修 士 論 文

動的ホスト設定機構の信頼性向上に関する研究

指導教官　篠田 陽一 助教授

北陸先端科学技術大学院大学
情報科学研究科情報システム学専攻

小高 英男

平成 10 年 2 月 13 日

目 次

図 目 次

表 目 次

第 1 章

はじめに

インターネットの飛躍的な拡大により、多くのユーザにとってネットワークを利用する機会が増加してきた。しかしネットワークを使うために必要な情報を設定するとき、ユーザはしばしば難ずかしさを覚えたり、戸惑うことがある。

このような不可欠な情報を自動的に設定する仕組みが動的ホスト設定機構である。この機構は、ホストを設定する情報を集中的に管理し配布するサーバとその設定情報を受け取るクライアントから構成されており、広く普及している。しかし現状では、そのサーバは故障に耐える仕組みを備えておらず、サーバが動作している計算機自体が故障したりサーバとクライアント間が通信不能になると、クライアントが動作している計算機はネットワークへの接続性を喪失する。

本研究では、LAN 上に分散した動的ホスト設定サーバ群が、ネットワークの分断やサーバの停止という障害が発生しても、クライアントに信頼性のあるサービスを提供できることを目的とする。

耐故障化の方法として、フォールトトレラント技術の漸次縮退系を本システムに適用する。これはシステム再構成技術の一つで、複数のサーバを用いてシステムを構成し一部のサーバの停止によってシステム全体が機能しなくなることを防ぐという方法である。そして障害が発生した場合、通信できないサーバをシステムから分離する。

しかし冗長なサーバ群がネットワークを介して構築されている事に注意を払わなければならない。なぜならあるサーバが他のサーバと通信ができない状況でも、他のサーバは動作し続けていることがあるからである。これは障害の原因がサーバの停止によるものなのかネットワークの分断によるものなのか区別がつかないためである。

各サーバはそれぞれが保持しているデータベースに基づいてクライアントにサービスを提供している。それゆえ上述の問題を認識せずサーバが勝手に動作するならば、不適切なサービスをクライアントに提供したりサーバの誤動作が発生してしまう。まさに動的ホスト設定機構の世界が混乱してしまうだろう。そこでこのような事態を回避するため、サーバ間ではアドレスを利用する際にサーバ間で合意を取り、データベースの情報を同期させる必要がある。各サーバはクライアントに貸し出したネットワークアドレスの状態を適切に制御しなければならない。

本稿では、 動的ホスト設定機構のサーバに有効な耐故障化の枠組を明確にし、システムの設計を行う。

本稿は 6 章から構成されている。各章の概要は次にようになっている。第 2 章では、本研究テーマに関連する分野として動的ホスト設定機構とそのサーバ間プロトコル、そしてフォールトトレラント技術について知識を整理している。本研究の背景や本稿で扱う用語の定義を確認する。第 3 章では、動的ホスト設定機構に理想的な耐故障化を行うときの要求仕様いて述べる。システムが目指す目標、それに伴う問題と解決するための機能を分析する。第 4 章では前章を踏まえてアドレス制御層に関する設計を行う。ここではアドレス制御層に特化して論じ、難解な問題を浮き彫りにして新たな解決の試みを提示する。そしてネットワークアドレスの状態を制御する設計について述べる。第 5 章では、本稿で提示したモデルと設計を評価する。最後に本研究で達成したことを確認した後、残された課題を第 6 章で述べる。

第 2 章

関連研究

本章では本研究と関連のある研究分野を概観し、後章で必要となる知識をまとめている。最後に本研究の位置付けとして、対象となる問題領域を明確にする。

2.1 動的ホスト設定機構

動的ホスト設定機構 [2, 7] は、各種の設定情報や IP アドレスをネットワークに接続された計算機に自動的に配布する仕組みである。これは 1993 年に RFC[1]1531[3] となったが、その基本的な概念は既に 1985 年に RFC0951[1] として提案されたブートストラッププロトコル (BOOTP) に基づいている。

動的ホスト設定機構は、ネットワーク上の共有資源を集中かつ動的に管理することを可能にしたということに意義があると考える。

構成要素として動的ホスト設定機構は次の二つを持っている：

1. 計算機に設定情報を配布するためのプロトコル
2. 計算機に貸し出す IP アドレスを制御する仕組み

特に上記 1 の要素を動的ホスト設定プロトコル (DHCP[2] と略す) と呼ぶ。

[1] Request For Comments

[2] Dynamic Host Configuration Protocol

2.1.1 モデル

DHCP はクライアントサーバ型のモデルであり、サーバ、クライアント、リレーエージェントという三種類の主体から構成される。それぞれの役割は、

- サーバ
  クライアントの要求に対して、ネットワークアドレスを貸し出したり設定情報を配布する。
- クライアント
  サーバに要求を送り、サーバからの設定情報を取得する。
- リレーエージェント
  クライアントが要求したメッセージがサブネットを越える時、そのメッセージを中継してサーバへ転送する働きをする。リレーエージェントを利用することで、各サブネットにサーバを配置すべき必要性を軽減している。

である。これらの関係図の一例を図 2.1に示す。ここでは、リレーエージェントはルータ上にあると仮定している。

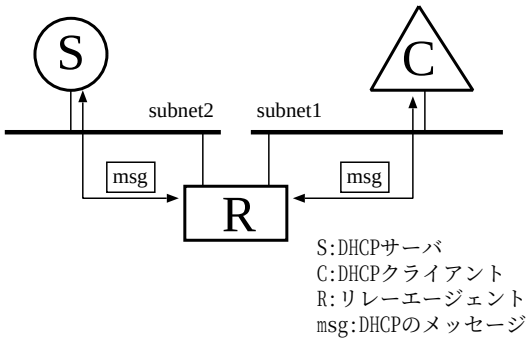

図 2.1: DHCP のモデル

2.1.2 プロトコル概要

DHCP はネットワークアドレスの割当て方式として「自動割当て」、「動的割当て」、「手動割当て」の三種類提供している。「自動割当て」は IP アドレスをクライアントに永久

的に割り当てる方針である。「動的割当て」はDHCPサーバがクライアントに期限付きでアドレスを貸し出すことである。「手動割当て」は、クライアントのIPアドレスはネットワーク管理者によって割り当てられる。

動的割当て方式は、一時的にネットワークを利用するクライアントへアドレスを貸し出す場合やアドレスを恒常的に使用しないクライアント群が限られたアドレスを共有するという場合には特に有効である。

表2.1に示すメッセージがプロトコルでは使われる。

表 2.1: DHCP メッセージ

メッセージ	用　途
DHCPDISCOVER	利用できるサーバ群を特定するために、クライアントがブロードキャストするメッセージ
DHCPOFFER	DHCPDISCOVERに返答するため、サーバが設定情報とIPアドレスを提案をする
DHCPREQUEST	クライアントは次の理由でこのメッセージを利用する。(1) 一つのサーバが提案した設定情報を要求し、他のサーバに対しては暗黙に提案の拒否を意味する (2) 以前に借りたアドレスの正しさを確認する (3) 特定のネットワークアドレスの期限を延長する
DHCPACK	承認したネットワークアドレスを含めた設定情報をサーバからクライアントへ送るとき
DHCPNAK	クライアントのネットワークアドレスが適切でない場合やアドレスの貸し出し期限が切れている場合、サーバがクライアントへこのメッセージを送信する
DHCPDECLINE	クライアントは、他のクライアントが特定のアドレスを既に使用していると認識すると、誤りをサーバに伝える
DHCPRELEASE	クライアントがアドレスの期限が切れる前に明示的にアドレスを返却することを伝える

動作例　図 2.2は、クライアントがサーバから IP アドレスを新規に取得し、アドレスを返却するまでの動作の流れを表している。二つのサーバが存在して、サーバ 1 の方が選択されたとする。DHCP の動作を右に記した数字毎に区切って説明していく。

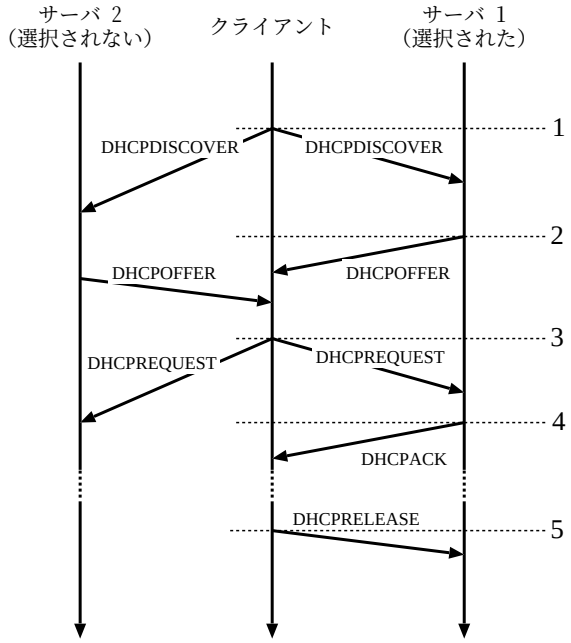

図 2.2: DHCP の動作の流れ

1. クライアントは物理的に繋がっているローカルサブネットに DHCPDISCOVER メッセージをブロードキャストする。

2. このメッセージを受信したサーバは、貸し出し可能な IP アドレスを含めた DHCPOFFER メッセージをブロードキャストする。

3. クライアントは複数の DHCPOFFER メッセージを受け取ることができる。それらのメッセージから一つを選択する。そして選択したサーバの IP アドレスを必ず含めた DHCPREQUEST メッセージをブロードキャストする。

4. DHCPREQUEST メッセージを受信したサーバは、メッセージ中に自分の IP アドレスが記述されていれば選択されたと理解し、それ以外は却下されたと判断する。

選択されたサーバは貸し出す IP アドレスをバインディングと呼ばれるデータベースに記録し、DHCPACK メッセージを返信する。

5. クライアントはアドレスを必要としなくなったら、DHCPRELEASE メッセージをサーバに送信する。

更に詳しいプロトコルの説明は、RFC2131[2] にある。

2.2 フォールトトレラント技術

2.2.1 フォールトトレランス

フォールトトレランスを述べるにあたり、障害、誤り、故障という用語に関してそれらの概念を明確に区別しておく。そして、フォールトトレランスの定義と目標を整理する。

概念定義

- 障害 (failure)
  システムが望まれるサービスを提供できないこと
- 誤り (error)
  システム構成要素の異常な出力であり、障害へと繋がるもの
- 故障 (fault)
  誤りを引き起こす原因

