Zero-CPU Collection with Direct Telemetry Access

Jonatan Langlet
Queen Mary University of London

Ran Ben-Basat
University College London

Sivaramakrishnan Ramanathan
University of Southern California

Gabriele Oliaro
Harvard University

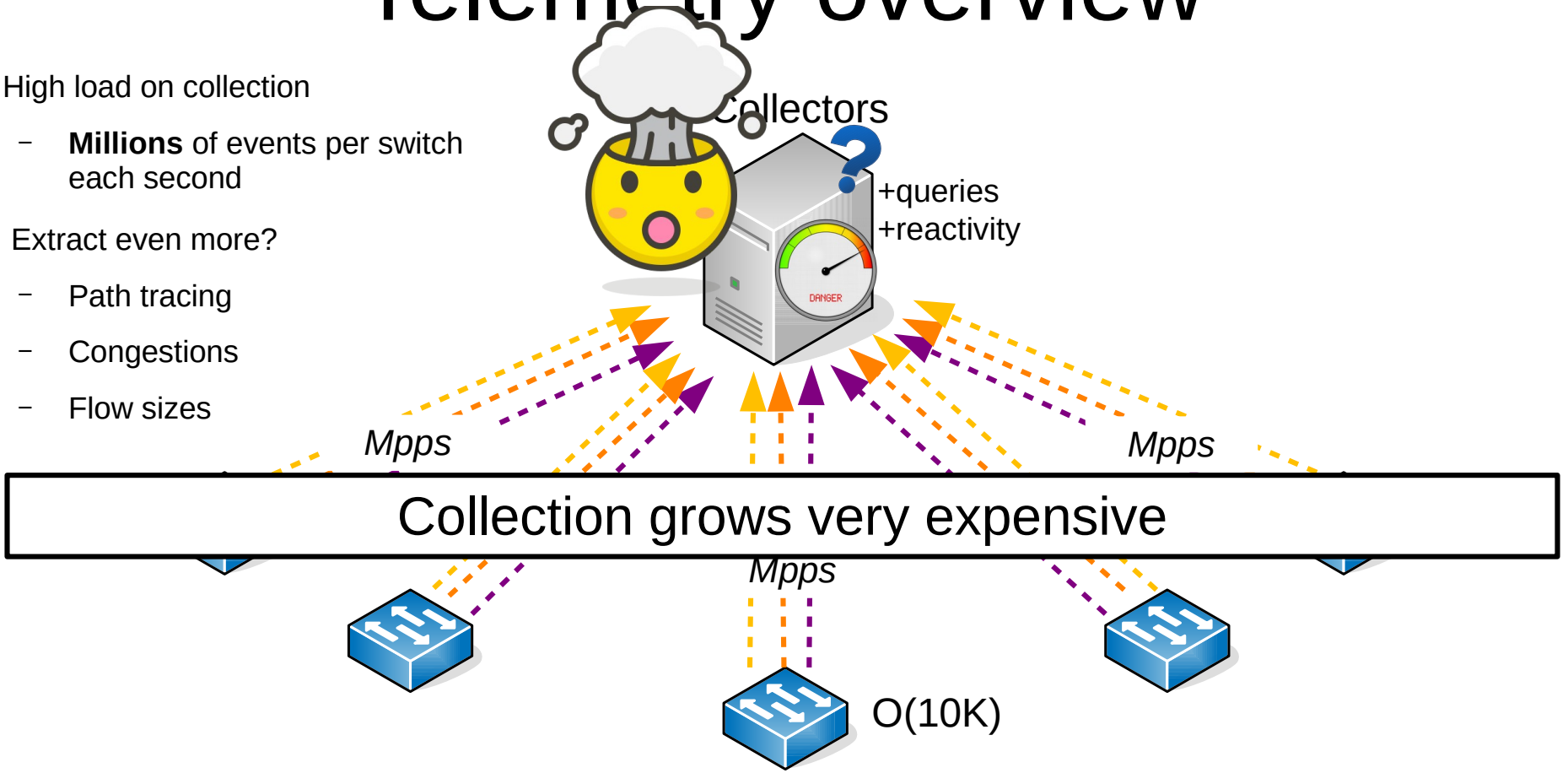
Michael Mitzenmacher
Harvard University

Minlan Yu
Harvard University

Gianni Antichi
Queen Mary University of London

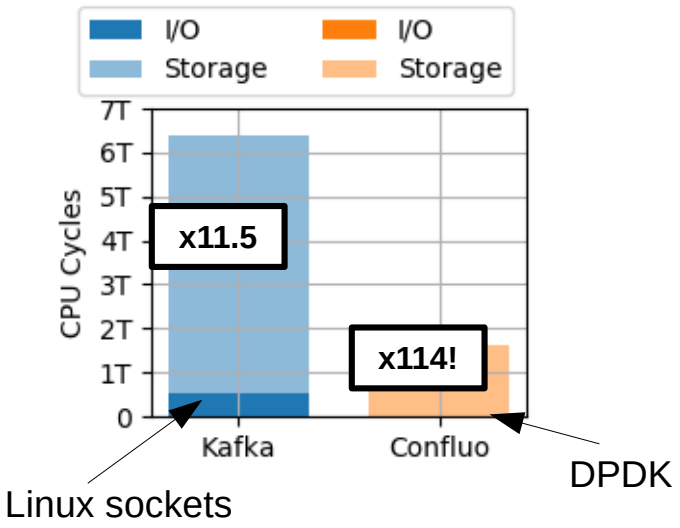
Telemetry overview

- High load on collection
  - **Millions** of events per switch each second
- Extract even more?
  - Path tracing
  - Congestions
  - Flow sizes



Zhou, Yu, et al., "Flow event telemetry on programmable data plane", SIGCOMM'20

CPU-based Collection

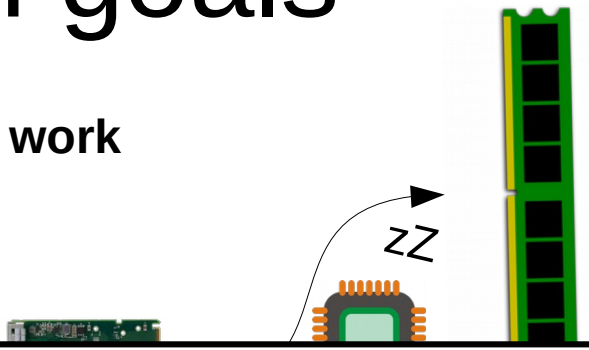


- Benchmarked two collectors
  - Broken into **I/O** and **Storage** cycles
- Packet I/O is not the issue
- The CPU really struggles with telemetry storage

Khandelwal, Anurag, et al., "Confluo: Distributed monitoring and diagnosis stack for high-speed networks", NSDI'19

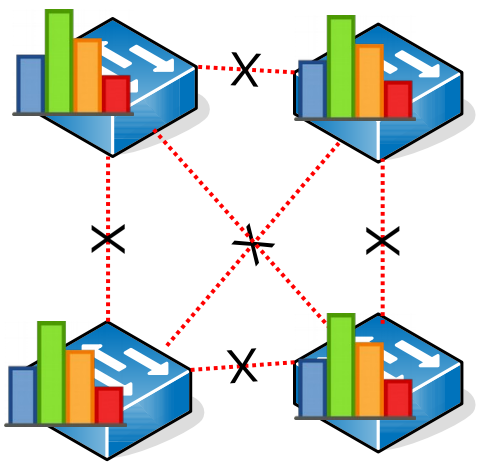
Design goals

- **We don’t want the CPU to do storage work**
  - No centralized data organization
  - No centralized collision handling

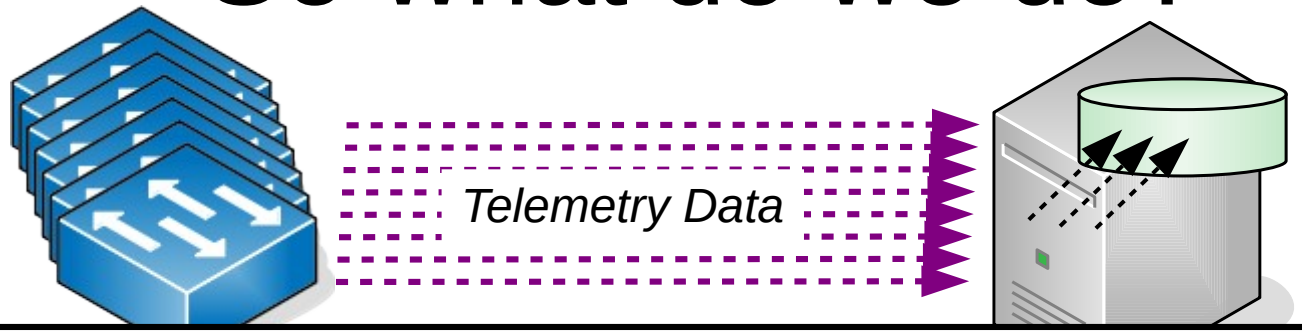


The main point is to **allow efficient centralized queryability**

- We want it **scalable**, with **low overheads**
  - **Very** low on-switch statefulness.
  - Minimal inter-switch collaboration
  - Fully in the data plane



So what do we do?

