情報処理学会論文誌 Vol.54 No.12 2413–2426 (Dec. 2013)

推薦論文

オーバ50 Gbit/s PCクラスタ型ストリームサーバの構成法

君山 博之[1,a)] 小倉 毅[1,b)] 丸山 充[2,c)]

受付日 2013年4月27日, 採録日 2013年9月13日

概要：近年，映像制作会社間で制作中の映像データを受け渡しする機会が増えている．特に制作技術の進歩により制作工程ごとの分業が進み，各工程の完了後，映像データを次の会社へ配送することが日常的に行われ，その配送にファイルサーバを利用するケースも増えている．しかし，1秒1ギガビットを超える非圧縮ハイビジョンや4K映像データをストレスなく扱うにはこれらのサーバは十分な性能を持っていない．そこで，高精細映像データを複数拠点へ同時配信可能なシステムの実現を目指し，複数のPCを使った超高速サーバ構成法を検討し，PC間を高速な内部ネットワークで接続したクラスタ型映像サーバを提案した．これまで16台のPC，8台の汎用ストレージ，内部結合ネットワークに20 Gbit/sのInfiniBandを採用し，最大24 Gbit/s（非圧縮ハイビジョン16本相当）の配信性能を持つ映像ストリームサーバを実現したが，さらなる性能向上要求に応えるためボトルネックである内部結合ネットワーク上の通信方式を新たに提案，実装を行った．その結果，24台のPC，12台のストレージを用いて世界初となる非圧縮ハイビジョン36本同時配信可能な54 Gbit/sの配信性能を持つストリームサーバを実現した．本論文では，本サーバの構成方法，実装方法および提案する内部結合ネットワーク上の通信方式を説明し，実機を使った性能評価結果について報告する．

キーワード：映像ストリームサーバシステム，クラスタ型サーバシステム，クラスタ内通信方式

Implementation Method for over 50-Gbit/s PC-cluster Based Stream Server System

Hiroyuki Kimiyama[1,a)] Tsuyoshi Ogura[1,b)] Mitsuru Maruyama[2,c)]

Received: April 27, 2013, Accepted: September 13, 2013

Abstract: Recently, video production sites often exchange video data for video-production in which there are many sub-process (ex. 3D modeling, color conversion) since each sub-process requires unique high-level skills. Video production sites use general file servers to exchange their video data via high-speed networks. However, these servers doesn't have enough performance to smoothly handle such high-quality video, such as uncompressed HD and uncompressed 4K videos, which have rates of over 1 gigabit data per second. Therefore, we researched PC-cluster-based video stream server system to handle such a high-quality video streams. We have developed video server system with 24-Gbit/s maximum throughput by combining 16 PCs, 8 general storage systems and 20-Gbit/s InfiniBand inter-cluster network. To respond to demands for delivering video data streams to more video production sites, we propose a new communication method over an inter-cluster network that offers higher total throughput. We implemented the new communication method on a PC-cluster-based server system prototype and evaluated system performance. By combining 24 PCs, 12 storages and one InfiniBand switch, we realize 54-Gbit/s throughput video server system. In this paper, we describe the proposed communication method, hardware and software implementation method on PC cluster, as well as evaluation results.

Keywords: video stream server system, cluster server system, inter-cluster communication method

1. はじめに

近年，放送や映画における映像制作がデジタル化され，制作途中あるいは制作が完了した映像を，映像データとして映像編集者や制作会社，放送局間で受け渡しする機会が増えている．映像制作における CG（Computer Graphics）技術や VFX（Visual Effects）技術，それらの技術を提供するツールは日進月歩で進化している．映像制作技術は高度に専門化され，これらを十分扱える技術者や会社は限られているのが現状である．そのため，映像制作では工程ごとあるいはシーンごとに作業者や制作会社が異なり，制作過程で生成される大容量の映像データが制作者間で頻繁に受け渡しされている．これまで映像データの受け渡しは，ハードディスクやテープなどの物理メディアを介して行われてきた．しかし，映像の高品質化，高精細化ニーズの高まりによって，使用する映像素材も 1 秒あたりのデータ量が 1 ギガビットを超える非圧縮ハイビジョン映像や 10 ギガビットを超える非圧縮 4K 映像に移行していくと考えられる．そのため，制作時間全体に対して物理メディアにコピーし移動する時間の割合が増えることが予想される．また，デジタル技術の向上により制作した CG や VFX の修正が容易になると同時に，修正のつどコピーと移動が発生することから，映像データ受け渡しの効率化がますます望まれてくると考えられる．

この問題を解決するために，ネットワークを使ったファイル転送サービス [1], [2] や，一般的なネットワークファイルサーバ，放送局のように自前で構築した専用ネットワークを使ったファイル共有システムが利用されはじめている [3]．たとえば，日常多くの映像を扱っている国内の放送局は 30 局程度で 1 つのグループ（系列）を作っており（文献 [4], [5]），各系列局は中央に設置したサーバに対して各地で取材した映像素材をアップロードしたり，自局のニュース番組制作のためにそれぞれが必要とする映像素材をサーバから日常業務としてダウンロードしたりしている．放送の場合は，特に，朝や夕方のニュース番組制作のために，アップロードやダウンロードが各系列局からほぼ同時に起こることが想定されることから，少なくとも，これらの系列放送局が同時にサーバとの間で受け渡しできることが必要であると考えられる．現在，放送局では，数 Gbit/s 程度の最大性能を持つサーバを使って，番組制作の日常業務として数十分の 1 に圧縮された映像データを時間をかけて送る運用を行っている．しかし，将来，使用する映像の高品質化や高精細化時代に同じ性能のサーバを使用し続けた場合，ファイル転送にかかる時間が増加していくことは容易に想定される．また，ファイル共有サーバを使ったシステムの場合，共有サーバから編集機や送出用のサーバにいったんダウンロードしないと，その映像を確認することができないことから，誤ってダウンロードした場合のリスクは非常に高くなり，急に映像素材を送りたくなっても，放送時間までに間に合わなくなることも想定される．同時に発せられたリクエストが増えるに従って送信速度が低下することは，締切りのある制作現場において映像制作が放送時間までに間に合わなくなることを意味する．そこで我々は，将来の映像の高品質化や高精細化に適用可能な高性能でかつリアルタイム配信機能を持つサーバシステムの実現方法を検討してきた．もし，中央のサーバにオンデマンドのリアルタイム配信機能があれば，各系列局では，局内のネットワーク型デコーダ（たとえば，文献 [6]）や IP パケットから映像信号に変換する装置（たとえば，文献 [7]）などを使ってサーバからの映像をいったん蓄積することなく好きなときに映像モニタに出力することができる．このことは，時間に追われる制作現場における作業時間の短縮と作業効率の向上に役に立つと考えられる．以上をまとめると，将来の映像制作には，以下の要件を満たすサーバシステムが必要になると考えられる．

(1) 高精細映像素材として毎秒 1.5 ギガビットの非圧縮ハイビジョン映像や毎秒 1 ギガビット弱の圧縮 4K ウルトラハイビジョン [8], [9], [10] 映像のリアルタイム送信が可能であること
(2) 少なくとも 30 カ所のクライアントから，サーバ内の任意の高精細映像素材データのオンデマンド配信リクエストを受けて，すべてのリクエストに対してリクエストされた映像の送信が可能であること

しかしながら，秒あたり 1 ギガビットを超える任意の高精細映像データを複数同時にオンデマンドでリアルタイム送信可能な性能を持ったサーバはない．そこで，我々はこれらのニーズを満たすとともに，高精細映像にも対応可能な映像サーバの構成方法について検討を行ってきた．一般的なファイルサーバや映像サーバは大容量のストレージを持った PC をベースに構成されているものが多い．ストレージについては複数の HDD や SSD を使い並列に読み書きすることによって高い性能を得られることから，映像サーバの性能限界は PC の内部転送速度の限界によって決まると考えられる．現在の PC の内部転送速度の理論的な限界速度は，ストレージ接続のためのホストバスアダプタ

[1] NTT 未来ねっと研究所
NTT Network Innovation Laboratories, Yokosuka, Kanagawa 239–0849, Japan
[2] 神奈川工科大学
Kanagawa Institute of Technology, Atsugi, Kanagawa 243–0292, Japan
a) kimiyama.hiroyuki@lab.ntt.co.jp
b) ogura.tsuyoshi@lab.ntt.co.jp
c) maruyama@nw.kanagawa-it.ac.jp

本論文の内容は 2012 年 7 月のマルチメディア，分散，協調とモバイル（DICOMO2012）シンポジウム 2012 にて報告され，マルチメディア通信と分散処理研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

（HBA）やネットワークインタフェースカード（NIC）を接続するための PCI Express バスの最大の転送速度であり，たとえば，Intel 社の最新の Xeon E5 プロセッサを使用した PC であれば，×8 PCI Express 2.0 の最大転送速度の 32 Gbit/s がそれに相当する [11].

