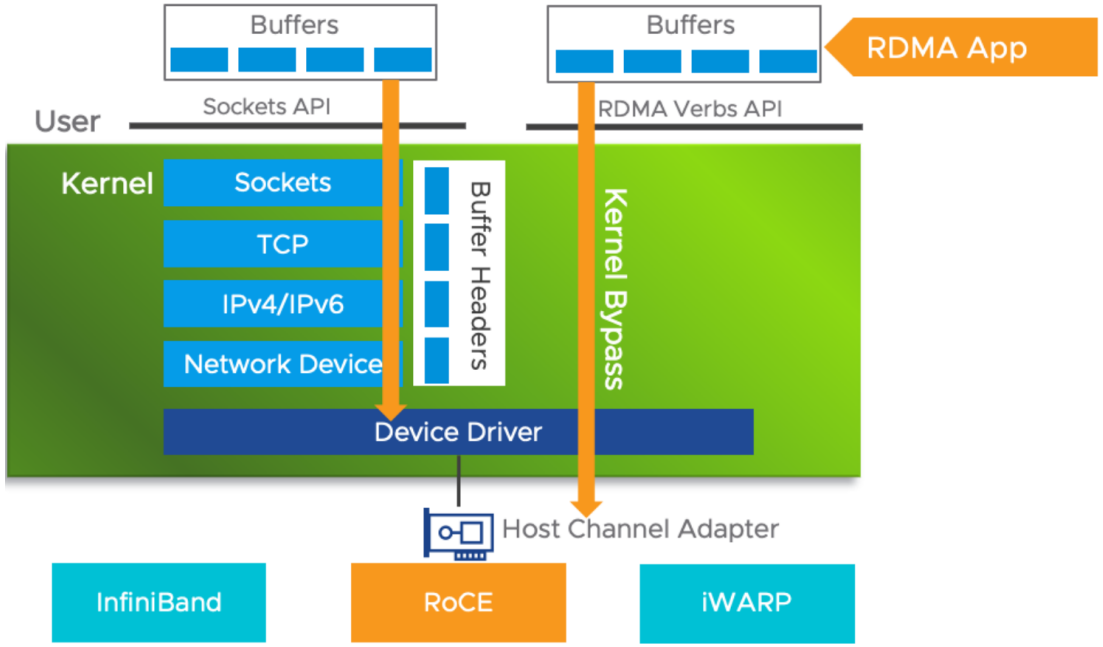
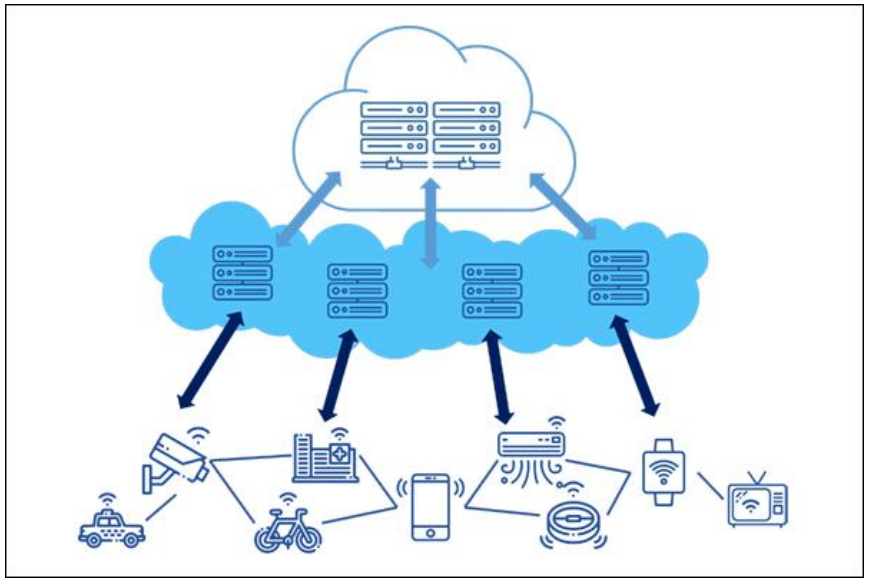
컨테이너 환경에서의 RDMA NIC 마이크로아키텍처 자원 고갈 영향 분석

김건우[1], 김진우[2,†], 박병준[2]
[1,2]광운대학교 (대학원생, 교수)

RDMA

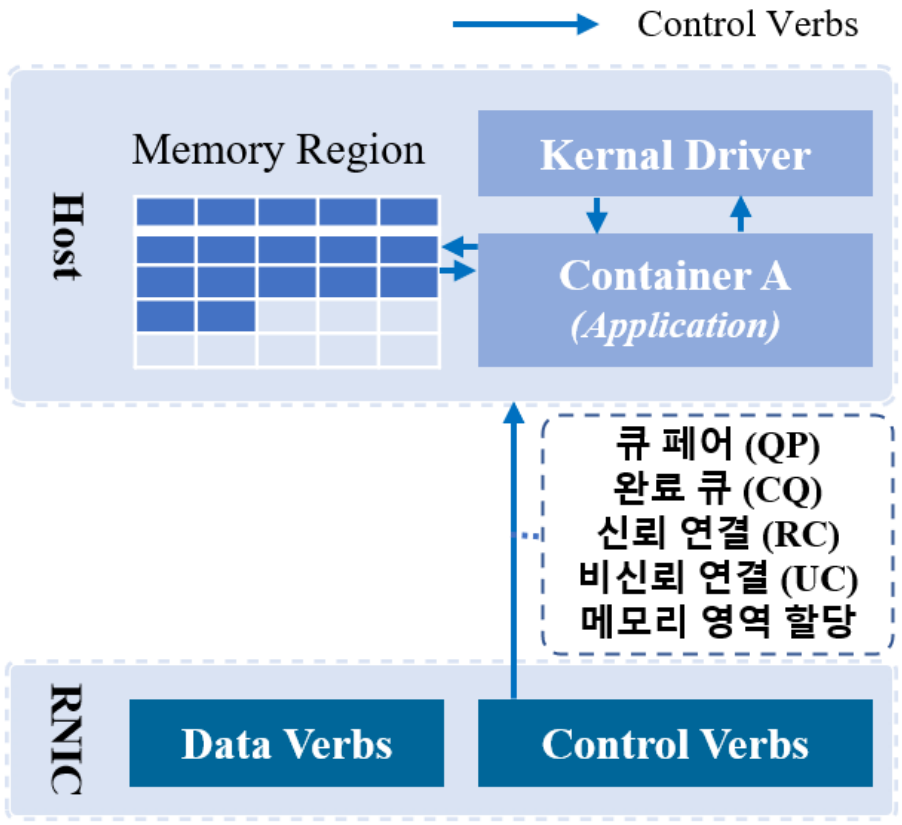
- **Remote Direct Memory Access (RDMA)**
  - Network Interface Card (NIC)을 통해 호스트 메모리에 커널을 우회하여 직접 접근하는 기술
  - Host CPU 적은 소모, 빠른 데이터 교환 등등… 클라우드 환경에 널리 도입되고 있다





