Linux Performance 2018

PERCONA LIVE

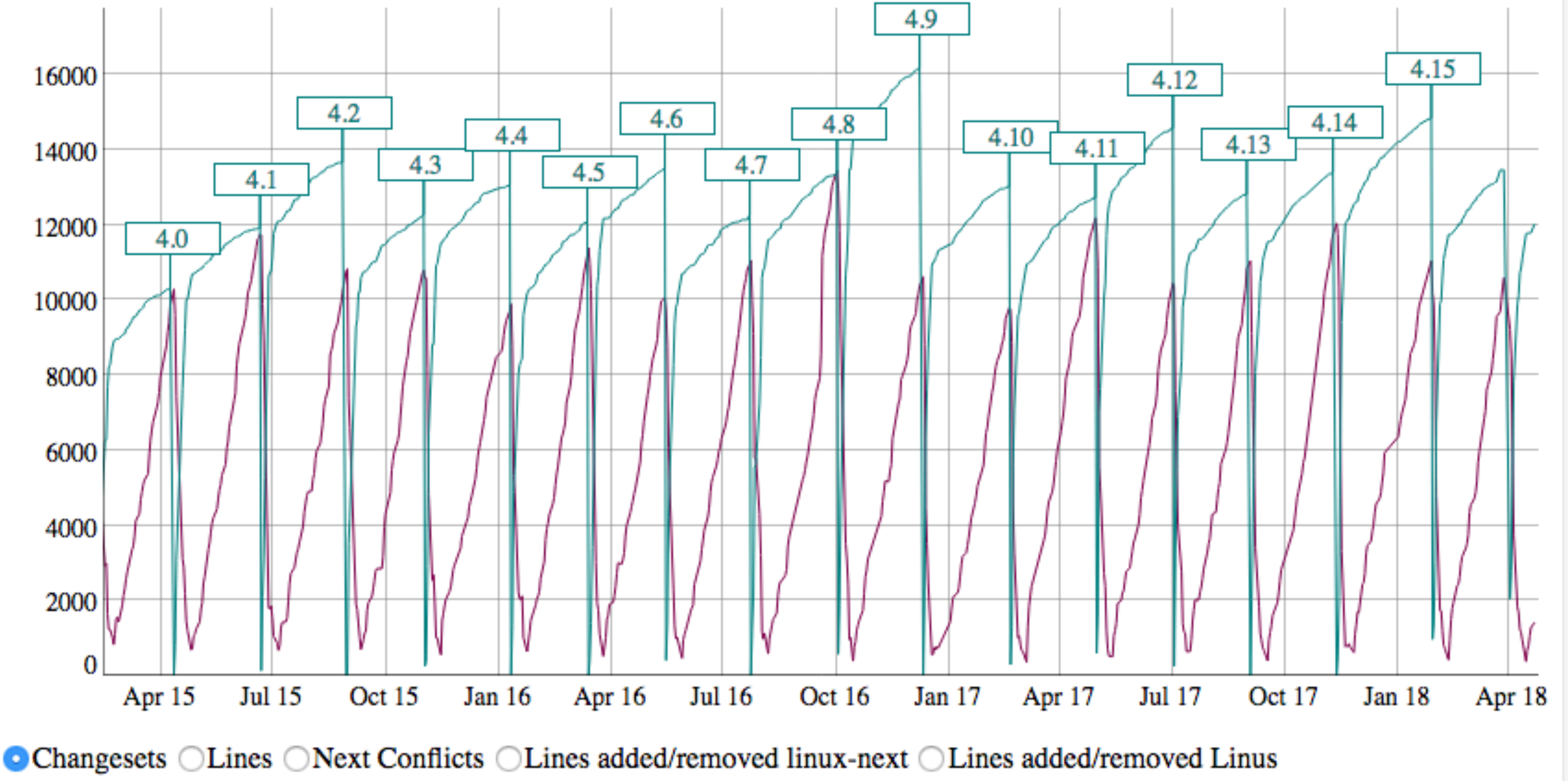
OPEN SOURCE DATABASE CONFERENCE

Apr 2018

Brendan Gregg

Senior Performance Architect

NETFLIX



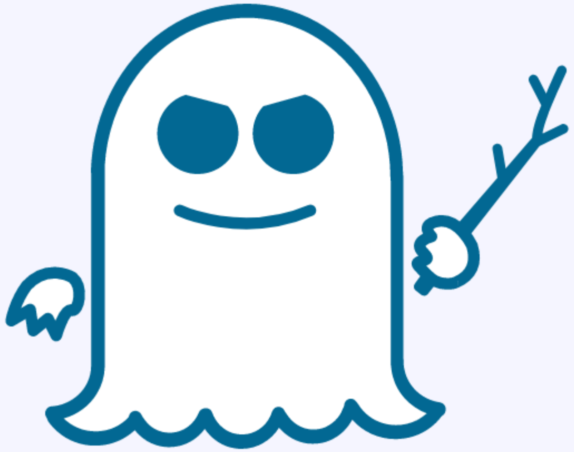
● Changesets ○ Lines ○ Next Conflicts ○ Lines added/removed linux-next ○ Lines added/removed Linus

http://neuling.org/linux-next-size.html

Post frequency:

