对象存储系统专题

施展 2019

课程信息

- 专题内容
  - 关于对象存储系统的研究
    - 一些基础
      - https://github.com/cs-course/iot-storage-experiment *物联网3年级下学期*
      - https://github.com/cs-course/obs-tutorial
    - 需要研究的问题
  - 关于研究的研究
    - 研究的基本要求和方法学习

课程信息

数据中心技术2019
扫一扫二维码，加入该群。

- 选课同学请实名入群
- 完成一个小投票
  - 本科学习过
    - 操作系统
    - 系统结构
    - 物联网数据存储实验
    - 均未学过

授课目标

- 研究对象
  - 对象存储系统
    - 理解经典存储系统的设计
- 研究者
  - 基础研究能力
    - 明白未来研究的主要任务

新手上路

- 内心里
  - 兴(MENG)奋(QUAN)的研究生生活开始了
- 然后
  - 我是什么人，我要做什么？
  - 研究的初心

新手上路

- 研究的初心

2.1 Motivation and Goals

- Why are you doing this? You must have a very good answer to this before you jump in to do a Ph.D. Don't do it if your main reason is "because I couldn't get a job" or "because there wasn't anything else I was particularly interested in doing at the moment…"
- It is very important to keep your goals and motivations in mind when in graduate school. Visualize what and where you would like to be in 5 years, and figure out what you need to do to get there.

https://www.ece.rutgers.edu/~pompili/index_file/extra/HowToDoResearch_ANRG_WP02001.pdf

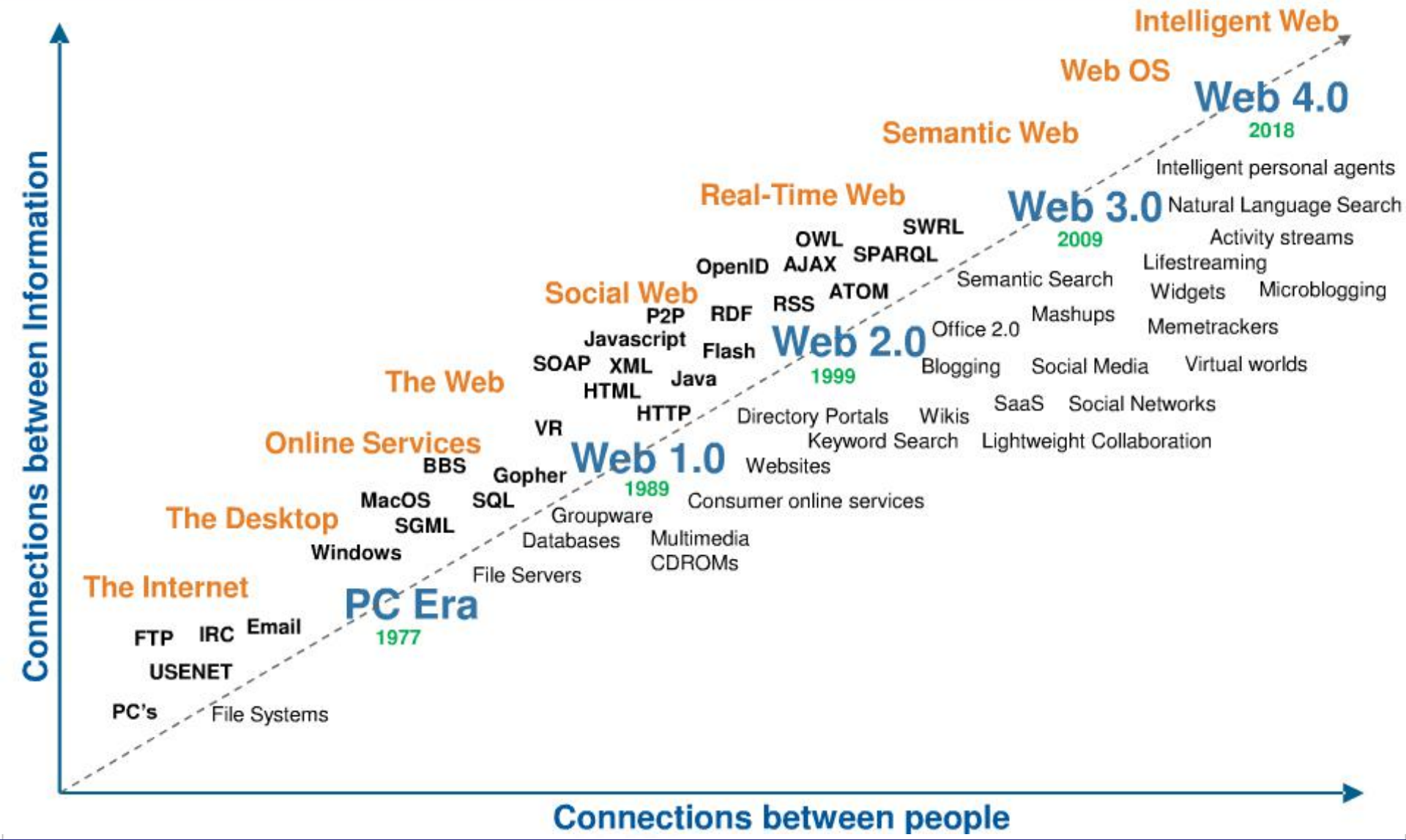
Read with a question in mind. "How can I use this?" "Does this really do what the author claims?" "What if...?" Understanding what result has been presented is not the same as understanding the paper. Most of the understanding

