No IO Left Behind: Emerging Platforms for In-Flight Data Processing

Philippe Cudré-Mauroux

Slides by Alberto Lerner
University of Fribourg – Switzerland

IFIP Meeting
September 2021

eXascale Infolab

eXascale Infolab (XI)

- Lab @ U. of Fribourg–Switzerland
- Data Infrastructures for social / scientific / AI applications

https://exascale.info/

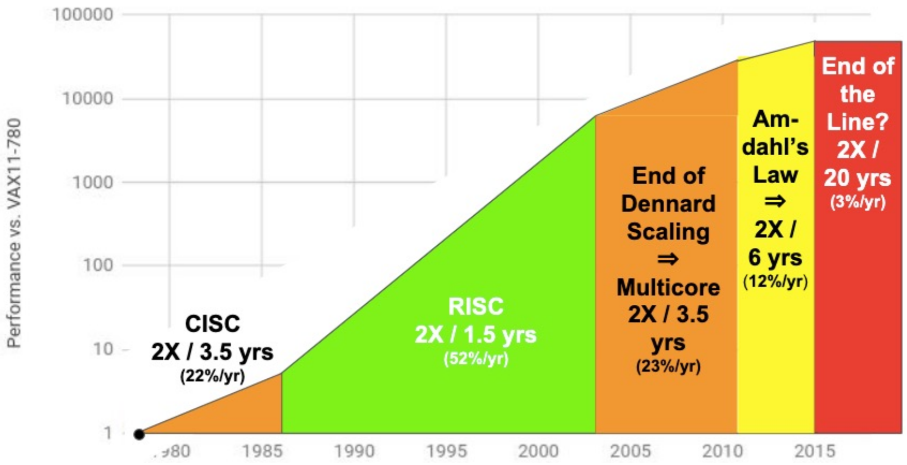
eXascale Infolab

erc European Research Council FNSNF HORIZON 2020 HASLER STIFTUNG SAP VERISIGN Google armasuisse amazon

Motivation

- End of growth of single program speed (Patterson and Hennessy Turing Award lecture @ ISCA'18)
- Specialization is the answer!

