Improving the effect of the short flows in the data center network with Multipath TCP

- Multipath TCP適用時のデータセンターネットワークでのショートフローに対する影響の改善

Shogo Fujii - sekiya labratory

14-MU

Outline

Today's requirement for large-scales cloud data center is getting higher and higher, for example web search engine and SNS(Social networking service) demands soft real-time response. Multipath TCP (MPTCP) improves throughput on modern topology which has multipath on host-to-host traffic. However, some researchers reported MPTCP affects short flow traffic negatively. In my research, I figure out why negative effect happens with two negative factors, bottleneck at aggregation switch and multiplexed flow patterns in one host.

Background

Big data growth matters seriously …

- Approach in data center

Scale-out, redundancy, cloud service

Improving datacenter network with MPTCP

MPTCP affects short flow traffic negatively[10]

- Why short flow?

commonly using distributed processing application in data center

- 80% of data center traffic is short flow[16]
- Partition /Aggregate structure[3]

Short flows in datacenter network is matter

Related work : approaching for switch, protocol etc…[9, 11]

In my research –

1. Using modern topology for massive resources
2. Seamless operation : without special implementation and device
3. Application friendly : optimizing the specified traffic patterns

Goal : Constantly high performance datacenter network with MPTCP

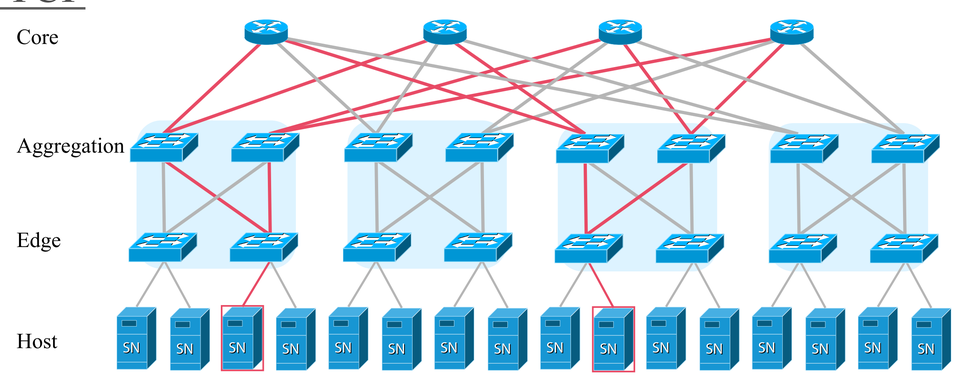

Fig.1 Fat tree topology(k=2)

Exploratory approach

- As the result of reproduction experiment[20],

1. MPTCP generates more traffic data, and causes delay with packet loss.
2. MPTCP has not achieved perfect effectiveness, only with 2 paths.

- **Hypothesis** : Using uncongested path, improving the performance of short flow.
- **Purpose**

1. validate the hypothesis.
2. figure out why negative effect for short flows happens.

- Simulation environment

Topology : 4-node(2-pairs) FatTree like topology

Simulator : ns-3 DCE(Direct Code Execution)

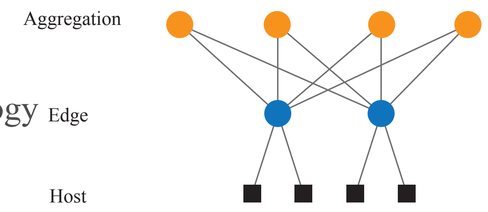

Fig.2 Simulation topology

- Benchmark traffic

Background flow with MPTCP and 2~70KB short flow on average 200ms (poisson arivals), and measuring FCT(Flow completion time) for short flow - 1000times on simulation.

