ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

VPP Accelerated High Performance & Scalable L3DSR L4 Load Balancer on Top Clos

Yusuke Tatsumi @ Yahoo Japan Corporation & Naoyuki Mori @ Intel

Acknowledgement:
Shunya Kitada, Yuta Kinoshita @ Yahoo Japan Corporation
Hongjun Ni, Ray Kinsella @Intel
Pierre Pfister, Jerome Tollet @Cisco

ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

Agenda

- Yahoo! JAPAN introduction, background, and issues
- FD.io VPP Overview
- VPP LB data-plane
- VPP LB control-plane
- Summary
- Call to Action

Yahoo! JAPAN introduction

ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

Yahoo! JAPAN introduction

Monthly Page Views
70+ Billion

Daily Unique Browser
90+ Million

Number of services
100+

Daily Unique Browser
(Only Smartphone)
60+ Million

Many Load Balancers support our services!

*Other names and brands may be claimed as the property of others.

https://s.yimg.jp/i/docs/ir/archives/jp/nenji/2017/jp2018integrated_report.pdf

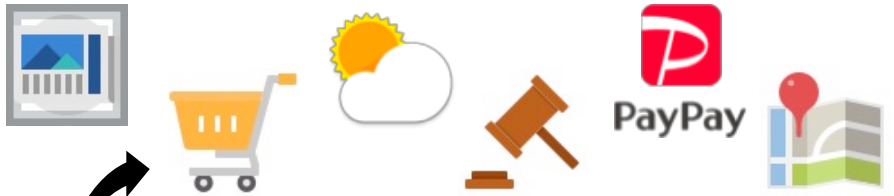
Background and issues of our LB

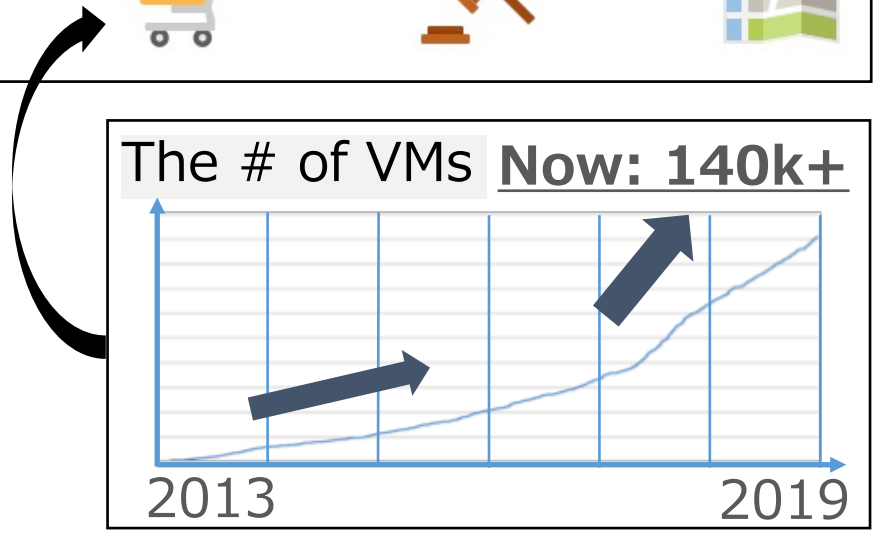
Total responsibility of LB =

Demands for speed:

[1] [2]

Explosive demands





More features are demanded of our LBaaS:

[3]

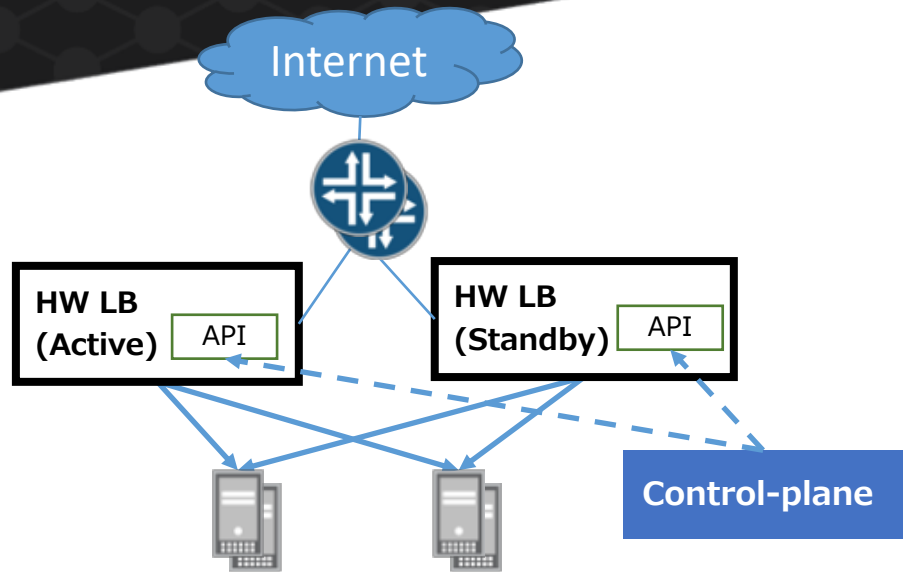
- Scale-in/out LB capability
- Robustness of LB system
- Elastic management of VIP

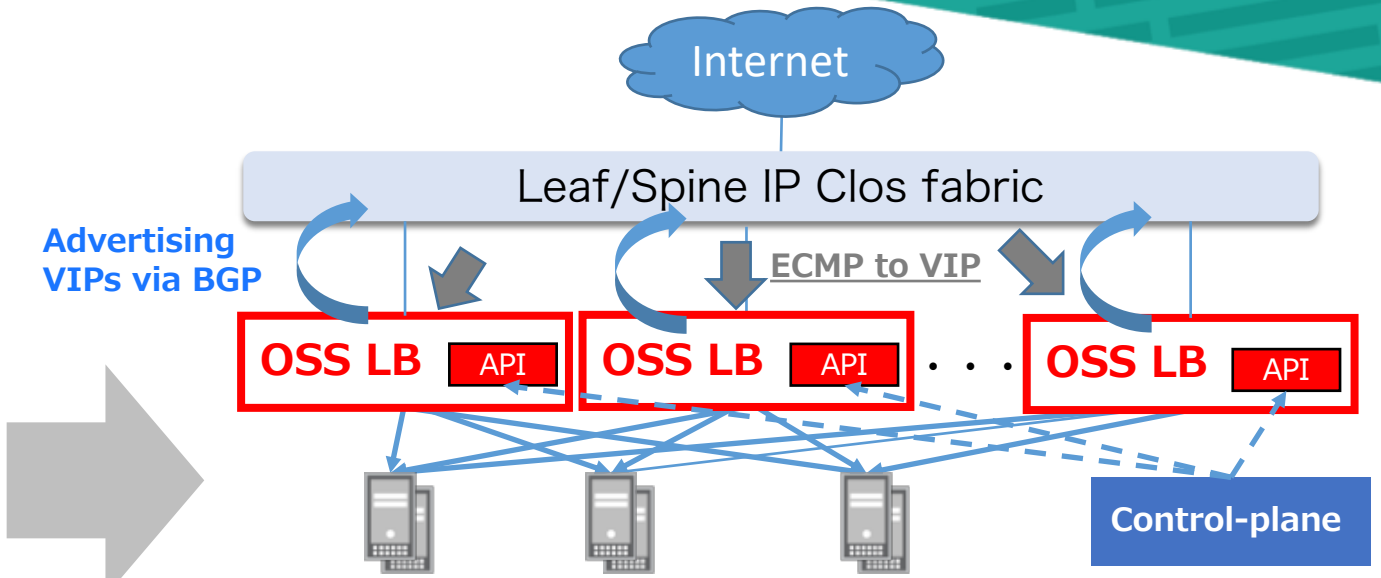
*Other names and brands may be claimed as the property of others.

