

## C . Mohan 访谈录

本专访主要介绍了关于 R\*系统和消息队列的技术，印度计算机科学发展状况，ARIES 的由来，以及 IBM 院士 C.Mohan 的研究生涯，等等

玛丽安 温丝莱

**玛丽安：**首先要感谢那些帮助我们系列专栏设计采访问题的人们。我们向你们承诺匿名，但是还是要在你们这里向你们致谢，没有你们的帮助，这个专栏就没办法成功。再次感谢你们在过去的几年里提供了那么多睿智的问题。

**玛丽安：**感谢 ACM SIGMOD Record 设置了数据库届备受尊敬的成员的系列访谈专栏。我是玛丽安 温丝莱，今天我们在 2003 年 SIGMOD 和 PODS 会议的举办地，圣地亚哥。今天采访的是 C.Mohan，它是 IBM Almaden 研究中心的 DBCache 项目的技术负责人。Mohan 因他在事务提交、日志管理和恢复方面的工作而闻名于世，他对现有数据库产品有着巨大的影响。对于 IBM 和其他运营商来说，如何把这些技术应用在现有数据库产品中是一个关键问题。Mohan 是 1992 年 TODS 上关于 ARIES 的文章的作者，ARIES 成为了每个数据库研究生资格考试的必考科目，这篇文章是唯一一篇超过 50 页限制被接收的 TODS 文章。Mohan 是 IEEE、ACM 和 IBM 的院士，并且获得了 SIGMOD 创新奖和 VLDB 10 年最有影响力论文奖。他在奥斯汀的得州大学获得博士学位。接下来欢迎 Mohan。

**C . Mohan：**谢谢玛丽安做了如此详细的介绍。

**玛丽安：**欢迎，Mohan，你的毕业论文是讲述理论的——在数据库加锁协议中管理死锁。提出这个理论方向之后，究竟是什么导致你毕业后走入企业？什么导致你从数据库理论转向到从事 ARIES 的实际工作中？

**C . Mohan：**实际上，虽然我的论文是理论方面的，但是不是因为我对理论感兴趣。当时在我就读的大学里还没有数据库领域的大规模系统项目，所以我做了一个能够尽快毕业的论文并且获得我的博士学位。但是，此时我对实际的事情比较感兴趣。我写了一篇关于 SSD 1 的一个评论报道，SSD 1 是一个分布式数据库系统，详细地分析了其设计，我把它发送到一些人手里，我渴望真正的脱离学校并且更多的参与到开发实际东西中。这就是为什么我没有申请任何一个高校工作的一个理由。我热衷于参加一个研究实验室，来获得更多操作实际系统的能力，进而改变它们，因此让产品体现出一些新技术。

**玛丽安：**当你到了 IBM，你参加到 R\*团队中，自那时起一直在 IBM。事后有没有想过，在 R\*所做的工作与你所希望的是否有所不同？

**C . Mohan：**在项目的中途，我加入到项目中。在那段时间，即使在 IBM 研究实验室，我也没有真正与产品开发人员一起工作过。R\*起初关注的是同构的分布式数据库。我们假设分布式数据库网络上的不同节点是所有系统 R 的节点，然后我们采用两阶段提交协议、复制、分布式查询编译和分布式优化。但是问题是既然我们只关注分布式数据库中的同构方面，当产品商业化的时候，证明了即使是 IBM 这样的企业，还是会推出具有某些不同功能的不

同关系数据库产品。解决 R\*研究项目中的异构问题会是一个更加富有成效的实践练习，也可以成为其他项目的基础。

**玛丽安：**异构情况比同构情况是一个更难解决的问题。

**C . Mohan：**的确是。

**玛丽安：**现在把注意力放在解决异构问题上是不是为时过早？

**C . Mohan：**我不这么认为，因为最近一段时间，许多公司，比如美国计算机公司（CCA）已经有人开始研究这个课题。我们在同构情况上的关注主要是因为开始 R\*项目的人们很熟悉系统 R，并且主机上的 DB2 仍然在开发过程中，直到 1984 年才开始发售。

**玛丽安：**当我比较 ARIES 和数据库领域内的其他重要工作时，我发现 ARIES 具有质的不同，它是一大堆需要被正确处理的细节的聚合体，而不是一个单一的较大创新。这点不同从侧面解释了为什么 ARIES TODS 文章相比于其他有影响力的文章较长。你能说一说这些质的不同，并且更泛化一点，我们在数据库世界是否能够达到一种状态——一些新创的想法不再起到重要作用？