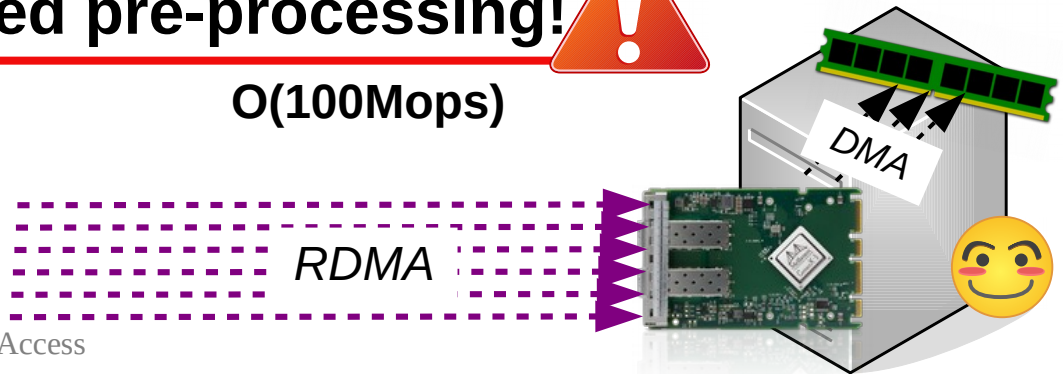


A packet specifies a location for a value
That is exactly where the value will be written

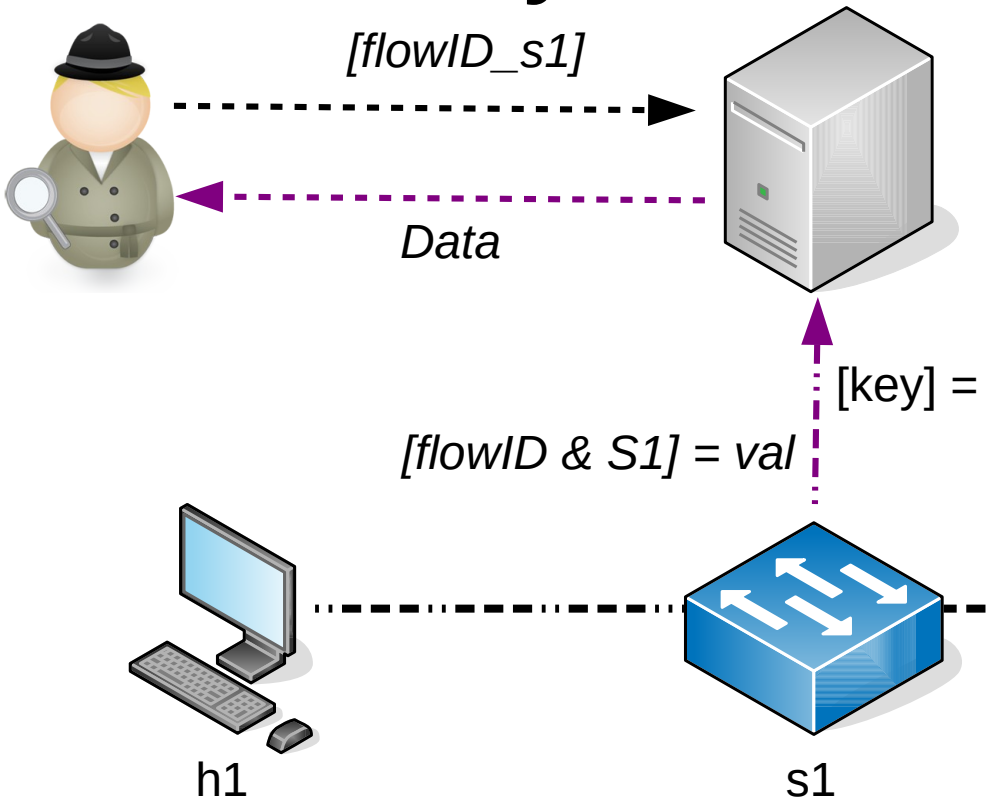
This is what **RDMA** is designed for!

No centralized pre-processing!

O(100Mops)



