for Teraflops machines!!**
- **Performance that holds up across machine sizes and problem sizes for a wide class of users sells computers in the long run.**

Scalability Requirements

- **A scalable system should be incrementally expanded, delivering linear incremental performance with a near linear cost increase, and with minimal system redesign (size scalability), additionally,**
- **it should be able to use successive, faster processors with minimal additional costs and redesign (generation scalability).**
- **On the architecture side, the key design element is the interconnection network!**

Problem Statement

- **The interconnection network must be able to : (1) increase in size using few building blocks and with minimum redesign, (2) deliver a bandwidth that grows linearly with the increase in system size, (3) maintain a low or (constant) latency, (4) incur linear cost increase, and (5) readily support the use of new faster processors.**

- **The major problem is the ever-increasing speed of the processors themselves and the growing performance gap between processor technology and interconnect technology.**
  - **Increased CPU speeds (today in the 600 MHz, tomorrow 1 GHz)**
  - **Increased CPU-level parallelism (multithreading etc.)**
  - **Effectiveness of memory latency-tolerating techniques. These techniques demand much more bandwidth than needed.**

- **Need for much more bandwidth (both memory and communication bandwidths)**

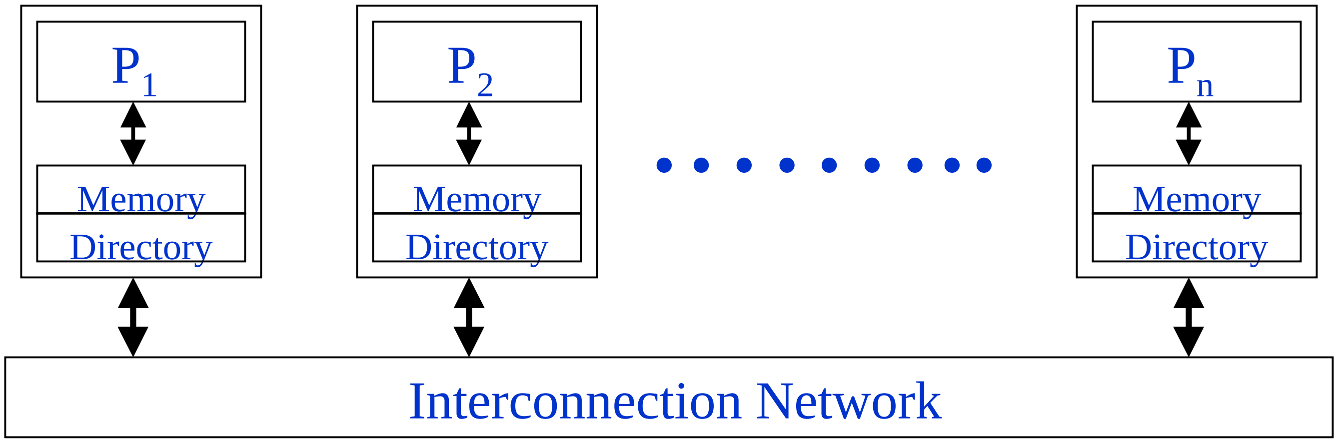
Current Architectures for Scalable Parallel Computing Systems

- **SMPs: bus-based symmetric multiprocessors: a global physical address space for memory and uniform, symmetric access to the entire memory (small scale systems, 8 - 64 processors)**

- **DSMs: distributed-shared memory systems: memory physically distributed but logically shared. (medium-scale 32 - 512 processors)**

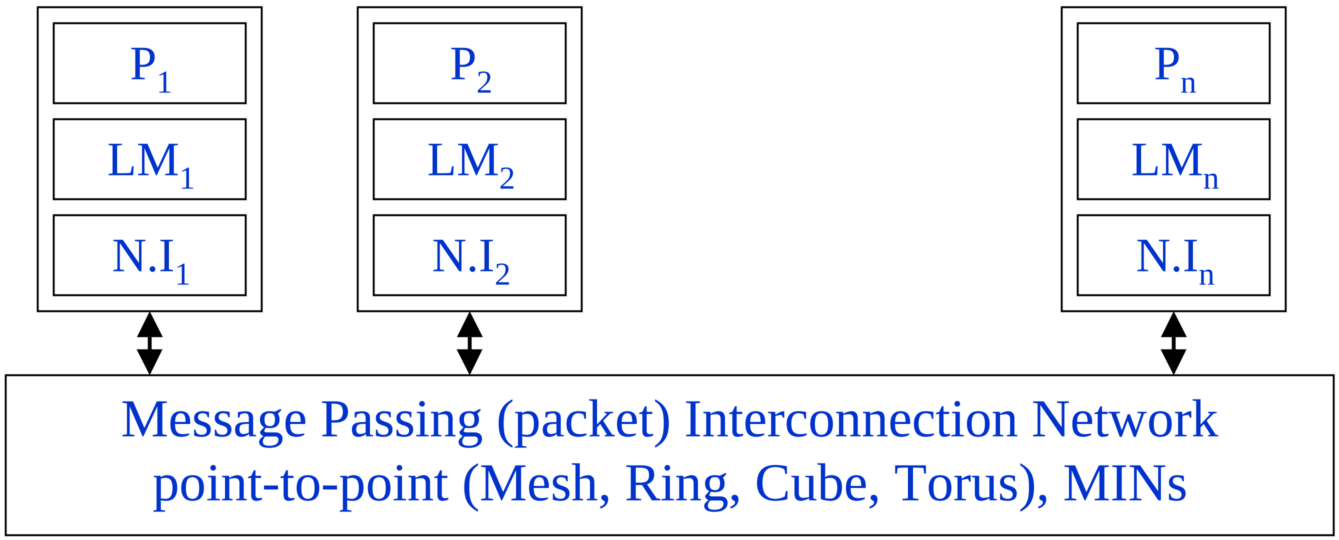
- **Message-Passing systems: private distributed memory. (greater than 1000 processors)**

Distributed Shared-Memory Systems



- **Memory physically distributed but logically shared by all processors.**
- **Communications are via the shared memory only.**
- **Combines programming advantages of shared-memory with scalability advantages of message passing. Examples: SGI Origin 2000, Stanford Dash, Sequent, Convex Exemplar, etc.**

No Remote Memory Access (NORMA) Message-Passing Model



- **Interprocessor communication is via message-passing mechanism**
- **Private memory for each processor (not accessible by any other processor)**
  - **Examples: Intel Hypercube, Intel Paragon, TFLOPS, IBM SP-1/2, etc.**

Fundamental Problems facing DSMs

- **Providing a global shared view on a physically distributed memory places a heavy burden on the interconnection network.**
- **Bandwidth to remote memory is often non-uniform and substantially degraded by network traffic.**
- **Long average latency: latency in accessing local memory is much shorter than remote accesses.**
- **Maintaining data consistency (cache coherence) throughout the entire system is very time-consuming.**

An Optical Solution to DSMs

- **If a low-latency interconnection network could provide a (1) near-uniform access time, and (2) high-bandwidth access to all memories in the system, whether local or remote, the DSM architecture will provide a significant increase in programmability, scalability and portability of shared-memory applications.**

- **Optical Interconnects can play a pivotal role in such an interconnection network.**

