Danaisis Vargas , Matthew Man, and William Dallaway



DUNE DAQ Core Software WG

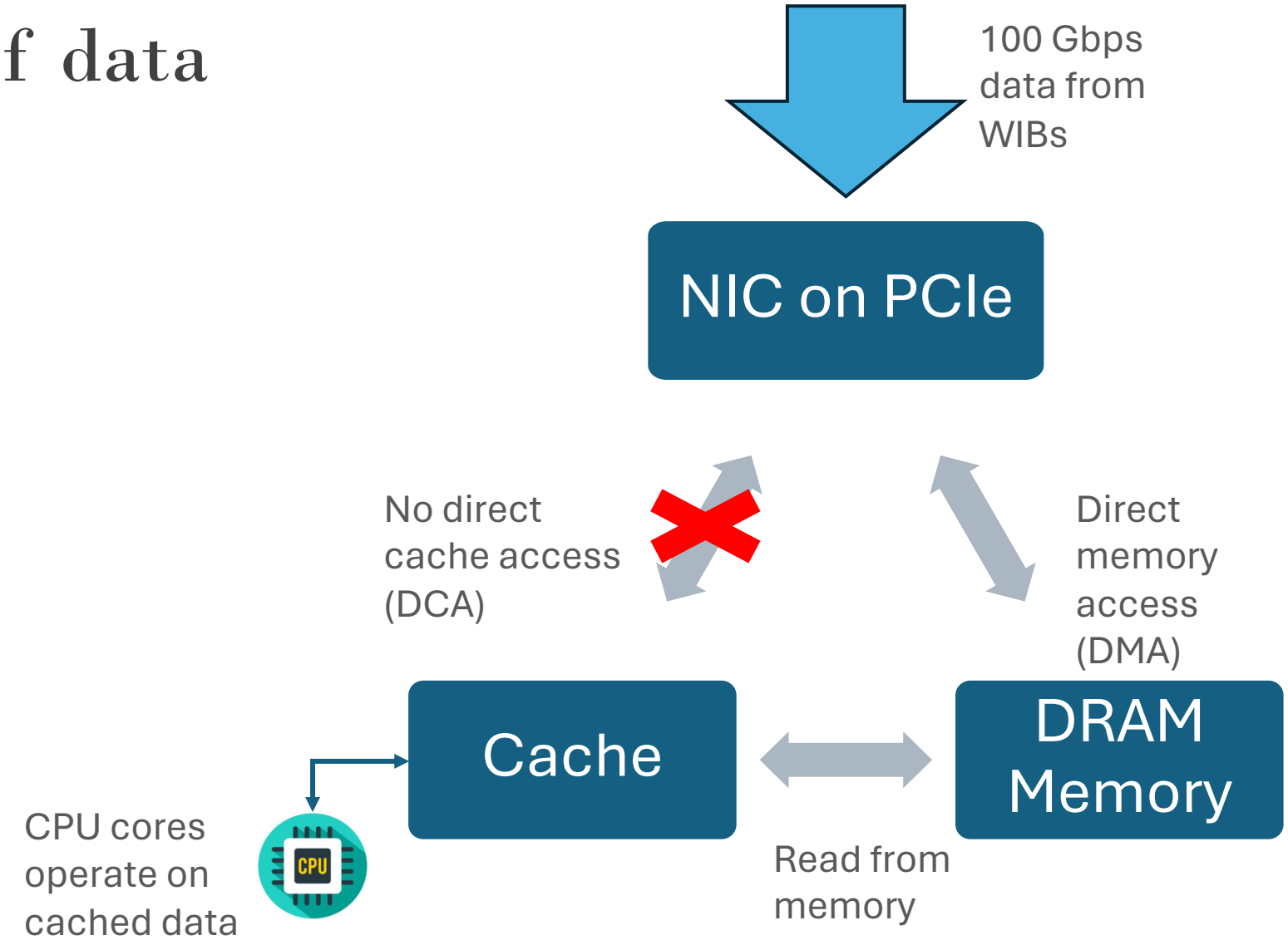
June 5th, 2024

Today in the talk