[1] https://www.linux-kvm.org
[2] https://kubernetes.io/

[3] LBaaS: Load Balancer as a Service

Limitations of our system and OSS possibilities





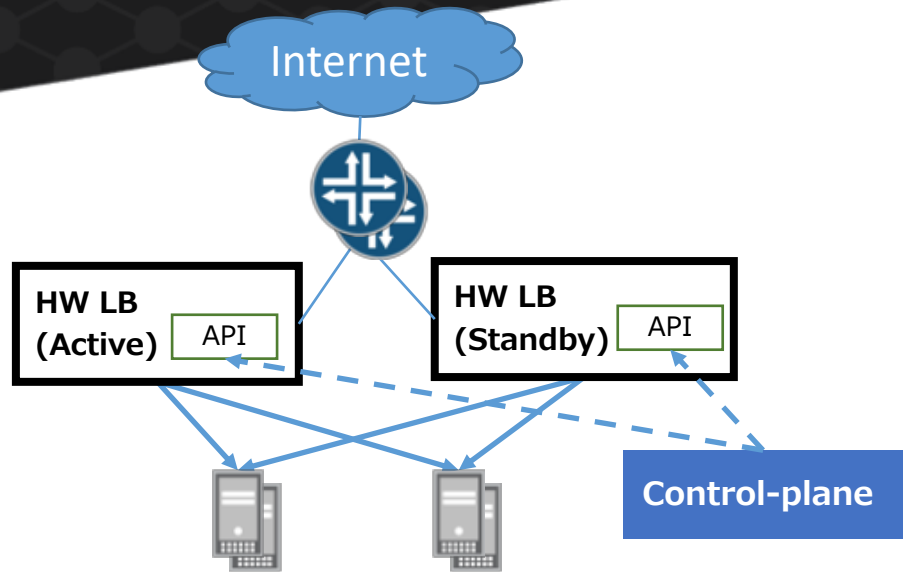
Our current appliance based LBaaS

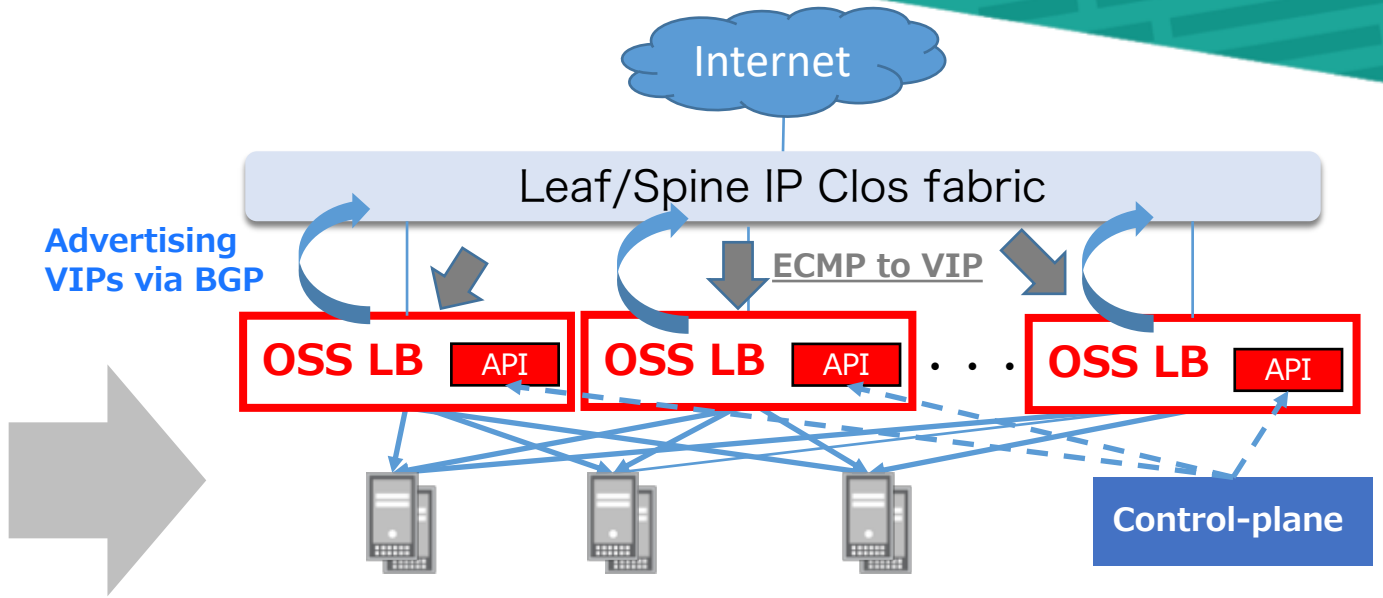
- Active-backup (2N) LB capability on top MLAG
  - hard to scaling-out
- Vender proprietary (Black box)
  - Lead time & Hardware EoL
- Careful operation (plan/exec) to place LB

Scalable LBaaS architecture with OSS

- Scaling-in/out (N+1) LB capability on top Clos
- Commodity server/NIC + OSS (White box)
- Easy operation to place LB with BGP

Limitations of our system and OSS possibilities





Strategy to break free of our current LBaaS limitations:

1. Collaborate with existing high performance OSS data-plane FD.io VPP
2. Develop control-plane from scratch based on our operational experiences

FD.io* VPP Overview

Universal data plane

- Layer 2 – 4 Network Stack
- CP, TM, Overlays and more …
- Linux (and FreeBSD) support
- Kernel Interfaces (Netmap, FastTap)
- Container and Virtualization support
- Appliance, Infrastructure, VNF & CNF

Extensible Modular Design

Architecture

- Pluggable, easy to understand & extend.
- Mature graph node architecture.

Plugins

- Full control to reorganize the pipeline.
- Fast, plugins are equal citizens

Fast, Scalable and Deterministic

- L2XC - 25+ Mpps per core

Deterministic

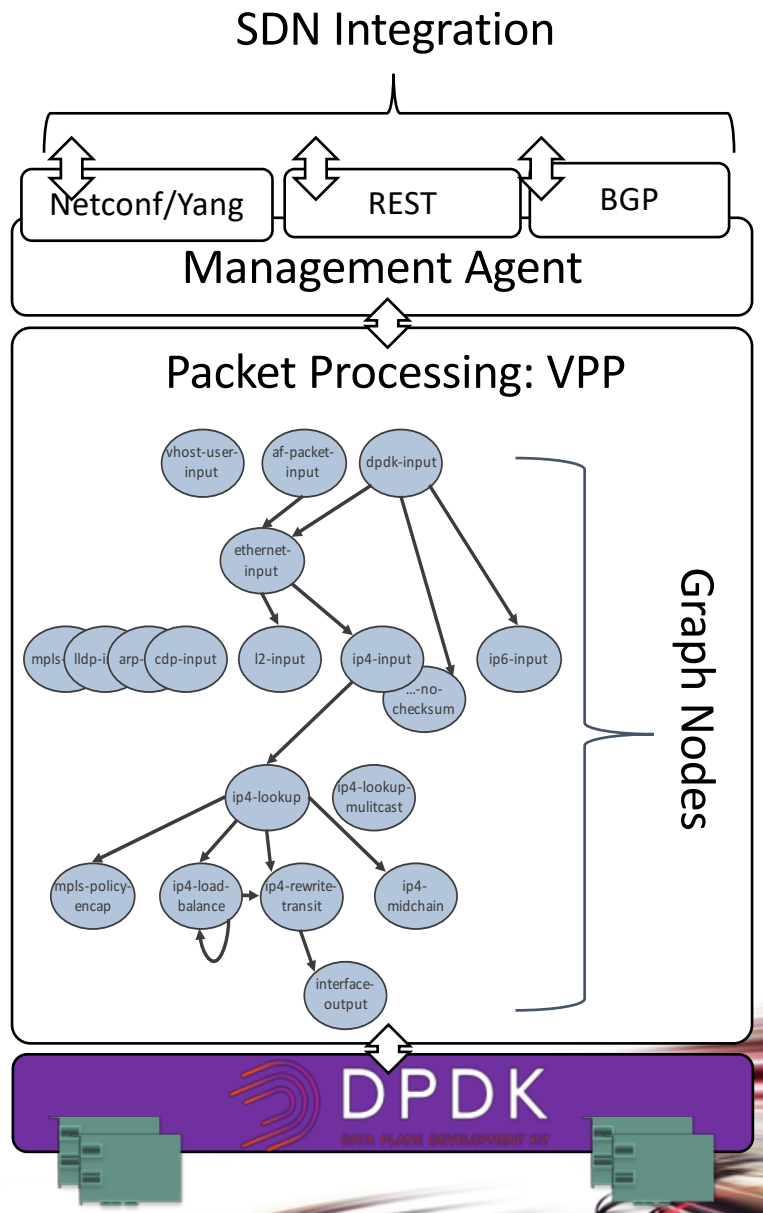
- 0 packet drops, ~15µs latency
- Continuous & extensive latency testing.

Scalability

- Linear scaling with core/thread count
- Supporting millions of concurrent L[2,3] tables entries.

Developer friendly

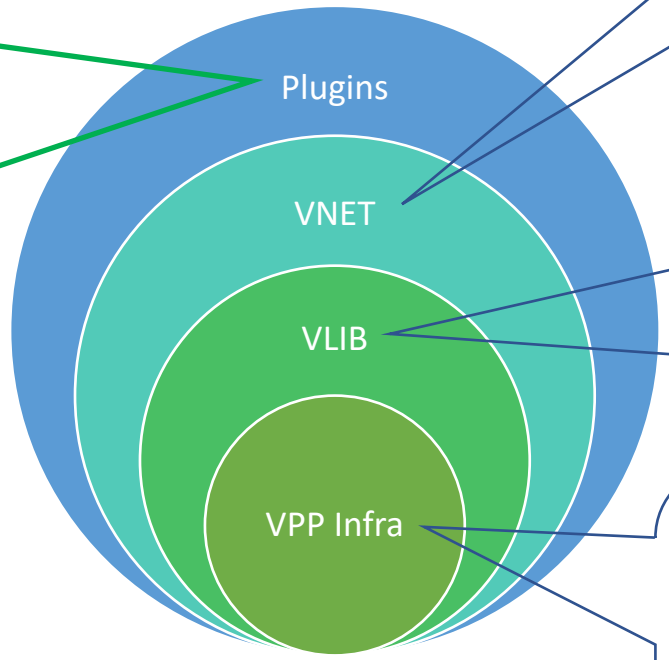
- Runtime counters for everything. (throughput, ipc, errors etc)
- Full pipeline tracing facilities.
- Multi-language API bindings.
- VPP command line introspection.



