CAM I/O Scheduler

M. Warner Losh

Netflix, Inc.

BSDCan 2015

http://people.freebsd.org/~imp/bsdcan2015/iosched-slides.pdf

http://people.freebsd.org/~imp/bsdcan2015/paper.pdf

NETFLIX

Outline

Overview / Motivation
- Graphs
- Roadmap

Background and Context
- Netflix OCA
- NAND Physics and SSD
- FreeBSD I/O Stack

Netflix I/O Scheduler

Recent Updates

NETFLIX

Outline

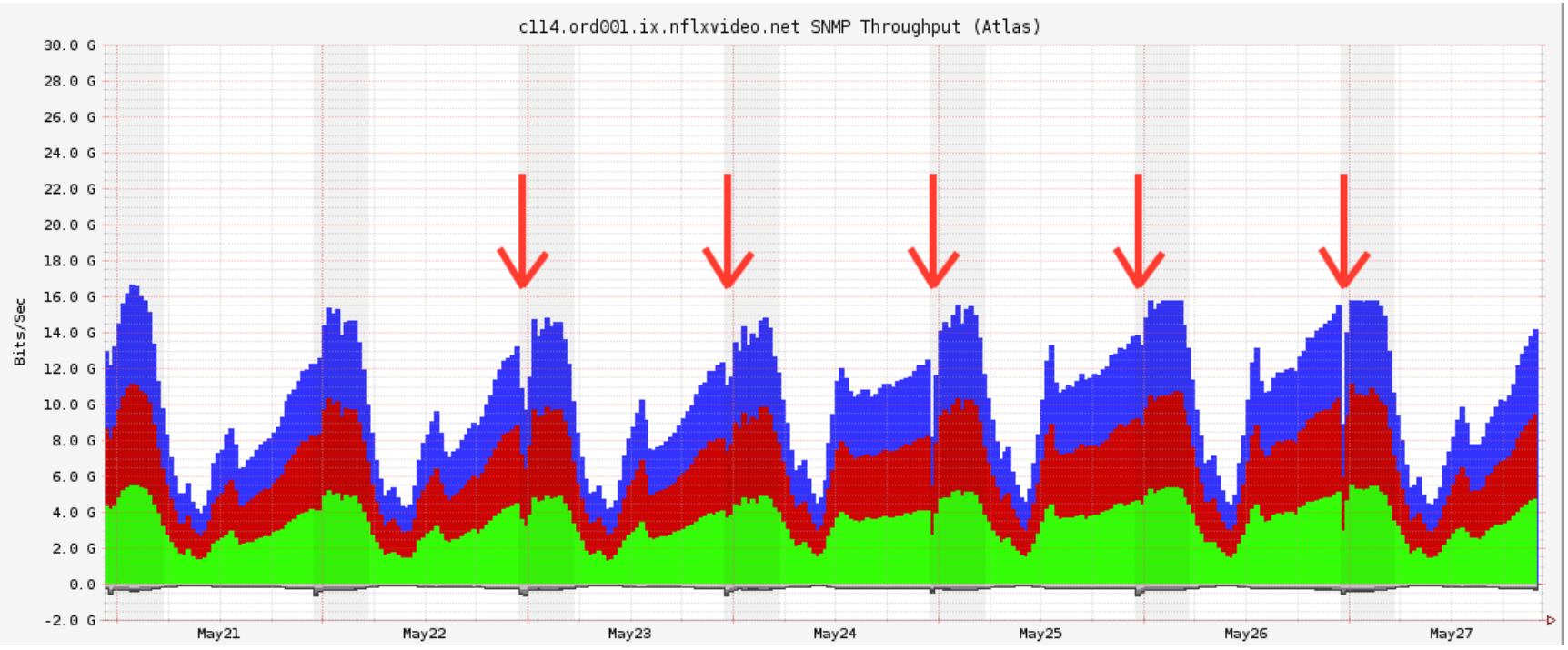
NETFLIX

Motivations

- Big performance hit in our operations
  - 20-50% drop in throughput
  - devstat reporting 100% busy for SSD with tiny use
  - Netflix's QoE metrics show customer problems
- Traced to adding / deleting content
  - Happened during content fill
  - Work around by idling server before fill
- Sometimes an issue, other times not
  - Some systems would have problems, others not
  - No fixed pattern to failure
  - Only affected flash caches
  - Seems related to write activity.