本稿では、上記の定義に基づき各用語を区別して使う。

システムを構成しているある要素に発生した障害をマスクすることがフォールトトレランス (fault tolerance) である [4]。マスク (mask) とは、誤りの発生を外から全く見えないようにすることである。システムに故障が存在したとしても、システムの障害を回避することがフォールトトレランス目的である。

2.2.2 システム再構成技術

フォールトトレランスを実現する基本的な原理は、冗長性を利用することである。主な方式としては、

- 故障の影響を完全にマスクする (静的冗長系)
- 故障要素をシステムから除去または隔離して、正常な予備要素に置換する (動的冗長系)

が挙げられる [8]。

動的冗長系は更に次の二つに分類される。

待機冗長系 通常は１つのモジュールのみを動作させている。故障検出と同時に故障モジュールをシステムから切り離して予備モジュールに切り換えることによりシステムを正常な状態に回復させるものである。

漸次縮退系 故障モジュールを切り離して残った部分システムだけで再構成し、許容された範囲で縮退されたサービスを提供し続けるシステム構成をいう。

2.3 DHCP サーバ間プロトコル

サーバ間プロトコルを考案する動機は、インターネットドラフト[3][5] によると、クライアントがアドレスを借りたサーバと通信できない状況が発生しても、クライアントはアドレスの貸し出し期限を延長できるようにしたいということである。

第 2.1節で述べた DHCP は複数のサーバを独立に扱うことはできる。そのため各サーバが保持するアドレスプール内の IP アドレスは、サーバ間で重複しないということが前提となっている。

そこでサーバ間プロトコルでは、冗長な DHCP サーバ群が協調動作できる仕組みを提供している。主眼としては、各サーバが蓄積しているバインディング情報を自動的に同期させるための機構の構築を目指していると考えられる。

2.3.1 モデル

各サーバはデータベースとしてアドレスプールとバインディングを保持している。共通のアドレスプールを保持しているサーバ群は、グループと呼ばれる。これがこのプロト

[3] これは、IETF(Internet Engineering Task Force) 内の Dynamic Host Configuration Working Group で進行中の議論を文章化したものである。そのため仕様がいつでも変更していく可能性があることに注意されたい。

コルの重要な考え方である。グループの概念図を図 2.3に示す。サーバは複数のグループのメンバーになることが可能である。バインディングデータベースの情報を同期させるため、各サーバは記録された情報が変更されると、グループ内のメンバーと更新情報を交換する。

サーバ間で交換しなければならない情報は次の 3 種類である：

- バインディング情報
  クライアントに貸し出した IP アドレスに関する記録情報
- アドレス管理情報
  IP アドレスの状態情報
- グループ管理情報
  グループのメンバー情報、グループが管理するアドレスプールの決定

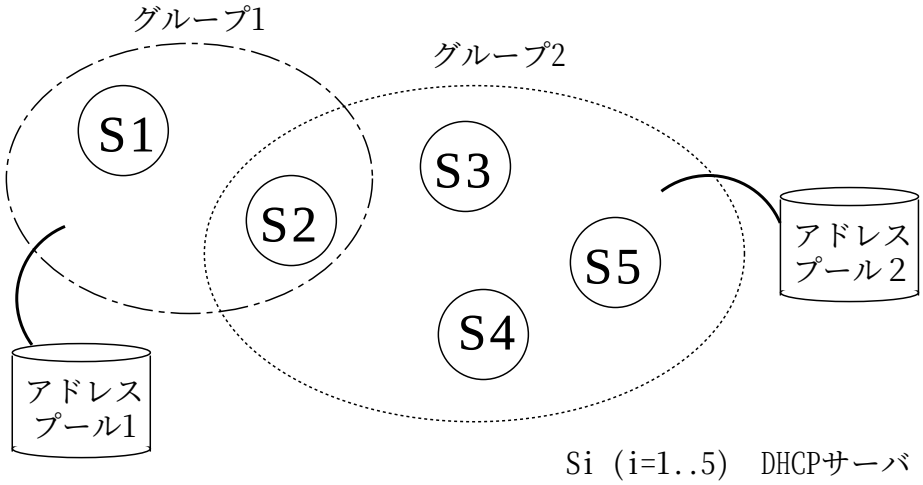

図 2.3: グループの概念図

2.3.2 プロトコル概要

サーバ間プロトコルでは、IP アドレスの状態をいかに適切に制御できるかが鍵となる。アドレスの状態に関する情報交換は、大きく二つの段階があると考えられる：

1. アドレスを利用可能な状態にする段階

2. アドレスを貸し出し、そのアドレス期限を管理する段階

サーバ間プロトコルのモデルでは、貸し出しに利用される IP アドレスを複数のサーバで共有化したため、上記 1 の段階が必要となった。この段階を設けることで、複数のサーバが一つのアドレスを別々のクライアントに重複して貸し出すことを回避する意図がある。ここまではグループ内のサーバ間内でのメッセージ交換である。後半は、サーバとクライアント間でのメッセージが交換が含まれる。

サーバ間ではメッセージを交換することによって、アドレスを制御している。以下にアドレスの状態遷移図を示し、状態を遷移させる操作との関係を確認する。

IP アドレスの状態遷移　RFC2131[2] では、アドレスの状態はあまり明確でなかった。しかしながら複数のサーバで共有化したアドレスプールを扱うためには、アドレスの状態と遷移を明示しなければならない。

一つの DHCP サーバで IP アドレスが管理されている場合は、アドレスの状態は 2 つで十分であり、状態遷移は図 2.4のようになった。サーバ間プロトコルでは更に 4 つの状態が加わった。サーバ間プロトコルは現在ドラフト段階で議論中であり、まだ仕様が完全に決定してない。そのためドラフトで記されている状態遷移は曖昧な点が多いため、ここでは説明を省略し、アドレスの状態を定義するに留める。

アドレスの状態　アドレスの状態は、DHCP とサーバ間プロトコルで共通の意味である。

- UNBINDABLE
  初期のアドレス状態、どのサーバも利用してない

- BINDABLE
  アドレスを貸し出すことができる状態

- POLLING
  合意を取るため、グループ内に問い合わせをしているアドレスの状態

- BOUND
  クライアントに貸し出された状態

- PUSHED
  クライアントに貸し出されたという事実をグループ内の全てのサーバが認識したアドレスの状態

- EXPIRED
  貸し出し期限が切れたアドレスの状態

状態遷移

- BINDABLE→BOUND
  クライアントから新規アドレスの貸し出し要求である DHCPREQUEST メッセージを受信したとき、状態遷移が発生し、サーバは DHCPACK を返す。

- BOUND→BINDABLE
  貸し出したアドレスの期限が切れたときに、この状態遷移が発生する。

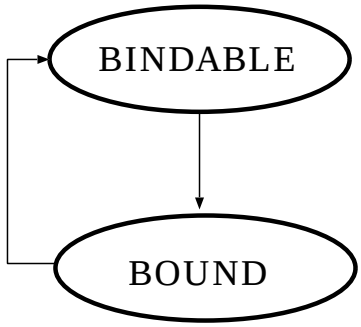

図 2.4: DHCP でのアドレス状態

2.4 本研究の位置付け

第 2.3節で述べたサーバ間プロトコルでは、各サーバによるバインディングの情報更新は常に成功すると仮定している。このプロトコルは、主眼として分散したデータベースの同期を取る仕組みを提供することにあると考えられる。一部のサーバが永久的に停止する

という障害には、冗長化というサーバ構成とグループからサーバを切り離すという方法で耐えることはできる。

しかしながらより現実的には、サーバの停止の他にネットワークの分断という障害が発生するが、ドラフトの方式ではこれらの障害時におけるアドレスの扱い方を明確にしていない。そのため既存の方式のままでは、

- グループが分離した場合、各グループが同じアドレスプールを保持し続けることになり、アドレスの重複貸し出しが起こる危険性がある
- ネットワークの分断が発生することを想定すると、分散しているサーバ間において、バインディング情報を完全に同期させることは困難である

という問題が浮上してくる。

本研究では、DHCP サーバがより現実的なネットワークの障害を扱えるようするため、障害としてサーバの停止とネットワーク分断に耐える仕組みを提供する。そしてこの二つ障害が同時に発生する場合も多重に起こる場合も想定する。システムが多段に分割され最悪の状況に陥っても、正しく動作できる枠組を検討する。

サーバ間の合意形成やアドレスプールの構造における新たな方式を提案し、サーバが提供するサービスの信頼性を保障できることを目指す。

第 3 章

動的ホスト設定機構の耐故障化への提案

本章では、本研究で提案するサーバ間プロトコルの詳細な目標設定を提示する。その上で複数の問題領域が存在することを明かにし、それらを達成するために必要な機能を対応付ける。本稿では耐故障化の方針として悲観的戦略を用る。そしてデータベースの同期、サービス、アドレスプールの 3 つの観点から具体的に耐故障化の基準を設定する。システムが障害によって最悪の状況に陥っても、サーバが矛盾したデータを乱用することを回避する。

3.1 概観

本稿が提案する機構は、ネットワーク上の障害に耐えサーバがサービスを継続的に提供できるように DHCP に耐故障機能を備えている。モデルとしては、扱うべき障害を限定し、本システムではそれらの障害に耐えるため DHCP サーバを冗長にする。その上で障害によって引き起こされるシステムの状況を分析し、耐故障の機能を明確にする。

3.2 モデル

本稿では、冗長な DHCP サーバの構成として第 2.2.2項で述べた漸次縮退系を適用する。これは複数のサーバはそれぞれクライアントにサービスを提供することができ、障害によって通信不能となったサーバは切り離される方式である。

複数の DHCP サーバで形成される世界をサーバグループまたは単にグループと呼ぶ。

この考えは、ドラフト draft-ietf-dhc-interserver-02.txt[5] で提案された概念を基に本システムで応用することにする。サーバグループの詳細は、第 3.3節で述べる。

プロトコルの要求事項

本研究で提案するシステムは、従来の DHCP プロトコルを変更無く扱えるようにする。そのため既存の DHCP クライアント、リレーエージェントと問題無く共存できることを考慮している。

ここで意図する耐故障化とは、引き起こされた障害の影響を軽減することである。

プロトコルの要求仕様を考える大前提として、システムの構成要素にはある時点で必ず故障が存在すると仮定している。その故障は誤りを引き起こし、そして誤りが障害となる。すなわち、サービスの異常が表層に出現する。

