

# **Interconnect Science & Art**

***Greg Chesson***

**Chief Scientist, R&D Division**

**Silicon Graphics, Inc.**

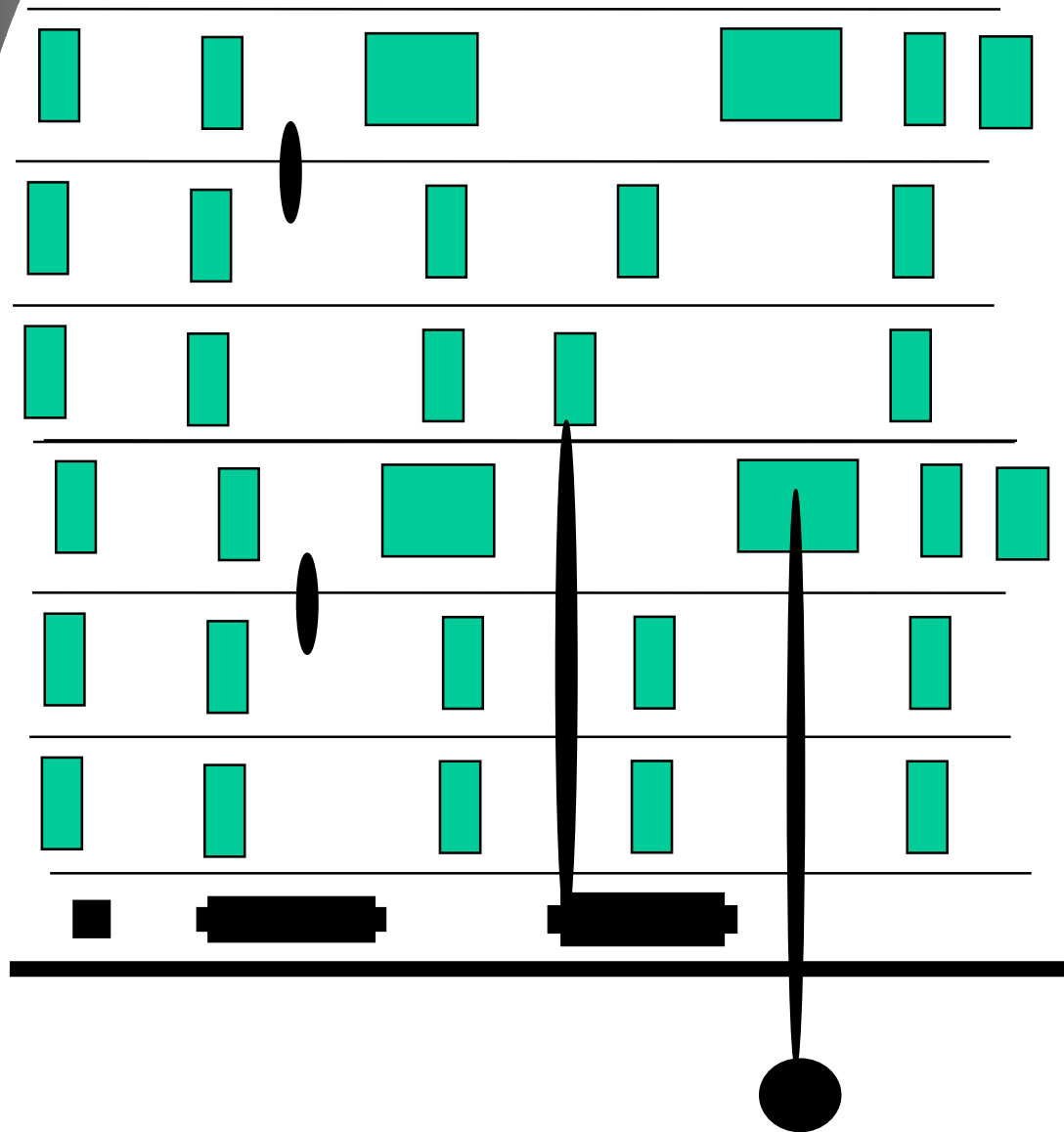
*[www.sgi.com](http://www.sgi.com)*

# Scaling Effects

## *Physics vs Chemistry*

- **Chemistry**
  - The “recipe” for dinner
- **Physics**
  - Dinner for 2
  - Dinner for 2000
  - Dinner in 20 minutes
  - Dinner in 2 seconds
  - Dinner for 2000 in 2 seconds

