分布共享数据服务理论与技术研究

魏恒峰
导师: 吕建　黄宇

南京大学软件所

July 4, 2016

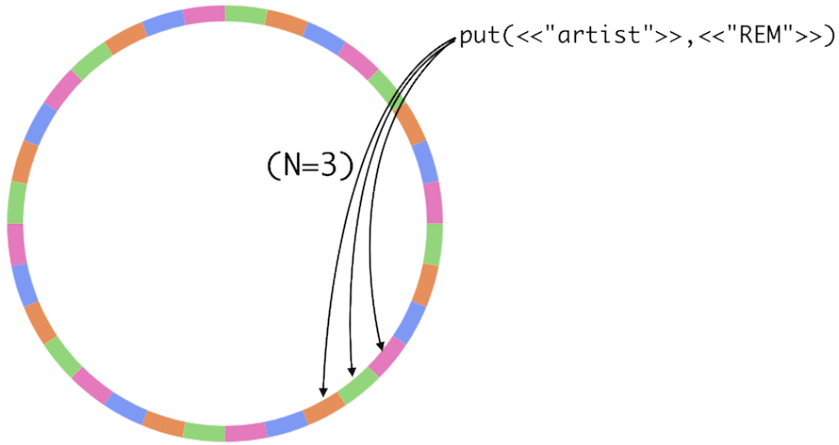
分布共享数据服务理论与技术研究

分布式应用

TODO: 动画: 分布部署

分布数据

(**TODO:** 动画: partition + replication)



distributed data : partition & replication

分布数据典型应用 (I)

图: 分布式存储系统 (开源 [左] & 商用 [右]).

应用需求 [Facebook@OSDI'10] vs. "分布数据":

低延迟: 就近访问副本数据

高可用性, 高容错性: 备份容灾

分布数据典型应用 (II)

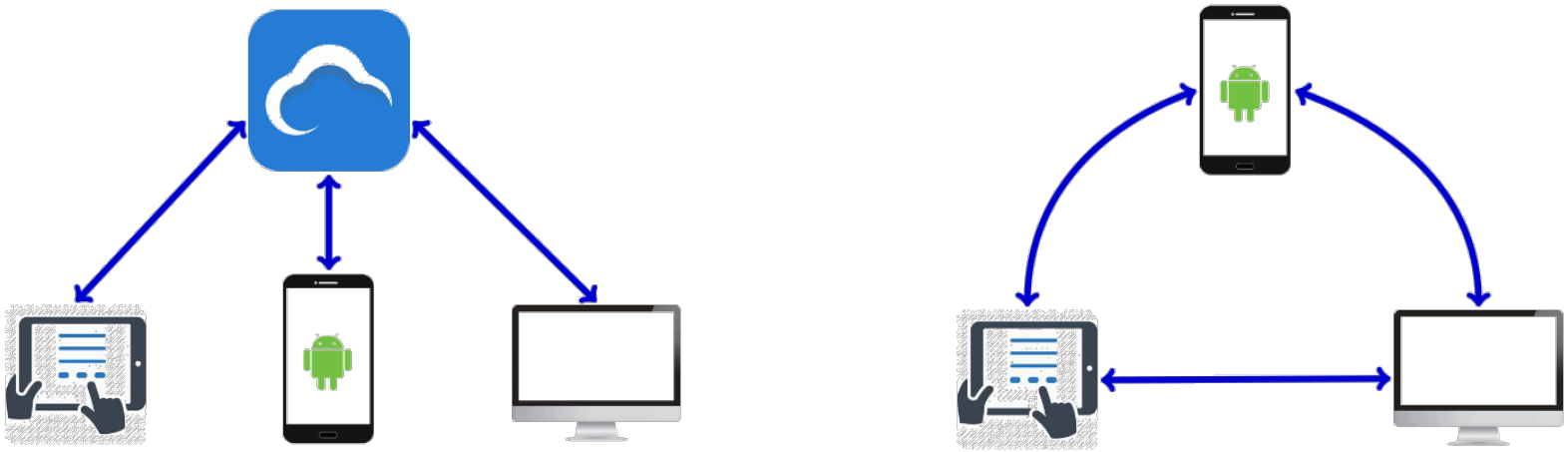
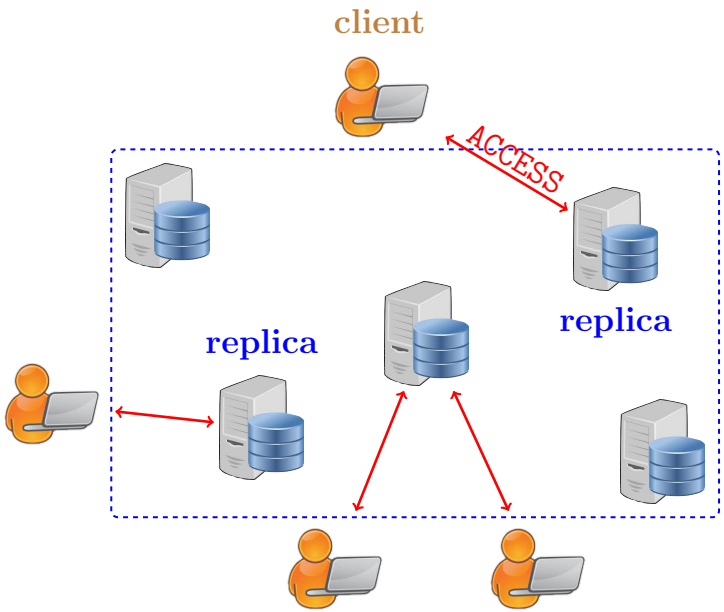


