Linux as the foundation of a networking OS

Matty Kadosh

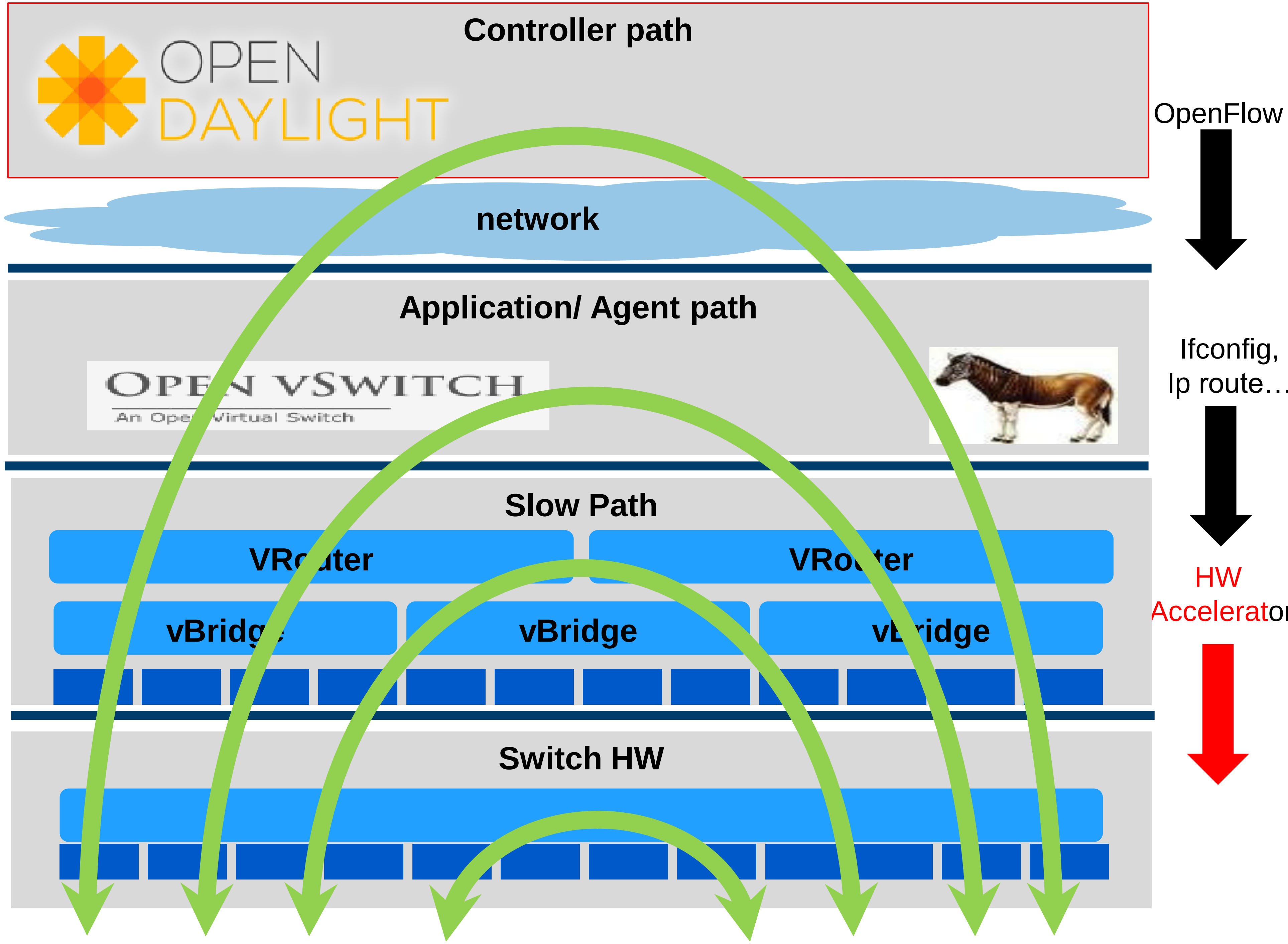
Connect. Accelerate. Outperform.™

Agenda - Linux as the foundation of a networking OS

- Provide a Linux-based open source networking OS
- a uniform OS for Ethernet switch/router boxes
- a uniform OS for eSwitch