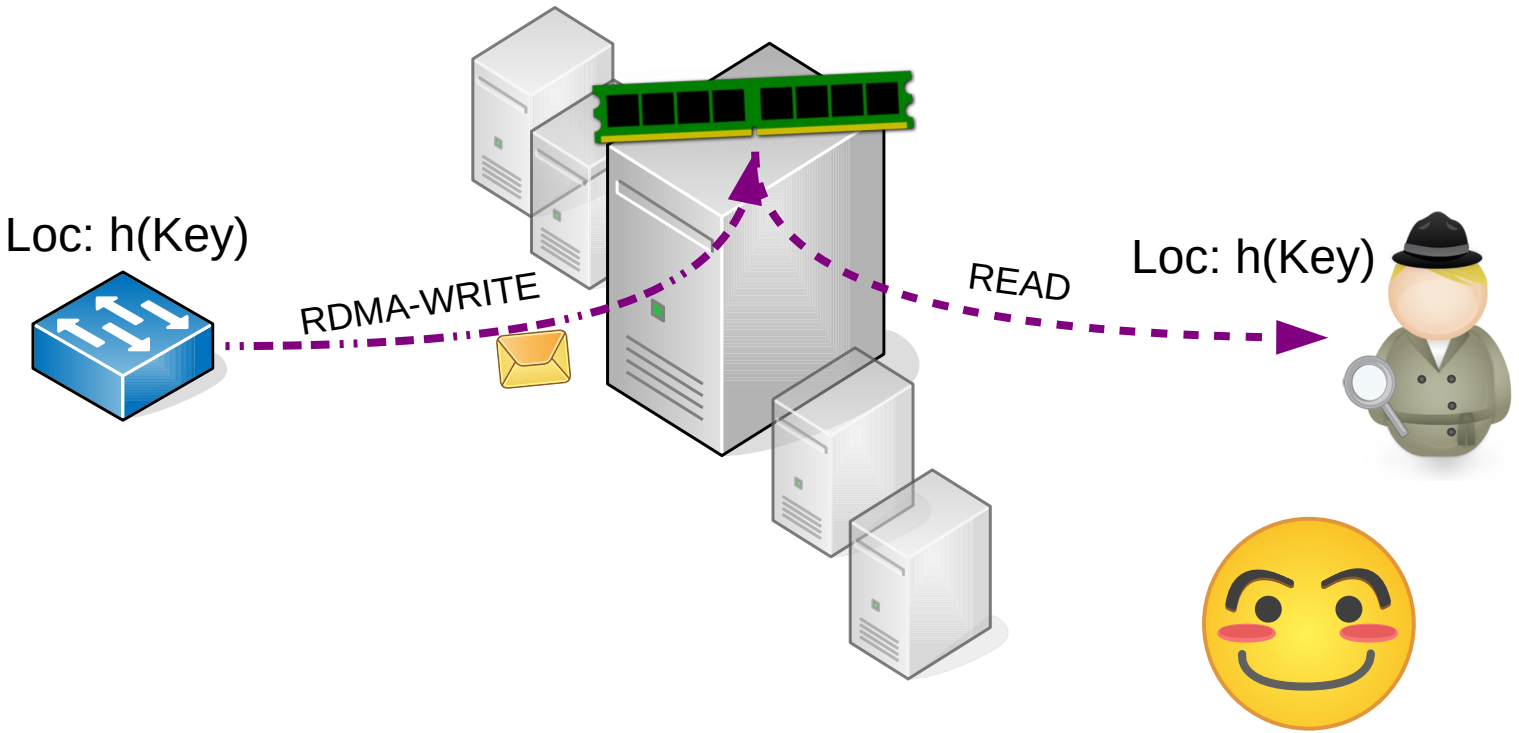
Key-value design



- E.g., Per-flow per-switch
  - Path tracing
  - INT postcarding

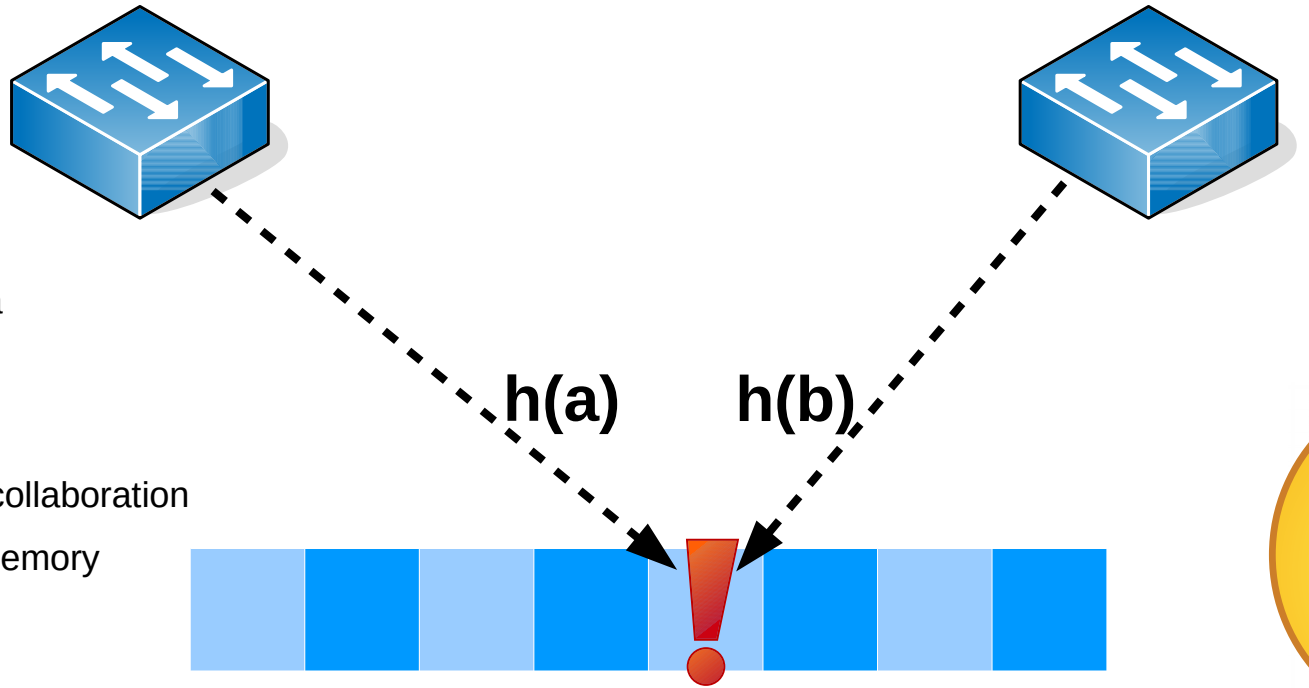
All telemetry data has a unique identifier: its **Key**

Global hash functions

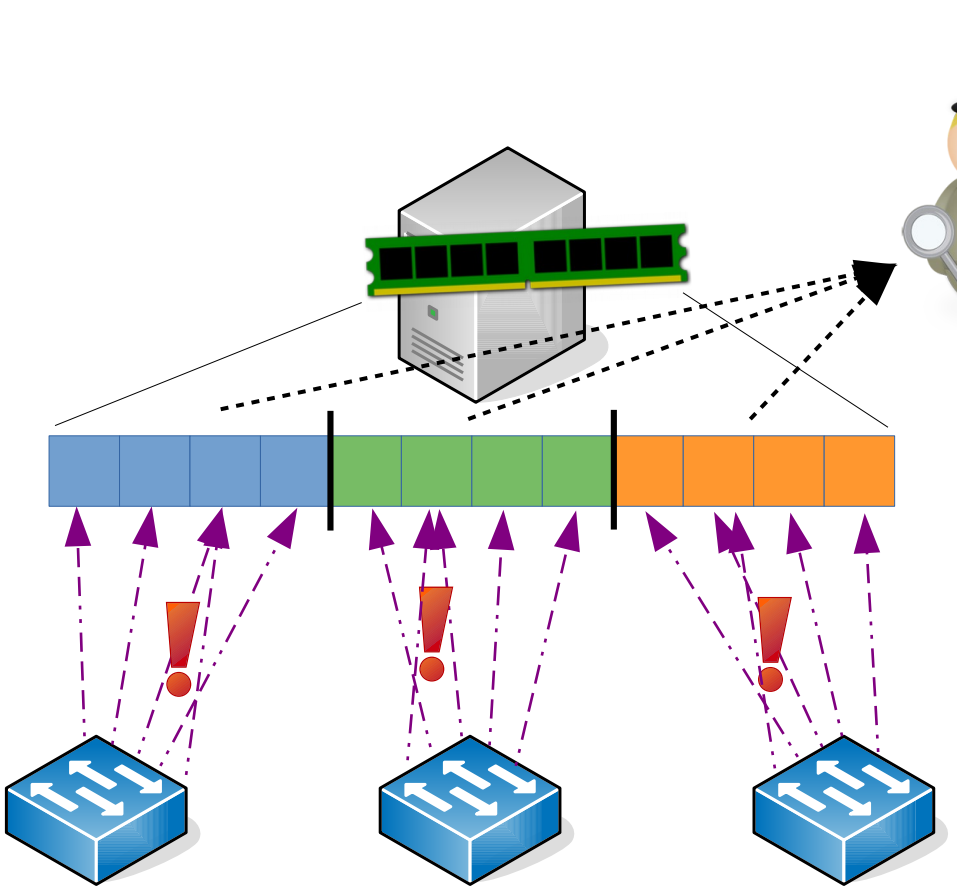


There is a problem!

- Hashes can collide!
  - Overwritten data
- But we want:
  - Near-stateless
  - No inter-switch collaboration
  - Shared global memory

