DarkHorse
a Proposed PetaScale Architecture (+)

Steve Poole

Los Alamos National Laboratory

Oak Ridge National Laboratory

HPEC06

September 19-22, 2006

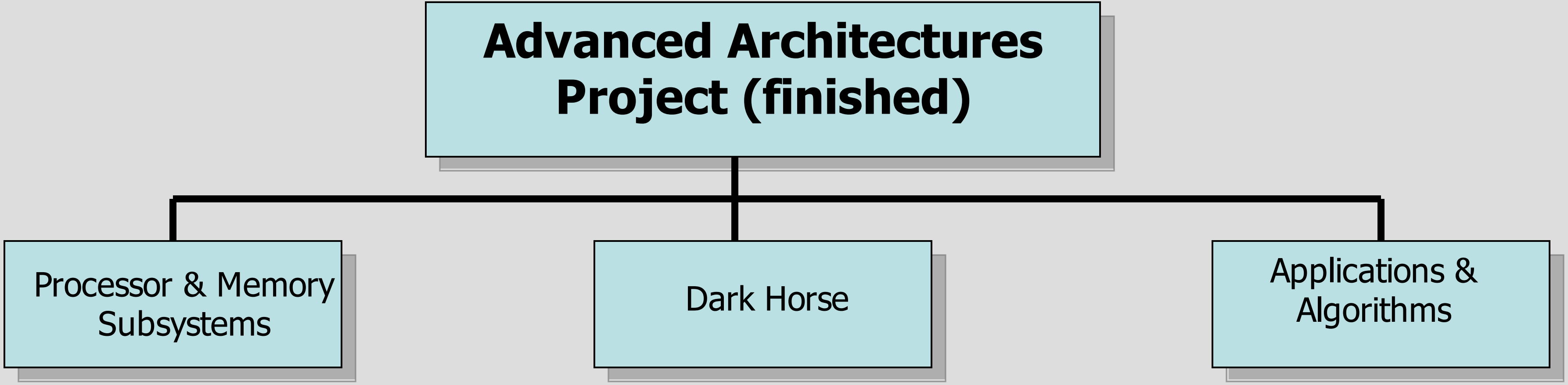
LA-UR-06

Advanced Architecture Team (LANL)

- LANL
  - Dave DuBois
  - Andy DuBois
  - Steve Poole
  - Chris Kemper

Some History of DH & 3D

- First basic ideas in 1997/1998
- HMM/GA Application (Kestrel, Sequence Alignment Modeling)
- Switch Application (SanNetworks, memory technology)
- 3D FPGA
- Potential Seismic Application (FD,RTM, A/E Modeling,XON)
- Specialized Search/Sort Problem (DB Problem)
- Started @ LANL 2001
  - 3D FPGA
  - 3D CAM
- Early processor disclosures in 2002



Advanced Architectures Project (finished)

Processor & Memory Subsystems

Dark Horse

Applications & Algorithms

- **Computer industry collaborations**
  - Understand and influence product roadmaps
- **Semiconductor industry collaborations**
  - 3D semiconductor stacking
- **Co-processor technologies**
  - FPGA accelerators
  - Graphics/Network processor accelerators

- **Determine the feasibility of developing a PF system in the ~FY08 time frame that is:**
  - based potentially on a variety of microprocessors,
  - computationally efficient for LANL algorithms, and
  - straightforward to program.
  - Balanced
  - First Principle

- **Minimizing time to solution for LANL computational workloads**
  - Adapt algorithms to different architectures
  - Develop new algorithms that take maximum advantage of computer architectures
  - Programming model(s)

P(FL)OPS Project Phases, 1-5 Years

- Initial Studies, finished (ARM, BCom, SPARC-8)
- **HW Phases:** FPGA/CAM Proof of Concept
  SOC design in FPGA
  Optical Switch Development
  Integrated Optics on Processor Stack
  Processor based SOC, ~1 TFLOPS, 128/256 GByte
  Initial Prototype
  Not just PIM, true SOC
- **SW Phases:** System Software, Middleware, SGPFS, Data Migration,
  Communication, Scientific Libraries, VIZ, Applications,
  Hybrid Programming Model, OS
- **Lvl of Effort:** 5 years Development for: (Not Free, now 3 years)
  3-D PE = 1-10 TF + 128/256/1024 GByte + Communications
  Optical/Copper Switch Fabric
  Software (Lots of work)
  Hybrid + Languages + OS

