

# 腾讯的云计算平台-台风系统简介

腾讯公司-基础架构部

朱会灿

陈峰



火龙果•整理  
[uml.org.cn](http://uml.org.cn)

# 大纲

- 云计算 Overview
- 台风平台Overview
- 具体项目介绍
- 应用情况

## 生活中的云计算

- 手机通讯录备份到云端
  - 再不担心手机丢失了
- 云存储，云硬盘：
  - 数据随时随地access。
- 互联网搜索取代图书馆
  - 海量数据存在云端，一个简单的检索，几千台机器为你服务

# 云计算平台意义

- 让开发者专注于核心业务
  - 管理存储和计算，自动数据备份，自动切换机器
- 更有效利用资源
  - 资源共享，提高资源利用率。
  - 资源池：建立大的**cluster**（**1000+** 机器）
    - 处理大规模数据更高效
    - 节约成本，弹性扩展，方便**capacity planning**
- 更有效地管理
  - 统一监控，维护，安全保护
    - 权限认证和**quota**：保证贡献资源者

# 挑战

- 安全：
  - 数据（文件）操作，通讯准入，进程隔离
    - Static partition v.s. dynamic allocation
    - Virtual Machine, Sandbox, Linux container
  - Authentication, Access control
- 公平：
  - 进程调度，带宽使用
  - Quota: 存储，计算资源
  - Hard limit: CPU, memory
  - 优先级