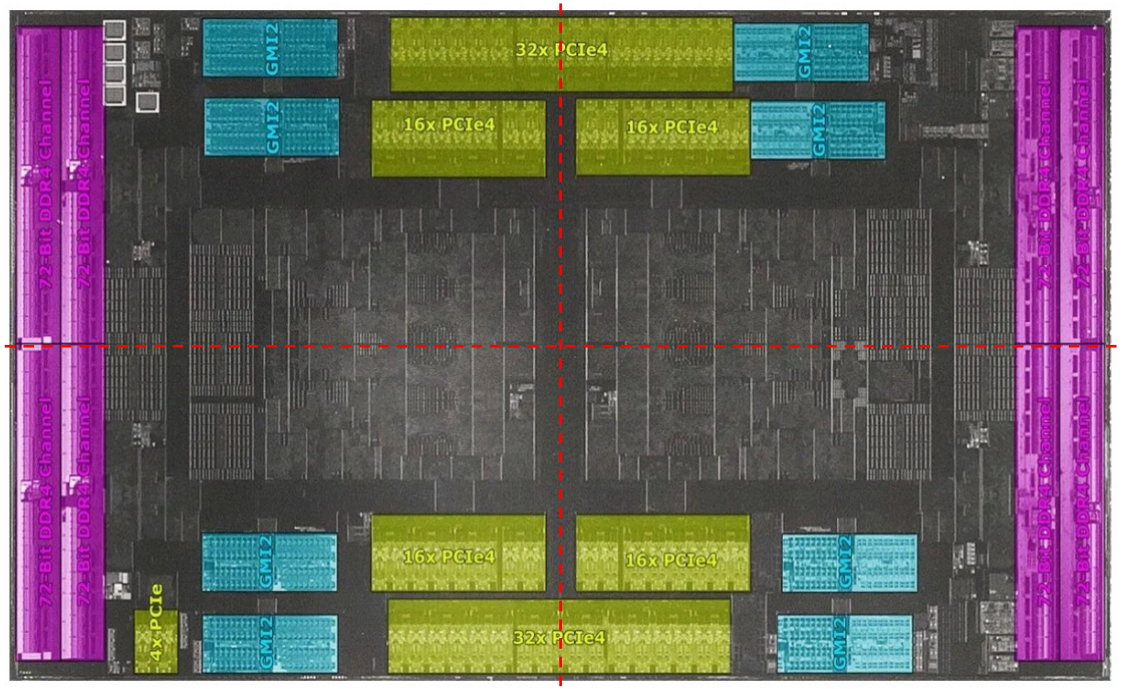
- AMD Zen3 Tuning: eliminating missed packets.
- np04-srv-005 storage server commissioning.
- Readout performance test summary.

