DarkHorse a Proposed Peta(FL)OPS Architecture

Steve Poole
Los Alamos National Laboratory

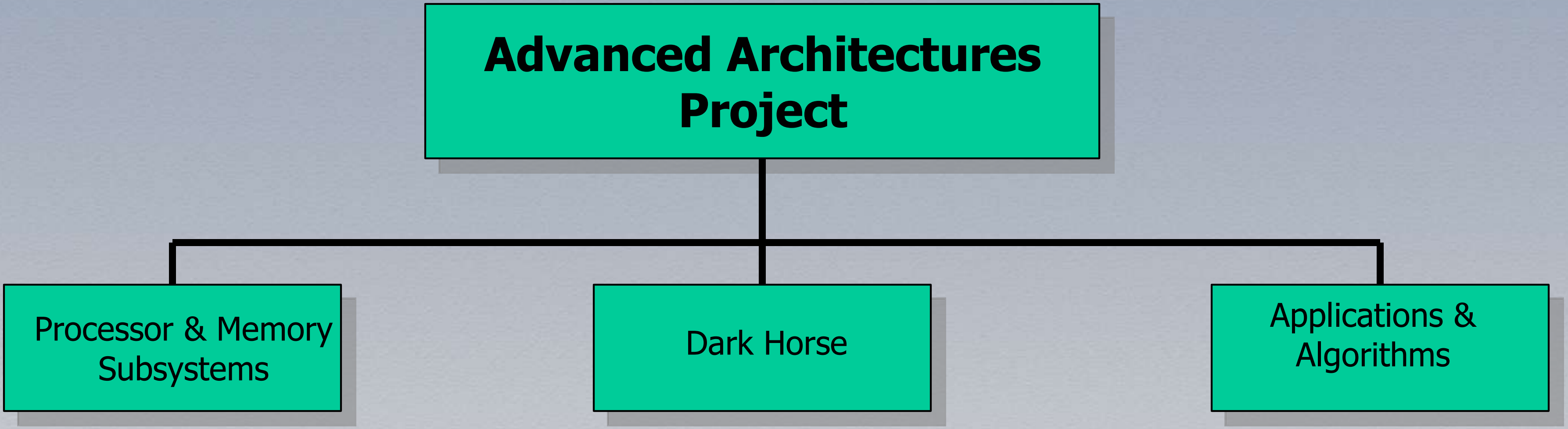
Salishan Conference on High Speed Computing
April 18-22, 2005
LA-UR-05-2745

Advanced Architecture Team

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

Some History

- 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

Processor & Memory Subsystems

Computer industry collaborations

- Understand and influence product roadmaps

Semiconductor industry collaborations

- 3D semiconductor stacking

Co-processor technologies

- FPGA accelerators
- Graphics/Network processor accelerators

Dark Horse

Determine the feasibility of developing a PF system in the ~FY08 time frame that is:

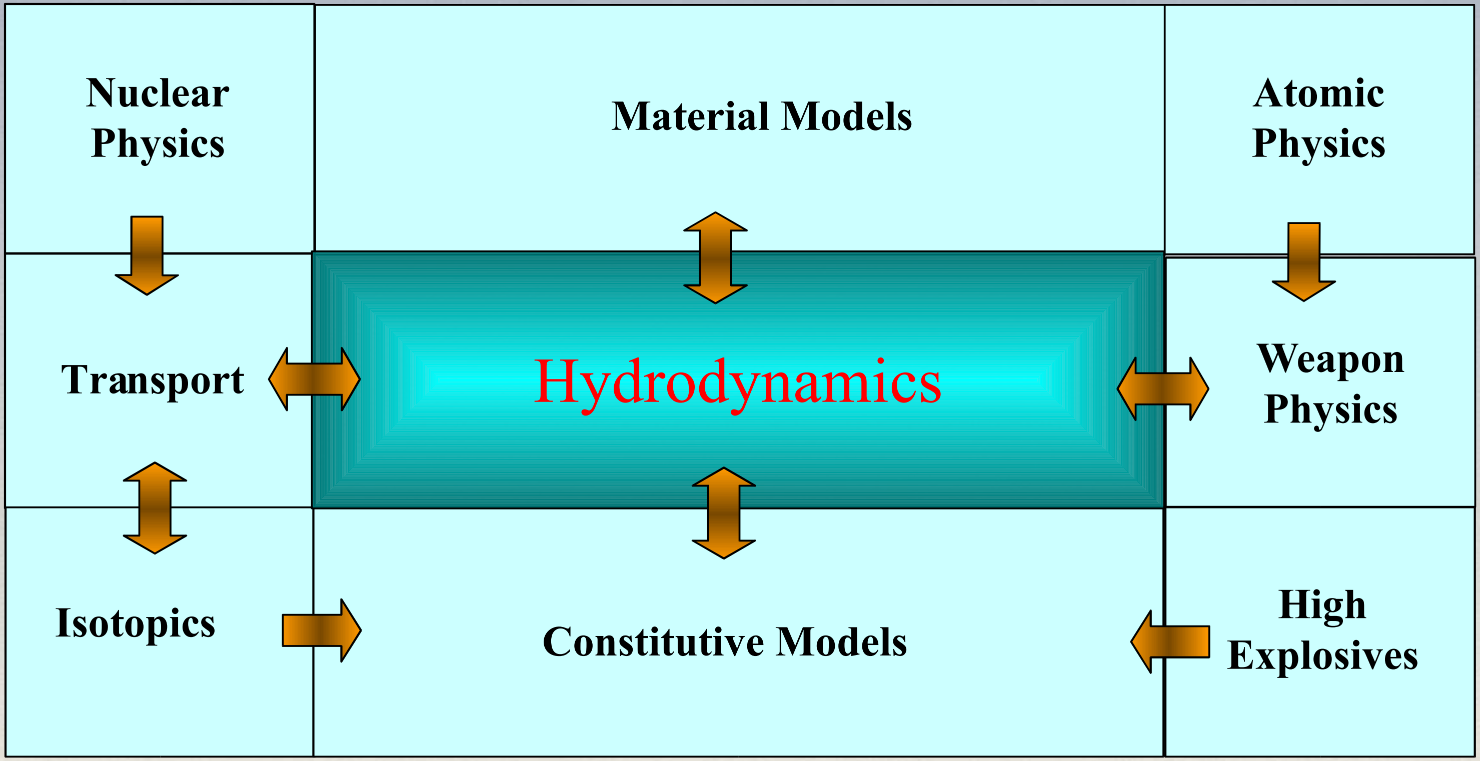
- based potentially on a variety of microprocessors,
- computationally efficient for LANL algorithms, and
- straightforward to program.
- Balanced
- First Principle

Applications & Algorithms

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)

