



海量设计文档分布式存储及负载均衡的研究与实现

苏李亮,王云福,侯 斌

(中广核工程有限公司 深圳 518052)

摘 要:在对海量设计文档的分布式存储技术进行分析比较后,选取最适合现有生产环境的解决方案,并且在分布式存储的基础上,提出与 Web 应用系统负载均衡技术相结合的系统架构方案。该方案避免用户访问远程存储资源时的漫长等待,解决大型企业局域网之间专线带宽瓶颈对网络应用限制的问题,减轻集中存储资源的负担,实现负载均衡策略从前端 Web 应用至底层数据存储的全局均衡,从而保证系统的高可用性。

关键词:分布式存储;负载均衡;海量存储

doi: 10.3969/j.issn.1000-0801.2013.12.009

Research and Implementation of Massive Design Documents on Distributed Storage & Load Balancing

Su Liliang, Wang Yunfu, Hou Bin

(China Nuclear Power Engineering Co., Ltd., Shenzhen 518052, China)

Abstract: The most suitable system architecture was presented, based on the distributed storage and load balancing, after the analysis and comparison of the technology of massive design documents in distributed storage. The solution avoids user taking a long wait to access storage resources, breaks the bandwidth bottleneck of enterprise Web application in the LAN, reduces the burden on central storage, and achieves the global load balancing between Web application and distributed storage, so as to ensure the high availability of system.

Key words: distributed storage, load balancing, mass storage

1 引言

信息化是核电工程设计单位实现发展与创新的重要手段,随着核电工程设计信息化建设的逐步深入,海量的非结构化、半结构化的工程文件信息给网络带宽和服务器响应速度带来了巨大的挑战。海量信息的分布式存储以及由负载均衡系统构建的 Web 服务器集群正是为满足这一需求而设计的,对企业应用的数据进行分布式存储,同时在现有网络结构上实现负载均衡,是一种能够提高网络带宽利用率、解决网络拥塞问题、提供空闲服务器响应、提高用户请求响应速度和避免单点故障的廉价的有效方法。

2 研究背景和现状

设计文档管理平台是中广核工程有限公司设计院建立的企业内容管理平台,其目标是实现设计文档的海量存储与管理,包括对各类结构化、半结构化工程文件的收集、分解、分类权限、审核、转换、发布、更新、删除、存储、版本控制等过程,并保证数据的准确性、可复用性、安全性。

设计文档管理平台的系统架构由 Web 应用服务器层、内容管理服务器(Documentum content server)层、存储库层以及由文件存储系统(file storage system)、数据库组成的数据存储层共 4 层组成。设计文档管理平台采用的分



层架构模式使得系统具有灵活的可扩展性,每一层均可以进行多个节点的部署,如图1所示。

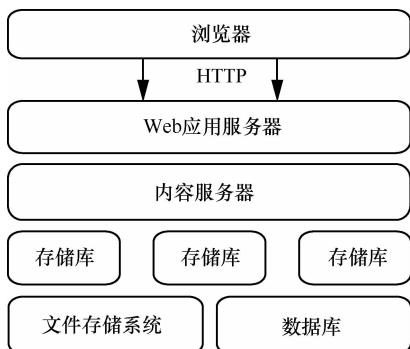


图1 设计文档管理平台系统架构

从图1可以得知,平台架构上的Web应用层以及数据存储层均以集中部署方式应用于整个企业。但随着企业的迅猛发展,用户规模大幅增长,设计平台最初的架构已很难适应公司业务多项目、多基地格局的要求,为此,本文提出对平台底层存储采用分布式技术,同时结合前端Web应用层的负载均衡对平台系统架构进行重构,以实现平台的高性能、高可用、可伸缩性。

3 分布式存储技术方案的比较与选择

分布式文件系统是分布式系统的重要组成部分,主要任务是统一管理所有分布式节点计算机上的文件资源,对一般用户和应用程序而言,它屏蔽了各个节点计算机底层文件系统的差异,对外提供统一的访问接口。分布式文件系统通常要考虑并解决以下3个关键问题。

第一,透明性问题,分布式文件系统一般由多个节点组成,需要集体协作才能对外提供服务,分布式文件系统的透明性是指通过统一管理分布在局域网(或广域网)中的所有节点服务器,使其对用户和应用程序的感觉好像只是一个高可用性的文件系统。