4 per year		https://kernelnewbies.org/Linux_4.15
4 per week	LWN .net News from the source	https://lwn.net/Kernel/
400 per day	LKML	http://vger.kernel.org/vger-lists.html#linux-kernel

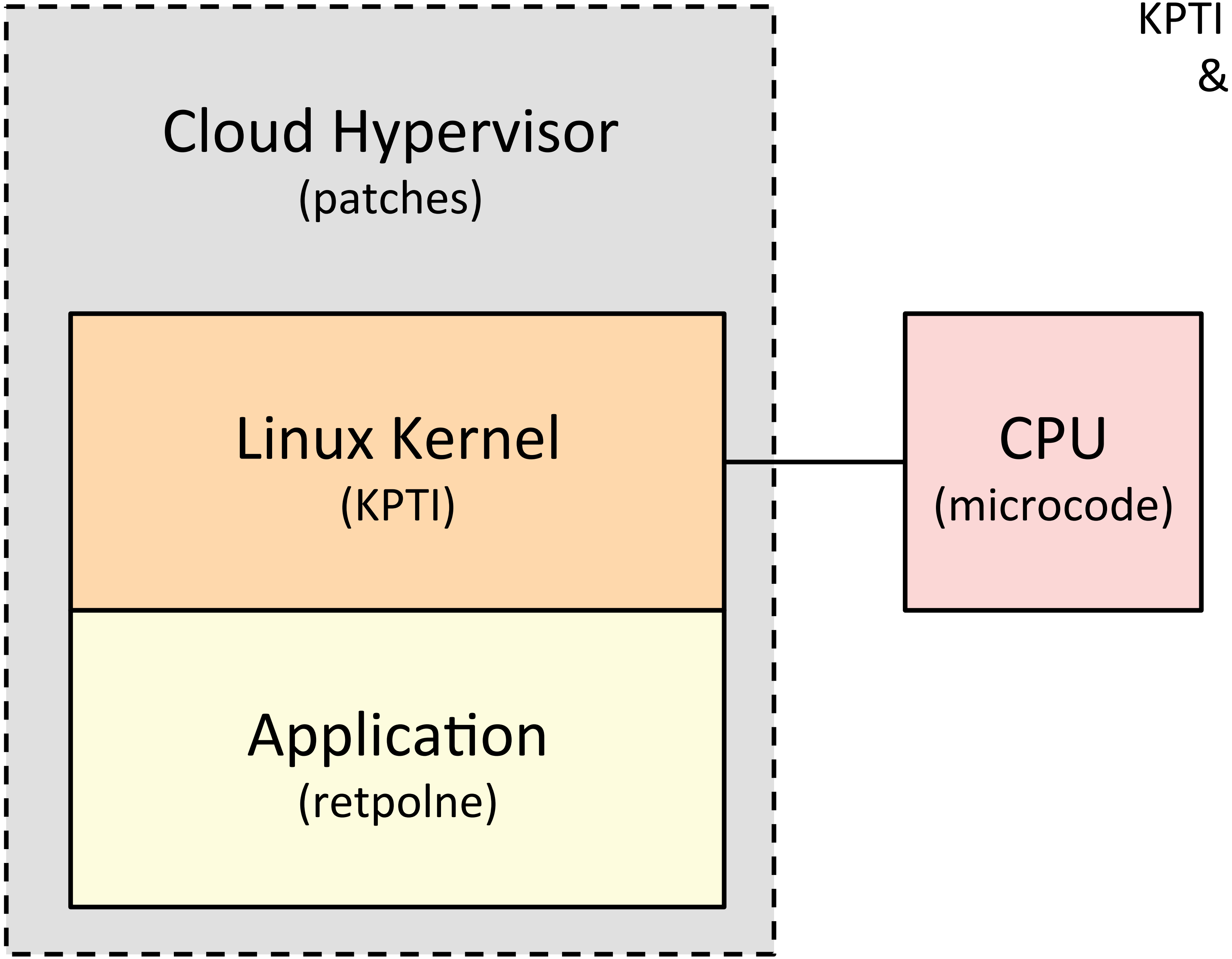
Meltdown



Spectre

https://meltdownattack.com/

KPTI Linux 4.15
& backports



Cloud Hypervisor
(patches)

Linux Kernel
(KPTI)

Application
(retpolne)

CPU
(microcode)

