CRANE:面向科学计算与企业信息化的云计算平台

CRANE: A Cloud Platform for Scientific Computing and Enterprises

徐骁麟 吴 松 石宣化

（ 华中科技大学计算机学院服务计算技术与系统教育部重点实验室/集群与网格计算湖北省重点实验室 武汉 430000 ）

1 总体概述

随着虚拟化技术的成熟以及人们对绿色计算的关注，云计算的优势逐渐显露出来。成本的低廉、资源利用率的提高、服务整合的节能以及服务使用的便捷，无不吸引着越来越多的厂商和用户投入到云计算的怀抱中。

作为云计算的核心，云计算平台的发展也是日新月异。随着亚马逊EC2的诞生及成功商业化运营，到Google App Engine提供的免费云计算服务，这些用途各异、门槛不同的云计算平台都极大推动了云计算技术的发展。国内云计算领域的发展也紧跟国际步伐，多个科研院校都积极进行云计算平台的开发工作。在此背景下，华中科技大学开发了CRANE云计算平台。

CRANE是服务计算技术与系统教育部重点实验室自主开发的一款新式云计算平台，紧密结合当前国内对云计算服务的需求，融合了IaaS及PaaS两个云计算服务层次，提出了企业信息化云服务和科学计算云服务这两大类特色PaaS平台。普通用户可以利用PaaS平台享用通用的开发环境，例如提交科学计算任务、部署企业信息化环境等；对于想高度定制操作环境的高级用户，可以通过IaaS平台申请所需资源，自主搭建满足自己需求的操作环境。CRANE平台采用虚拟化技术，在底层提供虚拟化资源，向IaaS用户和PaaS平台提供可定制的虚拟化环境支持。CRANE充分考虑了云平台的扩展性，提供了多数据中心协同接口及公共云扩展接口，为云规模的有效扩展提供了强大的支持。后文将对CRANE云平台做系统性的介绍，并指出CRANE下一步的发展规划。

2 基础架构及设计目标

CRANE云平台（如图1）分为驱动层、核心管理层、服务层以及接口层四个层次，同时还包括辅助功能以及云扩展等部分。驱动层增强了云平台的扩展性，能够支持多虚拟化平台、多存储系统，同时为大规模细粒度监控管理提供支持。核心管理层作为云平台的中心，提供重要的管理组件，负责对资源进行有效分配调度，使上层专注于服务提供。服务层包括三种常用服务，接口层则提供用户多种交互方式。

CRANE致力于提供模块化的云平台系统，以保证云平台的灵活可定制性。根据需求，用户可以有选择的安装或卸载有关组件，增加必要的第三方模块等。CRANE提供极具特色的多种服务。个性化的IaaS能够进行服务器自动整合，优化虚拟机放置；企业信息化PaaS采用分布式结构，保证web与数据库的负载均衡；科学计算PaaS提供多种科学计算环境，智能高效的作业托管。云监控系统可提供详细的细粒度监控数据，并能够对系统性能瓶颈进行分析诊断。CRANE不仅可以供企业客户使用，也可以供科研机构使用，不仅能够提供专用服务、公共服务，还可以提供个性化服务，极具实用性及发展潜力。

CRANE云平台的总体设计目标是：

（1）综合应用前沿技术。云计算平台基于虚拟化技术，为数据中心提供一套资源优化整合的综合解决方案，向最终用户提供灵活可定制的工作环境。充分整合绿色计算、虚拟化和云计算的核心思想及关键技术，打造一个集提供功能强大的IaaS、PaaS服务于一身的云计算平台。

作者简介：徐骁麟，博士研究生，研究方向为云计算；吴松，教授，博士生导师，研究方向为云计算、网格计算、计算系统虚拟化、网络存储等；石宣化，副教授，研究方向为云计算、并行分布式计算。

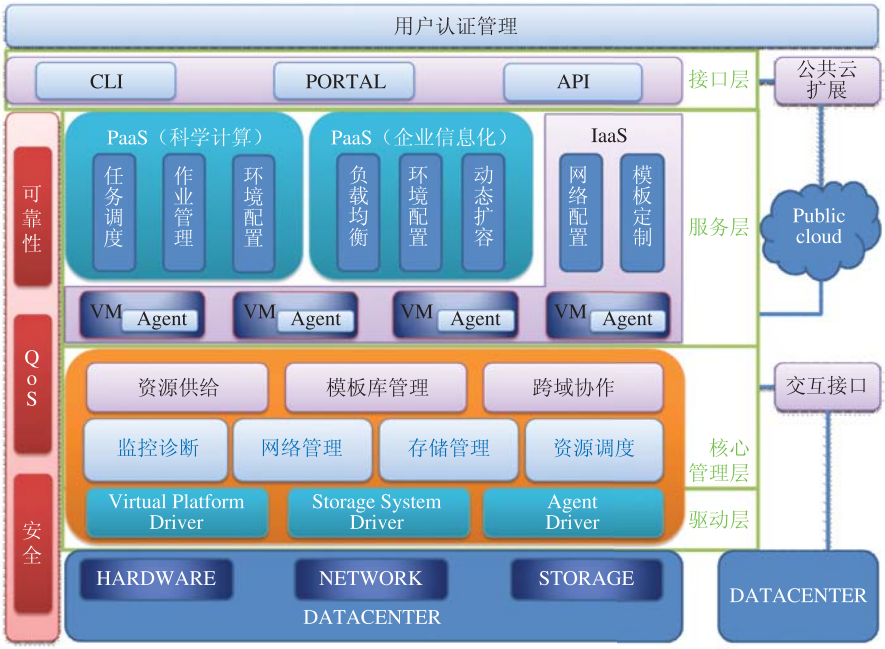

图1　CRANE云平台系统结构图

（2）可扩展的系统规模。云计算平台能够支持多个数据中心的资源整合，能够支持数百服务器、数千虚拟机的规模。既可以在科研机构等小型数据中心提供私有云计算环境，也可以在电信等大型数据中心提供公共云计算环境。

