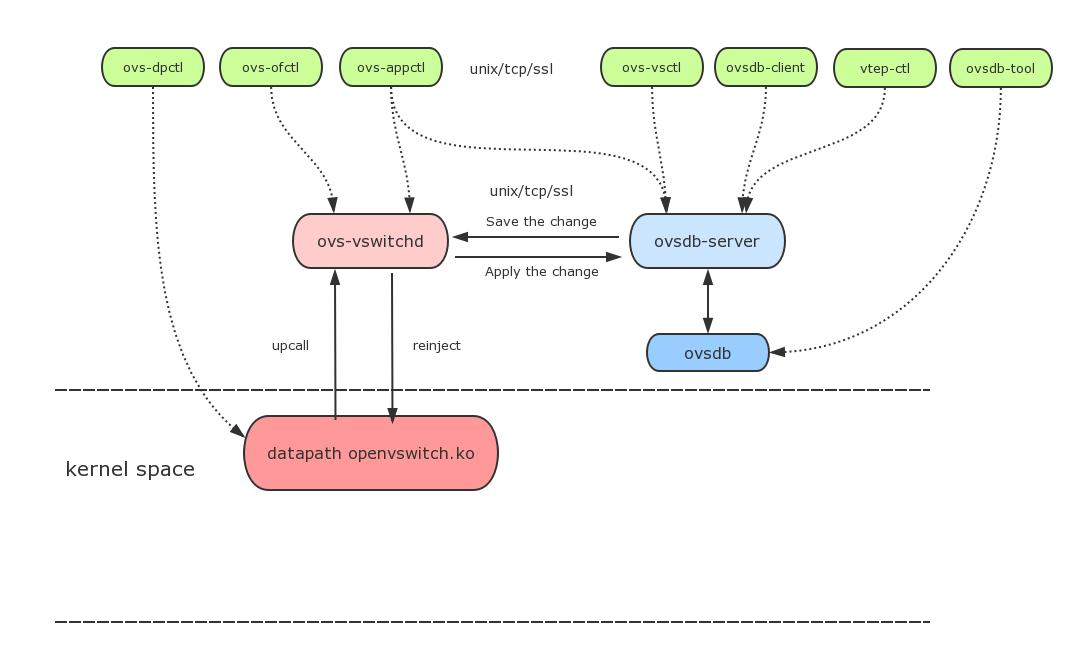
# ovs 的基本工作原理



# ovs 中的基本概念

todo

ofproto\_class 和 ofproto\_class

## dpif\_class 和 dpif

目前有两种类型的 dpif：dpif\_netlink 和 dpif\_netdev，它们都是在struct dpif 基础上扩展的。

dpif\_netlink 结构：

|  |
| --- |
| (lib/dpif-netlink.c)  /\* Datapath interface for the openvswitch Linux kernel module. \*/  struct dpif\_netlink {  struct dpif dpif;  int dp\_ifindex;  uint32\_t user\_features;  /\* Upcall messages. \*/  struct fat\_rwlock upcall\_lock;  struct dpif\_handler \*handlers;  uint32\_t n\_handlers; /\* Num of upcall handlers. \*/  /\* Per-vport dispatch mode. \*/  struct dpif\_channel \*channels; /\* Array of channels for each port. \*/  int uc\_array\_size; /\* Size of 'handler->channels' and \*/  /\* 'handler->epoll\_events'. \*/  /\* Change notification. \*/  struct nl\_sock \*port\_notifier; /\* vport multicast group subscriber. \*/  bool refresh\_channels;  }; |

dpif\_netdev 结构：

|  |
| --- |
| (lib/dpif-netdev.c)  /\* Interface to netdev-based datapath. \*/  struct dpif\_netdev {  struct dpif dpif;  struct dp\_netdev \*dp;  uint64\_t last\_port\_seq;  }; |

## dpif\_backer

dpif\_backer 表示交换机和 datapath 的接口，相同 type 的 datapath 共享一个 backer，Linux 内核支持两种 type 的 datapath ：”system” 和 “netdev”，所以就有两种类型的 datapath 接口：默认的 dpif\_netlink , 以及 DPDK 下使用的 dpif\_netdev。