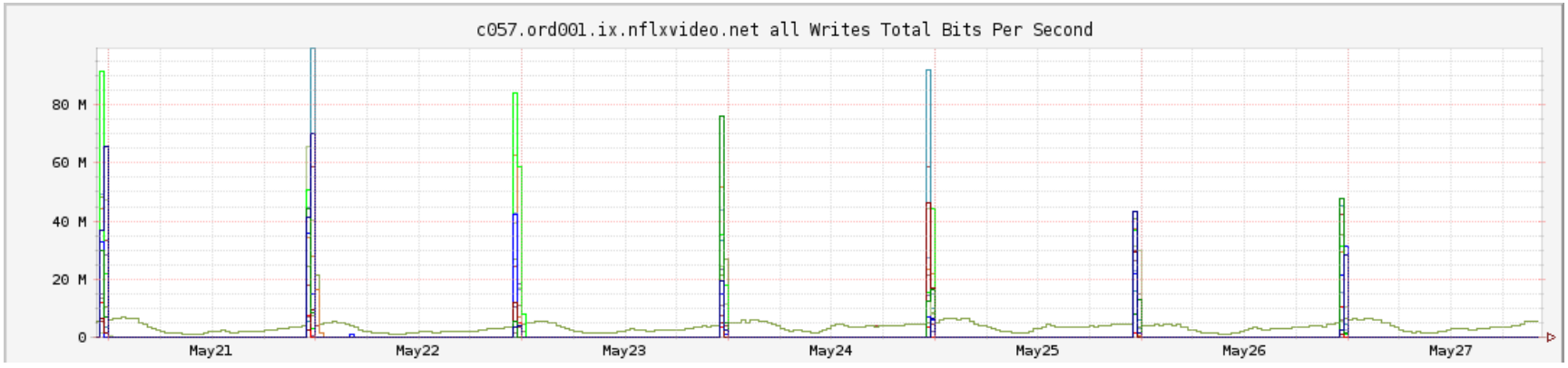
NETFLIX

Macro View

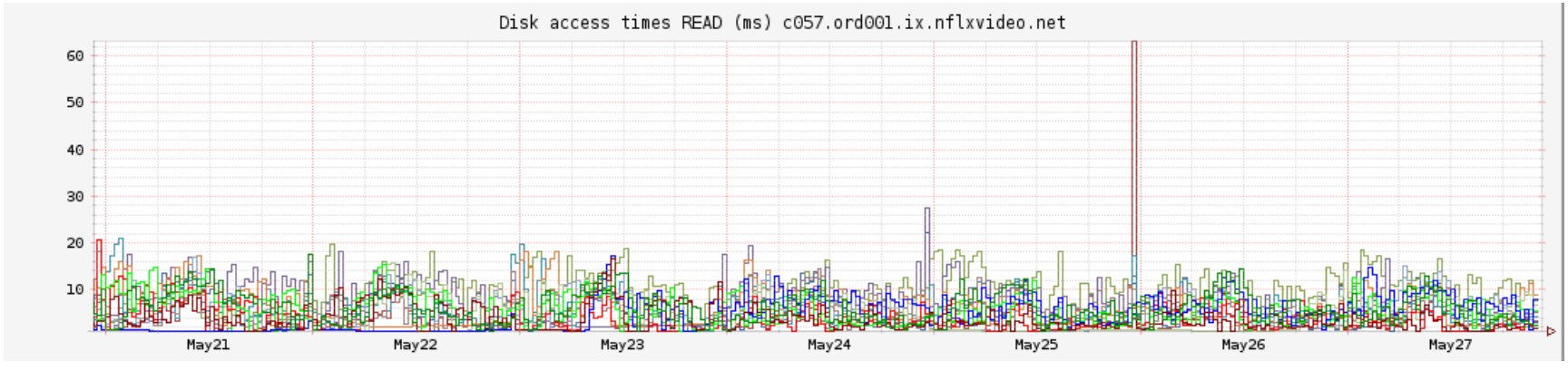


NETFLIX

Root Cause



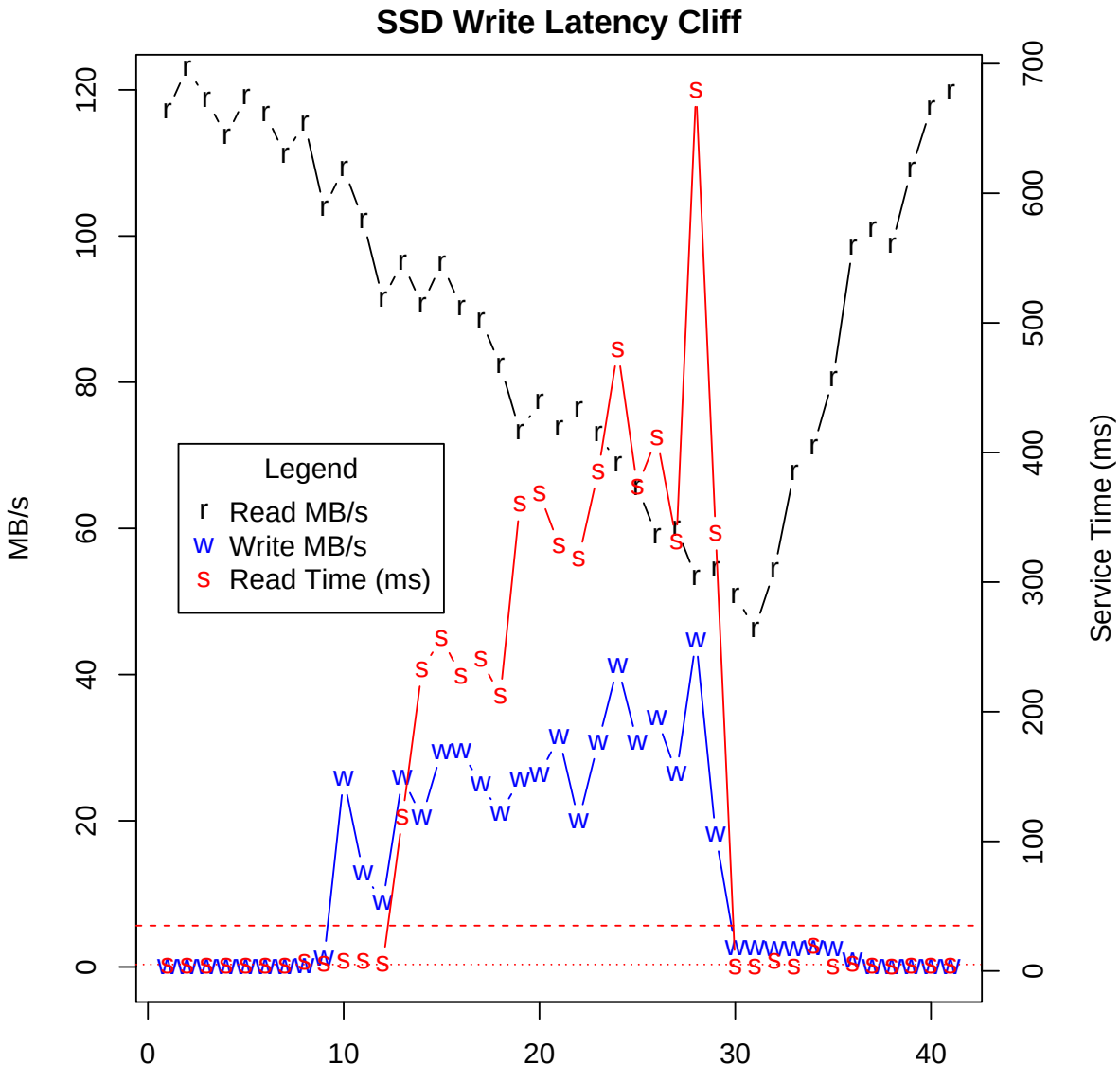
- High write workload



- Read latencies can spike

NETFLIX

One Graph View



Topics

- I/O Scheduler
- CAM
- I/O Stack
- SSDs
- NAND Physics
- Netflix Work load

NETFLIX

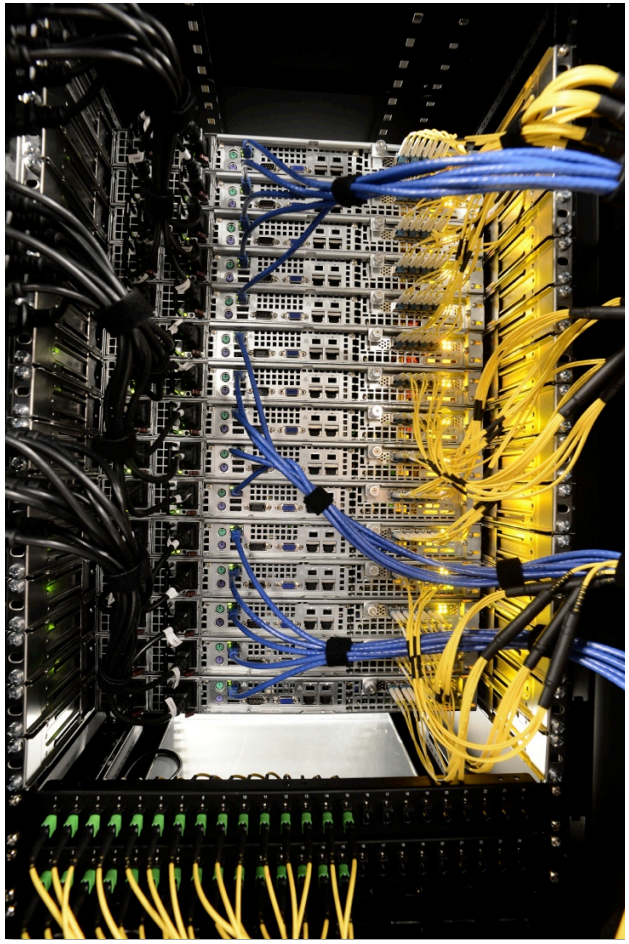
Outline

NETFLIX

Overview of Netflix's Open Connect Appliance (OCA)

