Embracing Open Source:
Practice and Experience from Alibaba

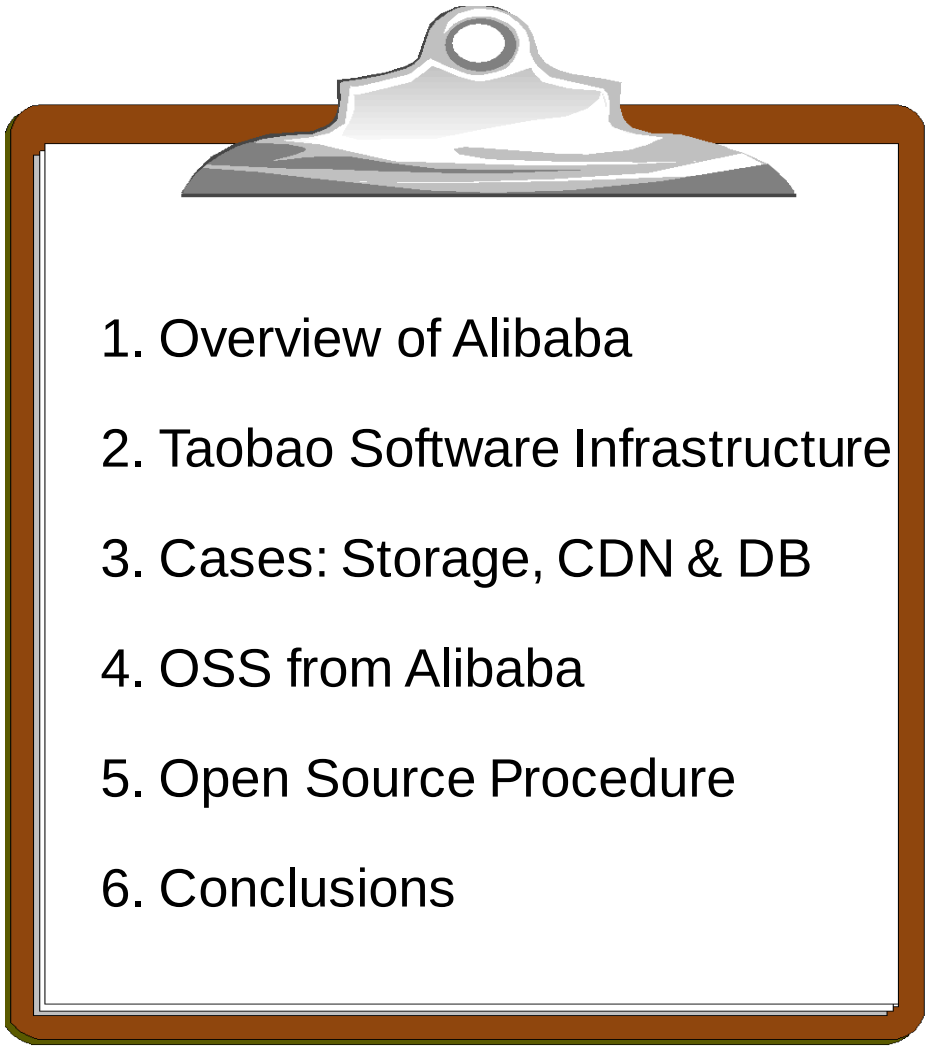
Wensong Zhang

Alibaba Group

The 11th Northeast Asia OSS Promotion Forum

2012.11.13

Agenda



1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

Alibaba Group Service Portfolio

淘宝网 Taobao.com

C2C

淘宝商城 mall.taobao.com

B2C

一淘 淘宝旗下网站

Search & pricing comparison

Alibaba.com Global trade starts here.™

Global e-commerce platform for small businesses

阿里巴巴 1688.com

E-commerce platform for Chinese small businesses

支付宝 Alipay.com

payment

聚划算 -juhuasuan.com- 品质团购每一天

Group Shopping

阿里云计算 Alibaba Cloud Computing

Cloud OS + Service

YAHOO! 中国雅虎

China Yahoo!

Taobao DataCube 数据魔方

Data Analytics

taohua 淘花

Digital Media
(video, music, book, like iTune)

Mobile e-commence
(Ali Cloud phone, Taobao & Alipay client)

