分布式实时数据库的通讯延迟模型与协议*

卢炎生 谢晓东 朱英武

(华中理工大学计算机系 武汉 430074)

摘要 本文通过建立分布式实时数据库系统中的一个网络通讯延迟模型,对它的时间开销进行了分析,指出实现实时通讯的可行性.同时给出了通讯协议ECSMA/CD(extend carrier sense multiple access/collision detection)的算法,为数据或消息的实时传送提供了一种机制.

关键词 分布式实时数据库系统,网络接口单元,网络通讯延迟模型,ECSMA/CD,介质访问控制层,逻辑链路控制层.

中图法分类号 TP311.13

分布式实时数据库系统必须能够处理具有时间限制的应用,而这些应用所涉及的某些数据又不在应用本地,所以不可避免地要与网络上的其它结点进行通讯,传送数据或消息.在分布式实时数据库系统中,不仅要求数据值正确,而且具有时间限制,即在规定的时间内,值正确的数据才是有效的.所以,实时通讯中,不仅要求数据或消息传送正确,而且要尽可能保证或必须保证数据或消息在应用可允许的时间范围内完成传送.在一些特定应用中,若处理或数据传送超时,则该数据对于实时应用已毫无意义,并且可能导致灾难性的后果.[1]

要满足分布式实时数据库系统中的时间限制,一个主要的问题是预先估价系统的时间开销,而分布式实时数据库系统中时间开销的一个重要方面是网络通讯的时间开销.因此,其时间开销模型及相应的通讯协议是一个极具挑战性的课题.

分布式实时数据库系统通常是在特定的网络系统支持下实现.它们之间的接口方法或直接利用网络通讯协议,或对其协议进行改进以满足应用的时间特性,我们采用的是后一种方法.

1 网络通讯延迟模型

在分布式实时数据库系统中,每个结点通过网络接口单元NIU(network interface unit)与通讯网络(Communication Network)相连.具体说来,NIU实现主机的逻辑链路协议、发送/接收数据的设备和通讯介质之间的互连.

网络通讯时,网络接口具有从输出设备读取数据和将数据存放到输入设备的功能.我们分别用t_{get},t_{put}来表示这两个动作花费的时间.在设备和通讯介质中传送的数据缓存在输入/输出缓冲区中,NIU控制单元控制缓冲区数据的存取、数据确认和流量控制.[2]我们用t_{ibuf},t_{obuf}表示数据在输入、输出缓冲区中等待的时间.

NIU把要传送的数据分割,加入帧头和帧尾,组成相应的数据帧,用t_{xcom}表示组帧所需的时间.类似地,当接收一个数据帧时,首先要确认数据帧,确认后,去掉帧头和帧尾,并进行CRC检测,用t_{dcom}表示这些动作所需的时间.

数据的传送首先要获得对通讯介质的控制.在带有冲突的载波监听多重访问CSMA/CD系统中,由于存在信道竞争,试图发送数据帧都有可能导致冲突;而在令牌环网中,发送数据帧前必须等待一个空闲的令牌.因此,我们用t_{wait}表示试图发送数据帧到可以发送所花费的时间,而用t_{xmit}表示成功发送数据帧所花费的时间,t_{prop}表示数据帧到达接收NIU的延迟时间,t_{recv}表示接收数据帧所花费的时间.

综合以上的分析,我们可以建立一个局域网(LAN)的网络通讯延迟模型,如图1所示.为了描述简单,我们先假设NIU和网络协议在各结点上都是相同的,即各结点是同质的.在后面的讨论中再放宽这个条件的限制.

LAN的体系结构定义为层次模型,OSI模型和IEE802标准都是建立在物理层之上.在OSI模型中,第2层是链路层,而在IEE802标准中则相应地表示为两层,即MAC(介层访问控制层)和LLC(逻辑链路控制层).所有这些层次

---

* 本文研究得到国家自然科学基金和国防科技预研计划基金资助.作者卢炎生,1949年生,教授,主要研究领域为信息系统,数据库系统.谢晓东,1974年生,助教,主要研究领域为数据库系统.朱英武,1972年生,硕士生,主要研究领域为数据库系统.

本文通讯联系人:卢炎生,武汉 430074,华中理工大学计算机系

本文1997-01-27收到原稿,1997-04-25收到修改稿

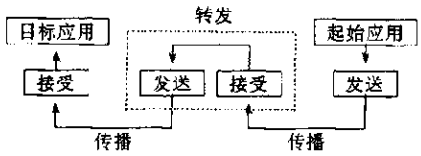

图1 LAN延迟模型

构成了LAN的协议,而更高的LAN层提供了模块化和灵活的设计环境,但也产生了延时和实时控制的问题.