VPP Layering for modular design

Plugins

- Plugins can be in-tree:
  - SNAT,
  - Policy ACL,
  - Flow Per Packet,
  - ILA,
  - IOAM,
  - **Load Balancer,**
  - SIXRD,
  - VCGN
  - Segment Routing
- Separate fd.io project:
  - NSH_SFC



VNET

VPP networking source

- Devices
- Layer [2, 3, 4]
- Session Management
- Overlays
- Traffic Management

VLIB

VPP application management

- buffer, buffer management
- graph node, node management
- tracing, counters
- threading
- CLI

and most importantly …

- main()

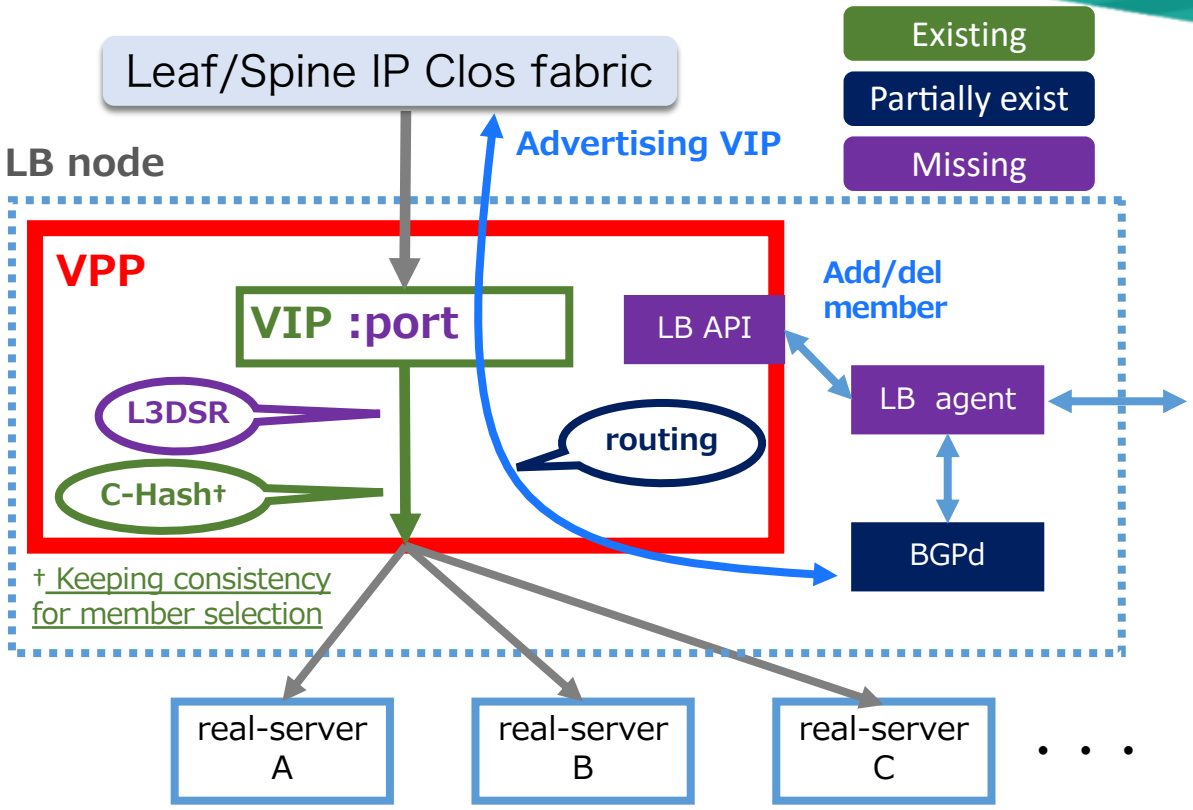
VPP Infra

Library of function primitives, for

- memory management
- memory operations
- vectors
- rings
- hashing
- timers

What's available? & What's missing?

- OSS based fast data-plane
  - ☑ **VPP**
- Scaling-in/out capability
  - ☑ **Consistent-hash feature in LB**
- Integration with ECMP/BGP on CLOS
  - ☑ **VPP Sandbox Router plugin**
- L3DSR LB (using IPv4 DSCP field) [1]
  - ☒ **VPP LB L3DSR support!**
- Management API
  - ☒ **VPP LB APIs enhancement!**



[1] https://www.nanog.org/meetings/nanog51/presentations/Monday/NANOG51.Talk45.nanog51-Schaumann.pdf

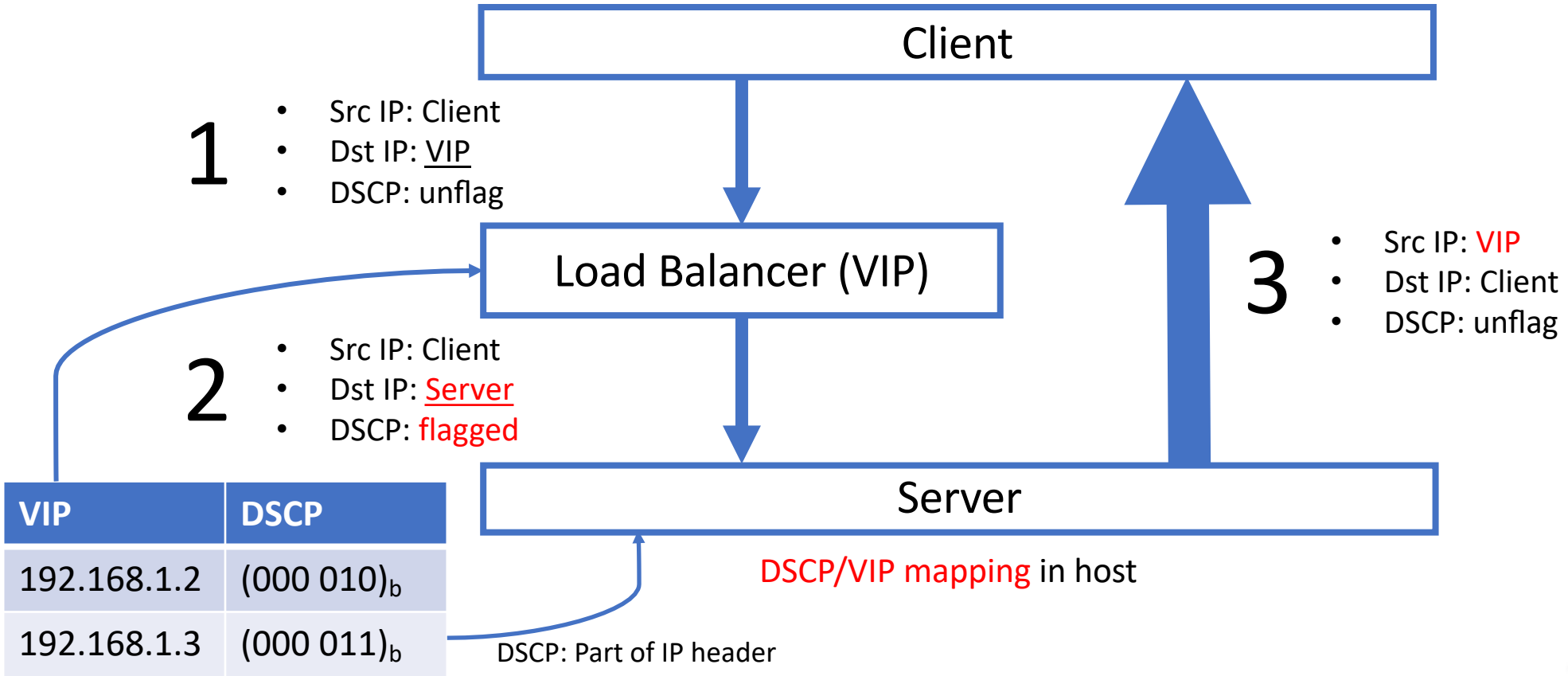
VPP LB plugin Enhancements

Category	Feature	Description	First appeared
Framework	Consistent Hash	Brings great scalability and redundancy w/ LB nodes	16.09
	Multi-Threading	Performance scalability w/ CPU cores	16.09
Protocols	GRE4/GRE6	Encap w/ GRE IPv4 and IPv6	16.09
	L3DSR ★	Layer 3 Direct Server Return	18.04
	NAT4/NAT6	NAT (originally for kube-proxy)	18.07
	Port # aware ★	Care TCP/UDP port number for LBaaS integration	18.10
API	Manipulation ★	Set/Get VIP/member over API from controller	18.10
	Statistics ★	Collect more detailed statistics like # of connections	On going

★ Added by this project

Why L3DSR (L3 Direct Server Return) ?

- No tunneling overhead with DSCP
- DSR: Saving LB bandwidth (*3*. is generally x2-10 larger than *1*.)

