Alibaba Cloud
Worldwide Cloud Services Partner

Deploying User-space TCP at Cloud Scale with LUNA

Lingjun Zhu, Yifan Shen, Erci Xu , Bo Shi, Ting Fu, ***Shu Ma***, Shuguang Chen, Zhongyu Wang, Haonan Wu, Xingyu Liao, Zhendan Yang, Zhongqing Chen, Wei Lin, Yijun Hou, Rong Liu, Chao Shi, Jiaji Zhu, and Jiesheng Wu

Alibaba Group

Background & Motivation

Design

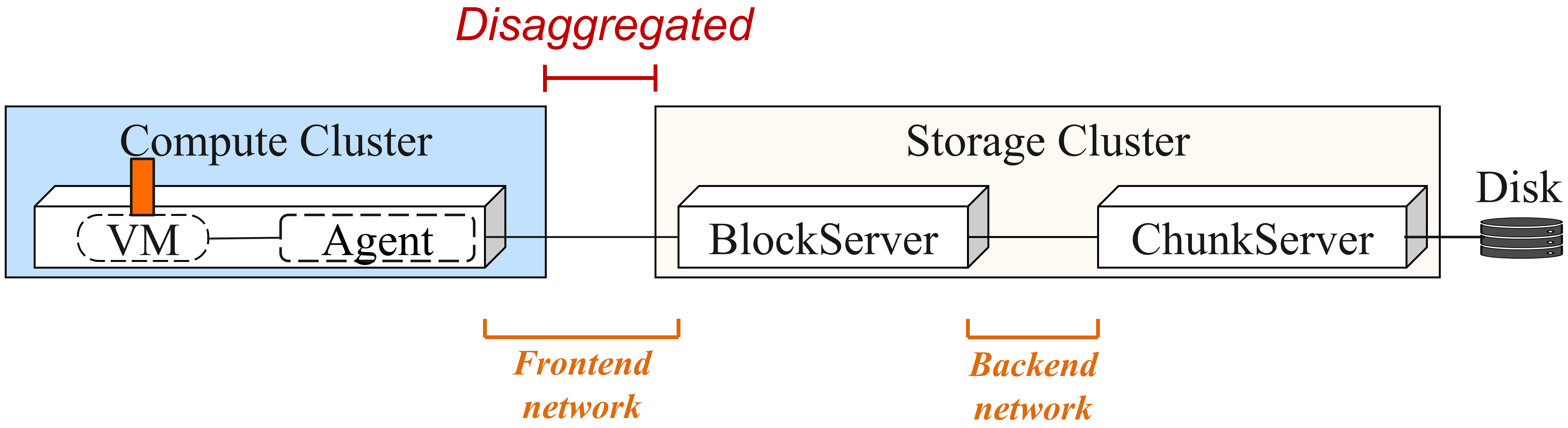
Evaluation

Discussion & Lessons

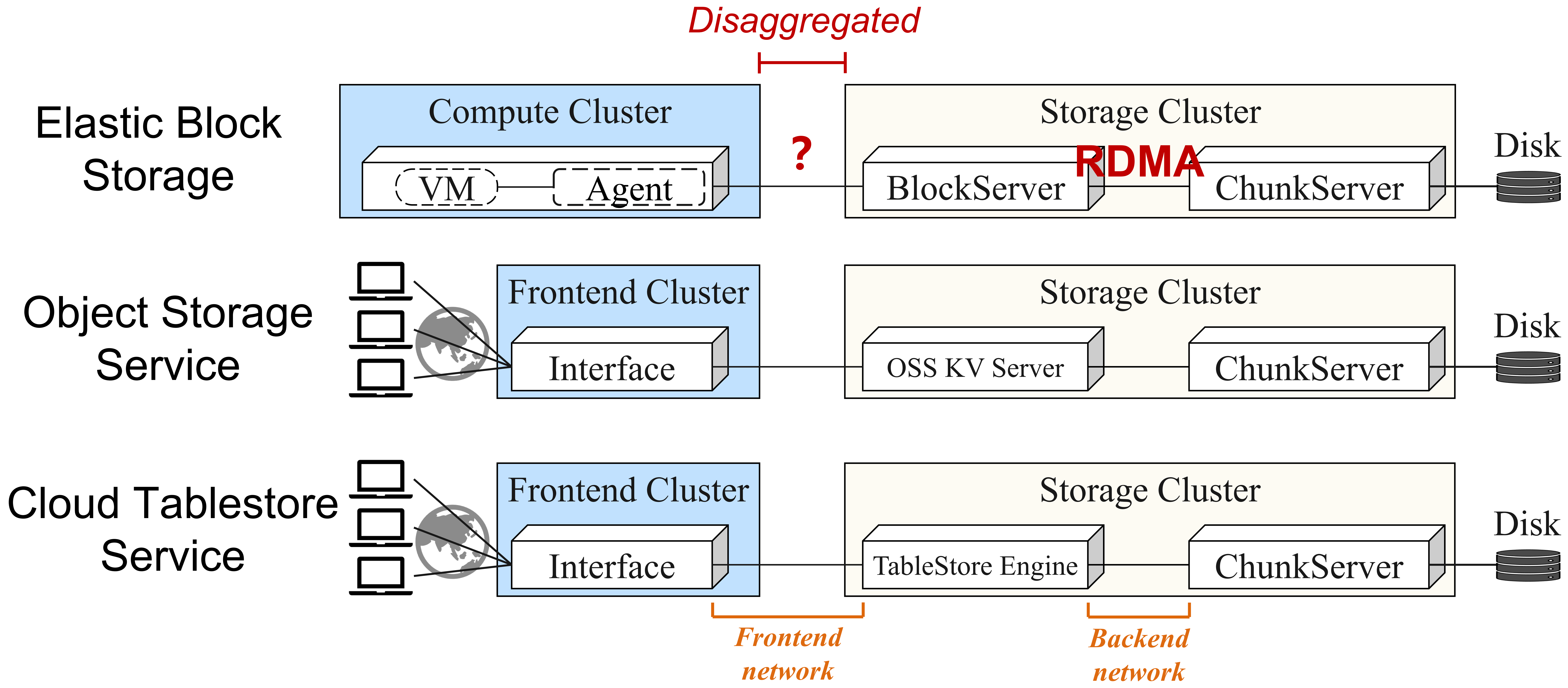
Conclusion

