

# 云计算的演进和挑战性问题

富丽贞 陆嘉恒

云计算是当前业界最火的词汇之一，它与很多或时髦或老牌的技术词汇联系到一起，比如虚拟化、网络、SaaS。云计算是一种新兴的共享基础架构的方法，最终将成为一种普及性服务。

利用这种提供软件服务和计算能力服务的基础架构，各服务提供商将巨大的系统池连接在一起提供各种IT 服务，用户可以在任何地方通过连接的设备访问应用程序。很多因素共同推动了云计算发展，包括连接设备、实时数据流、SOA采用，以及搜索、开放协作、社会网络和移动商务等Web 2.0应用急剧增长。硬件的性能攀升也使IT基础架构的规模大幅度提高，从而使提供云计算服务成为了可能。

## 云计算阶段性演进

云计算其实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。在上世纪末，分布式处理、并行处理和网格计算就已相当成熟，它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，这一概念又与云计算的关系最为密切。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只是一台虚拟服务器。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提供服务的两种形式。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。



云计算的逐步演进过程

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。开始流行的SaaS将虚拟化提升到了应用程序的层次，它使用的业务模型不是按消耗的资源收费，而是根据向用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。在用户看来，这样会省去服务器的固定硬件投入和软件授权开支；从供应商角度来看，这样只要维持程序就够了，能减少成本，Salesforce.com是迄今为止这类服务最为出名的公司，Google Apps 和Zoho Office 也是服务商提供的类似服务。云计算的概念直接源自公用计算和SaaS概念。“云”的优势在于其基础

架构管理，日益成熟的虚拟技术为这种管理提供了强大的技术支持，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。

## 引发数据管理挑战

作为一项有望大幅降低成本的新兴技术，云计算被认为是大势所趋，正日益受到众多企业的云计算服务提供商的追捧。与此同时，也随之产生了一系列新的挑战性问题，比如“云”之间的互联等等，对数据管理的挑战首当其冲。

云计算遭遇的一个难题是服务提供商要在功能和开发代价上做权衡。目前，早期的云计算提供的API比传统数据库系统的限制要多得多，只提供一个极小化的查询语言和有限的一致性保证。这给开发带来了编程负担。允许服务提供商提供更多的预期服务和更高级别协议，对于一个功能完备的SQL数据库来说也是很难达到的。在现有的云计算基础上，为了实现只做较少改动就使其功能更完备，需要业界积累更多经验，做更多探索。

易管理性在云计算中也极其重要。它带来的挑战在于，与传统系统相比，受有限的人工干涉、工作负载变化幅度大、多种多样的共享设备这三个因素的影响，云计算环境的管理更加复杂。在大多数情况下，没有基于云应用开发的数据库管理员和系统管理员，负载经常变化，甚至单一用户的负载随时间都会发生大幅度变化。

对于一个偶尔会用到比平常高出几个数量级资源的用户来说，云计算的可伸缩供应是经济的，在这种情况下调优是不可避免的。服务调优主要依赖共享设备的共享方式。例如Amazon的EC2用硬件级别上的虚拟机作为编程接口。而salesforce.com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。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又促进了颠覆性的自适应方法发展。

云计算的庞大规模也同样带来了新挑战。现有的SQL数据库不能简单地处理放置在云中的海量数据。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，但是仍需要做更多工作来融合现有的云计算机制中的优秀思想。就查询处理和优化而言，如果搜索一个涉及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。如何在云环境中编程也尚不清楚，业界需要更多地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，物理资源共享也带来了新的数据安全和隐私危机，安全很难再依靠机器或网络的物理边界得到保障。随着云计算越来越流行，预计会有新的应用场景出现，这也会引发一些新问题。例如，我们预测会出现一些需要预载大量数据集（如股票价格、天气历史数据以及网上检索等）的特殊服务。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，而联合云架构不会降低只会增大问题的难度。

