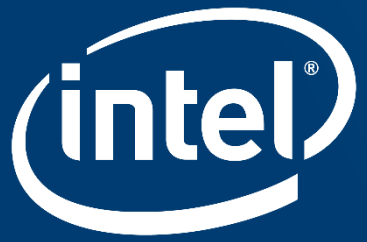


BUILDING A WHITE BOX LOAD BALANCER FOR THE CLOUD DATA CENTER

Open Networking Summit, San Jose, CA

April 2018

Jay Vincent, Platform Solution Architect

M Jay, Platform Application Engineer

Legal Disclaimer

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm%20 Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Celeron, Intel, Intel logo, Intel Core, Intel Inside, Intel Inside logo, Intel. Leap ahead., Intel. Leap ahead. logo, Intel NetBurst, Intel SpeedStep, Intel XScale, Itanium, Pentium, Pentium Inside, VTune, Xeon, and Xeon Inside are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Intel® Active Management Technology requires the platform to have an Intel® AMT-enabled chipset, network hardware and software, as well as connection with a power source and a corporate network connection. With regard to notebooks, Intel AMT may not be available or certain capabilities may be limited over a host OS-based VPN or when connecting wirelessly, on battery power, sleeping, hibernating or powered off. For more information, see http://www.intel.com/technology/iamt.

64-bit computing on Intel architecture requires a computer system with a processor, chipset, BIOS, operating system, device drivers and applications enabled for Intel® 64 architecture. Performance will vary depending on your hardware and software configurations. Consult with your system vendor for more information.

No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology is a security technology under development by Intel and requires for operation a computer system with Intel® Virtualization Technology, an Intel Trusted Execution Technology-enabled processor, chipset, BIOS, Authenticated Code Modules, and an Intel or other compatible measured virtual machine monitor. In addition, Intel Trusted Execution Technology requires the system to contain a TPMv1.2 as defined by the Trusted Computing Group and specific software for some uses. See http://www.intel.com/technology/security/ for more information.

†Hyper-Threading Technology (HT Technology) requires a computer system with an Intel® Pentium® 4 Processor supporting HT Technology and an HT Technology-enabled chipset, BIOS, and operating system. Performance will vary depending on the specific hardware and software you use. See www.intel.com/products/ht/hyperthreading_more.htm for more information including details on which processors support HT Technology.

Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, virtual machine monitor (VMM) and, for some uses, certain platform software enabled for it. Functionality, performance or other benefits will vary depending on hardware and software configurations and may require a BIOS update. Software applications may not be compatible with all operating systems. Please check with your application vendor.

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

Other vendors are listed by Intel as a convenience to Intel's general customer base, but Intel does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices. This list and/or these devices may be subject to change without notice.

