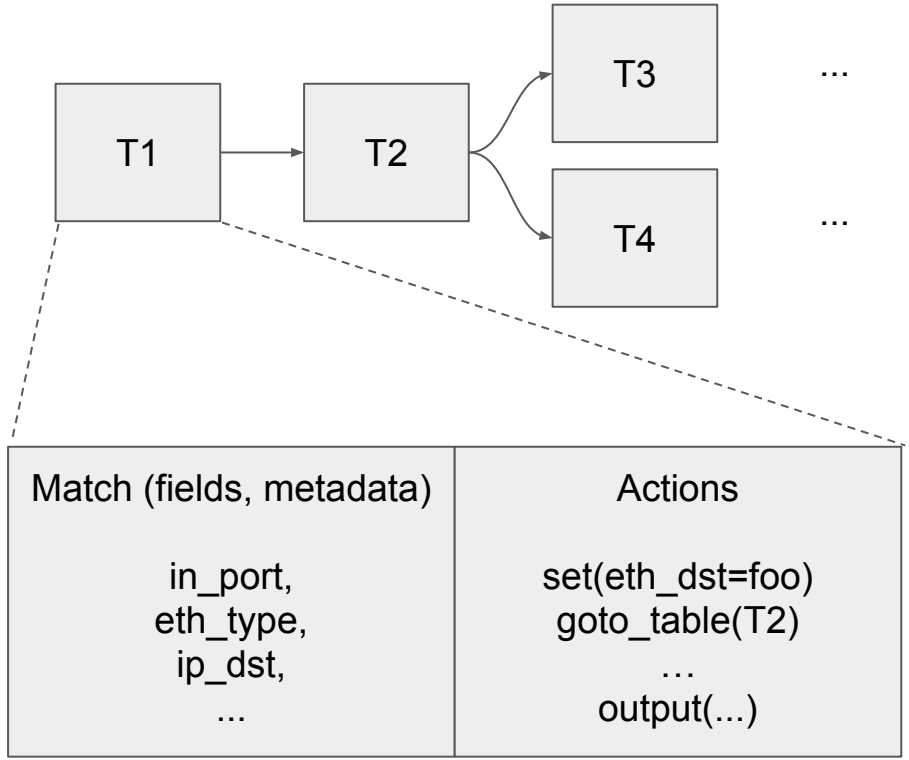
OVS and Netfilter

Joe Stringer, VMware

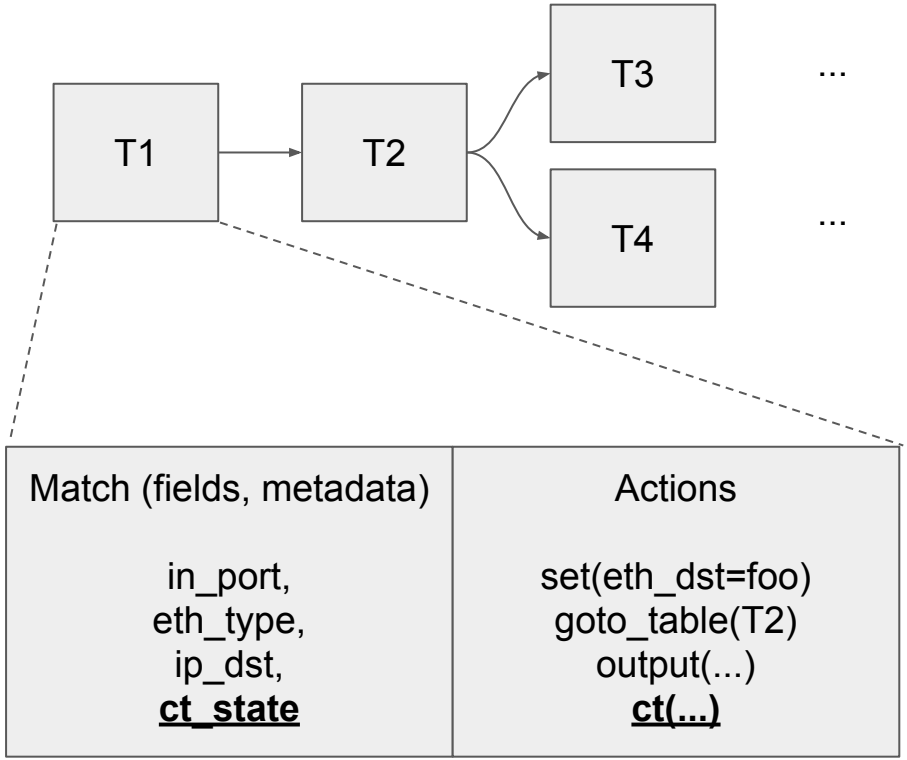
Overview

- OVS Connection Tracking Recap
- NAT
- Load balancing
- Wishlist

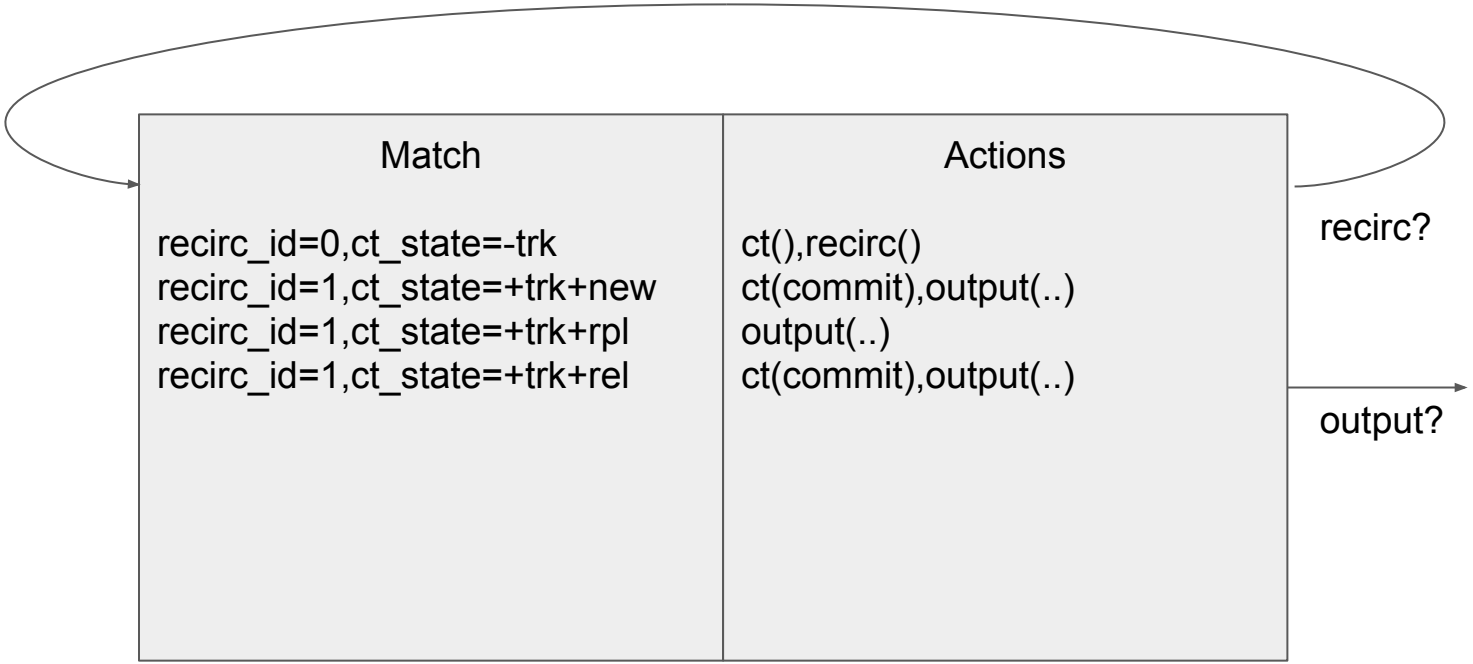
OVS: userspace



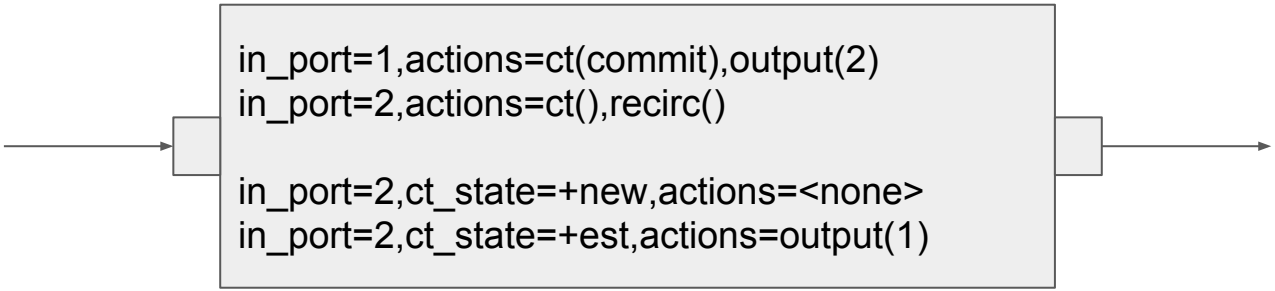
OVS: userspace



OVS: kernel


Match
recirc_id=0,ct_state=-trk
recirc_id=1,ct_state=+trk+new
recirc_id=1,ct_state=+trk+rpl
recirc_id=1,ct_state=+trk+rel
Actions
ct(),recirc()
ct(commit),output(..)
output(..)
ct(commit),output(..)
recirc?
output?

Firewall: Allow outbound connections



Matchable ct_state

- **trk** - Tracked - Been through the connection tracker
- **inv** - Invalid
- **new** connection
- **est** - Established connection
- **rpl** - Packet is in reply direction
