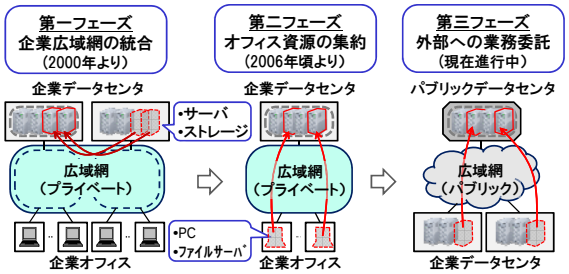
Solution Approaches for Wide-Area Distributed Systems toward Integration of Enterprise Networks and Computing Resources

小川 祐紀雄

大阪大学 大学院情報科学研究科
情報ネットワーク学専攻　先進ネットワークアーキテクチャ講座
（（株）日立製作所 横浜研究所 yukio.ogawa.xq@hitachi.com）

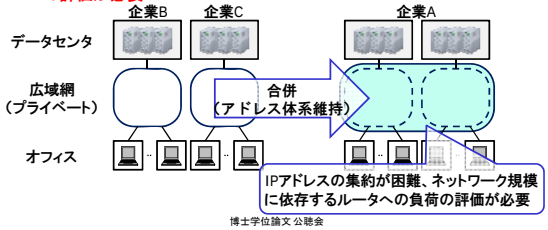
背景：企業ネットワークと計算資源の統合

- 全体コスト削減のため計算資源をデータセンタに集約
- 広域網の性能向上とともに三つのフェーズを経て進展



背景：統合に向けた広域網の課題（第一フェーズ）

- 第一フェーズ：合併にともなう企業広域網の大規模化
  - ✓ 企業レベルでは、IPネットワークは適切に構成すればスケーラブル
  - ✓ 既存アドレス体系を引き継ぐ場合は再構成が困難、
    ネットワーク規模に依存するルータへの負荷（トラヒック量、CPU利用率）の評価が必要

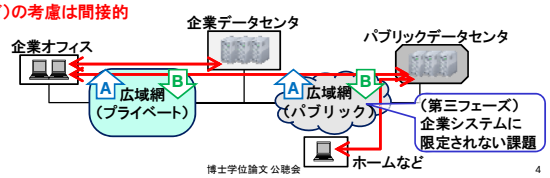


背景：統合に向けた広域網の課題（第二・三フェーズ）

- 第二・第三フェーズ：アプリケーションが広域網を経由

A 広域網の性能がアプリケーションの性能に影響
- ✓TCP改善方式はバルクデータフロー（FTPなど）が主対象
- ✓対話式データフロー（SSHなど）や混合フロー（リモートデスクトップなど）の考慮は間接的

B アプリケーショントラヒックが広域網の負荷・電力を増大
- ✓ スリープ部位を増やすための最適ルーティング検討が主流
- ✓ 地域性を表すトラヒックマトリクスを考慮した電力評価は対象外



研究の目的、博士論文の構成

- 各フェーズの未検討の課題に対し解決手法を確立

Chapter 2
Performance of a Large Enterprise Network for Updating Routing Information

大規模企業ネットワークの
ルーティング情報更新に必要なネットワーク性能

1. Y. Ogawa, T. Hirata, K. Takamura, K. Yamaha, S. Saitou, K. Iwanaga, and T. Koita, "Estimating the performance of a large enterprise network for updating routing information," *IEICE TRANSACTIONS on Communications*, vol. E88-B, pp. 2054-2061, May 2005.
2. Y. Ogawa, A. Nakaya, K. Takamura, K. Yamaha, S. Saitou, K. Iwanaga, and T. Koita, "Estimating the performance of a large enterprise network for the updating of routing information," in *Proceedings of IEEE Workshop on IP Operations and Management (IPOM 2002)*, pp. 161-165, Oct. 2002.