Agenda

Load Balancing Options

Software Based Load Balancers

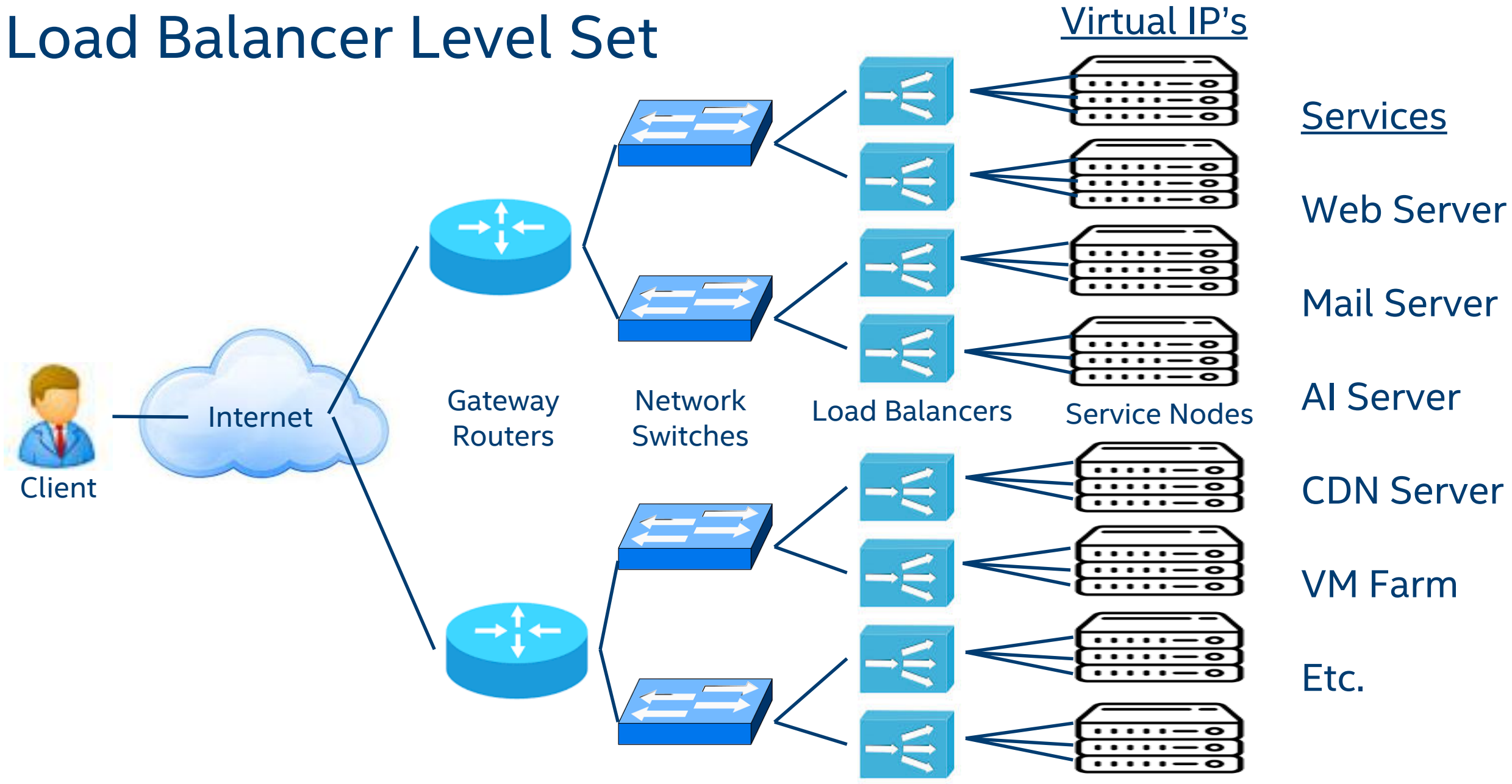
One Open Source Example

Key Performance Optimizations

Demonstration

Call to Action

Load Balancer Level Set



Load Balancer Options

Type	Advantage	Disadvantage	Cost Factors
Proprietary Physical Appliance	• Maximum Throughput • Robust Features	• Long Deployment Window • Static in Nature • Paying for stuff not used	• Large Capital Investment • Licensing and Support • Single Source Solution
Virtualized Proprietary Appliance	• Rapid Deployment • Robust Features	• Cost Increase with Scale • Paying for stuff not used	• Licensing and Support • COTS HW
Open Source Software	• Flexible • Scalable • Build what you need	• Limited Functionality	• Engineering • In House Support • COTS HW

Software Load Balancer Examples

Envoy

Maglev

Elastic Load Balancer

Jupiter

Katran

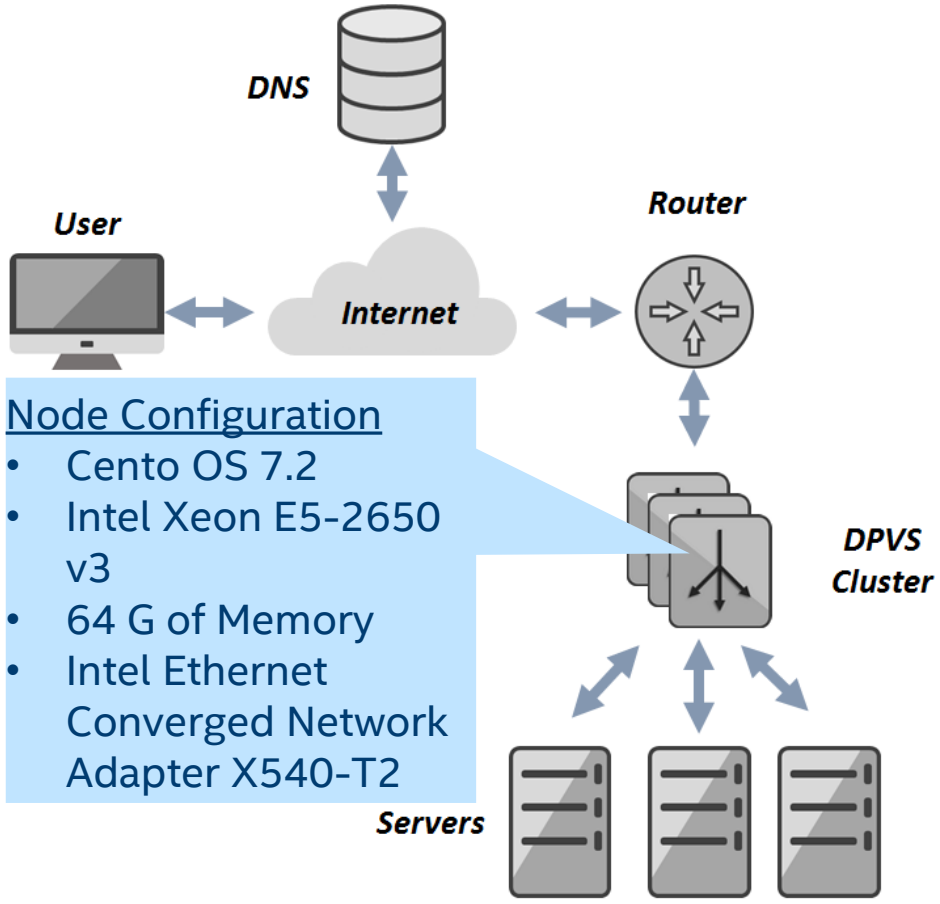
Load Balancer

FD.io VPP L3DSR

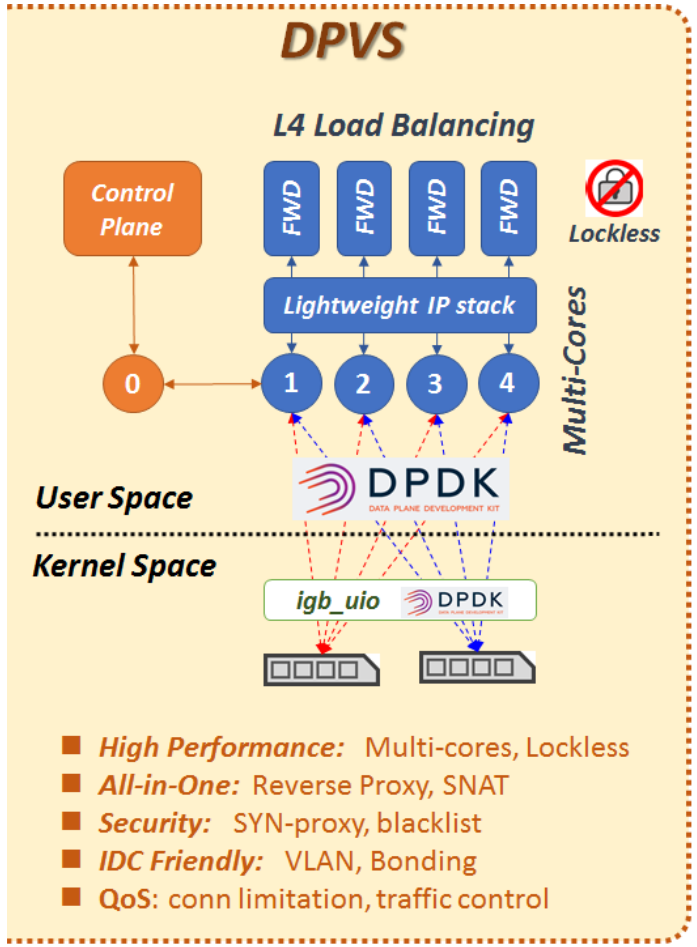
DPVS

Server Load Balancer

One Open Source Software Load Balancer



Node Configuration
- Cento OS 7.2
- Intel Xeon E5-2650 v3
- 64 G of Memory
- Intel Ethernet Converged Network Adapter X540-T2

