資料７－２
情報科学技術委員会
（第101回）H29.11.22

自己点検結果報告書（事後評価用）

課題名：　高機能高可用性情報ストレージ基盤技術の開発

１．課題の概要および目的

東日本大震災では、ストレージサーバの損壊・流失によって住基情報や医療情報等が喪失し、情報ストレージ技術の耐災害性の不備が露呈した。これらの脆弱性を改善し、多数の機器損壊が起こる広域災害下においても、単に情報を保存するだけでなく実際に利用できる「可用性」を高めた情報ストレージシステムを確立して、ストレージ情報サービスを保全できる技術の実現を目指す。このために、ストレージ情報を近隣のサーバ拠点間で空間的に分散させて冗長化し、クラウド環境と連携させたストレージアーキテキチャを開発する。

さらに災害時には、データ転送速度の大幅な低下が生じ緊急データ退避が損なわれての情報喪失のおそれがある。平時においても、急激な拡大を続ける情報量の増大に対応するため転送速度の高速化が求められているところである。このためにストレージ間とネットワークのスループットの高速化を実現する。

以上の開発技術の耐災害性能力に対する総合的な検証として投薬情報システム（電子お薬手帳）の実証試験を行う。東日本大震災では、避難先で治療の際に投薬情報の有無は、適切な治療の可否を決める重要な情報であった。本研究で実現するストレージシステムの高い可用性を示す実証試験として「お薬手帳」の情報が災害等でも失われないことを実証する。この際に、従来のクラウド技術では必ずしも満たされていなかったデータベース等を活用した高度な処理を行うための高機能プログラミングフレームワークを開発してストレージの高機能化を実現する。

２．研究開発目標

最終的な研究開発目標は、広域甚大災害で生じる情報喪失を防ぎストレージシステムの高い情報可用性を実現するための技術開発である。このために、ストレージ機器が広範囲で損壊しても蓄積されている情報自体には引き続きのアクセスが可能になるよう耐災害性の優れたストレージ技術を研究開発する。このためには機器損壊でも情報を保全できるよう適切な情報バックアップが必要である。また、東日本大震災においてはインターネットを含めて広域にわたる長期間の通信障害が発生した。今後の大災害においてもこの大規模な通信障害を想定すべきであるが、その際でも単に情報を保全して喪失がないだけではなく、近距離で情報を持ち合う分散ストレージにより機器損壊でも被災直後から引き続きの情報アクセスが可能な高可用性を実現することが目標である。

また、発災直後にしばしば起こる通信路輻輳による長時間の通信の混乱を回避することも耐災害性として重要な要因である。迅速なデータ退避を行うことで、発災あるいは予測発令の直後に回線中にデータが残存し輻輳によるデータ喪失を防ぐ必要がある。このためにストレージ機器自体とネットワーク転送の両者を連携してデータ転送速度の高速化を図る。これは平時においても今後大きく普及するビッグデータ処理やクラウドにおける大容量化に応えるものであり、高速データ転送レートのストレージ基盤技術を開発する。

具体的には、以下の技術分野について研究開発を展開する。これらは次世代ストレージ技術の開発としてプロジェクト内で連携を図りながら研究開発を進め、同時に、それぞれの技術課題ごとに定量的な目標を定めて分担と責任を明確にした分業化の効率性も併せて重視する。

（1）耐災害性高可用ストレージ技術

① 被災時のストレージシステム内での情報喪失確率が最小になるよう、各ストレージ拠点の災害リスクを定量化することで情報複製を最適化する技術を開発する。このときに従来の計算手法では拠点数の増加に伴って計算量が爆発的に増加するために求解が事実上困難であった最適組合せ問題に関して、本ストレージ用途に沿った最も適切な複製先を短時間に求めるアルゴリズムを確立する。これを学術的にも興味のある成果である。以上を利用して、分散情報ストレージ基盤において効果的な情報複製によるバックアップ機能を実現するストレージ制御ソフトウェア（リスクアウェア複製）とし

て実際のストレージサーバの制御ソフトとして実装してその可用性を実現する。

② 被災後の復旧では、複数ストレージ拠点でデータを保全した後にその個々のデータを回収して残存ストレージ拠点に応急的に再構成して復元することが必要になる。このデータ復元を実現するために、対象の残存データを保持する拠点ストレージ群を抽出して応急拠点にデータ転送を実行する機能を実現する。その際にできるだけ高速でデータ復旧を行うために、ファイルを分割して複数の経路で転送するマルチルートリストア技術を開発する。さらに、その有効性を確認するため、様々な災害状況を設定可能なシミュレーションプログラムでの検証を行う。

③ 災害脆弱性を改善し被災直後にも確実に情報にアクセスできる高可用性分散ストレージ基盤技術として、ストレージ情報をサーバ拠点間で空間的に分散冗長化しクラウド環境と連携させたストレージアーキテキチャを開発して、50 %のストレージ機器が広域にわたって損壊した場合にも 90 %以上の情報に継続してアクセスできる可用性を定量目標とし上記のリスクアウェア複製とマルチルートリストア技術により実現する。仙台市内の東北大学キャンパスの学内ネットワークを用いて実際にストレージシステムネットワークを実証試験用に構築して実証試験に臨む。

（2）ストレージ装置のデータ転送の高速化

① データがネットワークノードに残っていると、災害時にシャットダウンした際にデータ喪失のおそれがある。このため、緊急情報の退避を迅速に行うことが必要であり、対応が遅れると一時に多くの情報アクセスが集中する回線の輻輳により通信できない状況に陥る。災害発生時の緊急の情報退避では限られた時間内での高速データ転送が必須である。被災して回線や装置の損壊が起こっている場合は多数の通信経路が機能しないために、残存回線にアクセスが集中する。ストレージ間のネットワーク転送を高速化し迅速に情報を流してネットワークの占有時間を短縮することが必要である。これらの要請から、ストレージとネットワークのデータ転送速度を高速化する。このデータ転送の高速化は、平時においても、引き続き急激な拡大を続けている情報量とトラフィックの増大に応えることができる。本課題では、装置とネットワークの総合的なデータ転送速度を向上させる技術開発を行う。

② 具体的には、ストレージ機器自体の高速データ転送化とネットワークスループットの両者を高速化する必要がある。ストレージ装置（HDD）の高速化には線密度の向上と複数トラックを同時再生する並列記録方式とその信号処理技術を達成する。また、分散ファイル方式によるストレージ装置の並列化により 10 Gbps の転送レートを実現する。ストレージ間ネットワークの高速化には SDN（Software Defined Network）技術をベースにして、データレート配分を最適化するよう適応的にネットワーク構成を制御してスループットを向上させる。並列ストレージと SDN ネットワークにより総合的に従来の 5 倍の 10 Gbps の定量目標とする。

（3）高機能プログラミングフレームワーク

高度なデータベース処理を行うための高機能プログラミングフレームワークを開発して、東北大学で開発中の高機能言語 SML#を用いて、急速に拡大しているクラウドデータとこれまでの大きな蓄積があるデータベース情報の両方へのシームレスなアクセスを可能にすることを目標とする。これを、高可用ストレージシステムの実証試験のための高信頼なウェブアプリ開発のためのプログラミングフレームワークを基盤として提供する。

（4）耐災害性ストレージ基盤の実証試験

耐災害性高可用ストレージシステムを実際にストレージネットワークに実装してデータの高可用性を明らかにするために実証試験を行う。東日本大震災では避難先で診療の際に投薬情報の有無は適切な治療の成否を決める重要情報であった。これを踏まえて、電子化した投薬情報システム（電子お薬手帳）を例として、仙台市規模の１００万人規模の都市被災を想定したシナリオを策定し、上記高可用性耐災害ストレージシステム上に展開してその耐災害性を実証する。この試験環境における実験用としてより実用的な投薬情報アプリに関するソフトウェア技術開発を行い、実用的な投薬情報アプリを試作する。この投薬情報システムは、広く普及したスマートフォンやタブレットを端末機器とし、本プロジェクトで開発する高可用ストレージシステム上で情報を送受するアプリとして作成して実証試験に用いるものである。

この際に、東北大学の学内ネットワークを用いて仙台市規模の広域分散系ネットワークを用意しこの環境で動作する実用的なシステムで実証試験を行った。仙台市全人口を想定した100万クライアントからの多数のリクエストを想定する高負荷アクセス条件での実験を行い、災害シナリオに基づいた50%の機器損壊条件下で目標とする90%の可用性を実証する。

3. 課題の達成状況等

（1） 研究開発目標の達成状況等について

① 研究開発目標及び研究開発計画の変更理由と対応

変更する事項	変更理由	対応
低消費電力に関する目標の変更※1	予算削減のため	より重要な高可用性等の重点課題に注力するために開発を中止した
プログラミングフレームワークの研究対象に関する計画の変更※2	開発負荷に応じた研究体制にしたため	高信頼ウェブアプリ開発環境提供に特化した

※1当初、ストレージの「50%の低消費電力化を理論的に検証」が年次計画にあったが削除。

※2「従来のクラウド技術では必ずしも満たされていなかったデータベース等を活用した高度な処理を行うための高機能プログラミングフレームワークを開発してストレージの高機能化を実現する（平成 24 年度業務計画書）」としていたところ、ストレージインタフェース側の開発負荷が重くこれを集中的に推進させる必要が生じたため、プログラミングフレームワークとしては高機能化が求められ当初目的も逸脱しない実証試験用ウェブアプリ開発基盤に注力した。

② 採択時留意事項への対応/事前評価・中間評価指摘事項への対応

事前評価指摘事項	対応
他府省等の関連施策と十分な連携を図り、適切な役割分担により効果的に研究開発を推進すること	NEDO ストレージプロジェクトや総務省耐災害ネットワークプロジェクトなどと、本課題の独自の技術耐災害ストレージ技術を中心に情報提供や連携への情報交換をしてきた。
研究開発当初より、民間事業者や実用化後の運用主体及び想定される関係府省の協力を得て研究開発を進め、実用化に向けてのニーズを常に技術開発にフィードバックすること。また、ユーザ側、システム側、ソフト開発側の三者の視点を踏まえて具体的な要求定義がなされること	投薬情報アプリの開発において宮城県薬剤師会と連携し、専門家として開発仕様への積極的な助言を得、合同での被災状況を模擬した避難所における緊急調剤活動の実証試験を実施。開発した投薬情報アプリ設計の際に、使い勝手や保存情報の選別など実証試験用仕様に多数のフィードバックを行った。
民間の研究機関等における研究開発動向の把握に努め、それらの研究開発成果の積極的な活用を図ること	プロジェクトメンバー企業である日立製作所ストレージシステム研究所との密接な共同研究により最先端のストレージ技術動向の情報を得て十分な活用が図られた。
これらの動向に応じ、研究開発内容の変更等の柔軟な対応が可能となるような研究開発体制とすること	本課題の体制はストレージ基盤、ネットワーク、ソフトウェア、の横の連携と、産学連携による基礎と応用の縦の連携の両方を有する柔軟な開発体制を構築している。開発途中での仕様変更なども対応できている。
中間評価指摘事項	**対応**
本課題では、複数拠点にストレージを分散配置する手法により情報喪失確率の最小化を実現しようとしているが、ネットワーク稼働率の低下に伴うアクセス率の低下等、実際に起こりうる様々な災害状況を設定した	ネットワーク輻輳によるスループット低下は大きな課題であるが、本検討のソフトウェア制御ネットワークは経路状態をモニターして制御するので一定の対応が可能である。また、現実的な被災状況で議論するため東日本大震災の被災記録を参照して16種類にのぼる

上でより詳細な比較を行い、当該手法の有効性を検証しつつ、実用化を見据えた取組を行う必要がある。	精密な災害シナリオを作成した。電源復旧や通信復旧等の回復状況も実際の記録を元に考慮した。
本課題の実証実験に向けて「電子お薬手帳アプリ」を設計・試作しているが、時々刻々と変化するデータへの対応やプライバシーの問題等を踏まえた更なる検討が必要である。	データの同期は非同期ではあるが、実際には薬歴情報が入力されるとその時点で整合の取れたデータ複製が行なわれるので頻度の高いデータ登録があれば実用上問題のない更新がされる。プライバシーについて機微性の高い個人を特定できる情報と公開情報でもある薬局情報とを厳密に分離して、前者はセキュリティの高い管理ストレージに入れ、後者は秘匿の必要がないので端末に入れておくように設計している。

③ 研究開発目的の達成状況について

当初目的通り、東日本大震災による情報ストレージシステムへの広域被災を通じて明らかになった従来のストレージ技術の不備を改善し、高い耐災害性を有する高可用ストレージの技術を開発した。

従来用いられてきた機器故障からデータを復元する RAID 技術や遠隔地に情報をバックアップするクラウド技術では、拠点のストレージ機器が全壊しインターネットが停止する甚大被災には不十分であった。ストレージ拠点の損壊リスクを評価する耐災害ストレージシステム技術（リスクアウェア複製）と、応急拠点を緊急に復旧するデータ再構成技術（リストア）も開発した。これにより広域通信途絶下でも使える地域分散方式で共倒れリスクを最小にして、目標である 50 %の機器損壊で 90 %の可用性をシミュレーションと実証試験で確認した。この実証試験は、過去の被災状況をモデル化した16種類の被災シナリオを元に、大学内の実際のネットワークに耐災害ストレージシステムを配置して 100 万人クライアントの高負荷条件を仮想マシンで展開したウェブアプリであり、その開発環境は大学の独自プログラミングフレームワークを提供している。また、このシステムにおいてもネットワーク中にデータが残存した状態では被災時に喪失してしまうため、ネットワークの輻輳を避けてデータを高速に退避できるストレージとネットワークの高速ファイル転送技術を開発した。並列転送と分散ファイルシステムによる 10 Gbps の高速データ転送を達成し、ソフトウェア制御によるネットワークの高速化と併せて従来の5倍で 10 Gbps を超える転送レートを達成した。

④ 研究開発成果および目標達成状況について

1） 耐災害性高可用ストレージ技術

1.1 概要

ストレージシステムの耐災害性を大きく高める高可用ストレージシステムを東北大学と日立製作所で開発した。本課題におけるストレージの耐災害性は、ストレージ機器が大規模に損壊しても、大半の情報が残存機器に保全されるかバックアップ拠点で守られているかによって被災直後から情報にアクセスできる高い可用性を目標とする。本課題では、図 1-1 に示す安全な情報バックアップを行うリスクアウェア複製技術を開発した。甚大災害ではインターネットも被災して数週間にわたって利用できないことが起こるので、クラウドなどの遠隔地バックアップは前提とせず、市内か県内などの近隣にあって出向いて情報を入手できる近隣配置とする。一方で、近隣であるため同一の災害で共倒れになることを防ぎ、最小限のバックアップ容量でのデータ保護が目標である。個々のストレージ拠点の災害に対するリスクを配慮して最も安全な場所にあるサーバにデータ複製を行う点に独自性がある。これにより近隣バックアップでの懸念である同一災害による共倒れを回避する。一方で、この複製先の選択組合せは拠点数の増加とともに計算時間の爆発的拡大を招く。これを現実的な時間で行うアルゴリズムも開発した。また、図 1-2 に示すように、発災後に保全した残存データと残存機器とを再構成して応急代替拠点でデータ提供を可能にするマルチルートリストア技術も併せて実現した。

図 1-1 リスクアウェア複製の概念的説明

図 1-2 マルチルートリストア方式の概念的説明

この結果、半数のストレージ機器損壊でも 90 %の情報を保全でき、近距離バックアップの利点として、通信途絶下でも出向いて 3 時間以内にデータにアクセスできる可用性も目標通り達成した。以下、主要な成果を抜粋して説明する。

1.2 リスクアウェア複製アルゴリズムの原理設計（東北大、H24年度）

可用性向上のため、データを安全な場所に複製するという第一の観点と、生成されたデータを速く複製するという第二の観点の両方を実現した。第一の観点では、拠点ストレージの被災リスクを定量化し、そのリスクに基づいて最適な複製先を選択する。これがリスクアウェア複製の提案内容そのものであり、リスクをどのように定量化するか、そのリスク値に基づいた最適化問題の目的関数をどのように設定するかを研究課題として明らかにした。第二の観点であるデータの高速複製については、設計項目として「複製タイミング」と「複製順序」の二点が挙げられる。検討結果をまとめると、非同期複製方式で、複製頻度を高くして、ファイル単位もしくはチャンク単位の反復方式で複製を行うのが望ましい設計方針であることが明らかになった。以上の高可用性システムの原理設計を設定した。種々の災害ごとのリスクを考慮して複製先を選択するリスク計算アルゴリズムの可用性向上効果を検証し地震／津波、それぞれが発生した際の各拠点の被災状態状況を作りだし、データ生存割合を小規模なシステム条件で一次評価を行った結果、地震、津波どちらの場合においても、平均のデータ生存割合が目標の 90%を超える見通しを得た。

複製先拠点ストレージを決定するためのリスク計算アルゴリズムを検討した。まず、災害に対するリスクを定量的に表現すれば、リスクの小さい複製先拠点ストレージを探索する問題は数理計画問題として表現できる。さらに、数理計画問題を解くアルゴリズムとして3つのアルゴリズムを考案し最も優れた方法を選択した。

データを 1 つだけもつ n ヶ所の拠点ストレージが地理的に分散して配置されたフィールドを考え、災害に対するリスクとして、そのうち 2 つの拠点ストレージ i,j がある災害に同時に被災する確率 P_{ij} を定義する。拠点ストレージ i から拠点ストレージ j へデータを複製する場合 $\chi_j = 1$、拠点ストレージ

i から拠点ストレージ j へデータを複製しない場合を $\chi_{ij} = 0$ とる整数 χ_{ij}を使って、損壊するデータ量の見込み fは以下のように表せる。

$$f\left(x_{12}, \ldots, x_{n(n-1)}\right) = \sum_{i=1, i\neq j}^{n} \sum_{j=1}^{n} P_{ij} x_{ij}$$

リスクの小さい拠点ストレージを探索する方式開発には、この見込みの数 f をできるだけ小さくする解を求める問題と考えることができ、f を目的関数とした以下の整数計画問題で表せる。

$$\min f\left(x_{12}, \ldots, x_{n(n-1)}\right)$$

なお、ある災害で同時に被災する確率 P_{ij}は、地震の場合は拠点ストレージ i,j 間の距離が大きいほど小さくなるものとして、津波の場合は各拠点ストレージの浸水深さが大きいほど大きくなるものとしてモデル化を行ったものを用いた。さらに、この問題に、ストレージシステムのデータ複製にかかる特有の条件として、冗長度制約と容量制約を付加する。

この整数計画問題を解くために、分枝限定法（総当たり法）、Greedy 法（良い組合せから選択）、逆 Greedy 法（悪い組合せを先に除外）のアルゴリズムを用いた。分枝限定法、Greedy 法は既存アルゴリズム、逆 Greedy 法は報告者らによる独自の提案アルゴリズムである。リスクの高い拠点ストレージを複製先にしないことを目指し、問題を単純な部分問題に分けた後、部分問題毎に悪い解を順番に除いていくことで解が得られるので、耐災害高可用性のためのストレージ方式を開発できる。開発した災害シミュレーションを用いて被災状況をモデル化することで、検討したアルゴリズムの可用性効果を検証した。地震と津波に関する災害シミュレーションツールを開発し、このシミュレーションによって可用性とリスク計算に要する計算時間を明らかにした。

災害シミュレーションは、既存の災害研究成果を活用し、拠点ストレージの損壊を判定する。地震シミュレーションと津波シミュレーションについて、いずれにも適用できるツールを開発した。地震シミュレーションは、震源場所とマグニチュードが与えられると、まず、地震による地表面速度の距離減衰を用いて各拠点ストレージにおける地表面速度を求める。次に、ある地表面速度に対して損壊が正規分布に従う確率で起こる仮定を用いて各拠点ストレージの損壊を判定する。津波シミュレーションは、地震の震源場所とマグニチュードが与えられると、まず、マグニチュードに基づく津波高の予測モデルに基づき、与えられた地震に対する津波の、海岸到達時の高さを計算する。次に、津波を 1 周期の三角波と仮定し、海岸の堤防高さを越えて陸に流入する水量を求め、陸上の浸水域および各拠点ストレージの浸水深を計算する。最後に、被害関数による予測方法に基づき、ある浸水深に対して損壊は正規分布に従う確率で起こると仮定して、各拠点ストレージの損壊を判定した。

これらの災害シミュレーションを用いて可用性評価を定量的に行った。ここでは地震災害のみを示すが、各拠点ストレージは、データの冗長度 2、複製データ向けの空き容量 1 とした小規模なケースとした。このケースに対し全データ量に対して災害により損壊しなかった拠点ストレージのデータ量の割合を可用性として計測した。地震の場合は、拠点ストレージ群の設置場所付近で起こる直下型地震と、遠く離れた地点でおこる海溝型地震の 2 種類で行った。評価は、複製先の拠点ストレージを無作為に決定するランダム法での可用性と、既存アルゴリズムや提案法により作為的に決定する場合の可用性とを比較して行う。このシミュレーションによって図 1-3 に示す地震災害に対する定量的な可用性を得た。地震シミュレーションでは、最適化アルゴリズムである分枝限定法が最も可用性が高く、次いで逆 Greedy 法、Greedy 法、最後にランダム法という結果であった。本ケースでは、アルゴリズムを選べば、プロジェクトの目標とする 90%の可用性を達成可能である。

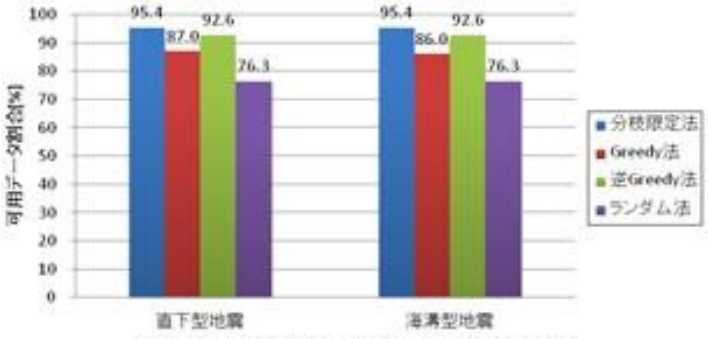


図 1-3 可用性評価結果(地震の場合)

一方、上記シミュレーションにおいて、複製先の拠点ストレージを決定するのにかかった計算時間は図 1-4 に示すように、Greedy 法、逆

Greedy 法、分枝限定法の順に計算時間が短い結果となった。分枝限定法の計算時間の拠点数に対する増加率は Greedy 法や逆 Greedy 法に比べて大きく、数千拠点以上ある大規模環境に適用する際には極めて大きな計算時間を要することが問題になると予想される。一方で、逆 Greedy 法は Greedy 法に比べてやや計算時間が大きいものの、その伸び率は同等である。可用性および計算時間の評価を合わせて考えると、リスクアウェア複製を大規模システムに適用可能になるよう、計算時間の短いアルゴリズムで目標とした 90%の可用性を達成することが今後の課題である。

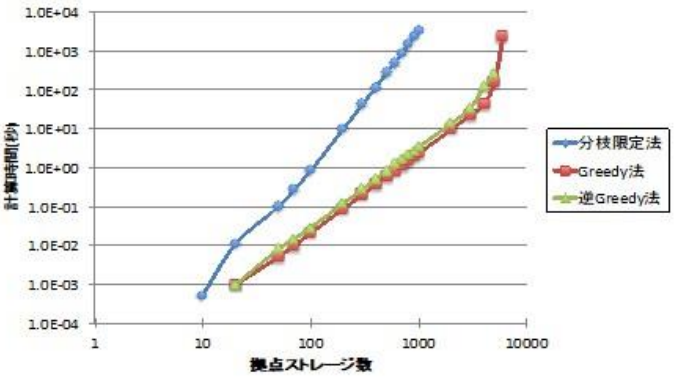


図 1-4 計算時間評価結果(地震の場合)

1.3 リスクアウェア複製の基本設計 (東北大、H25年度)

基盤システムの情報収集・複製先の決定・基盤システムへの通知を行うロケーションサーバを新たに追加することで、効率的なリスクアウェア複製方式を実現した。また、複製先決定を各回毎に段階的に行うことで、2 以上のファイル複製数に対し少ない計算時間で複製先を決定できる反復型複製先決定アルゴリズムを発案した。リスクアウェア複製機能をシステムとして提供するために、複製先を決定する機能に関するシステムアーキテクチャを検討し、その基本設計を行った。複製先を決定する機能のシステムアーキテクチャは大別すると、分散方式と集中方式がある。分散方式は各拠点ストレージが自身のデータ複製先を他の拠点ストレージに対して調停を行い決定する方式である。集中方式ではある一つのサーバ(ロケーションサーバ)が全ての拠点ストレージのデータ複製先を決定する方式である。両方式を比較検討した結果、本プロジェクトの目標の一つの指標である可用性向上効果を重視し、リスクアウェア複製システムの基本設計としてロケーションサーバを用いる集中方式が適当である。

