RDMA 101

David Peterson

T11-2019-00037-v001

BROADCOM®

RDMA 101

Basics

- Send
  - Push data to remote Queue Pair
  - Moves a single message (zero to 2^{31} bytes i.e., 2G)
- RDMA Write
  - Pushes data into remote virtual memory
  - Message is zero to 2^{31} bytes i.e., 2G
- RDMA Read
  - Pulls data out of remote virtual memory
  - Single RDMA Read Request size is zero to 2^{31} bytes i.e., 2G
- Atomic
  - Read/modify of remote memory location

RDMA 101

Base Transport Header (BTN)

bits bytes	31-24			23-16				15-8	7-0
0-3	OpCode			SE	M	Pad	TVer	Partition Key	
4-7	F/Res1 [a]	B/Res1 [a]	Reserved 6 [a]	Destination QP					
8-11	A	Reserved 7		PSN - Packet Sequence Number					

RDMA 101

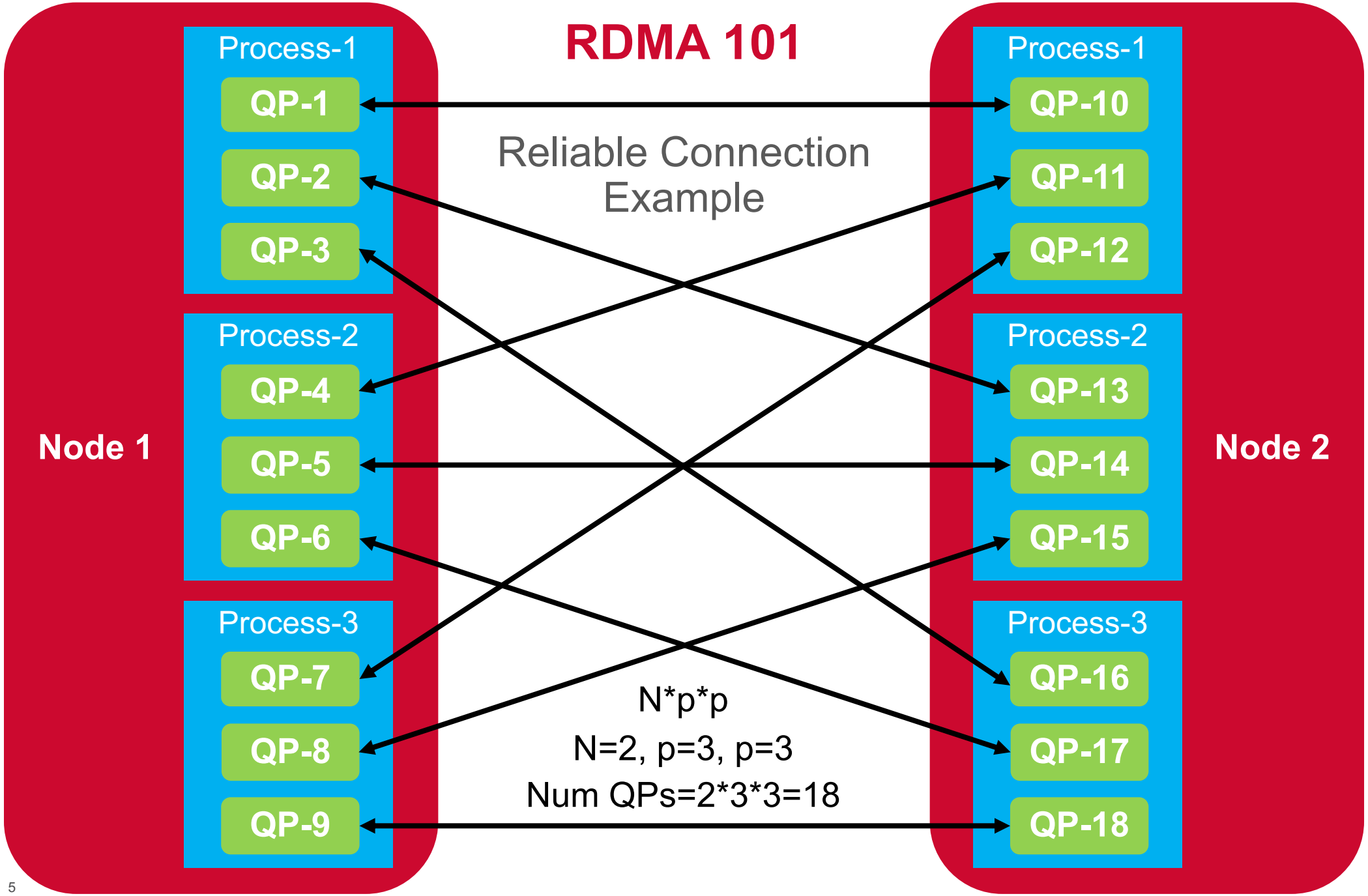
Transport Service Type

Opcode Field (000b) – Reliable Connection

With the Reliable Connected (RC) Transport Service, the number of QPs
required per endnode to achieve full process to process connectivity is
equal to N*p*p (where N is the number of nodes in the cluster and p the
number of processes per node). As the number of processes grows together with the number of cores per system, the number of RC QPs (and
its associated memory resources) start to become of significant impact.

