b

YD

中华人民共和国信息产业部 发布

××××-××-××实施

××××-××-××发布

**确定性IP网络的总体架构与技术要求**

**The Overall Framework and Technical Requirements**

**of Deterministic IP Network**

（讨论稿）

YD/T ××××—××××

中华人民共和国通信行业标准

ICS

M

目 次

[目 次 I](#_Toc9238327)

[前 言 II](#_Toc9238328)

[**确定性IP网络的总体架构与技术要求** 3](#_Toc9238329)

[1 范围 3](#_Toc9238330)

[2 规范性引用文件 3](#_Toc9238331)

[3 术语、定义和缩略语 3](#_Toc9238332)

[3.1 术语和定义 3](#_Toc9238333)

[3.2 缩略语 3](#_Toc9238334)

[4 确定性网络目标 4](#_Toc9238335)

[5 概述 4](#_Toc9238336)

[5.1 确定性时延 5](#_Toc9238337)

[5.2 应用需求 6](#_Toc9238338)

[5.3 挑战 7](#_Toc9238339)

[6 广域确定性网络框架 8](#_Toc9238340)

[6.1 概述 8](#_Toc9238341)

[6.2 用户网络接口 8](#_Toc9238342)

[6.3 资源预留信令 8](#_Toc9238343)

[6.4 确定性转发机制 9](#_Toc9238344)

[6.5 OAM工具集 12](#_Toc9238345)

1. 前 言

本标准按照GB/T1.1-2009给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由中国通信标准化协会提出并归口。

本标准起草单位： 华为技术有限公司，中国移动通信集团有限公司，中国电信通信集团有限公司，新华三技术有限公司，上海诺基亚贝尔有限公司，中兴通讯股份有限公司，中国联合网络通信集团有限公司

本标准主要起草人：强鹂，刘冰洋，谭佳瑶，耿亮，雷波，谢云鹏，万晓兰，陈端，宋雪雁，刘畅

**确定性IP网络的总体架构与技术要求**

* 1. 范围

本标准规定了确定性IP网络的基本概念、应用场景需求及挑战、设计目标、总体架构和功能模块，以及相应的技术要求。

本标准适用于确定性IP网络。

* 1. 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

* 1. 术语、定义和缩略语
     1. 术语和定义

下列术语和定义适用于本文件。

* + - 1. 确定性时延（Deterministic Delay）

指有界的时延及时延抖动。

* + - 1. 确定性IP网络（Deterministic IP Network）

确定性IP网络是指具有以下三个特征的单域网络：1）大量的网络设备，2）两个网络设备之间的距离很长，3）网络中有很多确定性数据流。确定性IP网络能够提供以极低的丢包率和有限的端到端传送时延来传送数据流的能力。

* + - 1. 确定性流（Deterministic Flow）

确定性流是对时延上界及抖动上界有要求的用户流量。

* + - 1. 确定性节点（Deterministic Node）

能够提供确定性服务的网络节点，包括确定性边缘节点、确定性传输节点。

* + 1. 缩略语

下列缩略语适用于本文件。

|  |  |  |
| --- | --- | --- |
| ABRW | Aggregated bandwidth reservation window | 聚合资源预留窗口 |
| DSCP | Differentiated Services Code Point | 区分服务编码点 |
| IEEE | The Institute of Electrical and Electronics Engineers | 电气电子工程学会 |
| IPv6 | Internet Protocol version 6 | 第六版互联网协议 |
| LAN | Local Area Network | 局域网络 |
| MPLS | Multi-Protocol Label Switching | 多协议标签交换技术 |
| NP | Network Processor | 网络处理器 |
| OAM | Operation,Administration and Maintenance | 操作、管理和维护 |
| PLC | Programmable Logical Controller | 可编程逻辑控制器 |
| QoS | Quality of Service | 服务质量 |
| RTT | Round Trip Time | 往返时间 |
| SDF | Scalable Deterministic Forwarding | 可聚合的确定性转发 |
| SR | Segment Routing | 区域路由(协议) |
| SRH | Segment Routing Header | 区域路由头 |
| SRR | Scalable Resource Reservation | 可聚合的资源预留 |
| SRv6 | IPv6 based Segment Routing | 基于IPv6的区域路由(协议) |
| TC | Traffic Class | 流量类型 |
| TSN | Time Sensitive Network | 时间敏感网络 |
| UDP | User Datagram Protocol | 用户报文协议 |
| UNI | User-Network Interface | 用户网络接口 |
| VR | Virtual Reality | 虚拟现实 |
| H | the number of hops | 跳数 |
| K | the size of aggregated resource reservation window | 集成资源预留窗口大小 |
| T | the length of a cycle | 周期长度 |

