Conserving Memory Bandwidth with Virtual Gather

Agur Adams

Stanford EE PhD Student
Advisor Dr. Philip Levis

DAM Winter Workshop

10 January, 2025

Note: Some results have been omitted from the original presentation. Please contact Dr. Philip Levis for access to the full presentation

Motivation

Network line rates have significantly increased

40 Gbps
2009

400 Gbps
2021

*Images from Memory4Less and HPC Wire

Motivation

Network line rates have significantly increased

- Increased speeds has led to **reduced visibility** for network analysis
- High volume of traffic **overwhelms** most analysis tools (e.g., Snort)

Example: Snort 3.0 would require 125-667 cores to support 100 Gbps throughput (or 4-21 servers) [1]

*Image from Snort.org

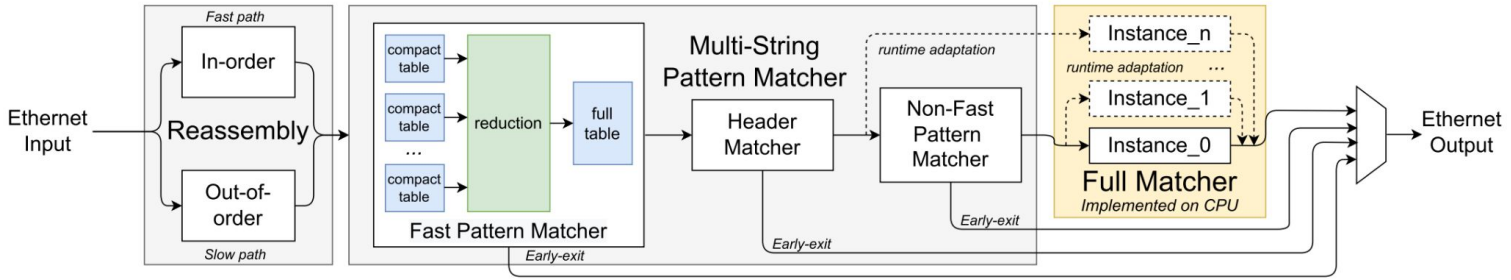
Motivation

Network line rates have significantly increased

- Increased speeds has led to **reduced visibility** for network analysis
- High volume of traffic **overwhelms** most analysis tools (e.g., Snort)

Complex analysis at line rates ≥ 100 Gbps requires on-NIC processing

Pigasus: Used an **FPGA-capable SmartNIC** to search over 10,000 Snort rules in over 100,000 concurrent connections at 100 Gbps [1]



Motivation

Network line rates have significantly increased

- Increased speeds has led to **reduced visibility** for network analysis
- High volume of traffic **overwhelms** most analysis tools (e.g., Snort)

Complex analysis at line rates ≥ 100 Gbps requires on-NIC processing

- Switches, routers, and fast packet processors support only simple analysis
- Specialized hardware can be costly and difficult to program (e.g., FPGAs)

Motivation

Network line rates have significantly increased

- Increased speeds has led to **reduced visibility** for network analysis
- High volume of traffic **overwhelms** most analysis tools (e.g., Snort)

Complex analysis at line rates ≥ 100 Gbps requires on-NIC processing

- Switches, routers, and fast packet processors support only simple analysis
- Specialized hardware can be costly and difficult to program (e.g., FPGAs)

SmartNICs have limited memory bandwidth relative to their line rates

- Copying data to reassemble application bytestreams is expensive
- Careful cache management required

