第 21 卷第 1 期
1998 年 3 月

南 京 气 象 学 院 学 报
Journal of Nanjing Institute of Meteorology

Vol. 21 No. 1
Mar. 1998

S^2MP 结构分析

陈杰伦[1)]　　朱　江[2)]

(1) 南京气象学院计算机与信息工程系, 南京　210044, 2) 同创集团系统部, 南京　210018)

摘　要　S^2MP 是SGI 刚刚推出的新| 代体系结构, 既具有 MPP 的可扩展性, 又具有 SMP 的可编程性, 将两者的优势集于| 身。将S^2MP 与历史上曾经出现的两种多机结构进行比较, 指出了 S^2MP 结构的特点。分析结果认为, S^2MP 是当代 MIMD 机器最具代表性的结构。

关键词　S^2MP, 可扩展性, 可编程性

分类号　TP303

1996 年 11 月, 在兼并了 Cray 公司短短数月之后, SGI 公布了它的新一代体系结构 S^2MP, 并推出了一种新型服务器——基于 S^2MP 的高性能 Origin 系列。

从多处理机结构的发展历史来看, 有两种主要的体系结构, 那就是共享存储型多处理机结构和消息传递型多处理机结构。典型的共享存储型多机结构被称为多处理机结构。典型的消息传递型多机结构被称为多计算机结构。从物理上看, 多处理机结构属于集中式存储, 多计算机结构属于分布式存储; 从逻辑上看, 多处理机结构共享存储器, 多计算机结构中各个处理机节点上的存储器是局部的。与多处理机结构(共享) 和多计算机结构(局部) 的两种不同存储机制相对应, 两种结构各自操作系统进程间的通信也有两种不同的机制: 共享变量和消息传递。共享存储型多处理机结构采用共享变量的通信机制, 它的程序设计以传统的高级语言为基础, 系统提供自动并行识别或增加并行语言成分。而消息传递型多机结构采用的是消息传递通信机制, 必须另一种编程环境(如 PVM、MPI 等) , 在程序中显式地写出消息的发送和接收。

可扩展性和可编程性是并行计算机有别于其他计算机的最重要的两个特性。一般认为, 共享存储型多机结构和消息传递型多机结构各侧重于一个方面: 共享存储型多机结构由于其存储器集中存放, 可扩展性不是很好, 但其采用了共享变量的通信机制, 可编程性良好; 而消息传递型多机结构正好相反, 由于其存储器分布在各个处理机节点上, 它所带来的可扩展性很好, 但它采用的是消息传递通信机制, 使得应用软件的编写非常困难, 给用户增加了很多负担, 其可编程性不好。

用今天的观点看, 共享存储型和消息传递型两种多机结构由于不能同时兼顾可扩展性和可编程性而显得都不够理想。共享存储型多机结构可编程性好, 方便了用户, 但其可扩展性不尽如人意。随着用户对机器计算性能要求的不断提高, 最终将被淘汰。消息传递型多机结构可扩展性好, 能通过不断扩展, 不断提高计算性能, 满足用户的要求。但众所周知, 计算机系统包

收稿日期: 1997- 06- 28; 改回日期: 1997- 12- 15

第一作者简介: 陈杰伦, 男, 1938 年 11 月生, 副教授

括两个部分: 一部分是硬件和系统软件, 另一部分是应用软件。其优良的可扩展性只能说明其系统平台的性能价格比很高, 而由于其可编程性不好, 致使用户在平台上编写或者移植应用软件的难度加大, 成本相应增高, 从而导致整个系统的性能价格比并不很高。