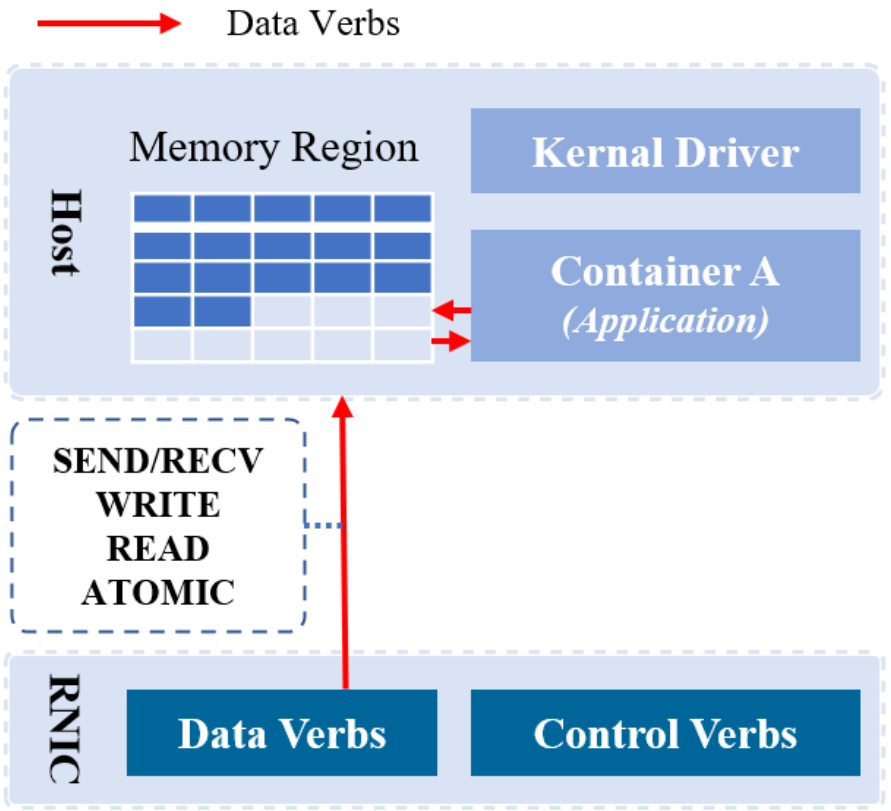
https://www.fibermall.com/ko/blog/rdma-key-technology-for-arithmetic-networks.htm

RDMA



- **Verbs**
  - Control Verbs
  - Data Verbs
- **초기화 과정**
  - QP, CQ와 같은 필요한 객체 생성
  - 신뢰, 비신뢰 연결 설정
  - 호스트의 DRAM 영역을 할당하고 가상 주소에서 물리적 주소로의 매핑 수행
- **Mellanox API**
  - ibv_ceate_qp, ibv_alloc_pd, ETC.
  - rdma_create, rdma_bind, ETC.

RDMA



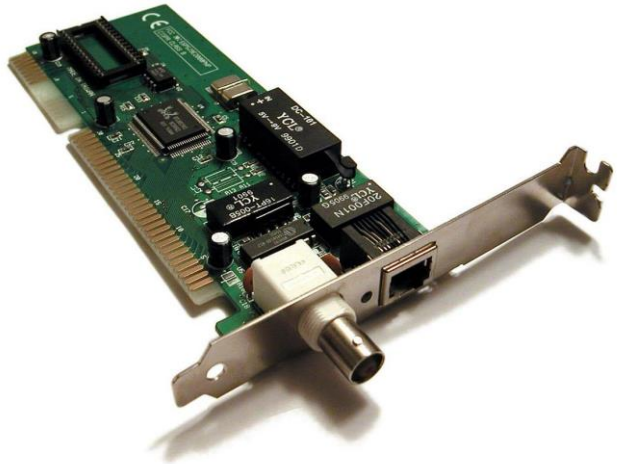
- **RDMA 작업 수행**
  - SEND/RECV
  - WRITE/READ
  - ATOMIC
- **Mellanox API**
  - ibv_post_send, ibv_post_recv, ibv_get, ETC.
  - rdma_read, rdma_write, ETC.

SmartNIC

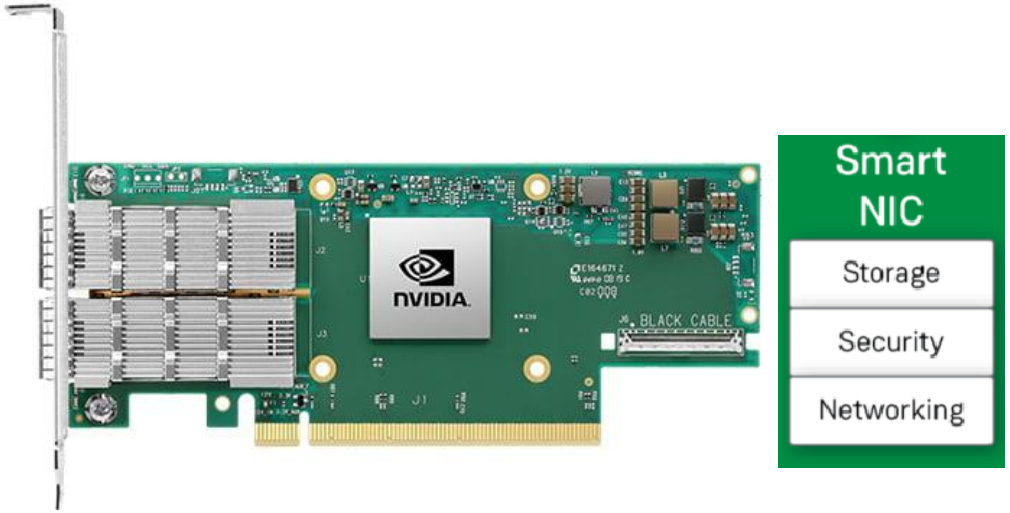
- **Network Interface Card (NIC)**
  – 컴퓨터를 네트워크에 연결하여 통신하기 위해 사용하는 전통적인 하드웨어 장치

- **Smart NIC (RDMA NIC, RNIC)**
  – 다양한 네트워크 관련 작업을 하드웨어로 오프로드하고 가속화하여 처리하는 장치
  – RoCEv2(RDMA over Converged Ethernet version 2)

