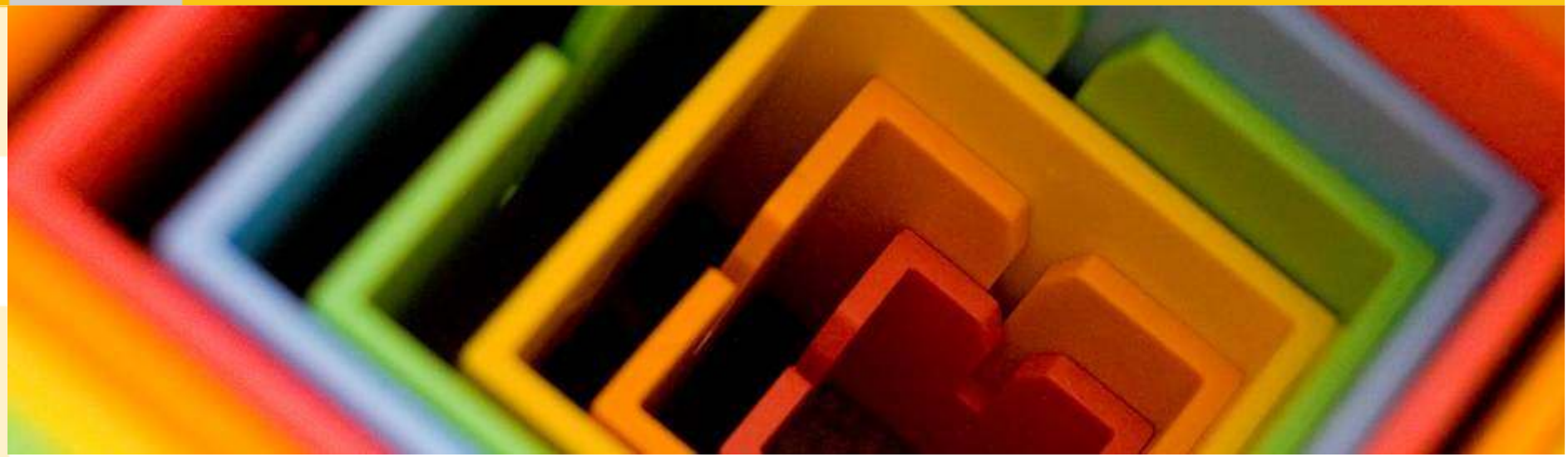
InfiniBand & Manycore Day

変える力を、ともに生み出す。
NTT DATAグループ

プログラマ目線から見たRDMAのメリットとその応用例について



2010年11月17日
株式会社NTTデータ
技術開発本部　伊藤雅典

INDEX

00 自己紹介

変える力を、ともに生み出す。
NTT DATAグループ

氏名

伊藤 雅典（いとう まさのり）

所属

株式会社NTTデータ 技術開発本部 ITアーキテクチャ＆セキュリティ技術センタ

担当業務

NTTデータの総合クラウドサービス BizXaaS™ （http://bizxaas.net/）の、
「フルOSSクラウド構築ソリューション」の開発ほかに従事

http://www.nttdata.co.jp/release/2010/040801.html

OpenStackやクラウドストレージ技術などに注力中

その他、活動領域

Open Cloud Campus、日本OpenStackユーザ会（JOSUG）、JEUG、
VIOPS InterCloud SIG、GICTF等でも活動中

00 自己紹介(続き)

Disclaimer

私の勤務先では、特に、InfiniBand に関する開発活動(製品、ソリューション等)は行っておりません。純粋に、ユーザの立場にあります。

Then, why me?

その昔(前職の頃)、仕事で InfiniBand 他の高速インタコネクトに関わっていたためです。具体的には・・・

- InfiniBand HCAカードの Linux 用デバイスドライバの開発
- SDPプロトコルの策定
- DAPL (Direct Access Provider Library) ライブラリの開発
- RDMA技術の応用の1つである、ICSC Socket Extension の仕様の策定

等々

01 概要

InfiniBand & Manycore Day シリーズの全体象

InfiniBand 関連　1. 技術編
　　2. 業界動向編
　　3. 応用編

Manycore 関連　調整中？

本日のプレゼンテーションの概要

以下のトピックについて、システムプログラマ視線でみた技術的なポイントをご紹介します。

- RDMA技術の代表格である、InfiniBandの機能概要
- RDMA技術のメリット
- RDMA技術および InfiniBand の応用例（SDP, vSMP）

