解説 4

ネットワーク境界のワークロード処理へのチャレンジ

ー The IBM Power Edge of Network™ processor ー

2010 年 2 月、IBM は飛躍的に増大するネットワーク境界のワークロード処理に対応する新しいタイプのプロセッサー・チップを発表しました。新しいプロセッサーはネットワーク境界で発生するワークロードを詳細に分析し、システム・アーキテクチャー、回路駆動方式、細部にわたる最適なトランジスタの選択など、総合的なアプローチで消費電力当たりの処理能力の最適化が図られ、チップ・レベルで最適なソリューションを実現します。このチップは、スタンダードで定められ、アルゴリズムの確立しているワークロードに対しては、ハードウェア・アクセラレーターを実装していますが、それはワークロードに最適化されたシステム設計という IBM サーバー製品共通のテーマに対する戦略的な取り組みが、チップ・レベルでも実践されているということを意味しています。本稿では、この新しいプロセッサーがどのような技術で開発され、どのような価値を提供するのかをご紹介します。

Article 4

A New Challenge: Handling Edge of Network Workloads

- The IBM Power Edge of Network processor -

IBM announced in February, 2010 a new type of processor for Edge of Network (EON) workloads, which have been recently increasing dramatically. Based on a deep analysis of EON workloads, a wide range of optimizations have been implemented, including of chip-level system architecture, together with the use of fine grained clock gating, the careful selection of threshold voltages for each transistor and other approaches. Accelerators implementation for the workloads, those were defined by standards and have established algorithms, is a kind of hybrid system. The workload optimized systems concept is a strategic approach for the IBM systems, the concept is applied to this processor chip also.

In this article, we introduce a strategic processor for the edge of network workloads, and explain how it can be applied in a wide range of next generation network products.

❶ はじめに

近年、光ファイバー網の整備や、3G、WiMAX などのモバイル・ネットワーク・インフラの進化により、ネットワークの広帯域化、低価格化が進んでいます。その結果 Web サーバーを介して提供されるコンテンツも、テキスト・ベースの情報量の小さいものから、写真、動画、音楽といった情報量の多いものへ広がり、インターネット電話やテレビ、オンライン・ゲームといった広帯域通信を前提とした新しいサービスも急速に普及し、データ通信量は飛躍的に増大しています。

同時に IPSec（IP Security Protocol）、ウイルス除去のためのパケット・フィルタリングのような、ネットワークに対するセキュリティー対策への要求や、通信帯域の効果的利用のためのデータ圧縮といったデータ処理が IP パケットに求められるようになり、広帯域ルーターやファイアウォール、ネットワーク進入検知・防御システムといった、ネットワークの境界に配備される製品群のネットワーク・パケット処理のワークロードも爆発的に増大しています。

また、あらゆるモノがネットワークに接続され、それらから発信される膨大な情報をリアルタイムに解析し、次のアクションに結び付け、スマートな社会を実現するためには、ネットワーク上の適切な場所で、情報の加工・統合を行わないと、あふれ出る情報を吸収することができなく

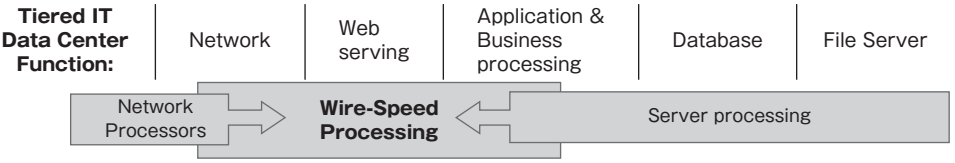

図1. ワイヤースピード・プロセッサーの位置付け

なってしまいます。

飛躍的に増大するワークロードを効果的に処理するため、ネットワークの境界ではネットワークとサーバーの融合が進みつつあり、従来ネットワーク製品群が処理していたワークロードと、サーバーで処理していたワークロードの両方を合わせて処理する製品群が立ち上がりつつあります。（**図 1**）

ネットワーク境界でのワークロードは、バックエンド・サーバーのようにネットワークのエンド・ポイントとしてデータを最終処理するものではなく、次々と流れてくるパケットのフィルタリング、圧縮伸張、暗号化、データの統合といった処理を行い、それを転送するインライン処理が主なものです。

