计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
42(6):1028~1033, 2005

大容量、高性能、高扩展能力的蓝鲸分布式文件系统

杨德志[1,2] 黄 华[1,2] 张建刚[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)
(yangdz@ict.ac.cn)

BWFS: A Distributed File System with Large Capacity, High Throughput and High Scalability

Yang Dezhi[1,2], Huang Hua[1,2], Zhang Jiangang[1], and Xu Lu[1]
[1](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)
[2](*Graduate School of the Chinese Academy of Sciences, Beijing* 100039)

Abstract With the increasing requirements of applications and developments in computer science, research on networking storage system (NSS) becomes hot spot of I/O subsystem research. As one of the core components of NSS, distributed file systems should be paid much attention to.

Based on the study of the existing research results, BlueWhale File System (BWFS) was designed by NRCHPC, the Institute of Computing technology, the Chinese Academy of Sciences. And it enables large capacity, high throughput and high scalability of BW1K NSS. In this paper, we described architecture of BWFS and its major characteristics are described and the test results of BW1K NSS are used to verify these characteristics.

Key words networking storage system; distributed file system; large capacity; high throughput; high scalability

摘 要 应用需求和计算机技术的发展使网络化存储系统成为网络服务器系统中 I/O 子系统研究的热点. 作为网络存储系统关键部件,分布式文件系统的研究具有非常重要的意义.

蓝鲸分布式文件系统(BWFS)是国家高性能计算机工程技术研究中心基于对国内外现有研究成果的分析和研究,自主设计实现的分布式文件系统. 它着重于大容量、高 I/O 吞吐率和高扩展能力等方面特性. BWFS 已经用到 BW1K 网络存储系统中,并通过 BW1K 的初步评测数据验证了这些特性.

关键词 网络存储系统;分布式文件系统;大容量;高吞吐率;高可扩展能力

中图法分类号 TP333;TP316.4

1 引 言

经济和技术的发展使高性能计算、商业计算、大规模数据处理等技术得到广泛的应用. 不断增加的应用需求对存储系统不断提出新的要求[1].

BWFS 是国家高性能计算机工程技术研究中心(NRCHPC)自主设计的用于海量网络存储系统 BW1K 的分布式文件系统. 它采用专用服务器模式,将文件访问的数据流与控制流有效分离,为系统

收稿日期:2005-01-12
基金项目:国家"八六三"高技术研究发展计划基金项目(2002AA112010)

客户提供高吞吐率和高扩展能力的数据访问.它采用灵活有效的机制管理系统存储资源,并支持存储资源的扩展.BWFS适合于构建大容量、高I/O吞吐率和高扩展能力的网络存储系统,能够满足高性能计算、核模拟、系统仿真计算、VoD、E-Mail等多种应用的需求.

本文在第2节介绍相关研究的情况;第3节介绍BWFS的系统服务器结构、存储资源管理机制、元数据管理机制和文件数据访问机制等;第4节通过对BW1K网络存储系统的初步评测,验证BWFS具有的高吞吐率和高可扩展性等特性.最后是总结及下一步的工作概况.

2 相关研究

经过20多年的研究,分布式文件系统的服务器结构发展到专有服务器系统占据主流的阶段.专有服务器系统采用专门的服务器提供文件系统元数据管理,专门的服务器提供文件数据存储服务,系统的应用服务器仅运行系统的各种应用服务.其典型代表有SliceFS/Duke[2],StorageTank/IBM[3]以及Lustre/ClusterFS[4]等等.

对于存储空间的管理,SliceFS采用共享块设备模型;Lustre采用对象存储技术[5];StorageTank则通过其虚拟存储技术管理磁盘阵列[6].对于元数据管理,SliceFS采用多个元数据服务器,并由位于应用服务器和元数据服务器集群间的μproxy提供元数据服务器集群的扩展能力,μproxy通过静态哈希函数和元数据位置固定的策略管理单个元数据的分布,响应并转发应用服务器所有的元数据请求;StorageTank采用多个元数据服务器构成的元数据集群处理元数据请求,它以一组元数据为单位,完成元数据映射、元数据迁移等工作;Lustre采用两个元数据服务器以"active-standby"方式,提供可用的元数据服务.