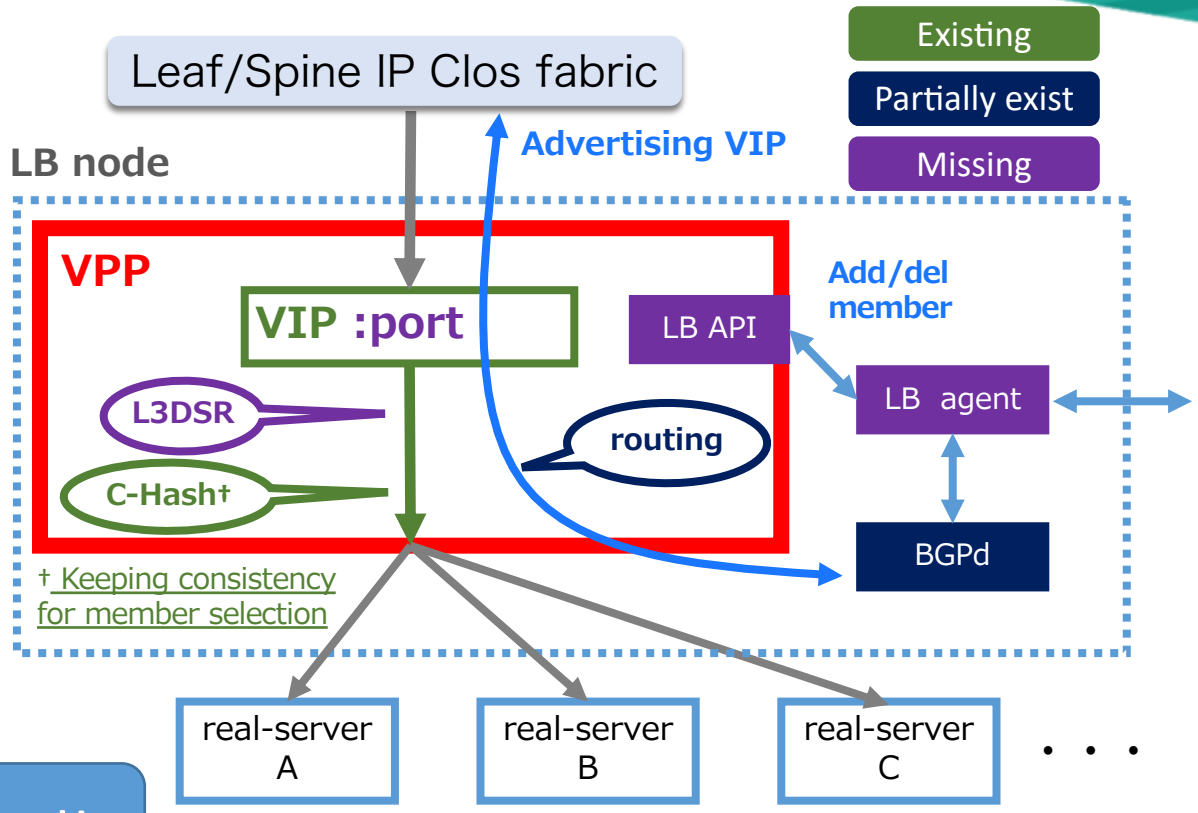


VIP	DSCP
192.168.1.2	$(000\ 010)_b$
192.168.1.3	$(000\ 011)_b$

Filling missing parts, build up system!

- OSS based fast data-plane
  - ☑ **VPP**
- Scaling-in/out capability
  - ☑ **Consistent-hash feature in LB**
- Integration with ECMP/BGP on CLOS
  - ☑ **VPP Sandbox Router plugin**
- L3DSR LB (using IPv4 DSCP field) [1]
  - ☑ **VPP LB L3DSR supported**
- Management API
  - ☑ **VPP LB APIs**

Successfully Developed!

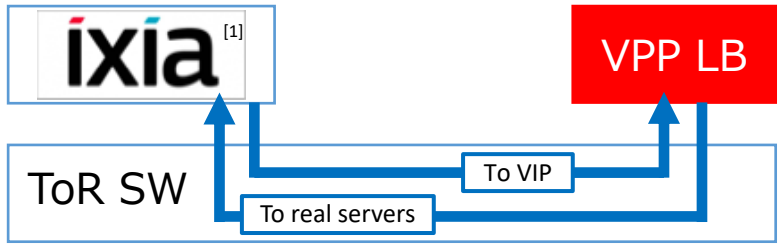


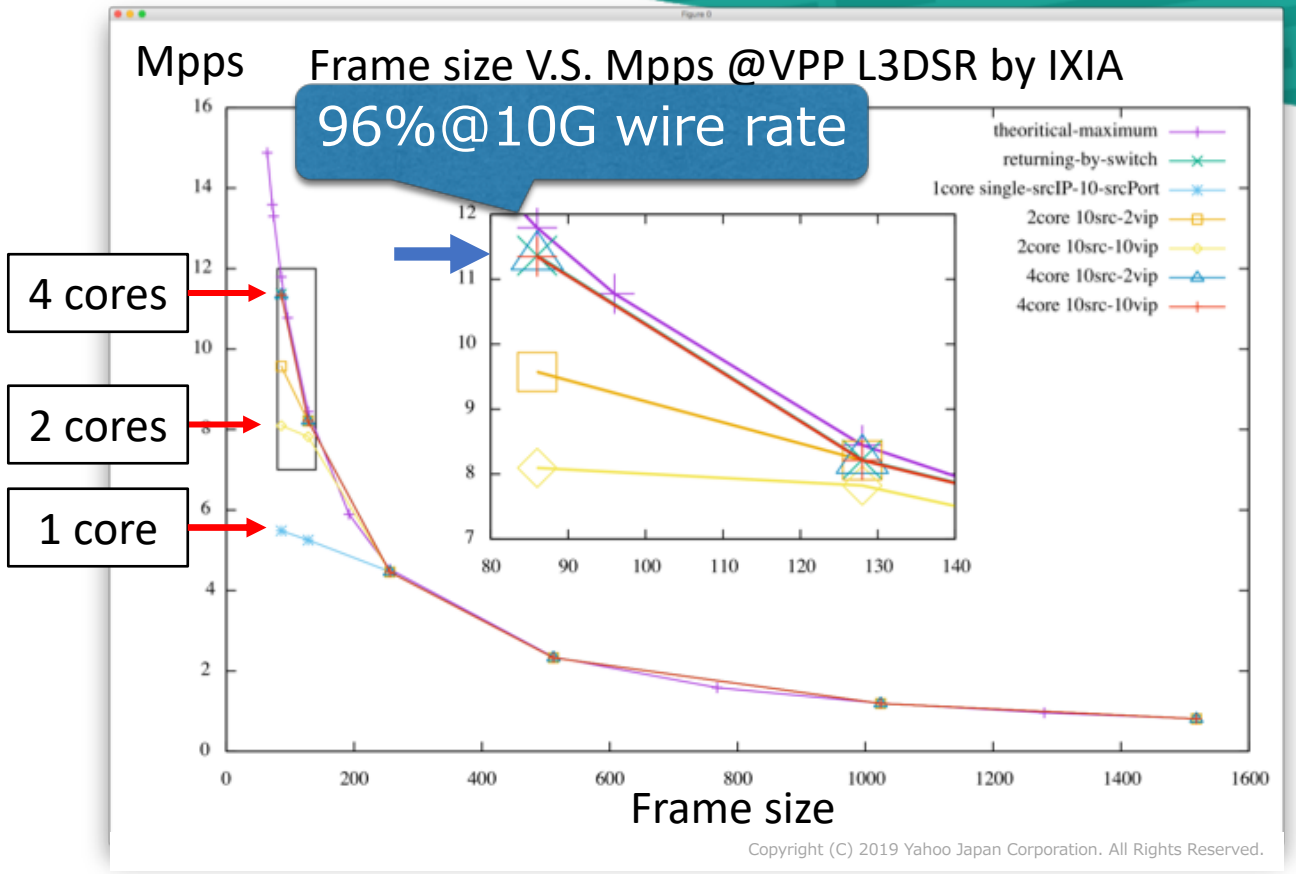
[1] https://www.nanog.org/meetings/nanog51/presentations/Monday/NANOG51.Talk45.nanog51-Schaumann.pdf

ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

VPP LB/node performance test @ 10Gbps

- Traffic generator: Ixia IxNetwork
- VPP Server Spec:
  - CPU: Intel® Xeon® Processor E5-2650L v3 * 2S
  - Memory: 384GB
  - NIC: Intel® Ethernet Network Connection X540-AT2 (10Gbps*2)





Close to 10G wire rate performance achieved!

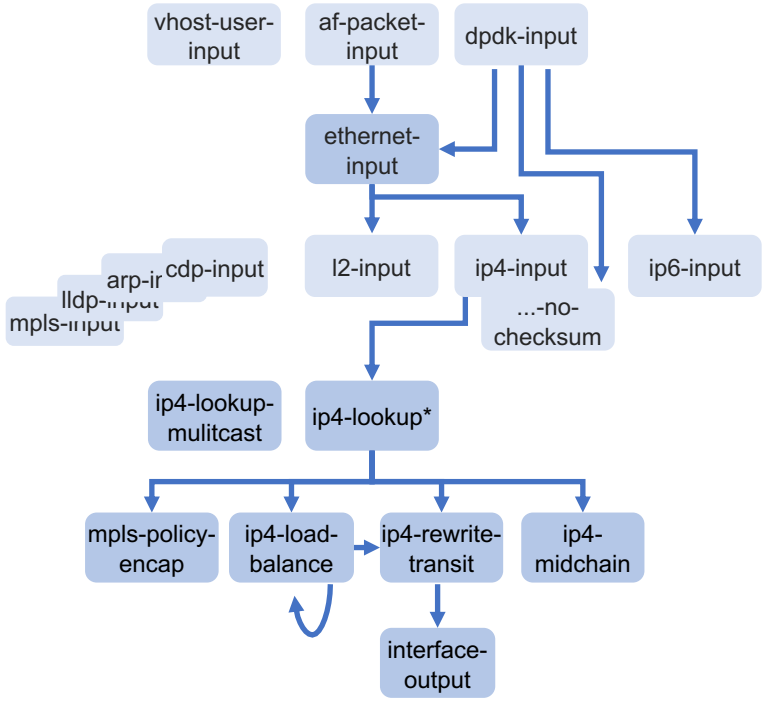
*Other names and brands may be claimed as the property of others.

