

# 分布数据一致性技术研究

魏恒峰

导师: 吕建 黄宇

南京大学软件所

November 3, 2016

# 分布数据一致性技术研究

① 研究背景

② 研究问题

③ 相关工作

# 分布数据一致性技术研究

① 研究背景

② 研究问题

③ 相关工作

# 分布式应用



新浪微博社交网站<sup>1</sup>:

- ▶ 日均用户近一亿
- ▶ 日均消息近一亿条

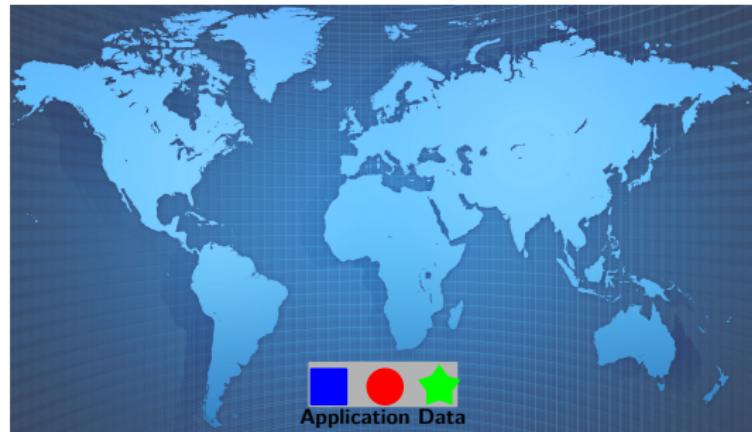
底层数据服务系统特性需求 ( $H^3 L$ ):

- ▶ 低延迟, 高可用性 (4 个 9<sup>2</sup>)
- ▶ 高容错性, 高可扩展性

<sup>1</sup> 2015 第三季度; 数据来自 China Internet Watch.

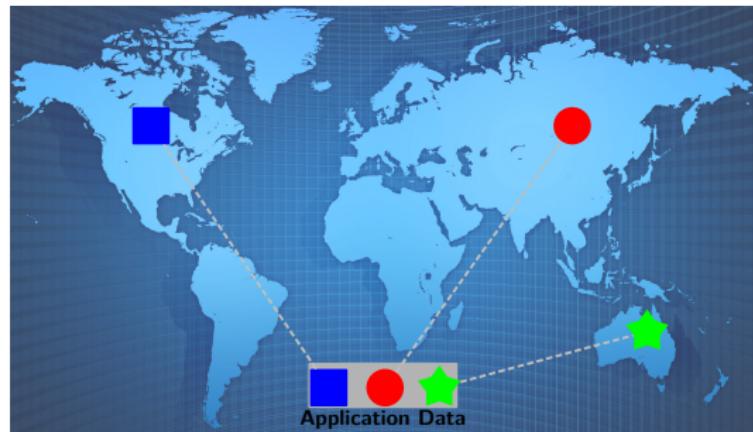
<sup>2</sup> 数据来自 InfoQ.

# 分布数据



应用数据:

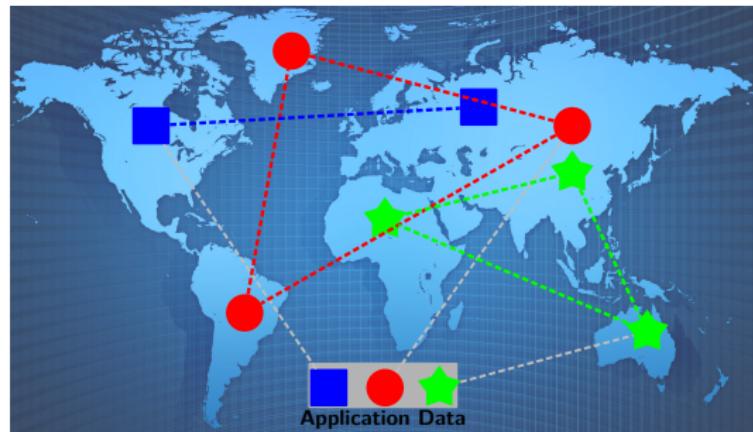
# 分布数据



应用数据:

1. 分区 (*partition*): 水平扩展

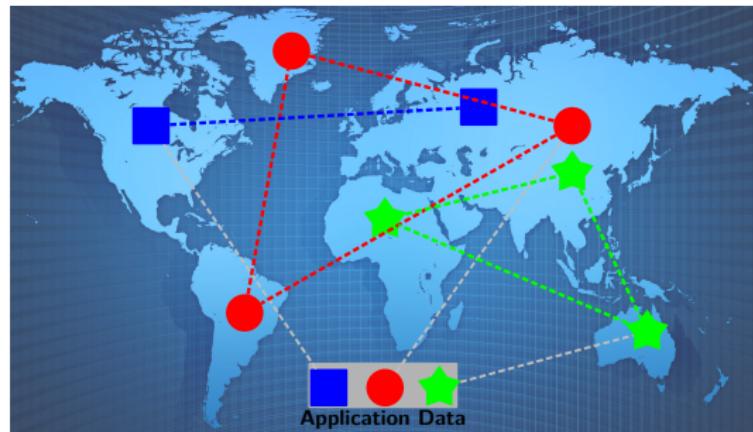
# 分布数据



应用数据:

1. 分区 (partition): 水平扩展
2. 副本 (replication): 就近访问, 容灾备份

# 分布数据



**分布数据** (distributed data):

1. 分区 (partition): 水平扩展
2. 副本 (replication): 就近访问, 容灾备份

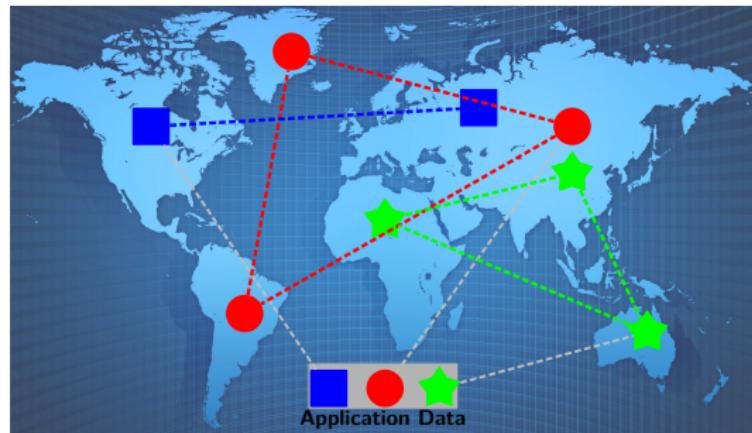
# 分布数据一致性技术研究

1 研究背景

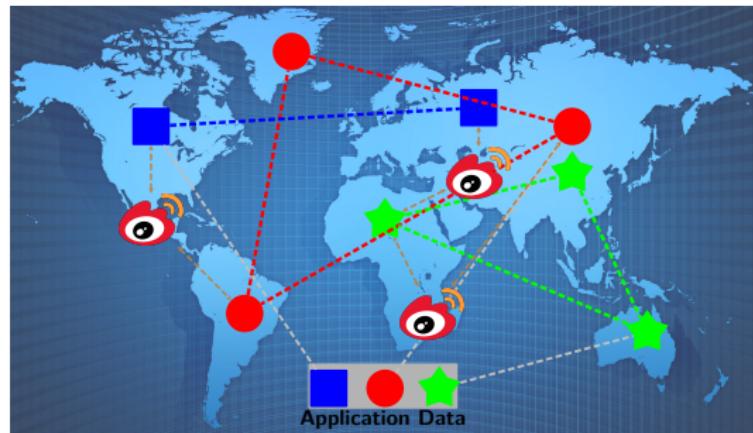
2 研究问题

3 相关工作

# 分布数据一致性问题

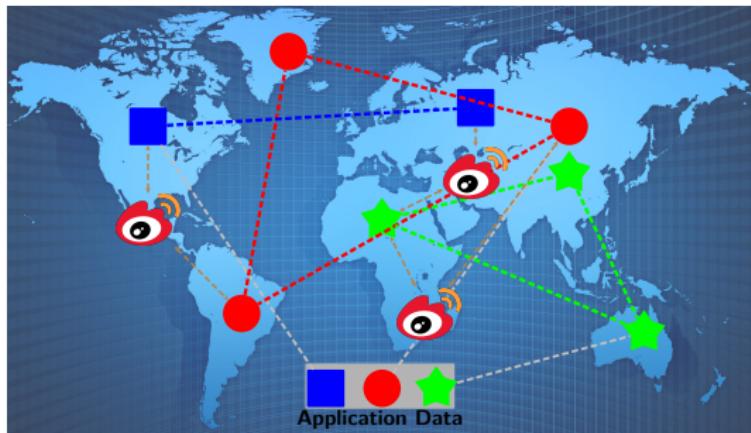


# 分布数据一致性问题



分布数据 (distributed data)  $\Leftarrow$  共享 (集中式) 数据 (shared data):

