ATCA 媒体资源服务器中存储刀片的实现①

马千里, 王朝营, 赵文贤

(中兴通讯 南京研发中心, 南京 210012)

摘 要: 媒体资源服务器是 IP 多媒体子系统中的一个重要实体, 而文件存储又是媒体资源服务器的一个关键功能. 本文介绍了 ATCA 存储刀片作为媒体资源服务器中存储资源的技术实现. 该方案还可以作为一种通用的设计, 在 ATCA 架构的产品中用来代替中小规模的磁盘阵列柜, 提高产品的集成度.

关键词: 存储刀片; 高级电信计算架构; 独立磁盘冗余阵列; IP 多媒体子系统; 媒体资源服务器

Realization of Storage Blade in ATCA Multimedia Resource Server

MA Qian-Li, WANG Chao-Ying, ZHAO Wen-Xian

(Zhongxing Telecom Equipment Corporation, Nanjing R & D Centre, Nanjing 210012, China)

Abstract: Multimedia Resource Server is a main entity of IMS, and file storage is a key function of Multimedia Resource Server. This paper introduces the technical realization of ATCA storage blade as storage resource in Multimedia Resource Server. The scheme can be also used as a general design to replace RAID in product based on ATCA, for increasing integration level.

Key words: storage blade; ATCA; RAID; IMS; multimedia resource server

IMS 是 3GPP 提出的 IP 多媒体子系统, 它是 ALL-IP 下的通信网络结构, 在承载、控制、业务分离的架构下, 基于 SIP 技术, 支持多种无线和固定的接入方式, 支持与多种网络互通, 提供全面的安全解决方案和完善的 QoS 保障, 提供开放的业务环境, 能够为固定和移动客户提供新颖多样、统一融合的多媒体业务体验.

图 1 是 IMS 网络从承载层、控制层再到应用层的简单示意.

AS(应用服务器)或 CSCF(呼叫会话控制功能)通过 SIP 协议控制 MRF(媒体资源功能)实现媒体资源的管理控制、检测分发和编解码转换, 如 DTMF 收发号, 互动式语音应答 IVR, 自动语音识别 ASR, 文本语音转换 TTS, 各种音视频播放、录制和多媒体会议等业务功能.

UE(用户设备)通过 RTP 协议与 MRF 设备相连, 实现终端与网络媒体的互通和交互, 中间会经过 SBC(会话边界控制器). UE 的注册信息保存在 HSS(归属用户服务器)中.

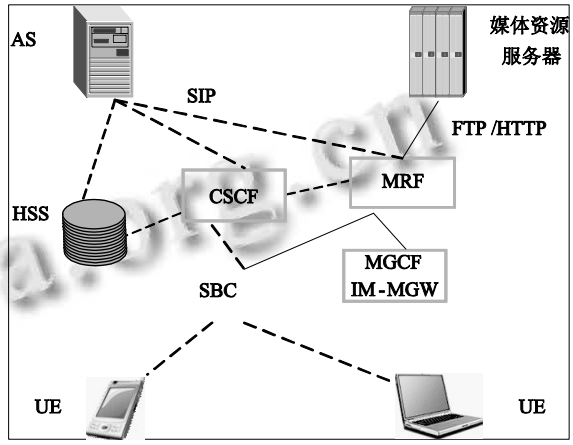

图 1 IMS 网络简单示意

IM-MGW(媒体网关)通过 RTP 协议与 MRF 设备相连, 实现 IMS 网络媒体与移动网络媒体互通和交互.

MRF 通过 SFTP 或 HTTPS 协议与媒体资源服务器相连, 实现从外部媒体资源服务器下载或上传媒体资源.

媒体资源包括 G.711、G.729、AMR 等音频编解码

① 收稿时间:2013-07-25;收到修改稿时间:2013-08-22

格式的 wave 或 MP3 音频文件, H.263、H.264、MPEG4 等视频编解码格式的 MPEG、MOV、AVI、3GP 视频文件以及数据文件.