Fundamental Problems facing Current Interconnect Technology

- **Chip power and area increasingly dominated by interconnect drivers, receivers, and pads**
- **Power dissipation of off-chip line drivers**
- **Signal distortion due to interconnection attenuation that varies with frequency**
- **Signal distortion due to capacitive and inductive crosstalks from signals of neighboring traces**
- **Wave reflections**
- **Impedance matching problems**
- **High sensitivity to electromagnetic interference (EMI)**
- **Electrical isolation**
- **Bandwidth limits of lines**
- **Clock skew**
- **Bandwidth gap: high disparity between processor bandwidth and memory bandwidth, and the problem is going to be much worse in future**
  - **CPU - Main memory traffic will require 10s of GB/s rate**
- **Limited speed of off-chip interconnects**

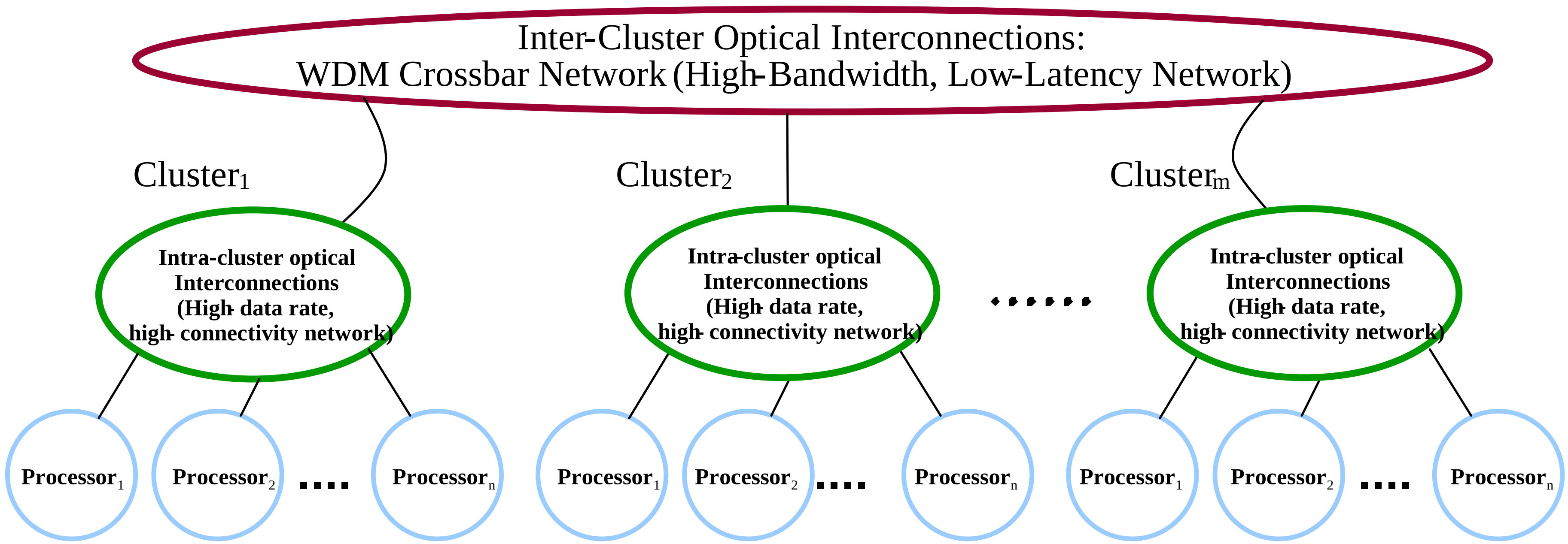
Optics for Interconnect

- Higher interconnection densities (parallelism)
- Higher packing densities of gates on integrated chips
- Fundamentally lower communication energy than electronics
- Greater immunity to EMI
- Less signal distortion
- Easier impedance matching using antireflection coatings
- Higher interconnection bandwidth
- Lower signal and clock skew
- Better electrical isolation
- No frequency-dependent or distance-dependent losses
- Potential to provide interconnects that scale with the operating speed of performing logic

SOCN for High Performance Parallel Computing Systems

- **SOCN stands for "Scalable Optical Crossbar-Connected Interconnection Networks".**
- **A two-level hierarchical network.**
- **The lowest level consists of clusters of *n* processors connected via local WDM intra-cluster all-optical crossbar subnetwork.**
- **Multiple (*c) clusters* are connected via similar WDM intra-cluster all-optical crossbar that connects all processors in a single cluster to all processors in a remote cluster.**
- **The inter-cluster crossbar connections can be rearranged to form various network topologies.**

The SOCN Architecture



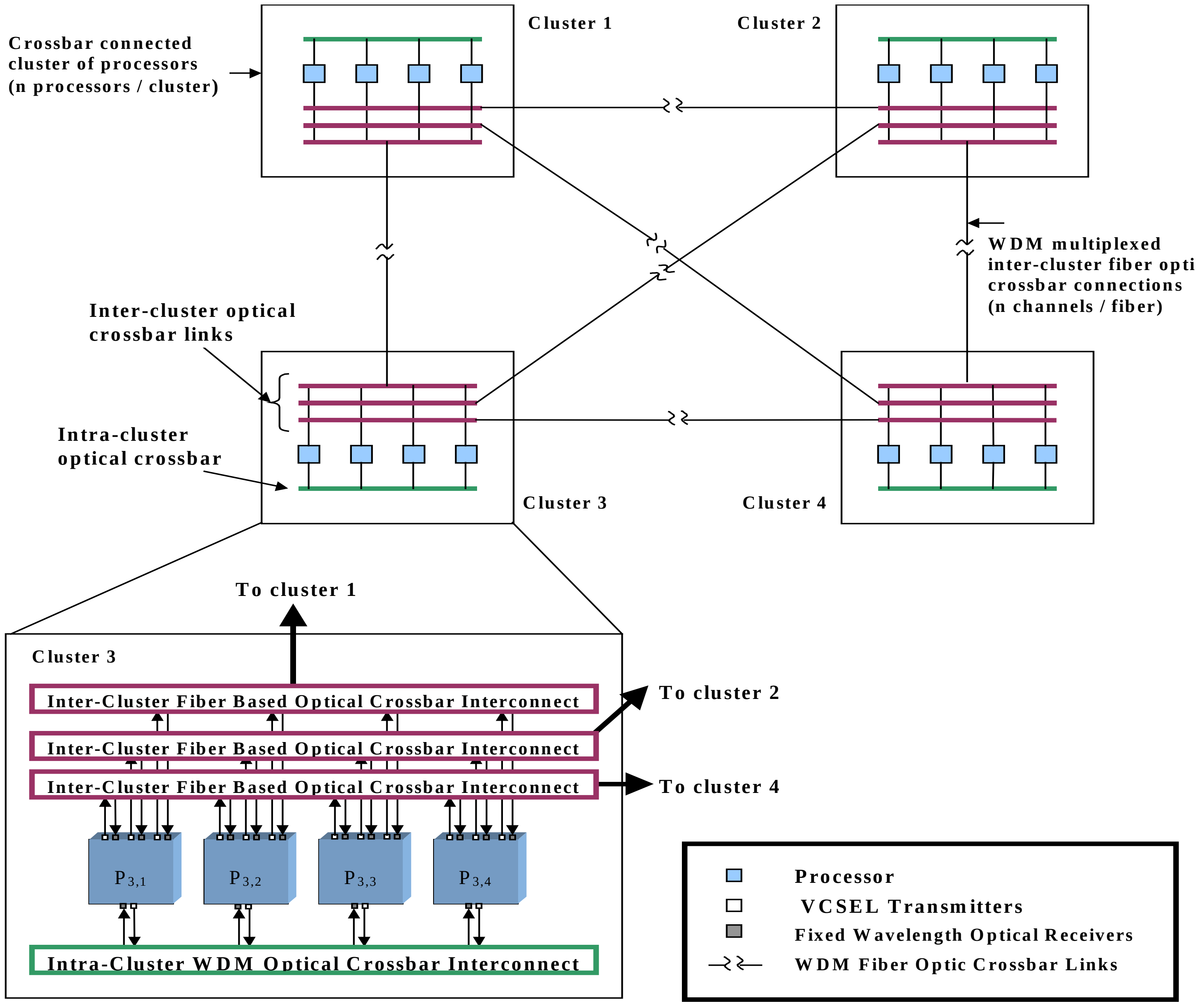
Both the intra-cluster and inter-cluster subnetworks are WDM-based optical crossbar interconnects.