（3）多层次的平台功能。云计算平台提供基础的IaaS服务，向用户提供高权限的操作系统环境，利用此服务能够满足不同类型用户的多种特色需求。同时，平台提供多种PaaS服务，企业信息化平台能够满足企业快速信息化的需求，科学计算平台能够满足科研机构动态可变的计算环境的需求。

（4）模块化的系统结构。云计算平台向开发者提供功能强大的开发接口，使得开发者能够为平台提供第三方功能模块或者插件，以自定义平台环境或者丰富平台功能。高度的模块化使得对平台功能的精简、丰富等变得非常简单。

（5）高可靠的运行环境。云计算平台提供容错恢复功能，使得系统或者工作环境出现故障后，能够及时恢复到正常状态；系统实时监控诊断功能，能够在系统出现故障时及时诊断出故障发生原因及地点，将系统宕机时间降到最低，并及时修复系统漏洞。

3　科学计算平台

3.1　背景及意义

当前High Performance Computing（HPC）应用主要面向科研机构、大型企业等，因为要获得HPC服务，必须具备如下条件：昂贵的硬件基础设置、充足的能源、专业管理人才，这使得获取HPC服务需要付出昂贵的代价，只有科研机构、大型企业等高端用户才能承受得起，而使得普通用户只能望‘HPC’兴叹。调研表明：缺乏HPC专业人才、建设和运维的高昂成本以及使用HPC应用的复杂度，是阻碍普通用户获取HPC应用的主要挑战，而云计算正是应对这些挑战的最佳途径。HPC与云计算结合，以云计算的方式提供HPC服务，这使得普通用户付出较小的代价即可获取HPC服务。

HPC Cloud这一模式刚刚提出，就在工业界、科学界引起强烈反响，一些企业或者科研机构纷纷推出HPC Cloud服务或者构建HPC Cloud系统。Amazon Elastic Compute Cloud(EC2) 已经开始了HPC Cloud尝试，前不久推出了针对HPC应用的Cluster Compute Instance，基于这种Instance，用户使用EC2的Amazon Web Services(AWS)可以很方便快速的构建Cluster，默认配置为8个Instances。荷兰高性能计算与e-science服务中心SARA开发了一个HPC Cloud系统，目前处于beta测试阶段，测试系统包括16个计算节点总共128个核。SARA系统主要是针对那些不能在超级计算中心运行但却可以在本地集群、工作站或者PC运行的HPC应用，例如那些需要专用或者定制函数库的应用软件。SARA宣称，到2010年底，将提供面向科学社区的产品级的HPC Cloud服务。公共云只是提

供了经过虚拟化的裸机，用户需要进行配置才能利用其进行高性能计算，而这个工作对大多数不具备足够HPC知识的普通用户而言难度不小。针对这一问题，MIT的一个研究小组开发了一个套件starCluster，用来在Amazon EC2上配置和部署高性能计算环境，包括了操作系统（Ubuntu Linux 9.10）、MPI（OpenMPI）、文件共享（NFS）、SSH、作业调度系统（SGE）和高性能数学库ATLAS等。

当前HPC Cloud主要提供计算资源，并在计算资源之上构建Virtual Cluster、部署MPI计算环境、作业调度系统等，这主要解决了从裸机到虚拟集群的资源聚合问题，方便传统HPC用户直接使用HPC Cloud服务。

目前像EC2、SARA提供的HPC Cloud，仍然属于IaaS，对于不具备专业知识的用户来讲，操作起来还是很不方便的。

3.2 系统结构及特色

CRANE 科学计算平台（CRANE MPI PaaS）是基于CRANE IaaS，针对MPI应用程序，集MPI环境部署、作业调度、作业管理于一身的PaaS平台。CRANE平台中MPI PaaS架构图如图2所示。

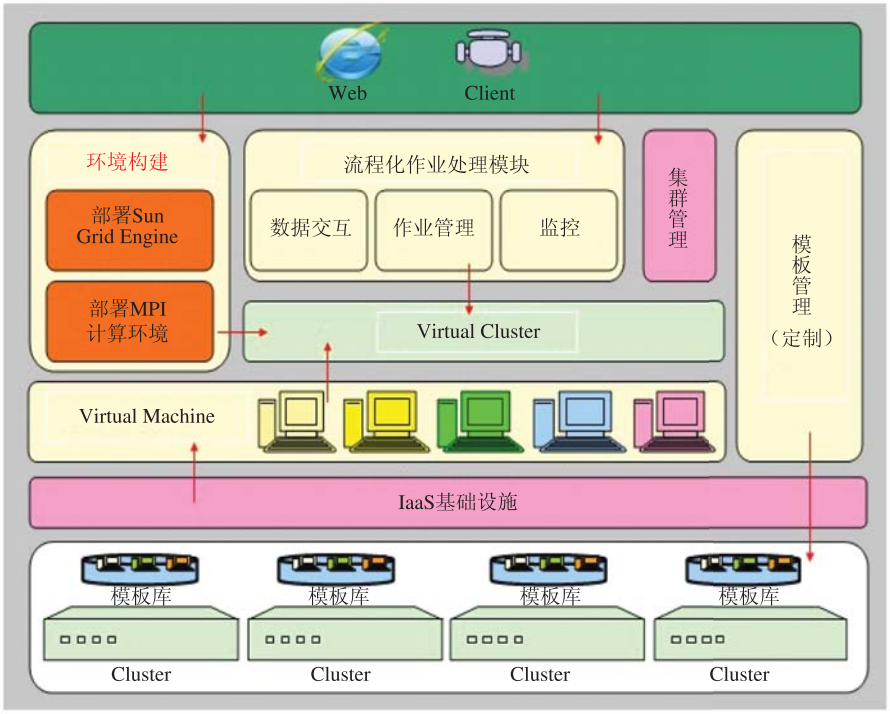

图2 CRANE MPI PaaS架构图

从架构图，我们可以看出MPI PaaS包括模板定制、环境构建、流程化作业处理等模块。CRANE为用户提供虚拟集群、模板定制、流程化作业处理等服务。

虚拟集群

CRANE平台为用户提供虚拟集群，虚拟集群上部署MPI计算环境、SGE作业调度器、SSH配置。用户获得虚拟集群的root权限，可以根据自己所需，部署MPI应用软件、定制特定函数库、计算HPC任务等。用户获取虚拟集群服务示意图，如图3所示。

