Windows Server 2016新機能
「記憶域スペースダイレクト」と
Lenovo x86サーバーで実現する
ハイパーコンバージド・インフラストラクチャ

実機による動作検証レポート

レノボ・エンタープライズ・ソリューションズ株式会社 ｜ マイクロソフト株式会社

検証協力
MKTインターナショナル株式会社
株式会社アクセライト
ハセゴーイノベーション
監修
日本仮想化技術株式会社

２０１７年3月作成

ITの効率性と生産性の向上

目次

はじめに …… 3
本検証の目的 …… 3

Lenovo x86サーバーとは …… 4

Windows Server 2016の特徴 …… 4

記憶域スペースダイレクト(Storage Spaces Direct, S2D)とは …… 5
記憶域スペースダイレクトの特徴 …… 5
記憶域スペースダイレクトを構成するソフトウェアスタック …… 6
様々な要件に対応できる回復性オプション …… 7
RDMAによるパフォーマンスの向上 …… 7

記憶域スペースダイレクトを使ったHCI構成の検証 …… 8
評価に利用した検証環境の概要 …… 8
検証環境構成 …… 8
ベンチマークで利用するロードジェネレーター …… 9
記憶域スペースダイレクトとHyper-VによるHCIの構成と基本性能の確認 …… 9
記憶域スペースダイレクトでのクラスター共有ボリュームの作成 …… 11
・記憶域スペースダイレクト機能の有効化 …… 11
・クラスター共有ボリュームの作成 …… 11
・回復性オプションごとのI/Oスループットの比較 …… 13
ハイパーコンバージド・インフラストラクチャの構成 …… 15
・仮想マシンの作成 …… 15
・仮想マシンの電源操作、移動 …… 16

ディスク障害時およびノード障害時の振る舞いの確認 …… 16
ドライブ障害時の振る舞い …… 16
ノード障害時の振る舞い …… 18
ノード障害時のHyper-V仮想マシンの振る舞い …… 19

システムメンテナンスに関する確認 …… 20
ノードのシャットダウン …… 20
クラスター対応更新による自動ローリングアップデート …… 20
ノードの追加 …… 21
データのリバランス …… 22

まとめ …… 24

推奨構成 …… 25

奥付 …… 26

はじめに

本ホワイトペーパーでは、Lenovo x86サーバーとWindows Server 2016 Datacenter Editionを組み合わせて、最新のソフトウェア定義ストレージ「記憶域スペースダイレクト」を用いたハイパーコンバージド・インフラストラクチャ（HCI）の実現性、およびその可能性について検証しました。

ここ数年、HCIが注目を集めています。なぜなら、従来は個別に構成し、運用する必要があったサーバー、ストレージ、ネットワーク、および仮想化ソフトウェアや管理ツールなどをあらかじめパッケージ化したHCIは、システムの構成もシンプルとなるため導入も容易で、短時間で使い始めることができるからです。また、システム内部の高度なアーキテクチャーが隠蔽されているため、容易に運用できる点も注目点のひとつです。現在、様々なベンダーから、HCIを実現するソリューションが提案され、それぞれの特徴に応じて、予算・用途・環境などに合わせて導入されています。

今日ITシステムの管理者を悩ませている大きな要素として、企業にとって、その消失が死活問題となるデータを保存する基盤となるストレージシステムがあります。ストレージシステムには、特に完全性、信頼性や可用性が求められます。また、現在のITシステムでは、ネットワークを通じて複数のシステムが情報を共有することも多く、データを必要なシステムに供給しつづけること自体がひとつの課題となっています。

このような背景のなか、HCIでは、ストレージシステムにソフトウェア定義ストレージ（Software Defined Storage：SDS）と呼ばれる、複数のサーバーに接続されたローカルストレージ（SSDやハードディスクドライブ）で分散ストレージシステムを構成するテクノロジーが利用されます。具体的なSDSの例として、最新のオペレーティングシステムであるWindows Server 2016が新規に搭載した記憶域スペースダイレクトが挙げられます。

本検証の目的

今回、Windows Server 2016 Datacenter Editionに新たに搭載された記憶域スペースダイレクト（S2D）を使ってWindows Serverの標準機能のみを利用したハイパーコンバージド・インフラストラクチャ（HCI）の検証を実施しました。

本検証は、Lenovo x86サーバー上でWindows Server 2016を使ったHCI構築を検討されているお客様に、安心して検討を進めていただくための情報提供を目的として実施しました。Windows Server 2016によるHCI構成のアーキテクチャーおよび基本的な構築手順を確認するとともに、信頼性の高いシステムを運用していく上で重要な、障害時の挙動と性能の推移を確認しました。また、ノード追加やメンテナンスの方法について実機で確認しています。

Lenovo x86サーバーとは

本検証では、Windows Serverシリーズとともに広く利用されているLenovo x86サーバーを使用しました。

Lenovo x86サーバーは、大型コンピュータなどの分野で長年にわたる実績を持つIBM社が手がけてきた高機能のx86サーバーです。IBM社のx86サーバー部門がレノボへと統合された現在、サーバー事業を主導していた米本社をはじめ、当時の製品開発を支えていた研究開発施設やエンジニア、さらには製造・供給体制までレノボへとそのまま継承されています。このように、すべての要素がレノボへと引き継がれたx86サーバーは、今もなおIBM社の伝統を受け継ぐ安心・信頼のサーバー製品です。

ITIC（Information Technology Intelligence Consulting）社によるグローバル・サーバー・ハードウェアの信頼性調査でも、Lenovo System x サーバーは2016年下半期まで8年連続で、もっとも信頼性が高く、また連続稼働時間の長いx86サーバーとして評価されています。

Windows Server 2016の特徴

Windows Server 2016は、高度なセキュリティ機能と、オンプレミスとクラウドのシームレスな連携を実現すると同時に、データセンター全体をソフトウェアで制御する、Software Defined Datacenter（SDDC）をOSの標準機能のみで実現することができる画期的なオペレーティングシステムです。レノボとMicrosoftは、長きにわたりパートナー関係を結んでいます。Windows Server 2016についても共同開発を実施しています。

記憶域スペースダイレクト(Storage Spaces Direct, S2D)とは

SDDCを実現するにあたり、核となる機能の一つが記憶域スペースダイレクトと呼ばれるソフトウェア定義ストレージ(Software Defined Storage, SDS)の機能です。

記憶域スペースダイレクトの特徴

記憶域スペースダイレクトは、Windows Server 2016 Datacenter Editionに新たに搭載されたもので、Windows Server 2016を実行する複数のサーバー上のローカルストレージと、そのサーバー間を繋ぐネットワークを用いて、分散ストレージを構築する機能です。

これにより、高価なNASやSANストレージを別途用意することなく、Windows Server、x86サーバーおよびそのローカルストレージだけで、高性能かつ信頼性が高い分散ストレージを低コストで構築できるようになりました。

さらに、記憶域スペースダイレクトが動作しているサーバー上で、Hyper-VやSQL Serverなどのアプリケーションワークロードを同時に実行する、いわゆるHCI構成も可能であり、システム構成の大幅な簡素化と高い柔軟性を得られるようになりました。

従来、高信頼性のクラスターを構築するには、x86サーバーに加えて、SAN(Storage Area Network)やDAS(Direct Attached Storage)を使った共有ディスクを設置する必要がありました。ただ、SANストレージの構築は一般的に高価であり、DASの共有ディスクはシステムの拡張性が制約されるなどの課題がありました。

Windows Server 2016の記憶域スペースダイレクトは、サーバーに内蔵されたローカルストレージをクラスターが共有する分散ストレージとして利用できるようにすることで、この課題を解決します。

また、ひとつのサーバークラスターが分散ストレージとアプリケーションクラスターを兼ねるHCI構成では、ストレージ容量やコンピューティング性能が不足した際には、サーバーを追加するだけで容量や性能を拡張できます。サーバーやディスクの故障時には、正常なコンポートに交換するだけでクラスターを正常な状態に復旧できます。このため、システムを構成するサーバーやストレージの運用がシンプルになり、運用の大幅な効率化を実現できます。

記憶域スペースダイレクトは、Windows Server 2016とx86サーバーの可能性をさらに広げ、データセンターの運用に革新をもたらすでしょう。

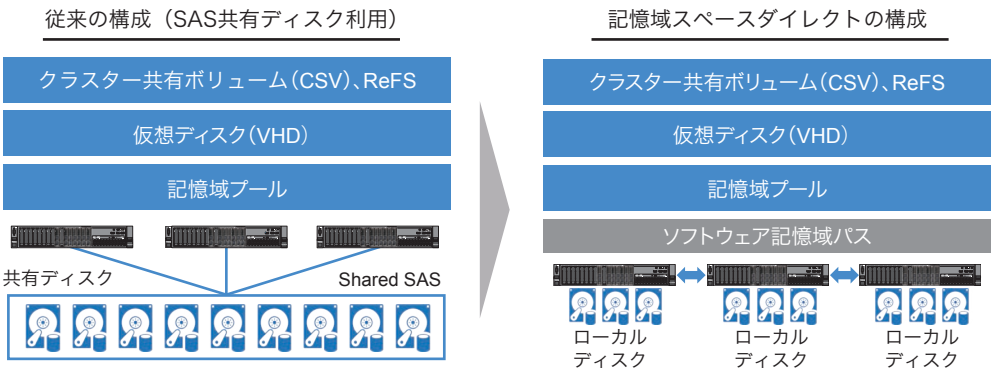


図１　共有ディスクを使う従来の構成と記憶域スペースダイレクトの構成の比較

記憶域スペースダイレクトを構成するソフトウェアスタック

記憶域スペースダイレクトは、新しい機能ではありますが、個々に利用されているソフトウェアスタックを見れば、これまで実績を積み重ねてきたWindowsのストレージ機能の組み合わせで実現されています。

