集群高性能通信系统综述^{*}

杨丰¹ 都志辉¹ 刘志强²

(清华大学计算机科学与技术系 北京 100084)¹ (河北大学科技处 保定 071000)²

摘要 本文分析对比了集群系统中几种主要的高性能通信系统: Myrinet, Infiniband 和 Quadrics。在硬件方面对比了链路、交换设备与主机适配器, 指出各自的组成与特点。在软件方面则描述了各自的软件实现: Myrinet/GM, Infiniband/VAPI 以及 Quadrics/Elanlib 的主要特点和功能。最后给出各个系统的实际性能, 结果表明 Infiniband 系统性能高, 结构简单, 对应的体系结构具有良好的前景。

关键词 集群, Myrinet, Infiniband, Quadrics

Review of High Performance Cluster Communication System

YANG Feng¹ DU Zhi-Hui¹ LIU Zhi-Qiang²

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)¹

(The Science and Technology Department, Hebei University, Baoding 071000)²

Abstract In this paper, several famous high performance cluster communication systems, Myrinet, Infiniband and Quadrics are compared and analyzed in hardware and software. For hardware issues, the physical link, network switch and host bus adaptor are compared and analyzed in component and characteristic. For software issues, the software package for each system, Myrinet/GM, Infiniband/VAPI and Quadrics/Elanlib are compared and analyzed in characteristic and function. Finally, based on the practical performance benchmark result, the Infiniband system has high performance, simple structure and the Infiniband Architecture can be employed widely in different fields.

Keywords Cluster, Myrinet, Infiniband, Quadrics

1 背景

20 世纪 90 年代以来, 集群计算(Cluster)^[1] 变得越来越重要和流行。高性能网络通信是它的关键组成部分。随着硬件的进步, 影响集群间通信能力的主要因素转向软件, 因此精简传输协议、减少系统调用和数据在内存间拷贝产生的开销是提高通信带宽、降低通信延迟的重要手段。用户层通信(user level communication, ULN)^[2] 在集群系统中广泛应用。和传统的核心层通信(kernel level communication, KLN)相比, ULN 减少了操作系统在通信关键路径上的影响, 避免了由操作系统的干涉而带来的软件开销, 实现了高带宽、低延迟的通信。虽然标准以太网技术也已经发展到了 10 GB 级别^[3], 但以太网使用 TCP/IP 通信协议, 存在内存拷贝, 依靠中断, 结构复杂等一系列的问题, 并不适合为集群系统提供高效率的互连功能。因此在前一阶段对用户层通信思想的研究基础上, 逐渐发展出了一系列专门面向集群的高性能通信系统, 它们的共同特点是在硬件上基于专门设计的高速链路以及主机适配器, 往往还拥有一定的运算以及存储资源, 能够独立完成与通信相关的操作; 同时使用专门设计的高速交换设备, 实现各种拓扑结构下的高效率互连互通; 此外还拥有针对自身硬件设计特点的软件系统, 基于用户层通信思想, 为上层应用提供完整的接口, 同时还可以兼容传统的 TCP/IP 通信方式。下面就针对几种比较成熟的集群高速通信系统: Myrinet^[4]、Infiniband^[5] 以及 Quadrics^[6] 的软硬件构成进行介绍。

2 集群高速通信系统硬件构成

2.1 通信链路

高速通信的基础是通信链路, 标准以太网使用双绞线, 而高速通信系统一般都使用专门设计的链路。对于 Myrinet 系

统, 它将独立的信道组合成全双工链路, 早期产品在数据传输时每个信道的传输率为 640 MBps, 一条链路的传输率达到 1.28 GBps, 目前的主流产品可以在一条链路上提供双向 2 GBps+2 GBps 的传输速率。在物理上, 早期 Myrinet 采用 18 芯扁平电缆, 每个方向 9 根, 用来实现双向传输。这种双工并行传输方式使 Myrinet 具有很好的并行性, 而且速度很快。而目前的产品则使用光纤互连。而对于最新一代的 Myrinet 10 GB 产品, Myrinet 使用了与 10G 以太网同样的物理链路。

