

科学研究費補助金成果報告書

平成 22 年 3 月 31 日現在

研究種目：基盤研究（C）

研究期間：2007～2009

課題番号：19500066

研究課題名（和文）：PC グリッドによる高信頼・高効率な分散仮想ストレージの研究

研究課題名（英文）：A Study on PC Grid based High Reliable and Efficient Distributed Virtual Storage

研究代表者

上原 稔（UEHARA MINORU）

東洋大学・総合情報学部・教授

研究者番号：70256775

研究成果の概要（和文）

既存アプライアンス製品により大容量ストレージを構成する方法は耐費用効果が低い。本研究では、PC グリッドの空き容量を集約することで容量効率の優れた大容量ストレージを構成する方法を提案する。本研究では 50%以上の容量効率を目標とする。このような PC データグリッドでは高い信頼性が求められる。そのため階層型 RAID など高度な信頼化技法を適用することで要求に応える。

研究成果の概要（英文）

In the way for constructing large storage by conventional appliance products, the cost performance is low. In this study, we propose the way for constructing large storage by PC data grid, which realizes large storage by collecting free disk spaces. Furthermore, it aims that its capable efficiency should be more than 50%. In such PC data grid, high reliability is required. We achieve it by applying advanced reliability techniques such as hierarchical RAID to PC data grid.

交付決定額

（金額単位：円）

	直接経費	間接経費	合計
2007年度	2,100,000	630,000	2,730,000
2008年度	700,000	210,000	910,000
2009年度	700,000	210,000	910,000
総計	3,500,000	1,050,000	4,550,000

研究分野：総合領域

科研費の分科・細目：情報学・計算機システム・ネットワーク

キーワード：分散システム、グリッド、RAID

1. 研究開始当初の背景

(1) クラウドなど大容量ストレージの必要性が高まっており、それを安価に構築するための研究が急務となっている。

(2) HDD の大容量化と OSS による 64 ビットファイルシステムの開発により大容量ストレージが身近になった。

(3) SAN(Storage Area Network)はその名の通りストレージに特化した特殊なネットワークを用いるため高価になる。よって価格が大規模化を妨げる。ここで簡単な試算を紹介する。現在、60TB のファイルサーバを SAN で構築するのに必要な経費は 2.5 億円(定価)である。しかし、純粋に必要な HDD だけの経費は 500 万円であり、その比は 50:1 になる。SAN システムの 98%は容量増に寄与していない。効率の悪いシステムだといえる。一方、60TB の容量を持つ PC グリッドは 0.5 億円で導入できるので、SAN との比は 5:1 となる。PC グリッドは計算グリッドにも利用できるため SAN より投資効果が高い。

(4) P2P に基づくストレージは基本的に複製方式に基づく。そのため容量効率は 50%を下回る。無料の外部資源を利用できるため、容量効率が低くてもコスト増にはつながらない。しかし、自前でストレージ資源を用意しなければならない場合は容量効率の低さが問題となる。

2. 研究の目的

(1) 本研究では、ネットワーク仮想ディスクによる階層型 RAID 構成技術を用いて、高信頼、高効率な大容量ストレージを PC グリッド上に実現する。ここで、高信頼性とは SAN と同程度の信頼性を指し、高効率とは 50%以上の容量効率と優れた価格性能比を指す。基盤となる PC グリッドが不安定なため SAN より高い信頼性が要求される。そこで階層型 RAID を用

いて耐故障を実現する。

3. 研究の方法

(1) 大容量ストレージを仮想ディスクで構成するツールキット VLSD(Virtual Large Scale Disks)を開発する。

(2) 500 台の PC からなる 60TB の大容量ストレージを試作する。実際に 500 台の PC を利用することは困難であるため、少数の PC 上で仮想的に 500 台分のサーバを稼働する。そのため、1 台の当たりの PC の容量を大きくする。また、ストレージサーバは仮想マシン(VMware)で動作させる。

