



面向HPC的光互连技术

——现状与前景

中科院计算所•智能中心

陈明宇

2003.10.14



内容

- 电信号的局限性
- 光互联技术
- 光交换技术
- 小雨点实验平台



Ghz时代的CPU

- Intel Pentium 4: 3.2G
- Intel Itanium: 1.5G
- AMD Athlon XP: 2.2G
- AMD Operton: 2.0G
- IBM PowerPC 970: 2.0G
- IBM Power 4+: 1.7G
- Sun UltraSparc III: 1.2G



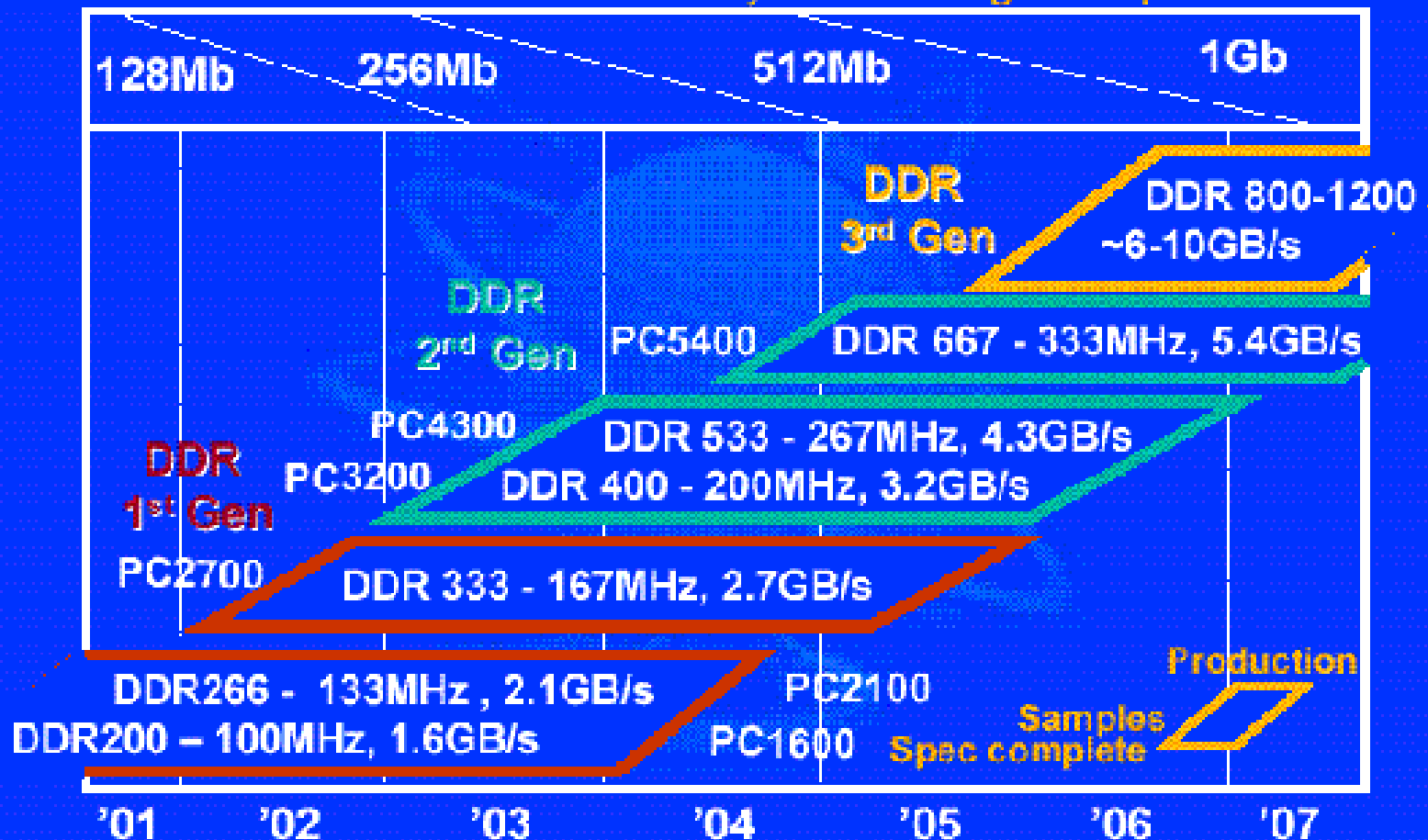
GHz时代的系统总线

- Pentium 4: 800Mhz 64bit 6.4GB
- Itanium2: 400Mhz 128bit 6.4GB
- Athlon XP: 400MHz 64bit 3.2GB
- AMD Operton: 1600MT 16Bitx2 6.4GB
- PowerPC G5: 1Ghz 32x2bit 6.4GB

GHz时代的DRAM

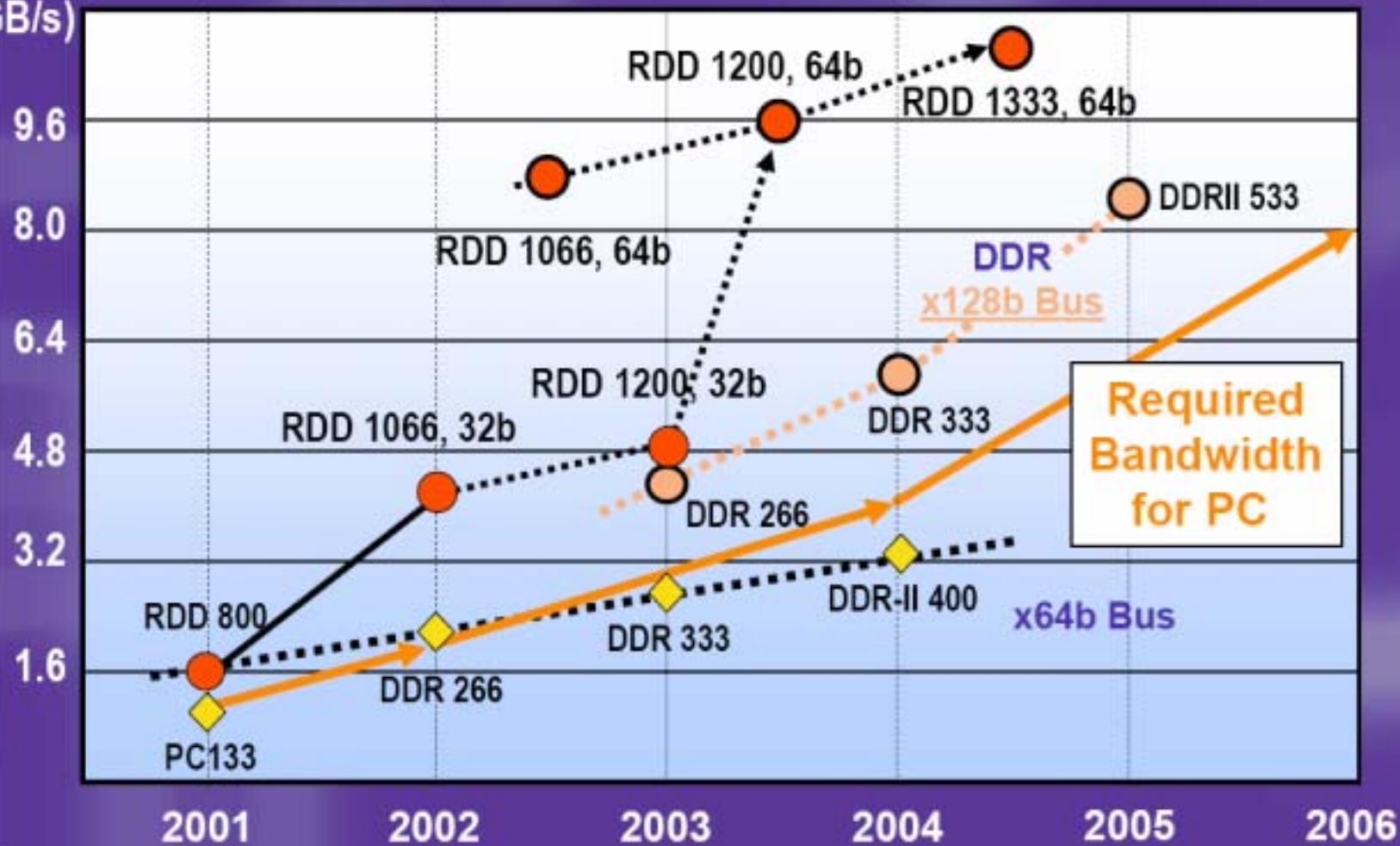
DDR Module Roadmap

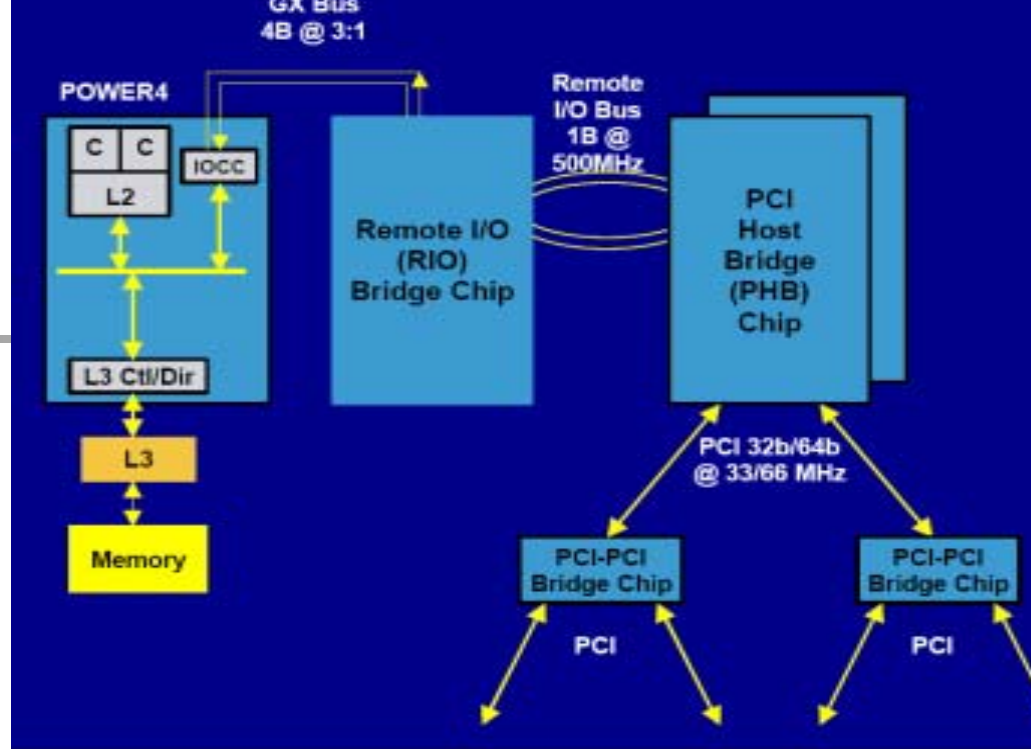
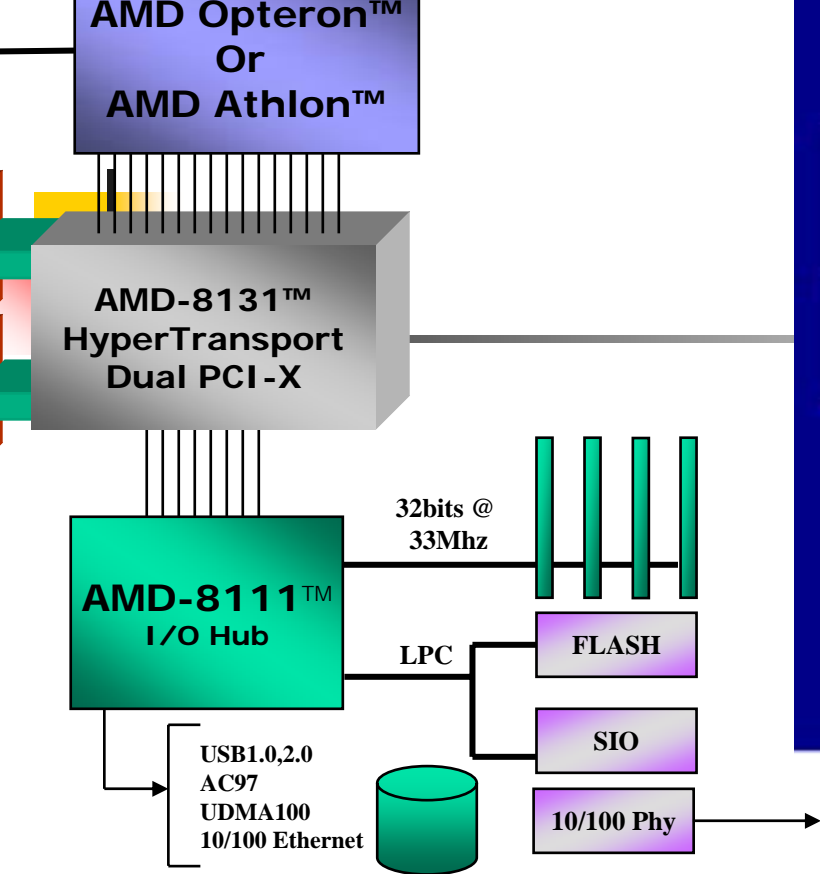
The "Mainstream" memory module migration path



