

基于 DDS 网络中间件的持久化数据管理问题研究

刘宏义, 杨 明

(解放军边防学院 战斗实验室, 陕西 西安 710108)

摘 要: DDS 网络中间件采用以数据为中心的实时发布/订阅通信模型, 具有无中心、数据多备份的特点, 因此, 给通信数据的管理带来了完整性和一致性的问题, 如何高效集约地管理和维护数据, 处理好负载分配问题, 使多个通信结点间保持历史数据一致, 成为 DDS 网络中间件研发设计中必须面对的挑战. 对此从 DDS 使用的通信模型入手, 研究分析了在此种模型下引发的持久化数据管理问题, 并给出了相应的解决方案.

关键词: DDS; 中间件; 发布-订阅; 持久化数据; 一致性; 负载分配

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2016)10-0163-04

Research on Persistent Data Management Problems Based on DDS Network Middleware

LIU Hong-yi, YANG Ming

(Battle Laboratory College of the PLA Border, Xi'an 710108, China)

Abstract: DDS network middleware uses data-centric real-time publish / subscribe communication model has the characteristics of non- center and multi-replica, therefore, it brings integrity and consistency challenges for communication data management. In order to efficiently manage and maintain data intensive, and handle load balancing, a plurality of communication between nodes consistent with historical data for real-time publish/subscribe network middleware research design must face the challenge. This paper starts from the communication model used by the DDS and then analyzes persistent data management problems caused by such model, given a kind of appropriate solutions.

Key words: DDS; middleware ; pub-sub; persistent data; consistency; load balancing

1 引言

以数据为中心的实时发布-订阅通信模型在通信数据管理和维护方面, 不同于 P2P 分布式存储. DDS 应用具有硬实时、数据量较少的特点, 其数据处理的核心问题是在不明显影响实时性的情况下, 保证分布式系统中数据的完整和一致. 仅从动态数据交换角度来说, 发布-订阅模型无疑是最适用于松耦合数据交换的一种通信模式, 但从静态数据访问角度来说, 这种多数据源的通信方式存在潜在的数据一致性问题 and 负载分配问题.

2 DDS 网络通信模型的基本原理

如图 1^[1-3]所示, DDS 使用以数据为中心的通信

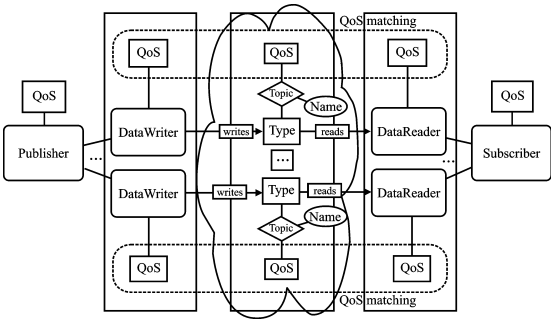


图 1 DDS 网络通信模型

模式, 数据的存在与否并不依赖于某一个具体的物理结点. 围绕着一个数据对象, 数据发送者、接收者被抽象成为这一类数据的发布者、订阅者. 应用程序将不用直接关心目的地的具体网络地址和端口, 只

要将需要共享的数据对象提交给 DDS 实现系统,由 DDS 负责基于消息内容进行无中心的传输,送达数据订阅端。

相反,在以对象为中心的通信中,基本概念是应用程序间的接口,接口由一组已知类型方法(方法参数的数量和类型)组成。以对象为中心的系统由接口服务器和客户端组成,通信则基于客户端在对应服务器所服务的命名接口上调用方法。以数据为中心和以对象为中心的通信方式是分布式系统中互补的两种模式,前者更适合实时通信,后者更适合应用程序。

