STORAGE DEVELOPER CONFERENCE

SDC 21

BY Developers FOR Developers

Virtual Conference
September 28-29, 2021

InfiniBand/RoCE RDMA Specification Update

Overview of August 2021 updates to the IB Specification, including the new Memory Placement Extensions

Chet Douglas, Intel

Specification update overview

- Volume 1, Release 1.5, published August 6, 2021
- The specification defines InfiniBand and RoCE
- Available to IBTA Members

- 2038 pages
- 22 issues addressed
- 57 new sections/features added

- InfiniBand NDR speeds
- QoS and bandwidth enhancements
- Virtualization section updated
- **Memory Placement Extensions**

INFINIBAND℠
TRADE ASSOCIATION

Support Enhanced Speeds

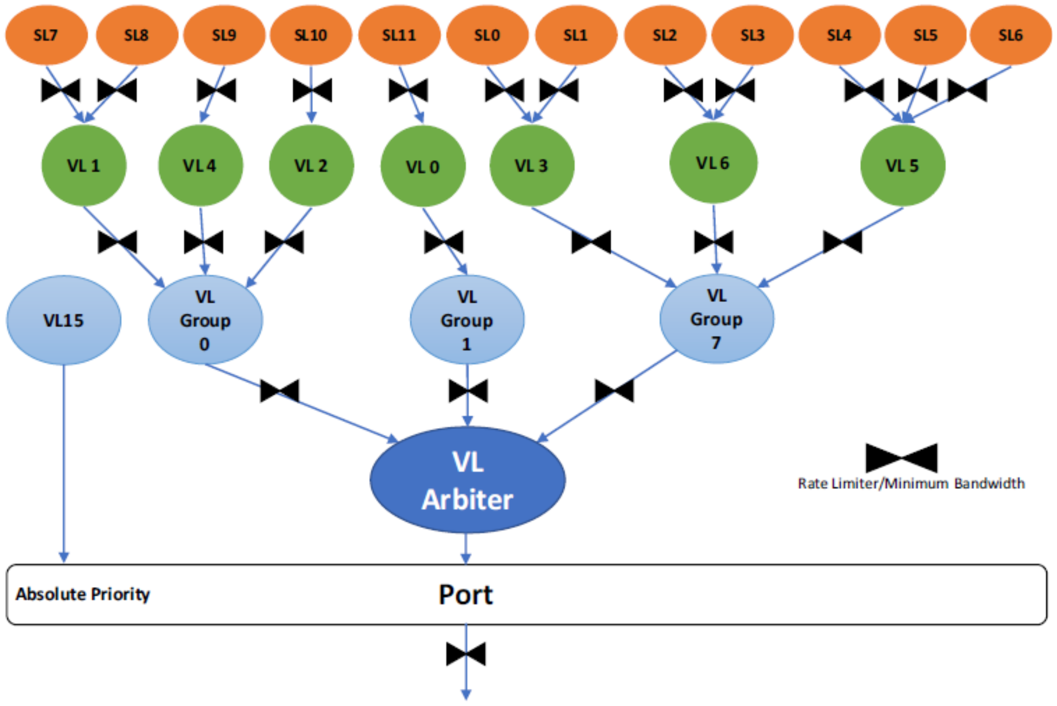
- 1.5 spec includes support for InfiniBand NDR speeds
- Including split support (2x)

Link Speed	Lane Speed	Signaling	2X throughput	4X throughput	8X throughput
EDR	25.78125 Gb/s	NRZ	51.5625 Gb/s	103.125 Gb/s	206.25 Gb/s
HDR	53.125 Gb/s	PAM4	106.25 Gb/s	212.5 Gb/s	425 Gb/s
NDR	106.25 Gb/s	PAM4	212.5 Gb/s	425 Gb/s	850 Gb/s

- RoCE follows the Ethernet Physical & Link Layer Standards
  - RoCE fully supports 400 Gb/s and future speeds