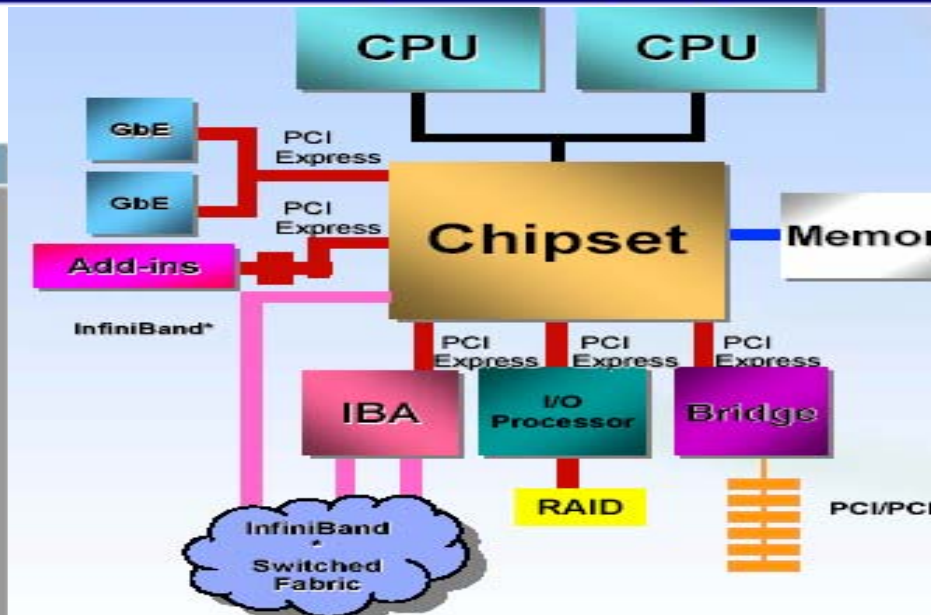
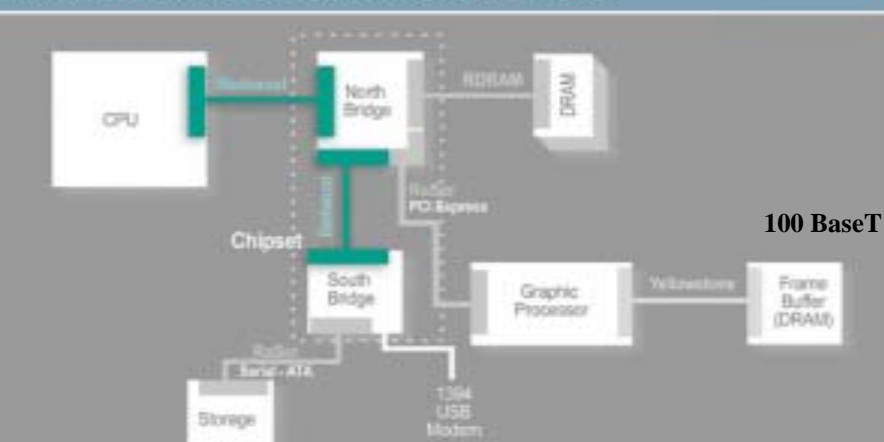
GHz时代的DRAM-续

1.2(GB/s)





POTENTIAL SOLUTIONS FOR COMPUTING APPLICATIONS





系统内部互联的趋势

- 交换式结构代替总线式
 - 高速串行点对点连接代替并行总线
 - 基于包交换的协议代替独立控制信号
 - 异步协议代替同步协议
-
- 传统意义上的互联走向通信模式？

Result: System I/O Shift to a Communications Paradigm



- Point-to-point: *Improves speed*
- Differential signaling: *Improves signal integrity*
- Packet transmission: *Lowers pin count*
- Asynchronous: *Relaxes routing rules*
- Layered protocols: *Allow rapid evolution*

Next generation interconnects have the essential elements for easy transition to a new medium



高速互连标准

- PCI x2.0 : 533Mhz并行
 - Rapid I/O : 1G并行/3.125G串行
 - HyperTransport : 1.6G/pin
 - PCI Express : 2.5G
 - Infiniband : 2.5G
-
- 3G是否是铜线连接的局限？



目前的技术水平(PCB)

- 并行信号200~600MHz
- 串行3G (FR4 > 1米)

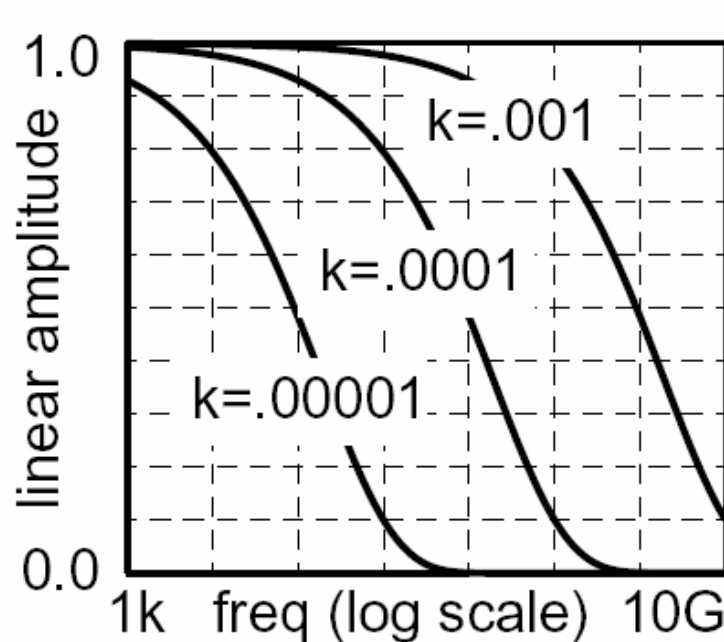
最新进展:

- 10G (FR4 24 Inch) : 预加重 , 后均衡
- 40G SiGe IC

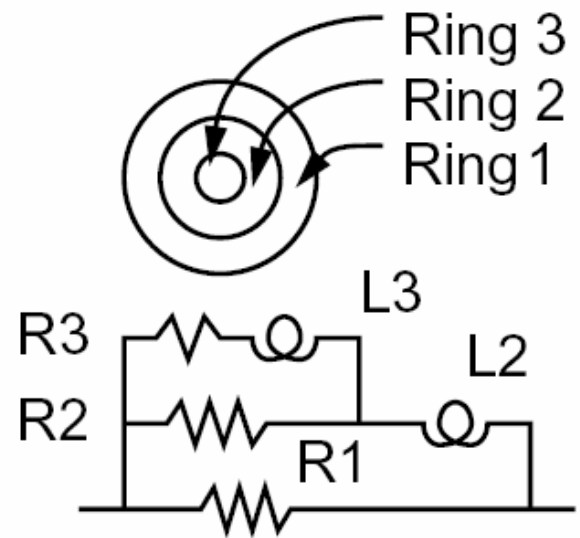
Skin Loss and Dielectric Loss

Nearly all cables are well modeled by a product of Skin Loss

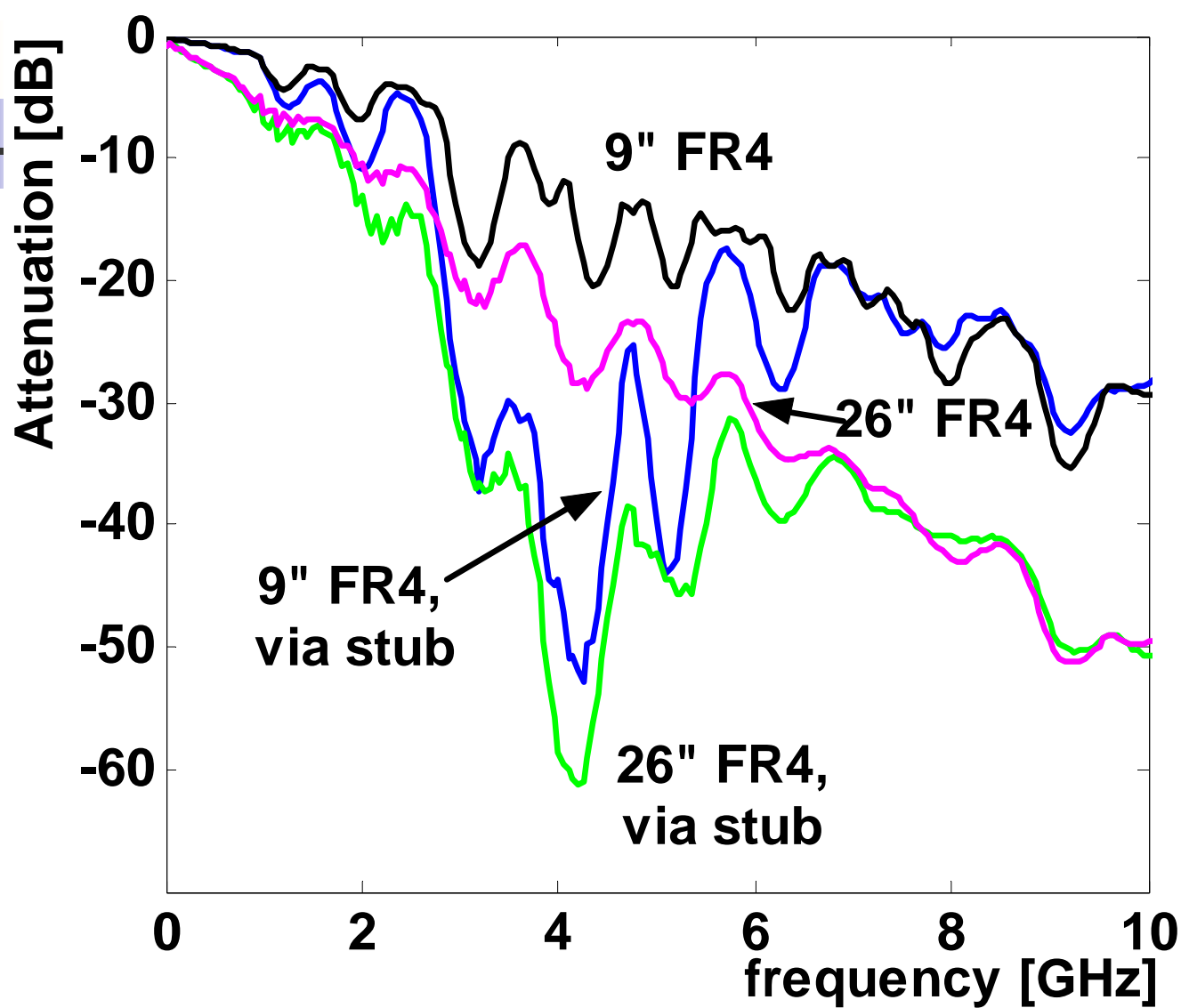
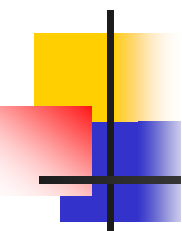
$S(f) = e^{-k_s(1+j)l\sqrt{f}}$, and Dielectric Loss $D(f) = e^{-k_d lf}$ with appropriate k_s, k_d factors. Dielectric Loss dominates in the multi-GHz range. Both plot as straight lines on log(dB) vs log(f) graph.



[YFW82]



Three-element equivalent circuit of a conductor with skin loss



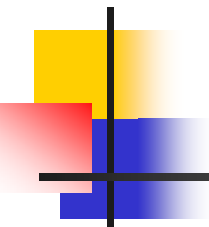
Conclusions

- Still much work to be done, but 1 Tb/s chip I/O seems an attainable target.
- 5Gb/s on 1meter PCB is the fastest that can be feasibly supported for the foreseeable future with *low latency*.
- Fiber seems to be progressing along either a 1-10-100-1000-10,000MHz or a 622-2488-10,000MHz evolutionary path. There may be an economically important need for 5Gb/s links.
- 10 Tb/s chip I/O is probably out of the question for current high-volume technologies (CMOS, FR-4 PCB). Computer designs and programs may have to give up cache coherency, and move towards cooperative computing architectures to break out of this limitation.



目前的技术水平（cable）

- 10GBaseT: 计划在超5类线，100米距离上实现10Gb/S ethernet
- 目前的结果：
 - 基于800Mhz PAM10
 - Cat 5e UTP < 20m
 - Cat 6 UTP < 30m
 - Cat 7 UTP 100m

- 
-
- 10G 是铜线连接的极限？
 - Sun Proximity Interconnect:
 - by Capacitance
 - 50Gb / pin
 - 并不能解决长距离传输问题