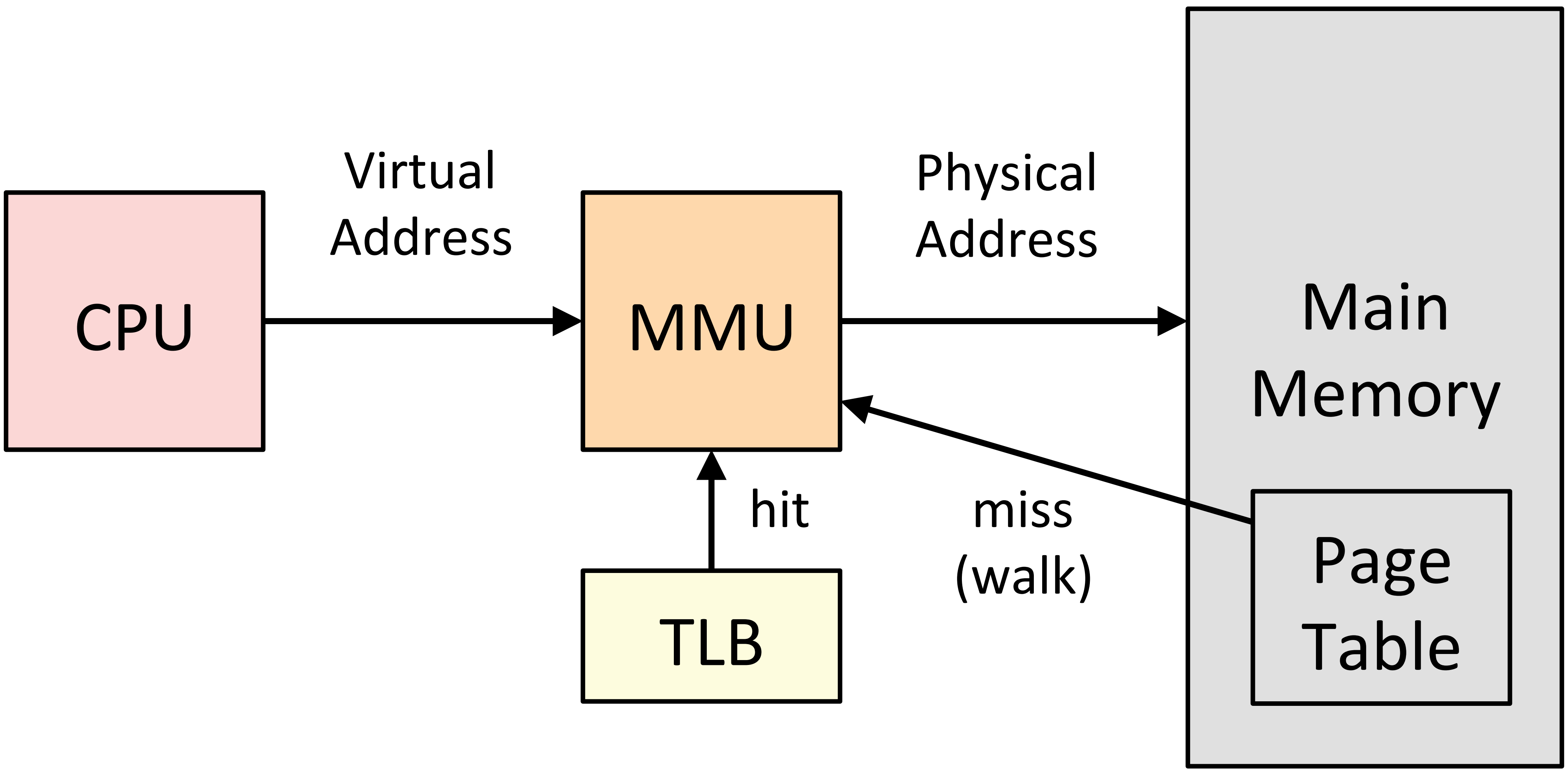
Server A: 31353 MySQL queries/sec

```
serverA# mpstat 1
Linux 4.14.12-virtual (bgregg-c5.9xl-i-xxx)     02/09/2018      _x86_64_        (36 CPU)
01:09:13 AM  CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest  %gnice   %idle
01:09:14 AM  all   86.89    0.00   13.08    0.00    0.00    0.00    0.00    0.00    0.00    0.03
01:09:15 AM  all   86.77    0.00   13.23    0.00    0.00    0.00    0.00    0.00    0.00    0.00
01:09:16 AM  all   86.93    0.00   13.02    0.00    0.00    0.00    0.03    0.00    0.00    0.03
[...]
```

Server B: 22795 queries/sec (27% slower)

```
serverB# mpstat 1
Linux 4.14.12-virtual (bgregg-c5.9xl-i-xxx)     02/09/2018      _x86_64_        (36 CPU)
01:09:44 AM  CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest  %gnice   %idle
01:09:45 AM  all   82.94    0.00   17.06    0.00    0.00    0.00    0.00    0.00    0.00    0.00
01:09:46 AM  all   82.78    0.00   17.22    0.00    0.00    0.00    0.00    0.00    0.00    0.00
01:09:47 AM  all   83.14    0.00   16.86    0.00    0.00    0.00    0.00    0.00    0.00    0.00
[...]
```