本稿では、DHCP サーバ間プロトコルに悪影響となる障害が何なのか、その障害に耐えるためシステムは何をすべきであるのか、という問題意識から出発している。

障 害

システムが扱える障害は下記の二つに限定している。ここでは、故障に関しては深く追求しない。

1. サーバの完全停止
   要求に対して全く返答しないサーバ。プロセス自体、またはプロセスが動作してる計算機のダウン

2. ネットワークの分断
   通信基盤の物理的損傷やルータ等の中継ホストが停止することによって引き起こされる

この二つの障害は独立している。それは、たとえネットワークの分断が起きたとしても、サーバは動作している可能性があるからである。

これらの障害は時には同時に発生しうるかもしれない。更に、障害が連続的に起こることも考慮しなければならない。一つの障害が発生し、その障害が回復する間もなく新たに障害が積み重なっていくことがあるだろう。

以後、本稿で単に「障害」という時は上記の 2 つを意味する。そして、それは障害の同時発生と多重発生という現象をも含む概念とする。

プロトコルの達成目標

1. サーバ群はどのような障害に陥っても、常に安全なサービスを提供する。グループから故障したサーバを適切に取り除くことができる。
2. 分断状況が回避されたりサーバが復旧した時、そのサーバはグループへ的確に統合され、かつ正しいデータベースの情報を再構築できる。
3. 各サーバが保持するデータベースの同期が保障される。
4. あるサーバと通信できるクライアントは、特定のサーバに借りた IP アドレスの期限延長などのサービスを受けられる。

3.2.1 前提条件

- システムの初期状態では、DHCP サーバは一つのグループに所属し、各サーバは同じグループのメンバーを認識している。
- サーバが所属するグループは一つとし、複数のグループに参加することは無いものとしている。
- 新たにサーバが増設されることは現時点では考慮しない。
- ネットワークの分断とは完全分離を意味する。グループが二つに分割した場合、グループ内のサーバ間では通信できるが、他のグループのサーバとは通信不能である状況を指す (図 3.1)。
- DHCP のクライアントとリレーエージェントは、従来の仕様を維持し修正を加えないこととする。

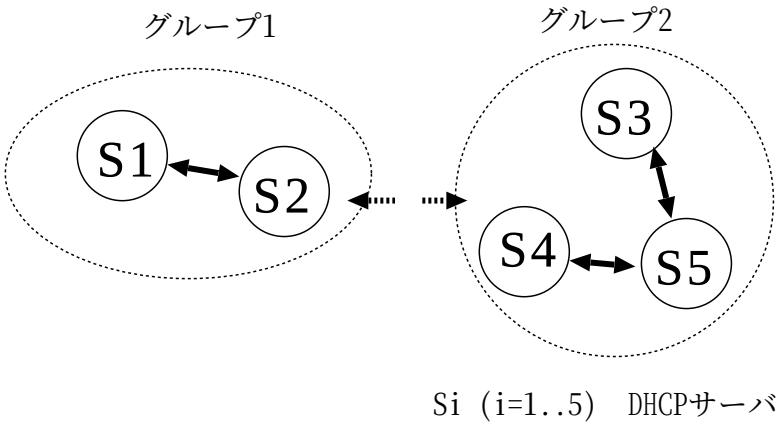

図 3.1: ネットワーク分断

3.2.2 耐故障化の方針

耐故障化全般に渡る方針としては、悲観的な戦略 (pessimistic strategy) [4] を採用する。これは障害が発生した時、サーバが保持しているデータの利用を制限するが、データの矛盾を回避することに重点を置いた方針といえる。

本稿が対象としている障害は、第 3.1節でその性質を分析したように、サーバが障害を特定することはできない。そのため、あるサーバから見て通信できないサーバでさえも実は動作していることがある。このような状況下では、分離したサーバ間において分離前から共有化していたネットワークアドレスを利用する時に、衝突が発生しうることが考えられる。

図 3.2に表す状況を想定してみる。ここでは、サーバ S1 の視点で他のサーバとクライアントの関係を描いている。分断後、S1 は S3 とグループが分かれたが、S3 が動作しているのか、どのクライアントと通信できているかなど S3 取り囲む環境を知ることはできない。そのためグループの分離後、分離したそれぞれのグループが同じアドレスプールを参照していると、各々のサーバが同一のアドレスを別々のクライアントに貸し出す可能性が生じる。

それゆえ本節では、悲観的な戦略を耐故障化の方針として決め、更に具体的に 3 つの観点について耐故障化の基準を設定していく。

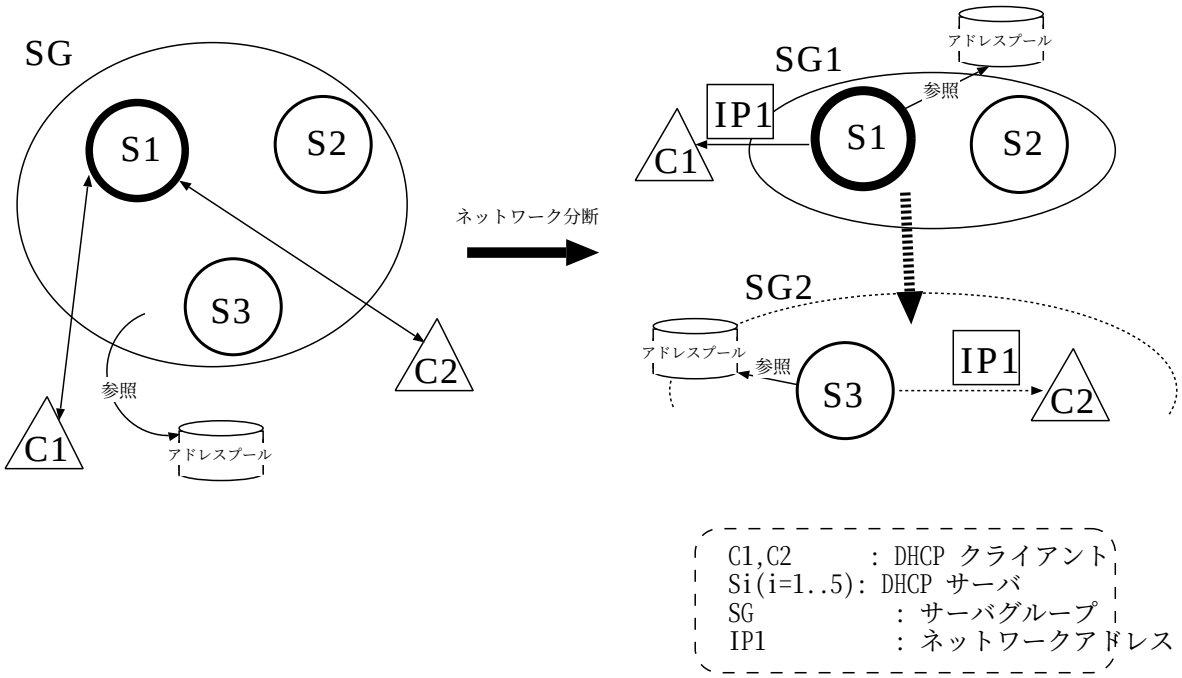

図 3.2: アドレスの重複貸し出し

データベースの同期

各サーバが同期を取るべきデータベースは、

- アドレス権限
- バインディング

の二つである。

アドレス権限のデータベースは、完全に同期する必要がある。なぜなら、サーバが権限を取得するにはグループの完全合意を得なければならないからである。

ここで、完全合意について説明を加えることにする。グループ内で合意が形成される過程を単純化したものを図 3.3に示す。3 台のサーバが存在する時、アドレスの権限取得が成立する場合を表している。図において時間は下向きに進む。

サーバがアドレスを予約する際に使うメッセージを表 3.1に記す。メッセージ OFFER を送信したサーバは、アドレスの予約が失敗したと認知した時に、OFFER_NG を送信しグループのメンバーに対して明示的に失敗した事を伝えるものとする。

表 3.1: アドレスの予約に関するメッセージ

メッセージ	意 味
OFFER	アドレスの予約申請
REPLY	問合わされてアドレスを使用しているか/いないかを答える
OFFER_NG	アドレスの予約失敗

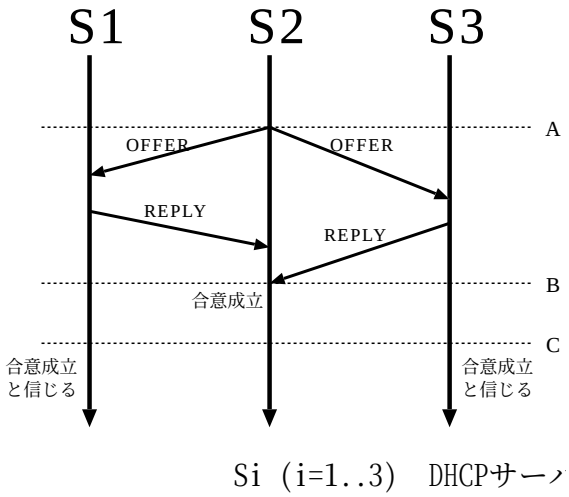

図 3.3: 合意成立

合意の過程を A、B、C の 3 つの時点で区切って説明する。

A 地点 S2 がグループにアドレスの予約を申請するため、メッセージ OFFER を送信した。

B 地点 グループ内の全てのサーバからメッセージ REPLY を受信する。REPLAY の内容から誰もアドレスを使用していないので、S2 はアドレスの予約に成功したと認識する。

C 地点 サーバ S1、S3 とも合意成立と信じる。

一方、バインディングデータベースの同期は、それほど厳密でなくて良いと考える。最低限必要なアドレスの更新情報を保持できる程度とする。仮に情報がサーバに伝わらないとしても、サーバグループの再構成がいずれ起きるからである。

更新情報は、アドレスの権限を取得したサーバ (以後、AS[1] と呼ぶ) が他のサーバ (以後、NAS[2] と呼ぶ) に伝達する。NAS は、通常、AS が発信した更新情報に基づきバインディングデータベースを更新していく。

サービス

サーバがクライアントに提供できるサービスの種類は、次の 4 種類である：

1. 新規アドレスの貸し出し
2. 貸し出し期限の延長
3. 再割当て
4. 再利用

AS は上記の全てのサービスを任意に提供できる。一方 NAS は、(1) と限定付で (3) のサービスを提供できる。(3) に関しては、AS が設定した延長可能な期限 (ELT) までは、NAS がクライアントからの延長要求に答えることができるとする。

[1] Authoritative Server

[2] Non-Authoritative Server

アドレスプール

障害が発生しグループ内のメンバーが分離されると、アドレスプール中で権限が発生していないアドレスを分割する必要がある。この分割は、初期設定による規則に基づいて行われる。分離したグループ間でアドレスプールは正確に分割されていなければならない。

障害が復旧しメンバーがグループに再統合されたら、それぞれが管理しているアドレスプールは一つにまとめなければならない。その際に、データベース情報の整合性をとることが必須の要求である。