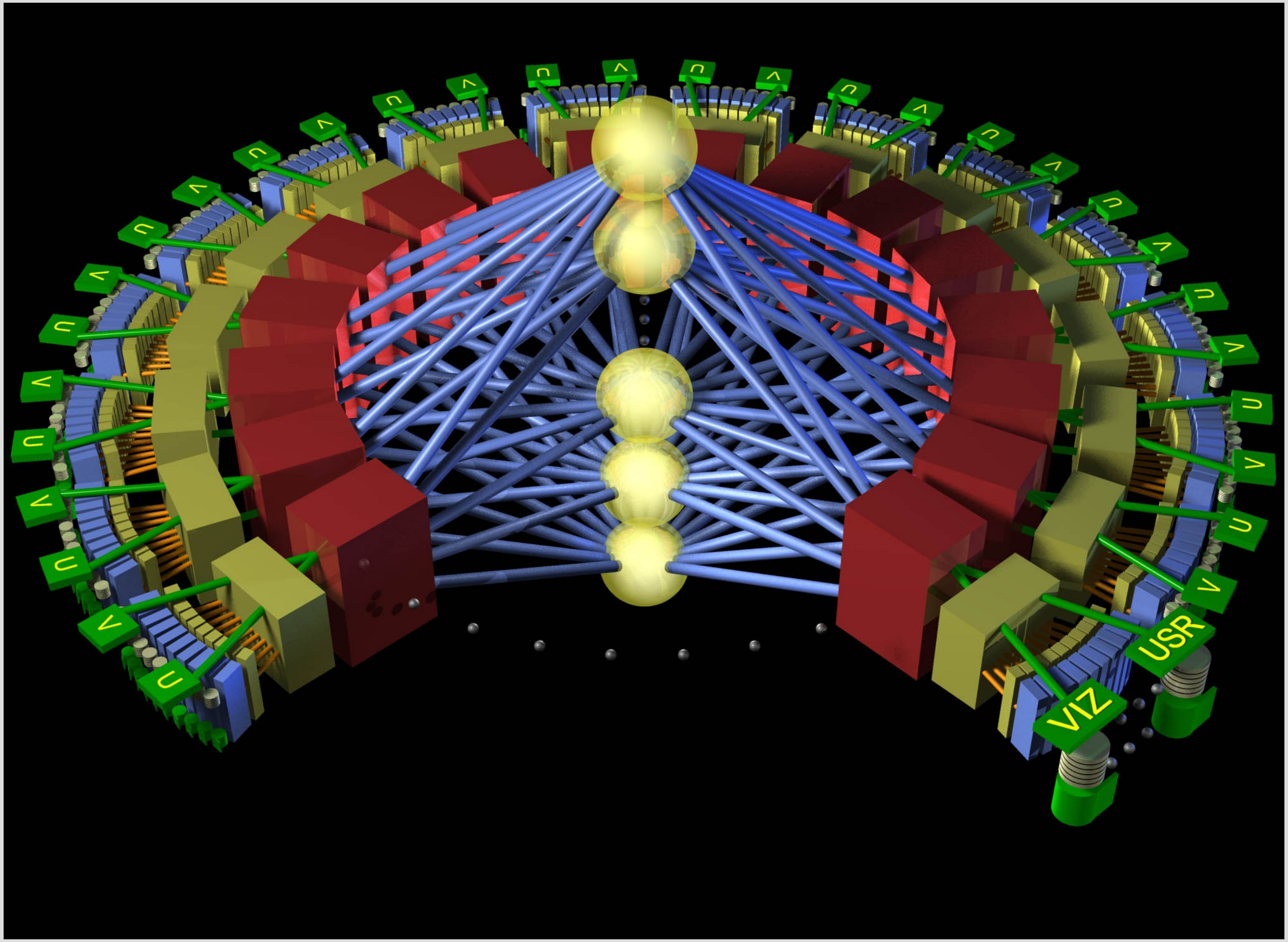
PFLOPS Advanced Architecture Design, Hardware and Software

Hardware	Elements	Technology	Software
Computing: 1K PEs, 1 PF=10^{15}FLOPS, 100 TB=10^{14}Byte	PE = System On Chip, "X" GHz, ~200 Functional Units, MultiCore, 128/256GB, Self healing, 4-8TB/s Vertical Internal BW, 1TF (128-256TB M)	3-D Stacking with N Layers/Split	OS: • Linux (64 bit), • K42 (multi-processor 64 bit all the way), • Plan9 (64 bit, secure) Programming Models: • asymmetric MP model, • function off-load models: 1) multi-stage and 2) parallel-stage, • computational acceleration models: 1) in core and 2) out of core scalar and vector parallelization
Communication: Compute Ratio B/F=0.2, Disk I/O Ratio B/F=0.1	PE-PE BW = 20 x 10GB/s PE-Blade BW = 10 x 10GB/s Blade/PE = 100	Optical/ Copper Switch Fabric	System Integration = Middleware

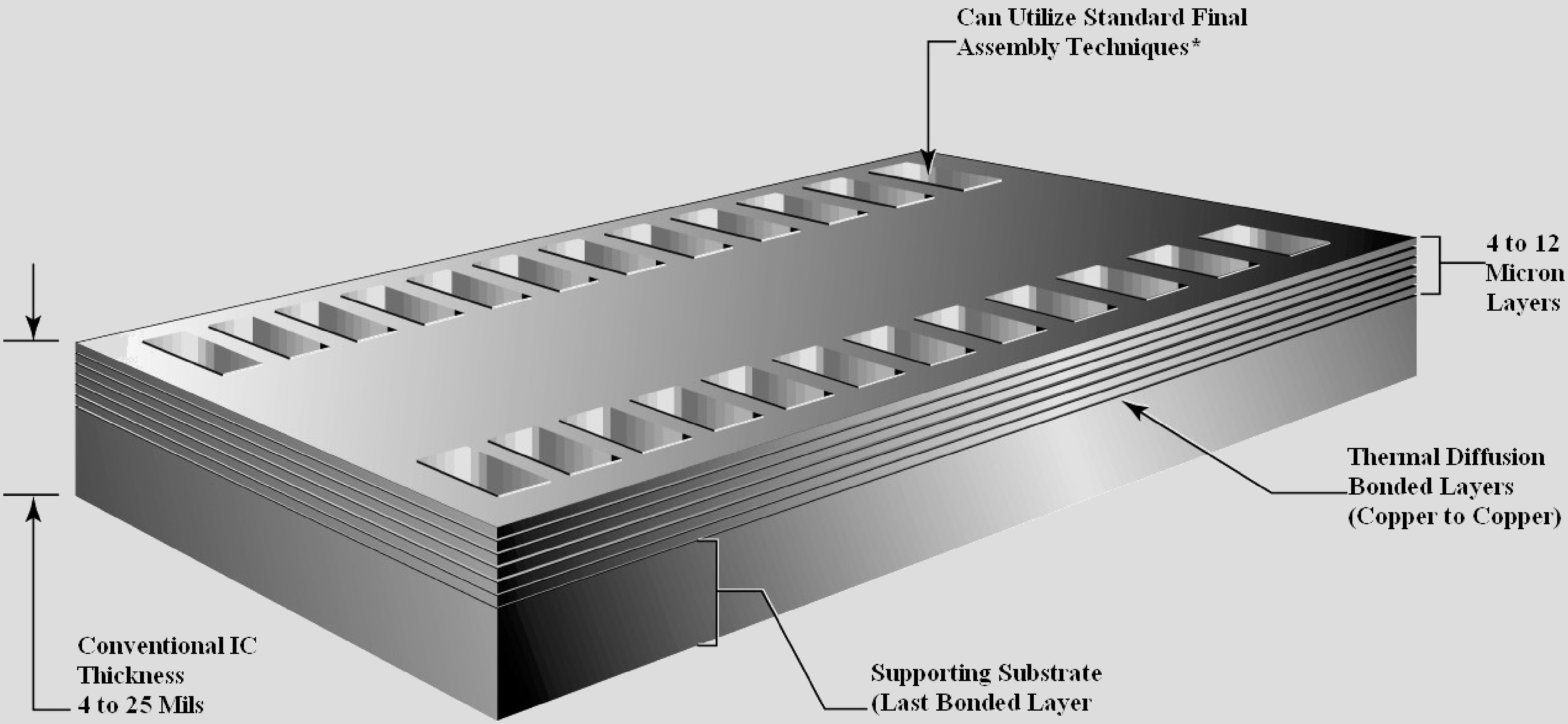
Data Storage:
Blade=10^5, Disk=10^{17}B,
Tape=10^{18}B
100PB
	Blade = 1 TB capacity, 1GB cache, 100 MB/s sustained BW	NASD OBSDs	Scalable Global Parallel File System,
			OBSD Disk to Tape Migration SW