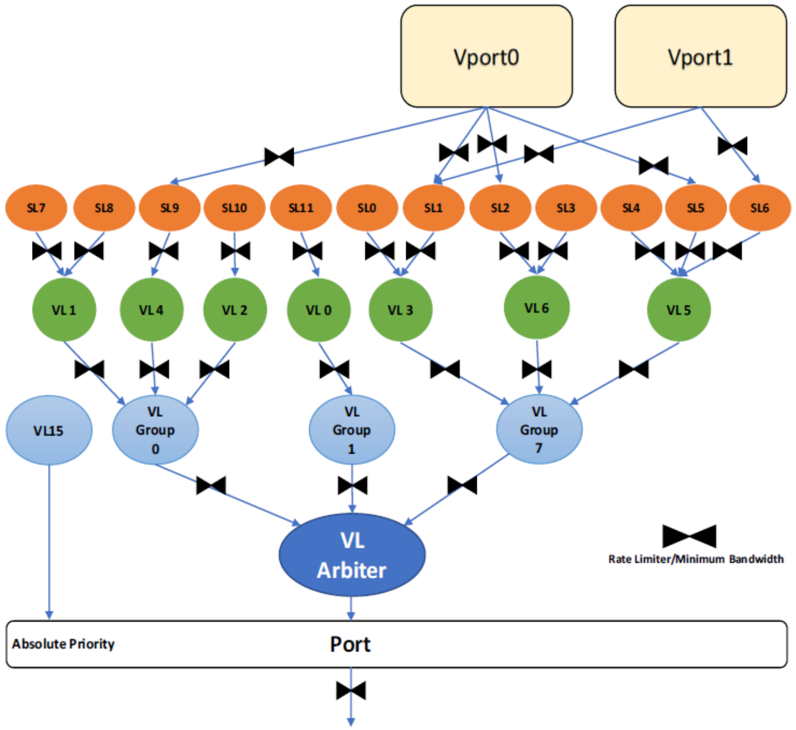
Minimum Bandwidth Added to Enhanced Port Arbiter

- New minimum bandwidth setting capabilities
- Allow setting a dynamic bandwidth increase
- Use cases:
  - Traffic that requires low latency and low jitter.
    - Management
    - Time sync
    - Heartbeat protocol



Virtual Ports QoS

- Introduce bandwidth sharing and rate limiting configuratior for {Vport,SL} on top of Standard or Enhanced InfiniBand port arbiter



