グリッド環境における通信遅延時間の確率分布を用いた集団通信時間の推定手法

甲斐島 武[†] 加藤 精一[††] 秋山 豊和[††]
野﨑 一徳[††] 水野（松本）由子[†††] 下條 真司[†,††]

これまでに計算およびネットワーク資源性能の動的変動性を正規分布で近似し並列アプリケーションの実行時間を推定する研究が行われているが，広域ネットワークにおける通信遅延時間は裾の長い確率分布となることが知られている．本稿では，通信遅延時間の確率分布に裾の長い分布であるパレート分布を適用することによって，グリッド環境をはじめとする広域分散計算環境における集団通信時間の推定方法を提案する．その後パレート分布を用いる推定手法との比較を行い，正規分布を用いる従来手法は，プロセッサ数の大きい場合に集団通信時間を過小に推定していること，および集団通信時間推定に必要な計算量が大きいことを解析的に示した．さらに，広域ネットワークにおける往復遅延時間データを用いて評価を行い，通信遅延時間の階段状の偏位が少なく推定時点からの経過時間が大きい範囲で提案手法が推定精度を向上させることができることを示した．

A Method for the Estimation of Collective Communication Time Using Probability Distribution of Communication Latency in Grid Environment

Takeshi Kaishima,[†] Seiichi X. Kato,[††] Toyokazu Akiyama,[††]
Kazunori Nozaki,[††] Yuko Mizuno-Matsumoto[†††]
and Shinji Shimojo[†,††]

There are many researchers on the estimation of execution time of parallel applications, and most of them assume that probability distributions of processor and network performance follow the normal distribution. However, many reports suggest that the communication latency in a wide-area network tends to follow a long-tailed distribution. In this paper, we propose an estimation method of collective communication time in grid environments and other wide-area distributed parallel environments using the Pareto distribution, a type of long-tailed distribution, as communication latency. As a result, we could analytically indicate that when compared to the Pareto distribution, the conventional method using the normal distribution tends to underestimate the collective communication time among many processors, and it needs more computational complexity of estimation. Furthermore, we evaluated the proposed method with the latency data of a wide-area network. The result showed that our method could improve the estimation accuracy of the collective communication time where there are few step changes in latency data and the communication is executed in a short period of time after the estimation.

1. はじめに

並列計算技術および計算機性能の向上により，気候変動シミュレーション[31)]等の地球環境科学分野や蛋白質の立体構造解析[8)]をはじめとする生物情報学分野の大規模計算が１組織内でも可能となってきている．しかし，現在運用可能な規模のスーパコンピュータや並列計算機では，研究者等のユーザが要求する規模の計算要求を満たすことが困難となってきている．たとえば，我々はこれまでに広域分散計算環境上で脳磁図信号解析による脳機能診断[18)]や，流体シミュレーションによる歯茎摩擦音解析[21)]といった大規模アプリケーションの実証実験を行ってきているが，実際の医療応

† 大阪大学大学院情報科学研究科
Graduate School of Information Science and Technology, Osaka University
†† 大阪大学サイバーメディアセンター
Cybermedia Center, Osaka University
††† 兵庫県立大学大学院応用情報科学研究科
Graduate School of Applied Informatics, University of Hyogo

用に必要な規模の計算量を得られていないのが現状である．そこで，他の組織が持つ計算資源の余剰能力を適切に利用しより大規模なシミュレーションを実現するために，グリッド技術をはじめとする広域分散計算技術の必要性が高まってきている[9)]．

このような広域分散計算技術を用いた大規模計算アプリケーションの例として，中期気象変動のアンサンブル予測[27)]，分子運動のモンテカルロ・シミュレーション[4)]，薬物候補化合物のスクリーニング[3)] 等のパラメータ・サーベイ型アプリケーションがあげられる．これらのアプリケーションでは広域ネットワークによる通信遅延時間の影響を抑えるため，可能な限りプロセッサ間の通信を低減させることが行われている．しかし，分子動力学計算[13)] やヤコビ繰返し法による線形連立方程式解法[22)] 等の同期型アプリケーションにおいてはプロセッサ間の頻繁な通信が必要となるため，これまで提案されてきた広域分散計算技術の手法では有効な並列化効率が得られない．効率低下は主に計算およびネットワーク資源の共有による資源性能の動的変動性に起因するため，性能や負荷状況を調査し個々の計算資源やネットワーク資源を含む全体の計算資源性能を推定した後に，適切なプロセス配置を決定する必要がある．したがって，広域分散計算環境上の計算資源を利用しユーザの計算要求を満たすためには，資源性能の動的変動性を考慮した正確な実行時間推定が必要である．

計算資源の性能および負荷状況の推定には，過去から計測してきた性能および負荷値からなる履歴を用いて性能および負荷値の発生頻度を何らかの確率分布に近似して行うことが一般的であり，これまでに提案されてきた性能推定手法は計算資源の性能および負荷値の確率分布が正規分布に従うと仮定しているものが多い[5),20),25)]．履歴から求められる平均値を性能や負荷の代表値とし標準偏差を動的変動分として扱うことがその代表例であり，同時に使用するプロセッサ数の小さい計算環境においては，数式の取扱いが容易な正規分布は有用である．

しかし，プロセッサ数が多く計算資源性能の動的変動性が大きい広域分散計算環境において計算資源の負荷や利用率は複雑な確率分布となることが多く，正規分布では近似できないことが示されている．一般に，都市規模，所得，単語出現頻度，地震規模等の社会的現象あるいは自然現象で観測される多くの確率分布は，小さな度数を持つ多数の事象と大きな度数を持つ少数の事象が共存する「べき法則」に従うことが知られている[1)]．「べき法則」に従う現象は大きな度数を持つ少数の事象により大勢が決定されることから，この確率分布の裾部分の正確な扱いが可能なパレート分布が地震学，土木工学，都市計画，保険等の分野で用いられている．同様に，パケット発生間隔の大きな分散を持つ通信が多数集約される広域通信においては，通信遅延時間の確率分布は裾の長い分布となることが報告されており[19),23)]，裾部分はパレート分布による近似が適当であることが示されている[11),17)]．並列計算や分散計算において，データの同報（broadcast）や簡約（reduce），分配（scatter），収集（gather）等を含む複数のプロセス間における通信を集団通信（collective communication）と呼び[22)]，複数のプロセッサへの通信時間の最大値により実行時間が決定されるため，発生頻度は小さいが大きな値をとる確率分布の裾部分の影響が無視できないほど大きくなる．

そこで本稿では，広域分散計算環境における同期並列型アプリケーション実行で大きな割合を占める[28)] 集団通信時間の推定精度を向上させるために，広域ネットワークにおける通信遅延時間に着目し，通信遅延時間の確率分布に裾の長い分布であるパレート分布を適用することによって，広域分散計算環境における集団通信時間の推定方法を提案する．その後，実際に観測された通信遅延時間データを用いた評価を行い，提案手法の特性および有効性を検討する．以下，2 章で周辺技術との比較を通じて，本稿で対象とする性能推定の問題を明らかにする．3 章では提案する性能推定手法について述べ，4 章で提案手法についての評価を行う．最後に 5 章でまとめと今後の課題について述べる．

2. 性能推定の周辺技術と問題

大規模計算の並列アプリケーションは，プロセス間のデータ依存性に着目し以下の 5 種に分類できる[29)]．

(1) **同期（*synchronous*）型**：多数のデータに対する均一な更新を行う問題（次の時点でのデータを求めるために現時点のすべてのデータが必要となる問題）

