基于自律分散系统的通信中间件研究与实现

姚 兰[1,3]，桂 勋[2]，谭永东[3]

(1. 成都信息工程学院控制工程学院 成都 610225; 2. 电子科技大学空天科学技术研究院 成都 610054;
3. 西南交通大学电气工程学院 成都 610031)

【摘要】介绍了自律分散中间件的逻辑抽象结构和概念，提出了具有高可靠性和容错性的自律分散中间件物理结构，并详细讨论了中间件采用的各项关键技术：为构建基于数据驱动的应用系统设计技术，提出了可支持多种通讯方式的归一化句柄编程模型和基于通信句柄的信息映射机制；为实现自律分散网络的在线扩展和在线维护特性，提出了广义混合队列模型；为实现自律分散网络的在线容错特性，提出了三阶段生存信号确认算法和基于发布定购模型的可靠组播通信算法。通过构建10个节点的分散系统，验证了自律分散网络的系统特性。

关 键 词 生存信号; 自律分散系统; 发布定购; 中间件

中图分类号 TP31 **文献标识码** A doi:10.3969/j.issn.1001-0548.2012.05.023

Research and Implement of Communication Middleware Based on Autonomous Decentralized System

YAO Lan[1,3], GUI Xun[2], and TAN Yong-dong[3]

(1. Deptement of Control Engineering, Chengdu University of information technology Chengdu 610225;
2. Institute of Astronautics & Aeronautics, University of Electronic Science and Technology of China Chengdu 610054;
3. Deptement of Electrical Engineering, Southwest Jiaotong University Chengdu 610031)

Abstract The logical abstract structure and concept of autonomous decentralized system (ADS) middleware are introduced. A physical structure of ADS middleware with high reliability and fault-tolerant capacity is presented and the related key technologies are discussed in detail. In order to construct data-driven application system design technology, the normalized handle programming model supporting several kinds of communication methods and information mapping mechanism are proposed. For the implementation of online expansion and online maintenance, generalized mixed-queue model are provided. The three-stage survival signal confirmation algorithm and reliable multicast communication algorithm based on the publish/subscribe model are put forward for the realization of online fault-tolerant capacity. Though the construction of 10 nodes ADS, the characteristics of ADS were verified.

Key words architecture; autonomous decentralized system (ADS); live signal publish/subscribe; middleware

自律分散系统(ADS)[1-4]是近年来发展的一种分散式系统模型，它打破了传统的集中式系统模型，又与现在流行的集散式系统有很大区别，是一种新型系统模型。在该系统中，各个组成部分是一个独立的整体，能不受外部控制而独立完成内部功能，同时还能够主动及时地向外部发送信息。利用该系统模型组建的网络具有自律控制和自律协调的能力，能够较好地实现系统在线扩展、在线维护和在线容错等功能。

日立公司在该理论基础上成功实现了中间件系统及相关通信协议ADP[5]，采用该中间件系统构筑了各种大型和超大型的分散式系统(如东京的交通运行系统ATOS[6-9])。ADS的相关标准已经通过了ISO相关组织的审核，成为了国际标准。

本文通过深入研究ADS及中间件理论，提出了一种基于ADS理论的中间件系统模型，基于该模型构造的中间件良好地运用于空中交通监控自动化系统等大型监控系统，证明了该模型的科学性和可行性。

收稿日期：2011 – 02 – 17; 修回日期：2012 – 02 – 23
基金项目：中央高校基本科研业务费专项资金(ZYGX2009J089)
作者简介：姚兰(1980 –)，女，博士生，主要从事自律分散、中间件技术方面的研究.