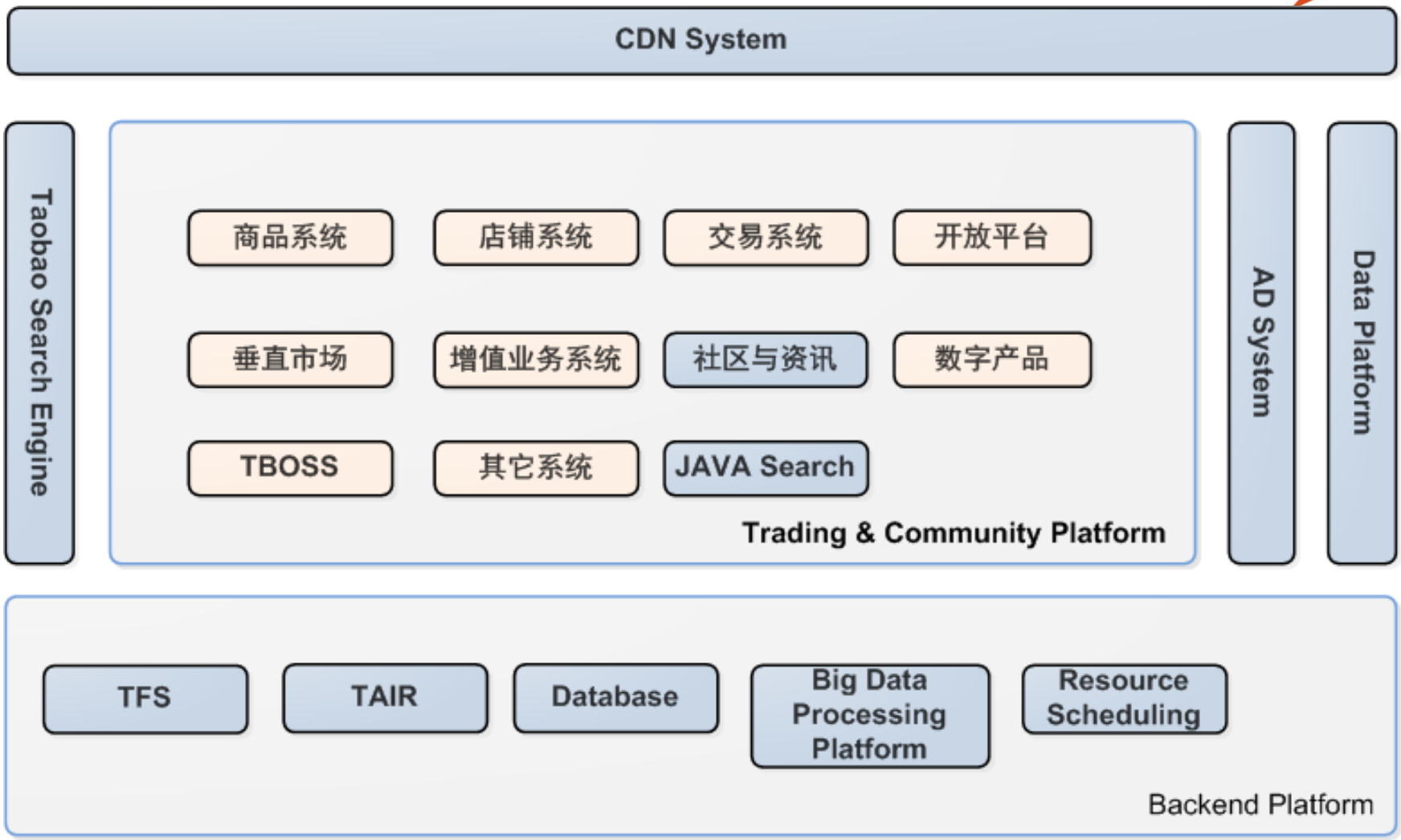
Taobao.com Overview

- The merchandise value of online shopping in China reached about 785 billion RMB in 2011, in which Taobao took about 80% market share
- Taobao.com create 2.7 million direct job opportunity in 2011.
- Alexa Traffic Rank: 13 （12～18）in the world
- In 12 December 2011, there were over 120M unique visitors, the peak traffic of CDN is 856Gbps
- Over 800 applications on the web sites

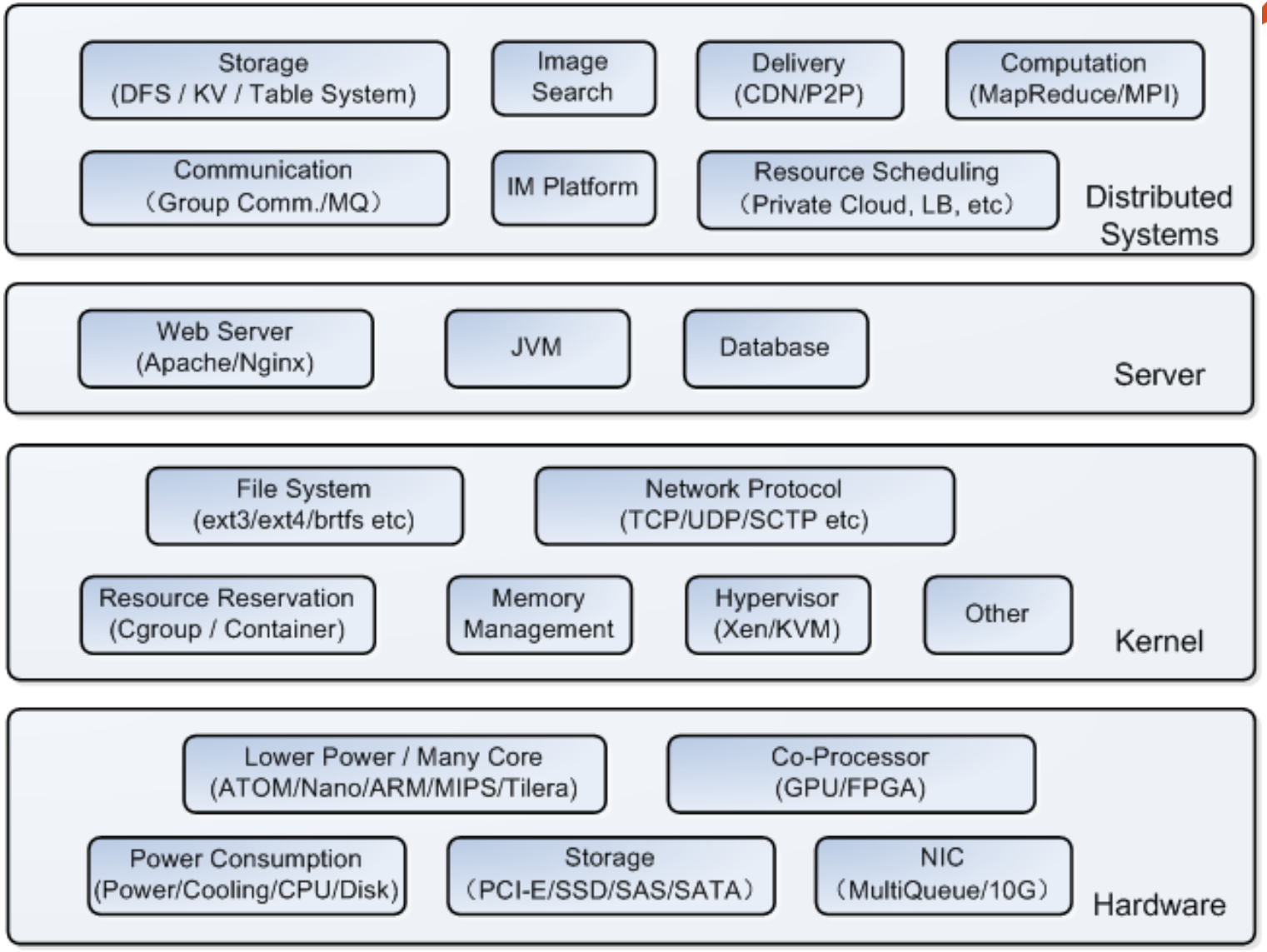
Agenda

1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

Taobao System Architecture



Software Infrastructure



Major Software

- CDN: the largest picture CDN system
  - based on open source LVS+Haproxy+Squid+Bind
  - capability of shipping 2400Gbps traffic
- TFS: distributed object storage (home grown)
  - Capacity of 12PB space, in which over 8PB is used
  - Every GB picture space cost 3.4RMB/Y, ->2.5 RMB/Y
- TAIR: distributed memory cache & K/V system
  - Integrate open source Redis and LevelDB
  - Provide disaster tolerant support
- OceanBase: Distributed Table System (home grown)
  - Support 100 billion records in a table, and transaction

Major Software（2）

- Big Data: Using Hadoop platform
  - One single Hadoop cluster has about 3000 servers
  - About 70PB space, and 48PB is used
  - Running over 150,000 jobs a day
- Database: optimizing MySQL with high speed non-volatile memory and multiple level tuning
- WangWang IM： Home grown, over 10M simultaneous users, 99.97% available in 2011
- Web Server Platform: Nginx is deployed for over 150 applications, over 4000 servers; TMD system; Tengine open source project

