遠隔データセンタと、仮想化技術を活用した次世代サービス基盤NHNの導入

NHN(Next Host Network)は、IIJが目指す次世代のサービス基盤です。NHNの導入によって、IIJでのサービス開発フローや設備増強のための需要予測が大きく変わろうとしています。ここでは、NHNを導入するにいたった背景、その設計思想や技術要素を説明します。

IIJでは、遠隔データセンタの活用によってデータセンタコストを圧縮しながら、運用自由度の高いサービス基盤を実現しています。現地作業の集約、ラックスペースや電力を無駄なく使う工夫を施し、運用コストの削減を目指した「NHN」(Next Host Network)を2008年度に設計し、新規サービスと既存サービスを移行する基盤として整備しました。ここでは、NHNの導入の背景、設計思想、技術要素について説明します。

4.1 NHN導入の背景

IIJではこれまで、自社サービス用に複数のデータセンタに分散して200ラック以上、数千台のサーバを運用してきました。各ホストはサービス単位でラックを確保して設備を構築していたため、それぞれのラックスペースやネットワーク機器に、将来の需要増に耐えうる余裕を設けており、基盤全体としてのロスが発生していました。このようなサービスごとの縦割りのシステム構成では、計画変更による拡張や、廃止時の設備の転用等が困難で、サービス基盤のコストがかさむうえ、ラックによって機材や配線が異なるため、運用作業の煩雑さも課題となっていました。

また、これまでは障害時の対応等を考慮した結果、物理システムへのアクセスの容易さから東京近郊で立地条件のよいデータセンタを利用し、サーバ運用の最適化を図ってきました。しかしながら、この数年、IAサーバの性能向上と低価格化を背景に、サーバ1台あたりの消費電力は高まり続け、ネットワーク、サーバ、ストレージ等の機器コストよりも、ラックスペース、冷却用の空調費、電気代といったファシリティコストの比率が高い状況になっています。東京都内のデータセンタは、お客様からの引合いが多く、機器の設置スペースはあっても、空調能力、UPS、非常用発電機の容量不足等の制約から十分にサーバを設置できない状況になってきました。

東京近郊のデータセンタにシステムが集中する構成に限界が見えてきたため、IIJでは場所に依存しないサービス用機材は、郊外等ファシリティコストの低い場所に移すことで、抜本的なコスト削減を目指すことにしました。検討にあたっては、コンテナ型データセンタ等、ファシリティコストの圧縮や電源事情の緩和につながるあらゆる方策を検討しています。多重化技術の進化によりネットワークコストも抑えられるようになってきたことも、郊外データセンタの検討を後押ししました。

また検討に際しては、遠隔データセンタを利用しても運用の自由度と機動性を確保できる構成への見直しも同時に進めました。

サービス用機材を郊外のデータセンタに持っていくことでスペース費用、人件費、電気代等を削減し、サービス基盤全体としてコスト削減に繋げること。同時に運用の自由度を向上させることを目指して、NHNの検討はスタートしました。

4.2 遠隔データセンタ利用を視野に入れた設計方針

NHNでは、サーバ運用者の経験を元に、重要視しなければならない部分と、保留できる部分を次のように切り分けました。

- ストレージに関しては、HDD単体故障等想定可能な故障は容認する。ただし、サービス停止に繋がる故障が起きないように、信頼性の高い構成にする必要がある
- サーバ機器はときどき故障する。故障ポイントは多岐にわたり、故障率を下げることには限界があるため、故障しても影響の少ない構成にすることが望ましい
- エッジで利用するネットワーク機器の故障は、これまでの経験上それほど多くない。NIC冗長化設定等は必要なときにのみ実施する