内容

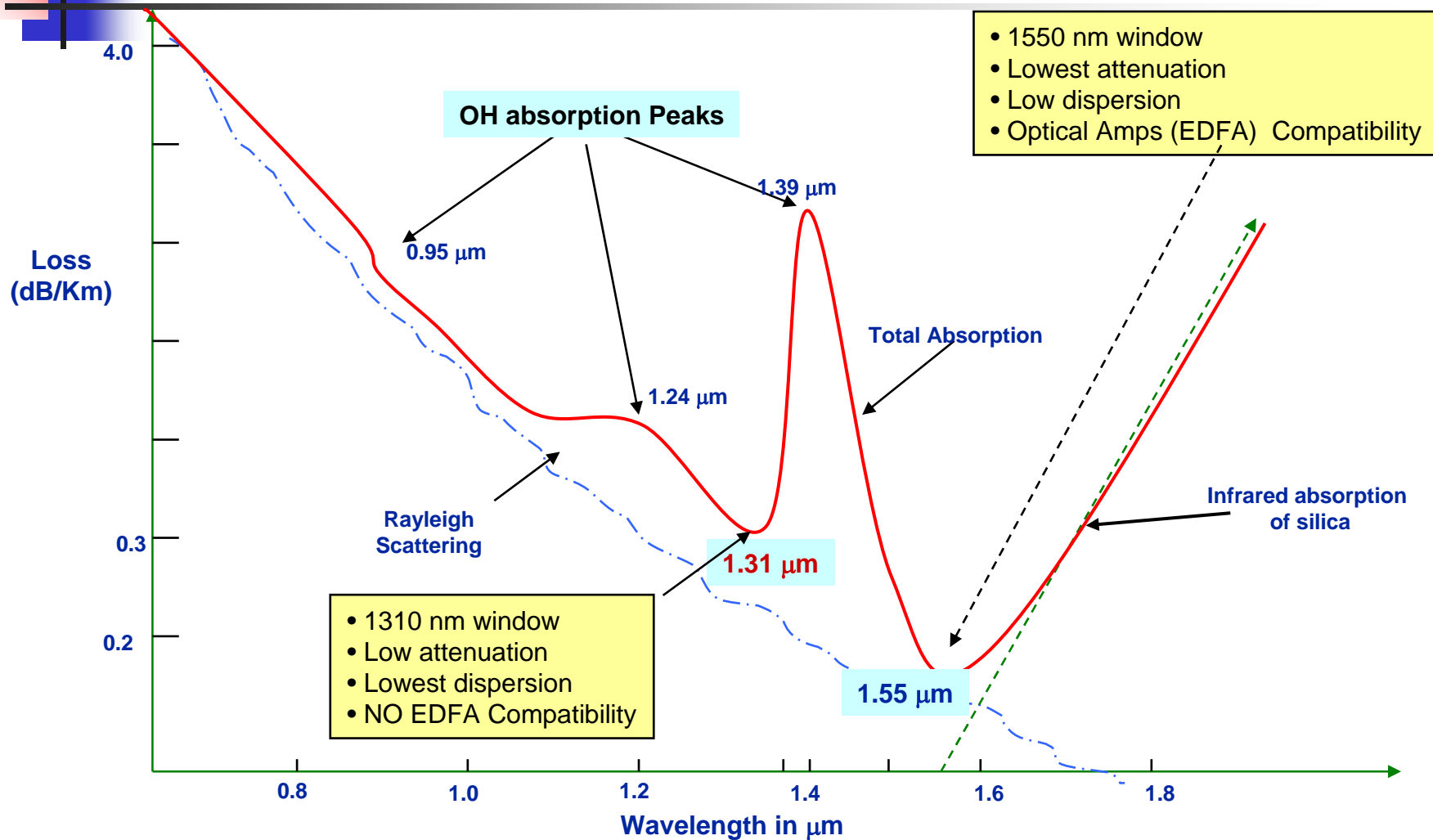
- 电信号的局限性
- 光互联技术
- 光交换技术
- 小雨点实验平台



光的传输潜力

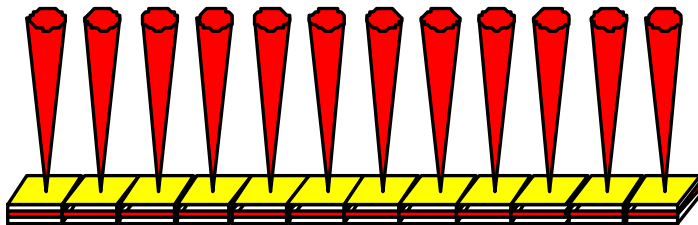
- 800nm --- 375 THz
- 光窗口：50T 可用带宽
- 空分复用的潜力：无串扰
- 传输不是问题，瓶颈在于如何产生信号

光纤传输中的光窗口



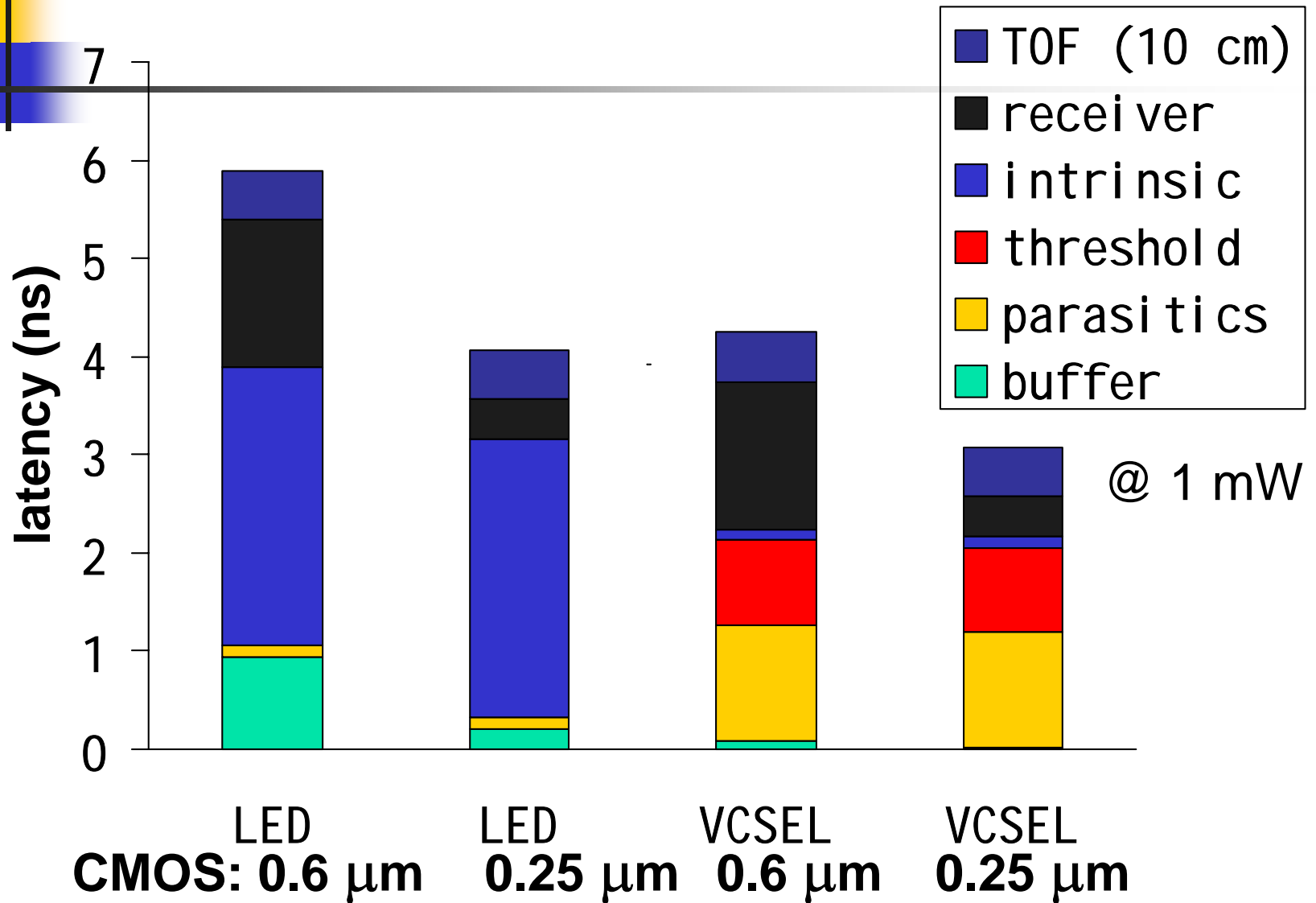
VCSEL: Vertical Cavity Surface Emitting Laser

- 谐振腔小，易产生微腔效应，低阈值激射
- 谐振腔比较短，动态调制频率高。
- 有源区截面呈园对称型，光束方向性好，易输出，易耦合。
- 出光方向垂直于衬底平面，适合于并行光互连和信息处理。
- 器件体积小($250\mu\text{m}$)，可实现高密度集成



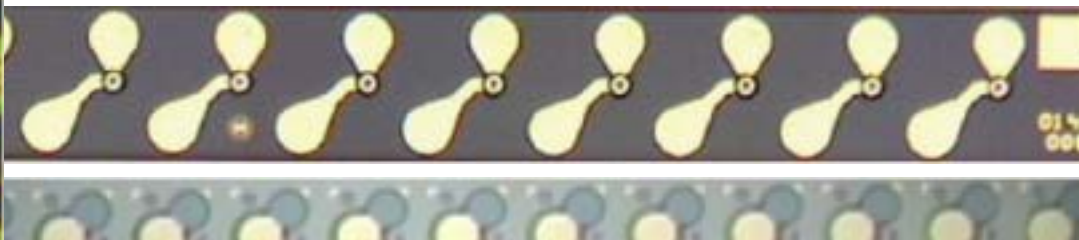
• GaAs工艺，还不能与CMOS结合

Total optical link latency

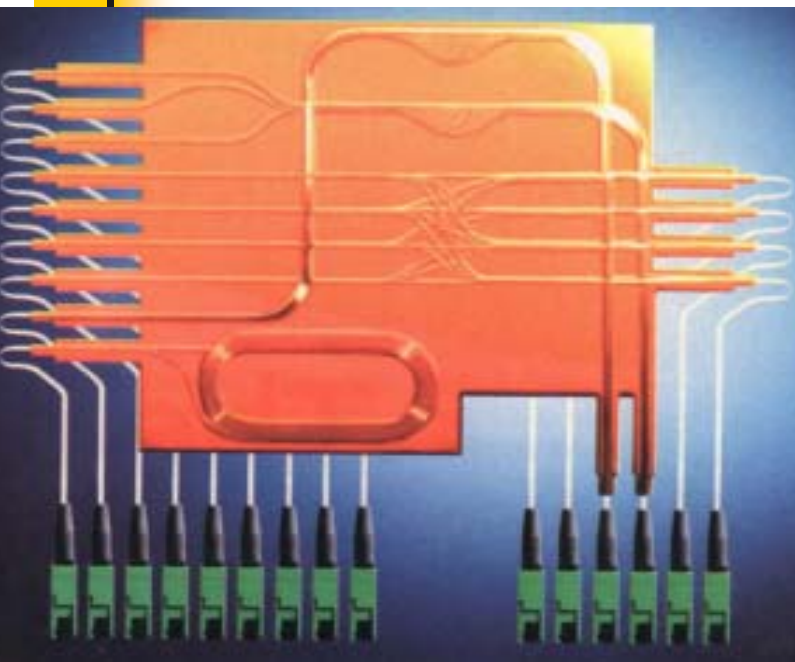


当前的技术水平

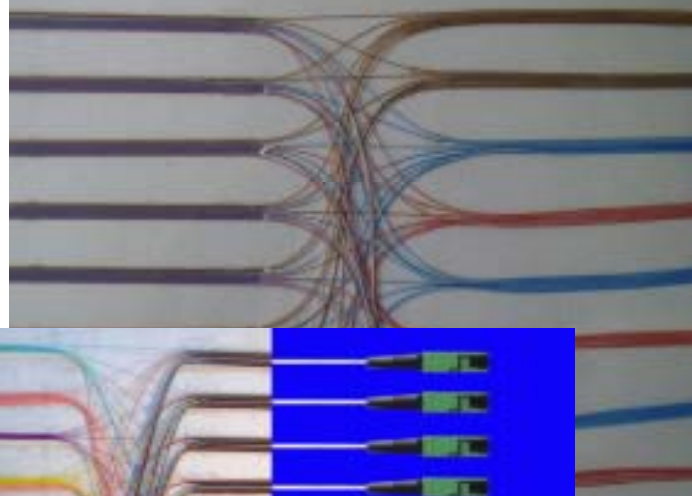
- 1x12 2.5~3.125G 主流
- 10G 产品化
- 40G 产品样品发布



