ホワイト ペーパー

# Cisco HyperFlex System および Cisco UCS M5 サーバと VMware vSphere 6.5 および Citrix XenDesktop 7.15 上の NVIDIA GRID 5.0 との統合

2018 年 2 月

# 目次

日本語は部分翻訳のみ、全文は英語版を参照下さい。

https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/whitepaper-c11-740243.pdf

## このドキュメントの内容

シスコは 2017 年に、Intel® Xeon® スケーラブル プロセッサ アーキテクチャに基づいた第 5 世代の Cisco UCS® B シリーズ ブレード サーバ、C シリーズ ラック サーバ、Cisco HyperFlex™ ハイパーコンバージド サーバを発表しました。また、ほぼ同時期に、新しいサーバ アーキテクチャの使用を念頭に設計された新しいハードウェアとソフトウェアが、NVIDIA 社より発表されました。

Cisco UCS B シリーズ ブレード サーバ、C シリーズ ラック サーバ、Cisco HyperFlex ハイパーコンバージド サーバにおける処理能力の増強により、高いグラフィック要件が求められるアプリケーションを仮想化できるようになりました。仮想デスクトップ インフラストラクチャ（VDI）における高パフォーマンスを実現し、グラフィック利用頻度の高いアプリケーションの配信能力を向上させるために、シスコでは、B シリーズ/C シリーズの PCI Express（PCIe）カードおよびメザニン フォームファクタ カードの Cisco Unified Computing System™（Cisco UCS）ポートフォリオにおいて、NVIDIA GRID P6、P40、P100、M10 カードのサポートを開始しました。

新しいグラフィック処理機能を追加したことにより、組織のエンジニアリング部門、設計部門、イメージング部門、マーケティング部門でアプリケーションを利用する際に、デスクトップ仮想化によるメリットを感じられるようになります。また、Microsoft Windows 10 や Office 2016 以降のグラフィックスが強化されたバージョンで、Cisco UCS C240 M5 ラック サーバや Cisco HyperFlex ハイパーコンバージド サーバに装着可能な NVIDIA M10 高密度グラフィック カードを利用できます。

この新しいグラフィック機能により、組織のグラフィック ワークロードやデータセンター内のデータを一元化できます。また、業務を地理的にシフトする必要がある組織にとっても、大きなメリットがあります。これまでは、画像ファイルはサイズが大きすぎて移動ができず、ファイルをローカルに保存して使用しなければなりませんでした。

Cisco UCS サーバの PCIe グラフィック カードには、次のような利点があります。

- 2 ラック ユニット（2RU）または 4RU フォームファクタで、NVIDIA GRID カードをフルサイズ（フルレングス）、フルパワーでサポート
- ハーフ/フル幅のブレード サーバで、メザニン フォームファクタ アダプタ グラフィック プロセシング ユニット（GPU）カードをサポート
- サーバおよび NVIDIA GRID カードの管理を Cisco UCS Manager に統合
- Cisco UCS Central Software や Cisco UCS Director など、Cisco UCS 管理ソリューションによるエンドツーエンド統合
- 2 つの NVIDIA グリッド カードを使用した Cisco UCS ブレード サーバ/ラック サーバによる効率的なラック スペースの活用。2 スロット、2.5 インチのラック ユニット ブレード タイプのデザインよりも効率的

Cisco UCS B シリーズ サーバのモジュール型 LAN on Motherboard（mLOM）フォームファクタ NVIDIA グラフィック カードには、次のような利点があります。

- サーバおよび NVIDIA GRID GPU カード管理用の Cisco UCS Manager 統合
- Cisco UCS Central Software や Cisco UCS Director など、Cisco UCS 管理ソリューションによるエンドツーエンド統合

このドキュメントの重要な要素は、VMware vSphere 6.5 の NVIDIA GRID 仮想グラフィック処理ユニット（vGPU）機能に関する VMware のサポートです。リリース 6.0 より前の vSphere では Virtual Direct Graphics Acceleration（vDGA）と Virtual Shared Graphics Acceleration（vSGA）のみがサポートされており、vSphere リリース 6.0 以降で vGPU がサポートされることで導入シナリオが大きく広がり、GRID カードの柔軟で効率的な構成が実現されます。

このドキュメントの目的は、パートナーおよびユーザでの、NVIDIA GRID 5.0 グラフィック処理カード、Cisco HyperFlex System、Cisco UCS B200 M5 ブレード サーバ、Cisco UCS C240 M5 ラック サーバ、VMware vSphere、Citrix XenDesktop 7.15 の vGPU モードでの統合利用を支援することです。

各モードでのカード、ハイパーバイザ、デスクトップ ブローカでサポートされているアプリケーションの一覧については、NVIDIA、Citrix、VMware の各パートナーにお問い合わせください。

ここでの目的は、サーバ、ハイパーバイザ、仮想デスクトップにグラフィック アプリケーションをインストールする準備ができるように、NVIDIA GRID P6、P40、M10 カードを搭載した Cisco UCS サーバと VMware vSphere、Citrix 製品との統合方法について、それぞれ説明することにあります。

## Citrix XenDesktop でのグラフィック展開に NVIDIA GRID vGPU を使用する理由

NVIDIA GRID vGPU により、複数の仮想デスクトップで単一の物理 GPU を共有し、また、論理的に複数 GPU を単一の物理 PCI カード上に定義することができます。それらがすべて vDGA パススルー グラフィックに関して 100 % のアプリケーション互換性を提供するとともに、複数のデスクトップが単一のグラフィック カードを共有するため、コストが抑えられます。Citrix XenDesktop の場合は一元化やプールが可能で、従来の複雑で高価なワークステーション/デスクトップを複数台利用することと比べて、より容易に管理できます。すべてのユーザ グループが、このような仮想化の利点を活用できるようになります。

GRID vGPU の機能により、仮想化ソリューションにおいて、NVIDIA ハードウェア アクセラレータ グラフィックの利点をフルに活用できます。この技術は、オンボードのグラフィック プロセッサを搭載した PC と同等の優れたグラフィック パフォーマンスを仮想デスクトップにも提供するものです。

GRID vGPU では、複数の仮想デスクトップ間で GPU ハードウェア アクセラレーションを実際に共有する上での、業界で最も先進的な技術を使用しており、グラフィック エクスペリンスを損なうことがありません。アプリケーションの機能と互換性は、ユーザがデスクトップを使用する場合とまったく同じです。

GRID vGPU テクノロジーでは、各仮想マシンのグラフィック コマンドがハイパーバイザによって変換されることなく、GPU に直接渡されます。複数の仮想マシンが仮想化サーバの単一 GPU の性能を直接使用することができるため、企業は仮想マシンの実際の GPU をベースとしたグラフィック アクセラレーションを使用できるユーザの数を増やすことができます。

サーバ上の物理的 GPU は、個別の vGPU プロファイルで構成できます。組織は、さまざまな種類のエンド ユーザのニーズに合わせて、最適なサーバ構成をとる上での大きな柔軟性を得ることができます。

vGPU のサポートにより、企業が NVIDIA GRID テクノロジーの力を使用してまったく新しいクラスの仮想マシンを作成し、エンド ユーザにより強力で変更ニーズに合わせたグラフィック エクスペリエンスを提供できるようになります。

### vGPU プロファイル

企業ごとに、個々のユーザのニーズは大きく異なります。GRID vGPU の主な利点の 1 つは、さまざまなクラスのエンド ユーザのニーズに応えるために設計された複数の vGPU プロファイルを柔軟に利用できるということです。

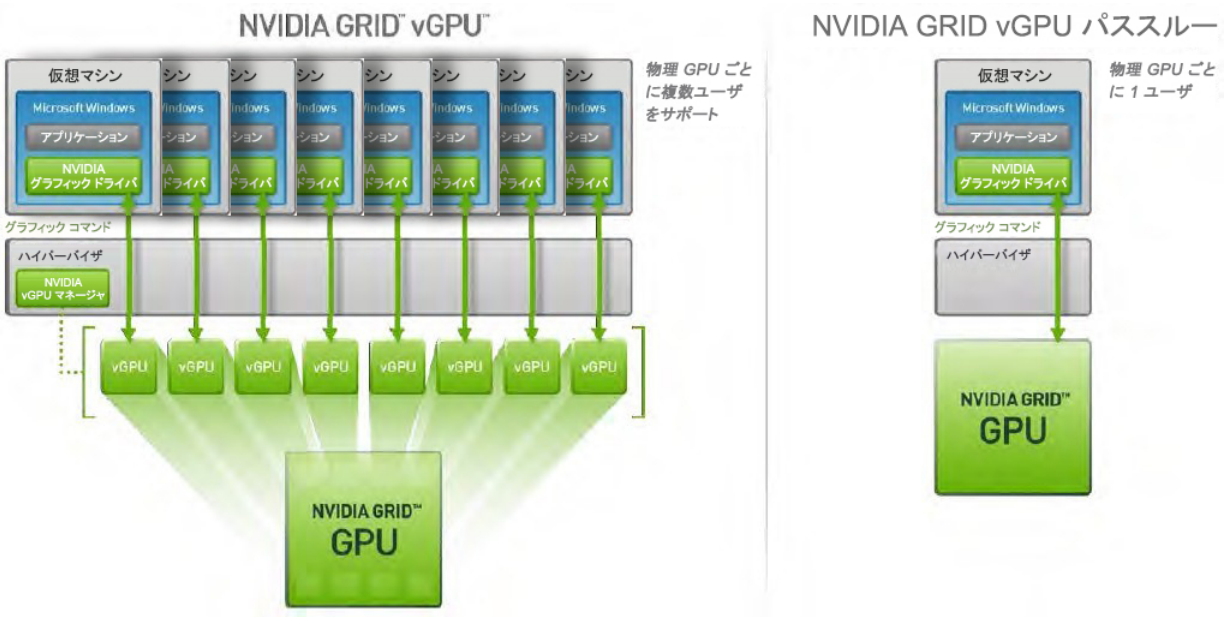
エンド ユーザのニーズは多様ですが、シンプルにするために、ユーザをナレッジ ワーカー、デザイナー、パワー ユーザのカテゴリにグループ化して捉えることができます。

- ナレッジ ワーカーの場合、最も重要な点は、効率的なオフィス アプリケーション、安定した Web エクスペリエンス、滞りないビデオ再生などです。ナレッジ ワーカーの場合、グラフィックに対する要求は最低限の負荷で済みますが、デスクトップ PC、ノート パソコン、タブレット、スマートフォンなど、ネイティブでグラフィック アクセラレーションが実行される最近のデバイスで普通に実現されている支障のない滑らかなエクスペリエンスを期待しています。
- パワー ユーザは、オフィス事務用ソフトウェア、Adobe Photoshop などの画像編集ソフトウェア、Autodesk AutoCAD などの主流のコンピュータ支援設計（CAD）ソフトウェア、製品ライフ サイクル管理（PLM）アプリケーションなどの、より要件の高いオフィス アプリケーションを使用する必要があるユーザです。これらのアプリケーションは、OpenGL や Direct3D などの API を完全にサポートした上で追加的なグラフィック リソースを必要とし、より要件が高くなります。