# VLSI density tracks metal layers



6-layer metal process

Wires vary

Vias, Bonds  
fight for area

substrate

# Internal Architecture

***“micro” interconnects problem mimic system interconnect problems***

- Many advances in metal, fab, packages, BUT
- Brick wall at chip boundaries
  - Inductive, capacitive, resistive, coupling effects
  - Pin density limitations
  - Power
- Introduces latency and bw limits
- Observe interconnect issues everywhere

# Extended Hierarchy

*Register - Memory*

*Reg - L1 - Mem*

*Reg - L1 - L2 - Mem*

*Reg - L1,2,3 - local Mem*

*Reg - L1,2,3 - local Mem - Remote Mem*

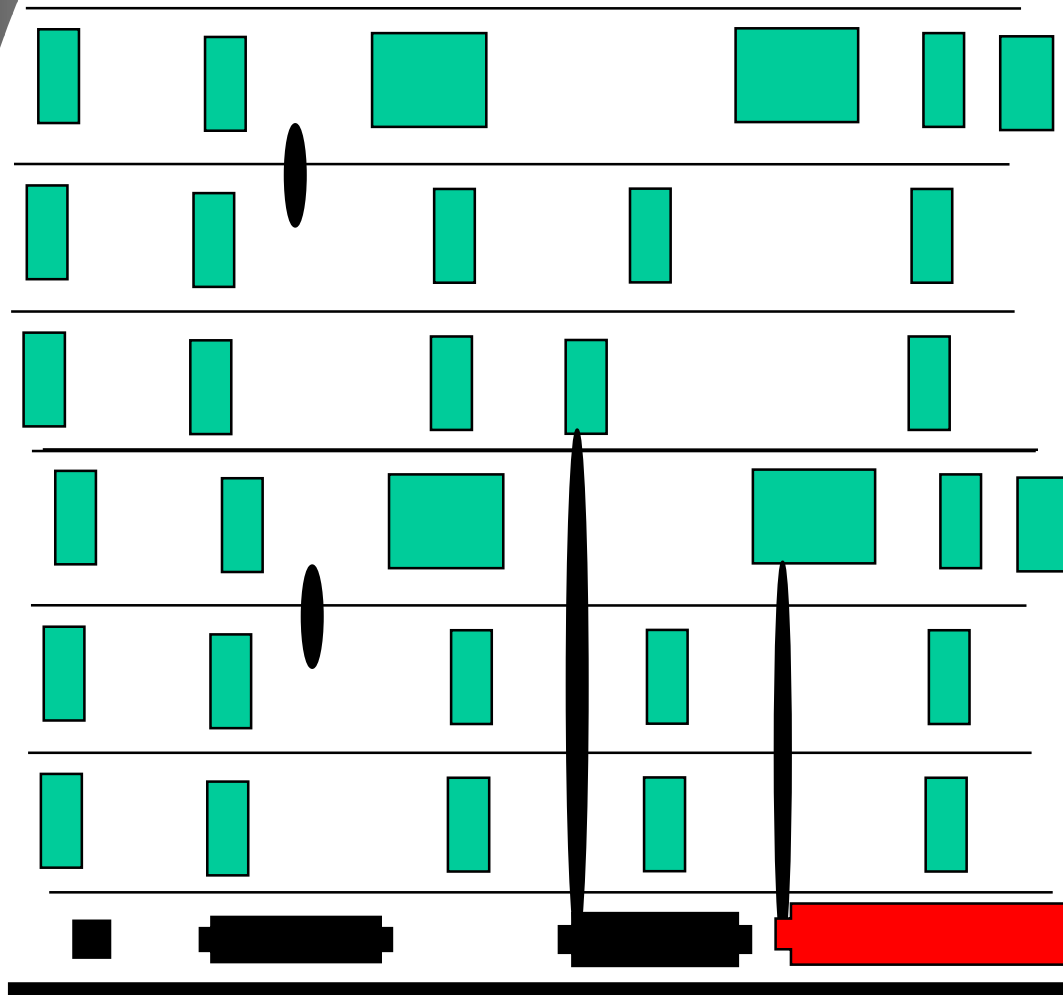
*Reg - cache - local - N1 remote - N2 remote*