AMD Zen3 Tuning: eliminating missed packets

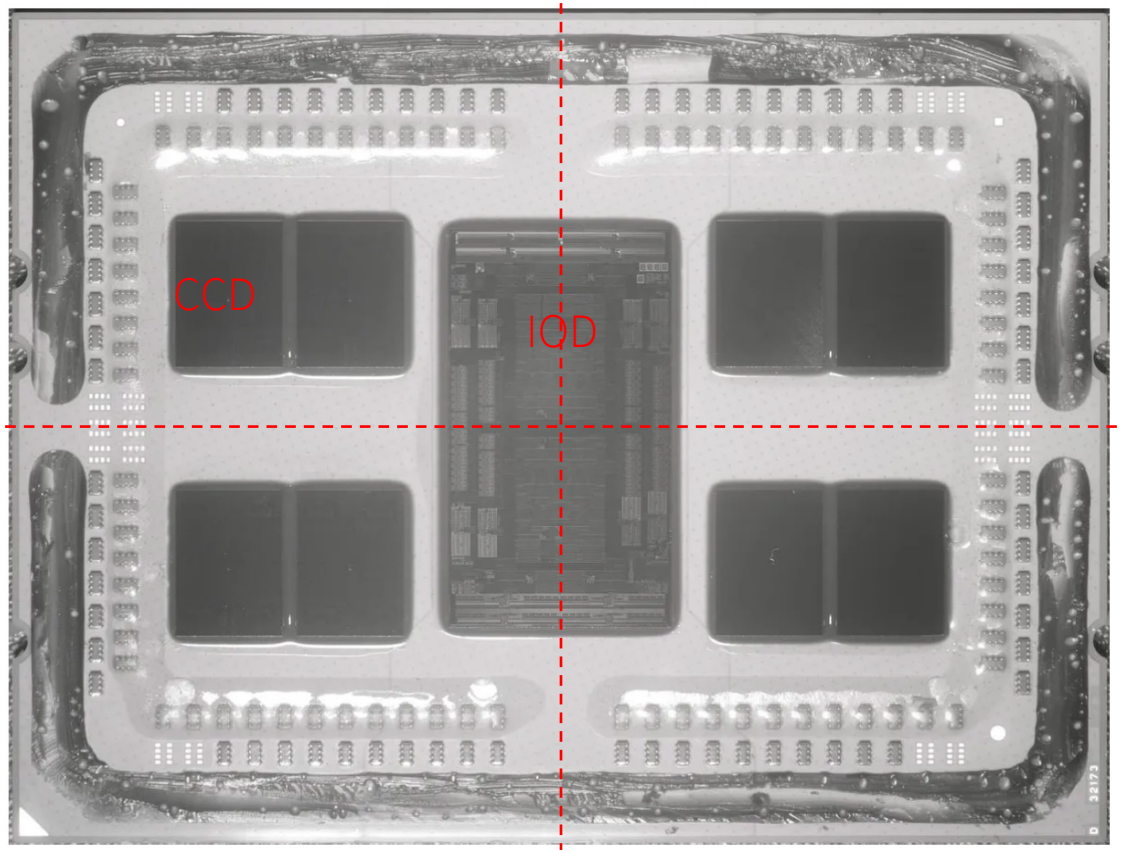
Flow of data


100 Gbps data from WIBs
NIC on PCIe
No direct cache access (DCA)
Direct memory access (DMA)
Cache
DRAM Memory
CPU cores operate on cached data
CPU
Read from memory

AMD Zen3 architecture



Single socket: connections



Single socket: dies

Note: pictured is Zen2 architecture but at this level they're equivalent

BIOS tuning

- np02-srv-001 and np02-srv-004 are same CPU family (Milan)
  - 001 has 4 cores x 4 CCDs per socket
  - 004 has 8 cores x 8 CCDs per socket
- On 001 the only BIOS setting necessary was to set the **LCKL frequency to maximum of 593 MHz**
  - Improves DMA from PCIe
- 004 needed more tuning, see table
  - Targeted at reducing memory latency
  - I included everything I changed, and highlighted the ones I think made the difference