- デザイナーは、ハイエンドの 3 次元 CAD ソフトウェアや専門のデジタル コンテンツ作成（DCC）ツールなどの要件の高い専門アプリケーションを使用する組織内のユーザです。例としては、Autodesk Inventor、PTC Creo、Autodesk Revit、Adobe Premiere が挙げられます。従来から、デザイナーはデスクトップ ワークステーションを使用しており、ハイエンドなグラフィックと専門的な CAD や DCC ソフトウェアの認定要件が必要であることから、仮想環境に組み込むのは難しいグループとされてきました。

vGPU プロファイルにより、GPU ハードウェアがタイムスライスで割り当てられ、非常に優れた共有仮想グラフィック性能を実現します（図 1）。

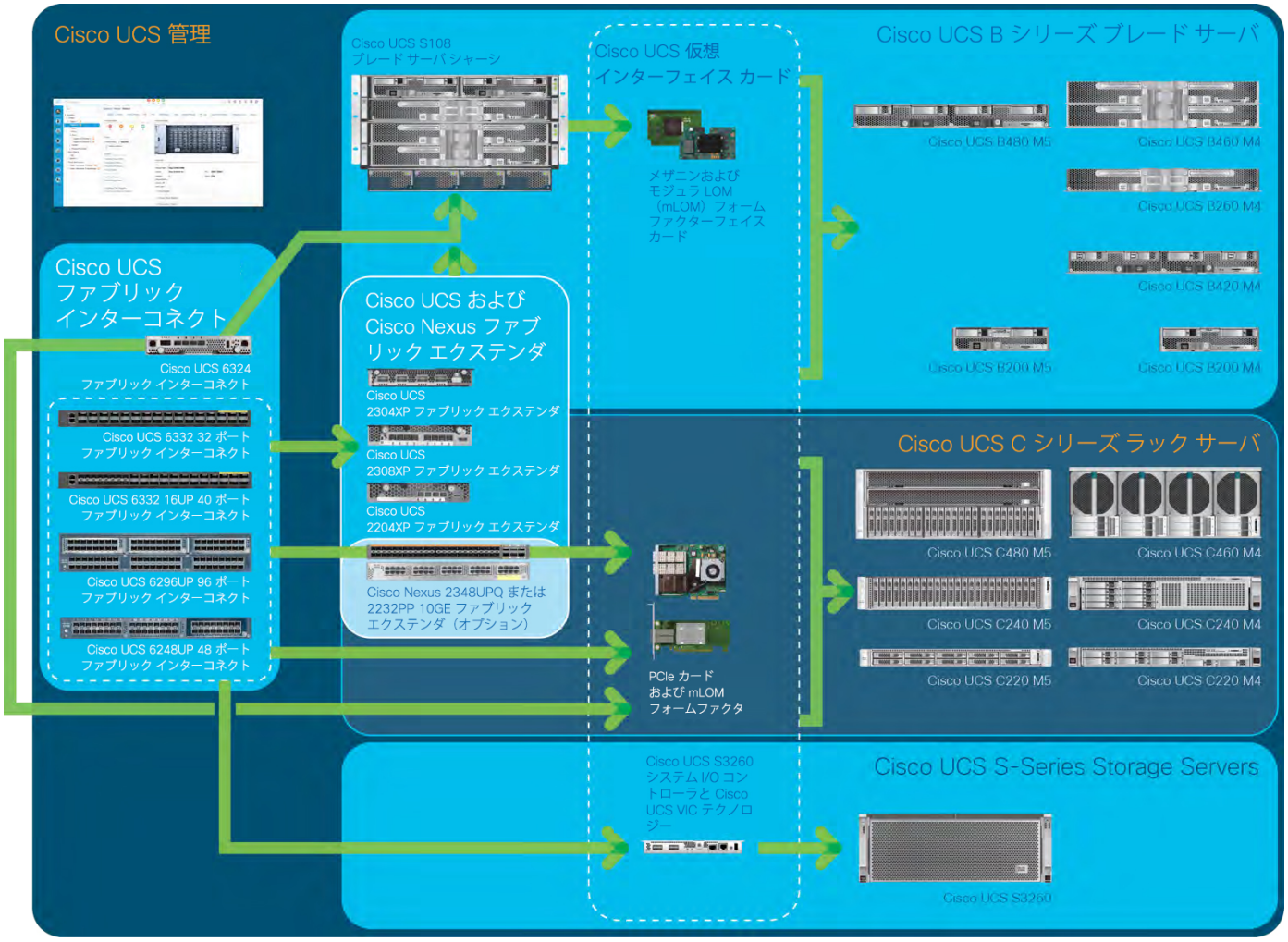
**図 1.** NVIDIA GRID vGPU システム アーキテクチャ




## Cisco Unified Computing System

Cisco UCS は、コンピューティング、ネットワーキング、およびストレージ アクセスを統合した次世代のデータセンター プラットフォームです。仮想環境向けに最適化されており、業界標準のオープン テクノロジーを利用して設計されたもので、総所有コスト（TCO）の削減とビジネスの俊敏性強化を目的としています。この UCS は、低遅延のロスレス 40 ギガビット イーサネットで統合されたネットワーク ファブリックと、エンタープライズクラスの x86 アーキテクチャ サーバをシステム化します。このシステムは、すべてのリソースが単一の管理ドメインに集約された、統合型の拡張性に優れたマルチシャーシ プラットフォームです（図 2）。

**図 2.** Cisco UCS のコンポーネント




Cisco UCS の主なコンポーネントは、次のとおりです。

- **コンピューティング：**Intel プロセッサをベースとしたブレード サーバとラック サーバのどちらもシステムに組み込んだ、まったく新しいクラスのコンピューティング システムをベースとする設計です。
- **ネットワーク：**このシステムは低遅延でロスレスの 40 Gbps ユニファイド ネットワーク ファブリック上で統合されています。このネットワーク基盤は、現状では LAN、SAN、高性能コンピューティング（HPC）ネットワークとして扱われている異なるネットワークを統合します。ユニファイド ファブリックにより、ネットワーク アダプタ、スイッチ、およびケーブルの数が減少し、電力と冷却の要件が緩和されるため、コスト削減につながります。
- **仮想化：**仮想環境のスケーラビリティ、パフォーマンス、および運用制御を強化することで、仮想化の可能性を最大限に活用します。シスコのセキュリティ、ポリシー適用、および診断機能は仮想化環境にまで拡張されており、変化するビジネス要件や IT 要件により適切に対応できます。
- **ストレージ アクセス：**ユニファイド ファブリックによってローカル ストレージ、SAN ストレージとネットワーク接続ストレージ（NAS）へのアクセスを統合します。ストレージ アクセスを統合することにより、Cisco UCS ではイーサネット、ファイバ チャネル、Fibre Channel over Ethernet（FCoE）、Small Computer System Interface over IP（iSCSI）プロトコルを使用して、ストレージにアクセスできます。この機能により、顧客はストレージ アクセスと投資保護を選択できるようになり、

サーバ管理者はシステムからストレージ リソースへの接続に関するストレージ アクセス ポリシーを事前に割り当てられるため、ストレージの接続と管理が単純になり、生産性も向上します。

- **管理：**Cisco UCS の特色は、あらゆるシステム コンポーネントを統合し、ソリューション全体を 1 つのエンティティとして、Cisco UCS Manager ソフトウェアから管理できることです。すべてのシステムの構成プロセスや動作を管理できる直感的な GUI、コマンドライン インターフェイス（CLI）、堅牢な API を備えています。

Cisco UCS では、次のことを実現できます。

- TCO を削減し、ビジネスの敏捷性を高める
- ジャストインタイムのプロビジョニングとモビリティ サポートにより、IT スタッフの生産性を向上する
- 単一の統合されたシステムにより、データセンターのテクノロジーを統合し、全体として管理、サービス提供できる
- 設計は数百台の物理サーバと数千台の仮想マシンに対応しており、要件に合わせて I/O 帯域幅を拡張できるため、拡張性を確保できる
- 業界のリーダーとのパートナー エコシステムを通じて、業界標準をサポートする

### Cisco UCS Manager

Cisco UCS Manager は、直感的な GUI、CLI、XML API によって Cisco UCS のすべてのソフトウェアおよびハードウェア コンポーネントを統合管理できる機能を備えています。一元管理機能を備えた単一の管理ドメインがあり、複数のシャーシや数千台もの仮想マシンを制御できます。Cisco UCS マネージャと NVIDIA GPU カードを緊密に統合することで、ファームウェアとグラフィック カード構成の管理を改善します。

### Cisco UCS 6332 ファブリック インターコネクト

Cisco UCS 6332 ファブリック インターコネクト（図 3）は、Cisco UCS B シリーズ ブレード サーバ、C シリーズ ラック サーバ、5100 シリーズ ブレード サーバ シャーシの管理と通信の中心となります。6332 ファブリック インターコネクトに接続されているすべてのサーバは、可用性の高い、統合された管理ドメインの一部として管理されます。

Cisco UCS 6300 シリーズ ファブリック インターコネクトは、ユニファイド ファブリックをサポートしているため、ドメイン内のすべてのサーバに対して LAN および SAN 接続を提供します。詳細については、次の URL を参照してください。
https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/6332-specsheet.pdf

6332 ファブリック インターコネクトでは、次の機能が提供されます。

- 最大帯域幅 2.56 Tbps の全二重スループット
- 1 RU フォームファクタでの 40 Gbps Quad Small Form-Factor Pluggable（QSFP+）ポート X 32
- 10 Gbps ブレークアウト ケーブル X 4 のサポート
- ラインレート、低遅延、ロスレスの 40 ギガビット イーサネットおよび FCoE の利用可能なポート
- Cisco UCS Manager による中央管理
- 効率的な冷却とメンテナンス性

図 3.　　Cisco UCS 6332 ファブリック インターコネクト


## Cisco UCS C シリーズ ラック サーバ

Cisco UCS C シリーズ ラック サーバは、Intel Xeon プロセッサの技術革新に合わせて、プロセッサ コア、周波数や、セキュリティと可用性の機能を向上させた最新のプロセッサを搭載しています。Intel Xeon スケーラブル プロセッサによって提供されるパフォーマンスの向上により、Cisco UCS C シリーズ サーバにおける価格対性能比を改善します。また、Cisco UCS の革新技術（標準ベースのユニファイド ネットワーク ファブリック、Cisco® VN-Link 仮想化サポート、Cisco Extended Memory Technology など）が業界標準のラックマウント フォーム ファクタにまで拡張されます。

これらのサーバはスタンドアロンの環境や Cisco UCS Manager 管理下でブレードサーバと合わせてコンフィギュレーション可能なサーバとしても動作するように設計されているため、サーバを必要なだけ使用して、組織のタイミングと予算に合わせた最適なスケジュールでシステムを徐々に拡張できます。Cisco UCS C シリーズ サーバは、スタンドアロン サーバとしても、Cisco UCS の一部としても導入可能なため、投資が保護されます。

多くの組織がラック マウント サーバを利用するのは、さまざまな内蔵ドライブと PCIe アダプタで幅広い I/O オプションを利用できるからです。Cisco UCS C シリーズ サーバは、シスコ製品はもちろんのこと、サードパーティ製のアダプタでサポートされているインターフェイスも含めて、多種多様な I/O オプションに対応しています。

## Cisco UCS C240 M5 ラック サーバ

Cisco UCS C240 M5 ラック サーバ（図 4、図 5、表 1）は、ビッグ データからコラボレーションに至るまで、ストレージを必要とするさまざまなインフラストラクチャの作業負荷に耐えうるパフォーマンスと拡張性を実現するように設計されています。

C240 M5 小型フォーム ファクタ（SFF）サーバは、Cisco UCS ポートフォリオを 2RU フォーム ファクタに拡張させた製品です。Intel Xeon スケーラブル プロセッサ ファミリ、2,666 MHz DIMM 対応の DIMM スロット（最大 24 枚/128 GB）、最大 6 本の PCIe 3.0 スロット、最大 26 台の内部 SFF ドライブが構成可能です。また、12 GB SAS ストレージ コントローラ カード用の専用内蔵スロット 1 個も装備しています。本製品では、10GBASE-T のマザーボード組み込み Intel x550 LOM ポート 2 個に加え、PCI スロットを使用することなくシスコ仮想インターフェイス カード（VIC）またはサードパーティ製ネットワーク インターフェイス カード（NIC）を設置できる専用の mLOM 内蔵スロットも搭載しています。

