第 35 卷 第 1 期 计 算 机 工 程 2009 年 1 月
Vol.35 No.1 Computer Engineering January 2009
·工程应用技术与实现· 文章编号：1000—3428(2009)01—0221—03 文献标识码：A 中图分类号：TP303

StarFabric 高速总线技术的研究与应用

黄亚雯，黎 想，刘海清，柴小丽
(华东计算技术研究所，上海 200233)

摘 要：在嵌入式多处理器系统互联中，为满足大量实时数据的传输，要求互联总线具备高带宽、低延迟、高可扩展性和高可靠性等特点，导致传统的共享总线方式成为数据交互的瓶颈。该文对比分析当前流行的互联总线技术，研究 StarFabric 互联技术的特点，包括网络拓扑结构设计和 StarFabric 软件技术等，设计并实现基于高速总线的简单通信协议。通过实际应用测试验证其具有高带宽、低延迟的特点。
关键词：高速总线；StarFabric 技术；高速互联

Research and Application of StarFabric High-speed Bus Technology

HUANG Ya-wen, LI Xiang, LIU Hai-qing, CHAI Xiao-li
(East China Institute of Computer Technology, Shanghai 200233)

【Abstract】 In embedded multiprocessor systems interconnection, the highway demands higher demands to satisfy the large number of real-time data transmission, so the traditional shared bus way becomes the interactive bottlenecks of data. This paper analyzes the popular bus technology, and focuses on the StarFabric interconnect technology features, which includes network topology design and StarFabric software technology. And it designs and realizes a high-speed bus-based simple communication protocol. Practical application proves its characteristics of high bandwidth and low latency.
【Key words】 high-speed bus; StarFabric technology; high-speed interconnection

1 新型高性能互联总线的特点

随着计算机技术的发展，传统的共享总线的数据传输方式已不能满足日益增多的设备互联需求。在 PCI 总线系统这类共享总线的体系结构中，所有设备都争用总线带宽，因此，设备越多，每个设备可用的带宽就越少，从而带来严重的总线瓶颈。克服传统总线的这一缺陷的途径是使用交换互联技术，在这种点对点的交换式总线结构中，数据传输是基于包格式的，不需要地址寄存器映射，每个设备通过网络连接到其他设备，大量设备可同时通信，提高了系统带宽。另外，使用交换式总线易于实现系统的高可靠性。

除此之外，传统的并行传输技术由于引脚多，带来了一定的电气和机械特性等问题，使信号频率和信号传输距离受到限制。新型高性能总线大多采用了串行 I/O 技术，由于互联信号线数量的减少，因此消除了由并行总线带来的信号偏移问题。

目前市场上主流的交换式总线技术包括 PCI Express, InfiniBand, Fibre Channel, StarFabric 和 RapidIO 等，它们各有特点，应用领域各有侧重。

2 各种流行的总线

PCI Express 总线是 Intel 在 2001 年春季推出的，旨在取代 PCI 总线连接内部芯片。每个 PCI Express 基本连接的单向带宽为 2.5 Gb/s, PCI Express 连接结构可以有×1, ×2, ×4, ×8, ×12, ×16 和×32 几种不同形式。×32 在每个方向上具有 32 个基本连接，可以进行 10 Gb/s 的传输，实现 8 Gb/s 的实际带宽。PCI Express 技术的发展空间集中在连接北桥和图形加速器以及连接南桥和 PC 外设，属于本地互联技术。

InfiniBand 技术主要针对服务器端的连接问题，它不仅能用于单台服务器，而且可用于集群服务器以及服务器之间的高速互联。使用 InfiniBand 的系统是由多个子网构成的，子网之间通过路由器以及网桥连接，一个子网最多可以由 6.4 万个节点构成。由于消除了内部 I/O 总线，InfiniBand 总线可以使服务器的占地面积减少 60%，从而取代体积庞大的服务器。随着应用的发展，InfiniBand 越来越多地被用于存储区域网络和集群技术领域，全球前 500 位的超级计算机中有很多系统都使用了 InfiniBand。

FC 是一种利用光纤和铜缆作为物理链路的高性能串行数据接口，整个 FC 从下至上分为：

(1)FC-0：物理层，定义了不同介质、传输距离、信号机制标准、光纤和铜缆接口以及电缆指标。

(2)FC-1：编/解码层，定义了编码和解码的标准以及出错控制。

(3)FC-2：信令协议层，定义了成帧、流控制和服务质量等。

(4)FC-3：提供链路捆绑和组播等通用服务。

(5)FC-4：协议映射层，定义了光纤通道和上层应用之间的接口，在此之上提供了用于存储的 SCSI 协议、用于网络的 IP 协议(网际协议)以及映射到网络架构上用于集群的虚拟接口(VI)协议。