1 中间件体系结构

1.1 中间件的逻辑抽象结构和概念

基于ADS体系结构理论，本文提出了从逻辑节点到作用域的由低层到高层的分散式系统抽象概念，形成以数据为驱动中心的系统设计模式，大幅简化复杂应用系统的设计过程。

1) 逻辑节点(logic node)。包含在数据域中的一台计算机，在一个网段上系统的逻辑节点在0～255间。2) 组播组(multicast group，MCG)。一个组播组是一组属于同一个数据域的节点集合，在该集合中的节点对其进行广播，只有该组播组中的节点能够接收。多个组播组成一个数据域，数据域负责管理内部的组播组。组播组负责在同一个数据域中执行各种分配而又相互协调的工作，是系统间相互协调自律工作的基础。3) 数据域(data field)。数据域是包含一个逻辑节点集合交换数据的抽象地点，一个数据域包含多个节点和多个组播组。数据域是比组播组更高的抽象层次。组播组合数据域如图1所示。4) 作用域(domain)。作用域是在数据域之上更高的抽象概念，一个作用域包含多数据域，通常一个作用域对应一个局域网。作用域如图2所示。5) 事务(transaction)。一个事务表示处理关于服务、实时事件等的一次抽象过程，在监控系统中，各个子系统间的一次业务数据交换就表示一个事务。6) 事务代码(transaction code，TCD)。标识一个事务的代码为事务代码，在监控系统中有许多事务代码，其中1～59 999为上层节点子系统间通讯所有的事务代码，而60 000以上为驻留代理间的各种相互协调的事务间的代码。

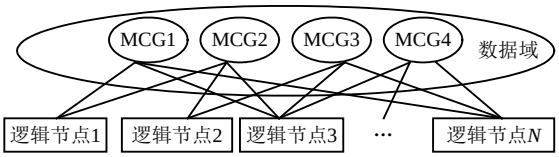

图1　组播组和数据域

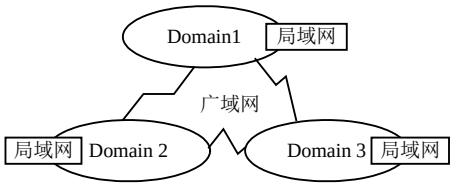

图2　作用域

1.2 中间件物理结构

为了尽可能保证可靠性和容错性，该模型在物理结构上把中间件系统设计为一个多进程结构，为进程间通信提供了一条基于本地令牌环网的松散耦合方式，形成了内部的一条软件总线。这样就可实现动态扩展，提高模型的可扩展性和容错性。中间件物理结构如图3所示。