次に、複製先決定アルゴリズムに関して述べる。リスクアウェア複製は、ある拠点ストレージのデータを他の近隣拠点ストレージに複製することを前提にしている。当初は複製数 1 でアルゴリズムを工夫して目標である 90%の可用性が満たされると判断した。しかし、その後災害パターンによっては、目標を達成するためには複製数1以上でのリスクアウェア複製の適用が望ましいこともあった。複製数が大きいと複製元拠点と複製先拠点の組み合わせ数が急激に増加するので、計算時間の大幅な増大を抑えることが課題となる。そこで、複製先決定を段階的に行うことで計算時間の増大を抑える方式を提案した。具体的には複製の序数毎に問題を分割する反復型複製先決定アルゴリズムである。提案アルゴリズムでは、第一の複製先を決定した後で、各拠点のデータ消失リスクを再計算する。そして再計算したデータ消失リスクに基づき、第二の複製先を決定する。第三以降もこれを繰り返す。このような処理手順にすることで、一括で複製先決定を行うアルゴリズム(一括型複製先決定アルゴリズム)に対して提案方式の反復型複製先決定アルゴリズムによって可用性の低下と計算時間の増大を抑えることができた。この検討結果により、複製数 1 の計算時間の複製数倍(例えば、複製数 2 の場合は 2 倍)の時間で複数の複製先を決定可能となる見込みを示した。

まず、複製元拠点と第一複製先候補拠点の同時被災リスク値を全ての組み合わせで算出する。これは第一複製先候補拠点に複製した場合のデータ消失リスク値を計算していることになる。次にそのリスク値に基づき単一複製先の決定アルゴリズムを用いて第一の複製先を決定する。次複製数が2以上では第二以降の複製先を決定する。この時点では第一複製先は決定しているので、複製元、第一複製先、第二複製先候補の3つの拠点の同時被災リスク値を全ての組み合わせで算出する。これは第二複製先候補拠点に複製した場合のデータ消失リスク値を計算していることになる。次にそのリスク値に基づき、単一複製先の決定アルゴリズムを用いて第二の複製先を決定する。

続いて、提案方式の反復型複製先決定アルゴリズムの具体例として単一複製先の決定アルゴリズムの一例である逆 Greedy 法の考え方をさらにブラッシュアップした Discreet 法を開発した。Discreet 法は第一の設計方針として過剰に安全または過剰に不安全な複製関係を作らないことであり、第二の設計方針は操作回数を有限にすることである。

1.4 反復型複製先決定アルゴリズムを用いた可用性評価（日立製作所、H25年度）

反復型複製先決定アルゴリズムをストレージ制御ソフトウェアとして実装し、可用性評価シミュレータによりその可用性向上効果と計算時間の評価を行った。可用性評価シミュレータの動作概要を図 1-5 に示す。

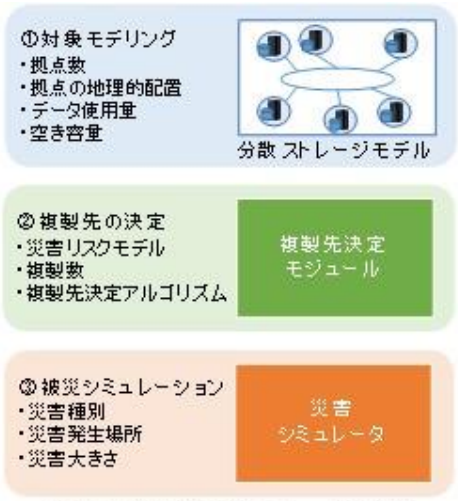


図 1-5 可用性評価シミュレータの動作

分散ストレージモデルは、システムを構成する拠点数、緯度・経度・高度で表される各拠点ストレージの地理的配置、各拠点ストレージにおいて利用者が格納したデータ使用量、各拠点ストレージにおいて複製データを格納するために用いる空き容量を用いて、対象システムの計算モデルを生成する。複製先決定モジュールは、地震や津波に対して拠点ストレージの安全性を測るための災害リスクモデル、各拠点ストレージが複製先とする拠点ストレージを意味する複製数、災害リスクと複製数に基づいて複製先にすべき拠点ストレージを決定する複製先決定アルゴリズムを用いて、各拠点ストレージの複製先を決定する。

表 1-1 各方式で達成できる可用性

#	測定条件		可用性
	複製数	複製先決定アルゴリズム	[%]
1	1	分枝限定法	90.23
2		Discreet 法	88.99
3	2	反復型分枝限定法	98.68
4		反復型 Discreet 法	98.42

災害リスクモデルは地震および津波に対するそれぞれのモデルが実装されており、シミュレータ利用者はいずれかを選択する。複製数は、シミュレータ利用者が与える。複製先決定アルゴリズムは、基本的な汎用アルゴリズムである分枝限定法、システム向けに開発した近似解法である Discreet 法などが実装されており、シミュレータ利用者はいずれかを選択する。それに加えて、シミュレータ利用者は、反復型複製先決定アルゴリズムを利用するか否かを選択する。

開発した可用性評価シミュレータを用いて、本年度の目標である 1000 拠点の大規模システムに対してリスクアウェア複製による可用性向上効果を検討した。可用性評価シミュレーションによる測定結果を表 1-1 に示す。この条件では、3 つのケースで目標とする 90%を達成した。複製数 2 の場合は、分枝限定法でも Discreet 法でも目標を達成できる。複製数 2 の場合は、ランダムに複製先を決定した場合の可用性は 87%程度であるため、数ポイントの可用性向上で目標達成が可能な、比較的達成容易な条件である。この条件では、いずれのアルゴリズムも 11 ポイント程度改善することで、目標を大きく超えて達成していることがわかる。一方、複製数 1 の場合は、ランダムに複製先を決定した場合の可用性は 75%程度である。この条件では、分枝限定法は 15 ポイント以上改善して目標を達成したものの、Discreet 法は 13 ポイント程度であったため、目標を達成しなかった。分枝限定法の可用性が高い理由は、分枝限定法が最適解を得られるアルゴリズムだからである。一方で、最適解を得るアルゴリズムは一般的に計算に時間がかかるため、拠点数が多いと計算時間が長く実用に耐えない問題がある。そこで、以降、分枝限定法に関する計算時間の評価を行った。

複製先決定にかかる計算時間の評価

反復複製先決定アルゴリズムによる複製先決定にかかる時間を計算し結果を図 1-6 に示す。同図から、複製数 2 一括の場合の計算時間は明らかに大きいことがわかる。これは、1 拠点あたり 2 拠点を選ぶ組み合わせの数が、拠点数の増加に合わせて爆発的に増えていくためである。

一方で、複製数 2 反復の場合の計算時間の増加は、複製数 1 と同程度であることがわかる。拠点数が増加しても計算時間は 2 倍程度に抑えられていることがわかる。すなわち、反復型複製先決定アルゴリズムは大規模な分散ストレージシステムに対しても実用的であることがわかった。

1.5 リスクアウェア複製への平均複製数アルゴリズムの導入と可用性評価（東北大、H26年度）

リスクアウェア複製についてシミュレータを用いて測定分析を行い、可用性に関してさらなる改善を図った。複製数を多くできない場合などのより可用性を確保するのに厳しい条件であっても、複製先決定アルゴリズムの改良や想定災害の限定などにより目標である 90%の可用性を実現手法を開発した。

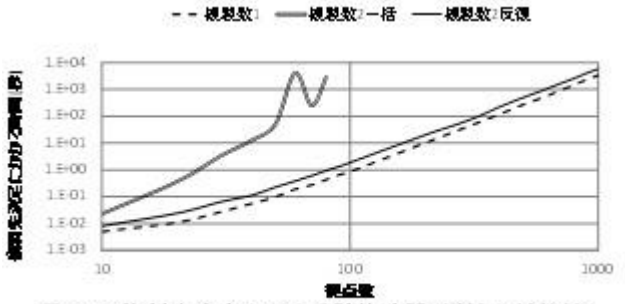


図 1-6 複製先決定にかかる計算時間（縦軸は対数軸）

複製先決定アルゴリズムの改良として、リスクベース反復複製方式を提案し、シミュレータを用いて可用性の評価を行った。リスクアウェア複製の従来の反復複製方式では、複製先を決定する際の整数計画問題の制約条件として、基本的には各拠点の複製数を固定値として設定していた。例えば、複製数 1、2、3 といった固定値を設定していた。従来の方法では、すでに安全な拠点ペアに対しても別の複製を強制的に行うことになるため、場合によっては過剰な複製になることがあった。そこで、図 1-7 に示すように津波リスクに対する山間部バックアップのように、適度に安全と判断された拠点ペアに対しては、さらなる複製を不要とする。例えば、第一複製を行った時点であらかじめ設定した安全度の閾値をクリアしていれば、第二複製は実施しない。これにより複製に伴うストレージコストの低減とデータの可用性の向上の両立が実現可能となる。

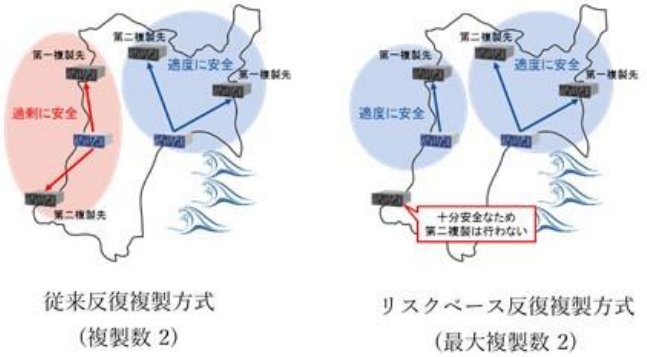


図 1-7 津波災害を想定した場合のリスクベース反復複製方式の概念

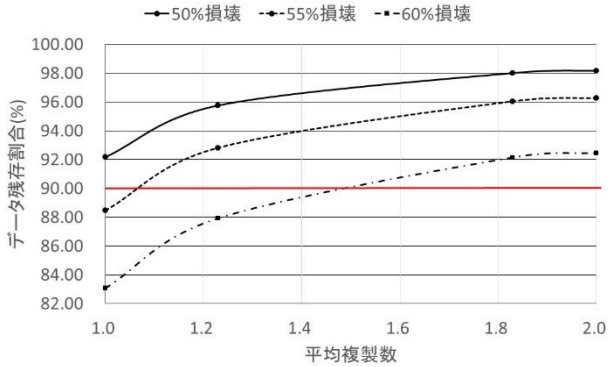


図 1-8 リスクベース反復複製方式の効果

図 1-8 にリスクベース反復方式の可用性の効果を示す。横軸は平均複製数、縦軸はデータ残存割合(可用性)である。ここで平均複製数とは、エリア全体で1拠点あたり何個の複製を持っているかを示しており、小さいほどエリア全体のストレージコストを低減できる。ここでは、拠点数が135の条件において、エリア全体での平均複製数を 1 から 2 まで変化させたときのデータ残存割合を示している。災害パターンは海洋型地震である。この災害において、拠点が 50%損壊の場合は目標を満たすが 55%損壊と 60%損壊では、複製数 1 では目標の 90%を達成できていない。90%の目標を達成するためには、従来反復複製方式では、複製数を2にする必要があった。これに対し、提案のリスクベース反復複製方式を用いれば、55%損壊ケースでは平均複製数 1.1 程度、60%損壊ケースでは平均複製数 1.5 程度の最小のコスト増で目標をクリアできる。

以上、複製先決定アルゴリズムの改良や想定災害の限定などにより目標である 90%の可用性を実現できるかを検証した結果、複製先決定アルゴリズムの改良の一つであるリスクベース反復複

製方式を用いることで 90%の可用性を実現できる見通しを得た。

1.6 マルチルートリストア機能とその評価シミュレーション（東北大、H27年度）

大規模災害によりその被災サイトに設置している情報機器が損壊してサービスが停止してしまった場合、速やかにそのサービスを再開する必要がある。このために、被災サイトもしくはその近隣サイト(以降、復旧サイト)に代替機器を設置し、その代替機器に対してバックアップデータをリストアする。この時、被災直後であるため、一部のバックアップデータ保持サイト(以降、バックアップサイト)も被災し、地域ネットワークも混雑、しかも、医療画像情報などバックアップデータのサイズが大きい緊急ファイルが流れる、など様々な懸念が考えられる。したがってこのような状況においても、リストア時間を短縮することが求められる。マルチルートリストアは上記課題に対応するリストア方式であり、具体的には残存する複数のバックアップサイトの機器から、復旧サイトの代替機器に並列でデータのリストアを行うことで高速化を図る。

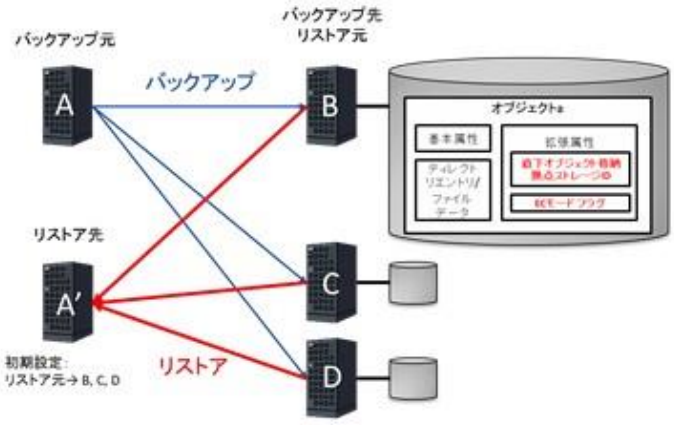


図 1-9 マルチルートリストアに対応したオブジェクトの構造

マルチルートリストア方式は下記のように実現される。まず、リストア対象のデータを保持する拠点ストレージ群を抽出する機能として、管理者がリストア対象のデータセットのトップディレクトリのデータ位置で一括して把握できるようにする。抽出した拠点ストレージ群の中でリストアデータの分担を決定する機能として、データを複数のチャンクに分割し、データチャンクと同サイズの複数のパリティチャンクを生成する Erasure Coding と呼ばれる消失データ補完方式を用いる。Erasure Coding を用いた場合、当該データが喪失してもパリティ訂正チャンクを用いて誤り訂正技術により、データチャンク数 m とパリティチャンク数 n の合計 m+n のチャンクのうち、消失したチャンクの数が n 以下であれば元のデータを復元できる。なお、リストアを実行する機能として、ファイルデータの付帯情報であるメタデータのリストアをまとめて行った後、ユーザーデータのリストアをまとめて行う方式が望ましい。これにより、バックアップサイトの数が徐々に回復していくケースで、高速なリストアが実現できる。このようなシナリオは東日本大震災発災直後の停電率の回復実績に基づいている。図 1-9 にマルチルートリストアに対応したシステムの構造を示す。A のデータを B,C,D に分割してバックアップし、被災後に残存した A' サーバにデータを集約させる。その際に、Erasure Coding の適用有無のフラグを格納する。

シミュレーションプログラムの基本設計として、前提とする条件と評価する方式検討を行った。まず、シミュレーションで前提とする条件についてのべる。災害後のネットワークは、大量のトラフィックが発生し、輻輳などにより応答性能が低下している可能性が高い。また災害直後は、停電などにより多数の拠点ストレージが一時的に停止し、その後電力の復旧に伴い拠点ストレージの稼働率が回復していくことが考えられるので、応答性能の低下と徐々に回復する拠点ストレージの稼働率を前提条件とした。

1.7 マルチルートリストアの評価（東北大、H28年度）

スループットの向上についてシミュレーションにより性能評価を行った。ファイルサイズ分布については、ファイルサーバのファイルサイズ分布として一般的によく知られる対数正規分布を用いた。また、平均ファイルサイズについては平均 80 KB と平均 8 MB の二つで行った。ファイルサイズ平均 80 KB の結果についてそれぞれ図 1-10 に示す。縦軸はリストアスループットであり、上ほど短時間にリストアが完了することを意味する。横軸はオリジナルデータに対するバックアップ容量比で、

右ほど容量が大きく高コストなシステムになる。グラフ中の表記で、kR の R は複製を意味し k がその複製数である。mDnP とは Erasure coding で、m がデータチャンク数、n がパリティチャンク数である。X:Y の表記は X がメタデータのバックアップ方法、Y がユーザデータのバックアップ方法を示す。例えば、8R:2D6P という表記はメタデータには 8 複製が適用され、ユーザデータにはイレージャーコーディングが適用され、そのデータチャンク数が 2、パリティチャンク数が 6 を示す。

従来方式である 1R:1R に対して、提案方式のマルチルートリストア 8R:2D6P は約 8 倍のリストアスループットを達成することが確認できた。つまり時間は従来方式の約 1/8 となる。この結果はファイルサイズが 8MB でも同様で高速化が確認できた。

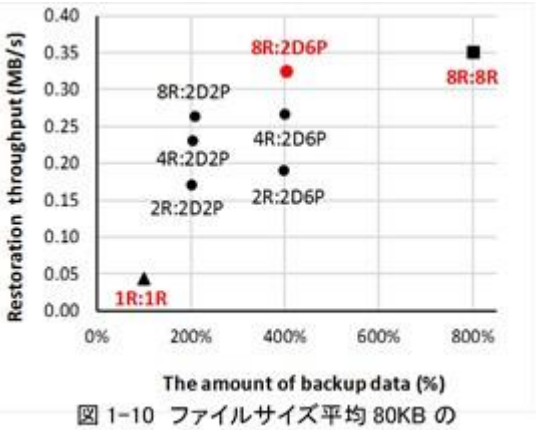


図 1-10 ファイルサイズ平均 80KB のリストアスループット

1.8 第二次実証試験システムの構築と総合可用性検証（日立製作所、H28年度）

プロジェクトでの第二次実証試験システムとして、仮想拠点を含む 108 拠点の基盤システムを分散環境で構築し、投薬システム（電子お薬手帳システム）と連動して動作することを確認した。また、第二次実証試験システムを用いて、拠点半数損壊時に 90%のデータが保全されていることと、リストア時間が従来方式と比べて短縮されていることを検証した。

1.8.1 第二次実証試験システムの構築

構築した第二次実証試験システムの構成概要を図 1-11 に示す。第二次実証試験システムは、開発したリスクアウェア複製機能とマルチルートリストア機能を、サーバ装置及びストレージ装置を用いた実システムへ導入して構築している。第二次実証試験システムには、耐災害性ストレージを提供する主装置として、ロケーションサーバと拠点ストレージがある。また、実証試験を円滑に進めるための副装置として、災害状況投入サーバがある。第二次実証試験システムは、1 台のロケーションサーバ、108 台の拠点ストレージ、1 台の災害状況投入サーバで構成される。ロケーションサーバは、本プロジェクトの主要開発機能であるリスクアウェア複製機能を搭載し、複製関係の生成および管理を行う、第二次実証試験システムの中核サーバである。第二次実証試験システムでは、108 台の拠点ストレージの情報を管理している。

また、第二次実証試験システム向けに、ロケーションサーバに、拠点ストレージの状態管理機能を搭載した。状態管理機能は、全拠点ストレージの性能低下発生の有無、サービス異常停止発生の有無、格納されたデータの健全性を、定期的にチェックする。チェックの結果、異常が検出された場合は、メールによってシステム管理者へ通知する。これにより、108 台という多数の拠点ストレージに発生した異常を早期に発見することができる。

拠点ストレージは、電子お薬手帳システムを構成するアプリケーションサーバへ、オブジェクトストレージ機能を提供するサーバである。108 台の内、7 台は物理サーバを用い、残りの 101 台は仮想マシンを用いた論理サーバを用いた。物理サーバは、星陵キャンパス、青葉山サイバーサイエンスセンタ、電気通信研究所菅沼研究室サーバルーム、電気通信研究所 IT21 センタサーバルームの 4 か所に設置した。論理サーバは、IT センタサーバルームに設置された仮想化サーバで動作するように構築した。このように、第二次実証試験システムでは、拠点ストレージを 3 キャンパス（4 サイト）へ設置した。

災害状況投入サーバは、実証試験において、拠点ストレージを疑似的に損壊させ、複製先へバックアップしたデータを活性化させる機能を提供するサーバである。災害状況投入のユーザインタフェースとして、第一次実証試験の際に実装した GUI（Graphical User Interface）を拡張し、RESTful API を追加した。RESTful API を用いると、スクリプトなどのコマンドライン処理によって、災害状況の投入を実行できる。その結果、災害状況を投入する自動化スクリプトを作成すること

で、様々な拠点ストレージの損壊パターンを容易に投入できるようになり、実証試験の効率が大幅に向上した。

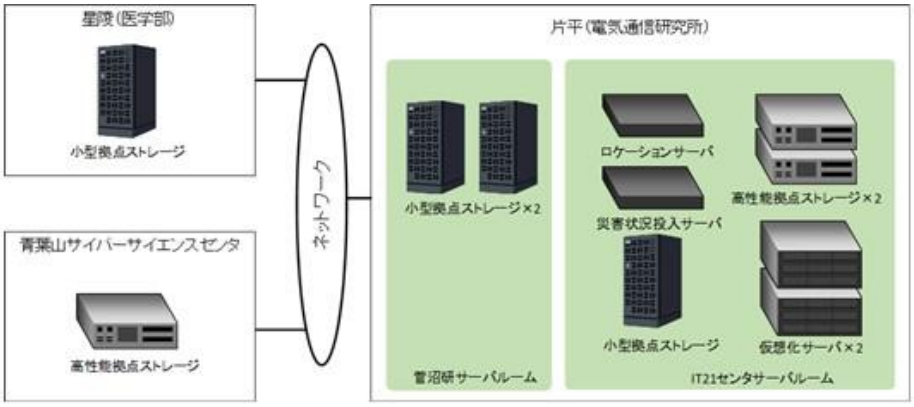


図 1-11　第二次実証試験システムの構成概要

第二次実証試験システムを構築後、規模の大きい第二次実証試験の高負荷性からいくつかの技術的な障害が発生したが、すべて実証試験開始までに対策できた。投薬情報システムと連動して動作することを最終的に確認した。3 キャンパス（4サイト）に設置された大規模な第二次実証試験システムを用いて、災害シナリオに基づく実証試験を容易に実施可能となった。さらに、大規模実証試験システムに発生した異常を早期に検知して対処できる運用管理を実現した。

1.8.2 拠点半数損壊時における 90%のデータ保全の確認

第二次実証試験システムを構成する 108 台の拠点ストレージが、半数損壊している場合における 90%のデータ保全について検証した。ただし、第二次実証試験システムを用いた 90%のデータ保全の確認は、後述する実証試験で網羅的に実施されるため、ここでの検証では、1 つの災害シナリオにおいて、90%以上のデータが保全されていることを確認した。なお、他のケースについては、シミュレーションによって 90%以上のデータ保全されていることを確認した。

検証では、災害状況投入サーバを用いて、54 台の拠点ストレージを停止させた。そして、停止した拠点ストレージのバックアップデータを持つ拠点ストレージのバックアップサービスを活性化し、そのバックアップデータへアクセスできるかを確認した。検証の結果、停止した 54 台の全バックアップデータへアクセスできた。従って、この災害シナリオにおけるデータ保全率は 100%であり、目標である 90%以上のデータ保全を確認した。

1.8.3 リストア時間短縮の確認

開発したマルチルートリストア機能によるリストア時間短縮の効果について、実環境を用いて評価した。バックアップデータを格納する拠点ストレージとリストア先の拠点ストレージは、論理サーバを用いて構築した。それらの拠点ストレージは、1Gbps で遅延 100 ミリ秒のネットワークを用いて接続している。

バックアップ方式は、Erasure Coding を用いた分割複製方式を用い、データチャンクとパリティチャンクを共に 4 とした。チャンクを格納するバックアップサーバ数は 8 台とした。バックアップリストア対象のファイルは 20GB の仮想マシンディスク（Virtual Machine Disk、VMDK）ファイルである。

リストア時間評価の結果を図 1-12 に示す。横軸は、評価項目を示し，縦軸は、リストア時間を示す。評価は、10 回測定した結果のうち、最大値と最小値を除いた 8 回分の平均値である。評価の結果，従来方式のシングルルートリストア（Single Route Restore，SRR）のリストア時間は、およそ 777 秒（27.4 MB/s）であった。一方、提案方式のマルチルートリストア（Multi Route Restore，MRR）のリストア時間は、およそ 265 秒（80.6 MB/s）であった。この結果から、MRR のリストア時間は、SRR と比べて 1/3 程度になること