媒体资源服务器是 IMS 网络中媒体资源的存储中心, 物理上既可以与 MRF 合设, 也可以分离.

1 系统概述

IMS 的媒体资源服务器硬件上可以采用 ATCA 架构来实现, 满足设备对高性能、大容量、高可靠、易扩展的需求.

ATCA 即高级电信计算架构, 它是国际 PICMG 组织为电信级应用制定的标准化的平台体系结构. PICMG3.0 是 ATCA 的基础标准, 它定义了 ATCA 系列规范的结构、供电、散热、互连[1], 以满足电信级应用的可靠性、可用性、可服务性、可管理性、开放性等重要特性需求.

图 2 是 ATCA 架构的原理框图示意, 整个系统通过控制消息交换平面 Base 平面、数据消息交换平面 Fabric 平面、以及管理消息交互平面 IPMB 连接在一起.

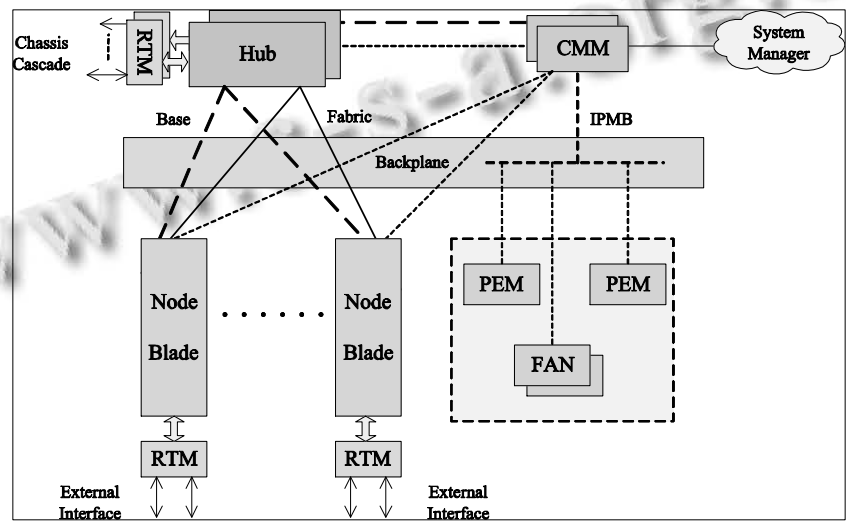

图 2 ATCA 原理架构

一对以太网交换板 Hub 构成双星互连拓扑结构, 既为若干个节点刀片 Node Blade 提供 Fabric 平面的交换, 也为节点刀片和机框管理模块 CMM 提供 Base 平面的交换.

一对主备的 CMM 为节点刀片、交换板、冗余的电源输入模块 PEM 和风扇模块 FAN 提供 IPMB 管理平面, 并由以太网连接到系统管理器 System Manager, 通过智能平台管理接口 IPMI 2.0 协议监控整个系统的硬件工作状态和进行数据管理.

Base 通道采用 1000Base-T 标准, Fabric 通道可以是 1000Base-BX、10G-XAUI、10G-KR、FC 或 PCIe 4×, IPMB 为 I^2C 总线.

节点刀片通过后插转换模块 RTM 提供对外接口, 如 GE、10G 以太网口, 或 FC 光纤通道. 配对节点刀片间有 Update 通道交互主备、负荷分担或状态更新等信息.

根据功能的不同, 节点刀片可以是服务器刀片, 或者存储刀片, 也还可以是 DSP 刀片、I/O 刀片等.

典型的 ATCA 机框有 14 个前插槽位, 其中 2 个为交换板槽位, 其余 12 个为节点刀片槽位. 机框之间通过交换板的 RTM 互联, 如图 2 所示.

