PASTE: A Networking API for Non-Volatile Main Memory

Michio Honda (NEC Laboratories Europe)

Lars Eggert (NetApp)

Douglas Santry (NetApp)

TSVAREA@IETF 99, Prague

May 22th 2017

More details at our HotNets paper: https://goo.gl/Ssreei

Motivation

SSICLOPS

- Non-Volatile Main Memories (NVMMs)
  - Persistent
  - Byte-addressable
  - Low latency
    - 10s-1000s of ns

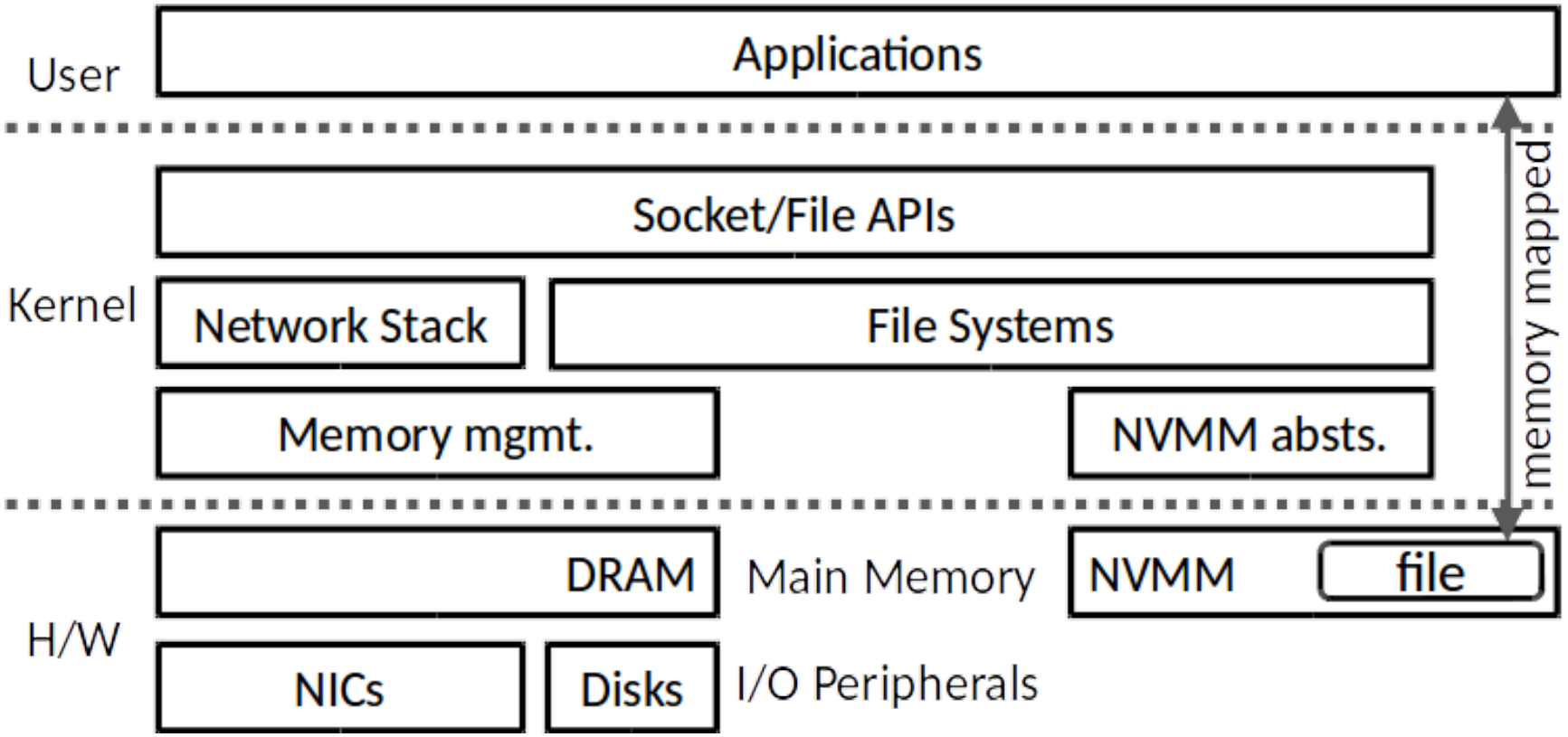
https://www.hpe.com/us/en/servers/persistent-memory.html

- A lot of work in databases and file systems
  - NVHeap [ASPLOS'11], NOVA [FAST'16], NVWal [ASPLOS'16], HiKV [ATC'17] etc

What are implications for networking?

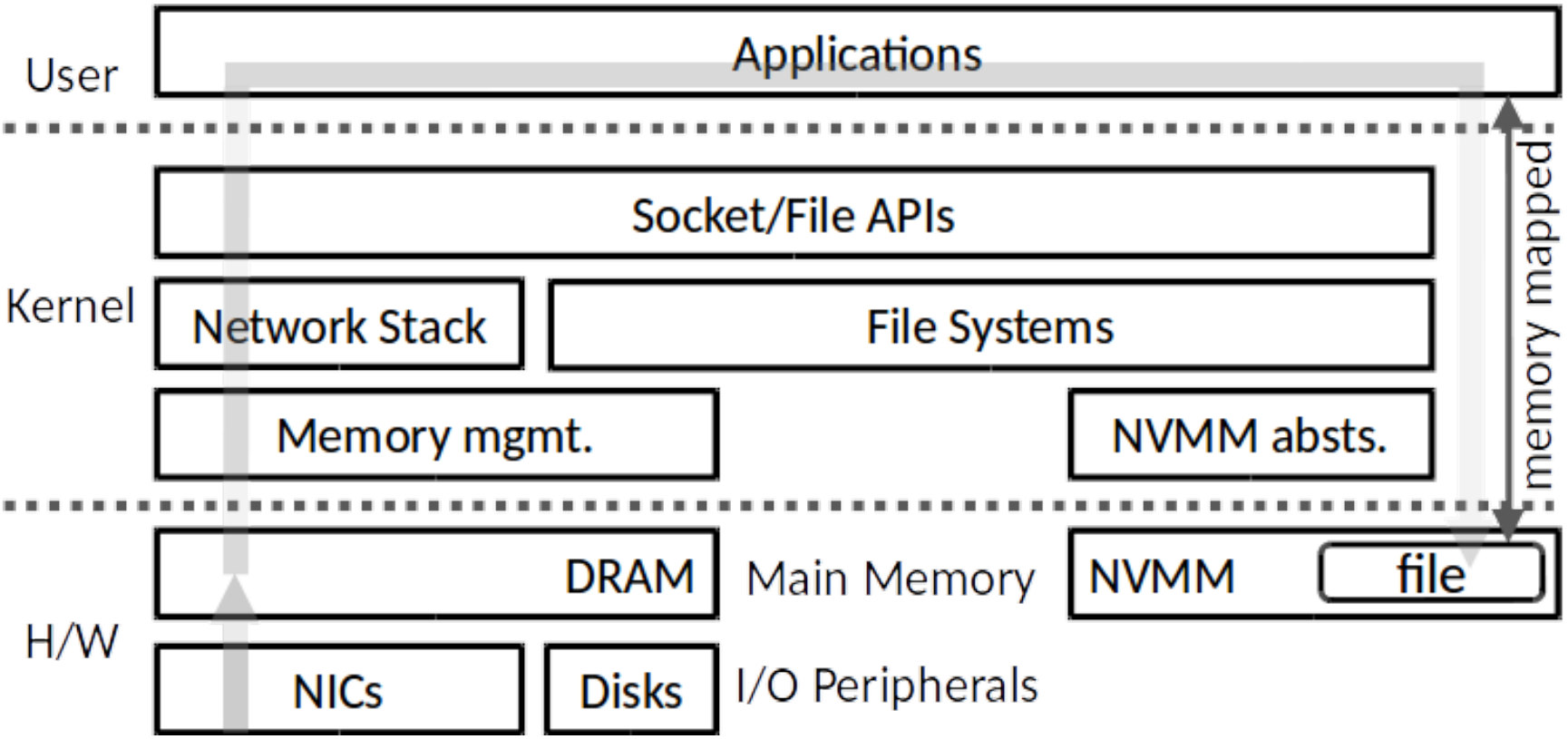
Review: Today's Stack with NVMM

SSICLOPS



Review: Today's Stack with NVMM – How data move from the network to the storage

SSICLOPS



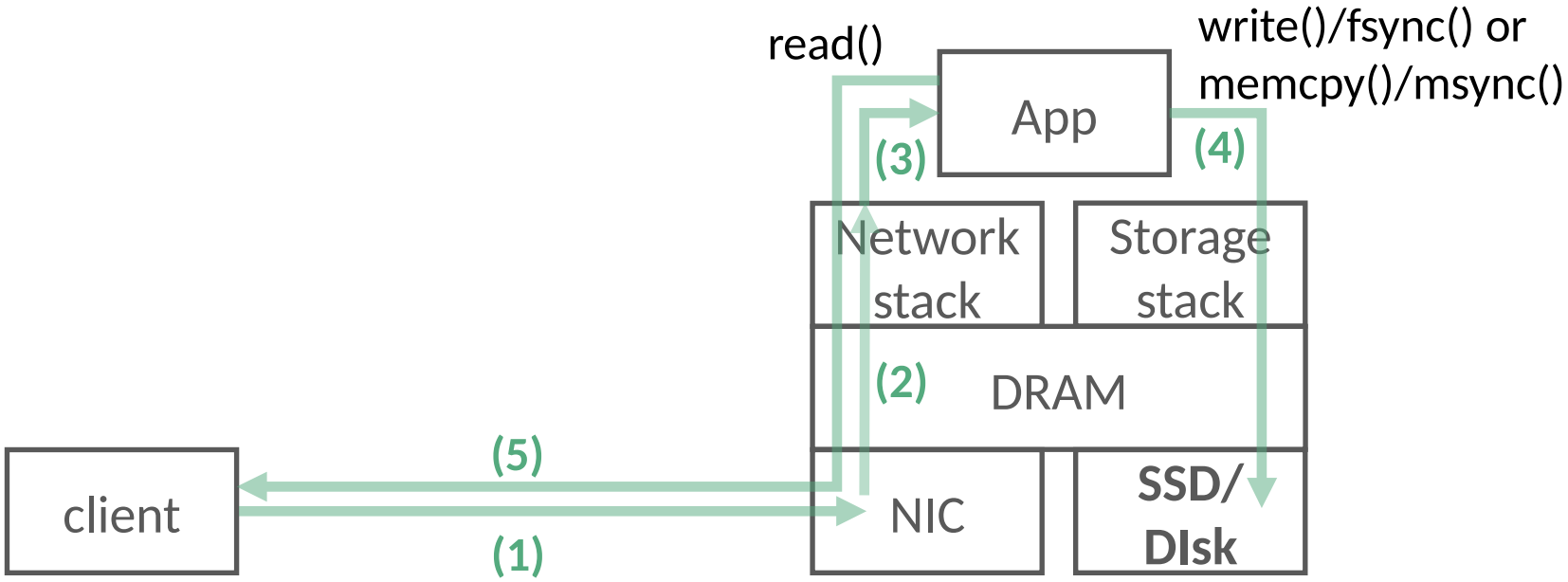
Case Study: Careful Data Transfer

- Server persists client's request prior to acknowledgment
- e.g., 1KB commit:

Case Study: Careful Data Transfer

SSICLOPS

- Server persists client's request prior to acknowledgment
- e.g., 1KB commit:

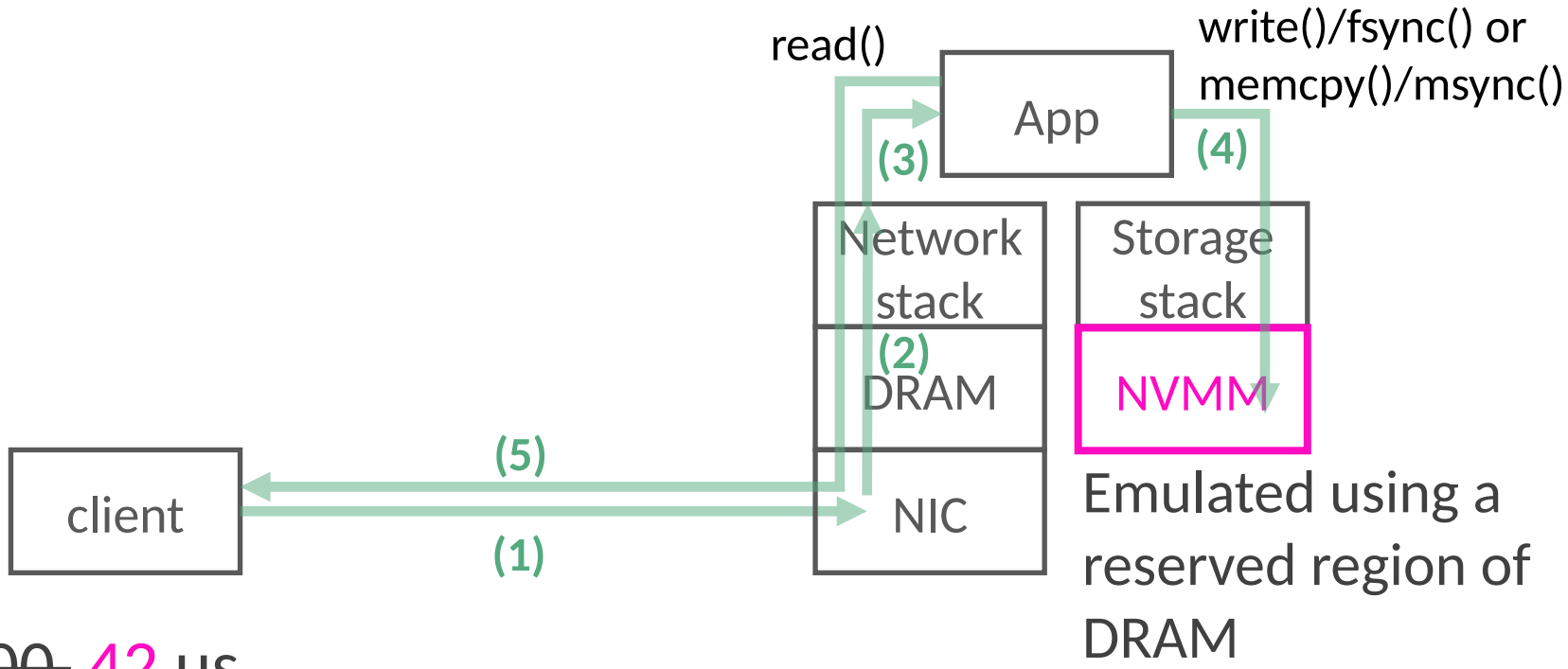


- 2030 us
  - Networking (w/o step (4)) takes 40 us

Case Study: Careful Data Transfer

SSICLOPS

- Server persists client's request prior to acknowledgment
- e.g., 1KB commit:

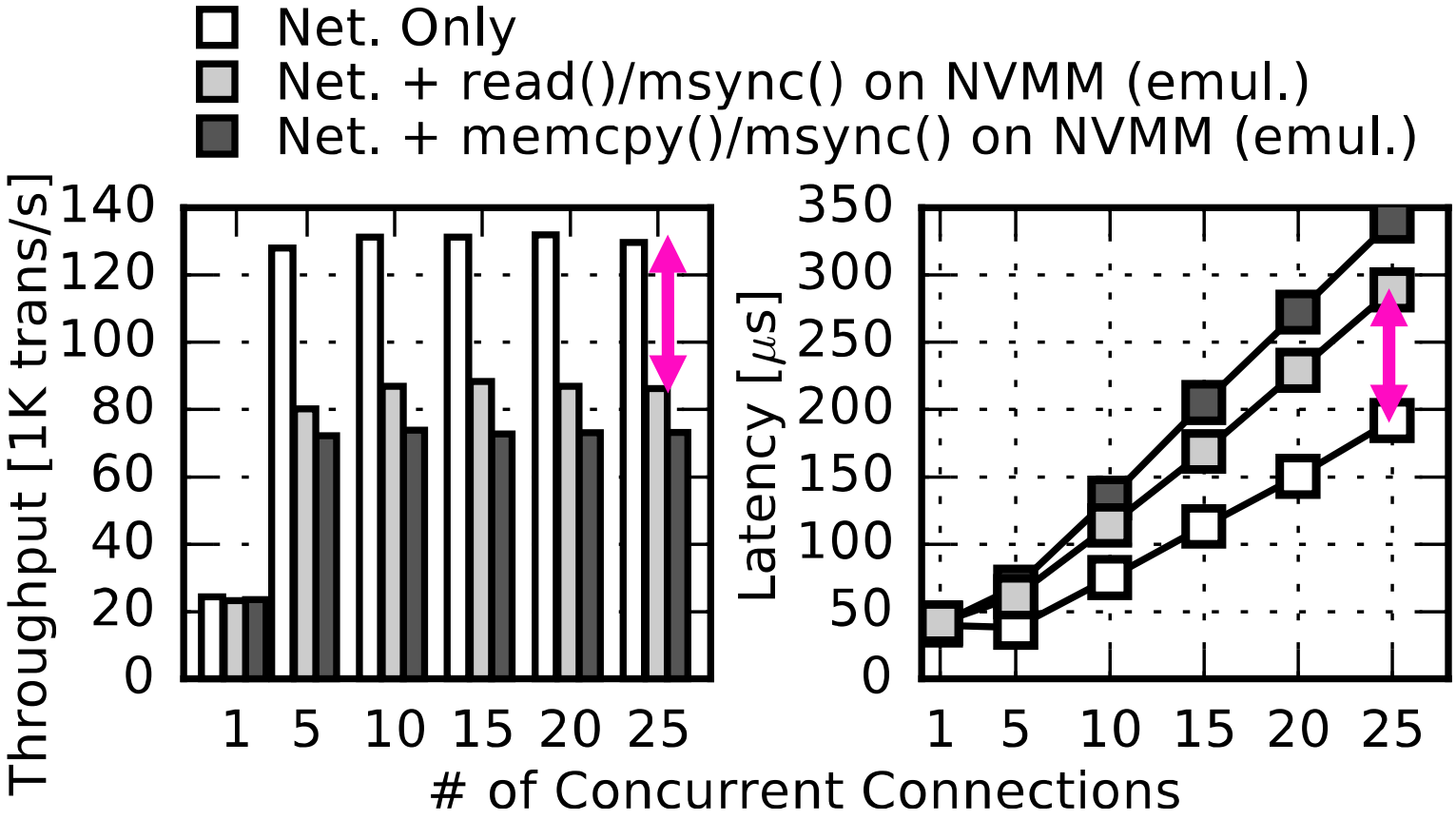


- ~~2000~~ 42 us
  - Networking takes 40 us
- This 2 us is not small

Case Study: Careful Data Transfer

SSICLOPS

- Parallel requests are serialized on each core

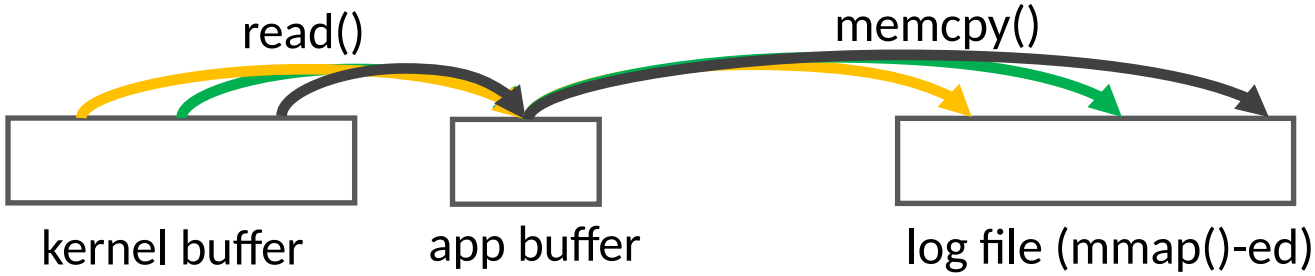


33 % throughput decrease, 50 % latency increase

Data Copies Matter

SSICLOPS

- **Cache Misses**
  - **Persisting data (e.g., to a log) always happens to a different destination**