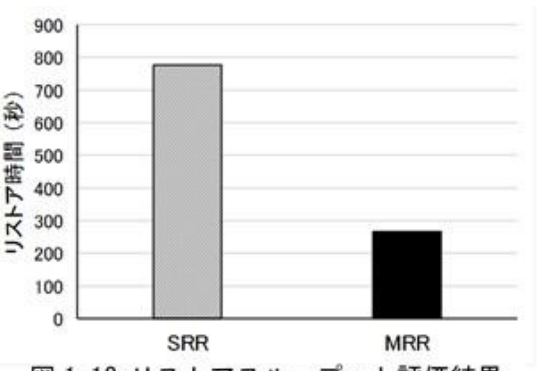


図 1-12 リストアスループット評価結果

を確認した。

以上に述べた 1.8.1～1.8.3 によって、実ネットワーク上に構築した 3 キャンパスに跨る大規模な第二次実証試験システムの構築を完了した。また、構築した第二次実証試験システムを構成する拠点ストレージを半数損壊させても、90%以上のデータが保全されていることを確認した。さらに、リストア時間が、従来方式と比較して短縮していることを検証した。

2） ストレージ装置のデータ転送の高速化

2.1 概要

情報喪失が許されない住基情報や医療情報等の重要情報は、災害時等の緊急時にはネットワーク中のデータをストレージ機器内に速やかに退避させたうえでシステムを適切にシャットダウンする必要がありこのために高速のデータ転送が必須である。一方、平時においても生成される情報量の指数関数的な急激な増加が続いており、ハードディスク装置（HDD）などのストレージ装置は大容量化され、これに応じて HDD に蓄積されるファイルやコンテンツも大容量化している。この結果、ファイル転送を迅速に行うために単位時間当たりのデータ送受信量（データ転送レート）の高速化が重要になってきている。

本課題では、ハードディスク装置（HDD）のデータ転送レートの高速化方式を検討している。最近のストレージはフラッシュメモリを用いる SSD が高速転送用に使われるようになっているが、大きな容量のファイルを扱う際には、ディスクからデータを読み出してネットワーク間を転送する時間自体が主要な転送時間の制約になる。HDD においてディスク回転やヘッドの機械的な運動によるオーバーヘッドは数ミリ秒であるのに対し、例えば、圧縮ハイビジョン映像(20 Mbps)は 1 時間の情報量でも 9 G バイトに達しそのファイル転送には 1 Gbps のデータ転送速度でも 9 秒を要する。このため、情報退避等の緊急対応でも 1 秒当たりにディスクからデータを読み書きするデータ転送レート自体が重要である。

本課題では、ストレージ機器のデータ転送レートをさらに高めるために HDD 内部で複数のトラックを並列して読み出すことでの高速データ転送化に加えて、分散ファイルシステムによる複数の HDD の並列転送化も検討し、目標通り総合的に従来の 5 倍に相当する 10 Gbps に増加させた。ストレージのシステム化に加えてネットワークの高速化によりストレージ間のデータ転送速度の高速化を達成した。

2.2 高速データ転送のための条件

HDD のデータ転送速度は 1 秒あたりに送り出す（受け取る）データ量であるが、これは十分に長い時間の操作ではヘッド下を通過するディスクの磁化ビットの総数に等しくなり、ドライブ装置を並列に使用するシステムの場合は総合的なデータ転送レートは、これに装置の数を乗じたものになる。ここで、これまでは一般的でなく本課題で取り組むトラック並列転送方式では、さらに同時に読み出すヘッド数（つまりトラック数）を乗じたものになり、

（データ転送レート）＝（ヘッド下を通過する磁化ビット数）×（同時並列ヘッド数）×（ドライブ数）

と表される。

すなわち、ストレージシステムの転送レート Xtotal は、個々の HDD の転送速度 Xdrive に HDD 並列化数 Ndrive を乗じた値である。前者の HDD データ転送速度 Xdrive は単位時間当たりにヘッドの下を通過してディスクから読み出されるデータ量であり、ディスクがヘッドの下を通過する線速度 $v=2\pi r \times rpm/60$（r と rpm はトラック半径と 1 分当たりのディスク回転数。）をディスク上の記録ビット長 b で除した値になる。この際に、複数のトラックを同時に読み出すことが可能であれば転送レートはそのトラック数 Ntrack に比例する。以上より、データ転送速度は近似的に以下の式で表現される。

$$X_{total} = N_{drive} \cdot X_{drive} = N_{drive} \cdot N_{track} \cdot \frac{\left\{ 2\pi r \frac{rpm}{60} \right\}}{b}$$

ここで、ディスク半径の増加は HDD の装置形状を大きくしてしまい、ディスク回転数の増加は消

費電力増や高速回転に伴うディスク面ブレ等のため実用上大きく増加させるのは難しい。これを考慮して、データ転送レートの向上には

1. 高線密度化による記録ビット長の短縮化
2. 同時読み出しトラック数の増加
3. 並列転送する HDD 数の増加

によるアプローチが可能である。

2.3 高線密度化

高密度磁気記録の研究からビット長をさらに短縮することでデータ転送の高速化を図った。これまでの高線記録密度化の検討から、ビット長の短縮には記録ヘッド磁界傾斜を改善して高い記録分解能を実現することが必要であることが分かっている。また、着実な磁気記録におけるノイズの改善や記録再生分解能の向上も重要である。特に、HDD の高密度記録のためにリソグラフィー等で記録層の微細加工を施してビットパターン媒体を用いて熱アシスト記録により等価的な記録磁界勾配を改善することで、現状の 5 倍程度の 5Tbit/inch2 程度の面記録密度が可能なことを明らかにしている。これにより線密度は 2 倍程度の向上が可能であり、データ転送レートもそれに応じて高速化できる。

本検討ではこれに加えて、HDD 単体でのデータ転送速度をさらに高速化するためにダウントラック方向とクロストラック方向の両方で符号化・復号化する新規の 2 次元信号処理方式により複数トラックを同時に読み出す手法を開発している。また、分散ファイルシステム(Gluster FS)を用いて一つのファイルを複数の HDD から同時並列転送するデータ転送レートの高速化も検討した。以上を通じてネットワークを含めたストレージ間転送において 10 Gbps を超える高速のデータ転送速度の可能性を検証した。

2.4 原理的ヘッドメディア配置

HDD は密に配置された同心円状のデータトラックを形成しているので、ディスクの複数トラックを同時に再生すれば単位時間当たりのビット数、すなわちデータ転送レートの向上が可能である。原理的には、図 2-1 の上部に示すように、それぞれのトラックに独立した専用ヘッドを配するアレイ型再生ヘッドを用いて両再生ヘッドの出力を加算して用いることを想定する。一方、再生ヘッドでは相反定理より一般的にトラック方向の加算性が成り立つので、同図下に示すような単一の幅広トラック幅再生ヘッドも利用できる。最近のトラック幅は 50nm 以下に微細化されているため再生ヘッドのトラック幅も狭トラックに加工する必要があり必ずしもヘッド作製が容易ではない。複数トラックを単一の幅広トラックのヘッドで読み出して信号処理によって各トラックの信号を分離復号できれば再生ヘッドのトラック加工精度を緩和できるので高い実用性が期待できる。なお、同図の幅広再生ヘッドについて、記録トラック端の再生磁束分布が再生素子内部で擾乱を起こして再生信号に歪を生じる可能性もあるが、その場合はアレイリーダーを要する。従来は困難とされてきたアレイヘッドであるが、最近トラック間クロストークを軽減する目的で複数の再生ヘッドを用いる議論が米国を中心に始まっており、本方式が実用的なものになる可能性がある。

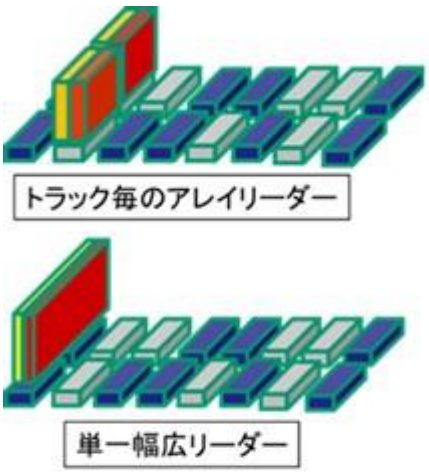


図 2-1 アレイヘッドによるマルチトラック再生(上)と単一幅広トラックヘッド再生方式(下)。橙の部品が再生素子。

これまでの HDD においてこのような検討例はなく本課題の独自方式である。複数トラックの同時読み出しは、単位時間当たりの読み出し情報量がトラック数に比例して増えるのでデータ転送の高速化を図ることができる。しかし、複数のトラックを同時に読み出す際に、互いに隣接する複数のト

ラックの再生信号を加算するため強いクロストークノイズ(トラック間干渉)を生じて通常の符号方式では復号不能に陥る。たとえば、片側トラックの記録磁化が 1 で他方が-1 の場合とその逆の場合ではいずれも並列ヘッドが読み出す信号は両磁化の和であるゼロになり区別がつかず復号不能である。このため複数トラックを同時再生する具体的な方式が提示されたことはこれまでなかった。

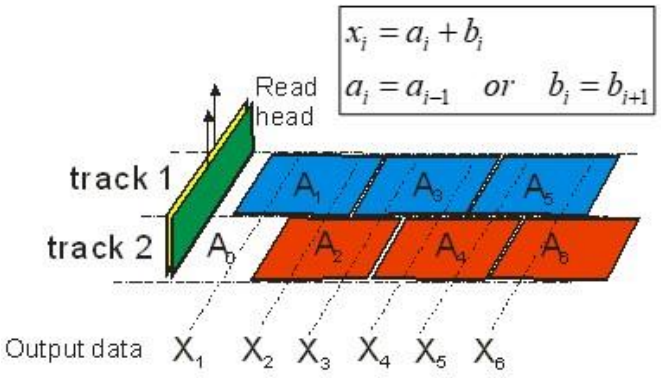


図 2-2 2 トラック同時再生におけるビット磁化配置とオーバーサンプリング検出点の位置。

また、最近提案された Shingle 記録(瓦書き記録)はディスク記録のオーバーライト記録の特徴を使って、トラック幅の広い記録ヘッドを用いながらヘッドをずらしながら次々とトラックを重ね書きすることで狭い記録トラック幅を実現できる。この記録再生方式においても狭トラック幅の再生ヘッドを同時再生トラック数だけ配置する必要があるが、狭トラック幅化によって多数のトラックを密に配置できるから複数のトラックを同時並列で読みだす本方式とは整合性が高い。この際に読出しヘッドの構成方法により 2 種類の方式について成果を得た。

2.5 再帰的復号方式の開発

2 トラック並列読み出しにおいては各トラックの記録磁化からの再生出力の和が再生応答になる。図 2-2 に示すように、記録磁化をヘッド走行方向に互いに半ビット長ずらして配置して 1 ビット内で 2 回サンプリングするオーバーサンプリングを行えば、隣接トラック間で 2 次元符号化することで多重化した再生信号からの復号が可能になる。同図で、検出点 X_1 で A_1 の先行ビットを復号した後に次のサンプリング X_2 で後続ビット A_2 を復号するのであるから、A_0 が予め決められていれば、加算信号から A_1 を知ることができる。この先行ビットの復号情報を用いて後続ビット A_2 も復号できる。このようにして、最初のサンプリングで直前ビットの情報を得て復号でき、それを用いて次のサンプリングで後続ビットを復号できる。なお、ここでは 2 トラックに限定して検討を進めているが、原理的には 3 以上の複数のトラック数に拡張できる。この多重記録再生方式はこれまでにない提案である。

入力	+1	-1	-1	+1	+1	+1	-1	-1
トラック #a	+1	+1	-1	-1	+1	+1	-1	-1
トラック #b	-1	-1	-1	+1	+1	+1	+1	-1
加算信号	0	0	-2	0	+2	+2	0	-2

復号信号＝加算信号－1ビット前復号

図 2-3 再帰的復号規則に沿った 2 トラック記録磁化の一例

入力	+1	-1	-1	+1	+1	+1	-1	-1
トラック #a	-1	-1	-1	-1	-1	-1	+1	+1
トラック #b	-1	+1	+1	-1	-1	-1	-1	-1
加算信号	-2	0	0	-2	-2	-2	0	0

復号信号: ±2:＝+1 0:＝-1

図 2-4 プリコーダにより加算信号と再生出力が対応するよう符号化した例。入力+1 に対して加算再生出力が+2 か-2 で入力-1 には 0。

図 2-3 で、トラック a とトラック b の 2 トラック並列読み出しを模式的に示している。ヘッド走行は左から右とし、ビットごとに 2 回サンプリングするので、数字 2 つ分で 1 ビットの磁化である。それぞれの 2 つのトラックを a_i と b_i とする。2 トラックを 1 つのヘッドで読み出した場合には再生信号 c_i は両者の和となる。任意の入力について加算再生信号 c_i から a_i と b_i を解くことができれば復号が可能になるが、単に 2 トラックの信号を加算しただけでは解くことができない。しかし、初期条件として例えば $a_0 = b_0 = -1$ を決め、同一トラックの信号は 2 回のサンプリングで同じである条件を入れれば解くことができる。すなわち、偶数ビットでは $a_i = a_{i-1}$ が成立し、奇数ビットでは $b_i = b_{i-1}$ となっているとすると、a_i か b_i が 1 ビット以前の信号 a_{i-1} か b_{i-1} から既知になるので上式は解くことができて復号が可能になる。同図はこれを時系列に復号する手順を図示しており、この磁化状態は2つの隣接ビットが交互に同じ値を取るので、最短の磁化転移間距離が2倍に広がって、記録磁化が互い

違いに配置されていることに相当する。ここで磁化転移間隔に対してビット間隔は半分に短縮できることである。通常、記録可能な最短磁化転移間隔が再生分解能制約を決めるが、同一のビット長に対して隣接磁化転移間の距離は 2 倍にしても良いから再生分解能制約を緩和できる可能性がある。ただし、この復号法では 1 ビット以前の情報を次ビットの復号に用いるためにビット誤りが伝搬する欠点があるが、以下のようにして解決できる。

両トラックの加算信号が直接復号値になるようにあらかじめプリコーダを用いて記録時の符号変換を施すことでエラー伝搬を防ぐことが可能である。磁化は正負の 2 値であるから、0 と 1 の入力データに対して+1 と−1 を記録磁化としその加算和が再生ヘッドで読み出される信号レベルになるためには例えば剰余演算を用いると復号が可能になる。その例を図 2-4 に図示する。最初の+1 の入力なので両トラック再生信号の加算値が+2 か−2 になるようにする。トラック 2 が−1 なのでトラック 1 には−1 を記録する。次は、入力が−1 なのでトラック a は−1 で決められているので、トラック2は加算して 0 になるように＋1をトラック b に記録すれば良い。

2.6 エラーレート特性

上述の並列再生方式について記録再生性能を検討した。記録再生分解能を考慮したヘッド再生波形をモデル化し、これを代数的に加算した信号を読み出して復号することで並列トラック再生方式における符号誤り率を計算するツールとしてシミュレーションプログラムの作成を完了している。本来、記録過程は複雑な非線形過程によって決定されるが、その記録再生分解能について、有限パルス幅で近似した孤立磁化転移波形を、その記録ビットに応じて加算した信号波形にランダム雑音を重畳した信号からクロックごとにサンプリングしてエラーレートを求めた。求めたエラーレート特性では、信号の SN 比が高い場合には正しく復号可能であり原理的に単一ヘッドによる並列同時再生方式が成立することを実証した。しかし、SN 比劣化に敏感にエラーレートが反映されることが示されており、方式によっては従来の単一トラック再生との比較では 2～3dB 程度高い SN 比が必要である。一方で、2 トラック並列での等価的な線密度で比較すると、プリコーダ復号では同一の転送レートで比較すると同じ SN 比でもより速いデータ転送レートが実現できることが示された。より現実的なエラーレート特性は次節以降で同時並列再生方式にPRML方式を組み合わせた場合に詳細に求める。

2.7 PRML方式を用いた2次元符号化とマルチトラック同時再生

複数の再生ヘッドを用いるアレイヘッド構成ではトラック数に比例した転送レートが期待できるが、複数トラック同時再生方について、1 ビットの磁化を 2 回オーバーサンプリングするためのオーバーヘッドがある。パーシャルレスポンス型の波形等化方式とビタビ最尤復号方式を導入してより現実的なＰＲＭＬ(Partial Response Maximum Likelihood)2次元符号を用いた記録再生方式を検討した。

図 2-5 は、各トラックの磁化ビット 1 つからの応答波形が重なる様子を示している。ここで実際は、複数のトラックからのそれぞれの磁化への応答を重ね合わせた応答になるものであるが、これを2次元に符号化してトラック1とトラック2が交互に現れる応答と考え、さらにこれを1トラックに縮退した密な応答に変換すれば復号が可能になる。

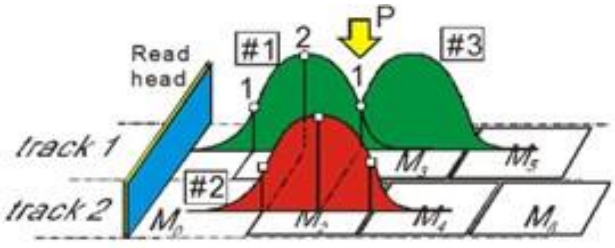


図 2-5 2 トラック同時再生の記録磁化分布と再生応答。トラック1と2に記録し単一再生ヘッドで再生する。

2.8 復号プロセス（状態推移とトレリス線図）

同図に示すように、各磁化からの再生パルスが3ビット分の広がりを持ち、検出点毎に(1,2,1)となるように等化された場合に2トラックでの各記録ビットの応答が重畳される様子を示している。各トラックでの1ビット分が(1,2,1)の応答になり、これが2トラック分の記録磁化の組合せとして位相が遅れた重畳応答になる。たとえば、ビット 1 が 2 回連続した場合にはその重畳した再生応答は

(1,2,1)が 1 検出点ずれて加算されるので、全応答は(1,3,3,1)であり、引き続き、次のビットが 1 であれば(1,3,4,3,1)である。

ここで、3 ビットに亘ってのパルス幅であるから最大のパルス重畳数は 3 パルスで、3 ビット重なっている検出点における最大振幅レベルは 4 である。もし、次のビットが−1 であればそのビット単独での再生応答は(−1, −2,−1)であるから、3 パルスの加算パルス応答は(1,3,2,−1,−1)となり検出点には 2 が現れる。応答長さが 3 であるから状態数は 4 であり、すべての状態で加算応答レベルを網羅することで状態推移が図 2-6(a)の通りに得られる。

図 2-6(b)はこれを時系列に展開したトレリス線図である。検出レベルに従う状態推移の軌跡の組み合わせからビタビアルゴリズムに従って最も確率の高い系列を選ぶことで復号系列が得られる。この PRML 方式は単一トラックの場合と同様であるが、検出点間隔が磁化ビットの半分であり、従来方式の磁化ビットと検出点間隔が等しい点が異なる。これは磁化ビットを等価的に長くできることに相当している。

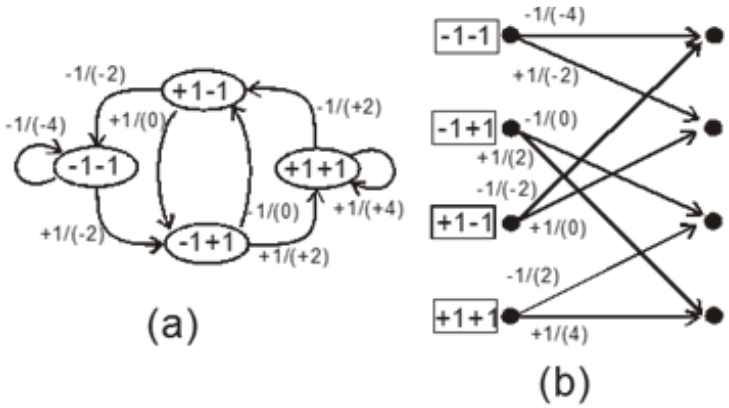


図 2-6 開発した複数トラック同時再生方式における PRML 復号の(a)状態推移図、(b)トレリス線図

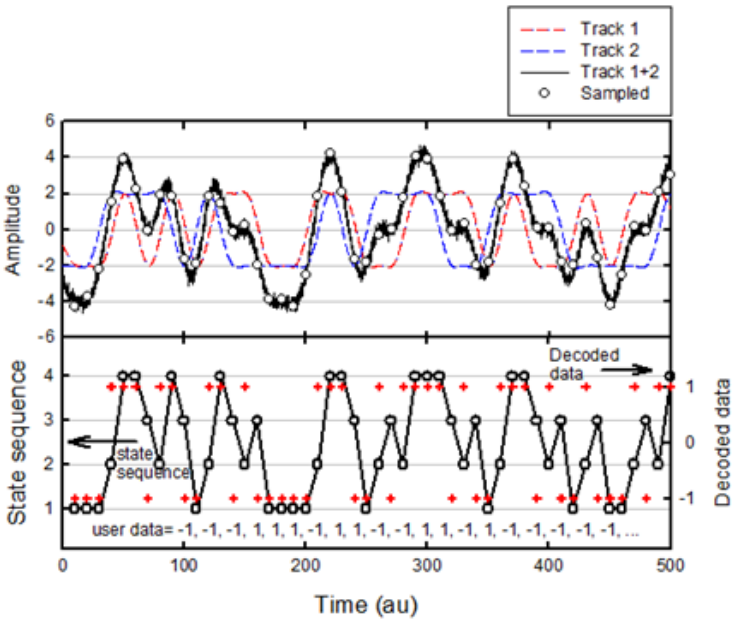


図 2-7 復号シミュレーションにおける複数ビット再生波（上）と検出された状態推移（下白丸）と復号データ（下赤クロス）

これをスピンスタンドの記録再生実験から得られている孤立磁化転移の再生波形が双曲線正接関数(tanh)で近似できるという知見を用いて本マルチトラック PRML についてのシミュレーションを行った。その複数ビットの再生波形と復号結果を、それぞれ図 2-7 の上と下に示している。上図に示すノイズを含む 2 トラック加算波形(白丸)から、下図の状態推移(白丸)を検出して復号データ(赤クロス)が正しく得られることが、同図下に示す符号入力と一致していることから確認できる。この妥当性を踏まえて、同シミュレーションを用いてより詳細なエラーレート特性の評価に進んだ。

2.9 複数トラック同時再生による高速化の記録再生方式の検証

新たな 2 次元復号方式を開発して、干渉を受けた再生信号からデータをトラックごとに分離復号する複数トラック同時再生を検討した。2 次元符号化方法を探索し復号可能な方式を確立するために、パーシャルレスポンス型の波形等化方式とビタビ最尤復号方式を導入して PRML 方式による実用的な 2 次元符号を用いた記録再生方式を提示した。これを発展させて性能の高い PRML 方式を探索した。当初の PRML 方式は各トラックの磁化ビットからの応答波形が 3 ビットのクロック点にまたがってそれぞれ(1 2 1)となるように再生応答波形を等化するものであった。より優れた高密度特性を得るには、これをより長い時系列にわたる高次のパーシャルレスポンス方式の可能性を探索する必要がある。実験によれば垂直磁気記録では再生応答は積分型であるので、図 2-

8 に示すような単峰性パルスに等化するのが素直である。また、等化目標波形のパルス幅が広いほど等化器でのノイズ強調を緩和できるので望ましい。これまで、検討したのは、4 ビット方式から PR(1 3 3 1)、5 ビット方式は PR(1 2 2 2 1)や PR(1 3 4 3 1)などである。高次になるほどビタビ復号における状態数が増えるので、信号処理系は複雑になる。ここでは、5 ビット 16 状態を上限としているが、昨今の LSI の進歩を考慮すると、将来はさらに高次パーシャルレスポンスを導入することも可能と考えられる。

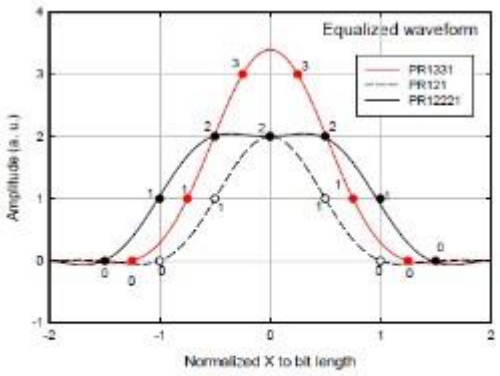


図 2-8 トラック PRML 方式におけるパーシャルレスポンス再生波形の例。3 ビット応答(121)、4 ビット応答(1331)、5 ビット応答(12221 を示す。)

