DRILL: Micro Load Balancing for Low-latency Data Center Networks

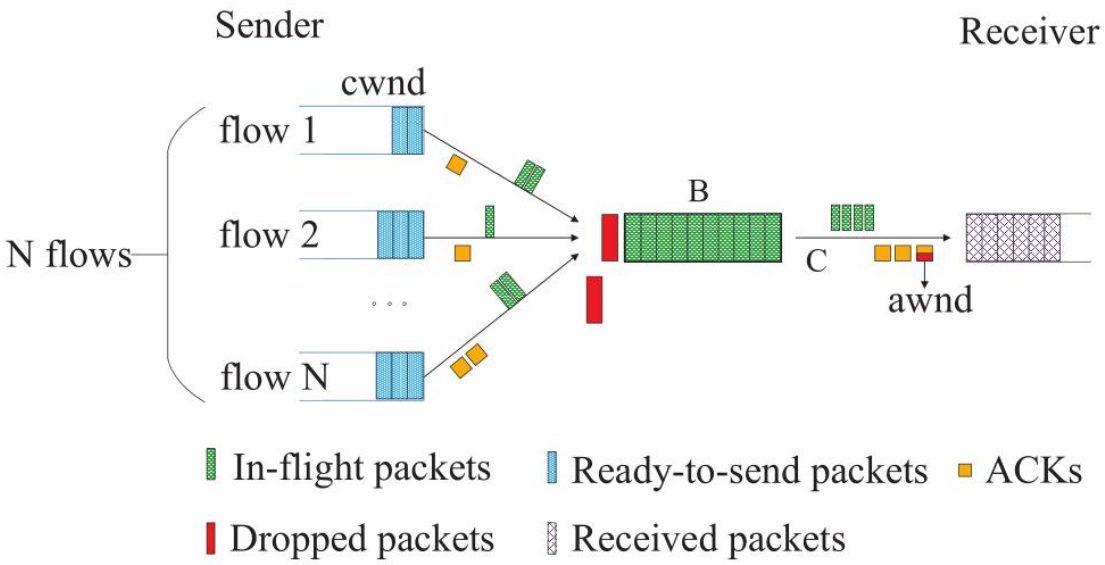
SIGCOMM '17

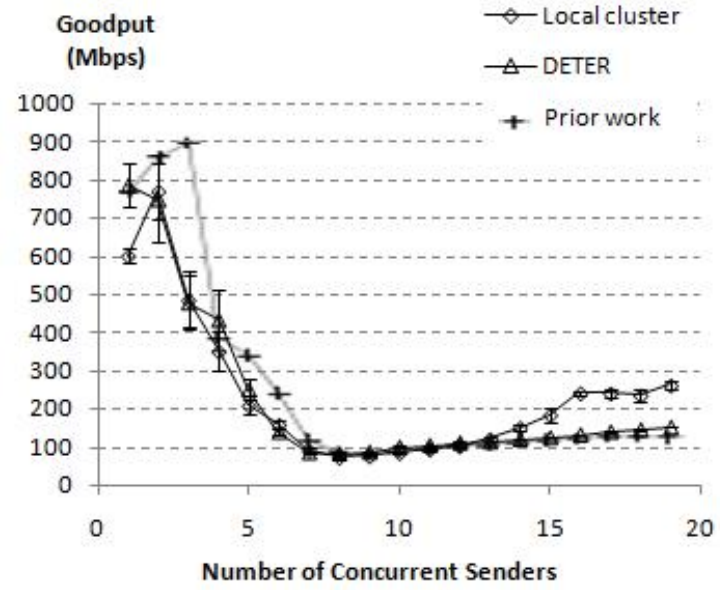
浙江大学
计算机系统结构实验室
Computer Architecture Laboratory
of Zhejiang University

徐天宇

周会 2021.07.15

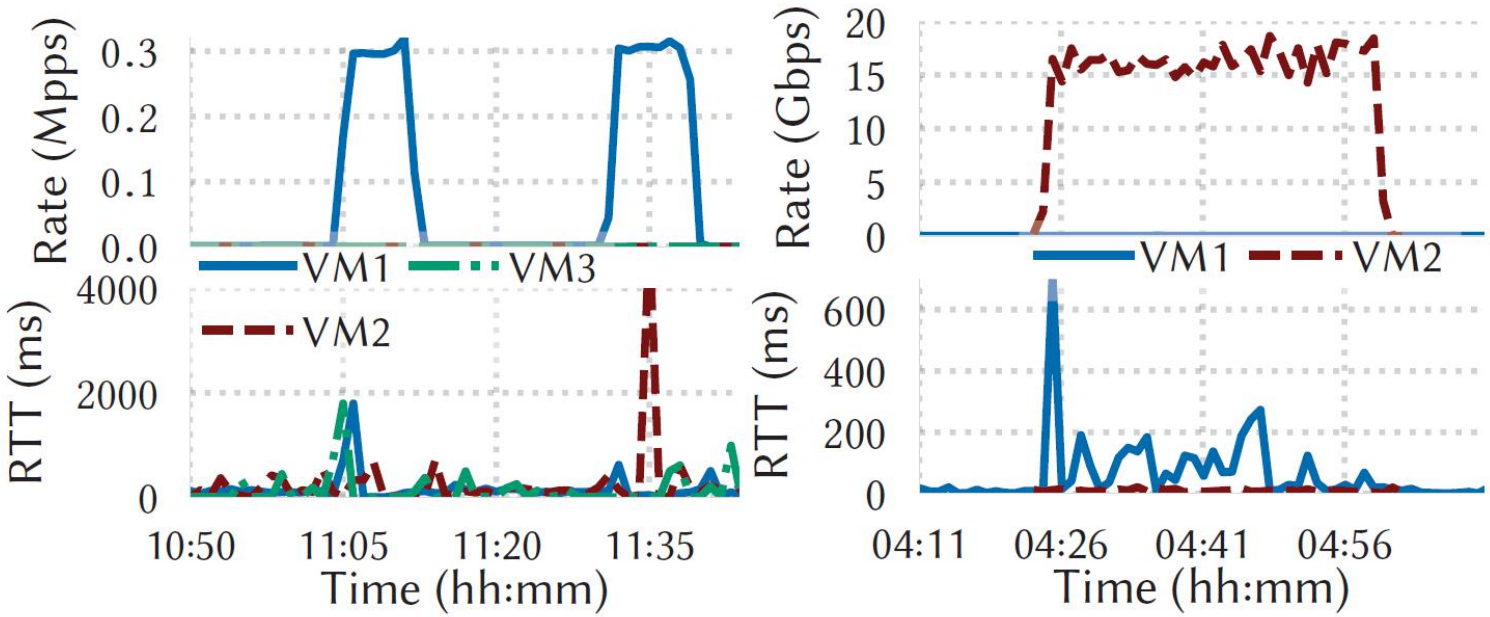
Incast问题





Incast：多对一传输场景下，缓冲区大小不够用，进而出现队列溢出丢包，导致数据传输速度/吞吐量大幅度降低

Incast问题



(a) Packet rate overload　　**(b)** Bandwidth overload

云网络平台下发生Incast，还容易导致性能隔离被打破，即当某一VM产生Incast问题时，同一Host的其他VM的性能也受到影响

Incast问题

- 在在线数据密集型应用（如大数据中心）中，分治（divide and conquer）方法被广泛应用的计算范式中，如MapReduce、Spark、Dryad、CIEL、TritonSort等多对一流量模式较为常见。但是，在汇聚期间多个发送方同时发送的数据很可能会淹没交换机缓冲区，使TCP的丢失恢复机制失效。然后会触发异常超时，导致灾难性的吞吐量崩溃。