さらに、C240 M5 は、優れたパフォーマンスと、内部メモリおよびストレージの卓越した拡張性を提供します。次の機能を備えています。

- パフォーマンスの向上と電力消費の削減を実現する最大 24 台、最大 2,666 MHz の DDR4 DIMM
- Intel Xeon スケーラブル CPU X 1、または X 2
- PCIe 3.0 スロット最大 X 6（GPU 用フルハイト、フルレングス X 4）
- ホットスワップ可能なファン（前面から背面への冷却用エアーフロー）X 6
- SFF 前面 SAS/SATA ハードディスク ドライブ（HDD）または SAS/SATA ソリッド ステート ドライブ（SSD）X 24
- オプションで、SAS/SATA ドライブの代わりに最大 2 台の前面 SFF NVMe PCIe SSD ドライブを装着可能。この NVMe ドライブは前面ドライブ ベイ 1 および 2 の前面にのみ装着し、ライザ 2 オプション C から制御する必要があります。
- オプションで最大 2 台の SFF、背面 SAS/SATA HDD/SSD、または最大 2 台の背面 SFF NVMe PCIe SSD ドライブを装着可能。
  - 背面 SFF NVMe ドライブは、ライザ 2、オプション B または C から接続
  - 12 Gbps SAS ドライブをサポート
- マザーボードの専用 mLOM スロットに、次のカードを柔軟に装着可能：
  - Cisco VIC 10/40Gbps FCoE 2 ポート
  - Intel i350 1 Gbps イーサネット RJ-45 4 ポート mLOM NIC
- 1/10 G イーサネット組み込み LOM ポート X 2
- 最大 2 つのダブルワイド NVIDIA GPU カードをサポート。より多くの仮想ユーザにグラフィックを活用した体験を提供可能
- ツール不要な CPU 挿入、使いやすいラッチ構造、ホットスワップおよびホットプラグ可能なコンポーネントで実現する優れた信頼性、可用性、有用性（RAS）の機能
- PCIe ライザ 1 にマイクロ SD カード用スロット X 1（オプション 1 および 1B）。
  - マイクロ SD カードは、Cisco Host Upgrade Utility（HUU）などのユーティリティ用ローカル リソース専用に提供されます。
  - 画像をファイル共有（ネットワーク ファイル システム（NFS）または Common Internet File System（CIFS））から取得し、後で使用するためにカードにアップロードできます。
- マザーボードのミニストレージ モジュール コネクタで、次のいずれかをサポート：
  - セキュア デジタル（SD）カード スロットを 2 つ備えた SD カード モジュール（容量の異なる SD カードは混在不可）
  - 2 本の SATA M.2 SSD スロットを備えた M.2 モジュール（容量の異なる M.2 カードは混在不可）

**注：**SD カードと M.2 モジュールを混在させることはできません。M.2 モジュールは、VMware では RAID 1 構成をサポートしていません。Microsoft Windows と Linux だけがサポートされます。

また、C240 M5 は、大容量ストレージを必要とする次のようなさまざまな用途のパフォーマンスを向上させ、お客様に幅広い選択肢を提供します。

- コラボレーション
- 中堅・中小企業（SMB）向けデータベース
- ビッグ データ インフラストラクチャ
- 仮想化と統合インフラ
- ストレージ サーバ
- 高性能アプライアンス

C240 M5 はスタンドアロン サーバ、または Cisco UCS 管理対象ドメインの一部として展開できます。Cisco UCS がコンピューティング、ネットワーキング、管理、仮想化、ストレージ アクセスを 1 つの統合型アーキテクチャにまとめるので、ベアメタル環境と仮想環境の両方においてエンドツーエンド サーバの可視化、管理、制御が可能になります。Cisco UCS 環境に実装した場合、C240 M5 は各種の標準規格に準拠したシスコ ユニファイド コンピューティングの革新的な機能を活用して、お客様の TCO（総所有コスト）の削減とビジネスの俊敏性の向上を実現します。

Cisco UCS C240 M5 ラック サーバの詳細については、次を参照してください。
https://www.cisco.com/c/dam/global/ja_jp/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c240m5-sff-specsheet.pdf

**図 4.** Cisco UCS C240 M5 ラック サーバ

**図 5.** Cisco UCS C240 M5 ラック サーバ背面図

**表 1.** Cisco UCS C240 M5 PCIe スロット

| PCIe スロット | 長さ | レーン |
|---|---|---|
| 1 | ハーフ | x8 |
| 2 | フル | x16 |
| 3 | ハーフ | x8 |
| 4 | ハーフ | x8 |
| 5 | フル | x16 |
| 6 | フル | x8 |

### Cisco UCS 仮想インターフェイス カード 1387

Cisco UCS VIC 1387（図 6）は、Cisco UCS C シリーズ ラック サーバに取り付けられたデュアル ポート拡張 Small Form-Factor Pluggable（SFP+）40 Gbps イーサネットおよび FCoE 対応 PCIe mLOM アダプタです。mLOM スロットを使用すれば PCIe スロットを使用せずに Cisco VIC を構成できるため、I/O の拡張性が向上します。シスコの次世代統合型ネットワーク アダプタ（CNA）テクノロジーを取り入れることで、将来の機能リリースにおける投資を保護します。このカードにより、ポリシーベースでステートレス、かつ敏しょう性に優れたサーバ インフラストラクチャが実現します。PCIe 標準準拠のインターフェイスを最大 256 個までホストに提供可能で、NIC またはホスト バス アダプタ（HBA）として動的に構成することができます。カードの特性は、ブート時に

サーバに関連付けられたサービス プロファイルを使用して動的に設定されます。PCIe インターフェイスの番号、タイプ（NIC または HBA）、ID（MAC アドレスおよび World Wide Name（WWN））、フェールオーバー ポリシー、帯域幅、Quality of Service（QoS）ポリシーは、すべてサービス プロファイルから決定されます。

VIC の詳細については、次を参照してください。https://www.cisco.com/c/ja_jp/products/collateral/interfaces-modules/unified-computing-system-adapters/datasheet-c78-736683.htm

**図 6.** Cisco UCS VIC 1387 CNA


### Cisco UCS B200 M5 ブレード サーバ

Cisco UCS B200 M5 ブレード サーバ（図 7）は、IT および Web インフラストラクチャや分散データベースなどのさまざまなワークロードで妥協のない優れた性能、汎用性、およびサーバ密度を実現します。エンタープライズクラスの B200 M5 ブレード サーバは、ハーフ幅のブレード フォーム ファクタでありながら、Cisco UCS ポートフォリオの機能を強化します。B200 M5 は、最新の Intel Xeon スケーラブル CPU を搭載し、最大 3072 GB の RAM（128 GB DIMM 使用）、2 台の SSD または HDD、最大スループット 80 Gbps での接続をサポートしています。

Cisco UCS B200 M5 サーバは、Cisco UCS 5100 シリーズ ブレード サーバ シャーシまたは UCS Mini ブレード サーバ シャーシに搭載できます。エラー訂正コード（ECC）の登録された DIMM（RDIMM）、または Load-Reduced DIMM（LR DIMM）用のスロットが合計 24 あり、イーサネットおよび FCoE を提供する Cisco UCS VIC 1340 アダプタ用のコネクタを 1 つサポートしています。

B200 M5 には、リア メザニン アダプタ スロットが 1 つあり、接続帯域幅追加用の Cisco UCS ポート エキスパンダ カード、または NVIDIA P6 GPU を構成できます。これらは、VIC 1340 に 4 ポート追加できるようになるハードウェア オプションです。VIC 1340 の全機能をネイティブのデュアル 40 Gbps インターフェイス、または、デュアル 4 X 10 ギガビット イーサネット ポート チャネル インターフェイスにすることができますまた、同じリア メザニン アダプタ スロットを NVIDIA P6 GPU で構成することもできます。

B200 M5 には、フロント メザニン スロットが 1 つあります。　フロント メザニン カードは、ストレージ コントローラまたは NVIDIA P6 GPU に対応可能です。

詳細については、次の URL を参照してください。https://www.cisco.com/c/dam/global/ja_jp/products/collateral/servers-unified-computing/ucs-b-series-blade-servers/b200m5-specsheet.pdf

**図 7.** Cisco UCS B200 M5 ブレード サーバ全面図

### Cisco UCS 仮想インターフェイス カード 1340

Cisco UCS 仮想インターフェイス カード（VIC）1340（図 8）は、Cisco UCS B シリーズ ブレード サーバの M5 世代に特化して設計された、2 ポート 40 Gbps イーサネットまたはデュアル 10 Gbps イーサネット FCoE X 4 対応の mLOM カードです。VIC 1340 を組み合わせることにより、ポリシーベースでステートレス、かつ敏捷性に優れたサーバ インフラストラクチャが実現します。このカード 1 枚で PCIe 標準準拠のインターフェイス合計 256 個、NIC または HBA をホストサーバに独立したリソースとして動的に構成することができます。さらに、VIC 1340 は Cisco UCS ファブリック インターコネクト ポートを仮想マシンに拡張し、サーバ仮想化の展開および管理を簡素化する Cisco Virtual Machine Fabric Extender（VM-FEX）をサポートします。

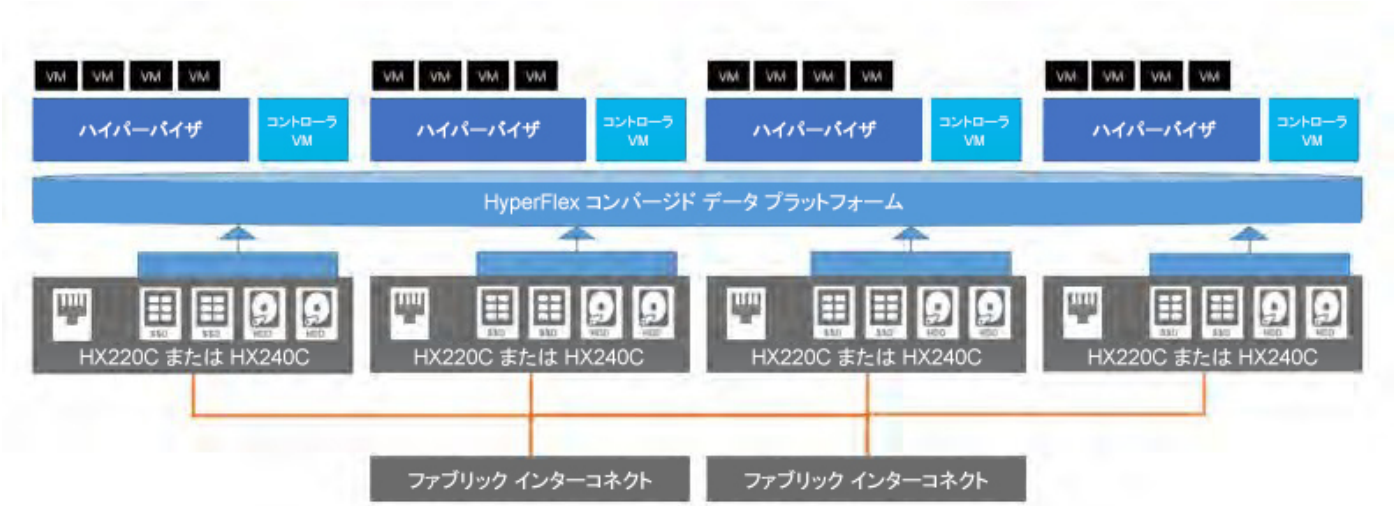
詳細については、次の URL を参照してください。https://www.cisco.com/c/ja_jp/products/interfaces-modules/ucs-virtual-interface-card-1340/index.html

**図 8.** Cisco UCS VIC 1340

## Cisco HyperFlex System

Cisco HyperFlex System は、ハイパーコンバージド インフラとして仮想サーバ ストレージ プラットフォームを提供します。コンピューティング、メモリ リソース、統合されたネットワーク接続、仮想マシン ストレージ用の分散型の高パフォーマンス ログ構造化ファイル システム、仮想サーバ用ハイパーバイザ ソフトウェアがすべて、単一の Cisco UCS 管理ドメインに含まれたシステムとして提供・構築されます（図 9）。

