Baidu’s Best Practice with Low Latency Networks

Feng Gao

IEEE 802 IC NEND
Orlando, FL
November 2017

Presented by Huawei

01

Low Latency Network Solutions

1. Background Introduction
2. Network Latency Analysis
3. Low Latency Network Solutions
4. Best Practice

Background Introduction

Artificial Intelligence

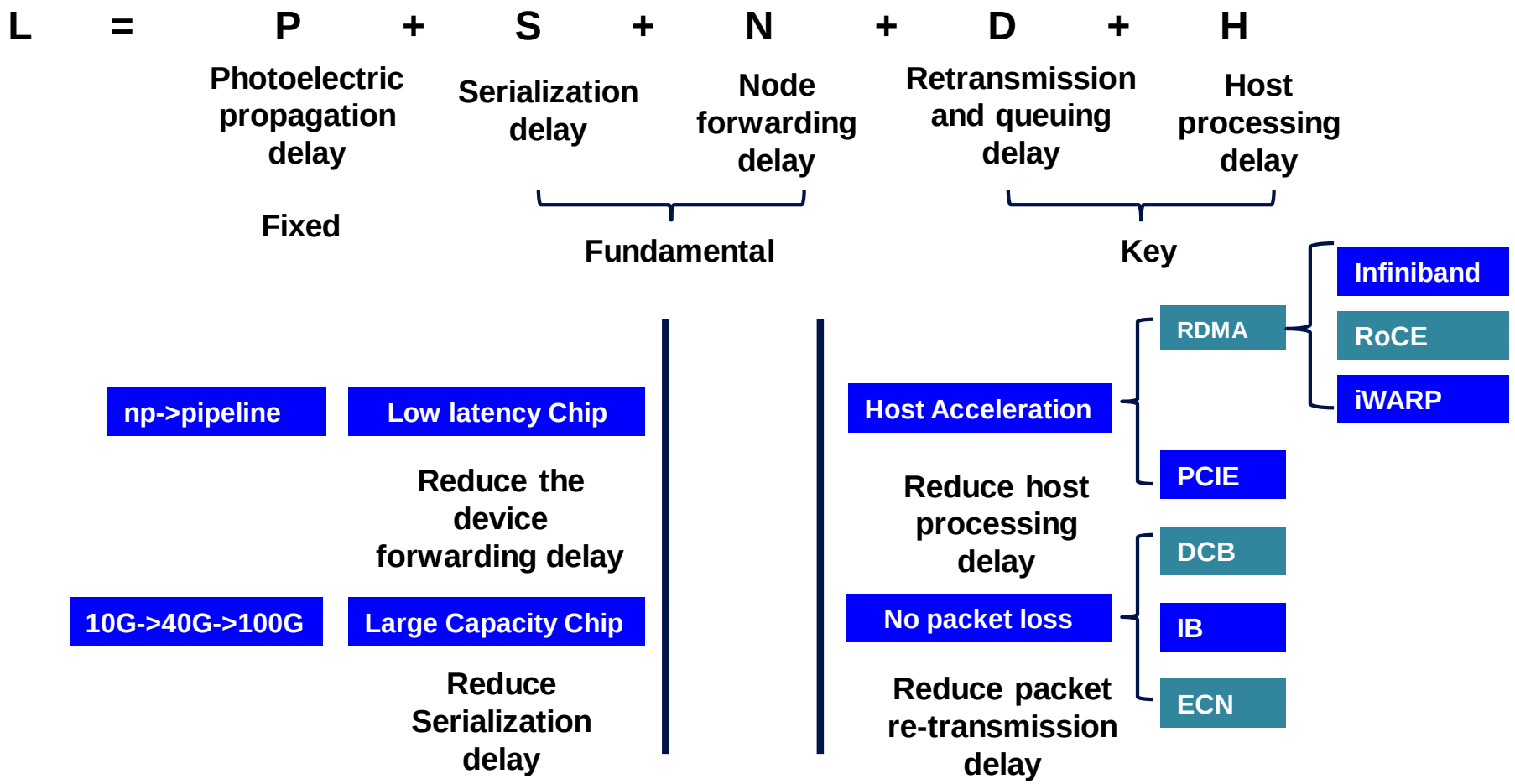
High Performance Computing Cloud

Real Time Big Data Analysis

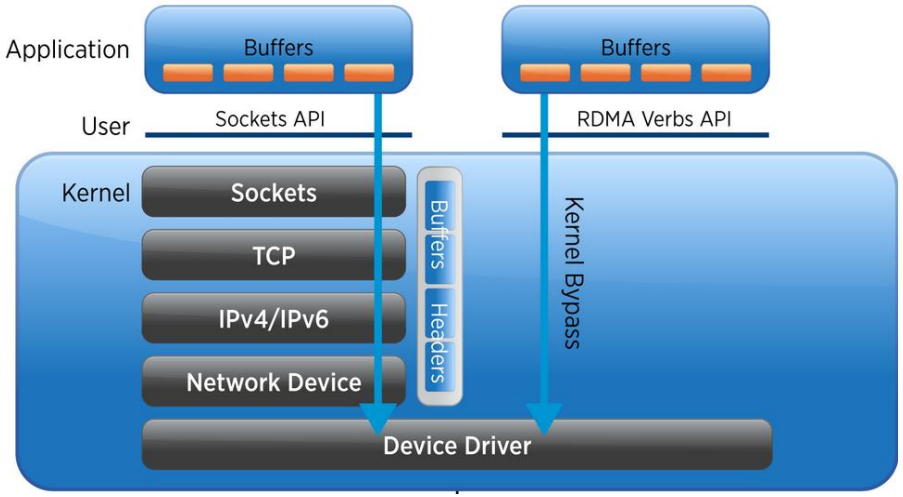
• Latency-sensitive applications are deployed and developed in Data Centers, from the simple pursuit of **high bandwidth, non-blocking** to the pursuit of **low latency, no packet loss**