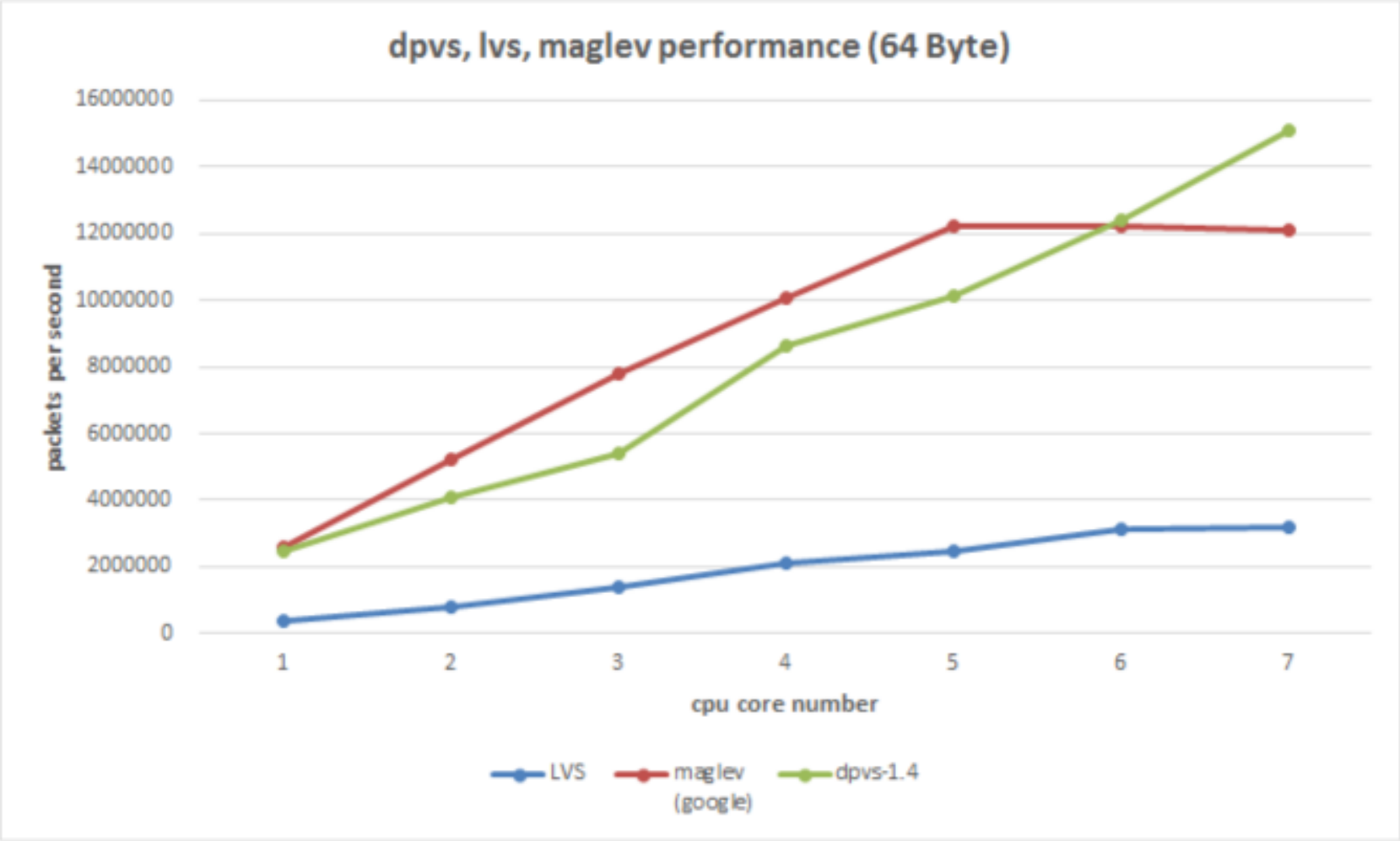


Benefits

- Consistent Hashing
- Connection tracking
- NIC Queue & CPU Bonding
- Must rapidly scale
- Equal Cost Multi-Path Routing
- Kernel Bypass using DPDK

