The SGI Origin: A ccNUMA Highly Scalable Server

摘要

SGI Origin 2000 是由 Silicon Graphics 公司设计和制造的一种 cache- coherence 非均匀内存访问(ccNUMA)多处理器。Origin 系统从头开始设计为一种多处理器，能够扩展到小型和大型处理器数量，而不需要任何带宽、延迟或成本。Origin 系统由 512 个节点组成，这些节点通过可伸缩的 Craylink 网络相连。每个节点由一个或两个 R10000 处理器、高达 4gb 的相干内存以及到 XIO IO 子系统的一部分的连接组成。本文讨论了创建 Origin 2000 的动机，并描述了其体系结构和实现。此外，还介绍了 NAS 并行基准 V2.2 和 SPLASH2 应用程序的性能结果。最后，将 Origin 系统与其他现代商业 ccNUMA 系统进行了比较。

1. 背景

这一段主要介绍了 Origin 系统产生的背景，以及本文的布局。

2. The Origin SMP Architecture

Origin 体系结构的框图如图 1 所示。

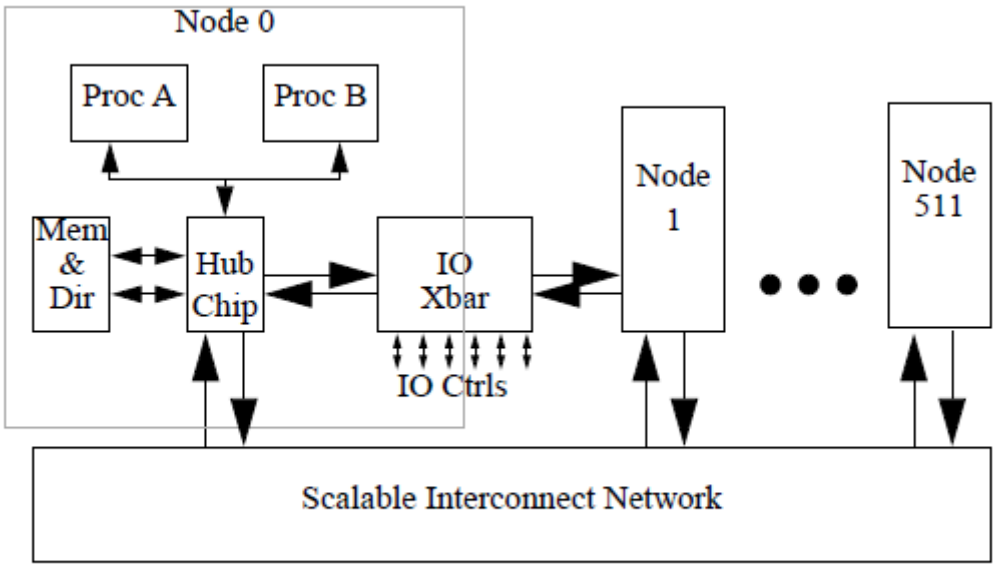

Figure 1 Origin block diagram

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，并且与 IO 子系统的一部分有连接。该体系结构支持最多 512 个节点，最多配置 1024 个处理器和 1 TB 的主内存。这些节点可以通过任何可扩展的互连网络连接在一起。Origin 系统使用的缓存一致性协议不要求点对点消息的有序传递，从而在实现互连网络时允许最大限度的灵活性。

3. Origin 系统的实现

3.1 网络拓扑

Origin 2000 系统采用的互连是基于 SGI SPIDER 路由器芯片。该芯片的框图如图 2 所示。该爬行器芯片的主要特点是:

- 每个路由器有六对单向链路
- 路由器低延迟
- DAMQ 缓冲结构以最大化利用负载
- 每个物理通道有四个虚拟通道
- 拥塞控制允许消息在两个虚拟通道之间自适应切换
- 支持 256 级的消息优先级，通过分组老化提高优先级
- 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
- 软件可编程路由表