そこで，我々は，将来のさらなる高精細映像や高品質映像のニーズに対応できるように，複数 PC により構成される PC クラスタアーキテクチャを適用することによって，スケーラビリティを持つ映像サーバの実現方式を検討してきた．PC クラスタアーキテクチャを映像サーバに適用するメリットは，映像データを複製することなく PC 1 台の性能限界を超えた配信能力を持つ映像サーバを実現できること，および，柔軟なシステム構成をとれることから，ユーザの要求条件にあった性能のサーバシステムを提供できることにある．我々は，その検討結果をもとに，16 台の汎用 PC とそれらを接続するための 1 ポートあたり 20 Gbit/s の InfiniBand switch を使った内部結合ネットワークおよび 8 台の汎用ディスクアレーを使って，24 Gbit/s の最大配信性能（非圧縮ハイビジョン最大 16 本を同時配信可能な性能）を持つ映像サーバシステムを実現した [12].

前述したように，国内放送局での利用を考えた場合，30 局近くの系列局がいっせいにリクエストを発した場合でもすべてのリクエストに応えられる必要がある．そのニーズを満たすために，1.5 Gbit/s の非圧縮ハイビジョン 30 本以上同時に配信可能な 50 Gbit/s 超の高速なサーバの実現を目指すとともに，1 映像ストリームあたりのシステムコストを下げる目的で，性能のボトルネックとなっている内部結合ネットワーク上の通信方式の再検討を行った．我々は内部結合ネットワーク上の新たな通信方式を提案，実装し，実機の PC サーバを使用しその効果を実測した．その結果，24 台の PC，1 台の InfiniBand switch，12 台の汎用ディスクアレーを用いて，世界で初めてとなる最大 36 本の任意の非圧縮ハイビジョン映像データをオンデマンドに配信可能な 54 Gbit/s の配信性能を持つサーバシステムを実現した．

本論文では，2 章において既存の 24 Gbit/s の配信性能を実現したクラスタ型映像サーバシステムの構成法について説明し，3 章では高速化のために新たに提案する内部結合ネットワーク上の通信方式について記述する．4 章ではその実装方式について説明し，実ハードウェアを使った性能評価結果を 5 章に示し，スケーラビリティに関する考察を 6 章に示す．最後に，まとめと今後の課題について説明する．

2. クラスタ型サーバ構成法

2.1 映像素材配信システム

クラスタ型映像サーバの構成について説明する前に，映像素材配信システムの構成について説明する．本論文で対象とする映像素材配信システムは，中央に映像が格納される映像サーバと高速なスイッチを使った高速なネットワーク，映像を再生するためのクライアントシステムから構成される（**図 1**）．高度な映像検索機能を付加するときには，映像の属性データ（メタデータ）を格納できる DBMS ソフトウェアがインストールされた PC を追加することもできる．

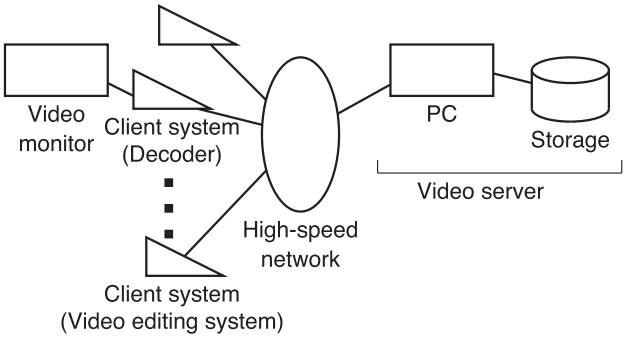

図 1 映像サーバシステムの構成
Fig. 1 Configuration of video server system.

クライアントシステムとして，リアルタイムに映像サーバから送信されてきた映像データを受信し，映像信号に変換する装置や，ネットワーク対応のデコーダ装置，あるいは映像編集システムを利用する．映像サーバは，映像データを格納するためのストレージと PC（またはワークステーション）から構成され，クライアントからのリクエストを PC が受け付け，映像データをストレージから読み出し，クライアントシステムに映像を送信する．通常，ストレージは PC の内部バスに接続されたホストバスアダプタカードを経由して接続されており，ストレージから読み出された映像データは，PC の内部バスを通じて PC 内部のメモリに書き込まれる．PC の内部メモリに書き込まれた映像データは，PC の内部バスに接続されたネットワークインタフェースカードまたはオンボードのネットワークコントローラを介してネットワークへ送信される．映像サーバの配信性能は，ストレージの読み出し（または書き込み）性能，ホストバスアダプタカード，ネットワークインタフェースカード，PC の内部バスの転送性能によって決まる．たとえば，前章で記述したように，PC の内部転送速度は最大で 32 Gbit/s であることから，これ以上配信性能を上げることは不可能である．この限界を超える性能の映像サーバを構築するためには，複数の PC を連動させて 1 つのサーバのように動作させる，つまり，クラスタ型サーバにする必要がある．次節では，これまで提案されているクラスタ型映像サーバの構成法について説明し，そのメリット，デメリットについて議論する．

2.2 ハードウェア構成

クラスタ型映像サーバは，複数の PC と複数のストレージ，それを接続するネットワークによって構成される．こ

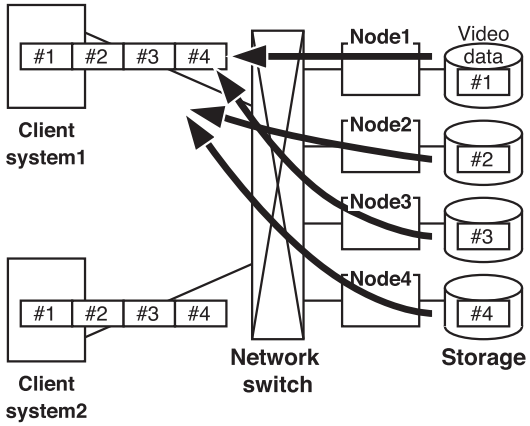

図 2 クラスタ型映像サーバ構成 1
Fig. 2 Cluster based video server system 1.

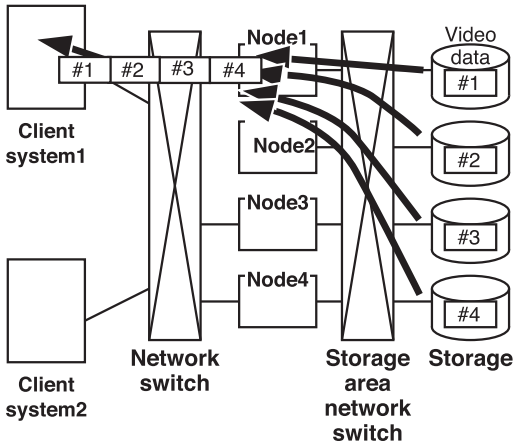

図 3 クラスタ型映像サーバ構成 2
Fig. 3 Cluster based video server system 2.

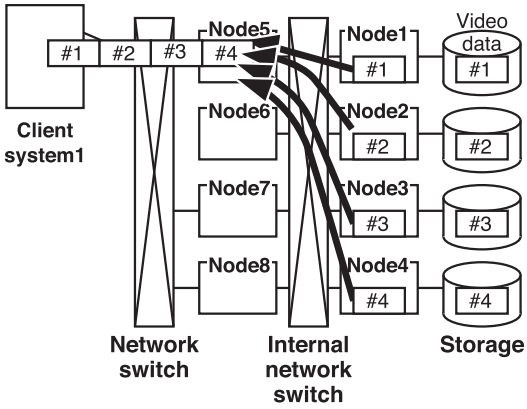

図 4 クラスタ型映像サーバ構成 3
Fig. 4 Cluster based video server system 3.

のサーバでは，各ストレージに対する負荷を低減させるために映像データを複数のストレージに分割して保存する．送信する際には，分割保存された映像データをストレージから並列的に読み出し，送信する．そして，クライアントへの送信経路上またはクライアントで 1 つの連続した映像データとして構築することによって，最終的に 1 つの映像ストリームとしてクライアントで再生することができる．このように映像データを分割保存することによって，映像データの複製を行うことなく，システム全体の性能を向上させることができる．複数の PC を使用したクラスタ型のサーバシステムはこれまで複数提案され，一部製品化されているものもある．これらのシステムの構成は，映像データをどこで 1 つの映像ストリームに組み立てるかによって 3 種類の構成に分類することができる．

構成 1 この構成は，**図 2** に示すようにストレージが接続された複数の PC（Node）をネットワークスイッチで接続する構成である．このシステムでは，蓄積時にクライアントにおいて映像データを時間軸方向（映像フレーム単位または映像セグメント単位）に分割し，Node 1，Node 2，... と順番に転送し，格納する．配信時には，格納した映像を Node 1，Node 2，... で順番に映像データを読み出し，クライアントに向けて送信し，クライアントにおいて映像データに組み上げることによって，リアルタイム再生を実現している [13], [14]．

