ESXi Network Internals

Overview of the Tcpip network stack

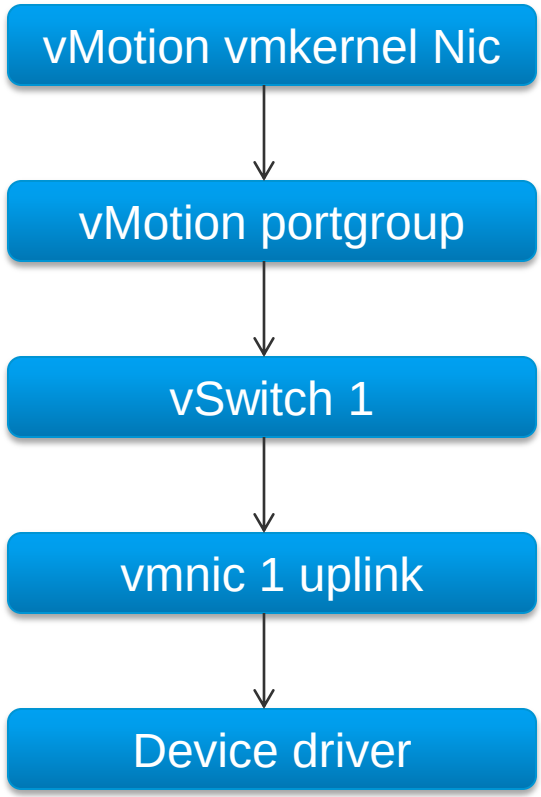
Arun Ramanathan

Agenda

Network stack

1. freebsd stack
   - TCP/IP
   - ARP, Ethernet
2. Tx/Rx dispatch layer
   - vmk glue code
   - Tx/Rx dispatch queues
3. Switching layer
   - VSS, DVS, port groups
4. Net Sched - IO control layer
   - Manages the uplink queue
5. vmk Linux layer
   - Linux networking, NAPI – networking APIs
   - Drivers – Eg. bnx2net

