# 1D－6　ワークステーションにおける高速プロトコル処理 (0–copy architecture) の実装と性能評価

北村浩　西田竹志
日本電気株式会社 C&C研究所
e-mail: kitamura@nwk.cl.nec.co.jp

## 1 はじめに

近年物理スピードが100Mbps以上の通信速度を持つ高速LANが実現してきている。また、RISC技術などの進歩により、CPUの処理速度は飛躍的に向上してきている。しかし、ホストの通信性能は伸びてきておらず、End-to-Endでの通信では高速LANの通信速度の半分にも満たないスループットしか出せない。

これに対し、旧来のデータとは質量共に異なる音声や画像などの時間的制約の厳しいバルクデータを扱うマルチメディア通信アプリケーションが増加しており、実際のスループットとして高い通信性能を持つワークステーション(WS)の出現が急務となってきている。

通信性能が上がらない原因は、通信処理の中で主記憶装置上でのデータのコピー処理の部分にある。これらの処理に要する時間は通信処理全体の半分以上の時間を占めている[1]。主記憶装置を構成するメモリの応答速度は数10nsec程度でCPUの速度に比べて極度に遅い上、CPUとメモリの速度差を吸収するキャッシュ機構も、大量で常に新しい通信データに対しては効果がほとんどないからである。海外でも同様の指摘があり、コピーを回避する方法が模索されている[2]。

本稿では、高速通信処理の障害となる問題を効果的に解決するために、WS内部で通信データのコピー処理を行なわない構造（**0-copy architecture**）を提案する。また、SVR4のUNIX WSでの実装について説明する。本方式は市販WSのTCPレベルで**220Mbps**の高いスループットを実現し、高速性を実証している。

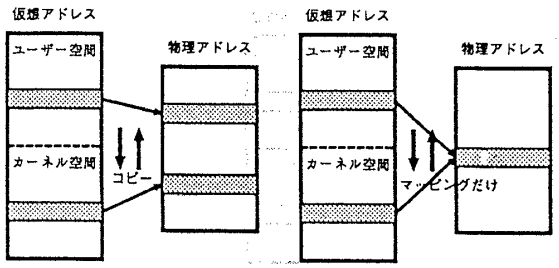
## 2 0-copy architectureの条件と実装方法

UNIXのユーザプロセスは、それぞれ独自のアドレス空間上で動作し、特権を持つカーネルプロセスもまた独自のアドレス空間を持つ構造になっている。通信を行なうアプリケーションでの処理は、ユーザ空間とカーネル空間の間を行き来し、現在の実装ではこれらの間にデータのコピー処理を行なっている。

現在のarchitectureでコピー処理が必要であったのには、二つの理由がある。**1.** データ管理上 カーネル空間とユーザ空間の間でデータ資源の管理を明確にすることと、独立した構造で管理を単純化するため。**2.** データの物理的メモリ配置 パケット化した離散データを物理的連続メモリ領域に配置するためや、オーディオデバイスのような特殊なメモリにデータを対象にするため。

データ管理上の要件を満たすには必ずしもコピーする必要はないが、現状ではコピー処理を行なっている。**0-copy architecture**はコピー処理を行なわずにデータ管理上の要件を満たし高速通信処理を実現している。

0-copy architectureの本質は、送信の場合は、アプリケーションが扱うメモリ領域が、直接ネットワークインタフェースドライバまで伝わり、受信の場合には、ドライバがデータを受け取ったメモリ領域が、直接アプリケーションプログラムからアクセスできる構造である。その実現法として、図1.bに示すようにUNIX OSで用いている仮想アドレスを応用して、メモリ空間を共有している。物理メモリ上で、現状の構造ではコピー処理を行なっているが、提案する構造(0-copy architecture)ではマッピングを行うだけでコピー処理は行なわない。




a: 現状の構造　b: 提案する構造

図1: メモリアドレス構造とコピー処理

0-copy architectureを実装したカーネルにはユーザプログラムが選択可能な二つの通信モードが存在する。

---

Implementation and Perfomance Evaluation of High Speed Protocol Processing (0–copy architecture) in Workstaion
Hiroshi KITAMURA　Takeshi NISHIDA
C&C Research Laboratories, NEC Corporation