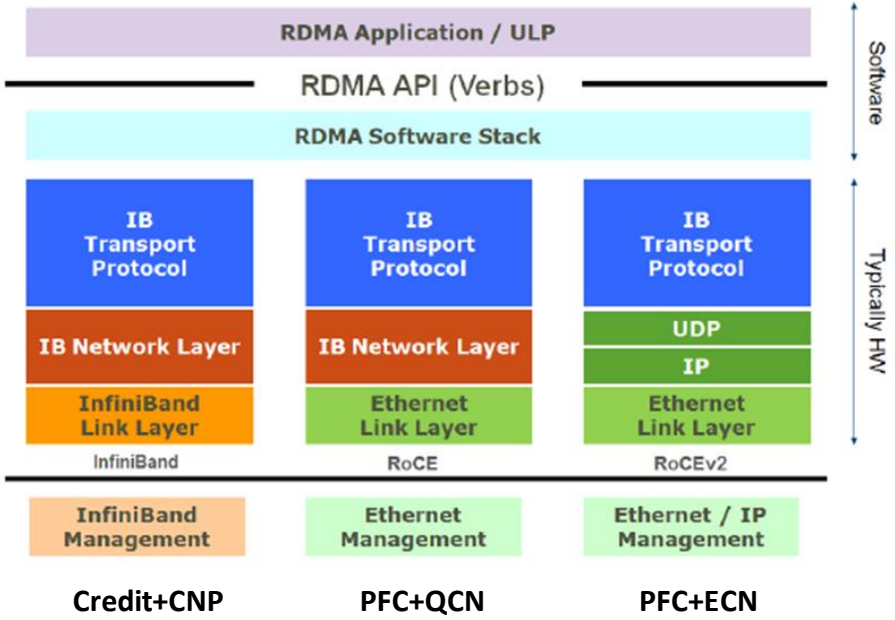
• **Bandwidth-centric** network design is switched to **latency-centric** design. Reduce the jitter of latency.

Network Latency Analysis



Low Latency Solution : Host Acceleration





RDMA vs TCP/IP

• Kernel Bypass brought by RDMA reduces the latency on the Host

RoCEv2

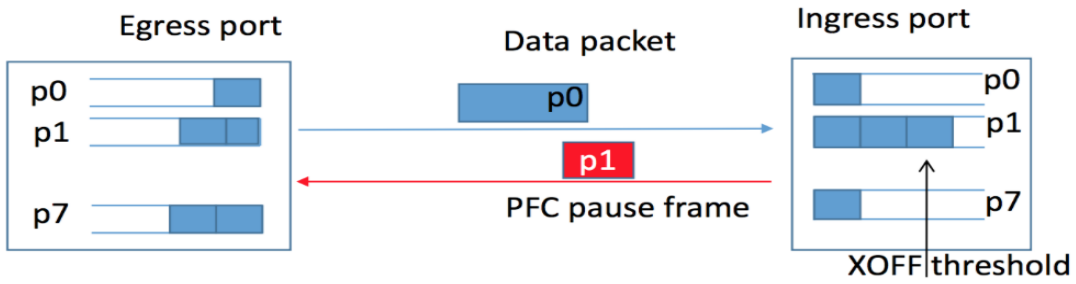
• Compatible with current Ethernet-based DCN

• Low CAPEX/OPEX

• Easy to deploy, easy to reuse the operation capability.

