证券交易的低延迟挑战

黄寅飞

上海证券交易所

2012年7月

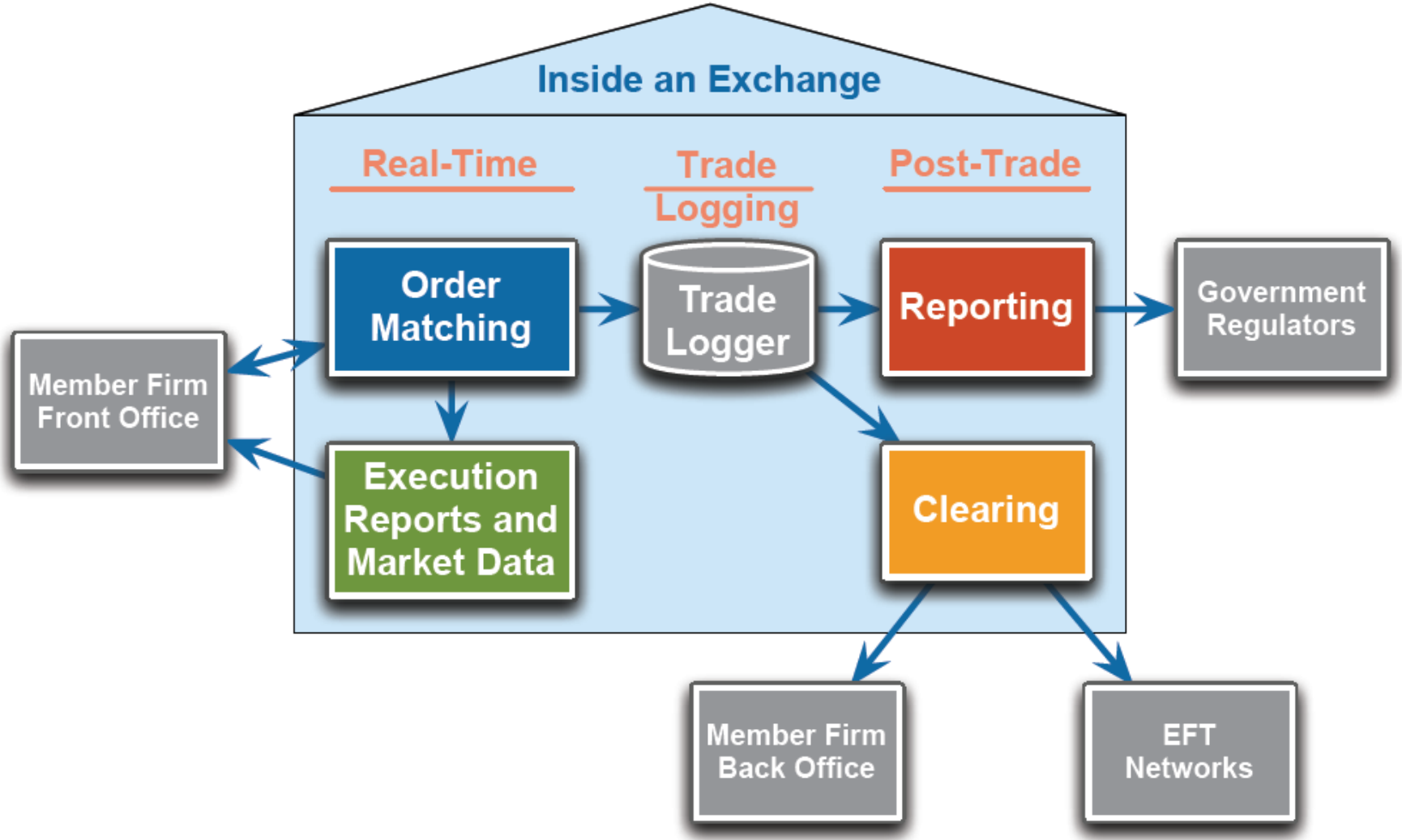
SSE SHANGHAI STOCK EXCHANGE

ADC Alibaba Developer Conference

大纲

- 行业背景
- 技术研发
- 低延迟算法
- 应用展望

证券交易系统示意图


Inside an Exchange
Real-Time
Trade Logging
Post-Trade
Order Matching
Trade Logger
Reporting
Government Regulators
Member Firm Front Office
Execution Reports and Market Data
Clearing
Member Firm Back Office
EFT Networks