第二,一致性问题,即在各个应用服务器访问分布式文件系统时,它们所访问的文件和目录内容之间的一致性关系。

第三,可扩展性问题,由于分布式系统的节点数量可能随时随地变化,因而分布式文件系统管理的文件对象也是随时变化的,应用对分布式文件系统在性能和容量方面的要求也在不断地增长。

基于设计文档管理平台的体系架构,采用分层模式,首先提出从存储层上实现分布式技术的应用,对系统进行

重构与扩展。

3.1 基于内容复制的分布式存储库

在基于内容复制的分布式存储库模式下,内容管理服务器负责每一个内容对象,包括文件实体和元数据的内部复制。内容管理服务器提供透明的服务,用户可以在与一个存储库建立会话后,实现对该存储库内容对象的操作,并相应更新至其他多个节点的存储库上。

用户可以访问来自不同节点存储库的数据,并可以通过联邦制的架构方式,使统治节点与各联邦成员的节点自动保持同步,解决不同节点之间用户、群组 and 访问控制列表的一致性问题。系统架构如图2所示。

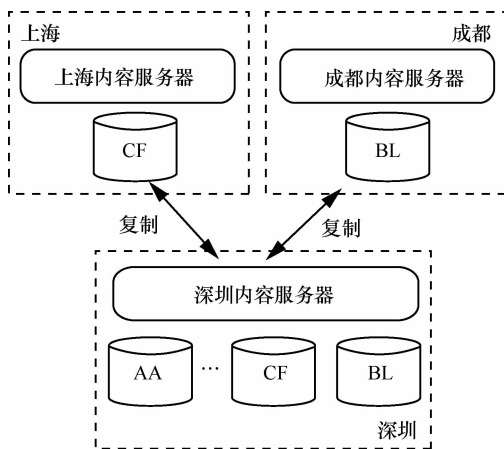


图2 基于内容复制的分布式架构

基于内容复制的分布式存储库方案具有远程节点访问数据速度快,同时支持Web和客户端形式的应用程序访问以及内容冗余增强系统数据安全等优点。但硬件成本及软件维护成本较高,各节点之间的数据复制以及一致性的保证需消耗大量系统资源,也是容易发生错误的地方,是该模式的不足之处。该模式比较适合节点产生大量需要纳入内容管理的数据,并且需要与系统配套的维护人员。

3.2 基于内容分布式存储的存储库

在基于内容分布式存储的存储库模式下,内容对象分散地存储于主节点与各分支节点服务器上,内容管理服务器通过部署中间件实现对内容对象的存储与复制。在主节点内容管理服务器上部署DMS(document messaging service)中间件,处理远程用户的异步写操作。而在每个远程节点都需完全部署一个内容管理服务器,仅负责处理内容对象的实体文件部分,同时需要部署ACS(accelerated content service)中间件,负责处理用户请求的实体文件内容。当远程用户访问内容数据时,主节点负责处理用户请

求内容数据的元数据,远程节点负责处理用户请求内容数据的实体文件,整个过程由部署在主节点与远程节点的中间件处理,解决了分布式系统的透明性问题。

内容对象的元数据由主节点内容管理服务器进行存储与维护,远程节点不包含内容对象的元数据,数据的一致性由主节点唯一处理,从而保证了所有节点内容对象的数据一致性。系统架构如图 3 所示。

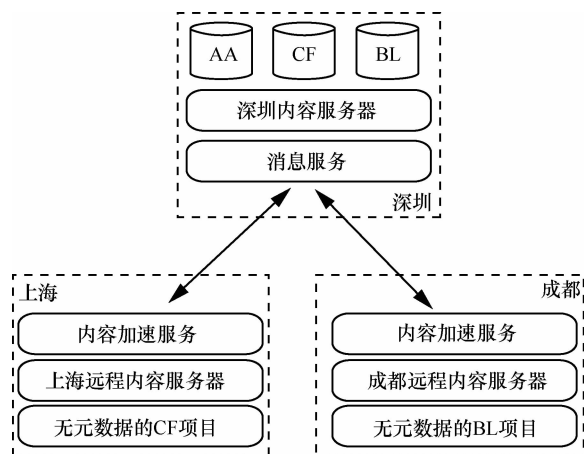


图 3 基于内容分布式存储的系统架构

基于内容分布式存储的存储库方案与前一方案相比,远程节点不维护内容对象的元数据,因此无法独立应用,数据的有效冗余度下降,硬件成本一致,但软件的维护成本则减少。该模式比较适合远程节点需要通过客户端形式的应用程序访问大量实体文件数据的情况。