3.3 サーバグループ

グループは、サーバ間でリンクが存在し、そのリンクで繋がったサーバの集合である。一つのサーバがグループ内の少なくとも一つのサーバとリンクが存在する時、そのサーバは同一のグループに所属する。

ここでは、グループの変化を考察する。先に第 3.1節で定義した障害の発生が、グループにどういう影響を与えるのかを予想しなければならない。それによって、グループの挙動に対する耐故障化を実現する仕組みが明確になると考える。

グループの変化としては、障害の発生によってグループまたはメンバーが分離する場合と障害の復旧によってグループまたはメンバーが一つに統合する 2 種類がある。

グループ分離

図 3.4は、DHCP サーバ S1 から見たサーバグループの分離状況である。図のように 3 種類に分かれる。(1) は正常状態である。(2) は、S3 と通信不能になった状況を表している。(3) は、S1 はグループ内のどのメンバーとも通信できない状況である。そのため、S2 と S3 が同じグループなのか別々なのかは全く認識できない。

グループ統合

同様に、サーバ S1 の視点でグループ統合の様子を図 3.5 に示す。図はは S1 と S2 で構成されるグループに S3 と S4 のグループが統合され一つのグループを形成したことを表している。S5 は依然、分離したままである。

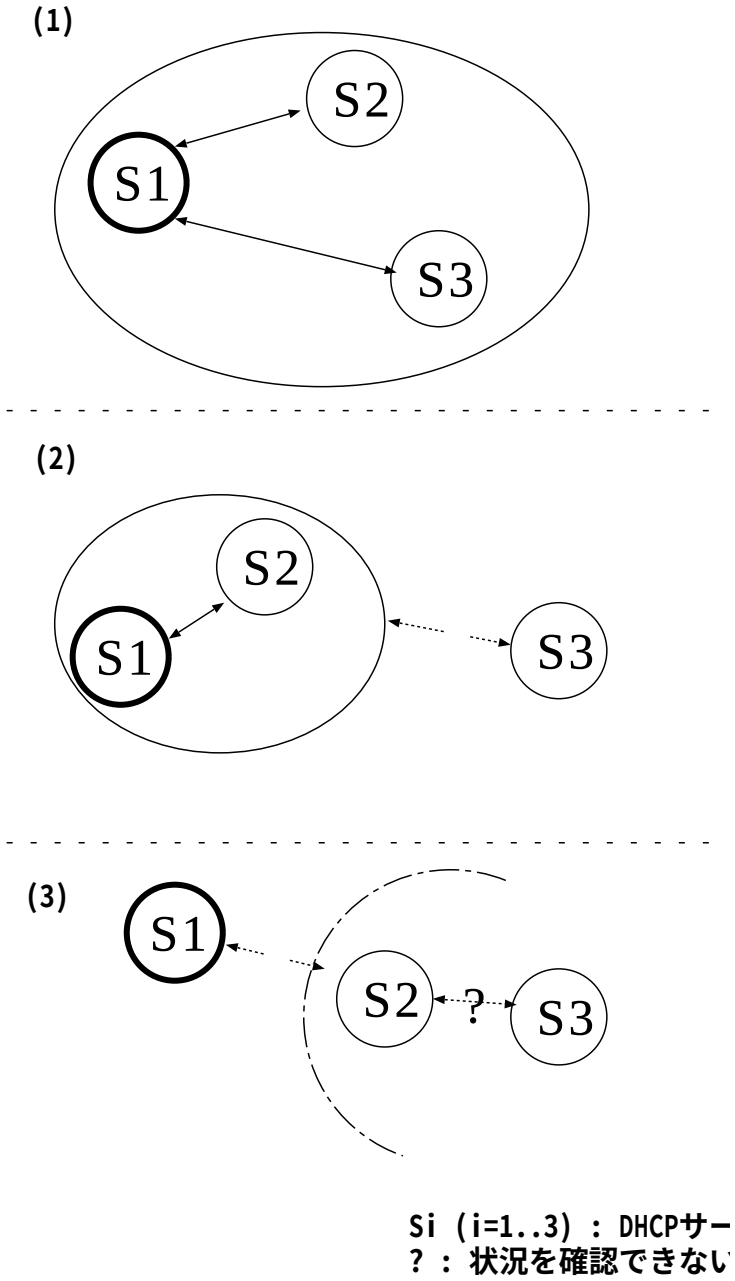

図 3.4: サーバグループの分離 (S1 の視点)

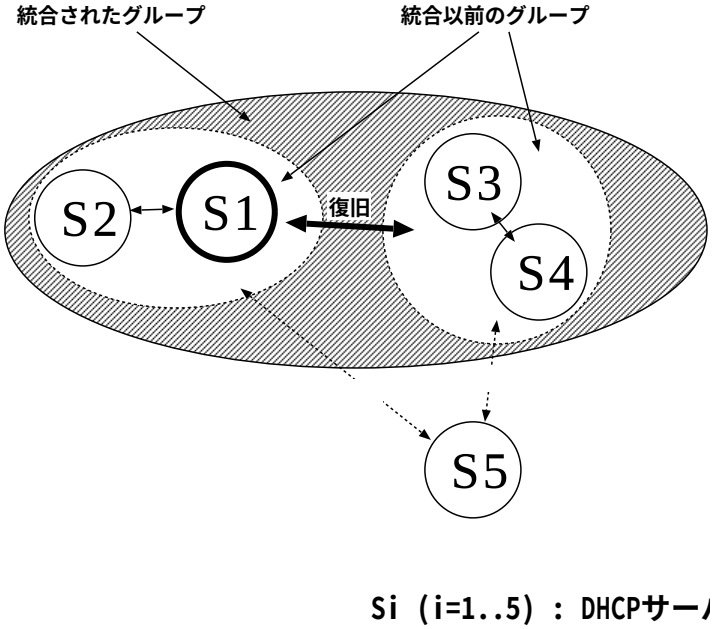

図 3.5: サーバグループの統合 (S1 の視点)

3.4 階層化

3.4.1 問題領域の階層化

本研究で提案する耐故障性を備えた動的ホスト設定機構は、耐故障化を実現するにあたり解決すべき問題が複数存在している。そこで、階層化という方法を用いて問題領域を分割する。この結果、 ネットワーク監視層、トポロジー制御層、グループ制御層、アドレス制御層の 4 つに分け、各層が扱う問題を明確にすることができる。問題領域を階層化した概念図を図 3.6に示す。

ネットワーク監視層

計算機がネットワークに接続しているか、データの送信ができるかを監視する層である。この層は既存のネットワークを意味している。TCP/IP プロトコルスウィートの 4 層モデル [6] におけるトランスポート層以下を指している。二つのホスト間のデータ送信形式をどのようにするかを問題にする。

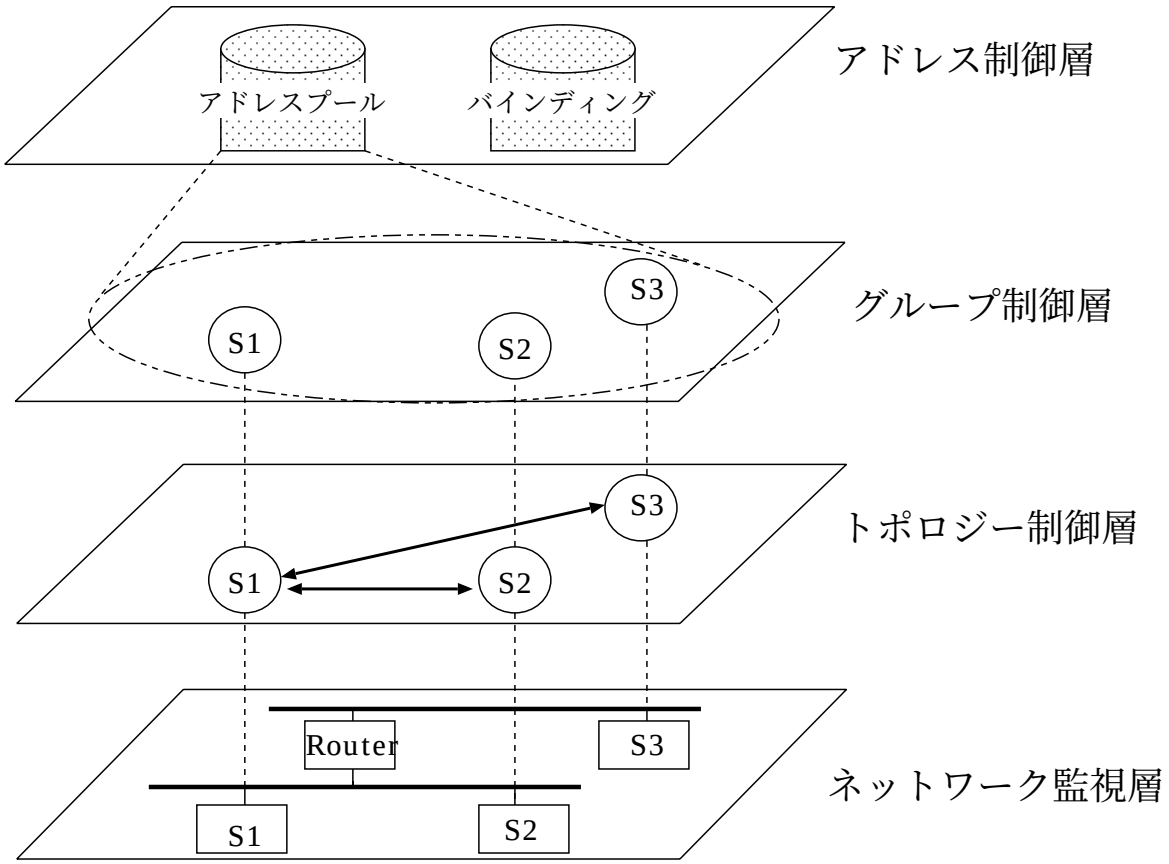

図 3.6: 問題領域の階層化

トポロジー制御層

この層では、DHCP サーバ群で形成される世界において、サーバ間の接続状況を把握する。仮想的にサーバ間におけるネットワークのトポロジー情報を認識しなければならない。

サーバへデータを送信する時、トポロジー情報に基づいて指定されたサーバ群に対するの最適な配送経路を決定する必要がある。

グループ制御層

この層は、同一グループのメンバーである DHCP サーバを認識することである。グループ全体に必要なデータが正しくメンバーに届いてるかを監視する。

アドレス制御層

この層では、時々刻々と変化するネットワークアドレスに係わるあらゆる情報を把握する。

- アドレスを含む提供できる資源は何か
- 共有化したアドレスをサーバ間でどうやって利用するか
- クライアントへのアドレスの貸し出しをどう扱うか

これらを解決しなければならない。

3.4.2 機能の階層化

本項では、第 3.4.1項で明らかにした各層の問題に対する解決策、すなわち提供する機能について述べる。

トポロジー制御層

DHCP サーバの世界では、サーバ間の連結は緩やかである。一つのサーバが全てのサーバとコネクションを確立している分けではない。そのため現時点のトポロジー情報から特定のサーバへの経路が無い場合、まず代替経路が存在するかどうかを確かめる。これが、