ここでのワークロード処理に求められるキーワードは、数十ギガビット／秒という通信速度で流れてくる大量のパケット・データを、ネットワークが保有している回線速度（スループット）を落とすことなく、かつ処理時間（レイレンシー）を最小限に、処理する“ワイヤースピード・プロセッシング”です。

ネットワークに関連するワークロードの特質は、小規模スレッドの大量処理であり、この分野において、汎用のサーバー向けプロセッサーは、同時処理可能なスレッド数がボトルネックとなるため適切なものではありません。また従来のネットワーク・プロセッサーも、組み込みプロセッサー・ベースの同時処理スレッド数が比較的小さなものなので、広帯域でかつ上位のサービス・レイヤーまでをカバーするアプリケーションには向きません。

ワイヤースピード・プロセッサーは、この領域をカバーするもので、IBM では現在 The IBM Power Edge of Network processor（PowerEN™）を開発中です。

概略について、2010 年 2 月 ISSCC（International Solid-State Circuits Conference）で発表済みですので、その内容をご紹介します［1］。

本チップは非常に大規模なもので、開発は日本を始め世界9カ国以上の IBM の研究所が密に連携し、最新の開発環境を駆使して進められています。

なお、今回ご紹介する内容はあくまで一次試作のものであり、今後開発が進むにつれて変更される可能性があります。

❷ The IBM Power Edge of Network processor (PowerEN)

IBM で現在開発中のワイヤースピード・プロセッサー（PowerEN）は、パケット処理などネットワークの下層レイヤーを受け持つネットワーク・プロセッサーの属性と、アプリケーションなどの上位レイヤーを受け持つサーバー・プロセッサーの属性を融合した新しいタイプのプロセッサー LSI（大規模集積回路）です。

具体的には、ネットワーク・プロセッサーの特長である、複数スレッドの同時処理が可能な低消費電力コア、ネットワーク特有のワークロードに特化したアクセラレーター、ネットワーク・オフロード・エンジン、メモリー・コントローラー、I/O などを統合し、小さなラインサイズでキャッシュやメモリーにアクセスする機能を備えており、同時に IBM サーバー・プロセッサーの特長である、IBM Power プロセッサー互換の命令セット、標準のプログラミング・モデル、OS とハイパーバイザーによる仮想化、さらには通常のサーバーと同様の RAS（Reliability、Availability、Serviceability）の体系をサポートしています。

チップ全体として、ネットワーク境界のワークロード処

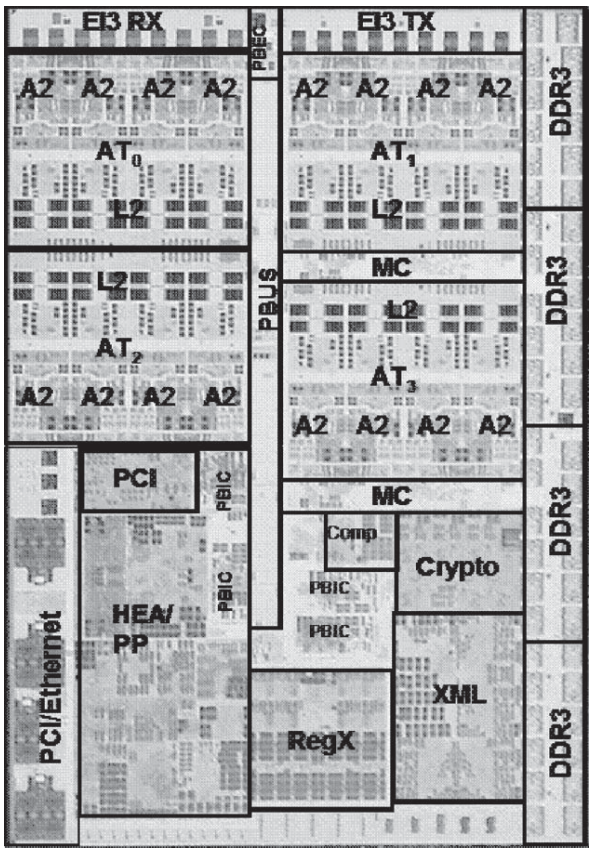

図2. The IBM Power Edge of Network processor

理に最適化したアーキテクチャーであり、ネットワーク・レイヤーからアプリケーション・レイヤーまで幅広い階層のワークロードをカバーし、通信速度はチップ当たり最大40G bit/sという高速のワイヤースピードを実現するものです。

