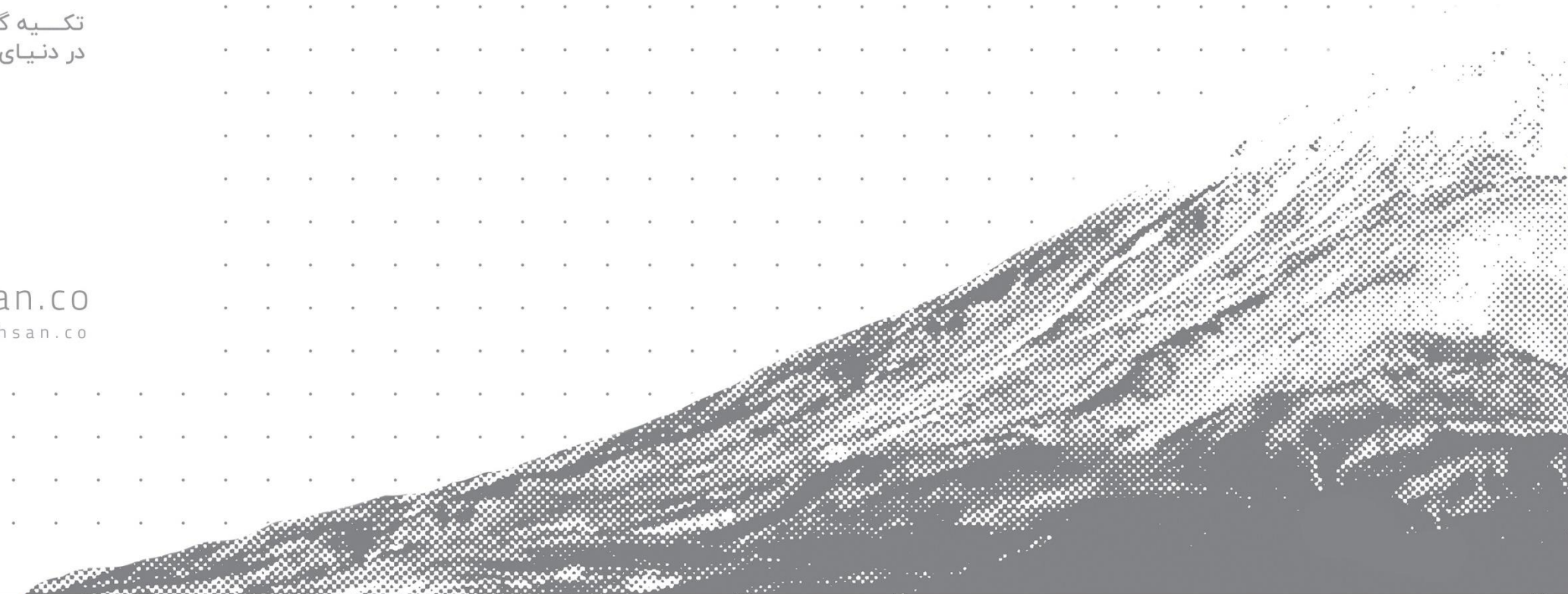
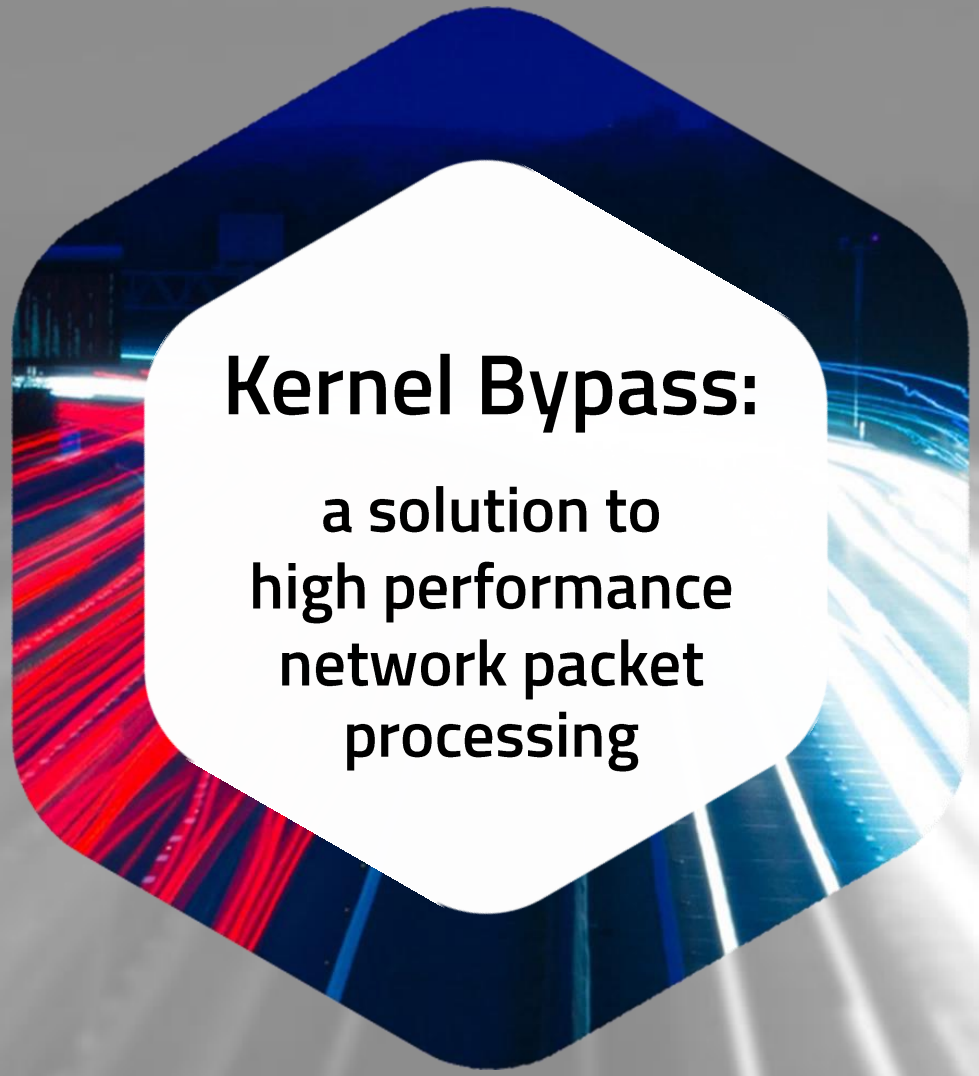
مهسان، تکیه‌گاه شما
در دنیای هوشمند

مهسان

تکیه گاه شما
در دنیای هوشمند

mahsan.co
info@mahsan.co





Kernel Bypass:

a solution to high performance network packet processing

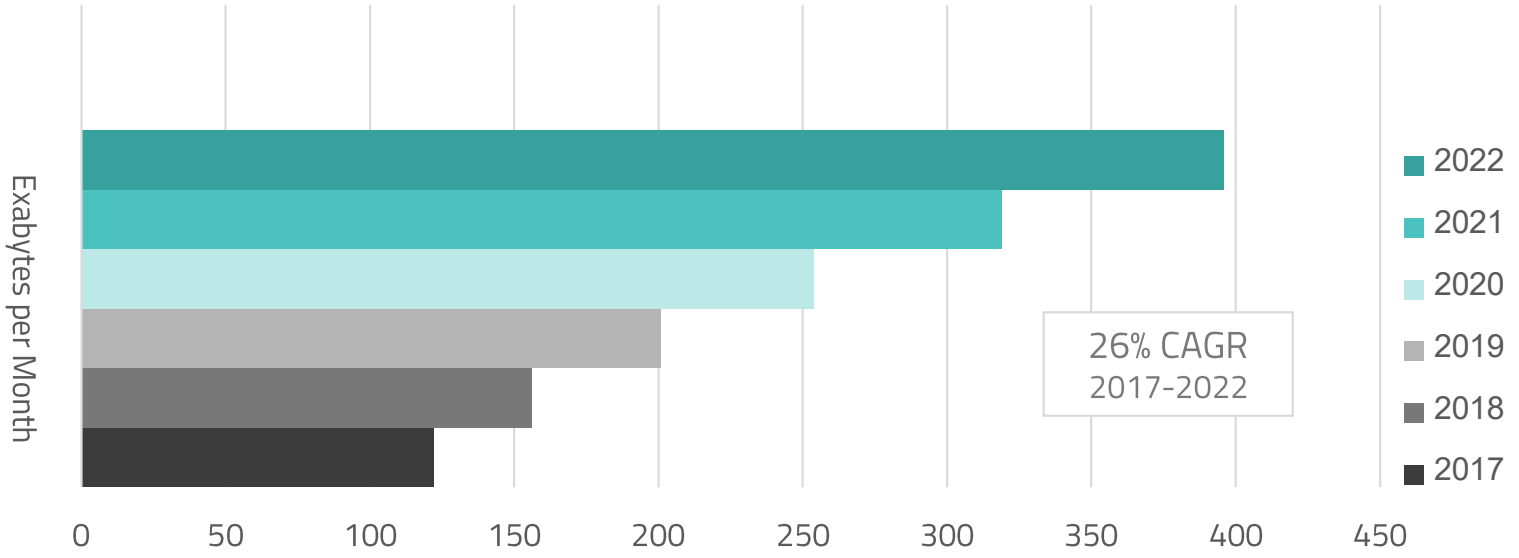
Outline

1| Kernel space limitations for high performance I/O

2| Introducing DPDK

3| Examples of DPDK usage

Global IP Traffic Forecasts

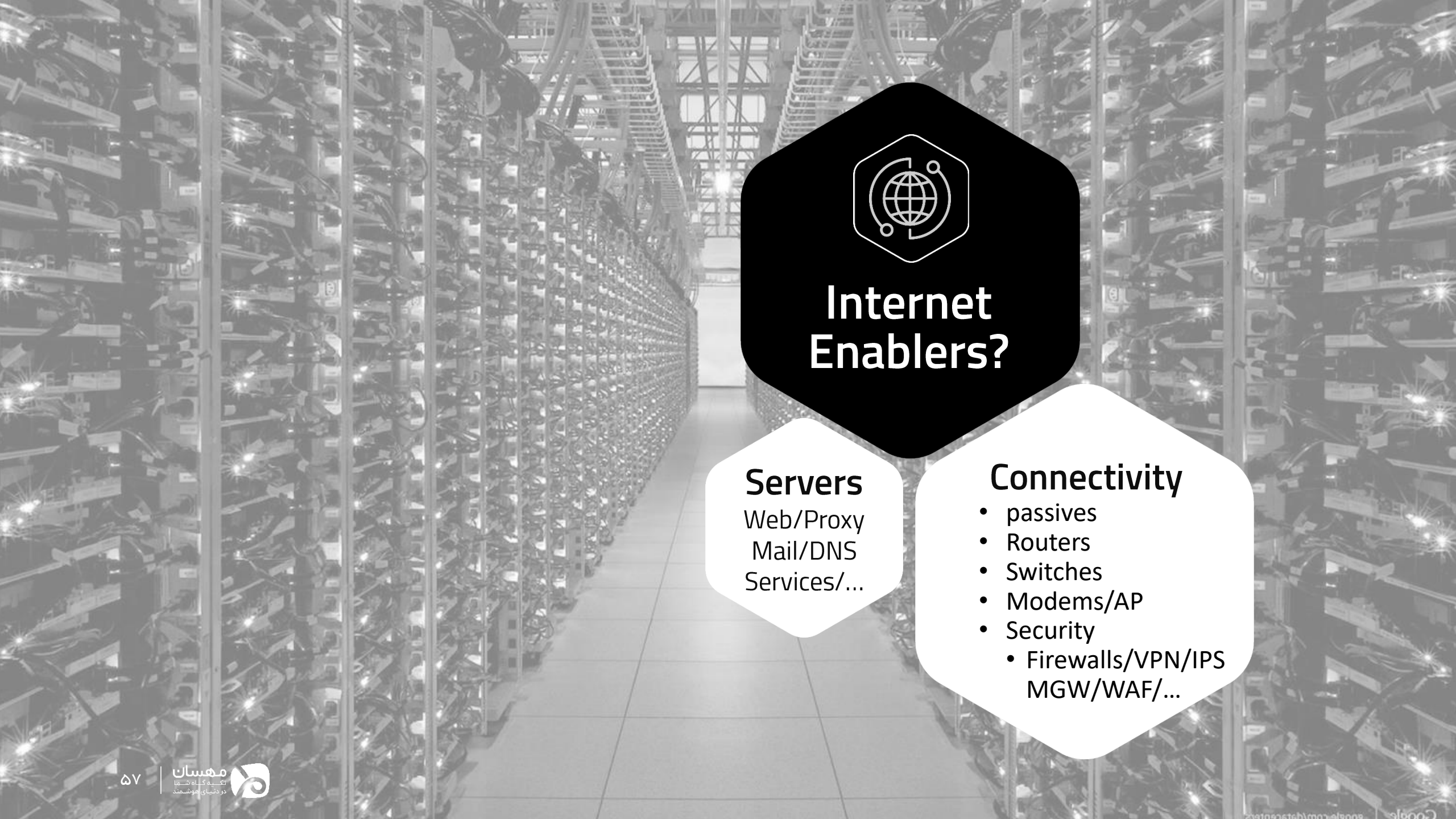


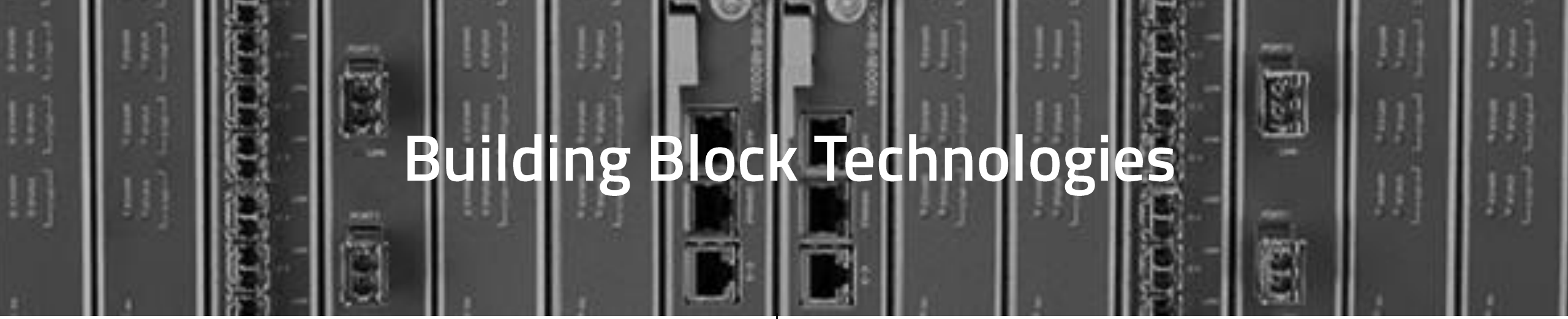
Source: Cisco VNI Global IP Traffic Forecast, 2017-2022

Historic Internet Contest

Year	Global internet traffic
1992	100 GB per day
1997	100 GB per hour
2002	100 GB per second
2007	2,000 GB per second
2017	46,600 GB per second
2022	150,700 GB per second

Source: Cisco VNI, 2018


Internet
Enablers?
Servers
Web/Proxy
Mail/DNS
Services/...
Connectivity
• passives
• Routers
• Switches
• Modems/AP
• Security
• Firewalls/VPN/IPS
MGW/WAF/...



Building Block Technologies

Proprietary HW

- Based on ASIC/FPGA
- High Quality
- Expensive

Commodity HW

- Based on CPU
- Flexibility
- Cheaper
- Modern innovations:
  - Multi-core CPU
  - Multi-queue NIC
  - NUMA architecture

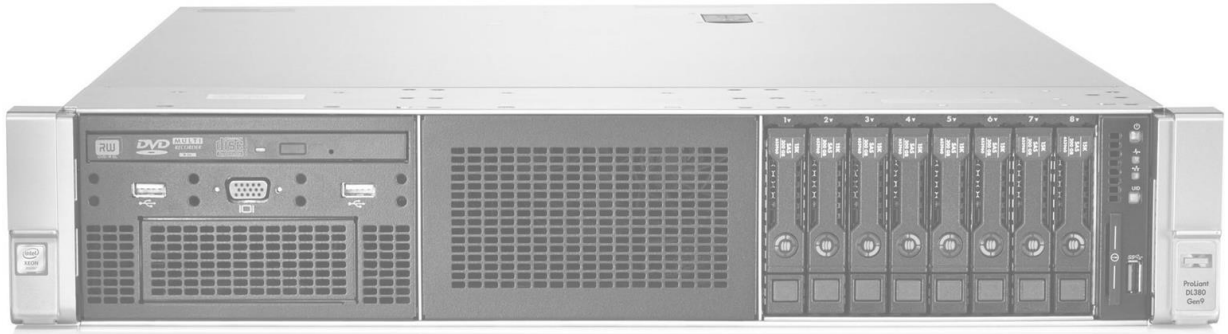
Software Solution Limitations

HW limitations

- General purpose architecture
- CPU/BUS/Mem speed

SW limitations

- Poor design
- Kernel vulnerabilities
- Kernel overhead



Kernel TCP/IP Stack Vulnerabilities

#	CVE ID	CWE ID	# of Exploits	Vulnerability Type(s)	Publish Date	Update Date	Score	Gained Access Level
1	CVE-2019-16089	476			2019-09-06	2019-09-09	7.5	None
An issue was discovered in the Linux kernel through 5.2.13. nbd_genl_status in drivers/block/nbd.c does not check the nla_nest_start_nofla								
2	CVE-2019-15927	125			2019-09-04	2019-09-05	7.2	None
An issue was discovered in the Linux kernel before 4.20.2. An out-of-bounds access exists in the function build_audio_procunit in the file sou								
3	CVE-2019-15926	125			2019-09-04	2019-09-05	9.4	None
An issue was discovered in the Linux kernel before 5.2.3. Out of bounds access exists in the functions ath6kl_wmi_pstream_timeout_event_ drivers/net/wireless/ath/ath6kl/wmi.c.								
4	CVE-2019-15925	125			2019-09-04	2019-09-05	7.2	None
An issue was discovered in the Linux kernel before 5.2.3. An out of bounds access exists in the function hclge_tm_schd_mode_vnet_base_cf								
5	CVE-2019-15924	476			2019-09-04	2019-09-05	4.9	None
An issue was discovered in the Linux kernel before 5.0.11. fm10k_init_module in drivers/net/ethernet/intel/fm10k/fm10k_main.c has a NULL alloc_workqueue failure.								
6	CVE-2019-15923	476			2019-09-04	2019-09-05	4.9	None
An issue was discovered in the Linux kernel before 5.0.9. There is a NULL pointer dereference for a cd data structure if alloc_disk fails in driv								
7	CVE-2019-15922	476			2019-09-04	2019-09-05	4.9	None
An issue was discovered in the Linux kernel before 5.0.9. There is a NULL pointer dereference for a pf data structure if alloc_disk fails in driv								
8	CVE-2019-15921	399			2019-09-04	2019-09-05	4.6	None
An issue was discovered in the Linux kernel before 5.0.6. There is a memory leak issue when idr_alloc() fails in genl_register_family() in net								
9	CVE-2019-15920	416			2019-09-04	2019-09-06	7.2	None
An issue was discovered in the Linux kernel before 5.0.10. SMB2_read in fs/cifs/smb2pdu.c has a use-after-free. NOTE: this was not fixed cc leak.								