带状光纤和光背板（当前）

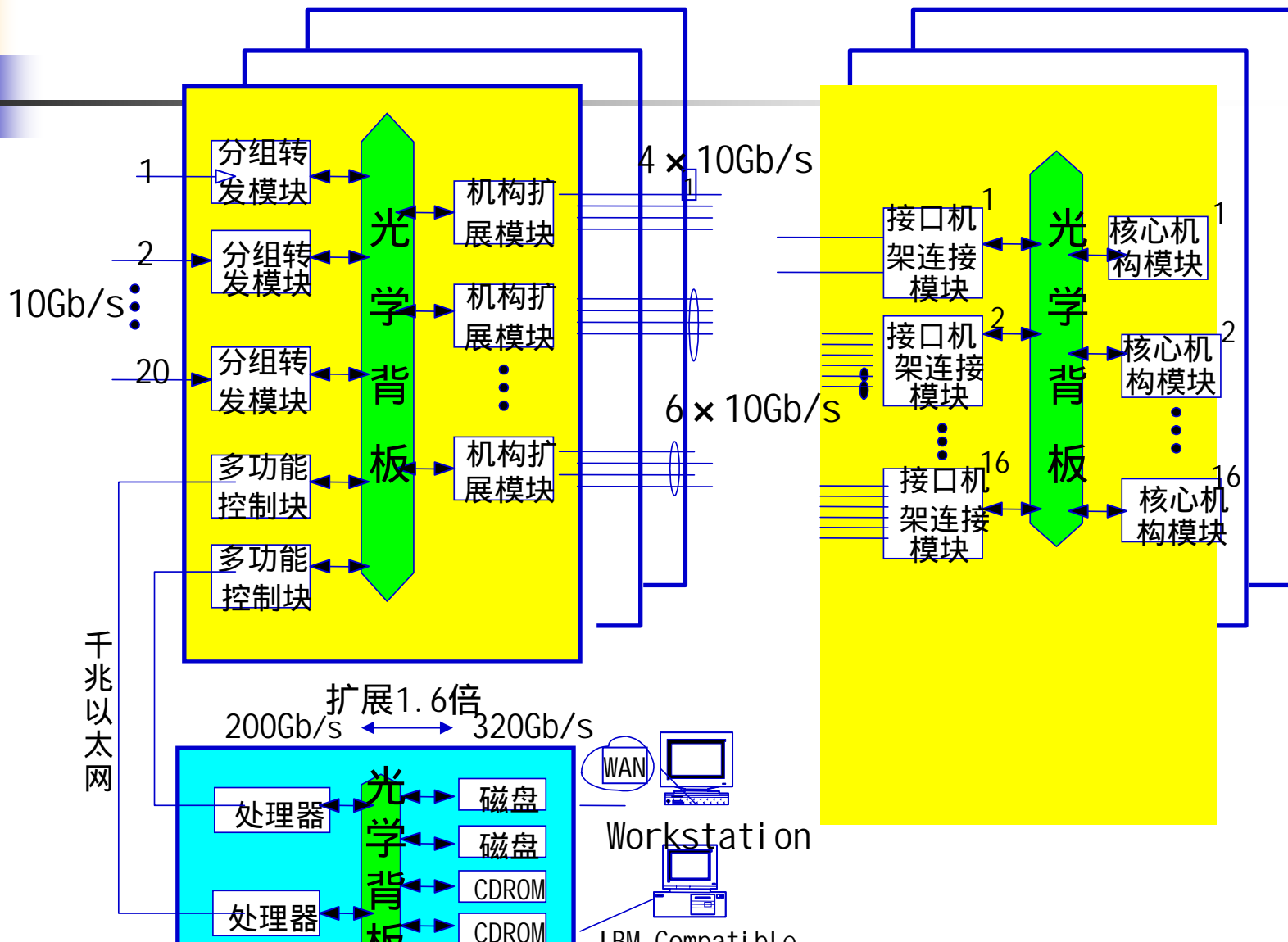


用于网络服务器或并行处理的
 8×8 光学 crossbar 网络

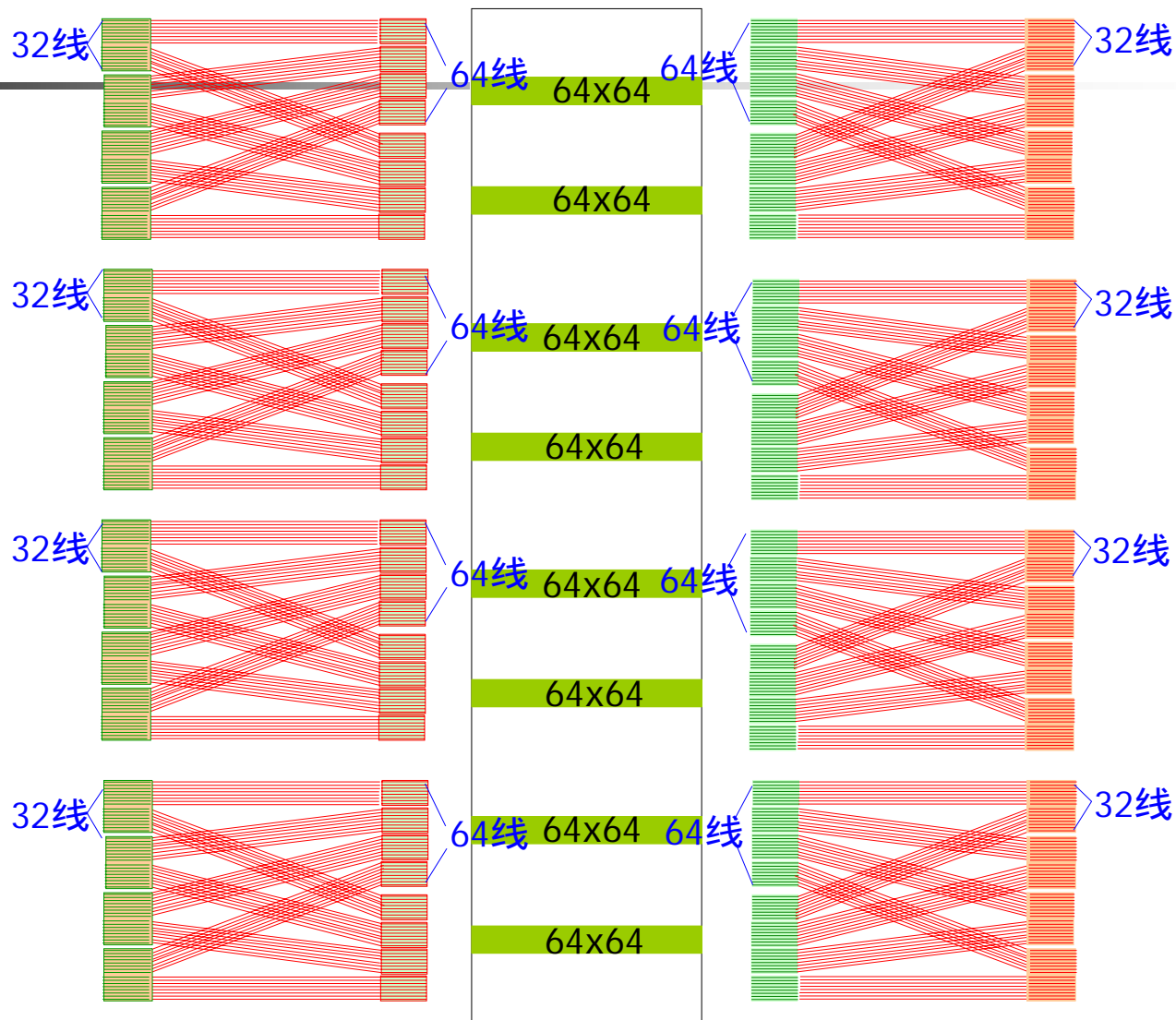


高速路由器中的光背板

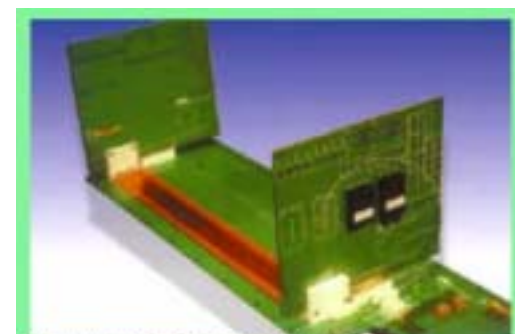
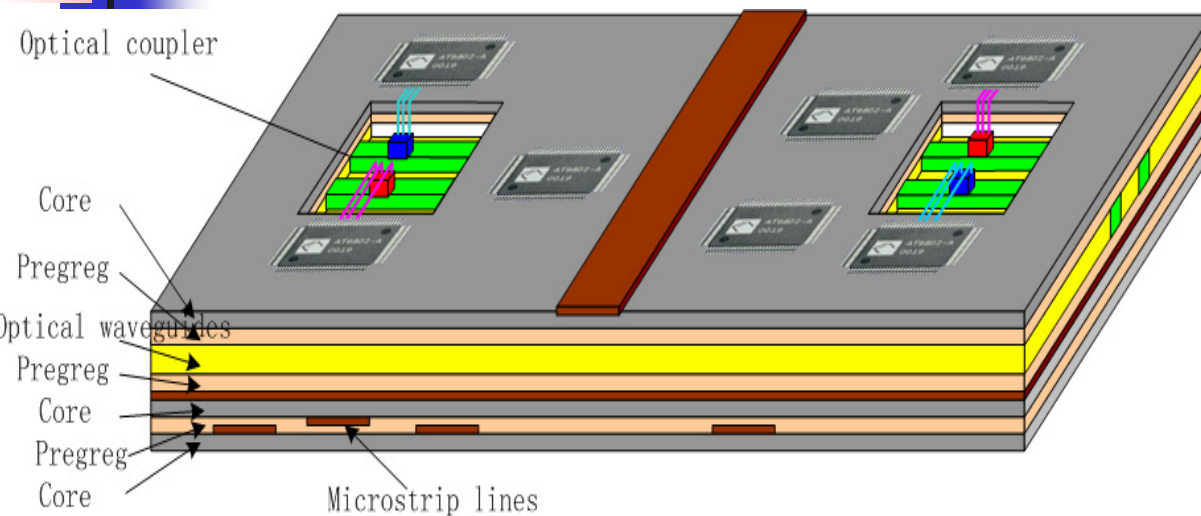
最多可达8个机构机架



320(x2)G扩展到1.28(x2)光纤互连网络背板



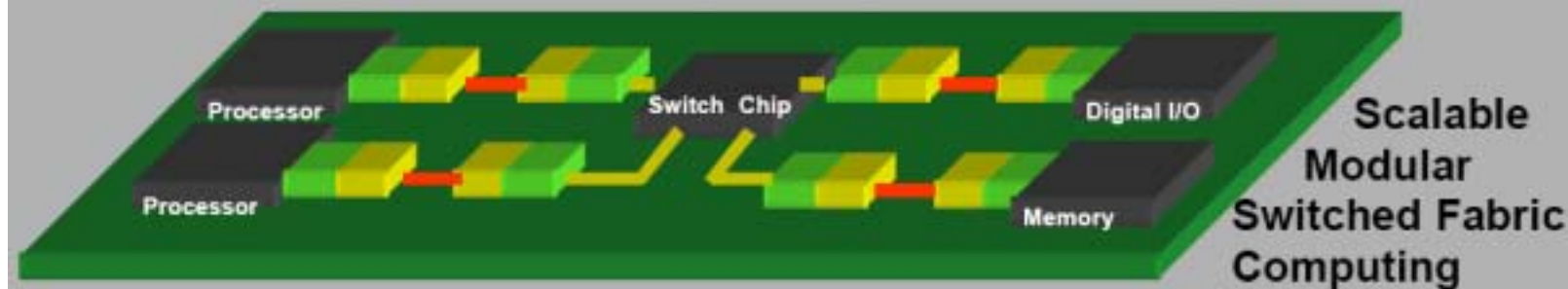
基于光波导的复合PCB/POB



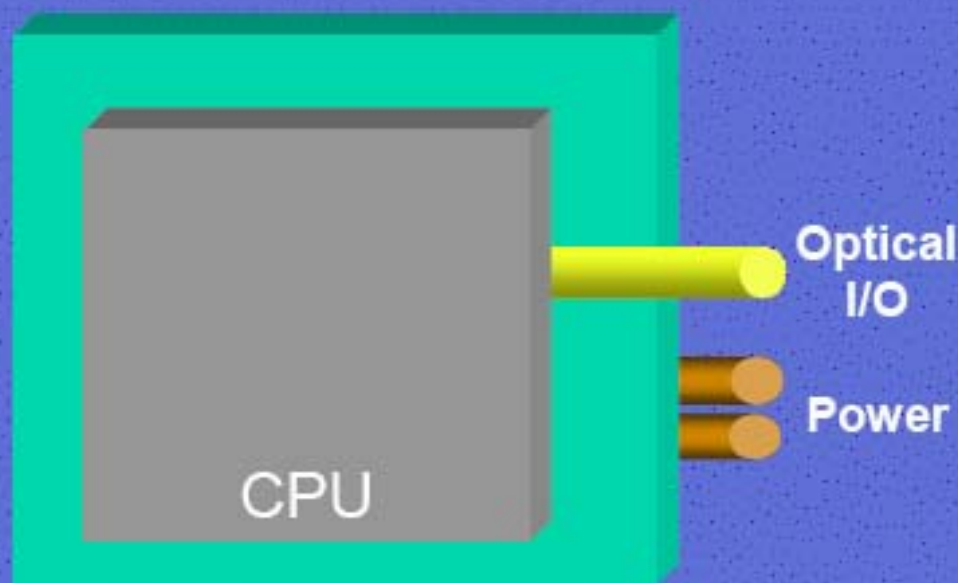
Courtesy of Optical Crosslinks, Inc.

