High Density Scalable Cloud Gateway for Cloud Networking

Ni Hongjun, Zhang Pan

Notices and Disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at www.intel.com.
- Intel processors of the same SKU may vary in frequency or power as a result of natural variability in the production process.
- Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
- Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804.
- The benchmark results may need to be revised as additional testing is conducted. The results depend on the specific platform configurations and workloads utilized in the testing, and may not be applicable to any particular user's components, computer system or workloads. The results are not necessarily representative of other benchmarks and other benchmark results may show greater or lesser impact from mitigations.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors.
  Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit www.intel.com/benchmarks.
- Performance results are based on testing as of 8/8/2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product or components can be absolutely secure.
- Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.
- The cost reduction scenarios described are intended to enable you to get a better understanding of how the purchase of a given Intel based product, combined with a number of situation-specific variables, might affect future costs and savings. Circumstances will vary and there may be unaccounted-for costs related to the use and deployment of a given product. Nothing in this document should be interpreted as either a promise of or contract for a given level of costs or cost reduction.
- No computer system can be absolutely secure.
-
- *Other names and brands may be claimed as the property of others.

Agenda

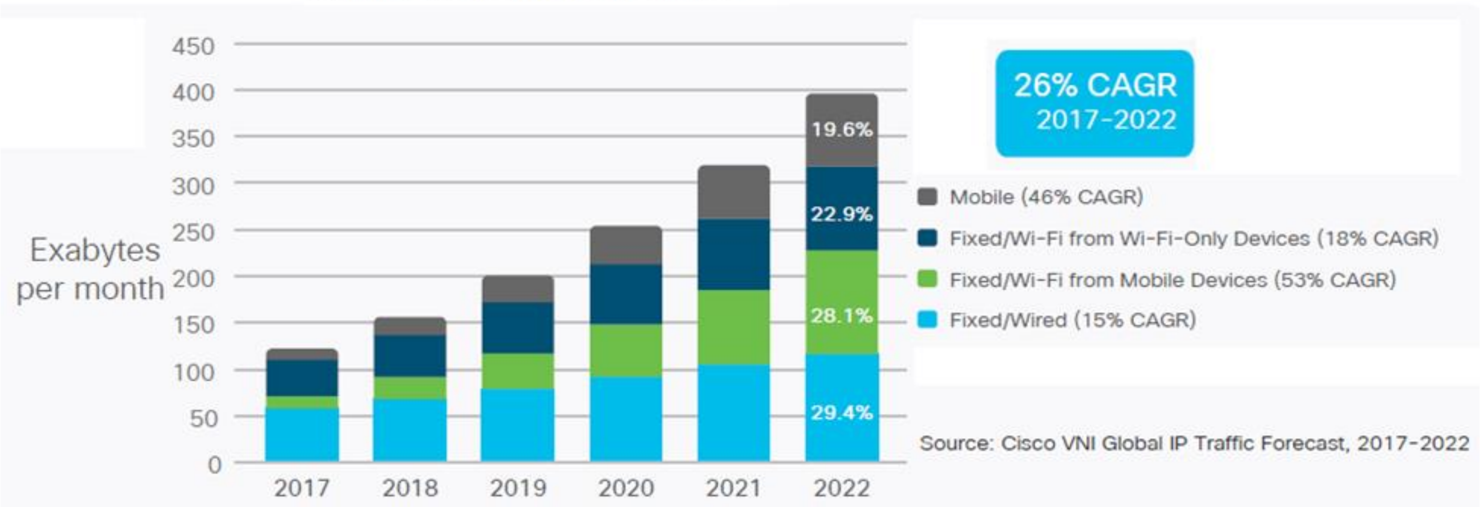
- Background
- Market Challenges
- Framework Refactor
- Key Features Optimization
- Newly Added Features
- Next Step
- Key Takeaways

