科技桥

科技桥栏目由本刊编辑部和清华大学科技开发部合办。其目的是推介清华大学和校友企业的科研成果，专利申报，报道院系科研团队、重点实验室和国际科技前沿动态，发布校企及校友企业新产品。

联系方式：《水木清华》编辑 010-62797884
科技开发部《科技桥》编辑 010-62785671
邮箱：smthkjq@tsinghua.org.cn 、kjq@tsinghua.edu.cn

项目推介

LoongStore 大规模集群云存储系统

清华大学信息技术研究院

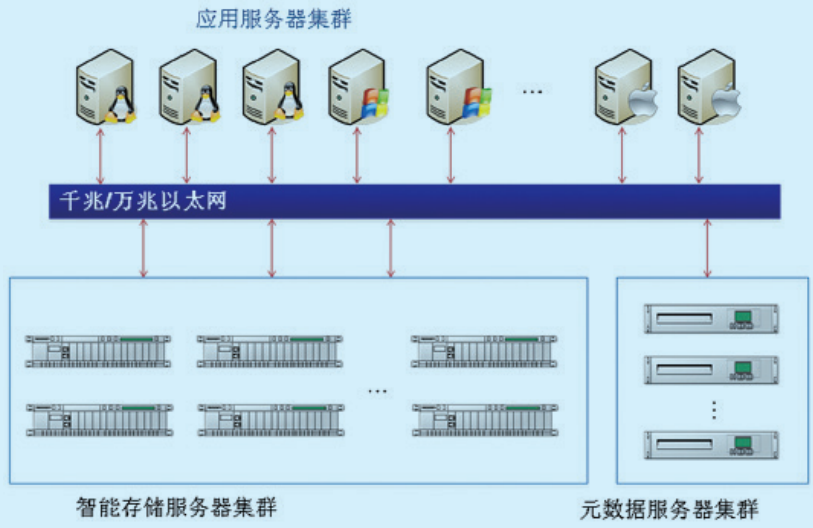
成果简介

网络存储技术从诞生到现在，已经经历了近二十年的历程。在这个过程中，无论是计算技术、网络通信技术还是信息储存技术都发生了翻天覆地的变化。计算技术领域，基于开放操作系统上的集群计算技术已经取代大型机；网络通信领域，开放架构的 TCP／IP 通信已经统治了一切；在信息存储技术上，单块磁盘的容量已经从 G 级别进入 T 级别。

传统的网络存储设备，无论是 FC SAN、NAS 还是 IP SAN，都是采用以满足单台服务器存储需要为目标的体系设计，在共享能力、性能、可靠性、扩展性以及总体拥有成本上，已经无法满足企业新的应用需求。如果仅仅通过外围手段来弥补，必将造成存储架构越来越复杂，用户管理难度越来越大，总体拥有成本越来越高。

为了满足 IT 技术发展的挑战，LoongStore 集群存储采用了代表计算技术、网络通信技术以及文件系统技术发展方向的体系架构，提供给用户跨平台共享、高性能、高可靠、可平滑扩展、使用和维护简单的高端存储系统。系统以开放架构平台智能存储服务器为基本节点，通过在千兆以太网基础上不受限制地添加节点，构成了业界最高的性能、可靠性以及极低总体拥有成本的存储集群。

