TRANSFORMING
COMMUNICATIONS

インテル® データ プレーン デベロップメント キット (インテル® DPDK)

インテル® アーキテクチャー上でのパケット処理

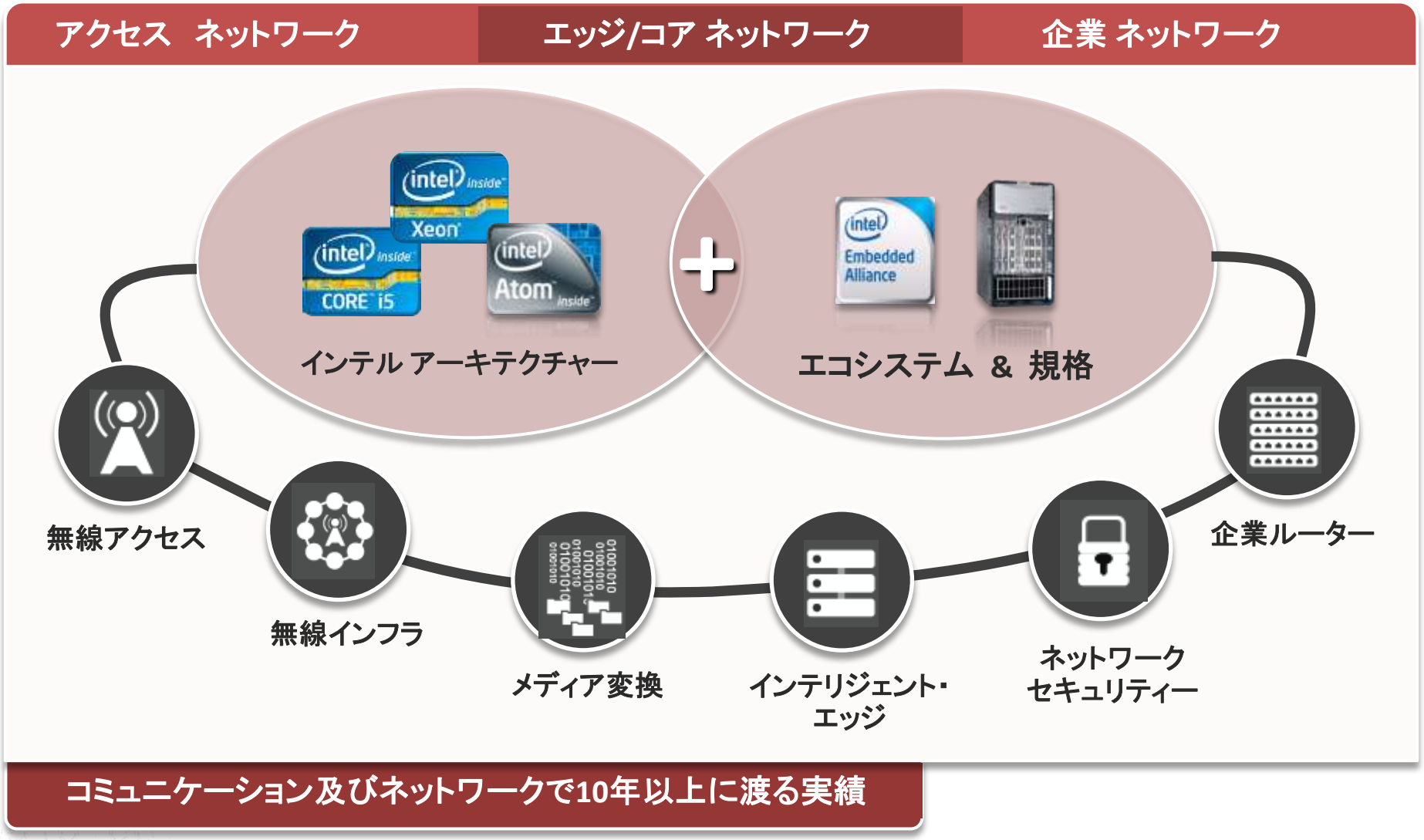
2012年11月20日

インテル株式会社
クラウド・コンピューティング事業本部
インテリジェント・システムズ・グループ
事業開発マネージャー
幸村　裕子

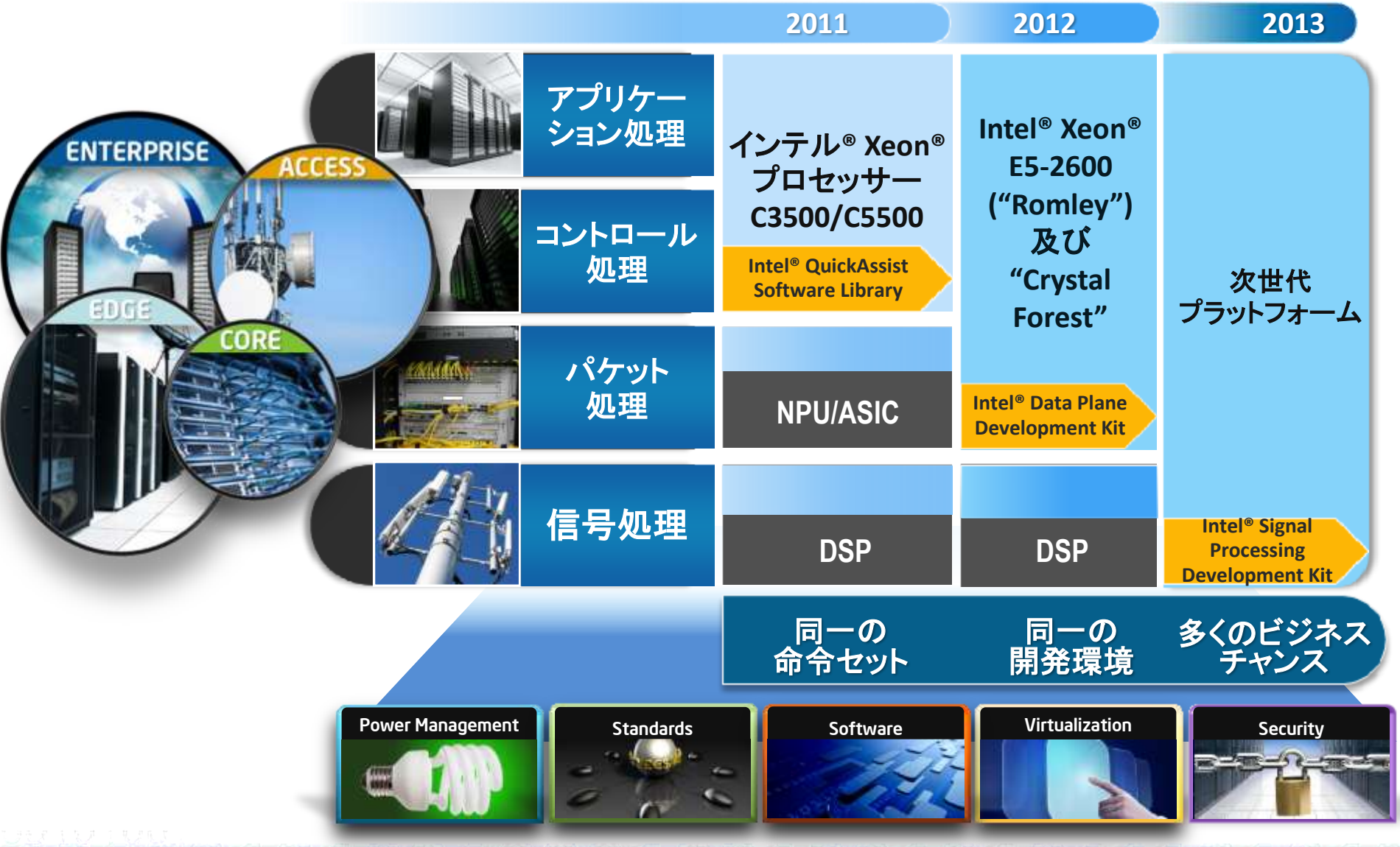
本日の内容

- 通信分野におけるインテルの活動
- 4:1 - ワークロードの統合
- パケット処理における課題
- インテル® データ・プレーン・デベロップメント・キット (インテル® DPDK)
- インテル® DPDKとインテル仮想ファンクション ドライバー (SRIOV)
- 参考資料