3 DDS 网络通信模型的数据管理问题

DDS 系统本质上是将传输层协议进行了封装,是一个典型的消息中间件。它直接提供给应用层程序标准的通信接口,让应用程序可以灵活地定义通信类型和通信质量需求(QoS),从根本上消除各种通信软件之间数据交换的障碍。DDS 采用的以数据为中心的发布/订阅模式可以良好的适应松耦合、无中心的分布式消息交换应用,数据发送、接收方可以自由加入和退出通信组。与有中心的 SAFMQ (Store And Forward Message Queue)不同,DDS 这种无中心的消息交换模型直接造成了对历史数据维护的困难,主要表现为数据一致性问题 and 负载分配问题两个方面。

3.1 数据一致性问题

由于 DDS 的通信模式是以数据为中心,许多个结点都可以向同一个通信组写入消息。

在不考虑结点失败的情况下,设在一个通信组内总数据集 $S = \{s_1, s_2, s_3 \cdots, s_n\}$, S 由 n 个数据块组成,并且分别由数据写入者 $W = \{w_1, w_2, w_3 \cdots, w_n\}$ 集合进行发布并记录在永久介质中,即每个写入者只维护有编号与之相同的数据块,读取者集合 $R = \{r_1, r_2, r_3 \cdots, r_m\}$,每一个订阅者都拥有完备的数据集 S 。物理结点集合 $N = \{n_1, n_2, n_3 \cdots, n_r\}$,所拥有的数据集 $\{S_1, S_2, S_3 \cdots, S_n\}$ 。

设: $n_i = \{R_i, W_i \mid R_i \subseteq R \text{ 且 } W_i \subseteq W, 1 \leq i \leq r\}$
 $R_i \cup W_i \neq \emptyset, n_i \cap n_j = \emptyset (1 \leq j \leq r \text{ 且 } i \neq j)$

当一个新的读取者 r_{m+1} 加入到读取者集合中,请求历史数据时,数据提供结点应当拥有完整的数据集 S (即 $S_i = S \ 1 \leq i \leq r$)。这样的结点需要满足:

$n_s = \{R_s, W_s \mid R_s \neq \emptyset \text{ 或 } W_s = W, 1 \leq s \leq r\}$ (条件 1)

由于网络环境不稳定,满足条件 1 的结点会因为某种原因不存在于系统,因此 r_{m+1} 将不能获得完整的数据集 S ,分布式结点中的数据视图将发生在同一时间点上不一致的情况。

3.2 负载分配问题

解决数据访问一致性问题的前提是在无中心的数据交换模型中选出一个保证具有完整数据集的结点(Master),所有数据请求全部由该结点负责传输。数据源的唯一性是数据访问一致性的必要保证。一个结点对历史数据的访问本质上是跨结点的内存访问。随着分布式系统的扩展,这种跨结点内存访问将会遇到结点 I/O 瓶颈,严重时将破坏数据传输的实时性。因此,不仅需要选择高效的存储方案,而且需要给上层应用程序提供负载分配的解决方案。

4 DDS 网络通信模型的数据管理解决方案

基于 DDS 的应用有其自身的特点:结构化的数据、较小却频繁的数据交换、硬实时性都使其区别于 P2P 的大规模数据传输。依据某种策略在某个通信范围内寻找到一个统一的历史数据维护结点(Master),负责向各个结点提供统一的数据视图,并保证 Master 结点的可替代性,将是问题的关键。

4.1 数据一致性解决方案

DDS 通信模型的基本操作单元是通信组,只有同一个通信组内的数据写入和读取者才能相互交换同种数据模型的数据。因此 DDS 数据一致性问题的原子问题即是某一个通信组下数据的一致性问题。考虑到 DDS 应用程序需要很高的容错性,需要通过增加不同结点中完整数据的备份个数来增强系统鲁棒性。

4.1.1 完整数据集获取策略

满足条件 1 的结点才会拥有完备的数据集 S , $W_i = W$ 条件与 DDS 松耦合性质冲突,应当通过满足 $R_i \neq \emptyset$ 这个条件来获取完整的数据集 S 。 R_i 需要满足另一个前提条件: R_i 必须在 S 由空集变为非空集之前就存在于结点 n_i 中,并且在 n_{m+1} 请求历史数据之前没有退出过系统,这样才能保证最终 $S_i = S$ 。但这样的条件是苛刻的,不能够容忍结点失败。因此需要引入数据校准的方式来将 S 进行传递,从而打破对于苛刻条件的依赖。

校准流程如图 2 所示。

(1) 通信域第一个结点 1 启动,将其选择为 Master 结点,并由系统产生一个 R_1 专门订阅通信组内的数据,以满足条件 1,此时数据通信开始,

Master 结点中的 W_1 开始生产数据 D_1 .