Motivation

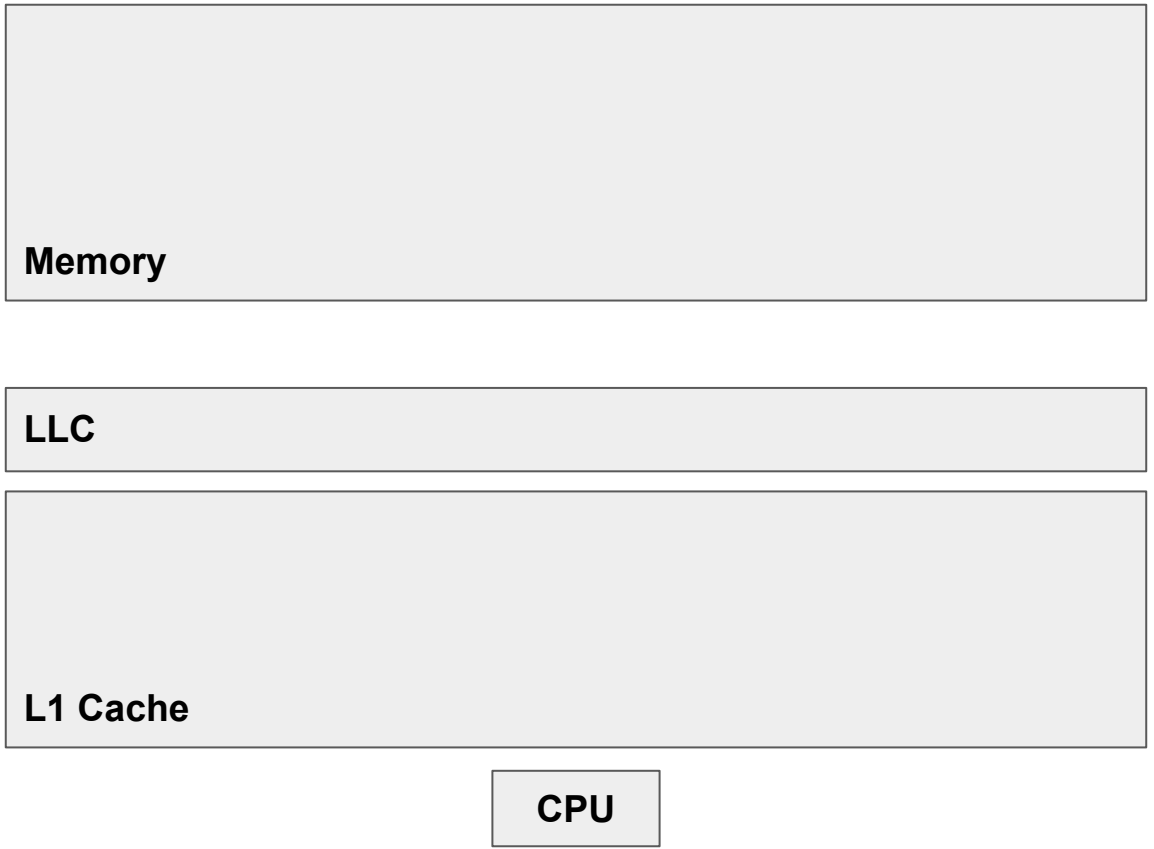
	Cores	NIC	DRAM	DRAM BW per core	NIC BW per core	Ratio
Google Cloud C3 2x Sapphire Rapids	176	200 Gbps	2x 8-ch DDR5	3.49 GB/s	0.14 GB/s	**24.93**
BlueField-3 SmartNIC DDR5	16	400 Gbps	2-ch DDR5	5.60 GB/s	3.13 GB/s	**1.79**

*Table data from *Lovelock: Towards Smart NIC-hosted Clusters* (2024) [2]

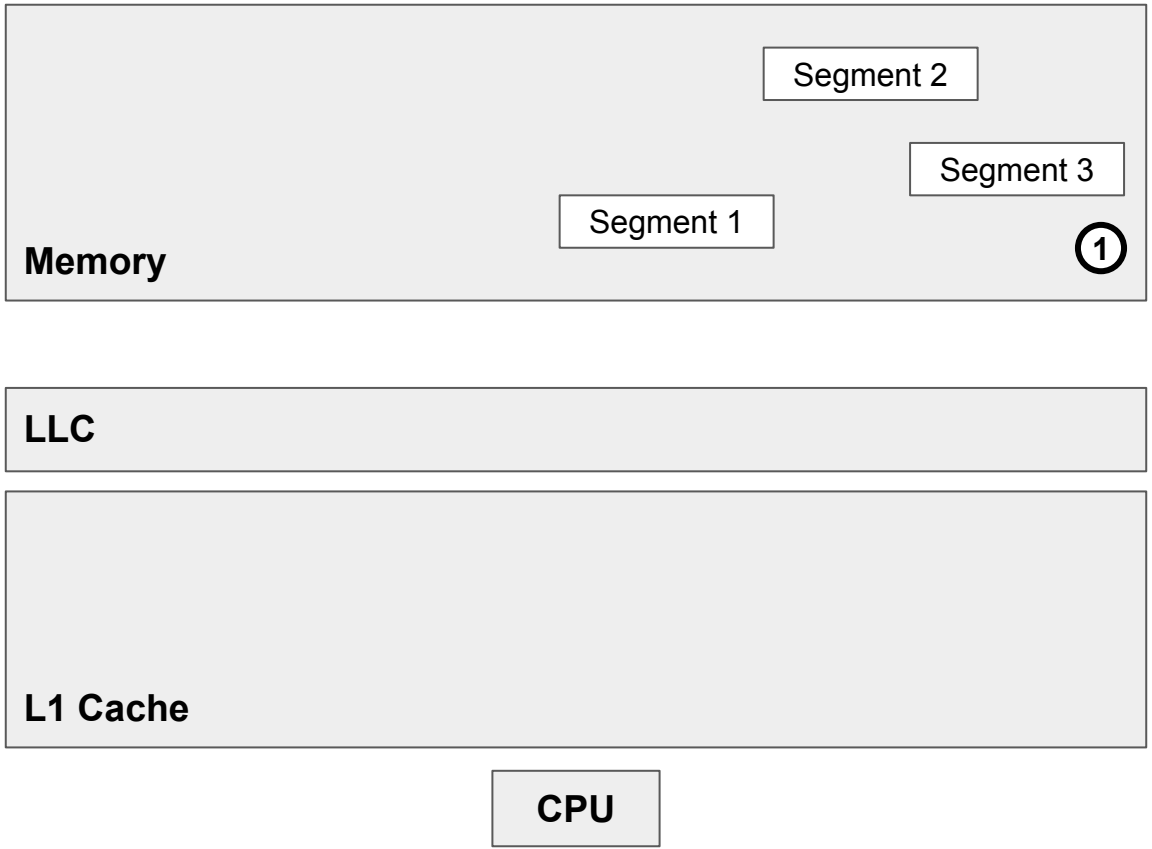
SmartNICs have limited memory bandwidth relative to their line rates