图: 个人多设备文件共享 ([基于云] C/S 结构 [左] & P2P 结构 [右]).

应用需求与特点 [Strauss@MIT Thesis'10] vs. "分布数据":

- 功能需求: 文件副本
- 网络断连: 备份容灾; 离线可用

分布式应用访问分布数据



TODO: 重绘: 显示 partition & replication

分布共享数据服务

- 简化上层应用的开发

TODO: 图: 分布共享数据服务

分布共享数据服务

TODO: 图：分布共享数据服务

- 简化上层应用的开发
- 提供共享数据的抽象

分布共享数据服务

TODO: 图: 分布共享数据服务

- 简化上层应用的开发
- 提供共享数据的抽象
- 屏蔽底层数据的分布性

分布共享数据服务理论与技术研究

数据一致性问题

TODO: 图：从分布到共享

读操作语义问题：在分布数据环境下，读操作允许返回什么值？

数据一致性问题

TODO: 图：从分布到共享

读操作语义问题：在分布数据环境下，读操作允许返回什么值？

数据一致性问题

数据一致性问题

数据一致与否是相对于应用逻辑而言的:

数据一致性问题

数据一致与否是相对于应用逻辑而言的:

- 数据一致性模型多样

数据一致性问题

数据一致与否是相对于应用逻辑而言的:

- 数据一致性模型多样
- 数据一致性模型有强弱之分

数据一致性问题

数据一致与否是相对于应用逻辑而言的:

- 数据一致性模型多样
- 数据一致性模型有强弱之分
- 应用规约数据一致性需求

数据一致性问题

数据一致与否是相对于应用逻辑而言的:

- 数据一致性模型多样
- 数据一致性模型有强弱之分
- 应用规约数据一致性需求
- 数据不一致导致应用异常 (anomalies)

数据一致性问题举例 (I)

