Vortex from UCloud

NFV in action on enterprise-grade IaaS cloud computing platform

徐亮 leo.xu@ucloud.cn

Agenda

- What and why is UCloud Vortex?
- How to implement it?
- Lessons learned in Ops

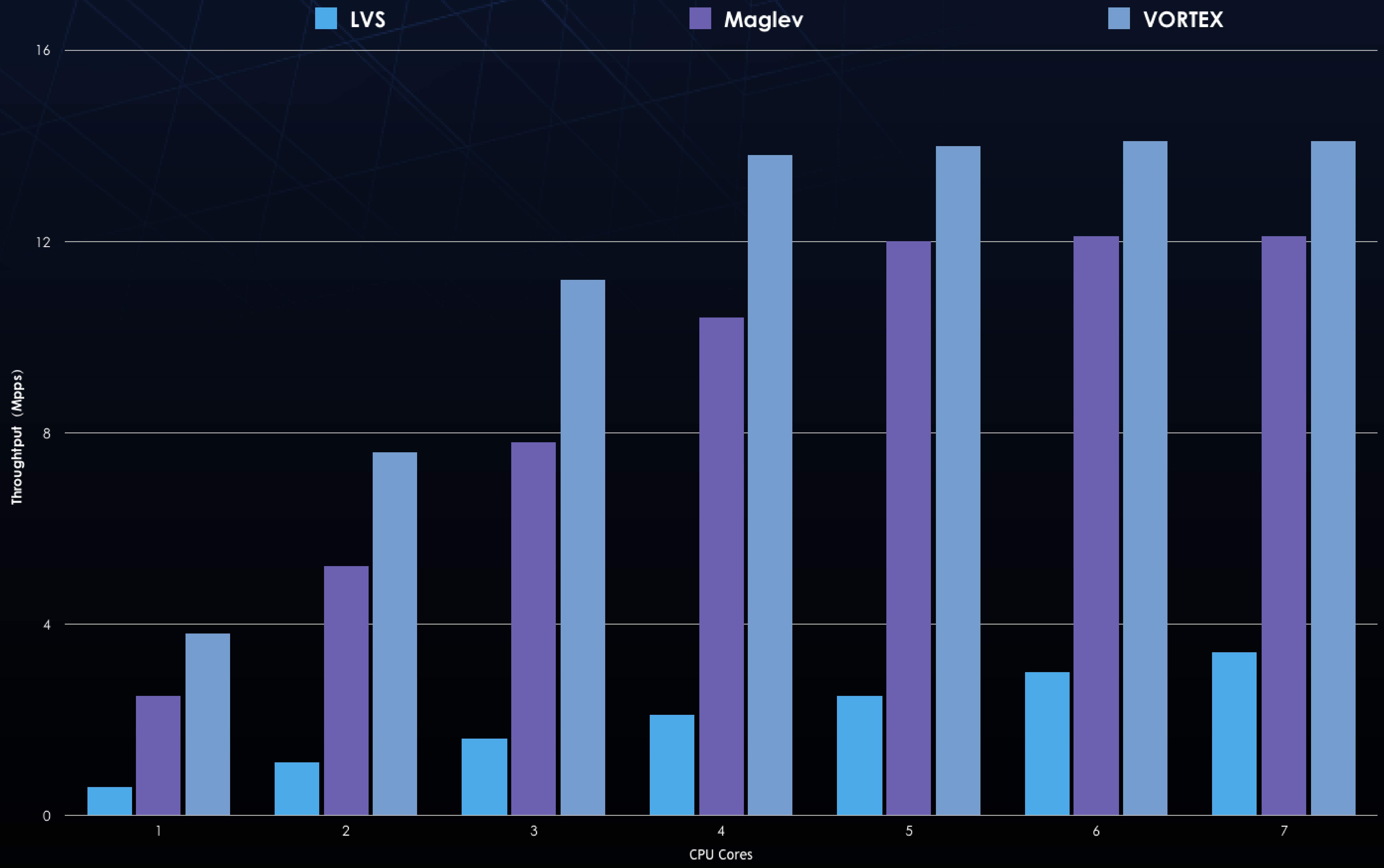
What is UCloud Vortex ?

