对象存储

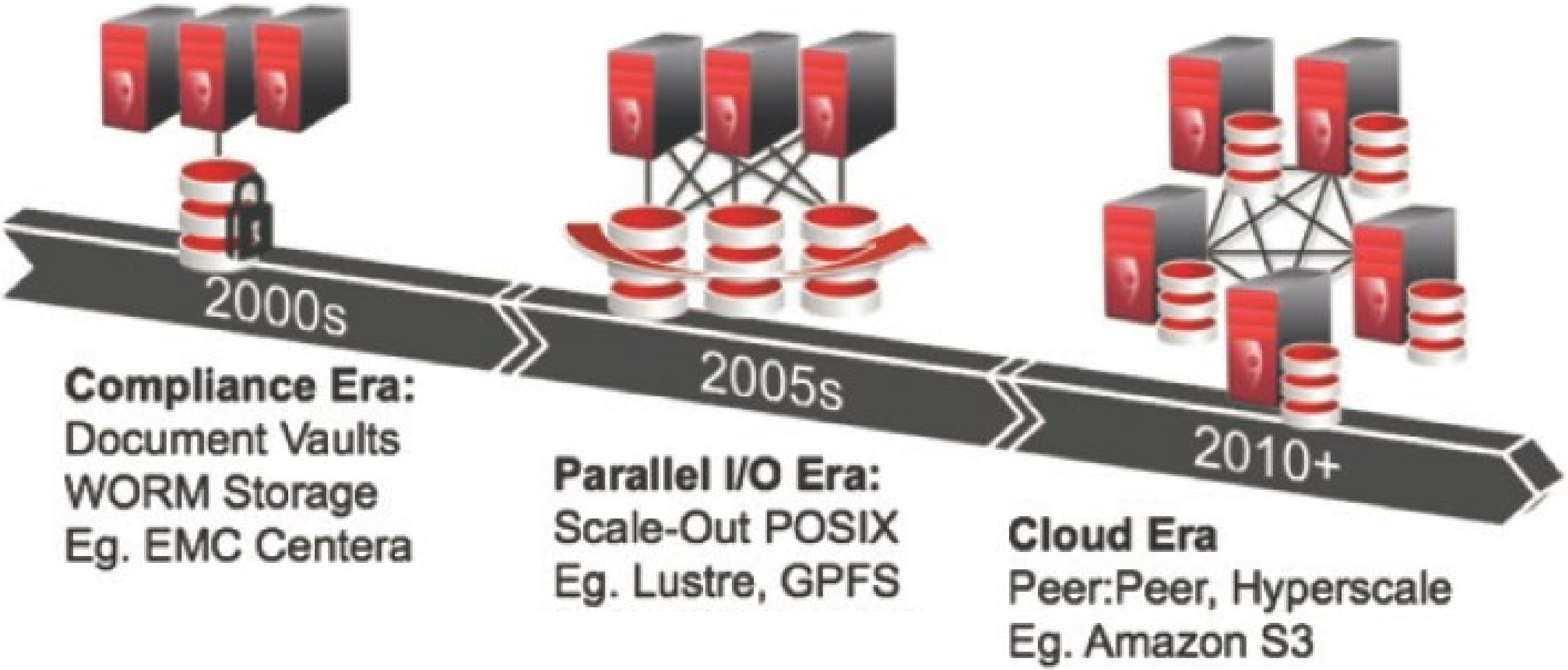
Object-Based Storage

块存储

文件存储

对象存储

History of Object Storage



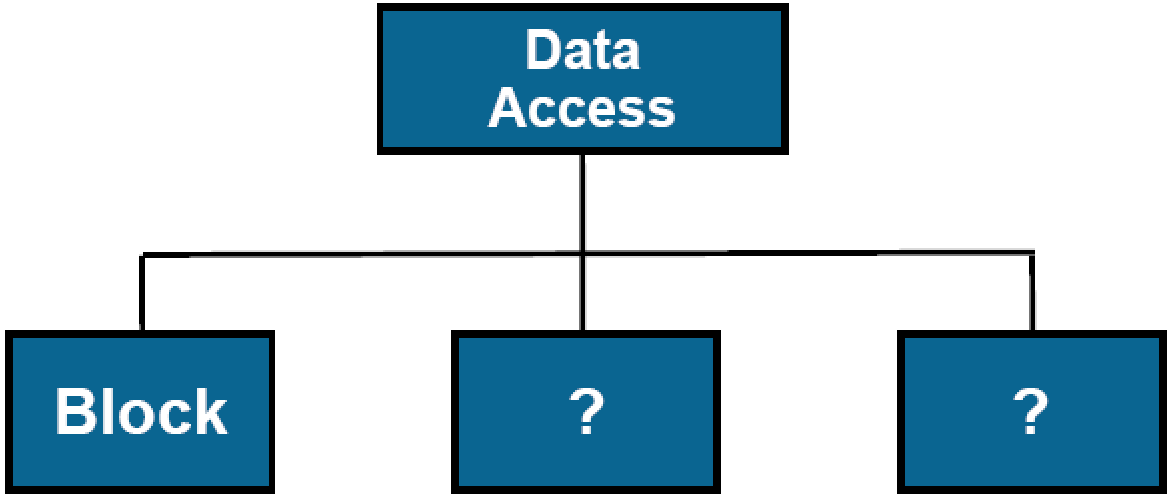
Compliance Era:
Document Vaults
WORM Storage
Eg. EMC Centera

Parallel I/O Era:
Scale-Out POSIX
Eg. Lustre, GPFS

Cloud Era
Peer:Peer, Hyperscale
Eg. Amazon S3

The Data Access Taxonomy

- The Block Paradigm

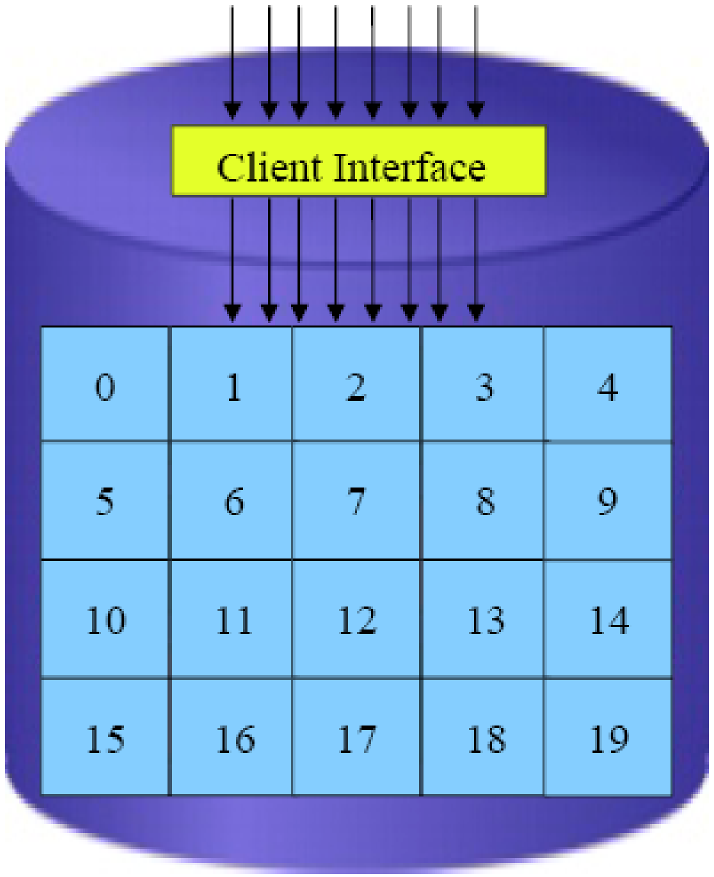


SCSI, ATA
SAS, SATA
FCP, iSCSI
SRP, NVMe…

DAS, SAN

The Block Paradigm

SCSI，ATA，SAS，SATA，FCP，iSCSI，SRP，NVMe…



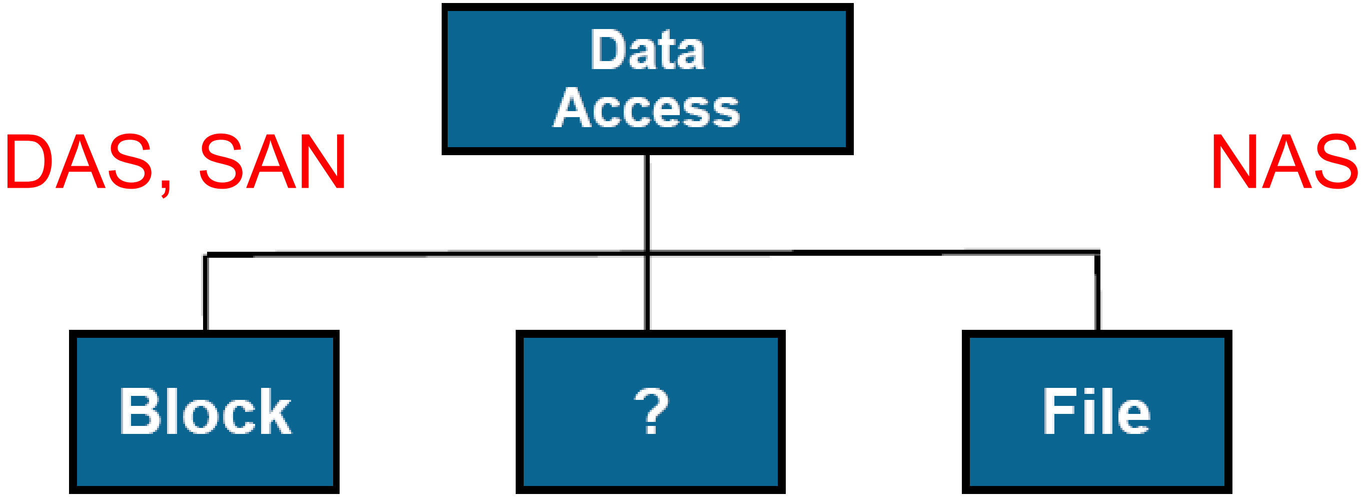
0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19

LBA：Logical Block Address

Physical Blocks:
e.g. 512 bytes

The Data Access Taxonomy

- The File Paradigm

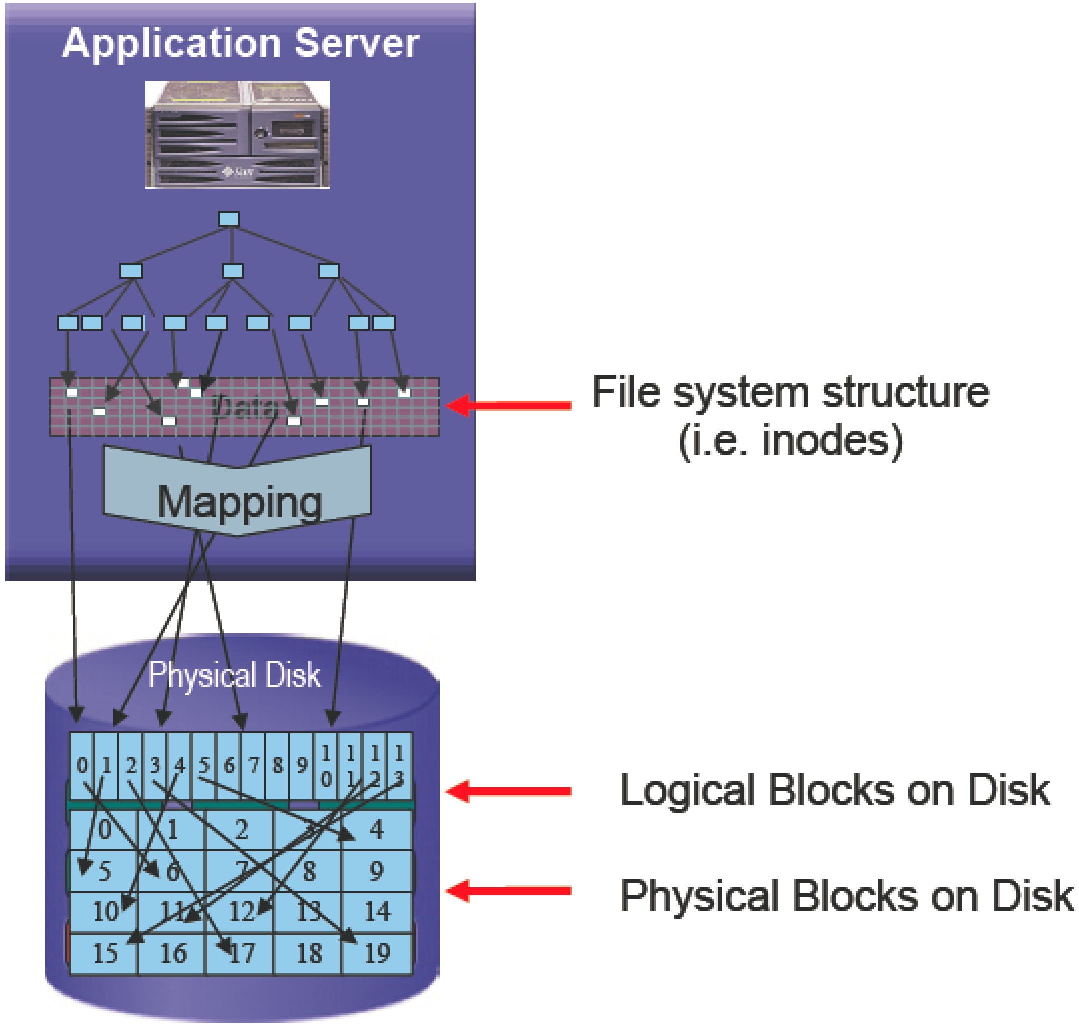


SCSI， ATA
SAS， SATA
FCP， iSCSI
SRP， NVMe…

Local FS,
Distributed FS: (NFS, CIFS, WAFS)
Shared FS: (Cluster FS, SAN FS)
Global, Distributed & Parallel FS

Local File Systems

- file/directory management(~**10%** of workload)
- block/sector management (~**90%** of workload)



File system structure (i.e. inodes)