Background

Welcome To The 5G/Cloud/IoT/Bigdata Era

Dramatically increased Network traffic/connections/throughput

Global IP traffic, wired and wireless



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

Picture from: Cisco Annual Internet Report (2018–2023) White Paper

Huge Network Throughput

Traffic amount increasing exponentially

Large Connections

Large connection number due to IoT devices and rich applications

100G Migration

100Gbps becomes mainstream network interface standards

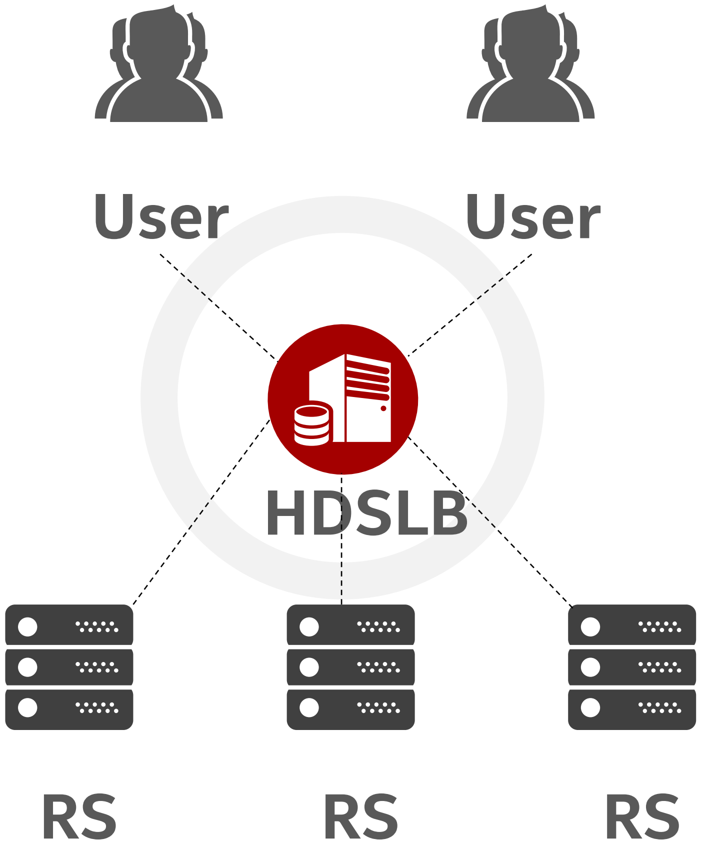
Elephant Flow

Video/storage applications generate huge throughput long-live connections

Market Challenges

New Challenges For Cloud Gateway

New business scenarios arise new challenges for the core access layer device-load balancer



