エントロピー・スロットリング：相互結合網のパケット移動度に着目した輻輳制御手法

横田隆史[†] 大津金光[†]
古川文人[††] 馬場敬信[†]

独立動作する多数のルータにより構成される相互結合網では，過大な転送負荷により系全体にわたる輻輳を生じる結果，スループット/レイテンシ両面での性能低下を避けることができない．相互結合網へのパケットの投入を適切に抑制することにより，最大の性能が得られる領域で相互結合網を運用できれば，性能低下を抑えられ効率の良い転送が可能になるはずである．本論文ではこうした観点から相互結合網へのパケットの投入を抑制（スロットリング）する方式を提案する．エントロピーにより相互結合網内の輻輳の度合いを表現できることを利用し，効果的に系内のエントロピーを求めスロットリングに反映させる．シミュレーション評価の結果，スループットを維持しながらレイテンシの増大を抑える著しい効果が得られることを示す．

Entropy Throttling for Maximizing Packet Mobility in Interconnection Networks

Takashi Yokota,[†] Kanemitsu Ootsu,[†] Fumihito Furukawa[††] and Takanobu Baba[†]

Large-scale interconnection networks, which are composed of many independent routers, have common but essential problem: excessive communication load causes heavy congestion which results in serious performance degradation. This problem is essential under excessive load conditions. In this paper, we introduce entropy measure which is able to represent congestion situation quantitatively. We propose practical entropy measuring method and injection limitation (throttling) method. Entropy value is measured by effective cooperation of routers, and each router throttles packet injection according to the entropy value. Our simulation results reveal that our method can preserve high throughput and reduce latency in heavy traffic load situations.

1. はじめに

大規模な並列計算システムを構築するには，多数の演算ノードを性能的・経済的に効率良く結合するための相互結合網の検討が必須である．集中制御の機構を持たない大規模相互結合網では，転送負荷の増大にともなって輻輳が生じる結果，転送性能が著しく悪化する現象が起きることが知られている[1),2)]．このために，過大な転送負荷の状況下であっても相互結合網の転送性能を保つための技術が求められる．

本論文では，独立に動作する要素（ルータ）を規則的に配置し相互に接続することにより構成する相互結合網（直接網）を前提とし，相互結合網の輻輳状態の検知と，演算ノードから結合網へのパケット投入の抑制（スロットリング）を行うことにより，上記の問題の解決を図る．

各ルータが独立してパケットの転送制御を行う状況では，配送途中のパケットの間で干渉（転送のブロック）が生じることは避けられない．パケット間の干渉は，システム全体の通信負荷の多寡に応じて発生の度合いが決まる．通信負荷が小さい場合，干渉は局所的に限られすぐに解消するが，通信負荷が大きい状況では，パケット間干渉が解消されず，相互結合網系全体を覆う輻輳の状態となる．

相互結合網が著しい輻輳の状態にあるとき，系全体の転送能力（スループット）は著しく低下し，パケット配送のレイテンシは急増する．こうした状況は，我々の日常生活において発生する交通渋滞と類似している．

† 宇都宮大学工学部情報工学科
Faculty of Engineering, Utsunomiya University
†† 帝京大学ラーニングテクノロジー開発室
Learning Technology Laboratory, Teikyo University

すなわち，全体の交通量が一定の水準を超えると急激に渋滞が発生し，流量が激減するとともに目的地までの所要時間が急増する．

こうした現象は，上述のように相互結合網が独立動作を行う多数のルータより構成され，輻輳に対する広域的な制御がなされない系においては本質的に発生しうるものである．物理的なバンド幅の向上やルーティングアルゴリズムの工夫によれば，輻輳への耐性を相対的に向上させることができるが，一定水準以上の転送負荷を加えた場合に輻輳状態になる問題自体が解決されるわけではない．

我々はこうした輻輳状態の問題に着目し，先行研究において輻輳の発生・成長のメカニズムを一部明らかにするとともに，輻輳の成長の結果，相互結合網系の性質が相転移ともいえる変化をなすことを明らかにした[3)~5)]．さらに，パケットの移動速度をもとに熱力学からのアナロジによりエントロピーを定義することで，相互結合網系の輻輳の度合いを表現できることを明らかにした．本論文では，こうした先行研究による知見を基盤とし，エントロピーとして表現される相互結合網系の輻輳度合いを測定し，その結果によりノードがパケットを生成し結合網に投入するのを抑制するスロットリング（throttling）の手法を考える．

以下，本論文は次のような構成をとる．まず 2 章で，先行研究で得られた輻輳の発生・成長に関する知見と，エントロピーについてまとめ，本論文で適用するためのエントロピーの近似手法を論じる．次に 3 章において，2 次元トーラス網を対象として，効率的にエントロピー値を求め，スロットリングする方式について検討する．4 章で非適応/適応ルーティングに本論文の提案方式を適用した場合の効果を評価する．その後，5 章で関連する研究との比較を行い，最後に 6 章でまとめる．

2. パケットの移動度とエントロピー

2.1 輻輳による相転移現象の発生

相互結合網では，大きな転送負荷により輻輳を生じ，その結果，系の性質が激しく変化することが経験的に知られていた．しかし輻輳のメカニズムは明確な形では提示されていなかった．我々はセルオートマトンの手法により単純化したモデルを用いることで輻輳の発生・成長・消滅のメカニズムを明らかにし，さらにエントロピーを導入することで輻輳の度合いを表現できることを示した[3)~5)]．

相互結合網内でパケット間の相互干渉による転送のブロックが複数のルータ間にまたがって発生する状況を**輻輳**と定義する．また，輻輳が生じない上限の転送負荷を**臨界転送負荷**と呼ぶ．臨界転送負荷の近辺では，系内のパケット存在密度のゆらぎにより，局所的に輻輳が発生するが，多くの場合持続せずに解消する．しかし，臨界転送負荷を超えた状況では，局所的な輻輳が解消せず周囲のパケットを巻き込んで成長し，ついには大きな塊状のクラスタ（輻輳領域）を形成する．こうした局所的な輻輳の頻繁な発生・消滅と，塊状をなす輻輳クラスタへの成長の挙動が見られることは，セルオートマトンにより単純化されたモデルだけではなく，2 次元トーラス等通常の直接網でも観測されることが明らかになっている[3),4)]．