Low Latency Solution : PFC + ECN

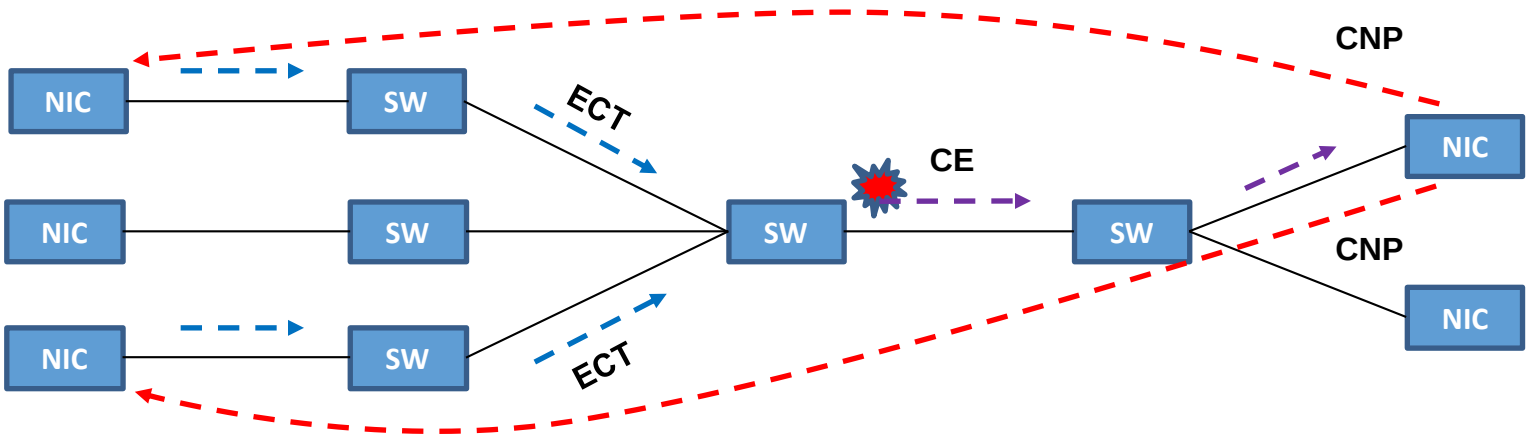
1、PFC(Priority Flow Control) is a kind of back-pressure protocol based on priority queues. Congestion node sends Pause frame to notify upstream node to stop sending to prevent buffer overflow and packet loss.



PFC problems

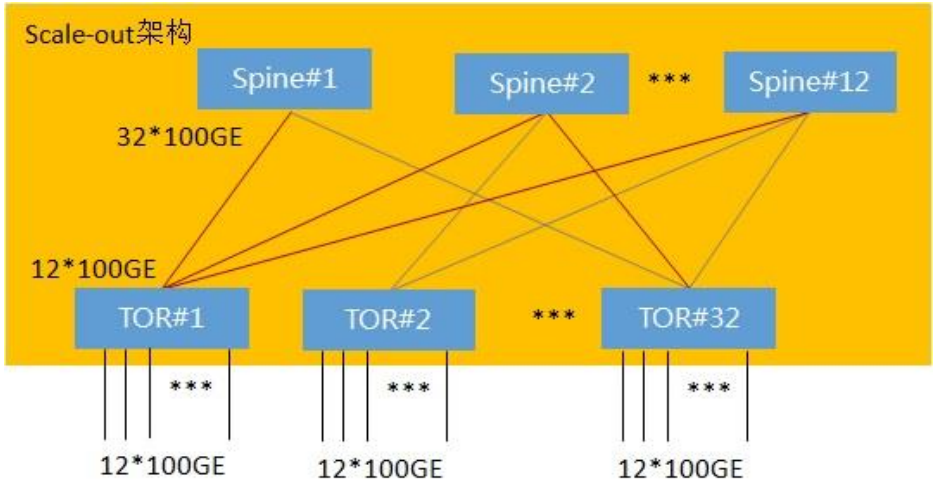
- HOL Blocking
- PFC deadlock
- PFC unfairness
- Pause Storm