2 ATCA存储刀片的实现

基于 ATCA 架构的媒体资源服务器中, 海量的音频、视频及数据文件的存储可以借助于商用的独立磁盘冗余阵列(RAID)柜如富士通的 DX60[3], 但这样需要占用单独的机柜空间, 不仅体积较大, 在 19 英寸标准机柜中占高 2U, 而且价格也高. 这里用两块 ATCA 存储刀片来实现磁盘阵列的功能, 以刀片的形式插在 ATCA 机框中, 供机框内其他的服务器刀片主机访问, 用以存储媒体资源数据, 可以显著提高产品的集成度, 同时也节约了成本, 并降低了系统功耗.

作为存储刀片, 数据吞吐量、存储能力和可靠性是

其重要指标, 这里按照 2 个 ATCA 机框级联即最多 22 块服务器刀片共享 1 对存储刀片的要求来设计存储刀片.

2.1 硬件原理

存储刀片功能上相当于一个中等规模的磁盘阵列柜, 但技术实现上与磁阵有较大区别, 它要求磁盘介质、RAID 控制器等都集成在刀片上, 并遵从 ATCA 的规范, 且在刀片间实现冗余备份功能.

Intel 近年推出的 IA 架构 Xeon 处理器增加了 RAS 即“可靠性、可用性、可服务性”来保证架构的稳定性和安全性; 不仅如此, 处理器还集成了多项存储特性, 如内置 RAID 加速器、增强的数据吞吐带宽、PCIe 非透明桥接以及异步内存自刷新; 加之处理器的架构-工艺按 Tick-Tock 交替持续演进, 这些特性所带来的高性能、高可靠、易集成、可扩展、开放化和标准化的优势助推了 Intel IA 处理器在中高端存储领域的应用, 促进了服务器和存储架构融合、以及统一存储的发展趋势.

本文的 ATCA 存储刀片即基于 Intel 存储平台而设计, 采用 Active-Active 双控工作方式, 图 3 是其硬件原理框图[2], 主要由四部分组成.