	Overall cache misses	Major Contributor
Networking only	0.0004 %	net_rx_action() (84%)
Networking + NVMM (read() + memcpy() + msync())	**4.4121** %	**memcpy() (98%)**

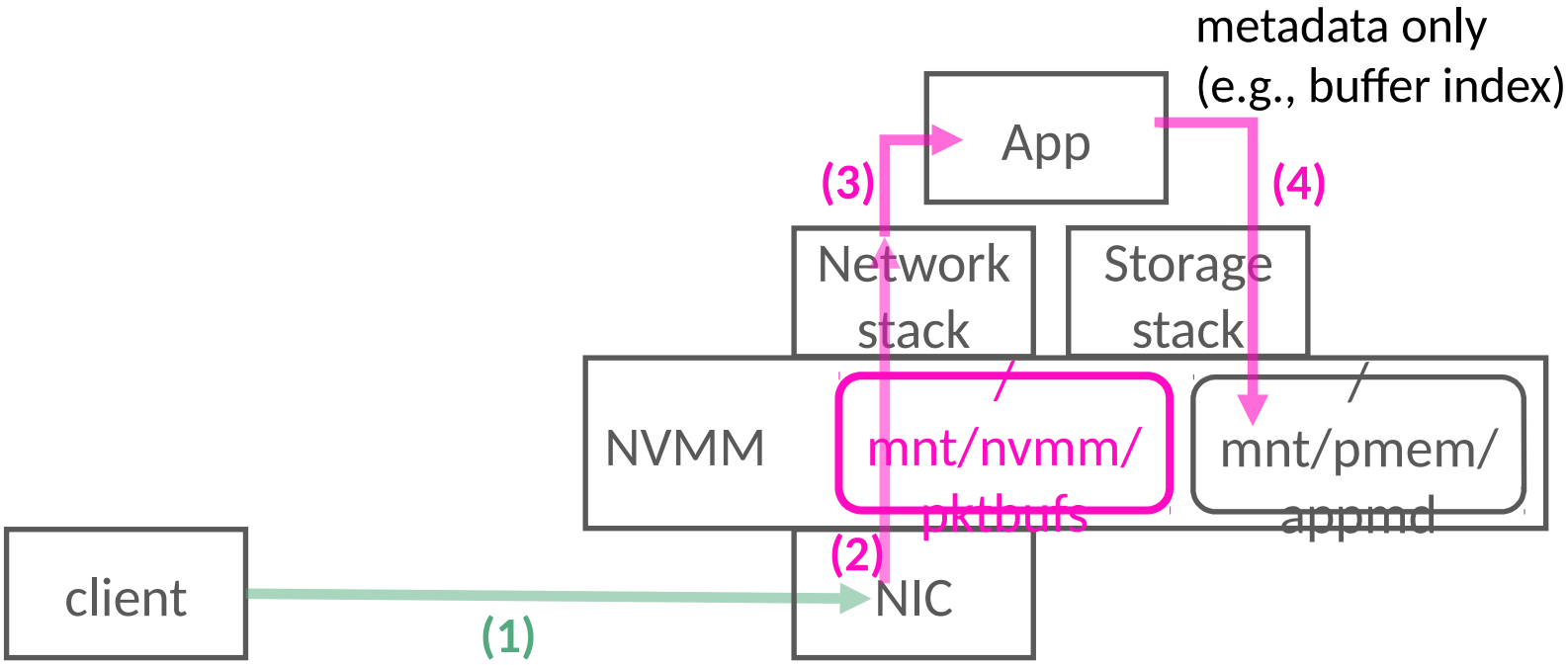
Reported by Linux perf

We must avoid data copy!

