计算技术研究所，微处理器中心

论文阅读报告　其3
The SGI Origin: A ccNUMA Highly Scalable Server

谭弘泽
201828013229048

October 18, 2018

Contents

1 摘要

SGI Origin 2000 是由Silicon Graphics公司设计生产的一个高速缓存一致的非均一内存访问（ccNUMA，cache-coherent non-uniform memory access）多处理器系统。Origin系统是作为一个既能够适应较少的处理器数量又能够适应较多的处理器数量，不会有任何带宽、延迟、功耗的悬崖，而从头开始设计的多处理器系统。Origin最多由512个可伸缩Craylink网络结点组成。每个结点由1到2个R10000处理器，最大4GB的一致性内存，和一个到XIO IO子系统的一部分的连接组成。这篇论文讨论构建Origin 2000的动机，然后描述了它的架构和实现。附加地，通过NAS Parallel Benchmarks V2.2和SPLASH2应用展示了性能结果。最后，Origin系统与同时代的其他商用ccNUMA系统进行比较。

2 背景

Silicon Graphics公司提供个很多代基于MIPS处理器的对称多处理器系统。从基于R3000的8处理器Power系列，到36处理器的基于R4000的Challenge和基于R10000的Power Challenge系统，这些对称多处理器系统的高速缓存一致的、全局可访问的内存架构为大型并行应用提供了一种便于编程的环境，同时还提供了为不论是并行还是基于吞吐的工作负载提供高效的执行力。
Power Challenge的下一代系统需要符合三个需求:

1. 需要能够扩大到超过Power Challenge系统的36处理器极限, 并且提供一个能够支持每个处理器有更高性能的基础设施。

   给定从Power系统到Power Challenge系统的处理器增长倍率4，希望下一代的系统最大能支持的处理器数至少翻到4倍。

2. 能够保持Power Challenge的高速缓存一致的全局地址内存模型。这个模型对实现高性能循环级并行代码以及支持已经有了Power Challenge的用户十分关键。

3. 初始和增加的系统功耗要比高性能对称多处理器系统低，使得功耗理想地接近一个工作站集团。

简单地造一个更大更快的基于总线监听的对称多处理器系统无法满足上面这三个要求。第二个或许还能达成，但是就需要权衡性能、处理器数量和功耗了。
Origin使用了分布式共享内存（DSM，distributed shared memory），通过基于目录的协议保持了Cache一致性。分布式共享内存有满足这三个要求的潜力：可伸缩性、易于编程、功耗。基于目录的一致性消除了限制基于总线监听的一致性的可伸缩性的广播瓶颈。全局可编址的内存模型被保留了下来，即使内存访问所花费的时间不再是均匀的了。在接下来的论文章节中，将会展示Origin的可伸缩共享内存多处理器(S^2MP,scalable shared-memory multiprocessing)架构。第3节详细描述Origin 2000的实现。第4节展示Origin 2000的性能。第5节将Origin系统与其他同时代的ccNUMA系统进行对比。最后，第6节总结这篇论文。

3 Origin S^2MP架构

Origin架构的框图如下：

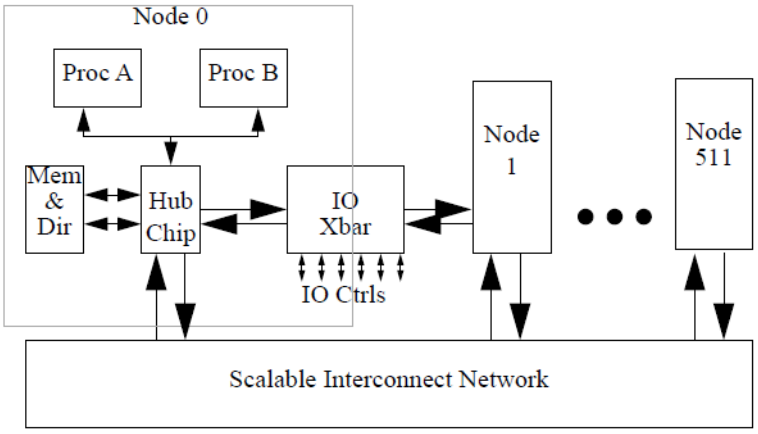

Figure 1 Origin block diagram

Origin系统的基本建造块是双处理器结点。附加在处理器上，一个节点包含最大4GB的主存以及相应的目录内存，还有到IO子系统的一部分的连接。
其中，目录内存是为了维护分布式共享内存的一致性协议而划分出来的内存。
这个架构最大支持512个节点，1024个处理器，以及1TB的主存。
Orgin系统使用的高速缓存一致性协议不需要点对点消息的按序传达，允许互联网络实现的最大灵活性。
还有许多诸如一致性协议宏带有干净独占的状态，DRAM带有奇偶校验码等特性。来提高对性能、功耗、可靠性可用性等都有所考量。