次に、IIJで利用している機器の故障率等を示し、上記の方針に至った経緯を説明します。

4.2.1 ストレージの故障率について

IIJで利用している約100台のDAS[*1]のログを調査したところ、ストレージの台数自体はあまり変わっていないにもかかわらず、HDDの故障数自体が年々減少傾向にあることが確認できました。また、この数年間の傾向として、ストレージの負荷とHDDの故障率に相関関係がほとんどみられなくなってきています。負荷の高さに関係なくHDDの故障は発生し、同一ストレージに搭載したHDDに故障が偏ることは稀なようです。

年	HDD故障数	同一RAIDでのHDD故障数
2005年	32台	7×1台、5×1台、4×1台、[*2] 2×2台、1×12台
2006年	22台	2×3台、1×16台
2007年	16台	2×1台、1×14台
2008年	7台	2×2台、1×3台
2009年	4台	1×4台(2009年9月現在)

HDDの故障率が年々下がっているため、HDD故障以外のストレージ障害、特にサービス停止に直結するものがより目立つようになってきました。具体的には、次のような障害によるものです。

- RAIDコントローラ上のキャッシュメモリの故障による機能停止
- RAIDコントローラの故障が原因と思われる不安定な動作
- SCSIカード等の接続インタフェースの故障
- RAIDコントローラ上のバッテリバックアップユニットの故障もしくは寿命による性能劣化

故障回数は、全体で1年間に数回程度ですが、RAIDコントローラの二重化や接続パスの二重化を行っていない機器で上記のような障害が発生してしまうと、現地で機器交換作業を行うまで復旧できずサービスの停止に直結してしまいます。遠隔データセンタを視野に入れると、ストレージのサービス停止は交換作業完了までに必要な時間が非常に長く致命的であると考えました。このため、現状よりハードウエアコストは増えますが、RAIDコントローラの二重化、接続パスの二重化を必須条件として機器選定を行いました。

4.2.2 サーバの故障率について

サーバの故障率に関しては、サービス復旧を優先して予防交換してしまうことが多く、その後に故障機器を持ち帰って試験しても再現できないこともあり、故障部位別の統計情報は取れていません。

2009年4月から9月までの6か月間で修理したサーバ数、現在予防交換して検査待ちになっている機材の台数から、故障率はおよそ1〜2%程度と推定しています。ただし、この値には、冗長HDDでの片側の故障といった冗長部品の故障は含んでいません。このため、サーバ機器に関しては、かなり多くの台数が故障すると推定しています。

＊1　Direct Attached Storageの略。IIJでは2008年までSCSI I/Fを経由して外部接続する形態のものを中心に利用していました。サーバ上にローカルHDDを載せる形式のHDD故障数は含めていません。
＊2　2005年の同じRAID上で7台、5台、4台の故障が起きているものは、同一タイプの製品かつ同一時期の導入品のため、ロット不良の可能性が高いです。2005年にサービス利用から外しました。

故障理由では、メモリ故障による停止や再起動が目立ちます。また、ファン故障による熱暴走、HDD等の接合部分のバックプレーン故障、電源(VRM)故障、プロセッサ故障等多岐に渡っていました。

ローカルHDDを搭載したサーバ機器で起動不可能な障害が発生した場合、現地に赴いてHDDを故障機材とは別の正常なサーバ機材に移設し復旧する必要があります。データセンタに上記のようなサーバ機器のメンテナンスができるオペレータを常時配置するためには多くの人件費がかかります。そのため、遠隔データセンタを視野に入れると、常時オペレータを配置するとコストが増大し、とはいえ、オペレータを配置しなければ、復旧までに多大な時間を費やさなければならないことを意味します。このため、サーバはローカルデータを持たないディスクレス運用を前提に機器選定を行いました。

4.2.3 ネットワーク機器の故障率について

ホスト等を収容するエッジL2スイッチの故障件数ですが、これまでの経験からあまり多くありません。コンデンサ不良等特定の不良ロットに該当したケースを除けば、稼働数から考えても1%未満です。