模板定制

CRANE平台为用户提供镜像模板定制功能，用户可以使用模板定制功能，将MPI应用软件、特定函数库部署到模板之中，并将模板保存到CRANE平台的模板库中。

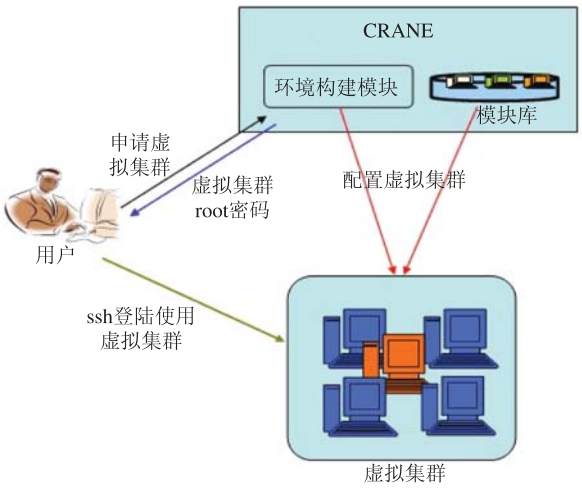

图 3 用户获取虚拟集群服务示意图

用户使用定制模板服务时，CRANE平台使用基础镜像模板（基础镜像模板中部署有MPI软件）启动一台虚拟机，并将root密码提供给用户。用户通过SSH登陆到虚拟机之中，配置虚拟机，配置完成后，使用CRANE平台的模板管理功能，将模板保存到CRANE平台之中，并按CRANE平台的要求对模板进行一个属性描述。用户定制镜像模板示意图如图4所示。

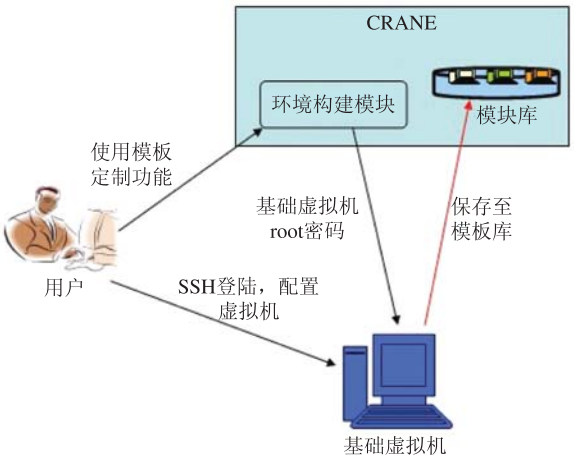

图4　模板定制示意图

流程化作业处理

用户完成一个MPI HPC作业需要哪些基本操作呢？据调研，基本操作如下：(1)申请一个虚拟集群；(2)部署MPI应用软件、或者程序；(3)传输数据到虚拟集群中；(4)启动作业；(5)计算结束，下载结果，释放申请的虚拟集群。流程化作业处理服务可以为用户完成这一系列操作，用户只需提供作业所需的参数即可。

流程化作业处理由流程化作业处理模块实现，如图5所示。

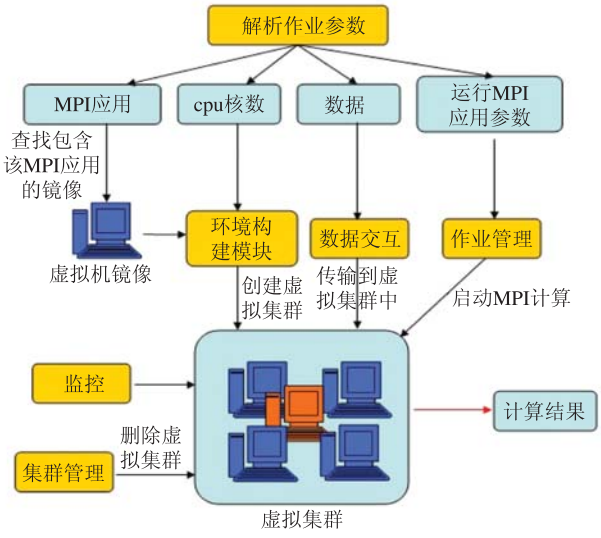

图5　流程化作业处理模块

流程化作业处理功能依赖于模板定制功能，用户通过流程化作业处理功能来完成MPI HPC作业，其所需的相应MPI应用软件必须预先部署到模板镜像之中。

目前，第一个版本的MPI PaaS处于测试之中，我们期待2010年年底之前，能推出一个完整、稳定的版本。

4　企业信息化平台

4.1　背景及意义

Web服务在企业信息化中起占据举足轻重的地位，是目前各类企业普遍采用的企业信息化途径。在网络和信息时代，海量的Web应用部署在分布于世界各地的服务器上，提供着各类Web服务；而云计算大大改变了人们使用计算机服务生活生产的方式，将这两者结合起来，研究如何利用云计算来更好地为人们提供Web服务，是一件很有意义的事情。