Packet Store (PASTE) Overview

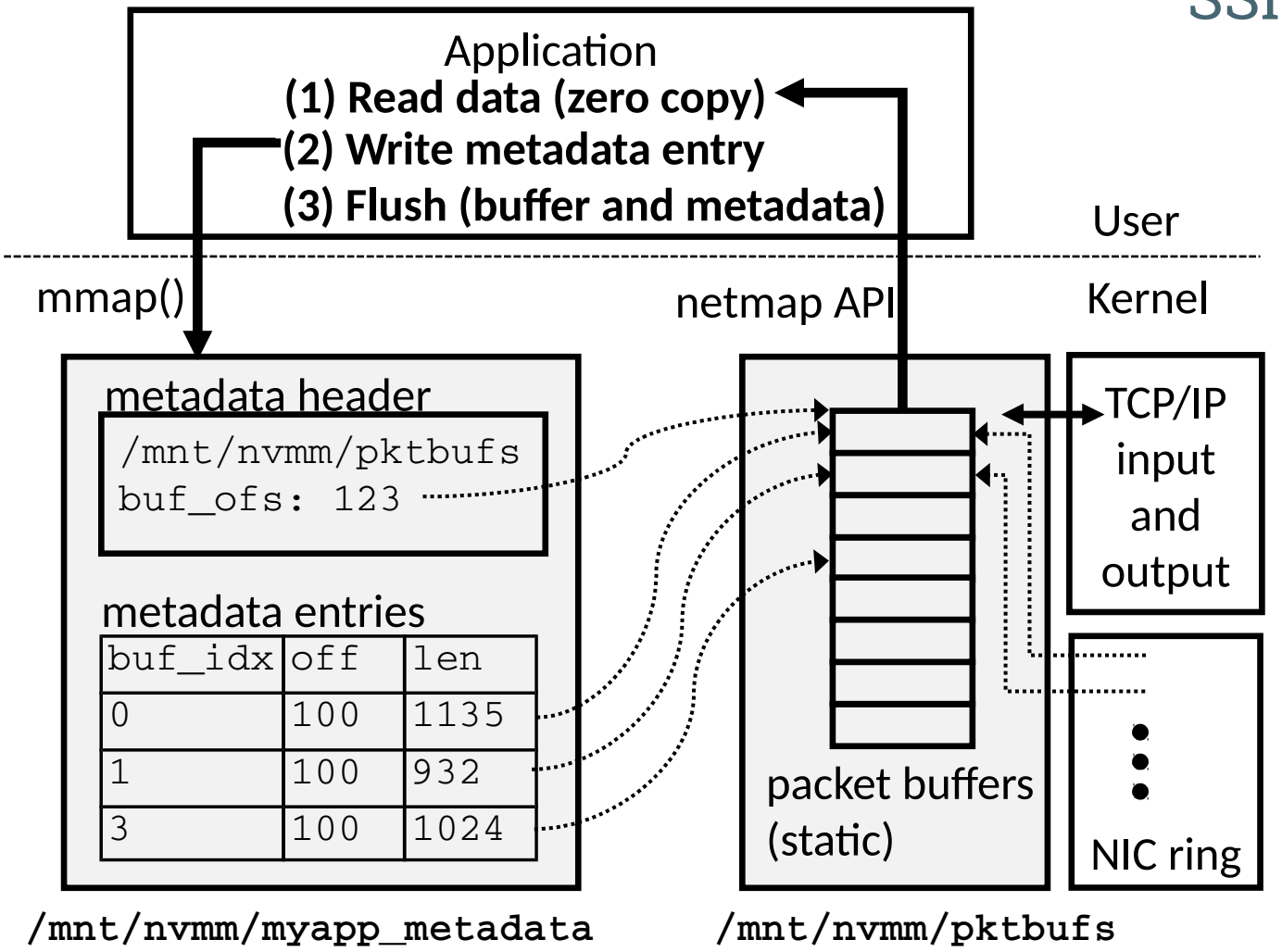
SSICLOPS

- **Packet buffers on a named NVMM region**
  - DMA to NVMM
- **Zero-copy APIs**



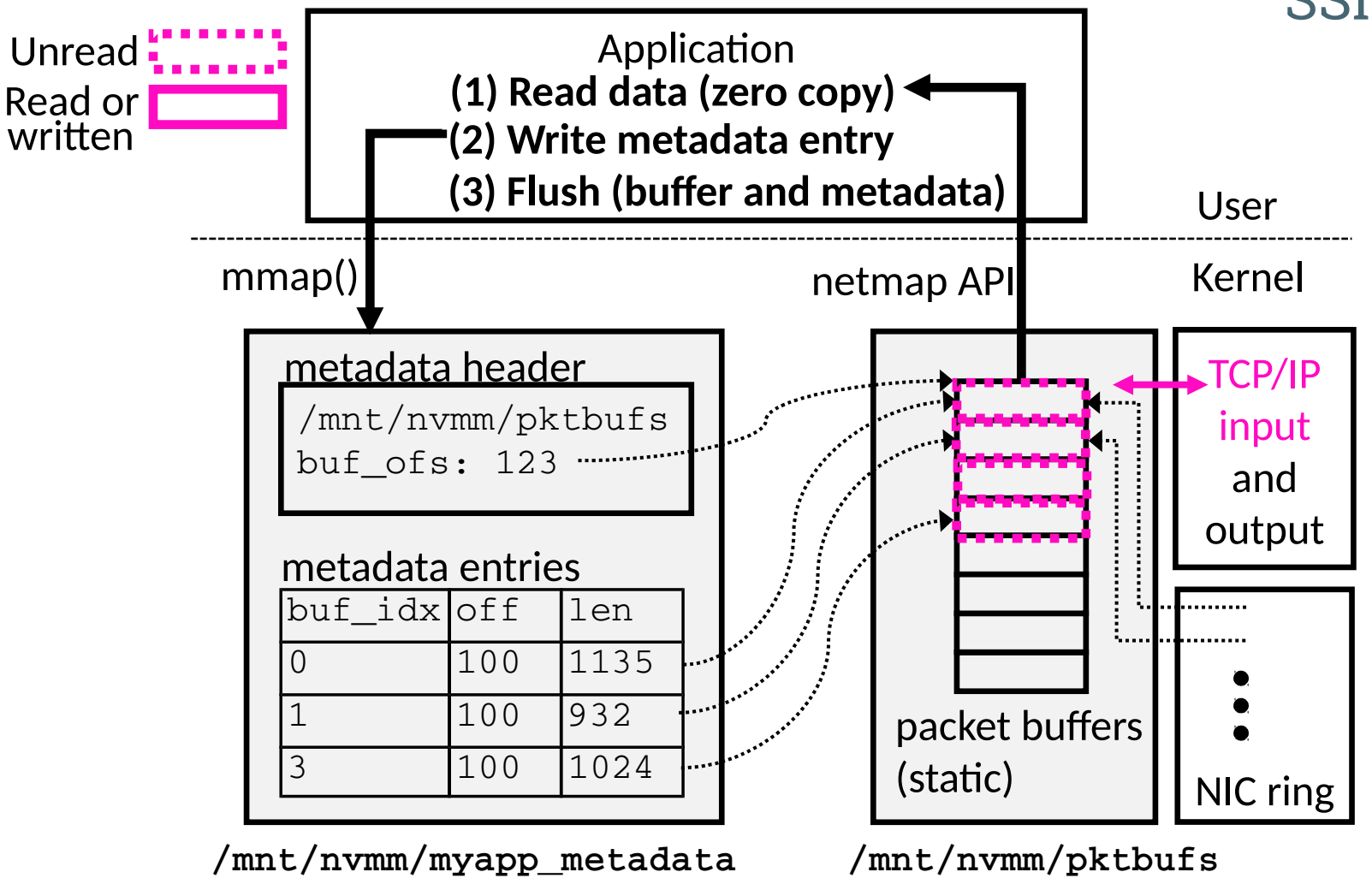
Fast Persistent Write with PASTE

SSICLOPS



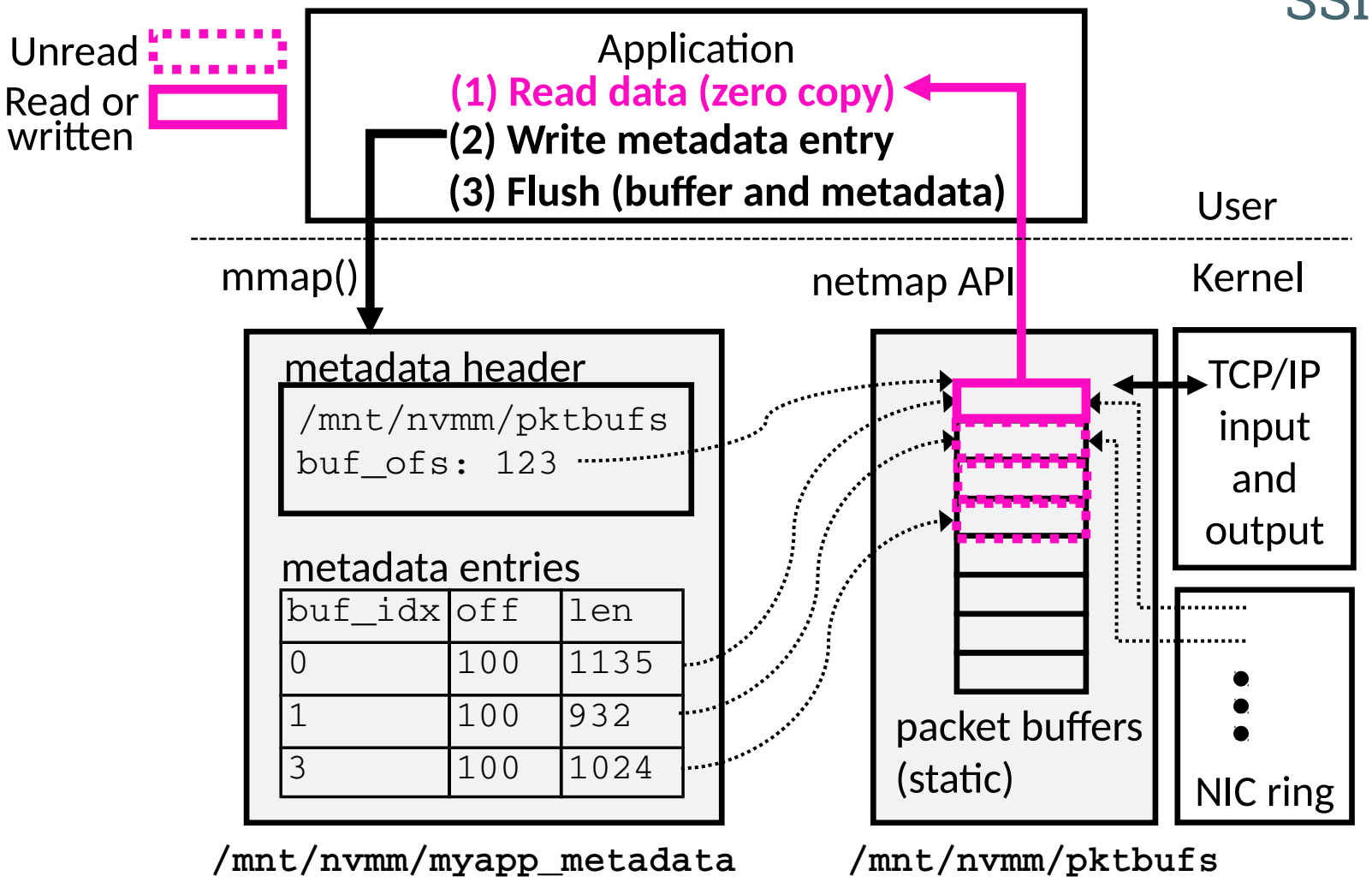
Fast Persistent Write with PASTE