Visualization:	GPU = General Purpose Unit	3-D on Chip	OpenGL (or successor) based Surface and Volumetric Rendering Software

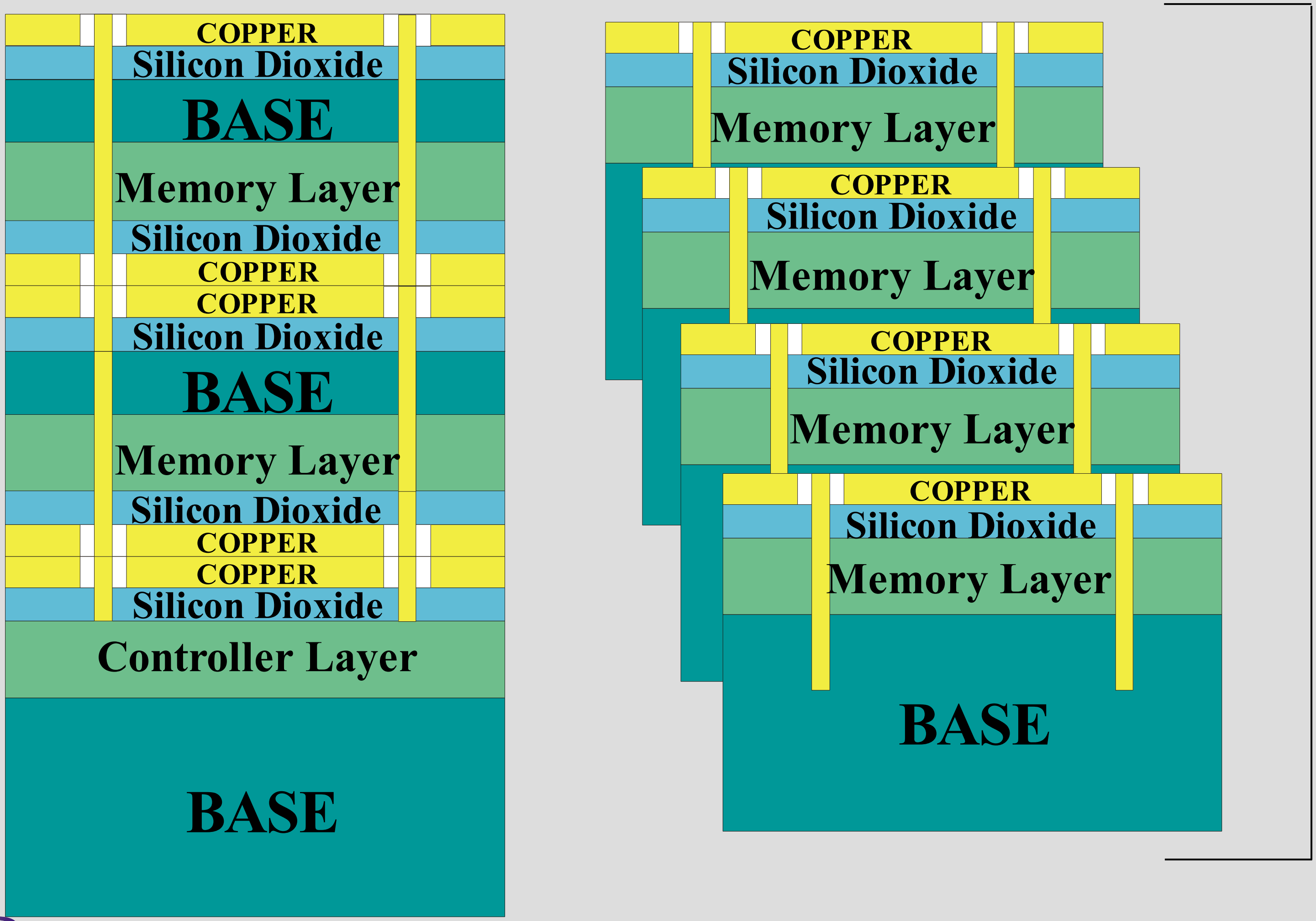
Supercomputer Interconnect Example



Building 3-D Chips



Stacking Multiple Thin Layers

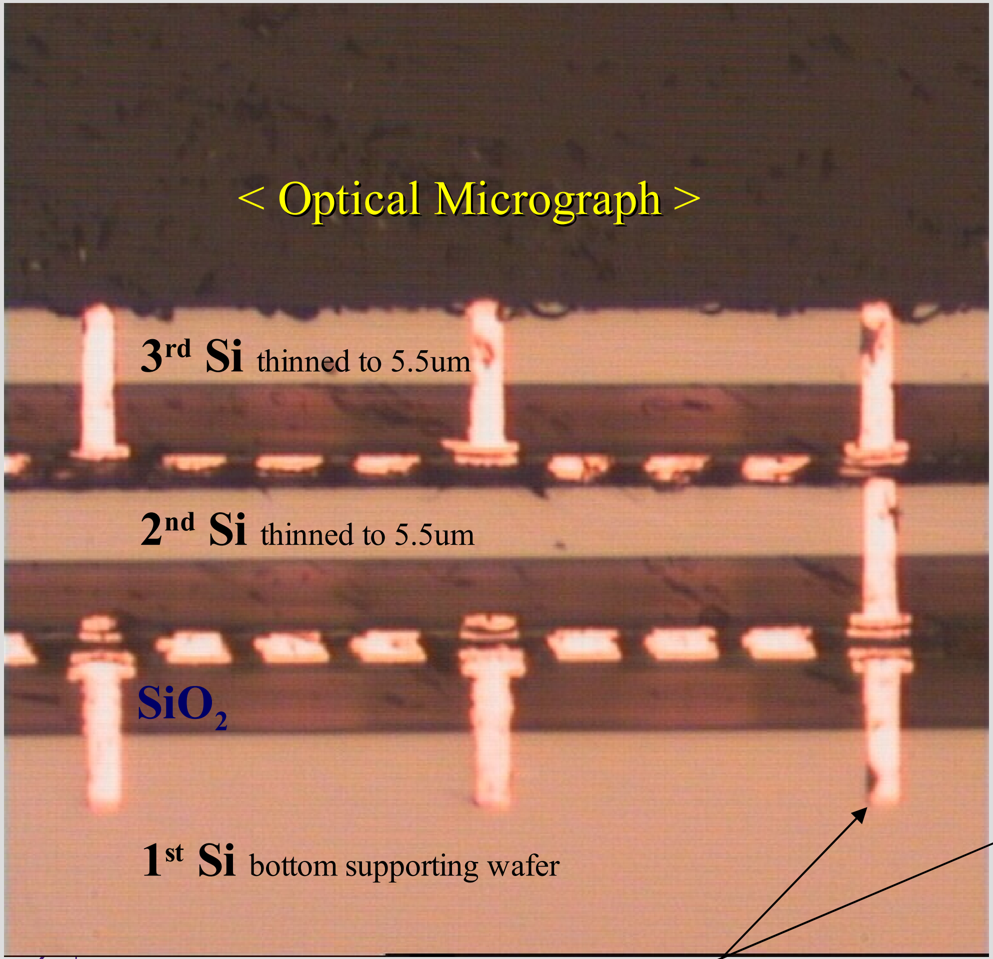