Architecture based on wavelength reuse.

Crossbar Networks

- The SOCN class of networks are based on WDM all-optical crossbar networks.
- Benefits of crossbar networks:
  - —Fully connected.
  - —Minimum potential latency.
  - —Highest potential bisection bandwidth.
  - —Can be used as a basis for multi-stage and hierarchical networks.
- Disadvantages of crossbar networks:
  - —$O(N^2)$ Complexity.
  - —Difficult to implement in electronics.
    - N^2 wires and switches required.
    - Rise-time and timing skew become a limitation for large crossbar interconnects.
- Optics and WDM can be used to implement a crossbar with O(N) complexity.

Example OC^3N



Optical Crossbar-Connected Cluster Network (OC^3N) Benefits

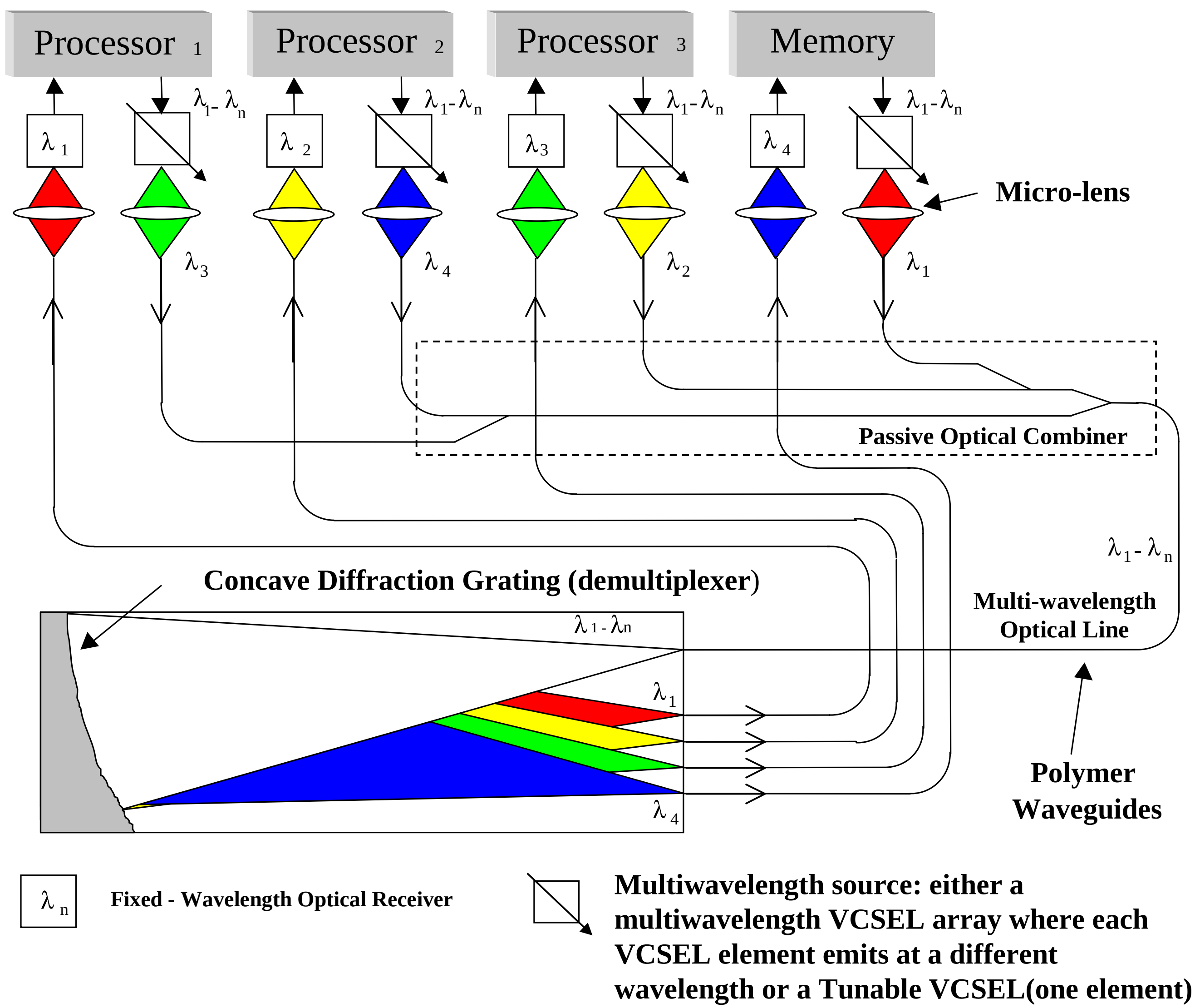
- **Every cluster is connected to every other cluster via a single send/receive optical fiber pair.**
- **Each optical fiber pair supports a wavelength division multiplexed fully-connected crossbar interconnect.**
- **Full connectivity is provided: every processor in the system is directly connected to every other processor with a relatively simple design.**
- **Inter-cluster bandwidth and latencies similar to intra-cluster bandwidth and latencies!**
- **Far fewer connections are required compared to a traditional crossbar.**
  - **Example: A system containing *n*=16 processors per cluster and *c*=16 clusters (N=256) requires 120 inter-cluster fiber pairs, whereas a traditional crossbar would require 32,640 interprocessor connections.**

OC^3N Scalability

- **The OC^3N topology efficiently utilizes wavelength division multiplexing throughout the network, so it could be used to construct relatively large (hundreds of processors) fully connected networks with a reasonable cost.**
- **# nodes** $N = n \times c$
- **Degree** $D_C = c = \frac{N}{n}$
- **Diameter** $K_C = 1$
- **# links** $L_C = (N^2/n^2 - N/n)/2$
- **Bisection width** $B_C = N^2/4 = (n \times c)^2/4,$
- **Avg. Message Dist.** $\bar{l}_C = 1$