02 InfiniBandの「機能」概要

Specification

- 最新版スペックのバージョンは1.2.1＋Errata
  - http://members.infinibandta.org/kwspub/spec/
  - 意外にバージョンが上がっていない
    - 機能としては十分成熟したということか。
- 構成
  - Volume 1：全体アーキテクチャ、リンク～トランスポート層、管理系　1727pp
  - Volume 2：物理層 834pp
- Volume 1+2 合計で2500page以上！
  - 物理層（L1）からトランスポート層（L4）、管理系、およびコネクタの規格等まで定められているため、膨大な量となっている

02 InfiniBandの「機能」概要

全体像の理解への近道

- スペックの Volume1、Chapter 3 “Architectural Overview” が一番網羅的でまとまっており、おすすめです。
  - 豊富な図版入りで、55page （ただし、英文）

システムプログラマの視線から見た特徴的な機能

- Addressing
  - パケットに埋め込まれているアドレスには２種類ある
  - サブネット単位でユニークな、LID（Local Identifier） ： 16bit
  - 全世界でユニークな、GID（Global Identifier） ： 128bit
    - GIDには、実は IPv6 が使用されている
    - つまり、InfiniBand “ルータ” という概念がある

02 InfiniBandの「機能」概要

- 豊富なトランスポート層の機能
  - ４種類のトランスポート
    - RC, RD, UC, UD
  - RDMA : Remote DMA
  - Atomic Operation
  - 以上はすべてハード(ファームも含む)で実装されている
- Congestion Control
  - Slow Drain問題の考慮等が入っている
- 管理系
  - 管理用に外付けのネットワークは要求しない。インバンドで管理