https://dspace.mit.edu/bitstream/handle/1721.1/41487/AI_WP_316.pdf

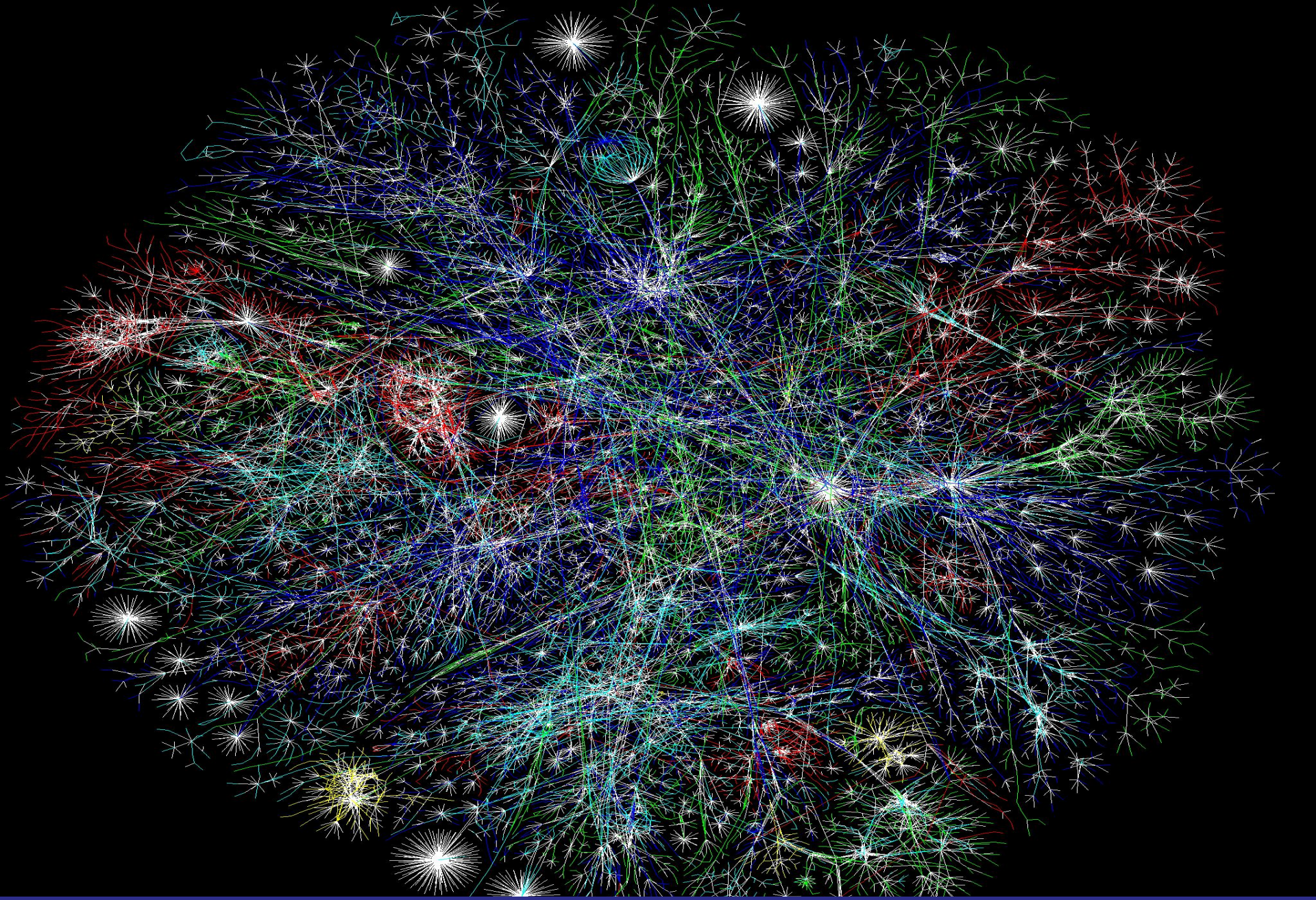
新手上路

- 首先
  - 从身旁熟悉的事物开始

The Intelligence is in the Connections



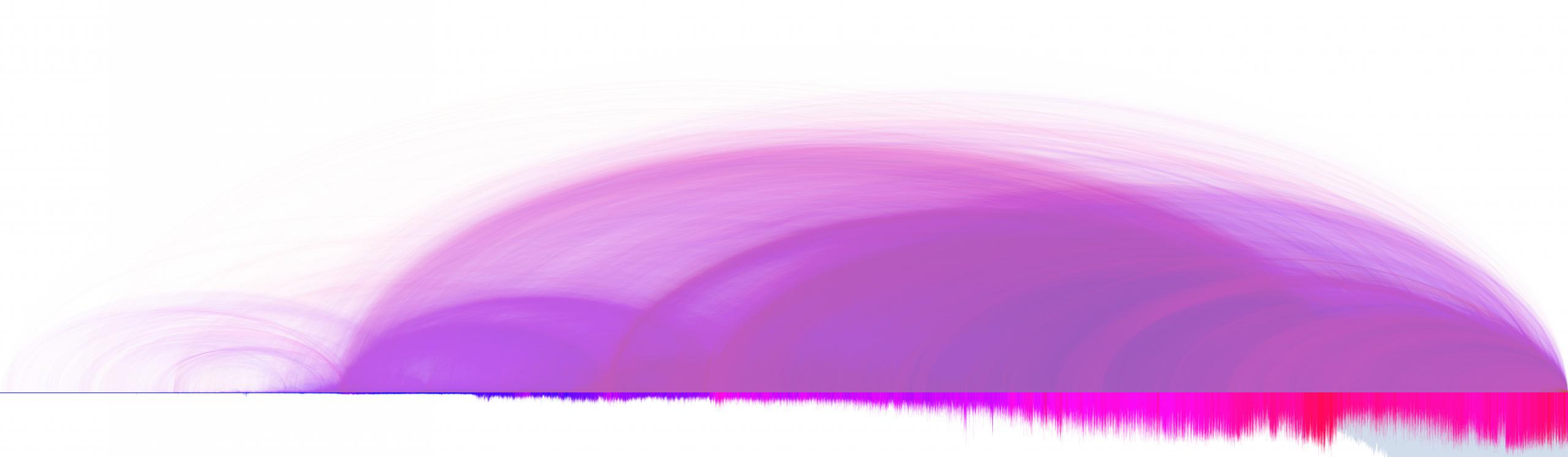
万维网持续进化



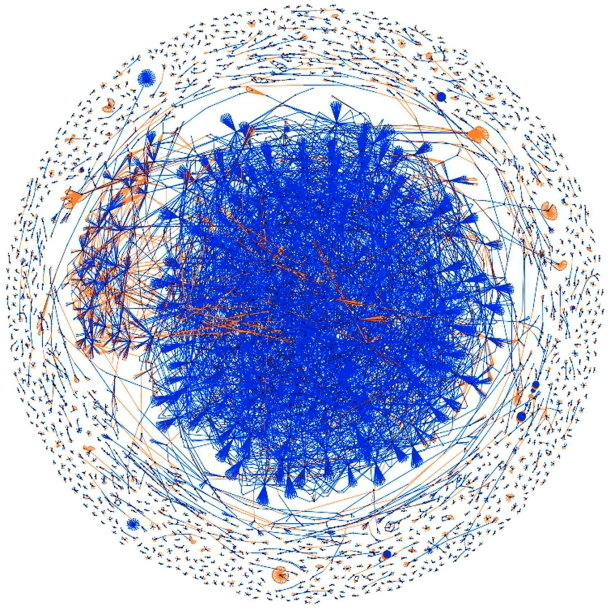
规模持续扩张

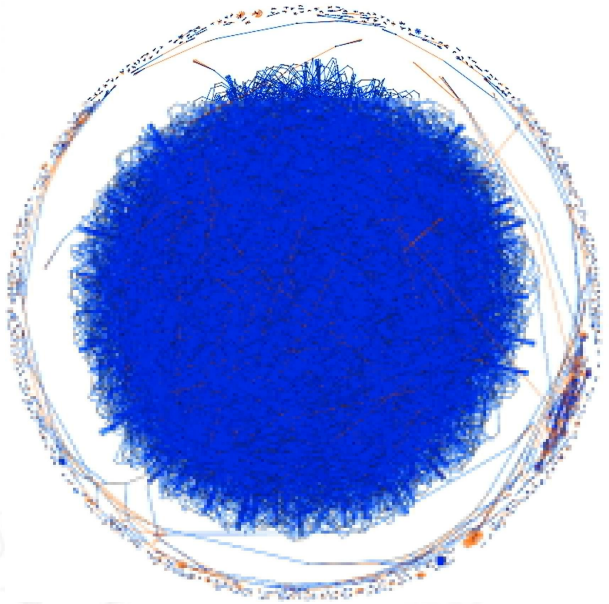

facebook
973,000
Logins
18 Million
Text
Messages
YouTube
4.3 Million
Videos Viewed
Google
3.7 Million
Search
Queries
NETFLIX
266,000
Hours
Watched
ANDROID APP ON
Google play
Available on the
App Store
375,000
Apps Downloaded
$862,823
Spent Online
174,000
Scrolling Instagram
60
SECONDS
2.4 Million
Snaps
Created
481,000
Tweets Sent

