FPGA を活用したスケーラブルな高速分散データベースの開発
―基盤改善によるデータベース応答速度の向上―

1. 背景

昨今、個々のハードウェア資源は急速な勢いでその性能を向上させており、相対的に各ハードウェア資源間を繋ぐバスインターフェースや、システムがハードウェアを利活用するためのドライバが全体の処理を律速するケースが発生しつつある。

データセンターにおいて普及しつつある 10G Ethernet 環境は、現在主流となっている Gigabit Ethernet と比較して帯域幅が 10 倍に増大する事に加え、内部通信のクロック上昇により遅延も減少する。また、ストレージ分野では NVM（Non Volatile Memory）と呼称される次世代のアーキテクチャが市場投入に向けて開発中であり、ハードディスクや SSD といった既存製品と比較して遅延と帯域の双方において大幅な性能改善が予想される。

このように各種分野において新規アーキテクチャが投入されつつある現状において、アプリケーションの処理全体に掛かる時間の内訳は変化してきている。I/O が律速となっているアプリケーションにおいてハードウェアの性能が改善されると、バスインターフェースやシステムプログラム（オペレーティングやデバイスドライバ等）における処理時間は無視する事ができなくなる。実際に、10G Ethernet 環境においてはモノリシックカーネル内のプロトコルスタックによる処理がハードウェア性能を十分に引き出せないため、ドライバとアプリケーションを垂直統合したユーザランドシステム（例：Intel Data Plane Development Kit、以降 Intel DPDK）が提案されており、オペレーティングシステムカーネルによって生じているオーバーヘッドの低減が求められている。

2. 目的

本プロジェクトは、今後急速に普及すると考えられる高性能ハードウェア資源を用いる環境において、ハードウェアアーキテクチャやデバイスドライバの改良によりデータベースという実際のアプリケーションにおいて処理時間を低減できる事の実証を目指した。

バスインターフェースにおける通信やデバイスドライバにおいて、ある程度のオーバーヘッドが存在する事は自明であるが、このオーバーヘッドが実際のアプリケーションに対してどの程度の性能低下を及ぼすかはアプリケーションの特性に依存する。具体的にはアプリケーションの処理時間が短くなるにつれ、これらのオーバーヘッドが全体の処理時間の中で支配的になる割合が高まる。本プロジェクトでは実用的に使われているアプリケーションとしてデータベースを用いた上で、その中でもこれまで挙げてきたオーバーヘッドが問題となるケースを取り上げ、そのようなケース下においてこれを改善する事により性能を改善させる事を目指した。

この改善が実現する事によって、現在は次世代アーキテクチャが本来の性能を十分に発揮できていない領域がその恩恵を被る事ができるようになる。つまり、ハードウェアの性能改善を可能な限りソフトウェアの性能向上につなげる事が本プロジェクトの副次的な目的であると言える。

3.　開発の内容

本プロジェクトでは独自のネットワークインターフェースカードと小規模なオペレーティングシステムを実装し、それぞれにデータベース処理を委任する事によって、KVS（Key Value Store）型データベースにおける 1 クエリ当たりの処理時間を改善した。

本プロジェクトの目的である、バスインターフェースやデバイスドライバのオーバーヘッド削減のため、ハードウェアからアプリケーションまで垂直統合したシステムを開発した。一般的な環境においては、ハードウェアとしてのネットワークインターフェースカード、オペレーティングシステム（デバイスドライバ）は汎用性のために独立している。これをアプリケーションであるデータベースに対して密結合する事で、オーバーヘッドの削減と、各要素によるタスクの分担が実現できる（図 1）。まず、オペレーティングシステムをデータベースに特化し、かつ密結合した物とする事により、必要な機能を絞って省力化できる他、OS とアプリケーションの間の通信によって生じるオーバーヘッドの削減が可能になる。同時に、ハードウェアではアプリケーションの機能の一部を代わりに実行する事によって、CPU リソースの削減のみならず、バスインターフェースの通信そのものの削減が可能になる。このソフトウェアレベルとハードウェアレベルのオーバーヘッド低減の双方がシステム全体の垂直統合により実現できた。KVS 型データベースのクエリにおける処理時間の大半は I/O 処理に律速されており、本システムにより KVS 型データベースの 1 クエリ当たりの処理時間が大幅に改善された。