Challenges :

- Fully functional SW base Data path (switch/router)
- HW as accelerator
  - HW driver/SDK will accelerate flows according to HW capability
  - Acceleration = optimization network should be fully functional without it
- Uniform API for Data path HW acceleration

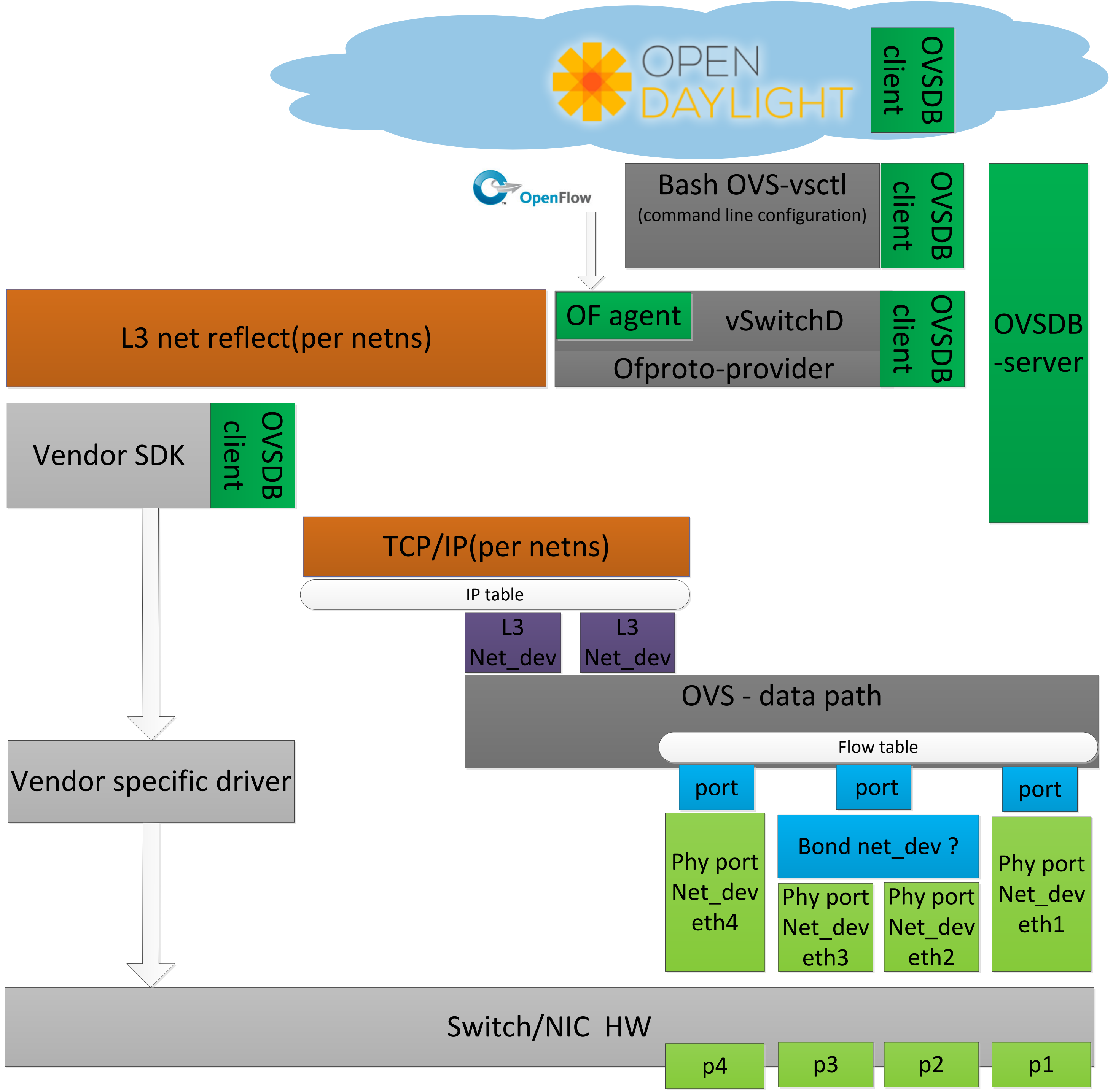


Networking data path building blocks

layer	Configuration	state	protocols
Router	static unicast router, static multicast router	unicast router, multicast router	OSPF, PIM, BGP, RIP, IGMP
L3 Interface	Interfaces, IP address, subnet, L3 type (vlan, port), MTU, static ARP, bond mode, LACP attributes, Sflow	interface state , ARP	VRRP, ARP, BFD, DHCP
Bridge	Ports, FDB aging time, static MAC, flood, broadcast, multicast FDB, MSTP vlan group, learning mode, span	Dynamic MAC table	IGMP snooping, xSTP, MLAG LACP
Port	Interfaces, vlan_mode, PVID, allowed_vlans, bond mode, LACP attributes, STP attributes, Sflow	state, STP state, STP rule, statistics	LACP
Phy Interface	Admin state, speed, MTU, Flow control, buffers, prio to buffer, storm control, Sflow, ETS, TC	State, statistics, LACP state	LLDP, DCBX, QCN, flow control

Linux as networking OS suggested solution - current view

- L2 based on OPEN VSWITCH An Open Virtual Switch
- Reflection to HW
  - OVSDB for switch configuration
  - ofproto-provider for OF
- L3 net-reflector
  - Receive route, ARP via net-link
  - Configure the HW accordingly
- Device driver should expose net-dev per HW port
  - control traffic (e.g STP,LACP)
  - Exception (e.g HW flow miss)
- Linux Bond vs OVS bond ?



Networking data path building blocks-gaps(Red-missing Black-duplicate)

