白皮书

数据平面开发套件（DPDK）
英特尔® 动态负载均衡加速器（英特尔® DLB）

使用英特尔® DLB技术实现高精度的网络限速

英特尔® 动态负载均衡加速器（Intel® Dynamic Load Balancer，英特尔® DLB）是一个硬件队列管理器和负载平衡器，通过数据平面开发套件（DPDK）的 Eventdev 抽象设备，从软件中卸载队列和调度任务。

目录

作者

王栋
软件应用优化工程师
英特尔

闭云峰
软件工程师
英特尔

Niall McDonnell
资深首席工程师
英特尔

郑春阳
云行业资深应用工程师
英特尔

徐恒阳
技术工程事业群网络研发总监
腾讯

方统浩
技术工程事业群网络架构师
腾讯

高文宾
技术工程事业群网络架构师
腾讯

庞玮
云架构首席工程师
腾讯

黎洁
云架构首席工程师
腾讯

引言

作为全球领先的云服务提供商之一，腾讯云*致力于向全球用户提供性能卓越的企业级网络服务。公有云对于服务质量有着严苛的要求，计算、内存、网络以及存储等各项资源的分配能否满足服务水平协议中所承诺的标准，都将直接影响最终用户的应用体验。对于云服务提供商来说，如何在充分利用以上资源，满足服务水平协议的前提下，尽可能减少额外资源开销，也是降低运营成本的关键因素之一。为在降低成本的同时保证优质的服务质量，腾讯云携手深度合作伙伴英特尔，基于腾讯云应用程序界面（Application Programming Interfaces, API）TGW 与腾讯专门的硬件工程实验室星星海实验室的创新软硬件结合方案，发挥 TGW 在网络领域的技术优势，针对网络资源调度及分配展开性能优化。

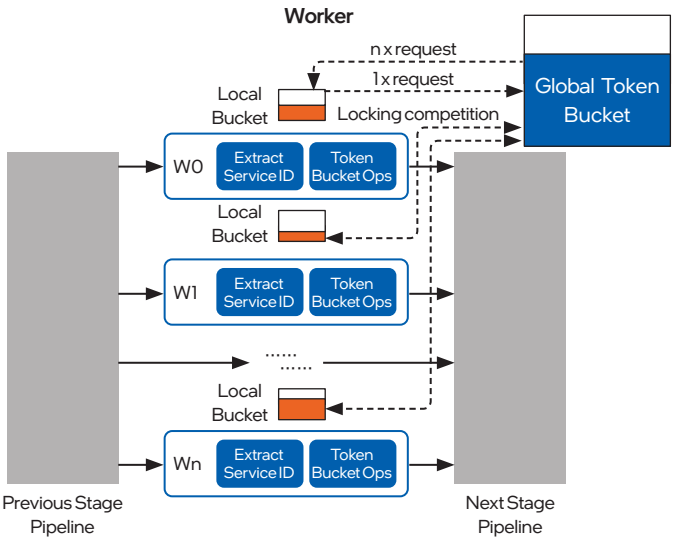
网络资源分配的常用方法是在网关对每个用户的带宽及并发控制和请求进行限速，以保护系统不会因为单位时间内的请求数量超载而造成拥塞。令牌桶算法是常见的限速机制之一，其工作原理是以一个设定的速率产生令牌并放入令牌桶，而每个用户请求都需要申请令牌，若令牌不足，则拒绝请求。但在多核处理器场景中，需要以原子的方式同步操作共享的令牌桶。因此，运行在多核处理器上的软件令牌桶方案，会使用“锁”对令牌桶加以保护。由于“锁”会概率性地降低转发性能，因此部分开发者使用了一种优化“锁”操作的方法，来降低“锁”对性能的影响。

另一种方法是使用网卡中类似 Flow Director 的技术，将属于同一个服务对象的网络数据包通过网卡分发到同一个核上，以此来消除“锁”。在这个方法中，每一个处理器核心的负载可能无法做到均衡，因为网络数据流中的服务对象的数量以及每个服务对象的网络流量会随着时间变化。当一个处理器核心过载时，报文因无法被及时接收而丢弃。因此，这种方法没有得到广泛使用。