* 1. 确定性IP网络应用场景需求及挑战
     1. 确定性时延概述

图1曲线表示网络时延的概率分布，确定性时延与低时延的定义不同，低时延是降低曲线的上限值，而确定性时延是把曲线变窄，即缩小抖动范围。传统IP网络由于存在时延长尾效应而无法实现确定性时延。技术上来说，确定性时延指的是低抖动。

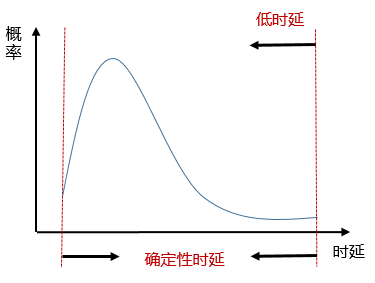


图1 低时延与确定性时延

虽然低时延和确定性时延的定义不同，但是它们并不是对立的。图2显示的是单跳时延的组成部分，主要包括链路时延和节点内时延。

链路时延是指数据包在链路上传输的时延，主要受链路长度和该链路的传输速率的影响。通常情况下，当网络设备部署后，它们的位置基本不会发生变化，即链路长度是固定的。同时，链路的传输速率是依赖于线缆媒介的，也是不变的。所以链路时延相对比较稳定，除非通过减少距离（比如部署边缘计算节点）或者线缆升级来降低时延。

节点内时延指的是设备内部操作造成的时间消耗，比如排队、NP处理等。节点内部时延变动比较大，图1所示的长尾效应主要就是由于节点内时延产生的。

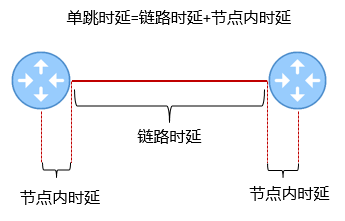


图2 单跳时延分析

因为链路时延基本稳定，所以降低时延的主要方式就是减少节点内时延，如图3所示。当节点内时延降低到一定水平时，意味着总时延变动很小，即压低最差时延上界使其无限逼近最小时延。所以，降低节点内时延能够同时获得低时延和确定性时延。

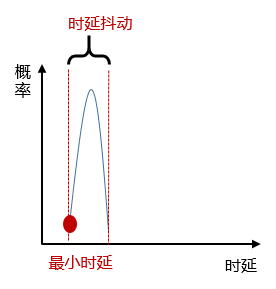


图3 降低节点内部时延能够同时获得低时延和确定性时延

* + 1. 应用场景及需求

确定性时延是协同或调度系统的前提，每个功能模块都应该在正确的时刻执行正确的动作。随着业务的日趋复杂，例如股票交易、电网继电保护、在线游戏、云PLC、远程医疗等场景，网络不仅要提供“及时”服务，还要“准时”。

以下简单分析了几个场景对确定性时延的需求：

* 远程医疗：大多数手术都需要多名医生和护士的配合。如果这些医生和护士都在远程操作，那么不同医生或护士发出的命令应该按顺序到达患者，不能太快或者太慢。 医生的操作指令通过前向链路传送到工作台，患者的画面及体征数据通过后向链路传送给医生。在远程医疗场景中，为了确保能顺利完成手术，端到端时延要求小于50毫秒，时延抖动小于200微秒。