(3) I/O のボトルネックを最小とするため、e-SATA2 HDD と 1G BASE LAN を用いる。

4. 研究成果

(1) VLSD

本節では大規模ストレージ構築のための VLSD(Virtual Large Scale Disk)ツールキットについて述べる。VLSD は大規模ストレージ構築のためのツールキットであり、Java によるソフトウェア RAID 実装と NBD 実装を含む。VLSD は 100% pure Java であり、Java が動作するプラットフォームの上なら VLSD も動作する。そのため Windows や Linux が混在する環境に適している。

VLSD を用いると OS に制約されることなく NBD デバイスと RAID を自由に組み合わせることができる。最低限必要な NBD デバイスはファイルサーバの 1 つである。

Linux の nbd-server コマンドや Windows の nbdsrvr コマンドは単一ファイルを仮想ディスクとして公開する。そのため 4GB の制約がある FAT32 で動作させた場合、120GB/2GB=60 プロセスの NBD サーバを稼働させる必要がある。VLSD は複数のファイルを単一の JBOD にまとめて公開することができる。

VLSD ツールキットには以下のクラスが含まれる。

- FileDisk
- VariableDisk
- JBOD
- RAID0,1,3,4,5,6
- NBDDisk
- RemoteDisk
- WebDisk
- CipherDisk

(2) 60TB ストレージ

図1にVLSDを用いて分散ストレージを構成した例を示す。

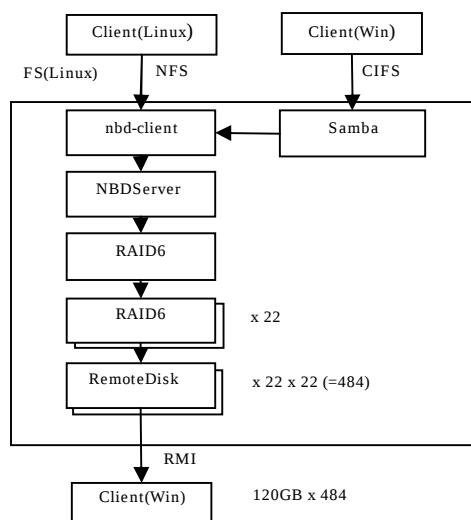


図1. 60TB 試作ストレージ

クライアントは500台存在し、そのOSはLinuxまたはWindowsである。それらはそれぞれNFS、CIFSで1台のファイルサーバと通信する。クライアントは同時にNBDサーバでもある。各クライアントでは空き容量を束ねた1つのNBDサーバが稼働する（従来のシステムでは複数のNBDサーバを稼働させなければならない場合があった）。ファイルサーバはSambaの稼働するLinuxマシンである。ファイルサーバでは、クライアントの分だけ

NBDDisk（後述）を作成し、22のNBDDiskから1つずつ合計22のRAID6を作成し、最後に22のRAID6から1つのRAID6を作成する。このRAID0FileをNBDサーバで公開し、自分自身のNBDデバイスで参照する。

(3) MMVD

単一大規模ディスクの問題点は、ファイルサーバがボトルネックとなることである。試作60TBシステムには、ファイルサーバが1台だけ存在する。500台のクライアントがそのファイルサーバに集中してアクセスする。試作システムでは、並行処理など十分な性能チューニングを行っていないため、ファイルサーバの負荷が高まり、性能が飽和してしまう。

このようなボトルネックを回避するために、大量小型仮想ディスク(Many Micro Virtual Disks, MMVD)を提案した。MMVDでは、クライアントごとに専用の小型仮想ディスクを用意し、それらを分散して配備する。試作システムでは500台のクライアントが存在するため、MMVDも500個必要になる。

MMVDの構成を図2に示す。各PCには小型のファイルサーバが組み込まれている。ファイルサーバ(FS)はLinuxで実装され、仮想マシン上で動作する。WindowsクライアントからはCIFSで、LinuxクライアントからはNFSでアクセスされる。また、各PCは120GBの空容量を持つ。これを2GBずつ60台の仮想ディスクに分割する。仮想ディスクごとにDiskServerが稼働する。ゆえに各PCで60のDiskServerが並行して稼働する。

