

ARCHITECTURES

PETAFL0P2 CONFERENCE

Steven J. Wallach

CenterPoint Venture Partners

wallach@centerpointvp.com

Panelists

- Tilak Agerwala, IBM Corporation
- Greg Chesson, Silicon Graphics
- Peter Kogge, University of Notre Dame
- Burton Smith, Tera
- Thomas Sterling, Caltech/JPL

To Get Things Going

- ❶ Objectives
 - 10 years out
 - ASCI Class System
 - ASCI Level Cost
- ❷ Use SIA Study
- ❸ Some gee-whiz stuff
- ❹ Multi-Die, Multi-Computing

PetaFlop in 10 years

- ASCI Standard
 - 8192 nodes (what is a node?)
 - costs \$100 million
- New wrinkle(s)
 - Order the subsystems from the WEB
- Software Environment
 - LINUX/NT
 - Fortran ‘09 or JAVA 2009

SIA Study

- Projects Semi-conductor technology over the next 10 to 12 years.
- Exceptional study
- 1997 study
 - <http://notes.sematech.org>
 - new study in the works

WHAT WE GET

Table 3 Performance of Packaged Chips

<i>YEAR OF FIRST PRODUCT SHIPMENT</i>	<i>1997</i>	<i>1999</i>	<i>2001</i>	<i>2003</i>	<i>2006</i>	<i>2009</i>
<i>TECHNOLOGY GENERATIONS DENSE LINES (DRAM HALF-PITCH) (nm)</i>	<i>250</i>	<i>180</i>	<i>150</i>	<i>130</i>	<i>100</i>	<i>70</i>
<i>ISOLATED LINES (MPU GATES) (nm)</i>	<i>200</i>	<i>140</i>	<i>120</i>	<i>100</i>	<i>70</i>	<i>50</i>
<i>Number of Chip I/Os</i>						
<i>Chip-to-package (pads) high-performance</i>	<i>1450</i>	<i>2000</i>	<i>2400</i>	<i>3000</i>	<i>4000</i>	<i>5400</i>
<i>Chip-to-package (pads) cost-performance</i>	<i>800</i>	<i>975</i>	<i>1195</i>	<i>1460</i>	<i>1970</i>	<i>2655</i>
<i>Number of Package Pins/Balls</i>						
<i>ASIC (high-performance)</i>	<i>1100</i>	<i>1500</i>	<i>1800</i>	<i>2200</i>	<i>3000</i>	<i>4100</i>
<i>MPU/controller, cost-performance</i>	<i>600</i>	<i>810</i>	<i>900</i>	<i>1100</i>	<i>1500</i>	<i>2000</i>
<i>Cost-performance package cost (cents/pln)</i>	<i>1.40-2.80</i>	<i>1.25-2.50</i>	<i>1.15-2.30</i>	<i>1.05-2.05</i>	<i>0.90-1.75</i>	<i>0.75-1.50</i>
<i>Chip Frequency (MHz)</i>						
<i>On-chip local clock, high-performance</i>	<i>750</i>	<i>1250</i>	<i>1500</i>	<i>2100</i>	<i>3500</i>	<i>6000</i>
<i>On-chip, across-chip clock, high-performance</i>	<i>750</i>	<i>1200</i>	<i>1400</i>	<i>1600</i>	<i>2000</i>	<i>2500</i>
<i>On-chip, across-chip clock, cost-performance</i>	<i>400</i>	<i>600</i>	<i>700</i>	<i>800</i>	<i>1100</i>	<i>1400</i>
<i>On-chip, across-chip clock, high-performance ASIC</i>	<i>300</i>	<i>500</i>	<i>600</i>	<i>700</i>	<i>900</i>	<i>1200</i>
<i>Chip-to-board (off-chip) speed, high-performance (Reduced-width, multiplexed bus)</i>	<i>750</i>	<i>1200</i>	<i>1400</i>	<i>1600</i>	<i>2000</i>	<i>2500</i>