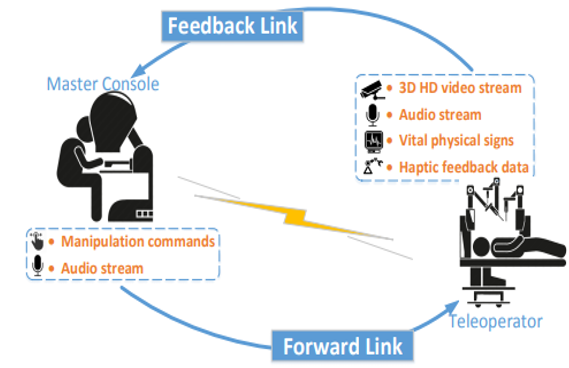


图4 远程医疗场景

* 在线游戏：为了公平起见，各个玩家同时发出的操作命令应该同时到达服务器并同时进行处理。这里的“同时”指的是相同的时隙，网络提供越快的传输和越低的抖动，服务器划分越细粒度的时隙，玩家越能体验到更顺畅的操作。
* 电网继电保护：在电源线的两端分别放置一个继电保护装置，它们向对端发送等量电流，并依据本地电流与从对端接收到的电流进行比较。如果两个电流的差异小于阈值，说明线路上没有发生故障；否则，说明线路某处存在故障。由于两端继电保护设备之间没有时间同步，所以需要利用RTT/2的时间来衡量A->B和B->A的单向时延。为了确保继电保护系统的准确性，两个方向之间的单向时延差应小于200微秒，抖动应小于50微秒[IEC 61850]。

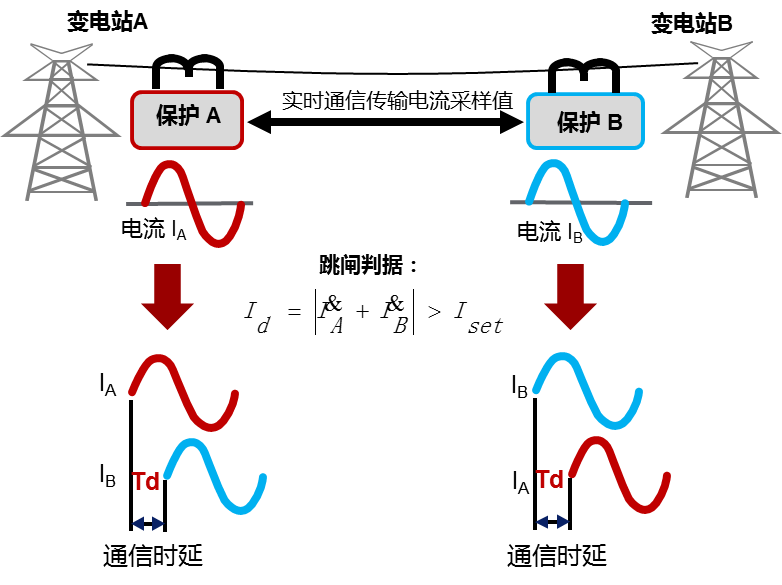


图5 继电保护场景

* + 1. 确定性IP网络挑战
       1. 概述

确定性IP网络是指具有以下三个特征的单域网络：1）大量的网络设备，2）两个网络设备之间的距离很长，3）网络中有很多确定性数据流。

这些特征将会带来很多挑战：1）端到端时延偏差容忍，2）新服务创建后的快速收敛，3）细粒度可扩展的资源预留，4）长链路传输时延容忍。

* + - 1. 端到端时间偏差容忍

IEEE TSN发布了一系列适用于LAN的确定性的标准。这些标准大部分都需要LAN内所有设备之间进行时间同步。由于大规模网络中存在大量异构设备，所有设备之间保持精确的时间同步很难而且成本昂贵，所以TSN技术无法直接应用于大规模网络。即使在非时间同步场景下，确定性IP网络架构也应该能够提供有界的延时。

* + - 1. 当新服务创建时快速收敛

VR通信等应用场景可能需要非常频繁地建立或断开确定性通信连接。在添加新的数据流时需要进行全局重新计算的机制，诸如IEEE 802.1 Qbv，不适合用于大规模网络。