各サーバーに接続されたローカルストレージは、過去のWindows Serverで実績をもつ「記憶域プール」により抽象化されて管理されます。それらをWindowsのファイル共有プロトコルであるSMBを用いて、サーバーをまたいで共有することでクラスター化し、分散ストレージが実現されます。

このクラスター化された「記憶域プール」から「仮想ディスク（VHD）」が作成され、その上にクラスター内で共有されるデータ領域である「クラスター共有ボリューム（CSV）」が作られます。クラスター共有ボリュームは1台のサーバーが停止したり、ローカルストレージが故障したりしても、アクセスが継続できる高信頼性のストレージとして利用できます。従来のクラスター共有ボリュームは、各サーバーが物理的に共有するDAS上に構成されるものでしたが、Windows Server 2016の記憶域スペースダイレクトでは、クラスター化された記憶域プールの上にも構成できるようになりました。

記憶域スペースダイレクトの主な利用用途として、Hyper-VやSQL Serverクラスターのデータ領域に利用しHCIを実現したり、スケールアウトファイルサーバー（SOFS）と併用して他のサーバーからアクセスを可能にしたりするなど、高信頼性かつ高性能のアプリケーションクラスターを構築する事などが挙げられます。

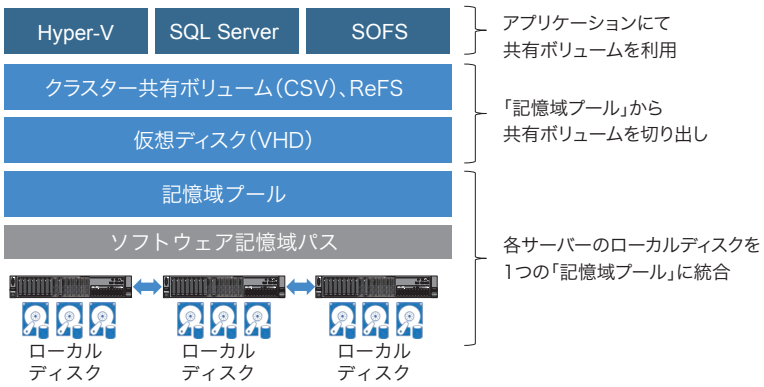


図２　記憶域スペースダイレクトのソフトウェアスタック概要

様々な要件に対応できる回復性オプション

記憶域スペースダイレクトでは、回復性オプションと呼ばれるデータ冗長化方式が複数用意されています。回復性オプションには主に3種用意されており、データの特性にあわせて選択できます。

1）双方向ミラー

双方向ミラーでは、データのコピーをクラスター上に2つ作成し、どちらか一方が障害となってもデータが失われることはありません。双方向ミラーでは、データのコピーがクラスター中に2重保存されるため、ディスクの物理容量のおよそ2分の1が実効容量となります。

2）3方向ミラー

3方向ミラーでは、合計3つのデータのコピーをクラスター上に作成するため、同じデータに対して二重障害に対する耐性を持ちます。双方向ミラーよりも高い可用性を示す一方、データのコピーがクラスター中に3重保存されるため、ディスクの物理容量のおよそ3分の1が実効容量となります。

3）パリティ

パリティは、保存対象データに加え、ディスクやノードが喪失した場合に備え、データを回復するためのパリティデータを計算し保存しておく方法です。従来、分散パリティの実装としてRAID-5、RAID-6が利用されてきましたが、Windows Server 2016ではより高効率で信頼性が高い、独自のイレージャーコーディングが適用されます。3台構成でシングルパリティを選択した場合は、実際に使える実効容量は物理容量のおよそ3分の2、ノード数が16台まで増えると最大80%の実効容量を確保でき、より多くのデータを格納できます。さらに、デュアルパリティ構成など、より多くのパリティデータを保持しておくことにより、多重障害に対して耐性を持たせることが可能です。

パリティの回復性オプションでは、データ書き込み時にパリティデータの計算が必要となり、双方向ミラーや3方向ミラーと比較すると書き込み性能が低い傾向にあります。このため、大量のデータを保存し、読み込み中心で利用するアーカイブのような用途にパリティの回復性オプションは向いています。定常的にディスク書き込みが発生する場合など、高い書き込みの性能が必要な場合にはミラー構成を選ぶとよいでしょう。

RDMAによるパフォーマンスの向上

記憶域スペースダイレクトは、一般的なファイル共有でも用いられるSMBプロトコルでストレージレイヤでのデータ通信を行います。Windows Server 2016ではSMB 3.0に対応しており、サーバー間のI/O性能を高めています。

記憶域スペースダイレクトは、高いI/O性能要件が求められる場合には、Windows Server間でRoCEもしくはiWARPと呼ばれるRDMA（Remote Direct Memory Access）機能を活用する「SMBダイレクト」が利用できます。

SMBダイレクトでは、RDMA機能を利用できるネットワークインターフェースおよびネットワークスイッチを利用することで、従来CPUで処理していた通信処理をネットワーク機器側にオフロードし、サーバーハードウェアが本来持っている高いポテンシャルを引き出せます。その結果、ディスクI/Oスループット、IOPS、そしてレイテンシー性能が向上するほか、CPU負荷も軽減され、システム全体の効率が改善します。

記憶域スペースダイレクトを使ったHCI構成の検証

今回、Lenovo x86サーバーを用いて、記憶域スペースダイレクトおよびHyper-Vを使ったハイパーコンバージド・インフラストラクチャ（HCI）の検証を実施しました。

評価に利用した検証環境の概要

検証環境構成

本検証では、記憶域スペースダイレクトを構成するサーバーとして、下記構成のサーバーを3台利用しました。

Lenovo System x3650 M5 (5462)	
CPU	E5-2620V3 2.4GHz 6core x2
Memory	128GB(16GB x8)
Disk	Boot : 120GB SSD(00YC365) x1 Data : SAS 300GB 12Gb 10K x6 Cache :NVMe P3700 400GB(00YA818)x2
HBA	N2215 SAS/SATA HBA(PCI-E)(47C8675)
Ethernet	40Gb 2port Mellanox ConnectX-3 Pro ML2 2x40GbE/ FDR VPIアダプター(00FP650) 1Gb 4port (Onboard)

図３　今回の検証構成

ディスクは、OSブート用のSAS接続SSDに加え、記憶域スペースダイレクトで利用するストレージ領域として、高速なNVM Express(NVMe)接続のSSDを2本、SAS接続のハードディスクを6本使用しました。

記憶域スペースダイレクトでは、NVMe SSDとハードディスクのように性能が異なるディスクが混在している場合、より高速なディスクがジャーナル（キャッシュ）として利用され、もう一方がデータ格納領域として利用されます。このため、本構成では、NVMe接続SSDがジャーナル、HDDがデータ領域として利用されます。

記憶域スペースダイレクトを構成するにあたり、HBAの選定には注意が必要です。記憶域スペースダイレクトではOSが直接ディスクを操作するため、RAIDカードではなく、直接ディスクを接続できるHBAの利用が必須要件となります。実際のシステム構成では、システムの起動領域をRAIDで保護するためにハードウェア処理のRAIDカードに接続し、記憶域スペースダイレクト向けのディスクはHBA経由で接続する、などと使い分けていくとよいでしょう。

管理用のネットワークにはGigabit Ethernet(GbE)、記憶域スペースダイレクトのノード間接続にはRoCEに対応した40 Gigabit Ethernet(40GbE)を利用しました。また、クラスター管理用に別途用意したActive Directoryサーバーを Gigabit Ethernetで接続しました。以下にネットワーク構成図を示します。

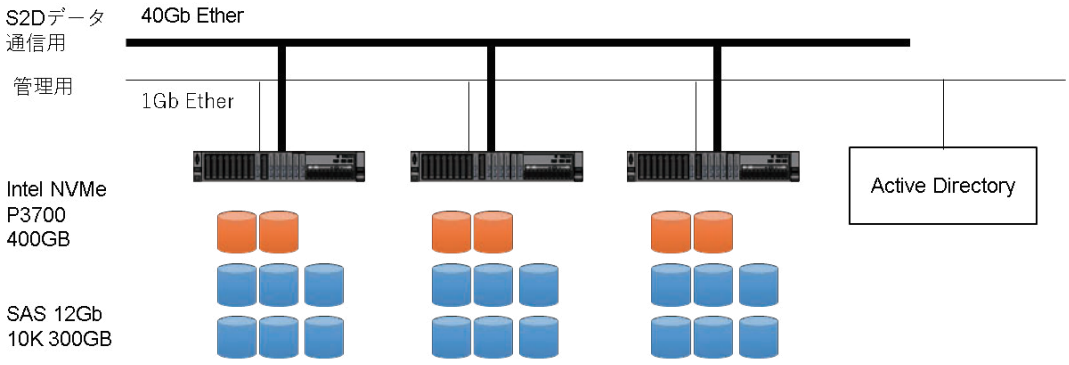


図4　検証時のネットワーク構成

ベンチマークで利用するロードジェネレーター

本検証では、ベンチマークにfio(Flexible I/O Tester)を利用しました。fioは、柔軟なパラメーター設定とログ出力ができ、かつ、高速なディスクに対し、十分な負荷をかけることができる、高性能なI/Oベンチマークソフトウェアです。

