3 The SGI Origin

cache-coherent non-uniform memory access

三个设计目标

1. 更多的核心，更高的性能
2. 保持一致的编程模型
3. 更低的功耗和成本

架构设计

最小节点是双处理器节点

支持最多 512 个节点，即 1024 个处理器

节点内双处理器多路复用总线，而非嗅探

降低绝对的内存延迟

降低远程内存访问的频率

高性能的局部和全局互联设计

满足可靠性和可用性需求

实现

关键 1：真正可扩展的系统

关键 2：很低的内存延迟

关键 3：无意外的带宽瓶颈

互联拓扑

基于 SGI SPIDER router chip

超立方体互联拓扑结构

缓存一致性协议

基于 DASH 协议改进

1）完全支持的 lean-Exclusive 状态

2）完全支持的 upgrade 请求

3）write-invalidate transaction 处理 IO DMA 数据

4）对网络顺序完全不敏感

5）采用更复杂的死锁避免方案

6）更好扩展性的目录方案

7）对页迁移的高效支持

节点设计

两个 MIPS R10000 处理器，194MHz

HIMM：horizontal in-line memory module

Hub Chip：连接处理器、内存和目录

Hub Chip 也分别连接到 router 和 IO 子系统

Hub Chip 包括 XB、II、NI、PI 和 MD

XB：交叉互联，crossbar

II：IO interface

NI：Network Interface

PI：Processor Interface

MD：Memory and Directory Interface

Hub 中所有的协议表都是硬布线实现

IO 系统

核心是 Crossbow（Xbow）ASIC

八个 XIO 端口

每个物理通道有两个虚拟通道

可以为每个 IO 设备单独申请带宽

产品设计

高度模块化设计

基础的模块是桌面级的：4 个节点、两个 router、12 个 XIO、1 个 CDROM 和最多 5 个 SCSI 设备

性能特点

两个实现性能和扩展的关键之处

1）fetch-and-op 原语是不缓存直接与内存交互的操作

2）软硬件对页迁移的支持

Origin 性能

使用 Microbenchmarks 测试延迟和带宽

使用 NAS 并行基准程序和 SPLASH2 套件测试并行应用的性能

相关的系统

Stanford DASH 协议

DASH 协议基于 SMP 节点

SMP 节点的主要优势：节点内的缓存共享

SMP 节点的劣势

劣势 1：局部内存延迟和成本较高

劣势 2：远程访问需要等到局部访问完成后进行

劣势 3：远程内存带宽将是局部内存带宽的一半

Sequent NUMAQ 和 DG NUMALiiNE

包含 63 个节点，每个节点是 4SMP

局部内存延迟很好，但最好情况的远程内存延迟也是局部延迟的 8 倍左右

Convex Exemplar X

节点的交叉互联，8 到 16 个处理器

并行环状互联实现 SCI 协议

分析：

节点内交叉互联的确降低了 SMP 带来的带宽损失

但对比 Origin 更小更集成的节点延迟还是更高

Overall Comparison of DSM Systems

Origin 最大的不同是有一个更加紧密集成的 DSM 结构

总结

Origin 2000 是一个高度可扩展的服务器设计

高度模块化的系统，入门成本和扩展成本低

超立方体互联提供了高带宽和低延迟

局部内存延迟低，远程内存访问少

软硬件支持的页迁移和快速的同步原语