

# サーバ群の運用における自動負荷予測手法

株式会社アイ・アイ・エム 技術部 塩川 孝雄

サーバの運用管理をしていると、その負荷は年・月単位の周期で右肩上がりを繰り返すパターンがあることに気付いてくる。この推移特性を実運用の経験則として初めから設定すれば、稼働データの整形と近似指数曲線を用いて、個々のサーバの多様な動きに合せた負荷予測を行うことができる。

システム運用管理者がこの予測作業を表計算ソフトにより自動的に行うことで、実運用における問題サーバの未然把握に役立つか、その管理手法の有効性について検証を試みた。

## The automatic prediction of the load on server clusters for system operation managers

Takao Shiokawa, Engineering Department, IIM Corporation

Through the experience of server management, I found out the pattern of their load which shows steadily growing in both a year and a month periods. If we assume such the pattern as a empirical rule from the beginning of prediction, we can get the prediction of server's load depend on each server's behavior with the reforming of operation data and fitted exponential curves.

This can be processed automatically with spreadsheet software. We test the effectiveness of this method for system operation managers to get to know the servers that will have failures preliminarily.

### 1. はじめに

システム運用管理者が負荷変動に伴う問題を未然に回避するには、処理件数の増減に関する情報を事前に収集して負荷予測を行う必要がある。

しかし、近年に見られるような多数のサーバを少人数で管理しなければならない状況下では、処理件数の事前把握自体が容易ではないのが現実である。

そこで、稼働データの傾向だけでおおよその妥当性のある負荷予測を自動的に行い、問題発生の可能性を事前に喚起する手法が望まれるところである。

不特定多数のエンドユーザが任意にアクセスするメールなどのサーバ群は、長期的に見ると自然増的な右肩上がり、かつ年・月など周期性のある負荷傾向を示すことが認識される。

最初に、この長期漸増型の負荷推移パターンをモデル負荷として、稼働データの整形プロセスと近似指数曲線を適用して予測計算を試みる。そして、この手法が他の負荷推移パターンについても実運用で有効な管理手法になり得るかを確認する。

そのための予測管理のシナリオとして、オンラインピーク時間帯での管理値越えサーバを未然に把握することを要点として考える。また、システム運用管理者が自らその仕組み作りを行うことを想定し、日常で使い慣れている表計算ソフトにより実用性を検証する。

### 2. 負荷推移のパターン分類

サーバ群の日常運用では、一般的に「近時点における負荷異常の未然防止」と「長期的な負荷改善対策」の二つの視点からみたキャパシティ管理が行われる。

多数のサーバの管理では、蓄積稼働データによりサーバ群全体の負荷傾向を把握して、上記の視点から問題を早期に見究める何らかの効率的な管理手法が不可欠となる。

特にサーバ台数が多い場合には、利用目的や負荷状況により事前にグループ分けを行い、重点的なキャパシティ管理を行うことが重要である。

このような中で、サーバの負荷を推移パターン別に判別して分類することはキャパシティ管理を一層行い易くする。サーバの負荷を代表するCPU使用率では、次のような推移パターンに分類することができる。

長期漸増型 長年に渡り少しずつ増加

短期急増型 最近において急速に増加

一定負荷型 システムループ等一定負荷が存在

過去高負荷型 過去に高負荷が存在

年周期型 年周期で大きなピークが存在

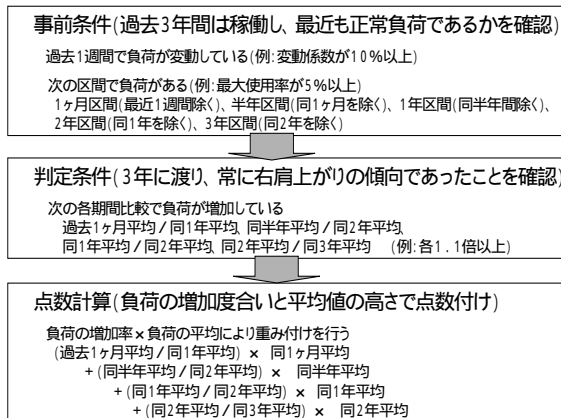
月周期型 月周期で大きなピークが存在

不定型 特に特徴が無い負荷推移

この負荷推移パターンを判別する計算式は、対象の稼働データをどのように切り出すかにより異なる。

例えば上記の長期漸増型の推移パターンでは、図-1のような手順で簡易な計算式を立てることにより、問題サーバ順に優先付けして自動表示することが出来る。[以降の試行で使用する稼働データは、オンラインピーク時間帯を想定した単位時間毎の平均CPU使用率で、土・日を除く3年間分のデータ(1時間平均×6時間/日)である]