内容异常丰富



https://bitinfocharts.com/visualization/Bitcoin-clean-2560x1024.png

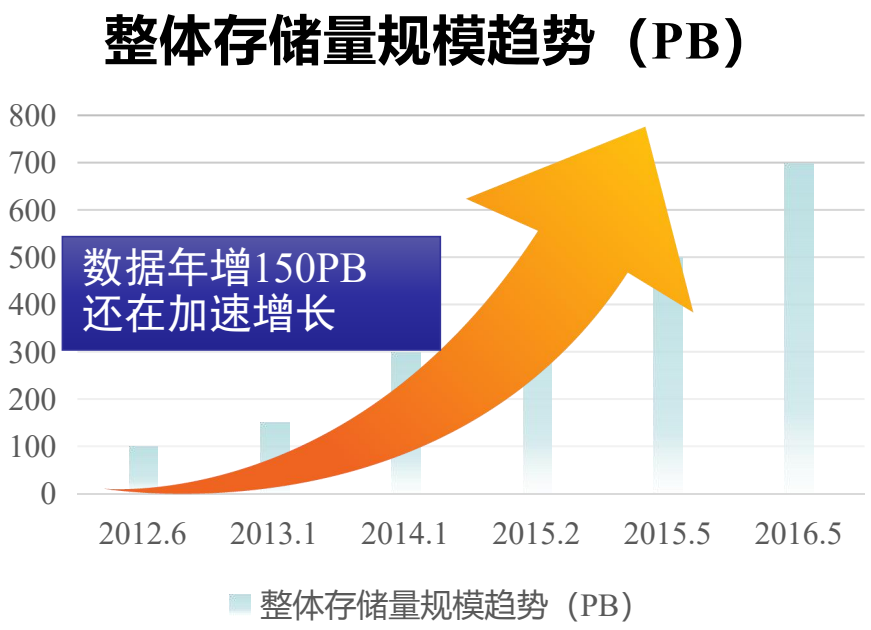




https://datastori.es/107-visualizing-bitcoin-with-dan-mcginn/

数据各处传播

身旁腾讯的现实



月活跃用户 >8亿

同时在线用户 >2亿

月活跃用户 >6亿

- 社交、游戏，访问密度高达100万次/秒/100GB量级的数据读写；
- 在线业务，应保证良好的用户体验，不论数据访问密集程度如何，均要求延迟在100毫秒以内

QQ空间相册

图片 4000亿张

存储量 200PB

微信朋友圈

日上传 10亿张

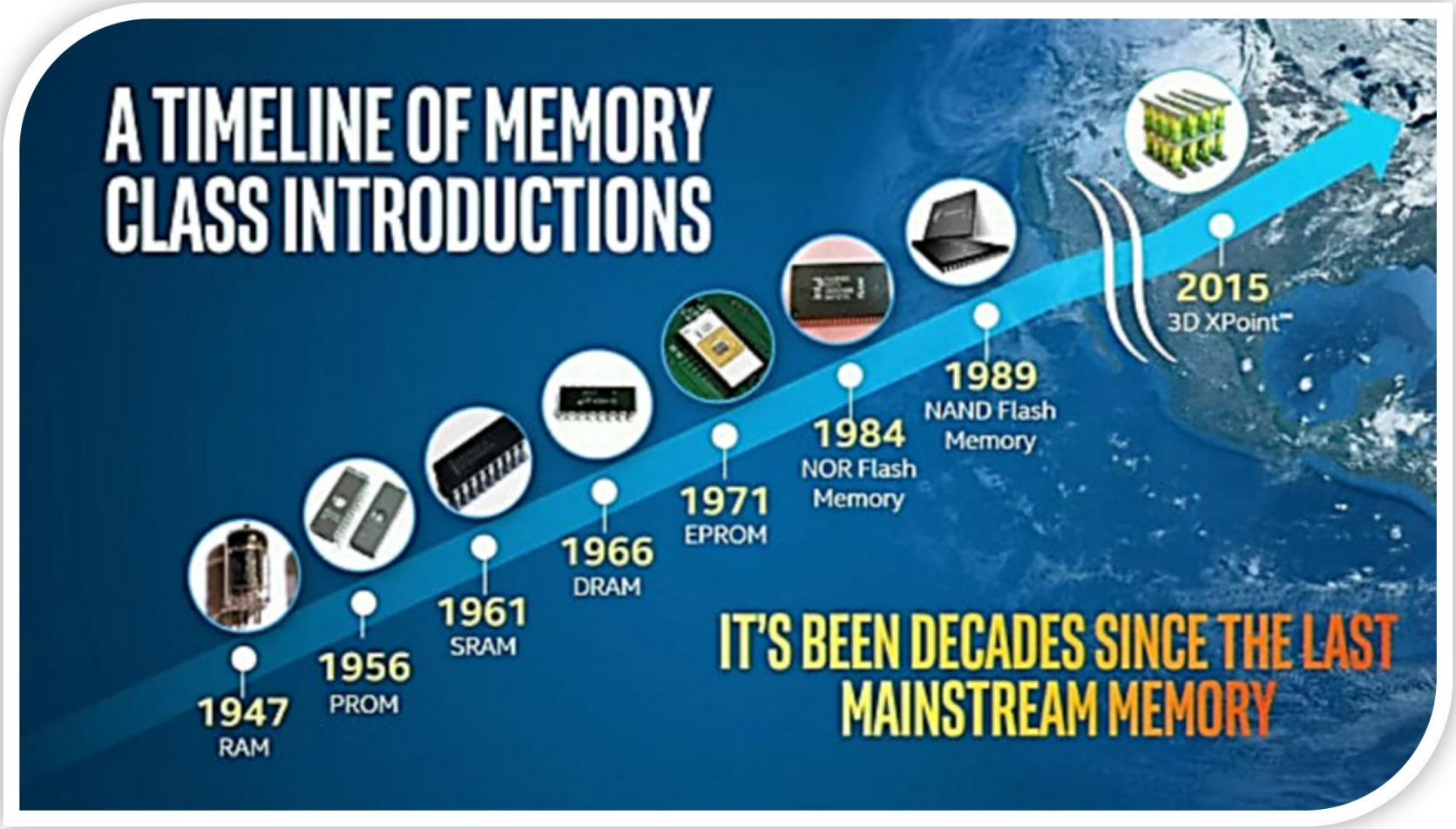
日下载 1200亿张

整个数字世界


数据存储挑战
庞大：规模持续扩张
复杂：内容结构丰富
4
BILLION
Connected People
$4
TRILLION
Revenue Opportunity
25+
MILLION
Apps
25+
BILLION
Embedded and
Intelligent Systems
50
TRILLION
GBs of Data
Source: Mario Morales, IDC

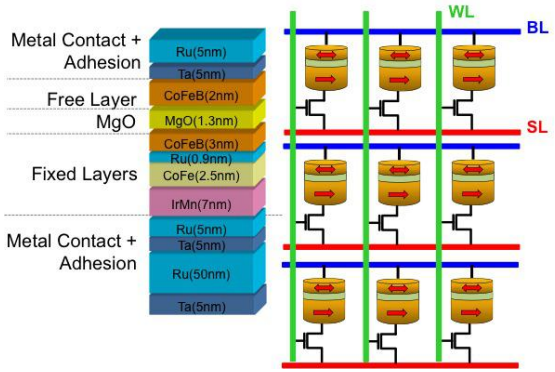
该怎么办？

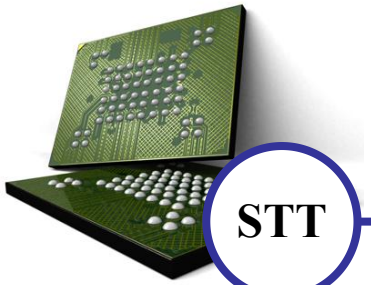
信息存储发展


A TIMELINE OF MEMORY CLASS INTRODUCTIONS
1947
RAM
1956
PROM
1961
SRAM
1966
DRAM
1971
EPROM
1984
NOR Flash Memory
1989
NAND Flash Memory
2015
3D XPoint™
IT'S BEEN DECADES SINCE THE LAST MAINSTREAM MEMORY

