第34卷 第21期 计 算 机 工 程 2008年11月
Vol.34 No.21 Computer Engineering November 2008
·博士论文· 文章编号：1000—3428(2008)21—0001—03 文献标识码：A 中图分类号：TP391

基于集中存储的刀片服务集群研究与设计

曲海平[1,2]，王秀文[1,2]，董玉敏[3]，冯 硕[1]，宋 春[1]，许 鲁[1]
(1. 中国科学院计算技术研究所，北京 100080；2. 中国科学院研究生院，北京 100039；
3. 新疆油田公司勘探开发研究院地球物理研究所，乌鲁木齐 830013)

摘 要：刀片服务器提供了强大的计算能力，但在其上构建多个服务环境时，过程繁琐且管理成本高昂。该文提出基于集中存储的Bladmin系统。该系统根据服务需求启动网络刀片，构建动态虚拟环境(VE)，满足了服务对刀片竞争性的需求，实现了VE的快速构建与动态调整。实验结果表明，Bladmin需要的构建时间较少且可管理性较高。

关键词：刀片集群；集群管理；服务部署

Research and Design of Blade Service Cluster Based on Centralized Storage

QU Hai-ping[1,2], WANG Xiu-wen[1,2], DONG Yu-min[3], FENG Shuo[1], SONG Chun[1], XU Lu[1]
(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate University of Chinese Academy of Sciences, Beijing 100039;
3. Geophysical Research Institute, Exploration and Development Institute, Xinjiang Oil Co., Urumqi 830013)

【Abstract】 Despite the computing power of blade servers, it is a tough process to build multi application environments based on blade servers, and needs high costs for management. This paper presents Bladmin system which is based on centralized storage and can build Virtual Environment(VE) dynamically by booting the blade from net according to the service requirement. Bladmin meets the need of blade rivalrousness for service, and supports the instant build and dynamic adjustment of VE. Experimental results show that Bladmin needs shorter build time and has higher manageability.

【Key words】 blade cluster; cluster management; service deployment

1 概述

刀片服务器可以使数据中心获得更高的单位成本性能、单位功耗性能和单位空间性能。与普通1U机架服务器相比，刀片服务器易于管理，TCO成本较低，处理器密度可以提高4倍，且功耗减少20%~30%，因此，被越来越多的用户接受。以刀片服务器构建的数据中心需要解决如下问题：

(1)服务的快速部署。刀片服务器处理的服务通常需要使用大量计算节点并行执行，现有的成熟软件可以完成计算任务在这些节点上的自动分发，但与之匹配的操作系统和相关配置信息的部署工作仍然难以统一、快速地完成，导致系统管理人员的负担和启动计算任务所需的准备时间极大增加。

(2)降低管理和维护成本。虽然兼具高性能与高可靠性的SAN或NAS设备已广泛应用于存储领域，但刀片服务器装备的本地磁盘无法达到与之匹配的可用性，磁盘和相关散热装置损坏已成为大规模集群管理与维护的主要负担。

(3)降低能耗。供电保障、冷却和总体能源消耗已成为超大型数据中心的主要成本负担和发展瓶颈之一，缩短部署、调度和等待时间，能从最大程度上提升系统单位能耗下的有效计算能力，并尽可能降低机房供电保障和冷却系统的负担。

随着计算规模的不断扩大，更有效的快速部署与调度成为充分利用计算资源、降低能耗与维护成本的必然要求，是保障数据中心计算环境高效、稳定工作的基础。

2 Bladmin系统设计

Bladmin(Blade admin)是为了解决当前刀片服务集群面临的诸多问题而开发的集群管理系统，它以SonD[1](Service on Demand)系统为基础，能按服务需求自主调度物理资源(计算和存储资源)，提高系统的服务能力，从而满足系统的服务质量需求。

2.1 SonD服务部署系统

SonD系统是一种典型的基于集中存储的服务部署系统，它由后端的虚拟共享卷管理系统VSVM[2]、前端的客户端系统和服务管理系统构成。虚拟共享卷管理系统负责维护服务模板及其“克隆”出来的快照网络磁盘。客户端系统负责与后端存储服务器和管理服务器的通信，以启动客户机。服务管理系统主要负责网络磁盘与客户机的动态绑定和对系统运行情况的监控等工作。SonD系统通过标准NBD协议或NVD[3]协议导出后端存储的预装系统的网络磁盘，实现客户机的无盘启动，从而完成服务的部署。基于集中存储，SonD实现了