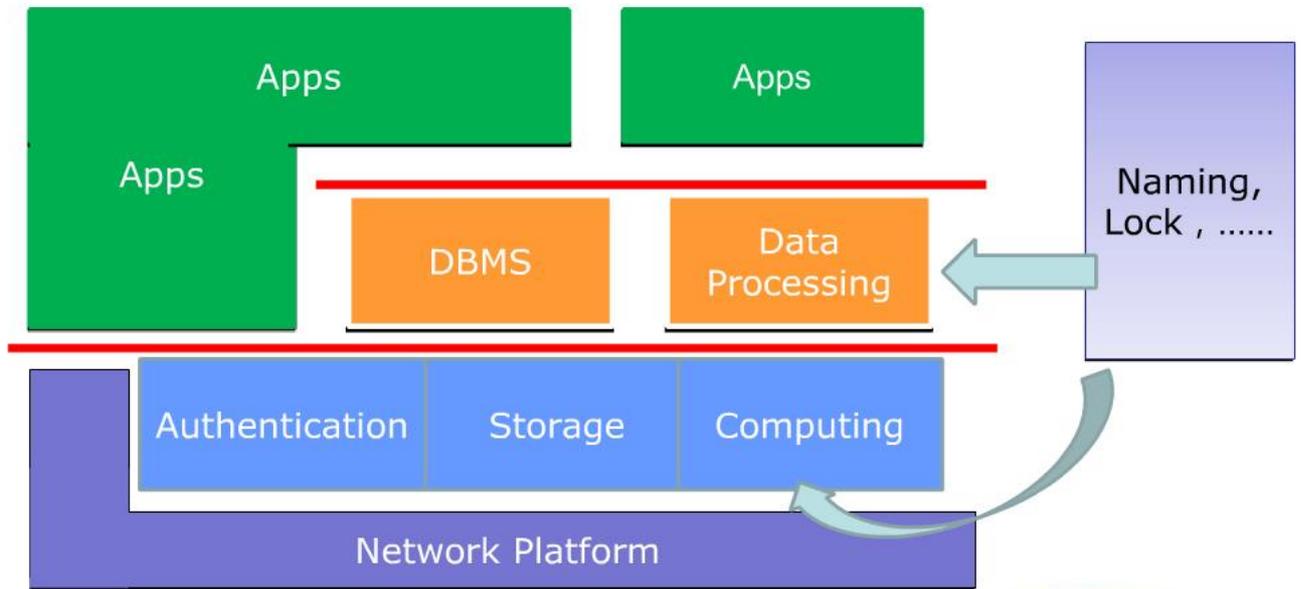
# 大纲

- 云计算 Overview
- 台风平台Overview
- 具体项目介绍

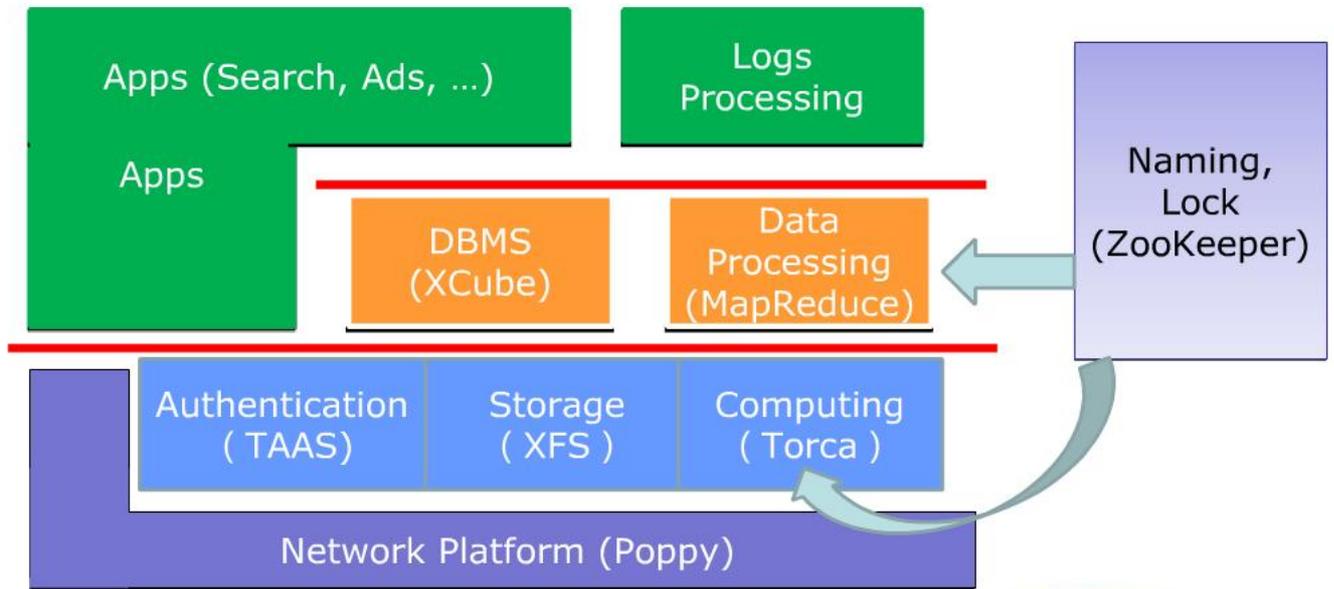
# 云计算平台作为OS

- 把云计算平台看做一个OS，则需要：
  - 文件系统（存储系统）
  - 进程和资源管理系统（运行程序，管理CPU，内存，等等）
  - 权限管理系统（账户，权限，认证，等）
  - 方便开发的系统软件（数据库等）
  - 程序开发 APIs（networking, threads, etc）

# 常规网络应用的项目层次关系



# 台风系统的层次关系



# 台风平台特性

## 清晰分层

- 模块界面
- 类似于Hadoop，但有独立的cluster management system

## 私有

- Target 公司内部业务：搜索，日志处理
- Use Linux Container

## 兼容性

- 主要采用C++开发，跟现有的代码无缝整合
- 通过SWIG/JNA等方式支持了多语言
- 支持Hadoop应用



# 台风平台目标

## 高效

- 处理大规模数据:  
Terabytes of data.

## 共享

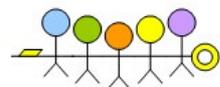
- Cluster 供多个业务和服务使用

## 公平

- 贡献资源者
  - 优先拿到资源
  - 资源不够时能抢到资源

## 安全

- 文件读写
- Job提交和控制
- 数据库Access



# 大纲

- 云计算 Overview
- 台风平台Overview
- 具体项目介绍：
  - Poppy
  - XFS
  - Torca
  - XCube
  - MapReduce

# 通讯平台



# Poppy

## • Poppy – 基于Protocol Buffer的RPC框架

### RPC

- Remote Procedure Call
- 调用远程服务的接口，就像调用本地函数一样简单
- 屏蔽了底层网络通讯和协议的实现细节

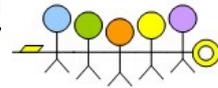
### Protobuf

- Google开源的序列化框架
- 高效可扩展的结构化数据存储方式
- 基于二进制，
  - 10% 于xml的序列化后数据大小
- 易于版本控制。前后向兼容

### HTTP

- Browser 浏览，提交
- JSON，PHP

Poppy = Protobuf + RPC + HTTP + m



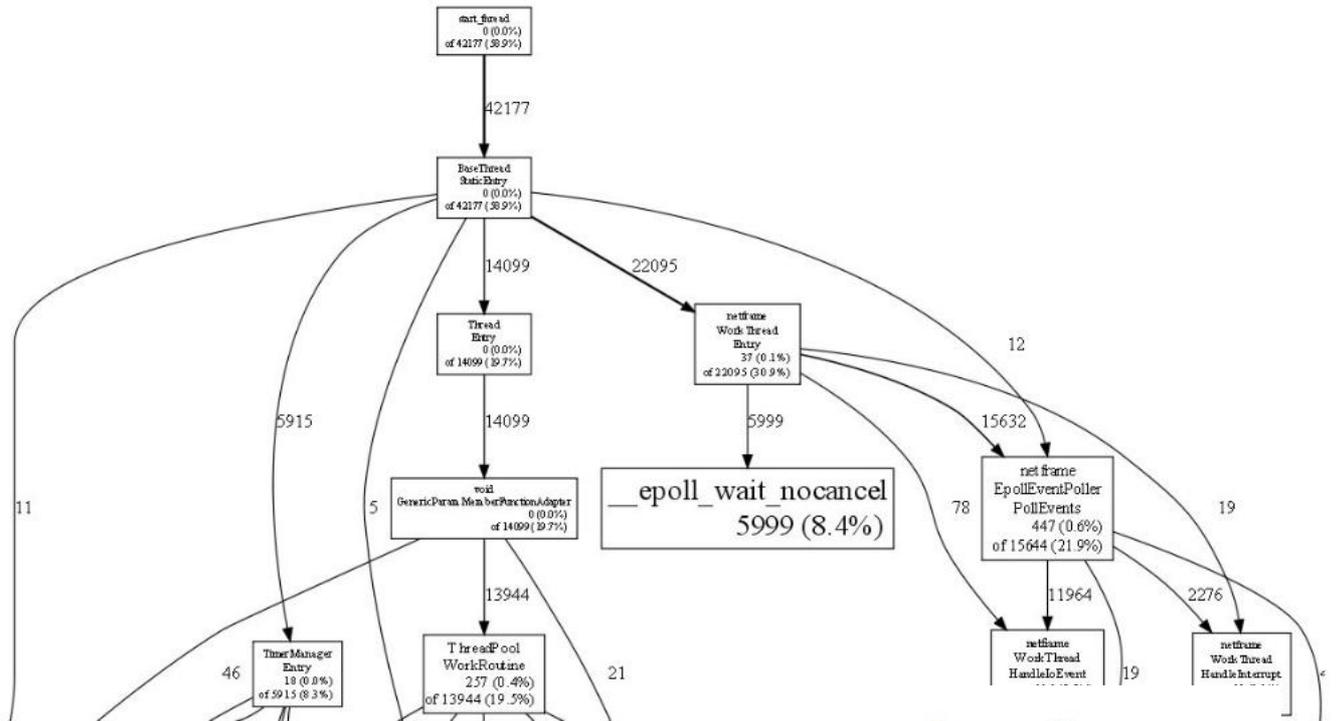
火龙果•整理  
[uml.org.cn](http://uml.org.cn)