WHAT WE GET

Table 24 Product Critical Level Lithography Requirements

<i>Year of First Product Shipment Technology Generation</i>	<i>1997 250 nm</i>	<i>1999 180 nm</i>	<i>2001 150 nm</i>	<i>2003 130 nm</i>	<i>2006 100 nm</i>	<i>2009 70 nm</i>	<i>2012 50 nm</i>
<i>Product Application</i>							
DRAM (bits)	256M	1G	—	4G	16G	64G	256G
MPU (logic transistors/cm ²)	4M	6M	10M	18M	39M	84M	180M
ASIC (usable transistors/cm ²)*	8M	14M	16M	24M	40M	64M	100M
<i>Minimum Feature Size (nm)**</i>							
Isolated lines (MPU Gates)	200	140	120	100	70	50	35
Dense lines (DRAM Half Pitch)	250	180	150	130	100	70	50
Contacts	280	200	170	140	110	80	60
Development capability (minimum feature size, nm)	140	120	100	70	50	35	25
Gate CD control (nm, 3 sigma at post-etch)**	20	14	12	10	7	5	4
Product overlay (nm, mean + 3 sigma)**	85	65	55	45	35	25	20

ARCH. - LONG TERM- 2009

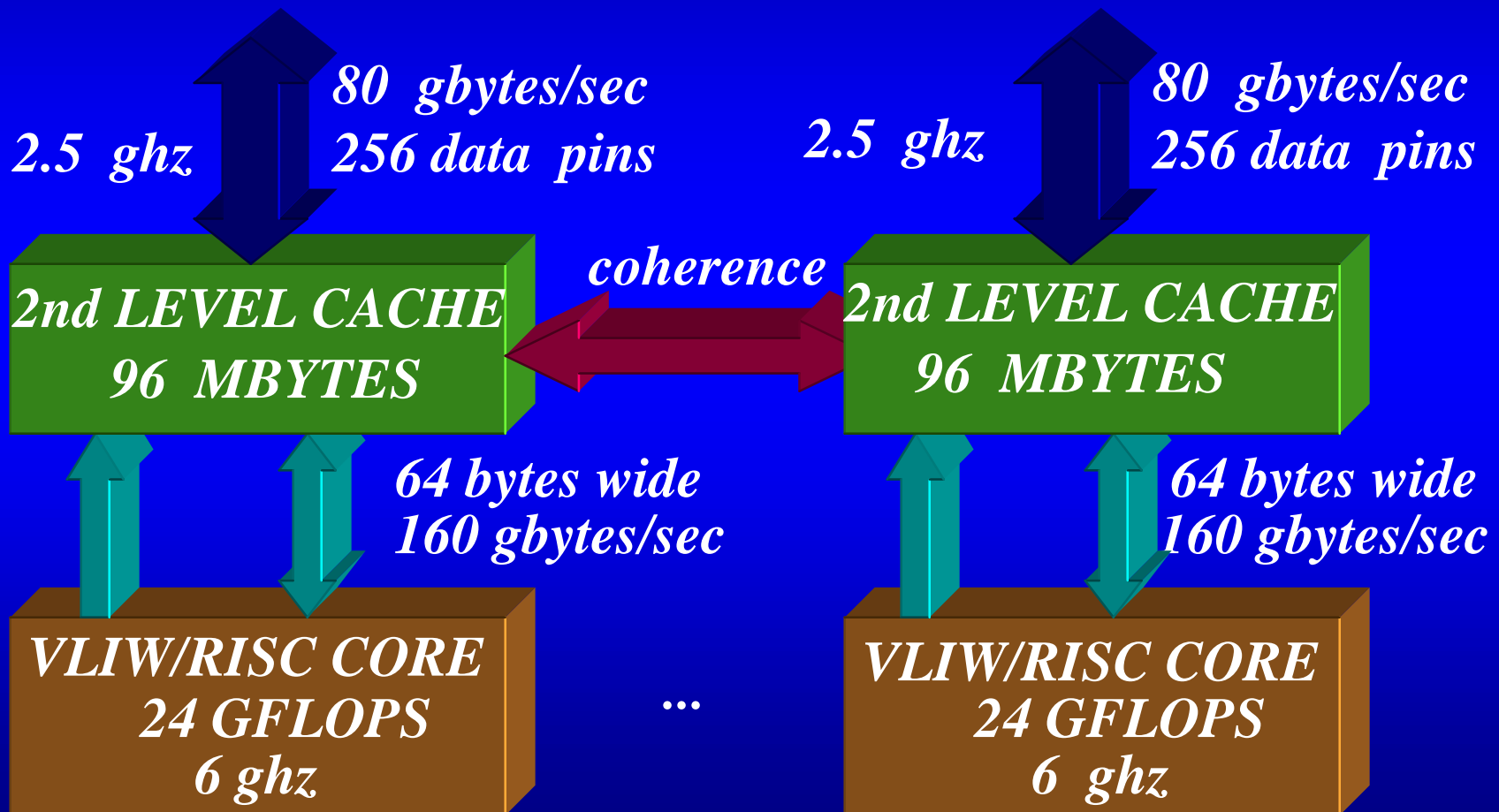
- **THE SIA STUDY TEACHES US:**
 - 64 gbits of dram - (8 gbytes)
 - 8 gbits of sram
 - 520 million MPU transistors
 - 70 nm lithography, 2.54 cm on-a-side
 - 6 ghz clock within vliw/risc core
 - 2.5 ghz across die
 - 2500 external signal pins