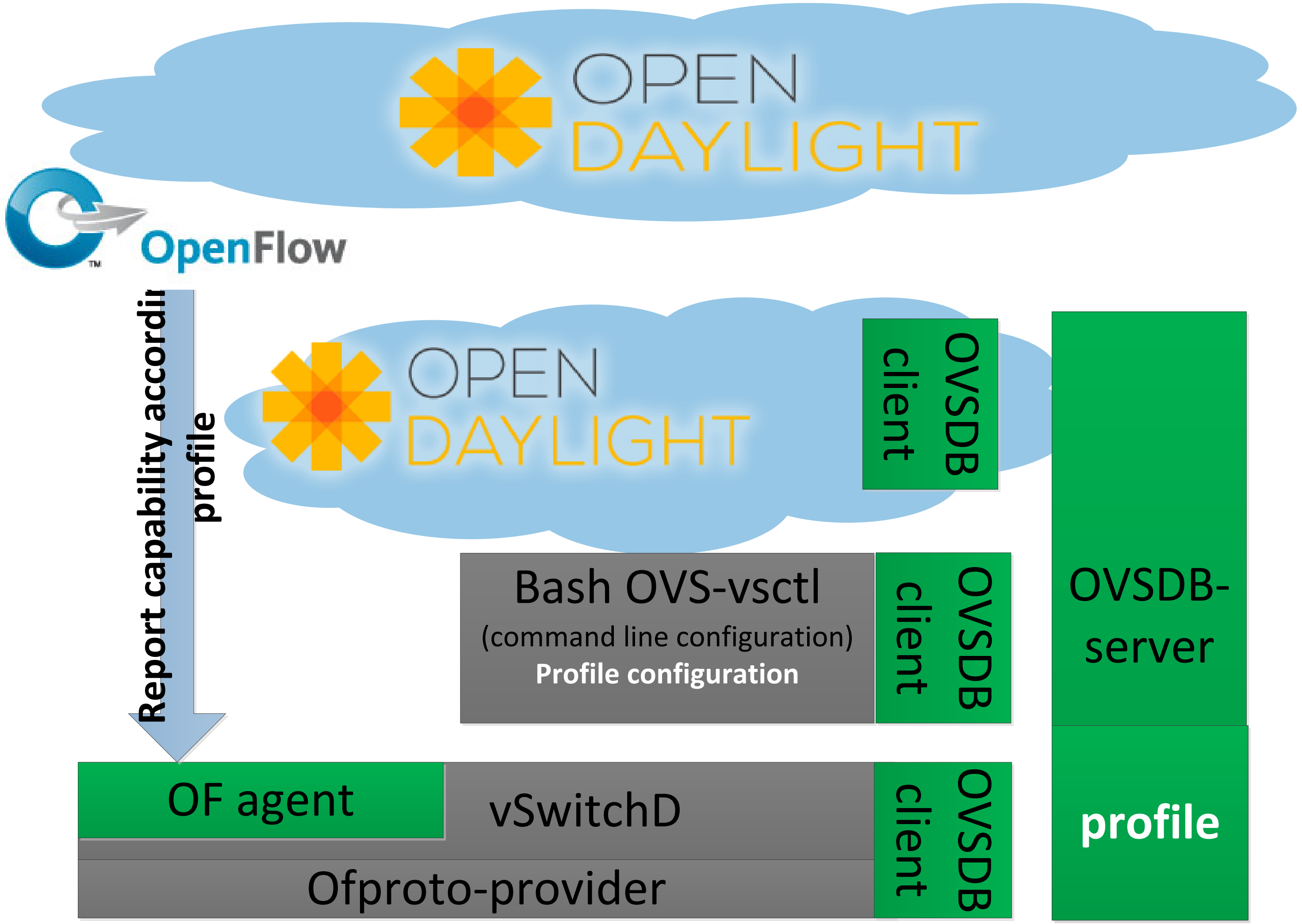
layer	Configuration	state	protocols
Router	static unicast router, static multicast router	unicast router, multicast router	OSPF, PIM, BGP, RIP, IGMP
L3 Interface	Interfaces, IP address, subnet, L3 type (vlan, port), MTU, static ARP, bond mode, LACP attributes, Sflow	interface state , ARP	VRRP, ARP, BFD, DHCP
Bridge	Ports, FDB aging time, static MAC, flood, broadcast, multicast FDB, MSTP vlan group, learning mode, span	Dynamic MAC table	STP, IGMP snooping, RSTP, MSTP, MLAG, LACP
Port	Interfaces, vlan_mode, PVID, ingress vlan filtering, ingress allowed vlans, egress allowed_vlans, bond mode, LACP attributes, STP attributes, Sflow	state, STP state, STP rule, statistics	LACP
Phy Interface	Admin state, speed, MTU, Flow control, buffers, prio to buffer, storm control, Sflow, ETS, TC	State, statistics, LACP state	LLDP, DCBX, QCN, flow control

Linux as networking OS - profiling Open vSwitch

- Different HW have different acceleration capability
- Admin should be able to control and profile the network e.g
  - Limit the SW base flow according to the HW capability
  - Limit the amount of “expensive” flows

