



专题：数据网络协议架构创新——NewIP

大规模确定性网络转发技术

强鹏, 刘冰洋, 于德雷, 王闯
(华为技术有限公司, 北京 100095)

摘要: 提出了一种适用于大规模网络部署的 3 层转发技术——LDN (large-scale deterministic network), 在保留传统 IP 转发技术统计复用的优势基础之上, LDN 技术可实现对端到端时延上界及抖动上界的严格保证, 为 5G uRLLC (ultra-reliable low-latency communication) 切片、工业互联网等未来应用场景提供网络服务支持。通过仿真实验对比了在相同网络环境下, 传统 IP 及确定性 IP 在端到端最差时延及抖动上的差异, 证明了 LDN 技术的有效性。

关键词: 确定性网络; 有界时延; 抖动

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019213

Large-scale deterministic network forwarding technology

QIANG Li, LIU Bingyang, YU Delei, WANG Chuang
Huawei Technologies Co., Ltd., Beijing 100095, China

Abstract: A layer 3 forwarding technology LDN (large-scale deterministic network) which was scalable for wide area deployment was proposed. LDN not only keeps the advantage of traditional IP's statistic multiplexing, but also be able to guarantee end-to-end bounded latency and bounded jitter required by 5G uRLLC (ultra-reliable low-latency communication) slice, industrial network, and other future scenarios. A simulation experiment was taken between traditional IP and deterministic IP (LDN enabled) on end-to-end worst case latency and delay variance (i.e., jitter), the comparison results successfully proved the effective and efficient of LDN technology.

Key words: deterministic network, bounded latency, jitter

1 引言

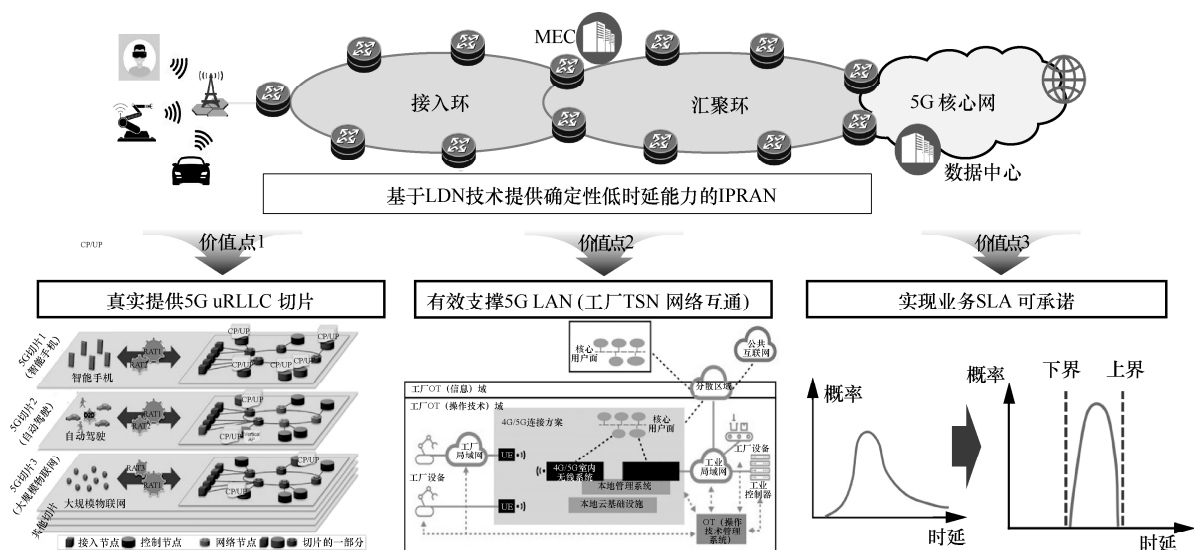
面对诸如工业互联网、远程医疗等时延敏感型业务^[1], 传统网络尽力而为的转发方式已无法满足此类业务对网络时延的需求, 确定性网络技