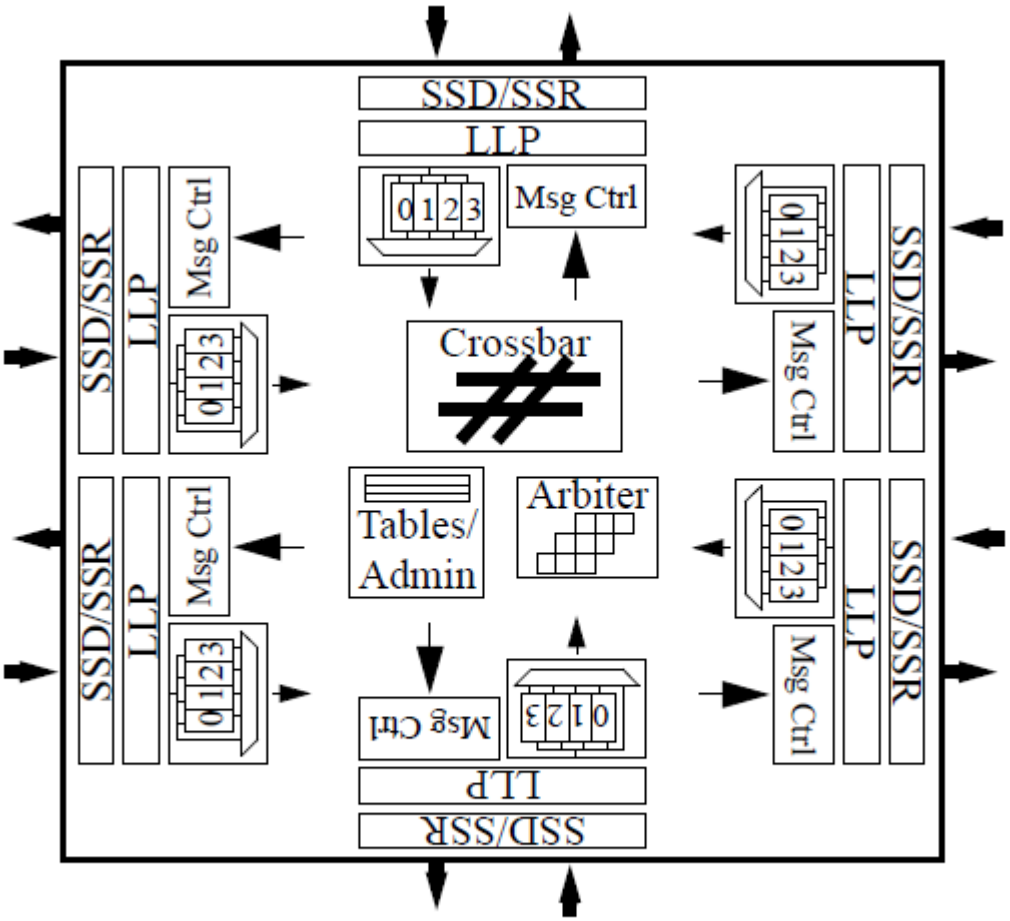

Figure 2 SPIDER ASIC block diagram

3.2 Cache 一致性协议

与 DASH 协议相比，Origin coherence 协议有一些改进。

首先，Clean-exclusive (CEX)处理器缓存状态(在 MESI 中也称为独占状态)完全由 Origin 协议支持。

超越 DASH 的 Origin 协议的第二个增强是完全支持升级请求，该请求将一行从共享状态移动到独占状态，而不需要传输内存数据的带宽和延迟开销。

Origin 协议对网络排序完全不敏感。消息允许在网络中相互绕过，协议检测并解决所有这些无序的消息传递。这使得 Origin 能够在其网络中使用自适应路由来处理网络拥塞。