構成 2 この構成は，**図 3** に示すように，Storage Area Network（SAN）スイッチを介して複数の PC（Node）と複数のストレージを接続する構成である．この構成では蓄積時に映像データをクライアントから 1 つの Node に送信し，その Node で時間軸方向に分割して SAN を介して各ストレージに格納する．送信する場合は各 Node が SAN を経由して，すべてのストレージからデータを集め，映像ストリームに組み上げてクライアントに送信することによって，クライアントにおけるリアルタイム再生を実現している [15], [16]．

構成 3 この構成は，**図 4** に示すようにクライアントに映像データを送信するための PC（Communication node，以降 CN），および，映像データを直結したストレージから読み出すための PC（Storage node，以降 SN）の 2 種類の PC を用意し，これらの PC を内部結合ネットワークスイッチを使って接続した構成である．蓄積時には，映像データをクライアントから 1 つの CN へ送信し，CN において構成 1 と同様に時間軸方向に分割して各 SN へ送信し，SN に接続されたストレージに格納する．送信する場合は SN がストレージからの読み出しを行い，CN が SN からデータを集めてストリームとして組み上げ，リアルタイムにクライアントに送信することによって，クライアントにおけるリアルタイム再生を実現する [17]．CN と SN 間は，ストレージから読み出されたすべての映像データが送受信されることから，十分大きな帯域が必要である．また，後述するように本構成では映像データトラヒックの競合による輻輳が問題となることから，配信用の外部ネットワークやストレージネットワークとは別に内部結合ネットワークを用意する必要がある．

ストリームサーバとして最も重要な機能は，リアルタイム配信機能，つまり，一定時間内に決められたデータ量を送信する機能である．この場合，問題となるのがネットワークにおけるトラヒック競合による輻輳である．分散構成の場合，前述したように，サーバとクライアントの間で，分散保存された映像を1つの映像ストリームにまとめる必要があり，そのためにネットワークスイッチを使わなければならない．もし，複数のNodeやストレージからの映像データの転送を同期をとらずに実行した場合，映像データが同時にネットワークスイッチの1つの出口向けに集まり，その出口においてトラヒック衝突による輻輳を起こす可能性がある．つまり，構成1の場合はクライアントにつながるネットワークスイッチ，構成2の場合はPCにつながるStorage Area Network Switch，構成3の場合はCNにつながる内部結合ネットワークの出口の帯域を超えるデータ量がNodeやストレージから集まった場合，輻輳を起こすことになる．輻輳は映像データの到着の遅れを引き起こすだけでなく，UDP（User Datagram Protocol）のような再送機能のないプロトコルで送信する場合は，最悪パケットロスを引き起こし，その結果，映像の乱れを生じることになる．ストレージやNodeを同期させて動かすことは文献[12]に示したように，同期が外れたときにシステム全体が停止するという問題があるため回避すべきである．この輻輳を回避するためには，できる限り出口帯域が大きいネットワークを使用するとともに，衝突により生じた映像データの到着遅延を吸収するためのバッファを用意する必要がある．さらに，Node数が増えれば増えるほどスイッチの出口において出口の帯域を超える量の映像データが集まる可能性が高くなり輻輳を起こす確率が高くなり，スケーラビリイティに大きな影響を及ぼすことになる．

上記の3つの構成に使えるネットワークについて価格，技術の成熟度合いから考えると，構成1のクライアントとPCをつなぐネットワークには10 Gigabit Ethernet，構成2のPCとストレージをつなぐネットワークには4 Gbit/sないし16 Gbit/sのFibre Channel，構成3の内部結合ネットワークでは20 Gbit/sまたは40 Gbit/sのInfiniBandが適当である．InfiniBandはスイッチ型のネットワークで，10 Gigabit Ethernetと比べてポート単価が安く，レイテンシが小さい．そのため，高いスループットが達成でき，かつ，上位プロトコルとしてTCP/IPをはじめとする様々なプロトコルをサポートしており，柔軟なシステム構成が実現可能なネットワークである．長距離伝送には向かないものの，同一ラックスペース内のPCどうしをつなぐのであれば十分適用可能である．これらの3つのネットワークを出口帯域の観点から比較すれば，構成3のInfiniBandを使う構成が有利であると考えられる．

さらに，各構成において，リアルタイム性を実現するための実装を考えた場合，以下のメリット，デメリットがある．構成1は，各Nodeから送られてきた映像データを，クライアントにある程度バッファを持たせることで組み上げて再生する構成である．構成1のメリットは使用する機材が少なくて済み，システムコストが一番低い構成だという点である．一般的に，クライアント装置は，たとえば，映像サーバが30拠点分の同時配信性能持っていたとしても，その数を超えて設置することが想定されるため，1秒あたりギガビットを超えるような高精細な映像データを配信する場合は，クライアントに大きなバッファを用意することはシステム全体のコストの面で不利となる．さらに，複数のPCからクライアントへの映像データの送信が完全に同期されていないと，クライアントに向かうネットワークスイッチの出口で衝突による輻輳を起こし，クライアントで必要とする時間までにクライアントに届かなかったり，信頼性のないUDPで送信する場合は，その輻輳により映像データそのものが失われたりする可能性がある．

構成2は，構成1に1台のSANスイッチ加えた構成であり，システムコストは構成1の次に低い．この構成において，複数のクライアント向けに複数のPCから読み出しリクエストが同時に行われた場合，ストレージでリクエストの競合を起こし，同時に行われたリクエストに対する応答が遅れる可能性があり，その場合，クライアントに対して一定時間内にデータを送ることができない可能性がある．

構成3の場合は，クラスタを構成するPCの数が倍になるものの，ストレージに対するリクエストをSNで順序制御することができることから，構成2のようなストレージに対するリクエスト競合を防止することができる．また，CNの入り口の帯域が，他の構成と比べて大きいので，スイッチの出口での競合が起きにくいというメリットがある．さらに，SNから転送されてくる映像データの順序制御とレート制御を別なNode（CN）で行うことで，負荷を分散し，他のプロセスの影響を受けずにリアルタイム性を確保しながらクライアントへ配信することができる．ストレージに対するアクセスを分離することで，たとえば，SN側に，リアルタイム送信に影響を及ぼしかねない高いCPUパワーが必要な暗号蓄積機能など付加価値を付けやすくなるメリットもある．

以上の結果から，我々は使用するPCの数は倍になるものの，他の構成よりもリアルタイム性を維持しやすい構成3をハードウェア構成として選択した[12]．しかしながら，構成3を採用した場合でも，すべてのPC（CNとSN）を完全に同期させない限り内部結合ネットワークスイッチの出口において衝突による輻輳が起こる可能性がある．すべてのPCを完全に同期させるのは，処理が複雑になるだけでなく，1つのPCの障害がシステム全体の障害に波及する可能性が高いため，同期をとる以外の衝突回避の方法が必要である．そこで，我々は，PC間の完全な同期をとることなく，通信の衝突を回避するための実装方式を提案し，

非圧縮ハイビジョン 16 本を同時に配信可能なサーバシステム（以降，16 多重サーバと呼ぶ）を実現した [12]．次節に，その提案方式の概要とさらなる性能向上を実現するうえでの課題について説明する．

2.3 16 多重サーバの実装方式と課題

前述した 16 多重サーバでは，CN から SN へのストレージからの読み出し処理の要求を SN へ等時間間隔で送り，その要求を契機に CN，SN 間のゆるやかな同期を実現した．その概略を**図 5** に示す，まず，CN がクライアントからのリクエストを受付ける（Step 1）．次に，CN から SN1 に対して，リクエストされた映像に該当する映像データの読み出しをリクエストする（Step 2）．そのリクエスト受付けた SN1 はストレージからの読み出しを行い，CN に対して映像データを転送する（Step 3）．引き続き，CN はクライアントに映像データを送信するとともに，次の SN（SN 2）に対して，映像データのリクエストを送り（Step 4），以降，同様に処理を進めることによって連続した映像データの配信を実現した．

我々は，16 多重サーバシステムの内部結合ネットワークを InfiniBand スイッチで構成した．InfiniBand スイッチとして，システム全体のスループットに対して十分高速なスイッチング性能（960 Gbit/s）を持つ QLogic 社製の SilverStorm 9024 InfiniBand スイッチ 1 台を使用し，PC とスイッチとの間は 20 Gbit/s（実効最大転送速度 16 Gbit/s）の DDR InfiniBand で接続した．事前に PC 内部の転送性能評価を行い，4.8 Gbit/s が本 PC の最大転送性能であることを確認し，それよりも十分高い帯域を持つ DDR InfiniBand を内部結合ネットワークとして採用した．また，InfiniBand のプロトコルとして，事前の評価結果 [18] から遅延時間の短い RDMA（Remote direct memory access）を採用した．前述したように，スイッチ型のネットワークを使ったこの構成では，スイッチング性能にかかわらず，複数の SN から同一の CN に対して映像データが同時に送信された場合，スイッチの出口で衝突が発生し，どちらかの転送が待たされることになる．この方式では CN，SN どうしを完全に同期させていないため，このような衝突が起こる可能性がある．このような衝突が起きた場合は，CN には，SN からの映像データが予定していたスケジュールよりも遅れて到着することになる．そのため，既存の 16 多重サーバシステムでは，衝突による遅延のゆらぎを CN 内のバッファリングによって吸収する方法を追加することによって，16 台の PC を使用し，非圧縮ハイビジョン 16 本を同時に配信できるクラスタ型サーバシステムを実現した．