TCP SACK kernel panic

- CVE-2019-11477

Remote code execution

- CVE-2019-11815

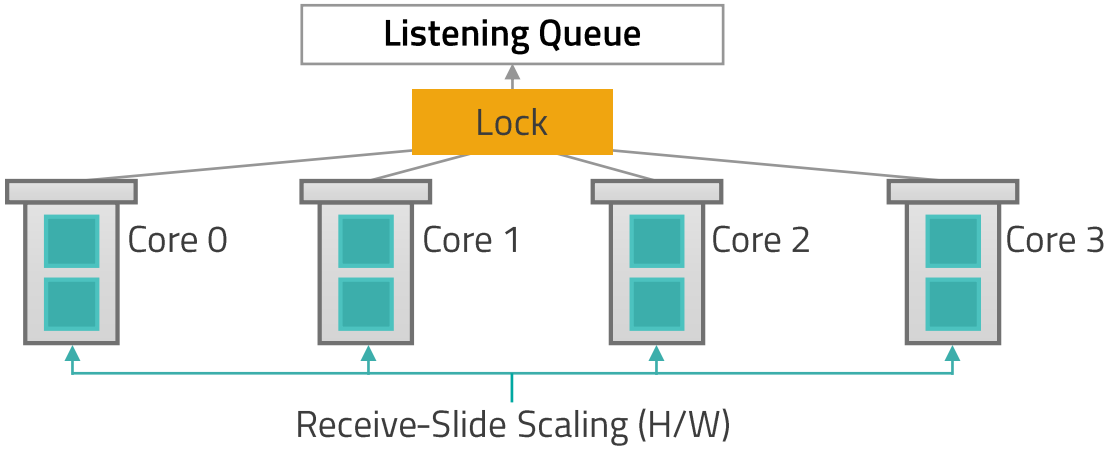
DoS attack

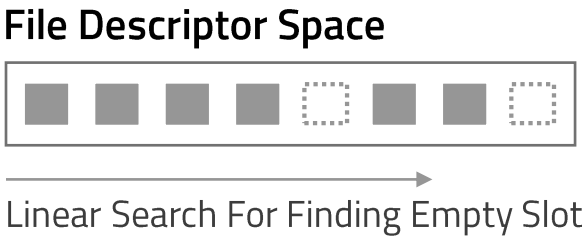
- CVE-2019-11478
- CVE-2019-11479
- CVE-2018-5390
- CVE-2019-9857
- CVE-2018-6922

...

Inefficiencies in kernel from shared FD

- Shared listening queue
- Shared file descriptor space

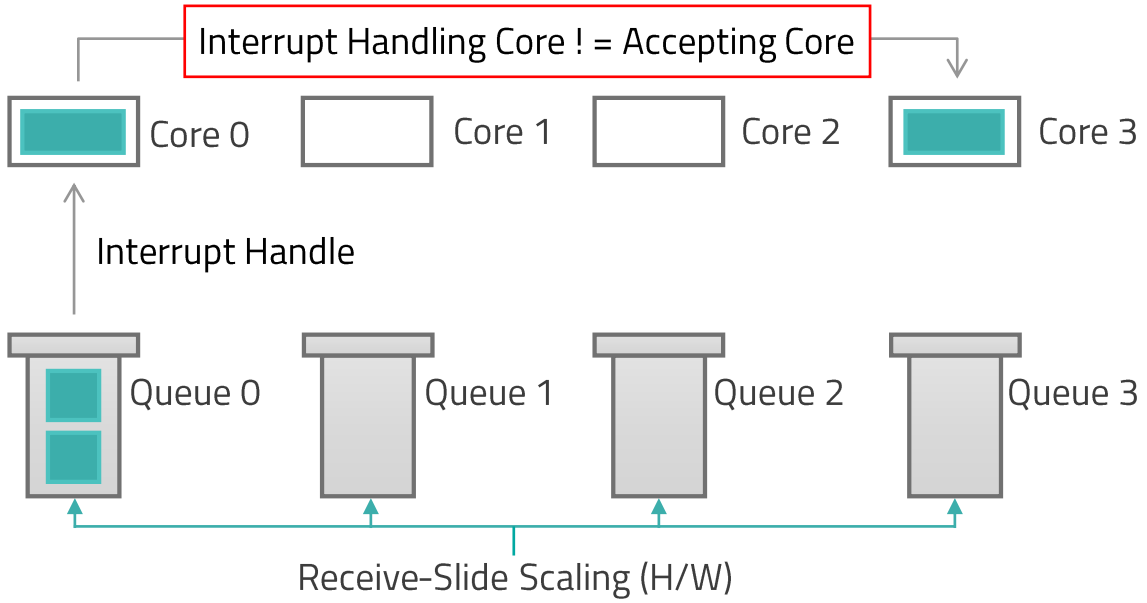




Per-core Packet Queue

Inefficiencies in kernel from broken locality

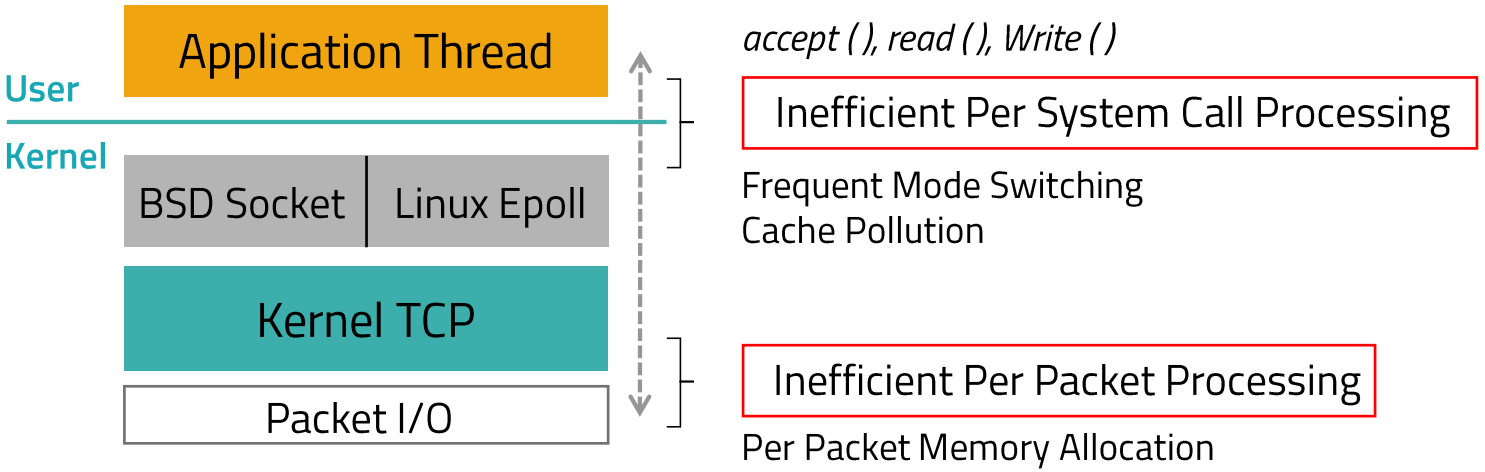
- Due to the nature of scheduler



accept ( )
read ( )
Write ( )

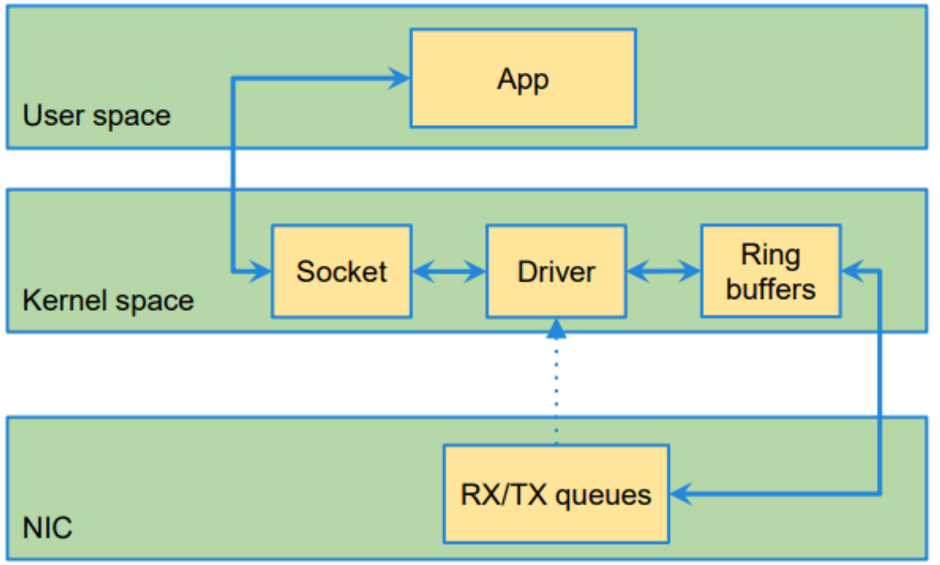
Per-core Packet Queue

Inefficiencies in kernel from lack of batch processing



Packet Processing Overheads in Kernel

- Interaction based on system calls
- Context switching on blocking IO
- Data copying from kernel to user space
- Interrupt handling in kernel
- Per packet processing