# Poppy的主要特性

- 高性能
- 支持多服务器负载均衡
- 支持传输内容压缩，内置压缩策略。
- 支持streaming，可流式分块传输大量数据。
- 支持多语言（Java, Python, PHP，以及任何支持JSON的语言）
- 支持TNS地址（基于Torca上的Job/Task寻址，当任务迁移时自动改到新的地址）

# Poppy特性：动态profiling

pprof \$binary\_file <http://host:port/>



# Poppy特性：Web提交

10.6.222.127:8080/rpc/form/rpc\_examples.EchoServer.Echo  
<< Back to method selection

Form for rpc\_examples.EchoServer.Echo

Hide Form

int32 id*	<input type="text" value="11"/>										
string query	<input type="checkbox"/>										
bool flag	<input type="checkbox"/>										
enum test_enum	<input type="checkbox"/>										
message nested_msg*	<table border="1"><tr><td>int32 id*</td><td><input type="text"/></td></tr><tr><td>string title</td><td><input type="checkbox"/></td></tr><tr><td>string url</td><td><input type="checkbox"/></td></tr></table>	int32 id*	<input type="text"/>	string title	<input type="checkbox"/>	string url	<input type="checkbox"/>				
int32 id*	<input type="text"/>										
string title	<input type="checkbox"/>										
string url	<input type="checkbox"/>										
message[] nested_msgs	<table border="1"><tr><td>+</td><td><table border="1"><tr><td>int32 id*</td><td><input type="text"/></td></tr><tr><td>string title</td><td><input type="checkbox"/></td></tr><tr><td>string url</td><td><input type="checkbox"/></td></tr></table></td></tr><tr><td>-</td><td></td></tr></table>	+	<table border="1"><tr><td>int32 id*</td><td><input type="text"/></td></tr><tr><td>string title</td><td><input type="checkbox"/></td></tr><tr><td>string url</td><td><input type="checkbox"/></td></tr></table>	int32 id*	<input type="text"/>	string title	<input type="checkbox"/>	string url	<input type="checkbox"/>	-	
+	<table border="1"><tr><td>int32 id*</td><td><input type="text"/></td></tr><tr><td>string title</td><td><input type="checkbox"/></td></tr><tr><td>string url</td><td><input type="checkbox"/></td></tr></table>	int32 id*	<input type="text"/>	string title	<input type="checkbox"/>	string url	<input type="checkbox"/>				
int32 id*	<input type="text"/>										
string title	<input type="checkbox"/>										
string url	<input type="checkbox"/>										
-											
int32[] rep_int	<input type="text"/>										
bytes b	<input type="checkbox"/>										
bytes[] bs	<input type="text"/>										
string query_ext	<input type="checkbox"/>										
uint64 long_ext	<input type="checkbox"/>										
request encoding	UTF-8										
response encoding	Same as request										
Send Request	Reset										

## Form for rpc\_examples.EchoServer.Echo

Show Form

```
{
  b: ,
  flag: false,
  id: 10001,
  long_ext: 0,
  nested_msg: {
    id: 222,
    title: ,
    url: ,
  },
  nested_msgs: [
    {
      id: 2222,
      title: ,
      url: ,
    },
  ],
  query: test,
  query_ext: test ext,
  rep_int: [
    1,
  ],
  test_enum: 1,
}
```