新锐存储技术

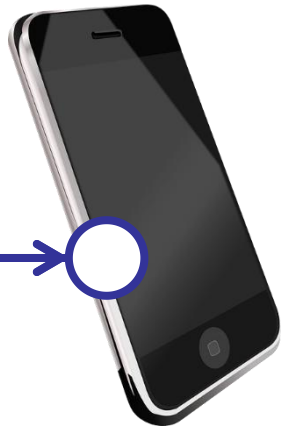
- 只靠这些？





STT

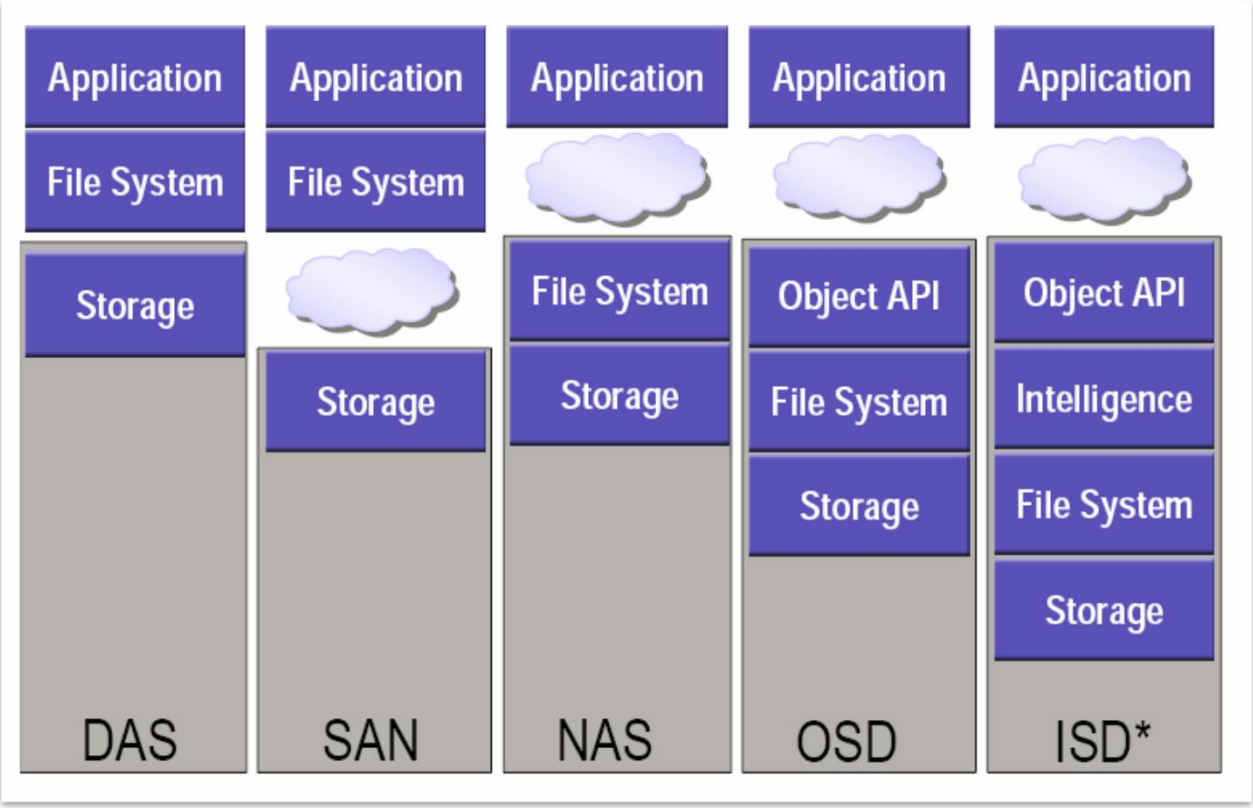
PRAM



还远远不够：数据规模、内容种类……

从战术上升至战略

- 系统层面还能提供什么干货？



第一部分

对象存储

Abstract

起源

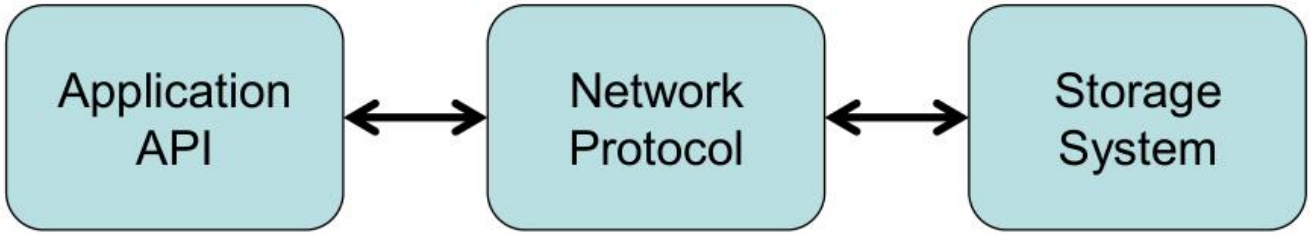
SNIA Education

- Object Storage is a generalized data container with uses in cloud storage, HPC file systems, and custom applications that provide their own indexing and metadata layers over objects. This tutorial provides a survey of these different kinds of objects, their APIs, and the applications that use them. The OSD (Object Storage Device) command set for SCSI provides a secure, general purpose mechanism used in HPC file systems via the NFSv4.1 pNFS (parallel NFS) protocol. The Amazon S3 object interface uses a web-based transport to provide cloud-based storage, and there are a number of similar interfaces in the open source community such as OpenStack Swift and Eucalyptus. The differentiating features of object storage systems include the access protocols (SCSI, RPC, REST), performance (high-speed LAN or WAN), security mechanisms, replication and reliability schemes, metadata and indexing services. As a result, different application domains have evolved their own Object Storage ecosystems.

Agenda

历史

- Object storage background and history
  - Abstract data containers
  - NASD, OSD, HTTP
- Objects and File Systems
  - Lustre, PanFS, Ceph, many others
- Objects and Web Storage
  - S3, Azure, Swift, many others



Two Paths to Objects 历史

SNIA Education

- Storage Devices
  - Move smarts into the device (NASD)
    - Network Attached Secure Disk
  - Raise the level of abstraction
    - Containers
    - Attributes
    - Security
  - SCSI Model
    - OSD command set
- Web Services
  - Add storage abstraction to a web-based system
    - Containers
    - Metadata
    - Security
  - REST Model
    - HTTP protocol

Some History

历史

- NASD (Network Attached Secure Disk)
  - 1990's research by Dr. Garth Gibson about moving intelligence into the storage device
  - Google cites NASD as inspiration for data node in its file system
- OSD (Object-based Storage Device)
  - Standards effort in the early 2000's created a SCSI command set for objects
  - There is a storage device behind this interface
- HTTP (HyperText Transfer Protocol)
  - A simple put/get protocol for the world-wide web
  - There is an arbitrary service behind this interface

基础概念

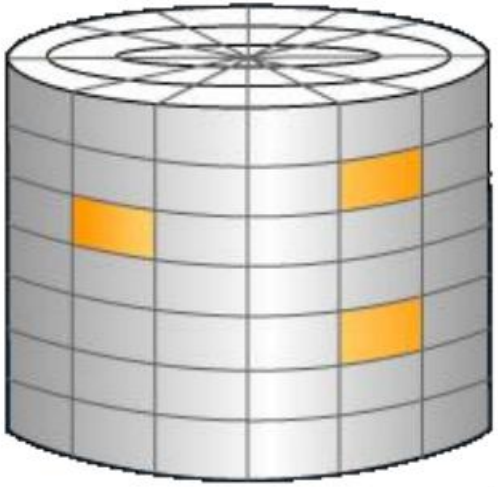
- Objects are containers for data and attributes
  - Every file system has an *inode* that is data blocks plus attributes
  - They are created, deleted, read, written, and have attributes