半導体テクノロジーとしては、IBMの最新の45nm SOI（Silicon on Insulator）によるSoC（System on a Chip）です。IBM Power ISA（Instruction Set Architecture）準拠で、4スレッド同時処理可能な64ビット・プロセッサー・コア（A2 Core）を16個搭載し、1チップで64スレッド同時処理が可能です。4個のコアごとに2メガバイトのL2（レベル2）キャッシュ・メモリーを共有するノードを構成しており、1チップ上に4ノード実装しています。このほかに、ネットワーク境界での典型的なワークロードを処理する4種類のアクセラレーター、DDR3メモリー・コントローラー、さらにはイーサーネット・パケット・プロセッシング・エンジン、PHY、PCIeインターフェース、最大4チップまでのマルチ・チップ構成をサポートするための高速バスから構成されています。

また、幅広いアプリケーションを想定したスケーラビリティーがサポートされており、ローエンドのアプリケーションではチップ内の4ノードのうち、1ノードのみをアクティブにすることができ、ほかの3ノードをディスエーブルすることで無駄な電力消費をなくすことが可能になっています。

ハイエンドのアプリケーションでは、4チップを高速バスで相互接続し、合計16ノード構成（64コア、32MB L2キャッシュ、256スレッド同時処理）まで拡張することが可能になり、イーサーネットの通信速度としては1チップで最大40 Gbit/s（10 Gbit/s x4）、4チップ構成では160 Gbit/sまでカバーされます。

図2と**表1**は一次試作チップの概略です。チップ面積が428mm2、トランジスタ数が約14億3千万個と巨大なもので、IBMの歴史の中でも最も複雑なLSIの1つとなりました。

数十Gbpsというワイヤースピードの実現のため、ネットワーク固有のワークロードで、アルゴリズムの確立したものに対してはハードウェア・アクセラレーターを実装しています。それらは、データの圧縮・伸張、暗号化、パターン・マッチング、XMLの構文解析の4種類です。これらのアクセラレーターは、ネットワーク環境のパケット化された

表2. アクセラレーター仕様

Accelerator Unit	Algorithm	# of Engines	Projected Bandwidth	
			Typical	Peak
HEA	network node mode	4	40 Gbps	40 Gbps
	endpoint mode	4	40 Gbps	40 Gbps
Compression	gzip（input bandwidth）	1	8 Gbps	9.2 Gbps
	gunzip（output bandwidth）	1	8 Gbps	16 Gbps
Encryption	AES	3	41 Gbps	60 Gbps
	TDES	8	19 Gbps	
	ARC4	1	5.1 Gbps	
	Kasumi	1	5.9 Gbps	
	SHA	6	23-37 Gbps	
	MD5	6	31 Gbps	
	AES/SHA	3	19-31 Gbps	
	RSA/ECC（RSA with 1024/2048 bit key）	3	45000/7260	
XML	Customer workload	4	10 Gbps	30 Gbps
	Benchmark workload	4	20 Gbps	
RegX	For typical pattern sets	8	20-40 Gbps	70 Gbps

表1. PowerEN 概略仕様

Technology	IBM 45nm SOI
Core Frequency	2.3GHz @ 0.97V（Worst Case Process）
Chip size	428 mm2（including kerf）
Chip Power（4-AT node） Chip Power（1-AT node）	65W @ 2.0GHz, 0.85V Max Single Chip 20W @ 1.4GHz, 0.77V Min Single Chip
Main Voltage（VDD）	0.7V to 1.1V
Metal Layers	11 Cu（3-1x, 2-1.3x, 3-2x, 1-4x, 2-10x）
Latch Count	3.2M
Transistor Count	1.43B
A2 Cores / Threads	16 / 64
L1 I & D Cache	16 x（16KB + 16KB）SRAM
L2 Cache	4 x 2MB eDRAM
Hardware Accelerators	Crypto, Compression, RegX, XML
Intelligent Network Interfaces	Host Ethernet Adapter/Packet Processor 2 Modes: Endpoint & Network
Memory Bandwidth	2x DDR3 controllers 4 Channels @ 800-1,600MHz
System I/O Bandwidth	4x 10G Ethernet, 2x PCI Gen2
Chip-to-Chip Bandwidth	3 Links, 20GB/s per link
Chip Scaling	4 Chip SMP
Package	50mm FCPBGA（4 or 6 layers）

