AMOS - Analysis of an Optically Interconnected Beowulf Cluster

G. A. Russell, K. J. Symington, J. F. Snowdon
Optically Interconnected Computing Group,
Heriot-Watt University, Edinburgh, UK
I. Gourlay, P. Dew
Informatics Research Institute,
University of Leeds, Leeds, UK

Acknowledgements

- Department of Physics, Heriot-Watt University.
  - J.F. Snowdon, A.C. Walker, K.J. Symington, T. Lim, B. Layet, J.J. Casswell, G.A. Russell
- School of Informatics, Leeds University
  - P. Dew, I. Gourlay, K. Djemame

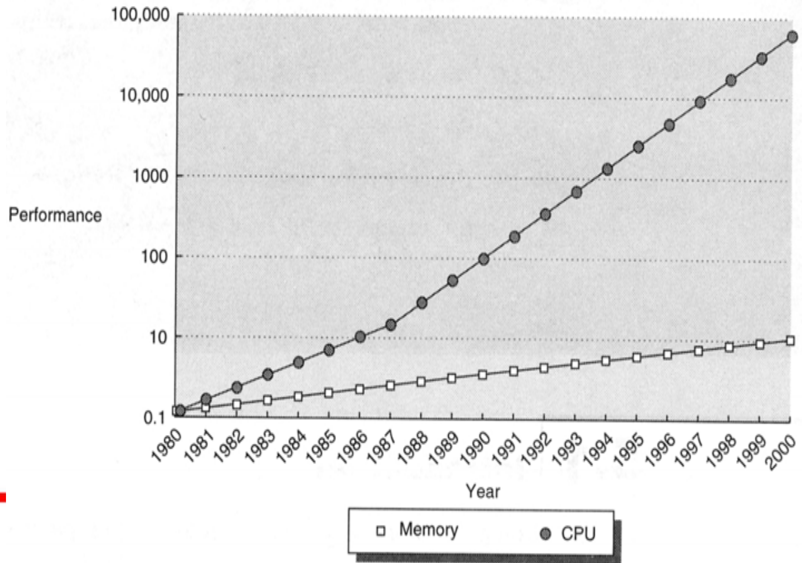
- EPSRC (OSI) funded AMOS project
- EPSRC funded PhD project

Key Points

- Modelling
  - Optical, Optoelectronic and Electrical
  - System Architecture and Software Abstraction
- Architectural Enhancements
  - Capacity
  - Distributed Network Intelligence
    - SPA Functionality
    - Shared Abstract Data Types
    - Load Balancing

The Problem

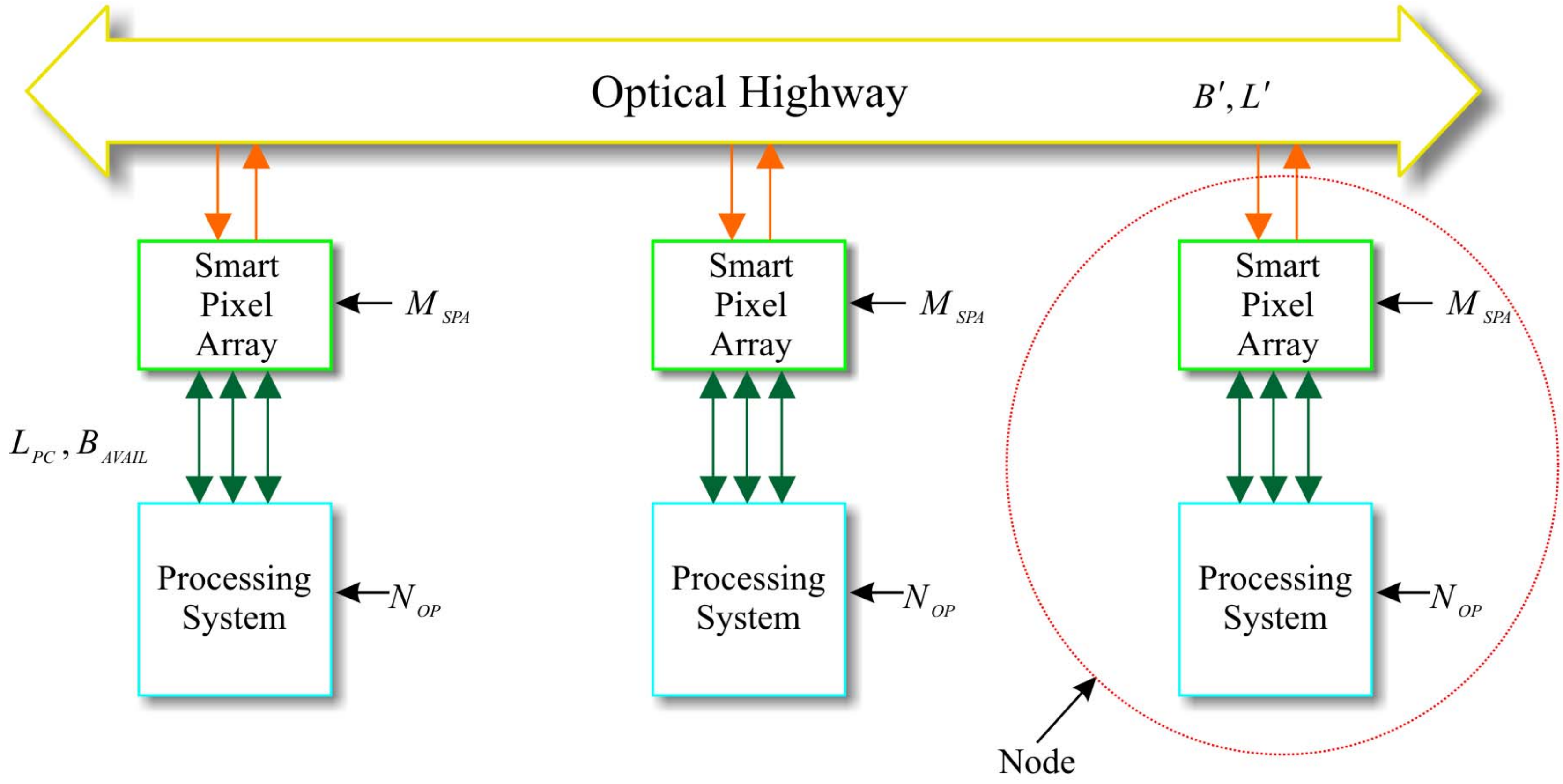
- Processors continue to increase in speed at equal or greater than Moore's Law
- Bandwidth continue to increase in speed at equal or greater than Moore's Law
- Processor Speed Increase >> Bandwidth Increase