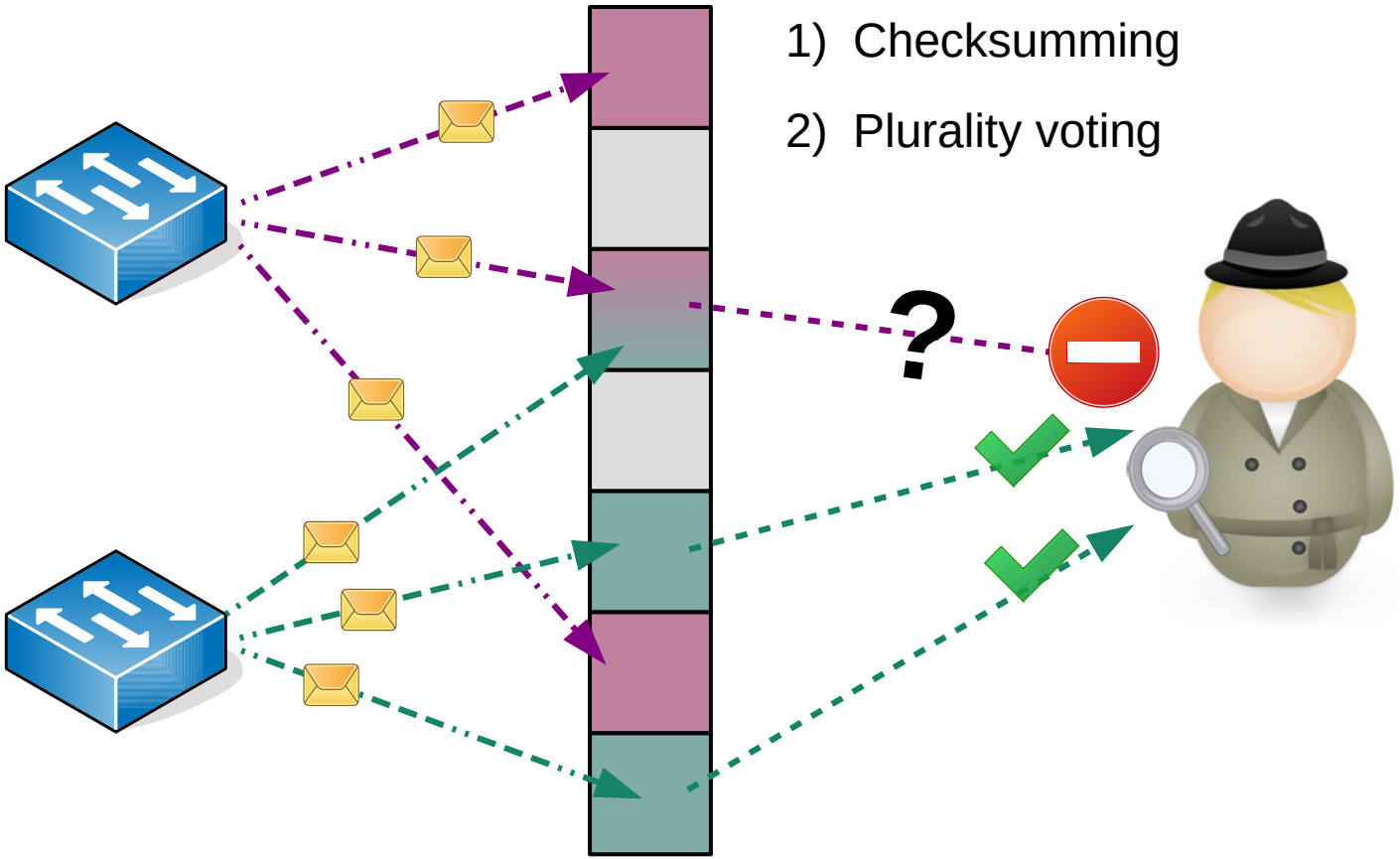


Dedicated memory?

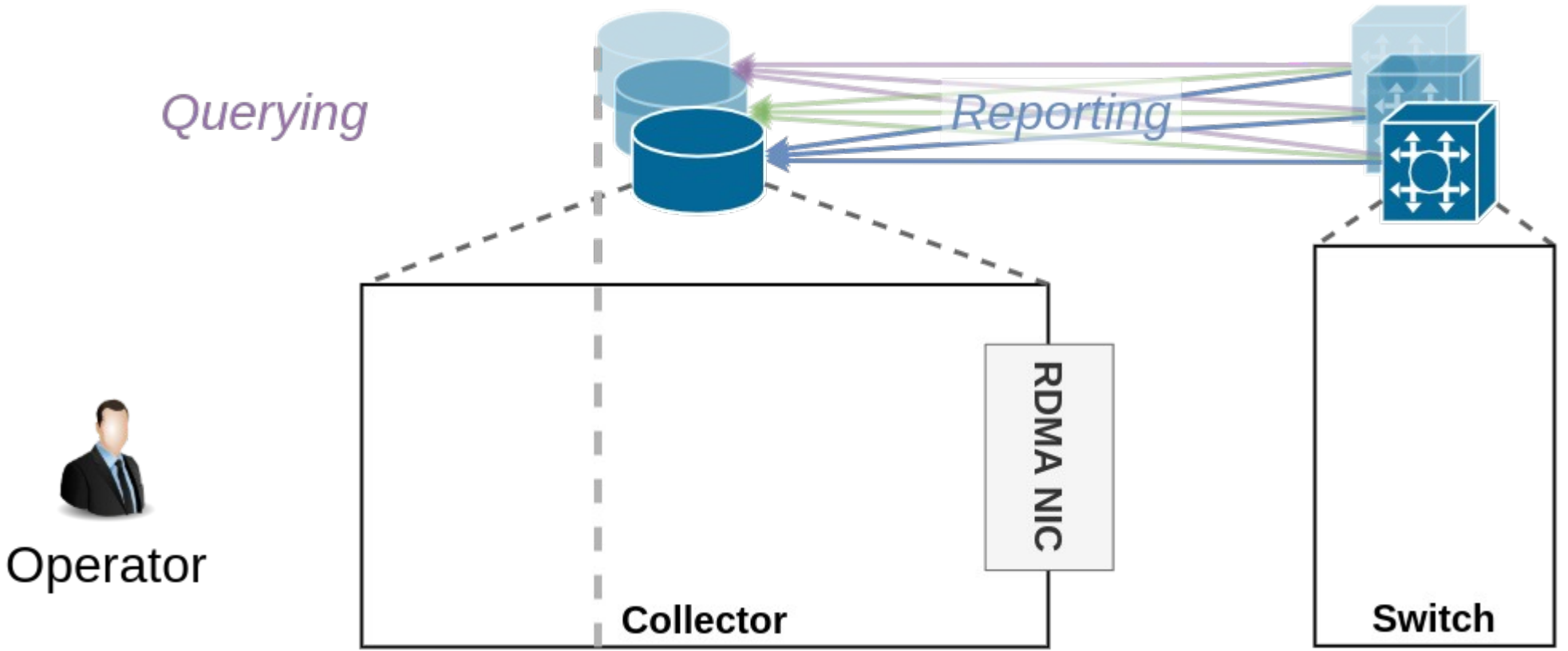


- Key collisions within each switch?
- Stateful solutions not feasible
  - Amount of memory
  - Increased complexity
- **Efficient queryability**
  - Which switch?
  - Where?