この 16 多重サーバシステムで 4 本の非圧縮ハイビジョンを配信する場合の CN，SN の処理タイムチャートを**図 6** に示す．この図において，横軸は時間を示し，塗りつぶされている領域は読み出し処理または転送処理が行われている時間である．このシステムは，文献 [12] に示したように，非圧縮ハイビジョン映像を 1 映像フレーム単位に分割して各ストレージに蓄積しており，配信する場合も，蓄積する場合も，リアルタイム性を維持するために 1/2 フレーム時間（毎秒 30 フレームの映像の場合は 1/60 秒）でその処理を完了する必要がある．もし，その時間を超過した場合，その処理以降の処理がシステム全体で遅れていくことになる．このような場合に備えて CN やクライアント側にバッファを追加することによって，この遅れをある程度まで吸収することは可能であるが，何回も遅れが続けば最終的にはクライアントに必要な時間までにデータが届かない，つまりアンダフローを起こすことになる．したがって，性能的に余裕を持たせた設計をする必要があるため，この方式

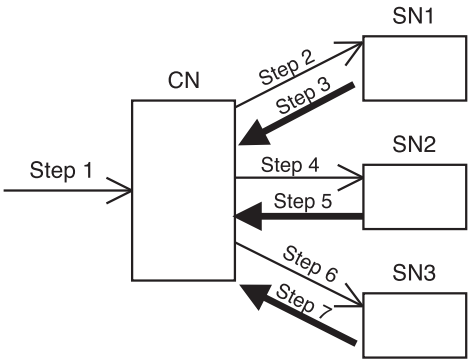

図 5 16 多重サーバの CN，SN 間転送方式

Fig. 5 Video data transmission protocol between CN and SN for 16-video stream server.

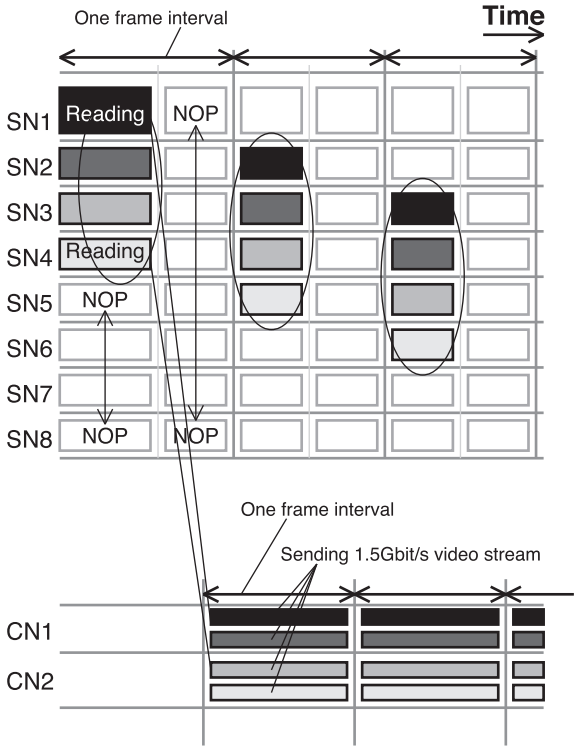

図 6 処理タイムチャート

Fig. 6 Time chart for processing on each node.

ではPCやストレージの性能を限界まで使用することができない．また，内部結合ネットワーク上での通信の衝突の確率は，CN，SNの台数が増えると指数関数的に増え，その増加に従ってCN，SN間の平均転送処理時間も指数関数的に増えることが確認されている．そのため，サーバシステム全体の送信能力を向上させるために，CN，SNを増やしても，ある時点で頭打ちになることが容易に想定される．内部結合ネットワークを2重化し出口帯域を増やすことによって衝突確率を減らすことはできるが，PCの内部転送速度より帯域を増やしても，PCの入り口で輻輳を起こすため性能向上は期待できない．また，コスト増につながるとともに故障点を増やすうえ，CN，SN数が増加していけばどこかで性能限界を迎えることは明白である．そこで，我々は，より少ないPC，ストレージ，内部結合ネットワークスイッチを使い，より多くの映像を配信できるシステムを実現するために内部結合ネットワーク上の通信方式を再検討した．

3. 内部結合ネットワーク通信方式の改良

前述した16多重サーバのCN，SN間通信方式は，図5に示すように，CNからSNに対してシーケンシャルに転送要求を行う方式である．前述したように，この方式はSNからの応答が一定時間内に返ってくることを前提としており，SNからの応答が1回遅延することによってこの映像ストリームの送信処理が遅延するだけでなく，システム全体の処理遅延が発生するという問題がある．そのため，PCやストレージの処理能力に対して，余裕を持たせるよう設計をする必要がある．このことは逆にいえば，処理能力の余裕分を有効活用することで，システム全体の処理能力を向上させられる可能性があると考えることができる．CN，SN間の通信処理に余裕を持たせるためには，バッファリングによりCNからクライアントへの映像データの送信開始を遅らせることにより実現可能であるが，シーケンシャルである限り処理遅延時間は蓄積されていくので，いずれどこかの時点でアンダフローを起こす可能性がある．

そこで我々は，CNからのリクエストをSNに対して1台ずつシーケンシャルに送信するのではなく，**図7**に示すようにすべてのSNに対して並列的に同時に送信する方式について検討を行った．**図8**に示すようにCN内にSNごとに複数のバッファを用意し，SNからの転送が完了し，次のバッファに空きがあれば，次のデータをSNの順番に関係なくSNにリクエストする．したがって，SNからCNに転送されてくる映像データは順番が入れ替わる可能性がある．CNは，すべてのSNからの最初の映像データが到着し，さらに数フレームバッファリングしてから，クライアントに向けて先頭の映像データから順番に送信を開始する．

図9に提案方式と従来方式のCNでのデータ受信処理の違いを示す．従来方式ではCNからSNに順番にリクエストを送信するので，その順番でCNにデータが到着し，CN内のバッファに転送される．提案方式では，並列的にリクエストが送信されるので，先に到着した順にCNでデータの受信が行われる．この方式によって，CNからリクエストを発してからSNからの映像データ受信受信完了までの時間が，1映像フレーム時間とSNの台数の積によって決まる時間を超えなければアンダフローは発生しないと考えられる．

しかしながら，この提案方式において，CNからすべてのSNに対して読み出しのリクエストを送信する場合，下記の3カ所でリアルタイム性を乱す恐れのある競合が起こ

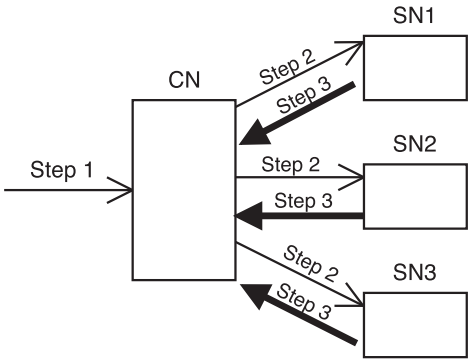

図7 CN-SN間転送方式
Fig. 7 New video data transmission protocol between CN and SN.

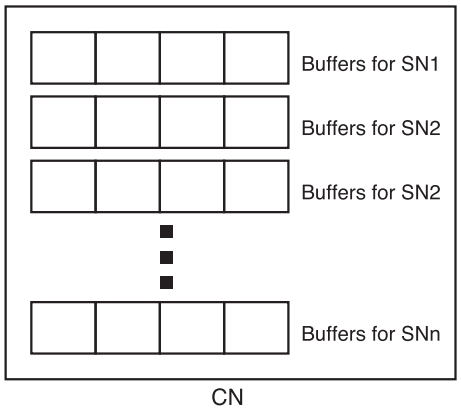

図8 CNバッファ構成
Fig. 8 Buffer structure for CN.

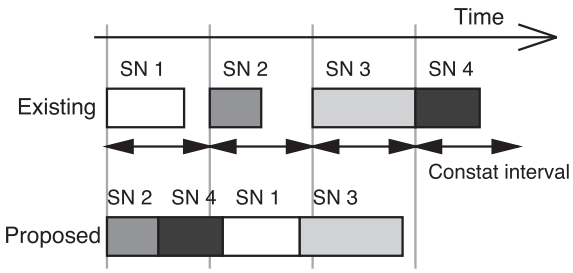

図9 CN上のSNからのデータ受信タイムチャート
Fig. 9 Time chart for CN during data receiving from SNs.