Printed Optical Board

Fiber to the Processor*

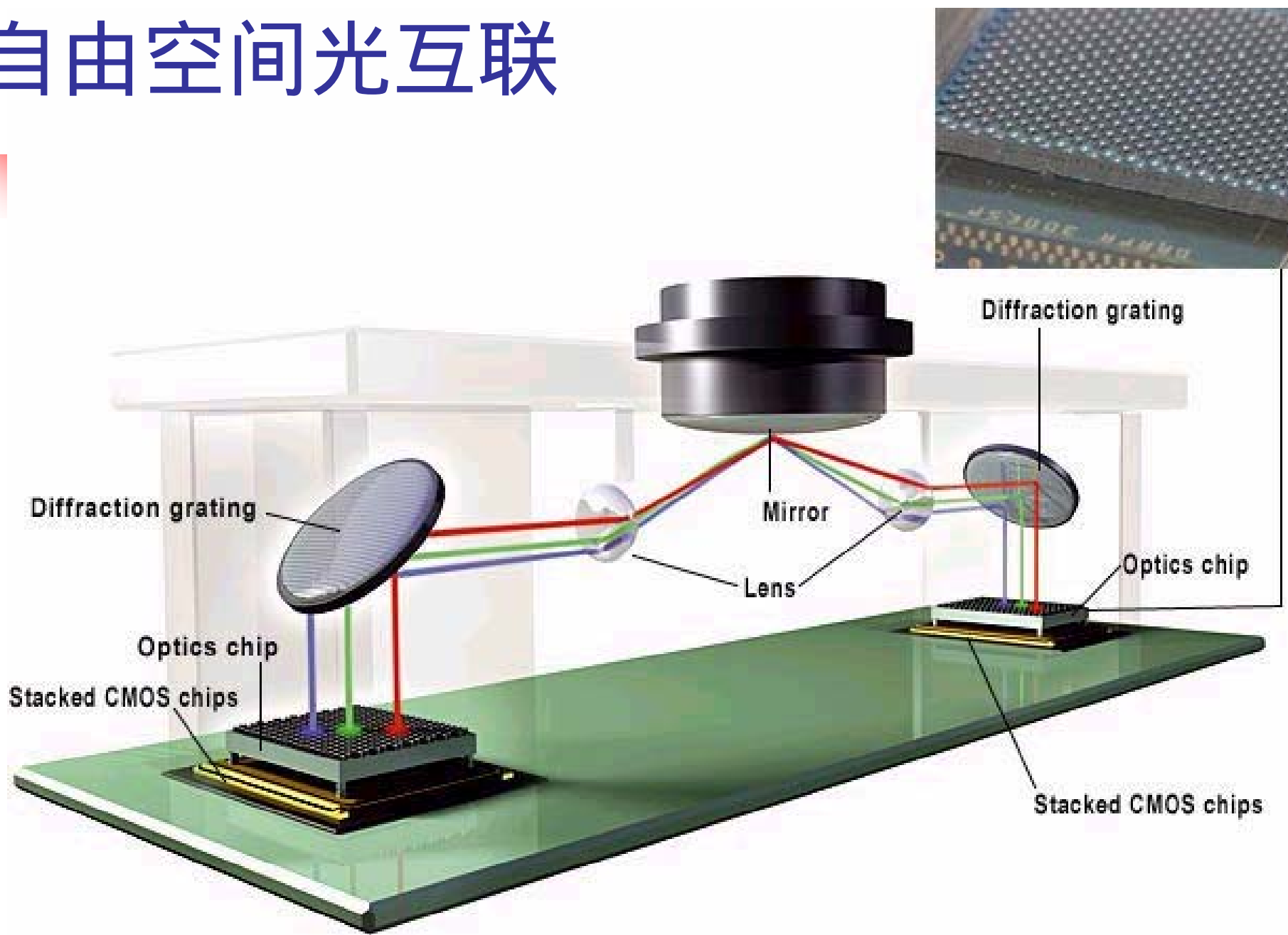


The Future: Encapsulated Processors



Complete Triad packaged together =
Optimum *system* performance

自由空间光互联

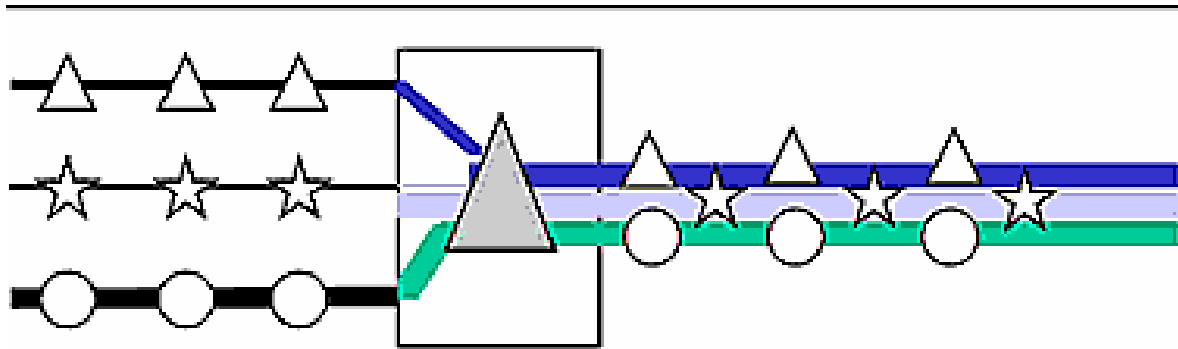




趋势

- 背板之间并行光纤互联技术已经成熟，并广泛用于通信领域高速交换机/路由器中
- 主流的3G产品相比于铜线连接并不具有很大的优势,10G以上光互联将显示优势
- 如果VCSEL与CMOS技术结合，将会大大降低光器件的使用成本

更多的带宽？ DWDM



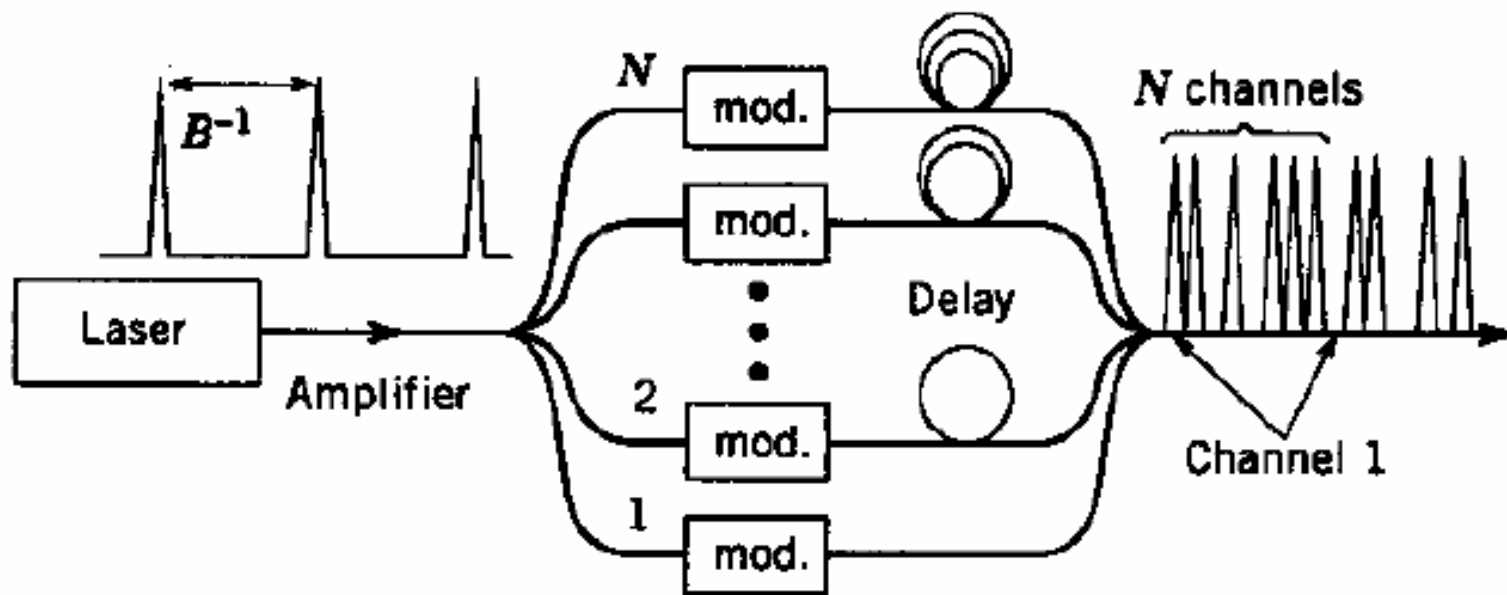
- Dense Wavelength Division Multiplexing
 - Merges optical traffic onto one optical fiber
 - Allows high flexibility in expanding bandwidth
 - Reuses existing optical signals
 - Not dependant on signal, bit rate or format



DWDM的潜力和可用性

- 单信道最高可达80G
- 200个以上信道
- 10Tb 以上容量
- 光放大器、复用/解复用、波长变换是一套复杂的系统，不适合近距离点对点的应用。
- 目前空分交换更经济实用。