Intra-Cluster WDM Optical Crossbar

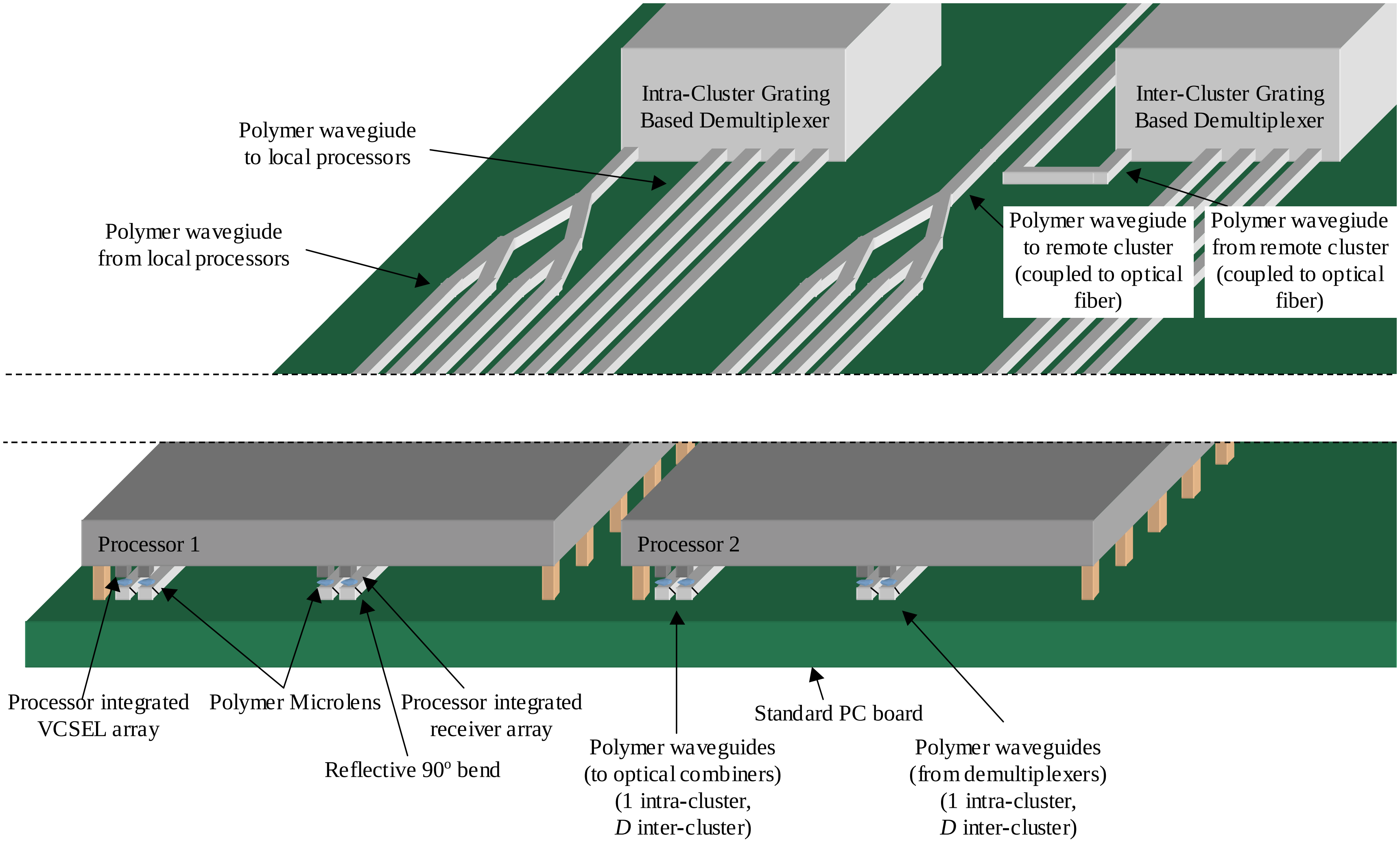
Applied Optics, vol. 38, no. 29, pp. 6176 - 6183, Oct. 10, 1999



WDM Optical Crossbar Implementation

- **Each processor contains a single integrated tunable VCSEL or a VCSEL array, and one optical receiver.**
- **Each VCSEL is coupled into a PC board integrated polymer waveguide.**
- **The waveguides from all processors in a cluster are routed to a polymer waveguide based optical binary tree combiner.**
- **The combined optical signal is routed to a free-space diffraction grating based optical demultiplexer.**
- **The demultiplexed optical signals are routed back to the appropriate processors.**