InfiniBand HCAの基本的な構造

IB spec. 1.2.1
Vol1. p96より
引用

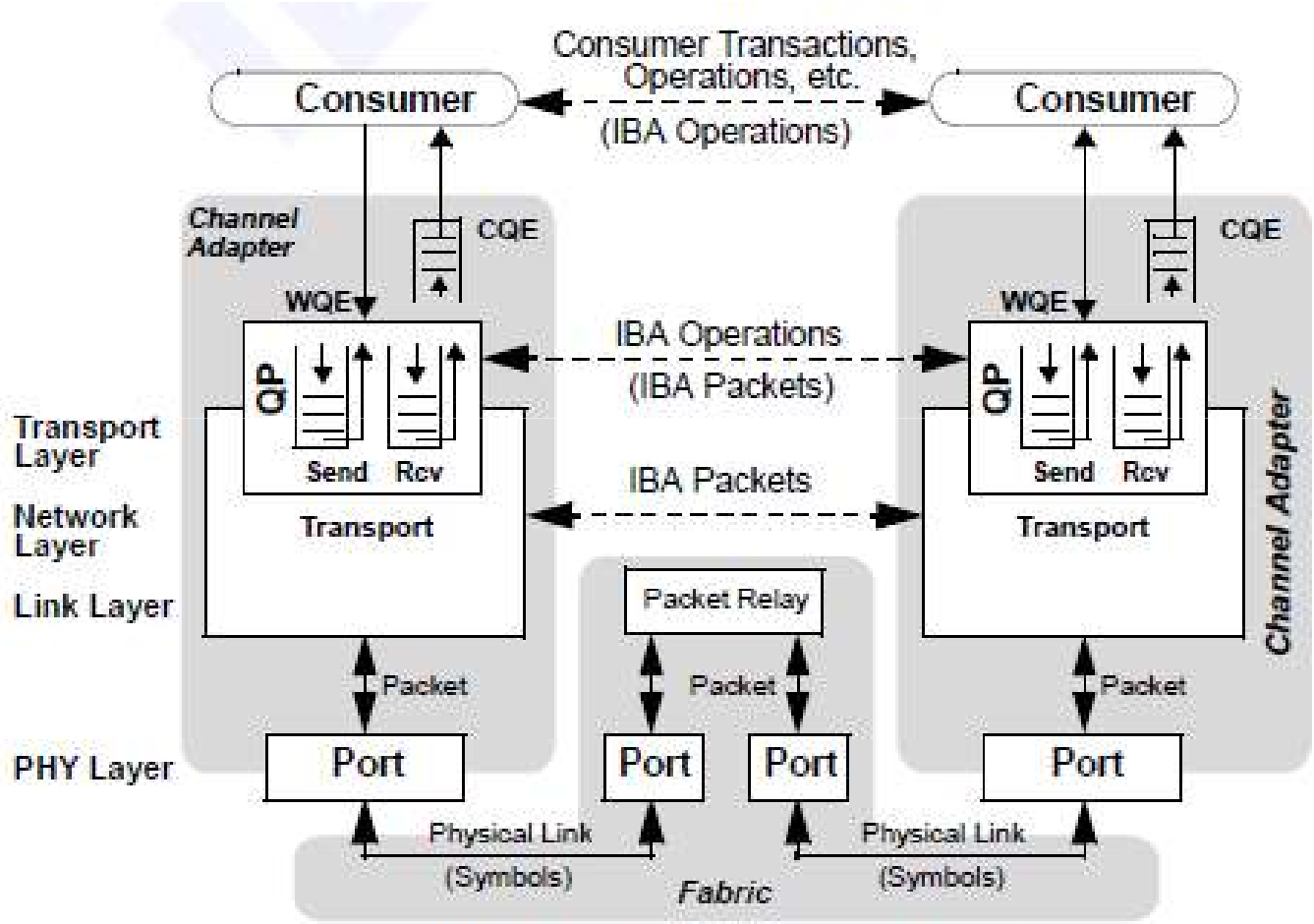


Figure 13 IBA Communication Stack

03 RDMA技術の「機能」概要

HCA	ホストチャネルアダプタの略。要するにNIC
WR	Work Request。要するに通信コマンド
QP	Queue Pair。プログラムが使う、通信エンドポイント 全二重のチャネルの送信側と受信側のコマンド処理キューに対応
CQ	Completion Queue。QPに投入したWR(通信コマンド)のうち、 処理が完了したWRに対応する情報エントリ(CQE)が格納される
PD	Protection Domain。アクセス保護のための単位
MR	Memory Region。HCAに登録された、仮想連続なメモリ領域
MW	Memory Window。MRを分割する単位

02 InfiniBandの「機能」概要

変える力を、ともに生み出す。 NTT DATAグループ

InfiniBandによる通信手順の概要

1. 事前準備
   - HCAとの接続、PD、CQの準備等
2. 通信エンドポイントを作成する(QPをつくる)
3. コネクションを張る(RCを使う場合)
4. 通信バッファをHCAに登録する　　2/3と4は逆でも構わない
5. 通信リクエストWRをQPにポストする
   - Send/Recv、RDMAなど
   - RDMAするリモートアドレスは、Send/Recvでコネクションの両側で通知しあうのが普通
6. CQから通信リクエストの完了を刈り取る

03 RDMA技術のメリット

広帯域通信？低遅延？

- 違います。それは、主にリンク層の性能によるメリット
  - リンク層の性能以外にも、I/Oバスの性能、I/Oブリッジの性能、DMAエンジンのI/Oバスの使い方の善し悪し、そもそも論としてソフトからの使い方の善し悪し等、さまざまな要素が関係してきます。

では何がRDMA技術のメリットなのか？

- コピーを減らせること
- システムバスの負荷が低いこと
  ≒SMPシステム全体の効率を上げられること

03 RDMA技術のメリット

まず、プログラマ視点で見た、RDMAを復習しておきましょう

- リモートノード上の、
- 仮想アドレス(user/kernel spaceを問わない)を指定して、
- 書き込み、読み込みができる
- しかもCPUを介さずにデバイスがやってくれる
- だから、Remote Direct Memory Access(RDMA)