(2)第二个结点启动后不论其上层应用是否请求历史数据,此结点都向 Master 结点请求其产生的所有历史数据,将完整的历史数据集复制于本地,同时由系统新建一个 R_2 . 这时结点 2 的 W_2 向网络写入数据 D_2 ,由于 Master 具有 R_1 ,所以 W_2 所写的数
据也将传给 Master 一份加入到 S 中,此时结点 1 和 2 的数据集都为 $S=\{D_1,D_2\}$.

(3)第三个结点启动,同样从 Master 处复制一份完整数据集 $S=\{D_1,D_2\}$ 到本地,同理建立 W_3 和 R_3 ,写入 D_3 .

(4)最后,通信系统中每一个结点都拥有完整数据集 $S=\{D_1,D_2,D_3\}$,当结点 1 失效后,依然可以保证通信组中完整数据集的存在.

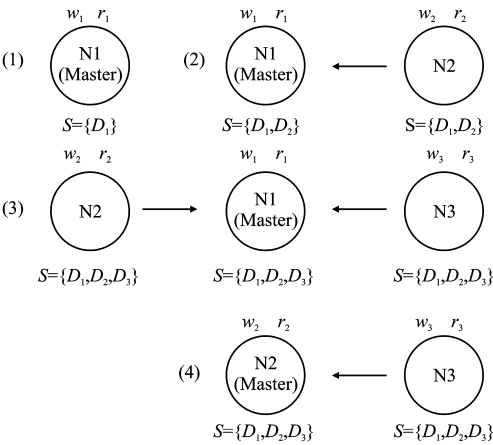


图2 数据校准实例

数据校准过程是为了避免数据集 S 的一致性随结点的进入和退出而损坏,与上层应用程序需要与否无关. 通信数据是否需要持久化存储,是可选的 QoS 保证之一,若一个数据不需要持久化,那么不会启用校准策略.

4.1.2 数据内容一致性

Paxos 算法很好地解决了分布式系统中多备份数据一致性问题^[4]. 它提供了完整的理论证明,操作请求结点接收到超过一半结点的回应后即可开始操作数据,并且不会破坏一致性. 为了避免 Paxos 算法的“活锁”问题,将采用其改进算法 Fast Paxos^[4].

Paxos 算法及其改进算法直接适用于小规模数据(Zookeeper^[5-8]建议小于 1 MB)一致性的达成. 但 DDS 应用程序所产生的历史数据量将会较大,不能直接使用 Paxos 算法去对所有历史数据进行一致性维护,否则将会随着数据量的增大严重影响到系统实时性. 因此需要一种方案来对某些系统关键部分

进行一致性处理.

除去网络传输差错影响,可以认为从同一个结点获取的完整数据集是相同的. 有了数据维护结点(Master),数据内容一致性问题就转化为了 Master 选取的一致性问题. 使用 Paxos 分布式一致性服务表现为将 Master 结点的 ID 号作为系统内一致性的操作对象,由 Fast Paxos 中的 Leader^[4] 来提交 Master 的建议值,并在通信系统各个结点中达成 Master 的最终一致.

网络传输的差错或者人为的篡改都会造成从同一结点获取的数据内容不正确,将 Master 处的完整数据集使用 MD5 摘要算法生成一个唯一值,将这个值也作为一致性服务的操作对象,如果某个结点发现本身具有的数据集的摘要值与 Master 的摘要值不同,需要重新从 Master 处得到一份数据. 保证自己维护的数据时刻与 Master 保持内容上的一致.