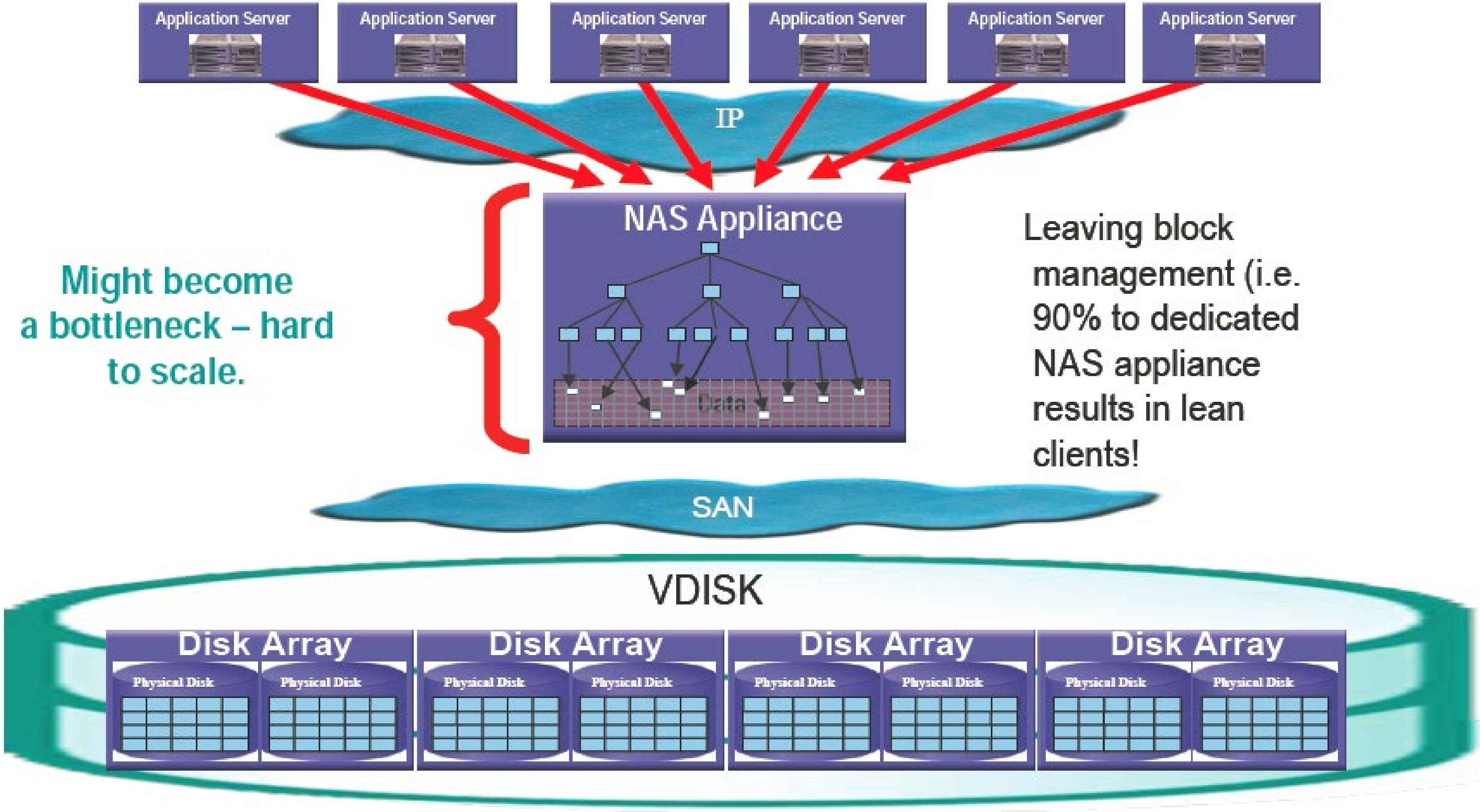
Logical Blocks on Disk

Physical Blocks on Disk

One more level of indirection

Distributed File Systems

e.g. NAS with NFS,CIFS Protocol

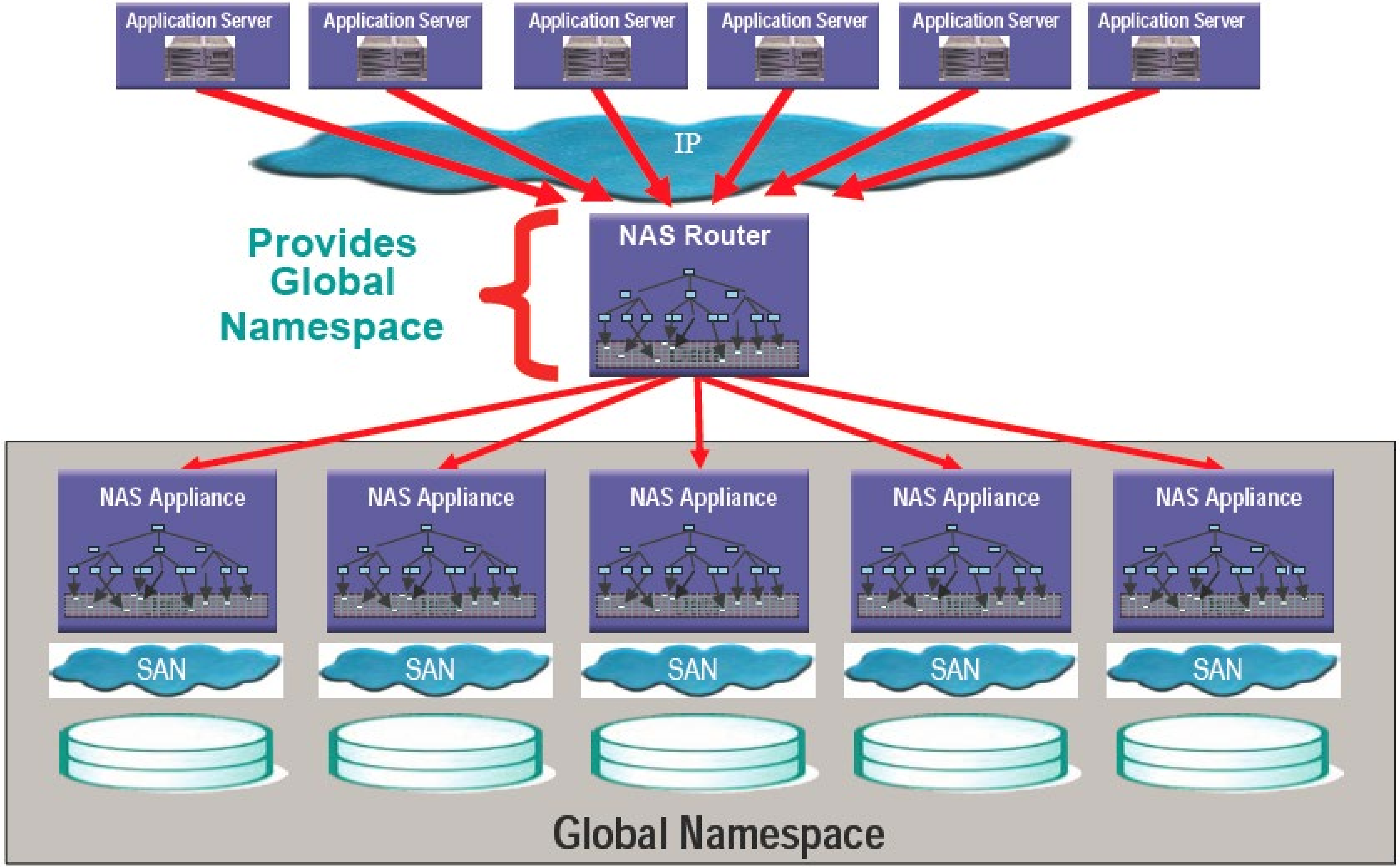


Might become a bottleneck – hard to scale.

Leaving block management (i.e. 90% to dedicated NAS appliance results in lean clients!

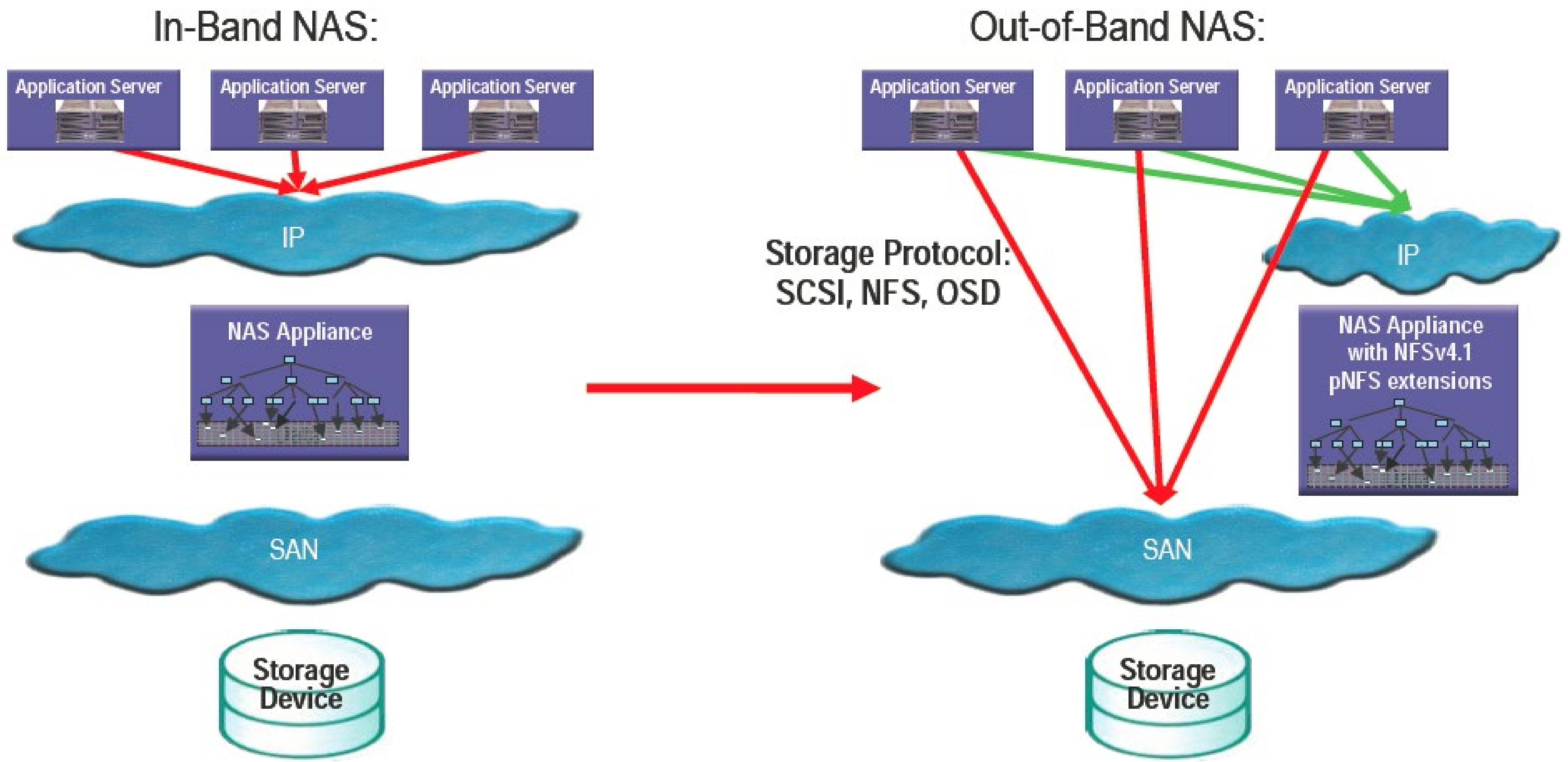
NAS Aggregation/Virtualization

Global Namespace



NAS Cluster

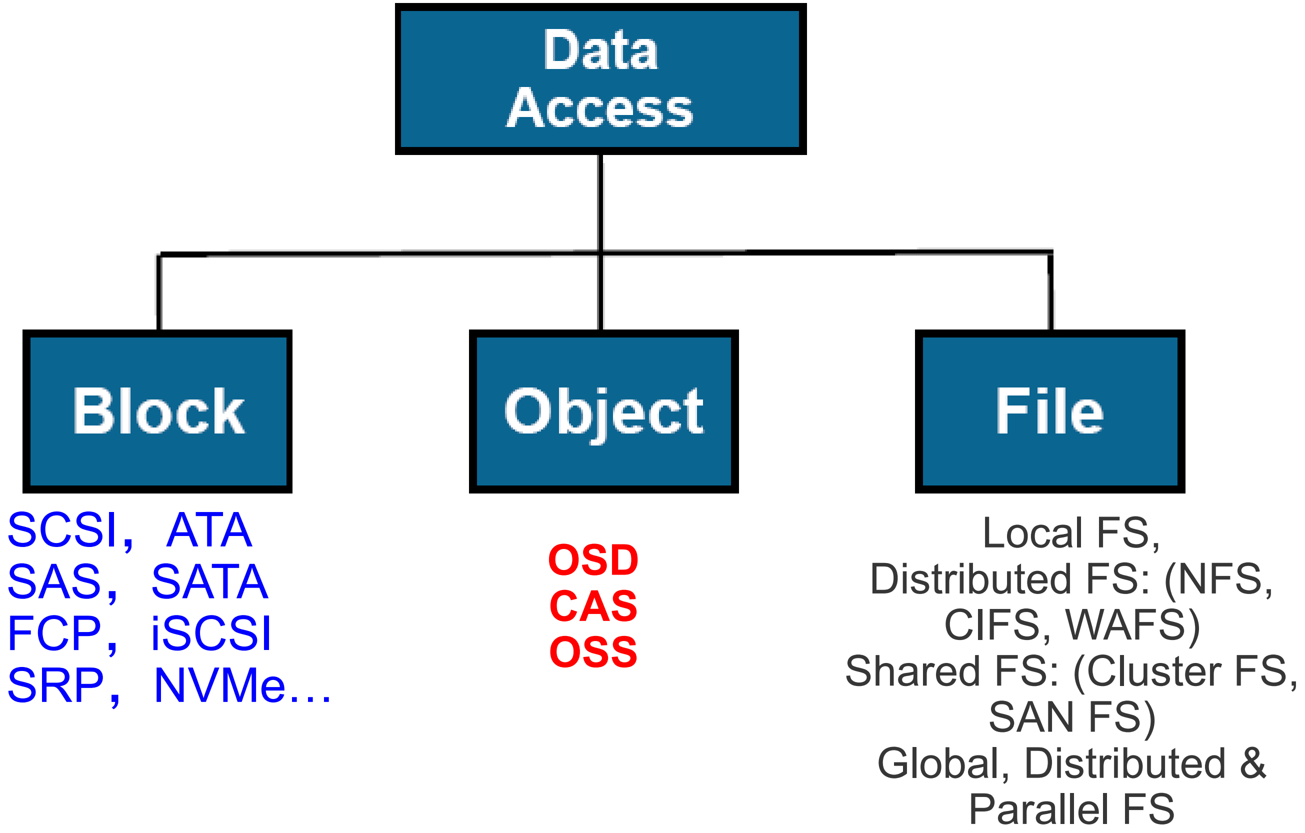
Loosely Coupled NAS: Global Namespace with NFSv4.1 and pNFS



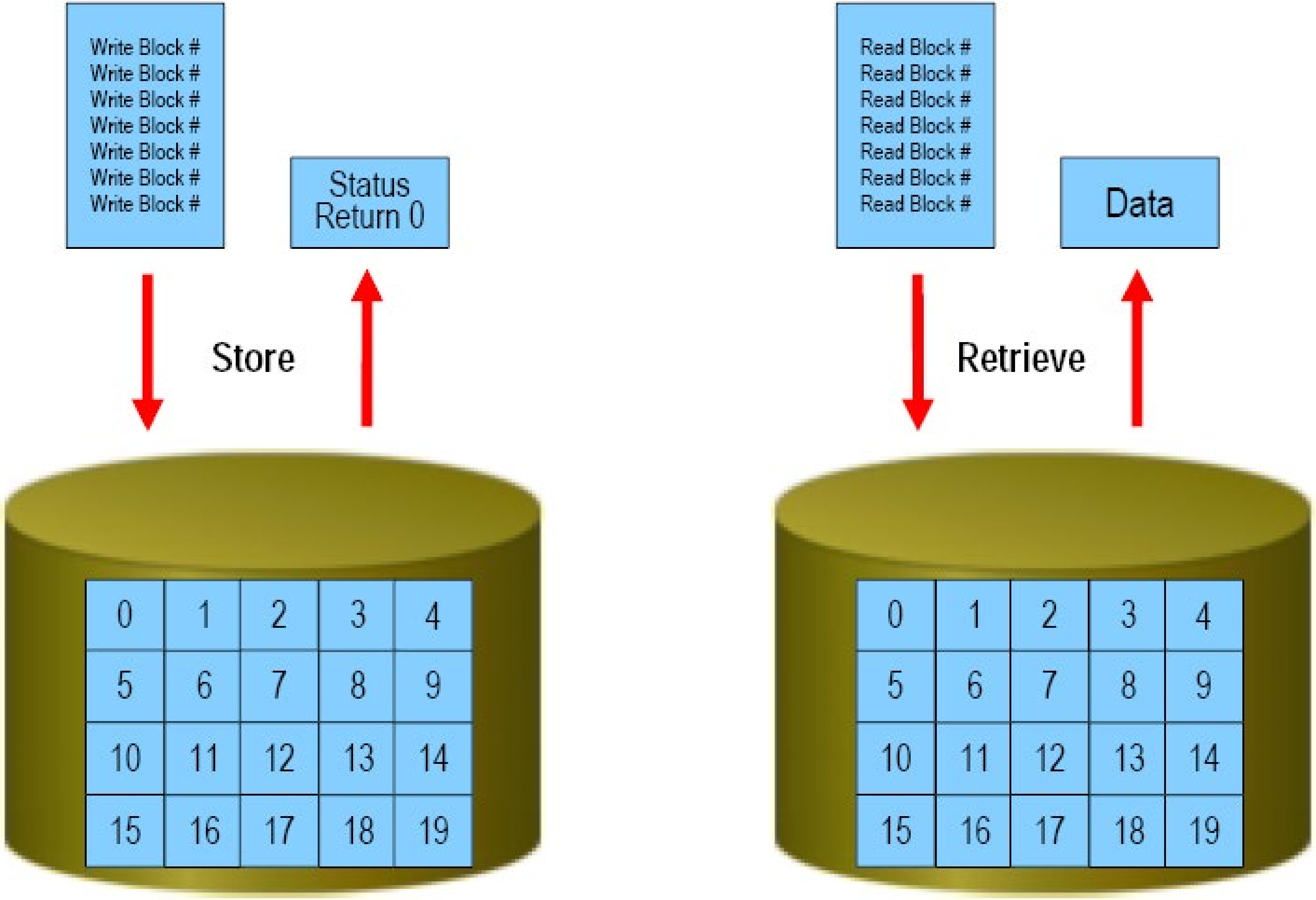
Scalable NAS
Loosely Coupled NAS Cluster

The Data Access Taxonomy

- The Object Paradigm


Data Access
Block
Object
File
SCSI， ATA
SAS， SATA
FCP， iSCSI
SRP， NVMe…
OSD
CAS
OSS
Local FS,
Distributed FS: (NFS, CIFS, WAFS)
Shared FS: (Cluster FS, SAN FS)
Global, Distributed & Parallel FS

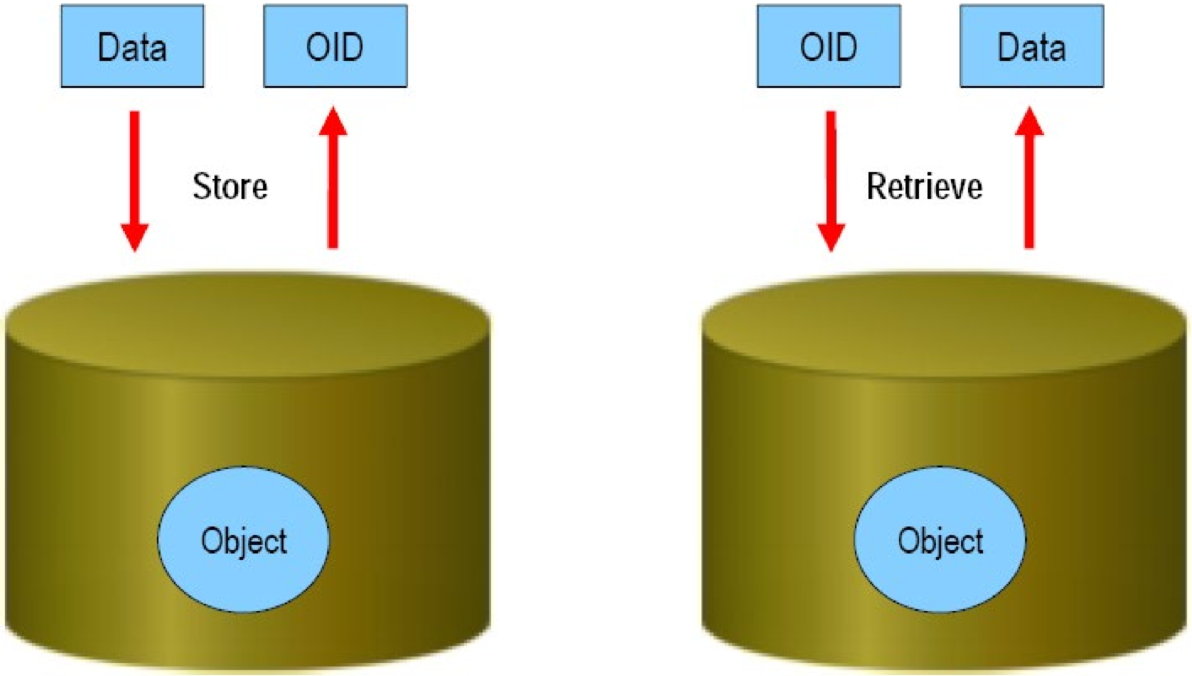
The Old Block Paradigm



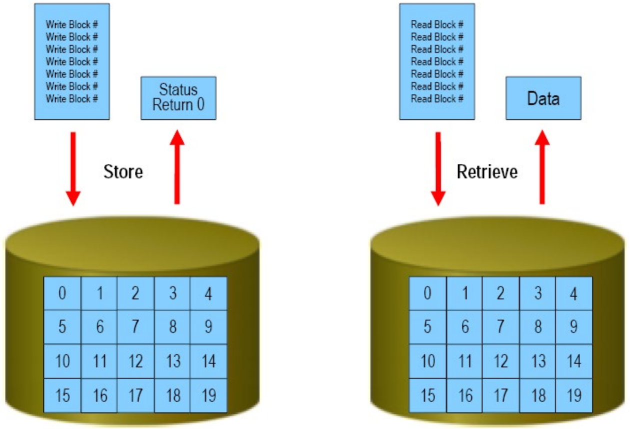
0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19