[1] https://www.ixiacom.com/

FD.io VPP – The “Magic” of Vectors

Compute Optimized SW Network Platform

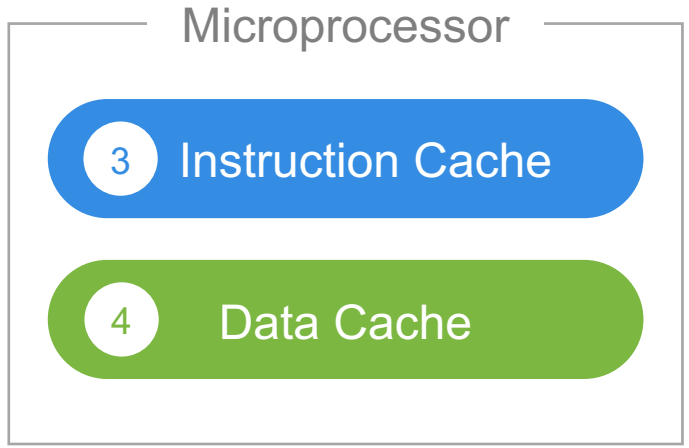
1 Packet processing is decomposed into a directed graph of nodes …



2 … packets move through graph nodes in vector …

Packet 0
Packet 1
Packet 2
Packet 3
Packet 4
Packet 5
Packet 6
Packet 7
Packet 8
Packet 9
Packet 10

3 … graph nodes are optimized to fit inside the instruction cache …

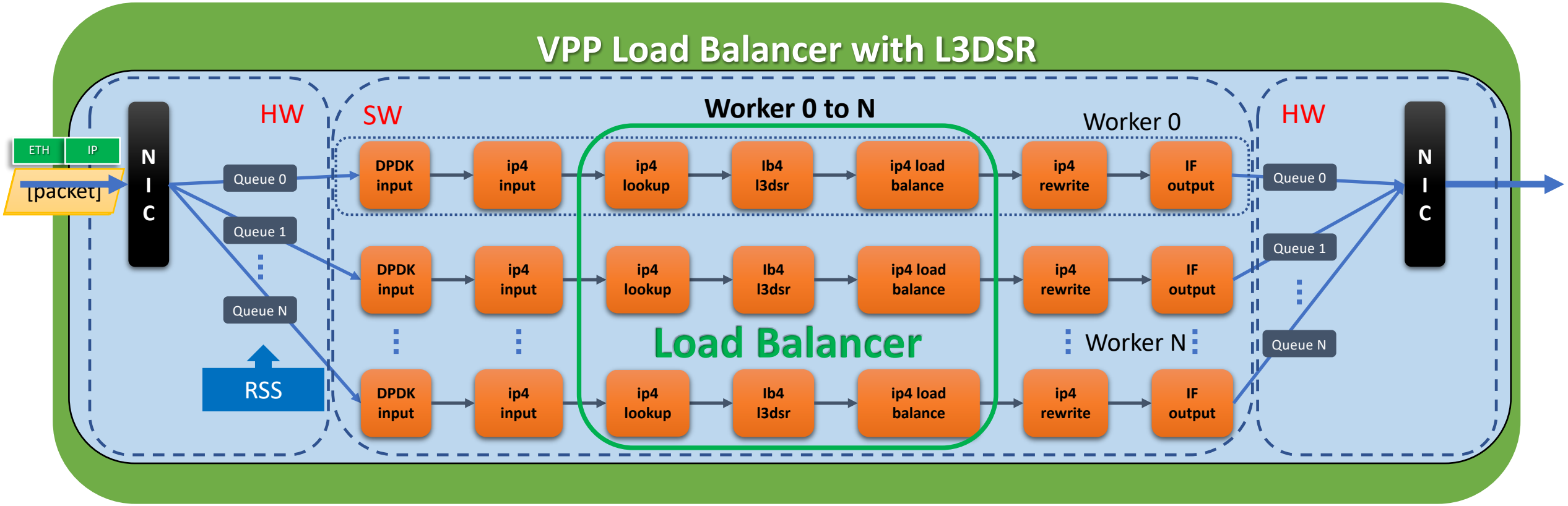


4 … packets are pre-fetched into the data cache.

* Each graph node implements a “micro-NF”, a “micro-NetworkFunction” processing packets.

Makes use of modern Intel® Xeon® Processor micro-architectures.
Instruction cache & data cache always hot ➔ Minimized memory latency and usage.

Multi-threading for performance scalability

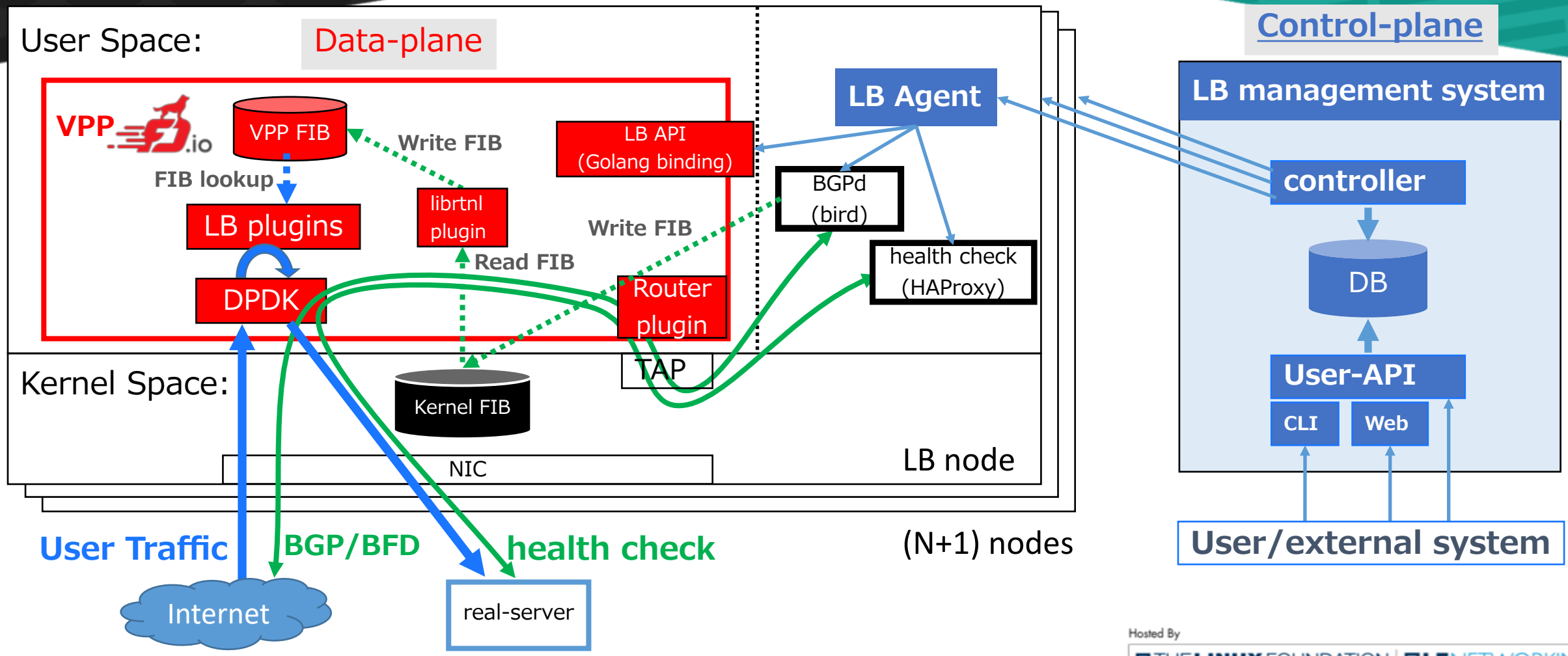


- RSS (Receive Side Scaling) enables traffic associated with one connection to a given thread.
- Well leveraging HW feature RSS and SW multi-threading for true scalability

Collaboration with VPP resulted in:

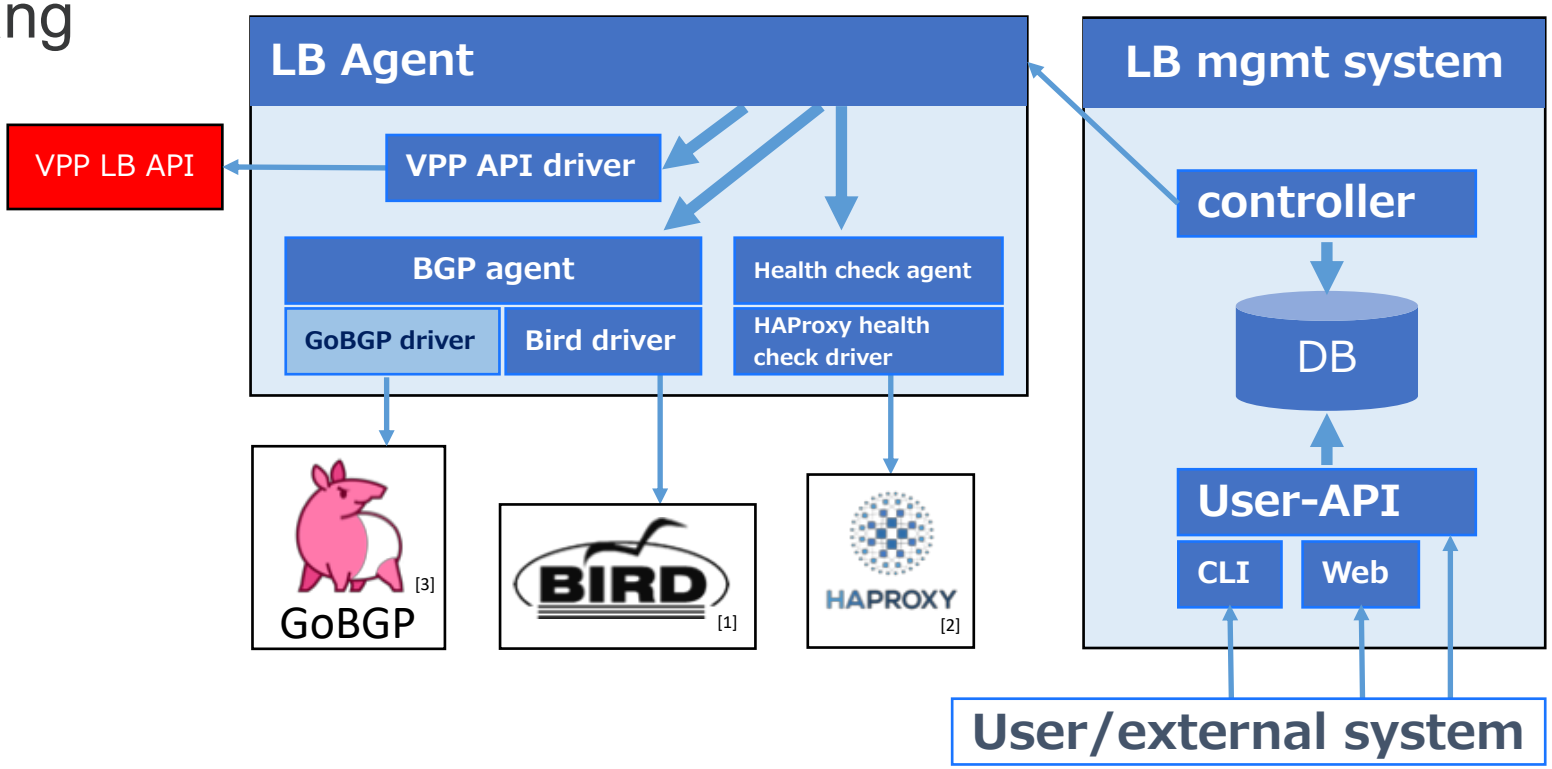
- OSS based stateless L3DSR L4 Load Balancer
- Scaling-in/out under N+1 capability
- On top Clos with BGP robustness
- Sufficient performance