NIC冗長化設定による冗長化構成を採ることもできますが、NICやネットワーク機器の追加が必要になりコスト増加の要因となります。また、ロードバランサ等を使ってサーバ間での冗長構成を採ったほうが運用が簡単になることが多いため、基本構成ではサーバ収容スイッチは冗長化せず、より高い信頼性が必要なときにのみNIC冗長化設定を実施する設計にしました。また、エッジL2スイッチの故障時の影響を抑えるときには、ロードバランサ等を使って別のサーバ収容スイッチ配下の機器と冗長構成を組むか、障害発生時に障害サーバ収容スイッチ以下のサーバ機器の中身を遠隔操作で別のサーバ収容スイッチ以下のサーバに移動して対応できるよう設計しました。

4.3 NHNの構成

NHNは、遠隔データセンタにも対応できる次のような構成にしました。

- サーバプール方式にして同一スペックのサーバを多数設置する。機器の設置、障害機器の物理交換等の現地作業は、計画作業として月1回等集約できるようにする。
- サーバ構成を画一化して構成管理のコストを抑えるために、物理搭載メモリ量の変更やローカルHDDの搭載等物理構成を変える作業は行わない
- 省電力サーバを利用してラックごとのサーバ収容数を向上
- サービス単位のラック割をやめラックごとのサーバ収容数向上
- 外部業者による機器設置や交換作業をスポットで委託することを想定して、機器や配線を画一化し作業ミスが起きにくいようにする
- Xen、OpenVZ等の仮想化技術と組み合わせ、集約度と自由度を上げる
- VLANを使い、現地での配線変更なしに論理ネットワークを構成できるようにする
- iSCSIを利用して安価なSANを構成する。ストレージは、極力サービス停止が起きないようにRAIDコントローラや接続パスの二重化を必須とする
- ディスクレスサーバは壊れることを前提にする。ハードウエアの障害発生時にはリモートから切り替え作業を行い一次対応が完了できるものを目指す
- ディスクレスサーバの故障発生時に予備サーバへの切り替え完了までの時間がサービス停止として許容できないものであったときには、あらかじめ系の異なる2つのサーバを準備しロードバランサ等を利用して冗長構成を採り、アプリケーション側での冗長構成を実現する
- OSのインストールを含めて、設置以降に必要となる作業をリモートから実行可能にする

次に、それぞれの技術要素について説明します。

4.3.1 iSCSIストレージを利用したIP SANの導入

遠隔データセンタとしての利用を視野に入れた場合、ストレージのサービス停止は長時間の障害に繋がるため、ストレージにはこれまで以上の高い信頼性を求めました。IIJは、IPネットワーク技術に長けていること、またFC SANに比べて安価にシステムが組めることから、iSCSIを利用したシステムを導入しています。そして、RAIDコントローラの二重化、接続パスの二重化を必須条件とすることで、サービス不能に陥るストレージの故障を極力減らすようにしています。

4.3.2 iSCSIストレージと省電力サーバを組み合わせたディスクレスサーバの実現

現行のサーバ機器は、iSCSIブートに対応しているものが少なく、すべてのサーバ機器にiSCSI HBAを装着するとコストがかかります。このため、ほとんどのサーバ機器のオンボードNICが対応しているPXE bootを使い、iSCSIを使ったSAN bootができるよう工夫しています。