2.2 エントロピーの定義

相互結合網の性質は上述の輻輳クラスタの有無により大きく異なる．輻輳クラスタが生じていない状態では，系全体にわたりおおむねパケットが流れている状況である．一方，輻輳クラスタの中ではパケットが相互にブロックし合うためにほとんど流れなくなる．このために，輻輳クラスタから脱出するパケットの流量と，周囲の領域にあったパケットが新たに組み入れられる流量とのバランスがとれた点で定常状態となる．相互結合網は，臨界転送負荷を境界として，こうした 2 つの状態をとる．

このような相互結合網の状態を定量的に表現するために提案されたのが，パケットの移動度をもとにしたエントロピーである．相互結合網の系内にある全パケットについて，時間 Δt 内にホップした回数を求める．ここで簡単化のため移動の方向は考慮せずにホップ数のみを考える．Δt をクロック単位で設定すれば，ホップ数は整数値になる．また，パケットの進行速度を 1 クロックあたり最大 1 ホップとすれば，個々のパケットの移動距離は範囲 $[0{:}\Delta t]$ の整数となる．

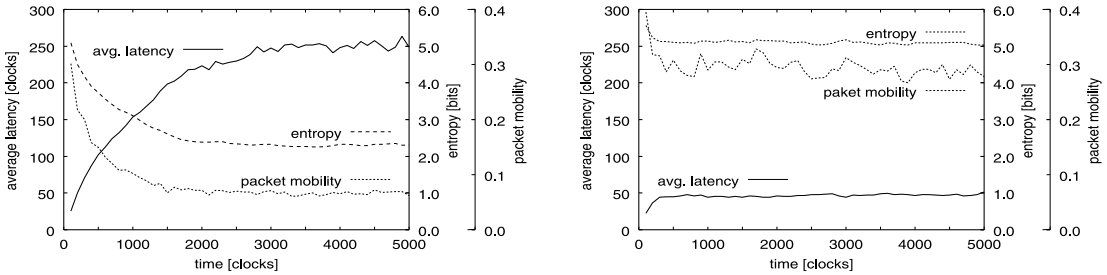
そして，移動距離の値ごとのパケットの個数の分布を求める．移動距離 h を持つパケットの数を n_h としたとき，$h \neq 0$ なる h に対して以下のようにエントロピーを定義する．

$$H_{\Delta t}(S) = \sum_{h=1}^{\Delta t} \frac{n_h}{N} \log_2 h \qquad (1)$$

ここで N は系内に存在するパケットの総数である．

2.3 エントロピーの近似表現

式 (1) によるエントロピー値を求めるには，時間 Δt 間での全パケットの移動ホップ数を計測しなければならず，実際の相互結合網で測定することを考えると現実的とはいえない．また，Δt を小さくすればエントロピーの測定は容易になるが，精度が低下する．この



(a) 輻輳状態を生じる場合 (b) 輻輳状態を生じない場合

図 1 平均レイテンシ，エントロピー，平均移動度の時系列変化の様子
Fig. 1 Time sequence of average latency, entropy and average packet mobility.

まま計測時間 $\Delta t \gg 0$ を前提として議論を進めるのは現実的ではない．したがって，できるだけ短時間で容易に測定できる方法が必要となる．

ここで式 (1) で同じ h 値を持つ n_h 個のパケットについて考える．個々のパケットはクロックごとに 1 ホップ進むかそこにとどまるかのいずれかを繰り返す．これらは Δt 間で h ホップ進んでいることから，各クロックでは $p_h = h/\Delta t$ の確率で進むと見なせる．こうして個々のパケットの移動確率を考える．

そこで，個々のパケット移動確率 p_i をもとにして，式 (1) を以下のように粗く近似する．

$$H(S) \approx \frac{\kappa}{N} \sum_{i=1}^{N} p_i \qquad (2)$$

κ は定数である．この式は，エントロピー値をパケットの移動確率の平均値で近似しようとするものである．もとのエントロピーの定義式からは大きく異なっているが，現実的には式 (1)，(2) との間には，リニアに近い良好な相関関係があることが示されている[4]．

2.4 相互結合網でのエントロピー

実際の相互結合網において，平均レイテンシ，エントロピー（式 (1) による），平均パケット移動度の各指標が時系列変化する様子をシミュレーション結果をもとに**図 1** に示す．32×32 の二次元トーラス網で minimum adaptive ルーティングを行った場合について，シミュレーション開始時から 5,000 クロックの間の様子である．パケット長 8 フリット，仮想チャネルのバッファ長 16 フリットで virtual cut-through フロー制御を行っている．いずれの指標も $\Delta t = 100$ クロックの時間窓で測定した平均値をプロットしている．

図 1 (a)，(b) は，各ノードからのパケット投入間隔が異なる．同図 (a) は投入間隔 32 クロックで，シミュレーション開始後から徐々に輻輳が進行している．同図 (b) はパケット投入間隔を 48 クロックにしたときのものであり，輻輳は生じていない．この図から，輻輳の度合いを表すパラメータとして，式 (1) によるエントロピー値ではなく，その近似である平均パケット移動度（式 (2)）を用いても問題ないことが確認される．

3. エントロピー・スロットリング

前章での結果をもとに，本章で現実の相互結合網に適用可能な方式を検討していく．議論の簡単化のため，ここでは 2 次元メッシュ/トーラス網を前提とする．まず，エントロピーを効率的に求める方式を検討し，そのうえで適切なスロットリングを行う方法を検討していく．

3.1 現実的なエントロピーの近似表現

前章において式 (1) によるエントロピーの近似として式 (2) が使えることを示した．これにより Δt 時間のパケットの移動ホップ数を測定する必要がなくなり，現実の相互結合網に適用しやすくなった．しかし，式 (2) では測定時刻における移動の有無を全パケットについて調べる必要があり，以下の 2 点において実装上の問題になる．

第 1 に，本論文では独立動作するルータによって構成される相互結合網を前提に検討しており，そこでは，パケットの位置や移動の有無を結合網全体にわたって管理するメカニズムが存在しない．このために，各ルータの協調動作によりパケット移動度を求めなければならない．

第 2 に，ルータは一般的に配送途中のパケットを保持するためのバッファを備えていることに留意しなければならない．パケットバッファの容量がパケット長より大きい場合には，1 つのバッファに複数のパケットが格納される可能性がある．このため，全パケットの移動の有無を正確に調べるには，パケットバッファの内容をそのつど精査しなければならない．

前者は独立動作するルータを前提にする限り避けられない問題である．このため，後者の問題について，ここでさらに簡略化の方法を検討する．

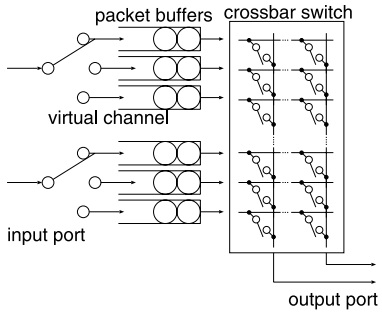


図 2 ルータのモデル
Fig. 2 Router model.