Major Software（3）

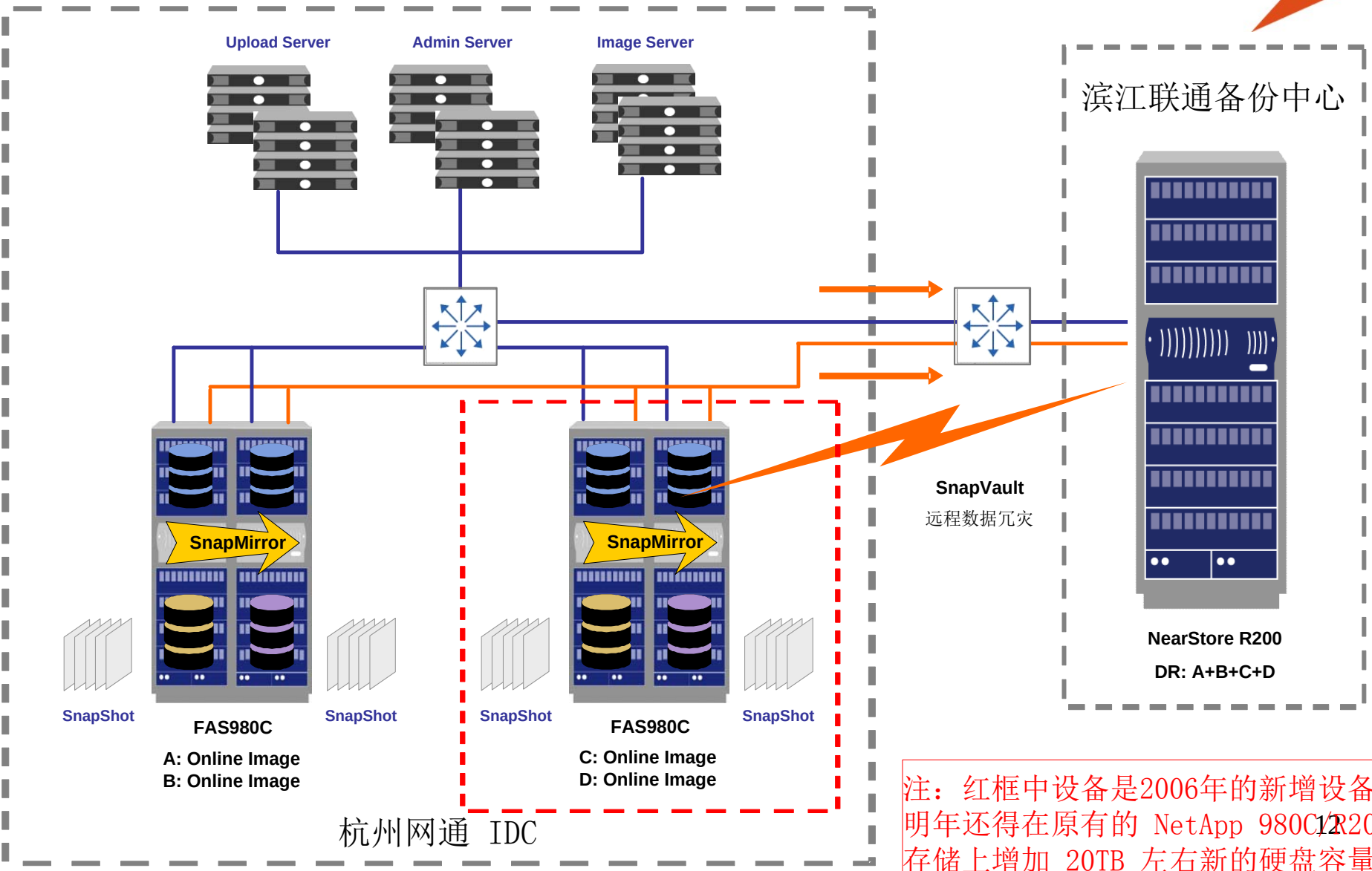
- Underlying Software
  - Develop Taobao JVM based on OpenJDK
  - Maintain Taobao Linux kernel based on Red Hat
  - KVM + Sheepdog for internal testing platform
  - Load balancing solution based on LVS
  - Network Mirror system based on open source software