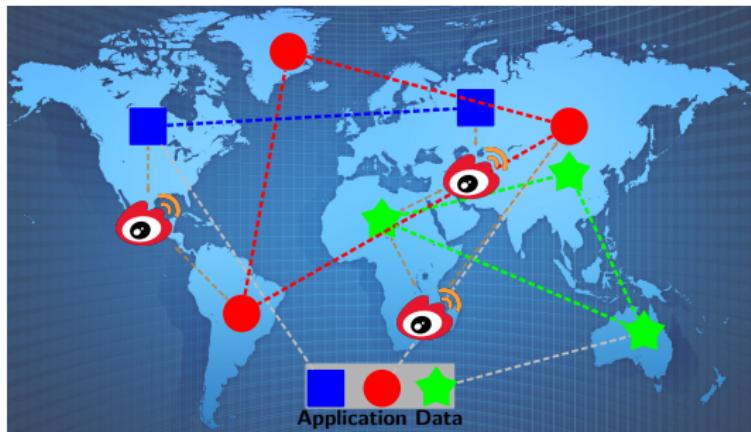
# 分布数据一致性问题



分布数据 (distributed data)  $\Leftarrow$  共享 (集中式) 数据 (shared data):

上层应用: 分布数据的语义是什么? 如何“方便”地使用分布数据?

# 分布数据一致性问题



分布数据 (distributed data)  $\Leftarrow$  共享 (集中式) 数据 (shared data):

上层应用: 分布数据的语义是什么? 如何“方便”地使用分布数据?

数据副本: 以何种顺序应用更新? 对应用提供什么保证?

# 分布数据一致性问题

Replica A      Replica B

图: 社交网络中, 消息-评论乱序 [Lloyd@CACM'14].

# 分布数据一致性问题

Alice: I've **lost** my ring.

Alice: I **found** it upstairs.

Bob: **Glad** to hear that.

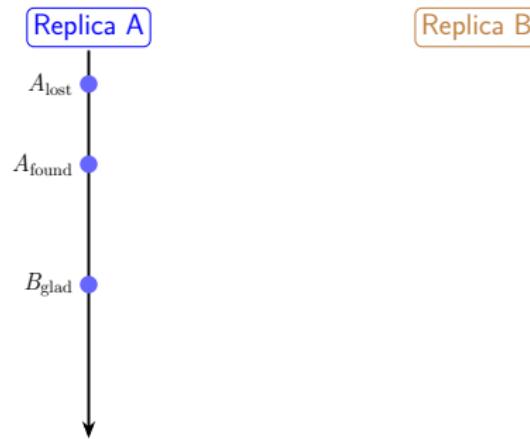


图: 社交网络中, 消息-评论乱序 [Lloyd@CACM'14].

# 分布数据一致性问题

Alice: I've **lost** my ring.

Alice: I **found** it upstairs.

Bob: **Glad** to hear that.

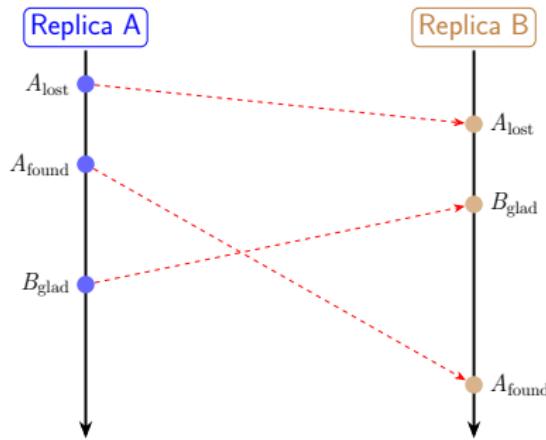


图: 社交网络中, 消息-评论乱序 [Lloyd@CACM'14].

# 分布数据一致性问题

Alice: I've **lost** my ring.

Alice: I **found** it upstairs.

Bob: **Glad** to hear that.

Alice: I've **lost** my ring.

Bob: **Glad** to hear that.

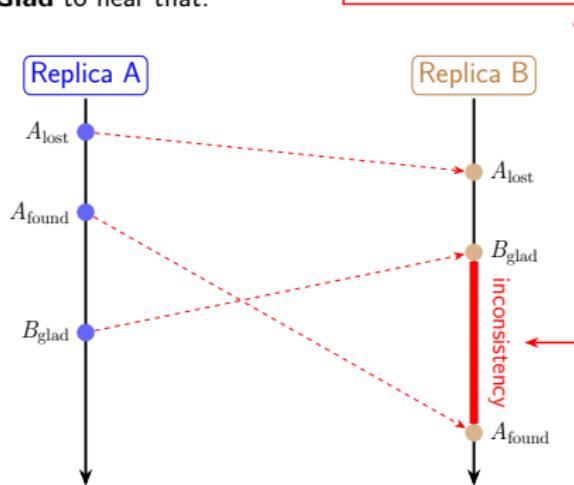


图: 社交网络中, 消息-评论乱序 [Lloyd@CACM'14].

