关于“Ram on Net”的设想

Version 0.1b

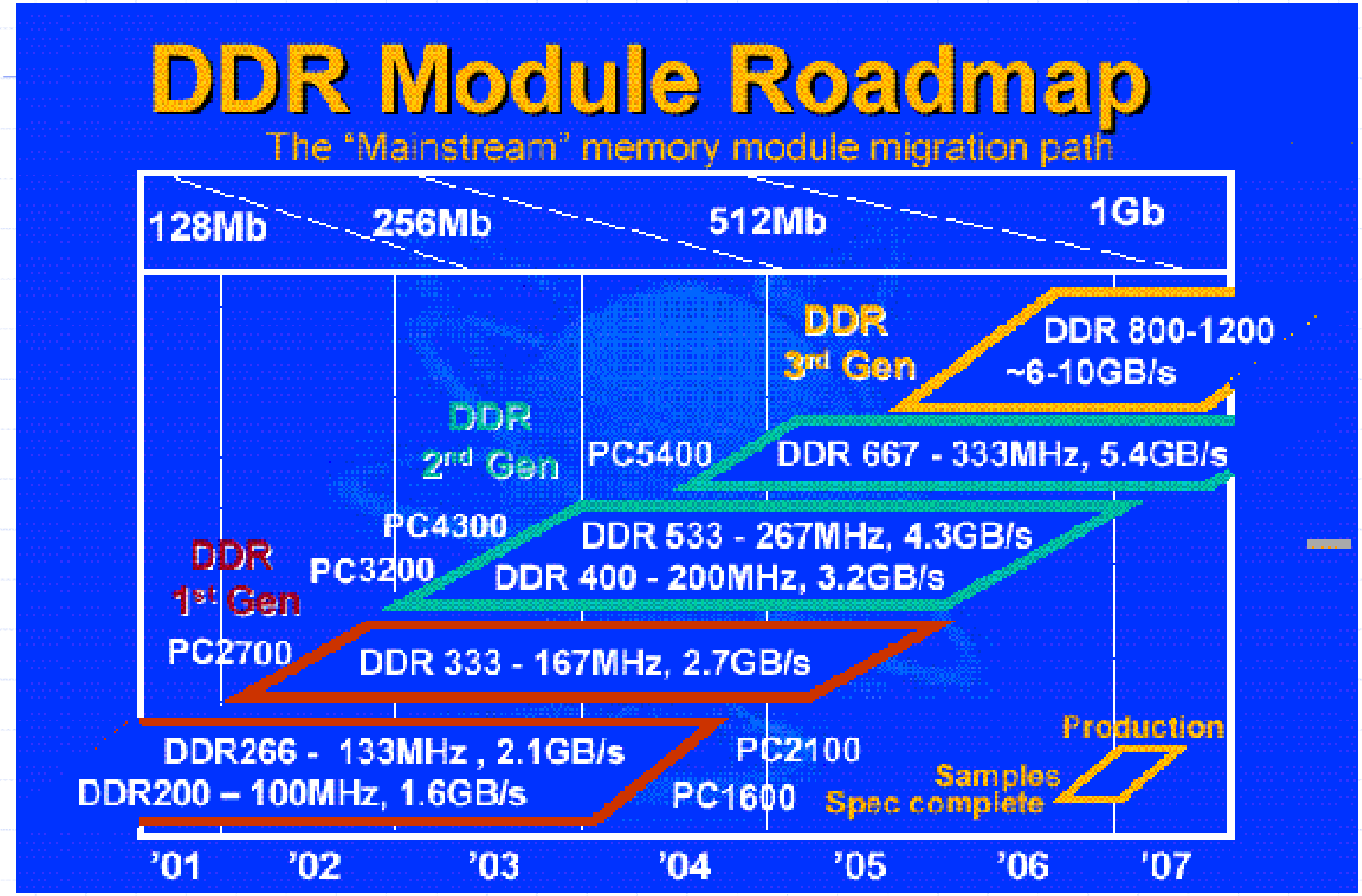
中科院计算所

2003.2.26

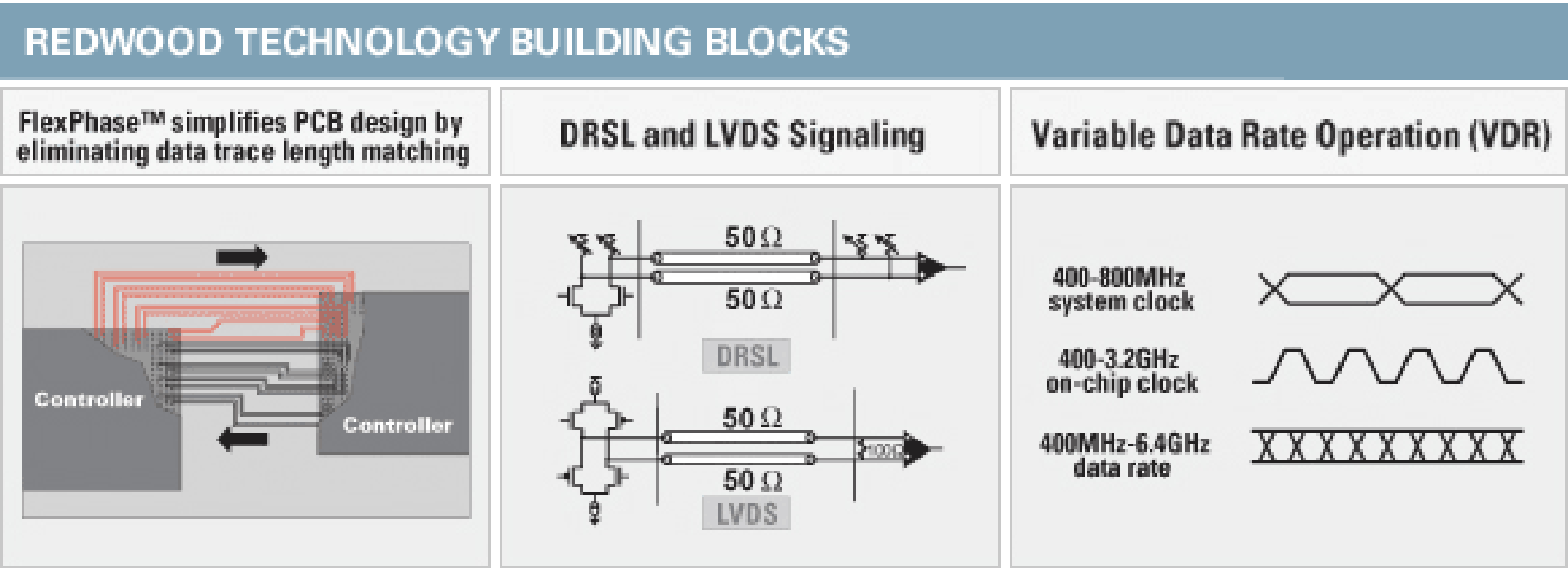
- 当前dram技术发展趋势
- 关于Ram on Net的三个设想
- Ram on Net对系统软件的需求探讨

2005年 主流DRAM性能预测

- 带宽
  - RDRAM 600MHz， 9.6 GB/s
  - DDR-II 400MHz， 6.4 GB/s
  - --主板频率提高受到限制
- 延迟
  - tRAC 20~40ns
  - --与传输带宽相关