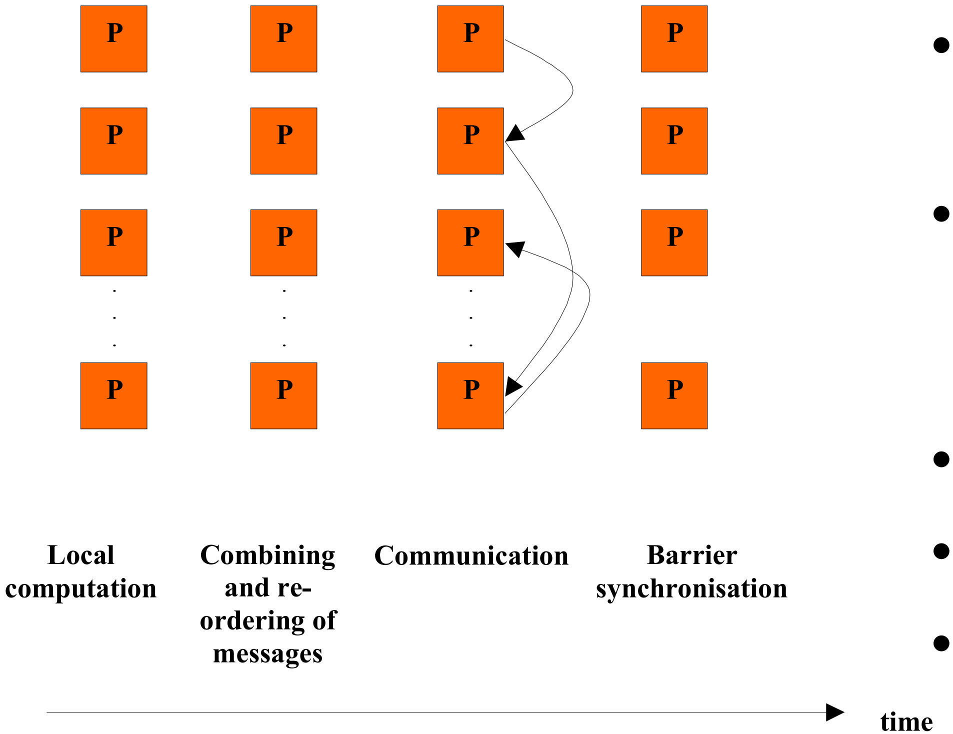
What is a Beowulf Cluster?

- Distributed memory multi-computer
- Commodity PC hardware
- Commodity OS (Windows, Linux)
- Message Passing Libraries (MPI)
- Excellent Cost vs. Performance

Modelling - System



Modelling Parallel Computers - BSP



- Write algorithm in terms of "Supersteps"
- Split communication and computational costs
- Small parameter set
- Measure parameters
- Maps well to Cluster architecture

The BSP cost model

Cost of a superstep:

$$C = w + gh + l$$

Local computation (→ w)

Communication (→ gh)

Barrier synchronisation (→ l)

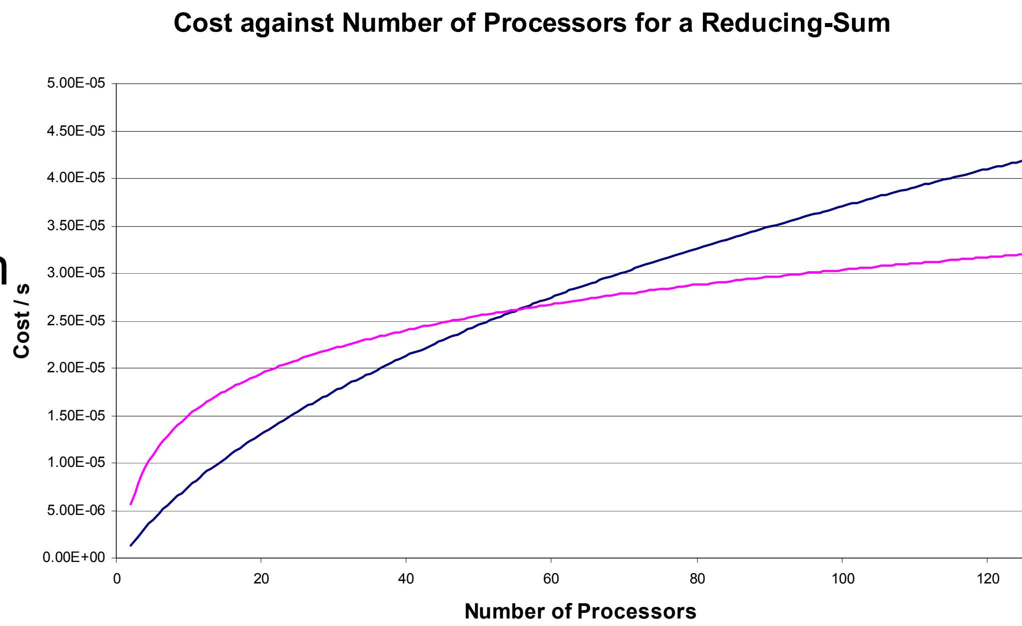
Reducing Sum Algorithm

- All processors start with 1 number. Want total on 1 machine.
- Each pair of machines add there numbers then each pair-of-pairs add and so on……..
- $Log_2(p)$ super-steps.
- Efficient on low connectivity.

- Implement on SPA layer.
- $Log_2(p)$-2 super-steps on SPA layer.
- Few computational or memory resources on SPA

SPA Functionality?

- Reducing-Sum
  - BSP
  - >50 processors
  - 32% performance gain at 128 processors
- Model Too Simple
- SADT?
- Load Balancing?



$$G = gh = 2\left(L_{PC} + L_{SPA} + R_{PC} + R_{SPA} + \frac{M_h}{B_{PC}} + \frac{M_h + M}{B_{SPA}} \right) + \left(\log_2(p) - 2\right)\left(L_{SPA} + R_{SPA} + \frac{M_h + M}{B_{SPA}} \right)$$