(2) **緩やかな同期（*loosely synchronous*）型**：繰返しごとの緩やかな同期と不均一なデータ更新を行う問題

(3) **驚異的並列（*embarrassingly parallel*）型**：各データが互いにきわめて独立している問題

(4) **非同期（*asynchronous*）型**：データ更新において明確な同期アルゴリズムが存在しない問題

(5) **複合（*metaproblem*）型**：上記 4 種類の複合型

それぞれ必要な並列化手法，並列アルゴリズム，プログラム記述性が異なり，それにともない適切な性能

推定を行う必要がある．このうち実行前後のデータ転送以外の通信が不要な驚異的並列型アプリケーション[3),4),27)]は，計算資源の動的変動性が大きい広域ネットワークの通信性能の影響を受けにくく，広域分散計算環境の資源を容易に活用することができる．一方，分割したプロセスどうしの同期や通信等の相互作用を必要とする同期並列型アプリケーション[13),22)]は，高速計算が必要とされる科学や工学におけるアプリケーションのうち約 70 % を占めるという報告がある[10)]．広域分散計算技術の適用分野を拡大するうえで，広域ネットワークを考慮した同期並列型アプリケーションの実現は重要な課題の 1 つである．

同期並列型アプリケーションを計算資源の動的変動性が大きい広域分散計算環境上で効率的に実行するためには，アプリケーションの実行順序や計算資源への配置を行うスケジューリングが必要である．資源性能や課金額等の制約条件を満たしながら実行時間やシステム効率等を最小化または最大化するための目的関数の違い，および広域分散計算ミドルウェアやオペレーティングシステムに対する動作の違いによって，様々な手法が提案されている．ただし，いずれのスケジューリング手法においても，現在および将来における実行性能および負荷を推定しそれに応じて時間的・空間的なアプリケーション配置を決定する必要があるため，性能推定の精度が高いほど実行時間やシステム効率等を向上させることが可能である．

並列計算における実行性能推定は主に，実際に実行を行い各命令ごとに要した時間を記録したトレースログを用いる方法と，トレースログを用いないで各命令間の依存関係と資源性能を分離し抽象化して行う方法の 2 つに分けられる．後者はさらに，各命令間の依存関係を抽象化し解析を行う方法と，資源性能の測定と推定を行う方法の 2 つに分けられる．

(1) **トレースログを用いる方法**：特定のプログラムについてチューニングを行う場合によく用いられる[32)]．実際の実行時間が得られるため正確であるが，環境が異なればトレースデータを測定しなおす必要があり可搬性に欠ける．

(2) **トレースログを用いない方法**

(2a) **プログラムの抽象化を行う方法**：個々の資源性能値をパラメータとして定式化し計算全体の実行性能の解析を行う方法である[6)]．いったん定式化を行えば環境が変わってもその式を適用できるが，定式化が十分でないと誤差が大きくなってしまう．スケジューリングに用いるためには次に示す資源性能値を代入する必要がある．

(2b) **各計算資源の性能推定を行う方法**：時々刻々と変化する個々の資源性能を測定および推定する方法である．測定項目はCPU負荷やネットワークの通信遅延時間等の代表的なものがよく用いられるが，プログラムの性質によって適切なベンチマークを用いる場合もある[7)]．推定には測定値の履歴を用いて線形時系列予測を行う手法が用いられている[30)]．

広域分散計算環境においては環境の変化が避けられないため，トレースログを用いる方法よりもプログラムの抽象化を行う方法が優れている．後者 2 つは実際利用のためには不可分であるが，広域分散計算環境においては資源性能の変動要因が多岐にわたり非線形的な挙動を示すため[23)]，特に長時間予測において，決定論的な線形時系列予測を行う方法よりも，各計算資源性能の動的変動性を確率過程と見なし統計値を用いて表し，それら統計値をパラメータとして性能解析を行う方法が優れていることが報告されている[25)]．このとき資源性能値の発生確率を何らかの確率分布で近似する必要があり，数学的取扱いの簡便性から，正規分布と仮定し平均値を代表値，標準偏差を変動成分と見なして性能推定を行った研究が多い[5),20)]．

広域分散計算環境における同期並列型アプリケーションの一般的な集合通信パターンを考慮すると，全体の実行時間は各プロセッサでの計算時間および各プロセッサへの通信時間の和の最大値により決定されるため，確率分布の裾部分が大きな影響を与える．実際，広域通信における通信遅延時間の確率分布は裾の長い分布となることが報告されている[19)]．一般に，正規分布をはじめとする裾が急激に減衰する確率分布は外れ値に対して敏感に反応するのに対し，パレート分布をはじめとする裾がべき乗的に減衰する確率分布は，データに外れ値があっても安定した推定値を与えるという利点を持つ．また，従来の正規分布を用いた性能推定では各確率分布の裾が短いため，裾の長いパレート分布を用いた性能推定と比較して，広域分散計算環境における全体の実行時間を過小に推定するという問題が発生しうる．

この問題を模式的に表したものが **図 1** である．図左は，確率変数 x に対するパレート分布および正規分布の確率密度関数 $p(x)$ であり，両確率分布の期待値および分散は等しい．これと同一の確率分布をとる複数の過程 $\{x_i \mid 1 \leq i \leq n\}$ が独立に生起するとき，それらの最大値が作る確率密度関数 $p(\max_{1\leq i\leq n} x_i)$ は図中央および図右となる．確率分布の数 n が大きくなるにつれて，正規分布よりもパレート分布が作る

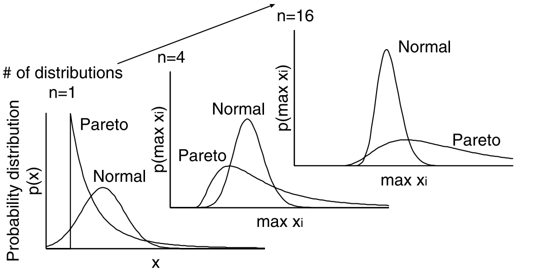

図 1 確率分布数の増大に対する最大値の確率密度関数
Fig. 1 PDF of max. value vs. # of distributions.

最大値の確率密度関数 $p(\max x_i)$ の裾部分の方が大きくなりその期待値 $E[\max x_i]$ が大きくなることを示している．すなわち，一般にプロセッサ数 n の大きい広域分散計算環境において，従来の正規分布を用いて推定した集団通信時間は，パレート分布を用いて推定した集団通信時間と比較して過小となることが予想される（3.2 節で詳述）．

3. 提案手法

パレート分布は確率変数 x に対して確率密度関数 $p(x)$ が

$$p(x) = \alpha k^{\alpha} x^{-\alpha-1} \quad (\alpha>0,\ k>0,\ x\geq k) \quad (1)$$

で与えられる確率分布であり，正規分布と比較して裾が長いのが特徴である（図 1 左）．ここで k は位置パラメータと呼ばれるもので確率変数のとりうる最小値を表し，α は尺度パラメータと呼ばれるもので減衰の傾きを表す．パレート分布の期待値 $E[x]$ および分散 $V[x]$ は

$$\begin{aligned} E[x] &= \alpha k/(\alpha-1) & (\alpha>1) \\ V[x] &= \alpha k^2/(\alpha-1)^2(\alpha-2) & (\alpha>2) \end{aligned} \quad (2)$$

である．

3.1 対象とする環境および推定の方針

一般的に，アプリケーション中において計算部分のみの性能推定に関する研究は多数行われており，特に驚異的並列型アプリケーションは広域分散計算環境においても実用可能な程度の負荷および計算時間の平均化が実現されている．これに対し，同期並列型アプリケーションでは計算および 1 対 1 通信に占める時間に比して集団通信に占める時間の割合が大きく[28]，通信性能の推定精度が実行時間の精度に大きく影響することが多い．このため，本稿では各計算資源上の計算時間は等しいものとし，以下では集団通信時間のみの推定を行うこととする．ただし，計算部分の時間をパラメータ c として与え，計算時間と通信時間の和の確率密度関数を $p(x)=\alpha k^{\alpha}(x-c)^{-\alpha-1}$ とすれば，計算時間を同時に考慮しながら以下の議論がそのまま適用可能であるが，本稿では計算部分については考慮しない．

実際の同期並列型アプリケーションの集団通信時間推定に際して，ネットワーク特性や通信プロトコル等を考慮する必要があるが，本稿では以下のような通信パターンについて検討した．