- UCloud.cn
  - Top Chinese IaaS service provider
  - 10 worldwide data centers
  - Tens of thousands of enterprise users
- UCloud Vortex
  - A layer-4 load balancing software
  - Just like LVS, but in Cloud scale
    - Multiple tenants
    - Overlay Network

Why ？

- Keep It Simple,Stupid!
  - UCloud Vortex = LVS + Multiple Namespaces + OVS
- Easy Customizable
  - Highest Random Weight (HRW) hashing
  - Active - Standby mode backends
- Faster
  - PPS：14M (10G, 64 bytes line rate)
  - CPS：200k+
  - Concurrent Connections：30M+

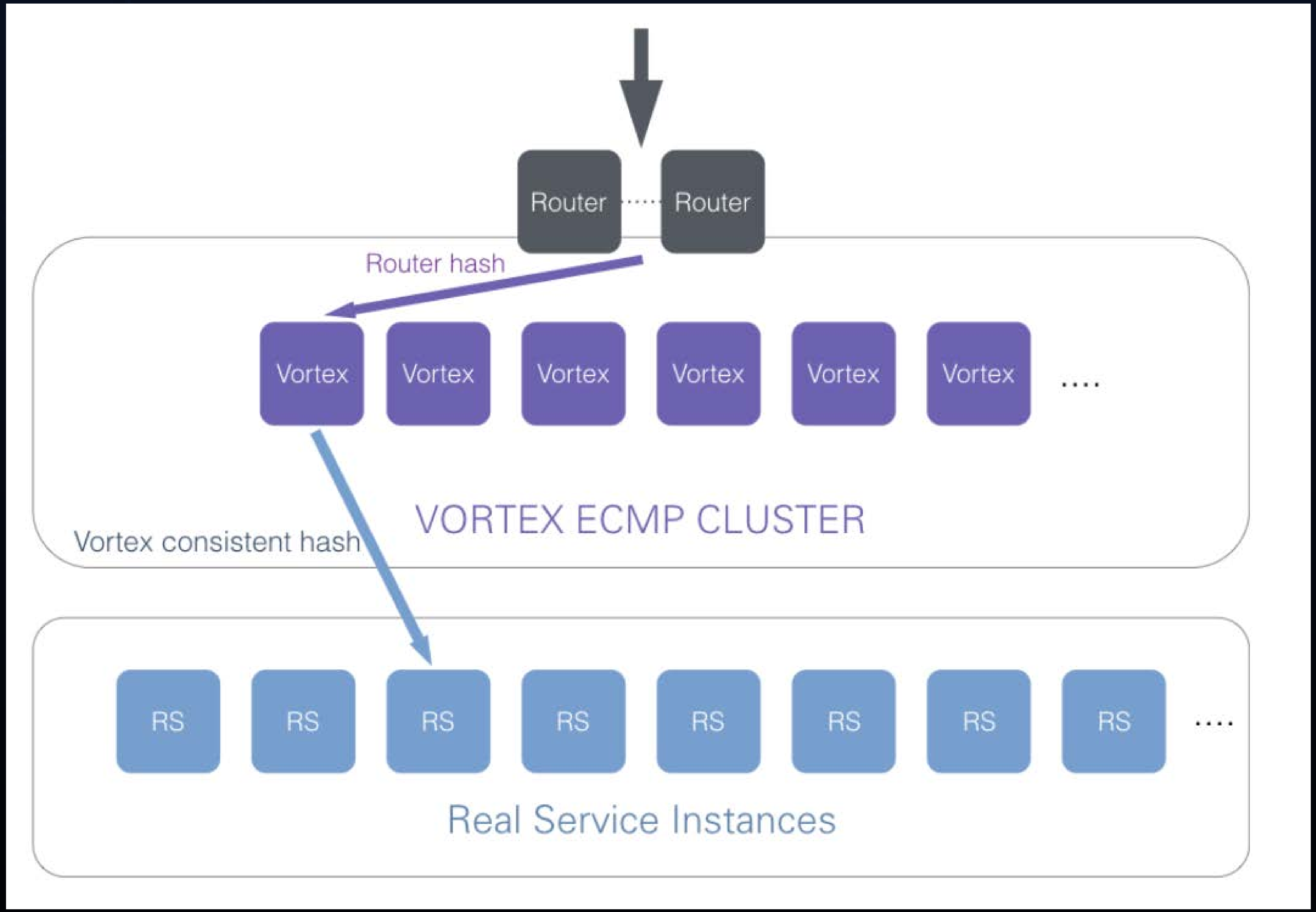
Performance



How ?

- Scale up by share nothing
  - RSS
  - Multiple-queues TAP device
  - Each core owns a session table