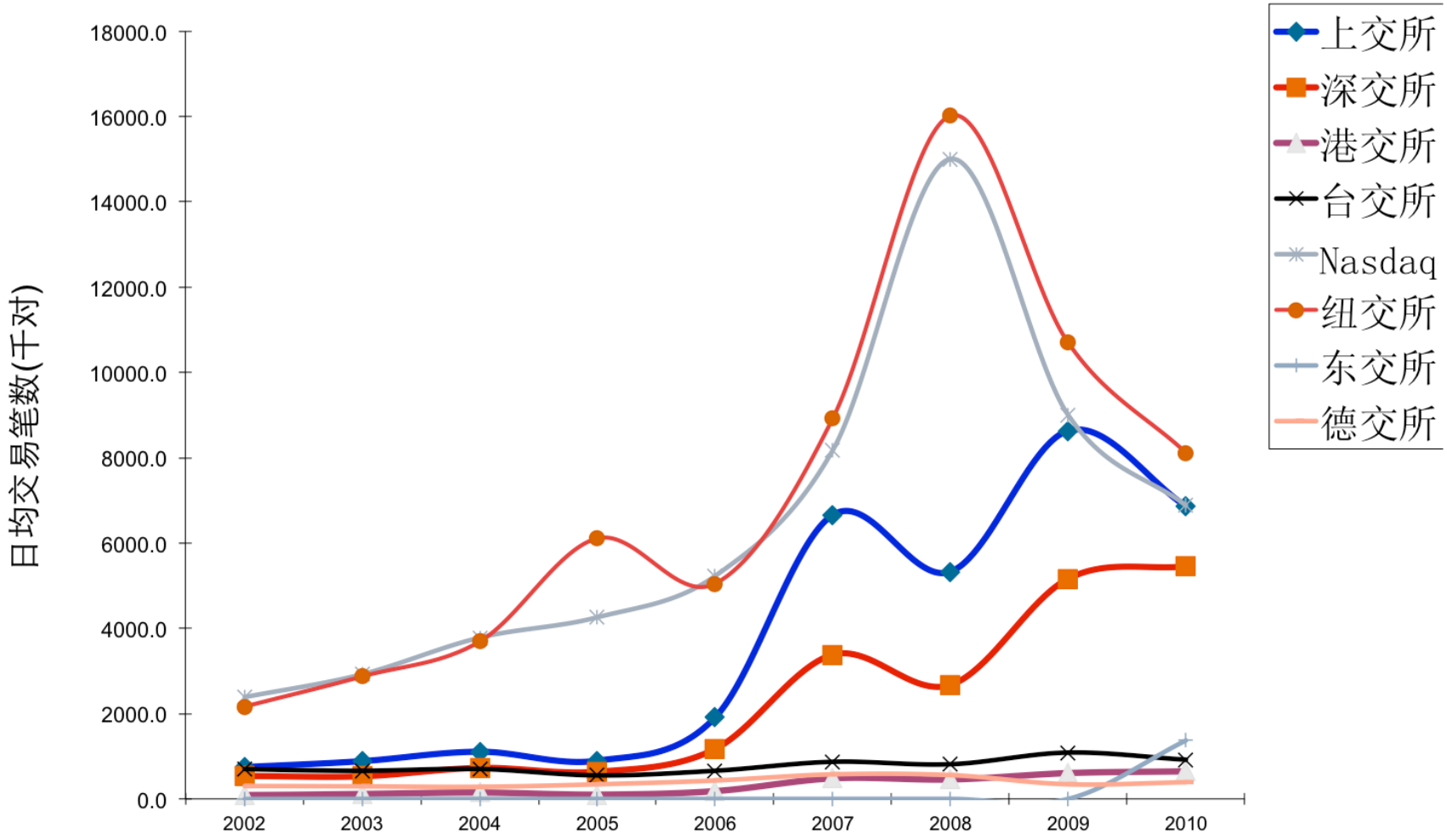
术语解释

- 交易所 – 证券公司 – 基金公司
- 行情 – 买 – 卖 – 高频交易
- 报价 – 做市商
- 集合竞价 – 连续竞价
- 内幕交易 – 市场操纵行为

全球证券交易所

- 纽约交易所**UTP/ARCA**
- 纳斯达克**INET**
- 伦交所**Millennium**
- 德交所**Optimise**
- **Direct-Edge**与**UME**
- 东京交易所**ArrowHead**
- 香港交易所**AMS**

全球第三大交易所


日均交易笔数(千对)
18000.0
16000.0
14000.0
12000.0
10000.0
8000.0
6000.0
4000.0
2000.0
0.0
2002
2003
2004
2005
2006
2007
2008
2009
2010
上交所
深交所
港交所
台交所
Nasdaq
纽交所
东交所
德交所

系统实际运行情况

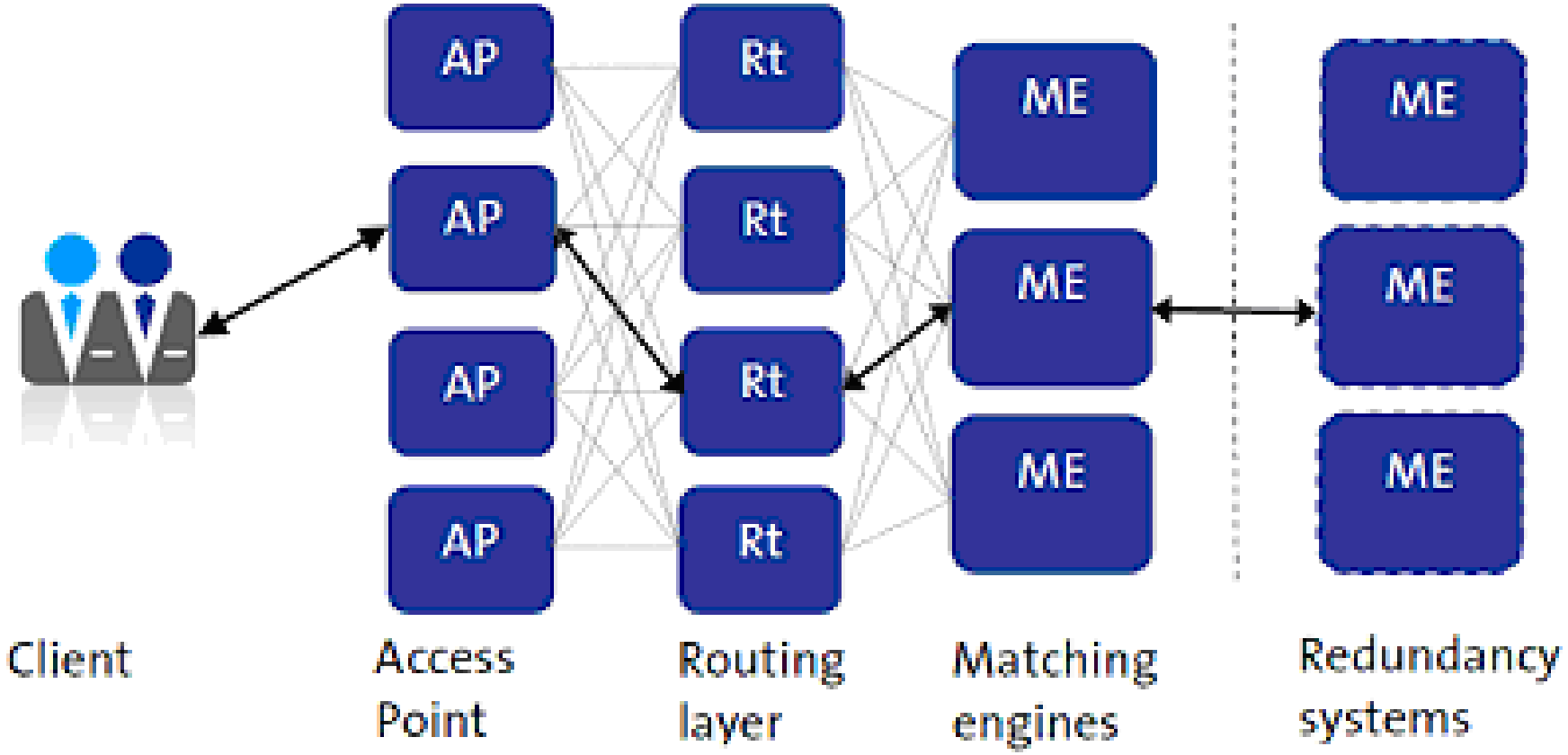
	上海证券交易所	纽约交易所	NASDAQ
日均成交对数（2010年）	687万	814万	692万

上海证券交易所	上线前系统	新交易系统
峰值处理能力（笔/秒）	20000	102131 * 2010.10.20

全球证券交易所技术架构特征

- 不直接使用商用数据库产品
- 操作系统转向**Linux**等开源系统
- 高速网络使用万兆以太网**/InfiniBand**
- 系统架构方面应用消息中间件
- 撮合引擎数据驻留内存
- 分布计算可线性扩展
- 灾难容忍能力

技术架构典型部署


AP
AP
AP
AP
Rt
Rt
Rt
Rt
ME
ME
ME
ME
ME
ME
Client
Access Point
Routing layer
Matching engines
Redundancy systems