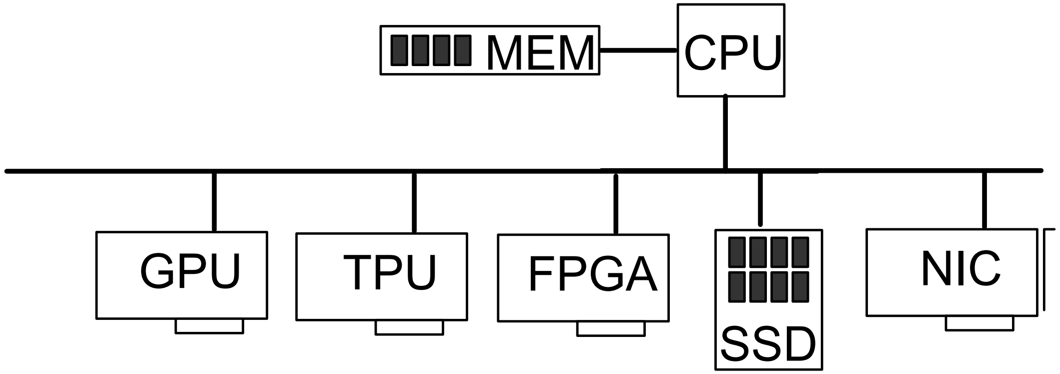
40 years of Processor Performance



eXascale Infolab

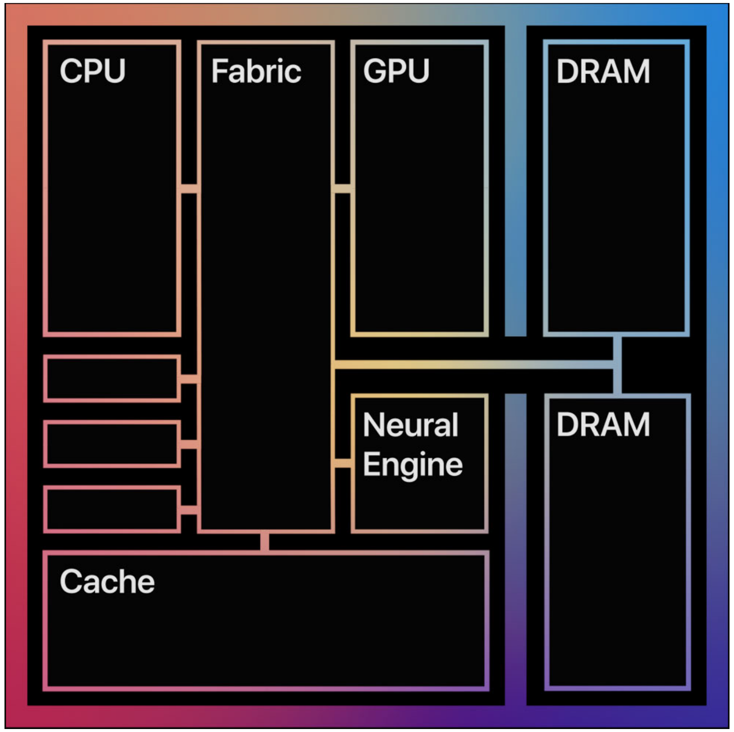
Specialization I

- Different computing units offer different functionalities



Specialization I

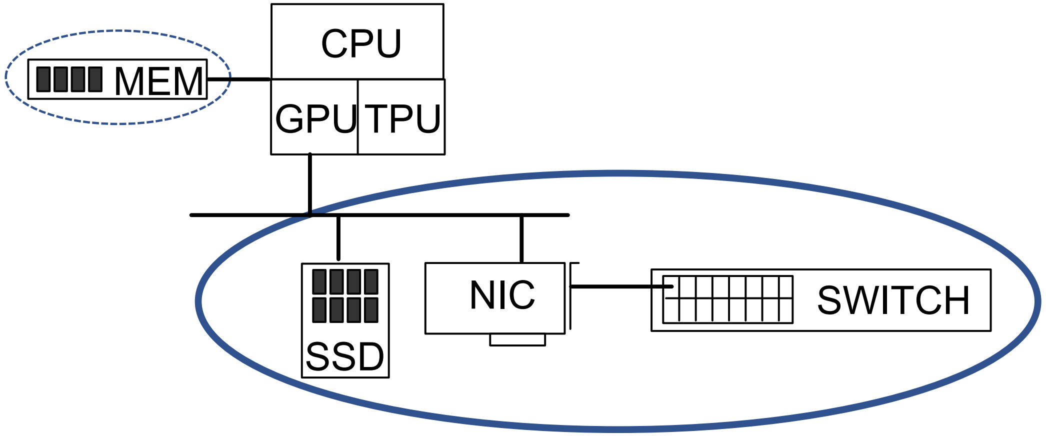
- Different computing units offer different functionalities
- A recent example: the M1 chip from Apple



Apple

Specialization II

- Different computing units offer different functionalities
- A recent example: the M1 chip from Apple
- Push functionality to units that were “passive” so far
  - No I/O should go untapped!



Agenda

- Network Switches as Accelerators
- Network Cards as Accelerators
- SSDs as Accelerators
- Research Agenda
  - Programming Models
  - New Execution Engines