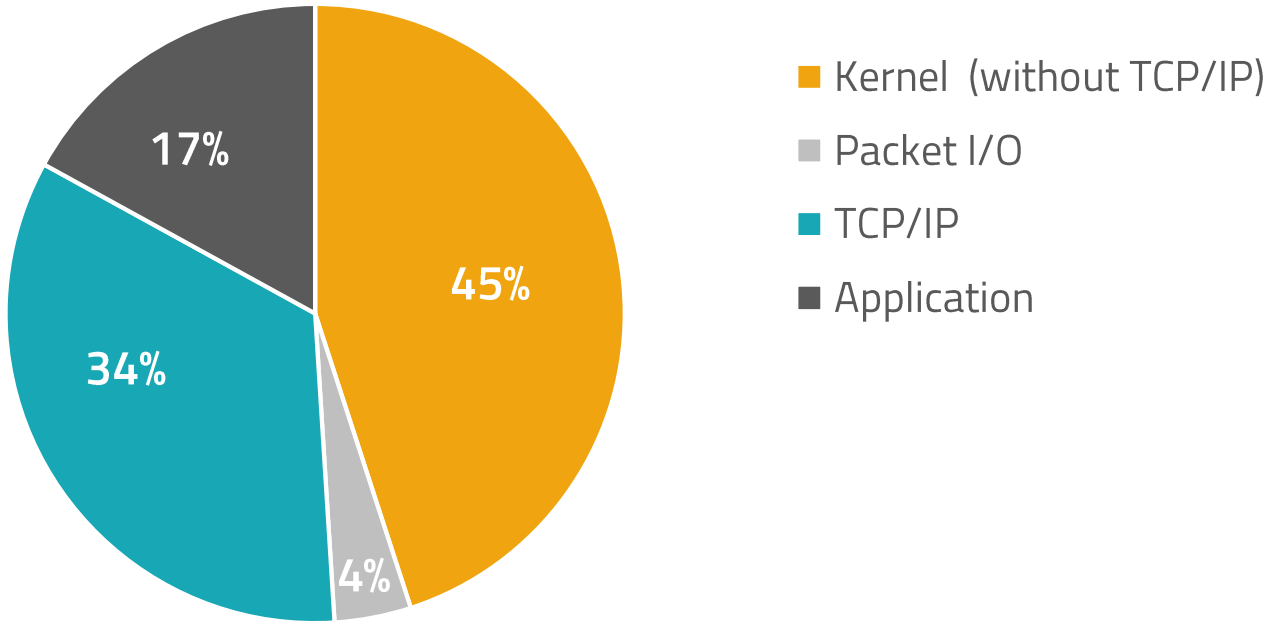


Packet Processing Overheads in Kernel

- Interaction based on system calls
- Context switching on blocking IO
- Data copying from kernel to user space
- Interrupt handling in kernel
- Per packet processing

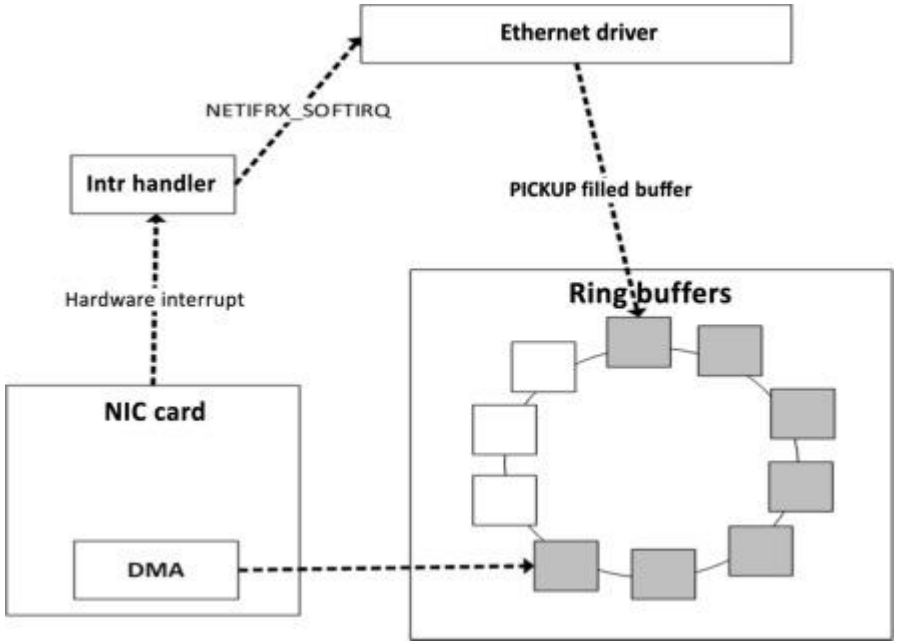
Function	Activity	Time (ns)
sendto	system call	96
sosend_dgram	lock sock_buff, alloc mbuf, copy in	137
udp_output	UDP header setup	57
ip_output	route lookup, ip header setup	198
ether_otput	MAC lookup, MAC header setup	162
ixgbe_xmit	device programming	220
Total		**950**

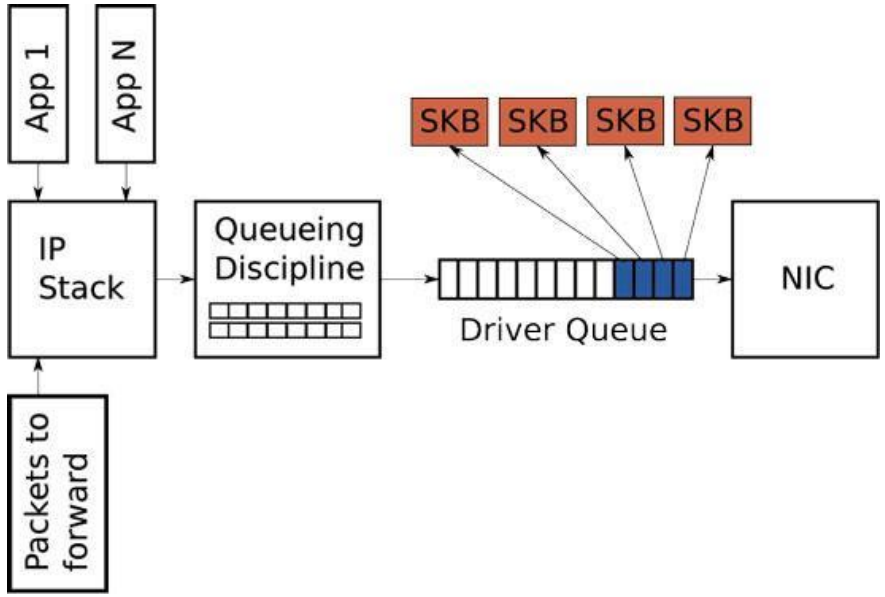
CPU Usage Breakdown Of Web Server



Web Server (lighttpd) Serving a 64 byte File / Linux-3.10

Packets RX/TX





What is the solution for better performance?

Kernel improvements

- SO_REUSEPORT
- SO_ATTACH*
- RPS
- Mega pipe
- XDP
- …

Bypassing kernel completely!

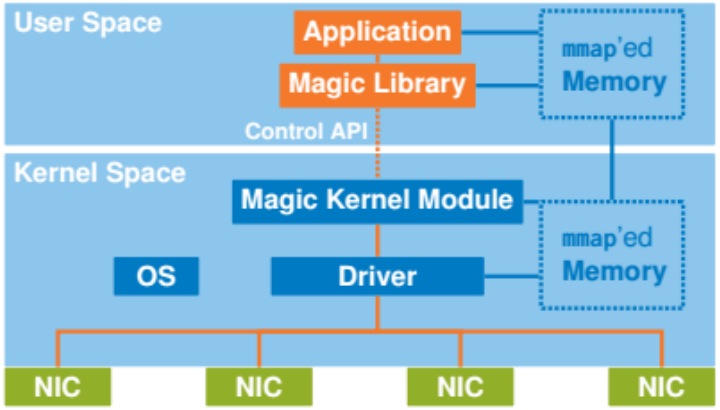
- Only for packet RX/TX

Solutions to bypass Kernel Overhead

Rely on a drive running in the kernel

- Netmap
- PF_RING
- Pfq
- OpenOnload

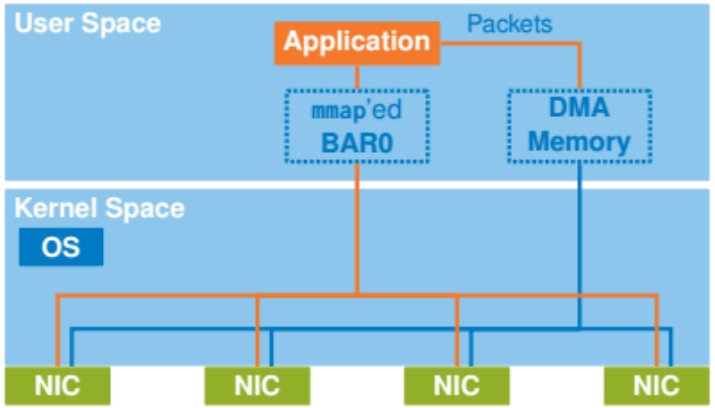
...