（1） 通信遅延の内訳は伝播遅延とキューイング遅延，再送遅延が主で，伝送遅延およびプロセッサ内の処理遅延は無視できる．

（2） 集団通信の内部的通信形態は 1 段同時通信である．

前者は，一般的な広域分散計算環境の通信遅延特性[16]を考慮したためである．後者については，通信遅延時間が小さい環境では多段の木構造通信で 1 プロセスあたりの通信数を低減させることが効果的であるのに対し，通信遅延時間の大きい環境ではある 1 つのプロセスを中心として 1 段で同時通信を行う方が集団通信時間が小さい[2] ことを考慮したためである．同様に，通信遅延時間の大きい環境において全縮約（allreduce）や全収集（allgather），全交換（all-to-all）等の全対全集団通信を行う場合も，多段であるバタフライ構造の通信よりも 1 段同時通信で簡約の後同報あるいは収集の後分配を行う方が通信時間が小さい．また，複数のクラスタにより構成された環境においては，クラスタ内およびクラスタ間の通信を階層的に行う実装がよく用いられる[14]．この場合は，同一クラスタに属する複数プロセスを 1 つにまとめて考慮し，クラスタ内通信遅延時間をパラメータ c_i として与えそれらの通信遅延時間の確率密度関数を $p_i(L_i-c_i)$ とすれば，以下の議論がそのまま適用できる．

このとき，同期並列型アプリケーションの集団通信のガントチャートは**図 2** のようになり，ルートプロセッサを中心にして行う．往路および復路の片側遅延時間が得られる場合は個々に推定した方が高精度の解析が可能であるが，本稿では通信遅延に往復遅延時間

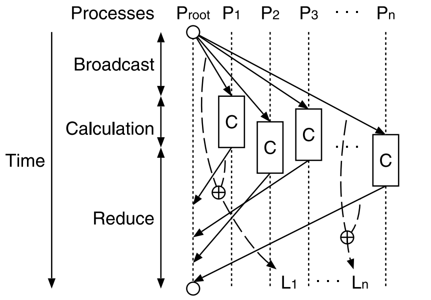

図 2 広域分散計算環境における集団通信のガントチャート
Fig. 2 Guntt chart of collective communications.

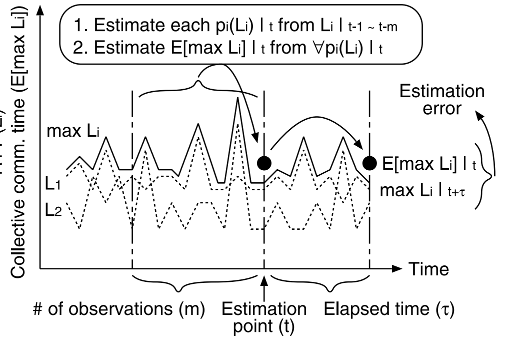

図 3　集団通信時間の推定および精度評価の手順
Fig. 3　Diagram of estimations & evaluations.

（Round Trip Time: RTT）を用いることとした．ただし，ルートプロセッサとその他プロセッサを別々に考慮するため，ルートプロセッサ自体は計算を行わないこととし，ルートプロセッサ自体も計算を行う場合はプロセッサ内の通信遅延時間を与えることとした．以下では，ルートプロセッサを除くプロセッサ数 n のときの，ルートプロセッサから第 i プロセッサへの通信遅延時間 L_i とその確率密度関数 $p_i(L_i)$ を用いて，全通信遅延時間の最大値 $\max_{1\le i\le n} L_i$ で表される集団通信時間を推定する方法について議論する．

以下，時刻を明示する場合に時刻 t における値を $\cdot|_t$ で表記すると，本稿で提案する手法は，通信遅延時間の観測値 $L_i|_{t-1\sim t-m}$ を用いて通信遅延時間の確率密度関数 $p_i(L_i)$ を推定し，それらを用いて集団通信時間の期待値 $E[\max L_i]|_t$ を推定する手法である．推定および評価の手順を**図 3** に示す．ただし，アプリケーション実行までの待ち時間やアプリケーション実行時間の経過のため，推定時刻から実際の通信が行われるまでに時間遅れが生じる．この時間遅れは，アプリケーション実行時間やスケジューリングの頻度等により決定され，一般的な広域分散計算環境では数分から数十時間程度となる．このとき，推定時刻 t から経過時間 τ を経た時刻における実際の集団通信時間 $\max L_i|_{t+\tau}$ と推定値 $E[\max L_i]|_t$ は必ずしも致せず，これらの値の差が推定誤差となるため，この推定誤差を用いて本手法の評価を行う．なお，集団通信時間 $E[\max L_i]|_t$ の推定に通信遅延時間の最大値の観測値 $\max L_i|_{t-1\sim t-m}$ を用いていないのは，各通信遅延時間 L_i が独立な過程から生起しておりそれらの最大値 $\max L_i$ の過程はより複雑であると考えたためである．

本稿では，確率密度関数 $p_i(L_i)$ については以下の特徴を持つと仮定して推定手法の検討を行う．

（1）　パレート分布に従う．

（2）　観測期間 m 中は定常．

（3）　各確率分布間は相互独立．

まず，第 i プロセッサへの通信遅延時間 L_i の確率密度関数 $p_i(L_i)$ および累積密度関数 $P_i(L_i)=\int_{-\infty}^{L_i} p(x)dx$ がパレート分布

$$
\begin{aligned}
p_i(L_i) &= \alpha_i k_i^{\alpha_i} L_i^{-\alpha_i-1} & (L_i \ge k_i)\\
P_i(L_i) &= 1 - k_i^{\alpha_i} L_i^{-\alpha_i} & (L_i \ge k_i)\\
p_i(L_i) &= P_i(L_i) = 0 & (L_i < k_i)
\end{aligned}
\tag{3}
$$

に従うとする．ただし，パラメータ推定に用いる履歴期間 m 中は定常であることを仮定するが，実際の計測データでは満足しない場合もある．このため，最大値の期待値推定に必要な最低条件を考慮して，各パレート分布の平均値が存在する条件 $^\forall\alpha_i > 1$ を仮定する．さらに，これら複数の確率分布を用いて解析する場合に各確率分布間の独立性が問題となるが，本稿では各確率分布はすべて互いに独立であるとした．ネットワークトポロジにおいてインタフェースや経路を共有する部分では依存性が生じるが，広域分散計算環境における通信遅延時間の内訳は伝播遅延が主で，かつ経路長が大きいため共有部分が少ないことを考慮したためである．

3.2　集団通信時間の解析

定常性を仮定すると履歴期間中の $\max L_i$ の期待値が推定時点の直後も継続すると考えられ，以下ではこの期待値 $E[\max L_i]|_t$ を用いて解析を行う．ただし，各 L_i について時間発展的な推定手法が適用可能ならば推定時点 t から τ 経過した時点の $^\forall L_i|_{t+\tau}$ から期待値 $E[\max L_i]|_{t+\tau}$ を求めることも可能である．

各確率分布間の独立性より，期待値 $E[\max L_i]$ は $\max L_i$ と全確率密度関数 $p_j(L_j)$ との積をパレート分布の定義域 $k_j \le L_j < \infty$ の区間で定積分することにより求められ，

$$
\int_{k_1}^{\infty}\int_{k_2}^{\infty}\int_{k_3}^{\infty}\cdots\int_{k_n}^{\infty} \max_{1\le i\le n} L_i \prod_{j=1}^{n} p_j(L_j)dL_j \tag{4}
$$

と表される．各 L_i が最大となることはその他すべての $L_{j|j\ne i}$ が L_i 以下となることと等しく，その確率は累積密度関数 $P_j(L_i)$ の積で表されるため，式 (4) は

$$
\sum_{i=1}^{n} \int_{k_i}^{\infty} L_i p_i(L_i) \prod_{j\ne i} P_j(L_i) dL_i \tag{5}
$$

となる．これは 3.1 節で仮定した条件 $^\forall\alpha_i > 1$ を満たすとき解析的に式が得られる．ただし $l = \arg\max_{1\le i\le n} k_i$ としたとき k_l に対してのみ非対称で，$n=2$ のときは

$$k_l\left\{\alpha_1\left(\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}}{\alpha_1-1}-\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}\left(\frac{k_2}{k_l}\right)^{\alpha_2}}{\alpha_1+\alpha_2-1}\right)\right.$$
$$\left.+\alpha_2\left(\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}}{\alpha_2-1}-\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}\left(\frac{k_1}{k_l}\right)^{\alpha_1}}{\alpha_2+\alpha_1-1}\right)\right\}, \qquad (6)$$