Web技术经过多年的发展，已经变得很成熟，但是绝大部分人还只是限于使用别人架好的Web服务，其中一个很重要的原因就是Web主机资源宝贵，大部分主机价格昂贵，一般人承受不起。而云计算利用相对集中的资源，在一个有较好的稳定性和扩展性的基础架构上，按需为人们提供各类服务，这就为解决传统Web主机扩展性差、资源利用率低、部署困难等问题创造了条件。CRANE云平台中的企业信息化平台正是基于此考虑，试图探讨如何利用云平台的IaaS基础设施为企业提供优质的信息化服务，让部署一个信息化服务变得廉价和快捷，并且具有良好的扩展性。

4.2　系统结构及特色

企业信息化平台在IaaS基础上实现，其提供了负载均衡、动态扩容、环境配置等功能，面向用户提供了企业网站搭建及运行环境。能够轻松应对访问量的突变及负载的不均衡，保证服务质量。

企业信息化平台服务架构如图6所示。Local域是外围的全局性的区域，内部由高速网络互联。Local域中包括多台物理资源，这些物理资源交给IaaS层统一管理和调度。在上面的PaaS层看来，所有的资源都是虚拟机形式的。Web PaaS层负责从IaaS层取得虚拟资源，并按照用户和应用将资源配置为一个个应用域，每个应用域实际上是一个配置好了相应应用环境的虚拟集群。应用域中的虚拟机包括三种角色，一个负载均衡器，多个Web服务器，一个存储服务器。负载均衡器为前端机，负责分发用户的请求；多个Web服务器构成后端机共同提供Web服务；存储服务器用

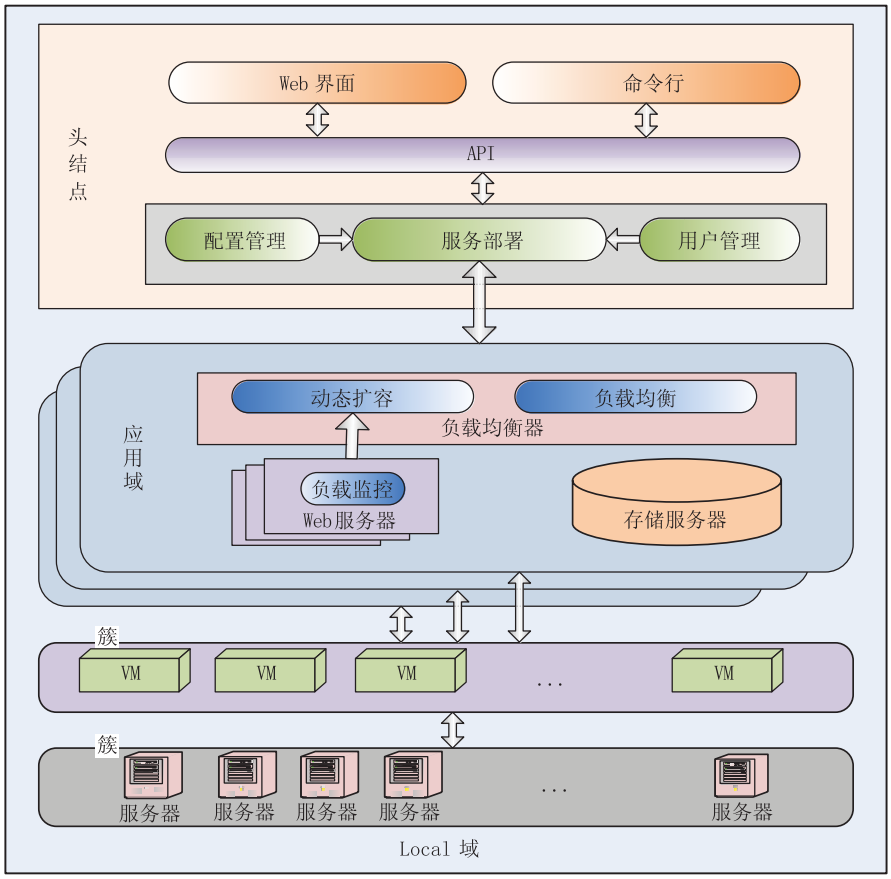

图6 企业信息化平台服务架构图

于存放数据库等需持久化的资源。

界面层放置在Local域的头结点上，也即Web PaaS用户使用服务的入口在头结点上，目前同时支持Web界面与命令行两种方式，均通过统一封装好的API函数来实现与下层功能的交互。此外，头结点上还包含配置管理、用户管理和服务部署模块。应用域负载均衡器包含动态扩容模块。而负载监控模块则放置在应用域Web服务器上。

配置管理模块负责管理用户与应用的配置，解析用户请求，生成相应的配置，维护应用的状态。用户管理模块负责接管通过平台验证的用户，维护用户、应用与资源的映射关系，同时创建和维护服务级别的用户权限信息。服务部署模块负责应用环境的部署，通过与IaaS层交互来申请应用域资源，通过与应用域机器的通信来完成服务组件的配置。动态扩容模块负责维护扩容策略，并根据相应的配置与负载信息来决定扩容的触发时机，然后与服务部署模块交互来实现服务的动态扩容与缩容（如图7）。负载监控模块负责监控和收集各Web后端的负载信息，并将信息传递给动态扩容模块。

动态扩容是Web PaaS的一大特色，也是它相对于传统主机服务的最大优势。传统的主机租用服务一般包括专用主机、虚拟主机、主机托管三大类，无论是哪一类，由于没有引入虚拟化技术，服务的容量一开始就固定了，无论负载是大是小，总是固定数量的机器在运行着。而Web应用的特点是负载变化较为频繁，这种静态的架构一方面可能无法满足较高的负载，另一方面可能在负载较低的时候由于空转而浪费资源。动态扩容的出发点就在于此，力图通过负载信息去实时监控Web服务器实时的负载情况，然后根据一套策略来进行自动地扩容和缩容。目前只取了