Linux KPTI patches for Meltdown flush the Translation Lookaside Buffer

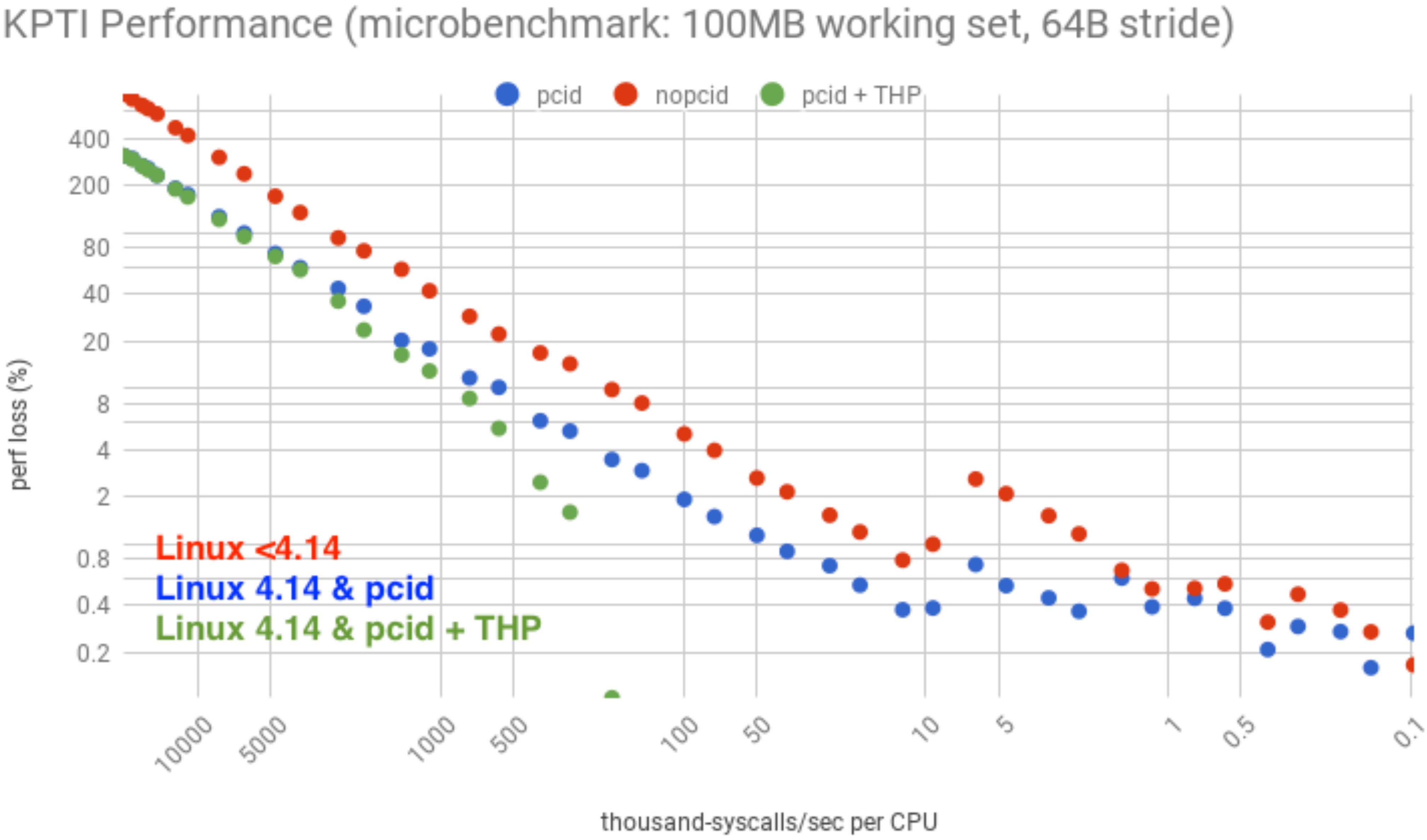


Server A: TLB miss walks 3.5%

```
serverA# ./tlbstat 1
K_CYCLES    K_INSTR       IPC DTLB_WALKS ITLB_WALKS K_DTLBCYC  K_ITLBCYC  DTLB% ITLB%
95913667    99982399     1.04 86588626   115441706  1507279    1837217     1.57  1.92
95810170    99951362     1.04 86281319   115306404  1507472    1842313     1.57  1.92
95844079    100066236    1.04 86564448   115555259  1511158    1845661     1.58  1.93
95978588    100029077    1.04 86187531   115292395  1508524    1845525     1.57  1.92
[...]
```

Server B: TLB miss walks 19.2% (16% higher)

```
serverB# ./tlbstat 1
K_CYCLES    K_INSTR       IPC DTLB_WALKS ITLB_WALKS K_DTLBCYC  K_ITLBCYC  DTLB% ITLB%
95911236    80317867     0.84 911337888  719553692  10476524   7858141    10.92  8.19
95927861    80503355     0.84 913726197  721751988  10518488   7918261    10.96  8.25
95955825    80533254     0.84 912994135  721492911  10524675   7929216    10.97  8.26
96067221    80443770     0.84 912009660  720027006  10501926   7911546    10.93  8.24
[...]
```