$n=3$ のときは

$$k_l\left\{\alpha_1\left(\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}}{\alpha_1-1}-\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}\left(\frac{k_2}{k_l}\right)^{\alpha_2}}{\alpha_1+\alpha_2-1}\right.\right.$$
$$\left.-\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}\left(\frac{k_3}{k_l}\right)^{\alpha_3}}{\alpha_1+\alpha_3-1}+\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}\left(\frac{k_2}{k_l}\right)^{\alpha_2}\left(\frac{k_3}{k_l}\right)^{\alpha_3}}{\alpha_1+\alpha_2+\alpha_3-1}\right)$$
$$+\alpha_2\left(\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}}{\alpha_2-1}-\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}\left(\frac{k_3}{k_l}\right)^{\alpha_3}}{\alpha_2+\alpha_3-1}\right.$$
$$\left.-\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}\left(\frac{k_1}{k_l}\right)^{\alpha_1}}{\alpha_2+\alpha_1-1}+\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}\left(\frac{k_3}{k_l}\right)^{\alpha_3}\left(\frac{k_1}{k_l}\right)^{\alpha_1}}{\alpha_2+\alpha_3+\alpha_1-1}\right)$$
$$+\alpha_3\left(\frac{\left(\frac{k_3}{k_l}\right)^{\alpha_3}}{\alpha_3-1}-\frac{\left(\frac{k_3}{k_l}\right)^{\alpha_3}\left(\frac{k_1}{k_l}\right)^{\alpha_1}}{\alpha_3+\alpha_1-1}\right.$$
$$\left.\left.-\frac{\left(\frac{k_3}{k_l}\right)^{\alpha_3}\left(\frac{k_2}{k_l}\right)^{\alpha_2}}{\alpha_3+\alpha_2-1}+\frac{\left(\frac{k_3}{k_l}\right)^{\alpha_3}\left(\frac{k_1}{k_l}\right)^{\alpha_1}\left(\frac{k_2}{k_l}\right)^{\alpha_2}}{\alpha_3+\alpha_1+\alpha_2-1}\right)\right\} \qquad (7)$$

となり，$n\geq 4$ のときも同様の規則的な式が得られる．

次に，集団通信時間の期待値 $E[\max L_i]$ のプロセッサ数 n のみに対する依存性は，式 (5) より 2 項係数と級数およびガンマ関数を用いて

$$nk\alpha\sum_{i=1}^{n}(-1)^{i-1}\frac{\binom{n-1}{i-1}}{i\alpha-1}=-\frac{nk\Gamma(-\frac{1}{\alpha})\Gamma(n)}{\alpha\Gamma(1-\frac{1}{\alpha}+n)} \quad (8)$$

（ただし $^\forall\alpha_i=\alpha$，$^\forall k_i=k$）と表される．$n\to\infty$ のときの漸近近似を考えると，級数展開の高次項を無視して

$$E[\max L_i]\approx-\frac{n^{\frac{1}{\alpha}}k\Gamma(-\frac{1}{\alpha})}{\alpha}\propto n^{\frac{1}{\alpha}} \qquad (9)$$

と表されほぼ $n^{1/\alpha}$ に比例して増加することが分かる．一方，正規分布を用いた場合における集団通信時間の期待値 $E[\max L_i]$ のプロセッサ数 n のみに対する依存性は

$$\mu+\sigma\sqrt{2\ln n-\ln\ln n-\ln 4\pi}$$
$$<E[\max L_i]<\mu+\sigma\sqrt{2\ln n-\ln\ln n} \qquad (10)$$

（ただし $^\forall\mu_i=\mu$，$^\forall\sigma_i=\sigma$）と近似できることが示されており[15]，式 (8) および式 (9) と比較して n に対する増加量はしだいに小さくなることが分かる．

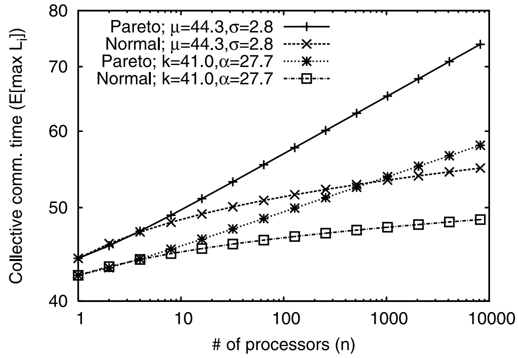

図 4 プロセッサ数に対する集団通信時間の変化
Fig. 4 Collective communication time vs. # of processors.

パレート分布を用いた式 (8) および正規分布を用いた式 (5) の数値積分値を**図 4** に示す．ただし図は両対数表示であり，また式 (5) で正規分布の場合の定積分区間の下限は $-\infty$ である．パラメータは 4 章で用いる評価データの代表値（**表 1** の中央値）2 組を使用し，式 (2) を用いてパレート分布および正規分布間で期待値および分散を同一とした．すなわち，$\mu=44.3$，$\sigma=2.8$ のとき $k=41.7$，$\alpha=16.9$ であり，$k=41.0$，$\alpha=27.7$ のとき $\mu=42.5$，$\sigma=1.59$ である．また近似式 (9) および (10) はこれらの曲線とほぼ一致するため図には示していない．両確率分布による推定値の交点はパラメータに依存するが，交点よりもプロセッサ数 n の大きい範囲でその増加につれて，正規分布を用いた推定値に比べパレート分布を用いた推定値の方が大きくなる傾向を示している．

3.3 パレート分布のパラメータ推定

実際の環境で確率分布を用いた性能推定を行うためには観測値から確率分布のパラメータを推定する必要があり，最尤法を用いてパレート分布のパラメータは以下により推定できる[12]．観測期間中の全観測値の大きい方から i 番目の観測値を X_i としたとき，尺度パラメータの推定値 $\hat{\alpha}$ は

$$\hat{\alpha}=m\left\{\sum_{i=1}^{m}\ln\left(\frac{X_i}{X_m}\right)\right\}^{-1} \qquad (11)$$

である．ここで，m は裾部分に属する観測値の数であり X_m は裾部分の観測値の最小値すなわち位置パラメータの推定値 $\hat{k}$ と等しい．一般的に，裾部分のみを対象とした解析においては全観測値から裾部分以外の観測値を無視してパラメータ推定を行うが，本稿で提案する手法では期待値を求めるために全観測値の確率分布の特性が必要であり m を観測期間中の全観

測数としてパラメータ推定を行った．ただし観測値の急激な変化がある場合で観測値の最小値に外れ値がある場合には，その値が位置パラメータ $\hat{k}$ となり敏感に反応する．

また，観測値からパラメータを推定する場合は 3.1 節で仮定した条件 $^\forall\alpha_i > 1$ を満たさない場合が起こりうる．このとき式 (5) は発散し期待値が得られないが，定積分区間の上限を x_0 で打ち切ることで積分可能となる．この場合，$n=2$ のときの期待値 $E[\max L_i]$ は式 (5) を用いて

$$k_l \left\{ \alpha_1 \left(\frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1} - \frac{k_1}{k_l}\left(\frac{k_1}{x_0}\right)^{\alpha_1-1}}{\alpha_1-1} - \frac{\left(\frac{k_1}{k_l}\right)^{\alpha_1}\left(\frac{k_2}{k_l}\right)^{\alpha_2} - \frac{k_1}{k_l}\left(\frac{k_1}{x_0}\right)^{\alpha_1-1}\left(\frac{k_2}{x_0}\right)^{\alpha_2}}{\alpha_1+\alpha_2-1} \right) + \alpha_2 \left(\frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2} - \frac{k_2}{k_l}\left(\frac{k_2}{x_0}\right)^{\alpha_2-1}}{\alpha_2-1} - \frac{\left(\frac{k_2}{k_l}\right)^{\alpha_2}\left(\frac{k_1}{k_l}\right)^{\alpha_1} - \frac{k_2}{k_l}\left(\frac{k_2}{x_0}\right)^{\alpha_2-1}\left(\frac{k_1}{x_0}\right)^{\alpha_1}}{\alpha_2+\alpha_1-1} \right) \right\} \quad (12)$$