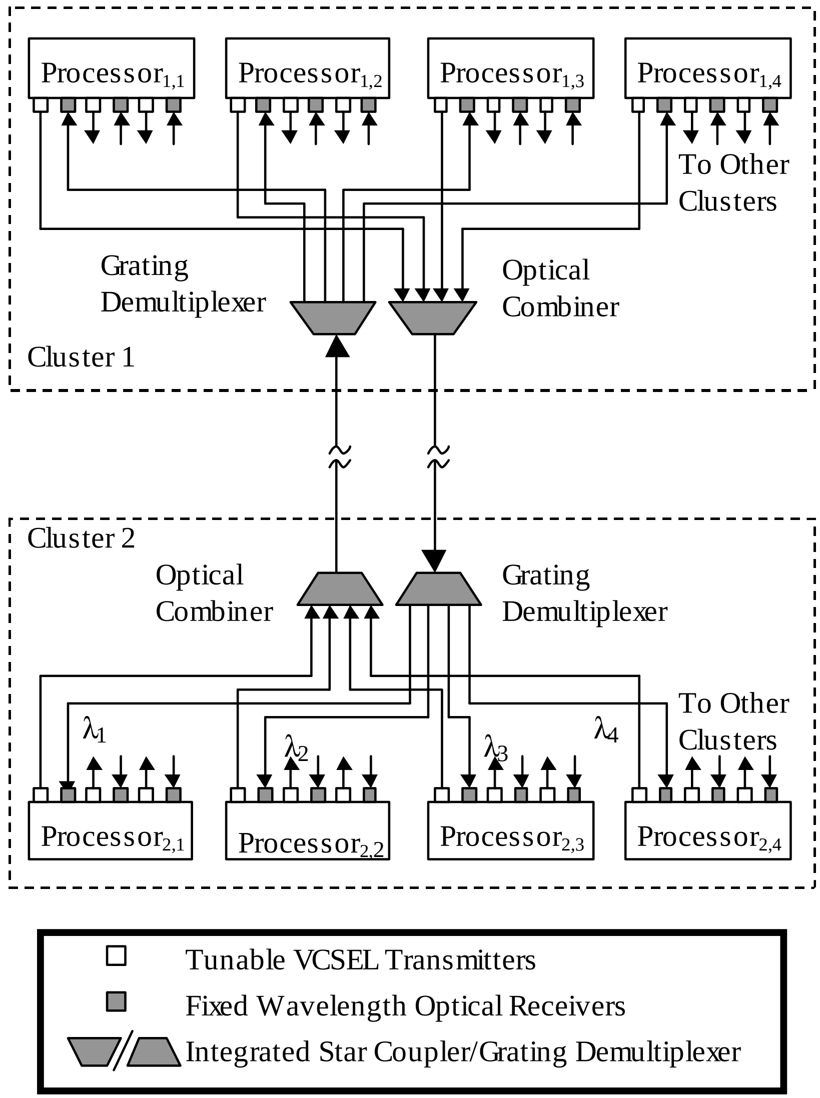
Polymer Waveguide Implementation



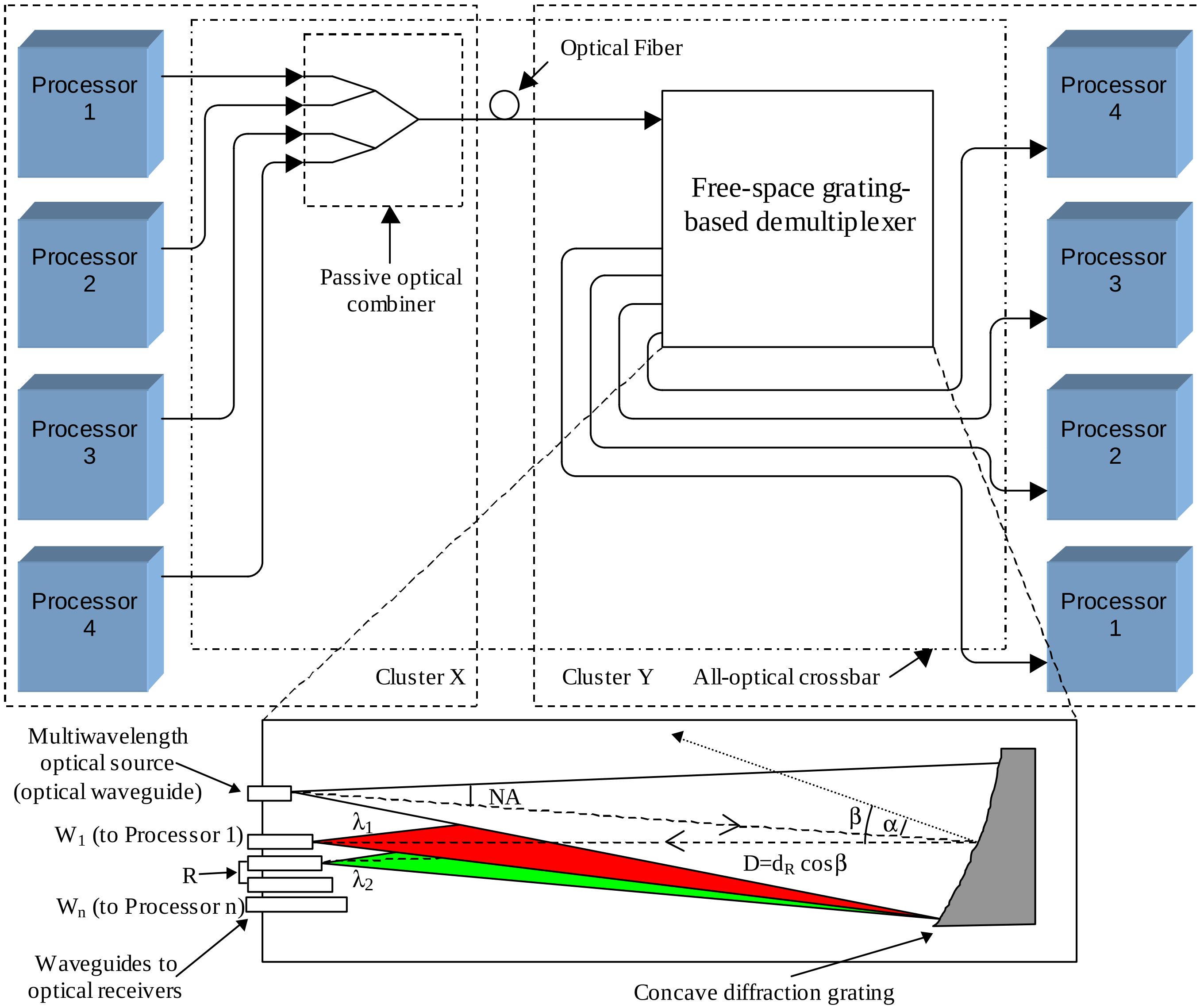
Inter-Cluster WDM Optical Crossbar

- **Inter-cluster interconnects utilize wavelength reuse to extend the size of the optical crossbars to support more processors than the number of wavelengths available.**
- **An additional tunable VCSEL and receiver are added to each processor for each inter-cluster crossbar.**
- **The inter-cluster crossbars are very similar to the intra-cluster crossbars with the addition of an optical fiber between the optical combiner an the grating demultiplexer. This optical fiber extends the crossbar to the remote cluster.**

