xsigo systems

xsigo systems

Infiniband Day03

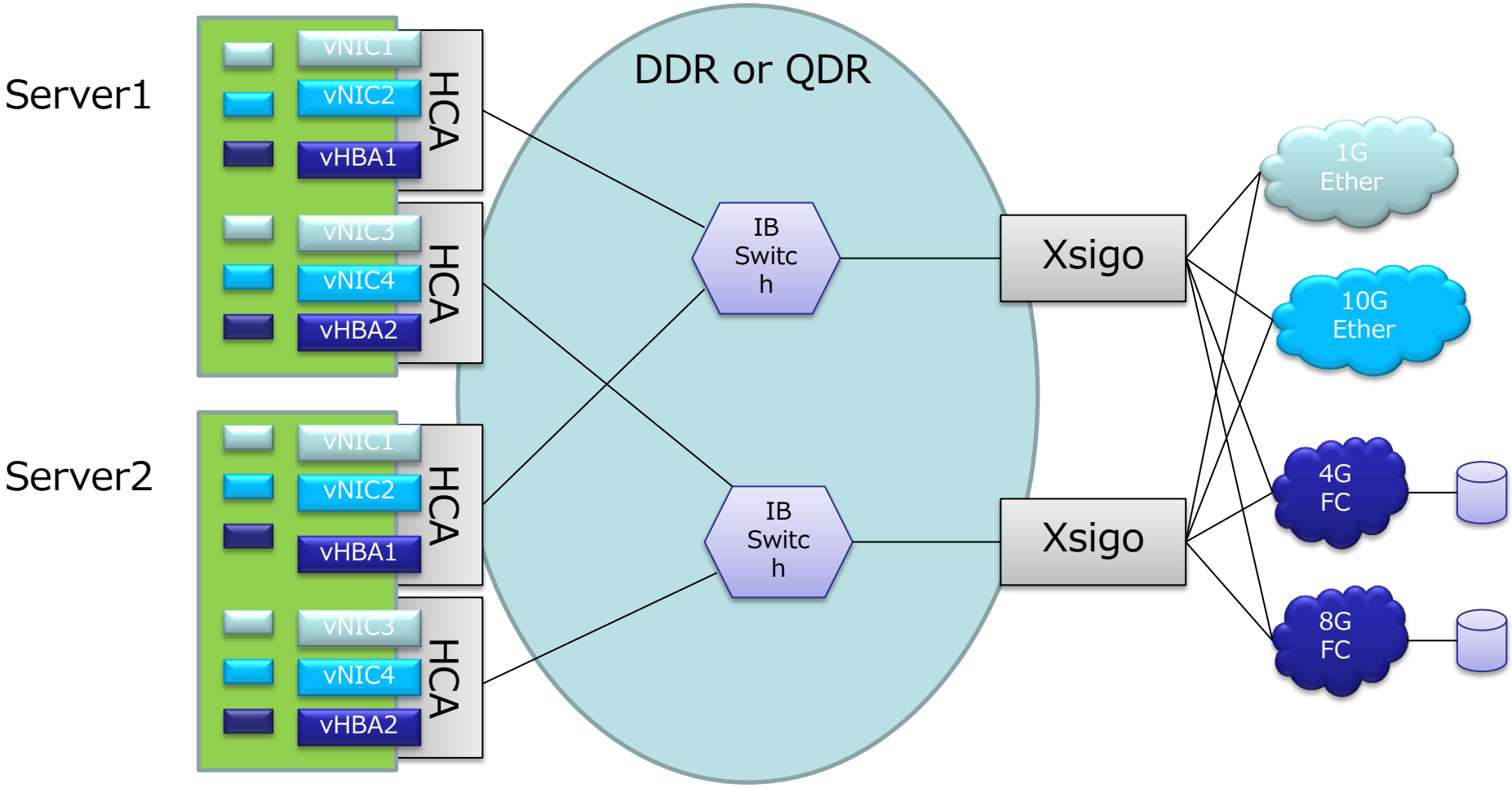
トランスポートレイヤー概要

2011年5月

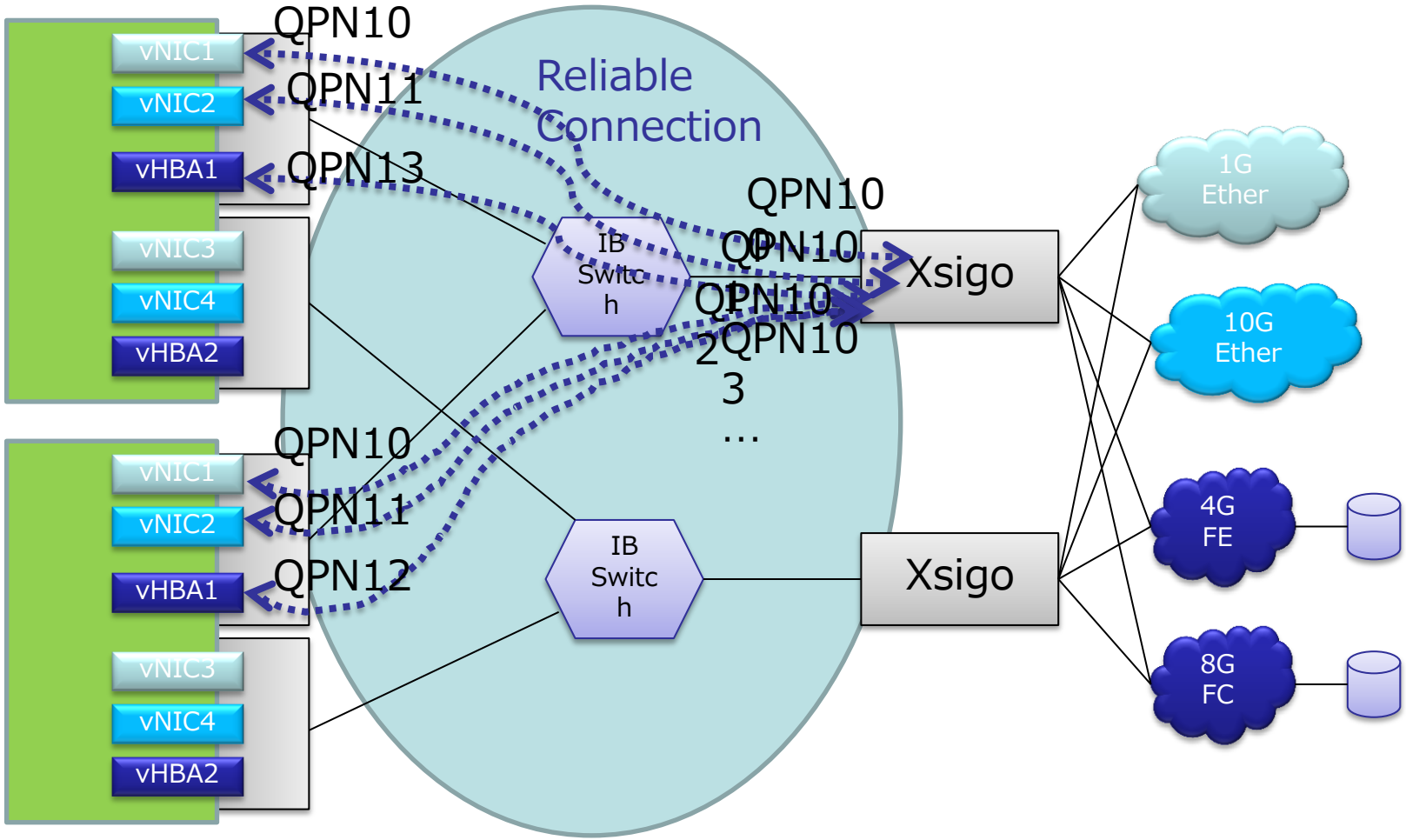
アジェンダ

- Xsigo紹介
- レイヤー・アーキテクチャー
- データ転送方式
- Work Request
- QP
- パケット・フォーマット
- Opcode
- Packet Sequece NumberとACK
- Inter Packet Delay
- メッセージ送信例

Xsigo紹介(1)