作者简介：黄亚雯(1983－)，女，硕士研究生，主研方向：嵌入式体系结构；黎 想，高级工程师；刘海清，工程师；柴小丽，高级工程师
收稿日期：2008-07-10 **E-mail：**slumyaya@126.com

StarFabric 技术由 StarGen 公司发起推广，主要用于嵌入式多处理机系统模块与模块之间背板级交换互联和底板与底板之间机柜级互联。StarFabric 技术的应用领域包括了 PCI 总线系统扩展、数据通信、存储系统、医疗影像、工业控制系统、军用声纳和雷达等。

RapidIO 技术主要应用在以下场合：DSP 连接，处理器和其他器件的点对点主/从连接，控制和数据背板连接，基带和 RF 板连接，芯片和处理器连接。RapidIO 可以实现效率高得多的应用，这是因为基础协议支持背板应用中的许多重要功能，如保证传送、读/写操作、消息传递、数据流、服务质量、数据平面扩展和协议封装，而这些功能并没有高层协议开销。同时作为高可靠性嵌入式系统的关键总线，RapidIO 通过物理层的握手机制、基于表的路由算法和支持热插拔等机制提供了很强的容错能力。

由于实际应用中本文选用了 StarFabric 总线技术，下面将着重分析 StarFabric 的技术特点和通信实现。

3 StarFabric 技术

StarFabric是一种高速、点对点的串行交换总线技术，主要面向实时应用的嵌入式领域的系统级互联，每一链路支持的带宽达 2.5 Gb/s，这些链路可以热插拔。具有拓扑结构简单、组网方便、高带宽(Gb/s)、低延迟(微秒级)和高可靠性等优点。其基本连接部件是桥接器(Bridge或Gateway)与交换机(Switch)。桥接器[1]的主要功能是进行PCI协议与StarFabric协议之间的转换，负责将主机接入StarFabric网络，主要器件为SG2010；交换机[2]的主要功能是实现StarFabric子网内部的高速数据交换，负责StarFabric网络的内部互联，主要器件为SG1010。StarFabric交换机总共可提供 6 对 2.5 Gb/s的双全工连接，共 30 Gb/s无阻塞的数据传输率。

3.1 拓扑结构设计

(1)网络基本结构

利用StarFabric器件(即SG2010 和SG1010)进行网络设计时，常见的拓扑结构有 3 种，如图 1~图 3 所示，分别为单行结构[3](非冗余结构)、并行结构[3]和复合路径结构[3]。

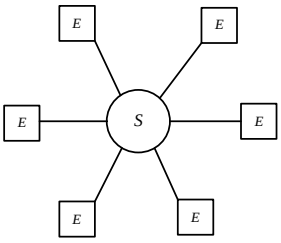

图 1 单行结构

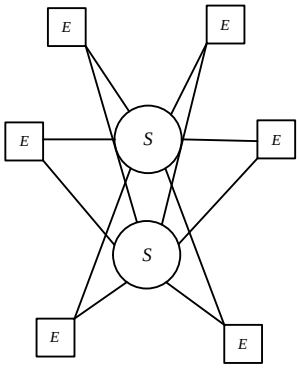

图 2 并行结构

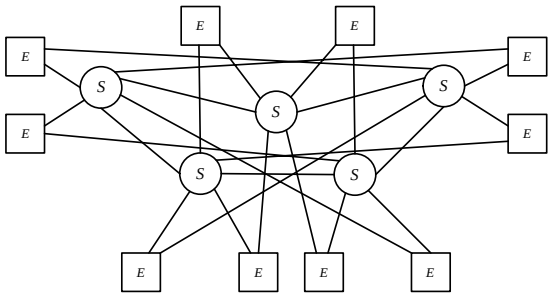

图 3 复合路径结构

在拓扑图中，SG2010 为 Edge(*E*)，SG1010 为 Switch(*S*)。Switch 有路由转发功能，Edge 不可以转发消息，只能作为消息发送的源或接收消息的目的。在 3 种结构中，单行结构不具有冗余的能力，并行结构和复合路径结构都具有冗余的特性。复合结构与并行结构的不同之处在于与边缘(Edge)节点同时连接的 2 个网络(Fabric)结构之间存在链路(link)连接。

