**调研报告**

1. **项目主题**

**基于分布式文件系统的NAS集群共享**

1. **项目背景**

**1、互联网大数据的高速发展**

**随着互联网的高速发展，信息技术的日益普及，人们处理的数据量正在成千上万的增长，对数据的计算能力和存储能力的要求不断提高，海量数据的时代已经来临。而作为这些数据信息的载体，web的易用性催生出了电子商务、社交网络、在线旅游等垂直性网站。根据中国互联网络中心发布的《第41次中国互联网络发展状况统计报告》【1】显示，截止至2017年12月，我国网民数已达7.72亿，相较于2016年增幅为4074万。该报告显示，手机终端的网络音乐、网络视频、网络游戏等网民规模相较2016年分别增长了9.36%、9.73%、15.77%，保持了较高的增长率。另外，以微博为代表的社交网站用户数也得到了迅速增长，2017年12月底，新浪微博活跃用户数已达3.76亿，网民中微博使用率达到了40.9%。无论是电子商务网站图片的展示，或社交网站图片的分享，图片数据量都达到了指数级别的增长。**

**以国内外互联网巨头来说，截至2014年底，Facebook用户上传的图片就已经达到了2500亿张，平均每天上传3.5亿张，在高峰期Facebook每秒能够处理55张照片。国外最大的图片分享网站Flickr则存储了接近6亿张图片，而且绝大多数图片还是高清的，单张图片大小一般达到了4-5M，公共占据了大约2.5PB存储空间，每秒的处理请求次数为38000次，每天新增的图片超过50万张。淘宝作为我国最大的电子商务网站，在线商品数量达到了10亿级别，图片存储服务器存储数量更是达到了290亿张图片，总量在1PB以上，而且每天仍在数以千万计的增长。因为图片传达信息的能力远大于文字描述，所以电子商务网站一般都特别注重上传时间，图片的质量和访问速度等问题。而根据淘宝的流量分析，淘宝所有的通信量中传输量更是达到了90%。我国另一社交公司巨头腾讯公司用户上传的图片更是达到了600多亿张，每人每天上传的图片为10张，图片的峰值访问次数更是达到了每秒50万次。**

**大数据时代的到来已经不可阻挡，面对数据的爆炸式增长，尤其是半结构化数据和非结构化数据，NoSQL存储系统和分布式文件系统成为了技术浪潮，得到了长足的发展。非结构化数据目前呈现更加快速的增长趋势，IDC研究报告分析指出，到2012年非结构化数据将占到数据存储总量的80%以上。集群NAS是一种横向扩展（Scale-out）存储架构，具有容量和性能线性扩展的优势，已经得到全球市场的认可。从EMC对ISILON、HP对IBRIX、DELL对Exanet等收购事件，以及IBM推出SONAS、NetApp发布Data ONTAP 8，都可以看出集群NAS已经成为主流存储技术之一。在国内，我们也看到UIT UFS、龙存LoongStore、九州初志CZSS、美地森YFS等集群NAS解决方案。集群NAS的未来潜在市场巨大，在高性能计算HPC、广电IPTV、视频监控、云存储等行业领域将逐步得到广泛应用。这些商业集群NAS系统固然很好，在性能、可用性、扩展性、易用性、管理性以及服务和支持方面都很不错的表现，但往往成本也很斐然，很多大企业都消受不起，更别提那些中小企业了。受益于近年来开源存储软件的蓬勃发展，我们完全可以使用普通的服务器、存储和网络设备，构建出低成本的高性能、高扩展、高可用基于分布式文件系统的集群NAS系统。**

**2、共享概念逐渐成熟**

**近几年共享经济发展迅速，但当下的互联网共享经济形态中，先不论项目好坏，“实物共享”仍然是主流形态，共享单车、共享租车、共享充电宝、共享雨伞、共享篮球等莫不如是。**

**诚然，不同用户在实物资源方面的闲置与缺失的矛盾是最突出的，例如出行经常面临无车或者不想驾车，也因此共享经济能够率先在实物领域突围。此外，技术上的桎梏也让共享实物这种最直接的方式能够最先落地。但随着信息互联网基础设施的完善以及人群生活中所接触的必要资源种类日益增多，技术、资源等非实物方面的共享需求将变得更大。**