从近年来涌现的众多并行计算机结构看, 业界的趋势是将可扩展性和可编程性两者的特点结合于一身。

S^2MP 正是符合这一趋势的一种多机结构。S^2MP 是 Scalable Shared Memory MultiProcessor 的简写, 可称其为 "可扩展共享存储器多处理机结构", 是一种分布式共享存储器结构。S^2MP 在物理结构上与消息传递型多机结构相类似, 存储器分布在各个处理机节点上, 可扩展性良好, 但不同的是 S^2MP 将存储器分布但不局部于系统。传统的消息传递型多机结构的整个系统的物理存储容量是各个处理机节点上分布存储器容量之和, 而 S^2MP 结构的机器的整个系统的物理存储容量就是一个处理机节点上分布存储器的容量。S^2MP 结构为系统中的分布存储器在逻辑上统一编址, 让所有处理机节点可以共享系统中每个存储单元, 这也就使得具有了与传统的共享存储型多机结构相同的可编程性。虽然为了获得这种可编程性而采取的一些技术措施使得其性能比传统的可扩展并行机有所降低, 但由于其兼有可扩展性和可编程性, 性能价格比甚高。

1　Origin 实现

基于 S^2MP 结构, SGI 1996 年 11 月推出了 Origin 系列服务器[1)]。其逻辑结构可参见图 1。

1.1　基本构件

Origin 服务器系统是一个模块化的系统, 节点是它基本的构造部件。节点间互连网络是系统最重要的部分, 路由器是互连网络的基本构件。

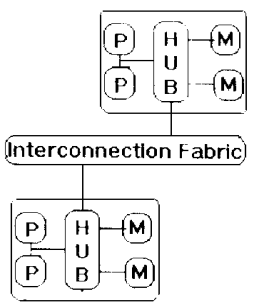

图 1　Origin 系统逻辑结构图

Fig. 1　Logical architecture of Origin system

1.1.1　节点　Origin 的节点是通用的, 每个节点的物理构造和逻辑功能都相同。每个节点内都有处理机、存储器、I/O 接口和互连网络接口, 并且用一个节点内网络将它们互连在一起。以 Origin 2000 为例, 在 Origin 2000 中, 节点被做成节点板 (Node Board), 每个节点板内含有 1 到 2 个 R10000 微处理器、一块 L2 Cache、主存储器、一个相关 cache 的目录存储器和两个接口(一个接 I/O 设备, 另一个接互连网络)。Origin 2000 节点板模块可参见图 2。

(1) 处理器　Origin 2000 上的节点板能支持 1 到 2 个 R10000 微处理器, 每个处理器以 HIMM (Horizontal In-line Memory Module) 方式安装在节点板上, 其中支持系统时钟 180 MHz R10000 微处理器的节点板上带有 1M 的 L2 cache, 支持系统时钟 195 MHz R10000 微处理器的节点板上带有 4M 的 L2 cache。

(2) 存储器　Origin2000 节点板上系统的主存储器和目录存储器使用 SDRAM 以 DIMM (Dual In_line Memory Modules) 模式安装。每个节点板上有八个存储器插槽, 其中至少应插满一条。每个插槽为主存 DIMMs 设有两个位置, 而为目录 DIMMs 设一个位置。对于 32 个处理器的配置, 其目录存储器是主存的一部分; 而若配置高于 32 个处理器, 则附加的目录存储器

1)　Origin200 and Origin2000 Technical Report. SGI 公司, 1996 年 12 月

将另置它槽。每个存储器插槽提供四路存储器交错。增加每个附加的存储器插槽都增加一个附加的 4 路存储器交错。一个具有完全八个存储器插槽的 Origin2000 节点板有 32 路存储交错。尽管这样, 具有 4 路存储交错的存储子系统已足够快, 可以提供完全的存储器带宽。使用更多的插槽能够降低重复引用同一个插槽中不同页的概率值, 并能对系统性能稍有提高。

