成蹊大学理工学研究報告
J. Fac. Sci. Tech., Seikei Univ.
Vol.47 No.1 (2010) pp.33-38

代理処理の範囲拡大によるサーバ電力使用量削減のための基礎検討

長名　保範[*1]，鈴木　顕[*2]，栗林　伸一[*3]

Preliminary study on a power reduction method for servers by an active use of proxy operation

Yasunori OSANA[*1], Ken SUZUKI[*2], Shin-ichi KURIBAYASHI[*3],

ABSTRACT： Most servers on the Internet is in 365x24 operation to provide service at anytime, while computers usually consume 50 to 100W even in their idle state. This work proposes a novel method to reduce servers' power consumption. In this method, an intelligent NIC with simplified TCP/IP support controls the sleep/awake state of the server. While the server is in sleep state, the NIC processes the arrival packet then respond by itself, transfer to the designated proxy server, or make the server awaken. This paper shows the result of preliminary statistics based evaluation of the method and reports current status of NIC implementation. From the evaluation, fraction of available time to sleep extends from 5-10% to 30-50% with this proposed method even on busier server.

Keywords：power reduction, servers, proxy operation.

（Received March 25, 2010）

１．はじめに

大学などの研究機関においては，研究成果公開などのために数多くのサーバが稼働しているが，これらのサーバの提供するサービスに対するアクセスは実際にはそれほど多くない場合もあり，省エネルギーの観点からは好ましくない。しかしその一方で，これらのサーバは高度に専門的なサービスを提供するためソフトウェアの構成が複雑である場合が多く，共用のサーバに集約することが困難である。

このようなサーバ群の電源を動的にオン・オフすることができれば消費電力を抑制することが可能になるが，その場合にはサーバの停止中も何らかの方法でネットワークの外部からみたリンク状態を維持し，必要に応じて起動する方法が必要になる[1)]。従来，ルータなど外部のネットワーク機器上にトラヒック監視機能をつけておき，長時間サービス要求がない場合にはサーバをスリープさせ，必要になったらネットワーク経由で起動の指令を行う方式[2)]などが採られてきたが，この方式ではネットワーク内部でのアクセスに対応できないなどの問題があり，より細かい粒度で柔軟な制御を行う方式が必要であった。

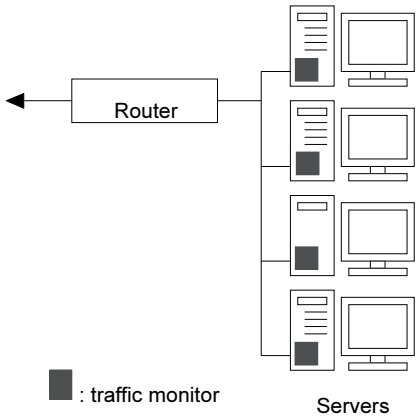

図１　提案手法 : 分散トラヒック監視

本研究では，IP レイヤの処理能力をもつインテリジェントなネットワークインタフェイスカード(NIC)を開発することで，図 1 のようにサーバごとにトラヒック監視を行い，必要のない場合には積極的にサーバをスリープさせることで消費電力の低減を図る方式の実現を目指している。この方式では，サーバのスリープ中にパケットが到着した場合に (1) NIC だけで処理を行う，(2) 他のサーバへ転送して代理処理を行う，(3) サーバのスリープ状態を解除してソフトウェアによる処理を行う，の 3

[*1]：情報科学科助教
[*2]：情報科学科学生
[*3]：情報科学科教授 (kuribayashi@st.seikei.ac.jp)

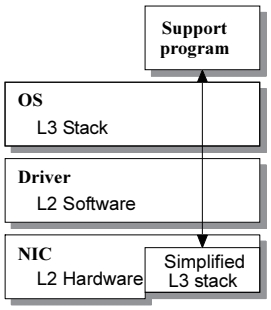

図２　提案するシステムのレイヤ構成

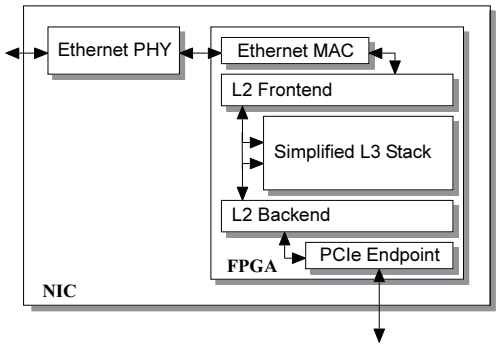

図３　提案する NIC の構成

通りを想定しており，特に(1)と(2)の処理範囲を拡大することで消費電力の低減を図ることを主眼としている。

全体の方式概要については既に文献[3)4)]に発表しているので，本稿では，開発中のNICの構成について述べるとともに，学内で測定したトラヒック統計に基づく，方式の有効性検証の結果を報告する。