- Opportunity for software exploitation
- Opportunity for interconnect-based memory
  - A “dual” to processor-in-memory
- Latency reduction vital for scaling

# Interconnect Scaling

- Scaling extends memory hierarchy ... and latency
- Direct result of bandwidth/latency “distance”
- Protocol/interconnect design issues at each level
- **Optical switching offers latency shortcut**
  - Increased jitter compromises clock rate
  - EO conversion is serious problem area
- **Electronic switching will evolve until ...**

... until



***FUND THIS***



**Optical channel**

**On sys/cpu chips**

# Petaflop Assumptions

- Multicomputer-style system @ some node size
- Only 1000 Teraflops, or 1000 1 TF nodes
- 1000-4000 port X2 interconnect, some X1 interconnect
- ~100 GB/s/port, or 10 planes at 10 GB/s/port
- Interconnect will provide: **IO, storage, messaging, memory, synchronization, fault-tolerance, admin**
- **Fusion** of internal and external interconnect technologies

$$\begin{array}{l} 10 \text{ GF} < \text{cpu} < 100 \text{ GF} \\ 10^5 > N_{\text{cpu}} > 10^4 \\ 2^{17} & \qquad \qquad \qquad 2^{14} \text{ (approx)} \\ 2^{10} \times 2^7 & \qquad \qquad \qquad 2^{8+6} \end{array}$$

assume 10 GF/cpu to begin

# Moore's Law

- Predicts commodity cpu/vlsi evolution
- Doesn't predict
  - Non-commodity
  - Non-cpu or optical technology
- Network/interconnect dependencies
  - Packaging (pinout) density, latency
  - CPU/logic trends protocol, function
  - Memory bandwidth/density buffers
  - Signaling distance, bw
  - Launch rate (!) cpu/ix arch (CPL?)
- All have different evolution rates