Alibaba Cloud Storage Network

Elastic Block Storage

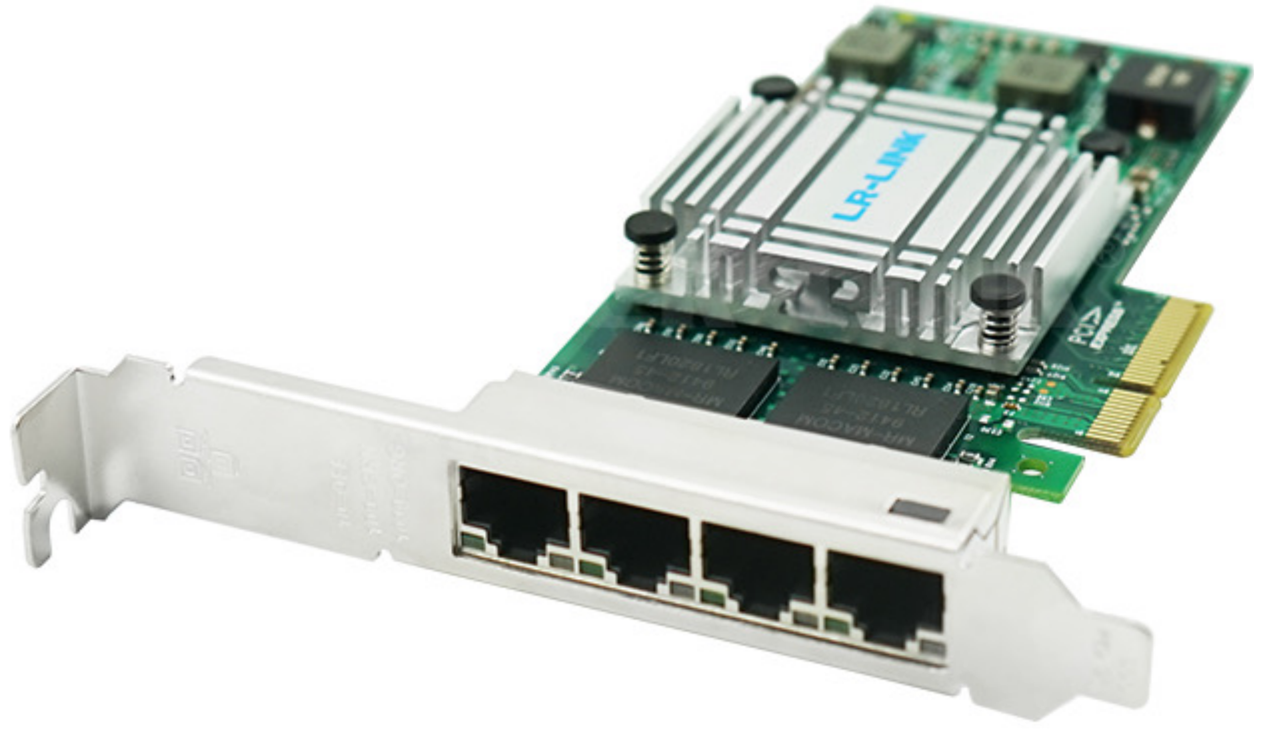


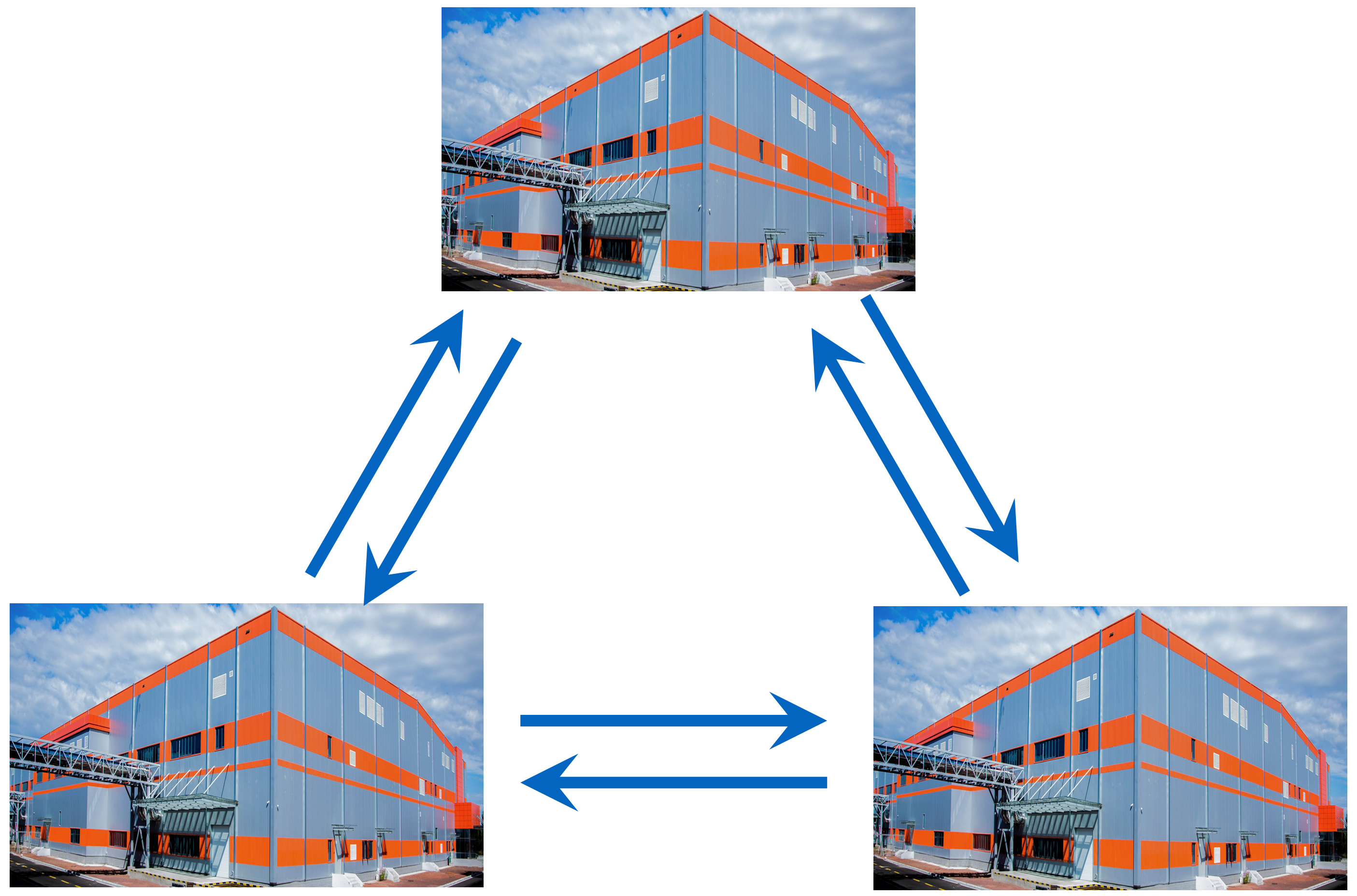
Alibaba Cloud Storage Network

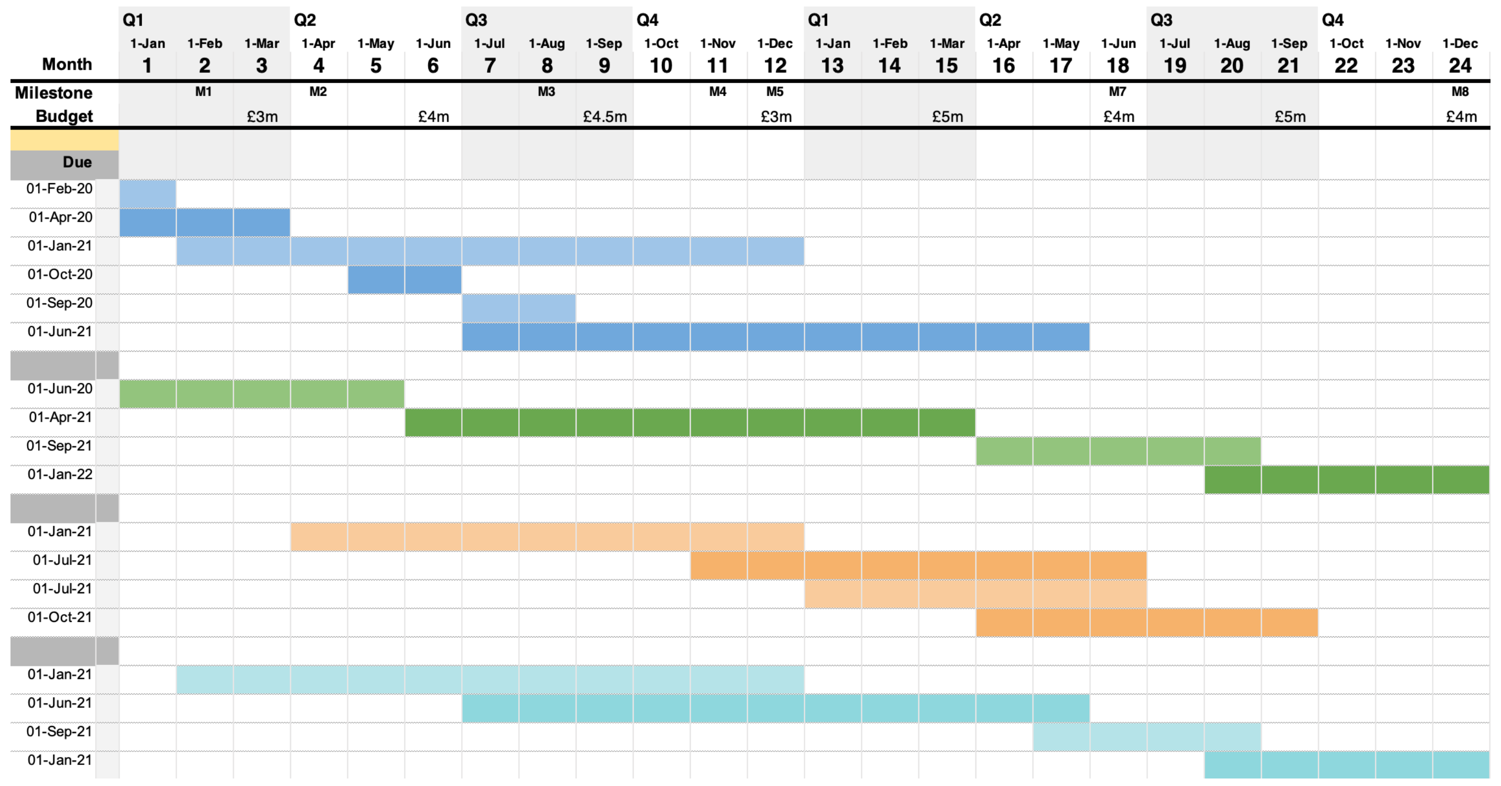


[1] NSDI 2021, When cloud storage meets RDMA

Requirements of Frontend Networks



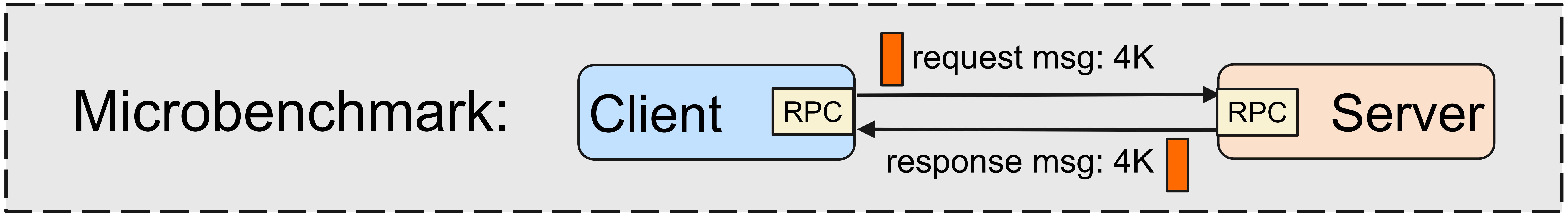




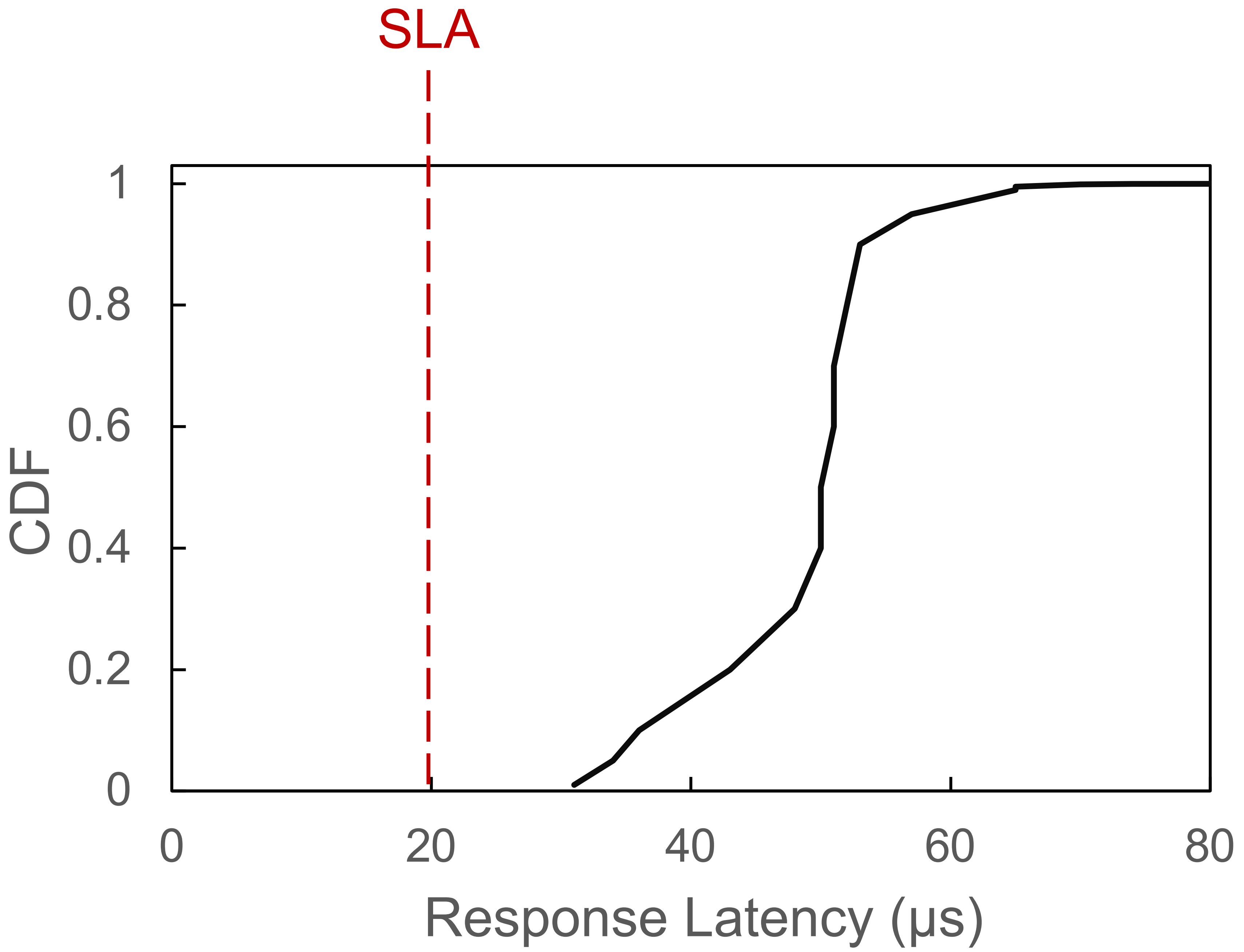
- ➢ **Diverse software and hardware**
- ➢ **Large amount of connections for inter-DC communications**
- ➢ **Engineering effort**