Block Based Device

Operations
Read block
Write block

Addressing
Block range

Allocation
External

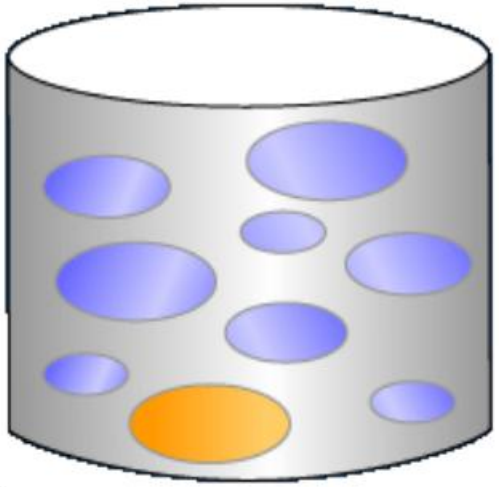


Operations
Create object
Delete object
Read object
Write object
Get Attribute
Set Attribute

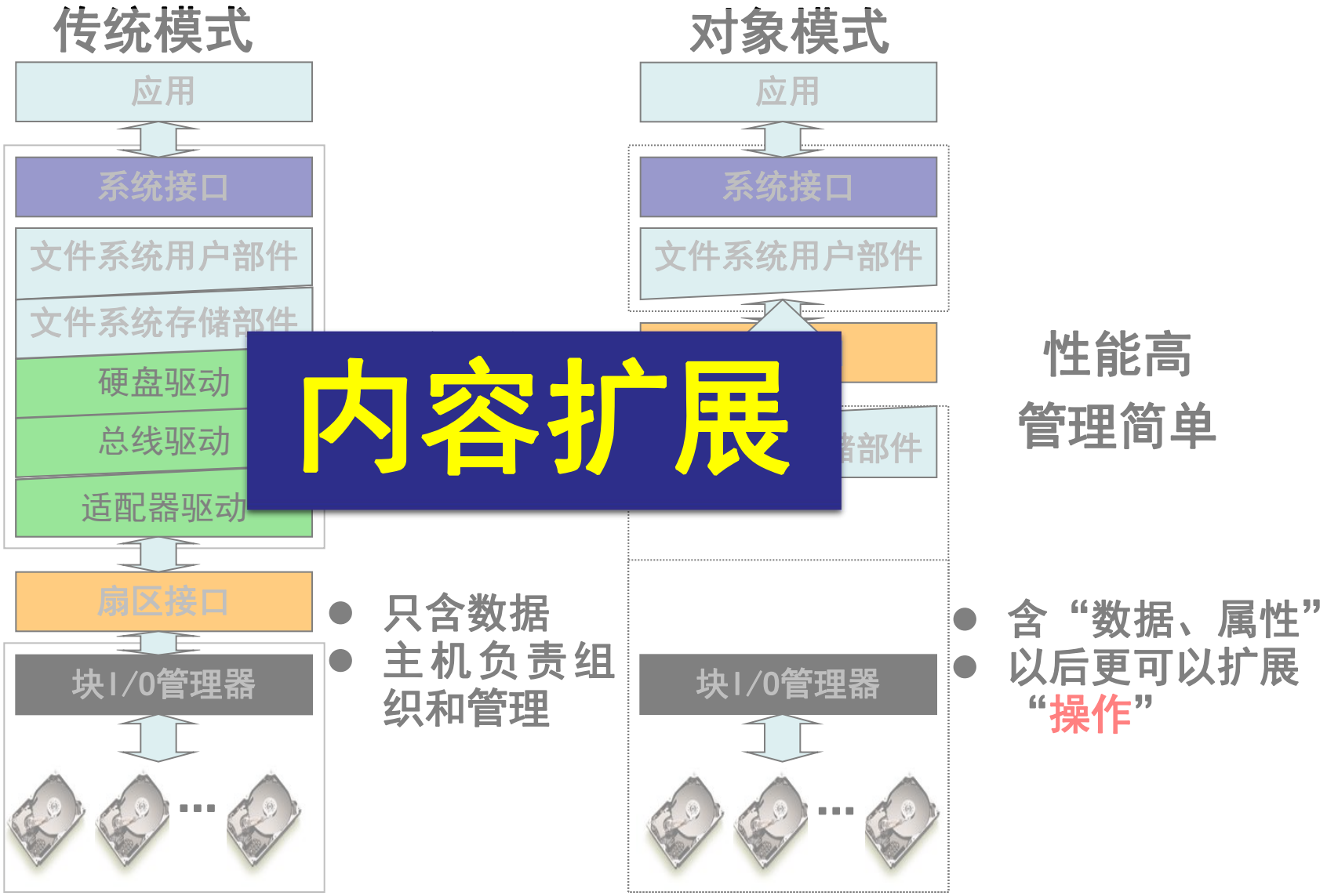
Addressing
[object, byte range]