SSICLOPS



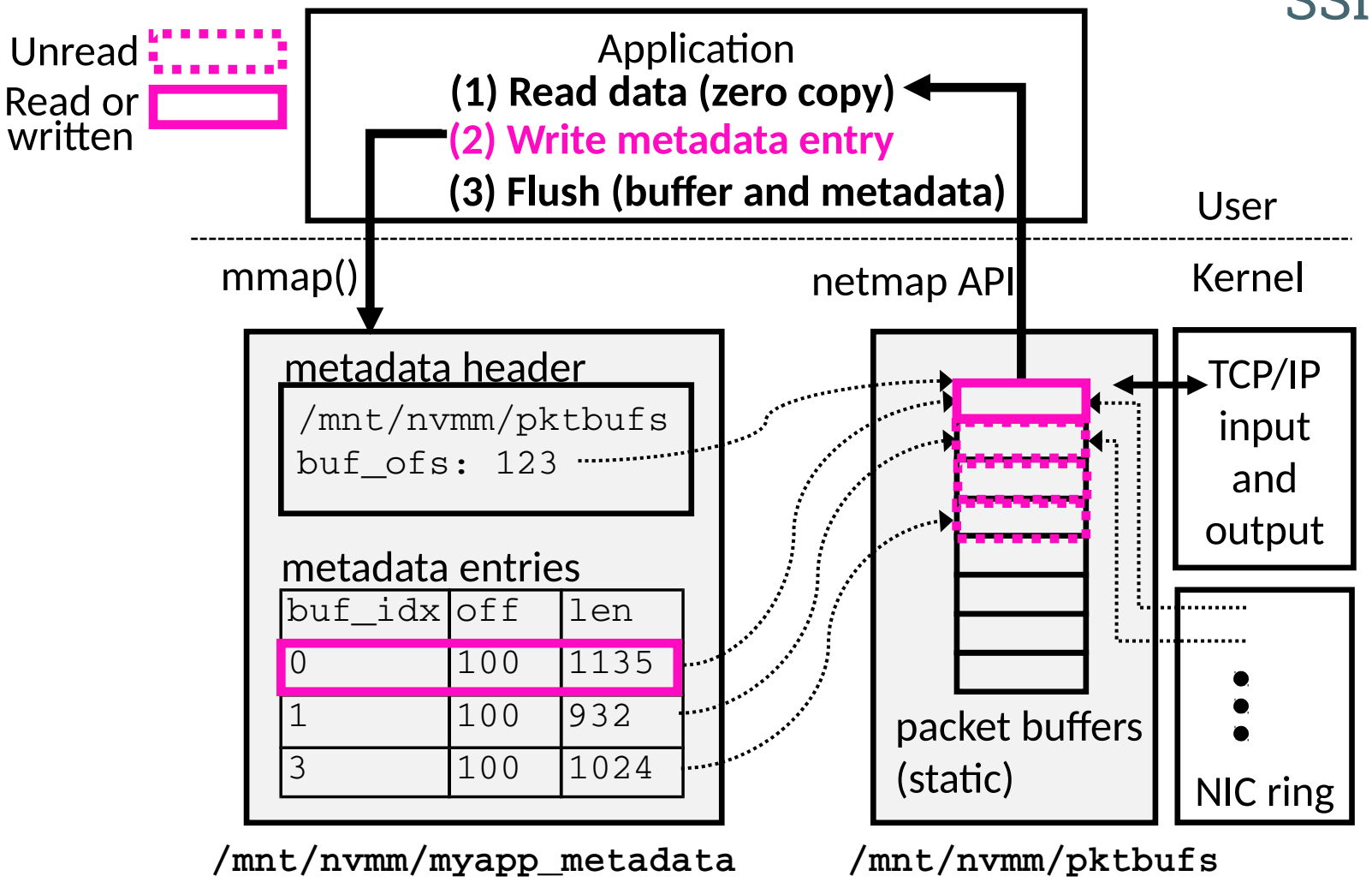
Fast Persistent Write with PASTE

SSICLOPS



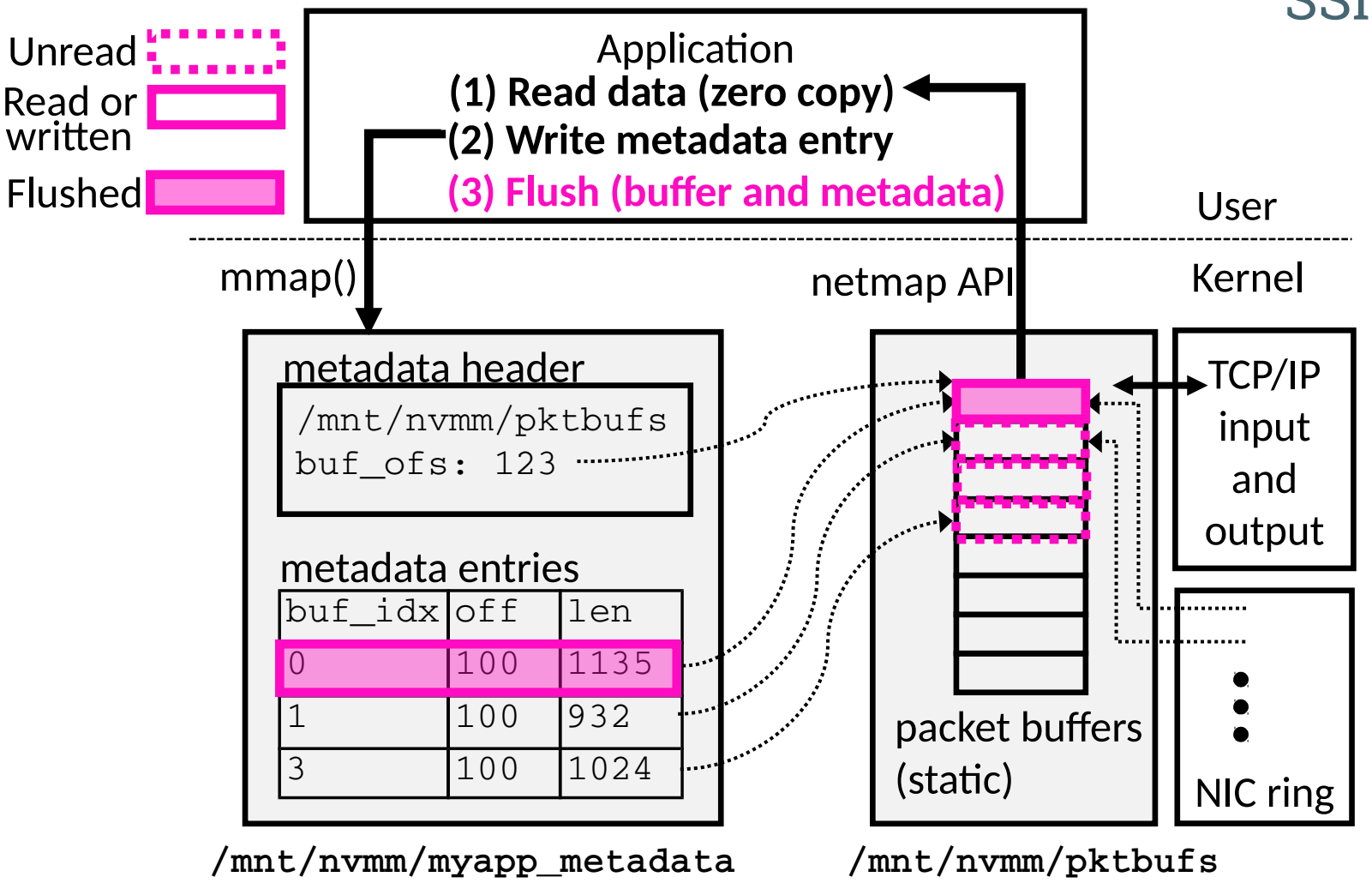
Fast Persistent Write with PASTE

SSICLOPS



Fast Persistent Write with PASTE

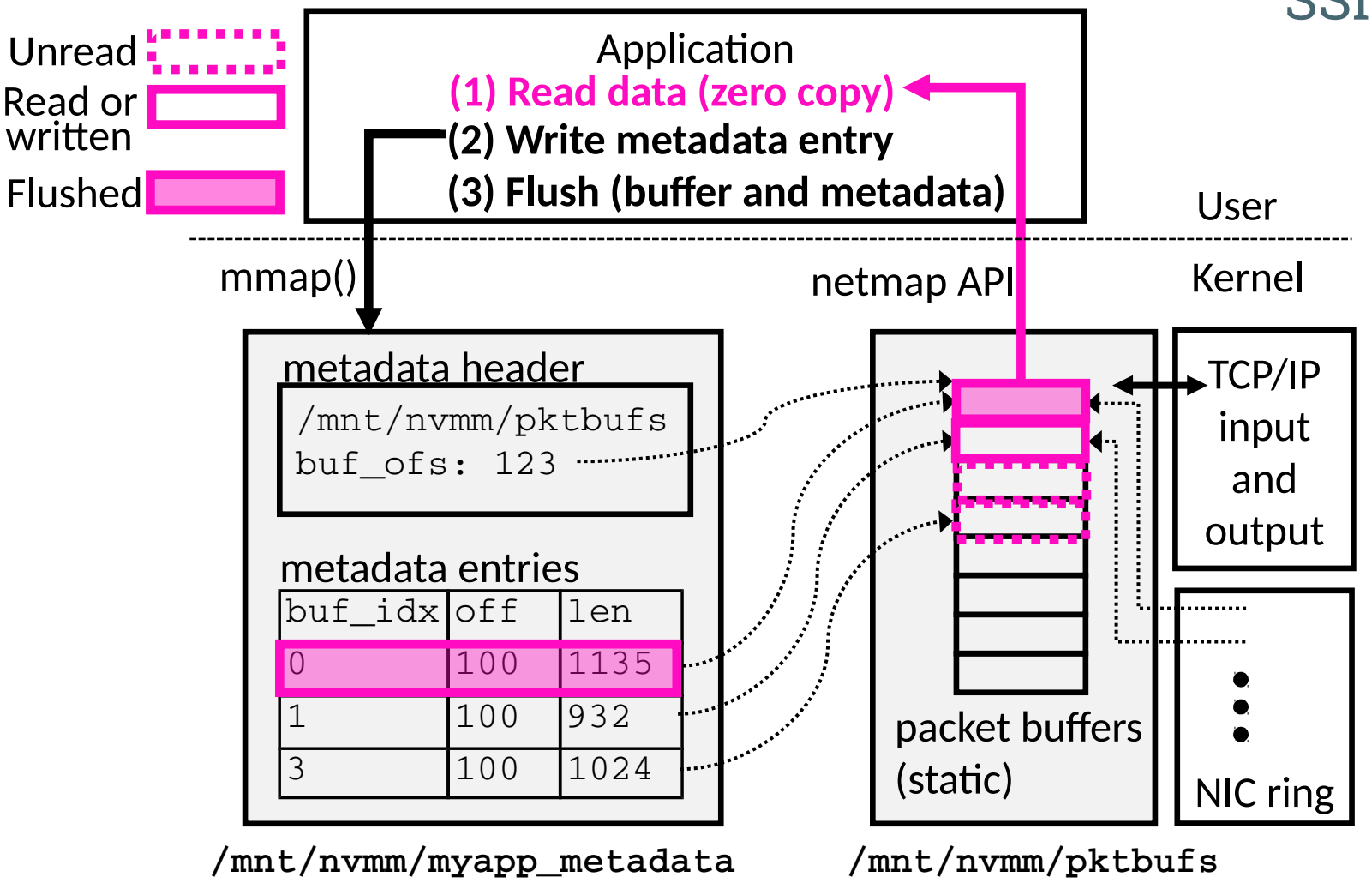
SSICLOPS



DMA is performed to L3 cache (DDIO)

