

Netkit

Specializing Linux Packet Delivery for Container Networks

Daniel Borkmann, Paul Chaignon · eBPF'26

개별연구 7주차 논문 리뷰

26.10.15 박수현

목차

1

문제 상황

2

설계

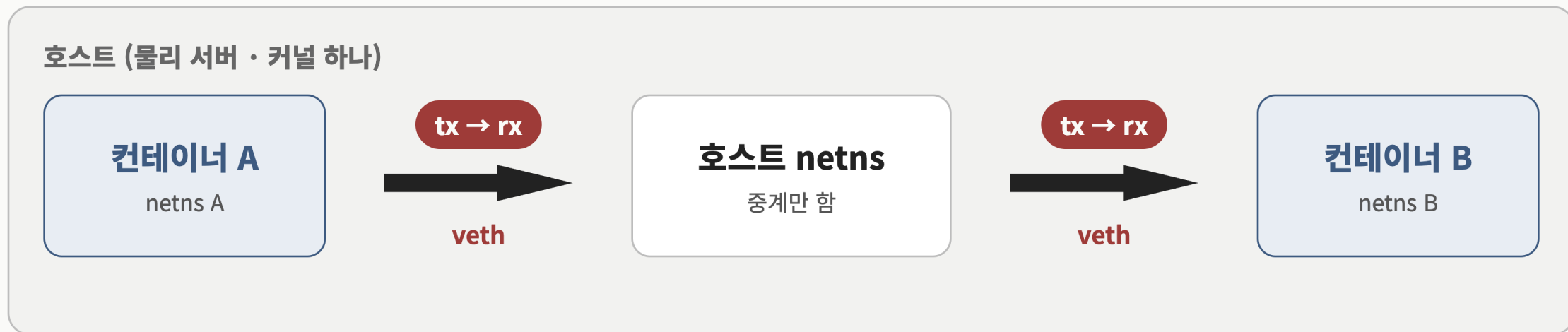
3

평가

4

실습

1. 문제 상황



송수신 전환이 두 번

1. 문제 상황: veth



veth 는 nic 역할, 쌍으로 존재
backlog 는 하드웨어 큐 역할, 송 수신 스택의 분리

1. 문제 상황: backlog 중첩

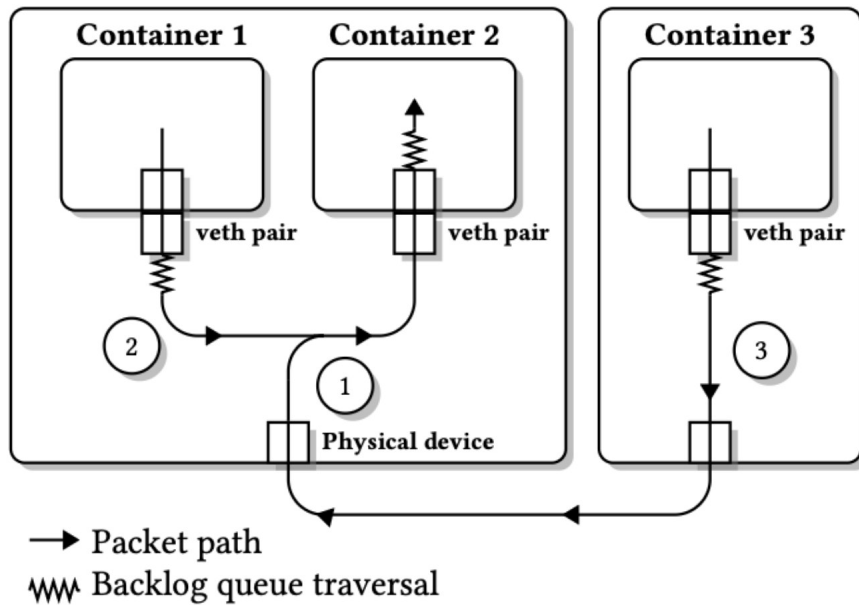


의미 없는 전환: 호스트 rx → tx

호스트는 받아서 바로 B로 다시 보내는 중계.
의미없는 backlog가 중복된다

1. 문제 상황: backlog 중첩

컨테이너 간 패킷 경로 (Fig. 2)