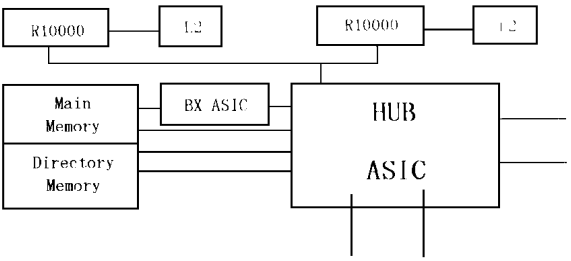

图 2 Origin 2000 结点板模块示意图

Fig. 2 Node board block diagram of Origin 2000

1.1.2 路由器(Router) 路由器是 Origin 系统互连网络的中心设备, 也可以说是整个系统的关键设备。Interconnection Fabric 就是由若干路由用链路互连构成的, 并且是路由器将节点与 Interconnection Fabric 相联。一个路由器主要由 SSD/SSR、LLP、Sender/Reciever、Crossbar、路由表(Routing Table)和路由局部块(Router Local Block)等部分组成。SSD/SSR(Source- Synchronous Drivers/Recievers)称为源同步驱动器/接收器。在路由的外部, 数据以 390Mhz 的频率、16 位的传输, 而在路由 Crossbar 内部, 数据则以 97.5 MHz 的核心频率、64 位的带宽传输。因此就用 SSD/SSR 来创建和解释路由 Crossbar 内部通信的高速、源同步信号。从外部到内部采用多路复用(Multiplex), 从内部到外部则采用多路分流(De- Multiplex)。

LLP(Link- Lever Protocol)与 SSD/SSR 接口, 主要作为路由内部与外部数据交换的链路管理设备, 提供芯片间的无错数据传输。它包括一个同步器与 ASIC 核心直接接口, 错误检测使用 CCITT CRC 码, 采用滑动窗口重发数据进行修正。同时支持 8 位和 16 位链接。Sender/Reciever 指的是路由发送器和路由接收器。路由接收器从 LLP 接收数据, 管理虚拟通道, 并向路由表和路由发送器转发数据。动态分配的存储单元对列(DAMQ)用于高负荷下的高效报文处理。分流逻辑提供给轻负荷下的性能。路由发送器为向其他芯片传输而驱动数据到 LLP。它也管理 CrayLink 信用, 其用于流控制。

路由 Crossbar 包括一系列手动优化的多路复用器, 用以控制从接收端口到发送端口的数据流。低负荷期间的报文分流允许一个报文以最低延迟通过路由器。当这不可能时, 一个波前仲裁器将决定优化路径。一个 "aging 协议 "给那些旧报文以更高的优先级, 超过更晚到达的报文。

路由表为那些通过互连网的报文提供静态的路径选择信息。为使路由延迟最小, 路由表查询采用流水线方式。每个路由器在报文进入下一个路由器时, 决定方向并由报文携带。路由表以不同于 2n 个节点提供灵活的路由和配置。

路由器局部块是路由的控制中心, 它提供对所有路由控制和状态寄存器的访问, 包括路由表、错误寄存器和保护寄存器。局部块也支持专用向量报文路由, 其在系统初始化期间使用。局部块同时也允许访问路由器性能错误比状态寄存器。

Origin 的 Interconnection Fabric 上所有的路由都被做在了路由板上, 一个路由板再加上若干电缆就是一个互连网络, 众多的节点板就插在路由板上, 组成整个系统。有四种路由板, 也就是指有四种不同的系统配置: (a) NULL 路由板: 一个路由, 接两个节点板(至多达 4 个 R10000 处理器), 该配置不能扩充。(b) STAR 路由板: 两个路由, 互连最多达 4 个节点板。(c) STANDARD 路由板: 用电缆联接 1 到 8 个封装(STAR 路由)内的 2 到 32 个节点板(内置 2

到 64 个 CPU)。(d) META 路由板: 用于联接标准路由器, 扩充系统, 从 33 到 64 个节点板(内置 65 到 128 个 CPU)。这些路由板允许配置不完全的系统(系统大小不为 $2n$), 这是因为可对路由表编程, 控制包通过 CrayLink 互连的路径选择, 并针对损坏的链接或不工作的模块重新配置。

