HiTactix/Symbioseの開発（1）

2Z－1　－概要－

岩嵜正明，中原雅彦，竹内　理，中野隆裕，田口しほ子，川田容子，児玉昇司
(株)日立製作所　システム開発研究所

1．背景

我々は，高品質連続メディア通信機能を提供可能なOS技術の実現を目的に，1994年よりHiTactixカーネルの設計・開発に着手し，これまでにアイソクロナス・スケジューリング技術やQoS保証通信技術（TTCP/ITM）等を完成した[1]．前者は，CPU負荷が90%を超える高負荷状態においても，周期スレッドのスケジューリング遅延を100マイクロ秒程度以下に保つことができる[2,3]．また，後者は，共有メディアである100Mbps Ethernet上で，40ストリーム以上のMPEG1データ（FBR,1.5Mbps）を，バックグラウンド・トラフィックが混在する環境下においても，極めて高品質に転送することができる[4,5]．

連続メディア処理向けに外部インタフェース仕様から新規設計したHiTactixは，このように優れたQoS保証性能を実現できるが，その反面，既存OSとの互換性を欠くという問題を抱えている．本稿では，この互換性問題の解決を始めとして，HiTactixを実用に耐える水準に押上げるべく推進中の技術開発の概要について報告する．

2．互換性問題解決へのアプローチ

2.1. 従来アプローチ

既存OSとの互換性を実現する手法としては，Machに代表されるマイクロカーネル方式が知られている．しかしながら，この方式はIPCオーバヘッド等の要因による性能低下が無視できない．また，IPCによるコンテクスト切り替えは，スレッド間でのスケジューリング属性の継承を困難にし，アプリケーションのリアルタイム実行の障害となる．

また，プロセッサ上で複数OSを時分割稼動させるVM（仮想マシン）方式やDARMA方式[6]もある．しかし，これらの方式にも，性能的なペナルティに加え，ネットワーク・デバイスの共有が困難であるといった問題がある．

2.2. ヘテロOS共存アプローチ

我々は，メモリ共有型マルチプロセッサを利用し，その各CPU上でひとつのOSを稼動させるヘテロOS共存方式により，互換性問題の解決を図っている．この方式では，例えば，デュアルCPUのPentium IIマシンを用い，一方のCPUでHiTactixを，他方のCPUでBSD/OSを稼動させる．両OSは，共有バッファやOS間通信機能を提供するソフトウェアSymbioseによって結合する．

各OSがCPUを占有する本方式は，マイクロカーネル方式やVM方式の様な性能上のペナルティを生じない．また，SMP（対称型マルチプロセッシング）方式のOSと比較すると，CPU間で共有する資源が少なく，ロック競合オーバヘッドやキャッシュ無効化ペナルティを低減でき，近年の高性能CPUアーキテクチャとの整合性に優れる．

本方式は，複数のOS間で機能分散処理が可能なアプリケーションに適している．例えば，我々は，本方式を応用した通信ゲートウェイ（Dynamic Gateway＝QoS保証ルータ＋負荷分散ルータ＋高機能ファイアウォール）を開発している．このゲートウェイでは，高性能を要求される下位層（ドライバやIP層）はHiTactixが処理し，上位層はBSD/OS上のアプリケーションが処理する．

我々は，高性能な通信ゲートウェイの実現に向け，性能を犠牲にせずに両OSのネットワーク通信機能が連動できる様にHiTactixのIP層を再設計した．同時に，BSD/OS側からはHiTactix側を仮想ドライバを経由してアクセスするEthernetデバイスと見なし，BSD/OSカーネル内部の改造を最小限に止めた．また，HiTactix内部での送受信処理や両OS間でのバッファ受け渡し処理を一括化することで，スループット性能の低下を防いでいる．

3．ファイルシステムのQoS保証

HiTactixは，当初より連続メディア応用向けに最適化したMedia File System（MFS）を備えている．このMFSは，ラージ・ブロックサイズの非同期入出力機能や，ゼロ・コピー入出力が可能なDirect Buffer Mapping機能を備えている[7]．しかしながら，TTCP/ITM等のリアルタイム通信技術の完成により，ネットワークへ送信可能なストリーム数が増加した結果，HiTactixを使用したビデオサーバでは，ディスク読み出し時のシーク待ちや回転待ちオーバヘッドがシステム全体の高性能化やQoS保証の障害となってきた．

3.1. Striped Media File System

上記の問題を解決するために，今回，Striped Media File System（SMFS）を開発した．SMFSは複数のディスクドライブにファイルをストライプし，並列読み出しを可能にする．さらに，このSMFSは読み出し遅延時間を保証する入出力スケジューリング機構を備えている．この入出力スケジューリング機構により，予測不能な入出力負荷が混在する状況下においても，数十ストリーム以上のメディアデータを遅延なく読み出すことが可能となった．

3.2. VFSサポート

HiTactixは上述のMFSやSMFS以外に，プログラムコード等を格納するための階層ファイルシステム（Local File System）やread-onlyのBSD File Systemをサポートしている．

A Development of HiTactix/Symbiose(1) - Overview -
Masaaki Iwasaki, Masahiko Nakahara, Tadashi Takeuchi, Takahiro Nakano, Shihoko Taguchi, Yoko Kawata, Shoji Kodama
Systems Development Laboratory, Hitachi, Ltd.

