Moore's Law
A Time for Reflection & Refraction
PetaFlops in 2009/10

Steve Wallach – Chiaro Networks

swallach@chiaro.com

LANL Contract: W-7405-ENG-36 (C7304)

Presentation

- Moore's Observation
- 2009/2010 COTS PetaFlop Supercomputers
  - Transistor Level
  - System Level
  - Grid Level
  - A new particle

Hardware (constant price)

- Moore's Law (really an observation)
  - doubling of processor performance every 18 months (actually the number of transistors)
- Dram
  - 4 times the capacity every 3 to 4 years
- Disk
  - doubling disk capacity every 1.5 years

Computer Advertisement - Nov. 1990

Option #2: Above w/40MB HD, Super VGA monitor and card. $1949

Option #3: Above w/89MB HD, Super VGA monitor and card. $2125

DELL COMPUTER CORPORATION

Dell System® 320LX
- 386SX 20MHz
- 1MB RAM - 40MB HD
- 1.2MB 5¼" floppy drive
- VGA Color Plus monitor/card
- One year on-site service

$2799

Option #2: With MS-DOS 5.3. Add: 99.95

Dell System and the Dell and Dell Computer logos are registered trademarks of, and Dell is a trademark of, Dell Computer Corporation.

ALR

PowerFlex Model 1 Computer
- 286 12.5MHz
- 1MB RAM
- 1.44MB 3½" floppy drive
- Upgradeable to 386/486SX

$849

Option #2: Above w/40MB HD, mono monitor and card. $1249

Option #3: Above w/40MB HD, VGA/color monitor and card. $1499

Microcom

Lotus

Dallas Morning News

Comparison Side by Side - 2003

• 386 - 20 MHz	150:1	• Pentium IV – 3 GHz
• 1 MByte Ram	1024:1	• 1024 MByte Ram
• 40 MByte Disk	5000:1	• 200 GByte Disk
• 2400 baud	416:1	• 1 Mbit (dsl or cable)
• DOS Based/VMS		• DOS Based/WNT

Double Every 18 months (4 times every 36 months) yields 128 times in 11 years

Triple Every 36 months yields 81 times every 12 years

The most important advance in software in the last 10 years is OPEN SOURCING – not technology

LinPack Analysis

- Compaq 386/SX20 SX with FPA - .16 Mflops
- Pentium IV – 2.80 GHz – 1317 Mflops
- 12 years ➔ 8231 (Doubling in less than a year, for 12 years)
- How
  - Clock = 140x (Moore's Law)
  - External Bus Width & Caching – 16 vs. 64 bits = 4x
  - Floating Point - 4/8 bits multi vs. 64 bits (1 clock) = 8x
  - Compiler Technology = 2x

Parallelism and Software's Impact as significant As Moore's Law

ITRS RoadMap- 2001/2002

Selected Figures/Tables

http://public.itrs.net

Transistor Density – Near Term

Table 1i High-Performance MPU and ASIC Product Generations and Chip Size Model—Near-term Years

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007
DRAM ½ Pitch (nm)	130	115	100	90	80	70	65
MPU/ASIC ½ Pitch (nm)	150	130	107	90	80	70	65
MPU Printed Gate Length (nm)	90	75	65	53	45	40	35
MPU Physical Gate Length (nm)	65	53	45	37	32	28	25
Logic (Low-volume Microprocessor) High-performance ‡							
Generation at production **	p01h	—	p03h	—	p05h	—	p07h
Functions per chip (million transistors)	276	348	439	553	697	878	1106
Chip size at production (mm^2) §§	310	310	310	310	310	310	310
High-performance MPU Mtransistors/cm^2 at production (including on-chip SRAM) ‡	89	112	142	178	225	283	357
ASIC							
ASIC usable Mtransistors/cm^2 (auto layout)	89	112	142	178	225	283	357
ASIC max chip size at production (mm^2) (maximum lithographic field size)	800	800	572	572	572	572	572
ASIC maximum functions per chip at production (Mtransistors/chip) (fit in maximum lithographic field size)	714	899	810	1020	1286	1620	2041

Transistor Density – Long Term

Table 1j High-Performance MPU and ASIC Product Generations and Chip Size Model—Long-term Years

Year of Production	*2010*	*2013*	*2016*
DRAM ½ Pitch (nm)	45	32	22
MPU/ASIC ½ Pitch (nm)	45	32	22
MPU Printed Gate Length (nm)	25	18	13
MPU Physical Gate Length (nm)	18	13	9
Logic (Low-volume Microprocessor) High-performance ‡			
*Generation at production ***	—	13h	—
Functions per chip (million transistors)	2212	4424	8848
Chip size at production (mm^2) §§	310	310	310
High-performance MPU Mtransistors/cm^2 at production (including on-chip SRAM)	714	1427	2854
ASIC			
ASIC usable Mtransistors/cm^2 (auto layout)	714	1427	2854
ASIC maximum chip size at production (mm^2)(maximum lithographic field size)	572	572	572
ASIC maximum functions per chip at ramp (Mtransistors/chip) (fit in maximum lithographic field size)	4081	8163	16326

Notes for Tables 1i and 1j:

I/O PADS

Table 3a Performance of Packaged Chips: Number of Pads and Pins—Near-term Years

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007
DRAM ½ Pitch (nm)	130	115	100	90	80	70	65
MPU/ASIC ½ Pitch (nm)	150	130	107	90	80	70	65
MPU Printed Gate Length (nm)	90	75	65	53	45	40	35
MPU Physical Gate Length (nm)	65	53	45	37	32	28	25
Number of Chip I/Os (Number of Total Chip Pads) — Maximum							
Total pads—MPU	3072	3072	3072	3072	3072	3072	3072
Signal I/O—MPU (1/3 of total pads)	1024	1024	1024	1024	1024	1024	1024
Power and ground pads—MPU (2/3 of total pads)	2048	2048	2048	2048	2048	2048	2048
Total pads—ASIC high-performance	3000	3200	3400	3600	4000	4200	4400
Signal I/O pads—ASIC high-performance	1500	1600	1700	1800	2000	2100	2200
Power and ground pads—ASIC high-performance (½ of total pads)	1500	1600	1700	1800	2000	2100	2200
Number of Total Package Pins—Maximum [1]							
Microprocessor/controller, cost-performance	480–1,200	480–1320	500–1452	500–1600	550–1760	550–1936	600–2140
Microprocessor/controller, high-performance	1200	1320	1452	1,600	1,760	1,936	2,140
ASIC (high-performance)	1700	1870	2057	2263	2489	2738	3012

Frequency/Gate Length 2001-2016

48 Overall Roadmap Technology Characteristics

Previous View | Next View

Table 4c Performance and Package Chips: Frequency On-Chip Wiring Levels—Near-term Years

YEAR OF PRODUCTION	2001	2002	2003	2004	2005	2006	2007
DRAM ½ Pitch (nm)	130	115	100	90	80	70	65
MPU/ASIC ½ Pitch (nm)	150	130	107	90	80	70	65
MPU Printed Gate Length (nm)	90	75	65	53	45	40	35
MPU Physical Gate Length (nm)	65	53	45	37	32	28	25
Chip Frequency (MHz)							
On-chip local clock	1,684	2,317	3,088	3,990	5,173	5,631	6,739
Chip-to-board (off-chip) speed (high-performance, for peripheral buses)[1]	1,684	2,317	3,088	3,990	5,173	5,631	6,739
Maximum number wiring levels—maximum	7	8	8	8	9	9	9
Maximum number wiring levels—minimum	7	7	8	8	8	9	9