Enhanced Port Arbiter

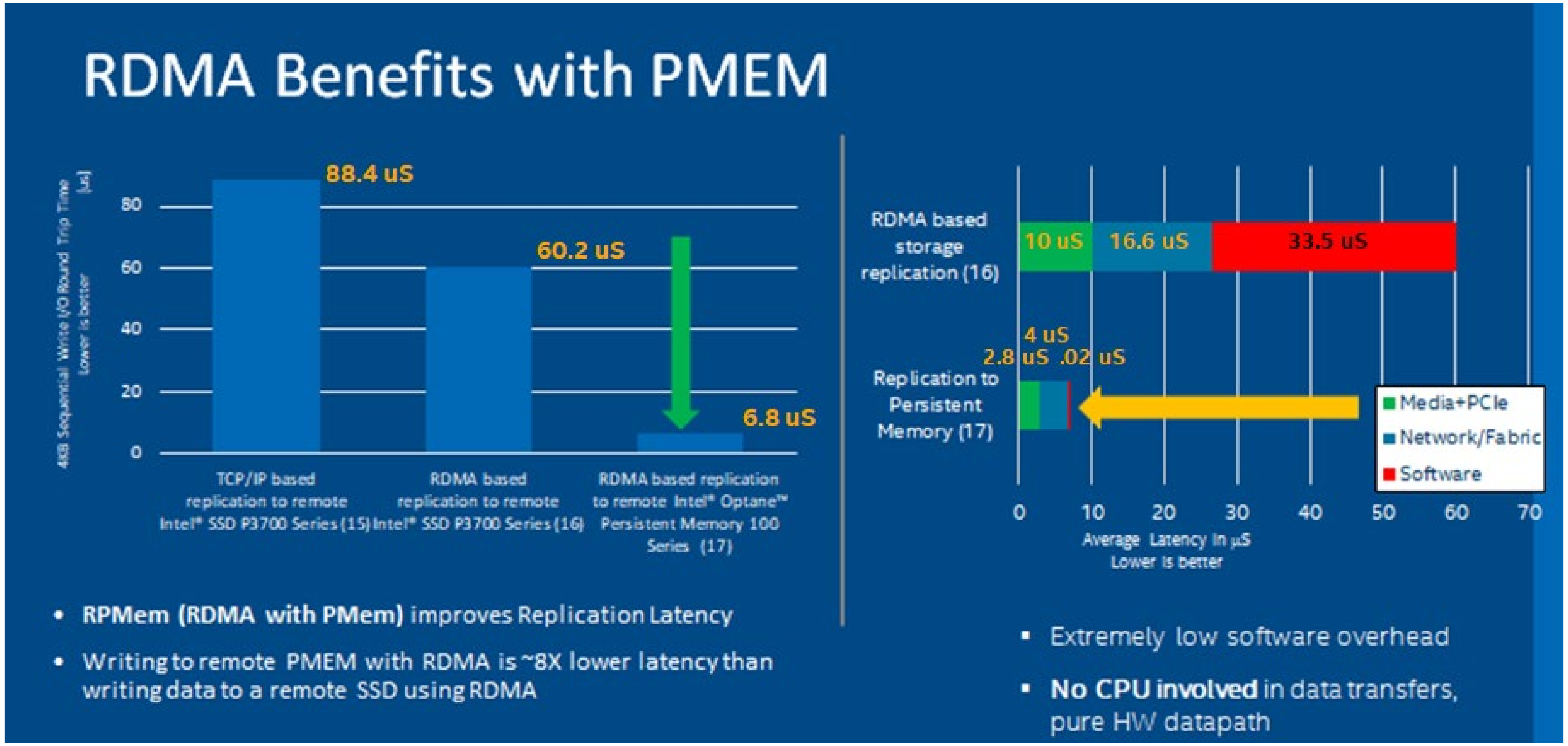
Memory Placement Extensions

"MPE"

STORAGE DEVELOPER CONFERENCE SDC 21

Memory Placement Extensions

- Motivation
  - Performance demonstration compared round trip latency for RDMA to an SSD vs RDMA to PMEM
  - Because the PMEM is byte addressable and is attached to the memory subsystem it is possible to transfer data to the final persistence domain via RDMA, with minimal CPU involvement



Memory Placement Extensions

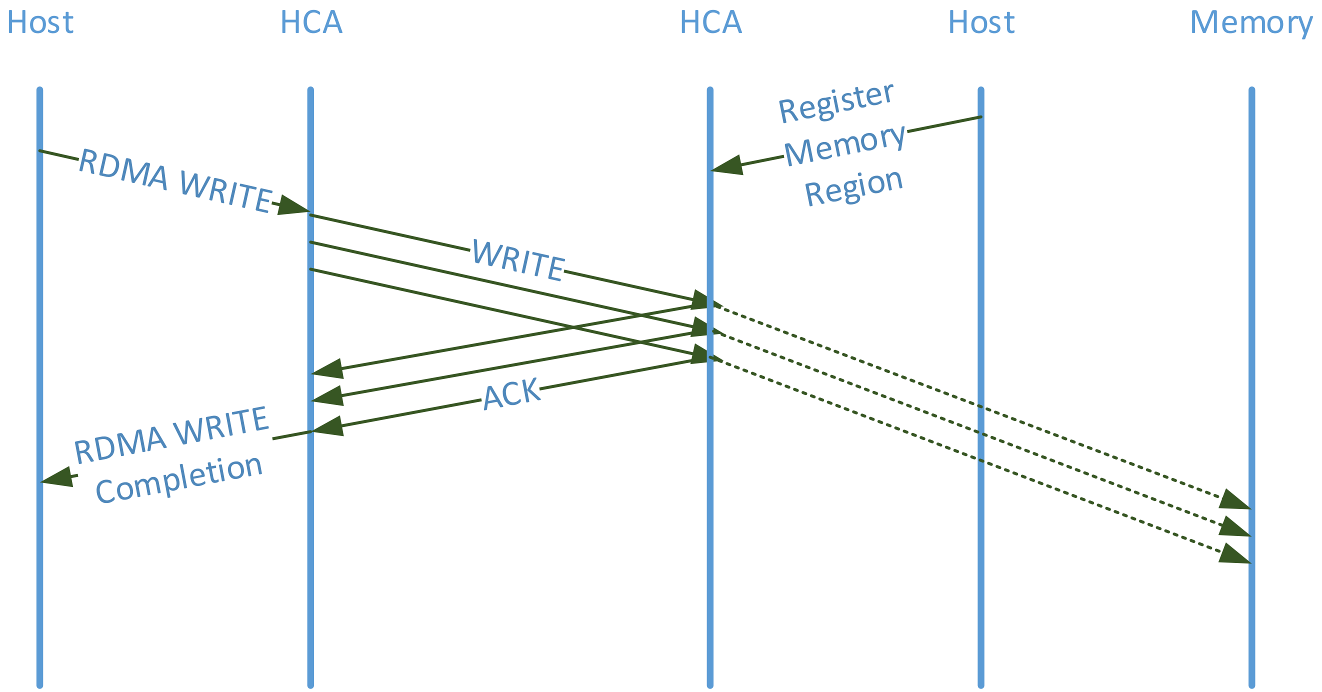
Motivation/problem statement

- With the advent of PMEM, storage is now connected to the memory subsystem and can be directly accessed using RDMA
- However, existing RDMA did not provide RDMA Write or Atomics reliability guarantees
  - RDMA Write completion does not guarantee data has reached remote host memory
- With pre-1.5 specification, any guarantees that the data has reached the persistence domain must be implemented by a ULP. This requires interrupts, pipeline stalls, adding additional latency to the transaction, and may not be scalable.