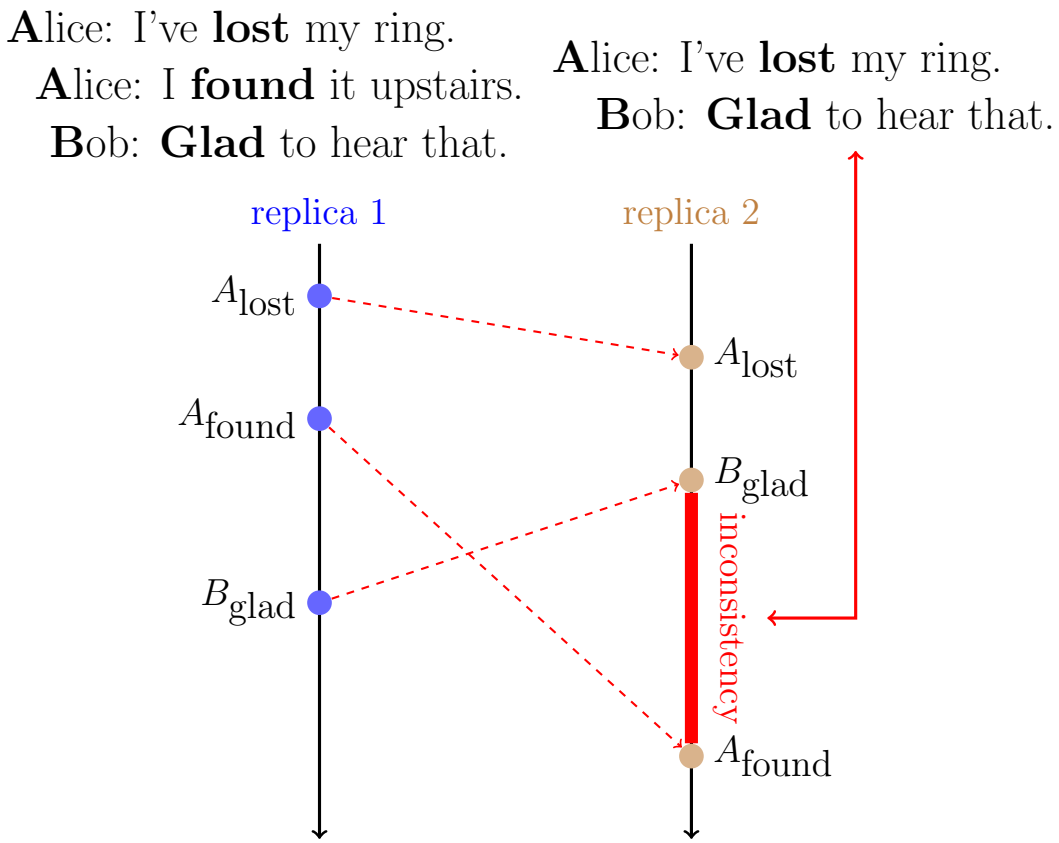


图: 社交网络中, 消息-评论乱序 [Lloyd@CACM'14].

数据一致性问题举例 (II)

图: 多设备文件共享时, 更新丢失 $(\#N = 3, \#W = 2, \#R = 1)$.

数据一致性问题举例 (II)

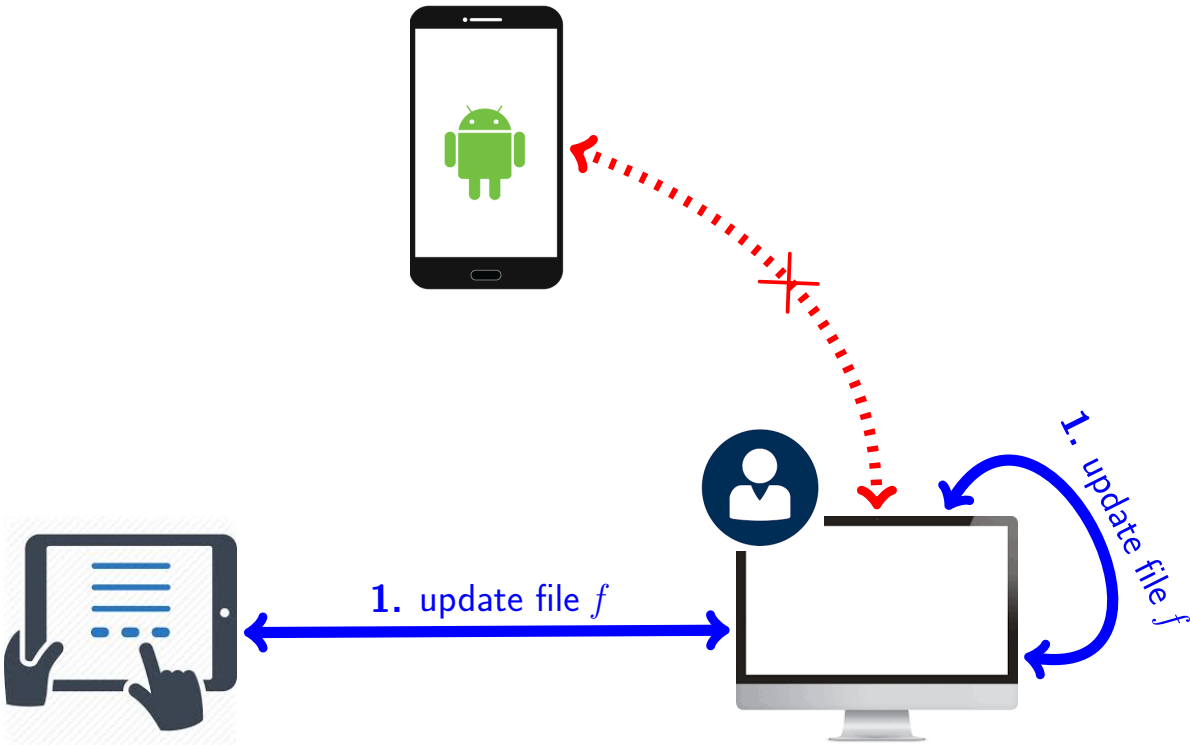


图: 多设备文件共享时, 更新丢失 $(\#N = 3, \#W = 2, \#R = 1)$.

