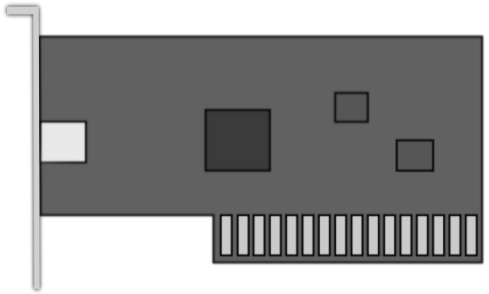
12th ANNUAL WORKSHOP 2016

FLASHNET

A Unified High-Performance IO Stack

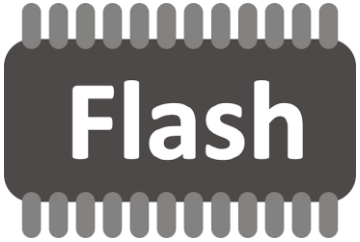
Animesh Trivedi, Nikolas Ioannou, Bernard Metzler, Patrick Stuedi,
Jonas Pfefferle, Ioannis Koltsidas
IBM Zurich Research

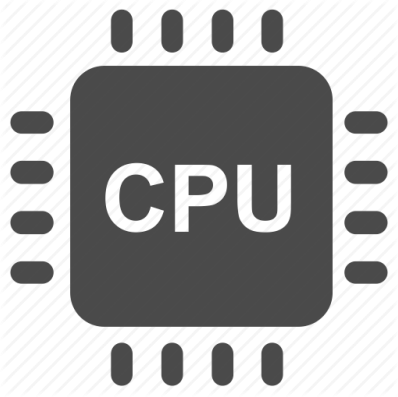
HIGH-PERFORMANCE IO



1 → 100 Gbit/sec, with ~1 usec link latencies

Rise of NVM devices, multi GBs/sec with ~10s usec device latencies



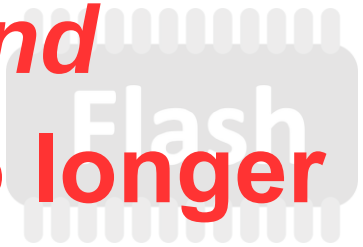


Marginal improvements

HIGH-PERFORMANCE IO

1 → 100 Gbit/sec, with ~1 usec link latencies

Rise of NVM devices, multi GBs/sec with ~10s usec device latencies



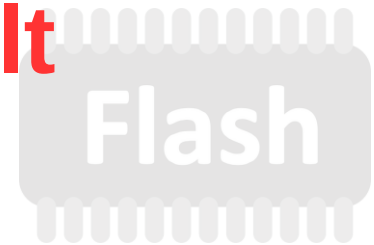
The notion of "*fast CPU and multiple slow IO devices*" is no longer valid

Marginal improvements

HIGH-PERFORMANCE IO

1 → 100 Gbit/sec, with ~1 usec link latencies

Rise of NVM devices, multi GBs/sec with ~10s usec device latencies



Traditional IO stacks built assuming slow IO and fail to deliver performance

Marginal improvements

WHY UNIFY NETWORK AND STORAGE IO?

- **Exposing high-speed networking performance to the user application:**
  - *Polling, direct hardware access, OS-bypass, zero-copy data movement, RDMA, DPDK…*
- **Exposing NVM device performance to the user application:**
  - *Polling, direct hardware access, OS-bypass,zero-copy data movement, NVMe, SPDK...*

Proposal : Unify network and storage IO

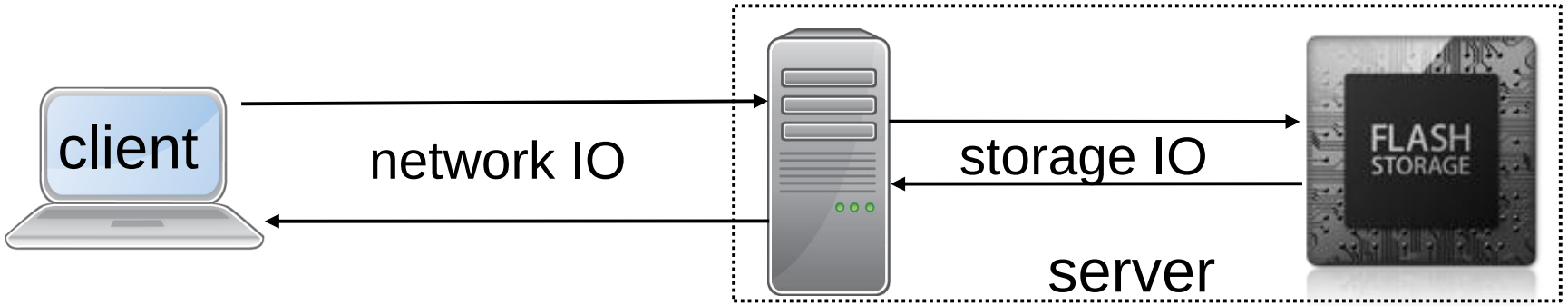
→ FlashNet !

THE PROBLEM SCENARIO

Key-Value Stores, Distributed Overlay File Systems e.g., Hadoop.

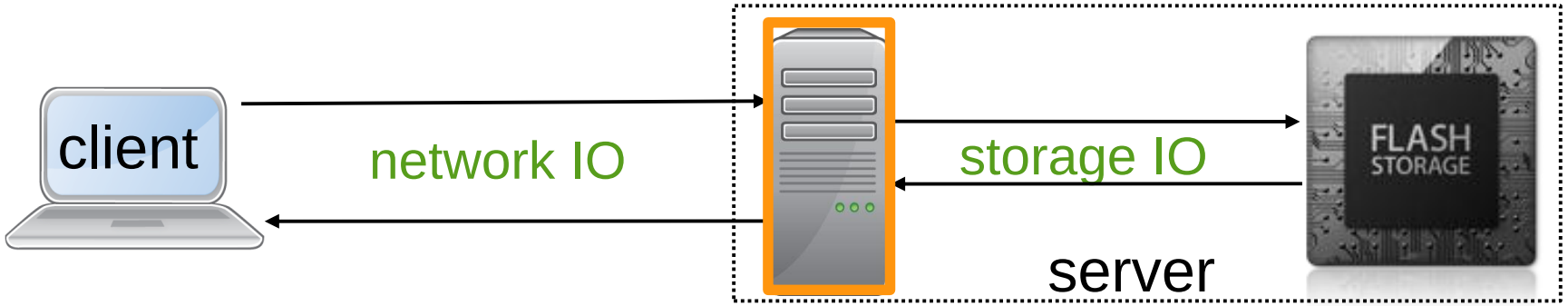
THE PROBLEM SCENARIO

Key-Value Stores, Distributed Overlay File Systems e.g., Hadoop.



THE PROBLEM SCENARIO

Key-Value Stores, Distributed Overlay File Systems e.g., Hadoop.