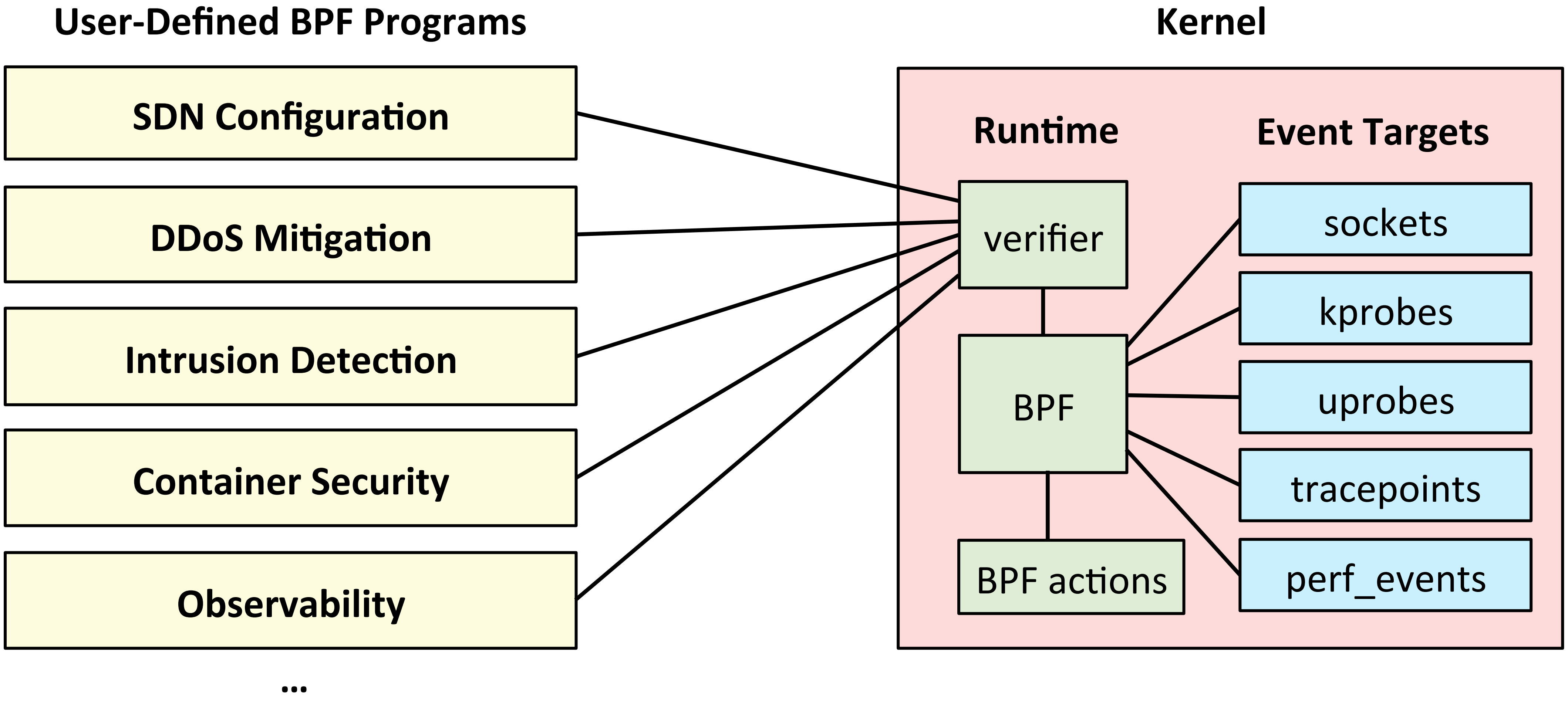


http://www.brendangregg.com/blog/2018-02-09/kpti-kaiser-meltdown-performance.html