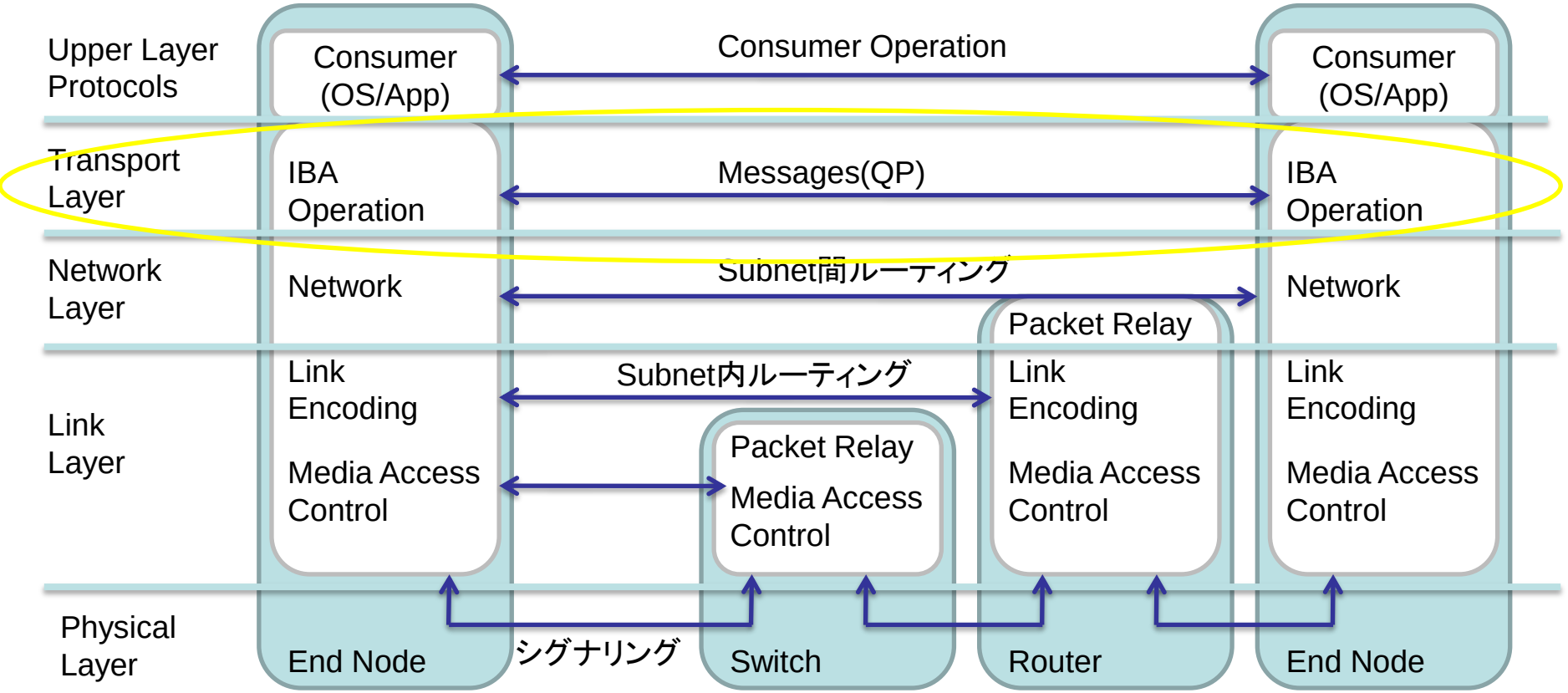

Server1
Server2
vNIC1
vNIC2
vHBA1
vNIC3
vNIC4
vHBA2
HCA
HCA
HCA
HCA
DDR or QDR
IB Switch
IB Switch
Xsigo
Xsigo
1G Ether
10G Ether
4G FC
8G FC

Xsigo紹介(2)


QPN10
QPN11
QPN13
vNIC1
vNIC2
vHBA1
vNIC3
vNIC4
vHBA2
Reliable
Connection
IB
Switc
h
QPN10
QPN10
QPN10
2QPN10
3
…
Xsigo
1G
Ether
10G
Ether
4G
FE
8G
FC
QPN10
QPN11
QPN12
vNIC1
vNIC2
vHBA1
vNIC3
vNIC4
vHBA2
IB
Switc
h
Xsigo

Infiniband Architecture(IBA) レイヤー・アーキテクチャ

- 本セッションではIBAのTransportレイヤについて概要を説明します。
- 詳細は、*"InfiniBand Architecture Specification"*をご参照ください。