- **Taobao Platform is completely built on open source and home grown software**

Agenda

1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

Picture Storage before 2007



注：红框中设备是2006年的新增设备，明年还得在原有的 NetApp 980C/R200 存储上增加 20TB 左右新的硬盘容量。

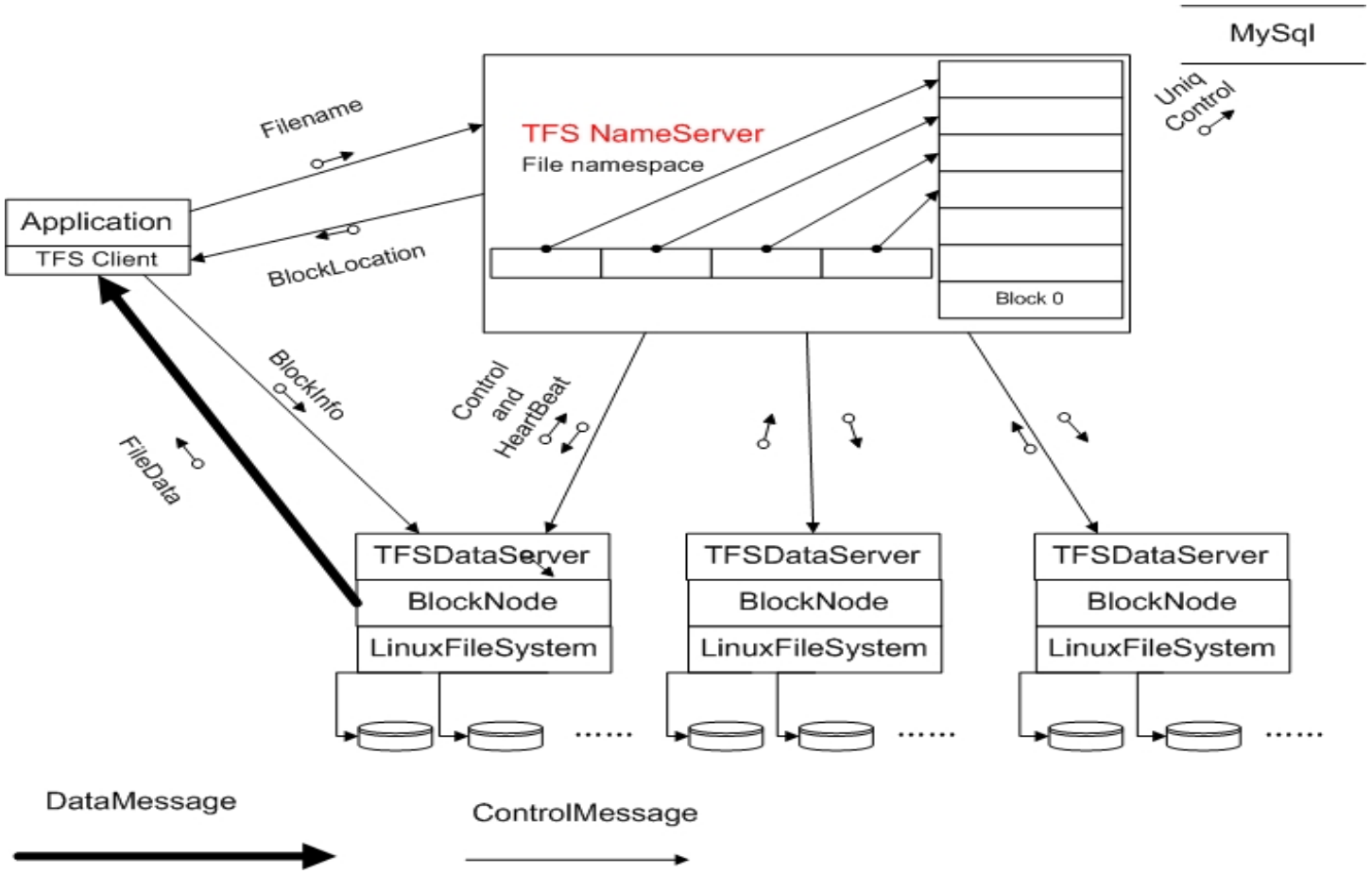
System Requirement

- System requirements
  - Data safety is becoming more and more important
  - Data is tripled every year
- Commercial storage products could not meet requirements any more
  - There is no optimization for little object
  - The number of files is too huge to support
  - The number of network connections reach its limit
  - The cost is too high, 10T NAS cost over 1M RMB at that time
  - There is no grantee in disaster tolerance

TFS 1.0

- June 2007
  - TFS（Taobao File System）1.0 went to production
  - Distributed storage for massive small objects
  - 200 PC Servers(146G*6 SAS 15K Raid5)
  - File Number: several billions
  - Deployed: 140TB
  - Used: 50TB

TFS 1.0 Architecture



TFS1.0 Features

- Master-backup Name Servers and multiple Data Servers
- Data Server runs on Linux servers with directly attached
- Each block is 64M in size
- **Filename (Object ID) contains metadata information, in order to keep metadata extremely small in Name Servers**
  - E.g. T2auNFXXBaXXXXXXXX_!!140680281.jpg include its block_no and object_no.
- Each block can have multiple copies.
- Use ext3 file system to store block

TFS 1.3

- June 2009

  TFS（Taobao File System）1.3 on production

- TFS Cluster（2010.8.22）
  - 440台PC Server (300G*12 SAS 15K RPM) + 30台PC Server (600G*12 SAS 15K RPM)
  - File Number: tens of billions
  - Deployed: 1800 TB
  - Used: 995TB
  - Name Server only uses 217MB memory for metadata

TFS Development

- TFS was released in open source in September 2010
- TFS 2.0 has already been used in production
  - Support large files
  - Support directory using external MySQL
  - Add resource center for QoS
- Onging
  - Performance tuning, and cost saving
  - Hybrid storage(SSD/SATA), object migration
  - Implementing Erasure Coding