https://github.com/**iqiyi/dpvs**

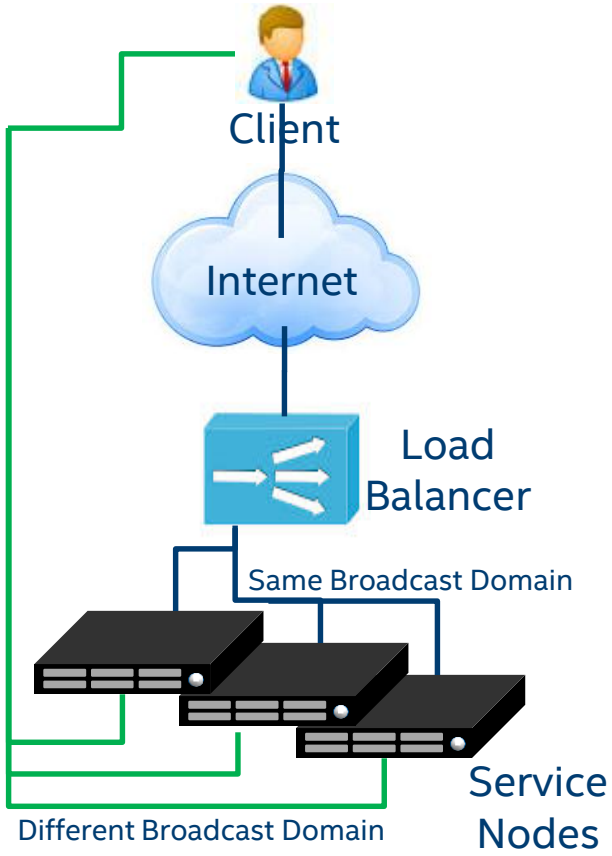
DPVS Performance



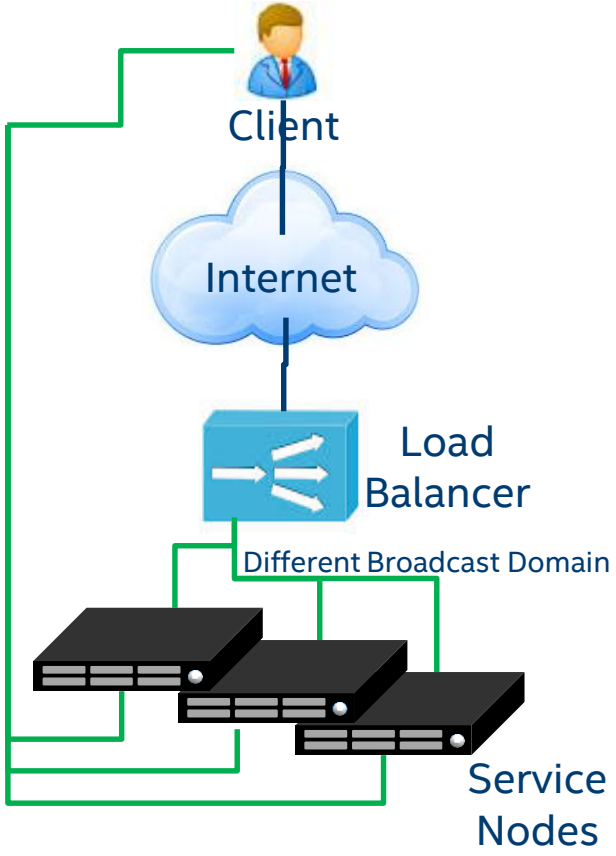
Source: https://github.com/**iqiyi/dpvs**

Three Modes of Operation

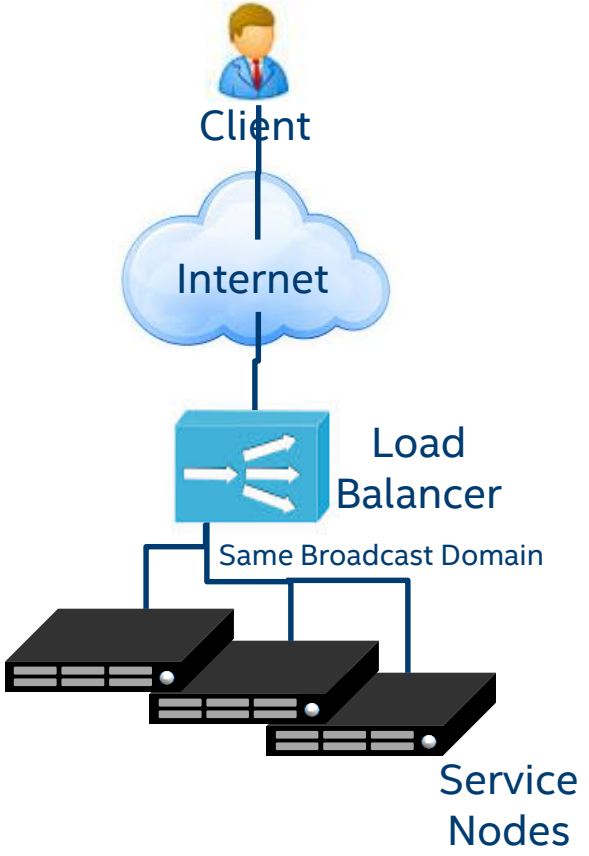
Direct Response



Tunnel



Full NAT



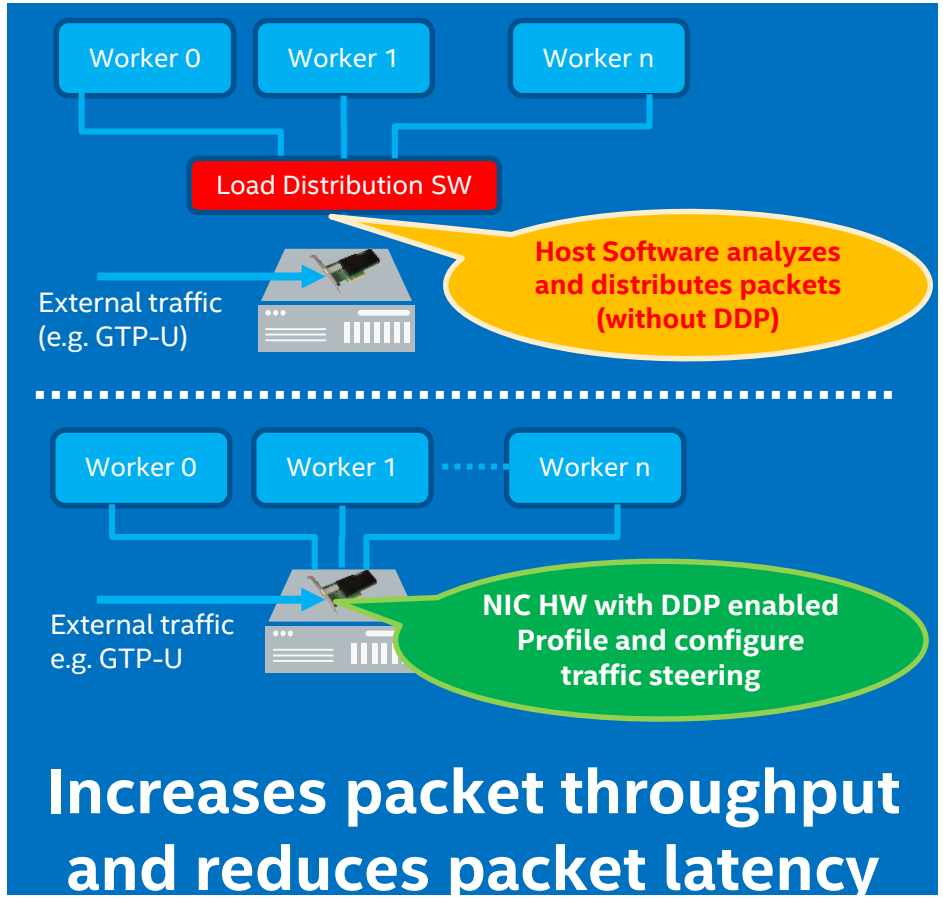
Performance Optimization

Intel Ethernet Adapters

- Dynamic Device Personalization
- Intel® Ethernet Flow Director
  - Advanced Traffic Steering