Performance Requirements for Single Node

01 100M Level Concurrent Conn

02 150Mpps/200Gbps Throughput

03 Single Session 10Mpps Level

Highlights

HDSLB Addressing These Challenges With Industry Leading Performance

Intel Processors and NIC Packaged Solution

Fully optimized

Handle 100M Level Concurrent Conn

Address the business challenges for large concurrent conns

Handle 150Mpps Level Throughput

Address the business challenge of huge traffic

Handle 10Mpps Level Elephant Flow

Address the business challenge of Elephant Flow

Up to 3x higher performance

Scaling for DNAT and SNAT

Common Features

01

HA

Advanced Session Sync

Easily nodes add/delete

02

LB Mode

FullNAT/DNAT/SNAT /DR

03

LB algorithms

Round Robin/Weighted
Lease Connection
Consistent Hash

04

Security

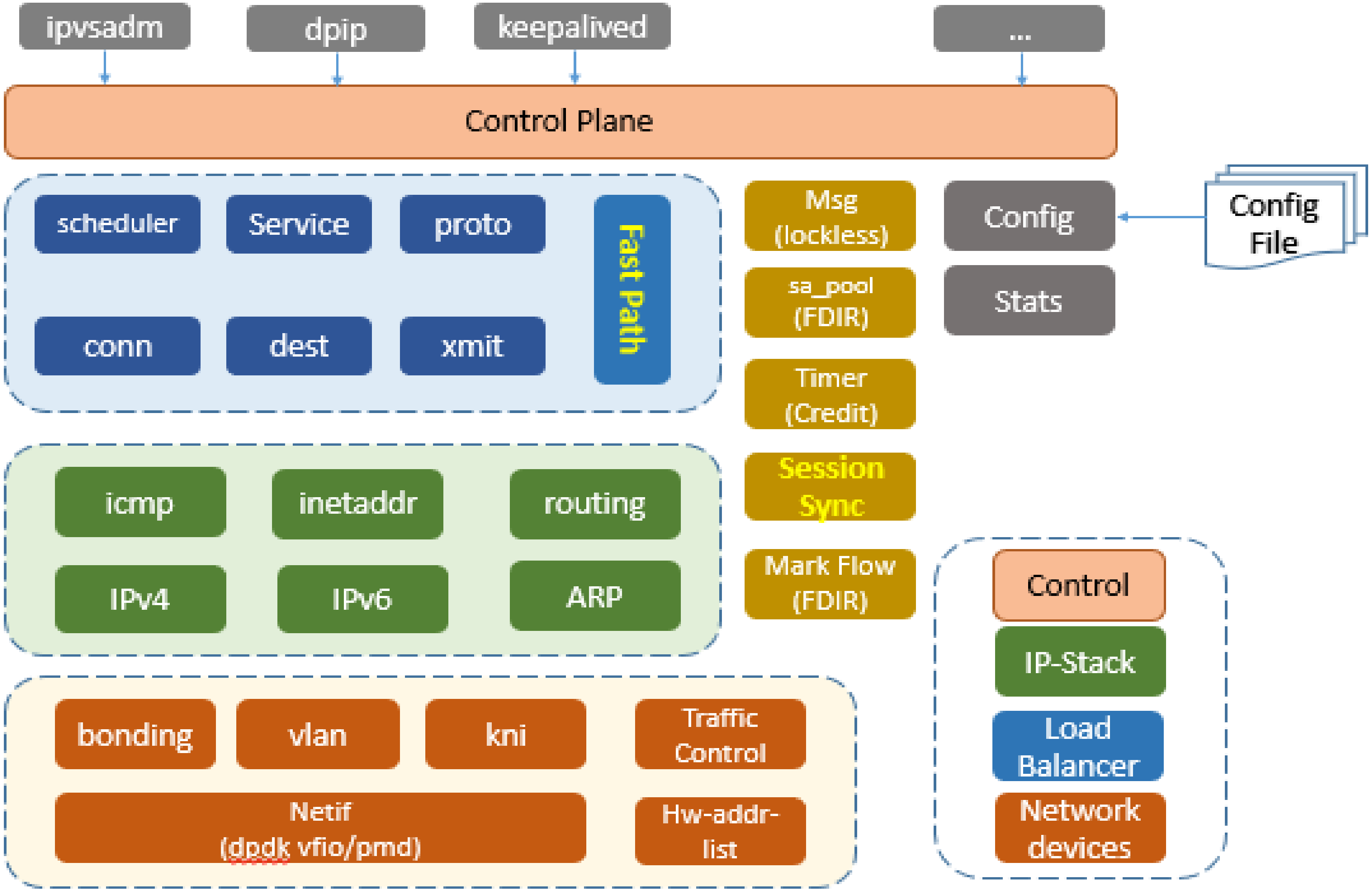
ACL/TCP SYN-proxy/QoS

05

Visibility/Observability

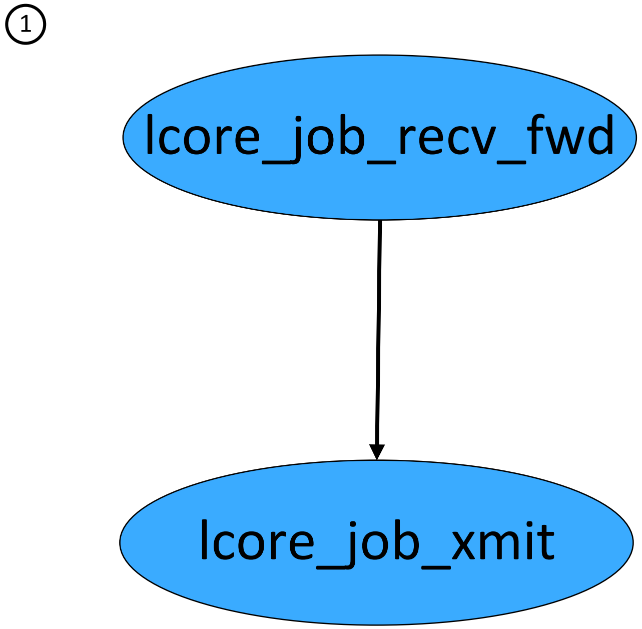
Detailed runtime telemetry

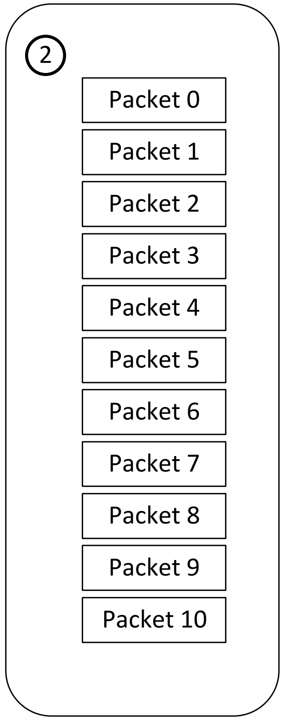
Refactor Framework



- Fast Path
- Session Sync
- Mark Flow
- Microjob
- DPDK 20.11

Key Optimization: Vectorize