#### 長期漸増型パターンの判定手順(3年間の例)



(図-1)

この判別作業において、パターン毎に他のサーバとの相対的な点数付けをしておく。この点数自身は絶対的なものではないが、負荷推移パターン毎に問題と思われる推移グラフを自動表示できる点で便利である。

### 3. 負荷予測計算の基礎検討

あらゆる業種に渡る数百台のサーバ群を対象に、上記の方法でパターン分類を自動的に行ってみた。

この作業を実際に行ってみると、推移パターン毎にグラフの右方向(将来方向)に延長線を引くことで、管理値を超える時期を予測できそうに見えてくる。

もし、これが可能であれば、上記の相対的な点数付けによる問題サーバの表示よりも、管理値越えまでの具体的な予測日数(予測年月日)を示すことで、より実用的な管理手法になり得ると考えられる。

以下に、負荷推移パターン毎に自動予測計算をする手法について検討した経過とその内容を述べる。

#### (1) 長期漸増型モデルによる検討

右肩上がりに負荷が増加している長期漸増型のサーバの多くは、エンドユーザが任意にアクセスするメールやOA業務用などのサーバである。

この長期漸増型の負荷推移パターンのサーバについて3年程度の稼働実績をみると、

高負荷になるに従って増加傾向が高まる  
強弱はあるが1年単位の周期性がある  
負荷増に伴い振幅が拡大する

といった共通の特徴が見受けられる。

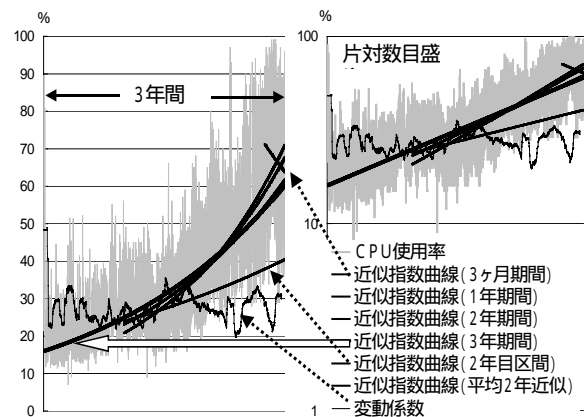
特に、統計的な視点でみると「平均値の変化に関係なく変動係数(CV)がほぼ一定」であるという特異な特徴とみることができる。

一般に、多くのエンドユーザが利便性を追求して任意に使用するサーバの負荷は、自然増的な要素を持つことから、統計的な手法によりある程度の予測が可能と考えられる。

そこで、上記の負荷推移の特徴を近似できる曲線として指数曲線「 $y = b m^x$  : yはCPU使用率、xは推移時間のカウント、bおよびmは定数」を適用してみることにした。

理想的なモデル負荷として、3年間を通して稼働し、右肩上がりでも近時点で管理値越え(例:CPU使用率90%4時間連続「以降の全試行で適用」)に至ったサーバの負荷推移を選定した。

このモデル負荷の分布が指数関数分布であるかどうかを確かめるために、普通の方眼グラフおよび片対数の方眼グラフで視覚的に確認したものが図-2である。



(図-2)

この結果、過去3年、2年および1年の各期間の稼働期間データにより作成した各近似指数曲線はほぼ同じ線に沿っており、全体の負荷分布と近似していた。また、近似の精度を表す決定係数( $R^2$ )は直線による近似に比べ明らかに大きかった。

なお、最近1年間を除く2年前の1年間(以下2年目区間)と過去3ヶ月期間は異なる方向を示していたが、これは負荷の年周期性および短期間データのためによるものと考えられた。