Enhanced BPF

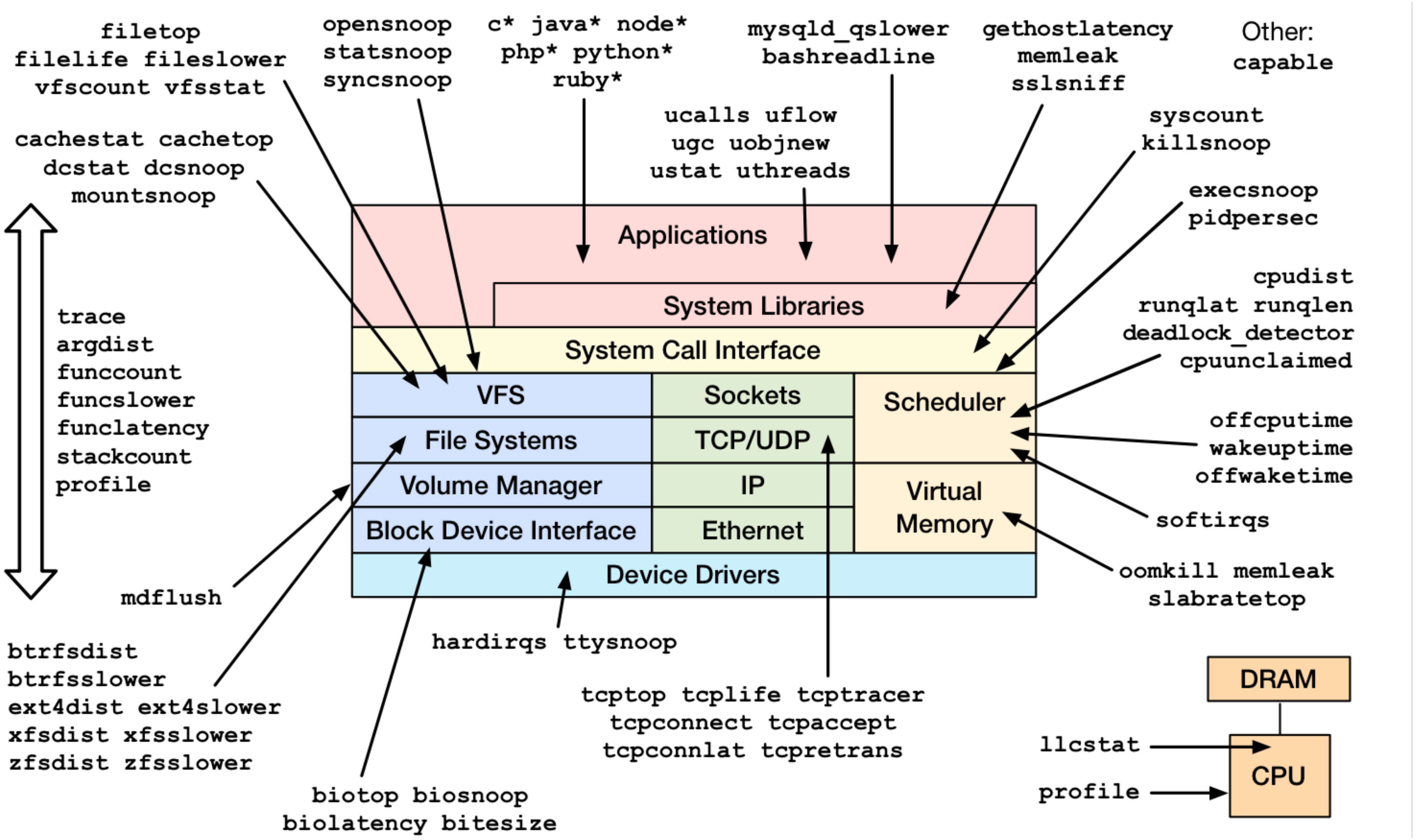
Linux 4.*

also known as just "BPF"



eBPF bcc

Linux 4.4+



https://github.com/iovisor/bcc

Identify multimodal disk I/O latency and outliers with eBPF `biolatency`

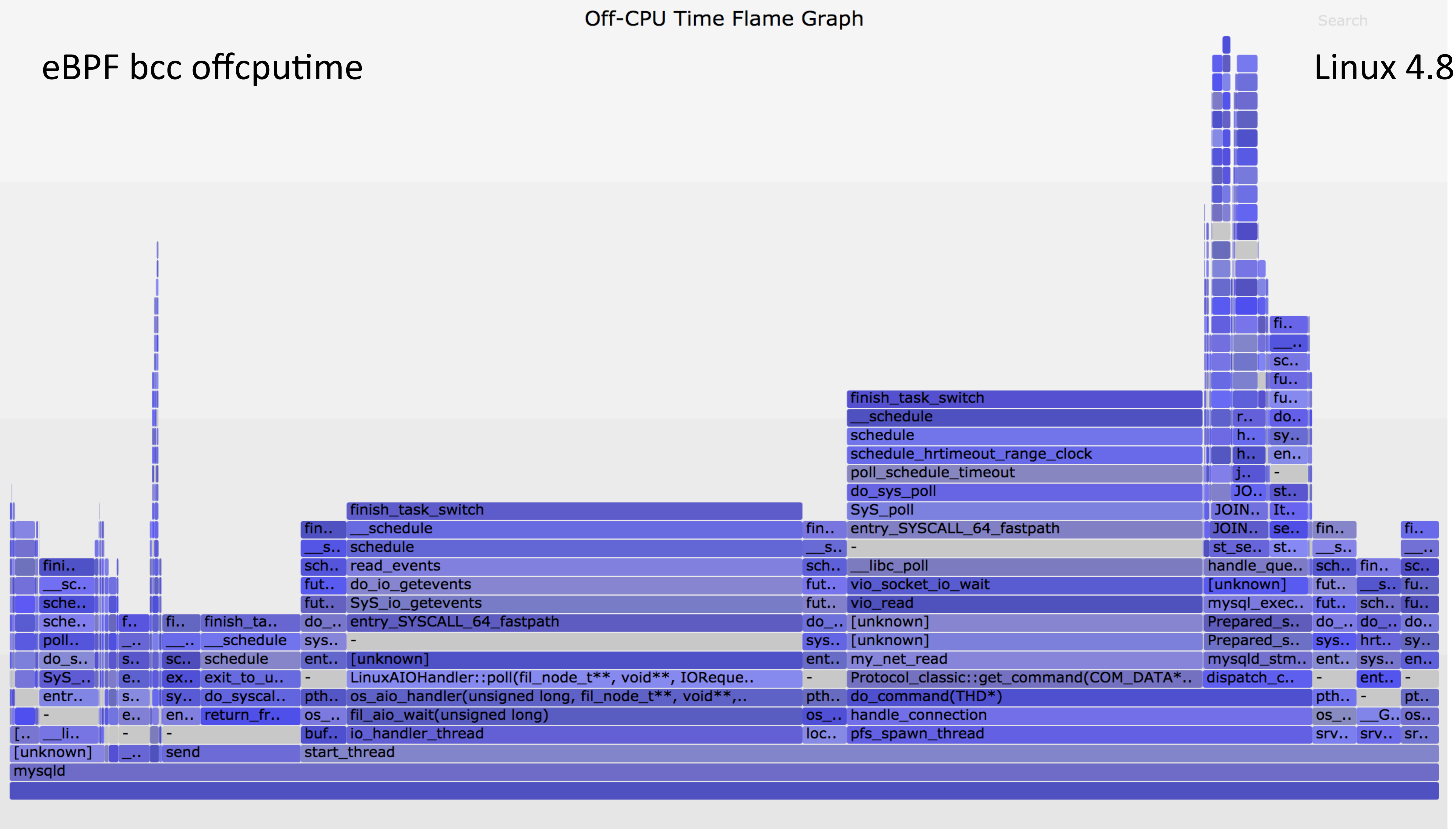
```
# biolatency -mT 10
Tracing block device I/O... Hit Ctrl-C to end.

19:19:04
     msecs               : count     distribution
         0 -> 1          : 238      |*********                               |
         2 -> 3          : 424      |*****************                       |
         4 -> 7          : 834      |*********************************       |
         8 -> 15         : 506      |********************                    |
        16 -> 31         : 986      |****************************************|
        32 -> 63         : 97       |***                                     |
        64 -> 127        : 7        |                                        |
       128 -> 255        : 27       |*                                       |

19:19:14
     msecs               : count     distribution
         0 -> 1          : 427      |*******************                     |
         2 -> 3          : 424      |******************                      |
[…]
```