Internal VPP LB node architecture and control-plane



LB control-plane

- Full-scratch written by Golang
  - 22k LoC with gRPC
- Components
  - LB agent and drivers
    - VPP API agent
    - BGP agent
    - Health check agent
  - LB management system
    - Controller & DB
    - User-API & Interface

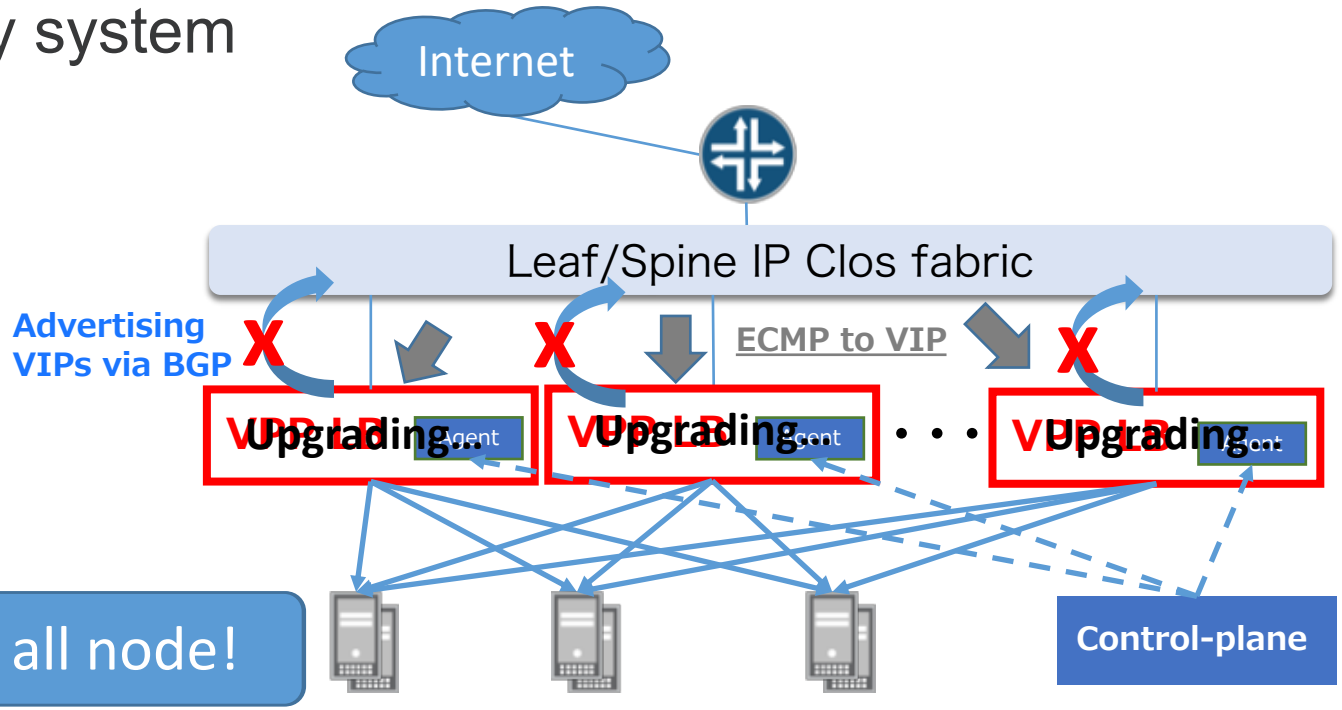


*Other names and brands may be claimed as the property of others.

[1] https://bird.network.cz/ [2] http://www.haproxy.org/ [3] https://osrg.github.io/gobgp/

LB control-plane (Cont.)

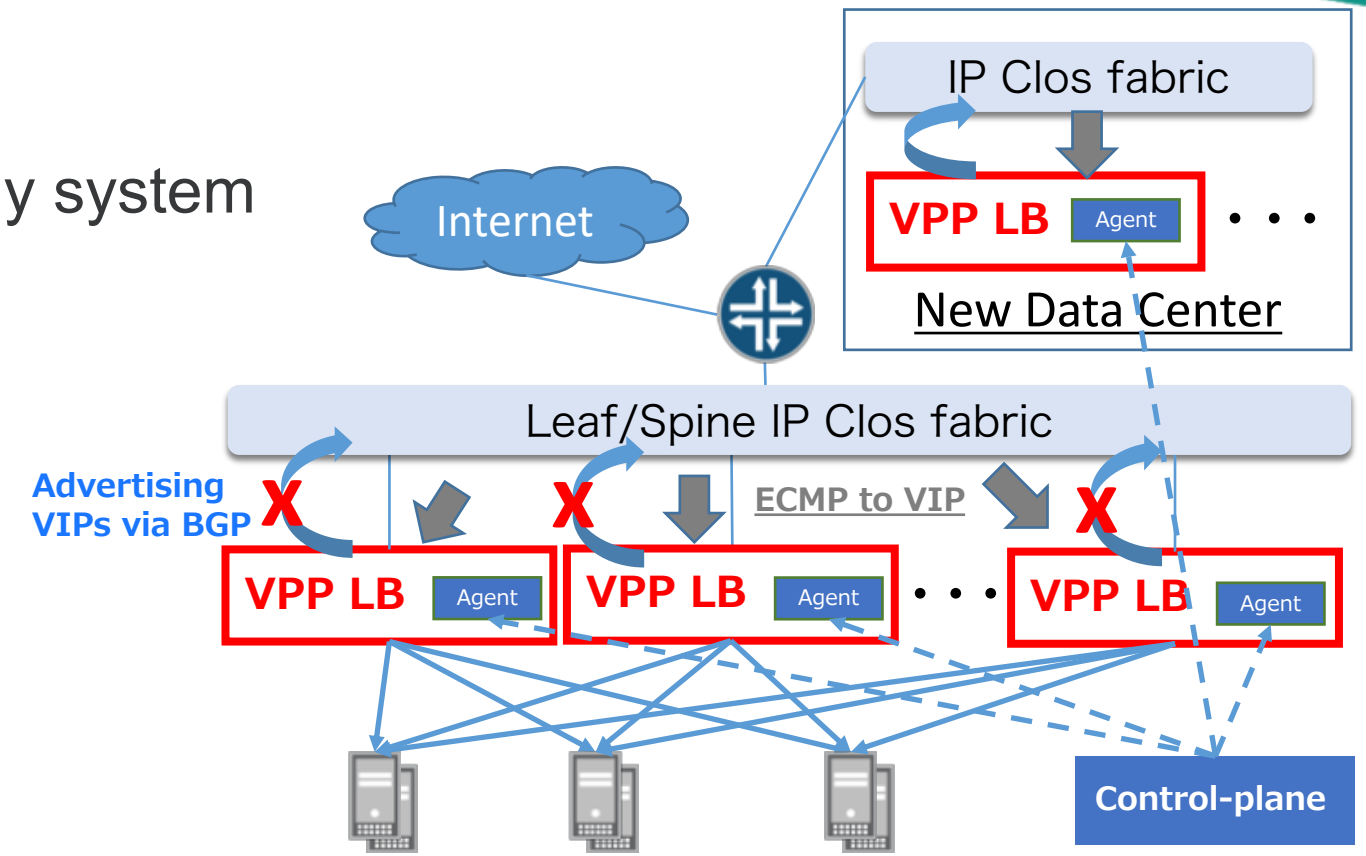
- Features
  - Serving API to integrate with any system
  - Zero-downtime upgrading
  - Life cycle management
    - HW EoL migration



Completed all node!