る可能性がある．

(1) SN は，複数の CN から非同期に読み取りリクエストを受信することになるため，SN 上で競合が起こる可能性がある．

(2) SN から CN への映像データの転送は，すべて内部結合ネットワークを経由することから，内部結合ネットワークスイッチ内部のスイッチ LSI 上で競合が起こる可能性がある．

(3) SN から CN へのデータ転送は，複数の SN から非同期に行われるため，スイッチから CN へ向かうネットワークまたは CN 上で競合が起こる可能性がある．

図 4 のハードウェア構成を用いた場合，たとえば，内部結合ネットワークとして InfiniBand を選択することによって，システム全体のスループットに対して十分高い転送（スイッチング）性能を確保できることから，内部結合ネットワークスイッチ LSI 上での競合は，無視できると考えられるが，それ以外の 2 つの競合に関しては無視することはできない．

この方式が利用可能かどうかを判断するためには，これら 2 つの競合について評価する必要がある．内部結合ネットワーク上の CN，SN 間通信は非常に複雑であり，それらを解析的に評価するのは困難である．そこで，我々は実機を用いて本方式の実装を行い，その有効性の評価を行った．

4. 実装

前述した内部結合ネットワーク上の通信方式の有効性を確認するため，汎用 PC サーバ，汎用ストレージおよび汎用 OS を使った実装方法の検討を行った．16 多重サーバシステムとの比較のため，ハードウェア（PC，ストレージ），OS，および InfiniBand プロトコルは同一のものを使用した．クライアントへ映像データを送信するためのネットワークインタフェースには一般的に広く使われている 10 Gigabit Ethernet を選択し，ストレージとのインタフェースには拡張性を考え，4 Gbit/s の Fibre Channel を選択し，SN とストレージとの間は 2 対の Fibre Channel を使って接続した．汎用 PC サーバと汎用の Linux を使ってリアルタイム処理を実現するために，疑似的なリアルタイム処理をサポートするミドルウェア（疑似リアルタイムモジュール：Quasi-real-time scheduler module）[12] を使用した．このミドルウェアは，OS の改版によってシステムコールが変わった場合やディスコンなどで OS そのものを変えざるをえない場合でも，大幅な変更なくアプリケーションを移植可能にするためのものである．本ミドルウェアはアプリケーションの各 Thread に対して一定時間間隔でソフトウェア割込みを送信し，その割込みを受け取った各 Thread がそれを契機に各処理を実行することによって，疑似的なリアルタイム処理を実現している [19]．CN および SN のソフトウェア構成を**図 10** および**図 11** に示す．

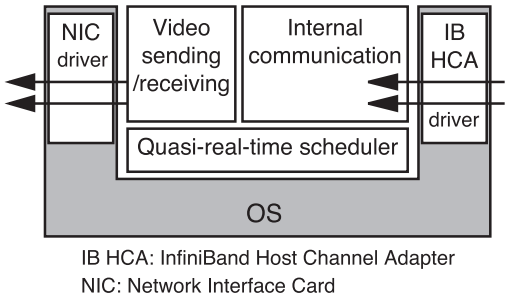

図 10 CN ソフトウェア構成
Fig. 10 Software configuration for CN.

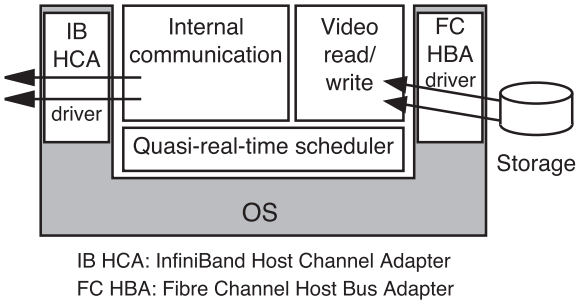

図 11 SN ソフトウェア構成
Fig. 11 Software configuration for SN.

まず，図 10 に示す CN のソフトウェアについて説明する．CN のソフトウェアは映像送受信処理モジュール（Video sending/receiving module）と内部通信処理モジュール（Internal communication module）から構成されている．映像送受信処理モジュールは，クライアントとのインタフェースの役割を持っており，各クライアントごとにその状態管理を行うとともに，疑似リアルタイムモジュールからの割込みにより，SN から送信されてきたバッファ上の映像データを一定時間間隔でクライアントに対して送信するものである．内部通信処理モジュールは，前述した CN，SN 間通信処理を実装したモジュールである．

図 11 に示す SN のソフトウェアは，映像読み書きモジュール（Video read/write module），CN との通信のための内部通信処理モジュール（Internal communication module）から構成されている．映像読み書きモジュールは，CN からのリクエストに応じて映像データを読み出す（または書き込む）モジュールである．少しでも読み出し処理の揺らぎを吸収できるように先読み機能を実装している．内部通信処理モジュールは CN と同様である．

また，CN から送られたデータがクライアントでオーバフローしないように，クライアントと CN との間には，文献 [20] で提案したクライアントのバッファ状態を CN にフィードバックし，そのバッファ状態をもとに送信レートを動的に変更するレート制御方式を実装した．

5. 評価

5.1 スループット評価

提案した内部通信方式の評価を行うため，前章のソフトウェアを実装し，**図 12** に示す最大 24 台の PC と 12 台の汎用ストレージを使った評価システムを構築し，評価実験を行った．CN および SN に使用したハードウェア，OS の環境を**表 1** に示す．この評価では，実際の再生可能な端末の代わりに，1 台あたり非圧縮ハイビジョン 3 本を同時に受信することが可能な汎用 PC サーバによって構築した端末エミュレータをクライアントとして用意した．

提案手法の効果を確認するため，CN と SN の台数を変えながらシステム全体のスループットを測定した．試験用映像データとして 1.5 Gbit/s の非圧縮ハイビジョン映像（約 5 分相当）をクライアント台数分用意し，それらをいったん蓄積し，各端末エミュレータから各 CN に対して映像配信リクエストを同時に 3 本送信し，CN から送られてくる映像データのスループットを映像ストリームごとに各端末エミュレータで測定した．文献 [12] に示したように，表 1 の PC とストレージを使った事前評価によって，ストレージからネットワークへの転送能力は最大 4.8 Gbit/s であることを確認している．その結果から，この PC は 1 台あたり 1.5 Gbit/s 非圧縮ハイビジョンを最大 3 本配信できる性能があると考えられる．そこで，本実験では，1 台の SN（CN）に対して最大負荷として 4.5 Gbit/s の負荷がかかるように，リクエストを送信して，その配信性能を評価した．

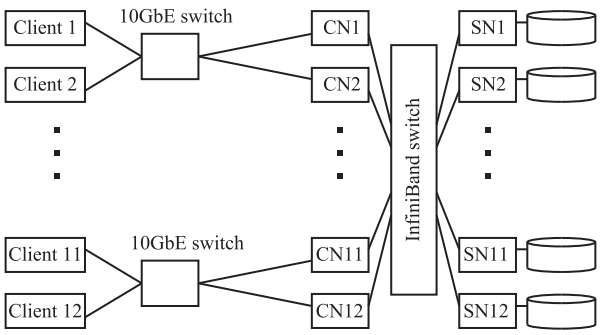

図 12 評価システム構成
Fig. 12 System configuration for evaluation.

表 1 CN，SN スペック
Table 1 Specifications of CNs and SNs.

Hardware	
CPU	Intel Quad Core Xeon 3.33 GHz x 2
Chipset	Intel 5000X
Memory	4 GByte
HCA	QLogic 7104-HCA-128LPX-DD
NIC（for CN）	NTT-IT PRESTA 10G NIC
HBA（for SN）	QLogic QLE2462
Software	
OS	RedHat Enterprise Linux 4

端末エミュレータで測定した平均スループットの総和，つまり，システム全体の平均スループットと CN（または SN）の台数との関係を**図 13** に示す．横軸は CN，SN の台数であり，縦軸は観測されたスループットの総和である．この図に示すように，本方式を実装したサーバにおいては，CN，SN の台数を 12 台まで増やしてもスループットをリニアに向上させられることが確認できた．また，5 分間にわたってスループットも一定であり，アンダフローが起きないことも確認できた．

5.2 内部映像転送処理時間評価

本提案方式は，最大負荷時（SN の台数 × 3 本の非圧縮ハイビジョン映像送信時）において CN，SN が単位時間あたりに転送すべき映像データ量は，システム構成（つまり CN，SN の台数）によらず一定であり，各 CN が SN に対して単位時間あたりに送信するリクエスト数も，各 SN が単位時間あたりに処理するリクエスト数も，システム構成によらず一定であるのが特徴である．しかしながら，CN の台数が増えることによって，SN に同時に到達する最大リクエスト数は増加するとともに，SN に対して複数のリクエストが同時に到達する確率も増加する．つまり，CN が 2 台の場合は，SN に同時に到達するリクエストは最大非圧縮ハイビジョン 6 本であるが，CN が 12 台になれば最大 36 本のリクエストが同時に到達する可能性がある．そのため，CN の台数が増えるに従って，リクエストが SN 上で衝突する確率も増加すると考えられる．そして，その衝突確率の増加は CN，SN 間の内部転送処理時間の増加として観測されると考えられる．

そこで，内部転送処理における衝突の評価を行うため，CN と SN の台数を変化させながら，最大負荷をかけたときの CN，SN 間の転送処理時間の測定を行った．平均転送処理時間，最大転送処理時間，最小転送処理時間の測定結果を**図 14** に示す．横軸は CN，SN の台数，縦軸は転

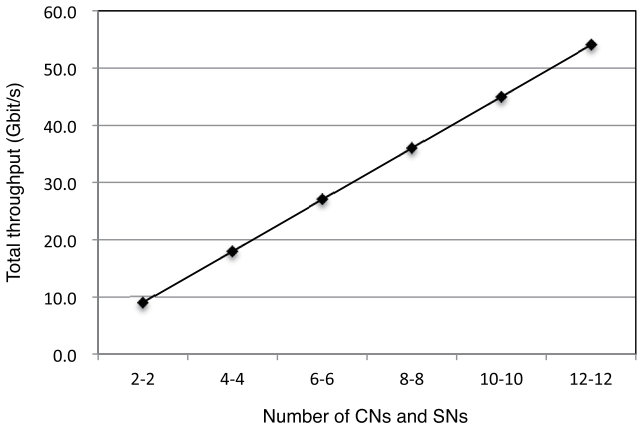

図 13 トータルスループット
Fig. 13 Total throughput.