Memory Placement Extensions

Motivation/problem statement details

- RDMA Acknowledge (and Completion)
  - Guarantee only that Data has been successfully received and accepted for execution by the remote HCA
  - Doesn't guarantee data has reached remote host memory
- Further guarantees are out of the scope of the standard
- There are platform specific ways to implement further guarantees
  - E.g. messaging + SW cache flushing
  - E.g. RDMA READ



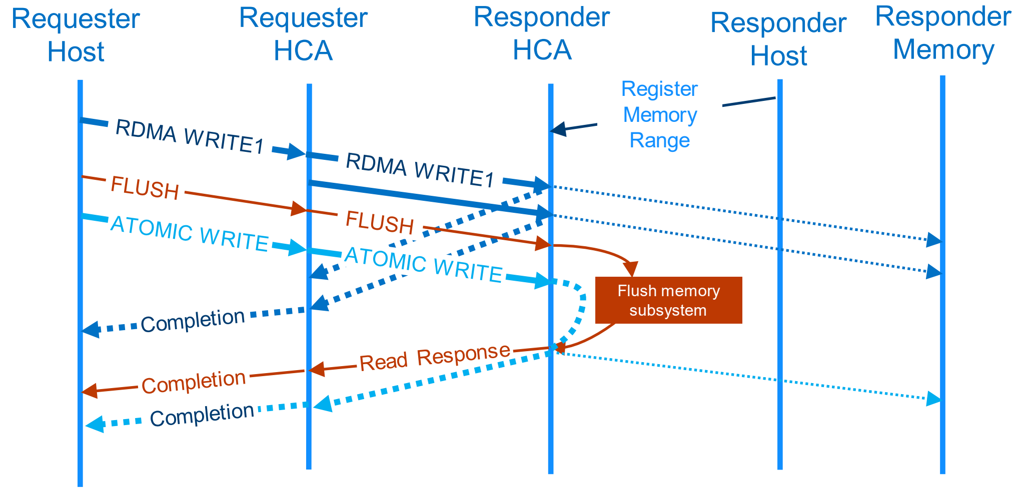
Memory Placement Extensions

- Two new command opcodes in the 1.5 specification
  - Flush
    - Flush all previous writes or specific regions, per QP
    - Provides acknowledgement that volatile writes have made it to Global Observability
    - Provides acknowledgement that persistent memory writes have made it to the power fail safe persistence domain
    - Pipelined operation
  - Atomic Write
    - Writes an aligned 8-byte value atomically
    - Provides guarantees for remote pointer updates to persistent or volatile memory
- Fully supported by InfiniBand and RoCE

Memory Placement Extensions

Synchronous Log Writing with MPE

- Synchronous log writing: Requester Host application wishes to send an 8 byte pointer update to PMEM only if the log data (write 1) was written to PMEM first
- Requester Host application issues a FLUSH operation after RDMA Write 1 to force the Responder HCA to flush the writes before completing the FLUSH response
- Without waiting for the FLUSH completion, the Requester Host application can queue an ATOMIC WRITE operation to update the 8 byte pointer
- Ordering rules prevent the Responder HCA from executing the ATOMIC WRITE until the previous FLUSH has completed
- This allows the Requester Host application to queue the RDMA WRITE, FLUSH, ATOMIC WRITE, in a pipelined manner, without waiting for this to be done by an ULP, and without stalling the pipeline