代替経路探索という機能である。さらにデータを DHCP サーバへ送信するための経路制御機能が必要である。

グループ制御層

この層では、グループメンバーの分離・再統合の機能を提供する。アドレスの扱いに関しては、全て上位層に委託してしている。

下位の層からコネクション変更の情報を受けると、その情報に基づいて各サーバが属すべきサーバグループを再構成する。

アドレス制御層

この層は、IP アドレスに関する管理を行う。具体的に提供すべき機能を以下に述べていく。

1. **バインディング情報の管理** クライアントからアドレスの貸し出し要求を受けると、サーバはクライアント名と貸し出す IP アドレス、その貸し出し期限、貸し出したサーバ名を記録したバインディグ情報を生成する。

   従来の記録内容に加え IP アドレスがどの状態に位置しているかも監視している。

2. **アドレスプールの管理** アドレスプールは、一つのサーバグループに対して一つ存在し、各サーバが共有している。

   ここではグループの分離が起こると、分断後もサーバがアドレスの貸し出しを継続できるように、アドレスプールの分割という機能を提供する。これにより、各サーバは分断後も利用できるアドレスを確保できることになる。

3. **権限管理** サーバはプールにあるアドレスを利用するに際に、サーバ間で合意を取らなければならない。それによってアドレスを利用できる権限を取得するのである。ここではこの権限を取得するため、交渉機能を提供する。

以上、これまで問題領域と提供する機能とを述べてきた。ここで第 3.4.1項と本項との対応を簡潔にまとめたものを図 3.7に示す。

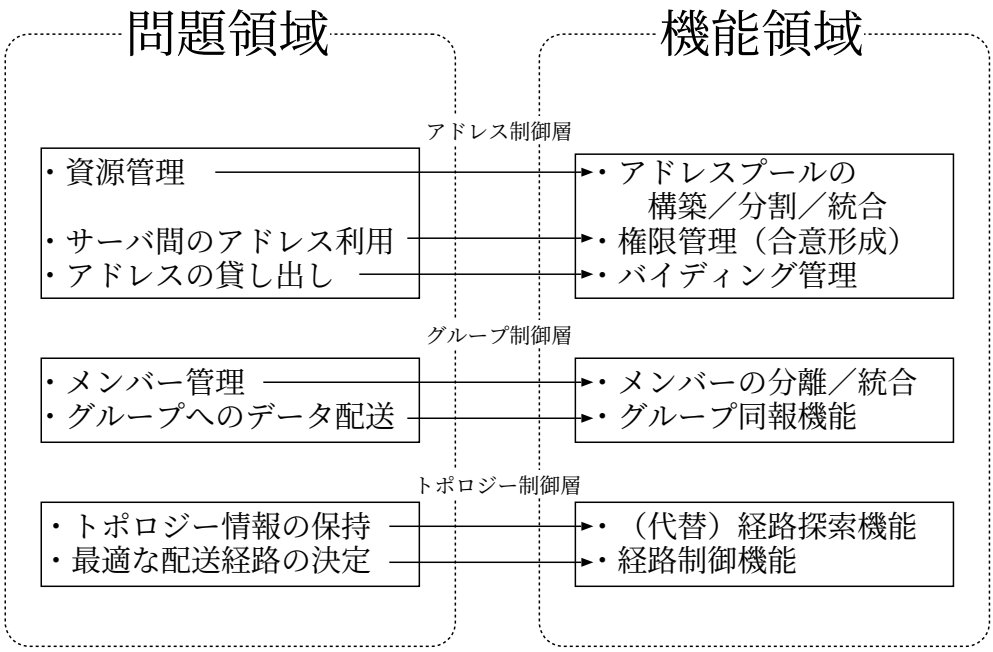

図 3.7: 問題領域と機能との対応

第 4 章

アドレス制御層の設計

本章では、本システムで最も複雑かつ核心部分であるアドレス制御層に焦点を当てる。まず、前章で述べた理想的な耐故障化を実現するにあたり、グループ内の合意形成が困難な問題であることを指摘する。その上で、より確実で曖昧性を排除した合意形成方法である「二者間の合意」を提案する。そして二者間の合意が常に成立するように、アドレスプールの構造を変換する。最後に、二者間の合意に基づく動作の流れを示す。

4.1 アドレス制御層の問題点

第 3.4.1項で既に述べたように、アドレス制御層ではデータベースの同期を行う。ここでは、特にアドレス権限の同期のみ取り上げ、権限取得の際に含まれる問題を明らかにする。

一つのサーバが資源としてのアドレスを利用する際には、完全合意を取らなければならない。それは各サーバが、アドレスプール内の未権限のアドレス領域を正しく認識できるようにするためである。グループが 3 台以上のサーバから構成されている場合、完全合意を形成することは深刻な問題になる。

4.1.1 完全合意の難しさ

合意に矛盾が生じる場合を図 4.1に示し、完全合意を形成する困難さを説明する。合意の過程を A、B、C の 3 つの時点に区切って状況を分析する。図の右側では、各時点におけるグループの状況を合わせて示している。

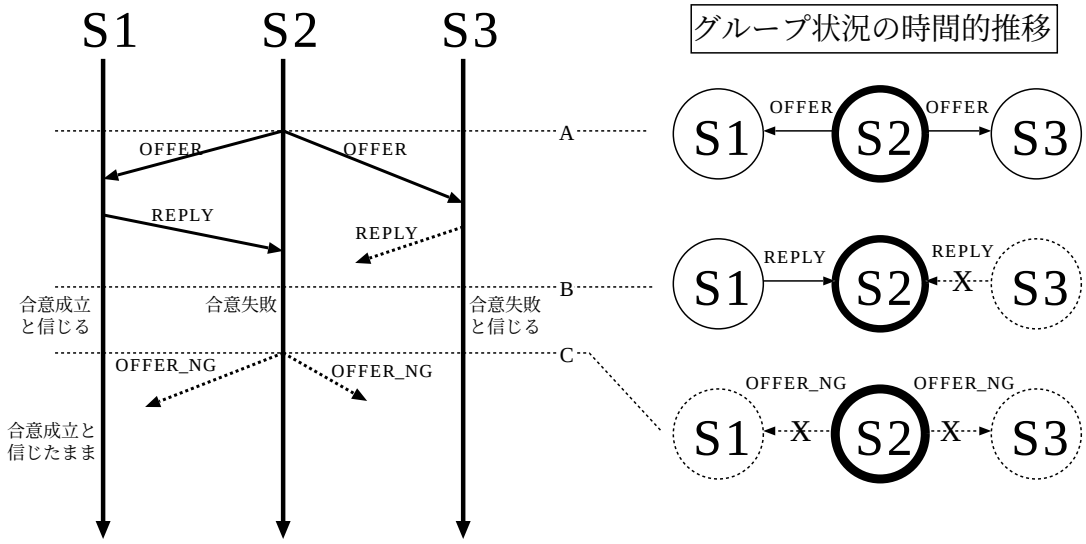

図 4.1: 合意の矛盾

A 地点 S2 がグループにアドレスの予約を申請するため、メッセージ OFFER を送信した。

B 地点 サーバ S1 からは REPLY メッセージを受信した。S3 からは応答が無い。最終的に先の申請に対して合意不成立と判断する。S1 は合意成立と信じている。S3 が動作している場合は、S2 とコネクションが張れないので、合意失敗と思う。

C 地点 S2 は合意失敗を伝えるため OFFER_NG メッセージをグループに送信しようと試みる。しかし S1 ともコネクションが張れない状況に陥ってしまった。S1 は依然、合意成立と信じたままである。

このように、合意の矛盾が生じることが分かる。このことはプールを分割する時に悪影響を及ぼす。結果として、分離したグループが認識している未権限のアドレス領域に食い違いが生じることになる。

4.1.2 二者間の合意

3 章で述べたモデルでは、前述した合意の矛盾が生じてしまう。アドレスの状態をより精緻に制御したとしても、必ず合意に曖昧性が含まれてしまう。これは、ネットワークの分断やサーバの停止といった故障が (1) 予期せず発生すること、(2) 故障の種類を特定できないという理由があると考えられる。

そこで、曖昧性を排除した最も確実な合意形成を創造する必要性がある。その方法は非常に単純であるが、二者の間で合意を結ぶことである。二者間であれば、合意が成立したか失敗したかのどちらかであり、第 3 者の意見を待つ必要は無い。

N 者間の完全合意は、二者間の合意を積み重ねることで求めることとする。

しかし二者の間では確実な取り決めが成立し易くなったといっても、グループメンバーが多数存在する場合、完全合意を形成することは依然として困難である。そのため、2 台のサーバだけでアドレスの権限交渉ができる枠組を創り出し、常に二者間の合意だけが必要な状況にする。そこでアドレスプールの構造に変更を加えることにする。これは次節で述べることにする。

4.2 アドレスプールの構造

新たに決定したアドレスプール構造の概念を図 4.2に示す。この図は 3 台のサーバが貸し出しアドレスを共有しており、各サーバが確保していくアドレスの予約領域が変化する様子を表している。

ここで、予約領域とはサーバ間の合意が取れたアドレスを意味する。一つのサーバだけが利用できるアドレスである。交渉領域は、まだどのサーバも権限を取得していないアドレスを指す。領域の責任者とは、交渉領域を管理する 2 つのサーバのことである。プールを分割するときは、この領域をサーバ間で分けることになる。