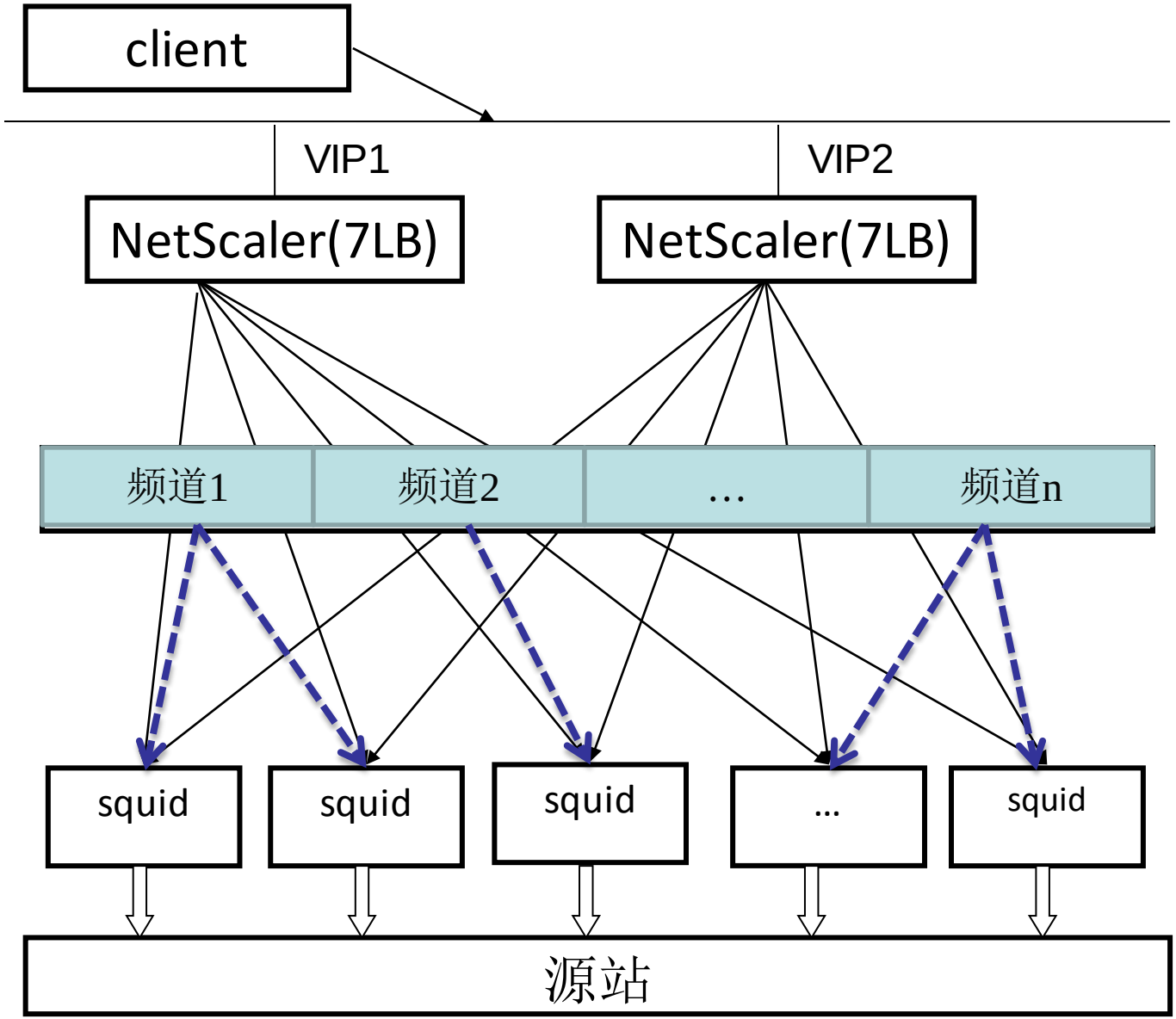
Agenda

1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

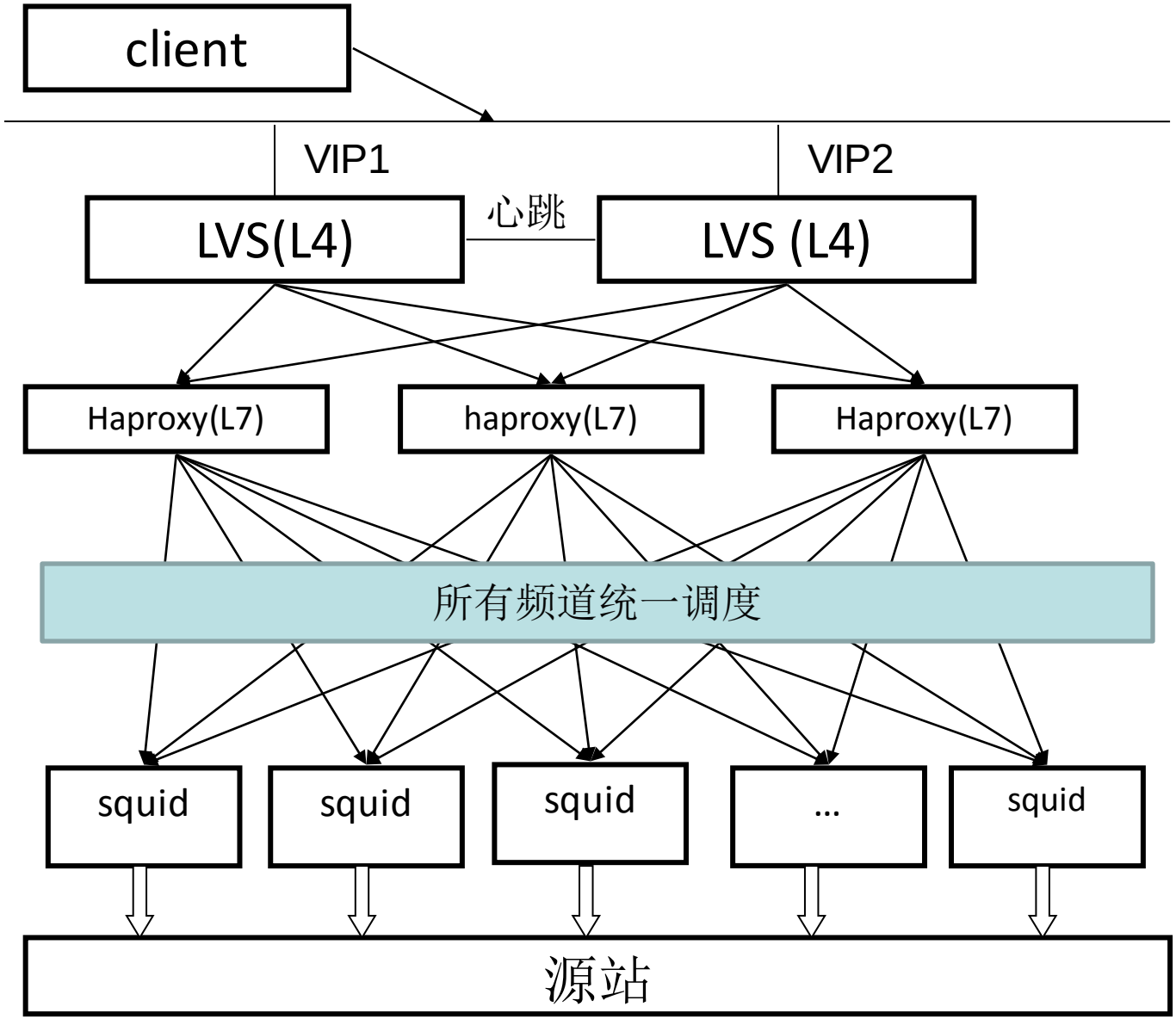
Taobao CDN System

- Issues
  - Commercial products: performance bottleneck, less features, and instability
  - Challenges in scale, performance, availability and manageability
- Develop own CDN system
  - New architecture and tuning on CDN site
  - CDN monitoring platform
  - Global load balancing system
  - CDN real-time object purging system
  - CDN configuration management system

CDN Site – Old Architecture



CDN Site – New Architecture



CDN Site Comparison

Features	New Architecture	Old Architecture
Traffic Distribution	☆☆☆☆☆	☆☆☆
Maintenance	☆☆☆	☆☆☆
Anti-attacking	☆☆☆☆	☆☆☆☆
Self-Control	☆☆☆☆☆	☆☆☆
Price	☆☆☆☆☆	☆☆☆
Scalability	☆☆☆☆☆	☆☆
Flexibility	☆☆☆☆☆	☆☆

- Scalability： one VIP address can ship 100G traffic with 10Gigabit NIC
- Flexibility: consistent hashing scheduling can make it easy to add/remove servers, where only 1/(n+1) objects needs moving

Squid Optimization

- Implement TCOSS based on COSS, FIFO + hot objects is kept, supporting 1T size file
- Squid memory optimization: 10M objects can save about 1250M memory in index
- Use sendfile to send objects in disk
- IO optimization: one request will need about 0.9 IO operation in average
- Use SSD+SAS+SATA hybrid storage, develop dynamic object migration algorithm

$$migration\_weight * \frac{frequency}{size^{migration\_power}}; \quad migration\_power \in (0, 1]$$

Agenda

1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

Commercial DB Solutions

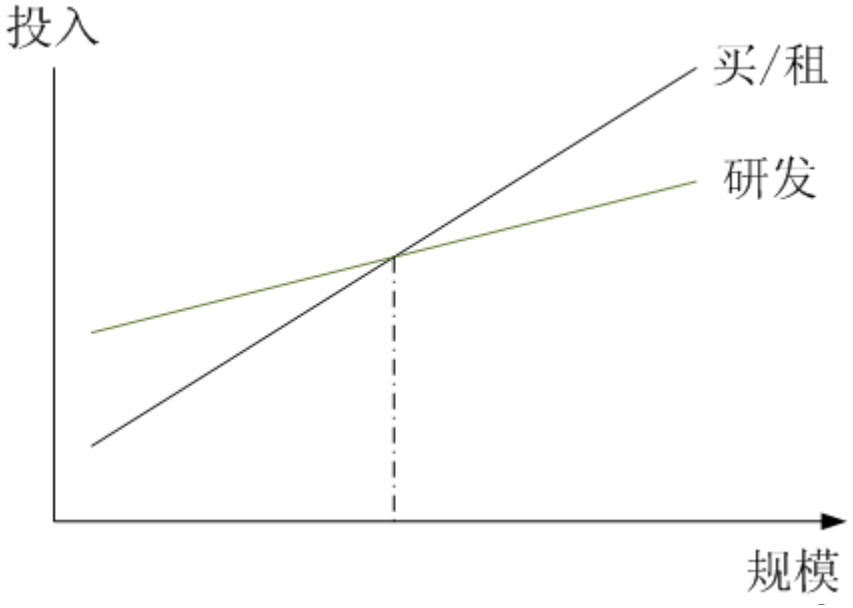
- IOE = IBM + Oracle + EMC
- Advantages:
  - Stable, and having a lot of features
  - Rich tools for operation
- Disadvantages:
  - License royalty is expensive, hardware cost is high
  - Centralized architecture, not scalable
  - Software black box