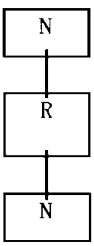

图 3　空路由板

Fig. 3　Null router

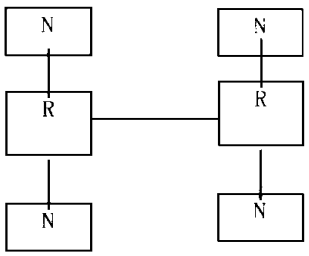

图 4　星路由板

Fig. 4　Star router

1.2　部件互连

在计算机系统中, 各部件之间的互连至为重要, 若互连得好, 则整个系统的性能会大大提高, 反之, 则可能成为系统的瓶颈。部件之间的互连涉及互连的物理特性和逻辑特性。

1.2.1　物理特性　Origin 服务器为互连不同部件共使用两种物理联接方式。一为 STL 链路, 一为 PECL 电缆。

第一种互连称为 STL 链路, 主要用于 ASIC 与 ASIC 之间的通信。STL 是一种 SGI 开发的低漂移 CMOS 信号技术, 其机器周期 2.5 ns, 其支持的通信既直接又高速。在系统中, 一般用 STL 做成一对 16 位单向链路, 其中每单向提供 780 MB/s 的峰值带宽(双向为 1.56 GB/s)。

第二种互连称为 PECL 电缆, 主要用于链接不同模块间的路由器。其原理是 STL 链路用不同的 ECL 发射机及接收机缓冲, 从而驱动电缆在封装外工作以联接模块。每根电缆是一根包含 50 对不同信号(总共 100 根导线) 的双屏蔽捆的双轴线。电缆的最大长度为 5 m。电缆可用于以如下长度支持系统最多达 64P: 13"、58"、65"、76"、88"和 108"。每个电缆联接器是一个 100 针片式联接器。一根 CrayLink 电缆的最小曲率半径是 1.25"。为确保不超过曲率半径的限制, RACK 系统装有电缆导管来为所有电缆布线。

1.2.2　逻辑特性　从整个 Origin 系统范围看, 有三种结构用于各部件之间的互连, 分别是: HUB Crossbar、XBOW Crossbar 和 Interconnection Fabric。

(1) HUB Crossbar　同一个节点上集成有处理器、存储器、I/O 接口以及互连网络接口, 有一个结构专用于将它们紧密耦合在一起, 它就是 HUB Crossbar。其实, 可将整个 Origin 系统看作是一个层次式机群系统的结构, 节点是一个群机, 此时, 如果将系统互连网络看作群间互连网络, 则节点板上的 HUB Crossbar 就是群内互连网络。

HUB Crossbar 结构将处理器、存储器、I/O 接口和互连网络接口四大部件联接在一起。它采用的不是传统 MPP 所用的单总线方式, 而是近似全互连的 Crossbar 结构。在传统单总线方式下, 存储器模块、处理器模块、I/O 模块都挂在这根总线上, 网络也通过路由设备与这根单总线相连, 其突出的缺点是本地存储器访问、I/O 操作、节点间通信以及远地存储器访问共享一根总线, 限制了几方面的可用带宽, 影响了并发操作。而在 HUB Crossbar 结构中, 每个部件有

两个缓冲,分别用作输入和输出。每个接口的输出引出一条链路,经三个分支分别连上另三个接口的输入。这样,任何一个部件的输出可同时抵达其他三个部件,任何一个部件的输入可同时接收另外三个部件的输入。本地存储器访问、I/O 操作、节点间通信以及远地存储器访问可并发执行,其可用带宽也不受限制。

在 Origin2000 节点板上完成 HUB Crossbar 功能的部件称为 HUB ASIC,每个 HUB ASIC 互连的部件通道都是 STL 链路,峰值带宽均为单向 780 Mb/s,双向 1.56 Gb/s。HUB ASIC 模块可参见图 5。

HUB ASIC 控制节点子系统内的节点间通信,也控制与别的节点 HUB ASIC 的节点间通信。XIO 或 CrayLink 互连端口采用一种外部消息格式,HUB ASIC 使用请求/应答格式发/收外部消息格式,将其转换为内部消息格式。所有的内部消息均由处理器和 I/O 设备初始化。