以上から、適当な近似指数曲線を選択することにより、翌年以降の負荷分布をある程度予測できるものと考えられる。

## (2) ピークの予測方法の検討

次に、ピークが管理値を超えるまでの負荷変動の様子を把握する方法として、上記の近似指数曲線の作成に加え、次のような工夫を行うこととした。

先に示したように、長期漸増型の特徴である「年周期性」平均値の変化に関係なく変動係数がほぼ一定であることを利用し、

前年同日に該当する実負荷と近似曲線との誤差値を算出する

その誤差値に近似曲線の伸び率を掛ける

その値を将来の該当日の近似曲線に対する予測誤差値としてプロットする

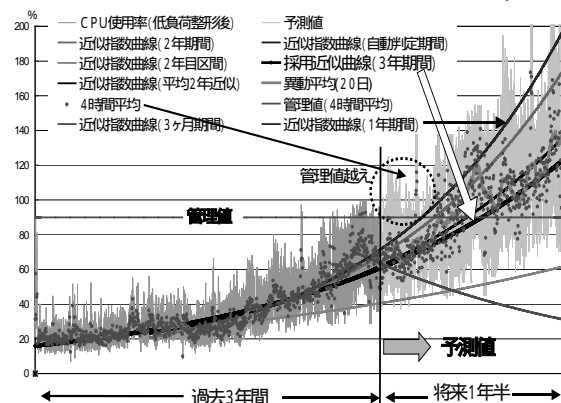
という計算手順により将来の負荷予測グラフを作成した。

また、実運用での設備改善検討などで必要と思われる3年先までの期間をシート上で計算して、管理値越えまでの日数をグラフに併記することとした。

これにより、他サーバの管理値越えまでの日数と比較し、日数の短い順に該当サーバの予測負荷グラフを自動表示することが出来る。

図-3は以上の検討に基づき、モデル負荷について3年期間の近似曲線により実際に負荷予測(CPU使用率100%を超えて予測)をしたものである。

機呼称: 13 局呼称: 社2 期間: 2004年11月20日 時刻: 09:58 2004年11月18日 より作成 単位: 30 超過予測 2004年11月23日 (115日後)



(図-3)

## 4. 予測計算の汎用化検討

上記の3項で検討した負荷予測の計算方法を基に、他の負荷推移パターンへの適用を試行して、汎用性を持たせることができるか以下に検証を行った。

この汎用化が可能であれば、全ての負荷推移パターンについて自動的な負荷予測ができ、より有効な管理手法になり得ると思われる。

## (1) 長期漸増型、過去高負荷型への適用

### a. 稼働データの事前整形

前項3で対象としたモデル負荷のように理想的な3年間の稼働データを採取できるケースはまれである。

近似指数曲線を作成する際に、祭日などの負荷が極端に低い日や稼働データの欠損が存在する場合は、負荷予測に影響しないようにデータを事前に整形しておく必要がある。

例：過去1ヵ月平均の30%（一定値）に満たない時は、その平均値に合せて該当データを整形

この他に、過去高負荷型の場合は何らかの理由で負荷改善がされているが、前後の平均値により整形することが出来れば、改善前のデータも予測計算の基礎データとすることが出来る。これは、整形すべき箇所の判定に使用する定数を適切に設定することである程度可能であった。

例：月単位に平均値をチェックし、過去に5ヶ月間連続で1.3倍（一定値）の該当月を判定

### b. 年周期による負荷のゆらぎの除去

サーバの負荷が年周期を有していることを考えれば、近似曲線の作成で採用する稼働データの期間毎にその予測の方向はゆらぐことになる。

3ヶ月期間、1・2・3年期間および2年目区間の5つの近似指数曲線について、長期漸増型および過去高負荷型の複数のサーバを対象に傾向を確認した結果、おおそ次のようであった。

#### 3ヶ月期間

年周期性に対し期間が短すぎ、近似曲線の方向が極端に変化する。短期予測用として活用。

#### 1年期間、2年目区間

年周期のピークが大きき場合、その予測時期により近似曲線の方向が大きく変化する。

#### 2年期間

に比べ方向の変化は減少するが、ピークが大きき場合は予測曲線が変化する。

#### 3年期間

に比べ方向の変化は減少する。しかし、3年間の間に負荷改善などによる変動が発生して予測用の基礎データとして使用に耐えられないケースが多い。経過確認用として活用。

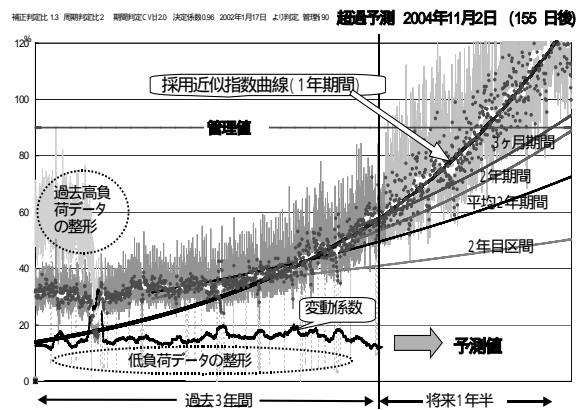
以上から、年の周期性が少ない場合には2年期間を採用した近似でも良いと思われるが、近似曲線の作成