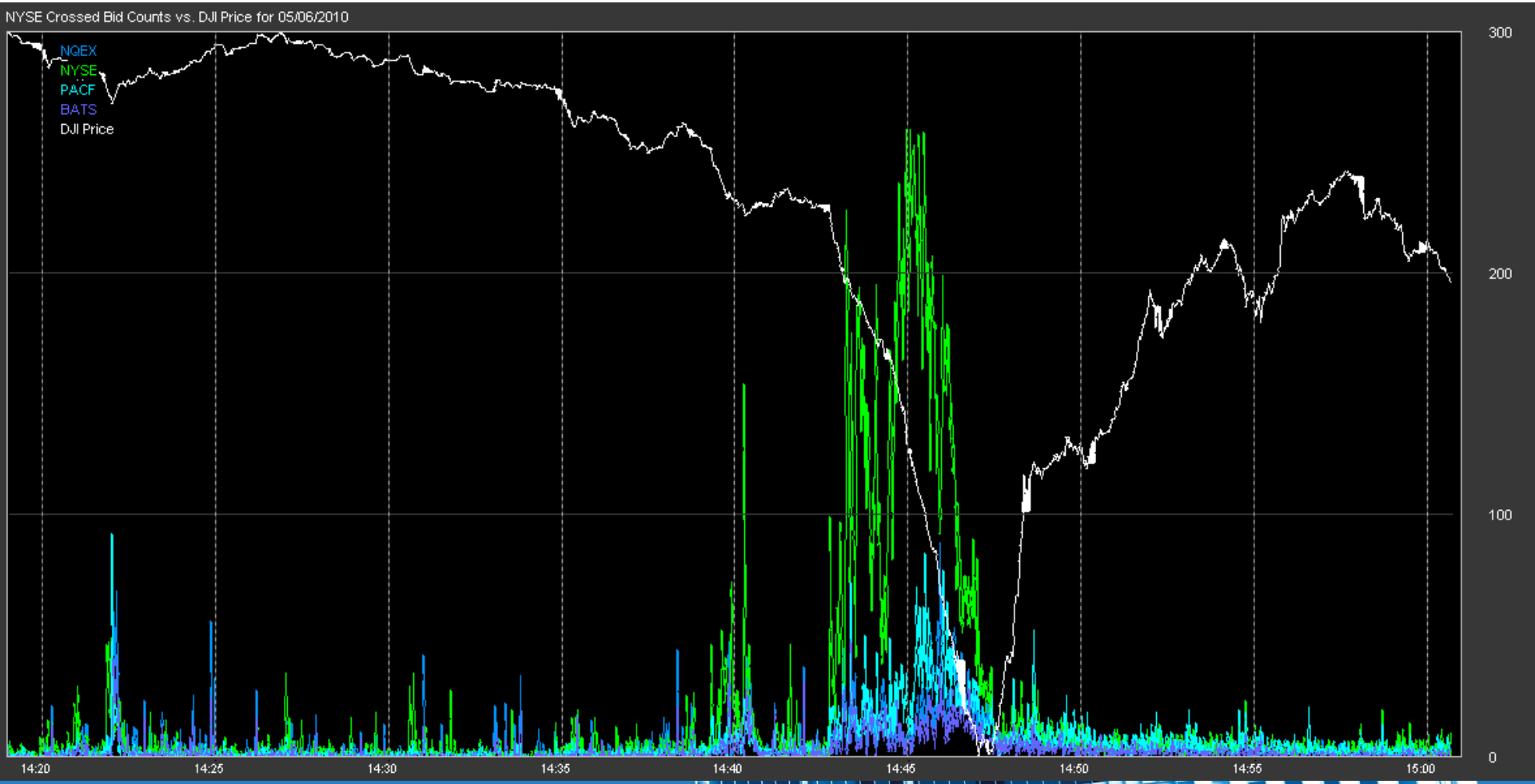
插播花絮

东交所乌龙指

- **瑞穗下错卖单交易巨亏成定局**
  - http://www.sina.com.cn 2005年12月20日10:36 财经杂志
- 12月8日，日本瑞穗金融集团旗下的瑞穗证券称，该公司当天在东

- **东京证券交易所因交易过失被判赔107亿日元**
  - 2009-12-04 21:06:45　来源: 人民网(北京)　跟贴 5 条 手机看新闻
- 人民网东京12月4日电 据共同社报道，关于瑞穗证券在2005年的股票交易中下单出错导致巨额损失一案，东京地方法院4日作出宣判，裁定东京证券交易所向瑞穗证券支付约107亿日元（约合8.26亿元人民币）。审判长松井英隆指出，下单出错的瑞穗证券过失重大，但东证提供的基础交易系统不完善，过失更大。法院算出东证也负共同责任的损失金额约为150亿日元，并裁定瑞穗证券和东证的过失比例为3比7。

2010年美股暴跌


NYSE Crossed Bid Counts vs. DJI Price for 05/06/2010
NQEX
NYSE
PACF
BATS
DJI Price
300
200
100
0
14:20
14:25
14:30
14:35
14:40
14:45
14:50
14:55
15:00

伦交所Millennium上线

- **伦交所股票交易系统故障 投资者升市无法沽货惹不满**
  - **http://www.17ok.com　　2008-09-09 10:16:39　　来源：汇港通讯**
- 英国伦敦交易所昨日开市后15分钟,交易系统突然故障,未能输入及执行买卖指令,无法交易,令投资者错过了华府接管两房令股市大涨的出货机会,惹起极大不满。
- **伦敦证券交易所因"技术故障"中断 备受业内指责**
  - **2011-02-27 10:08:00 来源：卫报　我要评论**
- 伦敦证券交易所正常交易时间为上午8时至下午4时30分。当天开盘前大约5分钟，交易系统“市场数据”部分出现故障，交易所停盘处理。中午12时15分，交易所正式开盘，但当天并未因中断而延长交易时间，仍在正常时间收盘。伦交所本月14日启用“千年信息技术”交易系统，这一系统先前故障频发。伦交所旗下的特阔伊斯交易平台去年10月率先启用“千年信息技术”交易系统，但启用第二天系统出现故障，导致交易中断大约2小时。

港交所披露易

- **港交所李小加称网站遭黑客侵袭 停牌为确保公平**
  - http://www.sina.com.cn 2011年08月10日 18:32 新浪财经
- 新浪财经讯 8月10日下午消息 香港交易所(109,-1.10,-1.00%,实时行情)(00388.HK)信息披露网站“披露易”今日出现技术故障，共有8只股票及债券被迫停牌。港交所行政总裁李小加表示，网站遭到黑客恶意侵袭，目前正在努力修复中，虽然停牌引发投资者不满，但能确保信息公平。

Facebook上市

- **传纳斯达克拟提交Facebook交易故障赔偿计划**
  - http://www.sina.com.cn 2012年06月06日 04:29 新浪科技
- 在Facebook股票于5月18日正式IPO上市时，纳斯达克OMX集团旗下交易系统的问题导致Facebook股票的开盘时间被推迟了30分钟，同时还导致股票经纪人成百上千万股的Facebook股票交易无法得到确认。直到Facebook股票开始交易的两个多小时以后，银行和交易公司才得以知道其订单的结果，其中一些公司从纳斯达克OMX集团那里收到通

- **纳斯达克拟4000万美元补偿Facebook投资者**
  - http://www.sina.com.cn 2012年06月07日 01:13 新浪科技
- 新浪科技讯 北京时间6月7日凌晨消息，纳斯达克(微博)OMX集团周三概略阐述了一项计划，内容是支付大约4000万美元的“一次性”款项，作为对某些在Facebook IPO(首次公开招股)首日因纳斯达克交易故障而蒙受了损失的金融公司的赔偿。