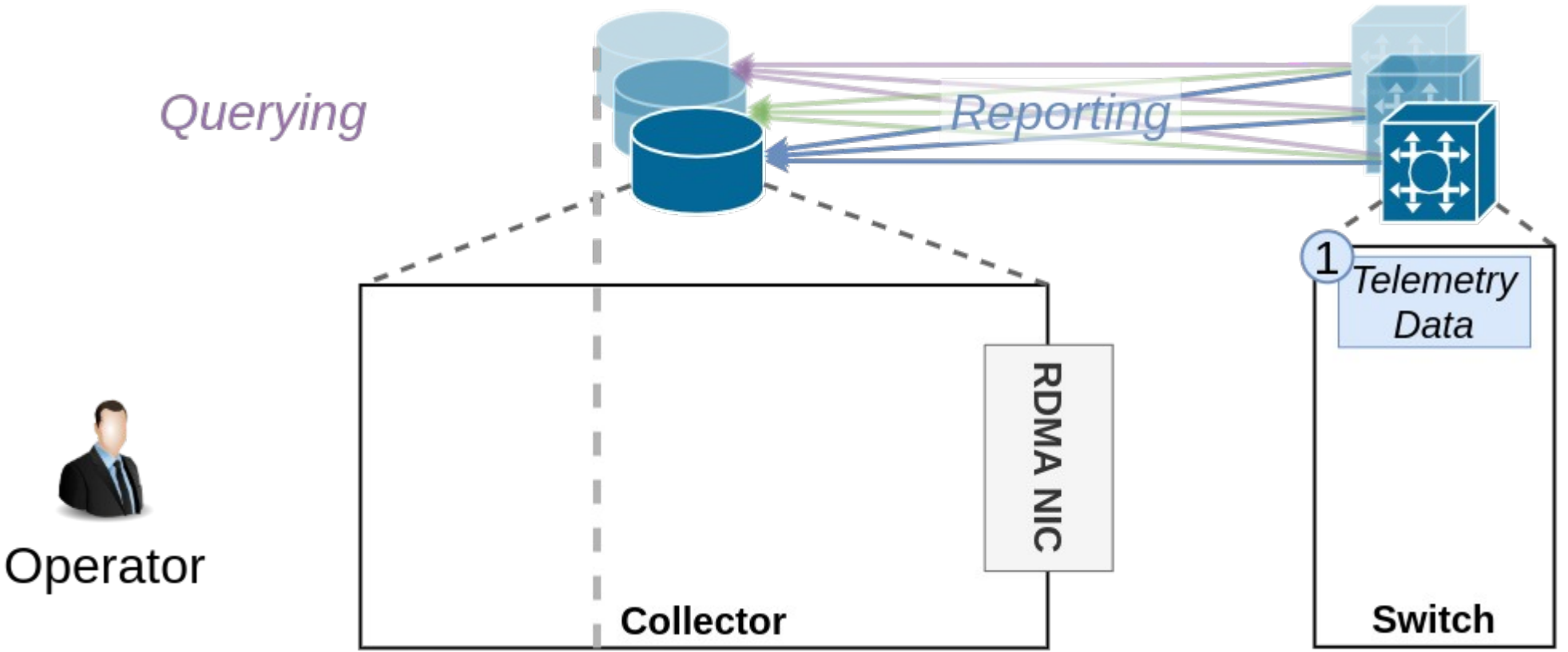
Built-in redundancies



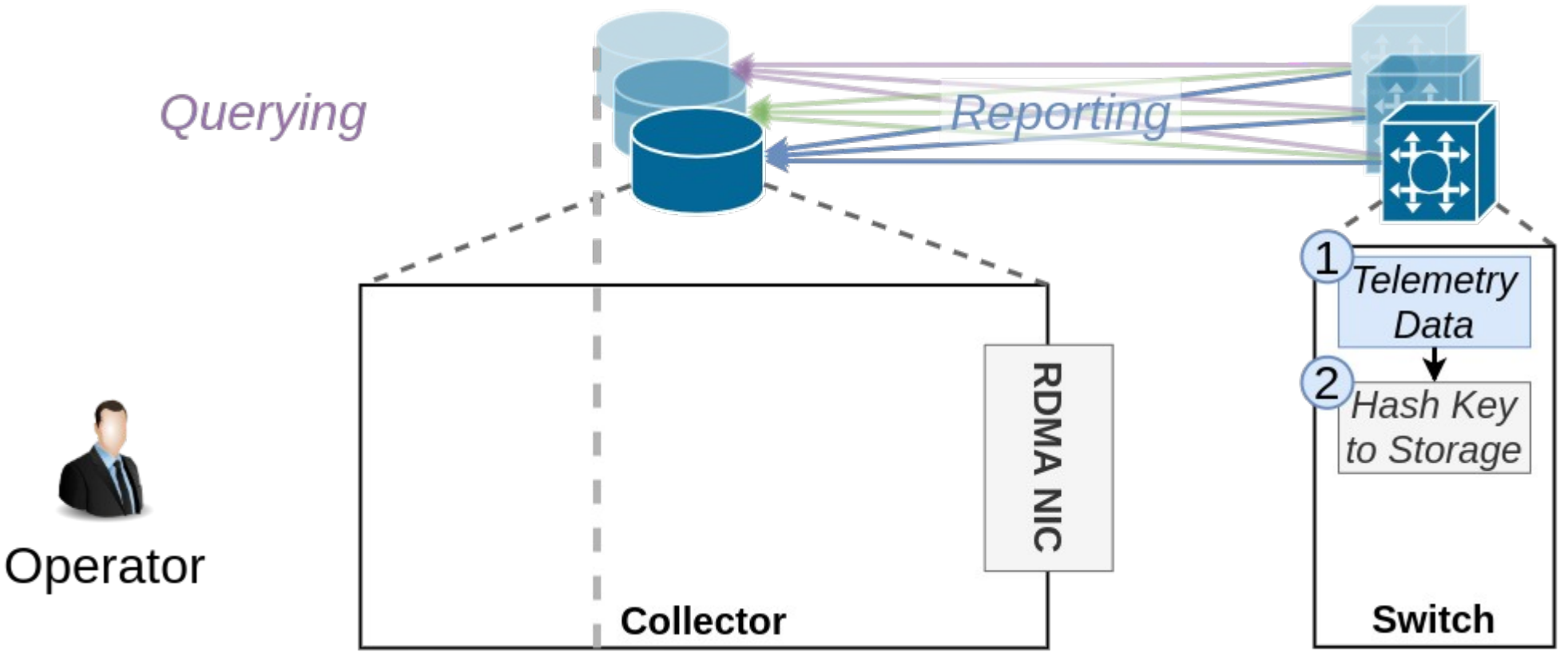
Overview



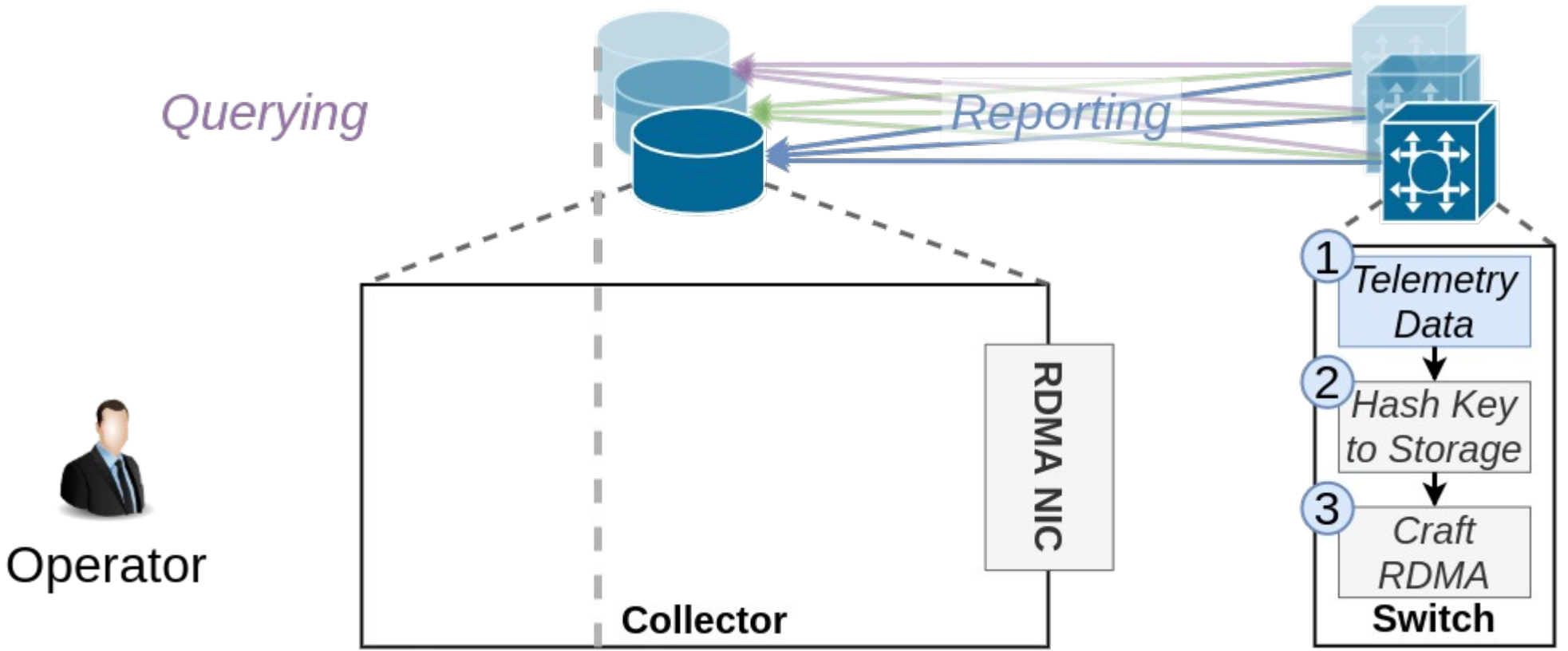
Overview


Querying
Reporting
1
Telemetry Data
RDMA NIC
Collector
Switch
Operator

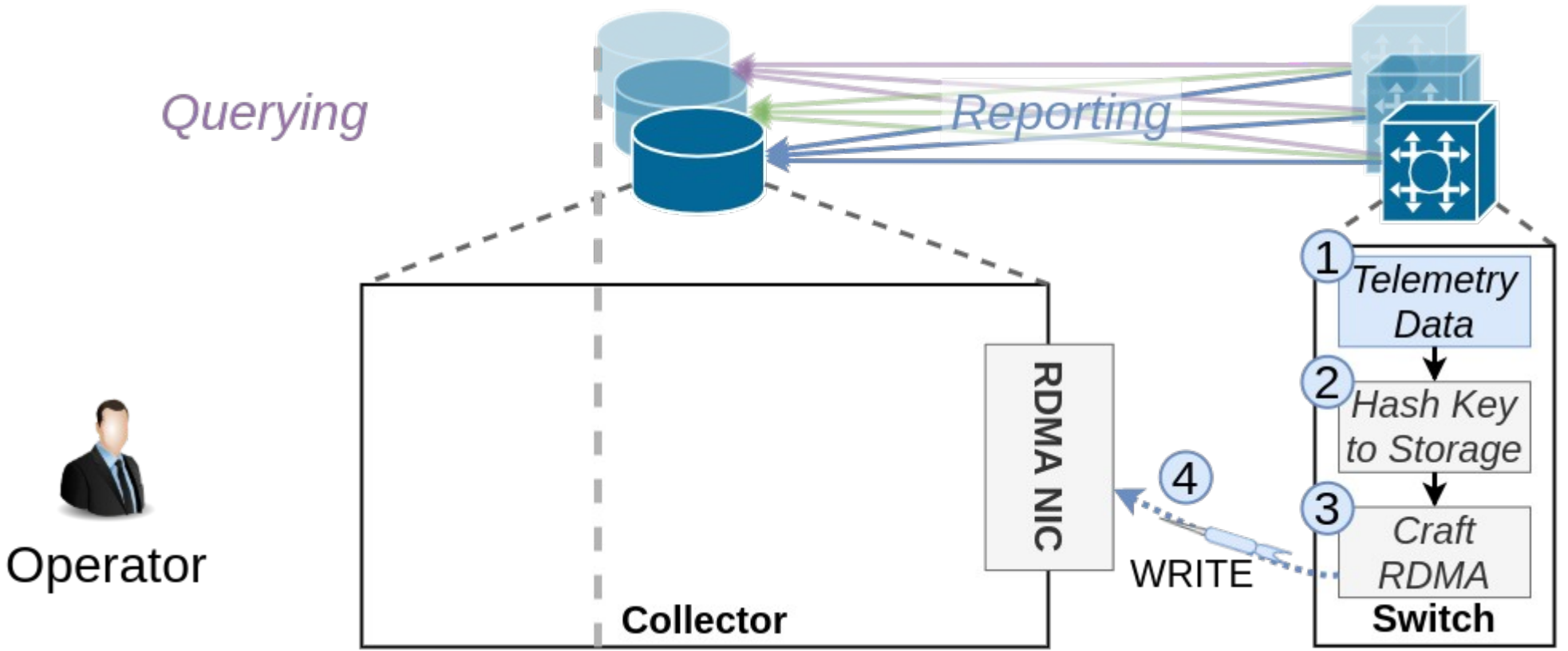
Overview


Querying
Reporting
Operator
RDMA NIC
Collector
1
Telemetry Data