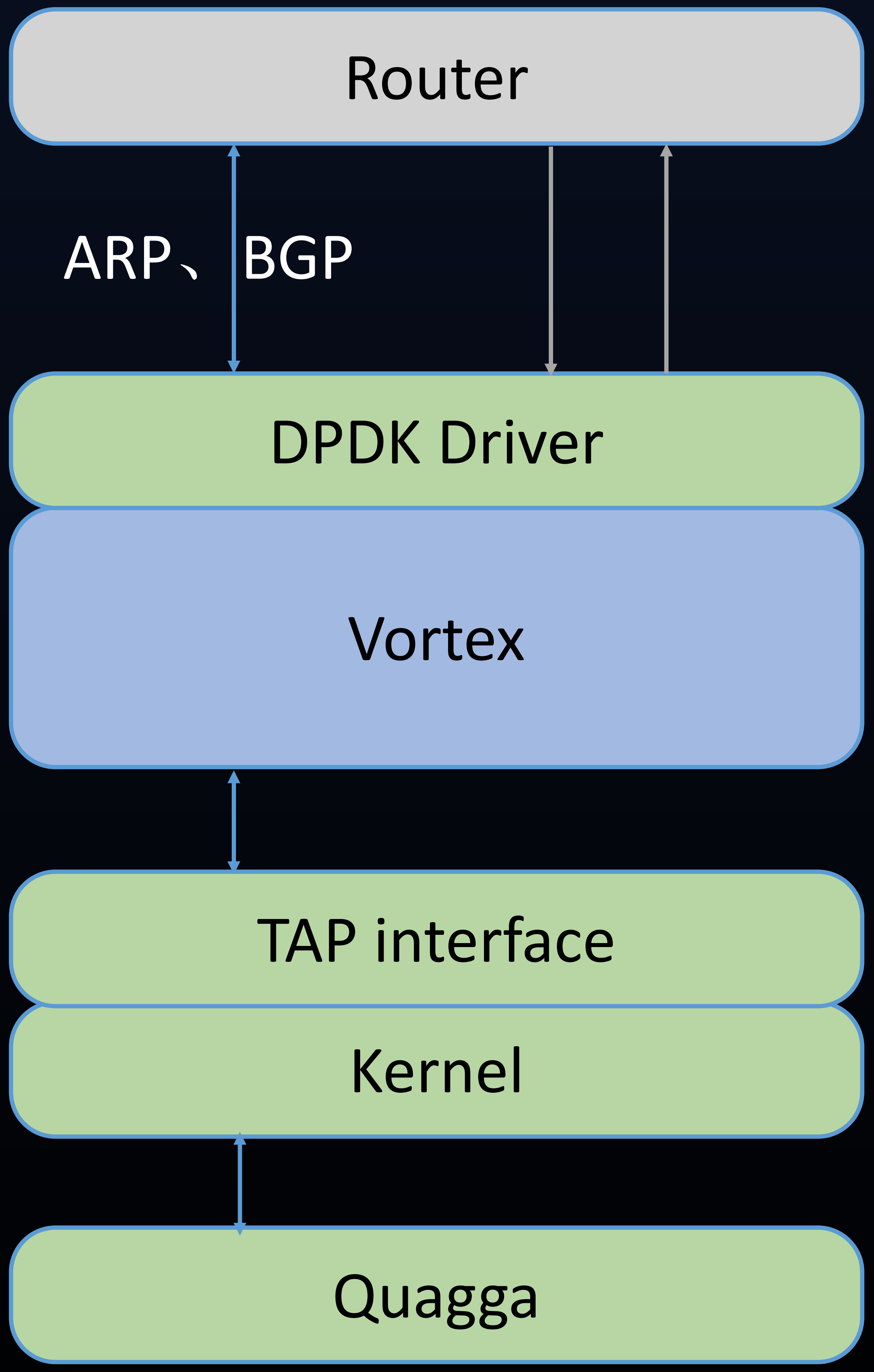
- Scale out by ECMP Cluster



UCLOUD
www.ucloud.cn

How ?

- ARP
  - Forward to kernel
  - Get the MAC address of router by Netlink from kernel neighbors table
- Local IP at TAP interface
  - Forward to kernel
  - BGP is handled by Quagga BGPD
- NVGRE
  - Directly forward by Vortex