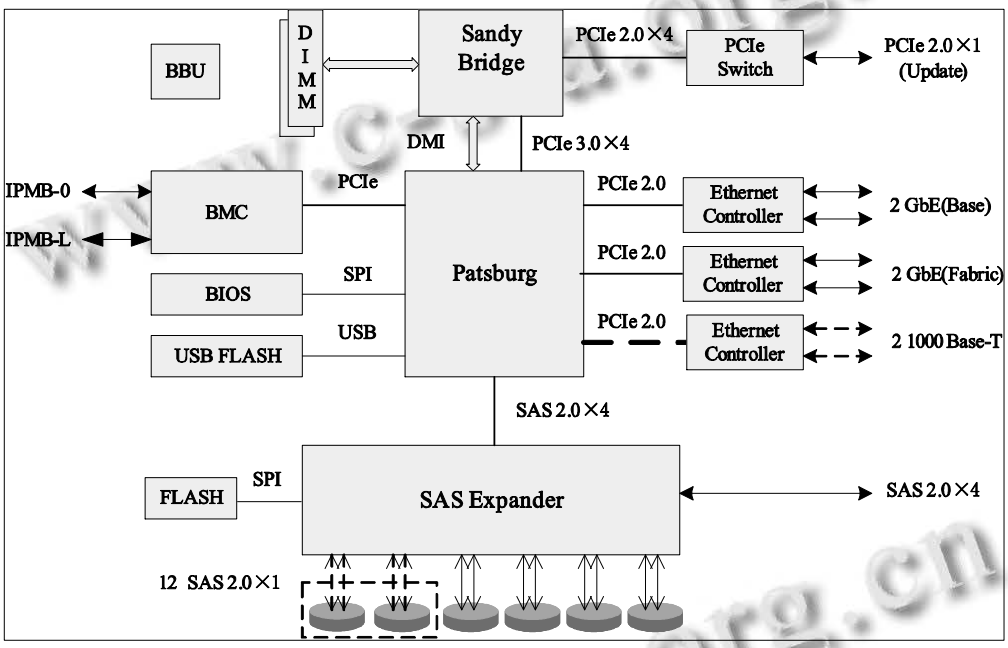

图 3　ATCA 存储刀片硬件原理

2.1.1 主控设计

存储刀片基于 Intel 的存储平台设计, 处理器采用 Sandy Bridge 系列低功耗 CPU, 内含 4 个硬件核, 主频 2.0GHz, 支持 3 通道的 DDR3 内存; 提供 6 个 Port 共 24 个 Lane 的 PCIe 3.0 接口, 支持非透明 NTB 模式; 通过 DMI 2.0 连接桥片.

图 3 中, CPU 扩展了 2 个 RDIMM, 提供 4GB 内存, 出一路 PCIe 2.0×4 的接口连到 PCIe switch, 交换芯片的 1 个 lane 设置为 NTB 方式, 接到背板 Update, 作为双控之间的 Cache 通道.

PCH 桥片选用内部集成了 SAS 控制器的 Patsburg-D, 其 SCU(Storage Controller Unit)支持 8 个 SAS 3.0 端口, 另外还集成了 PCIe 2.0, PCI, USB, SPI 以及 SATA 控制器接口.

Patsburg-D 与 Sandy Bridge 之间有一个独立的 PCIe 3.0 通道, 专门用于存储的加速. 图中利用该连接作为 CPU 和 SCU 之间的数据通道, 这样就不需要共享 DMI 总线的带宽, 以加快存储数据的读写.

PCH 的 SPI 接口连接 FLASH 作为刀片自举的 BIOS, PCH 的一个 USB 通道连接 USB FLAH, 作为刀片的版本存储空间.

2.1.2 前后端设计

PCH 出一路 PCIe 2.0×4 连到以太网控制器, 扩展出 2 个 GE 网口(根据需要, 也可以设计成 FC 通道或 10Gb 网口), 接到背板的 Fabric 面, 作为存储刀片的前端主机接口. 另外, PCH 还有一路 PCIe 2.0×4 连到

RTM, 通过以太网控制器扩展出 2 个 GE 网口(如图 3 中虚线所示), 既可以和 Fabric 面的 2 个 GE 口一起作为前端接口, 也可以用来将刀片接入存储集群中.

PCH 出一路 SAS 2.0×4 端口连到 SAS Expander, SAS 扩展器出 12 个 SAS 2.0×1 端口, 连到本板的 6 个 2.5 英寸 SAS 硬盘, 每个 SAS 硬盘各接 2 个 SAS 2.0×1 端口, 其中 4 个硬盘放在母板上, 另外 2 个硬盘放在 RTM 上(如图 3 中虚线所示), 6 个硬盘都支持热插拔操作.

双控之间的 SAS 扩展器还通过背板用 SAS 2.0×4 端口级联起来, 由此每个硬盘通过 2 个 SAS 2.0 端口分别对应到双控的 2 个 SAS 扩展器, 这样, 每个刀片的 SAS 控制器一方面通过本板的 SAS 扩展器控制本板的 6 个 SAS 硬盘, 还通过级联到对板的 SAS 扩展器控制对板的 6 个 SAS 硬盘, 所以每个存储刀片可以访问的硬盘总数达到 12 个.

2.1.3 BBU 设计

存储刀片通过电池备份单元 BBU 支持内存掉电保护. 利用 Sandy Bridge 处理器的内存自刷新功能, 当检测到电源异常时, CPU 控制 DDR3 内存进入自刷新模式, 同时将内存供电切换到 BBU, BBU 提供足够的电量, 刀片内存中的数据可以持续保持 24 小时以上; 电源正常后, 直接从内存中恢复之前缓存的数据, 不会造成重要数据丢失.

2.1.4 ATCA 兼容设计

双控之间的 Cache 通道占用了 Update 通道的 2 对差分线, Update 通道的其余 8 对差分线用作双控之间 SAS Expander 的级联通道.