NETFLIX

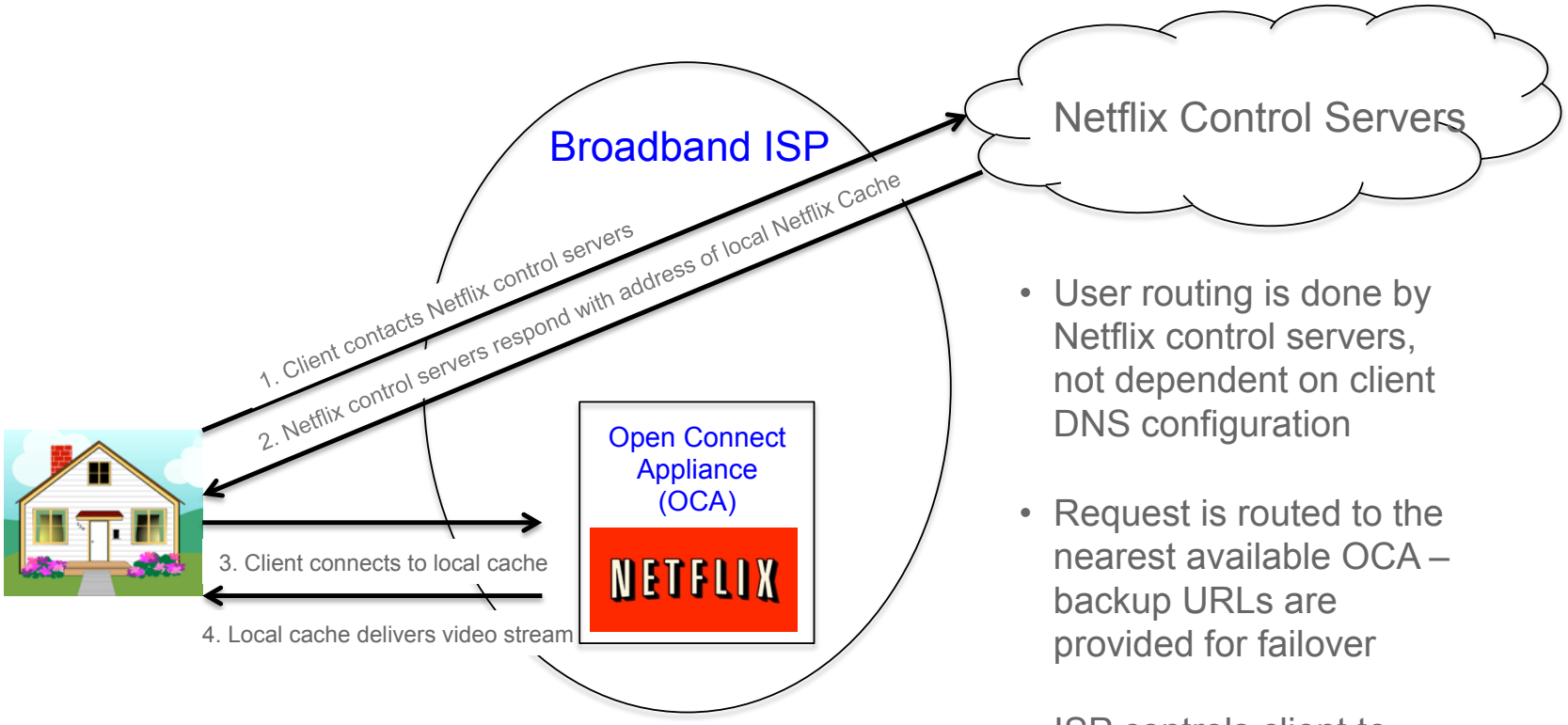
Netflix Open Connect Appliance (OCA)



Source: Netflix

NETFLIX

Netflix Open Connect Appliance (OCA)

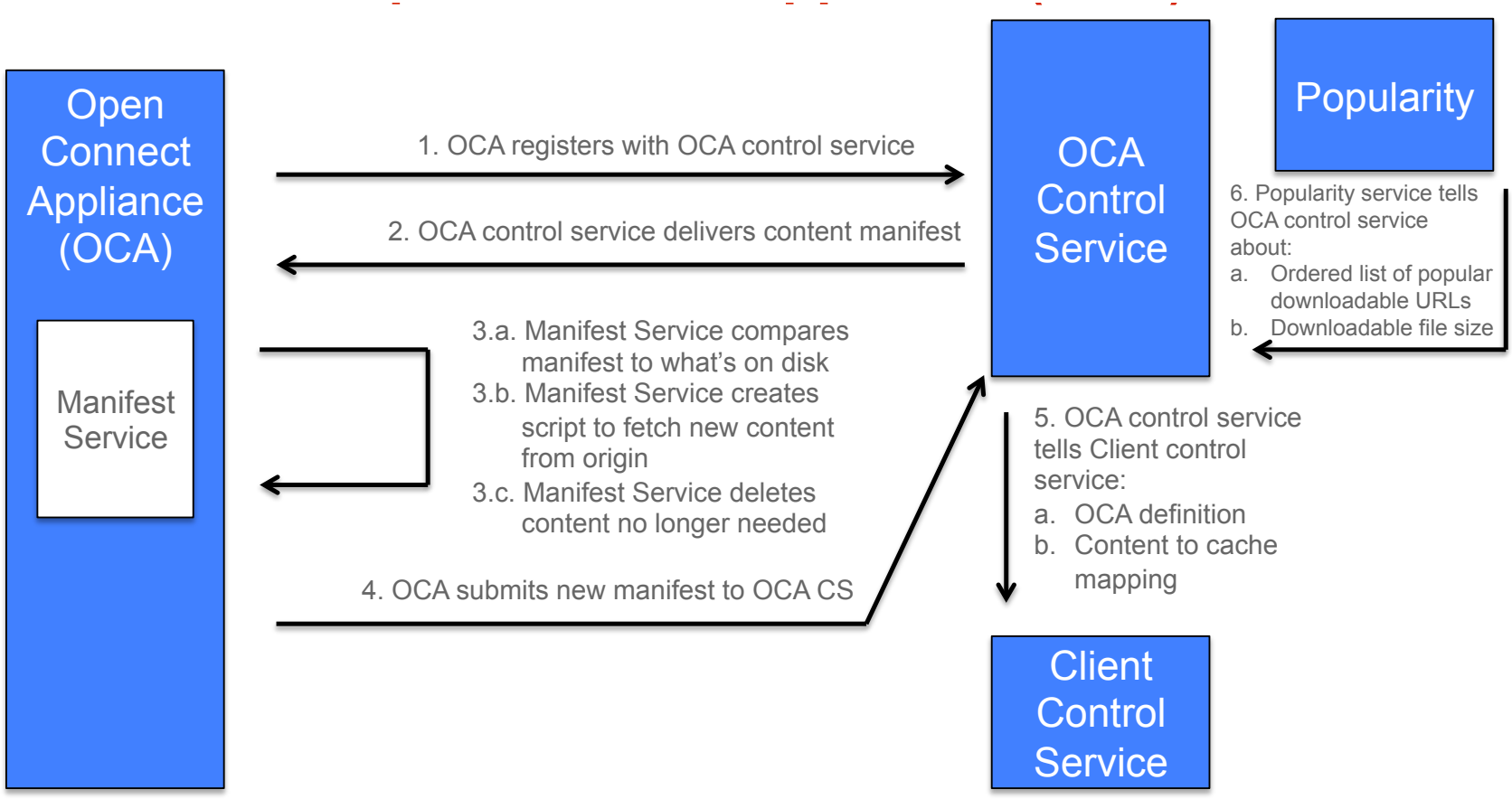


- User routing is done by Netflix control servers, not dependent on client DNS configuration
- Request is routed to the nearest available OCA – backup URLs are provided for failover
- ISP controls client to OCA mapping/clustering/ failover via BGP

Source: Netflix

NETFLIX

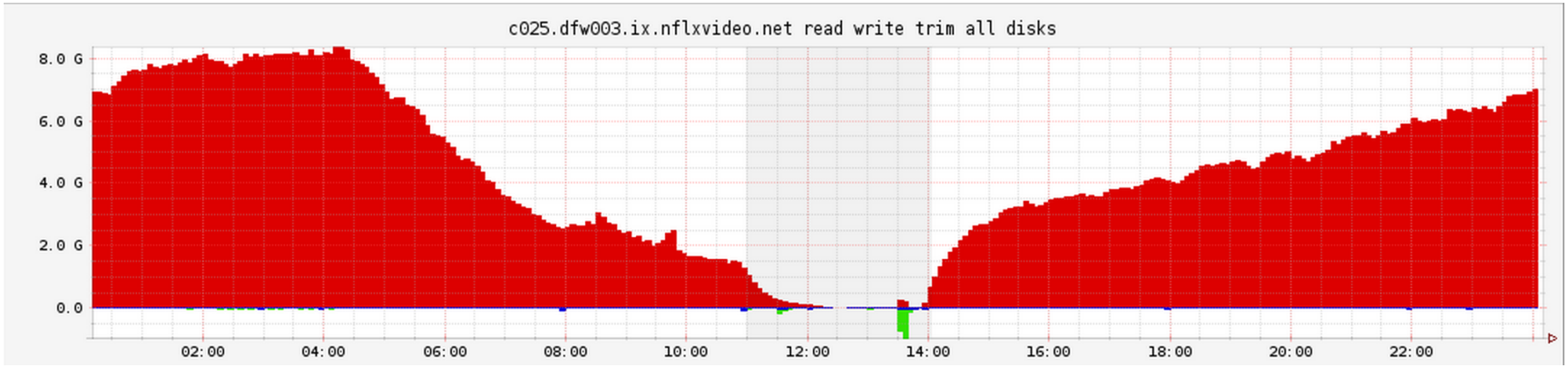
Netflix Open Connect Appliance (OCA)



Source: Netflix

NETFLIX

Typical Disk Activity (1 day)