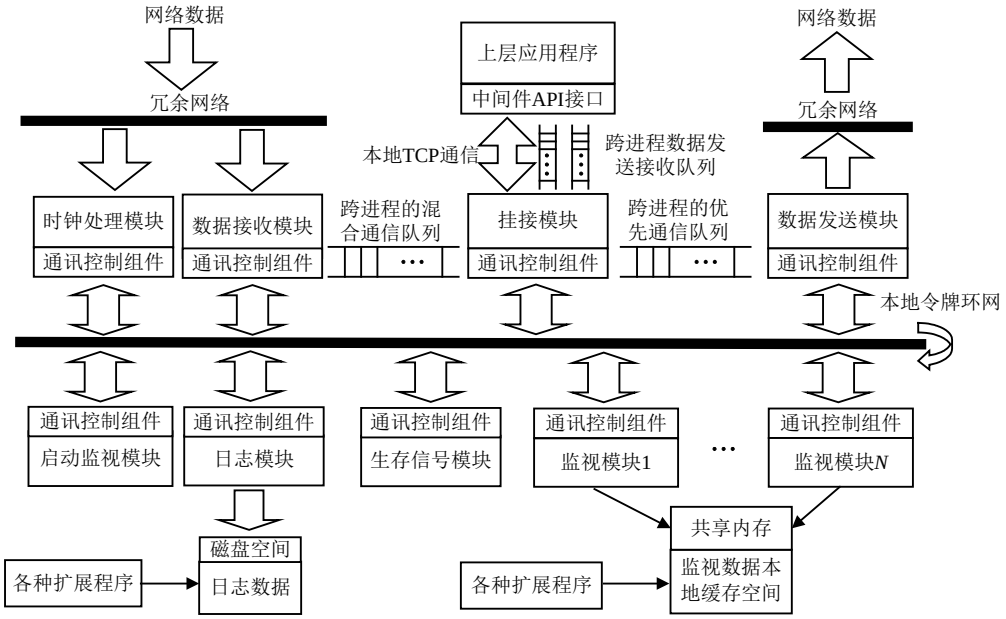

图3　中间件物理结构

由图3可见，系统由8个主要模块构成：挂接模块、发送模块、接收模块、启动监视模块、生存信号模块、时钟模块、日志模块和监视模块。每个模块都包含一个相同结构的通讯控制组件，该组件间可通过本地UDP协议栈进行通信，形成中间件内部的一条软件总线，从而实现模块间的信息共享和生存信号监视。同时，中间件内部形成一个相互监视机制，保证出现异常的模块被及时发现并重新启动。

其中，挂接模块、发送模块、接收模块为系统的核心模块，只要具有这3个模块就可完成通信工作；启动模块、生存信号模块为次核心模块；其他模块为外围模块。除了核心模块以外，其他模块都可根据配置进行加载。

1) 中间件的核心模块。

挂接模块内部维持了一个高效率的本地TCP连接池，上层应用程序模块通过本地TCP协议栈和挂接模块进行链接，完成各种初始化，并服务于系统生存信号的发送，以及跨进程输入输出队列的流量控制。发送模块和接收模块内各自都维持了一个远程TCP通信线程池和组播通信线程池，从上层应用程序模块发出的数据被汇流到一个跨进程的优先队列，然后由发送模块逐一发送出去。接收模块将通过冗余网络接收到的各种数据(包括普通数据、生存信号数据等)汇流到一个跨进程的混合通信队列，从队列的另一头逐一取出数据后分发到对应的上层应用程序模块。

2) 次核心模块。

启动监视模块负责根据系统配置文件加载中间件的各个进程，初始化内部需要交换数据的各个共享内存表，并作为生存信号仲裁者，根据生存信号状态结束并重启故障模块。生存信号模块负责接收和发送生存信号，当它监视到系统发生异常，会临时打开通信队列，向其内插入一条生存信号消息，让挂接模块及时通知上层应用程序模块。

3) 外围模块。

监视模块是一个可动态扩展的结构，不同的监视模块把各自采集的数据以自定义XML的方式进行发送。日志模块负责接收并保存系统内各个模块的状态信息，维持一个跨进程的日志信息接收队列。其他任意模块在需要保存状态时，都可向其内插入自定义的XML状态信息。时钟模块负责接收时间信息，以维护整个分散系统的时间同步。

2 中间件关键实现技术

自律分散系统的最大特点是基于数据驱动的系统设计模型，并且系统同时具备在线扩展、在线维护和在线容错特性，以下将讨论实现各项特性的关键技术。

2.1 基于数据驱动的应用系统设计模型关键技术

1) 基于句柄的编程模型。

不同于传统网络系统应用的设计逻辑，自律分散系统提出了基于数据驱动的应用系统设计模型。为了实现该模型，并避免传统网络应用系统实现技术所导致的代码耦合度、维护和升级成本高的缺点[10-12]，本文系统提供了以句柄为单位的编程模型，如图4所示。屏蔽底层TCP\UDP的通信细节，提出一套通用的编程概念，使开发人员只使用一套编程模型就可完成各种不同通信需求下的开发任务。针对各种通信情况，系统提供了4种通信句柄。

① 点对点通讯发送句柄：指定通信网络的情况下，底层采用TCP进行远程通信；② 组播通讯发送句柄：指定目的地数据域号、目的地节点号、目的地TCD的情况下通过组播方式进行远程通信；③ 多TCD接收句柄：接收到多个TCD中的一个或多个，句柄间可共享TCD；④ 指定单TCD接收句柄：接收到指定TCD，句柄间不共享TCD，功能模块由产生句柄的用户独占。