**図 9.　　Cisco HyperFlex System の概要**

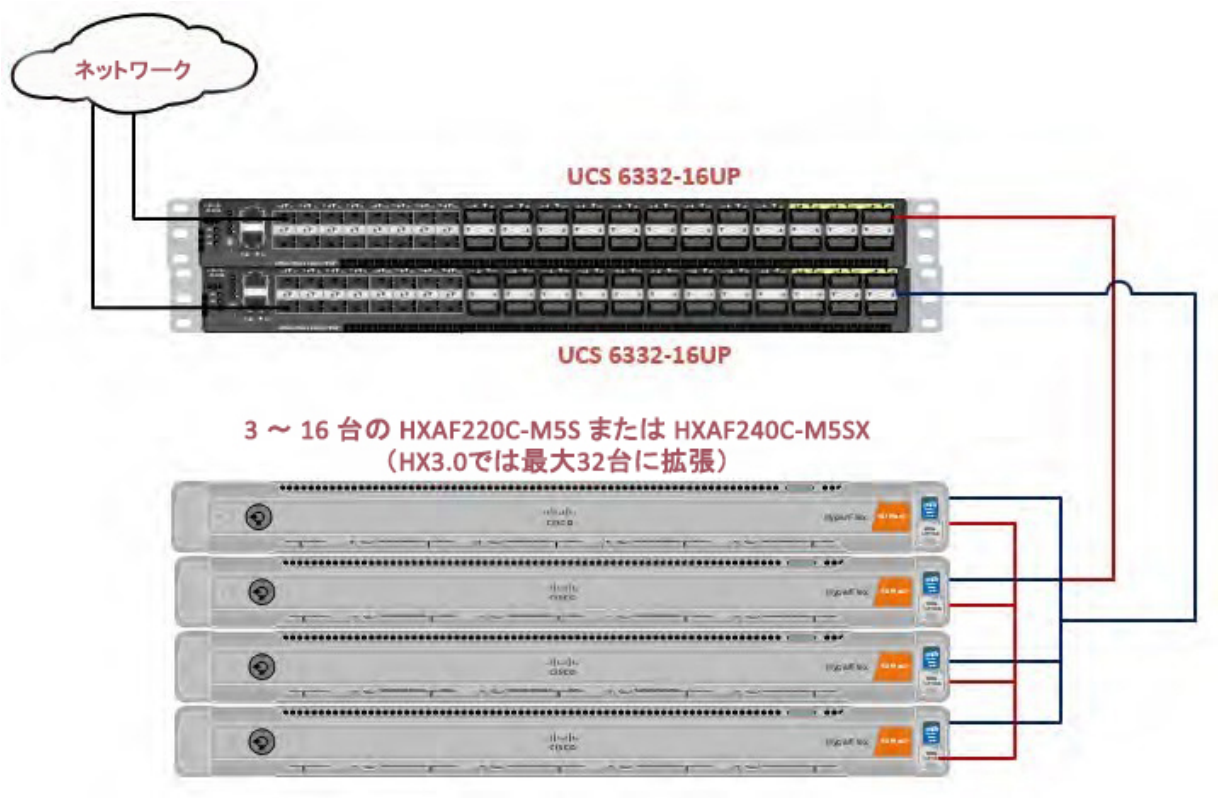



Cisco HyperFlex System は、1 組の Cisco UCS 6200 または 6300 シリーズ ファブリック インターコネクトから構成されており、クラスタあたり最大 16（HX3.0 より 32）の Cisco HyperFlex HX シリーズ オール フラッシュ ラック マウント サーバを備えています。さらに、クラスタごとに最大 16（HX3.0 より 32）のコンピューティング専用サーバを追加できます。Cisco UCS 5108 ブレード サーバ シャーシを追加することで、ハイブリッド クラスタ設計において追加のコンピューティング リソースに Cisco UCS B200 M5 ブレード サーバを使用できます。また、Cisco UCS C240 と C220 サーバを、追加のコンピューティング リソースに使用することもできます。1 組のファブリック インターコネクトに、最大 8 つの HX シリーズ クラスタを個別にインストールすることが可能です。ファブリック インターコネクトは、すべての HX シリーズ ラック マウント サーバとすべての Cisco UCS 5108 ブレード サーバ シャーシの両方に接続されます。インストール時には、ファブリック インターコネクトから顧客のデータセンター ネットワークに、ノースバウンド ネットワーク接続とも呼ばれるアップストリーム ネットワーク接続が行われます。

このドキュメントで使用されている構成では、Cisco UCS 6332-16UP ファブリック インターコネクトが、Cisco Nexus® 9372PX スイッチにアップリンクされています。

図 10 に、このドキュメントで使用されるハイパーコンバージド トポロジを示します。

**図 10.** Cisco HyperFlex の標準トポロジ




Cisco HyperFlex クラスタでのグラフィック サポートは、Cisco UCS B200 M5 ブレード サーバ、または Cisco UCS C240 M5 ラックサーバを、クラスタにコンピューティング専用ノードとして追加することで可能となります。

Cisco HyperFlex HX シリーズ サーバの詳細については、次の URL を参照してください。
http://www.cisco.com/c/ja_jp/products/hyperconverged-infrastructure/index.html

## NVIDIA GRID

NVIDIA GRID は、複数の仮想デスクトップとアプリケーション インスタンス全体で物理 GPU を共有するための業界で最も先進的な技術です。NVIDIA GPU の性能をフル活用することで、場所を問わず、すべてのデバイスに優れた仮想グラフィック エクスペリエンスを提供できるようになりました。NVIDIA GRID プラットフォームは、パフォーマンス、柔軟性、管理性、セキュリティにおいて最高のレベルを提供し、すべての仮想ワークフローにおいて適切な水準のユーザ エクスペリエンスを実現します。

NVIDIA GRID テクノロジーの詳細については、次の URL を参照してください。http://www.nvidia.co.jp/object/grid-technology-jp.html

### NVIDIA GRID 5.0 GPU

NVIDIA GRID ソリューションは、数々の業界評価実績のある、NVIDIA Maxwell および Pascal で動作する GPU で実行されます。これらの GPU は、ブレード サーバおよびコンバージド インフラストラクチャ用 NVIDIA Tesla P6、ラックおよびタワーサーバ用 NVIDIA Tesla M10、P40 のカードがブレード、ラックサーバに搭載可能です。

### NVIDIA GRID カード

デスクトップ仮想化アプリケーションの場合、NVIDIA Tesla P6、M10、P40 カードが高性能グラフィックに最適です（表 2）。

**表 2.** NVIDIA GRID カードの技術仕様



| GPU の数 | シングル ミッドレンジ Pascal | クワッド ミッドレベル Maxwell | シングル ハイエンド Pascal |
|---|---|---|---|
| NVIDIA Compute Unified Device Architecture（CUDA）コア | 2048 | 2560（GPU あたり 640） | 3840 |
| メモリ サイズ | 16 GB GDDR5 | 32 GB GDDR5<br>（GPU あたり 8 GB） | 24 GB GDDR5 |
| vGPU インスタンスの最大数 | 16 | 64 | 24 |
| 電源 | 75 W | 225 W | 250 W |
| フォーム ファクタ | モバイル PCI Express モジュール（MXM：ブレード サーバ）、X 16 レーン | PCIe 3.0 デュアル スロット（ラック サーバ）、X 16 レーン | PCIe 3.0 デュアル スロット（ラック サーバ）、X 16 レーン |
| 冷却ソリューション | ベア ボード | パッシブ | パッシブ |
| H.264 1080p30 ストリーム | 24 | 28 | 24 |
| ボードごとのユーザの最大数 | 16（1 GB プロファイル） | 32（1 GB プロファイル） | 24（1 GB プロファイル） |
| 仮想化使用例 | ブレードの最適化 | ユーザ密度の最適化 | パフォーマンスの最適化 |

### NVIDIA GRID 5.0 ライセンス要件

GRID ソフトウェア 5.0 では、同時ユーザ ライセンスと、ライセンスを管理するオンプレミス NVIDIA ライセンス サーバが必要です。ゲスト OS が起動すると、NVIDIA ライセンス サーバにアクセスし、同時使用ライセンスを 1 つ消費します。そして、ゲスト OS のシャット ダウン時に、ライセンスがプールに返されます。

また GRID 5.0 では、NVIDIA のサポート、更新、保守、サブスクリプション（SUMS）インスタンスに対して、同時ライセンスを 1：1 の比率で購入することも必要です。

NVIDIA Tesla GPU のライセンス製品としては、以下の NVIDIA GRID 製品があります。

- 仮想ワークステーション
- 仮想 PC
- 仮想アプリケーション

GRID 5.0 ライセンス要件のすべての詳細については、次の URL を参照してください。
https://images.nvidia.com/content/grid/pdf/GRID-Licensing-Guide.pdf [英語]

## VMware vSphere 6.5

VMware では、仮想化ソフトウェアが提供されます。VMware のサーバ向けエンタープライズ ソフトウェア ハイパーバイザ（VMware vSphere ESX、vSphere ESXi、vSphere）は、基盤となるオペレーティング システムを追加することなく、サーバ ハードウェア上で直接実行できるベアメタル ハイパーバイザです。vSphere 用の VMware vCenter サーバでは、クラスタ、ホスト、仮想マシン、ストレージ、ネットワーク、および仮想インフラストラクチャ上の他の重要な要素に対し、一元管理、完全なコントロール、可視性を提供します。

vSphere 6.5 により、vSphere ハイパーバイザ、VMware 仮想マシン、vCenter サーバ、仮想ストレージ、仮想ネットワークの機能が大きく強化され、vSphere プラットフォームのコア機能がさらに拡張されます。

vSphere 6.5 プラットフォームは次の機能を備えています。

- コンピューティング
  - **スケーラビリティの向上：**vSphere 6.5 では、より大きな最大構成サイズがサポートされています。仮想マシンでは、最大 128 の仮想 CPU（vCPU）と 6128 GB の仮想 RAM（vRAM）がサポートされます。ホストでは、最大 576 の CPU と 12 TB の RAM、ホストあたり 1024 の仮想マシン、クラスタあたり 64 のノードがサポートされます。
  - **サポートの拡張：**最新の x86 チップセット、デバイス、ドライバ、ゲスト オペレーティング システムに対する拡張サポートが提供されます。サポート対象のゲスト オペレーティング システムのすべての一覧については、VMware の互換性ガイドを参照してください。
  - **優れたグラフィック：**NVIDIA GRID vGPU により、仮想化ソリューションに対して、NVIDIA のハードウェア アクセラレータ グラフィックの利点がすべて提供されます。
  - **インスタント クローニング：**vSphere 6.0 に組み込まれた技術により、迅速なクローニングと仮想マシン展開のための基盤が提供され、現在可能な水準よりも最大で 10 倍高速になります。
- ストレージ
  - **仮想マシン ストレージの変換：**vSphere 仮想ボリュームにより、外部ストレージ アレイの仮想マシン対応を有効にします。ストレージ ポリシー ベース管理（SPBM）により、ストレージ階層およびダイナミック ストレージ サービス クラス（CoS）の自動化において、共通管理を行うことができます。これらの機能を同時に有効にすることで、データ サービス（クローンやスナップショットなど）を正確に組み合わせ、仮想マシン ベースでの効率的なインストールを行うことが可能となります。
- ネットワーク
  - **ネットワーク I/O 制御：**VMware 分散仮想スイッチ（DVS）の帯域幅の予約が新しく仮想マシンごとにできるようになり、帯域幅の確実な分離や制限が実現します。
  - **マルチキャスト スヌーピング：**VDS における IPv4 パケットの Internet Group Management Protocol（IGMP）スヌーピング、および IPv6 パケットのマルチキャスト リスナー検出（MLD）スヌーピングのサポートにより、マルチキャスト トラフィックにおけるパフォーマンスとスケーラビリティを改善します。
  - **VMware vMotion におけるマルチプル TCP/IP スタック：**vMotion トラフィック専用のデフォルト ゲートウェイを使用することで、vMotion トラフィックに専用のネットワーク スタックを実装し、IP アドレスの管理をシンプルにします。
- 可用性
  - **vMotion の機能強化：**仮想スイッチおよび vCenter サーバ間で、最大 100 ミリ秒（ms）のラウンドトリップ時間（RTT）以上の距離におけるワークロードのライブ マイグレーションを無停止で実施します。長距離での vMotion に対して、このような非常に長い RTT（サポート時間の 10 倍に増加）をサポートすることで、物理的にニューヨークとロンドンに位置するデータセンター間でワークロードを互いに移行させることが可能になりました。
  - **vMotion のレプリケーション支援：**2 つのサイト間でアクティブ - アクティブのレプリケーション設定を行っているユーザが、より効率的な vMotion 移行を行えるようになり、その結果として、データの移行量に応じて最大 95 % 以上に達する効率的な移行により、時間とリソースを大幅に節約できるようになります。

- **耐障害性（最大 4 vCPU）**：最大 4 つの vCPU を使用することで、ワークロードに対するソフトウェア ベースの耐障害性のサポートを拡張します。