- Flexible I/O Tester
  https://github.com/axboe/fio

クラスター共有ボリュームそのものをベンチマーク対象とする場合は、Windows版のfioを用いました。仮想マシン上のI/Oをベンチマーク対象とする場合は仮想マシンにCentOS 7を導入し、Linux版のfioを用いて測定を行いました。

記憶域スペースダイレクトとHyper-VによるHCIの構成と基本性能の確認

まず、今回の検証対象となるシステムで記憶域スペースダイレクトを構成しました。以下にシステムの構成を示します。

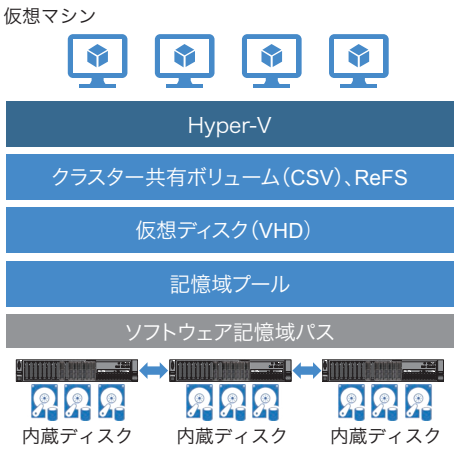


図5　記憶域スペースダイレクトとHyper-VによるHCI構成

構築手順の概要は以下の通りです。

① フェールオーバークラスターのセットアップ
▷ 各サーバーにWindows Server 2016 Datacenter Editionをインストールし、Active Directoryドメインに参加させる
▷ 各サーバーにHyper-Vの役割とフェールオーバークラスタリングの機能を追加し、フェールオーバークラスターを構成する
② 記憶域スペースダイレクトによるクラスター共有ボリュームの作成
▷ 記憶域スペースダイレクト機能の有効化
▷ クラスター化された仮想ディスク(VHD)を作成し、クラスター共有ボリューム(CSV)を作成する
③ Hyper-V仮想マシンの構成
▷ クラスター共有ボリューム上に仮想マシンファイルを配置し、高可用性仮想マシンを作成する

記憶域スペースダイレクトを構成するには、あらかじめフェールオーバークラスターを作成し、クラスターを構築する必要があります(①)。フェールオーバークラスターの作成方法については、下記のリソースなどをご参照ください。

- フェールオーバークラスターを作成する
  https://msdn.microsoft.com/ja-jp/library/dn505754(v=ws.11).aspx

注:Test-Clusterコマンドレットでテスト対象を指定する際、一部のテスト項目はローカライズ(日本語)された名称で指定する必要がありました。たとえば System Configurationを指定する場合、日本語版Windows Serverでは システムの構成 と指定する必要があります。

フェールオーバークラスターが構成できたら、記憶域スペースダイレクトを構成します(②)。またHCIを実現したい場合、記憶域スペースダイレクトに加えてクラスター共有ボリューム(CSV)を構成します。記憶域スペースダイレクトの構成には、PowerShellコマンドレットやフェールオーバークラスターマネージャーなどを使用します。

ここまでの準備ができたら、Hyper-Vマネージャなどから、Hyper-V仮想マシンをクラスター共有ボリュームに配置します(③)。Hyper-VやWindows Serverになじみがない方でも、比較的少ない学習コストでハイパーコンバージド・インフラストラクチャを構成することができるでしょう。

記憶域スペースダイレクトでのクラスター共有ボリュームの作成

《記憶域スペースダイレクト機能の有効化》

フェールオーバークラスターのセットアップ後、PowerShellからコマンドレットを実行し記憶域スペースダイレクトを有効化します。管理者権限が必要なため、スタートメニューの「Windows PowerShell」を右クリックし、ポップアップメニューから「管理者として実行」を選択します。シェルが起動したら、以下のコマンドレットを実行してください。

Enable-ClusterStorageSpacesDirect

これにより、記憶域スペースダイレクトの機能が有効化されるとともに、クラスター横断の分散ストレージ領域として自動的に「S2D on <クラスター名>」にて「記憶域プール」が作られ、「正常性プロバイダー」によるチェックが行われます。

自動的に作成された記憶域プールには、クラスターに存在する未使用のディスクがすべて追加され、結果がHTML形式のレポートとして出力されます。また、ディスク追加時に、速度の異なるディスクがある場合、より高速なドライブが自動的にキャッシュとして利用されます。ユーザ側での階層化等の設定は必要ありません。

《クラスター共有ボリュームの作成》

クラスター共有ボリュームを作成するためには、仮想ディスクを作成し、その仮想ディスクを初期化して読み書き可能なボリュームを作り、それをクラスター共有ボリュームに追加するという一連の作業が必要です。この作業には「New-Volume」コマンドレットを使用します。

New-Volume-StoragePoolFriendlyName<記憶域プール名>-FileSystem <FileSystem> -FriendlyName <ボリューム名> -ResiliencySettingName <冗長化方式、Mirror/Parityのいずれか> -Size <ボリュームサイズ> -PhysicalDiskRedundancy<冗長データの数>

以下にいくつかの実行例を示します。

・「S2D」で始まる名称の記憶域プールから「vol_mirror2」という名称で300GBの双方向ミラーのボリュームを作成する
New - Volume - Size 300GB - StoragePoolFriendlyName"S2D*" - FriendlyName "vol_mirror2" - FileSystem CSVFS_ReFS - ResiliencySettingName Mirror - PhysicalDiskRedundancy 1

・ 同様に「vol_mirror3」という名称で300GBの3方向ミラーのボリュームを作成する
New - Volume - Size 300GB - StoragePoolFriendlyName "S2D*" - FriendlyName "vol_mirror3" - FileSystem CSVFS_ReFS - ResiliencySettingName Mirror - PhysicalDiskRedundancy 2

・ 同様に「vol_parity3」という名称で300GBのパリティのボリュームを作成する
New-Volume - Size 300GB - StoragePoolFriendlyName "S2D*" - FriendlyName "vol_parity3" -FileSystem CSVFS_ReFS -ResiliencySettingName Parity - PhysicalDiskRedundancy 1

上記のコマンドレットを用いず、仮想ディスクの作成、ボリュームの作成、クラスター共有ボリュームへの追加の作業をそれぞれGUIから行うことも可能です。ただし、本ホワイトペーパー執筆時バージョンのWindows Server 2016では、GUIから操作した場合に回復性オプションを指定できない制約が確認されました。今後のWindows Serverのリリースで改善されていく点と考えられますが、現時点では、PowerShell上で作業するとよいでしょう。

仮想ディスクやクラスター共有ボリュームの作成に成功すると、クラスターのすべてのノードの「C:\ClusterStorage\」以下に「Volume1」、「Volume2」と作成順の連番で、クラスター共有ボリュームが作成されます。

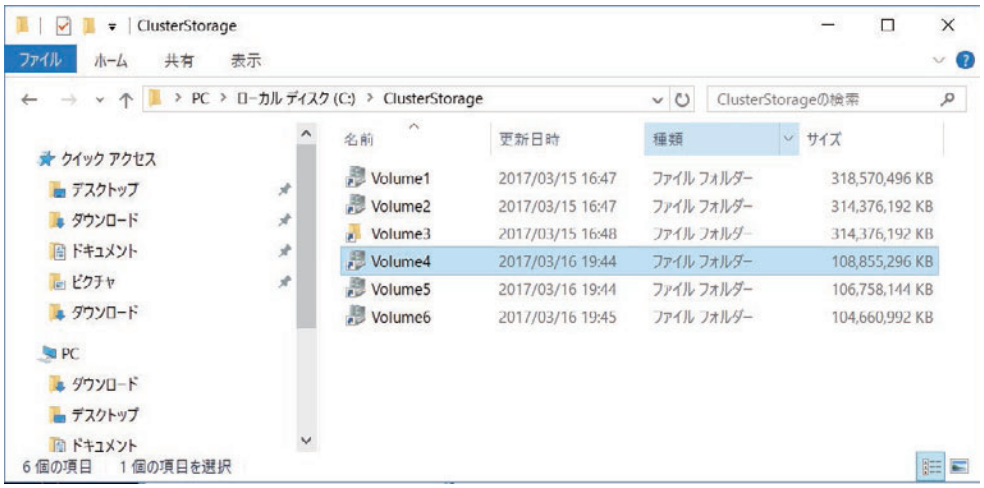


図6　クラスター共有ボリュームのエクスプローラー上の表示

エクスプローラー上からは、クラスター管理ボリュームも通常のフォルダと同じように見えますが、ここに書き込まれたデータは、仮想ディスク作成時の回復性オプションに従ってクラスター内に分散配置され、いずれのノードからも読み書きできます。