大規模企業ネットワークのルーティング情報更新に必要な性能

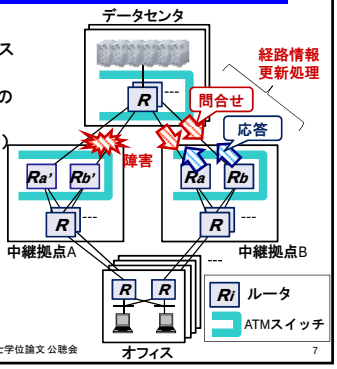
課題：広域網統合によるネットワーク規模の増大

- 本章の対象：国内*A*銀行
  - ✓ データセンタ～中継拠点～オフィス
  - ✓ 拠点数：2,100、ルータ数：5,000
  - ✓ 中継拠点ルータ(*Ra*,*Rb*,*Ra'*,*Rb'*)の隣接ルータ数：40～50

（世界最大規模の企業IPネットワーク）

- 経路情報更新量も増加
  - ✓ *Ra*～*Rb'* では処理負荷（パケット転送、経路情報処理）により経路問合せへの応答が遅延
  - ✓ 問合せ側ルータでタイムアウト発生（180秒）、経路情報リセット（通信不通障害が発生）

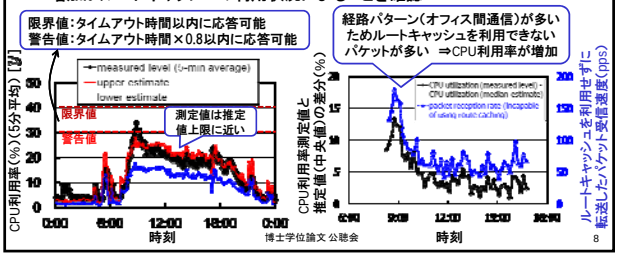
タイムアウトを発生させないルータの処理負荷推定が必要に



大規模企業ネットワークのルーティング情報更新に必要な性能

中継拠点ルータの処理負荷の評価

- タイムアウト時間内に応答可能な処理負荷（CPU利用率）、パケット受信によるCPU利用率の増加をモデル化、パラメータを実システムでの実験から算出
- 広域網統合直後に中継拠点ルータのCPU利用率測定値と、しきい値、パケット受信速度からの推定値を比較、警告値をこえるルータについてCPU利用率増加がルートキャッシュの利用状況によることを確認



大規模企業ネットワークのルーティング情報更新に必要な性能

まとめ

- 研究成果
  - ✓ 大規模企業IPネットワークの安定稼働のために、ルーティング情報更新を完了できるルータのCPU負荷を評価
  - ✓ CPU負荷しきい値をこえるルータについて、その増加の原因がルートキャッシュの利用状況によることを確認

（しきい値の設定により、事前の対策（トラヒックの時間分散、上位機種・モジュールへの更改）が可能に）

Chapter 3
Performance of a Thin-Client System
in a WAN Environment

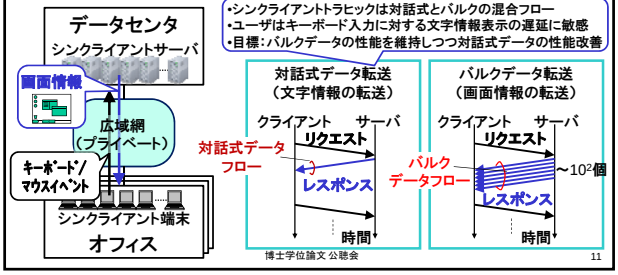
広域ネットワーク環境における
シンクライアントシステムの性能

1. Y. Ogawa, G. Hasegawa, and M. Murata, "A transport-layer approach for improving thin-client performance in a WAN environment," *International Journal of Internet Protocol Technology*, vol. 6, pp. 172-183, Nov. 2011.
2. Y. Ogawa, G. Hasegawa, and M. Murata, "Transport-layer optimization for thin-client systems," in *Proceedings of IEEE Workshop on Communications Quality and Reliability (CQR 2007)*, May 2007.
3. Y. Ogawa, G. Hasegawa, and M. Murata, "A transport layer approach for improving interactive user experience on thin clients," in *Proceedings of Australasian Telecommunication Networks and Applications Conference (ATNAC 2009)*, Nov. 2009.
4. Y. Ogawa, G. Hasegawa, and M. Murata, "Delay analysis and transport-layer optimization for improving performance of thin-client traffic," *Technical Report of IEICE*, vol. IN2008-56, pp. 75-80, Sept. 2008. (in Japanese).