- Copying data to reassemble application bytestreams is expensive
- Careful cache management required

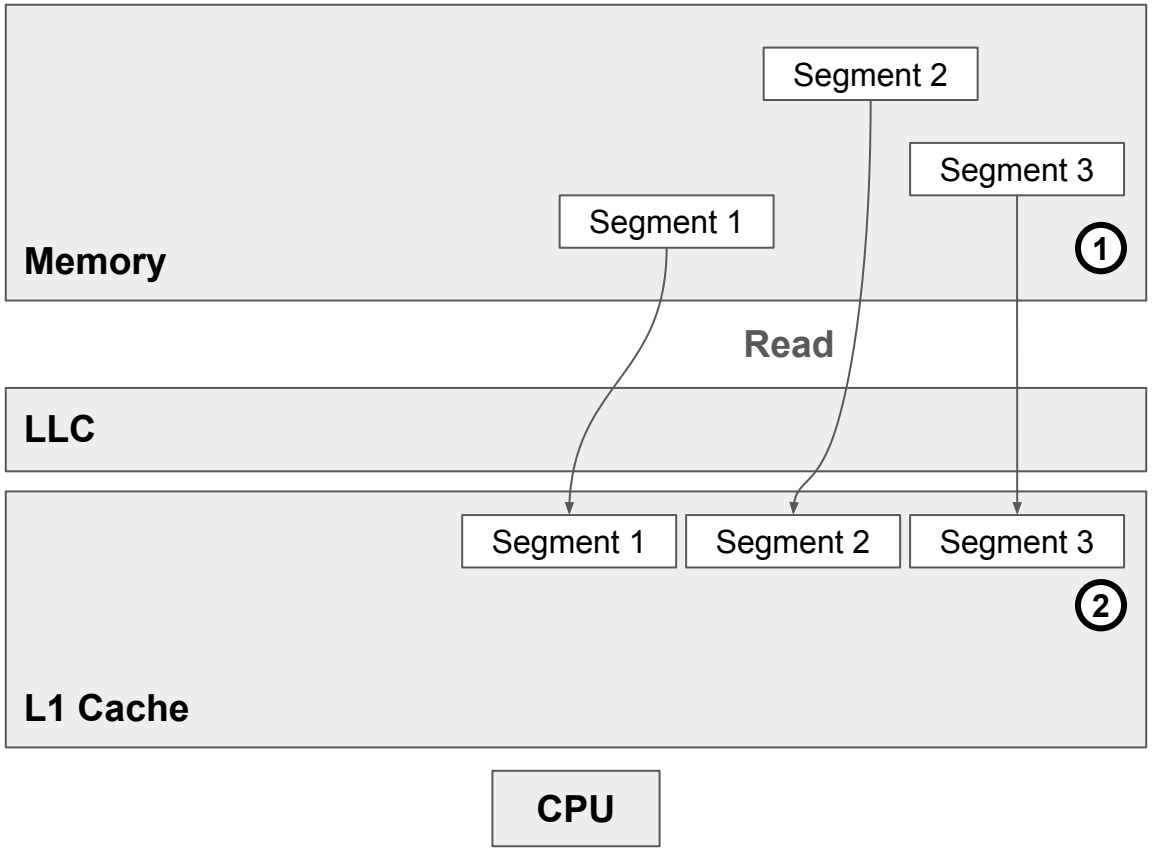
Example



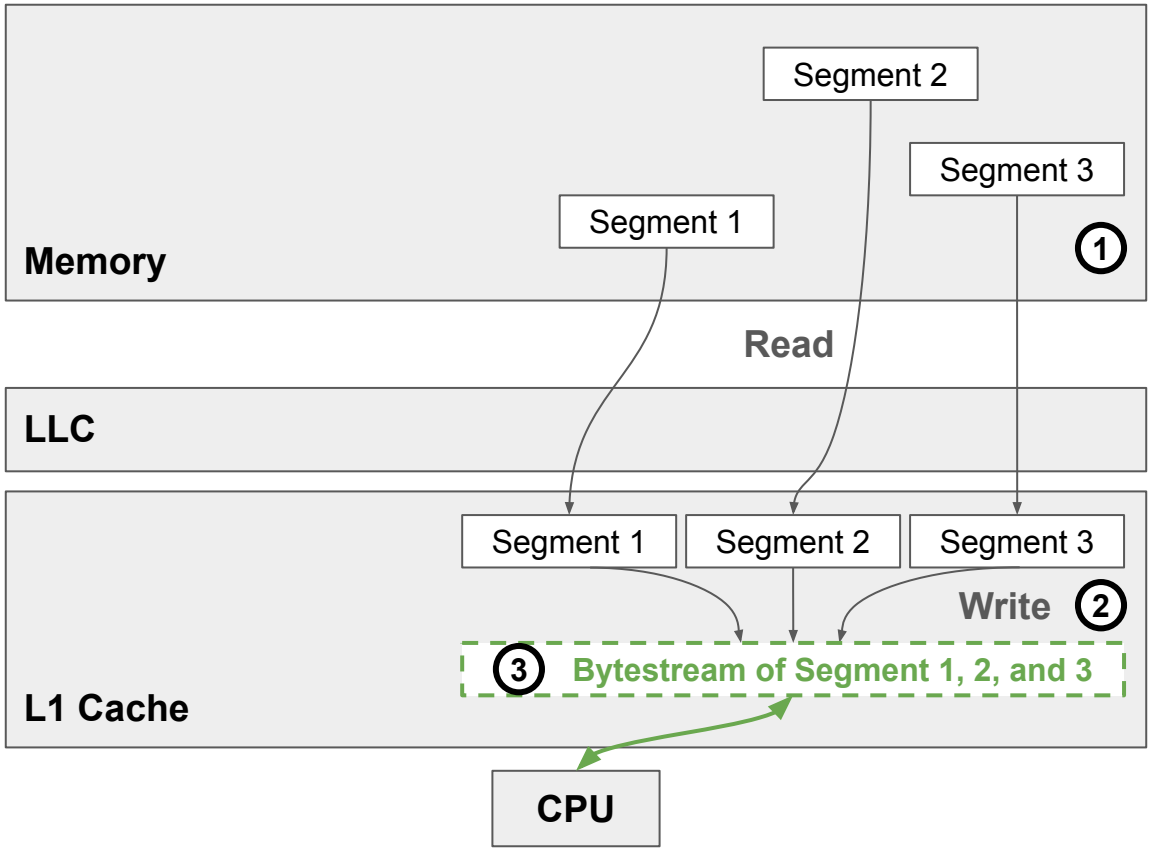
