素朴な疑問

計算機の性能評価はどうやってやるのですか？†

堀 川 隆††

「計算機の性能とは何か」，だれでも分かっていそうなことですが，こうして改めて聞かれてみると，実は，一本化された確固たる定義（指標）が存在していないことに気が付きます．

計算機の性能は，1)ハードウェアの性能(CPUを駆動する基本クロックの周期であるマシン・サイクル等)，2)計算機アーキテクチャの善し悪し，3)基本ソフトウェア（OSや言語処理系）の性能，4)アプリケーション・プログラムの種類，といった数多くの要因，および，それらの間の相性に依存する[1]ため，たとえば，マシンサイクルの短い計算機ほど高性能だというような，簡単な話では済みません．性能評価が1つの研究分野にまでなっている所以(ゆえん)でもあります．

複数の計算機の性能を比較評価する場合，直感的には，同じプログラムの実行を速く完了できる方が性能の良い計算機だといって間違いないように思えます．一方，同じ時間にこなせる仕事の量が多いほど性能の良い計算機だという考え方も正しいように思えます．この両者には関係はあるのでしょうか．前者の性能をレスポンス性能といい，後者をスループット性能といいます．レスポンス性能の良い計算機は，おおむねスループット性能も良いといえます．しかし，1台の計算機について考えた場合は，これらは相反する関係にあり，実際のところは，そう単純な話では済みません．というわけで，計算機性能を評価する場合，まず，どちらの側面を調べるのかを意識しておく必要があります．

一般の計算機ユーザが，性能を評価する方法として最も馴染みのあるのは，ベンチマーク・テストでしょう．ベンチマークというのは，計算機で実行させるジョブを代表するような標準的なテスト用プログラムのことです．少し前までは，プログラムから典型的な処理のパターンを抽出したものが利用されていました．しかし，このような小規模なベンチマークでは，コンパイラの出力するコードの微妙な差が極端な性能差になって表われてしまう場合があり，総合的な計算機性能を公平に比較するのが難しいという問題が認識されました．そこで，最近では，実際の応用プログラムそのものや，それに近いプログラムが用いられることが多くなっています．このプログラムの実行を完了するまでの時間（スループットの場合もある）を，様々な計算機について調べることにより，性能比較を行います．

現在，多用されている比較的新しいベンチマーク・プログラムには，SPEC CINT 92/CFP 92, Perfect Club, TPCがあります．SPECは，現在，unixワークステーションのCPU性能を比較するための標準的なベンチマークとして利用されています．これには，整数データのみを使用するプログラム群（CINT 92）と浮動小数点演算中心のプログラム群（CFP 92）があります．前者で調べた性能をSPECint 92，また，後者で調べた性能をSPECfp 92と呼び，両者を区別することになっています．Perfect Clubは，スーパコンピュータを含む数値計算用計算機の性能を調べるベンチマークです．TPCは，トランザクション処理性能を調べるベンチマークであり，現在，TPC-A～Cの結果が公表されています．

ベンチマーク・テストを利用する場合，単なる数値の比較だけを行うのではなく，ベンチマーク・プログラムの概要を理解しておくことが重要です．というのも，あるベンチマークで調べた性能が，計算機1＞計算機2となっても，別のベンチマークでは，この関係が逆転することが起こり得るためです．一例として，キャッシュ・メモリの容量よりもマシン・サイクル短縮を重視した計

† How to Evaluate Computer System Performance by Takashi HORIKAWA (C & C Research Laboratories, NEC Corp.).

†† NEC C&C研究所

算機と，逆に，キャッシュ・メモリの容量を重視した計算機の両者で，アクセスするデータ量の少ない小規模なベンチマークと，多量のデータをアクセスする大規模なベンチマークを実行させることを考えてみます．容量の少ないキャッシュでも，アクセスするデータのほとんどが入ってしまうような小規模なベンチマークでは，前者の計算機が良い性能を示します．逆に，大規模なベンチマークの場合，前者の計算機ではキャッシュ・ミス多発のために十分な性能を発揮できず，キャッシュ・ミスの少ない後者の計算機よりも性能が悪くなってしまうということが起こり得るのです．このように，計算機によって得意とするプログラム，不得意なプログラムがあるため，各自が計算機で実行させたい処理に近いベンチマークの結果を参考にする必要があります．とはいっても，すでに動作している計算機の性能を比較評価する目的には，ベンチマーク・テストが最も簡単であり，適しているといえます．