같은 머신 · 경로 ②

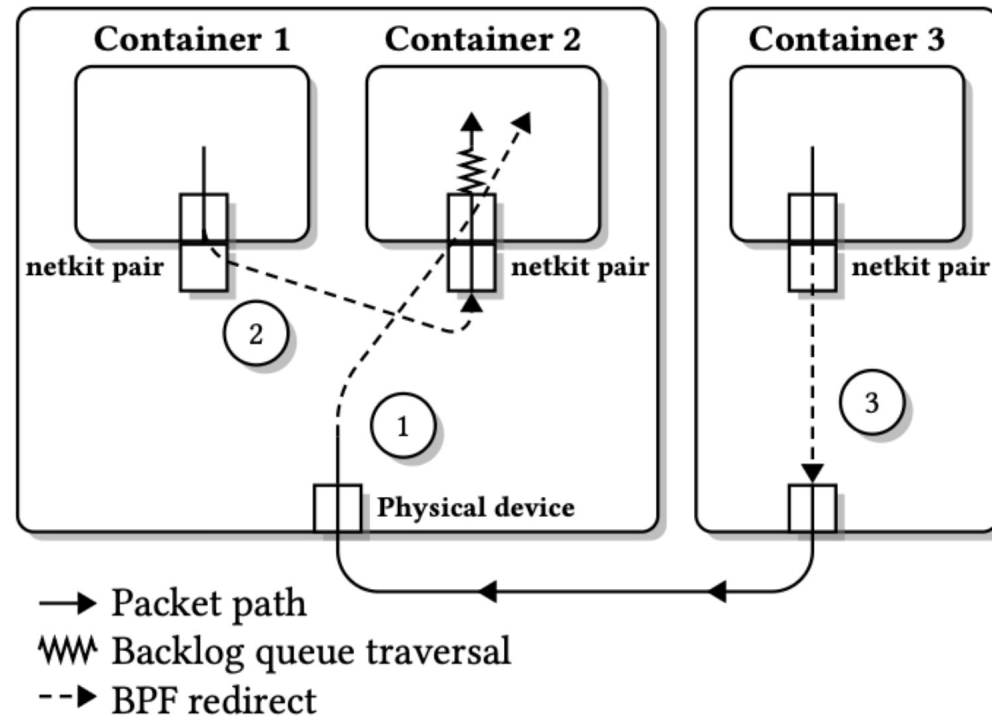
한 머신 안에서 backlog 2번
진짜 전환은 1번 → 1번 불필요

다른 머신 · 경로 ③ + ①

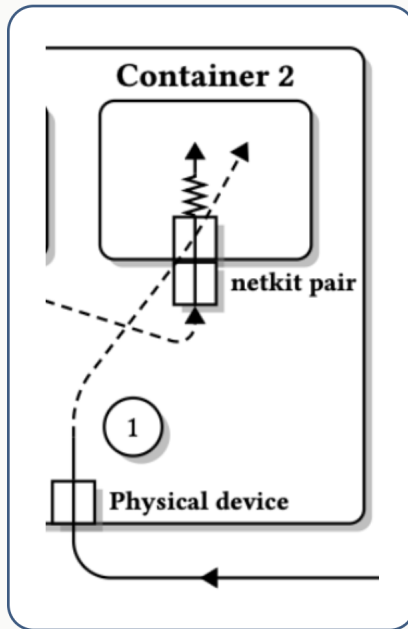
진짜 전환은 NIC 경계에서 일어남 → 2번 모두 불필요

2. 설계

적용 후 경로 (Fig. 4)



2.1 들어오는 쪽: bpf_redirect_peer()