3.3 分布式存储技术方案的实现

在现实生产环境中,设计文档管理平台仅采用 B/S 的模式,而上述两种技术方案所支持的 C/S 模式没有在远程节点应用的需求。因此,本文提出采用基于内容缓存分布式存储的存储库模式。

基于内容缓存分布式存储的存储库模式,内容对象以缓存的方式分散地存储于各个节点中间件服务器上,通过中间件的通信实现对内容对象缓存的存储与复制。每个远程节点仅需部署一个轻量级的 BOCS(branch office caching service,分支办公缓存服务)中间件服务器,负责对被请求内容对象的实体文件进行缓存。在主节点内容管理服务器上部署 DMS 中间件,负责处理远程用户的异步写操作。

当远程用户访问内容数据时,主节点负责处理用户请求内容数据的元数据,远程节点负责缓存并返回用户请求内容数据的实体文件内容。第一次处理时,BOCS 中间件对

该内容对象的实体文件部分进行缓存,当再次请求已缓存的内容对象时则直接返回,从而提高了用户访问数据的性能。整个过程由部署在主节点与远程节点中间件的服务器处理,解决了分布式系统的透明性问题。

内容对象的元数据及实体文件内容均在主节点内容管理服务器进行存储与维护,远程中间件节点仅处理对象的缓存数据,数据的一致性由主节点唯一处理,保证了内容对象数据的一致性。系统架构如图 4 所示。

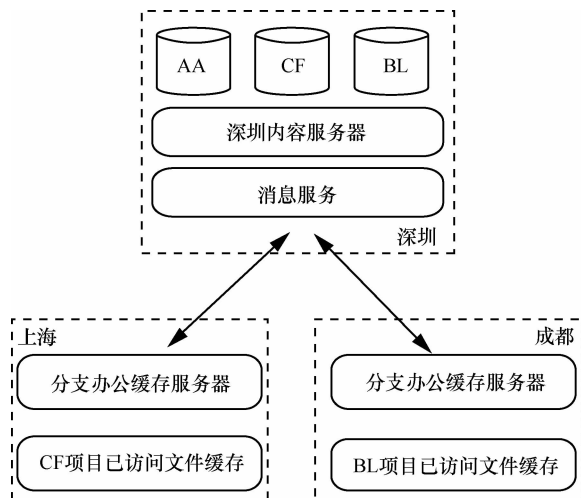


图 4 基于内容缓存分布式存储的系统架构

基于内容缓存分布式存储的存储库技术与上述两种技术方案相比,远程节点为不维护内容对象的元数据,仅是一个轻量级的缓存中间件服务器,无法独立应用,并且只可以处理 HTTP 或 HTTPS 方式的请求,不支持客户端应用程序。但在硬件成本上则大大节省,同时软件的维护成本也大大减少。该模式比较适合远程节点仅需要通过 Web 形式的应用程序访问大量实体文件数据的情况。

4 负载均衡技术方案的选择

常见的负载均衡技术按其实现方法进行区分,主要有 3 种:基于 DNS 的负载均衡、基于反向代理的负载均衡和基于 NAT 的负载均衡。

- 基于 DNS 的负载均衡技术: 为一个域名配置多个 IP 地址,用户在访问同一域名时会得到其中一个 IP 地址,不同的用户会得到不同的 IP 地址,从而访问不同的服务器,以达到负载均衡的目的。
- 基于反向代理的负载均衡技术: 使用代理服务器将



用户请求均匀地转发给多台内部服务器之一,从而达到负载均衡的目的。

- 基于 NAT 的负载均衡技术: 设置一个外部地址, 当用户访问该外部地址时, 地址转换网关会把用户请求转发给多台内部服务器之一, 并且把请求的源地地址改为自己的内部 IP 地址; 当内部服务器把响应发给地址转换网关时, 地址转换网关又把内部服务器的响应发给用户, 并且把响应的源地地址改为外部地址。