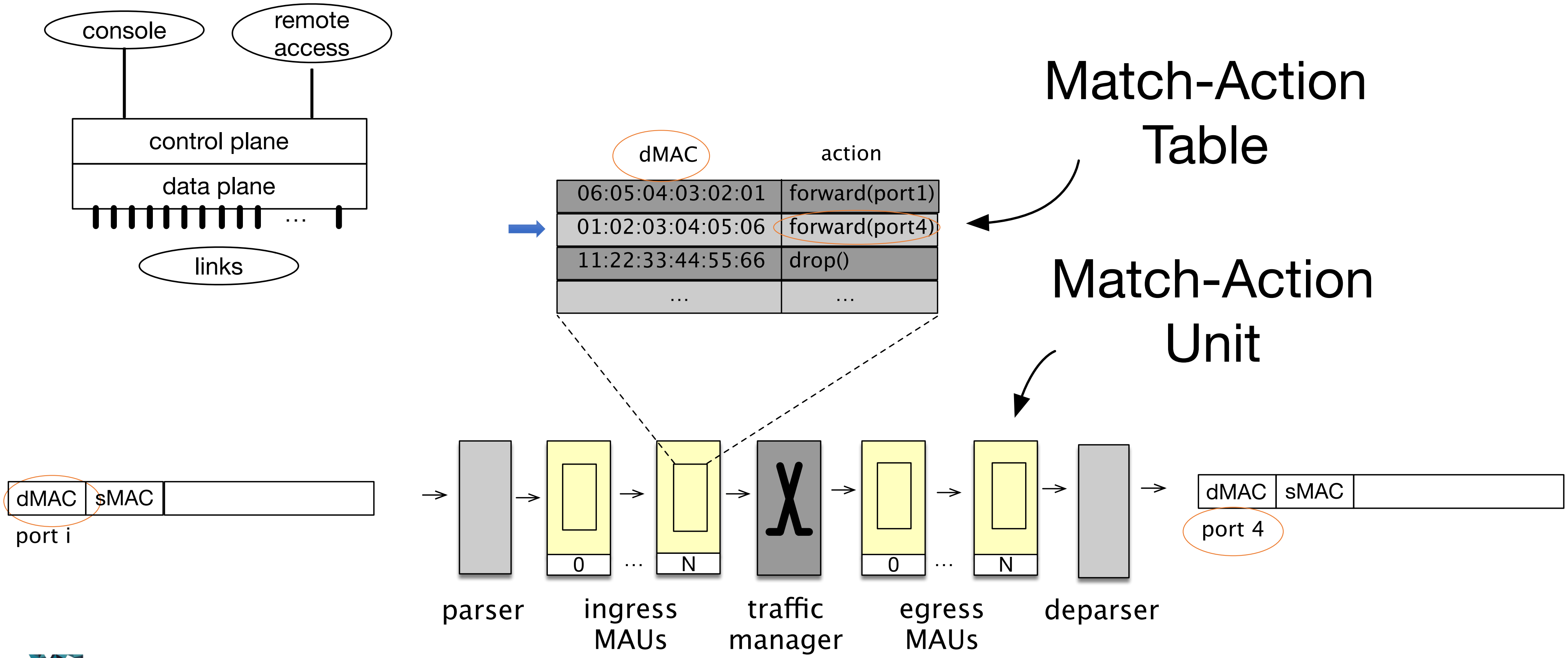
Alternative ***Programmable*** Platforms

Use Case(s)