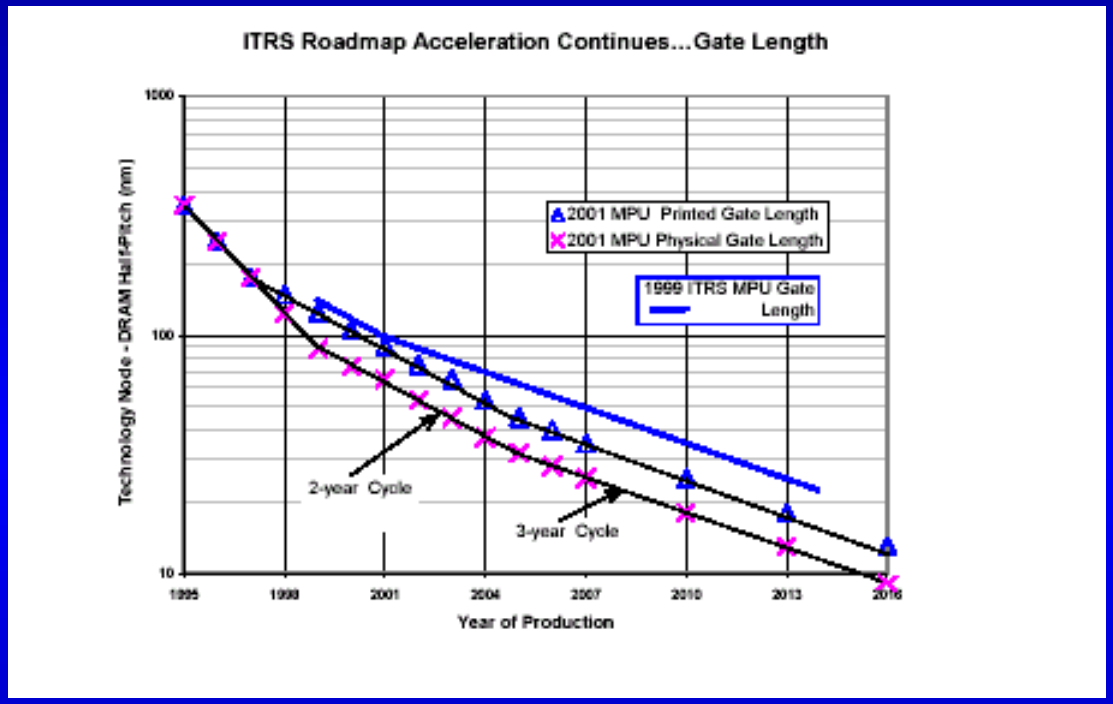
Table 4d Performance and Package Chips: Frequency On-Chip Wiring Levels—Long-term Years

YEAR OF PRODUCTION	2010	2013	2016
DRAM ½ Pitch (nm)	45	32	22
MPU/ASIC ½ Pitch (nm)	45	32	22
MPU Printed Gate Length (nm)	25	18	13
MPU Physical Gate Length (nm)	18	13	9
Chip Frequency (MHz)			
On-chip local clock	11,511	19,348	28,751
Chip-to-board (off-chip) speed (high-performance, for peripheral buses)[1]	11,511	19,348	28,751
Maximum number wiring levels—maximum	10	10	10
Maximum number wiring levels—minimum	9	9	10

ITRS CONCLUSION

- Approximately a tripling of I/O Pin transfer rates over the next 9 years.
- Overall
  - we see 9x the gate count
  - we see 2.3x the number of pins
  - we see 3x the I/O pin speeds (source synchronous, greater with clock recovery mechanisms) – how many?

CONCLUSION



Compared to 1999, technology advancements are increasing.

Economic Warfare – only the financially strong survive

Reflection and Refraction

E2 **When Will Optical Interconnects Appear on High-Performance Microprocessors?**

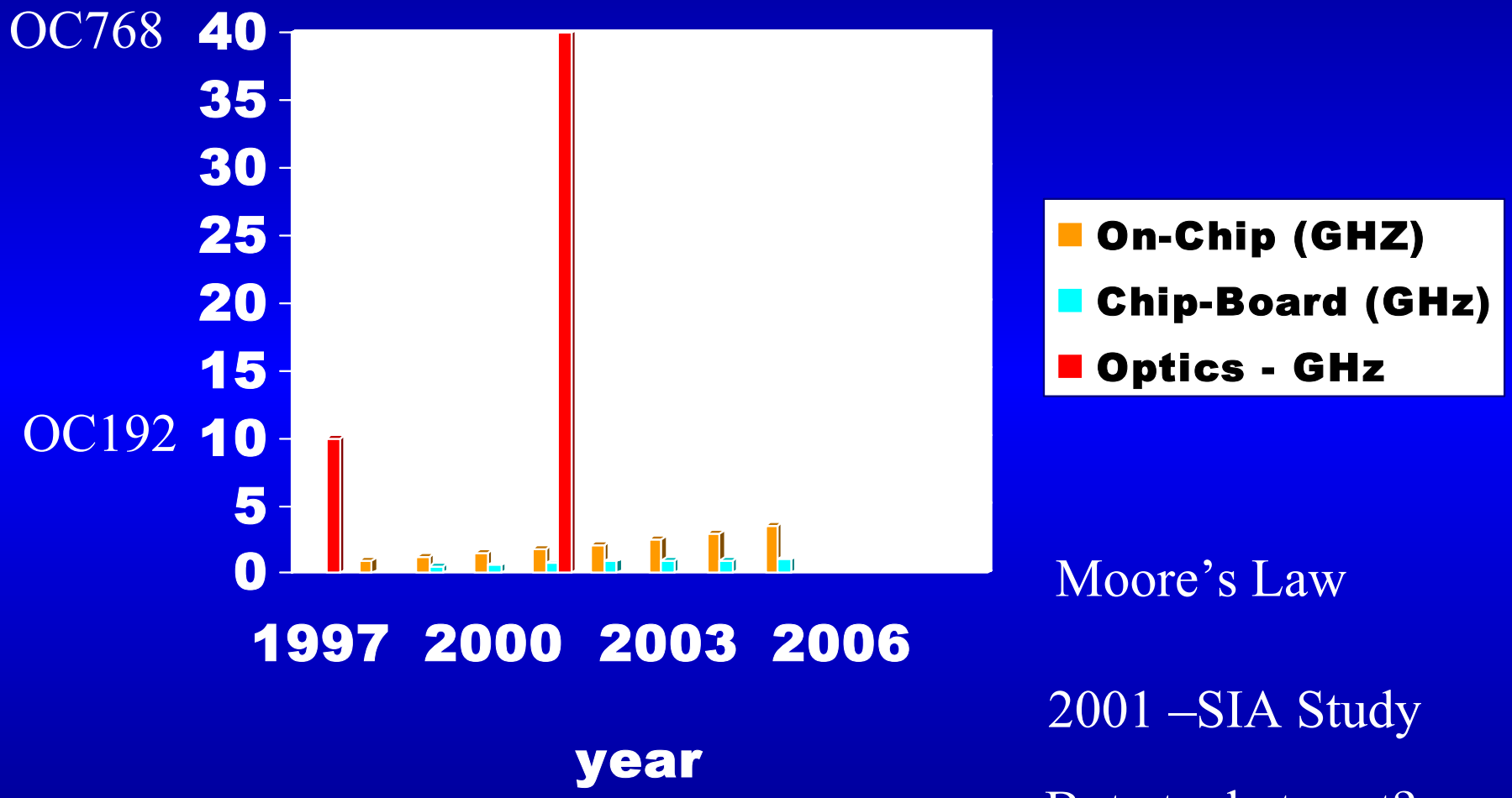
(Salon 7)

Organizer: Stefanos Sidiropoulos, Aeluros Inc, Palo Alto, CA
Moderator: Ian Young, Intel Corp., Hillsboro, OR

Performance of modern processors is limited not by intrinsic transistor speed, but by bandwidth and latency of internal and external interconnects. Optical interconnect technology promises virtually unlimited bandwidth. Will large-scale digital chips be forced to use optical interconnects on-chip and to the rest of the system? Will on-chip copper interconnects with mixed-signal I/O and DSP techniques enable microprocessors to remain purely electrical chips? Are the cost/performance advantages of opto-electronics attractive enough to overcome the inertia of a mature industry?