Traditional NIC

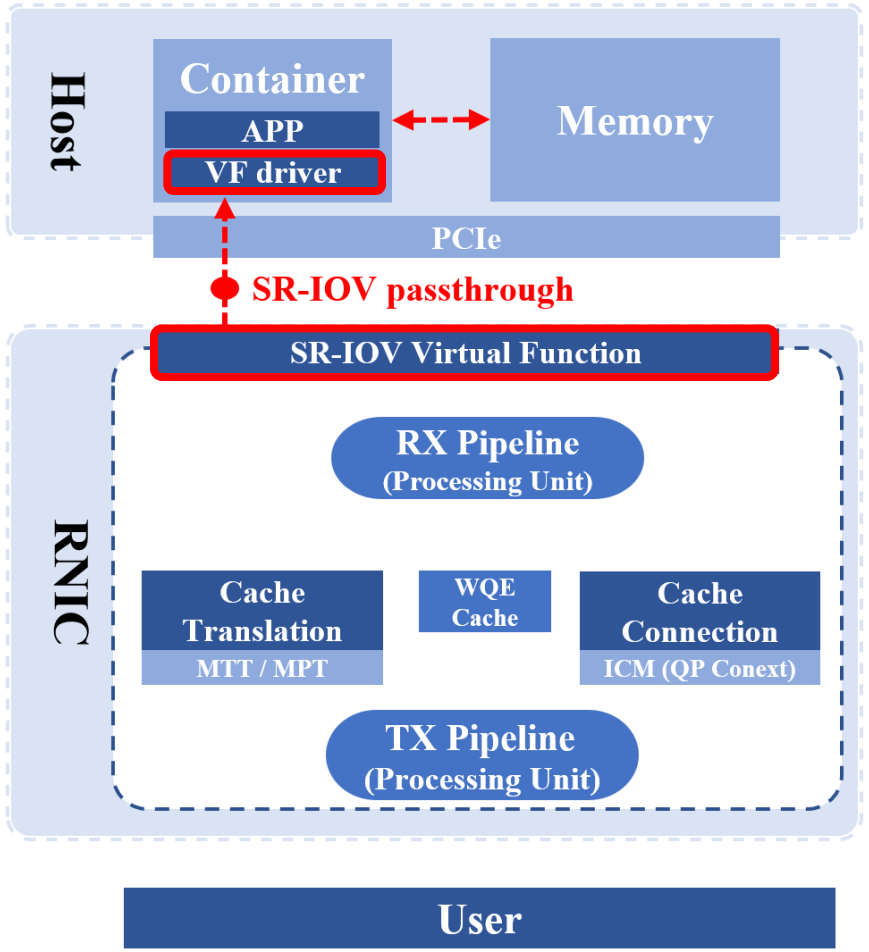


NIC



Smart NIC

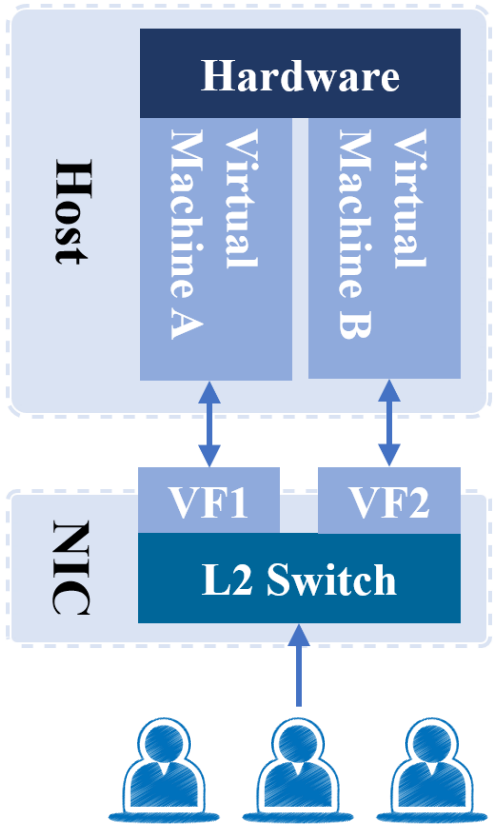
마이크로아키텍처 리소스



- RDMA NIC 내부에서 동작하는 하드웨어 자원
- 데이터 전송, 메모리 관리, 프로토콜 처리 등과 같은 작업을 하드웨어 수준에서 최적화.

- TX/RX Pipeline
  - 패킷 처리(패킷 분류, 스케줄링, 헤더 처리 등)를 담당
- MTT/MPT
  - 메모리 관리
- Workload Acceleration Engine
  - 데이터 작업을 가속화하기 위한 하드웨어 엔진

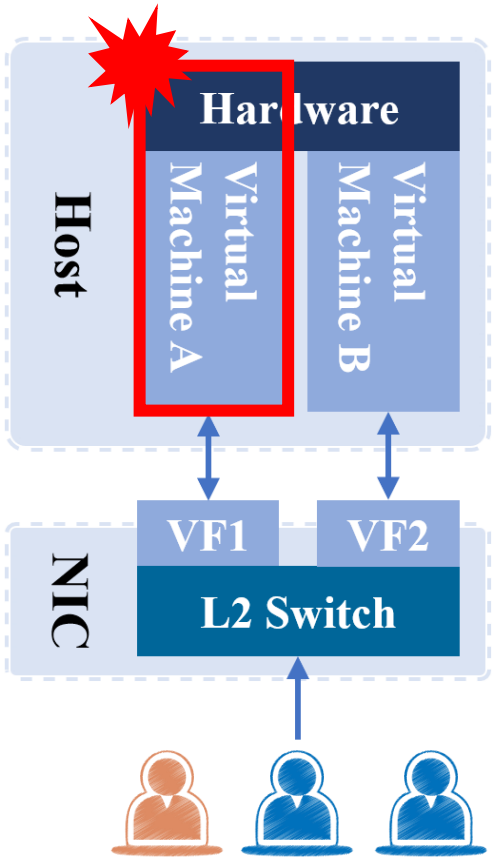
성능 격리 문제 [1]