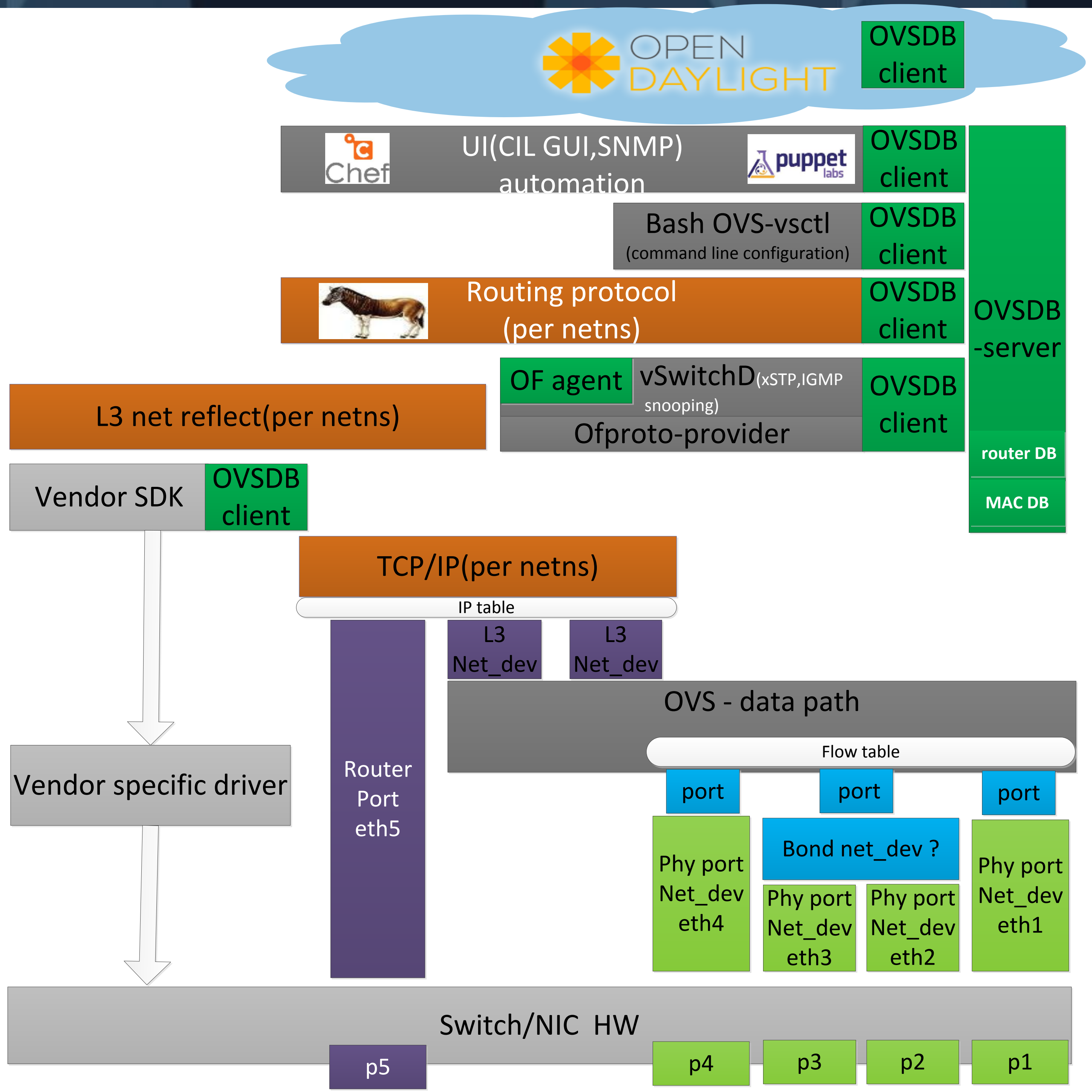
Solution: profile Open vSwitch

- expose the HW pipeline (e.g ACL, router, MAC table) & capability
- Limit the amount of tables/flow number and there action



