

云计算的演进和挑战性研究问题

富丽贞 陆嘉恒 孟小峰
中国人民大学

摘要: 本文首先介绍了云计算的演进过程,接着探讨了随着云计算的到来将会出现的新的挑战性问题,最后简单分析了引领云计算潮流的 Google 和 IBM 这两家公司的云计算基础架构框架。

经济的发展导致提供软件和计算能力服务基础架构的出现,俗称云服务或云计算。它是一种新兴的共享基础架构的方法,利用它用户可以在任何地方通过连接的设备访问其应用程序。可以将巨大的系统池连接在一起以提供各种IT服务。很多因素推动了对这类环境的需求,其中包括连接设备、实时数据流、SOA的采用以及搜索、开放协作、社会网络和移动商务等这样的Web 2.0应用的急剧增长。另外,数字元器件性能的提升也使IT环境的规模大幅度提高,从而进一步加强对一个由统一的云进行管理的需求。毫无疑问,云计算已经拥有了一个光明的前景。

云计算的演进

目前,云计算是 IT 行业的一个热点话题。但它并不是革命性的新发展,而是数据管理技术不断演进的结果,如图 1 所示。



图 1 云计算的演进

在上个世纪末,分布式处理(Distributed Computing)、并行处理(Parallel Computing)和网格计算(Grid Computing)已相当成熟。他们是云计算发展的技术基础。

上世纪 80 年代末,开始出现应用大量系统来解决单一问题(通常是科学问题)的情况,这就是网格计算的概念,而这种概念又导致向云计算的发展。网格计算的关注重点是将工作负载移到所需的计算资源所在位置的能力,大多数情况下这种位置都是远程的,而且持续可用。通常,网格是服务器集群,大型任务可拆分为多个小型任务,以便在这些服务器上并行运行。从这个角度来看,我们实际上可将网格视为仅仅是一台虚拟服务器。网格还要求应用程序符合网格软件的接口标准。

公共计算和 SAAS(软件即服务)可以看作是早期云计算的两种形式。现在云计算不只包括这两种形式,还包括网络服务、平台即服务以及 MSP(管理服务提供商)等其他形式。

到了上世纪 90 年代,虚拟化的概念已从虚拟服务器扩展到更高层次的抽象,首先是虚拟平台,而后又是虚拟应用程序。公用计算将集群作为虚拟平台,采用可计量的业务模型进行计算。最近,SAAS(软件即服务)将虚拟化提升到了应用程序的层次,它所使用的业务模型不是按消耗的资源收费,而是根据向订户提供的应用程序的价值收费。这种类型服务通过浏览器把程序传给成千上万的用户。在用户眼中看来,这样会省去在服务器和软件授权上

的开支；从供应商角度来看，这样只需要维持一个程序就够了，这样能够减少成本。**Salesforce.com** 是迄今为止这类服务最为出名的公司。**SAAS** 在人力资源管理程序比较常用。**Google Apps** 和 **Zoho Office** 也是类似的服务。

云计算的概念源于公用计算和 **SAAS** 概念。“云”的优势在于其基础架构管理，虚拟化技术的日益成熟和不断进步为这种管理提供了强大的支持，使“云”能够通过自动部署、重新构建映像、重新均衡工作负载、监控并系统地处理变更请求，以便管理并更好地利用底层资源。

云计算面临的挑战

作为一项有望大幅降低成本的新兴技术，云计算正日益受到一系列众多公司的追捧。但是同时也随之产生了一系列新的挑战性问题。

首先，云计算中一个跨领域问题就是供应商要在功能和开发代价上作权衡。目前，早期的云计算提供的 **API** 比传统的数据库系统的限制多得多。他们只提供一个极小化的查询语言和有限的一致性保证。这给开发带来更多的编程负担，但是允许服务供应商提供更多的预期服务和级别协议，这对于一个功能完备的 **SQL** 数据库也是很难达到的。在现有的云计算基础上，为了实现只做较少改动而使其功能更完备，我们需要更多的经验和做更多的工作。

其次，易管理性在云计算中极其重要，这也带来新的挑战。和传统的系统相比，受有限的人工干涉、工作负载变化幅度和多种多样的共享设备这三个因素的影响，云计算中管理更加复杂。大多数情况下，没有协助基于云的应用开发的数据库管理员和系统管理员。甚至是单一用户的负载随时间都会发生大幅度的变化。对于一个偶尔会用到比平常高出几个数量级的资源的客户来说，云计算的可伸缩供应是经济的。本来混合负载就很难调优，但在这种情况下调优是不可避免的。同时，服务调优主要依赖共享设备的共享方式。例如 **Amazon** 公司的 **EC2** 用硬件级别上的虚拟机作为编程的接口。而 **salesforce.com** 公司则在一个数据库系统上实现了具有多种独立模式的“多租户”虚拟机。其他的虚拟解决方案也是可行的。在负载之上平台之下，每一种方案都有不同的可见性和不同的控制彼此的能力。这些变化需要我们重新考虑跨层资源管理的传统角色和职责。