Net Stack



Software

Hardware

Net Stack

User world

vmkernel

vMotion

freebsd Layer

Tx/Rx dispatch layer

Switching Layer

Net Sched Layer

vmk linux layer

Device driver

Mbuf

vmk_PktHandle

vmk_PktHandle

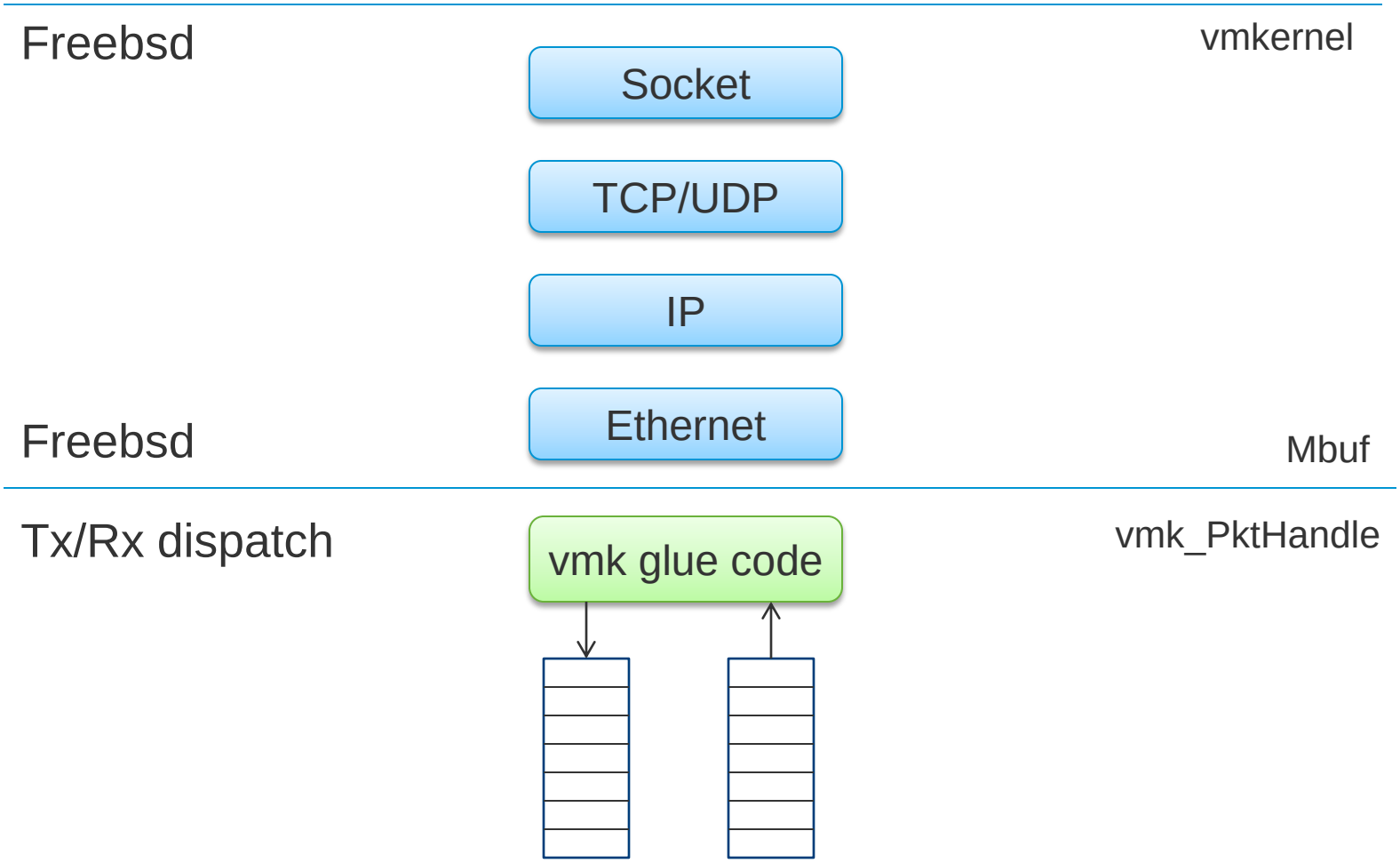
sk_buff

vmkernel

Hardware

Net Stack Layers

1. Freebsd Layer