数据一致性问题举例 (II)

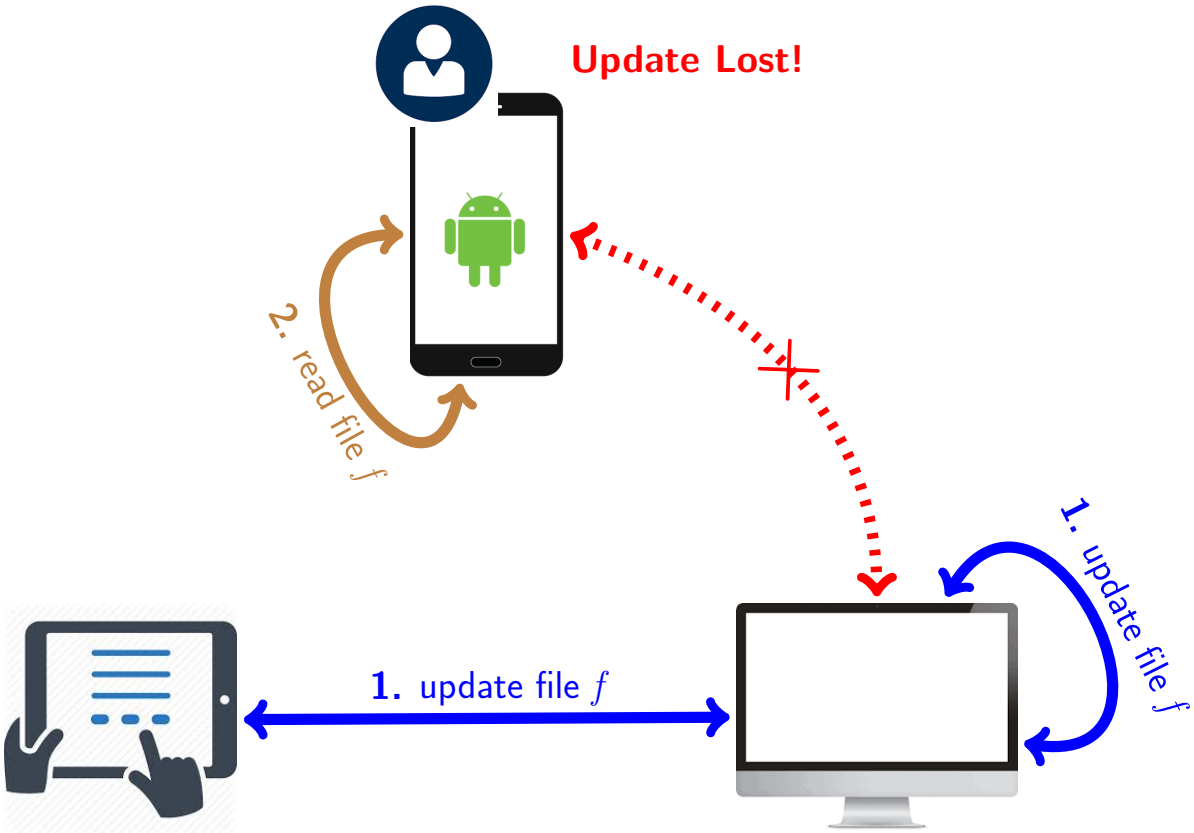
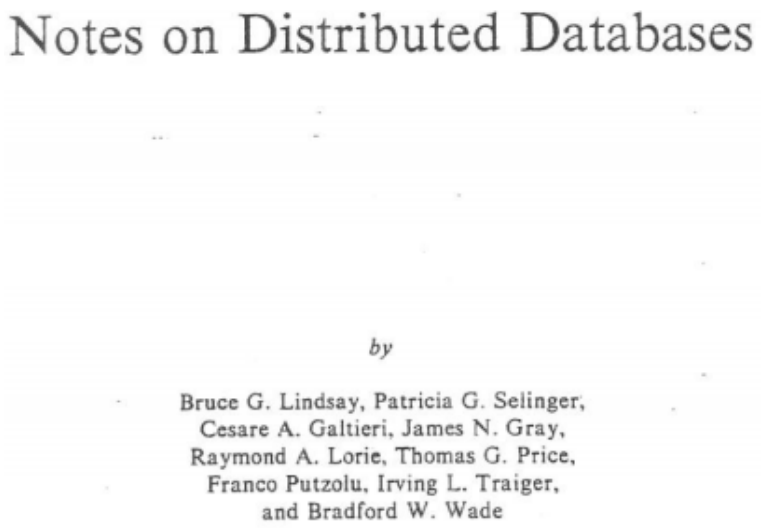