- 同时，微突发、高并发等特殊的流量模式（例如最近两年非常火的网络直播和流媒体业务）容易造成TCP Incast问题，严重降低了应用程序的性能。

负载均衡

- 数据中心(data center)作为云计算的硬件基础架构也在不断普及和应用.为了构建高可用、高性能、低成本的云计算基础存储和计算设施,数据中心通常部署了大量商用交换机和服务器。数据中心网络连接了大规模服务器集群,是传递计算和存储数据的桥梁。为了提供超高带宽,数据中心网络的拓扑结构普遍采用CLOS结构，在主机之间提供了多条可用路径.在网络高负载状态下,为了降低链路拥塞和数据包丢失的概率,数据中心负载均衡（Load Balance，LB）机制将网络流量分配到所有可用路径上,充分利用了网络中存在的冗余链路,提高网络传输性能.[1]

[1]数据中心负载均衡方法研究综述.软件学报’21

负载均衡

- 基于LB方案部署的位置，可以将LB方案分为三大类：
  - 基于中央控制器
  - 基于交换机
  - 基于主机
- 工作原理：
  - 感知全局拥塞/感知局部拥塞/本地队列/感知路径拥塞/不感知拥塞等
- 调度粒度：
  - 交换机/流级别/子流级别/包簇级别/包级别/动态切换等[1]

[1]数据中心负载均衡方法研究综述.软件学报’21

负载均衡

- DRILL：
  - 依据本地队列信息调度
  - 调度粒度：包级别

数据中心网络拓扑结构[1]