Example



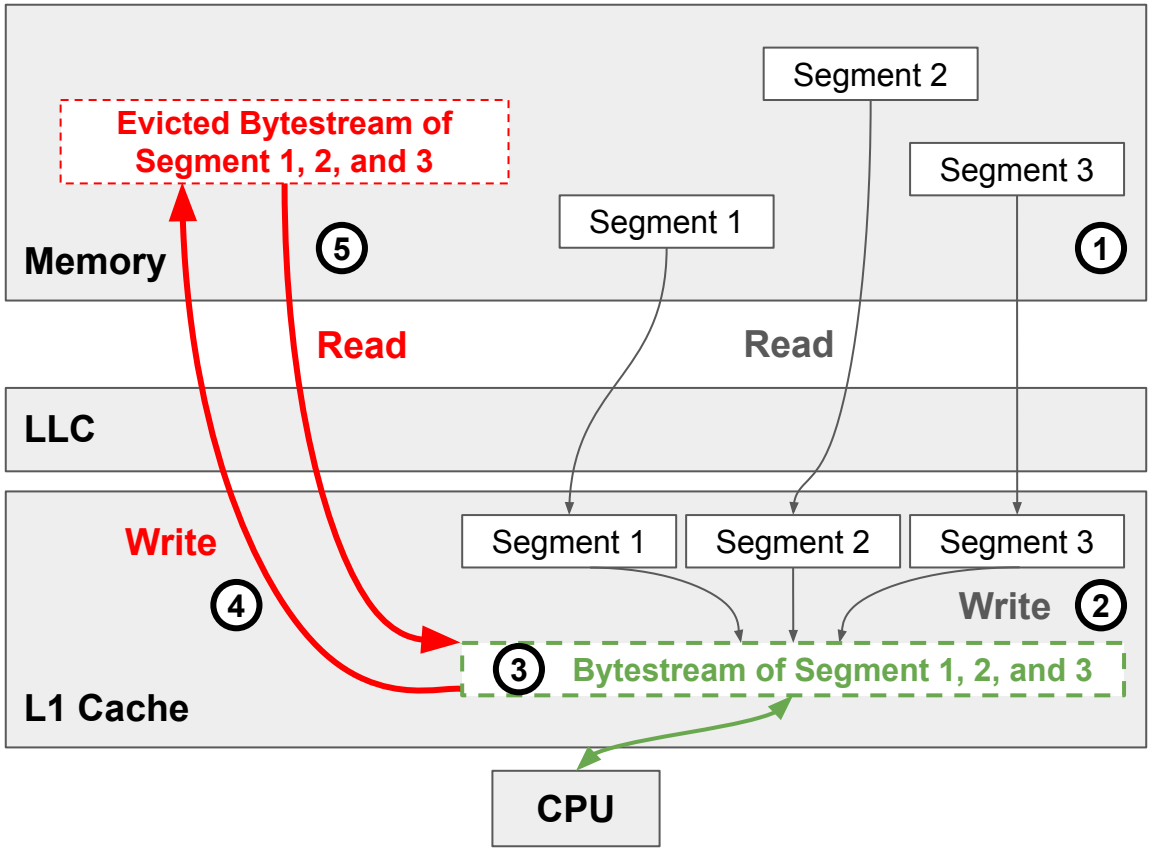
Example


Segment 2
Segment 3
Segment 1
Memory
1
Read
LLC
Segment 1
Segment 2
Segment 3
2
L1 Cache
CPU