PCH 出一路 PCIe 2.0 接到以太网控制器, 扩展出 2 个 1000Base-T 网口上背板的 Base 面, 作为刀片的控制通道.

前述的 2 个 Fabric 面 GE 用作 iSCSI 的数据通道.

刀片和 CMM 之间的通信由基板管理控制器 BMC(BaseBoard Management Controller)芯片来完成. BMC 通过 PCIe 连到 PCH 桥片, 通过 IPMB-L 总线连到刀片上的其他部件(如 CPU、温度传感器)及 RTM, 通过 IPMB-0 总线上背板与 CMM 相连. BMC 与 CMM 之间通过 IPMI 协议进行通信, 实现对整个刀片的工作管理, 包括对刀片运行状态的记录与上报, 上下电过程控制等.

ATCA 平台的高可用性完全可以满足磁盘阵列柜对于系统稳定性及安全性的要求.

2.2 RAID 机制

传统的磁盘阵列柜一般通过 SAS RAID 控制器实现 RAID 功能[3]. 图 3 中 PCH 桥片, 有的型号也支持 RAID 硬件加速. 这里的 ATCA 存储刀片, 基于 Intel 存储平台设计, 利用 CPU 强大的计算能力, 将 RAID 功能集成到 CPU 内部, 通过软件算法实现 RAID, 提供 RAID 0、1、10 以及 RAID5、RAID6 级别, 支持卸载 RAID 计算(XOR/P+Q).

和通常的磁盘陈列柜有所不同的是, 一对 ATCA 存储刀片的总共 12 块硬盘分处于双控工作的两块刀片上, 某些异常情况如一块刀片的 SAS 扩展器故障或一块刀片掉电时会导致有一半的硬盘访问不到, 因此对 RAID 级别须加以一定的约束. 比如对于 RAID1, 互为镜像的两组硬盘应分处于两块刀片上, 这样在一组刀片出现上述异常时不会引起任何数据的丢失.

从硬件角度来说, 服务器刀片主机读写存储刀片上的媒体资源数据流程如下: 对硬盘写数据时, 数据流经 Fabric 面的 GE 网口或 RTM 的 GE 网口进入存储刀片, 与此同时, 存储刀片也将数据流通过 Update 口发到双控的另一个存储刀片进行 Cache 同步. 存储刀片 CPU 对收到的数据流进行包括 RAID 运算在内的各种处理, 处理后的数据通过 PCIe 链路流入桥片内的 SAS 控制器, 然后再将数据经由 SAS 链路发给 SAS Expander, 最终由 Expander 将数据发送给硬盘. 读硬盘数据的过程与之相反.

2.3 软件实现

软件实现以 Linux 操作系统为底层平台, 上层模块主要包括 iSCSI 目标器模块、负载均衡模块、Cache 模块、RAID 模块、I/O 调度模块和管理控制模块, 软件结构如图 4 所示[3], 从上到下对应从前端主机接口到后端磁盘接口之间的 I/O 操作流程.

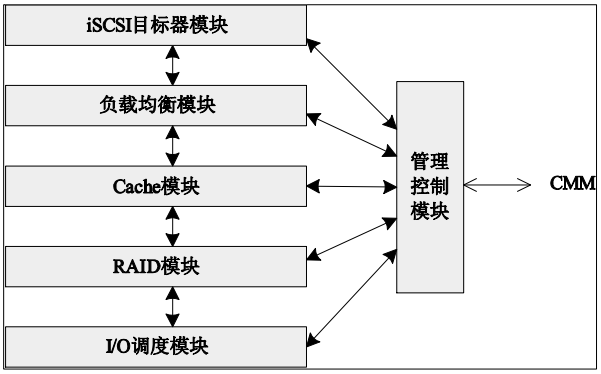

图 4 ATCA 存储刀片软件框图

① iSCSI 目标器模块: 负责接收主机的 SCSI 命令并返回应答和数据.

