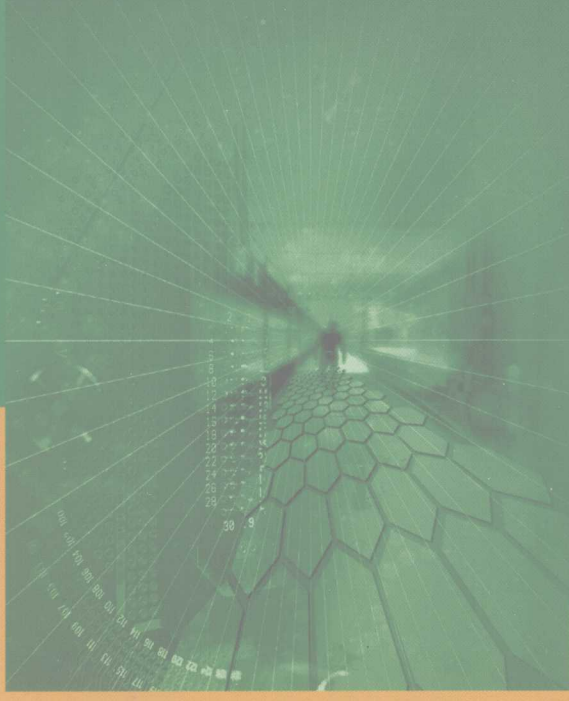
分布式实时数据库技术

肖迎元 ◎ 著



科学出版社
www.sciencep.com

分布式实时数据库技术

肖迎元 著

科学出版社

北京

内 容 简 介

本书以“系统模型与体系结构→分布式实时事务处理→故障恢复”为主线，论述了分布式实时数据库技术的主要概念、理论、技术与方法，是作者多年来在分布式实时数据库理论与技术方面研究工作的总结。

全书共10章，包括绪论、分布式实时数据库系统模型、分布式实时数据库数据交换策略、分布式实时事务调度策略、实时并发控制协议、分布式实时事务提交、分布式实时数据库系统故障恢复需求与正确性准则、基于日志的实时故障恢复、分布式实时数据库全局一致性备份等内容，涵盖了分布式实时数据库技术的各个主要方面。

本书适合作为高等院校计算机及相关专业研究生教材或参考书，亦可作为从事数据库研究或应用开发的研究人员、工程技术人员的参考书。

图书在版编目(CIP)数据

分布式实时数据库技术 / 肖迎元著. —北京：科学出版社，2009
ISBN 978-7-03-024655-4

Ⅰ. 分… Ⅱ. 肖… Ⅲ. 分布式数据库 Ⅳ. TP311.133.1

中国版本图书馆 CIP 数据核字(2009)第 088170 号

责任编辑：余 江 于宏丽 / 责任校对：陈玉凤
责任印制：张克忠 / 封面设计：耕者设计工作室

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

丽源印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2009年6月第 一 版 开本：B5(720×1000)
2009年6月第一次印刷 印张：9 1/2
印数：1—2 500 字数：175 000

定价：28.00元

(如有印装质量问题，我社负责调换〈长虹〉)

前　言

纵观数据库技术发展过程，计算环境和数据库技术基本保持着一种同步发展的态势，它们互相影响、互相促进。计算环境先后经历了集中式、分布式、网络等多种计算模式。相应地，数据库系统的发展也经历了集中式数据库系统、分布式数据库系统、B/A/S 多层结构的数据库系统和移动数据库系统等多个阶段。

分布式数据库系统作为数据库技术的一个重要发展阶段，多年来得到了广泛的发展，推动其发展的主要因素是不断增长的应用需求，如全球及我国范围内的航空/铁路/旅游订票系统、银行通存通兑系统、水陆空联运系统、连锁配送管理系统等都需要分布式数据库系统提供支持。近年来，计算机网络、分布式计算技术的进一步发展，使得实时存取分布在网络不同节点上的信息成为可能，于是分布式实时数据库技术便应运而生。分布式实时数据库是分布式数据库和实时数据库相结合的产物，是事务和数据都可以具有定时特性或显式定时限制的分布式数据库。分布式实时数据库系统在工业过程控制、电网调度、军事作战指挥系统、股票交易等时间关键型应用中具有广泛的应用前景。

分布式实时数据库系统集成了分布式数据库和实时数据库的功能，但并非二者在概念、技术、机制上的简单组合，而有一系列问题需要被研究和解决，如分布式实时事务模型、实时事务调度、实时提交机制、实时并发控制协议、实时故障恢复、安全性等。