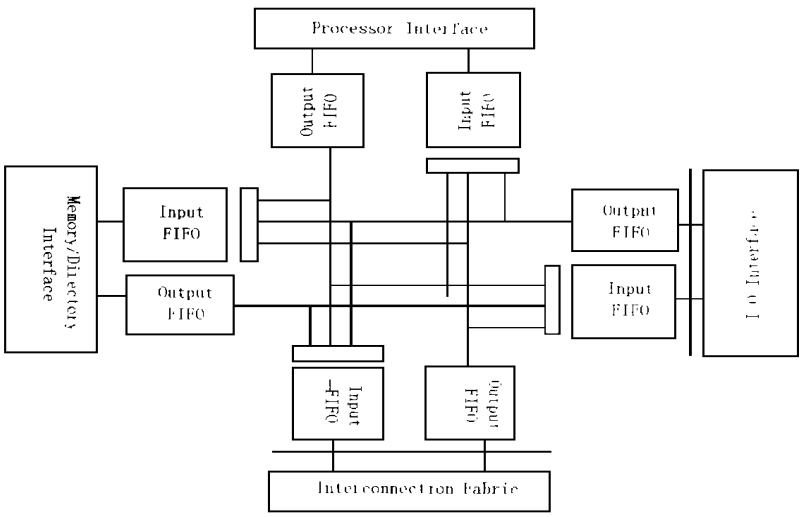

图 5 HUB ASIC 模块图

Fig. 5 Block diagram of HUB ASIC

四个部件接口充当各自子系统的独立控制器。每个接口从一个外部源取得输入,并将其转换为内部(HUB 内)消息。

每个 HUB ASIC 接口有两个 FIFO 缓存:其一用于输入消息,另一个用于输出消息。FIFO 在 HUB ASIC 与它所联接的设备之间提供缓冲。当缓存空时,FIFO 为更低的延迟提供旁路。

消息被分为请求式或应答式的。每个与 HUB ASIC 接口相联的输入和输出 FIFO 在逻辑上被分成两个队列:其一处理请求,另一处理应答。与独立的逻辑请求和应答路径一起的 Cache 相关协议保证能避免死锁的发生。

这些消息(读、写等)由各自的接口转换为针对相应目标——存储器或 I/O 接口的 CrayLink 互连请求。依赖于目录状态,目标或者对请求予以应答,或者向相同或不同 HUB 内的其他接口发出附加请求。

比如,假设一个处理器发出一个编程的 I/O 消息给本地 I/O 设备。消息由 HUB 芯片处理器接口接收,随即转换为 HUB 内格式,并通过 HUB 传到 I/O 接口。在 I/O 接口,消息被转换为 XIO 格式并置于本地 XIO 互连上。

(2) XBOW Crossbar Origin 系统的节点是通用的,不像 CM-5E 似的有 I/O 节点,Origin 系统的 I/O 设备是通过节点接在系统上的。基于若干原因,比如:Origin 的 I/O 子系统是

共享的; 使 HUB ASIC 以及节点更加模块化、性能价格比更高等等, 每个节点只有一个 I/O 端口, 而采用 XBOW Crossbar 结构接入多个 I/O 设备。

XBOW Crossbar 是一种动态纵横开关, 有八个端口, 其中两个接节点的 I/O 端口, 另外六个端口用来接 I/O 设备, 并且八个端口所采用的 I/O 协议一致, 为 XIO 协议。

SGI 将 XBOW Crossbar 结构做成 XBOW ASIC, 在 Origin 标准配置中, 总是每两个节点板使用一个 XBOW ASIC, 这样每两个节点板的两个 I/O 端口可以控制六个 I/O 设备。