2.10 エラーレートシミュレーションの実施

本並列トラック同時再生について実験からの知見を取り入れてのエラーレートシミュレーションを行った。まず、これまでスピンスタンドによる記録再生実験から孤立記録磁化転移形状が双曲線正接関数で近似できるという知見に基づいて、相反定理による再生応答を近似した。この再生ヘッド波形は目標とするパーシャルレスポンス波形等化されるが、その際に信号に含まれるノイズも強調するおそれがある。本シミュレーションではマルチトラック PRML についてこの再生波形等化を考慮して行った。この等化を実用されているトランスバーサルフィルタによる信号波形等化とディジタルフィルタによるノイズ低減方式を用いて実現した。以上のプロセスをブロックダイアグラムで図 2-9 に示す。ランダムに発生させた符号に対して相反定理により波形を生成し、その再生波形にランダムノイズを加算した上で等化器を通している。得られた波形をクロック点ごとにサンプリングしたのちにビタビ最尤復号アルゴリズムに従って復号した。ビット誤りの計数は復号された符号系列と入力符号をビットごとに比較して求めた。

図 2-9 エラーレートシミュレーションの計算フロー

2.11 長パルス幅波形を用いたマルチトラックPRMLのシミュレーション

マルチトラック PRML 方式による総合的なエラーレート特性を調べた。すでに(1 2 1)型のパーシャルレスポンス波形等化とビタビ最尤復号方式を導入する実用的な 2 次元符号を用いて記録再生方式が可能なことを示したが、さらにこれを発展させて 4 ビット以上の長い時系列にわたる性能の高い高次の PRML 方式についてその可能性を探索した。垂直磁気記録では再生応答は積分型であるのでその再生ヘッドからの出力応答は単峰性パルスになる。自然な形状を持つ 4 ビット方式から PR(1 3 3 1)、5ビット方式は PR(1 2 2 2 1)や PR(1 2 3 2 1)、PR(1 3 4 3 1)などである。高次になるほどビタビ復号における状態数が増えるので信号処理系は複雑になる。しかし、等化後の再生パルス幅を広げることができるので信号等化器での高周波域の強調を緩和することができ高線密度化には有利であ

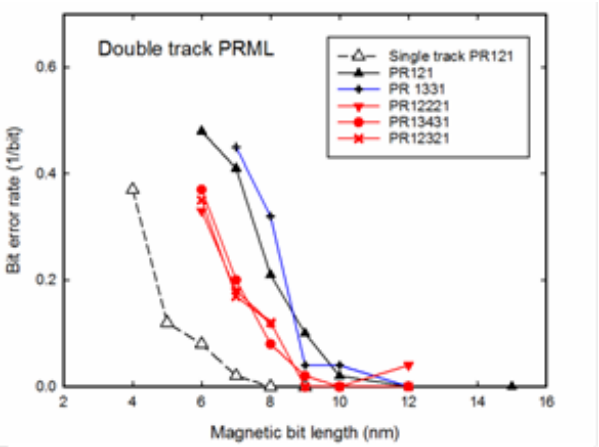


図 2-10 エラーレートシミュレーション。点線が比較用の単一トラック記録。入力信号 SN 比＝14dB

る。ここでは、5ビット16状態を上限としているが、今後の信号処理技術の進歩を考慮すると、将来はより複雑な信号処理でも実用的になることが予測されるのでさらに高次パーシャルレスポンスを導入して高密度化することも可能と考えられる。

以上のシミュレーションから、いくつかのPRML 方式に対して計算されたエラーレートの計算結果を図 2-10 に示す。図中PR121 single track がリファレンスの単トラックの場合である。同図は単トラックでのビット長なので 2 トラック並列転送では実効的にデータ転送を行う際のビット長は半分になる。これを考慮すると、エラーレートが増加し始めるビット長は 9nm 程度であり、実効的に 5 nm 弱である。さらに 4 ビット、5 ビットにすることでエラーレートの発生するビット長を短縮できる。単一トラック記録PR(1 2 1)での 7nm に比べて同一 SN 比の入力に対して等価的に短いビット長まで復号可能であることが示されている。単トラックの約 50 %増である。ここで、単トラックに対して2トラックでの転送レートは 150%程度まで改善されることが示された。隣接トラック干渉を考慮した新規 2 次元符号方式を開発して、干渉を受けた再生信号からデータをトラックごとに分離復号可能な複数トラック同時再生を確立した。

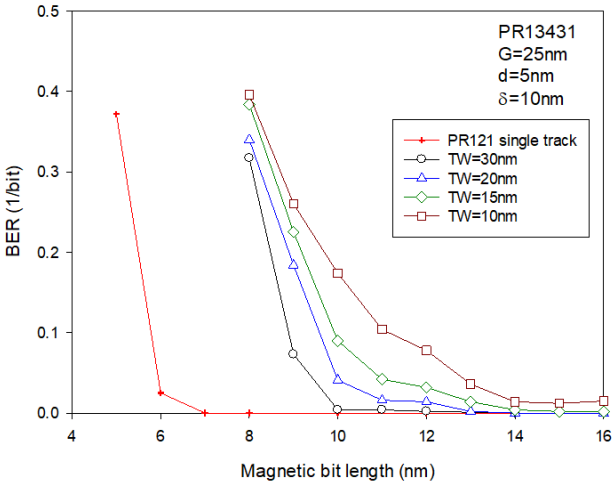


図 2-11 ビット長に対するエラーレート。PR121 が比較用の単一トラック記録(トラック幅 30nm)。複数トラック記録では夫々のトラックのビット長。

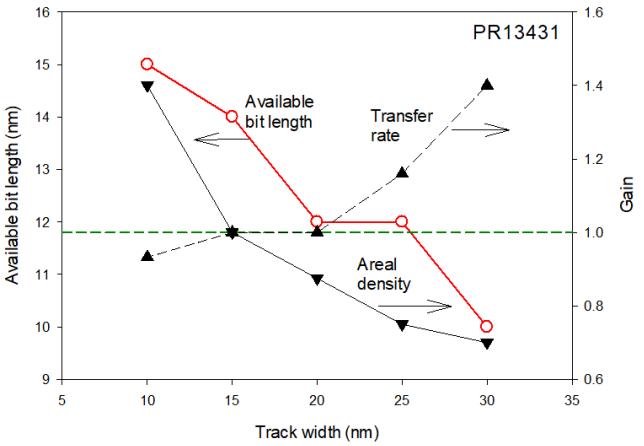


図 2-12 2トラック並列記録時のエラーのない記録ビット長(Available bit length)のトラック幅依存性(左縦軸)右縦軸は、等価 2 トラック分転送レート(Transfer rate)と面記録密度(Areal density))の単トラック記録に比べた増加分

一方で、図 2-11 は 5 ビットの PR(1 3 4 3 1)に固定し、トラック幅を変化させたときのビット長依存性である。トラック幅の低下とともに面密度は向上するが、信号の SN 比が劣化するので誤りのないビット長は大きくなる。例えば、トラック幅 30 nm の場合にはエラーの増加はビット長 10 nm 付近で始まるが、等価的に 2 トラックでは合計で 5 nm のビット長での転送が可能なことを示しており、単トラックの 7 nm に比べて短ビット長で高速データ転送が可能なことが示されている。また、この際に単一トラックでは 30 nm のトラック幅で 7 nm のビット長が使えているのに対して 2 トラックでは同じ 30 nm のトラック幅でエラーのないビット長は 10 nm まで増加するので面記録密度は 70 %に低下する。一方で、トラック幅を 10 nm に狭くした場合は可能なビット長が 15nm とほぼ倍増であってトラック幅は 3 分の一であるから面記録密度を大きくできる。

以上をトラック幅に対してまとめると図 2-12 を得る。図中 2 トラック並列記録でエラーなく記録可能な各トラックのビット長(Available bit length)はトラック幅 30 nm のときに 10 nm であり、その後トラック幅を狭めると増加してトラック幅 10 nm で約 15 nm になる。このビット長から 2 トラック分の転送レートは▲で Gain(右縦軸)として示すように、トラック幅を広げると単トラック記録に比べて 140%に改善される。この際に面記録密度(Areal density)は低下するが、トラック幅を狭めて転送レートの低下を許容すれば面記録密度を大きくできることが示されている。

2.12 分散ファイル方式によるHDD並列化による高速化

HDD 内部の高速化に加えて、並列分散ファイルシステムによって HDD 群を並列的に用いて転送速度の増加を図った。HDD の情報はファイルとして管理されることが多いのでデータ転送速度の向上はシステム的にはファイル転送速度の向上と考えてよい。ここで、一つのファイルを複数のチャンクに分割することができれば、これらを複数の HDD に収容し、データの転送の際には複数の HDD を並列運転してデータ転送速度を向上させることが可能である。実際すでに分散ファイルシステムと呼ばれるものが開発されており並列転送も実現されている。ここでは、ストレージの高速データ転送の立場から並列数と高速化の効果を検討した。

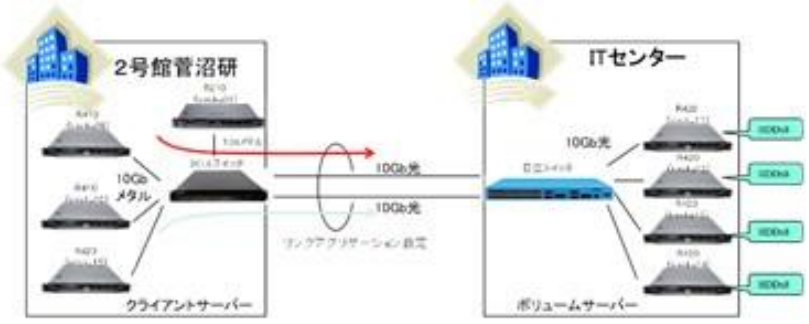


図 2-13 分散ファイルシステムを用いた HDD 並列運転によるファイル転送速度の測定系

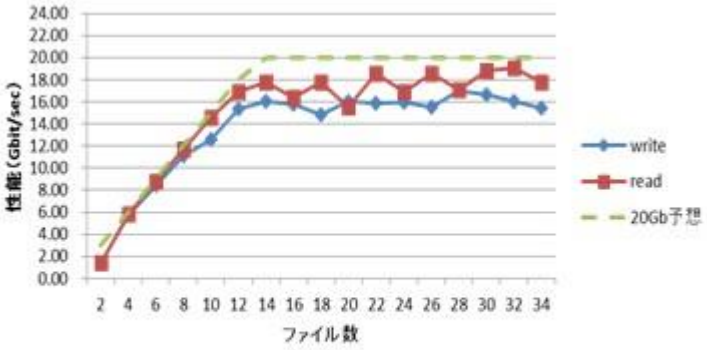


図 2-14 分散ファイルシステムを用いた HDD 並列運転での並列 HDD 数(多重度)に対するファイル転送速度の測定結果

ここでは Gluster FS を用いて分散ファイルシステムを導入して実験を行った。図 2-13 の測定系の構成図に示すように、ストレージサーバ群を 2 台のネットワークスイッチを介して 10 Gbps 光ファイバ 2 本で接続して HDD の同時並列数に対する総合的な転送速度を測定した。各ストレージサーバは複数の HDD を搭載しておりこれを多重化して HDD の並列数に対するデータ転送性能を測定した。図 2-14 にはその測定結果を示す。1 台の HDD のデータ転送速度は大体 1.5 Gbps 程度であり、同時並列転送の HDD 台数(多重度)に比例した転送速度の向上効果が確認され、8 台以上で 10 Gbps を越えることが分かった。HDD 並列数が 12 台程度で 14 Gbps に達してそれ以上ではネットワークの上限 20 Gbps に漸近した。ここでは 20 Gbps より少し小さい値で総合的な転送速度が飽和しているが、これはネットワークのデータ送信速度が 10 Gbps の光ファイバ 2 本分で 20 Gbps が上限となっているためであり、この制限を取り除けば HDD の並列度に比例したデータ転送速度が実現される。すなわち、HDD 多重の効果は大きく、現状でも 10 Gbps を越えることが可能である。

また、前述の短ビット長化とマルチトラック化によって所要 HDD 台数は更に低下可能である。分散ファイルシステムによる高速化と HDD 単体の高速化はそれぞれ独立に働く相乗的なものであるからその積でデータ転送速度が向上する。個々の HDD の転送速度が現状の 3 倍程度にできればHDD 多重度も2 台強で 10 Gbps にできる。この際にストレージ間通信はネットワーク越しに行われ、通常の使用ではネットワークのスループットが制約になることが多い。HDD 自体の高速化を実使用状態で効果を上げるにはネットワークとの相乗効果が重要である。

3) スマート分散ネットワーク技術

3.1 概要

本研究では、高機能高可用性ストレージ基盤技術の一部として、ネットワーク上に分散して配置されたストレージ間の通信を効率化・高速化することで、ストレージの高機能性、高可用性の向上

を支えるネットワーク基盤技術を研究開発することを目的とする。具体的には、冗長データ分割機能によりファイルを細分化して複製し、それらを小さなチャンク単位で並列転送するファイル転送の高速化に加え、その際にデータ転送経路が隘路とならないように、ネットワークの状態を動的に把握して、ソフトウェアで適切に経路を設定できる「高速 SDN 型ストレージ間通信機能(スマートルーティング)」を研究開発する。本機能により、主に高可用性を実現する際に重要となるレプリケーション等のネットワークを用いた遠隔ファイル転送時に、高速な回線の選択的利用や経路の多重化などによってネットワークレベルでの転送効率を向上することが可能となる。以上から、ネットワーク基盤の高度化を通じて本プロジェクトの成果目標の達成に資するものである。

本機能により、レプリケーションやバックアップ等で行われる大容量ファイル転送の高度化が実現できる。具体的には、平常時においては、ネットワーク状態を考慮したファイル転送スケジューリングが可能となり、ネットワーク全体の効率を向上しつつ、所望のファイル転送タスクの高速化が実現できる。また、地震等の災害時に有用な、極めて短くかつ厳密な時間要求内に所望のデータを複製できる 10 Gbps の転送高速化を実現できた。例えば、停電時のバッテリ運用の際、バッテリの枯渇でシステムの一部が停止する時間制約の条件下、停止前にバックアップを完了することなどが可能となる。

3.2 ファイル転送の高速化の検討

図 3-1 に高速 SDN 型ストレージ間通信機能(スマートルーティング)の概要を示す。本機能では、これまで静的・固定的に割り当ててきたストレージ間通信のためのネットワーク経路について、回線の利用状況を動的に把握し、優先度に応じて高速に転送できる経路を多重化を含めて高速かつ動的に設定することを可能とする。本機能の実現には、回線利用状況のリアルタイム観測と、それに基づく高速な転送経路の切換えが必須である。このため、柔軟にネットワーク構成を変更できるネットワーク基盤技術として、ソフトウェアによりネットワークをプログラミング可能な、ソフトウェア定義型ネットワーク(Software Defined Network: SDN)の導入を検討する。

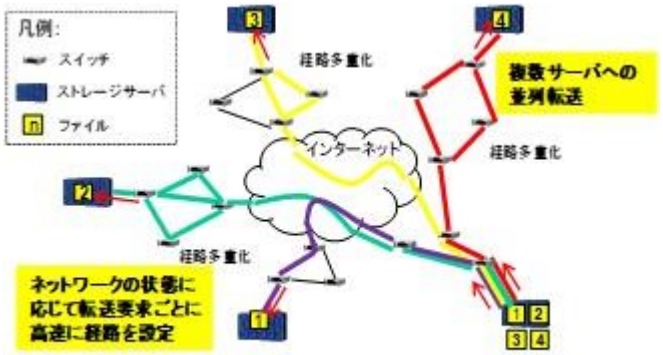


図 3-1 スマートルーティングの概要

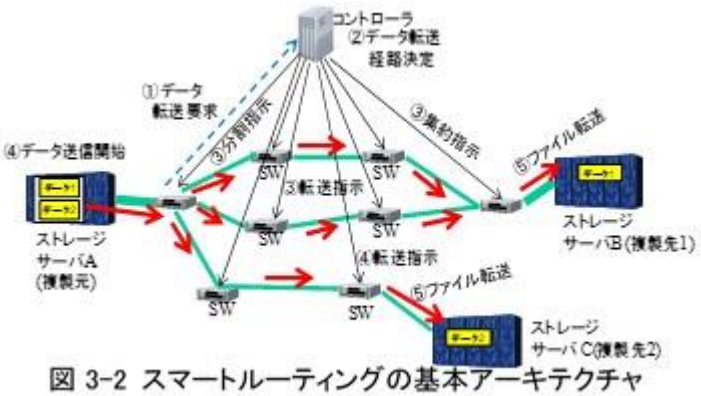


図 3-2 スマートルーティングの基本アーキテクチャ

平成 24 年度には、ネットワークの状態を動的に把握してソフトウェアで適切に経路を設定できる高速 SDN 型ストレージ間通信機能(スマートルーティング)の基礎検討を実施した。具体的には、まず、本方式の基本となる、並列転送最適化の基本動作手順およびアーキテクチャについて設計した。次に、スマートルーティングの基盤技術となるソフトウェア定義型ネットワーク(SDN)について、現在幅広く利用され標準化も進められている OpenFlow を採用して、本機能への適用可能性について検討した。

3.2.1 スマートルーティングのアーキテクチャ設計

本提案機能を実現するためには、データ転送を行うスイッチ/ルータを中央制御装置(コントローラ)により制御する中央集中型のアーキテクチャが適当である。従って本方式では、図 3-2 に示すような、コントローラと複数スイッチによる集中型アーキテクチャで設計した。

コントローラはスイッチから定期的に接続しているリンクのネットワーク状態を収集し、ネットワーク全体のトラフィック状況等を把握する。データ転送要求が発生した場合、データ転送先の情報

とトラフィック状況から、経路選択アルゴリズムにより適切な経路を選択し、スイッチに対して経路設定の指示(転送指示、分割指示、集約指示等)を行う。経路が設定したのち、データ転送が開始される。データ転送中は、コントローラはスイッチからの状況報告を受け、急激なトラフィック変動が発生した場合には対処する。

3.2.2 OpenFlow

SDN の具体的な設計・実装として、本方式では、OpenFlow を導入する。OpenFlow を採用する主な理由としては、本方式の基本アーキテクチャを実現するのに適している点、また標準化が進められているオープンな規格によってプロトコルが規定されているため、汎用性が高い、などの理由による。

OpenFlow は、プログラムの構成をすべてプログラムで制御し、「ネットワークの仮想化」を実現する概念である SDN のひとつである。OpenFlow では、動作をプログラムにより記述できることにより、従来と比べより柔軟なネットワーク構成と管理が可能となる。

OpenFlow は、OpenFlow コントローラと OpenFlow スイッチから構成される。OpenFlow コントローラは、経路計算などを行い、それを実現するためのルールが記述されたフローテーブルを OpenFlow スイッチに書き込む。OpenFlow スイッチは、OpenFlow コントローラに与えられたフローテーブルに従い、実際のフレームの転送等の処理を行う。

3.3 高速SDN型ストレージ間通信機能の設計とシミュレーション

平成 25 年度には、ネットワークの状態を動的に把握してソフトウェアで適切に経路を設定できる高速 SDN 型ストレージ間通信機能(スマートルーティング)の詳細設計とシミュレーション実験を実施した。具体的には、まず、本方式の基本となる、並列転送最適化の動作手順を詳細設計した。

設計した並列転送方式(スマートルーティング)の基本動作に関する基礎的シミュレーション実験を行うため、また次年度以降での提案方式の本実装と本実験の準備のため、ネットワークシミュレーション環境を構築した。具体的には、仮想環境上に OpenFlow コントローラ、スイッチのソフトウェア実装と、仮想ネットワーク環境を導入することで、ネットワークエミュレーション環境を構築した。コントローラには OpenDaylight、スイッチには Open vSwitch を用いた。また仮想ネットワーク環境として Mininet を使用した。本シミュレーション環境において、設計したスマートルーティングの基本動作に関する簡易実装を行い、転送効率の向上が実現可能であることを検証した。

3.4 高速SDN型ストレージ間通信機能の実装と評価

平成 26 年度には、高速 SDN 型ストレージ間通信機能の実装及び性能評価を実施した。具体的には、平成 25 年度に構築したシミュレーション環境上に、設計した複数チャンクを同時並列転送する際の経路選択アルゴリズム、SDN による経路制御手法、並びに経路多重化手法を実装した。さらにネットワーク上のトポロジー情報やフロー情報、トラフィック情報(スループット)をリアルタイムで可視化するツールを実装した。

また, シミュレーションによる性能評価実験を実施した。ここでは図 3-3 に示すようにホストが 7 台 (h1～h7)、OFS が 6 台 (s1～s6)から構成される仮想ネットワーク環境を構築した。スイッチ間のネットワーク帯域は 100 Mbps、スイッチ−ホスト間の帯域は 1000 Mbps とした。本実験では、平常時において、ホストからホストへバックアップのためにデータを高速転送することを想定したシミュレーション実験を行った。まず、ホスト h1 からホスト h2 へ利用可能な多重経路を経路探索アルゴリズムにより計算する。そして、データのバックアップのためのファイル転送として、ホスト h1 からホスト h2 に 1 Gbytes のデータを TCP で送信する。その 10 秒後にホスト h3 からホスト h4 に 50 Mbps のトラフィックを

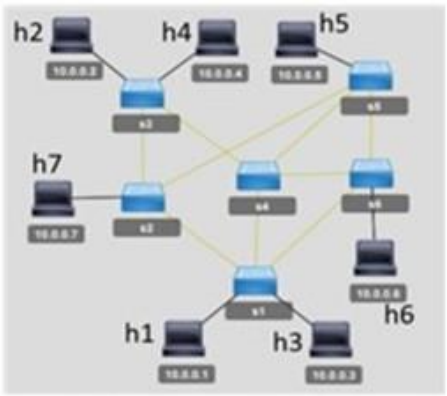


図 3-3 実機実験ネットワーク環境

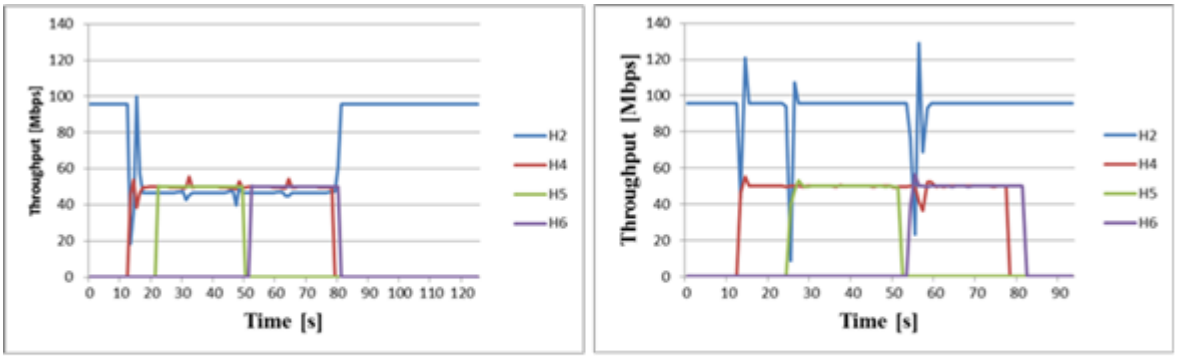


図 3-4 経路多重化なしの場合のスループット
（左：経路制御なし、右：経路制御あり）

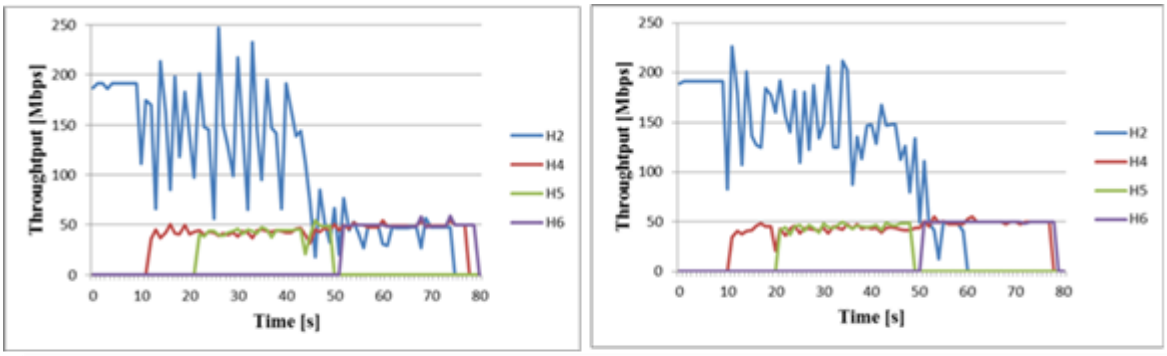


図 3-5 経路多重化ありの場合のスループット
（左：経路制御なし、右：経路制御あり）

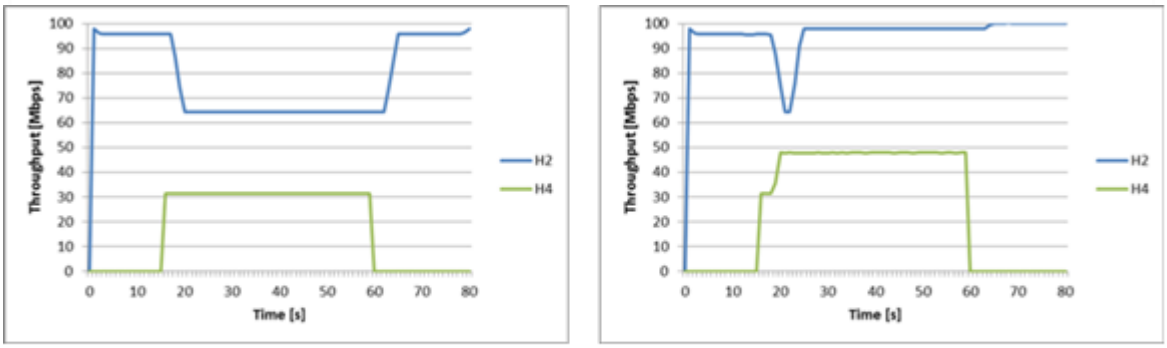


図 3-6 実機実験におけるスループット(左：経路制御なし、右：経路制御あり)