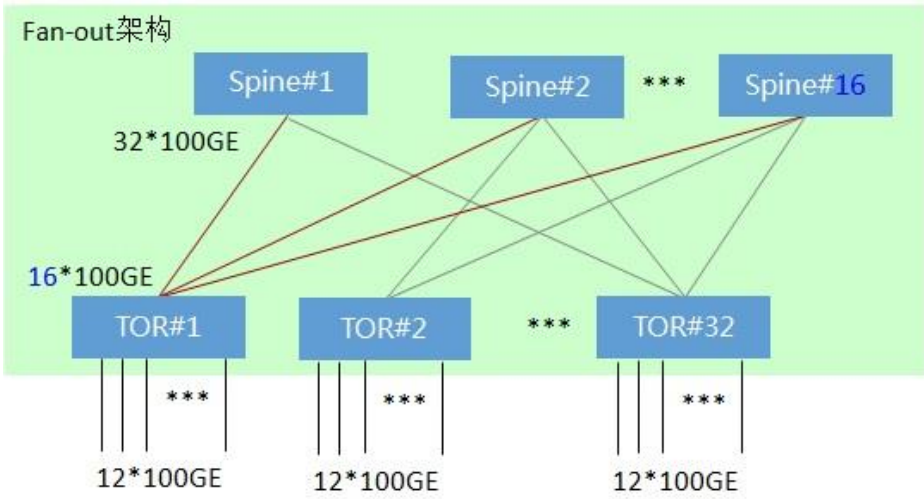
2、ECN(Explicit Congestion Notification) is a kind of end to end congestion control mechanism based on the flow.

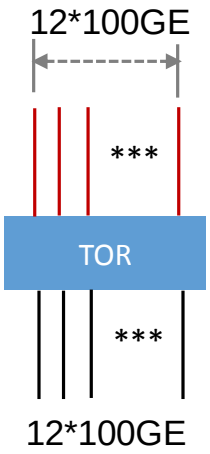


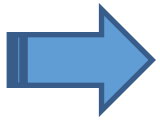
- Long control loop
- Randomness
- Complexity of setting threshold
- Different congestion control mechanisms on NICs

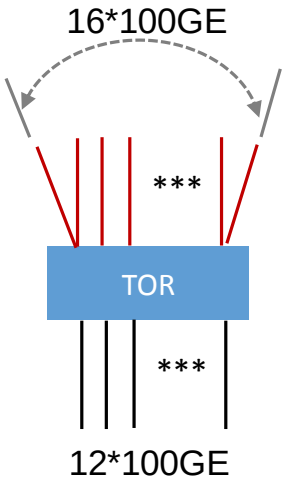
Low Latency Solution : Network Architecture Upgrade