**C . Mohan：**我认为 ARIES 论文比较长的原因是它试着综合了所有相关工作，并且对那些可能成为 ARIES 算法一部分的特征作出了解释。我翻阅了过去的工作，发现大部分问题都没有写到发表的论文中（即使它们可能提到了）。作为一个很细致的人，我感觉研究团队没有能够很好的把握整体——以一种高效可靠的方式包含并发控制、恢复和存储管理。我带着这些问题了解一些事情——不但阅读过去的论文还回顾了过去的系统 R 的代码，并且了解了一些没有被记录下来的特征。我也和系统 R 的人员聊过了，他们中的一些人仍然在 IBM 的研究实验室，我还回顾了一下层次数据库并阅读了代码，发现了 IBM 产品的某些特征。做完这些工作，我感觉研究团队和 IBM 产品开发团队将会受益于所有这些被收集的信息。所以我选择描述关于恢复、并发和存储管理的一些算法和可选项变量。

**玛丽安：**你在 ARIES 技术转化的经历中学到了什么？在今天的 IBM，你和产品组是什么样的关系？

**C . Mohan：**事实上，论文中的 ARIES 算法来源于与 Don Haderle 的交流，那个时候他是主机 DB2 的主要架构师。正是通过与他的交流，使我意识到 DB2 做某些事情的方式不同于系统 R 人员做那些事情，后来系统 R 作为 SQL/DS 被商业化。系统 R 使用基于影子页的恢复并且遗留下一个开放问题——怎么样做记录加锁和写前日志。当我试着理解这些事情为什么不同，我找到了一点感觉——真实问题是什么和解决方法的特点是什么，基于 Don Haderle 的关于 DB2 的真实生活经历，所以 ARIES 的产生与产品开发人员和研究人员的紧密合作是分不开的。

ARIES 工作导致了数据库技术学会的建立，作为一个框架模式，在这种框架模式下 IBM 的研究者和产品人员在一起工作。技术转化变得很容易，因为我们尽早的和产品开发人员一起工作，尤其是在我们研究项目的系统设计和实现阶段。从那时起，我开始密切关注产品开发。我努力试着做那些从一个研究者角度来说技术挑战的工作，论文顺便也就写出来了，并且对于真实问题来说它是适用的，因此把它并入到产品中。在 IBM 工作时，我也试着平

衡这两方面。我继续致力于面向产品的东西，不但在数据库领域，并且也有其他领域，比如说 WebSphere 和 Lotus Domino/Notes。给定的 ARIES 工作的基本类型主要用于那些管理永久保存数据的系统中，ARIES 算法的变体不止被包含到 IBM 关系数据库产品中也其他公司的产品所采纳，比如微软的 SQL Server。它们也被并入到信息系统中，比如 MQSeries 和 Lotus Domino 中使用基于日志的恢复。所以 ARIES 已经深入到很多地方中，我希望继续做该方面的 IBM 产品和研究。

**玛丽安：**分布式提交——是什么？为什么用户不喜欢它？

**C . Mohan：**这个工作是我加入 R\* 计划之后的第一个工作。给我的任务就是两阶段协调算法在 R\* 系统中的设计与实现。我查找原始的经典的兩阶段协议并且致力于他的变体，即所谓的假定放弃和假定提交。两阶段提交协议保证了分布式环境下的事务原子性。当一个事务对多个可回收的存储区域进行更新的时候（数据库节点或者可回收文件），如果事务提交，那么事务所做的更新会被永久记录下来。如果事务回滚，可能是因为用户的命令回滚，或者是因为系统崩溃，那么事务所做的所有更新就会被撤销。

当把这个协议实现到现实系统中，就会遇到问题，在某些环境下网络上不同节点上的数据的用户不愿意放弃他们的自治权——在他们的系统内数据是否该被提交还是被回滚，而让其他系统决定，这样可能需要一段时间延迟。在这段不固定周期的时间内，其他访问被修改而不被提交的数据的事务被拒绝。