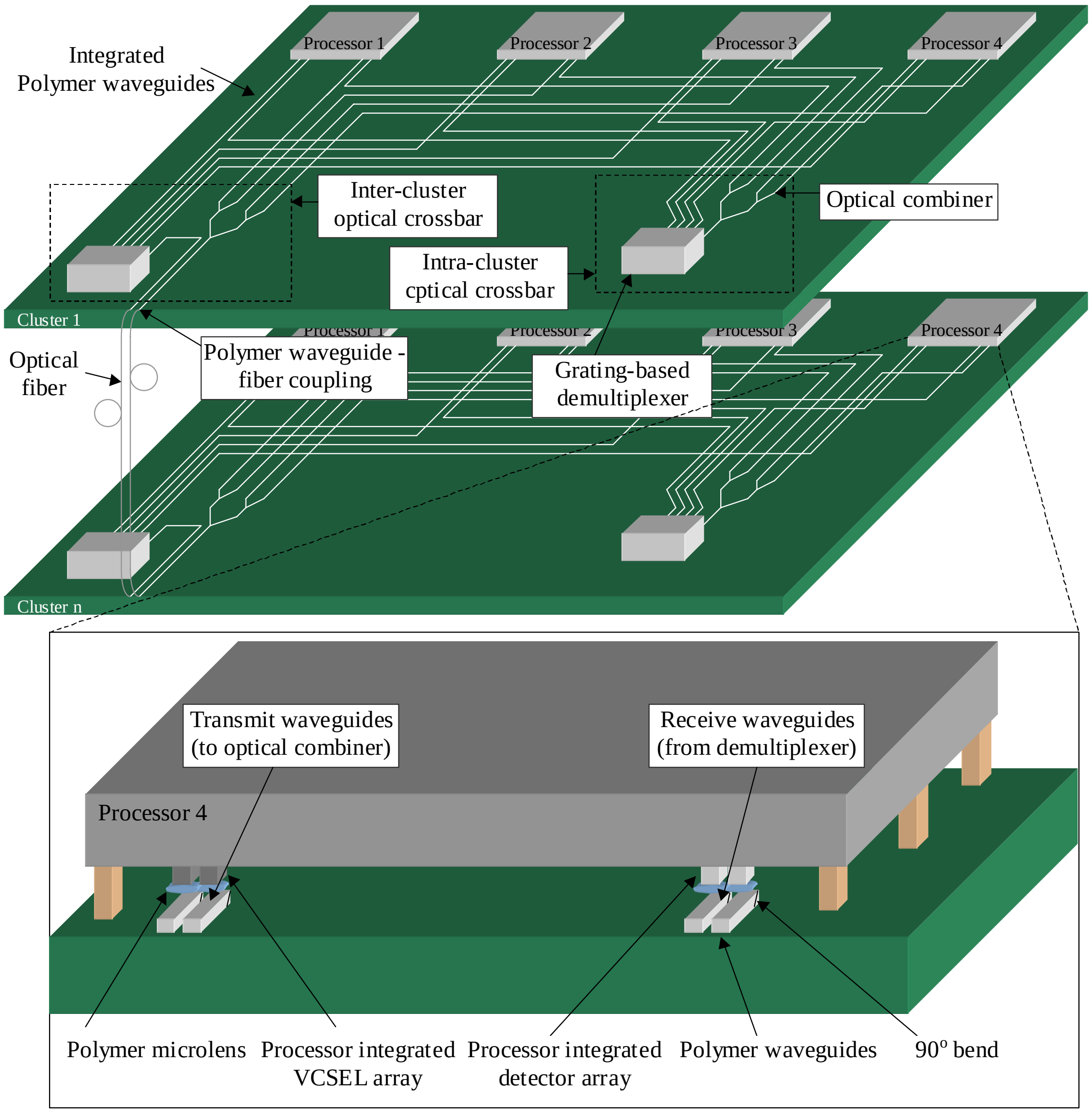
Inter-Cluster Crossbar Overview



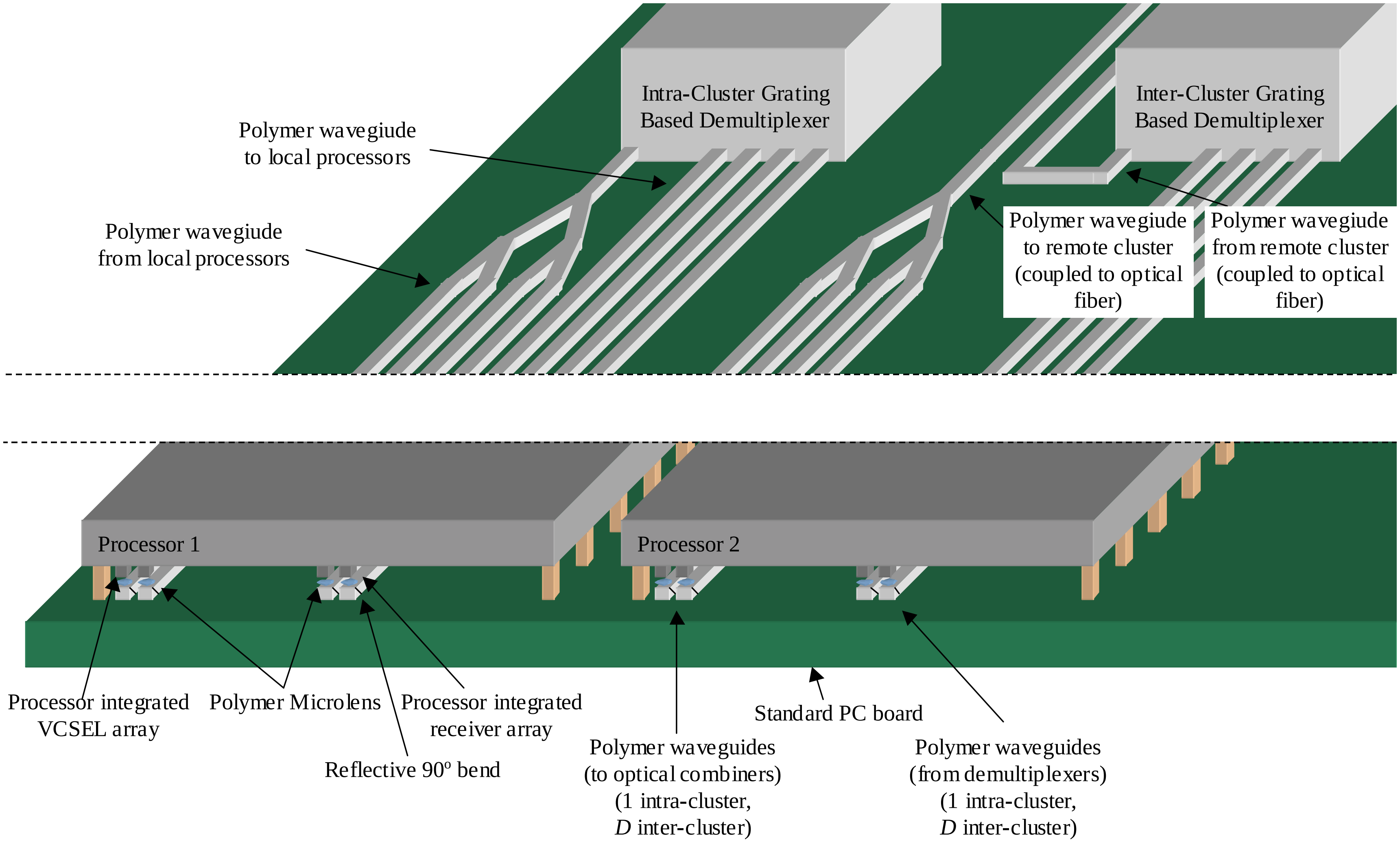
Inter-Cluster WDM Optical Crossbar



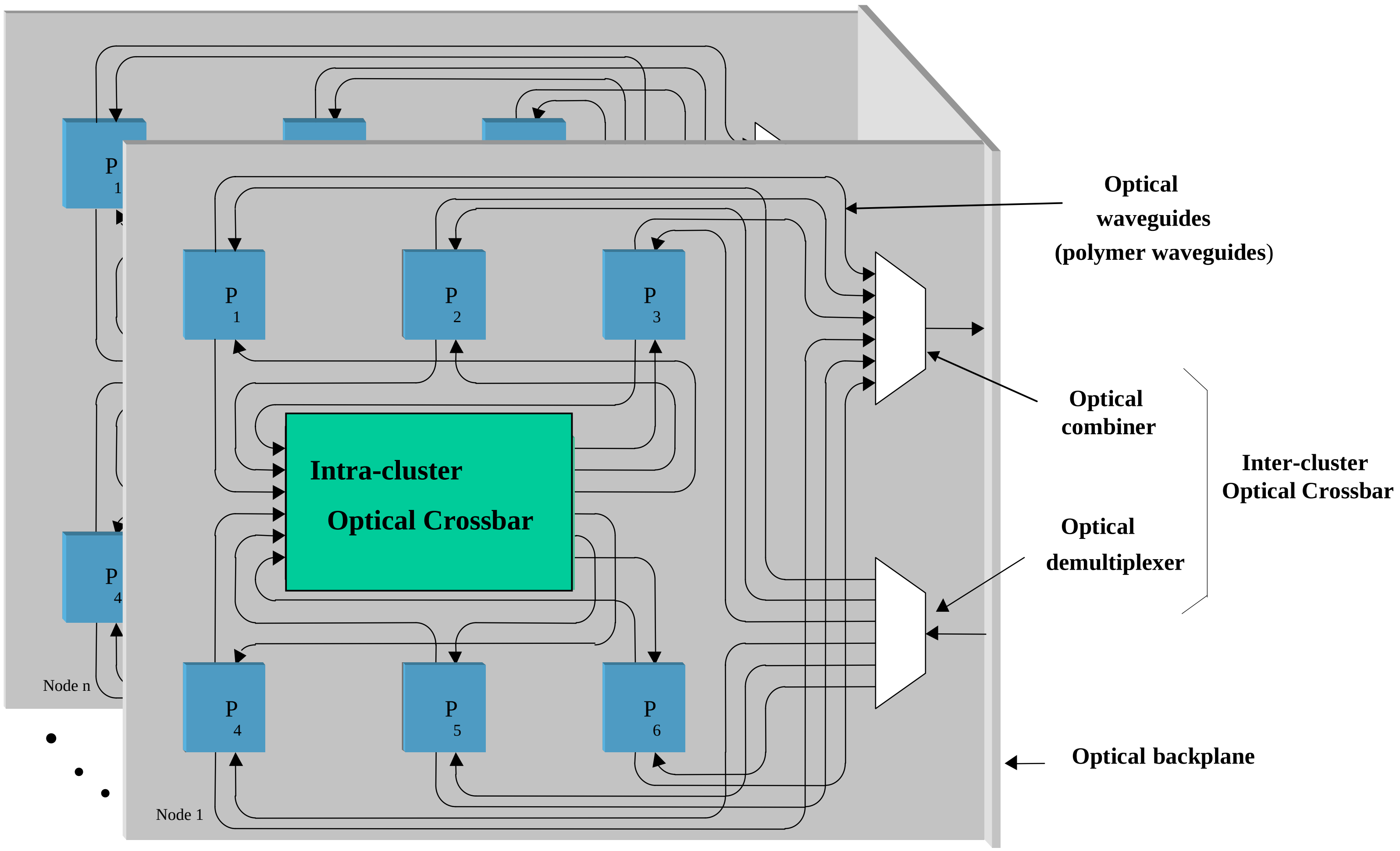
Possible Implementation



Polymer Waveguide Implementation



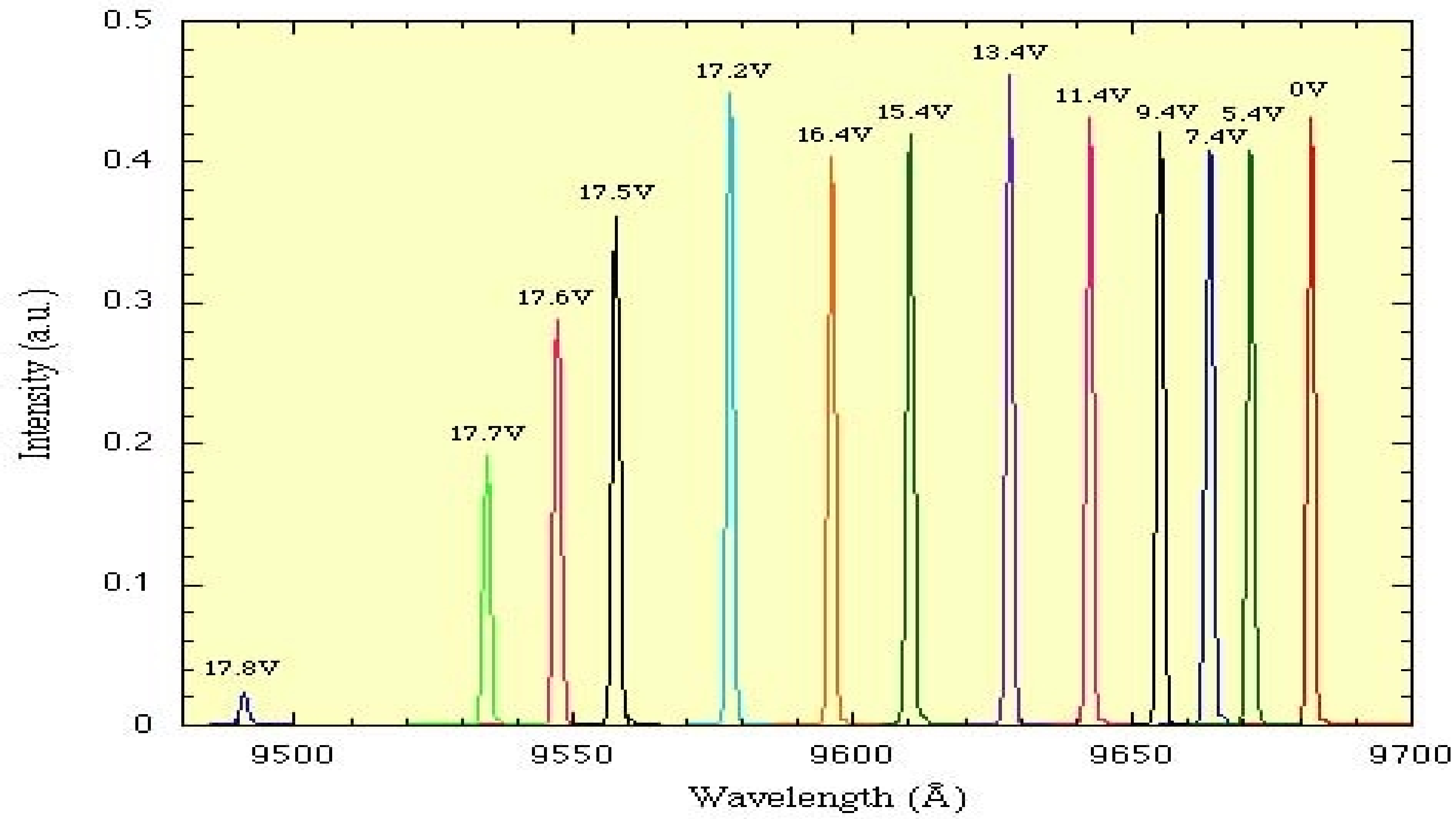
Overview of an Optical Implementation of SOCN



Emerging Optical Technologies which make SOCN a viable option

- **VCSELs (including tunable ones).**
  - —**Enable wavelength division multiplexing (WDM).**
  - —**Up to ~32nm tuning range around 960nm currently available.**
  - —**Tuning speeds in the MHz range.**
  - —**Very small (few hundred μm in diameter).**
- **Polymer waveguides.**
  - —**Very compact (2-200 μm in diameter).**
  - —**Densely packed (10 μm waveguide separation).**
  - —**Can be fabricated relatively easily and inexpensively directly on IC or PC board substrates.**
  - —**Can be used to fabricate various standard optical components (splitters, combiners, diffraction gratings, couplers, etc.)**

Tunable VCSELs