ストリーム・データ処理に最適化され、単純に寄せ集めて実装したのではなく、プロセッサーに新たなコマンド体系を定義し、データの入出力手法、パケットごとの処理結果のレポート、例外処理、仮想化、RASの体系などはすべて共通のインターフェースで実装されています。

個々のアクセラレーターがサポートしているアルゴリズムとスループットをまとめたものが**表2**です。

❸ 消費電力当たりの処理能力向上へのさまざまな工夫

3.1 アーキテクチャーの工夫による電力効率の向上

チップの低消費電力化の50%以上は、アーキテクチャーの工夫によるものです。ワイヤースピード・プロセッサーの、限られた面積に同時処理スレッド数をなるべく多く、かつ低消費電力でという要求仕様の実現に向けて、まずプロセッサー・コアに関しては、小型の高スレッド・プロセッサーが選択されました。小型の高スレッド・プロセッサーは、汎用のサーバー・プロセッサーに比べるとスレッド当たりのパーフォーマンスは低いのですが、消費電力当たりの処理能力の観点では2～4倍の性能を持っています［2］。また単位面積当たりに実装可能なプロセッサーの数は、小型の高スレッド・プロセッサーの方が2～4倍多くなるので、この場合は小型の高スレッド・プロセッサーが適しています［3］。

また、ネットワーク・ワークロードの中には、スタンダードで定められた画一的な処理が多く、これらはハードウェアでの処理に向いています。ハードウェア・アクセラレーターは、一般的に同じ処理をソフトウェアで行うより消費電力当たりの処理能力が10倍以上良くなります。ハードウェア・アクセラレーターを実装することで、パケット化されたデータをその単位でどのように処理するかという判断を高スレッド・プロセッサーで行い、それぞれのパケットの中身はアクセラレーターで処理するという最適なワークロード配分が可能になります。このほか、メモリー・コントローラーやシステムI/Oも含め、すべてを1チップに集積することは消費電力当たりの処理能力向上に非常に有効です。

以上のアーキテクチャー上の工夫に加え、ロジック回路、メモリーといった回路属性に最適化された電源電圧の設定や、ファンクション・ブロックごとに最適化された動作周波数の設定による消費電力の静的な削減、個々の回路の動作時以外はクロックの供給を止める（クロック・ゲーティング）ことによる消費電力の動的な削減、さらにはL2キャッシュを最新のeDRAM（embedded DRAM）で構成するといった実装技術の工夫も合わせて、2.0GHz動作時（動作電圧0.75V）で平均55Wの低消費電力（ワースト・ケース65W）を実現しています。