术受到越来越多的关注。本文提出了一种支持确定性 IP 服务的技术——大规模确定性网络 (large-scale deterministic network, LDN) 技术。LDN 技术旨在扩展传统 3 层转发技术, 以提供对传输时延上界及时延抖动上界有保障的确定性转

收稿日期: 2019-07-15; 修回日期: 2019-09-12

基金项目: 国家重点研发计划基金资助项目 (No.2018YFB1800100)

Foundation Item: The National Key Research and Development Program of China (No.2018YFB1800100)



■满足uRLLC业务需求, 实现端到端业务保障 ■使能更多业务场景, 进一步提升客户网络价值 ■企业专线场景下, 真正实现业务SLA的可承诺

图1 LDN用于5G承载场景

发技术。本文将有界的时延及时延抖动统称为确定性时延。

随着 3GPP Release 15^[2]的封版, 5G 技术面临着全面部署。其中网络切片, 尤其是 uRLLC (ultra-reliable low-latency communication, 超可靠低时延通信) 切片^[3], 对于时延的保证有严格的要求。uRLLC 切片要求不论在什么情况下 100% 地保证服务水平协议 (service level agreement, SLA), 即对于最差时延的上界是有保证的。LDN 技术可以有效支撑 5G uRLLC 业务, 如图 1 所示。此外 LDN 技术也可支持 5G LAN, 即利用 5G 网络连通工厂时延敏感网络 (time-sensitive network, TSN) 孤岛, 实现端到端精准控制。除 5G 承载之外, LDN 还可替代企业专线支持传统网络中的时延敏感型业务。

2 确定性 IP 技术背景

传统 IP 网络的时延概率分布曲线如图 2 所示, 在传统 IP 网络中由于存在时延长尾效应^[4]而无法实现确定性时延。为了消除长尾效应实现确定性时延, 首先需要分析时延的组成, 了解长尾

效应的产生原因。

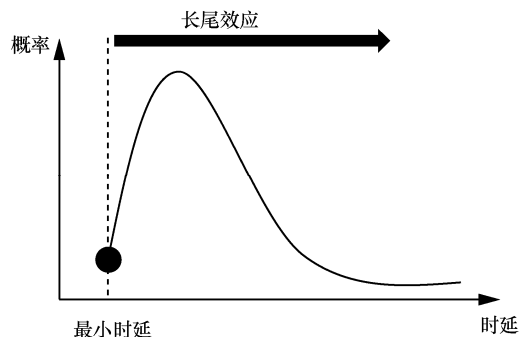


图2 传统IP网络的时延概率分布曲线

2.1 时延分析

以图 3 所示的单跳场景来分析时延的组成, 单跳时延=节点内时延+链路时延。其中, 链路时延是指数据分组在链路上传输的时延, 主要取决于网络设备之间的链路距离及链路传输速率。在一个稳定的网络拓扑中, 链路距离及链路速率相对稳定, 因此链路时延几乎没有什么变化的空间。节点内时延是指设备内部操作的耗时, 主要为排队时延。在不同负载等情况下, 节点内部时延的变动很大, 图 2 中的长尾效应主要就是由节点内部时延引起的。

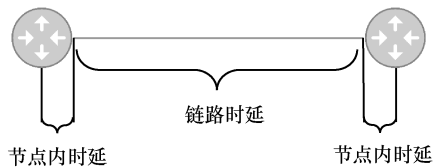


图3 单跳时延分析

基于以上分析可以知道一直以来通信行业追求的低时延与确定性时延的技术目标其实是一致的。

- 为了实现确定性时延，需要减少节点内时延以消除长尾效应。
- 为了实现低时延，由于链路时延在固定的网络拓扑下没有多少变动空间，因此能做的也是设法减少节点内时延。