1. Freebsd Layer

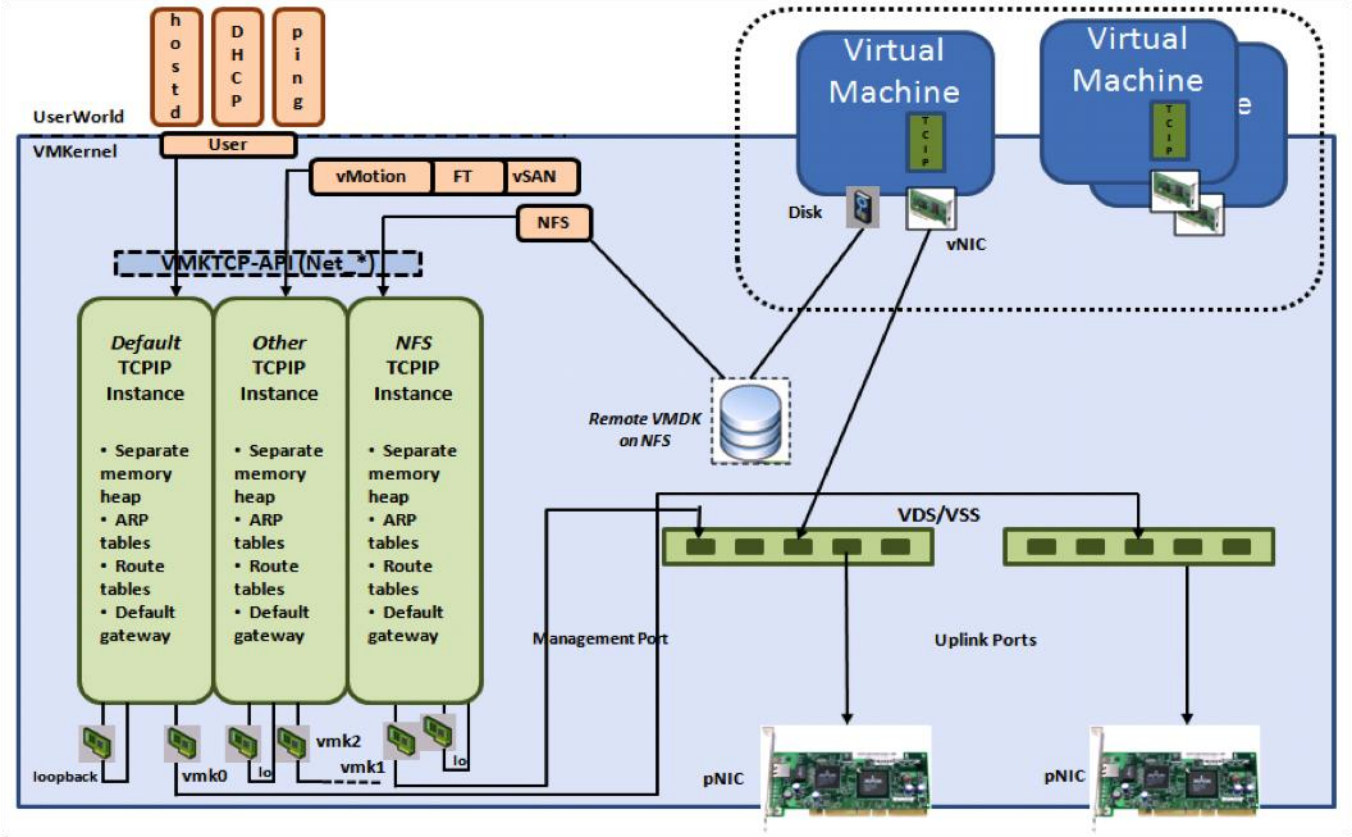
- Inherited from freebsd
- Consists of tcp/udp, ip and ethernet layers
- Responsible for constructing the ethernet frame
- Uses message buffer i.e Mbuf

- vMotion uses Net_ API to interact with freebsd layer
  - Net_SendTo, Net_SendMbufTo etc