- **rel** - Related - ICMP, eg “dst_unreach” response / helper “related” connection

Matchable ct_state

- **trk** - Tracked - Been through the connection tracker
- **inv** - Invalid
- **new** connection
- **est** - Established connection
- **rpl** - Reply direction
- **rel** - Related - ICMP response / helper “related” connection
- **src_nat** - Packet’s source address/port was mangled by NAT.
- **dst_nat** - Packet’s destination address/port was mangled by NAT.

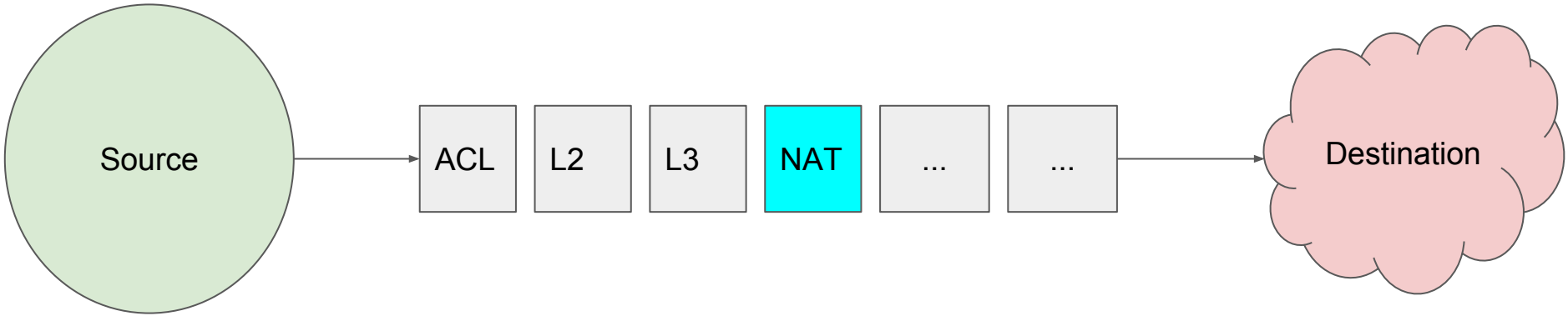
CT action

- **commit** the connection
- **zone** to commit the connection in (u16)
- **mark** to associate with the connection (masked u32)
- **labels** to associate (masked 128 bits)
- **helper** to apply
- **table** to continue processing (userspace-only; maps to recirculation)