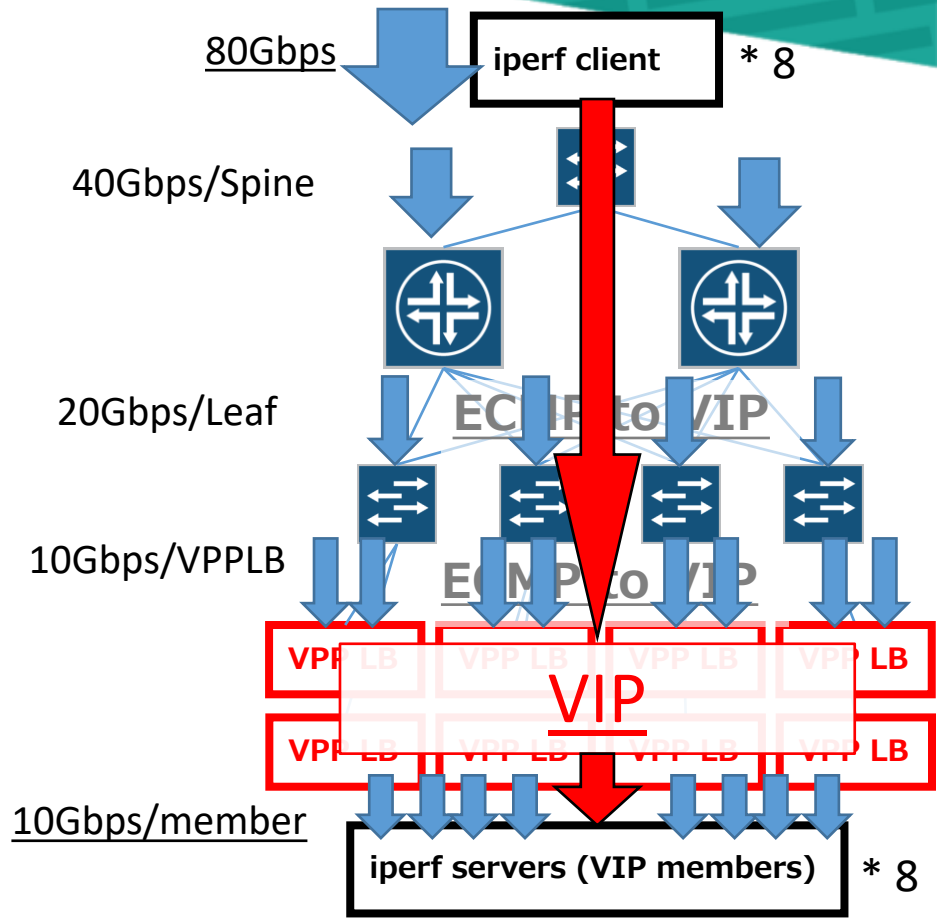
LB control-plane (Cont.)

- Features
  - Serving API to integrate with any system
  - Zero-downtime upgrading
  - Life cycle management
    - HW EoL migration



LB system performance @ 80Gbps

- Setting
  - Traffic generator: iperf 10Gbps * 8
  - 8 VPP-LB nodes: 10Gbps * 8
  - ECMP
    - Spine
    - Leaf
    - VPP-LB
- 80Gbps to the 1 VIP on 8 VPP-LB



ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

LB system performance @ 80Gbps

vpplb-node01
VIP: 172.27.156.248

vpplb-node02
VIP: 172.27.156.248

vpplb-node03
VIP: 172.27.156.248

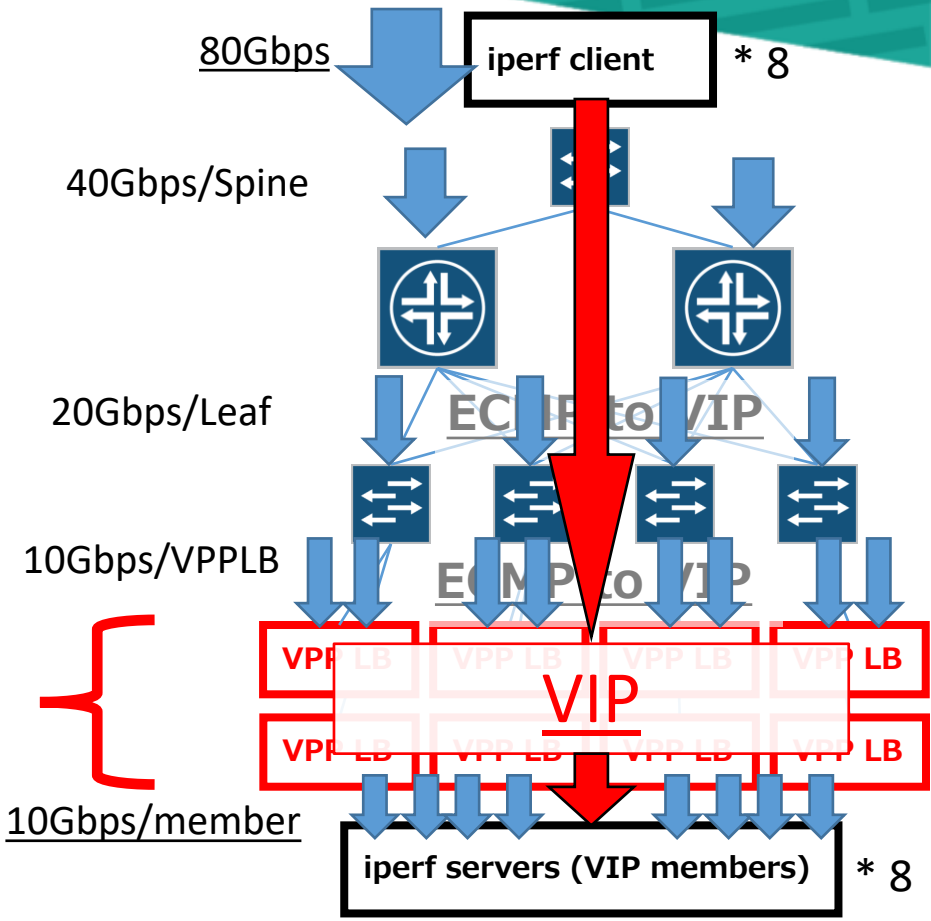
vpplb-node04
VIP: 172.27.156.248

vpplb-node05
VIP: 172.27.156.248

vpplb-node06
VIP: 172.27.156.248

vpplb-node07
VIP: 172.27.156.248

vpplb-node08
VIP: 172.27.156.248



Hosted By
THE LINUX FOUNDATION | LF NETWORKING

*Other names and brands may be claimed as the property of others.

ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

LB system performance @ 80Gbps

iperf –c VIP
=> 8.29 Gbps

iperf –c VIP
=> 9.42 Gbps

iperf –c VIP
=> 8.50 Gbps

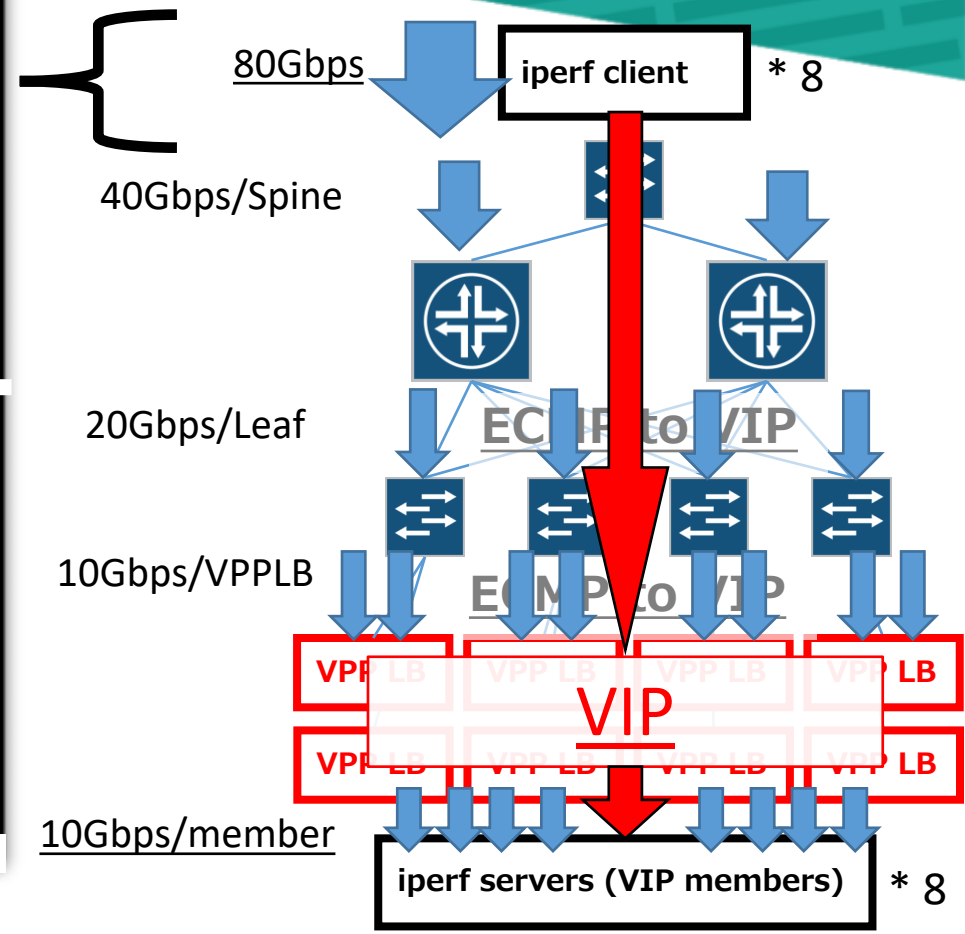
iperf –c VIP
=> 8.81 Gbps

iperf –c VIP
=> 8.98 Gbps

iperf –c VIP
=> 8.18 Gbps

iperf –c VIP
=> 9.34 Gbps

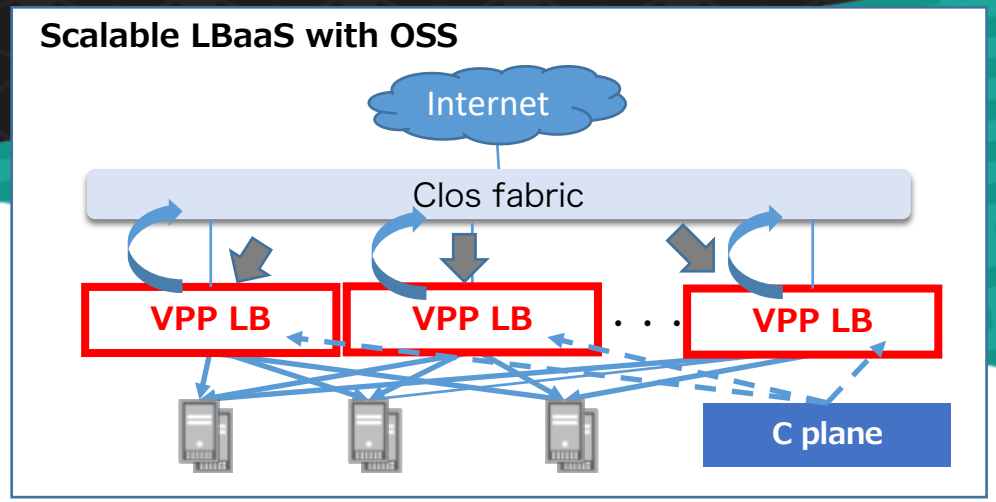
iperf –c VIP
=> 9.08 Gbps



80Gbps balancing via 8 VPP-LB on Top Clos/ECMP

Scale-out on multiple VPP-LB node & ECMP on top Clos

*Other names and brands may be claimed as the property of others.