また，BSD/OSとの連動によりマルチウィンドウ環境でのデバッグ等が可能となり，仮想端末機能も必要となってきた．このため，全ての入出力を統一したインタフェースで利用できるVFSの実装を検討している．

但し，MFSやSMFSは，非同期入出力機能やDirect Buffer Mapping機能，さらにはQoSパラメータ指定機能を提供している．現在，設計中のHiTactix用VFSインタフェースには，アプリケーション開発者がこれらの機能を利用できる拡張を施す予定である．

4．リアルタイム通信機能の拡張

HiTactixのリアルタイム通信技術（TTCP/ITM, RTIPSIG）は，共有型Ethernetのハードウェア仕様やTCP, UDP等のプロトコルスタックとの高い互換性を備えている．現在，その実用化に向けて下記の技術を開発している．

4.1. 他OSへのTTCP/ITM実装

TTCP/ITM方式のQoS保証効果を確実にするには，数ミリ秒周期毎のデータ送信量を予約帯域以下に抑制するTTCP/ITMモジュールを，共有セグメント上の全ノードに実装しなければならない．この要件を満足するため，Windows NTとWindows 2000用の中間ドライバとしてTTCP/ITMモジュールを開発している．

TTCP/ITMモジュールは，帯域マネージャ（TTCPサーバ）と交信して帯域の予約・解放を行うTTCPクライアントと，一周期毎のデータ送信量を抑制するITMドライバから構成する．TTCP/ITMモジュールは，IP層とデバイスドライバ層の間に挿入可能な設計としており，その実装には，OSがタイマ割り込みハンドラ等のインタフェースを提供し，デバイスドライバが複数パケットの送信要求を一括受理できるインタフェースを提供する必要がある．Windowsの中間ドライバ・インタフェースを始め，高速ネットワークに対応する最近のOSの仕様は，この条件を満足しており，容易にTTCP/ITMモジュールを組み込むことができる．

4.2. フェイルセーフ機構の実装

コネクション毎に帯域予約を行うリアルタイム通信システムにおいては，経路上のルータや帯域マネージャに障害が発生した場合，この障害発生を検出し，経路全体に渡って予約中の資源を自動解放するフェールセーフ機構が必要である．但し，この機構を集中管理方式で実現しようとすると，例えば，障害箇所を迂回した制御パケットの転送等が必要となり，実装が困難となる．

我々は，各コネクション毎に，経路上のルータ（帯域マネージャを兼ねる）が隣接ルータとハートビート・パケットを交換し，このパケットが一定時間に渡って送られて来なければ異常と判定し，該当コネクションの予約資源を解放する機構を実装している．

4.3. CPUクロック誤差の補正

数十分を超えるビデオデータ転送においては，サーバとクライアントのクロック誤差，即ち，水晶発振子の周波数誤差が累積し，バッファのオーバフローやアンダーフローが発生する．この問題を解決するため，クライアントのバッファ残量を監視し，サーバの送信レートを微調整し，サーバとクライアントのクロック誤差を補正する高精度レート整合技術を開発している．

5．今後の展開

以上，簡単にではあるが，現在推進中のHiTactixオペレーティングシステムの研究開発の概要について述べた．ここに述べた各機能は，既に開発済み，あるいは，遅くとも今年度中に開発を完了する予定である．

これらの機能の統合を完了後，実用化に向けて，幅広く共同研究開発パートナーを募ることも検討したいと考えている．

謝辞

本研究の一部は，IPAの次世代デジタル応用基盤技術開発事業「連続デジタルメディア処理向きアイソクロナス・カーネルの研究開発」として実施している．本研究開発の推進に当たり日頃から御指導・御助言をいただいているIPA技術応用事業部，大阪大学宮原研究室，九州大学谷口研究室，慶応義塾大学大岩研究室の方々に心より感謝申し上げます．

参考文献

[1] M.Iwasaki, et. al, "Isochronous Scheduling and its Application to Traffic Control," 19th IEEE RTSS'98, Dec.98

[2] 竹内他，「連続メディア処理向きOSの周期駆動保証機構の設計と実装」，情報処理学会論文誌，第40巻，第3号，99.3

[3] 中原他，「連続メディア処理向けマイクロカーネルにおける内部排他制御方式」，情報処理学会論文誌，第40巻，第6号，99.6

[4] 中野他，「Ethernet上でQoSを保証する通信方法の設計と実装」，情報処理学会，コンピュータシステム・シンポジウム論文集，pp.35～42, 97.11

[5] 竹内他，「アイソクロナススケジューラを応用したQoS保証型ルーティング方式の設計と実装」，情報処理学会，マルチメディア通信と分散処理ワークショップ論文集，98.11

[6] 新井利明，宮崎義弘，「異種OS共存技術『DARMA』の開発と制御システムへの適用」，日本工業出版，計測技術，第27巻，第7号，99.6

[7] 中野他，「連続メディア処理向きマイクロカーネルの開発(4)-入出力方式の設計と評価」，情報処理学会，第53回全国大会（平成8年後期）講演論文集(1), pp.147～148, 96.9

注1: Pentium IIはIntel Corporationの登録商標です．
注2: BSD/OSはBSDI社の商標です．
注3: Windows NT, Windows 2000はMicrosoft社の商標です．