Code[7-5]	Code[4-0]	Description	Packet Contents following the Base Transport header[a]
000 Reliable Connection (RC)	00000	SEND First	PayLd
	00001	SEND Middle	PayLd
	00010	SEND Last	PayLd
	00011	SEND Last with Immediate	ImmDt, PayLd
	00100	SEND Only	PayLd
	00101	SEND Only with Immediate	ImmDt, PayLd
	00110	RDMA WRITE First	RETH, PayLd
	00111	RDMA WRITE Middle	PayLd
	01000	RDMA WRITE Last	PayLd
	01001	RDMA WRITE Last with Immediate	ImmDt, PayLd
	01010	RDMA WRITE Only	RETH, PayLd
	01011	RDMA WRITE Only with Immediate	RETH, ImmDt, PayLd
	01100	RDMA READ Request	RETH
	01101	RDMA READ response First	AETH, PayLd
	01110	RDMA READ response Middle	PayLd
	01111	RDMA READ response Last	AETH, PayLd
	10000	RDMA READ response Only	AETH, PayLd
	10001	Acknowledge	AETH
	10010	ATOMIC Acknowledge	AETH, AtomicAckETH
	10011	CmpSwap	AtomicETH
	10100	FetchAdd	AtomicETH
	10101	Reserved	Undefined
	10110	SEND Last with Invalidate	IETH, PayLd
	10111	SEND Only with Invalidate	IETH, PayLd
	11000-11111	Reserved	undefined

RDMA 101



RDMA 101

Transport Service Type

Opcode Field (101b) –
Extended Reliable Connection

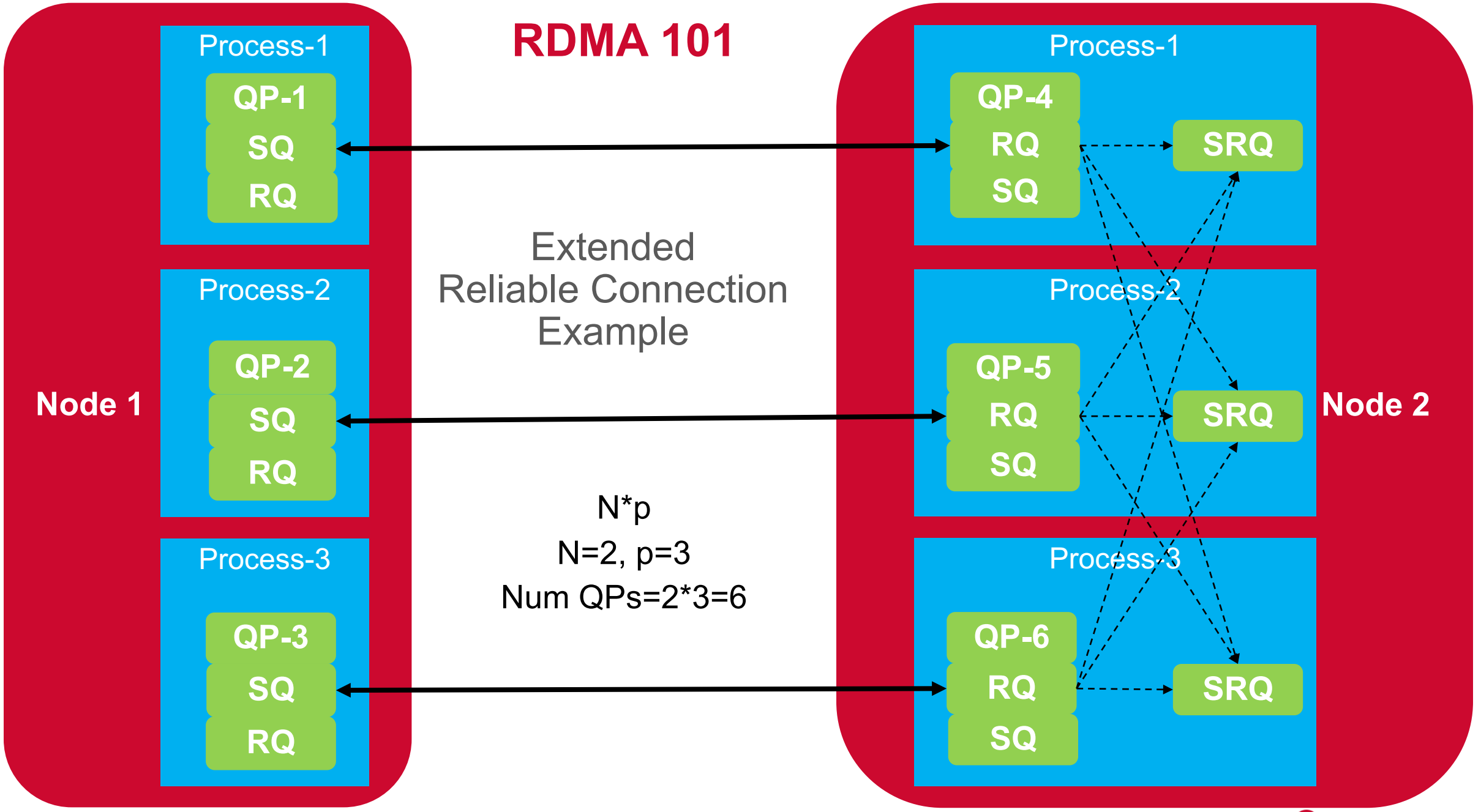
XRC allows significant savings in the number of QPs required to establish all to all process connectivity in large clusters.

A single (XRC INI) QP a process in one node can communicate with ALL processes on one remote node thus reducing by a factor of p the number of overall QPs required for full connectivity (as compared to when RC QPs are used).