在 open\_dpif\_backer() 中创建，在 close\_dpif\_backer() 中销毁。

|  |
| --- |
| （ofproto/ofproto-dpif.h）  /\* All datapaths of a given type share a single dpif backer instance. \*/  struct dpif\_backer {  char \*type;  int refcount;  struct dpif \*dpif;  struct udpif \*udpif;  struct ovs\_rwlock odp\_to\_ofport\_lock;  struct hmap odp\_to\_ofport\_map OVS\_GUARDED; /\* Contains "struct ofport"s. \*/  struct simap tnl\_backers; /\* Set of dpif ports backing tunnels. \*/  enum revalidate\_reason need\_revalidate; /\* Revalidate all flows. \*/  bool recv\_set\_enable; /\* Enables or disables receiving packets. \*/  /\* Meter. \*/  struct id\_pool \*meter\_ids; /\* Datapath meter allocation. \*/  /\* Connection tracking. \*/  struct id\_pool \*tp\_ids; /\* Datapath timeout policy id  \* allocation. \*/  struct cmap ct\_zones; /\* "struct ct\_zone"s indexed by zone  \* id. \*/  struct hmap ct\_tps; /\* "struct ct\_timeout\_policy"s indexed  \* by timeout policy (struct simap). \*/  struct ovs\_list ct\_tp\_kill\_list; /\* A list of timeout policy to be  \* deleted. \*/  /\* Version string of the datapath stored in OVSDB. \*/  char \*dp\_version\_string;  /\* Datapath feature support. \*/  struct dpif\_backer\_support bt\_support; /\* Boot time support. Set once  when vswitch starts up, then  it is read only through out  the life time of vswitchd. \*/  struct dpif\_backer\_support rt\_support; /\* Runtime support. Can be  set to a lower level in  feature than 'bt\_support'. \*/  struct atomic\_count tnl\_count;  }; |

# 问题：ovsdb 中的虚拟网桥和虚拟端口什么时候同步到内核？

ovs-vsctl 只会修改 ovsdb 中的信息，同步到内核是 vswitchd 做的。答案是在 vswitchd 中，会循环调用 bridge\_run() 和 netdev\_run()，这两个函数中会做同步：

|  |
| --- |
| （vswitchd/ovs-vswitchd.c）  int  main(int argc, char \*argv[])  {  …  while (!exiting) {  …  bridge\_run();  …  netdev\_run();  }  } |

# vswitchd 内存中网桥信息的更新(all\_bridges)

vswitchd 中使用的信息是内存中的副本，vswitchd 会从 ovsdb 从同步一份信息到内存中，然后使用内存中的信息。

vswitchd 在 bridge\_reconfigure() 函数中同步 ovsdb 的信息到内存。

## all\_bridges 与 bridge 同步

vswtichd 中把所有的 bridge 信息保存在 all\_bridges 中：

|  |
| --- |
| （vswitchd/bridge.c）  /\* All bridges, indexed by name. \*/  static struct hmap all\_bridges = HMAP\_INITIALIZER(&all\_bridges); |

这是一个哈希表，通过 bridge\_name 索引，struct bridge 结构通过 node 成员挂到哈希表上。

struct bridge 结构：

|  |
| --- |
| （vswitchd/bridge.c）  struct bridge {  struct hmap\_node node; /\* In 'all\_bridges'. \*/  char \*name; /\* User-specified arbitrary name. \*/  char \*type; /\* Datapath type. \*/  struct eth\_addr ea; /\* Bridge Ethernet Address. \*/  struct eth\_addr default\_ea; /\* Default MAC. \*/  // 指向 ovsdb 中对应的行  const struct ovsrec\_bridge \*cfg;  /\* OpenFlow switch processing. \*/  struct ofproto \*ofproto; /\* OpenFlow switch. \*/  /\* Bridge ports. \*/  struct hmap ports; /\* "struct port"s indexed by name. \*/  struct hmap ifaces; /\* "struct iface"s indexed by ofp\_port. \*/  struct hmap iface\_by\_name; /\* "struct iface"s indexed by name. \*/  /\* Port mirroring. \*/  struct hmap mirrors; /\* "struct mirror" indexed by UUID. \*/  /\* Auto Attach \*/  struct hmap mappings; /\* "struct" indexed by UUID \*/  /\* Used during reconfiguration. \*/  struct shash wanted\_ports;  /\* Synthetic local port if necessary. \*/  struct ovsrec\_port synth\_local\_port;  struct ovsrec\_interface synth\_local\_iface;  struct ovsrec\_interface \*synth\_local\_ifacep;  }; |

Bridge 同步在 add\_del\_bridges() 函数中，对于 vswitchd 内存中存在的但 ovsdb 中不存在的，调用 bridge\_destroy() 从 vswitchd 的内存中删除，对于 vswitchd 内存中不存在但 ovsdb 中存在的，调用 bridge\_create() 添加到 vswitchd 内存中。