Elements of an ASC Simulation Code



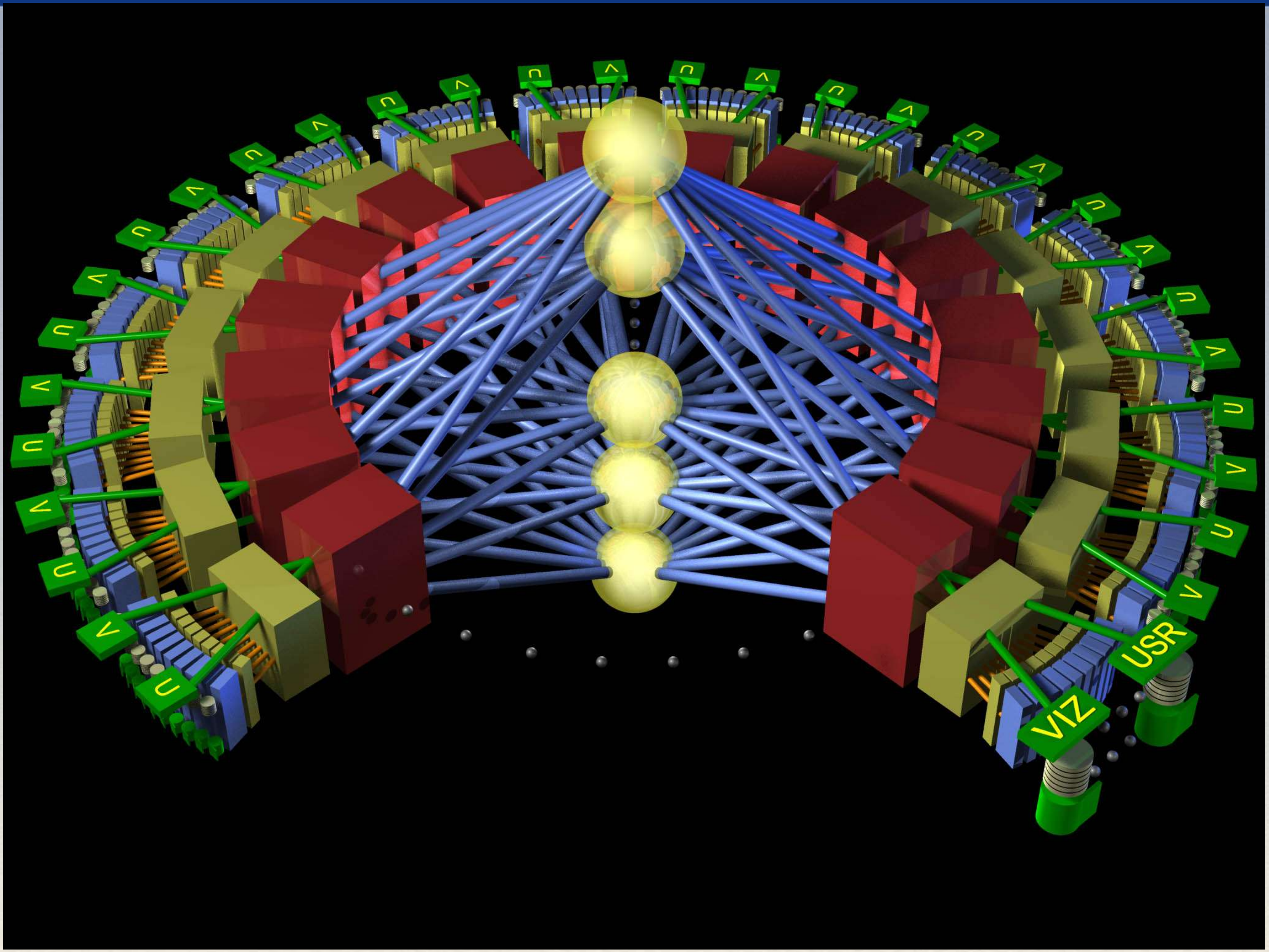
Time evolving coupled multi-physics simulations.

Some Unclassified Testbed Codes

Code	Association/Support
(S/R)AGE	Crestone Project
MCNP	Eolus Project
PARTISN	Sn Transport
SWEEP3D	Sn sweep strategy
TRUCHAS & TELLURIDE	Telluride Project

They are:

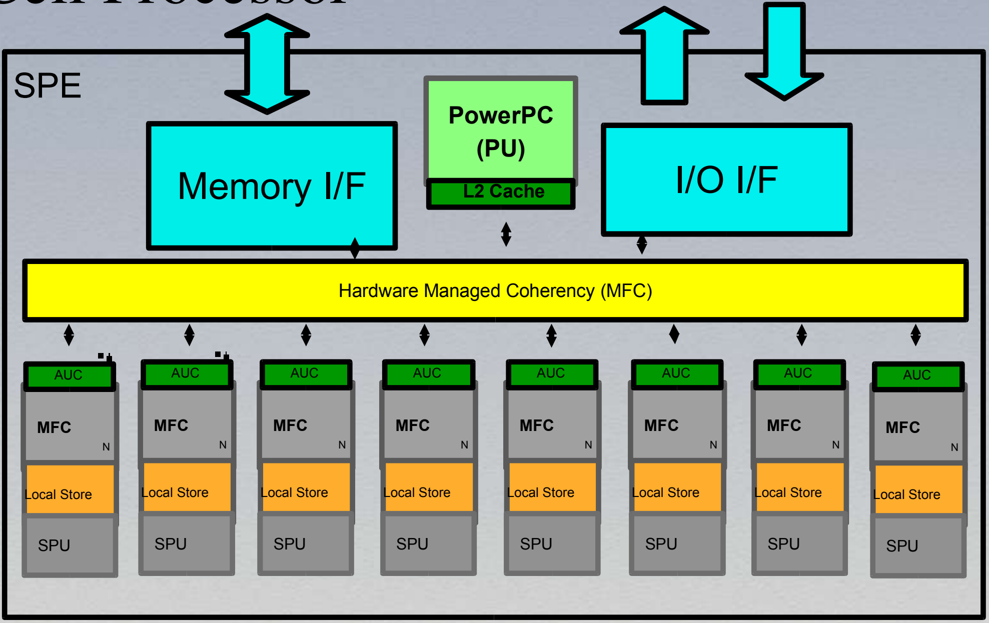
- **Representative** of computer science issues
- **NOT** parts of our classified codes
- used for unclassified applications
- used for methods & CS testing and R&D
- often export controlled