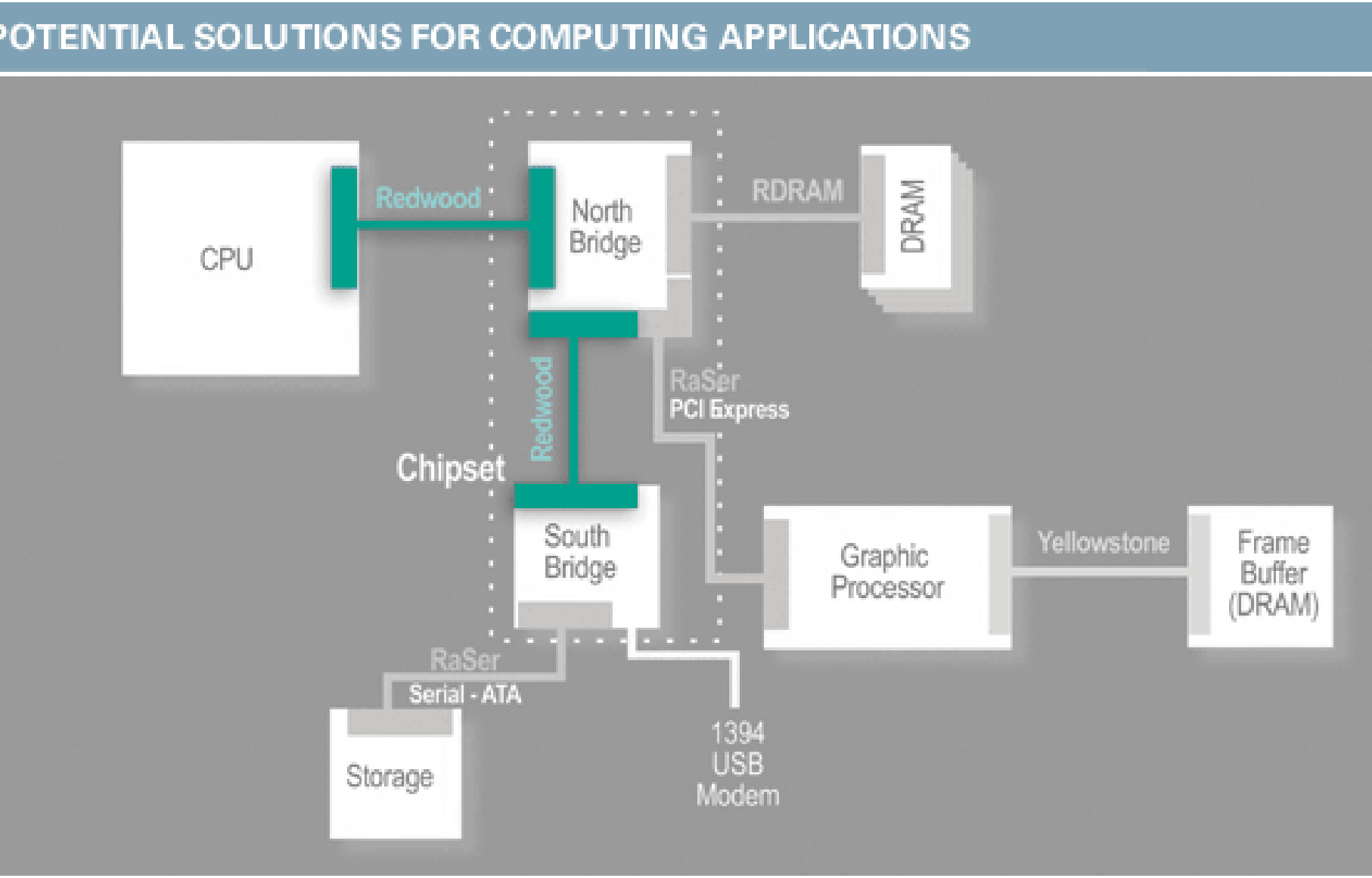

DDR Module Roadmap
The "Mainstream" memory module migration path
128Mb
256Mb
512Mb
1Gb
DDR
3rd Gen
DDR 800-1200
~6-10GB/s
DDR
2nd Gen
PC5400
DDR 667 - 333MHz, 5.4GB/s
PC4300
DDR 533 - 267MHz, 4.3GB/s
DDR
1st Gen
PC3200
DDR 400 - 200MHz, 3.2GB/s
PC2700
DDR 333 - 167MHz, 2.7GB/s
DDR266 - 133MHz , 2.1GB/s
DDR200 – 100MHz, 1.6GB/s
PC2100
PC1600
Production
Samples
Spec complete
'01
'02
'03
'04
'05
'06
'07