# 分布数据一致性问题

理想情况:

- ▶ one-size-fits-all 一致性模型
- ▶ 始终观察到最新副本

没有分布数据一致性问题

# 分布数据一致性问题

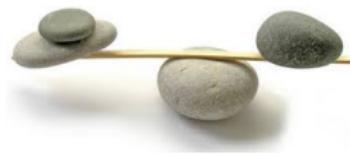
理想情况:

- ▶ one-size-fits-all 一致性模型
- ▶ 始终观察到最新副本

没有分布数据一致性问题

实际情况 (tradeoffs) [Guerraoui@TCDE'16]:

$H^3L$   
Partition-tolerance  
Convergence  
Churn  
...  
**Consistency**



# 分布数据一致性问题

理想情况:

- ▶ one-size-fits-all 一致性模型
- ▶ 始终观察到最新副本

~~没有分布数据一致性问题~~

实际情况 (tradeoffs) [Guerraoui@TCDE'16]:

$H^3L$   
Partition-tolerance  
Convergence  
Churn  
...

Consistency



以数据一致性为核心的权衡使得该问题具有挑战性

# 分布数据一致性问题

## 论文研究问题：

考虑到上述权衡，  
面向大规模分布式系统的  
分布数据一致性理论应具有什么特性？

# 分布数据一致性问题

## 论文研究问题:

考虑到上述权衡，  
面向大规模分布式系统的  
分布数据一致性理论应具有什么特性？

考察分布数据一致性问题研究的历史：

- ▶ 核心权衡与解决方案

# 分布数据一致性问题

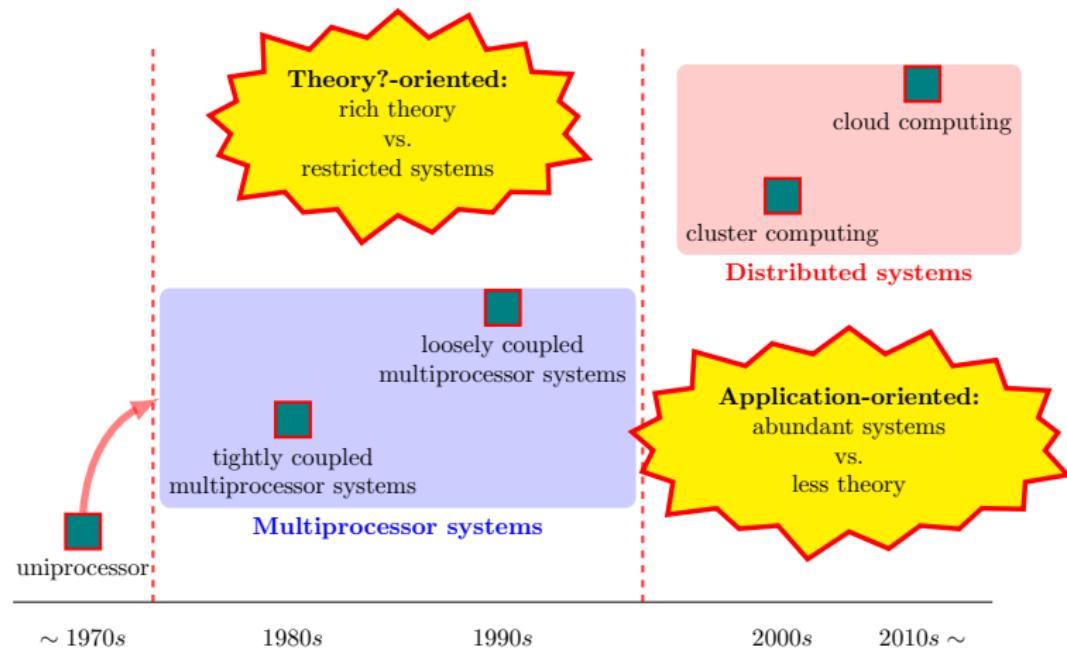
## 论文研究问题:

考虑到上述权衡，  
面向大规模分布式系统的  
分布数据一致性理论应具有什么特性？

考察分布数据一致性问题研究的历史：

- ▶ 核心权衡与解决方案
- ▶ 理论与系统

# 数据一致性问题研究的历史阶段



# 数据一致性问题研究的历史阶段 (多处理器系统)

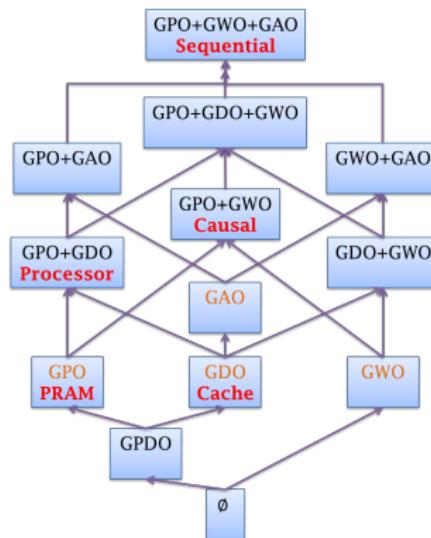
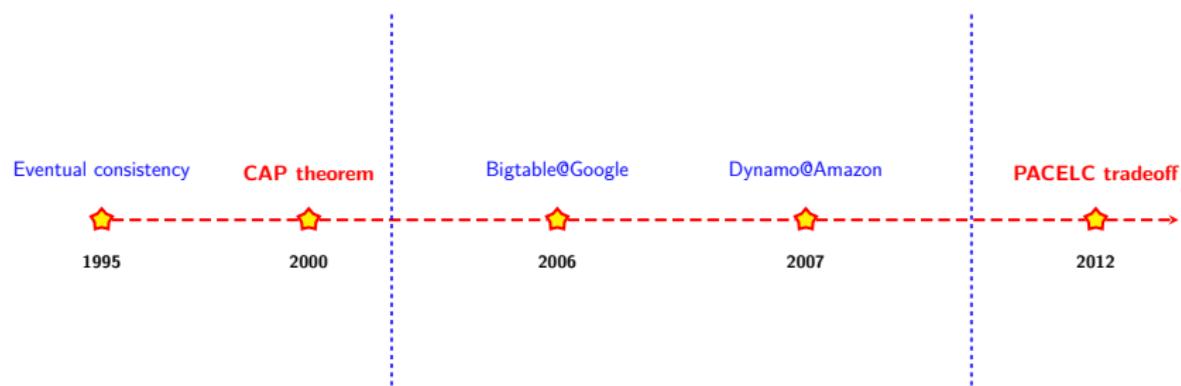


图: 一致性模型 (依据 [Steinke@JACM'04] 中 Figure 13 重绘).

核心权衡: 一致性模型的计算能力 vs. 系统性能

# 数据一致性问题研究的历史阶段 (分布式系统)



# 数据一致性问题研究的历史阶段 (分布式系统)



Eventual consistency



1995

CAP theorem



2000

Bigtable@Google



2006

Dynamo@Amazon



2007

PACELC tradeoff



2012

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

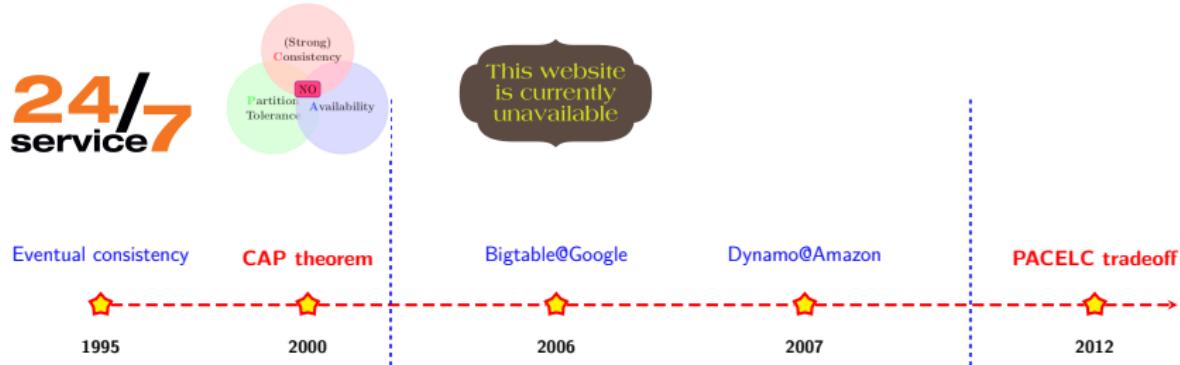
Douglas B. Terry, Marvin M. Thurber, Kara Barnes, Alan J. Demers,  
Mike J. Spreitzer and Carl H. Hauser

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

Towards Robust  
Distributed Systems

Dr. Eric A. Brewer  
Professor, UC Berkeley  
Co-Founder & Chief Scientist, Inktomi

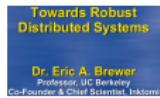
# 数据一致性问题研究的历史阶段 (分布式系统)



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

Douglas B. Terry, Marvin M. Thurber, Kara Barneser, 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.