２．提案手法とシステム構成

本研究報告で提案する手法の主眼は，NIC をインテリジェント化することで，ホストのスリープ中は NIC がホストの代わりに応答したり，あるいは NIC が他のホストに代理応答を依頼し，必要に応じてホストをスリープ解除することで，(1) 外部からの処理要求にシームレスに対応できることと，(2) 同様にホスト上で動作しているソフトウェアにも特別に手を加えることなく，スリープ状態と動作状態の切り替えを実現することにある。

提案するシステムの構成を図 2, NIC の構成を図 3 に示す。NIC は専用のハードウェアを実現する必要があるため，任意の論理回路を構成可能な LSI である FPGA (Field Programmable Gate Array)を搭載した PCI-Express カードを用いる。現在用いている FPGA ボードは Xilinx 社の Virtex-5 LXT[5)]を搭載した東京エレクトロンデバイス社製の PCI-Express 評価ボード(図 4)であり，PCI-Express と Gigabit Ethernet MAC を FPGA 上に搭載している。なお，最新の低価格帯デバイスである Spartan-6 LXT[6)]や Cyclone IV GX[7)] などでも同様の構成を実現できる。

FPGA のユーザロジック部には Ethernet MAC と PCI-Express を接続するための回路(L2 Frontend/Backend)を構成し，これはデバイスドライバ及び OS の IP over Ethernet スタックを経由して，通常の Gigabit Ethernet NIC と同様の働きをする。FPGA を用いて専用回路によるパケット処理を行うことで，マイクロプロセッサよりも低い動作周波数での処理が可能になり，電力効率が向上する。また，専用回路ではあるが再構成が可能なため，必要なプロトコルのための回路だけを選んで搭載したり，新たな機能を追加したりすることもできる。FPGA の消費電力は，PCI Express-Gigabit Ethernet のインタフェイスだけを構成した場合に 1.5W 程度とされており，追加の回路を含めても 2ー3W の範囲に収まると考えられる。

また，FPGA 上には簡単な L3 制御回路 (Simplified L3 stack) を搭載し，これを用いて通信状態の監視や，ホストがスリープしている間の代理応答とホストのスリープ解除などの制御を行う。この，FPGA 上の L3 制御回路と情報をやりとりするためのサポートソフトウェアをホスト OS 上で起動しておく。