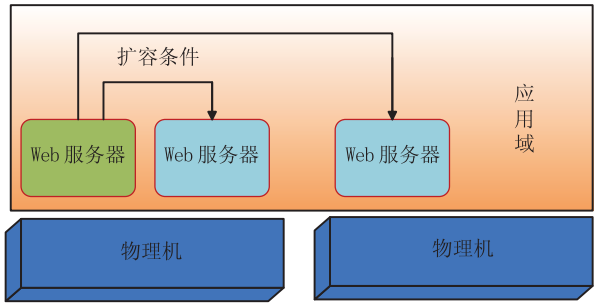

图7 企业信息化平台扩容示意图

内存平均利用率，策略也较为简单，只是当利用率超过某个临界值时就扩容，低于某个临界值时就缩容。事实上，不同的Web应用具有不同的负载特征，一般来说，Web应用对内存的消耗都是很大的，但对CPU的消耗和带宽的消耗却各有不同，如果能够根据应用类型制定相应的策略将更好，这是未来要做的工作。总之，动态扩容能够做到真正的需要多少资源就使用多少资源，对服务商来说可以提高资源利用率，而对用户来说则可以做到按使用付费。

5　个性化IaaS平台

5.1　背景及意义

用户通过Internet从计算机基础设施获得服务（如存储和数据库）。这类服务称为基础设施即服务（Infrastructure as a Service，IaaS）。IaaS可以粗略地定义为根据需要在用户环境之外以服务形式提供的可伸缩计算资源。用户只使用需要的资源，只为使用的资源付费。在任何时候，在 Internet 上的任何地方，都可以访问 “云” 中的任何资源，不需要关心资源在幕后是如何维护的。

Google、IBM、微软等云计算倡导者的产品涉及产业链的各个环节，从下面的数据中心到上面的应用，都是采用自己建设并提供服务的方式。基础设施层面的云计算是最底层的，也是最重要的，它给上层的应用及服务提供支撑。如果基础设施服务得不到保证，就无法给上层的应用及软件提供服务，从而影响到整个平台的服务质量。所以对于一个好的云平台来说，拥有一套独立的基础设施服务至关重要。

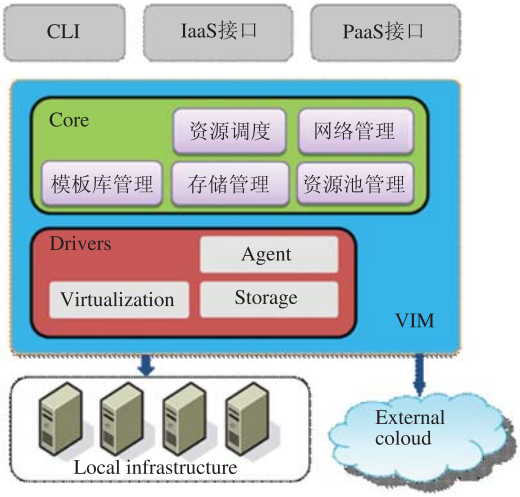

图8　CRANE个性化IaaS架构图

5.2　系统结构及特色

基础设施平台（如图8）采用虚拟化技术将物理资源转化为弹性的虚拟基础设施，向上层应用提供支撑，并面向用户提供基础设施服务。平台采用模块化的体系结构，向用户提供丰富的开发接口，使平台的扩展变得更加容易。本平台的主要特色有：同时支持多种虚拟化技术如xen、kvm等；动态增加基础设施如物理机、存储到系统；平台模块化设计，便于扩展；服务器后台自动整合。

首先，IaaS提供多种驱动（Driver），以支持系统在数据中心大量服务器上的部署。虚拟平台驱动屏蔽多种虚拟化技术的异构性，使得上层应用及服务能够在Xen，VMware，Kvm等不同虚拟化技术支持下正常运行，并对不同虚拟化技术的特性提供对应的增强功能。存储系统驱动使得云计算平台能够在本地存储系统或云存储服务提供商提供的存储上进行有效地存储管理，以增强扩展性及满足不同用户的需求，例如NFS，CSP等。该模块实现云存储中多个存储设备之间的协同工作，使多个的存储设备可以对外提供同一种服务，并提供更大更强更好的数据访问性能。Agent驱动将agent放于即将启动的虚拟机内部，在服务节点进入系统时，通过Agent进行信息采集、远程控制、服务器配置等工作；在系统运行过程中，维持服务器及虚拟机中有关模块功能的有效工作，并为故障节点提供恢复功能。

其次，IaaS实现了网络管理、存储管理、资源调度、跨域协作、模板库管理等多个系统核心功能，以保证系统的稳定运行以及其可扩展性。网络管理对整个系统的网络，包括物理网络、虚拟网络等进行统一管理，并提供基于IPV6的ip资源池管理。为动态增减物理资源、虚拟资源调度等系统功能提供支持。存储管理对系统所有存储资源进行统一管理，包括公共存储、私有存储等。任何一个授权用户都可以通过标准的公用应用接口来登录云存储系统，享受云存储服务。同时，存储也用于存储镜像模板。模板库管理为IaaS提供统一的模板库管理功能。将模板分为公共模板和私有模板，公共模板存放于公共存储区，私有模板存放于用户私有存储区，进行隔离管理，以保证用户数据的安全。同时为用户的模板定制功能提供支持。

对于云平台来说，不可缺少的一个模块就是云平台调度器。作为IaaS的主要特色之一，本系统将开发出一个独立于特定的云平台的调度器（如图9）。它可以动态配置，并按照相应的接口复用于其它云平

台。调度器接受用户对虚拟资源的使用请求，进行排队，综合参考请求优先级及系统负载等多种因素，向用户提供满足需求的资源。同时排队策略也像平台使用者开放第三方接口，使得能够定制合适的排队策略。