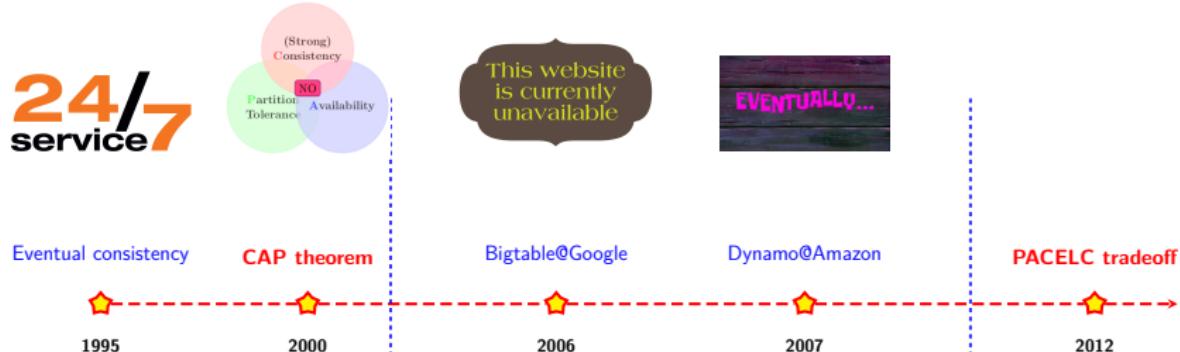


Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach  
Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber  
[jeff.dean,sanjay.ghemawat,wilson.hsieh,deborah.wallach,mike.burrows,tushar.chandra,andy.fikes,robert.gruber]@google.com

Google, Inc.

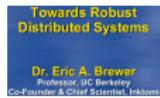
# 数据一致性问题研究的历史阶段 (分布式系统)



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

Douglas B. Terry, Marvin M. Thurber, Kara Barneser, 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.



Bigtable: A Distributed Storage System for Structured Data

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Heid, D. A. Hoffman, Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Ketukapat, Arvind Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber  
{fay,jeff,dean,sanjay,wilson,cheid,dahoff,giuseppe,deniz,madan,jampani,guna,arvind,alex,pilchin,swami,pete,mburrows,tushar,a...}@google.com

Google, Inc.

Dynamo: Amazon's Highly Available Key-value Store

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Ketukapat,<sup>1</sup> Arvind Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber  
Amazon.com

# 数据一致性问题研究的历史阶段 (分布式系统)



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

Douglas B. Terry, Marvin M. Thurber, Kara Barneser, 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.

Towards Robust  
Distributed Systems

Dr. Eric A. Brewer  
Professor, UC Berkeley  
Co-Founder & Chief Scientist, Inktomi

Bigtable: A Distributed Storage System for Structured Data

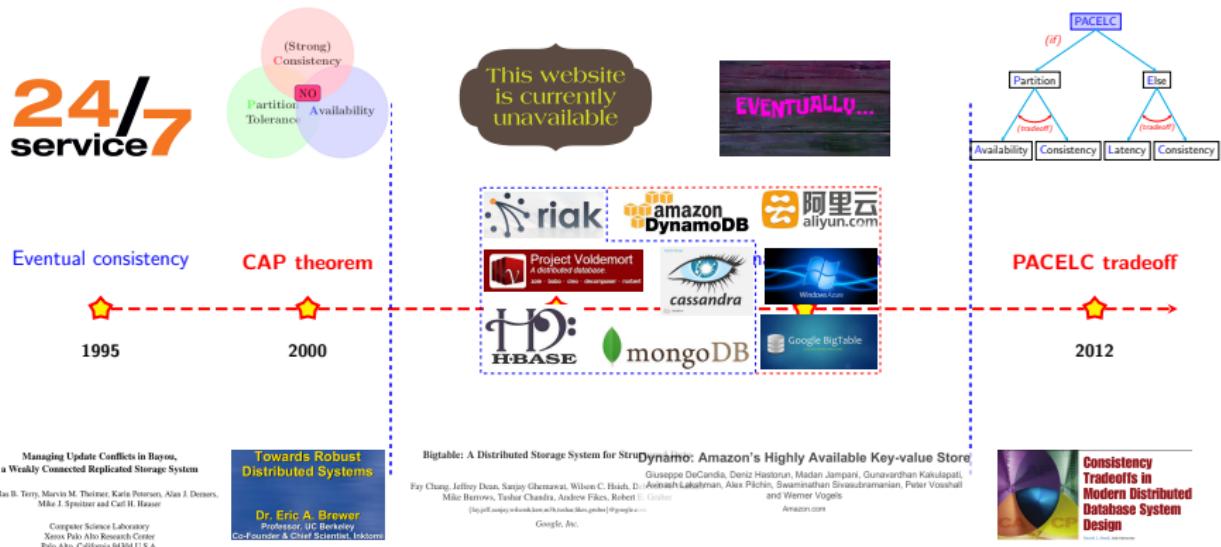
Dynamo: Amazon's Highly Available Key-value Store

Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Heid, D. Arvind Lakshman, Alex Pitkänen, Swaminathan Sivasubramanian, Peter Vosshall, Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, Giuseppe DeCandia, Dorin Hastorun, Madan Jampani, Gunavardhan Katulapati, and Werner Vogels

{fay,jeff,dean,sanjay,wilson,c.heid,d.arvind,alex,swami,peter,mike,tushar,andy,robert}@google.com

Google, Inc.

# 数据一致性问题研究的历史阶段 (分布式系统)



# 数据一致性问题研究的历史阶段 (结论)

新平台的两个特点：

需要什么样的数据一致性理论？

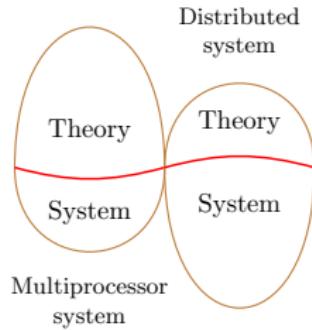
# 数据一致性问题研究的历史阶段 (结论)

新平台的两个特点:

- (1) 云计算新平台凸显应用价值观

需要什么样的数据一致性理论?

- (1) 与应用价值观相匹配



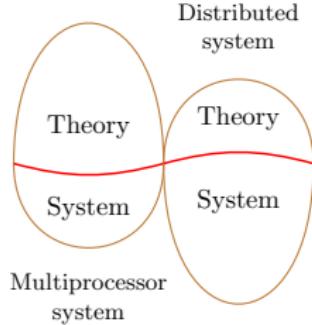
# 数据一致性问题研究的历史阶段 (结论)

新平台的两个特点:

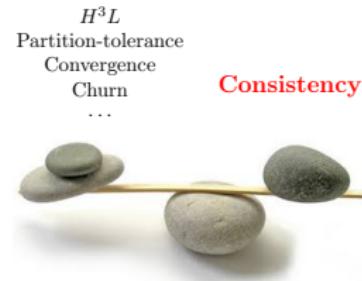
- (1) 云计算新平台凸显应用价值观
- (2) 应用价值观积极拥抱 tradeoffs

需要什么样的数据一致性理论?

(1) 与应用价值观相匹配



(2) 体现更丰富的 tradeoffs



# 数据一致性问题研究的发展趋势及我们的工作 (I)

## 购物车一致性需求

- ▶ 优先 `read-my-writes`
- ▶ 可接受 `any consistency`  
只要延迟低于 300ms

## 出租车实时位置查询一致性需求:

- ▶ 所有读请求都要满足 `2-atomicity`
- ▶ 违反 `atomicity` 的读请求低于 1%

# 数据一致性问题研究的发展趋势及我们的工作 (I)

## 购物车一致性需求

- ▶ 优先 `read-my-writes`
- ▶ 可接受 `any consistency`  
只要延迟低于 300ms

## 出租车实时位置查询一致性需求:

- ▶ 所有读请求都要满足 `2-atomicity`
- ▶ 违反 `atomicity` 的读请求低于 1%

## 应用价值观导向的数据一致性理论:

1. 多样化, 可调节
2. 精细化, 可度量

# 数据一致性问题研究的发展趋势及我们的工作 (II)

多样化: 从单一到融合 (mono- vs. multi-) [Terry@CACM'13]

- ▶ 融合强弱一致性: 不同操作, 不同一致性需求
- ▶ 融合一致与不一致: 容忍“有限度”的不一致



# 数据一致性问题研究的发展趋势及我们的工作 (II)

多样化: 从单一到融合 (mono- vs. multi-) [Terry@CACM'13]

- ▶ 融合强弱一致性: 不同操作, 不同一致性需求
- ▶ 融合一致与不一致: 容忍“有限度”的不一致

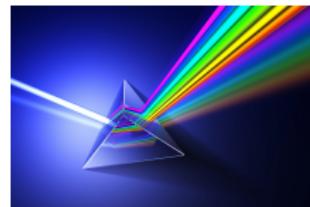


可调节: think *dynamically* [Terry@SOSP'13]

依据应用需求/系统状态调节数据一致性

# 数据一致性问题研究的发展趋势及我们的工作 (III)

精细化：从二元到连续谱 [Yu@TOCS'02]



# 数据一致性问题研究的发展趋势及我们的工作 (III)

精细化: 从二元到连续谱 [Yu@TOCS'02]



可度量: think *probabilistically* [Brewer@PODC'00]



量化系统执行, 后验系统对一致性的满足程度

# 数据一致性问题研究的发展趋势及我们的工作 (III)

论文研究问题：

如何在大规模分布式系统中  
落实“多样化, 可调节; 精细化, 可度量”  
这一体现应用价值观的数据一致性问题研究理念?