- 管理
  - **コンテンツ ライブラリ**：この一元的なリポジトリにより、仮想マシンのテンプレート、ISO イメージ、スクリプトなどのコンテンツの容易かつ効果的な管理を実現します。VMware vSphere のコンテンツ ライブラリを使用することで、コンテンツを一元的に格納/管理し、パブリッシュ - サブスクライブ モデルを通じてコンテンツを共有できます。
  - **vCenter 間でのクローニングおよび移行**：異なる vCenter サーバ上のホスト間での仮想マシンのコピーと移動が、単一アクションで可能です。
  - **強化されたユーザ インターフェイス**：VMware vSphere Web クライアントは、従来のものに比べ、より反応が早く、直感的で、シンプルになっています。

## Citrix XenDesktop および XenApp におけるグラフィック アクセラレーション

Citrix HDX 3D Pro により、OpenGL や DirectX に基づいた専用 3D グラフィック アプリケーションなどのハードウェア アクセラレーション用 GPU を使用した最高のパフォーマンスを、デスクトップやアプリケーションで利用できるようになります（標準的な仮想配信エージェント（VDA）は、DirectX の GPU アクセラレーションのみをサポートします）。

次に、3D プロフェッショナル アプリケーションの例を示します。

- コンピュータ支援設計（CAD）、製造（CAM）、エンジニアリング（CAE）アプリケーション
- 地理情報システム（GIS）ソフトウェア
- 医療画像用画像アーカイブ通信システム（PACS）
- OpenGL、DirectX、NVIDIA CUDA、OpenCL の最新バージョンを使用するアプリケーション
- 並列計算のために CUDA GPU を使用して、グラフィック処理以外の負荷の高いコンピューティングを行うアプリケーション

HDX 3D Pro は、あらゆる帯域で優れたユーザ エクスペリエンスを提供します。

- WAN 接続：1.5 Mbps と低い帯域幅でも WAN 接続を介したインタラクティブなユーザ エクスペリエンスを提供
- LAN 接続：100 Mbps の帯域幅で LAN 接続のローカル デスクトップと同等のユーザ エクスペリエンスを提供

グラフィック処理を集中管理用のデータセンターに移すことで、複雑で高価なワークステーションをよりシンプルなユーザ デバイスに置き換えることができます。

HDX 3D Pro では、Microsoft Windows デスクトップおよび Microsoft Windows サーバ用の GPU アクセラレーションが提供されます。VMware vSphere 6 と NVIDIA GRID GPU で使用する場合、HDX 3D Pro により、Windows デスクトップに vGPU アクセラレーションが提供されます。詳細については、次の URL を参照してください。https://www.citrix.co.jp/products/xenapp-xendesktop/hdx-3d-pro.html

### Microsoft Windows デスクトップにおけるグラフィック アクセラレーション

Citrix HDX 3D Pro を使用することで、ホストされたデスクトップ、もしくは Windows OS のデスクトップ マシン上のアプリケーションの一部としてグラフィックを多用するアプリケーションを配信できます。HDX 3D Pro は、物理ホスト コンピュータ（デスクトップ、ブレード、ラック ワークステーションなど）、GPU パススルー、VMware vSphere ハイパーバイザによって提供される GPU 仮想化技術をサポートしています。

GPU パススルーを使用することで、専用のグラフィック処理ハードウェアへの排他的なアクセスが可能な仮想マシンを作成できます。ハイパーバイザ上で複数の GPU をインストールし、各 GPU をそれぞれの仮想マシンに 1 対 1 で割り当てることができます。

GPU 仮想化を使用することで、複数の仮想マシンが単一の物理 GPU のグラフィック処理能力に直接アクセスできます。実際にハードウェア GPU を共有することで、複雑で要求の厳しい設計要件を持つユーザに適したデスクトップを提供します。NVIDIA GRID カードの GPU 仮想化は、仮想化されていないオペレーティング システムに展開される NVIDIA グラフィック ドライバと同じものを使用します。

HDX 3D Pro では、次の機能が提供されます。

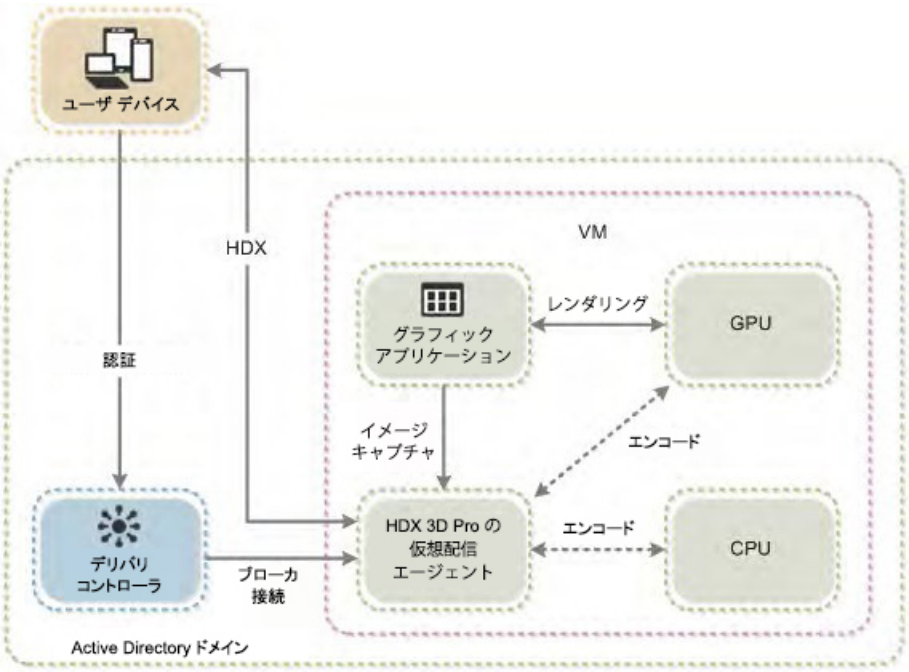
- **H.264 をベースとした適応型の圧縮技術（Adaptive H.264-based Deep Compression）による WAN およびワイヤレス パフォーマンスの最適化：**HDX 3D Pro では、エンコーディング時のデフォルトの圧縮手法として、CPU ベースの H.264 による全画面圧縮が使用されています。また、NVIDIA NVENC をサポートする NVIDIA カードを使用して、ハードウェア エンコードが行われます。
- **特別な使用例に対する無損失圧縮オプション：**HDX 3D Pro では、医療用画像などのピクセル単位での再現が必要なアプリケーションをサポートするために、CPU ベースの無損失コーデックが提供されています。非常に多くのネットワークと処理リソースが消費されることから、純粋な無損失圧縮は特別な使用例にのみ推奨されます。
  - 無損失圧縮が使用されるのは、次のような場合です。
    - 無損失インジケータのシステム トレイ アイコンが、表示画面のフレームに損失があるかないかをユーザに示します。この情報は、ビジュアル品質ポリシー設定で無損失ビルドを指定している際に有用です。送信フレームの損失がない場合は、無損失インジケータが緑に変わります。
    - 無損失スイッチにより、ユーザはセッション中の任意のタイミングで、常に無損失にするモードに変更することができます。セッション中の任意のタイミングで、[常に無損失（Always Lossless）] を選択、または選択解除するには、アイコンを右クリックするか、Alt+Shift+1 のショートカット キーを使用します。
  - 無損失圧縮のためには、HDX 3D Pro で、ポリシーで選択したコーデックとは別に圧縮用の無損失コーデックを使用します。
  - 有損失圧縮のためには、HDX 3D Pro で、ポリシーで選択したコーデックか、デフォルトのコーデックを使用します。
  - 以降のセッションでは、無損失スイッチの設定は保持されません。接続ごとに無損失コーデックを使用するには、ビジュアル品質ポリシー設定で [常に無損失（Always Lossless）] を選択します。
- **複数の高解像度モニタのサポート：**Microsoft Windows 7 および 8 のデスクトップの場合、HDX 3D Pro では、ユーザ デバイスで最大 4 つのモニタをサポートしています。ユーザは、任意の構成で自身のモニタを配置することができ、異なる解像度と向きのモニタを混在させることができます。モニタの数は、ホスト コンピュータの GPU の機能、ユーザ デバイス、使用可能な帯域幅によって制限されます。HDX 3D Pro は、すべてのモニタの解像度をサポートし、ホスト コンピュータの GPU の能力によってのみ制限されます。
- **動的解像度：**仮想デスクトップまたはアプリケーション ウィンドウのサイズを任意の解像度に合わせて変更できます。
- **NVIDIA Kepler アーキテクチャのサポート：**HDX 3D Pro では、GPU パススルーと GPU 共有のために、NVIDIA GRID K1 カードと K2 カードがサポートされています。GRID vGPU により、仮想化されていないオペレーティング システムに展開される同じ NVIDIA グラフィック ドライバを使用して、複数の仮想マシンが単一の物理 GPU に同時かつ直接にアクセスできます。
- **VMware vSphere と vDGA を使用する ESX のサポート：**リモート デスクトップ サービス（RDS）と VDI ワークロードの両方に対して、vDGA を備えた HDX 3D Pro を使用できます。vSGA を備えた HDX 3D Pro を使用する場合は、1 台のモニタのサポートに限られます。大規模な 3D モデルで vSGA を使用する場合は、API インターセプト技術が使用されるため、パフォーマンスの問題につながる場合があります。詳細については、VMware vSphere 5.1：Citrix の既知の問題を参照してください。

図 11 を参照：

- ホスト コンピュータは、配信コント ローラと同じ Active Directory ドメインに配置する必要があります。
- ユーザが Citrix Receiver にログオンし、仮想アプリケーションまたはデスクトップにアクセスする場合、コントローラはユーザを認証し、HDX 3D の VDA と通信して、グラフィカル アプリケーションをホストしているコンピュータへの接続を仲介します。

- HDX 3D Pro の VDA は、ホスト上で適切なハードウェアを使用し、デスクトップ全体またはグラフィカル アプリケーションのみを圧縮して表示します。
- デスクトップまたはアプリケーションの表示とユーザとの相互作用は、Citrix Receiver と HDX 3D Pro の VDA 間の直接の HDX 接続を通じて、ホスト コンピュータとユーザ デバイス間で転送されます。

**図 11.** Citrix HDX 3D Pro のプロセス フロー




### Microsoft Windows サーバにおける GPU アクセラレーション

Citrix HDX 3D Pro により、Microsoft Windows サーバのセッションで実行されているグラフィックを多用するアプリケーションで、GPU 上でのレンダリングが可能となります。OpenGL、Direct3D、Windows Presentation Foundation（WPF）レンダリングがサーバの GPU に移動することで、サーバの CPU がグラフィック レンダリングで遅くなることがなくなります。また、CPU と GPU で負荷が分けられるため、サーバでより多くのグラフィックが処理できます。

### Citrix XenApp RDS ワークロードの GPU 共有

RDS GPU 共有により、リモート デスクトップ セッションで、OpenGL および Microsoft DirectX アプリケーションによる GPU ハードウェア レンダリングが可能となります。

- 共有はベアメタル デバイスまたは仮想マシンで使用することができ、アプリケーションのスケーラビリティとパフォーマンスが高まります。
- 共有により、複数の同時セッションで GPU リソースを共有できます（ほとんどのユーザで、専用 GPU のレンダリング パフォーマンスが不要になります）。
- 共有には、特別な設定は必要ありません。

DirectX アプリケーションでは、デフォルトで GPU が 1 つだけ使用され、その GPU が複数のユーザで共有されます。DirectX を使った複数の GPU 間でのセッションの割り当ては試験的なもので、レジストリの変更が必要です。詳細については、Citrix のサポートに問い合わせてください。

ハイパーバイザ上に複数の GPU をインストールし、各 GPU にそれぞれの仮想マシンを 1 対 1 で割り当てることができます。つまり、1 枚のグラフィック カードに対して複数の GPU をインストールしたり、複数のグラフィック カードに対して複数の GPU をインストールしたりすることが可能です。なお、サーバ上の種類の異なるグラフィック カードの混在は推奨されません。