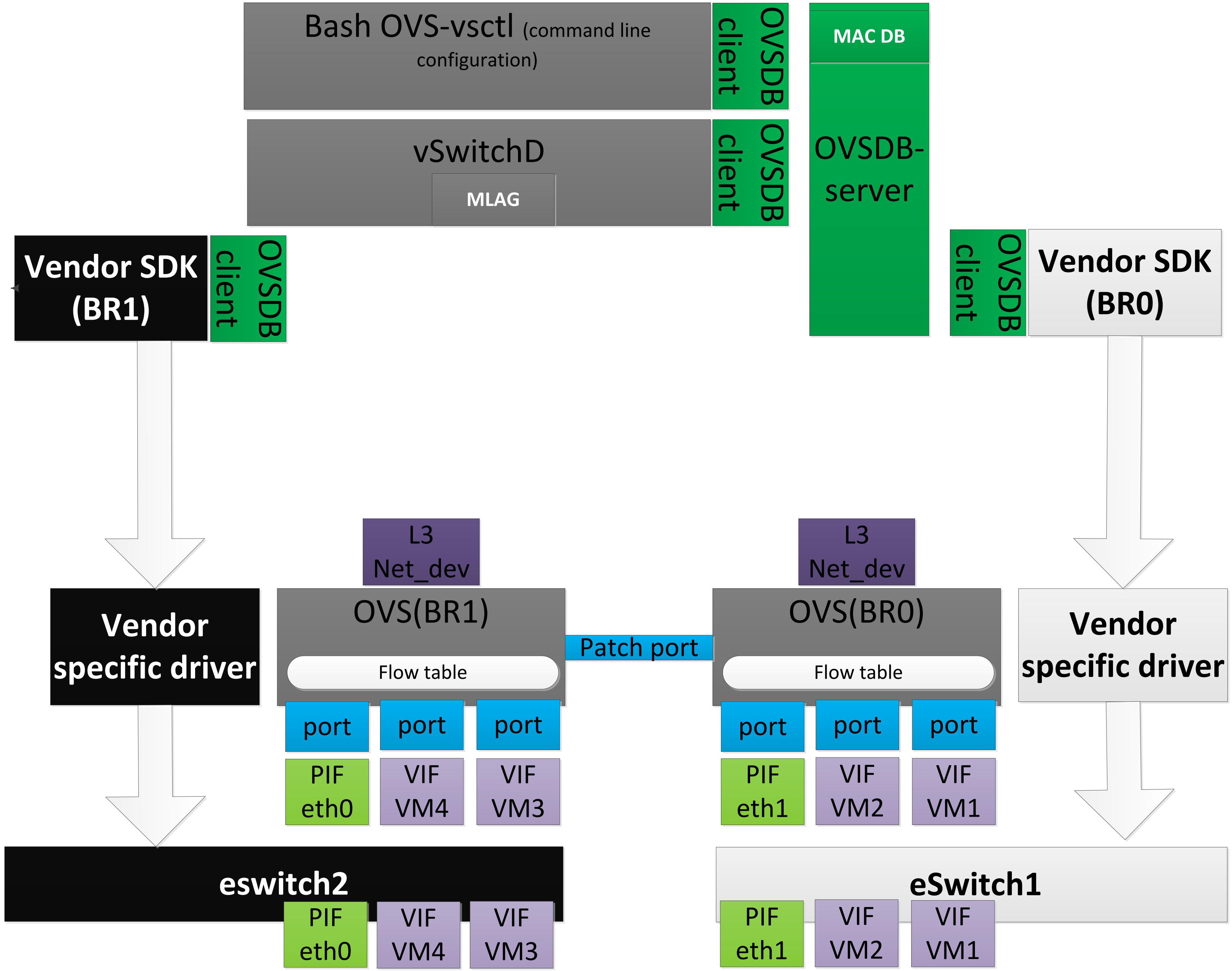
Linux as networking OS - next step

- profile Open vSwitch in order to expose the HW pipeline (e.g ACL, router, MAC table) & capability
- L3 interface modeling
  - In order to support router port over bond
  - L3 interface state reflection
    - Update L3 interface state according to Vlan port membership
- Full 802.1Q support
  - Ingress filtering configuration
  - Egress tagged / untagged membership
- Missing Protocol
  - IGMP snooping
  - RSTP/MSTP
  - MLAG
- Extend OVSDB
  - Static MAC
  - Router configuration
  - Box management



Linux as networking OS - SRIOV view

- Device driver should expose net-dev per VM
  - control traffic LLDP,EVB …
  - VXlan exception …
- MLAG (multi chassis link aggregation)in order to bond two PIF (eth0,eth1)



