2015中国DPDK开发者大会
China DPDK Summit 2015

Optimize Cloud Infrastructure with DPDK

孙成浩

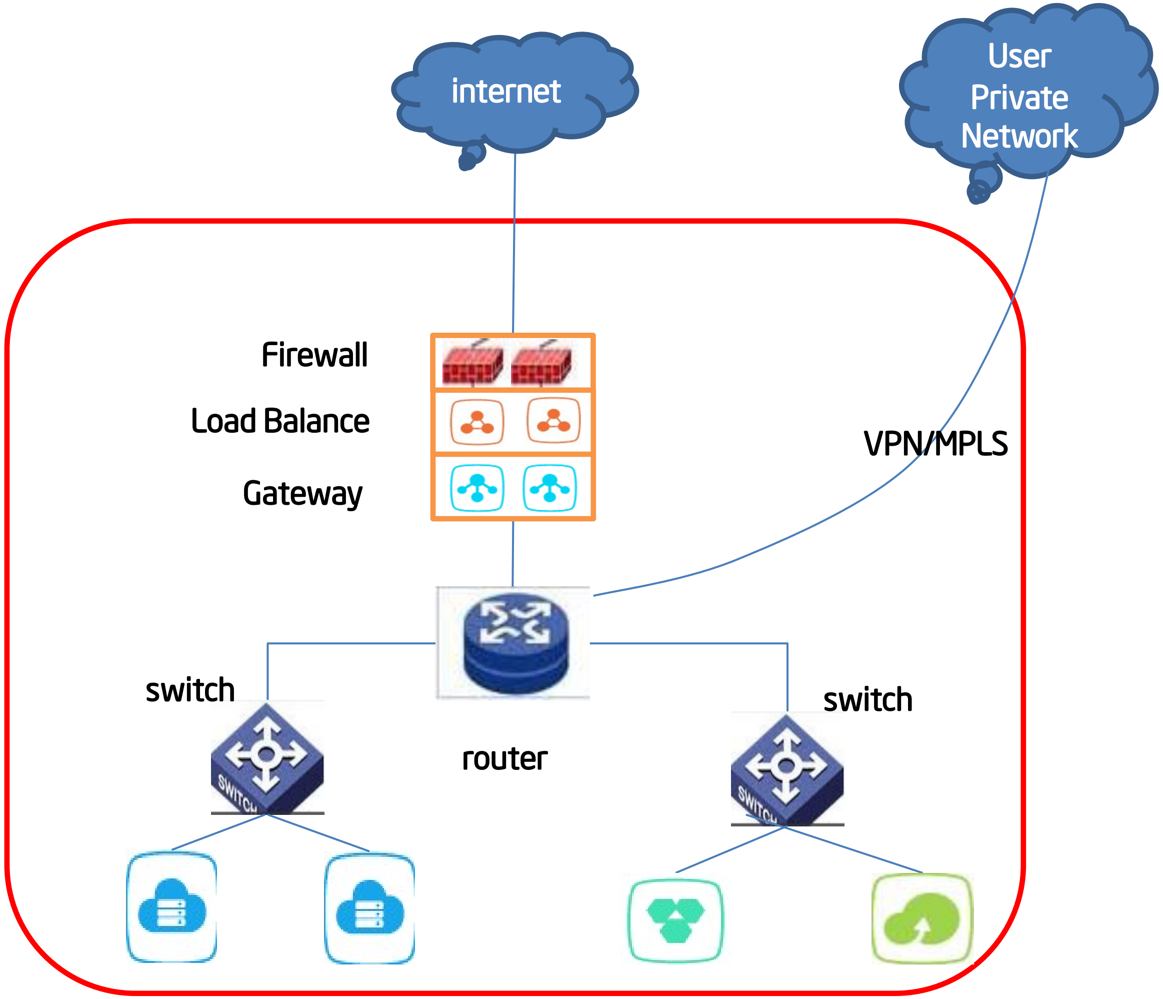
阿里巴巴技术保障部

2015.04.21

Agenda

- Cloud
- Stack
  - Layer2/Layer3
  - Layer4/Socket
  - Local Stack
- Application
  - Flow
  - VxLAN Gateway
  - Security
- Deployment
- Future Work

Cloud



Key Feature:

Inexpensive
Elastic/HA

Why DPDK

challenge:

- stable
- massive
- throughput
- latency
- flexible

?

DPDK Has:

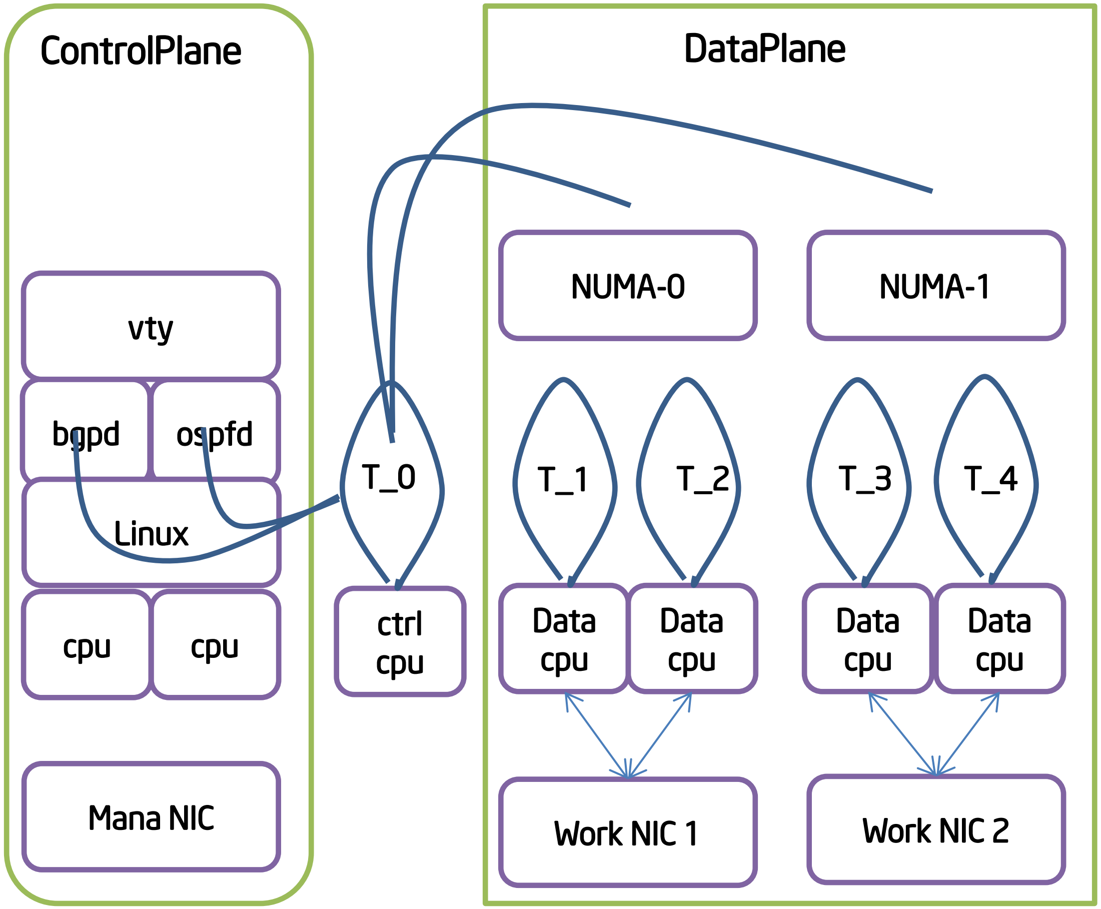
- User Space PMD Driver
- Run-to-completion Dataplane
- High Performance Libraries
- Developing Easily
- Cheap Servers

+

Our Work:

- TCP/IP Stack
- Application
- Deployment

Layer2/3 Stack: Fast



Light DP

- Run-to-completion
- No interrupt
- No preempt
- No Syscall
- Lock free

Coding

- CPU cache
- NUMA
- Per Core

Algorithm

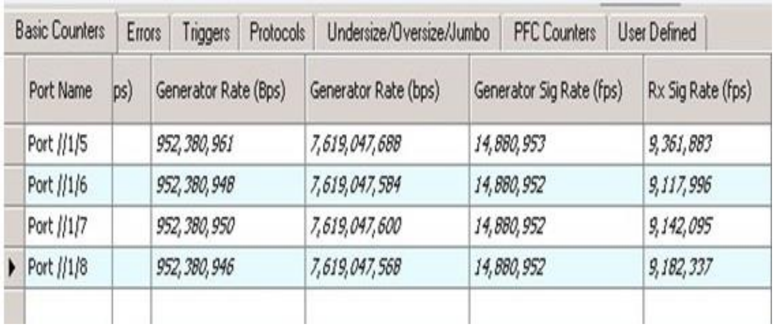
- hash
- LPM
- QOS
- Harp/Hipac

Throughput and Latency(82599)

Out stack supports:

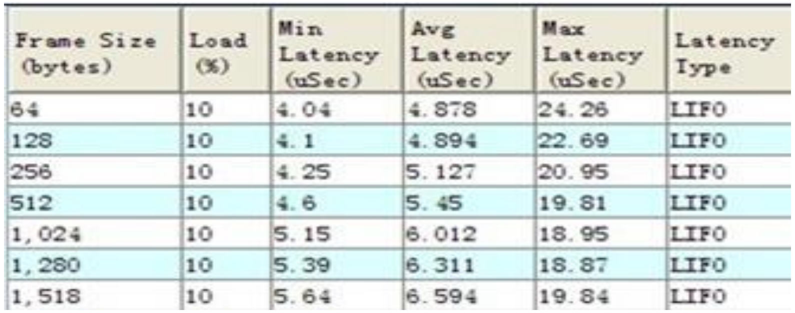
- vlan/bond/router
- mac table(16K)
- Arp table(16K)
- Route(10M)
- Multiple protocols(IPv6...)

64Byte(60%, l3 forward)



Port Name	ps)	Generator Rate (Bps)	Generator Rate (bps)	Generator Sig Rate (fps)	Rx Sig Rate (fps)
Port //1/5		952,380,961	7,619,047,688	14,880,953	9,361,883
Port //1/6		952,380,948	7,619,047,584	14,880,952	9,117,996
Port //1/7		952,380,950	7,619,047,600	14,880,952	9,142,095
Port //1/8		952,380,946	7,619,047,568	14,880,952	9,182,337

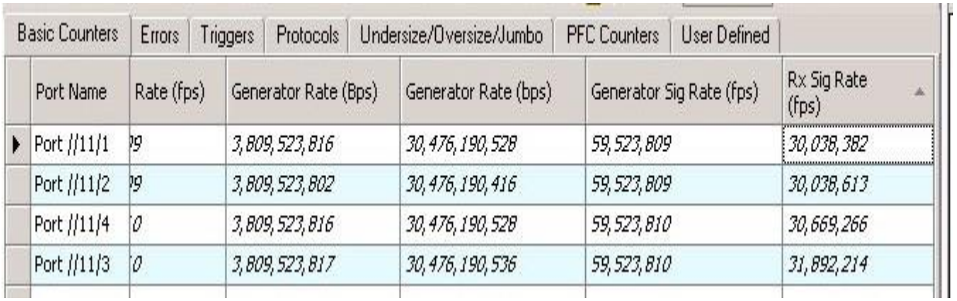
Latency(2544)



Frame Size (bytes)	Load (%)	Min Latency (uSec)	Avg Latency (uSec)	Max Latency (uSec)	Latency Type
64	10	4.04	4.878	24.26	LIFO
128	10	4.1	4.894	22.69	LIFO
256	10	4.25	5.127	20.95	LIFO
512	10	4.6	5.45	19.81	LIFO
1,024	10	5.15	6.012	18.95	LIFO
1,280	10	5.39	6.311	18.87	LIFO
1,518	10	5.64	6.594	19.84	LIFO