UDP で 70 秒間送信する。そして、実験開始から 20 秒後にホスト h3 からホスト h5 に 50 Mbps のトラフィックを UDP で 30 秒間送信し、実験開始から 50 秒後にホスト h3 からホスト h6 に 50 Mbps のトラフィックを UDP で 30 秒間送信する。この時、ホスト h1 からホスト h2 のスループットをリアルタイムで計測し、50 Mbps を下回った場合に、経路選択アルゴリズムにより経路の切り替えを実施する。ここでは、各経路の利用状況を観測し，最も利用率が低い経路を選択する。

そして、経路多重化なし（単一経路のみ利用）・経路制御なし、経路多重化なし・経路制御あり、経路多重化あり・経路制御なし、経路多重化あり・経路制御あり（提案手法）の 4 パターンで実験を行い、それぞれのスループットを比較した。図 3-4 と図 3-5 にそれぞれのスループットのグラフを示す。結果から経路を多重化し、かつ経路を動的に切り替えることで、全体の性能は約 2 倍に向上した。特に 10 秒から 30 秒にかけては平均のスループットは約 50Mbps から約 200Mbps へと 4 倍もの性能向上が見られた。

次に、実機上で実施した実験内容を以下に示す。実機上にも同様に実験環境として可視化ツールを構築し高速化効果を検証した。まず h1 から h2 に 50Mbps のトラフィックを 2 フロー、UDP で、

s1 経由で 80 秒間送信する。それから 15 秒後に h3 から h4 に 50 Mbps のトラフィックを s1 経由で，UDP で 45 秒間送信する。h3 から h4 のトラフィックが送信後、h1 から h2 のトラフィック(1 フローのみ)に対してスマートルーティングを適用し、s2 経由に経路を変更する。そして、経路制御なしと経路制御ありのそれぞれのスループットを比較した。実験結果のグラフを図 3-6 に示す。グラフから約 1.5 倍のネットワークの転送高速化を実現できた。

3.5 高速SDN型ストレージ間通信機能の改良と評価

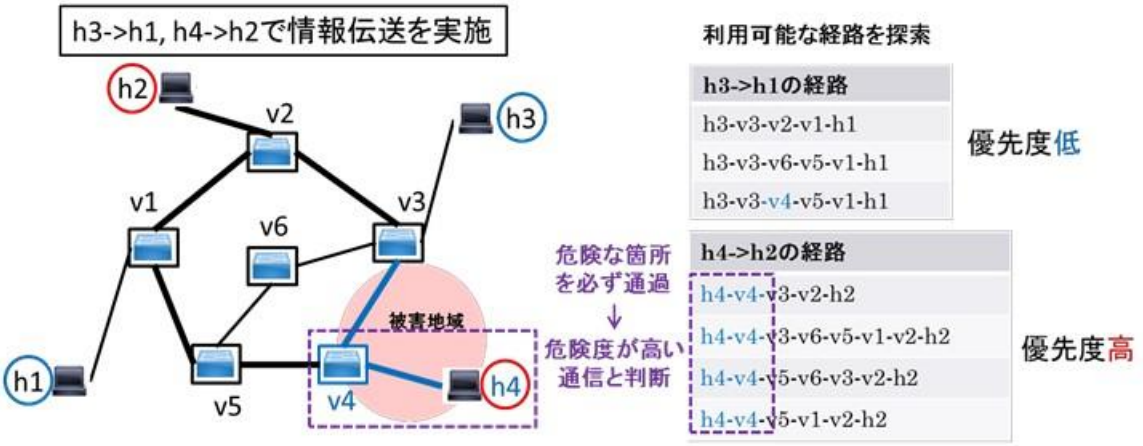


図 3-7 災害リスクに基づいた通信の優先制御

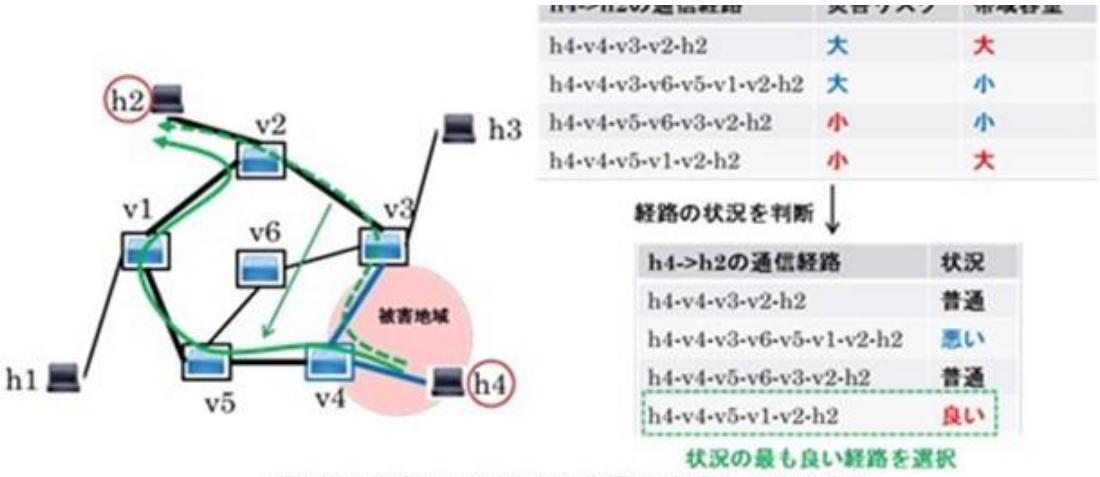


図 3-8 災害リスクと帯域容量を考慮した経路制御

平成 27 年度は、高速 SDN 型ストレージ間通信機能の改良・実装、及び性能評価を実施した。具体的には、平成 26 年度に実施したネットワークシミュレータ上及び実機上での高速化効果の検証結果に基づき、同通信機能の性能向上を目指した方式改良を行い、シミュレータ上及び実機上で実装・実験を実施して、より効果的な経路選択、経路多重化等の効果により、シミュレータ上で 7 倍、実機上で 3.5 倍程度(今年度の数値目標)の高速化が可能かを検証した。

まず、ノードのリスクに応じて複数

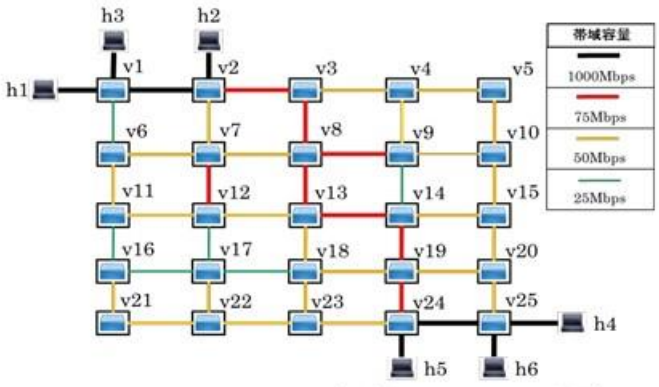


図3-9 シミュレーション実験ネットワーク環境

通信の優先制御を行う経路選択アルゴリズムを検討し、災害リスクに基づいた通信の優先制御と災害リスクと帯域容量を考慮した経路制御の設計及び実装を行った。災害リスクに基づいた通信の優先制御では、図 3-7 に示すように遮断される可能性が高い通信をより状況の良い経路に割り当て、情報伝送が完了する前に通信が途絶えることを防ぐ。災害リスクと帯域容量を考慮した経路制御では、図 3-8 に示すように状況の変化に応じて、高速かつ確実に情報を伝送可能な経路を随時選択し、トータルでの情報伝送量を増大させる。

さらに、改良した方式をシミュレータ上及び実機上に実装し、より効果的な経路選択、経路多重化等により、シミュレータ上で 7 倍、実機上で 3.5 倍程度の高速化効果が実現できるか否かを検証した。シミュレータ上で実施した実験内容を以下に示す。図 3-9 に構築したネットワーク環境を示す。ここではホストが 6 台、スイッチが 25 台から構成される仮想ネットワーク環境を用いて構築した。

本実験では、災害時において、ホスト h4 からホスト h1 へ、ホスト h5 からホスト h2 へ、ホスト h6 からホスト h3 へそれぞれバックアップのためにデータを転送することを想定したシミュレーション実験を行った。まず、各ホストは送信先までの利用可能な多重経路を経路探索アルゴリズムにより計算する。この時、時刻 0 で地震が発生し、ネットワーク内のノード（スイッチ）とリンクの損壊確率が時間と共に緩やかに上昇する。同時に各ホストはバックアップのために 1Gbyte のデータを TCP で送信を開始する。そして、時刻 200 で、ホスト h4 付近で津波が発生し、周辺のノードとリンクの損壊確率が急激に上昇する。本実験における地震と津波による損壊確率（災害リスク）は以下の通りである。

・地震による災害リスク

ノードi の災害リスク $n_i^{\{1\}}(t) = 0.002 + 0.00001 * t$

ノードi-ノードj間のリンクの災害リスク $l_{ij}^{\{1\}}(t) = 0.002 + 0.00001 * t$

・津波による災害リスク

$$n_i^{\{2\}}(t) = \begin{cases} 0 & (t \le t_i) \\ \dfrac{t - t_i}{30} & (t_i < t \le t_i + 30) \\ 1 & (t_i + 30 < t) \end{cases}, \quad l_{ij}^{\{2\}}(t) = \begin{cases} 0 & (t \le t_{ij}) \\ \dfrac{t - t_{ij}}{30} & (t_{ij} < t \le t_{ij} + 30) \\ 1 & (t_{ij} + 30 < t) \end{cases}$$

この時、従来手法として経路長のみを考慮した場合と提案手法のそれぞれで実験を行い、スループットと情報伝送量を比較した。各手法による情報伝送量を表 3-1 に示す。表から提案手法では、災害リスクを考慮して、災害の影響を受けやすい通信をより状態の良い経路に優先的に割り当てることで、ホスト h4 からホスト h1 への情報伝送量は約 4.5 倍、ホスト h6 からホスト h3 への情報伝送量は約 12 倍に向上したことを確認した。

表3-1 各通信の情報伝送量

	提案手法	経路長のみ考慮
h4-h1の通信の情報伝送量	226 Mbyte	1009 Mbyte
h5-h2の通信の情報伝送量	1035 Mbyte	933 Mbyte
h6-h3の通信の情報伝送量	72 Mbyte	886 Mbyte
総情報伝送量	1333 Mbyte	2828 Mbyte

次に、実機上に提案手法を適用し、3.5 倍程度の高速化効果が実現できるか否かを検証した。構築した実機実験用ネットワーク環境を図 3-10 に示す。実機の OpenFlow スイッチとして NEC PF5240 を利用した。また、スイッチ間の帯域は 100Mbps、スイッチ-ホスト間の帯域は 1000Mbps とした。

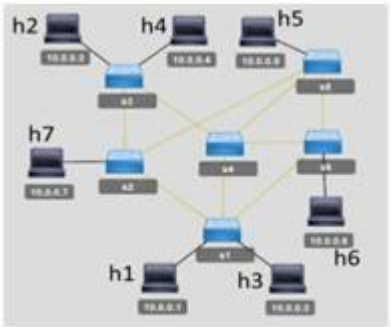


図 3-10 実機実験ネットワーク環境

実験手順を説明する。本実験では平常時にホスト h1 から h2 へバックアップデータを送信することを想定した。まず、ホスト h1 からホスト h2 への多重経路を計算し、1 Gbyte のデータを多重経路により、TCP で送信する。そし

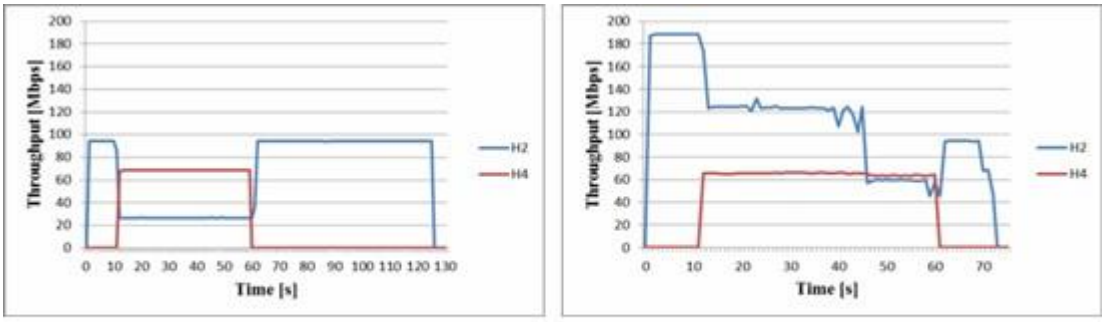


図3-11 実機実験結果(左:単一経路固定、右:提案手法)

表3-2 データ転送時間と平均スループット

	データ転送時間(h2) [秒]	10(s)-60(s)間の平均スループット [Mbps]		
		全体	h2	h4
既存手法 単一経路固定	127.1	95.2	26.6	68.6
提案手法 多重経路的変更	73.4	168.9	103.8	65.1

て、10 秒後にホスト h3 からホスト h4 に 70Mbps のトラフィックを UDP で 70 秒間、送信する。この時、h1 から h2 へのトラフィックに対して経路選択アルゴリズムにより、動的に経路を変更する。そして、各トラフィックのスループットと 1 Gbyte のデータの転送時間を計測し、既存手法として単一経路のみを固定して利用した場合と比較を行った。実験結果のグラフを図 3-11 に示す。グラフから多重経路を利用して、かつネットワークの利用状況に応じて経路を動的に変更することで、スループットが向上したことを確認した。また、表 3-2 に各トラフィックの平均スループットとデータ転送時間を示す。表から h2 へのトラフィックについては約 3.9 倍のスループットの向上が確認された。また、次年度に本格実施予定の実証実験に向けた実機による実験用 OpenFlow ネットワークシステムの構築を行った。具体的には、大学学内ネットワークを中心に、L1、L2 の様々な通信媒体、通信方式で OFS 間を接続するハイブリッドな実験用ネットワークを構築した。これにより、多様なネットワーク構成を疑似的に実現しつつ、実機による本方式の性能評価、実証実験が行えるネットワーク環境を実現した。

3.6 高速SDN型ストレージ間通信機能の総合評価

平成 28 年度は、平成 27 年度に開発を行ったスマートルーティングの総合評価を行った。この手法は遮断される可能性が高い通信をより状況の良い経路に割り当て、情報伝送が完了する前に通信が途絶えることを防ぐ。さらに経路の状況の変化に応じて、高速かつ確実に情報を伝送可能な経路を随時選択し、全体での情報伝送量を増大させる。さらに、シミュレーション実験による提案手法の総合評価を実施した。

シミュレータ上で実施した実験内容を以下に示す。図 3-12 に構築したネットワーク環境を示す。

ここではホストが 6 台、スイッチが 25 台から構成される仮想ネットワーク環境を構築した。本実験では、災害時において、ホスト h4 からホスト h1 へ、ホスト h5 からホスト h2 へ、ホスト h6 からホスト h3 へそれぞれバックアップのためにデータを転送することを想定したシミュレーション実験を行った。まず、各ホストは送信先までの利用可能な多重経路を経路探索アルゴリズムにより計算する。この時、時刻 0 で地震が発生し、ネットワーク内のノード(スイッチ)とリンクの損壊確率が時間と共に緩やかに上昇する。同時に各ホストはバックアップのために 1Gbyte のデータを TCP で送信を開始する。そして、時刻 200 で、ホスト h4 付近で津波が発生し、周辺のノードとリンクの損壊確率が急激に上昇する。

本実験における地震と津波による損壊確率（災害リスク）は前述の通りである。この際、損壊確率に基づいて破損したノードとリンクの場所と時間を記録し、これを災害シナリオとする。本実験では災害シナリオを 5 パターン用意し、実験を行った。そして、既存手法として、経路長のみを考慮し最短経路を利用する場合と帯域を考慮して広帯域の経路を利用する場合、災害リスクを考慮して最も災害リスクの低い経路を利用する場合、提案手法のそれぞれで、300 秒経過した際のデータ伝送量を比較した。

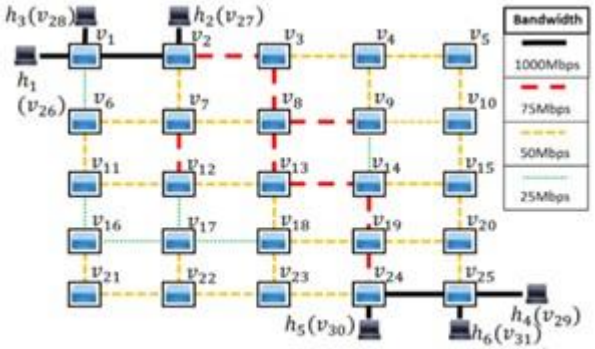
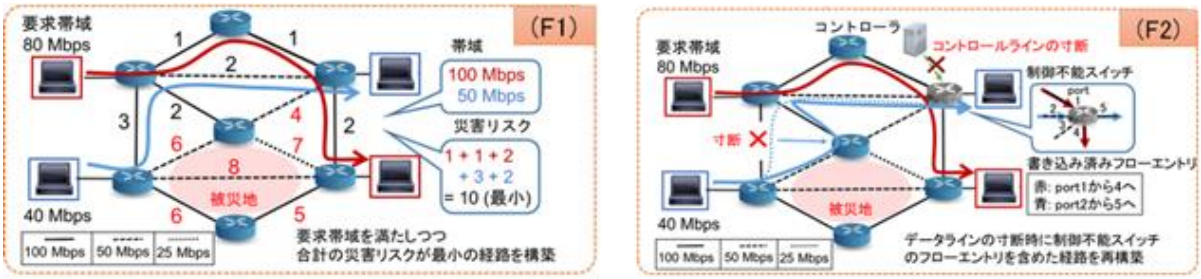


図 3-12 シミュレーション実験におけるネットワークトポロジ

表 3-3 各手法の総データ伝送量

災害シナリオ	総データ伝送量(Mbyte)			
	経路長のみ	広帯域利用	低リスク利用	提案手法
シナリオ1	1,587.8	1,172.7	1,424.0	2,005.6
シナリオ2	1,357.1	2,274.1	1,292.5	2,276.0
シナリオ3	1,410.0	2,455.5	1,458.6	2,537.4
シナリオ4	1,376.0	1,521.7	1,360.3	2,458.1
シナリオ5	1,315.3	1,905.1	1,315.1	2,121.5
全体平均	1,409.2	1,865.8	1,370.1	2,279.7

各シナリオでの総データ伝送量とその平均を表 3-3 に示す。実験結果から各シナリオの総データ伝送量を比較すると、提案手法はどのシナリオにおいても既存手法と比べて高い値となったことが確認できる。具体的には，データ伝送量は全体平均では提案手法では 2279.7MByte，最短経路のみ考慮した場合には 1409.2Mbyte，広帯域の経路を利用する場合には 1865.8Mbyte，最も災害リスクの低い経路を利用する場合には 1370.1Mbyte となり、提案手法の場合ではデータ伝送量が約 1.2～1.6 倍に増加した。



(F1) 制限時間を考慮した経路構築機能　　(F2)制御不能スイッチも活用した経路再構築機能

図 3-13 制限時間を考慮したスマートルーティングの改良

上記の結果に関して総データ伝送量に大きな差が出た災害シナリオ 4 に着目して考察を行う。経路長のみ考慮している場合には帯域容量を十分に利用できず、また通信の優先制御も行われていないため、災害リスクの高い h6 から h3 への通信がデータ伝送の途中で遮断されたため、データ伝送量は 1260.0Mbyte に留まった。広帯域の経路を利用する場合には帯域容量を十分に利用することができたが、災害リスクを考慮していないため、災害により通信が切断されやすくなり、その結果としてデータ伝送量は 110.6Mbyte に留まった。また、最も災害リスクの低い経路を利用する場合は、災害により通信が切断されにくくなるが、帯域容量を十分に利用することができないため、データ伝送量は 132.8Mbyte に留まった。一方で提案手法では、災害リスクに基づく優先度を考慮した通信制御機能により、災害リスクの高い通信に関して優先的に状況の良い経路へと割り当て、データ伝送を行うことで、通信の遮断前にバックアップ伝送が終了し、最終的なデータ伝送量を増加させるができた。特に h6 から h3 へのデータ伝送量は約 4 倍に向上することが確認された。

次にスマートルーティングの改良を行った。具体的には、(F1) 制限時間を考慮した経路構築機能と (F2) 制御不能スイッチも活用した経路再構築機能の設計を行った。その概要を図 3-13 に示す。(F1) 制限時間を考慮した経路構築機能では、制限時間内で各通信が転送するのに必要な帯域を持つ経路の中で，災害の影響を受けにくい経路を利用する。具体的には、複数の通信が同じ経路を使う場合は、制限時間内に転送できるように帯域制御を行うことで、寸断による経路の変更

回数を低減し、制限時間内で確実に転送する。(F2) 制御不能スイッチも活用した経路再構築機能では、制御不能スイッチを書き込み済みの経路表で動作させることで、経路表に従った固定のデータラインとして認識する。通信中断時には、制御不能スイッチの経路表も利用し、新しい経路を計算することで、データラインが繋がる制御不能スイッチを活用し、利用可能な転送経路を向上させる。

また、(F1) 制限時間を考慮した経路構築機能を実機環境上に実装して、実験を行った。実験環境を図 3-14 に示す。OpenFlow スイッチとして、Pica8 P-3295 を用いた。

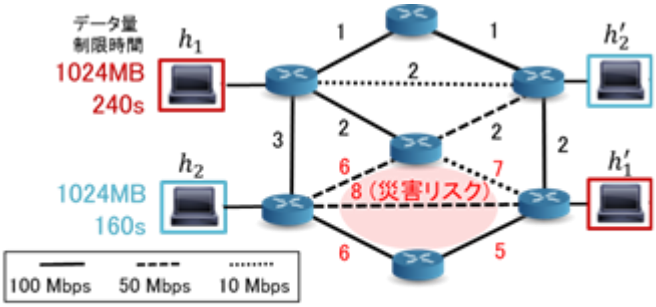


図 3-14 実機実験におけるネットワーク環境

表 3-4 各手法における転送量と寸断回数

	広帯域利用		提案手法(改良前)		提案手法(改良後)	
	転送量	寸断回数	転送量	寸断回数	転送量	寸断回数
h_1	1,024MB	0	1,024MB	0	1,024MB	0
h_2	716MB	1	848MB	0	1,024MB	0

まず、開始直後に、各ファイルサーバは TCP 通信を開始する。10 秒経過後に，災害リスクと制限時間を算出し，20 秒経過後に災害リスクが 5 以上の経路を遮断する。この時、帯域が最も大きい経路を選ぶ広帯域利用、提案手法(改良前)、提案手法(改良後)のそれぞれで実験を実施し、制限時間内の転送量と経路の寸断回数を比較した。

実験結果として各手法のスループットのグラフを図 3-15 に、制限時間内における転送量と寸断回数を表 3-4 にそれぞれ示す。提案手法では、災害リスクの高い経路を避け、制限時間の短い通信に広い帯域を割り当てることで、どちらの通信も全てのデータを完了させることができた。