np02-srv-004 BIOS settings

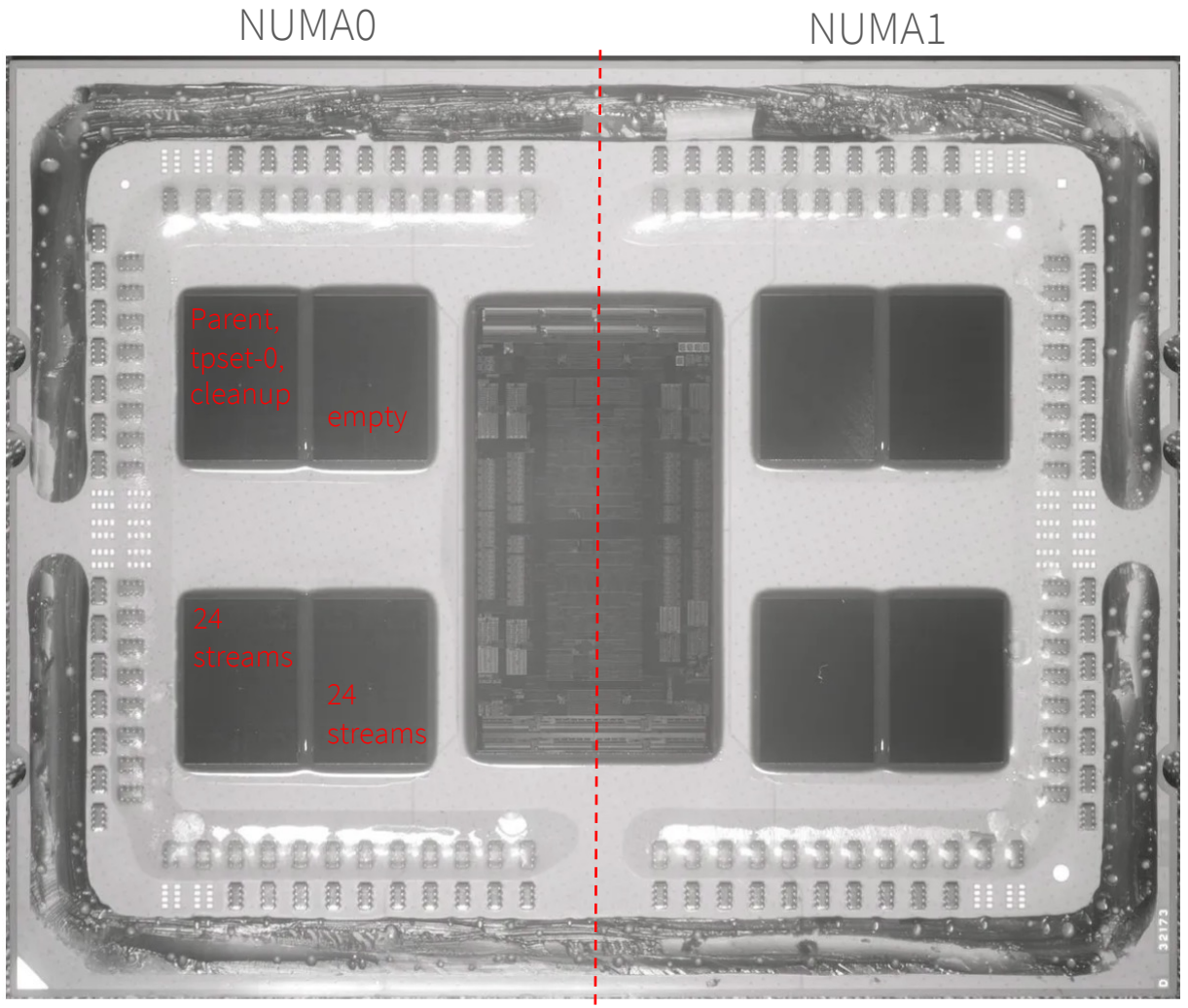
Setting	value	notes
LCKL frequency	593 MHz	
IOMMU	enable*	in kernel set iommu=pt and amd_iommu=on
relaxed ordering	disable	(on or off, similar performance)
determinism control	manual	
determinism slider	power	greater performance on CPUs
APBDIS	1	Enable fixed Infinity Fabric P-state control.
fixed SOC Pstate	P0	Highest-performing Infinity Fabric P-state. (try P1 if necessary)
DF C-states	disabled	turn off sleep states of infinity fabric
xGMI Link Width Control	manual	(dynamic link width management DLWM) manual control of inter-socket link width
xGMI Max Link Width Control	manual	for symmetric topology, we don't need inter-socket bandwidth, so set to min. 8 links instead of up to 16
xGMI Force Link Width Enable	force	
xGMI Force Link Width	1	1: 8 links 2: 16 links
xGMI Max Link Width	Auto	
ACPI SRAT L3 cache as NUMA domain	disabled	
NUMA nodes per socket	NPS2*	When changing NPS, make sure to add hugepages to `/opt/setup-100g-vfio.sh` script for each node

CPU pinning

Group by stream (48 WIB streams): rte-worker, post-proc, and recording handling the same streams grouped on same CCD

```
"--name runp02srv004eth0": {
  "parent": "1,2,3,4,129,130,131,132",
  "threads": {
    "rte-worker-16": "16",
    "rte-worker-20": "20",
    "rte-worker-24": "24",
    "rte-worker-28": "28",
    "tpset-0": "0,128",
    "cleanup-0": "0,128",
    "postproc-0-(100|101|102|103|116|117|118|119|132|133|134|135)": "17,18,145,146",
    "postproc-0-(104|105|106|107|120|121|122|123|136|137|138|139)": "21,22,149,150",
    "postproc-0-(108|109|110|111|124|125|126|127|140|141|142|143)": "25,26,153,154",
    "postproc-0-(112|113|114|115|128|129|130|131|144|145|146|147)": "29,30,157,158",

    "recording-(100|101|102|103|116|117|118|119|132|133|134|135)": "19,147",
    "recording-(104|105|106|107|120|121|122|123|136|137|138|139)": "23,151",
    "recording-(108|109|110|111|124|125|126|127|140|141|142|143)": "27,155",
    "recording-(112|113|114|115|128|129|130|131|144|145|146|147)": "31,159"
  }
},
```



What else is there to do?