Fast Persistent Write with PASTE

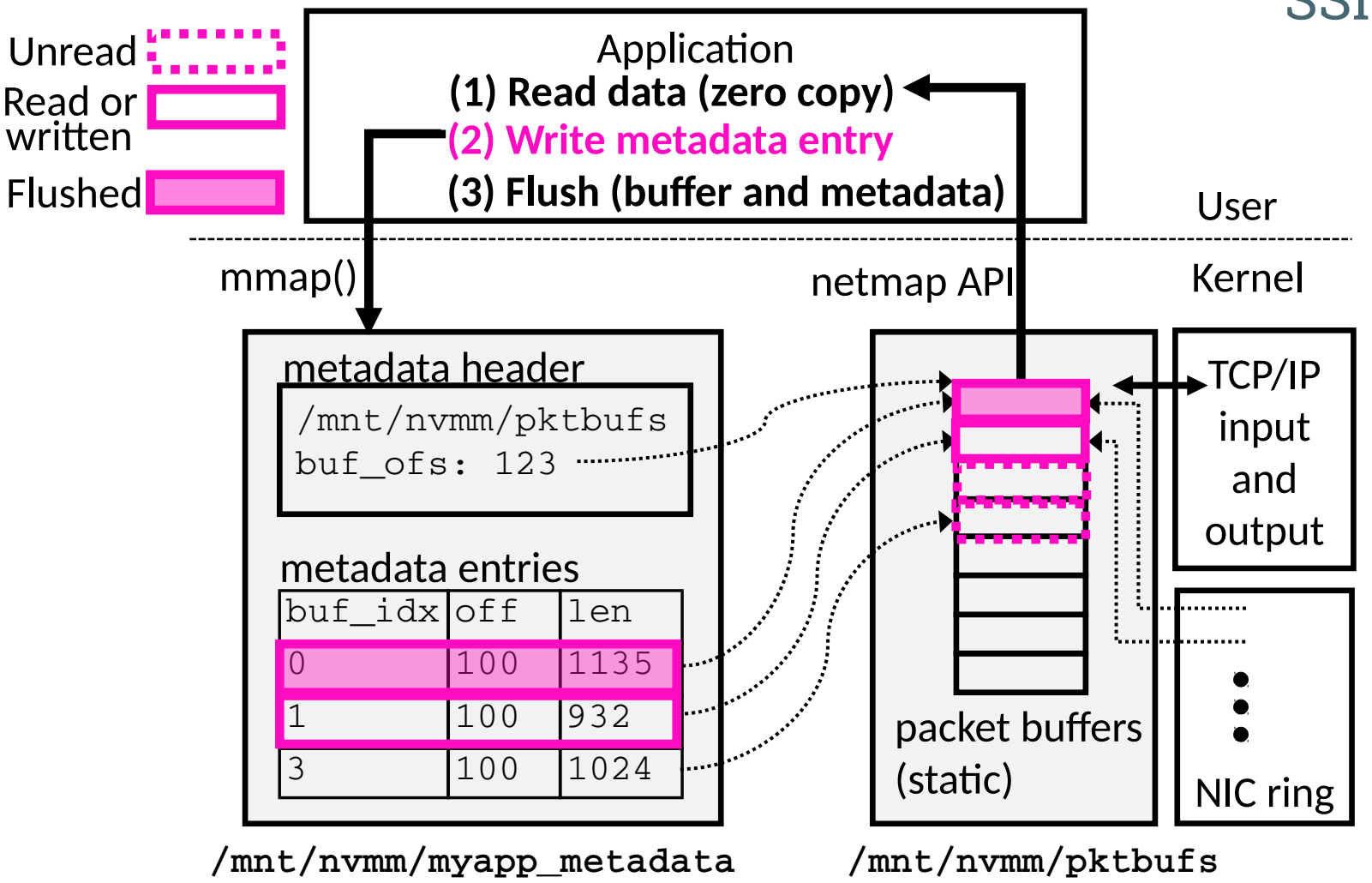
SSICLOPS



DMA is performed to L3 cache (DDIO)

Fast Persistent Write with PASTE

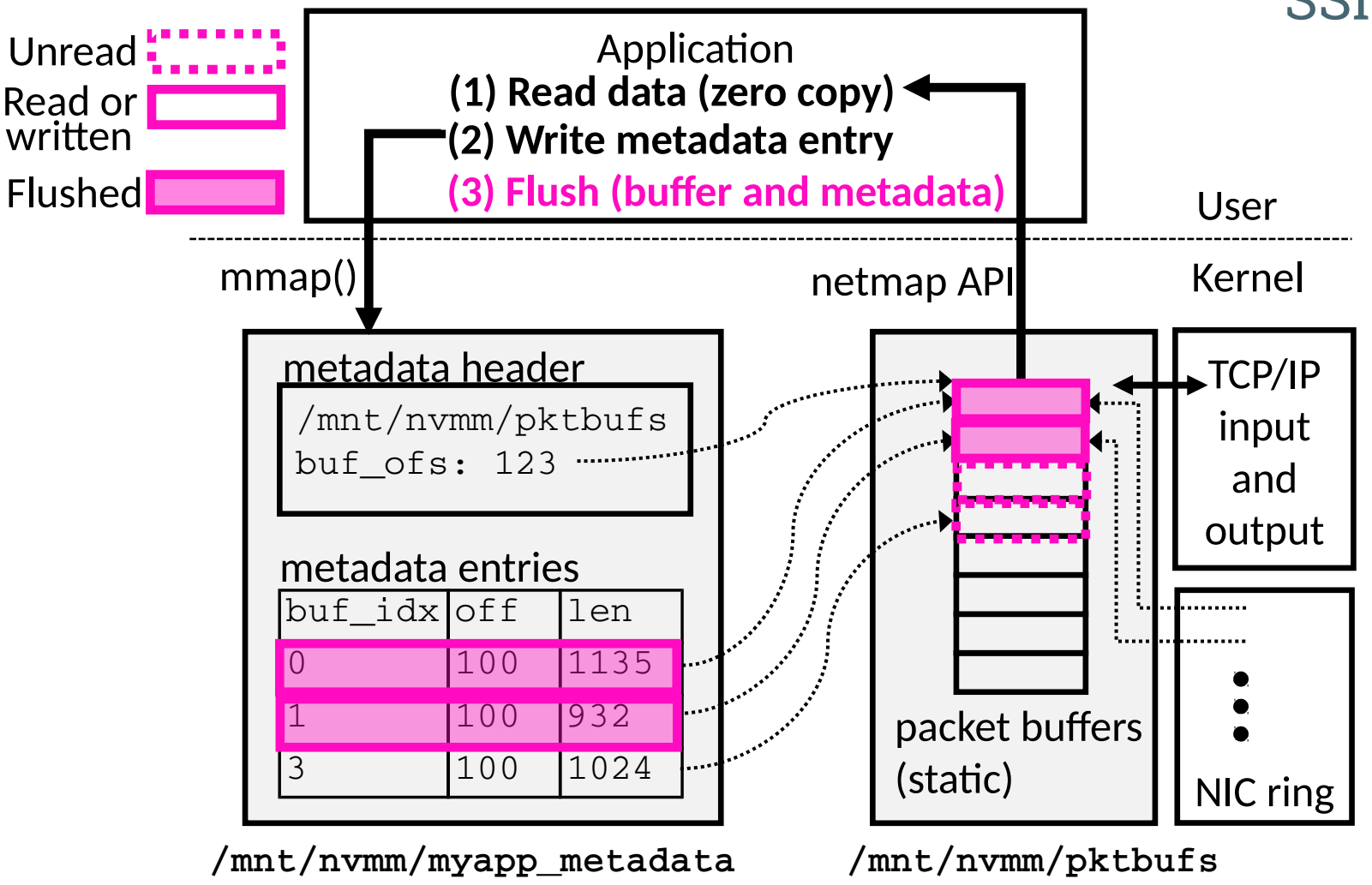
SSICLOPS



DMA is performed to L3 cache (DDIO)

Fast Persistent Write with PASTE

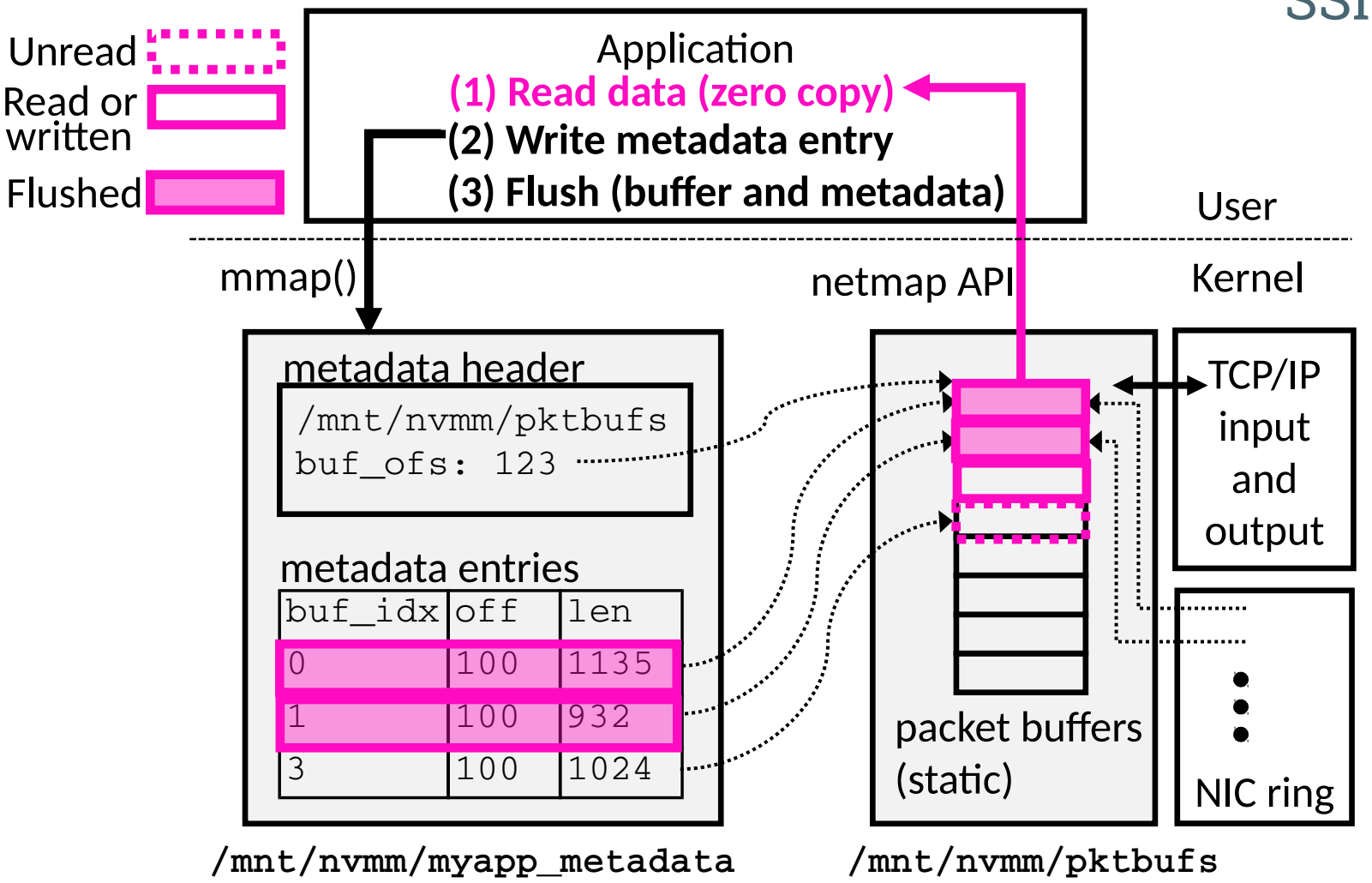
SSICLOPS



DMA is performed to L3 cache (DDIO)