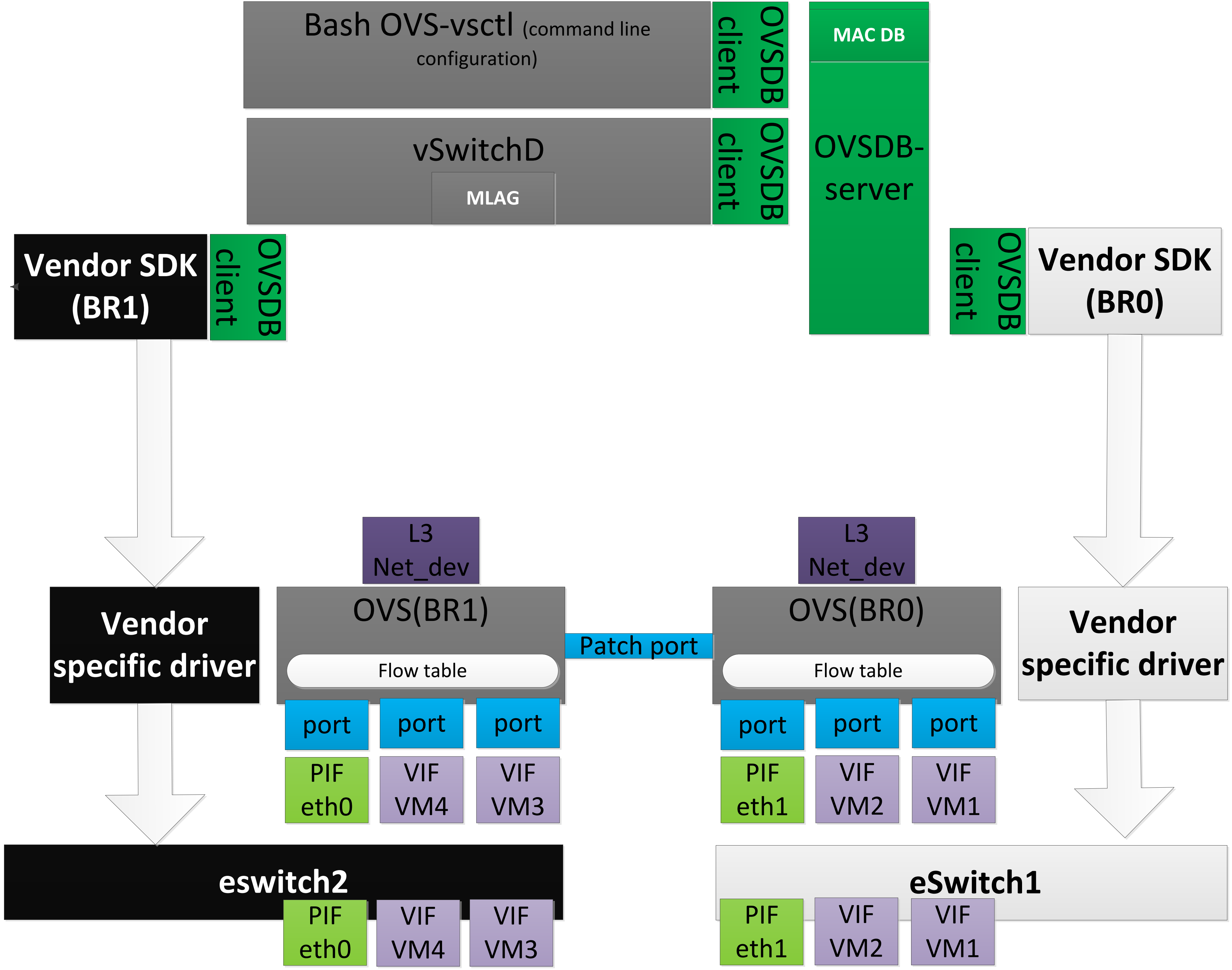
Thank You

Mellanox®
TECHNOLOGIES

Connect. Accelerate. Outperform.™

Linux as networking OS –OVS MLAG

- Device driver should expose net-dev per HW port
  - control traffic (LLDP)
  - VXlan exception …
- MLAG (multi chassis link aggregation)in order to bond two PIF (eth0,eth1)



Bash OVS-vsctl (command line configuration)
OVSDB client
MAC DB
vSwitchD
MLAG
OVSDB client
OVSDB-server
Vendor SDK (BR1)
OVSDB client
OVSDB client
Vendor SDK (BR0)
L3 Net_dev
L3 Net_dev
Vendor specific driver
OVS(BR1)
Flow table
Patch port
OVS(BR0)
Flow table
Vendor specific driver
port
port
port
port
port
port
PIF eth0
VIF VM4
VIF VM3
PIF eth1
VIF VM2
VIF VM1
eswitch2
PIF eth0
VIF VM4
VIF VM3
eSwitch1
PIF eth1
VIF VM2
VIF VM1