图: 多设备文件共享时, 更新丢失 $(\#N = 3, \#W = 2, \#R = 1)$.

数据一致性问题研究的历史阶段

TODO: 图：ps

数据一致性问题研究的历史阶段 (I; 1970s)



Notes on Distributed Databases

by

Bruce G. Lindsay, Patricia G. Selinger,
Cesare A. Galtieri, James N. Gray,
Raymond A. Lorie, Thomas G. Price,
Franco Putzolu, Irving L. Traiger,
and Bradford W. Wade

Tradeoff availability for consistency (and simplicity).

数据一致性问题研究的历史阶段 (II; Middle 1990s)

Eventual (weak) consistency [Terry@PDIS'94], [Terry@SOSP'95] driven by mobile computing.

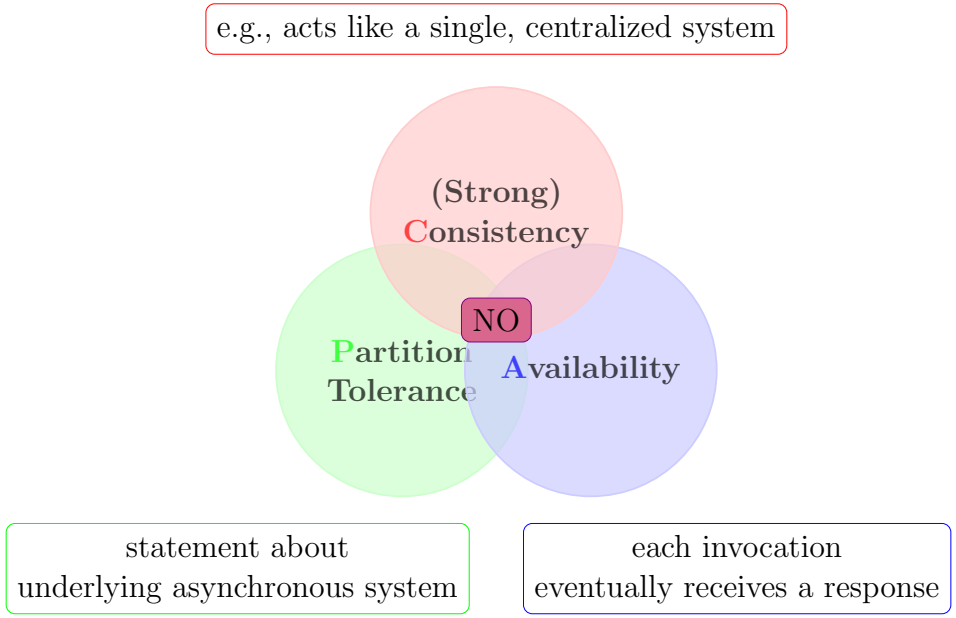
Managing Update Conflicts in Bayou,
a Weakly Connected Replicated Storage System

Douglas B. Terry, Marvin M. Theimer, Karin Petersen, Alan J. Demers,
Mike J. Spreitzer and Carl H. Hauser

Computer Science Laboratory
Xerox Palo Alto Research Center
Palo Alto, California 94304 U.S.A.

Tradeoff consistency for availability.