3 蓝鲸分布式文件系统(BWFS)

3.1 BWFS系统组成

BWFS采用专用服务器结构,所有服务器直接与高速互联网络连接.专门的元数据服务器集群负责文件系统元数据管理,专门的网络存储设备负责提供文件数据存储服务.应用服务器通过BWFS元数据访问协议,直接向网络存储设备进行文件数据的读写.与SliceFS不同的是,BWFS采用了一个位于元数据服务器集群后端的绑定服务器[7]完成元数据服务器的协同管理.

BWFS的系统结构如图1所示.它包括:①应用服务器(application server,AS)运行Web服务,Video服务,高性能计算等应用服务;②元数据服务器(metadata server,MS)处理元数据请求.多个元数据服务器通过绑定服务器协同,完成BWFS名字空间管理,为AS提供文件及文件系统的元数据服务;③绑定服务器(binding server,BS)主要完成文件系统元数据在MS集群中的分布决策.在后面的描述中,将元数据服务器和绑定服务器统称为系统服务器;④网络存储设备(storage node,SN)为BWFS提供存储服务.BWFS支持多个网络存储设备,通过虚拟化技术,将多个网络存储设备物理地址虚拟成统一的逻辑块设备地址空间;⑤管理服务器(administrator,AD)负责BWFS和服务器状态的管理;⑥高速互联网络将系统的应用服务器、系统服务器、网络存储设备以及管理服务器连在一起,构成完整的应用系统.

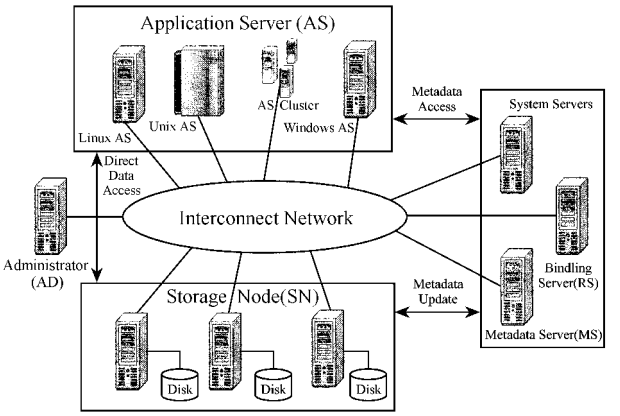

Fig. 1 Architecture of BWFS.

图1 BWFS系统结构图

3.2 BWFS资源管理机制

BWFS采用分布式分层资源管理模型进行系统资源的管理,如图2所示.

为提供单个存储设备物理资源的扩展能力,各个SN将自己管理的物理存储资源组织成一个统一的物理存储空间.然后,这个存储空间映射到一个使用64位块号的虚拟块设备地址空间,虚拟设备的块资源称为逻辑资源.系统的一个专门的服务负责逻辑资源的管理,支持SN的动态加入,完成新加入的SN的资源映射等工作,为系统资源的扩展提供支持.

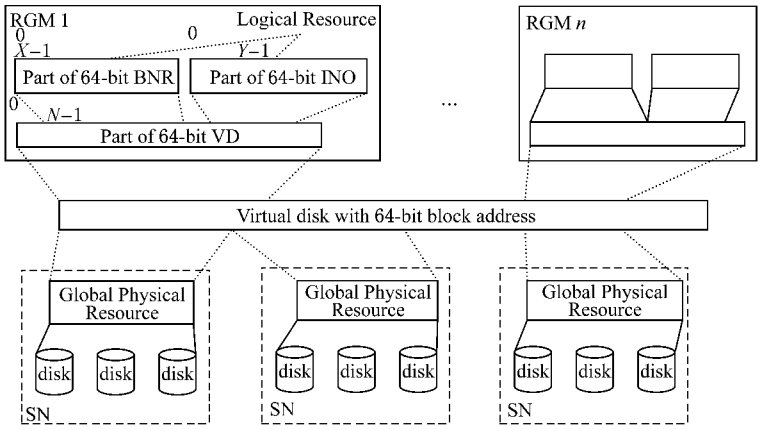

Fig. 2 BWFS resource management model.

图 2 BWFS 分层虚拟化资源管理模型

BWFS 从逻辑资源管理服务获得逻辑资源,并根据获得的资源属性将逻辑资源划分成资源组(resource group,RG),每个资源组由相应的资源组管理器(RGM)负责组织和管理 RG 的资源及其变化信息.

BWFS 采用 64 位的索引节点(inode)号和文件块,提供大规模的文件系统管理能力. 索引节点和文件块到设备块的映射工作由 RGM 完成.