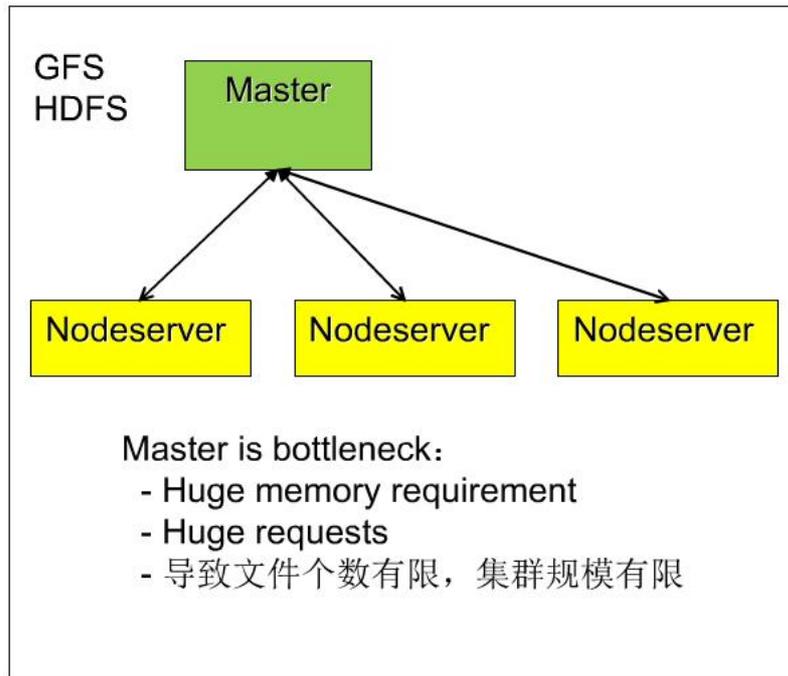
# Poppy项目目前使用情况

项目	描述
Typhoon	台风平台所有项目
Cocktail	长尾广告
新网页搜索	新的网页搜索采用Poppy做通讯
其他	社区搜索，网络平台

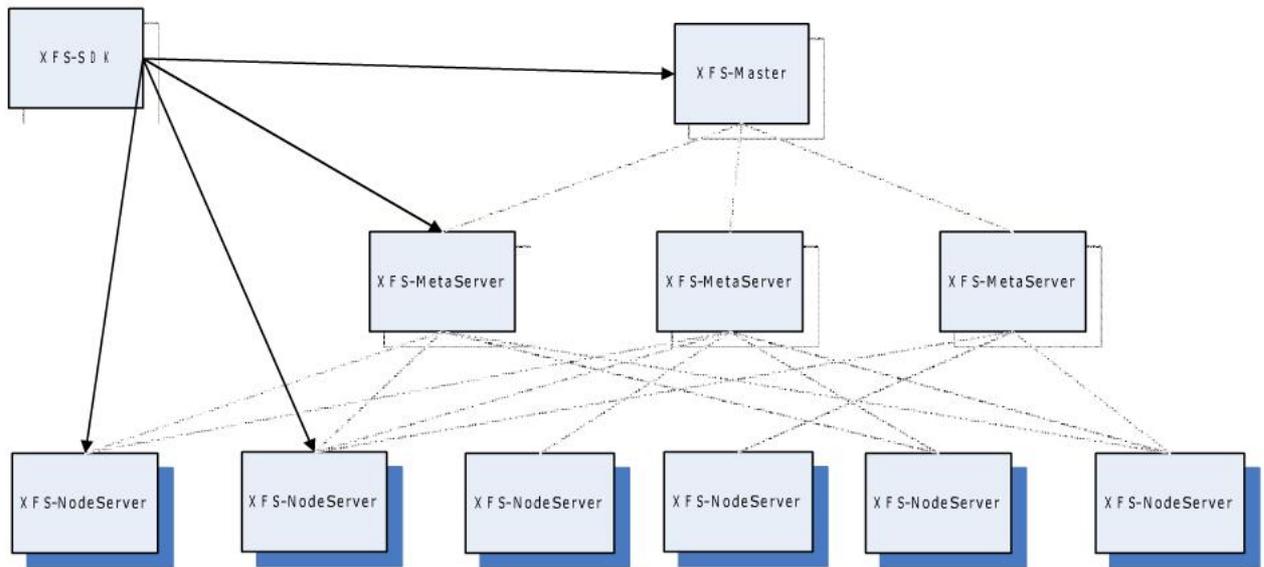
# 大纲

- 具体项目介绍：
  - Poppy
  - XFS
  - Torca
  - XCube
  - MapReduce

# 业界分布式文件系统



# XFS架构设计



## 安全性 ( Safety & Security )

### 主备Master

- LSM-Tree: Commit Log + checkpoint
- 两阶段提交

### 主备Metaserver

- 无状态
- 主备切换的高可用机制

### 文件块备份

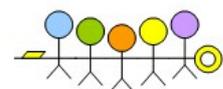
- 缺省3备份
- Checksum

### 回收站

- 删除的文件自动放到回收站里，可恢复。

### 权限管理

- Ownership，读写权限，Quota限制



## 系统性能数据

- 读：从多个nodeserver并发读
- 写：以pipeline 方式同时写到多个nodeserver
  
- 支持多达**5亿**个文件
- 单client读大数据：90M/s
- 单client写大数据：66M/s

## 文件操作相关函数

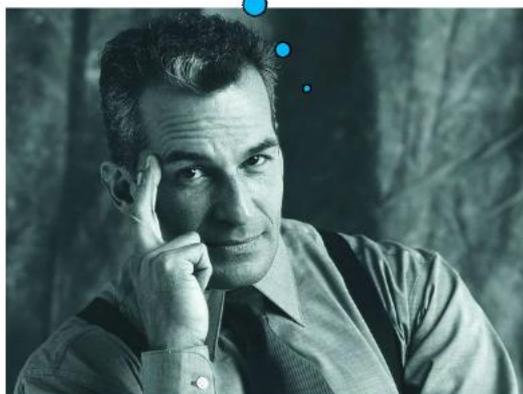
- Open: 同步, 异步, 备份数
  - Close
  - AsyncRead
  - Read
  - ReadLine
  - AsyncWrite
  - Write
  - Copy
  - Seek
  - Tell
  - Flush
  - LocateData: 获取文件的各数据块所在的机器ip
- Move
  - Rename
  - AddDir
  - List
  - GetMachingFiles
  - GetSize
  - Chmod