**玛丽安：**那么解决方法是什么？

**C . Mohan：**我们已经在致力于找到解决方案方面花了太长的时间，虽然在商业化方面，它还不能进行的很好。该解决方案采用了高级事务模型的概念：让节点独立的提交数据的改变。如果你需要回滚数据的变化，那么需要使用日志指出什么被改变了并且应用合适的撤销操作。对于撤销操作，定义了补偿事务的概念并且用它们来逻辑的撤销单节点事务提交的数据改变。有很多文献是关于这个论题的，但是在实际实现的过程中，即使在不完整的原型中，也没有一个完全的实现。

**玛丽安：**那么在产品中呢？

**C . Mohan：**在产品中，当然也缺少。但是在工作流管理系统的背景下，最终一些想法开始拨云见日。标准化的努力，比如说 Web 服务事务（WS-TX）和 Web 服务协调（WS-C），也正试图给人民提供一些特征——用来建造用户自己的高层事务概念。随着这些规范变得标准化，公司 IBM、BEA 和微软实现了这个标准化，你可以看到越来越多的采用做事务的方式。

**玛丽安：**最后呢？

**C . Mohan：**是的，会主导未来很长时间。

**玛丽安：**队列系统：在数据库世界中，我们把它放在分布式应用的什么位置？

**C . Mohan：**在这个领域，已经有很多商业化的支持，这种支持以一种 IMS 的队列事务处理的方式存在很久了，比如说数字产品，并且很多其他公司都在发行自己的事务消息队列系

统。在上个世纪 90 年代初期，IBM 为了做到一种异步的事务处理引进了它的 MQSeries 产品，但是对于做应用内的协调和分布式行为的同步 RPC 方法来说，它是可选的。因为人们不愿意采用两阶段提交协议作为分布式计算的实现方法，基于消息的分布式事务工作方式在现实世界中很流行。而许多研究团队都忽视了这个论题，只有少数论文是关于这方面的研究。

最近我们发现商务处理集成已经变得越来越重要，随着不同公司大规模的使用 IT，并且组织间的工作越来越多的使用自动化，尤其是随着 Web 产生且公司内事务通过 Web 完成，基于通信的方式完成分布式计算变得越来越有意义。一些 DBMS 运营商已经引入一些高级技术到他们的产品中，使得基于数据库的消息队列系统的实现成为可能。

**玛丽安：**那么在消息队列系统中研究问题有哪些？

**C . Mohan：**事实证明原有的事务并发被引入到关系系统中的那些特点不够好，尤其是遇到增加的并发，这个时候就需要消息 API 来支持。例如，当一个用户试着从队列获得一个消息，即使有一些比较旧的消息处于未提交状态，这些信息必须被跳过，从而为用户提供一个最近的已提交的可用的信息。能够跳过未提交数据而为用户提供快速的请求应答的想法是很难被支持的，即使给定了我们的数据库系统中存在的当前事务隔离特性。所以这是一个研究问题。

另一个研究问题是消息可以具有广泛多种多样的形式。如果你看到过 Java 的报文通信服务，它让用户自己定义新的报文头部。不同的信息运营商可以添加自己的头文件。所以信息的格式可以都不同，各种各样。当你把这些信息映射到关系上时，你就会发现同样的问题，类似于当你试着在关系系统中对 XML 文档进行建模。此外，消息可以很大，所以记录日志对性能的某些影响也需要我们解决。

第三个研究问题来自于实际，消息进入和离开系统很快。一些消息可能永远存在，而其他消息可能不会永存，并且使用这种方式的关系系统的效率目前并不是很理想。

**玛丽安：**从代码上看，这些消息是否等价于 RPC 调用？

**C . Mohan：**是的

**玛丽安：**为什么有些需要永久保存？

**C . Mohan：**RPC 开始的时候没有事务的概念，后来有了事务 RPC 的概念，这些 RPC 同步的执行。如果你想使用一个异步的基于消息的方式获得等价于事务 RPC 的功能，那么你不得不保证消息分发时的“一次且仅一次”的语义。一旦我给你发送一个请求需要做一些操作，那么无论发生什么情况，消息都会递交给你。一旦你产生一个应答或者针对我发给你的消息做一些操作，那么你需要反馈一些响应信息给我。不管失败还是其他什么情况，我们都需要确认所有这些信息没有丢失。这就是为什么要永久保存信息的缘故。