ISSC 2002 – San Francisco

OPTICAL SWITCHING - WHY



Moore's Law

2001 –SIA Study

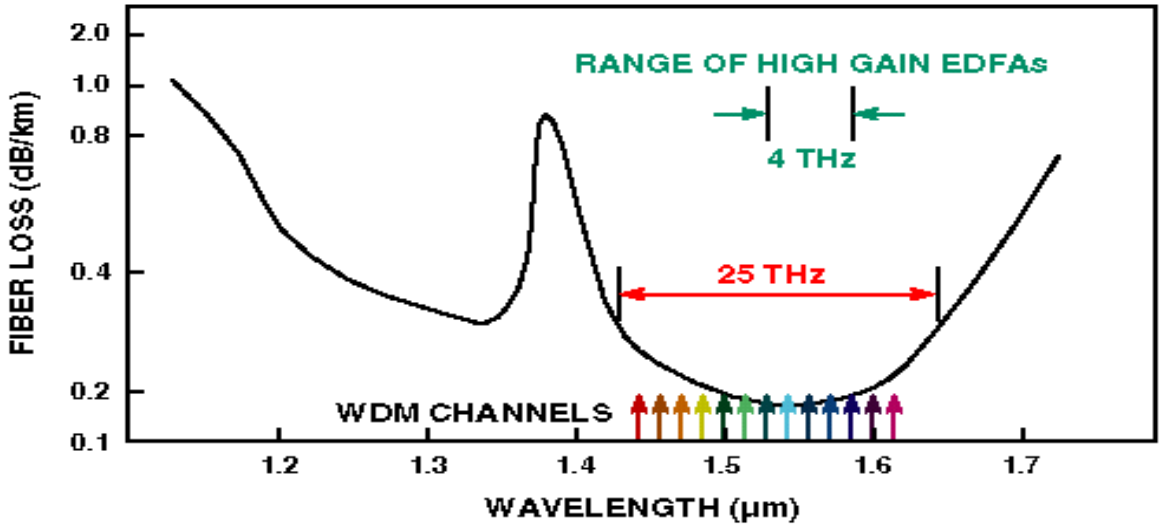
But at what cost?

AMPLIFIER TECHNOLOGY

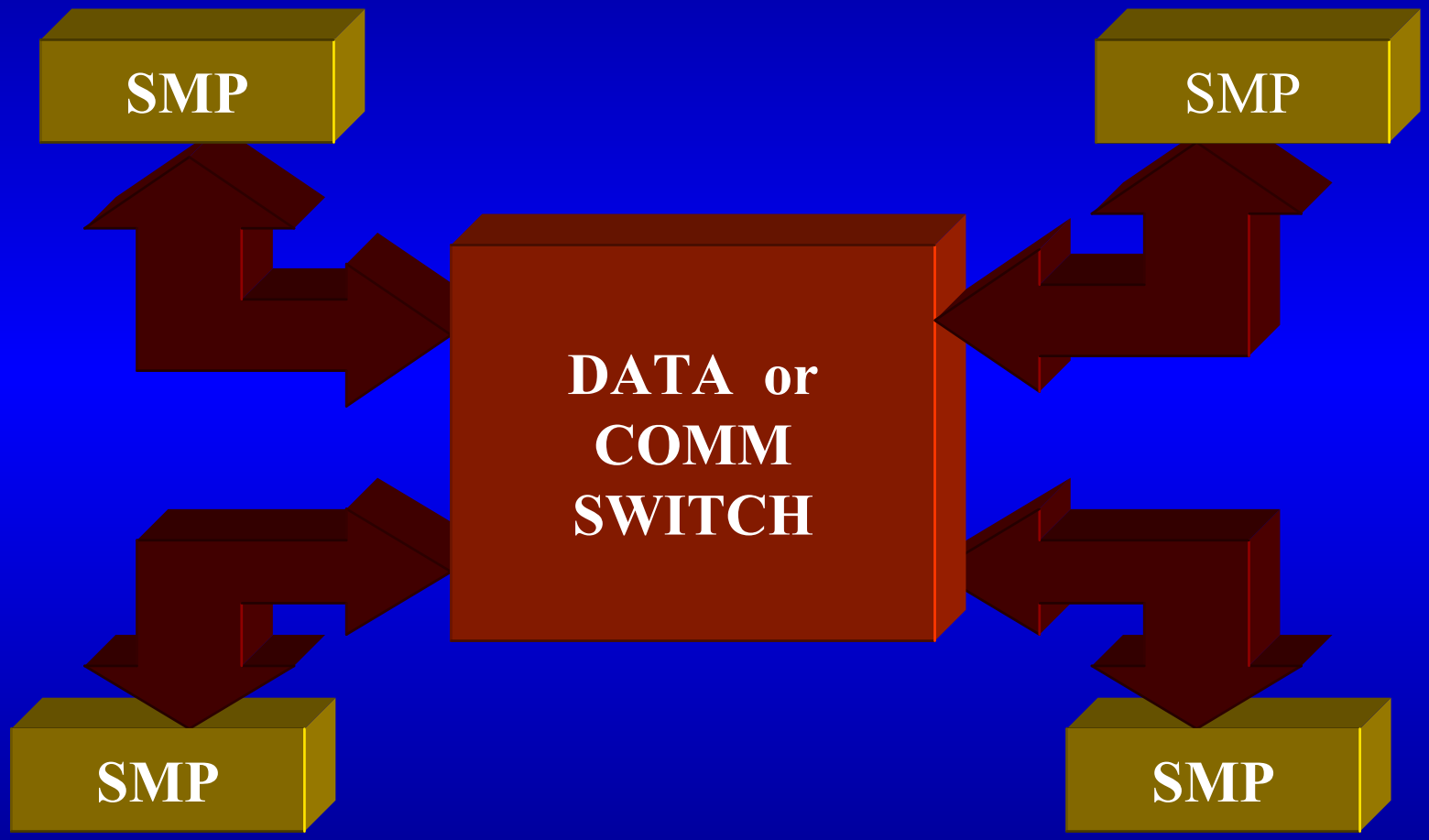
WAVELENGTH DIVISION MULTIPLEXING (WDM)

- EXPLOITS
  - ENORMOUS BANDWIDTH OF SILICA FIBER
  - HIGH-GAIN WIDEBAND OPTICAL AMPLIFIERS

$$c = \lambda * f$$



SYSTEM ARCHITECTURE



ABOVE/BELOW THE LINE

- 5 Levels (x,y,z,t,iteration)
- DO IT =

EXPLICIT - MPI, THREADS, HPF, OBJECTS

  - DO TIME =
    - DO X=
    - DO Y=
    - DO Z=

AUTOMATIC - FORTRAN/C

Reflections on Software

- Scalar → Vector
  - Vectorization, restructure
- Vector → SMP
  - Parallelization, restructure
- SMP → MPP
  - Data distribution, HPF, restructure
- SMP → Cluster of MPP
  - Data Distribution, MPI, low-latency interconnects, restructure
- Speed of Light is the ultimate limiting factor (does not scale over time).
- ***The C2K problem***

C2K Problem

- 50 Meter Radius (100 Meter Round Trip)
- Lower Bound of 500 ns Latency (Hardware - General Case)
- Thus over time, Latency gets proportionally worse
  - Same Problem exists in Processor Clock Vs. Memory Access Time.
- Solutions
  - Better Spatial locality (Bigger SMP's)
  - Spatially/Topologically Systems
    - Remember 250 Watts/Sq. Ft Cooling (air)
    - LANL 40,000 Sq. Ft => 10 Megawatts
  - Better Latency Tolerant Algorithms

SOFTRON – Particle Physics

- Hardware has
  - Electrons, photons, etc.
  - Well behaved and understood rules
- Software has
  - None of the above
  - Search for the Softron

SOFTRON – Fundamentals

- No Mass
- No Charge
  - Actually accumulative over a 5 year period
- High Energy Accelerators need to be built to determine its fundamental structure
  - Secondary priority at CERN

SOFTRON - Principles

- Half life about 20 years
- Fixing a bug is inversely proportional to the time it takes to find it
- Debugging the first 95% is easy, the second 95% is more difficult
- The closer to Redmond, WA, the more difficult to determine the softron existence

What is needed

- A realization that software is an engineering discipline by both developers and managers
- A fundamental new approach for development and verification