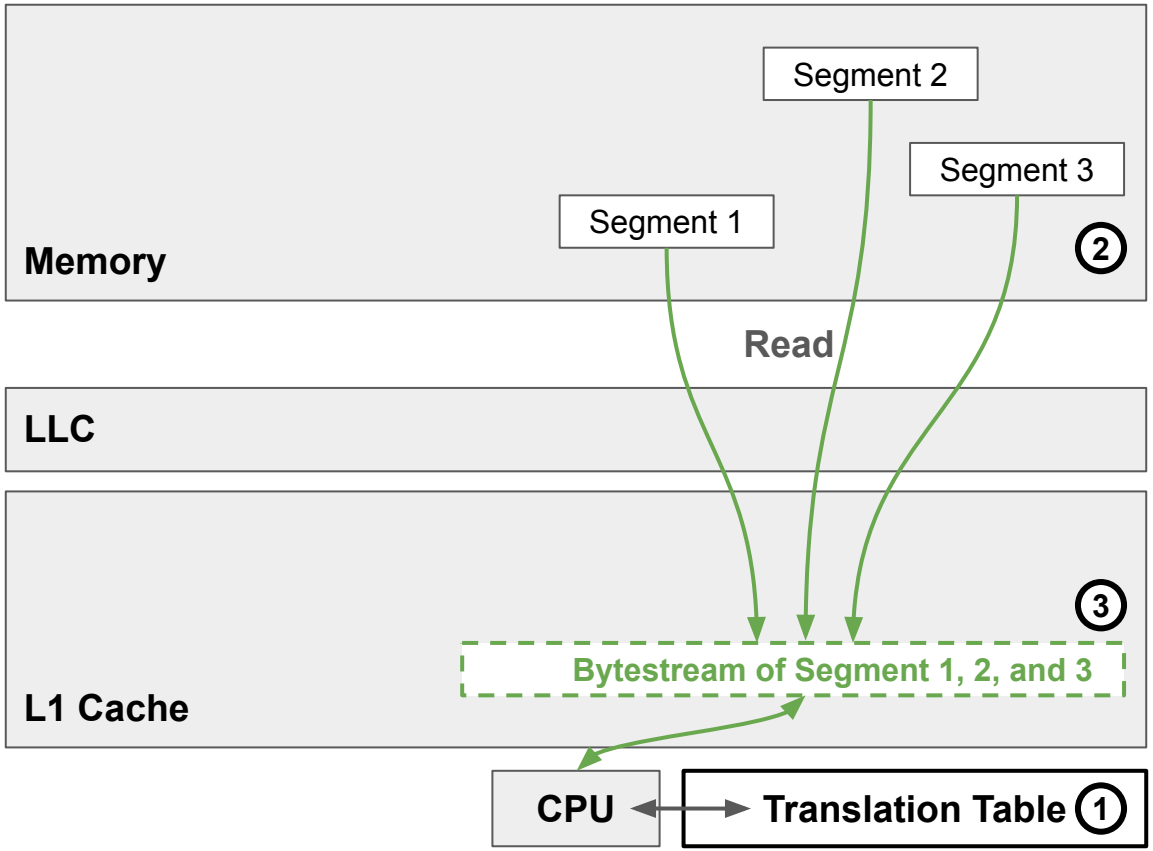
Example


Segment 2
Segment 3
Segment 1
Memory
1
Read
LLC
Segment 1
Segment 2
Segment 3
Write
2
3
Bytestream of Segment 1, 2, and 3
L1 Cache
CPU

Example


Segment 2
Evicted Bytestream of Segment 1, 2, and 3
Segment 3
Segment 1
Memory
5
1
Read
Read
LLC
Write
Segment 1
Segment 2
Segment 3
4
Write
2
3
Bytestream of Segment 1, 2, and 3
L1 Cache
CPU

Potential Solution



Virtual Gather

Using **virtual gather**, a SmartNIC can make the payloads of a series of packets addressable as a contiguous region in memory without a copy

The SmartNIC populates and manages a **translation table** of all transport-layer segments of a flow

Translation Table

Addr	Len	Off

Example: 4015 bytes received in 3 TCP segments (assume max segment size of 1460 bytes)

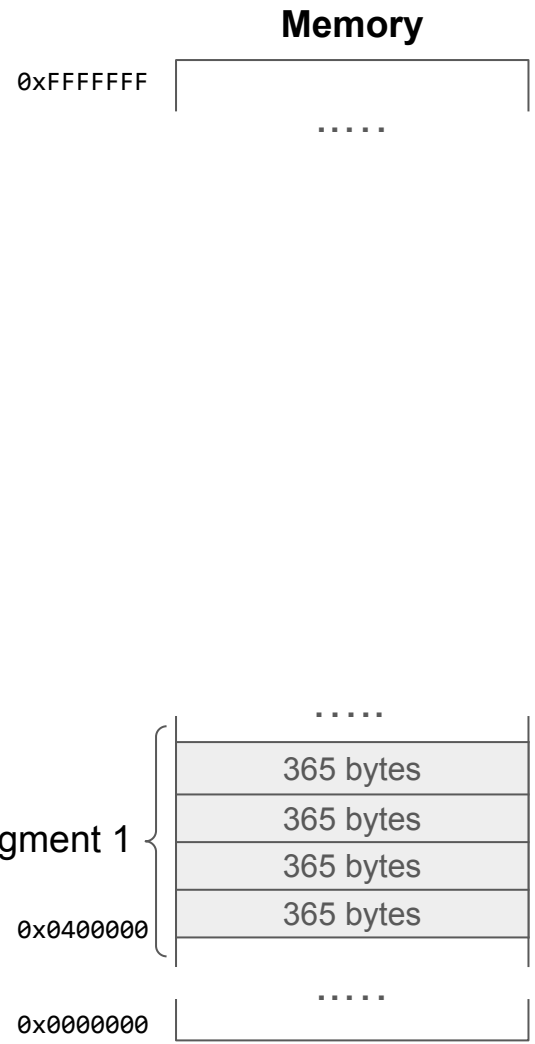
Virtual Gather

Using **virtual gather**, a SmartNIC can make the payloads of a series of packets addressable as a contiguous region in memory without a copy