広域ネットワーク環境におけるシンクライアントシステムの性能

課題：PC集約、広域網経由通信による性能低下

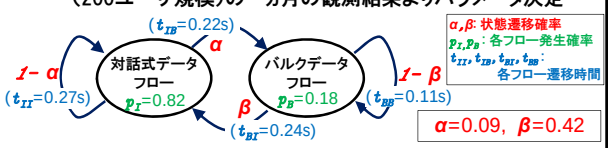
- データセンタにPC集約（シンクライアント化）
- シンクライアント通信はTCPベース、広域化により性能低下
  - ✓ 広域網の遅延が鍵に、特にTCPの作用による影響は未検討



広域ネットワーク環境におけるシンクライアントシステムの性能

シンクライアントトラヒックのモデル化

- レスポンストラヒックを二状態系でモデル化、実システム（200ユーザ規模）の一ヵ月の観測結果よりパラメータ決定



•フロー数では対話式データが支配的
•バイト数ではバルクデータが支配的

	対話式データフロー	バルクデータフロー
フロー数	82%(=p_I)	18%(=p_B)
パケット数	39%（1個/フロー）	61%（7.4個/フロー(=c_B)）
バイト数	8%（137バイト/パケット）	92%（936バイト/パケット）

広域ネットワーク環境におけるシンクライアントシステムの性能

ユーザビリティの基準

- 「人間の応答時間」を基準に、対話式データフロー(レスポンスパケット)転送時間に対し二段階のしきい値[1]を設定
  - ✓「無知覚」 $< (150 - t_0)$ ms
  - ✓「生産性低下」 $> (1000 - t_0)$ ms　t_0：リクエストパケット転送時間
- 二状態系モデルより、対話式データのしきい値以上のレスポンスパケット発生頻度($r_{I_{150}}, r_{I_{1000}}$)から、定常状態において、しきい値を超えるレスポンスの発生サイクル時間($T_{I_{150}}$, $T_{I_{1000}}$)を導出

遷移確率　遷移時間

$$T_{I_k} = \frac{(1-\alpha)^{I_{II}}}{r_{I_k}} + \frac{\alpha t_{IB}}{r_{I_k}} + \frac{\beta p_B^{I_{BI}}}{c_B p_I r_{I_k}} + \frac{(1-\beta) p_B^{I_{BB}}}{c_B p_I r_{I_k}} \quad (I_k = I_{150} \text{ or } I_{1000})$$

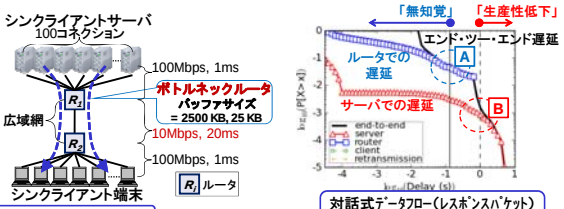
一サイクル中の対話式データフロー数
〃　バルクデータフロー数

[1] N. Tolia, D. G. Andersen, and M. Satyanarayanan, "Quantifying interactive user experience on thin clients," IEEE Computer, vol. 39, pp. 46-52, Mar. 2006.

広域ネットワーク環境におけるシンクライアントシステムの性能

計算機シミュレーションによる遅延の解析

- 実データを用いて、対話式データフローの遅延の原因
  - A「無知覚」(約150ms)付近：ボトルネックルータ(広域網接続ルータ)における他コネクションのバルクデータによるバッファリング遅延
  - B「生産性低下」(約1000ms)付近：サーバにおける自コネクションのバルクデータによるバッファリング遅延



広域ネットワーク環境におけるシンクライアントシステムの性能

既存方式を適用したときの問題点

- 既存方式を適用
  - ✓ ボトルネックルータのバッファサイズ大：対話式データフローの優先制御
    - 対話式データの割合(バイト数)は8%、バルクデータは非優先でも遅延は小
  - ✓ ボトルネックルータのバッファサイズ小：TCP SACKオプションの適用
    - パケットドロップからの早期回復により、サーバでのバッファリング時間低下