利用規則

交渉領域は 2 台のサーバが管理しているため、基本的にはその 2 者間で合意が成立すればアドレスを利用できるようにする。

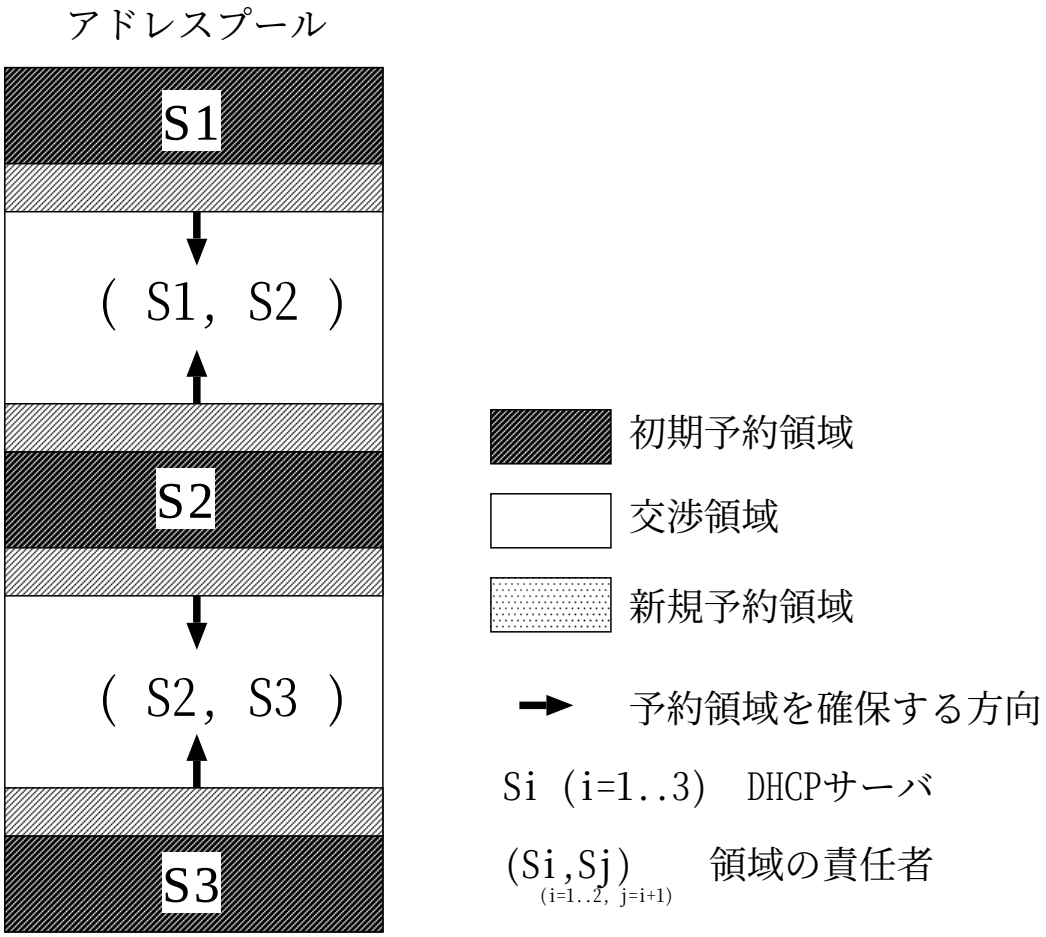

図 4.2: アドレスプールの構造

4.3 機能定義

- DvideAddrPool()

  説明 この関数は事象 PART_MEMBAER によって引き起こされる。アドレスプール中のアドレスを規則にしたがって分割する。規則は初期設定で決まる。

  返り値 OK または NG

- IntegrateAddrPool()

  説明 この関数は、一つはメンバー統合という事象によって、二つ目は AS と分離後 AS から UPDATE を受信することによって引き起こされる。AS と同じグループに復帰した時は、AS からの正しい状態情報を用いてアドレス状態情報ファイルを更新する。権限が発生していない情報は、各グループが保持しているアドレスプールを統合する。

  返り値 OK または NG

4.4 IP アドレスの状態制御

アドレスの状態を適切に制御できることがアドレス制御層の核心部であることは既に述べてきたことである。本節はそのアドレスの振る舞いを定義する。そのためアドレスが取りうる状態を定義し、次に外部からの刺激である事象と反応を明確にする。最後に、それらと上述した機能とを合わせ抽象化したアドレスの状態遷移図と状態遷移関数としてまとめる。

4.4.1 アドレス状態定義

- IDLE
  これはアドレスの初期状態を意味している。どのサーバもこの状態のアドレスに対する権限を交渉していくことができる。

- BINDABLE
  二者間の合意が成立したアドレスの状態である。サーバがアドレスの権限を取得したという時のアドレス状態を指す。

- BOUND
  クライアントにアドレスが貸し出された状態である。
- NO_AUTHORITY BOUND
  アドレスの権限を取得していない側において、アドレスがクライアントに貸し出されたことを表している。
- LP EXPIRED
  AS 側においてアドレスの貸し出し期限が切れた状態である。
- ELT EXPIRED
  AS 側においてアドレスの貸し出し期限が切れた状態である。
- NO_AUTHORITY LP EXPIRED
  NAS 側においてアドレスの貸し出し期限が切れたことを認知した状態である。
- NO_AUTHORITY ELT EXPIRED
  NAS 側においてアドレスの延長可能期限が切れたことを認知した状態である。
- FAILURE EXPIRED
  AS 側においてアドレスの期限が切れ、その上グループの分離が起きたときのアドレスの状態である。
- NO_AUTHORITY FAILURE UNKOWN
  他のサーバがアドレスの権限を取得していて、AS とグループが分離した後の状態。
- NO_AUTHORITY FAILURE UNKOWN2
  サーバが AS と分離し、ELT が切れたアドレスの状態である。クライアントからの延長要求は却下する。
- NO_AUTHORITY FAILURE BOUND
  グループは AS とは分離したが、アドレスはクライアントに貸し出されていることを表している。
- INVALID
  アドレスプール内の交渉領域を分割した結果、自身が管理すべきでないアドレス群がこの状態に留まる。

4.4.2 事象定義

- RESERVED
  交渉領域にある特定のアドレスの権限を取得できたことを表す
- PART_MEMBER
  現行のグループ状態からあるサーバ (AS 以外) が完全分離したことを意味する
- INTEGRATE_MEMBER
  現行のグループ状態から分離していたサーバ (AS 以外) が復帰したことを意味する
- PART_AS
  NAS が AS とグループから分離したことを意味する
- ENOUGH_QT_ADDR
  グループの共有資源である IP アドレス数が十分にあることを表している
- LACK_QT_ADDR
  グループが共有している IP アドレス数が減少し十分な量がないときや枯渇したことを表している
- LP_EXPIRED
  貸し出し期限 LP(lease priod) が切れたこと
- ELT_EXPIRED
  延長可能タイマー ELT(extension limit timer) が切れたこと
- INVALID_ADDR
  所有していたアドレスが、アドレスプールの分割によって他のグループの資源に移ったことを意味する
- TRANSFER
  アドレス権限をグループ内の他のサーバに委譲する
- AS_IS_IDLE
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が IDLE 状態にあることを示す。そのアドレスを IDLE 状態に遷移させなければならない。

- AS_IS_BOUND
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が BOUND 状態にあることを示す。そのアドレスを NO_AUTHORITY BOUND 状態に遷移させなければならない。

- AS_IS_LPEXP
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が LPEXP 状態にあることを示す。そのアドレスを NO_AUTHORITY LP EXPIRED 状態に遷移させなければならない。

- AS_IS_ELTEXP
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が ELTEXP 状態にあることを示す。そのアドレスを NO_AUTHORITY ELT EXPIRED 状態に遷移させなければならない。

- AS_IS_INVALID
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が INVALID 状態にあることを示す。そのアドレスを INVALID 状態に遷移させなければならない。

- AS_IS_FAILEXP
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が FAILURE EXPIRED 状態にあることを示す。そのアドレスを NO_AUTHORITY FAILURE UNKNOWN 状態に遷移させなければならない。

- NAF_BOUND
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が NAS 側で NO_AUTHORITY FAILURE BOUND 状態にあることを示す。そのため、そのアドレスは AS 側で BOUND 状態に遷移させなければならない。

- NO_BOUND
  アドレス状態に関する更新メッセージから、特定されたアドレスの状態が NAS 側で NO_AUTHORITY FAILURE BOUND 状態でないことを示す。

以下は、クライアントから受信したメッセージの違いによる事象である。括弧内は、クライアントの状態である。

- DHCPDISCOVER
- DHCPREQUEST[selected]
- DHCPREQUEST[renewing]
- DHCPREQUEST[rebinding]
- DHCPREQUEST[init_reboot]

4.4.3 アドレスの状態遷移と遷移関数

アドレスの状態遷移図

2 つの視点から見た IP アドレスの状態遷移図をそれぞれ図 4.3、4.4 に示す。図 4.3は、アドレスプール内の交渉領域からアドレスを予約できた場合を表している。図 4.4は、他のサーバからバインディング情報を取得したときの状態遷移を示している。図 4.5は、図 4.4の一部を詳細に表した図である。

状態遷移関数

状態遷移関数は、$Q \times \Sigma$ から Q への写像を表している。ここで、Q は第 4.4.1項で述べた状態の集合で、Σは第 4.4.2項で定義した事象の集合である。この写像は、現在のアドレス状態と入力された事象に対して定まる次のアドレスの状態を意味する。この遷移関数に基づき実装する。

4.5 モジュール関係

上述してきたアドレス制御層は、WIDE 版の DHCP 上に実装することを想定している。本プロトコルにおける耐故障化の機能の一部を取り上げ、既存の実装との親和性を示し、手続きの呼び出し関係を視覚化する。