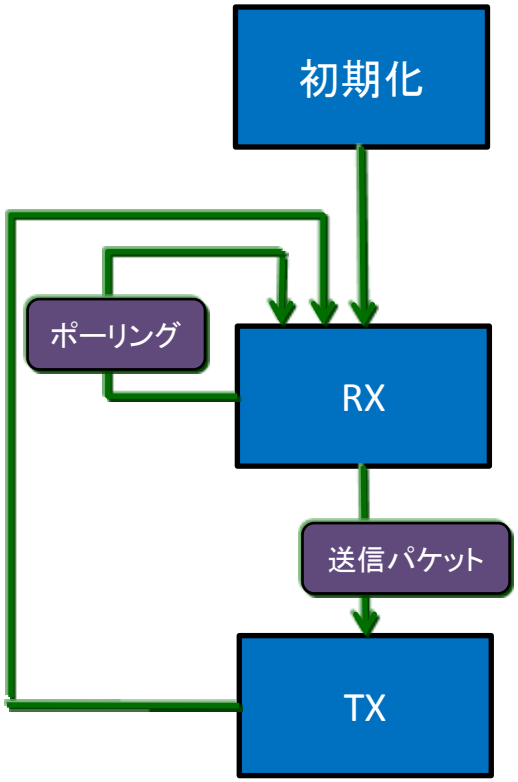
世界のコミュニケーションをiAで



4:1 ネットワーク・ワークロードの統合



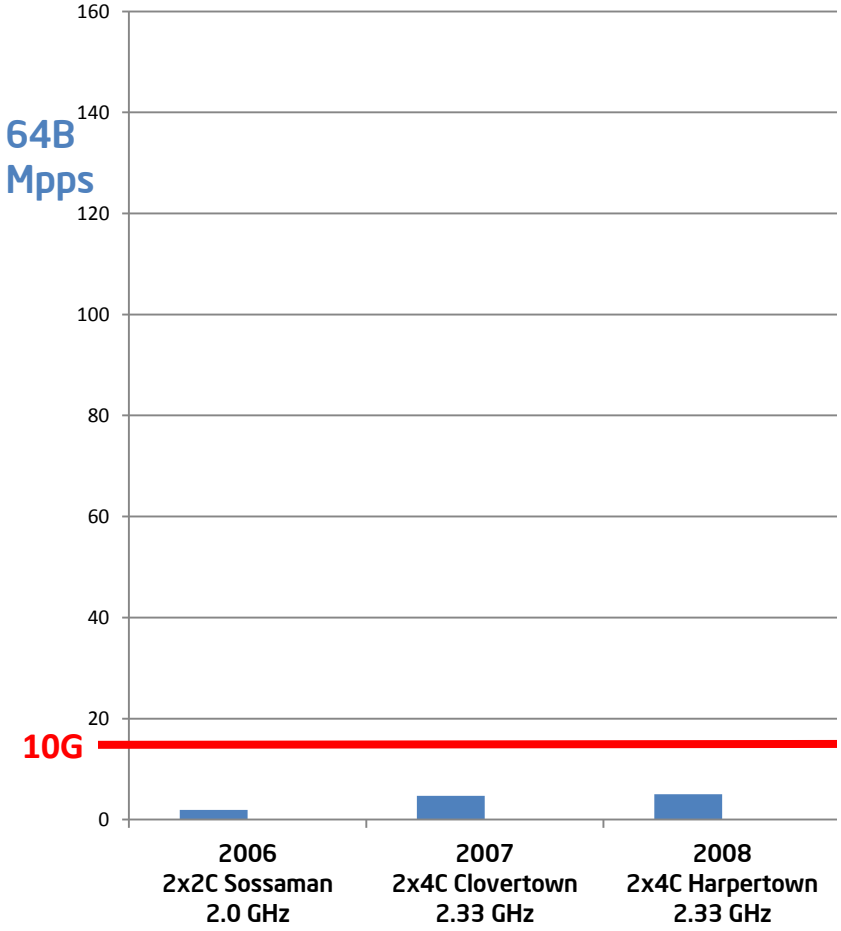
30,000 フィート上空から見たパケットの流れ



1. **初期化**
   - メモリの初期化(ゾーン・およびプール)
   - デバイスおよびデバイスのキューの初期化
   - パケットフォワーディングアプリケーションの開始
2. **パケット受信 (RX)**
   - 機器のRXキューをポーリングしてパケットをバーストで受け取る
   - RXバッファを新たにキューごとにメモリープールから割り当て、デスクリプタに保存
3. **パケット送信 (TX)**
   - RXで受け取ったパケットを送信
   - パケットを保持していたメモリーを開放