- **传统的大三层网络结构**
- 核心层：

核心层是网络的高速交换主干，对整个网络的连通起到至关重要的作用

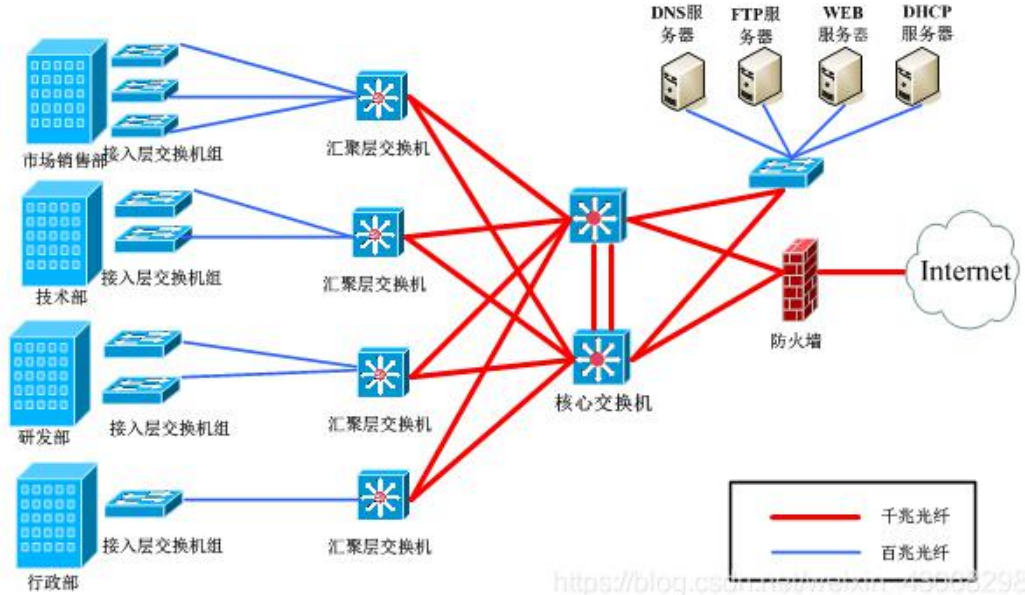
核心层应该具有如下几个特性：可靠性、高效性、冗余性、容错性、

可管理性、适应性、低延时性等；

- 汇聚层：

汇聚层是网络接入层和核心层的“中介”，就是在工作站接入核心层

前先做汇聚，以减轻核心层设备的负荷。



汇聚层具有实施策略、安全、工作组接入、虚拟局域网（VLAN）之间的路由、源地址或目的地址过滤等多种功能。

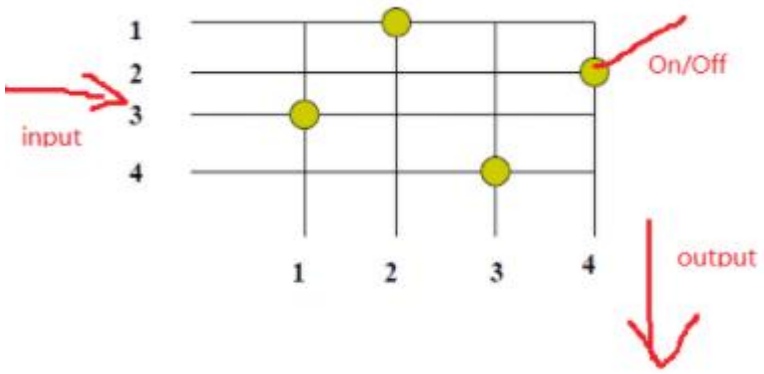
- 接入层：

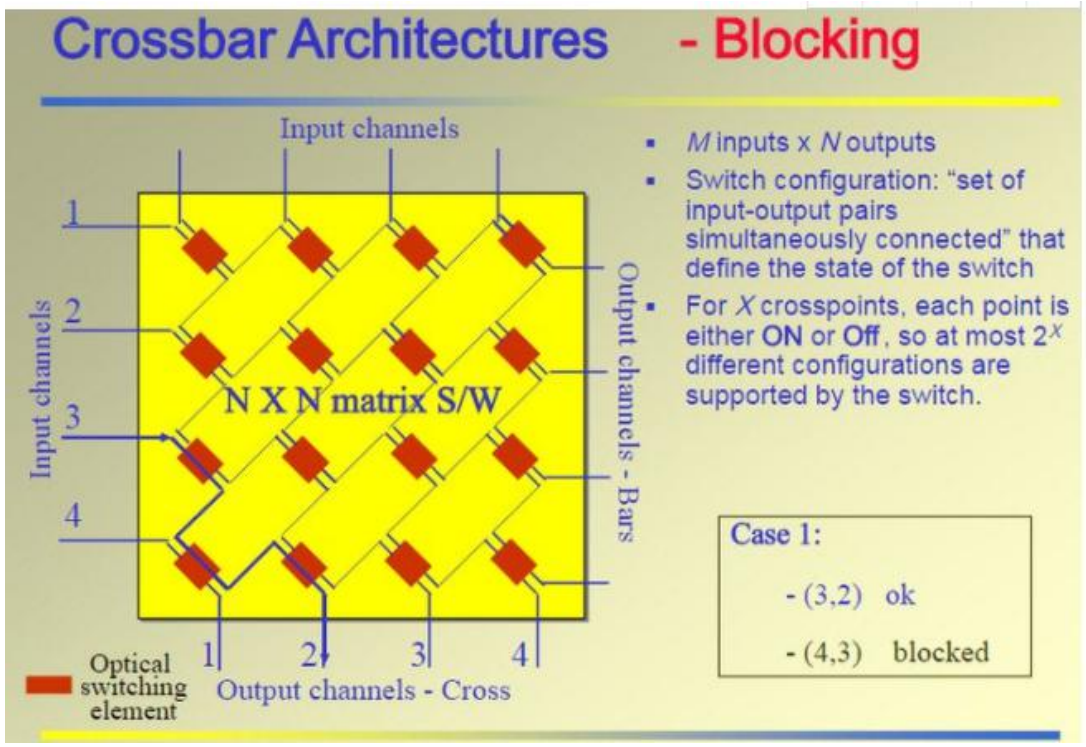
接入层向本地网段提供工作站接入。在接入层中，减少同一网段的工作站数量，能够向工作组提供高速带宽。

[1]云数据中心网络与SDN.机械工业出版社

数据中心网络拓扑结构[1]

- 向二层网络演进
  - CrossBar（Fabric）
  - 结构简单、控制简单
  - 规模有限，没有冗余通路、易阻塞

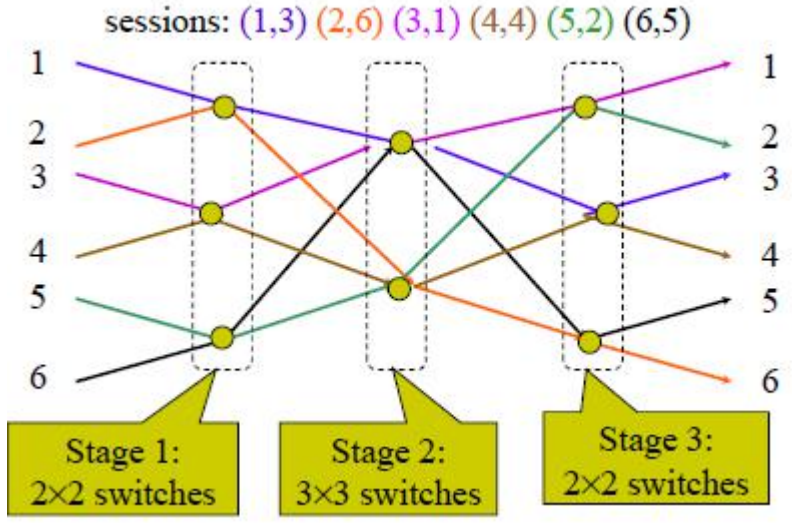


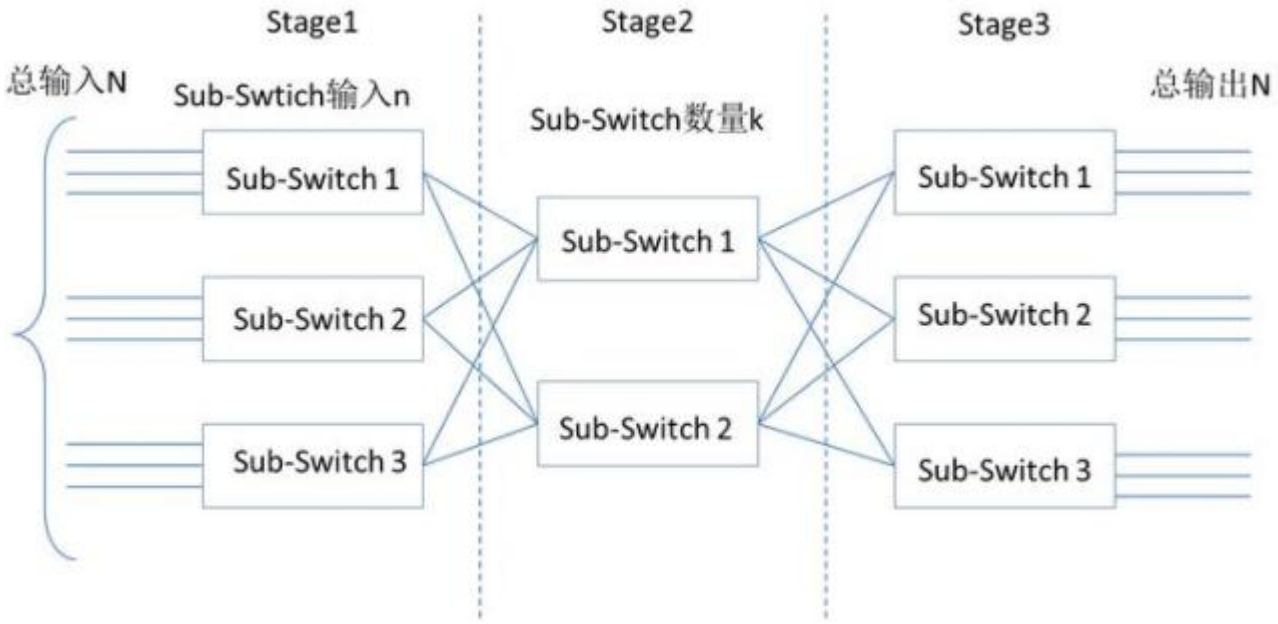


[1]云数据中心网络与SDN.机械工业出版社

数据中心网络拓扑结构[1]