- QPs = Nnodes x Nprocesses
- Decrease from Nnodes x Nprocesses^2

101 Extended Reliable Connection (XRC)	00000	SEND First	XRCETH, PayLd
	00001	SEND Middle	XRCETH, PayLd
	00010	SEND Last	XRCETH, PayLd
	00011	SEND Last with Immediate	XRCETH, ImmDt, PayLd
	00100	SEND Only	XRCETH, PayLd
	00101	SEND Only with Immediate	XRCETH, ImmDt, PayLd
	00110	RDMA WRITE First	XRCETH, RETH, PayLd
	00111	RDMA WRITE Middle	XRCETH, PayLd
	01000	RDMA WRITE Last	XRCETH, PayLd
	01001	RDMA WRITE Last with Immediate	XRCETH, ImmDt, PayLd
	01010	RDMA WRITE Only	XRCETH, RETH, PayLd
	01011	RDMA WRITE Only with Immediate	XRCETH, RETH, ImmDt, PayLd
	01100	RDMA READ Request	XRCETH, RETH
	01101	RDMA READ response First	AETH, PayLd
	01110	RDMA READ response Middle	PayLd
	01111	RDMA READ response Last	AETH, PayLd
	10000	RDMA READ response Only	AETH, PayLd
	10001	Acknowledge	AETH
	10010	ATOMIC Acknowledge	AETH, AtomicAckETH
	10011	CmpSwap	XRCETH, AtomicETH
	10100	FetchAdd	XRCETH, AtomicETH
	10101	Reserved	Undefined
	10110	SEND Last with Invalidate	XRCETH, IETH, PayLd
	10111	SEND Only with Invalidate	XRCETH, IETH, PayLd
	11000-11111	Reserved	undefined

RDMA 101



RDMA 101

Extended Reliable Connection – A few more details

- XRC Domain
  - Attribute used to associate XRC TGT QPs and XRC SRQs. XRC packets can only target XRC SRQs in the same XRC Domain as the XRC TGT QP that they are destined for.
- XRC INI QP
  - XRC Initiator QP. This is the initiator Queue for XRC operations. XRC INI QPs are used to issue XRC outgoing requests and do not have a responder side. XRC incoming requests will be handled by XRC TGT QPs.
- XRC SRQ
  - This is the Receive Queue where Receive WQEs are posted for incoming XRC requests. XRC request packets carry in an extended header (XRCETH) the XRC SRQ number that is being targeted and from which a receive WQE will be fetched if required.
- XRC TGT QP
  - XRC Target QP. This is the responder for XRC operations. XRC TGT QPs (together with XRC SRQs) are used to process incoming XRC requests. XRC TG QPs do not have a requester side. XRC outgoing requests are issued through XRC INI QPs.
- XRCETH XRC
  - Extended Transport Header. Present in XRC request packets.

RDMA 101

Transport Service Type

Opcode Field (011b) – Unreliable Datagram

Code[7-5]	Code[4-0]	Description	Packet Contents following the Base Transport header[a]
011 Unreliable Datagram (UD)	00000-00011	Reserved	undefined
	00100	SEND only	DETH, PayLd
	00101	SEND only with Immediate	DETH, ImmDt, PayLd
	00110-11111	Reserved	undefined

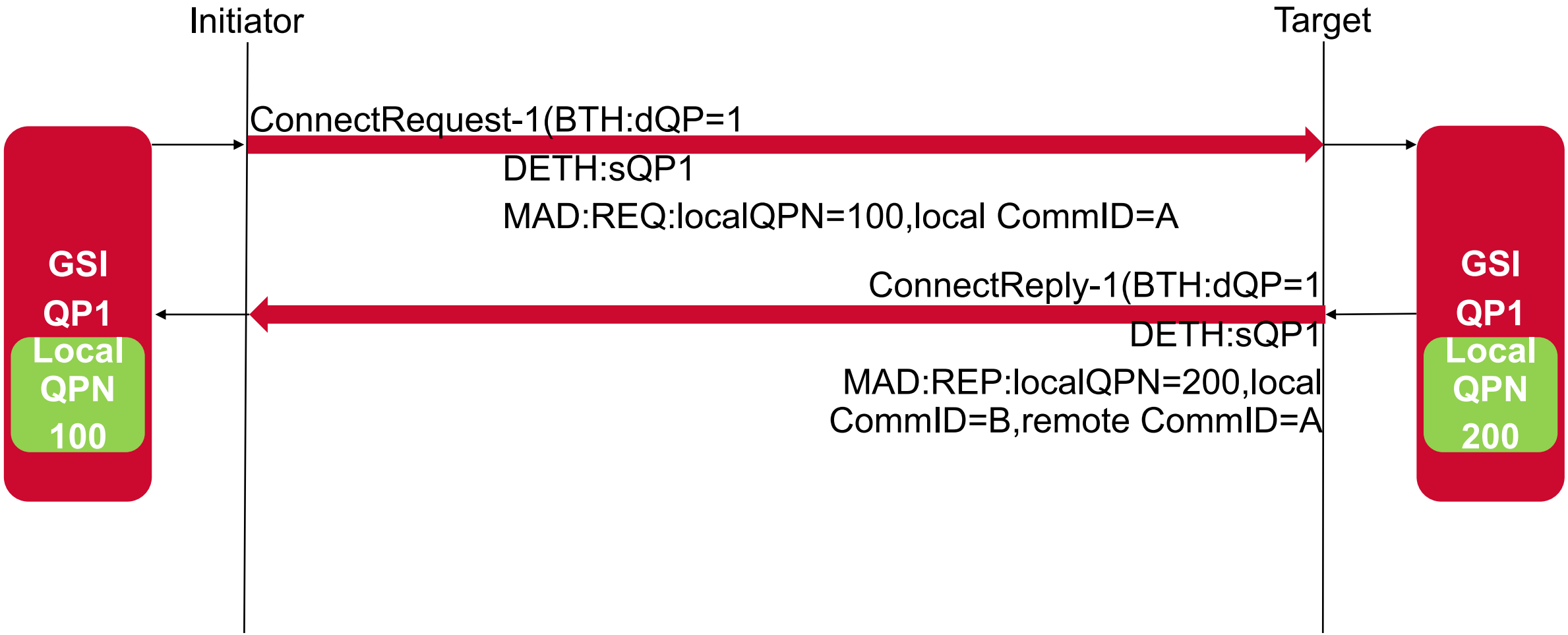
RDMA 101

Overview

- Hosts initialize context and register memory regions
- Establish connection
- Use Send/Receive model to exchange memory region keys between peers
- Post read/write operations
- Disconnect