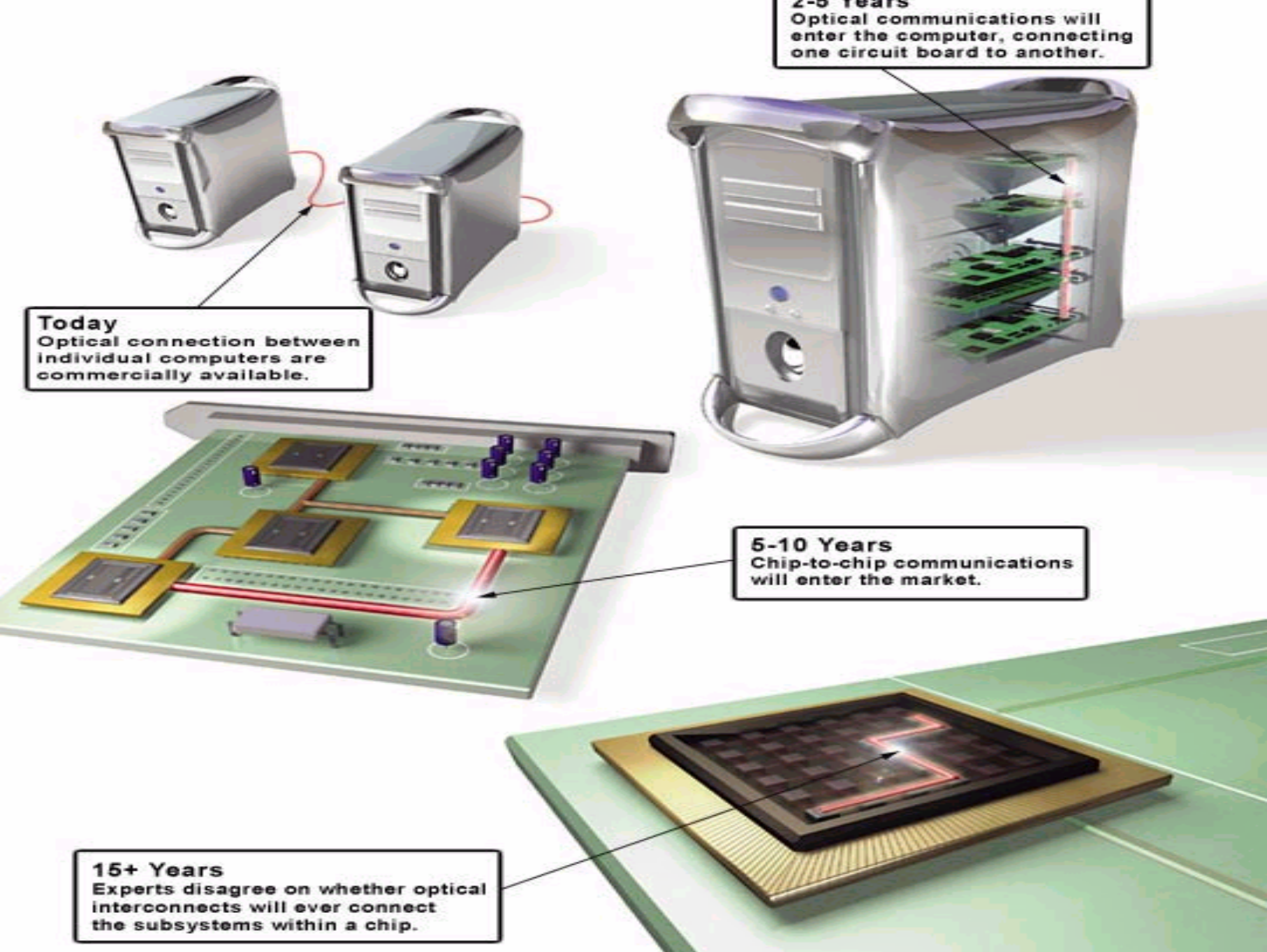
铜线连接的极限？



- FlexPhase: per bit, on-chip alignment of data and clock
- DRSL/LVDS: enabling low-power operation, reduced electro-magnetic radiation (EMR)
- VDR: data rates of 1-10 times the speed of the clock

未来的连接是什么形式？


POTENTIAL SOLUTIONS FOR COMPUTING APPLICATIONS
CPU
Redwood
North Bridge
RDRAM
DRAM
Redwood
RaSer
PCI Express
Chipset
South Bridge
Graphic Processor
Yellowstone
Frame Buffer (DRAM)
RaSer
Serial - ATA
Storage
1394
USB
Modem


2-5 Years
Optical communications will enter the computer, connecting one circuit board to another.
Today
Optical connection between individual computers are commercially available.
5-10 Years
Chip-to-chip communications will enter the market.
15+ Years
Experts disagree on whether optical interconnects will ever connect the subsystems within a chip.