기본

NIC 수신



호스트 veth 송신



backlog



컨테이너

bpf 헬퍼 함수

호스트 netns

NIC 수신 → tc-bpf



bpf_redirect_peer()

skb→dev 교체
= 네트워크 네임스페이스를 교체

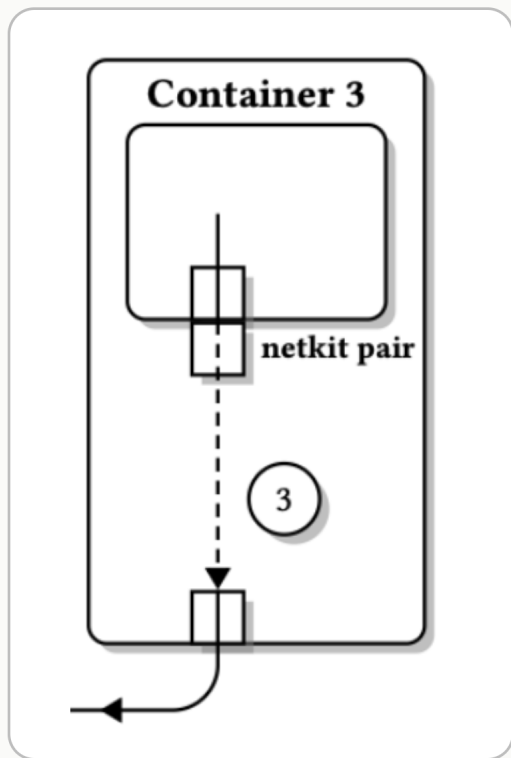


컨테이너 netns

IP → TCP → 소켓

dev만 바꿔 수신 함수를 한 번 더 실행

2.2 나가는 쪽: netkit 디바이스



문제점

1. 컨테이너 내부에서 외부의 nic 을 조회할 수 없다.
2. egress 에서 네임스페이스를 전환하는 bpf함수를 지원하지 않는다.

primary 호스트 쪽 디바이스 · **peer** 컨테이너 쪽 디바이스 -- pair로 구성되어있다.