performance = network IO + storage IO + server time

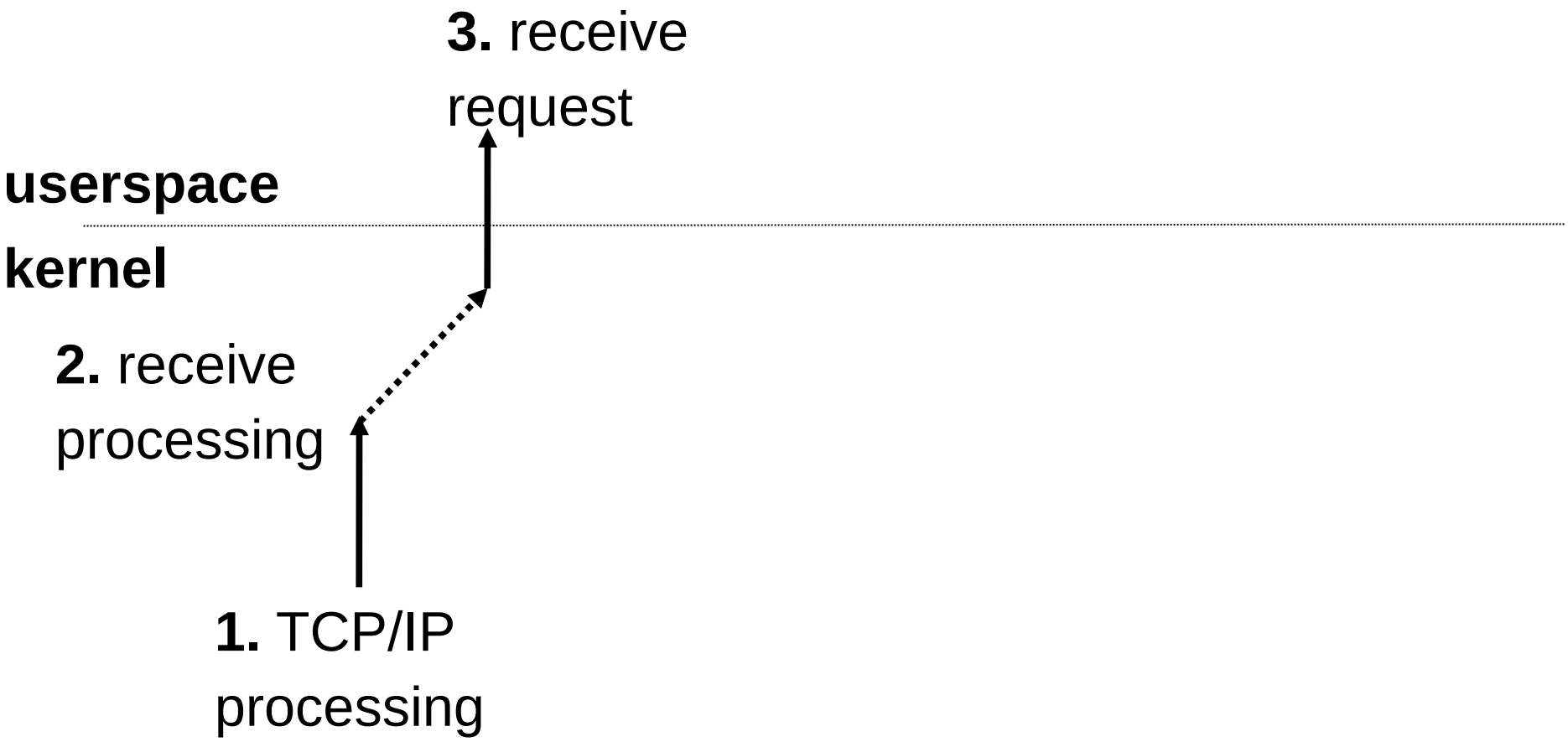
SERVER SIDE - A DETAILED LOOK: SEND

userspace

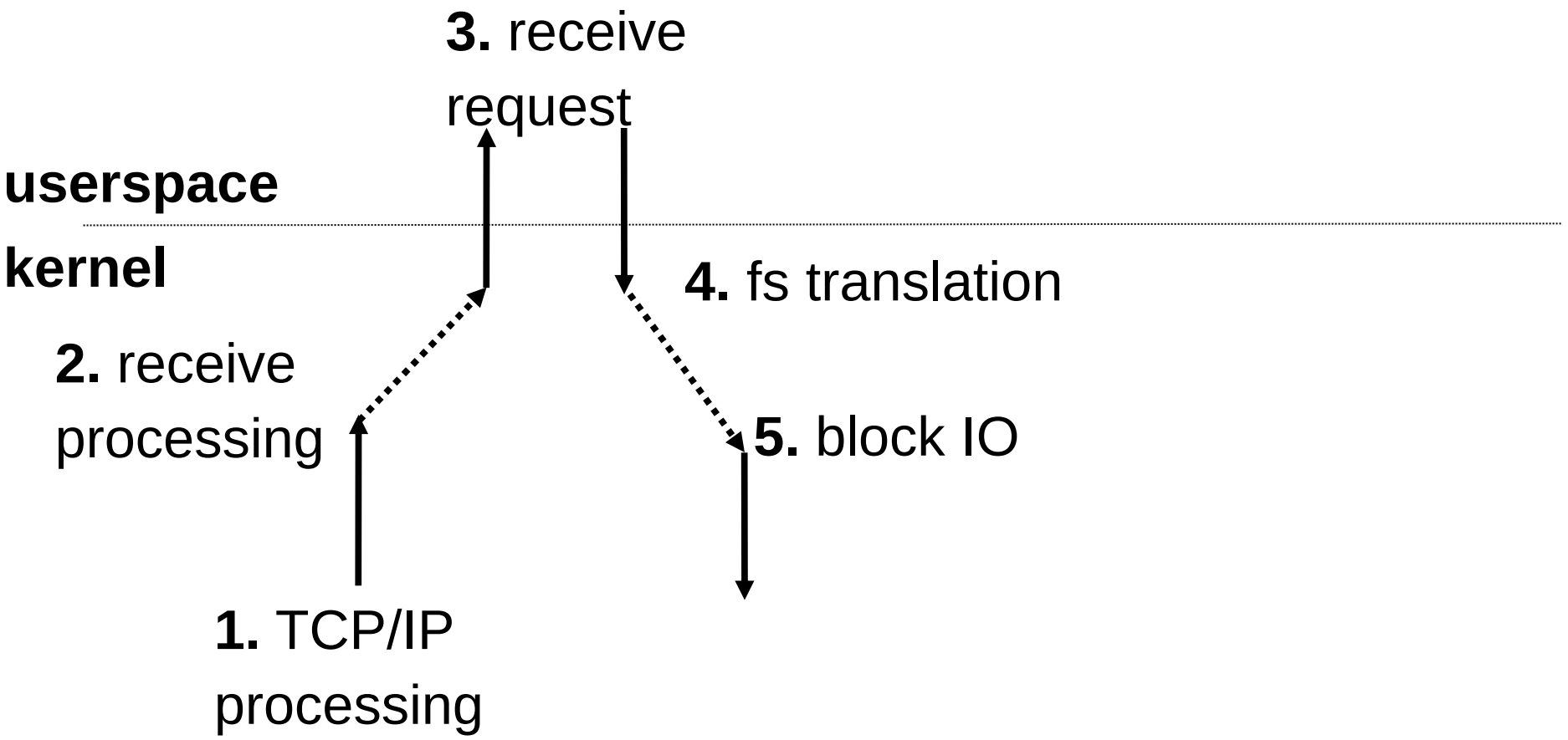
kernel

1. TCP/IP processing

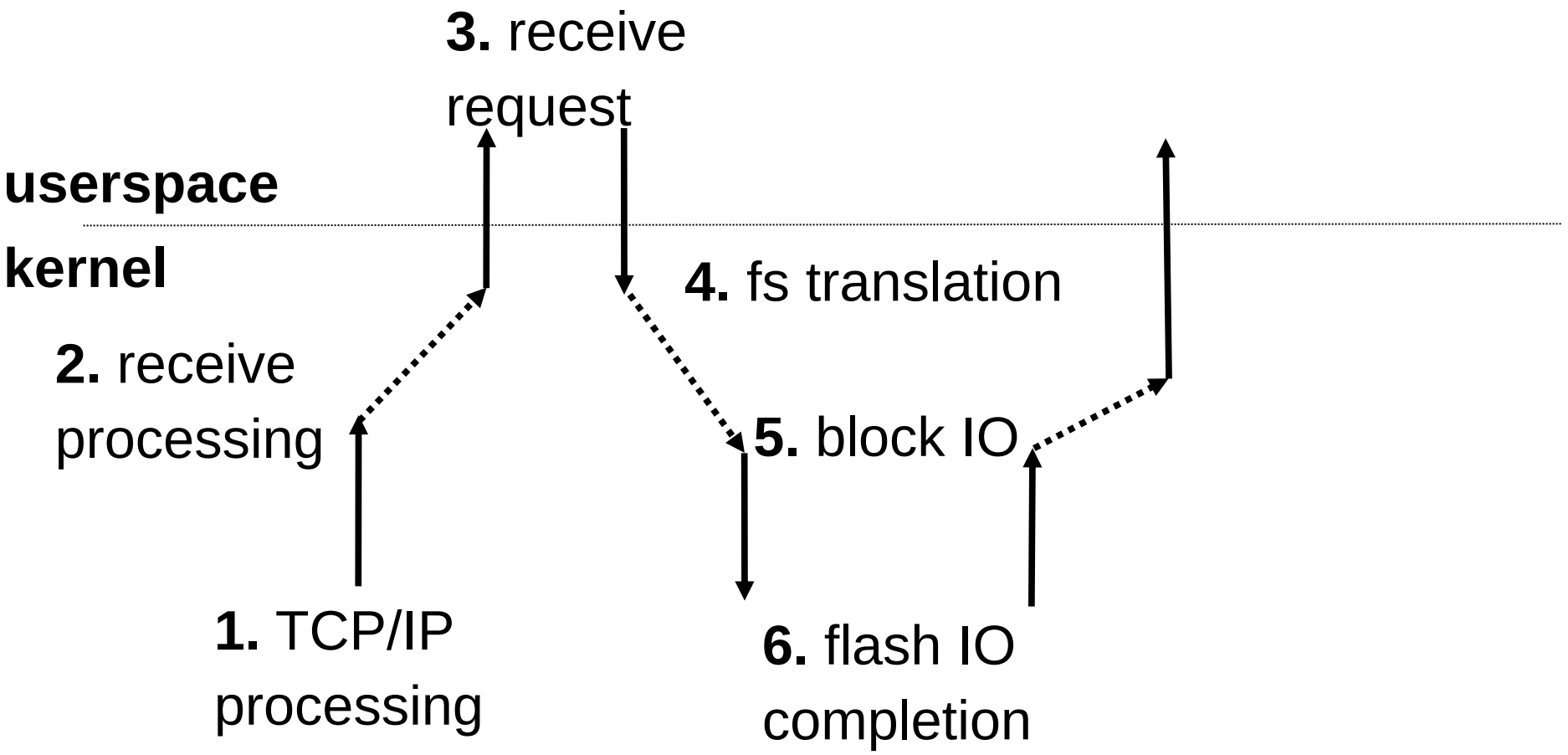
SERVER SIDE - A DETAILED LOOK: SEND



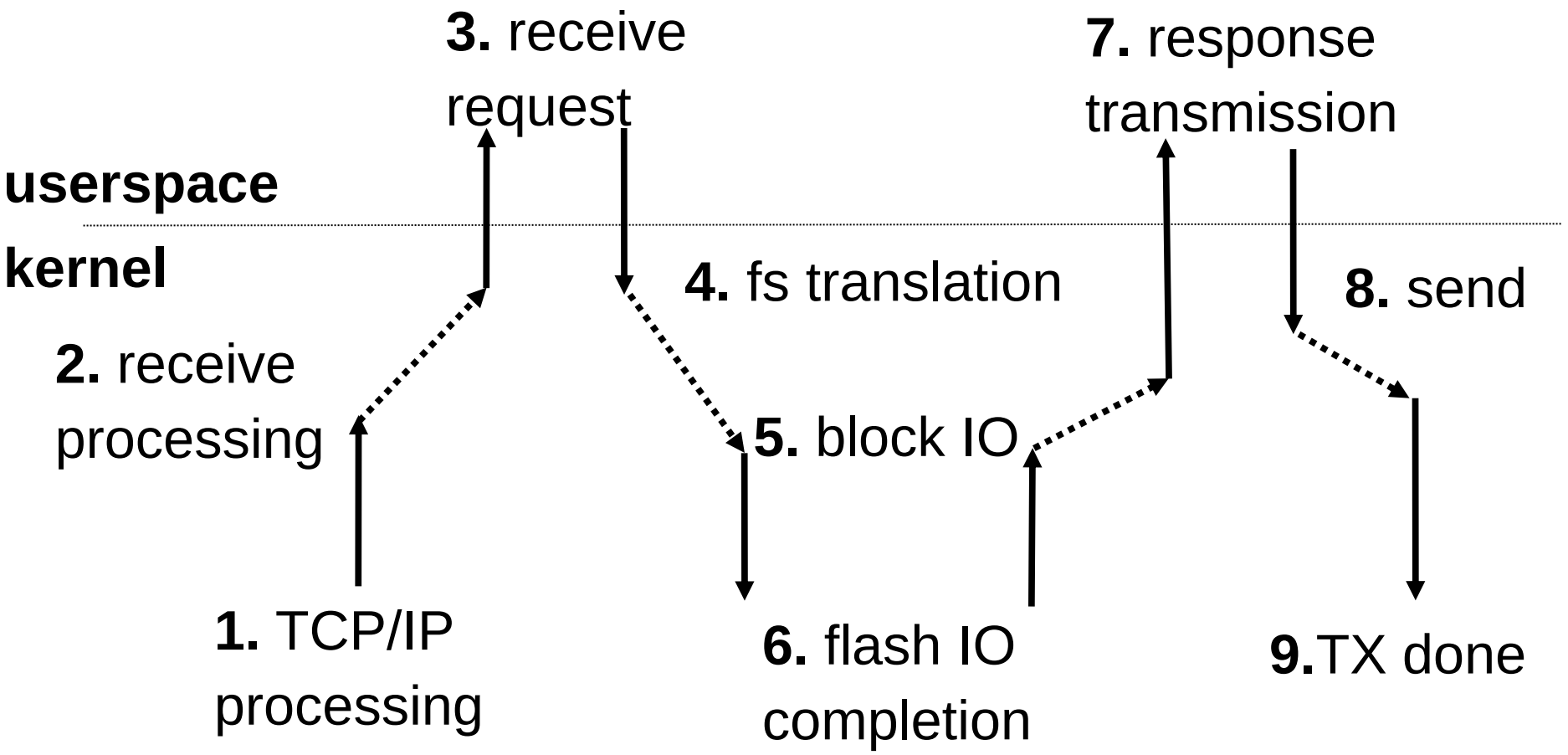
SERVER SIDE - A DETAILED LOOK: SEND



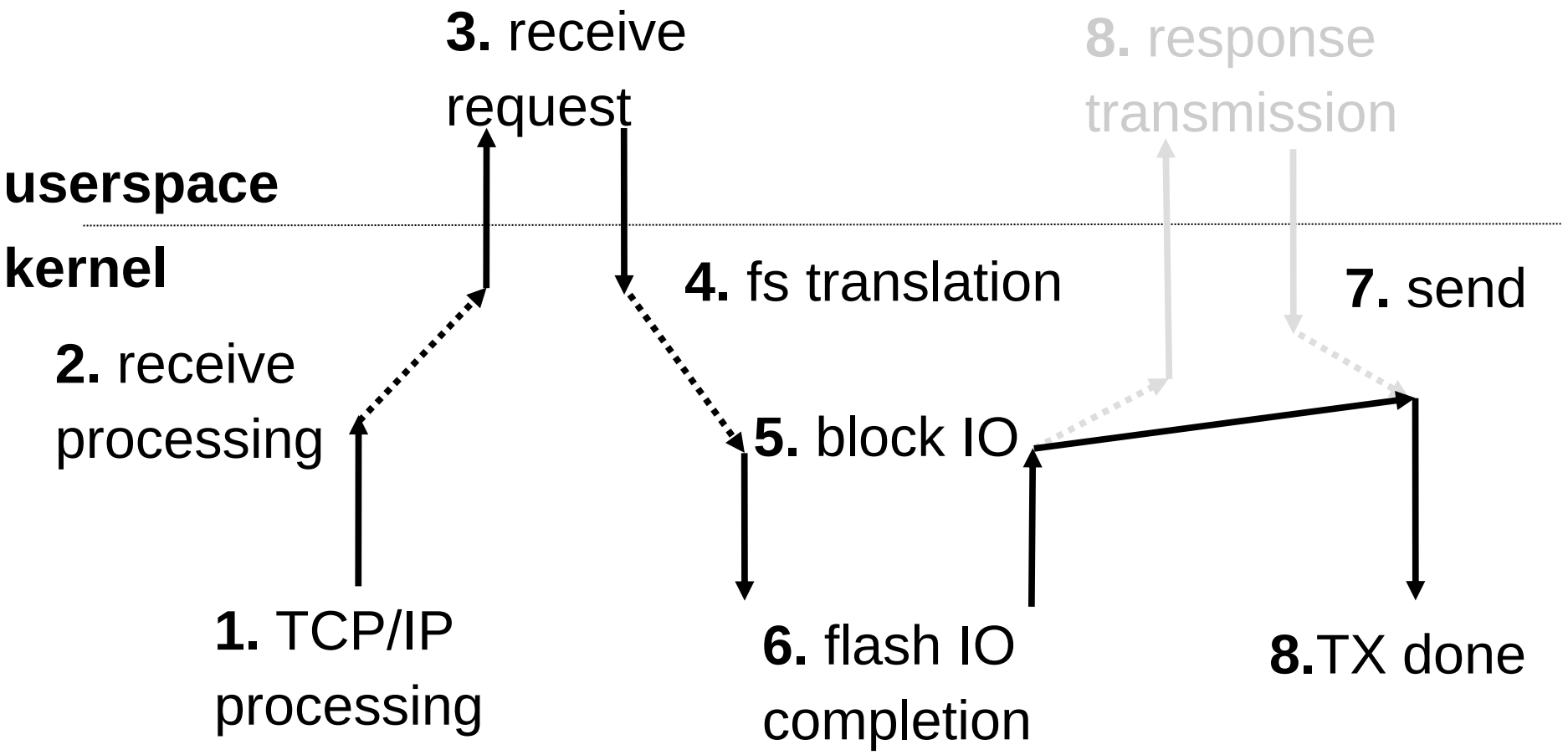
SERVER SIDE - A DETAILED LOOK: SEND



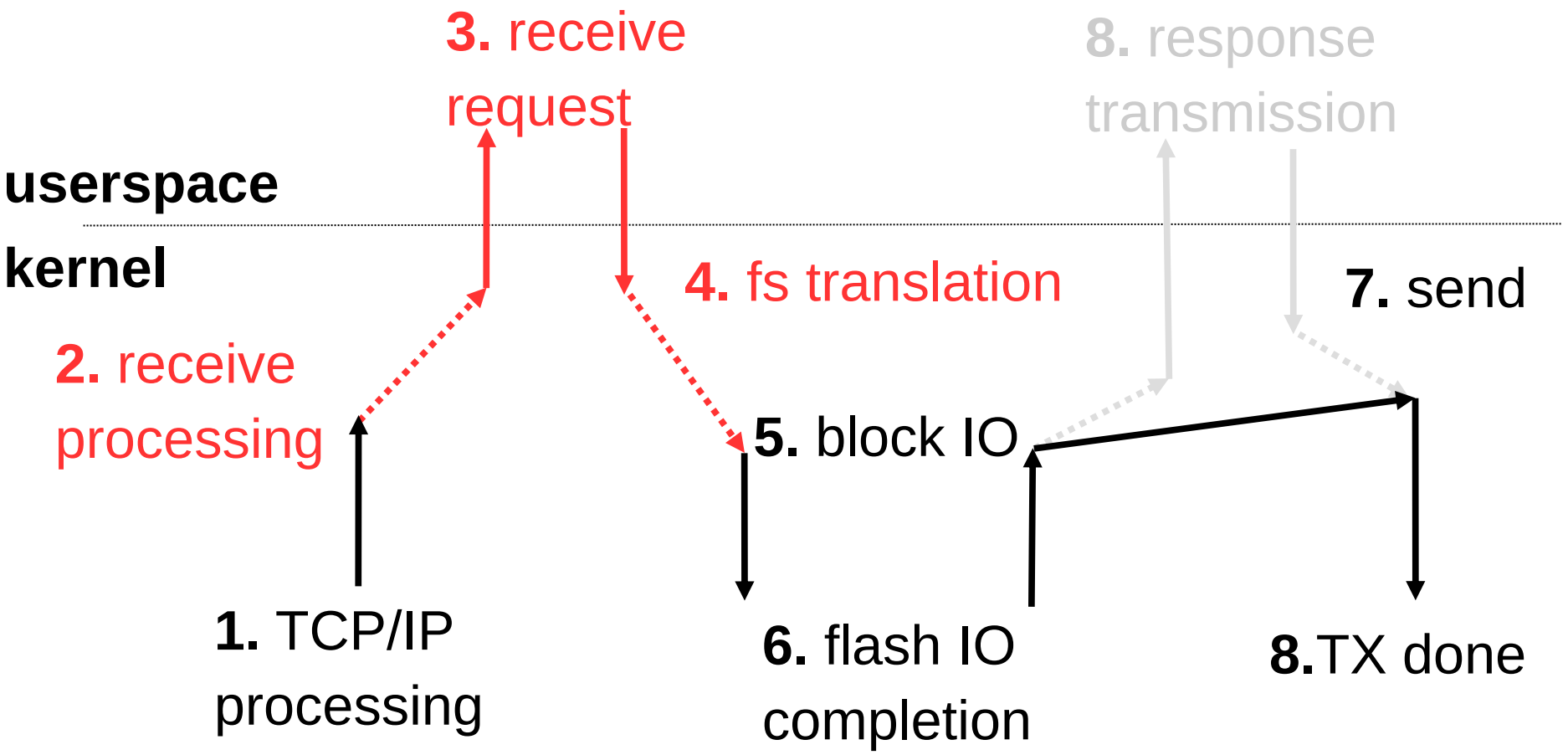
SERVER SIDE - A DETAILED LOOK: SEND



SERVER SIDE - A DETAILED LOOK: SENDFILE

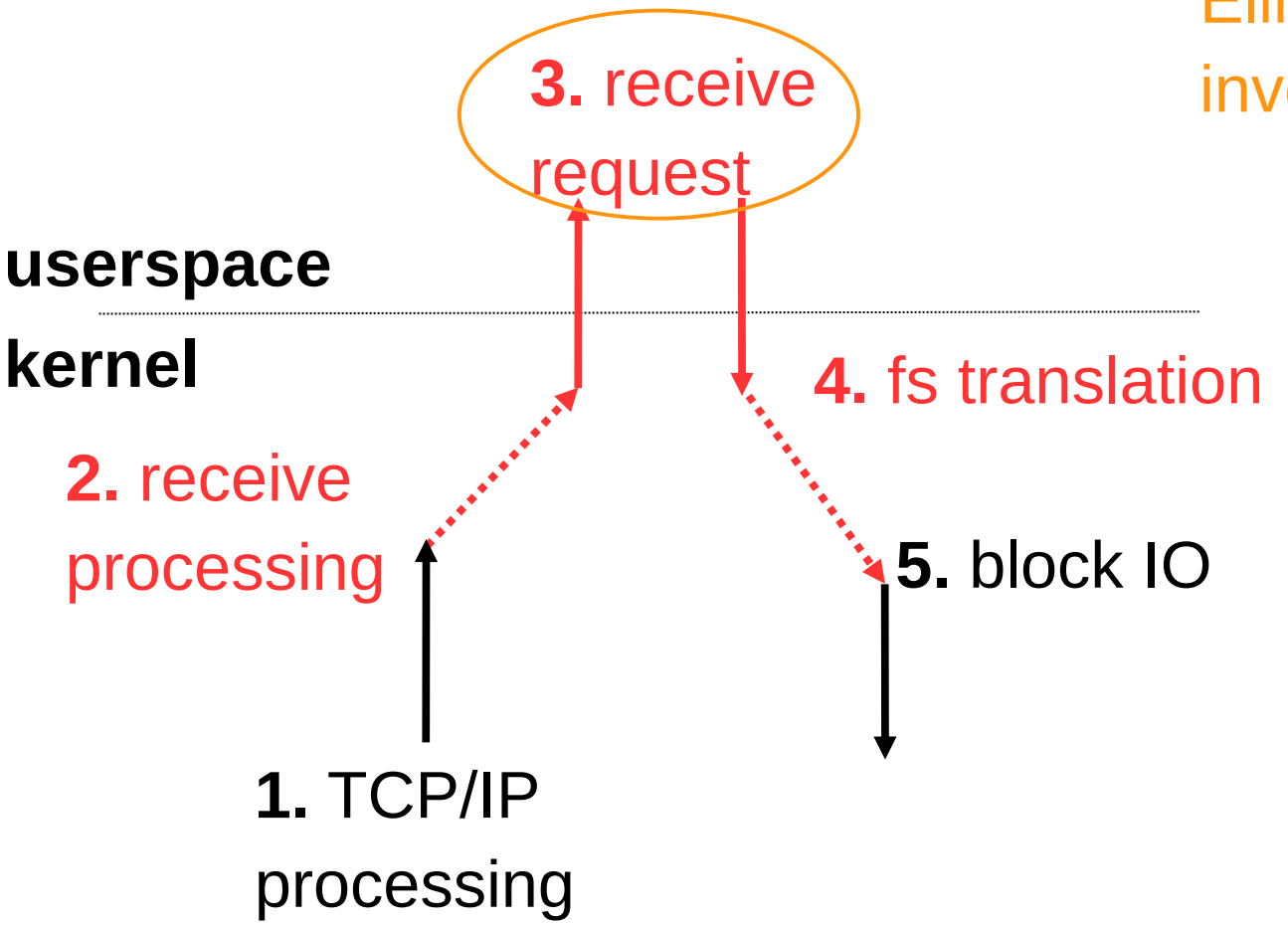


SERVER SIDE - A DETAILED LOOK: SENDFILE

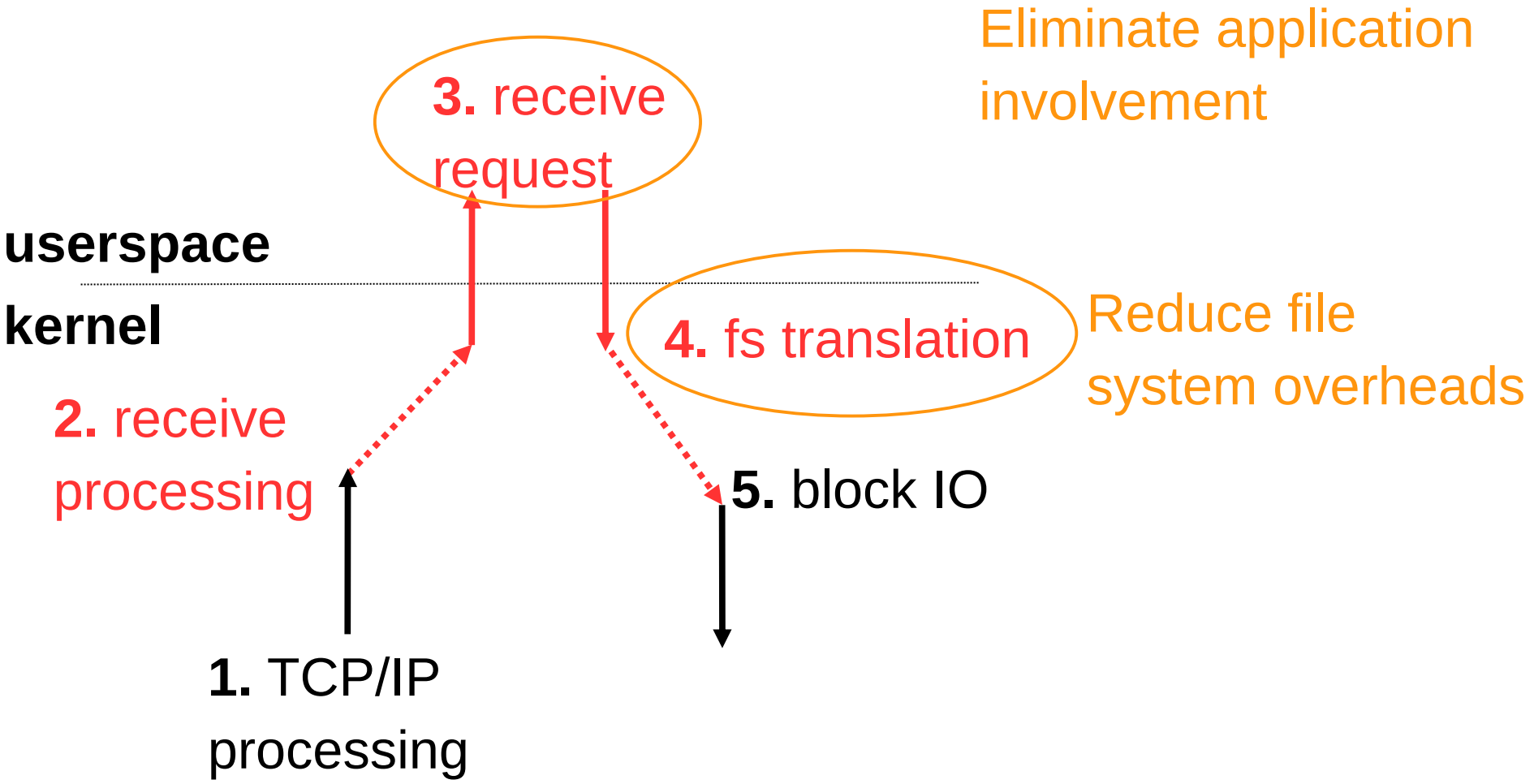


THE FLASHNET APPROACH

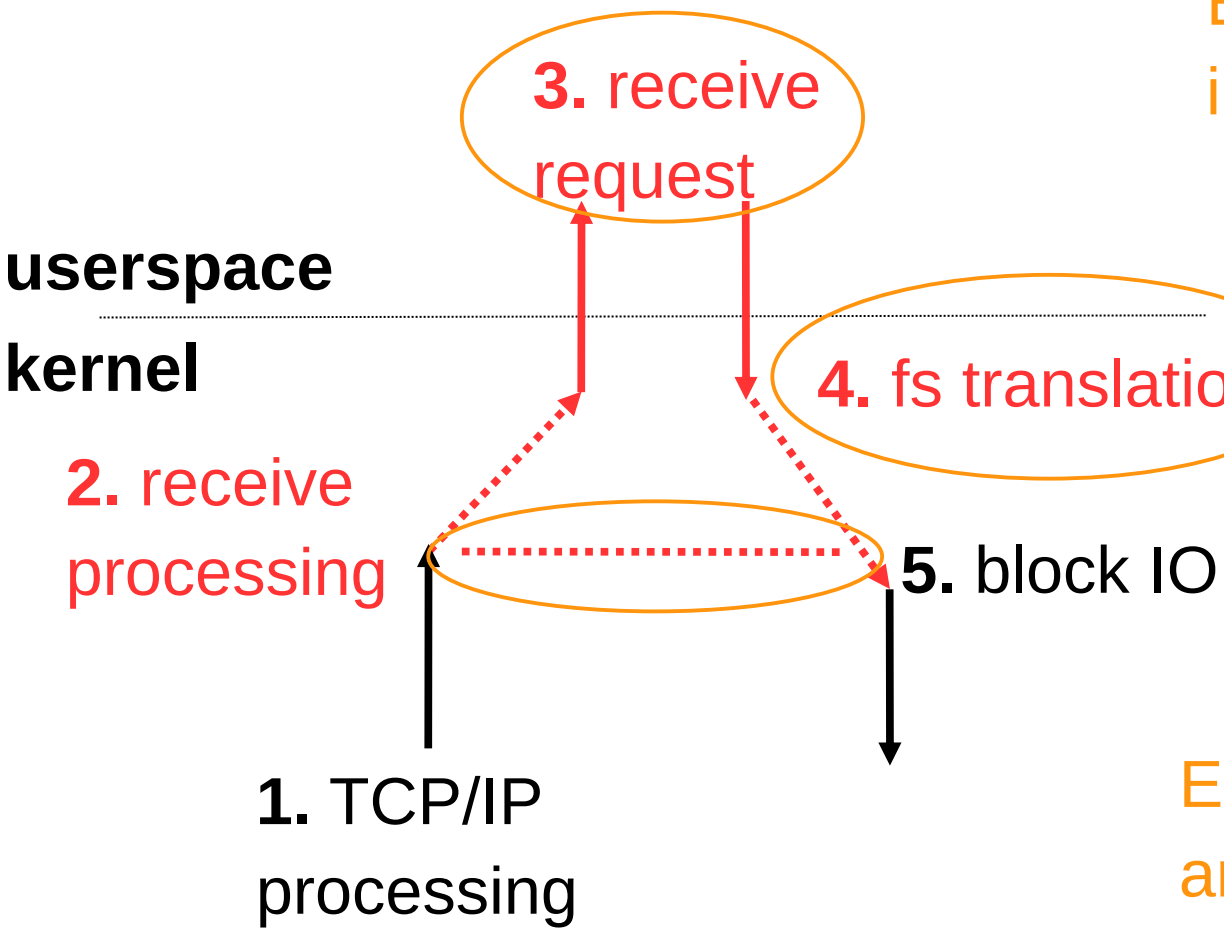
Eliminate application involvement



THE FLASHNET APPROACH



THE FLASHNET APPROACH



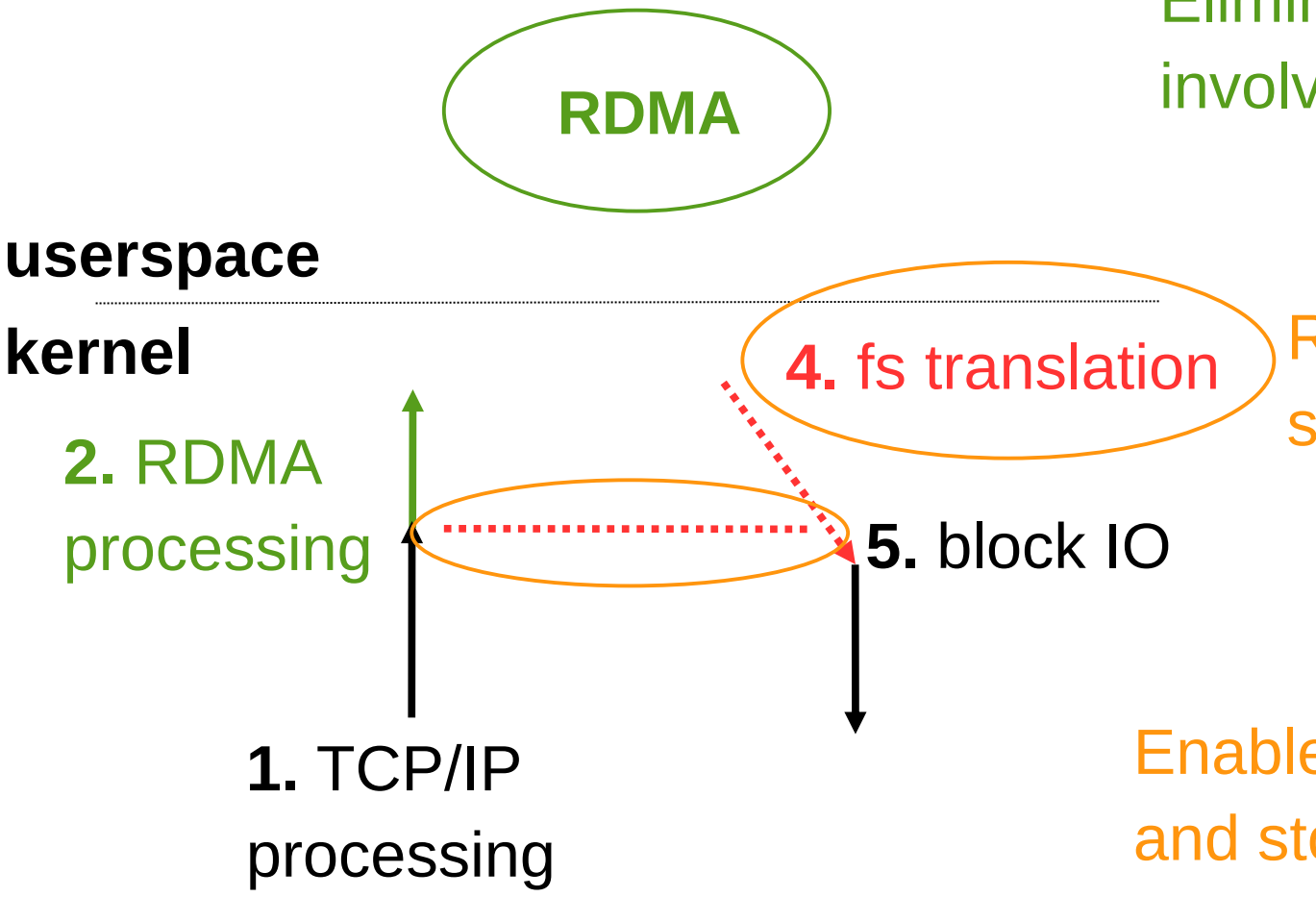
Eliminate application involvement