LoongStore 为一种大规模的文件共享存储系统，它面向海量数据、高并发访问的环境。其产品架构图如图所示：



LoongStore 产品架构

LoongStore 产品架构支持：

- 单文件系统 100PB
- 支持 2000 个以上节点并发
- 支持超过 500 亿文件

产品主要特点：

- 产品架构上：实现了横向扩展。从原来的单节点增加 CPU、内存、存储等等的纵向扩展方案往增加节点数量的横向扩展方案的转变，使得容量扩展更加灵活、同时也实现了更低的起始投入成本。
- 产品工作流上：实现了更高效的数据读写方式。通过实现元数据与数据通道分离、将数据条带化存储在不同的存储服务器上，从而高效实现数据的并行读写。
- 产品性能上：实现了高聚合带宽。克服了以前文件集中式存储的瓶颈，实现了文件切片分布式存储方案，实现了高聚合带宽。
- 产品性能上：实现了性能近线性增长。如：
（1）聚合带宽随着存储集群规模扩大而线性增长；
（2）充分发挥硬件性能，如网卡性能利用率达到了 80%；
（3）扩展到 100GB/s 以上的聚合性能。
- 产品可靠性上：实现了无单点故障。通过
（1）为不同目录的数据设置不同的冗余度；
（2）自动故障探测与恢复；
（3）设备恢复速度是 RAID 的 5 倍，750GB SATA 盘恢复时间 15 分钟；
（4）元数据服务器每两台互备等，实现了产品的无单点故障。
- 在产品扩展性方面：实现了在线业务扩展容量。支持不停业务在线扩容，并自动迁移均衡数据，从而实现不需要关停业务的方式轻易扩展容量。
- 在产品管理方面：实现了单点配置管理。产品通过单点实现了所有的配置管理功能，通过内置各种自动化处理流程，减少人为参与。
- 产品兼容性方面：兼容各种 X86 架构的服务器和 IP 网络，应用服务器端支持 Linux/Windows/Mac 等等。

应用说明

LoongStore 大规模集群云存储系统能够帮助企业加速信息到价值的转换，目前已经在互联网、广电、能源、电信等数据密集型行业帮助企业实现了突破性的价值。产品主要应用行业与应用案例有：（1）互联网：视频，SNS 等，例如中国国家网络电视台（CNTV），人人网，搜狐；（2）能源：大庆油田，中石油勘探开发研究院总院；（3）动漫和广电：中影集团，教育电视台等，《建党伟业》、《建国大业》、《孔子》等大片的后期制作就是采用该存储系统；（4）教育；（5）军工，航天等等领域。

LoongStore 大规模集群云存储系统

技术指标

- 支持 PB 级的存储容量规模，最大容量可超过 100PB；
- 可提供几十甚至上百 GByte/s 的聚合数据读写带宽，而且性能可随存储规模的扩展而线性提升；
- 可高效支持海量小文件的存储和管理，单套系统可高效管理超过 500 亿个文件，现有在线系统已经超过百亿级文件数量规模，而且每天仍然以 2000~5000 万个新文件的速度递增，每天的下载量为 10~20 亿个文件；
- 可支持超过 2000 多个计算节点（或者应用节点）同时高效并发访问同一套存储系统；
- 提供完全符合 POSIX 标准的文件级接口，支持全局共享访问，无须加入第三方共享软件。前端可同时部署 Linux、Windows、Mac 等异构计算节点。所有应用可不经过任何修改和移植即可使用存储服务；
- 支持在线扩展存储容量，无需终止应用的正常服务，新增容量即插即用。同时提供在线数据迁移功能，扩容后可均衡所有设备上的存储负载；
- 内置数据高可用机制，一旦发现硬件故障将自动进行数据恢复，无需停止服务。除了能够冗余磁盘故障之外，即使整个存储服务器或者磁盘阵列故障都不影响应用正常服务，也不会破坏应用数据。传统存储只能冗余磁盘故障；
- 如果系统中正常设备上的空闲空间总和不少于故障设备上已经使用的空间即可进行数据恢复，无须加入新的设备，传统存储采用 RAID 机制必须加入新的存储设备才能进行恢复。确保了系统无间断持续运行；
- 在线控制存储设备的负载。在观察到某块存储设备有异常现象时，可以在线手动禁止这块磁盘的空间分配，设置后系统不再往被禁止分配空间的磁盘内写入新的数据，原有的数据仍然可用。当设备正常时即可在线放开；
- 提供主动数据迁移功能，方便硬件在线更换。当某些设备使用时间超过一定年限，或者是发现设备有安全隐患时，可以主动将该设备上的数据迁移到其它存储设备，迁移完成后即可将设备移除下线，不影响应用的正常服务；
- 简易管理，通过单一图形化界面即可管理和监控整个存储系统；
- 高硬件兼容性，既可以采用传统的 SAN 或者盘阵作为存储单元，也可以采用通用的存储服务器进行构建，硬件选型范围广，成本低，适合用于构建海量的云存储系统。

合作方式 商谈

所属行业领域 信息领域