⇩

- 「無知覚」(約150ms)をこえるレスポンス発生頻度は低下
- 「生産性低下」(約1000ms)をこえるレスポンス発生頻度は増加
  - ✓ ボトルネックルータのバッファサイズ大：バルクデータと対話式データの切替りで、パケットドロップなしにタイムアウトが発生
    - ルータでのバルクデータ入力によるキュー長の急な変化に、サーバのRTO が追随しない
  - ✓ ボトルネックルータのバッファサイズ小：タイムアウト発生後にパケットを送信しない区間が発生
    - TCP SACKの実装により再送タイムアウト発生後に輻輳ウィンドウサイズを増加させず

広域ネットワーク環境におけるシンクライアントシステムの性能

既存方式改善のための提案手法

- 対話式/バルクデータフローの切り替わり時にRTOが追随するよう、サーバのRTO算出時の重み付け定数を変更

RTO:再送タイムアウト時間

$$srtt \leftarrow (1-g)srtt + g \cdot rtt$$

加重平均した往復遅延時間　重み付け定数　往復遅延時間

$$g = \begin{cases} 7/8 & (rtt \geq srtt) \\ 1/8 & (rtt < srtt) \end{cases}$$

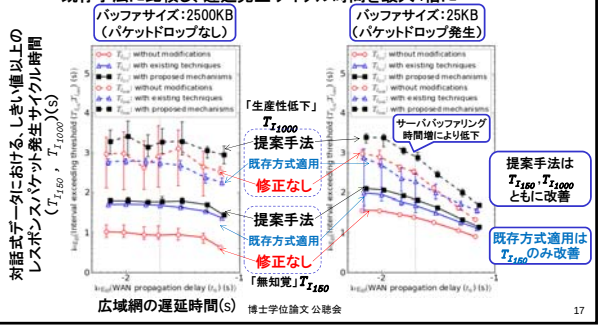
サーバのRTOは、
✓バルクデータフロー切替り時にすばやく増加
✓対話式データフローに戻ってもすぐに縮小しない

- TCP SACKの隘路を回避するよう、タイムアウト発生後に一時的にTCP SACK制御をオフに(パケットドロップ回復後に再度TCP SACK制御をオン)

広域ネットワーク環境におけるシンクライアントシステムの性能

広域網の遅延の影響(シミュレーション実験結果まとめ)

- 提案手法適用により、「生産性低下」(約1000ms)のしきい値において既存手法に比較し、遅延発生サイクル時間を最大4倍に



広域ネットワーク環境におけるシンクライアントシステムの性能

まとめ

- 研究成果
  - ✓ シンクライアント通信のレスポンストラヒックを対話式/バルクデータの混合フローによる二状態系でモデル化
  - ✓ 対話式データフローの遅延要因を分析し、既存方式の適用では改善しないことを指摘
  - ✓ 提案手法により対話式データフローの遅延発生サイクルを評価環境では最大4倍(発生頻度を1/4)に改善

(ただし、対話式データに対するバルクデータの影響を完全になくすためには、両者のコネクション分離が必要)

Chapter 4

Power Consumption of a WAN
with the Aid of Distributed Computing

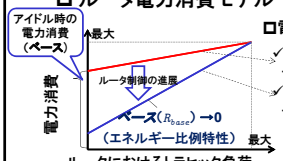
分散コンピューティングを適用した広域ネットワークの電力消費

1. Y. Ogawa, G. Hasegawa, and M. Murata, "Power consumption evaluation of distributed computing network considering traffic locality," *IEICE TRANSACTIONS on Communications*, 2012. [submitted for publication].
2. Y. Ogawa, G. Hasegawa, and M. Murata, "Effect of traffic locality on power consumption of distributed computing network," in *Proceedings of 9th International Conference on Communications (COMM 2012)*, June 2012. [submitted for publication].
3. Y. Ogawa, G. Hasegawa, M. Murata, and S. Nishimura, "Performance evaluation of distributed computing environment considering power consumption," *Technical Report of IEICE*, vol. IN2009-172, pp. 169-174, Mar. 2010. (in Japanese).