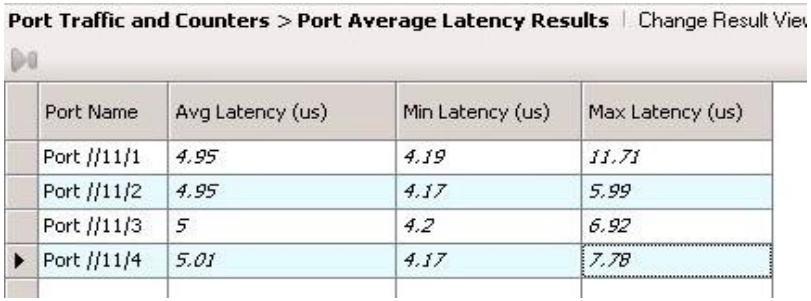
Throughput and Latency(Fortville)

64Byte(50%,l3 forward)



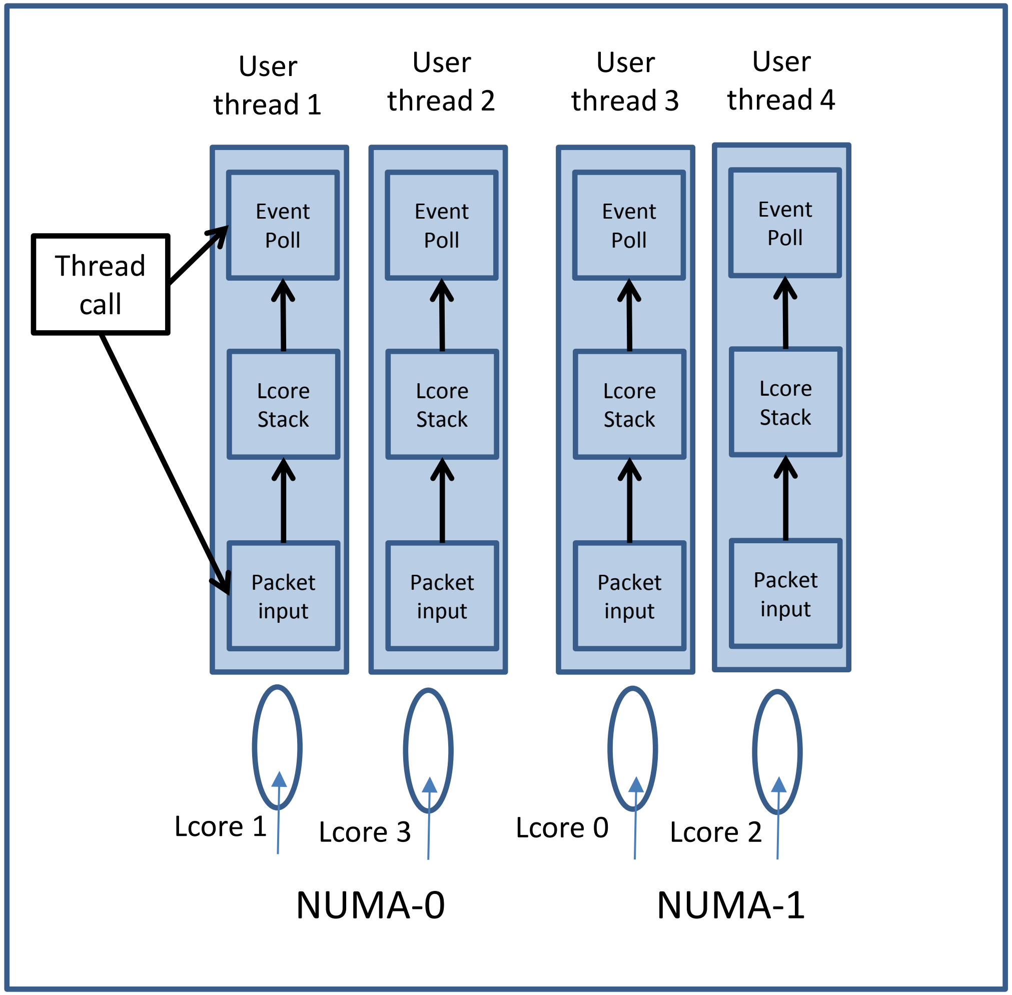
Port Name	Rate (fps)	Generator Rate (Bps)	Generator Rate (bps)	Generator Sig Rate (fps)	Rx Sig Rate (fps)
Port //11/1	9	3,809,523,816	30,476,190,528	59,523,809	30,038,382
Port //11/2	9	3,809,523,802	30,476,190,416	59,523,809	30,038,613
Port //11/4	0	3,809,523,816	30,476,190,528	59,523,810	30,669,266
Port //11/3	0	3,809,523,817	30,476,190,536	59,523,810	31,892,214

Latency


Port Traffic and Counters > Port Average Latency Results | Change Result Vie

Port Name	Avg Latency (us)	Min Latency (us)	Max Latency (us)
Port //11/1	4.95	4.19	11.71
Port //11/2	4.95	4.17	5.99
Port //11/3	5	4.2	6.92
Port //11/4	5.01	4.17	7.78