- 向二层网络演进
  - Clos
  - 节约成本、增加效率



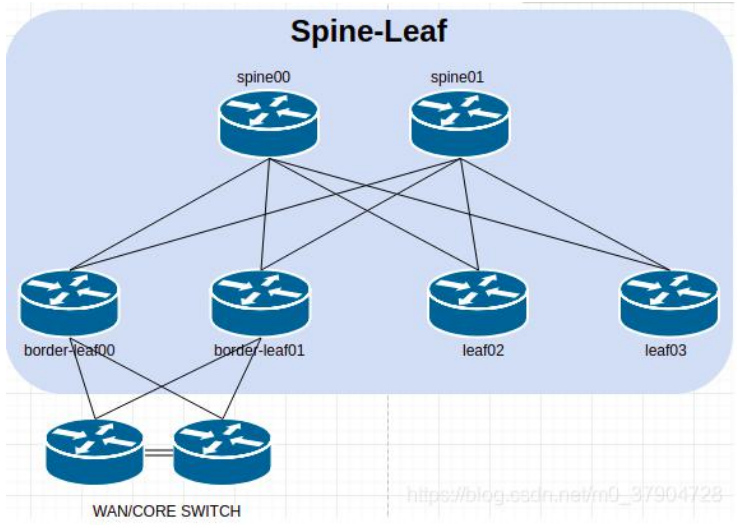


[1]云数据中心网络与SDN. 机械工业出版社

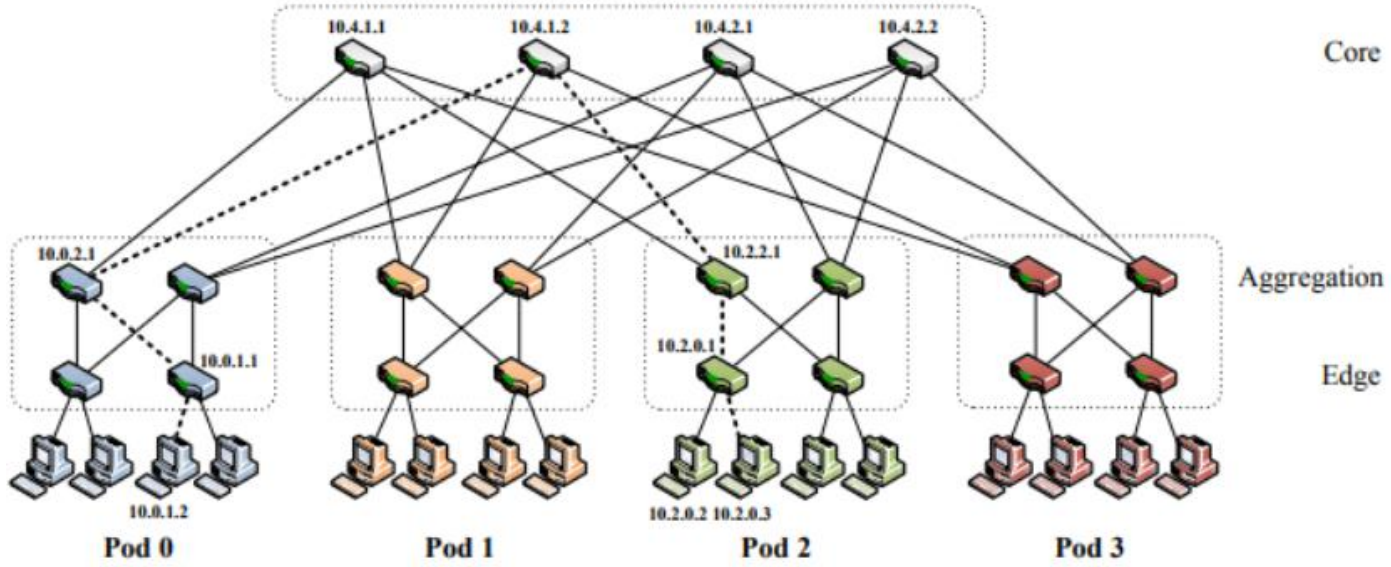
数据中心网络拓扑结构[1]

- 向二层网络演进
  - Clos

Leaf-Spine



Fat-Tree



[1]云数据中心网络与SDN.机械工业出版社

数据中心网络拓扑结构[1]

- 向二层网络演进
  - Clos
  - Facebook的网络拓扑

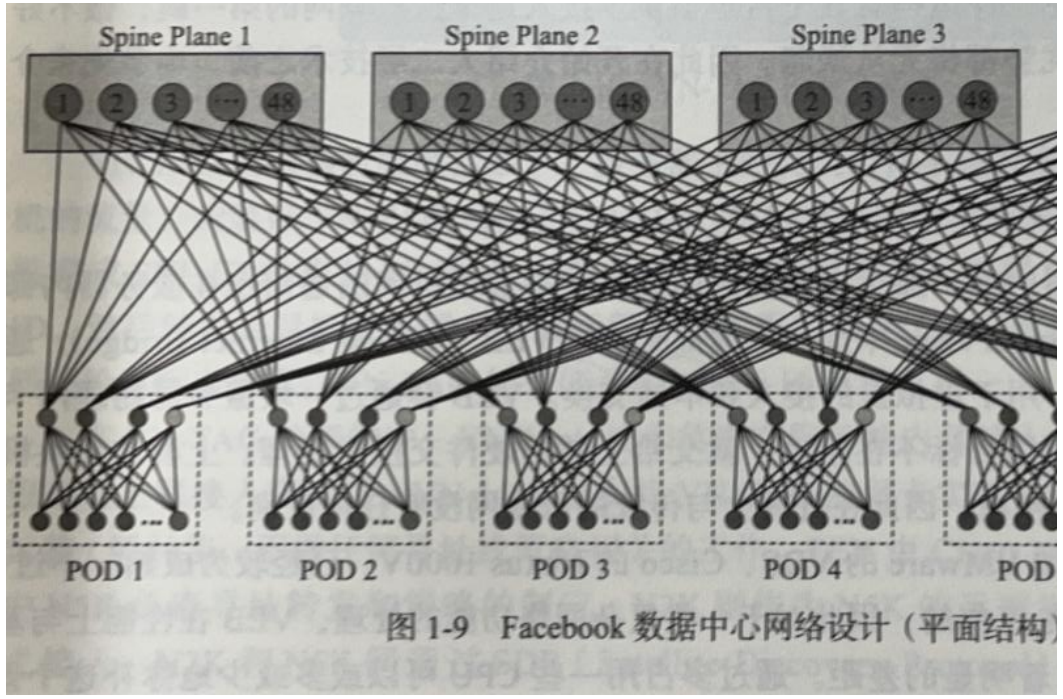
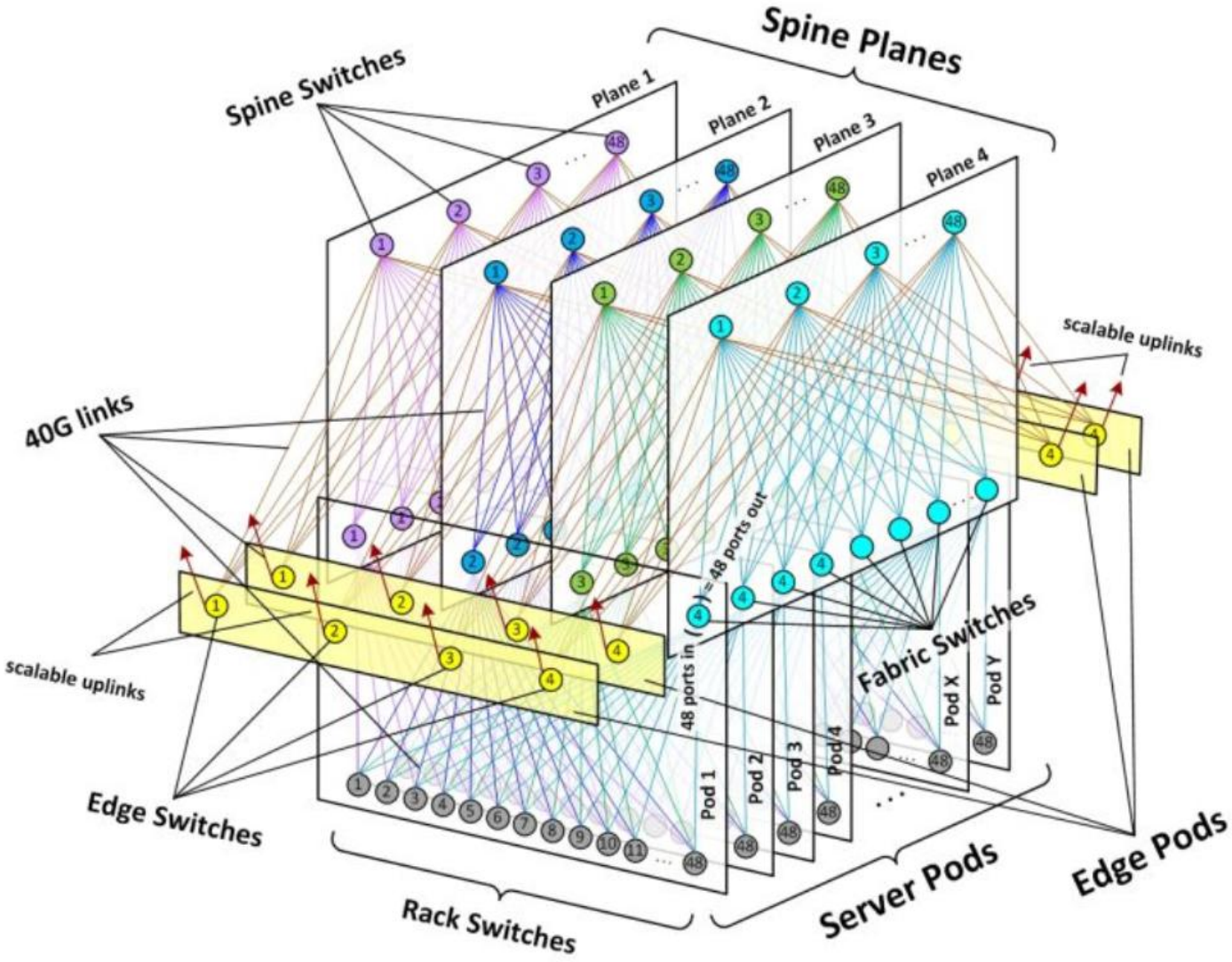