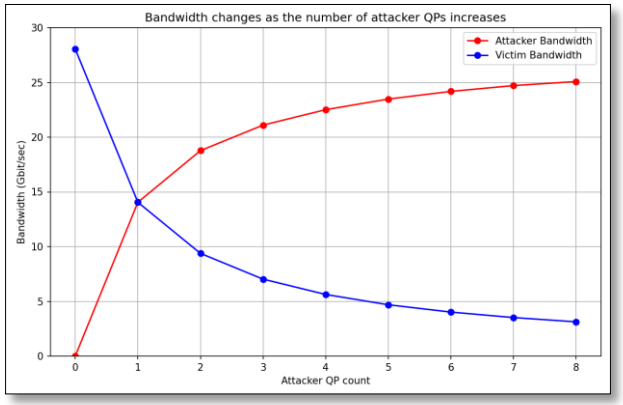


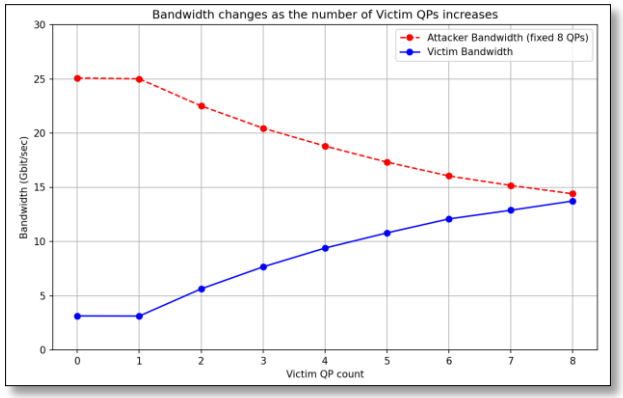
- **Multi-tenant**
  - 여러 테넌트가 동일한 하드웨어 자원을 공유하여 독립적으로 운영되는 환경

[1] Grant, Stewart, et al. "SmartNIC Performance Isolation with FairNIC: Programmable Networking for the Cloud", ACM, July, 2020

성능 격리 문제 [1]







경쟁에 따른 Bandwidth 변화

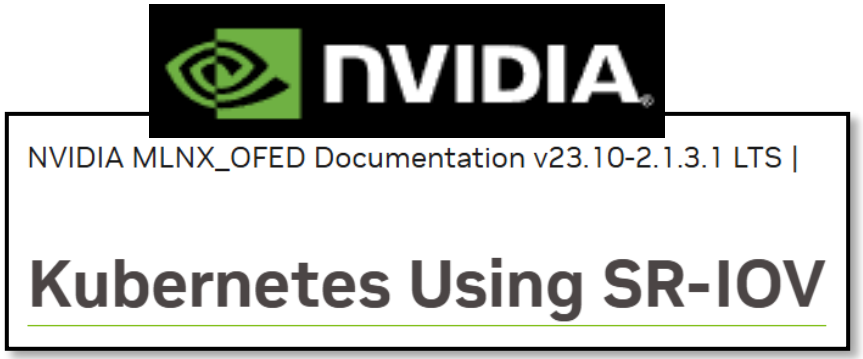
- **Multi-tenant**
  - 여러 테넌트가 동일한 하드웨어 자원을 공유하여 독립적으로 운영되는 환경
- **자원 격리 문제**
  - 악의적인 의도를 가지거나 의도치 않은 행동을 한 테넌트가 있을 때
  - 자원을 고갈시키거나 경쟁하여 다른 테넌트의 성능을 저하시킨다.

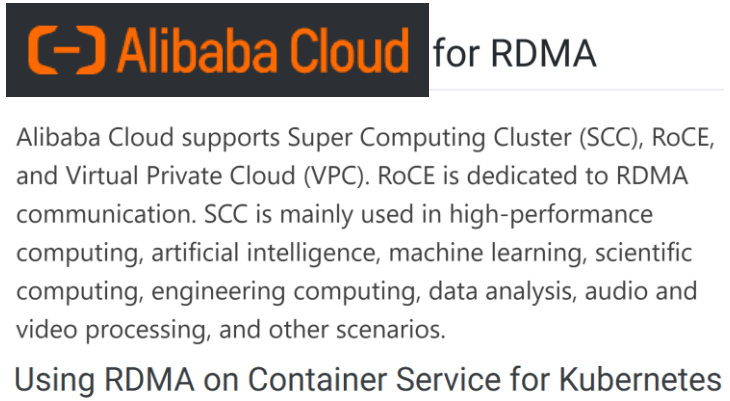
[1] Grant, Stewart, et al. "SmartNIC Performance Isolation with FairNIC: Programmable Networking for the Cloud", ACM, July, 2020

관련 연구 및 동향

- **RDMA의 성능 격리 문제는 여러 이전 연구에서 다루어졌다 [1].**
  – 이들은 대부분 가상머신 또는 베어메탈 환경에서 분석되었다는 한계점이 있다.

- **최근 클라우드의 동향**
  – 컨테이너를 사용한 클라우드 네이티브 아키텍처를 지향하기 때문에 RDMA 또한 컨테이너에 맞게 도입하려는 움직임을 보이고 있다.