更多的带宽？OTDM



- 仍处于试验阶段
- 有潜力达到1Tb/s



内容

- 电信号的局限性
- 光互联技术
- 光交换技术
- 小雨点实验平台

OEO vs OOO

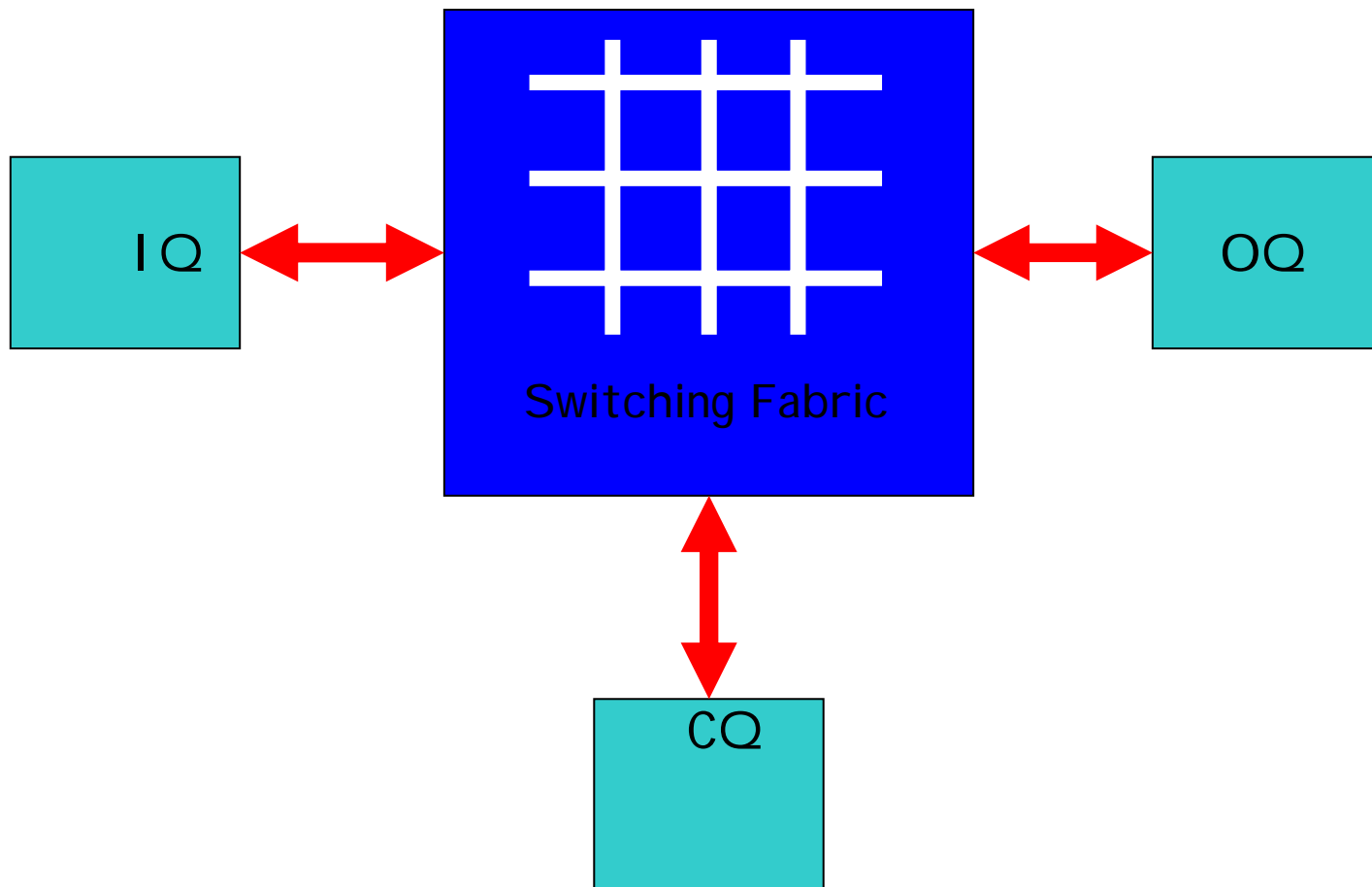


- 成熟、低成本（但不包括光/电转换）
- 灵活，包交换
- 带宽受限（40G），端口受限



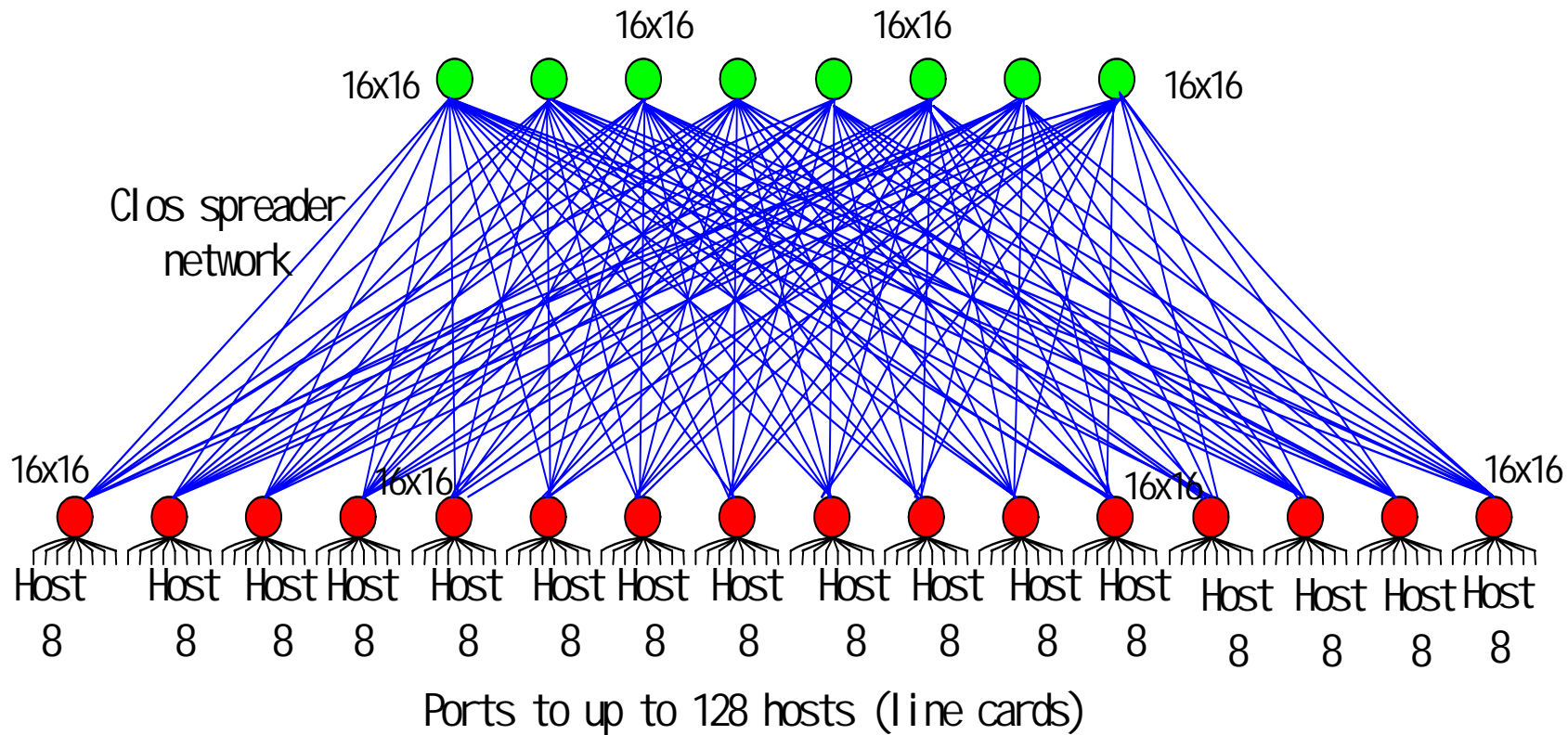
- 粗粒度，波长变换
- 无带宽限制，透明传送

NxN 交叉开关——交换的核心



基于XBar16的128端口Myrinet 交换机结构

Spine of the Clos Network

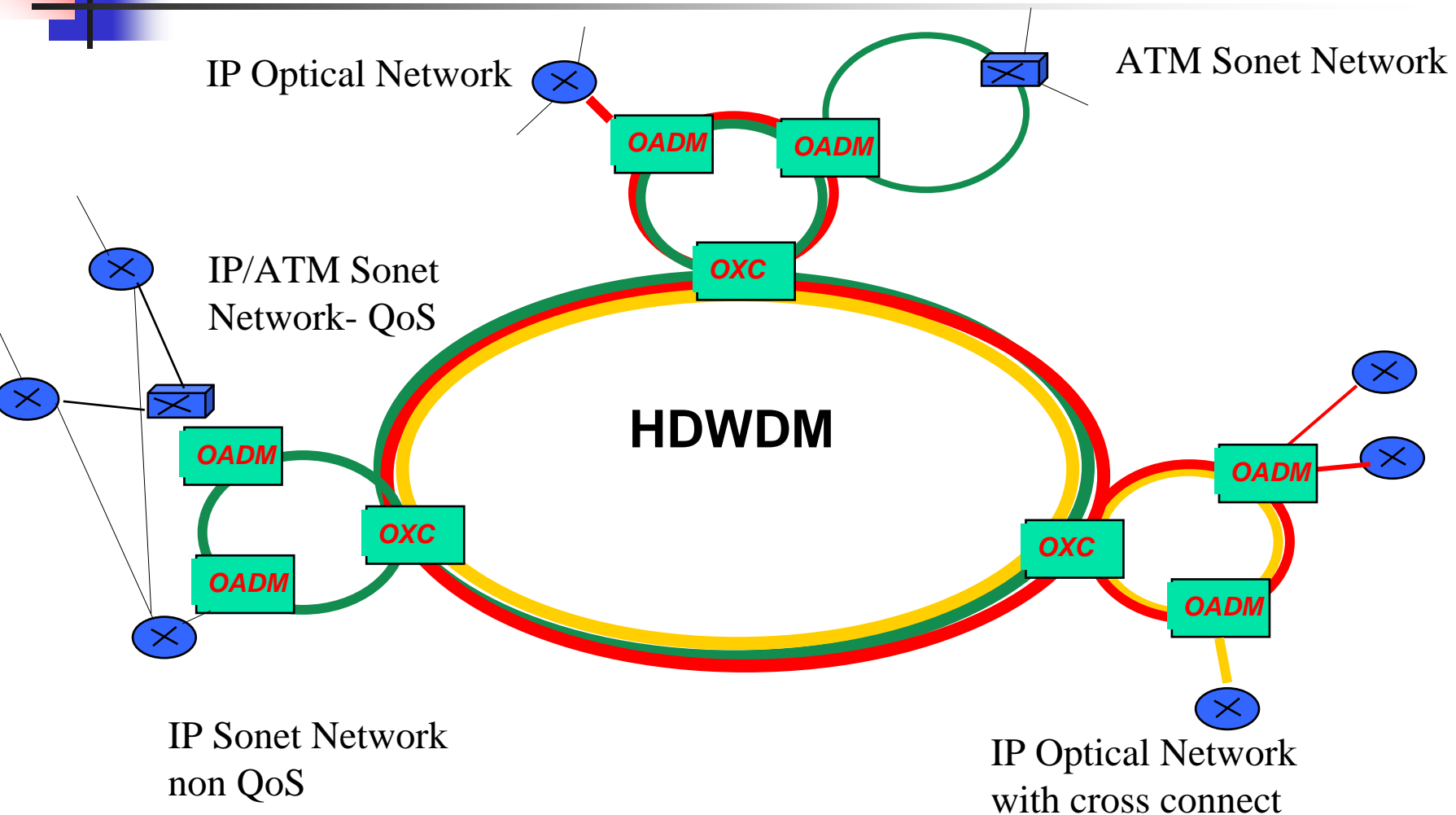




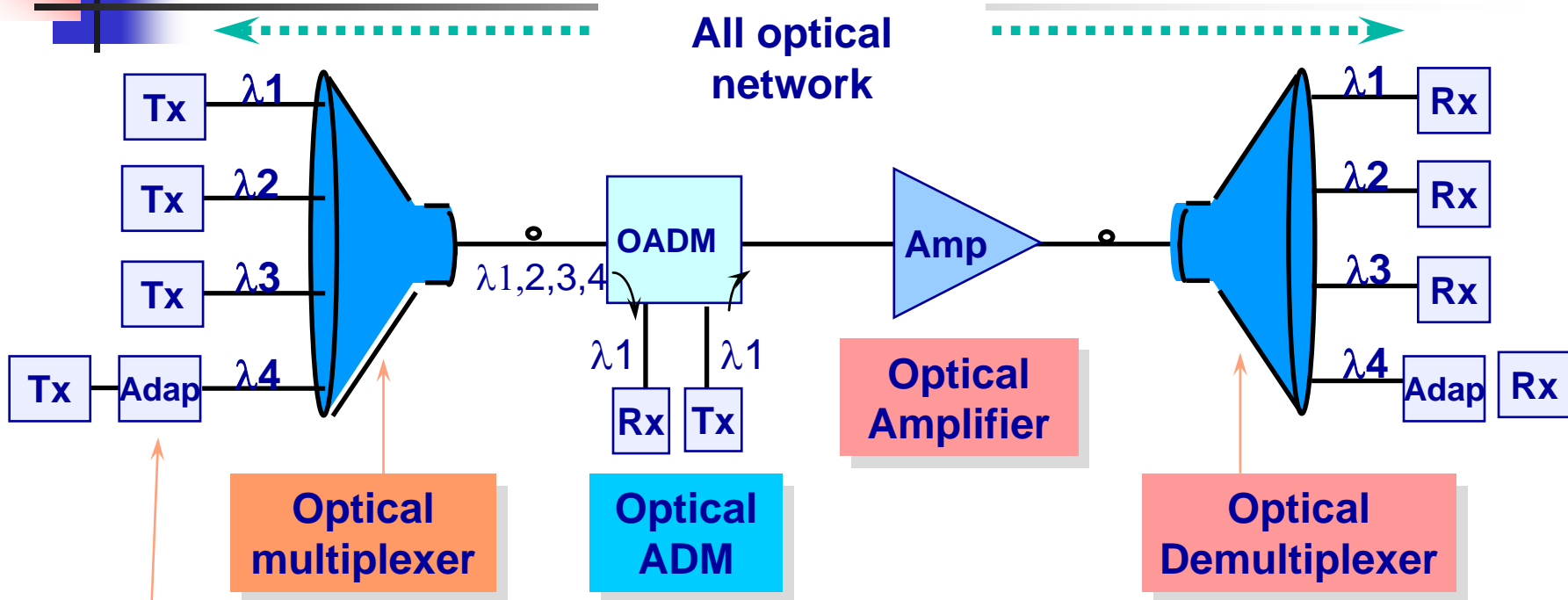
电交叉开关目前的技术水平

- 每路2.5~5Gbps
- 产品：144x144x5G
- 预测：可达1000端口左右
- Tb/s 是极限？

Future Optical Internet



Optical Add-Drop Multiplexer



OADM – Optical Add-Drop Multiplexer
Allows one or more wavelengths to be dropped or added to the linear system

Transponder

Transponder - or wavelength adaptor needed if the Tx equipment does not generate WDM ready wavelength



Optical Cross-Connect(OXC)

Free Space

- Mechanical

- Liquid Crystal

- holograph

Guided-Wave Integrated Optics

- Thermo-optic

- Electro-optic

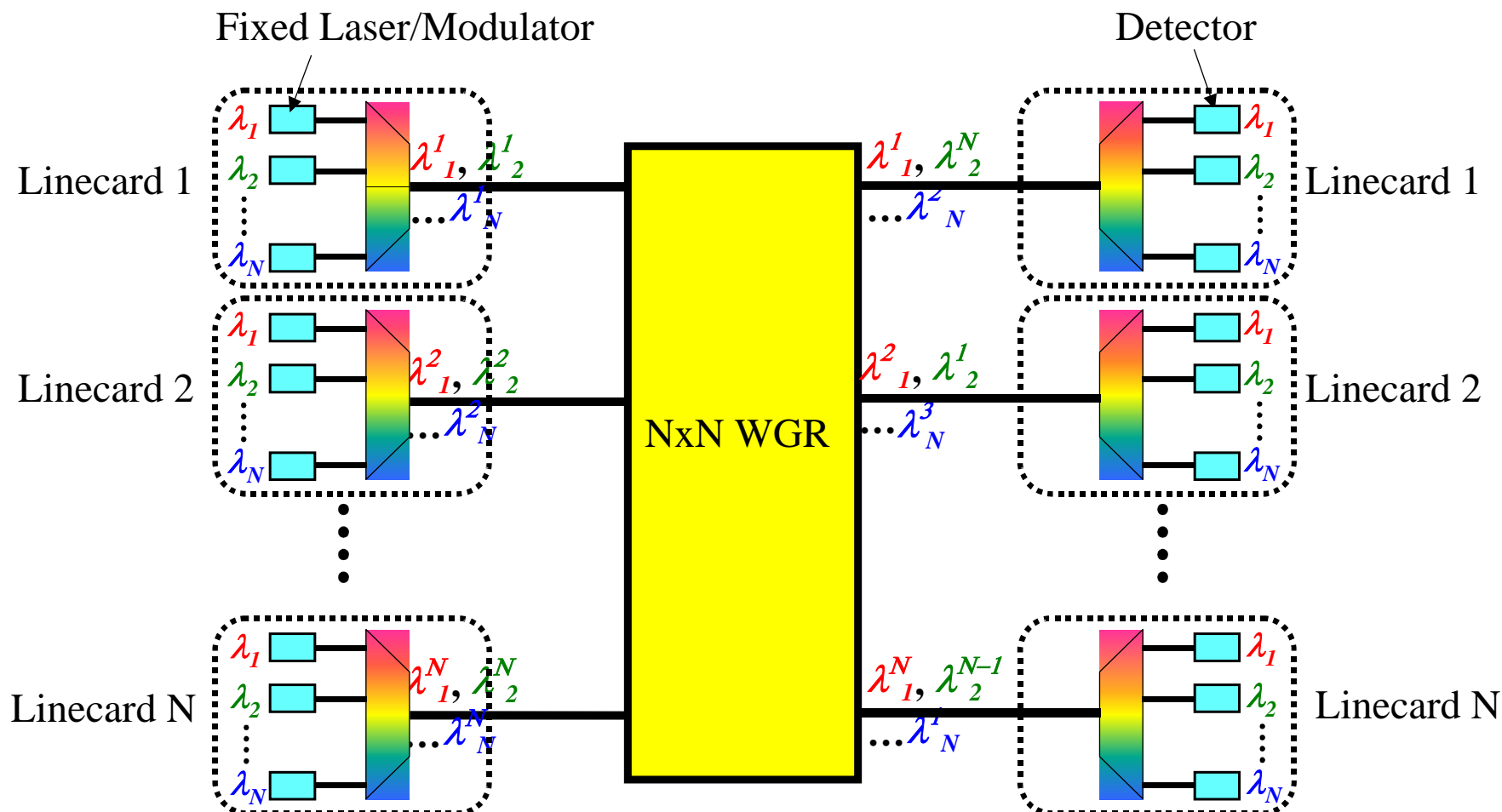
- AWGR

Guided-Wave Active Component

- SOA space switches

- MEMS

Waveguide Grating Router



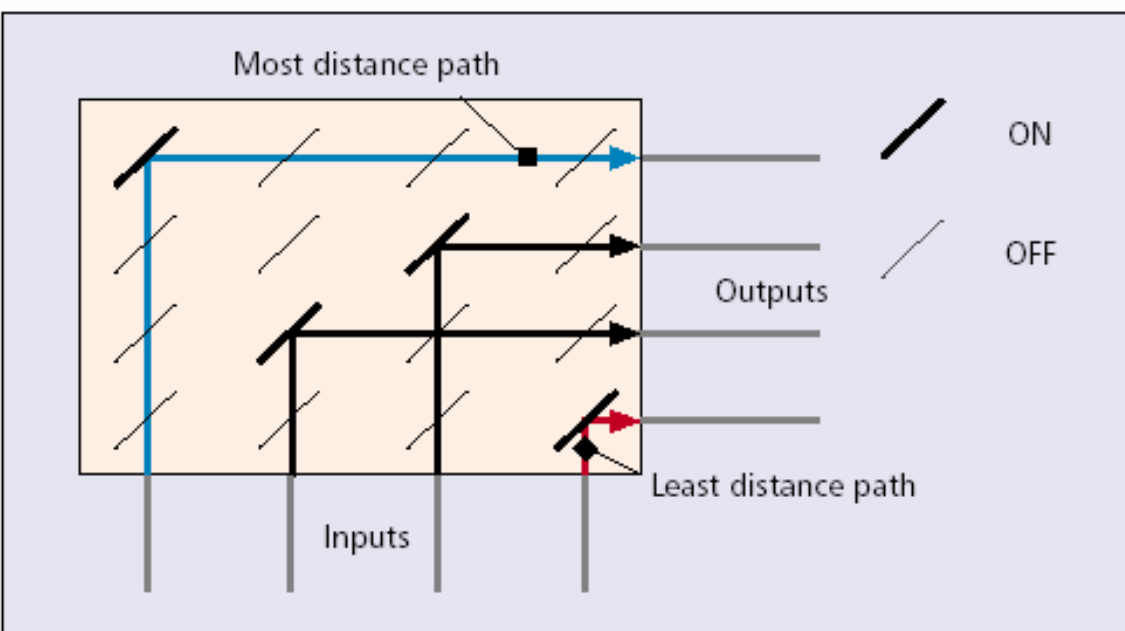


Figure 3. A 2D crossbar switching architecture.

