IBM　ハイパフォーマンス・コンピューティングフォーラム　2000　報告 ～その２～

The Report About IBM HPC Forum 2000 (part2)

長谷　佳明
Yoshiaki NAGAYA

Abstract: This is the report about "IBM HPC FORUM 2000",and especially I report the architecture of "BLUE GENE ".

1　はじめに

昨月大阪にて行われた「IBM HPC FORUM 2000」, 特に「BLUE GENE Project」について報告を行う. そして私は, その中でも「BLUE GENE」の持つ特異なアーキテクチャを中心に報告をする.

2　BLUE GENE Project の技術的側面

先に角さんより報告があったように,BLUE GENE は,IBM が手がける研究プロジェクトである. その目的は, 昨今一般的にも注目を集めるようになった GENE, つまり遺伝子に関連している. その実は, 遺伝子から作られるたんぱく質の構造の解析にある. しかしながら, その解析には膨大な計算が要求され, その計算量そのものが問題となっている. その問題を解決するために IBM は独自のハード的, ソフト的アプローチで新たな計算機を開発した.

3　BLUE GENE 専用機その開発

3.1　開発プランとその問題性

前セクションでも述べたとおり, このプロジェクトはたんぱく質の構造解析に関する計算をするために超高性能演算が求められ, その概算では,「1 Petaflops/s」つまり「$1 \times 10^9 \times 10^6$」が必要とされている. それを解決するためには, 現段階の技術におけるワンプロセッサ処理では到底解決することができない. その解決法の基盤となるのが, 多数のプロセッサを協調させることで処理を飛躍的に高める並列処理技術である. その方法はさらに, 共有メモリマシン型, そして分散メモリ型に分割される. 昨今では, その構造を階層化させ, 各ノードは SMP であるが, それを高速のネットワークで結んだ分散メモリ型マシンも存在する. そして, 現段階での高性能コンピュータの実現方法としては, 以下の 2 点が見つかるであろう.

1. Connect a Million PCs in One Machine Room
2. Build a Larger SP

しかしながら, 上記の 2 点とも大きな問題を抱えている.

3.1.1　Problem about Connect a Million PCs in One Machine Room

この,PC を 100 万台コネクトするという方法は, 確かに, 最高の状態でいかなる問題も起こらなければ, 目標の演算性能を得ることができるかもしれない. しかし, これらは以下のような問題をはらんでいる.

1. 電力問題 - 要求されるのは 100 Megawatt !!
2. 設置場所の問題 - 20000cabinets,5 acres !!
3. 管理コストの問題 - 単に PC を直すために 1000 人の専任技術者！
4. システムそのものの信頼性 - PC の"failure"が 0.1 秒ごとに生じる！

電力の値としては, これは京都市の電力消費量に値してしまう. これらの問題から 100 万台の PC をコネクトしたマシンは現実的ではない. 実は IBM の技術者の方は, 更なる問題点として,「One Million MS Windows licences(Oh no!)」というものも挙げられていた.(アメリカンブラックジョークだと思われるが)

3.1.2　Build a Larger SP

SP とは,IBM が販売している最上位機種である RS/6000 である. このマシンは最高で 512 ノードまで接続が可能なスケーラビリティを持っている.[1]しかしながら, このシステムをもってしても 1Petaflops を実現することは現段階では難しい.IBM としても 3〜5 年後には, 各チップの性能向上により実現が可能となる可能性を示唆しているが, 以下の問題をはらんでいる.

1. 500MW の電力
2. 10acres の敷地

[1]max である 512 ノード接続したマシンとして "ASCI White 8192CPUs,6.2TB memory,160TB disk,12.3TFlops/s"が有名である

3. 1M 個の Chip

4. 5 万もの UNIX Sytem Image

これらから,SP を用いても目標とするパフォーマンスを発揮できないと判断された.

3.2 一般的な問題

ここで, 一般的な汎用機をこのプロジェクトで使用することについての問題を挙げておく.

プロセッサの問題 汎用プロセッサは, その処理のサイクルに対する最適化が行われている. その最適化の複雑性故に, プロセッサを増やして協調処理させることを困難にしている.

Von Neumann bottleneck これは, 有名な問題であり, 今日のすべてのコンピュータの抱える問題である. つまり, プロセッサの速度に対して,I/O や Memory へのアクセス速度が極端に遅いことである.

OS 自体の抱える問題 OS が, あまりにも多くの問題を対処するために, あまりに多くの命令を必要とし, それに応じて OS の処理自体が複雑になる. このため,"failure"が頻発する.

3.3 The Answer

前のセクションまでの問題を考慮した上で次のような解決法が考えられた.

1. スーパーコンピュータをスクラッチから開発

2. 実は"Unbalance"なシステム

「1」とは, 先の問題を解決するためには, スクラッチ, つまり全面的にそのアーキテクチャの根本から見直すと決めたことをさしている. 「2」とは, メモリと CPU の関係の"unbalance"を避けていた伝統的なシステムを見直し, そのシステムを不均衡にしてはどうか？という逆転の発想である.