图 1-9 Facebook 数据中心网络设计（平面结构）



[1]云数据中心网络与SDN.机械工业出版社

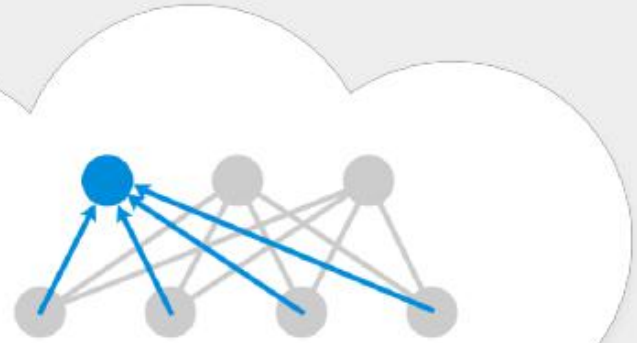
DRILL

A STARTING POINT : "EQUAL SPLIT FLUID" (ESF)

Equally split all incoming flow to other leaves along all shortest outgoing paths.

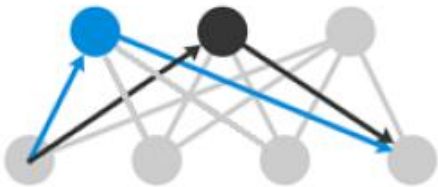
DRILL

A STARTING POINT : "EQUAL SPLIT FLUID" (ESF)



Each of n spines receives
1/n of all inter-leaf traffic.

A STARTING POINT : "EQUAL SPLIT FLUID" (ESF)



Therefore, any two paths between the same source and destination experience the same utilization (and mix of traffic) at all corresponding hops.

ESF is optimal for all traffic demands.

DRILL

◆ 算法Drill(d,m)——**对称路由**：

1. 通过ECMP+OSPF找到该switch的所有等价路径
2. 准备一个m大小的储存区
3. 当一个数据包到来时，随机选择N种可能的路径中的d个
4. d+m中选出load最小的，发送
5. 选择load最小的m条路径，更新m储存区

DRILL

◆ 确定d与m

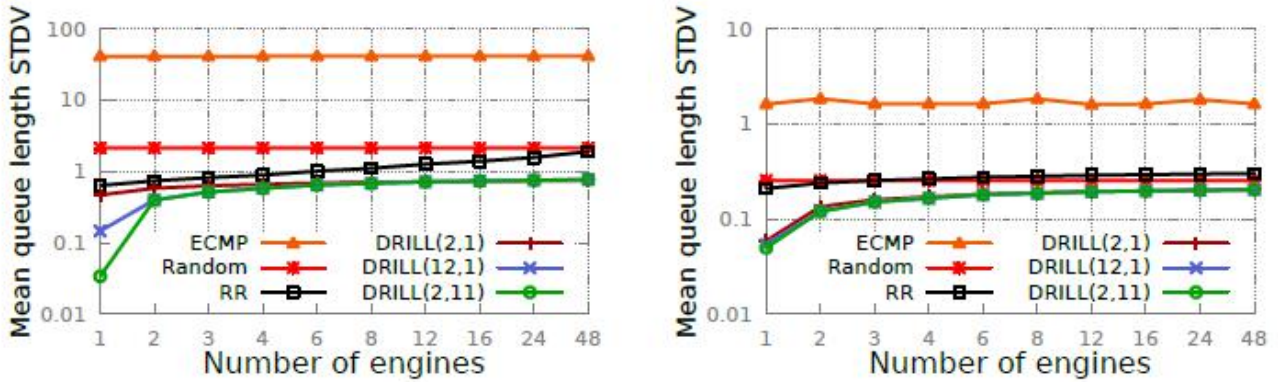


Figure 2: (a) 80% load. (b) 30% load. Adding a choice and a memory unit improves performance dramatically.

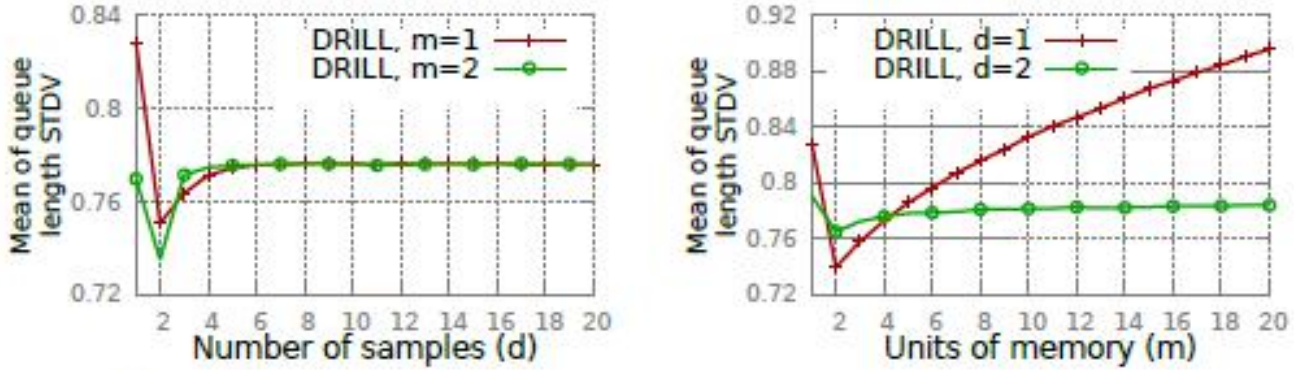


Figure 3: Excessive choices & memory cause sync effect.