2
Hash Key to Storage
Switch

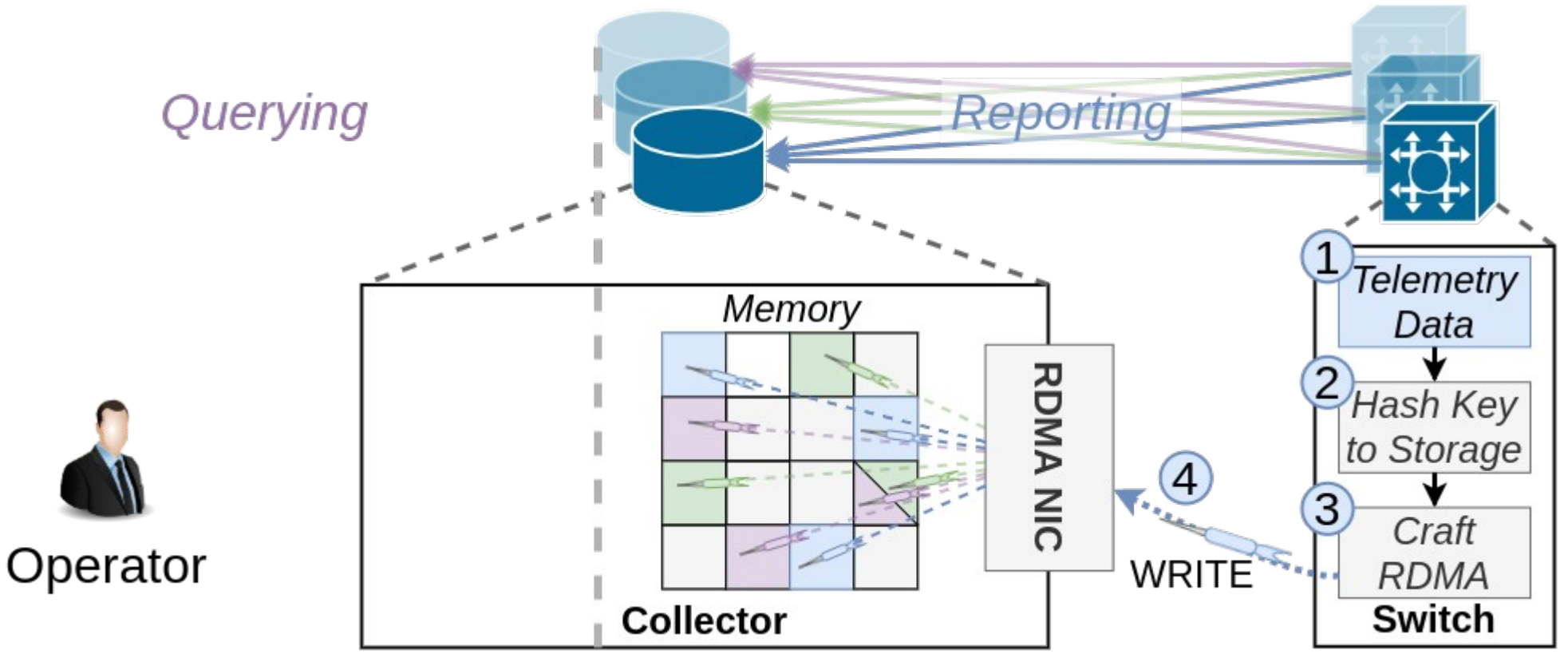
Overview


Querying
Reporting
Operator
Collector
RDMA NIC
1
Telemetry Data
2
Hash Key to Storage
3
Craft RDMA
Switch

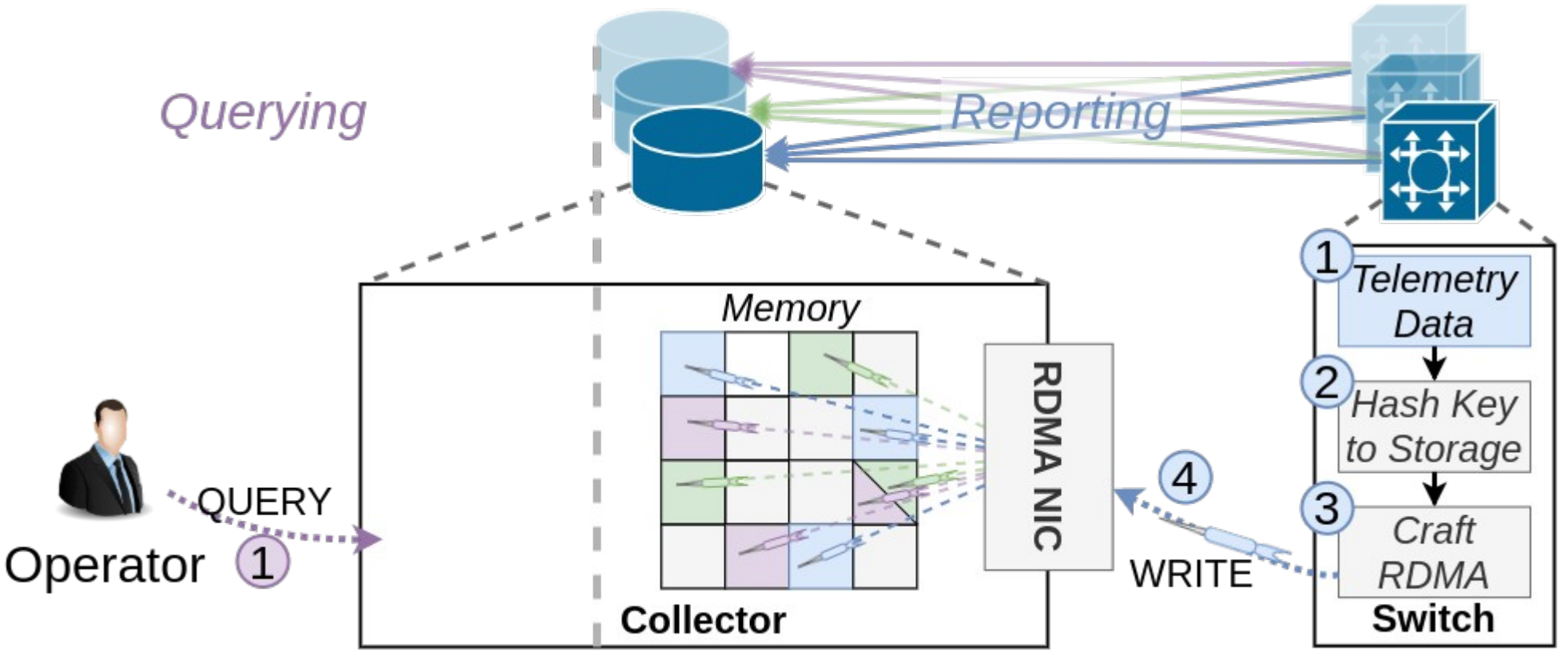
Overview


Querying
Reporting
Operator
Collector
RDMA NIC
WRITE
1
Telemetry Data
2
Hash Key to Storage
3
Craft RDMA
Switch
4