簡略化したルータの構成を**図 2** に示す．入力ポートから入力されたパケットは仮想チャネルごとに設けられたパケットバッファに格納される．パケットバッファは一般的に FIFO（First-In First-Out）により構成され，パケットは到着順に格納され到着順に取り出される．ルータからの出力の対象となるのはパケットバッファの先頭にあるパケットのみである．先頭パケットはクロスバスイッチによって適切に出力方向を選択され出力される．

こうしたルータの一般的な構成を考慮し，中にパケットの内容を 1 片でも保持しているバッファの数（N_{ob}: number of occupied buffers）と，それらのバッファのうち，先頭パケットがブロックされずに出力されている状態にあるものの数（N_{ab}: number of active buffers）を求め，それらの比 N_{ab}/N_{ob} を求める．

配送途中のパケットは必ずパケットバッファ内に保持されるから，結合網中のパケットの数 N が多いほど N_{ob} は大きくなる．また，バッファ内の先頭パケットのみが出力の対象となり，しかもそのパケットがブロックされず転送されている状況は，そのパケットを保持しているバッファが出力状態にあることを意味している．以上から，式 (2) をさらに近似したものとして N_{ab}/N_{ob} を用いることができることが分かる．この近似の妥当性は，4 章で評価する．

以上の議論をもとに，本論文では，相互結合網系の輻輳の度合いを表すエントロピーの近似として，1 フリット以上のパケット要素を保持しているバッファの個数 N_{ob} と，先頭パケットを出力しているバッファの個数 N_{ab} をもとに N_{ab}/N_{ob} の値を用いる．

3.2 エントロピーの測定方法

ここで実際に N_{ob} と N_{ab} の値を求める方法について検討する．各パケットバッファにパケットの構成要素が含まれているか否か，そのパケットバッファからパケットが出力中であるかブロックされているのか，の各情報は容易に得ることができる．そこで，各ルータごとに，自ルータ内の occupied buffer，active buffer の数（n_{ob}，n_{ab}）を保持するためのレジスタを用意する（mobility counter register: MCR と称する）．各ルータは 1 つの MCR を持ち，各 MCR は (n_{ob}, n_{ab}) の 2 つの数の組によって構成される．$N_{ob} = \sum n_{ob}$，$N_{ab} = \sum n_{ab}$ である．

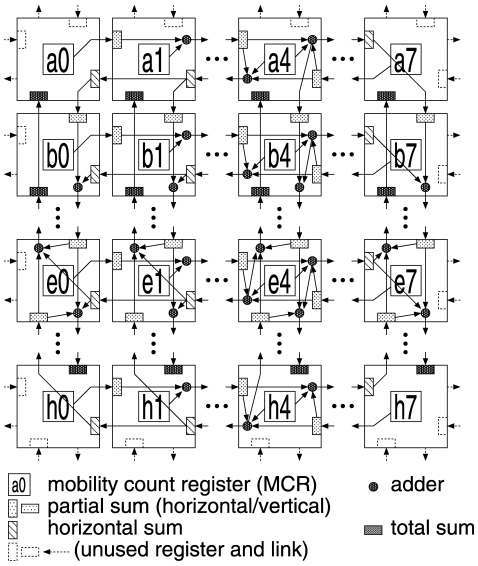


図 3 エントロピー計測回路
Fig. 3 Entropy measurement circuit.

相互結合網全体での N_{ob}，N_{ab} の値を求めるには，全ルータについて MCR の値の総和を求めればよい．このために，隣接ルータの間で適切な情報を送り，結果的に $\sum MCR$ を求める方法を考えた．本論文で前提としている 2 次元メッシュ/トーラス網において総和を求めるための回路の例を**図 3** に示す．この図では，8×8 のトーラス網の例を示している．

まず各ルータは自己の MCR 値を求める（図中，各ルータの中心部に配した a0〜h7 で表している）．この値は，各パケットバッファの状態をもとに簡単な組合せ回路により求めることができる．次に各ルータは，各々の MCR 値を水平方向中心部のルータに向け転送する．たとえば図 3 で 1 行目にあるルータは MCR 値を水平方向に a4 のルータに向け転送する．このとき各ルータは，隣接ルータから受け取った MCR 値に自己の MCR 値を加え，反対側の隣接ルータに転送する．これにより，中心部にあるルータ（図の場合 a4 のルータ）で，同一行にある全ルータの MCR 値の総和（水平方向総和値）が求められる．この水平方向総和値を，外側に向けて順次転送していくことによって，同一行にあるルータにブロードキャストすることができる．水平方向の総和値が求められれば，同様の転送

を行うことで垂直方向の総和を求めることができる．こうして，全体のルータの MCR 値の総和を求め，全体にブロードキャストすることができる．図 3 は，こうした一連のプロセスを，2 次元メッシュのトポロジの中に埋め込んだ形で効率良く実行できることを示している．

さらに，図 3 に示す回路は環状構造をなしていないため，上述の動作を行うために全ルータで同期をとる必要がない．各ルータは，自己の MCR レジスタ値の更新や隣接ルータ間での通信を随時行うことが可能である．このため，図 3 は各ノードの MCR 値を入力とし N_{ob}，N_{ab} を求めるための系全体にまたがった大きな順序回路となっている．

このようにして MCR の総和を求めるため，ルータには専用のレジスタと加算器が必要になる（図 3 中に記載）．MCR として保持・転送しなければならない情報の量は，各ルータに搭載されているパケットバッファの数に依存するが，一般的には大きな数にはならない．各ルータが保持するパケットバッファの数は，たかだか仮想チャネルの数に入出力ポート数を乗じたものである．たとえば 2 次元トーラス網で 3 本の仮想チャネルを用いる場合は，4 つの入力ポート（上下左右）から $3 \times 4 = 12$ であり 4 ビットで表現できる．したがって $32 \times 32 = 1,024$ ノードのシステムでは，MCR の N_{ob}，N_{ab} の各成分を求めるのに 14 ビットあれば十分である．この程度であれば，MCR 等のレジスタや図 3 を実現するための回路コストは大きな負担にはならない．