Modelling - Optoelectronic

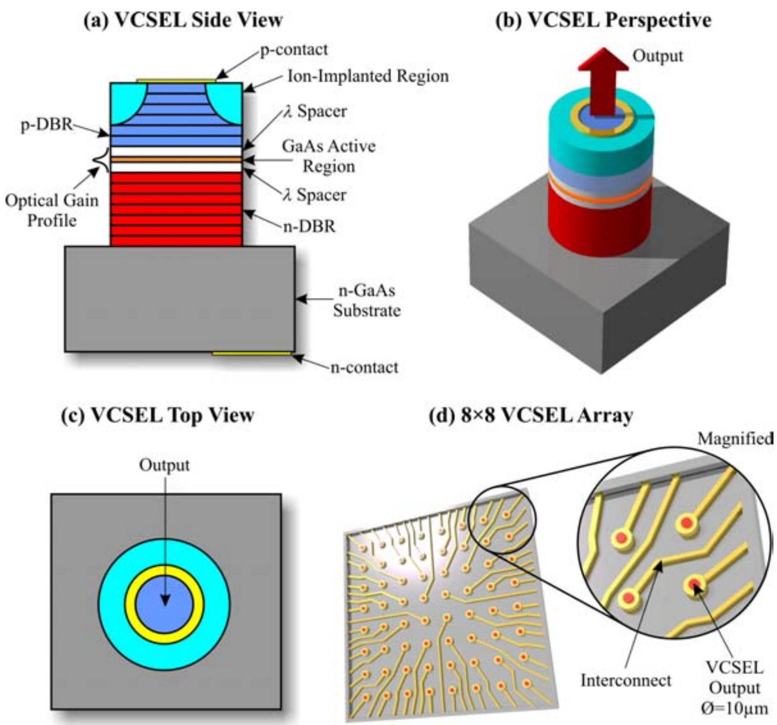
- Number of Channels Limited by Power
- Signal to Noise Ratio of Photodectector
- Optical Power of VCSEL
- Efficiency of Optical Elements
- Semiconductor Density

$$p_{power} = \frac{2}{\ln(\xi)} \ln\left(\frac{NEP\sqrt{\dfrac{8B'}{\xi_{ed} C_W \xi_{mirror}^2 \xi_{QWP}^2}}}{P_{VCSEL}} + 1 \right)$$

$$P_{VCSEL} = \frac{\dfrac{\eta_{li}}{V_{th}}}{\left(1 - \dfrac{\eta_{li}}{V_{th}}\right)} \left(P_{Velec} - I_{th} V_{th}\right)$$

$$p_{device} = \frac{N_d}{2 \cdot C_w}$$

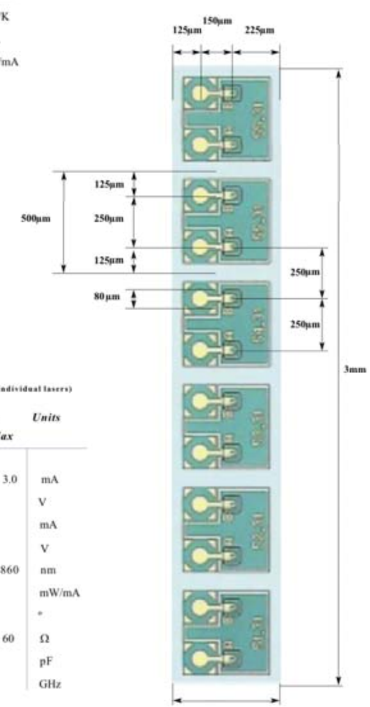
Experimental - Optoelectronics



Parameter	Symbol	Ratings Min	Ratings Typ	Ratings Max	Units
Temperature tuning coefficient	$\delta\lambda/\delta T$		0.06		nm/K
Threshold current variation 0 to +85°C	ΔI_{th}		0.6		mA
Current tuning coefficient	$\delta\lambda/\delta I$		0.2		nm/mA

Ordering information

APA1101120000 850nm multi-mode array 1x12



Absolute maximum ratings

Parameter	Symbol	Rating	Units
Optical output power	P_{max}	7	mW
Peak forward current	I_{max}	10	mA
Electrical power dissipation	P_{tot}	20	mW/laser
Reverse voltage	V_R	5	V
Operating temperature	T_{op}	0 to +85	°C
Storage temperature	T_{stg}	-40 to +100	°C

(T=25°C)

Electro-optical characteristics (for individual lasers)

Parameter	Symbol	Conditions	Ratings Min	Ratings Typ	Ratings Max	Units
Threshold current	I_{th}		1.0	1.8	3.0	mA
Threshold voltage	V_{th}			1.6		V
Operating current	I_{op}	typ.P_{out} = 1.5 mW		5		mA
Operating voltage	V_{op}	typ.P_{out} = 1.5 mW		1.8		V
Emission wavelength*	λ	I_F = 5m A	840	850	860	nm
Slope efficiency	η	I_F = 5m A		0.45		mW/mA
Beam divergence	θ	FWHM, I_F=5m A		16		°
Differential Resistance	R_{Diff}	I_F = 5m A	30	45	60	Ω
Capacitance	C	I_F = 5m A		0.8		pF
Bandwidth	f_{3dB}	I_F = 5m A		>3		GHz

*Tighter wavelength specifications available on request (T=25°C)

Modelling - Optical

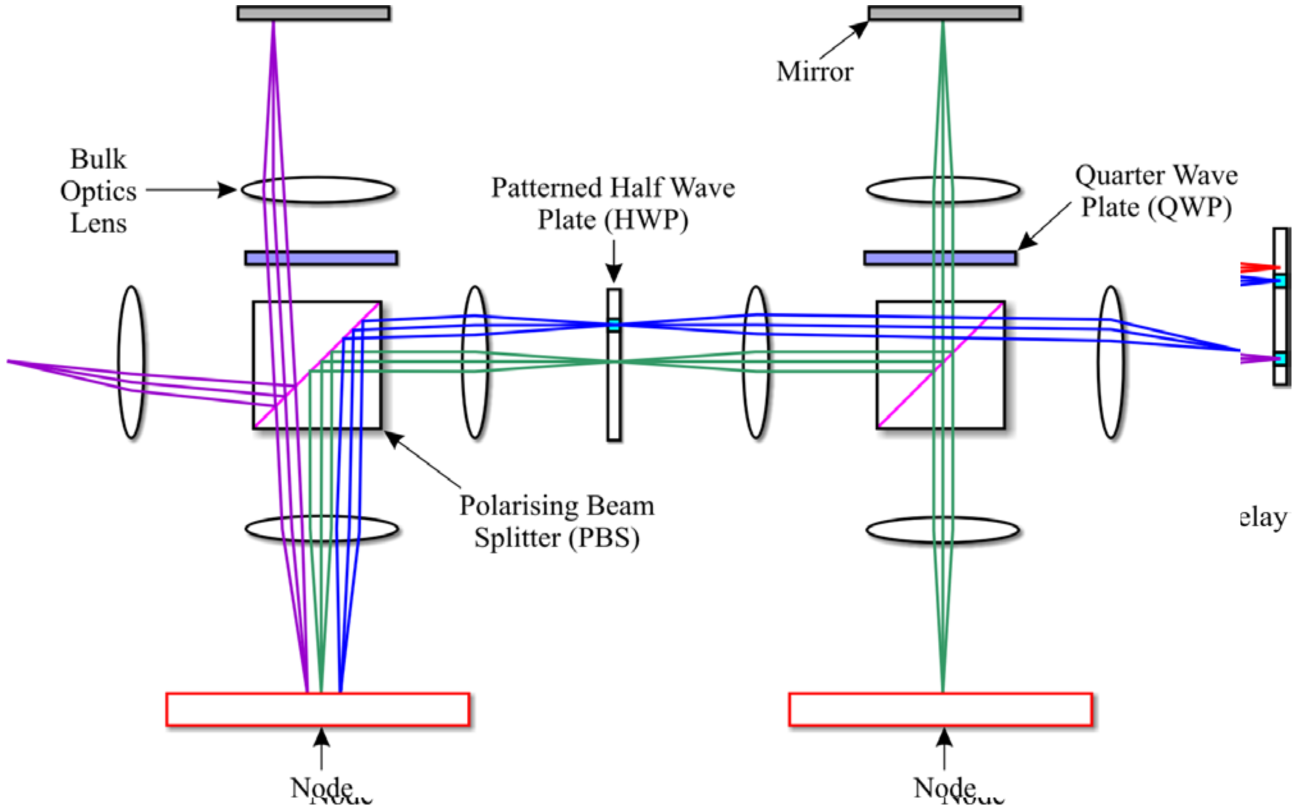
- Number of Channels Limited by Aberration, e.g. Spherical.
- For CCN h_{max} is also function of p_{ab}
- Close approximation to Code V simulation of small number of stages