* + - 1. 细粒度可扩展的资源预留方法

为了保证确定性流的QoS，网络必须为每个确定性流进行资源预留。路径上的所有设备应该为每个独立的确定性流维护资源预留状态。由于在大规模网络上存在大量的确定性流，这种资源预留方法是不可扩展的。

* + - 1. 容忍长链路传输时延

IEEE 802.1 Qch提出了一种典型且有效的周期转发机制，能够使端到端抖动小于2\*T（T为一个周期）。图6展示了IEEE 802.1 Qch的方法。节点A是节点B的上游节点，节点A在周期x发送的数据包会在同一周期内被节点B接收。一个周期应该包括链路传输时延。考虑到端到端抖动的上限是2\*T，所以周期T的长度不能太大。相应地，链路传播时延不能太大。这种限制对大规模网络中的长距链路并不友好。

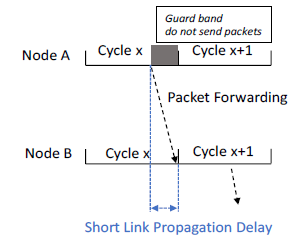


图 6 IEEE 802.1 Qch

* 1. 确定性IP网络目标

确定性IP网络旨在加强现有IP能力使其提供确定性服务，即保障传输时延上界及时延抖动上界。

* 1. 确定性IP网络总体架构
     1. 概述

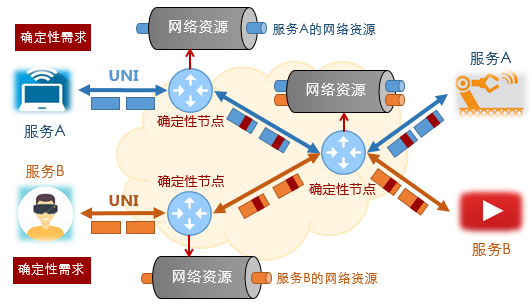


图7 确定性IP网络总体架构

图7显示了确定性IP网络的总体架构， 主要包括四个关键的功能组件：1）用户网络接口，2）资源预留信令，3）确定性转发机制，4）OAM工具集。

* + 1. 用户网络接口

用户通过UNI表达对网络的需求。至少有两种实现UNI的方式：

1）通过修改协议栈，让应用程序自己表达确定性需求；

2）通过用户端和网络端之间的代理设备来表达确定性需求。

* + 1. 资源预留信令

网络接收到用户的确定性需求后，要为这些确定性流进行相应地资源预留。目前有三种进行资源预留的方式：1）带内信令，2）带外信令，3）预配置。

* + 1. 确定性转发机制
       1. 概述

图8显示了一个完整的确定性转发方案，包括转发面的可扩展确定性转发（Scalable Deterministic Forwarding,SDF）和控制面的可扩展资源预留（Scalable Resource Reservation,SRR）。

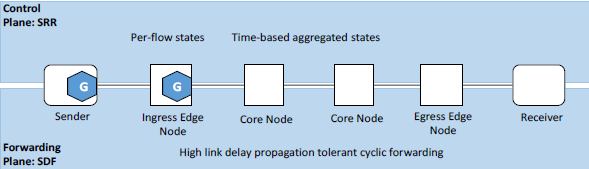


图8 确定性转发架构

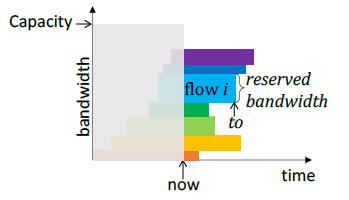
在SRR中，边缘节点记录每个流的资源预留状态，核心节点基于时间聚合所有流的状态。在SDF中，所有节点都是频率同步的，而且入口边缘节点或发送者都具有门控的功能，能够进行流量整形到特定模式。数据包是以周期队列和调度方式进行转发的。