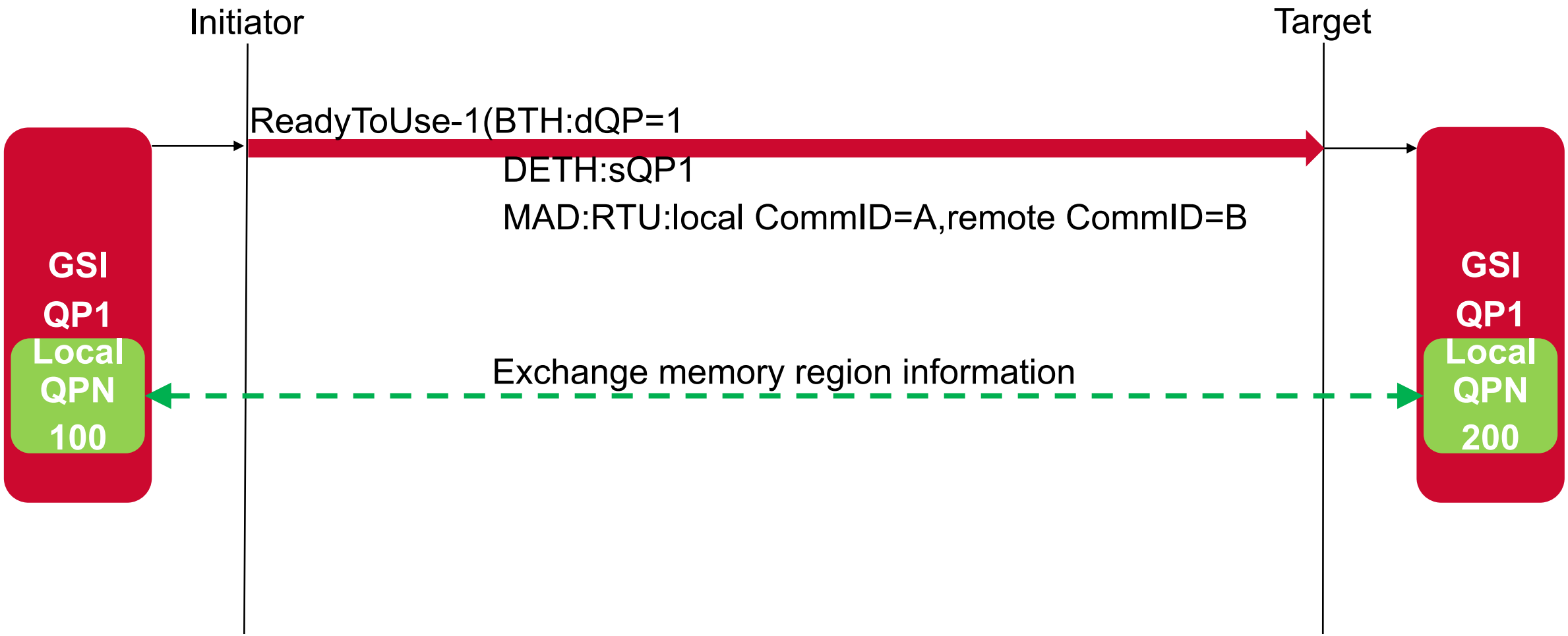
RDMA

Example IB Initialization & Connection Setup-1-Create “Admin” Connection & QP



RDMA

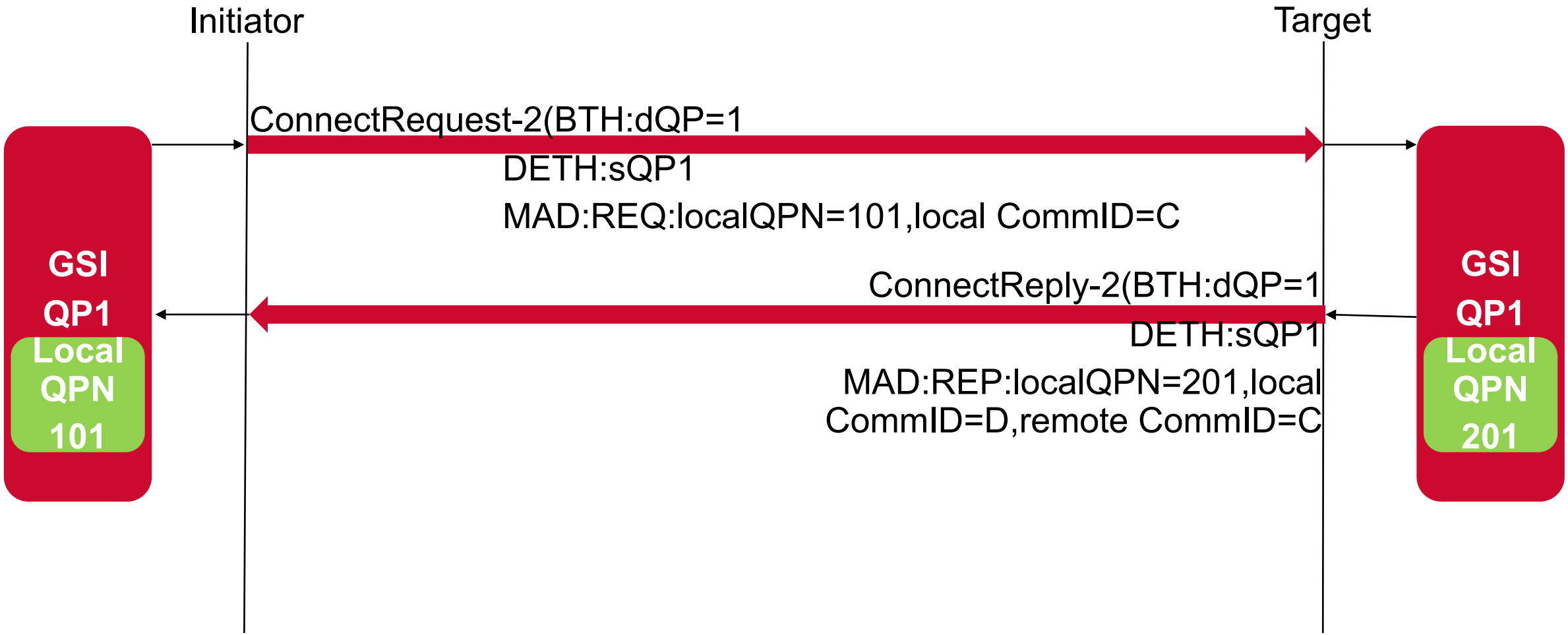
Example IB Initialization & Connection Setup-2-Send Ready To Use & Exchange Info