NIC 上の L3 制御回路は，送受信するパケットを監視したり，ホストのスリープ中には到着パケットの処理を

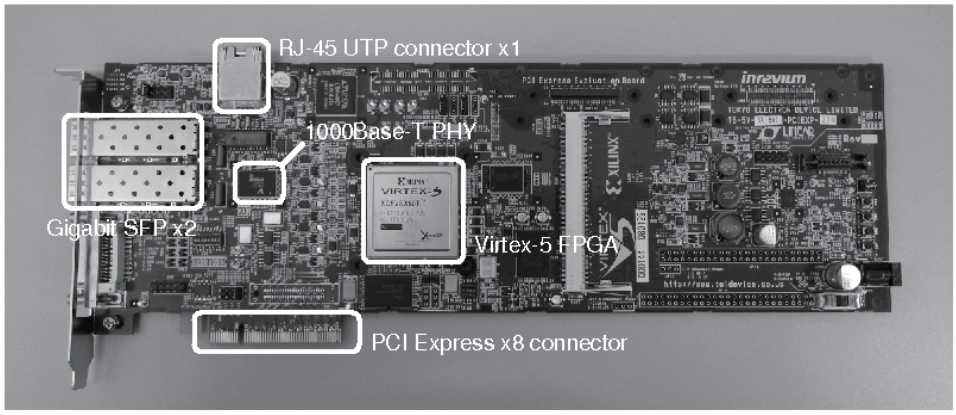

図４　Virtex-5 搭載 PCI Express 評価ボード

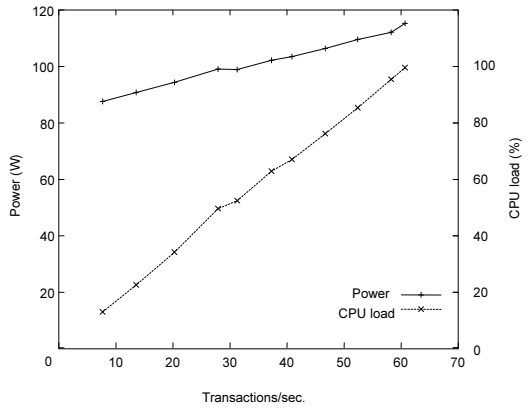

図５　サーバへのリクエスト発生量と消費電力

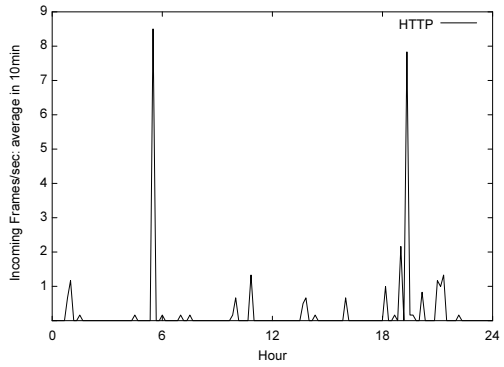

図６　サーバ A における 24 時間の
受信 HTTP フレーム数

行う。ヘッダの解析や，応答パケットの生成などの処理は当初，基本的に専用の回路で実現する予定であるが，将来的にはより複雑なプロトコルなどに対応するために，MicroBlaze などのソフトコアプロセッサを FPGA 上に導入してソフトウェアで処理することも考えられる。

なお，ホストが起動している間の通信については，FPGA 上の L3 制御回路は透過的に動作し，OS とネットワークの間の通信には介入しない。

３．予備評価と考察

3．1　サーバ負荷と消費電力

予備評価の最初のステップとして，ネットワークの負荷状況とサーバの CPU 負荷，消費電力の関係を調べた。計測に用いたサーバの構成は，

- CPU: Intel Celeron D 2.53GHz
- RAM: 512MB DDR-2 SDRAM
- OS: FreeBSD 8.0-RELEASE
- HTTPD: Apache 2.2.11

のようになっている。このサーバでクライアントからの

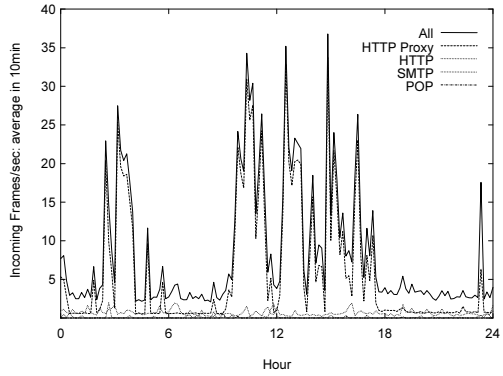

図７　サーバ B における 24 時間の受信フレーム数

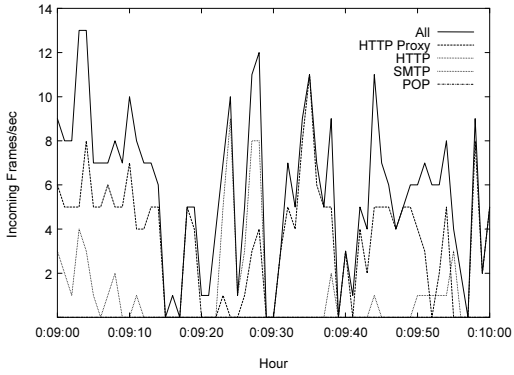

図８　図７の１分間分の拡大図

HTTP のトランザクション発生数を変化させて図 5 を得た。CPU 負荷は時間当たりのトランザクション数にほぼ比例する一方, 消費電力は CPU 負荷が低くなってもピーク時の 2/3 以上である。このサーバはスリープ状態に移行した場合には 5W 程度の消費電力になるので，必要がない場合にはサーバを停止，あるいはスリープさせて消費電力を削減することが効果的であることがわかる。

3．2　スリープ可能な時間の見積もり

ここでは, A, B 2 台のサーバの到着パケットを 24 時間にわたってキャプチャし，その分類を行った。なお，測定は Ethernet のフレーム数単位で行った。