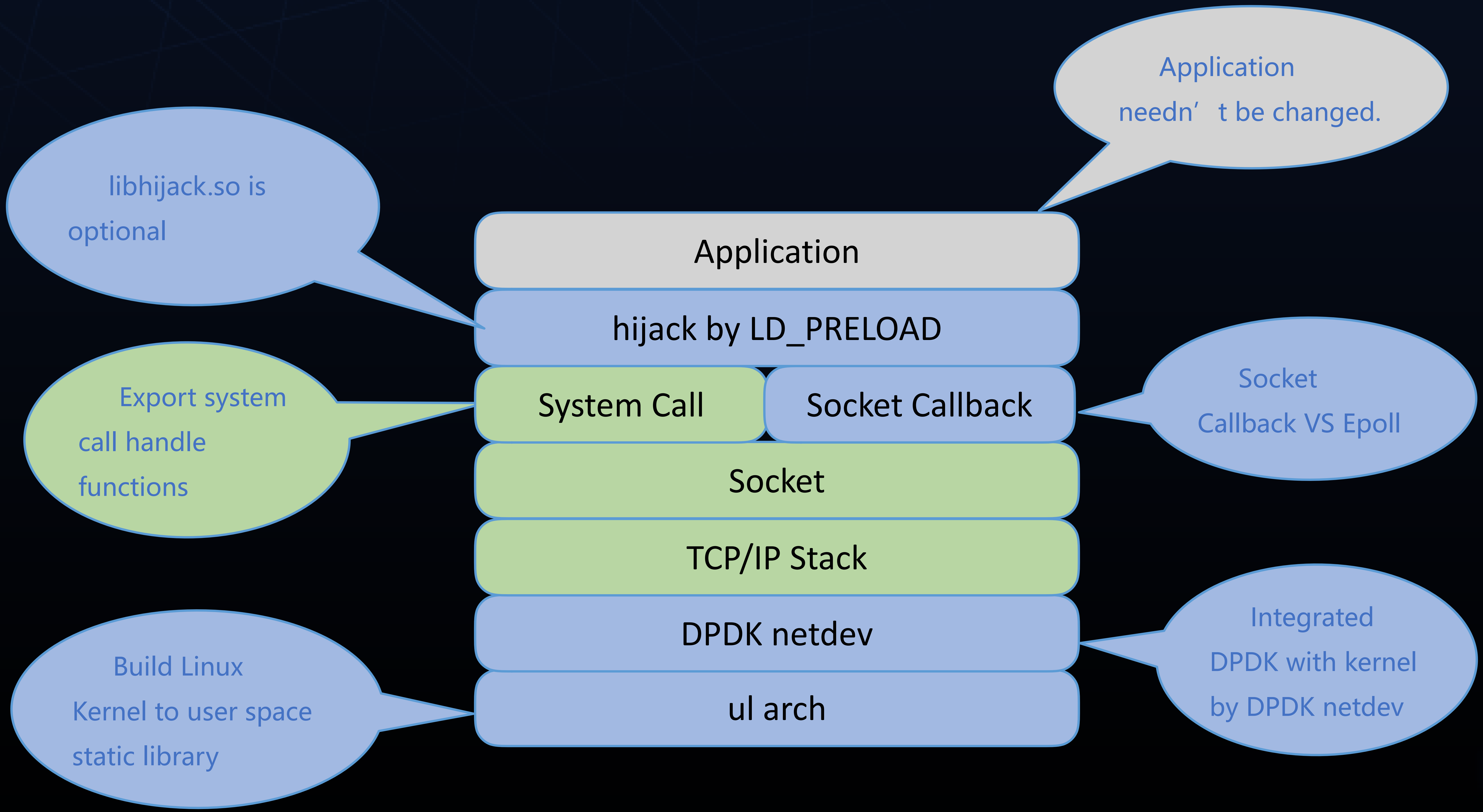


How ?

- HTTPBench
  - A HTTP client and server for Vortex benchmark
  - An Unikernel application with DPDK and Linux TCP/IP stack
  - 50k+ HTTP CPS / core

Libul Architecture

- Inspired by Linux Kernel Library，NS3 DCE, User Mode Linux, RUMP



Application

hijack by LD_PRELOAD

System Call

Socket Callback

Socket

TCP/IP Stack

DPDK netdev

ul arch

Application needn’t be changed.