潜在的解决方案——光互连

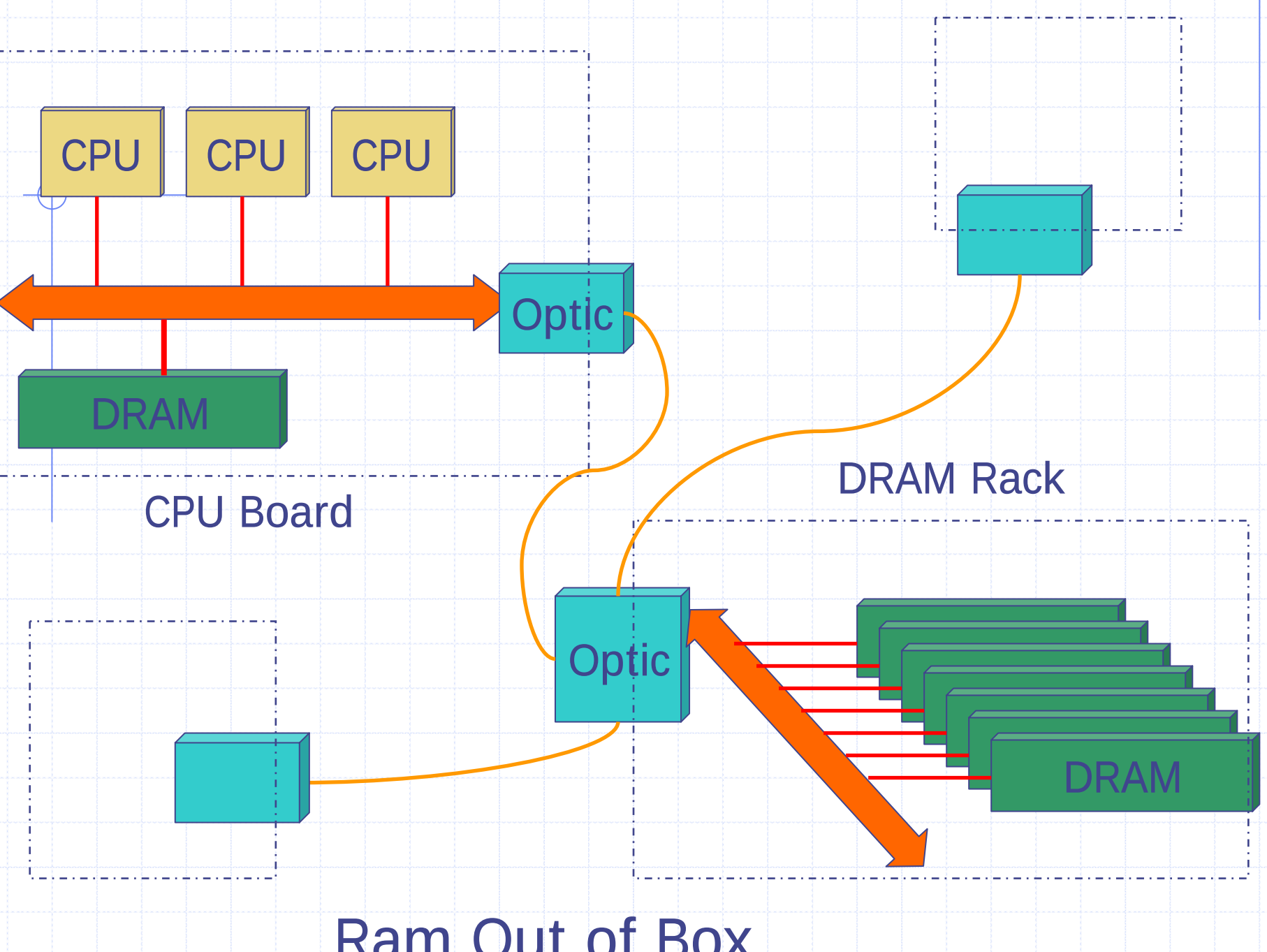
- 带宽：
  - 理论上TB级+，远未达到极限
- 延迟：
  - 光电转换延迟：　<ns
  - 串并转换延迟：　 ns
  - 距离延迟：
    - 0.3米/ns，300米/μs，300公里/ms
  - 接口电路延迟：
  - 交换/路由延迟：

问题：DRAM可以离CPU多远？

- 10ns：3米 相当于 local DRAM
- 100ns~1μs：30-300米 相当于NUMA
- 10μ：3公里 相当于机群消息通信
- 100 μ~1ms: 30-300公里 相当于局域网 tcp/ip
- 10ms: 相当于磁盘随机访问延迟

第一个设想：Ram Out of Box

- 想法：将DRAM组作为可独立安装的部件。(独立机架，独立供电,光连接)
- 与NUMA的不同：
  - CPU只能访问少量本地高速缓存
  - 多CPU共享海量内存
  - 远程内存访问不依赖其他CPU的节点/板