U
V
USR
VIZ

One Processor's attributes

IBM Cell Processor

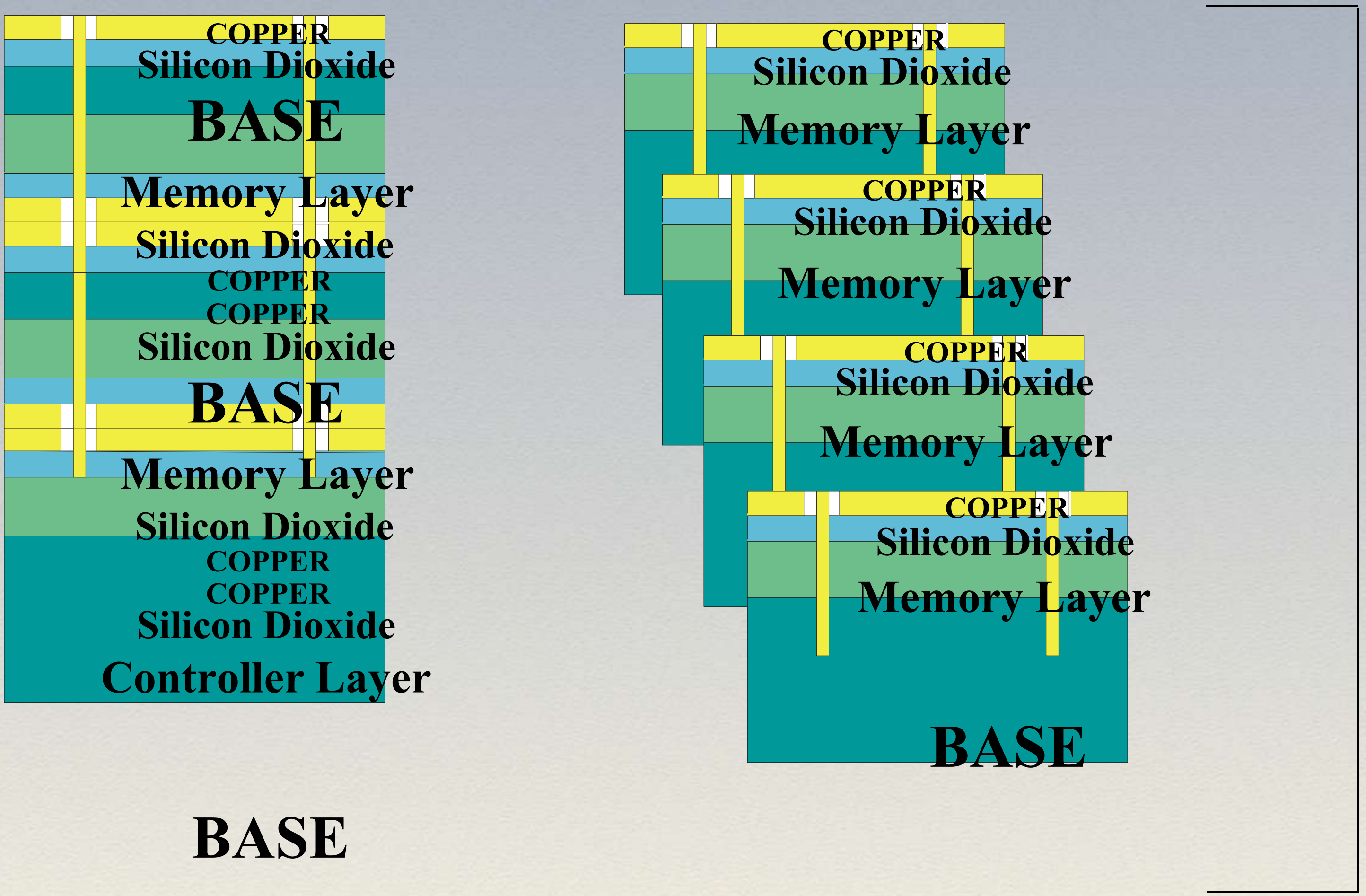
- SMP on a chip
  - 1 PowerPC
    - L1 (32kB+32kB)
    - L2 (512kB)
    - Coherent load/store
  - 8 SPUs
    - 256 GFlops (SPtotal)
    - 256 kB Local Store
    - Coherent DMA
- High-B/W Memory
  - 25.6 GB/s (data)



- Configurable I/O interface
  - Up to 35GB/s out
  - Up to 25GB/s in
  - Coherent interface or I/O