と求められ，$n \geq 3$ のときも同様の規則的な式となる．ここで x_0 の選択は観測値の確率分布を考慮する必要があるが，観測数 m のとき観測値から求められる上側累積密度関数 $1-P_i(L_i)$ の最小値が $1/m$ であり，精度の制約から上側累積密度関数と $1/m$ が等しくなる

$$x_0 = k_i m^{\frac{1}{\alpha_i}} \quad (13)$$

を定積分区間の上限とした．

さらに，観測数および観測精度によっては全観測値が等しくなり，式 (11) より尺度パラメータ α_i が ∞ となる場合が起こりうる．このとき，対応する k_i を k_δ とし，確率密度関数 $p_i(L_i)$ をデルタ関数 $\delta(L_i - k_\delta)$ で近似することにより，$n=3$ のときの期待値 $E[\max L_i]$ は，$k_\delta = k_l$ のとき，式 (5) を用いて

$$k_\delta \left\{ \left(1-\left(\frac{k_2}{k_\delta}\right)^{\alpha_2}\right)\left(1-\left(\frac{k_3}{k_\delta}\right)^{\alpha_3}\right) + \alpha_2 \left(\frac{\left(\frac{k_2}{k_\delta}\right)^{\alpha_2}}{\alpha_2-1} - \frac{\left(\frac{k_2}{k_\delta}\right)^{\alpha_2}\left(\frac{k_3}{k_\delta}\right)^{\alpha_3}}{\alpha_2+\alpha_3-1} \right) + \alpha_3 \left(\frac{\left(\frac{k_3}{k_\delta}\right)^{\alpha_3}}{\alpha_3-1} - \frac{\left(\frac{k_3}{k_\delta}\right)^{\alpha_3}\left(\frac{k_2}{k_\delta}\right)^{\alpha_2}}{\alpha_3+\alpha_2-1} \right) \right\} \quad (14)$$

と求められ，$n=2$ や $n \geq 4$ のときも同様の規則的な式となる．ただし，$k_\delta \neq k_l$ のときはその確率分布を除きプロセッサ数を $n-1$ とした場合の式と等しい．複数の確率分布がデルタ関数となる場合は，$k_\delta \neq k_l$ であるデルタ関数の確率分布を除きその分プロセッサ数を減じた場合の式と等しい．

表 1　評価データのパラメータ分布
Table 1 Parameter distributions of the evaluation data.

	平均値	最小値	中央値	最大値
平均値 μ	50.1	1.0	44.3	249.0
標準偏差 σ	10.3	0.0	2.8	361.2
位置パラメータ k	46.5	1.0	41.0	238.0
尺度パラメータ α	–	0.9	27.7	∞

4. 評　　価

本章では，実際の広域ネットワーク上で観測された通信遅延時間データを用いて，まずパレート分布および正規分布への近似の程度を χ^2 適合度検定を用いて調べる．観測数 m および経過時間 τ を変化させて推定したとき，パレート分布を用いた提案手法と正規分布を用いた従来手法および推定直前の値を用いる手法に対してそれぞれ実際の値との誤差を求め，推定精度の指標として評価を行った．

4.1 評価に用いる通信遅延時間データ

評価に用いた観測データは，米国立応用ネットワーク研究所（National Laboratory for Applied Network Research: NLANR）が推進する能動的計測プロジェクト（Active Measurement Project: AMP）から提供されているものを用いた．全米および国際 115 サイトに観測点が存在し 13,110 サイト間におけるミリ秒精度の往復遅延時間データが 1 分間隔で計測されており☆，本稿では太平洋標準時 2005 年 3 月 20 日から 27 日までの 1 週間すなわち 10,080 分間のデータを用いた．ただしデータ欠測による評価誤差を抑えるため，全観測期間を通じてパケット損失が 10% 以上のデータを除外し 102 サイト 9,841 サイト間のみを評価データとして使用した．パケット損失時は直前の観測値が得られる時点の値を使用した．全期間を用いてパラメータ推定した平均値 μ，標準偏差 σ，位置パラメータ k，尺度パラメータ α の全観測データについての特性を表 1 に示す．平均値と位置パラメータの差，標準偏差および尺度パラメータから，評価データは分散が小さくネットワーク利用率が小さいと考えられる．一般に，パケットの輻輳回避の相互作用が原因で通信時間の確率分布の裾が長くなることが示されており[23)]，ネットワーク利用率の大きい方がパレート分

☆ 2005 年 3 月の値．http://amp.nlanr.net/から取得可能．

布の適合度が大きくなることが広域環境においても確認されている[11),17)]．一般的な広域環境における通信遅延時間分布の尺度パラメータ α は 1–2 程度であり，実際の広域環境では本稿で示したよりも推定精度を向上させられることが十分考えられる．

評価データをパレート分布および正規分布で近似するにあたり，各確率分布の近似の程度を調べるため χ^2 適合度検定を行った．適合指標 $\hat{\lambda}^2$ は観測数 m のとき階級数 N の度数分布を用いて，第 i 階級内に含まれる観測値数 Y_i，観測値から推定した各確率分布における第 i 階級内に含まれる期待数 mp_i より次式で与えられ，値が小さいほど適合度が高いことを示す[24)]．

$$\hat{\lambda}^2 = \frac{\sum_{i=1}^{N}\frac{(Y_i - mp_i)^2}{mp_i} - \sum_{i=1}^{N}\frac{Y_i - mp_i}{mp_i} - df}{m-1} \quad (15)$$

ここで $df = N-1-est$ は自由度を表し，est は推定パラメータ数である．裾部分の度数が小さいため観測値の対数変換を行った後，その標準偏差 $\hat{\sigma}$ を用いて次式より階級幅 w を決定した[26)]．

$$w = 3.49\hat{\sigma}m^{-\frac{1}{3}} \quad (16)$$

各サイト間のデータについてパレート分布および正規分布で近似したときの適合指標 $\hat{\lambda}^2$ を比較した結果，適合指標の差 $\hat{\lambda}^2_{pareto} - \hat{\lambda}^2_{normal}$ の全データにおける分布の {最小値，下側四分位数，中央値，上側四分位数，最大値} がそれぞれ $\{-1.47\times10^{292}$，-6.39×10^{6}，-22.4，10.3，$8.35\times10^{15}\}$ となった．これらの値が負でその絶対値が大きいほどパレート分布の適合度が高く，逆にこれらの値が正でその絶対値が大きいほど正規分布の適合度が高いことを示す．すなわち，正負の分布から正規分布よりもパレート分布による近似が適合する場合の方が多く，また絶対値の大きさから正規分布よりもパレート分布の適合度の方が高いことを示している．

4.2 評価条件と評価指標

本稿では，各 102 サイトからデータ品質が十分な全対向サイトまでの間の往復遅延時間データ L_i を用いて，集団通信時間 $E[\max L_i]$ の推定を行った．評価条件として，推定精度に与える影響が大きいと考えられるパラメータ推定に用いる観測数 m および推定時点からの経過時間 τ の 2 つを変化させた（図 3）．実際のスケジューリングに利用するには，計算量を小さくするため観測数が小さくかつプロセス実行の経過に対して推定誤差が大きくならない方が望ましく，観測数 m および経過時間 τ に対する推定精度の評価が必要なためである．ただし，評価データの観測間隔が 1 分であるため観測時間 m および経過時間 τ を 1 分単位とし，各推定時刻 t について評価を行った．