0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19

The New Object Paradigm

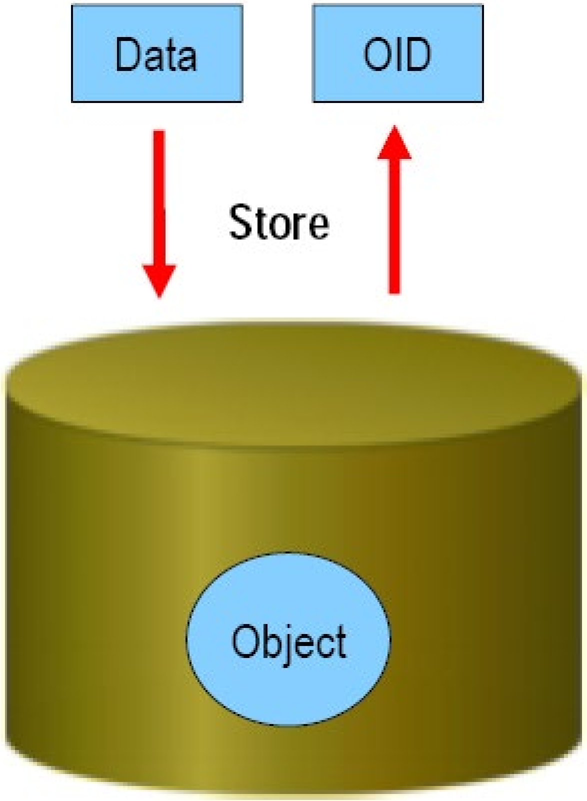


Block Paradigm



The New Object Paradigm

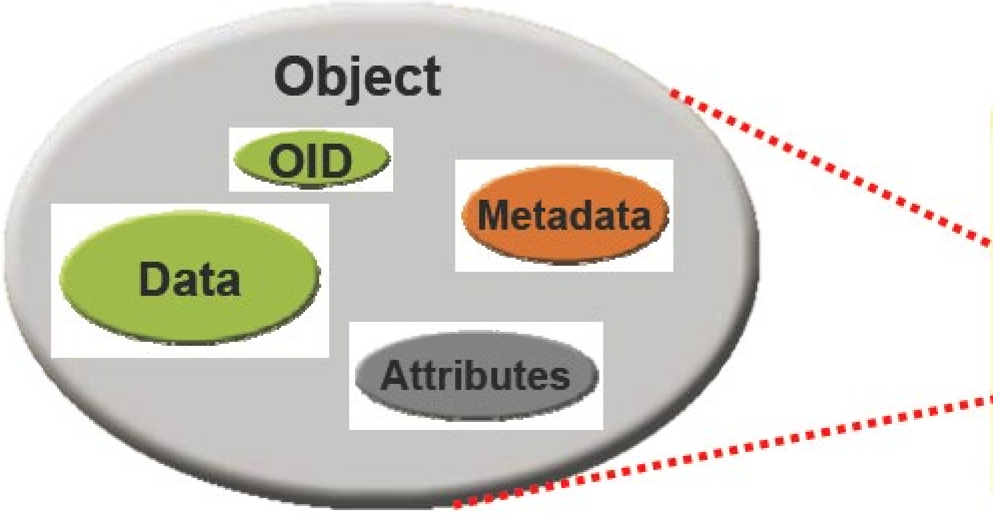
- •WRITE 26,763 Bytes
- •QoS= High
- •Description = “X-Ray”
- •Retention = 50 years
- •Access Key = *&^%#
- •Data Payload........



Object Storage Responsibilities:

- Space Management
- Access Control (Identity Mgmt)
- QoS Management
- Cache, Backup
- Policy Migration, Retention

Self-Contained Objects

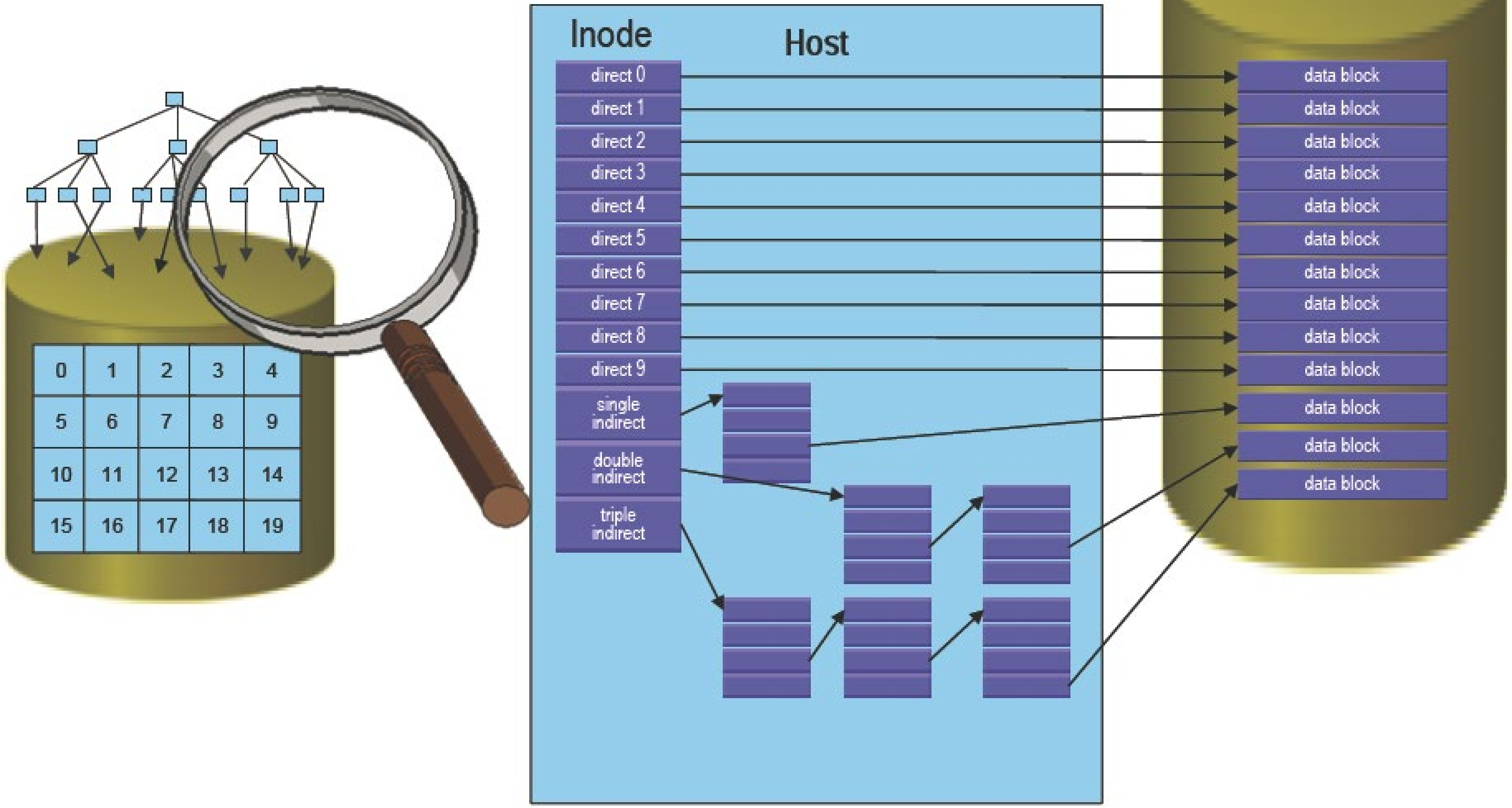


self-contained snail

Metadata:
creation date/time; ownership, size...
Attributes –inferred:
access patterns, content, indexes...
Attributes –user supplied:
retention, QoS…

Block Access - Inodes

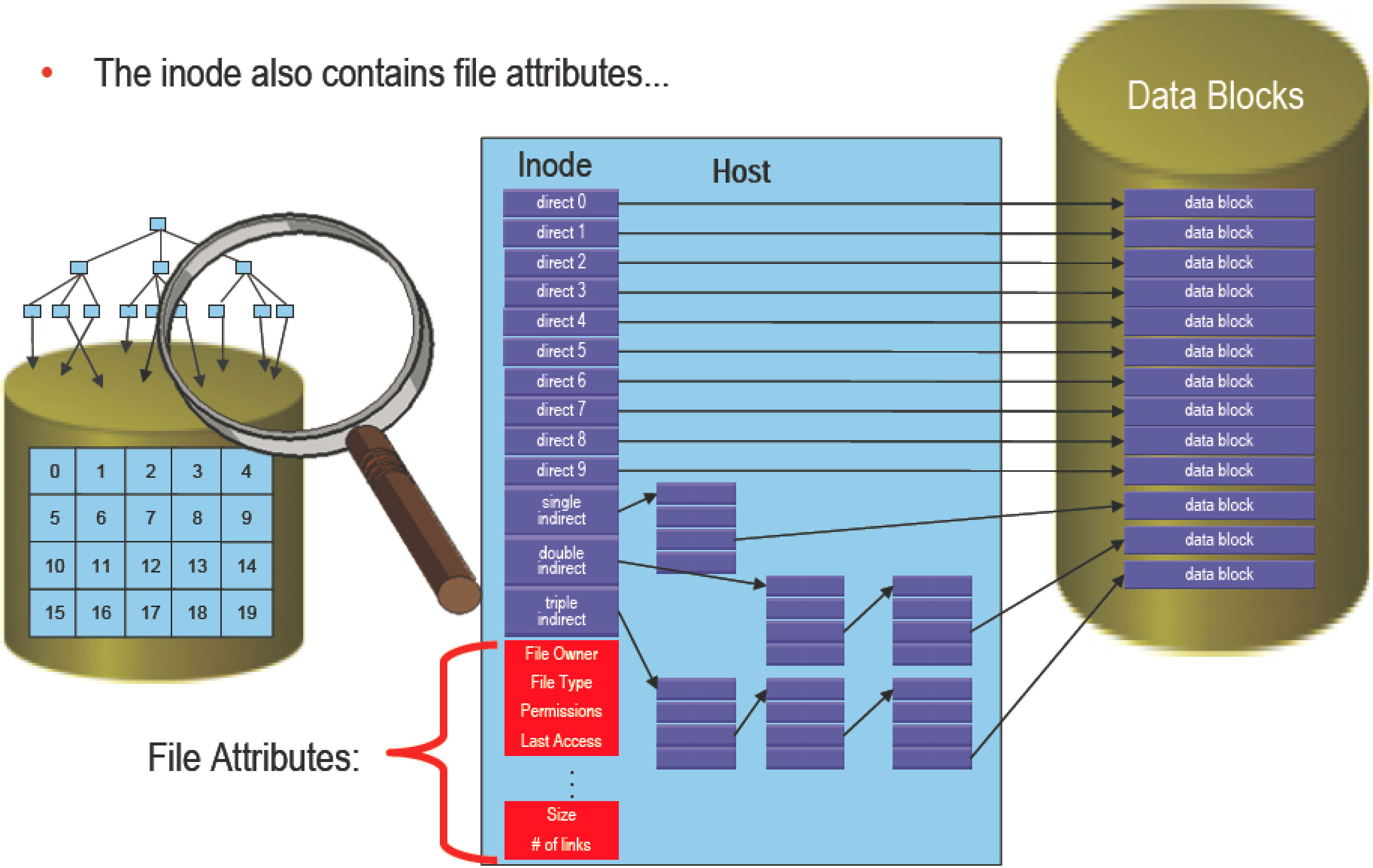
The inode contains a few block numbers to ensure efficient access to small files. Access to larger files is provided via indirect blocks that contain block numbers



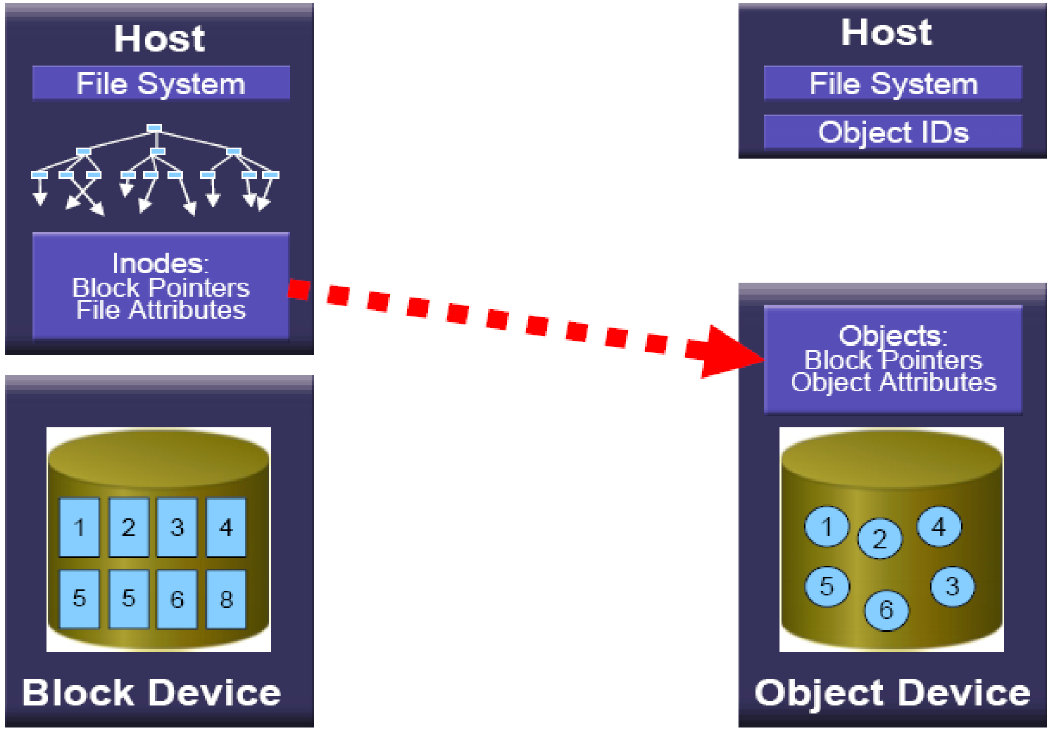
0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19

Block Access - Inodes

- The inode also contains file attributes...