IA パフォーマンスの年々の向上

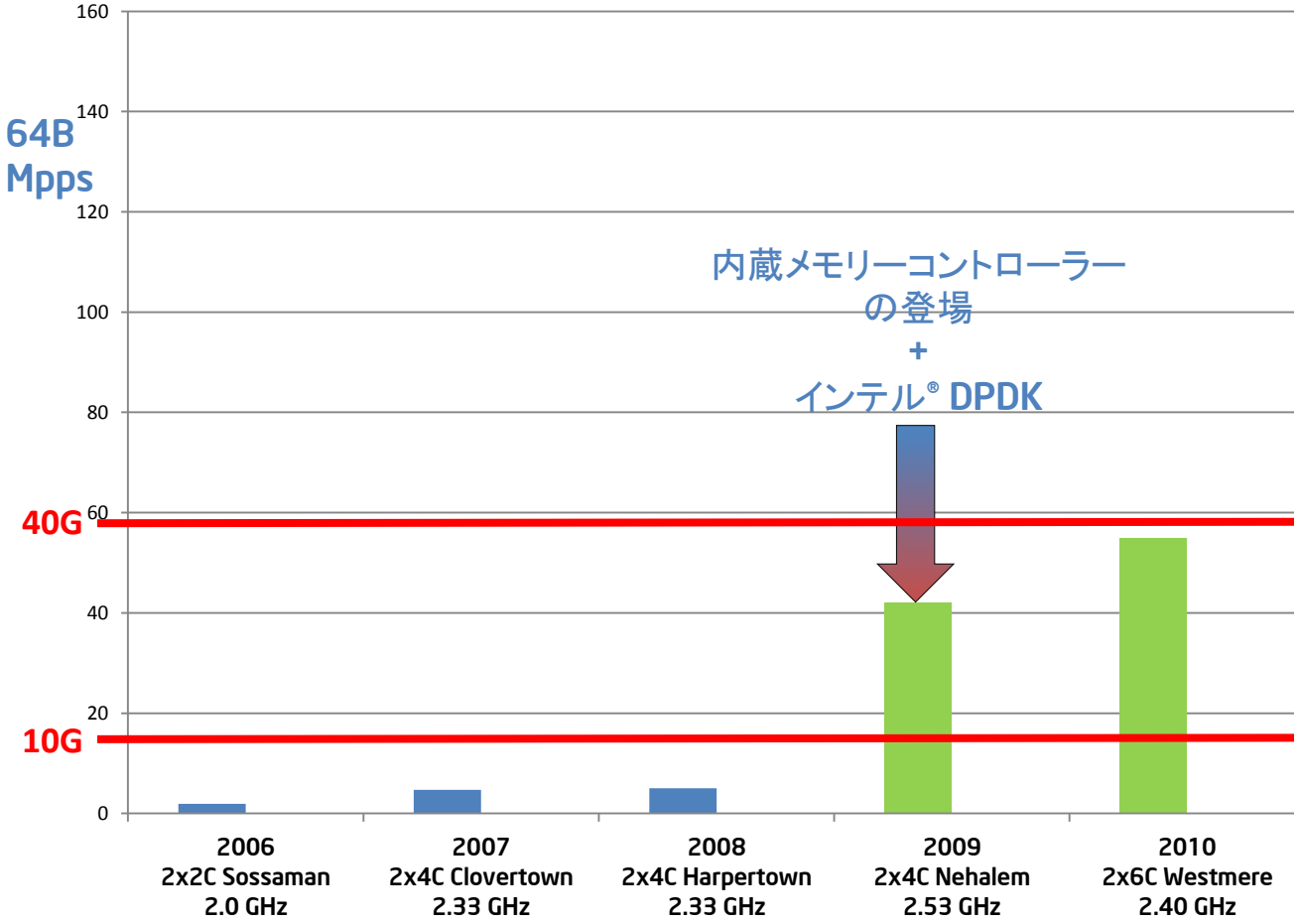
iAでのPv4 レイヤー3 フォワード性能



今までのソフトウエアでのアプローチではインテル　アーキテクチャーは通信のパケット処理で他のアーキテクチャーに見劣りする

IA パフォーマンスの年々の向上

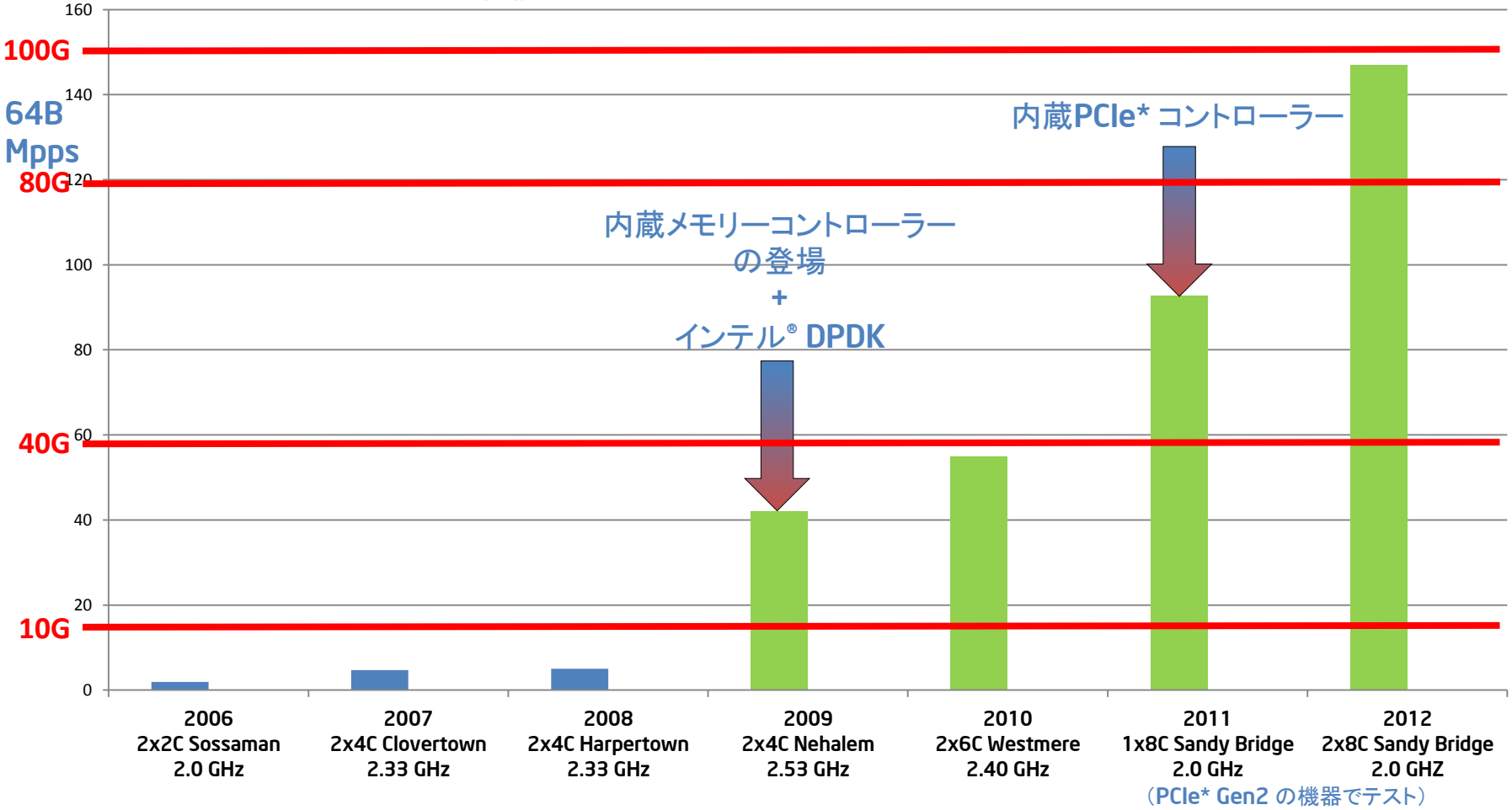
iAでのPv4 レイヤー3 フォワード性能



CPUコアとメモリアーキテクチャーの革新、およびインテル　DPDKの登場で処理能力は飛躍的に向上

IA パフォーマンスの年々の向上

iAでのPv4 レイヤー3 フォワード性能



標準の“off-the-shelf” IAプラットフォームで魅力的な性能を提供します

80Mpps達成へのハードル

メモリおよび PCIe のアクセスがCPUの動作に比べると非常に遅い

複数パケットを、それぞれのソフトウエア処理の間に処理し、メモリやPCIeアクセスにかかる時間を短縮

データが必要なときにCPUから見て遠くに位置し、CPUが待たなくてはならない