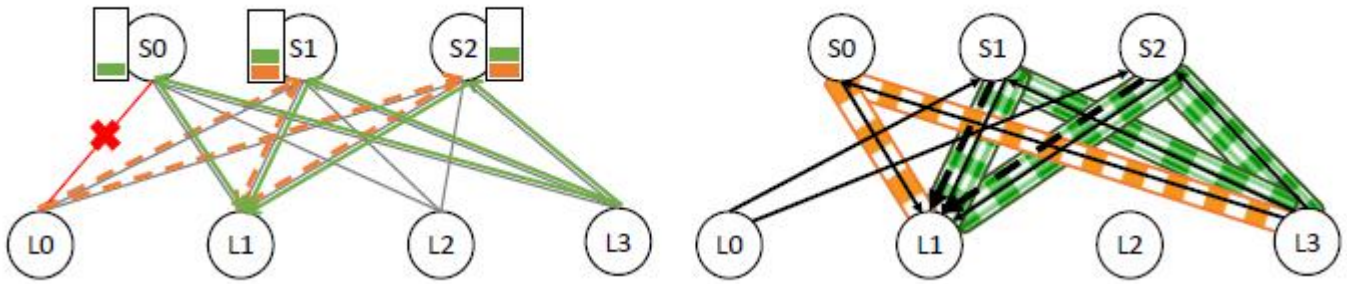
DRILL

- 包重传问题
  - 使用DRILL，只有当线路延迟 >> 排队延迟时，才会触发包重传（TCP-ACK机制）。
    但是数据中心大部分延迟来自于排队延迟，DRILL保证了良好的负载均衡，
    线路延迟 << 排队延迟,其包重传问题并不显著。
  - 同时DRILL保留了解决包重传问题的方案——在终端主机引入stack来对包重排序。

DRILL

◆ 路径非对称问题

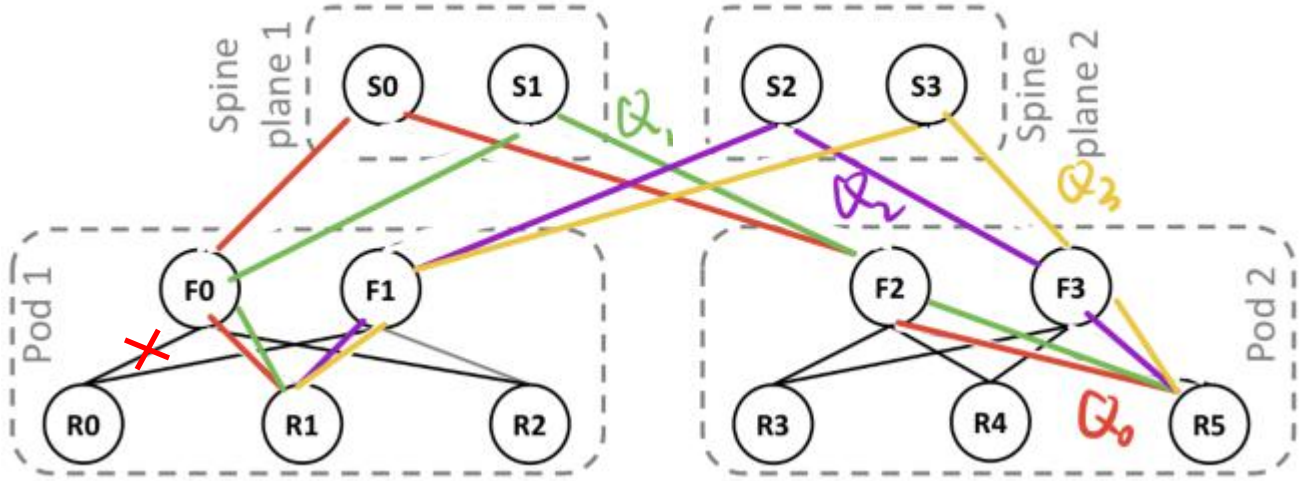
- 在控制平面对路径进行评估，近似地分解为微小的对称路径
- （先构建带标签多向图Quiver，然后分解）



DRILL

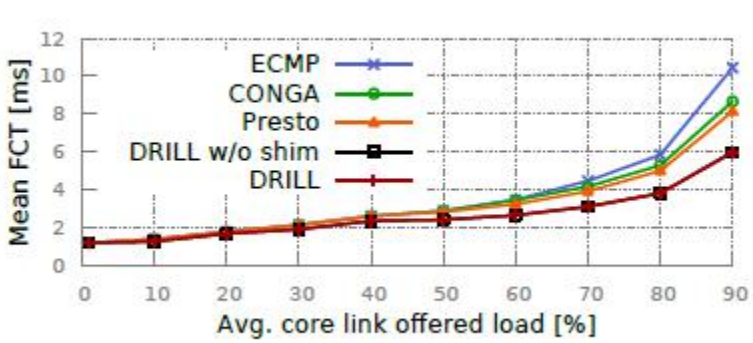
◆ 路径非对称问题

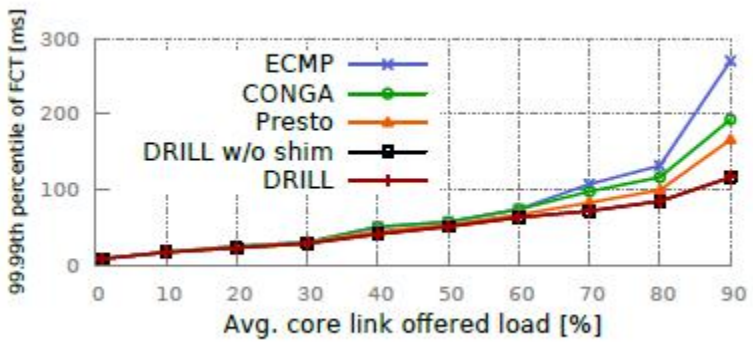
- 在控制平面对路径进行评估，近似地分解为微小的对称路径
- （先构建带标签多向图Quiver，然后分解）