03　RDMA技術のメリット

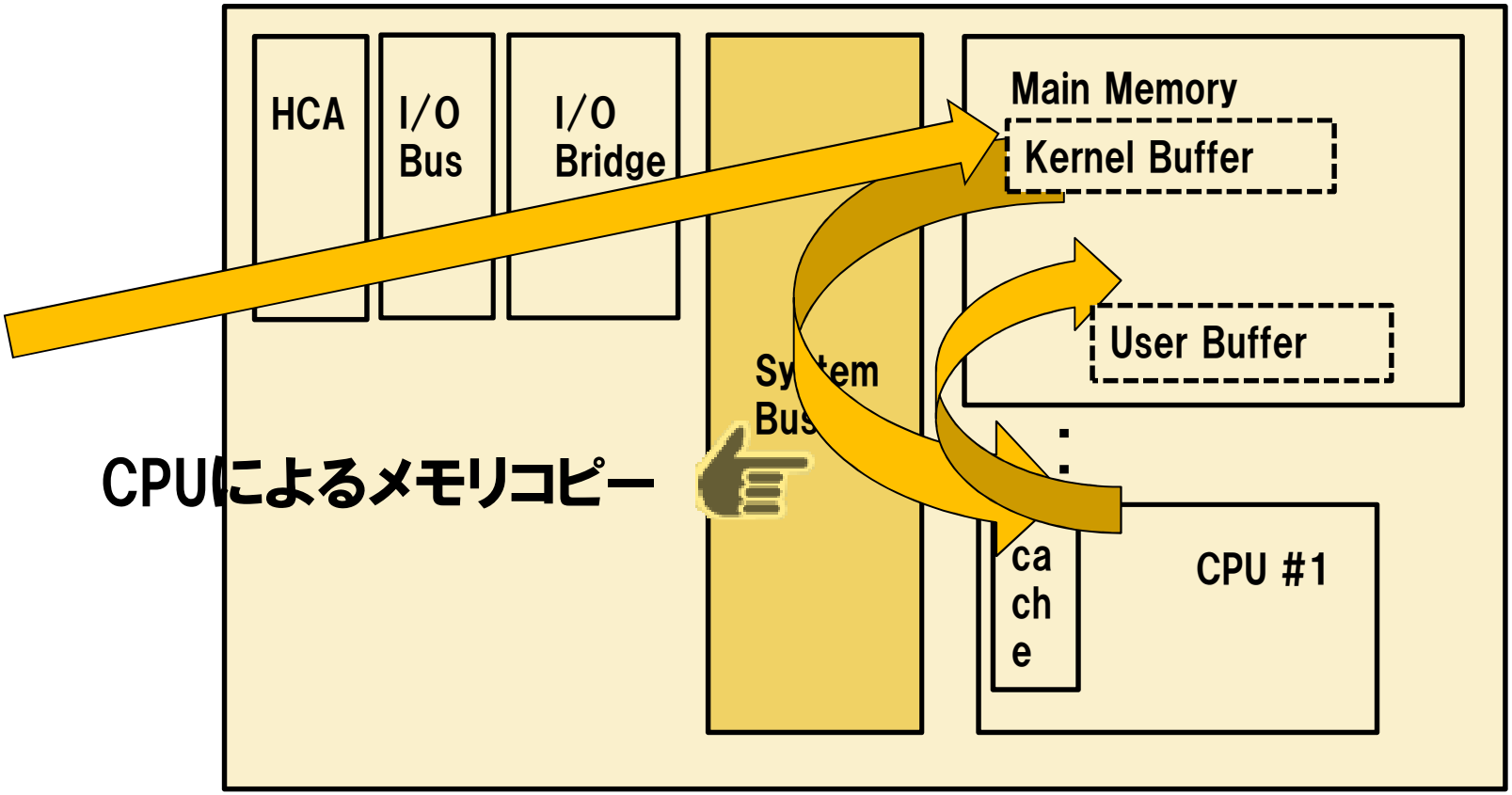
「コピーを減らせる」とはどういうことか？

以下の2つのケースにおける、受信側でのデータの動きを比較してみます

1. 通常の socket 通信のケース
2. RDMAによるユーザレベル通信 (ULT) のケース　0コピー通信？

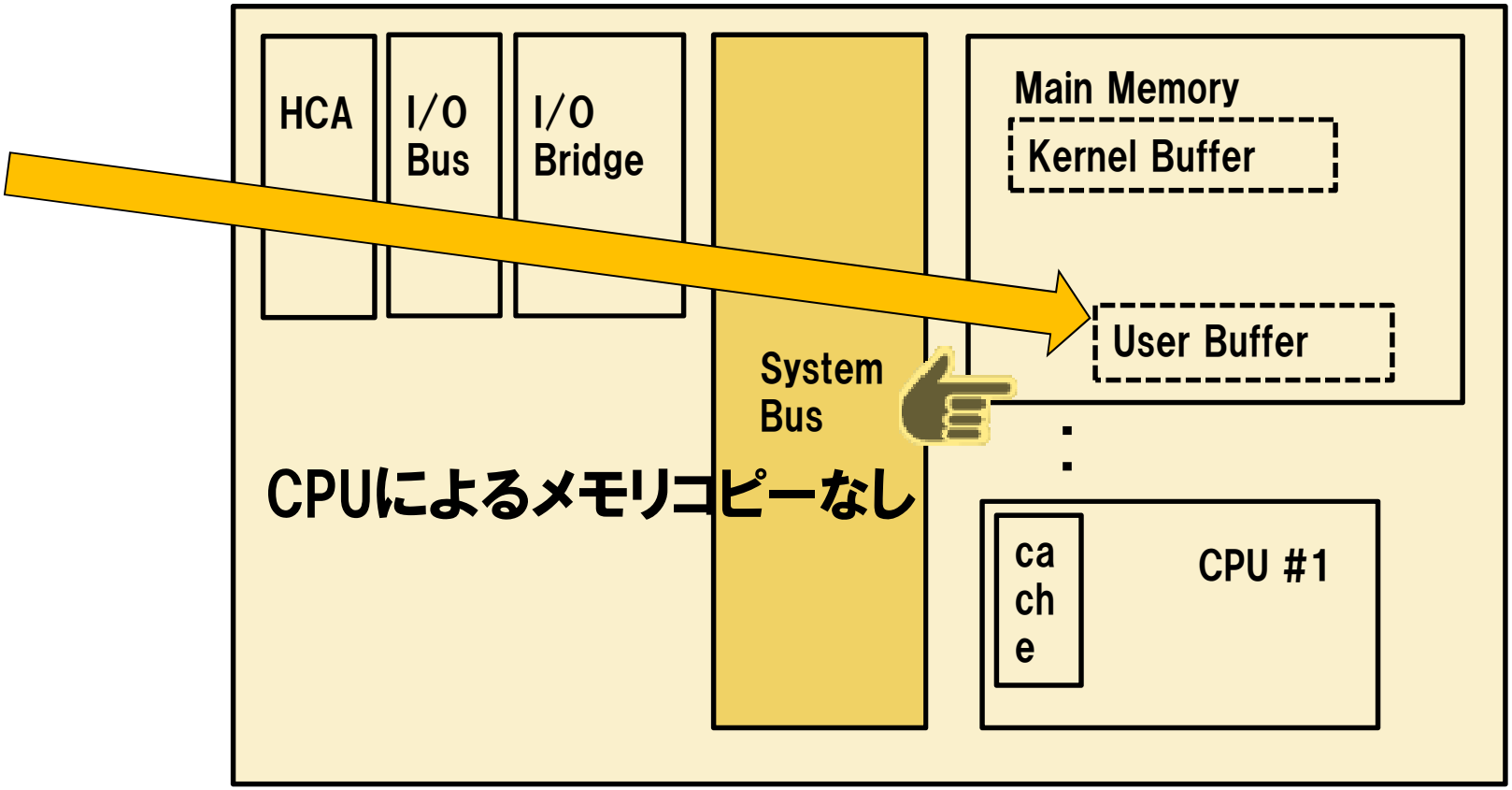
03 RDMA技術のメリット

1. 通常のSocket通信などのケース
送受信バッファを使って Send/Recv する



03 RDMA技術のメリット

2. RDMAによるユーザレベル通信 (ULT) のケース
リモートのユーザバッファへ直接RDMA Write する



03 RDMA技術のメリット

これら2つのケースにおける、受信側でのデータの動きの違い

1. 通常の socket 通信などのケース
   - デバイス -> I/Oバス -> I/Oブリッジ -> システムバス -> メモリ (kernel buffer) -> メモリ(user buffer)
2. RDMAによるユーザレベル通信 (ULT) のケース
   - デバイス -> I/Oバス -> I/Oブリッジ -> システムバス -> メモリ (userbuffer)