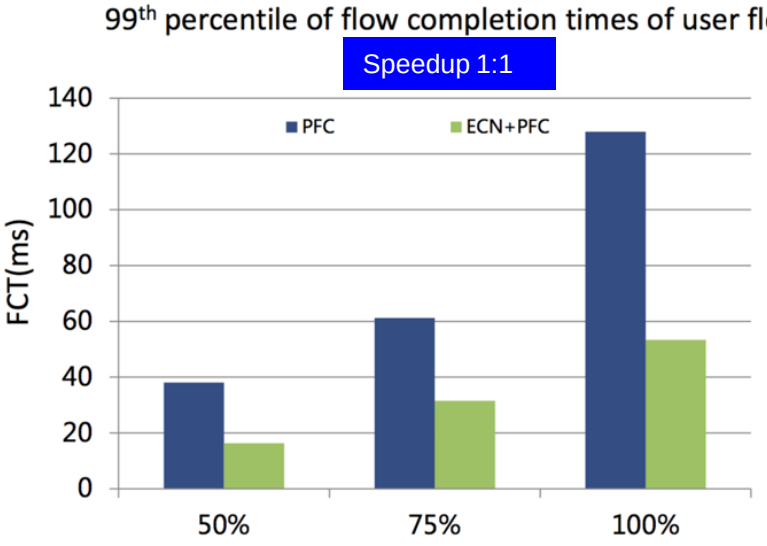


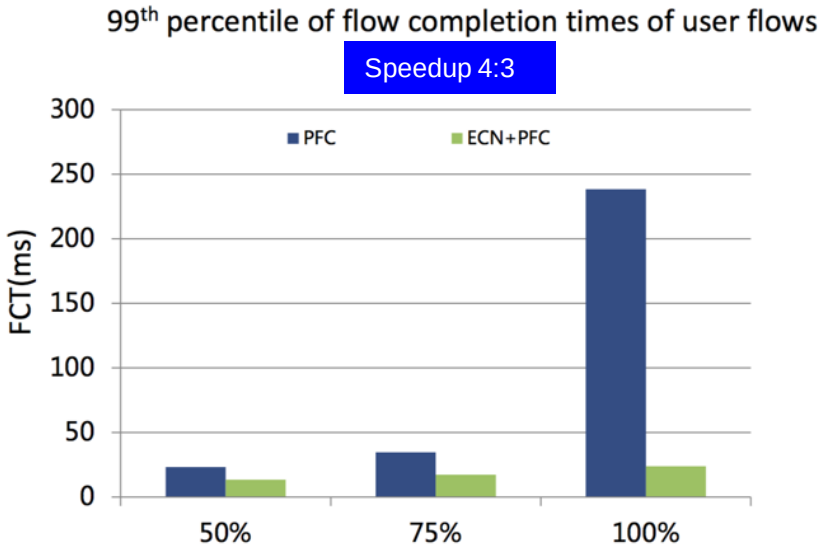
1,Non-blocking when scale-out;
Speedup = 1:1

2, Speedup > 1 when Fan-out;
Speedup = 4:3

Best Practice -1

Evaluation objects:
1. PFC only and ECN+PFC
2. Under different network utilization and speedup ratio



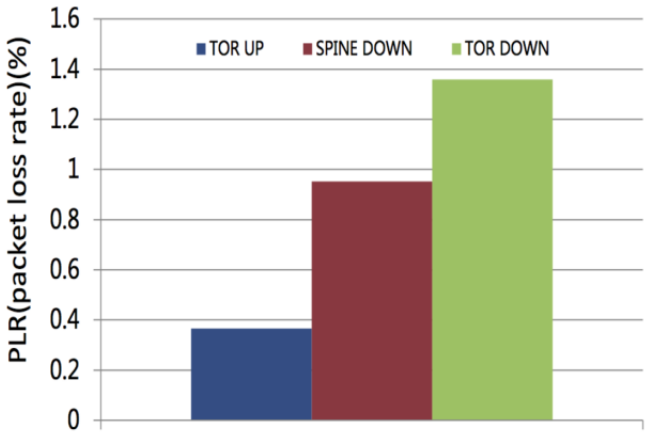


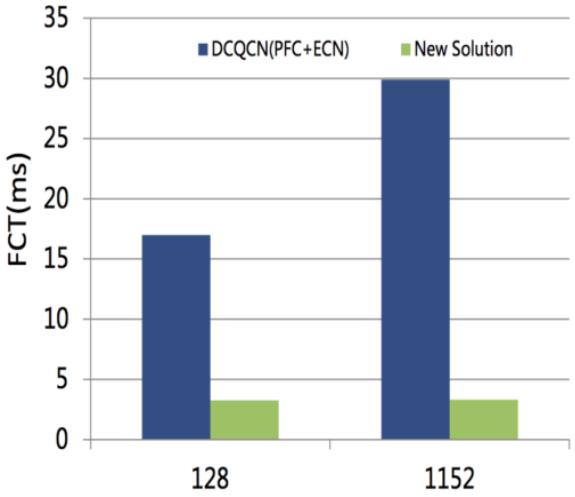
Conclusions：

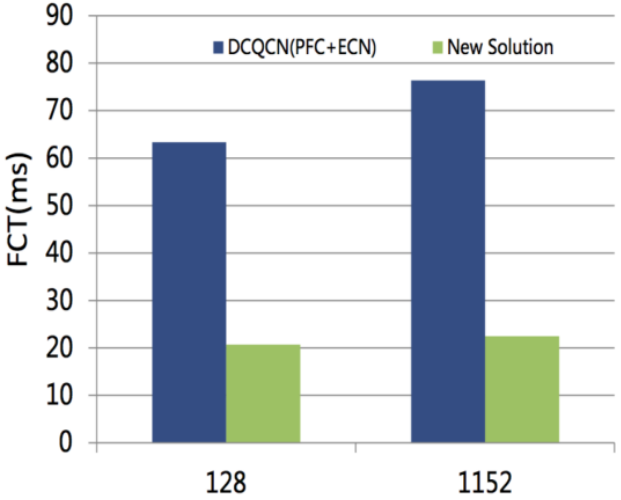
1、ECN+PFC outperforms PFC under different kinds of network utilization.
2、Speedup ratio profits the efficiency of the network: the higher, the better.
3、Threshold should be configured properly：provided the headroom, PFC threshold should be set as high as possible. ECN threshold should be set based on traffic pattern.