# 大纲

- 具体项目介绍：
  - Poppy
  - XFS
  - Torca
  - XCube
  - MapReduce

# With Torca

只考虑自己的业务



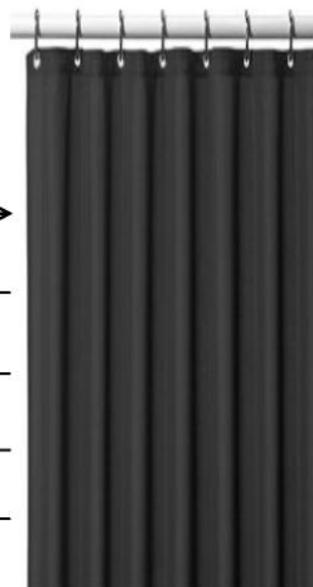
资源申请

给我资源

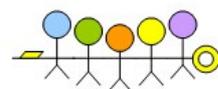
启动程序

监控程序

展示程序

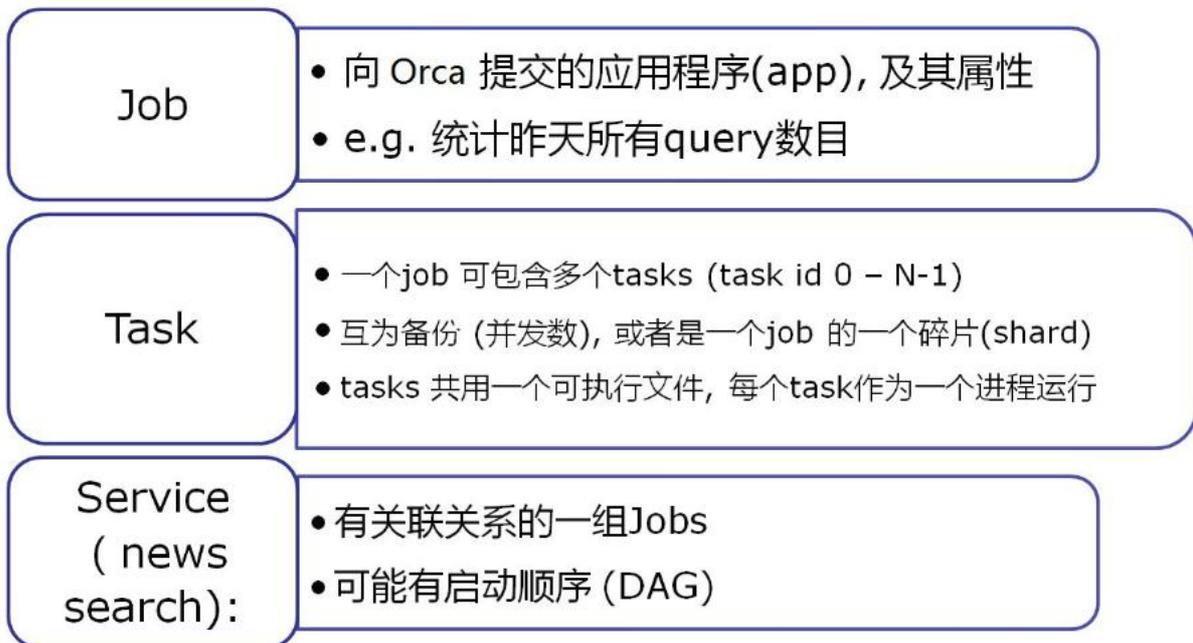


用户不需要知道幕布后是什么...



火龙果•整理  
[uml.org.cn](http://uml.org.cn)

# Service/Job/Task



# Job Description File

## 用于描述用户对其作业的资源要求

```
cmd = /home/user/bin/querycounter
```

```
Requirements = server_type=="TS5"
```

```
Task_count = 10
```

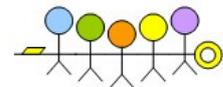
```
CPU = 2
```

```
Memory = 2G
```

```
Args = ...
```

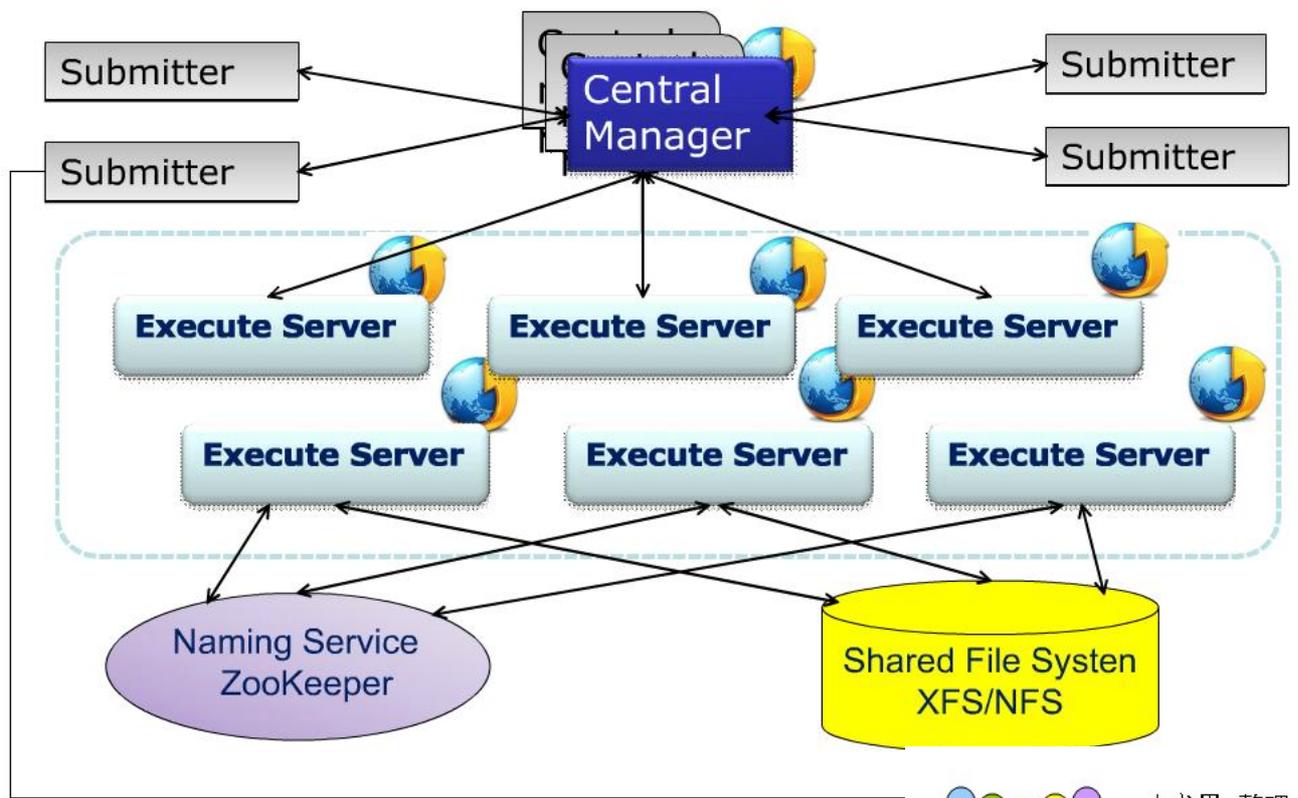
如何把**10**个数加起来:

- 手工
- **task 0** 等待其它**task** 完成, 通过网络(文件)收集结果
- 起一个新的**job**, 只有一个**task**, 读出前面写出的结果, 加起来 (**service**)



火龙果•整理  
[uml.org.cn](http://uml.org.cn)

# Architecture of Torca



# Torca实现

## 资源共享

- 服务器复用
- 统一分配port , memory , cpu , hard disk 等
- Quota + 抢占调度

## 自动容错

- Failed task : 重启一定次数及迁移到其他机器重启。
- Failed machine : 自动迁移它上面的所有task

## 任务隔离

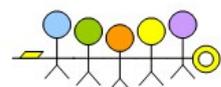
- 不同任务由Resource Container隔离 (hard limit)
- Job permission , access control

## 适应性强

- 灵活的资源匹配策略 : cpu负载、cpu核数、硬盘, Memory、内网流量、Hertz、机架等
- Map HDFS to XFS, 支持Hadoop job scheduling

## 规模庞大

- Hard work !