本白皮书研究了在网络带宽限速中常用的针对令牌桶的“锁”的优化方案，分析了其在精度上的不足之处。随后，介绍了基于第四代英特尔® 至强® 可扩展处理器集成的英特尔® DLB 技术所实现的无锁令牌桶方案，以及其实现原理与对比测试数据，展现了无锁令牌桶方案相比现有优化方案所具备的优势。

轻量化锁限速方案

当多个处理器核心同时对一条网络数据流做限速时，可能存在多个核心同时对同一令牌桶加锁以使某个核心获得令牌桶的所有权，随之产生的“锁”竞争是导致性能下降的主要原因。开发者通过更改令牌桶的使用方式，配合一定的算法，降低“锁”竞争的概率，减少“锁”对性能的影响，这种方法称为轻量化锁。详见图一。



图一 轻量化锁限速方案

轻量化锁限速方案阐述

轻量化锁的限速方案由一个全局令牌桶，以及对应不同处理器核心的多个本地令牌桶组成。全局令牌桶根据设定的速率产生令牌，本地令牌桶以批量预取的方式从全局令牌桶获得令牌，令牌最终在本地桶被消费掉。

在令牌从产生到消耗的过程中，只有从全局桶到本地桶的预取操作需要加锁，且每次预取的数量会大于每个包实际消耗的令牌数。在处理同等数量的报文时，轻量化锁的方案对令牌桶加锁的次数明显低于传统的单一全局令牌桶方案。因此，随着处理器核心数量的增加，轻量化锁限速方案能够在一定程度上减少“锁”竞争，而获得较好的性能。

轻量化锁限速方案的局限性

轻量化锁限速方案包含两个关键参数:

一是全局令牌桶产生令牌的速率，即限速后的目标速率；

二是批量大小，当本地桶中令牌数量不足时，从全局桶预取令牌的数量。

全局令牌桶产生令牌的速率较低时，存在一种情况，即在单位时间内产生的令牌数无法满足所有本地令牌桶的批量预取请求。无法得到补充的本地令牌桶将因没有足够的令牌而导致报文被丢弃。然而，其他的本地令牌桶中却可能仍有未消耗的令牌，这些被丢弃的报文并没有超出限定的速率，导致限速后的速率低于目标速率。

以上原因会带来限速后的速率波动，让精度成为限速方案优化时必须关注的问题。详见图二。

基于英特尔® DLB 技术的无锁限速方案

■ 英特尔® DLB 技术介绍

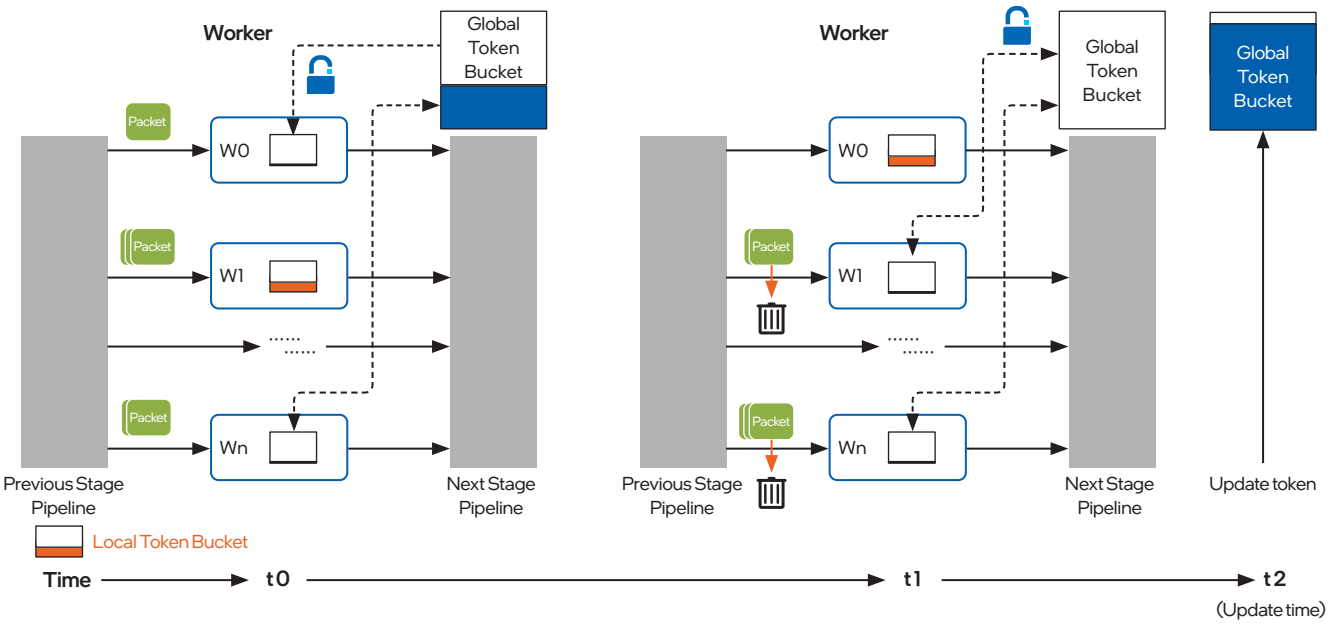
得益于技术的进步，每一代 CPU 的核心数量都较前一代有大幅提高。充分利用多核的优势，需要软件具有更好的并发度，而这给软件优化带来了巨大挑战。为此，第四代英特尔® 至强® 可扩展处理器中引入了英特尔® DLB 技术，可有效地解决高并发软件架构遇到的性能挑战。