## port 同步

add\_del\_bridges() 函数只是同步了 bridge 信息，下面看下 port 信息的同步。

在 struct bridge 中用一个 hmap 保存了所有 port 信息：

|  |
| --- |
| （vswitchd/bridge.c）  struct bridge {  …  /\* Bridge ports. \*/  struct hmap ports; /\* "struct port"s indexed by name. \*/  …  } |

port 在 bridge 中的添加和删除分别由函数 port\_create() 和 port\_destroy() 中完成：

port\_create()：

|  |
| --- |
| （vswitchd/bridge.c）  static struct port \*  port\_create(struct bridge \*br, const struct ovsrec\_port \*cfg)  {  struct port \*port;  port = xzalloc(sizeof \*port);  port->bridge = br;  port->name = xstrdup(cfg->name);  port->cfg = cfg;  ovs\_list\_init(&port->ifaces);  // 插入 port 到 br->ports 中  hmap\_insert(&br->ports, &port->hmap\_node, hash\_string(port->name, 0));  return port;  } |

port\_destroy()：

|  |
| --- |
| （vswitchd/bridge.c）  static void  port\_destroy(struct port \*port)  {  if (port) {  struct bridge \*br = port->bridge;  struct iface \*iface;  if (br->ofproto) {  ofproto\_bundle\_unregister(br->ofproto, port);  }  // 遍历 destroy port 下的 interfaces  LIST\_FOR\_EACH\_SAFE (iface, port\_elem, &port->ifaces) {  iface\_destroy\_\_(iface);  }  // 从 br->ports 中删除  hmap\_remove(&br->ports, &port->hmap\_node);  free(port->name);  free(port);  }  } |

port\_create() 调用链条为：

* bridge\_reconfigure() （vswitchd/bridge.c）
* bridge\_add\_ports() （vswitchd/bridge.c）
* bridge\_add\_ports\_\_() （vswitchd/bridge.c）

port\_destroy() 调用链条为：

* bridge\_reconfigure() （vswitchd/bridge.c）
* bridge\_del\_ports() （vswitchd/bridge.c）

## datapath 同步

vswitchd 中所有的 datapath 也保存在一个名为 all\_datapaths 的全局变量中：

|  |
| --- |
| （vswitchd/bridge.c）  /\* All datapath configuartions, indexed by type. \*/  static struct hmap all\_datapaths = HMAP\_INITIALIZER(&all\_datapaths); |

这也是类似 all\_bridges 的一个哈希表，key 是 type。

Datapath 通过 datapath\_create() 和 datapath\_destroy() 进行同步。

datapath\_create()：

|  |
| --- |
| （vswitchd/bridge.c）  static struct datapath \*  datapath\_create(const char \*type)  {  struct datapath \*dp = xzalloc(sizeof \*dp);  dp->type = xstrdup(type);  hmap\_init(&dp->ct\_zones);  // 添加到 all\_datapaths  hmap\_insert(&all\_datapaths, &dp->node, hash\_string(type, 0));  smap\_init(&dp->caps);  return dp;  } |

datapath\_destroy()：

|  |
| --- |
| （vswitchd/bridge.c）  static void  datapath\_destroy(struct datapath \*dp)  {  if (dp) {  struct ct\_zone \*ct\_zone;  HMAP\_FOR\_EACH\_SAFE (ct\_zone, node, &dp->ct\_zones) {  ofproto\_ct\_del\_zone\_timeout\_policy(dp->type, ct\_zone->zone\_id);  ct\_zone\_remove\_and\_destroy(dp, ct\_zone);  }  // 从 all\_datapaths 中删除  hmap\_remove(&all\_datapaths, &dp->node);  hmap\_destroy(&dp->ct\_zones);  free(dp->type);  smap\_destroy(&dp->caps);  free(dp);  }  } |

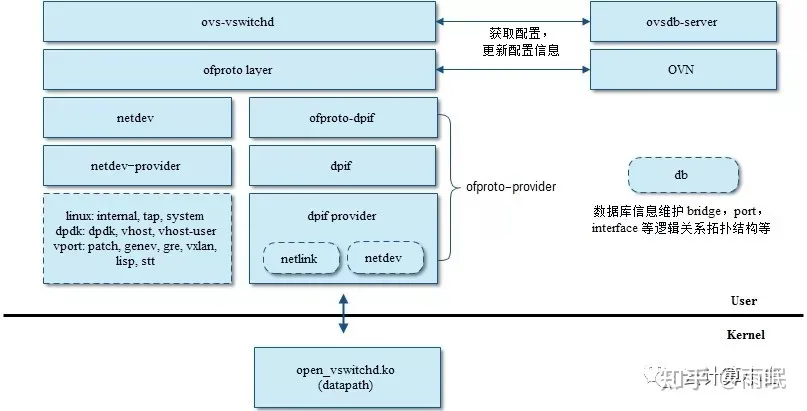
datapath\_create() 调用链条为：

* bridge\_reconfigure() （vswitchd/bridge.c）
* datapath\_reconfigure() （vswitchd/bridge.c）

datapath\_destroy() 调用链条为：

* bridge\_reconfigure() （vswitchd/bridge.c）
* datapath\_reconfigure() （vswitchd/bridge.c）