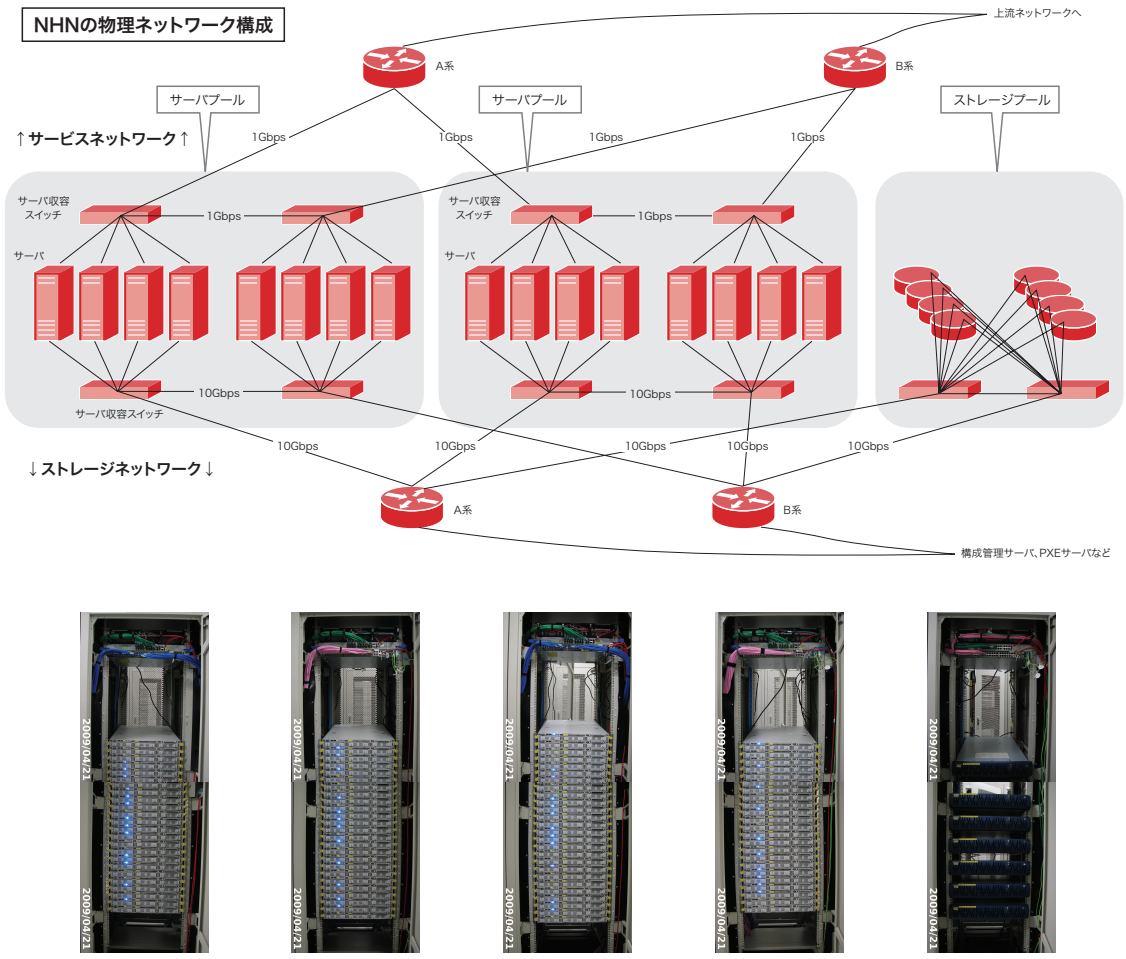
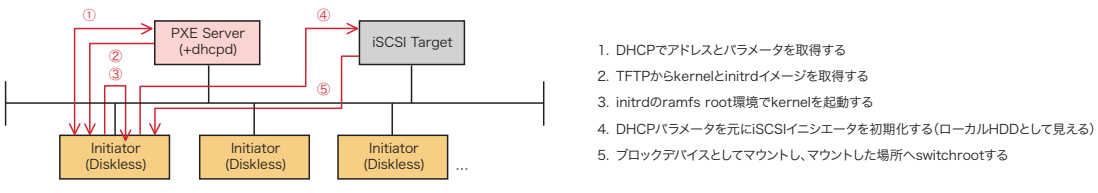


図-1　NHNの物理構成



1. DHCPでアドレスとパラメータを取得する
2. TFTPからkernelとinitrdイメージを取得する
3. initrdのramfs root環境でkernelを起動する
4. DHCPパラメータを元にiSCSIイニシエータを初期化する（ローカルHDDとして見える）
5. ブロックデバイスとしてマウントし、マウントした場所へswitchrootする