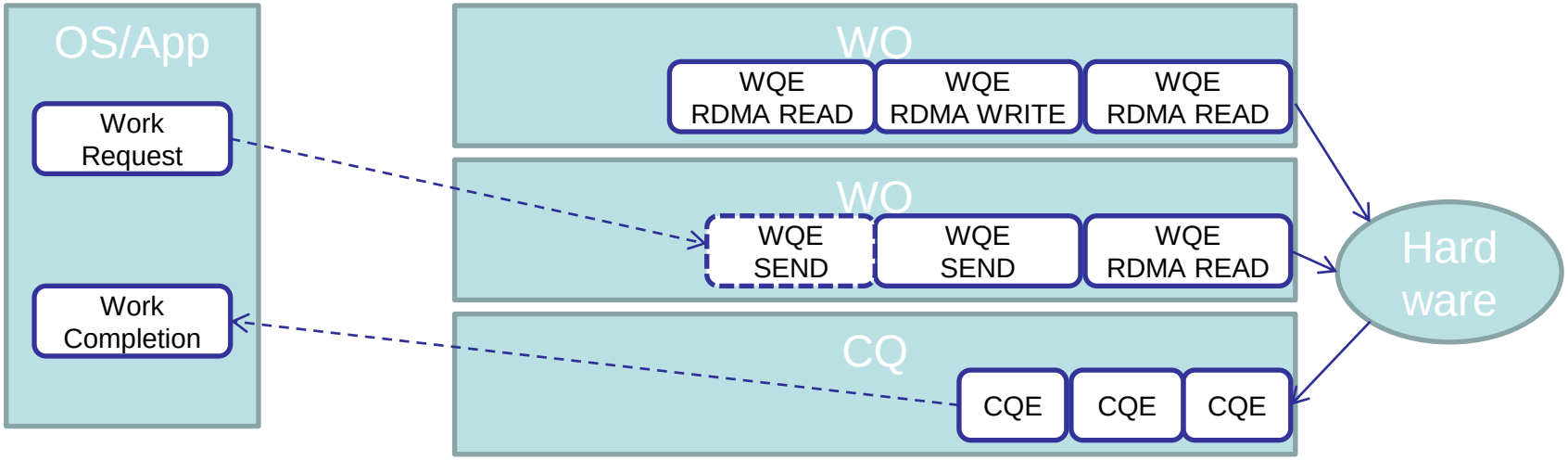


Infiniband Architecture(IBA) データ伝送方式

	Reliable Connection (RC)	Reliable Datagram (RD)	Unreliable Connection (UC)	Unreliable Datagram (UD)	Raw Datagram (IPv6&Ether)
Ack	有	有	無	無	無
コネクション	有 ローカルQPとリモートQPがひとつづつ割り当てられます。	無 End-to-Endコンテキスト(EEC)と呼ばれるノード間の接続上で多重化されます。EECは、各RD QPがEECを確立した任意のノードの任意のQPと通信できます。複数のQPが同じEECを使うことができ、ひとつのQPが複数のEECを使うこともできます。	有 ローカルQPとリモートQPがひとつづつ割り当てられます。ACKがないので、パケットがロストしたり、壊れたりしても再送しません。そのようなパケットは単純にドロップされます。それ以外はRCと似ています。	無 任意のノードの任意のアンリライアブル・データグラムQPへも通信できます。	無 ロー・データグラム・パケットではQP番号は指定されません。
データ転送保証	有	有	無	無	無
データ順序保証	有	有	無	無	無
データ・ロス検知	有	有	有 受信側でエラー検知されます。送信側には通知されません。	無	無
メッセージサイズ	< 2Gバイト	< 2Gバイト	< 2Gバイト	< 4Kバイト	< 4Kバイト
マルチキャスト	無	無	無	有	有
RDMA Write/Read	可	可	Write:可、READ:不可	不可	不可

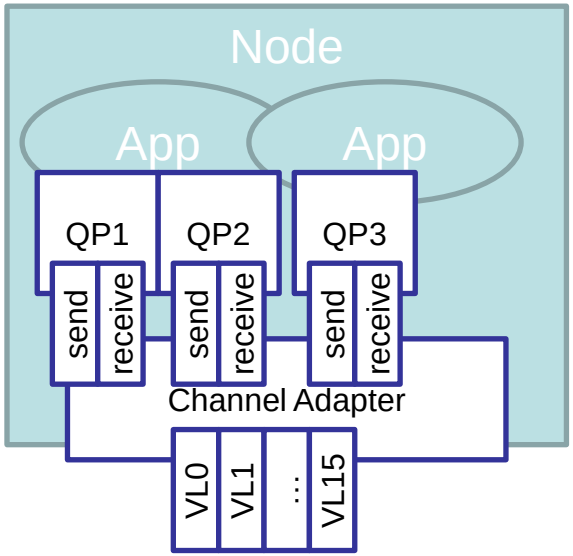
Work Request

- Work Request（WR)は、ソフトウェア・トランスポートレイヤーでの基本的な要素です。
- WRは、HCA内でWork Queue Element(WQE)として、QPのSendキュー、Receiveキューへキューイングされます。QP毎のWQで順序が保たれます。
- リクエストの処理が行われた後は、WQEがCompletion Queue(CQE)に移され、OS/AppはWork Completionを受け取ります。