本书研究适合分布式实时数据库特性和需求的新策略、新技术、新机制和新方法，着重对分布式实时事务模型、实时事务调度策略、实时并发控制协议、实时故障恢复技术、全局一致性备份技术进行深入的研究。

本书作者多年来一直从事分布式实时数据库系统的研究与应用开发，在攻读博士学位期间作为主要技术负责人参与了国产实时数据库管理系统的研发，本书中的许多内容都是作者在攻读博士学位期间研究成果的总结和扩展，在此要感谢导师刘云生教授在作者攻读博士学位期间给予的悉心指导。在本书的撰写过程中参阅了该领域大量的研究成果，也得到了天津理工大学副校长张桦教授的热情鼓励和帮助，在此表示衷心感谢。

本书得到了天津市自然科学基金(08JCYBJC12400)、中小企业创新基金(08ZXCXGX15000)、天津市高等学校科技发展基金(2006BA16)、智能计算及软件新技术天津市重点实验室的资助，在此表示感谢，也感谢科学出版社给予的大力支持与帮助，特别感谢余江编辑为本书出版付出的辛勤劳动。

本书可作为计算机及相关专业研究人员的参考资料，也可作为从事数据库研究或应用开发的研究人员、工程技术人员的参考书。由于作者水平有限，书中难免有不足之处，欢迎读者批评指正。

肖迎元

2009 年 1 月

目　录

第1章 绪 论

自20世纪70年代E. F. Codd提出数据库的关系模型后，关系数据库理论在随后多年的发展中不断成熟。然而在关系数据库管理系统(Relation Database Management System，RDBMS)的实现和产品开发中，却遇到了一系列的技术问题：数据库规模越来越大；数据库的结构越来越复杂；在越来越多的用户共享数据库的情况下，如何保障数据库的完整性、安全性、并发性以及故障恢复的能力。以上问题成为数据库产品是否能够进入实用并最终为用户所接受的关键因素。Jim Gray为解决这些重大技术问题，提出了事务概念及相关处理技术，为关系数据库管理系统的成熟以及其顺利进入市场发挥了关键作用。概括地说，解决上述问题的主要技术手段和方法是，把对数据库的操作划分为“事务”的基本单位，一个事物要么全做，要么全不做(即all-or-nothing原则)；用户在对数据库发出操作请求时，需要对有关的不同数据“加锁”，防止不同用户的操作之间互相干扰；在事务运行过程中，采用“日志”记录事务的运行状态，以便发生故障时进行恢复；对数据库的任何更新都采用“两阶段提交”策略。

进入20世纪90年代后，一些新的应用领域，如工业过程控制、空中交通管制、计算机集成制造、智能交通、电子商务、股票交易系统等，不断向数据库技术提出新的要求和挑战，从而出现了一批适用于特定应用领域的现代数据库，如实时数据库、内存数据库、主动数据库、时态数据库等。同时，计算机领域本身的发展也给数据库技术带来了深刻的变化，计算机网络技术、分布式计算技术的迅速发展，使得实时地存取分布在网络不同节点上的信息成为可能，于是分布式实时数据库技术便应运而生。分布式实时数据库涉及信息处理技术、分布式计算技术、实时处理技术等多个学科领域，已受到这些领域研究者的关注，特别是在数据库领域，分布式实时数据库已成为新的研究热点。

1.1 分布式实时数据库系统概述

随着计算机网络技术、分布式计算技术的迅速发展，一些新的应用(时间关键型应用)，如电话交换、电力或数据网管理、空中交通管制、雷达跟踪、指挥控制系统、证券交易等，对传统商务或事务型分布式数据库技术提出了新的要求。这些应用一方面要求维护大量的共享数据和控制知识；另一方面其应用活动有很强的时间性，要求在一定的时刻或一定的时期内自外部环境采集数据，按彼此之间的联系

存取已获得的数据和处理采集的数据，再及时作出响应。同时，它们所处理的数据往往是“短暂”的，即只在一定的时间范围内有效，过时则对当前的决策或推导无意义。显然，面向商务或事务型的传统分布式数据库技术已经不能满足上述时间关键型应用的需求，这就需要分布式实时数据库系统的支持。