Stacking Multiple Thin Layers

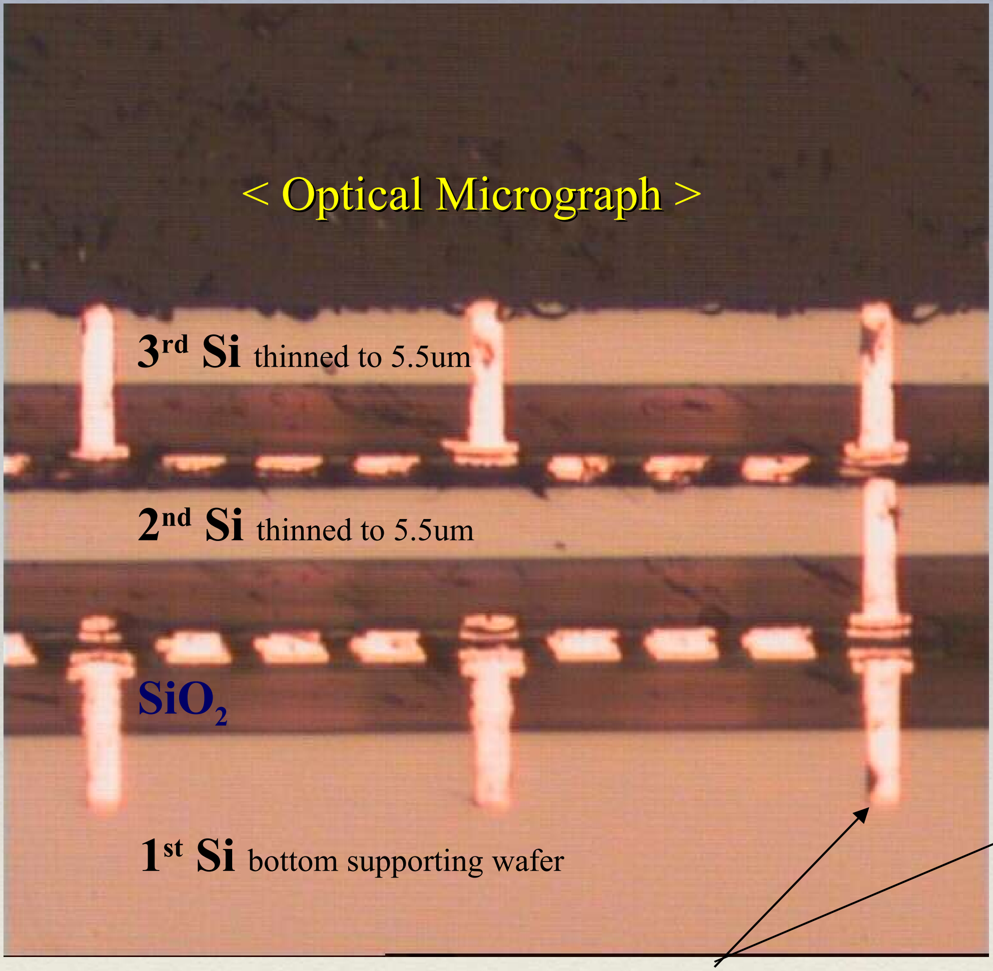
Repeat - One Wafer at a Time



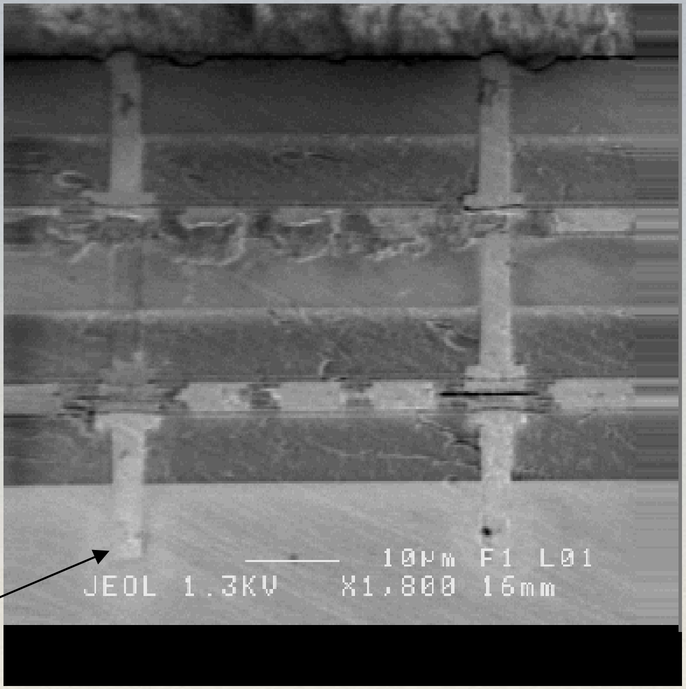
Additional Memory Layers to be stacked

Stacking Process

Three wafers successfully aligned and stacked

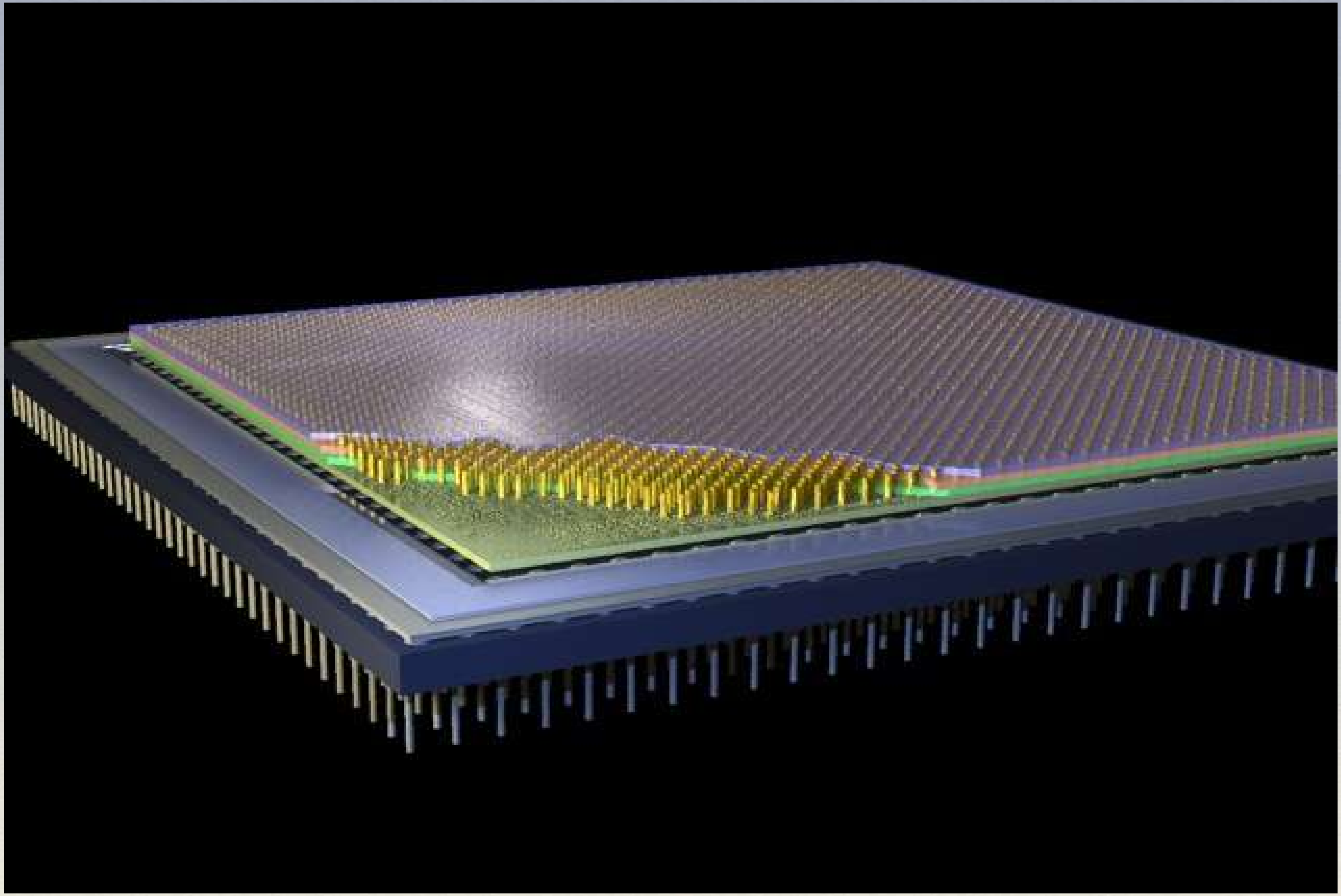


< Scanning Electron Micrograph >



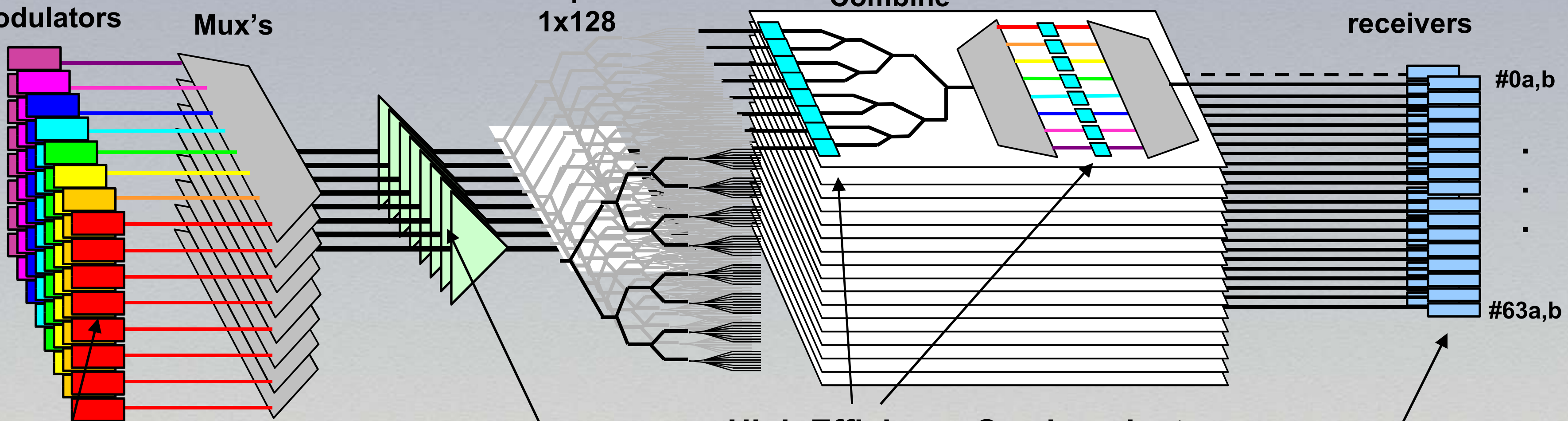
"Super Via" 4um in diameter and 12um in height

Chip Stacking



Bufferless Crossbar Design:

Implemented via Broadcast and Select Architecture