对于 Infiniband 系统, 它的链路部分称为 IB Link。在物理上它既可以使用铜芯电缆也可以使用光纤。无论使用何种材质, 其链路均为双工传输, 信号速率是 2.5G 波特, 能够提供每方向 250 MB/s, 聚合 500 MB/s 的带宽。与此同时, IB Link 还支持多条链路捆绑形成一条单一链路, 目前支持 1~12 条链路捆绑形成一条新链路(对应 Infiniband 的 1X~12X 工作模式), 总带宽最高可以达到 6 GB/s。一般在本机适配器上使用 4X 模式, 一个端口由 4 条独立链路(2 对, 每方向一对)聚合而成, 对应双向带宽为 2 GB/s。此外在链路长度上铜芯电缆最长为 17 米而光纤则可以达到 10 公里。

对 Quadrics 系统, 其链路部分基于 DDR 技术, 每半个时钟周期进行一次数据传输, 数据位宽为 10 位, 其中 9 位为实际数据 1 位为时钟信号。每条链路提供两个虚拟通道并为之提供相应的缓存, 同时每条链路使用了 128 路, 16 位的 FIFO 缓存进行流控。其链路带宽为每方向 400 MB/s。而其下一代产品 QsNetII 及 Elan-4 的链路带宽则达到了聚合 2 GB/s。

2.2 交换设备

对于面向集群系统的高速通信系统, 由于集群网络环境比较统一, 不存在不同网段等复杂情况, 而且由于集群通信高性能、高效率的要求, 因此在交换设备上普遍采用多端口全连通结构, 同时具有能够直接实现切通(Cut-Through)或者虫蚀

^{*} 本课题得到北京市自然科学基金(4042018)资助。杨丰 硕士研究生, 主要研究方向为计算机体系结构和集群高性能通信技术; 都志辉 副教授, 主要研究领域为网格计算与集群式计算; 刘志强 教授, 主要研究领域为高性能计算。

(Worm-Hole)等高效路由功能。例如 Myrinet 的交换机 Myrinet-2000 由 8 端口的基本模块构成,各端口全联通且每端口均可以实现链路最高速(2 GBps 或者下一代产品的 10 GBps),再通过背板连接形成支持 8-128 端口的交换设备,而不同的交换设备间还可以继续级联形成更复杂的拓扑结构,同时也避免了传统的局域网中以太网和 FDDI 网络采用总线或环形结构通信造成的信道拥挤,整个网络可以拥有很高的截面带宽(Bisection Bandwidth)。

Infiniband 使用的交换设备 InfiniScale 也是一个具有 8 端口全联通的交换设备,每端口均可以达到 10 GBps 的带宽,各个交换设备间也可以进一步级联,而且交换设备之间的互连可以支持最高 12X 的 IB Link 工作模式,确保不会出现瓶颈。同时 Infiniband 还可以支持不同 Infiniband 子网间的互连,此时使用专门的 Router 设备,既可以使用 IB Link 进行连接,也可以使用 MAN 或者 WAN 进行连接。

Quadrics 使用的交换设备 Elite 同样也是一个具有 8 端口的全联通交换设备,每个端口还支持两个虚拟通道,因此其实际联通规模为 $16 * 8$,每端口支持最大带宽为每方向 400 MB/s,其下一代产品则可以达到每端口聚合 2 GB/s。此外还硬件支持广播以及自适应的路由功能等。在扩展性方面其基于 4 叉胖树的拓扑结构,可以有效地进行多维的扩展。

2.3 主机适配器

由于高速互连系统要实现高性能的通信,在主机适配器的设计上比标准的以太网适配器要复杂得多,特别是要满足用户层通信的要求,在主机适配器上往往具有相应的运算部件与存储部件。同时为了满足不同的要求,因此主机适配器一般都是可编程控制的。对 Myrinet 系统,其主机适配器 Myrinet 网卡上集成了 SRAM 存储器和 LANai 芯片。SRAM 存储器用以存放 Myrinet 控制程序(MCP, Myrinet Control Program)和数据缓冲区,而 LANai 芯片则包含了 RISC 处理器和 DMA 引擎等部分。DMA 引擎用来实现主机与网卡和网卡与链路之间的快速 DMA 传输,而 RISC 处理器则运行存储在 SRAM 中的 MCP 程序。因此,MCP 可以完全控制 Myrinet 网卡的工作方式和流程,只要设计不同的 MCP,就可以改变 Myrinet 网卡的行为,具有很大的弹性。目前主流的 Myrinet 主机适配器基于 PCI-X 总线,133MHz, 64 位工作模式,提供每方向 250 MB/s 的带宽,而下一代 Myrinet 10G 产品则开始基于 PCI-Express 总线。