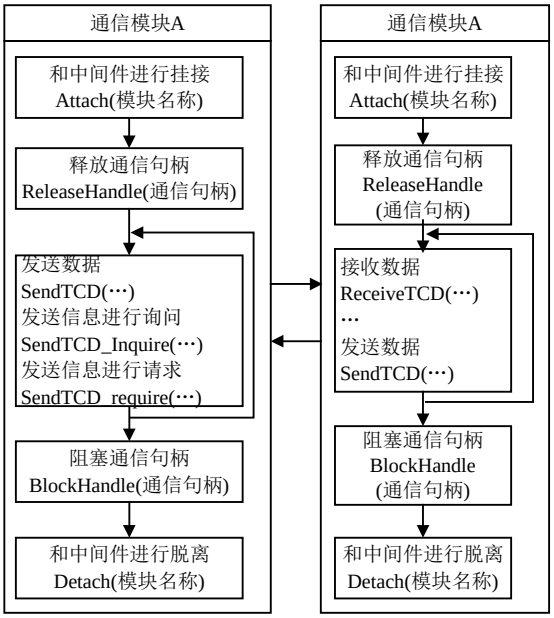

图4 中间件的编程模型

系统以4种句柄为基础提供了4种通信模式：点对点的请求应答通讯、多点请求一点应答、多点请求多点应答、一点请求多点应答，以满足各种通信需求。在编程开发上设计了一个对4种模式都通用的模型，如图4所示，通信的任意一方都需要在模块生命周期开始和结束的时候，明确地向中间件进行挂接和脱离。挂接以后模块将在中间件内部注册，中间件会为其分配资源，脱离后将资源回收。通信中每个句柄也要显示地进行释放和阻塞操作。通过这种方式来确定句柄状态，保证在数据发送和接收过程中句柄对应的通信通道是可用的。

2) 基于通信句柄的信息映射机制。

句柄信息映射机制如图5所示，该机制目的是为

了屏蔽中间件底层TCP和组播通信细节，通过配置文件在共享内存中建立的两张表(物理通信网络环境表和通信句柄表)来实现的。物理通信网络环境表保存了实现具体网络通信需要的套接字及其相关信息，其对应于系统的物理通信信息层。通信句柄表保存了实现不同业务而需要的抽象通信句柄及其通信属性，对应系统的逻辑通信信息层。这两层采用物理通信号进行关联，物理通信信号和通信句柄是一对多的映射关系，即多个通信句柄可使用同一个套接字来发送数据。

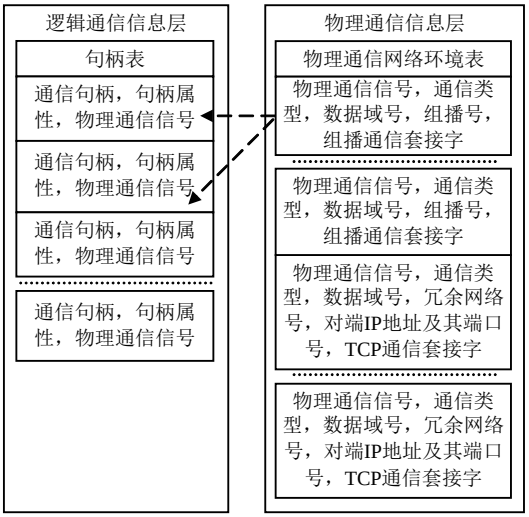

图5　句柄信息映射

1) 数据发送映射过程。

上层应用程序通过挂接模块发送的数据只包含了通信句柄和数据优先级信息。系统先根据通信句柄在逻辑通信信息层中查找到相应的句柄属性及其物理通信信号，然后在物理通信信息层中通过物理通信信号找到对应的套接字，组成一个完整的数据包插入跨进程的优先通信队列中，由发送模块取出队列中的数据包，根据其中的套接字进行发送。

2) 数据接收映射过程。

数据接收模块接收到数据信息后，在模块内直接获取数据包自带的句柄属性，然后直接到逻辑通信信息层查找和其属性匹配的本地通信句柄，并在模块注册表内确认使用该本地通信句柄的模块已经运行，之后调整数据包并把已匹配的本地接收句柄写入数据包内，最后把该数据包插入跨进程的数据接收队列，由挂接模块从队列中取出数据，送入对应的模块接收队列内。