Fast SOA Fiber-Selector Gates
Fast SOA Color-Selector Gates
Laser-Integrated Modulators
Mux's
EDFAs
Star Couplers 1x128
8x1 Combine
40 Gb/s Packet receivers
#0a,b
#63a,b

High Efficiency Semiconductor Optical Amplifier Gates

High Capacity TX:

Low Latency High Power Optical Gain

High Sensitivity Cyclic RX

S: Multiple fibers (8 scaling to 40+)

: Multiple colors per fiber (8 scaling to 100+)

T: Switching time (~2 ns scaling to <0.1ns)

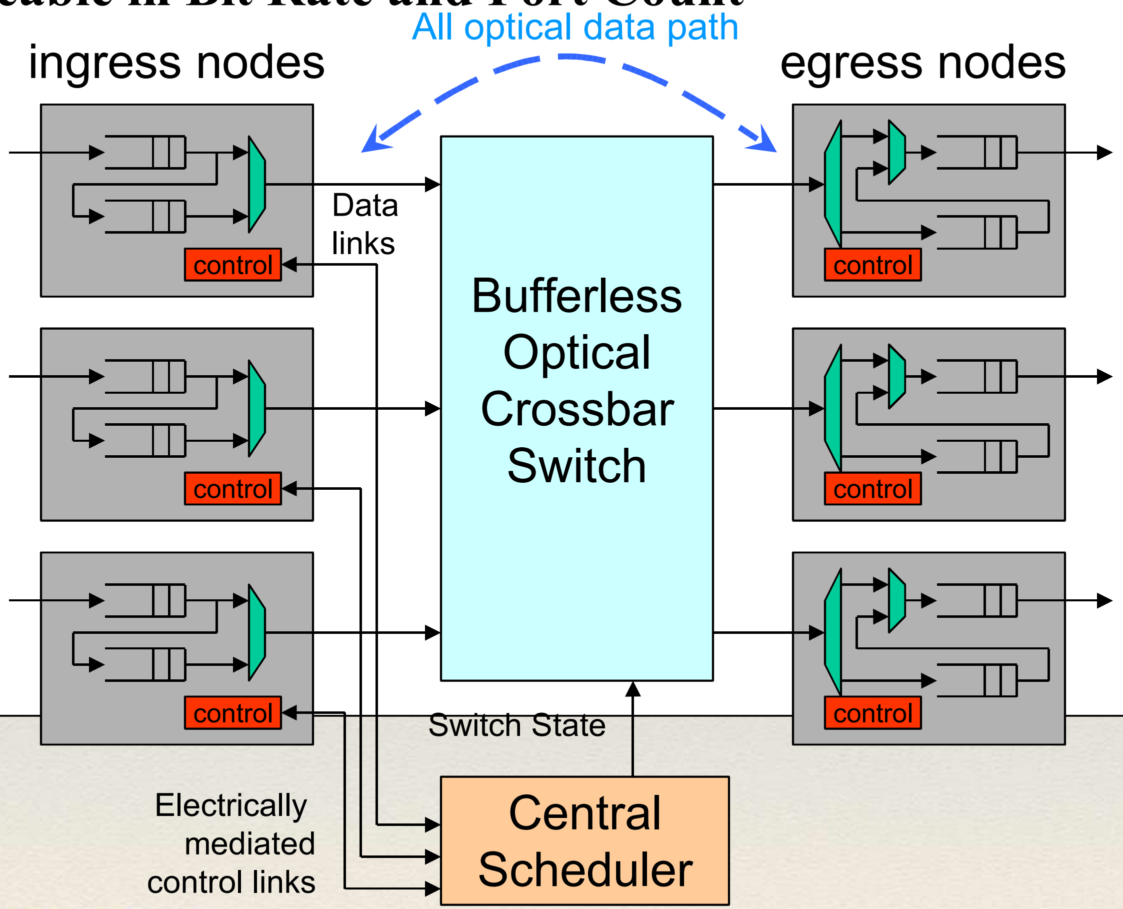
High bit rates (40G scaling to 100G+)