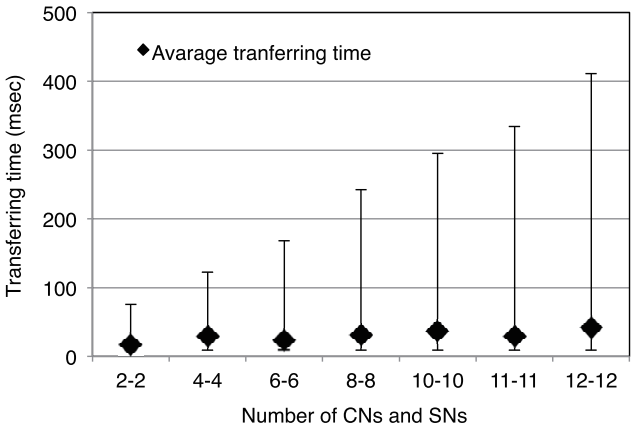

図 14 内部結合ネットワーク転送処理時間
Fig. 14 Transferring time over the internal network.

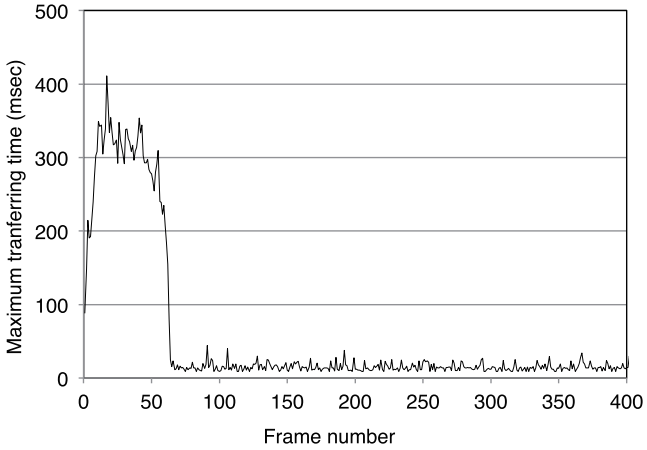

図 15 CN7 におけるフレームごとの転送処理時間
Fig. 15 Transferring time per frame for CN7.

送処理時間である．転送処理時間には，SN へリクエストを送信し，SN においてストレージからの読み出しを行い，SN から CN へ映像データが転送されてくるまでのすべての時間が含まれている．この図から分かるように，最小転送処理時間は CN，SN の台数に依存せずほぼ一定値であった．これはまったく衝突が起きない場合の転送処理時間であると考えられる．最大転送処理時間に関しては，CN と SN の台数にほぼ比例して増加することが確認された．最悪のケースでは，すべての CN から転送リクエストがほぼ同時にすべての SN へ送信されるため，最も待たされる場合，CN は SN の台数分処理を待たされることになる．したがって，SN の台数に比例して最大転送処理時間が増加するのは妥当であると考えられる．しかしながら，平均転送処理時間に関してはこれら 2 つのように単純には評価できないため，次章で近似を使った評価方法について示すとともに，本サーバシステムのスケーラビリティについて考察を行う．

3 章で記述したように，CN では SN の台数分の映像フレームを受信した後に，あらかじめ決めておいた時間（初期バッファリング時間）分の映像データを追加でバッファリングしてから，クライアントに向けて送信を開始する．初期バッファリング時間が 200 ミリ秒であれば，すべての SN からの最初の映像フレームを受信した後に，追加で 6 映像フレームを受信してから，クライアントへの送信を開始することになる．クライアントにおいてアンダフローを起こすケースは，SN から CN への転送処理時間のうち，SN の台数 ×1 フレーム時間（この試験では 1 フレーム時間は 1/29.97 秒に相当）を超えた時間の累積が初期バッファリング時間を超えた場合である．つまり，j 番目の CN，先頭フレームから SN の台数 N_{SN} を除く i 番目のフレームに対して，以下の式で計算される T_{ij} が，初期バッファリング時間 T_{buf} を超える場合である．

$$T_{buf} > T_{ij} \tag{1}$$

ただし，

$$T_{ij} = \sum_{k=1}^{i} (t_{kj} - N_{SN} \times t_{ave}) \tag{2}$$

ここで，t_{kj} は j 番目の CN で観測された k 番目のフレームの転送処理時間であり，N_{SN} は SN の台数であり，t_{ave} は 1/29.97 秒である．図 14 から分かるように，$N_{SN} = 12$ の場合の最大転送処理時間は，$N_{SN} \times t_{ave}$ である約 400 ミリ秒を超える 411 ミリ秒が観測されている．**図 15** に最大転送処理時間を観測した CN7 の先頭から 400 フレームのフレームごとの転送処理時間を示す．3 章で記述したように，CN にバッファがある限り SN に対して読み出しをリクエストするのが，本方式の特徴である．したがって，この図からも分かるように，最大転送処理時間の 411 ミリ秒が観測される前にすでに 11 ミリ秒を超えるデータがバッファリングされていることからアンダフローを観測することはなかった．

実際の実装では，T_{buf} の値として，余裕を持たせるため 3 フレーム時間（つまり約 100 ミリ秒）を設定したが，各フレーム，各 CN に対して上記の T_{ij} を計算した結果，この値が T_{buf} を超えることはなく，さらに，T_{ij} がゼロを超えることもなかった．本提案方式では，SN の台数が増えるに従ってバッファリング量が増えるため，先頭の映像データがクライアントに向かって送信開始されるまでの時間が増加する．そのため，クライアントにおける応答時間が悪化することから，今後，最適なバッファ量に関して評価，検討していく必要があると思われる．

以上の結果から，提案した内部結合ネットワーク通信方式を導入することによって，PC1 台あたりの最大配信性能（スループット）を 1.5 Gbit/s から 2.25 Gbit/s まで，1.5 倍に引き上げることが可能となり，さらに，最大で 12 台の SN，12 台の CN，つまり合計 24 台の PC を使ったシステムを構築した場合でも，内部ネットワークにおける衝突によるアンダフローを回避し，スケールアップできることが

確認できた．本提案方式を適用することによって，システム全体で，既存システムの 2 倍を超えるリアルタイムストリーム配信能力を持つクラスタ型サーバの構築が可能であることを確認した．

6. スケーラビリティに関する考察

本サーバシステムのスケーラビリティ，つまり，PC（CN，SN）の台数とシステム全体のスループットとの関係について考えるうえで，CN，SN 間の転送処理時間が CN，SN の台数が増えることによってどのように変わるかを理論的に把握する必要がある．3 章で記述したように，CN で観測される転送処理時間は，CN から SN への転送リクエストの送信から始まり，SN においてストレージからの映像データの読み出し処理が行われ，CN へデータが転送されて，CN で受信完了するまでの時間である．CN，SN 間の転送処理の過程で競合が起きるのは，前述したとおり (1) SN への読み出し処理が複数の CN から同時に届くケース，(2) SN から CN への映像データ転送が同時に同一の CN に対して行われるケースであると考えられる．これら 2 つのケースが転送処理時間に対してどのような影響を与えるかを以下に考察する．

SN における映像データの読み出し処理は，リアルタイム性を保持するために同時に処理されず，リクエストはいったんキューイングされ，1 リクエストずつ実行される [12]．そこで，このケースに対しては SN を窓口が 1 の待ち行列と見なして待ち行列理論を用いて考察する．待ち行列理論で解析するためには，SN に対するリクエスト到着時間間隔分布と SN での処理時間分布が必要である．これらの分布は一般分布となるため，1 台の SN を窓口 1 の待ち行列と見なして GI/G/1 の待ち行列システムを解析すればよいことになる．定常状態において，CN からのリクエストを受信し SN での読み出し処理が完了するまでの平均処理時間 E_{SN} は次式で表すことができる [21]．ここで，CN から SN へのリクエストの到着時間間隔の平均と分散をそれぞれ E_{it}，V_{it}，SN での読み出し処理時間分布の平均と分散をそれぞれ E_{rd}，V_{rd} とした．

$$E_{SN} = \frac{\left(\frac{V_{it}}{E_{it}^2} - 1\right) + \rho_{SN}\left(\frac{V_{rd}}{E_{rd}^2} + 1\right)}{2(1-\rho_{SN})\mu_{SN}} + E_{rd} \qquad (3)$$