- サーバ A
  - Web (データベースを含む)
- サーバ B
  - Web proxy
  - Web
  - メール (SMTP/POP)
  - DNS

サーバ A は単独のサービスのみを提供するサーバ，サーバ B は学科内で共有されさまざまなサービスを提供しているサーバである。

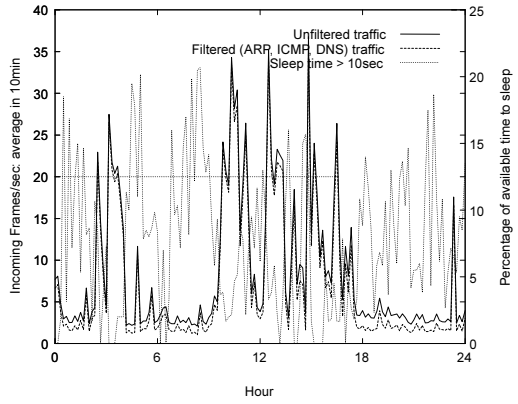

図９　フィルタを適用したトラヒック量と
スリープ可能な時間の割合

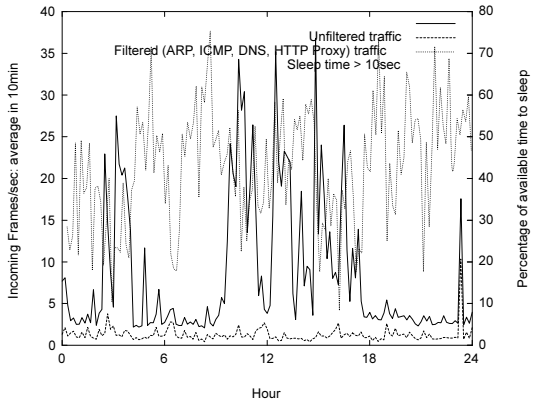

図 10　図９をさらにフィルタした場合

3. 2. 1　サーバ A の場合

サーバ A は，ある公開データベースを HTTP 経由で提供するために稼働しており，アクセス数はそれほど多くない。その 24 時間の到着フレーム数のうち，HTTP のものは図 6 のようであった。このグラフは 10 分ごとのフレーム数を表しており，このうち 10 分間の到着フレーム数が 0 である区間は 85% であった。この結果から，85% の時間はスリープ状態に移行できる可能性があることがわかるが，このサーバの例は非常に負荷が低い例と考えられるので，次節ではより多くのサービスを提供しており負荷の高いサーバ B の測定結果について述べる。

3. 2. 2　サーバ B の場合

サーバ B では複数のサービスが稼働しており，サーバ A に比べるとアクセスが多い。このサーバの 24 時間の受信フレーム数の総量と，そのうち主要なプロトコルの占める量を図 7 に示す。このグラフを拡大して，ある 1 分間の分だけを示したのが図 8 である。図 8 から，フレームを 1 秒間まったく受信しないことが何度かあり，それ以外にもいくつかのプロトコルに関わる通信だけが行われている時間があることがわかる。後者については，それらのプロトコルを NIC や代理サーバで処理することができれば，サーバはスリープ状態に移行してもよい。

そこで，NIC や代理サーバによる処理を適用した際の効果を調べるために，NIC 上の回路での応答が容易と考えられる ARP, ICMP, DNS query といったプロトコルのトラヒックを図 7 から除去したものが図 9 である。この図にはまた，フィルタ適用後に 10 秒以上受信がない場合はスリープ状態に移行可能であるとし，その時間の占める割合を示しており，比較的トラヒック量の少ない時間帯であれば単純な代理処理であっても 5ー10%の時間はスリープ状態に移行して消費電力を落とすことが可能であると考えられる。

表１　NIC の各モジュールの回路規模

	Slice	BlockRAM
FPGA 全体	7,200	60
DMAC	1,990	11
Routing table	205	0
ARP	484	10
Ethernet Tx	115	1
Ethernet Rx	159	1
Ethernet MAC	388	5
合計	3,341	28