Bufferless Optical Crossbar

Compact, Integrated

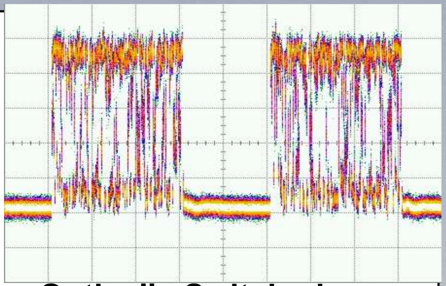
Low Optical Impairment

Scaleable in Bit Rate and Port Count

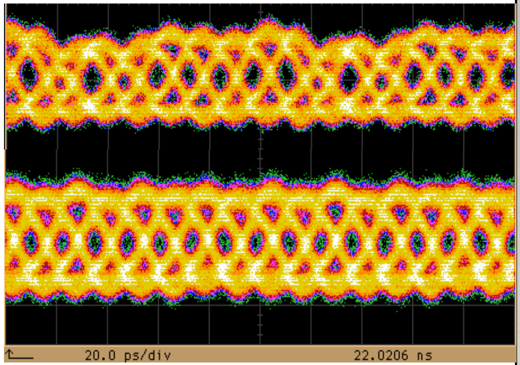


Switched 40G
Optical Packets

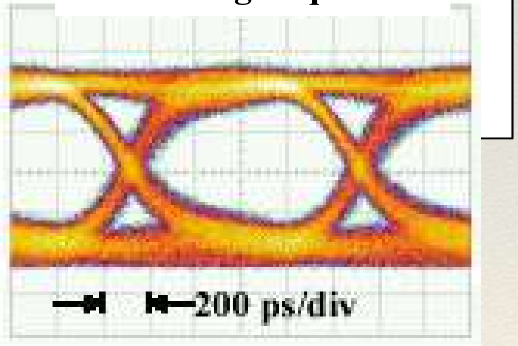
5 ns/div



Optically Switched
SOA @ 80GHz

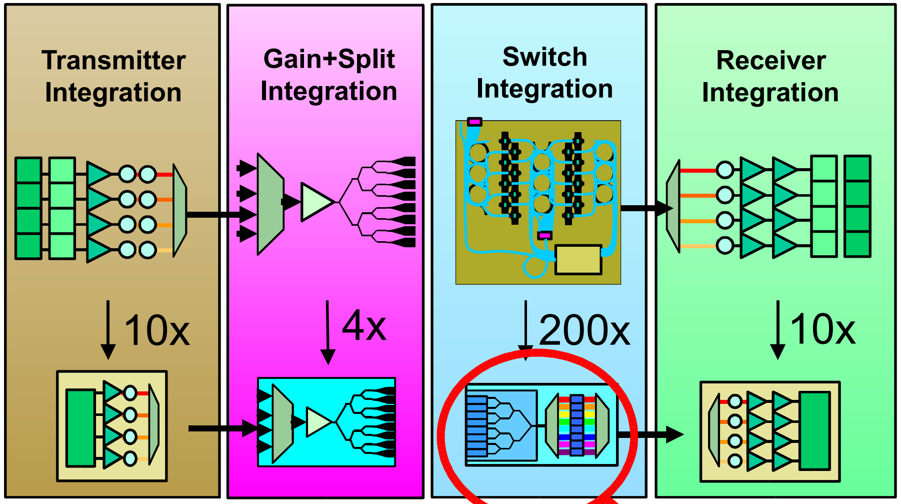


Large Signal 2^{31}-1 PR
Switching Sequence



OSMOSIS will integrate to be cost competitive with conventional OEO

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



Modular chassis

Octal switch port blade

Integration achieves

200:1 Parts count reduction

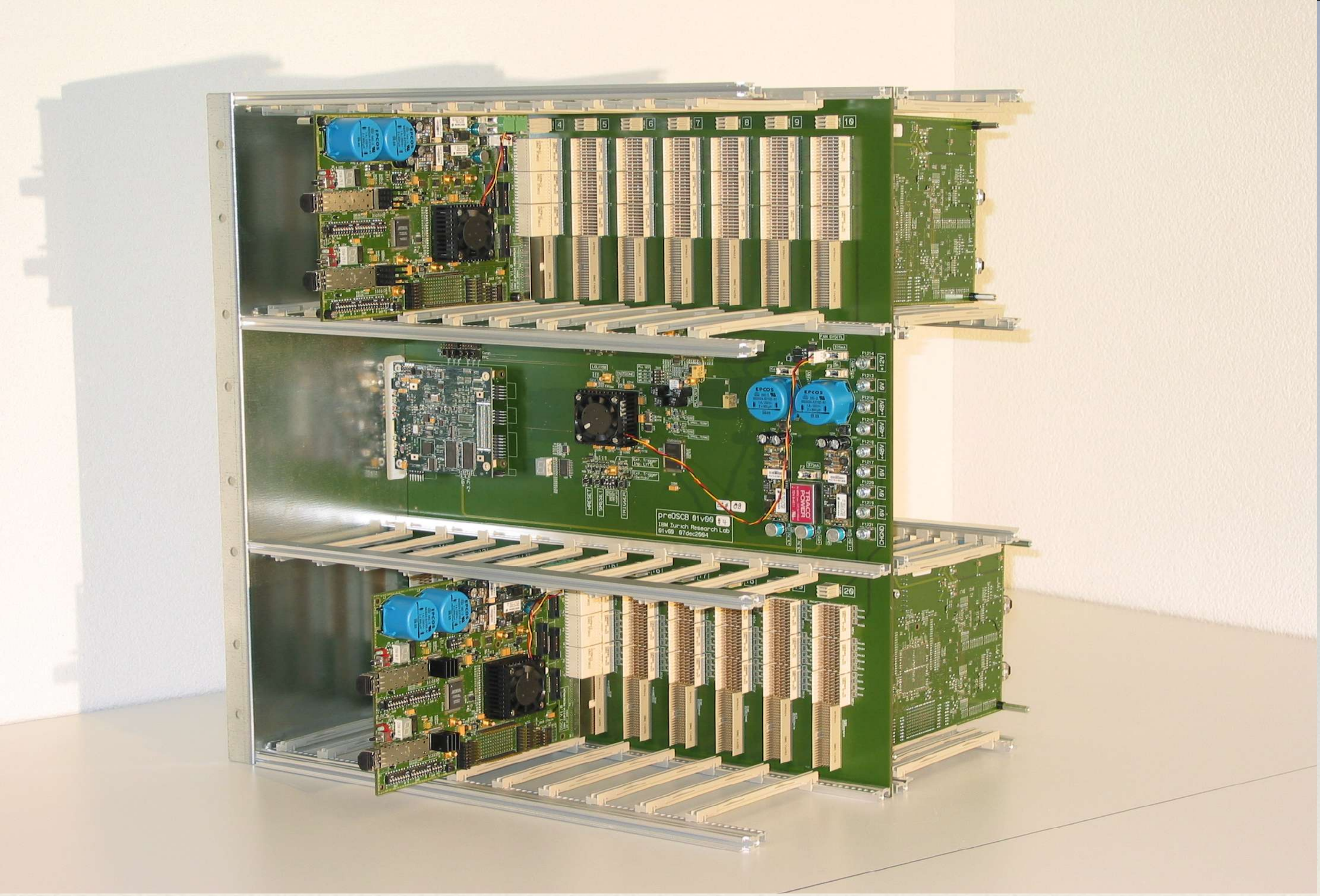
10X Power reduction

Provisioning in 16 port increments

10 Terabit/sec form factor

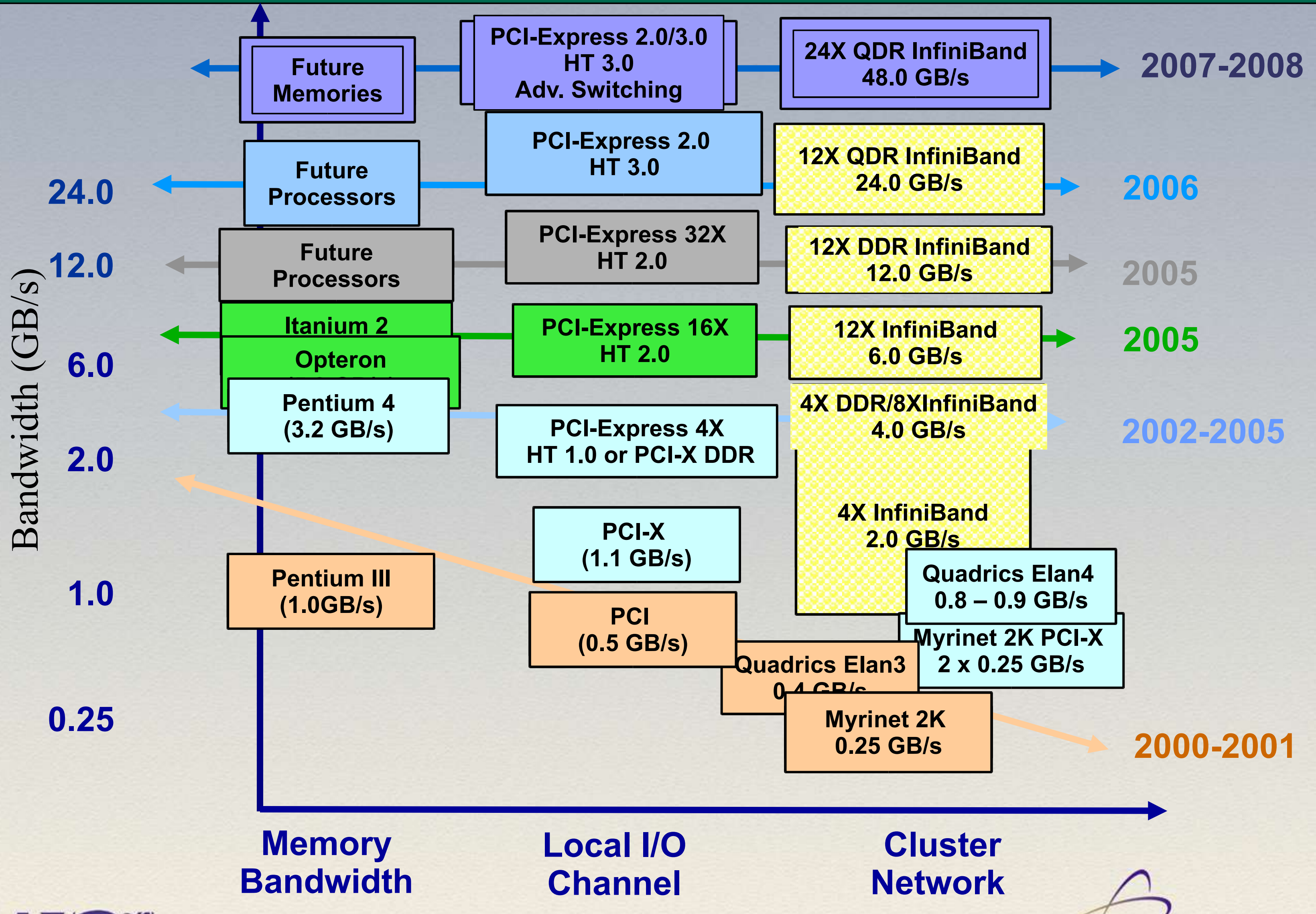
- 1.28 Gigapackets/sec in 64 port switch module
- Cell-oriented error correction supports 10^{-21} BER
- Goal: 10 Tbit/sec in a single stage module @ first commercial release

Development Switch



As of this month (Test Vehicle/Prototype)

InfiniBand Roadmap



Future Memories

PCI-Express 2.0/3.0
HT 3.0
Adv. Switching

24X QDR InfiniBand
48.0 GB/s

2007-2008

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

2005

Pentium 4
(3.2 GB/s)

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