同时，调度器在系统底层进行对用户透明的资源调度、服务器整合等工作，以提高系统资源利用率、降低能耗。本平台中，所有作业都运行于虚拟机之中，资源调度主要针对虚拟机级别的资源进行调整，且调度是基于cpu及内存使用率的二维调度，主要包括虚拟机放置，虚拟机迁移，动态改变内存、vcpu数等方式。同时提供第三方调度策略接口，使得平台用户能够定制适合自身的调度策略。

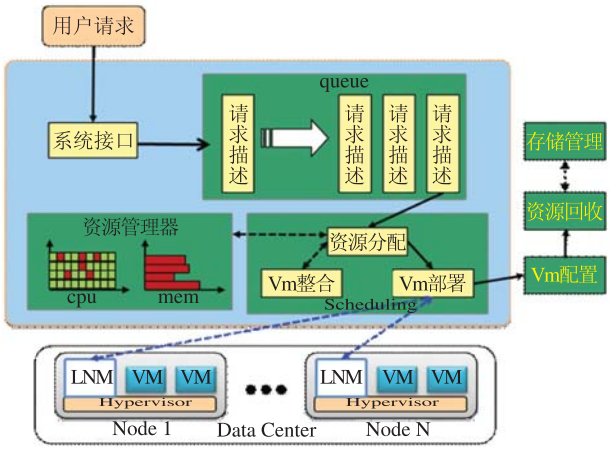

图9 CRANE平台IaaS调度器结构图

6 云监控系统

6.1 背景及意义

云平台往往需要管理较大规模的基础设施，因此监控系统必不可少。目前一些流行的开源云计算平台如Eucalyptus、OpenNebula等仅提供有限的监控功能，主要提供可用性方面的信息或为调度提供必要信息。然而作为以提供服务为目的的管理平台，一组详细的监控数据不仅可以帮助管理者和用户了解系统是否可用，更可以呈现系统及服务运行状态，为QoS的保障提供基本信息，辅助问题排查及产能规划。这对于系统管理者和系统用户双方都有着重要的意义。如系统管理者可以通过详细监控信息更快速地发现和修复QoS问题，一个使用云平台资源运营特定业务的用户可以根据监控数据更好地在业务增长时应对资源部署规划的需求。

大规模系统的监控是一个传统问题，在该领域已经有许多成熟的商业或开源解决方案。Ganglia是传统集群监控软件的代表，主要用于收集系统级的信息如CPU利用率、磁盘空间等。Ganglia由于其良好的性能和简单直观的特点受到广泛欢迎。Nagios是一款网络主机监控软件，能监视本地或远程主机以及服务，同时提供异常通知。相比Ganglia，Nagios的特点在于不仅提供主机资源的监控，还能监控网络服务状态（SMTP, POP3, HTTP, NNTP, ICMP, SNMP, FTP, SSH），并且可以通过其插件系统轻松地进行扩展。然而这些传统监控软件不能很好地适应云计算平台监控的需要。传统解决方案不是针对虚拟化数据中心设计的。在虚拟化数据中心监控的对象从过去的资源层、应用层两个层次变为物理资源层、虚拟资源层到应用层三个层次。由于虚拟化技术的引入，系统状态信息的收集方法和数据的组织方式都需要进行相应的改变。例如在Xen/linux平台主机上，使用传统方法在domain0收集到的数据不能准确反映整个物理主机的负载状况。另一方面，在虚拟化技术的支撑下，虚拟主机可以方便地被创建、销毁甚至伸缩、迁移。云计算之所以应用虚拟化技术，正是利用这一优势实现资源的高效、弹性配置以提高资源利用率和降低成本。但这同时也带来了监控的困难，因为虚拟机这一类监控对象会动态地增加、减少和改变配置和宿主机，监控系统需要知道虚拟机何时启动和停止，跟踪虚拟机的迁移。简单地说，监控系统需要知道任意时刻物理资源和虚拟资源的对应关系，而这一对应关系在云平台中是由资源调度管理模块维护的。所以云平台的监控系统必须能够从资源管理层获得调度信息。

6.2 系统结构及特色

云平台监控系统的总体架构图如图10所示。整个监控系统采用了监控代理收集信息，中心服务器处理存储数据的架构。

在数据收集方面，本系统使用部署到监控对象（物理机或虚拟机）上的监控代理程序agent收集需要的各类状态信息。通过将agent部署到虚拟机内部，实现gray-box式的监控，可以获得更详细的状态信息，并可以进一步实现对应用的监控。我们在设计上确保agent的轻量化、可配置性和可扩展性。一些分布式系统的监控方案将监控数据在被监控主机本地存储，而在我们的方案中agent只负责收集数据和发送数据到中心服务器。收集具体指标数据的程序被定义为模块，经过简单配置agent可以与不同的模块装配，从而对应不同的监控对象。另外，agent的数据收集模块易于扩展，可以方便地增加新的监控指标。

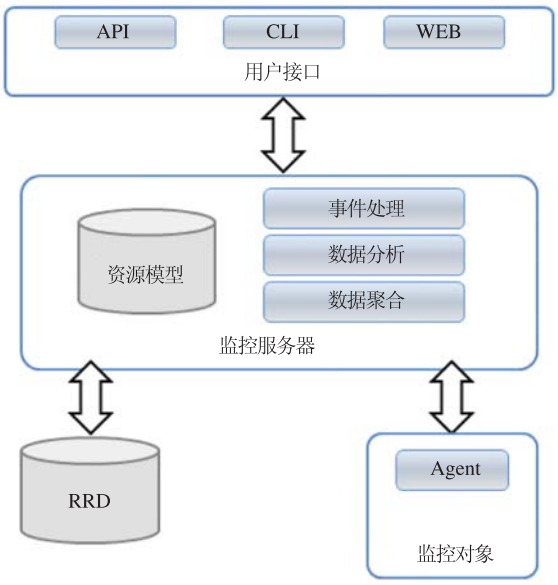

图10　CRANE云监控系统架构

本系统使用一个监控服务器作为主控模块。监控服务器维护着系统资源模型（资源的属性及相互关系）。资源模型相当于资源的元数据，记录了各个资源部件的各类属性及资源部件之间的相互关系，监控服务器与IaaS调度器交互实时更新该模型。通过维护资源模型信息，解决了虚拟资源动态变化给监控带来的难题。本系统在数据收集和呈现之外加入了数据分析和事件管理系统，根据预定义的过滤规则发现异常并触发相应事件。事件处理模块执行被触发事件的处理过程（如发出通知）。