これらのボリュームは、フェールオーバークラスターマネージャーから以下のように確認できます。

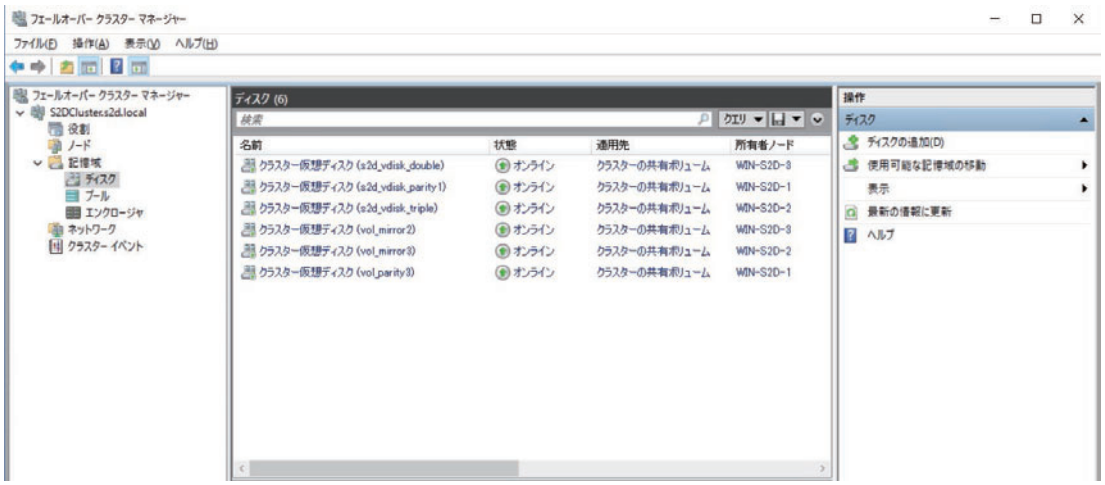


図7　フェールオーバークラスターマネージャーでのクラスター共有ボリュームの表示

《回復性オプションごとのI/Oスループットの比較》

先に構成した記憶域スペースダイレクトのクラスター共有ボリュームに対してI/O負荷をかけ、スループットを測定しました。なお、このスループットは本検証環境における参考値です。記憶域スペースダイレクトの最大性能を示すものではないことにご注意ください。

本検証環境にて、スレッド数を増加させながら、読み込みおよび書き込み速度を測定した結果を以下に示します。

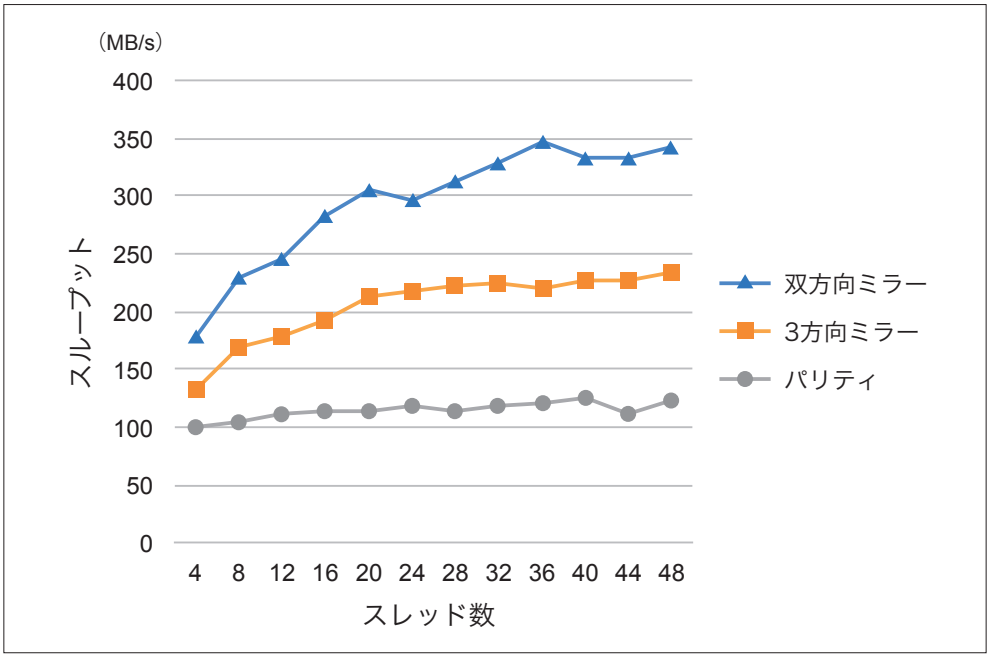


図８　記憶域スペースダイレクトに対する書き込みスループット

【ベンチマーク条件】
・ワークロード(fioオプション)
モード:ランダム書き込み(--rw=randwrite)
スレッド数:4～48(--numjobs=4～48)
ブロックサイズ:1MB(--bs=1M)
メモリキャッシュを利用しない(--direct=1)

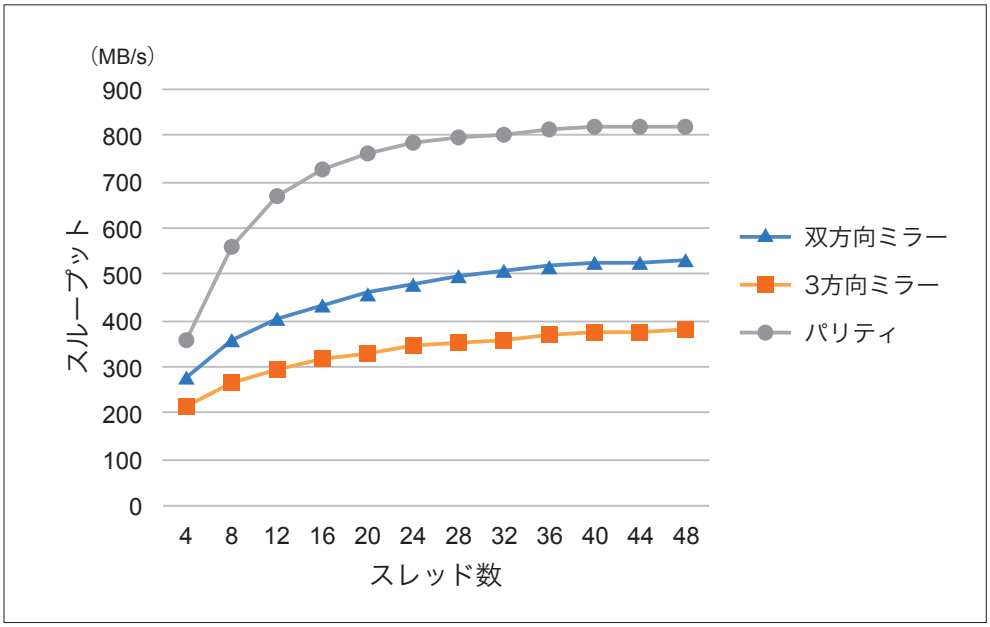


図 9　記憶域スペースダイレクトに対する読み込みスループット

【ベンチマーク条件】
・ワークロード（fioオプション）
モード：ランダム書き込み（--rw=randwrite）
スレッド数：4～48（--numjobs=4～48）
ブロックサイズ：1MB（--bs=1M）
メモリキャッシュを利用しない（--direct=1）

今回の条件では、2方向ミラーが高い書き込みスループットを示しました。一方、読み込みスループットではパリティ（シングルパリティ）が高い値を示しました。

3方向ミラーは、今回測定したなかで二重障害に対する耐性を唯一持つ回復性オプションですが、I/O性能と対障害性の両方でバランスがとれた結果となりました。

ハイパーコンバージド・インフラストラクチャの構成

HCIにおける利用を想定したテストとして、仮想マシンを共有クラスターボリューム上に作成できることを確認しました。ここでは、仮想マシンを記憶域スペースダイレクトの共有クラスターボリューム上に作成する手順、およびその結果例を示します。

《仮想マシンの作成》

仮想マシンは、フェールオーバークラスターマネージャーより作成します。フェールオーバークラスターマネージャー上に表示されているクラスターから「役割」を選択、右ペインの「仮想マシン...」と表示されているメニューをクリックし、「仮想マシンの新規作成(V)」を選択します。

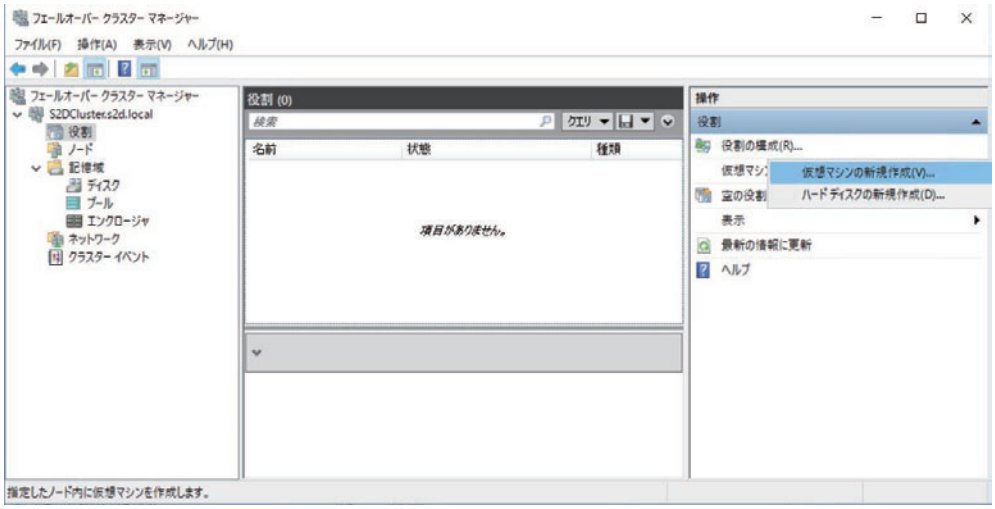


図 10　仮想マシンの作成(フェールオーバークラスターマネージャー)

