

複数 NIC を連携制御する既存手法の問題点 Problem of Existing Methods Coordinating Multiple NICs

吉田泰三[†]
Taizo Yoshida

谷口秀夫[†]
Hideo Taniguchi

1 はじめに

高速な通信を可能にする通信路とネットワークインタフェースカード (NIC) が次々に登場している。しかし、最新の高速な機器は高価である。そこで、廉価な低速の機器を複数利用し、高速な通信を実現する手法が提案されている。提案手法は、複数 NIC を仮想的に一つに見せる手法、および NIC と通信パスを固定対応させる手法に分類できる。

本稿では、これらの既存手法について送信処理に着目し、その問題点を明らかにする。

2 複数 NIC を仮想的に一つに見せる手法

2.1 方式

複数 NIC を仮想的に一つの NIC に見せる手法について、方式や特徴などを表 1 に示す。複数 NIC を仮想的に一つに見せる手法は、ラウンドロビン分散と負荷分散に分類できる。

ラウンドロビン分散方式は、ラウンドロビンで送信パケットを各 NIC に振り分ける。この方式の特徴は、制御オーバーヘッドが少ないことである。例として、FreeBSD の Lagg[1] 機能の Round-robin モード、Linux の Ethernet チャンネルボンディング [2] の balance-rr モードがある。

負荷分散方式は、各 NIC の送信負荷情報をもとにパケット送信する NIC を決定する。この方式の特徴は、送信負荷情報に基づく NIC 負荷の均一化ができることである。しかし、制御オーバーヘッドが大きい欠点がある。例として、Linux の Ethernet チャンネルボンディングの balance-tlb モードや balance-alb モードがある。

2.2 考察と問題点

ラウンドロビン分散方式の問題は、受信側ではパケットの送信順序と到着順序が一致しないことである。この問題は、送信側から受信側までの通信経路が異なることで発生する。パケットを送信すると、通信経路が同一の場合、パケットは入れ替わることなく受信側に到着するため、送信順序と到着順序は一致する。一方、通信経路が異なる場合、パケットが速やかに送信される通信経路とそうでない通信経路が存在することにより、送信順序と到着順序が一致しないことがある。この場合、再送が発生し、通信効率は低下する。

負荷分散方式では、送信負荷情報の更新周期の長短により、異なる問題が生じる。この方式では、送信負荷情報が更新されるまでは、更新前の送信負荷情報にもとづいて、パケットを送信する NIC を決定する。このため、送信負荷情報の更新周期が長い場合、各 NIC から送信されるデータ量を逐次反映することができない。

したがって、NIC 負荷の偏りが生じやすい。一方、送信負荷情報の更新周期が短い場合、送信負荷情報の更新が頻繁に行なわれ、ラウンドロビン分散方式と同様に受信側ではパケットの送信順序と到着順序が一致しないことが発生しやすくなる。例えば、Linux の balance-tlb モードの送信負荷情報の更新周期は、10 秒である。このため、多数パケットを送信する場合は、更新周期が長いと判断でき、少数パケットを送信する場合は、短いと判断できる。

3 NIC と通信パスを固定対応させる手法

3.1 方式

NIC と通信パスを固定対応させる手法について、方式や特徴などを表 2 に示す。

通信パスを固定した分散方式は、特定情報をもとにパケットを送信する NIC を決定する。特定情報は、MAC アドレス、IP アドレス、およびポート番号の三つである。この方式の特徴は、受信側において、パケットの送信順序と到着順序が一致することである。例として、FreeBSD の Lagg 機能の Cisco Fast EtherChannel (FEC) モード、IEEE 802.3ad Link Aggregation Control Protocol (LACP) モード、および Loadbalance モード、Linux の Ethernet チャンネルボンディングの balance-xor モードや 802.3ad モードがある。

3.2 考察と問題点

NIC と通信パスを固定対応させる手法の問題は、使用する特定情報により、問題が異なる。

MAC アドレスを使用する場合、相手計算機単位でパケットを送信する NIC を決定する。例えば、balance-xor モードは、送信側の各 NIC の MAC アドレスを同一に設定する。このため、パケット送信に使用する NIC は、受信側の MAC アドレスに依存して決定する。計算機の MAC アドレスは一意なものであるため、NIC を計算機単位で分散して使用する。したがって、分散の単位は、計算機単位であり、大きいといえる。一方、ルータを介した通信では、ルータの MAC アドレスが使用される。このため、特定の NIC を使用することになり、負荷分散はできない。

IP アドレスを使用する場合、ルータを介するか否かにかかわらず、相手計算機単位でパケットを送信する NIC を決定する。したがって、分散の単位は、計算機単位であり大きい。

ポート番号を使用する場合、コネクション単位でパケットを送信する NIC を決定する。例えば、Linux は、受信側とコネクションを確立するときに、送信側のポート番号を OS が動的に決定する。したがって、分散の単位は、コネクション単位であり、小さい。