分散コンピューティングを適用した広域ネットワークの電力消費

課題：広域網の電力消費増加

- □ 大量の計算資源がパブリックデータセンタに集約
  - ✓ 広域網のトラヒック負荷、電力消費が増加
  - ✓ 分散コンピューティング（例、CDN）による負荷削減 ⇒ 電力消費削減
- □ アプリケーションによりトラヒックマトリクスは変化
  - ✓ 例、スマートシティでは「地産地消」：地域の生成データを地域で利用
  - ✓ トラヒックの地域性の影響を検討
- □ ルータ電力消費モデル －エネルギー比例特性[2]



- □ 電力消費削減のアプローチ
  - ✓ トラヒック負荷に依存せず
    - 経路制御によりスリープ可能なノードを増やす
  - ✓ トラヒック負荷に依存
    - 分散コンピューティングによりトラヒック量削減

本章のアプローチ

[2] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "A power benchmarking framework for network devices," in *Proceedings of the 8th International IFIP-TC 6 Networking Conference (NETWORKING 2009)*, pp. 795-808, May 2009.

分散コンピューティングを適用した広域ネットワークの電力消費

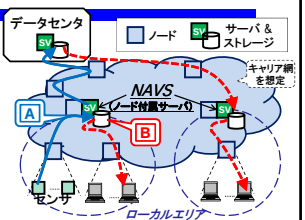
アプリケーションシナリオと電力消費モデル

- □ ノードに取り付けたサーバ機能（***NAVS***）によりデータセンタとのトラヒック削減

NAVS：Node-attached virtual server（ノード付属サーバ）

A アップロード：データ圧縮

B ダウンロード：データキャッシング



システム電力消費量を最小にするNAVS（P個）の組合せ最適化問題

システム全電力消費量 / ノードiからjへのリンクの電力消費量 / ノードiの電力消費量 / ノードiのNAVSの電力消費量 / データセンタサーバの電力消費量

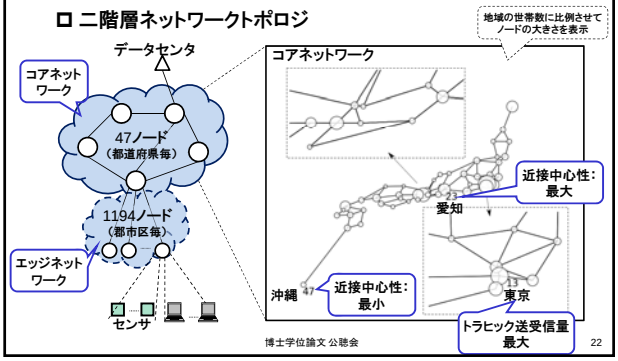
$$E = \sum_{i\in N}\sum_{j\in N} a_{ij}(f_{ij})x_{ij} + \sum_{i\in N} b_i(g_i) + \left(\sum_{i\in N} c_i(g_i)y_i + c_e(g_0)\right) \quad (\forall e\in N)$$

ノード集合 / ノードiからjへのリンクのトラヒック量 / ノードiのトラヒック量 / データセンタサーバのトラヒック量

分散コンピューティングを適用した広域ネットワークの電力消費

ネットワークモデル

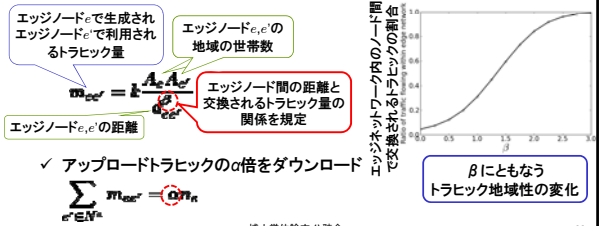
- □ 二階層ネットワークトポロジ



分散コンピューティングを適用した広域ネットワークの電力消費

トラヒックモデル

- □ アップロードトラヒック
  - ✓ 各エッジノードから地域（郡市区）の世帯数に比例した量（w_e）を生成
- □ ダウンロードトラヒック
  - ✓ 需要と供給の関係を重力モデルによりモデル化