Fast Persistent Write with PASTE

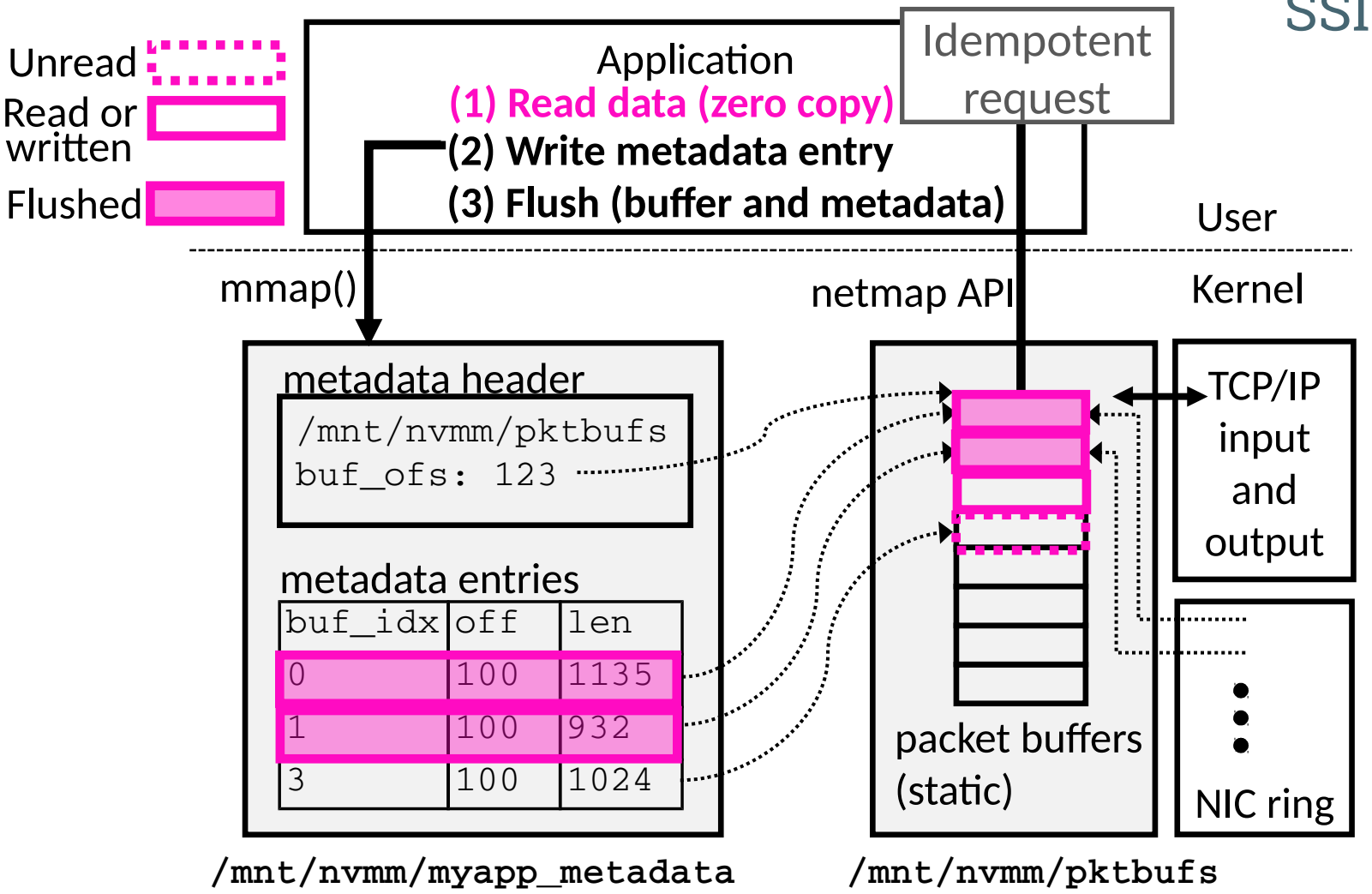
SSICLOPS



DMA is performed to L3 cache (DDIO)

Fast and Selective Persistent Write with PASTE

SSICLOPS

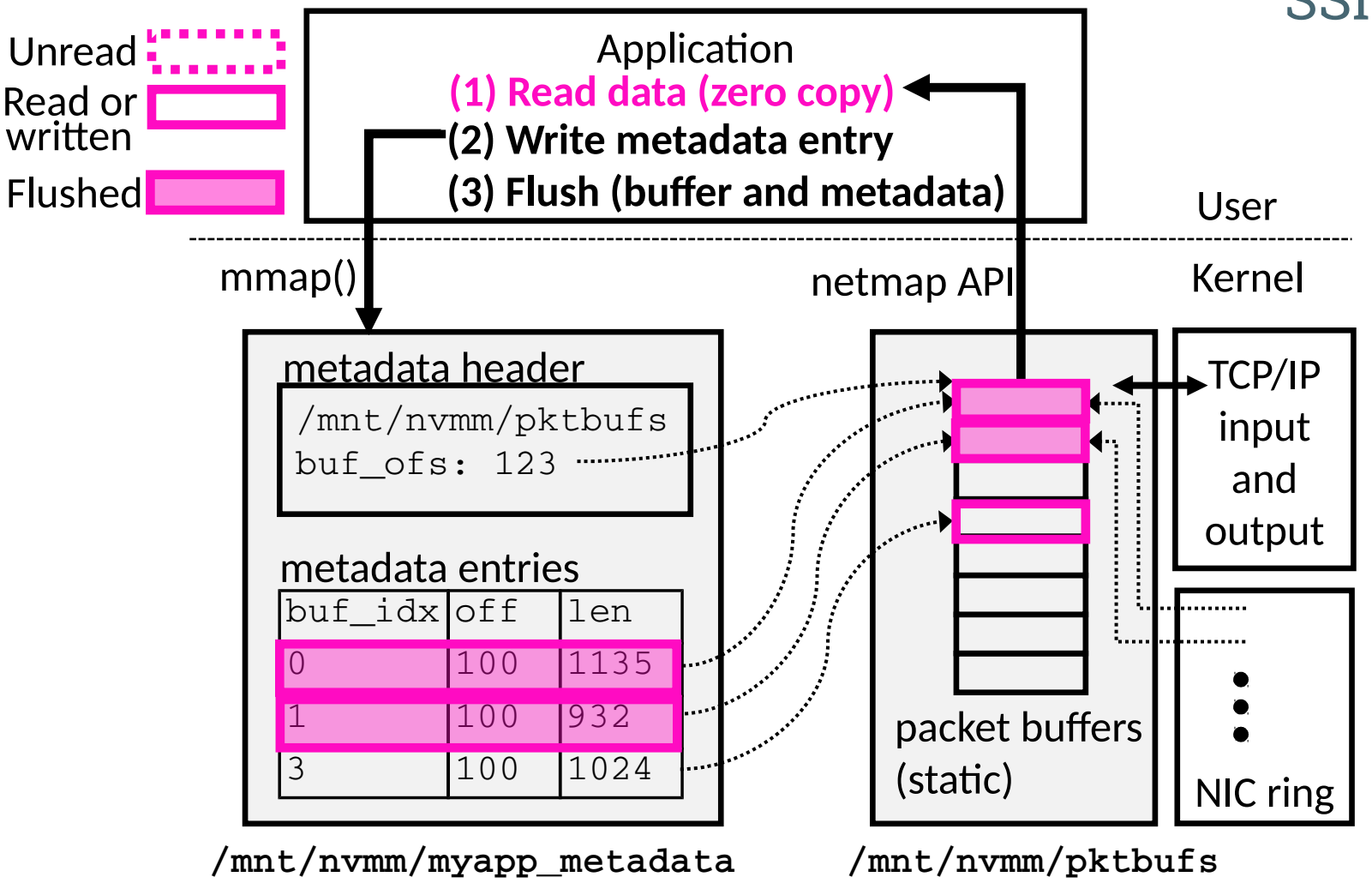


DMA is performed to L3 cache (DDIO)

Unnecessary data is not flushed to DIMM

Fast and Selective Persistent Write with PASTE

SSICLOPS

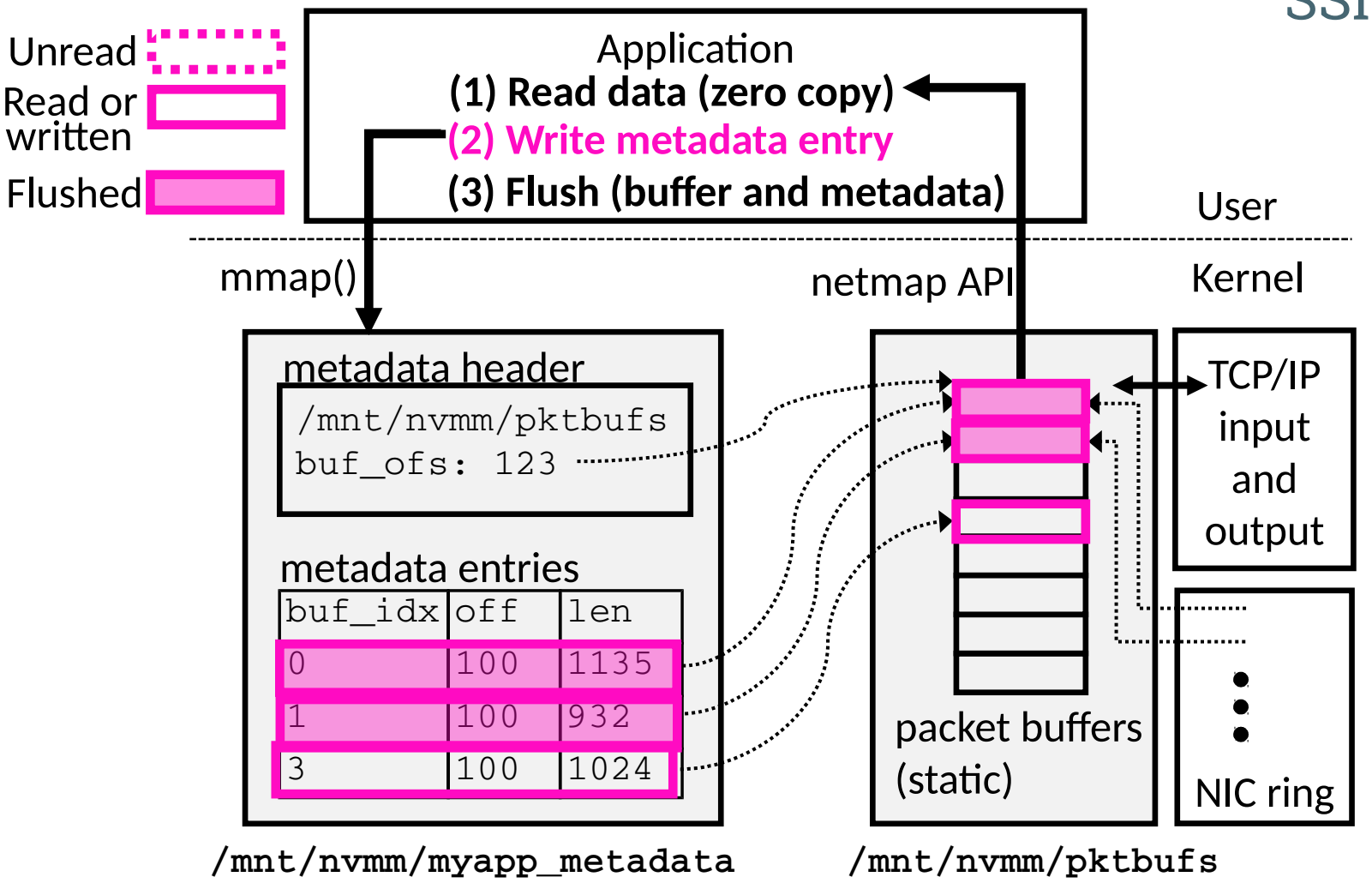


DMA is performed to L3 cache (DDIO)