実は違いは最後のメモリコピーしかない

- メモリコピーはシステムバスを2回通るのに注意

つまり、RDMAの有無で、システムバス上のトラフィックは(最悪)3倍違う！

- これがSMPシステム全体の効率に効く。

03 RDMA技術のメリット(余談)

I/Oバスに対する DMA READ/WRITE とRDMA Read/Writeの関係

- RDMA Writeソース側のデバイス上のHCAのDMAエンジンからは、主記憶上からのデータの読み出しが必要
  - DMA READ を実行
- RDMA Writeターゲット側のHCAのDMAエンジンからは、主記憶へデータの書き込みが必要
  - DMA WRITE を実行

- つまり、立ち位置によって、READとWRITEが逆転して見えることがあるので混乱しないようにしましょう

04 RDMAの応用例(1):SDP

RDMA技術の問題点の一つ

- ハードの設定を気にするのは面倒だし、慣れたソケット通信でプログラムを書きたい

解決策

- ソケットライブラリのすぐ下の層(ユーザ空間)で、通信処理を横取りし、RDMAを利用した低コスト・高性能通信処理にすげかえてしまえばいい!
- この種のアイデアは、古くは、UCB の FastSockets の研究[1]にさかのぼります

[1] Steven H.Rodrigues, Thomas E.Anderson, and David E.Culler. High-performance local area communication with fast sockets. *USENIX, 1997.*

04 RDMAの応用例(１):SDP

SDPの処理の概要

- TCPの代わりに独自プロトコル(Socket Direct Protocol)を定義。上位プログラムには通常の byte-stream通信を見せる。
- ショートメッセージ、Recv Postingが間に合わない場合はSend/Recvを使用して通信
- ロングメッセージの場合は、プログラムが渡したバッファをメモリ登録してからRDMAを使用してOコピー通信

使い方

通常のsocket通信を行うプログラムの起動時に、LD_PRELOADに libsdp.so を指定するだけ

04 RDMAの応用例(１):SDP

効果

ベンチマークすると、物理層の実効性能の9割以上は達成できます。

例:SDRの測定例で、例えば7.6Gbpsくらい。

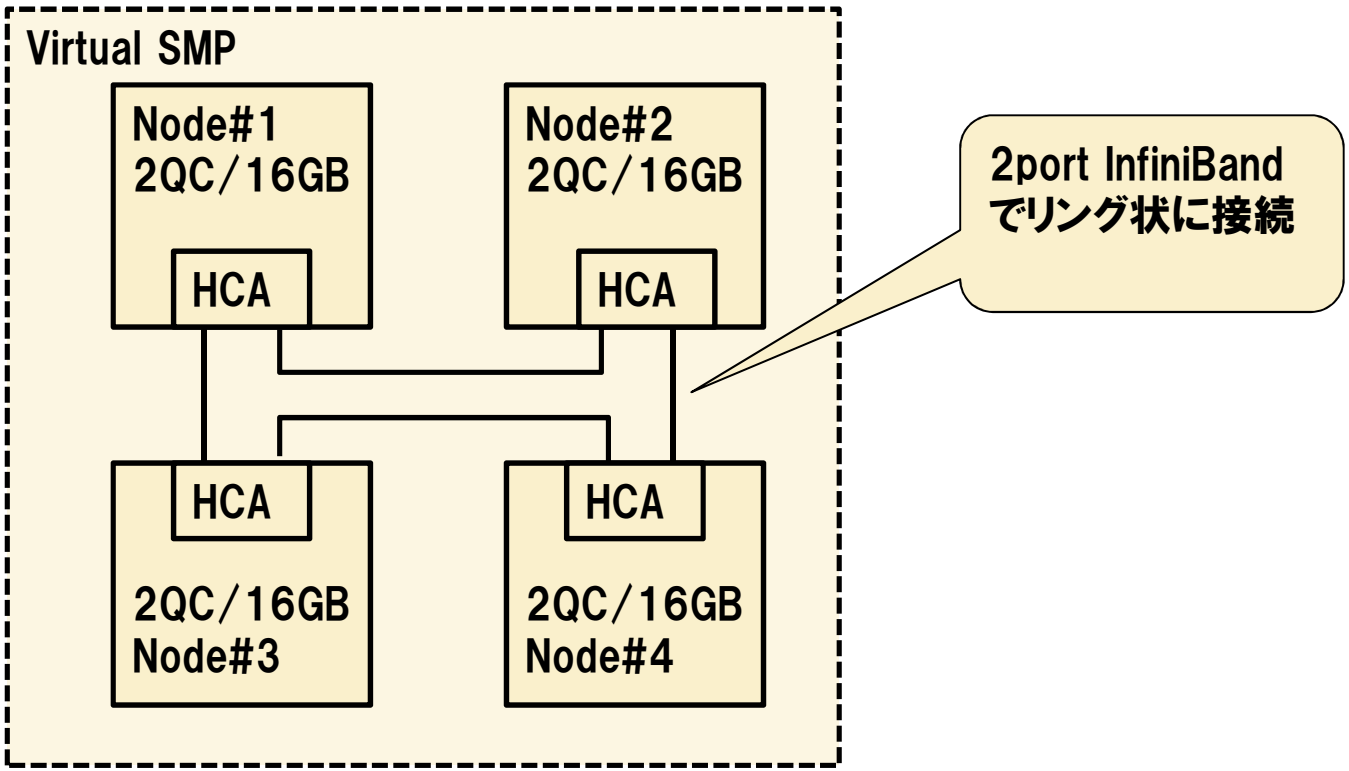
特性(というか、留意事項)