$$p_{ab} = \frac{\pi\phi^2}{2C_w\left(h_{\max}\left(A_T^2 + \frac{\lambda^2}{16}\right) + s_{VCSEL}^2\right)}$$

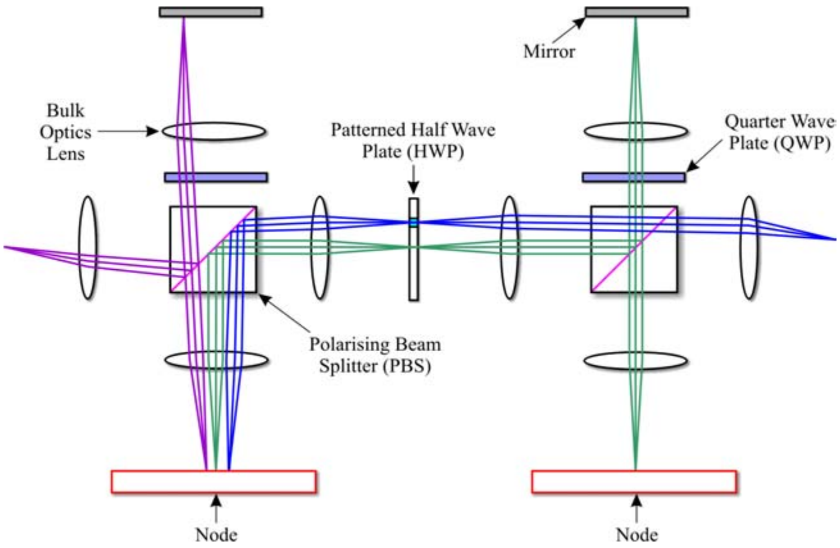
$$A_L = \frac{kf}{(f/\#)^2}$$

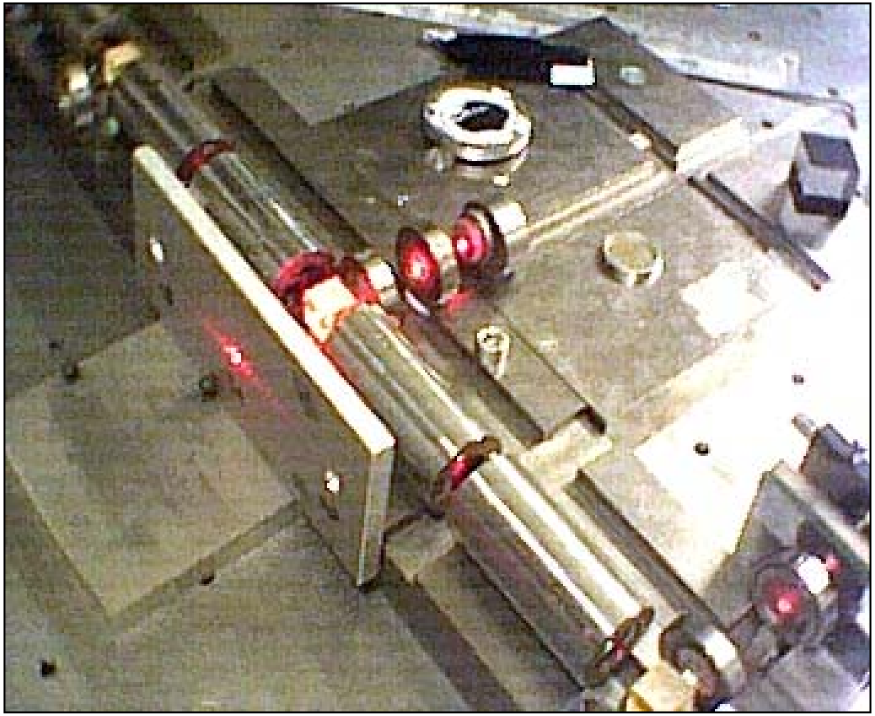
$$A_T = \frac{A_L\phi}{(f - A_L)}$$

Experimental - Optics



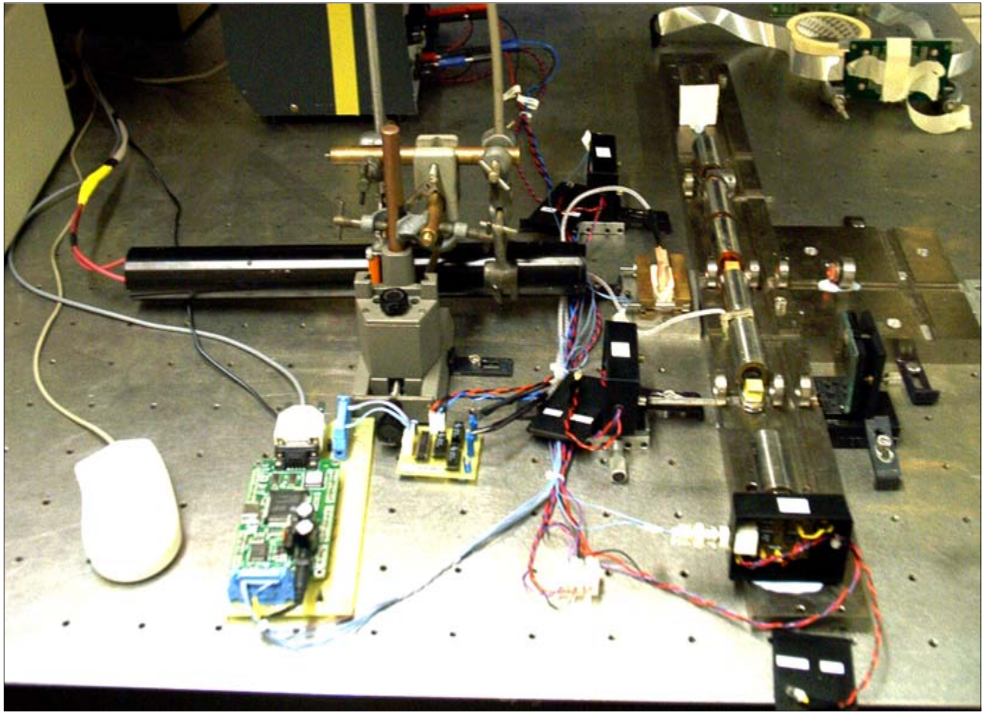
Experimental - Optics





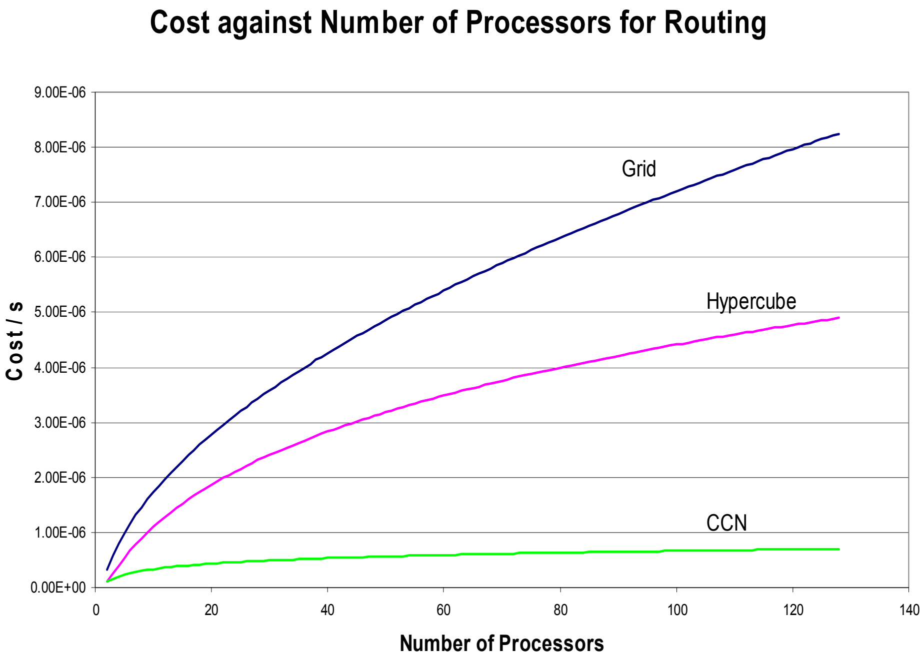
Experimental - Optics

- Massively parallel interconnect
- Polarisation controlled
- Reconfigurable via Liquid Crystals
- Demonstrator currently under construction



Capacity

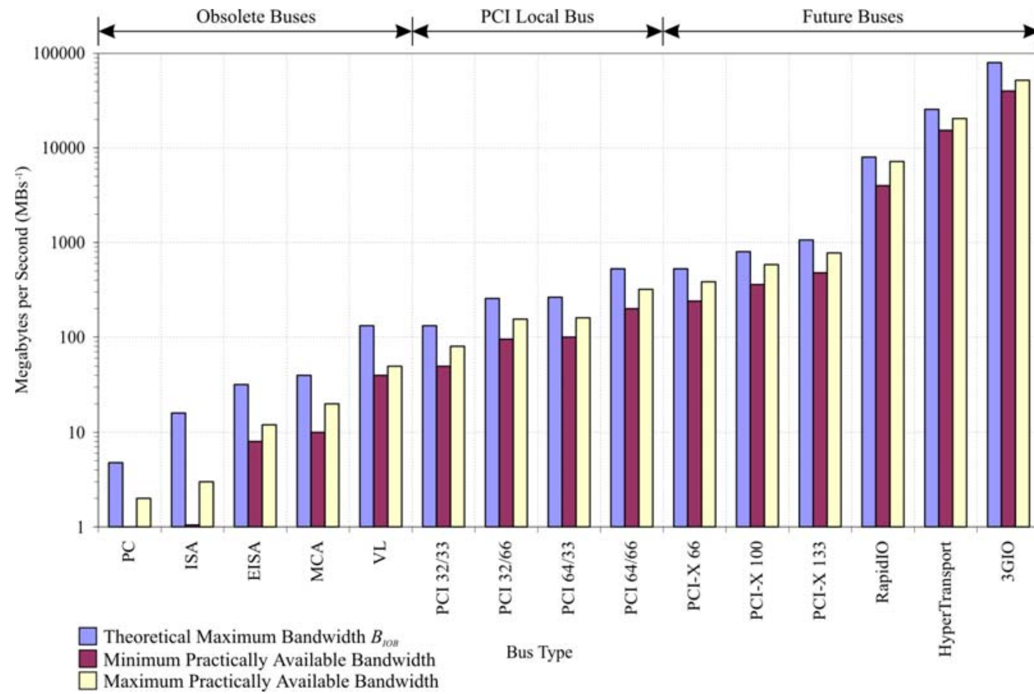
- Massive Parallelism
  - 1024 Processor Clusters
  - Huge Routing Cost
- High Connectivity
  - Reduces Routing
  - Increases Pin-out



- G. A. Russell , J. F. Snowdon, T. Lim, I. Gourlay, P. M. Dew, ***Modelling Of Optical Interconnects For Parallel Processing***, Conference Proceedings from PREP 2001 at University of Keele, UK, ISBN 1899371281, pp. 29-30, April 2001.

Modelling - System Bus

- Memory and IO bus efficiencies
  - Signalling Overheads
  - Protocol Overheads
- Processor Utilisation
  - DMA
  - Cache Stalls
- Transmission Lines