The SmartNIC populates and manages a **translation table** of all transport-layer segments of a flow

Translation Table

Addr	Len	Off
0x0400000	0x5B4	



Example: 4015 bytes received in 3 TCP segments
(assume max segment size of 1460 bytes)

Virtual Gather

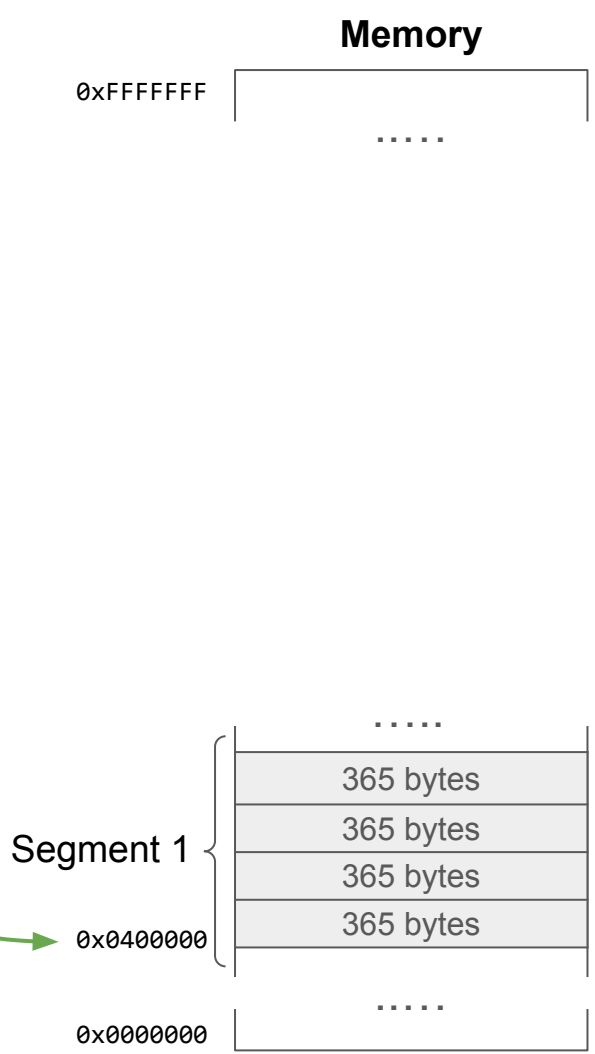
Using **virtual gather**, a SmartNIC can make the payloads of a series of packets addressable as a contiguous region in memory without a copy

The SmartNIC populates and manages a **translation table** of all transport-layer segments of a flow

Last Off	0x000
+ Len	0x5B4
New Off	**0x5B4**

Translation Table

Addr	Len	Off
0x0400000	0x5B4	**0x5B4**



Example: 4015 bytes received in 3 TCP segments
(assume max segment size of 1460 bytes)

Virtual Gather

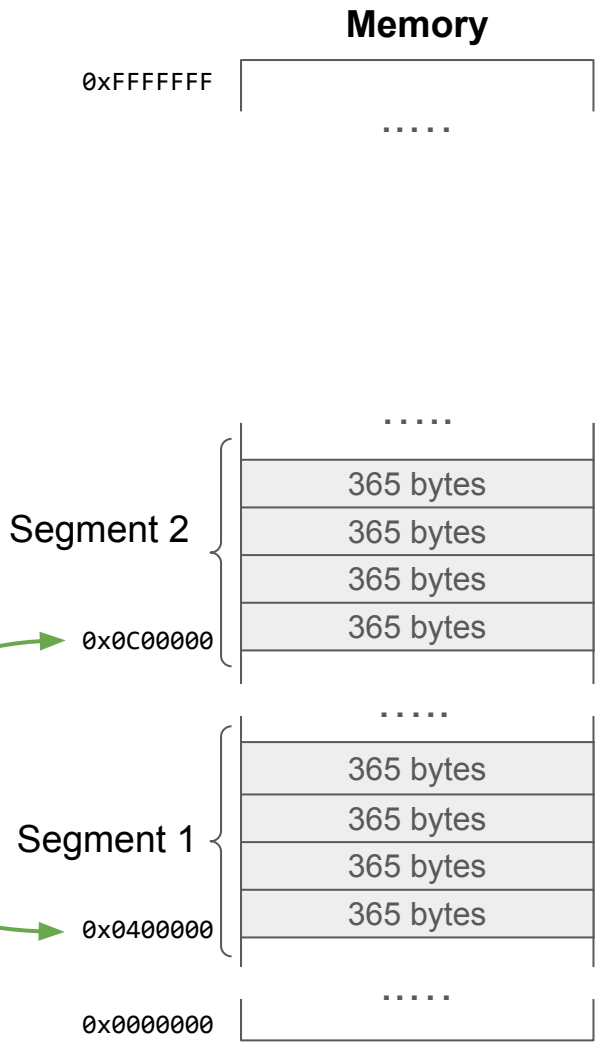
Using **virtual gather**, a SmartNIC can make the payloads of a series of packets addressable as a contiguous region in memory without a copy