時点で年周期によるゆらぎを完全に除去することにした。

これは、2年目区間および1年区間の各1年間の2つの平均値を計算し、個々の該当負荷をこの数値に置き換えた後に、2年間の平均近似指数曲線を作成することで可能である。

この平均2年近似指数曲線を採用する効用については、後に述べるピーク負荷の顕著な年周期型パターンで再確認する。

図 - 4 は過去に高負荷のある長期漸増型の負荷パターンでの予測例である。



( 図 - 4 )

## (2) 短期急増型および不定型への適用

### a. 有効な稼働データ期間の自動判定

稼働を開始して期間の短いサーバはわずかな稼働データしか存在しない。

このため、自動予測計算における重要なデータ整形のもう一つの方法として「近似曲線の作成で採用する稼働データの期間を自動判定する方法」を検討した。

これには、長期漸増型での特徴で述べた「平均値の変化に関係なく変動係数がほぼ一定である」という特性を利用する。変動係数の値がある範囲から大きく変化（例えば2倍以上）する時点を検出する方法をとることとした。

この結果、大きな負荷変動やデータ採取がされていない過去の時点から現在までの期間を近似曲線作成の採用期間として判定することが可能となった。

次に、表計算ソフトのシート上のセル間で3年区間を連続計算して上記の採用期間を判定する計算式の例を示してみる。

< 採用期間判定式の事例（一部） >

計算式の主要部分を言葉で表現すると、「(過去の該当日時から更に1ヶ月前の前後1ヶ月間の変動係数)を(現在から該当日時までの変動係数の平均)で割った値が一定値以上(N)となる日時までを近似曲線作成で採用する期間とする」となる。

例えば全3年間(約5700セル)のうち700番目のセルに設定する計算式は次のようになる。

「=IF(R830<>0, IF(R830/AVERAGE(R\$2:R700)>予測グラフ!G\$1,3,2),0)」

「R700」「R830」は該当日時の前後1ヶ月間変動係数のセル名、「予測グラフ!G\$1」は変動係数の変化を判定する一定値(N:例えば2.0)のセル名である。

### b. 自動判定期間の近似を含む近似曲線の選択

上記により、有効な稼働データの期間を判定して近似曲線を作成できることから、複数の近似曲線の中から次の基準による期間での近似曲線を採用して負荷予測をすることとした。

自動判定期間が2年区間より長い場合

平均2年区間を基本として採用する、但し、1年区間がより上方への傾向にある場合は1年区間を採用する。

自動判定期間が1年以上2年未満の場合

自動判定期間を基本として採用する。但し、1年区間がより上方への傾向がある場合は1年区間を採用する。

自動判定期間が1年未満の場合

自動判定期間を基本として採用する。

3ヶ月区間がより上方傾向にある場合は3ヶ月区間を採用することも考えられるが、短期で不安定な場合が多いため採用しない。

なお、近似曲線の方向比較は、前述した指数曲線計算式の係数(m)の大小比較により簡易的に行うこととした。

ここで、上記の基準において上方傾向の近似曲線を選択する理由は、管理値超えの予測期日を早めることで、問題サーバの発見を早期にすることを配慮したためである。しかし最終的にどの近似指数曲線を使用するかは、ユーザのシステム環境に合わせてシステム運用管理者が決定すべきものと考えられる。

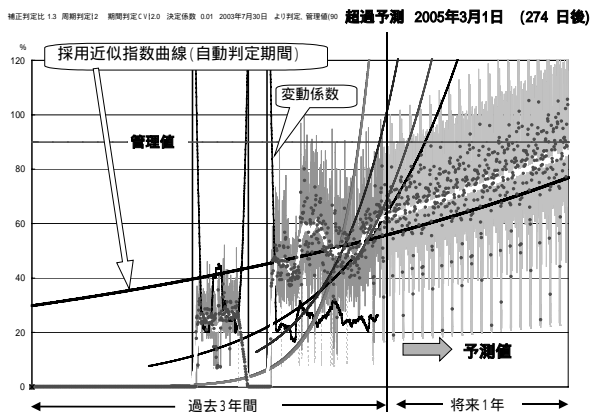
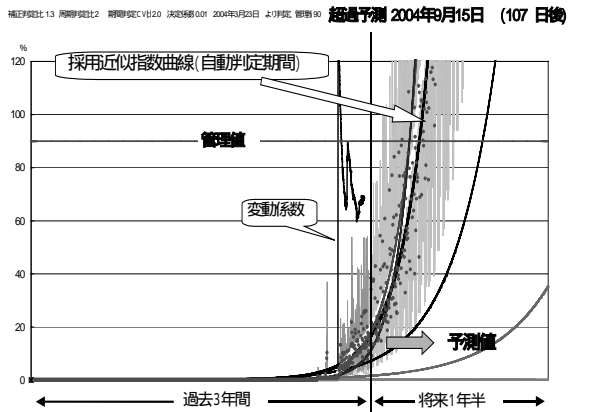
このため、各近似指数曲線(6本)を推移グラフに併記して、システム運用管理者が視覚的に予測の方向の妥当性を確認し易くする方法にした。