# 数据一致性问题研究的发展趋势及我们的工作 (III)

## 论文研究问题:

如何在大规模分布式系统中  
落实“多样化, 可调节; 精细化, 可度量”  
这一体现应用价值观的数据一致性问题研究理念?

## 论文主要贡献:

理念: 提出“以应用为导向的”、“多样化, 可调节; 精细化,  
可度量”的一致性问题研究理念

# 数据一致性问题研究的发展趋势及我们的工作 (III)

## 论文研究问题:

如何在大规模分布式系统中  
落实“多样化, 可调节; 精细化, 可度量”  
这一体现应用价值观的数据一致性问题研究理念?

## 论文主要贡献:

理念: 提出“以应用为导向的”、“多样化, 可调节; 精细化,  
可度量”的一致性问题研究理念

VPC:

# 数据一致性问题研究的发展趋势及我们的工作 (III)

## 论文研究问题:

如何在大规模分布式系统中  
落实“多样化, 可调节; 精细化, 可度量”  
这一体现应用价值观的数据一致性问题研究理念?

## 论文主要贡献:

理念: 提出“以应用为导向的”、“多样化, 可调节; 精细化,  
可度量”的一致性问题研究理念

VPC:

PA2AM:

# 数据一致性问题研究的发展趋势及我们的工作 (III)

## 论文研究问题:

如何在大规模分布式系统中  
落实“多样化, 可调节; 精细化, 可度量”  
这一体现应用价值观的数据一致性问题研究理念?

## 论文主要贡献:

理念: 提出“以应用为导向的”、“多样化, 可调节; 精细化,  
可度量”的一致性问题研究理念

VPC:

PA2AM:

RVSI:

# 分布数据一致性技术研究

1 研究背景

2 研究问题

3 相关工作

# “多样化, 可调节” 的研究理念 (一)

“多样化, 可调节” 的读写寄存器一致性模型 (多处理器系统):

# “多样化, 可调节” 的研究理念 (一)

“多样化, 可调节” 的读写寄存器一致性模型 (多处理器系统):

典型: Hybrid consistency [Attiya@SIAM J. Comput.'98]

思想: 将操作分为强弱两类

# “多样化, 可调节” 的研究理念 (一)

“多样化, 可调节”的读写寄存器一致性模型 (多处理器系统):

典型: Hybrid consistency [Attiya@SIAM J. Comput.'98]

思想: 将操作分为强弱两类

其它: “带同步的” 一致性模型 [Dubois@IEEE Computer'88] [Steinke@JACM'04]

特点: 强调正确性 (properly synchronized)

总评: 相关工作丰富; 理论扎实

# “多样化, 可调节” 的研究理念 (二)

“多样化, 可调节” 的读写寄存器一致性模型 (分布式系统):

思想: 借鉴并发展 Hybrid consistency 的思想

典型:

- ▶ Causal/immediate/forced consistency [Ladin@TOCS'92]
- ▶ RedBlue consistency [Li@OSDI'12]
- ▶ Apache Cassandra<sup>3</sup> [Facebook@SIGOPS OSR'10]
- ▶ Pileus [Terry@SOSP'13]

<sup>3</sup><http://cassandra.apache.org/>

# “多样化, 可调节” 的研究理念 (二)

“多样化, 可调节” 的读写寄存器一致性模型 (分布式系统):

思想: 借鉴并发展 Hybrid consistency 的思想

典型:

- ▶ Causal/immediate/forced consistency [Ladin@TOCS'92]
- ▶ RedBlue consistency [Li@OSDI'12]
- ▶ Apache Cassandra<sup>3</sup> [Facebook@SIGOPS OSR'10]
- ▶ Pileus [Terry@SOSP'13]

特点: 更细粒度的多一致性模型共存、更能容忍数据不一致

总评: 已成趋势; 缺少基础理论工作

<sup>3</sup> <http://cassandra.apache.org/>

# “多样化, 可调节” 的研究理念 (三)

“多样化, 可调节” 的**事务一致性模型** ([分布式系统](#)):

# “精细化, 可度量”的研究理念

	读写寄存器	事务
--	-------	----

表: 相关工作二: “精细化, 可度量”的研究理念