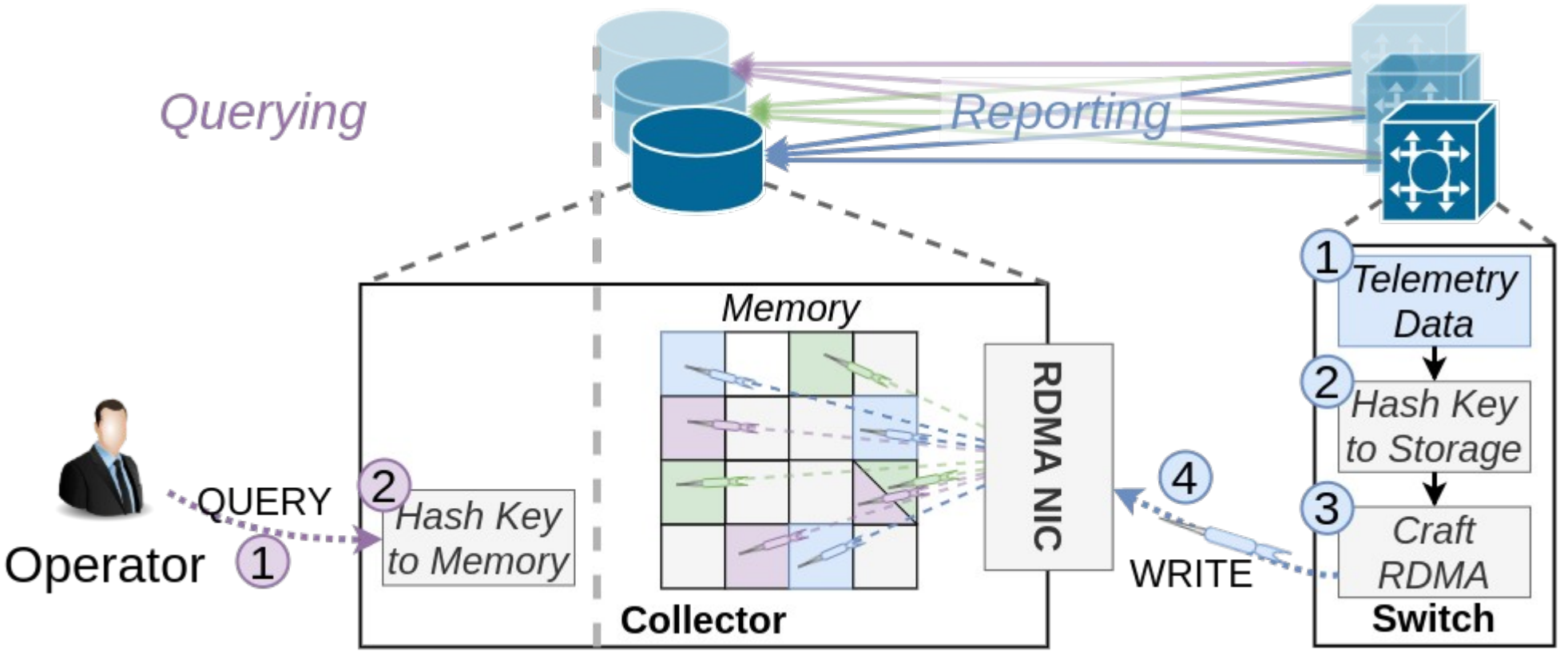
Overview


Querying
Reporting
Operator
Memory
RDMA NIC
Collector
1
Telemetry Data
2
Hash Key to Storage
3
Craft RDMA
Switch
4
WRITE

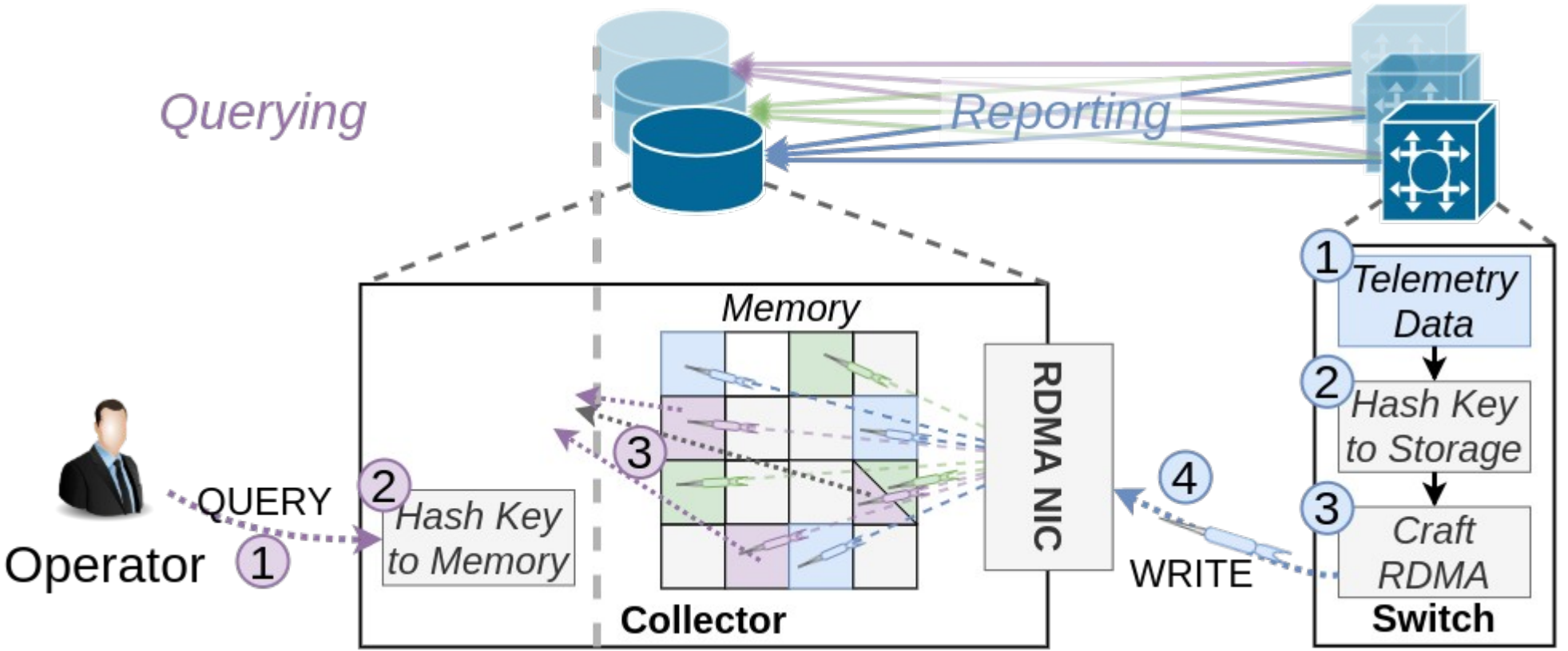
Overview


Querying
Reporting
Memory
RDMA NIC
Collector
Operator
QUERY
1
4
WRITE
1
Telemetry Data
2
Hash Key to Storage
3
Craft RDMA
Switch