4X DDR/8XInfiniBand
4.0 GB/s

2002-2005

PCI-X
(1.1 GB/s)

4X InfiniBand
2.0 GB/s

Quadrics Elan4
0.8 – 0.9 GB/s

Pentium III
(1.0GB/s)

PCI
(0.5 GB/s)

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

Quadrics Elan3
0.4 GB/s

Myrinet 2K
0.25 GB/s

2000-2001

Distance from CPU

Conclusions

- DarkHorse pushed many design envelopes
  - It is the I/O, NOT FLOPS
  - 3D Memories
    - Self Healing
  - 3D Stacking (S/MOC)
  - 3D FPGA/CAM Designs
  - Optical Interconnects
    - Networking (OSMOSIS)
    - Chip-to-Chip (ZRL)
  - Interconnects
    - 12X-QDR Infiniband
    - 32X-ODR Infiniband (Future)

Conclusions (cont)

- 3D Memories will improve Power/Performance
  - Non-DRAM
- Currently modeling codes against DH design
  - Some new algorithms (Sparse)
  - Potentially new language approaches
  - Future HW/SW designs
- The design is feasible
  - Most of the sub-components exist
- We have started the design of the follow-on

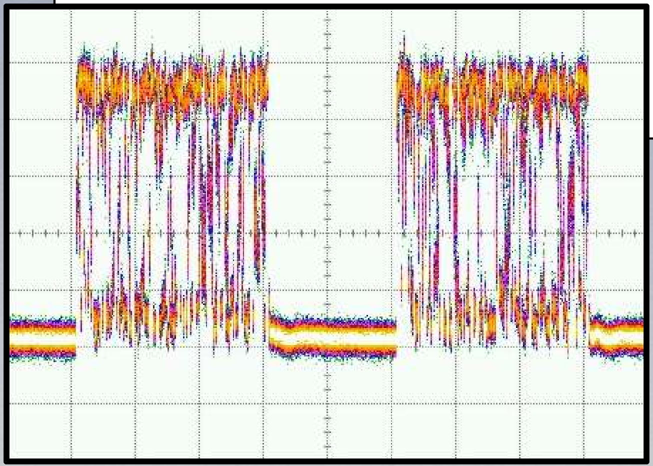
Special Thanks (DH)

- LANL
  - Gary Grider
  - Karl-Heinz Winkler
  - John Morrison
  - James Peery
  - Ken Koch
  - Rich Graham
  - Mike Boorman

- SNL
  - Bill Camp
  - Jim Tompkins
  - Matt Leininger
- Mellanox
- IBM
  - (ZRL,POK,TJW,ARL,STIDC)
- Corning
- Many others…