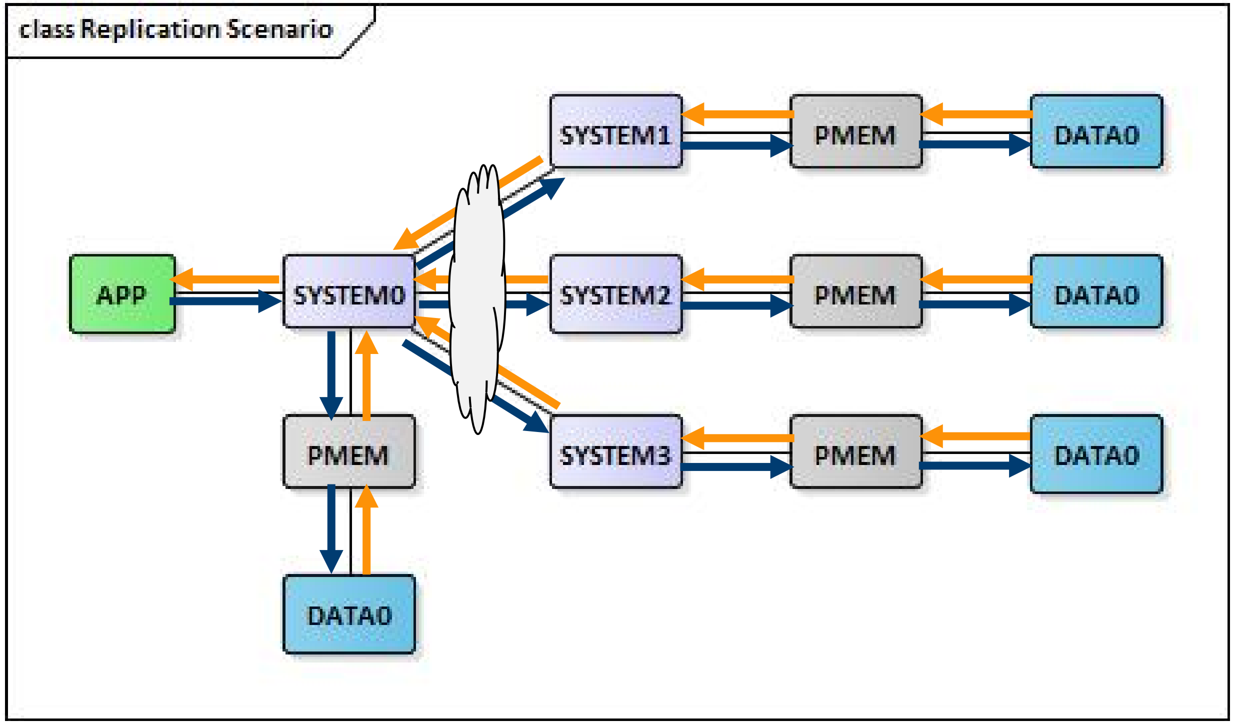


Transport Flow for synchronous log writing to PMEM – With MPE

Memory Placement Extensions

Scenarios where MPE can improve performance

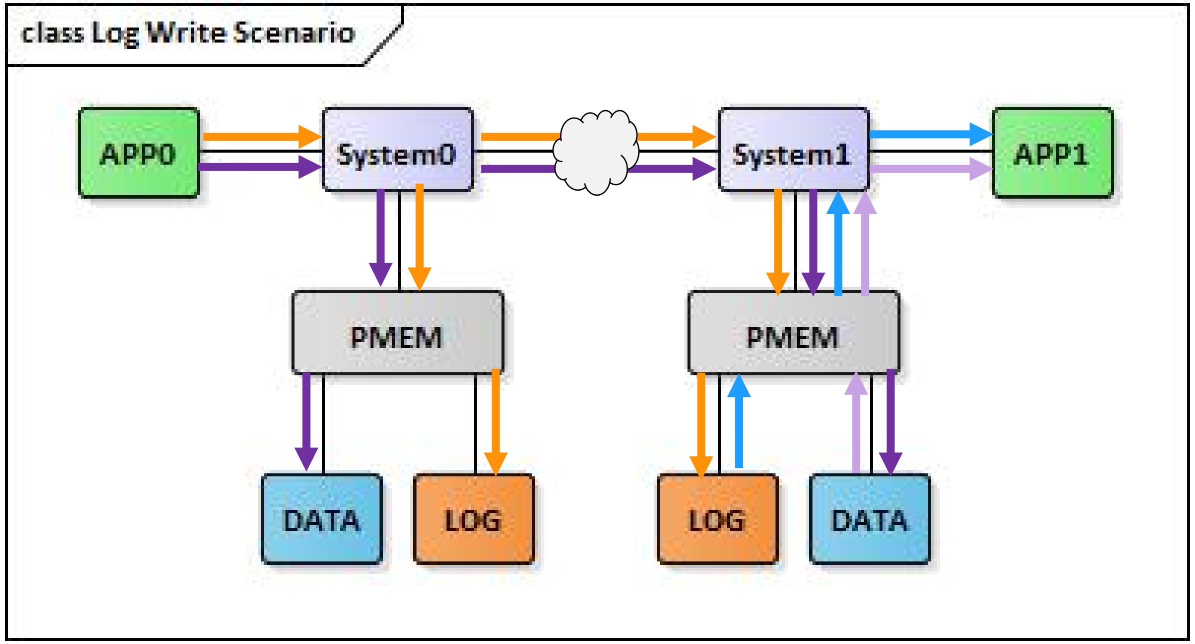
- ***Synchronous* Replication**
  - Multiple physical copies of all data are replicated on several systems before the original data is considered Highly Available (HA)
  - High priority network latency sensitive scenario for datacenters adopting PMEM
  - Advantageous to perform consistency checks on replica data in-line, pipelined with the remotes writes & flushes



Memory Placement Extensions

Scenarios where MPE can improve performance

- ***Synchronous* Log Writing**
  - Shared data is distributed amongst multiple systems
  - Log Files keep track of transactions applied to the data
  - High priority network latency sensitive workload for SQL adoption of PMEM
  - Advantageous to perform consistency checks on log data in-line, pipelined with the remote writes & flushes