英特尔® DLB 是集成在 CPU 内部的硬件队列管理器，软件通过入队、出队的方式与英特尔® DLB 进行交互。其中，入队方称为生产者，出队方称为消费者。

英特尔® DLB 有两个主要的特点，即动态与负载均衡。负载均衡要解决的是因为待处理数据在处理器核心之间分发不均匀，导致的处理器核心负载不均衡的问题。与一些软件方案所使用的静态调度算法不同，英特尔® DLB 在分发待处理数据的过程中，能够根据每个处理器核心的负载情况，动态地选出最合适的核心，并将数据分发给其进行处理。

为了实现动态特性，英特尔® DLB 设计了四种队列模型，来应对不同应用场景的需求：

- **Direct Queue:** 适用于多个生产者但只有一个消费者的场景，无负载均衡；
- **Unorder Queue:** 适用于多个生产者以及多个消费者的场景，不关心任务的先后顺序，将每个任务调度给当前负载最低的处理器核心去处理；
- **Order Queue:** 适用于多个生产者及多个消费者的场景，关心任务的先后顺序；当多个任务被多个处理器核心处理完时，需要按照原始顺序重新排列；



图二 轻量化锁限速方案引发限速速率波动

- **Atomic Queue:** 适用于多个生产者以及多个消费者的场景，任务按照一定的规则进行分组；处理这些任务时使用同一组资源，关心同一分组内的任务先后顺序。

详见图三。

基于英特尔® DLB 技术的无锁限速方案阐述

如上所述，现有的优化限速方案性能的方法，集中于降低“锁”的开销，也因此引入了精度问题。另外一种思路是使用无锁的限速方案，这种方案通过给网卡下发特定规则或是在软件中按照预定的算法，将同一条流的网络报文调度到同一个处理器核心，通过在同一个处理器核心上访问同一个令牌桶，实现无锁的限速方案。这些方案的问题在于报文的调度规则是静态的，无法根据处理器核心的负载情况做出动态调整，极易因网络突发流量导致部分处理器核心过载，进而产生丢包的情况。