# Vswitchd



从图中可以看出 OVS 的分层结构，最上层 vswitchd 主要与 ovsdb 通信，做配置下发和更新等；中间层是 ofproto ，用于和 OpenFlow 控制器通信，并基于下层的 ofproto provider 提供的接口，完成具体的设备操作和流表操作等工作。netdev 层实现了对网络设备（如 Ethernet）的抽象，基于 netdev provider 接口实现多种不同平台的设备，如 Linux 内核的 system, tap, internal 等，dpdk 系的 vhost, vhost-user 等，以及隧道相关的 gre, vxlan 等。

vswitchd是ovs主要的用户态程序，它从ovsdb-server读取配置并发送到ofproto层，也从ofproto读取特定的状态和统计信息并发送到数据库；

ofproto是openflow的接口层，负责和Openflow controller通信并通过ofproto\_class与ofproto provider底层交互；

ofproto-dpif是ofproto接口类的具体实现；目前openvswitch只支持ofproto-dpif，但是用户可以很容易的实现其他ofproto provider。

dpif 层实现对流表的操作。

netdev是ovs系统的网络设备抽象（比如linux的net\_device或交换机的port），netdev\_class定义了netdev-provider的具体实现需要的接口，具体的平台实现需要支持这些统一的接口，从而完成netdev设备的创建、销毁、打开、关闭等一系列操作；

## ofproto\_class 、ofproto 和 ofproto\_dpif

在 struct bridge 中包含一个 ofproto 成员，其表示 OpenFlow 网桥的实现：

|  |
| --- |
| （vswitchd/bridge.c）  struct bridge {  ...  /\* OpenFlow switch processing. \*/  struct ofproto \*ofproto; /\* OpenFlow switch. \*/  ...  } |

ofproto 层可以有多个实际实现，代码中分为了接口和数据两部分，其中 struct ofproto\_class 对应接口，struct ofproto 对应数据。ofproto\_class 和 ofproto 都有具体的实现，目前只有一个实现，对应的接口是 ofproto\_dpif\_class，数据是 ofproto\_dpif。ofproto\_dpif 定义为：

|  |
| --- |
| （ofproto/ofproto-dpif.h）  struct ofproto\_dpif {  ...  struct ofproto up;  ...  } |

ofproto\_dpif 中封装了 ofproto，此外还封装了一些实现相关的内容。ofproto\_dpif 和 ofproto 的关系，类似内核中 inet\_sock 和 sock 结构的关系。

由于这种关系，ofproto 实际是由 ofproto\_class 中的 alloc 函数进行分配的，如 ofproto\_dpif\_class 的实现：

|  |
| --- |
| （ofproto/ofproto-dpif.c）  static struct ofproto \*  alloc(void)  {  struct ofproto\_dpif \*ofproto = xzalloc(sizeof \*ofproto);  return &ofproto->up;  } |

可以看到，实际分配的是 ofproto\_dpif 结构，只不过返回时返回的是其中的 ofproto 地址。而且 ofproto\_class 的接口中也都是使用的 ofproto 结构。

**bridge->ofproto 初始化**

bridge->ofproto 初始化在 bridge\_reconfigure() 函数中：

|  |
| --- |
| （vswitchd/bridge.c）  static void  bridge\_reconfigure(const struct ovsrec\_open\_vswitch \*ovs\_cfg)  {  …  /\* Finish pushing configuration changes to the ofproto layer:  \*  \* - Create ofprotos that are missing.  \*  \* - Add ports that are missing. \*/  HMAP\_FOR\_EACH\_SAFE (br, node, &all\_bridges) {  if (!br->ofproto) {  int error;  // 调用 ofproto\_create() 初始化 bridge->ofproto  error = ofproto\_create(br->name, br->type, &br->ofproto);  if (error) {  VLOG\_ERR("failed to create bridge %s: %s", br->name,  ovs\_strerror(error));  shash\_destroy(&br->wanted\_ports);  bridge\_destroy(br, true);  } else {  /\* Trigger storing datapath version. \*/  seq\_change(connectivity\_seq\_get());  }  }  }  …  } |