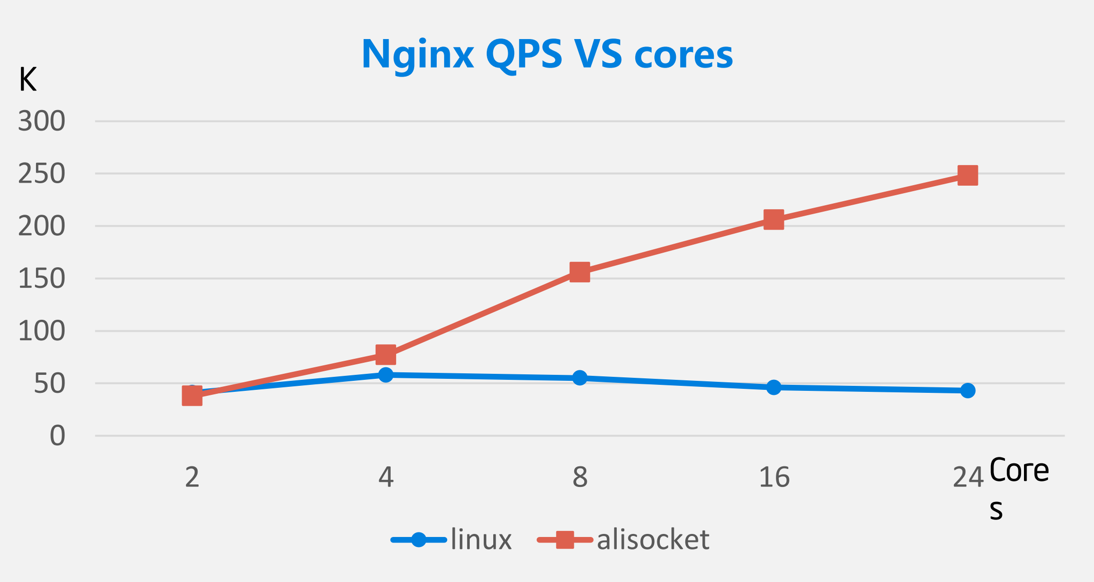
Layer4/socket



Key Feature:

Per thread listen
Per core flow table
Run-to-completion
Syscall hijack

QPS



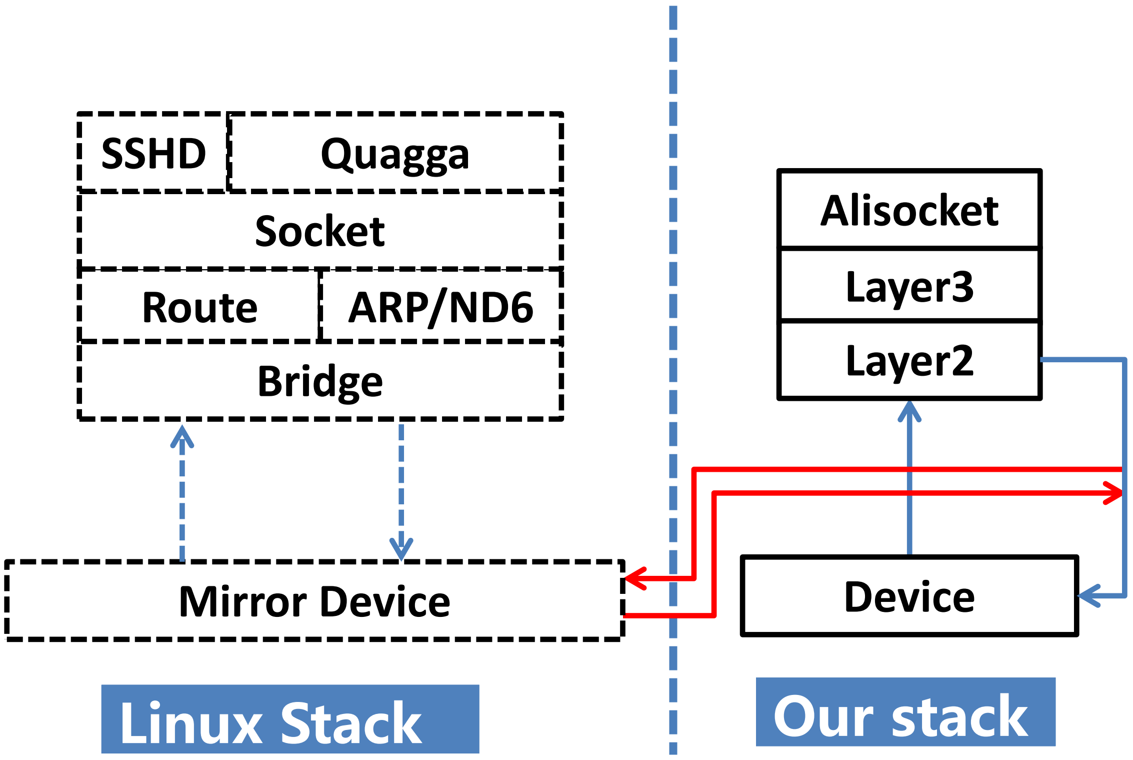
DUT：

- Xeon E5-2630 @2.30GHz
- 82599 10G X 2
- Linux 2.6.32-131.21.1.tb93
- Nginx 1.7.8
- Page size 612Bytes

Test Tools：

- Spirent Avalanche

Linux Stack: Simple



We Support
Everything Linux
Supports

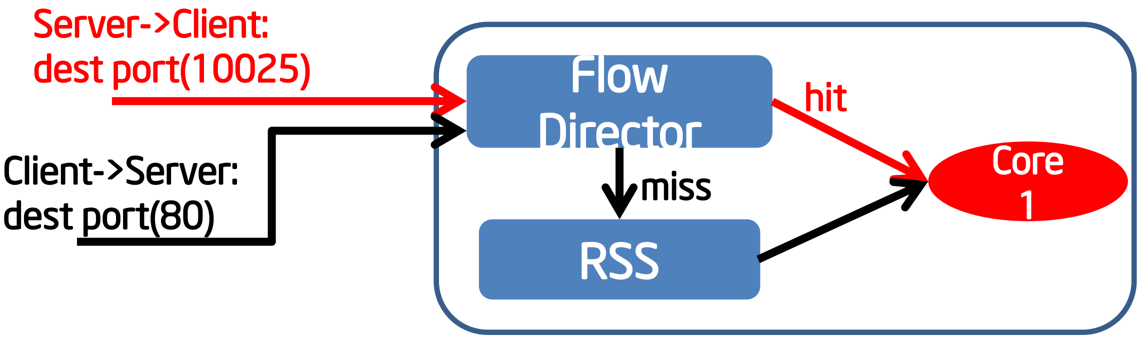
Snat: flow

Flow Dir

- 1W—2W(core 1)
- 2W—3W(core2)
- 3W—4W(core3)
- 4W—5W(core4)