libhijack.so is optional

Export system call handle functions

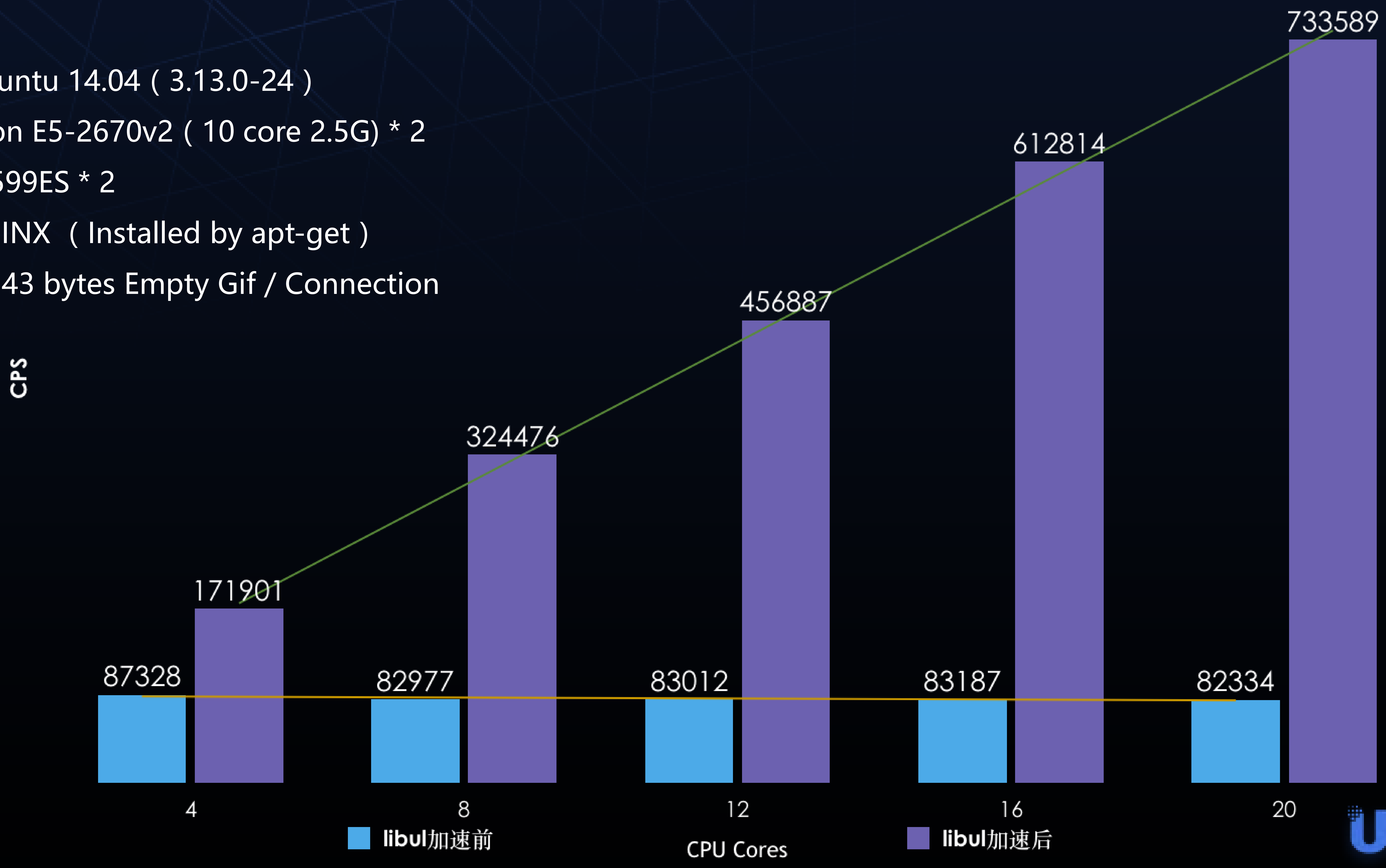
Socket Callback VS Epoll

Build Linux Kernel to user space static library

Integrated DPDK with kernel by DPDK netdev

Libul Performance

- OS : Ubuntu 14.04 (3.13.0-24)
- CPU: Xeon E5-2670v2 (10 core 2.5G) * 2
- NIC: 82599ES * 2
- App : NGINX (Installed by apt-get)
- Request : 43 bytes Empty Gif / Connection



UCLOUD
www.ucloud.cn

Ops - CPU Usage

- CPU usage always 100% in OS
- Calculate CPU load by application
  - Summary effective CPU cycles when received and processes packets
  - CPU load = effective CPU cycles / rte_get_timer_hz
  - Adjust by CPU Frequency