4 Origin的实现

分布式共享内存系统已经在学术组织中有所使用了，但是成功地商业化还需要让系统真正地可伸缩，低访存延迟并且没有意想不到的带宽瓶颈。
作者从系统的全局互联开始介绍。

4.1 网络拓扑

在Origin 2000系统中使用的互联基于SGI SPIDER路由芯片。下面是这个芯片的框图。
主要特征有：

1. 每个路由器有6对无向连接
2. 虫洞路由的低延迟（41ns pin-to-pin）
3. 全局仲裁的DAMQ（dynamically allocated multi-queue）缓冲结构来最大化负载下的效用。

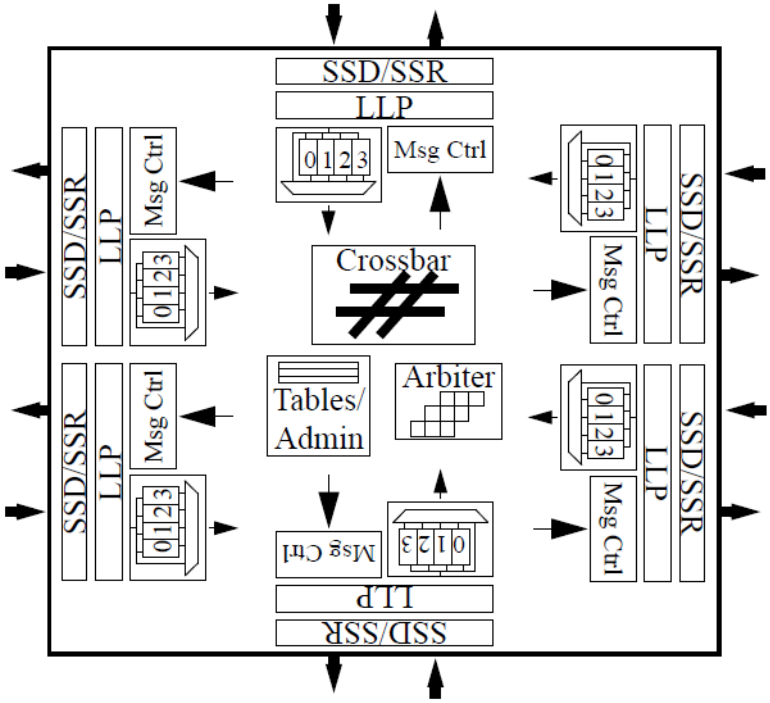

Figure 2 SPIDER ASIC block diagram

4. 每个物理通道有四个虚拟通道
5. 允许两个虚拟通道间自适应地交换消息的拥塞控制
6. 支持256级消息优先级并通过包的时间提升优先级
7. 对每个包的CRC校验并且通过回退n滑动窗口协议错误重传
8. 软件可编程路由表

Origin 2000使用SPIDER路由器创建多刺宽大超立方（bristled fat hypercube）互联拓扑。32个处理器和64个处理器的连接方式如下图：

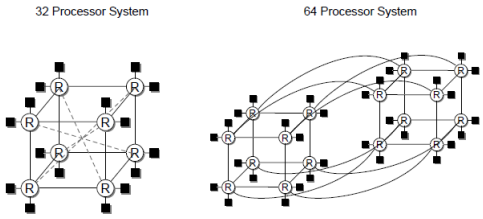

Figure 3 32P and 64P Bristled Hypercubes

128个处理器的连接方式如下图：

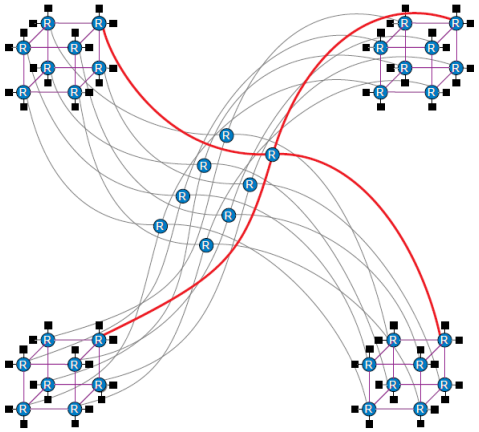

Figure 4 128P Heirarchical Fat Bristled Hypercube

每个32处理器的立方体作为一组，每组对应位置的处理器通过元路由器（meta-router）相连。
要扩大到1024个处理器，每个128处理器系统的元路由器都要被替换乘一个5维超立方。其中，5维超立方有32个路由器节点。这32个节点每个上面有一个接口分别连到32组32处理器的立方体每组中的某一个上面，刚刚好。