RSS

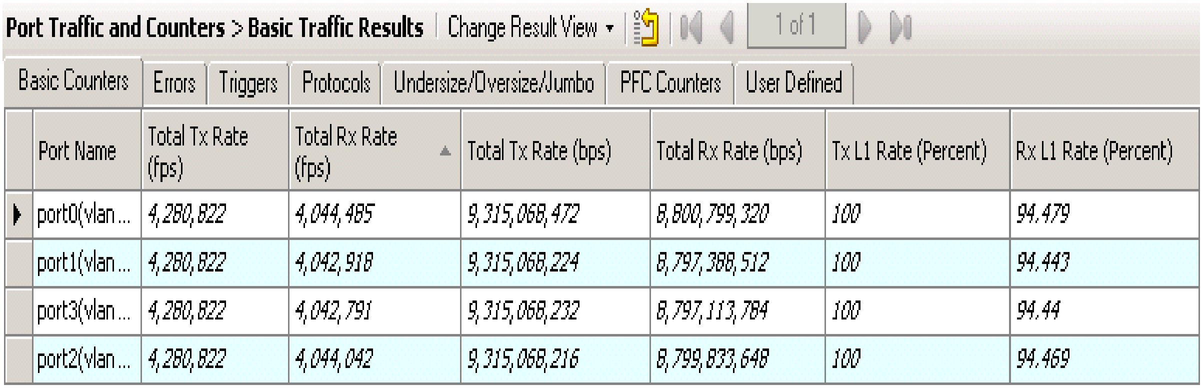
- Source Port
- Same initial key



Performance

- Per core session flow table
- Packets from same flow go to same core in one server
- Packets from same flow go to same core of different server when failover
- Flow sync packets go to same core of different servers.
- Packets from client to server miss Flow Director and match RSS.

cocurrency 800W sessions
200Wsps per server



Port Name	Total Tx Rate (fps)	Total Rx Rate (fps)	Total Tx Rate (bps)	Total Rx Rate (bps)	Tx L1 Rate (Percent)	Rx L1 Rate (Percent)
port0(vlan...	4,280,822	4,044,485	9,315,068,472	8,800,799,320	100	94.479
port1(vlan...	4,280,822	4,042,918	9,315,068,224	8,797,388,512	100	94.443
port3(vlan...	4,280,822	4,042,791	9,315,068,232	8,797,113,784	100	94.44
port2(vlan...	4,280,822	4,044,042	9,315,068,216	8,799,833,648	100	94.469

Cloud: Vxlan Gateway

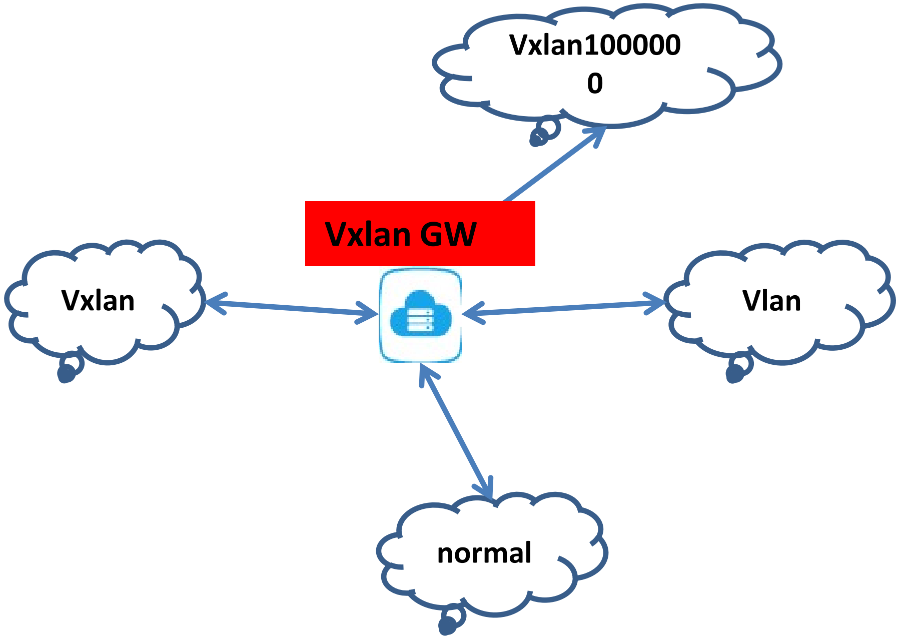
Stack Enhancement:

VxLAN Encap/Decap

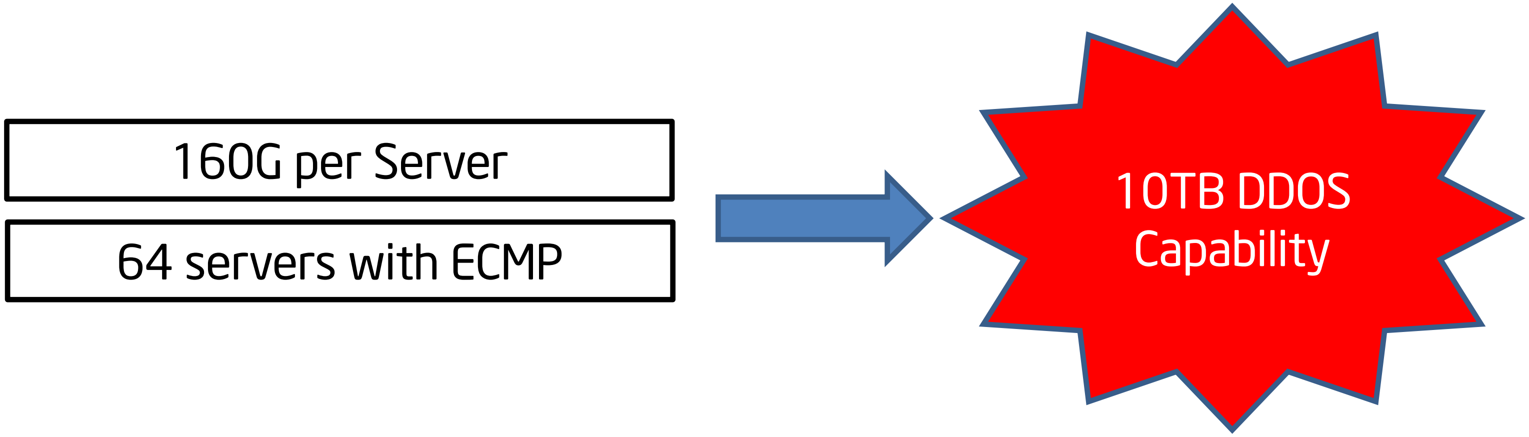
VxLAN Router Interface

VxLAN ACL

Traffic between VxLANs



Security: DDOS



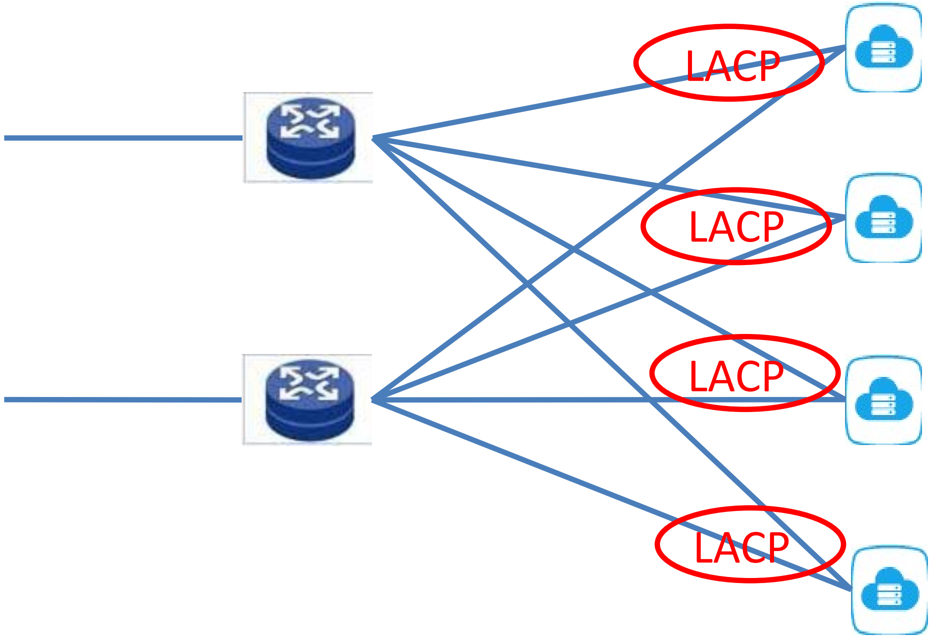
160G per Server

64 servers with ECMP

10TB DDOS Capability

Layer2 HA (lag)