仮想マシンをクラスター共有ボリューム上に作成するため、仮想マシンの新規作成ウィザードで最初に表示される「名前と場所の指定」ページにて、「仮想マシンを別の場所に格納する(S)」をチェックします。格納先として、先ほど作成した「C:\ClusterStorage\」以下のクラスター共有ボリュームのパスを指定します。その他のパラメーターについては、従来のHyper-V仮想マシンを作成する場合と同様です。

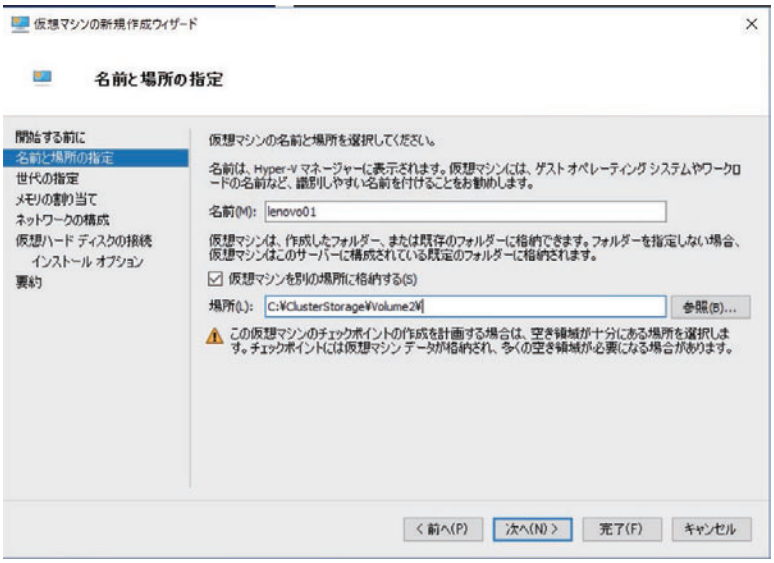


図 11　高可用性仮想マシンの領域としてクラスター共有ボリュームを指定

《仮想マシンの電源操作、移動》

クラスター共有ボリュームに作成した仮想マシンは従来どおりパワーオン／オフでき、仮想マシンを別のクラスター内のノードに無停止で移動（ライブマイグレーション）することもできます。さらに、仮想マシンが動作している物理サーバーに障害があった際も、別のノードで同じ仮想マシンを立ち上げ直すことでシステムを復旧でき、対障害性を高められます。

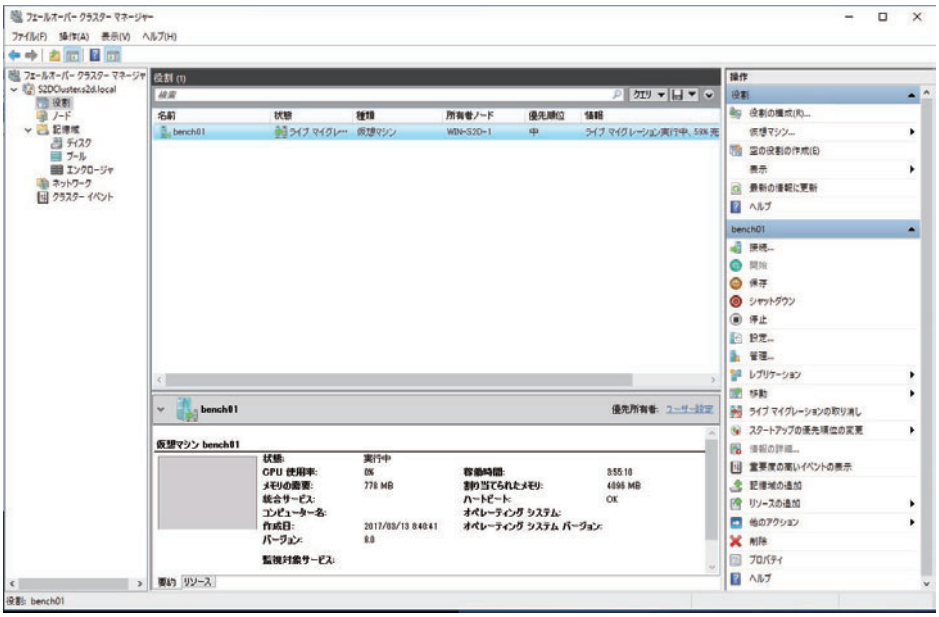


図12　ライブマイグレーション動作中のフェールオーバークラスターマネージャー

以上の手順で、Windows Server 2016を用いたハイパーコンバージド・インフラストラクチャの構築は完了です。

ディスク障害時およびノード障害時の振る舞いの確認

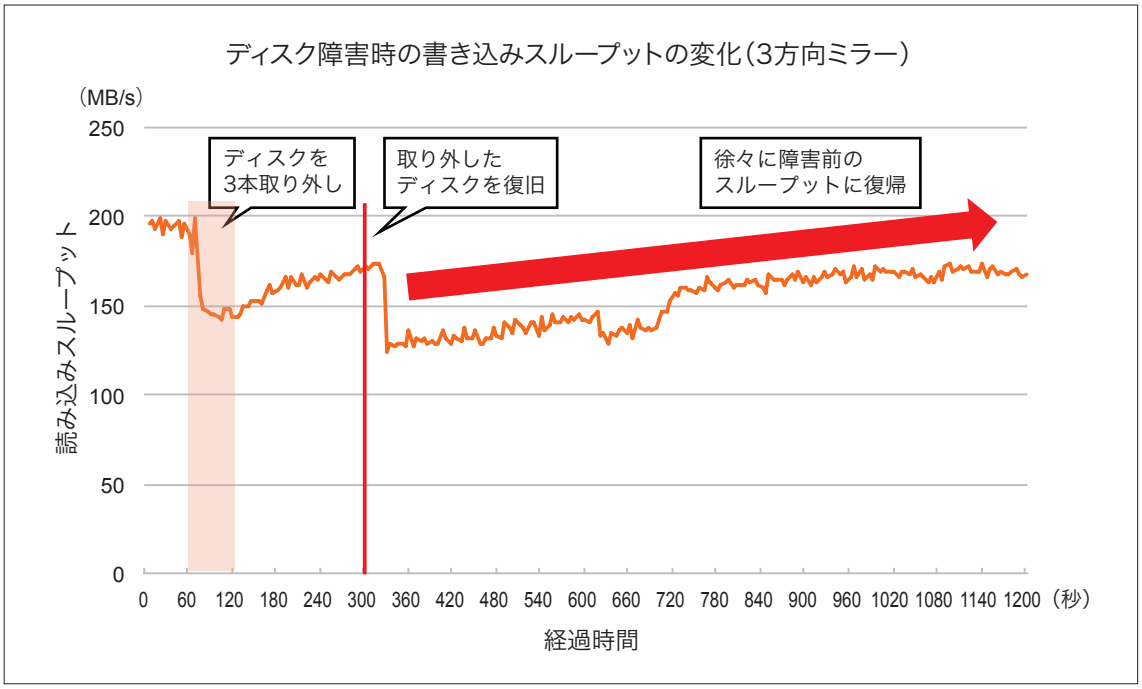
分散ストレージで管理者が頭を悩ませる課題が、システム障害です。特に、記憶域スペースダイレクトのように多数のドライブを統合し単一のストレージとして集中管理するアプローチでは、記録メディアであるSSDやハードディスクドライブの障害を運用フェーズのイベントとして織り込んでおく必要があります。

今回、記憶域スペースダイレクトがドライブ障害やクラスターノード障害の際にどのようにふるまうかを検証しました。

ドライブ障害時の振る舞い

記憶域スペースダイレクトを構成したクラスターに対するドライブ障害時の動作確認として、特定のドライブへのアクセスが失われた場合のI/Oスループットの変化を確認しました。

クラスター共有ボリューム（3方向ミラー）に対して書き込み負荷をかけた状態で、ディスクの抜き差し、および、クラスター中のノードの1台に擬似的な障害を発生させた結果を次に示します。



【ベンチマーク条件】
・ワークロード（fioオプション）
モード：ランダム書き込み（--rw=randwrite）
スレッド数：4（--numjobs=4）
ブロックサイズ：1MB（--bs=1M）
メモリキャッシュを利用しない（--direct=1）