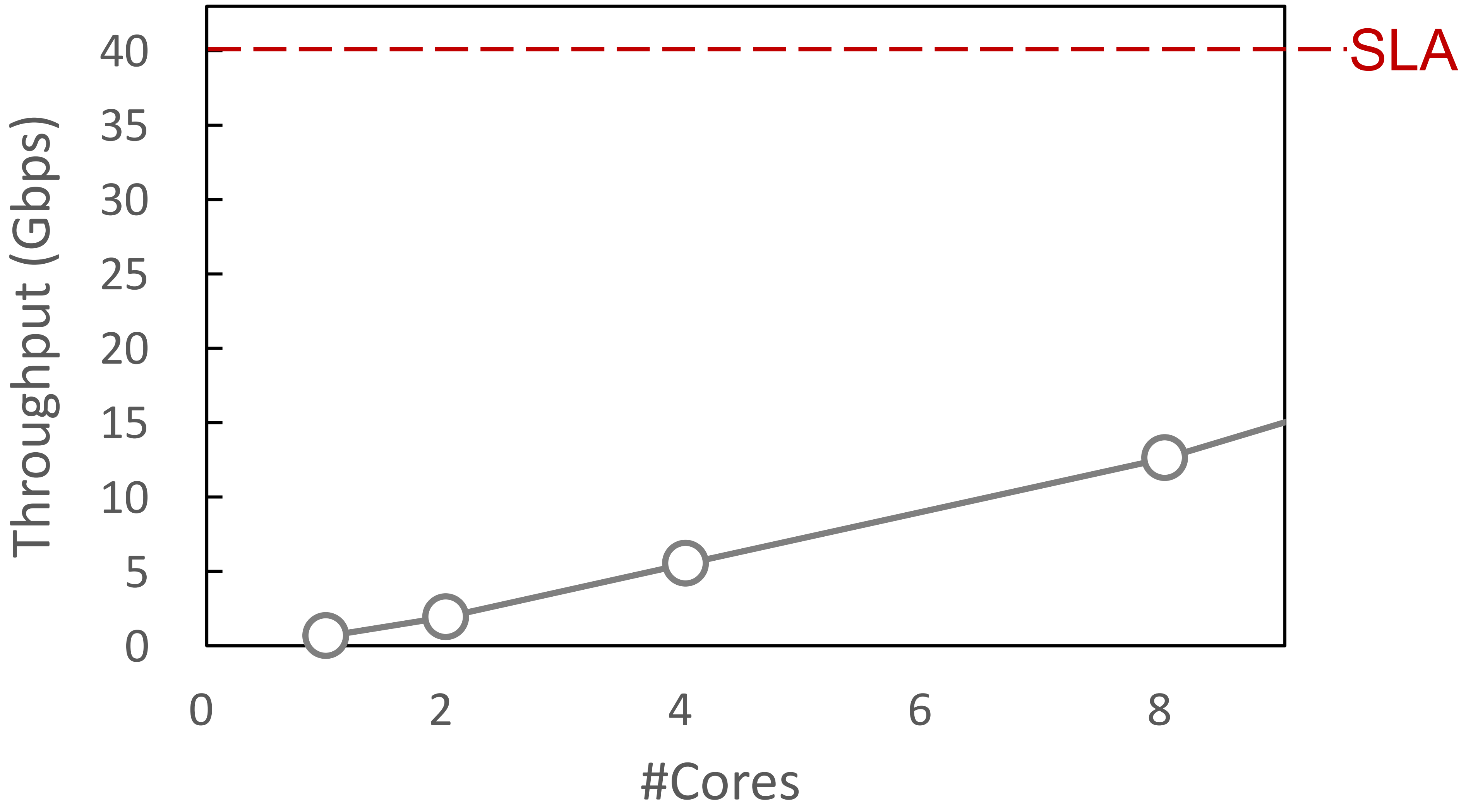
Kernel TCP can satisfy them all, but ...

Revisiting Kernel TCP



Microbenchmark:





No longer suitable for cloud storage

Beyond Kernel TCP

RDMA

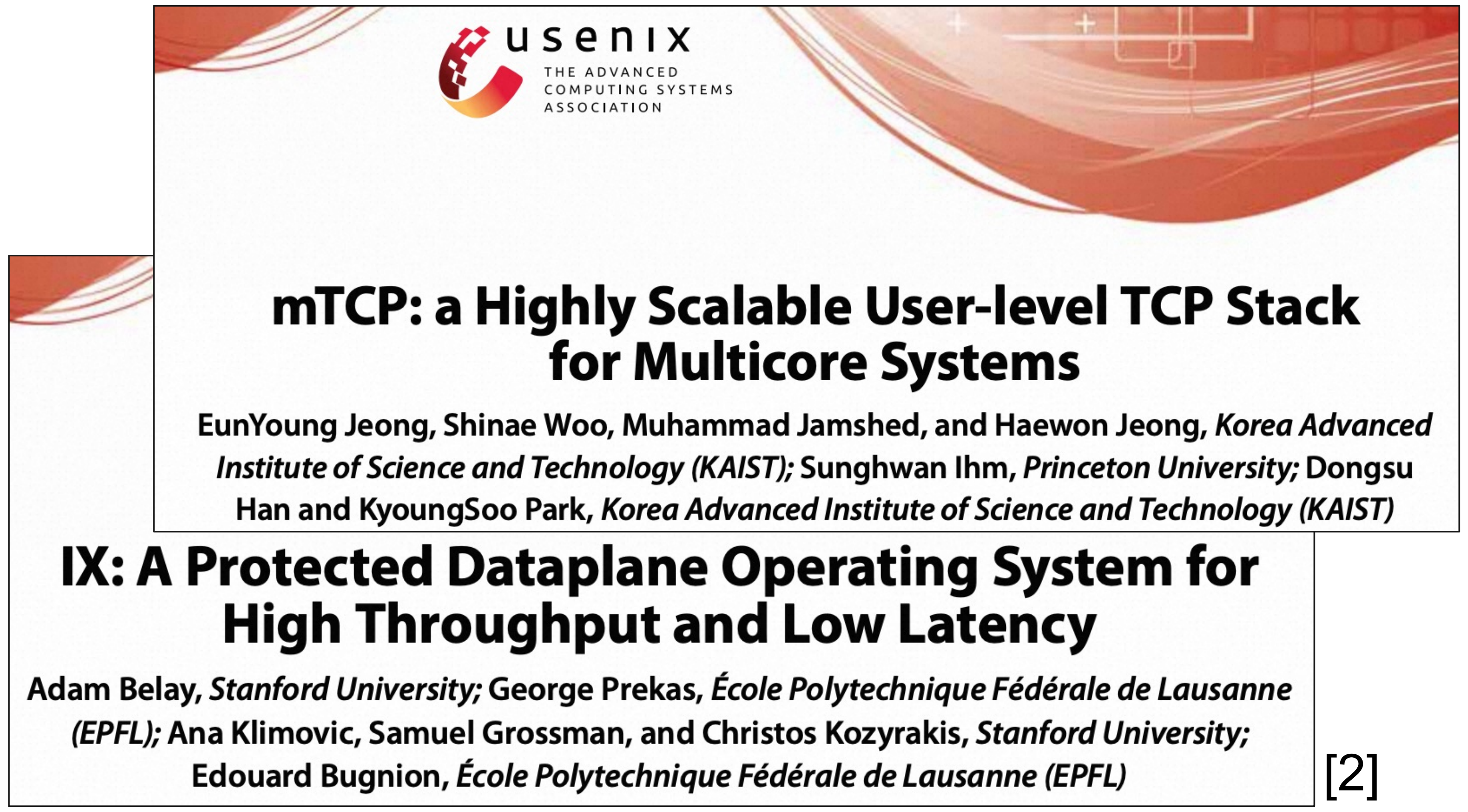
When Cloud Storage Meets RDMA

Yixiao Gao, *Nanjing University and Alibaba Group;* Qiang Li, Lingbo Tang, Yongqing Xi, Pengcheng Zhang, Wenwen Peng, Bo Li, Yaohui Wu, Shaozong Liu, Lei Yan, Fei Feng, Yan Zhuang, Fan Liu, Pan Liu, Xingkui Liu, Zhongjie Wu, Junping Wu, and Zheng Cao, *Alibaba Group;* Chen Tian, *Nanjing University;* Jinbo Wu, Jiaji Zhu, Haiyong Wang, Dennis Cai, and Jiesheng Wu, *Alibaba Group*

[1]

- ✗ Legacy devices
- ✗ Large amount of connections

Existing User-space TCP



mTCP: a Highly Scalable User-level TCP Stack for Multicore Systems

EunYoung Jeong, Shinae Woo, Muhammad Jamshed, and Haewon Jeong, *Korea Advanced Institute of Science and Technology (KAIST);* Sunghwan Ihm, *Princeton University;* Dongsu Han and KyoungSoo Park, *Korea Advanced Institute of Science and Technology (KAIST)*