LBaaS Summary:

Coordination of our Control-plane with VPP data-plane

- Easy operation to place LB node with BGP
- Serving API to integrate with any system
- Zero-downtime upgrading
- Life cycle management (EoL migration)
- OSS based stateless L3DSR L4 Load Balancer
- Scaling-in/out under N+1 capability
- On top Clos with BGP robustness
- Sufficient performance in operation perspective

Realization of LBaaS:

- **Scale-in/out LB capability**
- **Robustness of LB system**
- **Elastic management of VIP**

Call to Action

- FD.io: A Good open community resolving real-world networking issues and developing new features
  - Quick response from community
  - Deepening networking knowledge in community
- Collaboration
  - Sharing your ideas/issues
  - Coding and verifying from your own perspective
  - Open discussion within the community

Let's work together!

- ytatsumi@yahoo-corp.jp
- naoyuki.mori@intel.com

Legal Disclaimers

- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.

  Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.
- Cost reduction scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.
- All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.
- Intel, the Intel logo and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.
- *Other names and brands may be claimed as the property of others.
- © 2018 Intel Corporation.

*Other names and brands may be claimed as the property of others.

ons

NORTH AMERICA

OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

Hosted By

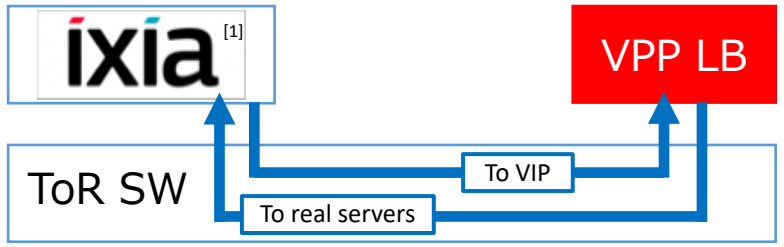
THE LINUX FOUNDATION | LF NETWORKING

ons
NORTH AMERICA
OPEN NETWORKING //
Enabling Collaborative
Development & Innovation

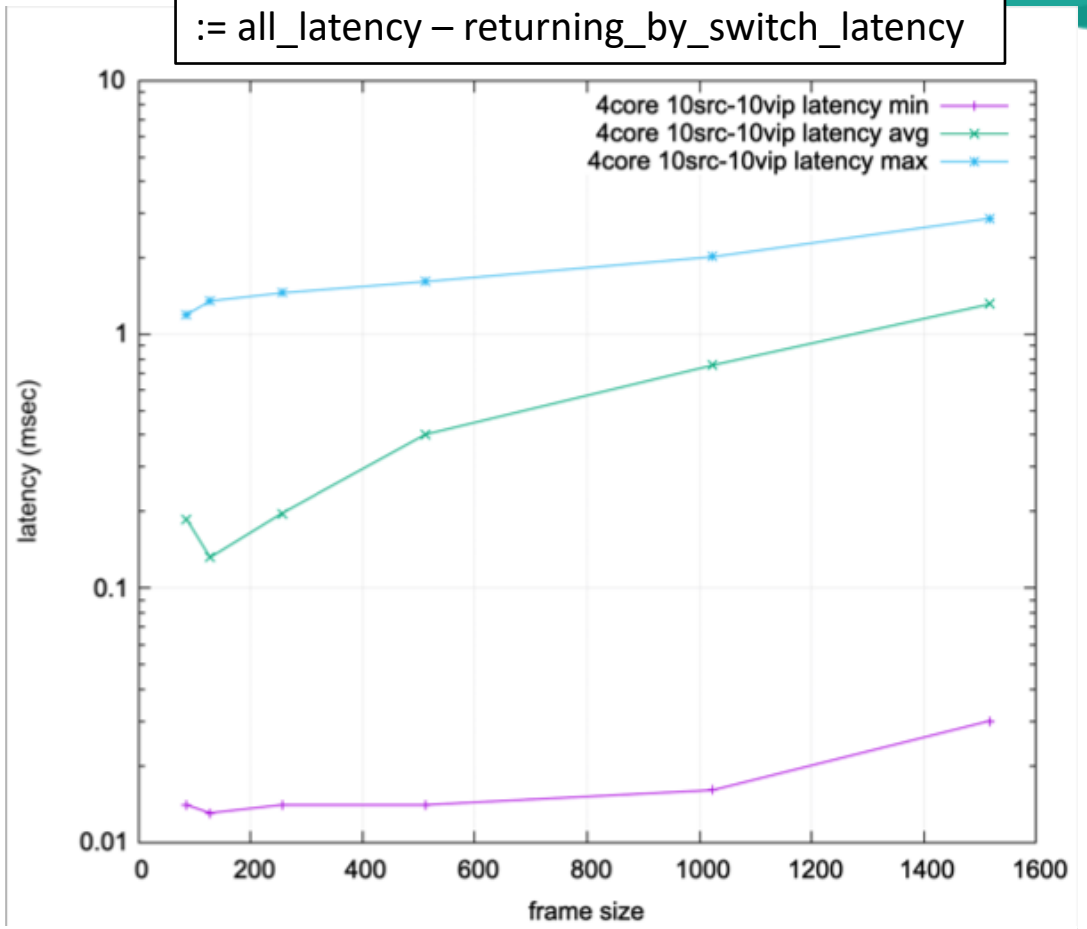
Backup

VPP LB/node performance test @ 10Gbps

- Traffic generator: Ixia IxNetwork
- VPP Server Spec:
  - CPU: Intel® Xeon® Processor E5-2650L v3 * 2S
  - Memory: 384GB
  - NIC: Intel® Ethernet Network Connection X540-AT2 (10Gbps*2)



VPPLB Additional latency
:= all_latency – returning_by_switch_latency

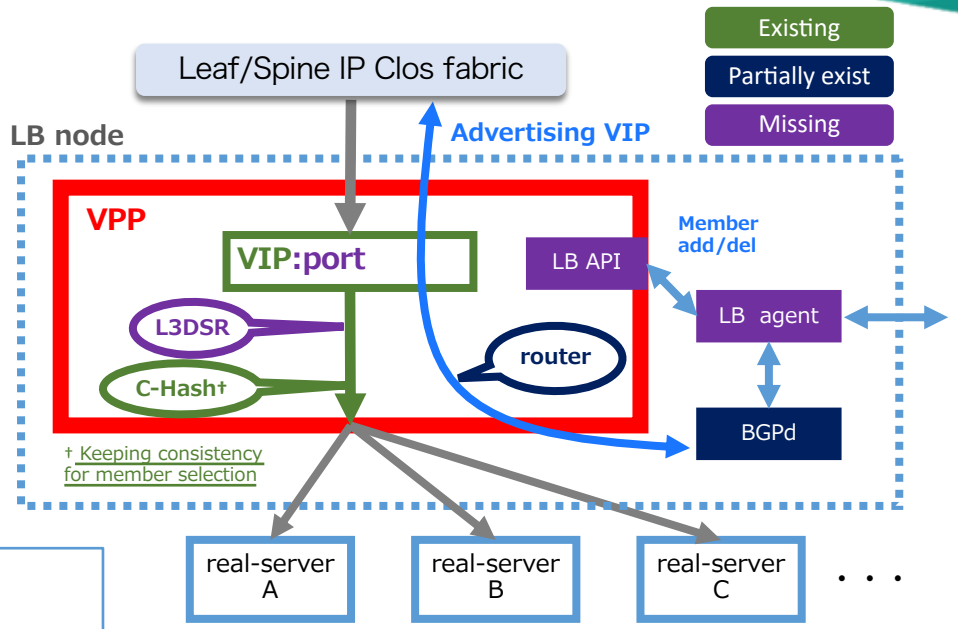


*Other names and brands may be claimed as the property of others.

[1] https://www.ixiacom.com/

L3DSR LB configuration example

- L3DSR LB (using IPv4 DSCP field) [1]
  - ☑ **VPP LB L3DSR supported**
- Management API
  - ☑ **VPP LB APIs**



L3DSR LB configuration example:

lb vip 192.168.1.1/32 protocol tcp port 80 encap l3dsr dscp 23 new_len 1024
lb as 192.168.1.1/32 protocol tcp port 80 172.16.1.1
lb as 192.168.1.1/32 protocol tcp port 80 172.16.1.2
lb as 192.168.1.1/32 protocol tcp port 80 172.16.1.3 * Added by this project

[1] https://www.nanog.org/meetings/nanog51/presentations/Monday/NANOG51.Talk45.nanog51-Schaumann.pdf