3.3 节点设计

原点节点的设计适合于单个 16x11 印刷电路板。原点节点板图如图 5 所示。

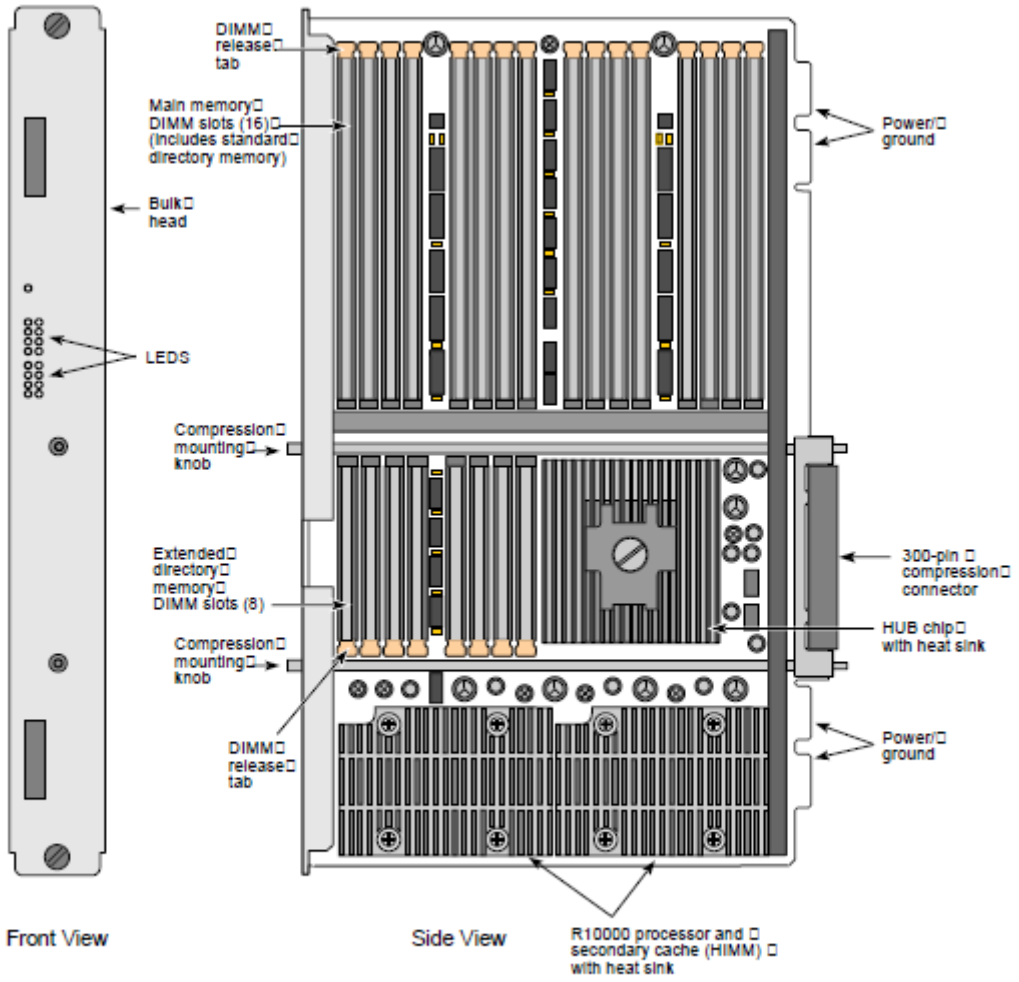

Figure 5 An Origin node board

图 6 显示了集线器芯片的框图。集线器芯片分为五个主要部分:横杆(XB)， IO 接口(II)，网络接口(NI)，处理器接口(PI)，内存和目录接口(MD)。所有接口通过连接到横梁的 FIFOs 相互通信。