さらにこの考えを推し進めて，ARP, ICMP, DNS query に加えて Web proxy に関連するトラヒックをフィルタしたものを図 10 に示す。Web proxy のキャッシュは，トラヒックの多い時間帯にはネットワークの混雑を緩和するために有効な役割を果たすが，トラヒックが少なくなるにつれてその重要性は薄くなる。したがって，トラヒックが少ない場合には Web proxy のキャッシュサーバを停止して，Web proxy へのリクエストは NIC や代理サーバが処理することで，キャッシュの有効性よりも消費電力の削減を優先することは妥当な選択であると考えられる。Web proxy を除いた場合には，図 10 から，10 秒以上トラヒックのない時間の割合は 30ー50%程度になるので，早朝・夜間は NIC や代理サーバによる代理応答の範囲を拡大することで，サービス品質を落とさずにさらに消費電力を削減できることがわかる。

3. 3　NIC の実装状況

図 4 に示した FPGA 評価ボードは Xilinx 社の Virtex-5 シリーズ FPGA である，XC5VLX50T-1FF1136 を搭載しており，これに NIC の各種機能を実装中である。

FPGA 上の回路はすべて 125MHz で動作し，

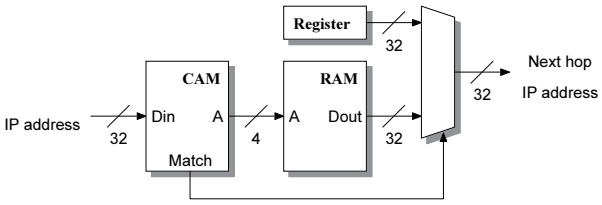

図 11　ルーティングテーブルの構成

PCI-Express 側は 64bit 幅 (最大 8Gbps), Ethernet 側は 8bit 幅 (最大 1Gbps)の，それぞれ送受信別の信号線を使ってデータの送受信を行う。現在までに，

- PCI-Express を介してホスト PC とのデータを送受信する DMA controller (DMAC)
- IP パケットの送受信を行うためのルーティングテーブルおよび ARP テーブル管理機構
- Ethernet パケットの生成・分解と送受信を行う機構
- Ethernet MAC

の実装がほぼ完了しており，それぞれの回路規模は表 1 に示す通りになっている。現状では FPGA の持つ論理ブロック (Slice) とメモリブロック (BlockRAM)をそれぞれ半分弱使用しており，上位層のプロトコルを実装するために充分な回路リソースの余裕が残っている。

以下で, L3 スタックのうち実装が完了しているルーティングテーブルと，ARP テーブル管理機構の構成について説明する。IP パケットを送信するにはルーティングテーブルを参照して次ホップの IP アドレスを得た後に，ARP テーブルで宛先の Ethernet MAC アドレスを参照し，そこへと送られるため，これらふたつの機構は L3 スタックの実装においてきわめて重要である。

3. 3. 1　ルーティングテーブル管理機構

図 11 にルーティングテーブルの構成を示す。ルーティングテーブルは，ルータのアドレスを格納する通常のメモリと, 宛先 IP アドレスからこのメモリを参照するポインタを得るための連想メモリ(CAM)から成る。CAM の出力と RAM のアドレスを接続しておくことで, CAM に宛先 IP アドレスを入力することで，RAM の出力から次ホップのルータのアドレスを得ることができ，入力から出力までの遅延は CAM, RAM ともに 1 クロックサイクルで，合計 2 クロックサイクルである。

CAM と RAM はそれぞれ 16 エントリであり, CAM は 0, 1 の値の他に don't care を保持することのできる三値連想メモリを用いてサブネットマスクの取り扱いを実現している。また，デフォルトルートの取り扱いは専用のレジスタにルータのアドレスを格納しておき，CAM にマッチしなかった場合にこれを次ホップのアドレスとして出力することで実現している。また，図には示されていないが, 宛先として自サブネット内の IP アドレスを与えた場合には，次ホップのアドレスとして宛先アドレスがそのまま出力される。

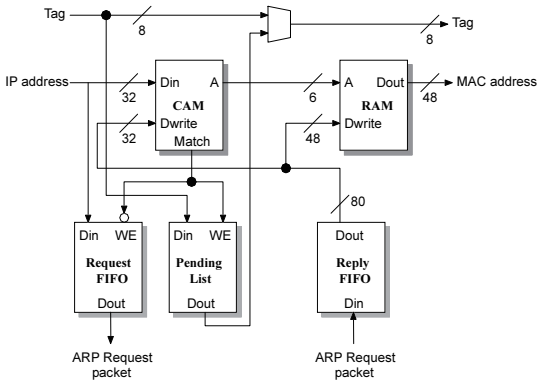

図 12　ARP テーブル管理機構の構成

ルーティングテーブルでは，少ないエントリ数と，三値連想メモリの必要性を考慮して，FPGA に搭載されているメモリ機能専用ブロックである BlockRAM は用いず, 通常のロジックに用いる LUT をメモリとして使用している。