Ram Out of Box

ROB可能带来的优点

- 可扩展性：
  - dram组可动态、单独进行升级和扩容。不受CPU节点自身设计时的限制。
  - 需要时可以以增加DRAM组的方式扩展。
- 稳定性：
  - DRAM的故障不影响CPU节点
  - DRAM的故障更容易进行动态屏蔽
  - 可利用冗余DRAM组进一步提高可靠性
- 智能性：
  - DRAM组可以进行I/O(RDMA)功能
  - DRAM组可以执行一些简单算法，如diff,search等。

ROB的可能带来的优点(2)

- 延迟理论上可以接近NUMA的访问速度
- 如果充分利用光传输潜力，带宽足够

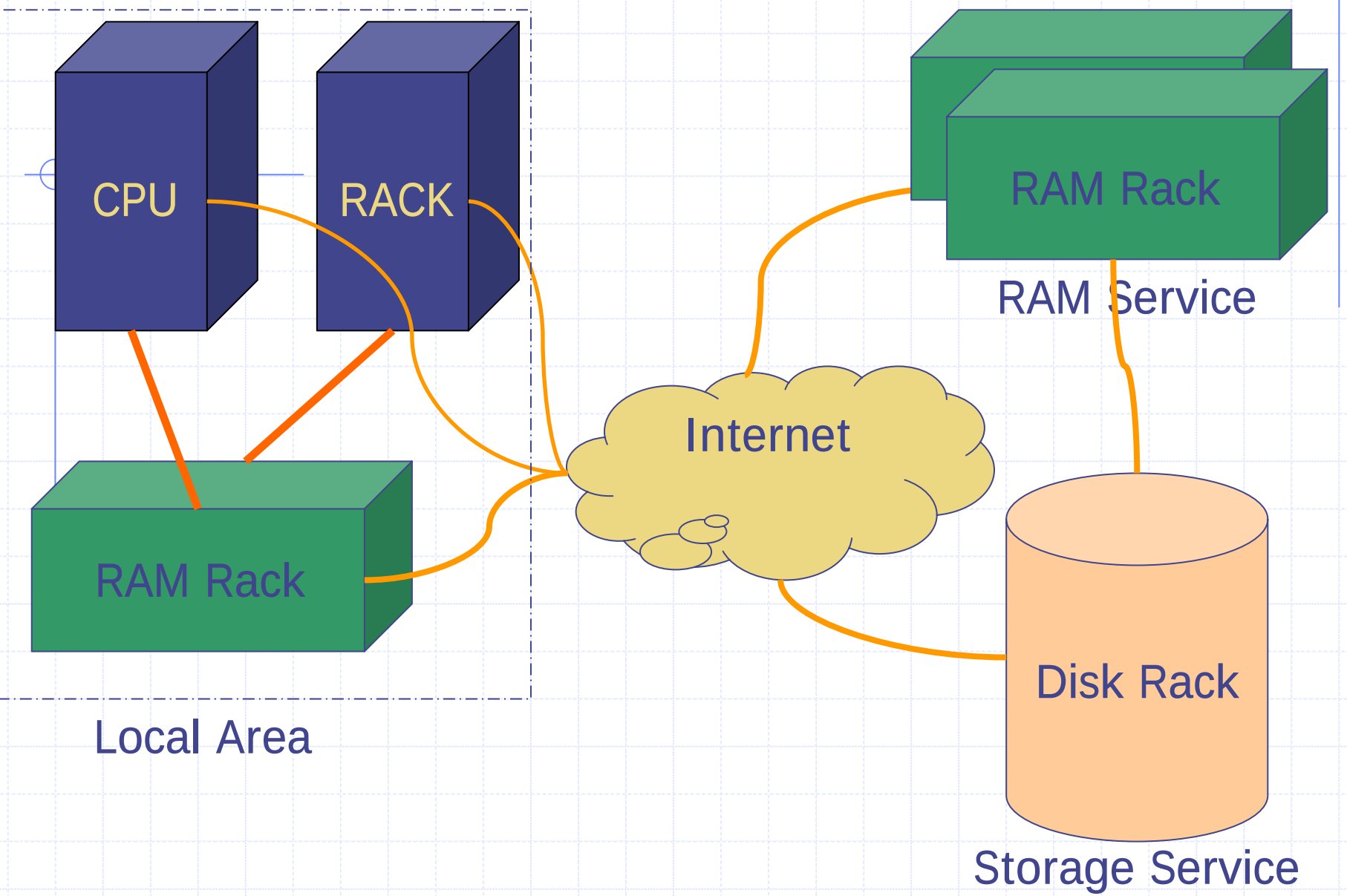
- CPU,RAM进一步模块化，可进行动态屏蔽和热插拔更换。减少故障修复时间。
- CPU RACK和RAM RACK结构趋于单一化，便于简化主板设计以及高密度和冷却系统设计