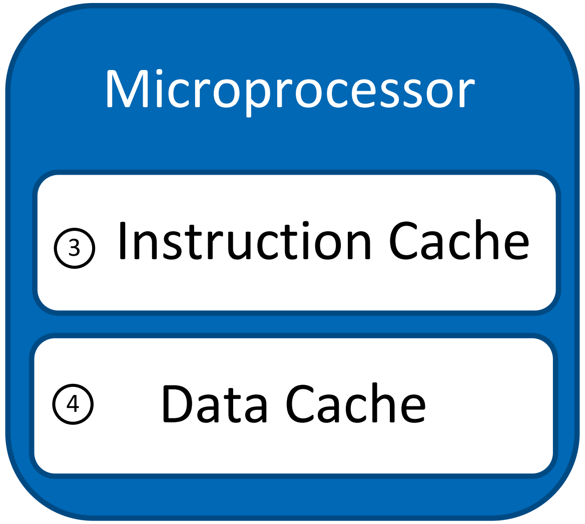


Packet processing is decomposed into more microjobs…

… packets moved through microjobs in vector …

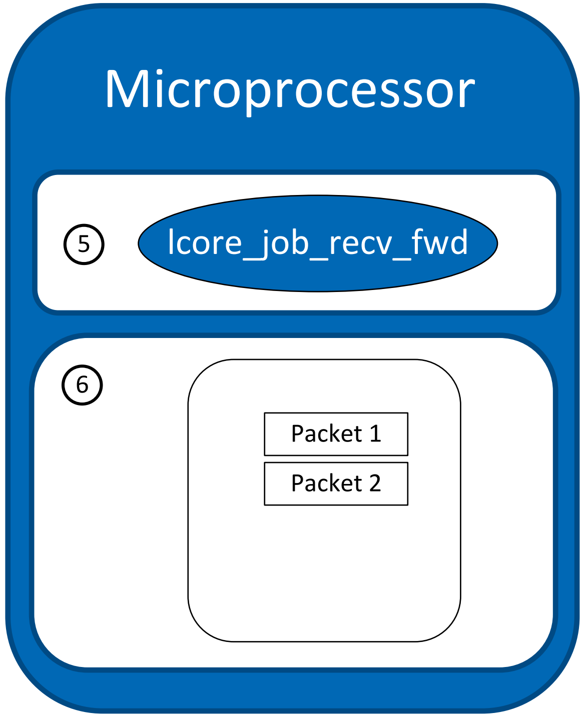
Key Optimization: Microjob

Microjobs: microjobs are optimized to fit inside the instruction cache …



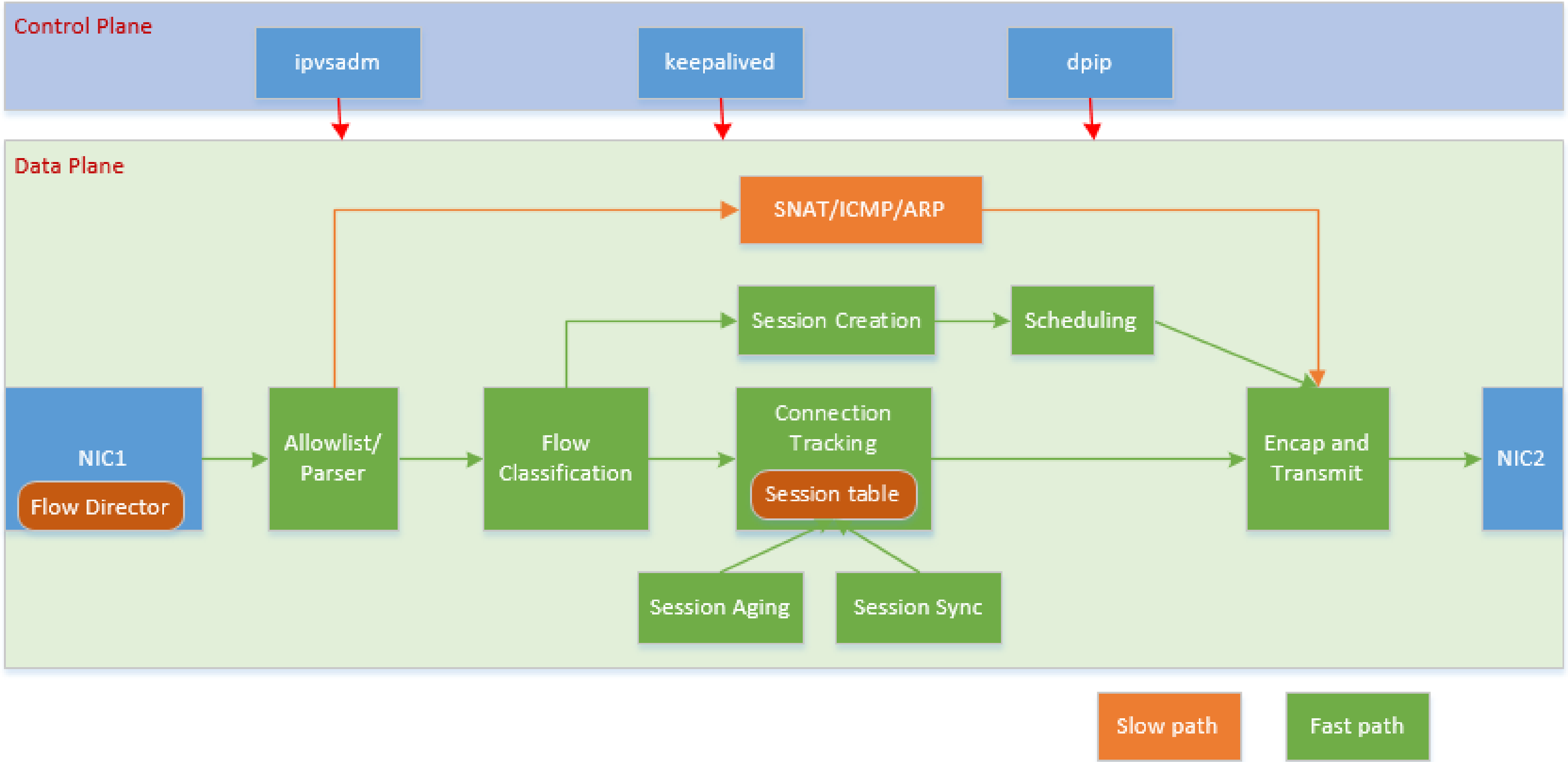
… packets are pre-fetched, into the data cache …

… instruction cache is warmed with instructions from a single microjob …



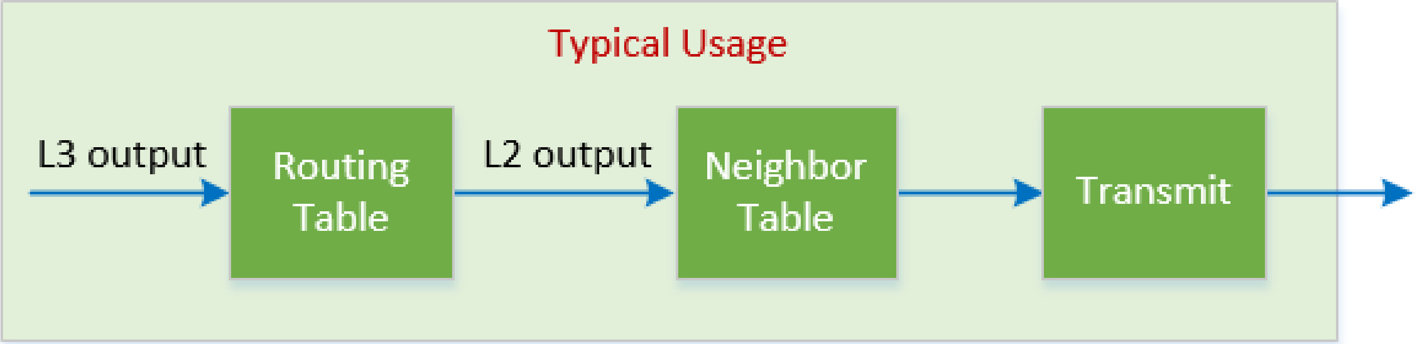
… data cache is warmed with a small number of packets …