以上のほか、3項(2)のピークの予測方法で述べた近似曲線と実績値の差による予測値の展開については、短期間での予測が可能なよう次のようにすること

とした。

- 自動判定期間が1年期間より長い場合  
前年同日に該当するデータを年展開する。
- 自動判定期間が1年期間より短い場合  
前月同日に該当するデータを月展開する。

図 - 5 に短期急増型、不定型の負荷パターンの予測例を示す。



(図 - 5)

### (3) 月周期、年周期型への適用検討

月周期型の負荷推移パターンは、これまで汎用化した予測手法をそのまま適用して予測グラフを得ることができた。

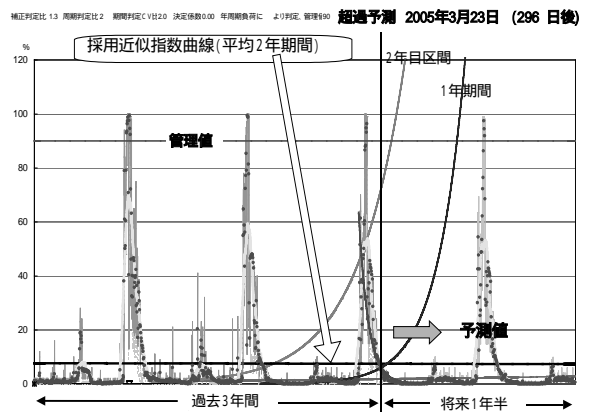
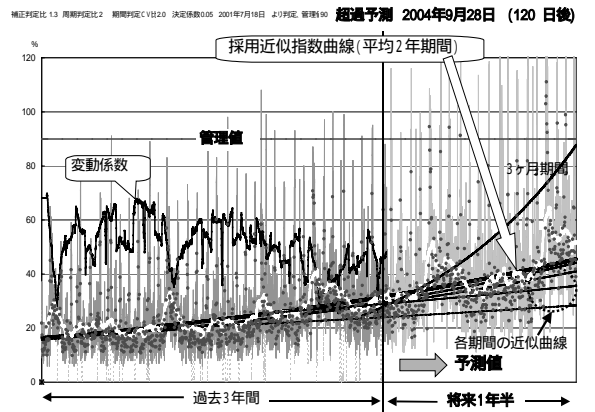
ピークが顕著に現れている年周期型の負荷推移パターンについても、先に4項(1)bで整理したように、1年間毎の平均負荷でならして近似曲線を作成することで、ほぼ妥当と思われる予測が可能であった。

しかし、この機能を自動的に行う場合には問題があり、工夫が必要であった。それは、顕著なピーク負荷がある場合、変動係数も高くなることから、そのピーク時点を有効ではない期間として優先判定をしてしまうためである。(例：通常の変動係数は40%以下、ピーク時の変動係数は80%以上)

この対処として、近似曲線の採用期間の判定を無視して平均2年指数近似曲線より負荷予測を行う機能を追加した。(年周期型かを事前に判定する：平均対ピークの負荷比2倍)

これにより、年周期型の負荷パターンについても汎用的に負荷予測を行えた。

図 - 6 に月および年周期型の予測事例を示す。



(図 - 6)