4 The architecture of "BLUE GENE"

4.1 The Unbalance System に対して

従来のシステムにおいては,Table.1 のようにシステム構成されていた.Balanced System が適している処理形態は,N 倍の CPU パワーで元の N 倍の数のジョブもしくはトランザクションを処理することである. しかしながら,N 倍の CPU パワーで元々解いていた問題と同じかもしくはより大きな単一の問題を解くには不向きである. そして, このプロジェクトが対象としているたんぱく質構造解析では, その問題の大きさそのものが拡張する傾向がある. さらに, 対象の大きさを N 倍するに対して命令数として 2 乗のオーダで増加する傾向を持っている. つまり, 以下の方程式が成り立っていた.

Table 1　従来のシステム (Balanced System)

	Memory	I/O	CPU Power
Balanced System	1byte	1bit/s	1op/s

$$"NtimesCPUPower"^2 \quad \rightarrow \quad "NtimesMemory" \tag{1}$$

つまり,CPU のパワーが N の 2 乗倍になって初めて N 倍の問題を処理することが可能となっていた. そのため,Unbalance System が必要と考えられる. つまり, 従来の CPU パワーとメモリの関係を取り払ったものである.

4.2 CPU の処理配分

CPU の処理能力は確かに, コンピュータの全体の処理のキーである. しかしながら, その処理バランスの中身を解析すると以下の Table.2 のような結果が出ている. これらからも明らかなように,CPU の処理の大半はデータの移動に費やされ, 本質の演算には使われていない.

Table 2　CPU の処理配分

	Moving Data	Another
CPU Power	90 %	10 %

4.3 Solutions

Von Neuman bottleneck への対策 キャッシュの大容量化といったメモリの階層化による方法ではなく,CPU チップ内に高性能の主記憶 (16MB) を内臓した.

メモリアクセス速度の問題 なお残るメモリアクセス速度と,CPU 速度のギャップであるが, これに対しては,CPU 内にマルチスレッドユニットを 8 つ乗せることでパフォーマンスを向上させた.

並列化によるプロセッサ間のシグナル遅延問題 プロセッサを単純化し, 一つのチップ上に複数のプロセッサをグリッドで配置した. これによりこれまで以上の並列化が可能となった.

プロセッサの開発時間の問題 プロセッサを抜本的に見直し, 命令数を極限まで減らした.[2]

[2] この背景には, 実は最も開発に時間をかけた命令が最も使われていなかったという教訓に基づいている.

OS の廃止 専用計算機を設計するため OS による汎用性は犠牲にした. これにより,OS のオーバヘッドの消失. プログラムのコンパイルには Linux マシン上でのクロスコンパイルを用いることとした.

5 Building BLUE GENE

5.1 Design

CPU などの配置とさらにそれに伴う性能向上は以下の Fig.1 のようになる. そしてその性能は,Table.3 のようになる.

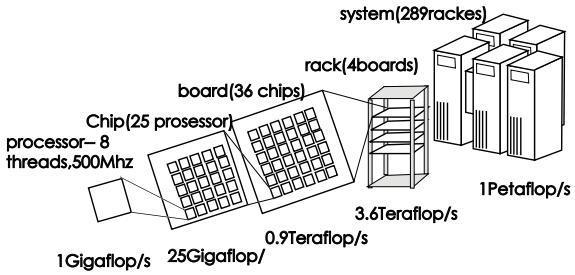

Fig. 1 スケーラビリティ

Table 3 BLUE GENE

Threads	Prosessors	chips	memory	Power
8M	1M	40k	0.5TB	1MW

さらに,Chip から Chip へのコミュニケーションは, ハイパーキューブ型にコネクションされ, 入力方向 500Mb/s, 出力方向 500Mb/s を実現している.

5.2 Error Handling

これら並列処理マシンにおいて”failure”との関係は重要な課題である.BLUE GENE においては, チップが数日で”failure”を起こすことを想定している. しかしながら, これらのエラーを全て捕らえることは不可能である. そこで,「Application Checkpoint」なるものを設けることでエラーチェックをしている. その実とは, まず System レベルで自身の正しさを自己的にテストを行うことでチェックを行う. さらに,Application レベルでもそれ自体が自己の状態が正しいのか定期的にテストを行うことでチェックを行う. その裏で, 定期的にファイルという形でジョブをデータごとダンプをしておく. そしてもし, 各レベルでエラーを観測した際には, 最新のスナップショットからシステムを再構築することでシステムのハザードを防いでいる.

5.3 HOUSE of BLUE GENE

BLUE GENE は, 地下一階に置かれており, 空調を制御するのみならずシステム冷却のために, 地下一階と一階の間に, 配水管を設置することで冷却に一役買っている.

6 IBM HPC Forum 2000 を通じて

今回は, ハイパフォーマンスコンピューティングに関して話を聞く機会を得た. ここで説明された BLUE GENE の公演は大変エキサイティングな内容に満ちていた. 直接的にこの技術を知ることで我々が得られる知識は少ないかもしれない. しかしながら, 世界の最先端の研究についてその第一線で働く方から説明を受けたことは大きな刺激となった. 彼らの話ぶりは, 自身があふれ自己の実現しようとするものに対する情熱と誇りに満ちていた. 私自身にとっては, 初めてこれら本格的講演会で生の英語のプレゼンに触れたことも貴重な体験となった. 付け足しではあるが, 他の公演についてはプレゼンの最後に簡単に説明する予定である.

7 参考文献

参考文献

1) 『IBM HPC FORUM 2000 Lecture Note』(IBM, 2000)