DPDK Network Acceleration

- Kernel bypass
  - Faster interface with the kernel net stack
  - Polling instead of interrupts
  - Facilitates using standard Linux* userspace net tools (tcpdump, ftp, and so on)
  - Eliminate the copy_to_user and copy_from_user operations



Demo Choices

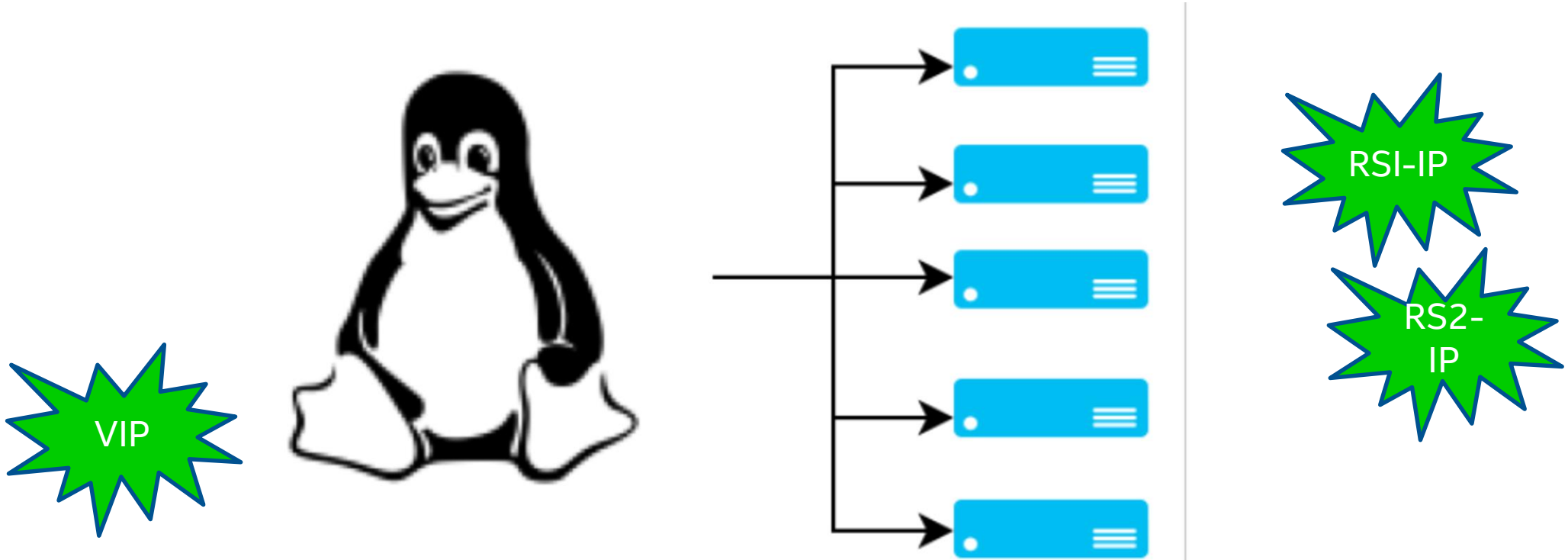
Full NAT Versus Direct Response?

Kernel Network Interface (KNI)?

Why to have man in the middle (Load Balancer) ?