基金项目：国家“973”计划基金资助项目(2004CB318205)；国家自然科学基金资助项目(60373045)
作者简介：曲海平(1981－)，男，博士研究生，主研方向：网络存储，集群管理；王秀文，博士研究生；董玉敏，工程师；冯 硕、宋 春，研究实习员；许 鲁，研究员、博士生导师
收稿日期：2008-02-06 **E-mail：**quhaiping@nrchpc.ac.cn

服务模板的快速“克隆”，并通过无盘启动技术实现服务环境的快速构建。

2.2 Bladmin 系统框架

Bladmin 是一个可以快速构建、动态调整的集群虚拟系统，它为每个服务创建一个用于运行该服务的虚拟集群系统，能满足该服务的要求，且不降低服务质量。此虚拟集群系统称为该服务的虚拟环境(Virtual Environment, VE)。实现对 VE 的管理是 Bladmin 的主要目标，具体分为 3 个模块：对 VE 进行申请与审批的 VE Queue，对 VE 的运行情况进行监控的 VE Monitor，对 VE 的操作进行管理的 VE Scheduler。上述 3 个模块和充当 VE 执行单元的 SonD。Bladmin 系统框架如图 1 所示。

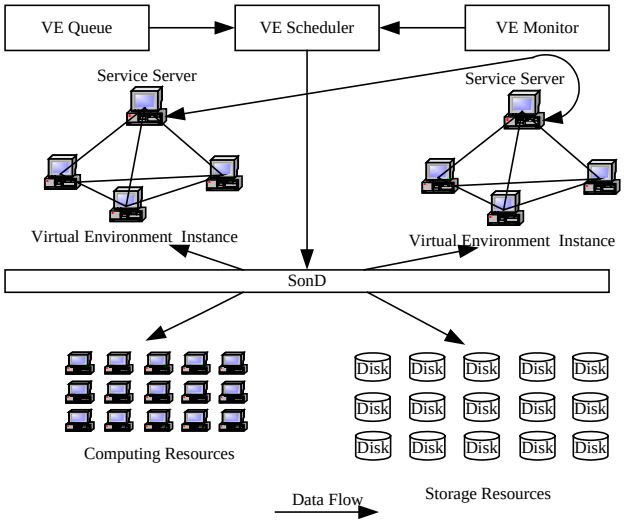


图 1 Bladmin 系统框架

2.3 Bladmin 的主要功能模块

Bladmin 的主要功能模块如图 2 所示。

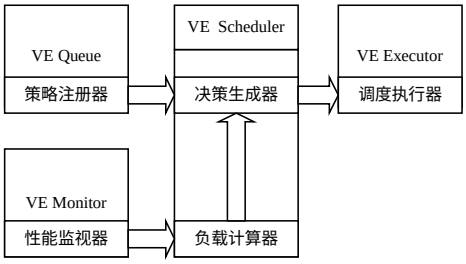


图 2 Bladmin 的主要功能模块

性能监视器获取并保存系统的原始负载数据，负载计算器根据该服务的负载计算公式计算当前 VE 的综合负载。

策略注册器提供了服务的调度策略接口，给出负载上下阈值与调度级别的定义。负载上下阈值是判定 VE 是否重载或轻载的依据，调度级别标志该 VE 是否可以进行动态调整以及调整的粒度。

决策生成器在比较当前综合负载与上下阈值的前提下，结合调度级别给出调度命令，并发送给调度执行器执行对 VE 的调整。

3 VE 的生命周期

下文以 VE 的生命周期为顺序描述了 Bladmin 系统的工作流程。

3.1 VE 的准备

(1)VE 的申请。用户通过 VE Queue 提交作业申请，主要包括对计算节点数量、硬件限制、服务模板的操作系统与运行软件、作业运行时间等服务属性的要求，以及通过策略注册器注册相关的调度策略。

(2)VE 的批准。管理员结合当前资源使用情况，通过 VE Queue 对提交申请的作业进行审批，给出同意或拒绝的批示。

3.2 VE 的构建

(1)服务模板的制作。Bladmin 提供的转移工具可以方便地把本地硬盘的软件环境拷贝到后端的网络磁盘上并设置为服务模板。

(2)服务的部署。VE Scheduler 为准备完毕的作业选择相应的计算节点，并向 SonD 发布构建命令。SonD 根据服务模板快速“克隆”出多个快照，并把快照部署给计算节点。