ただし，

$$\rho_{SN} = \frac{E_{rd}}{E_{it}} \qquad (4)$$

$$\mu_{SN} = \frac{1}{E_{rd}} \qquad (5)$$

次に，SN から CN への映像データ転送について考える．SN から CN への転送は，SN は CN のリソースが空くまで待ち，空いてから転送を開始する．CN において複数の映像データが同時に受信可能かどうかは，InfiniBand HCA のハードウェアやデバイスドライバの実装に依存すると考えられる．CN において n 本の映像転送が同時に可能である場合は GI/G/n の待ち行列と見なすことができるが，この特性を解析的に求めるのは困難である．CN において n 本の映像を同時に受信する場合は，1 本の映像しか受信できない場合と比べて，受信処理にかかる時間は n 倍になるものの平均待ち時間は減るため，SN から CN への転送時間は短くなることが想定される．実際，到着時間の分布がポアソン分布，処理時間分布もポアソン分布の場合，同時に処理できる数が多い方が平均待ち時間が少ないことが分かっている [22]．本評価の目的は，本方式のスケーラビリティの限界，つまり，最大接続可能台数の評価である．そこで，CN において同時に 1 本の映像転送しかできないと仮定して，SN と同様に GI/G/1 の待ち行列として評価すると，評価された値は実際よりも過大評価となるものの，スケーラビリティの限界に対する指標を得ることはできると考えられる．SN におけるストレージからの読み取り処理と同様に，SN から CN への転送処理が完了するまでの平均処理時間 E_{CN} は次式で表すことができる．ここで，SN への映像データの転送処理が生起する時間間隔分布の平均と分散をそれぞれ E_{dt}，V_{dt} と，SN から CN への転送処理時間分布の平均と分散をそれぞれ E_{tr}，V_{tr} とした．

$$E_{CN} = \frac{\left(\frac{V_{dt}}{E_{dt}^2} - 1\right) + \rho_{CN}\left(\frac{V_{tr}}{E_{tr}^2} + 1\right)}{2(1-\rho_{CN})\mu_{CN}} + E_{tr} \qquad (6)$$

ただし，

$$\rho_{CN} = \frac{E_{tr}}{E_{dt}} \qquad (7)$$

$$\mu_{CN} = \frac{1}{E_{tr}} \qquad (8)$$

我々が必要としているのは，CN，SN の台数が増加したときに SN へのリクエスト送信から CN への転送終了までの処理時間がどのように変化するかである．このシステムでは，シーケンシャルアクセスが多いことから，ディスクからの読み出し処理に先読み処理を導入している [12]．そのため，連続した映像データを読み出す場合，CN からのリクエストが到着した時点で先読み処理によりディスクからの読み出しが完了している場合がほとんどである．したがって，ディスクからの読み出し処理時間は，ほぼ一定の分布をするものと考えられる．よって，SN での読み出し処理時間は CN の台数には依存しないことから，E_{rd} は CN に依存せずほぼ一定，V_{rd} はゼロで近似することができる．また，リクエストの到着時間間隔の平均値 E_{it} も，本アーキテクチャから CN の数には依存しない．したがって，E_{SN} のうち，CN（SN）の台数に依存するのは V_{it} だけとなり，CN に依存しない定数を A_{SN}，B_{SN} とすると，式 (3) の E_{SN} は以下のように表すことができる．

$$E_{SN} = A_{SN} \times V_{it} + B_{SN} \qquad (9)$$

$$A_{SN} = \frac{E_{rd}}{2E_{it}(E_{it} - E_{rd})} \quad (10)$$

$$B_{SN} = \frac{E_{rd}}{2} \quad (11)$$

同様に，SN から CN への転送処理について，同時に 1 映像データの転送しか行われないと仮定したことから，つねに最大レートで CN に送信されると見なすことができる．したがって，E_{tr} は，CN（SN）の台数に依存せずほぼ一定であると見なすことができ，同様に V_{tr} はゼロで近似することができる．また，SN から CN への映像データの転送処理が生起する時間間隔分布の平均値 E_{dt} は，定常状態では SN，CN 間で一定レートで送信されていることから，CN（SN）の台数に依存せず一定である．このことから，E_{CN} も A_{CN} および B_{CN} を CN の台数に依存しない定数とすると，式 (6) は次式のように表すことができる．

$$E_{CN} = A_{CN} \times V_{dt} + B_{CN} \quad (12)$$

$$A_{CN} = \frac{E_{tr}}{2E_{dt}(E_{dt} - E_{tr})} \quad (13)$$

$$B_{CN} = \frac{E_{tr}}{2} \quad (14)$$

SN への読み出しリクエストが CN から発せられ CN への映像データ転送が完了するまでの時間は，E_{SN} と E_{CN} の和であることから，CN の台数によって CN，SN 間の転送処理時間がどのように変化するかを推測するためには，(1) CN，SN 間の CN からのリクエストの到着時間分布の分散と，(2) CN への転送時の待ち時間分布の分散が，CN の台数に応じてどのように変化するかを考察すればよい．もし，SN での先読み処理が十分働いているのであれば，E_{rd} はメモリ間の平均転送時間となるので，A_{SN} は A_{CN} に比べて十分小さいと考えられる．ゆえに，CN から SN へリクエストを送信し，その応答である映像データを受信し終わるまでの時間と CN の台数の関係は，V_{it} と CN の台数の関係と同じ傾向を示すと考えられる．

一方，前章の CN，SN 間の転送処理時間測定データから CN，SN 間の転送処理時間分布の分散を計算し，その値をプロットしたものを**図 16** に示す．この図の横軸は CN（SN）の台数であり，縦軸が実測された分散の値である．この図から転送処理時間の分散は CN の台数とともに指数関数的に増加していくことが確認できた．

そこで，転送処理時間の分散と CN の台数の関係と同じ傾向を示すと仮定して，図 14 に示した平均転送処理時間と CN の台数との関係を示すグラフを指数関数でフィッティングを行い，その曲線を外挿したものを**図 17** に示す．CN の台数を N_{CN} としたときに，この平均処理時間が $N_{CN} \times t_{ave}$ を超えた時点でバッファリングによる救済は不可能となるため，そこが本アーキテクチャの理論限界であると考えられる．この図から，40 台を超えて CN と SN を接続しても，1 台あたり 2.25 Gbit/s のストリームを

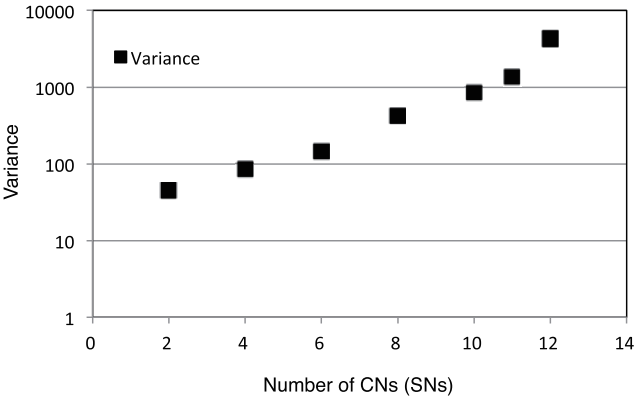

図 16 CN-SN 間転送処理時間の分散と CN（SN）台数との関係
Fig. 16 Relationship between variance of transferring time and number of CNs (SNs).

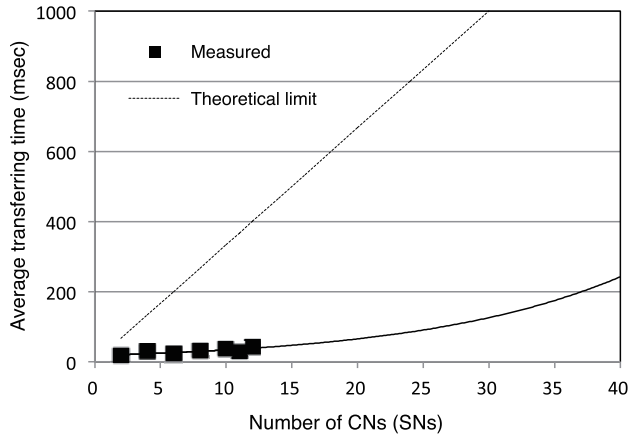

図 17 平均転送処理時間と CN（SN）台数との関係
Fig. 17 Relationship between average of transferring time and number of CNs (SNs).