确定性 IP 的技术目标如图 4 所示，即压低最差时延上界使其无限逼近最小时延。

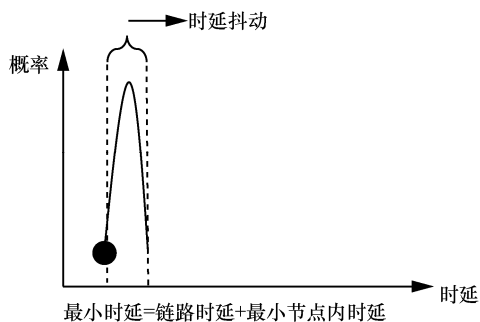


图4 DIP网络的时延概率分布曲线

2.2 现有技术分析

近几十年来有很多的技术探索去保障 IP 的服务质量（quality of service, QoS），比如优先级队列调度、资源预留等。可是这些技术依然无法实现确定性时延，其根本原因是微突发。以一个简单的例子来进行解释，如图 5 所示，假设节点 A 有 10 个入接口，1 个出接口，所有接口的速率都是一样的。假设每个入接口上都有一条流，并为此流预留了 5% 的接口带宽资源。假设 10 个接口的 10 条数据流都是同一优先级，10 条流在节点 A 的出接口汇聚时总共占用出接口最多 50% 的带宽资源，因此不存在拥塞的情况。然而由于对每个

数据分组的到达时间没有控制，因此最差的情况是 10 条流的数据同时到达，在出接口排队等待调度。排队的产生会挤压原本流内报文的间隙，形成微突发。假设下游节点 B 有 10 个像 A 一样的上游节点，则微突发的情况会进一步累积，形成微突发迭代。多跳之后，确定性时延将无从保证。

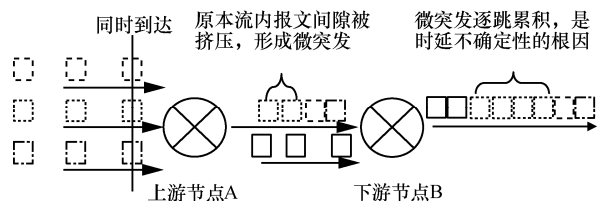


图5 微突发及突发迭代

图 5 中的例子说明，传统 IP 转发即使配合了资源预留及优先级调度，也无法实现确定性时延。其主要原因是由于对每个数据分组的行为缺乏控制，导致了排队形成了微突发。为此需要控制每个数据分组的行为，不能容许数据分组随意发送（接收），而且为每个数据分组分配一个特定的发送（接收）时间周期。通过控制每个数据分组的行为来避免冲突，从而控制节点内部排队时延，最终消除长尾效应实现确定性时延。

