1、

块存储主要用于结构化数据类应用，在相对小的容量空间上可以提供更高的IO性能，所以从后端存储产品形态上来看，一般是多控多活的紧耦合集中式SAN架构，存储介质大多采用高速的SAS和近年来已成主流的固态硬盘，前端多采用FC光纤协议组网。另外，存储双活、存储虚拟化、存储复制等技术也支撑着这块的生态。

对象存储近年来在国内市场逐渐火热起来，主要是特定行业的海量非结构化数据的应用场景。由于对象存储采用扁平的文件组织方式，所以在文件量上升至千万、亿级别，容量在PB级别的时候，这种文件组织方式下的性能优势就显现出来了，文件不在有目录树深度的问题，历史和近线数据有同样的访问效率。另外，对象存储多采用分布式架构，可以在商用x86服务器上轻松构建对象存储，磁盘介质也大多采用低速的SATA盘，所以在成本上也具优势。

文件存储主要是集中式或分布式的NAS类应用，以灵活的共享文件系统为特色，承载百万文件量和TB级别的文件类应用。

主要区别：

块存储，是最底层的存储，它关注的是磁盘的基本存储单元——块（Block），块存储可以没有软件服务器。而文件存储（像NAS）和对象存储（S3、七牛云），是软件层面的存储，底层也是基于块存储的。而文件存储和对象存储的区别在于，他们组织文件的方式，文件存储是有目录树的、有标准属性的（权限、用户、读写），而对象存储是扁平的、没有目录的、附加属性是灵活的，他们的接口协议也不一样。对象存储接口主要以S3与Swift为代表，其接口就是简单的GET、PUT、DEL和其他扩展，文件存储的话是以POSIX接口为主，以libcephfs为代表。

且三种存储分别对应了不同的协议：

（1）块存储，主要操作对象是磁盘。以 SCSI 为例，主要接口有 Read/Write/Read Capacity/Inquiry 等等。FC，iSCSI，也是块存储协议。和文件存储相比，没有文件和目录树的概念，一般协议也不会定义磁盘的创建和删除操作。协议更注重传输控制。

（2）对象存储，主要操作对象是对象（Object）。以 S3 为例，主要接口有 PUT/GET/DELETE 等。和块存储相比，没有随机读写的接口。和文件存储相比，没有目录树的概念。协议更注重简洁。

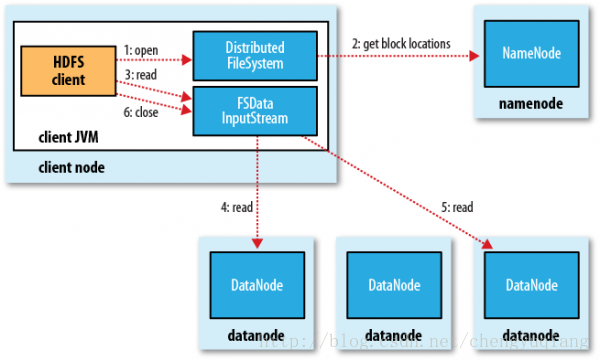
（3）文件存储，主要操作对象是文件和文件夹。以 NFS 为例，文件相关的接口包括：LOOKUP/ACCESS/READ/WRITE/CREATE/REMOVE/RENAME 等等，文件夹相关的接口包括：MKDIR/RMDIR/READDIR 等等。同时也会有 FS\_STAT/FS\_INFO 等接口用于提供文件系统级别的信息。POSIX，SAMBA 等也是文件存储协议。协议更注重接口的灵活，以及访问权限控制。

2、《Hadoop Distributed File System》

(1)客户端读取HDFS系统中指定文件指定偏移量处的数据时，工作流程？

HDFS通过使用RPC来调用nameNode，以确定文件起始块的位置。对于每个块，nameNode返回存有该块副本的datanode地址。datanode将通过和客户端的距离来排序。

DistributedFileSystem类返回一个FSDataInputStream对象给客户端读取数据FSDataInputStream类封装该对象。接着，客户端对这个输入流重复调用read()方法，将数据从datanode传输到客户端。到达块的末端时，关闭连接，寻找下一个最近的datanode，客户端只需要读取连续的流。



（2）客户端向HDFS指定文件追加写入的工作流程？

首先，打开已有的HDFS文件，会调用append方法获取文件最后一个数据开的位置信息，如果文件的最后一个数据块写满则返回null，然后会调用DFSOutputStream.newStreamForAppend()方法创建到这个数据块的

DFSOutputStream输出流对象，获取文件租约，并将构建方法包装为HdfsDataOutputStream对象。

其次就是建立数据流管道，判断最后一个数据块是否写满，如没有写满，则根据ClientProtocol.append()方法返回的该数据块的位置信息建立到该数据块的数据流管道，如果写满了，就调用addBlock()向NameNode申请一个新的空的数据块之后建立数据流管道

第三步：通过数据流管道写入数据，然后调用close()关闭输出流。

（3）新增加一个数据块时，HDFS如何选择存储该数据块的物理节点？

当一个新块被创建时，HDFS把第一份副本放在writer的节点，第二份和第三份存放在不同机架的不同节点，当副本数小于两倍机架数时，每个节点上的副本不多于1份，每个机架上的副本不多于2份。把第2、3份副本放到不同机架能够在集群中更好的分发单个文件的块副本。如果任意文件的头两份副本放在同一个机架，2/3的块副本会在同一个机架上。

（4）HDFS采用了哪些措施应对数据块损坏或丢失问题？

每个DataNode上都允许了一个block scanner，定期地扫描块副本并验证相关的校验和当读客户端或block scanner检测到坏块时，会通知NameNode。NameNode将块标记为损坏，但是不会马上安排删除副本，而是复制出一个好的副本。仅在好副本数达到复制因子，损坏的副本才会被安排删除。这个策略旨在尽可能久地保存数据。所以如果所有的副本都损坏了，该策略允许用户从损坏的副本中读取数据。

（5）HDFS采用了什么措施应对主节点失效问题？

和CHeckpointNode类似，BackupNode也有能力创建定期检查点，但除此之外还在内存中维护文件系统命名空间最新的image，这总是和NameNode的状态一起同步，如果NameNode发生了故障，BackupNode内存中的image和磁盘上的checkpoint就是最新的命名空间状态。

（6）NameNode维护的“数据块-物理节点对应表”需不需要在硬盘中备份？为什么？

不需要，因为HDFS将整个命名空间存放在RAM中。文件块位置信息只存储在内存中，是在DataNode加入集群的时候，NameNode 询问DataNode得到的，并且间断的更新。所以当“数据块—物理节点对应表”失效时可通过向NameNode请求得到最新的文件块位置信息。