ハードウエアおよびソフトウエアによるプリ・フェッチを行う。PCIeではDDIOを使ってデータを直接キャッシュメモリに読み込む

共有のデータストラクチャーへのアクセスがアプリケーションのボトルネックとなる

共有の度合いを減らすスマートなアクセス方法を考案 (例. メッセージの受け渡しにロックレスなキューを使う)

システムが、RXパケットによる割り込みの量についていけない

割り込み処理のネットワークデバイスドライバーをポーリングモードのドライバに変更

Linuxのもともと持っているスケジューラーのタスクスイッチにかかるオーバーヘッドが大きい

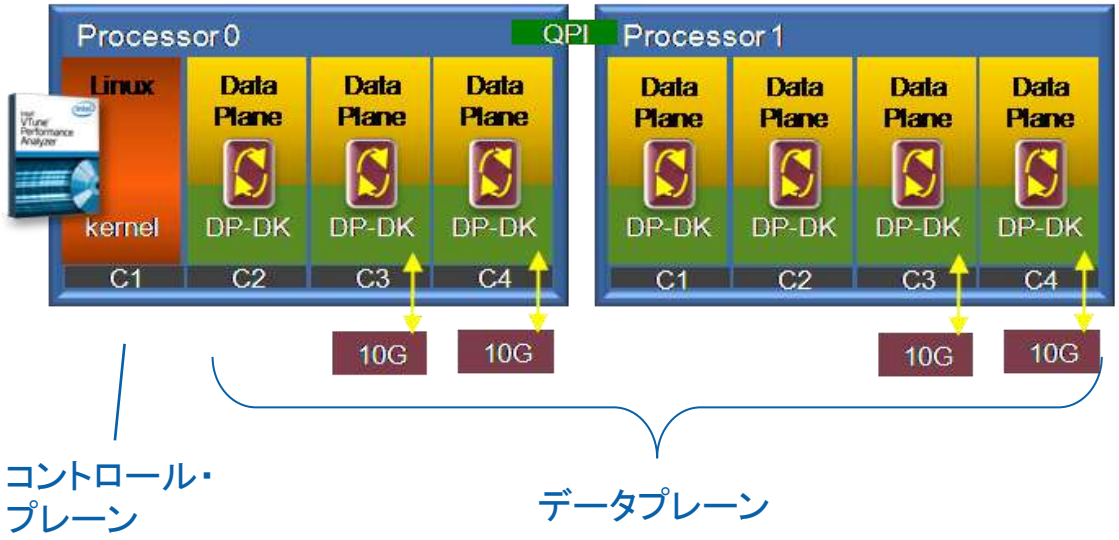
論理コアひとつに対してひとつのソフトウエアスレッドを割り当てる

PTU によるとページテーブルは常に書き換えられている(DTLB のミス-アップデート)

Linuxが大きなページを使えるようにする (2MB, 1GB)

これらのハードルを、ハードウエアでのアシストおよびプログラミングの効率化で解決!!