Best Practice - 2

Evaluation objects:
1. DCQCN：PFC+ECN
2. New Solution：TOR downlinks enable ECN, Per-Packet Load Balancing







Conclusions：

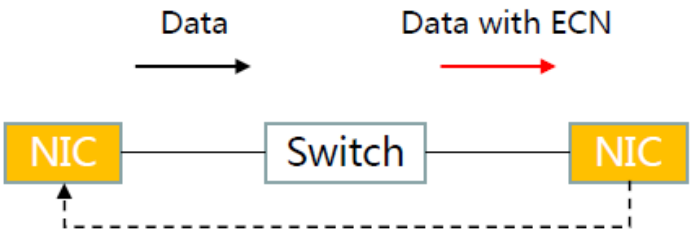
1、Need to involve an ideal load balancing algorism: increase the speedup ratio could mitigate the congestion of Fabric's internal ports, but the packet loss caused by uneven distribution of traffic still exists.

02

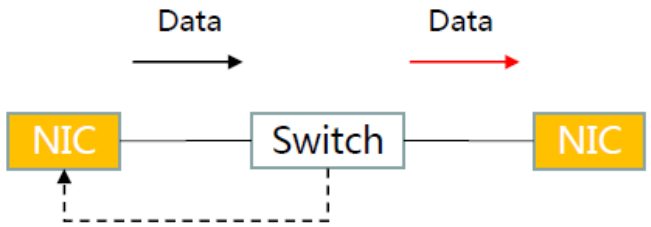
Innovations on Low Latency Network Technology

1. Control Plane – Feedback Mechanism
2. Data Plane – Multipath Load Balancing
3. Management Plane – Self Adaptive Network
4. Function enhancement : Queuing Optimization

Control Plane – Feedback Mechanism Optimization



Traditional Congestion Notification



Congestion Notification / Packet Loss Notification

Feedback info is simple
• Only mark congested/uncongested, no quantized congestion information.

Notification loop is long
• NIC generates the congestion notification, the control loop is long.
• Congestion notification packet is mixed with normal traffic, without prioritization design.

Notification Message improvement
• Involve congestion notification mechanism with more quantized levels, not two status.

Multiple ways to accelerate
• Switch feedback the congestion/packet loss directly, shorten the control loop
• Set a higher priority to notification message
• TCP fast retransmission

Data Plane – Multipath Load Balancing

The traditional hash algorithm distributes traffic unevenly

• In multi-path scenario, as using flow 5-tuple based hash algorism, elephant flows may map to the same link, introducing persistent congestion on the link.