MEMS: Micro-Electro-Mechanical-Systems

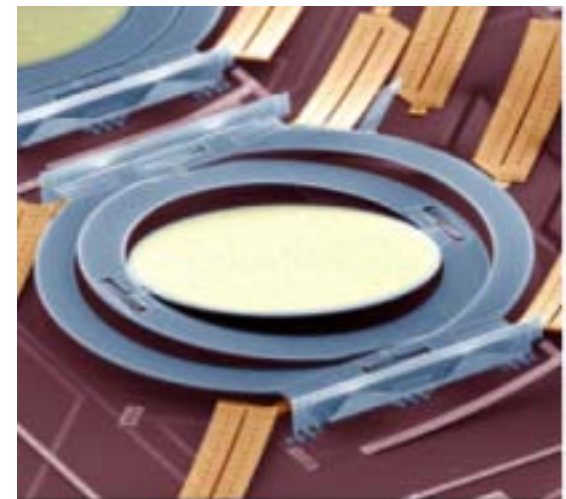
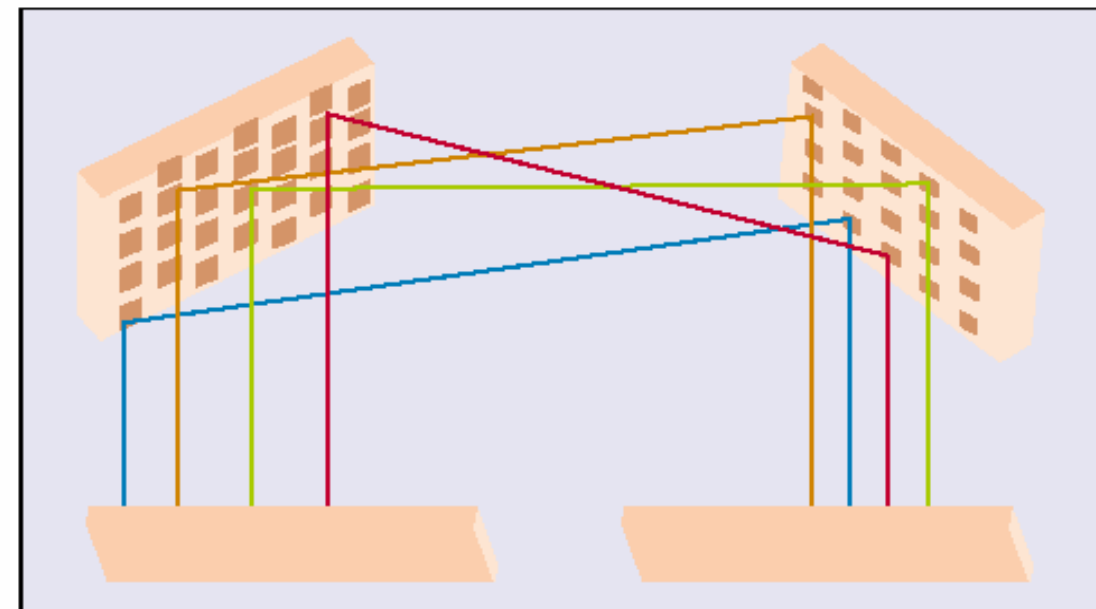
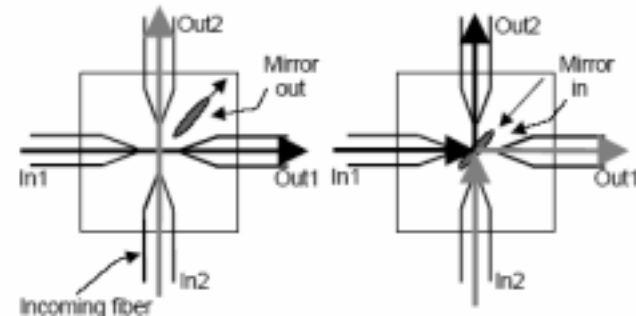
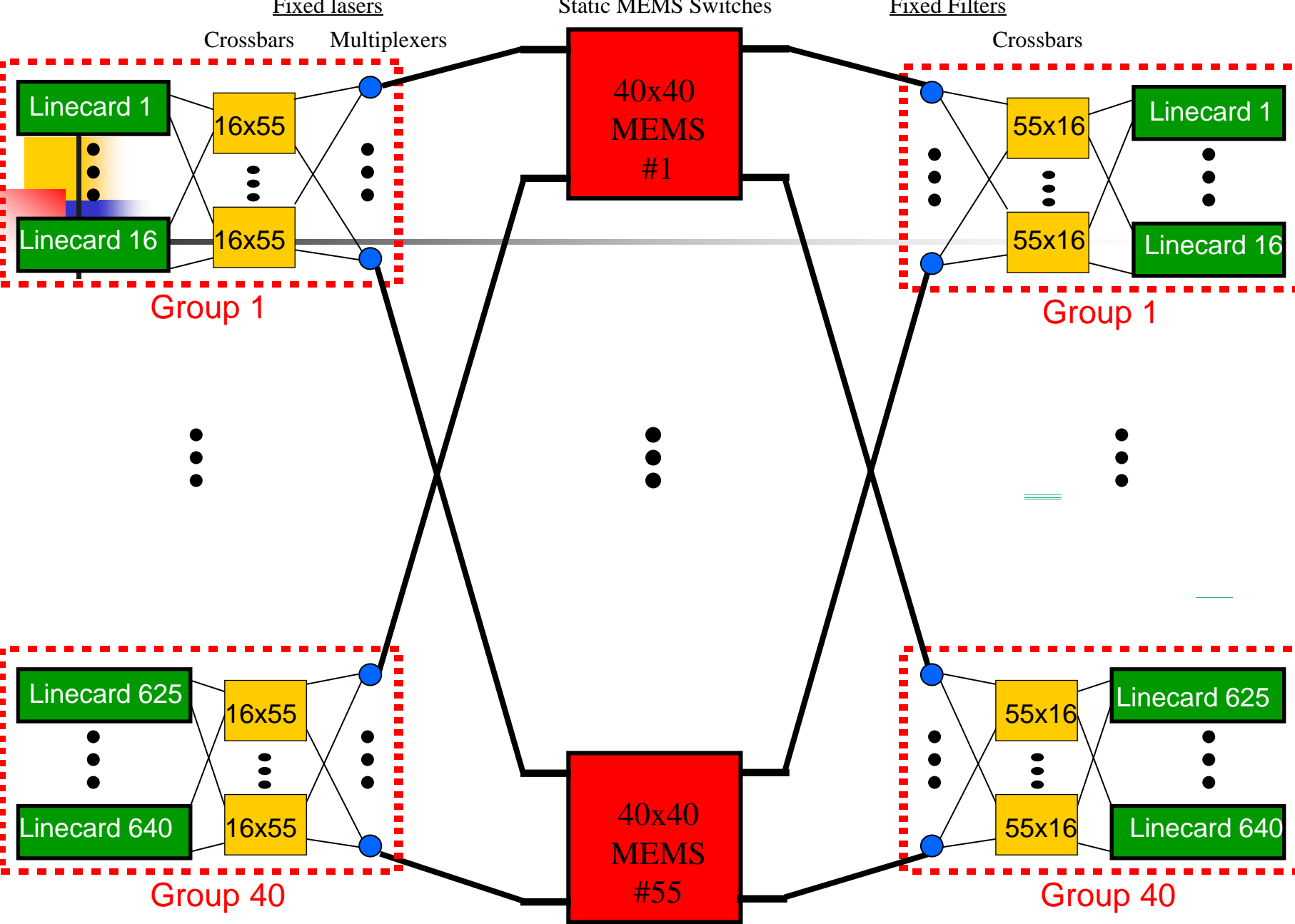


Figure 5. A 3D schematic illustration of a multi-layered optical switch.



当前的技术水平

- 1296*1296 MEMS (lucent)
- 模块：8x8 (10 ms) 32x32 (20 ms)
(OMM inc)



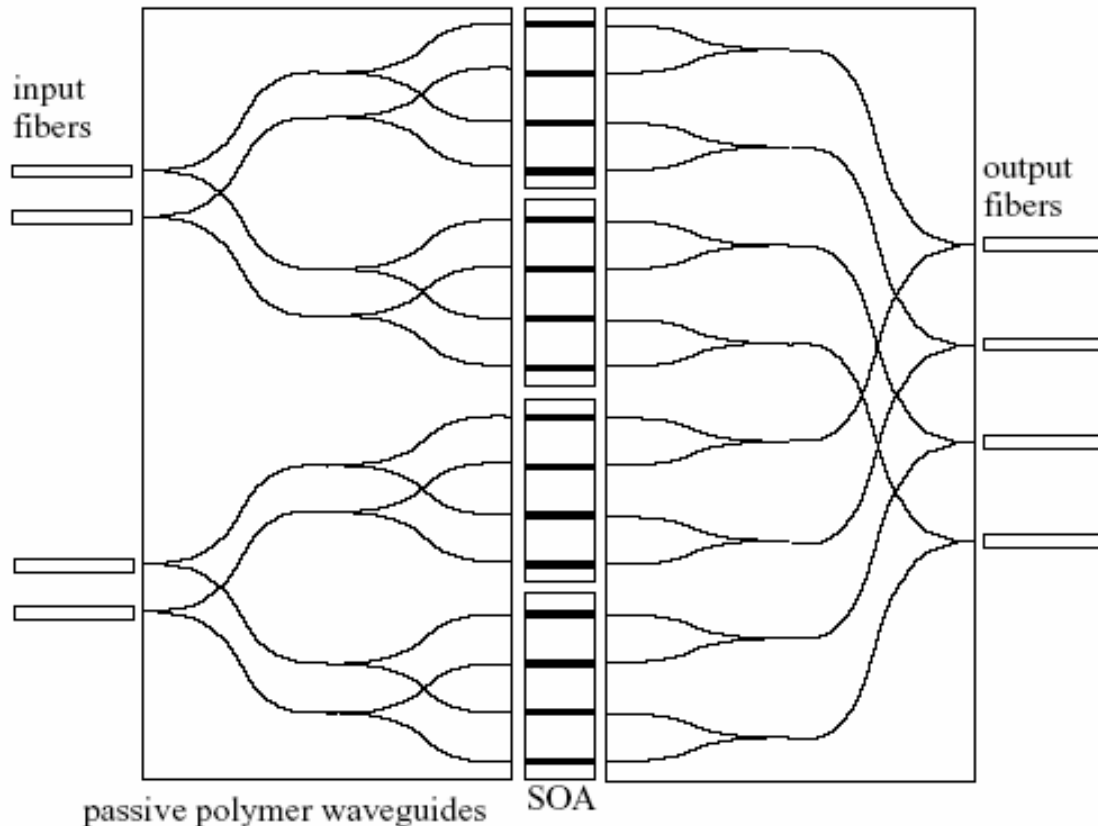
A 625x160G switch fabric 100Tb



MEMS 的其他应用

- 传感器: 压力、加速度
 - 喷墨打印头: HP
 - 数字投影仪: TI DLP/DMD
 - 高清晰显示: Iridigm > 200 DPI
 - ...
-
- 基于半导体技术，是否会有类似Moore定律的发展？

Semiconductor Optical Amplifier(SOA) switch



- Active
- 高速
(2ns~200ps)
- 噪声/失真大
- 仍处于研究阶段



光交换的应用前景？

- 目前不具有高速动态包交换的能力
- 交换带宽不受限制，潜力很大
- 有可能在ms量级改变连接结构
- 可选技术多，有待进一步成熟/淘汰
- 成本有待下降
- 看好MEMS？



内容

- 电信号的局限性
- 光互联技术
- 光交换技术
- 小雨点实验平台



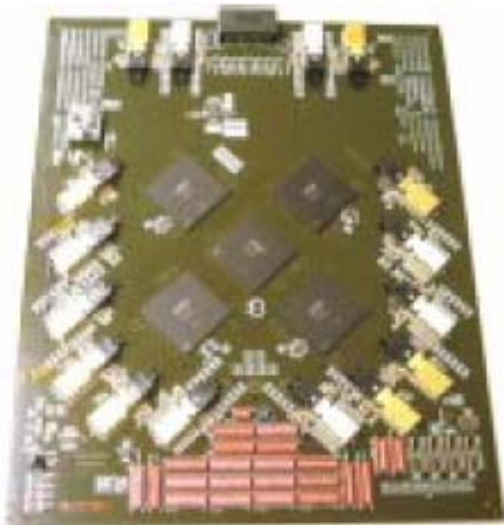
小雨点光互连试验系统

- 验证光互连技术的可行性
 - 点对点带宽 $\geq 16\text{Gb/s}$
 - 实验光电/电光转换延迟的影响
- 验证CPU和Memory分离的思想
 - 从Memory总线开始联接光信号
- 进行新OS存储管理系统实验