4.1.3 数据操作一致性

DDS 给上层应用不仅要提供获取数据的功能,还应当具备数据操作的接口. 完整数据集 S 在整个分布式系统中将会有很多备份,多个备份之间必须保持一致才能保证历史数据的一致. 所有的数据操作请求应首先提交给 Master 结点,Master 结点通知所有的结点按照此操作请求来操作数据. Master 会对每一个数据操作进行编号,其他结点若没有收到值较小的操作不会进行较大编号操作,防止了操作命令乱序. Master 会在每个校准流程开始时给请求结点发送一个数据包,包括此数据副本的最近操作编号,结点校准完毕后,不再接受小于此编号的操作命令. 每一次对数据集的操作都会通过 Master 进行处理,这样的策略给事务的提交和回滚提供了可能. 如图 3 所示.

4.2 负载分配策略

随着结点扩展,如果依然按照单个通信组的处理方式选举出一个控制整个通信域的 Master 结点来保证数据的完整和一致,将会造成通信域中 Master 结点的过载,造成系统响应缓慢,失去了实时性的保证. 因此需要一种合理的方案或者手段来降低结点负载量.

4.2.1 两种 Master 选举方法

(1)全局唯一 Master:全局 Master 为了实现功能将会加入通信域中所有的通信组. 这种策略在结点数量不多的情况下还能适用,但系统的可扩展性将会很差. 随着 DDS 应用结点的增多,通信域中任

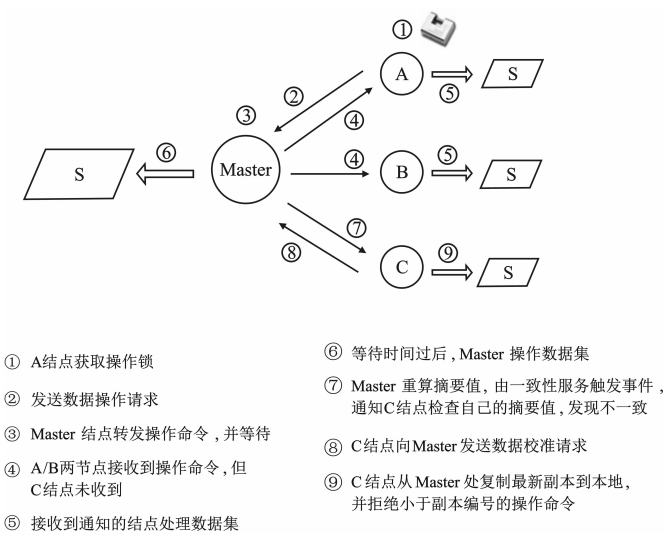


图3 数据操作一致性实例

何一个结点都无法达到需要的 I/O 速度,这就导致一个请求有可能很久后才得到回应,破坏了实时性.

(2)每个通信组都包含一个 Master: 这样的策略可以通过控制 Master 选举过程来将不同通信组的 Master 分配到尽量分散的结点,从而降低具体 Master 结点的负载量. 这样的策略理论上很好的增强了系统的可扩展性,但在实际情况中却并非如此. Master 的选举牵涉到通信域内的一致性,所以当通信组很多的情况下,选举过程所耗费的时间将严重影响到系统实时性.

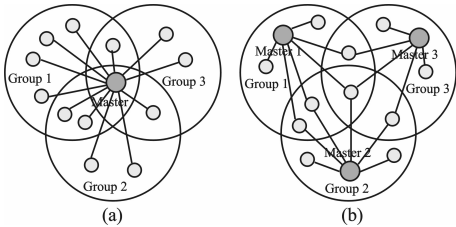


图4 不同 Master 的情况

Master 的选举范围是问题的关键,应当使用一种折中的策略来划分 Master 的选举范围,由使用 DDS 实现系统的系统管理员来灵活的配置 Master 选举范围,从而让 Master 负载问题和 Master 选举延时问题对系统实时性的影响降到最低. 为了解决这个问题,需要引入命名空间的概念. 不同 Master 的情况如图 4 所示.