MySQL Development

- 2008: Using MySQL in non-critical applications
- 2010: Building MySQL development team
  - Combining High-speed non-volatile memory and multiple layer system optimization
  - TDDL data sharding middleware was getting mature
  - Some core applications started to migrate to MySQL
- 2011: all core databases were migrated to MySQL
  - UIC and IC Databases have MySQL 16*2 cluster each
  - Transaction Center Database: MySQL 16*2 cost is near 4M RMB (IOE is over 20M RMB), TPS was increased from 9000 to 128K

Experience

- Commercial software cannot meet requirements of large-scale systems
- Open source software plus self development can have good control, and good scalability
- Economies of scale, R&D investment is rewarding
- Optimization is long-term
- “If we build, they come” vs “if they come, we build”, the latter is more smooth



Agenda

1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

OSS from Alibaba

- Alibaba have released over 100 pieces of software in open source so far, including frontend, backend, database, file system, hardware, and so on
- Alibaba also contribute changes back to upstream

TFS

- Taobao File System
- http://tfs.taobao.org
- Distributed Object Storage System
  - Optimized for small files/objects
  - High Availability, Reliability
  - Big Data, and high concurrency
  - Lost cost
  - Linear scalability

Tair

- Taobao Pairs
- http://tair.taobao.org
- Distributed K/V storage system
  - Auto disaster-tolerance
  - Support memory-cache and persistence
  - Support multiple storage engine for different applications
    - Mdb
    - Redis
    - LevelDB
    - …

OceanBase

- Distributed database system
- http://oceanbase.taobao.org
- Features
  - Support 100 billion records in one table
  - Support database transactions
  - Support linear scalability
  - Compatible with MySQL protocol
  - Proven in many Taobao big data applications

KISSY

- A powerful JavaScript framework
- http://docs.kissyui.com
- Features
  - Modularization
  - Extensible
  - Full features
  - Rich components

Linux Kernel

- http://kernel.taobao.org
- One of Ext4 core development team
- Over 150 patches accepted by kernel
- Company contribution rank: 115
  - Statistics from 2006
- Virtualization project Sheepdog
  - Major contributors

No. 112	MathEmbedded Consulting	160 (0.06%)
No. 113	General Electric	158 (0.06%)
No. 114	Real-Time Remedies	156 (0.06%)
No. 115	Tao Bao	140 (0.05%)
No. 116	Open Nandra	137 (0.05%)
No. 117	Tilera	135 (0.05%)
No. 118	US National Security Agency	134 (0.05%)
No. 119	Barco	129 (0.05%)
No. 120	CSR	128 (0.05%)
No. 121	secunet Security Networks AG	124 (0.05%)
No. 122	MSC Vertriebs GmbH	123 (0.05%)
No. 123	RisingTide Systems	118 (0.04%)
No. 124	VIA Technologies, Inc.	115 (0.04%)
No. 125	OKI SEMICONDUCTOR	114 (0.04%)
No. 126	Realtek	113 (0.04%)
No. 127	Ksplice	109 (0.04%)
No. 128	Eukrea Electromatique	108 (0.04%)
No. 129	Solid Boot	106 (0.04%)
No. 130	Nuvoton Technology	104 (0.04%)
No. 131	Fusion-io	103 (0.04%)
No. 132	Adaptec	101 (0.04%)
No. 133	AudioScience	98 (0.04%)
No. 134	Selenic Consulting	94 (0.03%)
No. 135	CISCO	89 (0.03%)
No. 135	MEV Limited	89 (0.03%)
No. 135	Gaisler Research	89 (0.03%)
No. 135	Holoscopio Tech.	89 (0.03%)
No. 139	Candela Tech.	88 (0.03%)
No. 140	GNU	80 (0.03%)
No. 141	Telargo	79 (0.03%)
No. 141	Tower Technologies	79 (0.03%)
No. 141	IDT	79 (0.03%)
No. 144	OMICRON electronics	77 (0.03%)
No. 145	Apple	73 (0.03%)
No. 145	Digi International	73 (0.03%)
No. 147	Tensilica	70 (0.03%)
No. 147	SMSC	70 (0.03%)
No. 149	Wacom	67 (0.02%)
No. 149	Infor	67 (0.02%)
No. 149	University of Queensland	67 (0.02%)
No. 152	KFKI Research Institute	66 (0.02%)
No. 152	emlix Gmbh	66 (0.02%)
No. 154	Coraid	65 (0.02%)
No. 154	Microgate	65 (0.02%)

Hadoop

- Maintaining own version of Hadoop, including HDFS, MapReduce, Hbase and Hive, supporting Pig and Mahout from eco-system。
- Hadoop official Chinese document translation
- Hive
  - Contributed over 20 patches, about mutli-distinct aggregation, JDBC interface and authentication etc.
- HBase
  - Contributed about 47 patches
  - Short the recovery time of HBase

MySQL

- Maintaining own branch of MySQL
  - Performance improvement
  - New features
- http://mysql.taobao.org
- About 3000 MySQL servers
- Working with Oracle, Percona and Mariadb, having good cooperation.
- Signed OCA with Oracle, over 20 patches were accepted.