Unnecessary data is not flushed to DIMM

Fast and Selective Persistent Write with PASTE

SSICLOPS

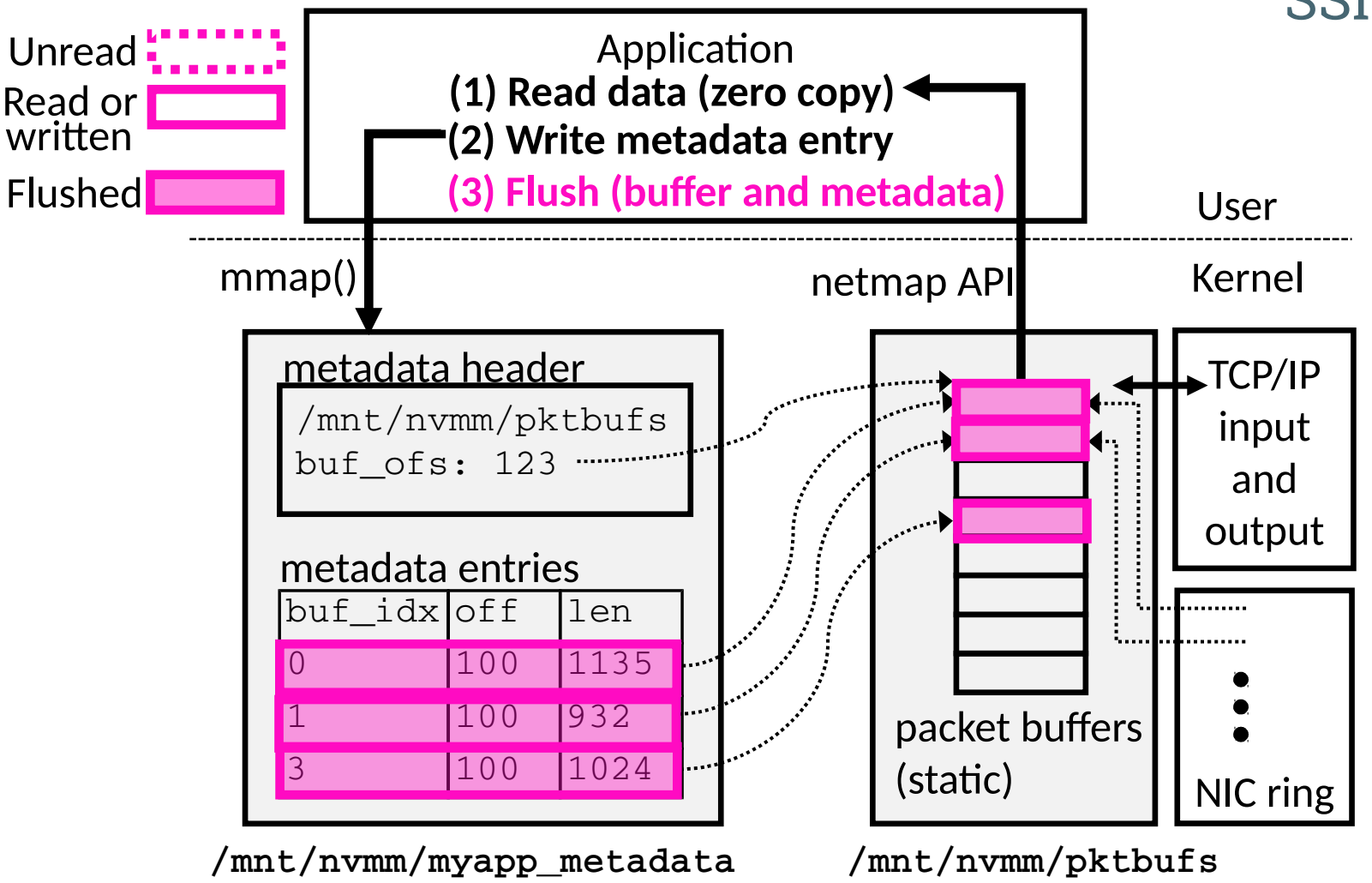


DMA is performed to L3 cache (DDIO)

Unnecessary data is not flushed to DIMM

Fast and Selective Persistent Write with PASTE

SSICLOPS



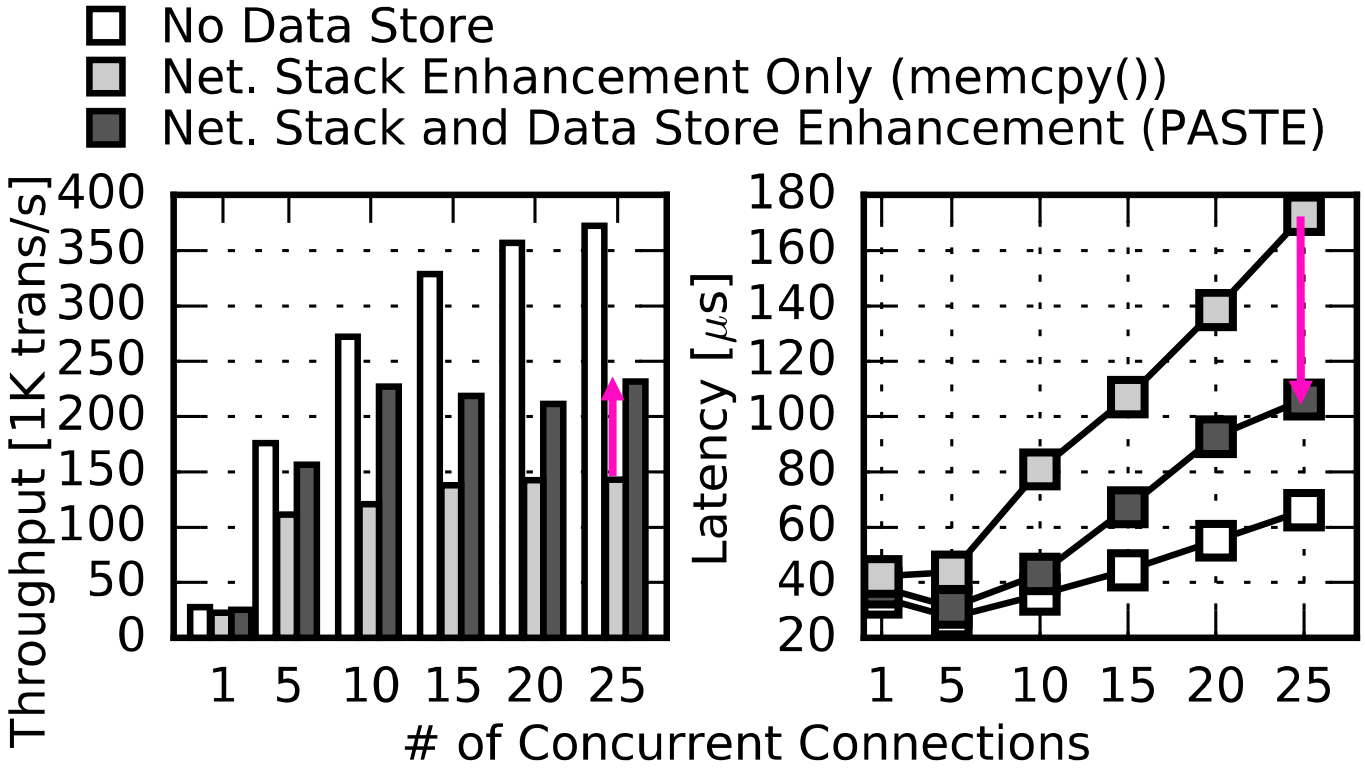
DMA is performed to L3 cache (DDIO)

Unnecessary data is not flushed to DIMM

Preliminary Results

SSICLOPS

- Implementation
  - Extend the netmap framework
    - TCP/IP is also supported



10-88 % throughput increase, 9-46 % latency reduction

Related Work

- Enhanced network stacks
  - MegaPipe (OSDI'12), Stackmap (ATC'16), Fastsocket (ASPLOS'16)
  - IX and Arrakis (OSDI'14), mTCP (NSDI'13), Sandstorm (SIGCOMM'14), MICA (NSDI'14)

No NVMM aware

- NVMM filesystems
  - BPFS (SOSP'09), NOVA (FAST'15)
- NVMM databases
  - NVWAL (ASPLOS'15), REWIND (VLDB'15), NV-Tree (FAST'15)

No networking aware

Conclusion

SSICLOPS

- Networking APIs are now a bottleneck for durably storing data
- Ongoing work
  - Brushing up APIs
    - Smooth integration with netmap API
  - Forming a filesystem with packet buffers
    - Fast data movement between files
  - Use case applications