The SmartNIC populates and manages a **translation table** of all transport-layer segments of a flow

Last Off	0x5B4
+ Len	0x5B4
New Off	**0xB68**

Translation Table

Addr	Len	Off
0x0400000	0x5B4	0x5B4
0x0C00000	0x5B4	**0xB68**



Example: 4015 bytes received in 3 TCP segments
(assume max segment size of 1460 bytes)

Virtual Gather

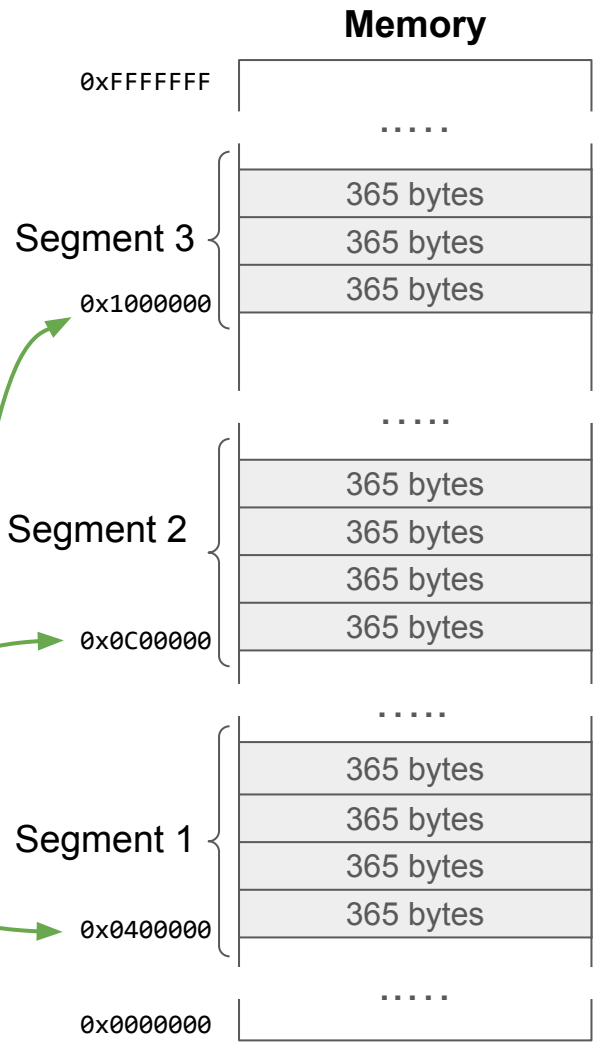
Using **virtual gather**, a SmartNIC can make the payloads of a series of packets addressable as a contiguous region in memory without a copy

The SmartNIC populates and manages a **translation table** of all transport-layer segments of a flow