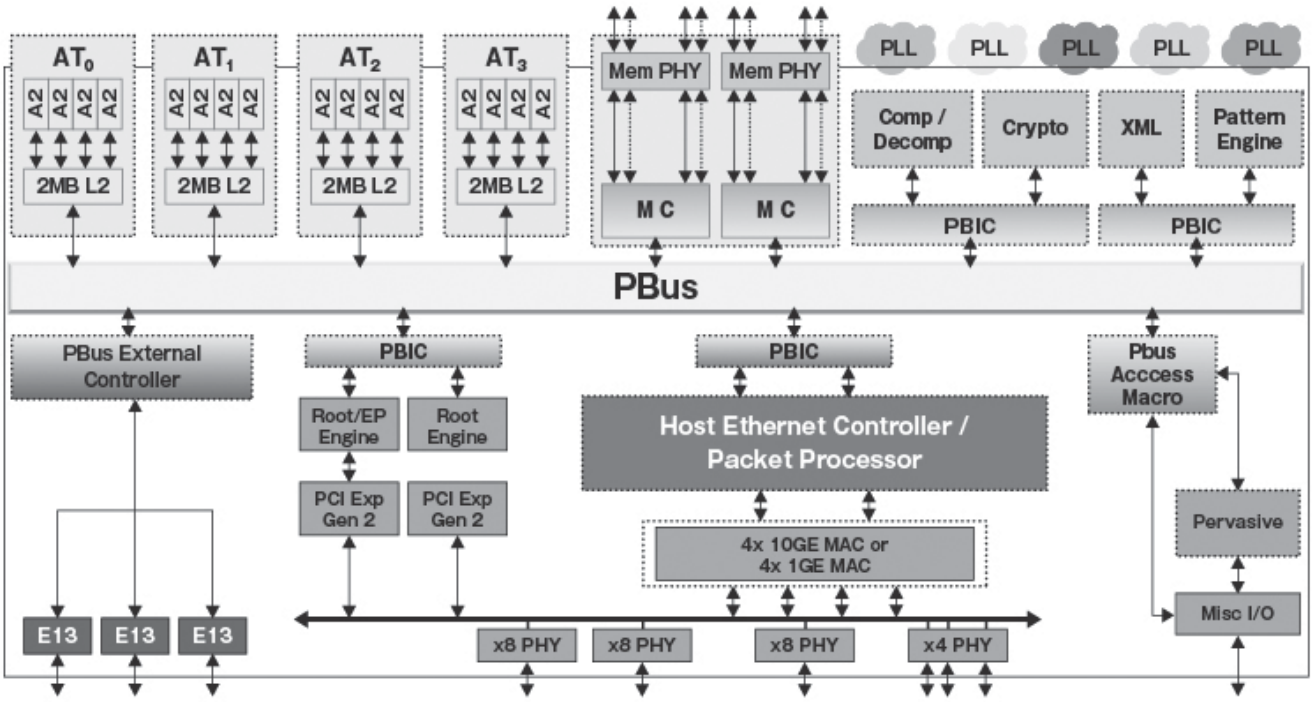

図3. ブロック・ダイアグラム

図 3 にチップ全体のブロック・ダイアグラムを示します。

図のブロック単位で最適な動作周波数が定められ、それらを右上の 5 種類の PLL（Phase Locked Loop）でカバーしています。

消費電力の削減には低電圧化が有効ですが、定められた動作速度を実現するために必要な電源電圧は、製造ばらつきからチップによって微妙に異なります。さらには動作周波数と消費電力への要求事項はアプリケーションごとに異なるため、個々のチップに必要な最低電圧はチップの製造工程上のテストで決定され、チップ内部の不揮発性メモリーに保存され出荷されます。そのデータはチップの電源投入時に読み出され、外部の電源回路が消費電力の観点から最適な電源電圧を供給するのに使われます。この技術を Adaptive Power-Supply（APS）と呼びますが、これはプロセス分布全体での平均消費電力の低減に効果的です［4］。

チップ内部へのクロックの供給に関しては、位相のばらつきを最小限にすることが、局所的な回路の不必要な高速化の削減につながり、電力消費の低減に非常に有効です。このチップではバランスの取れた H 型ツリー構造を採用することでクロック位相のばらつきを最小限に抑えています。

3.2 クロック・ゲーティングによる低消費電力化

動作時の電力消費を減らすため、チップ全体にクロック・ゲーティングを行っています。具体的にはローカル・クロック・バッファが、4 ～ 32 個のフリップ・フロップ単位のきめ細かさで配置されており、この単位でクロックのオン・オフのコントロールを行っています。

クロック・ゲーティングの有効性は、クロック・ゲート制御がなされているフリップ･フロップの割合で決まりますが、このチップでは、ファンクション・ブロックごとに動作時とスタンバイ時のクロック・ゲート制御の割合に、それぞれ 70% と 90% という設計目標値を設けました。

実際に実装された値は、ファンクション・ブロックによって異なりますが、全体としては目標の 95% が実現され、結果として最大 32%（40W）の電力消費の削減が達成されています。

またクロック・ゲーティングのような動的なクロック制御とは別に、アプリケーションによって、あるファンクションが不要の場合、該当ブロック全体のクロックを止めることで、そのブロック全体の電力消費を削減することも可能になります。