Source: "Micromachined Tunable Vertical Cavity Surface Emitting Lasers," Fred Sugihwo, et al., *Proceedings of International Electron Device Meetings*, 1996.

Existing Optical Parallel Links based on VCSELs and Edge Emitting Lasers

	Fiber	Detector	Emitter	Data rate	Capacity
SPIBOC	SM	PIN	12 edge	2.5 Gb/s	30 Gb/s
OETC	MM	MSM	32 VCSEL	500 Mb/s	16 Gb/s
POINT	MM	-	32 VCSEL	500 Mb/s	16 Gb/s
NTT	MM	PIN	5 edge	2.8 Gb/s	14 Gb/s
Siemens	MM	PIN	12 edge	1 Gb/s	12 Gb/s
Fujitsu	SM	PIN	20 edge	622 Mb/s	12 Gb/s
Optobahn 2	MM	PIN	10 edge	1 Gb/s	10 Gb/s
Jitney	MM	-	20	500 Mb/s	10 Gb/s
POLO	MM	PIN	10 VCSEL	800 Mb/s	8 Gb/s
Optobus II	MM	PIN	10 VCSEL	800 Mb/s	8 Gb/s
P-VixeLink	MM	MSM	12 VCSEL	625 Mb/s	7.5 Gb/s
NEC	MM	-	6 edge	1.1 Gb/s	6.6 Gb/s
ARPA TRP	SM	-	4 edge	1.1 Gb/s	4.4 Gb/s
Oki	MM	-	12 edge	311 Mb/s	3.7 Gb/s
Hitachi	SM	PIN	12 edge	250 Mb/s	3 Gb/s

Ref: F. Tooley, "Optically interconnected electronics: challenges and choices," in Proc. Int'l. Workshop on Massively Parallel Processing Using Optical Interconnections, (Maui Hawaii), pp. 138-145, Oct. 1996