Allocation
Internal

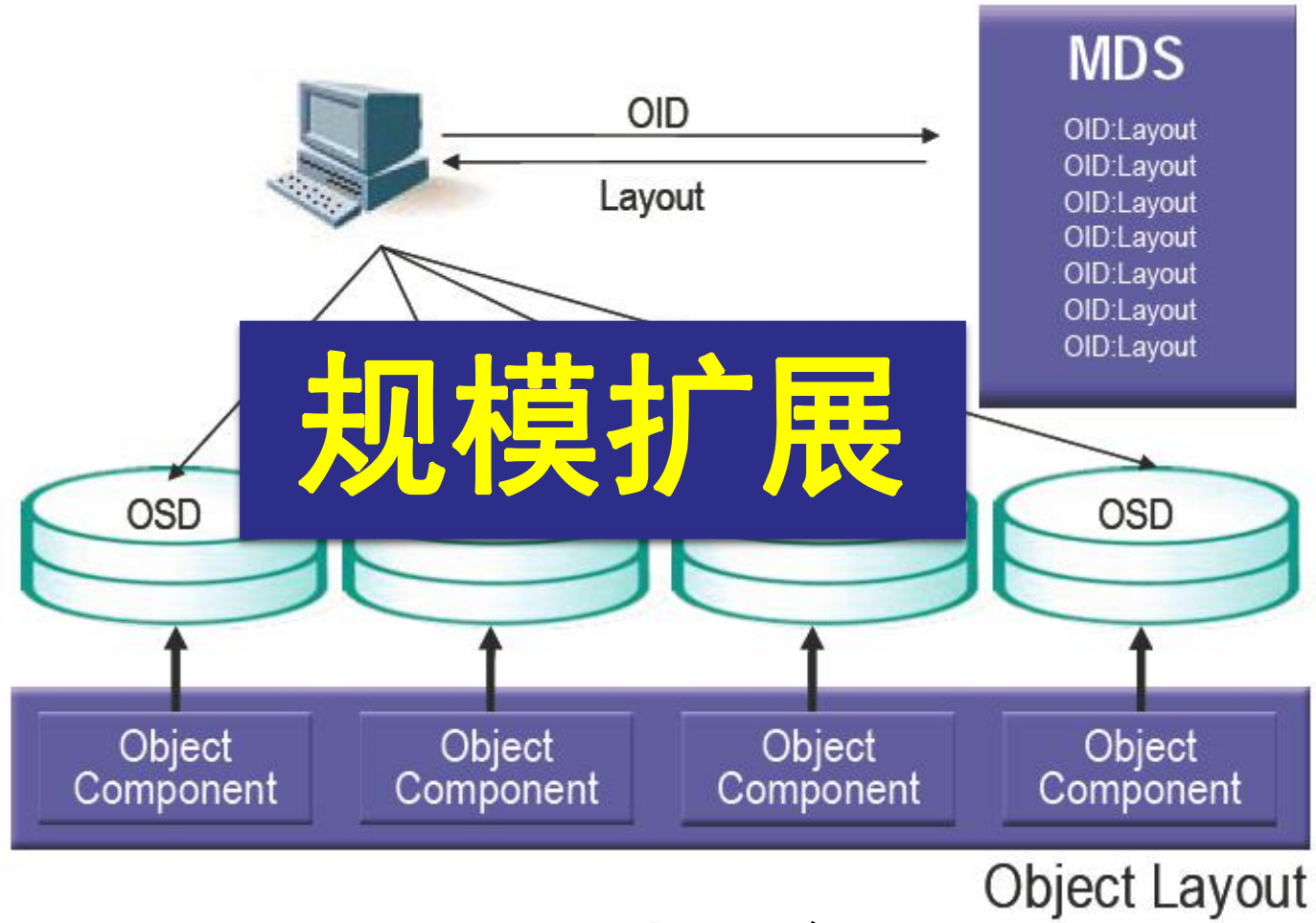
Object Based Device



对象存储技术

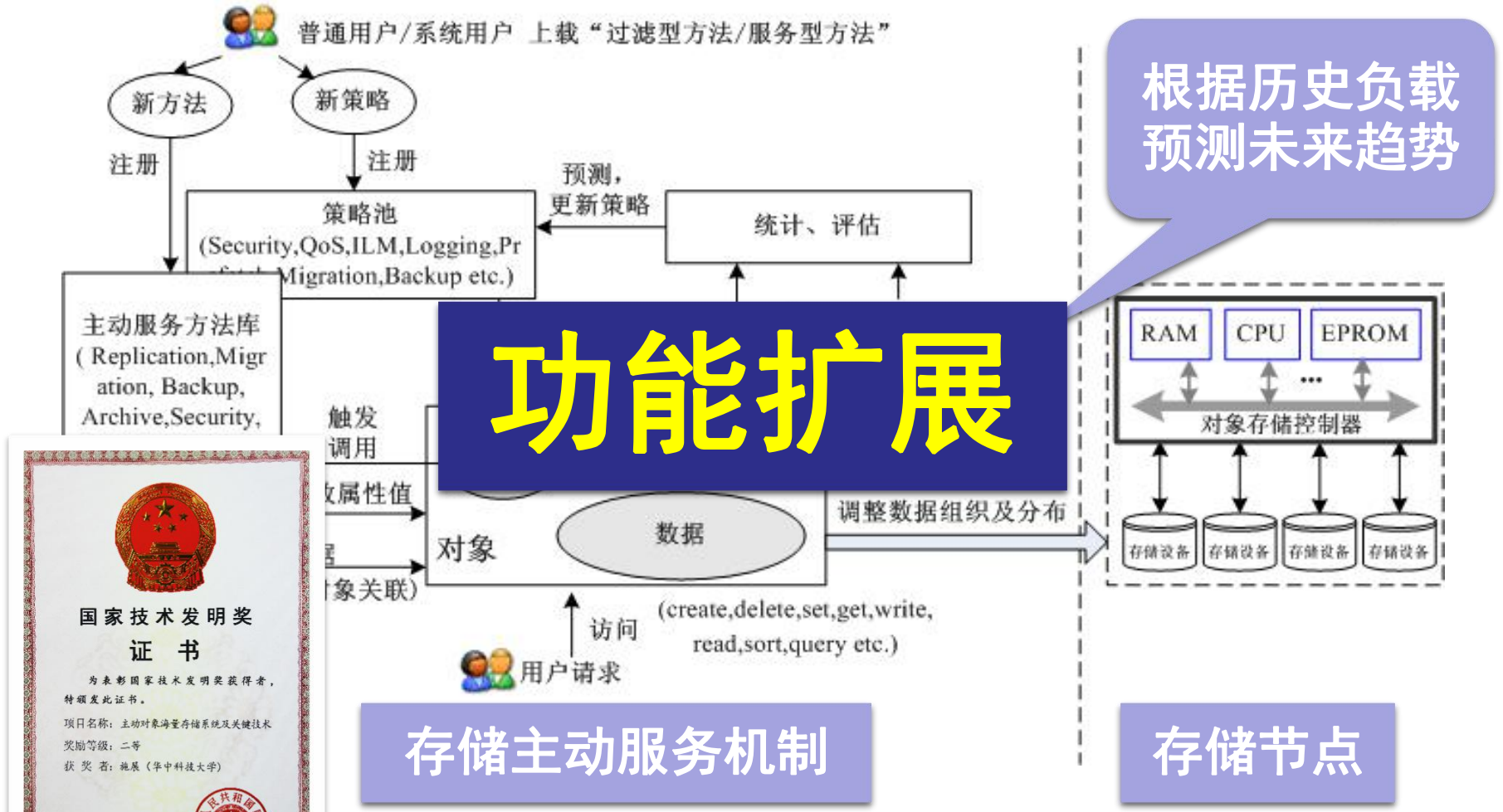

传统模式
对象模式
应用
系统接口
文件系统用户部件
文件系统存储部件
硬盘驱动
总线驱动
适配器驱动
扇区接口
块I/O管理器
主机
存储系统
内容扩展
性能高
管理简单
只含数据
主机负责组织和管理
含“数据、属性”
以后更可以扩展“操作”

对象存储系统



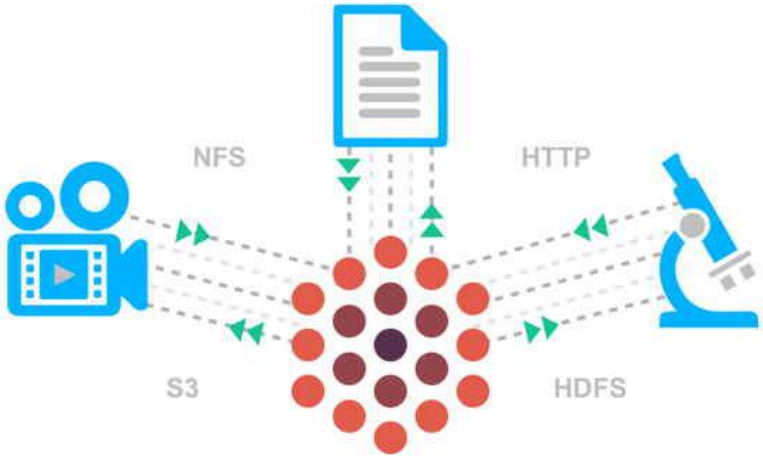
元数据由此解耦

主动对象存储


普通用户/系统用户 上载“过滤型方法/服务型方法”
新方法
新策略
注册
注册
预测，
更新策略
策略池
(Security,QoS,ILM,Logging,Pr
Migration,Backup etc.)
统计、评估
主动服务方法库
(Replication,Migr
ation, Backup,
Archive,Security,
触发
调用
属性值
象关联)
对象
数据
调整数据组织及分布
访问
(create,delete,set,get,write,
read,sort,query etc.)
用户请求
功能扩展
根据历史负载
预测未来趋势
RAM
CPU
EPROM
对象存储控制器
存储设备
存储设备
存储设备
存储设备
存储主动服务机制
存储节点

国家技术发明奖
证 书
为表彰国家技术发明奖获得者，
特颁发此证书。
项目名称：主动对象海量存储系统及关键技术
奖励等级：二等
获 奖 者：施展（华中科技大学）
证书号：2014-F-30901-2-04-R03