* + - 1. 控制面可扩展资源预留

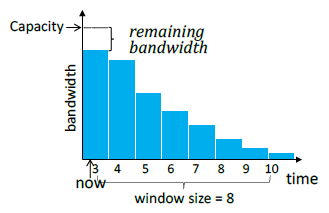
控制面应记录每个流的资源预留状态，包括保留多长时间的资源（如带宽）。每流状态至少应包含信息（flow\_ID，reserved\_bandwidth，start\_time，end\_time），如图9（a）所示。维持每个流的状态在边缘节点是可行的，因为每个边缘节点的流的数量不会太多。但是，当大量数据流汇聚到核心节点时，会导致核心节点负载过重。所以必须在核心节点进行聚合以实现可扩展性。

图9（b）显示了SRR中的聚合方法：

1. 将时间划分为时隙，每个流的状态为（flow\_ID，reserved\_bandwidth，start\_time\_slot，num\_time\_slot）。
2. 核心节点计算各个时隙中带宽预留的总和。对于同一时隙，核心节点只需要维护一个变量。假设核心节点可以维持K个时隙的状态，那么聚合带宽预留窗口（Aggregated Bandwidth Reserved Window,ABRW）的大小是K。
3. 仅当路径上有足够的资源时，新的资源预留请求才会成功。资源应以时隙为单位进行预留。即使数据流只需要使用1.7个时隙的资源，核心节点仍然要为它保留2个时隙的资源。一个数据流每次最多可以请求K个时隙的资源。如果需要超过K个时隙的资源，应在K个时隙到期之前发送续订请求。
4. 核心节点根据边缘节点维护的每流状态刷新其ABRW。请求者也可以积极拆除其资源预留。



(a) 边缘节点记录的每个流的资源预留状态



(b) 核心节点集成的带宽预留窗口

图9 可扩展资源预留

* + - 1. 转发面可扩展确定性转发

如图10所示，周期转发不感知每个流的状态，并且可以控制抖动的边界。在周期转发中，时间被划分成几个周期，数据包按周期发送。发送者的周期数和接收者的周期数是固定的，但周期内的确切时间是未知的。所以最好的情况是数据包在发送周期结束时发送并在接收周期开始时接收，而最坏的情况是数据包在发送周期开始时发送但在接收周期结束时接收。因此，端到端抖动的上限等于最坏情况的端到端时延减去最好情况的端到端时延，即2\*T.

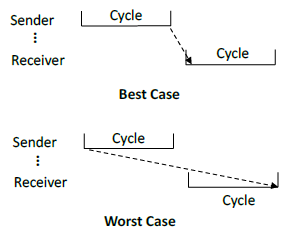


图10 周期转发

TSN-Qch是经典的周期转发机制之一，但是该方案需要所有节点之间进行时间同步，而且链路传播时延受限。这些特点都不适合应用于大网中。SDF在突破TSN-Qch的限制的同时继承了周期转发的优势。如图11所示，在SDF中所有节点都是频率同步的，并且允许长链路传播时延，端到端排队时延上界是2\*T\*跳数。

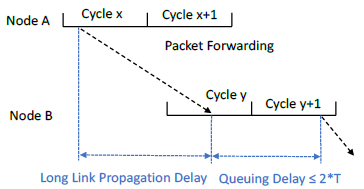


图11 可扩展确定性转发（SDF）

在不考虑非排队传播时延变化的情况下，每对相邻节点都具有稳定的周期映射关系，用于表示数据包的转发时间。如图12所示，节点A是节点B的上游节点，A和B之间的周期映射关系可以表示为周期x→周期y+1。也就是说，节点A在周期x发出的数据包将由节点B在周期y+1重新发送出去。由于它们之间没有时间同步，所以节点B可以接收两个周期的分组。因此，每个输出端口需要2个接收队列和1个发送队列。这样的话，每个数据包就需要2比特来携带周期标识，以此表示它应该进入的队列。

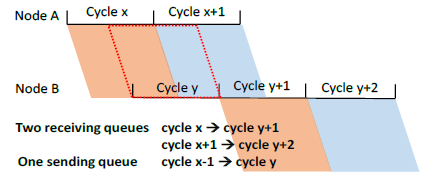


图12 SDF中的三个队列

有几种方式可以携带这2比特的周期标识：IPv4包头的区分服务字段（DSCP），IPv6包头的Traffic Class，MPLS头的Traffic Class，IPv6扩展头，UDP选项，SRv6的SID，SRH的Reserved Field，SRv6的TLV，SR-MPLS头的TC，等等。