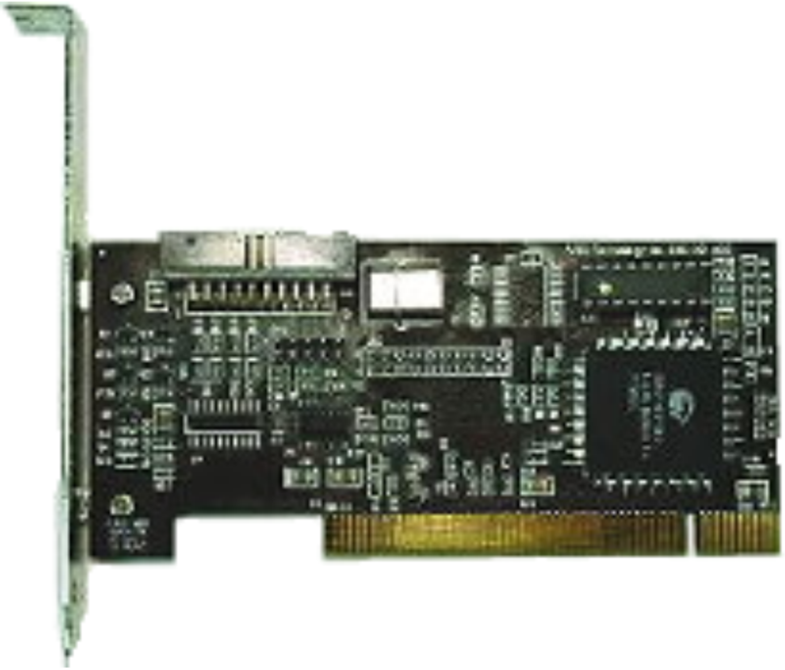


$$N_{op} = N_{\max} - N_{\max} O_{PS} \left(\frac{B_{per} + B_{act}}{B_{mem}} \right) - N_{ov}$$