Reduce file system overheads

Enable direct network and storage interaction

THE FLASHNET APPROACH

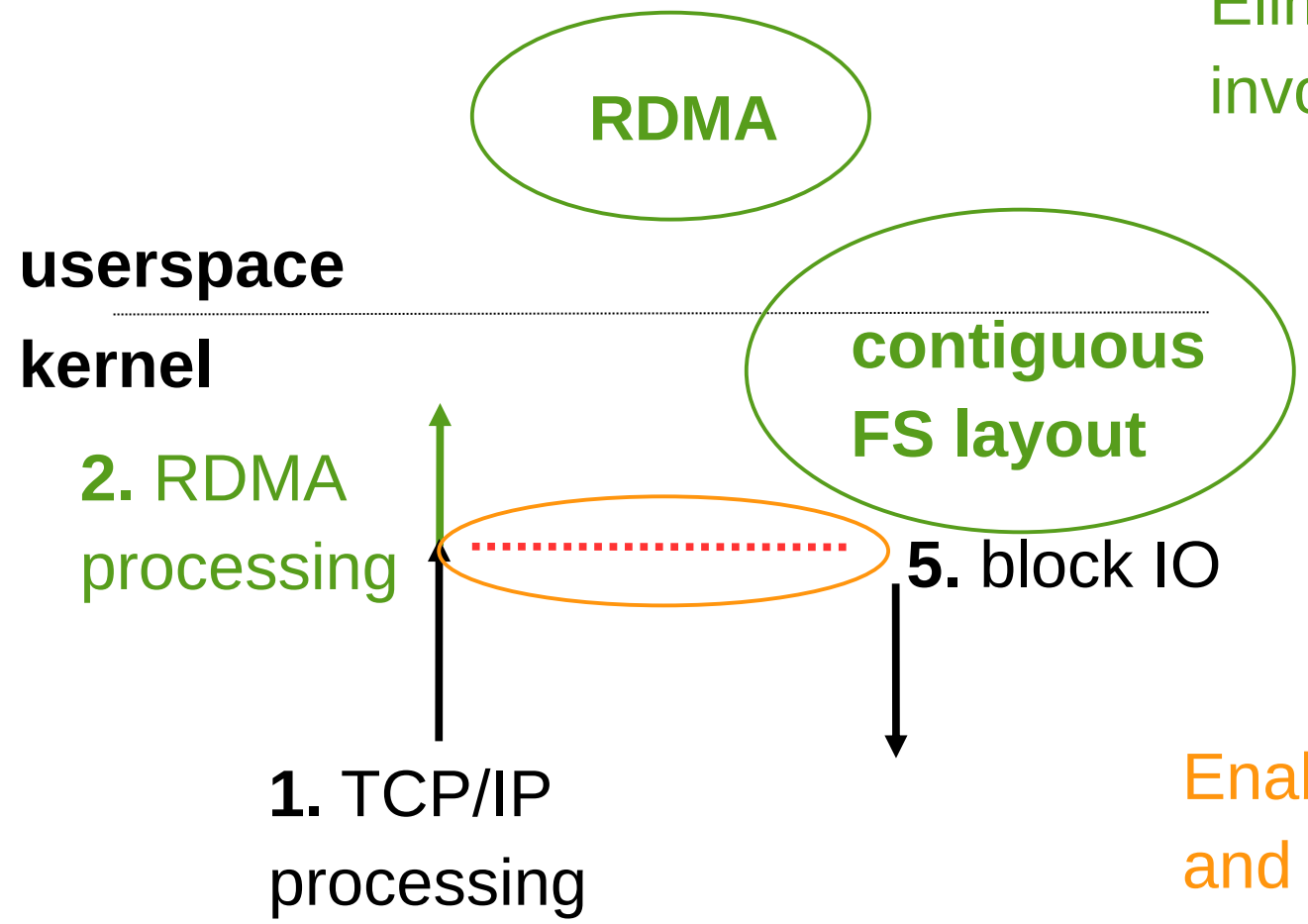
Eliminate application involvement



Reduce file system overheads

Enable direct network and storage interaction

THE FLASHNET APPROACH

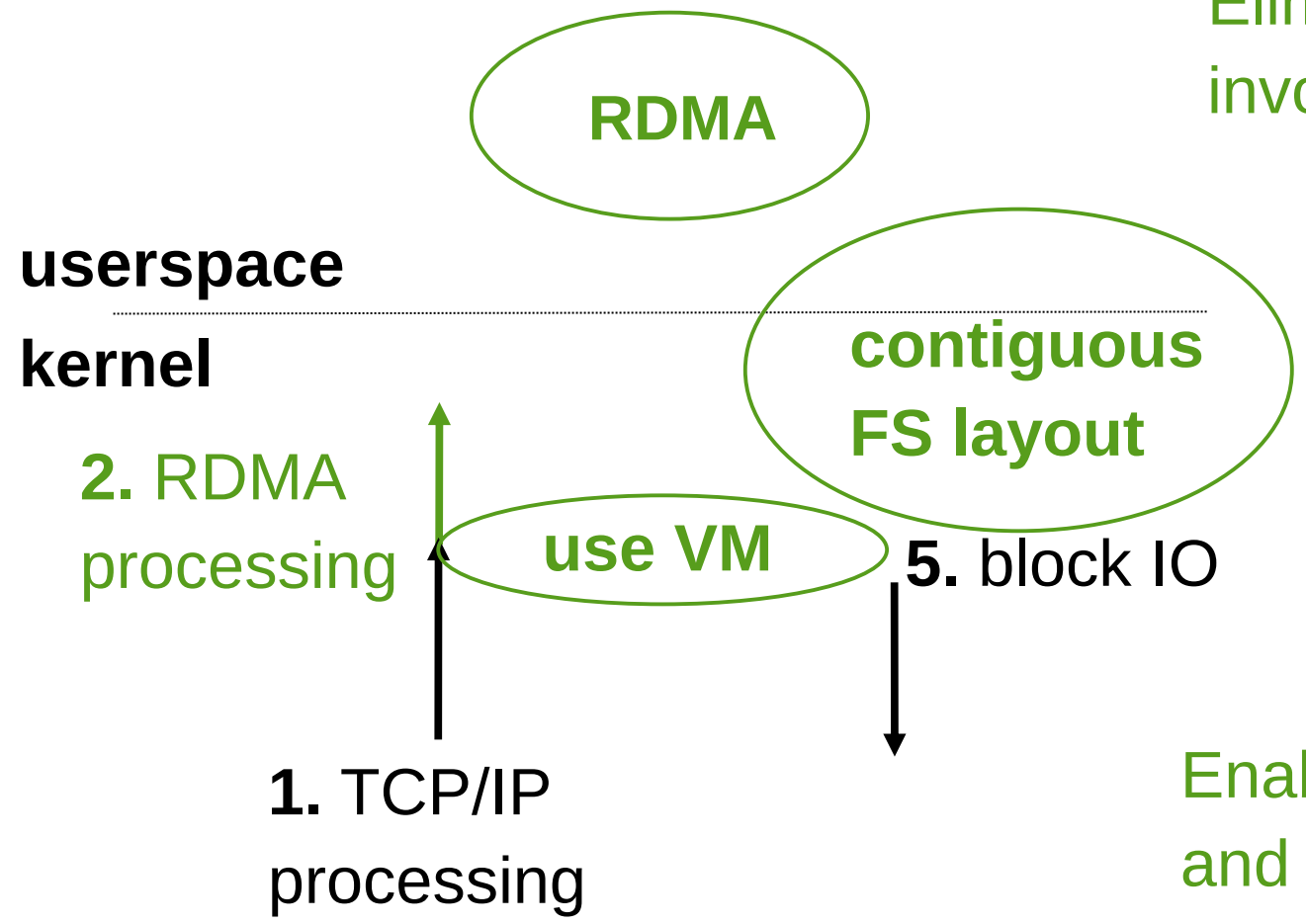


Eliminate application involvement

Reduce file system overheads

Enable direct network and storage interaction

THE FLASHNET APPROACH

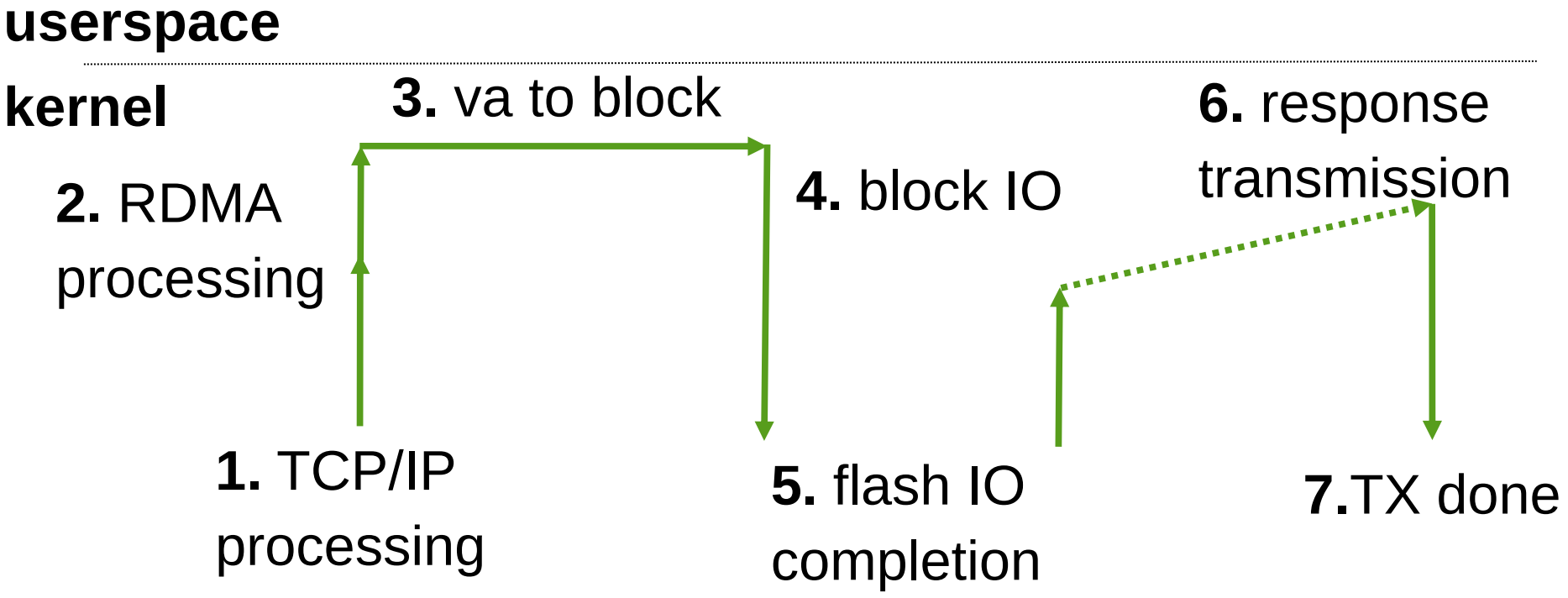


Eliminate application involvement

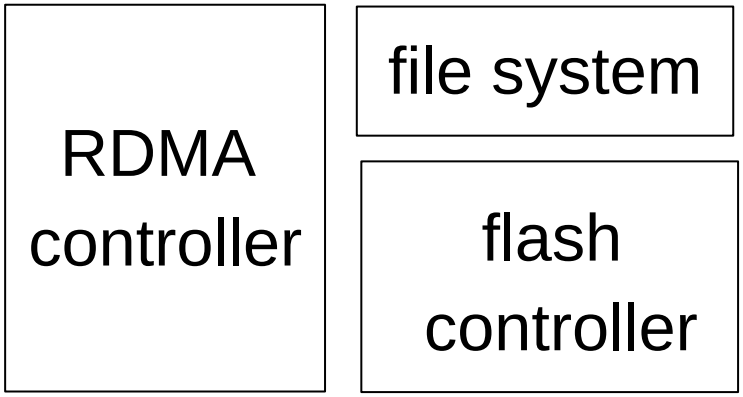
Reduce file system overheads

Enable direct network and storage interaction

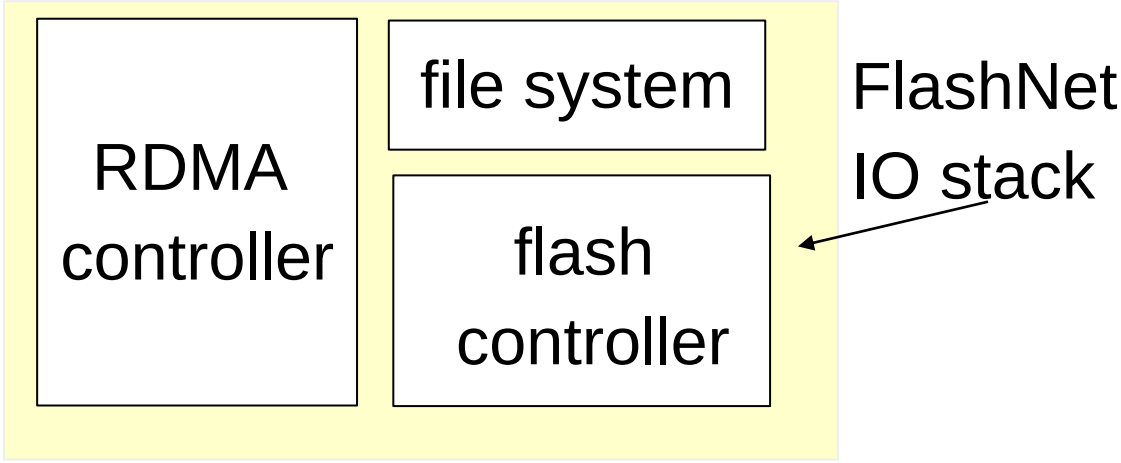
FLASHNET IO OPERATION



FLASHNET: A UNIFIED IO STACK



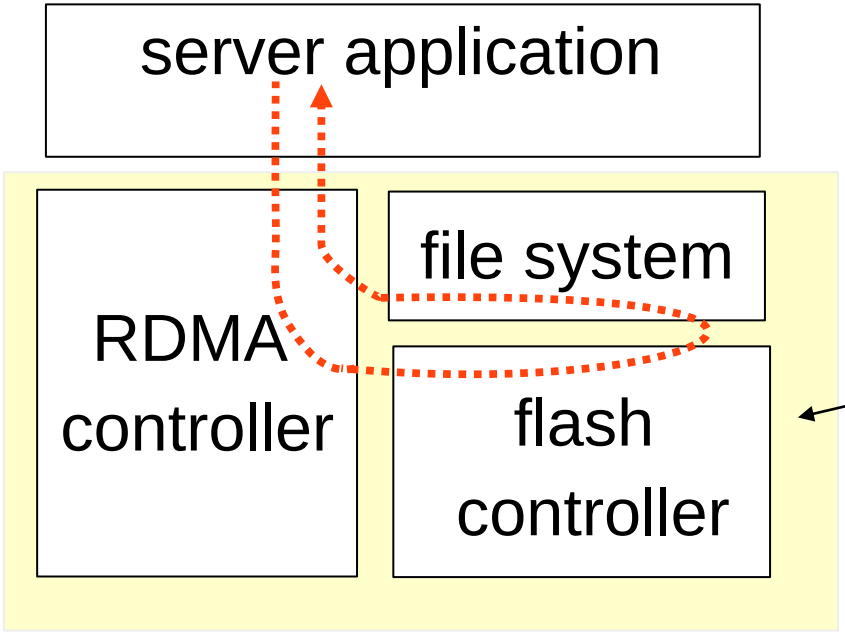
FLASHNET: A UNIFIED IO STACK



[1] SoftiWARP: Software iWARP kernel driver and user library for Linux, Metzler et al, https://github.com/zrlio/softiwarp