図 3 の回路により N_{ob}，N_{ab} を求めるための時間的なコストについても，ここで検討したい．図 3 の回路は，元となる相互結合網のトポロジをそのまま用いて構成しており，各ルータの MCR 値が全体に反映されるまでの遅延は，$N \times N$ メッシュ/トーラスで $O(N)$ である．当該結合網の直径が同様に N $(2N)$ であることから，上述の $O(N)$ の時間コストが許容可能か否かが論点となる．

我々の先行研究[3)~5)] で得られた知見によれば，当初小規模だった輻輳が周囲の配送途中にあるパケットを吸収し，系を覆うほどに成長する．ルータはつねにパケットを転送しようとするから，輻輳領域の周辺ではブロック状態から開放されるパケットと，その領域に新たに吸収されるパケットとが併存する．輻輳領域の成長ないし消滅は，これらの収支によって決まる．このため，最初の種になる輻輳領域が発生してから，それが成長して結合網全体の性能を低下させるまでには，一般に $O(N)$ より長い時間を要する．このような輻輳の発生・成長のメカニズムを勘案すれば，図 3 の回路によるエントロピー測定の時間コスト $O(N)$ は実際上軽微なものと考えられる．

3.3 スロットリングによるパケット投入の抑制手法

前節に述べた手法により相互結合網系全体の N_{ob}，N_{ab} 値が求められれば，両者の比 $R_m = N_{ab}/N_{ob}$ を求めることにより系内の輻輳状態を判定することができる．ここで $0 \leq R_m \leq 1$ である．

本論文では，系内で 1 つの閾値（R_{th}）を設け，R_m 値が閾値 R_{th} より下回る場合に演算ノードからのパケット投入を抑制する☆．適切な R_{th} の値は，次章での評価の結果により定める．

パケットが相互結合網の系全体にわたって一様に分散している状況では，上記のような単純な方法でも制御できることが期待できる．ここでは，相互結合網内に隘路が存在し，パケット分散に著しい偏りが生じる場合について検討しておく．

たとえば，生成されるパケットの一定以上の割合のものが同一の受信先ノードとなっているホットスポット通信について考える．たとえば 32×32 の 2 次元トーラス網において全体の 5%のパケットが中心の位置にあるノードに向けて送信される状況を仮定しよう．時間の経過とともに中心ノードに向かうパケット（ホットスポットパケットと呼ぶ）のために著しい輻輳が発生する．輻輳の発生とともに上記の R_m 値が低下するから，その値が閾値 R_{th} を下回ったところでスロットリングによりノードからのパケット投入を抑制すればよい．

しかし，スロットリングにより新たなパケットの投入を抑制しても，隘路が解消されない限り R_m 値は改善されない．**図 4** はこうした事態を模式的に示した例である．相互結合網の系内にホットスポットパケットが一定数以上存在する限り，中央部が隘路となりパケット移動度が改善されず，したがって R_m 値が低いままになる．このとき，ストッロリングによりパケットの投入が抑制されているから，同図中に示したように，中央部以外のルータでは配送途中のパケットが存在せず，有効に働かない状況が生じてしまう．

こうした事態を避けるため，本論文では相互結合網系内のパケット総数が少ない場合には有効な R_m 値とは認めずスロットリングしない対策を講じる．具体的には，N_{ob} 数が系内の全ルータ数の一定割合を下回るか否かで判断する．この基準割合を P_n とし，$N \times N$

☆ たとえば，演算ノードからの入力のポートにあるパケットバッファの状態を強制的に busy にすることで実現できる．

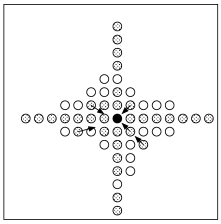


図 4 ホットスポット通信の模式図
Fig. 4 A simplified example of hot-spot traffic.

の系で $N_{ob} < P_n \cdot N^2$ のときはスロットリングを強制的に解除する．P_n を強制解除パラメータと呼ぶ．

以上のようにして，N_{ob}，N_{ab} の値からこれらの比によってエントロピーを近似した相互結合網の輻輳の度合いを表す指標を求め，その大小によって演算ノードから相互結合網へのパケットの投入を抑制する手法を，**エントロピー・スロットリング**と呼ぶ．

4. 評　　価

エントロピー・スロットリングの効果を確認するため，相互結合網シミュレータに前章で示した機能を実装した．

図 3 に示したエントロピー測定回路は，ルータ間を接続する専用の通信線路を追加せず，2 次元トーラスのトポロジに埋め込む形で実現した．パケットの送受信に用いる通信リンクを使用し，図 3 中のレジスタの値が更新された後，ルータがパケットを送受信していないタイミングを利用して必要な情報を送る．このとき，もしリンクが通常のパケットの転送に使用されていれば，空くまで待ちパケット転送にはいっさい干渉しない．このため，エントロピー測定の動作自体がパケットの転送性能を抑制することはない．このやりかたは文献 6) で行っているのと同様の手法である．この方式によれば，元の結合網のトポロジをそのまま使えるうえに専用のリンクを加える必要がないため，実装性に優れる利点がある．一方で隣接ルータ間での情報の伝送に遅延が生じるが，これについては 4.5 節で評価する．

スロットリングの強制解除を行うためのパラメータ P_n（3.3 節参照）は，以下の評価で $P_n = 0.25$（25%）とした．この値の妥当性についても 4.4 節において議論する．

ルーティングアルゴリズムは，パケットを x 軸 → y 軸の順で転送する次元順ルーティングと，パケットが x 軸，y 軸のどちらにも進めるときには受け入れ可能である方向を選択する minimum adaptive ルーティング（以降，単に適応ルーティングと呼ぶ）の 2 通りで試みた．次元順ルーティングでは，パケットは必ずチャネル 0 番地に投入され，その後ラップアラウンド線を通過するごとにチャネル番号を 1 だけ増す．仮想チャネル数は 3 である．後者の minimum adaptive ルーティングでは，我々が文献 6) で用いた仮想チャネル制御方式を用いた．すなわち，6 本の仮想チャネルを用意し，パケット投入時に送信元ノードから受信先ノードへの方向によって初期仮想チャネル番号を変える．送信元ノードを中心として受信先ノードが第 1，第 3 象限の方向にあるとき仮想チャネル番号 0 を割り当て，第 2，第 4 象限にあるときはチャネル番号 3 の仮想チャネルを割り当てる．その後は，最短経路の条件化で適応ルーティングを行い，パケットがラップアラウンド線を超えるごとに仮想チャネル番号を 1 だけ増す．