パレート分布を用いた提案手法および正規分布を用いた従来手法の推定精度を評価する際に，推定時点の直前の最大値 $\max L_i|_{t-1}$ を推定値とする手法との比較を行った．直前の最大値を用いる手法は通信遅延時間の動的変動性をまったく考慮しておらず，通信遅延時間の動的変動性を確率分布で扱うパレート分布および正規分布を用いた手法との差異が明らかになると考えたためである．評価指標には，推定時点 t から経過時間 τ を経た時点での観測値の最大値 $\max L_i|_{t+\tau}$ と推定値 $E[\max L_i]|_t$ の差から平方根平均自乗誤差（Root Mean Square Error: RMSE）を求め推定誤差の評価指標として用いた．

ただし正規分布を用いた手法では式 (5) が解析的に求められないため，数値計算ライブラリ GNU Scientific Library☆ 1.6 を用いて Gauss-Kronrod 則に基づく適応的数値積分により推定値を求めた．また，パラメータ推定において標準偏差 σ_i が 0 となる場合があり，このとき 3.3 節で述べた手法と同様に確率密度関数 $p(L_i)$ をデルタ関数 $\delta(L_i - \mu_i)$ として求めた．

4.3 推定誤差の評価

各手法の推定傾向を調べるため，あるサイトから対向 3 サイトへの往復遅延時間データ $L_1|_t$，$L_2|_t$ および $L_3|_t$ と，それらを用いて観測数 $m=16$ で推定した集団通信時間 $E[\max L_i]|_t$ の時間変化の一部を**図 5** に示す．14:00 前後で正規分布を用いた推定値が過大となっているのは，正規分布による近似が大きな値の外れ値に敏感に反応するためである．また 12:30 前後でパレート分布を用いた推定値が過大となっているのは，図 5 上部にその期間を示すように，尺度パラメータ $^{\exists}\alpha_i \le 1$ となっているためである．$^{\exists}\alpha_i \le 1$ となるのは観測期間中に階段状の通信遅延時間の偏位がある場合に多く，偏位前後で異なる 2 つの確率分布を 1 つの確率分布としてパラメータ推定することに起因し，3.1 節で述べた定常性の仮定が成立しない場合に対応する．特に，偏位期間と観測期間 m が近い場合で偏位の開始時および終了時には，全観測値中で偏位期間外の観測値が僅かに含まれる場合において，観測値の最低値 $\hat{k}$ に敏感なパレート分布を用いた推定の誤差を大きくしてしまう．それ以外の時間帯では両確率分布を用いた推定法でほぼ同じ時間変化であるが，最大値を示す通信遅延時間だけでなくそれ以下の通信遅延時間の裾部分の影響をより考慮するパレート分布を用いた推定法の方がわずかに大きな値を示している．これ

☆ http://www.gnu.org/software/gsl/

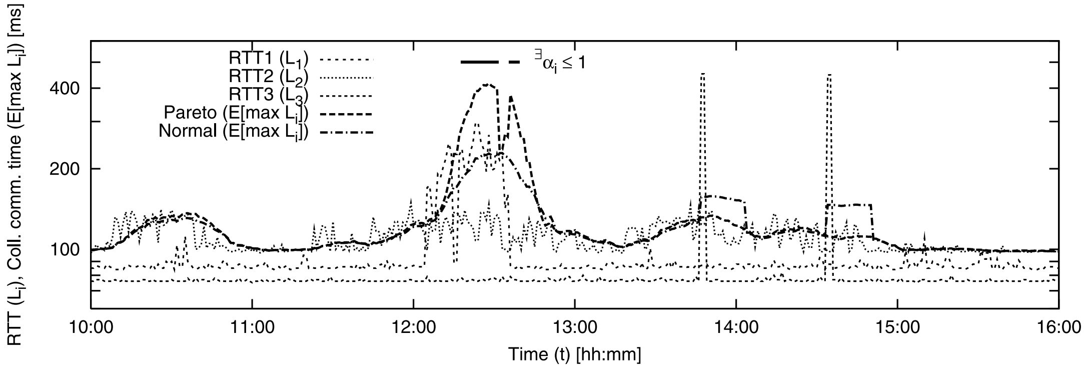

図 5　往復遅延時間データと集団通信時間の推定値の時間変化の一例
Fig. 5　An example trace of observed RTTs and estimated collective comm. time.

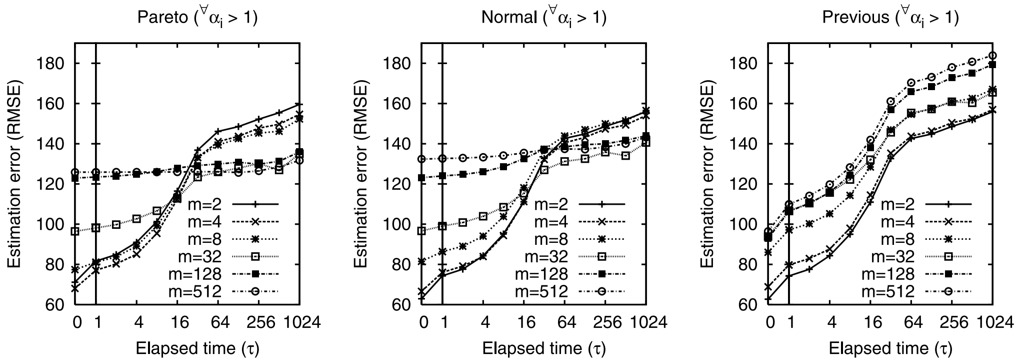

図 6　観測数および経過時間を変化させたときのパレート分布，正規分布，直前値の推定誤差（$^\forall\alpha_i > 1$）
Fig. 6　Estimation error of each method vs. # of observations and elapsed time ($^\forall\alpha_i > 1$).

表 2　全データに占める $^\exists\alpha_i \le 1$ となる時刻の割合
Table 2　Percentage of the data points as $^\exists\alpha_i \le 1$.

観測数 m	2	4	8–128	256	512	1024
割合 [%]	15.6	6.0	4.1	2.8	2.3	1.1

らの傾向は評価条件を変化させた場合についても同様であった．このように尺度パラメータ α_i の値によって各手法の傾向が異なることから，以下では $^\forall\alpha_i > 1$ の部分と $^\exists\alpha_i \le 1$ の部分に分けて検討する．ただし**表 2** に示すように全データ中で $^\exists\alpha_i \le 1$ となる割合は小さい．

まず $^\forall\alpha_i > 1$ の部分について，評価データの全期間全サイトで平均した推定誤差の観測数 m および経過時間 τ に対する推定誤差の変化を**図 6** に示す．図 6 は経過時間の増加に対して推定誤差が単調増加することを示しており，観測数の増加に対して確率分布を用いる手法は推定誤差の変化が小さくなることを示している．正規分布を用いる手法（図 6 中央）とパレート分布を用いる手法（図 6 左）を比較すると，観測数 m=2，4 以外のときパレート分布を用いた手法の

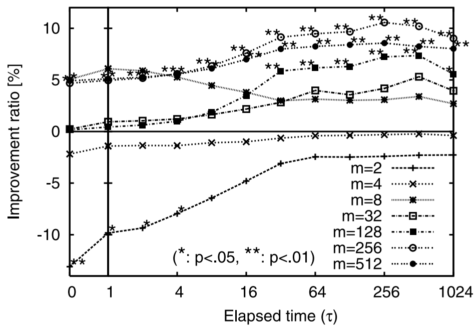

図 7　正規分布の手法に対するパレート分布の手法の誤差改善度
Fig. 7　Error improvement rate of Pareto to Normal.