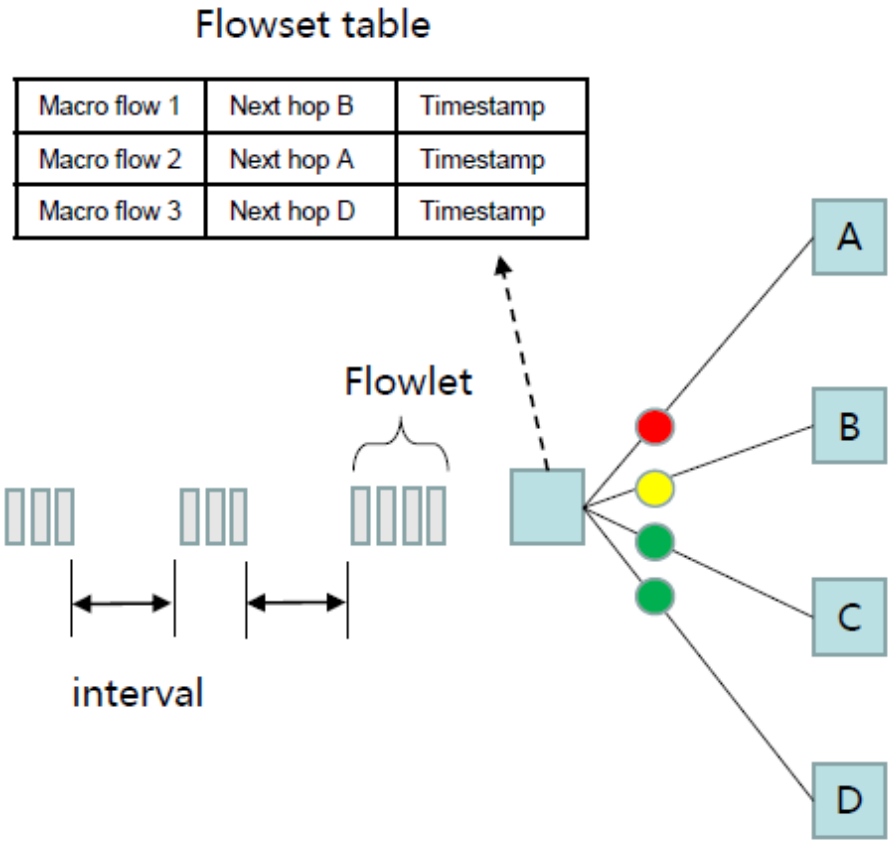
New multi-path load balancing

• Select a idle path based on measured load of multi paths

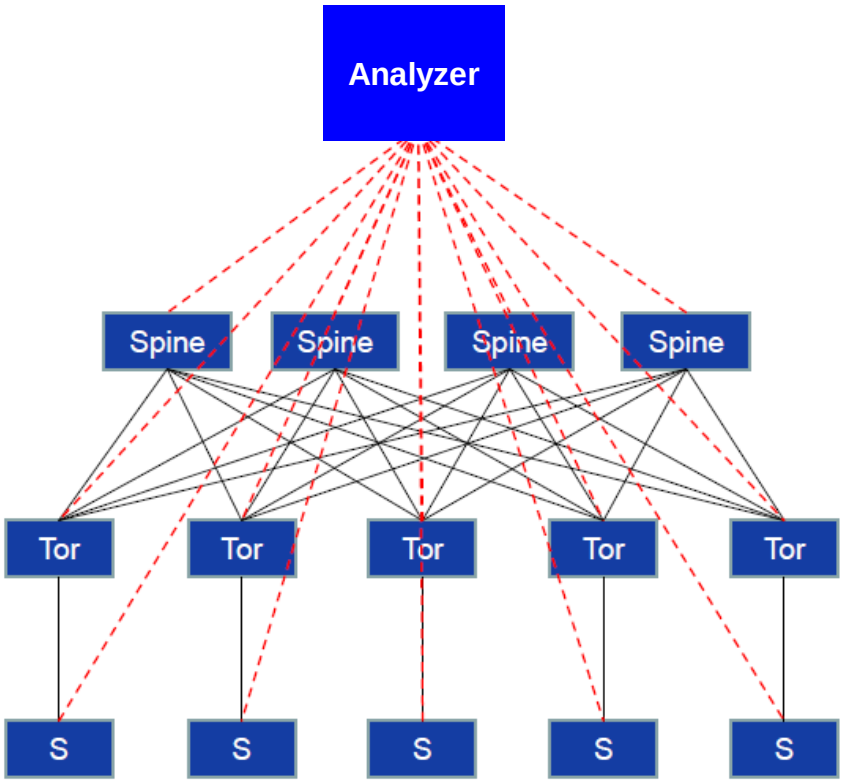
• Use the length of the egress queue as a hash key of load balancing algorithm

• Cut elephant flows into flowlets, schedule to different paths and make sure no out-of-order.

Dynamic load balancing



Management Plane – Adaptive Network



Low latency network puts forward higher requirements for operation and maintenance management automation

• According to the severe requirements of packet loss and latency, the network configuration needs to be dynamically adapted to ensure the online configuration is always best.

Effect of the Adaptive Network

1. Detection and discovery

• Traffic measurement, mark the information along the network nodes (timestamp, ingress port, egress port, queue)

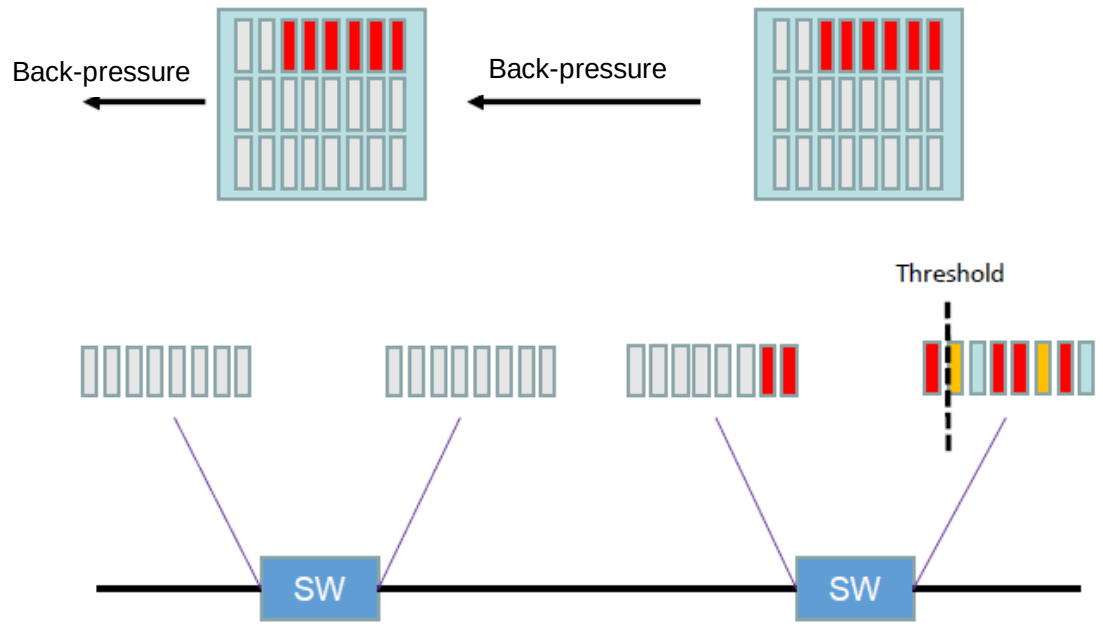
2. Computing and characteristic analysis