$$B = 5 \times 10^{14} \frac{A}{D^2}$$

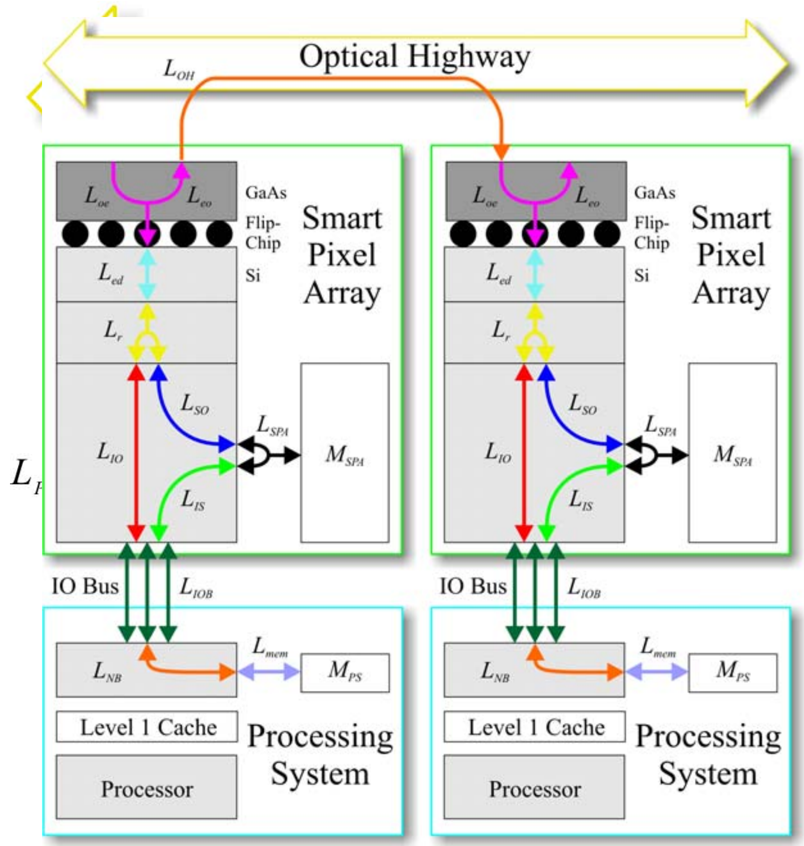
Experimental - System Bus

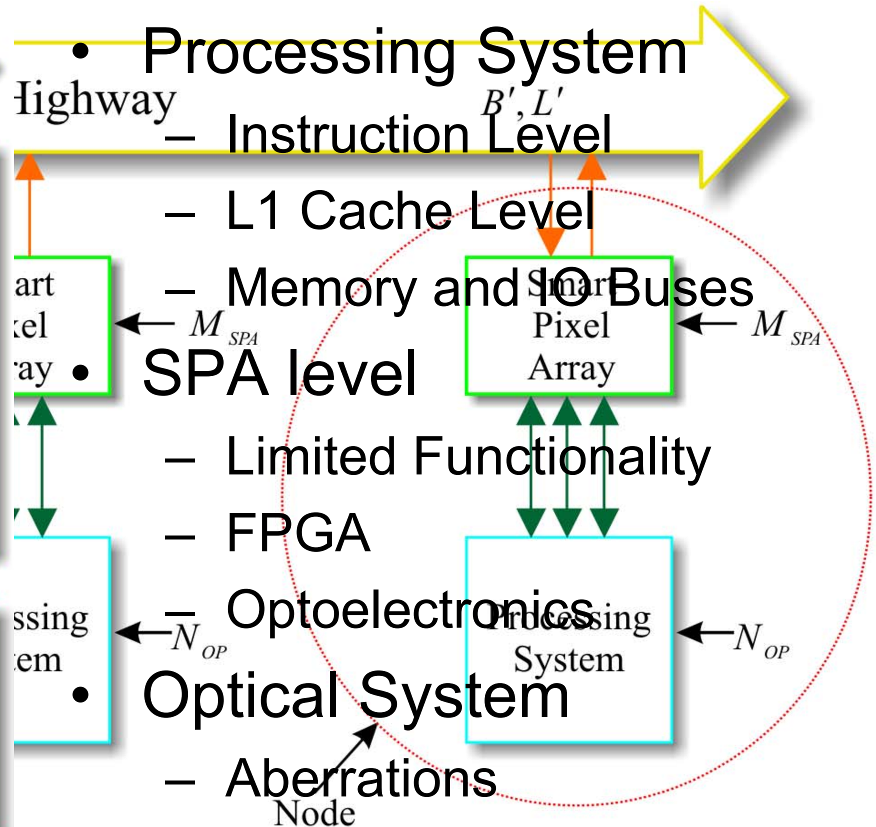
(a) RD2 PC Geiger PCI Card



(b) RD2 PC Geiger Bay Panel

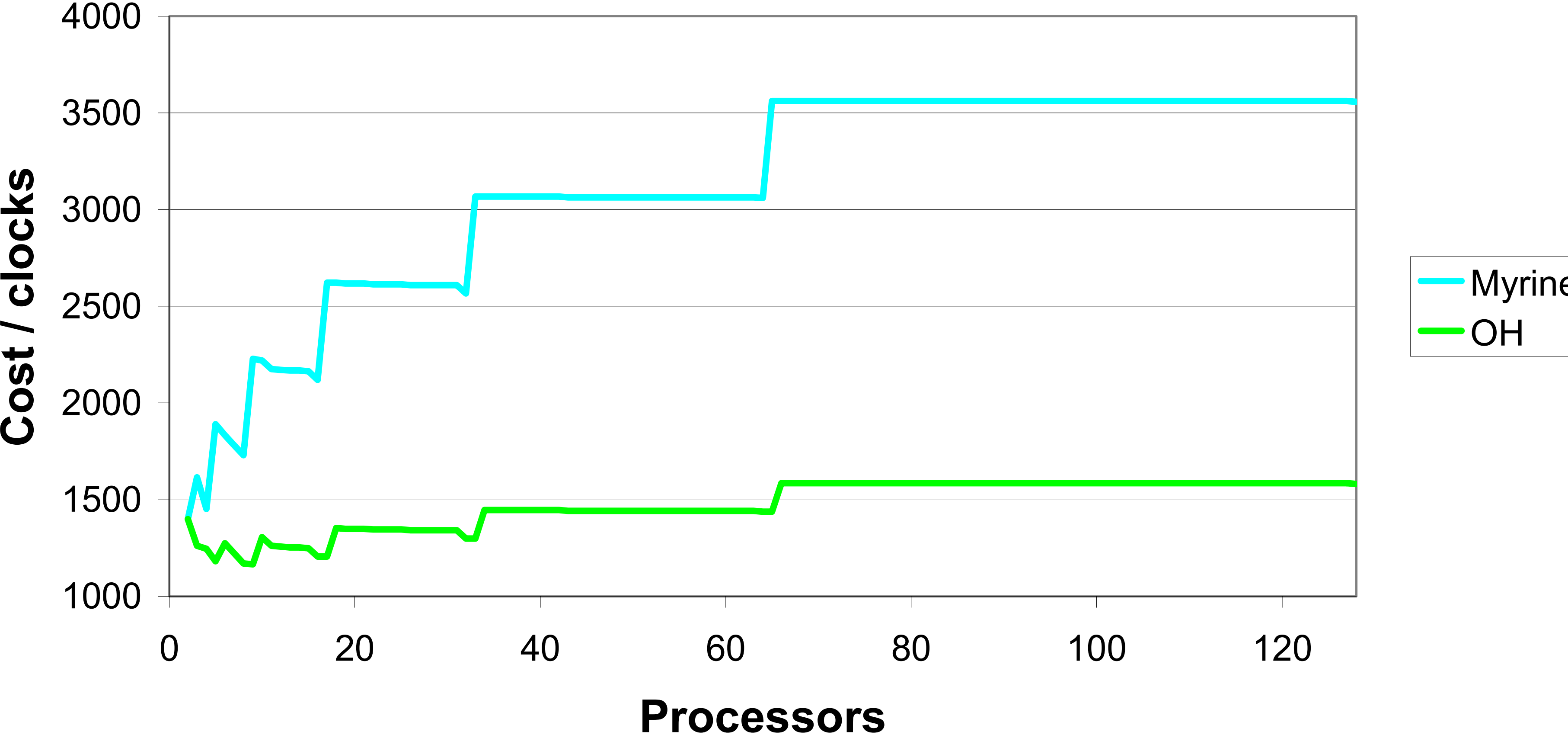
Putting It Altogether



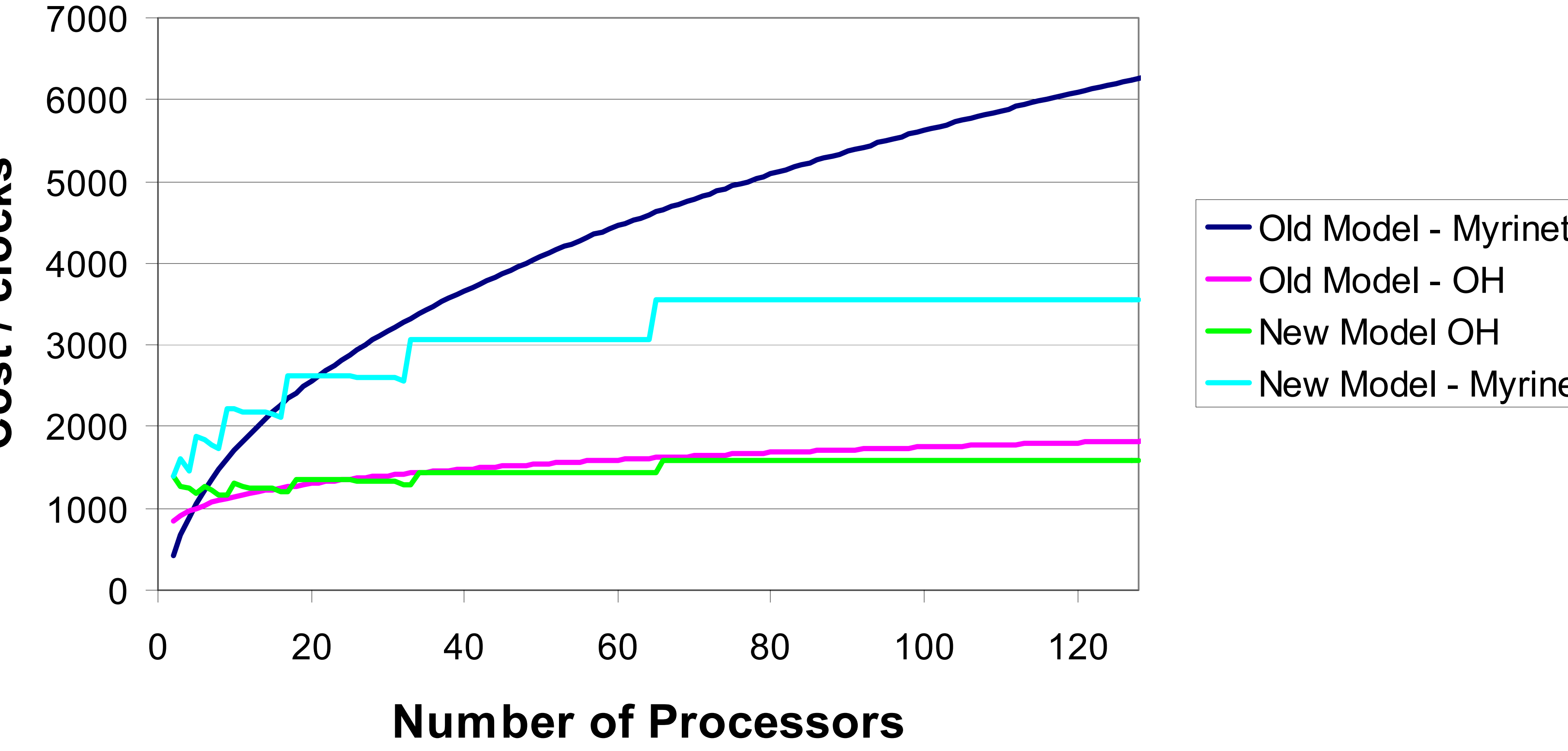


- Processing System
  - Instruction Level
  - L1 Cache Level
  - Memory and IO Buses
- SPA level
  - Limited Functionality
  - FPGA
  - Optoelectronics
- Optical System
  - Aberrations
  - Power Lose

Cost against Number of Processors for a Reducing Sum



Cost against Number of Processors for a Reducing-Sum



Sorting

Parallel sorting by regular sampling:

- This algorithm was chosen due to:
  - Suitability for combining large messages prior to communication.
  - Coarse-grain PC computations, allowing data to be sent from smart-pixel layer to PC layer during local computation.
- Results:
  - For large data sets, algorithm can be implemented, with effective communication bandwidth close to the optical interconnect bandwidth.