从上述介绍与分析中可以得出, 基于 DNS 的负载均衡、基于反向代理的负载均衡, 均无法满足会话保持的要求, 而设计文档管理平台是基于 J2EE 体系的 Web 企业应用, 实现内容管理等复杂业务逻辑, 不是简单的静态内容, 用户访问与操作应用系统时必须与同一台前端 Web 应用服务器建立并保持会话。因此, 本文提出采用业界较为先进的负载均衡器 F5 BIG-IP 提供的解决方案。

F5 BIG-IP 是基于 NAT 的负载均衡技术的扩展, 实现了 L4/L7 交换。F5 BIG-IP 负载均衡处理主要包括以下几个过程。

(1) 截获和检查分析流量: 保证只有合适的数据分组才能通过。

(2) 服务器监控和健康检查: 随时了解服务器群的可利用性状态。

(3) 负载均衡和应用交换功能: 通过各种策略或负载均衡算法将访问请求导向到合适的服务器, 这一过程包括目标服务器的选择及 NAT 过程。

(4) 会话保持: 通过会话保持, 保证一系列相关联的会话不会被负载均衡到不同的服务器上。

在设计文档管理平台架构, 采用 F5 BIG-IP 负载均衡的 iRule 策略将远程节点的访问通过 IP 规则分配至远端的 Web 应用服务器, 并用 least connection 策略对 Web 应用服务器集群进行监控与健康检测。例如, 上海的用户访问时, 对请求进行截获并应用 iRule 规则检查, 发现符合 IP 规则, 则将请求发送到上海的 Web 服务器, 再采用 least connection 策略对所有 Web 服务器进行监控和健康检查, 如果 Web 服务器 SH1、SH2 均为健康状态, 并且 SH2 的当前连接数小于 SH1, 则负载均衡器将请求发送至 SH2, 并维持用户与 Web 服务器 SH2 的会话直至用户的会话结束。在设计文档管理平台上, 负载均衡应用方案如图 5 所示。

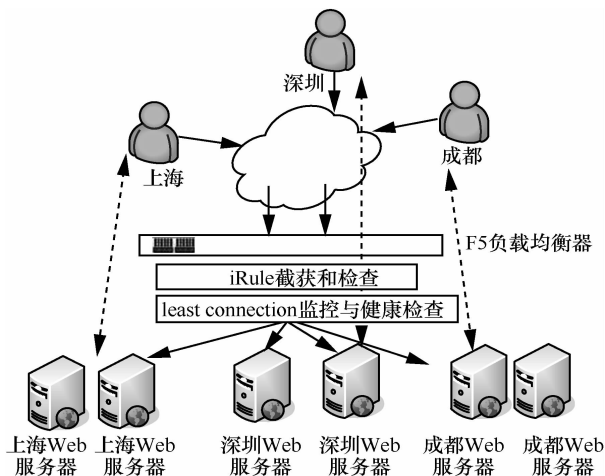


图 5 F5 负载均衡应用方案

5 解决方案的应用

综合上述分布式存储与负载均衡技术的分析与研究, 提出将两者结合到设计文档管理平台的系统架构上, 实现系统架构从前端的负载均衡到后端内容数据分布式存储的统一。目前采用新的系统架构方案已经在企业的生产环境中进行应用。改进后的设计文档管理平台的系统架构如图 6 所示。