Additional Memory Layers to be stacked

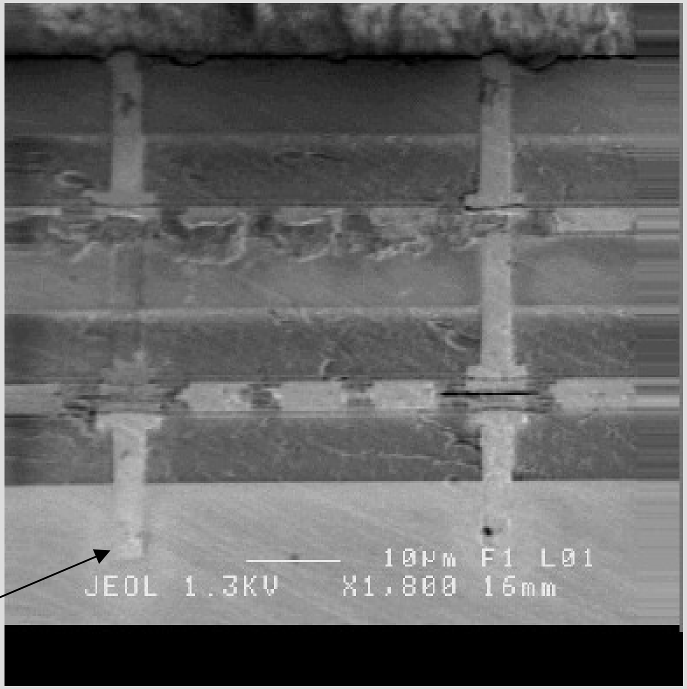
Repeat - One Wafer at a Time

Stacking Process

Three wafers successfully aligned and stacked

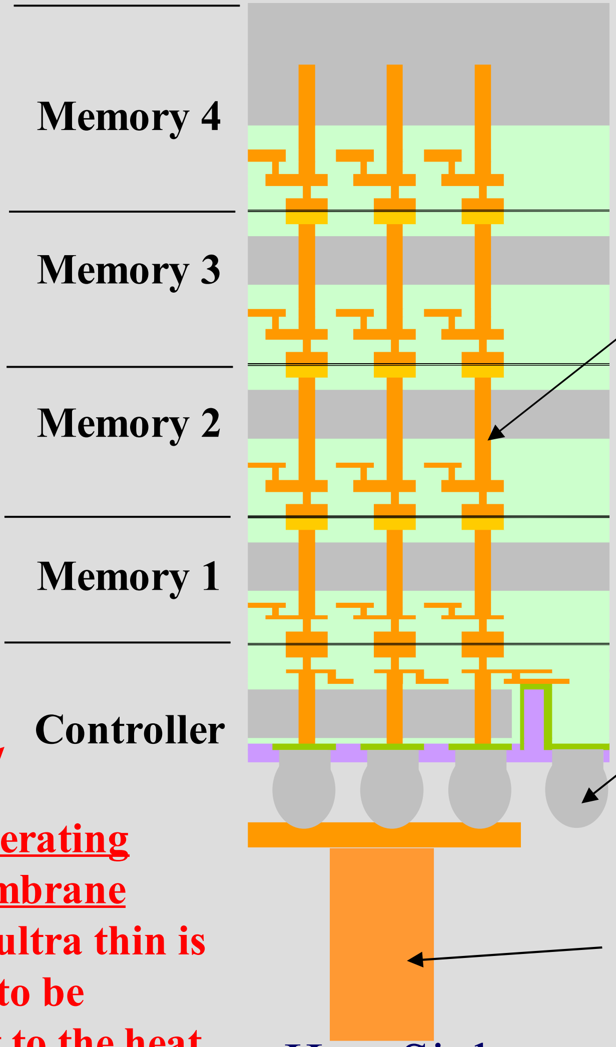


< Scanning Electron Micrograph >



"Super Via" 4um in diameter and 12um in height