netkit

send → TCP → IP



netkit_xmit: skb→dev = primary



BPF redirect: skb→dev = NIC

netkit: veth와의 차이

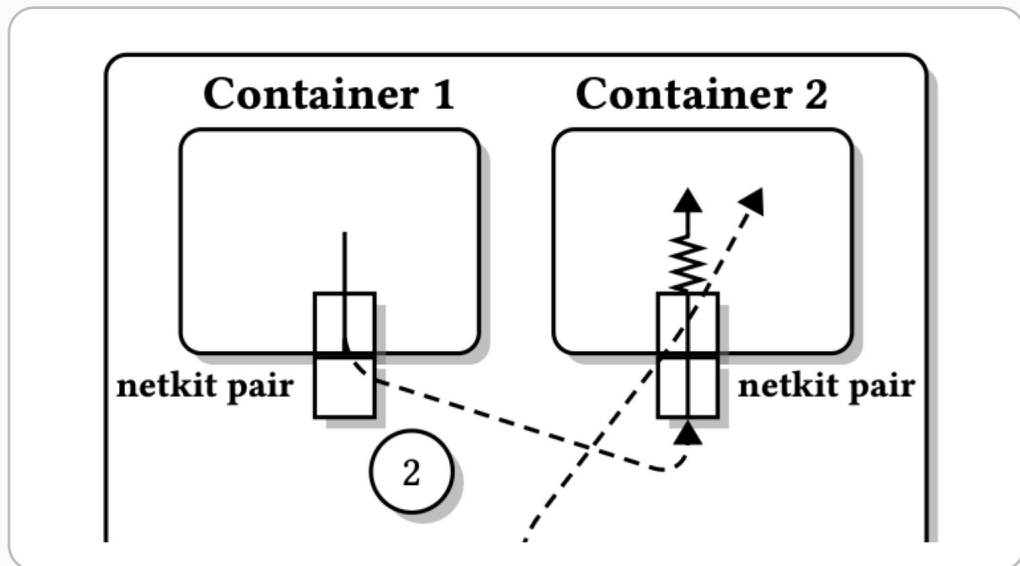
① 송신 경로에서 netns 전환

② 비대칭 구조로 호스트가 관리 - 격리 강화

veth: 양 끝 대칭 → 컨테이너가 자기 쪽 BPF를 붙이거나 땜

netkit: BPF 등록은 primary로만 → 컨테이너 netns에서 접근 불가

2.3 같은 머신: 컨테이너 → 컨테이너



veth · backlog 2번

1 송신 → backlog → 호스트 → backlog → 2 수신