さらに、改良した方式を分散ストレージシステムに適用して、高速化通信方式の検討を行った。この方式において各サーバは複数のストレージを保持し、ストレージ間でデータを送受信する。その際に SDN を用いてスイッチ間で利用するリンクを動的に・柔軟に制御を行う。具体的には、コントローラがスイッチからネットワークトポロジ、トラフィックの情報を収集する。そしてトラフィックの本数を基に各トラフィックが利用するリンクを設定し、フローテーブルを生成し、各スイッチに設定する。また、通信中、コントローラはスイッチからポート毎や通信毎のトラフィック量を収集する。そして各リンクの利用率を計算し、利用率の低いリンクに通信を送信するようにフローエントリを生成して、スイッチのフローエントリを更新する。そしてこの通信方式によるスループット向上の検証のための実機実験を行った。具体的には図 3-16 に示すように実機の SDN スイッチを 2 台用意して、各スイッチにホスト(H1, H3)とサーバ(H2, H4)を接続した。SDN スイッチ間は 1Gbps のリンクを 11 本接続し、SDN スイッチとホスト・サーバ間はそれぞれ 10Gbps と 1Gbps のリンクを接続した。そして、H1 から H2 に 1Gbyte のデータを伝送するトラフィックを 10 本送信し、その 10 秒後に H3 から H4 へ 800Mbps のストリーミングデータを送信した。この時、リンクアグリゲーションにより 8 本の物理リンクを 1 つの論理リンクと設定した場合(LAG)、SDN により各トラフィックが利用するリンクを動的に選択した場合について、それぞれ実験を行い、各ホストのスループットを計測した。

図 3-17(a)にリンクアグリゲーション時の各ホストのスループットのグラフを示す。グラフから、特

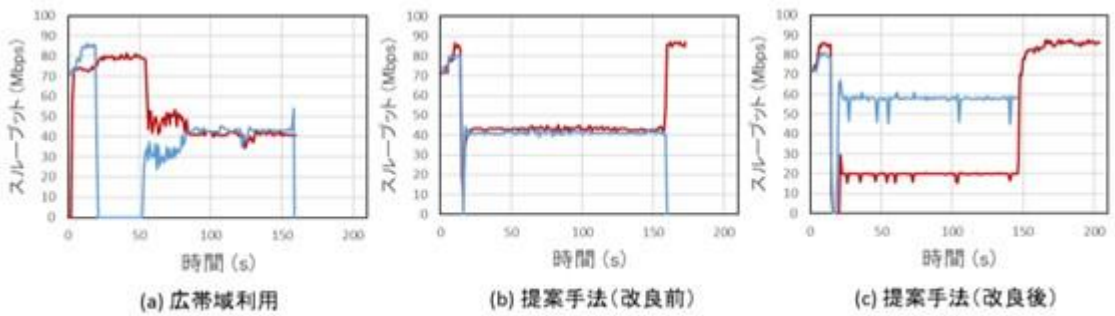


図 3-15 各手法におけるスループット

定のリンクにトラフィックが偏っているため、十分なネットワーク性能を確保できていないことが確認できる。二つのトラフィックが混在している 10 秒から 70 秒の間、ネットワーク全体の平均スループットは 1.75 Gbps に留まった。図 3-17(b)に SDN により動的にリンクを選択

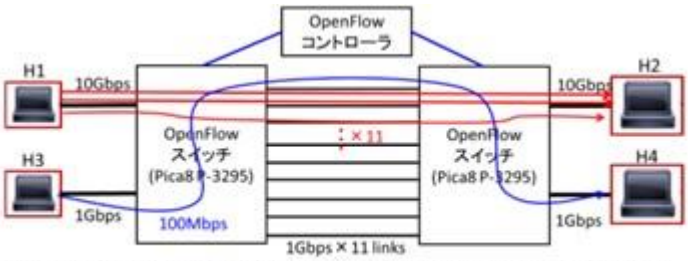


図 3-16 分散ストレージシステム向け高速化通信方式の実験環境

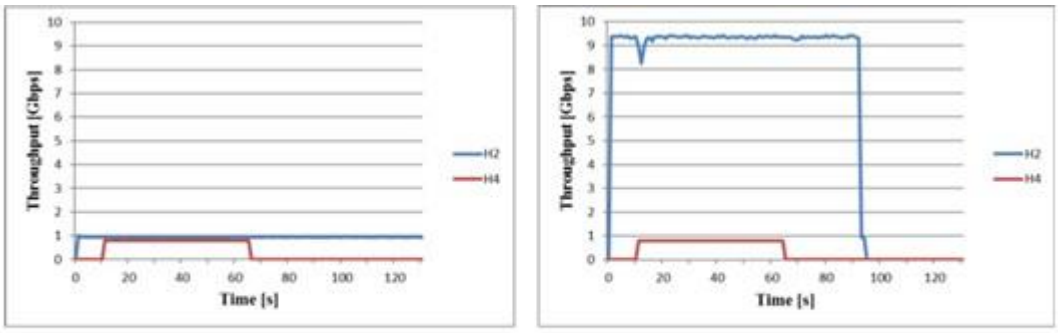


図 3-17 分散ストレージシステム向け高速化通信方式の実験結果（各ホストのスループット）

した場合のスループットのグラフを示す。グラフから、トラフィックの偏りを減らし、十分なネットワーク性能を確保でき、他のトラフィックが発生した場合もその影響を考慮して、他の空いているリンクを利用することでスループットの低下を防ぐことができた。10 秒から 70 秒の間のネットワーク全体の平均スループットは 10.1Gbps となり、従来よりも最大 5.8 倍の性能を得ることができた。

この手法によりトラフィック状況に応じて利用するリンクを動的に選択することで、スループットの向上が確認できた。従って、提案手法により分散ストレージにおけるデータ伝送の効率化が期待できる。以上より、改良した方式を実機上に実装し、従来の 5 倍となる 10Gbps の高速データ転送を実現することができた。

4） 高機能プログラミングフレームワーク

4.1 本項目の背景と目的

本項目の目的は、本プロジェクトが実現する大容量かつ高速なストレージハードウェア技術およびそれによって実現される高可用性ストレージシステムを、種々のソフトウェア開発の中で使いこなすためのソフトウェア技術の確立である。

本項目では、この一般的な目的の達成のために、以下の具体的な技術的目標を設定した。

- クラウド系データと関係データベースの双方を系統的かつ安全にアクセスする方式を構築する。
- 同方式を、次世代高信頼プログラミング言語SML#にシームレスに統合する。

より具体的には、高可用ストレージ技術を活用した高度な次世代の情報処理システム開発におけるソフトウェアの生産性と信頼性を飛躍的に高めるために、上記の目標の達成によって実現される拡張されたSML#を、我が国のソフトウェア生産の現場である産業界へ展開することである。

我が国の学術研究によって実現された次世代高信頼プログラミング言語が産業界で使用される例は殆ど存在しない現状から推測される通り、この目的の達成には産業界との連携が不可欠である。

本プロジェクトにおける本項目の目的は、本プロジェクトが実現する高可用ストレージシステムの実証実験である投薬情報システム（お薬手帳システム）開発の生産性および信頼性等の向上に貢献するために、上記目標の達成によって実現されるSML#言語をお薬手帳システム開発のため

のソフトウェア基盤として提供し、高機能プログラミングフレームワークを実現する次世代言語の実用性を検討・検証することである。高機能プログラミングフレームワークの、産学連携で推進する本プロジェクト内部でのソフトウェア生産への貢献は、将来の我が国のビッグデータ解析等を含む高付加価値のソフトウェア生産への貢献の第一歩であり、その重要なステップである。

以上が、基礎研究である本項目の、産学官プロジェクトへの貢献であり、その意義と位置づけである。このような性格から、本項目は、本プロジェクト経費の分配を受けずに実施した基盤的基礎的研究である。本プロジェクトからの研究費の補助は受けていないが、本項目の研究は、本プロジェクトの中で、日立ソリューションズ東日本殿の開発のサポートやそれにともなう討論などから得られた種々の研究シーズ、さらに本プロジェクトを通じた日立ソリューションズ東日本殿との緊密なディスカッションなどの寄与のもとに達成されたものである。特に、以下の 28 年度の（3）の2の JSON データの高水準かつ型安全なアクセス方式は、本プロジェクトでの産学連携から得られたシーズをもとにして達成されたものと言える。なお、それら基盤研究成果の一部は、本項目の従事者が所属する東北大学の基盤的経費や競争的外部資金等の補助をも受けて遂行されたものである。

以下，成果の概要を報告する。その詳細は、中間報告および各年度の報告の通りである。

4.2 平成25年度までの成果の概要

平成25年度までに予定通りの成果をあげた。具体的には、

1）高機能プログラミングフレームワークのためのデータモデルと型理論の構築（平成24年度）

2）高機能プログラミングフレームワークのための言語機能の構築（平成25年度）

の課題を達成した。これら両者は、技術的な展開として連続したものであるので、統一した記述として記述する。

前述の通り、本研究項目が開発を目指す高機能プログラミングフレームワークの目的は、これらデータベースとデータ構造を、次世代高信頼プログラミング言語 SML#にシームレスに統合することによって、完全かつ高機能なプログラミング環境を提供することである。この目的を達成するための最初の課題は、その理論的基礎となるデータモデルとその型理論を構築することである。データモデルは、操作対象のデータの構造とその操作系、その性質等を定義するものであり、その定義には必要とする機能や操作性を考慮した検討が必要である。データモデルの型理論とは、データとその操作系をプログラミング言語に統合することを可能にするための型導出システムに関する理論である。この両者を確立することにより、高機能プログラミングフレームワークが目指す高機能かつ安全なデータ構造の設計と実装が可能となる。

この課題達成のため、まず本プロジェクトの実証実験として構築予定のお薬手帳実証システムを高可用性ストレージ活用アプリケーションの典型と捉え、その実現に必要な機能を、実証システムの開発を担当する株式会社日立ソリューションズ東日本と共同で検討した結果、必要な機能はローカルストレージ上で SQL を通じて提供される関係データベースへのアクセスおよび高可用ストレージ基盤が REST インターフェースを通じて提供するキーバリュストアへのアクセスであり、さらに関係データベースには集約関数等の機能が有用であるとの結論を得た。そこで、現状の SML#の SQL 連携を集約関数で拡張するための型理論を構築し、さらに型付きのキーバリューストアクセスのためのデータモデルとその型理論を構築した。さらにそれら双方に対して、構築した型理論を基礎として、それらシステムを言語に取り入れるための中心となる技術であるコンパイルアルゴリズムを開発し、言語機能を構築した。

前者は、我々が SML#の SQL 連携のために構築した一連の基礎理論の技術的な拡張である。後者は、本プロジェクトにおいて着手した新たな研究トピックであり、我々の現段階の認識では、世界的にも未だ確立していない先端テーマである。本研究で構築した新たな機能は、SML#の SQL 統合という先端機能を実用化する上で重要である。その詳細は中間報告で詳述の通りである。

4.3 平成26年の成果の概要

平成26年度の研究計画は、前年度までに構築した高水準キューバリューストア機能を含む高機能プログラミングフレームワーク機能のプログラミング言語SML＃へ統合であった。この統合の

意義は、その結果得られる拡張されたSML＃言語が、平成27年度から本格的に開始予定とされる高可用ストレージの実証実験であるお薬手帳システム開発のためのプログラミング言語として貢献すること以外、存しえない。

本年度は、この観点から、まず、お薬手帳システムと高可用ストレージシステムの関係及びお薬手帳システムのソフトウェア構造の分析を行なった。この分析の結果、以下の点が明らかとなった。

- お薬手帳システムは、クライアント・サーバ型の Web アプリケーションである。
- 多数のリクエストへのスケーラビリティを考慮し、クライアントとお薬手帳サーバ間の実装は、REST 方式を基本とする。
- 高可用ストレージシステムは低レベルのブロックアクセスインターフェイスを提供するのみであり、高水準のプログラミングインターフェイスは、本プロジェクトでは、開発の対象ではない。
- お薬手帳システムからの高可用ストレージの利用は、高可用ストレージシステムが提供する低レベルインターフェイスを直接使用する。

以上の分析の結果、お薬手帳システムを高信頼関数型言語で開発する上で要求される言語機能等の分析として以下の点が明らかとなった。

- キーバリュストアのための高水準プログラミング機能は高可用ストレージの高水準な扱いを意図したものであり、その高可用ストレージ向けの実用化には高可用ストレージ側のシステム開発が必要であるため、本プロジェクトでは実施困難である。
- お薬手帳システムを高水準言語で開発するには、高信頼関数型言語における WEB アプリケーション開発環境の整備が必要である。
- お薬手帳システムが扱うデータのプログラミングサポートとしては、高可用ストレージアクセスは S3 準拠の低レベルのもので十分であるのに対して、REST 方式の Web アプリケーション実現に必要な、クライアント・サーバ間でやり取りされる複雑なデータに対するプログラミングサポートが重要となる。
- さらに、お薬手帳システムサーバが関係データベースを自由にアクセスできれば、お薬手帳システムの機能を拡張できる可能性を持つ。

以上の分析を踏まえ、26年度の研究の方針を以下のように拡大・変更した。

- 高水準の型付きキーバリューストアアクセス機構のSML＃言語への組込みは産業界のニーズに従ったもとして実現して初めて意味をもつものであることを考慮し、産業界側が連携してこれら機能を使ってシステム開発を行う体制が整った時点で行うこととする。
- お薬手帳実証実験を高信頼関数型言語で実現するための必要とされる Web アプリケーション開発基盤を新たに研究開発する。

この方針の下で、以下の具体的な目標を設定して研究を実施した。

1. SML＃内での WEB サーバを実現するための機能を研究開発する。
2. REST 型のサーバクライアント連携を高水準に実現するために、SML＃と外部との高水準なデータのやり取りの基礎として、JSON データの高水準の型付きアクセス機構を研究開発する。

これら研究成果の概要を以下に示す。

4.3.1 SML#での高水準な Web アプリケーション開発環境

今日広く用いられている Web アプリケーションフレームワークは、プログラムの自動生成やファイル名の命名規約など、プログラミング言語の範囲を超えた記述を要求するものが多い。この方法は、典型的なアプリケーションの開発において生産性が高いと考えられている一方、実行されるコードの全体像を把握することが難しく、フレームワークの想定を越える細やかな制御等を必要とするアプリケーションを書くことを難しくしている。

この問題は、Web アプリケーションを構成する全ての要素を単一のプログラミング言語で記述する環境が整備されるなら解決できるはずである。より具体的には、高階関数やモジュール言語などを含み、柔軟で高い記述力を持つ関数型言語が、Web アプリケーション開発に必須なデータ

ベースアクセスやスケーラブルなサーバを実現する上で重要なマルチコア CPU 上のスレッド機能をサポートするなら、Web アプリケーションを通常のプログラムと同様、一つのアプリケーションとして記述する環境を提供できるはずである。Cとの直接連携を通じたネイティブスレッドやシームレスなSQL統合を実現しているSML＃は、このアプローチを実践する出発点として最適と考える。そこで本研究では、以上の洞察に基づき、SML＃による高水準な Web アプリケーション開発の可能性を分析し、単独プロセスの Web サーバとして動作するユーザープログラムとして Web アプリケーションを開発する枠組みを提案し、実用的な Web アプリケーションを施策することによって、その可能性を検証した。

本研究は、Web アプリケーション開発実績を持つ日立ソリューションズ東日本殿からの受託研究員と共同で実施した。

本研究は、関数型言語での高水準な Web アプリケーション開発環境構築の第一歩であり、多くの課題が残されているが、実際の試作を通じて、従来の Web フレームワークが採用しているメタな技術を用いることなく、全ての要素をSML＃のプログラムとして記述することを実証できた。具体的には、Web アプリケーションの基盤となる HTTP サーバ機能は、SML＃の C との連携機構を活用し既存の HTTP サーバーライブラリをインポートし、SML＃の高階関数と組み合わせることで、高性能かつ高水準な Web サーバーライブラリを実装することができた。また、アプリケーション本体は、REST アーキテクチャを採用することで、HTTP 通信と関数の類似性から、個々のリソースに対するリクエストを処理するコントローラ関数の集まりとして実装することができた。データベースとの連携を司るモジュールは、SML＃のデータベース統合機構により、データベースへの問い合わせをラップした単純な構造として記述することができた。試作全体を通じて、SML＃単独による Web アプリケーション開発の高い生産性を確認することができた。

今後の課題として、更なる開発項目を列挙する。まず、Web に関連する様々な技術を抽象するライブラリの整備が必要である。例えば、Digest 認証などの認証機構や、クッキーの取り扱い、マルチパートフォームデータの処理などは、本試作では扱っていない。多くの Web アプリケーションフレームワークでは、これらの機能は組み込みで提供されている。本発表で提案した開発方式では、これらの機能はSML＃ライブラリとしてユーザに提供するのが望ましい。また、本試作は純粋な RESTful アプリケーションであるから、ステートレスな通信のみを行っているため、セッション管理を取り扱っていない。一方、現実のアプリケーションでは、セッション管理を必要とする場面も多い。本開発方式においてセッションの概念をどのように取り扱うかは、注意深い設計を要する今後の課題である。最後に、スレッドプールを利用したリクエストの並行処理のサポートは、Web サーバの更なる効率化のためには必須である。スレッドプールによる並行処理は libmicrohttpd がすでにサポートしており、SML＃はマルチスレッド対応と組み合わせるならば、リクエストの並行処理は自然に達成できるはずである。

4.3.2 JSON データの型付き高水準部アクセスのための分動的レコード型を持つ計算系

本研究では、静的型システムをもつ高信頼関数型言語が JSON データを高水準にしかも型安全に扱うための言語機能を研究開発した。この機能を実現する上での主な障害は、JSON データが本質的に持つ種々のデータが混在した集合の存在である。静的型システムの理論が表現可能なリストなどの集合データは、その要素が共通の型を持つものに限定される。しかるに、JSON の配列データは、典型的には、共通なフォールドを含むが、各要素は、それぞれ独自のフィールドを含む異種のレコードの集合である。この性質が、WEB アプリケーションなどでの JSON データの一つの強みとなっていると同時に、静的な型を与えることを困難にしている。本研究では、Buneman と Ohori によって1996年に提案された「部分的に動的に型付けられたレコード」の考え方を基礎に、静的な型制約と JSON 配列の持つ異種データ集合が共存する型付き言語を提案し、その実装方式を構築した。部分的に動的に型付けられたレコードは、フィールドの部分集合が静的に推論可能な、動的な型を持つレコードである。この機構によって、JSON データを型付きデータとして扱うことが可能となる。この機構を下に、SML＃言語に JSON データを導入するための型システムを構築した。JSON データは、SML＃言語の特長であるレコード多相性を使って柔軟に処理することができる。

本研究は、Web アプリケーション開発の実績を持つ日立ソリューションズ東日本殿からの受託研究員と共同で実施した。

本研究が達成した具体的な成果は、以下の通りである。

- JSON データの型システムを構築した。
- 部分レコード型を持つ多相型言語の型システムの構築し、部分レコード型の実行時データを型付きJSONデータとみなすことによって、JSONデータを扱う多相型システムを構築した。
- SML#に上記の JSON 機能を導入するための型主導コンパイル方式を開発した。

以上の結果を用いれば、SML#言語を拡張し、従来要求されていた煩雑な構文構造の解析などを必要とせず、JSON データが型付きデータ構造として高水準に操作できる言語が実現可能なはずである。

4.4 平成27年度の成果

平成 27 年度は、平成 26 年度に引き続き研究を遂行し以下の研究成果をあげた。

4.4.1 高水準キーバリュストアの型理論とコンパイル方式

本項目は、主に平成 25 年度までに進めていた研究項目である。本年度は、この項目を収束させ、雑誌論文として発表した。

4.4.2 SML#での高水準な Web アプリケーション開発環境の構築

平成 26 年度に行った SML#での Web サーバ開発の枠組みの開発を進め、スレッドプールを利用したリクエストの並行処理のサポート環境を整えるなどの完成度を高めた。今後さらなる研究開発を進め、開発環境を日立ソリューションズ東日本へ提供予定である。さらに、この Web アプリケーション開発環境を実際に使用して

Webアプリケーションを試作し、それらの有用性の確認と問題点の分析を行った。その結果、従来の Web アプリケーション開発における幾つかの重要な問題点が明らかとなった。それらの成果の一部は、研究発表で報告している。

4.4.3 JSON データの高水準かつ型安全なアクセス方式

平成 26 年度から開始した Web アプリケーションフレームワークの理論と実装技術の研究開発の中心課題の一つである。平成 27 年度は、平成 26 年度の成果を発展させ、理論と実装技術を構築し、それらを SML#言語へ実装した。

4.4.4 Web アプリケーションの整合性と型安全性を実現する理論と実装技術

現在の Web アプリケーション構築方法論に内在する脆弱性の原因の分析と、それ らを取り除き、Web アプリケーションの信頼性と安全性を飛躍的に向上させる目 的で、Web プログラムの参照整合性と型健全性を保証する枠組みの研究を開始し、初期的な結果を得た。

4.5 平成28年度の成果

平成 28 年度の計画は以下の通りである。

「平成 27 年度に引き続き高信頼な WEB アプリケーションフレームワークの構築のための基本方式、型理論、およびそれらを関数型言語SML＃で実現するための実装技術の研究開発を行い、これら成果を、日立ソリューションズ東日本が、お薬手帳システム開発にて活用するためのサポートを行う。」

この計画のもと研究開発を推進してきた、その研究成果の主なものは以下の3点である。

4.5.1 SML＃での高水準な Web アプリケーション開発環境の構築

平成 26 年度から推進してきたSML＃での Web サーバ開発の枠組みの開発を平成 27 年度の成果を発展させさらに進め、次項で報告する JSON データの扱いを取り入れたリクエストのより安

全な処理機構、お薬手帳の開発のサポートの中で明らかになったクエリ文字列からのデータベース問い合わせ式の生成のためのより柔軟な SQL 表現を取り入れた Web サーバ開発の枠組みのプロトタイプを完成させた。柔軟な SQL 表現は、SML＃が世界で初めて達成し 2011 年の関数型言語のトップコンファレンス ACM ICFP にて発表した関数型言語への SQL 問い合わせ言語のシームレスな統合の機構を、今回のプロジェクトでの実際の使用実績を踏まえ、その理論とコンパイル技術ともに大幅に拡張したものである。具体的には、SQL 式の SELECT 節、FROM 節、WHERE 節を独立の高階多相型関数として表現し、SQL 式を、これら関数の合成として自由に定義可能にし、さらに、それら組み合わせに対して正確な多相型を推論し、型推論が成功した型の正しい SQL 式を、データベースサーバーに送るためのコマンドにコンパイルする技術である。

4.5.2 JSON データの高水準かつ型安全なアクセス方式

JSON 等の外部データの高水準かつ片安全なアクセス方式は、平成 26 年度から開始した WEB アプリケーションフレームワークの理論と実装技術の研究開発の中心課題の一つである。平成 28 年度は、平成 27 年度にほぼ完成をみた理論と実装技術、さらにSML＃言語へ実装し実際に開発環境を日立ソリューションズ東日本でのお薬手帳の次世代版プロトタイプの開発の使用に供し、その機能の検証や評価を行い、その過程で明らかになった種々の拡張を行った。とくに、27 年度のシステムでは実現していなかったレコードやリストからなる複合データを、JSON に変換する多相関数機能を、動的型付けの理論を使い実装した。この 26 年以来開発を続け、完成させた本項目の研究成果は、オブジェクト指向プログラミング分野の最高峰の国際会議である ECOOP2016 に発表された[1]。

4.5.3 関数型言語を用いたお薬手帳システムの開発のサポート。

平成 27 年度は、日立ソリューションズ東日本は、本プロジェクトの一貫で、次世代のより堅牢な WEB アプリケーション基盤の構築にむけ、SML＃を中核とする高機能プログラミングフレームワークによるアプリケーション開発基盤の開発を、お薬手帳システムをターゲットとして実施した。

1の機能を装備したSML＃を、2016 年 7 月に Version 3.2.0 版としてリリースするとともに、このSML＃を中核とする開発環境を日立ソリューションズ東日本へ提供した。

さらに、本基礎研究をよりスムーズに反映し、産学連携のシナジーを達成するために、日立ソリューションズ東日本とほぼ週毎の技術ミーティングを開催し、SML＃の新機能を使いこなすノウハウの提供、システム設計やプログラムコードのレビューを実施し、先端のプログラミング技術を活かした開発のサポートを実施した。

5） 耐災害性ストレージ基盤の実証試験

5.1 概要

本実証実験は、東日本大震災を想定した甚大広域災害に対して開発する高可用性ストレージシステムの性能を実証することを目的として、高可用ストレージにアクセスするスマートフォンアプリを実証実験用に新たに開発して試験するものである。その実験条件として、