4.2 高速缓存一致性协议

Origin使用的高速缓存一致性协议和Stanford DASH协议相似，但是有几处显著的性能提升。
向DASH协议一样，Origin高速缓存一致性协议是非阻塞的。
作者介绍了包括引入新的状态等的几个优化，并且详细描述了处理器读写请求的流程。

4.3 结点设计

Origin结点的设计放进了一块单独的“16 × 11”印刷电路板中。草图如下：

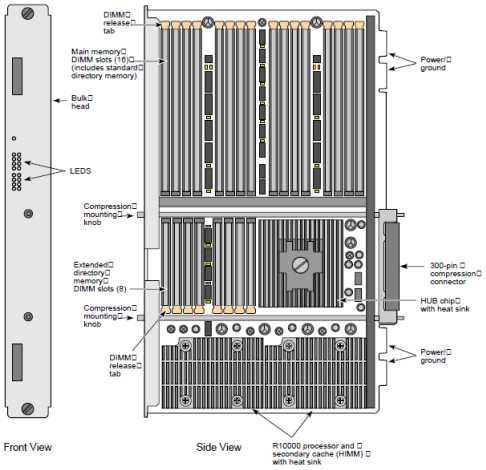

Figure 5 An Origin node board

下图是集线器芯片的框图，分成了五大块：

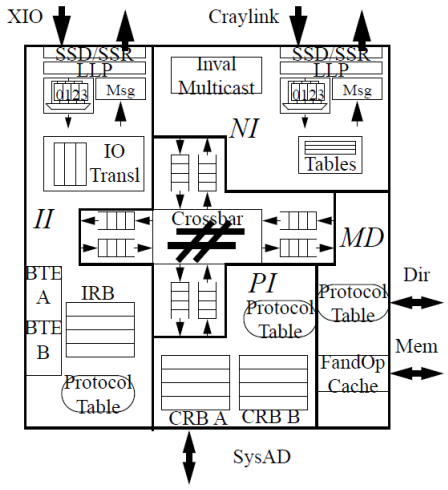

Figure 6 Hub ASIC block diagram

集线器芯片带宽的原始数据列在了下表中：

Port	SysAD	Mem	XIO	Craylink
GB/s	0.78	0.78	1.56	1.56

Table 1 Hub ASIC port bandwidths

各个单元的尺寸列在了下表中：

Section	XB	IO	NI	PI	MD
K gates	246	296	56	133	77

Table 2 Hub ASIC gate count

这个芯片用得很多，节约尺寸是有必要的。

4.4 IO子系统

下图展示了连接两个结点的IO卡的一种可能的配置：

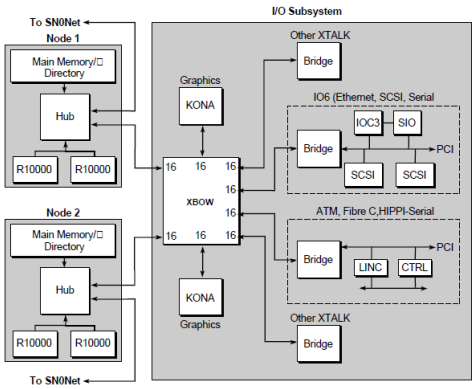

Figure 7 Example IO subsystem block diagram

一排结点可以通过节点上的两个接口以及I/O子系统上的两个接口线性串联。
下表列出了共同的XIO卡：

Board	Number of Ports
Base IO	2 Ultra SCSI, 1 Fast Enet, 2 serial
Ultra SCSI	4
10/100 Enet	4
HiPPI	1 serial
Fibre Channel	2 Cu loops
ATM OC3	4
Infinite Reality Gfx	1
Standard PCI Cage	3
VME Adapter	1