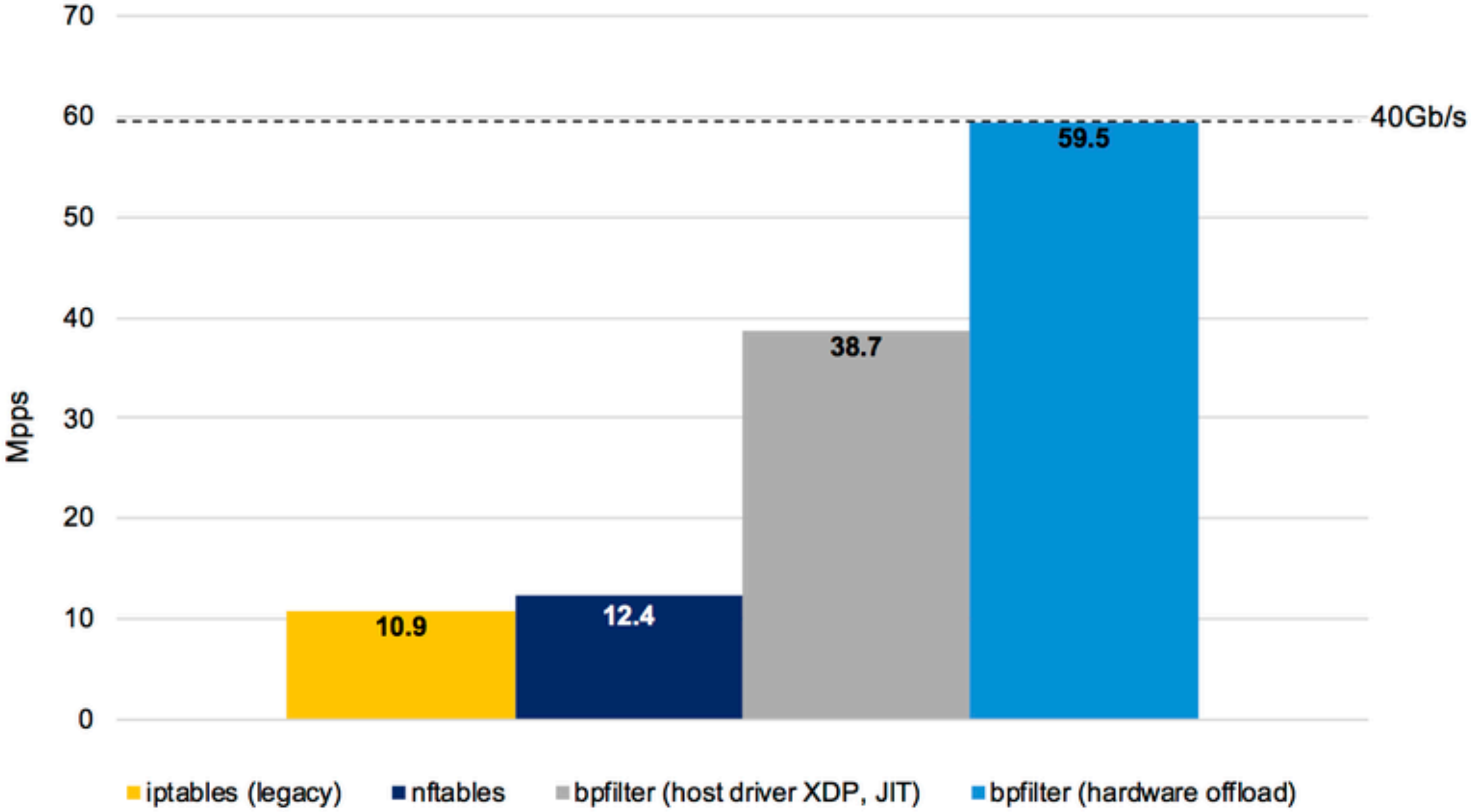
eBPF bcc offcputime

Linux 4.8+



eBPF XDP

Linux 4.8+



https://www.netronome.com/blog/frnog-30-faster-networking-la-francaise/

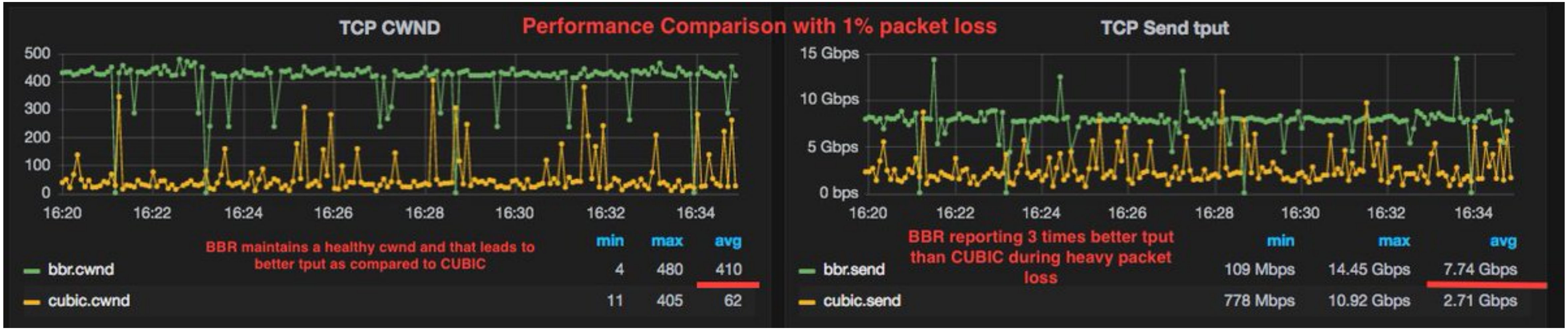
Linux 4.9

BBR

TCP congestion control algorithm

Bottleneck Bandwidth and RTT

1% packet loss: we see 3x better throughput



Performance Comparison with 1% packet loss

TCP CWND

BBR maintains a healthy cwnd and that leads to better tput as compared to CUBIC

	min	max	avg
bbr.cwnd	4	480	410
cubic.cwnd	11	405	62

TCP Send tput

BBR reporting 3 times better tput than CUBIC during heavy packet loss

	min	max	avg
bbr.send	109 Mbps	14.45 Gbps	7.74 Gbps
cubic.send	778 Mbps	10.92 Gbps	2.71 Gbps

https://twitter.com/amernetflix/status/892787364598132736

https://blog.apnic.net/2017/05/09/bbr-new-kid-tcp-block/ https://queue.acm.org/detail.cfm?id=3022184

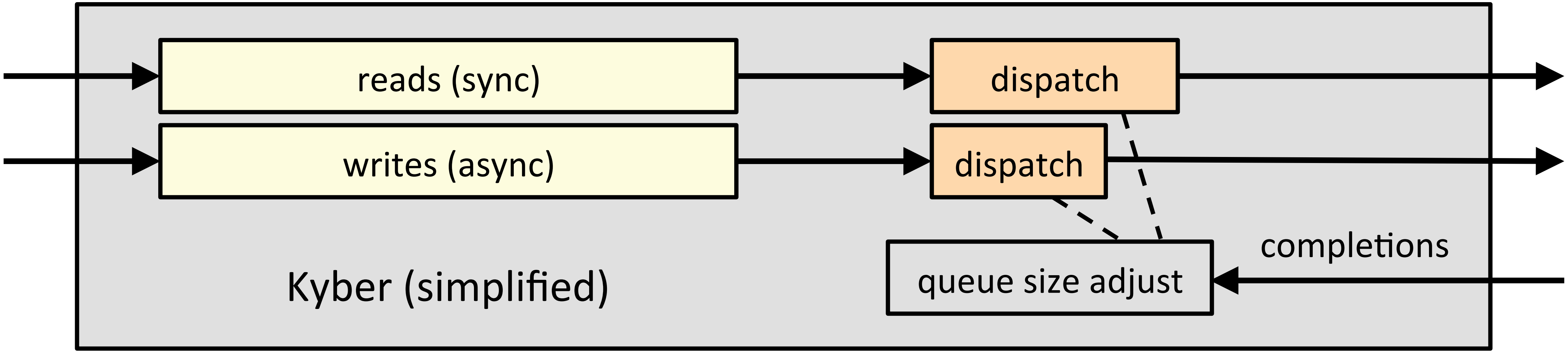
Linux 4.12

Kyber

Multiqueue block I/O scheduler

Tune target read & write latency

Up to 300x lower 99th latencies in our testing



https://lwn.net/Articles/720675/

More perf 4.4 - 4.16 (2016 - 2018)