RWCP RHINET-3 SW

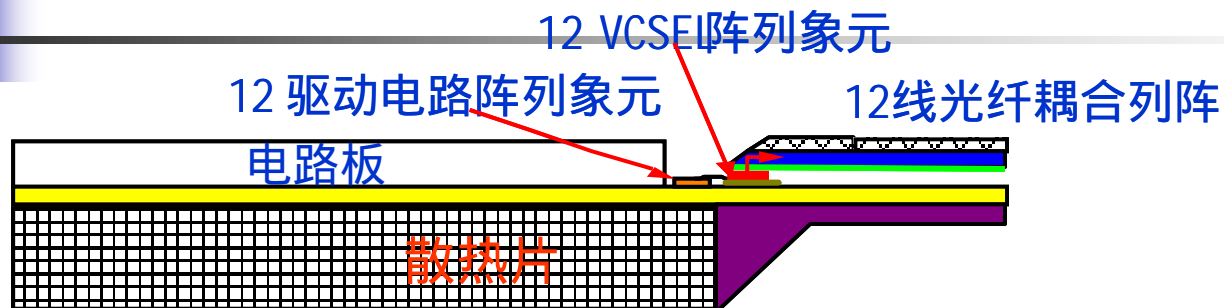


- 8X8 10Gbps(1.25x12) crossbar
- 0.14 CMOS ASIC SW LSI
- Martini/DIMMnet-1 64bx133M

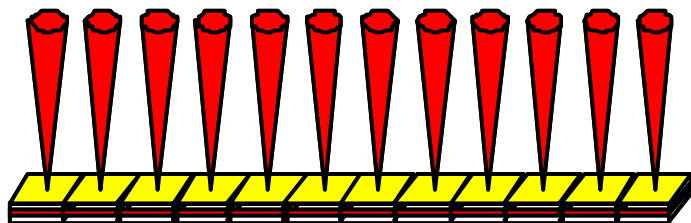


华中科技大学的光互联技术基础

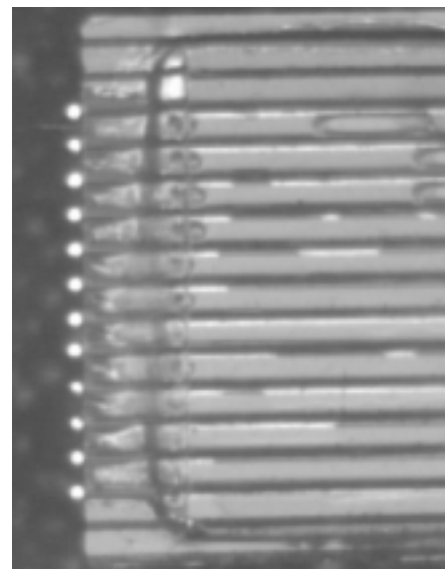
已研制的 2.5Gbps 1x12 VCSEL/PIN 发射和接收模块



VCSEL 并行发送模块

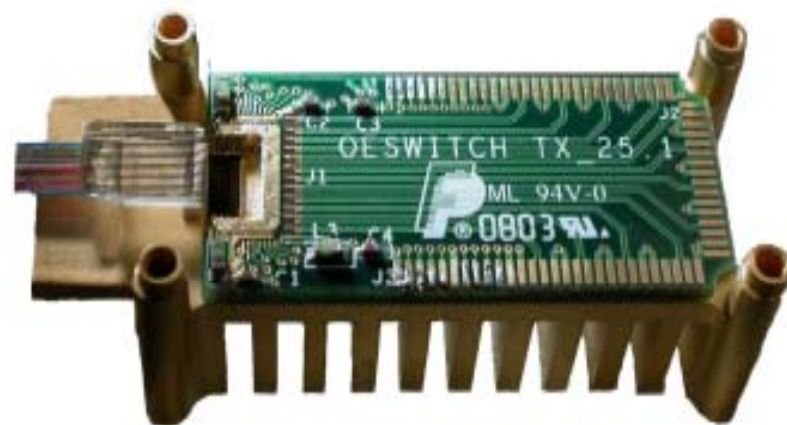
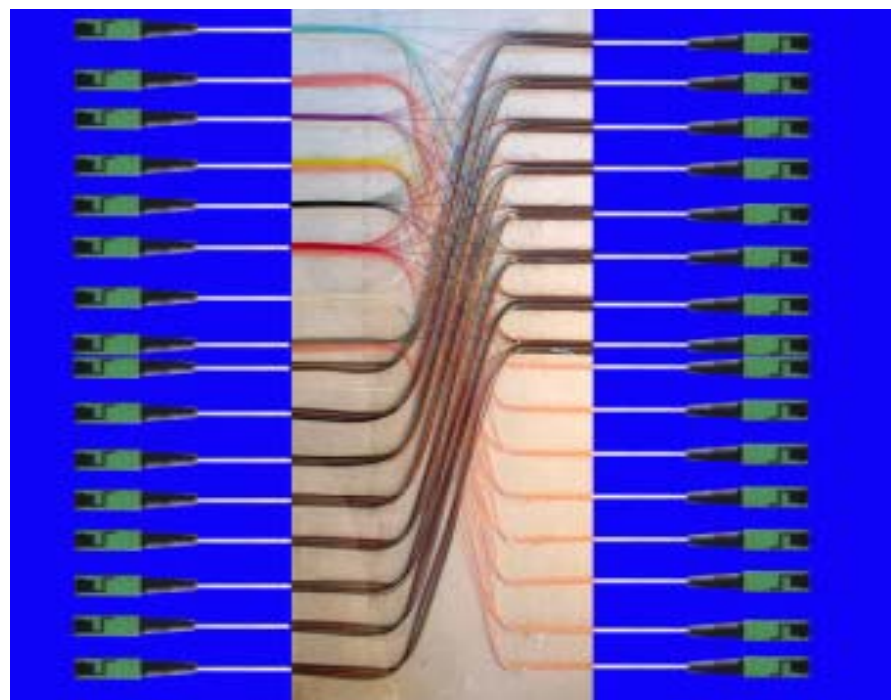


1xN VCSEL面发射激光阵列



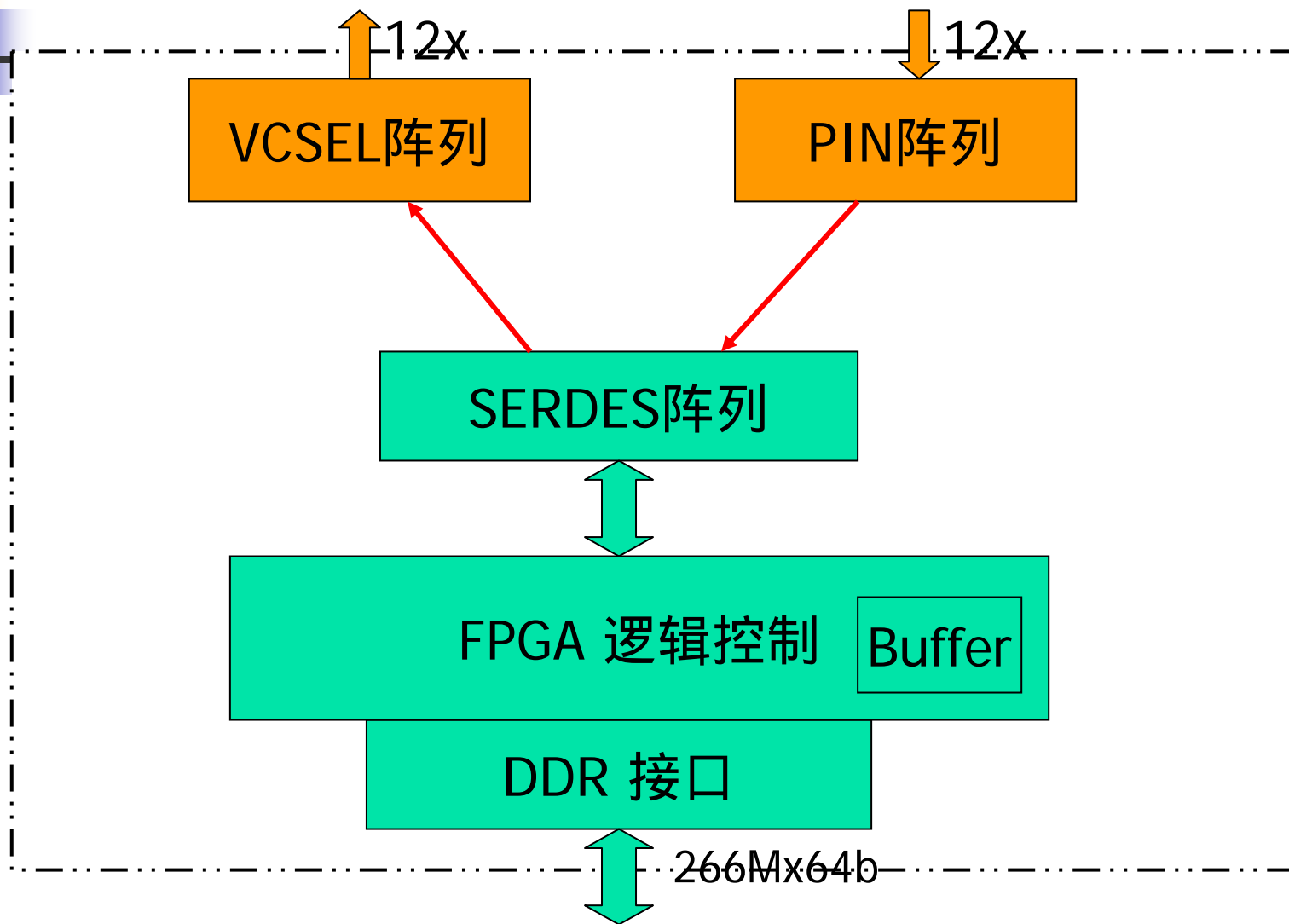
带45度楔角的一维
光纤列阵耦合芯片

From 华中科大激光重点实验室



12x2.5G VCSEL/PIN

小雨点通信板结构



接口选择的考虑

- 目前最好的PCIx $133\text{M} \times 64\text{b} = 8\text{Gb/s}$
- AGP 8x: $32\text{b} \times 133\text{M} \times 4 = 16.8\text{Gb/s}$:
 - I/O接口，更适合通信(中断/DMA)
 - 较新，实现复杂
- DDR 266: $133\text{M} \times 2 \times 64 = 16.8\text{Gb/s}$
 - 优点：简单，有先例，适用广，可扩展
DDR333 +
 - 被动设备，难实现有效消息到达通知/DMA
- Hyper Transport: N/A



主要难点

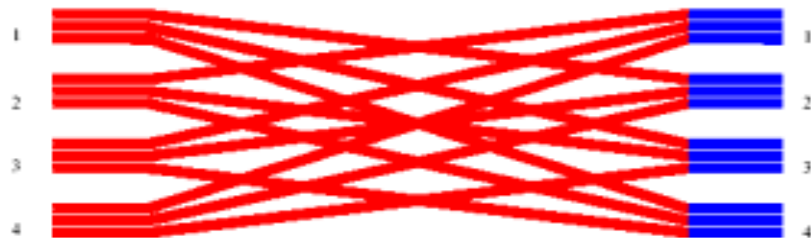
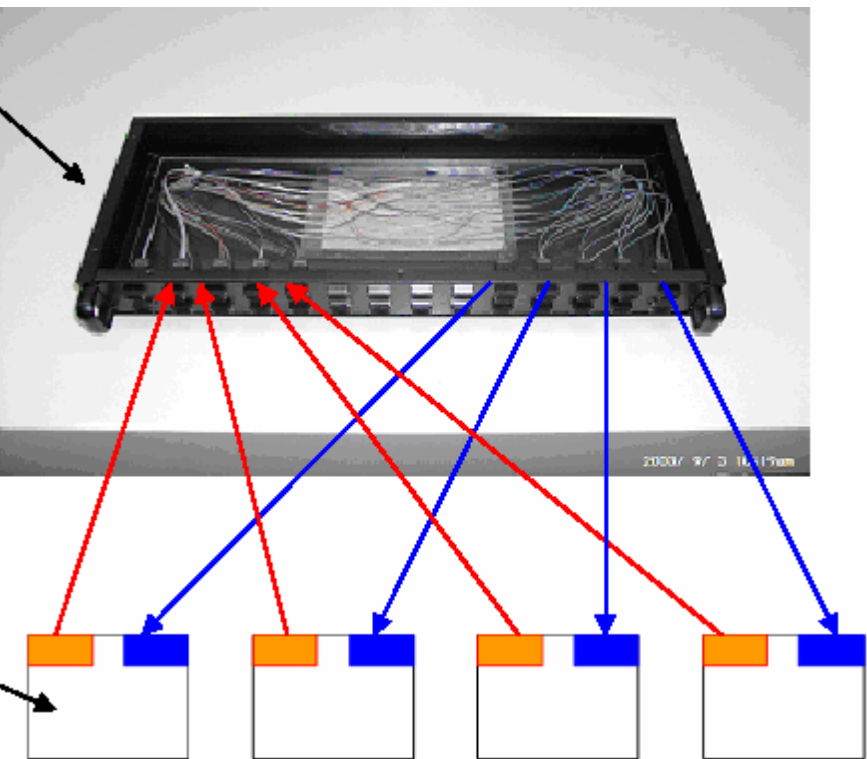
- 用FPGA实现高速 DDR接口
- 板上多路高频信号布线
- 操作系统存储管理模块的修改



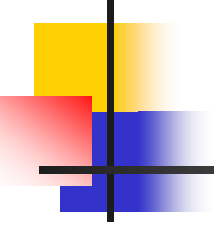
Xilinx VirtexIIpro 2VP40

- 4万个逻辑单元
- 3Mb RAMbit
- 2 PowerPC 405 core
- 12 Rocket I/O(2.5G)
- FF1152封装
- \$\$\$!

小雨1号



- 1:1 16G
- 4x4 x8G



问题？