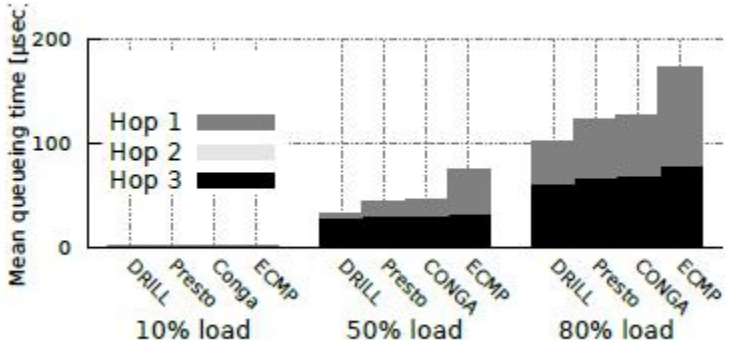


DRILL

- Evaluation
  - symmetric clos



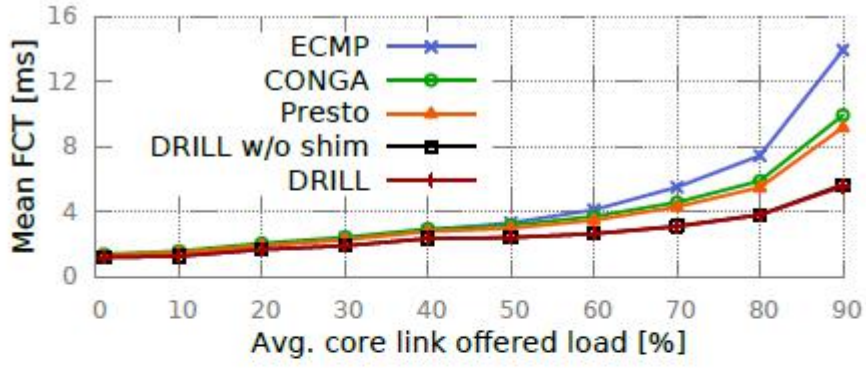




DRILL

◆ Evaluation

● scalability



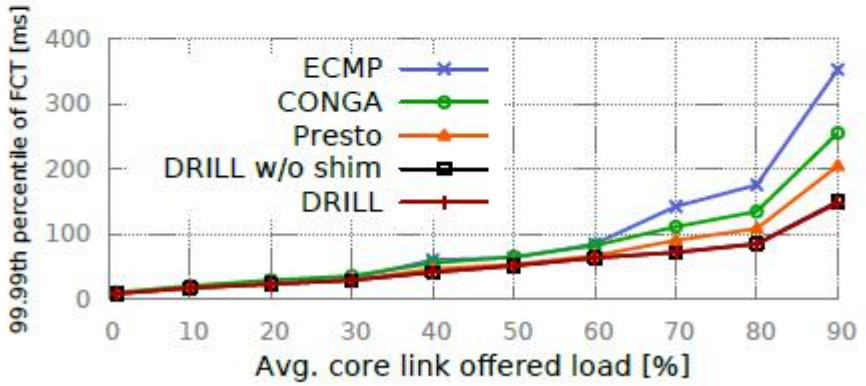
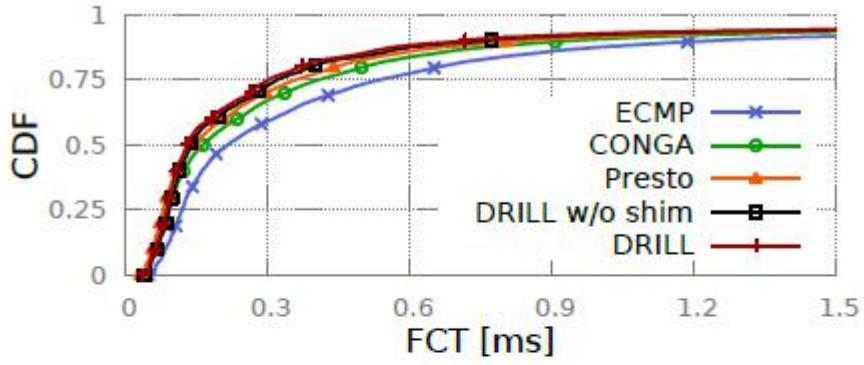


Figure 7: DRILL handles scaling-out gracefully.

CDF (cumulative distribution function)



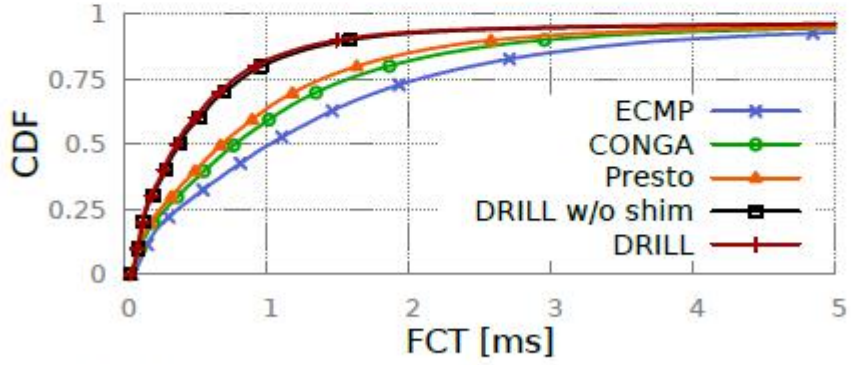
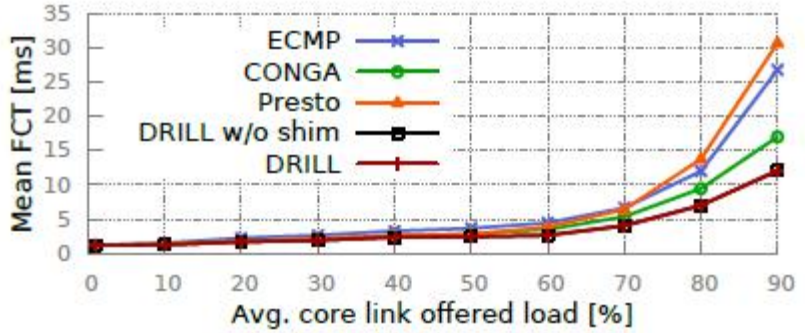


Figure 8: Scale-out topology with (a) 30% load (b) 80% load. DRILL's improvement is greater under heavy load.