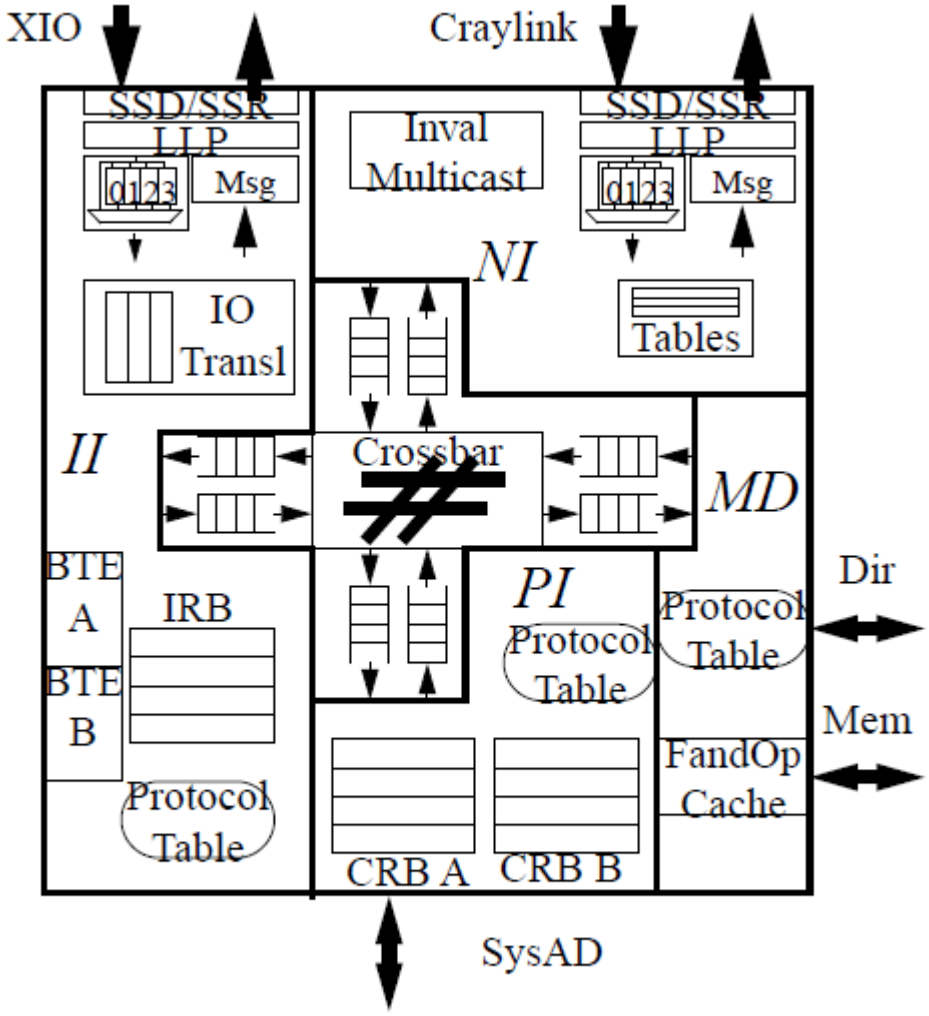

Figure 6 Hub ASIC block diagram

3.4 IO 子系统

IO 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：

- 8 个 XIO 端口，从原点连接到 2 个节点和 6 个 XIO 卡
- 每个物理通道有两个虚拟通道
- 路由器低延迟
- 支持从特定设备分配的消息带宽
- 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

4. Origin 性能

本节使用微基准测试来测量延迟和带宽，以及使用 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，来检查 Origin 系统的性能。

4.1 微处理器基准测试程序

第一个微基准测试检查原始内存系统的延迟和带宽。表 4 显示了单独测量的内存引用的延迟。

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

Table 4 Origin 2000 latencies

在 Origin 2000 上，每个处理器可以有效地利用节点上可用的内存带宽的一半以上。因此，我们在图 11 中加入了流结果，每个节点只运行一个线程，图 12 中每个节点只运行两个线程。