Switches

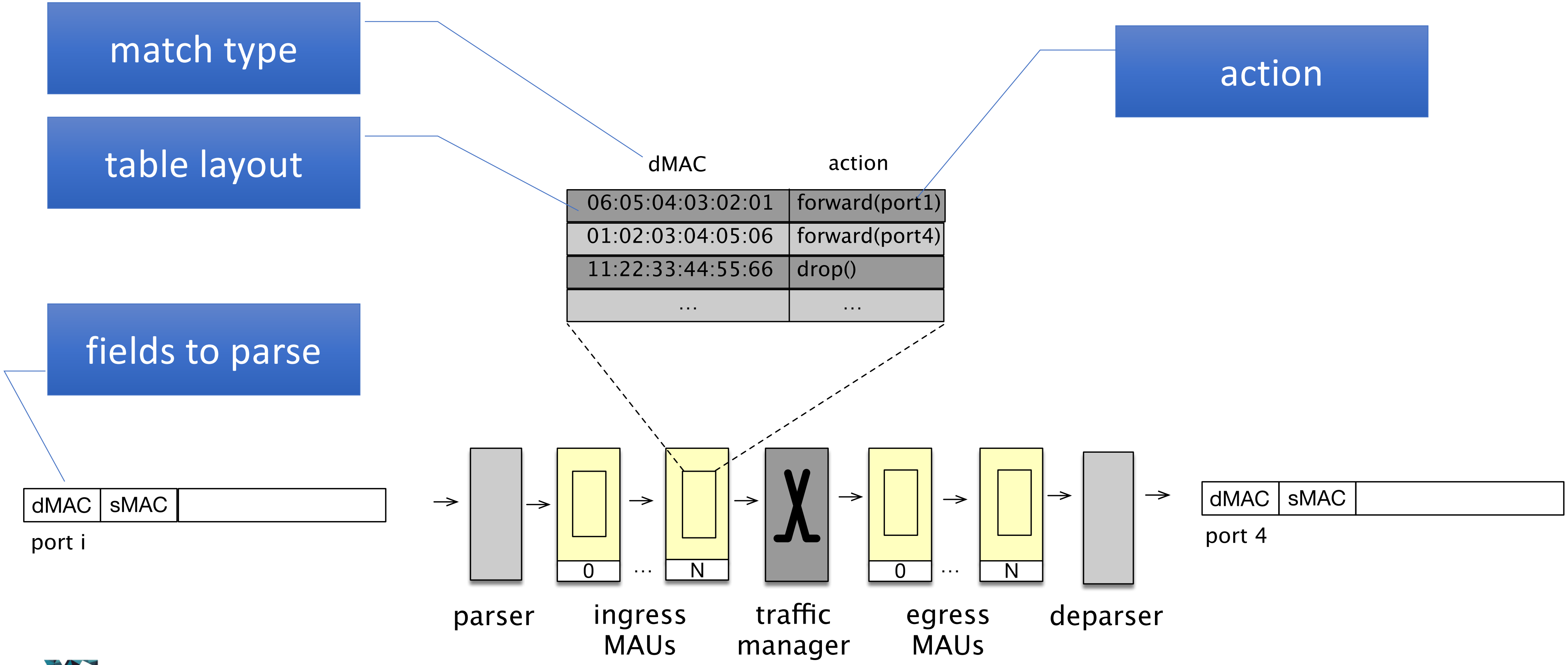
Compute while transmitting

Anatomy of a Programmable Switch



dMAC	action
06:05:04:03:02:01	forward(port1)
01:02:03:04:05:06	forward(port4)
11:22:33:44:55:66	drop()
…	…

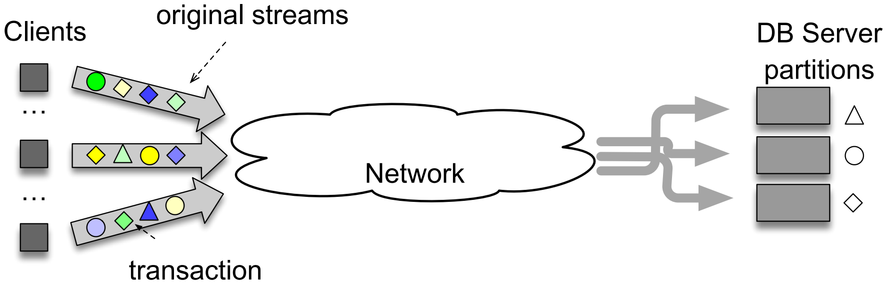
What Is Programmable?



dMAC	action
06:05:04:03:02:01	forward(port1)
01:02:03:04:05:06	forward(port4)
11:22:33:44:55:66	drop()
…	…