しかし、分散単位に関係なく、いずれの場合も次の情報を考慮していないため、新たな問題も抱えている。

[†] 岡山大学大学院自然科学研究科, Graduate School of Natural Science and Technology, Okayama University

表1 複数NICを仮想的に一つに見せる手法

方式	概要	特徴	問題点	例
ラウンドロビン分散	ラウンドロビンで送信パケットを各NICへ振り分ける。	制御オーバーヘッドが少ない。	受信側では、パケットの送信順序と到着順序が一致しないことが発生する。	(1)FreeBSDのRound-robinモード (2)Linuxのbalance-rr
負荷分散	各NICの送信負荷情報をもとにパケットを送信するNICを決定する。	(1)送信負荷情報にもとづくため、NIC負荷が均一化できる。 (2)制御オーバーヘッドが大きい。	(1)送信負荷情報の更新周期が長い場合、NIC負荷の偏りが生じやすい。 (2)送信負荷情報の更新周期が短い場合、NIC負荷の偏りは少なくなるが、受信側ではパケットの送信順序と到着順序が一致しないことが発生する。	Linuxのbalance-tlbモードやbalance-albモード

表2 NICと通信パスを固定対応させる手法

方式	概要	特徴	問題点	例
通信パスを固定した分散	通信パスの特定情報をもとにパケットを送信するNICを決定する。 ・MACアドレス ・IPアドレス ・ポート番号	受信側において、パケットの送信順序と到着順序が一致する。	送信データ量、コネクション接続時間、およびNIC性能を考慮できない。	(1)FreeBSDのFECモード、LACPモード、およびLoad-balance (2)Linuxのbalance-xorモードや802.3adモード

- 送信データ量
- コネクション接続時間
- NIC性能

分散単位の送信データ量が異なる場合、NICあたりの分散単位数や接続時間が同様であっても、各NICの負荷は異なる。例えば、各NICに同数のソケットが割り当てられており、各ソケットの接続時間は同じとする。このとき、各ソケットを流れる送信データ量が異なると、送信データ量の多いソケットが使用しているNICは、送信負荷が大きくなってしまう。このため、新たなソケットによるデータの送信要求を単位時間あたりに転送するデータ量が最も多いNICで処理する場合、単位時間あたりに転送するデータ量が最も少ないNICで処理する場合と比較して、応答時間は長大化する。

分散単位の接続時間が異なる場合、NICあたりの分散単位数や送信データ量が同様であっても、各NICの負荷は異なる。例えば、各NICに同数のソケットが割り当てられており、各ソケットを流れる送信データ量は同じとする。このとき、各ソケットの接続時間が異なると、接続時間の長いソケットが使用しているNICでは、通信処理が共存する可能性が高い。このため、新たなソケットによるデータの送信要求をコネクション接続時間の長いソケットが使用しているNICで処理する場合、コネクション接続時間の短いソケットが使用しているNICで処理する場合と比較して、各通信処理の応答時間は長大化する。

NIC性能を考慮しない場合、NICあたりの分散単位数、分散単位の送信データ量、および接続時間が同様であっても、各NICの使用率は異なる。ここで、使用率とは、NICの単位時間あたりに転送可能なデータ量の最大値に対する単位時間あたりに転送するデータ量の割合である。計算機に搭載されているNICは、必ずしも同一の通信性能とは限らず、それぞれの通信性能を考慮しなければ、各NICを効率的に使用できない。

例えば、各NICに同数のソケットが割り当てられており、各ソケットを流れる送信データ量と接続時間は同じとする。このとき、NIC性能に違いがあると、単位時間あたりに処理できるデータ量はNIC性能により異なる。このため、新たなソケットによるデータの送信要求を使用率が大きいNICでデータを送信する場合、使用率が小さいNICでデータを送信する場合と比較して、応答時間は長大化する。

4 おわりに

複数NICを連携制御する既存手法の問題点について述べた。複数NICを連携制御する方式は、複数NICを仮想的に一つに見せる手法、およびNICと通信パスを固定対応する手法に分類できる。

複数NICを仮想的に一つに見せる手法は、送信負荷情報の更新周期の長短によってNIC負荷の偏りが発生する。また、受信側パケットの送信順序と到着順序が一致しないことが発生し得る。NICと通信パスを固定対応させる手法は、送信データ量、コネクション接続時間、およびNIC性能を考慮できない。このため、通信処理が共存することにより、各通信処理の応答時間が大きく異なることが起こる。

残された課題として、送信データ量、コネクション接続時間、およびNIC性能を考慮し、複数NICを効率的に制御する方式の検討がある。

謝辞 本研究の一部は、科学研究費補助金基盤研究（B）（課題番号：24300008）による。

参考文献

- [1] “The FreeBSD Project,” <http://www.freebsd.org/>
- [2] Thomas Davis, “Linux Channel Bonding,” <http://sourceforge.net/projects/bonding/>,