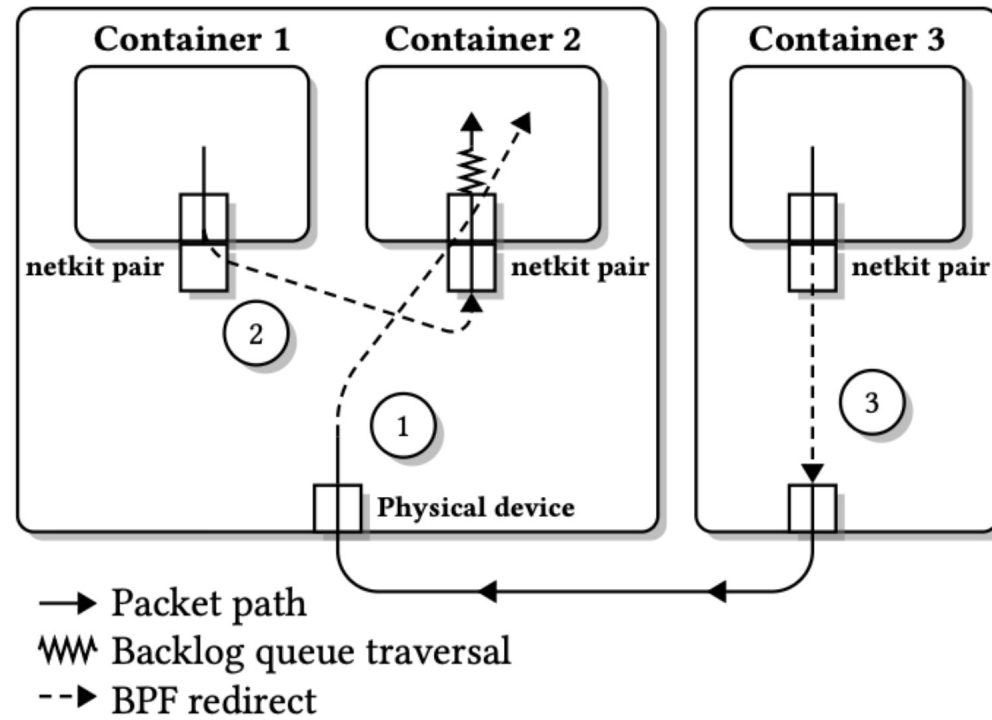
netkit · backlog 1번

1 송신 → 1 송신 netkit BPF → 2 송신 → backlog → 2 수신

2 송신 → 2 수신, 진짜 전환은 없앨 수 없음

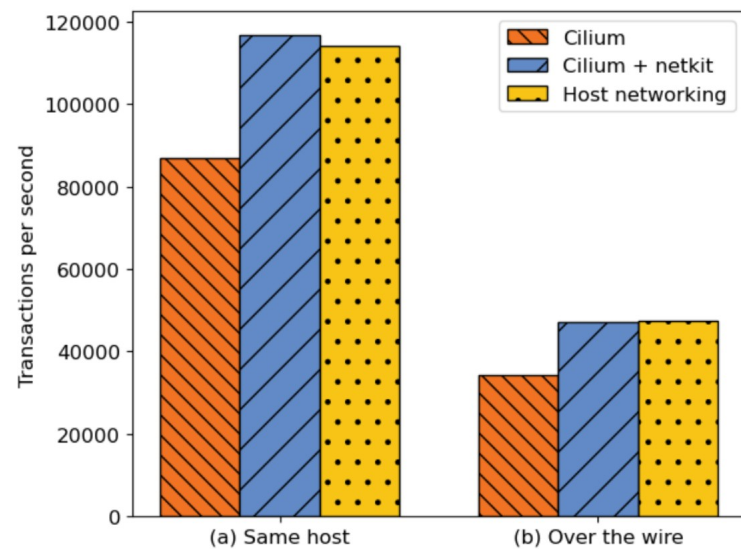
2. 설계

적용 후 경로 (Fig. 4)

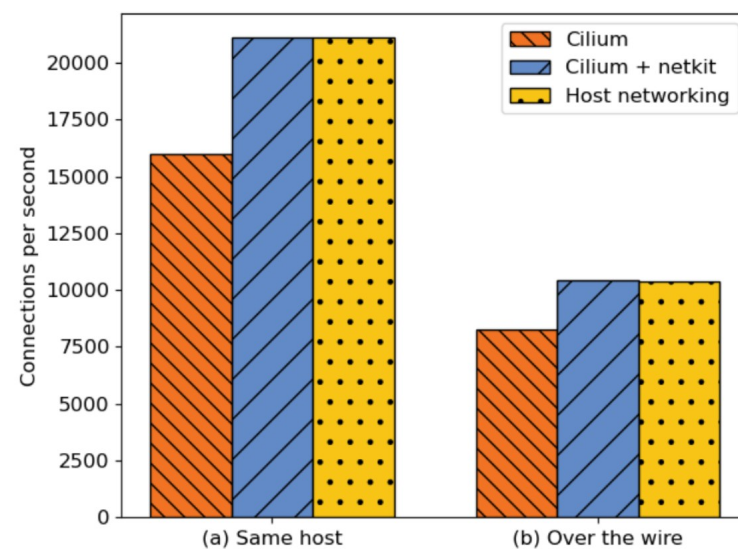


3. 평가: 처리량

TCP_RR · 초당 트랜잭션 (Fig. 5)



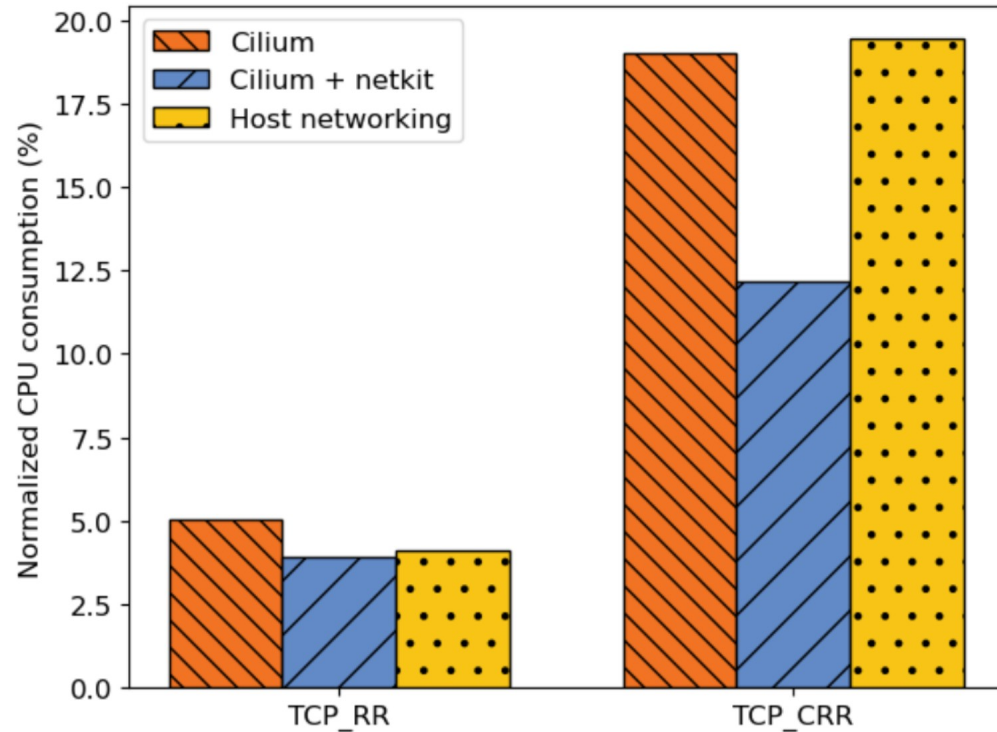
TCP_CRR · 초당 연결 수 (Fig. 6)