2.2　在线扩展和在线维护特性关键实现技术

中间件的在线扩展和在线维护特性的实现依赖于图6所示的广义混合队列模型，其采用了多消息分发器、单主队列和二级队列多种形式来处理各种消息，并以此为基础形成了一个具备良好在线性能的分散系统。

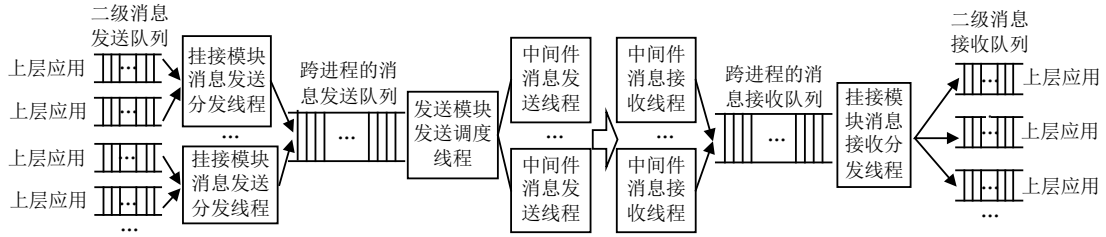

图6　中间件队列模型

如图6所示，每个上层应用都有一个发送队列和接收队列，一个挂接模块的消息发送分发线程负责处理多个上层应用(缺省为2个)的二级消息发送队列，多个分发线程的联合工作构成了挂接模块的消息发送分发处理机制。该机制可灵活地处理上层应用发送海量数据时的情况，不会造成二级消息队列的溢出。发送模块由TCP发送线程池和组播发送线程池构成，并管理这两个线程池的发送调度线程。该调度线程负责从消息队列中取出数据，然后唤醒相应的发送线程发送消息。在中间件的接收方，接收模块内部的各个消息接收线程收到消息后，立刻将消息放入消息接收队列内部。挂接模块端由一个独立的消息接收分发线程负责从队列获取数据，然后分发到各个二级消息接收队列。该模型的核心是在入队列一端采用各个消息分发器，这样有利于消息的及时响应，而在出队列的一端采用独立的队列处理线程。该方法比采用多个队列处理线程效率更高，因为出队列的数据所对应的上层应用是随机的，独立处理线程避免了不同消息的重复判断过程。

2.3　在线容错特性关键技术

1) 三阶段生存信号确认算法。

生存信号是中间件系统各个节点间相互协调的基本信息源，其可靠性对于整个系统而言是非常重要的。为保证生存信号的可靠性，系统提出了基于三阶段确认的生存信号确认算法，如图7所示。所谓三阶段就是分三种完全不同的方式确认某个节点的

生存状态，只有三阶段都判定节点死亡，生存信号模块才判定节点死亡，并且通知上层感兴趣应用程序模块。具体过程如下：

第一阶段：每隔一个固定时间间隔检查是否收到某节点发送的生存信号。假如连续*N*(缺省值为4)次没有收到，就初步判定该节点已经死亡。这种由被测节点主动发送生存信号的方法为“推”。

第二阶段：当“推”法判断为节点死亡以后，就由监测节点主动向被测节点发送询问消息，被测节点收到后必须立刻回复，假如在连续*M*(缺省值为3)次询问以后，监测节点等待的结果都是超时，则此时进入第三阶段。这种由监测节点主动询问的方法为“拉”。

第三阶段：该阶段由生存信号模块主动向被测节点的相邻节点寻问，要求其返回对被测节点的监测结果。假如其他节点的回应确认该节点在第二阶段就已死亡，生存信号模块就更新本地节点内的被测节点状态为死亡，并且通知上层感兴趣的应用程序模块。假如该过程中只要有一个节点返回被监测节点在第一或第二阶段有生存信号的信息，生存信号模块就更新本地节点内的被测节点状态为繁忙。假如中间件检测到网络环境中只有两个节点，就跳过第三阶段。这种主动询问其他相邻节点被测节点状态的方法为“问”。