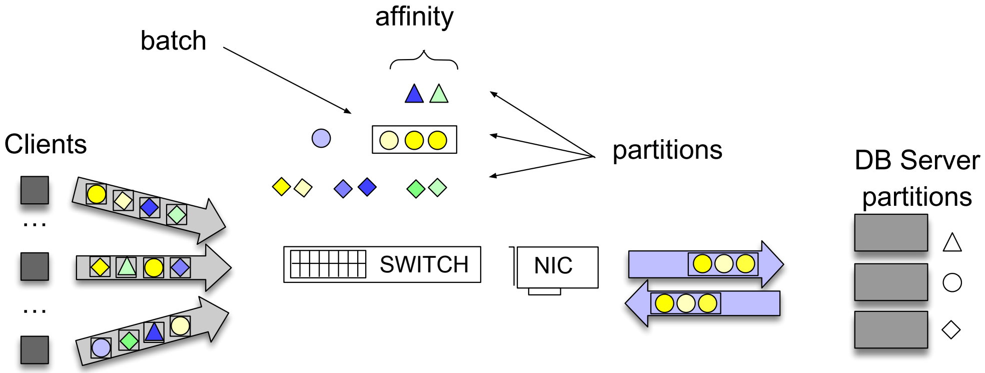
Use Case: Overhead in OLTP workloads

- In-memory databases partition data across cores/threads
- Clients issue streams of small transactions at random
- They are delivered to an arbitrary database core/thread
- Networking overhead in TPC-C is 53%; even higher in YCSB
  - Yes, RDMA helps but consumes CPU that would otherwise be running transaction processing!



Transaction Triaging

- Coordinate switch and NIC to deliver transactions [VLDB'21]
- This means delivering transactions
  - In batches
    - Both requests and responses
  - Separated by partitions
  - Ordered by affinity
- Results:
  - 7.95x faster than UDP networking
  - 1.9x faster than RDMA networking



Theo Jepsen, Alberto Lerner, Fernando Pedone, Robert Soulé, and Philippe Cudré-Mauroux. "**In-Network Support for Transaction Triaging**." In *Proceedings of the VLDB Endowment*, 14:1626–39, 2021.

Other Opportunities

- Offloading query processing [CIDR'19]
- Offloading graph analytics [In Preparation]
- External Memory
- …

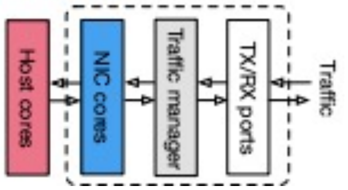
Alberto Lerner, Rana Hussein, and Philippe Cudré-Mauroux. "**The Case For Network Accelerated Query Processing**." In *CIDR 2019, 9th Biennial Conference on Innovative Data Systems Research*, 2019.

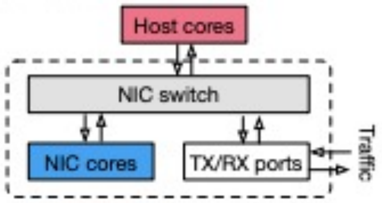
NICs

Compute between the Server and the Wire

Smart NIC Ecosystem

- Many different specialties but most can be classified by this quadrant



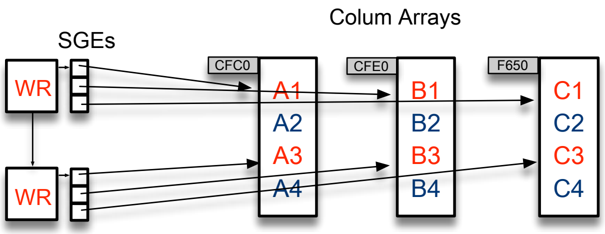


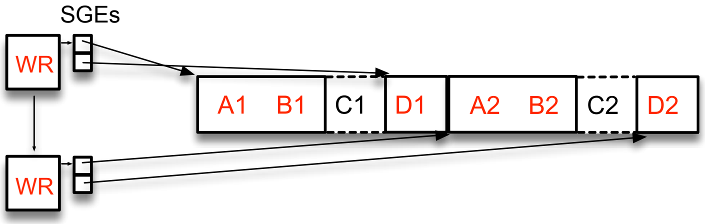
[Liu’19]

	NPU (software)	FPGA
On Path	Netronome Agilio	NetFPGA Corundum
Off Path	Mellanox Bluefield	Mellanox Inova