ROB实现上的主要问题

- 因为取代LocalRAM需要保证一定的速度，基本上必须保证全部硬件实现。
- DRAM组将有很大的聚合带宽，内部交换网络的设计也是一个问题。
- 因为CPU本地缓存受限，必须采用更有效的数据预取、预测算法。现有的编译器和CPU结构是否足够？
- 本地内存减少和远程内存延迟增加对Cache Coherence问题有什么影响？

第二个设想:Ram On Net

- 想法：广域网上使用远程DRAM
- 基础：
  - 采用光的广域网的带宽将可与本地相比
  - 本地Dram(<1 μs)和Disk延迟(10ms)相比具有相当一段的空白
  - 即使延迟相同,大访问量下DRAM随机访问性能也优于Disk。


CPU
RACK
RAM Rack
RAM Rack
RAM Service
Internet
Disk Rack
Local Area
Storage Service

Ram on Net

RON可能带来的好处

- 对需要处理海量数据的计算任务，提供一种比磁盘快（10倍以上？）的海量存储设备。
- 软件性能允许的条件下，真正实现内存使用的On Demand资源共享，提高资源利用率。
- 网络DRAM服务器如果处于良好的管理下，将比独立用户提供更好的可靠性。能更好的保存计算的中间状态。
- 在两处以上的DRAM服务器上冗余保存数据在一定程度上可以取代磁盘稳态存储的地位？

RON实现上的主要问题

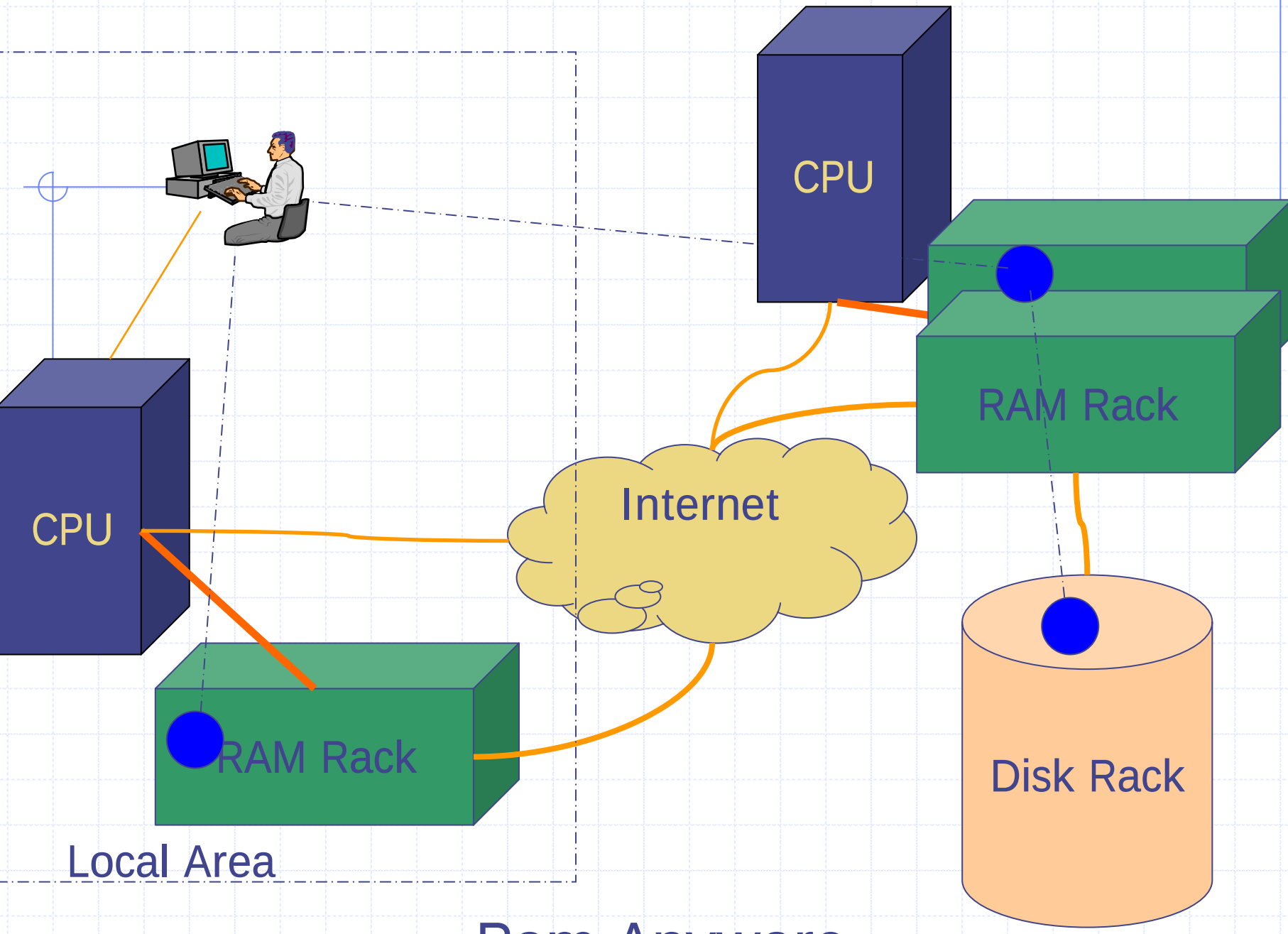
- 因为距离造成的必然延迟，大型任务必须有足够的本地RAM服务
- 广域网中路由器或交换机将带来多大的额外延迟？是否依赖Qos或电路交换？
- RON的延迟允许一定程度上使用软件
  - 是采用透明的MMAP/FileSystem/SWAP机制，还是采用程序参与的主动预取
  - MMAP由本地DRAM服务器实现还是由CPU节点实现？是否需要和何时需要(什么延迟下)额外的硬件