**例如，很多人的电脑硬盘存储空间大多数时候用不完，过剩也造就了资源的浪费，同时，与之相关的企业数据中心等场所又面临巨大的存储压力。为何不把过剩的存储空间共享出来？这样不仅资源的利用率将会大大增加，同时存储成本也能大大降低。**

1. **立项依据**

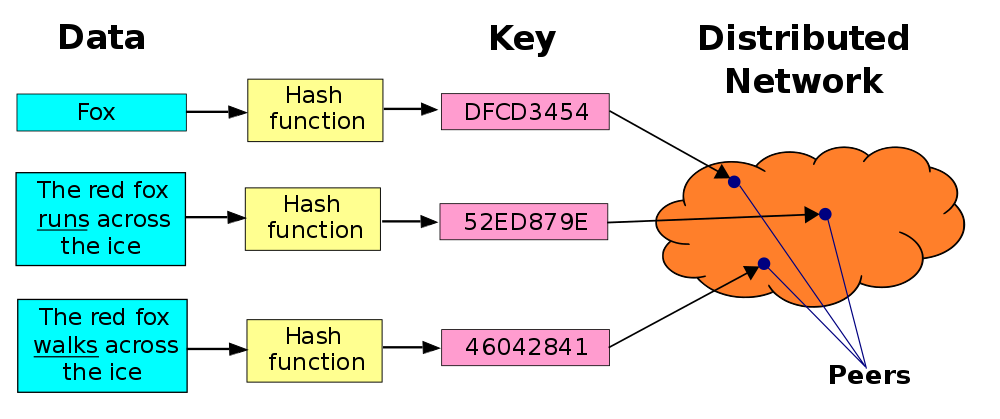
**此处有点不好写。。。。。可以说是老师让选的题吗。。。。**

1. **前瞻性\重要性分析**

**面对互联网时代数据的爆炸增长，利用开源的软件或协议，使用普通的服务器、存储和网络设备，构建出低出本的高性能、高扩展、高可用基于分布式文件系统的集群NAS来存储海量数据。同时在基于分布式文件系统的NAS集群上应用共享的概念，让用户贡献存储空间而不是搭建一些集中服务器提供存储空间，让存储的成本更低，效率更高，与未来互联网的发展趋势相符，无疑会有广阔的发展前景。**

1. **相关工作**
2. **DHT（分布式哈希表）**

**A distributed hash table (DHT) is a class of a decentralized distributed system that provides a lookup service similar to a hash table: (key, value) pairs are stored in a DHT, and any participating node can efficiently retrieve the value associated with a given key. Responsibility for maintaining the mapping from keys to values is distributed among the nodes, in such a way that a change in the set of participants causes a minimal amount of disruption. This allows a DHT to scale to extremely large numbers of nodes and to handle continual node arrivals, departures, and failures.**

****

1. **IPFS[2]**

**IPFS is a distributed file system which synthesizes successful ideas from previous peer-to-peer sytems, including DHTs, BitTorrent, Git, and SFS. The contribution of IPFS is simplifying, evolving, and connecting proven techniques into a single cohesive system, greater than the sum of its parts. IPFS presents a new platform for writing and deploying applications, and a new system for distributing and versioning large data. IPFS could even evolve the web itself. IPFS is peer-to-peer; no nodes are privileged. IPFS nodes store IPFS objects in local storage. Nodes connect to each other and transfer objects. These objects represent ﬁles and other data structures. The IPFS Protocol is divided into a stack of sub-protocols responsible for diﬀerent functionality:**

**1. Identities - manage node identity generation and verification.**

**2. Network - manages connections to other peers, uses various underlying network protocols. Configurable.**

**3. Routing - maintains information to locate specific peers and objects. Responds to both local and remote queries. Defaults to a DHT, but is swappable.**

**4. Exchange - a novel block exchange protocol (BitSwap) that governs eﬃcient block distribution. Modelled as a market, weakly incentivizes data replication. Trade Strategies swappable.**

**5. Objects - a Merkle DAG of content-addressed immutable objects with links. Used to represent arbitrary datastructures, e.g. file hierarchies and communication systems.**