ARCH. - LONG TERM - 2009

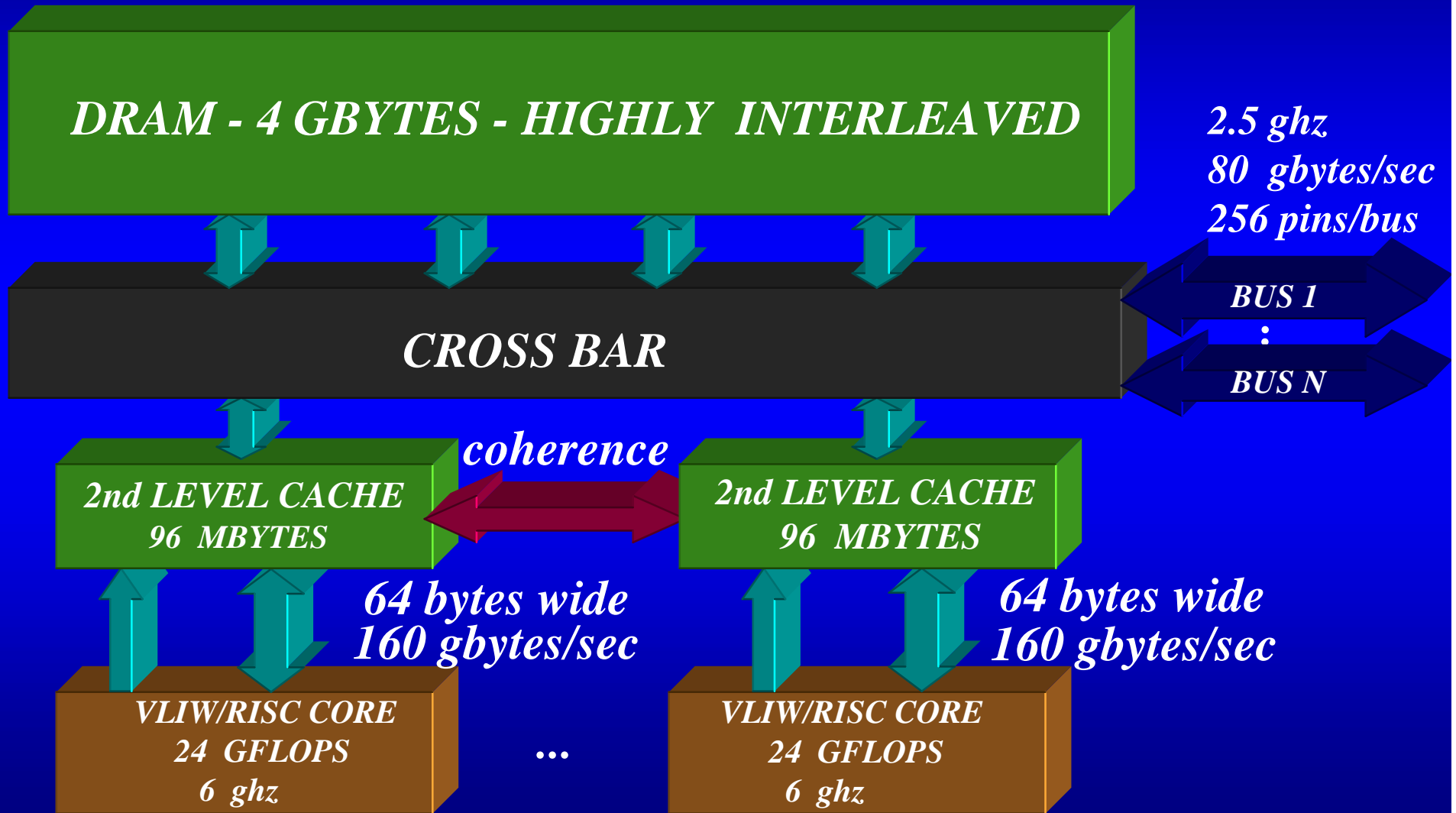
DESIGN ASSUMPTIONS

- 9 million transistors - vliw/risc core with first level cache.
- 2nd. Level cache - rule of thumb. 1/4 to 1/2 mbyte per 100 mflops peak.
- 96 mbyte 2nd. Level (6 Inst, 90 data)
- 170 watts
- .6 to .9 volts power supply