RON实现的主要问题(2)

- 对计算程序编写者，远程RAM的延迟速度是可选择的，需要对性能/延迟/价格/容量进行平衡，如何隐藏这种差异
- Ron为进程检查点或进程状态迁移提供了一种更快的设备。如何有效利用这一优点实现可靠计算

第三个设想：RAM anyware

- 想法：
  - 不只面对HPC用户,可以针对个人用户
  - 对用户来讲，隐藏本地Ram、远程Ram、磁盘的差别。RAM就是存储。
  - 因为RAM服务提供一定可靠性，可以免去shutdown/login概念。用户可以在任何状态下暂停并转到另一个地方继续工作。由服务方隐藏数据迁移、状态存储和恢复等过程。



Ram Anyware

RON对系统软件的要求

- 减少系统自身对RAM的依赖
- 对因距离带来的延迟进行屏蔽和优化调度，充分利用新体系结构的特性。
- 增加对可靠性的支持，保证进行大规模任务时局部部件的故障不影响整体的进度。

精简系统的考虑

- 尽量少使用有限的本地缓存
- 去除低级I/O功能和设备支持,使用RDMA和数据推送
- 简化资源访问接口,是否需要文件系统？是否强化远程调用?
- 专用I/O和系统功能服务器，向MPP的回归？

延迟隐藏的考虑

- Cache Size的选择
- 透明cache策略和主动cache 策略(预取)
- CPU MultiThread机制的利用
- Memory Consistency 语义的研究
- memory map 以及文件系统语义研究

可靠性考虑

- Remote RAM为状态备份提供了一种快速有效的设备，使得备份的速度和频率都可以有效地提高
- 减少local RAM相应减少对检查点的开销
- 现有的检查点、状态迁移、快照等技术如何利用和发展
- 是否可以形成一种新的编程模式？

- 文件系统与数据库接口的统一

RON下的程序设计

- 能否简单的扩展PRAM模型？
- 需要什么样的同步语义
- 需要什么样的资源申请策略
- 如何利用RAM 内建的计算功能

HPC-OG：拆散的哲学

- 与SOC,blade,PIM走完全不同的方向：
  不同功能部件分开，相同功能部件集中，通过综合部件级服务组织计算的体系
- 光连接将是联络其中的纽带