What Next

- Two Views (2007 vs. 2009)
  - Crawford – (presented at 2002 Microprocessor Forum
    - "Billion-Transistor Budget, A Different Kind of Real Estate Development", Oct 15.2002

http://www.intel.com/pressroom/kits/events/mpf_2002/index.htm

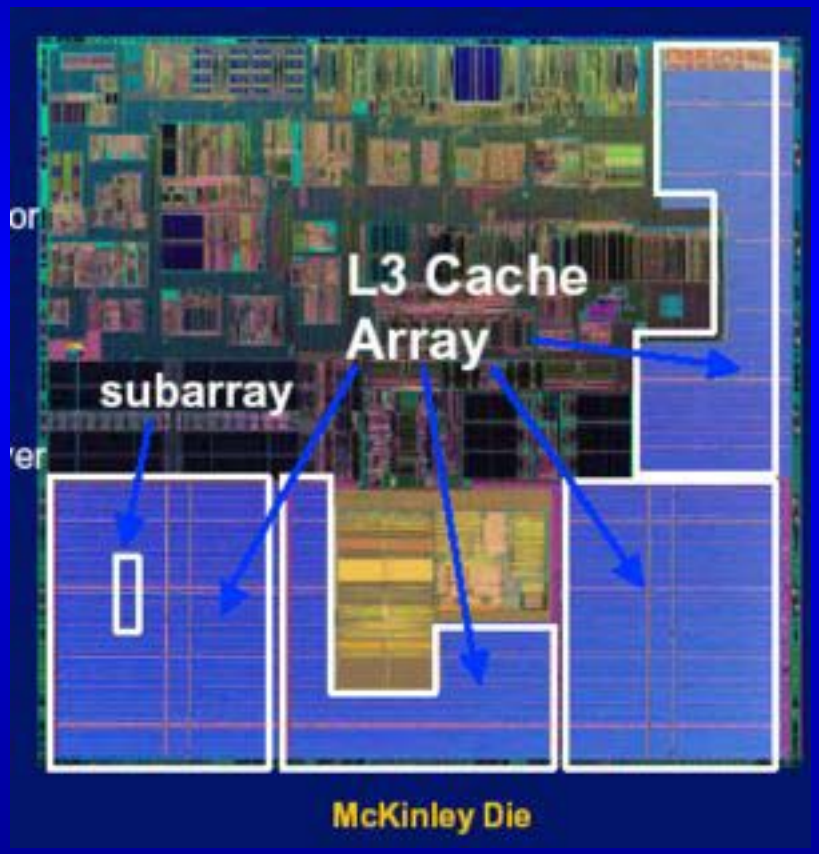
  - Wallach

- Compare/Contrast

Architecture 1

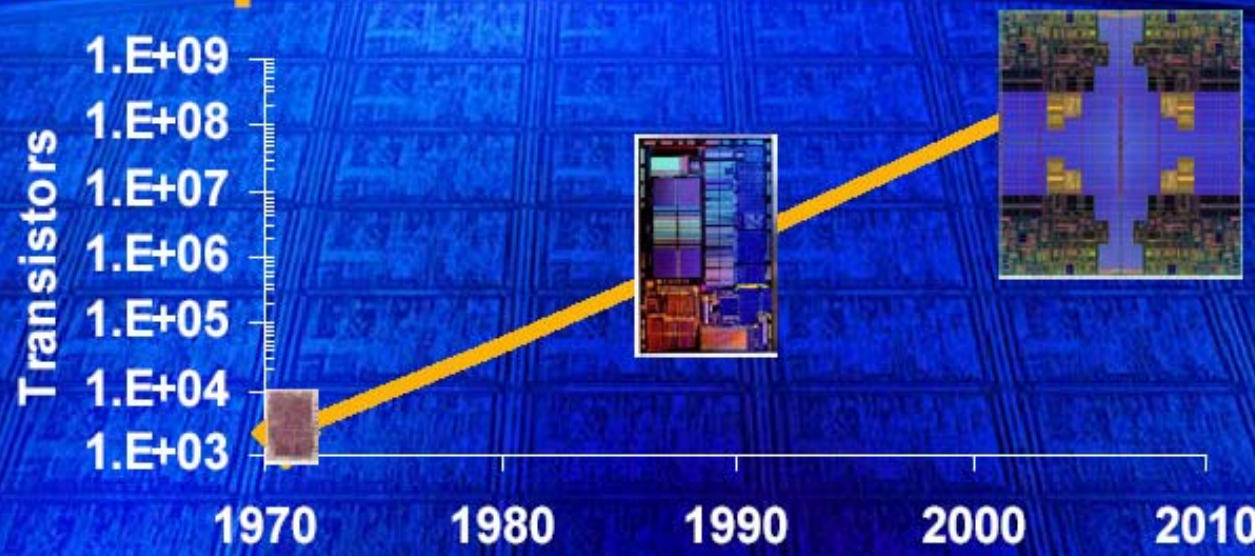
- Shared Common L3 Cache
- SMP (4 processors)
- One External Memory Bus

McKinley / IA-64



- 221 million transistors
- 3 MByte – L3 Cache
  - 175 Million Trans
  - 25 Million VLIW Core
  - 21 Million L1/L2 Cache
- 6.4 GBytes/sec I/O

Microprocessor Evolution



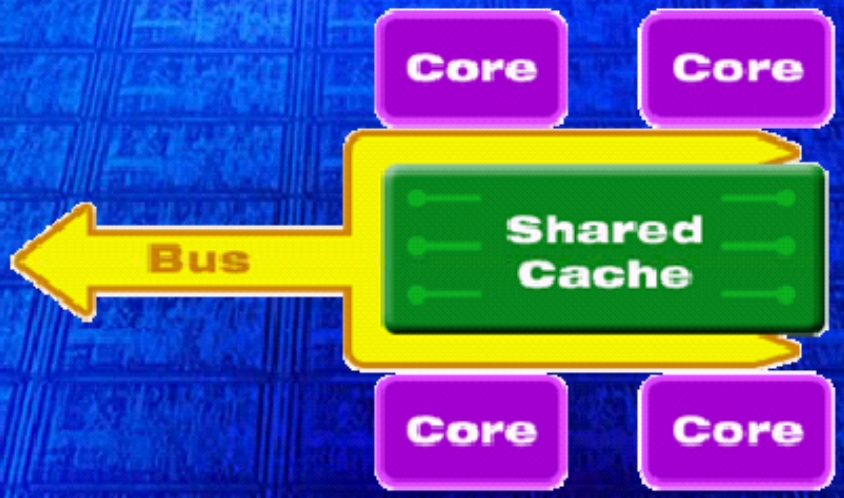
	4004	Intel® 486 processor	Future processor
Date	1971	1989 (18 yrs)	2007 (+18 yrs)
Transistor Count	2300 transistors	1.2M (500 X)	1000M (800 X)
Mfg. Process	10um process	1.0 um (1/10 X)	65nm (1/15 X)
Wafer Area	50mm wafer	100-150mm (4-9X area)	300mm (4-9 X area)
Die Size	12mm2	174mm2 (14 X)	400mm² (2.3 X)
Core Frequency	108 kHz	25 MHz (250 X)	6 GHz (240 X)

intel

Moore's Law delivers 1 Billion Transistors circa 2007

5

Multi-Core Enterprise Processor



4 Processor system on a chip, Integrating:

- 4 Itanium 2 processor Cores ~120 M transistors
- Shared Cache 12-16 MB 700-950 M transistors
- Leaf interconnect