MMVDでは、ユーザは専用のユーザディスクを割り当てられる。ユーザディスクの容量は12GBで、2GBの仮想ディスク6台から構成される。1つのユーザディスクは1つのRAID6に対応する。

各 PC は 120GB の空容量を持つ。この容量は 12GB のユーザディスク 10 台分に相当する。それゆえ、1 つの FS は 10 ユーザの専用ファイルサーバとなる。FS は、異なる 6 台の PC から 2GB ずつ計 12GB のユーザディスクを 10 台構成する。ユーザディスクは多くとも 6 台の仮想ディスクで構成される。それゆえ同時故障数は多くないと仮定できるため、RAID6 で十分と考えられる。

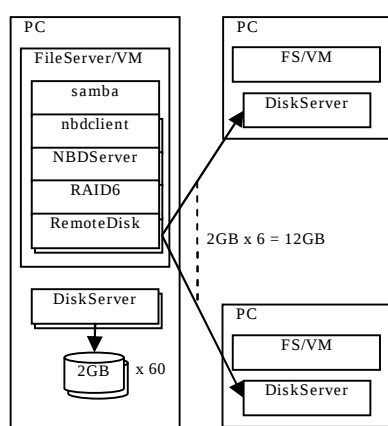


図 2. MMVD

(4) NaryRAID

我々は RAID55 より容量効率の優れた 3 耐故障 RAID 方式として、NaryRAID を提案した。ただし、その実装方式は本論文と異なりオブジェクト指向設計に基づくものではなく、すべてを 0 から作成するオールインワン方式である。NaryRAID では、ディスク番号の N 進数における各桁の数値が等しいディスクを集めてパリティグループを構成する。各ディスクが少なくとも 3 つのグループに所属すれば 3 耐故障となる。NaryRAID の基数を N とし、データディスク数 D を N^n としたとき、パリティディスク数 P は Nn である。NaryRAID は基数 N と次数 n によって特徴づけられる。よって、一般化して NaryRAID(N, n) と表す。図 1 に基数 2、次数 3 の NaryRAID(2,3) を示す。

NaryRAID(2,3) では、8 台のデータディスクと 6 台のパリティディスクで構成される。基数 N が小さいと容量効率はあまり大きくないが、基数 N が十分大きければ P は無視できるほど小さくなる。よって、容量効率は高い。しかし、容量効率の高さは信頼性の低さでもある。RAID の設計においては、容量効率と信頼性はトレードオフの関係にあり、信頼性を追求すると容量効率は小さくなり、コストも増加する。よって、要求に応じた信頼性を確保しながら容量効率を高めることが重要となる。NaryRAID は、容量効率を高める技法として有効である。

(5) VDML

VLSD を管理するために、我々は VDML (Virtual Disk Modeling Language) を提唱した。VDML は XML で表現した仮想ディスクのドメインモデルである。VDML タグは VLSD のクラスに対応する。各クラスのフィーチャーは属性および要素で表現される。

5. 主な発表論文等

[雑誌論文] (計 4 件 : すべて査読あり)

(1) Erianto Chai, Katsuyoshi Matsumoto, Minoru Uehara, Hideki Mori: "Virtual Large Scale Disk Based on PC Grid", International Journal of Scalable Computing: Practice and Experience (SCPE), Vol. 9, No. 4 (accepted)

(2) Erianto Chai, Minoru Uehara, Hideki Mori: "Case Study on the Recovery of a Virtual Large Scale Disk", Springer LNCS Volume 5186/2008 Network-Based Information Systems(NBIS2008), pp.149 -158 (2008.8.21)

(3) Kenichi Tanaka, Minoru Uehara, Hideki

Mori: "Performance Evaluation of a Grid using Windows PCs", Inderscience publisher, International Journal of Web and Grid Services, Vol.4, No.4, pp.395-417, (2008.7)