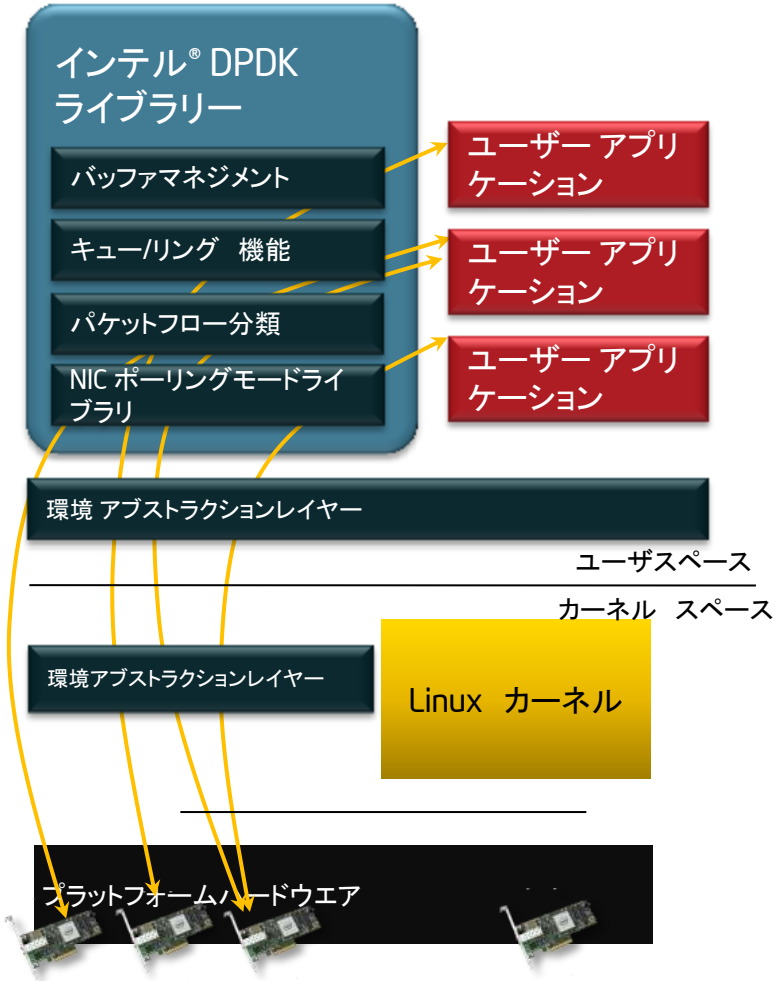
インテル® DPDK 理念



- **すべてのIA CPU上で動作**
  - インテル® Atom™ プロセッサーから最新の インテル® Xeon® プロセッサー・ファミリーまで
  - iAの価値提供手段
- **ファスト・パスに特化をした高速化**
  - 多くのパケットをLinuxカーネル/GPOSに送るとシステム全体の性能低下を起こす
- **ネットワークでパフォーマンス低下が懸念されるポイントの参考ソフトウエアコードを提供**
  - 最適なソフトウエアアーキテクチャの例示
  - データ構造およびデータ保管のティップス
  - コンパイラが最適なコードを生成する手助け
  - 80 Mppsへのハードルを解消

インテル® データプレーン デベロップメント キット

- **データプレーンライブラリーおよび最適化されたNICドライバ**
  - キューと バッファの管理、パケットフロー分類、　ポーリングモードのNICドライバー等
  - シンプルなAPI でのインターフェース, 標準のツールチェーンのサポート(gcc/icc, gdb, プロファイリング・ツール)
- **ランタイム環境**
  - オーバーヘッドの少ない 最も高速なデータプレーン処理に最適化されたrun-to-completion モデル
- **環境のアブストラクションレイヤーおよびブートコード**
  - プラットフォームに特化したブート・ガイドラインと初期化コード
- **BSD-ライセンス且つソースコードもインテルよりダウンロード可能**
  - ユーザーアプリケーションに単独でも組み込み可能なソリューションですが業界をリードするエコパートナー の商用データプレーンソリューションの一部としてもご利用いただけます



インテル® DPDK はユーザーおよび広く業界に柔軟な BSD のライセンス・モデルに基づいて無料で提供され、スタートポイントとしてお使いいただけます

インテル® DPDK ライブラリ 及びドライバ

メモリ・マネージャー: “Object Pool” にメモリを割り当てる役割をもつ。“Pool”は大きなページもメモリスペースとして作成され、リング構造で“Free Object”を保管する。また“Object”がDRAMのチャネルに均等に分散されるようにパディングを行う“アライメント ヘルパー”機能も提供する。

バッファ・マネージャー: OSがバッファをアロケートしたりアロケートを取り消したりすることに費やす時間を大幅に減らす。インテル® DPDKはメモリプール内に一定サイズのバッファを事前にアロケートする。

キュー・マネージャー: “spinlock”を使用する代わりにロックしない安全なキューを実現し、不要な待ち時間無く別個のソフトウエア・コンポーネントがパケット処理することを可能にする。

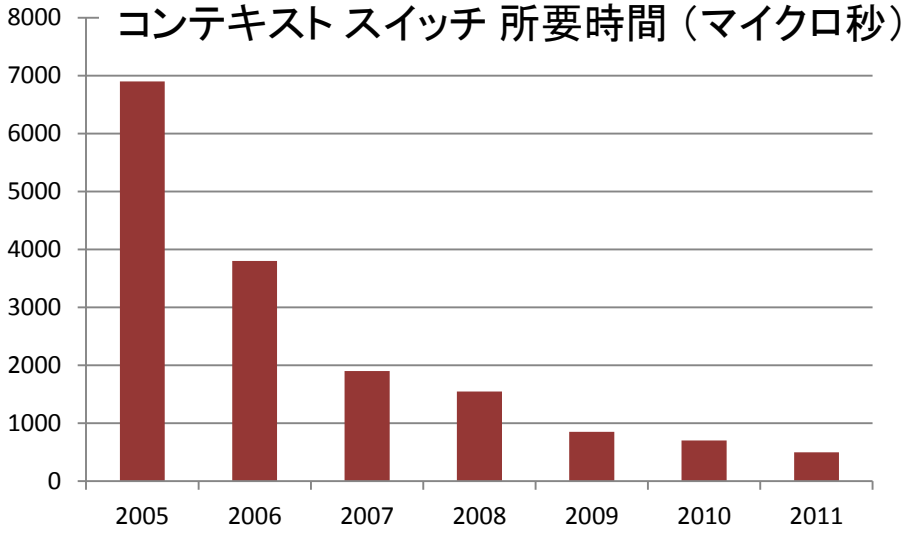
フロー分類: インテル® ストリーミング SIMD 拡張命令 (インテル® SSE) を用いて効率的にtuple情報のハッシュを作成し、パケットがフローにすばやくまわされ、処理されるようにして大幅にスループットを改善する。