intel

1B Transistors Can and Will be Used

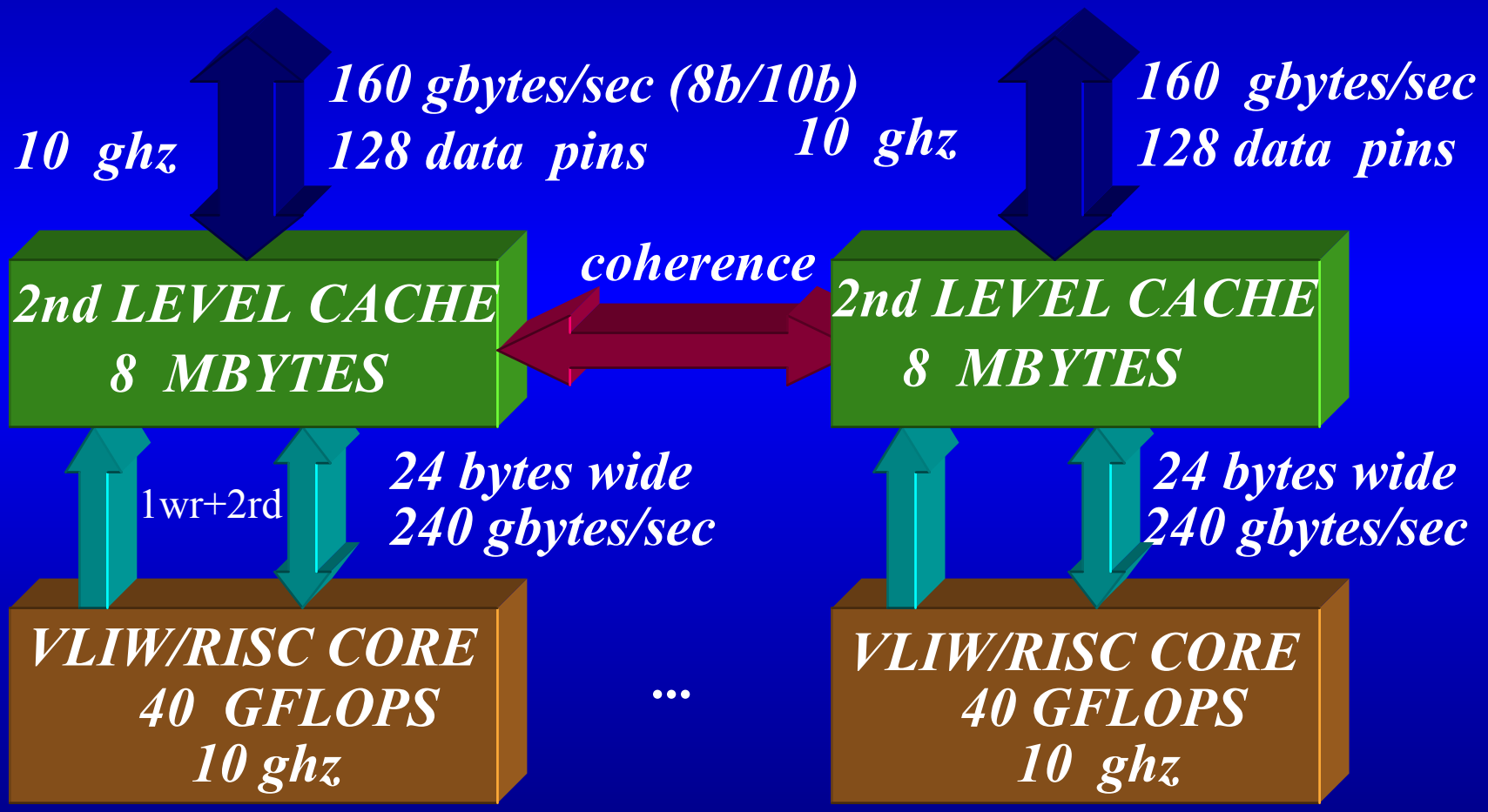
ARCH. 2- LONG TERM- 2009

- **THE ITRS STUDY TEACHES US:**
  - 10 gbits of dram - (1.25 GBytes)
    - 2.35 Gbits (DRAM)/cm^2
  - 200 Mbits of sram – 24 MBytes
  - 1200 million MPU transistors
  - 65 nm lithography, 2.00 cm on-a-side (400 mm^2)
  - 10 ghz clock within vliw/risc core
  - 10 ghz across die (CDR within chip?)
  - 1024 external signal pins

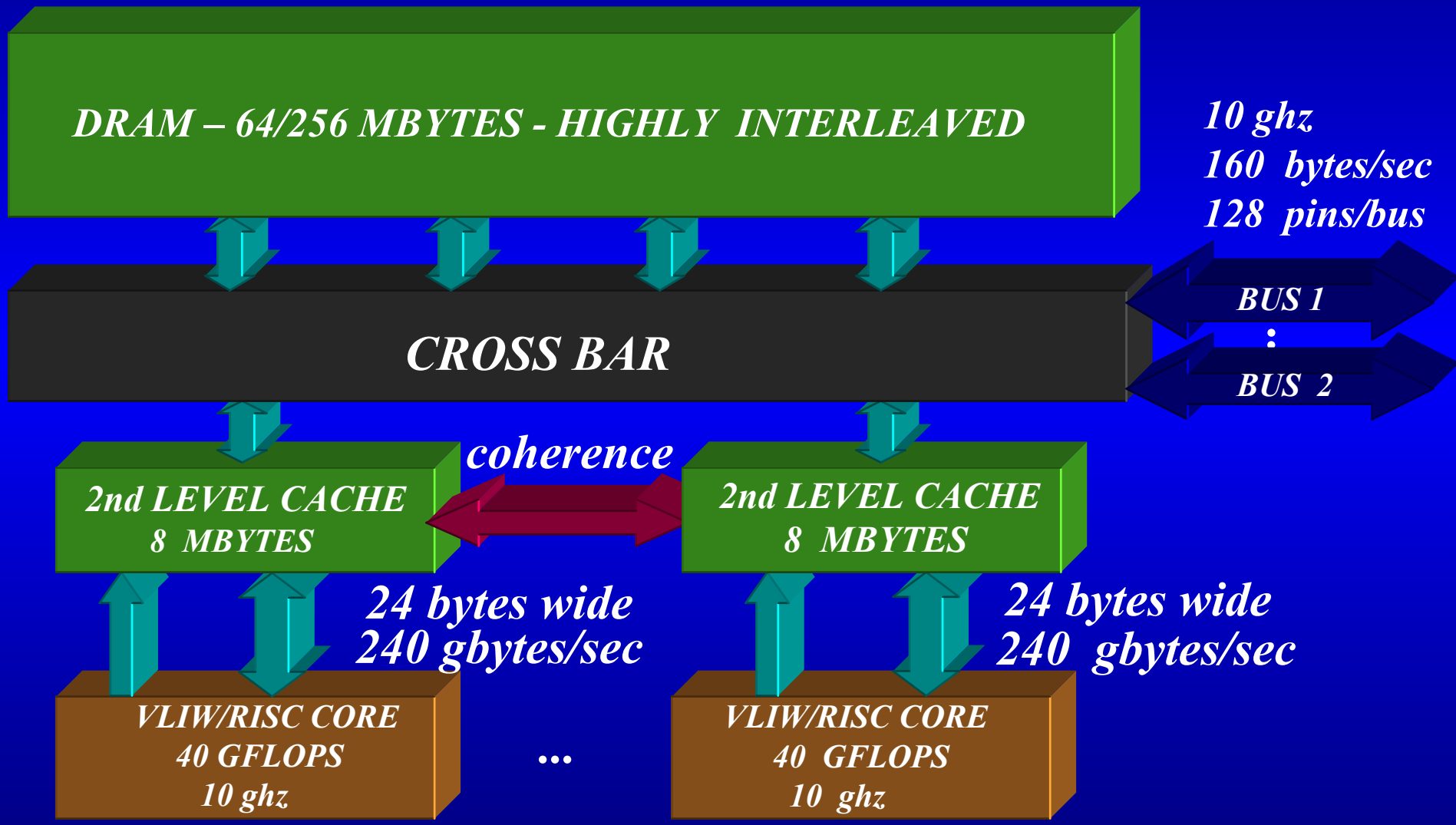
Architecture 2

- Distinct L3 Cache
- Multiple External I/O Busses
- SMP (4 Processors)