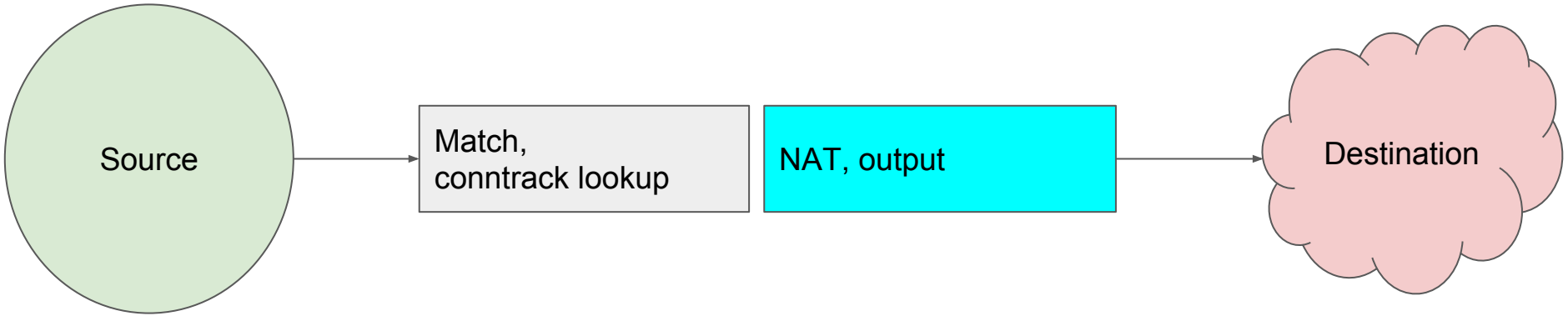
CT action

- **commit** the connection
- **zone** to commit the connection in (u16)
- **mark** to associate with the connection (masked u32)
- **labels** to associate (masked 128 bits)
- **helper** to apply
- **nat** action - Without nested attrs, performs reverse NAT based on CT
  - Exclusive **SRC**/**DST** flags to indicate translation type
  - **IP_MIN**/**IP_MAX** for L3 address range
  - **PROTO_MIN**/**PROTO_MAX** for L4 port range
  - **PERSISTENT** flag - retain consistent IP mapping, even across reboots
  - **PROTO_HASH** - Maps to NF_NAT_RANGE_PROTO_RANDOM (implemented as hash)
  - **PROTO_RANDOM** - fully-random L4 port mapping