本节简要介绍分布式数据库系统的体系结构、实时数据库系统、分布式实时数据库系统及分布式实时事务的特性。

1.1.1 分布式数据库系统的体系结构

计算机网络通信技术的迅速发展，以及地理上分散的公司、团体和组织对数据库更为广泛的应用需求，在集中式数据库系统的基础上产生和发展了分布式数据库系统(Distributed Database System，DDBS)。分布式数据库系统包括分布式数据库和分布式数据库管理系统。分布式数据库是一组数据集，逻辑上它们属于同一系统，但物理上它们分散在用计算机网络连接的、具有场地自治能力的多个站点上，并统一由一个分布式数据库管理系统管理。分布式数据库管理系统(Distributed Database Management System，DDBMS)是建立、管理和维护分布式数据库的一组软件，是分布式数据库系统的核心和基础。分布式数据库管理系统负责实现分布式数据库的建立、查询、更新、复制、维护等功能，包括提供分布透明性、查询优化、协调全局事务的执行、协调各站点共同完成全局应用、保证数据库的全局一致性、执行并发控制、实现更新同步和全局恢复等。图 1-1 给出了一个分布式数据库系

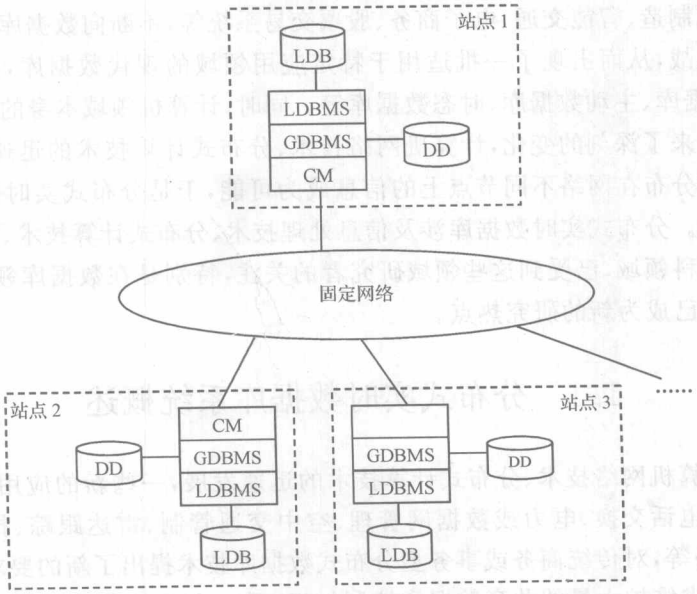


图 1-1 分布式数据库系统体系结构

统的体系结构。图中各组成部件的说明如下：

(1) 本地数据库管理系统(Local Database Management System，LDBMS)：主要功能是建立和管理本地数据库，提供场地(站点)自治能力，执行本地应用及全局查询的子查询，提供本地恢复功能。

(2) 全局数据库管理系统(Global Database Management System，GDBMS)：主要功能是提供分布透明性，协调全局事务的执行，协调各本地数据库管理系统以完成全局应用，保证数据库的全局一致性，执行并发控制，实现更新同步，协调各本地数据库管理系统以实现全局恢复功能。

(3) 通信管理器(Communication Manager，CM)：主要功能是在分布式数据库系统各场地之间传送消息和数据，提供可靠通信服务功能，提供支持分布式事务管理的通信机制。

(4) 数据字典(Data Directory，DD)：包括全局数据字典(Global Data Directory，GDD)和局部数据字典(Local Data Directory，LDD)。全局数据字典用来存放全局概念模式(Global Conceptual Schema)、分片模式(Fragmentation Schema)、分布模式(Allocation Schema)的定义以及各模式之间映像的定义，以及有关用户存取权限的定义与数据完整性约束条件的定义。局部数据字典的功能与集中式数据库的数据字典类似。

(5) 本地数据库(Local Database，LDB)：存放由本地数据库管理系统管理的本地数据。在分布式数据库系统中，全局数据库(分布式数据库)为各站点上本地数据库的逻辑集合。

1.1.2 实时数据库系统

实时数据库系统(Real-time Database System，RTDBS)是传统的实时系统和数据库系统相结合的产物，适用于对时间有特殊需求的数据管理应用领域。国外(特别是美国、英国、瑞典)对于实时数据库系统进行了多年的研究，目前仍在积极进行着。