[2] SALSA: A unified stack for SSDs and SMR disks, Koltsidas et al. http://ibm.biz/salsa-whitepaper

FLASHNET: A UNIFIED IO STACK

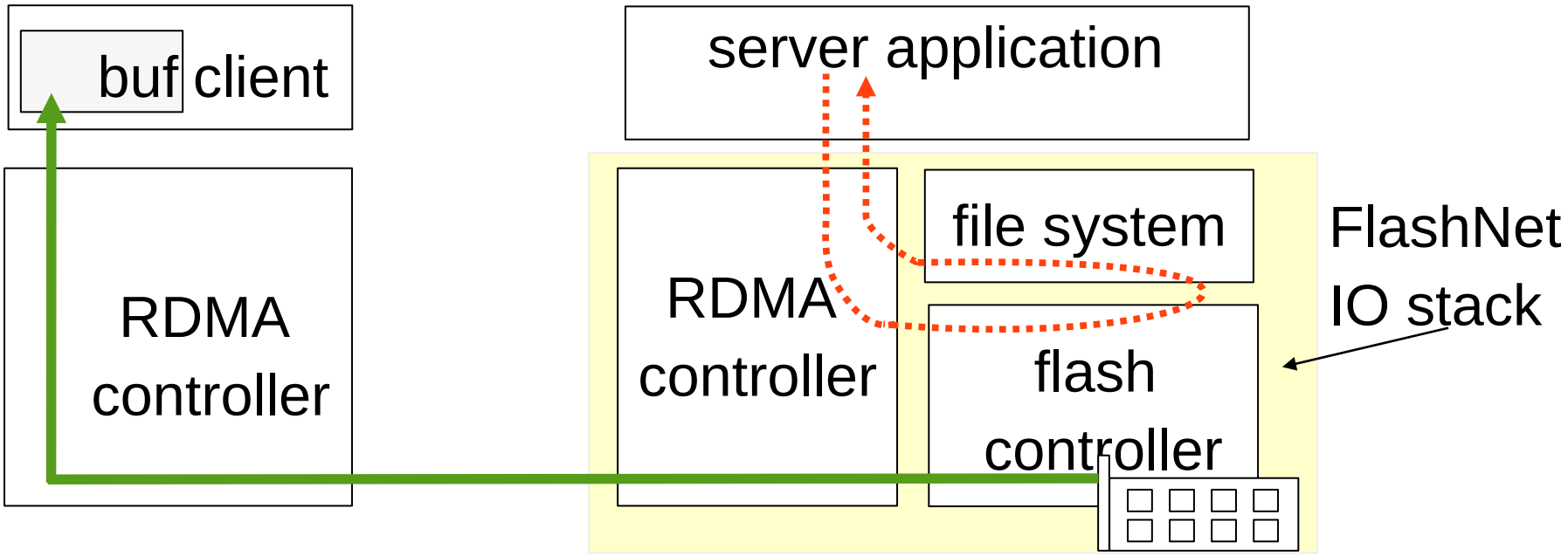


network control setup expanding to storage

[1] SoftiWARP: Software iWARP kernel driver and user library for Linux, Metzler et al, https://github.com/zrlio/softiwarp

[2] SALSA: A unified stack for SSDs and SMR disks, Koltsidas et al. http://ibm.biz/salsa-whitepaper

FLASHNET: A UNIFIED IO STACK



network control setup expanding to storage

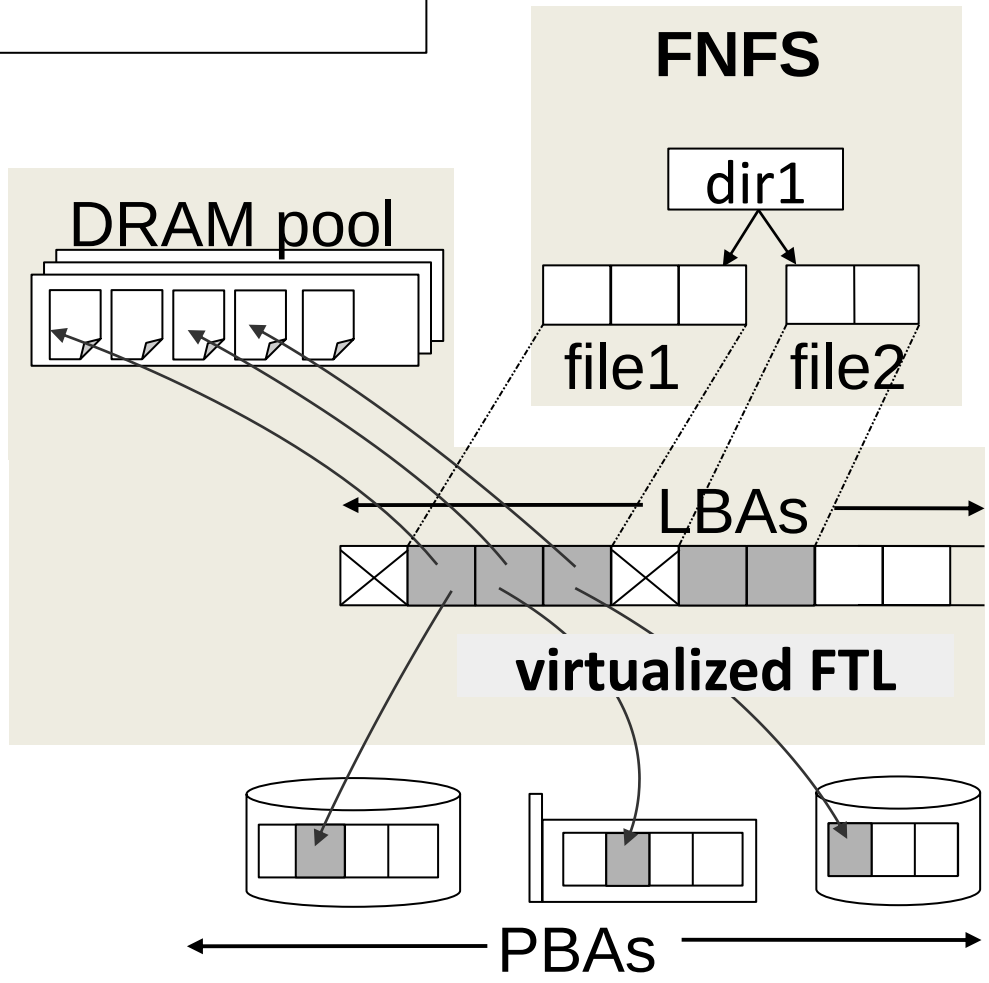
data path from a flash device to a client buffer

[1] SoftiWARP: Software iWARP kernel driver and user library for Linux, Metzler et al, https://github.com/zrlio/softiwarp

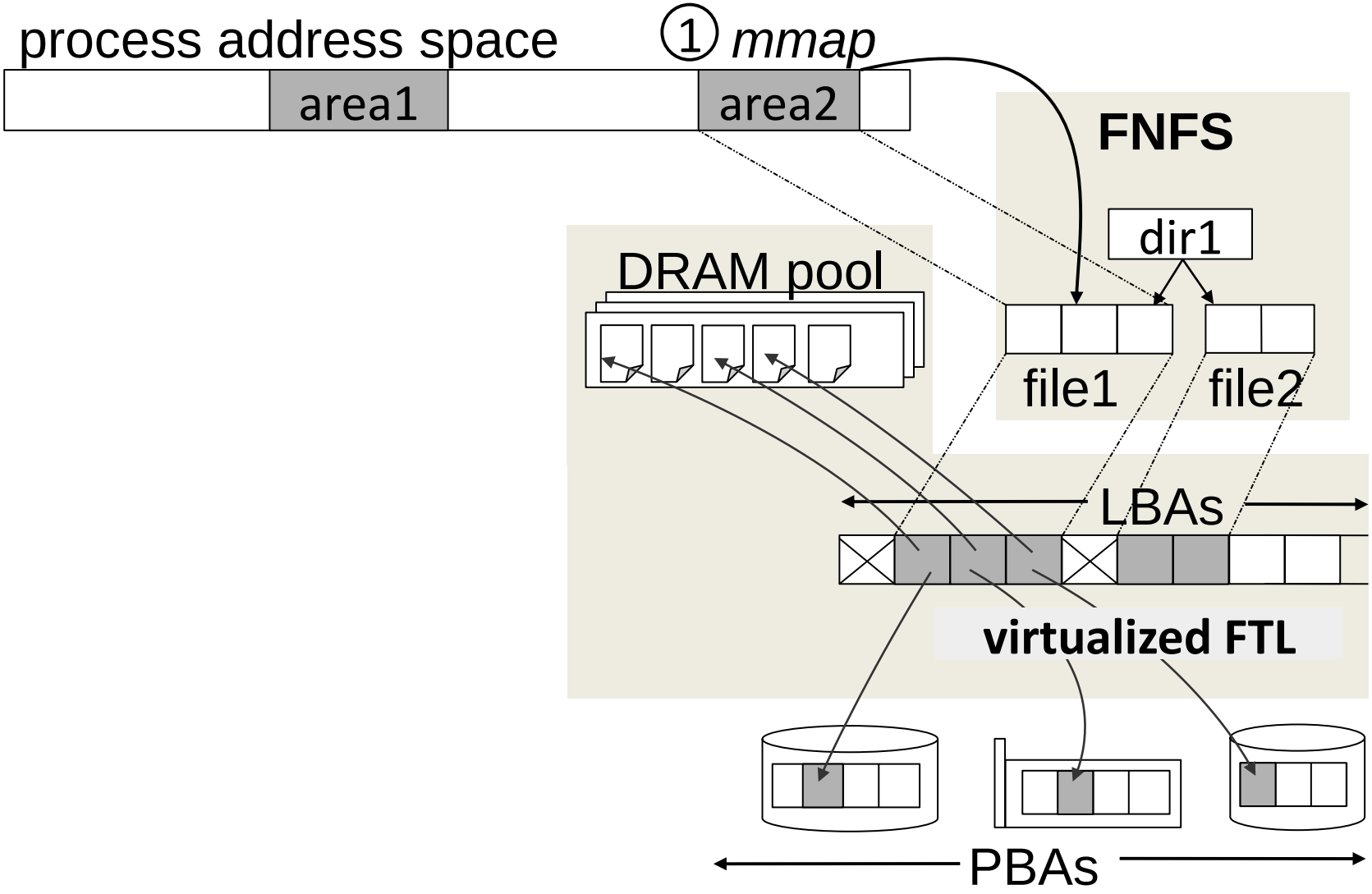
[2] SALSA: A unified stack for SSDs and SMR disks, Koltsidas et al. http://ibm.biz/salsa-whitepaper