3.3 BWFS 元数据管理机制

BWFS 采用传统 Unix 文件系统的树形结构和文件命名规则,为系统客户端提供全局文件系统名字空间(global file system namespace),每个用户能够以访问本地文件的方式访问 BWFS 中任何一个具有访问权限的文件.

为减少影响系统性能和错误恢复的因素,BWFS 采用"集中决策,分布处理"的原则管理文件系统的数据和元数据. BWFS 元数据管理采用 BS 集中决策元数据的分布,每个元数据服务器直接为 AS 提供元数据服务的机制.

BS 对 BWFS 活跃元数据进行单一映射,当前负责文件元数据更新的 MS 称为元数据宿主 MS. 活跃元数据与宿主 MS 是一一映射关系,映射关系建立过程被称为"绑定".

BWFS 的 AS 缓存文件元数据、文件块与设备块的映射关系以及元数据宿主服务器信息等等. 它采用基于超时和回叫(callback)相结合的机制维持元数据缓存一致性. 文件元数据更新通过 MS 完成. BS 的全局元数据分布表(GMDT)和 MS 上缓存元数据分布信息的局部元数据分布表(LMDT)共同完成元数据分布信息的管理. BS 的 GMDT 记录 BWFS 当前所有活跃元数据在元数据服务器集群中的分布. MS 的 LMDT 记录 MS 当前知道的活跃元数据的分布信息. 它是 GMDT 的部分信息的缓存. BWFS 当前的实现中,MS 根据分布信息的访问情况采用 LRU 机制管理 LMDT. 如果某个活跃元数据的分布信息因为缓存管理机制被释放掉,MS 与 BS 通信重新获得该分布信息. 只有在一个元数据不再活跃时,宿主 MS 才通知 BS 放弃对该元数据的映射管理.

当 AS 元数据缓存失效时,它按照如下步骤进行:①它根据自己缓存的元数据宿主服务器信息,向可能完成请求处理的 MS 发起请求;②MS 收到请求后,首先查找 LMDT,明确自己是否是本次请求涉及到的元数据的宿主服务器. 如果 LMDT 中没有相应的信息,它向 BS 发出请求,刷新 LMDT;③BS 收到 MS 发来的请求后,查找GMDT. 如果该元数据没有绑定,BS 根据动态灵活的策略进行绑定操作. 它首先明确是否存在预先设定的用户特定的绑定策略. 如果有,则按照设定的策略进行后续的操作. 其次,它明确与该元数据在元数据请求中存在关联关系的元数据当前的宿主服像器,比如,本元数据对应文件的父目录元数据. 这样做是为了减少后续元数据操作可能涉及到多个元数据服务器的概率. 如果该元数据服务器负载很大,则选择一个当前的负载值和未来预期的负载值都较小的元数据服务器作为该元数据的宿主服务器,完成绑定过程. BS 将绑定结果返回给 MS;④MS 根据 BS 返回的结果判断. 如果元数据绑定在自己,则进行相应的处理,并返回处理结果给 AS. 否则,它将新的宿主 MS 信息返回给 AS;⑤AS 根据返回信息进行处理. 并在请求完成

后,更新自己缓存的元数据宿主服务器信息.

为减少文件系统出错恢复时间,BWFS 采用分布式的日志技术提供对出错文件系统快速恢复能力的支持. BWFS 的分布式日志系统由各个 MS 的本地日志系统构成. 在绝大多数情况下,各个 MS 在各自的日志空间中,采用本地日志技术记录自己负责的元数据操作的执行情况. 只有在一个活跃元数据的宿主元数据服务器发生改变时,分布式日志系统才通过 BS 协同,要求其原来的宿主服务器处理日志的内容,以防止过期的内容覆盖新的内容.

当一个 MS 出错后,其他的服务器使用它记录的日志内容进行离线恢复操作. 在恢复过程中,日志内容涉及到的元数据处于暂时不可访问状态.

3.4 BWFS 文件数据访问机制

BWFS 将客户端文件访问的数据流与控制流分离,以减小各个服务器的负载,提高客户数据 I/O 吞吐率及其扩展能力.

BWFS 文件访问的"集中决策,分布处理"原则体现在 MS 负责管理和控制文件数据块在 SN 上的分布,各个 SN 直接为客户端提供数据读写请求处理.