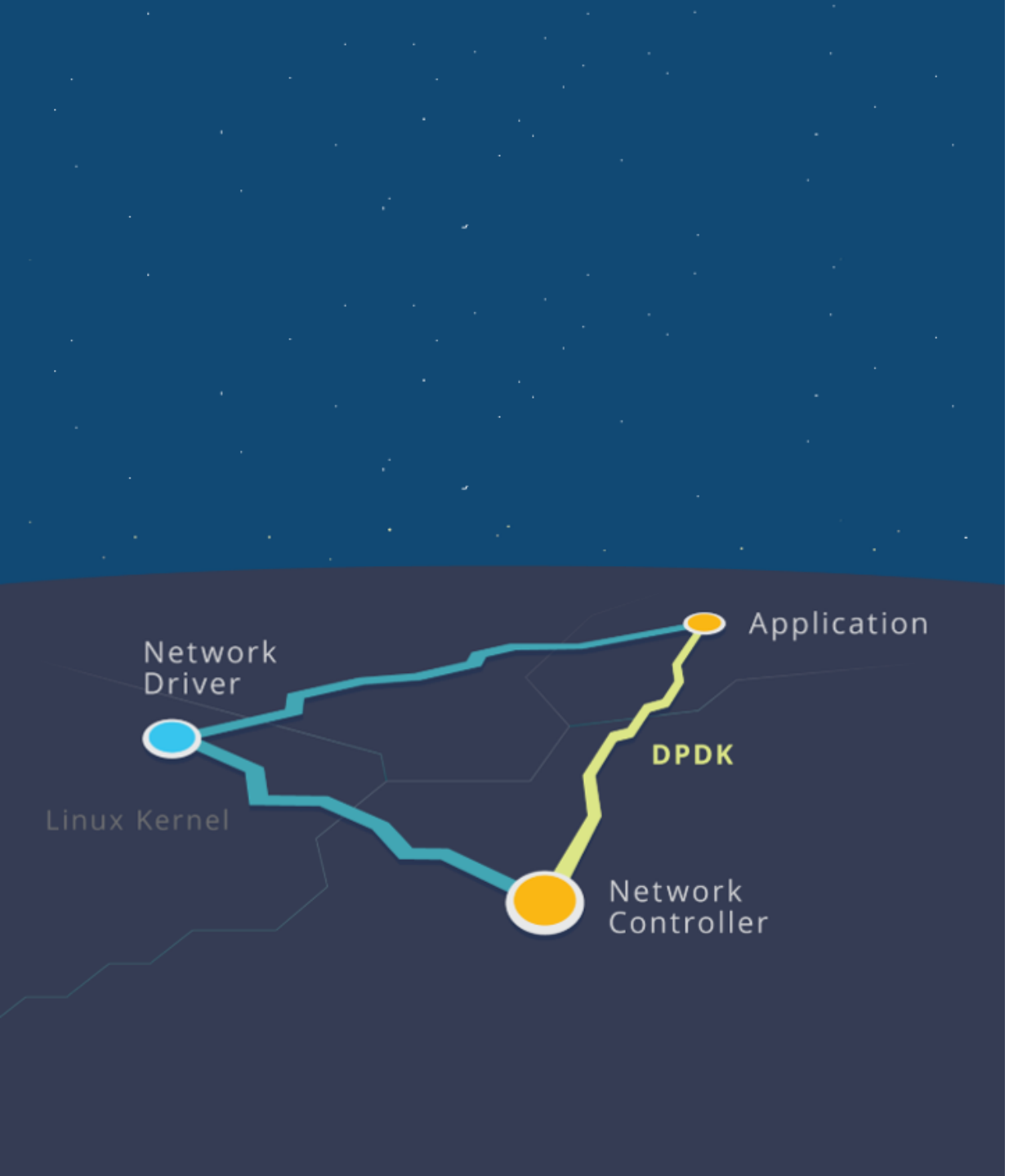


Solutions to bypass Kernel Overhead

Re-implement the whole driver in userspace

- DPDK
- Snabb





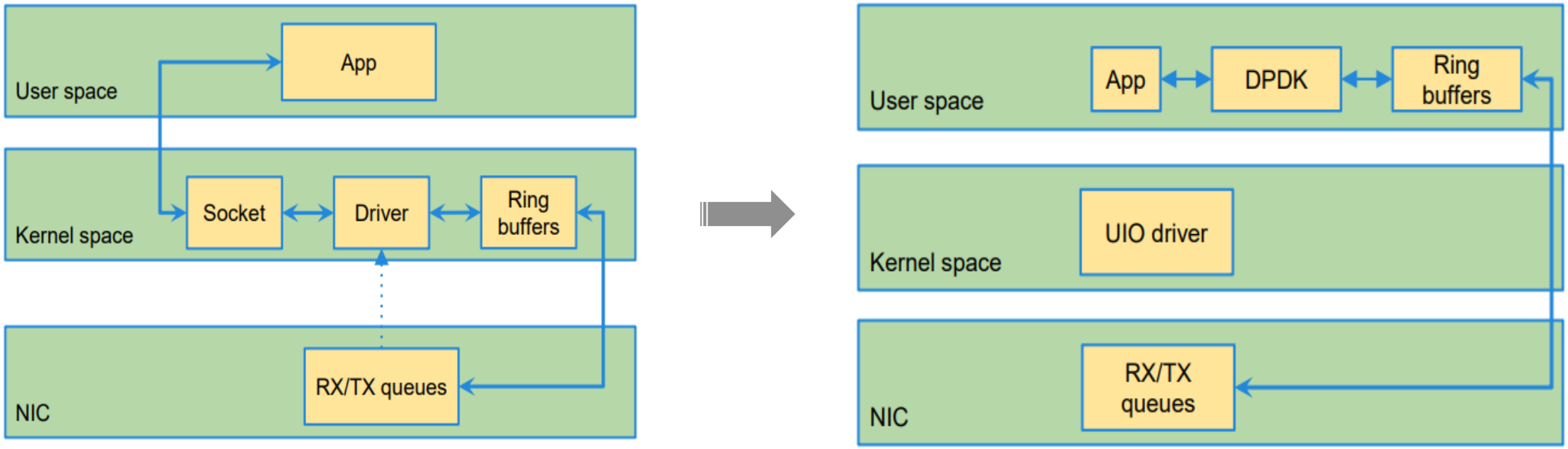
Data Plane Development Kit

- a framework for high performance packet processing
- was created in 2010 by Intel
- open source community was established at DPDK.org in 2013
- Currently is a Linux Foundation project (since 2017)

What is inside DPDK?

- Processor affinity (separate cores)
- Huge pages (no swap, TLB)
- UIO Driver (zero copy)
- Poll mode (no interrupt overhead)
- Lockless synchronization (avoid wait)
- Batch processing (versus single packet processing)
- SSE (Streaming SIMD Extensions)
- NUMA awareness
- Pre-allocation, ring API, …
- Libraries: reassembly, flow classification/filtering, QoS, bpf, cryptodev,…
- Kernel interface (KNI)