BROADCOM

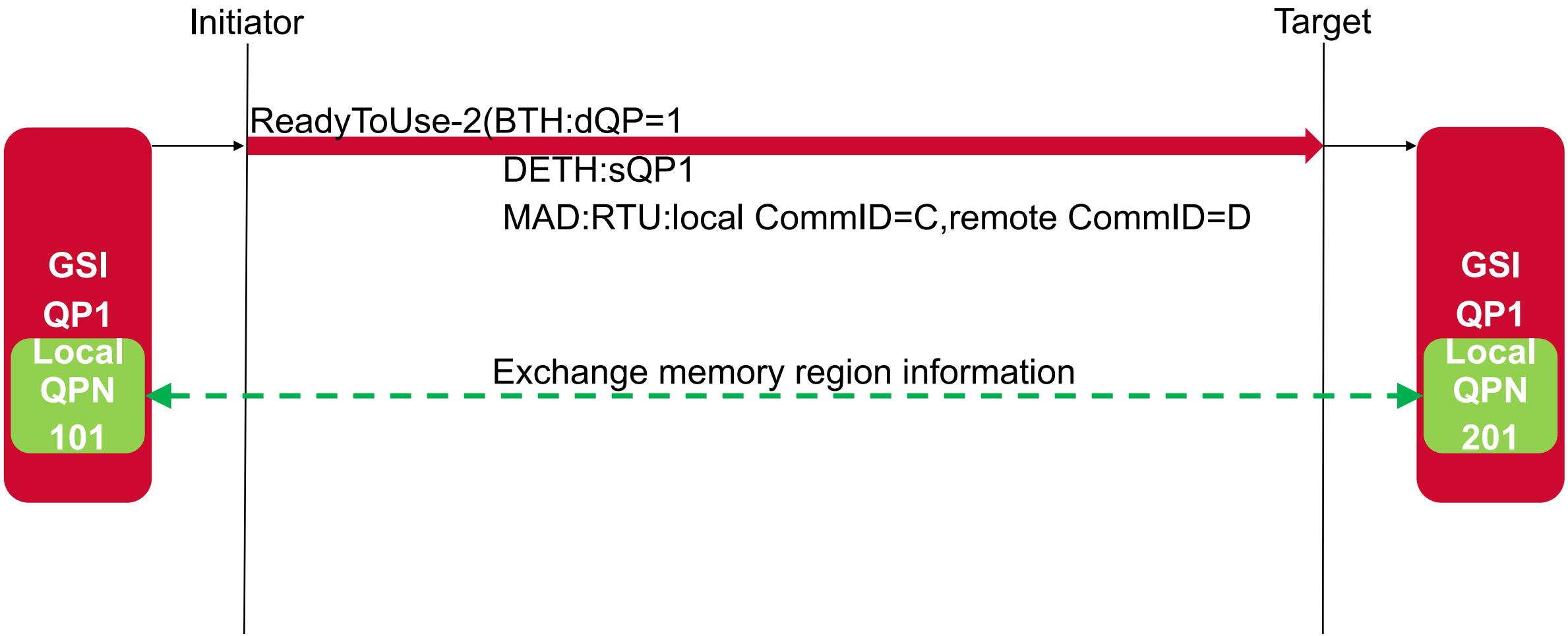
RDMA

Example IB Initialization & Connection Setup-3-Create “I/O” Connection & QP



RDMA

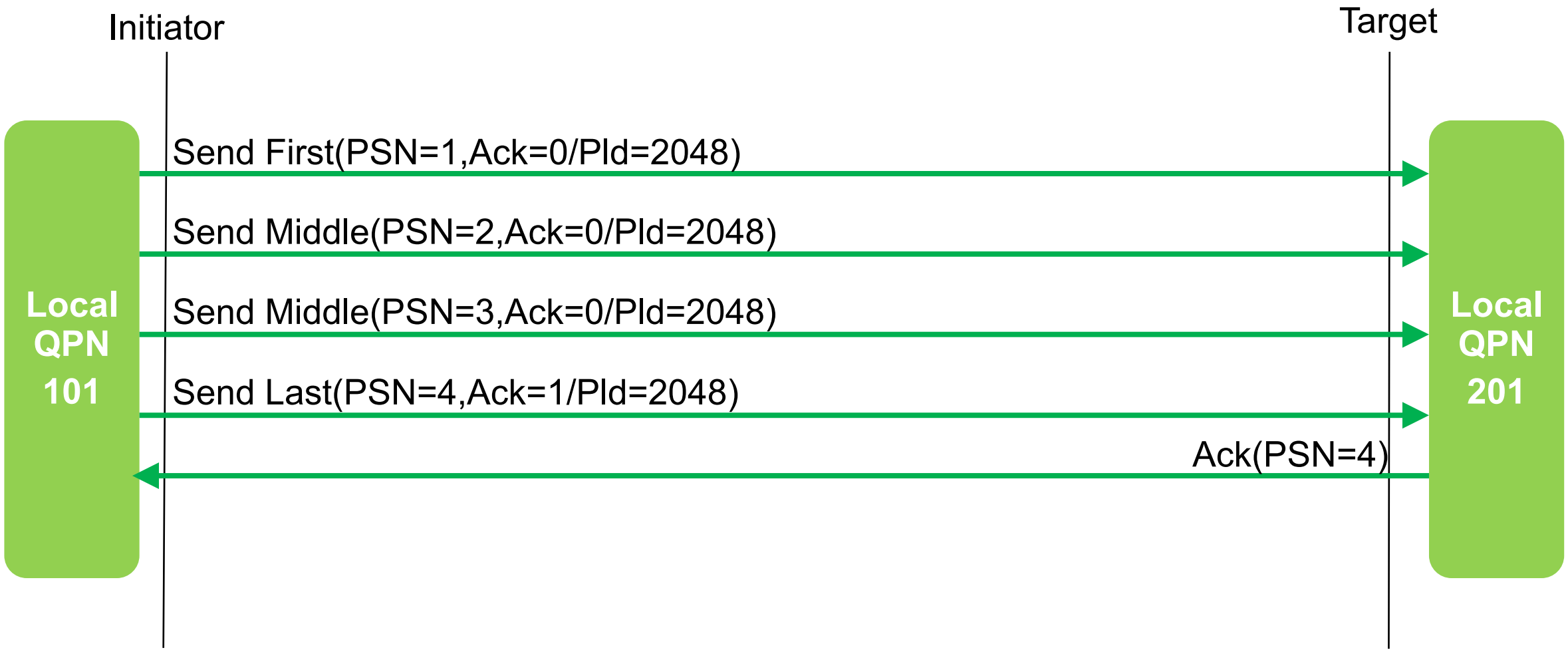
Example IB Initialization & Connection Setup-4-Send Ready To Use & Exchange Info