对 Quadrics 系统,其主机适配器 Elan 的结构更加复杂,以 Elan3 产品为例,其片上同时集成了两个 32 位的 RISC 运算核心,分别为微指令核心与线程核心。其中微指令处理核心支持 4 个硬件线程,分别负责处理数据输入,DMA 引擎,对线程处理核心的调度以及对主机用户层命令的处理,每个线程均可以独立地对内存进行基于流水线的申请调用。而线程核心可以协助执行一些高层消息传递协议(例如 MPI),因此它以额外的指令集实现构建网络数据包,处理事件,高效地调度线程以及有效保存与恢复线程上下文环境等功能。此外 Elan 产品的最大特点之一就是还在片上集成了一个内存控制单元(Memory Management Unit, MMU),它可以将片上集成的相关芯片的虚拟地址空间(32 位)与片上集成的内存地址空间(64 MB, 28 位)以及总线地址空间(PCI, 48 位)进行相互转换,特别是可以把片上芯片的 32 位地址空间与主机地址空间(32 位或者 64 位)进行转换,并与主机 MMU 表同步,而且这些操作对用户是透明的,这使得 Quadrics 系统具有良好的适应性。此外该适配器上还集成有路由表以及 8 KB 的高速缓存。目前的 Elan3 产品基于 PCI 总线,66MHz 64 位模式,提供每方向 400 MB/s 的带宽。而其下一代产品 Elan4 基

于 PCI-X 总线,提供聚合 2 GB/s 的带宽指标。

对 Infiniband 系统,其主机适配器 InfiniHost HCA 的结构是相对比较简单的,除了提供对不同总线结构的支持(PCI-X 133MHz 64 位 以及 PCI-Express 8X 模式),还包括一定的缓存以及 CPU Offload 功能,通过总线直接访问主机内存,通过 firmware 可以对 HCA 的相关功能进行控制。一块 HCA 上可以提供两个 IB Link 端口,每个端口均可以工作在 4X 模式,一块 HCA 就可以提供每方向 2 GB/s 的带宽。不过只有基于 PCI-Express 8X 总线的第三代产品才可以真正实现上述指标。

3 集群高速通信系统软件构成

以上分析了几种高速互连系统在硬件上的构成,下面则分别描述一下其对应的软件构成。这些配套的软件构成的共同特点就是都针对其对应产品的硬件特点进行优化,以达到最好的性能与效率。

3.1 Myrinet/GM

GM(Glenn's Message)是由 Myricom 公司专门设计与 Myrinet 配合使用的系统软件。它由 MCP(Myrinet 控制程序),DRIVER(驱动程序)和 LIBGM(GM 函数库)三部分构成。

LIBGM 为用户程序提供编程接口,其中最重要的是发送和接收函数,此外,还提供一整套通用函数,如初始化、DMA 内存管理、令牌管理、哈希管理、CRC 校验等。为提高整个系统的性能,LIBGM 与 MCP 绕开操作系统直接交互,避免了包括在用户态和核心态之间切换,从用户空间到核心空间数据拷贝等各种开销。因此,从发送和接收操作的路径上看,GM 采用的是用户层通信。

DRIVER 的功能是完成 MCP 的初始化,其次是提供一些辅助性功能。在 Linux 环境下 DRIVER 以 Module 的形式实现,当被加载到内核的时候,DRIVER 完成一系列初始化工作,包括:查找并自动配置 PCI 网卡设备,把 MCP 从主机加载到 LANai 的 EPROM 上,并且对 LANai 进行配置,发出 reset 指令,驱动 MCP 运行。LIBGM 通过调用 DRIVER 的 ioctl 功能来完成各种辅助功能。

MCP 是 GM 的关键部分,它运行在 Myrinet 的网卡的 LANai 芯片上,它由 DRIVER 加载到 LANai 上。MCP 借助 LANai 上的 CPU 和 SRAM 承担了大部分的通信处理工作,包括:在主机内存和 LANai SRAM 之间的数据 DMA 传送、数据包的拆分和组装、CRC 校验、通信的应答管理、端口和令牌资源的分配、调度管理等。主机方面的通信开销几乎为零。