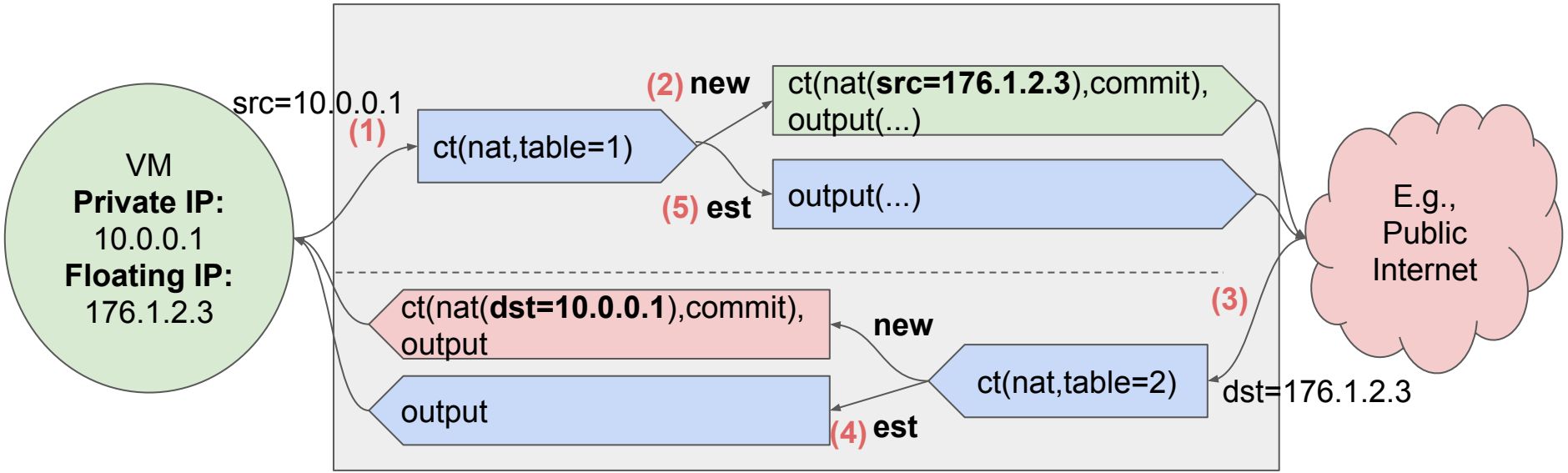
NAT pipeline



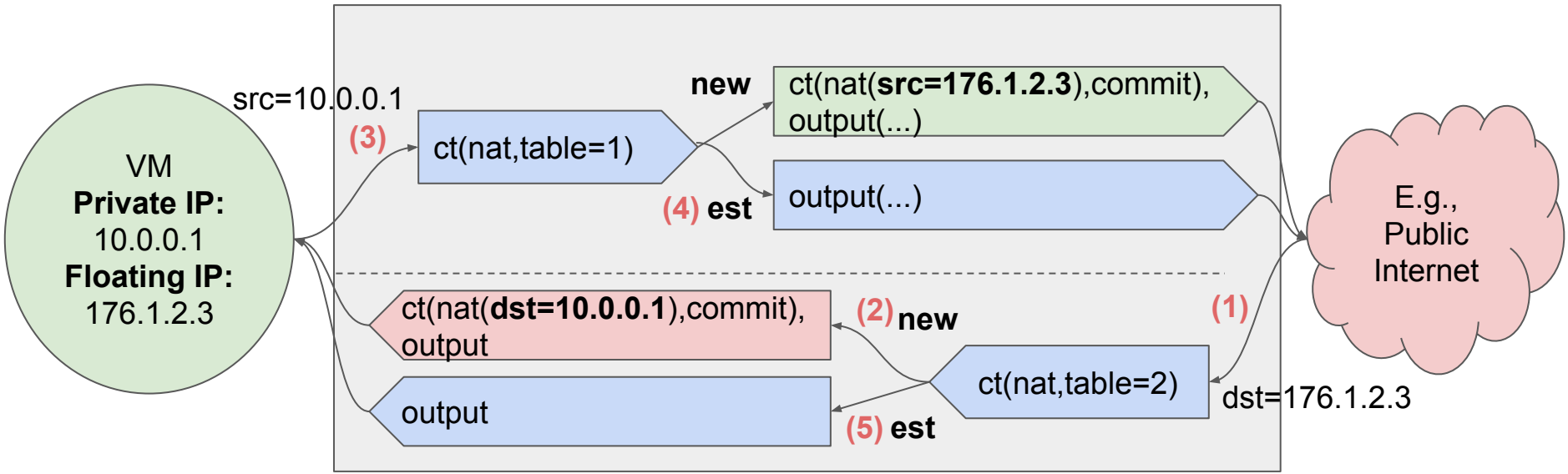