問題なく処理可能，つまり，システム全体で 180 Gbit/s までスケールアップ可能であることが確認できた．しかしながら，図 16 と図 17 に示したフィッティングによる曲線の指数の乗数はそれぞれ異なっている．今後，さらなる高速化や効率化を図ることを考えた場合，上記のモデルと近似方法の精度を高め，限界を精度良く把握する手法を確立する必要があると思われる．

7. まとめ

本論文では，ネットワークを使った映像素材共有のための中核装置である映像サーバの実現方法として，PC クラスタをベースにした分散型映像ストリームサーバの 3 種類の構成法について説明した．我々は，蓄積された映像の試写が必須であるという観点から，リアルタイム配信機能を実現するため CN と SN という 2 種類の PC を用意し，それらを高速な内部結合ネットワークで結合し，CN によってリアルタイム性を保持しながら，クライアントに映像を送信する構成方法を採用してきた．その構成を使って映像ストリームサーバを実現するために，過去に提案した実装

方式について説明した．そして，さらなる性能向上の要求に応えることを目的に，性能向上のボトルネックとなっている従来の内部結合ネットワーク通信方式を改良した新しい通信方式を提案した．

この提案方式を実装し，実機を使い性能評価を行った．その結果，本方式により，従来方式と比較して PC 1 台あたり 1.5 倍の配信性能を得られることを確認するとともに，24 台の PC，12 台の汎用ストレージ，1 ポートあたり 20 Gbit/s の InfiniBand スイッチを使用することによって，ストリームサーバとしては最速の 54 Gbit/s（非圧縮ハイビジョン 36 本）の同時配信性能を持つサーバシステムが実際に構築できることを確認した．さらなる拡張性について検討するため，内部結合ネットワーク上の通信について，CN，SN 間の転送処理をモデル化し，待ち行列理論に基づいた考察を行い，CN（SN）を 40 台まで増設した場合でも本アーキテクチャは問題なく適用可能であることを示した．

今後，初期バッファ量の最適化を行い応答性の向上を図っていくとともに，次世代のウルトラハイビジョンやスーパハイビジョンへの対応を目指し，内部結合ネットワーク内の通信モデルについて，その高精度化を検討し，さらなる拡張の可能性について検討を行っていく予定である．また，最新のハードウェアに対して，本アーキテクチャを適用することで，さらに高速でかつコストパフォーマンスの高い映像サーバシステムの実現可能性を検証していく予定である．

謝辞 本研究の一部は，独立行政法人情報通信研究機構の「高度通信・放送研究開発委託研究ダイナミックネットワーク技術の研究開発」の一環として実施しました．

参考文献

[1] Aspera Inc.: Customer – Universal Pictures, available from ⟨http://asperasoft.com/customers/customer/view/Customer/show/universal-pictures/⟩ (accessed 2013-04-22).
[2] Signiant Inc.: | Mainichi Broadcasting System | Signiant, available from ⟨http://www.signiant.com/customers/case-studies/mainichi-broadcasting-system/⟩ (accessed 2013-04-22).
[3] 上村 明，菊池秀彦，星 英之：報道素材のファイル伝送運用，放送技術，Vol.64, No.5, pp.65–68 (2011).
[4] 株式会社 TBS テレビ：JNN Web, 入手先 ⟨http://www.tbs.co.jp/jnn/⟩（参照 2013-04-22）.
[5] 日本テレビ放送網株式会社：国内ネットワーク | 会社概要 | 企業・IR 情報 | 日本テレビ，入手先 ⟨http://www.ntv.co.jp/info/outline/domestic.html⟩（参照 2013-04-22）.
[6] NTT エレクトロニクス株式会社：HV9100 シリーズ | AVC/H.264 対応 HDTV/SDTV エンコーダ/デコーダ | NTT エレクトロニクス，入手先 ⟨http://www.ntt-electronics.com/digital_video/products/hv9100/index.html⟩（参照 2013-07-24）.
[7] 鍋谷栄展，小林正之，田中篤史，山崎恭啓：Qool Tornado QG70—非圧縮超高精細映像の高速 IP ネットワーク伝送で切り開く映像新時代，PFU テクニカルレビュー，Vol.24, No.1, pp.1–8 (2013).
[8] International Telecommunication Union: Ultra High Definition Television: Threshold of a new age, available from ⟨http://www.itu.int/net/pressoffice/press_releases/2012/31.aspx⟩ (accessed 2013-04-26).
[9] Sony Corporation: XAVC Specification Overview | XAVC | Sony, available from ⟨http://www.xavc-info.org/xavc/XAVCSpecificationOverview.html⟩ (accessed 2013-08-03).
[10] 白川千洋，野村 充，石丸勝洋，山口高弘，藤井哲郎：デジタルシネマを支える NTT の技術，NTT 技術ジャーナル，Vol.18, No.4, pp.51–55 (2006).
[11] Intel Corporation: インテル®C600 シリーズ・チップセット，入手先 ⟨http://www.intel.co.jp/content/www/jp/ja/chipsets/server-chipsets/server-chipset-c600.html⟩（参照 2013-07-10）.
[12] 君山博之，小倉 毅，丸山 充，伊藤秀一：自立分散型オーバ 10Gbit/s 映像ストリーム PC クラスタサーバシステムの構成法，電子情報通信学会論文誌 B，Vol.J93-B, No.7, pp.965–976 (2010).
[13] Kamezawa, H., Nakamura, M. Tamatsukuri, J., et al.: Inter-Layer Coordination for Parallel TCP Streams on Long Fat Pipe Networks, *Proc. 2004 ACM/IEEE conference on Supercomputing* (*SC 04*), IEEE Computer Society, pp.24–33 (2004).
[14] Liao, X. and Jin, H.: A new distributed storage scheme for cluster video server, *J. Syst. Archit.*, Vol.51, No.2, pp.79–94 (2005).
[15] Silicon Graphics International Corp.: SGI – Products: SGI InfiniteStorage: Storage Management: SGI InfiniteStorage Shared Filesystem CXFS, available from ⟨http://www.sgi.com/products/storage/software/cxfs.html⟩ (accessed 2012-05-10).
[16] 竹内 理：ストリーム蓄積・配信処理向け SAN ファイルシステム DJMFS の実装，情報処理学会研究報告，Vol.2006, No.44(OS-102), pp.77–84 (2006).
[17] Fukazawa, K., Suzuki, H. and Sasaki, C.: Distributed Video Server Using a New Striping Technique, *Proc. Multimedia Networks: Security, Displays, Terminals, and Gateways*, *SPIE*, pp.147–157 (1998).
[18] 小倉 毅，君山博之，釘本健司，川野哲生，清水健司，丸山 充：InfiniBand を用いた PC クラスタ型高速ストリームサーバアーキテクチャ，信学技報，Vol.NS2005-32, pp.33–36 (2005).
[19] Kimiyama, K., Shimizu, K., Kawano, T., et al.: Real-time Processing Method for Ultra-high-speed Streaming Server based on PC Linux, *Proc. 18th International Conference on Advanced Information Networking and Applications* (*AINA-2004*), IEEE Computer Society, Vol.2, pp.441–446 (2004).
[20] 釘本健司，小倉 毅，君山博之，川野哲生，清水健司，丸山充：非圧縮 HDTV-IP 伝送におけるフィードバック制御の検討：応答性の高いストリーミングサーバの実装と評価，信学技報，Vol.CQ2006-37, No.239, pp.13–18 (2006).
[21] 高橋敬隆：電子情報通信学会知識ベース 3 章拡散近似法（非マルコフモデル近似式導出法），入手先 ⟨http://www.ieice-hbkb.org/files/05/05gun_01hen_03.pdf⟩（参照 2013-04-22）.
[22] 木暮 仁：M／M／s 型＜待ち行列＜オペレーションズ・リサーチ＜ Web 教材＜木暮，入手先 ⟨http://www.kogures.com/hitoshi/webtext/or-que-mms/index.html⟩（参照 2013-04-22）.

推薦文

映像データを送信する複数のノードと映像データをストレージから読み出す複数のノードを内部結合ネットワークで結合し，非圧縮ハイビジョンを 36 本同時配信可能な，54GBit/s というきわめて高速な配信性能を実現するストリームサーバの構成法を提案している．大容量のストリームサーバを，安価な PC クラスタシステムで実現したことは高く評価でき，実用的な面で高い貢献が認められる．よって，本研究会からの推薦に値する．

（マルチメディア通信と分散処理研究会主査　勝本道哲）

君山 博之 （正会員）

1990 年東北大学大学院工学研究科原子核工学専攻修士課程修了．博士（工学）．同年日本電信電話株式会社入社．クラスタ型高速並列ストリームサーバシステムの研究に従事．電子情報通信学会，ACM 各会員．

小倉 毅 （正会員）

1994 年神戸大学大学院工学研究科システム工学専攻修士課程修了．同年日本電信電話株式会社入社．ネットワークプロトコル，並列処理アーキテクチャ等の研究に従事．電子情報通信学会会員．

丸山 充 （正会員）

1985 年電気通信大学大学院応用電子工学専攻修士課程修了．博士（工学）．同年日本電信電話株式会社入社．2012 年より神奈川工科大学情報学部教授．リアルタイム指向ネットワークコンピューティング構成技術の研究に従事．電子情報通信学会，IEEE 各会員．