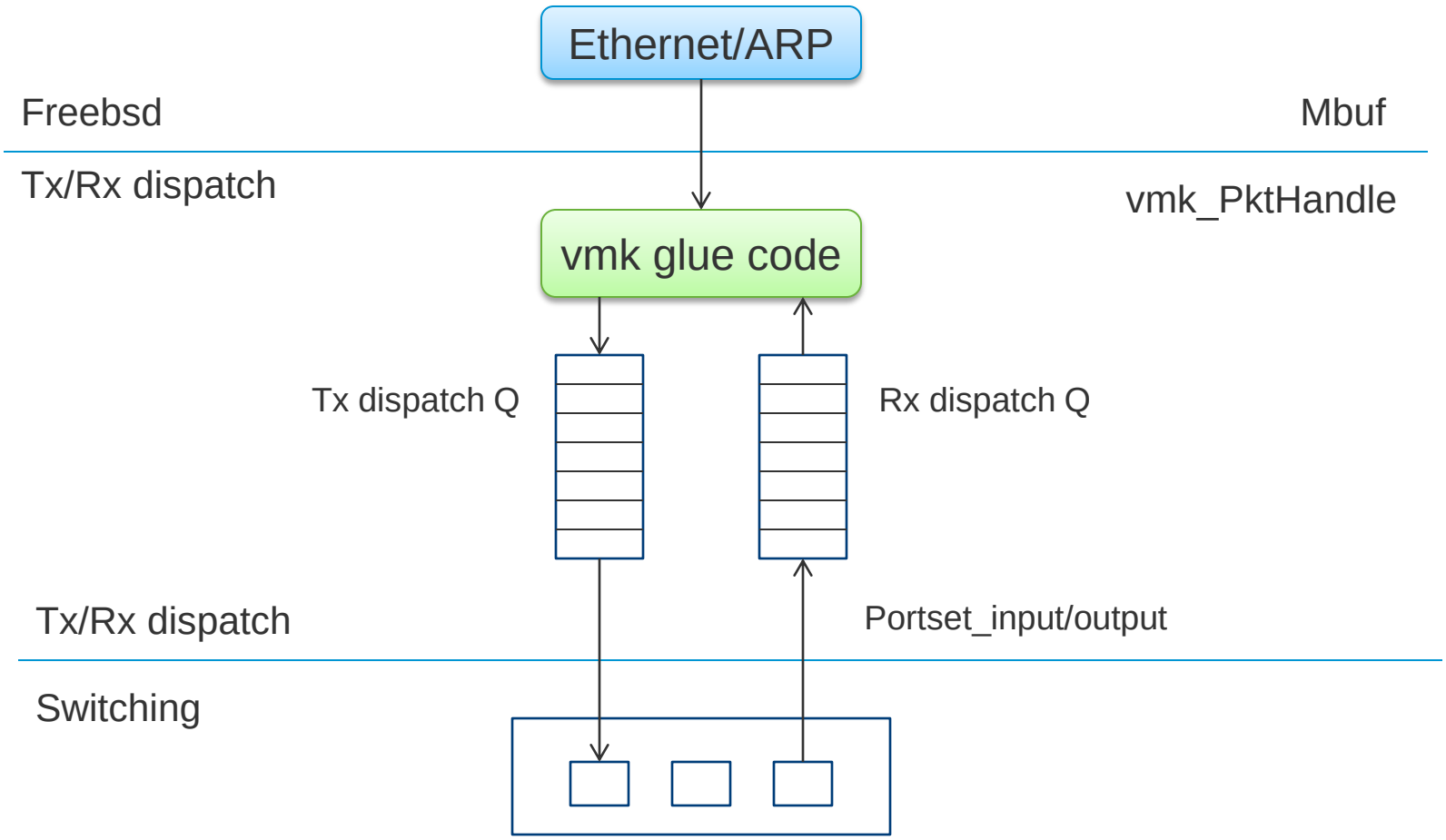
- Virtualized: Supports multiple netstack instances
  - Now vMotion has its own instance