サーバグループ用の設定

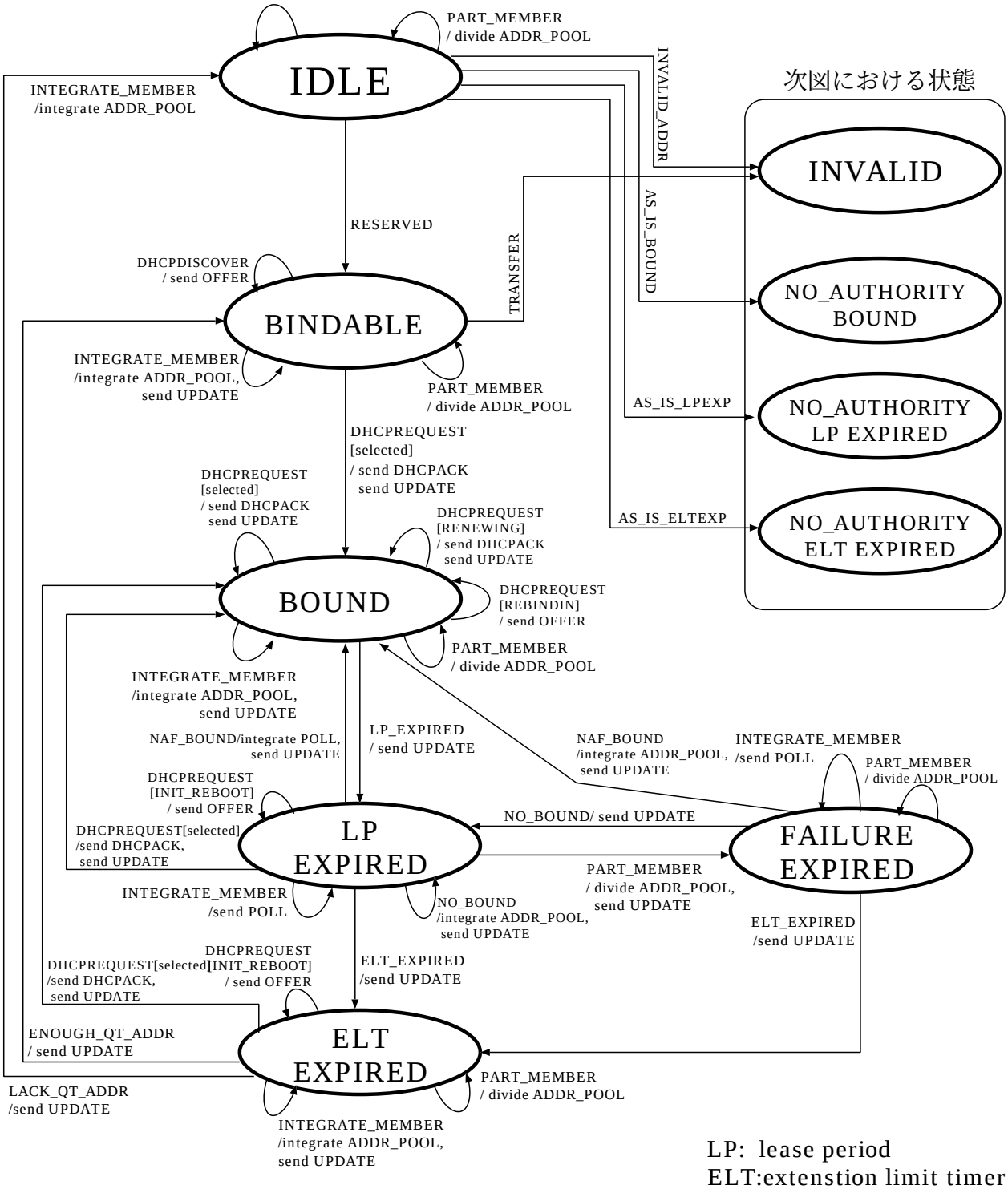

図 4.3: AS から見たアドレス状態遷移図

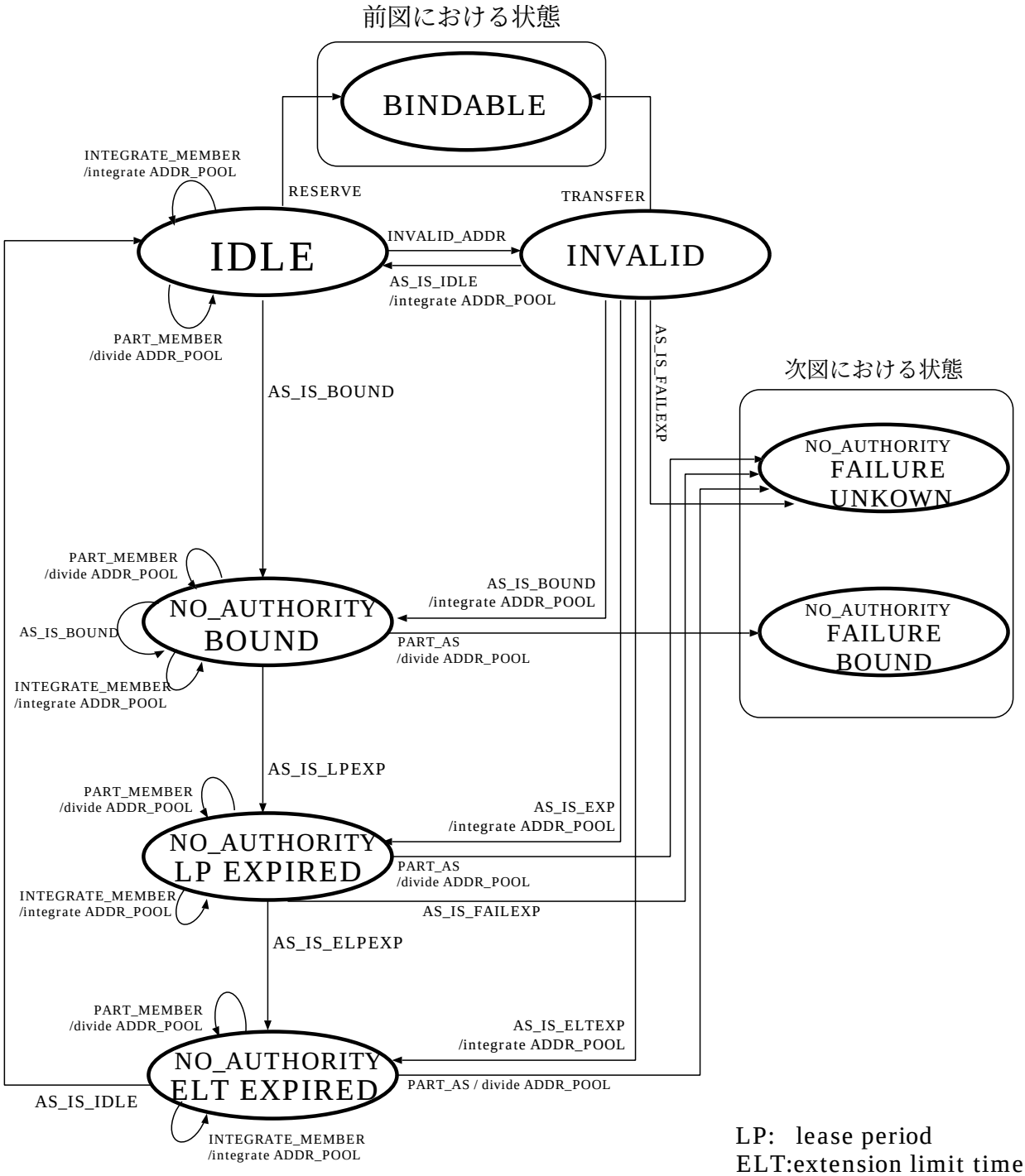

図 4.4: NAS から見たアドレスの状態遷移図

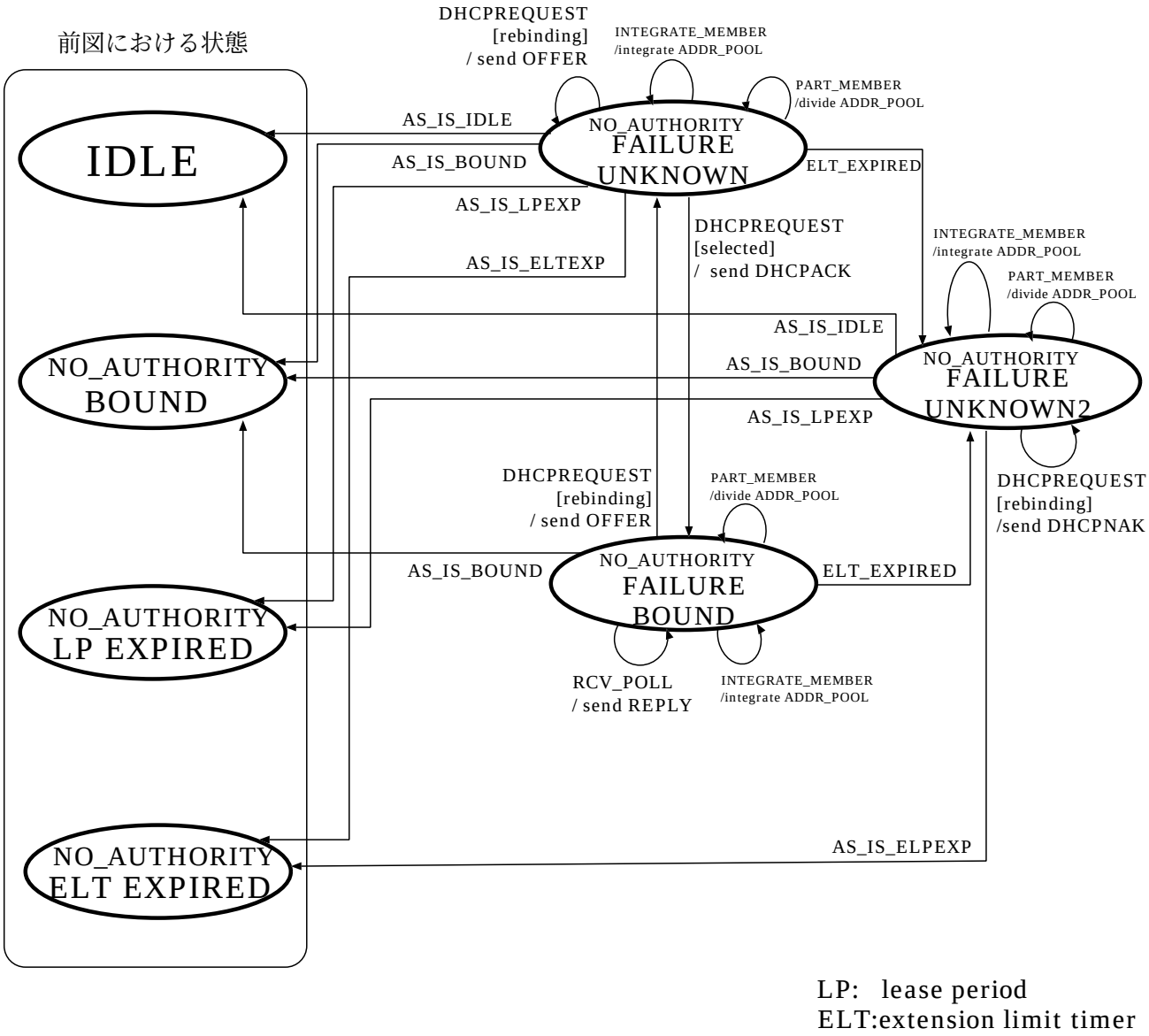

図 4.5: AS と分離したアドレスの状態遷移図

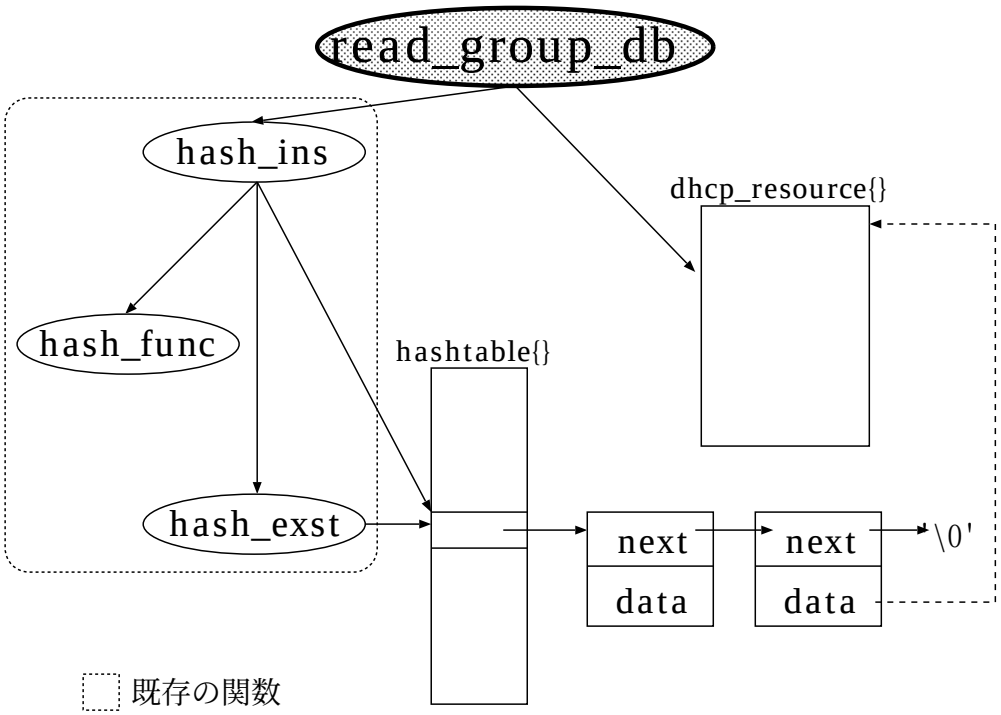

図 4.6: グループ設定

- read_addr_db()
  グループ用の初期設定ファイルを読み込み、資源リストである dhcp_resouce{} に書き込む。これは内部で次の三つの関数を呼んでいる。
  - hash_ins()
  - hash_func()
  - hash_exst()

第 5 章

評 価

評価としては、3 章で述べた理想的な耐故障化の機能と 4 章で述べた実現性を高めた機能とを比較して行う。実現性を高めたことによる利点、それに伴った損失を議論する。次の二つの視点からモデルを評価する。

1. 合意形成の相違

2. アドレスプールの空間の相違

前者はモデルの質を評価の基準とし、後者は数量的な観点である。

合意形成の相違

理想的なモデルでは、完全合意を成立するために一斉合意を行う。この合意方法は、グループ内の各サーバに一度に問い合わせて、各サーバからの返事を待つのである。この方法は、ネットワークの分断という障害を想定する場合、サーバの数が増えれば増える程困難になってくる。非常に効率が悪い。さらに各サーバ間では、利用可能なアドレスの領域を認識することに一貫性が欠けるという事態が起こりうる。データの安全性という視点では、信頼性があまりない言える。

一方、実現可能なモデルでは二者間で合意を成立させる。このことは、何よりもまず確実である。二者間での交渉により、合意成立か不成立かが正確に決定される。ゆえに、上記のような各サーバが保持するデータベース情報に曖昧性が生じることはない。

安全な情報に基づいてサービスを提供する方針は、クライアントの要求を常に満たすことを制限することにも繋がる。しかしこれは、本研究の第一の目的である誤りの無いサービスを提供することを優先した結果であると考える。不利益の副産物が生じた。

アドレスプールの空間

理想的なモデルも実現可能なモデルもどちらもサーバグループで資源としてのネットワークアドレスを共有している点は同じである。しかし実現可能なモデルでは、二者間の合意が常に成立するようにアドレスプールの構造を変更した。これによってアドレスプールの空間の大きさに違いが生じた。

まず、アドレスプールを複数のサーバで利用することを表した概念図を示す。図 5.1は理想的モデルの方を表し、図 5.2は実現可能なモデルにおけるアドレス空間を表している。

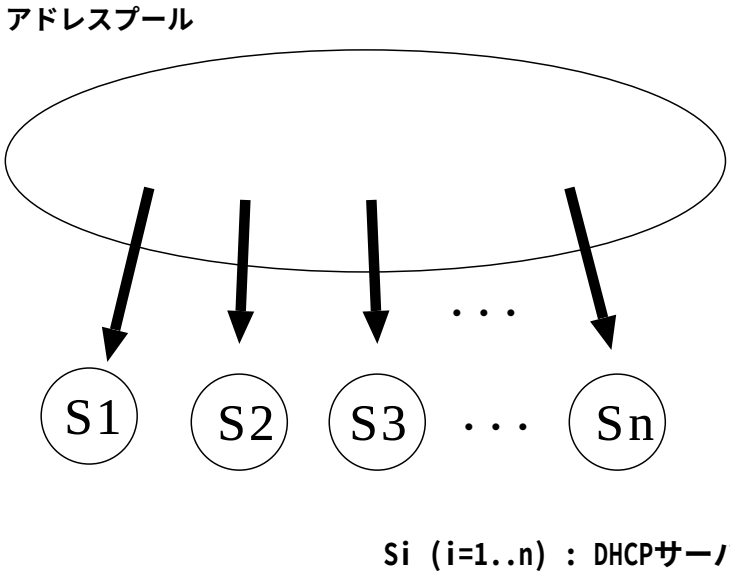

図 5.1: アドレスプールの空間 (理想モデル)

これらの図から明らかなように実現可能なモデルでは、一つのサーバが保持するアドレスプールの空間は縮小した。しかしながら障害が発生した場合、実現可能なモデルの方は各サーバがアドレスプールの領域を明確に認識しプールを適切に分割できるため、アドレスの利用効率は向上する。特にネットワーク分断が発生した場合は、分断した各グルー

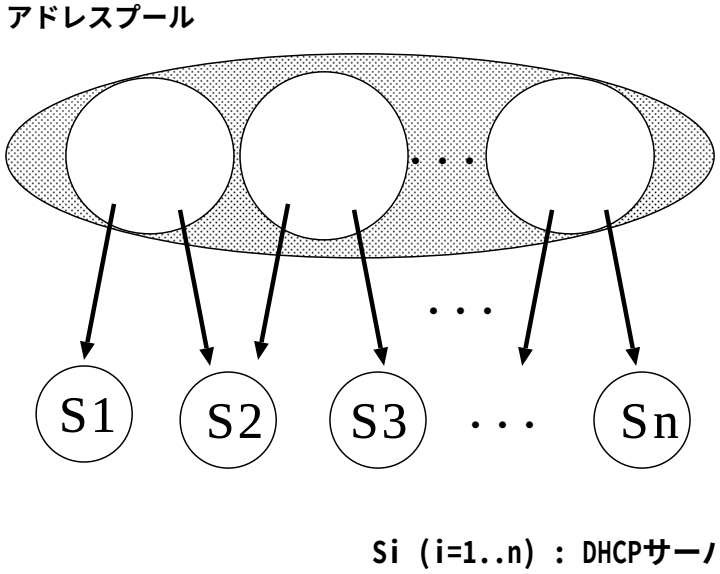

図 5.2: アドレスプールの空間 (実現モデル)

プが分割したプールに基づきネットワークアドレスの貸し出しを継続することが可能となる。

第 6 章

おわりに

以上、障害に耐えられる動的ホスト設定機構の仕組みを研究してきた。ネットワーク上に障害が発生すること想定して、ネットワークアドレスの状態を制御することに問題領域を限定してきた。アドレスの状態を適切に管理するため、障害によってサーバグループがどのように分割されるを分析し、分割されたグループ内の各サーバの視点毎にアドレスの状態制御を設計してきた。