AS 读写文件时,首先在缓存中查找请求对应的设备块号. 如果没有找到,它将请求以〈*offset*, *count*〉的方式发给宿主 MS. MS 收到请求后,如果是已经分配的块,获取块号. 如果需要新分配数据块,它根据灵活的资源分配策略,比如,为提高数据读写的并行度而采用"条带化"技术将分布在不同的 SN 上等等,进行块分配. MS 将块号以及负责块内容读写的 SN 的信息返回给 AS. 为减少块号查询请求的数量,BWFS 的当前实现是一次从 MS 获得 256 个块映射信息,AS 与 MS 的一次通信就可以获得连续 256×4KB 的文件内容的块信息. AS 获得块号和映射信息后,直接向相应的 SN 发起读写请求.

AS 使用操作系统的数据缓冲区缓存文件数据,并采用简单高效的机制保证多个 AS 的数据一致性. 为降低数据缓存一致性保障机制对系统性能的影响,BWFS 目前的实现仅对写文件操作提供数据一致性保障,没有保证读操作获得的内容结果的有效性. AS 在开始写之前,首先与 MS 进行通信,申请写权限. MS 收到请求后,检查该文件当前存在的写权限. 如果没有冲突,它授权给该 AS. 否则,它可以让本请求等待,也可以要求相应的 AS 将更改过的内容写到设备,释放写权限. MS 将结果返回 AS. AS 根据获得的结果进行相应的写操作处理.

4 BWFS 特性的初步验证

BWFS 通过将文件数据访问控制流和数据流有效地分离,为系统客户提供高吞吐率、可扩展的数据服务;BWFS 采用高效灵活的资源管理机制,能够提供大容量、可扩展的存储服务;BWFS 根据用户设定的要求、元数据间的关联关系和元数据服务器的负载情况,动态地进行元数据分布决策,保证了元数据服务器能够动态加入到服务器集群中,元数据以非常平稳的方式分布到新的元数据服务器上. 同时,根据应用的需求情况,系统能够动态地决定参与应用请求处理的元数据服务器. 系统的元数据服务具有很高的扩展能力,并表现出根据用户需求动态分配系统资源的动态处理能力.

BWFS 已经运用到 BW1K 网络存储系统中. BW1K 目前采用千兆以太网作为高速互联网络,系统由 128 个应用服务器、2 个元数据服务器、1 个绑定服务器和 32 个网络存储设备构成,存储容量达到 512TB. BW1K 已经在实际应用环境中使用,稳定地为应用提供存储服务.

我们通过对 BW1K 的系统聚合 I/O 吞吐率随 AS 和 SN 规模变化情况的测试,初步验证 BWFS 的 I/O 吞吐率及其扩展能力.

测试 1. 首先测试 1 个 SN 的 BWFS 与 NFS 的大文件读写聚合 I/O 吞吐率,验证 BWFS 的高吞吐率特性. 然后测试小规模 AS 时,BWFS 聚合 I/O 吞吐率随 SN 数量变化的扩展能力. 它采用的硬件环境是 Intel® Xeon™ CPU 2.40GHz,Intel® 82545EM 千兆以太网控制器. 每一个 SN 都配置 3ware® 9500 SATA 磁盘阵列控制器,12 块 160GB 的 Seagate® SATA 硬盘做成 RAID10. SN 配置 2GB 内存,其他服务器为 1GB 内存. 所有这些节点通过两个 NETGEAR JGS524 千兆交换机互联. 测试将 AD,BS 和 MS 放在一个服务器,称为系统服务器. NFS 服务器是一个使用 EXT3 的 SN. BWFS 的 AS 使用 RedHat 9.0,SN 和系统服务器使用 RedHat8.0. BWFS 2.0 优化版. 每个客户端单进程使用 1MB 大小的块进行 dd 操作读写 20GB 的文件. 测试结果如图 3、图 4 所示.

图 3 是 NFS 和使用 1 个 SN 的 BWFS 在不同 AS 规模下的读写聚合 I/O 吞吐率的比较. 在这 5 种规模下,BWFS 的读写 I/O 吞吐率均高于 NFS,验证了 BWFS 具有比 NFS 高的聚合 I/O 吞吐率的特性.