技术研发

挑战

- 高可用
- 安全
- 低延迟
- 高吞吐
- 风险控制
- 敏捷开发
- 持续交付
- 低耦合高内聚
- 人员培训
- 应急处置

应对

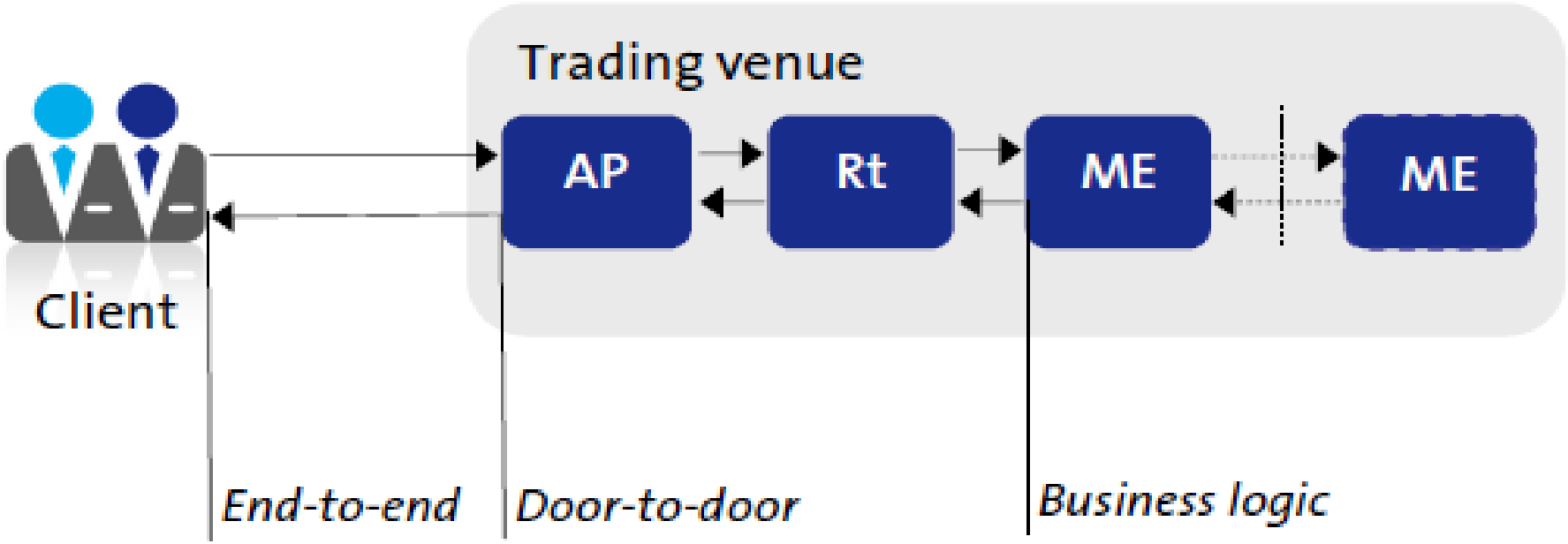
- 群狼计划：PC服务器、开源平台
- 猎豹计划：持续提速、消息接口
- 狡兔计划：两地三中心、主机托管
- HAP、VHAP
  -

科研活动

- 上证联合技术专题
  - 开源、低延迟、云计算、集群、通讯协议
  - 算法交易、文本挖掘、代码检查、FIX
- 国家科技支撑计划
  - 证券核心交易系统研发
  - 证券业云平台研发与运营
  - 《交易系统：更新与跨越》@武剑锋

低延迟算法

延迟度量



端到端　　门到门　　主机延迟

复制算法分类

- 系统的高可用性通常通过对核心组件冗余复制保证
- 对复制技术根据复制媒介的分类
  - 基于共享存储的复制 (文件复制方案)
  - 基于消息传递的复制 **(网络复制)** √
- 两类基于消息传递的复制方法:
  - 基于法定人数集(Quorum)的状态机复制→**Paxos算法**
  - 组通信系统→**虚同步 (Virtual Synchrony)**

虚同步

- 虚同步含义：在保证从应用层看上去与真正同步一样的前提下，允许调整消息顺序以提高性能
- 主要优点：
  - 性能极佳，几乎达到与IP多播相当的性能
- 主要不足：
  - 容错能力较弱，数据一致性无理论上的证明

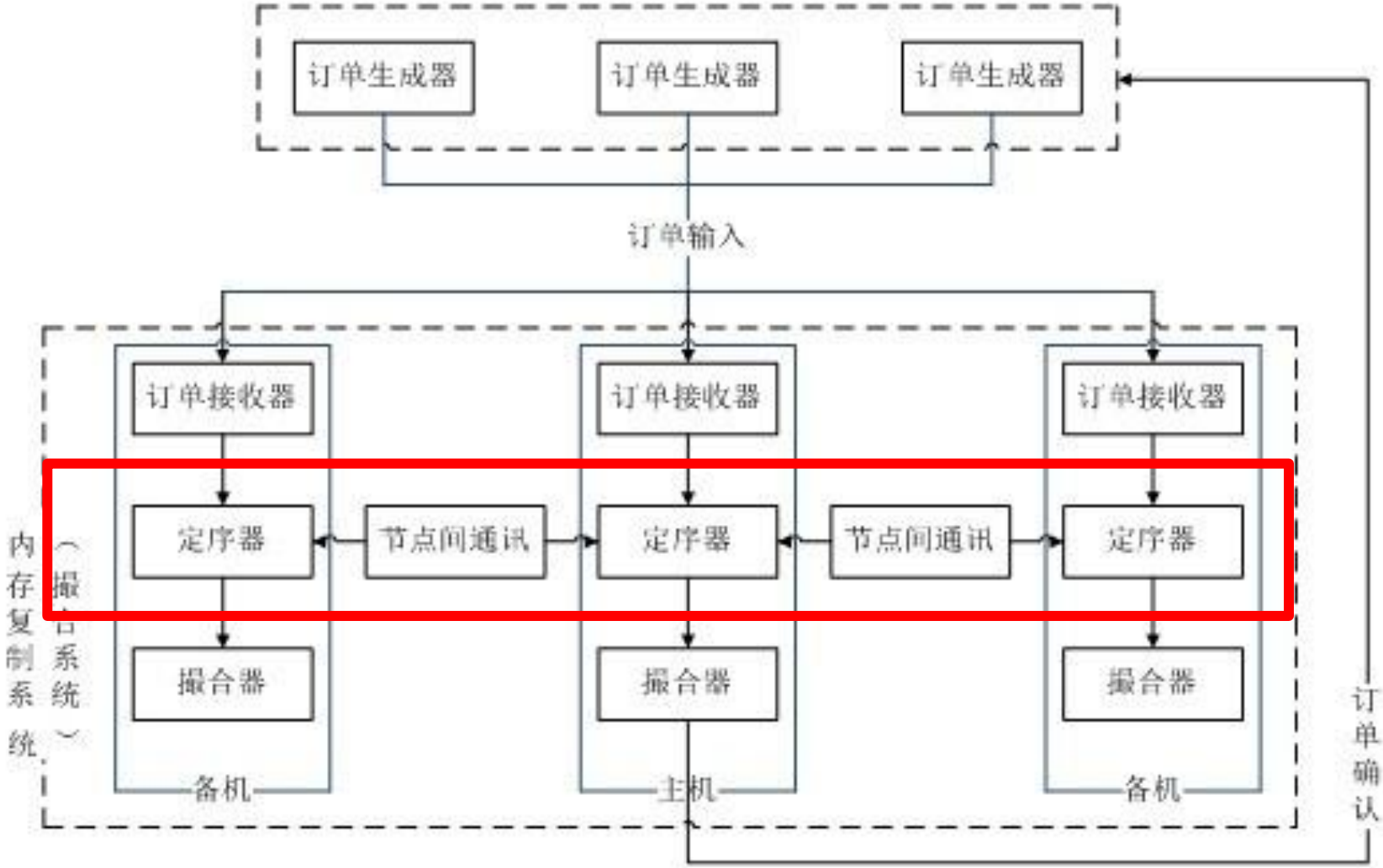
状态机复制

- 基本思想
  - 将各个撮合主机看作确定性状态机，内存数据为状态，订单消息是输入。
  - 相同初始状态->相同订单序列->相同最终状态
- 保证各结点上订单序列的一致，是保证数据一致性的关键