- AMD Zen3 Tuning: eliminating missed packets.
  - Successful understanding and control of AMD architecture.
  - For PRR we should not be limited by Intel vs AMD CPU architectures.
  - I believe Intel machine 031 and AMD machine 004 are both capable of reading out 4 CRPs.
  - Technical point: DPDK needs to be rebuilt to allow rte-workers to use CPUs >128.
  - Power consumption/bare-minimum resources.
    - Are there BIOS settings that are unnecessary and can save power? (DF C-states, fixed P states)

np04-srv-005 storage server commissioning

Status

- All nvme drives prepared
  - 4 RAID level 10 arrays
  - /data2, /data3, /data4, /data5
  - Added into fstab to persist on reboot
- Kurt mentioned file-transfer metadata-creation script relies on /data0-3 pattern.
  - Do the RAID data directories need to be renamed?
  - How are /data0-1 used and how are /data2-5 used?

Testing

- Tested data writing with NP02 and NP04
- With four data writers we were able to sustain ~7Hz trigger rate from CRP4+5
  - the bottleneck is the network
- We used twelve datawriters for NP04
  - bottlenecks were from the trigger

What else is there to do?

Readout performance test summary

Tests

Emulation mode

- Stream scaling performance tests (with scaling for 8, 16, 24, 32, 40, and 48 streams).
  - with TPG enable
  - with raw recording enable
  - with TPG and raw recording enable
  - before and after applying kernel params

- Stream scaling performance tests (with scaling for 8, 16, 24, 32, 40, and 48 streams).
  - with TPG and raw recording enable
- Reading one APA / CRP per server (using only one NUMA node) when possible.
  - using differents TPG methods: AbsRS, SilverBullet, and SimpleThreshold
  - before and after applying kernel params
- Reading two APAs / CRPs per server when possible.
  - 2 x 100 Gb NICs
- Reading four APAs / CRPs.
  - 2 x 200 Gb NICs

Readout performance test summary

Test	np02-srv-001	np02-srv-002	np02-srv-003	np02-srv-004	np04-srv-031	np04-srv-021	np04-srv-022	np04-srv-028	np04-srv-029
Emulation									
Stream scaling w TPG	X	X	X	X					
Stream scaling w raw recording			X	X					
Stream scaling w TPG and recording			X	X					
Stream scaling before and after applying kernel params				X					
Data									
Stream scaling w TPG and recording			X						
TPG methods						X	X	X	X
Reading one APAs / CRPs	X	X	X	X	X	X	X	X	X
Reading two APA / CRP	possible	possible	X	X	X				
Reading four APA / CRP				possible	possible				

What else is there to do?

- Readout performance test summary.
  - The performancetest app (https://github.com/DUNE-DAQ/performancetest) is up to date.
    - How to conduct benchmark and performance tests.
    - How to process and present the results.
  - All configuration are on np04 (https://gitlab.cern.ch/dune-daq/online/np04daq-configs.git)
  - Reading two **APAs / CRPs**. We would like to test:
    - np02-srv-001 (AMD) with 2 x 100 Gb NICs
    - np02-srv-002 (Intel) with 2 x 100 Gb NICs
  - Reading four **APAs / CRPs**. This will be possible with servers:
    - np04-srv-031 (Intel) with 2 x 200 Gb NICs
    - np02-srv-004 (AMD) with 4 x 100 Gb or 2 x 200 Gb NICs
  - Minimum resources test

References

AMD tuning guides:

https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/data-plane-development-kit-tuning-guide-amd-epyc7003-series-processors.pdf