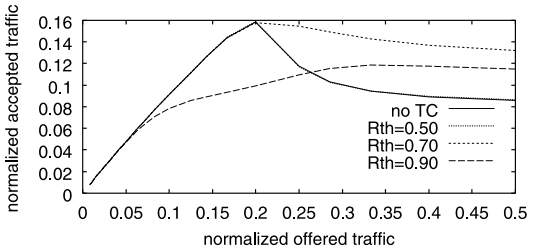
シミュレーションでは，パケット長 8 フリット，パケットバッファの容量 16 フリットとし，フロー制御として virtual cut-through 方式を用いる．パケットは進行方向をブロックされない限り，1 クロックで 1 ホップ進行する．

提案のエントロピー・スロットリングの手法により，過大な転送負荷を与えたときのスループットの低下やレイテンシの増大が抑えられることがシミュレーションにより確認できれば，提案手法の有効性が示せたことになる．

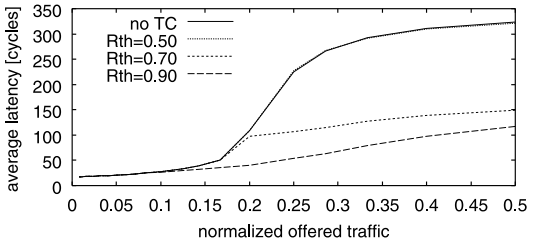
4.1 ランダム通信性能

図 5 にランダム通信での次元順ルーティングの結果を示す．“no TC” がスロットリング機能を用いない場合の性能であり，他のプロットはスロットリングの ON/OFF を行う閾値 R_{th} を図中の値に設定した場合の性能である．図 5 (a) が投入転送負荷（正規化した値で表示）に対して実際に転送されたスループット（正規化スループット）を表しており，同図 (b) が平均レイテンシを表している．以下のグラフも同様である．

図 5 (a) から，適切な閾値（この場合 $R_{th} = 0.7$）を設定することにより，過大な転送負荷の下でもスループットの大幅な低下を抑えられていることが分かる．また同図 (b) から，スループットだけではなく平均レイテンシもおよそ 1/2 まで軽減できていることが分かる．なお，図 5 で $R_{th} = 0.5$ の曲線は “no TC” とほとんど重なっており，図中では判別できない．$R_{th} = 0.9$ のときは，スロットリング制御が過度に働く結果，投入転送負荷が比較的小さい段階でスループットが飽和している．また $R_{th} = 0.5$ の場合は，スロットリングの効果がほとんど得られていない．

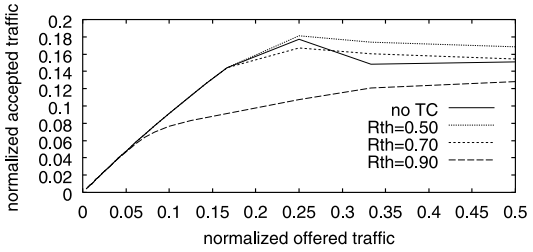


(a) 投入転送負荷 対 正規化スループット

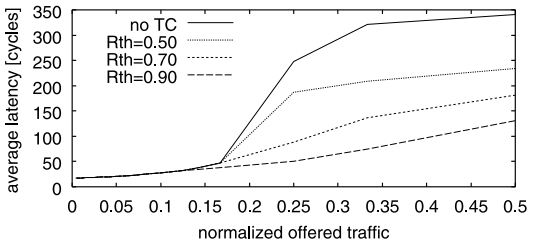


(b) 投入転送負荷 対 平均レイテンシ

図 5 次元順ルーティング，ランダム通信での性能
Fig. 5 Performance of dimension-order routing under random traffic.

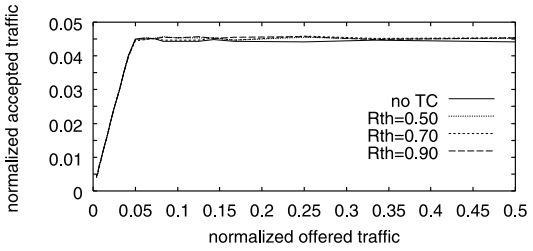


(a) 投入転送負荷 対 正規化スループット

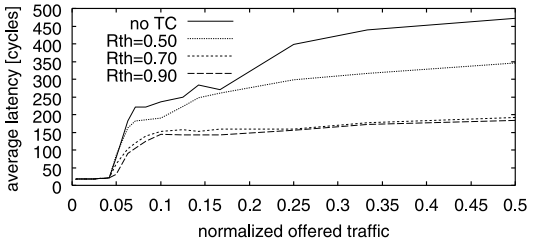


(b) 投入転送負荷 対 平均レイテンシ

図 6 適応ルーティング，ランダム通信での性能
Fig. 6 Performance of adaptive routing under random traffic.

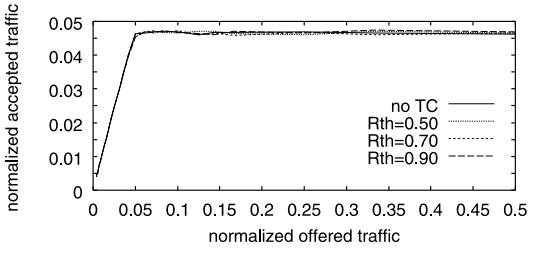


(a) 投入転送負荷 対 正規化スループット

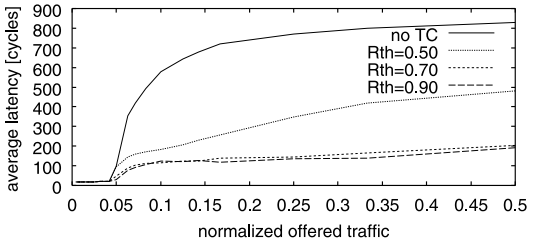


(b) 投入転送負荷 対 平均レイテンシ

図 7 次元順ルーティング，2%ホットスポット通信での性能
Fig. 7 Performance of dimension-order routing under 2% hot-spot traffic.