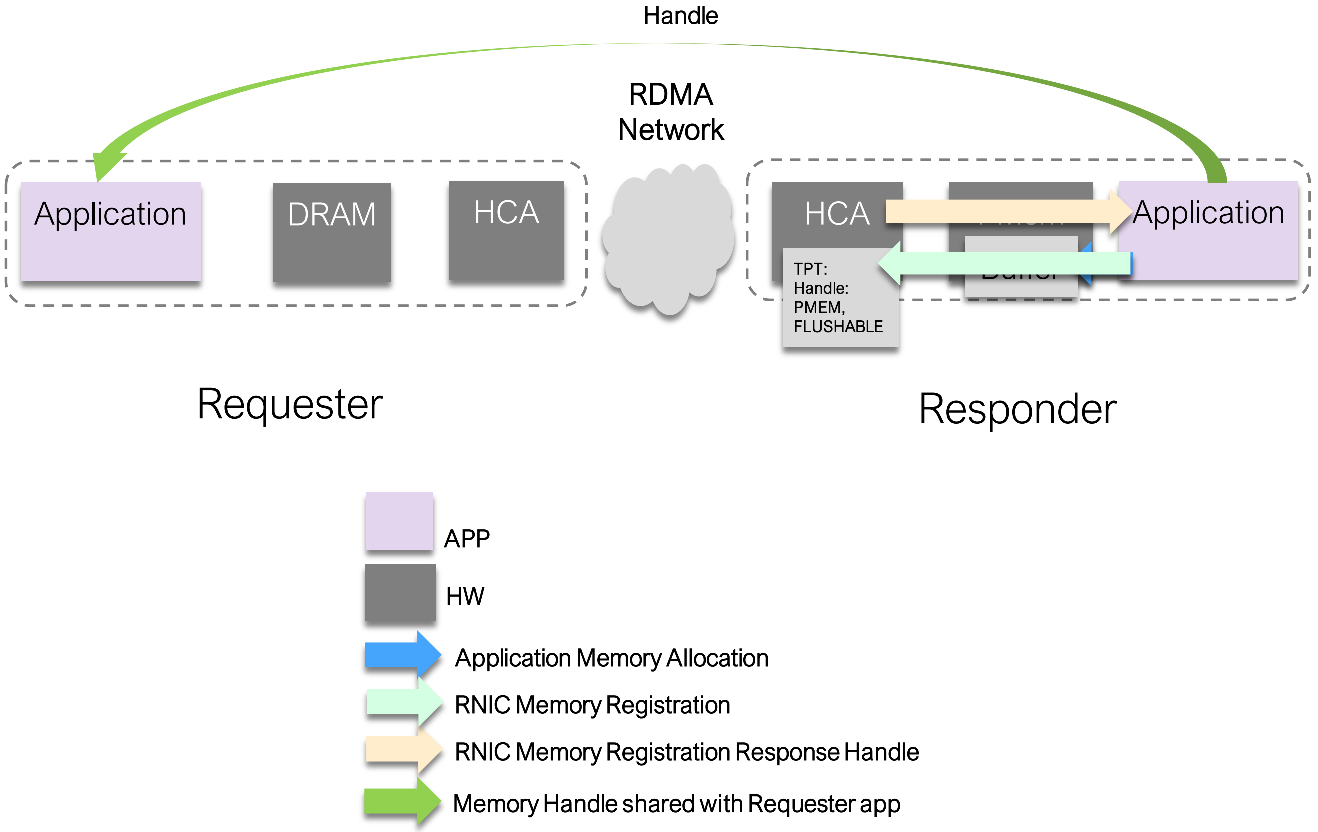


MPE Details

FLUSH and ATOMIC WRITE

Memory Placement Extensions

- MPE Initialization



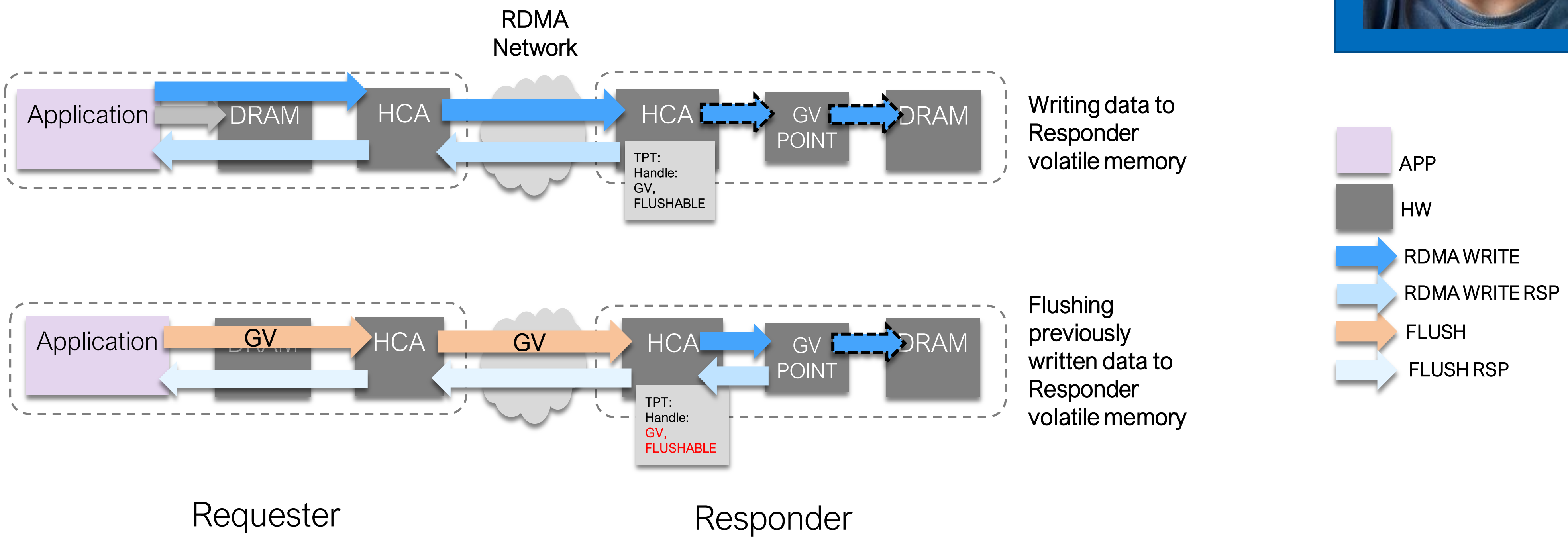
Allocate persistent memory
buffer on Responder

Registering memory with Responder HCA:
-Address
-Length