3 确定性 IP 的使能技术 LDN

3.1 基于 LDN 的承载解决方案

基于 LDN 的承载解决方案如图 6 所示，其中包括终端 CE、入口 PE 节点、支持 LDN 能力的 P 节点、控制器等，整体流程如下：

（1）所有网络设备（PE 节点及 P 节点）需要保持微秒级周期相对固定；

（2）对每个准许接入的确定性业务需要进行路径计算、资源预留等操作；

（3）（1）（2）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；

（4）入口网关需要对用户报文进行流量整形，

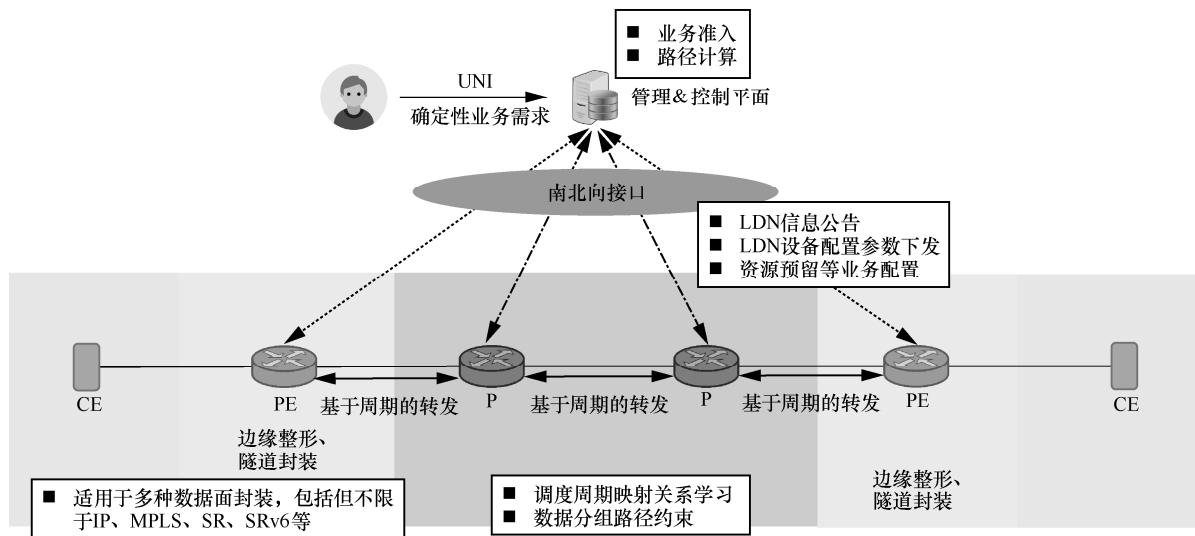


图 6 基于 LDN 的承载解决方案

并为报文打上初始周期标签，正式进入 LDN 周期转发流程；

(5) 数据分组携带周期标签发往下游设备，当数据分组抵达 LDN 路由器后，路由器根据本地维护的周期映射关系表，替换数据分组中的周期标签并将数据分组送入相应的队列等待转发；

(6) 每个 P 节点及 PE 节点都维护着特定数量的 LDN 队列，并对这些队列进行周期门控调度。

下面将对整体架构中涉及的技术细节进行描述。

3.2 等长周期划分

为了避免盲目转发而造成的微突发，LDN 技术需要控制每个报文在每一跳的转发行为。如果将网络设备视为一个黑盒子，那么报文经过一跳网络设备的行为即可描述为：报文进入该跳设备的时间及报文出该跳设备的时间。控制报文在每一跳的转发行为则相应地可以描述为：一旦给定报文进入某跳设备的时间，则该报文出该跳设备的时间即可确定。将报文进、出网络设备的时间控制到一个精确的时刻是很困难的，但是控制到一段时间区间要相对容易。基于这个思想，LDN 设备将各自的时间以 T 为单位划分为等长周期，并为每个数据报文合理安排进、出本跳的周期。

3.3 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 $U=1\ 250\ \text{byte}$)，时间周期的长度记为 T (例如 $T=10\ \mu\text{s}$)，则最小资源预留量为 U/T (例如 $U/T=1\ 250\ \text{byte}/10\ \mu\text{s}=1\ \text{Gbit/s}$)。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 $0.5\ \text{Gbit/s}$ 的带宽资源，由于在此例中最小预留资源量为 $1\ \text{Gbit/s}$ ，则依然需要为这条流预留 $1\ \text{Gbit/s}$ 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，也不得为其他的确定性流所用，然而这部分空闲资源却可以用于转发尽力而为 (best effort, BE) 报文。