* + 1. OAM工具集

细粒度的可视化OAM工具集应该提供给用户，以便用户验证网络是否确实提供了所请求的确定性服务。这种细粒度的OAM工具集不应该引入任何可扩展性问题。所以带内方法可能是最合适的方法。

* 1. 确定性IP网络技术要求
     1. 概述

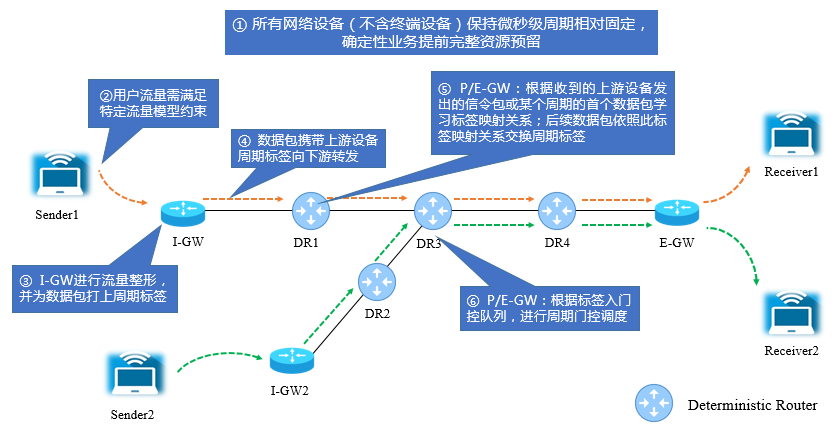


图13 确定性IP网络技术要求总览图

图13显示了确定性IP网络总体技术要求，主要包括发送端（Sender）、入口网关（Ingress Gateway，I-GW）、支持确定性转发能力的路由器设备（Deterministic Router，DR）、出口网关（Egress Gateway，E-GW）及接收端（Receiver），整体流程如下：

* 首先所有网络设备（不含终端设备）需要保持微秒级周期相对固定；
* 其次对每一条确定性业务均需预留沿途所有的带宽资源；
* 前两条完成之后，可以开始传送确定性业务的报文。用户报文的流量模型需要满足资源预留的约束；
* 入口网关需要对用户报文进行流量整形，并为报文打上初始周期标签，正式进入确定性IP网络周期转发流程；
* 数据包携带周期标签发往下游设备，当数据包抵达DR路由器后，路由器根据本地维护的周期映射关系表，替换数据包中的周期标签并将数据包送入相应的队列等待转发；
* 每个DR路由器及出口网关都维护着特定数量的确定性转发和发送队列，并对这些队列进行周期门控调度。
  + 1. 微秒级周期固定

设备将各自的时间划分为等长的时间周期，并为每个数据包安排特定的发送周期。如图14所示，网络中所有设备将各自的时间轴以T为单位划分为若干周期。不同设备的周期可以是从不同时间开始的，然而周期前沿的差距（例如D1、D2等）必须是维持稳定不变的，可以容许在微秒级的范围内抖动。



图14 设备周期固定

* + 1. 资源预留

每条确定性流在正式发送数据报文之前，都需要为其预留沿途的所有资源。资源预留的过程可以利用现有技术完成，例如通过SDN控制器进行集中算路及配置的方式，或者RSVP-TE的方式等。特殊之处在于，每条确定性流都有一个最小预留资源量。

最小预留资源量的目的是为了确保每条流在每个周期内都能发送至少一个数据报文。假设数据报文的大小为U（例如U=1000Bytes），时间周期的长度记为T（例如T=10us），则最小资源预留量即为U/T （例如U/T=1000Bytes/10us=10M）。所有确定性流的资源预留量都不得小于最小资源预留量。例如有一条流只需要1M的带宽资源，由于在此例中最小预留资源量为10M，则依然需要为这条流预留10M的带宽。