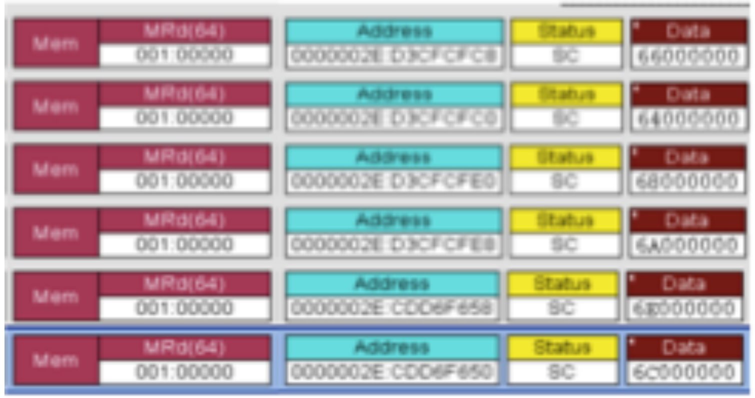
Use Case: *Actual* Zero-Copy RDMA

- Databases often transmit data that is very fragmented





- The card does not optimize the transfers



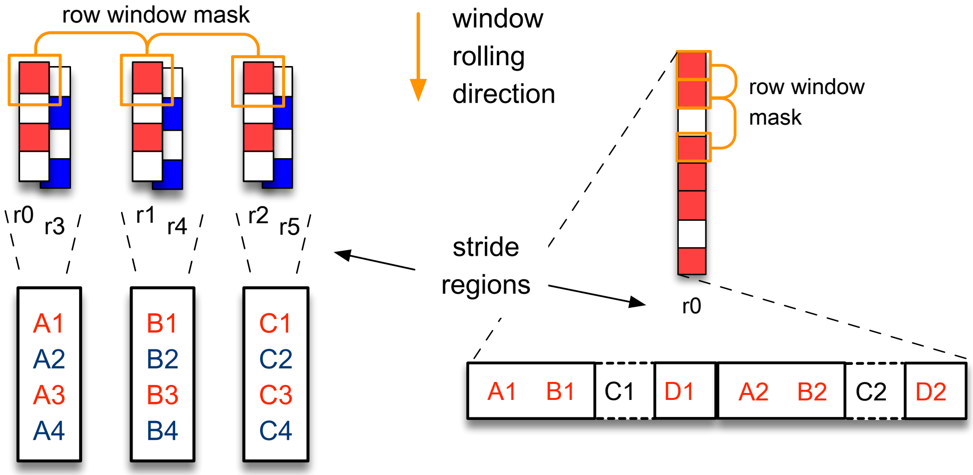
```
Mem  MRd(64)  001:00000   Address  0000002E:D3CFCFC8   Status  SC   Data  66000000
Mem  MRd(64)  001:00000   Address  0000002E:D3CFCFC0   Status  SC   Data  64000000
Mem  MRd(64)  001:00000   Address  0000002E:D3CFCFE0   Status  SC   Data  68000000
Mem  MRd(64)  001:00000   Address  0000002E:D3CFCFE8   Status  SC   Data  6A000000
Mem  MRd(64)  001:00000   Address  0000002E:CDD6F658   Status  SC   Data  6B000000
Mem  MRd(64)  001:00000   Address  0000002E:CDD6F650   Status  SC   Data  6C000000
```

- Databases copy the data to a contiguous buffer prior to transmitting

eXascale Infolab

D-RDMA: Optimize the DMA schedule

- Extension to RDMA protocol [Submitted]
- NIC receives what to transmit instead of how to transmit
  - Declarative!
- NIC decides how to best DMA the data
- Preliminary Results:
  - Copy buffer: 100% CPU->18Gbps
  - D-RDMA: virtually no CPU->98Gbps