Separating Fast Path from Slow Path

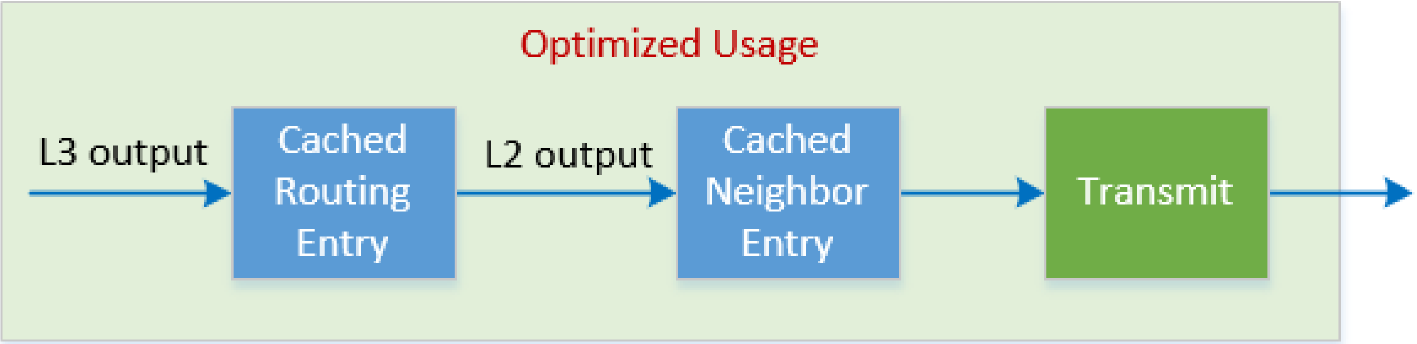


- **Slow Path** is used to handle ICMP/ARP etc.
- **Fast Path** is used for session creation, scheduling, connection tracking, session aging, etc.