- Result

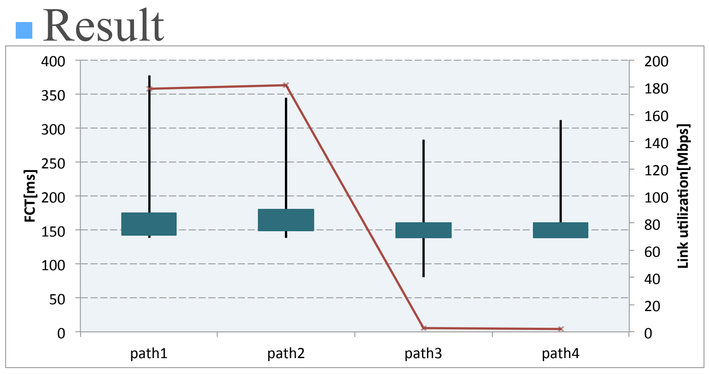

Fig3: Flow completion time and link utilization for 70kb benchmark traffic

Table1 : network parameter

Parameter	Value
aggr-edge	200Mbps
edge-host	100Mbps
RTT	0.5ms
Buffer	50kb

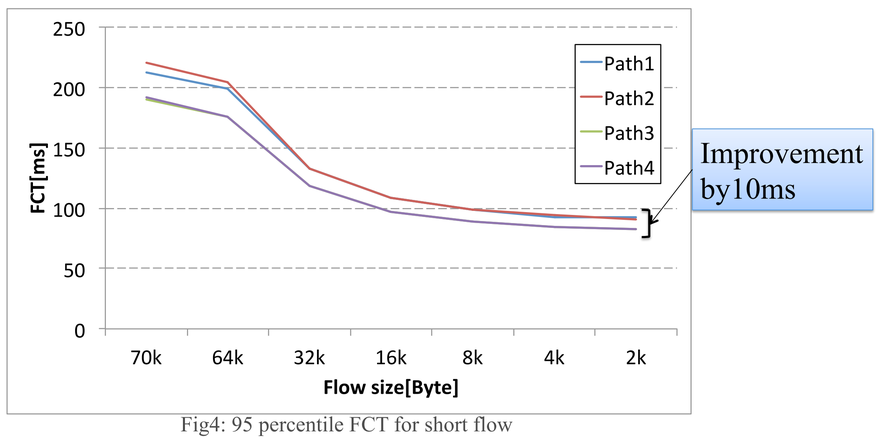

Fig4: 95 percentile FCT for short flow

Analysis

- Effect of link utilization and short flow interval

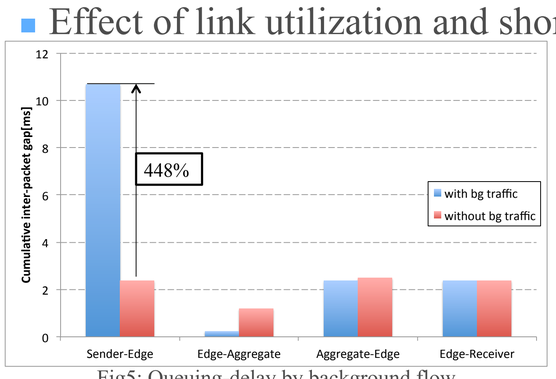

Fig5: Queuing-delay by background flow

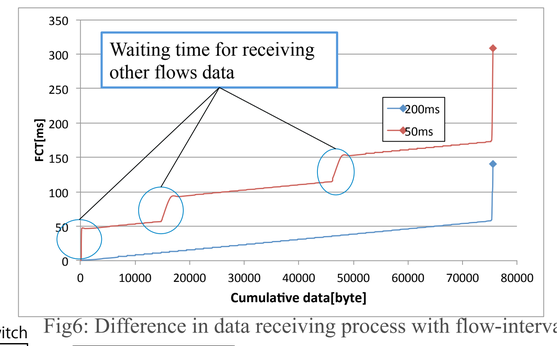

Fig6: Difference in data receiving process with flow-interval

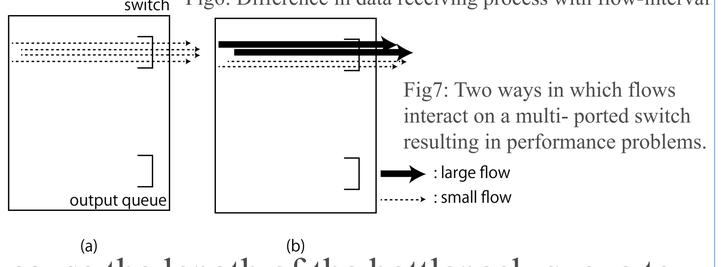

Fig7: Two ways in which flows interact on a multi- ported switch resulting in performance problems.

- What happens?

1. Queue buildup-Fig7(b)

Long-lived background flow cause the length of the bottleneck queue to grow.

2. Incast-Fig7(a)

Many flows coverage on the same interface of aggregate switch over a short period of time, synchronously.

- Impairment on the short flows:

1. queue-buildup delay, they are in queue behind packets from the large flows, even when no encounter with bursty traffic.
2. packet loss.

- Solution :

Reducing the size of the queues at aggregation switch

Conclusion

- Validated the hypothesis from reproduction experiment, and improving the performance of short flow by selecting uncongested path.
- Clarified short flow delays arise under what circumstances; Queue buildup and Incast patterns, and gave the direction for solutions.
- Reproducing these problems in a real world, and proposing the method for avoiding pressuring aggregating switch queue.
- Analyzing the unclear point of short flow delay, application or hardware side etc. as my future work.