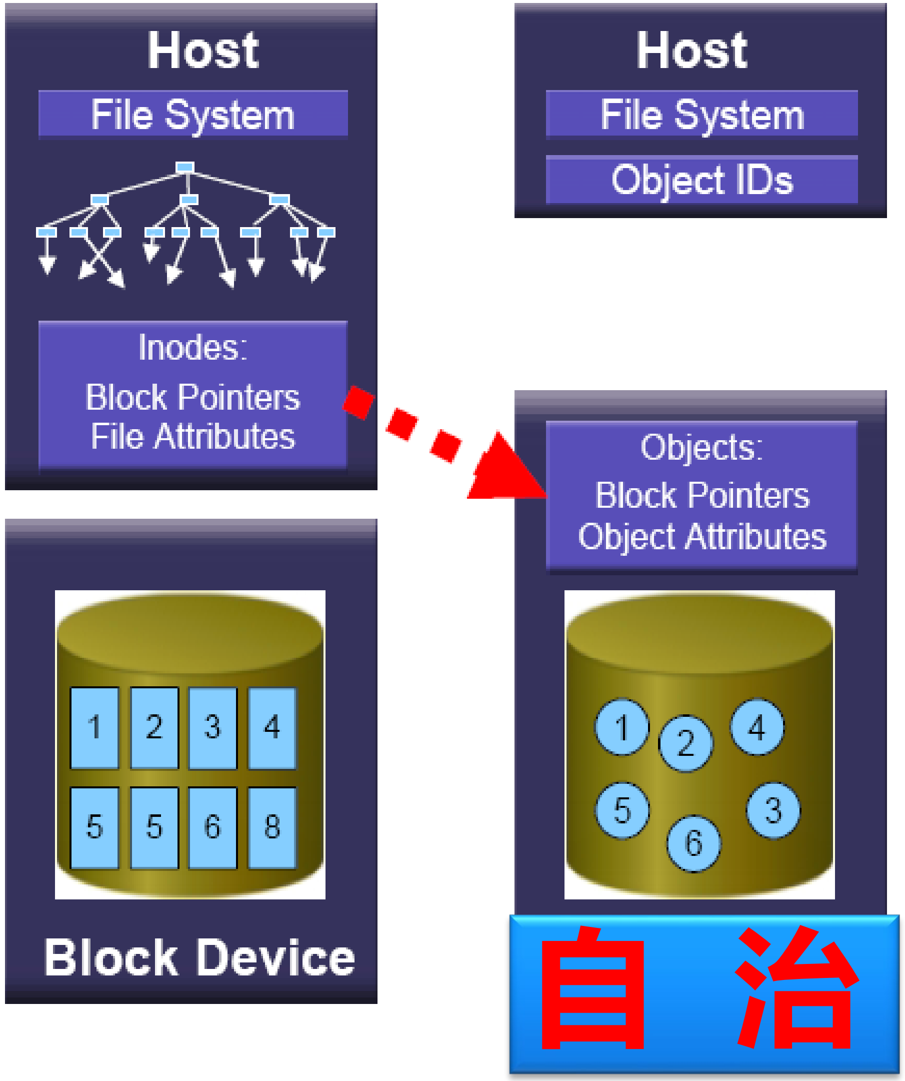


Inodes vs. Objects



- ➢ Abstract some of the lower layers of storage away from the administrators and applications.
- ➢ Inclusion of rich custom metadata **within** the Object.
  - ✓ specific information(from user or app.) for better indexing purposes
  - ✓ Support data-management policies
  - ✓ Centralize management of storage across many individual nodes and clusters
  - ✓ Optimize metadata storage and caching/indexing independently from the data storage

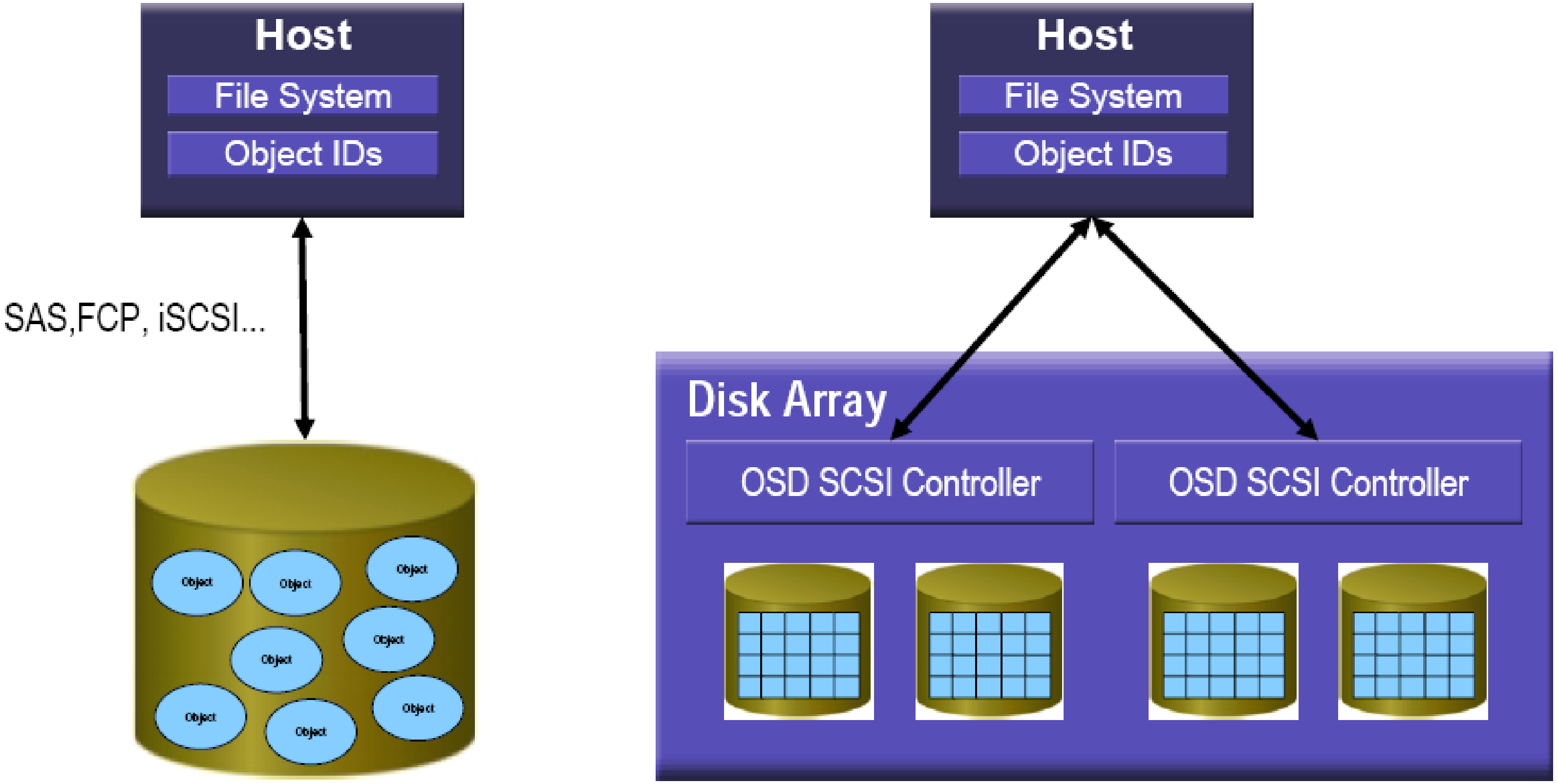
Object Autonomy



Storage becomes autonomous:

–capacity planning
–load balancing
–backup
–QoS, SLAs
–understand data/object grouping
–aggressive pre-fetching
–thin provisioning
–search
–compression/de-duplication/encryption
–strong security
–compliance/retention/secure delete
–availability/replication
–audit
.
.
.

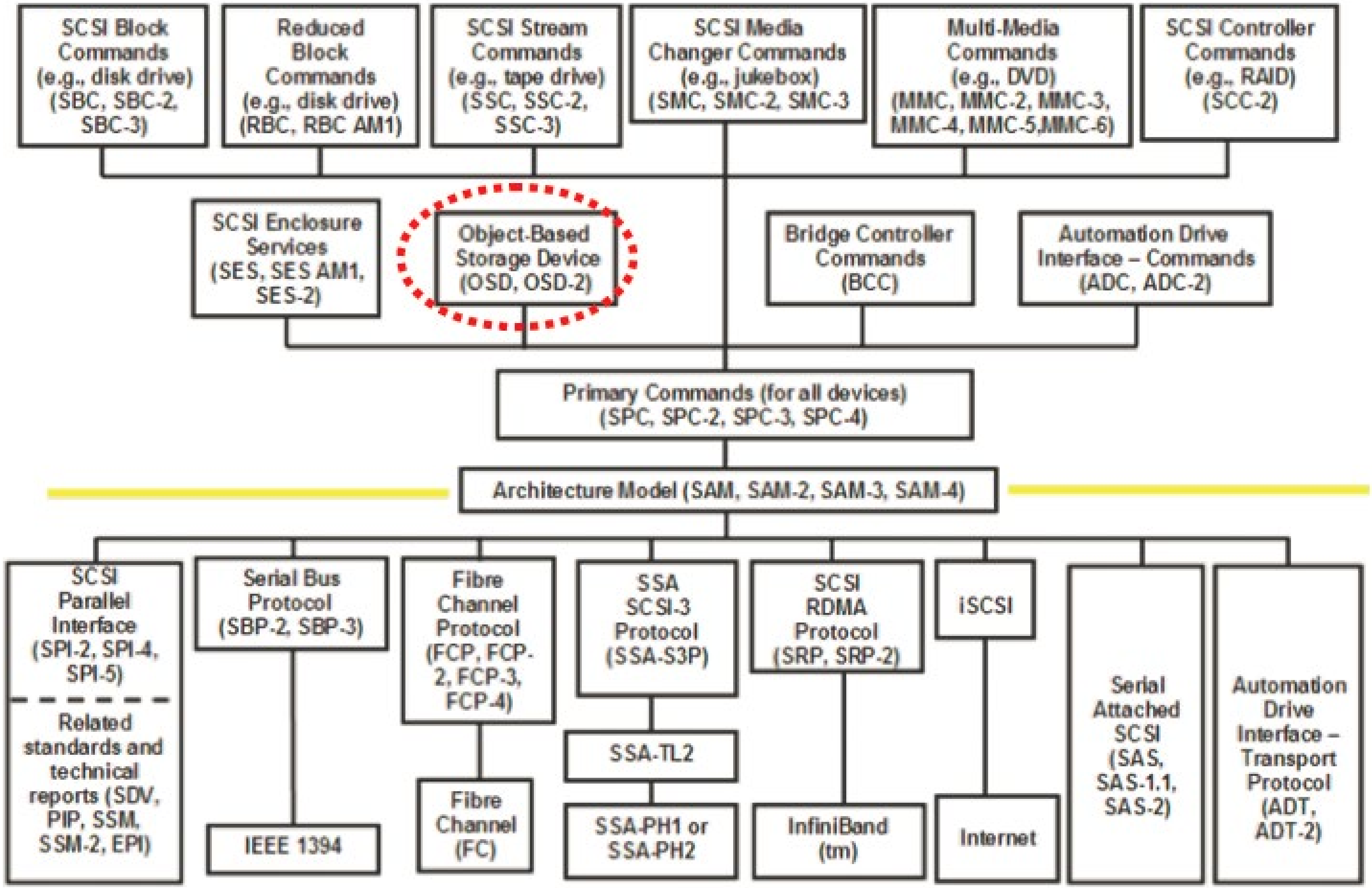
ANSI T10 OSD SCSI Targets



对象存储是一种将数据作为对象进行管理的计算机数据存储体系结构。与文件存储和块存储不同，每个对象通常包括**数据本身**，数量可变的**元数据**和**全局唯一标识符**。对象存储可以在多个级别上实现，包括**设备级别**，系统级别和接口级别。

SCSI Standards Architecture

设备级
对象存储



SCSI Block Commands (e.g., disk drive) (SBC, SBC-2, SBC-3)

Reduced Block Commands (e.g., disk drive) (RBC, RBC AM1)

SCSI Stream Commands (e.g., tape drive) (SSC, SSC-2, SSC-3)

SCSI Media Changer Commands (e.g., jukebox) (SMC, SMC-2, SMC-3

Multi-Media Commands (e.g., DVD) (MMC, MMC-2, MMC-3, MMC-4, MMC-5,MMC-6)

SCSI Controller Commands (e.g., RAID) (SCC-2)

SCSI Enclosure Services (SES, SES AM1, SES-2)

Object-Based Storage Device (OSD, OSD-2)

Bridge Controller Commands (BCC)

Automation Drive Interface – Commands (ADC, ADC-2)

Primary Commands (for all devices) (SPC, SPC-2, SPC-3, SPC-4)

Architecture Model (SAM, SAM-2, SAM-3, SAM-4)

SCSI Parallel Interface (SPI-2, SPI-4, SPI-5)

Related standards and technical reports (SDV, PIP, SSM, SSM-2, EPI)

Serial Bus Protocol (SBP-2, SBP-3)

IEEE 1394

Fibre Channel Protocol (FCP, FCP-2, FCP-3, FCP-4)

Fibre Channel (FC)

SSA SCSI-3 Protocol (SSA-S3P)

SSA-TL2

SSA-PH1 or SSA-PH2

SCSI RDMA Protocol (SRP, SRP-2)

InfiniBand (tm)

iSCSI

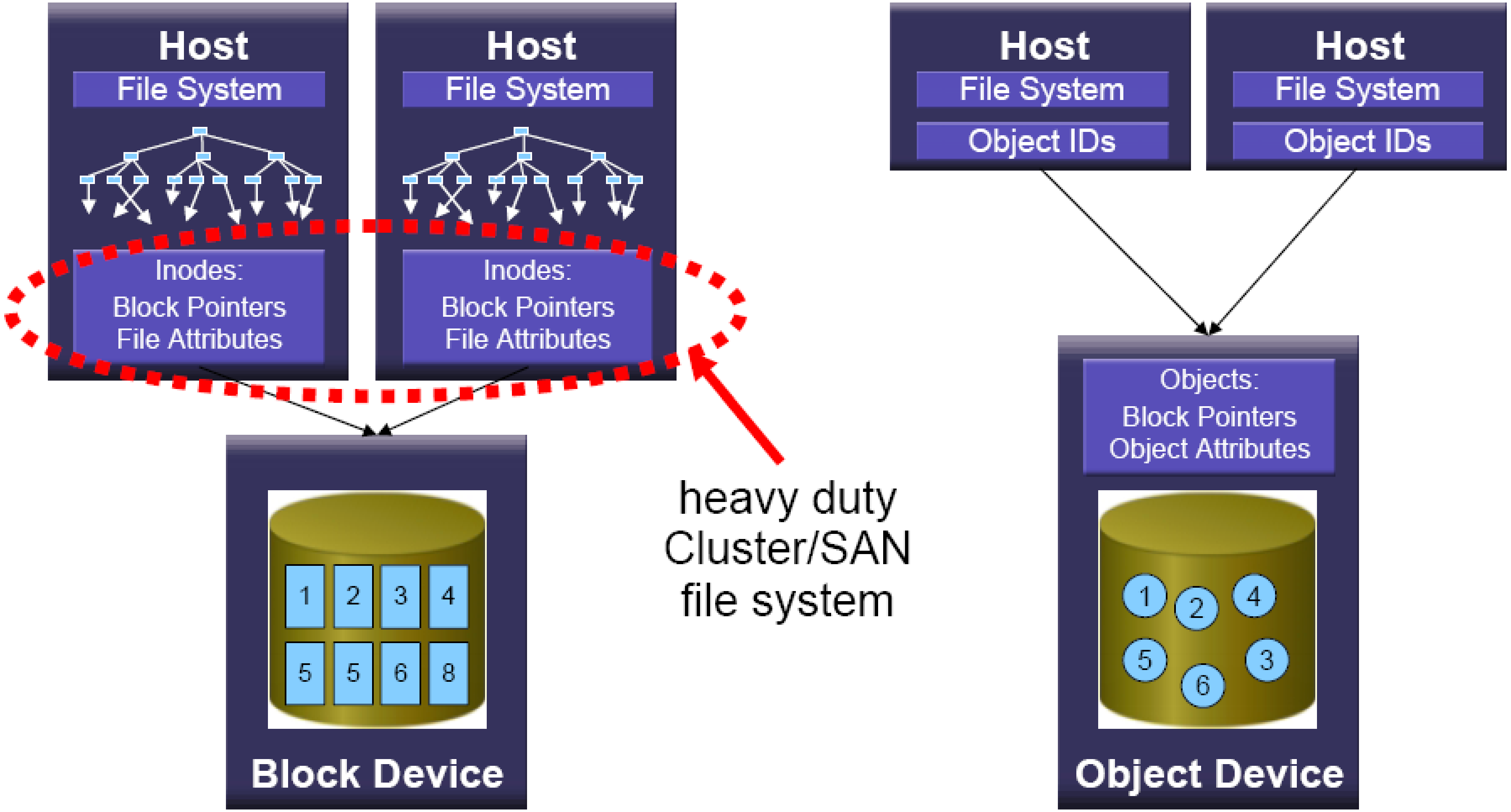
Internet

Serial Attached SCSI (SAS, SAS-1.1, SAS-2)

Automation Drive Interface – Transport Protocol (ADT, ADT-2)