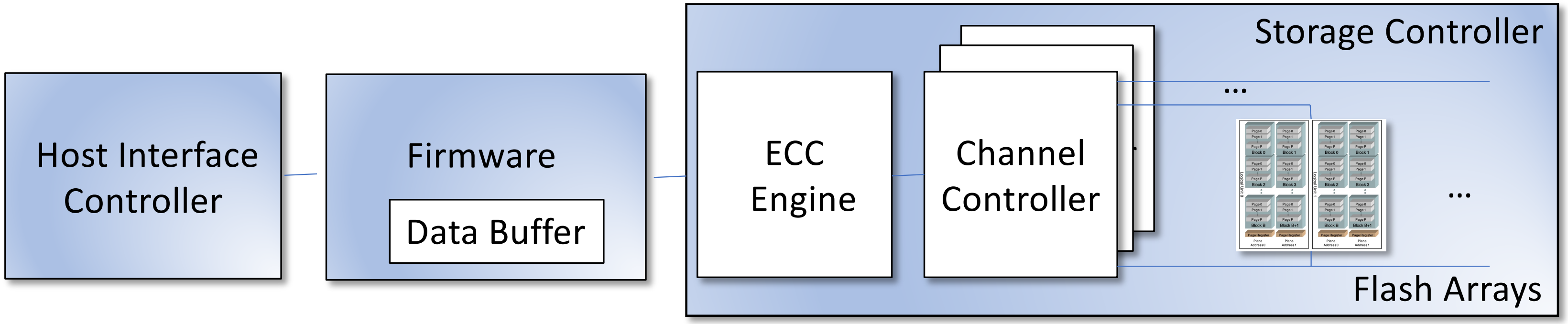


André Ryser, Alberto Lerner, Alex Forencich, Philippe Cudré-Mauroux. “**D-RDMA: Bringing Zero-Copy RDMA to Database Systems** .” Submitted.

SSDs

Compute between the Server and the Flash Array

SSDs Are Powerful Devices



HIC

- Implements the NVMe controller
- Performs data transfers in and out of the device for 100’s K cmds/sec

Firmware

- Implements the FTL (page mapping, wear leveling, and GC)
- Not only FTL, but also:
    - Low-level scheduling
    - DMA control, etc

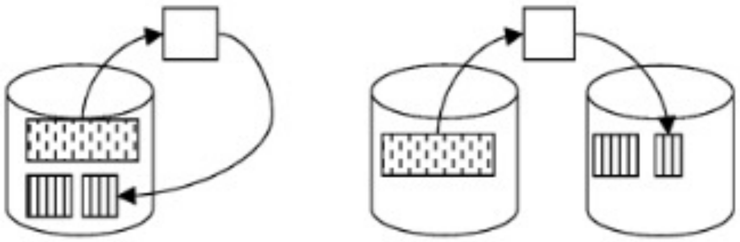
Storage Controller

- Interfaces with Flash packages
- Performs scrambling and ECC

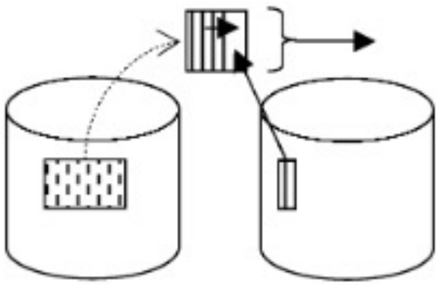
eXascale Infolab

Use Case: Sort and “Spilling” Operators

- External sort is the third most important IO pattern, after the transaction log and buffer manager flushes
- During an external sort, several *runs* are generated on disk
- The runs are later merged
- There are many optimization opportunities if the device is aware of the run-generation/merge pattern

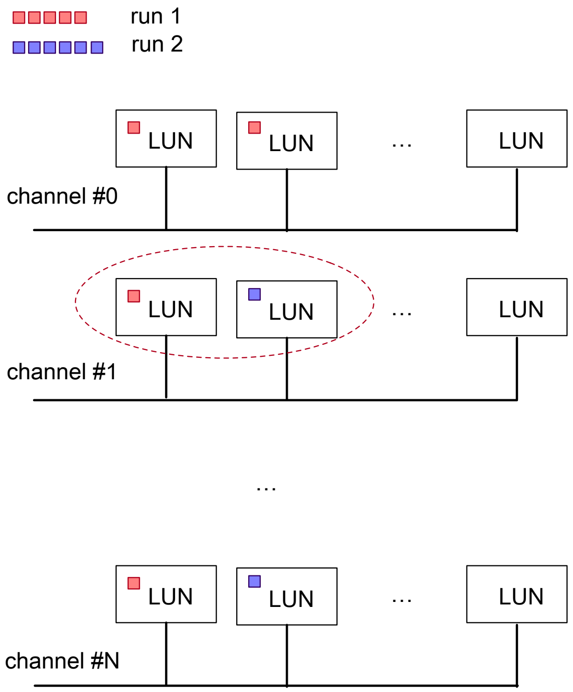


[Graefe’06]