ANATOMY OF A FLASHNET OPERATION

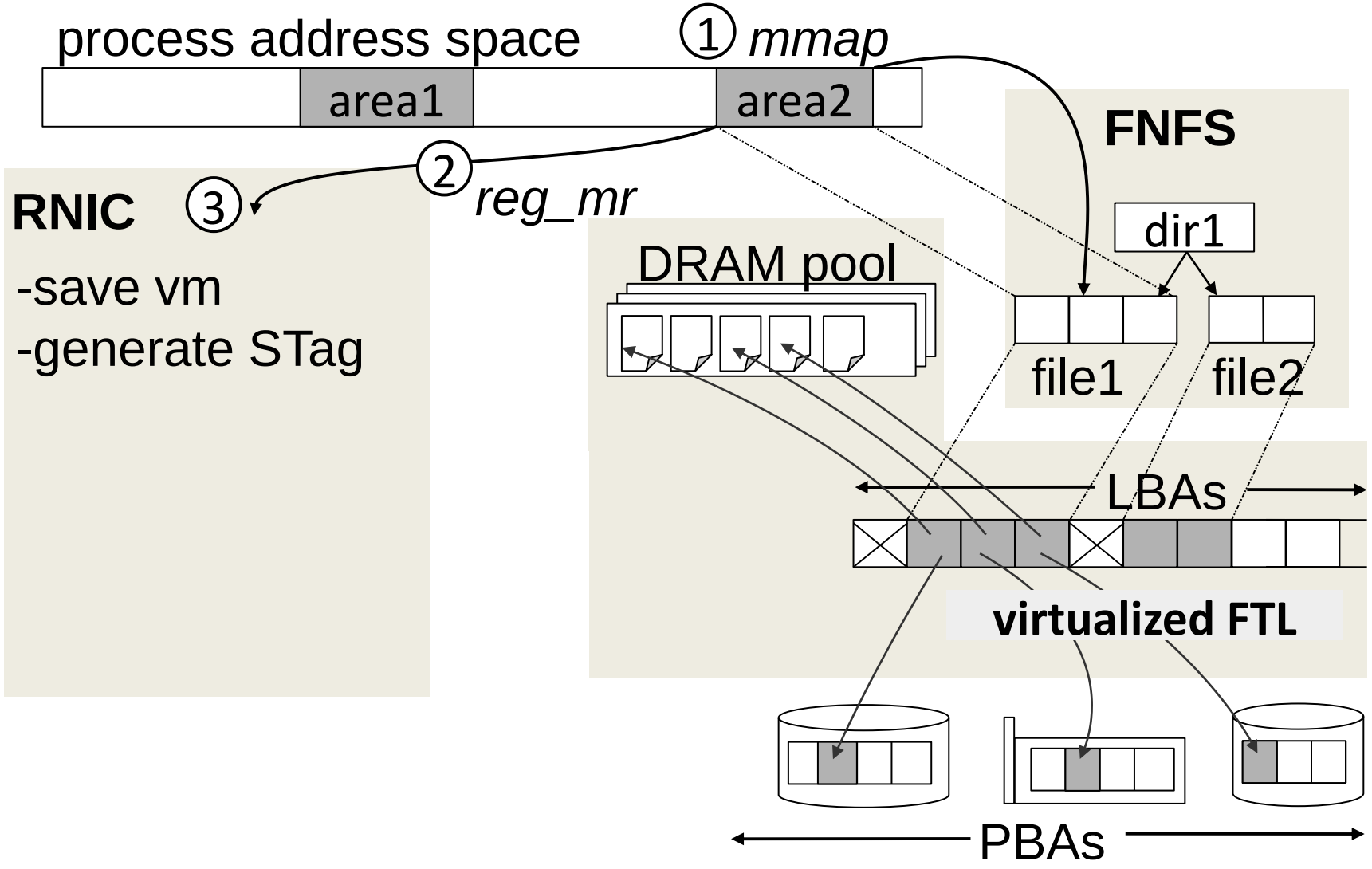
process address space



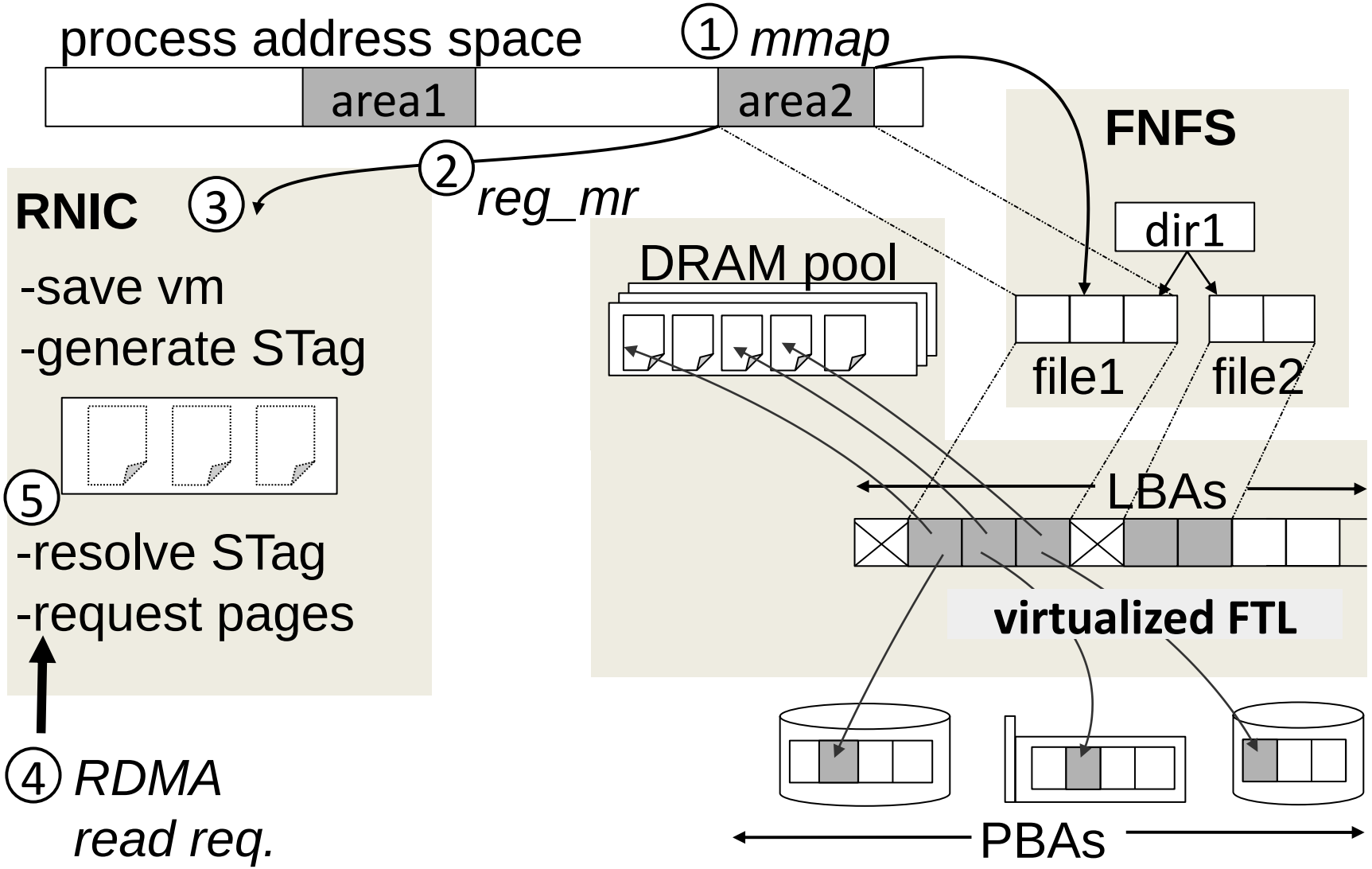
ANATOMY OF A FLASHNET OPERATION



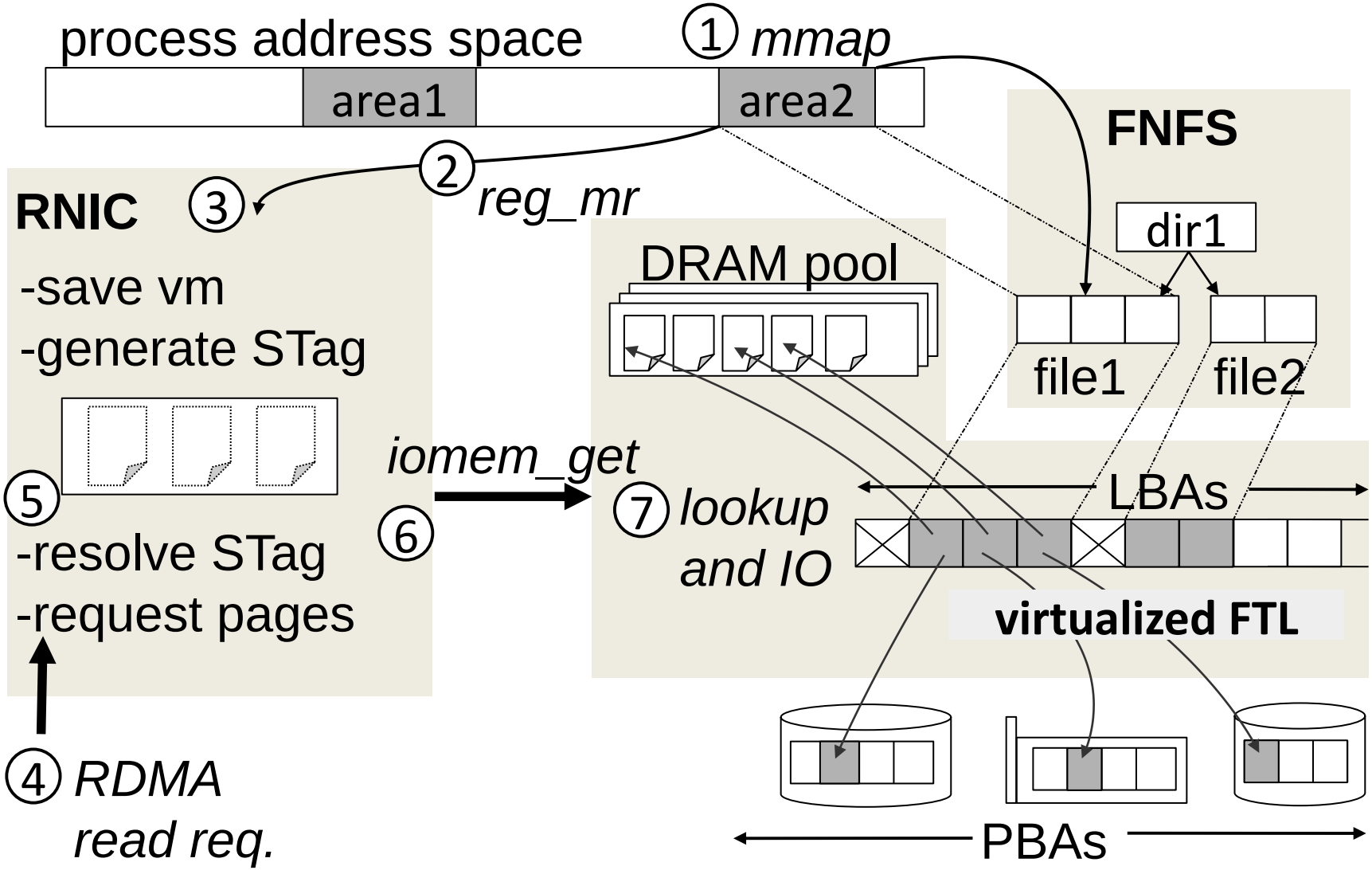
ANATOMY OF A FLASHNET OPERATION



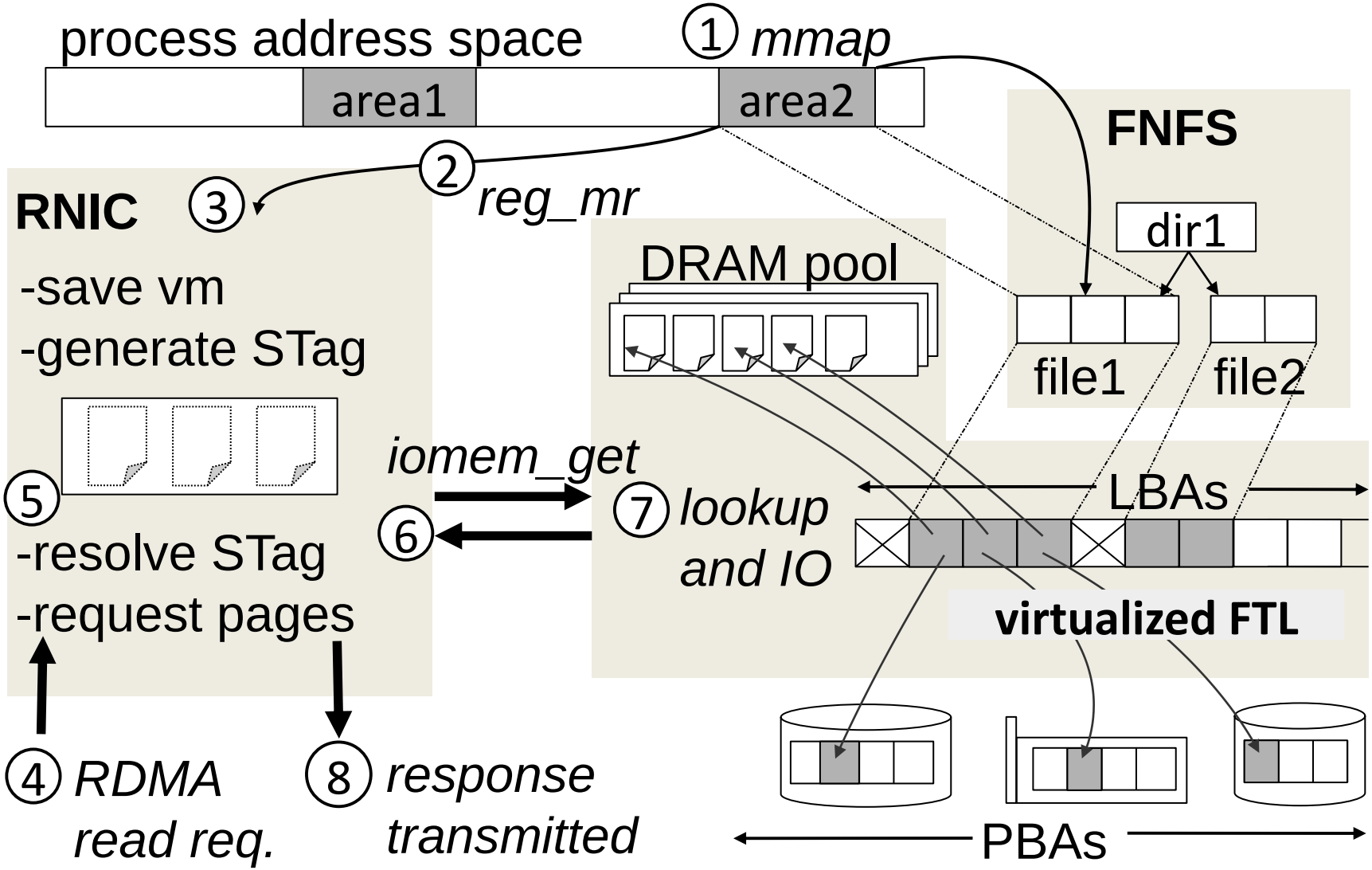
ANATOMY OF A FLASHNET OPERATION



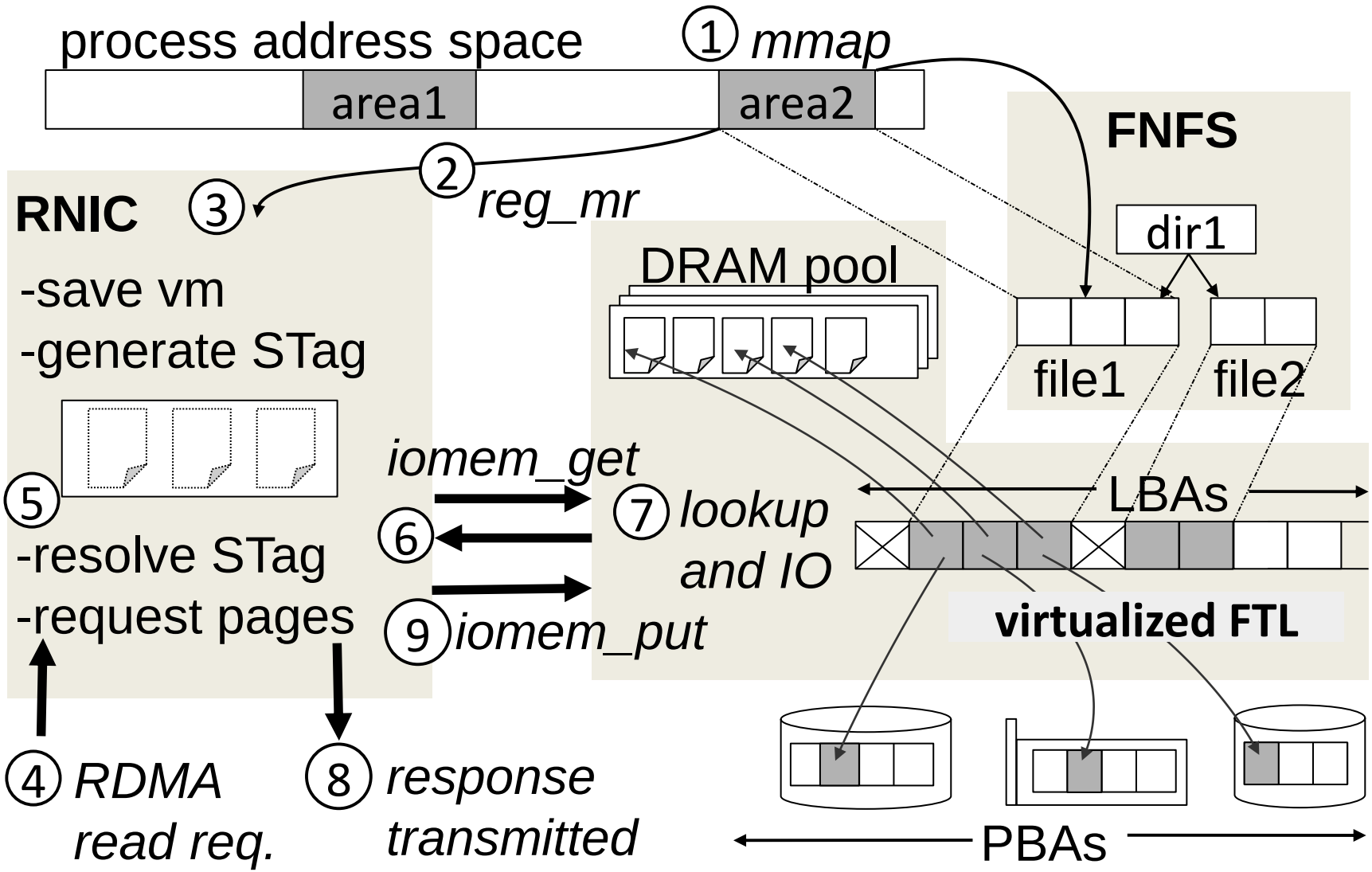
ANATOMY OF A FLASHNET OPERATION



ANATOMY OF A FLASHNET OPERATION



ANATOMY OF A FLASHNET OPERATION



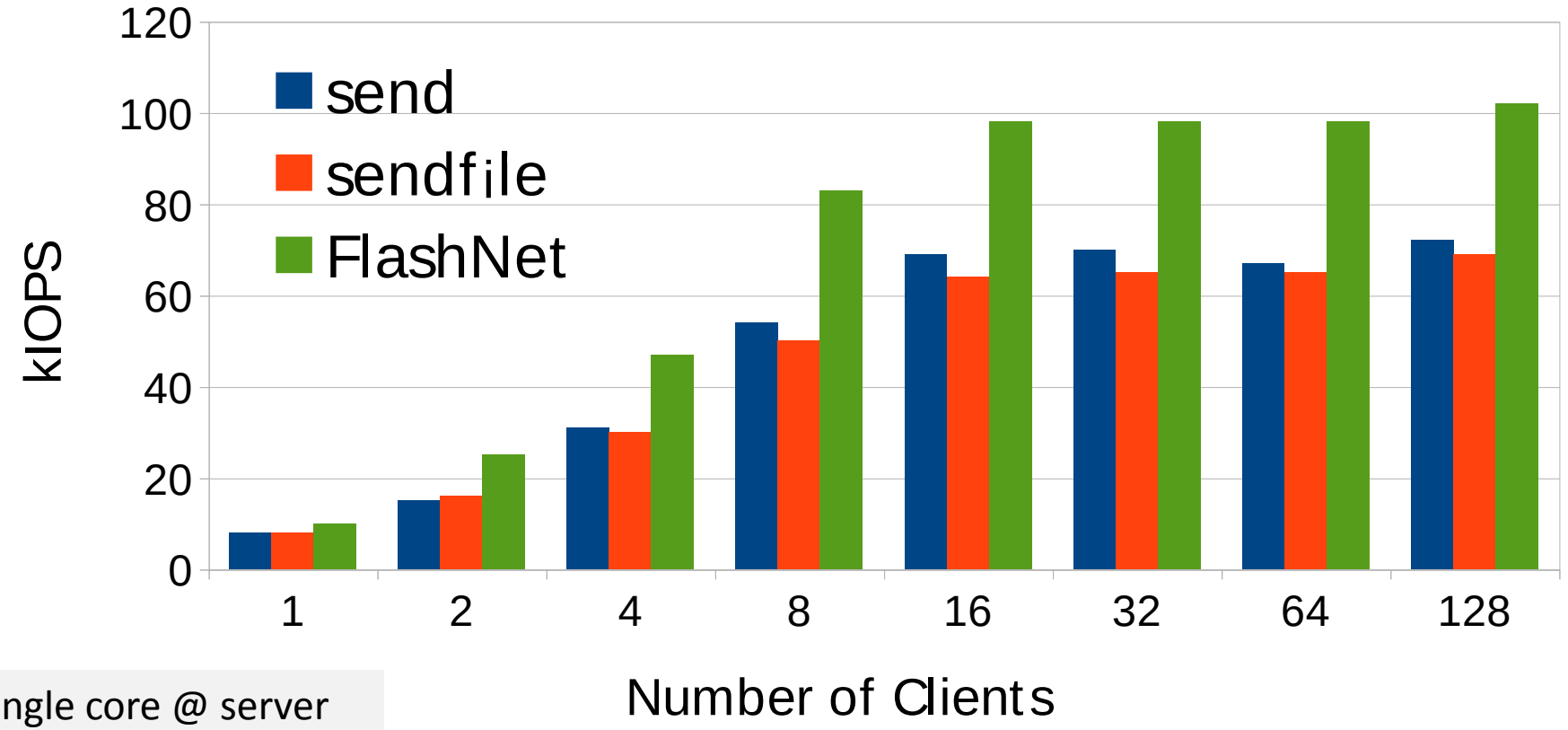
PERFORMANCE EVALUATION

How efficient is FlashNet's IO path?