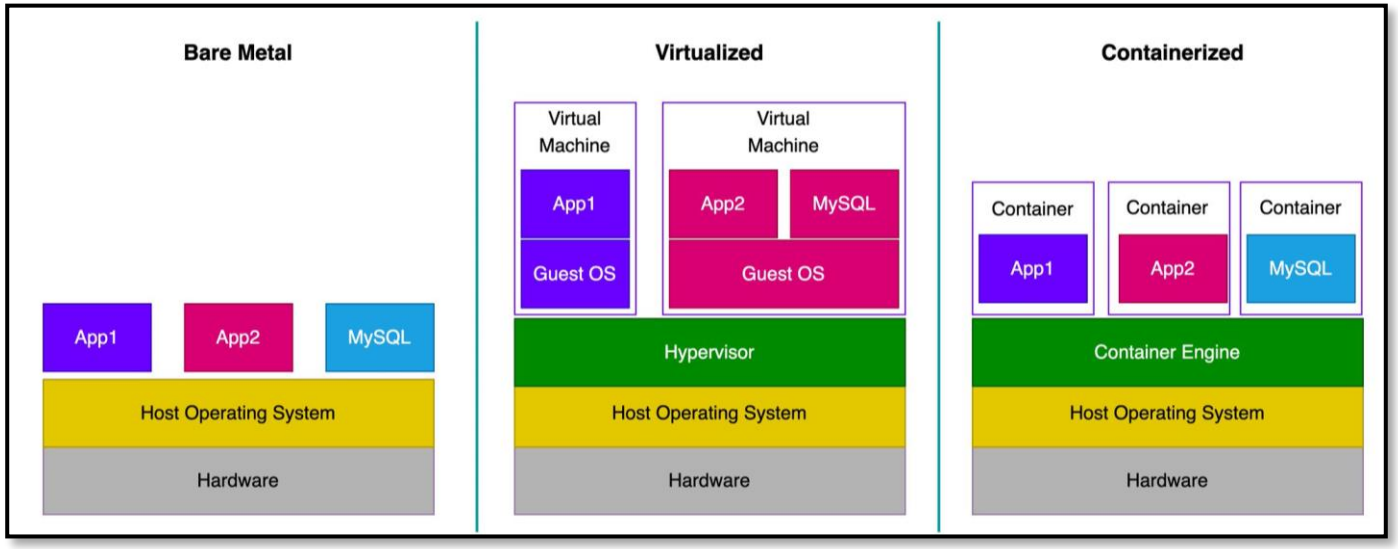




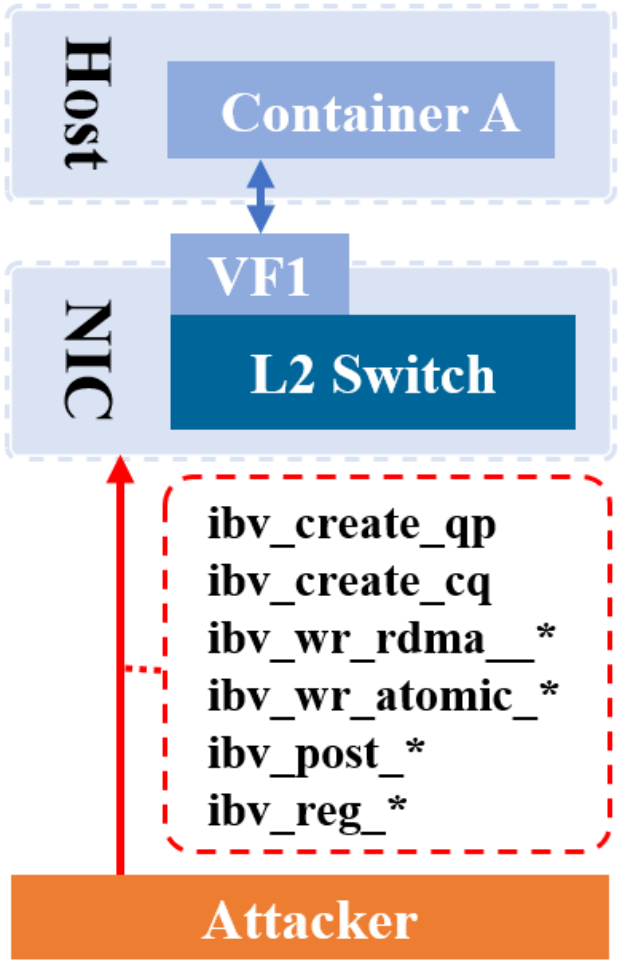
[1] Grant, Stewart, et al. "SmartNIC Performance Isolation with FairNIC: Programmable Networking for the Cloud", ACM, July, 2020

연구 목적

- 컨테이너 환경에서 RNIC의 마이크로아키텍처 자원이 고갈되었을 때의 영향을 분석한다.



위협 모델



- **위협 모델**
  - 공격자는 특정 Control/Data Verbs 남용할 수 있다고 가정한다.
    - 과도한 큐 페어(QP) 및 완료 큐(CQ) 생성.
    - 새로운 메모리 위치에 접근하여 연산 수행 가능.
  - 의도적으로 예외 상황을 유발시킴으로써 RDMA서비스 불능으로 만들 수 있음을 가정한다.

공격 시나리오

- **공격 시나리오**

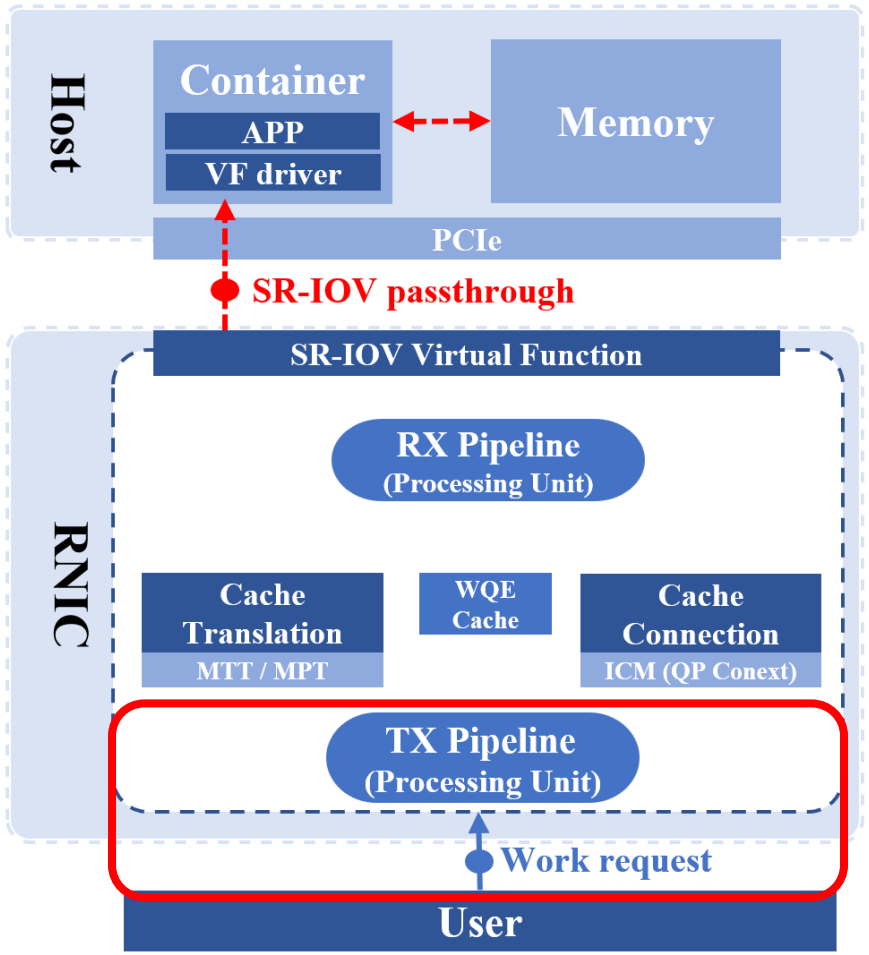
  1) 큐 플러딩 공격

  다수의 큐 페어(QP) 또는 완료 큐(CQ)를 통해 RNIC의 TX/RX PU의 처리 능력 초과.