# Torca应用

- <http://news2.soso.com/>

**SOSO 新闻**  **搜搜** 搜搜新闻每3分钟自动更新

首页 国内 国际 社会 军事 财经 娱乐 体育 科技 教育 健康 女性 汽车 房产 互联网 游戏

## 山西代县铁矿垮塌事故已发现9人死亡4人受伤

[山西代县一铁矿发生垮塌事故 9...][图文详情] 山西代县铁矿事故...][突发事件] 山西代县一铁矿发...]



· 为什么个税起征点还是3000元? 17:09  
· 审计署: 去年央行10个分支行违规发津... 15:06  
· 铁道部: 京沪高铁投资回报定位保本微利 18:07  
· 坚决树立反隔必脏的信心 13:52  
· 新闻出版总署: 不允许建立所谓记者“黑... 16:44  
· 全面部署 多措并举 加强高校招生录取管理 8:53  
· 记公安一线共产党员群体: 做人民喜爱... 13:32  
· 中国拟修改兵役法 大学生入伍放宽至24岁 9:55

世界最大规模蓝精灵集会 1 2 3

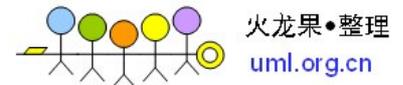
**热搜榜**

1. 长春 日晕奇观
2. 悬浮视察照
3. 小狗戴墨镜安全帽
4. 拯救 面具娃娃
5. 一年3次闪婚诈骗
6. 起征点3000元
7. 河南天价过路费案
8. 抢救官员奖40万
9. 警察 失明 坚持巡逻
10. 纪敏佳 整容争议

**分享排行榜**

1. 组图: 长沙再发枪击案 警察全城缉捕... (1567次)
2. 中国人争得比美国少物价比美国高引... (1132次)
3. 中铁建沙特轻轨项目净亏41.48亿将由... (1126次)
4. 昆明警方发布最大规模通缉令抓捕500... (1111次)
5. 昆明警方发布最大规模通缉令 共500... (197次)
6. 长沙遭暴雨袭击积水严重 市区交通瘫痪 (96次)
7. 高清: 京沪高铁贵宾候车室亮相 (94次)
8. 在重庆女方有“裸婚”特权? 网友质疑... (94次)

**新闻追踪** 张柏芝 | 北京暴雨 | 塑化剂 | 高铁



# 大纲

- 具体项目介绍：
  - Poppy
  - XFS
  - Torca
  - **XCube**
  - MapReduce

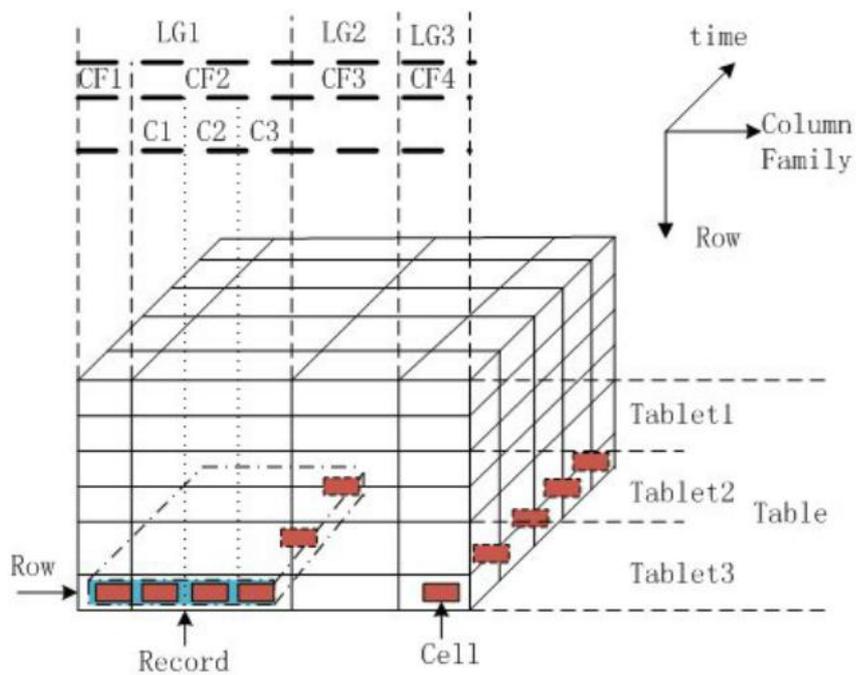
# XCube是什么

- XCube是分布式NoSQL数据库

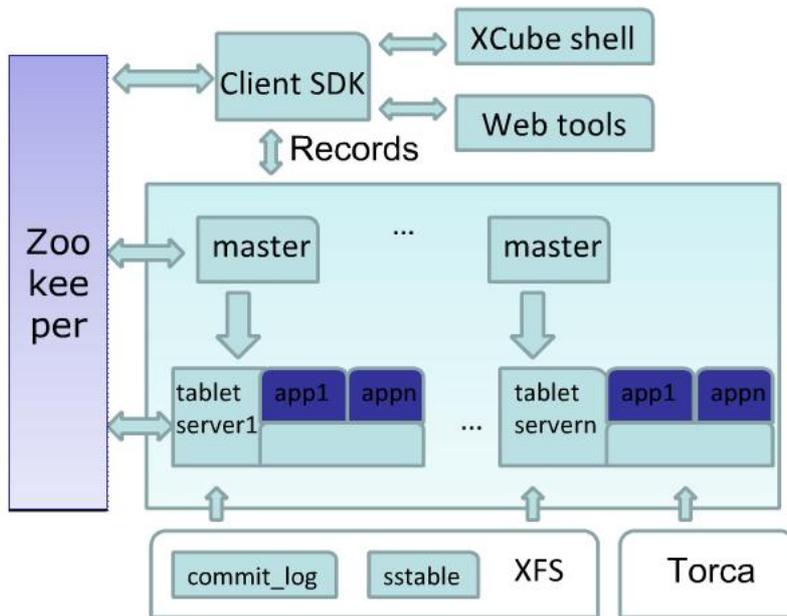
	SQL	NoSQL
结构	2 维表	3维+
操作	SQL 语句	Customized API
存储	按行存储	按列存储
数据模型	ACID	XCube支持单行内 atomic ops
Scalability	Low	High