"FaStack" Cross-Sectional Diagram



Super-via acting as thermal via

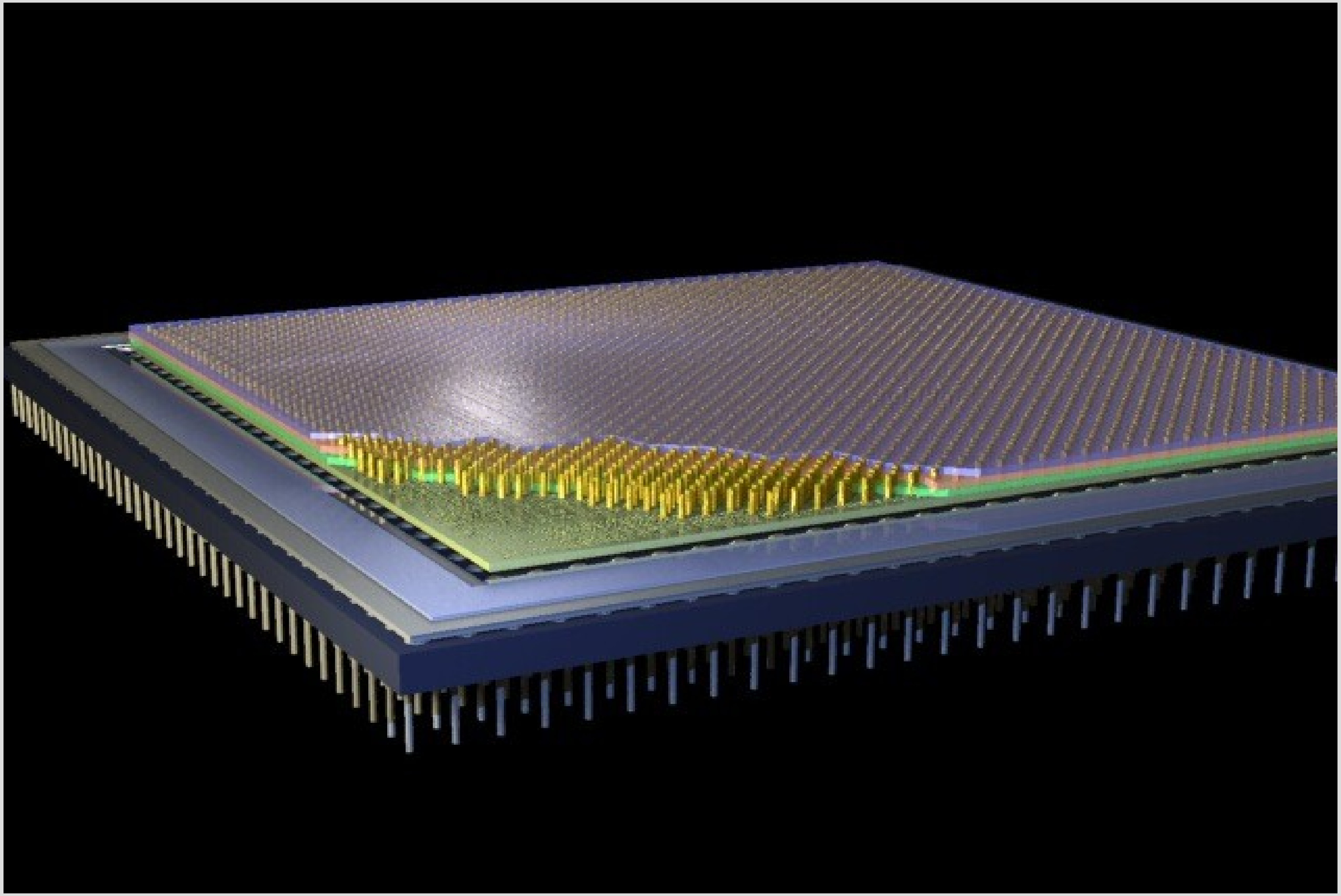
BGA

Package level thermal via

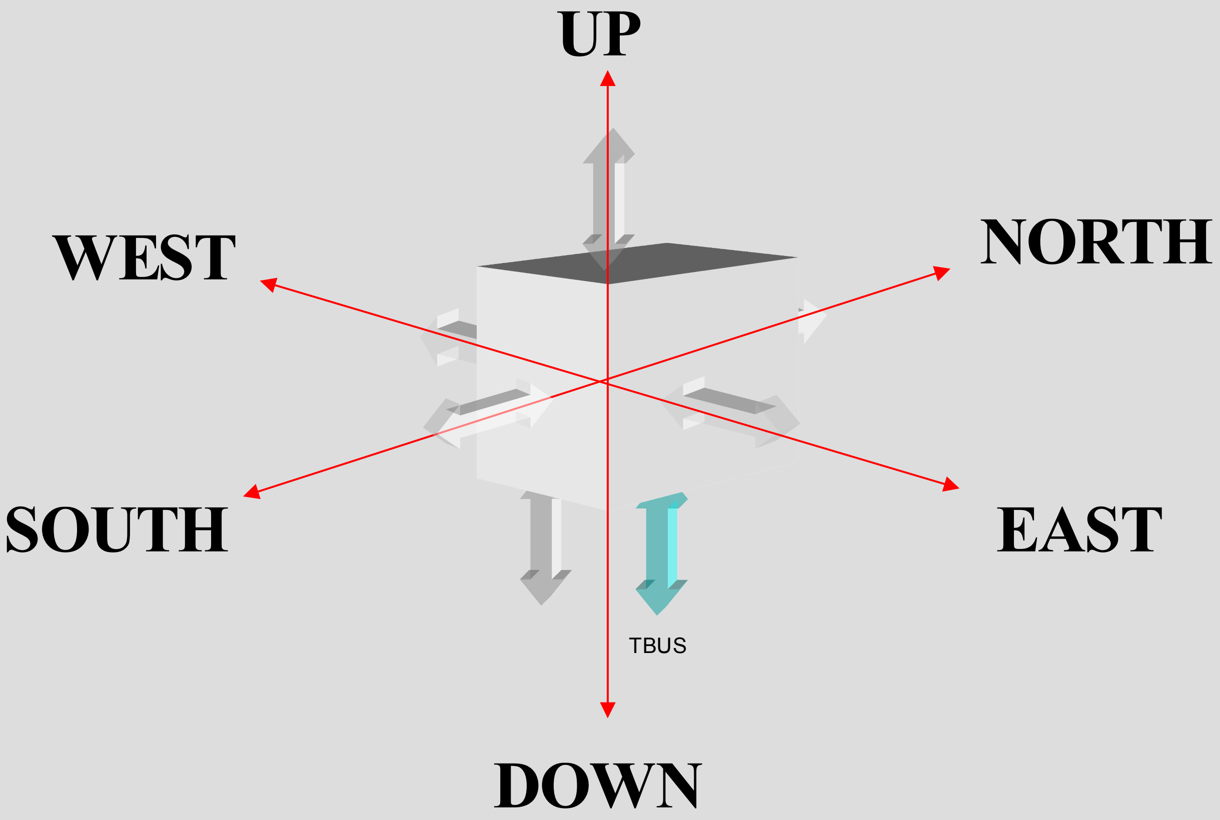
Heat Sink

Most heat generating controller membrane besides being ultra thin is also designed to be located closest to the heat sink