## 3 0-copy architecture 実装方法の要素

- カーネル、ユーザ間でメモリ空間を共有する方式は、確実で性能の出る一旦カーネル内で確保した領域をユーザ空間にマップする部分共有方式を用いる。
- ユーザプロセス内に 0-copy 通信を行なうための特別な領域をデバイスセグメントとして設け、その領域をマップしユーザプログラムの 0-copy 通信処理は全てそこを通して行なう。このことによりメモリ管理の制御はデバイスの制御の形で行なえる。
- 0-copy 通信で扱うデータは、その形式をデータ位置のポインタ集合である入出力ベクタ型へ変換してユーザ、カーネル空間の間を行き来する。
- データの形式の変換とマッピング制御の機能は、一般性がありユーザが選択的にモジュールの構成を変えることの出来るストリームモジュールとして実装し、カーネル内での処理を行なう。
- 0-copy 通信の入出力は、新しいセマンティクスを導入するために新しいシステムコールを設ける。

## 4 性能評価結果

必要最低限のデーモンのみを動かし、送信装置と受信装置のみの閉じたネットワーク上で再現性の高い環境で、3Gbyte のデータを TCP/IP を用いたソケット間通信プログラムで通信して評価を行なった。

物理ネットワークとして **Ethernet** と **FDDI** を用いた通信が考えられるが、**Ethernet** の場合は物理帯域が、**FDDI** の場合は I/F ボードのバス性能が通信性能を決定しており、0-copy architecture の特性を示す媒体として用いることが出来ない。そこで、物理的制約が比較的少ないループバック **I/F** を用いる。

ループバック **I/F** を使った通信は IP モジュールの下に送信したデータが受信するデータになるように折り返すモジュールを通しての通信で、同一装置内で送信と受信両方のプロセスが同時に動くために、資源の競合などが起こるという制約が発生する。

**測定値**　標準出荷 OS では **25Mbps** 程のスループットである。そこで、0-copy architecture に適するように、モジュールの water mark 値を大きくするチューニングを行なった後の性能評価結果を表 1 に示す。

0-copy での read はカーネル処理をモジュールとして実装しているためフロー制御が難しいため、それほど性能が出ないが、0-copy での write は予想通りの成果を示した。0-copy architecture を用いることによって、通常の通信より約 **33%** の性能が向上する。今後更に CPU 能力が向上し、メモリやバス速度のと差が大きくなれば、もっと顕著な成果が現れると予想される。

| 送信側 | 受信側 | 通信速度 |
|---|---|---|
| 通常の write | 通常の read | 45 Mbps |
| 0-copy での write | 通常の read | **60** Mbps |
| 通常の write | 0-copy での read | 39 Mbps |
| 0-copy での write | 0-copy での read | **54** Mbps |

表 1: ループバック I/F での性能

**TCP 層での checksum 処理 off の場合**　ユーザ、カーネル空間の間でコピー処理の次に問題となるのは、コピー処理の半分に相当するメモリから CPU へのデータ移動処理である TCP 層での checksum 処理である。

そこで、この処理を off にした状態で 0-copy 通信を行なった。結果は、**最大値で 110 Mbps** 通信速度を記録した。同一の装置で送受信処理を行なっているので、一般的な考えに従えば、送受信どちらか一方だけの処理では、二倍の **220Mbps** の通信速度を出せることが原理的に可能であることが、実験的に証明された。

## 5 まとめ

WS 内部で通信データのコピー処理を行なわない構造 **0-copy architecture** の実装方法の概略と、この構造を用いた通信での性能評価について報告した。

予想通りの高い性能を示しており、通常の 33% の性能向上をし、checksum 処理 off で原理的に 220Mbps の通信速度が出ることが示せ、0-copy architecture の有用性が実験的にも証明された。また、マルチメディア通信では必要不可欠な技術であることが明確になった。

今後は、さらに安定的に 0-copy architecture が動作するよう改良を加えるとともに、0-copy 通信実現の後の問題である TCP 層での checksum 処理の高速化についても検討していきたい。

## 参考文献

[1] 北村浩、前原一之 “ワークステーションにおける高速プロトコル処理を目指した性能評価” 信学会技報 SSE92-38 Sep. 1992

[2] D. Banks, M. Prudence, “A High Performance Network Architecture for a PA-RISC Workstation,” *IEEE Journal on Selected Areas in Communications*, February 1993, Volume 10, No.1 pp 191-202.