  2) 캐시 소진 공격

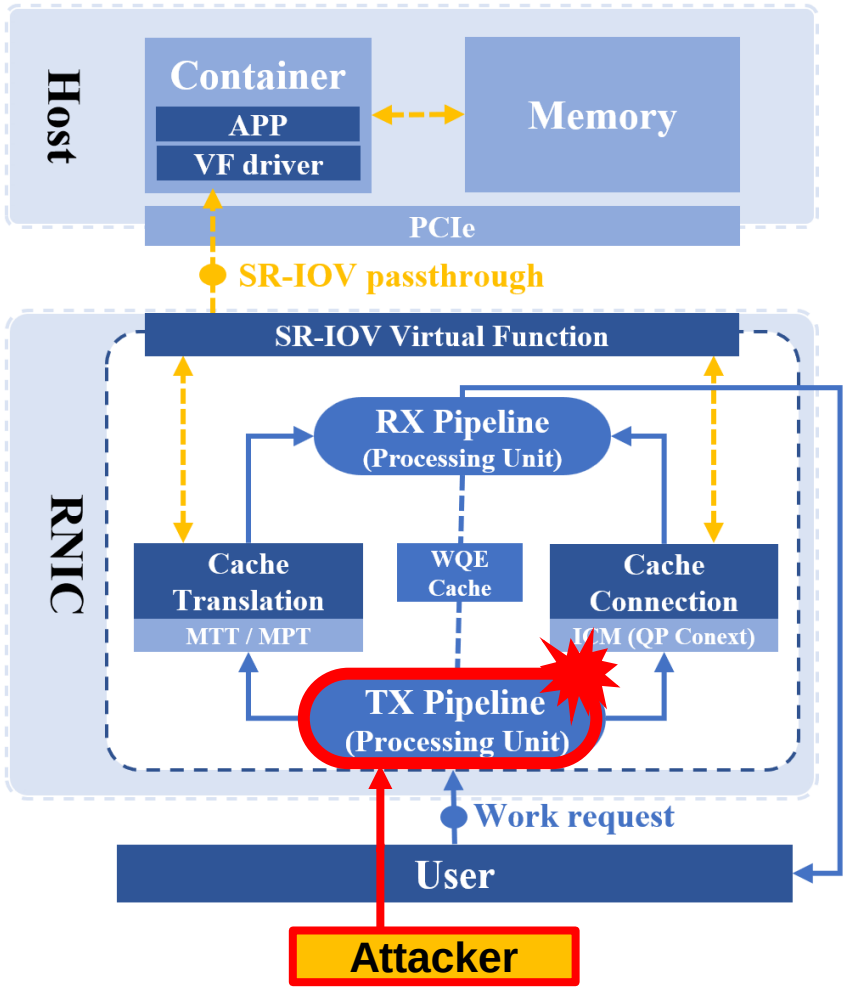
  새로운 메모리 위치에 접근하는 RDMA 연산 수행을 통해 캐시 처리 능력 초과.

공격 시나리오 : 1) 큐 플러딩 공격



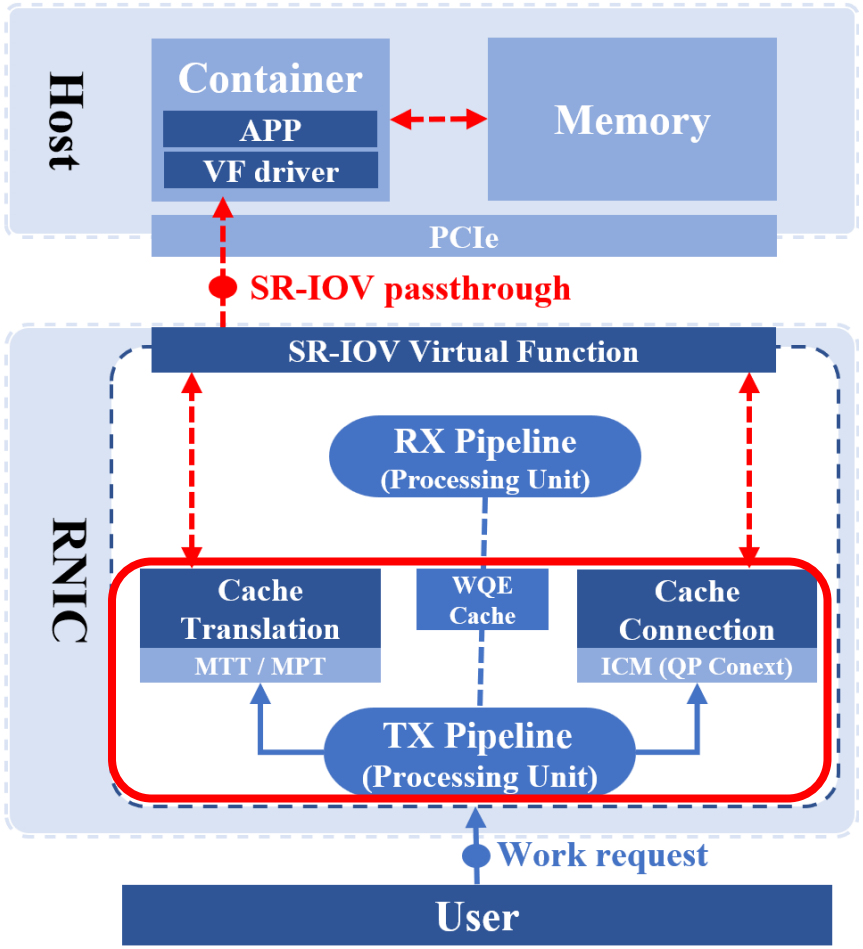
- **정상 작업 요청**
  - TX Pipeline, QP.

공격 시나리오 : 1) 큐 플러딩 공격



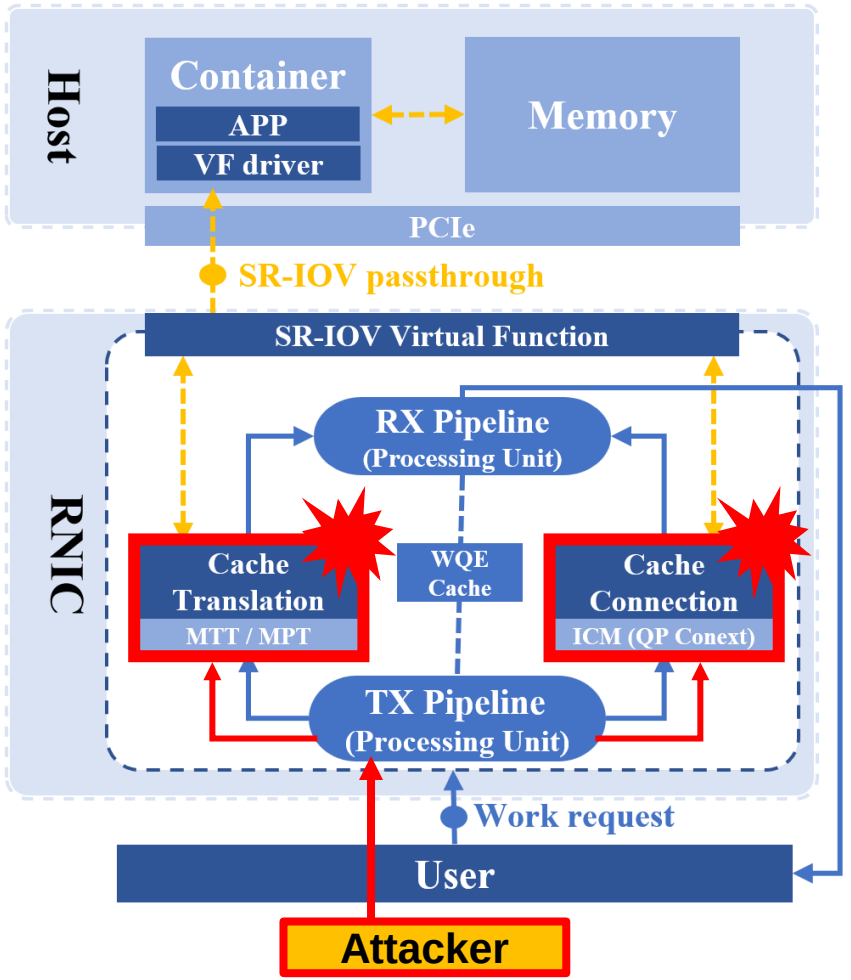
- **정상 작업 요청**
  - TX Pipeline, QP.
- **공격 방식**
  - 다수의 큐 페어(QP) 또는 완료 큐(CQ)를 통해 지속적인 데이터 송수신.
- **공격 영향**
  - TX/RX PU의 처리 능력을 초과시킨다.
  - 정상적인 데이터 전송 작업의 지연을 유발시켜 전체 대역폭을 감소시킨다.