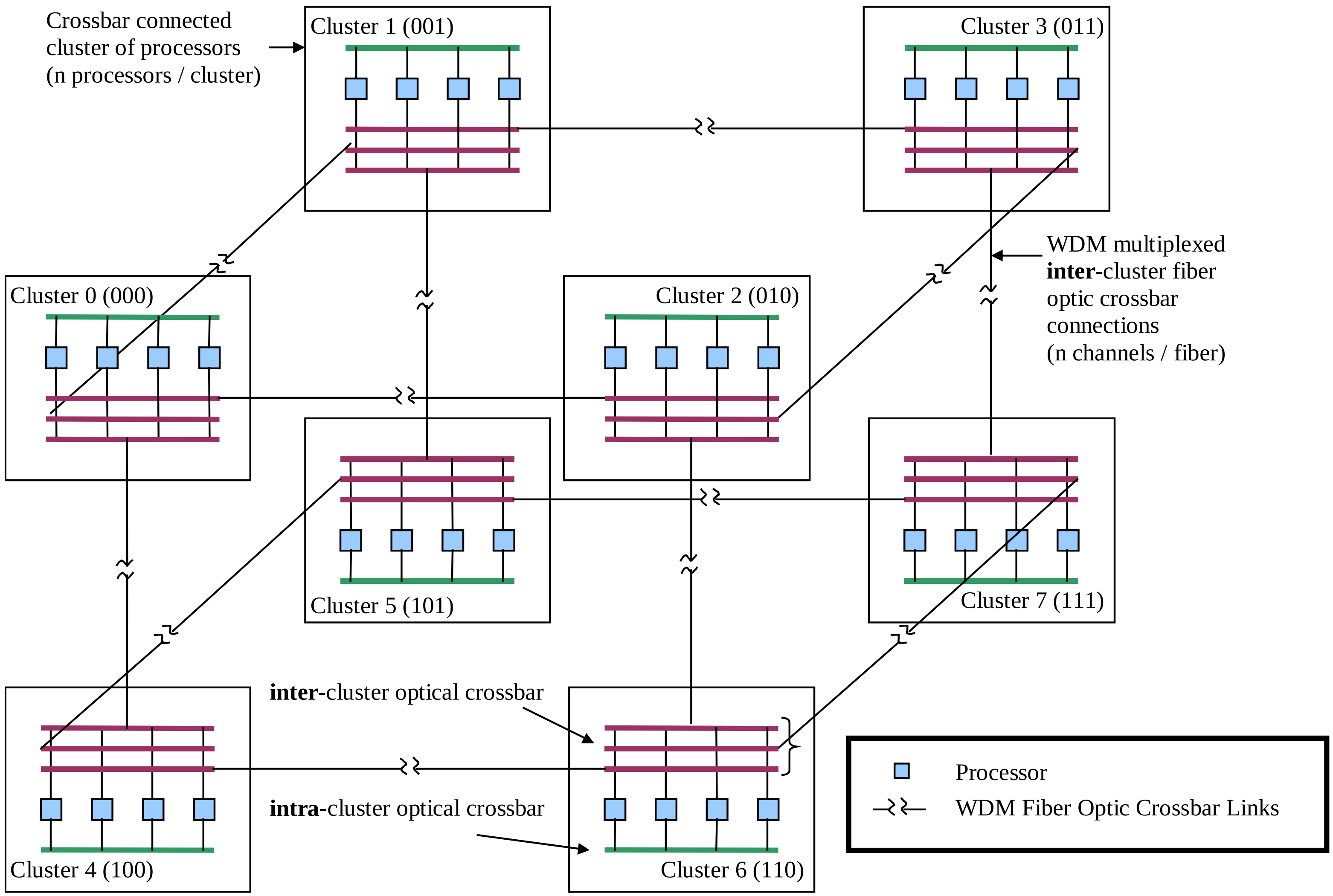
Architectural Alternatives

- **One of the advantages of a hierarchical network architecture is that the various topological layers typically can be interchanged without effecting the other layers.**
- **The lowest level of the SOCN is a fully connected crossbar.**
- **The second (and highest) level can be interchanged with various alternative topologies as long as the degree of the topology is less than or equal to the cluster node degree.**
  - **Crossbar**
  - **Hypercube**
  - **Torus**
  - **Tree**
  - **Ring**

Optical Hypercube-Connected Cluster Network (OHC²N)

- **Processors within a cluster are connected via a local intra-cluster WDM optical crossbar.**
- **Clusters are connected via inter-cluster WDM optical links.**
- **Each processor in a cluster has full connectivity to all processors in directly connected clusters.**
- **The inter-cluster crossbar connecting clusters are arranged in a hypercube configuration.**

Example OHC^2N (N = 32 processors)



OHC^{2}N Scalability

- The OHC^{2}N does not impose a fully connected topology, but efficient use of WDM allows construction of very large-scale (thousands of processors) networks at a reasonable cost.
- # nodes $N_H = n \times 2^d$
- Degree $D_H = d + 1$
- Diameter $K_H = d = log_2\left(\frac{N}{n}\right)$
- # links $L_H = \frac{1}{2}2^d d = \frac{N}{2n} log_2\left(\frac{N}{n}\right)$
- Bisection width $B_H = n2^{d-1} = N/2$
- Avg. Message Dist. $\bar{l}_H = \frac{1}{N-1}\left[\frac{N log_2\left(\frac{N}{n}\right)}{2} + (n-1)\right]$

Hardware Cost Scalability

- **A major advantage of a SOCN architecture is the reduced hardware part count compared to more traditional network topologies.**

	OC^3N	OHC^2N
VCSEL's (tunable)/ processor	O(c)	$O(\log_2(c))$
Detectors / processor	O(c)	$O(\log_2(c))$
Waveguides / processor	O(c)	$O(\log_2(c))$
Demultiplexers / cluster	O(c)	$O(\log_2(c))$

* c = # clusters = N/n

OC^3N and OHC^2N Scalability Ranges

- **An OC^3N fully connected crossbar topology could cost-effectively scale to hundreds of processors.**
  - —**Example: n = 16, c = 16, $N = n \times c$ = 256 processors. Each processor has 16 tunable VCSEL's and optical receivers, and the total number of inter-cluster links is 120. A traditional crossbar would require $(N^2-N)/2$ = 32,640 links.**

- **An OHC^2N hypercube connected topology could cost-effectively scale to thousands of processors.**
  - —**Example: n = 16, L = 9 (inter-cluster links / cluster), N = 8192 processors. Each processor has 10 tunable VCSEL's and optical receivers, the diameter is 10, and the total number of inter-cluster links is 2304. A traditional hypercube would have a diameter and degree of 13 and 53,248 inter-processor links would be required.**

Conclusions

- **In order to reduce costs and provide the highest performance possible, high performance parallel computers must utilize state-of-the-art off-the-shelf processors along with scalable network topologies.**
- **These processors are requiring much more bandwidth to operate at full speed.**
- **Current metal interconnections may not be able to provide the required bandwidth in the future.**
- **Optics can provide the required bandwidth and connectivity.**
- **The proposed SOCN class provides high bandwidth, low latency scalable interconnection networks with much reduced hardware part count compared to current conventional networks.**
- **Three optical interconnects technologies (free-space, waveguide, fiber) are combined where they are most appropriate.**