[3]

IX: A Protected Dataplane Operating System for High Throughput and Low Latency

Adam Belay, *Stanford University;* George Prekas, *École Polytechnique Fédérale de Lausanne (EPFL);* Ana Klimovic, Samuel Grossman, and Christos Kozyrakis, *Stanford University;* Edouard Bugnion, *École Polytechnique Fédérale de Lausanne (EPFL)*

[2]

- ✗ Performance issues
- ✗ Compatibility

Solution: building our own user-space TCP - Luna

[1] NSDI'21, When Cloud Storage Meets RDMA
[2] NSDI'14, mTCP: A Highly Scalable User-level TCP Stack for Multicore Systems
[3] OSDI'14, IX: A Protected Dataplane Operating System for High Throughput and Low Latency

Background & Motivation

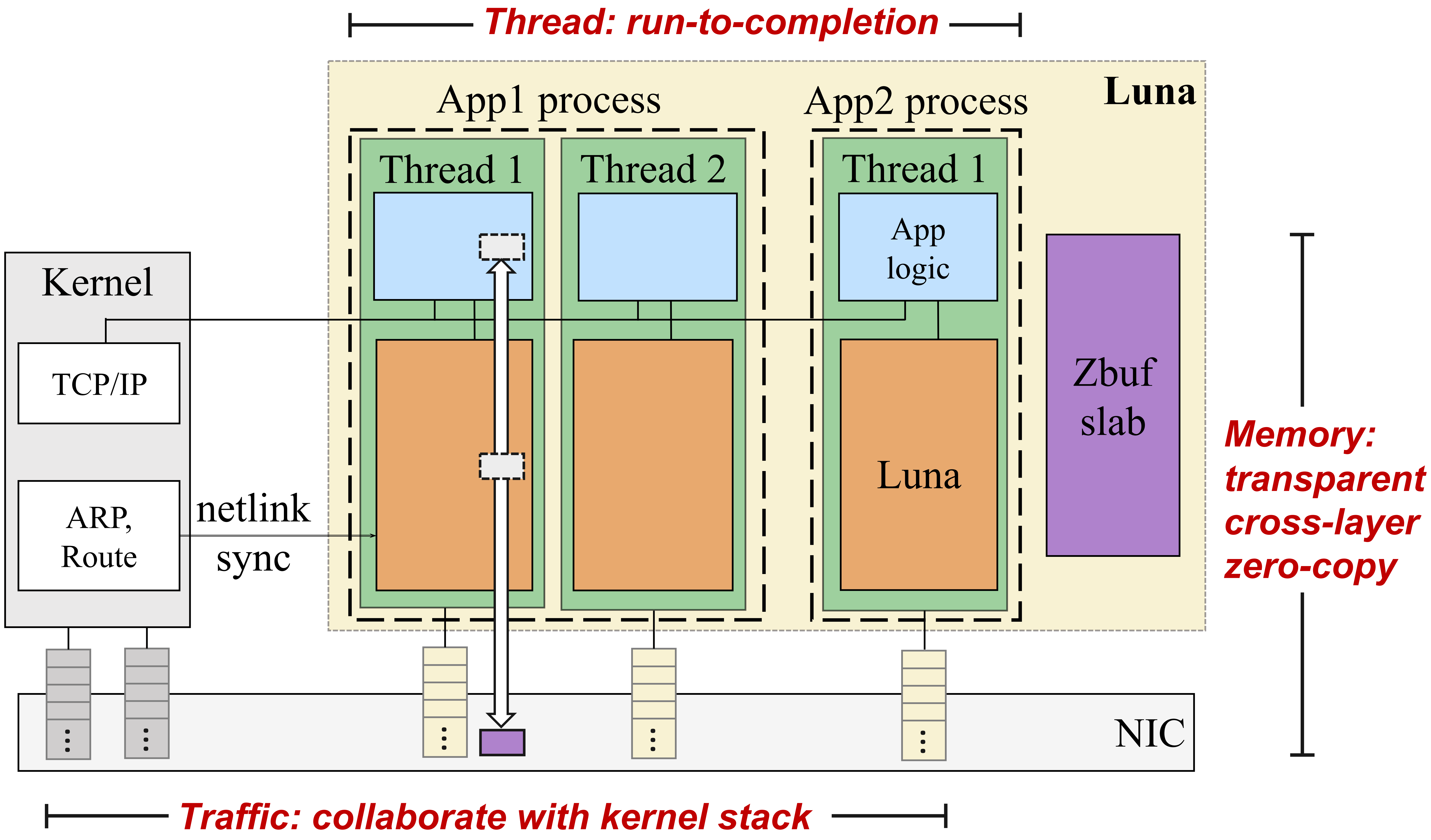
Design

Evaluation

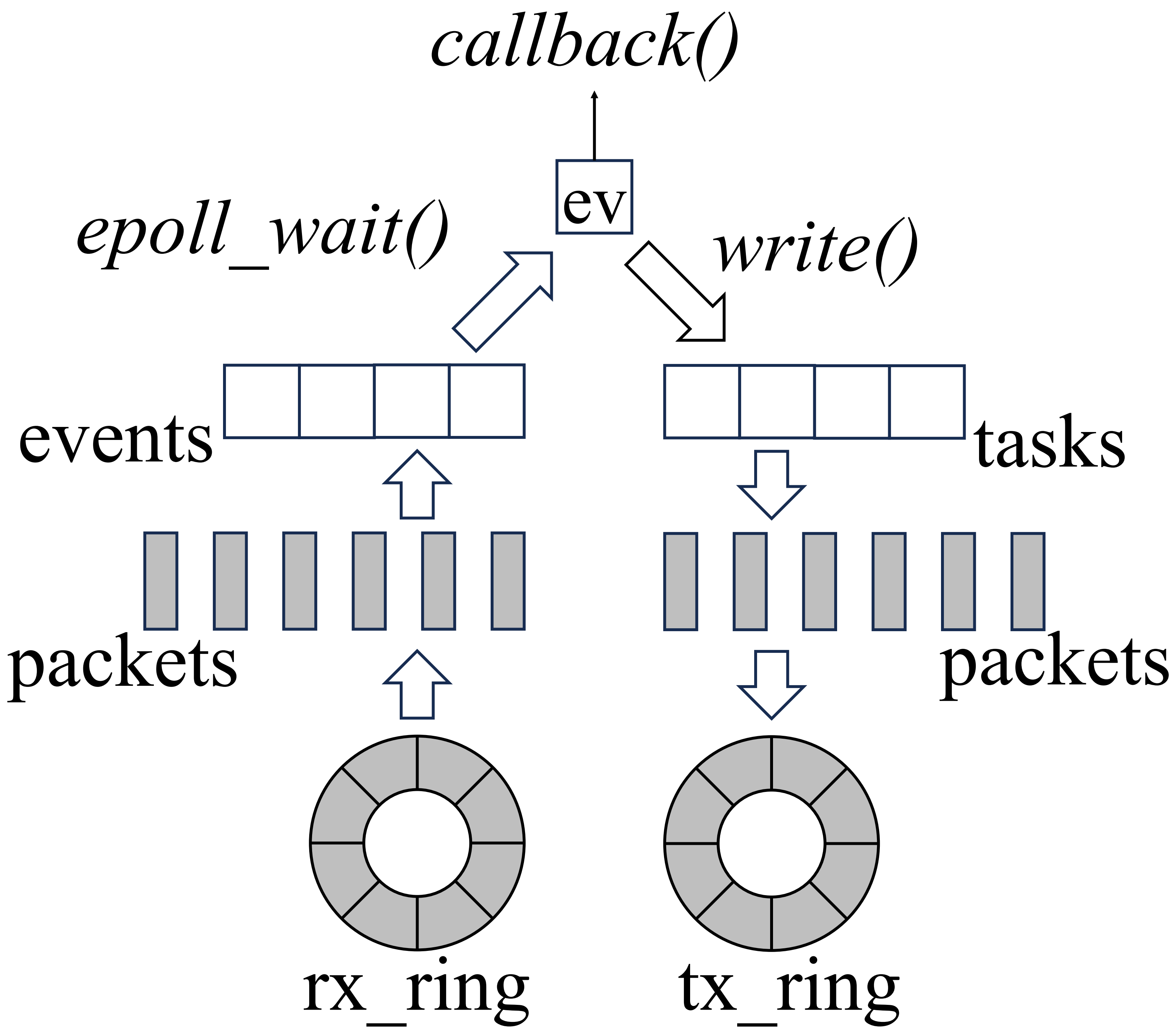
Discussion & Lessons

Conclusion

Luna Architecture

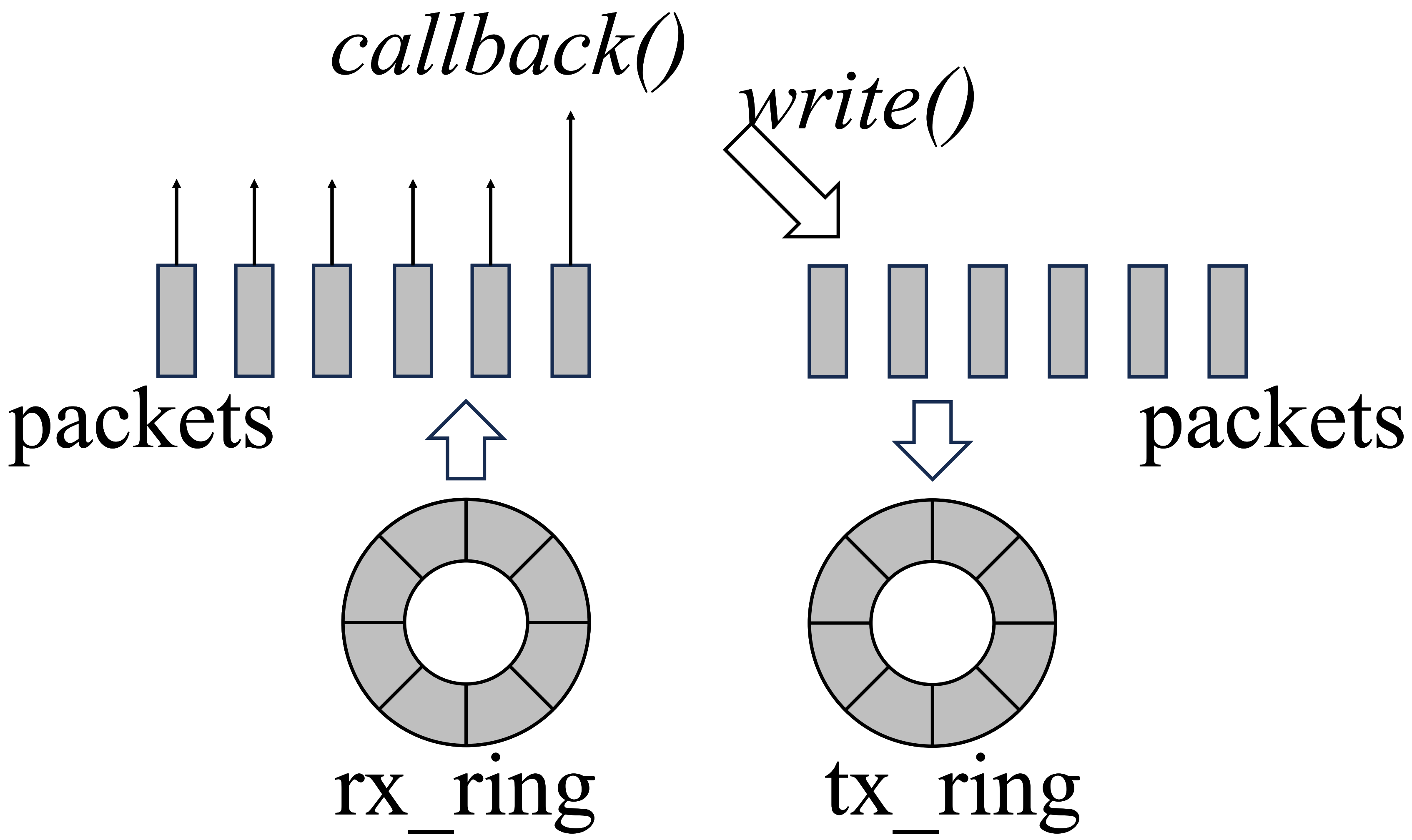


Thread Model: Run-to-completion



Batch-r2c

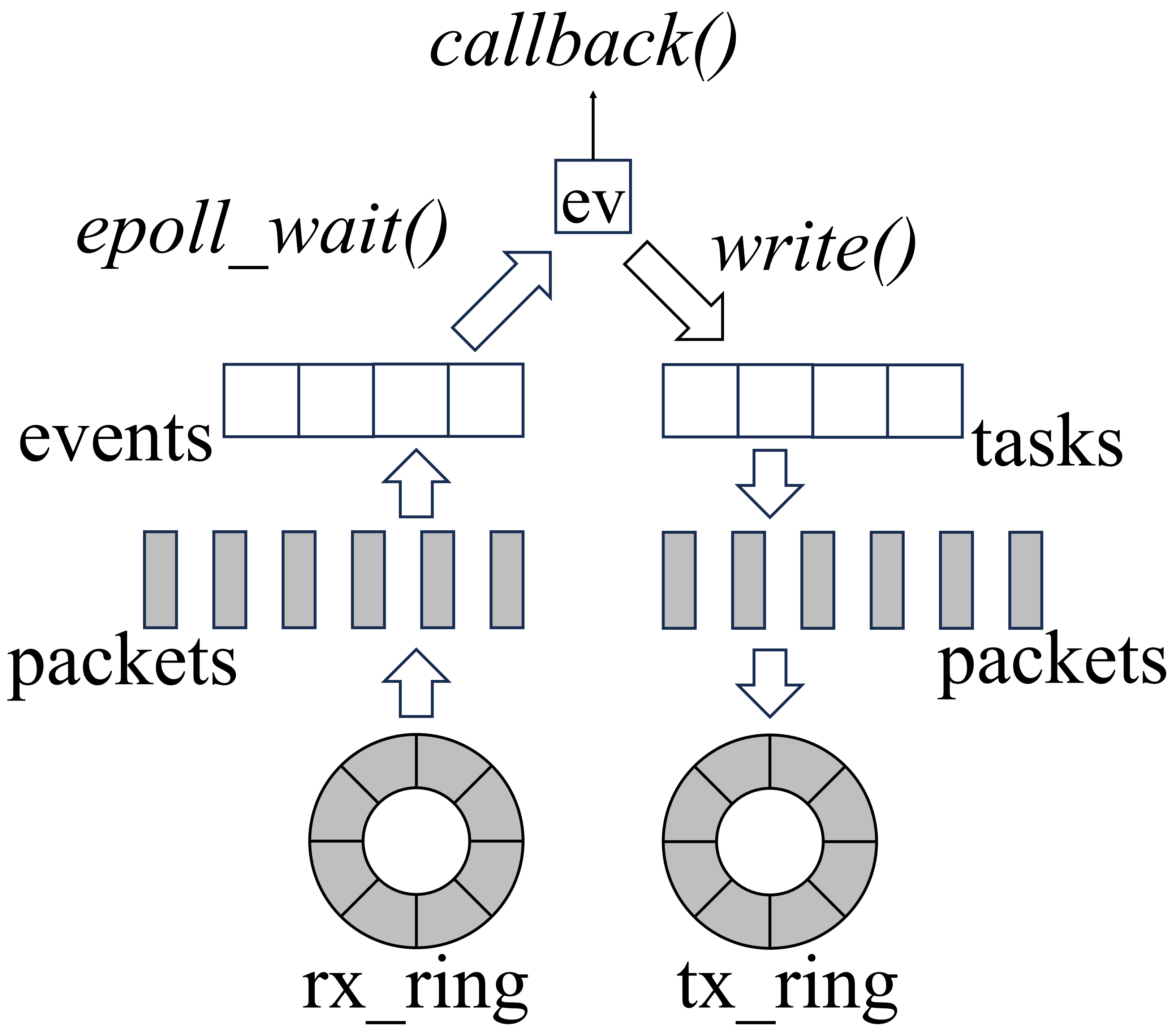
(similar to mTCP[1], IX[2])



Inline-r2c

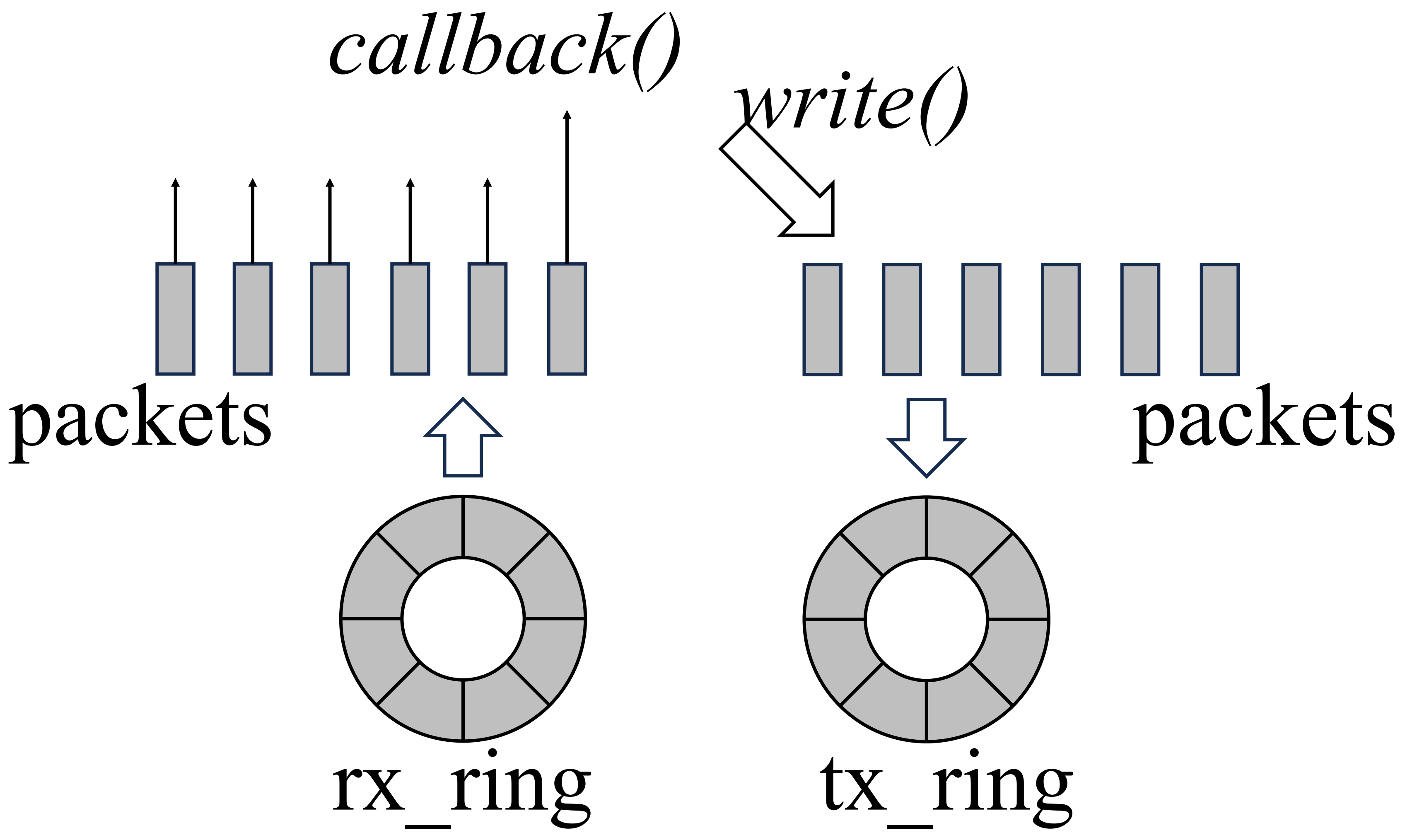
[1] NSDI'14, mTCP: A Highly Scalable User-level TCP Stack for Multicore Systems
[2] OSDI'14, IX: A Protected Dataplane Operating System for High Throughput and Low Latency