공격 시나리오 : 2) 캐시 소진 공격



- **정상 작업 요청**
  - Cache Translation, Cache Connection.
  - 메모리 접근 및 관리.

공격 시나리오 : 2) 캐시 소진 공격



- **정상 작업 요청**
  - Cache Translation, Cache Connection.
  - 메모리 접근 및 관리.
- **공격 방식**
  - 지속적으로 새로운 메모리 위치에 접근하는 RDMA 연산 수행.
- **공격 영향**
  - 캐시 히트율을 낮추고 Cache Translation를 처리하는 캐시를 고갈 시킨다.
  - 이를 통해 Latency와 Cache Miss가 증가하게 된다.

실험 환경

- **성능 격리 시뮬레이션 환경**
  - Bluefield-3 RNIC
    - RoCEv2 프로토콜 기반 RDMA 통신
  - SR-IOV
    - VF를 통해, 독립된 큐 페어, 완료큐 사용
  - Docker
    - Container

- **성능 벤치마킹 도구**
  - Perftest (ib_write_bw, ib_read_lat 등)
  - Perf

perftest Public
Infiniband Verbs Performance Tests
C 618 292

GitHub

공격 영향 분석 : 1) 큐 플러딩 공격

- 공격자가 5초마다 QP 및 CQ 자원을 과도하게 사용함.
- Victim의 대역폭이 26.61 Gbit/sec 에서 1.64 Gbit/sec으로 약 93.9% 급격히 감소.
- 이는 TX/RX 처리 유닛이 점진적으로 과부하되며 대역폭이 크게 저하됨을 의미.

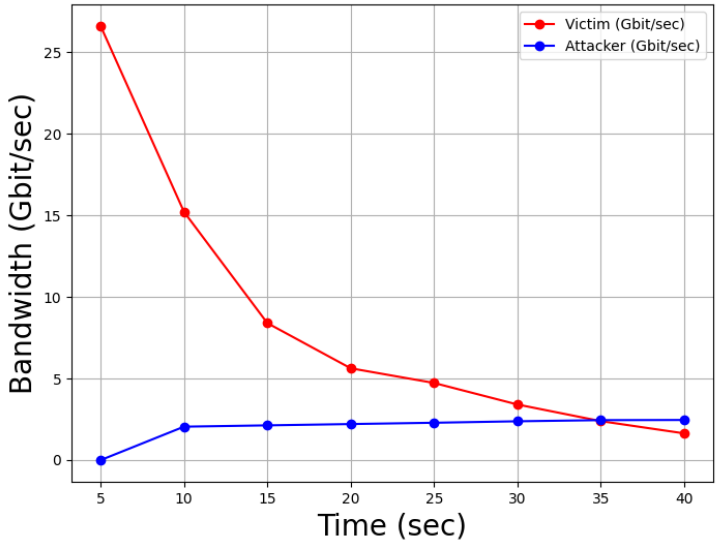


Fig. 3 Bandwidth changes due to TX/RX processing unit overload

공격 영향 분석 : 2) 캐시 소진 공격

- 공격자가 연산에서 새로운 메모리 위치에 지속적으로 접근하여 캐시 고갈 유도.
- 초기 지연 시간 1.56μs 에서 1,746.34μs로 약 1,117배 급격히 증가.
- 캐시 미스 비율 14.48%에서 31.07%로 약 115% 증가하였다.