BROADCOM

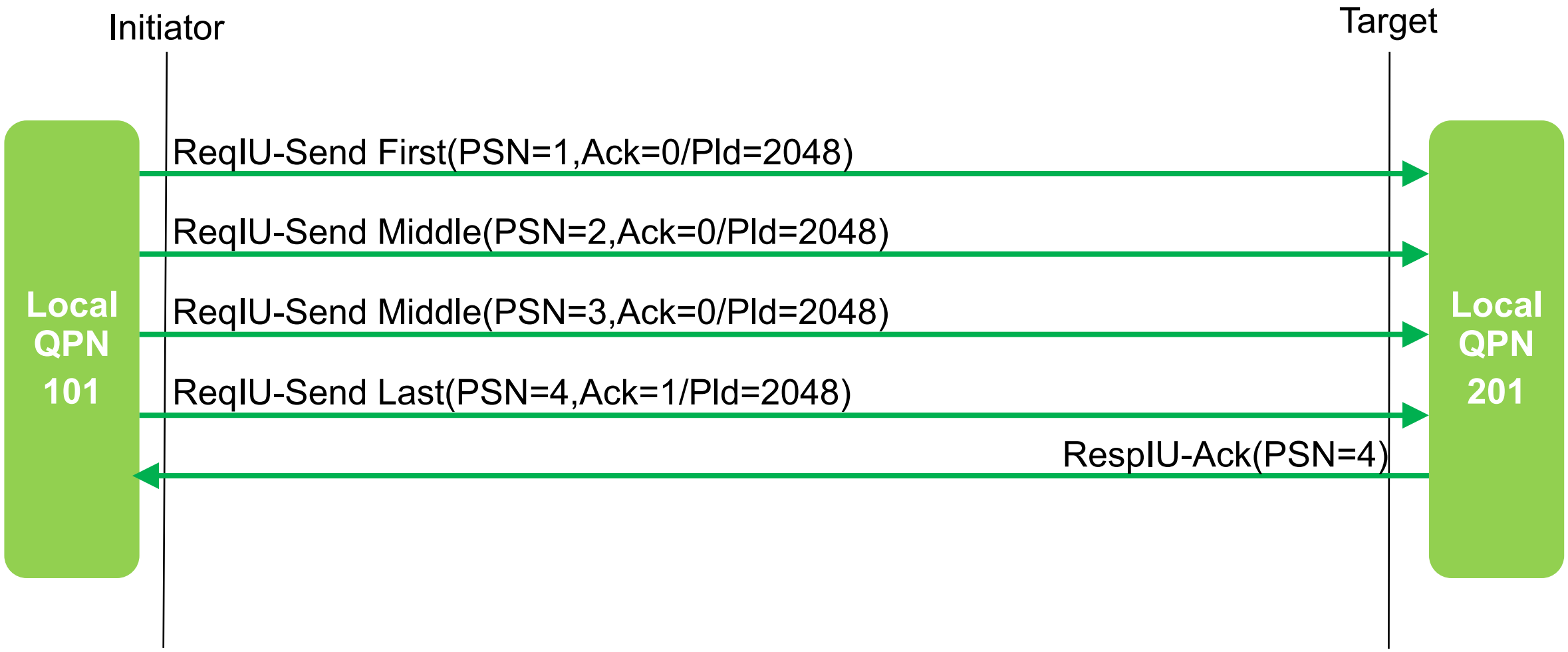
RDMA 101

Example Send – 8k



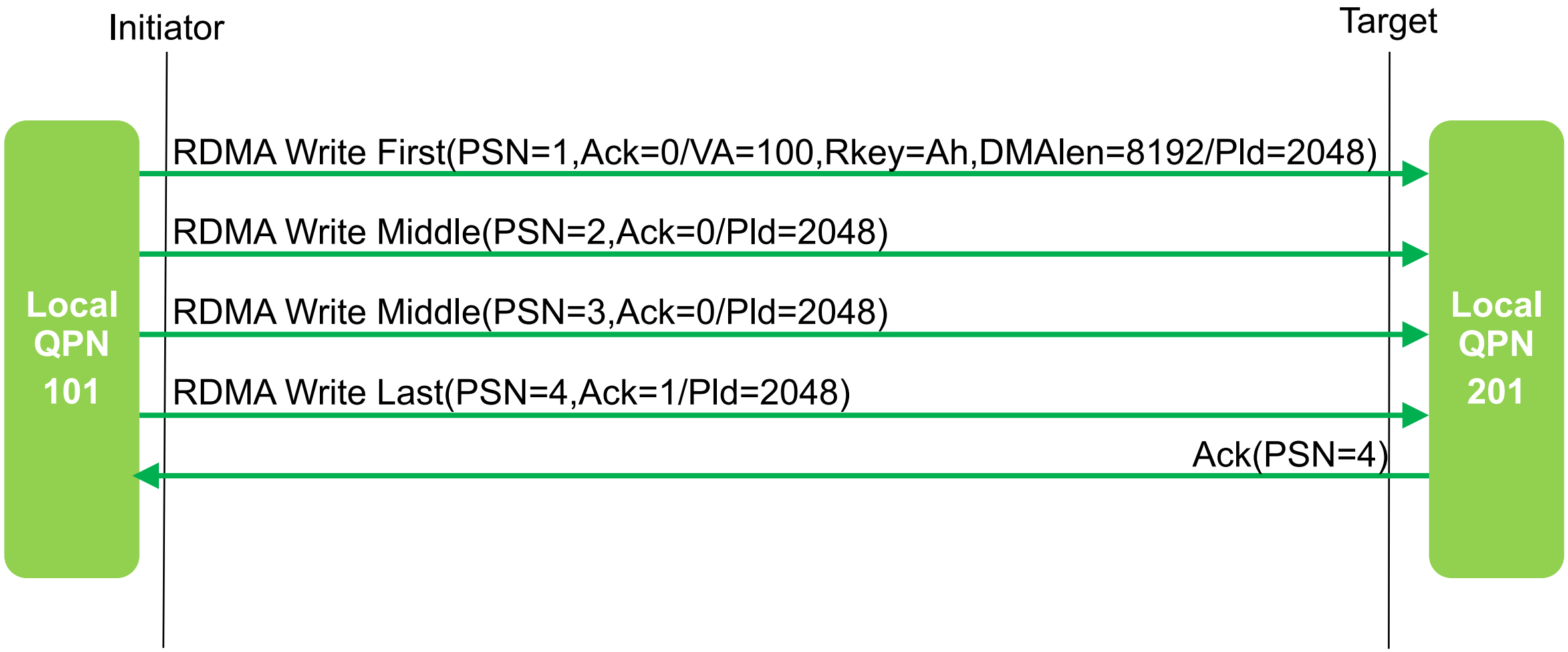