(a) 投入転送負荷 対 正規化スループット



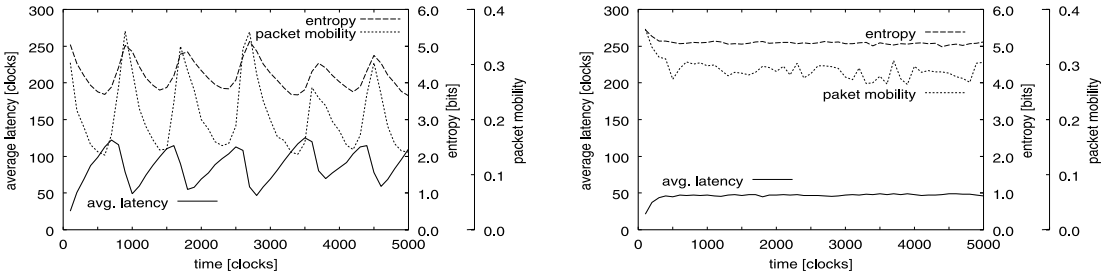
(b) 投入転送負荷 対 平均レイテンシ

図 8 適応ルーティング，2%ホットスポット通信での性能
Fig. 8 Performance of adaptive routing under 2% hot-spot traffic.

図 6 は同様にランダム通信での適応ルーティングの結果である．この場合も次元順ルーティングの場合と同様のことがいえる．ただし，$R_{th} = 0.5$ でも効果が現れていることが図 5 との違いである．図 5 の結果では，$R_{th} = 0.5$ の場合の結果がスロットリングなしの場合とほとんど変わらなかったが，図 6 では違いが明確になっている．また，ここでも $R_{th} = 0.9$ のときスロットリング制御が過度に働き比較的小さい投入転送負荷でスループットが飽和している．

4.2 ホットスポット通信性能

全体の 2%のパケットの受信先を中央のノードにしたホットスポット通信のパターンでの性能を**図 7**，**図 8** に示す．図 7 が次元順ルーティングの場合で，図 8 が適応ルーティングの場合である．図 7 (a) および図 8 (a) によれば，ホットスポット通信の場合はエントロピー・スロットリングを施してもスループットの向上は得ら



(a) 輻輳状態を生じる場合 (b) 輻輳状態を生じない場合

図 9 平均レイテンシ，エントロピー，平均移動度の時系列変化の様子（$R_{th} = 0.7$ でのエントロピー・スロットリングによる）

Fig. 9 Time sequence of average latency, entropy and average packet mobility (under Entropy Throttling with $R_{th} = 0.7$).

れていない．しかし，図 7 (b) および図 8 (b) から，適切な閾値（図 7 (b)，図 8 (b) とも $R_{th} = 0.7$）を選ぶことで輻輳による平均レイテンシの増加をおおよそ半分以下に削減できることが分かった．図 7，図 8 とも，$R_{th} = 0.5$ のときに比べ $R_{th} = 0.7$，0.9 のほうが平均レイテンシを抑えられている．

ここで，ランダム通信，ホットスポット通信ともに最適な閾値が $R_{th} = 0.7$ であることに注目したい．本手法は，相互結合網の輻輳状態をエントロピーとして表現することを理論的基盤としている．このエントロピーは定義上，系内の輻輳状態を定量的に表現するものであり，メッシュ/トーラスといったトポロジや，適応/非適応ルーティング方式，また，通信パターンに依存しない．本手法で用いている N_{ab}/N_{ob} 比は，3.1 節で述べたようにエントロピーの近似表現であり，結合網の方式によらず輻輳の状態を定量的に表現する．本評価では，ランダム通信，ホットスポット通信という 2 つの通信パターンについて最適な閾値を求め $R_{th} = 0.7$ を得ており，N_{ab}/N_{ob} 比での近似が妥当であったことが確認できる．なお，この閾値 $R_{th} = 0.7$ が，本評価と異なったトポロジや通信パターンにおいてどの程度適用可能かの検討は，今後の課題としたい．

4.3 スロットリング効果の観測

さらに，スロットリングによりパケット投入を抑制することの効果を時系列で確認するため，図 1 と同条件において，エントロピー・スロットリングを適用したときの状態を調べた．結果を**図 9** に示す．図 9 (a) から，輻輳を生じる条件では 3 章に示した方法により輻輳状況を把握し適切にスロットリング制御できていることが分かる．スロットリングを行わない図 1 (a) では，時間の経過とともに結合網が輻輳状態となったが，図 9 (a) では輻輳状態の検出によりスロットリングされている様子が分かる．

図 9 (a) のグラフが大きく振動しているのは，以下の理由によるものと考えられる．スロットリングによりパケット投入を抑制すると結合網内のパケットの数が減っていき輻輳状態が緩和される．パケット投入が抑制されている間，ノードはパケットの投入を行うことができずに，そのまま再開されるのを待つ．結合網の輻輳の緩和とともにパケットの投入がいっせいに再開され，結合網内のパケット数が急激に増える．しかし輻輳が発生・成長し，その結果が図 3 による測定エントロピー値に反映されるまでには遅延がある．こうした遅延のために振動が生じるものと考えられる．

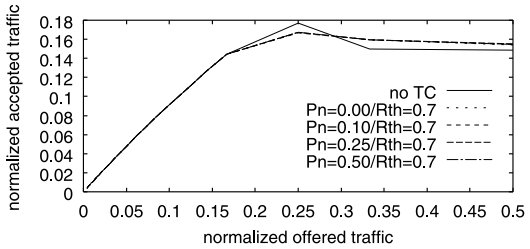
なお，輻輳を生じない状態（図 9 (b) の場合）では，スロットリングを行うことがないため，図 9 (b) と図 1 (b) とで，ほぼ同じグラフが得られている．

4.4 強制解除パラメータの評価

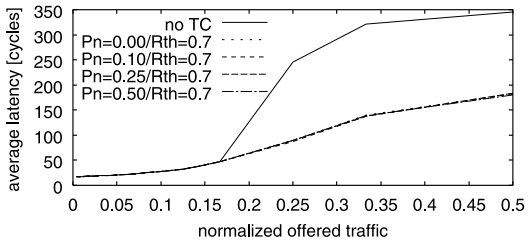
3.3 節において，著しく偏った通信パターンの場合，スロットリングにより過度にパケット投入を抑制することを避けるために導入した強制解除パラメータ P_n について評価する．

P_n の値を変えて転送特性を測定した結果を**図 10**，**図 11** に示す．図 10 はランダム通信の場合であり，図 11 は 2%ホットスポット通信の場合である．図中，“no TC” がスロットリングしない場合であり，他のプロットは強制解除パラメータ P_n とスロットリングの閾値 R_{th} の値を示している．