- Lots of read traffic
- Little write traffic
- 100:1 read:write or more ratio
- "Fill Window" needed to avoid problem

NETFLIX

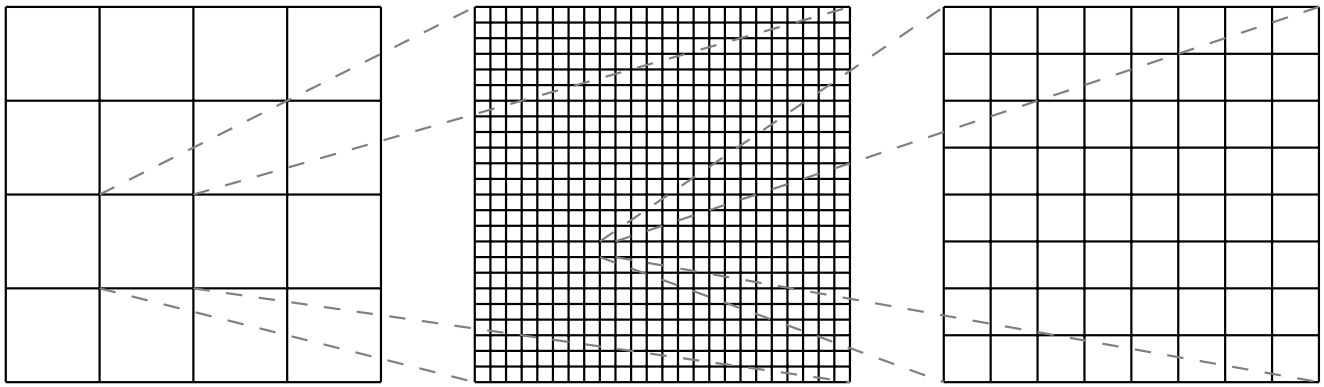
NAND Flash

A Brief Primer on Why and How NAND Flash "Presents Challenges" and How SSDs Hide That Through Lies and Deceit

Source: `http://pretty-little-liars.wikia.com/wiki/Season_5`

NETFLIX

NAND Geometry



16 chips/package
CE Parallel

4000 blocks/chip
Erase unit

500 pages/block
Read / Write unit

NAND typical geometry

Pages are 4kiB-32kiB (typically 16kiB) plus OOB

NETFLIX

NAND Limitations

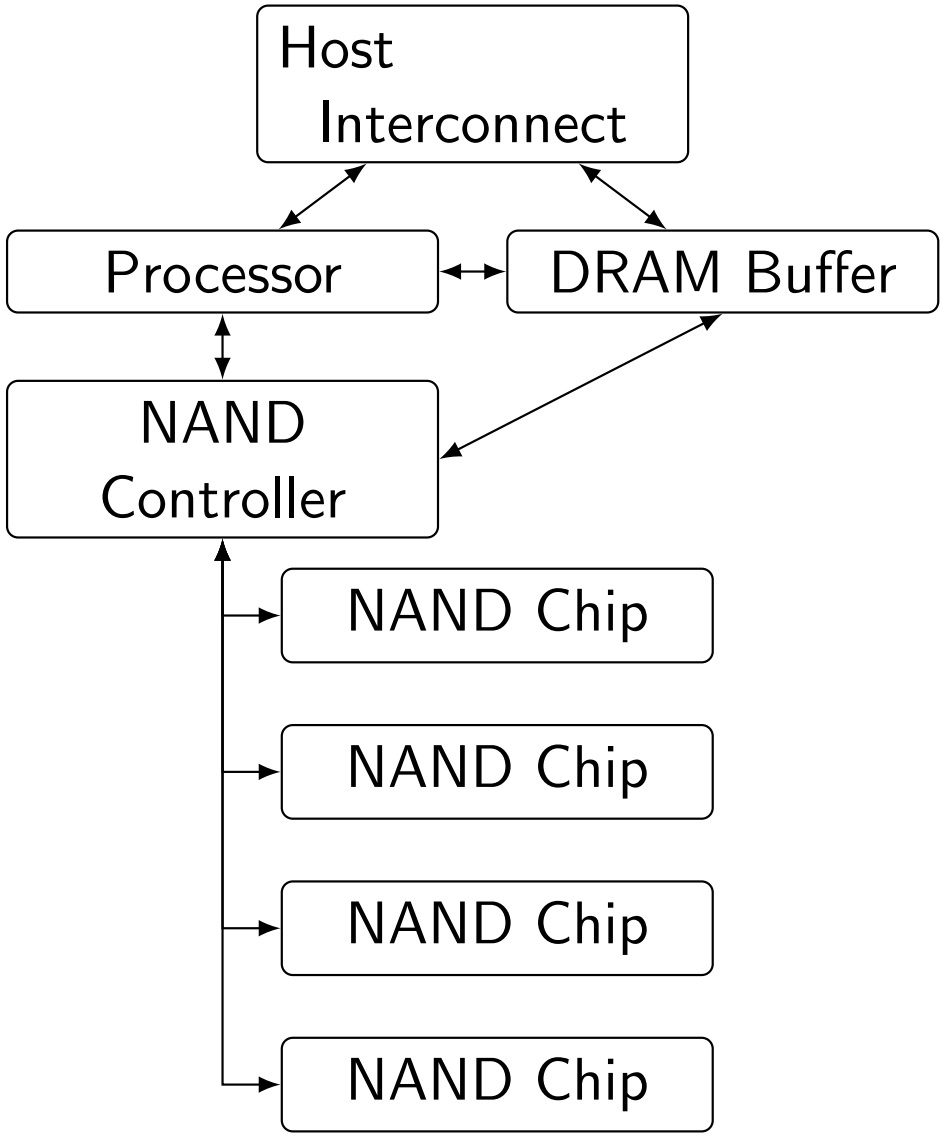
- Read page at a time
- Write page at a time
- Write pages in order
- Erase entire block
- Single Duplex, No queuing
- Low endurance (~3k P/E cycles)
- Unreliable Media (ECC / LDPC needed)
- Parallelism through Banks / CE lines
- NAND Flash and FreeBSD