The RUN Directive

- Observe the different IO optimized versions on an instrumented SSD[CIDR'20]
- We expect runs to be striped across LUNs
  - Good for writes!
  - But interference between reads during the merge phase
- Design an NVMe directive (w/ Philippe Bonnet)
  - Stripe I/O on range, trying to collocate runs that will be merged



Alberto Lerner, Jaewook Kwak, Sangjin Lee, Kibin Park, Yong Ho Song, and Philippe Cudré-Mauroux. "**It Takes Two: Instrumenting the Interaction between In-Memory Databases and Solid-State Drives**." In *CIDR 2020, 10th Conference on Innovative Data Systems Research*, 2020.

Research Agenda

How long has it taken GPUs to become mainstream in DBs?

Programming Model/Abstraction

- Main critique: current models are too low level or inexistent!
  - E.g., the unit of computation for a NIC/switch is a packet
  - In practice, however, a packet may have several inputs or be part of a larger input
- Proposing viable programming models depends on understanding the opportunities and limitations of each platform
  - Get more experience by offloading some selected computations “by hand”

New Execution Engines

Variants

- As with GPUs, algorithms need to accommodate different hardware variants
- A Database's hardware platform may evolve step-by-step by adding new accelerators

Scheduling

- (Re-)Loading query logic is not trivial in certain devices
- Competing queries may share a given device and should be isolated from each other

Conclusion

- An I/O event is a viable opportunity to offload applications' computations
- Because of a rare confluence of factors, database and hardware codesign is becoming increasingly accessible
- Current programming models and execution environments are inadequate;
    - In particular there is little work on unified programming models

Q&A

Thank you!

eXascale Infolab

References

- André Ryser, Alberto Lerner, Alex Forencich, Philippe Cudré-Mauroux. “**D-RDMA: Bringing Zero-Copy RDMA to Database Systems** .” Submitted.
- Theo Jepsen, Alberto Lerner, Fernando Pedone, Robert Soulé, and Philippe Cudré-Mauroux. “**In-Network Support for Transaction Triaging**.” In *Proceedings of the VLDB Endowment*, 14:1626–39, 2021.
- Alberto Lerner, and Philippe Bonnet. “**Not Your Grandpa’s SSD: The Era of Co-Designed Storage Devices**.” In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD ’21)*, 2021.
- Nadeen Gebara, Alberto Lerner, Mingran Yang, Minlan Yu, Paolo Costa, and Manya Ghobadi. “**Challenging the Stateless Quo of Programmable Switches**.” In *Proceedings of the Nineteenth ACM Workshop on Hot Topics in Networks, HotNets20*, 2020.
- Alberto Lerner, Rana Hussein, André Ryser, Sangjin Lee, and Philippe Cudré-Mauroux. “**Networking and Storage: The Next Computing Elements in Exascale Systems?**.” *IEEE Data Engineering Bulletin* 43, no. 1 (March 2020): 60–71.
- Alberto Lerner, Jaewook Kwak, Sangjin Lee, Kibin Park, Yong Ho Song, and Philippe Cudré-Mauroux. “**It Takes Two: Instrumenting the Interaction between In-Memory Databases and Solid-State Drives**.” In *CIDR 2020, 10th Conference on Innovative Data Systems Research*, 2020.
- Alberto Lerner, Rana Hussein, and Philippe Cudré-Mauroux. “**The Case For Network Accelerated Query Processing**.” In *CIDR 2019, 9th Biennial Conference on Innovative Data Systems Research*, 2019.

- Goetz Graefe, “**Implementing sorting in database systems.”** In *ACM Computing Surveys, 38(3), 2006.*
- Ming Liu, Tianyi Cui, Henry Schuh, Arvind Krishnamurthy, Simon Peter and Karan Gupta. “**Offloading Distributed Applications onto SmartNICs Using Ipipe**.” In *Proceedings of the ACM Special Interest Group on Data Communication (SIGCOMM’19), 2019.*