Ops - Power Management

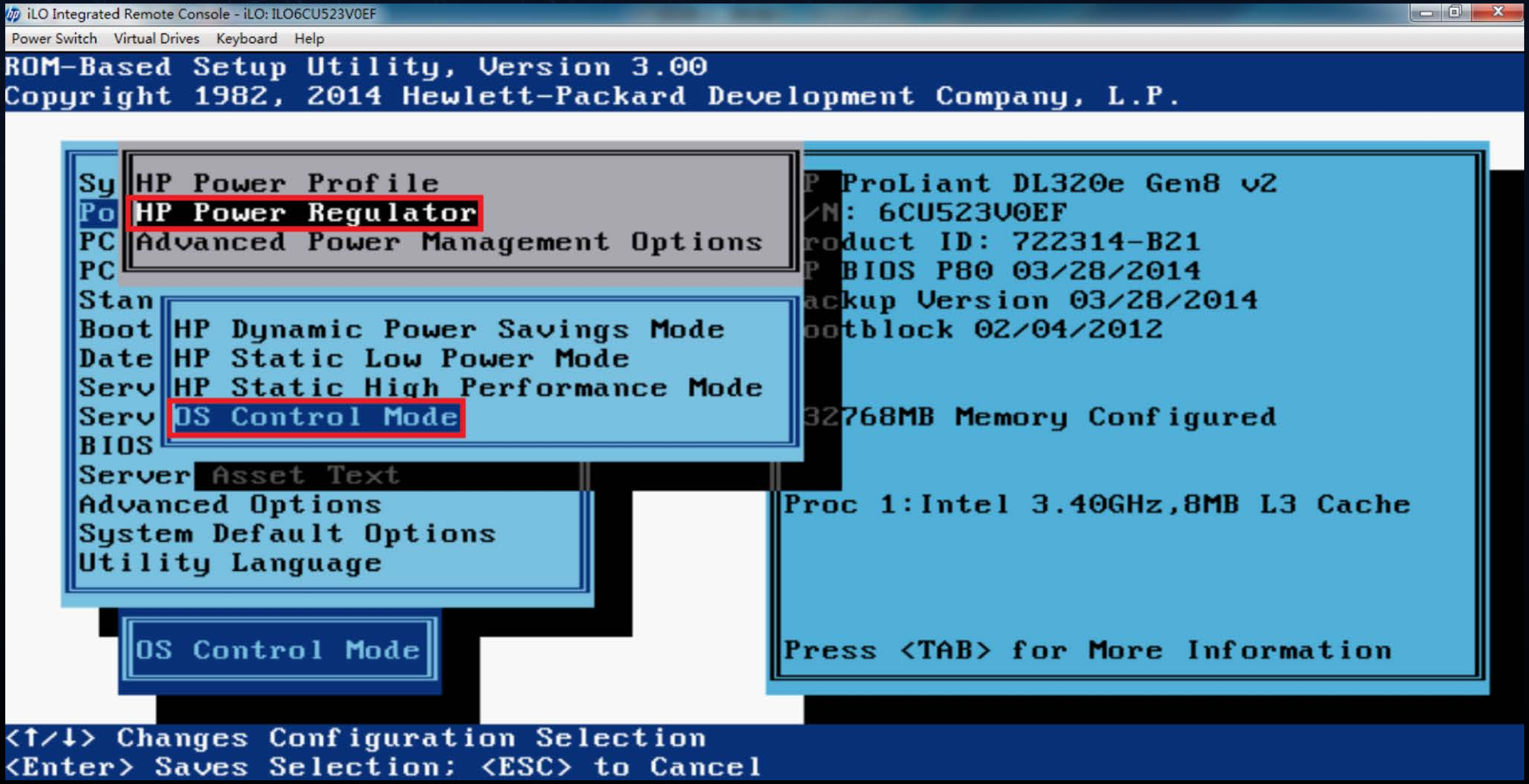
- scaling_available_frequencies is not created in /sys

  POWER: File not openned

  POWER: Cannot get available frequencies of lcore 2

  POWER: Unable to set Power Management Environment for lcore 2
- Set “HP Power Regulator” to “OS Control Mode”
- Add “intel_pstate=disable” into kernel command line

Ops - Power Management



```
iLO Integrated Remote Console - iLO: ILO6CU523V0EF
Power Switch  Virtual Drives  Keyboard  Help
ROM-Based Setup Utility, Version 3.00
Copyright 1982, 2014 Hewlett-Packard Development Company, L.P.

Sy  HP Power Profile                    P ProLiant DL320e Gen8 v2
Po  HP Power Regulator                  /N: 6CU523V0EF
PC  Advanced Power Management Options   roduct ID: 722314-B21
PC                                      P BIOS P80 03/28/2014
Stan                                    ackup Version 03/28/2014
Boot HP Dynamic Power Savings Mode      ootblock 02/04/2012
Date HP Static Low Power Mode
Serv HP Static High Performance Mode
Serv OS Control Mode                    32768MB Memory Configured
BIOS
Server Asset Text
Advanced Options                        Proc 1:Intel 3.40GHz,8MB L3 Cache
System Default Options
Utility Language

     OS Control Mode                    Press <TAB> for More Information

<↑/↓> Changes Configuration Selection
<Enter> Saves Selection; <ESC> to Cancel
```

Ops - Single Cable Server

- Single Cable Server in legacy IDC
- SRIOV
  - Linux owned PF port
  - DPDK use VF port
- Only 2 queues
  - 4 queues after upgrade Linux ixgbe to 4.1.5

Ops - Troubleshooting

- Dynamic configurate dump conditions
  - rte_pktmbuf_dump
    - hard to read
  - forwards to debug tap device
    - lost information, such as input port

- Any Suggestions?

THANK YOU

UCLOUD

www.ucloud.cn