上世纪 90 年代末，研究学者们开始研究自我管理技术。对易管理性的需求加速了这一技术的发展。云计算系统需要自适应的在线技术，反过来系统中新的架构和 **API**（包括区别与传统 **SQL** 语言和事务语义的灵活性）又促进了颠覆性的自适应方法的发展。

接着，云计算的庞大规模同样带来了新的挑战。现有的 **SQL** 数据库不能简单地处理放置在云中的成千上万的数据。在存储方面，是用不同的事务实现技术，还是用不同的存储技术，或者二者都用来解决一些限制性问题还不确定。在这个问题上，目前在数据库领域内有很多提议。现有的云计算已经开始探索一些简单的实用性方法，但是我们仍需要做更多的工作来融合现有的云计算机制文化中的好思想。就查询处理和优化而言，如果搜索一个涉及到数千条处理的计划空间需要花费很长时间，那么这是不可行的，所以需要在计划空间或搜索上设限。最后如何在云环境中编程还尚不清楚。我们需要更多的了解云计算的现实问题（包括性能限制和应用需求）来帮助设计。

此外，在云基础架构中，物理资源共享带来新的数据安全和隐私危机。他们不能再依靠机器或网络的物理边界得到保障。因此云计算为合成和加速这方面现有的工作提供了丰富的机遇。要想成功关键在于我们能否准确瞄准云的应用场景以及能否准确把握服务供应商和顾客的实际动向。

最后，随着云计算越来越流行，预计会有新的应用场景出现，也会带来新的挑战。例如，我们预测会出现一些需要预载大量数据集（像股票价格、天气历史数据以及网上检索等）的特殊服务。从私有和公共环境中获取有用信息引起人们越来越多的注意。这样就产生新的

问题：我们需要从结构化、半结构化或非结构的异构数据中提取出有用信息。同时，这也表明跨“云”服务必然会出现。在科学数据网格计算中，这个问题已经很普及。即便在一门学科中，也会需要大量位于不同地理位置的共享数据服务器。大体上，在大多数企业亦是如此。而联合云架构不会降低只会增加问题的难度。

云计算实例分析

本文最后对引领云计算潮流的两大公司 google 和 IBM 的云计算平台做一个简单的分析。

短时间内Google的在云计算上的地位依然不可撼动，其开放式的平台体现了云计算模式的精髓。Google的云计算服务所需要的绝大部分基础软件都是开源的，这意味着用户可以自由的得到那些代码并修改。从 2003 年开始，Google连续几年在计算机系统研究领域的最顶级会议与杂志上发表论文，揭示其内部的分布式数据处理方法，向外界展示其使用的云计算核心技术。Google的云计算技术实际上是针对Google特定的网络应用程序而定制的。针对内部网络数据规模超大的特点，Google提出了一整套基于分布式并行集群方式的基础架构，利用软件的能力来处理集群中经常发生的节点失效问题。Google使用的云计算基础架构模式包括四个相互独立又紧密结合在一起的系统。包括Google建立在集群之上的文件系统Google File System，针对Google应用程序的特点提出的Map/Reduce模式(映射/化简编程模式)，分布式的锁机制Chubby以及Google开发的模型简化的大规模分布式数据库BigTable。

GFS (Google文件系统)是为Google应用程序本身而设计的。一个GFS集群包含一个主服务器和多个块服务器，可以被多个客户端访问。

为了让不熟悉分布式系统人们能够有机会将应用程序建立在大规模的集群基础之上，Google 还设计并实现了一套大规模数据处理的编程规范 Map/Reduce 系统。这样，非分布式专业的程序编写人员也能够为大规模的集群编写应用程序而不用去顾虑集群的可靠性、可扩展性等问题。应用程序编写人员只需要将精力放在应用程序本身，而关于集群的处理问题则交由平台来处理。Map/Reduce 通过把对数据集的大规模操作分发给网络上的每个节点实现可靠性；每个节点会周期性的把完成的工作和状态的更新报告回来。如果一个节点保持沉默超过一个预设的时间间隔，主节点（类同 Google File System 中的主服务器）记录下这个节点状态为死亡，并把分配给这个节点的数据发到别的节点。每个操作使用命名文件的原子操作以确保不会发生并行线程间的冲突；当文件被改名的时候，系统可能会把他们复制到任务名以外的另一个名字上去。