订单定序器

- 引入一个定序器(Sequencer)模块
  - 模块职能：为所有的订单消息选择一个全局统一的顺序，各结点按此顺序接收订单。
  - 订单处理过程中，需先由定序器进行定序，完成定序后才可进行撮合
  - 为避免单点故障，系统中需配备多个定序器。目前设计方案为各主机都配备一个定序器，相互协作完成订单信息的定序与复制。

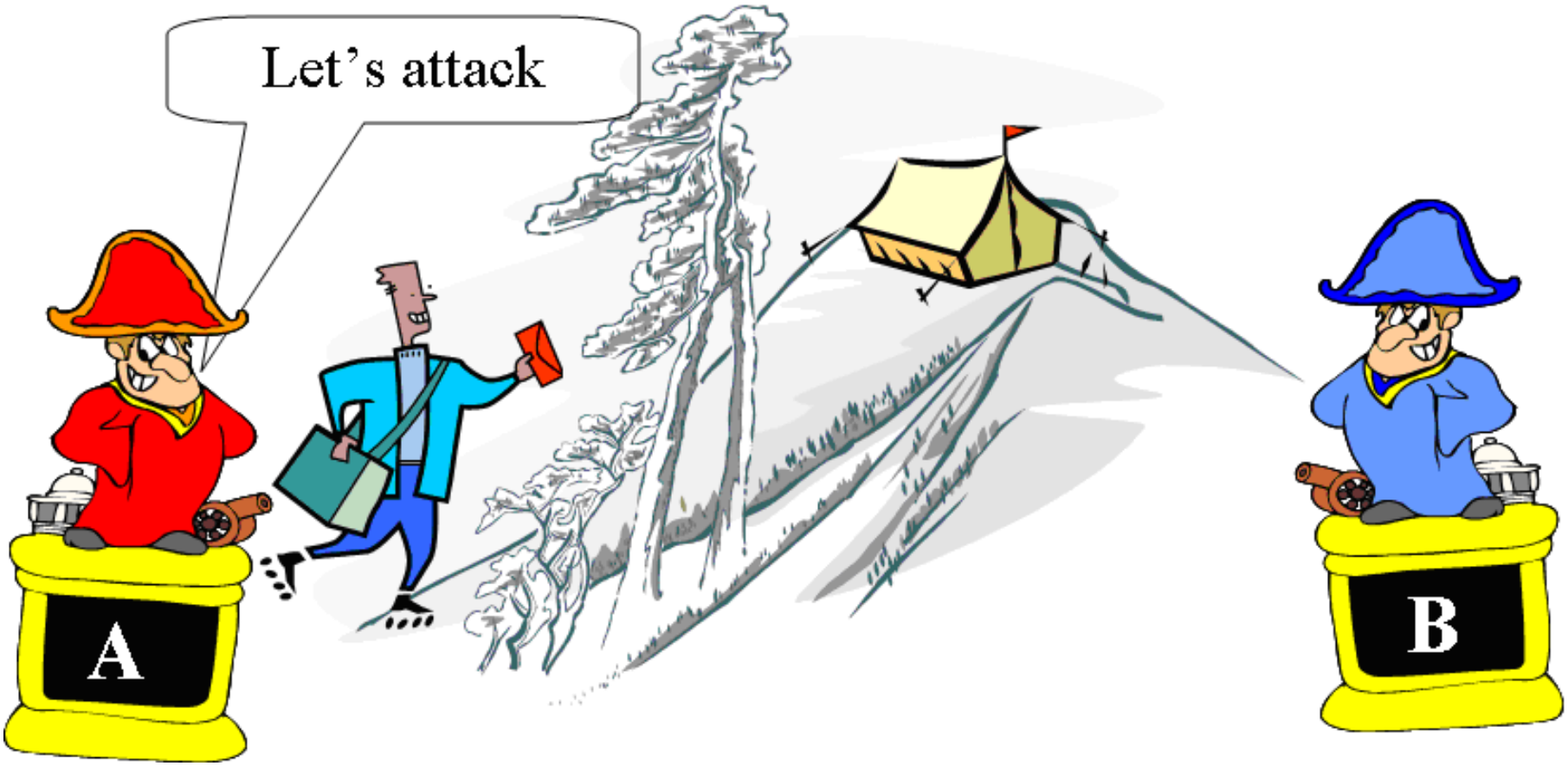
架构示意图


订单生成器
订单生成器
订单生成器
订单输入
订单接收器
订单接收器
订单接收器
定序器
节点间通讯
定序器
节点间通讯
定序器
撮合器
撮合器
撮合器
备机
主机
备机
内存复制系统（撮合系统）
订单确认