Queue Pairs (QP)

- QPとは、ハードウェアが上位のConsumerに提供する仮想的なインタフェース
- リモートノードに対して、仮想的な送受信のそれぞれの通信ポートを提供します。リモートノードのQPとそれぞれ対に作成されます。
- ひとつのChannel Adapter(CA)につき、最大で2^24(16M) QPまでサポートされます。
- 各QPは独立して動作し、ほかのQPの動作の影響を受けません。



パケット・フォーマット

- Local Routing Header(LRH): 宛先LID、送信元LID、サービスレベル（VL)を指定。
- Global Routing Header(GRH): 異なるサブネット間でのルーティング。RouterはVCRCを再計算。
- Base Transport Header(BTH): 宛先QP番号、パケットシーケンス、オペレーションコード（Opcode）を指定。
- Extended Transport Header(ETH): オペレーションコードなどに依存。
- Invariant CRC(ICRC): ファブリック内で不変のCRC（GRH以外を対象）
- Variant CRC(VCRC): GRHも含め対象

Opcode

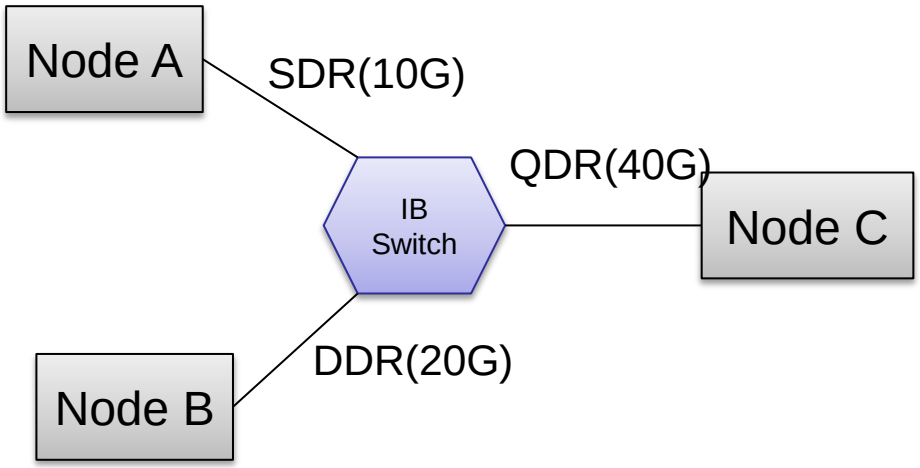
Code[7-5]	Code[4-0]	Description	Packet Contents following the Base Transport header[a]
000 Reliable Connection (RC)	00000	SEND First	PayLd
	00001	SEND Middle	PayLd
	00010	SEND Last	PayLd
	00011	SEND Last with Immediate	ImmDt, PayLd
	00100	SEND Only	PayLd
	00101	SEND Only with Immediate	ImmDt, PayLd
	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

Packet Sequence NumberとACK

- Packet Sequence Number(PSN)は、コネクション確立時に初期化され、QPでパケット生成時にひとつづつ増加されます。
- 受信側QPでは、PSNによりパケットロストを検知します。
- Reliableサービスでは、ACKまたはNAKパケットを返信し、送信側へ正しく受信できたかどうかを通知します。送信側ではエラーを検知すると再送を行います。