RDMA 101

Proposed FC-RDMA Send – 8k



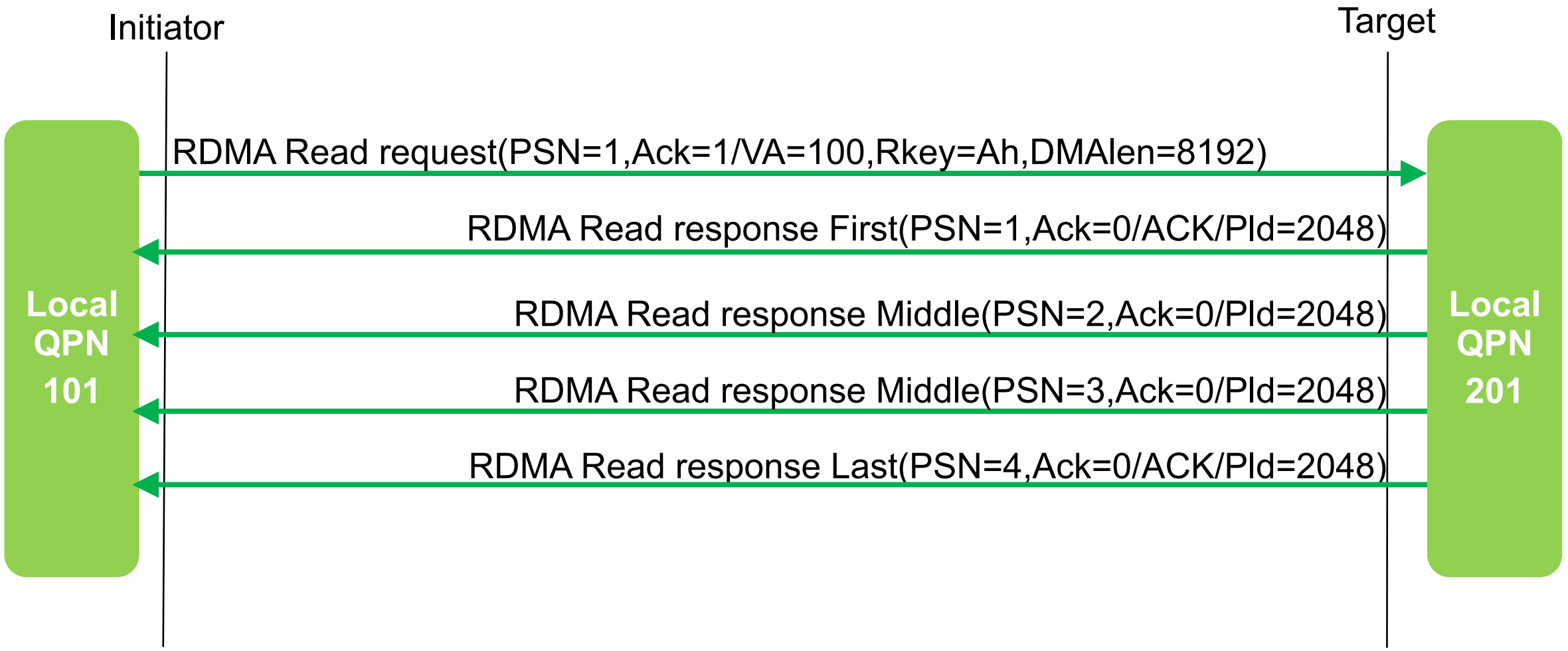
RDMA 101

Example RDMA Write – 8k