1. 東日本大震災クラスの広域甚大災害を想定し、現実的な被災モデルと実験シナリオを策定。
2. 実際の被災時に情報保全の重要性が示されている投薬情報システム（いわゆる「お薬手帳」）を対象とし、電子化を想定した実験用ウェブアプリを開発する。
3. 半数の拠点ストレージが損壊する被災状況でも情報の90%は発災直後からアクセスできる定量目標を設定。
4. インターネットが停止してクラウドへのアクセスができないことを仮定する。企業内や地方自治体のイントラネットがつながっていることが望ましいが、一切のネットワークが停止した状態でも3時間以内に情報にアクセスできること。
5. まず一次実証として沿岸部の10万人規模の都市を想定し、最終的には二次実証として仙台市規模の100万人都市での高アクセス負荷条件でも目標を達成できること。

の条件を課した。

実際の運用時に近い水準で可用性を評価するためには、災害時に被災者に利用されるサービスをストレージ基盤上に試作し、ユーザによるシステムの利用や拠点の被災状況に関して可能な限り実際の状況を反映した条件下で実験することが求められる。そこで、実証実験仕様の策定から被災者の薬歴を管理する投薬情報システムの試作を行い、実際の公開ハザードマップなどを参照して現実性の高い被災シナリオを策定した。特に投薬情報システムの試作では、高可用ストレージ基盤が近隣の拠点間でデータを複製し合うという特長を活かし、災害により拠点損壊や広域通信障害が発生しても被災者が自身の薬歴を参照可能なシステムとして構築し、開発したストレージシステムの可用性を検証し目標の 50 %の機器が損壊した場合でも 90 %の可用性を実証した。

5.2 被災モデル化と実証実験仕様の策定

最終目標である仙台市規模(100 万人)ユーザによる投薬情報システムの利用を通したストレージ基盤の可用性実証実験を実施するため、まず実験構成の検討やユーザのシステム利用シナリオ、ストレージ基盤の被災シナリオの作成、および評価指標を設計した。実験構成の検討では、仙台市内にある 100 拠点のストレージ間でデータがリスクアウェア複製された場合の可用性を検証するため、投薬情報システムの運用を想定してユーザの薬歴データを格納する仙台市の 100 の医療機関(60 の病院と 40 の診療所)を選定し、当該拠点にストレージが配備されて 100 万ユーザがシステムを経由してデータの格納や取得を行う状況を実現するための物理的なシステム構成を明確にした。

図 5-1. 100 万ユーザによる投薬情報システムの利用状況

システム利用頻度シナリオの作成では、 東日本大震災が発生した 2011 年 3 月を含む 2011 年 1 月～5 月における宮城県の処方箋発行枚数に基づき、仙台市 100 万ユーザによる各月の薬歴の登録回数と参照回数を設定した(図 5-1 参照)。

2011 年 1 月～2 月は震災前の平常時のシステム利用状況を表し、3 月は災害発生時の利用状況を、4 月～5 月は復旧途上の利用状況を表す。災害発生直後にアクセス頻度は最大で 20 %ほど増えている。

被災状況とシステム使用のシナリオ策定

利用シナリオ作成では、図 5-1 に示す月次のシステム利用状況をもとに、より実際に近い利用状況を再現できるよう日次、および各時間帯での登録・参照回数への詳細化も検討した。日次の変動としては、多くの医療機関で休み明けと休み前は他の曜日と比べて多数の患者が通院する傾向が見られる。そのため、上記の変動を考慮した比率で月次の登録・参照回数を日次に按分した。各時間帯の変動としては、開院直後の朝の時間帯に最も通院患者が多く、その後日中にかけて減少するものの午後にまた増加するような傾向が見られる。このような変動を考慮した比率で日次からさらに各時間帯の登録・参照回数に按分するようにした。なお各時間帯の登録や参照の発生タイミングは、按分された値を平均とするポアソン過程にしたがうものと設定した。

ストレージ基盤の被災シナリオの作成では、内陸型地震と海溝型地震の 2 パターンの広域災害による仙台市 100 拠点の被災状況を検討した。災害による被災対象は各拠点と拠点間のネットワーク、およびインターネットの 3 つとし、最終目標である「広域災害により半数の拠点が損壊しても 90%の可用性」を検証可能な被災シナリオとした。実証実験では、システムの利用シナリオと

ストレージ基盤の被災シナリオを組み合わせ、100 万ユーザが利用シナリオに沿ってシステムにアクセスしながら、災害が発生したら被災シナリオにしたがってストレージ基盤の被災状況を設定していくことで、実際の運用時に近いレベルで可用性を評価できる。評価指標の設計では、利用シナリオにしたがって 100 万ユーザがアクセスする月ごとの可用性指標を定めた。参照リクエストの失敗は被災によってデータが喪失して薬歴の参照ができなかったリクエストを意味する。実証実験では、災害が発生した月であっても 90%以上となるかどうかを検証する。作成した実証実験仕様については、東北大学医学部付属病院や宮城県薬務課、宮城県薬剤師会の有識者からのレビューをいただき、内容の改善を図った。以上のことから、仙台市規模(100 万)ユーザによる投薬情報システムを用いたストレージ基盤の可用性実証実験の仕様を策定した

5.3 投薬情報システム(電子お薬手帳ウェブアプリ)の開発

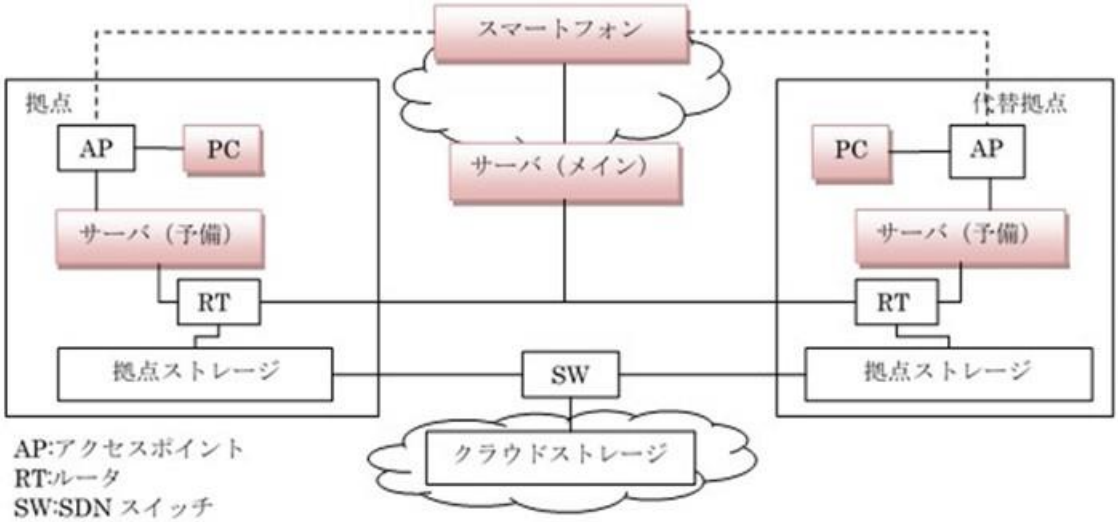


図 5-2. 投薬情報システムのハードウェア構成（赤い四角が開発対象）

実証実験に用いる題材として、高可用ストレージ基盤上に投薬情報システムを開発した。平成 24 年度にシステムの機能設計を実施し、平成 25 年度に内部設計と実装、およびストレージ基盤との接続テストを実施した。機能設計では、投薬情報システムのユースケースを分析して薬歴の登録や参照などの主機能を明らかにし、主機能を実現するためのシステムアーキテクチャやデータモデル、および画面等を設計した。特にシステムアーキテクチャ設計では、ストレージ基盤の特長を活かして高可用なシステムを実現するハードウェア構成と、セキュリティや結合度を考慮したモジュール構成となるよう工夫している。

システム構成

ハードウェア構成を図 5-2 に示す。投薬情報システムは、クライアント／サーバ／ストレージ基盤の 3 層構成となっている。クライアントはスマートフォンアプリおよび拠点内 PC のブラウザである。スマートフォンアプリはユーザが通常使用するクライアントであり、ブラウザはユーザが災害で負傷したなどスマートフォンを利用できない緊急時に医療従事者等により利用されることを想定した。

Web アプリケーションサーバはクラウド上のメインサーバと、各拠点にある予備サーバから構成される。インターネットに接続できる平常時には、ユーザはクラウド上のメインサーバにアクセスして薬歴の登録や参照を行う。メインサーバはユーザのデータを保持する拠点ストレージを特定して格納する。格納されたデータはストレージ基盤により複製先拠点に複製される。一方、災害によって拠点が損壊した場合には、メインサーバはユーザの複製データを保持する代替拠点ストレージを特定してそこからデータを取得する。さらにインターネットが途絶した通信障害時には、ユーザ自身がデータを保持する近隣の拠点に直接移動し(図 5-2 点線部分)、拠点内の LAN 経由で代替サーバにアクセスして薬歴を参照する。このように投薬情報システムは、近隣拠点間でデー

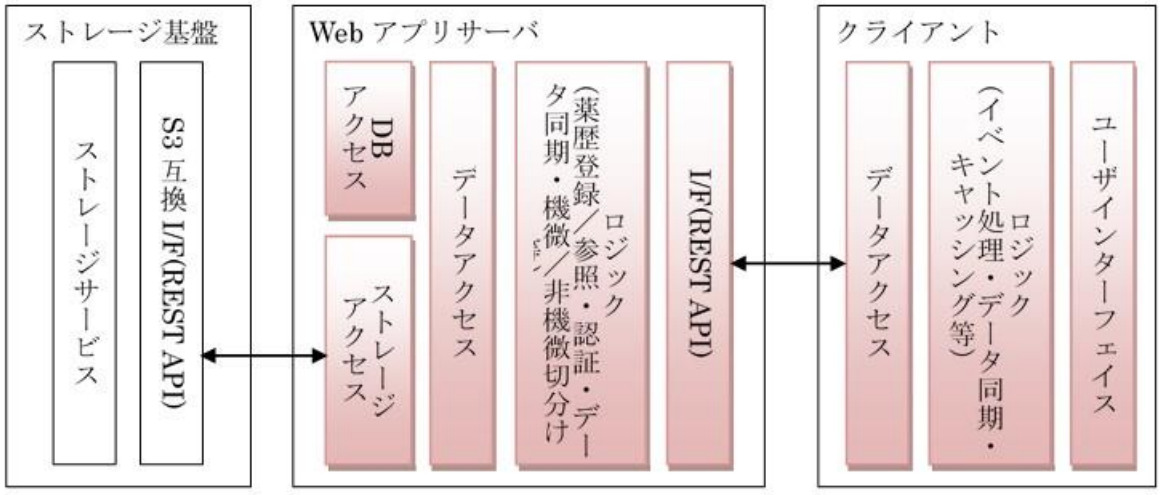


図 5-3. 投薬情報システムのモジュール構成（赤い四角が開発対象）

タを複製し合うストレージ基盤の特長を活かし、大規模災害で拠点損壊や広域通信障害が発生してもユーザが薬歴を参照可能な耐災害情報システムとなっている。

次にシステムのモジュール構成を図 5-3 に示す。クライアントとサーバの間およびサーバとストレージ基盤の間は、REST API による一般的なプロトコルを用いたデータの送受信で実現する設計としており、各層の独立性を高める構成とした。また、サーバ側のロジックでは、クライアントから送信されるデータの機微／非機微を判定し、判定結果に応じて格納先を切分ける機能を設計した。個人の薬歴のような機微データは情報漏洩リスクが小さいストレージ基盤に格納し、拠点の情報などの機微でないデータはローカル DB に格納することで、システムの情報セキュリティレベルを向上させている。

5.4 10万ユーザによるアクセスを想定した第一次実証実験

5.4.1 第一次実証試験の実施

10 万ユーザが投薬情報システムを利用する場合の高機能高可用性情報ストレージ基盤の耐災害性実証試験を実施した。仙台市および周辺自治体にある 24 の医療機関に拠点ストレージが配備された状況の下で、大規模な自然災害により半数の拠点ストレージが損壊したときの耐災害性を評価した。

実証試験で考慮した自然災害は、拠点ストレージ配備エリアで今後 30 年以内に発生しうる全ての地震災害と、津波連動型宮城県沖地震の 2 つである。本報告では、発生しうる全ての地震災害による半数拠点損壊時の耐災害性実証を中心に成果を記述する。評価期間はインターネット接続が概ね復旧する被災後から 5 日間とした。東日本大震災では、被災後 5 日間でスマートフォンによる主要なポータルサイトのページビューが震災前と同程度以上に回復している。ストレージ基盤の耐災害性評価指標には、ユーザによる薬歴参照リクエストの成功率として定めるデータ可用性と、各リクエストのターンアラウンドタイムと拠点までの移動時間の和として定めるデータ取得時間の 2 つを用いた。各指標の目標値は、大規模災害による半数拠点ストレージ損壊および広域通信障害が発生した場合に、データ可用性が 90%以上、データ取得時間が 3 時間以内と設定した。第一次実証試験では、プログラムを用いた理論上の耐災害性評価と実機を用いた観測上の耐災害性評価の 2 種類の試験を実施した。

理論上の耐災害性評価は、地震災害によって生じうるあらゆる半数拠点損壊パターン（24 拠点の内どの 12 拠点のストレージが損壊するかの組合せ）を被災シナリオとして考慮し、各損壊パターンの発生確率による重みを付けたストレージ基盤の統計的耐災害性を求めるために実施した。本試験では実証環境をプログラムとして再現するソフトウェアを別途作成し、当該ソフトウェアを利用した理想的な環境の下で試験した。本試験の前提条件を表 5-1 に示す。

表 5-1 耐災害性実証試験の前提条件

項　目	内　容
利用ユーザ数	10 万ユーザ（仙台市規模の 1 割）。
拠点ストレージ	仙台市と周辺自治体にある医療機関 24 拠点に設置
実験期間	発災時から 5 診療日
ネットワーク可用率	発災時 発災直後から日ごとに 52%→52%→79%→100%→100%と推移
データ複製	リスクアウェア複製とランダム複製の 2 つ。複製数はともに 1。
被災シナリオ	24 拠点のうち 12 拠点が損壊するパターン 2,704,156 通り全てを使用。シナリオ再現期間中は、損壊状況は変わらない。
利用シナリオ	シナリオ再現期間中に約 3 万回の参照が均等間隔で発生。どのユーザが参照するかは一様ランダムに決定。参照ユーザがインターネット途絶対象の場合はデータ格納拠点まで移動。
評価指標	データ可用性とデータ取得時間。

5.4.2 第一次実証実験の成果

理論上の耐災害性評価から得られた 2 つの評価指標の確率分布を図 5-4 に示す。

左のデータ可用性分布の結果から、提案手法であるリスクアウェア複製は従来手法であるランダム複製よりも高い可用性で頻度が高い傾向が見て取れる。実際に、データ可用性の平均と標準偏差はリスクアウェア複製で 0.809 と 0.071、ランダム複製で 0.756 と 0.045 となり、t 検定により 1%水準での有意差が認められた。一方、データ可用性の目標値である 0.9 で評価すると、リスクアウェア複製であっても可用性平均は目標値を下回る結果となった。しかし、リスクアウェア複製では、半数拠点が損壊する地震災害が 100 回発生した場合にそのうち 16.5 回は 0.9 を超える可用性が得られると期待でき、これはランダム複製の 16.5 倍も高い。

また、右のデータ取得時間分布の結果から、近隣拠点ストレージ間でデータを複製しておけば、インターネット途絶が最も深刻な災害直後であっても拠点への移動を含め 22 分以内でデータを取得できることがわかる。この値はデータ取得時間の目標値である 3 時間以内を十分に達成している。右図からは、災害後のインターネット接続の復旧に応じて、移動するユーザの数が減少してデータ取得時間が短縮されていくこともわかる。なお、ランダム複製の方がより近隣の拠点にデータ複製される場合が多いため、リスクアウェア複製よりもより短時間でデータの取得が可能となっている。

2 つ目の観測上の耐災害性評価は、膨大な半数拠点損壊パターンの中から比較的発生確率の高いパターンを被災シナリオとして抽出し、10 万ユーザがシステムを利用した場合の耐災害性を求めるために実施した。本試験では抽出した 5 つの被災シナリオそれぞれに対してダミークライアントプログラムによる利用シナリオに沿った 10 万ユーザのアクセスを実行した。そのうち 1 つの被災シナリオでは 10 万のダミーユーザに加えて実ユーザ 9 名も参加し、インターネット途絶対象者としてデータ格納拠点への移動と電子お薬手帳アプリを用いた薬歴参照を行った（ただし、9 名中 3 名はヒアリングでの参加）。

本試験の詳細な前提条件は、表 5-1 に示すのと概ね同じ項目であり、相違点は利用シナリオの参照が均等間隔ではなくより現実の発生状況に近いポアソン過程で生起するようにした点、被災シナリオを 5 つに限定した点、データ複製をリスクアウェア複製のみとした点の 3 つである。

観測上の耐災害性評価の結果を表 5-2 に示す。データ可用性観測値は被災シナリオ 5 を除いて 90%以上の可用性が得られることがわかる。左列には理論上の値としてストレージ基盤のデータ残存割合も併記しているが、全てのシナリオで観測値とのわずかな差異が認められた。これは、ポアソン過程により参照間隔が短く決定されて一時的にリクエストが集中したことによるクライアント側のタイムアウト発生が主要因と考えられる。また、データ取得時間の観測値については 5 つの被災シナリオで 3 時間以内となった。しかし、表に示した観測値はシナリオ再現期間に行われた全参照のデータ取得時間の平均値であるため、実際には取得に 3 時間以上要した場合もあった。特に被災シナリオ 5 で実ユーザが実際にデータ格納拠点まで移動する際に 3 時間以上かかることが多かった。これは、現状の利用シナリオで定めたダミーユーザのデータ格納拠点まで

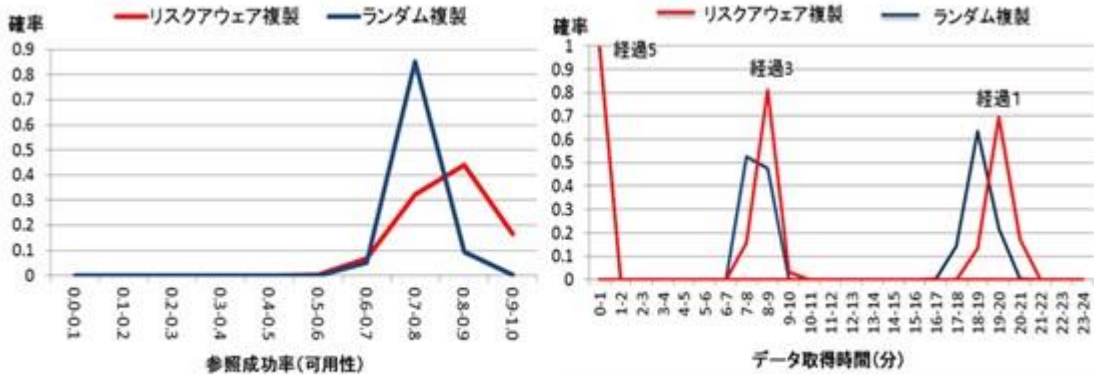


図 5-4 理論上の耐災害性評価結果(左:データ可用性、右:データ取得時間)

の移動時間と、実ユーザが徒歩や車でデータ格納拠点に移動した場合の移動時間の差異が大きいために生じている。第二次実証試験に向けて実際の移動経路や交通状況を踏まえたより現実に近い移動時間を利用シナリオに定める必要がある。また, 被災シナリオ 5 ではデータ残存率が 90%に達していない。このため, 発生確率の高い被災シナリオで 90%以上のデータ残存率を実現できるように複製方式の改善が必要であることもわかった。

表5-2 観測上の耐災害性評価の結果

被災シナリオ	基盤のデータ残存率	データ可用性観測値	データ取得時間観測値
平常時シナリオ(参考)	100.00%	99.99%	0.17分
被災シナリオ1	100.00%	99.98%	13.76分
被災シナリオ2	100.00%	99.97%	29.59分
被災シナリオ3	91.67%	91.62%	13.35分
被災シナリオ4	91.67%	91.16%	33.89分
被災シナリオ5 (実ユーザ9名も参加)	83.33%	83.25%	103.11分

以上により、ダミークライアントプログラムを利用して 10 万ユーザが投薬情報システムを利用する状況を再現し、発生確率の比較的高い複数の被災シナリオに対して 90%以上のデータにアクセスできることを実証した。また、投薬情報システムのアーキテクチャや移動時間に関する利用シナリオ、データ複製方法の改善点を明確にした。

5.5 第二次実証試験計画の立案と実証試験環境の改良

投薬情報システムを用いた高機能高可用性情報ストレージ基盤の耐災害性実証試験の第二次実証試験の試験計画の見直しを行った。第一次実証試験の結果を踏まえて、被災シナリオや評価方法などの試験計画を見直し、実証試験仕様を修正した。

5.5.1 被災モデルの改良

第一次実証試験では、防災科学研究所の J-SHIS で公開されている確率論的地震動予測地図により今後 30 年間に発生しうるすべての地震を考慮した地震動の超過確率を使用し、被災シナリオを設定した。これにより災害の網羅性を考慮できるが、すべての地震を考慮した地震動の超過確率を用いるため、1 つの地震により発生する地震動の強さを表していない。つまり、震源域の違う個々の地震についての拠点損壊確率の和集合の値を用いることになり、局地性震源では同時損壊が起こり得ない拠点の組合せの損壊が評価対象に含まれる課題があった。

また、第一次実証試験では、すべての半数拠点損壊パターンによる統計的評価とストレージ拠点を複製する手法の比較のために、各シナリオや拠点間の複製関係をプログラムによりシミュレートする理論値計算を実施した。本プログラムによりデータ可用性の期待値を計算したところ、す

べての半数拠点損壊パターンによるシミュレーションでは震源域の局地性の視点からは同時損壊が起こり得ない拠点の組合せの損壊が評価対象に含まれ、期待値が 90%に達しない可能性があるという課題が生じた。第二次実証試験では、前述の「すべての地震」の他に、海溝型地震の震源，内陸活断層型地震の震源について地震と津波による拠点の損壊確率を用いて、震源ごとに損壊確率の最も高い半数拠点損壊パターンを選択することとした。

5.5.2　1を超える最大複製数の導入

半数拠点損壊時に 90%以上のデータ可用性を実現するための複製数及び複製アルゴリズムの決定を目的として可用性理論値の計算実験を実施した。実験条件を表 5-3 に示す。表 5-3 の条件ですべての半数拠点損壊パターンのデータ可用率をプログラム計算で求め、データ可用率の期待値を求める。その結果を第 2 次実証試験の複製数・複製アルゴリズムの選定に利用する。

表5-3 理論値計算実験の条件

項　　目	内　　容
拠点ストレージ数	24 拠点
損壊拠点数	半数拠点損壊(12 拠点)
複製アルゴリズム	リスクアウェア複製(拠点間距離) リスクアウェア複製(同時被災確率加算)
1 拠点当たりの複製数	1, 1.5, 2, 2.5

表5-4 データ可用性の期待値

	同時被災確率加算方式	拠点間距離
最大複製数 1.0	0.879	0.759
最大複製数 1.5	0.939	0.870
最大複製数 2.0	0.963	0.907
最大複製数 2.5	0.969	0.280

ストレージ基盤が採用しているリスクアウェア複製ではリスクの算出方法により，データの可用性に影響がある。そこで，リスクの算出に拠点間距離を用いる方法と同時被災確率加算方式の 2 パターンを使用し、データ可用性を評価した。計算結果を図 5-5 と表 5-4 に示す。どの複