NAT pipeline: kernel



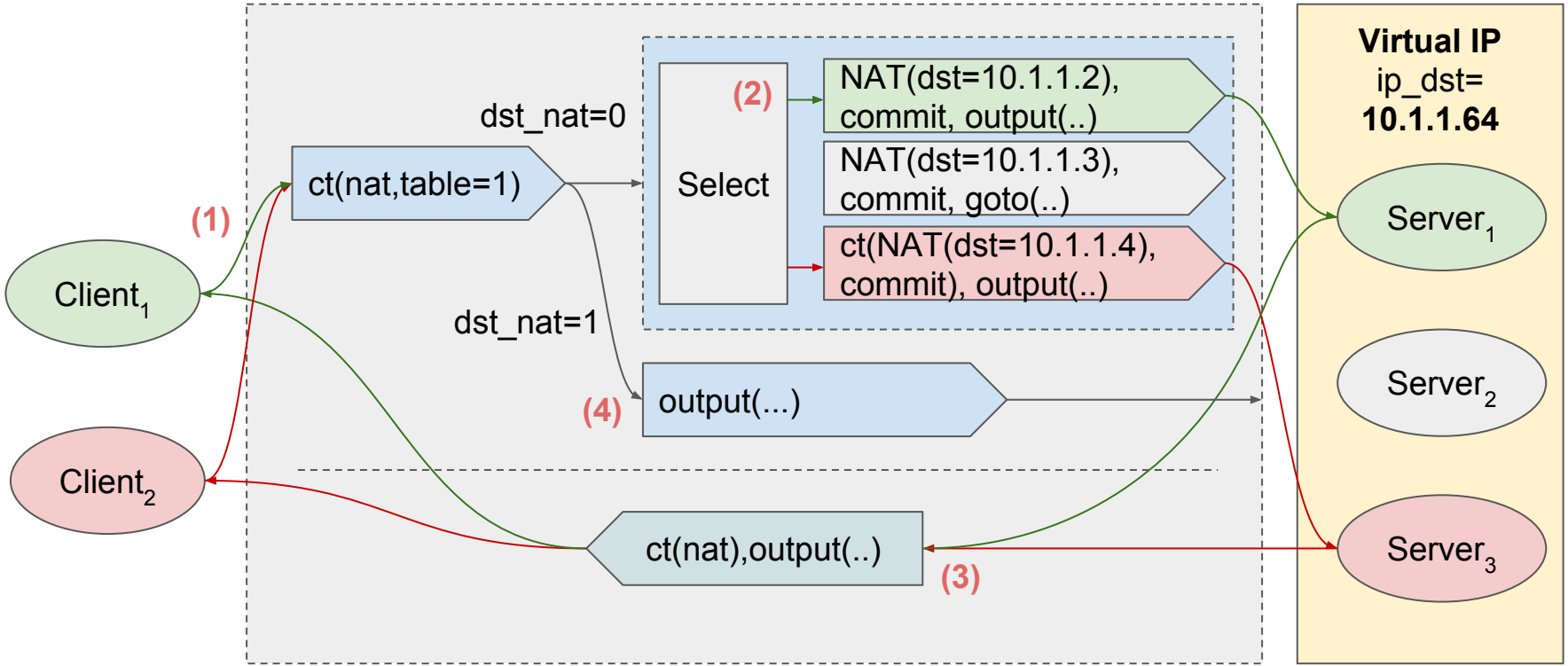
NAT for OpenStack Floating IPs: outgoing