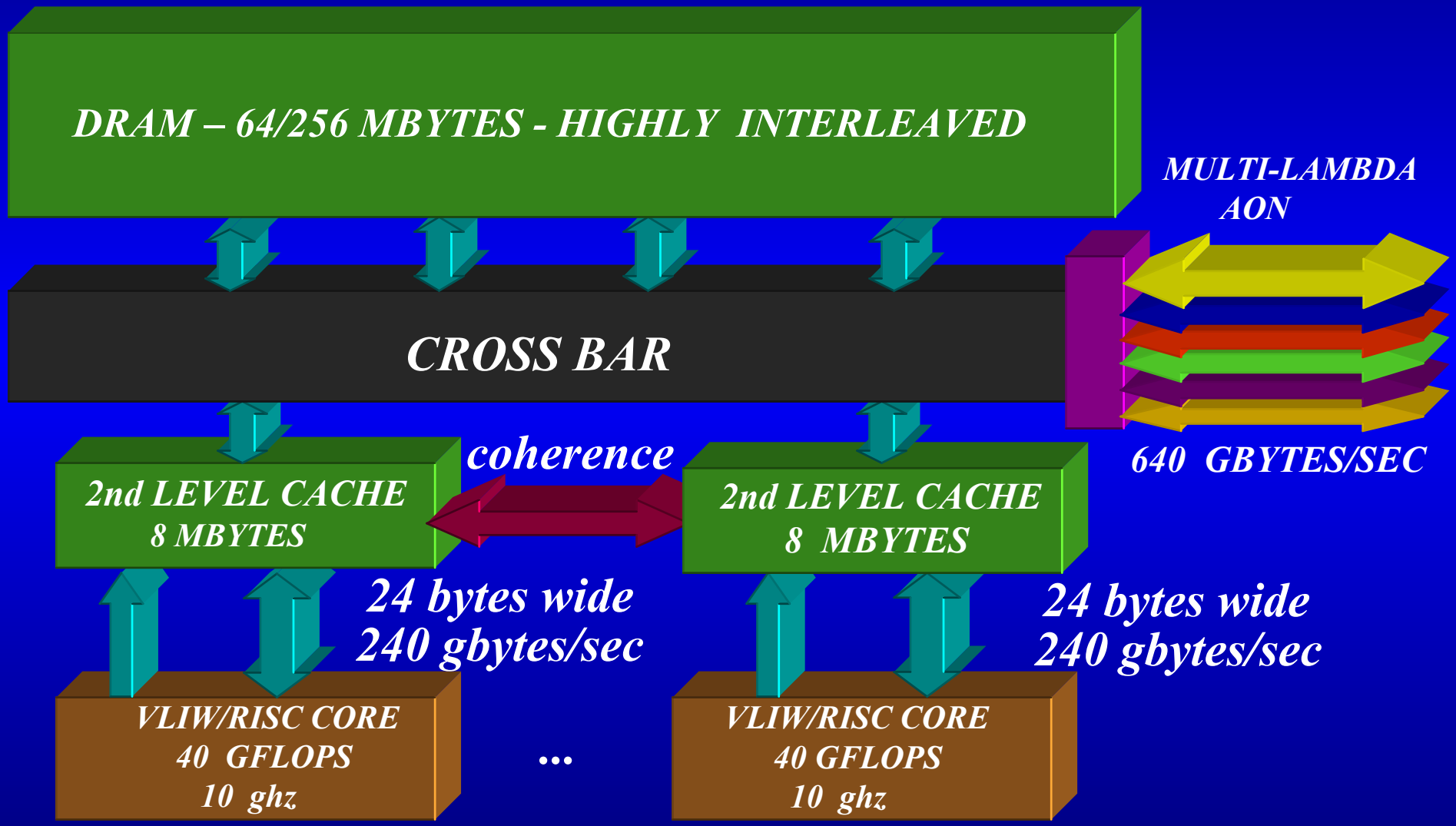
MAXIMUM PIN-USE EXTERNAL SMP- 2/4 CPU'S



INTEGRATED SMP – 2 CPU



INTEGRATED SMP - WDM



Compare & Contrast

- One Bigger Shared L3 Cache. Adaptive to dynamic Allocation of data
- Higher Latency due to cross-bar in front of Cache
- No cache coherent issues (shared cache)

- Many Smaller L3 Caches. What Cache Blocking Approach does the compiler use?
- Lower Latency – direct access from processor
- L3 Cache coherent issue

Compare & Contrast

- Physical Design may be more complex due to merging of L3 requests
- One I/O
  - Simplifies Interface Design
  - Lower Bandwidth
- Single L3 cache single point of failure

- Step and repeat cpu/cache hierarchy
- Potential Multiple I/O
  - More Pin Out (Can this be implemented?
  - Higher Bandwidth
- More Fault Tolerant

Sobering Calculation (Common to Both Architectures)

- 100 Gflops Peak Performance (4 cpu's)
- 400 GBytes/sec I/O Desirable (Real *8 Saxpy)
- Perhaps 80 GBytes/sec Achievable (ratio =1/5). McKinley is approx .4 = (6.4 GB/s /(4 Gflops x 4 B))
- Yields 640 Gbits/sec
- 64 10 Gbit/sec CDR Interfaces(?) (Point to Point Interface – NOT BUS)
  - Multi-Level Signaling for greater than10 Gbits/sec(?)
  - Can't share bus with CDR signaling
- What does main memory system look like? (with inputs from multiple die – each input 64 CDR interfaces)

COTS PetaFlop System -2009

- 8192 Dies (4 CPU/die-minimum)
- Each Die is 160 Gflops (external memory)
- 1.3 PetaFlop Peak
- Power 8192 x200 Watts = 1.6 MegaWatts
- Main Memory yields in excess of 3 MegaWatts (512 TBytes)

COTS PetaFlop System

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

COTS PetaFlop System

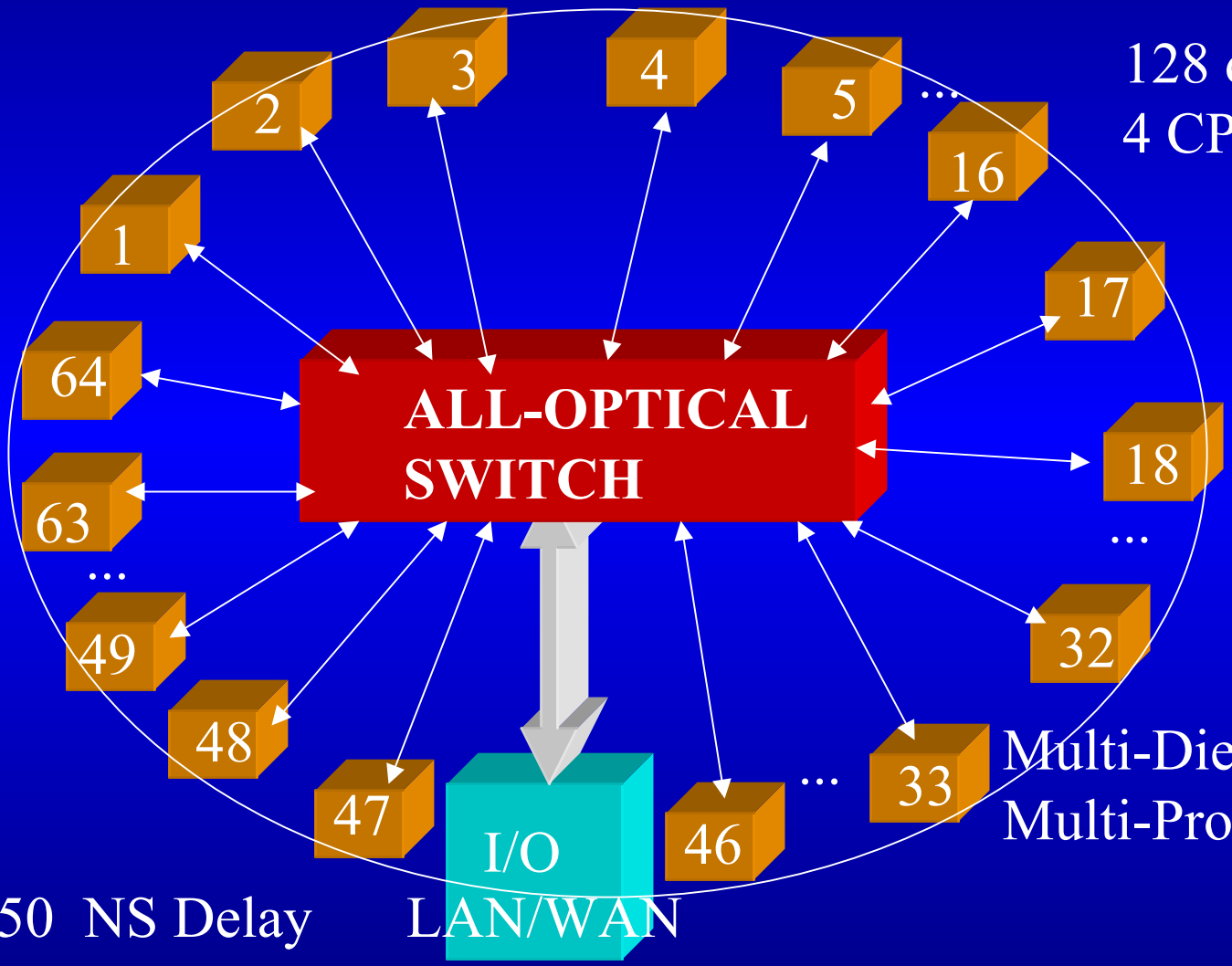
- Optical Interconnect
  - OC768 Channels (40 GHz)
  - 128 Channels per Fiber (DWDM)-5.12 THz
  - 2004/2005 Commercial Transmission
  - ALL Optical Switching
- Bisection Bandwidth of 128 TBytes/sec (1 Pbit > 1 PFlop)
  - 20 TFlops/rack*.1bytes/flop/sec*64 racks
- Rack Bandwidth - 20 TFlops*.1= 2 TBytes/sec (full duplexed)
- 3-4 OPTICAL MT-RJ Connectors or equivalent per rack in each direction (3 x 5.12 or 4 x 5.12)