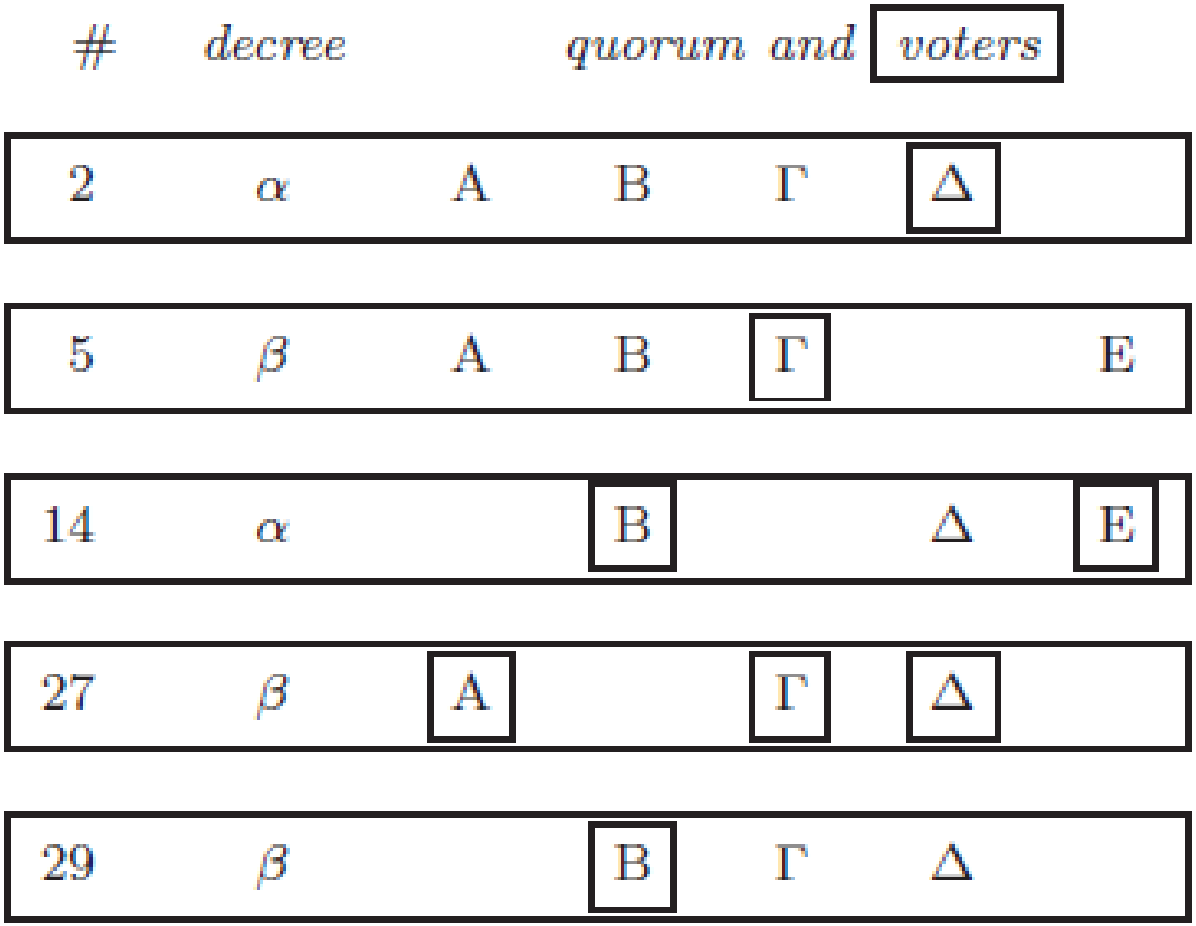
Paxos算法

- Paxos算法基于投票机制，一条订单需得到超过半数结点的赞成投票才能被全局选定。
- 协议中根据职责划分了角色：
  - Client：发送订单请求
  - Proposer：发起对订单的投票 (只保留一个，称为Leader)
  - Acceptor：响应对订单的投票
  - Learner：学习订单投票结果
- 基本概念：
  - 提案值(Proposal)：候选订单
  - 决议(Agreed Value)：已选定订单
  - 批准(Accept)：投票赞成
  - 通过(Choose)：选定某个订单，不允许再修改

Leslie Lamport



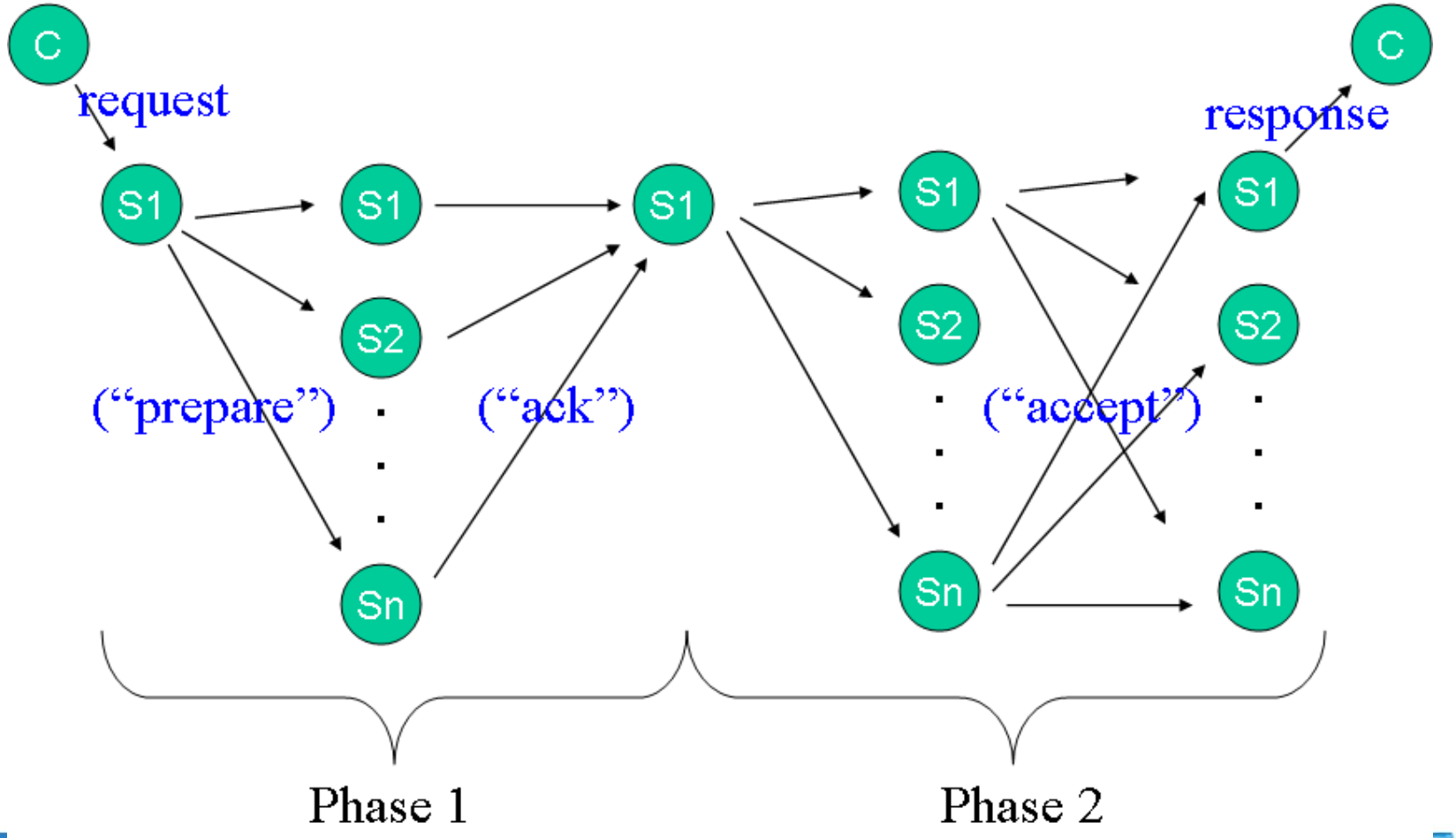
考古碎片


 # decree quorum and voters
2 α A B Γ Δ
5 β A B Γ E
14 α B Δ E
27 β A Γ Δ
29 β B Γ Δ

Paxos算法内容

- Proposer与Acceptor选定订单的过程，需通过一个两阶段的协议进行：
  - Prepare阶段：为即将发起的投票收集信息，避免发起投票的订单值与之前已通过的订单值冲突
  - Accept阶段：正式进行投票，并收集投票结果，形成决议
- Learner需对投票结果进行学习，获知决议