在实时数据库系统中，事务和数据都可以具有定时限制，系统的正确性不仅依赖于事务执行的逻辑结果，还依赖于逻辑结果产生的时间。实时数据库系统中事务的定时限制典型地表现为事务的截止期(Deadline)，它定义为事务在给系统带来最大价值的前提下可最晚提交的时间。事务若不能在规定的截止期内完成，将丧失其价值，甚至还可能带来灾难性的后果。数据的定时限制表现为时态数据对象的有效期。除了持久数据对象(其值在数据对象的整个生命周期内始终有效)，实时数据库系统还需要处理“短暂”有效的时态数据对象，时态数据对象建模外部不断变化的客观环境，是外部客观环境在计算机内的逻辑表示。时态数据对象的值通过各种传感器(如温度传感器、压力传感器等)获得，并随着有效时间的截止而变得无效，即时态数据对象的值仅在规定的时间(有效期)内有效。

实时数据库系统中事务可以从不同的侧面进行分类：

1）按使用数据的方式分类

像传统数据库一样可以有如下分类：

（1）只写事务——收集关于现实世界的信息并写入数据库。

（2）只读事务——读取数据库中的数据值并设置执行控制部件的参数。

（3）更新事务——基于现有数据值导出新的数据值，故它可能既读又写。

显然，这种分类便于设计或改造并发控制方案。

2）按关键性分类

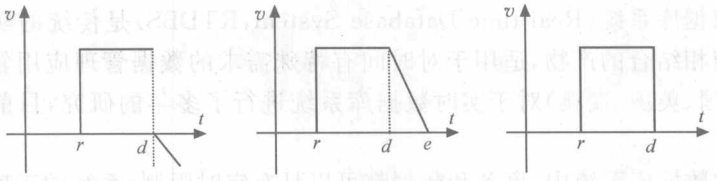
也就是按事务时限（截止期）的性质，即事务超截止期对系统带来的影响分类，而这种时限的性质可以很好地用价值函数来建模。超截止期的事务称为超时事务，此时事务将产生不可预测的结果。按照事务超截止期对系统带来的影响可将事务分为三类：

（1）硬实时事务——超截止期会导致恶果（价值函数取大的负值）。它对应于安全危急性活动。

（2）软实时事务——超截止期仍有一定的价值，但不断下降，直到某一时刻（称为最终有效时间），其价值降到零，此后保持为零。

（3）固实时事务——一旦到达截止时间，其价值立即降为零，此后固定为零。

图 1-2 中（a）～（c）分别给出描绘硬、软、固实时事务的示意图。其中 v、t 两坐标轴分别为价值函数与时间，d 为截止期，e 为最终有效时间，r 为放行（Release）时间。当 $t<r$ 时，$v(t)=0$ 意思是在事务未准备好以前起动是无价值的。对于固实时事务，它是软实时事务在 e 与 d 重合情况的特例。



(a) 硬实时事务　(b) 软实时事务　(c) 固实时事务

图 1-2　实时数据库中基于关键性的事务示意图

3）按功能分类

实时数据库系统以两种方式直接与现实世界交互作用，一是关于现实世界状态或事件的信息被记录到数据库中；二是事务可以触发各种影响现实世界的活动。这就给予我们一种如下事务分类：

（1）数据接收事务——记录现实世界的状态或发生的事件到数据库中。它是简单地只写事务且应是被立即执行（不能等待和阻塞）的硬实时事务。

(2) 数据处理事务——类似传统数据库的事务,用来执行用户或应用程序发起的数据库操作请求。

(3) 控制事务——能触发外部世界中相关活动的事务。控制事务通过向受控子系统发送控制消息,来触发改变外部世界的活动。

4) 按到达的时间分类

按事务到达系统的时间分类,有:

(1) 周期事务——以一定的周期循环地到达和被执行。例如,对应现实世界的数据采样(亦即上述数据接收)事务是周期事务。

(2) 非周期事务——由外部事件(如现实世界状态变化)或内部事件(如特定的时钟行为、数据库状态的变化)动态驱动。如上述的数据处理事务。

(3) 零星事务——偶尔地一次性执行。它们是非预先安排的,故一般是非实时的,如通常的即席查询。

1.1.3 分布式实时数据库系统

一般认为,分布式实时数据库系统(Distributed Real-Time Database System, DRTDBS)是分布式数据库系统和实时数据库系统相结合的产物,是事务和数据都可以具有定时特性或显式的定时限制的分布式数据库系统。分布式实时数据库系统并不是简单地把实时数据库分散地实现,而是具有自己的性质和特征的系统。一个分布式实时数据库系统应该具有如下的特点:

1) 物理分布式性

分布式实时数据库系统中的数据不是存储在一个站点上,而是分散存储在由计算机网络连接起来的多个站点上。

2) 逻辑整体性

在分布式实时数据库系统中,物理上分布的数据在逻辑上是一个整体,它们被分布式实时数据库系统的所有用户共享,并由一个分布式实时数据库管理系统统一管理。

3) 站点自治性

站点自治性也称场地自治性,各站点上的数据由本地数据库管理系统管理,对于本地应用(局部应用)具有自治处理能力。

4) 时限性

不同于传统的分布式数据库系统,在分布式实时数据库系统中事务和数据都可能有定时限制。事务的定时限制典型地表现为事务有执行时间限制,事务若在规定的截止期后完成,结果将变得毫无价值(对固实时事务而言),甚至这可能带来灾难性的后果(对硬实时事务而言)。数据的定时限制通常表现为数据的有效期,事务所存取的数据对象的值必须是在其有效期范围内的,过期的数据是没有意义

的。显然，适合传统分布式数据库系统的事务调度策略、并发控制机制、提交协议、恢复模式等已经不能满足或不能很好地满足分布式实时数据库系统的时限需求。

集中式数据库的许多概念和技术，如数据独立性、数据共享和减少冗余度、并发控制、完整性、安全性和恢复等，在分布式实时数据库系统中都有了不同的、更加丰富的内容。

5）数据独立性

数据独立性是数据库系统追求的主要目标之一。在集中式数据库系统中，数据的独立性包括两个方面：数据的逻辑独立性和数据的物理独立性。在分布式实时数据库系统中，数据独立性具有更多的内容。除了数据的逻辑独立性与物理独立性外，还有数据的分布独立性亦称分布透明性(Distribution Transparency)。分布透明性指用户不必关心数据的逻辑分片，不必关心数据物理位置的分布细节，也不必关心重复副本一致性问题，同时也无须关心局部场地上数据库支持何种数据模型。

6）集中与自治相结合的控制结构

数据库是多个用户共享的资源。在集中式数据库系统中，为了保证数据库的安全性和完整性，对共享数据库实施集中控制。在分布式实时数据库系统中，数据的共享有两个层次：

(1) 局部共享。即在本地数据库中存储本场地上各用户的共享数据，这些数据是本场地用户常用的。

(2) 全局共享。即在分布式实时数据库系统的各个场地也存储供其他场地的用户共享的数据，支持系统的全局应用。

因此，相应的控制机构也具有两个层次：集中和自治。分布式实时数据库系统通常采用集中和自治相结合的控制机构。各分布式实时数据库系统可以独立地管理本地数据库，具有自治功能。同时，系统又设有集中控制机制，协调各本地数据库管理系统的工作，执行全局应用。当然，不同的系统，其集中和自治的程度不尽相同。

7）适度的数据重复

在集中式数据库系统中，尽量减少数据冗余度是系统目标之一。原因是冗余数据不仅浪费存储空间，而且容易造成各数据副本之间的不一致；为了确保数据的一致性，系统要付出一定的维护代价。而在分布式实时数据库系统中却希望有适度的数据重复，在不同的场地存储同一数据的多个副本。其目的如下：

(1) 提高系统的可靠性、可用性。当某一场地出现故障时，系统可以对另一场地上相同副本进行操作，不会因为一处故障而影响整个系统的可用性。

(2) 提高系统性能。系统可以并行操作或选择离用户最近的数据副本进行操作，以减少通信开销，改善整个系统的性能。

但是，这种数据重复也会和数据冗余一样带来与操作集中式数据库系统一样的问题。虽然重复数据增加存储空间的问题随着存储介质容量的增大、价格的下降而得到解决，但多个副本间数据一致性维护问题是分布式实时数据库系统必须着力解决的问题。一般来讲，数据重复方便了检索，提高了系统的查询速度、可用性和可靠性，但不利于更新，增加了系统维护的代价。

8) 全局的一致性、可串行性和可恢复性

分布式实时数据库系统中除了各本地数据库应满足集中式实时数据库的一致性、并发事务的可串行性和本地数据库的可恢复性以外，还应保证数据库的全局一致性、全局并发事务的可串行性和系统的全局可恢复性。