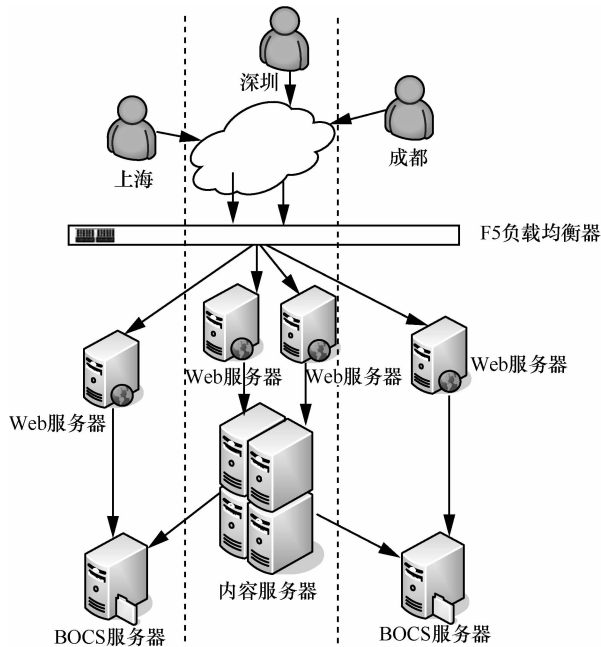


图 6 改进后的设计文档管理平台系统架构

6 结束语

本文所研究的分布式存储与负载均衡技术, 从解决大型企业局域网中分支机构访问与存储非结构化内容数据

资源问题的角度出发,提出了轻量级的内容缓存分布式存储技术方案,同时,结合负载均衡技术实现了系统的全局负载均衡,提高了系统的可用性与可靠性。目前系统已经在提供设计文档相关的应用服务,运行效果良好。

随着系统规模的增长,存储服务器有可能成为系统性能的瓶颈和失效点,然而,虚拟化和云计算技术的结合带来了全新的资源整合和使用模式,资源的按需分配和动态流动对于提高资源利用率、降低用户总体拥有成本具有十分重要的意义。基于本文现在的工作,未来的一个研究方向将是利用虚拟化和云计算技术探索更加高效、高可用、低成本的系统架构方案。

参考文献

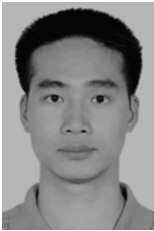
1 吴思宁,贺劲,熊劲等. DCFS 机群文件系统服务器组的设计与实现. 2002 全国开放式分布与并行计算学术会 (DPCS2002),武汉,中国,2002

2 王云岚,李增智,薛军等.基于 DNS 的负载均衡算法研究.计算机工程与应用,2002,38(4):11~13

3 高国柱,将东兴,金磊.一种基于反向代理的 Web 资源访问控制系统的设计.计算机工程,2002,28(12):244~246

4 户现锋,张大陆.NAT 技术及其在防火墙中的应用.微型机及应用,2000,19(6):32~33

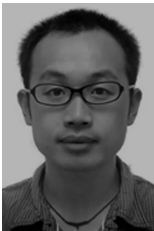
[作者简介]



苏李亮,男,中广核工程有限公司工程师,主要从事核电文档管理、核电工程设计管理、协同设计管理等信息系统方面的研究工作。



王云福,男,中广核工程有限公司高级工程师,主要从事核电信息生产系统的设计、架构、运营等方面的研究工作。



侯斌,男,中广核工程有限公司助理工程师,主要从事核电文档管理、J2EE 信息系统方面的开发与研究工作。

(收稿日期:2013-08-01)



·简讯·

2013 年 10 月通信业主要指标完成情况(一)

指 标 名 称	单 位	本年本月止累计到达	比上年同期累计	本月
电信营业收入	亿元	11 198.1	12.3%	1 152.4
其中:电信主营业务收入	亿元	9 713.3	8.6%	1 004.6
电信固定资产投资完成额	亿元	2 307.9	-6.3%	230.3
固定本地电话通话时长合计	万分钟	25 350 344.5	-15.8%	2 379 316.4
区间电话通话时长	万分钟	2 662 540.6	-20.4%	252 274.5
区内电话及拨号上网通话时长	万分钟	22 687 803.9	-15.2 %	2 127 041.9
固定长途电话通话时长合计	万分钟	4 928 109.4	-16.4 %	469 755.2
国内长途电话通话时长	万分钟	4 813 683.7	-16.4%	459 678.0
国际电话通话时长	万分钟	69 819.5	-17.1%	6 134.4
港澳台电话通话时长	万分钟	44 606.2	-17.6%	3 942.8
移动电话通话时长合计(含本地)	万分钟	240 856 816.7	5.2 %	24 843 410.5
移动电话国内长途通话时长	万分钟	55 679 497.9	5.8%	5 878 272.3
移动电话国际电话通话时长	万分钟	108 860.2	8.0%	11 879.0
移动电话港澳台电话通话时长	万分钟	80 829.4	-3.7%	8 117.6
移动短信业务量	万条	76 934 395.9	2.8%	7 230 383.7
移动互联网接入流量	万 GB	105 545.5	68.9%	12 869.2

注:(1)收入增长率按可比口径计算。(2)固定长途电话通话时长和移动电话通话时长均包含相应的 IP 电话通话时长。(3)通话时长各项指标均为去话通话时长。

(来源:工业和信息化部运行监测协调局)