A Ratio of .2 Bytes/flop/sec is desirable

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



128 die/box
4 CPU/die

Multi-Die
Multi-Processor

10 meters= 50 NS Delay

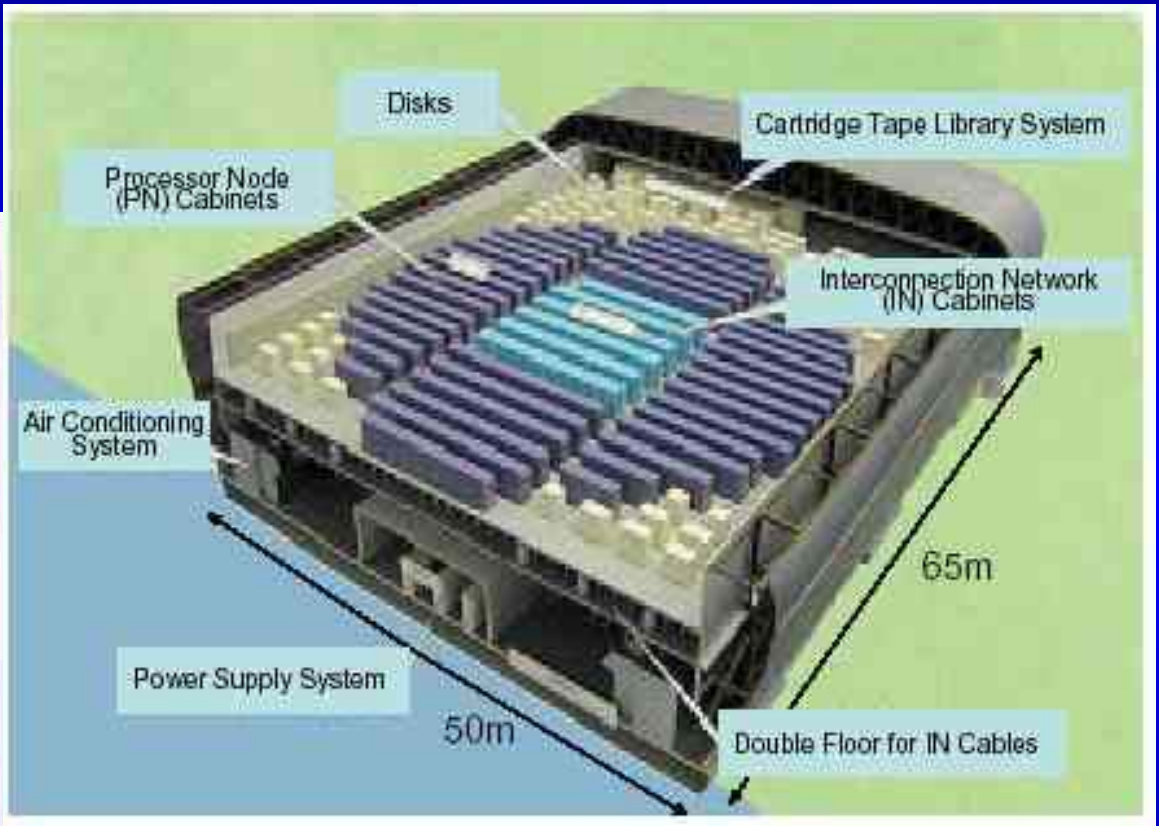
A Computer Room

SCC @ Los Alamos

40,000 sq. feet (4,000 sq. meters)

A Computer

Another Computer



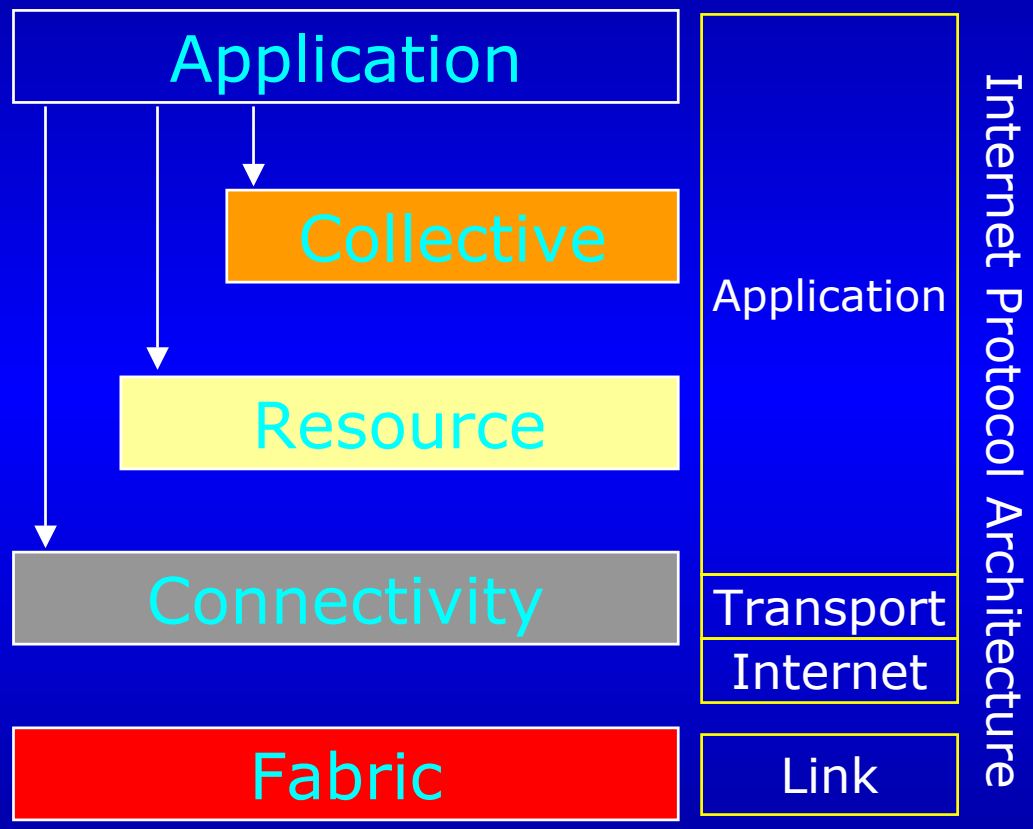
Layered Grid Architecture

"Coordinating multiple resources": ubiquitous infrastructure services, app-specific distributed services