# 内核部分

## datapath

一个网桥对应一个内核态的 datapath 结构。

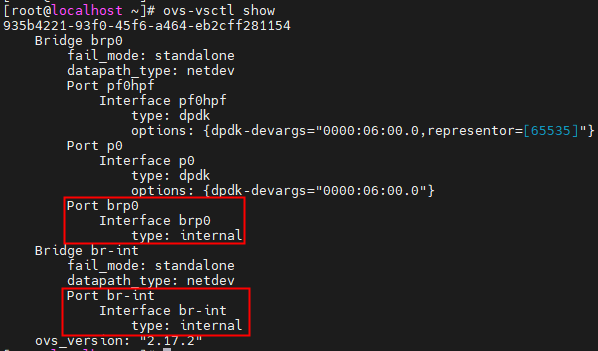
## vport

端口的类型定义在：

|  |
| --- |
| （linux/compat/include/linux/openvswitch.h）  enum ovs\_vport\_type {  OVS\_VPORT\_TYPE\_UNSPEC,  OVS\_VPORT\_TYPE\_NETDEV, /\* network device \*/  OVS\_VPORT\_TYPE\_INTERNAL, /\* network device implemented by datapath \*/  OVS\_VPORT\_TYPE\_GRE, /\* GRE tunnel. \*/  OVS\_VPORT\_TYPE\_VXLAN, /\* VXLAN tunnel. \*/  OVS\_VPORT\_TYPE\_GENEVE, /\* Geneve tunnel. \*/  OVS\_VPORT\_TYPE\_LISP = 105, /\* LISP tunnel \*/  OVS\_VPORT\_TYPE\_STT = 106, /\* STT tunnel \*/  \_\_OVS\_VPORT\_TYPE\_MAX  }; |

以下是各种类型端口的作用：

* OVS\_VPORT\_TYPE\_INTERNAL：每个OVS交换机有个可以用来处理数据报的本地端口，可以为这个网络设备配置 IP 地址。(比方在把eth0增加某个bridge的时候，它的IP地址就失效了。能够把IP地址赋给br，这就是internal port的地址)等等。每个网桥创建时会默认创建一个同名的 internal 端口，例如：



## OVS\_VPORT\_TYPE\_INTERNAL 相关逻辑

下面是 internal 类型的 vport 在内核模块的逻辑。

### 创建和删除

如前文所述，internal 类型的 port 是和网桥的生命周期绑定的，因此，其创建和删除是在 datapath（内核态表示网桥）的创建和删除中。

**创建**

ovs\_dp\_cmd\_new() 中（响应 OVS\_DP\_CMD\_NEW 的 netlink 消息）：

|  |
| --- |
| （datapath.c）  static int ovs\_dp\_cmd\_new(struct sk\_buff \*skb, struct genl\_info \*info)  {  …  /\* Set up our datapath device. \*/  // 创建 internal port  parms.name = nla\_data(a[OVS\_DP\_ATTR\_NAME]); // 名称和 datapath 相同  parms.type = OVS\_VPORT\_TYPE\_INTERNAL; // 类型是 internal  parms.options = NULL;  parms.dp = dp;  parms.port\_no = OVSP\_LOCAL; // 端口编号是特殊的 OVSP\_LOCAL  parms.upcall\_portids = a[OVS\_DP\_ATTR\_UPCALL\_PID];  …  vport = new\_vport(&parms); |

**删除**

ovs\_dp\_cmd\_del()->\_\_dp\_destroy() 中（响应 OVS\_DP\_CMD\_DEL 的 netlink 消息）：

|  |
| --- |
| （datapath.c）  /\* Called with ovs\_mutex. \*/  static void \_\_dp\_destroy(struct datapath \*dp)  {  int i;  // 先删除非 internal 的端口  for (i = 0; i < DP\_VPORT\_HASH\_BUCKETS; i++) {  struct vport \*vport;  struct hlist\_node \*n;  hlist\_for\_each\_entry\_safe(vport, n, &dp->ports[i], dp\_hash\_node)  if (vport->port\_no != OVSP\_LOCAL)  ovs\_dp\_detach\_port(vport);  }  list\_del\_rcu(&dp->list\_node);  /\* OVSP\_LOCAL is datapath internal port. We need to make sure that  \* all ports in datapath are destroyed first before freeing datapath.  \*/  // 删除 internal 端口  ovs\_dp\_detach\_port(ovs\_vport\_ovsl(dp, OVSP\_LOCAL));  /\* RCU destroy the flow table \*/  call\_rcu(&dp->rcu, destroy\_dp\_rcu);  } |