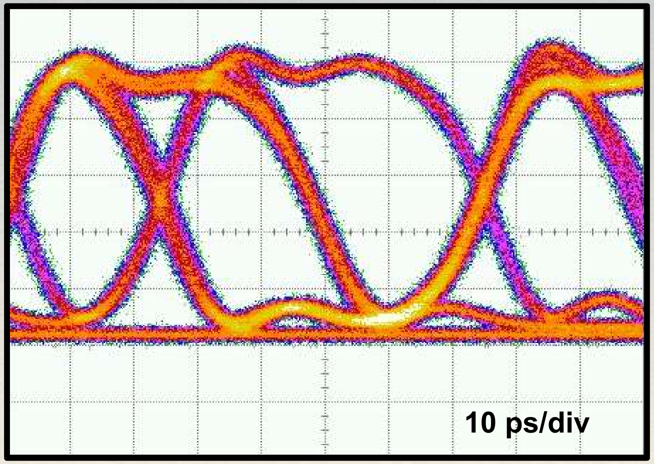
Backup (Movie)

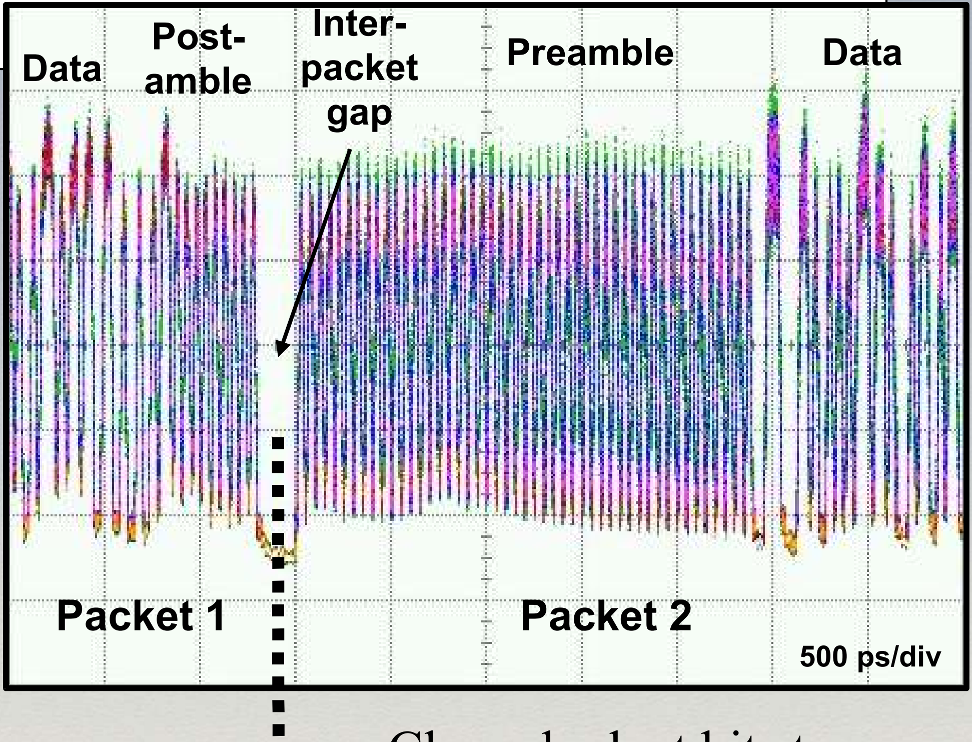
Data packet bit streams/eye diagrams



5 ns/div

40 Gb/s data packet



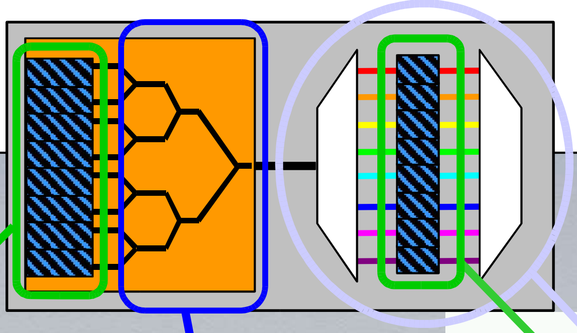


Closer look at bit stream

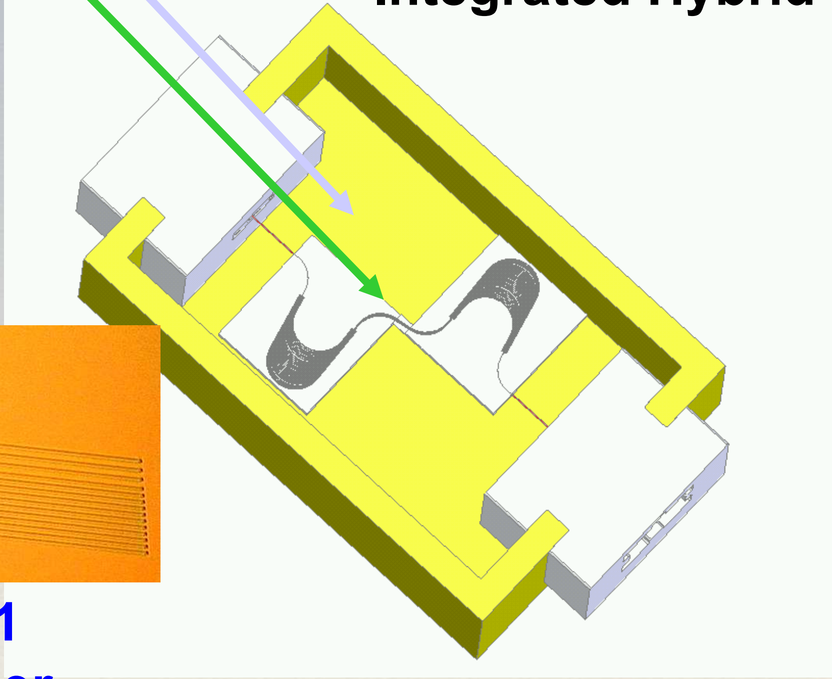
Fast integrated optical fiber and color selector

Cost reduction:

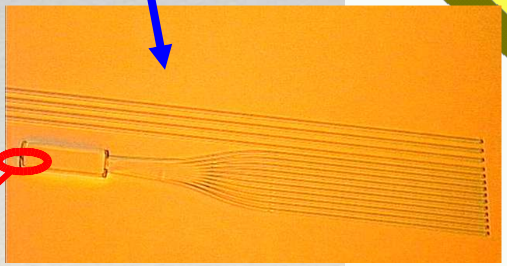
- planar integration
- parts reduction
- wide fabrication tolerances → high optical margins



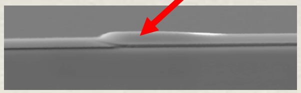
Fast SOA Color Select Integrated Hybrid



Monolithic SOA Array



InP 8x1 Combiner



Monolithic Optical Interface