図 10 では，スロットリングしない場合（“no TC”）とそれ以外のプロットのみが判別可能であり，転送性能が P_n の値に依存しないことが分かる．一方，図 11 では，正規化スループットに大きな差は生じないものの，P_n の値により平均レイテンシに差が生じている．$P_n = 0$ または 0.1 のとき，投入転送負荷に対して漸増している．$P_n = 0.25$ の場合は臨界転送負荷付近で急増するが，それ以外の負荷では安定している．



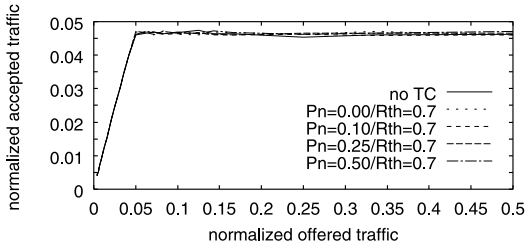
(a) 投入転送負荷 対 正規化スループット



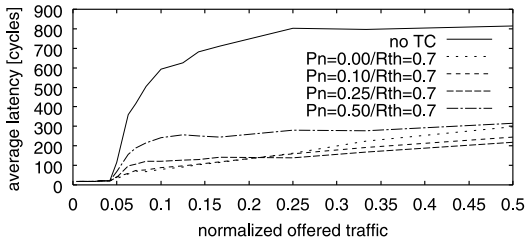
(b) 投入転送負荷 対 平均レイテンシ

図 10 強制解除パラメータの違いによる転送特性の変化（ランダム通信）

Fig. 10 Performance changes on the force-unthrottle parameter (random traffic).



(a) 投入転送負荷 対 正規化スループット



(b) 投入転送負荷 対 平均レイテンシ

図 11 強制解除パラメータの違いによる転送特性の変化（2%ホットスポット通信）

Fig. 11 Performance changes on the force-unthrottle parameter (2% hot-spot traffic).

$P_n = 0.5$ でも同様だが，$P_n = 0.25$ に比べレイテンシが大きい．以上から，$P_n = 0.25$ とするのが妥当であることが分かる．

4.5 エントロピー測定回路の応答特性

本評価では，図 3 のエントロピー測定回路を専用の接続信号線を用いず，物理リンクを通常パケットの転送と共用し，通常パケットを転送していない時間にのみエントロピー測定回路の機能を働かせるようにした．このように結合網の資源を有効に利用することで低コストな制御が可能になるが，一方で，通常パケットを優先することによる影響を評価しておく必要がある．

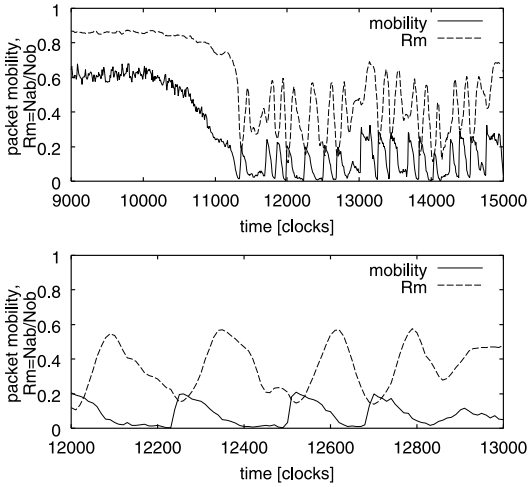


図 12 通信パターンを変えた場合の応答特性

Fig. 12 Transient behaviors.

投入転送負荷を 0.1 とし，シミュレーション開始から 10,000 サイクル間ランダム通信を行い，その後 2%のホットスポット通信を開始した．この通信パターンの切替え前後の時系列状況を**図 12** に示す．下のグラフは上の一部を時間拡大表示したものである．パケット移動度（mobility）の変化に対して，エントロピー測定回路による $R_m = N_{ab}/N_{ob}$ 値は，おおよそ 100 クロック以下となっている．この遅延量は結合網の直径 32 よりも大きいが，図 12 から分かるように，系の挙動変化（この場合 mobility 値の変化）に十分追従できており，エントロピー計測の時間コストが軽微であることが分かる．

5. 関 連 研 究

本論文は，(1) 輻輳の度合いを定量的に表現する指標としてエントロピーを導入し，(2) これをスロットリングに使用したこと，を主な新規性としている．以下，上記の観点から関連する研究についてまとめる．

エントロピーを指標として通信網の制御に供する研究として，文献 7) があげられる．インターネット等でのルーティング経路を選択するためのパラメータ群によりエントロピーを定義し，これを最大化するように制御する手法である．これは，本論文でのパケットの移動度を表すエントロピーとは異なるうえ，輻輳の問題の解決を図った手法ではない．

これに対し，スロットリングに関してはいくつかの既往研究がある．Baydal らは，個々のルータが内部の局所的な情報を用いてスロットリング制御する手法 U-Channels，ALO，INC を提案している[8]．U-Channels，ALO はブロックされていないチャネルの数を用い，INC は一定時間内のパケット流量を計る手法である．両者とも局所的な情報のみであり系全体の輻輳状況を表していない．López らの DRIL[9] や Obaidat らによる CLIC[10] も同様にルータ内部の局所的な情報を用いるものである．輻輳箇所に流入するパケットを抑制することが必要であるため，局所情報では十分ではなく，本論文の手法が有利である．

一方，Thottethodi らの手法[11] はバッファ内にたまっているパケットの量の多寡によりスロットリングするが，meta-packet により系内の輻輳情報を収集し，スロットリングの ON/OFF に用いる閾値を動的に変える．最大の転送性能が得られる条件を求めるために，hill climbing により閾値を摂動しながら，さらに局所解（local maxima）に陥るのを防いでいる．この方法は，系全体の情報を収集する点で本論文の方式と類似するが，一方で，meta-packet による副作用（バンド幅の使用と遅延）を避けるため，一般のパケット用とは別のリンクを前提としているほか，摂動により最適閾値が求まるまでの時間が本論文と比較して長いことの 2 点が大きく異なる．本論文では一般パケットとリンクを共用しながら効率的に輻輳情報を収集している．また閾値等のパラメータの微調整が不要であり応答が速く安定した効果が得られている．文献 11) では 5,000〜10,000 サイクルの周期で性能が大きく揺らぐ現象が見られるが本論文の方式の揺らぎは 1,000 サイクル以下であり（図 9 参照），素早く安定した制御が行えていることが分かる☆．