第三个云计算平台就是 Google 关于将数据库系统扩展到分布式平台上的 BigTable 系统。为了处理 Google 内部大量的格式化以及半格式化数据，Google 构建了弱一致性要求的大规模数据库系统 BigTable。除了这三个部分之外，Google 还建立了分布式程序的调度器，分布式的锁服务等一系列相关的云计算服务平台。

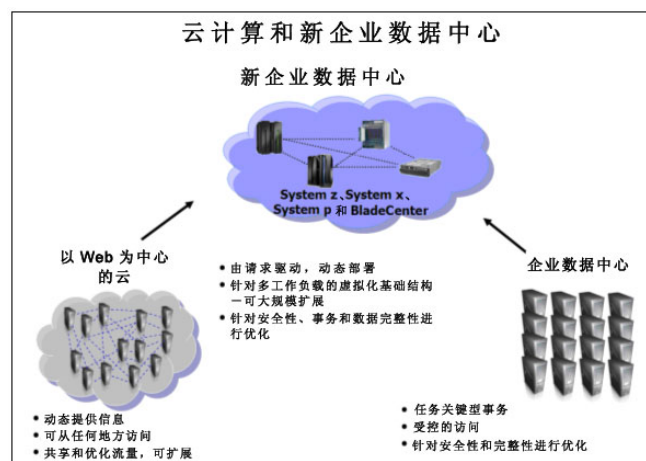


图 2. 云计算和新企业数据中心

蓝色巨人IBM对此也投下了重注，并为此命名为“蓝云”计划。IBM具有发展云计算业务的一切有利因素：应用服务器、存储、管理软件、中间件等等，因此IBM自然不会放过这样一个成名机会。最近又推出了“新企业数据中心”的设想，该设想结合了以 Web 为中心的云计算模型和当前的企业数据中心的优势。中国云计算网的一篇文章给出了“新企业数据中心”模型（如图 2 所示）以及其基础架构服务框架（如图 3 所示）。

新企业数据中心将是虚拟化、高效管理的中心，它将使用以 Web 为中心的云所采用的某些工具和技术，并进行一般化以便被范围更广的客户采用，另外还进行增强以支持安全的事务性工作负载。通过高效且共享的基础架构，企业能够对新的业务需求迅速做出反应，实时解析大量信息，而且还能根据实时数据做出明智的业务决策。新企业数据中心是一种演进的新模型，能提供有助于使 IT 和业务目标保持一致的高效且动态的新方法。如图 3 所示，从高级别的架构角度来看，新企业数据中心的基础架构服务在逻辑上可分为不同的层次。物理硬件层已虚拟化，以便能提供灵活且适应性强的平台，从而提高资源利用率。接下来的两层是虚拟化环境层和管理层，它们是新企业数据中心基础架构服务的关键。通过把这两层结合起来，可以确保数据中心内的资源得到有效的管理，并可以快速部署和配置。另外，新企业数据中心旨在处理混合模式的工作负载。

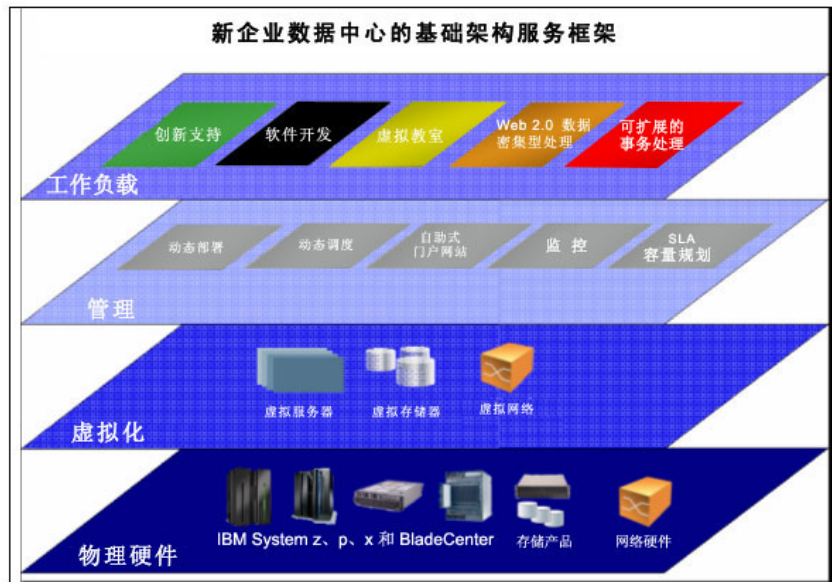


图 3. 新企业数据中心的基础架构服务框架

参考文献：

- [1] The Claremont Database Research meeting : <http://db.cs.berkeley.edu/claremont/>
- [2] IBM: <http://www.ibm.com/developerworks/websphere/zones/hipods/>
- [3] 计世网: <http://www.ccw.com.cn>
- [4] 中国云计算网: <http://www.cloudcomputing-china.cn>