3.4 周期映射关系学习

在 LDN 系统中，每对邻居 LDN 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

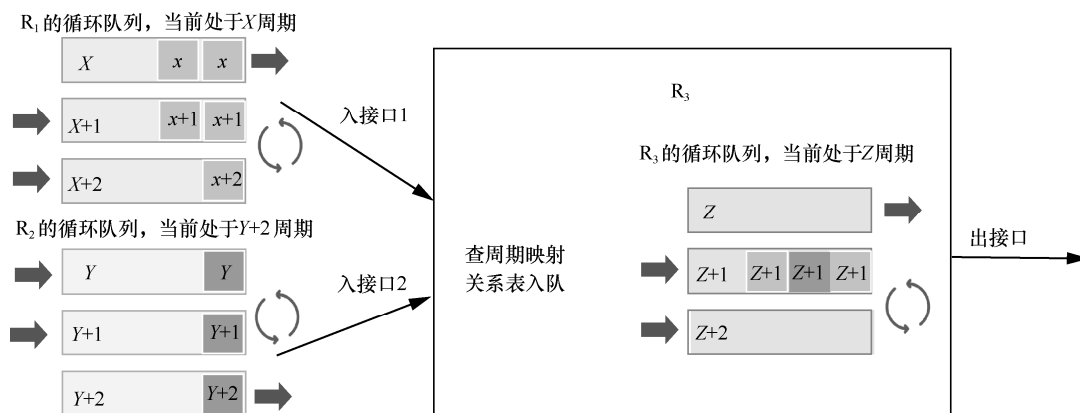


图7 循环队列调度

指导着后续 LDN 路由器的数据分组转发行为。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。通过学习报文获得周期映射关系之后，后续的数据分组均可以简单地据此映射关系转发。

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。

3.5 循环队列调度

如第3.2节所述，每个LDN路由器将自己的时间划分为等长的周期。LDN路由器对周期进行循环编号，每个周期对应着一个队列。在一个周期内只有其对应的队列会打开，存储于该队列中的数据报文得以发送，其余队列处于关闭的状态，只能用于接收数据分组。

图7给出了一个循环队列调度的例子， R_1 与 R_2 是 R_3 的两个上游节点。所有节点都有相同数量 Q 个循环队列，在本例中假设 Q 等于3。当下游节点 R_3 收到数据报文之后，读取报文上携带的周期编号信息，并据此查找周期映射关系表。随后 R_3 将报文中的周期信息替换为查表所得到的出标签信息（比如入周期标签 X 对应的出周期标签 $Z+1$ ），并将报文送入本跳相应的队列中等待发送。

4 理论分析

4.1 时延

首先分析一跳之内数据报文的时延，图8和图9分别给出了LDN的最差单跳时延及最小单跳时延分析。若上游节点最后1 bit到达下游节点的时间刚好超出了前一个周期一点，则出现最差情况。最差情况下数据报文需要在下游节点中等待约 $2T$ ，才能继续向下转发，如图8所示。若上游节点最后1 bit到达下游节点的时间刚好落在某一周期的后边沿，则在紧随其后的一个周期内即可向下游转发，此为最好情况，如图9所示。在最好情况下数据报文需要在下游节点中等待 T 的时间。

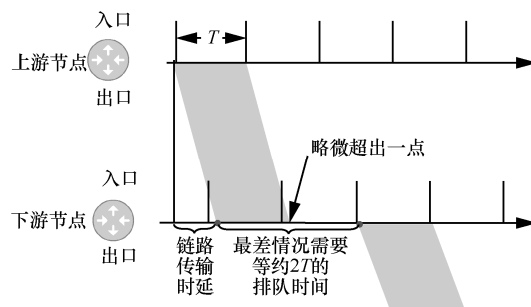


图8 LDN最差单跳时延

由上述分析可知LDN系统中，单跳的排队时间为 $T \sim 2T$ ，因此：

- LDN系统中端到端的最差时延=总链路传输时延+ $2T \times$ 跳数；

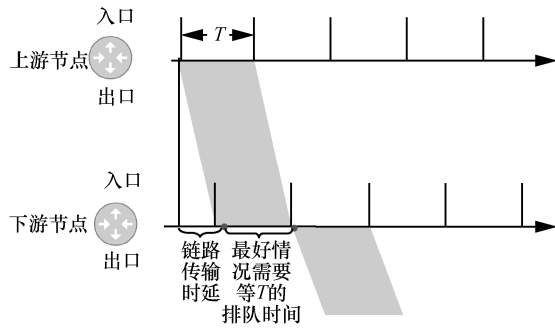


图9 LDN 最好单跳时延

- LDN 系统中端到端的最好时延=总链路传输时延+ $T \times$ 跳数。

4.2 抖动

由于每对邻居节点之间都有一个周期映射关系, 因此一旦数据报文在源点的发送周期确定了, 则该报文在目标点的接收周期也就可以确定下来。

唯一无法确定的是, 数据报文在发送周期的具体哪个时刻发送, 并在接收周期的具体哪个时刻接收。图 10 给出了最好及最差情况:

- 最好情况是数据报文在发送周期的最末尾才发出, 在接收周期的一开始就收到了;
- 最差情况是数据报文在发送周期的最开始就发出, 在接收周期的最末尾才收到。

则端到端时延抖动 $\leq$ 最差情况-最好情况 $=2T$ 。

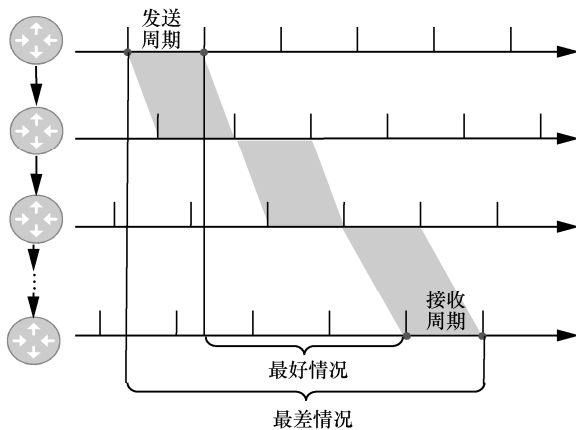


图10 LDN 端到端抖动

4.3 队列开销

由第 3.5 节可知每个 LDN 设备维护着若干

LDN 循环队列, 每个 LDN 队列对应一个周期, 用于存储一个周期内的数据报文。假设 LDN 出接口的接口速率为 S , 则每个 LDN 队列的存储开销为 $S \times T$ 。因此每个 LDN 节点只需要维护 $Q \times S \times T$ 大小的队列空间, 其中 Q 为一台设备内 LDN 循环队列的数目。

5 仿真实验

为了验证 LDN 技术的性能, 本文对比了传统 IP 与 DIP (LDN enabled) 在相同网络环境下的端到端最差时延及抖动。实验采用如图 11 所示的拓扑^[5], 共涉及 315 个网络节点。假设每个网络节点下挂一台主机, 且任意两台主机之间都可以建立一条业务流, 则最多有 $315 \times 314 = 98\,910$ 条流同时存在。考虑到实际情况中网络大多轻载, 因此实验假设网内仅有 5 000 条流共存, 其中 1 000 条目标待测流、4 000 条背景流。实验参数见表 1。实验开始阶段首先将图 11 所示拓扑导入 OMNest 中, 随后利用 INET Framework 安装协议栈完成路由发现, 实验对比结果如图 12 与图 13 所示。

假设目标待测流在 DIP 模式下以每周期 1 个报文的节奏进行转发, DIP 与传统 IP 在不同突发条件下的端到端最大抖动对比如图 12 所示。突发数据报文均一次性发送, 例如“100 个数据分组/ms”即意味着每 1 ms 一次性发送 100 个数据报文, “200 个数据分组/2 ms”即每 2 ms 一次性发送 200 个数据报文, 以此类推, 因此不同突发条件下的背景流速其实是不变的。由图 12 可知, 在 DIP 模式下端到端抖动始终小于 $20\, \mu\text{s}$ (即 $2T$), 且不随突发的加剧而改变, 而 IP 模式下的最差抖动随着突发的加剧而显著增加。在每 10 ms 有 1 000 个报文的突发情况下, IP 模式的最差抖动是 DIP 模式的最差抖动的近 300 倍。由此可见 DIP 可以有效降低数据报文 3 层转发的时延抖动。