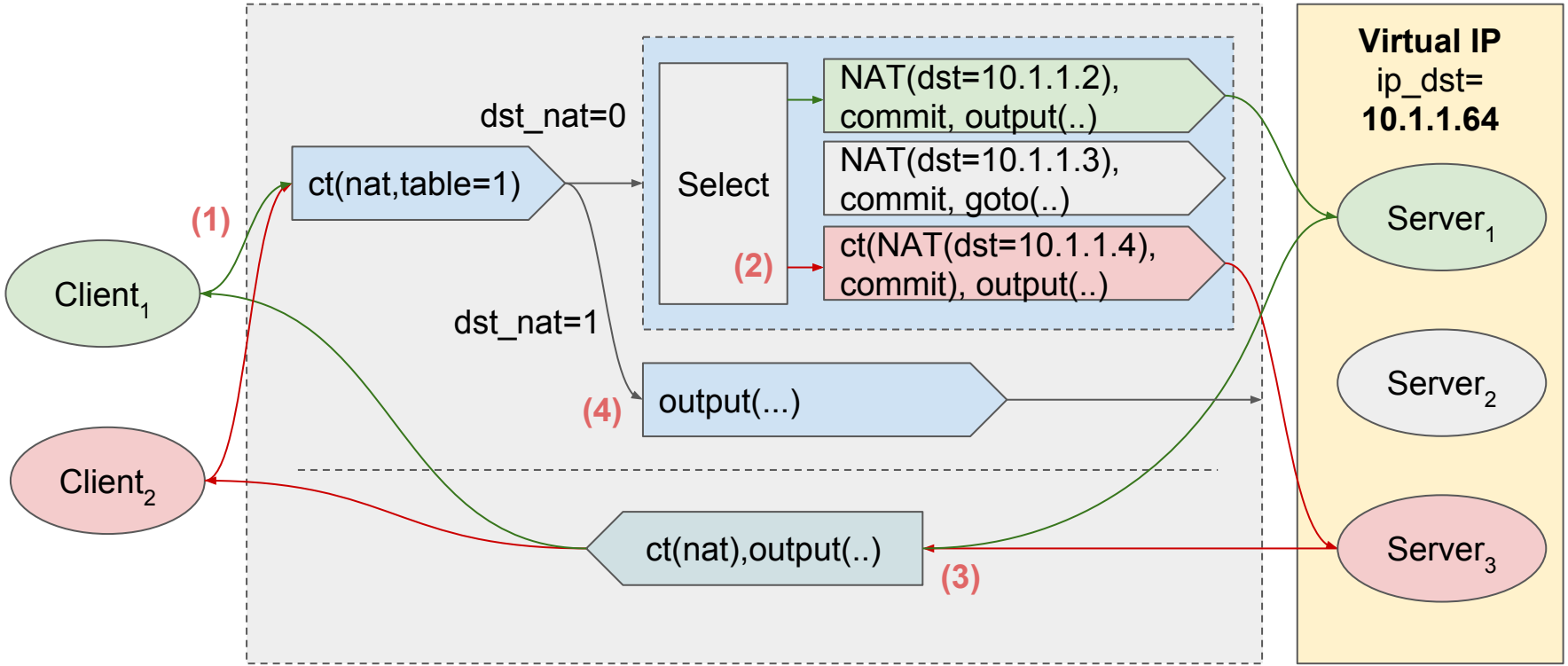
NAT for OpenStack Floating IPs: incoming



DNAT Load Balancing



DNAT Load Balancing



Wishlist: Zones

- Directionality
- Per-zone resource limits
- Dumping
  - All conntrack entries for a zone
  - Filter, eg based on specific labels
- Flush individual zones
- Delete events are hardcoded not to provide mark/labels
  - Selectively log terminated connections when CT is deleted
- Some sort of tied-fate deletion of master -> delete children

Thanks for listening