Routing and Neighbor Optimization

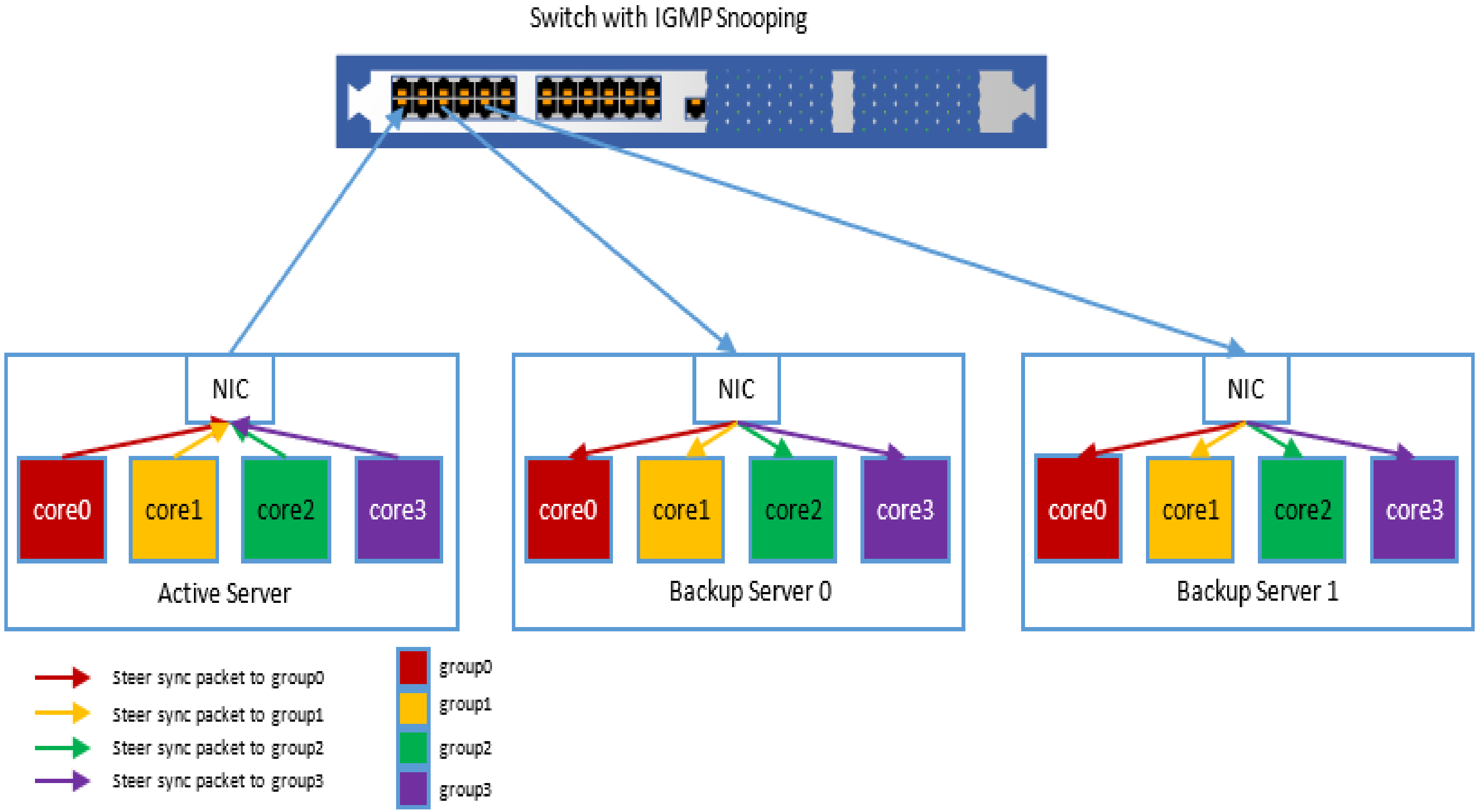


Cache routing and neighbor entry:



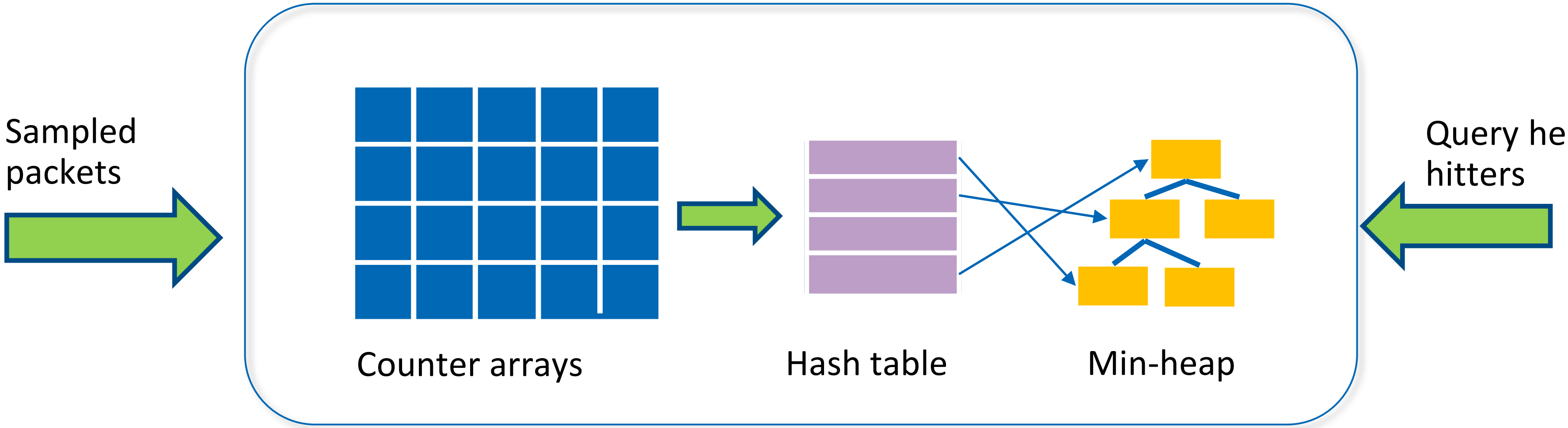
- Caches when creating a new session
- Updates when routing/neighbor change