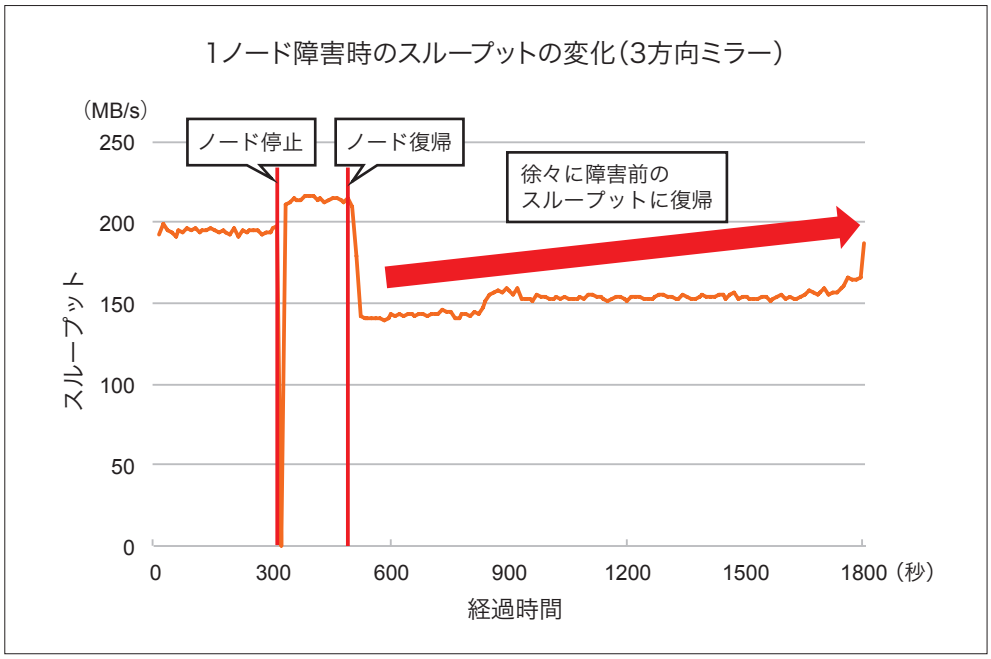
・測定条件
書き込み先に3方向ミラーのクラスター共有ボリュームを利用。ベンチマーク開始1分後、1分30秒後、2分後にそれぞれ1本ずつディスクをスロットから取り外し、ベンチマーク開始から5分後に取り外したディスクをもとのスロットに復旧

ディスクを取り外した時点でスループットが4分の1程度低下しましたが、ボリュームへの書き込みは継続されていることがわかります。

改めてディスクを取り付けると、記憶域プール上でディスクがオンライン状態に復旧しました。その際、一時的に書き込みスループットが落ちましたが、時間が経過すると徐々に障害前のスループットまで回復する様子が確認できました。

ノード障害時の振る舞い

クラスター共有ボリューム(3方向ミラー)に対し、書き込み負荷をかけた状態でノード障害を発生させた結果を以下に示します。



【ベンチマーク条件】
・ワークロード(fioオプション)
モード:ランダム書き込み(--rw=randwrite)
スレッド数:4(--numjobs=4)
ブロックサイズ:1MB(--bs=1M)
メモリキャッシュを利用しない(--direct=1)

・ 測定条件
書き込み先に3方向ミラーのクラスター共有ボリュームを利用。ベンチマーク開始 5分後にサーバーの電源を落とし、疑似的に障害を発生させ、その直後に再起動し、ノードをクラスターに復帰させた。

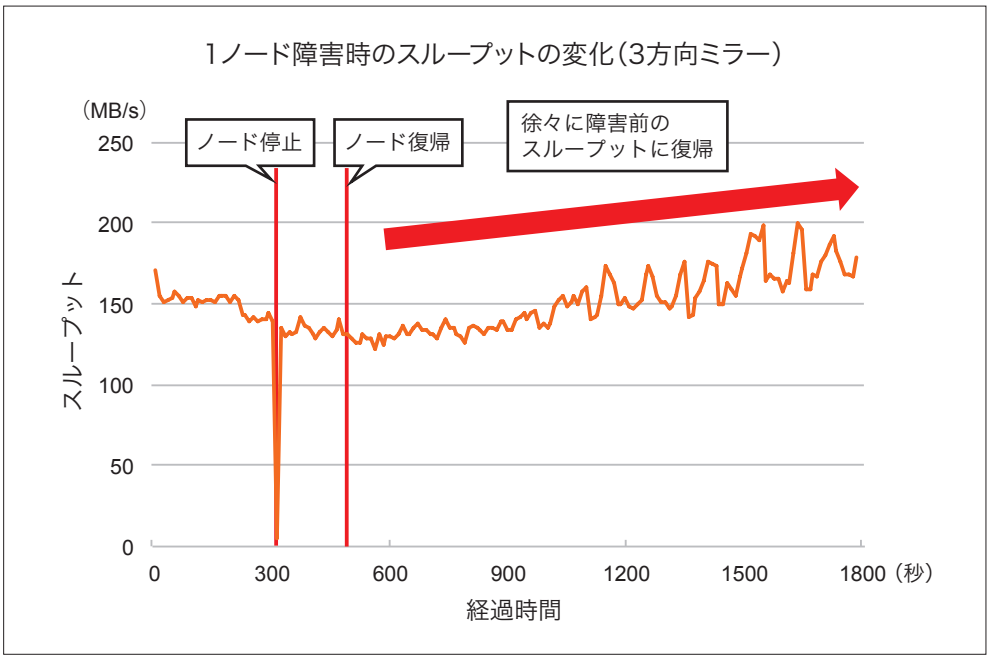
障害が起こった瞬間、1～2秒ほど書き込みスループットがゼロになる瞬間が確認されましたが、その後すぐ、障害前に近いスループットに戻りました。書き込みレイテンシーにシビアなユースケースでなければ、ノード障害がアプリケーションに与える影響は軽微と考えられます。

障害となったノードが復帰したタイミングで一度スループットが落ちますが、こちらも徐々に回復します。

ノード障害時のHyper-V仮想マシンの振る舞い

HCIとしての利用時を想定し、クラスター共有ボリューム上のHyper-V仮想マシンを実行しながらノード障害を発生させ、振る舞いを確認しました。

クラスター共有ボリューム（3方向ミラー）に対し、書き込み負荷をかけた状態でノード障害を発生させた結果を以下に示します。



【ベンチマーク条件】
・ワークロード（fioオプション）
モード：ランダム書き込み（--rw=randwrite）
スレッド数：4（--numjobs=4）
ブロックサイズ：1MB（--bs=1M）
メモリキャッシュを利用しない（--direct=1）

・ 測定条件
3方向ミラーのクラスター共有ボリューム上に仮想マシン（CentOS7）を構築。仮想マシンにSSHで接続し、上記の条件のベンチマークを仮想マシン内で実行。ベンチマーク開始 5分後に物理サーバーのうち1台の電源を落とし、疑似的に障害を発生させ、その直後に再起動し、ノードをクラスターに復帰させた。

結果は、仮想マシン上のI/Oも前項と同様の傾向となりました。ノード停止時の瞬間はI/Oが一瞬停止するものの、仮想マシンはシャットダウンなどされることなく動作を続けました。障害を発生されたクラスターノードが復旧した際、一時的に性能低下が起こるものの、一定時間経過後は障害前の性能に戻ることが確認できました。

このように、記憶域スペースダイレクトの可用性機能によって、障害時にもI/O機能が維持されることが確認できました。

システムメンテナンスに関する確認

記憶域スペースダイレクトによる分散ストレージやHCIを運用するにあたり、ソフトウェア・アップデートやデータ量の増加によるクラスターの拡張作業は織り込んで考えておくべきシナリオです。

ここでは、記憶域スペースダイレクトを構成するクラスターノードのアップデート方法、およびクラスターノードの追加方法について検証しました。

ノードのシャットダウン

システムを運用していく上で、再起動を伴うサーバーのメンテナンスが必ず発生します。たとえば、オペレーティングシステムを最新に保つためソフトウェア更新を適用したり、新しいファームウェアをハードウェアにロードしたりする必要があります。また、これら以外にも運用上の理由によりクラスターノードを一時的にシャットダウンする場合が考えられます。

クラスターノードのメンテナンスが必要となった場合、従来通りOSに対してシャットダウン操作すれば、該当ノードはクラスターから自動的に切り離された上でシャットダウンされます。また、該当ノード上で仮想マシンが動いている場合には、ほかの稼働中ノードに仮想マシンがライブマイグレーションされた上でノードがシャットダウンされます。

クラスター対応更新による自動ローリングアップデート

Windows Server 2016には、記憶域スペースダイレクトを構成するクラスターノードを無停止でメンテナンスができる機能として「クラスター対応更新」があります。

フェールオーバークラスターマネージャーから「クラスター対応更新」機能を使うと、自動的にクラスターのメンバーノードに対して順番にWindows Updateを適用する、いわゆるローリングアップデートが行われます。各クラスターノードの再起動が必要になったタイミングで、ノードがクラスターから自動的に切り離され、ノードの再起動後、クラスターに自動的に復帰します。この仕組みにより、システム管理者によるメンテナンス作業が省力化されます。

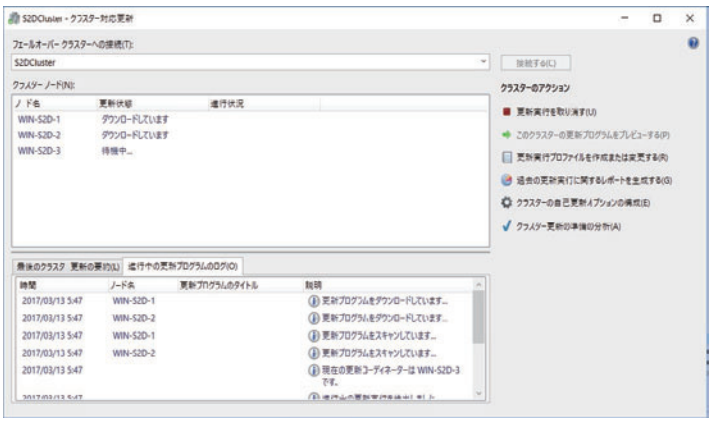


図 13　クラスター対応更新

今回の検証環境でも、クラスター対応更新を実施し、上に示す画面から実行状況を確認しながら、ローリングアップデートによりサービス無停止でアップデートできることを確認しました。