数据一致性问题研究的历史阶段 (III; Early 2000s)



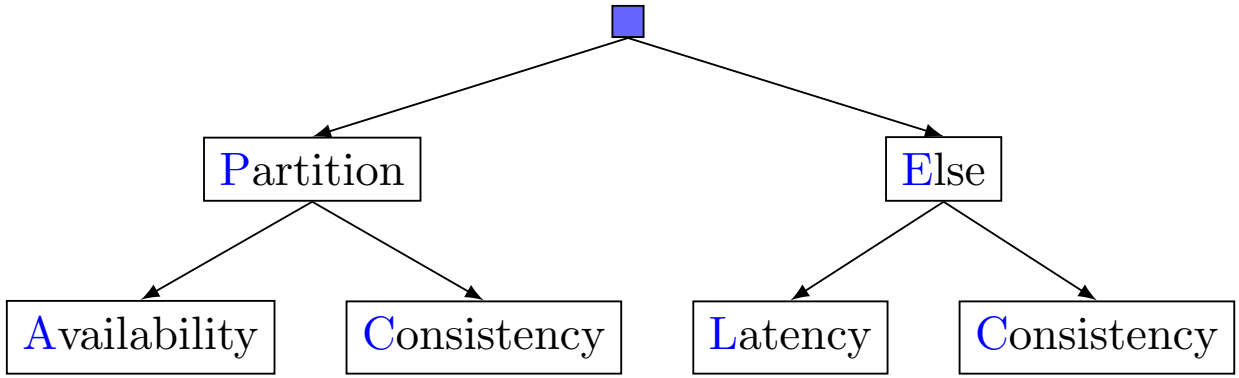
TODO: 图: 重绘

CAP 定理: 分布式系统无法同时满足强一致性, 可用性和分区容错性.

[Brewer@PODC'00] [Gilbert@SIGACT'02]

数据一致性问题研究的历史阶段 (IV; Late 2000s)

数据一致性问题研究的历史阶段 (V; 2010s)



分布共享数据服务理论与技术研究

1. 研究背景
2. 研究问题
3. 研究方法
   - 理论模型: 分布共享数据
   - 技术途径: 三维框架

分布共享数据服务理论与技术研究

1. 研究背景
2. 研究问题
3. 研究方法
   - 理论模型: 分布共享数据
   - 技术途径: 三维框架

分布共享数据 (I)

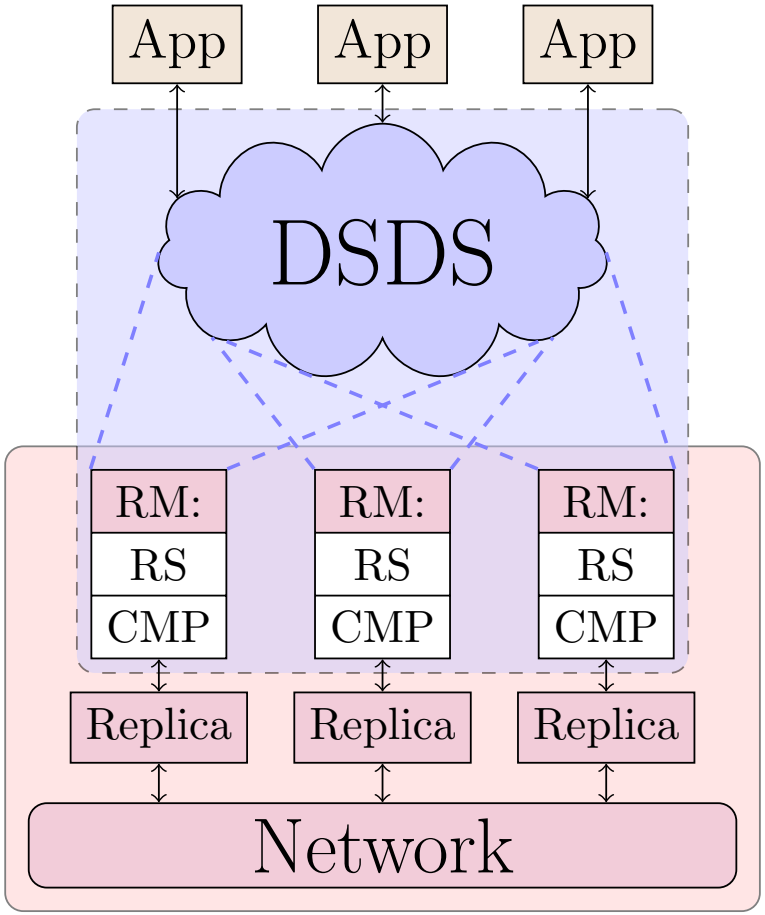
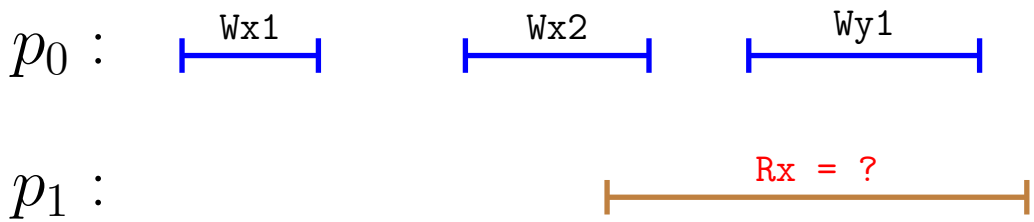


图: 分布数据共享服务.