仮想マシンでは、VMware vSphere 6 で使用可能な GPU への直接パススルーが必要となります。Citrix HDX 3D Pro が GPU パススルーで使用されている場合、サーバの各 GPU では、マルチユーザ仮想マシンが 1 つサポートされます。

RDS GPU 共有を使用する場合の拡張性については、いくつかの要因に依存します。

- 実行されているアプリケーション
- アプリケーションが消費するビデオ RAM の容量
- グラフィック カードの処理能力

一部のアプリケーションは、ビデオ RAM の不足に対して、他のアプリケーションより適切に処理を行います。ハードウェアが極端にオーバーロードした場合、システムが不安定になるか、グラフィック カードのドライバに障害が発生します。このような問題を避けるには、同時接続ユーザ数を制限します。

GPU アクセラレーションが発生していることを確認するには、GPU-Z などのサード パーティ製ツールを使用します。GPU-Z は http://www.techpowerup.com/gpuz/ [英語] で入手できます。

### Citrix HDX 3D Pro の要件

アプリケーションをホストしている物理/仮想マシンは、GPU パススルー、または vGPU を使用できます。

- GPU パススルーは、Citrix XenServer、VMware vSphere および ESX で利用でき、Virtual Direct Graphics Acceleration（vDGA）と呼称されています。また、Microsoft Windows Server 2016 の Microsoft Hyper-V でも利用でき、Discrete Device Assignment（DDA）と呼称されます。
- vGPU は、Citrix XenServer と VMware vSphere で利用できます。次の URL を参照してください。https://www.citrix.co.jp/products/xenapp-xendesktop/hdx-3d-pro.html
- Citrix では、ホスト コンピュータに少なくとも 4 GB の RAM とクロック速度 2.3 Ghz 以上の仮想 CPU が 4 つ搭載されていることが推奨されています。

GPU の要件は次のとおりです。

- CPU ベースの圧縮（無損失圧縮など）の場合、Citrix HDX 3D Pro は、配信されているアプリケーションと互換性があるホスト コンピュータのディスプレイ アダプタをすべてサポートします。
- NVIDIA GRID API を使用する仮想化グラフィック アクセラレーションの場合は、HDX 3D Pro はサポート対象の GRID カードで使用できます（NVIDIA GRID 参照）。GRID により、高いフレーム レートが提供され、非常にインタラクティブなユーザ エクスペリエンスが実現します。

ユーザ デバイスの要件は次のとおりです。

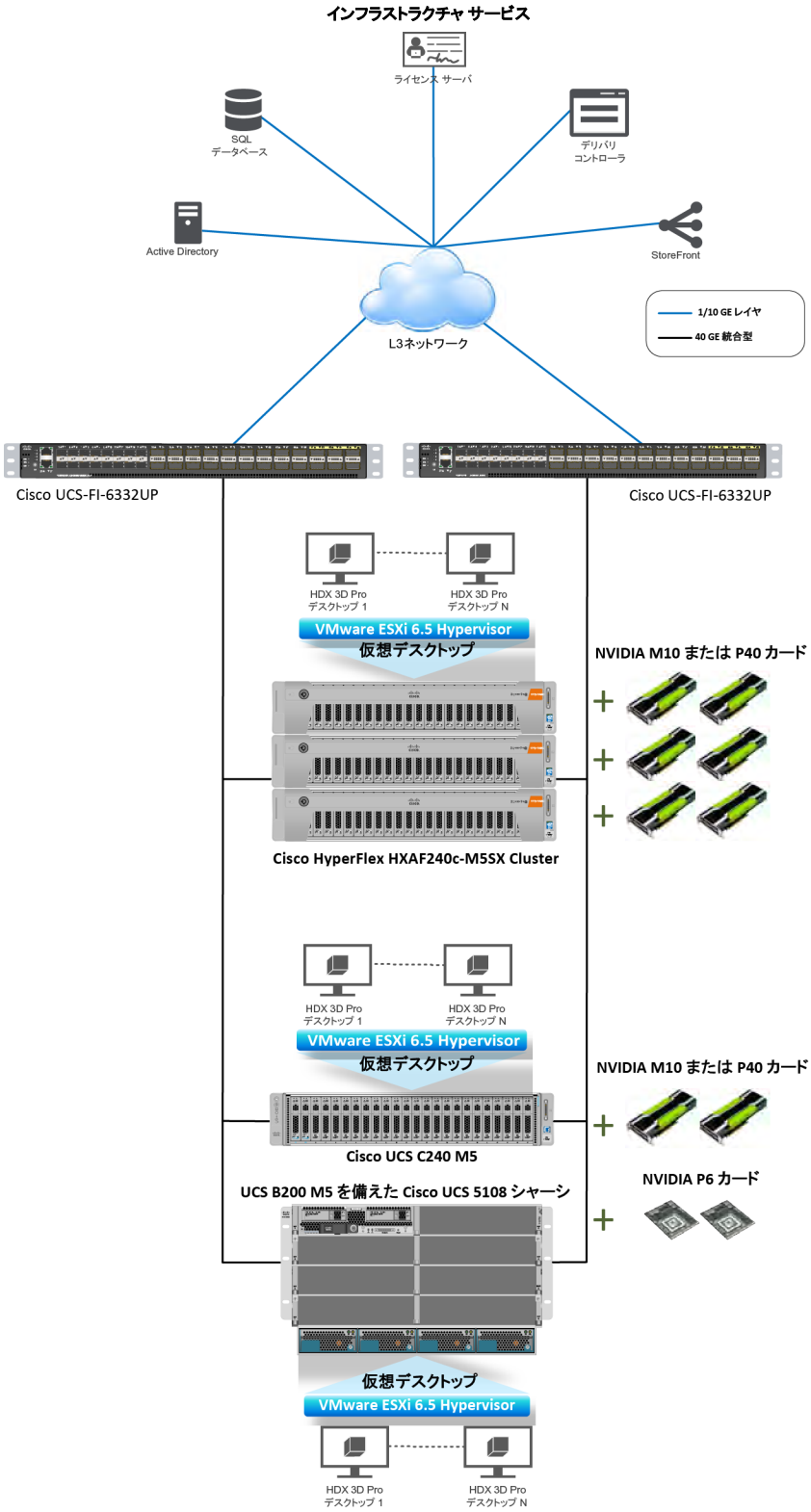
- HDX 3D Pro は、ホスト コンピュータの GPU でサポートされるすべてのモニタの解像度をサポートします。ただし、最小推奨のユーザ デバイスおよび GPU 仕様でパフォーマンスを最適化する場合、Citrix では、ユーザ デバイス用の最大モニタ解像度として、LAN 接続で 1920 X 1200 ピクセル、WAN 接続で 1280 X 1024 ピクセルが推奨されています。
- Citrix では、ユーザ デバイスに少なくとも 1 GB の RAM とクロック速度 1.6 GHz 以上の CPU が搭載されていることが推奨されています。低帯域幅での接続に必要なデフォルトの Deep Compression コーデックを使用する場合、ハードウェアで復号が行われない限りは、より強力な CPU が必要になります。最適なパフォーマンスのためには、Citrix では、ユーザ デバイスに少なくとも 1 GB の RAM とクロック速度 3 GHz 以上のデュアルコア CPU が搭載されていることが推奨されています。
- 複数モニタを利用する場合は、Citrix では、クワッド コア CPU を搭載したユーザ デバイスが推奨されています。
- ユーザ デバイスで HDX 3D Pro を使用して、デスクトップや配信アプリケーションにアクセスする場合には、GPU は必要ありません。
- Citrix Receiver をインストールする必要があります。

詳細については、次の Citrix HDX 3D Pro の記事を参照してください。https://www.citrix.co.jp/products/xenapp-xendesktop/hdx-3d-pro.html および https://docs.citrix.com/ja-jp/xenapp-and-xendesktop/7-15-ltsr/graphics/hdx-3d-pro.html

## ソリューション構成

図 12 に、ソリューション構成の概要を示します。

**図 12.** 参照アーキテクチャ



このソリューションのハードウェア コンポーネントは次のとおりです。

- Cisco UCS C240 M5 ラック サーバ（2.20 GHz Intel Xeon Gold 5120 CPU X 2）、メモリ 768 GB（64 GB X 12 DIMM、2,666 MHz）
- Cisco UCS B200 M5 ブレード サーバ（2.20 GHz Intel Xeon Gold 5120 CPU X 2）、メモリ 768 GB（64 GB X 12 DIMM、2,666 MHz）
- Cisco HyperFlex HX240c M5SX オール フラッシュ ハイパーコンバージド サーバ（2.20 GHz Intel Xeon Gold 5120 CPU X 2）、メモリ 768 GB（64 GB X 12 DIMM、2,666 MHz）
- Cisco UCS VIC 1387 mLOM（Cisco UCS C240 M5 ラック サーバおよび Cisco HyperFlex HX240c M5S オール フラッシュ ノード）
- Cisco UCS VIC 1340 mLOM（Cisco UCS B200 M5 ブレード サーバ）
- Cisco UCS 6332 ファブリック インターコネクト（第 3 世代ファブリック インターコネクト）X 2
- NVIDIA Tesla M10、P40、P6 カード
- Cisco Nexus 9372 スイッチ（オプションのアクセス スイッチ）X 2

**注：**高負荷グラフィック アプリケーションの場合は、Intel Xeon プロセッサ 6154 CPU（3.0 GHz）と NVIDIA Tesla P40 か P6 カードのペアが推奨されます。

このソリューションのソフトウェア コンポーネントは次のとおりです。

- Cisco UCS ファームウェア リリース 3.2（2 c)
- Cisco HXDP 2.6(1a)
- VDI ホスト向け VMware ESXi 6.5（5969303）
- Citrix XenApp と XenDesktop 7.15
- Microsoft Windows 10（64 ビット）
- Microsoft Server 2016
- Microsoft Office 2016
- NVIDIA GRID ソフトウェアとライセンス：
- NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-1OEM.650.0.0.4598673
  - 385.90_grid_win10_server2016_64bit_international
  - vGPU ライセンス サーバ バージョン 5.0.0.22575570（Quadro-Virtual-DWS ライセンス）

## Configure Cisco UCS

This section describes the Cisco UCS configuration.

### Install NVIDIA Tesla GPU card on Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash server

Install the M10 or P40 GPU card on the Cisco UCS C240 M5 Rack Server and Cisco HyperFlex HX240c M5 All Flash Node.

Table 3 lists the minimum firmware required for the supported GPU cards.

**Table 3.** Minimum server firmware versions required for GPU cards

| Cisco Integrated Management Controller (IMC) | BIOS minimum version |
|---|---|
| NVIDIA Tesla M10 | Release 3.1(1) |
| NVIDIA Tesla P40 | Release 3.1(1) |

Mixing different brands or models of GPU cards in the server is not supported.

The rules for configuring the server with GPUs differ depending on the server version and other factors. Table 4 lists rules for populating the Cisco UCS C240 M5 with NVIDIA GPUs.

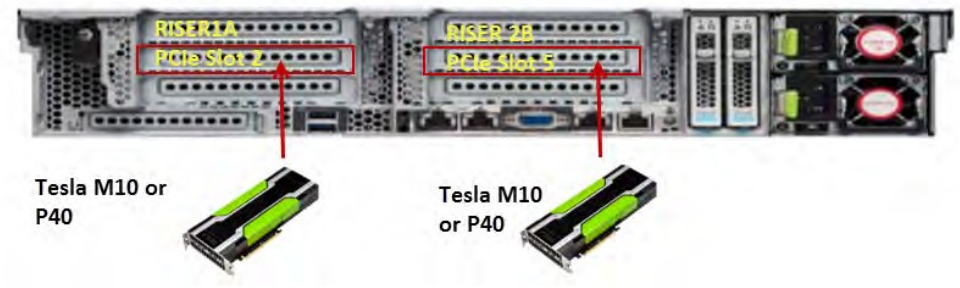
Figure 13 shows a one-GPU installation, and Figure 14 shows a two-GPU installation.