在物理层中,数据帧的发送、传播和接收的延迟依赖于物理层的以下特性:带宽、拓扑结构和传输介质.根据这3个特性和一定数据帧的大小,我们可以把数据帧的发送、传播和接收表示为 $t_{xmit}+t_{prop}+t_{recv}$. 而MAC层的发送活动包括:(1) 数据帧发送前的组帧阶段;(2) 数据帧发送前的等待阶段;(3) 数据帧的确认和帧的分解阶段.所有这些活动的延迟表示为 $t_{wait}+t_{xcom}+t_{dcom}$. LLC层的活动包括:(1) 从设备传送数据到NIU;(2) 填充输出缓冲区;(3) 填充输入缓冲区;(4) 从NIU传送数据到设备.所有这些活动的延迟表示为 $t_{get}+t_{obuf}+t_{Ibuf}+t_{put}$.

对如图1所示的延迟模型,我们作以下假定:(1)一个应用起动数据传送到另一个应用;(2)起始应用和目标之间需要 $n-1$ 次数据转发(Relay);(3)由于传送的数据具有时间限制,所以没有任何恢复机制来重发丢失的数据;(4)t_{wait} 包含了信道竞争的因素.

由以上假定,我们可以得到以下结论:当结点 i 发送一数据到结点 j 时,它的延迟可用 UI 表示,即

$$UI=t_{pros}+n\times(t_{pmit}+t_{wait}+t_{xmit}+t_{prop}-t_{recv}+t_{prev})$$

其中 t_{pmit} 和 t_{prev} 分别为发送前和接受后的处理时间估算,它们又分别表示为

$$t_{pmit}=t_{get}+t_{obuf}+t_{xcom}\qquad t_{prev}=t_{dcom}+t_{Ibuf}-t_{put}$$

现在取消前面的条件限制,数据传送从起始结点 i 到终止结点 j 需经过中途结点 r_1 到 r_{n-1} 的数据转发,它的延迟为 UI.

$$UI=t_{pros}^{[i]}+t_{xCOM}^{[i]}+t_{get}^{[i]}+t_{obuf}^{[i]}+t_{Wait}^{[i]}+t_{xmit}^{[i]}+t_{pros}^{[j]}+t_{dCOM}^{[j]}+t_{Ibuf}^{[j]}+t_{put}^{[j]}-t_{prop}^{[j]}+t_{recv}^{[j]}+\sum_{k=r_1}^{r_{n-1}}(t_{xCOM}^{[k]}+t_{Wait}^{[k]}+t_{xmit}^{[k]}+t_{prop}^{[k]}+t_{recv}^{[k]}+t_{dCOM}^{[k]}+t_{get}^{[k]}+t_{put}^{[k]}+t_{Ibuf}^{[k]}+t_{obuf}^{[k]})$$

上式 UI 中,包含了可确定部分(Deterministic)和不可确定部分(non-Deterministic).可确定部分在实时系统中我们可以充分利用;而对于不可确定部分,它必须具有上限,以确保数据的实时传送.可确定部分对于某一特定的系统配置是可预测的.我们用 DUI 表示通讯延迟 UI 中的可预测部分,而 ΔUI 为不可确定部分,即 $UI=DUI+\Delta UI$. 而 $\Delta UI\approx\sum_{k=1}^{r-1}t_{wait}[k]$,即在每个结点上的 t_{wait} 是通讯延迟 UI 最主要的不可确定因素.为了确保数据的实时传送,我们必须补偿不可确定因素.

参考TCP/IP协议中的IP数据报格式,它有一个8比特的服务类型(Service Type)字段,如图2所示.它规定了对本数据报的处理方式,一共分为5个子域,其中优先权子域指示数据报的优先权,表示数据报的重要程度.它提供了一种拥塞控制的手段.例如,当网络发生拥挤时,无优先权的拥塞控制信息必然受到拥塞的影响,从而影响拥塞控制的效率.假如网络软件服从优先权,则可以给拥塞控制信息赋予较高的优先权,从而设计出不受拥塞影响的拥塞控制算法.[3]

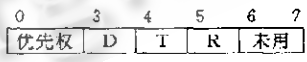

图2

我们可以从TCP/IP协议中的IP数据报格式中受到启发,在分布式实时系统中,给具有时间限制的数据/消息赋予较高的优先权,高优先权的数据/消息优先发送,从而使通讯延迟中的不可确定部分 $\Delta UI\approx\sum_{k=1}^{r-1}t_{wait}[k]$ 得到补偿,实现数据/消息的实时传送.在下面部分中,我们将给出一个实时通讯协议.

2 实时通讯协议

带有冲突检测的载波监听多重访问CSMA/CD协议广泛实用于介质访问控制层(MAC),它是在发送数据的同时进行冲突检测,一旦发现冲突,立刻停止数据发送并等待冲突平息,然后再进行CSMA/CD,直到将数据成功地发送完毕.CSMA/CD有两个关键问题,它们涉及到等待时间策略:(1) 监听信道忙时,是否继续监听载波,监听到信道空闲时是否立即发送数据;(2) 一旦检测到冲突,需要等待多长时间再进行CSMA/CD.