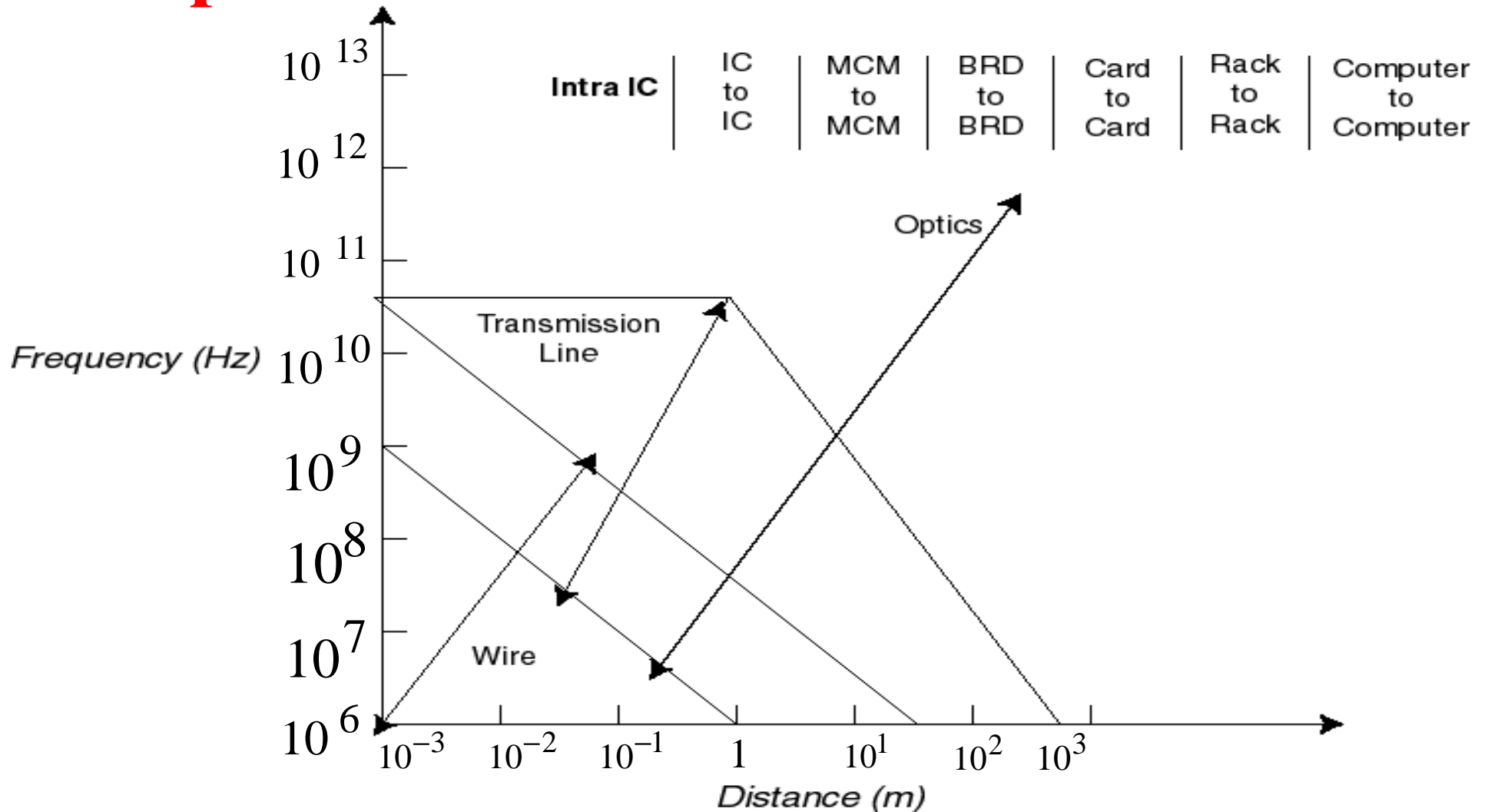
MAXIMUM PIN-USE EXTERNAL SMP- 6/8 CPU'S



INTEGRATED SMP - 4 CPU



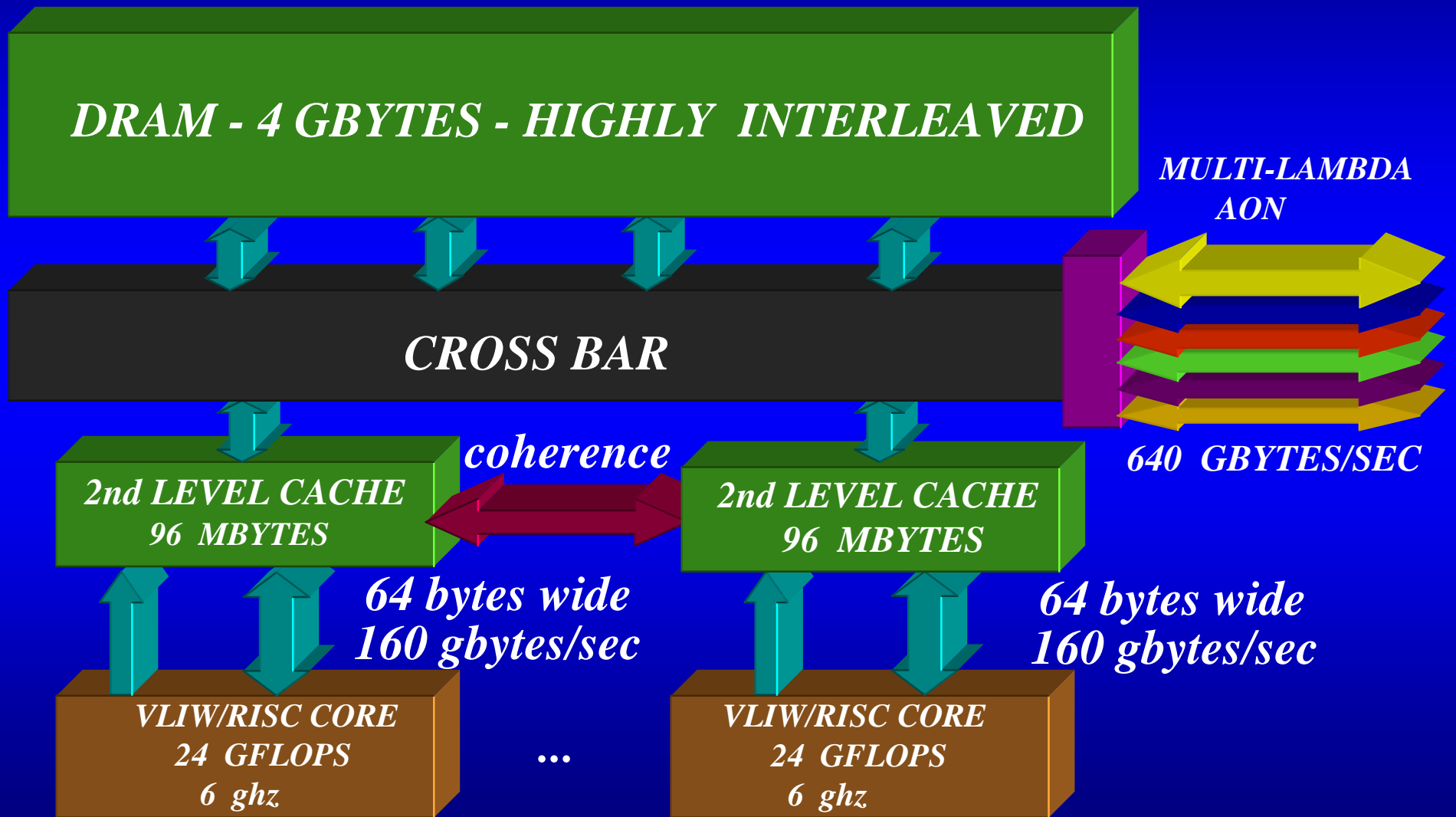
Optics: The Preferred Interconnect



Optics: The preferred interconnect technology for the higher frequency and longer distance applications [Feldman:88a] [Tsang:90a]

www.hpcc.gov/talks/petaflops-24june97

INTEGRATED SMP - WDM



COTS PetaFlop System

- 8192 Dies (4 CPU/die-minimum)
- Each Die is 120 GFlops
- 1 PetaFlop Peak
- Power 8192 x200 Watts = 1.6 MegaWatts
- Extra Main Memory yields in excess of 3 MegaWatts (512 TBytes)

COTS PetaFlop System

- 15.36 TFlops/Rack (128 die)
- 30 KWatts/Rack - thus 64 racks - 30 inch
- Common System I/O
- 2 Level Main Memory
 - local on chip
 - off-chip within same rack
 - same bandwidth/longer latency
 - meets Byte/Flop metrics
 - reduces external rack bandwidth by a factor of 10