4.2.2 命名空间

设通信组的集合 $G = \{g_1, g_2, g_3 \cdots, g_n\}$, 命名空间集合 $P = \{p_1, p_2, p_3 \cdots, p_m\}$, $p_i = \{G_i \mid G_i \subseteq G, 1 \leq i \leq m\}$, $G_i \cap G_j = \emptyset$. 一种 DDS 应用或许会涉

及多个数据类型,每个数据类型都划分出一个通信组,命名空间从数据相关性的角度对多个通信组再做出一个划分,同一个命名空间下的通信组会选择出同一个 Master,拥有统一处理的数据集合. 给上层应用提供这种手段后,系统管理员可根据具体需要对通信组进行划分,找出实时性最好的划分方案.

该方案应用在了清华大学实时服务总线 (RTSB)原型系统上. Master 重选延迟和通信域结点数及命名空间个数的关系如图 5 所示. 结点数越多或者同一时刻需要重选 Master 的命名空间越多,Master 选举完成的延迟越长. 因此命名空间的划分原则应当是在单一结点服务能力不会出现瓶颈的情况下,尽量少的划分命名空间,才能保证系统的高可伸缩性.

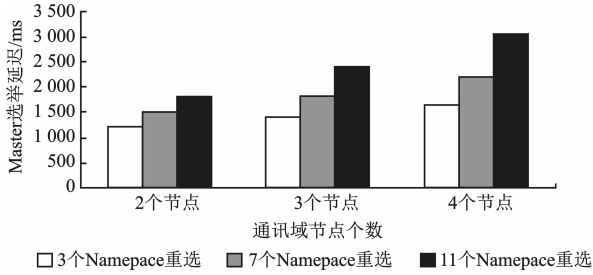


图5 Master 重选延迟

在实际的应用中,就算引入了命名空间的概念,依然不能真正降低 Master 选举的时间,这就给系统响应实时性带来了隐患. 命名空间中具有完整数据集的结点将会有很多,如果某个结点仅是为了获取数据而不需要操作数据,那么任何一个具有完整数据集的结点都可以作为数据的提供者,而不需要等待到选出 Master 后才响应数据请求. 因此在一个命名空间中没有 Master 时,系统应当选取一个具有命名空间完整数据集的结点来作为数据提供结点,进一步增强系统实时性.

5 结束语

本文从以数据为中心的发布-订阅模型入手,研究分析了在这种通信模型下持久化数据的管理存在的问题,结合 DDS 通信模式,设计了灵活的数据管理域划分机制,选取了合适的数据管理结点选择策略,在不严重影响系统实时性的情况下应用 Paxos 算法较好解决了 DDS 网络通信中间件在网络通信中历史数据一致性问题.

(下转第 172 页)

体现在以下三点:一是对比了 DSR 算法中 $A(t)=D(t)$ 和 $A(t)=D(t)$ 情况下,对 $\hat{X}(t)$ 解收敛性和准确性的影响;二是改进了 DSR 算法的计算模型,增加帧间动态信息的传递和获取,消除了原 DSR 算法处理结果中的“齿状”效应,并减少了原 DSR 算法的运行时间;三是通过增加空间域的边缘指导信息,更好地保护了边缘信息,以及改善图像边缘锯齿,同时不增加过多额外的时间开销.实验表明,本文提出的算法与传统超分辨率算法相比,能够有效恢复更多图像细节,且运行时间相较于 DSR 大大减少,可得到在复杂度和图像质量上都占优的处理结果.下一步工作是针对去模糊正则项进一步深入研究,提高该算法的鲁棒性.