- Possible queuing in the ethernet layer if ARP is not found

1. Freebsd Layer

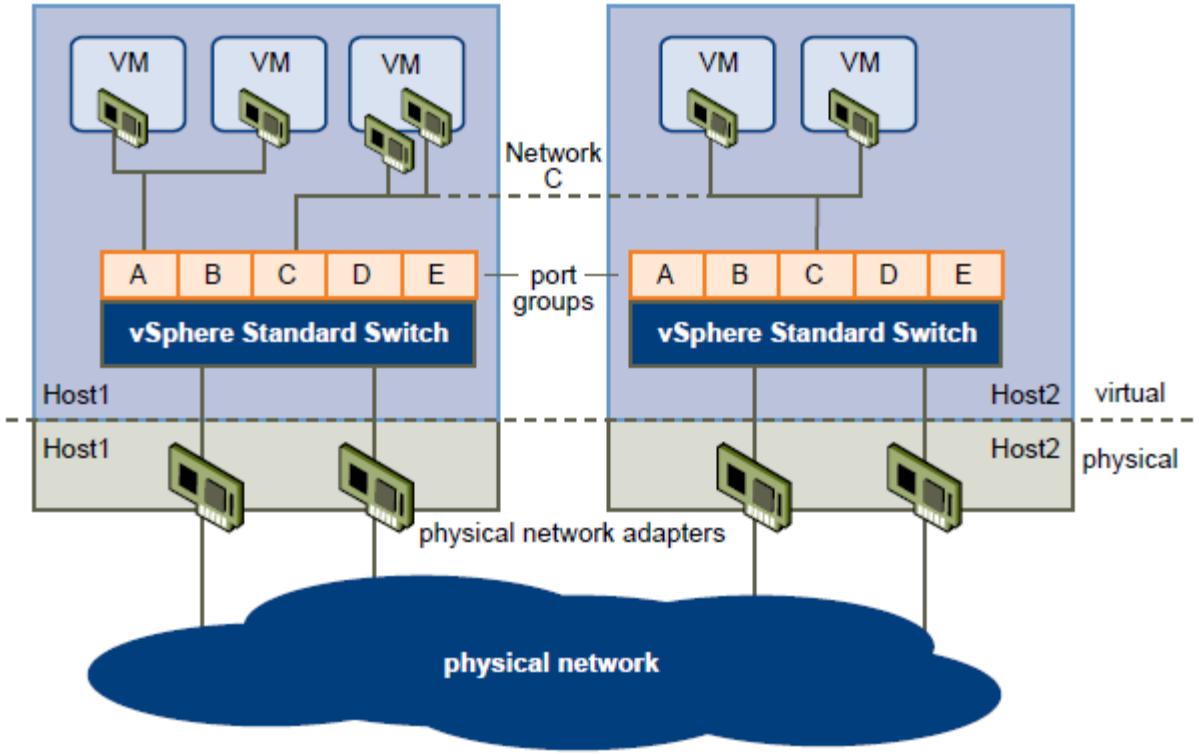


2. Tx/Rx Dispatch Layer

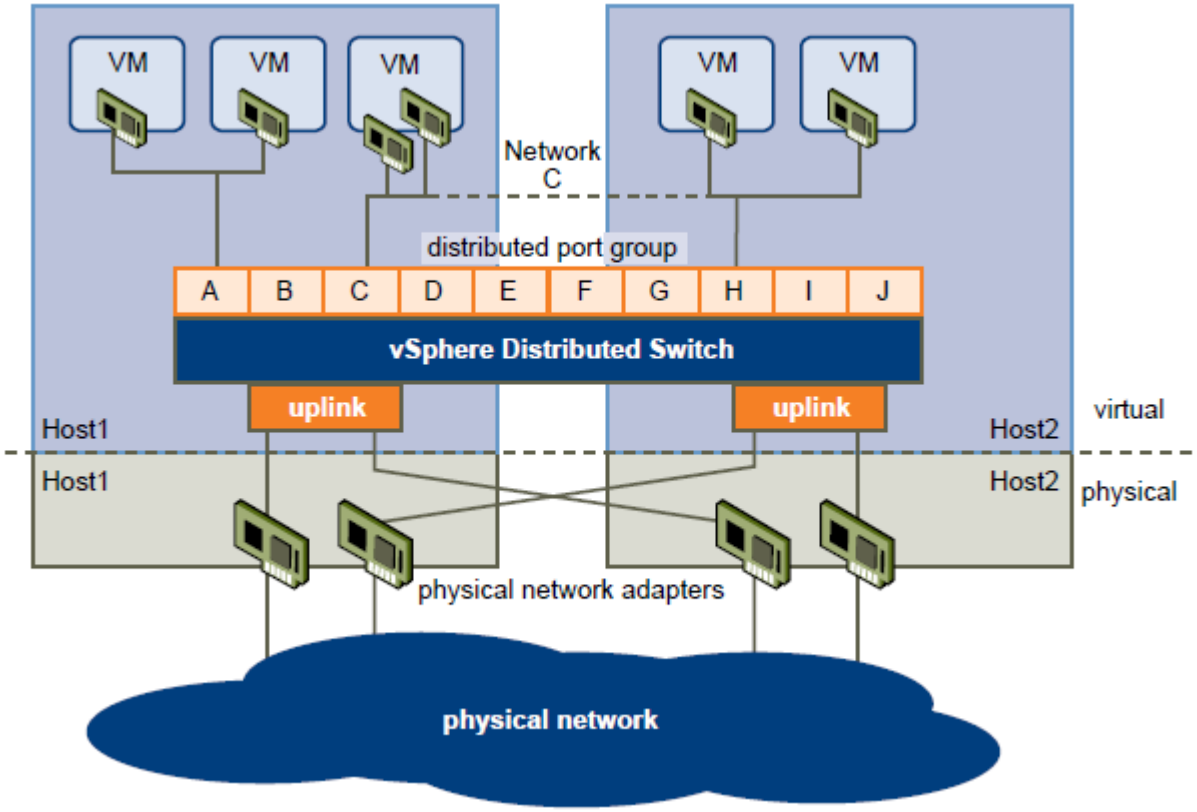