-PMEM – Used for GV/PMEM FLUSH selector
-FLUSHABLE – Indicates FLUSH of the region is allowed

Handle for memory registration made
available to Requester application

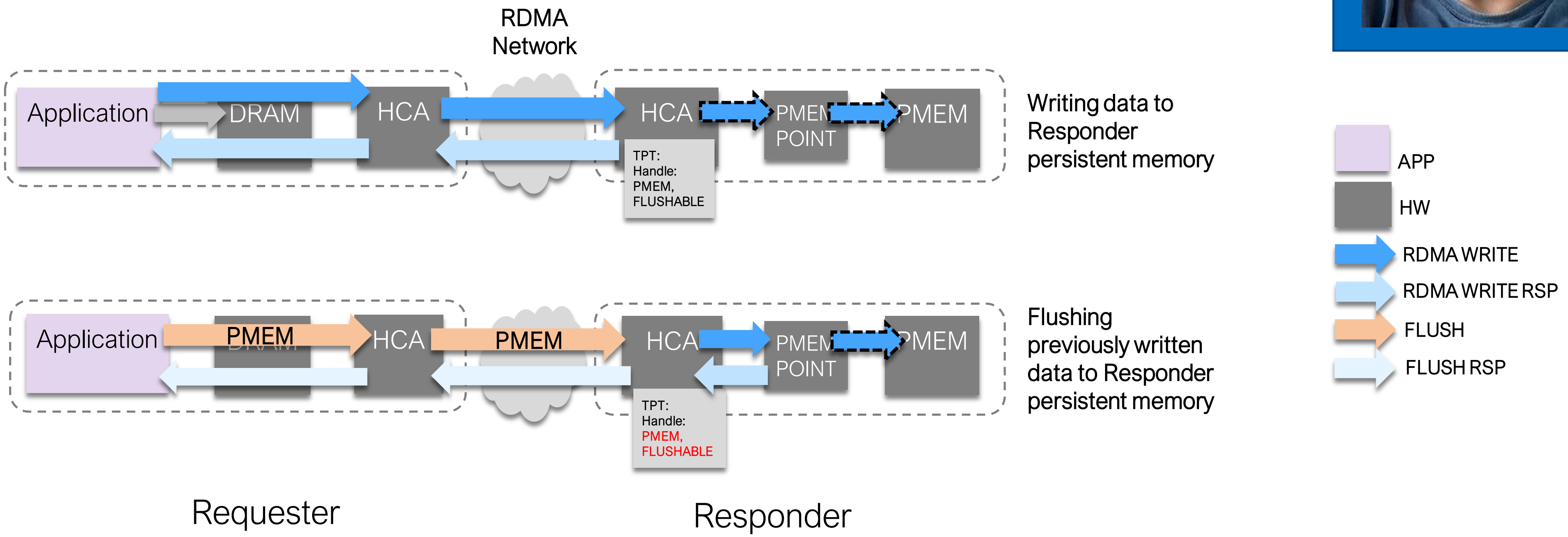
Memory Placement Extensions

- FLUSH to Global Visibility (GV)