veth 대비 크게 향상, Host와는 오차 범위 내 동률

3. 평가: CPU 사용량

수신 측 정규화 CPU (Fig. 7, 다른 머신)



conntrack: 연결마다 상태를 기록하는 테이블(netfilter)

연결 추적 · 연결 생성·삭제 비용이 큼

	Cilium (veth)	Cilium + netkit	Host
커널 conntrack 우회	O	O	X
backlog 제거	X	O	O

RR: conntrack 영향이 작음

연결 1번 → conntrack 작업 적음

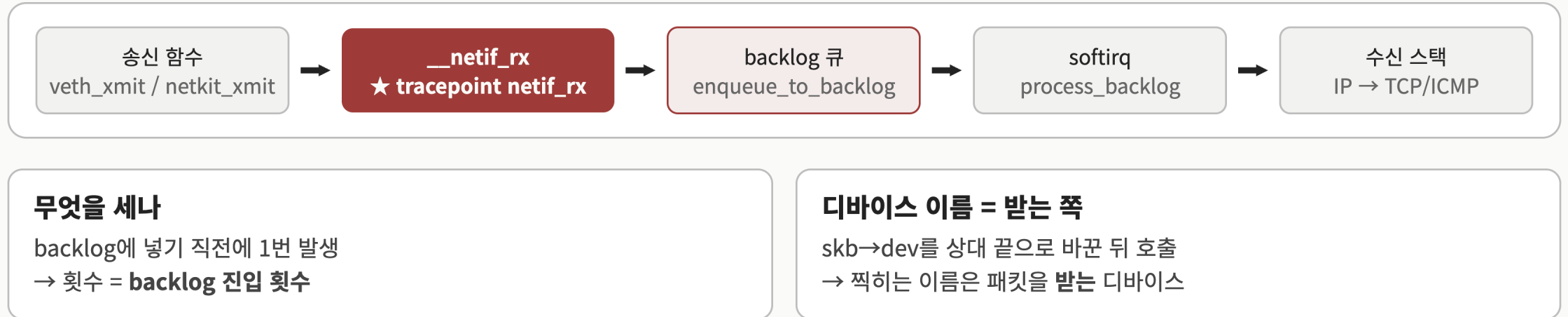
CRR: conntrack 영향이 큼

연결 반복 → conntrack 비용 커짐

4. 실측: 실험 구성



4. 실측: 관측 지점 netif_rx



4. 실측: backlog 진입 횟수

① BPF 없음

```
psh@psh-MS-7E28:~/netkit-lab$ sudo bpftrace -e 'tracepoint:net:netif_rx { @[str(args->name)] = count(); }'  
Attached 1 probe  
^C  
  
@[enp6s0]: 1  
@[nkb0]: 10  
@[nka0]: 10  
@[nkb1]: 10  
@[nka1]: 10  
@[lo]: 544
```

nka0 10 · nkb0 10 · nka1 10 · nkb1 10 → 총 40

② peer 슬롯 redirect BPF

```
psh@psh-MS-7E28:~/netkit-lab$ sudo bpftrace -e 'tracepoint:net:netif_rx { @[str(args->name)] = count(); }'  
[sudo: authenticate] 비밀번호 :  
Attached 1 probe  
^C  
  
@[nkb1]: 10  
@[nka1]: 10
```

nka1 10 · nkb1 10 → 총 20