- SDPでは、プログラムにTCPソケットに見えるパス１本について１つ、RCのコネクションをはります。なので、低通信量でコネクション確立・切断を繰り返すような使い方をすると、性能が出ません。
- 送受信バッファを十分大きくする必要があります。でないと、0-copy通信が走らないことがあります。
- RCコネクションはハード資源です。なので、多量のコネクションを張るようなプログラムは性能劣化を起こすことがあります。

05 InfiniBandの応用例(1):vSMP

ScaleMP

- リングトポロジのIBで接続した仮想SMP(vSMP)なるもの
  - 以下の例では 8core、16GB の4台を束ねて、32core、64GB の SSIにできる



05 InfiniBandの応用例(１):vSMP

リングトポロジでvSMPを実現するにあたって気になること

- 鶏と卵問題
  - このNノードなら、 IBサブネットもN個存在する
  - サブネットごとに Subnet managerが必要
  - Subnet manager は SSI になった後に起動するのか？
  - でも、Subnet Manager がいないと、SSIになれないはず
  - では、vSMPってどうやって起動するんだろう？
  - もしかしたら、各ノードごとにOSあげるのか？
    - そのあとでSSI化するなら納得できる

05 InfiniBandの応用例(1):vSMP

考えるべきこと

- spinlock の実装
  - spinlock を獲得しようとしたプログラムが動作している物理ノードの以外のノードも含めて、システムワイドでCAS操作を行う必要がある。InfiniBand の Atomic Operation で実現するのか？
- ある物理ノードで動いているプロセスが、他ノードのメモリにアクセスしたらどう処理するのか？
  - 別ノードにコンテキストを移動させる？
- 別物理ノードに接続されているI/Oアダプタに対してI/O要求が出た場合にどう処理するのか？
- etc.

某所のScaleMP環境で実機検証・調査をしたいと思います ☺

06 今後の予定(？)

- 今後の InfiniBand & Manycore Day でご紹介したいこと
  - InfiniBand以外のRDMA系技術の動向
  - NFS/RDMAの概要とメリット
  - Oracle　RACが使用する RDS : Reliable Datagram Socket の概要とメリット
  - DAPL等、RDMA技術を利用可能な通信ライブラリの概要
  - RDMA技術を活用するためのSocket API拡張

07 まとめ

- システムプログラムからみたRDMAのメリット
  - メモリコピーを減らすことが、通信性能だけでなく、システム性能全体に効く（意外と認知されていない）

- 今後も、RDMA技術の応用例やメリットなど、ご紹介していきたいと思います
  - 一緒に実機評価作業をしてくれるボランティア求む！ ☺

ご清聴ありがとうございました

変える力を、ともに生み出す。

NTT DATAグループ

Q&A

References

- InfiniBandの規格
  - http://www.infinibandta.org/
  - メンバ登録(無料)をしなくてもスペックは以下から入手できるようです。
    - http://members.infinibandta.org/kwspub/spec/
- OpenFabrics
  - http://www.openfabrics.org/
- InterConnect Software Consortium
  - http://www.opengroup.org/icsc/