  `https://www.youtube.com/watch?v=lj0XAE6C6-k`

  `https://people.freebsd.org/~imp/bsdcan2014.pdf`

NETFLIX

Typical SSD / NVMe



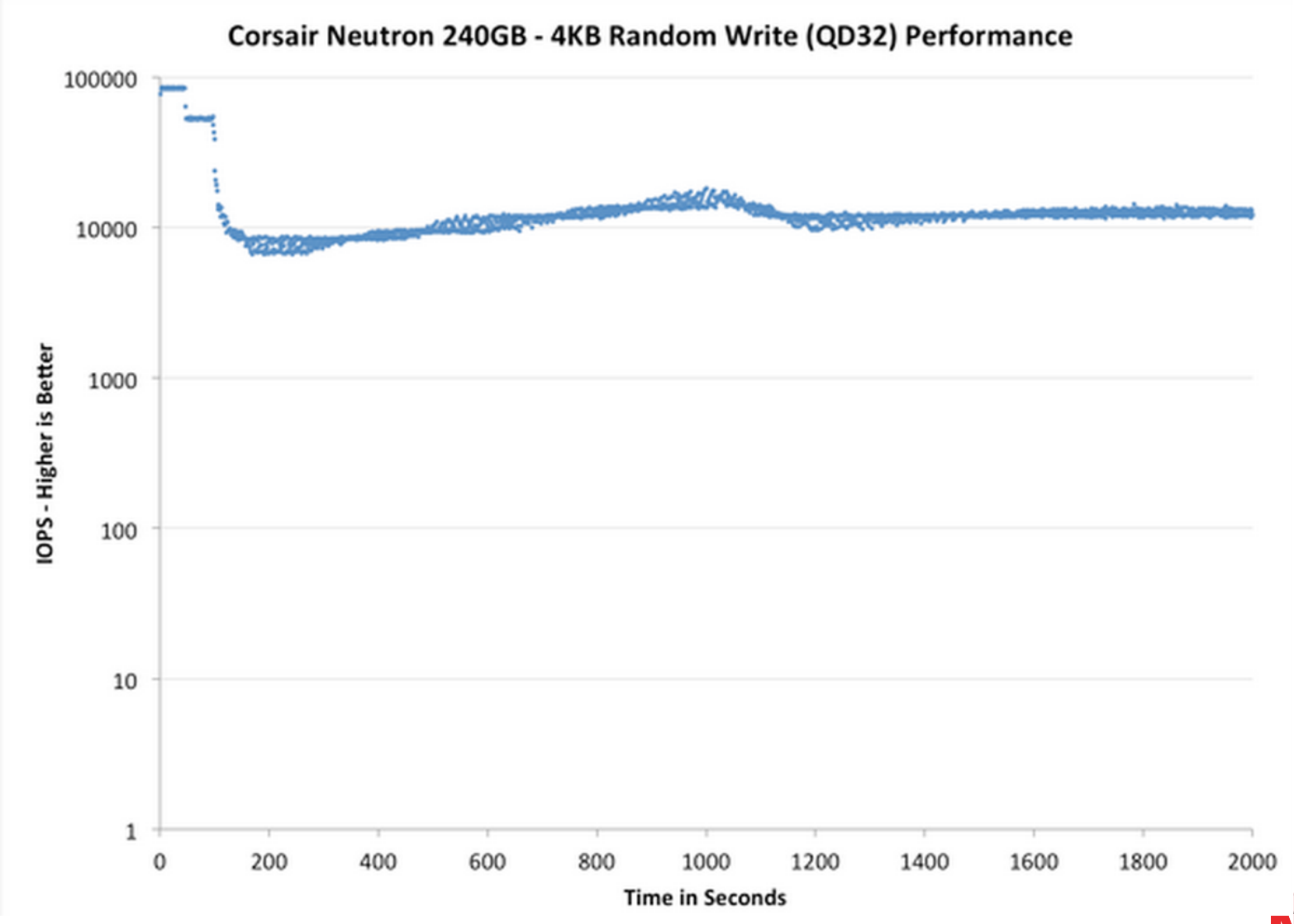
Generalized block diagram of flash storage device.

SSD / NVME Firmware

- Flash Translation Layer (FTL)
  - LBA to PA translation
  - Metadata for log / NAND
- Wear Leveling
  - Where to start writing at block boundary
  - Which block to garbage collect
- Reliability
  - Retention (data too old or read too much)
  - Wear out (block too worn with high RBER)
  - Program / Erase error processing
- Garbage Collection
  - Moves data forward
  - Extra reads and writes (Write Amplification)
  - Can affect performance