Data Sharing

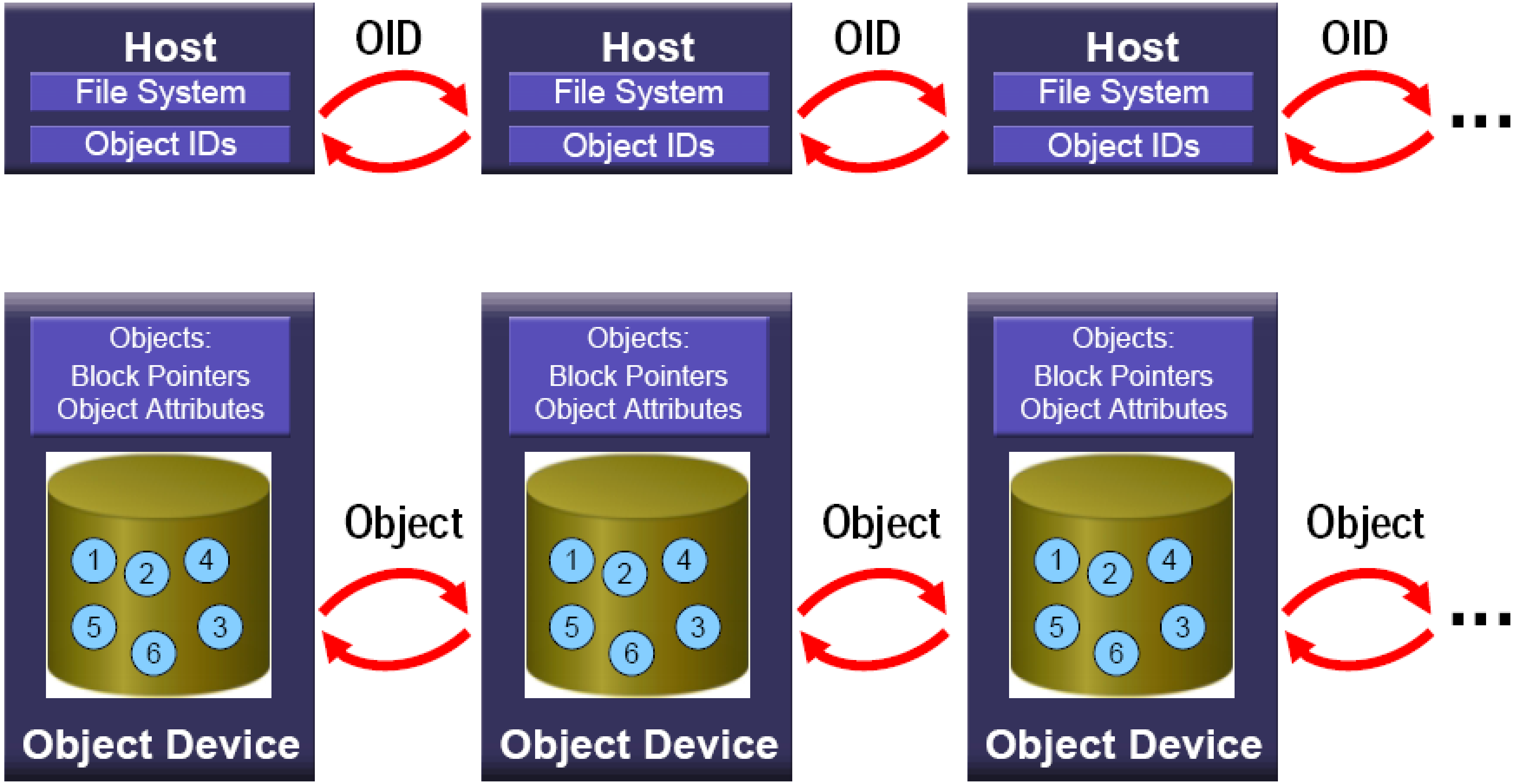
Homogeneous/Heterogeneous



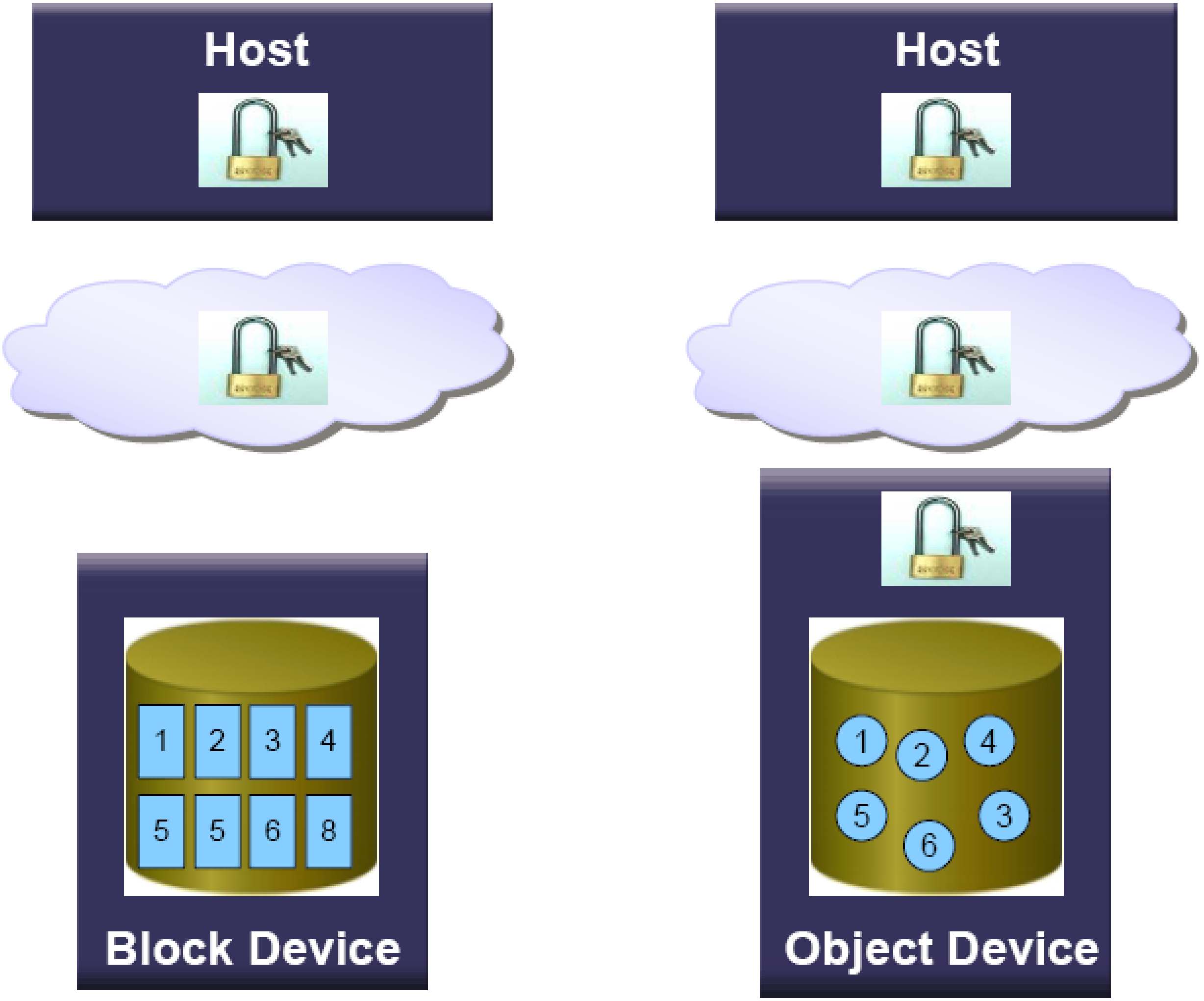
heavy duty
Cluster/SAN
file system

Data Migration - ILM

Homogeneous/Heterogeneous



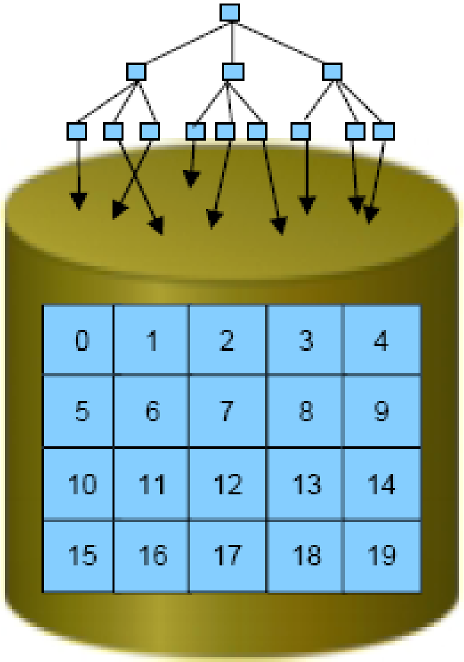
Additional Layer of Security



- strong security via external service
  - –authentication
  - –authorization
  - ....
- fine granularity
  - –per object

Living in a Flat Namespace

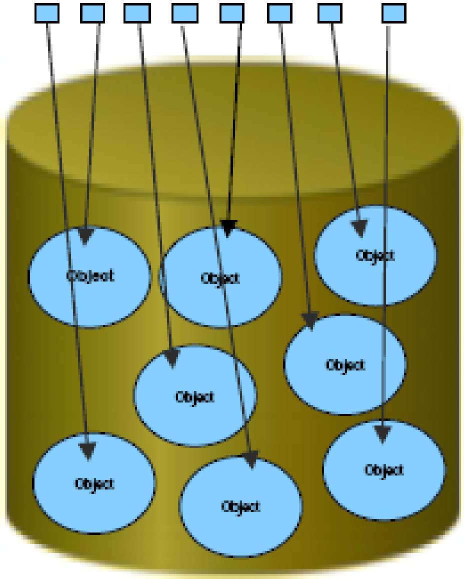
File names / inodes



0	1	2	3	4
5	6	7	8	9
10	11	12	13	14
15	16	17	18	19

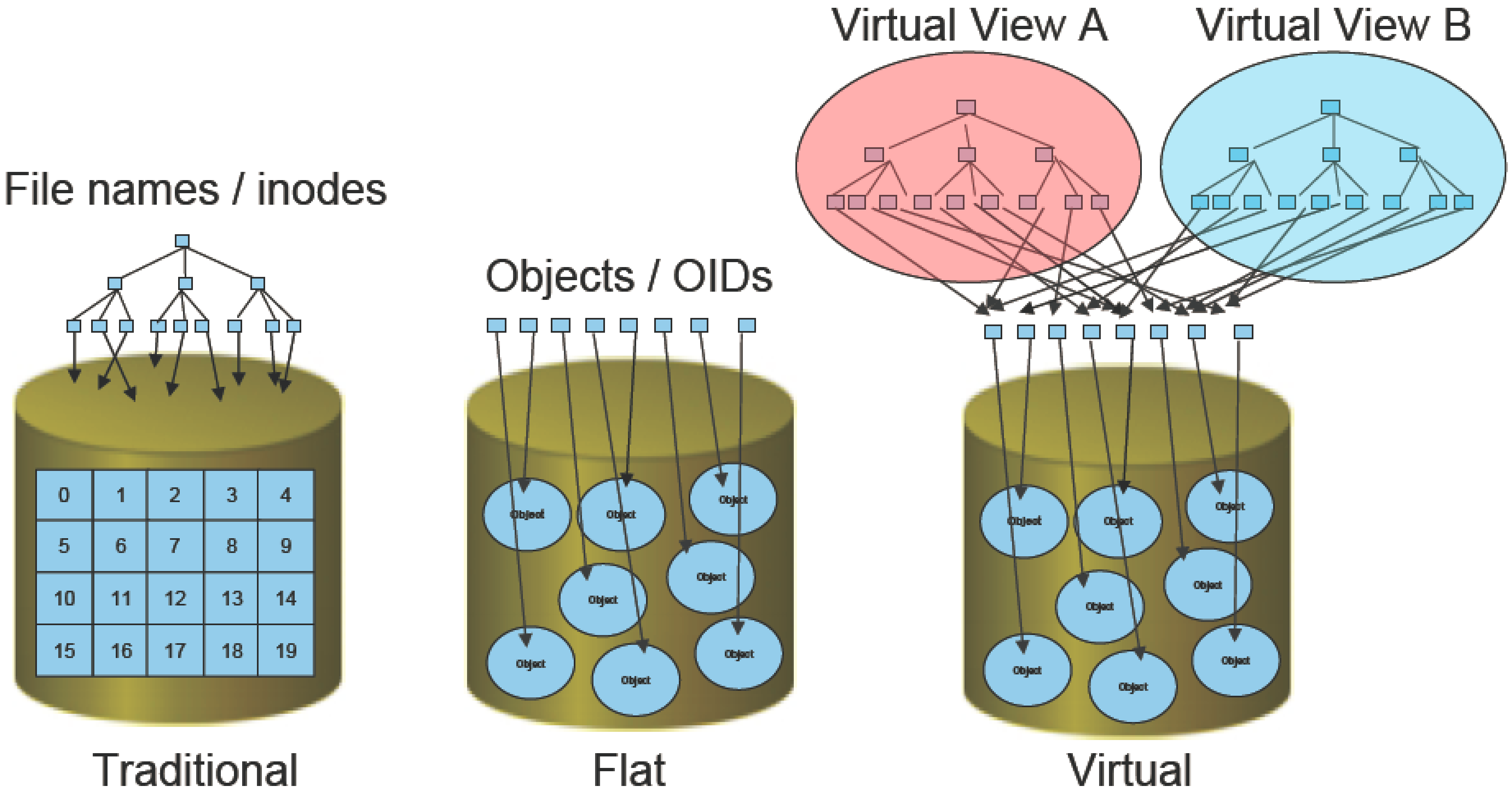
Traditional
Hierarchical

Objects / OIDs

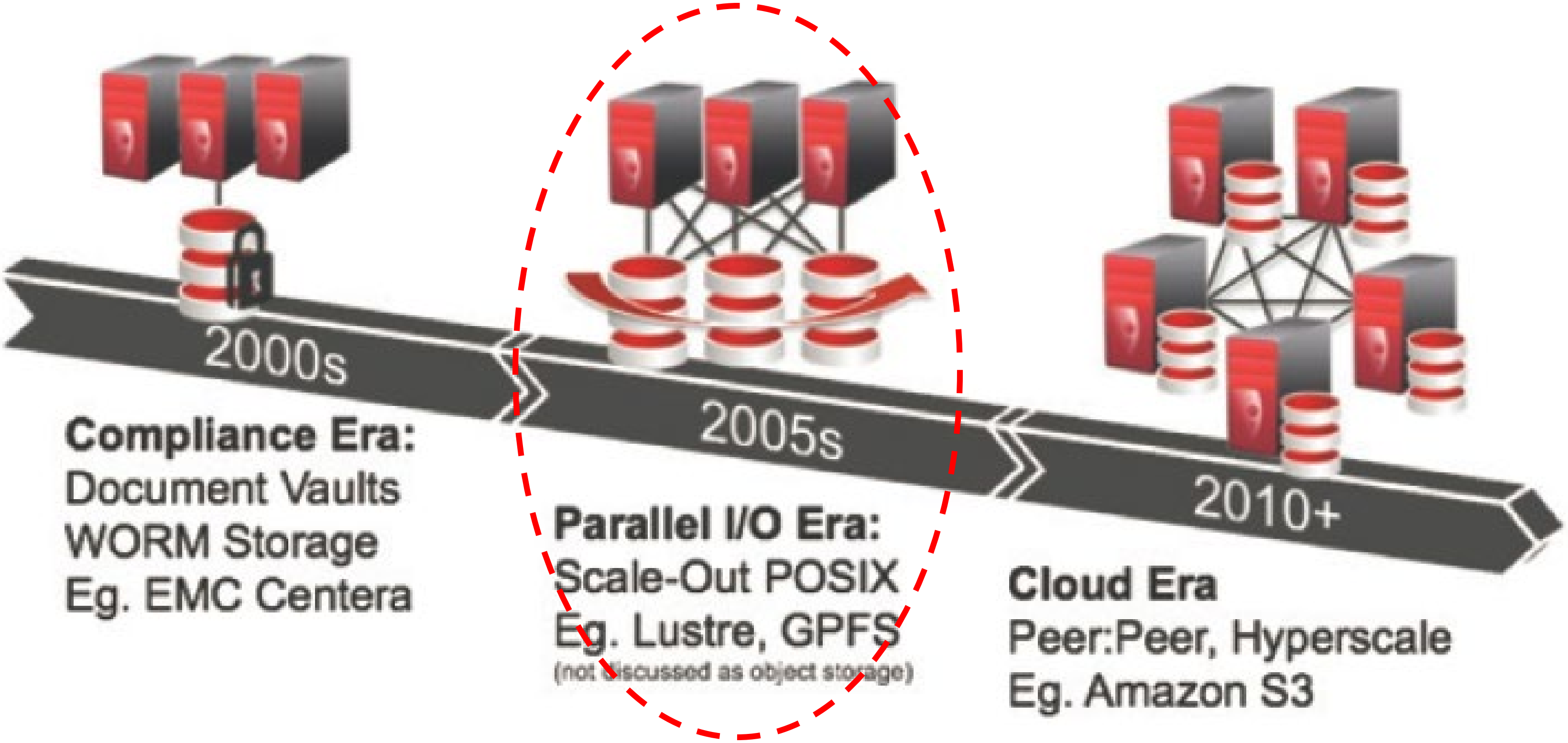


Flat

Virtual View / Virtual File Systems



The First Generation of Object Storage（Device-level）



2000s

Compliance Era:
Document Vaults
WORM Storage
Eg. EMC Centera

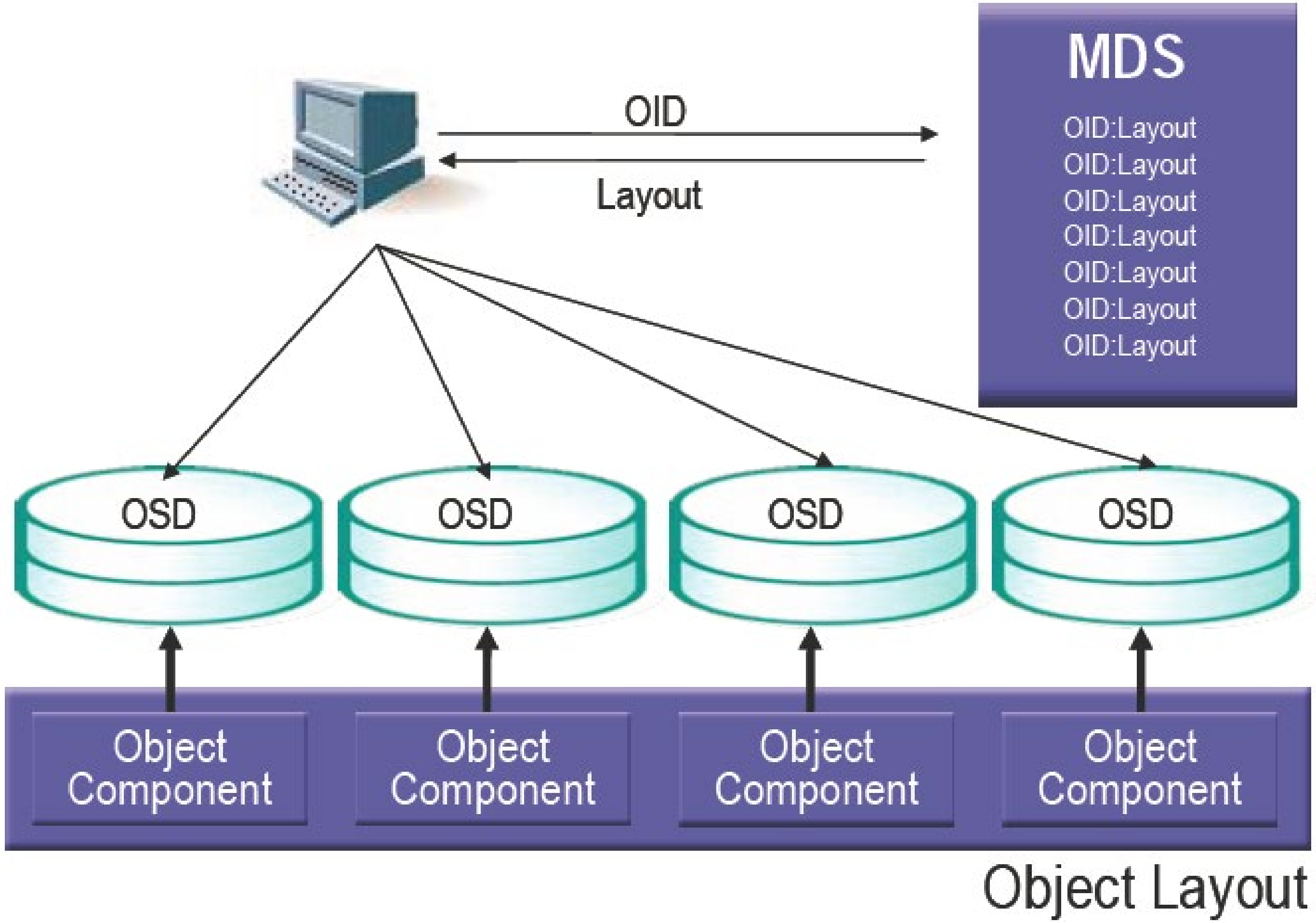
2005s

Parallel I/O Era:
Scale-Out POSIX
Eg. Lustre, GPFS
(not discussed as object storage)