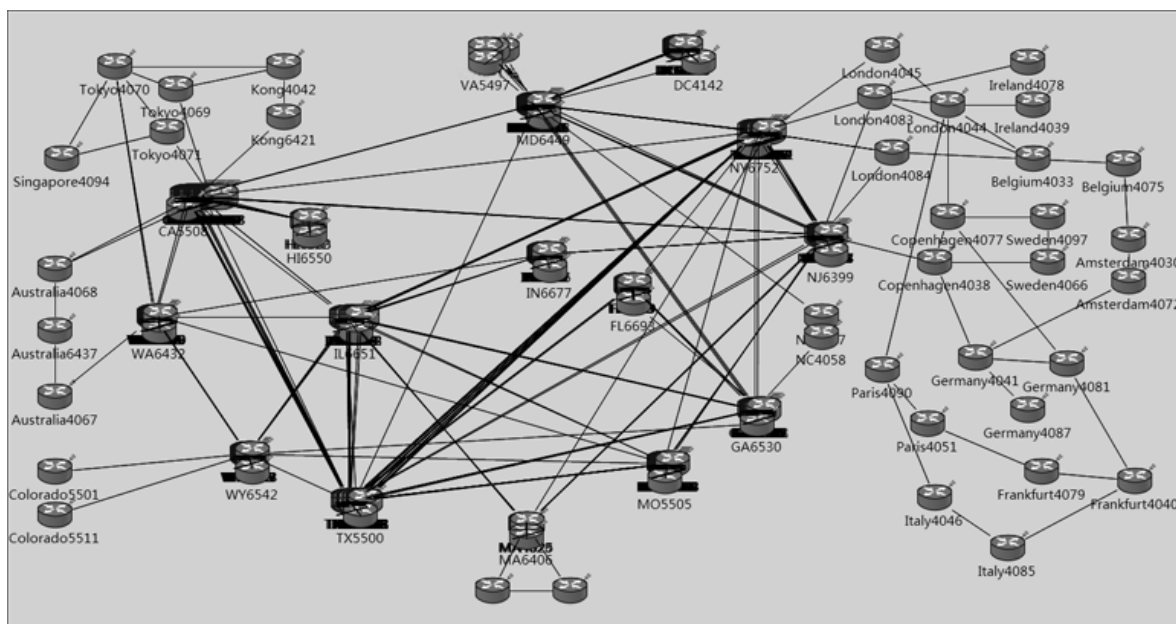


图 11 实验拓扑

表 1 实验参数

参数类型	参数值
周期长度 T	$10\ \mu\text{s}$
数据报文大小 P	$64\ \text{byte} \leq P \leq 1\ 500\ \text{byte}$
每周期发送报文数 N	1
链路速率	100 Gbit/s
实验环境	OMNest + INET Framework

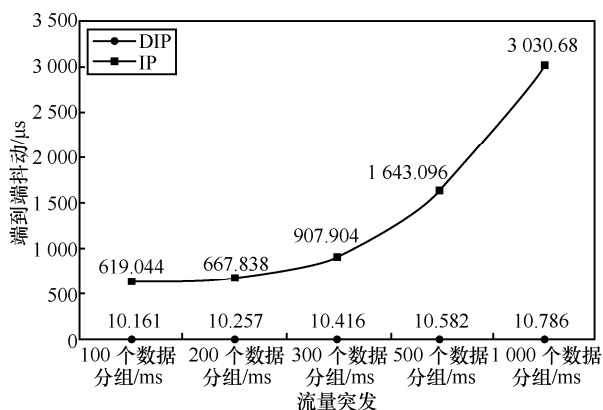


图 12 DIP 与 IP 在不同突发条件下的端到端抖动对比

DIP 与传统 IP 在不同突发条件下的端到端最差时延对比如图 13 所示。由图 13 可知，在 DIP 模式下的端到端最差时延几乎不受突发的影响，

始终维持在一个相对稳定的数值。而传统 IP 模式下的端到端最差时延随着突发的加剧而显著增加。在每 10 ms 有 1 000 个报文的突发情况下，IP 模式的最差时延是 DIP 模式的最差时延的近 30 倍。由此可见 DIP 可以有效地减少时延长尾效应，提供 SLA 保证。