NETFLIX

Effects of Garbage Collection

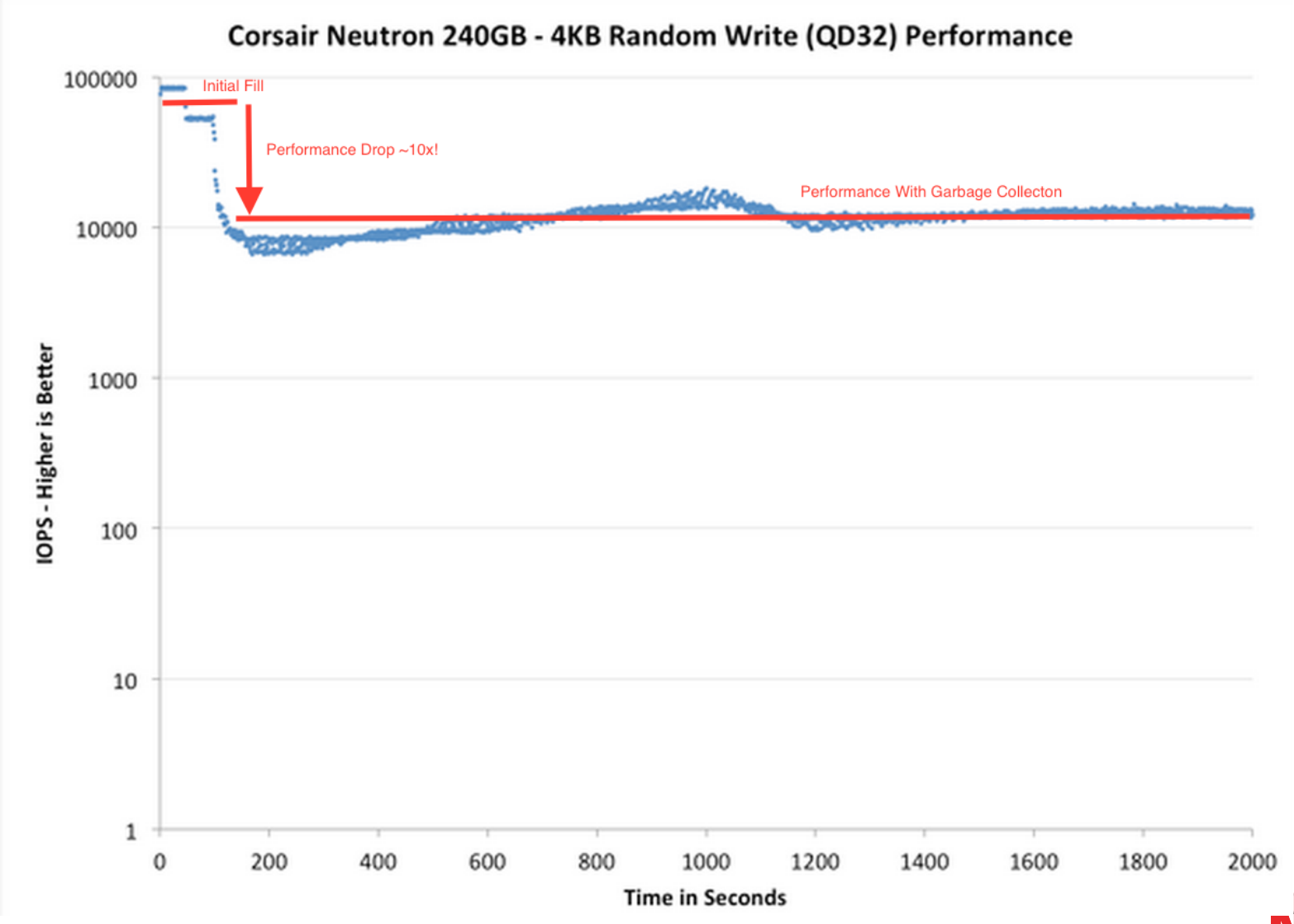


Source

http://www.howtogeek.com/165542/

NETFLIX

Effects of Garbage Collection

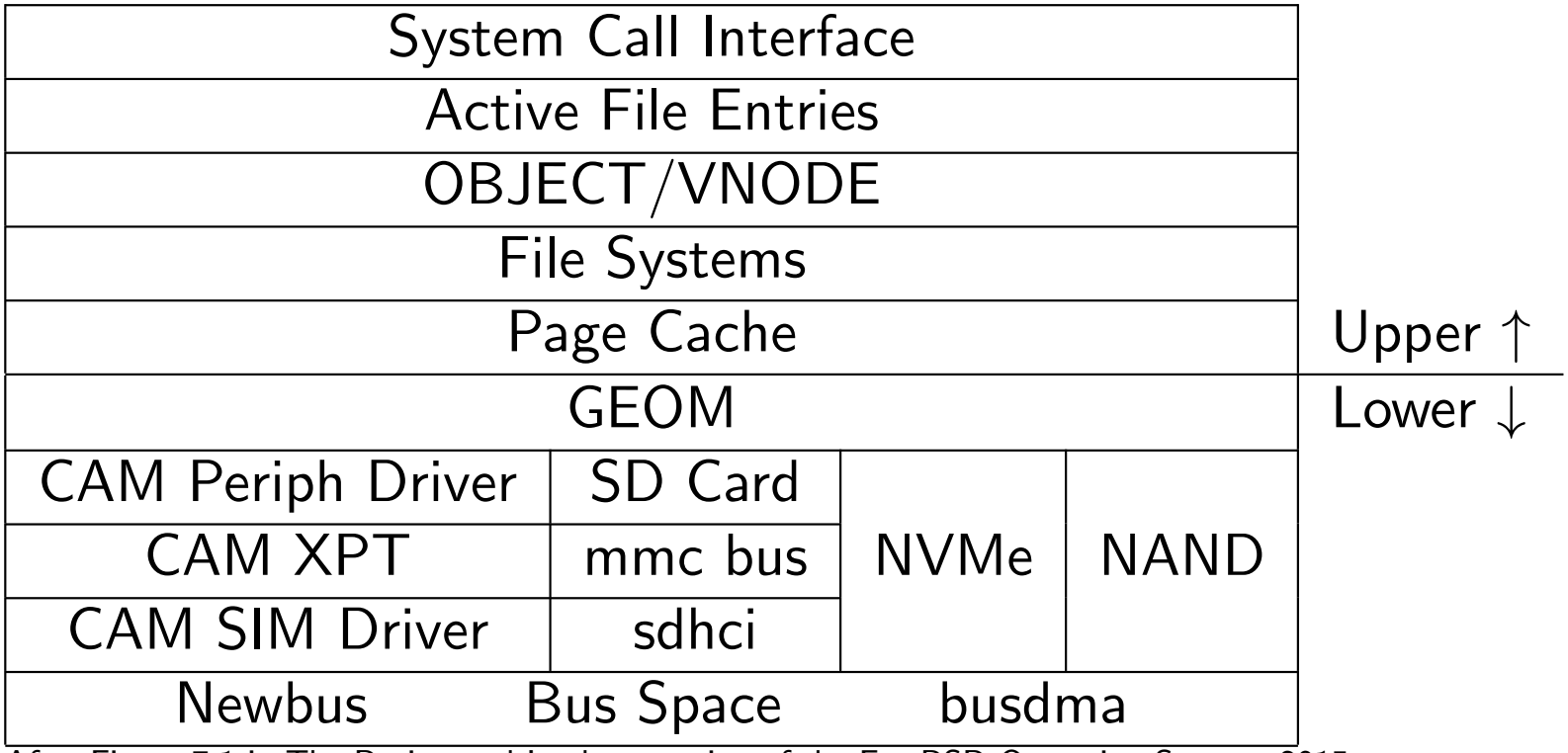


Source

http://www.howtogeek.com/165542/

NETFLIX

FreeBSD I/O Stack



After Figure 7.1 in The Design and Implementation of the FreeBSD Operating System, 2015.

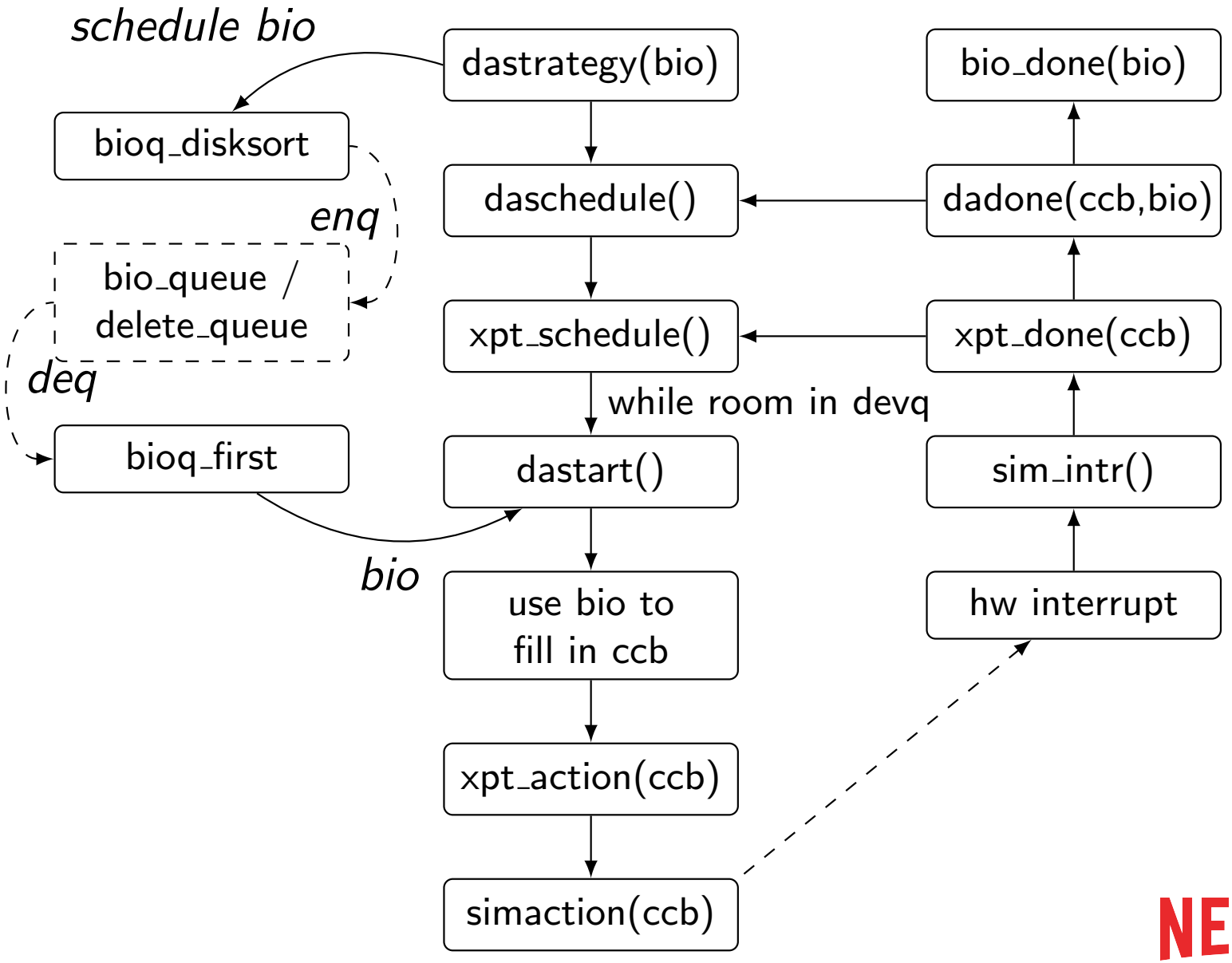
NETFLIX

FreeBSD I/O Stack

- Upper half of I/O Stack focus of VM system
  - Buffer cache
  - Memory mapped files / devices
  - Loosely coupled user actions to device action
- GEOM handles partitioning, compression, encryption
  - Filters data (compression, encryption)
  - Muxes Many to one (partitioning)
  - Muxes One to Many (striping / RAID)
  - Limited Scheduling
- CAM handles queuing
  - Shapes flows to device
  - Limits requests to number of slots
  - Enforces rules (eg tagged vs non-tagged)
  - Multiplexes shared resources between devices