6. おわりに

本論文では，過重な転送負荷の条件下でも輻輳の発生を抑え，パケットの流量を最大限に保つことで，輻輳にともなうスループットやレイテンシの著しい悪化を防ぐ手法について検討した．

まず，相互結合網系の輻輳の度合いを定量的に表現できるエントロピーを導入した．そしてエントロピー指標を大規模直接網において扱いやすい形式に近似し，1 個以上のパケット片を保持しているパケットバッファの個数 N_{ob} と，そのうち先頭パケットを出力しているものの個数 N_{ab} との比を指標とした．この近似により，輻輳指標が 2 次元メッシュ/トーラス網で効果的に求められることを示した．そして一定の閾値 R_{th} を設け N_{ab}/N_{ob} の値が R_{th} を下回った場合に新たなパケットの投入を抑制するエントロピー・スロットリングの手法を提案した．

シミュレーション評価の結果，提案手法によれば輻輳にともなうスループットやレイテンシの悪化を効果的に防げることが明らかになった．負荷の状況や転送パターン等に追随する微調整も不要であり，比較的短かい時間で安定した効果が得られる．

謝辞 有益なコメントをいただいた査読者に感謝いたします．本研究は，一部日本学術振興会科学研究費補助金（基盤研究（B）18300014，同（C）16500023，若手研究（B）17700047）の援助による．

☆ さらに，文献 11) の評価は 16 × 16 トーラスの場合であり，本論文は 32 × 32 トーラスであることも付記しておく．

参考文献

1) Duato, J., Yalamanchili, S. and Ni, L.: *Interconnection Networks: An Engineering Approach*, Morgan Kaufmann Pub. (2003).
2) Dally, W.J. and Towles, B.: *Principles and Practices of Interconnection Networks*, Morgan Kaufmann Pub. (2004).
3) 横田隆史，大津金光，古川文人，馬場敬信：セルオートマトンによる大規模相互結合網シミュレーションの試み，信学技報，Vol.105, No.453, pp.31–36 (2005).
4) 横田隆史，大津金光，古川文人，馬場敬信：セルオートマトンによる相互結合網の輻輳の解析，情報処理学会論文誌：コンピューティングシステム，Vol.47, No.SIG 7 (ACS 14), pp.21–42 (2006).
5) Yokota, T., Ootsu, K., Furukawa, F. and Baba, T.: Phase Transition Phenomena in Interconnection Networks of Massively Parallel Computers, *Journal of Physical Society of Japan*, Vol.75, No.7, p.074801 (2006).
6) 横田隆史，西谷雅史，大津金光，古川文人，馬場敬信：大域的な情報を用いる相互結合網方式 Cross-Line，情報処理学会論文誌：コンピューティングシステム，Vol.46, No.SIG 16 (ACS 12), pp.28–42 (2005).
7) Bernstein, H. J.: Some Comments on Highly Dynamic Network Routing, Technical Report 371, Computer Science Dept., New York Univ. (1988).
8) Baydal, E., López, P. and Duato, J.: A Family of Mechanisms for Congestion Control in Wormhole Networks, *IEEE Trans. Parallel and Distributed Systems*, Vol.16, No.9, pp.772–784 (2005).
9) López, P., Martínez, J.M. and Duato, J.:

DRIL: Dynamically Reduced Message Injection Limitation Mechanism for Wormhole Networks, *Proc. 1998 International Conference on Parallel Processing*, pp.535–562 (1998).
10) Obaidat, M.S., Al-Awwami, Z.H. and Al-Mulhem, M.: A new injection limitation mechanism for wormhole networks, *Computer Communications*, Vol.25, pp.997–1008 (2002).
11) Thottethodi, M., Lebeck, A.R. and Mukherjee, S.S.: Exploiting Global Knowledge to Achieve Self-Tuned Congestion Sontrol for k-Ary n-Cube Networks, *IEEE Trans. Parallel and Distributed Systems*, Vol.15, No.3, pp.257–272 (2004).

(平成 18 年 1 月 26 日受付)
(平成 18 年 5 月 2 日採録)

横田　隆史（正会員）

1983 年慶應義塾大学工学部電気工学科卒業．1985 年同大学院電気工学専攻修士課程修了．同年三菱電機（株）に入社，中央研究所，先端技術総合研究所，産業システム研究所に所属．主席研究員．1993 年 12 月から 1997 年 3 月まで新情報処理開発機構（RWCP）に出向．2001 年 4 月より宇都宮大学工学部助教授．計算機アーキテクチャ，設計方法論等の研究に従事．工学博士．ICCD Outstanding Paper Award（1995），FPGA/PLD Design Conference 審査委員特別賞（2002），PDCAT'05 Outstanding Paper Award（2005）各受賞．電子情報通信学会，IEEE 各会員．

大津　金光（正会員）

1993 年東京大学理学部情報科学科卒業．1995 年同大学大学院修士課程修了．1997 年同大学院博士課程退学，同年より宇都宮大学工学部助手となり現在に至る．計算機システムの高性能化に関すること，特にマルチスレッドアーキテクチャ，バイナリ変換処理，実行時最適化等に興味を持つ．

古川　文人（正会員）

1998 年宇都宮大学工学部情報工学科卒業．2000 年同大学大学院博士前期課程修了．2003 年同大学院博士後期課程修了．同年 4 月より宇都宮大学ベンチャー・ビジネス・ラボラトリー非常勤研究員．2005 年 4 月より帝京大学ラーニングテクノロジー開発室助手．博士（工学）．高性能計算機システム，授業改善のためのラーニングテクノロジーに関する研究に従事．

馬場　敬信（フェロー）

1970 年京都大学工学部数理工学科卒業．1975 年同大学大学院博士課程単位取得退学．同年より電気通信大学助手，講師を経て，現在宇都宮大学工学部教授．工学博士．1982 年より 1 年間メリーランド大学客員教授．計算機アーキテクチャ，並列処理等の研究に従事．1992 年情報処理学会 Best Author 賞，2002 年 FPGA/PLD Design Conference 審査委員特別賞，PDCS2002 国際会議 Best Paper Award 各受賞．著書 "Microprogrammable Parallel Computer"（MIT Press），『コンピュータアーキテクチャ』（改定 2 版）（オーム社），『コンピュータのしくみを理解するための 10 章』（技術評論社）等．電子情報通信学会，IEEE 各会員．