DRILL

- Evaluation
  - packet reorder && link failure



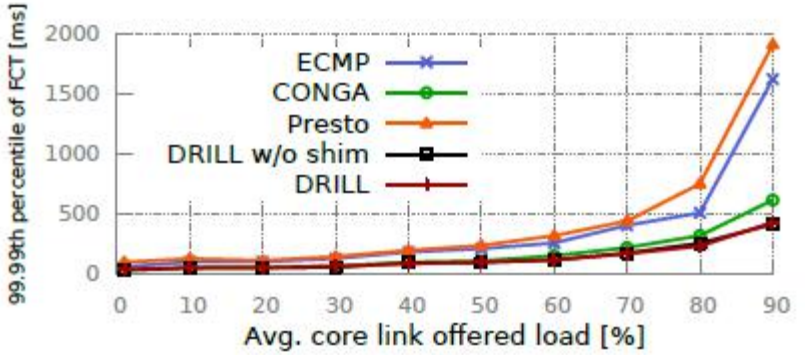
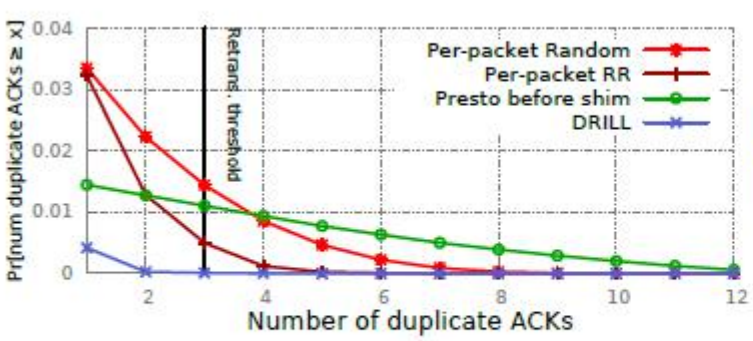
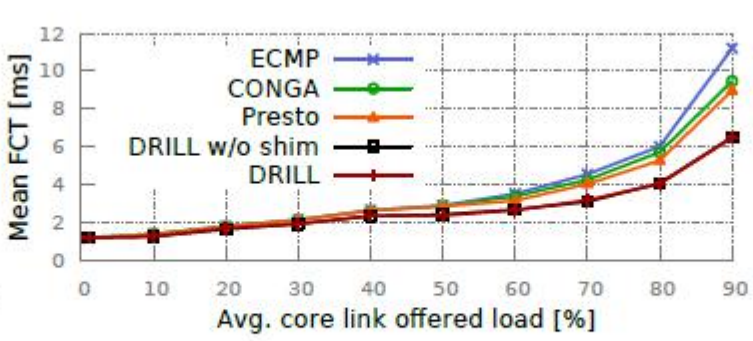


Figure 12: DRILL handles 10 link failures.





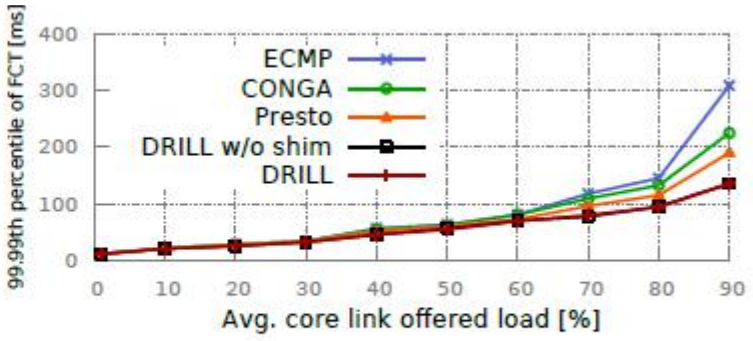


Figure 11: (a) Less than 0.1% of flows with DRILL hit TCP retrans. threshold, (b,c) DRILL handles single link failure.

DRILL

- Evaluation
  - asymmetric

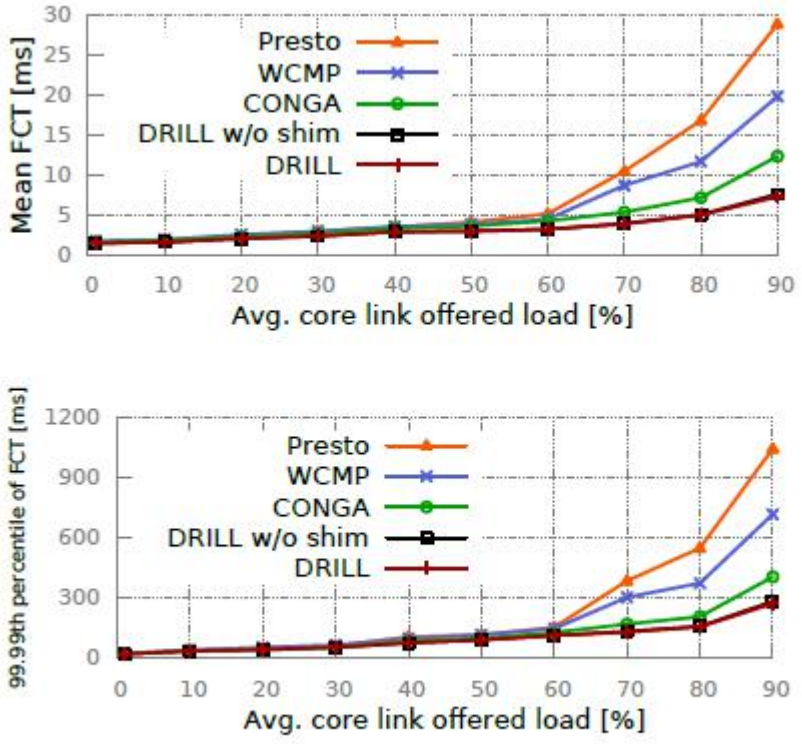


Figure 13: DRILL is efficient in heterogeneous topologies.

DRILL

◆ Evaluation

● incast problem

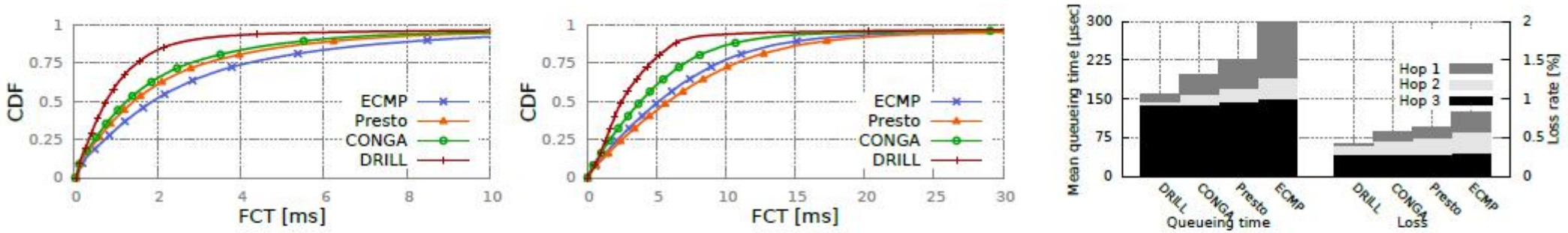


Figure 14: Incast: (a, b) DRILL cuts the tail latency under 20% and 30% load, (c) Where queueing and loss happen under 20% load, across hop 1 (first leaf upward to spine), hop 2 (spine downward to leaf), and hop 3 (leaf to host).

	Stride			Bijection			Shuffle		
	CONGA	Presto	DRILL	CONGA	Presto	DRILL	CONGA	Presto	DRILL
Elephant throughput	1.55	1.71	1.8	1.46	1.62	1.78	1	1.1	1.1
Mean FCT	0.51	0.41	0.21	0.71	0.63	0.45	0.95	0.91	0.86
99.99th percentile FCT	0.2	0.15	0.04	0.22	0.18	0.08	0.86	0.79	0.68

Table 1: Mean elephant flow throughput and mice FCT normalized to ECMP for the synthetic workloads.