本文在并行结构的基础上，设计了冗余的交换模块，实现冗余路由的功能。数据首先通过 Edge 的 0 链路发出，经主交换模块传递到目的节点，一旦链路 0 或主交换模块发生问题，数据可以马上通过 Edge 的链路 1 发出，使用备份模块接替工作，还可以现场更换问题模块(FRU)，从而提高了系统的可靠性。

(2)交换结构

如果只使用桥接器，则最多可连接 3 个 PCI 总线；而使用交换机(SG1010)后，可以组成非常灵活的互联网络。对于SG1010而言，通常交换设备的3个端口(Link)连接边缘(Edge)节点，其他 3 个端口(Link)连接交换(Switch)节点。这样设计的目的是为了保证每一路数据都有一条链路可用来交换，从而确保系统具有良好的实时性。图 4 是本文设计的交换结构的拓扑图，这是一个 5×5 的交换结构，可以实现大端口高速数据交换。

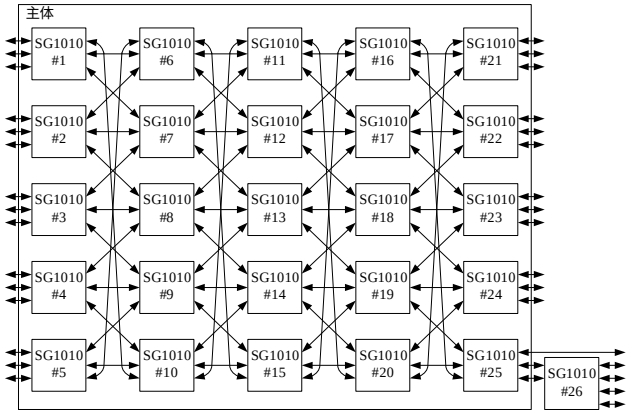

图 4 5×5 交换网络结构

3.2 协议层组成

StarFabric 协议将控制信息和数据通信集成在一个协议中，数据和控制信息在 Fabric 中是以帧的形式传递的。它规定了以下 3 层协议层：

(1)物理层规定电气信号、时钟、解码、传输介质。每条链路(端口)由 4×2 对差分信号组成物理实体，可以提供 2.5 Gb/s×2 发送和接收全双工带宽。2 个链路(端口)可以通过绑定作为一个端口使用，提供双倍的端口带宽，即 10 Gb/s。

(2)链路层规定帧格式、错误编码和恢复、链路同步。

StarFabric 使用可变长度帧格式，分别为 128 Byte 和 64 Byte。以 128 Byte 为例，端口最大输入有效带宽计算如下：

622 Mb/s×4=2.49 Gb/s

因为使用 8B/10B 编码法，所以要扣除 2 成的时脉信号：

2.49 Gb/s×0.8=1.99 Gb/s

再扣除帧头：

1.99 Gb/s×0.89=1.77 Gb/s

64 Byte 帧则由 0.8 的帧效率产生 1.59 Gb/s 的数据率。

(3)网络层规定路由方式、流控制、错误处理、错误通知、端口/链路映射、帧仲裁。StarFabric 支持 3 种交换路由方式，分别为标准的 PCI 地址路由、路径路由和多播路由方式。通常桥设备具有 2 种功能：桥功能和网关功能。地址路由方式对应于桥功能，路径路由方式对应于网关功能。StarFabric 还支持 8 种服务类别，包括异步、同步、高优先级异步、多播、地址路由、高优先级同步、预留和特殊。

3.3 StarFabric 软件体系结构

StarFabric 软件体系结构如图 5 所示。

图 5　StarFabric 软件体系结构

StarFabric的硬件层由StarFabric链路、StarFabric交换开关设备及StarFabric桥设备 3 个部分组成。FPL(Fabric Primitives Library)[4]层提供了配置和维护Strafabric硬件的功能，向上提供驱动层、内核和用户层访问硬件的接口，可以很好地屏蔽底层硬件特性。驱动层使用IOCTL API接口。用户层可以直接调用FPL提供的接口，也可以通过IOCTL系统调用方法访问驱动层。IOCTL可以完成很多功能，如IOCTL中的SF_IOCS_MAKE_ CONN功能，可以在 2 个节点间建立连接，为数据传输做准备。SF_IOCS_REG_VAL 和 SF_IOCG_REG_VAL功能可以对硬件的寄存器进行读写操作。SF_IOCS_SENDLWR_ MSG功能能够给对方节点发送中断事件。