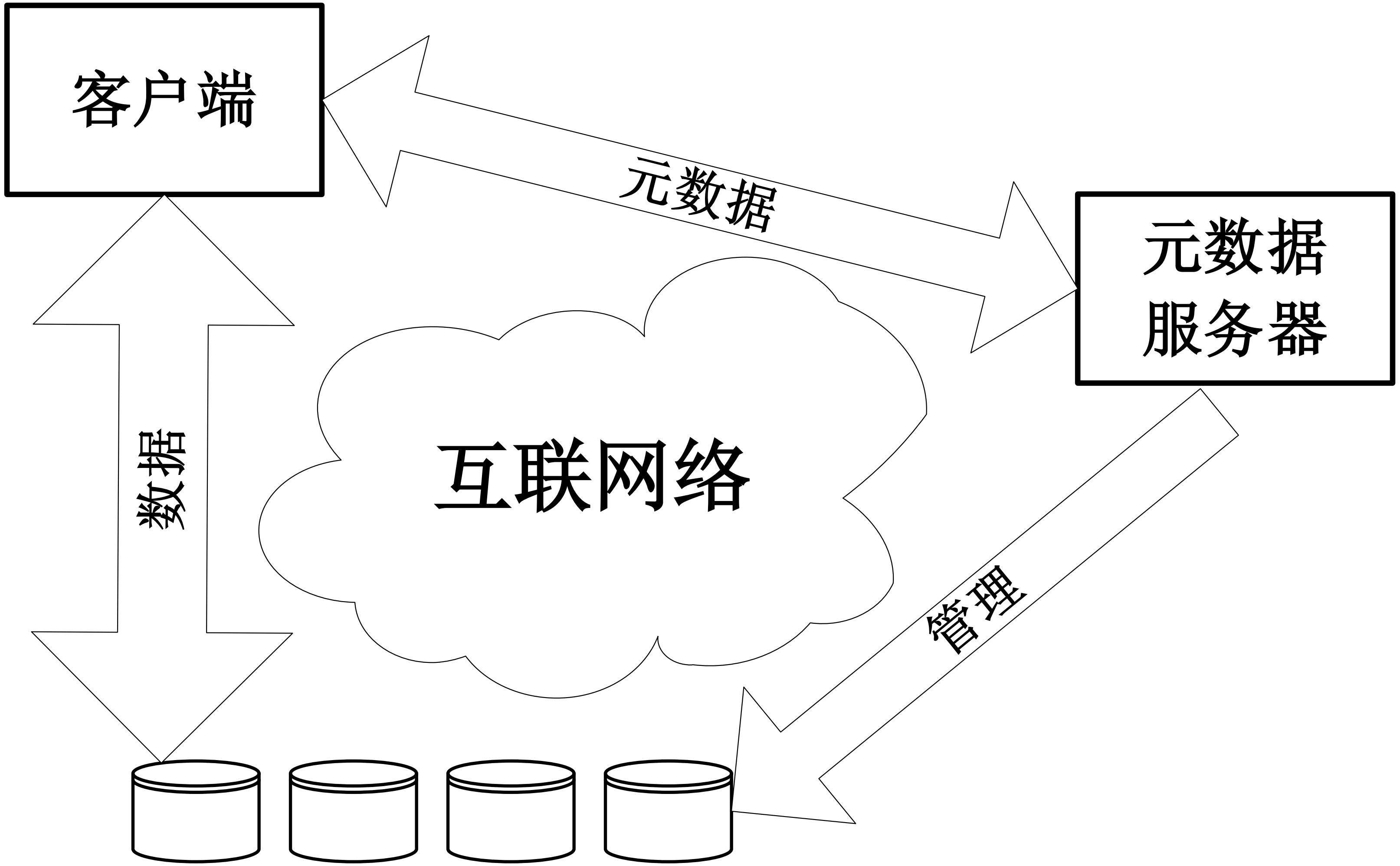
2010+

Cloud Era
Peer:Peer, Hyperscale
Eg. Amazon S3

Object Decomposition

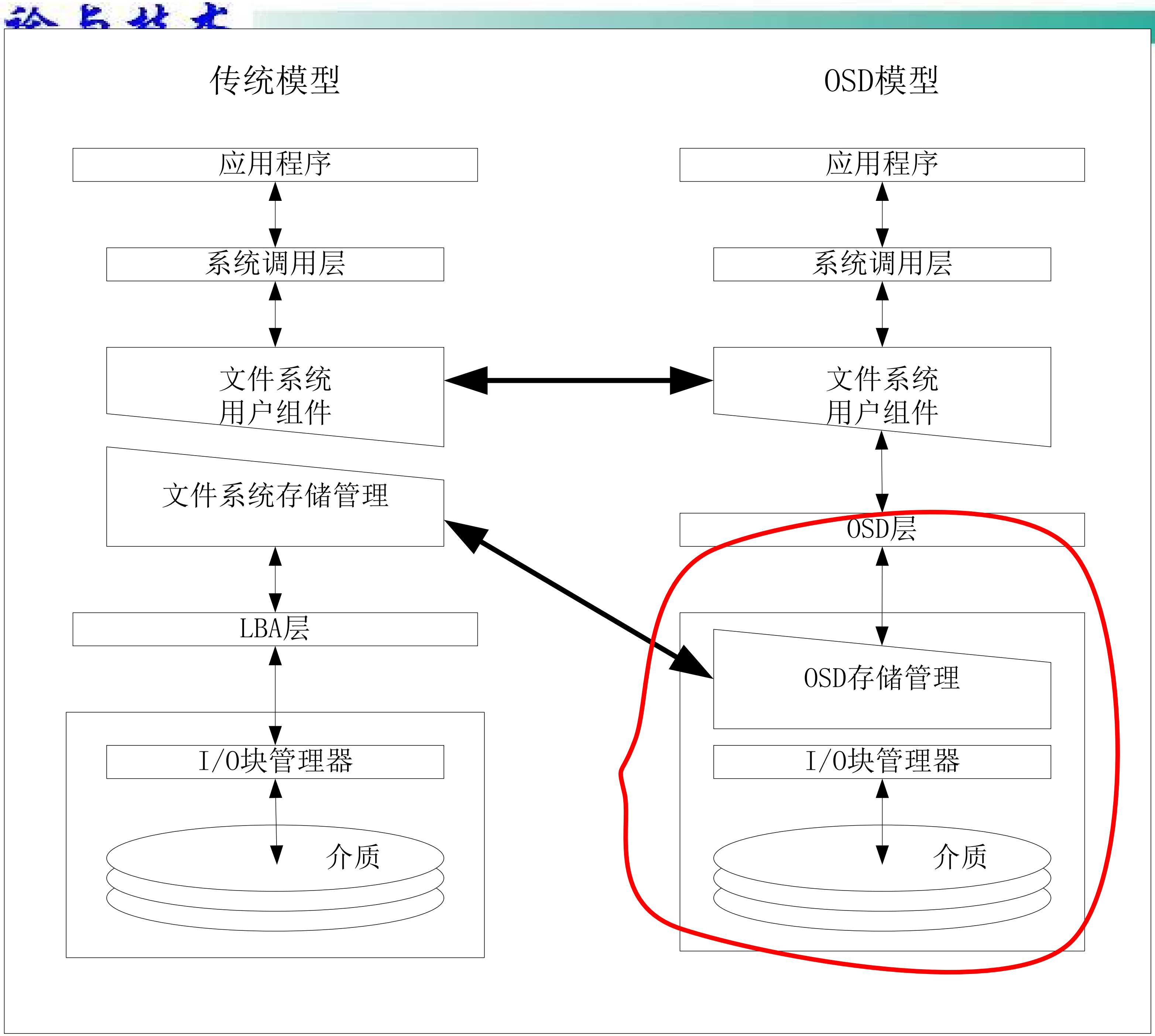


对象存储系统结构



对象存储系统组成

- **对象(Object)**
  - 包含了文件数据以及相关的属性信息，可以进行自我管理
- **OSD（Object-based Storage Device)**
  - 一个智能存储设备，是Object的集合
- **文件系统**
  - 文件系统运行在客户端上，将应用程序的文件系统请求传输到MDS和OSD上
- **元数据服务器(Metadata Server，MDS)**
  - 系统提供元数据、Cache一致性等服务
- **网络连接**


传统模型
OSD模型
应用程序
系统调用层
文件系统
用户组件
文件系统存储管理
LBA层
I/O块管理器
介质
应用程序
系统调用层
文件系统
用户组件
OSD层
OSD存储管理
I/O块管理器
介质

对象存储与传统存储的对比

	存储接口	存储系统	优点	缺点
块级存储	块	块存储设备	提供高性能的随机I/O和数据吞吐率，如：SAN	可扩展性和可管理性较差、价格较高、不能满足成千上万CPU 规模的系统
文件储存	文件	块存储设备 +文件系统	扩展性好、易于管理、价格便宜，如：NAS	开销高、带宽低、延迟大，不利于高性能集群中应用
对象存储	对象	块存储设备 +文件系统 +定位逻辑 +应用程序	支持高并行性、可伸缩的数据访问， 管理性好、安全性高、适合高性能集群使用	处于发展阶段，相应的硬件、软件支持有待进一步完善

对象存储的特性（总结）

- 性能优势
- 存储设备的智能化
- 数据的共享更容易
- 管理更方便
- 更好的安全性

是大数据存储理想的选择

对象存储系统实例：Lustre

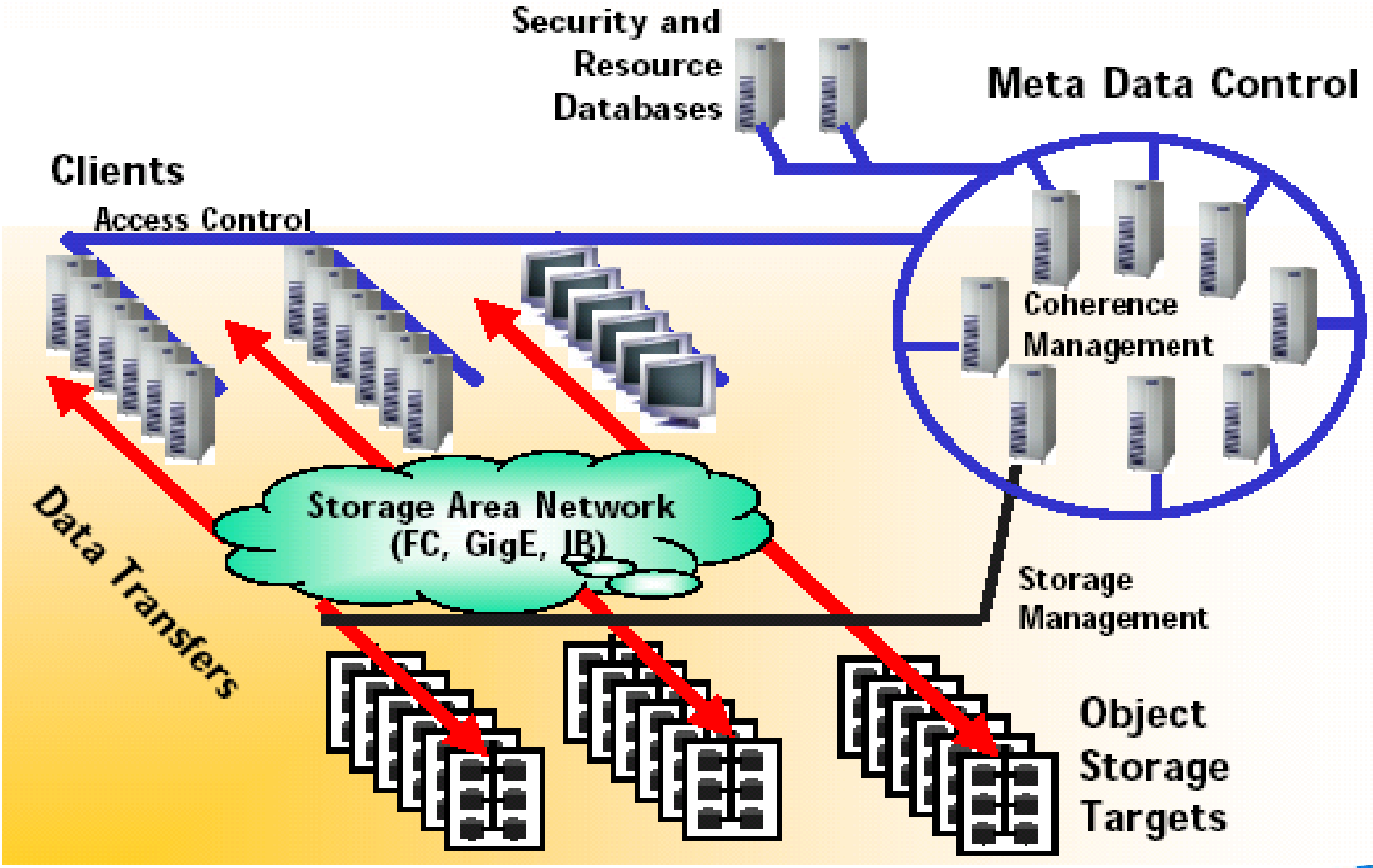
- A shared file system for HPC clusters
  - Open Source software (GPL)
  - linux cluster
- Very high metadata and I/O performance
  - 5,000 file creations/sec in 1 dir, 1,000 nodes
  - Single clients up to 290MB/sec.
  - Aggregate up to 11GB/sec
- Scalable to 1,000’s of nodes
- In production now on such clusters

Lustre Retrospective

- 1999 Initial ideas @CMU
- Seagate: management aspects, prototypes
  - Much survives today
- 2000 National Labs
  - Can Lustre be next generation FS?

    100 GB/sec, trillion files, 10,000’s clients, secure, PBs
- 2002 – 2003
  - Many partners: Dell, HP, Cray, LNXI, DDN others
  - Production use, 1.0 released
- April 3rd, 2018, Lustre 2.11.0 released

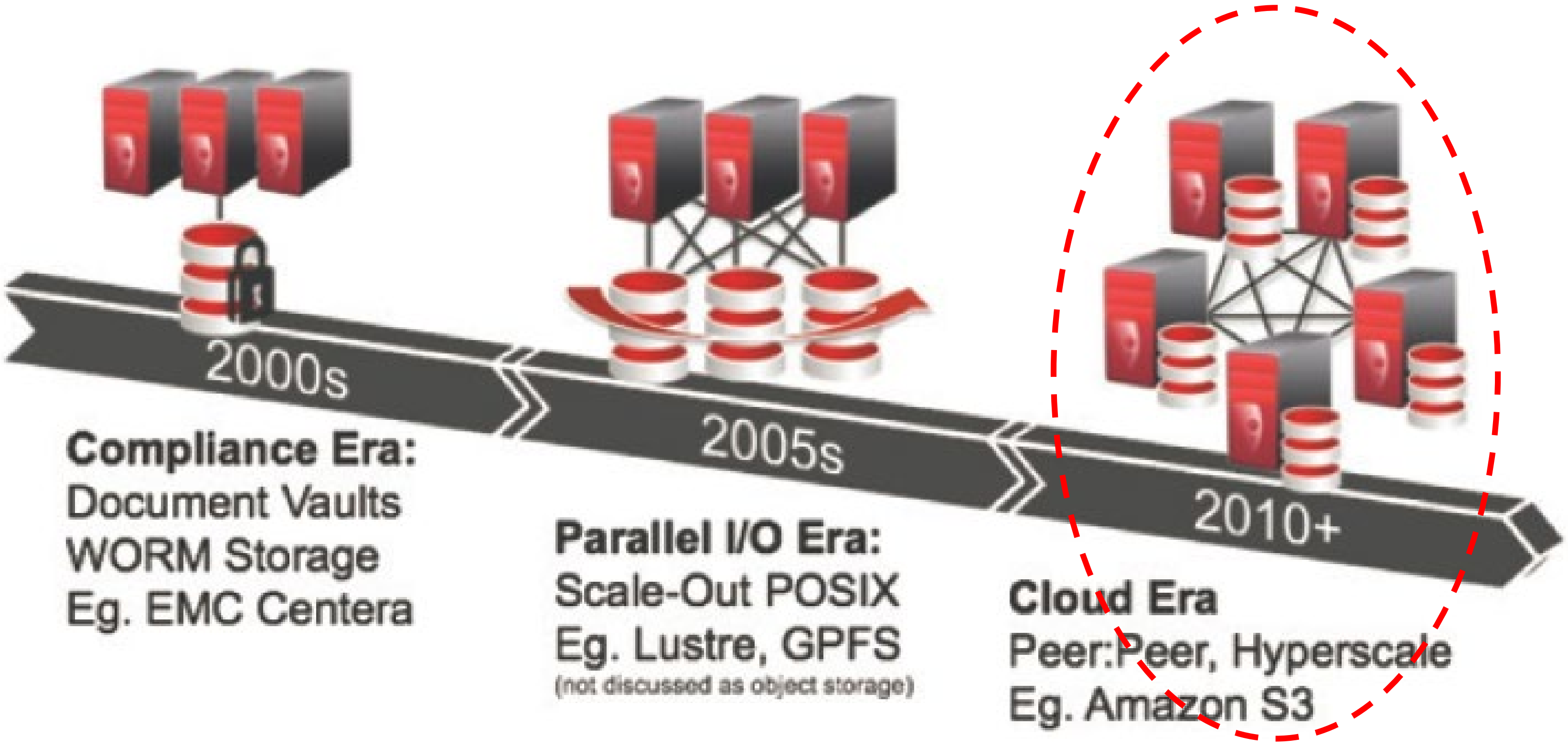
A Lustre Cluster



Reference

- Lustre: A SAN File System for Linux
  - http://lustre.org/documentation/
- Several presentation materials from Dr. Peter J. Braam