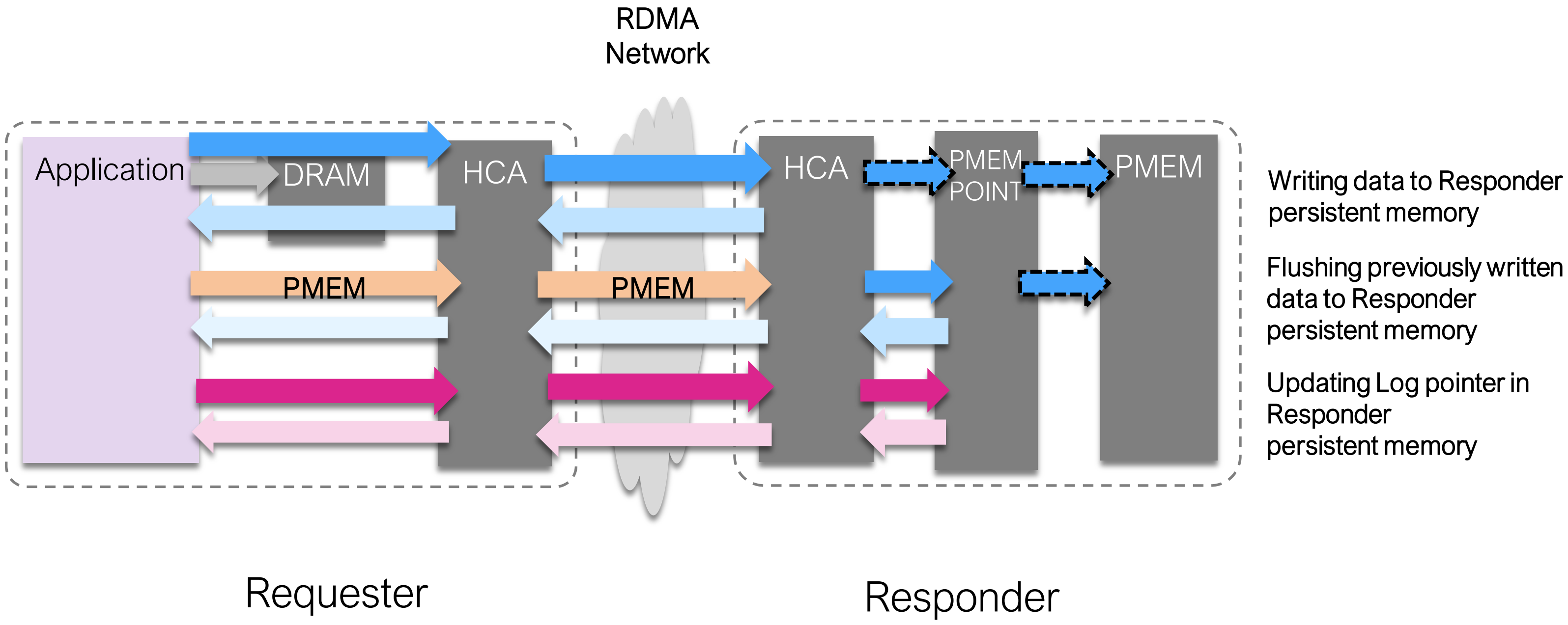
Memory Placement Extensions

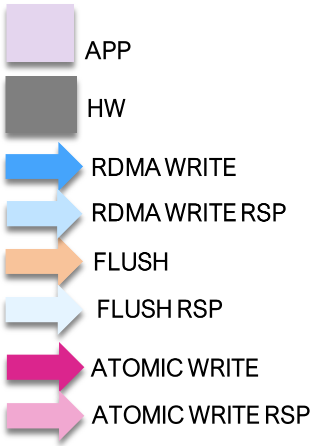
- FLUSH to Persistence (PMEM)



Memory Placement Extensions

- Pipelined Log Writing Example





Next steps

- Future version of spec considering:
  - Opcode extension to support additional operations
  - VERIFY operation to provide remote integrity checks
  - Continued integration of extensions into base document text

For more information

https://www.infinibandta.org/ibta-specification/

- RDMA vendors:
  - Implement MPE in your InfiniBand and RoCE adapter(s)
- RDMA users:
  - Enhance your application(s) and ULP(s) to leverage MPE

Please take a moment to rate this session.

Your feedback is important to us.