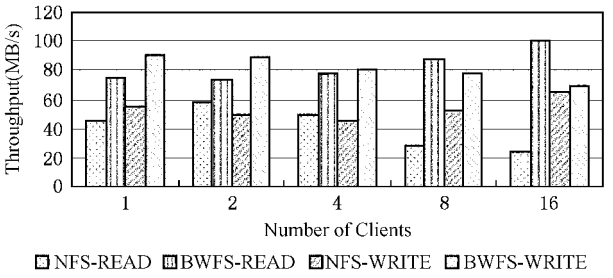

Fig. 3 Throughput comparison of NFS and 1-SN BWFS.

图 3 NFS 和 1 个 SN 的 BWFS 聚合读写吞吐率对比

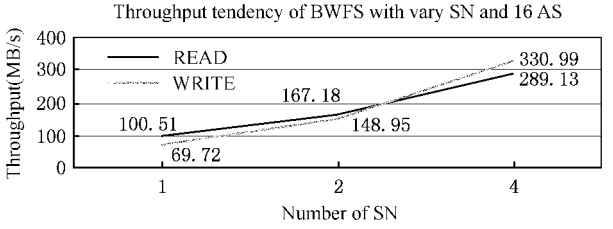

Fig. 4 Throughput tendency of BWFS with vary SN.

图 4 BWFS 吞吐率随 SN 变化情况

图 4 是 16 个 AS 时,文件读写的聚合 I/O 吞吐率随 SN 数量的变化趋势图. BWFS 针对大文件读写采用的条带化块分配等优化策略使文件的块分布在不同的 SN 上,数据读写在 SN 的高度并行.SN 的增加既能够扩大系统的存储空间,又能够提高系统的大文件读写聚合 I/O 吞吐率.

测试 2. 针对 BWFS 在 AS 数量较大,并且 AS 和 SN 数量变化时,系统的聚集 I/O 吞吐率的变化情况进行测试. 测试采用的硬件环境是,AS 配置为 Intel® Xeon™ 2GHz,2GB 内存,RedHat8.0;SN 的配置 Intel® Xeon™ 2.4GHz,1GB 内存,6 块 120GB 的 IDE 硬盘,采用 3ware® 的 RAID 卡,配置成 RAID0,RedHat8.0;所有服务器通过千兆以太网连接. 测试使用我们自己设计的测试用例. 它模拟石油探测高性能计算的模式,测试读大文件的聚合I/O吞吐率. 测试首先生成一个大文件. 然后多次对这个大文件进行读操作,测试系统读文件的聚合I/O吞吐率. 图 5 是在 32 个 AS 和 1 个 SN,32 个 AS 和 2 个 SN 以及 52 个 AS 和 2 个 SN 这 3 个系统规模下,多次测试的系统 AS 读文件聚合 I/O 吞吐率的最小、最大和平均值.

图 5 的数据表明:① SN 的增加将提高聚合 I/O 吞吐率. 在 32 个 AS 时,两个 SN 的平均聚合 I/O 吞吐率比一个 SN 时提高近 30%;② AS 超过一定的规模时,BW1K 的聚合 I/O 吞吐率平均值有所下降,但仍维持在一个相对稳定的数值. 在 52 个 AS、两个 SN 的系统规模的测试过程中,从 SN 上收集到的磁盘 I/O 利用率在 65%~75%,CPU 的利用率在 60%~65%,网络带宽利用率在 85%~97%. 这表明,SN 服务器本身的处理能力并没有达到饱和,而网络的处理能力基本达到极限,网络的处理能力限制着 SN 处理更多的请求. 综合图 4 和图 5 的数据,BWFS 对存储资源和应用规模的扩展的支持得到了初步的验证,并初步明确网络将成为限制系统 I/O 吞吐率扩展的主要因素.

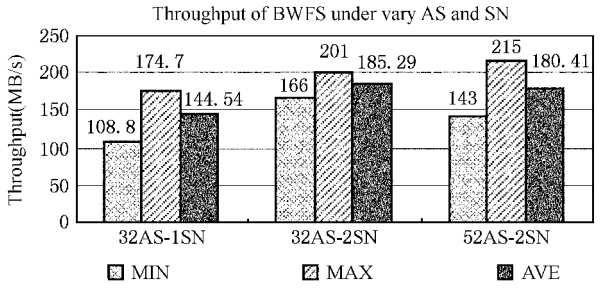

Fig. 5 Throughput variation of large scale BWFS.