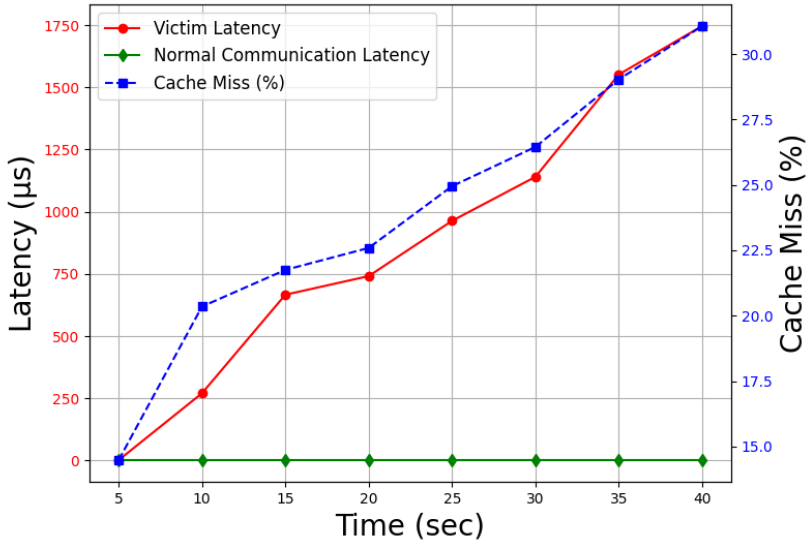
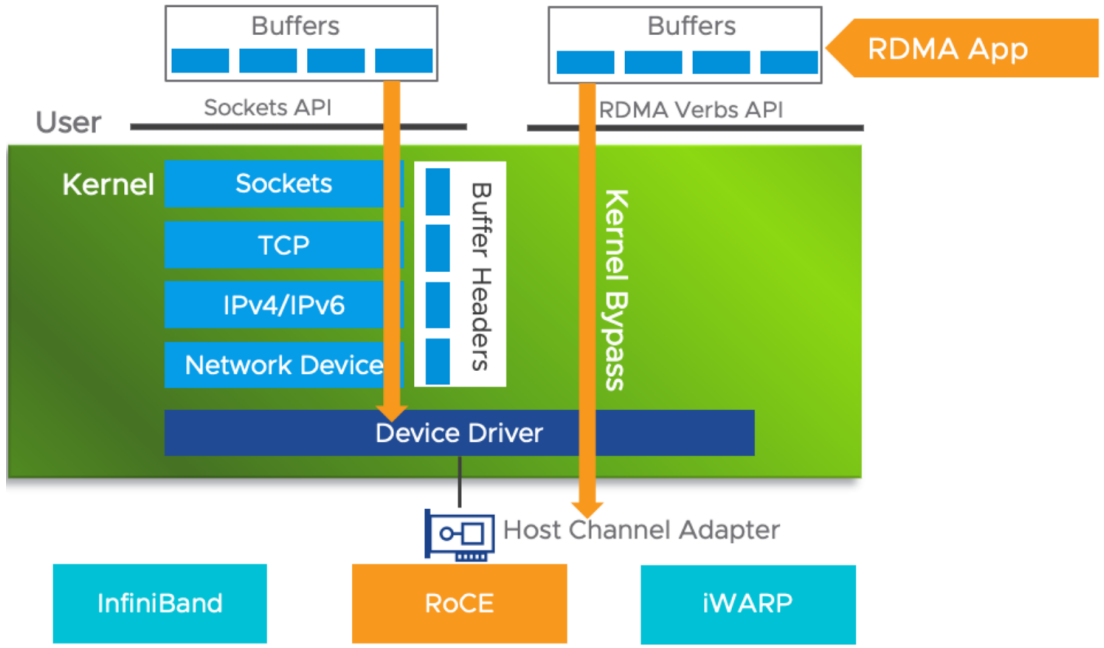


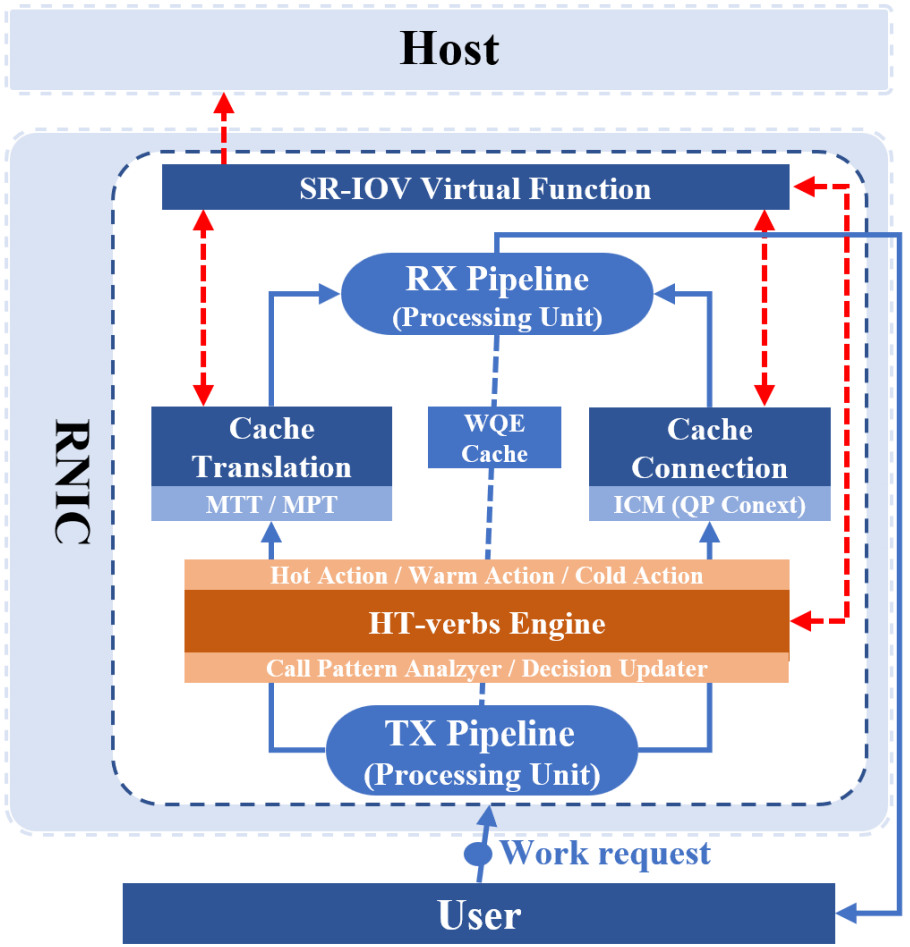
Fig. 4 Latency and cache miss change rates due to resource depletion

공격 탐지의 어려움

- **기존 일반적인 모니터링 방식으로는 공격을 탐지하는데 한계가 있다.**
  - 호스트의 CPU, 메모리를 우회하여 NIC 자원에 접근하기 때문.



HT-verbs: Threshold 기반 자원 관리



- **HT-verbs**
  - 자원 고갈 공격을 완화하기 위해 Threshold 기반 자원 관리 방식을 제시.
  - RDMA Verbs를 Hot, Warm, Cold로 분류 .
- **흐름**
  - 분류 전, RDMA Verbs의 호출 기록을 주기적으로 집계한다.
  - 특정 캐시 접근과 소모율에 대한 패턴을 분석한다.
  - 과거/현재 패턴 변화를 분석하여 분류 기준을 업데이트 한다.
- **주기적으로 분류에 맞는 동작 수행**
  - Hot : 우선 순위가 낮은 컨테이너 접근 시 지연을 유도하거나 제한한다.
  - Warm : 자원 접근 시 동일한 격리 제공한다.
  - Cold : 높은 성능을 제공한다.

결론

- **연구 목적**
  - **컨테이너에서** RoCEv2 기반 BlueField-3 RNIC의 성능 격리 문제를 실험적 분석하였다.

- **실험 결과**
  - TX/RX Processing Unit 과 내부 Cache 자원 고갈.
    - 대역폭: 약 93.9% 감소, 지연 시간: 약 1,117배 증가, 캐시 미스: 115% 증가.

- **해결 방안 제시**
  - HT-verbs : Threshold 기반 자원 관리.
    - RDMA verbs 사용 패턴을 기준으로 Hot, Warm, Cold 분류 및 동작.

- **향후 연구**
  - RDMA Verbs 관점에서 RNIC의 마이크로아키텍처에 미치는 영향 분석.
  - HT-verbs 시스템을 구축하여 실제 환경에서 적용 가능성 검증.

감사합니다.