Endpoint Application based on DPDK



LACP

LACP

LACP

LACP

Failover

With LACP, MTTR < 3S

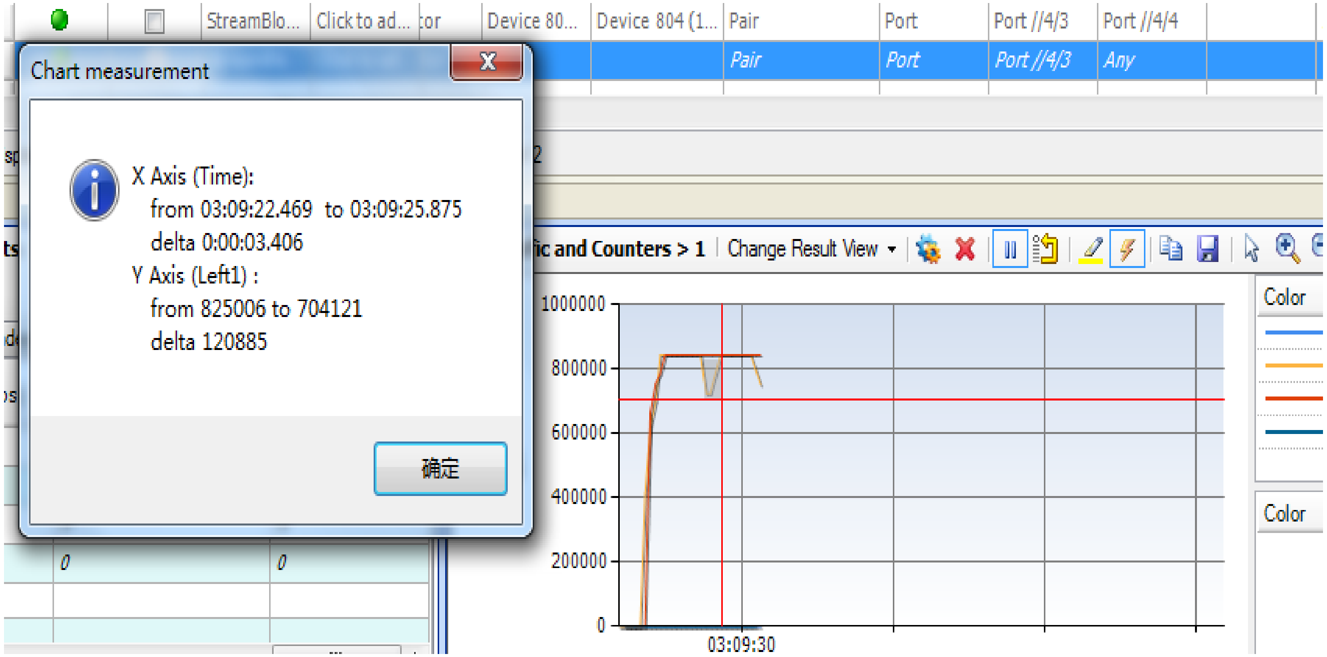
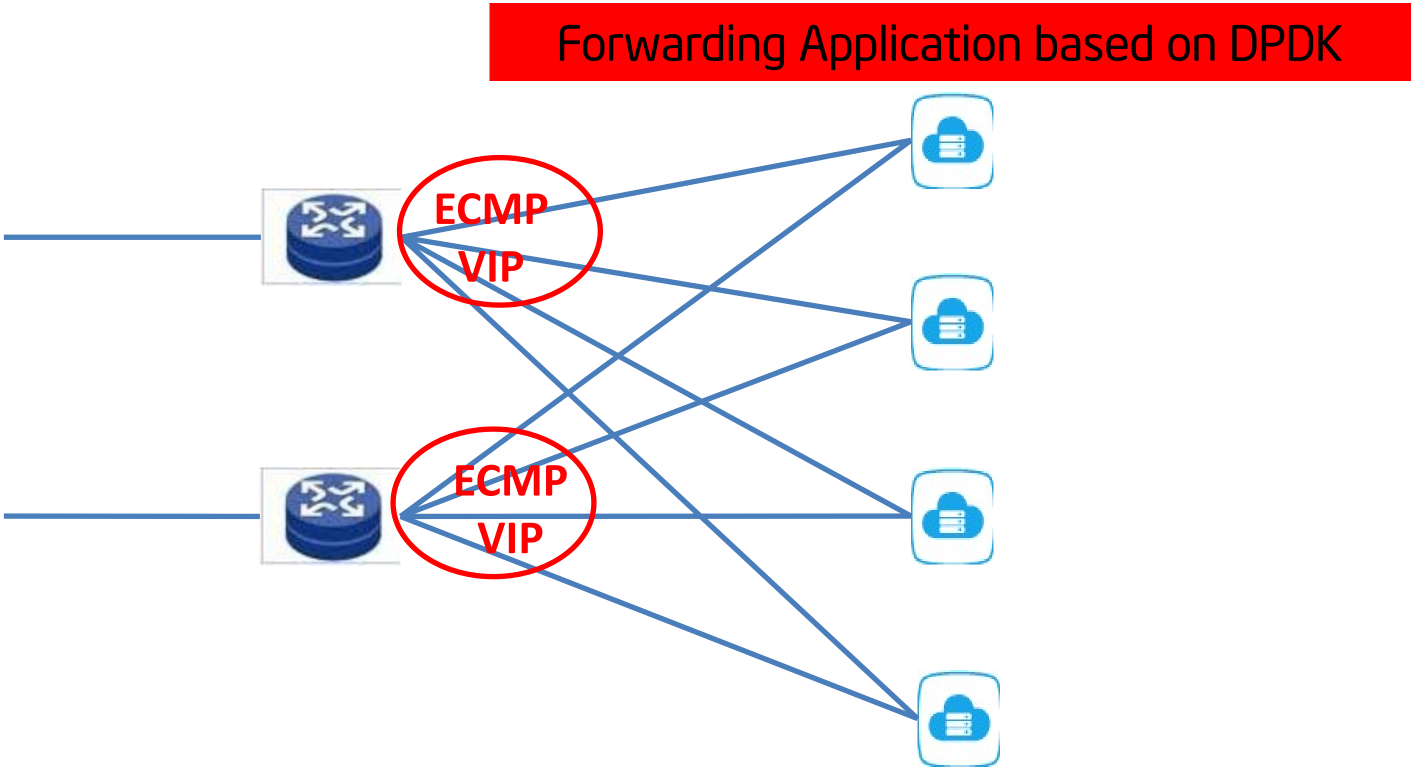


Chart measurement

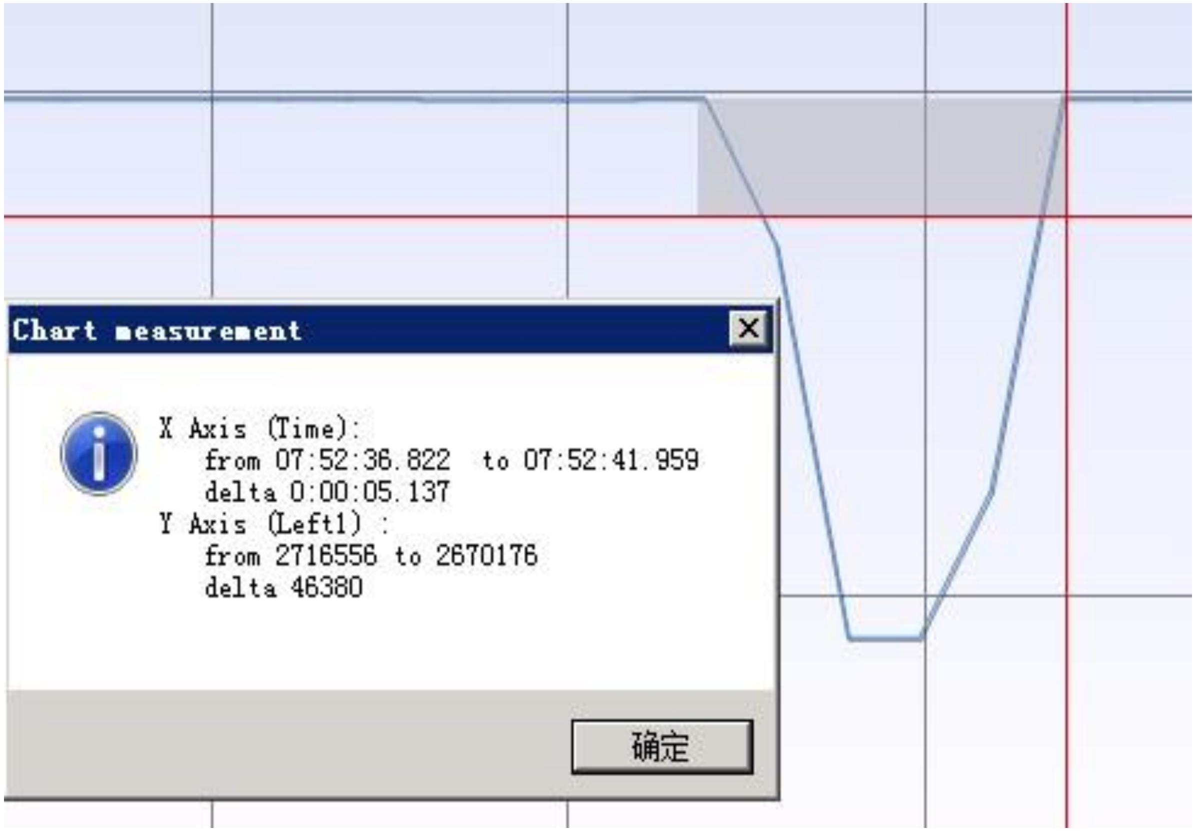
X Axis (Time):
from 03:09:22.469 to 03:09:25.875
delta 0:00:03.406
Y Axis (Left1) :
from 825006 to 704121
delta 120885