工业界广泛应用



简单
粗暴
有效

标准化

Google Cloud Platform

Windows Azure

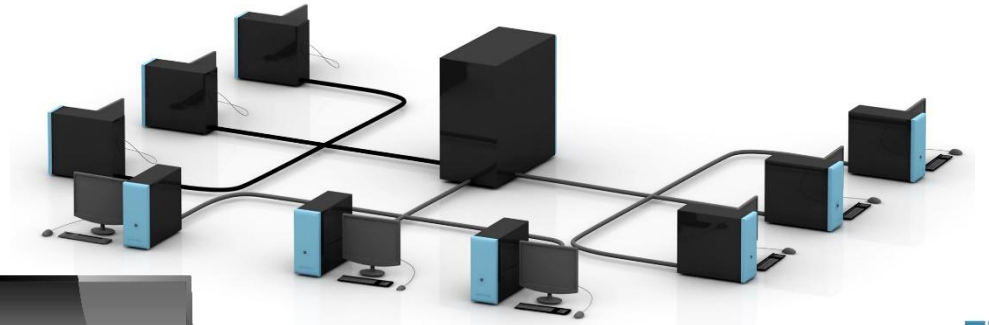
阿里云
aliyun.com

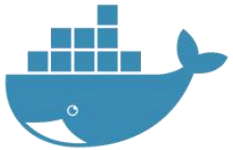
腾讯云

百度云

初步实践

- 说明
  - https://github.com/cs-course/obs-tutorial/blob/master/README.zh_cn.md
- 原材料
  - Git、Github/Gitlab/Bitbucket
  - Python、Go、Java
  - 可用虚拟机或容器



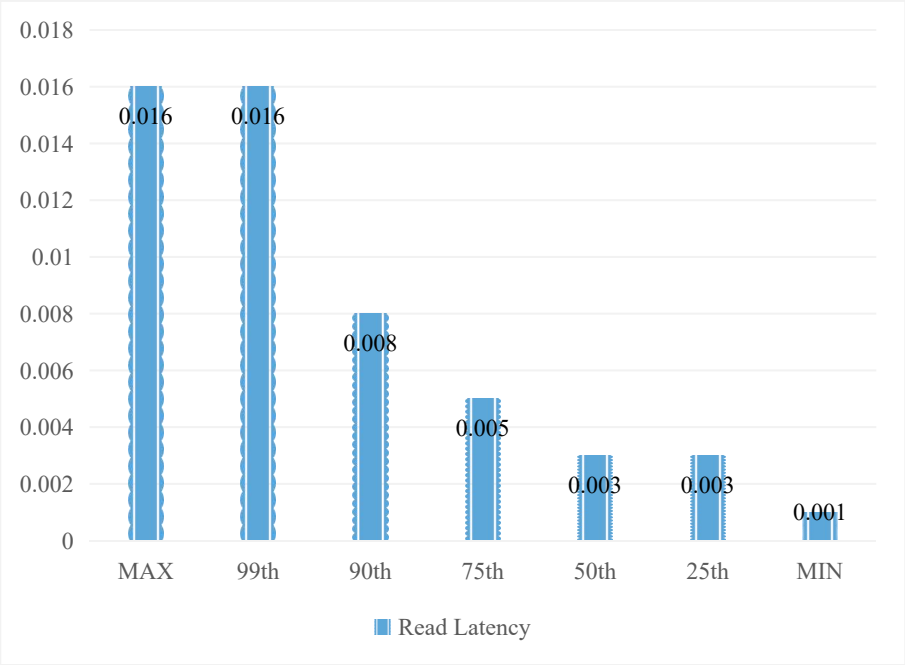


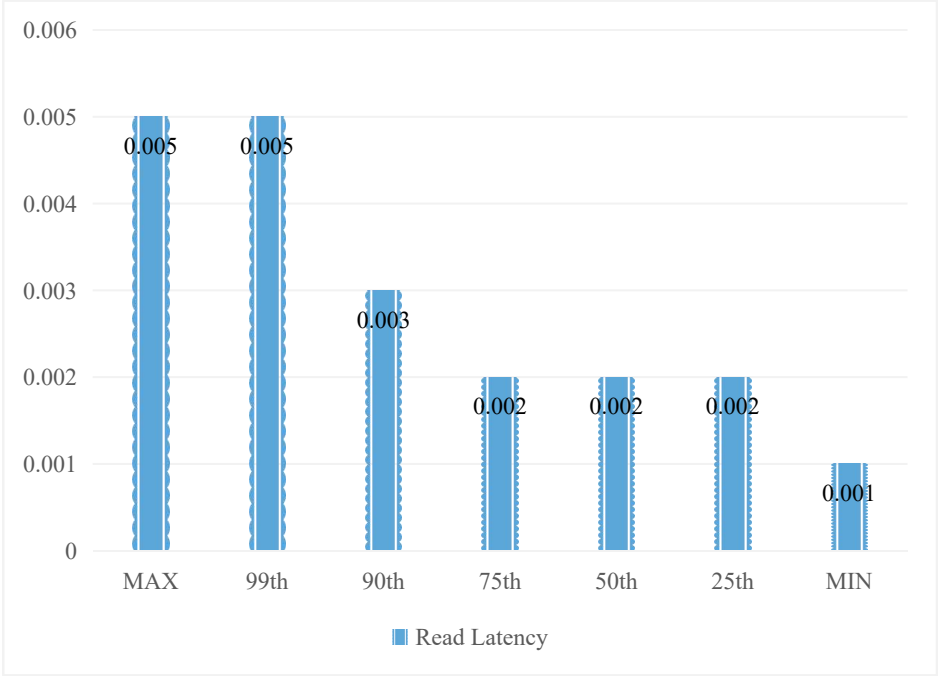
性能观测

```
Test parameters
endpoint(s):      [http://127.0.0.1:9000]
bucket:           loadgen
objectNamePrefix: loadgen
objectSize:       0.0010 MB
numClients:       10
numSamples:       100
verbose:       %!d(bool=false)


Results Summary for Write Operation(s)
Total Transferred: 0.098 MB
Total Throughput:  0.26 MB/s
Total Duration:    0.381 s
Number of Errors:  0
------------------------------------
Write times Max:       0.062 s
Write times 99th %ile: 0.062 s
Write times 90th %ile: 0.053 s
Write times 75th %ile: 0.042 s
Write times 50th %ile: 0.034 s
Write times 25th %ile: 0.030 s
Write times Min:       0.019 s


Results Summary for Read Operation(s)
Total Transferred: 0.098 MB
Total Throughput:  2.18 MB/s
Total Duration:    0.045 s
Number of Errors:  0
------------------------------------
Read times Max:       0.016 s
Read times 99th %ile: 0.016 s
Read times 90th %ile: 0.008 s
Read times 75th %ile: 0.005 s
Read times 50th %ile: 0.003 s
Read times 25th %ile: 0.003 s
Read times Min:       0.001 s


Cleaning up 100 objects...
Deleting a batch of 100 objects in range {0, 99}... Succeeded
Successfully deleted 100/100 objects in 94.8232ms
```

```
Test parameters
endpoint(s):      [http://192.168.3.85:9000]
bucket:           loadgen
objectNamePrefix: loadgen
objectSize:       0.0156 MB
numClients:       8
numSamples:       256
verbose:       %!d(bool=false)


Results Summary for Write Operation(s)
Total Transferred: 4.000 MB
Total Throughput:  1.54 MB/s
Total Duration:    2.604 s
Number of Errors:  0
------------------------------------
Write times Max:       0.145 s
Write times 99th %ile: 0.145 s
Write times 90th %ile: 0.100 s
Write times 75th %ile: 0.089 s
Write times 50th %ile: 0.076 s
Write times 25th %ile: 0.068 s
Write times Min:       0.060 s


Results Summary for Read Operation(s)
Total Transferred: 4.000 MB
Total Throughput:  58.21 MB/s
Total Duration:    0.069 s
Number of Errors:  0
------------------------------------
Read times Max:       0.005 s
Read times 99th %ile: 0.005 s
Read times 90th %ile: 0.003 s
Read times 75th %ile: 0.002 s
Read times 50th %ile: 0.002 s
Read times 25th %ile: 0.002 s
Read times Min:       0.001 s


Cleaning up 256 objects...
Deleting a batch of 256 objects in range {0, 255}... Succeeded
Successfully deleted 256/256 objects in 21.493393ms
```