3.3 デバイスの最適化による低消費電力化

微細加工の進んだ最近の CMOS トランジスターは、漏れ電流による電力消費が増大する問題を抱えていますが、これに対してはオン・オフの閾値（しきいち）電圧（Vt）を変えることにより、動作速度と漏れ電流の相反するパラメーターを選択的に使い分けることで対応しています。低消費電力の観点から、それぞれの回路に求められる必要最小限の動作速度のトランジスタを効果的に選択する必要があります。

IBM の 45nm のテクノロジーは多種の閾値電圧（Vt）をサポートしており、このチップでは Regular Vt（RVT）、High Vt（HVT）、Super High Vt（SVT）の 3 種類を使い分けています。RVT は最も高速ですが、漏れ電流が大きいため、限定的に使用されています。通常は論理合成時に HVT を使って合成され、フリップ・フロップ間のタイミング検証の後、動作速度に余裕のある経路上のトランジスタが、速度は遅いのですが漏れ電流のより少ない SVT に置き換えられます。チップ全体として RVT 5%、HVT 20%、SVT 75% を目安として設計され、低消費電力化が図られています。

ノードごとに実装された 2MB の L 2 キャッシュには、SRAM ではなく Deep trench（DT）embedded DRAM（eDRAM）が採用されています。この eDRAM の採用により同じ面積の SRAM に比べ 3 倍以上のメモリー容量を持ちながら、消費電力を約 20% の低消費電力で実装することが可能になりました［5］。

❹ 終わりに

今回はテクノロジー最前線として、ネットワークの境界で現在のデータセンターが、どのような課題を抱えており、それに向けてどのような技術革新が進みつつあるかを、まだ開発段階にある IBM のワイヤースピード・プロセッサー（PowerEN）を例にご紹介しました。

IBM ではワークロードに最適化されたシステム設計というサーバー製品共通の戦略的な取り組みを進めていますが、それは PowerEN のような半導体チップのレベルから実践され、お客様の問題解決に取り組んでいます。本プロセッサーを実装した製品が、近い将来お客様のお役に立てるものと確信しています。

[参考文献]

[1] Charles Johnson, David H. Allen, Jeff Brown, Steve Vanderwiel, Russ Hoover, Heather Achilles, Chen-Yong Cher, George A. May, Hubertus Franke, Jimi Xenedis, Claude Basso, "A Wire-Speed PowerTM Processor: 2.3GHz 45nm SOI with 16 Cores and 64 Threads" , Proc. 2010 IEEE International Solid-State Circuits Conference, pp.14-16, (2010.2).

[2] S. Balakrishnan, R. Rajwar, M. Upton, K. Lai, "The Impact of Performance Asymmetry in Emerging Multicore Architectures" , ISCA '05, pp. 506-517, (2005.6).

[3] D.M. Tullsen , S.J. Eggers, H.M. Levy, "Simultaneous multithreading: maximizing on-chip parallelism" , ISCA '95, pp. 392-403, (1995.6.22).

[4] M. Keating, D. Flynn, R. Aiken, A. Gibbons, K. Shi. "Low Power Methodology Manual." Springer, 2007, pp. 21-24.

[5] J.Barth et al, "A 1 MB Cache Subsystem Prototype With 1.8 ns Embedded DRAMs in 45 nm SOI CMOS," IEEE J. Solid-State Circuits, vol. 44, no. 4, (2009.4).

[6] 宮武久忠 , "検索時パリティチェック機能を有する高速並列型連想メモリ" , PROVision 67 号 , 日本アイ・ビー・エム株式会社 , pp.87-93, (2010.11).

日本アイ・ビー・エム株式会社
開発製造　大和システム開発研究所
シニア・テクニカル・スタッフ・メンバー（技術部長）

西野 清志　**Kiyoshi Nishino**

[プロフィール]
日本 IBM 入社以来ワークステーション、組み込み製品、ストレージ製品、ハイパーフォーマンス・コンピューターなど幅広い製品開発に従事。研究開発部門のテクニカル・ストラテジー担当を経て、2008 年よりワイヤースピード・プロセッサーの開発に従事。IBM Academy of Technology 会員。