3. Switching Layer

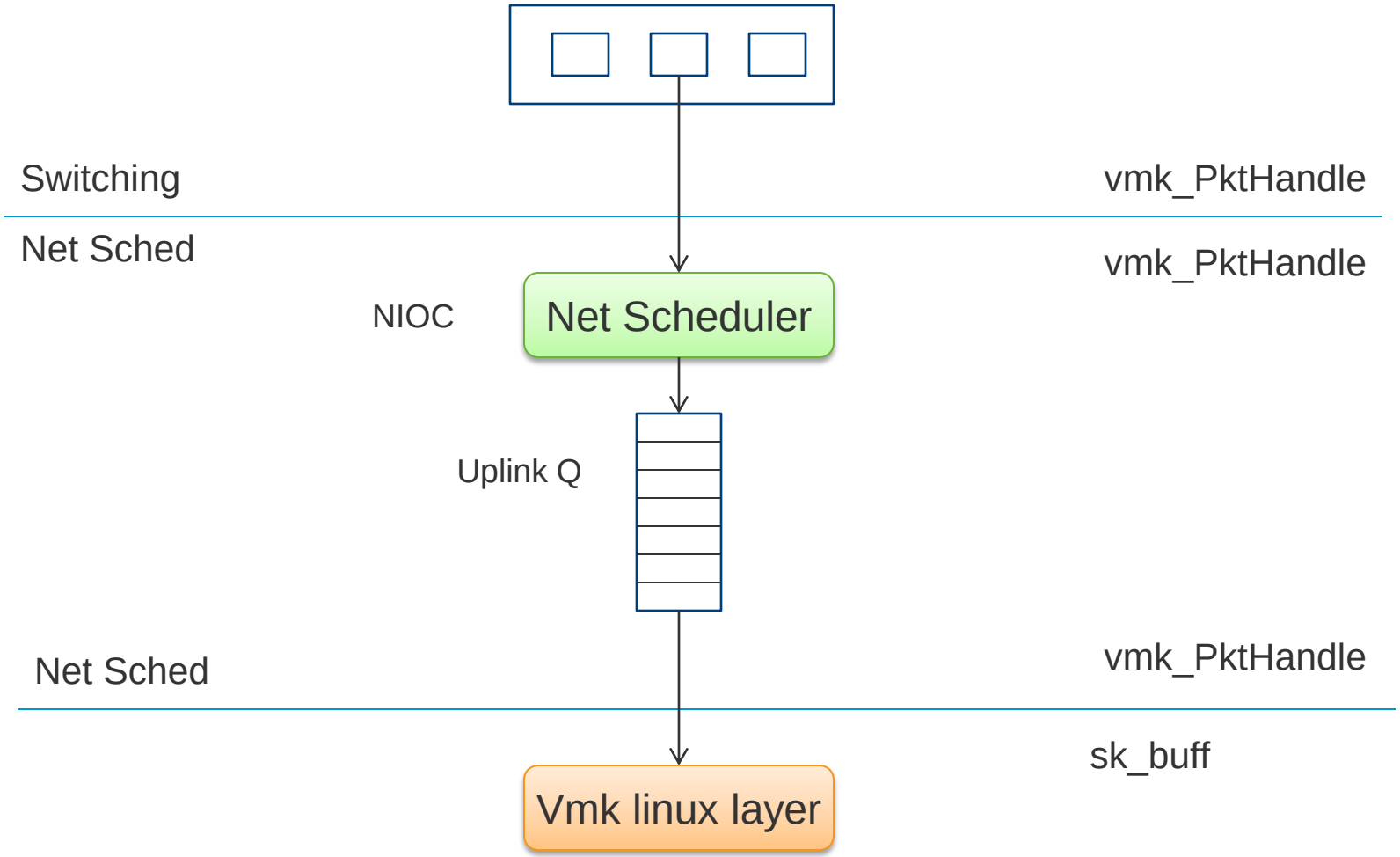
vSphere standard switch



3. Switching Layer



4. Net Sched Layer



5. Vmk Linux layer