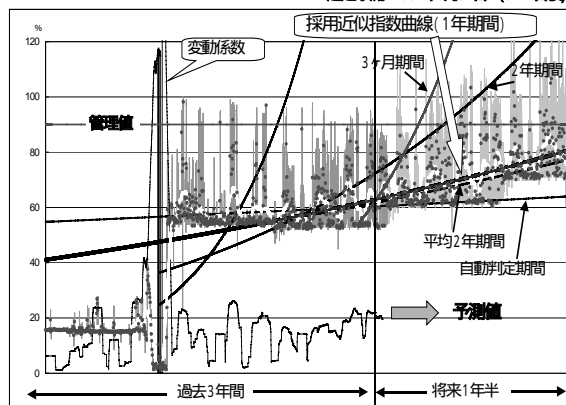
### (4) 一定負荷型への適用

一定負荷型の負荷推移パターンは、システムループが内在しているか、一部のアプリケーションプログラムなどの負荷がCPUを占有している等の問題を既に含んでいると考えられる。

このため、負荷推移パターンの分類の段階で問題サーバとしてアラート扱いにして良いと思われる。

しかし、一定負荷型では予測計算がややそぐわないものの、割り切って他の負荷推移パターンと同様に自動予測を行うことで、ある程度の予測効果と実運用での効率化が期待できる。

図 - 7 に一定負荷型パターンの予測例を示す。



(図 - 7)

## 5. 実用面からの考察

本稿の負荷予測手法は、問題サーバの未然把握を容易に行えるよう汎用性に重点を置いた。

その汎用性を持たせるための方法として、

稼働期間の判定 (変動係数の変化: 例 2 倍)

過去高負荷部の判定 (負荷減比: 例 1.3 倍)

年周期型の判定 (平均対ピークの負荷比: 2 倍)

のように一定の値 (定数) の設定を行ったが、ユーザによっては、これらの設定値の調整が必要である。

例えば、上記の過去高負荷部の判定では「負荷改善率が低い場合の負荷減」と「周期性による緩やかな負荷減」の間の判別が難しい。このため「負荷減比の定数」を大きくし、不要なデータ整形を避けようとすると、逆に過去高負荷型の整形が出来ず、結果として本来よりも下方向に予測しすぎることになる。

これを定常的に改善するには、「負荷の周期性による負荷減」と「人為的な要素による単発的な負荷減」を判定する新たな方法を検討することが考えられる。

おそらく、サーバ負荷の実態を更に見極め「負荷推移パターン別分類」「データ整形」のプロセス改善により、汎用性と予測精度に関わる同類の問題を改良していけると思うが、今後の継続的な課題といえる。

最後に、本稿で試行した負荷予測手法での処理速度および仕組み作りに要した工数について参考として載せた。

### (1) 予測計算の処理速度

使用したパソコンは4 CPU、2 GHz、5.24 MB RAMである。また使用した表計算ソフトはMicrosoft社のExcel 2000™である。

予測計算の速度は、1サーバ当たり15秒程度であ

った。しかし、シートでのデータ展開、計算式やマクロおよびグラフ式により、1サーバ当たりの表計算ファイルは13MBと大容量であった。このためディスクへの読み書きを含め30秒程度を要した。

このため、多数のサーバを対象とする場合は、事前に負荷パターン分類(500台で30分程度)を行い、優先順位付けの高いサーバから予測計算するなど、効率的運用を工夫する必要がある。

### (2) 自動化の工数

予測計算方法の検討1ヶ月、表計算による自動化作業1ヶ月、検証テスト1ヶ月の合計3ヶ月を要した。

実際にユーザ環境へ適用する場合は、稼働データの常時採取の設定や本稿内で述べた各定数の調整期間が必要である。

## 6. おわりに

本稿で検証した負荷予測手法は簡易的なものであるが、システム運用管理者が机上で鉛筆を舐めながら作成する負荷予測グラフに近似している。

これは、近似曲線と実際の稼働値の誤差を組み合わせ、グラフ上でプロットしたことで、あたかも実績値のような予測値を表現できたことによると思われる。

統計手法としては多少の不満もあるが、一般にサーバ負荷が年周期性を有した企業活動の結果であることを考えれば、実運用上ではこの程度の予測手法でも効果があると思われる。

なお、本稿での予測管理手法を実際に適用する場合は、各ユーザのシステム運用環境に合わせた対象データの期間・時間帯、管理値などのシナリオ設定が必須となる。これは、予測計算の条件設定として必要なものであるが、設定値の推奨化や予測値のフィードバック機能などの改善を図り、更に、自動負荷予測手法を汎用化していくことが、今後の大きな課題となる。

以上

### 参考文献:

- 基本統計学「有斐閣ブックス:宮川公男」
- ゆらぎの世界「講談社:武者利光」
- 蓄積稼働データによるキャパシティ予測管理へのアプローチ「情報処理学会研究報告 2004-EVA-11:塩川孝雄」