NETFLIX

CAM – Data I/O data path through CAM



NETFLIX

Outline

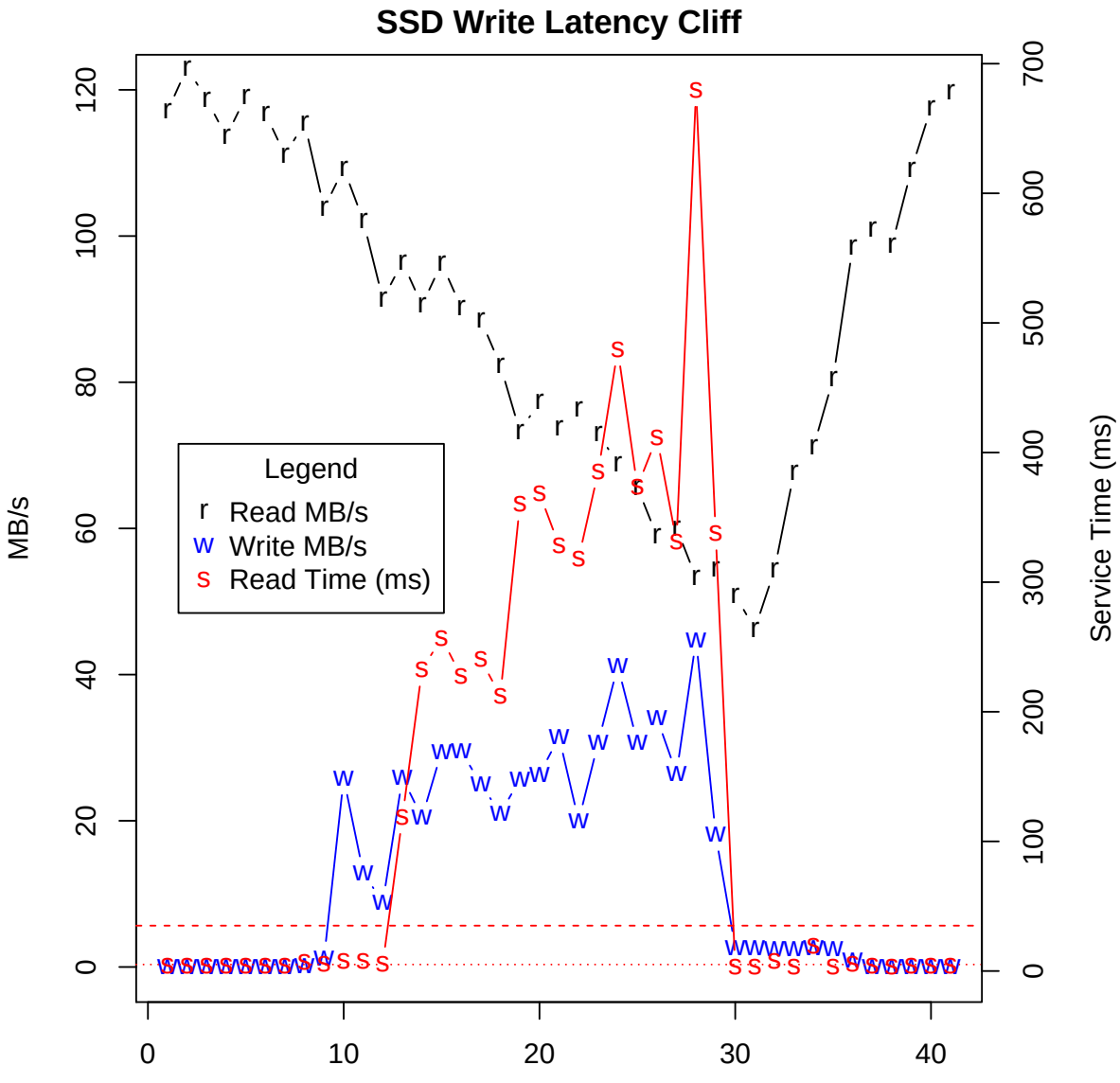
NETFLIX

FreeBSD Default I/O Scheduler

- No differentiation of I/O (except `BIO_DELETE`)
- Implements two ordering policies
  - In order
  - elevator
- Implemented in the CAM PERIPH drivers
  - Most flexible
  - Duplicated code
  - Partial lie: SIM drivers also involved (NCQ)
- Generally performs well for well behaved devices
  - SSDs break the rules
  - Different SSDs break different rules
  - Assumed cost symmetry often not the case

NETFLIX

Another Look



Houston, we have a problem

NETFLIX

Netflix I/O Scheduler Theory

- Restricting write rate reduces write amp disturbance
- Fewer concurrent writes leaves more banks for reads
- Elevated latency OK within limits
- Lowest later needed since PERIPH knows about device
  - GEOM layer too high
  - GEOM filters requests, but can't force PERIPH

NETFLIX

I/O Scheduler Changes

- Create abstract interface to scheduler
- Convert da and ada PERIPHs to new interface
- Make sure no regressions

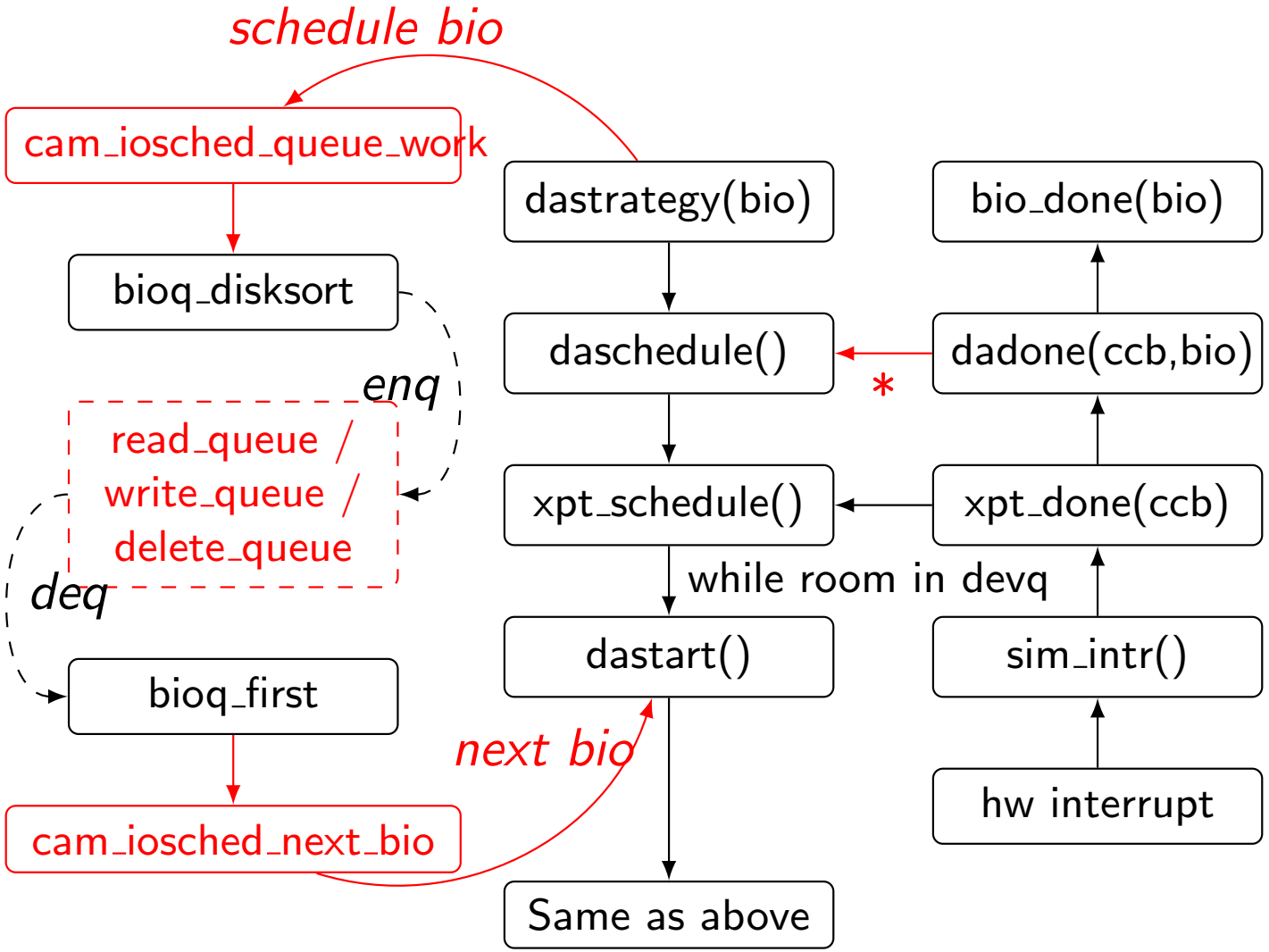
NETFLIX

Netflix I/O Scheduler Changes

- Separate read, write and delete queues
- Lots of statistics
- Ability to limit number of I/O in device at a time
- Adjustments needed for scheduling