ECSMA/CD也是一种带有信道竞争的通讯协议.它在系统中每个结点上定义了两个时钟:一个是实时时钟 $T_p(t)$,一个是虚时时钟 $T_v(t)$.实时时钟 $T_p(t)$ 在不断地计时,而虚时时钟 $T_v(t)$ 在信道忙时便停止计时,一旦信道空

闲时便和 $T_p(t)$ 同步，当 $T_v(t)$ 在计时时，它的计时速率比 $T_p(t)$ 要快，即速度 $a_v(t)>a_p(t)$，其中 $a_p(t)\approx 1$. 在 ECSMA/CD中，$a_v(t)/a_p(t)$ 是一个参数，我们可以根据实际应用的需要进行设置.

在分布式实时数据库系统中，我们将具有时限的数据或消息从结点 i 传送到结点 j 的时间限制表示为 d，用 l 表示数据/消息的时间长度，而 DU_i^j 为通讯延迟，则数据或消息发送的最迟开始时间为 $\text{begin}_{\max}=d-DU_i^j-l$，那么在时间 t，数据/消息实时发送的宽裕时间为 $x_t=d-DU_i^j-l-t$. ECSMA/CD 的发送原则是时间最少宽裕者优先(Minimal-Laxity-First). 由于 $a_v(t)>a_p(t)$，所以发送者的 $T_v(t)$ 一旦和 $\text{begin}_{\max}$ 相等，便立即发送数据/消息. 在 ECSMA/CD中，有一个等待队列 wait-queue，队列中的数据/消息按 x_t 从小到大排列. 下面就给出 ECSMA/CD 协议的类 C 语言的算法描述.

数据/消息 m 的发送算法.

```
void Transmit M(m)
{
  begin_max^m = d_m - DU_i^j - l_m;
  if (信道空闲且 T_v(now) ≥ begin_max^m)
      send(m); /* 立即发送入 */
  else /* 插入等待队列 */
  {
    x_t^m = begin_max^m - T_p(now);
    insert (m, x_t^m), wait-queue);
  }
}
```

当检测到信道空闲时：

```
void Do When Idle ( )
{
  /* 重置虚时时钟和全局变量 P */
  T_v(now) = P = T_p(now);
  /* 从队列中移去超时的数据/消息 */
  for (i-header (wait-queue); i 不为空; i-next (wait-queue))
    /* 等待队列每个元素 */
    if (begin_max(i) < T_p(now))
     /* 从等待队列中移去 i */
    remove (I, wait-queue);
}
```

当发送数据/消息 m 时，检测到冲突后，采用随机算法重发数据/消息 m，该算法描述如下：

```
void Random Retansmit (m)
{
  begin_max^m = random(T_p(now), d_m - DU_i^j - l_m);
  x_t^m = begin_max^m - T_p(now);
  insert (m, x_t^m, wait-queue);
}
```

上述算法中用到的虚时时钟的计时算法为：

```
void Tick ( )
{
  while 信道空闲时
    {
      T_v(now) = a_v(now) × (T_p(now) - p);
      /* candidate 取为等待队列的第 1 条数据/消息 */
      candidate = header (wait-queue);
      if (begin_max(candidate) = T_v(now))
          /* send by minimal-laxity-first */
          {
          /* remove candidate from wait-queue */
          remove (candidate, wait-queue);
          send (candidate);
          }
    }
}
```

3 结束语

对于分布式实时数据库系统中的实时通讯，我们经过通讯延迟模型的分析，得出了数据/消息实时通讯的可行性，并给出了一种扩展的CSMA/CD协议算法，为数据/消息的实时传送提供了根据.但困难的是延迟模型的不可确定因素的参数化或补偿方法，这需要根据具体系统背景进行大量的实验与分析.目前，我们拟在UNIX网络环境下开展实验研究，初获结论后，将另文发表.

参考文献

1 卢炎生等.实时数据库管理系统研究.见：徐秋元编，全国第11届数据库会议论文集.西安：西北工业大学出版社，1993.452～457
(Lu Yan-sheng *et al*. A study of real-time database management system. In: Xu Qiu-yuan ed. Proceedings of the 11th China Conference on Database. Xi'an: Northwestern Polytechnical University Press, 1993. 452～457)

2 Agrauala A K, Levi S T. Objects architecture for real-time, distributed, fault tolerant operating systems. IEEE Workshop on Real-time Operating System, 1987,7(1):142～148

3 周明天，汪文勇.TCP/IP网络原理与技术.北京：清华大学出版社，1993
(Zhou Ming-tian, Wang Wen-yong. Principle and technology of TCP/IP netware. Beijing: Tsinghua University Press, 1993)

The Network Communication Delay Model and Communication Protocol of Distributed Real-Time Database System

LU Yan-sheng XIE Xiao-dong ZHU Ying-wu

(*Department of Computer Science Huazhong University of Science and Technology Wuhan* 430074)

Abstract This paper is proposed the time expenditure of network through a network communication delay model and guarantees the feasibility of real-time communication. An algorithm of the extend CSMA/CD(carrier sense multiple access/collission detection) protocol is also presented to implement the real-time transmission of data or message.

Key words Real-time database, network interface unit, network communication delay model, extend CSMA/CD, medium access control layer, logical link control layer.

Class number TP311.13