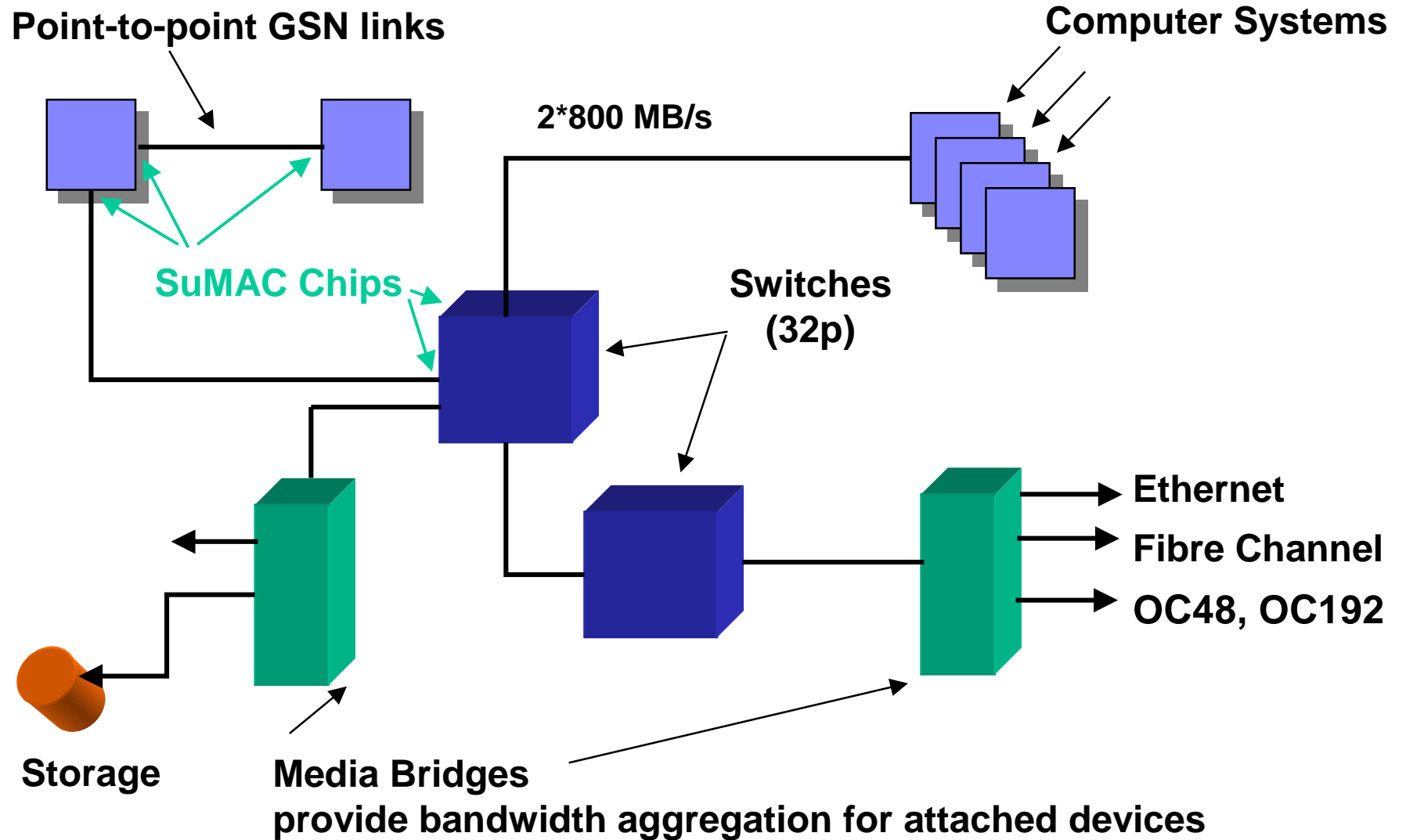
# Baseline Technologies

- **Packaging**      **1500 ->      3000-5000**
  - Limited by die area and yield
  - Allow 16-port (or more) routers
  - Latency of  $\log_{16}(N)$  hops
- **Logic**            **1            ->      4-10 Ghz**
  - Inject clock all over die
- **Memory**         **1-4        ->      10-80 GB/s on chip**
- **Signalling**      **1-2            4+ Ghz/channel**
- **Launch rate**    **1.5 M->    10 M/sec**

# Contemporary IX Design

- GSN (HIPPI-6400) Interconnect:
  - OS Bypass messaging
  - Storage
  - Media conversion, bandwidth distribution
  - ASCI Blue cluster interconnect
  - ASCI PathForward adaptor develop for DEC, IBM, pci
  - ANSI standard ST protocol carries OS Bypass and storage plus non-coherent NUMA access (MPI-2)
- Next Generation
  - Extended switches
  - Extended adaptors
  - Optical links