有些情况，如果能够提高性能，信息内容可以不必永久保存，就像在研究团队中比较流行的基础能力，即按照应用分组。依赖于什么信息经过消息传播出去，你是做一些事务工作还是传播一些像传感器数据以及类似的信息，你可以、也可以不关心被处理信息是否被永久保存。

**玛丽安：**Mohan，作为 56 个 IBM 院士中的一员，你认为什么使得一个人有资格成为一个院士？成为院士之后怎样改变你所做的事情？

**C . Mohan :** 既然 IBM 有 30 万雇员，所以成为 56 个被选中的人中的一个感觉很好，因为一个 IBM 院士是一个人在公司可以获得的最高技术职位。但是同时也会感觉很有责任感，因为公司期望我们成为公司的高级顾问。公司期望我们不要只关注一个项目或一个很窄的问题，而是希望我们监督公司的很多个部分，并且希望我们在进行组织内部协调时起到领导者的作用。对于那些公司内资历比较浅的年轻人来说，公司也期望我们对他们起到指导者的作用，尤其是那些参与到较深的技术活动中的年轻人。所以，我们就不能花太多的时间在某一个问题上。我们需要能够遍历很多，因为 IBM 的业务遍布全球。即使你限制自己在某个研究领域，但是仍然有很多研究实验室遍布全球。

成为一个IBM院士就是不同的人处理不同的事情。我已经选择了，不但要关注数据库领域，还要致力于与WebSphere和IBM整个软件组相关的活动，IBM整个软件组主要负责Lotus产品和Tivoli产品等等。通过关注IBM整个软件组的架构，我可以和IBM内任何地方的任何人进行交互。我的工作还涉及到了IBM全球服务（IBM Global Services, IGS），它能做当前很多盛行的事情，像系统集成项目、顾客意见征求和面向全球多种用户的IT业务——外包业务。

**玛丽安：**现在数据库理论家应该致力于什么问题？

**C . Mohan :** 这个很难说。昨天在小组会议[SIGMOD2003]上我们也讨论了这个问题。如果你看到在产品方面当前什么比较盛行，那么就是XML，目前在这个领域已经有很多动作。在形式化的基础之上，使用过去应用的方法仍然有许多工作可以做，因为会遇到一定的新问题，他们需要被形式化。整个XML领域，在DBMS背景下尤其是查询处理上怎么样处理，以及当被存储为原始形式的时候怎么做并发控制（而不是转化为关系记录），这些都是很大的研究领域，需要借助理论家的帮助，提供一个更可靠的基础。

**玛丽安：**为什么你确定其他致力于数据库领域的研究实验室，比如说AT&T、贝尔实验室和惠普实验室，现在做的不够好？

**C . Mohan :** 我认为这个问题实际上是一个事实：至少就惠普来说，他们的确拥有一个产品，但是他们不会以严肃的方式出售他们的产品。但是对于惠普实验室来说，仍然有机会通过惠普产品把惠普成果商业化，但是就AT&T和Lucent来说，他们根本就不是真的计算机公司。更重要的是，他们不是软件产品生产公司。我一直想知道，当他们有大量的数据库研究者的时候，怎样使得他们所做的工作拨云见日，能够被普通用户使用，仅仅是因为这些公司没有---

**玛丽安：**的确是，很好的一点。好，那么微软呢？你怎么比较？

**C . Mohan :** 当然微软不同于其他。微软一开始实际上并不真正做数据库方面的工作，即使他们出售一个数据库产品，因为他们出售Sybase的重新标识版本。一旦他们决定占有那个产品则改变它的内核等等使其成为微软的产品，然后他们就会看到研究组的需要。所以微软研究人员在做技术转化过程中能够获得更大的成功。但是如果比较IBM研究和微软研究，很明显我们做的研究时间更长。

我们这里传承下来了关系模型的发明者的遗产，并且我们和数据库领域的产品人具有一个长期的合作关系。而微软，产品形成是在研究组成立之前，所以他们需要一段很艰难的时间来建立它们与产品使用组织之间的信誉。