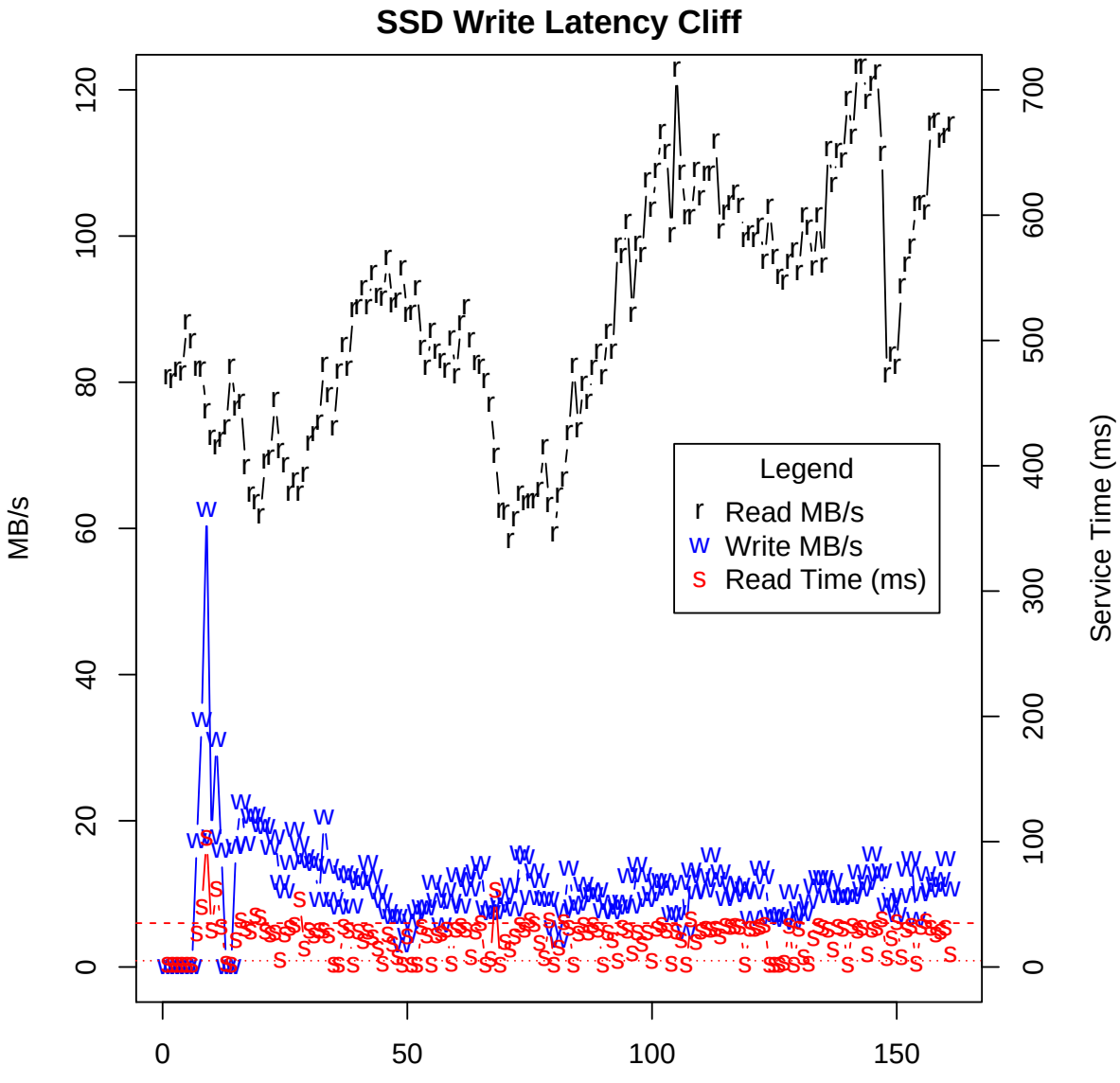
NETFLIX

Netflix I/O Scheduler Changes



Changes to default scheduler in red.

Results



NETFLIX

Outline

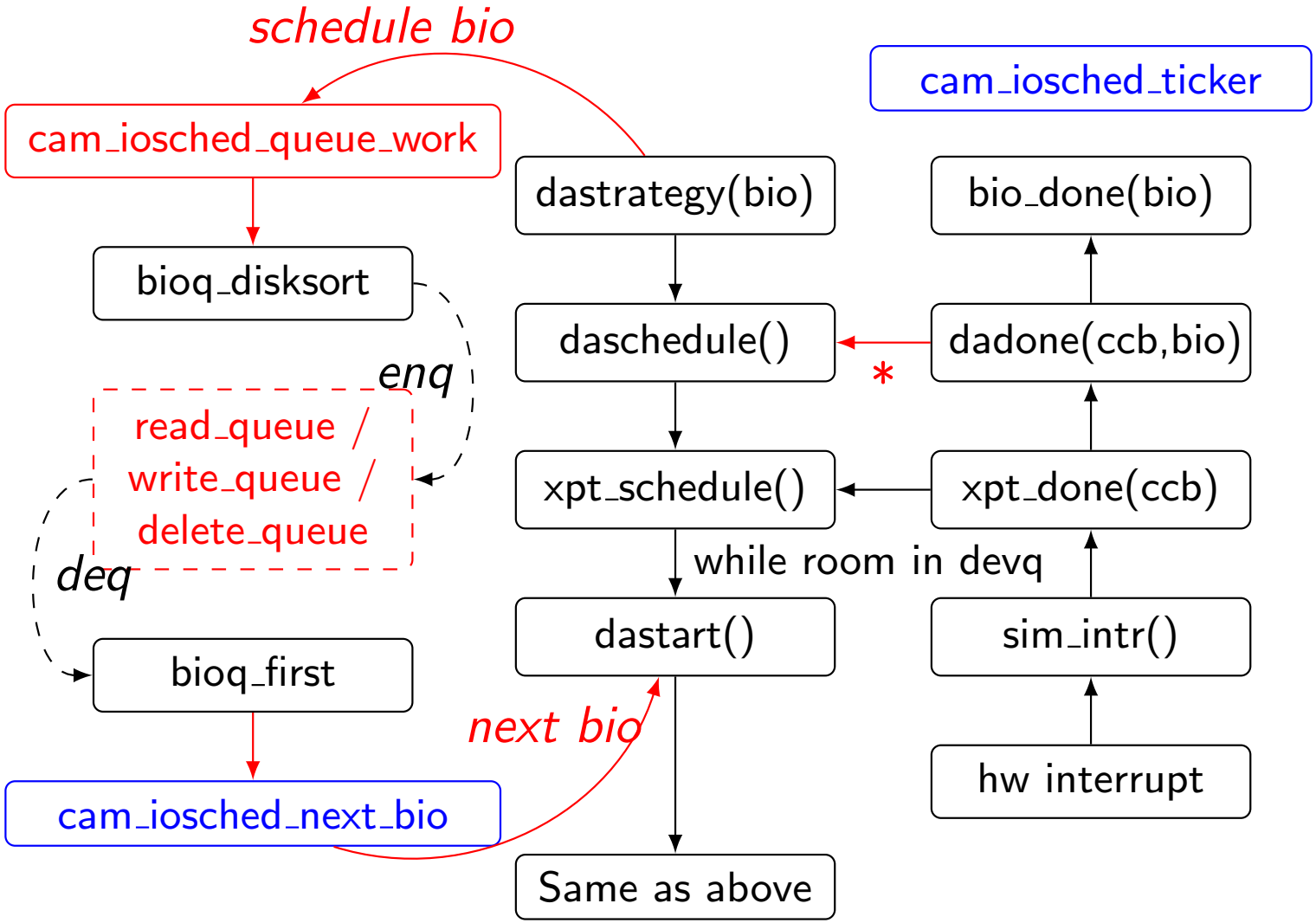
NETFLIX

Added Features

- Added Bandwidth and IOPS limits
- Added timeout to implement quanta scheduling
- Added dynamic steering of limits

NETFLIX

Code Changes



Changes to default scheduler in red. Changes to prior I/O scheduler in blue.

NETFLIX

Recent issues

- Large quanta produce large latencies
- Rate limited drives report 100 percent busy
- Dynamic loop not yet tuned

NETFLIX

Questions

Questions?
Comments?

Warner Losh
wlosh@netflix.com
imp@FreeBSD.org

http://people.freebsd.org/~imp/bsdcon2015/iosched-slides.pdf
http://people.freebsd.org/~imp/bsdcon2015/paper.pdf
http://people.freebsd.org/~imp/asiabsdcon2015/works

NETFLIX