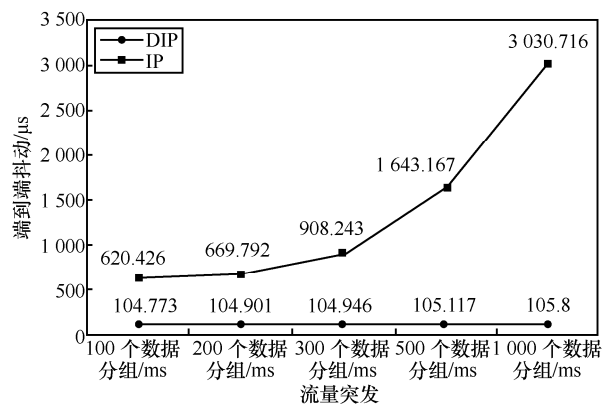


图 13 DIP 与 IP 在不同突发条件下的端到端最差时延对比

6 结束语

LDN 技术是确定性 IP 的主要使能技术。LDN 在基于传统 IP 的基础之上引入周期转发的思想，通过控制每个数据分组在每跳的转发时

机来减少微突发, 消除长尾效应, 最终实现端到端时延确定性。LDN 技术可以保证在最差情况下的端到端时延依然有界, 且最差时延与最好时延之间的差距仅为 $2T$ 。以 μs 级周期转发为例, 则最差时延无限逼近于最好时延, 两者之差仅为微秒级。

此外 LDN 技术中核心节点无逐流状态, 设备之间不需要精准时间同步, 因此具有良好的大网可扩展性。同时, LDN 技术的芯片开销低, 以 4 个循环队列为例, 对于 10GE 接口, $10\mu\text{s}$ 周期调度下每出接口仅需要 50 KB 的 LDN 队列资源。最后报文的开销较小, 数据报文中仅需要携带 2 bit 额外信息, 用于区分不同的循环队列即可。

综上, LDN 技术不仅可以实现对端到端时延及抖动的有界保障, 且适用于大规模部署, 是应对 5G 承载、工业互联网等未来场景时极具竞争力的一项技术。

参考文献:

- [1] IETF. Deterministic networking use cases: RFC 8578[S]. 2019
- [2] 3GPP. Summary of rel-15 work items: TR 21.915[S]. 2019.
- [3] POPOVSKI P, TRILLINGSGAARD K F, SIMEONE O, et al. 5G wireless network slicing for eMBB, uRLLC, and mMTC: a communication-theoretic view[J]. IEEE Access, 2018(6): 155765-55779.
- [4] TUCKER C, ZHANG J. Long tail or steep tail? a field investigation into how online popularity information affects the distribution of customer choices[J]. Working Papers, 2007, 6(1): 6.
- [5] MAHAJAN R, SPRING N, WETHERALL D, et al. Inferring

link weights using end-to-end measurements[C]//ACM Sigcomm Internet Measurement Workshop, Nov 6-8, 2002, Marseille, France. New York: ACM Press, 2002: 231-236.

[作者简介]



强鹏 (1988-), 女, 博士, 华为技术有限公司高级工程师, 主要研究方向为网络切片与确定性网络。



刘冰洋 (1985-), 男, 博士, 华为技术有限公司主任工程师, 主要研究方向包括网络架构、网络安全及可行、路由和命名解析、确定性网络等。



于德雷 (1977-), 男, 华为技术有限公司技术专家, 主要研究方向为未来网络架构、网络切片、确定性网络等。



王闯 (1975-), 男, 华为技术有限公司技术专家, 主要研究方向为未来网络架构与技术探索。