是否存在一种方法，可以在多核处理器中，既能去掉保护全局令牌桶的“锁”，又能保证多核的负载均衡？

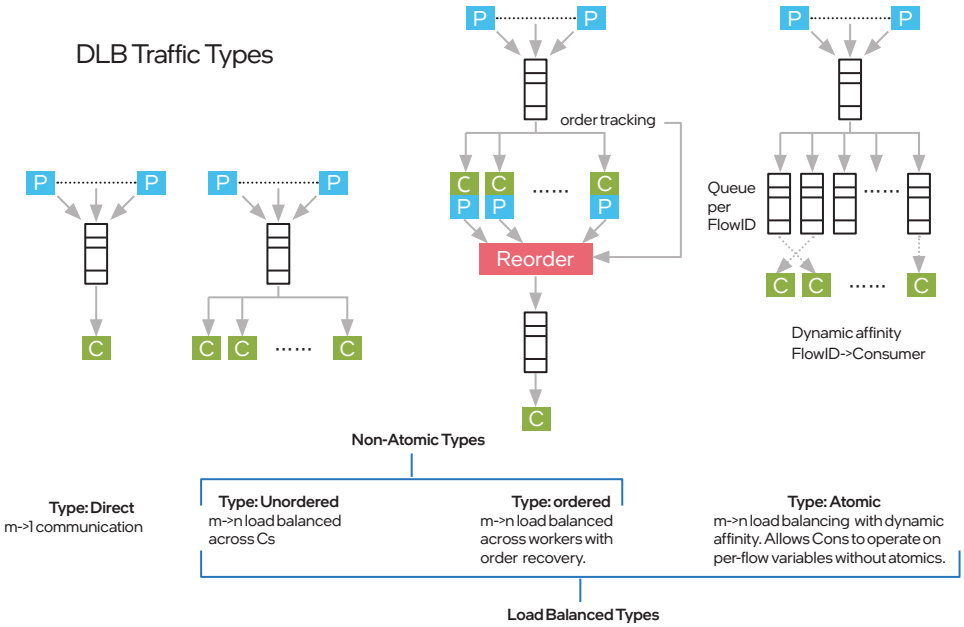
利用英特尔® DLB 的 Atomic Queue 特性，即可以在多核心的场景下实现无锁限速方案。

将待处理的网络报文按照其所属的限速网络数据流进行分组，英特尔® DLB 的 Atomic Queue 能够把属于同一分组的报文调度到同一个处理器核心进行处理；另外，Atomic Queue 还会为每一条流动态地选择处理器核心，当有多条网络数据流时，流量能够较为均匀地分散到各个处理器核心，确保处理器中多个核心的负载均衡。

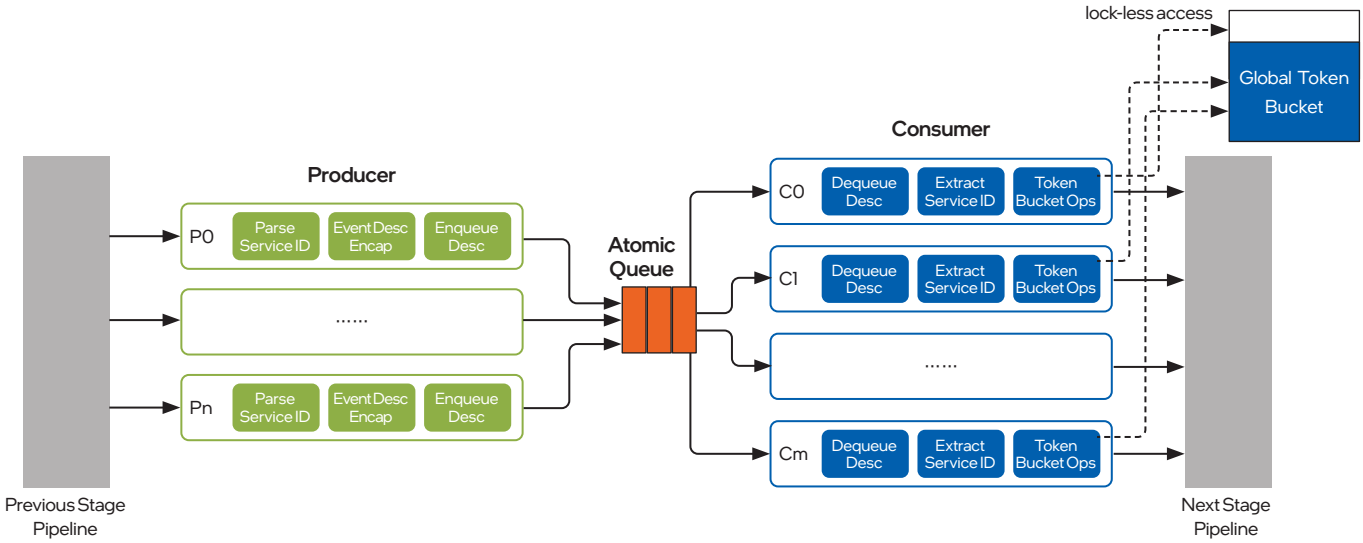
在无锁限速方案中，处理器核心被分成了两组，从队列操作的角度，分别被称为生产者和消费者。生产者为每个报文生成 Atomic Queue 所需的 Flow ID，随后将报文入队到 DLB 的 Atomic Queue 中。DLB 在消费者线程间分发消息，同时保证原子性。消费者从 Atomic Queue 获取报文之后，以无锁的方式安全地访问 Flow ID 对应的全局令牌桶，完成限速相关操作。