図-2　PXE bootとiSCSIによるディスクレスサーバ（Linux）の起動

iSCSIブートに必要なiqn[*3]、IPアドレス等の情報は、DHCPオプションを使って情報を渡します。これによって、構成情報を書き換えて予備サーバを起動することで、故障したサーバ機器と予備機を入れ替えることができます。

サーバ故障に伴う現地作業を回避するため、NHNではサーバ自体をディスクレス構成にしローカルデータを持たないようにしました。仮にサーバ機器の1台に故障が発生したときには、リモートから構成情報を書き換えて再起動することで、すでに設置済みの予備サーバを故障機器のストレージの内容を保持したまま起動できます。

4.3.3 VLANを使った配線変更不要な仮想ネットワークの実現

NHNでは、1本の物理配線に複数のネットワークを同居させることで、現地での配線変更を不要なものにし、リモートのみでの対応を可能にしています。技術的には、サーバごとにaccess VLANとtrunk VLANを使ってネットワークを構成し、必要に応じて1本の物理配線に複数のネットワークを同居させます。

VLAN自体は目新しい技術ではありませんが、NHNではIIJが管理する構成情報と連動して自動的に該当ホストを収容しているサーバ収容スイッチのVLAN設定を書き換える仕組みを実装しました。これにより、ネットワーク機器のVLAN設定のミスを減らし、運用コストを削減しています。

執筆者:
牧野　泰光(まきの やすみつ)
IIJサービス事業統括本部 システム基盤統括部 システム基盤運用課 課長
IIJ法人サービス、個人サービスのサーバインフラ設計、運用業務に従事。2008年度よりサービスホストの設備調達、現地構築業務等をシステム基盤運用課に集約し、基盤システム化していくことで設備の集約、運用の効率化を推進。

花高　信哉(はなたか しんや)
IIJ サービス事業統括本部 システム基盤統括部 システム基盤運用課
小林　直(こばやし ただし)
IIJ サービス事業統括本部 システム基盤統括部 システム基盤運用課

4.4 導入効果

IIJでは、2008年度に今後の遠隔データセンタ利用を視野に入れ「NHN」というキーワードでサービスホスト構成の大幅な見直しを実施し、新規サービスと既存サービス移行のための基盤整備を行いました。

NHNでは、サーバ運用者の視点と経験を元に最新の技術を取り入れ、仮想化、iSCSIを使ったディスクレスサーバ、ホスト情報の集中管理、ネットワーク機器のVLAN設定の動的変更等を導入しました。

NHNを社内サービスに導入した当初、それまでのサービスホスト構成との違いが浸透していなかったため、次のような意見を聞くことがありました。

- 仮想化サーバでは不安。物理サーバにしてほしい
- これほどの性能は必要ないので、もっと安価なサーバにしてほしい
- メモリ量が多すぎるので減らしてほしい
- ローカルディスクに比べてディスク単価が高い

当初はこのような意見が聞かれましたが、個々のサーバが多少高価なものであっても、仮想化の導入による集約、ファシリティ、運用コストまで含めたコストを考慮すると安価になることが理解され、現在では社内導入の敷居は低くなっていると思います。

また、サーバを要求した数日後には利用可能になり、不要になった時点でデータセンタでの物理作業なしに返却できるメリットが浸透してきています。それまでは、サービスごとにばらばらの機器を使って設備を構築していたため、計画変更による他サービスへの転用が難しかったり、急に機器が必要になっても購入待ちや設置待ちですぐに利用できませんでした。NHNの導入によってサービスの開発フローや設備増強の需要予測の仕組みも変わりつつあります。

IIJは、引き続きセキュリティ部分やI/Oの仮想化部分等を詳細に検討し、IIJ GIO等のお客様への提供に向けて完成度をより高めていきます。

＊3　iSCSI Qualified Nameの略。今回はiSCSIターゲットを一意に識別するために利用します