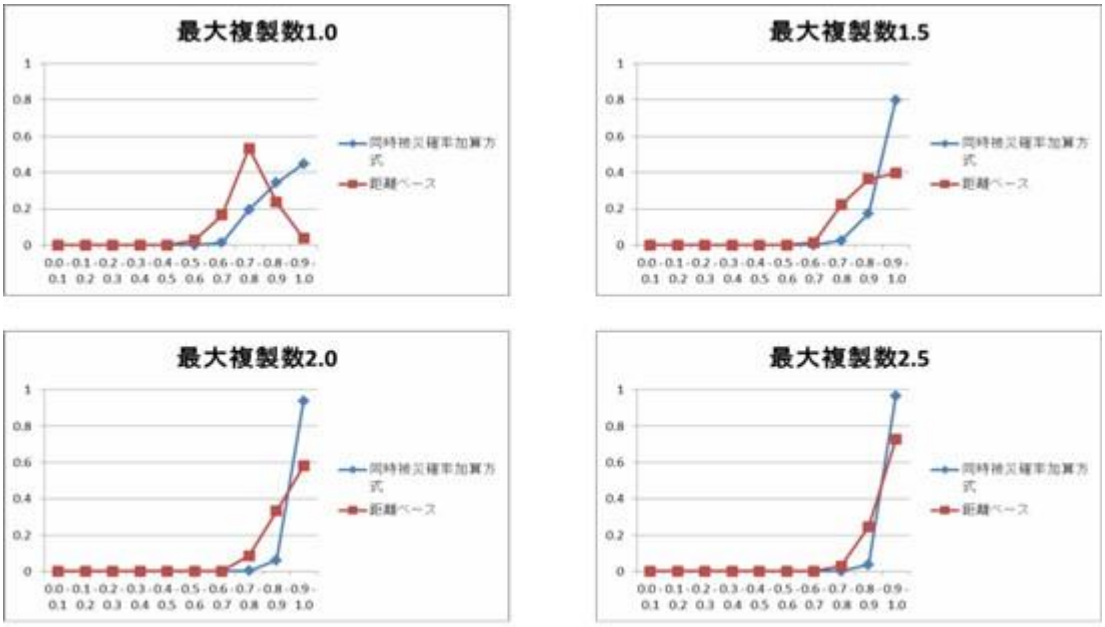


図 5-5 複製アルゴリズムの評価

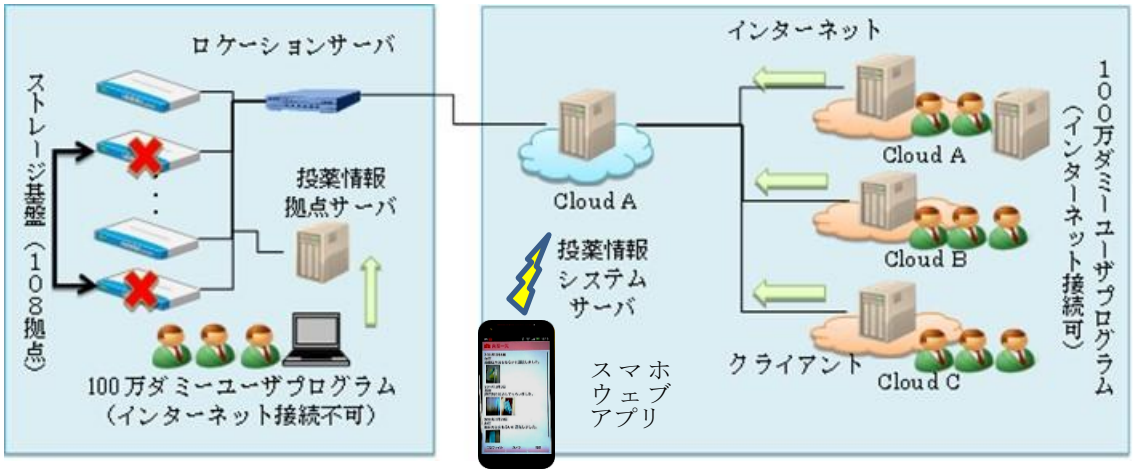


図 5-6 第二次実証試験の実験環境

製数でも同時被災確率を加算したリスクを用いた複製方式がより高いデータ高可用性を示している。また，同時被災確率加算方式の複製アルゴリズムを利用すると、複製数が 1.5 以上であれば、可用性の期待値は 90%以上となることがわかった。一方、拠点間距離のリスクを用いた場合、最大複製数 2.0 で期待値は 0.9（90%）を超える。以上の結果により、第二次実証試験では以下の複製アルゴリズムと最大複製数を用いることとした。

試験計画の見直しを反映するため投薬情報システムを改良し、第二次実証試験で実施する 100 万人規模のユーザの利用を想定し実験環境を構築した。第二次実証試験の実験環境を図 5-6 に示す。インターネット上のクラウド上には投薬情報システムサーバと、クライアントとしてサーバへのデータアクセスを再現するダミーユーザプログラムを配置する。東北大学学内の実ネットワークに 108 の拠点ストレージ基盤と拠点ストレージを制御するロケーションサーバのほか、災害によるインターネット途絶時のデータアクセスを想定した投薬情報システムの拠点サーバと、拠点サーバへのデータアクセスを再現するダミーユーザプログラムを配置する。被災シナリオにより設定した半数のストレージ基盤を停止させることで拠点損壊をシミュレーションする。

5.6 第二次実証試験の実施

第二次実証試験では大規模災害時の発災直後から 5 日間の 100 万人規模のユーザの利用をシミュレーションし、108 のストレージ拠点の半数が損壊した場合でも 90%以上のデータにアクセスでき，データ取得時間が 3 時間以内であることを実証した。

5.6.1 実験方法

第二次実証試験は，東日本大震災で甚大な被害を受けた仙台市および周辺地域(以下実証フィールドと記す)で高機能高可用性情報ストレージ基盤および投薬情報システムが利用されることを想定し、表 5-5 に示す条件で実施した。発災からの時間経過とともに、設備や電源の復旧により向上していくネットワーク可用率については東日本大震災における実際の被災地域でのネット

表 5-5 実験条件

項目	内容
利用ユーザ数	実証フィールド内の 100 万人のユーザ(仙台市規模)
拠点ストレージ	実証フィールド内の 108 の医療機関に設置
実験期間	発災時から 5 診療日
ネットワーク 可用率	平常時 100% 発災時 発災直後から日ごとに 52%→52%→79%→100%→100%と推移
データ複製	リスクアウェア複製，平均複製数ベース反復方式，平均複製数 1.5
データ参照回数	平常時 30 万回，災害時 35 万回
損壊拠点数	54 拠点(108 拠点の半数)

ワーク可用率の推移に基づいて決定した。データ参照回数は主たる実証フィールドである仙台市の処方箋発行枚数、外来患者の平均通院間隔から推定した。

5.6.2 被災シナリオ

第一次実証試験で使用した被災シナリオには震源域の局地性の視点からは同時損壊が起こり得ない損壊拠点の組合せが評価対象に含まれるという課題があった。この課題を解決するために、第一次実証試験で使用した「今後 30 年以内に起こりうる地震の重ね合わせ」のほかに、海溝型地震や内陸活断層の地震など、震源域の局地性を考慮した災害を評価するためのシナリオを追加し，計 15 シナリオで実証することとした。被災シナリオを表 5-6 に示す。

表 5-6 被災シナリオ

#	視点	種別	被災シナリオ
1	網羅性	すべての地震	今後 30 年間に起こりうる地震の重ね合わせ
2		内陸活断層	活断層の地震の重ね合わせ
3	局地性	海溝型地震	プレート間大地震・福島県沖地震
4			東北地方太平洋沖型地震
5		海溝型地震＋津波	プレート間大地震・福島県沖地震＋津波
6			東北地方太平洋沖型地震＋津波
7		内陸活断層	福島盆地西縁断層帯
8			利府長町線断層帯
9			長井盆地西縁断層帯
10			双葉断層
11			旭山撓曲帯
12			遠刈田断層帯
13			愛島推定断層
14			作並屋敷平断層帯
15			山形盆地断層帯南部
16	ー	ー	平常時(損壊無)

5.6.3 データ可用性の評価

データ可用性の実験結果を表 5-7 に示す。実験の結果、半数の拠点が損壊した状況で、データ可用性は 94%～100%となり、全ての被災シナリオで目標値の達成を確認した。実験結果の可用率と理論上導き出される値との差が生じることがあった。この理由は、参照ユーザがランダムに選択されるため、取得拠点が必ずしも均等には選択されないためである。また、今回の実験で

表 5-7 データ可用性の実証実験結果

#	被災シナリオ	データ可用性	
		理論値	実験値
1	今後 30 年間に起こりうる地震の重ね合わせ	100.00%	99.96%
2	活断層の地震の重ね合わせ	100.00%	99.66%
3	海溝寄りのプレート間大地震と福島県沖地震	100.00%	100.00%
4	東北地方太平洋沖型地震	100.00%	99.67%
5	作並屋敷平断層帯	94.44%	94.45%
6	山形盆地断層帯南部	96.30%	96.04%
7	福島盆地西縁断層帯	96.30%	95.86%
8	利府長町線断層帯	97.22%	97.06%
9	長井盆地西縁断層帯	97.22%	97.18%
10	双葉断層	97.22%	96.88%
11	旭山撓曲帯	97.22%	96.99%
12	遠刈田断層帯	97.22%	97.19%
13	愛島推定断層	98.15%	97.87%
14	海溝寄りのプレート間大地震と福島県沖地震＋津波	100.00%	99.76%
15	東北地方太平洋沖型地震＋津波	100.00%	99.72%
16	平常時(損壊無)	100.00%	100.00%

は、ネットワークが途絶したユーザが拠点に移動した際にアクセスする、3 台の投薬情報システムの予備サーバを 1 台の PC 上の仮想マシンで実行している。このため、リクエスト数が多い時間帯で仮想マシンの CPU が高負荷となり、リクエストを処理できずに情報を提供できないことも原因であることがわかった。災害時も低スペックのサーバを運用する場合、輻輳のため求められる性能を満たさない可能性があるため、拠点で運用するサーバの選定には留意が必要である。

5.6.4 データ取得時間

シナリオごとのデータ取得時間の分布を図 5-7 に示す。データ取得時間の評価ではリスクを統計的に評価する際に利用されるバリューアットリスクと期待ショートフォールを用いた。データ取得時間の 99.5%バリューアットリスクは 1 時間 37 分 18 秒～2 時間 49 分 25 秒となり、99.5%の信頼水準で、3 時間以内にデータが取得できることを確認した。また，期待ショートフォールも 3 時間以内である。これらの評価により、全ての被災シナリオで、データ取得時間は目標値である 3 時間以内を達成したことを確認した。

第 2 次実証実験では情報ストレージ基盤を構成する拠点の半数損壊時の耐災害性について、実際のストレージ基盤を用いた実験環境上で、東日本大震災クラスの広域甚大災害をできるだけ精密に想定した 15 の被災シナリオを用いた実証実験を実施した。その結果、ストレージ基盤の対災害性の目標値であるデータ参照成功率（可用性）が 90%以上であることと、ユーザが 3 時間以内にデータを取得できることを確認した。このことから、情報ストレージ基盤が高い耐災害性を持つと評価した。

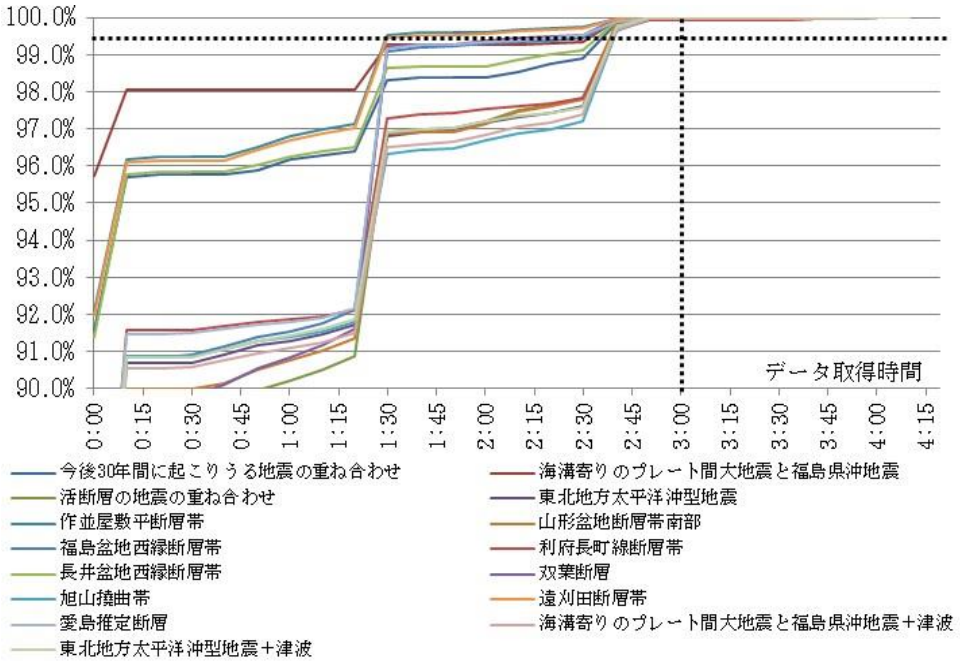


図 5-7　シナリオごとのデータ取得時間の分布

⑤ 独創性・優位性について

情報ストレージの高可用性化はこれまでも重要な技術課題であり様々な検討が続いてきた。しかし、東日本大震災では深刻な情報喪失が発生してその技術的な不備が露呈した。これは、従来技術の想定を超える大規模な広域甚大被災だったことに起因している。建屋ごと、あるいは地域ごと損壊・流失し長時間の復旧を要する事態はこれまでは現実的なものとは考えられなかった。この従来技術の不備を補う取組み自体に本課題の独自性がある。一方で、高速ストレージ装置とネットワークのデータ転送レートの高速化についても特徴ある技術開発を行った。ストレージ、ネットワーク、ソフトウェアの関連分野で学術的にも幅広く成果があった。

1） 耐災害性高可用ストレージシステム

従来のストレージの主な可用性向上技術として RAID 技術とクラウドによるディザスタリカバリ技術がある。RAID 技術は同じストレージ装置内にある複数のディスク装置にデータを分散複製させる方式である。同技術は同一装置群内の数台～十数台でデータ保全を行うため、これを超える規模の被災では機能せずデータを失う。一方、クラウドによるディザスタリカバリ技術は、バックアップデータを遠隔地にあるデータセンタに格納し、被災時には広域ネットワーク経由でデータを送受信するため、甚大災害でしばしば発生するインターネットや広域ネットワークの障害が発生すると、データ自体は消失しないがデータへのアクセスが不能となる。中小のユーザにとっては導入が高コストになる傾向もある。

さらに可用性向上に関する最近の技術動向として、Google File System や Cassandra があげられる。Google File System はラック(電源境界)を意識して複製を行うことができるが、この技術ではデータセンタ全体が被災するような大規模災害には対応できない。Cassandra はラック(電源境界)に加えてデータセンタを意識して複製を行うことができる。データセンタを意識した複製を用いた場合でも、複製先は遠隔地へのデータセンタを想定しており、広域ネットワーク障害ではアクセスが不能となる。地域ネットワークで接続する近隣のデータセンタへ複製する運用も可能だが、同時被災してデータが消失する可能性がある。

これに対して、本課題で提案しているリスクアウェア複製機能を有する分散ストレージシステムは、災害リスクを考慮することで近隣に位置しておりながら安全である拠点を見出して複製を行うことで可用性を向上させるもので、国際的にも従来なかった独創性の高い技術である。本技術は広域ネットワーク障害下でも被災直後からデータへのアクセスもしくは復元が可能であり、可用性を向上させる優位性がある。オペレーションズリサーチを応用し、拠点災害リスクを考慮した確率的アプローチによる新規な複製先決定アルゴリズムは学術的にも価値がある。先に述べた従来技術である Cassandra を地域のデータセンタ群へ適用した場合と比較すると、複製数 1、半数のデータセンタが損壊の条件において従来技術は 75%のデータにしかアクセスできないが、提案技術は 90%以上のデータにアクセス可能である。

2） ストレージシステムの高速データ転送

本課題では、複数のトラックを一括して再生することで単独トラックでは限界のある転送レートを増加できる点に優位性があり、特に、これまで実用されたことのない複数トラックを同時に読み出して転送レートを倍増させる方法を初めて提示した。通常ディスク装置においては複数トラックを同時に再生すると隣接トラックの信号が重畳される強い信号干渉を生じる。これは通信路でのクロストークに相当するもので、一般的には復号不能に陥る。本方式はこの点を回避できる点で学術的な独創性があり、このための複数のトラックの記録情報を重畳して再生することを前提に2次元符号化による符号化と書き込みを行い、それを同時再生する点が新規である。

2次元符号化自体はこれまで広く研究が進められているが、トラック間の信号干渉を軽減して記録密度を向上させることが目標であり、複数トラックを単一ヘッドで読み出して、重畳された信号から個々のトラックの情報を分離再生して復号する検討は行われてこなかった。情報量やファイルの容量増は著しく、今後も高速データ転送の引き続き要求が高まるのは間違いがないが、その際に本課題における転送レートの増加はストレージにとって有用である。

3） 高速 SDN 型ストレージ間通信

甚大災害時には経路切断や輻輳など様々な障害がネットワークに発生するため耐災害性を高めるには柔軟性の高い経路制御が求められる。最近になって普及が進んでいる SDN (Software Defined Network)はソフトウェアで適応的にネットワークの経路制御が可能であり、耐災害ネットワークのために有用である。本研究では、この SDN を用いて分散ストレージ間のファイル転送通信に特化した経路制御の高度化を目指している点に特長がある。経路被災を適応的に検出して最も高いスループットを探索する制御手法が重要であり、被災を模擬して適応経路制御により5倍を超える 10 Gbps のスループットを実現した。

4） 高機能プログラミングフレームワーク

高機能プログラミングフレームワークが基礎とする次世代関数型言語 SML#は、SQL のシームレスな統合やCとの直接連携などの先端機能を世界ではじめて達成した我が国発の ML 系高信頼関数型言語であり、日本学術会議の「第 22 期学術の大型研究計画に関するマスタープラン」にても SML#を基礎一つの中核とした高信頼ソフトウェア工学基盤の提案が採択されるなど、我が国の次世代のソフトウェア工学への貢献が期待されている最先端プログラミング言語である。本研究が達成を目指すSML＃への高機能プログラミングフレームワークの統合は、今後重要性を増すビッグデータ分析などを含む高度なデータ処理を要求される次世代ソフトウェアの生産性と信頼性を飛躍的に高めると期待されている。

5） 投薬情報システムを用いた高可用ストレージ実証実験

耐災害性ストレージサーバ群を都市部の実際のネットワーク上に展開し、スマートフォンに対応した試験アプリからアクセスする実証実験は、都市部のユーザが平常時や災害時に実際にアプリを利用している状況の下で可用性を検証できる実用性の高い実験であるという点で優位性が認められる。

（2） 研究開発体制について

ストレージ機器、ストレージシステム、ネットワーク、ソフトウェア、等に経験の深い大学研究者と企業研究者からなる研究開発グループが、東北大学電気通信研究所における集中研方式として垂直統合型プロジェクト体制を構築して研究開発を推進した。

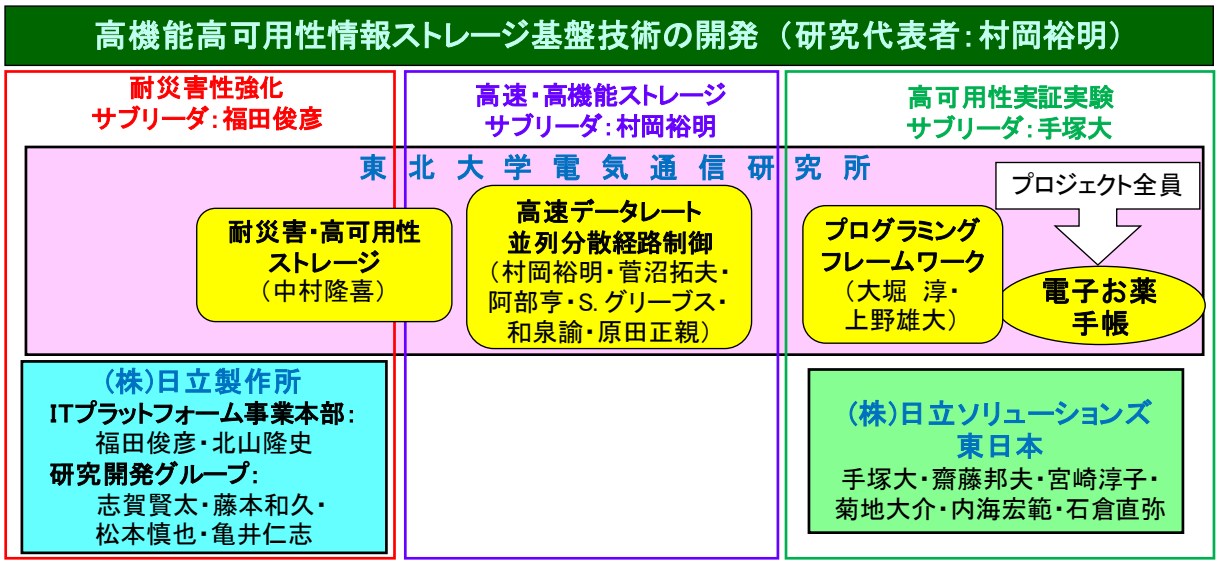
東北大学では、主として各技術分野のストレージ基盤要素技術の研究開発を担当し、日立製作所がストレージ基盤を具体化するためのシステム作りを担当している。また、日立ソリューション東日本では、構築されたシステム・ネットワーク上で動作する投薬情報システムを開発している。この実証実験に関しては、各グループの成果を統合して推進している。

PO の出席のもと、全開発メンバーが出席するプロジェクト推進会議を年2回開催してPOと有識者への報告と指導を仰いだ。一方、プロジェクトレビュー会議を毎月1回開催して日日の実務上の進捗状況・問題点・解決策等を討議し、各グループのより密接な連携の推進を図った。本プロジェクトは地域の支援を重視したプロジェクト実施体制としており、地域産業界が組織した震災復興企業コンソーシアム（東北新生 IT コンソーシアム）など地域活動にも参加して進めている。

下図には分担の連携を示す。学術面の基盤研究と実用性の高い技術開発を並行して進めるため、分

野別技術の横のつながりと各分野内での基盤と応用の縦の連携を兼ね備えた研究開発組織を構成した。

研究実施体制



（3） 成果の波及効果について

① ストレージシステムに関する最大の国際業界組織である SNIA (The Storage Networking Industry Association)の Storage Developer Conference で本課題の高可用ストレージ技術の技術紹介を行っている。今後、国際的な連携に発展できる可能性がある。

② 総務省、内閣府プロジェクト[*1, *2] で開発したネットワークの災害復旧技術について、本ストレージプロジェクトと試作システムの紹介や意見交換を行った。無線ネットワークの緊急復旧は耐災害ストレージシステムの可用性を改善する効果が高く、優れたシナジー効果が期待できる。

*1 総務省プロジェクト 「情報通信ネットワークの耐災害性強化のための研究開発」

*2 内閣府 SIP プロジェクト 「レジリエントな防災・減災機能の強化」

③ 第 3 回国連防災会議（平成 27 年 3 月 14-18 日開催）のパブリック・フォーラム企画「IT とラジオのシンポジウム」に参加し、ポスター及びデモ展示により、耐災害情報ストレージ技術の情報発信を行った。本シンポジウムの参加者は 219 名以上にのぼり、外国人を含む数多くの方々に情報発信を行うことができた。

4. 今後の展望

（1） 研究結果を踏まえた今後の展望、予想される効果・効用

① 今後の展望、予想される効果・効用

1） 宮城県薬剤師会とは被災時の薬局情報復旧について数回の議論を経て共同の実証実験を行っている。今後の耐災害技術の発展への共同での取り組みなどの展開が期待される。

2） 民間企業への技術移転を開始している。民間企業なのでビジネス環境が整うことが必要であるが、それ次第で製品開発に移行できる。共同研究を実施した日立製作所と日立ソリューションズ東日本はプロジェクトの技術成果を直接持ち帰っている。また、技術紹介など日立グループでの広報を行っている。

3） 耐災害ソリューションやストレージ装置の高可用化機能に興味を示した民間企業３社等に技術内容を紹介している。

４）本課題における先進的な要素技術開発は今後のストレージネットワークやソフトウェア技術に波及効果が考えられる。

- 高可用ストレージ技術は複製データを分散保存する技術であり、今後のファイル分散ストレージなどの先端的技術に展開が可能である。
- マルチヘッド記録再生や二層記録層記録など複数のヘッドディスクを用いる磁気記録研究が進展している。同時並列トラック再生は連携可能である。
- プログラミングフレームワークやソフトウェア制御ネットワーク(SDN)の有効性を示すことができた。今後のストレージ技術への波及効果が期待できる。

②事業終了後の実用化や自立的な取組の継続に向けた方策

1） ビッグデータに代表される巨大情報量に対する新しい情報処理を研究するヨッタインフォマティクス研究センターが東北大学で立ち上がっている。ここで、本課題の分散ストレージ技術や高速ストレージ技術を開発技術の展開として継続的に取り組む。

2） 従来のエンタープライズ型の専用大型ストレージシステムから SDS (Software Defined Storage)や IoT など小規模な情報装置をスケールアウトしてシステム化へのイノベーションが進んでいる。本プロジェクトで確立した分散ロケーションのストレージシステムの技術が展開できる。