Alibaba Group 阿里巴巴集团

JVM

- Tuning OpenJDK VM
  - Performance improvement
  - Customize JVM from Taobao application requirements
- The deployment of own JVM version is 14% in 2011, will reach 100% in taobao.com and tmall.com in 2012.
- Contributed 16 general patches to Oracle JVM
- Contributed developers to Oracle JVM core team as well :)

WebX

- A Powerful Java Web Framework
- http://www.openwebx.org/
- Features
  - Based on Java Servlet API
  - Good extensibility
  - Reliability is proved in large-scale web sites

FastJSON

- A very efficient JSON library written in Java
- http://code.alibabatech.com/wiki/display/FastJSON/Home
- From the 3rd party testing report, FastJSON is faster than Jackson, Gson, JSON-Lib and Hessian
- Ease to use
- Easy to extend

Alibaba Group 阿里巴巴集团

Druid

- The best JDBC connection pool written in Java
- https://github.com/AlibabaTech/druid/wiki
- Features
  - SQL Monitoring / Web-Spring-SQL Monitoring
  - Filter-Chain mechanism
  - Support monitoring statistics and log
  - Prevent from SQL injection attack
  - Resolve the memory issue of Oracle PSCache

Dubbo

- A distributed service framework empowers applications with service import/export capability and high performance RPC
- http://code.alibabatech.com/wiki/display/dubbo
- Features:
  - Remote service invocation
  - Dynamic service discovery
  - Load-balancing/failover/clustering capabilities.
  - Event publish/subscription

Alibaba Group 阿里巴巴集团

Tengine

- A web server based on Nginx
- http://tengine.taobao.org
- Features:
  - Dynamic loadable module support
  - Input filter mechanism
  - Javascript / CSS comb
  - Health checking on backend servers
  - Support pipe and syslog logging and sampling
  - Monitoring load to protect system
  - Features for ease use and maintenance

Alibaba Group 阿里巴巴集团

SeaJS

- A Module Loader for the Web
- http://seajs.org
- Features:
  - SeaJS pursue simple and nature coding style
  - Good maintenance for JS code

Alibaba Group 阿里巴巴集团

Arale

- An open, simple, easy-to-use JS class based on SeaJS
- http://aralejs.org
- Features:
  - A simple and consistent coding style
  - Components

QLExpress

- A lightweight script engine
- http://code.taobao.org/p/QLExpress/wiki/index/
- Features:
  - Support standard Java syntax
  - Support self defined operator, operator overloading, function definition, macro definition, data delay loading
  - Compile and execution for performance
  - Used in a lot of systems in Taobao, and outside too.

TBSchedule

- A job scheduler for tasks running in JVM among multiple hosts
- http://code.taobao.org/p/tbschedule/wiki/index/
- Features
  - Job can be dynamically created, stopped, and run on multiple servers
  - Configure time period for task execution
  - A unified control console, which can adjust parameters
  - One simple interface
  - Used in a lot of systems in Taobao, and outside too.

Tsar

- Monitoring and data collection tool for operation
- http://tsar.taobao.org
- Feature
  - Simple, ease to deploy, stable, low overhead
  - Modularization design, easy to add more modules
  - Support real-time data collection and display
  - Can work with Nagios, and send alert messages

Alibaba Group 阿里巴巴集团

TProfiler

- A Java performance analysis tool, widely used in Taobao production systems
- http://github.com/taobao/tprofiler
- Features:
  - Support analyzing and sampling
  - Record method execution time and times
  - Generate hot methods and analysis report

TimeTunnel

- Real-time data transmission platform based on thrift
- http://timetunnel.taobao.org
- Features:
  - High performance
  - Real time
  - Sequential
  - Reliable
  - High available
  - Easy to extend

TaoKeeper

- ZooKeeper monitor
- http://code.taobao.org/p/taokeeper/src/
- Features
  - Monitor ZooKeeper connection numbers, watcher numbers, node numbers
  - Can set threshold for monitoring alerts
  - Can generate daily/weekly monitoring report
  - Can check current status of machines

MetaQ

- Java Massaging middleware
- http://metaq.taobao.org/
- Features:
  - High performance, TPS can reach 45000 with 2Kbytes message on Gigabit ethernet
  - Support ordered messages
  - Support message filtering
  - Support transaction in message consumption at consumer side
  - Support distributed deployment

TDDL

- TDDL(Taobao Distributed Data Layer) is a distributed data access engine for data shreding
- https://github.com/taobao/tb_tddl
- Features:
  - Data access routing
  - Support one write and multiple read
  - Easy to extend, and auto data migration

Automan

- An auto testing framework
- http://code.taobao.org/p/AutoMan/wiki/index/
- Features:
  - Design based on Page-Model, separate “element lookup” and “component operation”
  - Easy to use and maintain

Toast

- An auto-test job scheduling platform
- http://toast.taobao.org
- Features:
  - A general-purpose scheduling platform
  - Support timed run, and manual run
  - Provide monitoring and management on testing machines

SQLAutoReview

- A SQL auto-auditing tool
- http://code.taobao.org/p/sqlautoreview/src/
- Features:
  - Parse SQL statement from SQLMap
  - Build create index script for every SQL statement
  - Merge all create index script with exist indexes on these tables

Diamond

- A persistent configuration management center
- http://code.taobao.org/p/diamond/wiki/index/
- Features
  - Simple
    - HTTP interface to get configuration
  - Reliable
    - Multiple level protect and disaster-tolerance support
    - Reliable notification on configuration update, to ensure client to get the latest data

Lower Power Server

- Each NODE has 3 HDDs
- 24 x 2.5” SATA/SSD
- 2U 8 nodes
  - Hot pluggable
  - Lower power comsuption
  - Lower cost
- Server node configuration：
  - Intel® Atom™ D525 with 2 cores
  - Intel® ICH9R Chipset
  - 4GB memory DDR23 800MHZ SO-DIMM w/o ECC
  - LAN: Intel 82574L 2*1GB
  - HDD:
    - 1* SSD 80G ,
    - 2* 2.5” SATA 500GB

Green Computing

- Open source green computing
- http://www.greencompute.org/

开源绿色计算

English Version

首页 | 项目介绍 | **设计规范** | 合作赞助 | 论坛讨论 | 联系方式 | 新闻公告

设计规范名称	采用CPU型号	版本	发布时间	下载地址	面向应用
主板设计规范	Intel Atom D525	V1.0	2011-9-27	中文版 英文版	CDN的缓存服务应用
机箱和电源设计规范	Intel Atom D525	V1.0	2011-9-27	中文版 英文版	CDN的缓存服务应用
服务器测试规范	Intel Atom D525	V1.0	2011-9-27	中文版 英文版	CDN的缓存服务应用

Agenda

1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

Open Source Committee

- Found Alibaba Open Source Committee with ten persons, who is from
  - Developers mainly
  - Lawyer
  - Security analysis
- The goal of Alibaba Open Source Committee is to help promote open source
  - Ensure open source procedure
  - Give advice on open source

Alibaba Open Source Procedur

- Engineers initiate open source request
- Need approval from their manager and the head of big division
- Security analysis team will check document and code for any security issues
- Alibaba Open Source Committee record this software in open source list
- Offer advices on open source license issue
- Host project on Taocode and also synchronize it to github