Major features:

- TCP listener lockless (4.4)
- copy_file_range() (4.5)
- madvise() MADV_FREE (4.5)
- epoll multithread scalability (4.5)
- Kernel Connection Multiplexor (4.6)
- Writeback management (4.10)
- Hybrid block polling (4.10)
- BFQ I/O scheduler (4.12)
- Async I/O improvements (4.13)
- In-kernel TLS acceleration (4.13)
- Socket MSG_ZEROCOPY (4.14)
- Asynchronous buffered I/O (4.14)
- Longer-lived TLB entries with PCID (4.14)
- mmap MAP_SYNC (4.15)
- Software-interrupt context hrtimers (4.16)

Many minor improvements to:

- perf
- CPU scheduling
- futexes
- NUMA
- Huge pages
- Slab allocation
- TCP, UDP
- Drivers
- Processor support
- GPUs

Take Aways

1. Run latest

2. Browse major features

eg, https://kernelnewbies.org/Linux_4.15

Some Linux perf Resources

- http://www.brendangregg.com/linuxperf.html
- https://kernelnewbies.org/LinuxChanges
- https://lwn.net/Kernel
- https://github.com/iovisor/bcc
- http://blog.stgolabs.net/search/label/linux
- http://www.brendangregg.com/blog/2018-02-09/kpti-kaiser-meltdown-performance.html

OPEN SOURCE DATABASE CONFERENCE

PERCONA LIVE