Key Configuration for Demo

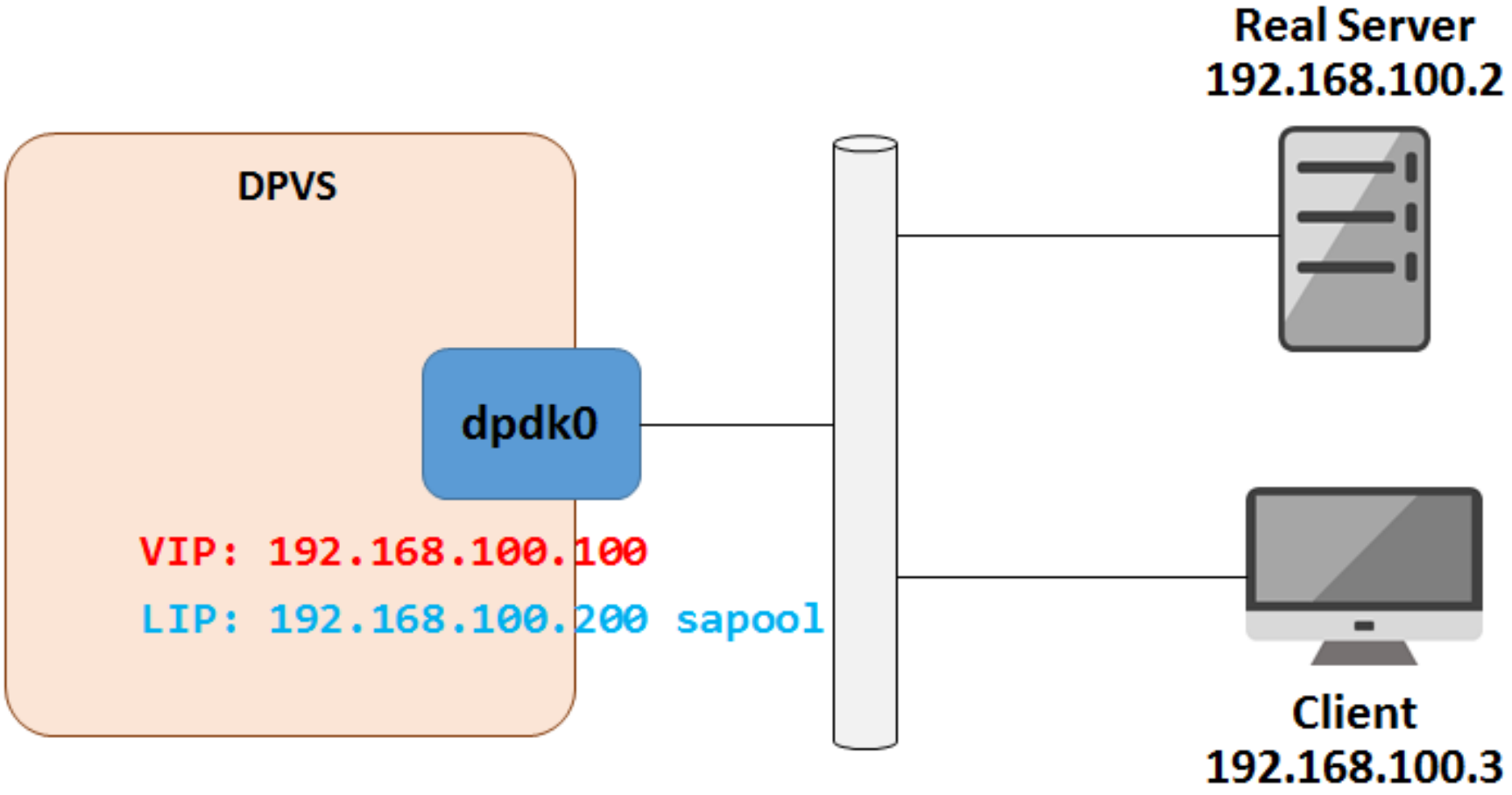
L4 Load Balancer –Full NAT



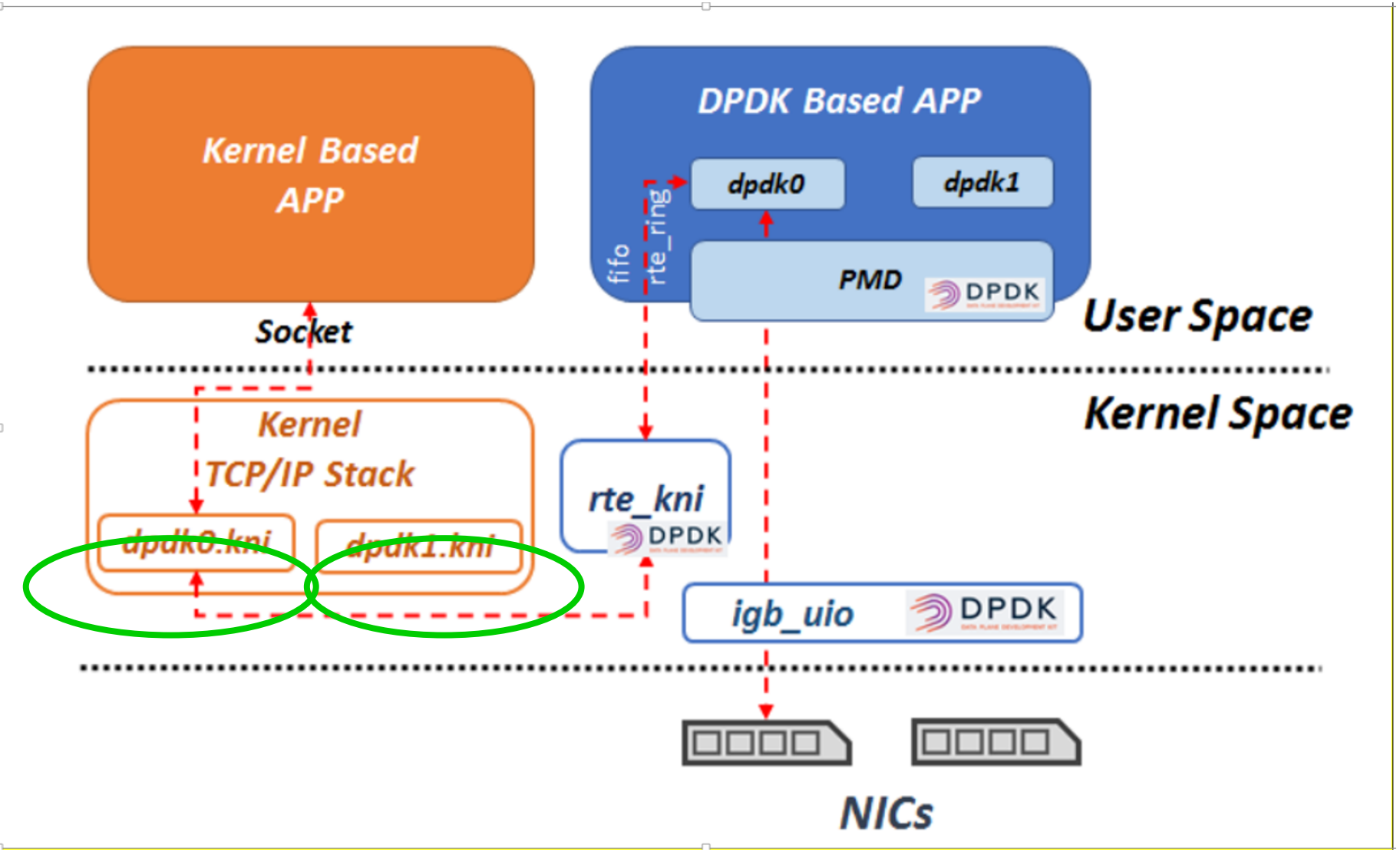
VIP – “Virtual IP” = “Service IP; ”RS1, RS2 – “Real Server IP”

The benefit of Full NAT – Servers can be used as is. No change needed

Direct Response Server Configuration



DPVS - Data Path (DPDK) + Control Path (IPVS)



Same Domain

- Both clients & Servers
  - On same side
- Dpdk0.kni alone enough

Different Domain

- Both clients & Servers
  - On same side
- Dpdk0.kni & DPDK1.kni

Spread To Servers? Or Benefit From Stickiness?