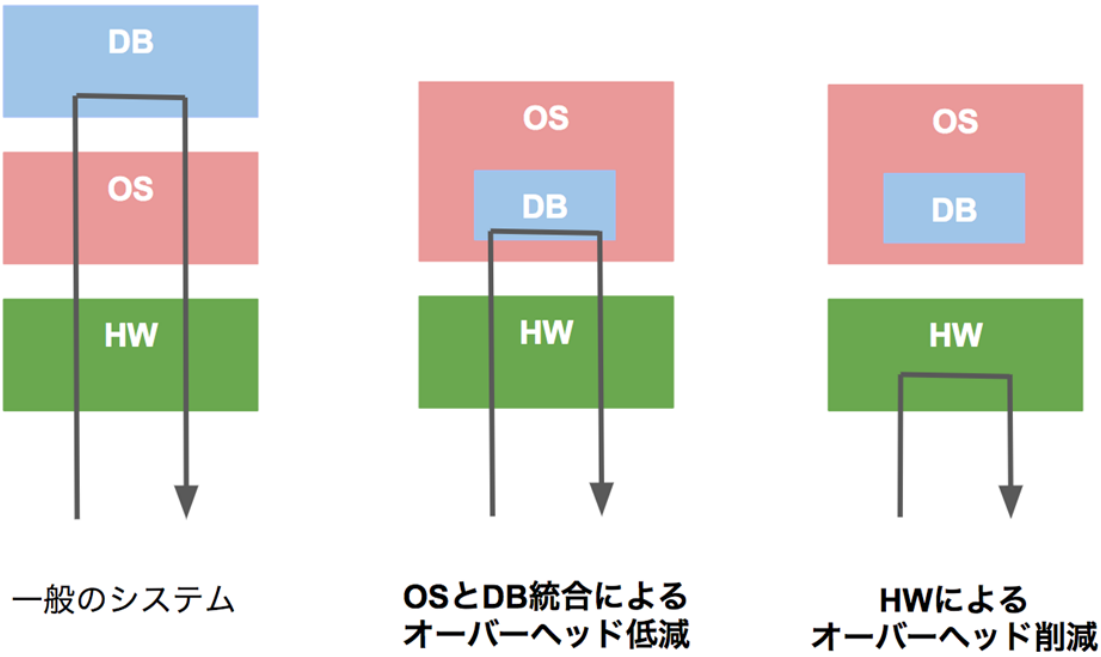

図 1 本プロジェクト開発物による目標達成イメージ

このような垂直統合型システムを実現するため、本プロジェクトではハードウェアとしてのネットワークインターフェースカード及びシステムソフトウェアであるオペレーティングシステムの双方を独自に開発した。前者は PCI Express インターフェースと Ethernet ポートを搭載した FPGA（Field-Programmable Gate Array）上で動作し、後者は x86 アーキテクチャにおいて汎用的に動作する。

4.　従来の技術（または機能）との相違

現在直面しつつあるバスインターフェースやデバイスドライバによる性能律速は、アーキテクチャを構成する階層（ハードウェア、システムソフトウェア、アプリケーション）ごとの独立した性能改善（一般的に広く行われている手法）では解決できないため、本プロジェクトではシステム全体を俯瞰し、ハードウェアからアプリケーションまで全ての要素を垂直統合したシステム全体での性能改善を行った点が特徴的であると言える。この手法を用いる事により、データベース自体の性能の改善では解決できない根本的な性能上のボトルネックを解決する事ができた。

システムを垂直統合した改善の例としては、Intel DPDK を用いたアプリケーション開発が存在する。しかしながら、この手法ではハードウェアレベルの改善を行う事ができず、Intel 社製のネットワークインターフェースカード上で動作する環境におけるデバイスドライバレベルでのオーバーヘッド削減しか実現できない。本プロジェクトにおいて中心に位置づけているハードウェアレベルでのデータベース処理の移譲プロセスは、Intel DPDK で実現する事ができないため、本プロジェクトにおいてはシステムソフトウェアも含め全て再実装を行った。

5.　期待される効果

本プロジェクトによって実証された手法を用いる事により、データベースを根幹としている各種アプリケーションの高速化、及びそれらのアプリケーションを活用する環境における次世代アーキテクチャの本格導入が期待できる。

一般的に、新規開発されたアーキテクチャはその普及までに時間が掛かる。実際に 10G Ethernet は、各種周辺機器の高価格化を理由にデータセンタレベルでも依然として普及途上である。次世代アーキテクチャ環境におけるアプリケーションの性能の改善は、それらのアーキテクチャのコストパフォーマンスを下げ、ハードウェア導入のきっかけを生む。ハードウェアの売上の増大は量産化による価格低下を生じ、更なる普及に繋がるという相乗効果を引き起こす。つまり、次世代アーキテクチャ下におけるアプリケーション性能の向上は、アーキテクチャの世代交代の最初の契機となり得る。本プロジェクトでは実験に用いたハードウェアの制限や現時点で製品化されていないアーキテクチャを想定したため、実際の数値として次世代アーキテクチャにおける性能改善を示す事はできなかったが、提案したモデルや、計測結果から本提案が次世代アーキテクチャにこそ有用である事は十分に示されている。

6.　普及（または活用）の見通し

本プロジェクトでの開発成果を実際の製品システムに組み込む事に関しては依然として課題が残るものの、今後更なる性能向上手法を実装する事により、主に Web サービス等を提供している企業における利活用が見込まれる。

本プロジェクトでの開発成果を直接サービス内に組み込む場合、memcached といった既存のキャッシュシステムを置き換える事例が想定される。この場合、キャッシュシステムが必要となる程の大規模な Web サービスにおいて利活用される事になる。サービスの規模にはよるものの、おおよそ数百万から一千万程度のユーザー数を誇るサービスがその対象となると予想される。

本システムはハードウェア及びシステムソフトウェアの改善によってアプリケーションの性能向上を達成しているため、ハードウェアやシステムソフトウェアをサービス運用者自身が管理する環境への導入を当初想定していた。しかしながら、昨今の Web サービスはそのバックエンドとしてクラウドサービス（例：Amazon Web Services, Microsoft Azure）を用いる事が多く、それらのサービスの基盤部分に本システムを導入する事で、アプリケーションサービス提供者に影響を与える事なく、多くのシステムにおいて本システムの利活用ができるようになる。

将来の展望として、本システムを SQL 等に代表される構造化データベースのバックエンドとして動作するように改良する事で、キャッシュ用途のみならず、アプリケーションデータを格納するデータベースの高速化が図れる。構造化データベースは内部に KVS 型データベースに比較的近いストレージエンジンを包含しており、これを本システムで置き換える事は比較的容易であると予想される。これにより比較的小規模な Web サービスであっても、本システムを利活用できるようになる。

7. クリエータ名（所属）

粟本 真一（東京大学理学部情報科学科）
包 含（東京大学理学部情報科学科）
関 祥吾（東京大学理学部情報科学科）

（参考）関連 URL

本プロジェクト内で開発したシステムのソースコード等

https://github.com/Raphine/