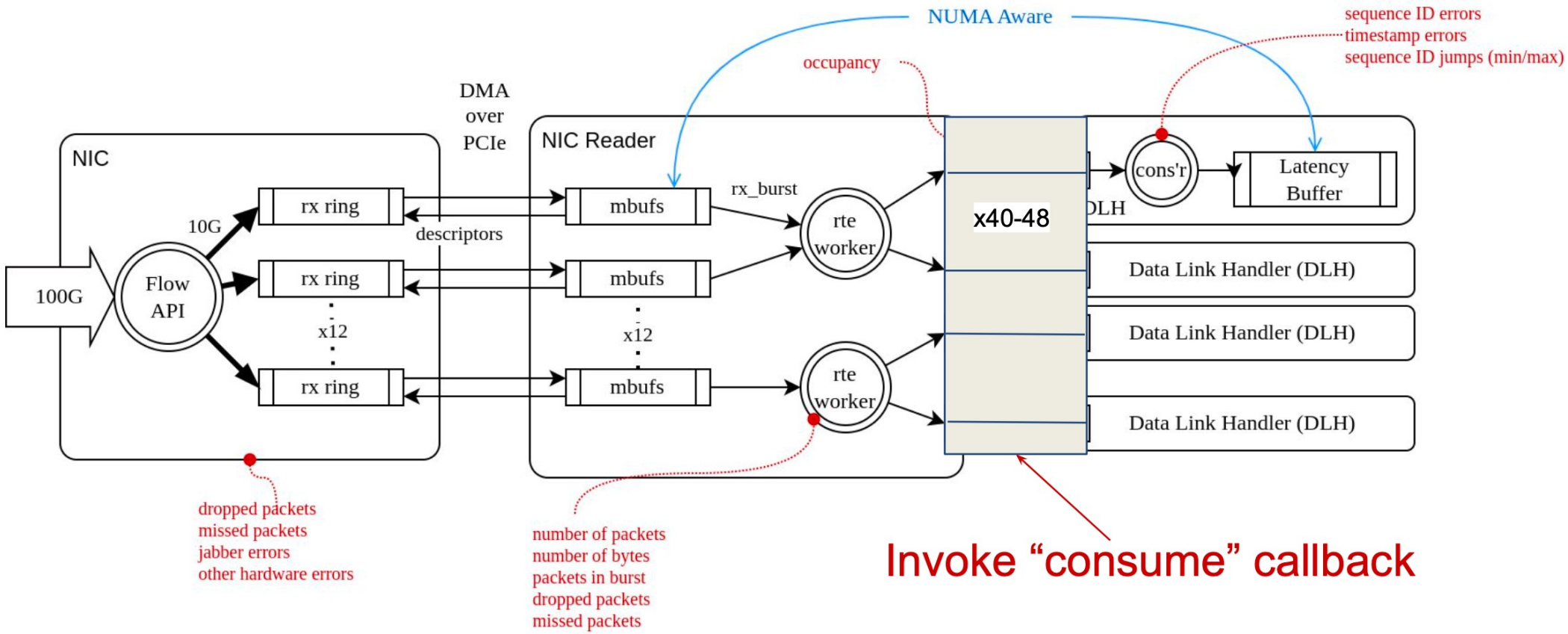
https://www.amd.com/content/dam/amd/en/documents/epyc-technical-docs/tuning-guides/amd-epyc-7003-tg-workload-57011.pdf

AMD topology:

https://www.anandtech.com/show/16529/amd-epyc-milan-review/4

Thank you

Callbacks


NUMA Aware
sequence ID errors
timestamp errors
sequence ID jumps (min/max)
occupancy
DMA over PCIe
NIC
NIC Reader
100G
Flow API
10G
rx ring
rx ring
x12
rx ring
descriptors
mbufs
mbufs
x12
mbufs
rx_burst
rte worker
rte worker
x40-48
cons'r
Latency Buffer
DLH
Data Link Handler (DLH)
Data Link Handler (DLH)
Data Link Handler (DLH)
dropped packets
missed packets
jabber errors
other hardware errors
number of packets
number of bytes
packets in burst
dropped packets
missed packets
Invoke “consume” callback