3.4 网络初始化和驱动加载

系统上电后主要完成网络发现和驱动加载的过程。桥设备既可以作为根，也可以作为叶子。网络结构可以配置成硬根启动模式或者伪根启动模式。在硬根模式下，系统中只存在一个根节点，其余被配置成叶子节点。系统复位后，根节点进行网络枚举，为各叶子节点分配一个唯一的 FID，用以标识节点。然后通过 sfInit()完成网络发现和驱动注册。至此，可以发现设备表中已经成功安装的 StarFabric 设备，并允许用户使用 open, close 和 ioctl 等标准接口函数操作设备。在伪根模式下，所有的节点都设置成叶子节点。只有一个节点由软件伪装成根节点。伪根通过 sfInit_proot()接口发现和初始化 Strafabric 驱动。所有叶子节点在发现网络之前，都要等待伪根节点为其分配 FID。

节点的上电顺序如下：在硬根模式下，如果桥设备是使用地址路由的，则叶子节点必须在根节点之前上电。在伪根模式下，如果设计者需要更改伪根节点，则需要重新启动各个节点，然后由新的伪根为其分配 FID。这是因为网络中不允许有非唯一的 FID。

3.5 通信实现

StarFabric 提供 2 种通信机制：

(1)共享内存数据通信

StarFabric 设备可以把远程物理内存映射到本地 PCI 地址空间，也称为建立连接。通信基于远程直接存储访问(RDMA)技术。执行 StarFabric 驱动程序提供的映射操作后，对本地已映射地址的读写操作实际作用于远程主机的 RAM，并且该读写操作是阻塞的，即只有在操作实际完成后，函数调用才返回。

(2)写消息事件机制

StarFabric 提供事件通知机制。此传输方式适用于通过中断通知其他节点某种条件或情况发生，即发给对方消息并触发对方的中断，消息只有 4 Byte。本文设计了消息事件传输的接口函数，以该功能为基础，用户程序可以建立完善的异步通知机制。

本文基于这 2 种机制设计了简单的基于同步握手信号的通信协议，如图 6 所示。发送准备好数据后，向接收方发一个发送请求命令；接收方收到请求后，回复一个接收确认消息；接收方收到确认消息后，开始发送数据。假设数据的传输条件为：系统在初始化后，已经为各节点间建立了共享内存的连接空间。握手信号的发送使用本文设计的写消息事件接口来实现，其中定义数据 0x10101010 为发送请求信息，0xa0a0a0a0 为确认信息。另外，为了提高数据传输的效率，使用了 DMA 的传输方式。

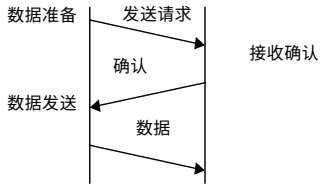

图 6　基于同步握手信号的传输方式

最后，通过带宽和延迟测试程序测试得到：在 66 MHz, 64 bit 的 PCI 环境下，使用本文的共享内存 DMA 数据传输接口函数，数据传输峰值带宽可达 210 MB/s，延迟为 6.8 μs，体现了 StarFabric 高带宽、低延迟的高速总线特性。

4　结束语

本文在研究当前主流高速总线技术的基础上，对高速互联点对点总线 StarFabric 进行了探索，并实现了高速数据通信。综上所述，StarFabric 是一种高带宽、低延迟的新型总线，基于 StarFabric 构建的网络系统具有高带宽、可扩展性好、拓扑结构构建灵活等优点。利用 StarFabric 提供的 FPL 可以实现对 StarFabric 设备的操作。另外，远程直接存储器访问技术的使用带来了通信的便利性和高效性。

参考文献

[1] StarGen Inc.. SG2010 PCI to StarFabrie Bridge Data Sheet[Z]. (2003-08-10). http://www.stargen.com.

[2] StarGen Inc.. SG1010 Starfabrie Switch Data Sheet[Z]. (2003-10-08). http://www.stargen.com.

[3] StarGen Inc.. StarFabric Architecture Specification Revision 1.0[Z]. (2007-01-15). http://www.stargen.com.

[4] StarGen Inc.. Fabric Programmer's Manual[Z]. (2003-09-09). http://www.stargen.com.