Thread Model: Run-to-completion



Batch-r2c

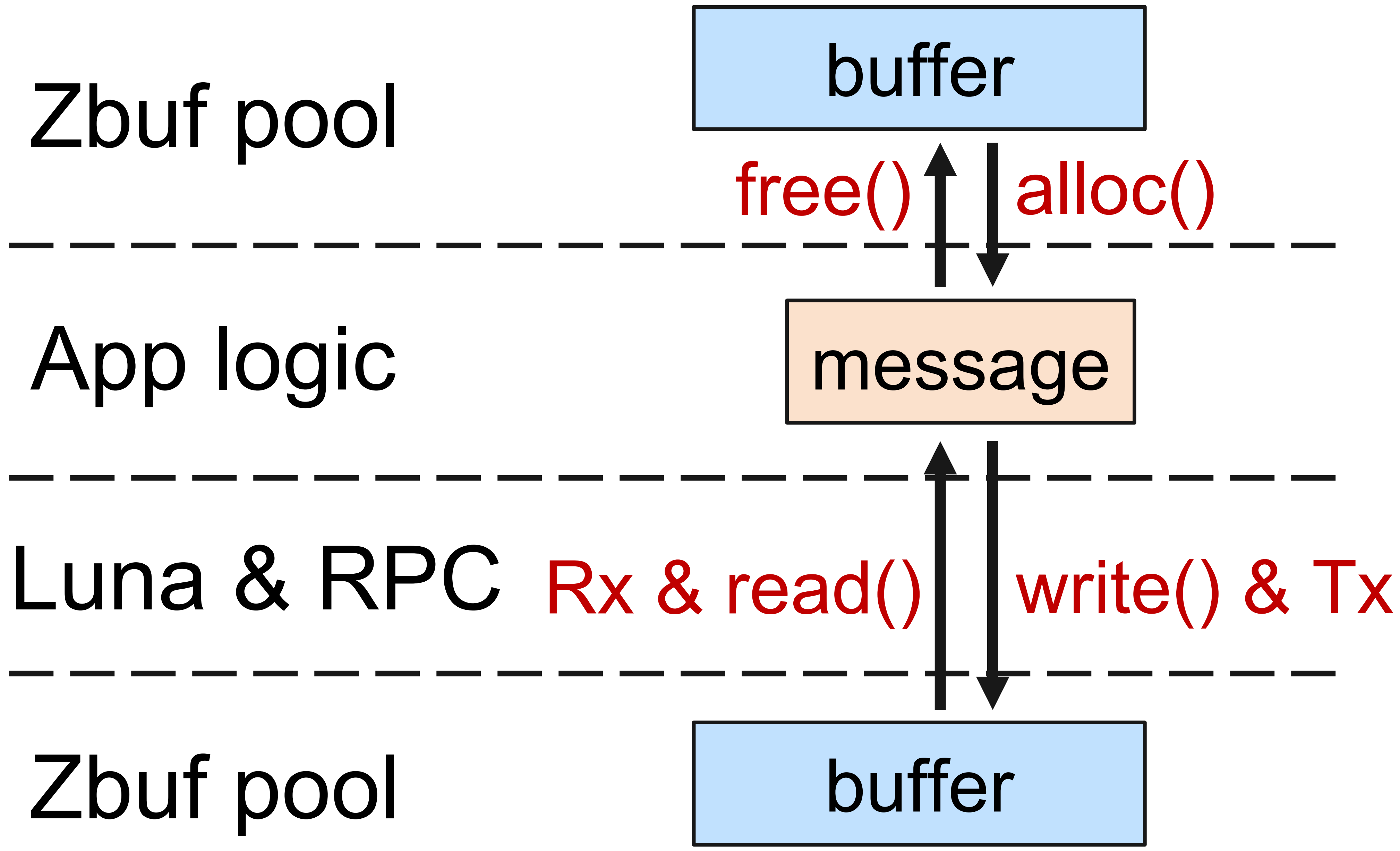
- ✓ Pros: Compatibility
- ✗ Cons: Event framework overhead