9-machine cluster testbed

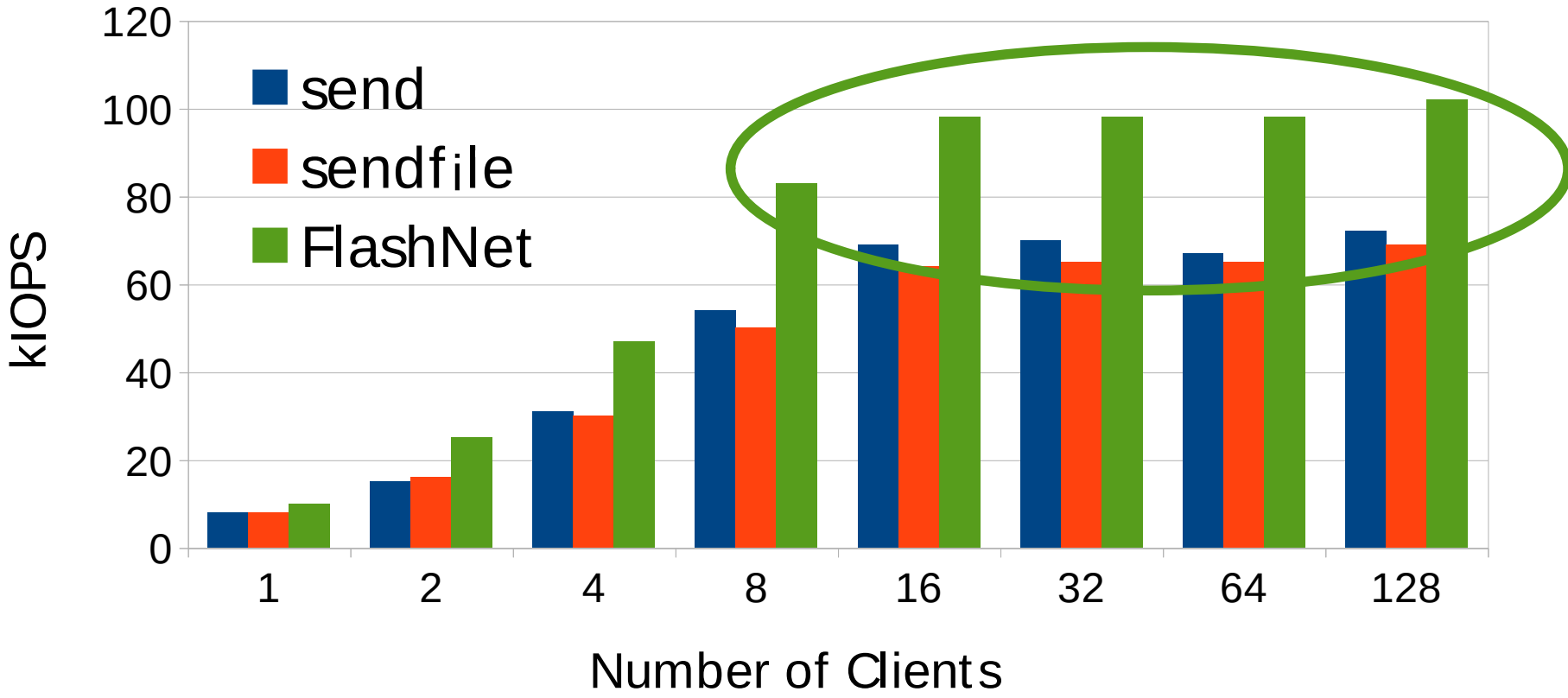
CPU : dual socket E5-2690, 2.9 GHz, 16 cores
DRAM : 256 GB, DDR3 1600 MHz
OS : Linux 3.19 kernel
NIC : 40Gbit/s Ethernet
Flash : 1.3 GB/sec (read), 680 MB/sec (write)
peak read IOPS: 360K, chip latency: 50µsec

PERFORMANCE - IOPS EFFICIENCY



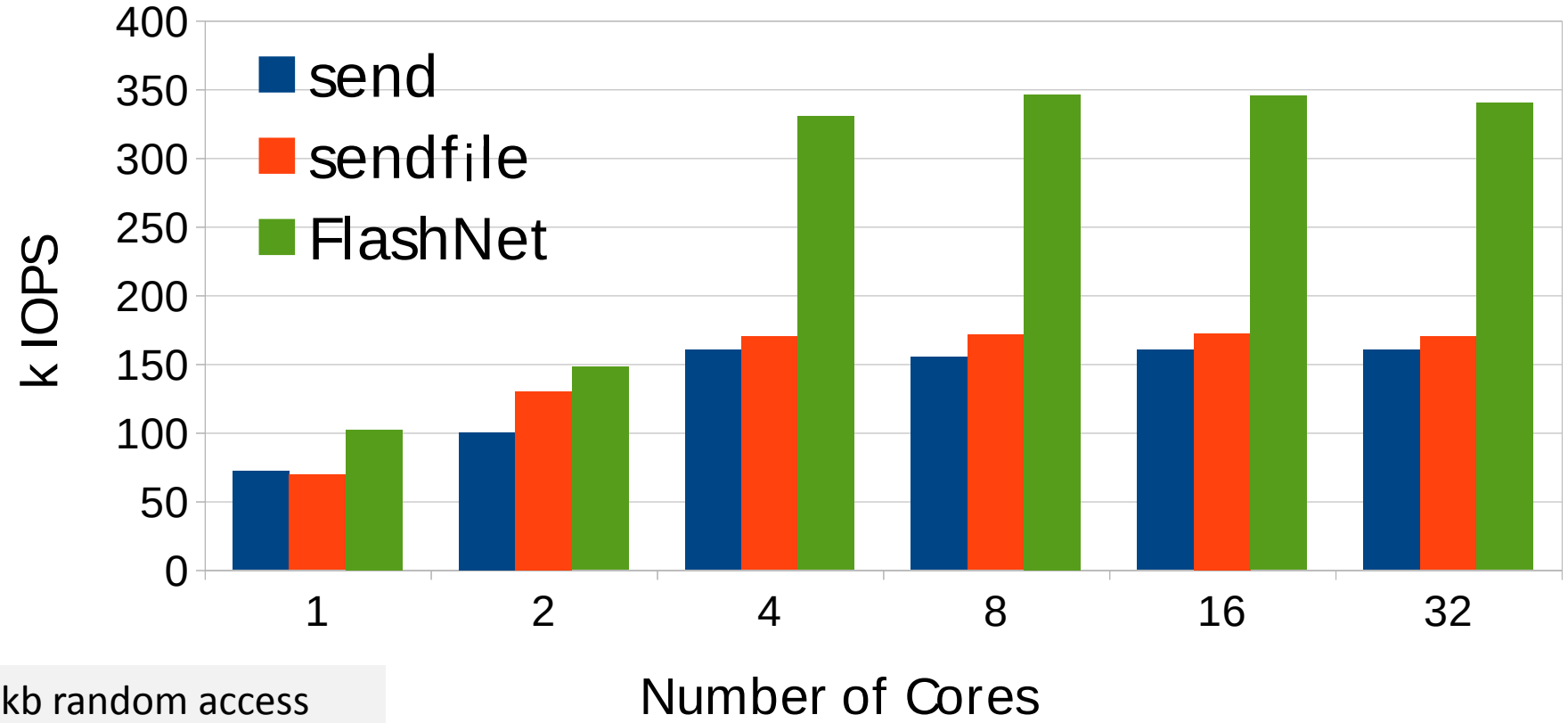
single core @ server
4kb random access
one outstanding req
ext4/fn fs
1 server, 8 client hosts

PERFORMANCE - IOPS EFFICIENCY



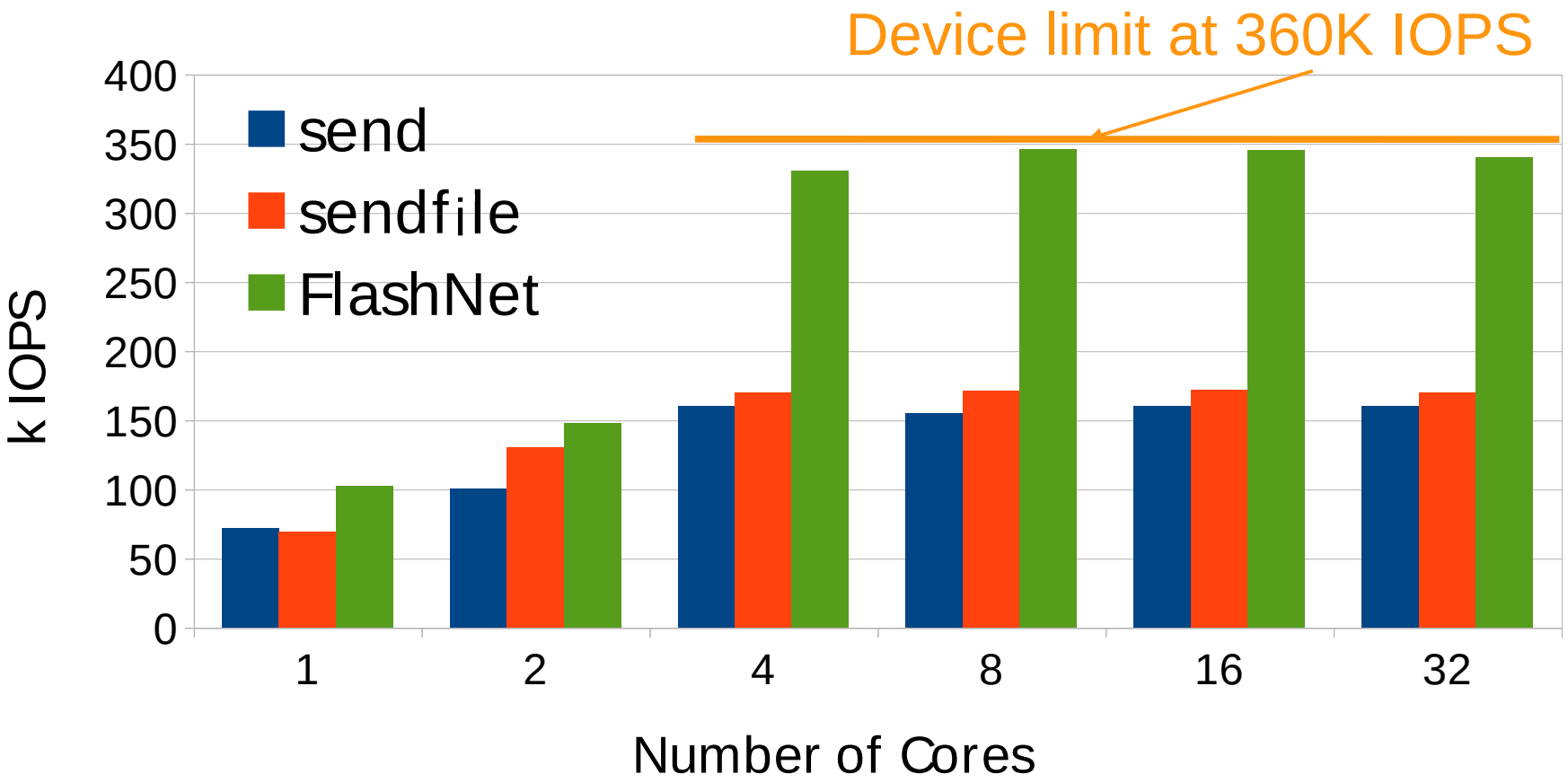
- FlashNet reads are almost 50% more efficient