COTS PetaFlop System

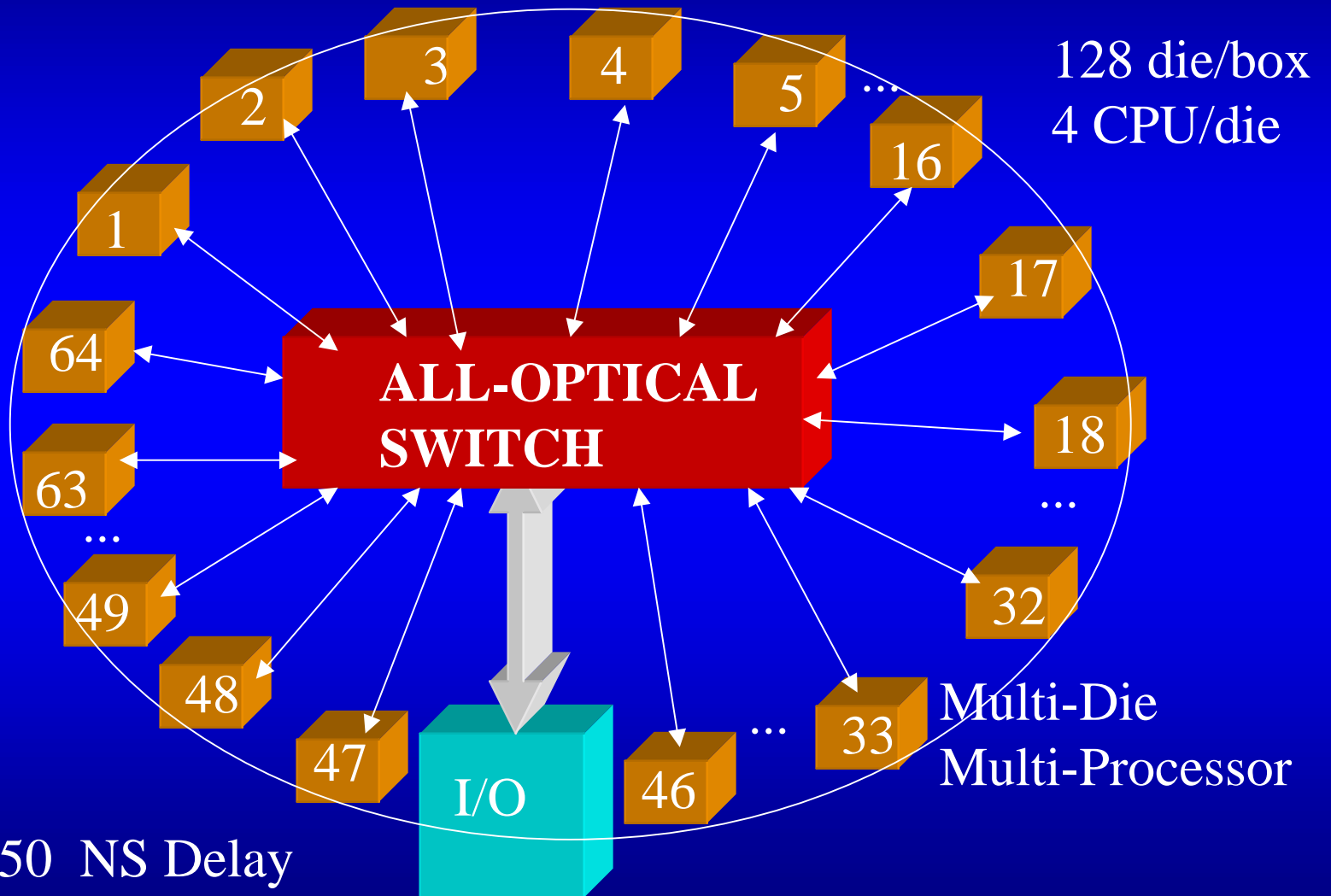
- Optical Interconnect
 - OC768 Channels (40 GHz)
 - 128 Channels per Die (DWDM)-5.12 THz
 - ALL Optical Switching
- Bisection Bandwidth of 50 TBytes/sec
 - 15 TFlops/rack*.1bytes/flop/sec*32 racks
- Rack Bandwidth - 15 TFlops*.1= 12 THz
- 2-4 OPTICAL MT-RJ Connectors or equivalent per rack

COTS PetaFlop System

Memory Hierarchy

- Physical Memory
 - Cache
 - Local Memory (Max 2 levels)
 - Global Memory
- I/O
- Substantial Bandwidth (like a vector system)

COTS PetaFlop System



10 meters = 50 NS Delay