在无锁限速方案中，由于只使用了全局令牌桶，因此不存在低速率时本地令牌桶预留令牌导致的限速后速率偏低，以及预取令牌导致的限速后速率偏高的精度问题。

CPU	Pre-production 4th Generation Intel® Xeon® Scalable Processor
Memory	512GB
NIC	E810-C for QSFP x 2
NIC Firmware Version	2.30 0x80005d1b 0.0.0
Microcode	0x8f000320
Operating System	Ubuntu* 20.04 LTS
Linux* Kernel Version	5.4.0-67-generic
DPDK Version	20.11+DLB patch



图三 英特尔® DLB 的四种队列模型



图四 基于英特尔® DLB 技术的无锁限速方案

详见图四。

方案对比测试

本节介绍对比测试的原理、拓扑结构和配置。

■ 测试原理与拓扑

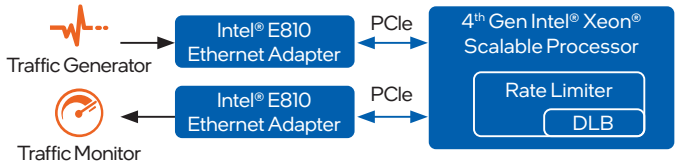
测试中以目的 IP 地址区分不同的需要限速的网络数据流，通过网络测试仪向被测设备 (Device Under Test, DUT) 发送不同目的 IP 地址的网络数据流，数据包在被测设备处理后返回给网络测试仪；网络测试仪每 2 秒统计一次数据包的接收速率，连续统计 20 次 (共 40 秒)，然后记录结果。

限速软件限制每个目的 IP 的速率为 1Mbps，网络测试仪发送的待观测数据流速率超过 1Mbps，使限速软件丢弃部分网络报文，以便观察限速精度。

测试分为两次。第一次，在被测设备运行使用轻量化锁限速方案的软件，记录测试结果；第二次，在被测设备运行使用无锁限速方案的软件，记录测试结果。依据两次测试的结果，比较两个方案的限速精度。

■ 测试配置

被测设备使用了集成英特尔® DLB 的第四代英特尔® 至强® 可扩展处理器 (预生产版本)，以及两张 E810 网卡，每张网卡使用一个 100Gbps 接口，分别连接到网络测试仪的两个测试端口。详见图五。



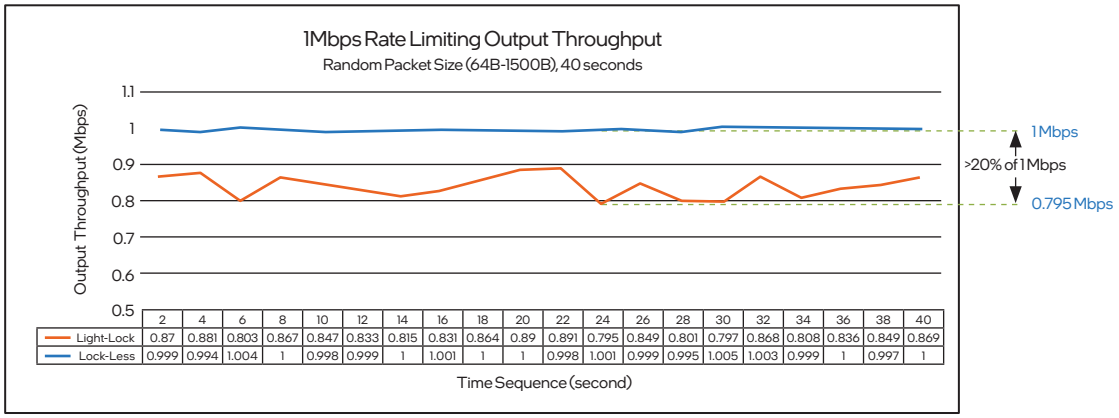
图五 测试原理与拓扑结构

被测设备的具体配置如下:

流量组	IP 数	随机 DIP 设置	组总带宽 (Mbps)	单用户平均带宽(Mbps)
普通用户组	16384	192.168.0.0 / 0.0.63.255	8192	0.5
恶意用户组	4096	192.168.64.0 / 0.0.15.255	6144	1.5
观测组	1	192.168.192.10 / 0.0.0.0	2	2