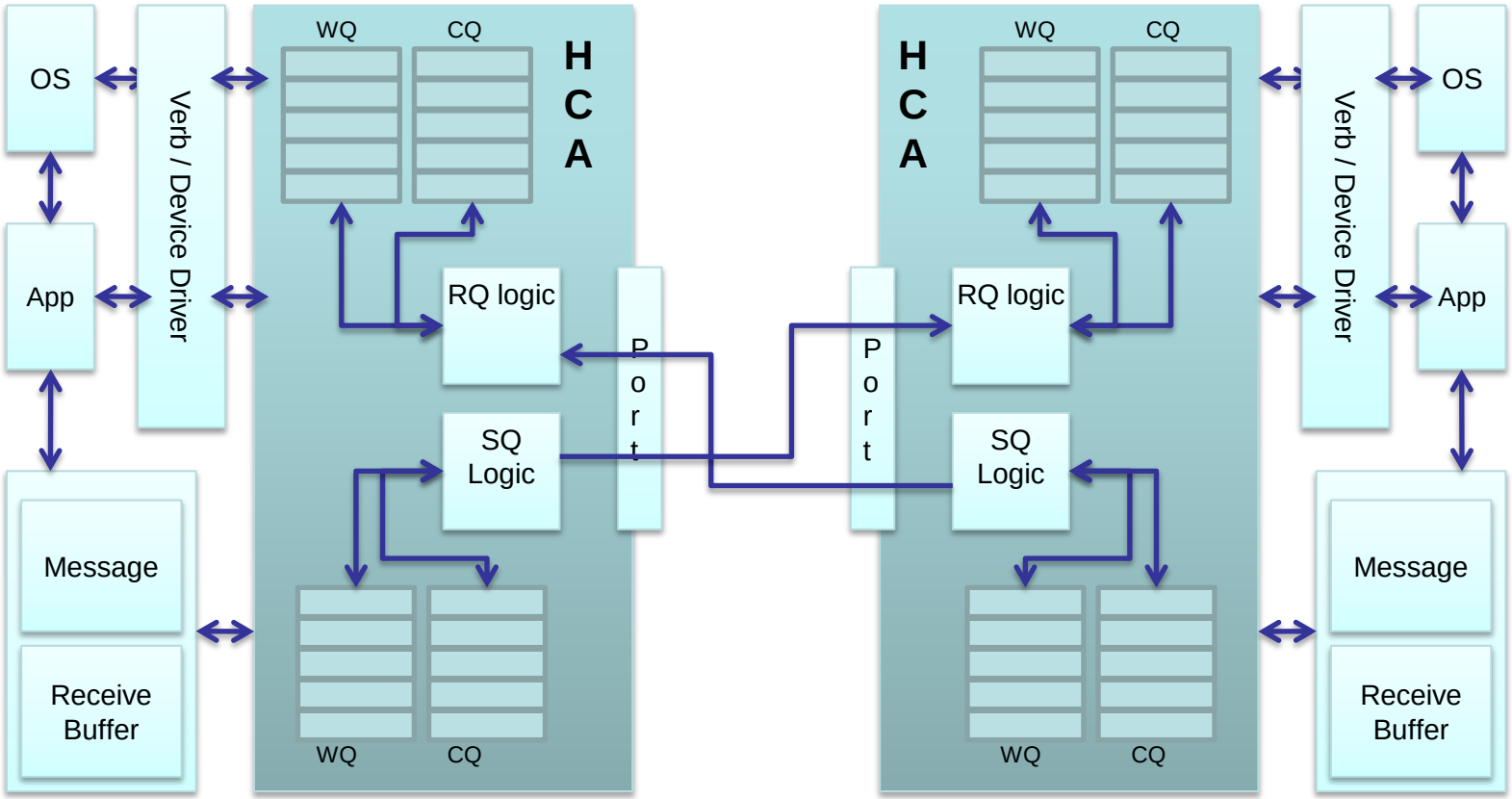
Inter Packet Delay (IPD)

- 異なるリンクスピードとの通信においては、送信パケット間に適当なDelayが指定され、バッファのオーバフローを防ぎます。
- QP ContextやEE Context内では、各宛先に対するPathRecordが参照されます。
- PathRecordには、Path MTUとともにPath Speedの情報が含まれます。



IPD	rate	Comment
0	100%	Suited for matched links
1	50%	
2	33%	Suited for 30 Gbps to 10 Gbps conversion
3	25%	Suited for 10 Gbps to 2.5 Gbps conversion
11	8%	Suited for a 30 Gbps to 2.5 Gbps conversion

メッセージ送信例:
5KB MessageのRC Send



Mesg 5KB	Create QP	WR	WQE	Data2KB	cPSN 100

ePSN 100	ePSN 101	Data 2KB	Data 2KB	PSN 101	Send Middle	Data 1KB	Data 1KB	PSN 102	Send Last	CQE	WRC

Buff 5KB	Create QP	WR	WQE	Data 2KB	PSN 100	Send First	Ack 100	unAcked PSN 100	cPSN 101	Ack 101	unAcked 101	cPSN 102	ePSN 102	Ack 102	unAcked 102

CQE	WRC