(3)VE 的启动。VE Scheduler 通过网络远程启动节点，SonD 通过部署的绑定关系启动系统后，节点开始进行作业。

3.3 VE 的监控

VE Monitor 采用 Nagios 开放式的监控架构，使管理员对整体有总体把握，并可以了解特定节点特定进程的工作状态。它通过外部插件来报告主机和服务的状态信息，问题发生时能通过多种方式(邮件、短信等)通知管理员。监控中采用的层次化资源模型能很好地分析并定位故障点。

(1)系统负载的获取。采用 Nagios 的 host 与 hostgroup 监控机制实现对节点、VE、集群 3 层负载数据的实时获取。

(2)服务状态的保证。用 Nagios 的 service 和 servicegroup 监控机制实现对各个 VE 中指定服务进程的监控。

3.4 VE 的调整

VE Scheduler 根据 VE Monitor 的监控数据和特定插件对 VE 进行相应的动态调整，以满足服务的质量要求。VE Scheduler 提供了易于扩展的插件机制，可以满足特定服务的需求。

(1)VE 的负载计算。本文提供了将各个节点 CPU 负载加权平均，从而得到 VE 负载的简单计算机制，开放了监控数据的处理权限，使用户可以另行添加更符合需要的负载计算公式。

(2)VE 的决策生成。VE Scheduler 提供了扩充(GROW)、收缩(SHRINK)、不变(DO_NOTHING)3 种决策机制，通过系统负载与上下阈值的比较给出对应决策。

(3)VE 的调度机制。VE Scheduler 为决策的执行提供了 EnforcePolicyHandler 和 RecommendPolicyHandler 这 2 种调度机制。前者发布调度命令给 SonD，实时进行对 VE 的调整，后者仅发布一个 alert 建议给管理员，由管理员决定是否进行实际调整。

3.5 VE 的终止

VE 可以通过作业的运行时间属性设置终止时间，也可以由管理员人为终止。VE 的终止包括删除相关快照以释放后端存储空间、删除 VE 的相关记录以释放计算节点。

4 实验结果与分析

笔者将 Bladmin 系统应用于新疆油田公司勘探开发研究院地球物理研究所的刀片服务集群上，进行了 Paradigm 作业的实地测试。

4.1 实验环境搭建

地球物理研究所的刀片服务集群共使用了 4 种刀片服务器，分别为 IBM BladeCenter 8832I1C(32 bit), RackSaver BladeRack I, IBM BladeCenter 8843IFG(EM64T)和 Verari BladeRack 2。使用传统的镜像服务进行部署，先制作一个镜

像放到镜像服务器上，然后通过网络把镜像分发到节点的本地硬盘。在 Paradigm 作业过程中，一旦某个节点宕机，整个作业就会停止并作废，只能重新计算。由于硬盘的原因，使得节点运行很不稳定。

本文加入一台 Bladmin 服务器，其硬件配置为 Intel Pentium4 2.8 GHz CPU，Intel S3000AH 主板，2x Kingston 1 GB DDR533 内存，2x Intel e1000 网卡, Areca 1230 的 RAID 卡，带 12 块 WD 250 GB SATA 硬盘。作业 Paradigm 的操作系统是 AS3.5，每种刀片先找一个节点，把这些节点上的本地系统制作成模板，创建 VE 对节点进行快速部署，然后启动节点并在 VE 下进行 Paradigm 运算。因为 Paradigm 运行节点的不可中断性，所以其 VE 策略采用当前的负载计算机制，默认阈值及不可调度的设置。

4.2 VE 构建时间的比较

VE 的构建包括模板制作、服务部署和并发启动 3 个步骤，图 3 给出了原环境与 Bladmin 环境在 64 个节点、10 GB 模板规模下的 VE 构建时间。

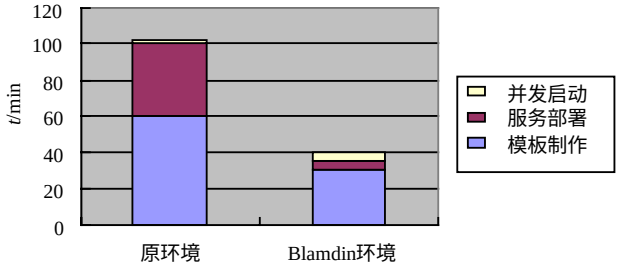


图 3 VE 的构建时间

在原环境下制作一个 10 GB 的镜像模板需时 1 h 左右，且操作复杂。在 Bladmin 环境下制作模板的时间极大减少，10 GB 的本地系统盘进行转移只需 30 min 左右。