Paxos算法示意图


C
request
S1
S1
S1
S2
Sn
("prepare")
("ack")
S1
S1
S2
Sn
("accept")
C
response
S1
S2
Sn
Phase 1
Phase 2

Google Chubby

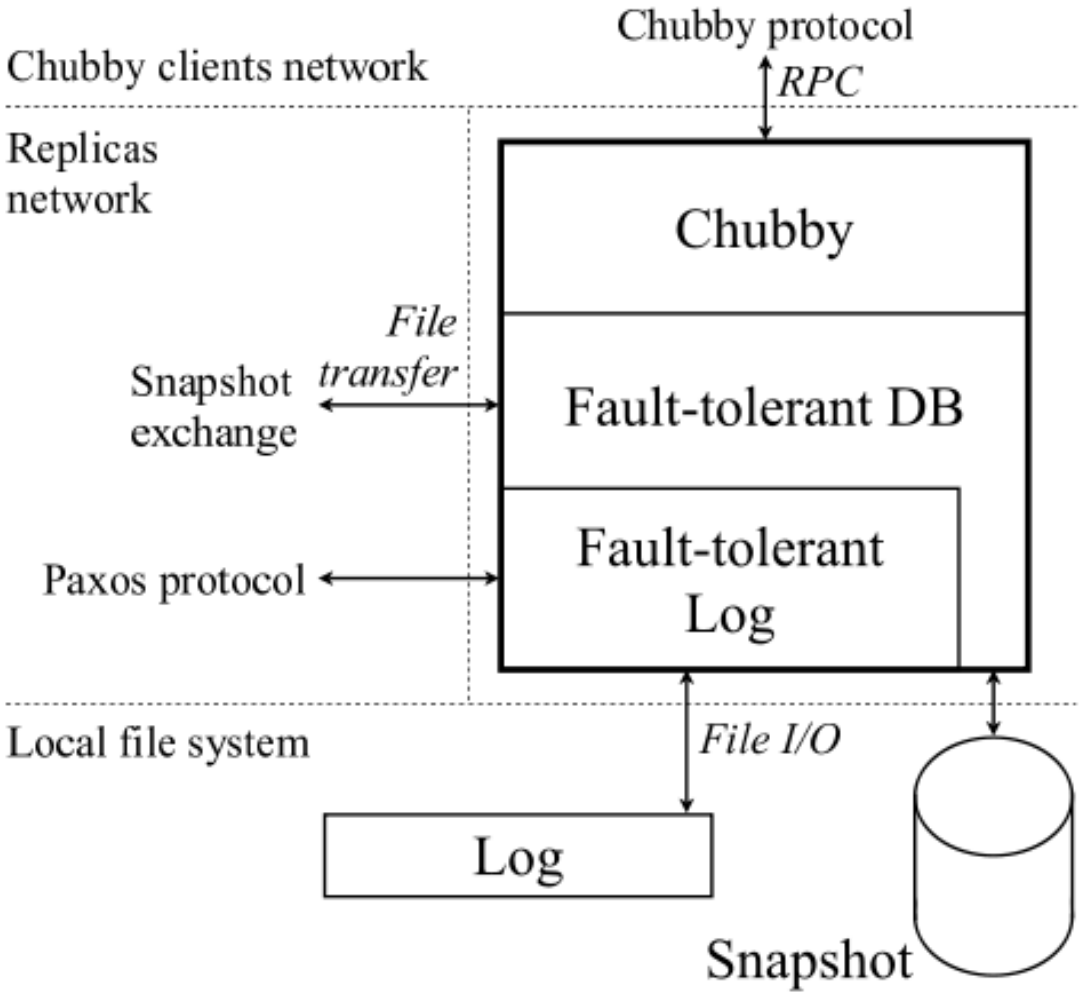
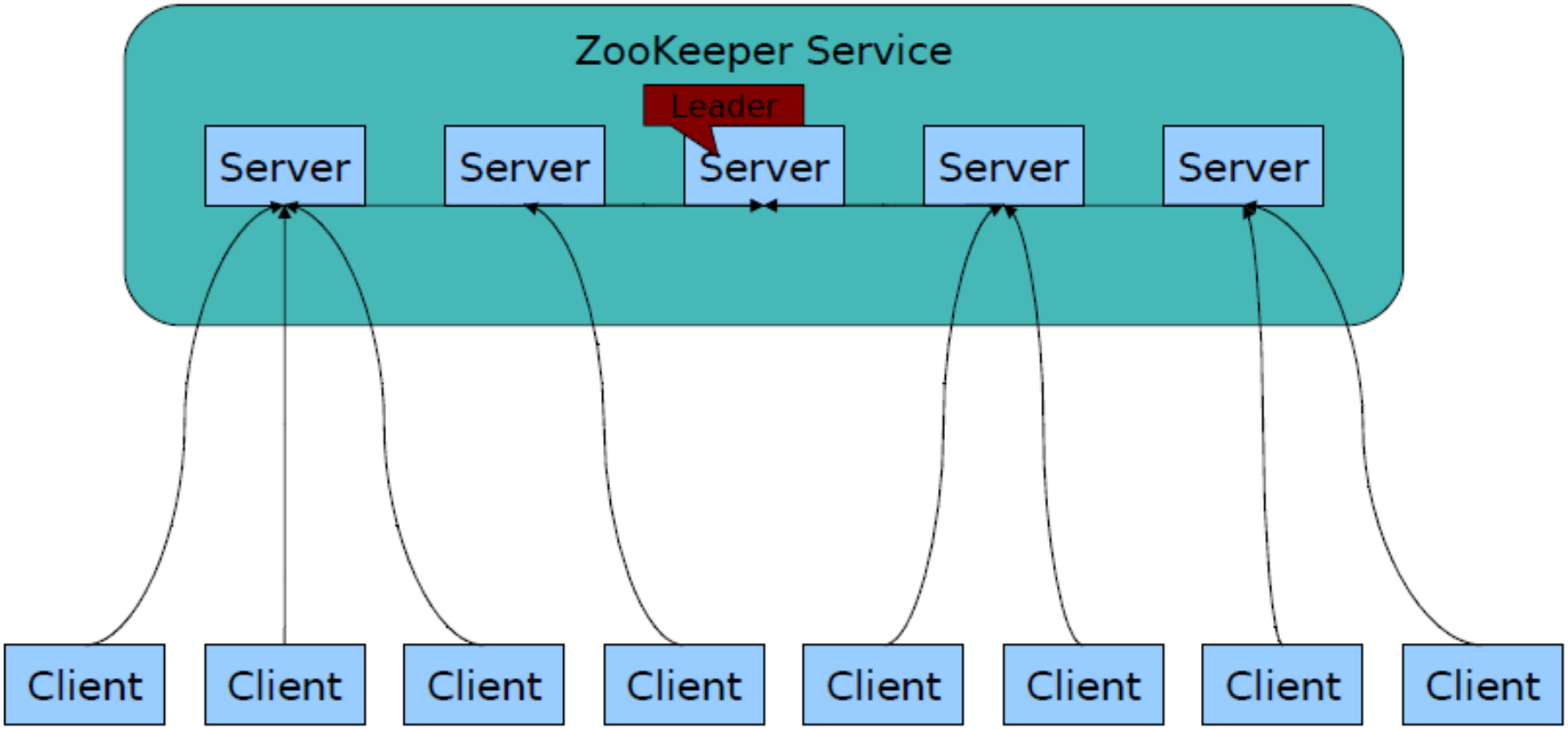


Figure 1: A single Chubby replica.