なお，ルーティングテーブルの初期化はホストからドライバを介して行う。

3. 3. 2　ARP テーブル管理機構

ARP テーブルはルーティングテーブルと異なり動的に更新されるため, 図 12 のように複雑な構成をとっている。CAM と RAM を組み合わせて宛先のアドレスを得る点はルーティングテーブルと同様であるが，該当するエントリが CAM に存在しない場合は ARP の query packet を発行し，それに対する応答を待つ点が大きく異なる。ここで, ARP の query を行う場合に, 応答があるまで ARP テーブルがブロックされてしまうと性能の低下を引き起こす可能性があるため，応答パケットを待っている間にも処理を続けられるような設計になっている。

したがって，入出力の順序が保証されないため，ARP テーブルを参照する場合には宛先 IP アドレスと共に 8bit のタグを入力し，出力される MAC アドレスには参照時に入力されたタグを付与することで入出力間の対応をとっている。

ARP テーブルは CAM, RAM ともに 64 エントリで，BlockRAM を用いている。CAM は動作周波数を維持するためにパイプライン化されており，2 クロックサイクルの遅延を伴うため，CAM に該当するアドレスがあった場合には 3 クロックサイクルで宛先 MAC アドレスを得ることができる。

該当するエントリが CAM に存在しない場合は，図 12

の“Pending list”で示される部分に存在するメモリにタグが記録され, ARP query パケットが送信される。ARP query に対する応答パケットが届いた際には, CAM, RAM に必要な情報が保存されるとともに，Pending list に記録されているタグが MAC アドレスとともに出力される。また，Pending list は最大 16 個の IP アドレスに対する未解決の ARP リクエストを保持することができる。ARP のアドレス解決のためのパケットは，Pending list の処理中でも取りこぼしが発生しないように, FIFO を用いて送受信される。

４．まとめ

本稿では，分散制御による細粒度な電力管理を行うための代理応答能力を持つ NIC の実装と，その適用効果に関する予備評価の結果を示した。

本学内で稼働中のサーバのトラヒック情報から適用効果を見積もった結果，単一のサービスを提供する平均負荷の低いサーバでは 85% 程度，また，複数のサービスを提供する平均負荷の高いサーバの場合には通常時で 5〜10%，トラヒックの少ない時間帯においては代理応答の範囲をさらに拡大することで30〜50%程度の時間はサーバをスリープ状態に移行できる見込みであることを示した。後者のサーバは本学内では比較的稼働率の高い部類に入ると考えられ，他のサーバでは同様の手法を用いることでさらに長時間スリープ状態で運用することが可能になると考えられる。

NIC は FPGA を用いて実装され, L3 スタックの基礎となる部分は開発済みである。今後，いくつかのプロトコルを処理できる専用の回路を組み込んでも最大 2〜3W 程度の消費電力で動作する見込みである。

今後は提案したNICのハードウェアおよび関連する基本的なソフトウェアの開発を進め，提案方式の有効性を検証するとともに，さらに消費電力を削減するための方式について検討していく予定である。

参考文献

1) Bruce Nordman and Kenneth Christensen. Proxying: the next step in reducing IT energy use. IEEE Computer, Vol. 43, No. 1, pp. 91–93, Jan. 2010.
2) Kenneth J. Christensen, Ghamara Gunaratne, Bruce Nordman, and Alan D. George. The next frontier for communications networks: power management. Computer Communications, Vol. 27, pp. 1758–1770, Aug. 2004.
3) 長名保範, 栗林伸一. 代理応答・処理の範囲拡大によるサーバ電力使用量削減方式の提案. ICM2009-45, 電子情報通信学会, Mar. 2010.
4) 栗林伸一, 長名保範. エンドシステム, ネットワーク, 電力網の連携による ICT 総体の電力使用量ならびに発電量の削減法. ICM2009-47, 電子情報通信学会, Mar. 2010.
5) Xilinx Inc. Virtex-5 FPGA User Guide, v5.2, Nov. 2009.
6) Xilinx Inc. Spartan-6 FPGA User Guide, v1.1, Dec. 2009.
7) Altera Inc. Cyclone IV Device Handbook, v1.1, Jan. 2010.