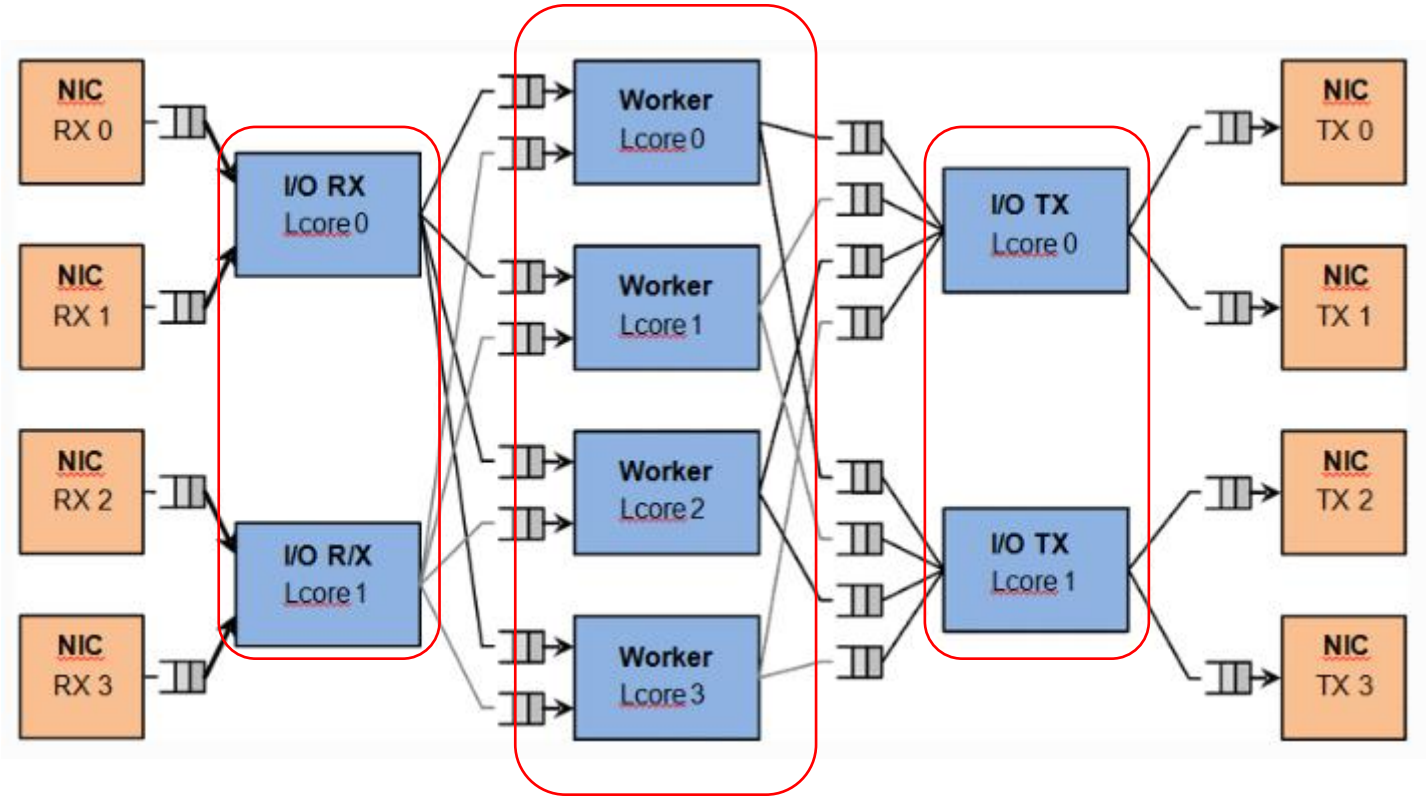
Packet handling with DPDK



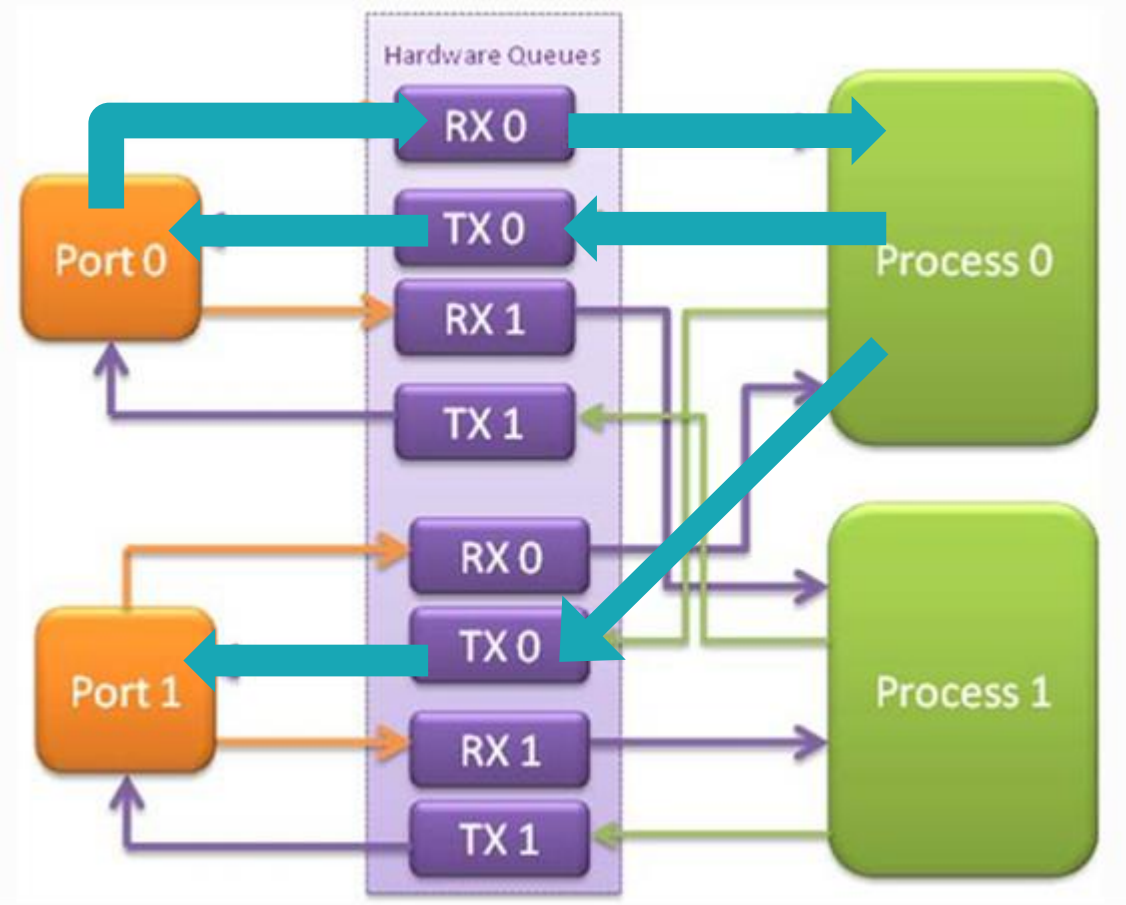
L2 Basic Forwarding

```
for (;;) {
    RTE_ETH_FOREACH_DEV(port) {

        /* Get burst of RX packets, from first port of pair. */
        struct rte_mbuf *bufs[BURST_SIZE];
        const uint16_t nb_rx = rte_eth_rx_burst(port, 0,
                bufs, BURST_SIZE);

        if (unlikely(nb_rx == 0))
            continue;

        /* Send burst of TX packets, to second port of pair. */
        const uint16_t nb_tx = rte_eth_tx_burst(port ^ 1, 0,
                bufs, nb_rx);

        /* Free any unsent packets. */
        if (unlikely(nb_tx < nb_rx)) {
            uint16_t buf;
            for (buf = nb_tx; buf < nb_rx; buf++)
                rte_pktmbuf_free(bufs[buf]);
        }
    }
}
```