本系统使用开源数据存储工具RRDTool存储监控数据。相比其他数据库方案，RRDTool将数据存储在文件，读写更加高效，且自动循环存储空间，不会出现数据存储不断增大的情况。

如何呈现数据直接决定了监控系统的易用性。本监控系统提供三类用户接口：基于Web Service标准的API，命令行接口以及Web图形化接口。许多云计算用户将部分业务部署到云端，以一种“混合”的模式使用云计算资源，他们需要将本地资源与云端资源置于统一视图下管理，API接口便于将监控组件集成到用户自己的管理工具中。命令行工具方便高级用户快速地查看、下载监控信息。基于Web的图形化接口是本系统主要特色之一。系统针对不同的用户角色提供不同的信息可见度：管理员用户可以查看全部监控信息，而普通用户只能获得自己使用的资源（虚拟机）或服务的信息。为了便于管理者理解资源和应用的关系，本系统提供三类视图：物理机——虚拟机视图、用户——虚拟机视图、应用——虚拟机视图，将物理资源、虚拟资源和应用三个层次联系起来。

7　用户管理及多域支持

CRANE实现了share-nothing架构，通过分布式任务队列实现了多计算集群的管理。采用LDAP集中管理用户信息，减少管理成本，增强安全性，并提高数据的一致性。CRANE平台中的各主要部件均能够独立地运行于多台服务器，各部件间通过消息队列实现异步通信，因此，CRANE具有较好的横向扩展性，如图11所示。

API Server向用户或第三方程序提供云计算服务调用接口。API Server接受HTTP请求，解析请求，验证用户身份，检查用户授权；将授权用户的合法请求转译为CRANE平台命令，转发至CRANE Controller。

CRANE Controller是平台的中枢，负责系统全局状态管理、分发与调度请求、与User Manager交互等。包含一个 AMQP 消息队列，CRANE Controller的各项功能均通过消息队列实现，有效的防止各组件的间的阻塞。当用户发出一个请求后，API Server会把这个请求放入任务队列中，然后把任务队列中的任务传递到消息队列服务器中，Cluster Manager从消息队列服务器发出来的消息提取出任务，然后任务由Cluster Manager进行处理，结果处理完后把结果和任务状态存储在Cache/Persistent Store,供用户查询。

在处理用户请求的过程中采用任务队列和消息队列。消息队列采用AMQP，一个高级消息队列协议，是应用层协议的一个开放标准，为面向消息的中间件设计。RabbitMQ 是一个实现了AMQP协议的消息服务器，它由分布式、高可靠性、并发和异步等特征。支持消息持久化和崩溃恢复，如果RabbitMQ崩溃了，消息并不会丢失，重新启动RabbitMQ服务后，队列和消息都可以恢复。此外，它可以和Python无缝结合。如果采用其他的消息队列，则不支持消息持久化，当消息队列宕掉的时候，队列中的消息则丢失，无法恢复。这样用户的请求也就会得不到相应。

任务队列采用celery，celery是一个基于分布式消息传递的异步任务队列。因为使用RabbitMQ的消息队列，具有消息持久化和崩溃恢复的功能，因此不用担心因为celery出现错误而照成任务的丢失；它是实