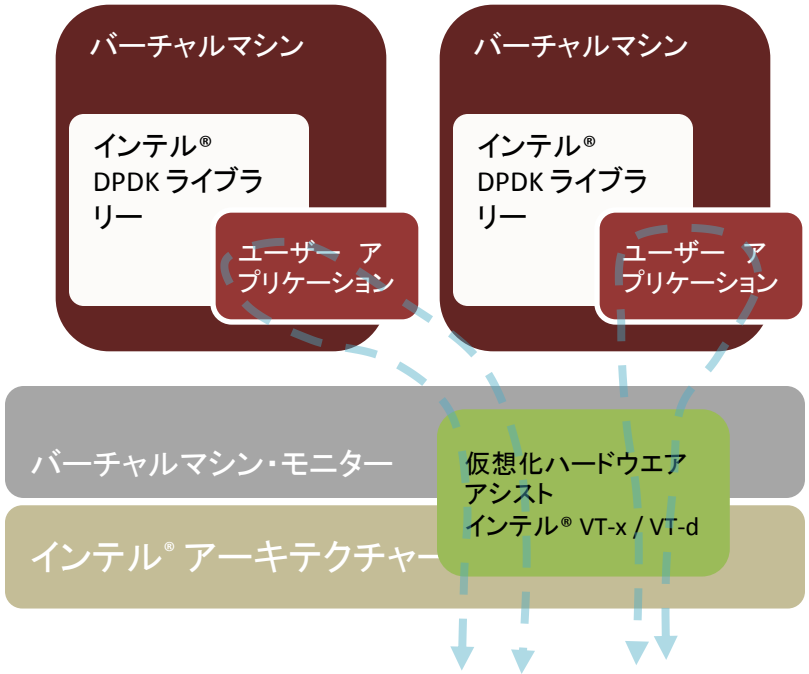
ポール・モード・ドライバ: インテル® DPDK は1 GbE 及び10 GbEのイーサネット* コントローラーのポール・モード・ドライバを持っている。非同期な割り込みベースの通知方式をなくすことでパケットのパイプライン処理を大幅に高速化する。

インテル® DPDK 仮想化-インテル® VT ハードウエアアシスト

仮想化のハードウエアアシスト:

- *拡張ページテーブル*
- *アドレス・トランスレーション・サービス*
- *仮想プロセッサ ID*
- *デスクリプタ・テーブル Exit*
- *VMX プリエンプション タイマ*
- *ループ停止 Exit*
- *割り込みのリマップ*
- *キュード インバリデーション*
- *インテル® VT-d ラージテーブルのサポート*





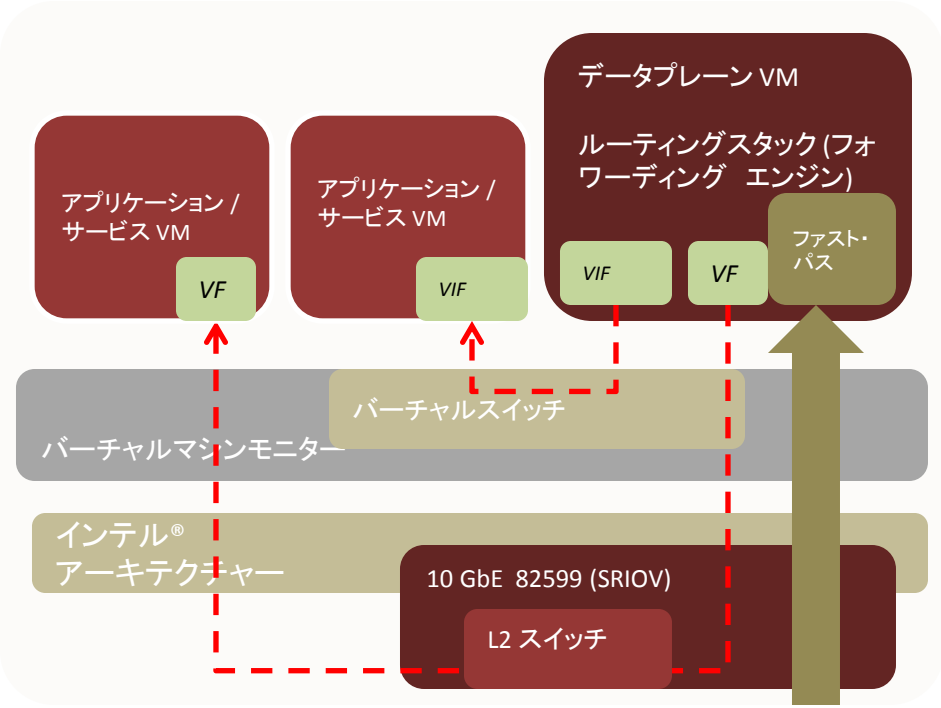
データプレーン・アプリケーションの変更無しに仮想化

バーチャルマシン間の通信

ゲスト間の通信を行う際の選択肢:

- *ハイパーバイザのバーチャルスイッチを介して通信する*
- *82599 / 82756のL2 スイッチを利用する*

上記の選択肢は2者択一ではない

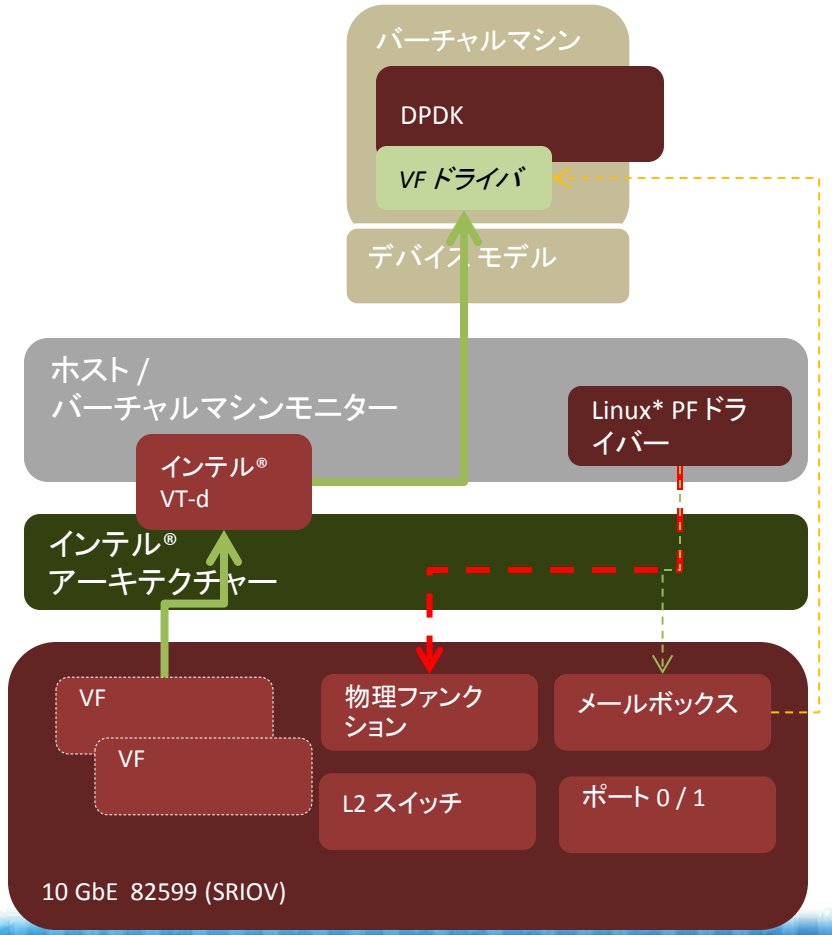


インテル® 82599 10 Gb イーサネット 仮想ファンクションドライバー

- 物理 ファンクションのは伝統的なドライバの機能と同じ機能を担当 (リンク・セットアップ)
- 物理ファンクションがVFの仮想MACを作成しVFとの通信をメールボックスを通じて管理する:
  - VLAN のVM 設定
  - マルチキャストアドレスの設定
  - VF リセットの要求

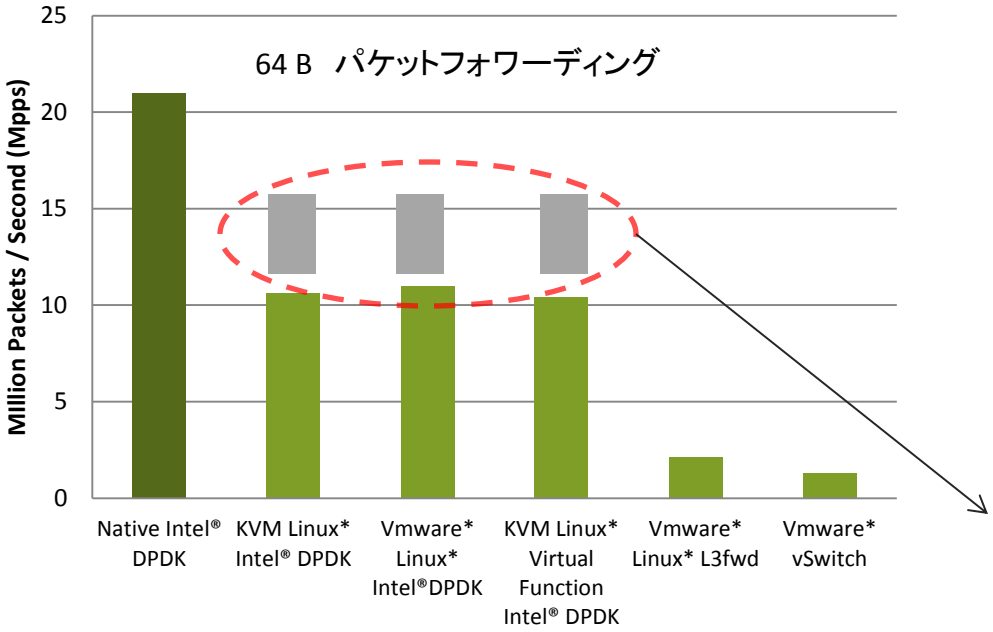
```
root@dpdkguest-desktop:/root/phymem/phymem# lspci
00:00.0 Host bridge: Intel Corporation 440FX - 82441FX PMC [Natoma] (rev 02)
00:01.0 ISA bridge: Intel Corporation 82371SB PIIX3 ISA [Natoma/Triton II]
00:01.1 IDE interface: Intel Corporation 82371SB PIIX3 IDE [Natoma/Triton II]
00:01.3 Bridge: Intel Corporation 82371AB/EB/MB PIIX4 ACPI (rev 03)
00:02.0 VGA compatible controller: Cirrus Logic GD 5446
00:03.0 Ethernet controller: Realtek Semiconductor Co., Ltd. RTL-8139/8139C/8139C+ (r
ev 20)
00:04.0 Ethernet controller: Intel Corporation 82559 Ethernet Controller Virtual Func
tion (rev 01)
root@dpdkguest-desktop:/root/phymem/phymem#
```

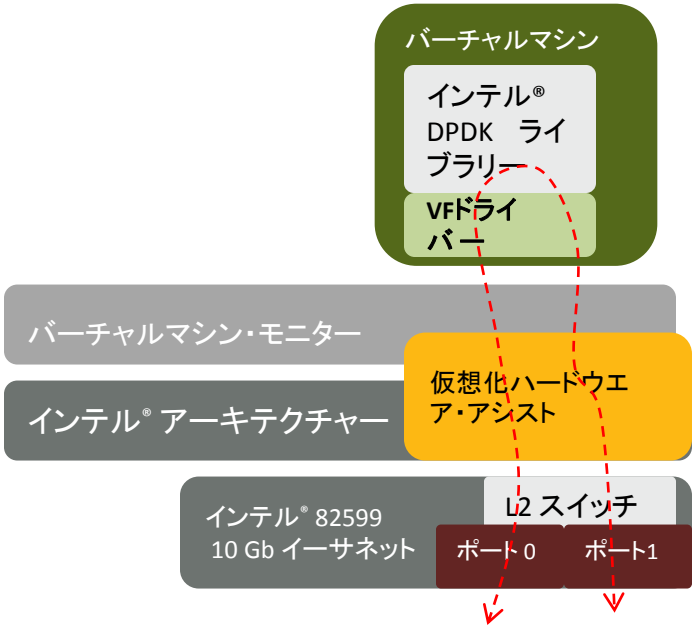
- MACアドレスもしくはVLANタグに基づきレイヤー2 スイッチに Classifier / Sorter を組み込む
- Pool と Pool (VF と VF) の橋渡しをサポート
- Anti-Spoofingのサポート