为某条确定性流所预留的资源，哪怕在某段时间没有被使用，也不得为其它的确定性流所用，但是这部分空闲资源可以用于转发BE报文。

* + 1. 边缘整形及标签初始化

当确定性流的数据报文到达入口网关I-GW之后，I-GW需要根据该确定性的资源预留情况进行整形。例如某条确定性流的资源预留量为10M，则意味着该流在每个周期内最多发送不超过1000Bytes的报文，I-GW需要记录该流在一个周期内已经发送了多少字节，如果已经达到了一个周期内所容许的字节上限，则需要对该流的数据报文进行缓存或者丢弃。



图15 标签初始化

同时，I-GW需要对收到的数据报文进行标签初始化。初始化过程如图15所示，I-GW对自己的周期进行循环编号，周期编号的取值范围为0-3。数据报文在某个周期发送出去之前，打上当前周期对应的编号。

* + 1. 周期映射关系学习

在确定性IP网络中，每对邻居路由器之间都有一个稳定的周期映射关系。该周期映射关系指导着后续路由器的数据包转发行为。周期映射关系的构建可以通过SDN控制器进行配置，也可以通过自适应的方式学习得到。



图16 周期映射关系自学习

图16解释了周期映射关系自学习的过程，上下游节点A与B均以0-3对各自的周期进行循环编号。上游节点A从某个周期（例如1号周期）的开始发送一个学习报文，学习报文经过链路传输到达下游节点B。B记录下收到第一个bit的时刻（假设该时刻落在节点B的3号周期）。B从收到第一个bit的时刻开始往后推一个周期长度T的时间，得到的时刻值意味着所有从上游节点A的1号周期内发出的数据包均已到达下游节点B。考虑到芯片及链路上可能存在一定的抖动，因此再往后累加一个抖动∆，最终得到一个时间点。该时间点对应着上游节点A最后一个1bit最晚到达的时间，在图8的例子中该时间点落在节点B的1号周期内。从该时间点开始寻找第一个完整周期，本例中对应的是2号周期，据此构建A与B之间的周期映射关系1→2。

该映射关系意味着上游节点A在1号周期内发出的数据报文在下游节点B的2号周期内完整的转发出去。同理，节点A在2号周期内发出的报文，节点B需要在本地的3号周期内转发出去，以此类推。根据此映射关系，数据报文在每一跳发送之前只需要携带发送设备的当前周期编号。下游节点收到数据报文后根据本地维护的周期映射关系表，即可确定收到的报文需要在本地哪个周期再次转发出去。

* + 1. 门控队列调度

确定性路由器中每个周期对应着一个队列。在7.5中我们已经提到，每个路由器用0-3对自己的周期进行循环编号。这就是说每个路由器对应着4个确定性接收和转发队列。图17解释了4个队列的由来，假设上下游节点之间的周期映射关系为0→3, 1→0, 2→1。由于确定性路由器之间的周期可以从不同的时间开始，因此下游节点的一个周期内可能会收到分属上游不同周期的数据包。例如图17中下游节点在周期3内会收到来自上游节点1号周期及2号周期内的数据包。根据上下游节点之间的周期映射关系，属于上游节点1号周期的数据包需要在下游节点的0号周期内转发出去，而属于上游节点2号周期的数据包需要在下游节点的1号周期内转发出去，因此下游节点需要2个接收队列来分别存储两个不同周期的数据包。同时下游节点在3号周期还需要转发来自上游节点0号周期的数据报文，因此还需要1个发送队列。由此可见下游节点需要至少3个DIP循环队列，其中2个接收队列1个发送队列。然而3个DIP队列是在不考虑任何芯片及链路抖动情况下的需求，为了吸收节点内芯片及节点间的抖动，还需要额外增加一个DIP队列。这就是每节点（准确的说是每个LSDN出接口）4个DIP队列的由来。通过如图18所述门控的方式循环调度这4个DIP队列。



图17 4个DIP队列



图18 门控循环调度