网络测试仪流量模型

为了使测试接近实际场景，测试过程中使用 64 字节至 1,500 字节范围内的随机长度的报文，流量模型中包含背景流量和观测流量。背景流量由 80% 的不超速访问的普通用户和 20% 的超速访问的恶意用户构成。受网络测试仪能力所限，无法对所有 IP 的流量进行统计，故添加一观测组流量，该流量中只包含一个目的 IP 地址，通过网络测试仪统计观测组流量的接收速率，实现对限速效果的观测。



	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34	36	38	40
Light-Lock	0.87	0.881	0.803	0.867	0.847	0.833	0.815	0.831	0.864	0.89	0.891	0.795	0.849	0.801	0.797	0.868	0.808	0.836	0.849	0.869
Lock-Less	0.999	0.994	1.004	1	0.998	0.999	1	1.001	1	1	0.998	1.001	0.999	0.995	1.005	1.003	0.999	1	0.997	1

图六 测试对比结果

测试结果对比与分析

测试于 2021 年 5 月进行。图六为测试结果图表，其中将轻量化锁限速方案的 20 个数据采样点连成橙色的线，无锁限速方案的 20 个采样点连成蓝色的线。

从图中可以看到，无锁方案整体限速非常稳定且准确，整体误差小于 1%；而轻量化锁方案，限速后的流量速率偏小，且有大幅度波动，甚至出现了大于 20% 的误差。

以上测试说明，使用基于英特尔® DLB 的无锁限速方案相比轻量化锁限速方案，能够获得更高的限速精度。

■ 无锁限速方案的灵活性和通用性

利用英特尔发布的 DLB 软件开发包 (SDK) 以及 DPDK 软件库，可以基于英特尔® DLB 优化现有在 Linux 内核空间、用户空间或者 DPDK 框架中开发的应用程序。因此，使用英特尔® DLB 技术实现的无锁限速方案，也可以应用于多种形态的应用程序。

同时，英特尔® DLB 也支持英特尔® Scalable-IOV 和 SR-IOV 这些 IO 虚拟化技术，使得单个英特尔® DLB 设备可虚拟出多个虚拟设备，共享给多个虚拟机或容器使用。

总结

英特尔® DLB 作为第四代英特尔® 至强® 可扩展处理器引入的新型加速技术，具有丰富的硬件队列管理功能，为软件优化带来了新的可能。

使用英特尔® DLB 的 Atomic Queue 特性所实现的无锁限速方案，能够避免了使用“锁”来保护全局令牌桶的需求，因此无需再考虑对于“锁”的优化。通过与现有的基于轻量化锁的方案进行测试对比，本白皮书展现了无锁限速方案在限速精度和性能上的优势。

基于上述阐述和测试，可以得出结论，那就是基于英特尔® DLB 技术的无锁限速方案，借助英特尔® DLB 软件开发包提供的丰富开发库，可在 Linux 系统的内核态、用户态以及 DPDK 框架中实现精准限速，且可灵活地应用于不同场景。

鸣谢

本白皮书中的应用场景和性能数据是在以下英特尔同事的帮助下完成的:

曹亚辉，李怡文，梁存铭，Pravin Pathak，Edward Pullin，Jay Vincent，许茜，朱河清

缩略词

API	Application Programming Interface	应用程序接口
CPU	Central Processing Unit	中央处理器
DIP	Dynamic IP	动态 IP
DPDK	Data Plane Development Kit	数据平面开发套件
DUT	Device Under Test	被测设备
Intel® DLB	Intel® Dynamic Load Balancer	英特尔® 动态负载均衡加速器
IOV	Input/output Virtualization	I/O 虚拟化
LTS	Long Term Support	长期支持
NIC	Network Interface Card	网卡
QSFP	Quad Small Form-factor Pluggable	四通道 SFP 接口
SDK	Software Development Kit	软件开发工具包
SLA	Service Level Agreement	服务水平协议
TGW	Tencent Gateway	腾讯云网关

相关资料

文档名称	文档编号/链接
英特尔® DLB 编程指南	613545
英特尔® DLB 软件开发包	686372
DPDK Eventdev (英特尔® DLB) 开发指南	http://doc.dpdk.org/guides/eventdevs/dlb2.html