# GSN Interconnect

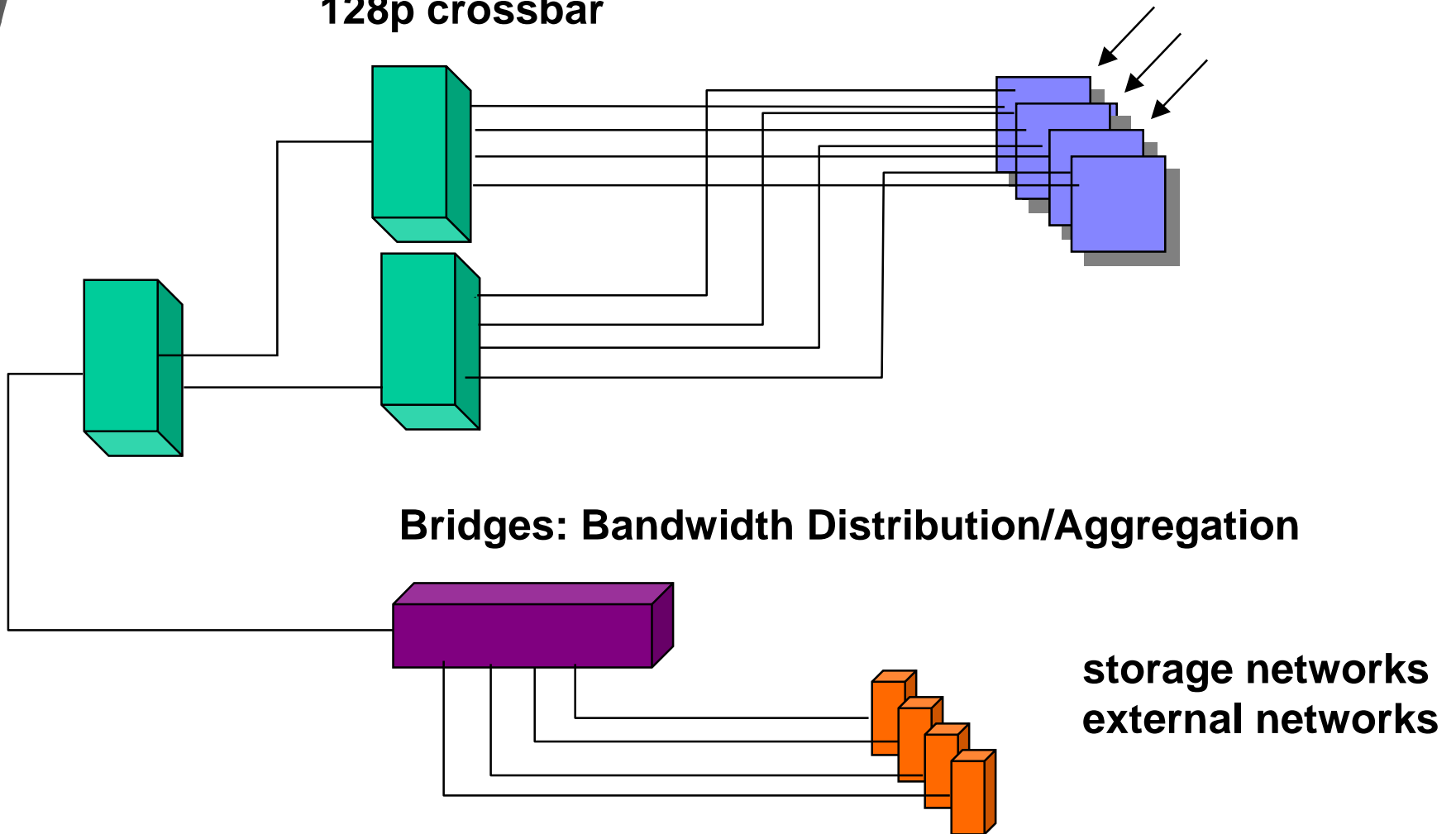


# Peta Sys

(optical signaling, elect. or opt. switch)

**Switches**  
32p-1024p  
128p crossbar

**Computer Systems**  
100 nodes @ 10 TF



**Bridges: Bandwidth Distribution/Aggregation**

**storage networks**  
**external networks**

# Scaling up Interconnect/OS

- **Depend on Meta-OS overlay**
  - membership, fault-tolerance, services, scheduling
- **Depend on base OS**
  - cluster FS/IO, scheduling, group process control
- **Depend on target protocol/API**
  - hw half-life is brief
  - protocol/API serves as target “assembly language”
- **Depend on vm, tlb, proc, queue management**

Software must evolve with processors, interconnect  
Software nearly always follows hardware  
Software key to efficiency levels

# Scaling Summary

- **Peta interconnect configurations possible within normal interconnect evolution**
- **1-10 GB/s single-stage switched interconnects**
- **Use optics if/when ready, need hybrid vlsi**
- **Interconnect protocol provides**
  - memory
  - storage
  - messages
- **Need new generation of OS, metaware, application software**