The Current Generation of Object Storage



Compliance Era:
Document Vaults
WORM Storage
Eg. EMC Centera

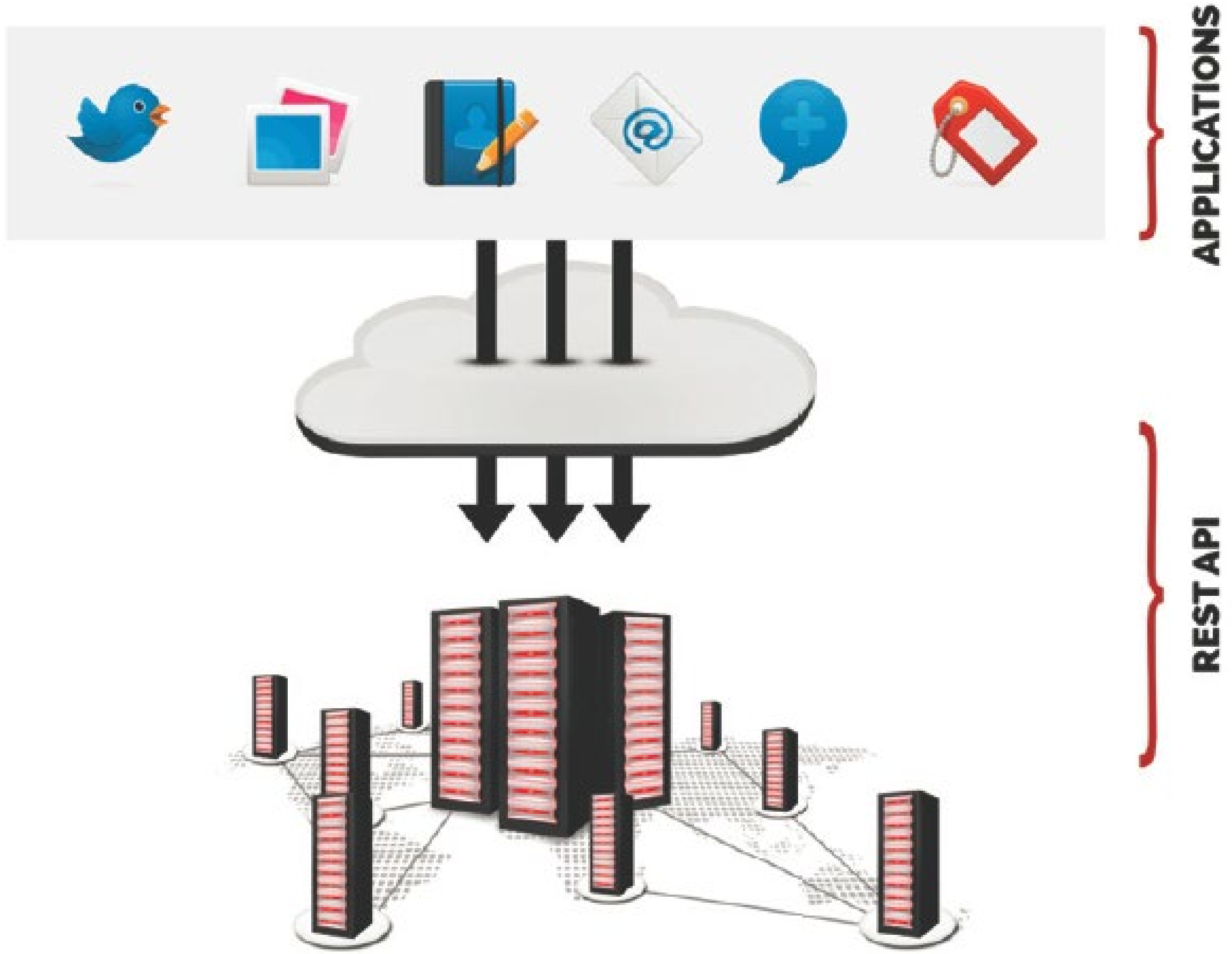
Parallel I/O Era:
Scale-Out POSIX
Eg. Lustre, GPFS
(not discussed as object storage)

Cloud Era
Peer:Peer, Hyperscale
Eg. Amazon S3

Cloud Storage, Storage Clouds, Object Storage

- **Cloud Storage** is the storage used for Compute Cloud infrastructures
  - Compute Clouds are very IOPS intensive and usually **block storage** is used in these applications
- **Storage Clouds** are “storage in the cloud”, whether public or private
  - Storage Clouds are simply storage capacity that is made available through the Internet
  - Most of today’s storage clouds use **object storage** technologies

Scale out object storage with simple REST API



REST API's

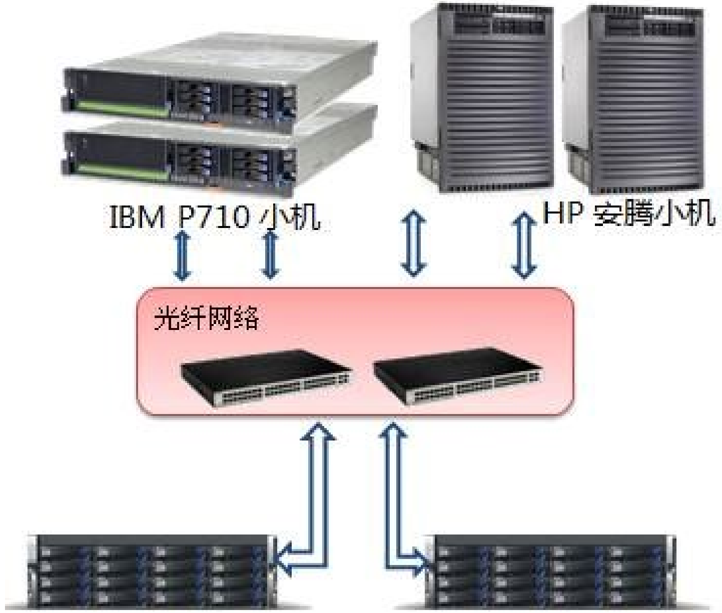
- REST stands for Representational State Transfer
- It is a software architecture that is used for distributed application environments
- REST API's have become the predominant interface for cloud applications to connect to the cloud
- For storage-centric cloud applications, a REST API is the interface between the application and the object storage platform
- PUT GET DELETE…

Current Object Storage Summary

- Data is stored as objects in one large, scalable pool of storage
- Objects are stored with metadata – information about the object
- An Object ID is stored, to locate the data
- REST is the standard interface, simple commands used by applications
- Objects are immutable; edits are saved as a new object

块、文件和对象

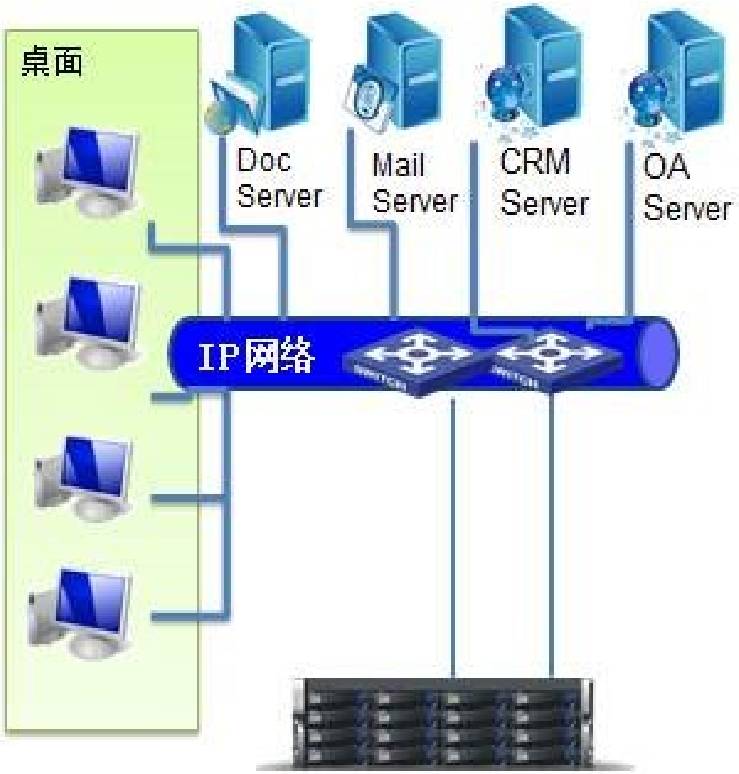
- **块存储**（DAS/SAN）
  - FC、iSCSI协议
  - 专有的系统中，高读写性能和高可靠性
  - 一套存储只服务一个应用系统，例如如交易系统，计费系统。典型行业如金融，制造，能源，电信等



块、文件和对象

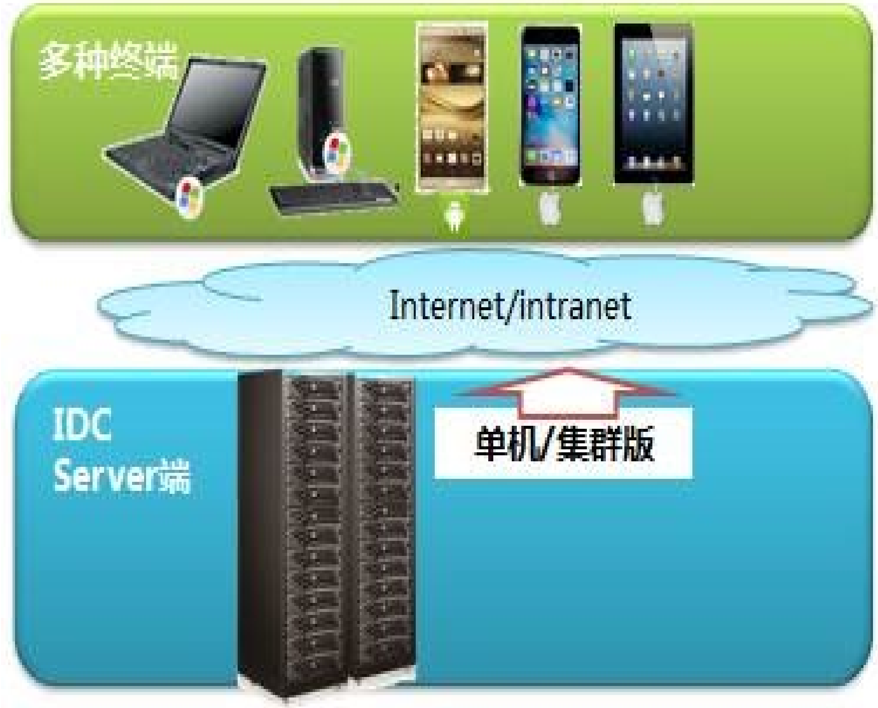
□ 文件存储（NAS）

- ✓ NFS/CIFS协议，IP网络
- ✓ 兼顾多个应用和更多用户访问，同时提供方便的数据共享手段
- ✓ 中小企业市场，CRM系统，SCM系统，OA系统等



块、文件和对象

- ◆ **对象存储**（Object）
  - ➢ OSD，HTTP协议
  - ➢ 互联网或者公网
  - ➢ 海量数据，高并发访问
  - ➢ 常见的适配应用如网盘、媒体娱乐，医疗PACS，气象，归档等数据量超大而又相对“冷数据”和非在线处理的应用类型



块、文件和对象

块存储

文件存储

对象存储

Throughput easy, latency hard

Throughput is easy

Latency is hard

Throughput is an engineering problem, latency is a physics problem!

对象存储与键值存储

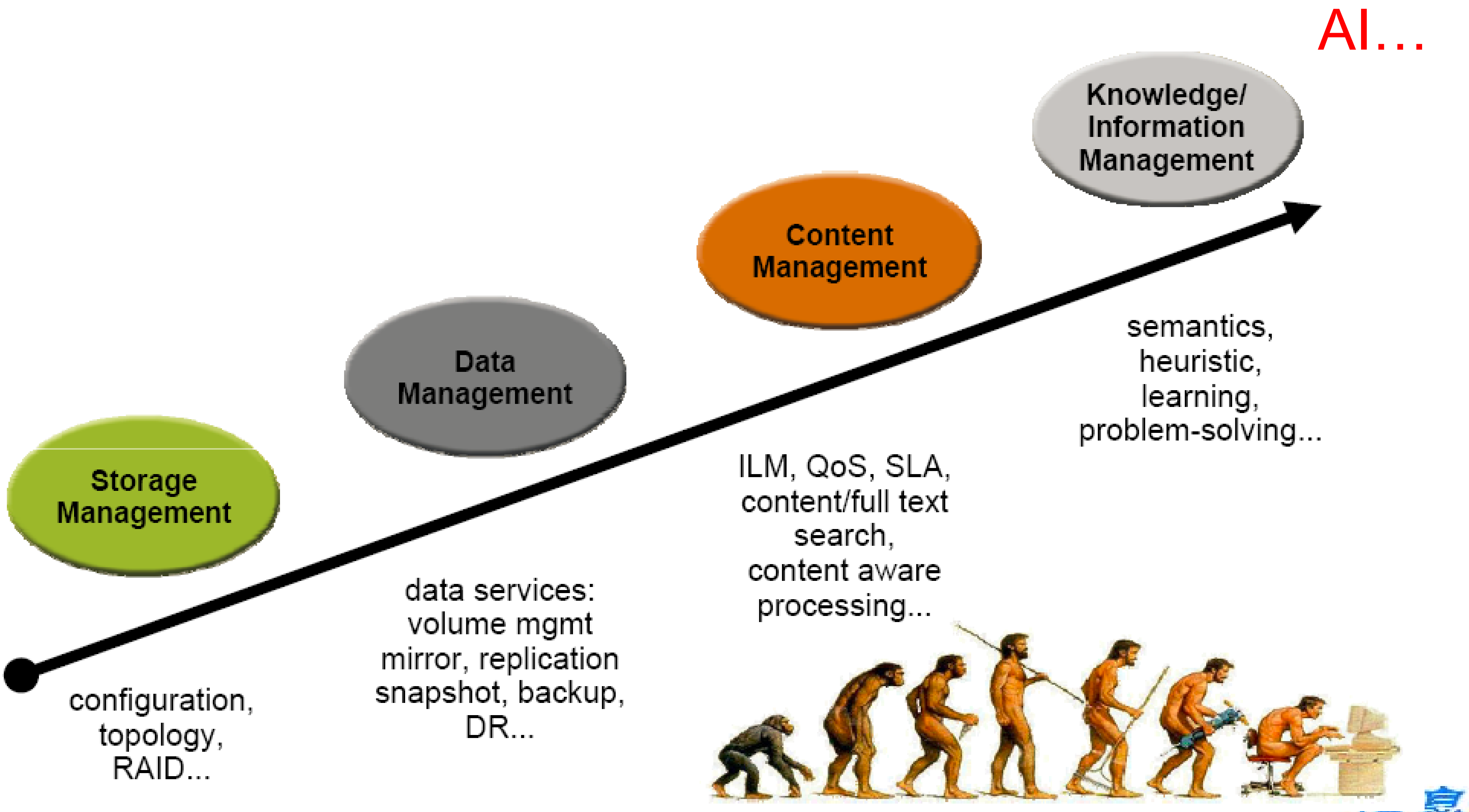
◆ 相同点

- ➢ 对象ID 类似于Key（任意字符串）
- ➢ 数据可以具有任意大小

◆ 不同点

- ➢ 对象存储还允许将一组有限的属性（元数据）与每个(数据)对象相关联
- ➢ 对象存储针对大量数据（数百兆甚至千兆字节）进行了优化，而对于键值存储，其值则相对较小（千字节）
- ➢ 对象存储通常提供较弱的一致性保证，例如最终的一致性，而键值存储则提供较强的一致性。