3.2 Quadrics/Elanlib

Elanlib 是针对 Quadrics 配套使用的软件系统。它为底层的 Elan 硬件提供了一种用户层的访问机制。特别是在内存访问方面,基于片上集成的 MMU 控制器的对应功能,它可以把在并行环境中的每个用户进程通过主机 MMU 访问内存对应的标识符替换成一种更广泛的上下文标识(Virtual Process Identification, VPID),通过这种统一的标识,它把各个进程的地址空间集成为一个全局的虚拟地址空间,每一个用户进程均可以用这个标识和相应的虚拟内存地址通过 Elan 适配器以 DMA 访问另外结点的内存空间,也可以决定自己的地址空间如何提供给别的进程进行访问。如果用户进程是多线程的,则这些线程共享这个标识代表的内存空间,这称为 Global Virtual Memory。

具体地说,Elanlib 主要由 Elanlib、Tport (Tagged message port)和 Elan3lib(针对具体的 Elan 适配器)组成。其中 Elanlib 是位于上层的用户接口,例如 MPI 以及 Cray 的 Shmem 等就在 Elanlib 的基础上运行。Tport 提供了基本的

点对点的消息传递机制,实现阻塞与非阻塞的消息发送与接收,并且支持同步与异步两种消息传递方式。Tport 实际上是一种 MPI 实现^[7],这里 Quadrics 将其作为传输层,基于其主机适配器上的运算核心来运行。

而 Elan3lib 则处于底层,它为实际的硬件适配器提供用户层接口,它的主要功能之一就是上面提供的 Global Virtual Memory 功能,使不同的进程在一个统一的虚拟地址空间内实现互相通信,同时又维护每个进程的独立性。此外 Elan3lib 提供了一种基于事件的进程/线程同步机制,例如远程 DMA 操作完成就是一个事件,这些事件就存储在 Elan 本机适配卡的内存中。Elan 适配卡上运行的线程与主机中运行的进程就可以通过这些事件完成相应的同步操作。此外 Elan3lib 还可以提供一种完全零拷贝、零缓冲和无操作系统干预的远程 DMA 操作,这也是基于其适配器上具有的命令与线程处理核心以及可以把 DMA 描述符通过片上 MMU 直接放入片上内存的设计。

3.3 Infiniband/VAPI

IBTA 定义了 Infiniband 体系结构,定义了其各个层次的功能实现,但没有直接给出一种 API 接口定义,只是给出了

一种称为 Verbs 的概述。VAPI(Verbs Application Programming Interface)是针对 Infiniband 的实际产品 Infinihost 设计的软件接口。基于指定的硬件设计,它可以提供 S/R 传输模式和 RDMA 传输模式,同时支持可靠的基于连接的与不可靠的基于数据报的服务,此外还包括内存保护,用户虚拟内存与 DMA 转换等。此外还包括对主机适配卡的初始化以及相关的维护功能。

在具体实现上,VAPI 基于队列机制,包括发送队列与接收队列,用于存放发送数据与指定接收数据的相关指令。同时还包括工作队列(Work Queue, WQ),用于存放通信指令,还包括完成队列(Completion Queue, CQ),用于存放通信过程完成的相关标志。应用程序通过查询完成队列了解相关通信过程的完成情况(包括 RDMA 方式)。

4 性能分析

以上分析了 Infiniband、Myrinet 与 Quadrics 的软硬件构成,下面给出它们的实际产品情况及其性能表现,这些数据来自文[8,9]。表 1 说明各个系统的实际产品情况,表 2 是各个产品的实际性能表现。

表 1 不同集群高速通信系统的设备构成

名称	交换设备	主机适配器	主机总线	总线模式	理论指标
Myrinet	Myrinet-2000	M3F-PCIXD-2	PCI-X	133MHz 64bit	2 GBps 每方向
Quadrics	Elite 16	Elan3 QM-400	PCI	66MHz 64bit	400 MB/s 每方向
Infiniband 2 nd	InfiniScale	InfiniHost MT23108(双端口)	PCI-X	133MHz 64bit	每端口每方向 1 GB/s
Infiniband 3 rd		InfiniHost MT25208(双端口)	PCI-E	PCI-E x8	每端口每方向 1 GB/s