Flow Pinning has the benefit of Locality
Caching Benefits thereof.

When Spreading the Load, say for instance **Round Robin**
You trade locality for overall utilization.

What Other Scheduling Algorithms Are Out There?

Least Connected

Tagging Weight To The Server

That Is About Scheduling. What About Data Flow?

Forwarding Methods - Thin Request. Bulky Response

For Single Request, server returns multiple objects.

So Why Have Load Balancer in the path of bulky Response?

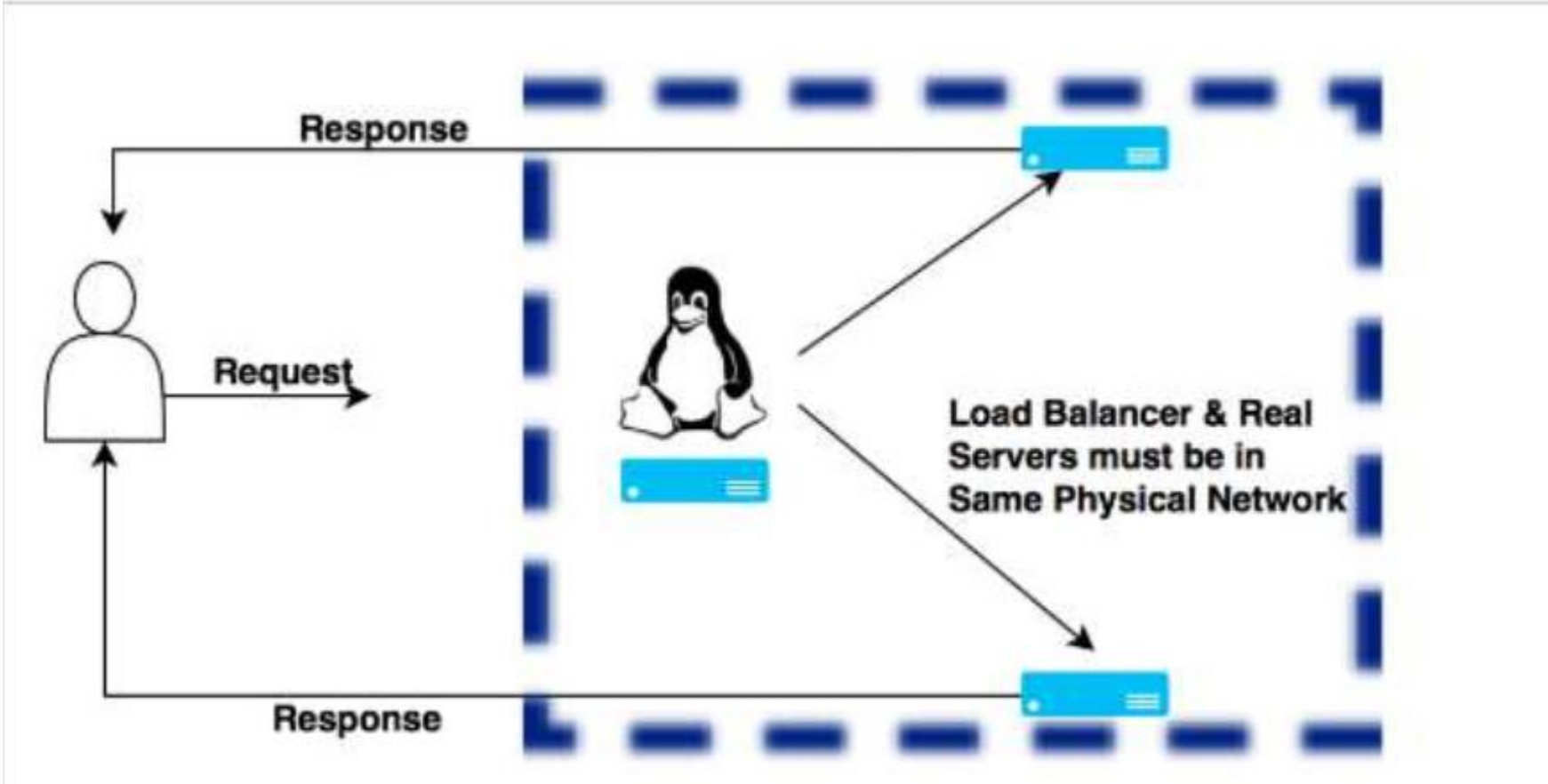
Why not Load Balancer Become Transparent During Server Response?

That is **Direct Response Server – DRS**

How?

Why To Have Man In The Middle During Response?

Request Is Served By Load Balancer But Not Response



Ready For Quiz?

