The SGI Origin: A ccNUMA Highly Scalable Server

对于多处理器系统，比较流行的有 3 种模式，对称多处理器模式（SMP）,非均存存储访模式（NUMA）和规模并行处理模式.SMP 模式即两个或两个以上的同样的处理器连接到一个共享的主存上。在 SMP 系统中，所有的处理器可以同时访问一个物理存储器，即同时运行一个操作系统，因此也被成为均匀性存储访问系统，这种结构比较简单，但是由于是共享存储器，容易在访问时产生系统瓶颈，可扩展性也比较差。MPP 是分布式存储器模式，可扩展性好，但是需要并行编程和并行编译，在软件系统构建上比较复杂，使用不方便。NUMA 架构将若干单元通过专门的互联设备联结在一起组成分布式和共享内存空间。每一个处理器可以访问自己的存储器，也可以访问其他处理器或者共享存储器，所有的访存有远、近之分，延迟有长有短，称为非均匀存储访问。在某个处理器访问空间上比较远的处理器时，会有很大的延迟，为了缓解这个问题，通过高速缓存一致性使得处理器访问存储器的几率大大降低，在某种程度上提高了系统效率，这种架构称为 CC-NUMA。这种架构继承了 SMP 和 MPP 系统的一些优点，在处理器个数，内存大小、I/O 能力和宽带上有很大的伸缩性，又保持了 SMP 单一系统，简单编程和易于管理的优点。

CC-NUMA 架构应用于 SGI 公司的 ORIGIN 系列，SGI 公司很好的发展和扩展了 CC-NUMA 技术，其基本架构被广泛应用。ORIGIN2000 的基本原理图，每一个节点拥有 2 个处理器，2 个二级缓存，主存，用于互联的 HUB 芯片，1 个 I/O 接口，1 个互 联网络的路由器接口，它的每个节点可以看作是一个 SMP，通过互联网络可扩展至128 个处理器的多处理器系统。Origin 2000 的所有结点 通过 CrayLink 高性能互联网络相互联接，路由器是构成 CrayLink 的基本单位，它包含 6 个端口，内部采用交叉开关实现端口间的全互联。 每个路由器的 2 个端口用于联接结点，其余 4 个端口实现路由器间的互联，形成互联网络拓扑结构。该 CrayLink 的半分带宽与结点个数成线性递增关系，对任意 2 个结点，至少能提供两条路径，保证了结点间的高带宽、低延迟联接和互联网络的稳定性和容错能力。

Origin 系统的基本构件是双处理器节点。除了处理器之外，节点还包含 4 GB 的主内存及其 对应的目录内存

Origin 2000 系统采用的互连是基于 SGI SPIDER 路由器芯片。该芯片的主要特点是:

1. 每个路由器有六对单向链路
2. 路由器低延迟
3. AMQ 缓冲结构以最大化利用负载
4. 每个物理通道有四个虚拟通道
5. 拥塞控制允许消息在两个虚拟通道之间自适应切换
6. 支持 256 级的消息优先级，通过分组老化提高优先级
7. 通过 go-back-n 滑动窗口协议，对每个包进行 CRC 检查，并重新传输错误信息
8. 软件可编程路由表

O 子系统的核心是 Crossbow(Xbow) ASIC，它与 SPIDER 路由器有许多相似之处。Xbow 的一些主要特点是：

1. 8 个 XIO 端口，从原点连接到 2 个节点和 6 个 XIO 卡
2. 每个物理通道有两个虚拟通道
3. 路由器低延迟
4. 支持从特定设备分配的消息带宽
5. 通过 go-back-n 滑动窗口协议，对每个包进行 CRC 检查，并重新传输错误信息

origin 系统性能

使用微基准测试来测量延迟和带宽,并通过使用 NAS 并行基准 V2.2 和 SPLASH2 套件来测量一组并行应用程序的性能。

内存的延迟测量

Memory level	Latency (ns)
L1 cache	5.1
L2 cache	56.4
local memory	310
4P remote memory	540
8P avg. remote memory	707
16P avg. remote memory	726
32P avg. remote memory	773
64P avg. remote memory	867
128P avg. remote memory	945

共享内存带宽测量

递增操作全局共享寄存器