# XCube数据模型



# XCube系统架构—子模块设计



子模块名	功能
Master	负责管理和维护系统当中的所有TabletServer，以及表操作和权限验证过程。
Tabletserver	负责组织和管理结构化数据，提供读、写、扫描、删除等服务。
Client SDK	提供给用户整个系统的读、写、扫描、删除等服务的API。
XCube shell	提供，表操作和少量数据操作的Shell工具。
Web tools	提供整系统的监控和运维工具

# XCube项目目前使用情况

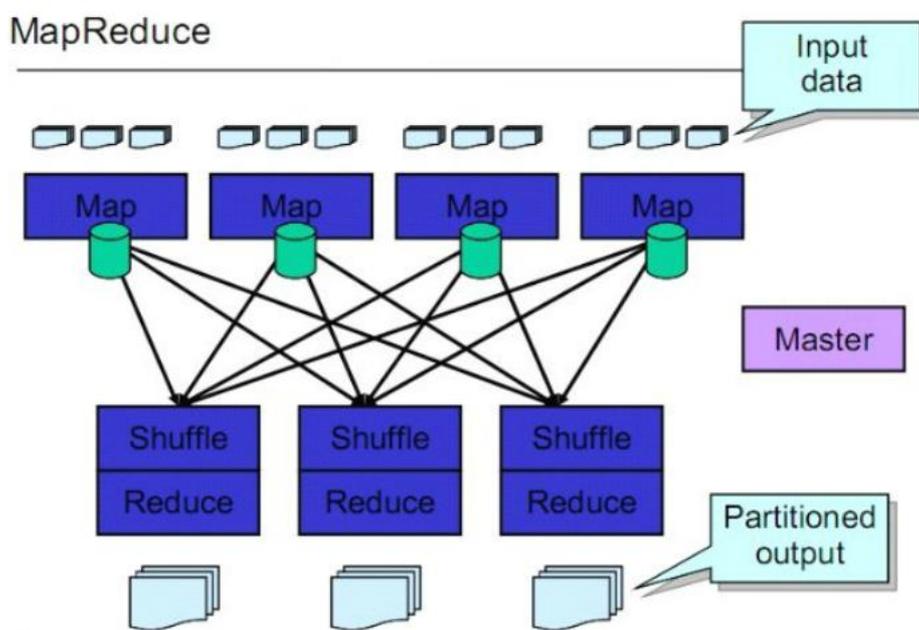
集群名	已接入应用	使用情况
websearch	网页索引项目	测试数据上表现稳定： URL table : ~1000亿行， 100万亿 cells ; 100TB 数据



# 大纲

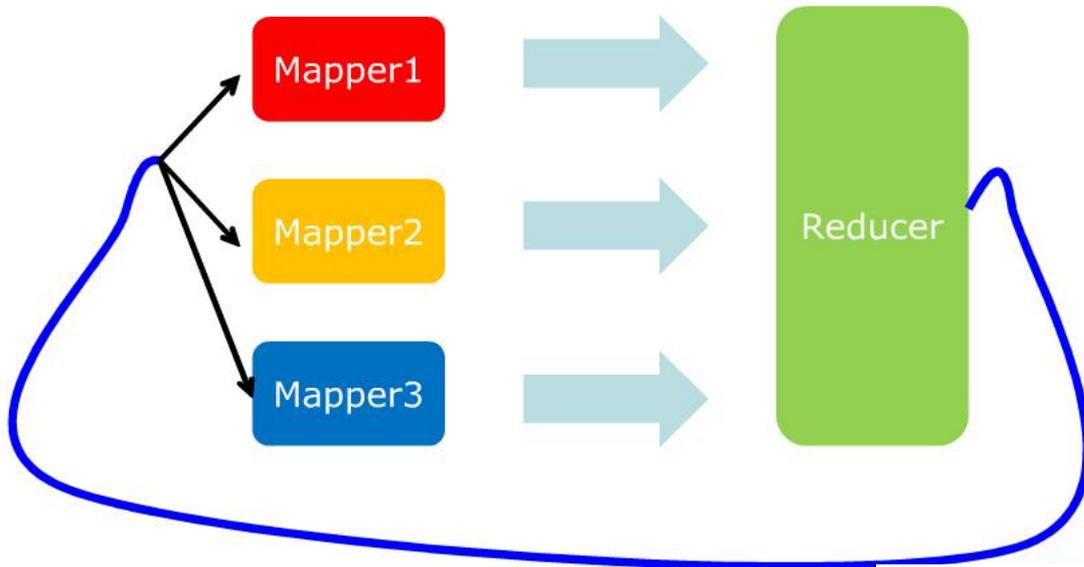
- 具体项目介绍：
  - Poppy
  - XFS
  - Torca
  - XCube
  - **MapReduce**

# MapReduce简介



## MultiMap

- 用户在一个MapReduce任务中可以指定多个Map Class，用于对不同格式的文件进行处理。



# 运行时监控

## MapReduce Status

job状态: kRunningJob 中间数据预测大小: 136.74T counters

开始时间: 2012-02-17 18:21:21 运行时间: 18h 6m 33s

Map Task	总数: 19325	完成数: 19056	剩余数: 269	失败数: 4	-	map预测结束时间: 2012-02-18 12:43:00
进度						map: 98.62%