(3) Interconnection Fabric　Origin 系统的节点间互连采用了超立方体的结构, SGI 称其为 Interconnection Fabric。超立方体是一种二元 n–立方体结构, 一个 n–立方体由 $N = 2^n$ 个顶点组成, 这 N 个顶点分布在 n 维上。比如, 8 个顶点的 3–立方体。(之所以用 "顶点" 这一词, 是为了与 S^2MP 中的 "节点" 一词相区别。)

在 Origin 的实现中, 一个路由 ASIC 就是 n–立方体的一个顶点。每个路由 ASIC 有 6 个路由端口, 其中 3 个端口是 STL 端口, 另外三个是 PECL 端口。3 个 STL 端口中, 只有两个 STL 端口直接接节点板, 这就产生了 Origin 的第一种配置: NULL ROUTER。NULL ROUTER 中, 只有一个路由 ASIC, 只用到了该路由 ASIC 的两个 STL 端口, 分别接了 2 个节点板、共 4 个处理器。3 个 STL 端口中余下的那个 STL 端口用来接另一个路由 ASIC(被称作 MODULE 内路由 ASIC) 的一个 STL 端口, 从而使这两个路由 ASIC 构成一个模块(MODULE)。这就形成了 Origin 的第二种配置: STAR ROUTER。STAR ROUTER 中, 有两个路由 ASIC, 每个路由 ASIC 各自有两个 STL 端口分别接两个节点板, 各自还有一个 STL 端口相连。该配置共有 2 个路由 ASIC、4 个节点板、8 个处理器。它就被 SGI 称为 MODULE, 可将每个路由 ASIC 中余下的 3 个 PECL 端口相连, 以此方法来扩充。Origin 的第三种配置: STANDARD ROUTER 就是用 8 个 STAR ROUTER 扩充而来的。STANDARD ROUTER 将 8 个 MODULE 各自的 3 个 PECL 端口互连而成, 共有 16 个路由 ASIC、32 个节点板、64 个处理器。Origin 还有一种最大的配置, 即 META ROUTER。共有路由 ASIC 48 个、64 个节点板、128 个处理器。

从超立方体的角度总结一下 Origin4 种配置的结构。(a) NULL ROUTER: 0–立方体, 1 个顶点。(b) STAR ROUTER: 1–立方体, 2 个顶点。(c) STANDARD ROUTER: 4–立方体, 16 个顶点。(d) META ROUTER: 5–立方体, 32 个顶点。

2　结束语

MIMD 机器对气象业务现代化具有较为深远的影响, 我们始终在关注着它各方面的进展, 希望通过对各种结构的 MIMD 系统的研究和剖析有助于在气象业务领域中的应用。经过上述简略分析, 我们认为, S^2MP 结构是当代 MIMD 机器中最具活力的结构, 对气象业务部门将会产生重大的影响, 关于该结构在可扩展性和可编程性两方面更具体的表现, 我们将另文介绍。由于我们手头掌握的有关 S^2MP 的资料有限, 不足之处欢迎读者指正和共同探讨。

参 考 文 献

1 Kai H W. 高等计算机系统结构 并行性 可扩展性 可编程性. 北京: 清华大学出版社, 1995

2 郑世荣, 李晓峰. 大规模并行处理系统互连通信的新技术研究. 计算机研究与发展, 1996, (6): 402～407

3 陈树清. 并行计算机的现状与发展趋势. 计算机世界(周报), 1997, (15): 97～99

4 David E C, Richard M, Karp A. Practical Model of Parallel Computation. Communication of the ACM, 1996, 39(11): 78～85

ON S^2MP STRUCTURE

Chen Jielun

(Department of Computer Science and Information Engineering, NIM, Nanjing 210044)

Zhu jiang

(Department of System, TouTru Group, Nanjing 210018)

Abstract The S^2MP structure is compared to two types of multi-machine structure, indicating the merits of the former. The achievement of the servicer is described and the S^2MP behaviors in the respect of expansibility and programibility are objectively investigated alongside their realization.

Keywords S^2MP, expansibility, programibility