Quiz

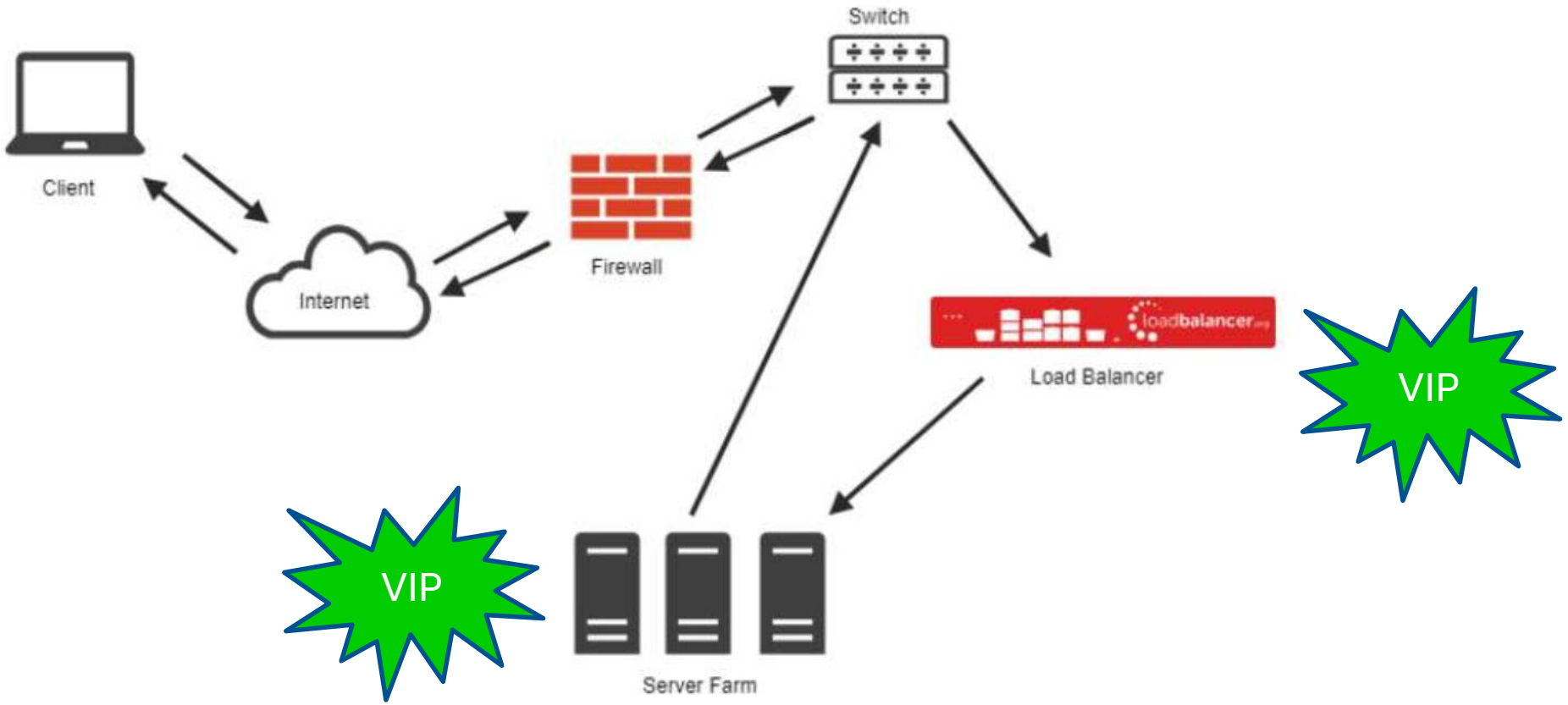
Quiz

Can you have 2 devices with same IP addresses in same subnet?

If no, Why not?

If yes, how can it?

Load Balancer & Server Farms Have Same Service IP Address



PRO: Great For Scalability

Consideration: Real Servers must host the service at VIP (rather than in RS IP)

Key Configuration for Demo

Role Of Dummy Interface In Direct Routing:

Packets from end users are forwarded directly to the real server. The IP packet is not modified

So, the real servers must be configured to accept traffic for the virtual server's IP address.

This can be done using a dummy interface or packet filtering to redirect traffic addressed to the virtual server's IP address to a local port.

The real server may send replies directly back to the end user.

Thus, the Load Balancer does not need to be in the return path.

net.ipv4.conf.lo.arp_ignore=1

net.ipv4.conf.lo.arp_ignore=1

```
[root@localhost user]# ./02_2-Machines_DRS_NGINX_Setup.sh
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
0.0.0.0         192.168.0.1     0.0.0.0         UG    100    0        0 enp129s0
172.16.0.0      0.0.0.0         255.255.255.0   U     100    0        0 enp59s0f0
192.168.0.0     0.0.0.0         255.255.255.0   U     100    0        0 enp129s0
192.168.122.0   0.0.0.0         255.255.255.0   U     0      0        0 virbr0
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet 172.16.0.10/32 scope global lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
net.ipv4.conf.lo.arp_ignore = 1
Kernel IP routing table
```

Call to Actions

1. Start evaluating Open Source Software Load Balancers, we can help
2. Looking for collaborators to develop and refine scalable solutions
3. Next Steps - Explore a micro-services implementation

Contacts

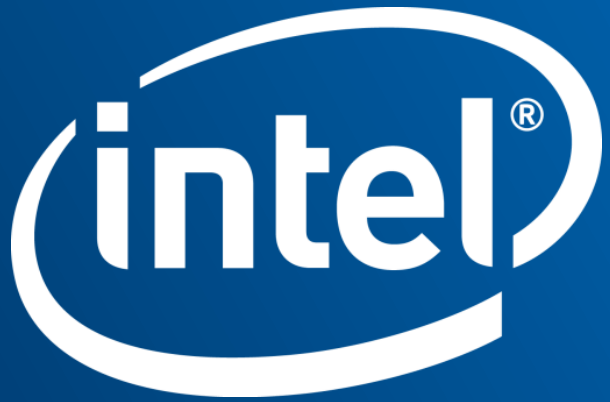
Jay Vincent - jay.l.vincent@intel.com
M Jay - Muthurajan.Jayakumar@intel.com

References

https://www.slideshare.net/ThomasGraf5/linuxcon-2015-linux-kernel-networking-walkthrough
http://www.linuxvirtualserver.org/software/ipvs.html#kernel-2.6
https://docs.fd.io/vpp/17.07/lb_plugin_doc.html
https://software.intel.com/en-us/article/get-the-dpdk-cookbook
https://ai.google/research/pubs/pub44824
http://ja.ssi.bg/#lvsgw

Additional Sessions

- Friday 1:50 – Accelerated Container Networking using DPDK – Sujata Tibewala & M. Jayakumar, Intel
- Friday 2:30 – VPP Accelerated High Performance & Scalable L3DSR L4 Load Balancer on Top Clos – Yusuke Tatsumi, Yahoo Japan Corp & Naoyuki Mori, Intel


intel

Demo Block Diagram - DPVS With IPVS