另外，在没有发生网络故障的前提下，中间件发现除了自己以外已经没有其他节点处于生存状态，就立刻向上层感兴趣的节点发送节点关闭消息。

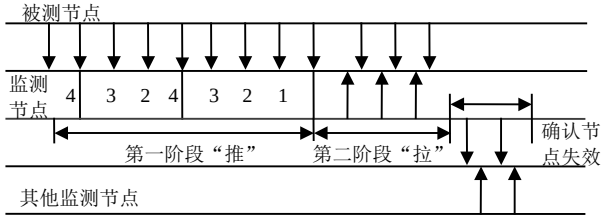

图7 节点生存信号确认过程

2) 基于发布定购模型的可靠组播通信算法。

发布定购模型是一个在空间和时间上都松散的异步通信模型。空间上是松耦合的是因为使用该模型发送方不需要指定接收方的地址；在时间上的松耦合是因为发送方和接收方不需要同时处于活跃状态。采用该模型中间件可实现良好的动态扩展特性，并且在该模型上巧妙实现了可靠的组播通信算法。

图8为可靠的组播通信算法。图中，发送方首先通过组播方式对外发送当前将要发送的事务代码TCD组。由于具体的事务代码所对应的事务信息可通过配置文件获取，所以接收方可判断这些TCD是否是上层应用需要的TCD。假如接收方判断为需要，就向发送方“定购”自己感兴趣的TCD组。接收方一直发送定购信息直到接收到从发送方发送来的消息编号组信息。然后接收方就依据消息编号，向发送方要求发送编号消息，当接收完毕以后就再次要求发送新的编号组信息，该过程如图8所示的3)、4)、5)、6)过程，接收方和发送方持续该过程一直到发送方通知数据发送完毕为止。

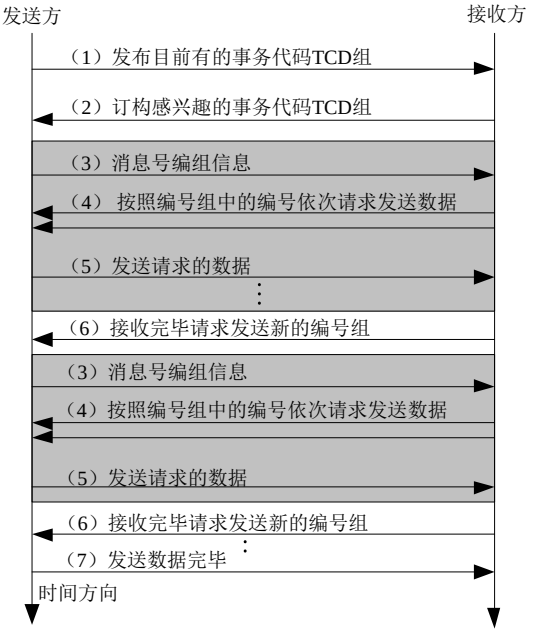

图8 可靠的组播通信算法

3 中间件特性试验分析

为验证该基于自律分散理论中间件的有效性，构建了由10台计算机构成的百兆带宽分散系统。其各项试验结果如表1、表2所示。

表1 基于句柄的通信试验结果

试验内容	试验描述	试验结果
一对一的发送接收通信句柄	用一个TCD发送句柄及其接收句柄进行大型文件传输	大型文件传输成功
一对多的通信句柄	用1个TCD对多个节点广播数据	对此TCD感兴趣的节点(模块在本地注册TCD)均可接收到此数据。
多对一的接收句柄	不断注册或删除感兴趣TCD	接收方可根据上层应用需求的TCD，接收感兴趣的数据

基于句柄的通信试验每项持续时间为60 min，从表1试验结果可见通过该中间件提供的编程模型，可大幅简化上层应用系统的研制难度，容易实现基于数据驱动的系统设计。在线性能试验每项分别试