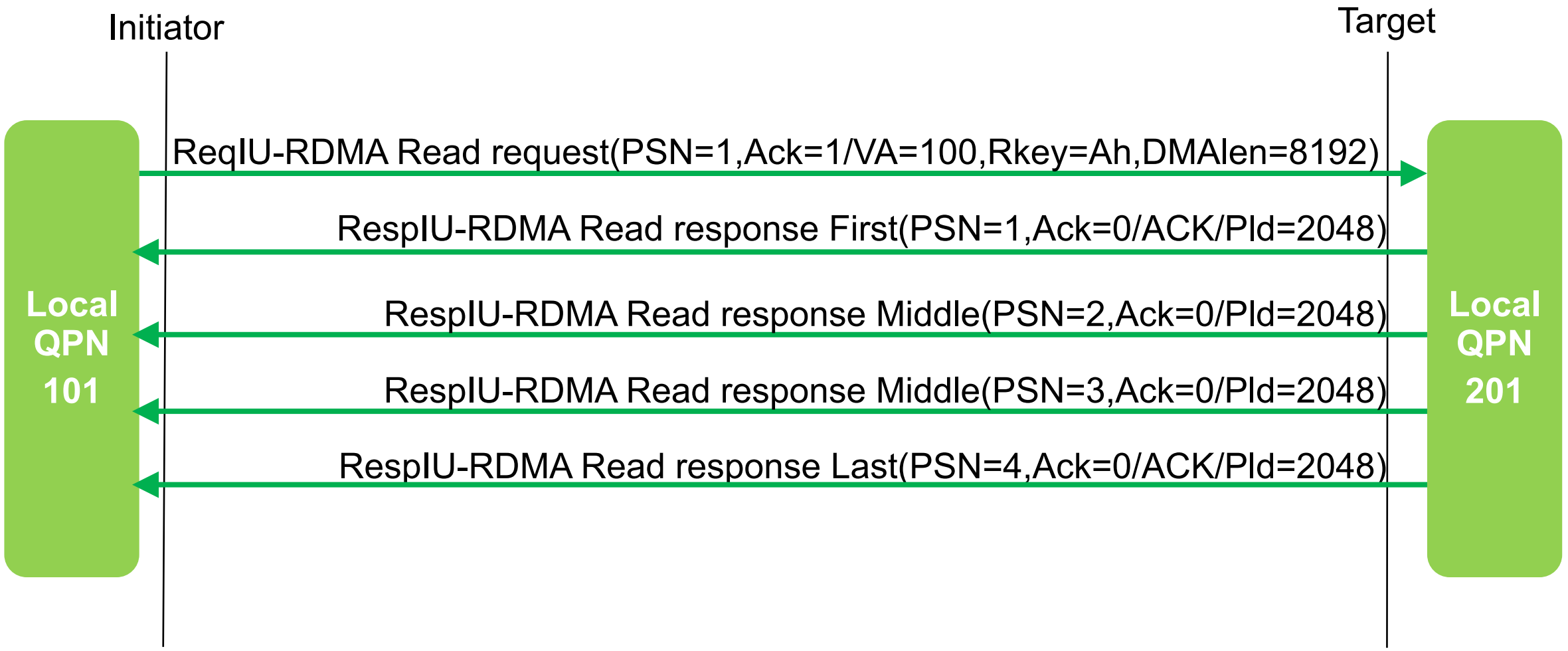
RDMA 101

Example RDMA Read – 8k



RDMA 101

Proposed FC-RDMA RDMA Read – 8k



Thank You