ノードの追加

ハイパーコンバージド・インフラストラクチャでは、コンピューティングリソースもしくはストレージが不足した際、ノードを追加していくことで性能を向上させられることが大きな特徴のひとつです。

記憶域スペースダイレクトは、最小2台から構成できることから、今回の検証環境内で2台のフェールオーバークラスターを構成し、残りの1台を後から追加できることを確認しました。

今回のように2台でクラスターを構築する場合、クラスターノードがお互いと通信できなくなる「スプリットブレイン」が発生した際にノードの生存状況を判断するため、外部にクォーラムを用意する必要があります。

Windows Server 2016では、外部クォーラムとしてAzureクラウド上のストレージが利用できます。Azure上のストレージをクォーラムに使う場合は、「クラスタークォーラムの設定の構成」ウィザードにて「クラスター監視を構成する」を選択し、Azureのストレージアカウントを設定します。

図 14　「クラスタークォーラムの設定の構成」ウィザードでのクラウド監視の設定

クラスターへノードを追加するには、新しくメンバーノードとして追加したいサーバーをActive Directoryに参加させ、既存のクラスターノードと同様の機能と役割をインストールしておきます。

クラスターにノードを追加するには、フェールオーバークラスターマネージャーの左側ペインで「ノード」を選択し、右側ペインの「ノードの追加...」をクリックし、ウィザード中で追加したいノードを指定します。ノードをクラスターに参加させると、ノードに搭載されている未使用ディスクが自動的に記憶域スペースダイレクトの記憶域プールに追加されます。

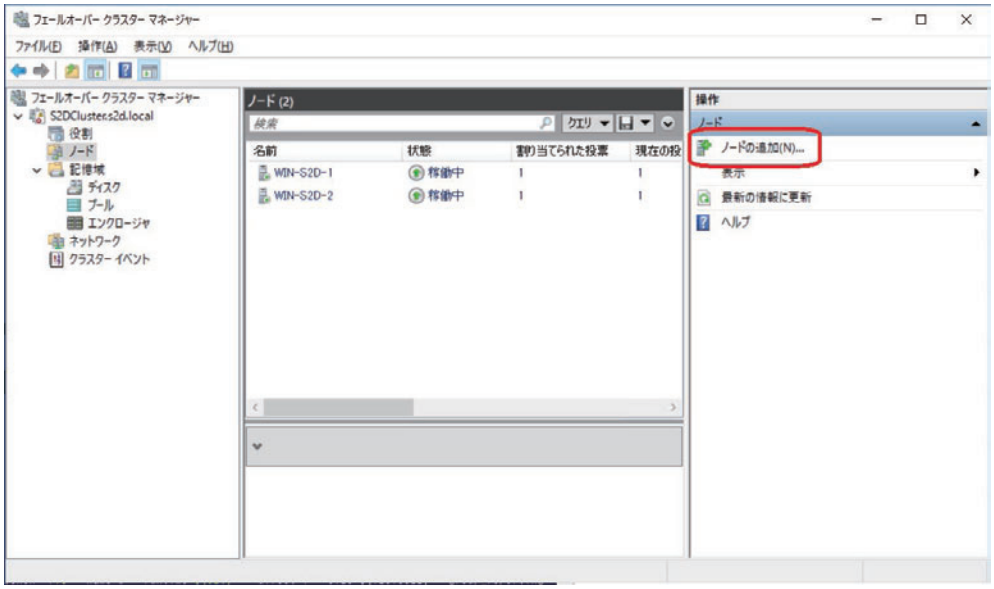


図 15　フェールオーバークラスターマネージャーからのノード追加

データのリバランス

クラスターに新たなノードを追加した時点では、新しいノードのローカルディスクにデータは保存されていません。このため、クラスター内でデータ分散のバランスが取れていない状態となっています。

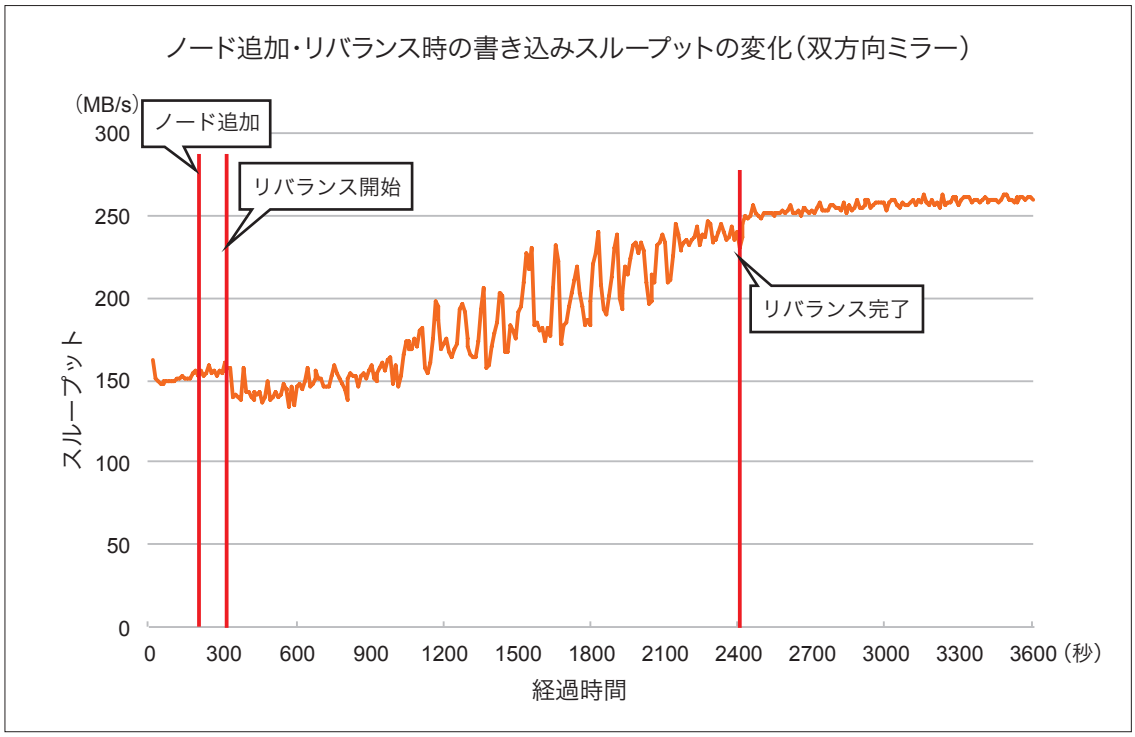
以下のコマンドを実行すると、新しいノードを含めたデータのリバランスが行われ、性能、および回復性オプションにあわせてデータの配置が最適化されます。

Optimize-StoragePool –FriendlyName <記憶域プール名>

最適化ジョブの実行状況は以下のコマンドで確認できます。

Get-StorageJob |? Name -eq Optimize

実際に本コマンドレットを実行した際の性能の変化は以下の通りとなりました。



【ベンチマーク条件】
・ワークロード（fioオプション）
モード：ランダム書き込み（--rw=randwrite）
スレッド数：4（--numjobs=4）
ブロックサイズ：1MB（--bs=1M）
メモリキャッシュを利用しない（--direct=1）

・測定条件
双方向ミラーのクラスター共有ボリューム上に上記の条件のベンチマークを実施。ベンチマーク開始　3分後にノード追加のオペレーションを実施、またベンチマーク開始5分後に、ノード間データのリバランスコマンドを発行（リバランスはコマンドレット発行から約35分で完了）

リバランス開始直後は、若干ですが一時的にスループットが低下しました。しかし、リバランス中からノード追加による性能向上がみられ、アプリケーションI/Oへの悪影響は非常に限定的でした。

リバランスが完了したのちは安定した性能が得られました。クラスターノードが2ノードから3ノードへ追加され、データが分散配置されることにより、ノード追加前と比較してスループットが向上しています。

本検証結果より、クラスターノード追加の手順が確認できました。また、ノード追加後にリバランスを実行することで、クラスターノード間でデータを再配置し、記憶域プールを最適化できることも確認できました。

今回の検証では、ノード追加およびデータのリバランスの最中もアプリケーションに対する影響は限定的でした。実際の運用では、I/O頻度が低い時間帯、もしくはメンテナンス時間などがあれば、それらのタイミングに作業をスケジュールするとよいでしょう。

まとめ

本検証の結果、Windows Server 2016およびLenovo x86サーバーを組み合わせることで、慣れ親しんだWindows Serverの操作で、信頼性が高く、かつ、柔軟な分散ストレージを構築することができました。さらに、その分散ストレージを使ったハイパーコンバージド・インフラストラクチャ（HCI）を構築できることが明確になりました。

メンテナンスやノード障害の際の縮退運転時も、ある程度の性能低下が認められたものの、サービスは無停止で継続されました。ノード復帰後には自動的に元の性能に戻り、高信頼性クラスターを運用する上で重要な点も問題なく動作することが確認できました。

また、自動ローリングアップデートを可能とする「クラスター対応更新」、コンピューティングリソースもしくはストレージ領域が不足した際のノード追加、追加したノードを含むリバランスの機能はどれも簡単に操作できました。Windows Server 2016 Datacenter Editionを導入することにより、なじみ深いWindowsの操作だけで、過不足なくHCIを構成、運用できます。