インテル® DPDK 仮想化環境での性能

レイヤー 2　スイッチを用いた場合のインテル® 82599 10 Gb イーサネット・コントローラの仮想ファンクションドライバーの性能





仮想環境上での性能は“仮想化フレンドリーな”プログラミング方法の指針に沿うことでスモールパケットのスループットを更に高めることも可能

インテル® DPDK VFはインテル® DPDK PMD(ポーリングモードドライバー）性能と一致する

インテル® DPDK 公開ウエブサイト

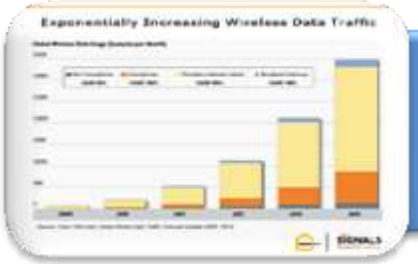
www.intel.com/go/dpdk (英語)
http://www.intel.com/p/ja_JP/embedded/hwsw/technology/packet-processing (日本語)

- 資料や記事、ホワイトペーパーやビデオなど
- 協業各社様の情報および記事
- サポートフォーラム

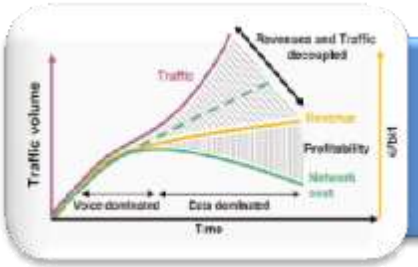
まもなく・・・

- インテル® DPDK を公開します
- ソースコードもご提供

まとめ

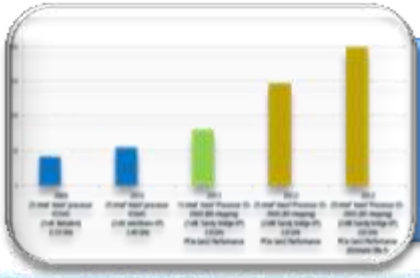


データ/通信 の融合により必要とされるデータ・プレーンの処理能力は指数関数的に増えています



TCO への関心が高まり、R&Dの効率化とすばやい市場投入がもとれまれています

インテルの 4:1 単一アーキテクチャーによる複数の負荷の処理は、半導体プロセス革新とCPUのマイクロ・アーキテクチャーの革新がご享受いただけます



データプレーンのソフトウエアの最適化によりインテル・アーキテクチャー・プラットフォームの能力を最大限に発揮いたします

参考資料

- **通信・ネットワークでのインテル製品について：メインページ**
  http://www.intel.com/content/www/us/en/embedded-developers-engineers/communications-overview.html

- **通信インフラに使われる Intel 製品**
  www.intel.com/go/commsinfrastructure

- **Crystal Forest プラットフォームによるワークロードの統合**
  www.intel.com/go/commsplatform

ビデオ:

4:1 ワークロードの統合
: http://www.intel.com/content/www/us/en/communications/4-to-1-communications-workload-consolidation.html

インターネットの1分:
http://edc.intel.com/Video-Player.aspx?id=5599

インテル® データ プレーン デベロップメント キット:
http://edc.intel.com/Video-Player.aspx?id=5378

サービスエッジ:
http://www.intel.com/content/www/us/en/communications/intel-service-edge-animation.html

法務情報

この文書は現状のまま提供され、いかなる保証もいたしません。この保証には、商品適格性、他者の権利の非侵害性、特定目的への適合性、また、あらゆる提案書、仕様書、見本から生じる保証を含みますが、これらに限定されるものではありません。インテルはこの仕様の情報の使用に関する財産権の侵害を含む、いかなる責任も負いません。また、明示されているか否かにかかわらず、また禁反言によるとよらずにかかわらず、いかなる知的財産権のライセンスも許諾するものではありません。

本資料に掲載されている情報は、インテル製品の概要説明を目的としたものです。本資料は、明示されているか否かにかかわらず、また禁反言によるとよらずにかかわらず、いかなる知的財産権のライセンスを許諾するためのものではありません。製品に付属の売買契約書『Intel's Terms and Conditions of Sale』に規定されている場合を除き、インテルはいかなる責を負うものではなく、またインテル製品の販売や使用に関する明示または黙示の保証（特定目的への適合性、商品性に関する保証、第三者の特許権、著作権、その他、知的所有権を侵害していないことへの保証を含む）をするものではありません。インテルの製品は、医療、救命、延命措置などの目的への使用を前提としたものではありません。

性能に関するテストや評価は、特定のコンピューター・システム、コンポーネント、またはそれらを組み合わせて行ったものであり、このテストによるインテル製品の性能の概算の値を表しているものです。システム・ハードウェア、ソフトウェアの設計、構成などの違いにより、実際の性能は掲載された性能テストや評価とは異なる場合があります。システムやコンポーネントの購入を検討される場合は、ほかの情報も参考にして、パフォーマンスを総合的に評価することをお勧めします。インテル製品の性能評価についてさらに詳しい情報をお知りになりたい場合は、http://www.intel.co.jp/jp/performance/resources/benchmark_limitations.htm を参照してください。

パフォーマンスの推定値は変更される場合があります。

インテル製品は、予告なく仕様や説明が変更される場合があります。

機能または命令の一覧で「留保」または「未定義」と記されているものがありますが、その「機能が存在しない」あるいは「性質が留保付である」という状態を設計の前提にしないでください。これらの項目は、インテルが将来のために留保しているものです。インテルが将来これらの項目を定義したことにより、衝突が生じたり互換性が失われたりしても、インテルは一切責任を負いません。

.本資料に記載されているすべての製品、数値、日付は、現在の予想に基づくものであり、予告なしに変更する場合があります。

Intel、インテル、Intel ロゴ、Intel Inside、Intel Inside ロゴ、Intel Core、Core Inside、Intel vPro、Intel vPro ロゴ、Pentium 、Intel Atom、Ultrabookはアメリカ合衆国およびその他の国における Intel Corporation またはその子会社の商標または登録商標です。

インテルの商標を外部向けに使用する際は、インテルからの許諾が必要です。インテル製品の広告およびプロモーションにおいてインテルの商標を使用する際は、商標に関する適切な脚注が必要です。

* その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。