Cost for Sorting

- Most communication intensive part
  - Each processor receives *O(p)* sorted blocks of data, each of size n/p^2
  - Blocks must be sent from SPA layer to PC and merged
  - By data streaming $f \approx 1$
  - Send a fraction (1-*k*) of received blocks to the PC to start merging
  - If merging 2 blocks takes an/p^2 then require

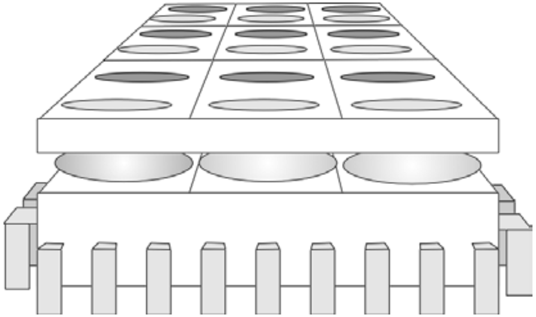
$$4^{\left(\frac{1}{aB_{pc}}\right)} \leq (1-k)p \quad \text{and} \quad \frac{(1-k)}{k} << 1$$

Bottom Line: We can exploit the bandwidth for large n

POCA

FPGA

- Reconfigurable
- High Bandwidth required for internal operation.



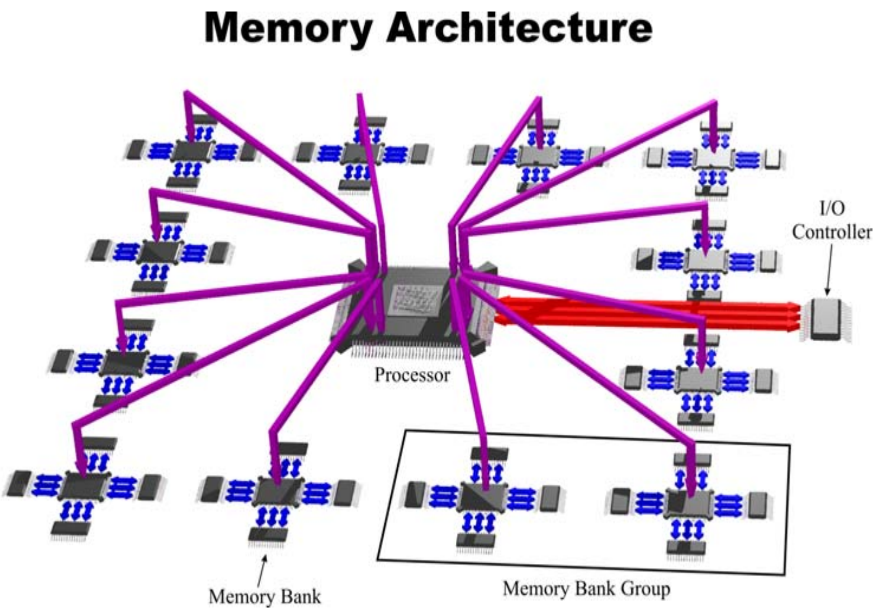
Optics

- Reconfigurable
- High Bandwidth
- Geometric Mapping

Applications

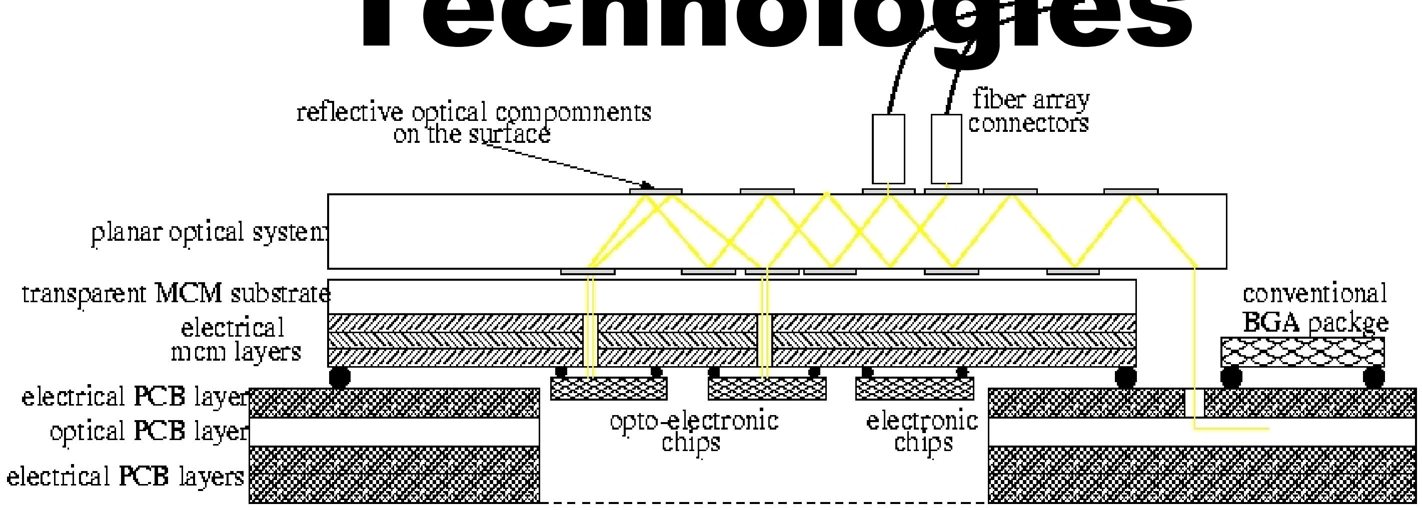
- Optical Interconnect
- Optical Neural Network
- Internet Backbone

HOLMS - The Future



- Multiple Optical Technologies
  - Planar Waveguide
  - Fibre
  - Free-space
- Custom Memory Controllers
- Mephisto (ARM) Processor?

HOLMS Optical Technologies



- Planar ‘Free-space’ optics
- Optical PCB
- Fibre

Conclusions

- Developed a Parameterised Model of an Optically Interconnected computer System at Algorithm, System and Component Levels.
- Network Capacity and Intelligence can allow Optical Bandwidth to be used in a Beowulf system.
- …..BUT Beowulf was the Wrong Architecture.
- Models will Hold for the RIGHT Architecture.

OIC Website

http://www.optical-computing.co.uk

AMOS

HOLMS

POCA

NOSC

OFPGA