1.1.4 分布式实时事务的特性

如同集中式数据库系统，在分布式实时数据库系统中同样需要确保分布式实时事务的 ACID 特性，即原子性(Atomicity)、一致性(Consistency)、隔离性(Isolation)和持久性(Durability)。分布式实时事务的原子性要求构成分布式实时事务的每一个子事务要么都提交，要么都不提交；一致性要求分布式实时事务的执行结果必须是使全局数据库从一个一致性状态变到另一个一致性状态，即当全局数据库只包含成功事务提交的结果时，就说全局数据库处于一致性状态；隔离性是一个分布式实时事务的执行不能被其他事务干扰，即一个分布式实时事务内部的操作及使用的数据对其他并发事务是隔离的；持久性是指一个分布式实时事务一旦提交，它的每一个子事务对数据库中数据的改变就应该是永久性的，即使故障也不应该对其执行结果有任何影响。分布式实时事务的一致性和隔离性主要由分布式实时数据库系统的并发控制机制来确保。分布式实时事务的原子性和持久性则由分布式实时数据库系统的提交协议和恢复机制来保证。

分布式实时事务与传统的集中式数据库系统中事务相比，存在如下两个方面的不同：① 传统的集中式数据库系统中事务在单个数据库服务器上执行，而分布式实时事务的执行被分解为多个分布在不同站点(数据库服务器)的子事务的协同执行；② 传统的集中式数据库系统中的事务无执行时间限制，而分布式实时事务具有定时限制。上述区别决定了分布式实时事务除了具有事务的 ACID 特性外，还具有如下不同于集中式数据库系统中事务的特性：

(1) 执行特性。由于分布式实时事务执行时被分解成多个子事务，因此每一个分布式实时事务必须创建一协调者进程，以协调各子事务的执行，确保全局数据库的一致性。

(2) 操作特性。在分布式实时事务中，除了对数据的存取操作外，还包括大量的通信原语，以负责协调者进程和各子事务间的数据传送，以及子事务之间的数据传送。

(3) 控制报文。在分布式实时数据库系统中,除了数据报文外,还增加了控制报文。因为除了对数据进行存取操作外,还要对各子事务的操作进行协调。

(4) 结构特性。分布式实时事务通常具有复杂的结构,事务内部和事务之间可能存在着各种结构,如分层、嵌套、分裂、合并、彼此通信合作等。

(5) 定时特性。分布式实时事务的执行具有时间限制,这种时间限制典型地表现为事务的截止期。

1.2 支持分布式实时事务的内存数据库

20 世纪 80 年代中期以来,内存数据库(Main Memory Database,MMDB)引起了越来越多的数据库研究者们的兴趣。两大因素决定着内存数据库的研究与发展。一是现代应用,如工业过程控制、空中交通管制、计算机集成制造、智能交通等,要求数据库有强的功能和高的性能,特别是像保证"硬实时"这样的性能要求,而传统的常驻磁盘数据库系统已无能为力。二是随着 VLSI 技术的高速发展,半导体存储器的成本不断下降,而密度却指数级地增长,使得存储量很大而廉价的内存普遍使用,因而,将整个数据库或它的大部分放置在内存中成为可能。

内存数据库要求数据库"主版本"或"工作版本"常驻内存,磁盘版本(外存版本)仅作为内存"工作版本"的后援。由于所有的数据库读、写操作都是针对内存"工作版本",因而内存数据库能消除事务执行过程中的磁盘输入输出,使得系统具有较好的可预测性和实时性。针对上述内存数据库的本质特征,我们给出了内存数据库的如下定义:

定义 1.1 内存数据库:设有数据库系统 DBS,DB 为 DBS 中的数据库,DBM(t)为在时刻 t DB 在内存的数据集,DBM(t)$\subseteq$DB。TS 为 DBS 中所有可能事务的集合,AT(t)为在时刻 t 处于活动状态的事务集,AT(t)$\subseteq$TS。$D_t(T)$为事务 T 在时刻 t 所操作的数据集,$D_t(T) \subseteq$ DB。若在任一时刻 t,均有

$$\forall T \in \mathrm{AT}(t)(D_t(T) \subseteq \mathrm{DBM}(t))$$

成立,则称 DBS 为一内存数据库系统,DB 为一内存数据库,简记为 MMDB。

显然,内存数据库系统需要一定的内存容量,但并不要求整个数据库都常驻内存。

根据上述内存数据库的定义,活跃事务只与内存数据库的"工作版本"打交道,当事务提交时,事务提交的结果不一定立即反映到内存数据库的外存版本,但应在非易失性存储器上记录日志。当系统故障发生,可依内存数据库外存版本和日志对内存"工作版本"进行恢复。

在分布式实时数据库系统中,为了更好地满足分布式实时事务的定时限制,要求系统能较准确地估计事务的执行时间。但对于基于磁盘的数据库系统而言,由