Table 3 Origin IO boards

4.5 产品设计

Origin 2000有高度模块画的设计。基本建造块是含有4个节点板，2个路由板和12个XIO版的插槽的台式机模块。模块还包含光盘（CDROM，Compact Disc Read-Only Memory）和最多5个Ultra SCSI（Small Computer System Interface）设备。
下面是台式机模块的框图：

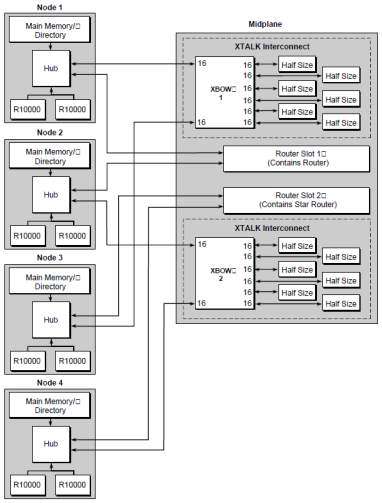

Figure 8 Deskside module block diagram

下面是台式机模块的后视图：

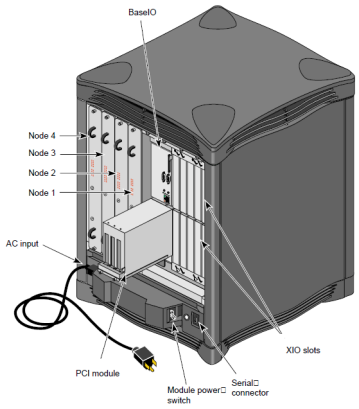

Figure 9 Deskside module, rear view

然后要连接16处理器系统的时候可以像下图这样:

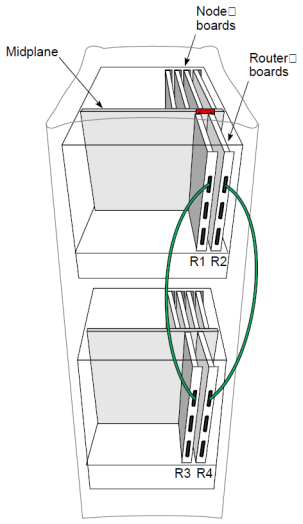

Figure 10 16 processor Origin system.

4.6 性能特征

Origin系统有两个非常重要的特性。
第一个是fetch-and-op原子操作延迟很小。
第二个是Origin提供支持页迁移的软件和硬件。
下表给出了Origin系统各个存储层次的延迟:

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

5 Origin的性能

这节使用微基准测试检验Origin系统的延迟和带宽，包括NAS Parallel Benchmark V2.2和SPLASH2组合。

5.1 Microbenchmarks

文章给出了STREAM测试在每结点单线程和每结点2线程的情况下，STREAM测试的结果:

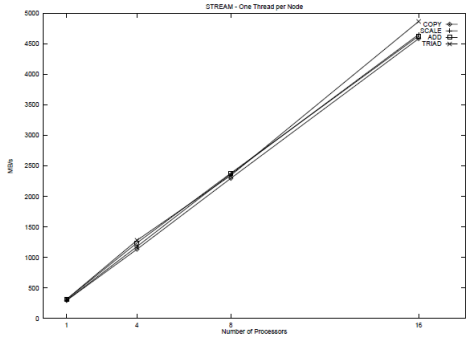

Figure 11 STREAM results - one thread per node

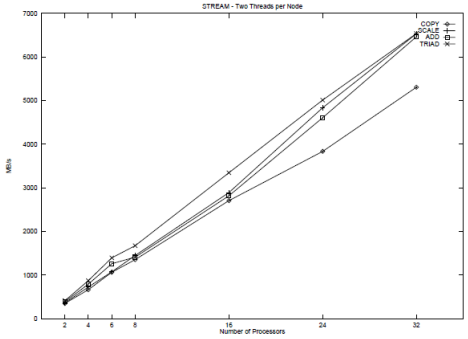

Figure 12 STREAM results - two threads per node

除了COPY在每结点两线程的情况下表现不佳以外，整体还可以。

下表对比了fetch-and-op和load-link/store-condition两种原子操作。LL/SC随着线程数增加，成功率逐渐趋近于0，以至于随证线程数增加，性能逐渐趋于0。而fch-inc的成功次数却能随着处理器个数的增长而增长。