推定精度が小さくなっていることが分かる．この場合は，観測数が小さすぎるために裾の長いパレート分布に近似させたパラメータ推定が不適切であることが原因であり[11]，経過時間 $\tau \le 16$ で正規分布および直前値を用いた手法（図 6 右）の方が推定誤差が小さい．正規分布を用いた手法の推定誤差に対するパレート分布を用いた手法の推定誤差の改善度の平均値を**図 7** に示す．さらに，平均値の差が有意に認められるかどう

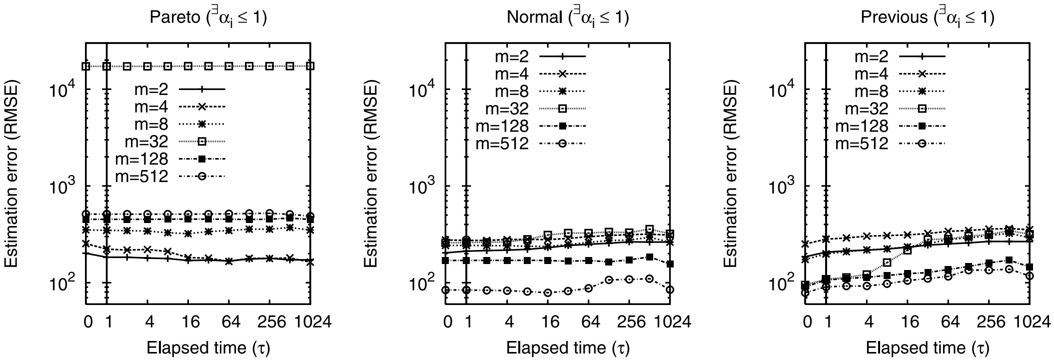

図 8 観測数および経過時間を変化させたときのパレート分布，正規分布，直前値の推定誤差（$^\exists\alpha_i \le 1$）
Fig. 8 Estimation error of each method vs. # of observations and elapsed time ($^\exists\alpha_i \le 1$).

かを検定するため Mann-Whitney の U 検定を行い，その有意水準も図に示した．観測数 $m \ge 8$ で改善度が正となっているが，有意性が認められるのは観測数 $m = 128$ かつ経過時間 $\tau \ge 32$ のときと観測数 $m \ge 256$ のときであり，特に観測数 $m = 256$ 経過時間 $\tau = 256$ において改善度の最大値が 11%であることを示している．

次に $^\exists\alpha_i \le 1$ の部分について，評価データの全期間全サイトで平均した推定誤差の観測数 m および経過時間 τ に対する推定誤差の変化を**図 8** に示す．図 6 では経過時間の増加に対して，推定誤差の増加が $^\forall\alpha_i > 1$ の部分と比較して比較的小さいことを示しており，偏位前後の階段状変化の後では推定誤差が大きい状態で安定していると考えられる．また観測数の増加に対して，正規分布および直前値を用いた手法は推定誤差が減少するのに対して，パレート分布を用いた手法は推定誤差が増加することを示している．これは，正規分布および直前値を用いた手法は観測数の増大に対し平均の効果が大きくなること，およびパレート分布を用いた手法は観測数の増大に対し観測値の最小値の影響が大きくなることが原因であると考えられる．ただしパレート分布を用いた手法において観測数 $m = 32$ の場合の推定誤差が大きくなっているのは，評価データは観測数 m に近い期間の偏位が多く存在したため推定誤差が大きくなったことが原因である．

以上の結果より，$^\forall\alpha_i > 1$ で経過時間 $\tau \le 16$ のときは観測数 $m = 2$ で正規分布および直前値を用いる手法，$^\forall\alpha_i > 1$ で経過時間 $\tau \ge 32$ のときは観測数 $m \ge 32$ でパレート分布を用いる手法，$^\exists\alpha_i \le 1$ のときは観測数 $m = 512$ で正規分布を用いる手法が推定誤差を最小にできることが分かる．

4.4 実際利用に対しての考察

4.3 節の結果より，各条件によってパレート分布，正規分布および直前値を用いた手法の推定傾向が異なるため，つねに同一の手法を用いるよりも各条件によってこれらの手法を切り替える方法が有効である．経過時間 τ および観測値の階段状の偏位をとらえるために尺度パラメータ α_i を条件として，推定誤差が最小となるような手法および観測数 m を決定することが考えられる．ただし，長時間の運用においては切替え点が変化することが予想されるため，アプリケーション実行中にも推定と誤差評価を行いながら変動に適応させていくことが必要である．あるいは適合度指標 $\hat{\lambda}^2$ が最小となる分布を選択する方法も考えられ，この場合はアプリケーション実行中に適合度指標 $\hat{\lambda}^2$ を求める必要がある．

また集団通信時間推定のために必要な計算量を考慮すると，正規分布を用いた手法は解析的な式が得られないため，数値積分を行う必要がある．その他の裾の長い確率分布である対数正規分布，ガンマ分布，ワイブル分布等を用いた場合も式 (5) の定積分が解析的に求まらず数値積分が必要である．本稿で用いた数値積分は，積分区間を適応的に分割し近似値がある誤差範囲に入るまで収束させていく手法であり，分割点ごとに被積分関数の値を求める必要がある．計算量は収束速度にも依存するが，誤差範囲および分割数の限度が存在するため，プロセッサ数 n に対し式 (5) の被積分関数を求めるための $\mathcal{O}(n)$ と見積もられる．一方，パレート分布を用いる手法は式 (6)，(7) より四則演算およびべき算のみであるが，項数は

$$n \sum_{i=1}^{n} \binom{n-1}{i-1} = 2^{n-1} n \tag{17}$$

であるため，計算量は $\mathcal{O}(2^{n-1} n)$ と見積もられる．両手法における集団通信時間推定の時間を比較するため，**表 3** の環境における観測時間 $m = 16$ のときの計測結

表 3　集団通信時間推定に用いた計算機の仕様
Table 3　Specification of the estimation machine.

CPU	Intel Pentium4 2.53 GHz
Motherboad	ASUS P4S533-MX
Memory	512 MB DDR333
OS	Linux 2.4.20 (RedHat 9)
Compiler	gcc 3.2.2 (option: -O3)

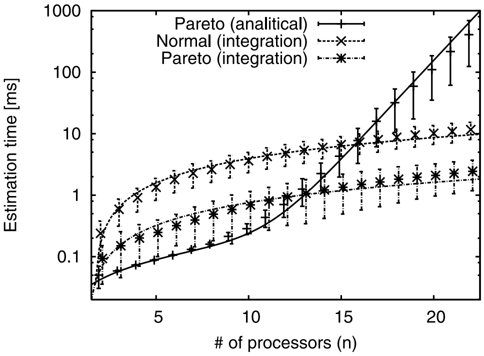

図 9　集団通信時間推定に要する時間
Fig. 9　Estimation time vs. # of processors.

果を**図 9** に示す．プロセッサ数 $n \leq 15$ では正規分布を用いた方法よりパレート分布を用いた方法の推定時間が小さいのに対し，$n \geq 16$ では逆となることを示している．ただし，パレート分布においても数値積分を行う方法が適用でき，図 9 より $n \geq 13$ で数値積分を行えばつねに正規分布を用いる手法よりも推定時間を小さくできることが分かる．また，複数のクラスタによる階層構造を考慮した場合，クラスタ内通信遅延時間 c_i で通信遅延時間の確率密度関数 $p_i(L_i - c_i)$ を用いて，式 (5) の定積分が解析的に求まらないため数値積分を行う必要があり，推定時間はパレート分布を用いて数値積分を行う場合とほぼ等しい．

5. ま と め

本稿では，広域ネットワークにおける通信遅延時間に着目し，通信遅延時間の確率分布に裾の長い確率分布であるパレート分布を適用することによって，グリッド環境をはじめとする広域分散計算環境における集団通信時間の推定方法を提案した．提案手法より，プロセッサ数の大きい広域分散計算環境において，パレート分布を用いる提案手法と比較し，正規分布を用いる従来手法では集団通信時間を過小に推定していること，および推定に必要な計算量が大きいという問題を解析的に示した．さらに，運用中の広域ネットワークの往復遅延時間データを用いて評価を行い，観測値の階段状の偏位が少なく推定時点からの経過時間が大きい範囲で正規分布を用いる従来手法よりも推定精度を向上させることができることを示した．

さらに，実際の多種にわたる並列アプリケーションおよび広域分散計算環境に適用するためには，同期型以外の並列アプリケーションの考慮，集団通信だけでなく計算時間の考慮等が必要である．実証実験による評価と並行してこれらの課題を解決していきたいと考えている．