**6. Files - versioned file system hierarchy inspired by Git.**

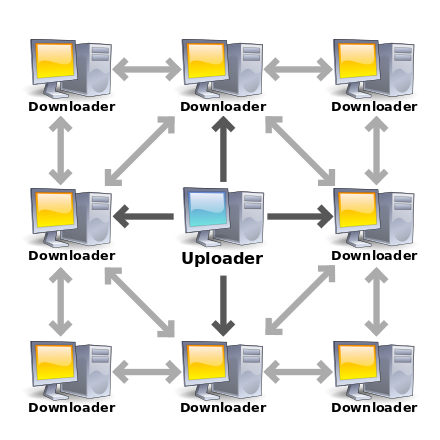
**7. Naming - A self-certifying mutable name system.**

**These subsystems are not independent; they are integrated and leverage blended properties. However, it is useful to describe them separately, building the protocol stack from the bottom up.**

1. **BitTorrent**

**BitTorrent (BT) is a communication protocol for peer-to-peer file sharing ("P2P") which is used to distribute data and electronic files over the Internet.**

**BitTorrent is one of the most common protocols for transferring large files, such as digital video files containing TV shows or video clips or digital audio files containing songs. Peer-to-peer networks have been estimated to collectively account for approximately 43% to 70% of all Internet traffic (depending on location) as of February 2009.[3] In November 2004, BitTorrent was responsible for 25% of all Internet traffic.[4] As of February 2013, BitTorrent was responsible for 3.35% of all worldwide bandwidth, more than half of the 6% of total bandwidth dedicated to file sharing.[5]**

****

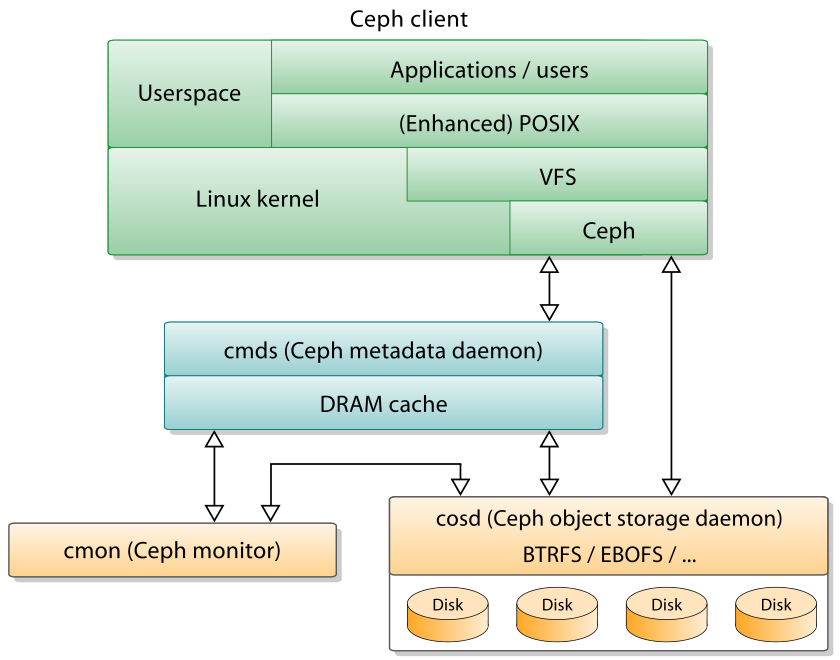
1. **Ceph**

**Ceph is a free-software storage platform, implements object storage on a single distributed computer cluster, and provides interfaces for object-, block- and file-level storage. Ceph aims primarily for completely distributed operation without a single point of failure, scalable to the exabyte level, and freely available.**

**Ceph replicates data and makes it fault-tolerant,[6] using commodity hardware and requiring no specific hardware support. As a result of its design, the system is both self-healing and self-managing, aiming to minimize administration time and other costs.**

**On April 21, 2016, the Ceph development team released "Jewel", the first Ceph release in which CephFS is considered stable.[citation needed] The CephFS repair and disaster recovery tools are feature-complete (snapshots, multiple active metadata servers and some other functionality is disabled by default).[7]**