图 5 不同系统规模 BWFS 的聚合 I/O 吞吐率

综合以上数据,BWFS 的大容量、高吞吐率和高扩展能力特性得到了验证.

5 结束语

本文描述了 BWFS 基于网络可扩展的系统结构、灵活高扩展能力的资源管理机制、高效可扩展的元数据访问和数据访问机制,并通过对 BW1K 的初步测试,验证 BWFS 的高聚合 I/O 吞吐率和高扩展能力.

应用的需求决定了网络存储系统的发展方向. 网络存储系统正向高吞吐量、高峰值数据传输率、高增容率、高检索速度、高数据安全性、高可用性等方向发展. 在未来的工作中,BWFS 将在高扩展元数据服务、I/O 吞吐率的进一步优化、服务的高可用、异构的应用服务器环境支持和应用请求的 QoS 保证等方面进行大量的工作,以适应更多应用的需求.

参 考 文 献

1 SIMS of UC Berkeley. How Much Information. http://www.sims.berkeley.edu/ how-much-info/, 2000-11-10

2 D. Anderson, J. Chase, A. Vahdat. Interposed request routing for scalable network storage. Duke University, Tech Rep:CS-2000-05, 2000

3 J. Menon, D. A. Pease, R. Rees, *et al*. IBM storage tank—A heterogeneous scalable SAN file system. IBM Systems Journal, 2003, 42(2): 250~267

4 P. J. Braam. The lustre storage architecture. http://www.lustre.org/docs/lustre.pdf, 2003-08

5 R. O. Weber. Information Technology—SCSI object-based storage device command. Technology Committee 10 Drafts. http://www.t10.org/ftp/t10/drafts/osd/osd-r10.pdf, 2004-07

6 J. S. Glider, C. F. Fuente, W. J. Scales. The software architecture of a SAN storage control system. IBM Systems Journal, 2003, 42(2): 232~249

7 Tian Ying, Xu Lu. Technology of load balancing in distributed file system. Computer Engineering, 2003, 29(19): 42 ~ 44 (in Chinese)

(田颖,许鲁. 分布式文件系统中的负载平衡技术. 计算机工程, 2003, 29(19): 42~44)

Yang Dezhi, born in 1977. He is now a doctoral candidate of ICT, CAS. And his current research interests include scalable distributed file system metadata service, large scale distributed file system and networking storage systems.

杨德志,1977 年生,博士研究生,主要研究方向为海量网络存储系统、分布式文件系统、分布式文件系统可扩展元数据服务等.

Huang Hua, born in 1978. He is now a doctoral candidate of ICT, CAS. And his current research interests include large scale distributed file system, networking storage systems.

黄华,1978 生,博士研究生,主要研究方向为分布式文件系统、海量存储、网络存储等.

Zhang Jiangang, born in 1971. He is now an associated professor of ICT, CAS. His current research interests include large scale distributed file system, networking storage systems.

张建刚,1971 年生,博士,副研究员,硕士生导师,主要研究方向为分布式文件系统、海量存储、网络存储等.

Xu Lu, born in 1962. He is now a professor of ICA, CAS. His current research interests include virtual storage system, large scale distributed file system, networking storage systems.

许鲁,1962 年生,博士,研究员,博士生导师,主要研究方向为分布式文件系统、海量存储、虚拟化存储等.

Research Background

With the increasing requirements of enormous applications, storage subsystem becomes the center of computer architecture. And in recent years, researches of networking storage systems have become tendency of that of storage subsystem.

There have been many efforts on large scale distributed storage system, such as Slice/Duke University, Lustre/ClusterFS, StorageTank/IBM, and so on. And there are few researches of large scale distributed storage systems in China now. With supports from the National High Technology Research and Development Program ("863" Program, No. 2002AA112010), BlueWhale 1000 networking storage system (BW1k) was designed by the National Research Center for High Performance Computers of the Institute of Computing Technology, Chinese Academy of Sciences.

BW1k consists of the share-disk storage subsystem, VSDS, the large distributed file system, BWFS, and the management subsystem. With the global file system provided by BWFS, BW1k provides file-level sharing of storage resources to all clients with storage resources independently and transparently. And now BW1k represents with higher I/O throughput and larger scalability than NFS systems in many high performance computing applications. It provides large capacity, high aggregated I/O throughput and high scalability storage services to clients. In the future, it will function more efficient with added characteristics in a large quantity of metadata-intensive and I/O-intensive applications.