② 负载均衡模块: 两个存储刀片工作在对称式双主模式(Symmetric Active-Active), 之间实现高速互联的通讯. 负载均衡模块将主机对存储系统的 I/O 操作均衡地分配到两个存储刀片上进行处理, 系统不需要主机端的多路径软件参与即可自动实现负载均衡, 不仅有效提高数据传输带宽, 而且还提供冗余功能: 一旦检测到一条 I/O 路径出现故障, 即将 I/O 操作重新路由到其他正常路径上.

③ Cache 模块: 包括缓冲和镜像两部分. 对于读请求, 先判断要读数据是否已在缓存中, 如果命中就从缓存中读取, 否则即通过 RAID 模块从硬盘读取. 对于写请求, 在将数据存入当前存储刀片的写 Cache 中的同时, 也同步写入到另一存储刀片的镜像 Cache 中; 若当前存储刀片发生故障, 主机可通过读取另一个存储刀片的镜像 Cache 数据来确保数据的安全和完整. 当写 Cache 中的一部分数据成功写入硬盘后, 即将镜像 Cache 内相同的数据及时删除, 以保证数据的一致性[4].

④ RAID 模块: 通过软件算法支持多种 RAID 级别和重构, 将对逻辑磁盘的读写转换为对物理磁盘的读写, 实现磁盘阵列逻辑的处理, 包括正常模式的读写、降级模式的读写、硬盘重构及回拷、LUN 管理等.

⑤ I/O 调度模块: 接收下发到硬盘的 I/O 请求并进行分析、合并或重组, 双控模式下还根据链路情况选择在本端下发还是到对端下发, 实现特定的调度功能, 提高 I/O 通道的数据传输效率.

⑥ 管理控制模块: 包括存储资源管理和刀片运行控制两部分. 存储资源管理和上述模块均有交互, 它通过命令行或 Web 界面实现对整个存储资源的分配、管理和监视, 包括阵列管理(各级别 RAID 的创建、删除和扩容等)、卷管理(逻辑卷的创建、删除、扩容和映射等)[5]; 刀片运行控制遵从 ATCA 规范, 将刀片的运行状态上报给 CMM, 配合 CMM 完成刀片的运行控制.

3 结语

综上所述, 本方案的 ATCA 存储刀片为双控设计, 采用 Intel 新一代的 Sandy Bridge CPU, 内含 4 个硬件核, 主频 2.0GHz; 2GB 的高速缓存, 支持掉电保护; 两块刀片物理上最大可容纳 12 块 2.5 英寸 SAS 硬盘, 逻辑上支持 RAID0、RAID1、RAID10、RAID5、RAID6 等级别; 提供 Web 及 CMM 管理接口; 提供对内对外共 8 个 1Gb iSCSI 通道. 经 IOMETER 实测, 存储刀片最大数据传输带宽 2700MB/s 全双工, 最大 IOPS 可达 18 万.

在此基础上, ATCA 存储刀片还可以扩展出快照、远程镜像、虚拟化存储等丰富的软件功能[6], 在 ATCA 产品中用以替代中小规模的磁盘阵列柜的应用, 具有集成度高、成本低、功耗小等独特优势.

参考文献

1 PICMG3.0, AdvancedTCA (Base Specification) Rev3.0, 2008.

2 Intel Corporation, Sandy Bridge Platform Design Guide, 2010.

3 Fujitsu Limited, ETERNUS DX60/DX80/DX90 SA/SE Disk Storage System Handbook, 2010.

4 万亚平,冯丹,刘立,申宏建.一种基于 iSCSI 的双控制器 RAID.计算机工程,2010,36(10):8–10.

5 鲁士文.存储网络技术及应用.北京:清华大学出版社, 2010:302–311.

6 张继平.云存储解析.北京:人民邮电出版社,2013:70–78.