• Analyze real-time service characteristics, calculate the optimal scheduling strategy

3. Instruction distribution and continuous optimization

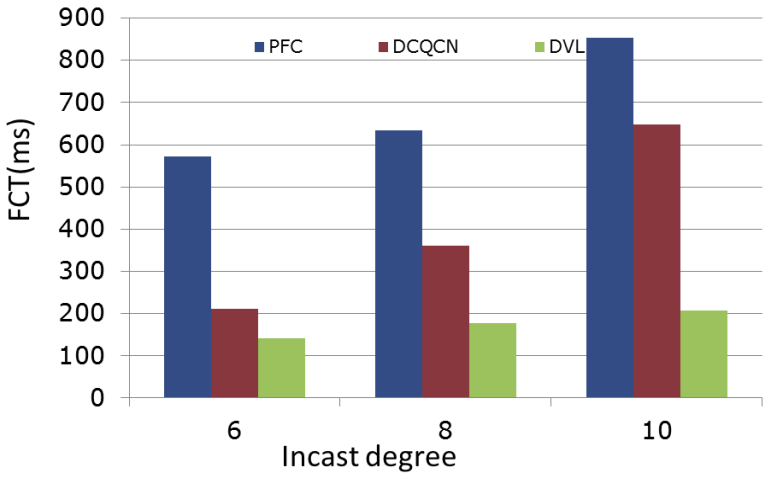
• According to the traffic pattern, self configure and dynamically tune the parameters.

Function enhancement : Queuing Optimization



Advantages of technical solutions

- low latency: isolate the congested flows, make non-congested flows low latency.
- High throughput: buffer congested flows a long the path, fully utilize the link capacity, not slow down the mice flow
- Quick response: zero hop response



Traffic characteristics

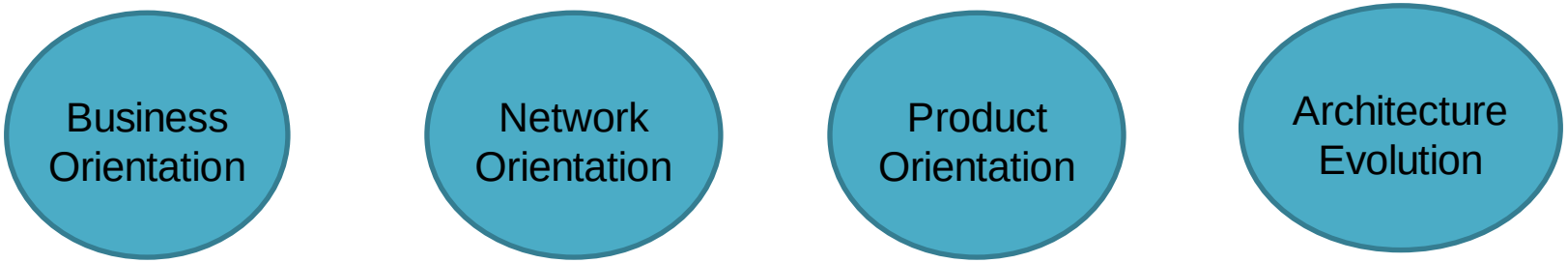
Elephant flows: contribute 80 percent of total traffic. Packet loss has little influence to the whole performance. Latency non-sensitive.

Mice flows: contribute 20 percent of the data traffic load. Packet loss has serious influence to the whole performance. Latency sensitive.

03

Summary

Summary



Internal requirements from Baidu Cloud & Artificial intelligence applications

Under the overall layout of the network, achieve network acceleration within the partial data center network

Promote industrial development, products need to be optimized

Invest in small-scale . Optimize and iterate gradually

THANKS