表 2 不同集群高速通信系统的实际性能对比

名称	峰值带宽	延迟(8 字节)	主机开销
Myrinet	单向 236 MB/s	6.5 μs (S/R)	0.5 μs
	双向 471 MB/s	7.3 μs (RDMA)	
Quadrics	单向 314 MB/s	2.0 μs	0.7 μs
	双向 319 MB/s		
Infiniband 2 nd	单端口单向 831 MB/s	7.8 μs (S/R)	2.0 μs
	单端口双向 901 MB/s	6.0 μs (RDMA)	
Infiniband 3 rd	单端口单向 968 MB/s	5.2 μs (S/R)	—
	单端口双向 1916 MB/s		
	双端口单向 1486 MB/s		
	双端口双向 2729 MB/s		

可以看到各个产品都基本达到了设计要求,高带宽,低延迟,其中 Quadrics 以及 Infiniband 第二代产品由于受总线带宽限制在双工模式下性能提高有限,不过在更换新一代总线后性能就会有较大的提高,根据文[10],Quadrics 的下一代基于 PCI-X 总线的 Elan4 单向和双向均可以达到 800 MB/s 左右的带宽。特别是 Infiniband 的第三代产品基于 PCI-E x8 总线(双向 4 GB/s),完全可以满足其双端口通信的要求,因此其双工模式及双端口模式下性能有接近 100%的提升,同时由于避免了原 PCI-X 桥接芯片的影响,可以基于 PCI-E 总线直接与主机内存控制器通信,因此第三代 Infiniband 产品在延迟指标上也有了一定的提升。而 Quadrics 产品借助其特有的集成 MMU 架构在延迟上取得了一定的优势。

总结 以上分析了 Myrinet, Infiniband 以及 Quadrics 这几种面向集群的高性能通信系统的软硬件构成。通过对比可以发现,Infiniband 产品拥有实际性能高、结构简单等许多优点,而且 Infiniband 是一种开放的标准,不是某一家公司的产品。实际上随着目前集群环境中节点主机逐渐向多核心 CPU、集成内存与 I/O 控制器、高性能串行总线方向演化,Infiniband 为

代表的这类具有结构简单、架构灵活可扩展的通信系统在集群系统中具有更好的前景,目前 Infiniband 产品有望成为一些集群节点的标准配件,同时为了统一不同产品,已经有一些通用的 API 被提出,例如 User Direct Access Programming Library (UDAPL)^[11]以及 Remote Direct Data Placement (RDDP)^[12]等。因此,对于面向集群的高性能通信系统,采用简单的片上策略及灵活的规模弹性是一种较好的方向。

参考文献

- Anderson T E, Culler D E, Patterson D. A case for now (networks of workstations)[J]. IEEE Micro, 1995, 15(1): 54~64
- Bhoedjang R A F, Ruhl T, Bal H E. User-level network interface protocols[J]. Computer, 1998, 31(11): 53~60
- Hurwitz J, Feng W -C. End-to-end performance of 10-gigabit Ethernet on commodity systems[J]. IEEE Micro, 2004, 24(1): 10~22
- Myricom Inc. Myrinet www.myri.com
- Mellanox Technologies. Mellanox InfiniBand InfiniHost Adapters. http://www.mellanox.com
- Quadrics, Ltd. QSNET. http://www.quadrics.com
- Beecroft J, Homewood M, McLaren M. Meiko CS-2 interconnect Elan-Elite design[J]. Parallel Computing, 1994, 20(10-11):1627~1638
- Liu Jiuxing, Chandrasekaran B, et al. Micro-benchmark level performance comparison of high-speed cluster interconnects[A]. In: Proceedings of 11th Symposium on High Performance Interconnects[C], Columbus, OH, USA, 2003. 60~65
- Liu Jiuxing, Mamidala A, Vishnu A, et al. Performance evaluation of InfiniBand with PCI Express[A]. In: Proceedings of 12th Symposium on High Performance Interconnects[C], Columbus, OH, USA, 2004. 13~19
- Brightwell R, Doerfler D, Underwood K D. A comparison of 4X InfiniBand and Quadrics Elan-4 technologies[A]. In: Proceeding of 2004 IEEE International Conference on Cluster Computing [C], Albuquerque, NM, USA, 2004. 193~204
- DAT Collaborative. uDAPL: User Direct Access Programming Library [EB/OL]. http://www.datcollaborative.org/uDAPL.v11.pdf
- IETF RDDP Working Group [EB/OL]. http://www.ietf.org/html.charters/rddp-charter.html