これに対し，設計開発段階にある計算機の性能評価（予測）は，どのようにするのでしょうか．この段階では，評価対象となる計算機が存在していないわけですから，実機でベンチマーク・プログラムを実行させる単純な方法では評価できません．したがって，評価方法の工夫が必要となる等，一般に，性能評価は困難です．しかし，計算機を設計開発する立場からすると，でき上がってから後の段階での評価よりも，この段階で性能を評価し，問題点を早めにフィードバックすることが重要です．この段階における性能評価をおろそかにすると，「でき上がった計算機が思ったとおりの性能を発揮しない」といったトラブルに直結してしまいます．

このような段階における性能評価は，対象計算機の動作を表現した評価モデルを作成して行います．評価モデルには，待ち行列網等を用いて解析的に表現したものや，シミュレータとして動作を直接表現したものがあります[2]．このうちシミュレータの方が対象計算機の動作を詳細に表現できますから，計算機の設計を評価する場合，多少，作成に時間がかかっても，シミュレータを用いて性能評価を行うことが多いようです．シミュレータをさらに分類すると，どのようなデータを与えるかによって，1)確率的に発生させたイベント・データを入力とする離散イベント型，2)既存の計算機の動作系列を記録したトレース・データを入力とするトレース・ドリブン型に分類できます．プロセッサのパイプライン方式や，キャッシュ・メモリ方式の評価には，開発している計算機と同様のアーキテクチャを持つ計算機（たとえば，1世代前の機種）から採取したトレース・データを用いるトレース・ドリブン・シミュレーションが必須です．

さて，これまで，どちらかというと，1台の計算機を1ユーザ（1本のプログラム）で使用する場合の性能を評価する話が中心でしたが，複数のユーザで共用するサーバでは，多数のユーザが並行して使用する状況下での性能も重要です．問題となるのは，1人のユーザが独占して使用するときの性能に比べ，多数のユーザで使用する場合の性能はどのようになるか，という点です．

すなわち，複数のユーザが共用するサーバでは，あるジョブを実行するためにCPUが必要になったとしても，別のジョブを実行するためにすでにCPUを使っている場合は，CPUが利用可能になるまでの間，待つ必要が生じてしまいます．サーバの性能を評価する場合，このような待ち時間を見積ることが重要となっており，これまで，待ち行列網を利用した方法が多用されています．この方法では，CPUやディスクといった計算機資源を使用する際に発生する待ちに着目して評価対象をモデル化し，ユーザ数とレスポンスやスループットの関係を評価します．このモデルへの入力となるのは，1ジョブの実行を完了するために使用するシステム資源の量（CPU時間，ディスク・アクセス時間）です．待ち行列網による性能評価では，このデータの正確さが評価精度を決める重要な要素になります．そこで，我々は，測定によって得た正確なデータを待ち行列網モデルに入力して評価を行う方法（シングル・プロファイル法[3]と命名）を開発し利用しています．

以上，説明してきましたとおり，一口に「計算機性能の評価」といっても，様々な局面があり，それぞれに応じた評価手法が用いられています．現時点では，各々の評価手法の特徴を把握した上で，これらを使い分けていく必要がありそうです．もちろん，これらを体系的に整理することも非常に重要で，徐々に行われていくものと思い

ます．

参 考 文 献

1) 富田眞治他訳：ヘネシー&パターソン，コンピュータ・アーキテクチャ，日経BP社 (1992)．
2) Heidelberger, P. and Lavenberg, S. S.: Computer Performance Evaluation Methodology, IEEE Trans. Computer, Vol. c 33, No 12, pp. 1195-1220 (1984)．
3) 堀川 隆，紀 一誠：DBMS動作特性の測定・解析手法，情報処理学会研究報告，94 ARC-104-4/94 OS 62 4 (1994)．

(平成7年3月6日受付)

堀川 隆（正会員）

1959年生．1981年同志社大学工学部電子工学科卒業．1983年京都大学大学院工学研究科修了．同年日本電気(株)入社．以来，マイクロ・プロセッサや計算機システムのアーキテクチャ，計算機性能評価の方法論やツール等の研究開発に従事．現在は同C&C研究所システム基礎研究部研究専門課長．