"Sharing single resources": negotiating access, controlling use

"Talking to things": communication (Internet protocols) & security

"Controlling things locally": Access to, & control of, resources



Source: NSF Teragrid

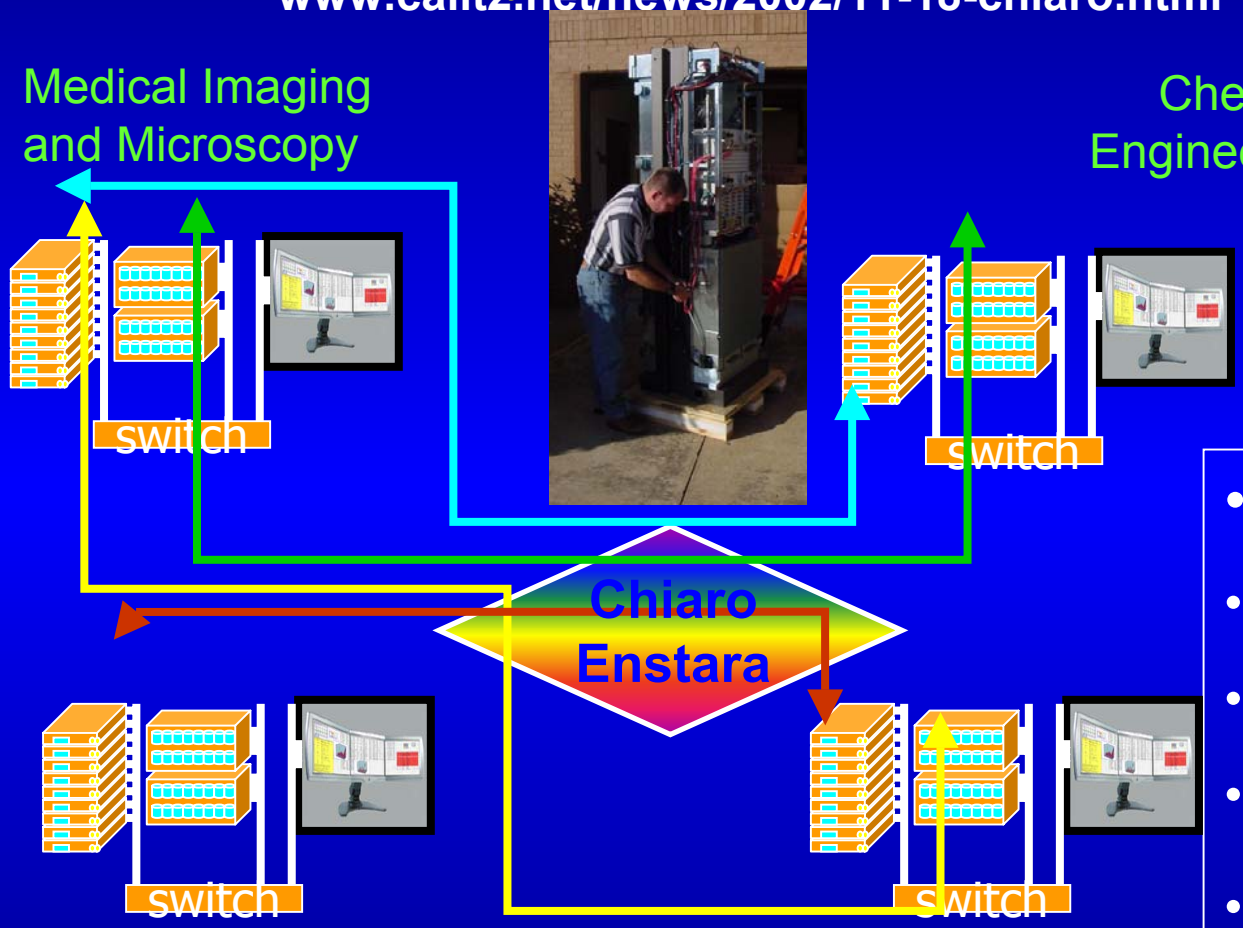
Building Something New

One Organization (merge institutions)	The TeraGrid (a Grid hosting environment)	Loose Collaboration (current situation)
▪ One sysadmin team ▪ One mgmt team ▪ Distributed machine room, centralized control ▪E.g. Google data centers	▪ Single devl env. ▪ Single stack to learn ▪ Develop here, run there ▪ Run here, store there	▪ Different MPIs ▪ Hit-and-miss grid software: ▪Globus version? ▪Condor-G? ▪G2? ▪ Unique devl env.
Not a Grid	**Applications are developed for the Grids because the barriers are low and the return large**	**Not a Grid, but with significant user investment, Grid apps can be developed**

Pete Beckman <beckman@mcs.anl.gov>

UCSD OptIPuter LambdaGrid Enabled by Chiaro Networking

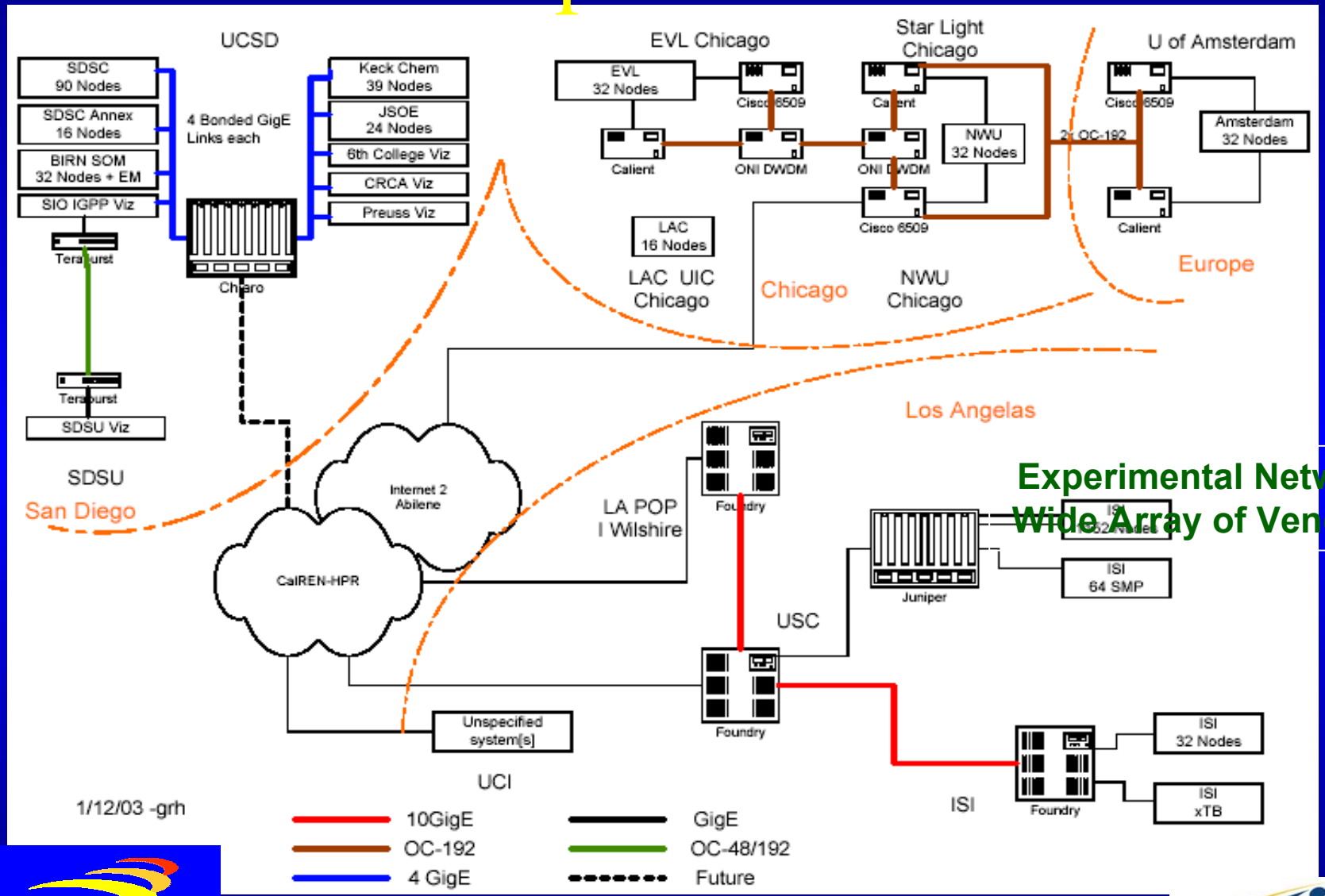
www.calit2.net/news/2002/11-18-chiaro.html



- Cluster – Disk
- Disk – Disk
- Viz – Disk
- DB – Cluster
- Cluster – Cluster

Image Source: Phil Papadopoulos, SDSC

The OptIPuter 2003



Experimental Network
Wide Array of Vendors

CONCLUDING

- David Nolte, “Mind at Light Speed: a New Kind of Intelligence”; Free Press, Nov. 2001

- “..will the silicon transistor go the same way of the vacuum tube”

Concluding

SAN FRANCISCO, Oct. 27 — STMicroelectronics plans to announce a breakthrough on Monday in light-emitting silicon that could lead to a new generation of more powerful computing processors and more efficient automobile components as well as potentially higher-speed optical data-transmission systems.

New York Times; John Markoff