謝辞　本研究は，科学研究費補助金特定領域研究（C）課題番号 13224059，および若手研究（B）課題番号 15700055 の助成を受けて行われた．

参 考 文 献

1) Adamic, L.A. and Huberman, B.A.: Zipf's law and the Internet, *Glottometrics*, Vol.3, pp.143–150 (2002).
2) Bernaschi, M. and Iannello, G.: Collective communication operations: Experimental results vs. theory, *Concurrency Pract. Ex.*, Vol.10, No.5, pp.359–386 (1998).
3) Buyya, R.: In brief: The Virtual Laboratory project, *IEEE Distrib. Syst. Online*, Vol.2, No.5 (2001).
4) Casanova, H., et al.: The virtual instrument: Support for grid-enabled scientific simulations, *Int. J. High Perform. C.*, Vol.18, No.1, pp.3–17 (2004).
5) Charney, M.: The role of network bandwidth in barrier synchronization, *J. Parallel Distr. Com.*, Vol.28, No.2, pp.202–212 (1995).
6) Culler, D.E., et al.: LogP: A practical model of parallel computation, *Comm. ACM*, Vol.39, No.11, pp.78–85 (1996).
7) Dongarra, J.J., et al.: *LINPACK users' guide*, Society for Industrial & Applied Mathematics (1979).
8) Fitch, B.G., et al.: Blue matter, an application framework for molecular simulation on Blue Gene, *J. Parallel Distr. Com.*, Vol.63, No.7–8, pp.759–773 (2003).
9) Foster, I., et al.: The anatomy of the grid: Enabling scalable virtual organizations, *Int. J. Supercomput. Ap.*, Vol.15, No.3, pp.200–222 (2001).
10) Fox, G.C., et al.: *Parallel computing works*, Morgan Kaufmann Publishers (1994).
11) Fujimoto, K., et al.: Statistical analysis of packet delays in the Internet and its application to playout control for streaming applications, *IEICE T. Commun.*, Vol.E84-B, No.6, pp.1504–1512 (2001).
12) Hill, B.M.: A Simple general approach to in-

ference about the tail of a distribution, *Ann. Stat.*, Vol.3, No.5, pp.1163–1174 (1975).
13) Kalé, L., et al.: NAMD2: Greater scalability for parallel molecular dynamics, *J. Comput. Phys.*, Vol.151, No.1, pp.283–312 (1999).
14) Karonis, N.T., et al.: MPICH-G2: A grid-enabled implementation of the message passing interface, *J. Parallel Distr. Com.*, Vol.63, No.5, pp.551–563 (2003).
15) Lai, T.L. and Robbins, H.: Maximally dependent random variables, *P. Nat'l Acad. Sci. USA*, Vol.73, No.2, pp.286–288 (1976).
16) Lee, C. and Stepanek, J.: On future global grid communication performance, *Proc. HCW 2001* (2001).
17) Loguinov, D. and Radha, H.: End-to-end Internet video traffic dynamics: Statistical study and analysis, *Proc. IEEE INFOCOM 2002*, pp.723–732 (2002).
18) Mizuno-Matsumoto, Y., et al.: A grid Application for an evaluation of brain function using independent component analysis (ICA), *Proc. CCGrid2002*, pp.111–118 (2002).
19) Mukherjee, A.: On the dynamics and significance of low frequency components of Internet load, *Internetworking*, Vol.5, No.4, pp.163–205 (1994).
20) 新家正総：Latency や gap のゆらぎを考慮した LogP モデルの検討，情処研報，99-CPSY-62, pp.25–32 (1999).
21) Nozaki, K., et al.: The first grid for oral and maxillofacial region and its application for speech analysis, *Method. Inform. Med.*, Vol.44, No.2, pp.253–256 (2005).
22) Pacheco, P.: *Parallel programming with MPI*, Morgan Kaufmann Publishers (1997).
23) Paxson, V. and Floyd, S.: Wide-area traffic: The failure of Poisson modeling, *IEEE ACM T. Network.*, Vol.3, No.3, pp.226–244 (1995).
24) Pederson, S. and Johnson, M.: Estimating model discrepancy, *Technometrics*, Vol.32, No.3, pp.305–314 (1990).
25) Schopf, J.M. and Berman, F.: Performance prediction in production environments, *Proc. IPPS/SPDP 98*, pp.647–653 (1998).
26) Scott, D.W.: On optimal and data-based histograms, *Biometrika*, Vol.66, No.3, pp.605–610 (1979).
27) Tanaka, Y., et al.: Climate simulation using Ninf-G on the ApGrid testbed, *Proc. CCGrid2003* (2003).
28) Vetter, J.S. and Mueller, F.: Communication characteristics of large-scale scientific applications for contemporary cluster architectures, *J. Parallel Distr. Com.*, Vol.63, No.9, pp.853–865 (2003).
29) Wilson, G.V.: *Practical parallel programming*, MIT Press (1995).
30) Wolski, R.: Dynamically forecasting network performance using the network weather service, *Cluster Computing*, Vol.1, No.1, pp.119–132 (1998).
31) Yokokawa, M., et al.: 16.4-Tflops direct numerical simulation of turbulence by a Fourier spectral method on the Earth Simulator, *Proc. SC2002*, p.50 (2002).
32) Zaki, O., et al.: Toward scalable performance visualization with Jumpshot, *Int. J. High Perform. C.*, Vol.13, No.3, pp.277–288 (1999).

(平成 17 年 4 月 28 日受付)
(平成 17 年 8 月 11 日採録)

甲斐島　武

1978 年生．2003 年大阪大学大学院工学研究科情報システム工学専攻修士課程修了．同年 4 月より同大学院情報科学研究科マルチメディア工学専攻博士課程に在籍．

加藤　精一

2002 年東京大学大学院理学系研究科天文学専攻博士課程修了．同年大阪大学サイバーメディアセンター教務職員を経て，2004 年より同センター助手．宇宙ジェットの磁気流体シミュレーション，グリッドや P2P 技術による仮想研究環境に関する研究に従事．理学博士．天文教育普及研究会，日本天文学会各会員．

秋山　豊和（正会員）

1999 年大阪大学院工学研究科修士課程修了．2000 年同大学院博士課程中退後，同大学サイバーメディアセンター助手を経て，2005 年より同センター講師．広帯域ネットワークにおける分散データベースシステムに関する研究に従事．工学博士．電子情報通信学会，IEEE 各会員．

野﨑　一徳（正会員）

2000 年北海道大学歯学部卒業．2004 年大阪大学大学院歯学研究科博士課程修了．同年 4 月より同大学サイバーメディアセンター教務職員．口腔領域の医学とその臨床を促進するグリッド技術を用いた情報システム開発に関する研究に従事．歯学博士．

水野（松本）由子

1996 年大阪大学大学院医学研究科博士課程精神神経科学修了．2003 年同大学院工学研究科博士後期課程情報システム工学専攻修了．1996 年同大学医学部機能画像診断学医員．1998 年同大学大学院基礎工学研究科ポスドクリサーチアソシエイト．1999 年米国ジョンズ・ホプキンス大学ポスドクリサーチフェロー．2000 年大阪城南女子短期大学助教授．2004 年より兵庫県立大学大学院助教授．医学博士，工学博士．臨床神経生理学，信号処理工学，情報科学を用いた脳機能解析や精神機能解析に関する研究に従事．日本臨床神経生理学会，日本精神神経学会，日本生体磁気学会，日本医療情報学会各会員．

下條　真司（正会員）

1986 年大阪大学大学院基礎工学研究科博士課程修了．同年同大学基礎工学部助手．1989 年同大学大型計算機センター講師．1991 年同センター助教授．この間米国カリフォルニア大学アーバイン校客員研究員．1998 年大阪大学大型計算機センター教授．2000 年より同大学サイバーメディアセンター教授．マルチメディア応用システム，peer-to-peer コミュニケーションネットワーク，ユビキタスネットワークシステム，グリッド技術等の研究に従事．工学博士．志田林三郎賞，日本医用画像工学会論文賞，大阪科学賞受賞．日本学術振興会インターネット技術第 163 委員会副委員長．電子情報通信学会，IEEE CS，ACM 各会員．