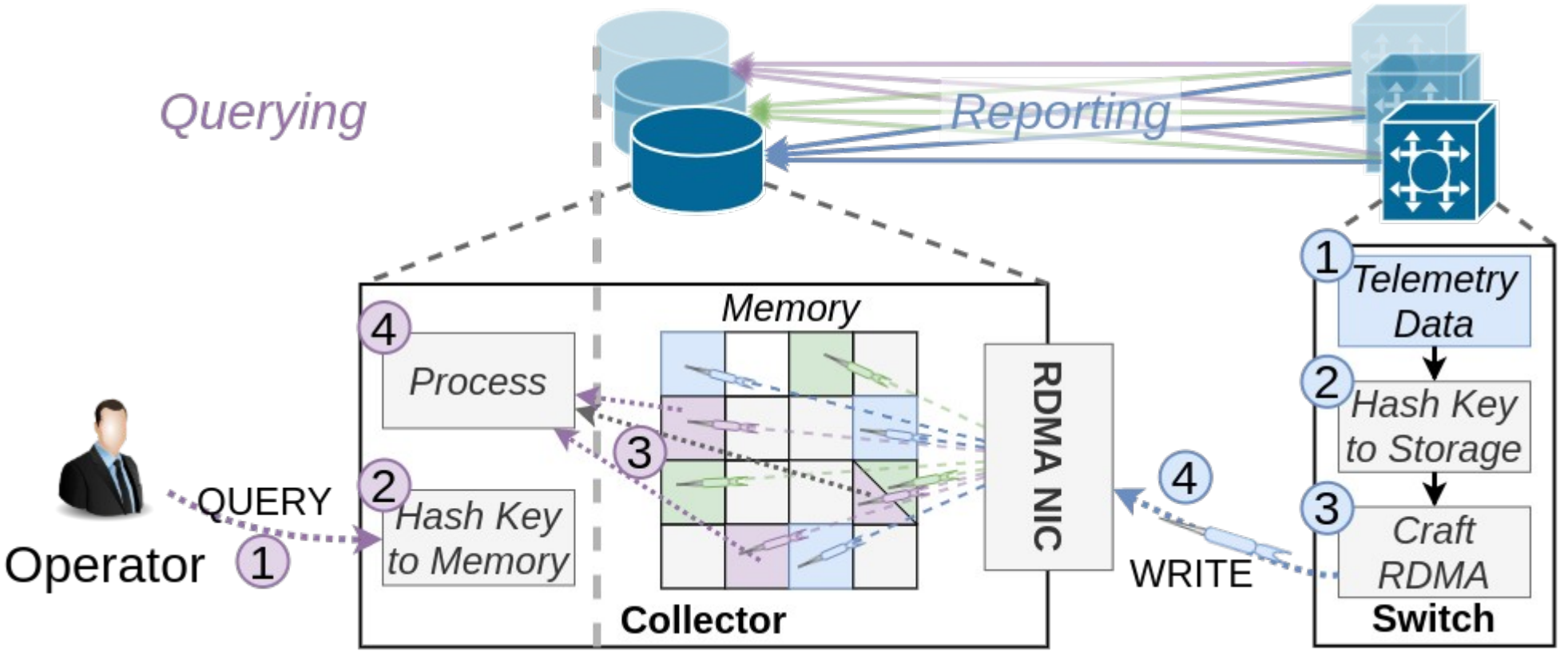
Overview


Querying
Reporting
Memory
RDMA NIC
Collector
Switch
Operator
QUERY
1
2
Hash Key
to Memory
1
Telemetry
Data
2
Hash Key
to Storage
3
Craft
RDMA
4
WRITE

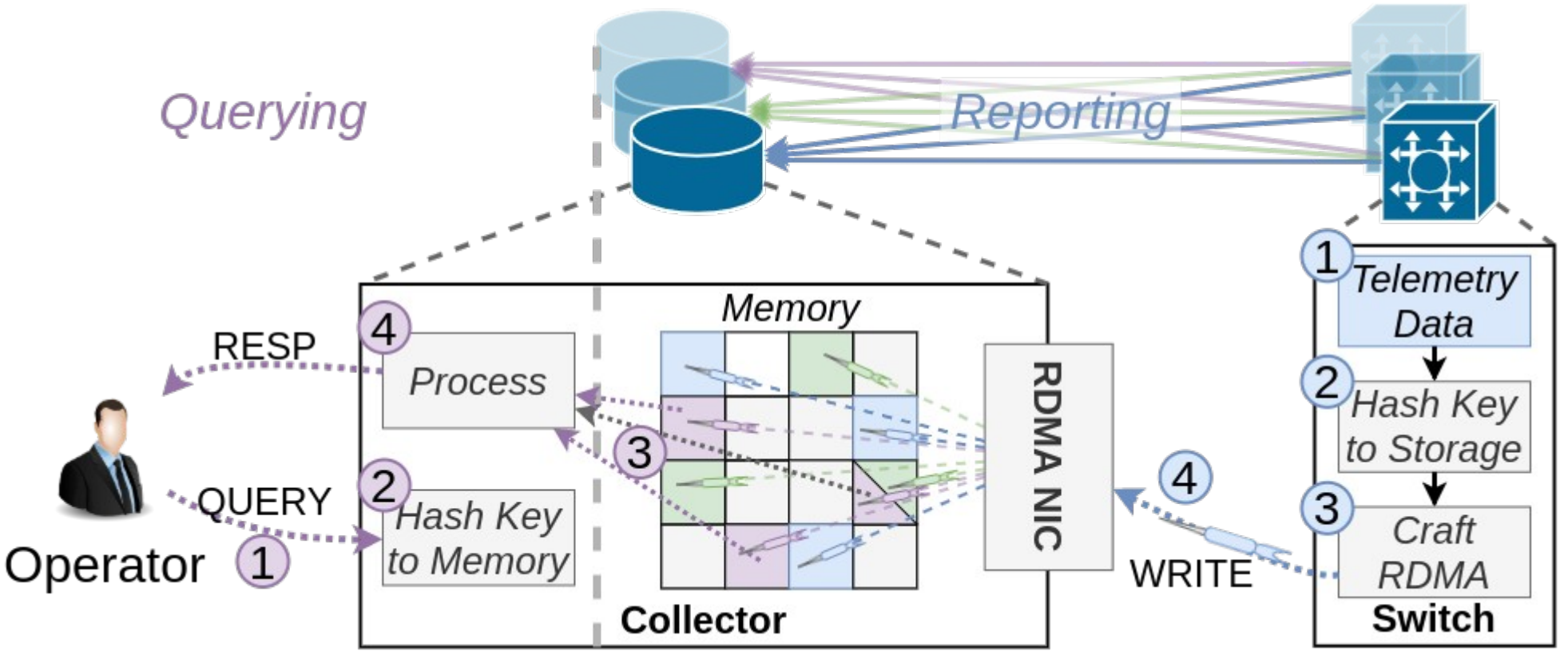
Overview


Querying
Reporting
Memory
RDMA NIC
Collector
Switch
Operator
QUERY
1
2
Hash Key
to Memory
3
4
WRITE
1
Telemetry
Data
2
Hash Key
to Storage
3
Craft
RDMA

Overview

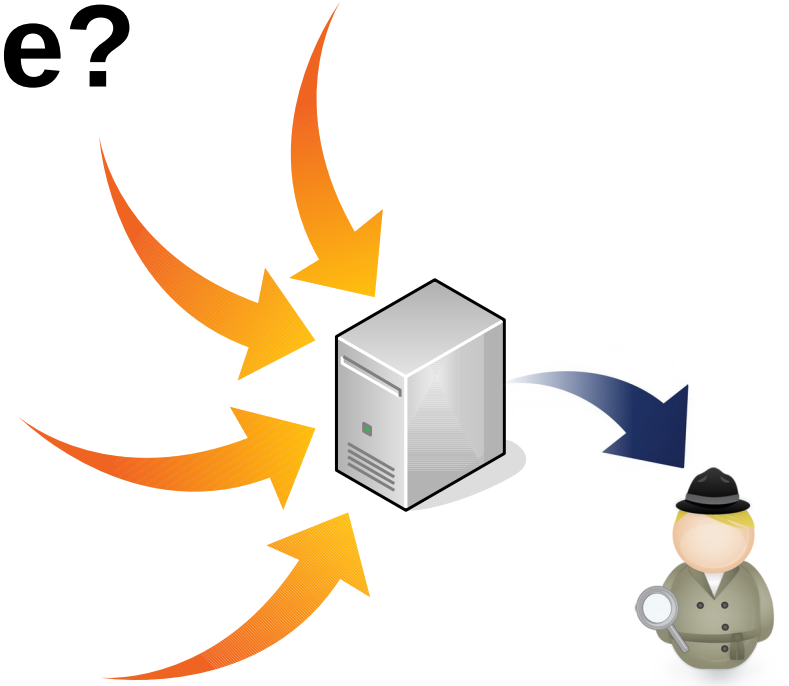

Querying
Reporting
Memory
4
Process
2
Hash Key
to Memory
3
QUERY
1
Operator
Collector
RDMA NIC
4
WRITE
1
Telemetry
Data
2
Hash Key
to Storage
3
Craft
RDMA
Switch

Overview


Querying
Reporting
Memory
RESP
4
Process
3
2
Hash Key
to Memory
QUERY
1
Operator
RDMA NIC
4
WRITE
1
Telemetry
Data
2
Hash Key
to Storage
3
Craft
RDMA
Switch
Collector

Future?

- **Generalize** beyond key-value queries
  - E.g., aggregated network states
- How to **query the unknown**?
  - Iterate over data
- **Immediate** controller reactivity?
  - Real-time controller response to events
- **Tailored RDMA** for telemetry?
  - New primitives
    - E.g., multi-write for redundancy
  - Remove the need for a super-reliable network



Vision: a transport protocol for telemetry data