Stacked Chip Interconnect/Thermal Paths



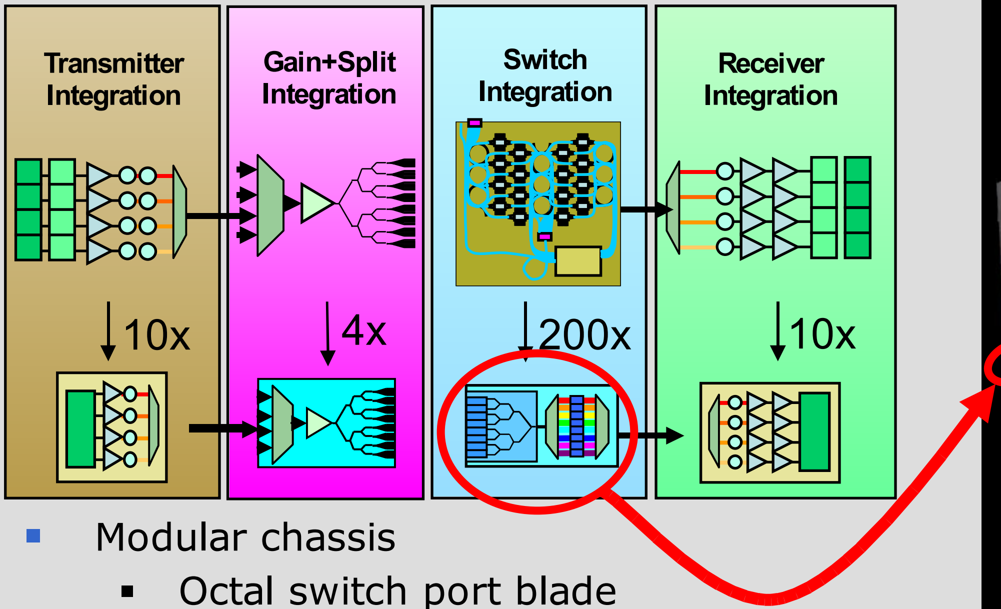
Interlayer Interconnnect



Interlayer Interconnect

- Minimum interlayer delay
- Flexible width and routing
- One-to-many or many-to-one
- Complex topologies feasible
- Designed and delivered to LANL 2003

- Today ~$50K/port →
- ~$1.5k/port for commercial

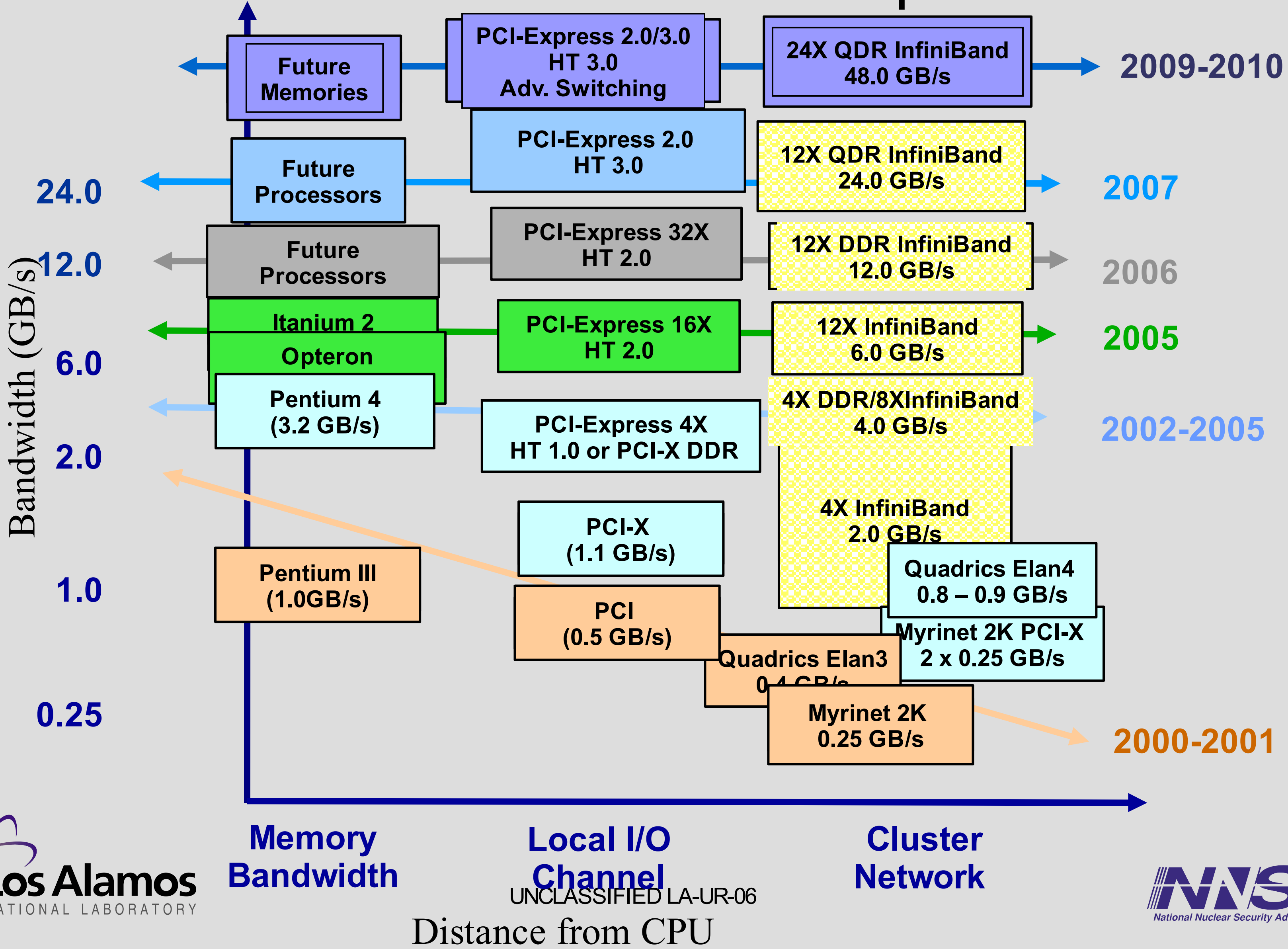


- Modular chassis
  - Octal switch port blade
- Integration achieves
  - **200:1 Parts count reduction**
  - **10X Power reduction**
- Provisioning in 16 port increments