結果として、

- 提供するサービスの信頼性が向上
  アドレスの権限を取得した AS サーバは、障害が発生したとしても全てのサービスをクライアントに提供し続けることができる。ネットワーク分断によって AS サーバと分離した NAS サーバは、アドレスの貸し出し期限を延長することについては貸し出し期限から一定の時間 ELT 以内まではサービスを提供できるが、その期間以降はできない。同じグループであるなら、期間 ELT が更新されていくため貸し出し期限を延長することができる。

- 新たに提案した二者間合意がサーバ間における合意の矛盾を回避
  アドレスを割当て可能状態 BINDABLE にならなければ、クライアントにアドレスを貸し出すことができない。二者間合意により各サーバはメッセージを 1 回送信することで、アドレスの利用に関して合意に達したかが分かる。更に本稿では二者間合意が常に成立するようにアドレスプールを構造化している。

ということが分かった。

本方式の残された課題としては、第 3.4.2節で既に述べたトポロジー制御層とグループ制御層を設計し実現することである。現時点では各層の機能概要を分析したに過ぎない。本稿では障害を二つに限定しているため、耐えられる障害の範囲を拡大することは重要である。第 4 章で述べたアドレスプールの構造は、最も単純に設計したものである。二者間の合意を成立させることを前提にして、アドレスの最適な利用効率を達成する構造を考える必要がある。

謝 辞

本稿は多くの方々のご支援無くしては完成することはなかったであろう。特に次の方からのご支援には最大の感謝を申し上げる。

まず、指導教官である篠田 陽一先生から適切なご指導と助言を戴かなかったら、研究を進める過程で発見した多くの障壁を乗り越えることはなかったであろう。さらに、より基本的な部分である正しい日本語の話し方・書き方などまで幾度となく指摘してくださった。本当に、心から感謝申し上げる。

研究室の方々にも大変お世話になった。毎回のゼミでは多くの貴重な意見を戴いた。発表練習の時、何人かの先輩と M1 の方には審査当日の朝まで内容を詰めてくださり、生涯忘れなれない出来事となった。皆さんに深く感謝申し上げたい。

距離こそ離れているが多くの友人からの励ましが、研究を進める推進力になったことは疑いない。ありがとう。

そして最後に、我が両親には多大な支援と最高の激励を絶えず戴いた。本当に感謝に耐えない。

参考文献

[1] W.J. Croft and J. Gilmore. RFC0951:Bootstrap Protocol, September 1985.

[2] R. Droms. RFC2131:Dyanamic Host Configuration Protocol, March 1997.

[3] R. Droms. RFC1531:Dyanamic Host Configuration Protocol, October 1993.

[4] Pankaj Jalote. *Fault Tolerance in Distributed Systems*. Prentice Hall, 1994.

[5] R. Cole K. Kinnear and Droms. An Inter-server Protocol for DHCP ⟨draft-ietf-dhc-interserver-02.txt⟩, March 1997.

[6] W. Richard Stevens. *TCP/IP Illustrated, Volume1 -The Protocals-*. Addison-Wesley, 1994.

[7] R. Droms S. Alexander. RFC2132:DHCP Options and BOOTP Vendor Extensions, March 1997.

[8] 藤原 秀雄当麻 喜弘. フォールトトレラントシステムの構成と設計. 槙書店, 1991.