**The August, 2017 release (codename "Luminous") introduced the production-ready BlueStore storage format which avoids many shortcomings of the old filesystem-based filestore, providing better latency and additional storage features.[8]**

****

**Ceph employs four distinct kinds of daemons:[9]**

* **Cluster monitors (ceph-mon) that keep track of active and failed cluster nodes**
* **Metadata servers (ceph-mds) that store the metadata of inodes and directories**
* **Object storage devices (ceph-osd) that uses a direct, journaled disk storage (named BlueStore[10], since the v12.x release) or store the content of files in a filesystem (preferably XFS, the storage is named Filestore).[11]**
* **Representational state transfer (RESTful) gateways (ceph-rgw) that expose the object storage layer as an interface compatible with Amazon S3 or OpenStack Swift APIs**

**All of these are fully distributed, and may run on the same set of servers. Clients directly interact with all of them.[12]**

**Ceph does striping of individual files across multiple nodes to achieve higher throughput, similar to how RAID0 stripes partitions across multiple hard drives. Adaptive load balancing is supported whereby frequently accessed objects are replicated over more nodes.[citation needed] As of September 2017, BlueStore is the default and recommended storage type for production environments[9], which is Ceph's own storage implementation providing better latency and configurability than the filestore backend, and avoiding the shortcomings of the filesystem based storage involving additional processing and caching layers. The Filestore backend is still considered useful and very stable; XFS is the recommended underlying filesystem type for production environments, while Btrfs is recommended for non-production environments. ext4 filesystems are not recommended because of resulting limitations on the maximum RADOS objects length.[13]**

1. **Glusterfs**

**GlusterFS is a scale-out network-attached storage file system. It has found applications including cloud computing, streaming media services, and content delivery networks. GlusterFS was developed originally by Gluster, Inc. and then by Red Hat, Inc., as a result of Red Hat acquiring Gluster in 2011.[14]**

**In June 2012, Red Hat Storage Server was announced as a commercially supported integration of GlusterFS with Red Hat Enterprise Linux.[15] Red Hat bought Inktank Storage in April 2014, which is the company behind the Ceph distributed file system, and re-branded GlusterFS-based Red Hat Storage Server to "Red Hat Gluster Storage".[16]**

1. **参考文献**

**【1】第41次中国互联网络发展状况统计报告**

**【2】Juan Benet. IPFS - Content Addressed, Versioned, P2P File System (DRAFT 3)**

**【3】"Does bittorrent support SHA-2 256/512 or future SHA-3? - Quora".**

**【4】Jones, Ben (7 June 2015). "BitTorrent's DHT Turns 10 Years Old". TorrentFreak. Retrieved 5 July 2015.**

**【5】 "Application Usage & Threat Report". Palo Alto Networks. 2013. Archived from the original on 31 October 2013. Retrieved 7 April 2013.**

**Retrieved 7 April 2013. (registration required)**

**【6】Jeremy Andrews (2007-11-15). "Ceph Distributed Network File System". KernelTrap.**

**【7】 Sage Weil (2016-04-21). "v10.2.0 Infernalis Released". Ceph Blog.**

**【8】 Sage Weil (2017-08-29). "v12.2.0 Luminous Released". Ceph Blog.**

**【9】M. Tim Jones (2010-06-04). "Ceph: A Linux petabyte-scale distributed file system" (PDF). IBM. Retrieved 2014-12-03.**

**【10】 "BlueStore". Ceph. Retrieved 2017-09-29.**

**【11】"Hard Disk and File System Recommendations". Retrieved 2017-03-17.**

**【12】Jake Edge (2007-11-14). "The Ceph filesystem". LWN.net.**

**【13】"Hard Disk and File System Recommendations". ceph.com. Retrieved 2017-06-26.**

**【14】Timothy Prickett Morgan (4 October 2011). "Red Hat snatches storage Gluster file system for $136m". The Register. Retrieved 3 July 2016.**

**【15】 Timothy Prickett Morgan (27 June 2012). "Red Hat Storage Server NAS takes on Lustre, NetApp". The Register. Retrieved 30 May 2013.**

**【16】"Red Hat Storage. New product names. Same great features". redhat.com. 20 March 2015. Archived from the original on 2 April 2015. Retrieved 20 March 2015.**