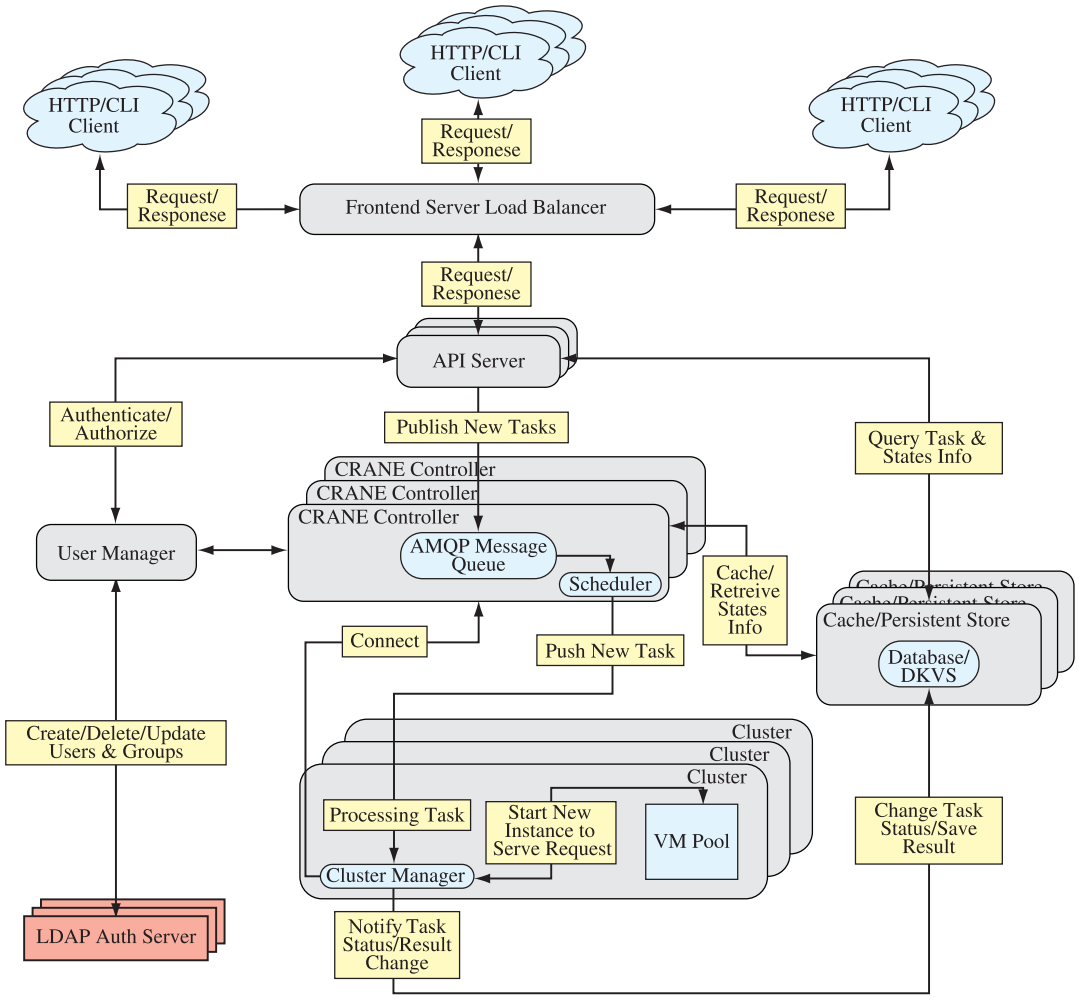

图11 CRANE任务队列及用户管理示意图

时的，也支持调度，可以在一个指定的时间运行制定的任务；支持并发，可以使用多处理模块并发执行任务，同时任务的执行结果可以存储在制定位置，也可以作为HTTP回调，实现跨语言的通信。也支持速率限制，可以使用令牌桶算法限制每个任务的速率，这样可以避免因为单个任务占用很多网络资源而照成系统响应低的影响。

采用这种任务队列和消息队列的方式，可以使多个Cluster Manager同时处理多个用户的请求，因为用户的请求会作为任务加入任务队列中，然后传给消息队列进行传递到集群后台进行并行处理并把返回的结果回调给用户。而任务和消息的处理都是异步并行执行的，因此可以对多个用户的请求有较好的响应。

User Manager实现了用户认证与授权。User Manager 通过使用轻量级目录访问协议（LDAP）构建集中的身份验证系统。User Manager 的 后端是一个LDAP 服务器集群，该LDAP集群支持自动同步，因此使用 LDAP 可以减少管理成本、增强安全性、避免数据复制的问题，并提高数据的一致性。

分布式的 Cache/Persistent 维护 CRANE 的全局系统状态信息、用户会话信息、任务状态与结果以及共享缓存等。

由图11可知，通过横向扩展（增加CRANE各组件的部署数量），CRANE可以支持管理多个计算集群，同时share-nothing 的系统架构和异步消息处理较好的保证了系统在规模扩大之后不会出现较大的性能下降。

8 未来发展规划

CRANE云平台目前只推出了一个测试版本，部署在CNGI华中科技大学节点上。对于CRANE平台整体而言，将进一步加强平台稳定性、可靠性及安全性，提高QoS，加强云互操作功能，增强多域统一用户管

理、权限控制功能等。同时，将提供完整API及完善的系统文档，以便系统的发布以及二次开发。

针对科学计算平台，目前虚拟集群内部通过以太网相连，对HPC任务来说，目前系统的1Gbs的以太网还难以满足要求。目前，虚拟化平台XEN还未能支持低延迟、高吞吐率的Infiniband网卡，使得infiniband这种适用于HPC的技术未能集成到CRANE之中。但是，据我们了解，Mellanox公司将在未来数月之内发布针对虚拟化平台XEN的Infiniband驱动，那时XEN上的虚拟机将可以直接使用Infiniband，这对提高HPC性能将有极大帮助。未来，Infiniband应用于XEN技术成熟之后，CRANE平台将使用Infiniband，以提高HPC性能。MapReduce已经成为一种越来越流行的大规模数据处理模式，未来可以尝试集成开源的MapRuduce实现hadoop到CRANE平台中。

针对企业信息化平台，目前仅支持PHP应用，由于很多企业级应用都是用Java语言实现的，因此下一步将增加对Java应用的支持。目前只是用单一的虚拟机作数据库服务器，这对数据库的安全性和扩展性都不利，未来将考虑采用分布式数据库。文件的存储目前采用的是NFS共享，这在数据读取量大时会导致效率低下，未来将考虑采用完全的分布式文件系统。目前的扩容策略只是单一地考虑内存利用率，未来将在实验和测试的基础上，进一步探讨各种负载（内存利用率、CPU利用率、网络数据传输率等）与QoS的关系，并区分出应用负载类型，如CPU密集型、带宽密集型、综合型等，从而制定更加合理的扩容策略。增强可配置性，同时降低模块耦合度，尝试将一些模块做得更加通用，将程序接口化与插件化，便于系统以后的扩展。

针对个性化IaaS，下一阶段的主要工作将放在独立调度模块的设计与实现以及存储管理上。 本系统将会做成一个可同时提供千台虚拟机创建请求，为用户提供基础设施服务，并可轻松地扩展以及收缩计算能力的平台。

针对云监控系统，目前已完成了监控代理agent，监控服务的数据聚合和数据存储，以及一个基于Web的图形化界面。下一步的主要工作是开发事件管理系统，并在对监控数据进行初步分析的基础上预定义一批事件，帮助用户及时发现和诊断异常。此外将进一步完善用户接口，以改善用户体验。