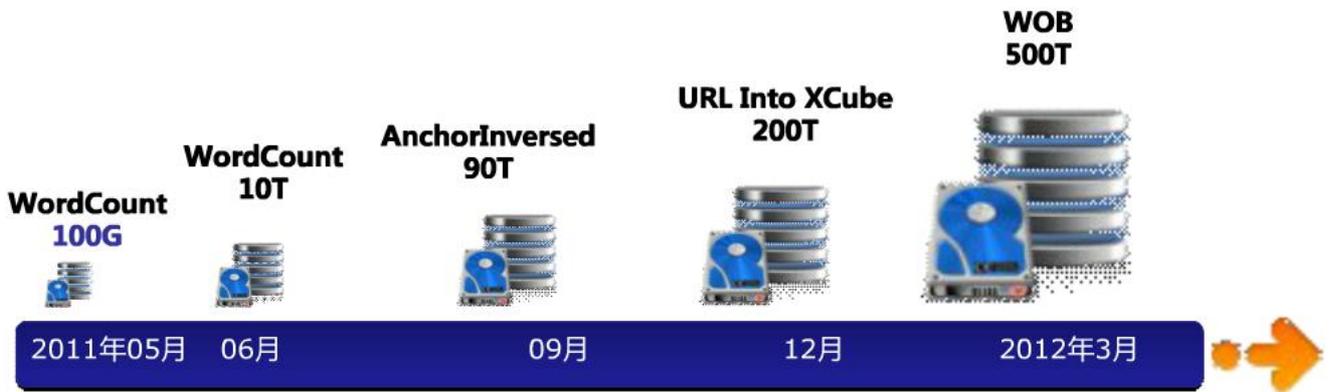
Reduce Task	总数: 700	完成数: 0	剩余数: 700	失败数: 12	shuffle预测结束时间: 2012-02-18 13:28:38	-
进度						shuffle: 94.48% reduce: 0%



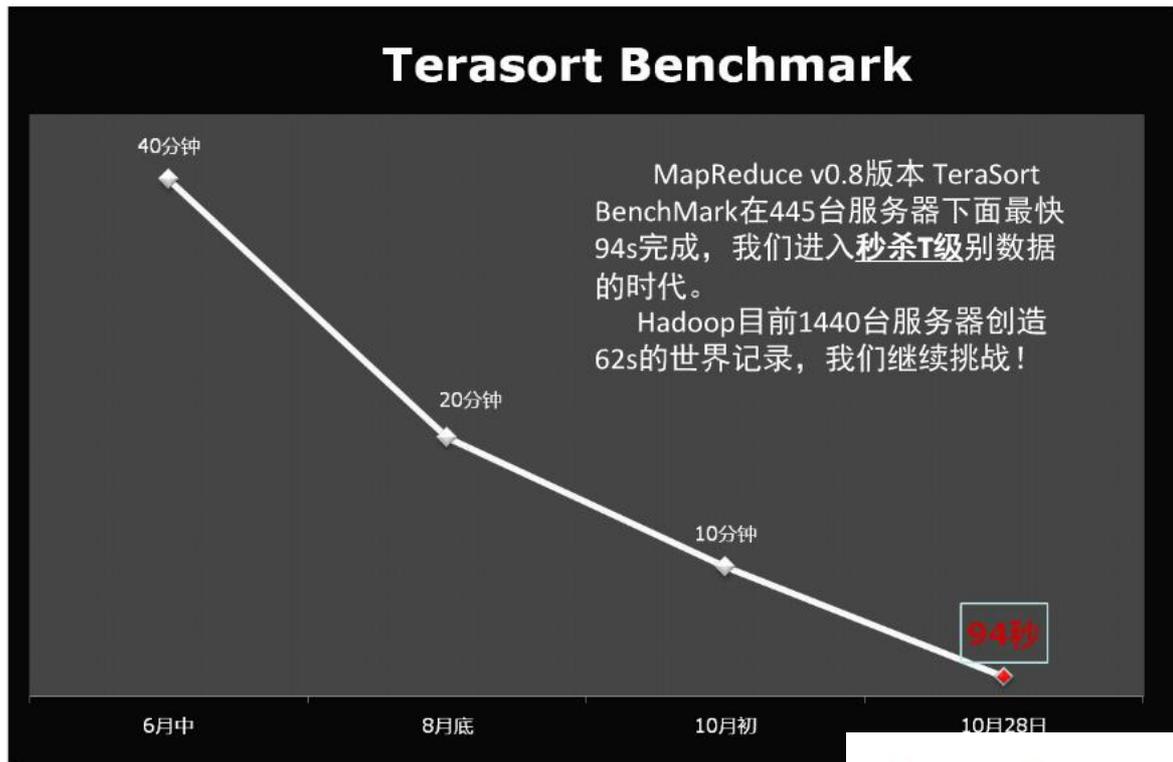
火龙果•整理  
uml.org.cn

# MapReduce 数据规模

MapReduce 4月份完成第一个版本开发，5月份成功完成100G数据规模的运算，2012年初，单次作业运算最大数据规模突破500T。



# MapReduce 成就—TeraSort



火龙果•整理  
[uml.org.cn](http://uml.org.cn)

# 大纲

- 云计算 Overview
- 台风平台Overview
- 具体项目介绍
- **应用情况**

# Typhoon 系统促进创新

- 自动管理机器和存储：
  - 开发者只需全力关注功能和产品
- 快速试验成为现实：
  - AntiSpam, Click Model, 锚文本处理：
    - 处理大规模数据。
    - 迭代一次：2星期 → 几个小时
      - 快速大规模试验成为现实
      - 可并行跑几个试验

## Cluster部署情况

- 在多地都有cluster部署
- 数据挖掘和处理业务：
  - click model, anti-spam: 1星期→几个小时
  - Anchor text 反转: 2星期 → 12 小时
- 线上服务：
  - 基于新架构的新闻搜索 (news2.soso.com)
  - 内容广告, 统一下载系统, websearch indexing

## 项目规模

- 100万行源代码，包含大量的测试代码，覆盖率85%
- 70名开发人员
- 历时两年
- 严格的开发流程和规范
- 开发过程中还有CodeReview，构建工具等副产品产出。
- 还在持续发力中

# Q & A