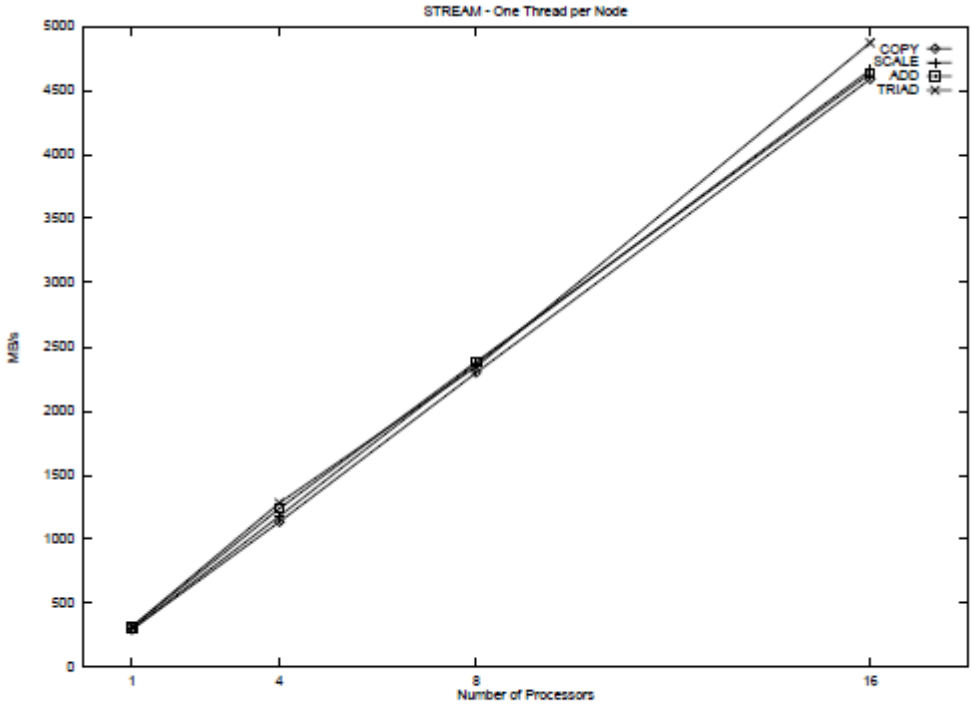

Figure 11 STREAM results - one thread per node

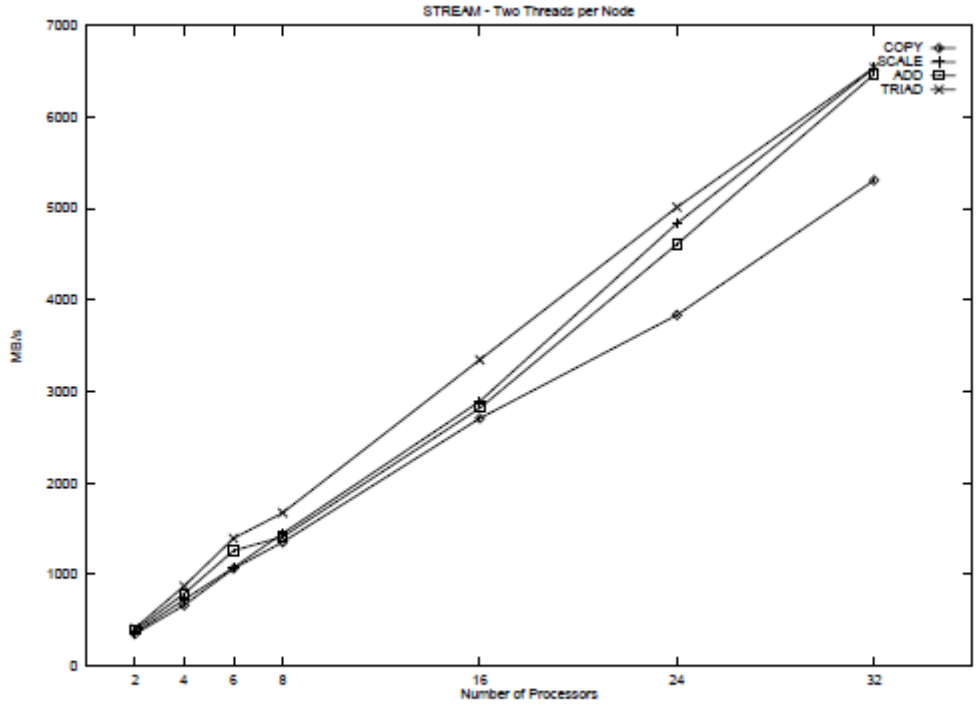

Figure 12 STREAM results - two threads per node

表 5 显示了获取和递增操作在实现全局共享计数器方面的有效性。

M op/s	1 P	2 P	4 P	8 P	16 P	32 P
fch-inc	4.0	7.4	6.1	10.0	19.3	23.0
LL/SC	6.9	2.3	0.84	0.23	0.12	0.09

Table 5 Comparison of LL/SC and fetch-and-op for atomic increments

4.2 应用

图 14 显示了使用最多 32 个处理器的类 A 数据集的 NAS 并行基准测试的性能。总的来说，NAS 基准上的加速非常好。对于几个基准测试，由于随着处理器数量(以及节点数量)的增加，总缓存大小和可用的内存带宽都比较大，因此实现了超线性加速。