- ✓ アップロードトラヒックのα倍をダウンロード

$$\sum_{e'\in N_e} m_{ee'} = \alpha m_e$$

分散コンピューティングを適用した広域ネットワークの電力消費

評価結果：トラヒック地域性の影響

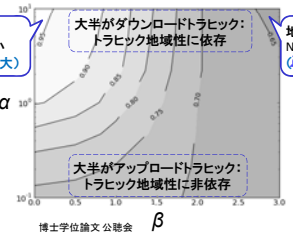
- □ 総電力消費を最小化するP個のNAVSの配置を貪欲交換法により求める
  - ✓ NAVS におけるアップロードトラヒックの圧縮率は0.1
  - ✓ トラヒック量総和は2.6Tbps（2011年の日本のインターネットトラヒック量）
  - ✓ コアルータ：Cisco® CRS-3シリーズ、エッジルータ：Juniper MXシリーズ
  - ✓ 第一NAVSはデータセンタが接続するコアルータに固定（データセンタへの冗長な経路を回避）
  - ✓ データセンタは東京に設置、ルータのアイドル時の電力消費の割合（R_{base}）は0.01

α,βにともなうE_{min}/E_1の変化

地域性小：NAVS効果小（E_{min}は最大）

大半がダウンロードトラヒック：トラヒック地域性に依存

地域性大：NAVS効果大（E_{min}は最小）



- E_1：データセンタにのみNAVSを設置したときの電力消費（初期状態）
- E_{min}：NAVSを必要個数設置し最小化されたときの電力消費

分散コンピューティングを適用した広域ネットワークの電力消費

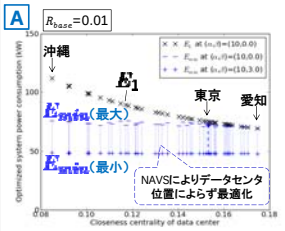
データセンタ位置、ルータのエネルギー比例特性の影響

A NAVS分散配置によりデータセンタへの転送にかかる電力を緩和

B エネルギー比例特性小：NAVSに電力削減効果よる効果小

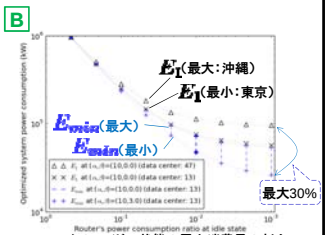
エネルギー比例特性大：NAVSによる効果大（配置なしに比較し最大約30%）

A



データセンタ接続コアノードの中心近接性

B



ルータのアイドル状態の電力消費量の割合（R_{base}）

分散コンピューティングを適用した広域ネットワークの電力消費

まとめ

□ 研究成果

- ✓ 分散コンピューティングによる広域網の電力削減効果に着目、トラヒック地域性の観点から評価。システム電力消費、広域網トポロジー、トラヒックをモデル化し効果を算出
- ✓ アップロードトラヒックが支配的であれば、トラヒックの地域性に関係せず。ダウンロードトラヒックが支配的であれば、トラヒック地域性の増加に従い、電力削減効果が大
- ✓ ルータのエネルギー比例特性が向上するにつれ、データセンタ位置の影響があらわれるが、分散コンピューティングにより影響をなくすことができ、評価環境では最大約30%に

本研究のまとめ、今後の課題

□ 企業網や計算資源の統合に向けた各フェーズの未解決の課題に対し、モデル化を通じて事象の定量化を実施

- ✓ 第一フェーズ（企業の広域網統合）：
  ルータの経路情報処理のモデル化を行い、ルータの経路情報処理スケーラビリティを定量化
- ✓ 第二フェーズ（オフィスPCのデータセンタへの集約）：
  シンクライアントシステムのレスポンストラヒックを対話式データとバルクデータの二状態系でモデル化、単一コネクション中の対話式データに対するバルクデータの影響を示した
- ✓ 第三フェーズ（パブリックデータセンタへの資源集中）：
  広域ネットワークの電力消費やトラヒック地域性をモデル化、システム電力消費に対するトラヒック地域性やデータセンタ位置などの影響を示した

□ 今後の課題

- ✓ 第三フェーズ後に、少数のデータセンタへ資源が過剰に集中したときの資源の最適配置