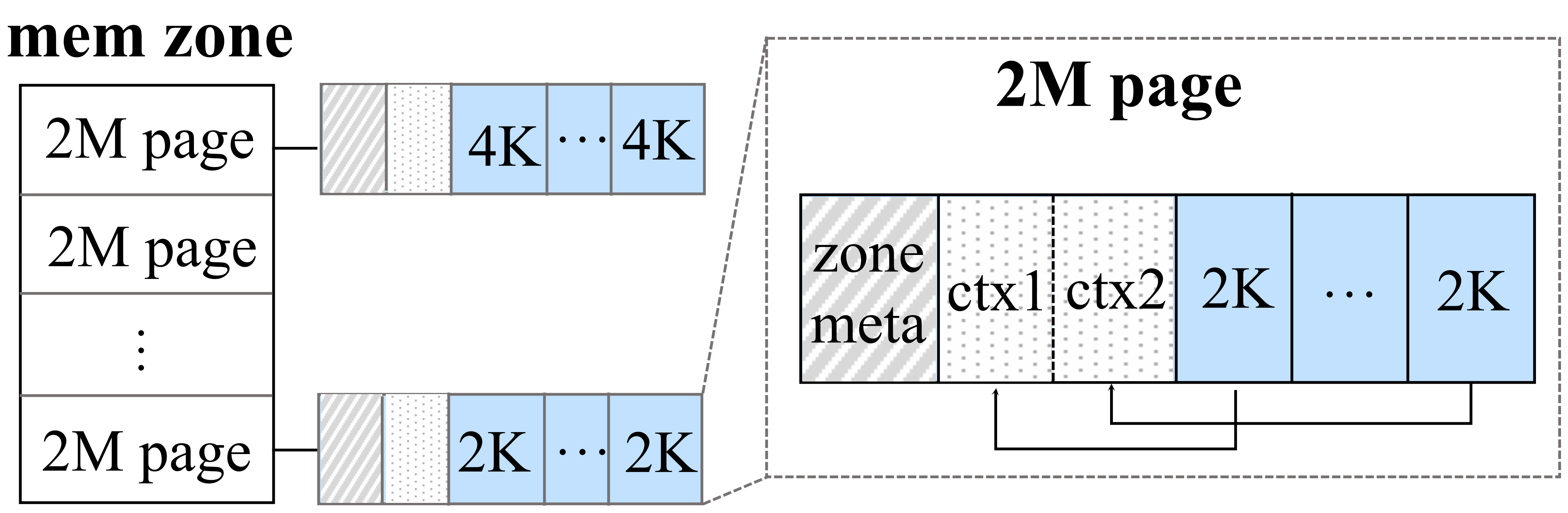


Inline-r2c

- ✓ Pros: Performance
- ✗ Cons: Programing model change

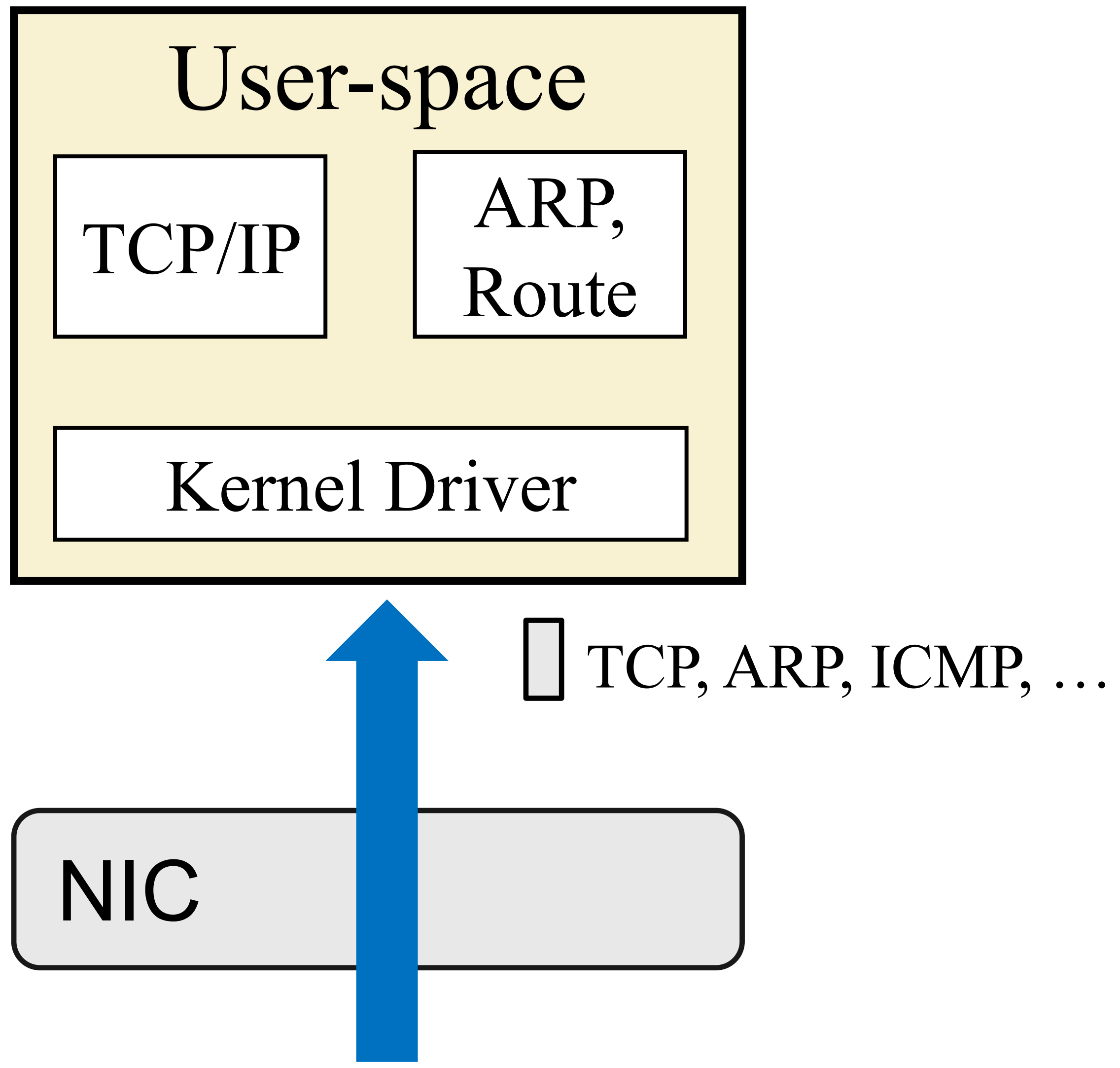
Memory Model: Cross-layer Zero-copy





Transparent buffer management —— ***just like heap*** 😊

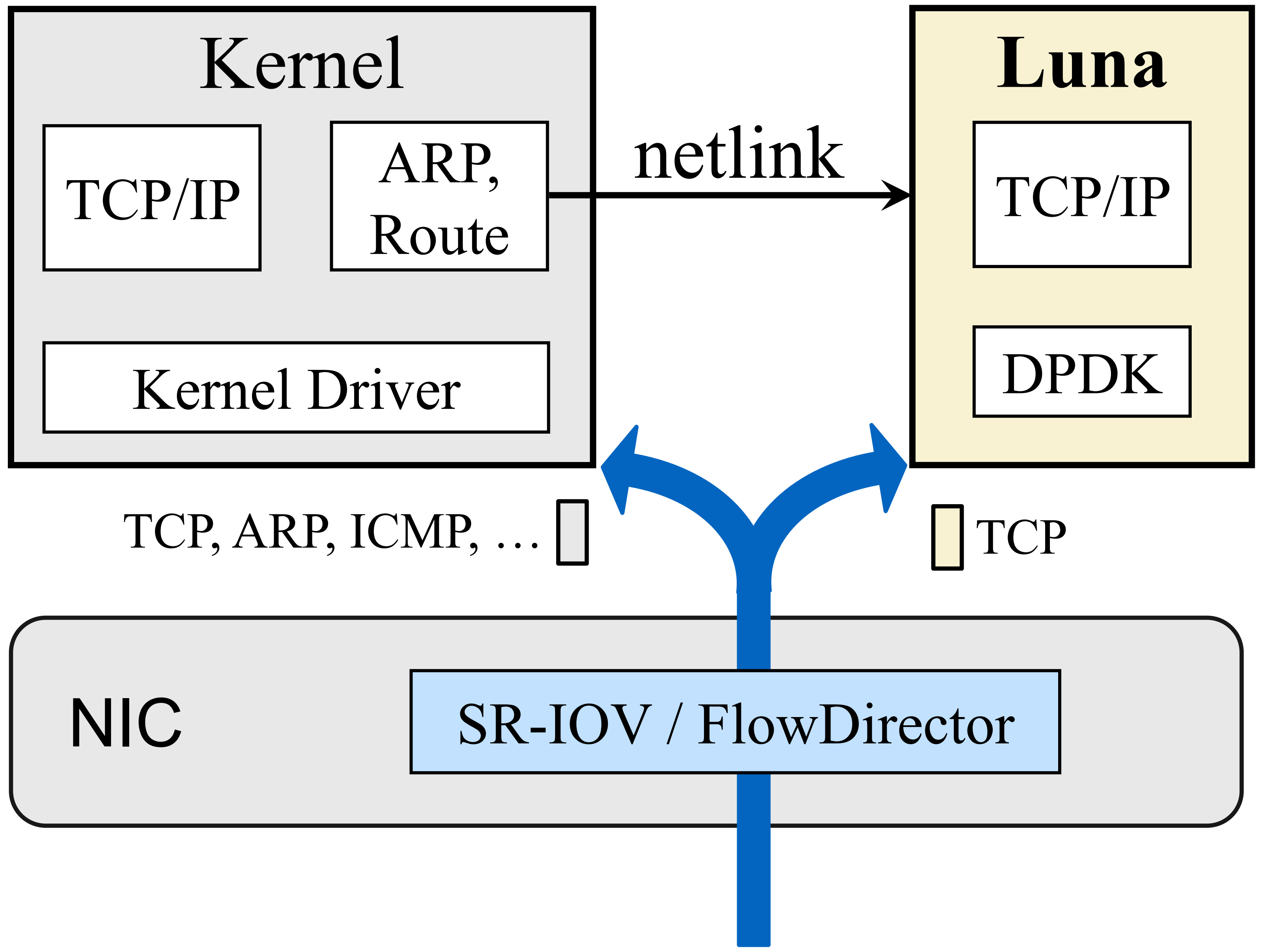
Traffic Model: Collaborate with Kernel Stack