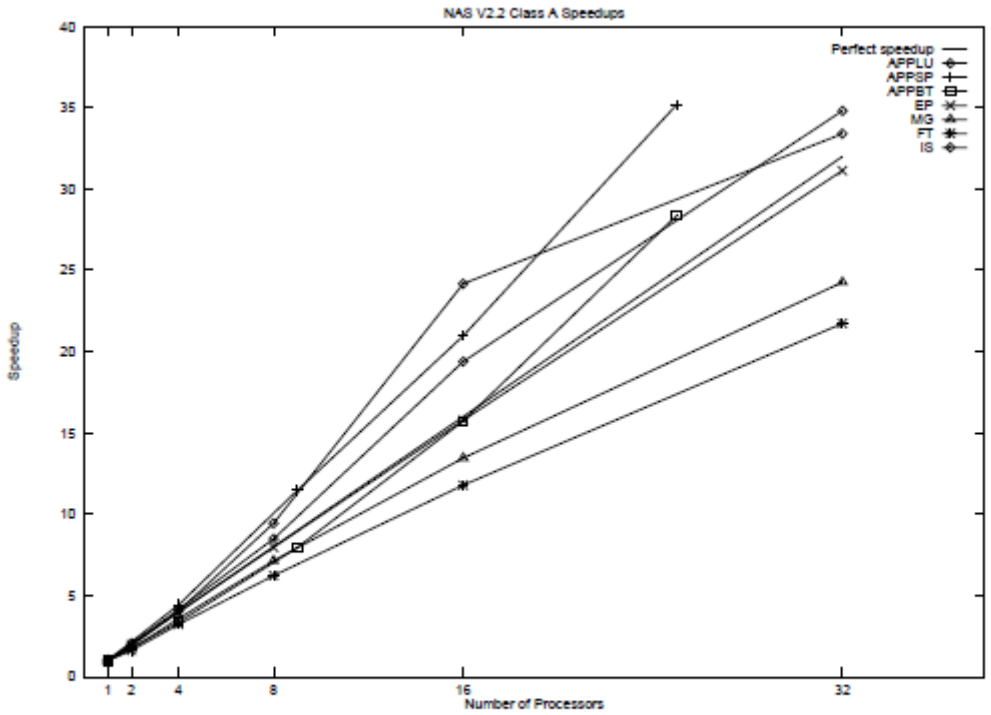

Figure 14 NAS Parallel V2.2 Speedups

表 6 列出了从 SPLASH2 套件运行的应用程序及其命令行参数。

Application	Command Line
radiosity	-batch -room
raytrace	balls4.env
lu	-n2048 -b16
ocean	-n 1026
barnes	< input.512

Table 6 SPLASH2 Applications

图 15 显示了 Origin 系统上的四个 SPLASH2 应用程序和一个 SPLASH2 内核的加速速度。

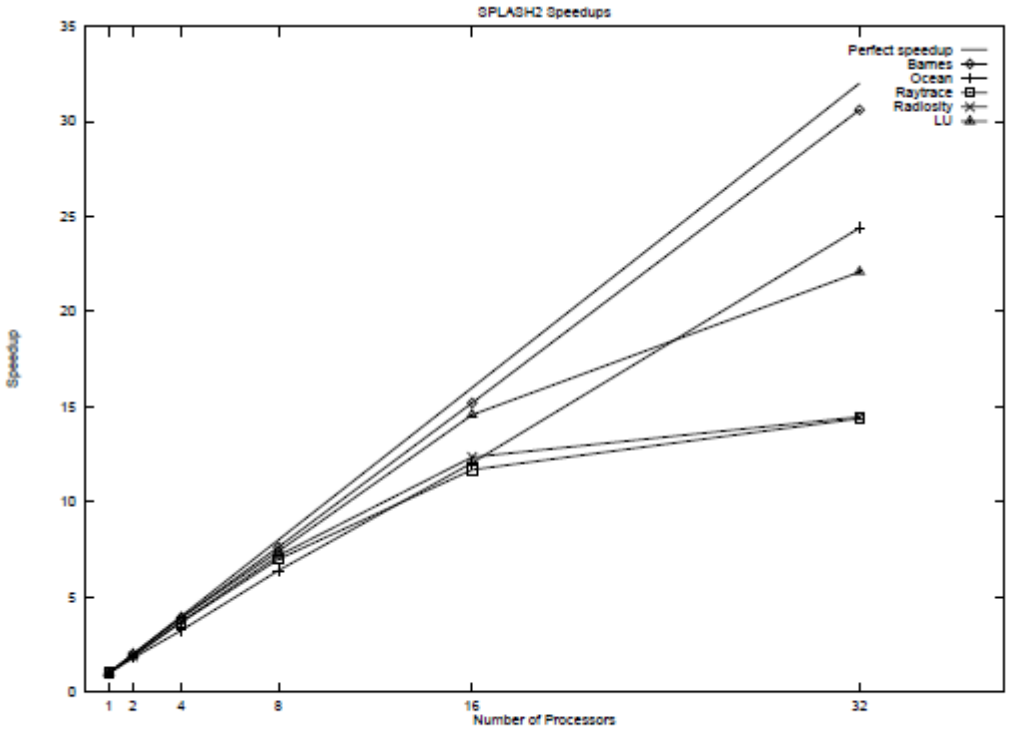

Figure 15 SPLASH2 Speedups

5. 总结

Origin 2000 是一个高度可扩展的服务器，设计来满足技术和商业市场的需求。这是通过提

供一个高度模块化的系统，低起点和增量成本来实现的。本地内存的低延迟和远程到本地内存的低延迟比率允许现有的应用程序库轻松地将其应用程序从现有SMP挑战和Power挑战系统的统一访问迁移到 NUMA 源系统。Origin 还提供了一些特性来帮助这些应用程序的性能，包括对页面迁移和快速同步的硬件和软件支持。