PERFORMANCE - CORE SCALING



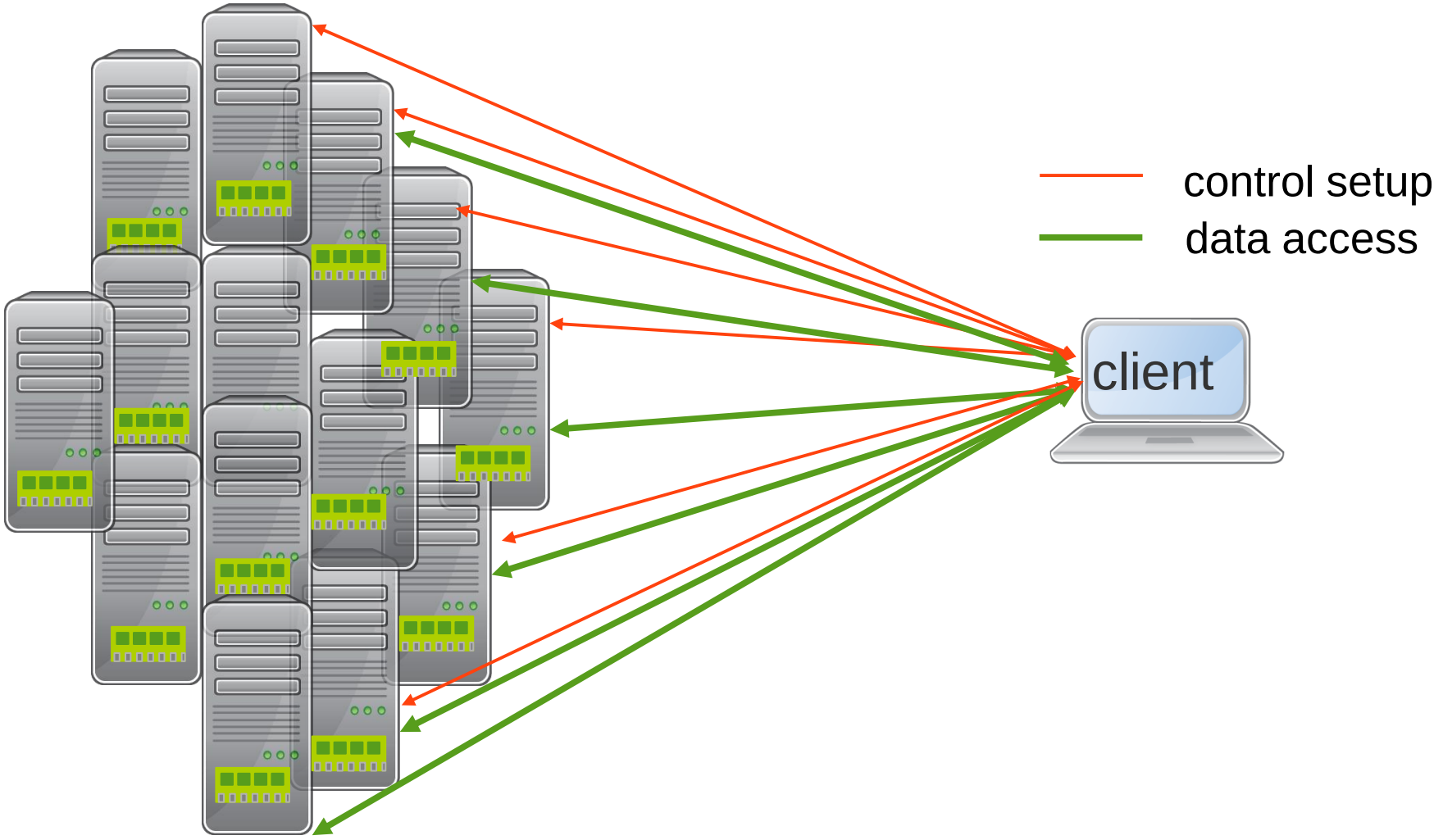
4kb random access
one outstanding req
ext4/fn fs
1 server, 8 client hosts
128 client processes

PERFORMANCE - CORE SCALING



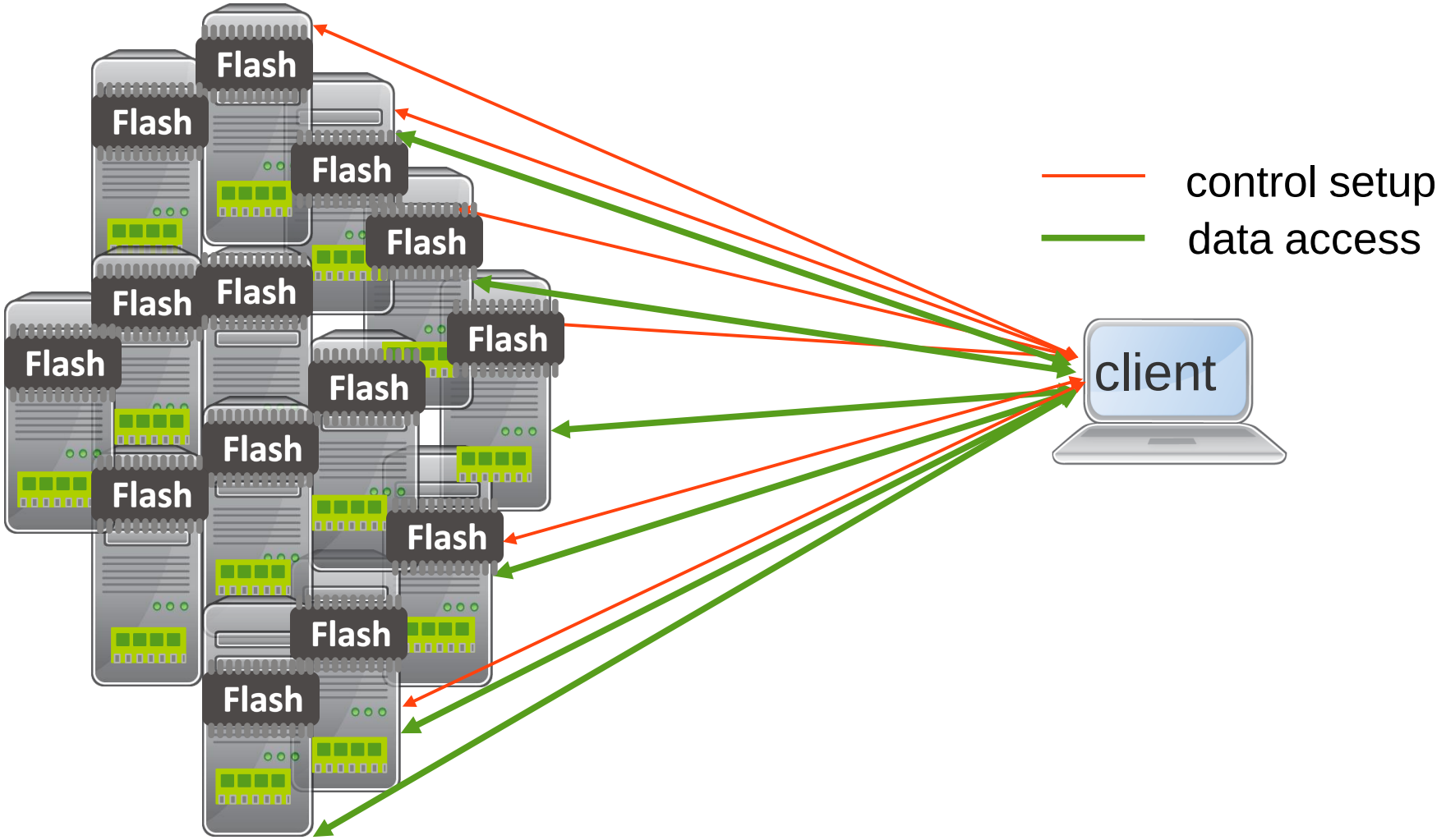
- FlashNet IO operations scale better wrt per-core scaling

APPLICATION: RSTORE ON FLASHNET



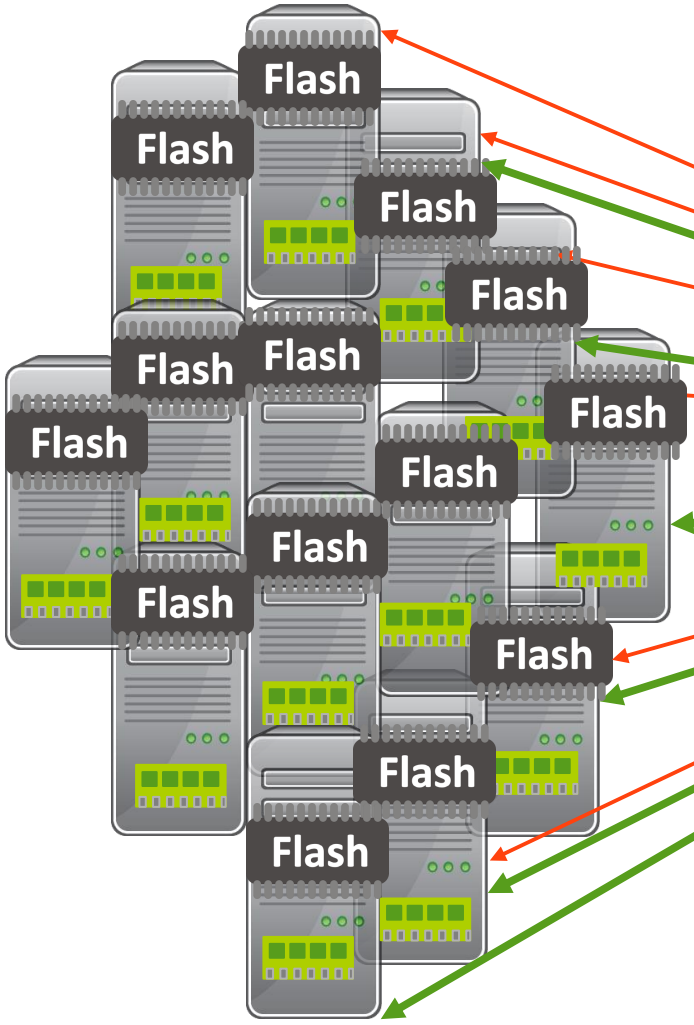
[1] RStore: A Direct-Access DRAM-based Data Store, Trivedi et al., ICDCD'15

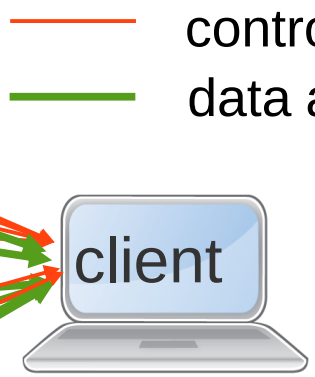
APPLICATION: RSTORE ON FLASHNET



[1] RStore: A Direct-Access DRAM-based Data Store, Trivedi et al., ICDCD'15

APPLICATION: RSTORE ON FLASHNET





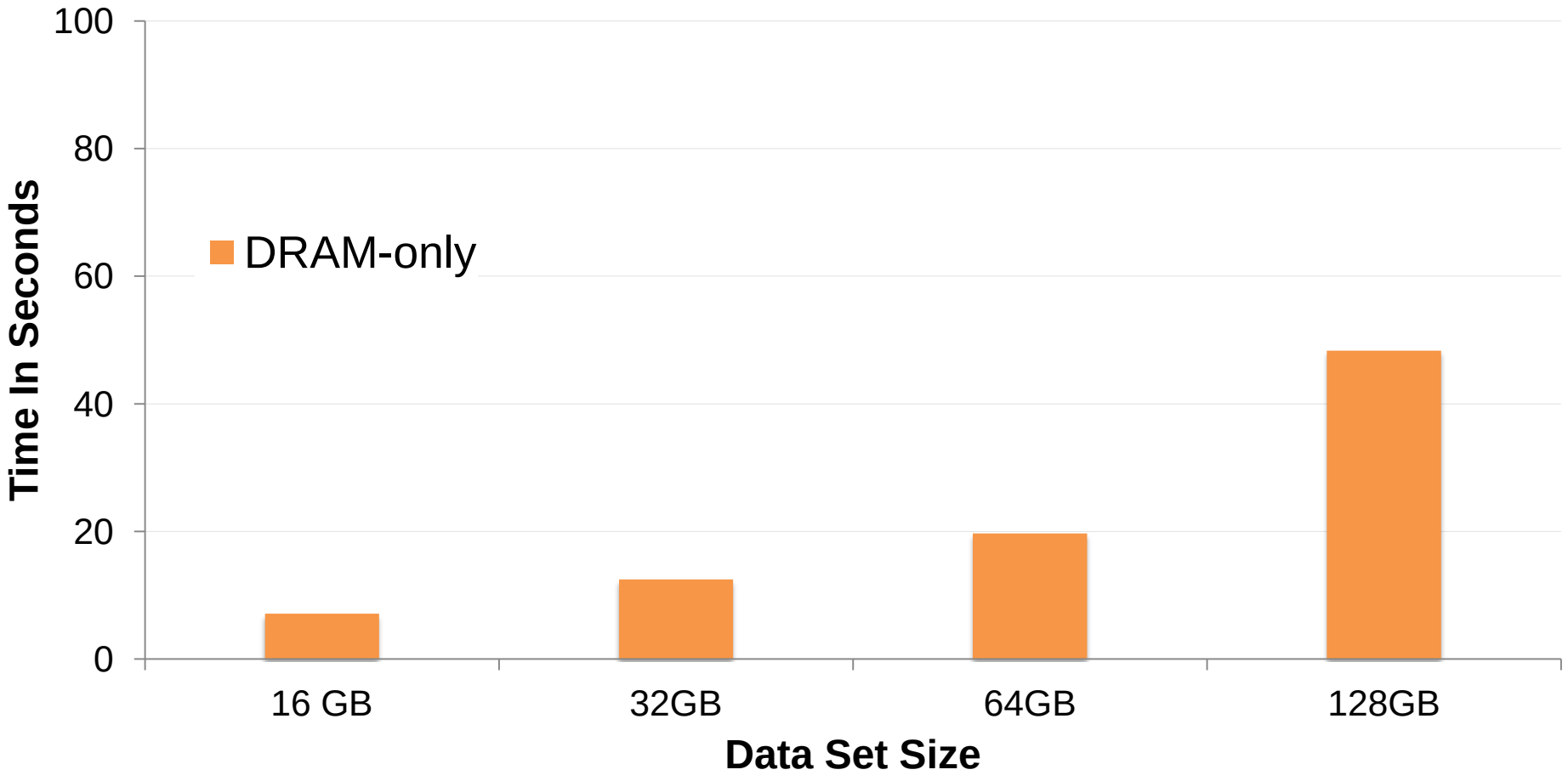
control setup

data access

- **4 lines** to enable FlashNet
- Flash managed by RStore's control API
- Flash accessed by RStore's data API
- Run RSort on FlashNet/Rstore
- 4 servers with direct-attached SSDs