**Table 4.** NVIDIA GPU population rules for Cisco UCS C240 M5 [[SHOULD THIS SAY "M5"?]] Rack Server

| Single GPU | Dual GPU |
|---|---|
| Riser 1, slot 2<br>or<br>Riser 2, slot 5<br>supported in all riser options | Riser 1, slot 2<br>and<br>Riser 2A/2B, slot 5 |

**Figure 13.** One-GPU scenario


**Figure 14.** Two-GPU scenario



For more information, refer to these configuration documents:

- https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/c/hw/C240M5/install/C240M5.pdf
- https://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/hyperflex-hx-series/hxaf-240c-m5-specsheet.pdf

### Install NVIDIA Tesla GPU card on Cisco UCS B200 M5

Install the P6 GPU card on the Cisco UCS B200 M5 server.

Table 5 lists the minimum firmware required for the GPU card. Figure 15 shows the card in the server.

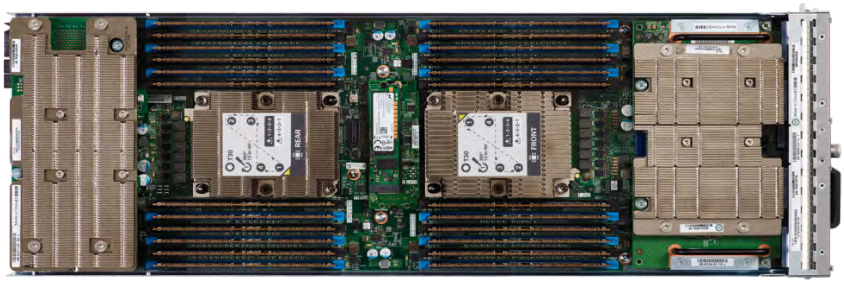
**Table 5.** Minimum server firmware versions required for GPU card

| Cisco Integrated Management Controller (IMC) | BIOS minimum version |
|---|---|
| NVIDIA Tesla M6 | Release 3.2(1d) |

Before installing the NVIDIA P6 GPU, do the following:

- Remove any adapter card, such as a Cisco UCS VIC 1380 or 1280 or a port extender card, from mLOM slot 2. You cannot use any other card in slot 2 when the NVIDIA P6 GPU is installed.
- Upgrade your Cisco UCS system to a version of Cisco UCS Manager that supports this card. Refer to the latest version of the release notes for Cisco UCS software at the following URL for information about supported hardware: http://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-manager/products-release-notes-list.html.

**Figure 15.** Cisco UCS B200 M5 Blade Server with two NVIDIA GRID P6 GPU cards



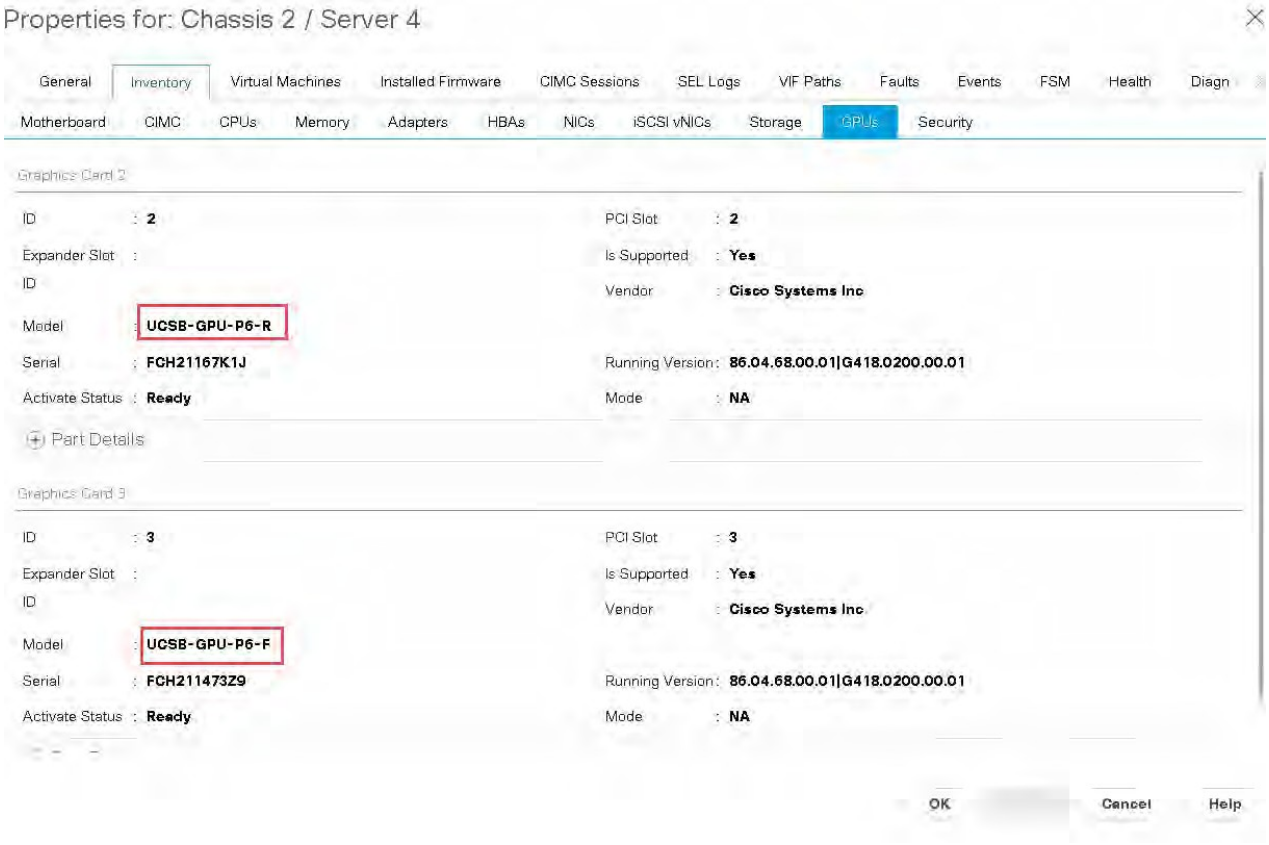
For more information, refer to this configuration document:
https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/hw/blade-servers/B200M5.pdf.

### Configure the GPU card

Follow these steps to configure the GPU card.

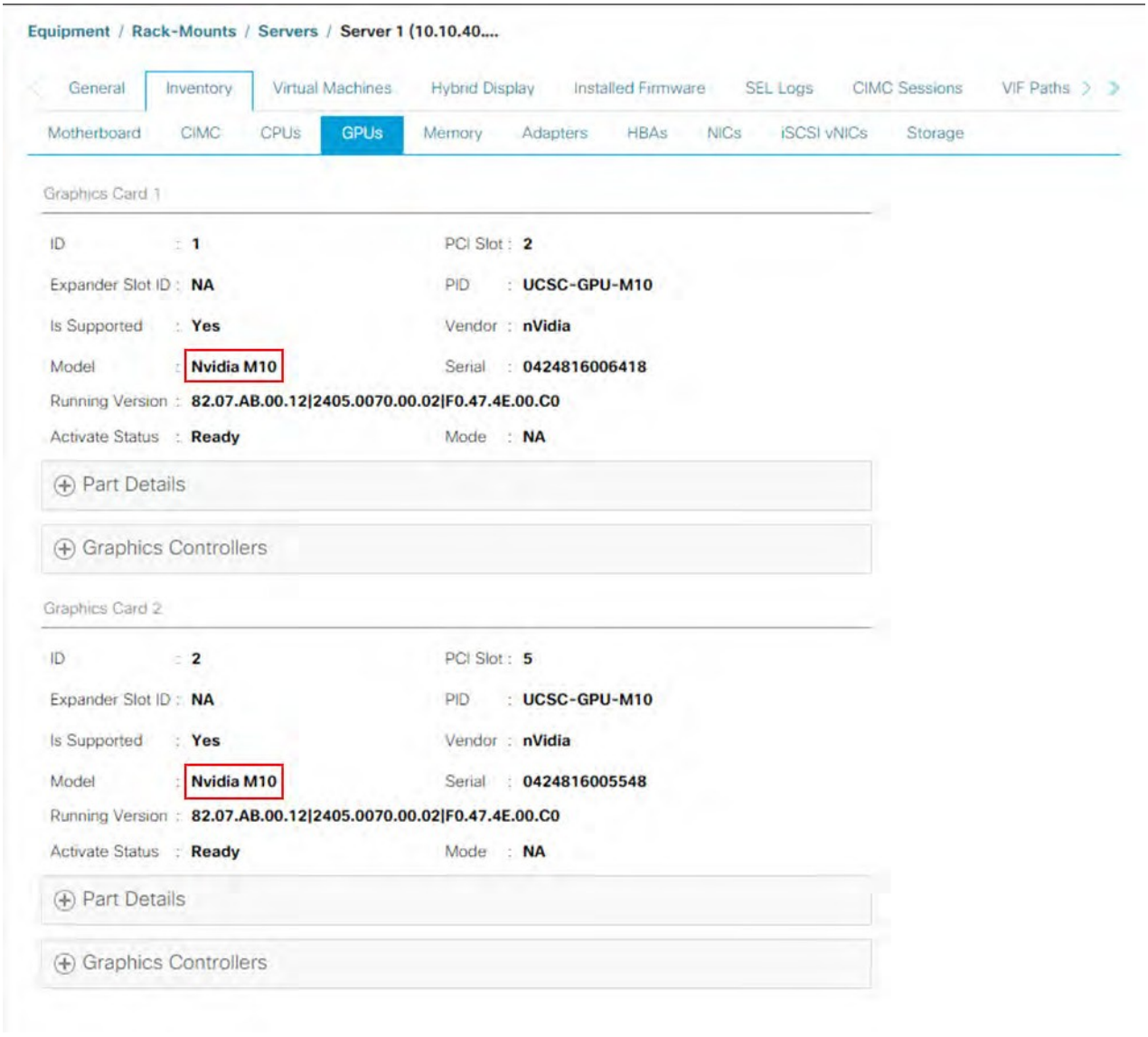
1. After the NVIDIA P6 GPU cards are physically installed and the Cisco UCS B200 M5 Blade Server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 16, PCIe slots 2 and 3 are used with two GRID P6 cards.