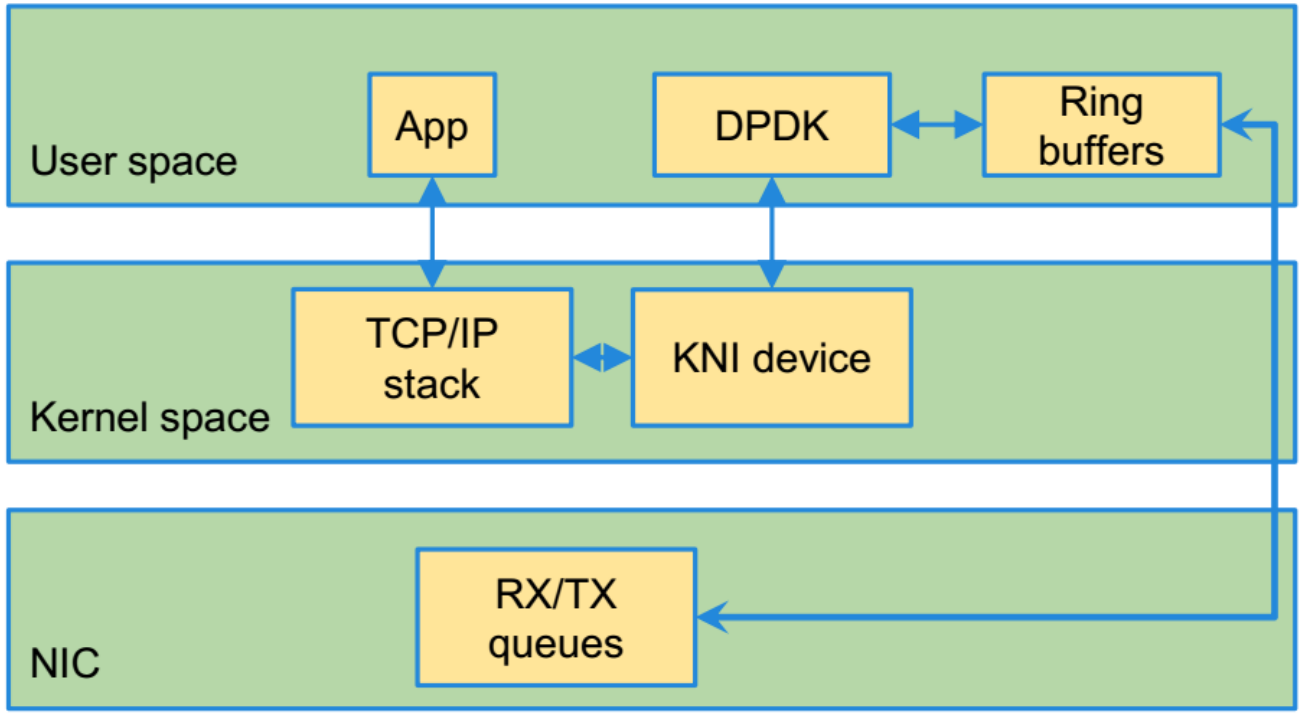
Load Balancer



Independent Processes

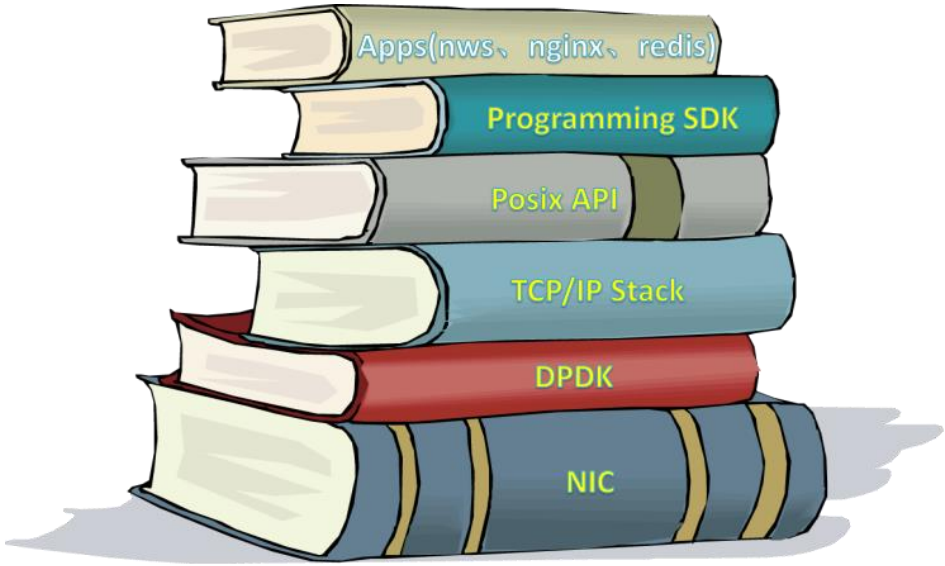


Interface to kernel

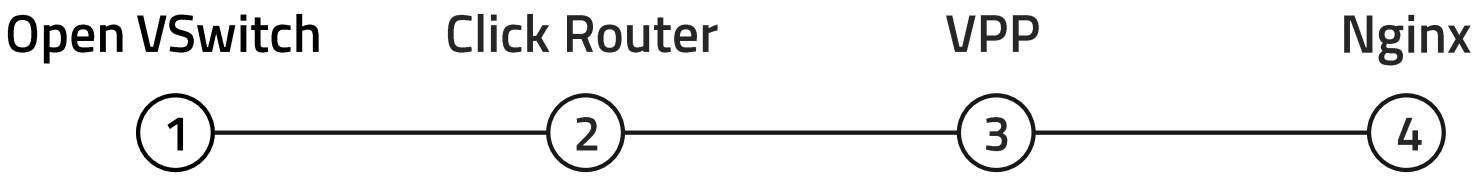


Userspace TCP/IP stack compatible with DPDK

- mTCP
- F-Stack
- TLDK
- ANS
- OpenFastPath
- ...