Other user-space TCP

✗ Exclusive with kernel stack Apps

✗ Complex control plane logic and configuration



Luna

✓ Cohost with other Apps

✓ Get rid of complex control plane implementation

Background & Motivation

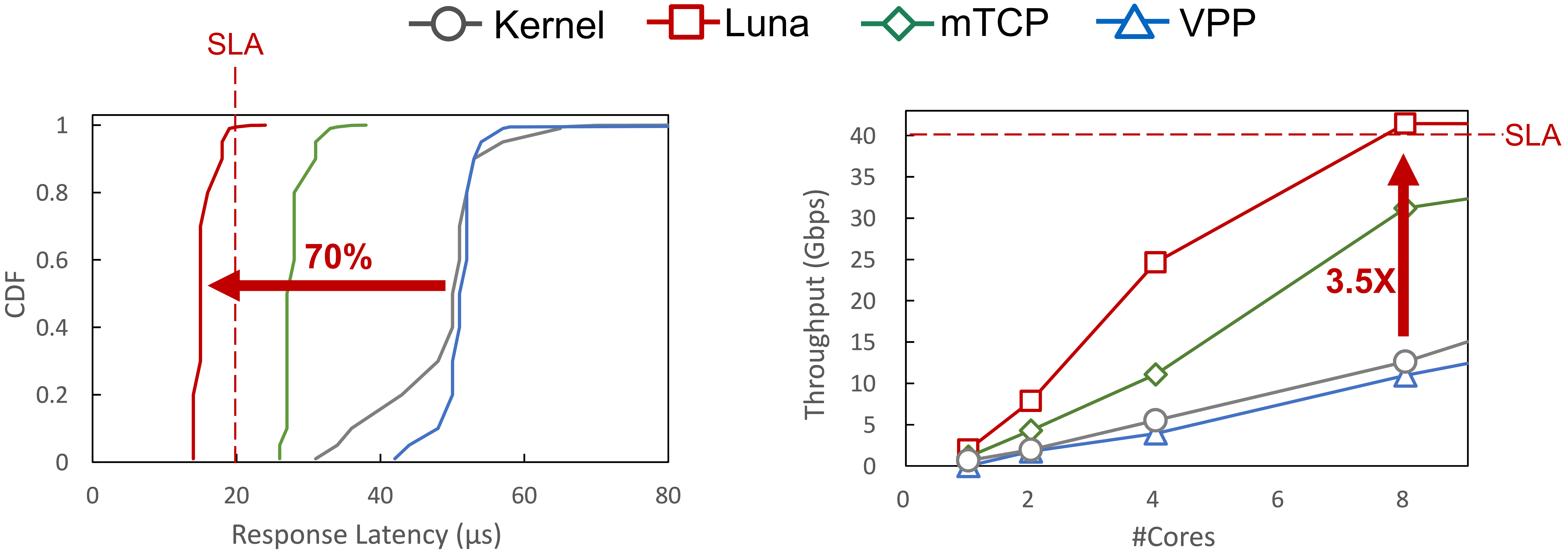
Design

Evaluation

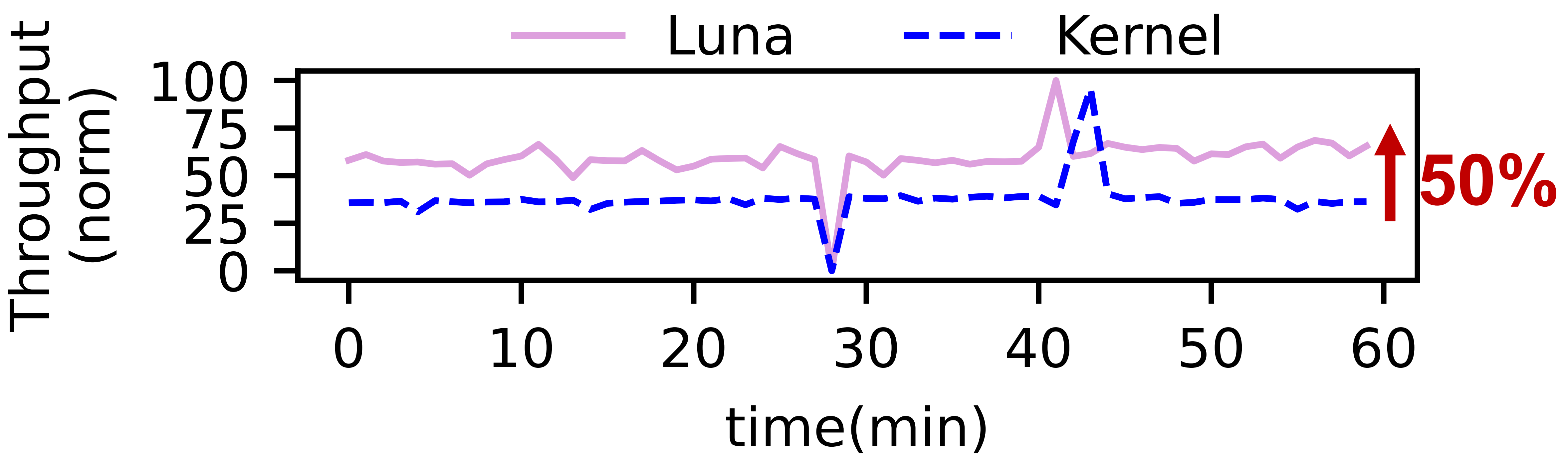
Discussion & Lessons

Conclusion

Evaluation - Microbenchamrk



Evaluation - in the field



Elastic Block Storage

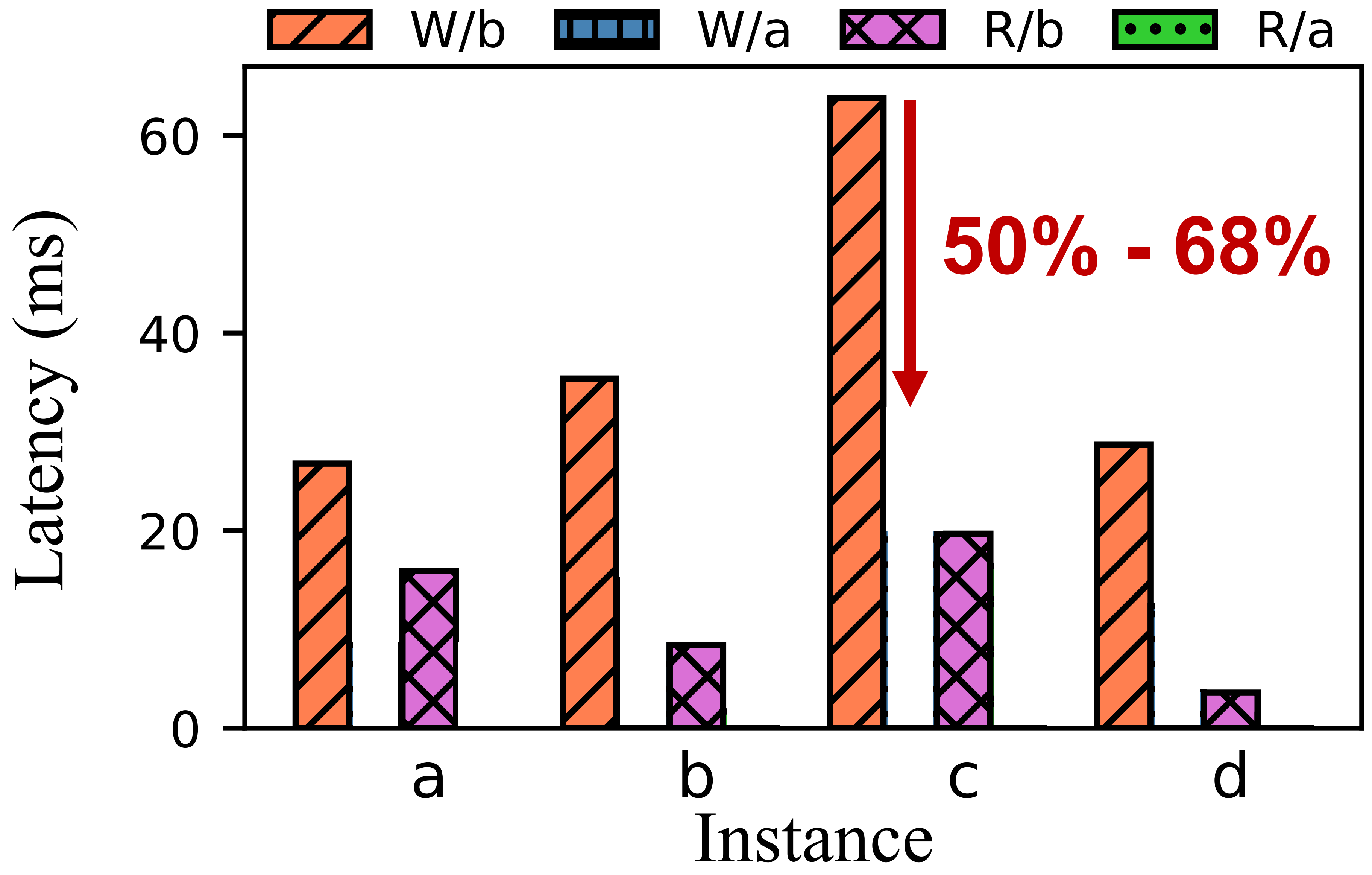


Table Store Service

Background & Motivation

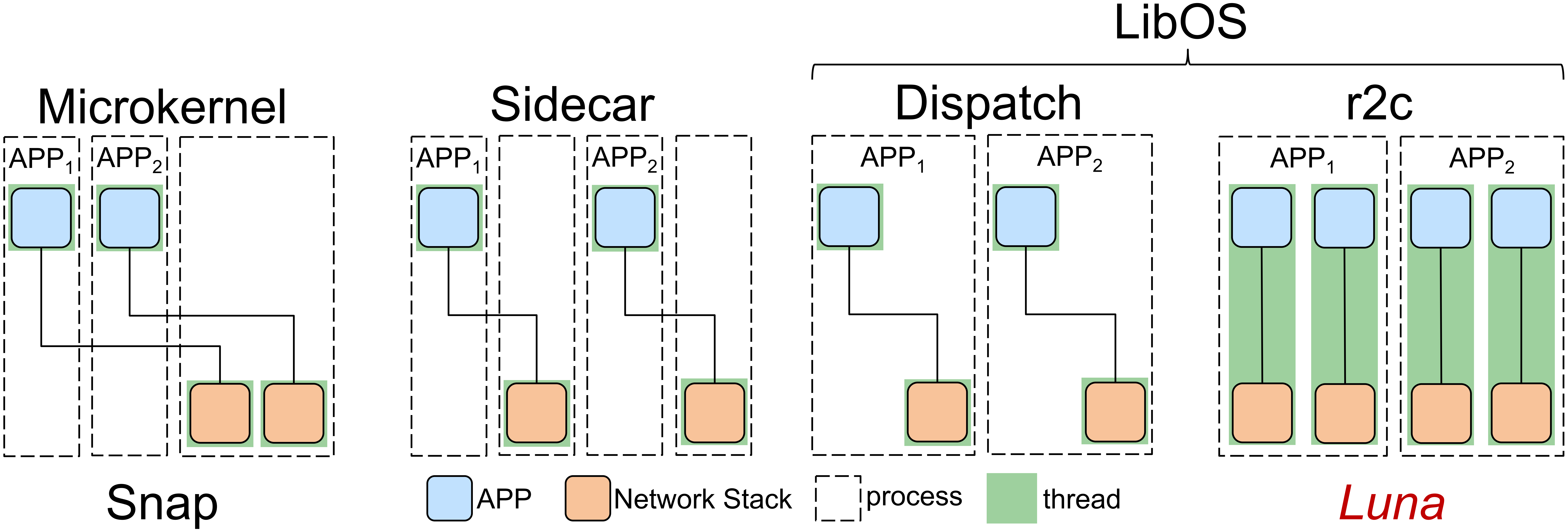
Design

Evaluation

Discussion & Lessons

Conclusion

Design Choices: LibOS with R2C



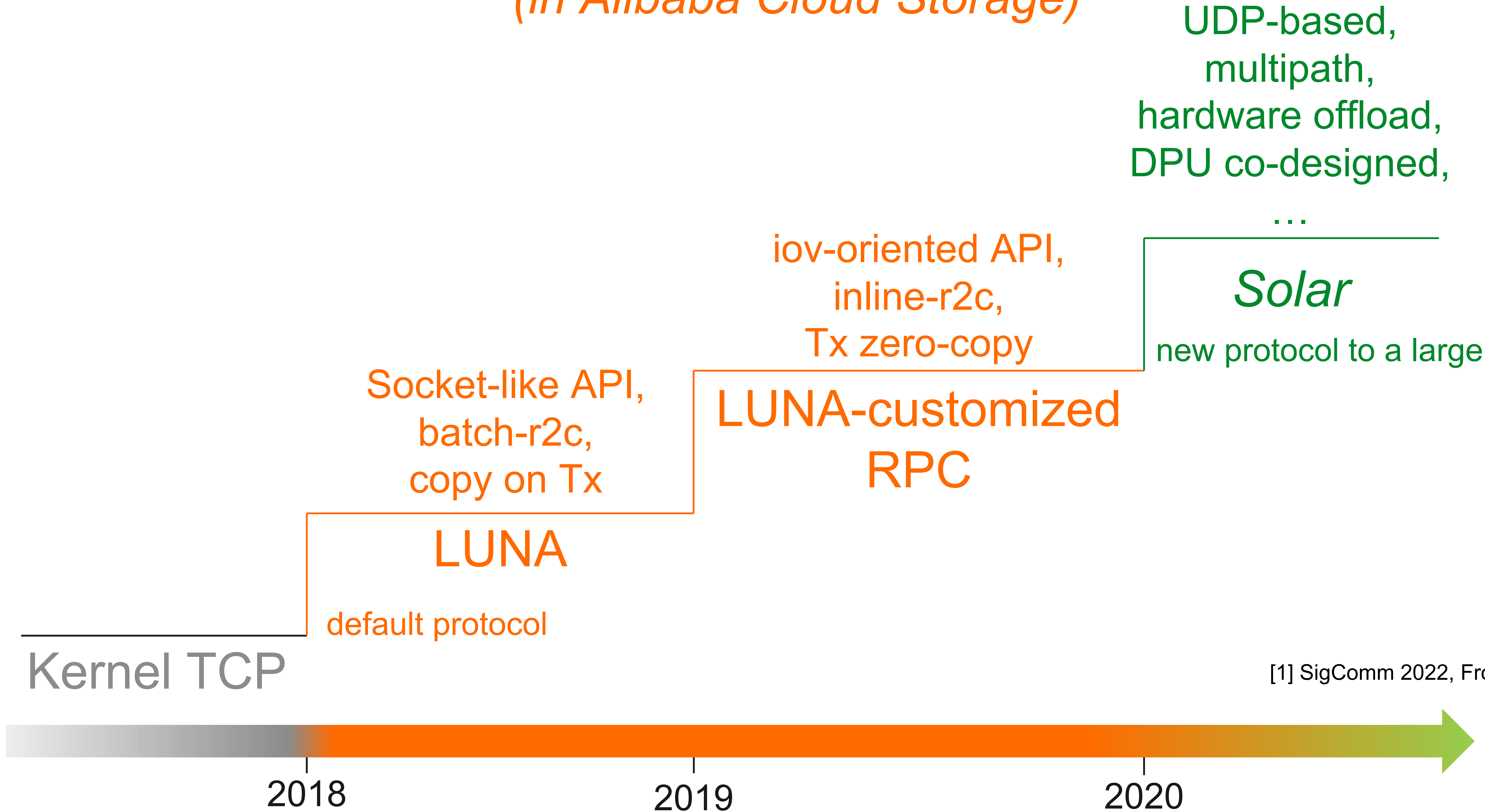
- ➢ Performance is the most critical
- ➢ No frequent upgrades

Design Choices: Customized TCP

New transport protocols	pFabric, pHost, …	High-performance, but complex
Customized TCP	Simplified fast path NewReno Fast recovery	Simple, but works

Network Evolution

(in Alibaba Cloud Storage)



Kernel TCP

LUNA

default protocol

Socket-like API,
batch-r2c,
copy on Tx

LUNA-customized RPC

iov-oriented API,
inline-r2c,
Tx zero-copy

Solar

new protocol to a larger scale

UDP-based,
multipath,
hardware offload,
DPU co-designed,
…

[1] SigComm 2022, From luna to solar

Background & Motivation

Design

Evaluation

Discussion & Lessons

Conclusion

Conclusion

- Luna, a user-space network designed for cloud storage service
- Network architecture of Alibaba Cloud Storage
- Three design pieces
  - Thread model: run-to-completion
  - Memory model: cross-layer zero-copy
  - Traffic model: collaborate with kernel stack
- Discussion
  - LibOS vs. Microkernel
  - TCP and tailoring
  - Evolution

Alibaba Cloud
Worldwide Cloud Services Partner

Thanks for Listening!
Q&A

Email: mashu.ms@alibaba-inc.com