参考文献:

- [1] Baker S, Kanade T. Limits on super-resolution and how to break them [J]. PAMI, 2002, 24 (9): 1167-1183.
- [2] Wang L F, Xiang S M, Meng G F, et al. Edge-directed single-image super-resolution via adaptive gradient magnitude self-interpolation[J]. IEEE Trans on Image Processing, 2013, 23(8): 1289-1299.
- [3] Freedman G, Fattal R. Image and video upscaling from local self-examples [J]. ACM Transactions on Graphics, 2011, 30(2): 474-484.
- [4] Irani M, Peleg S. Improving resolution by image registration[J]. Cvgip Graphical Models & Image Processing, 1991, 53(3): 231-239.
- [5] Schultz R, Stevenson R. Extraction of high-resolution frames from video sequences[J]. IEEE Transactions on Image Processing, 1996, 5(6): 996-1011.

- [6] Farsiu S, Robinson D, Elad M, et al. Fast and robust multiframe super resolution[J]. IEEE Transactions on Image Processing, 2004, 13(10): 1327-1344.
- [7] 华书娜,王培康,朱高,等.基于聚类和高斯过程回归的超分辨率重建[J].微电子学与计算机,2012,29(10): 26-29.
- [8] Farsiu S, Elad M, Milanfar P. Video-to-Video dynamic super-resolution for grayscale and color sequences [J]. EURASIP Journal on Applied Signal Processing, 2006, 2006(1): 1-15.
- [9] Elad M, Feuer A. Super-resolution restoration of image sequence [J]. IEEE Transactions, 1999, 21 (9): 817-834.
- [10] Elad M, Feuer A. Super-resolution restoration of an image sequence: adaptive filtering approach on Image Processing [J]. IEEE Transactions, 1999, 8 (3): 387-395.
- [11] Simonyan K, Grishin S, Vatolin D. Fast video super-resolution via classification[C]//Proc IEEE International Conference on Image Processing (ICIP). California, USA; IEEE Press, 2008: 349-352.

作者简介:

汤晓莉 女,(1993-),硕士研究生.研究方向为基于运动补偿的视频去噪,视频超分辨率.

E-mail: xxtxl@mail.ustc.edu.cn.

韩睿 女,(1985-),博士,助理研究员.研究方向为图像去噪、图像缩放、视频帧率转换.

郭若杉 女,(1977-),硕士,高级工程师.研究方向为图像分割、目标识别、图像融合.

(上接第 166 页)

参考文献:

- [1] Angelo Corsaro. Addressing the Data-Distribution Challenges of Next-Generation Business and Mission-Critical Systems[EB/OL]. [2016-03-02]. <http://www.dis.uniroma1.it/tucci/corsi/SD08/SD08-dds.pdf>. 2008.
- [2] Gerardo Pardo - Castellote. OMG Data-Distribution Service (DDS) Achitectural Overview[C]// IEEE International Conference on Distributed Computing Systems. Canada: IEEE, 2003.
- [3] Woonchul Kang. RDDS: A Real-Time Data Distribution Service for Cyber-Physical Systems[J]. Industrial Informatics, 2012, 8(2): 393-405.
- [4] Leslie Lamport. Fast paxos[J]. Distributed Computing, 2006, 19(2): 79-103.
- [5] Flavio P Junqueira, Benjamin C Reed, Marco Serañi. Zab: High-performance broadcast for primary-backup systems[C]// IEEE/IFIP 41st International Conference on Dependable Systems & Networks. Canada:

IEEE, 2011.

- [6] Marco Primi. Paxos made code: Implementing a high throughput Atomic Broadcast [D]. Luxembourg: University of Lugano, 2009.
- [7] The Real-time Publish-Subscribe Wire Protocol DDS Interoperability Wire Protocol Specification Version 2.1. OMG Available Specification formal [EB/OL]. [2016-01-12]. <http://www.omg.org/spec/DDS/2.1>. 2009.
- [8] 任昊利,李旺龙,张少杨,等.数据分发服务—以数据为中心的发布/订阅通信[M].北京:清华大学出版社,2014.

作者简介:

刘宏义 男,(1972-),副教授.研究方向为作战仿真和虚拟现实技术. E-mail: newsolder@Sohu.com.

杨明 男,(1972-)副教授.研究方向为指挥自动化和作战模拟.