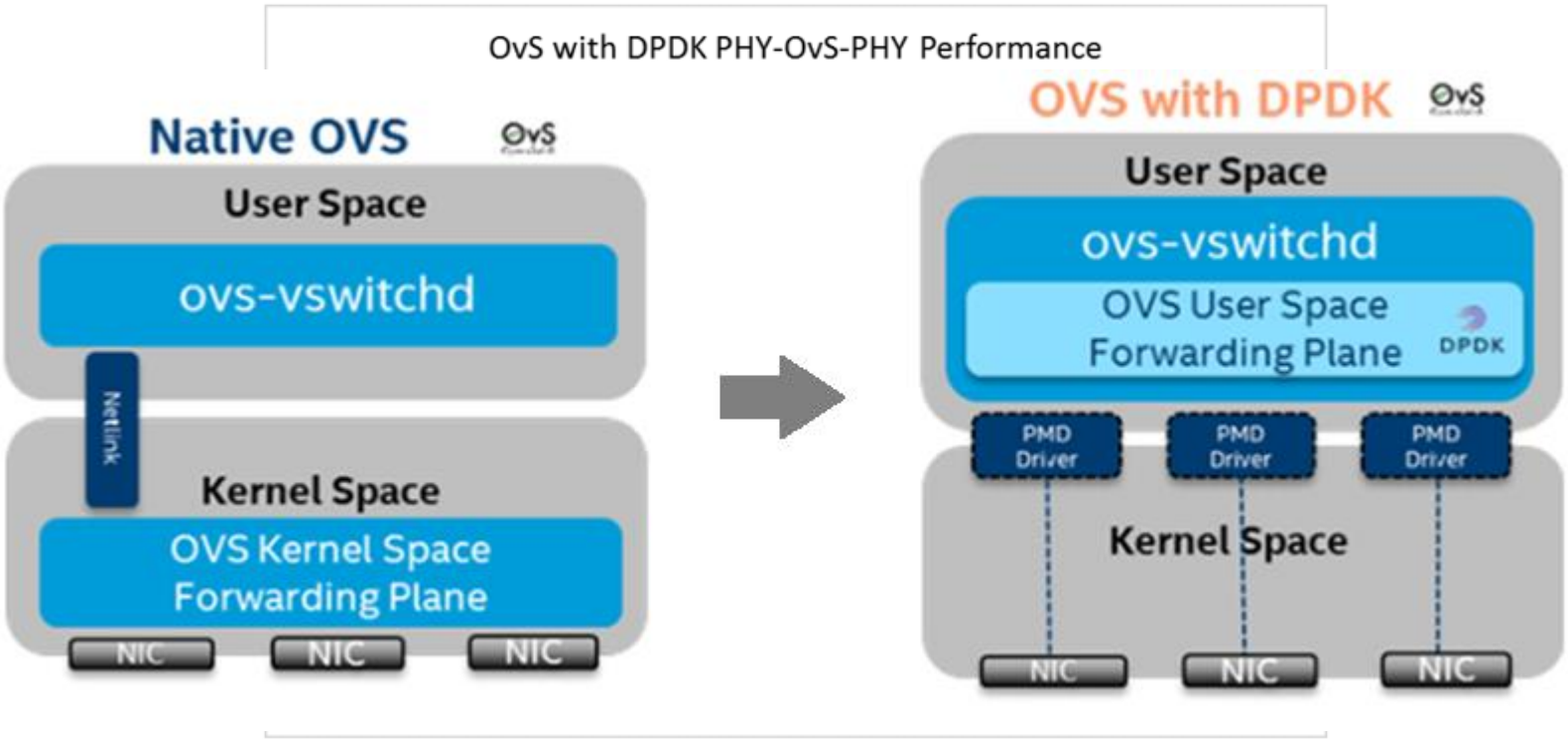


Examples of DPDK Usage

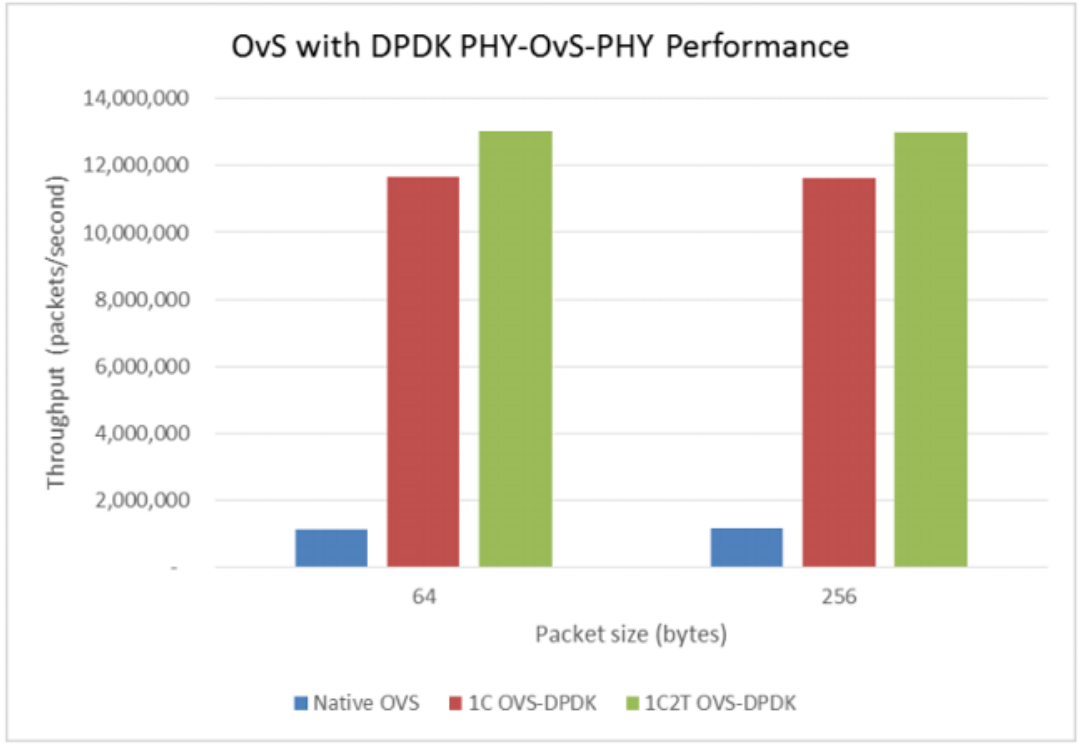


Open VSwitch 1 — Click Router 2 — VPP 3 — Nginx 4

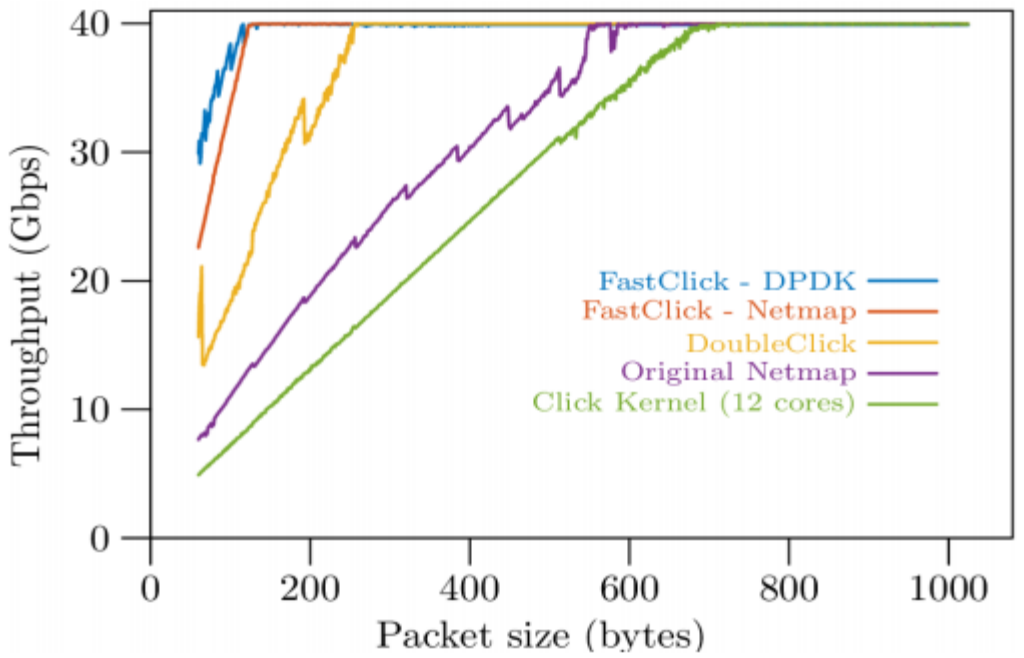
Improving OvS performance



Improving OvS performance



Improving Click Router Performance

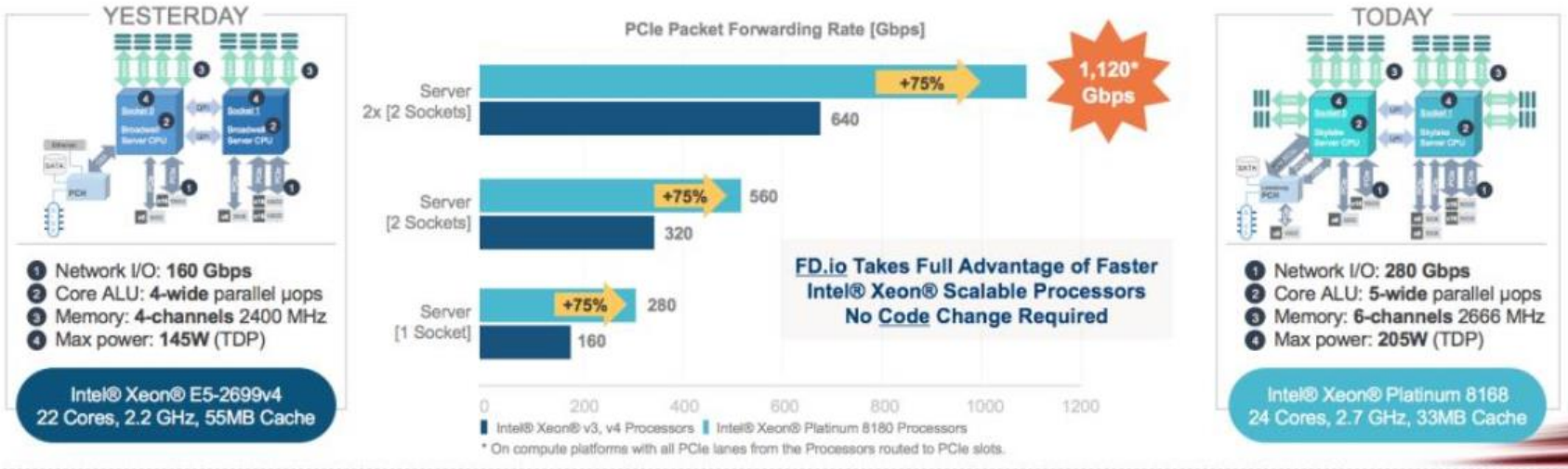


Open VSwitch — Click Router — **VPP** — Nginx

1 — 2 — 3 — 4

FD.io Benefits from Intel® Xeon® Processor Developments

Increased Processor I/O Improves Packet Forwarding Rates

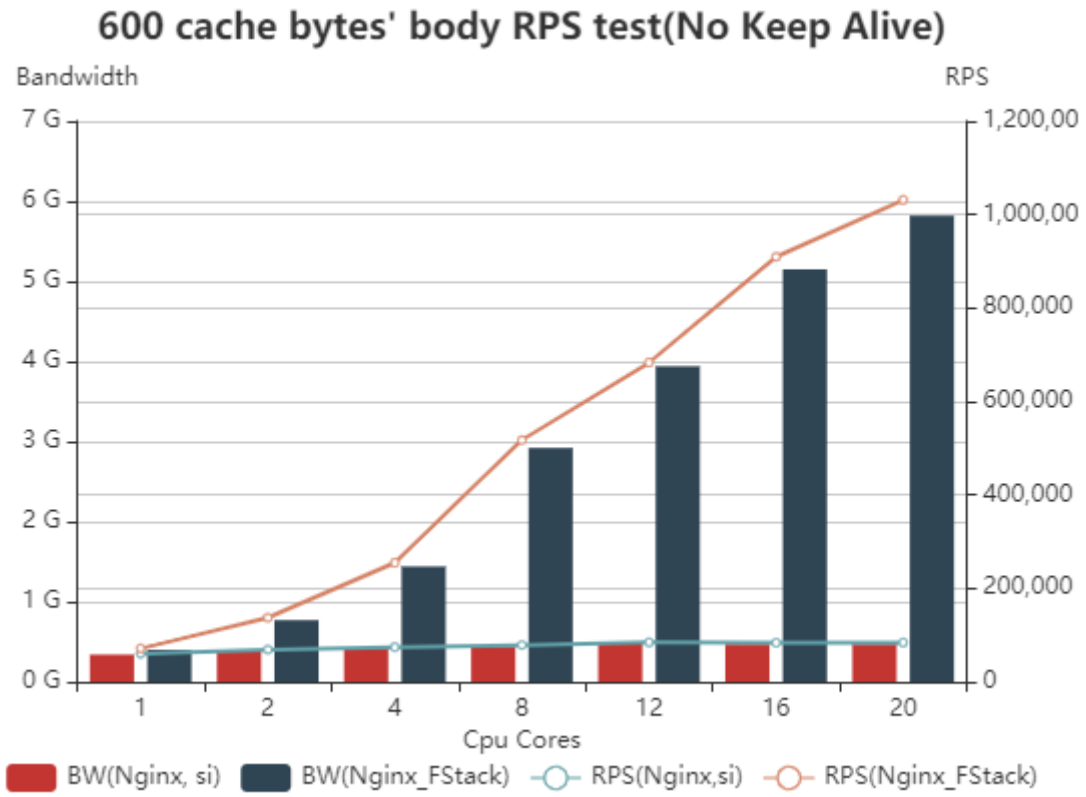


Breaking the Barrier of Software Defined Network Services
1 Terabit Services on a Single Intel® Xeon® Server !

Open VSwitch Click Router VPP **Nginx**

1 — 2 — 3 — 4

Improving Nginx with F-Stack



مهسان به عنوان شرکتی فعال در حوزه سامانه‌های ارتباطی نوین، تلاش دارد به معتمدترین ارائه دهنده راهکارهای موثر و کارآمد در حوزه فعالیت خود تبدیل گردد.

از این رو تمام تلاش خود را برای رسیدن به آن به کار بسته‌ایم و در این مسیر جمعی از بهترین‌ها ما را همراهی می‌کنند.

۱ | جلوگیری از نشت اطلاعات بارو
Baroo DLP

۲ | دروازه امنیتی بارو
Baroo Gate

۳ | مدیریت و نظارت بر دسترسی راه دور
Wina PAM

۴ | یکسوکننده جریان داده بارو
Baroo Data Diode

۵ | مدیریت تلفن همراه سازمانی دژ
Dej EMM

۶ | سامانه پایش تیام
Tiyam Monitoring System

با تشکر

ما همواره در مهسان پذیرای همکارانی متخصص و توانمند با ایده‌هایی جذاب هستیم

HR@mahsan.co

مهسان
تکیه گاه شما
در دنیای هوشمند

تهران، خیابان بهشتی، خیابان قنبرزاده، کوچه دهم، پلاک ۳۶
تلفن تماس:۸۸۵۲۱۲۳۸ info@mahsan.co