Live Migration



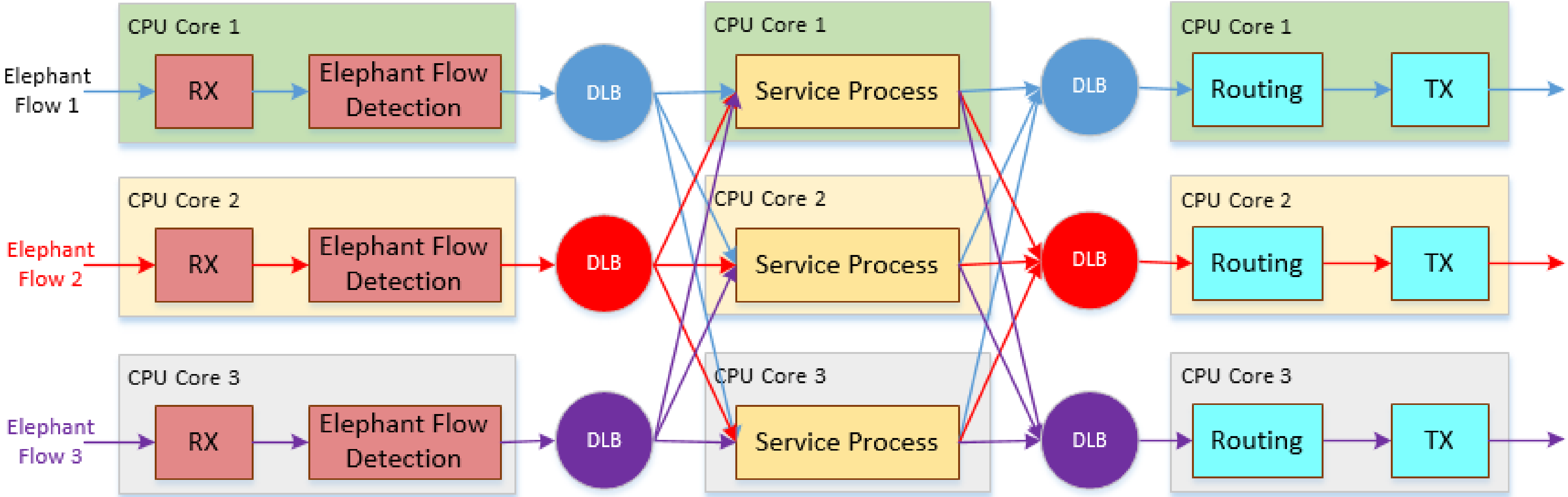
- Multicast for core x
- ToR for forwarding
- Sync via Multicast
- New Session Sync
- All Session Sync

Heavy Hitter Detection Algorithm



- The algorithm profiles and reports heavy flows with their estimated packet counts.
- The data structure is small enough to reside in local data cache.
- Only a small percentage of total packets needs to be sampled (e.g. 1%, configurable).
- Leverages a hash table to optimize the heap lookup time.
- Collaborating with Professor Liu, the author of Nitrosketch to further improve the algorithm.

Elephant Flow Processing



- Some CPU Cores are bound to create a CPU Group.
- Elephant flows per CPU are detected through an innovative algorithm.
- Then distributes elephant flows to CPU Cores in this Group through a DLB queue for Service Process.
- Packets are aggregated through DLB into the same CPU Core as RX, and then perform Routing and TX.

Next Step

- Elephant flow detection and distribution with Hardware DLB. Work In Progress
- IPv6 routing lookup optimization using novel algorithm. Work In Progress
- Wireguard support and optimization. Work In Progress
- Inline data inspection.
- Threat detection and defense.
- More is coming …

Key Takeaways

- HDSLB is a **High Density** and **Scalable** Cloud Gateway running on x86 servers.
- It **refactors** DPVS project, and leverages **DPDK 20.11**.
- It fully leverages **HW capabilities** of CPU and NIC.
- It separates **Fast Path** from Slow Path to boost performance.
- **Vectorize and Microjob** helps to get more performance gain.
- It addresses the challenges of **performance**, **scalability** and **live migration**.

Acknowledgement

DPDK Community

Jay Vincent @ Intel

Li Jokul @ Intel

Wang Yipeng @ Intel

Niall McDonnell @ Intel

DPVS Community

Li Baoqian @ Intel

Xu Qian @ Intel

Alan (Zaoxing) Liu @ Boston University

Zhu TaoX @ Intel

intel®