次世代 HPC・Enterprise 統合基盤のためのアプリ上の共通の特徴

田邊 昇　　冨森苑子　高田雅美　城和貴
(株)東芝　　奈良女子大学

1. はじめに

近年，Big Data 解析へのニーズの高まりに伴い，Enterprise 用途の計算基盤への要求に大きな変化が生じてきた．その結果，Enterprise 用途と HPC 用途の基盤には共通の特徴が求められるようになってきた．基盤統合は改良の恩恵の波及範囲を拡げるとともに，量産による経済的メリットが期待できる．本ポスターでは Enterprise 用途と HPC 用途の計算基盤統合を目指すにあたり，従来のベンチマークではあまり活性化できなかったアプリケーション上の共通の特徴を考察する．

2. ベンチマークにおける変化と共通点

表 1 に HPC 用途と Enterprise 用途のベンチマークの特徴の概要を示す．HPC 向けシステムの主要ベンチマークは Top500[1] (HPL)だった．HPL は密行列処理で，キャッシュベースの CPU 上で効率的に動作できる．Top500 は疎行列の反復解法の HPCG [2]に切り替わろうとしている．これはキャッシュベースの計算機が苦手とする処理である．

一方，従来の Enterprise 用途の基盤設計においては用いられてきた SPEC ベンチマークはキャッシュアーキテクチャにとってあまり厳しくなかった．TPC ベンチマークはストレージアクセスが最大のボトルネックだった．一方，Big Data 解析のニーズの高まりを背景に，従来の Top500 が活性化できなかった処理のベンチマークとして Graph500[3]が誕生した．これも HPCG 同様に疎行列処理である．その行列は巨大でランダム性が高い．

以上のような主要ベンチマーク上のワークロードの変化が HPC と Enterprise の両方においてほぼ同時に起きている．それらは巨大な疎行列に対する処理という共通の方向性をもっている．つまり，今後の計算基盤設計においては，従来の適応性を維持しつつ，上記の共通の方向性に向けた改善を行ない，HPC と Enterprise の両方を効率的に実行できる統合された基盤とすることが重要である．

表 1. 各種ベンチマークにおける変化と共通点

	HPC		Enterprise		
	Top500		SPEC	TPC	**Graph 500**
	旧	**新**			
	HPL	**HPCG**			
カーネル	密行列	**疎行列**	多様	DB	**疎行列**
サイズ	小～中	**小～大**	小	大	**小～大**
アクセスパターン	規則的	**規則的 or 不規則**	多様	やや規則的	**不規則**
キャッシュ	適	**不適**	適	中間	**不適**

3. Big Data 解析におけるランダム性

大容量データに対するランダムアクセスは，現在のキャッシュベースの計算機システムにおいて不得意な処理である．なぜなら，空間的局所性と時間的局所性が乏しいアクセスになり，階層記憶が有効に機能しないためである．単一ノードに入りきらない大きなデータへのランダムアクセスはバースト長の短い通信を誘発するため，ネットワークへの負荷が高まる．

ところが，Big Data 解析においては，2 つの観点からのランダム性が本質的に存在する．第一は疎行列で表現される対象に含まれるランダム性である．例えば SNS における友人関係，WWW におけるリンク関係など，サイトや個人ごとの任意性が本質的に存在する．この特質は Graph500 のワークロードに反映されている．

第二はハッシングに伴うランダム性である．Hadoop 等がアクセスするデータは Key Value Store(KVS)に Key によってランダム化された位置のデータをアクセスして負荷分散を実現する．この仕組みが本質的に大容量データへのランダムアクセスを発生させる．なお，この仕組みは HPC でのニーズが高い Gausian が用いる Linda と共通である．

局所性が無いメモリアクセスのスループットはキャッシュライン内の有効データ率により制約される．4 バイトの整数へのランダムアクセスを行なう場合は主記憶の最大バンド幅の 1/16～1/32 に実効バンド幅が低下する．Big Data 解析向け基盤においても HPC 向けの基盤と同様に主記憶バンド幅が高いプロセッサへのニーズが高まる．

一方，局所性が無い通信のスループットはその並列システム上の二分割バンド幅によって制約される．よって，十分なスケーラビリティを得るためには，ネットワークも Infiniband を用いた Fat Tree などの HPC 向けのネットワークと同様の性質が求められる．

以上のように，Big Data 解析は HPC と同じ方向性に向かって計算基盤を進化させる必然性がある．

4. Enterprise のインメモリ化の影響

多くの HPC 系アプリと同様に，Graph500 はインメモリで動作する．例えば世界の人口が約 70 億人であるように，現実世界の解析対象には本質的な上限があることが多い．8 バイト/人の BFS 木格納配列を世界の人口分確保するとしても 56GB に過ぎず，多くの現実的な解析対象グラフを単一または複数ノードのインメモリで処理できる．

従来の Hadoop は実行途中にストレージアクセスがあるため，ボトルネックはストレージにあり，SSD のニーズが高まった．しかし，インメモリで動作する Hadoop 互換環境[4]が存在し，主記憶容量の増加に伴い，今後はインメモリ解析が普及するであろう．その結果，ボトルネックはストレージではなく，メモリとネットワークに移動する．

ストレージネックの典型例であるデータベースマネージメントシステム(DBMS)もインメモリ化が進行中である．急速に普及しつつあるインメモリ DBMS である SAP HANA[5]は解析系の高速化のために列優先のデータ構造と，トランザクション処理の高速化のために行優先のデータ構造も併せ持ち，行優先から列優先に格納しなおす．この操作は主に主記憶上の行列の転置に相当し，キャッシュの効果が低い．Enterprise 用途である DBMS においても，HPC の FFT と共通の特性が基盤に求められると考えられる．

5. おわりに

大容量データへのランダムなアクセスが，現在の計算機システムが抱える弱点であるとともに，HPCとEnterpriseの双方が必要とする改善点であることを論じた．Enterpriseのインメモリ化によりEnterpriseのボトルネックはストレージからメモリとネットワークに移動すると考えられる．今後は上記の知見に基づいた適切な解決策の研究開発に，重点を移して行くことが望まれる．

謝辞　本研究の一部は総務省戦略的情報通信研究開発推進制度(SCOPE)の一環として行われたものである．

参 考 文 献

[1] Top500: http://www.top500.org/.
[2] M. A. Heroux and J. Dongarra: "Toward a New Metric for Ranking High Performance Computing Systems", Sandia National Lab. Report, SAND2013-4744 (2013).
[3] Graph500: www.graph500.org/.
[4] ScaleOut Software Inc.: "ScaleOut hServer", http://www.scaleoutsoftware.com/products/scaleout-hserver/
[5] V, Sikka et.al.: “Efficient transaction processing in SAP HANA database”, Proc. SIGMOD’12, pp.731-741 (2012)