## 先行者实例借鉴

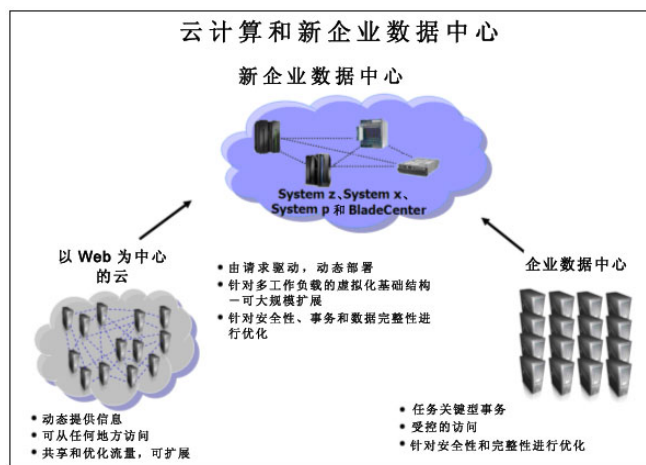
短时间内Google 在云计算上的地位依然难以撼动，其开放式平台体现了云计算模式的精髓。Google 的云计算服务需要的绝大部分基础软件都是开源的，这意味着用户可以自由得到那些代码并修改。从2003 年开始，Google 连续几年在顶级学术会议与杂志上发表论文，揭示其内部的分布式数据处理方法，向外界展示其使用的云计算核心技术。

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。

为了让不熟悉分布式系统的人们能有机会将应用程序建立在大规模集群的基础之上，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大规模集群编写应用程序，而不用顾虑集群的可靠性、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。

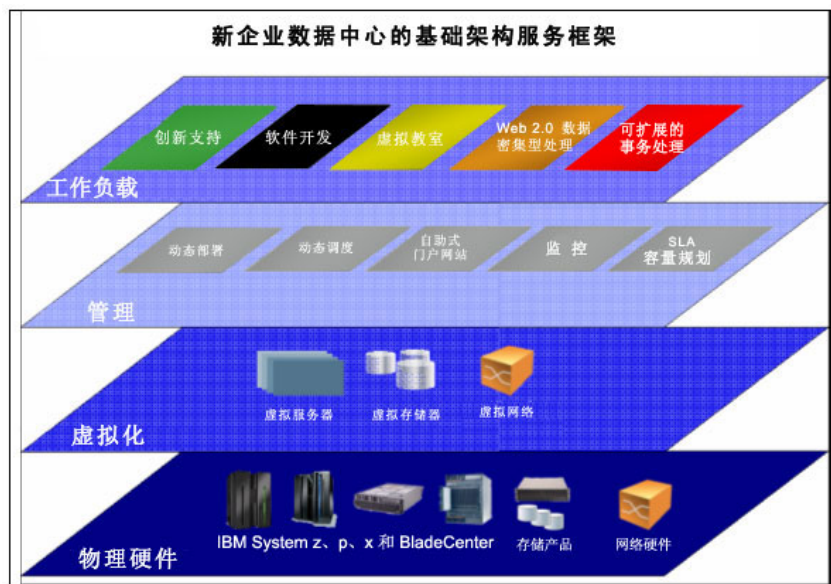
Map/Reduce 通过把对数据集的大规模操作分发给网络上的每个节点实现可靠性，每个节点会周期性地完成的工作和状态的更新报告回来。如果一个节点保持沉默超过一个预设的时间间隔，主节点（类同Google File System 中的主服务器）记录这个节点状态为死亡，并把分配给这个节点的数据发到别的节点。每个操作使用命名文件的原子操作以确保不会发生并行线程间的冲突；当文件被改名的时候，系统可能会把它们复制到任务名以外的另一个名字上去另一个重要的云计算平台实践就是Google 关于将数据库系统扩展到分布式平台上的BigTable系统。为了处理Google 内部大量的格式化以及半格式化数据，Google 构建了弱一致性要求的大规模数据库系统BigTable。除了以上技术之外，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。

IBM 也是云计算的另一个重要的推动者，并在内部将云计算命名为“蓝云”计划。IBM 具有发展云计算业务的一切有利因素，比如应用服务器、存储、管理软件、中间件等，因此，IBM 也拥有天然的技术优势。IBM 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。



云计算与新企业数据中心的的关系

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。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”采用的某些工具和技术，并进行了普及化处理，以便被范围更广的客户采用；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。



新企业数据中心的基础架构服务框架

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。物理硬件层已全面虚拟化，以便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。通过将这两层结合起来，可以确保数据中心内的资源得到有效管理，并可以快速部署和配置。