**玛丽安**：Mohan，我们知道你是从ITT Madras（马德拉斯，印度理工学院）毕业获得学位的。在过去十年间，印度发生了很大变化。假设你今天从ITT毕业，那么你会做和现在所做的不一样的东西吗？

**C . Mohan**：1972年到1977年之间，当我在印度理工学院马德拉斯分校学习的时候，没有关于计算机科学的毕业生项目。即使我是一个化学工程学生，但是我还是对计算机科学具有很浓厚的兴趣。所以如果我现在仍然在那，我会重新选择计算机科学学位，并且从某种形式意义上来说，在获得博士学位之前，我能够学到很多关于计算科学的知识。为了获得尽可能多的知识，在我的空闲时间里，我不得不靠自己获得参考资料，发邮件给美国高校和研究实验室。

在印度，目前学习计算机科学的学生有很多选择，谈到方式，它们可以采用暑期实习的方式。以前没有太多的印度软件公司有能够给本科生提供帮助的工作使其在工作中获得经验。但现在有很多与西方的具有研究实验室和产品的软件公司相关的活动。在印度理工学院的德里分校校园内，IBM有自己的研究实验室。所以对于计算机科学的学生来说，有更多的机会获得实际操作的经验。

**玛丽安**：那么对于数据库研究呢？

**C . Mohan**：是的。事实上，印度理工学院孟买分校，相比于其他地方的计算机科学部门，在工作人员人数方面已经是最大的数据库研究组。

**玛丽安**：哇。

**C . Mohan**：相当于威斯康辛过去所拥有的地位。印度科学院已经有一个组在VLDB和SIGMOD等类似会议上发表过很多高水平工作。如果印度学生想读一个数据库方面的硕士或博士，在印度就有做世界级研究的场所。

在工业研究实验室方面，在数据库领域，IBM是唯一一个进入印度的公司。纵然如此，数据库小组仍然很小。因为，如果一个人热衷于工业并且作数据库研究，在一个足够大的组形成之前就要花费很长时间。但是对于学者，有更多的机会。

**玛丽安**：在所有过去的研究工作中，有没有一个你喜欢的工作但是没有做到众所周知？

**C . Mohan**：之前我向你描述的我在IBM做的第一个工作是假定放弃的两阶段提交协议。当它以多种标准的形式被工业界广泛采用时，X-Open XA协议和OSI DTP协议和其他带有OTS和JTS的，并且被实现在多个公司的产品中，实际上它不被认为是R\*所要做的。研究团队也没有意识到那么多，但是它是个很基础的工作。有时在分布式系统课程中讲它。

我的第二个比较喜欢的工作，考虑ARIES的情况下，是索引并发控制和恢复。虽然ARIES的基本恢复模式已经在教科书和课堂中占用很大篇幅和时间，但是更复杂的方面——索引并发控制和恢复，没有得到足够的讲述。那个工作以短文的方式发表在SIGMOD92上，描述了极其基础的特征。所以需要更多的教与学。

我的第三个比较喜欢的工作是提交LSN（日志序列号）概念，一种简单方法，用以识别一页所有数据是否提交。这个是VLDB90上的一篇文章。

依我看来，这些工作都很吸引人，很好，很重要，在将来需要被研究团队进一步采用和利用。

**玛丽安**：如果你有额外时间做一件其他事情，不是现在所做的，你想做什么？

**C . Mohan** : 后来, 我就没有像发表ARIES算法工作那个巅峰时候那样发表那么多论文。我更喜欢能够坐下来写更多的我想到的和理解的东西。我没能找到时间, 或者说有时我不能确定我是否愿意经历写作这个痛苦过程, 写出来的东西要具有容易理解的形式, 然后发表出去。

**玛丽安** : 你需要一个学生。你的确需要一个学生。

**C . Mohan** : 我愿意花很多时间和更多的人一起工作, 而不是之后遍历这些工作。我作为IBM院士经常参与到活动中来就已经拿走了做这些事情的机会。

**玛丽安** : 作为一个计算机科学研究者, 如果你可以改变发生在你身上的一件事情, 你想改变什么?

**C . Mohan** : 这个很难回答。我猜可能是在某些特殊的软件的某些细节作的更深入, 并且更加深入的了解他们, 就像我过去一直做的。同时我更希望自己能够以一种更容易理解的方式解释更复杂的概念。以摘要的方式描述算法, 就像其他研究者所能做的一样, 我不是真正的精通, 这就是为什么许多人找到我的论文说读起来很复杂。

**玛丽安** : 好的, 非常感谢。

**C . Mohan** : 谢谢!

(范玉雷译, 富丽贞效)