(4) Erianto Chai, Minoru Uehara, Hideki Mori, Nobuyoshi Sato: "Virtual Large-Scale Disk System for PC Room", LNCS 4658, Network-Based Information Systems, pp.476-485, (2007.9.3-4)

[学会発表](計12件:すべて査読あり)

(1) Minoru Uehara: "Combining N-ary RAID to RAID MP", In Proc. of 1st International Workshop on Information Technology for Innovative Services (ITIS2009) in conjunction with 2009 International Conference on Network-Based Information Systems (NBIS2009), pp.451-456, (2009.8.19-21), USA

(2) Kenichi Tanaka, Minoru Uehara, Makoto Murakami, Motoi Yamagiwa: "Web-Disk-Based Efficient Application Delivery Method for Grids", In Proc. of 2009 International Conference on Network-Based Information Systems (NBIS2009), pp.391-397, (2009.8.19-21), USA

(3) Erianto Chai, Minoru Uehara, Makoto Murakami, Motoi Yamagiwa: "Online Web Storage using Virtual Large-Scale Disks", In Proc. of the Third International Workshop on Engineering Complex Distributed Systems (ECDS-2009), pp.512-517, (2009.3.16-19), Japan

(4) Kenichi Tanaka, Minoru Uehara, Makoto Murakami, Motoi Yamagiwa: "Office Grid

based on Windows PCs", In Proc. of the Third International Workshop on P2P, Parallel, Grid and Internet Computing (3PGIC-2009), pp.427-432, 2009.3.16-19, Japan

(5) Kenichi Tanaka, Minoru Uehara, Hideki Mori: "Building a Linux Grid on a Virtual Machine using a Windows Grid", Springer LNCS Volume 5186/2008 Network-Based Information Systems (NBIS2008), pp.132-141 (2008.9.1-5), Italy

(6) Minoru Uehara: "Security Framework in a Virtual Large-Scale Disk System", In Proc. of IEEE 10th International Workshop on Multimedia Network Systems and Applications (MNSA2008), pp.30-35, (2008.6.20), China

(7) Katsuyoshi Matsumoto, Minoru Uehara: "N-ary RAID: 3-resilient RAID based on an N-ary number", In Proceedings of 23rd International Conference on Advanced Information Networking and Applications (AINA2009), pp.249-255, (2008.5.26), UK

(8) Erianto Chai, Minoru Uehara, Hideki Mori: "Evaluating Performance and Fault Tolerance in a Virtual Large-Scale Disk", In Proceedings of 22nd International Conference on Advanced Information Networking and Applications (AINA2008), pp.926-933, (2008.3.28), Japan

(9) Kenichi Tanaka, Minoru Uehara, Hideki Mori: "A Case Study on Linux Grid on Windows using Virtual Machine", In Proceedings of 4th International Symposium on Frontiers in Networking with Applications (AINA2008), pp.195-200, (2008.3.25), Japan

(10)Kenichi Tanaka, Minoru Uehara, Hideki Mori: "Parallel Computing of CG using an Open Source Windows Grid", In Proceedings of 4th International Symposium on Frontiers in Networking with Applications (AINA2008), pp.173 -180, (2008.3.25), Japan

(11)Minoru Uehara: "A Toolkit for Virtual Large Scale Storage in a Learning Environment", In Proc. of 21th International Conference on Advanced Information Networking and Applications Workshops/Symposia 2007, Vol. 1, pp.888 -893, (2007.5.23), Canada

(12)Erianto Chai, Minoru Uehara, Hideki Mori: "A Case Study on Large Scale Disk System concatenating Free Space", In Proceedings on IEEE 2nd International Conference on Innovative Computing, Information and Control(ICICIC2007) (2007.9.5 -7), Japan

[図書](計 0 件)

[産業財産権](計 0 件)

6 . 研究組織

(1) 研究代表者

上原 稔 (UEHARA MINORU)

東洋大学・総合情報学部・教授

研究者番号 : 70256775