验了50次，从表2试验结果可见该中间件已具备了自律分散系统优越的在线扩展、在线维护和在线容错特性。

表2　在线性能试验结果

试验内容	试验描述	试验结果
在线扩展	多个节点加入分散系统	系统很快(平均10 s以内)感知到新节点并融合为一个更大的分散系统
在线维护	多个节点进入维护状态，维护结束后正常运行	系统很快(平均1 s以内)感知各个进入维护状态的节点，并约束其它节点对维护状态节点的访问，维护结束后，系统很快感知并结束约束。
在线容错	多个节点，网络异常，时断时续。	系统很快(平均10 s以内)感知节点异常，分散系统随网络状态进入伸缩状态。

4　总　结

目前，该中间件已应用于某新一代基于Linux的AirNet ATC系统的设计中，改变了传统ATC系统中雷达数据全部直接采用UDP进行编程，网络数据流混乱且不可靠的局面，使系统中的网络数据流程变得非常清晰，提高了通信可靠性。在AirNet ATC系统中让中间件同时承受16部雷道发送的海量数据，系统可稳定运行72 h以上。由于中间件具有在线可扩展性、在线可容错性、在线可维护性，从而使得在中间件上构筑的AirNet ATC系统具有比国内同类系统更加灵活和可靠的特性，并且降低了系统维护和升级的成本。

当前自律分散理论在我国交通行业的应用越来越广，中间件的研究与实现对该理论的应用做了有益的探索。

参　考　文　献

[1] INJI M, HIROKAZU I, KAWANO K, et al. Autonomous decentralized software structure and its application[C]//Proc of FJCC'86. Dallas, USA: IEEE Computer Society, 1986: 1056-1063.

[2] INJI M. Autonomous decentralized systems: concept, data field and architecture and future trends[C]//Proc of 1993 the 1st int Symp on Autonomous Decentralized Systems. Kawasaki, Japan: IEEE Computer Society, 1993: 28-34.

[3] INJI M. Applications in rapidly changing environments[J]. IEEE Computer, 1998, 31(4): 42-43.

[4] INJI M. Heterogeneous autonomy under evolutionary situation[C]//Proc of 1999 the 4th int Symp on Autonomous Decentralized Systems. Tokyo, Japan: IEEE Computer Society, 1999: 406.

[5] Distributed Manufacturing Architecture Committee Japan FA Open Systems Promotion Group, MSTC/JOP. Specifications for autonomous decentralized protocol R 3.0[S]. Tokyo，Japan: 1999.

[6] UMIO K, IWAMOTO T, KIKUCHI K, et al. Widely-distributed train-traffic computer control system and its step by step construction[C]//Prof of 1995 the 2nd Int Symp on Autonomous Decentralized Systems. Phoenix, USA: IEEE Computer Society, 1995: 93-102.

[7] UMIO K, KAMIJOU K, KAKURAI Y, et al. Phased-in construction method of ATOS[C]//Prof of 1999 the 4th Int Symp on Autonomous Decentralized Systems. Tokyo, Japan: IEEE Computer Society, 1999: 415-424.

[8] AZUO K, EISUKE I, SHINICHI K, et al. Hitachi's initiatives in addressing the challenges of 21st century railway systems[J]. Hitachi Review, 1999, 48(3): 126-133.

[9] AZUO K, EISUKE I, SHINICHI K. Assurance technology for growing system and its application to Tokyo metropolitan railway network[J]. IEICE Transactions on Information and Systems, 2001, E84-D(10): 1341-1349.

[10] AN Yong-dong, GUI Xun, QIAN Qing-quan. Key technologies for decentralized control platform with dynamic evolution towar ds railway transportation system [J]. Journal of Southwest Jiaotong University, 2006, 14(3): 301-307.

[11] 孟辉, 廖建新, 王纯, 等. 移动智能网中消息中间件的性能建模与分析[J]. 北京邮电大学学报, 2006, 29(3): 76-80.
YANG Meng-hui, LIAO Jian-xin, WANG Chun, et al. Performance modeling and analysis of message middleware in mobile intelligent network[J]. Journal of Beijing University of Posts and Telecommunications, 2006, 29(3): 76-80.

[12] CHMIDT D C. The adaptive communication environment (ACE)[EB/OL]. [2010-06-10]. http://www.cse.wustl.edu/~schmidt/ACE.html.

编　辑　漆　蓉