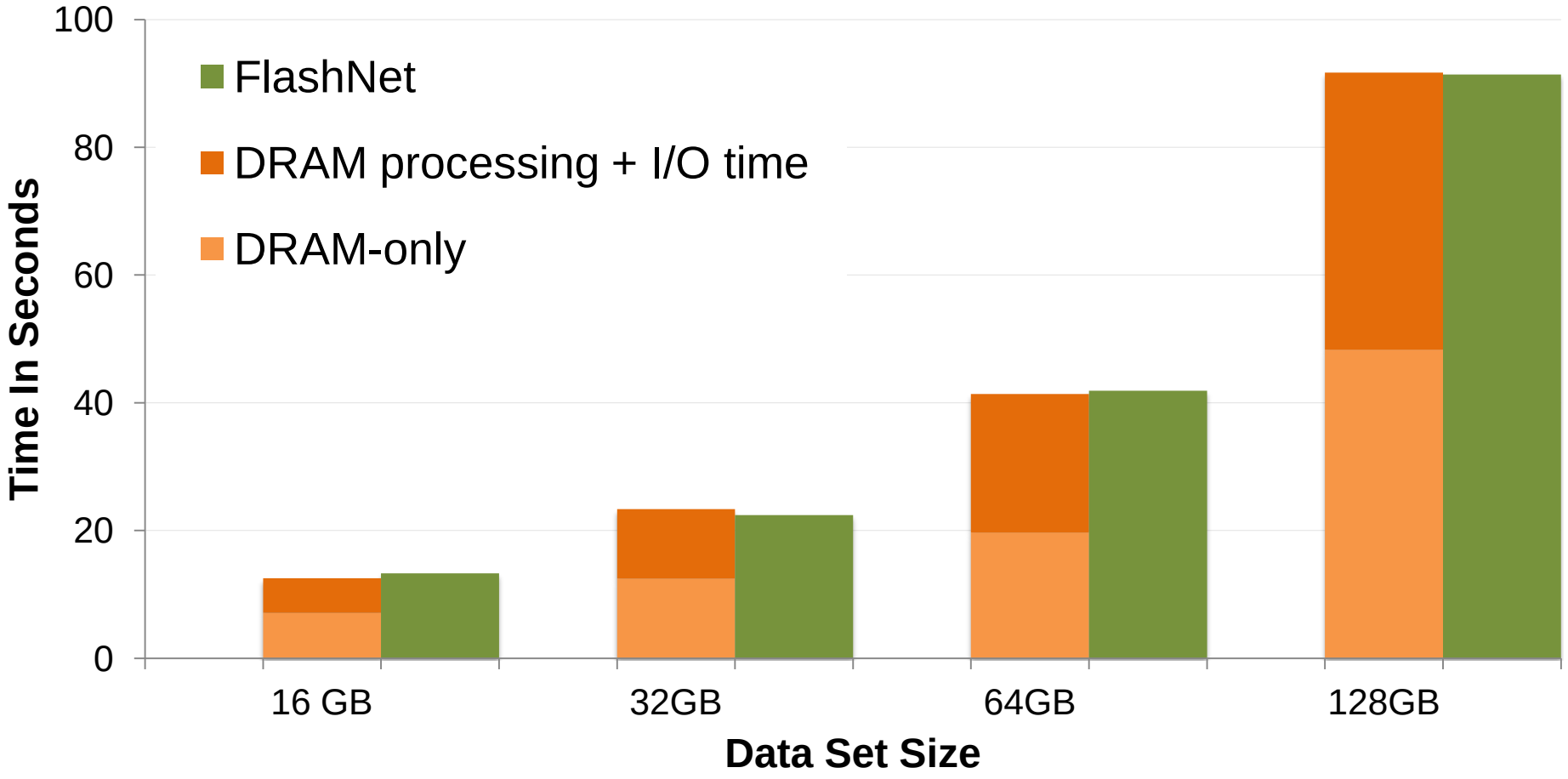
[1] RStore: A Direct-Access DRAM-based Data Store, Trivedi et al., ICDCD'15

RSTORE ON FLASHNET: TERASORT



RSTORE ON FLASHNET: TERASORT

RSTORE ON FLASHNET: TERASORT



RSTORE ON FLASHNET: TERASORT

- FlashNet adds minimum overheads to RDMA-ready applications

CONCLUSIONS

- **CPU-centric IO stacks incur overheads**
- **Solution: Apply unified path separation (ctrl/data) and RDMA access models to both storage and network IO stacks**
- **Implemented software prototype that benefits from unified storage/network access semantics**
- **Demonstrated performance gains for**
  - a unified end-host network-storage stack
  - a distributed data store.
- **More FlashNet benefits:**
  - Client transparent
  - Byte-granular access to storage
  - Easy storage tiering
  - Obsoletes overhead of network storage access protocol
- **Exploring HW implementation**

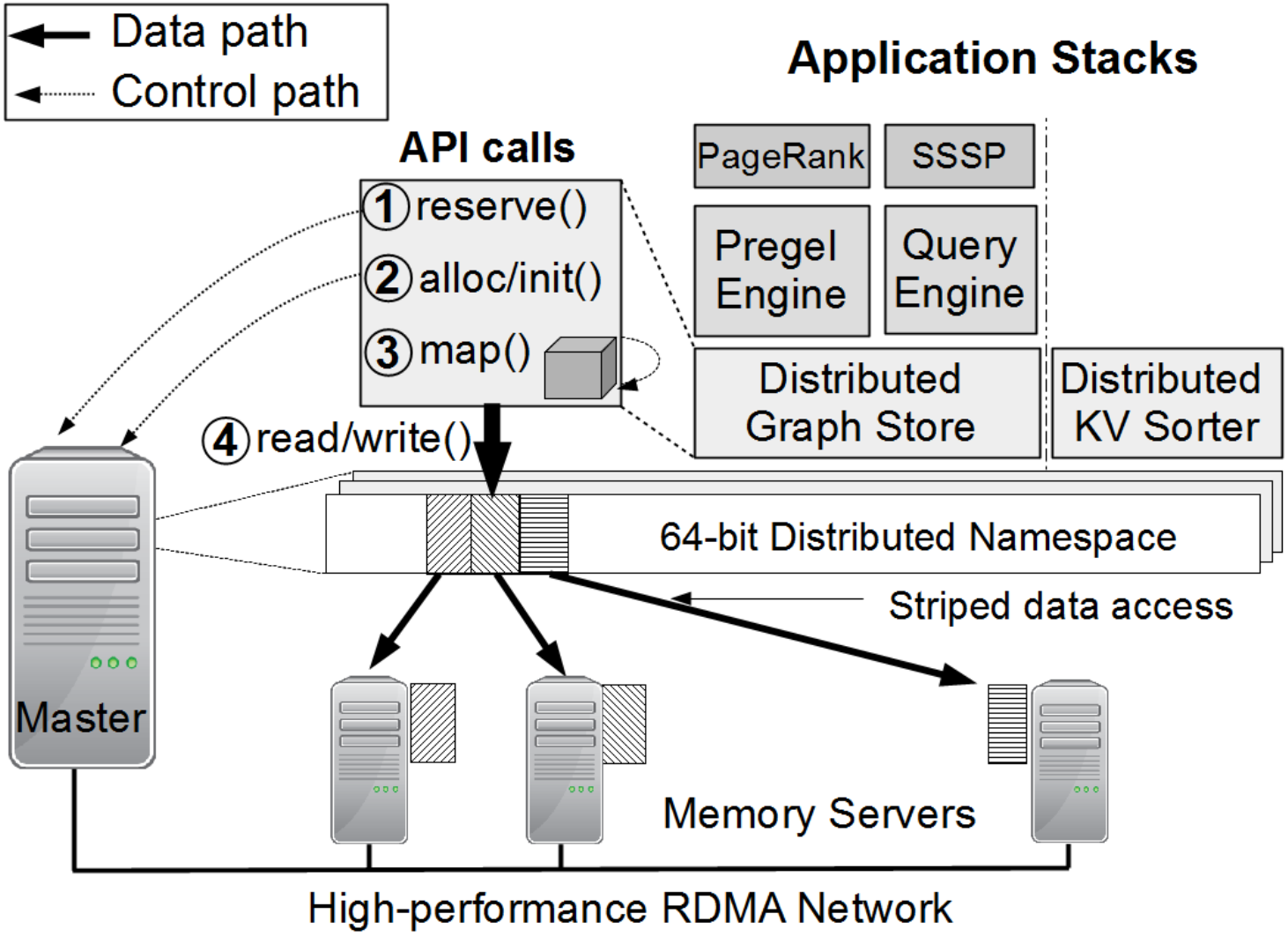
OPENFABRICS
ALLIANCE

12th ANNUAL WORKSHOP 2016

THANK YOU

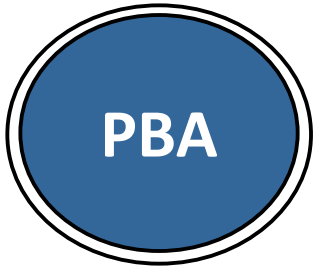
BACKUP

RSTORE



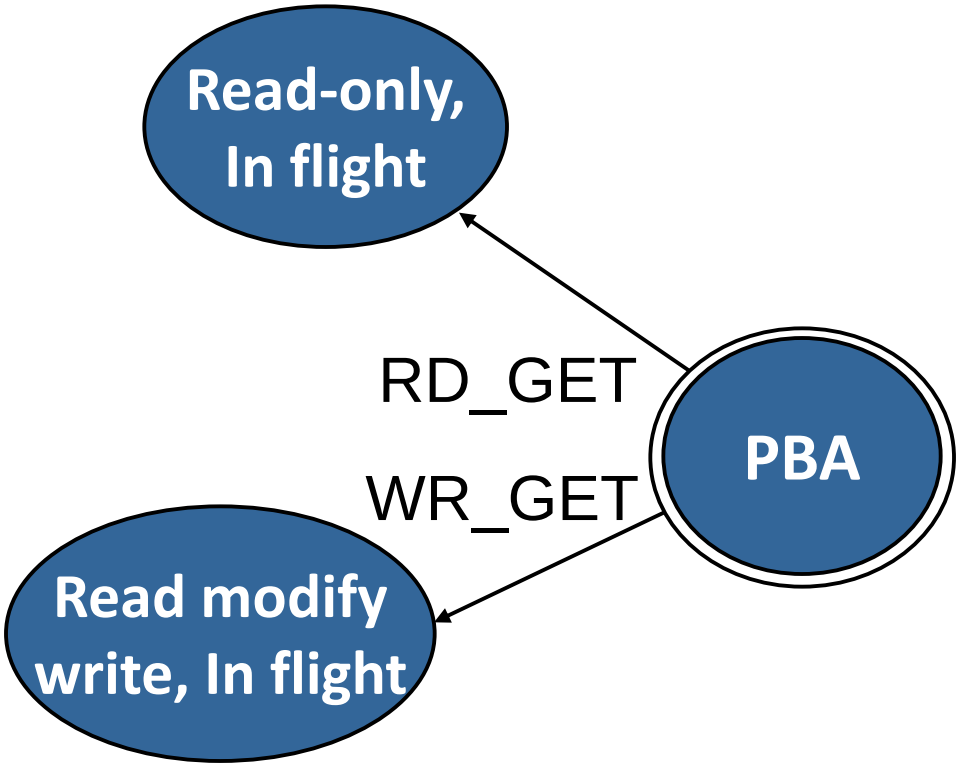
[1] RStore: A Direct-Access DRAM-based Data Store, Trivedi et al., ICDCD'15

FLASHNET: FLASH CONTROLLER

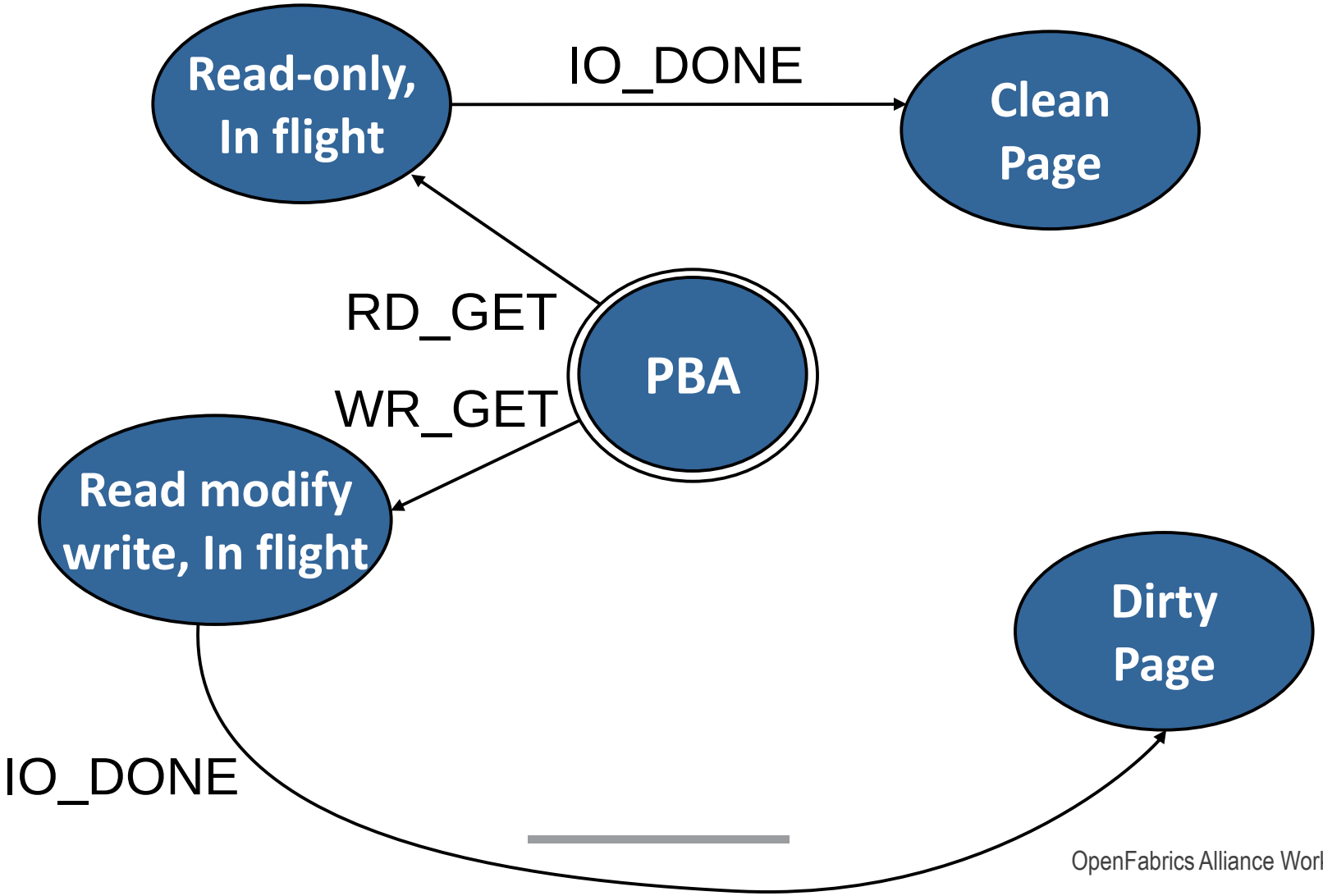


PBA state says a flash Logical Block Address (LBA) is stored on a Physical Block Address (PBA) on a device not in a DRAM page

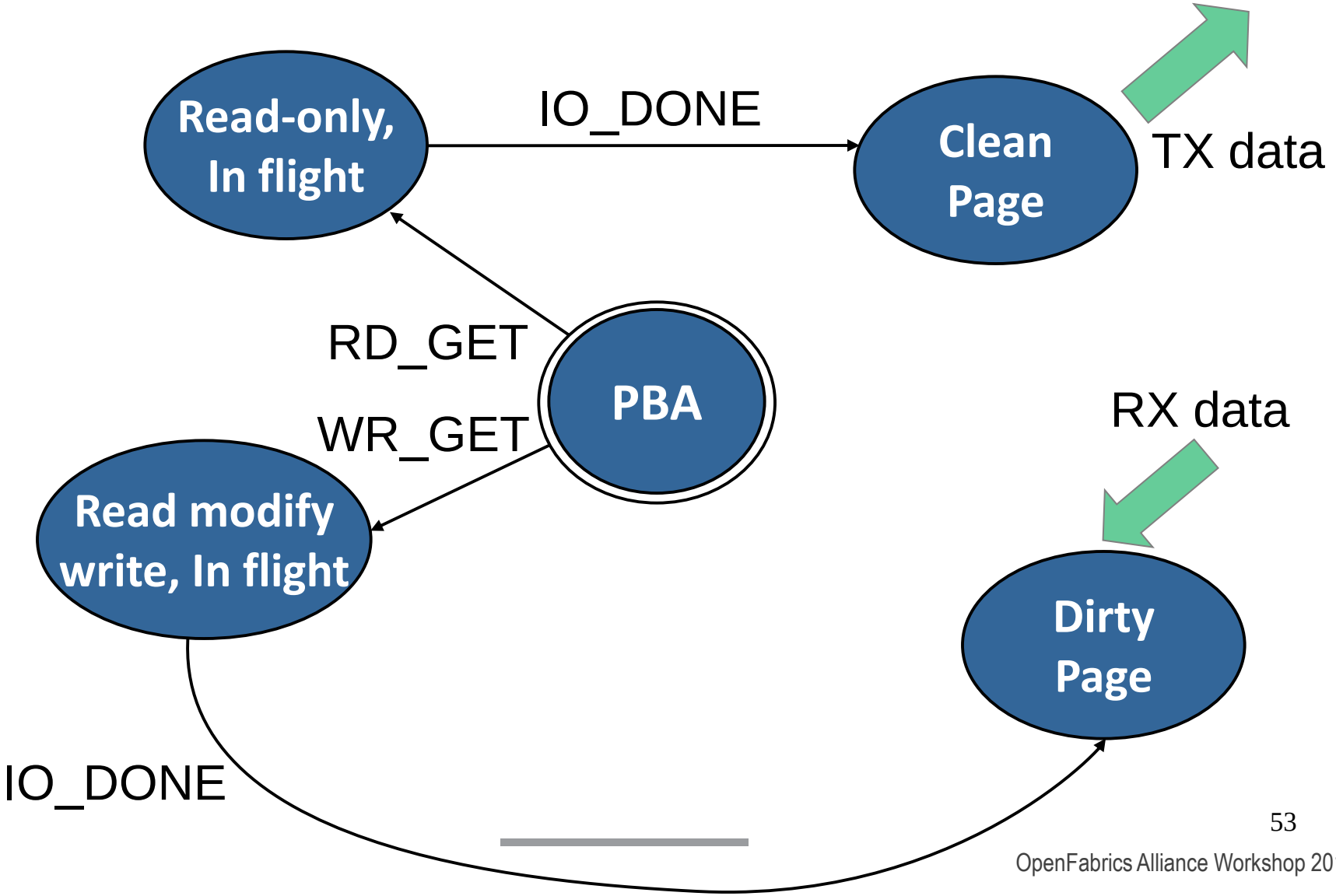
FLASHNET: FLASH CONTROLLER



FLASHNET: FLASH CONTROLLER



FLASHNET: FLASH CONTROLLER



FLASHNET: FLASH CONTROLLER