Layer3 HA



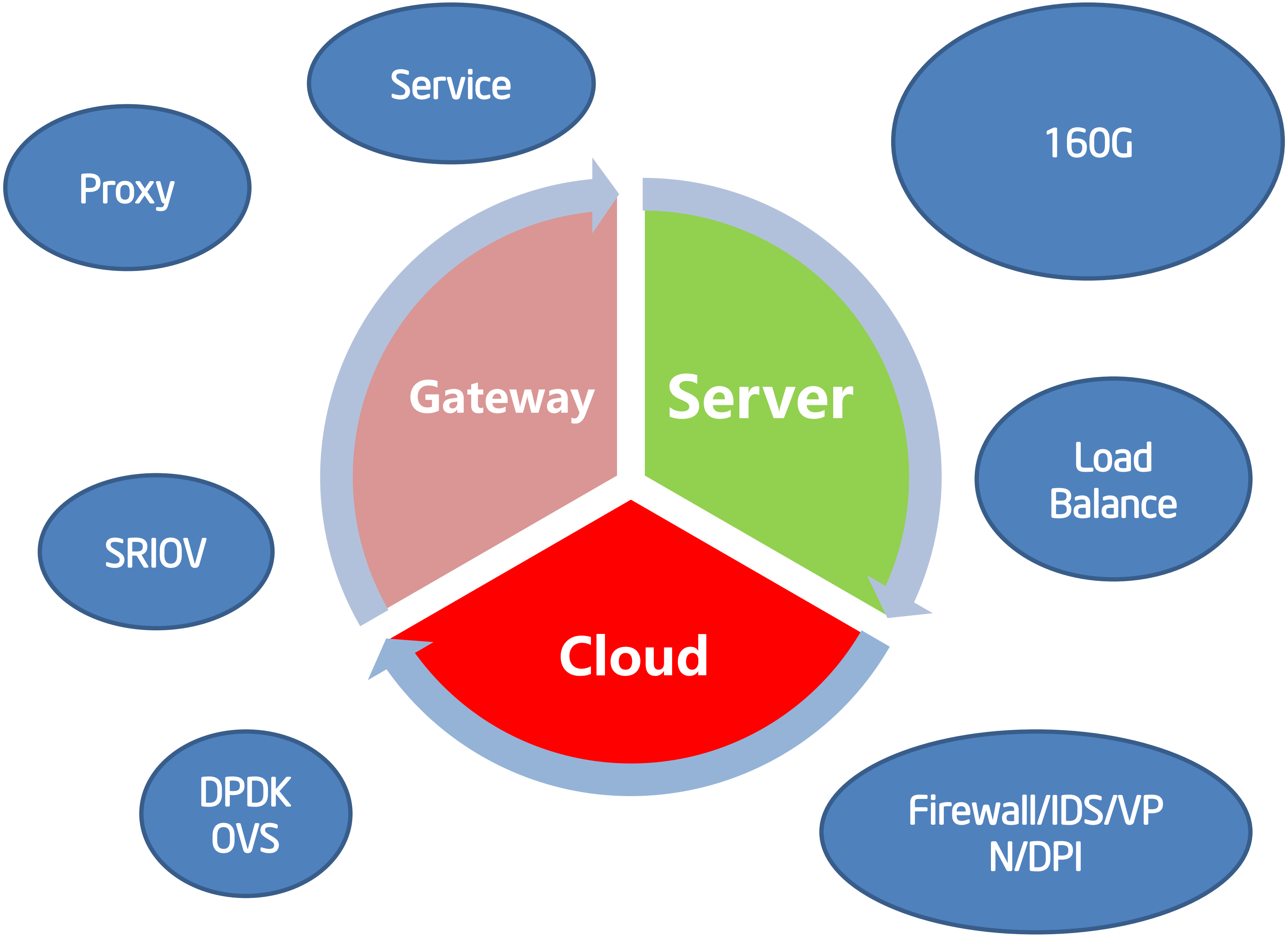
Failover

With OSPF/BGP/VRRP, MTTR < 10S



```
X Axis (Time):
    from 07:52:36.822   to 07:52:41.959
    delta 0:00:05.137
Y Axis (Left1) :
    from 2716556 to 2670176
    delta 46380
```

Future work



Service

Proxy

160G

Gateway

Server

Cloud

Load Balance

SRIOV

DPDK OVS

Firewall/IDS/VPN/DPI

About us

阿里巴巴 Alibaba.com.cn 阿里技术保障 Ali Infrastructure Service

人才	1000+ 工程师 杭州 · 北京 · 上海 · 深圳 · 青岛 · 香港 · 美国
机会	从自学成才的草根，到顶尖学府的精英 阿里为每个人提供发光发热的舞台
前瞻	国家级博士后工作站 现代物理方向

Thanks