分布共享数据 (II)

x, y : 共享变量　　p_0, p_1 : 客户进程

多进程并发提交 (读/写) 操作:



问题: 读操作允许返回什么值?

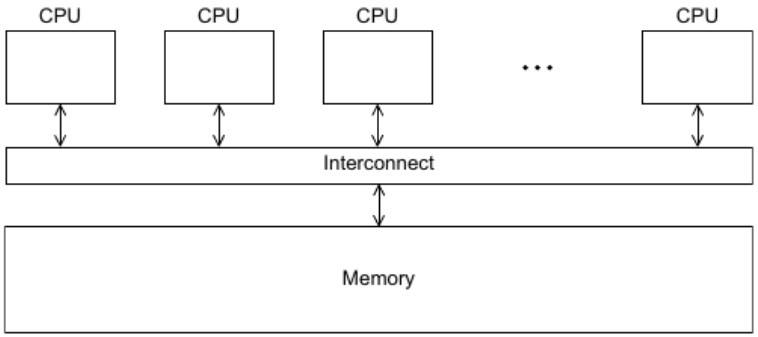
不同一致性 $\underset{\text{定义}}{\overset{\text{规定}}{\rightleftharpoons}}$ **不同合法返回值**

分布共享数据 (III)

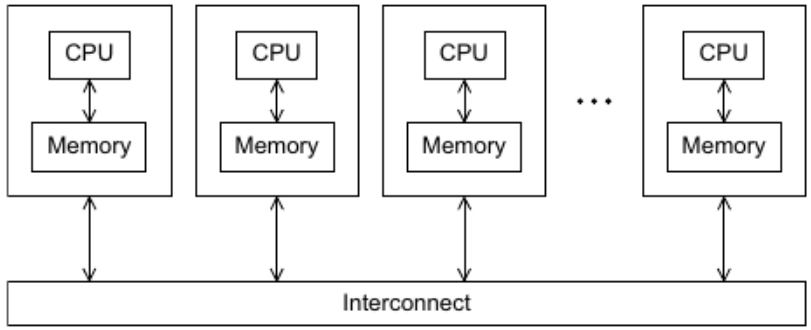
基本定位: 传统概念应用于新型平台

分布共享数据服务: 分布共享内存模型 + 分布数据系统

分布共享内存 (I)



(a) 共享内存系统.



(b) 分布内存系统.

图: 多处理器系统体系结构 **TODO:** 重绘.

分布共享内存 (II)

分布内存系统实例: 按系统耦合度分类

分布共享内存 (III)

分布共享内存: 在分布内存之上提供共享内存的假象

TODO: 图: 分布共享内存 (from Kai Li)

分布共享数据 (IV)

问题空间: 传统问题, 新平台, 新挑战 (**TODO: 总结**)
目的: 并行编程
实现手段: 硬件/操作系统

分布共享数据服务理论与技术研究

1. 研究背景
2. 研究问题
3. 研究方法
   - 理论模型: 分布共享数据
   - 技术途径: 三维框架

分布共享内存中的数据一致性问题

数据一致性问题的三个层面:

1. 虚拟共享数据有什么? ▶ 数据类型
2. 上层接口语义是什么? ▶ 一致性模型
3. 底层消息传递为什么? ▶ 一致性保障

研究框架

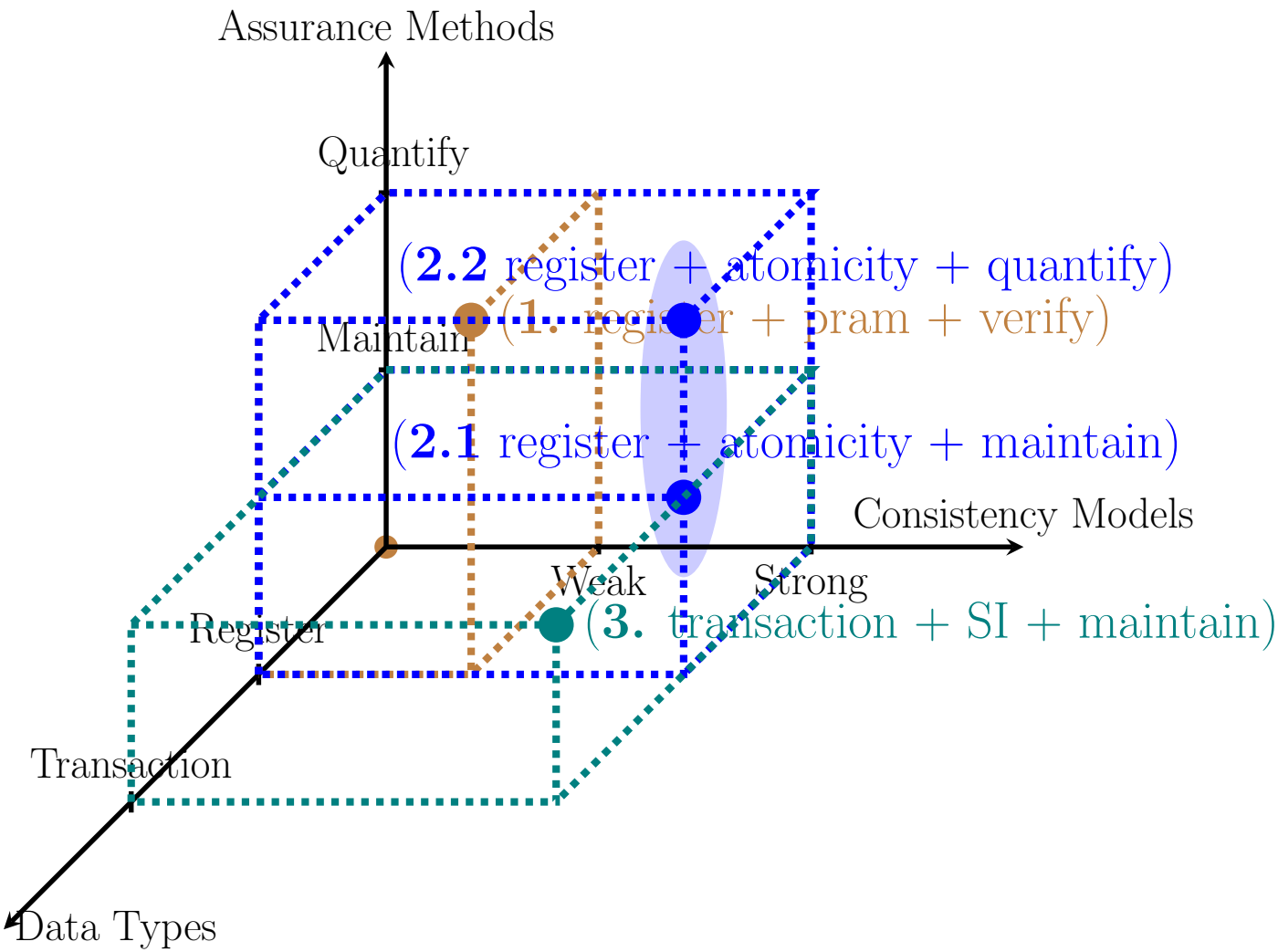


图: 数据一致性及保障技术研究框架

研究挑战

用户对一致性的需求:

1. 多样化, 可定制 [Terry@CACM'13]

2. 精细化, 可度量 [Bailis@VLDB'12]

研究挑战

用户对一致性的需求:

1. 多样化, 可定制 [Terry@CACM'13]
   - 多样化:
     - 一致性族: causality; read-your-writes (RYW)
     - 参数调节: 提供 "有限度" 的不一致 [Yu@TOCS'02]
   - 可定制: 混合使用, 运行时可变
2. 精细化, 可度量 [Bailis@VLDB'12]

研究挑战

用户对一致性的需求:

1. 多样化, 可定制 [Terry@CACM'13]
   - 多样化:
     - 一致性族: causality; read-your-writes (RYW)
     - 参数调节: 提供 "有限度" 的不一致 [Yu@TOCS'02]
   - 可定制: 混合使用, 运行时可变
2. 精细化, 可度量 [Bailis@VLDB'12]
   - 精细化: "在大多数情况下, 访问到一致数据"
   - 可度量: 量化系统执行, 后验系统对一致性的满足程度