Windows Server 2016には、今回評価した機能以外にも、遠隔地のサーバーとデータを同期し、ディザスタリカバリを実現する記憶域レプリカをはじめ、エンタープライズのITインフラに求められる多くの機能を標準機能として搭載しています。

Windows Server 2016 DatacenterEditionとLenovo x86サーバーの組み合わせにより、トレンドとなっているHCI環境やソフトウェア定義ストレージ（SDS）を取り入れられます。プリインストール版、Reseller Option Kit（ROK）など、レノボはさまざまな形態で Windows Server オペレーティングシステムを提供しています。

ぜひ、Windows Server 2016とともにLenovoサーバーの導入をご検討ください。

本レポートはレノボ・エンタープライズ・ソリューションズ株式会社の依頼により、MKTインターナショナル株式会社、株式会社アクセライト、ハセゴーイノベーション、日本仮想化技術株式会社（監修）の四社が第三者検証協力として、記憶域スペースダイレクト（S2D）を使ったWindows Serverの標準機能のみを利用したハイパーコンバージド・インフラストラクチャ（HCI）の実現性および可能性評価を実施し、Lenovo製品の推奨構成を付与したものです。

【MKTインターナショナル株式会社について】

MKTインターナショナル株式会社は、「知識をオープンに共有し、価値を創造する使命を持って、お客様、社会から信頼される企業を目指す」ことをミッションに掲げ、2011年4月の設立以降、日本で随一の法人向けマーケティング活動を支援してきました。

会　社　名　：　MKTインターナショナル株式会社
所　在　地　：　神奈川県川崎市
設立年月日　：　2011年4月28日
代　表　者　：　代表取締役社長　赤井 誠
主な事業内容　：　事業企画及びマーケティングコンサルティング、マーケティング委託業務等
ホームページ　：　http://www.mkt-i.jp/

【株式会社アクセライトについて】

会　社　名　：　株式会社アクセライト
所　在　地　：　東京都文京区本郷
設立年月日　：　2010年11月
代　表　者　：　代表取締役社長　板垣 貴志
　　執行役員CTO　石田 精一郎
主な事業内容　：　調査研究コンサルティング、システム開発事業、医療データリポジトリ開発、学会事務局運営
ホームページ　：　https://accelight.co.jp/

【ハセゴーイノベーションについて】

屋　　号　：　ハセゴーイノベーション
所　在　地　：　神奈川県川崎市
代　表　者　：　長谷川 猛
主な事業内容　：　各種コンサルティングおよびコンピュータソフトウェア開発等

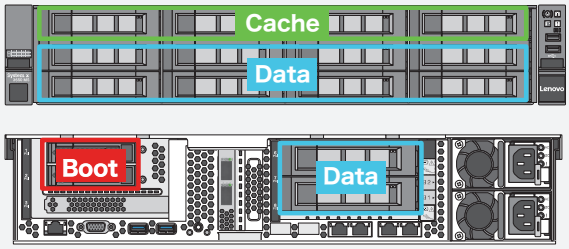
【日本仮想化技術株式会社について】

会　社　名　：　日本仮想化技術株式会社
所　在　地　：　東京都渋谷区
設　　立　：　2006年12月
取　締　役　：　代表取締役社長兼CEO　宮原 徹
　　取締役CTO　伊藤 宏通
ホームページ　：　http://virtualtech.jp/

S2D 検証構成・推奨構成

推奨構成1：普通の仮想サーバーはこれ！SSD構成

Lenovo Pressをベースとしたssd構成です。SSDのバランスやコスト等最適化されており3方向ミラーの構成で約50TBの実行容量を持たせることができます。一般的な仮想サーバーで使用するためのS2Dの構成に推奨いたします。



System x3650 M5 3.5型　4台構成

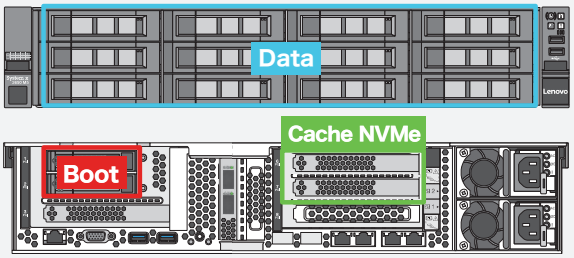
8871D4J	System x3650 M5(HS 3.5)/XeonE5-2630v4(10) 2.20GHz-2133MHz×1/PC4-19200 16.0GB(16×1) (Chipkill)/RAID-M5210/RAID-5200-1GF/POW(750W×1)/ OSなし/3年保証24x7(CRU)/SS90	4
00YJ198	XeonE5-2630v4(10) 2.20GHz-2133MHz	4
46W0829	16GB 2R PC4-19200 RDIMM CK	60
00YK005	4TB 7.2K 12Gbps NL SAS 3.5型 Gen2 HS HDD	40
00YC345	S3710 800GB SATA 3.5型 eMLC HS Enterprise Performance SSD	16
00WG690	600GB 10K 12Gbps SAS 2.5型 Gen3 HS HDD	8
00FK658	追加2HDDキット(2.5型/背面ベイ)	4
00FK659	追加2HDDキット(3.5型/背面ベイ)	4
46C9114	ServeRAID M1215 SAS/SATA コントローラー	4
01GR250	Mellanox ConnectX-4 Lx 2x25GbE SFP28 アダプター	4
47C8675	N2215 SAS/SATA HBA(PCI-E)	4
00FK934	高効率 750W(200-240V) リダンダント電源機構	4
46M2593	NEMA5-15P to IEC C13 電源ケーブル (2.8m)	8
01GU576	Windows Server 2016 Datacenter ROK レノボ版 (16コア)	4
01GU634	Windows Server 2016 Datacenter 追加ライセンス (4コア)	4
01GU642	Windows Server CAL 2016 (10ユーザー)	4

※SFP28のトランシーバー・ケーブルは含めておりません。

推奨構成2：高速なIOが必要なときは！NVMe構成

本検証をベースとし、NVMeのPCI Expressアダプターを搭載した構成です。3方向ミラーの構成で約65TBの実行容量を持たせることができます。

より高速なI/Oが必要となる場合に、こちらの構成を推奨いたします。



System x3650 M5 3.5型　4台構成

8871D4J	System x3650 M5(HS 3.5)/XeonE5-2630v4(10) 2.20GHz-2133MHz×1/PC4-19200 16.0GB(16×1)(Chipkill)/ RAID-M5210/RAID-5200-1GF/POW(750W×1)/OSなし/ 3年保証24x7(CRU)/SS90	4
00YJ198	XeonE5-2630v4(10) 2.20GHz-2133MHz	4
46W0829	16GB 2R PC4-19200 RDIMM CK	60
00YK005	4TB 7.2K 12Gbps NL SAS 3.5型 Gen2 HS HDD	48
00WG690	600GB 10K 12Gbps SAS 2.5型 Gen3 HS HDD	8
00FK658	追加2HDDキット(2.5型/背面ベイ)	4
46C9114	ServeRAID M1215 SAS/SATA コントローラー	4
01GR250	Mellanox ConnectX-4 Lx 2x25GbE SFP28 アダプター	4
47C8675	N2215 SAS/SATA HBA(PCI-E)	4
00YA812	P3700 1.6TB NVMe Enterprise Performance Flash アダプター	8
00FK934	高効率 750W(200-240V) リダンダント電源機構	4
46M2593	NEMA5-15P to IEC C13 電源ケーブル (2.8m)	8
01GU576	Windows Server 2016 Datacenter ROK レノボ版 (16コア)	4
01GU634	Windows Server 2016 Datacenter 追加ライセンス (4コア)	4
01GU642	Windows Server CAL 2016 (10ユーザー)	4

※SFP28のトランシーバー・ケーブルは含めておりません。

あらゆる環境に、信頼性が打ち勝つ。

レノボの飽くなきチャレンジとイノベーションが、
時代のスタンダードに。

効率性の向上やIT投資を最大限に活かすことができます。
マイクロソフトのテクノロジーは、適正価格で大企業のビジネスツールをお客様に提供します。

サーバー・ストレージ関連の最新情報はこちら

http://www.lenovojp.com/server/

エンタープライズ・ソリューションに驚異的なパフォーマンスと高い汎用性を、
最高レベルの信頼性とセキュリティと共にご提供するラック型2Uサーバー

System x3650 M5

System x3650 M5は、最大2個のインテル® Xeon® プロセッサー E5-2600 v4製品ファミリーと、2ソケット・サーバーでは屈指の大容量ストレージ(最大116TB)のサポート、およびストレージの選択肢により、仮想デスクトップ環境はもとより、クラウドからビッグデータまで、あらゆるワークロードを最適化し、System x Trusted Platform Assuranceによる最高レベルのセキュリティーと信頼性と共に、ビジネスを加速させます。

お電話やメールでのお問い合わせはこちら！

法人のお客様向け見積依頼
・ご購入相談窓口

0120-68-6200

hojin_jp@lenovo.com

受付時間：月曜日から金曜日9時から17時30分
（祝日および年末年始休業日を除く）

レノボ・ジャパン株式会社
〒101-0021
東京都千代田区外神田四丁目14番1号　秋葉原UDX

http://www.lenovojp.com/business/

ITの効率性と生産性の向上

Windows Server