10 Terabit/sec form factor

•1.28 Gigapackets/sec in 64 port switch module
•Cell-oriented error correction supports 10^{-21} BER
•Goal: 10 Tbit/sec in a single stage module @ first commercial release

InfiniBand Roadmap



2009-2010

Future Memories

PCI-Express 2.0/3.0
HT 3.0
Adv. Switching

24X QDR InfiniBand
48.0 GB/s

2007

Future Processors

PCI-Express 2.0
HT 3.0

12X QDR InfiniBand
24.0 GB/s

2006

Future Processors

PCI-Express 32X
HT 2.0

12X DDR InfiniBand
12.0 GB/s

2005

Itanium 2

Opteron

PCI-Express 16X
HT 2.0

12X InfiniBand
6.0 GB/s

2002-2005

Pentium 4
(3.2 GB/s)

PCI-Express 4X
HT 1.0 or PCI-X DDR

4X DDR/8XInfiniBand
4.0 GB/s

4X InfiniBand
2.0 GB/s

PCI-X
(1.1 GB/s)

Quadrics Elan4
0.8 – 0.9 GB/s

Myrinet 2K PCI-X
2 x 0.25 GB/s

2000-2001

Pentium III
(1.0GB/s)

PCI
(0.5 GB/s)

Quadrics Elan3
0.4 GB/s

Myrinet 2K
0.25 GB/s

CPU/Memory Stack

- R8051 CPU
  - XXX MHz operation; 140MHz Lab test (VDD High)
  - 220MHz Memory interface
- IEEE 754 Floating point coprocessor
- 32 bit Integer coprocessor
- 2 UARTs, Int. Cont., 3 Timers, …
- Crypto functions
- 128KBytes/layer main memory
- Codes Running…

- Completely synthesized, placed and routed in 3D with standard Cadence tools. Runs slightly better than predicted by models and tools. We are working with this currently.

3D FPGA’s

- Why 3D?
- Why FPGA’s?
- Technology Elements
- Project Structure
- Project Objectives
- Cost and Schedule

Why 3D?

- 2D chip performance is limited by chip size
  - Large chip size means defects, wire delay
- 2D function is limited by chip size
  - Even with better lithography, there’s never enough room for memory
  - I/O is a perpetual problem
- 2D systems require high package count
  - Board real estate is precious for ALL applications

3D ADDRESSES ALL THESE ISSUES!

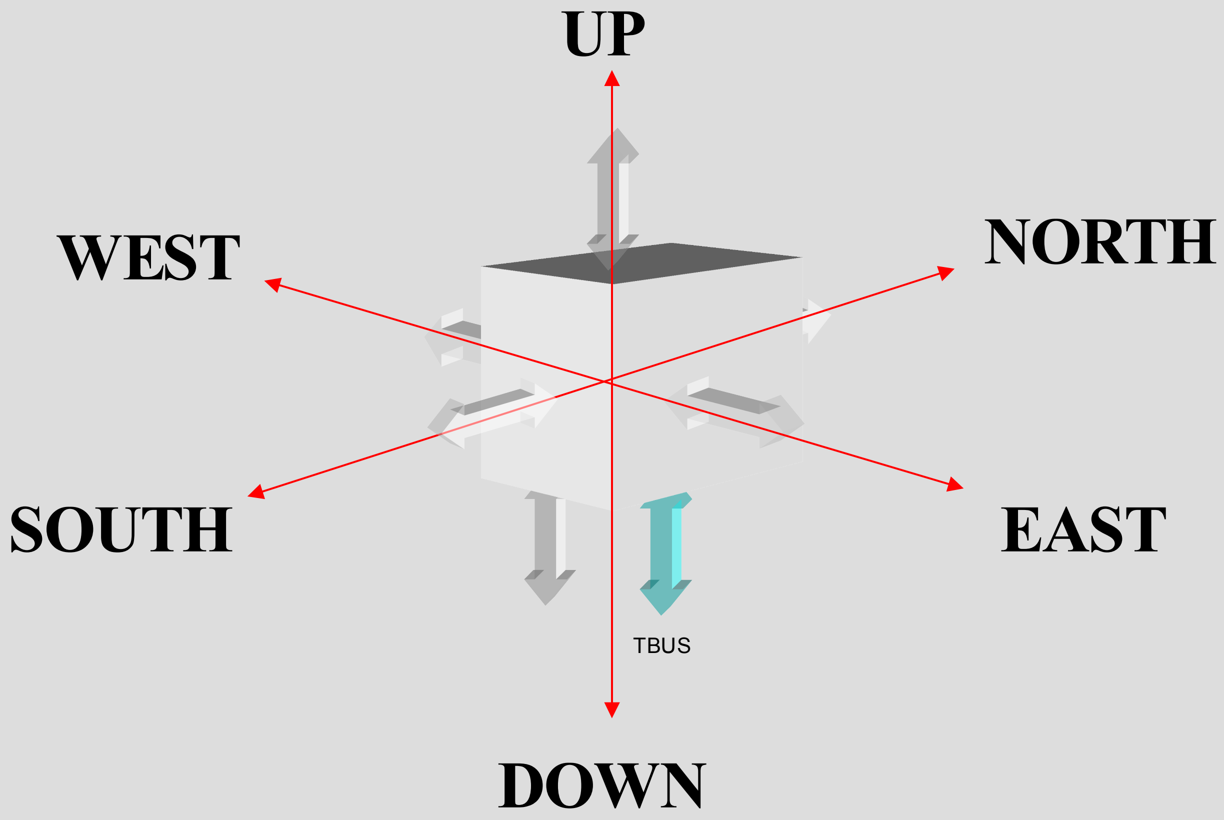
Why FPGA's?

- ASIC's and Processors are too expensive to prototype
- Present FPGA's are too slow
- Better potential for rad hardening
- Properly designed 3D FPGA can implement complete SoC.