找出问题



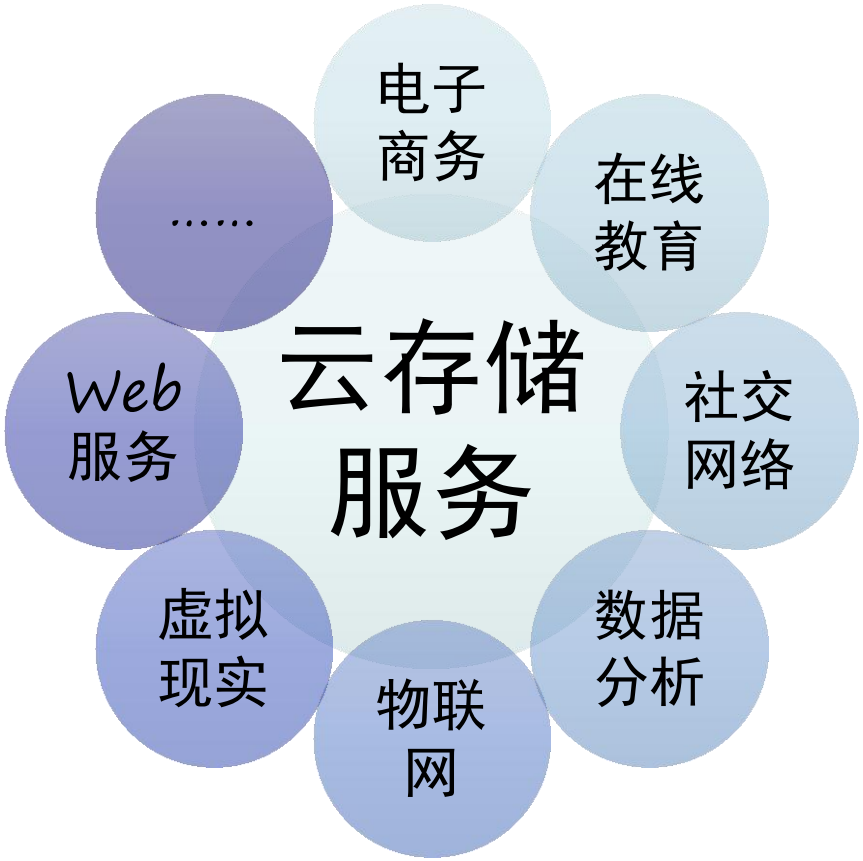



```
Results Summary for Read Operation(s)
Total Transferred: 0.098 MB
Total Throughput:  2.18 MB/s
Total Duration:    0.045 s
Number of Errors:  0
-------------------------------------
Read times Max:       0.016 s
Read times 99th %ile: 0.016 s
Read times 90th %ile: 0.008 s
Read times 75th %ile: 0.005 s
Read times 50th %ile: 0.003 s
Read times 25th %ile: 0.003 s
Read times Min:       0.001 s

Cleaning up 100 objects...
Deleting a batch of 100 objects in range {0, 99}... Succeeded
Successfully deleted 100/100 objects in 94.8232ms
```

```
Results Summary for Read Operation(s)
Total Transferred: 4.000 MB
Total Throughput:  58.21 MB/s
Total Duration:    0.069 s
Number of Errors:  0
-------------------------------------
Read times Max:       0.005 s
Read times 99th %ile: 0.005 s
Read times 90th %ile: 0.003 s
Read times 75th %ile: 0.002 s
Read times 50th %ile: 0.002 s
Read times 25th %ile: 0.002 s
Read times Min:       0.001 s

Cleaning up 256 objects...
Deleting a batch of 256 objects in range {0, 255}... Succeeded
Successfully deleted 256/256 objects in 21.493393ms
```

意味着什么

- 现代应用的交互性需求日益增强，要求其底层的云存储服务具有低响应延迟



iQIY
淘宝网 Taobao.com
优达学城 UDACITY
小米 xiaomi.com
Microsoft HoloLens
HORTONWORKS
CLOUDERA

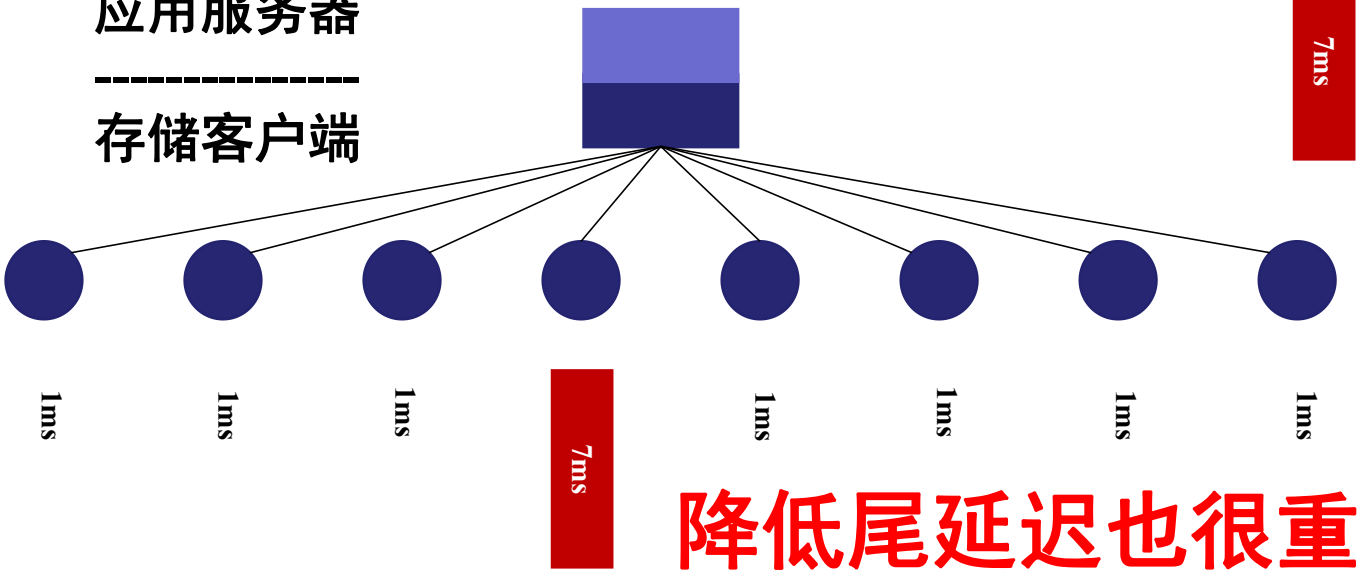
意味着什么

- 面
- 应

的

应用服务器

存储客户端



用户感知的延迟

存储请求

存储响应延迟

降低尾延迟也很重要！

本讲小结

- 对象存储基础知识
  - 为云存储数据
- 用于解决什么问题？
  - 扩展性
- 当前存在什么问题？
  - 性能保障

研究对象

要有矿

没挖光

研究目标