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

下图继续对比fetch-and-op和load-link/store-condition两种原子操作：

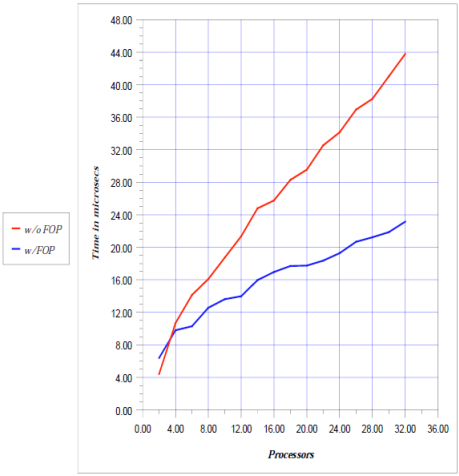

Figure 13 Comparison of LL/SC and fetch-and-op for a null doacross loop

LL/SC虽然能够保证操作原子性，但是如果多核同时操作时时候进行放弃的实现方式，那么随着核数的增大效率可能极为低下。

5.2 应用

使用NAS Parallel Benchmarks V2.2 Class A和SPLASH 2套装来代表应用性能。
下图是使用NAS Parallel Benchmarks V2.2 Class A测出的并行加速。

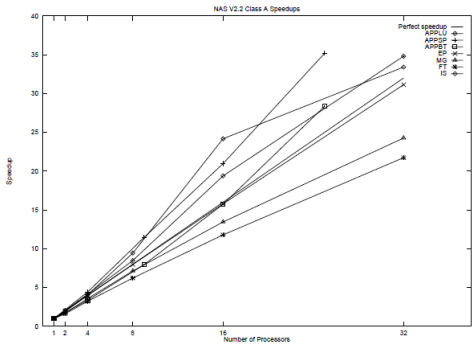

Figure 14 NAS Parallel V2.2 Speedups

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

Table 6 SPLASH2 Applications

文章按照以上参数测试了SPLASH 2的5个应用，测试结果如下：

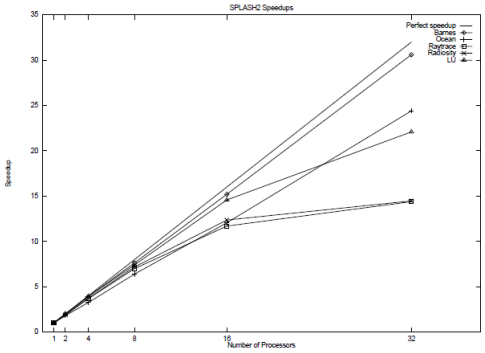

Figure 15 SPLASH2 Speedups

有一些数据型的应用到一个机箱的16个节点以后就很难通过增加结点数目来提升性能了，但是也有一些增加到32个节点能提升性能的应用。

6 相关系统

6.1 Stanford DASH

高速缓存一致性协议上，Origin和Stanford DASH类似但是做了一些改良。架构上，区别在于Stanford DASH每个结点是4处理器的。

6.2 Sequent NUMAQ 和DG NUMAiiNE

Sequent NUMAQ最多有63个结点，每个结点4处理器。
DG NUMAiiNE和Sequent NUMAQ比较类似。

6.3 Convex Exemplar X

一开始的配置支持64处理器，但是架构可以被放大到512处理器。

6.4 DSM系统的总体比较

Origin和这些最主要的不同在于Origin使用了把本地访问当作总的DSM内存访问的一种优化的更加紧凑的DSM结构。
那个模型更适合使用取决于几个因素，比如对系统可伸缩性的需求，工作负载是不是吞度量导向的，对全局可寻址的需求。

7 总结

Origin 2000是一个为同时符合技术和商业市场需求而设计的高度可伸缩的服务器。这些通过提供提供一个高度模块化的低进入点和功耗增长率的系统来实现。一种叫做bristled fat hypercube network(多刺宽大超立方网络)可以被用来提供高对分带宽和低延迟互联。到本地内存的低延迟和低的远程对本地延迟比允许已有应用方去简单地从已有的均匀访问的对称多处理器Challenge 和Power Challenge系统移植到NUMA Origin系统。Origin也有几个能够助于提高这些应用性能的几个特性，包括硬件和软件对页迁移和高速同步的支持。

8 附录A-单词翻译对照表

英文	中文
cliff	峭壁，悬崖
from the ground up	从头开始
consist	由...组成
portion	一部分（强调这一部分作为个体）
contemporary	同时代的
infrastructure	基础设施
hub	集线器
retain	保持
modular	模块化
bristled	多刺毛的
rear-view	后视
deskside	台式机