Last Off	0xB68
+ Len	0x447
New Off	**0xFAF**

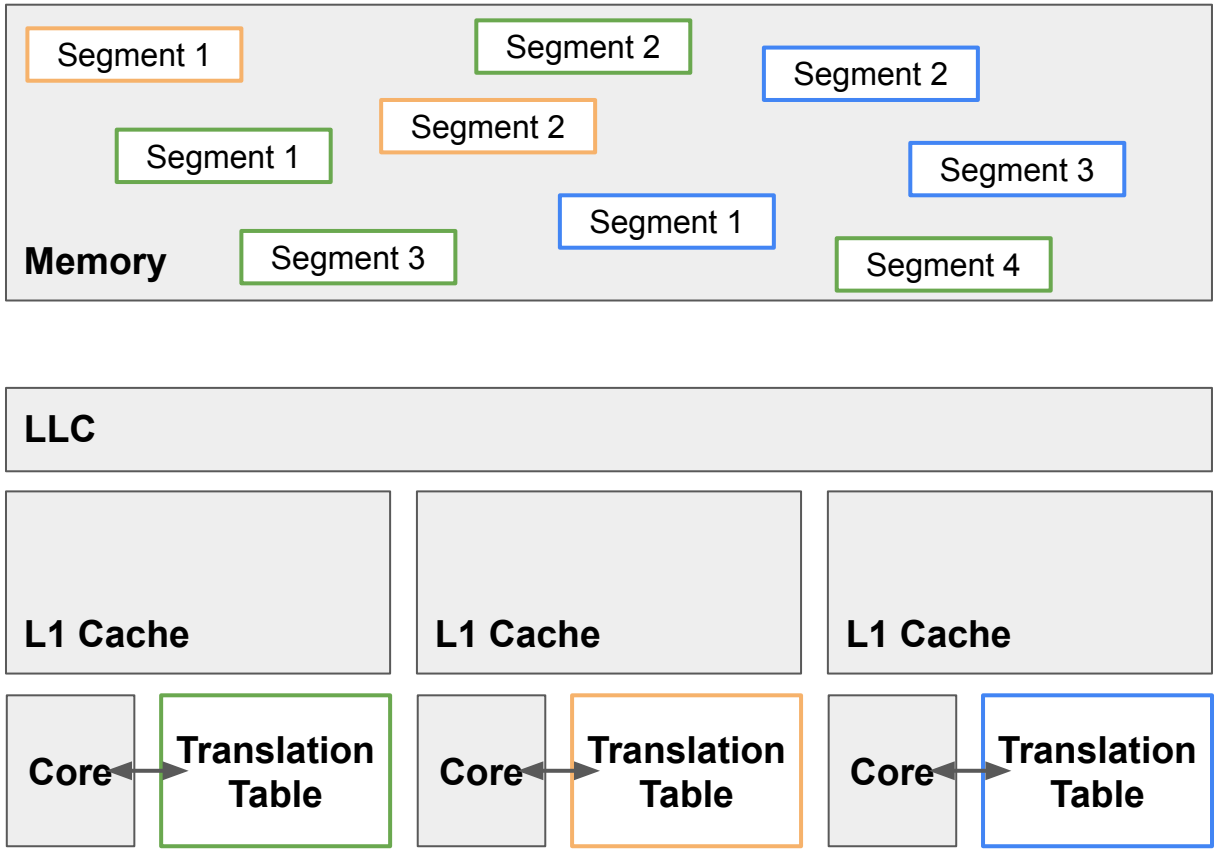
Translation Table

Addr	Len	Off
0x0400000	0x5B4	0x5B4
0x0C00000	0x5B4	0xB68
0x1000000	0x447	**0xFAF**



Example: 4015 bytes received in 3 TCP segments
(assume max segment size of 1460 bytes)

Virtual Gather



Virtual Gather

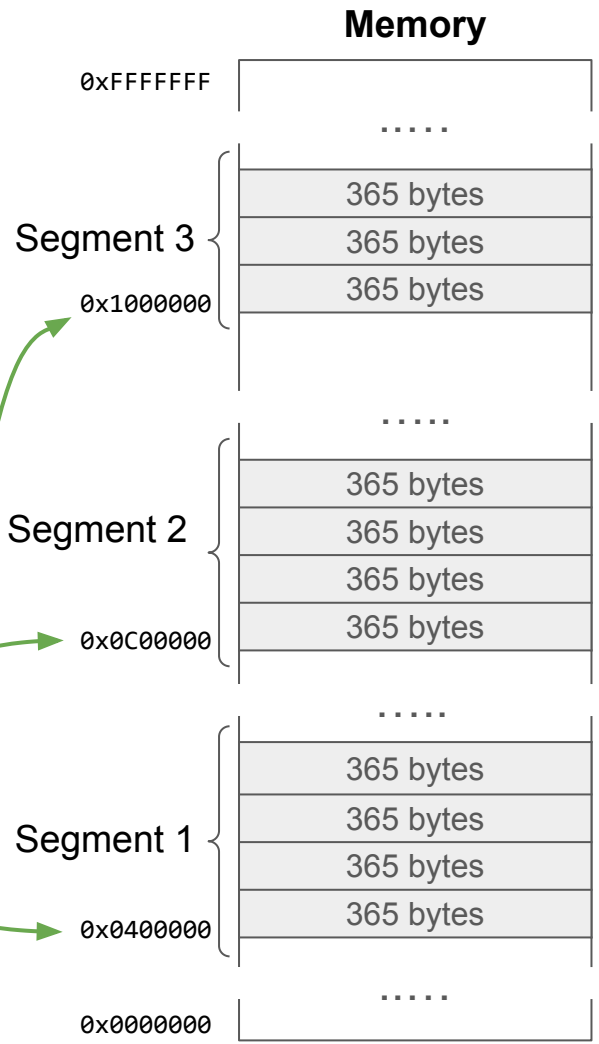
What makes this challenging? *22,000 new flows/sec*

How do you manage the translation table at line rate?
~16 nanosecs per packet at 400 Gbps

What network applications benefit?

Translation Table

Addr	Len	Off
0x0400000	0x5B4	0x5B4
0x0C00000	0x5B4	0xB68
0x1000000	0x447	0xFAF



Example: 4015 bytes received in 3 TCP segments
(assume max segment size of 1460 bytes)

Conclusions

- SmartNIC memory bandwidth is an emerging problem in packet processing for network analysis
- Packet processing for network analysis requires bytestream reassembly
- Copying data to reassemble the application bytestream is expensive
- With **virtual gather**, a SmartNIC can make the payloads of packets addressable as a contiguous region in memory without a copy

References

[1] Z. Zhao, H. Sadok, N. Atre, J. Hoe, V. Sekar, and J. Sherry, *Achieving 100Gbps Intrusion Prevention on a Single Server*, OSDI 2020

[2] S. Park, R. Govindan, K. Shen, D. Culler, F. Özcan, G-W. Kim, H. Levy *Lovelock: Towards SmartNIC-hosted Clusters*, HotCarbon 2024