The Evolution of Data Processing



AI…

Knowledge/ Information Management

semantics, heuristic, learning, problem-solving…

Content Management

ILM, QoS, SLA, content/full text search, content aware processing…

Data Management

data services: volume mgmt mirror, replication snapshot, backup, DR…

Storage Management

configuration, topology, RAID…

Moving Data to the Processor is Costly

One floating-point calculation

17 picojoules

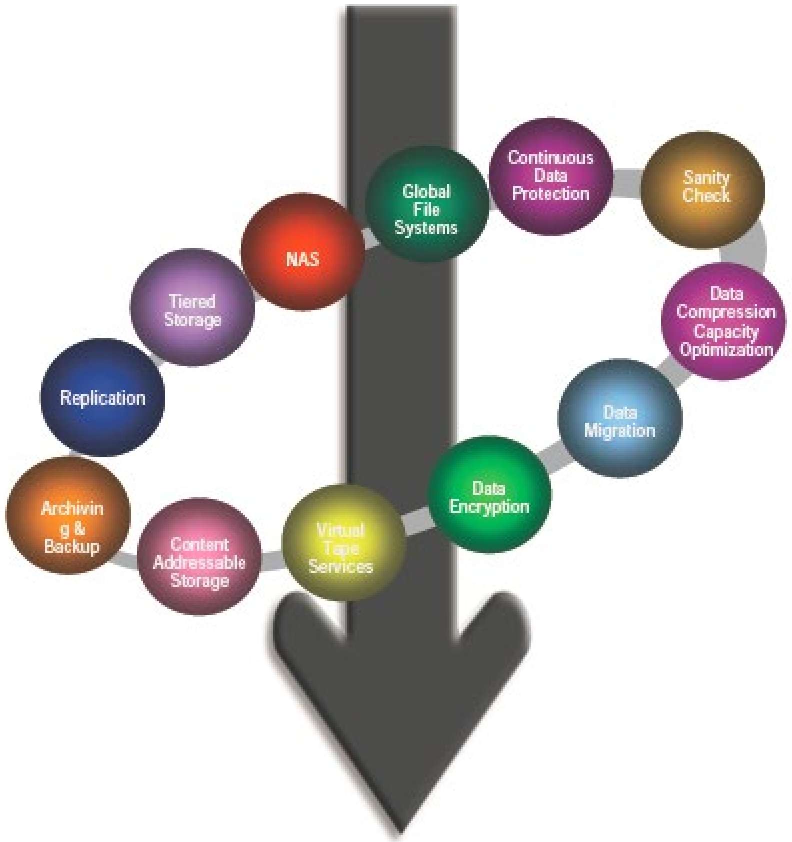
Moving data from DRAM to CPU

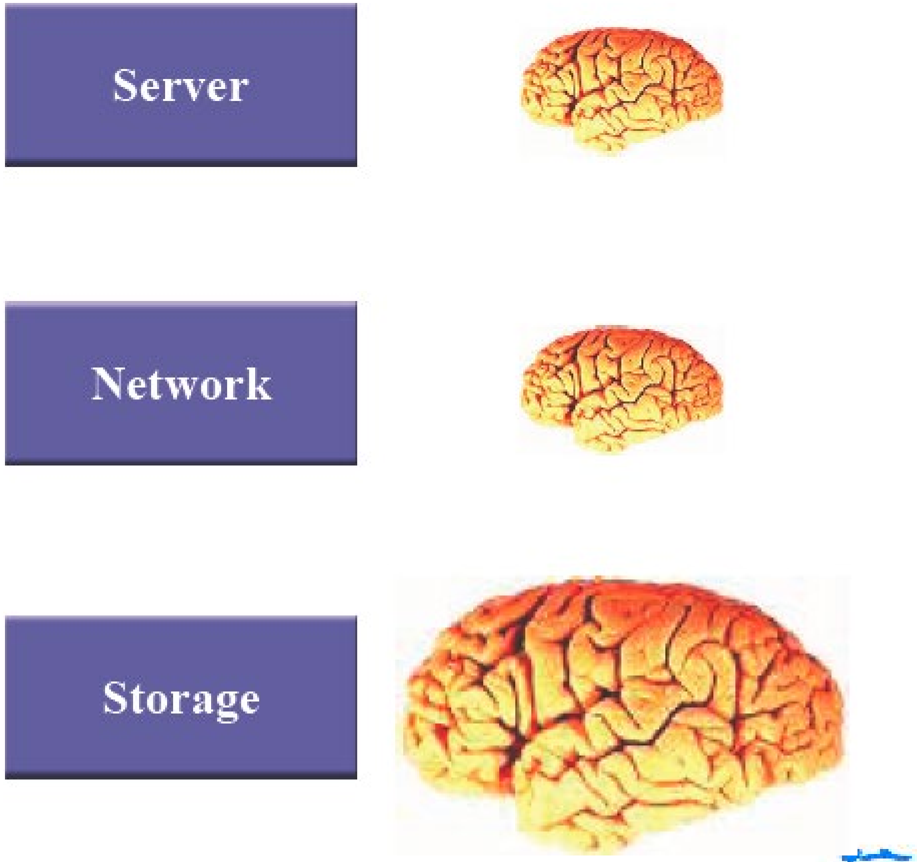
17,000 picojoules

Opportunities for **1000x improvement** are increasingly rare

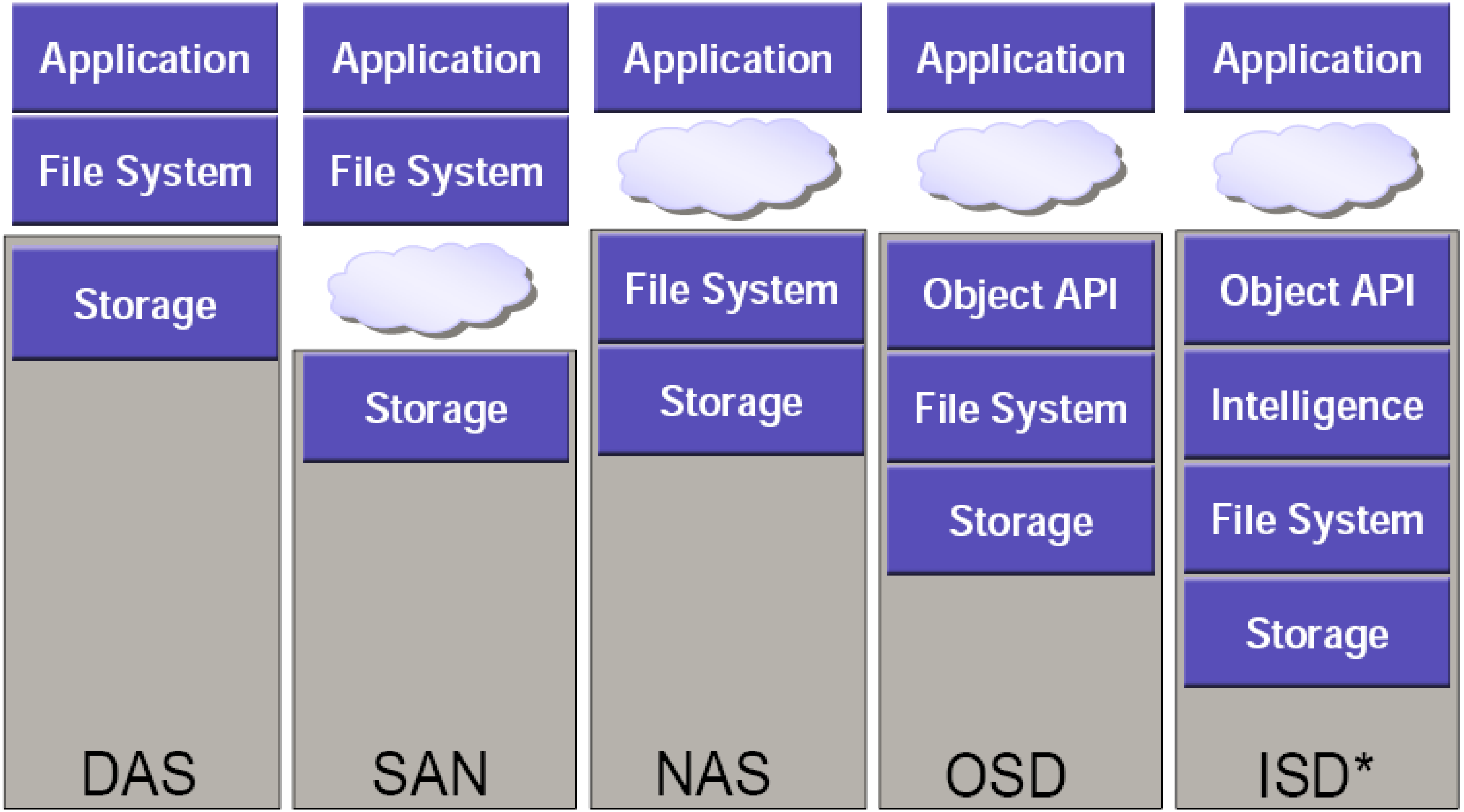
Migration of Storage Application

- Process the data where it lives...



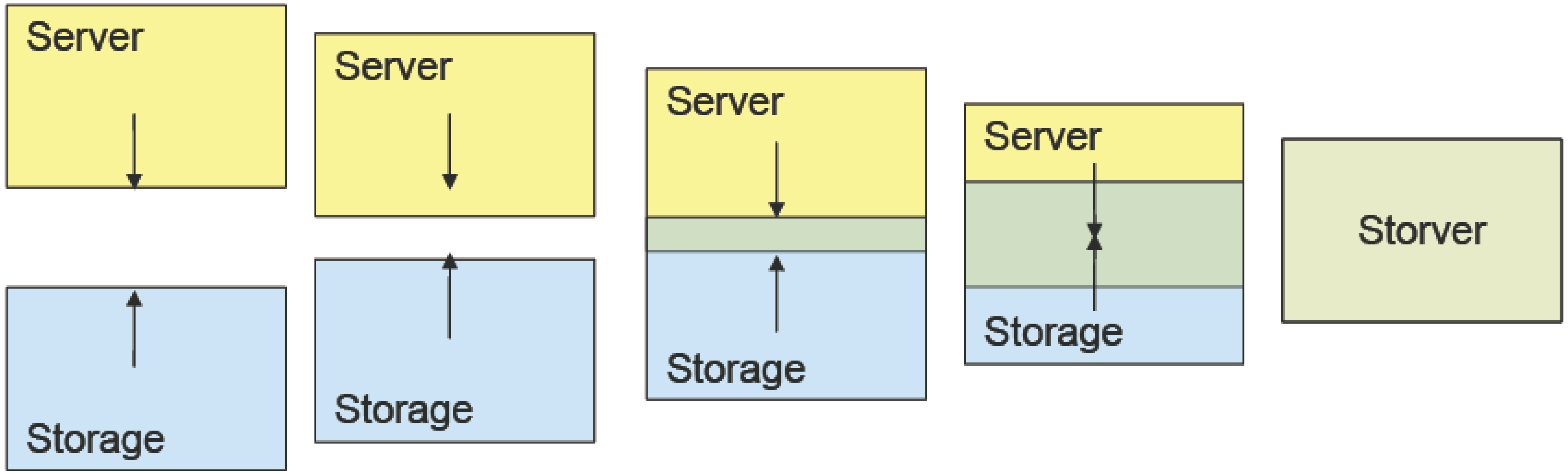


The Evolution of Storage



The Vertical Consolidation

- Storage and server
- Migration of data processing applications
- **No I/O is best I/O**



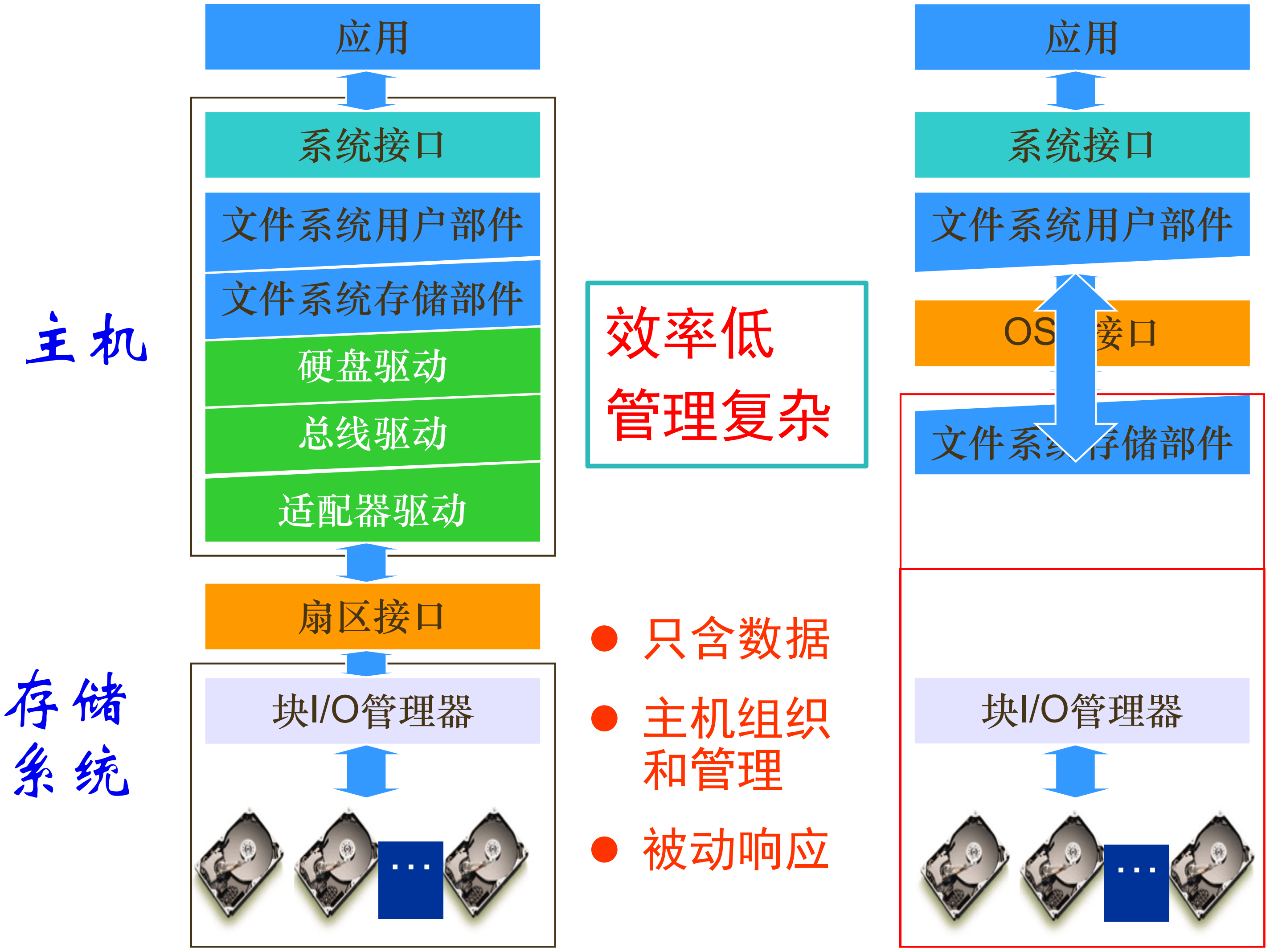
主动对象存储

- ◆传统存储系统被动响应服务请求
- ◆对象具有智能性

智能的系统能够提供主动服务

存储组织结构对比

传统模式　　　　主动对象模式



效率低
管理复杂

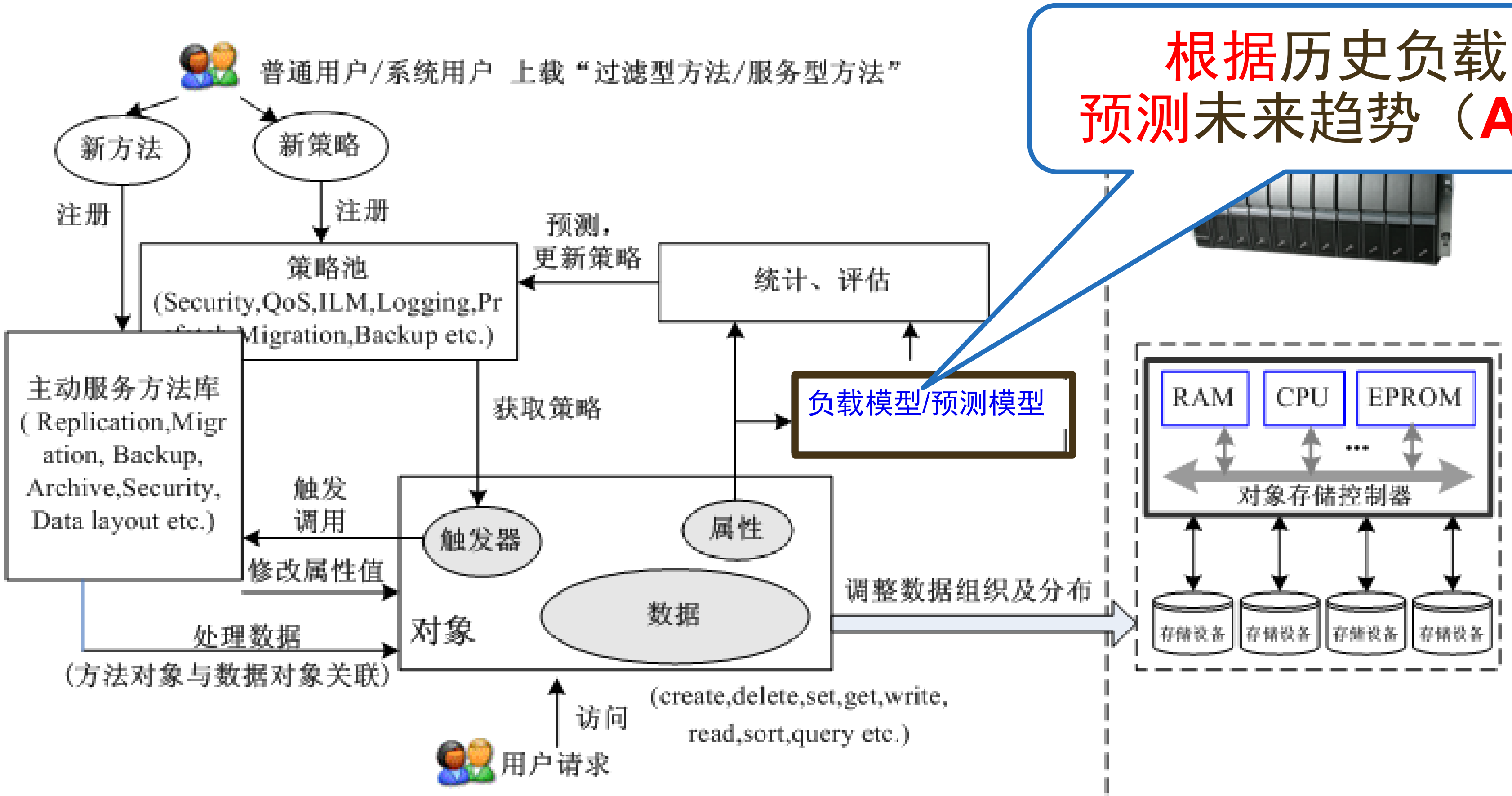
- 只含数据
- 主机组织和管理
- 被动响应

性能高
管理简单

- 含“数据、属性”和“**操作**”
- 自组织和自管理
- 主动服务

主动对象存储服务机制

根据历史负载
预测未来趋势（AI）



存储主动服务

存储节点