ZooKeeper


ZooKeeper Service
Leader
Server
Server
Server
Server
Server
Client
Client
Client
Client
Client
Client
Client
Client

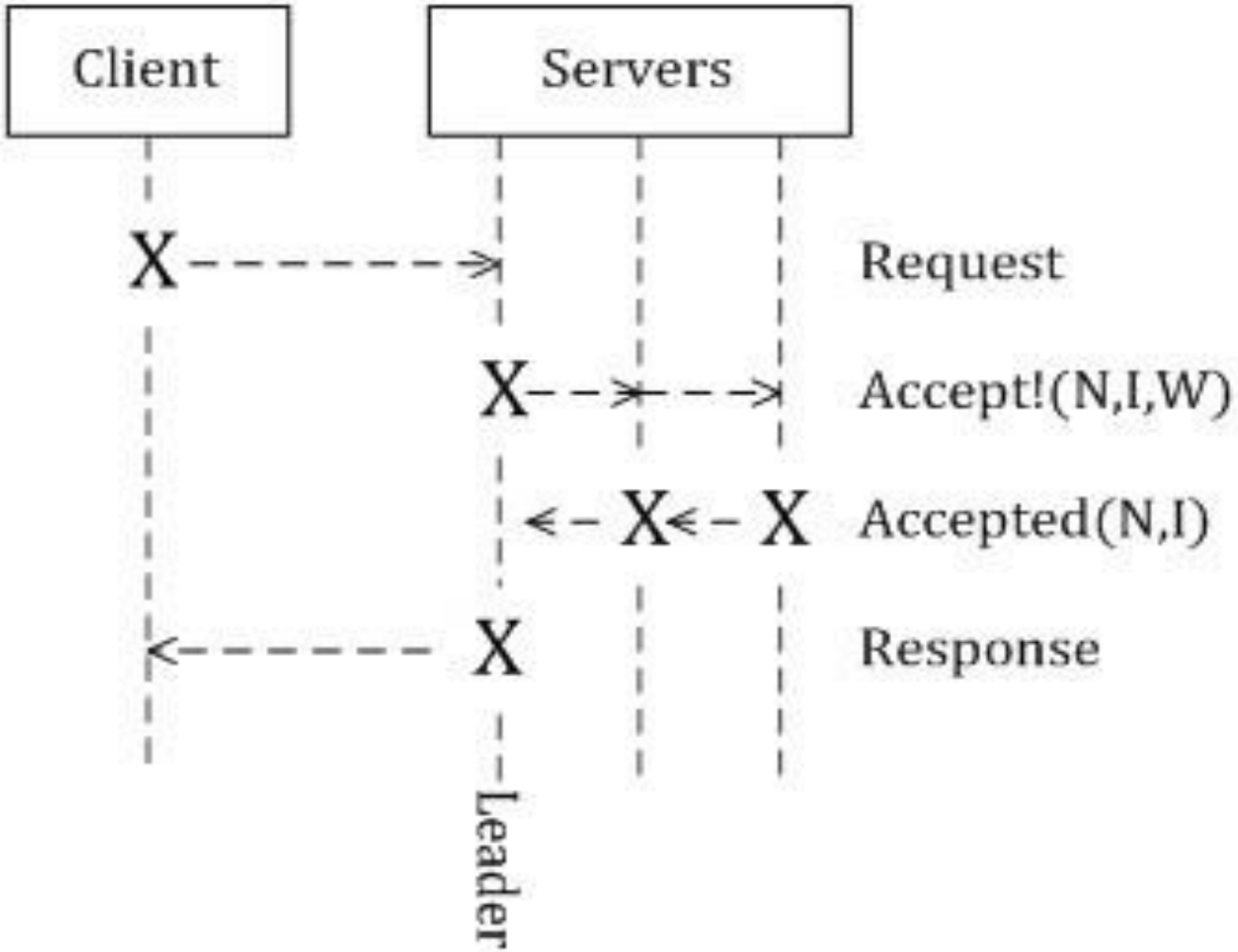
Paxos算法适用环境

- 分布式系统故障分类
  - 拜占廷故障(恶性故障）
  - 非拜占廷故障(良性故障)
- 处理拜占廷故障代价较高，证券交易系统采用专网，发生恶性故障概率低。
- 基本的Paxos算法适用于非拜占廷故障：
  - 处理器：以任意速度运行，可能故障，但不会串通、说谎以及进行使协议转向的尝试
  - 网络：消息可能花费任意长的时间到达，可能丢失、乱序或重复传送，但不会被篡改。

一致性证明

- **Paxos算法核心在于通过两个多数集合至少有一个公共结点的性质来保证一致性。**
  - **一个订单Va被选定后，则有超过半数的结点在Accept阶段对其投过赞成票**
  - **若想选定另一订单Vb，必需先在Prepare阶段收集到超过半数的响应，其中必然含有Va，从而使得Accept阶段只能发送Va**
- **Paxos算法用2N+1个结点容忍最多N个结点发生非拜占廷故障，保证数据的一致性。**

Collapsed Multi-Paxos



算法性能测算

- 采用Collapsed Multi-Paxos方案，在千兆或万兆以太网中将订单复制延迟降到百微秒级

单位：微秒	千兆以太网	万兆以太网
Client发送请求	100	50
接收及相应处理	10	10
Leader向其它结点发送提案	100	50
其它结点接收提案并进行相应处理	10	10
其它结点发送确认，Leader接收	150	75
Leader处理接收到的确认	30	30
提案通过表决后Leader进一步处理	100	100
处理完成Leader向Client发确认	100	50
网络通讯总用时	450	225
主机处理总用时	150	150
总计	600	375

万兆以太网中实验：
测试了复制消息大小在256B - 4KB之间，进行3结点以及5结点复制的延迟

主结点上测得延迟基本都在**100微秒到200微秒**之间；
客户端测得延迟基本都在**200微秒到350微秒**之间。

与估计值基本相符

应用展望

开源平台

- **HAP基础库**
- 高速并发内存数据访问库
- 审计日志高效事务处理库
- 高可用多机备份路由管理库

高速网络选型

- 万兆以太网：**10G/40G/100G**
  - 通用性好，适用范围广
- **InfiniBand：40G/80G**
  - 专用于站点内高速总线，技术成熟

RDMA

- 远程直接内存访问
- 不占用CPU资源
- 编程接口较复杂

可靠组播

- PGM
- 兼具UDP的快速和TCP的可靠保序
- ACK与NAK

消息中间件

- 通讯模块 => 消息中间件
- 清晰的应用/架构分层
- 保证会话层面的消息可靠送达
- 配合应用层面的订单和广播重传

开放 – 合作

- 开放的思维
- 开放的交易机制
- 开放的平台
- 开放的基础设施
- 开放的新技术研究

THANKS!

SSE SHANGHAI STOCK EXCHANGE
Address: SECURITIES TOWER
NO.528 South Pudong road Shanghai 200010 PR.China