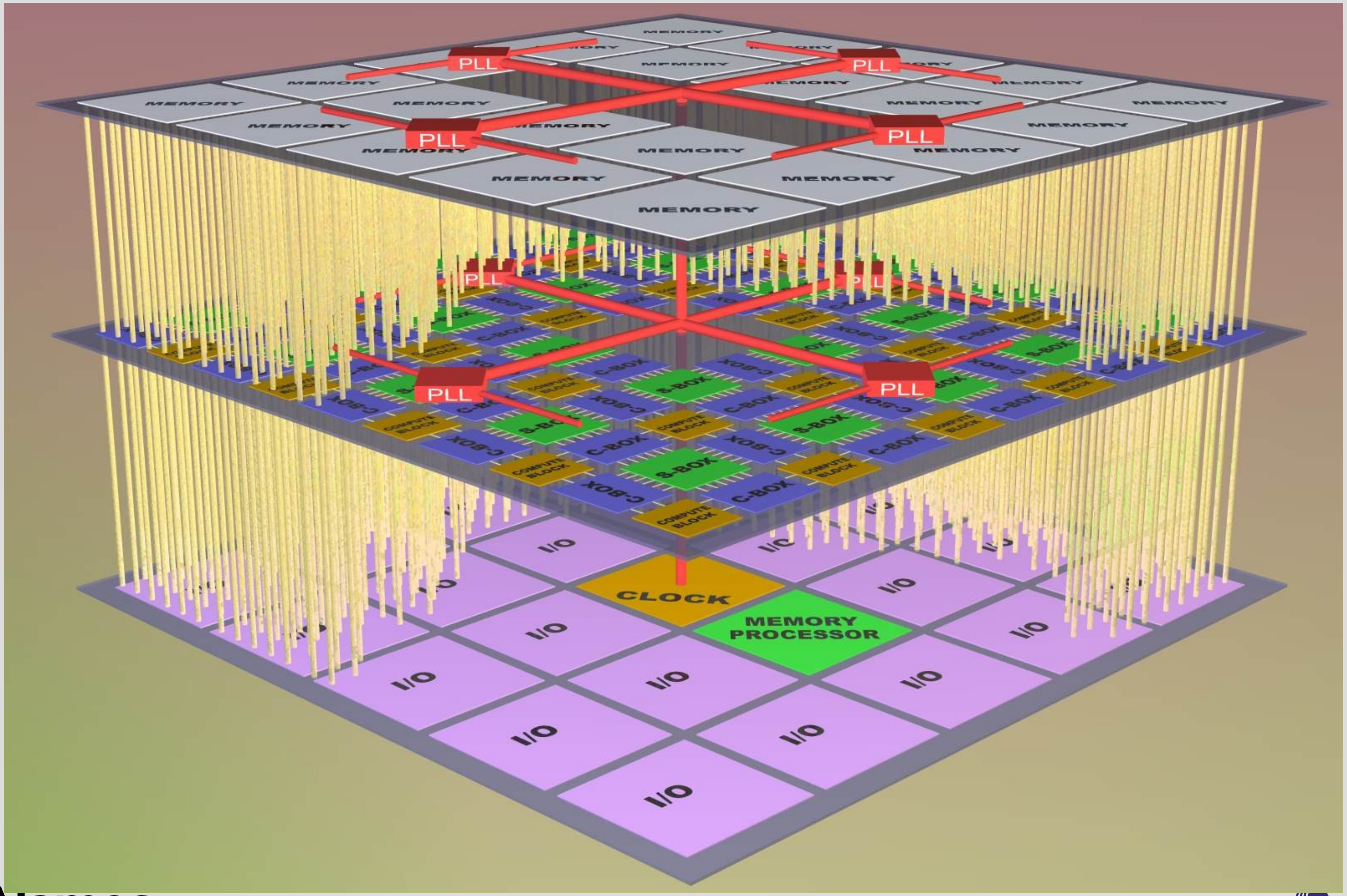
Technology Elements

- 3D wafer-scale bonding technology is perfected and DEMONSTRATED
- FPGA design methodology is well known, readily adaptable to 3D
- FPGA software is available
  - This the most difficult problem

Interlayer Interconnnect



Conceptual Device Structure



Interconnect Limitations

- The interconnect is Achilles' heel of present FPGA's
- Major source of performance restriction, both on-chip and off-chip
- Resource limited, delay inducing
  - Wire length
  - Capacitance
  - Limited I/O resources

Interconnect Strategy

- Examine Supercomputer Interconnect Architectures
- Review academic approaches to FPGA interconnects
- Examine new IC interconnect concepts
- Evaluate interconnect performance in FPGA use
- Adopt elements that work, as demonstrated in simulation

Project Structure

- Phase I: Architecture/Proof of Principle
  - Simulated interconnect strategy
  - Test cells for each layer
- Phase II: Small Working Prototype
  - Constructed at final line width
  - Assembled with S-O-A chip stacking
- Phase III: Scaled-up Limited Volume Production
  - Assembled with full wafer bonding

Project Objectives: I

- USABLE 1.25-2.0 GHz FPGA
- Targeted to processor architectures
- Provision for future SRAM/DRAM layers
- Minimum 4X I/O Resources of conventional parts

Project Objectives: II

- Produce a USABLE prototype 3D FPGA
- Result in 24-36 months
- Phased with defined milestones
- Fully US project; no foreign design participation

Conclusions

- DarkHorse pushed many design envelopes
  - It is the I/O, NOT FLOPS (ITIOS)
  - 3D Memories
    - Self Healing
    - 4-8TB/S Memory BW
  - 3D Stacking (S/MOC)
  - 3D FPGA/CAM/OC-768 Device Designs
  - Optical Interconnects
    - Networking (Optical & Copper)
    - Total Optics off-chip
    - Optics on board (Chip to Chip)
    - OSMOSIS (see refs)

Conclusions (cont)

- Interconnects
  - 12X-QDR Infiniband
  - 32X-ODR Infiniband (Future)
- 3D Memories will improve Power/Performance
  - Non-DRAM (Hybrid)
  - Some new memory technologies (around the corner)
  - S/G
- Currently modeling codes against DH design
  - Some new algorithms (Sparse)
  - Libraries
  - Potentially new language approaches
    - PGAS/DGAS
  - Future HW/SW designs

Conclusions (cont)

- The design is feasible
  - Most of the sub-components exist or have been proven
  - Cooling technologies exist for >1KW socket
    - Liquid Metal
    - Microchannel Cooling
    - Liquid Immersion
- Work has started the design of the follow-on
  - Pegasus (~800PF - 1EF)
- Design expandable
  - ~10TF/Socket
  - ~10TF/4sq in.
- DOE/ASC does not like to fund disruptive technologies