Encourage Open Source

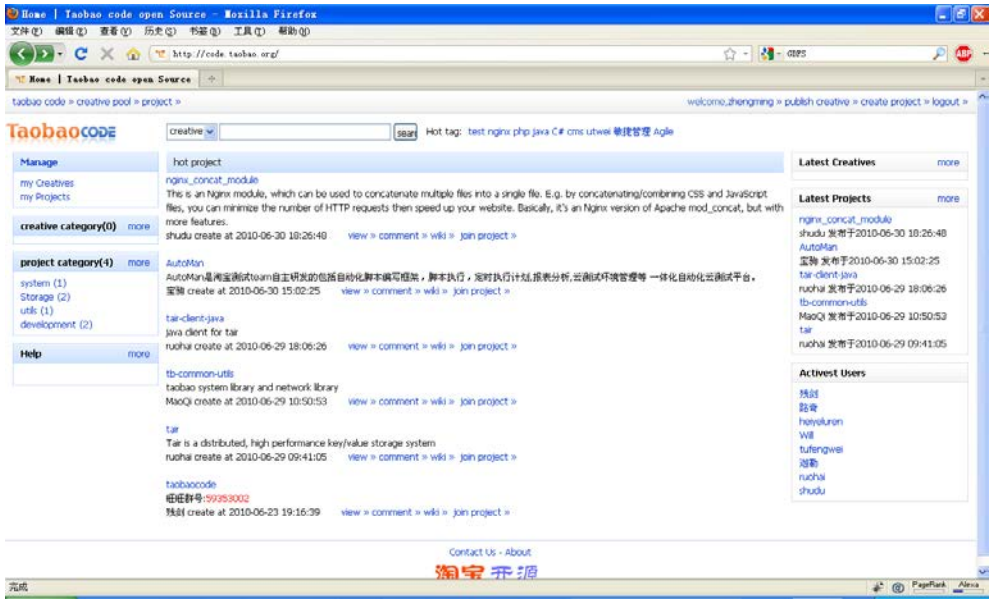
- Accomplishment in open source project is the best encouragement to engineers.
- Host internal open source meetings from time to time, for sharing and brain storming.
- Reward excellent open source projects.

Open Source Licenses

- GPL vs. BSD vs. MIT vs. Apache
  - GPL is better to guarantee development of open source project
- Alibaba's consideration on licenses
  - Most projects choose GPL。
  - Some libraries use BSD or Apache license
  - Some just follow the license from upstream
- Alibaba Group is the copyright owner
  - (C) 2007-2012 Alibaba Group Holding Limited

Open Source Platform

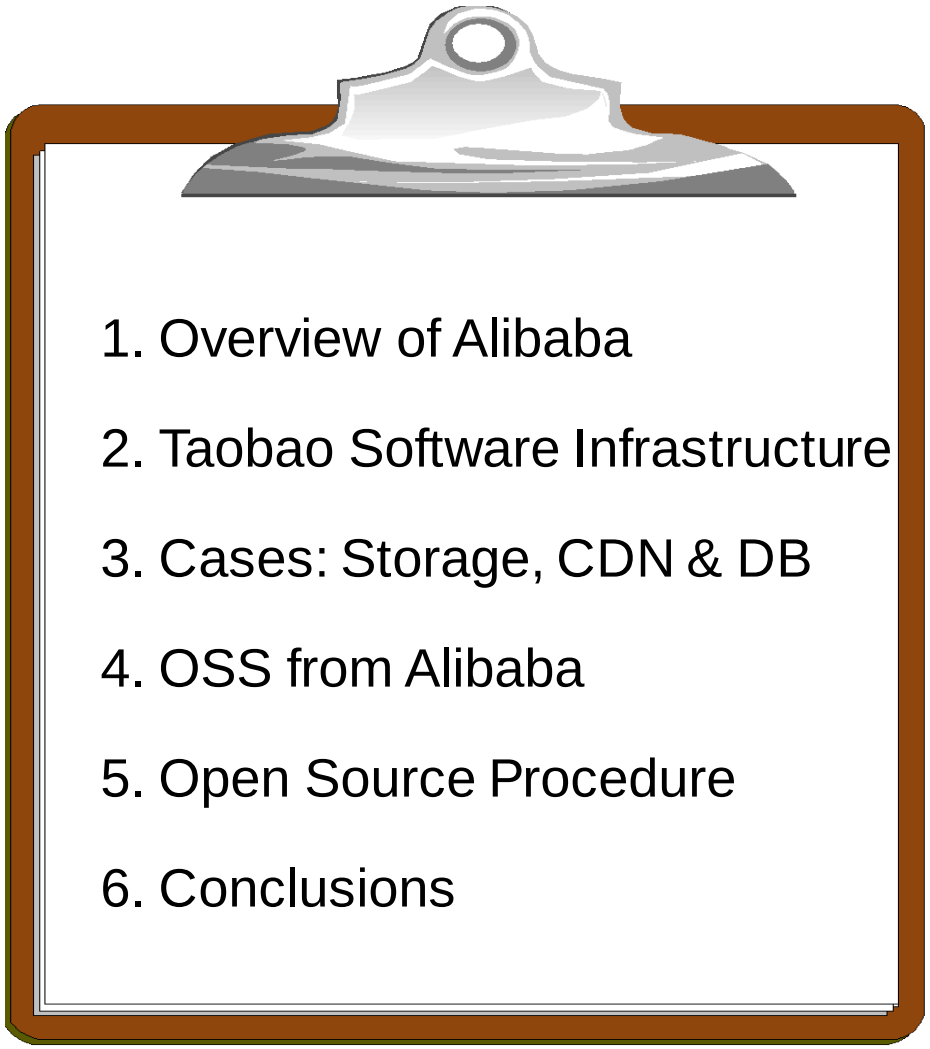
- code.taobao.org
- Goal
  - The platform itself is open source too
  - Fast access in China
- Status
  - There are 373 projects
  - Mature open source projects is mainly from Taobao
  - Non-Taobao projects are becoming active, some is good.



Open Source Platform

- code.alibabatech.com

Agenda



1. Overview of Alibaba
2. Taobao Software Infrastructure
3. Cases: Storage, CDN & DB
4. OSS from Alibaba
5. Open Source Procedure
6. Conclusions

From Engineer Side

- Alibaba engineers gain some benefits from doing open source projects
  - More accomplishment
  - Working with more developers and hackers, good for skill improvement
  - Getting more users
  - Make a longer lifecycle of their code
  - Happy to see how their code is used in different ways

From Company Side

- Improve software quality through open source
  - More user requirements and feedbacks
  - More users tesing and bug reports
  - Patches from external developers
- Receive industry recognition on technology capability and open spirit
  - Engineers have strong sense of identity
  - Attracting more talents to join Alibaba

Conclusion

- Alibaba Group is the beneficiary of open source system, and actively participate in building open source eco-system.
- Hope that the company can accumulate better reputation, and attract more talents to meet the greater technical challenges in the future.
- Alibaba Group hopes to do more technology innovation with the industry in a more open manner.

Q&A

Q&A

Thanks