原环境下的部署时间是分发镜像的时间，分发镜像给 Racksaver 刀片(节点 1~节点 6)时，需要 30 min 左右，而分发给 ibm32 刀片(节点 1~节点 64)时，需要 40 min 左右。Bladmin 环境下的部署时间是快照的创建时间，64 个快照的创建时间在 5 min 以内。

节点从本地硬盘上启动大约需要 2 min，且不受并发启动节点数的影响。而 Blamdin 环境读取网络磁盘的数据，并发节点启动数目会影响其节点启动时间。图 4 给出了 Bladmin 环境下的节点并发启动时间，共测试 11 组，每组以 12 个节点递增。可以看出，在 5 min 内能同时启动 64 个节点，同时启动 100 个以上的节点需要 10 min 左右。

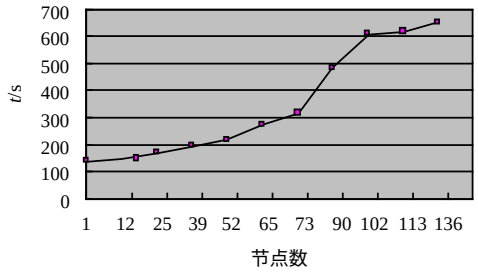


图 4 Bldmin 节点并发启动时间

由图 3、图 4 可以看出，即使在百台规模的 VE 创建中，Blamdin 的构建速度也比原环境提高了一倍以上，缩短了作业的准备时间。

4.3 Paradigm 作业的运行

Bladmin 实现节点上的系统启动，除了 swap 和 scratch 挂载的是本地硬盘，其他程序都使用网络硬盘，包括/opt 目录下约 5.6 GB 的 Paradigm 应用程序和计算作业 Inline57-160。表 1 给出了该作业在原环境 64 个节点与 Bladmin 环境 64 个、128 个节点上的运行时间。

表 1 Paradigm 作业的运行时间

	总计用时/h
原环境 64 个节点	30.27
Bladmin 环境 64 个节点	30.34
Bladmin 环境 128 个节点	16.38

由表 1 可以看出，当节点数相同时，运行同一个作业，原环境与 Blamdin 环境所需时间相近；在同样的 Bladmin 环境下，128 个节点的运行时间是 64 个节点的一半左右，节点个数与运行时间满足线性关系。可见，使用 Bladmin 对 Paradigm 的应用几乎没有影响。

由作业在 128 个节点上运行时的 Blamdin 服务器负载可以看出，此应用模式对服务器的压力很小，CPU 通常处于空闲状态，如图 5 所示。由此可见，Bladmin 系统可以满足 Paradigm 作业的运行需求。

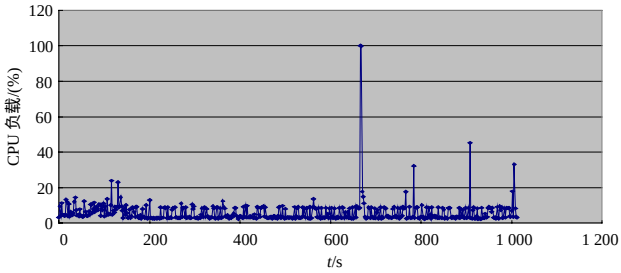


图 5 Bladmin 服务器的 CPU 负载

5 结束语

本文提出的 Bladmin 集群管理系统能根据用户提交的作业申请自动、快速、灵活地构建所需的 VE，其基于 SonD 的动态服务部署机制极大缩短了 VE 的部署时间。开放的监控架构和灵活的调度机制在降低管理和维护成本的同时，提高了整个数据中心的资源利用率。

参考文献

[1] 刘振军, 许 鲁, 尹 洋. 蓝鲸 SonD 服务部署系统[J]. 计算机学报, 2005, 28(7): 1110-1117.

[2] Yin Yang, Liu Zhenjun, Yang Shuqing, et al. Vsvm-enhanced: A Volume Manager Based on the Evms Framework[C]//Proc. of Grid and Cooperative Computing Workshops. [S. l.]: IEEE Press, 2006.

[3] Tang Haiying, Feng Shuo, Zhang Huan. NVD: The Network Virtual Device for HA-SonD[C]//Proceedings of International Workshop on Networking, Architecture, and Storages. Shenyang, China: [s. n.], 2006.