**Figure 16.** NVIDIA GRID P6 card inventory displayed in Cisco UCS Manager



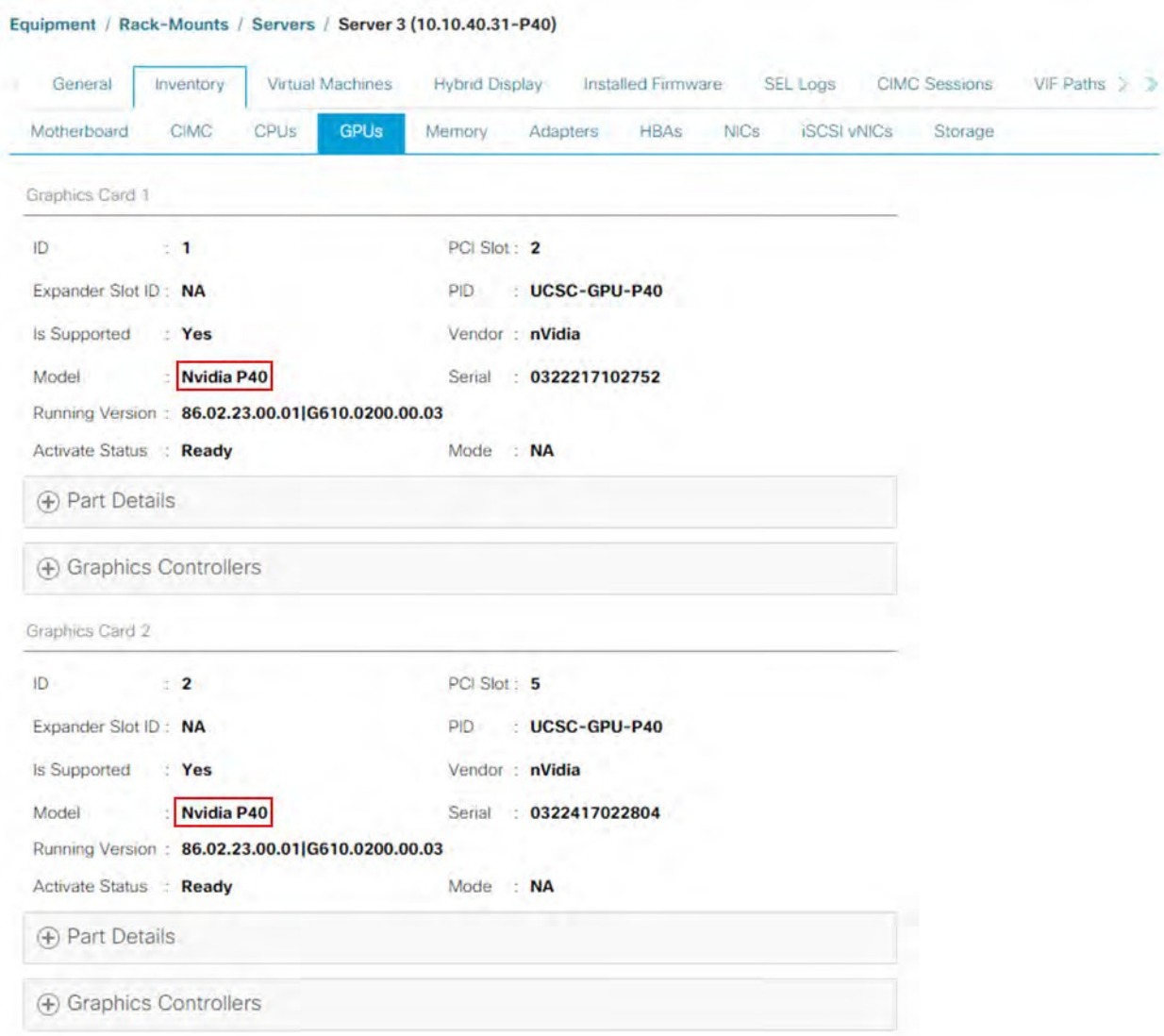

2. After the NVIDIA M10 GPU cards are physically installed and the Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 17, PCIe slots 2 and 5 are used with two GRID M10 cards.

**Figure 17.** NVIDIA GRID M10 card inventory displayed in Cisco UCS Manager




3. After the NVIDIA P40 GPU card is physically installed and the Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash server is discovered in Cisco UCS Manager, select the server and choose Inventory > GPUs. As shown in Figure 18, PCIe slots 2 and 5 are used with the two GRID P40 cards.

**Figure 18.** NVIDIA GRID P40 card inventory displayed in Cisco UCS Manager




You can use Cisco UCS Manager to perform firmware upgrades to the NVIDIA GPU cards in managed Cisco UCS C240 M5 and Cisco HyperFlex HX240c M5 All Flash servers.

**Note:** VMware ESXi virtual machine hardware Version 9 or later is required for vGPU and vDGA configuration. Virtual machines with hardware Version 9 or later should have their settings managed through the VMware vSphere Web Client.

## Install the NVIDIA GRID license server

This section summarizes the installation and configuration process for the GRID 5.0 license server.

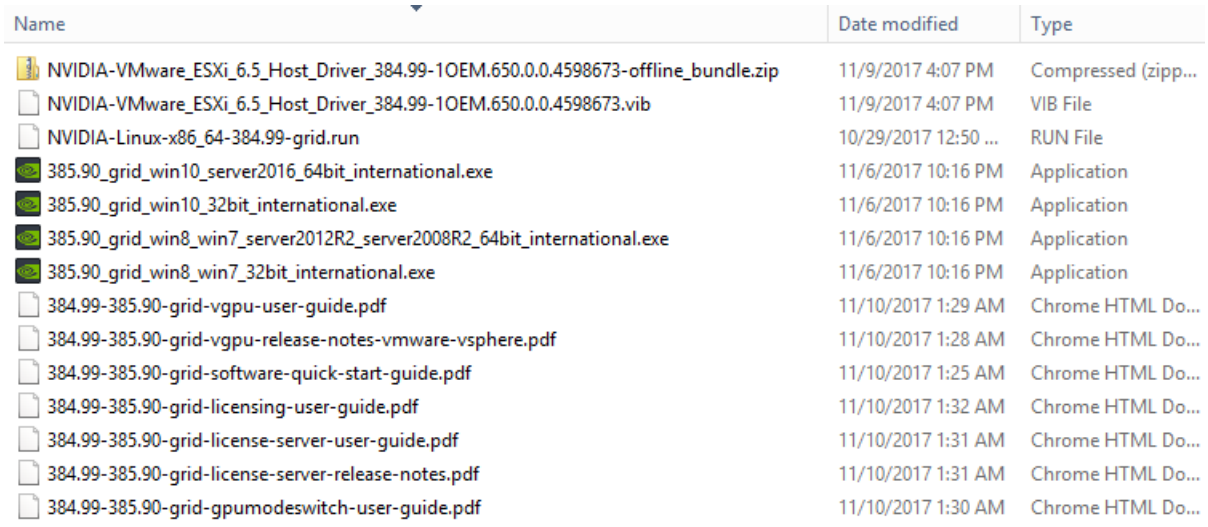
The NVIDIA GRID vGPU is a licensed feature on Tesla P6, P40, and M10 cards. A software license is required to use the full vGPU feature set on a guest virtual machine. An NVIDIA license server with the appropriate licenses is required.

To get an evaluation license code and download the software, register at http://www.nvidia.com/object/grid-evaluation.html#utm_source=shorturl&utm_medium=referrer&utm_campaign=grideval.

Three packages are required for VMware ESXi host setup, as shown in Figure 19:

- The GRID license server installer
- The NVIDIA GRID Manager software, which is installed on VMware vSphere ESXi; the NVIDIA drivers and software that are installed in Microsoft Windows are also in this folder
- The GPU Mode Switch utility, which changes the cards from the default Compute mode to Graphics mode

**Figure 19.** Software required for NVIDIA GRID 5.0 setup on the VMware ESXi host



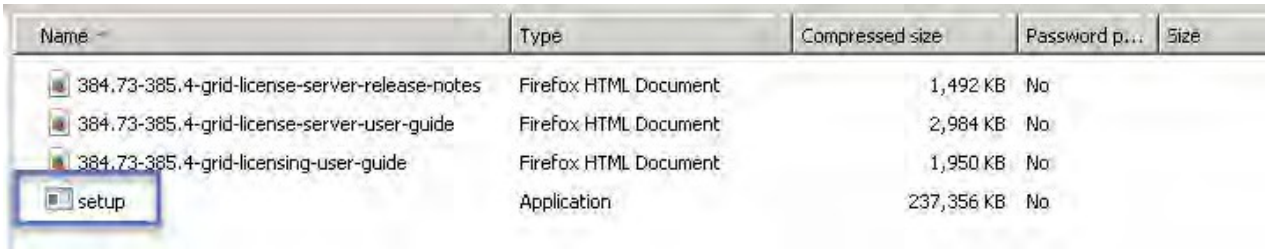

## Install the NVIDIA GRID 5.0 license server

The steps shown here use the Microsoft Windows version of the license server installed on Windows Server 2016. A Linux version of the license server is also available.

The GRID 5.0 license server requires Java Version 7 or later. Go to Java.com and install the latest version.

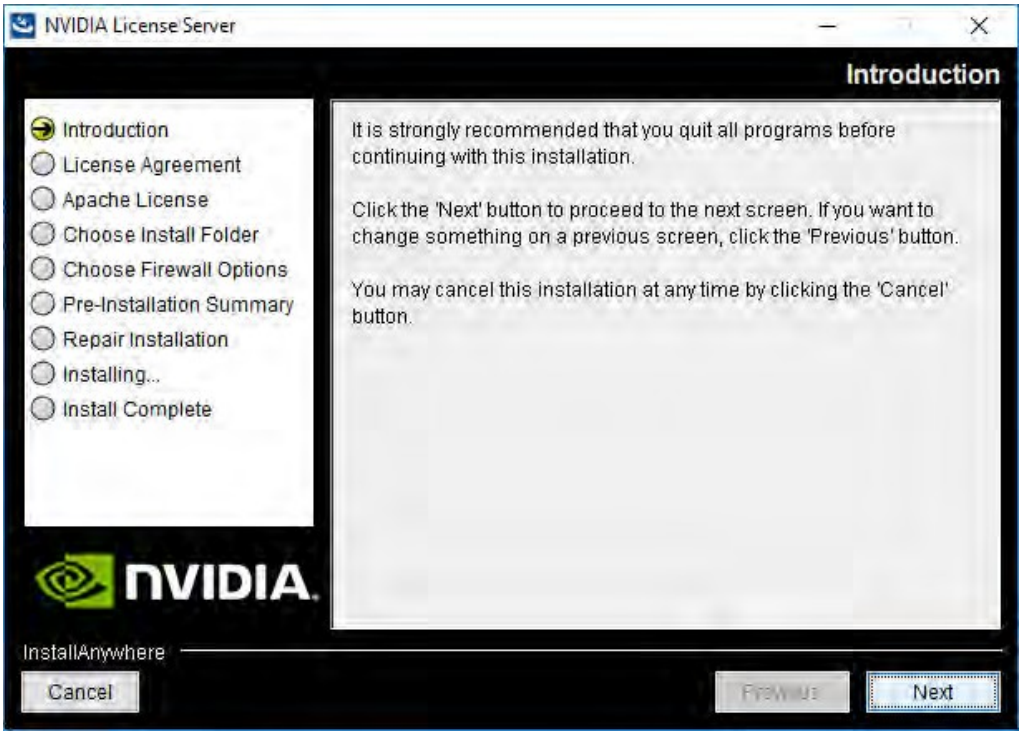
1. Extract and open the NVIDIA-ls-windows-2017.08-0001 folder. Run setup.exe (Figure 20).

**Figure 20.** Run setup.exe



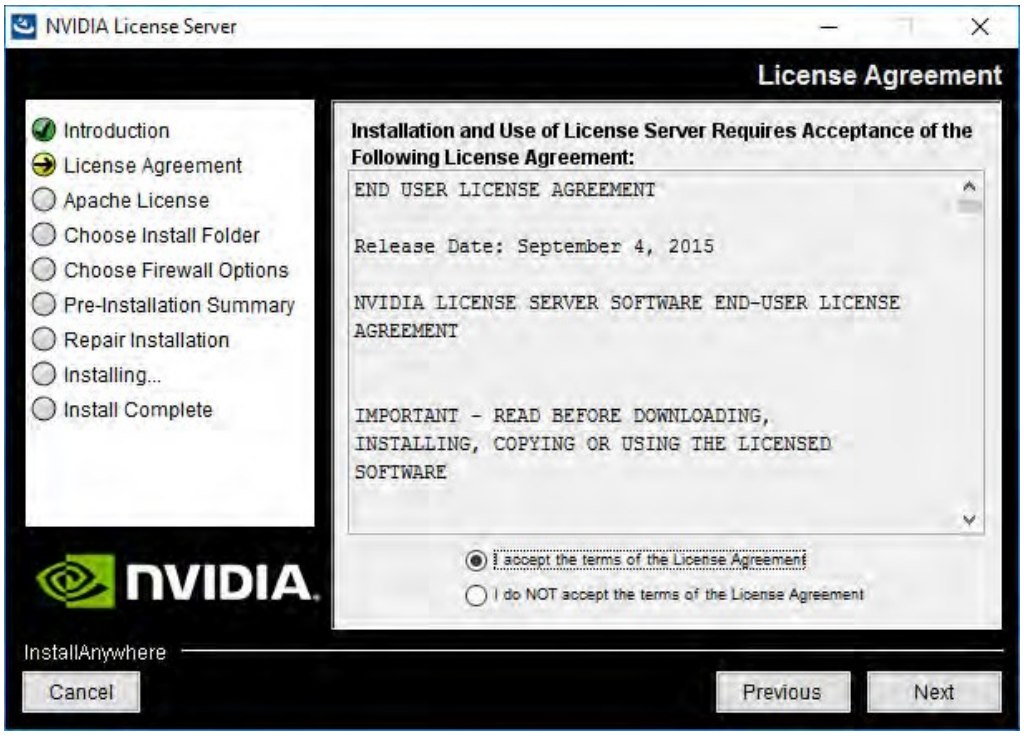
2. Click Next (Figure 21).

**Figure 21.** NVIDIA License Server page



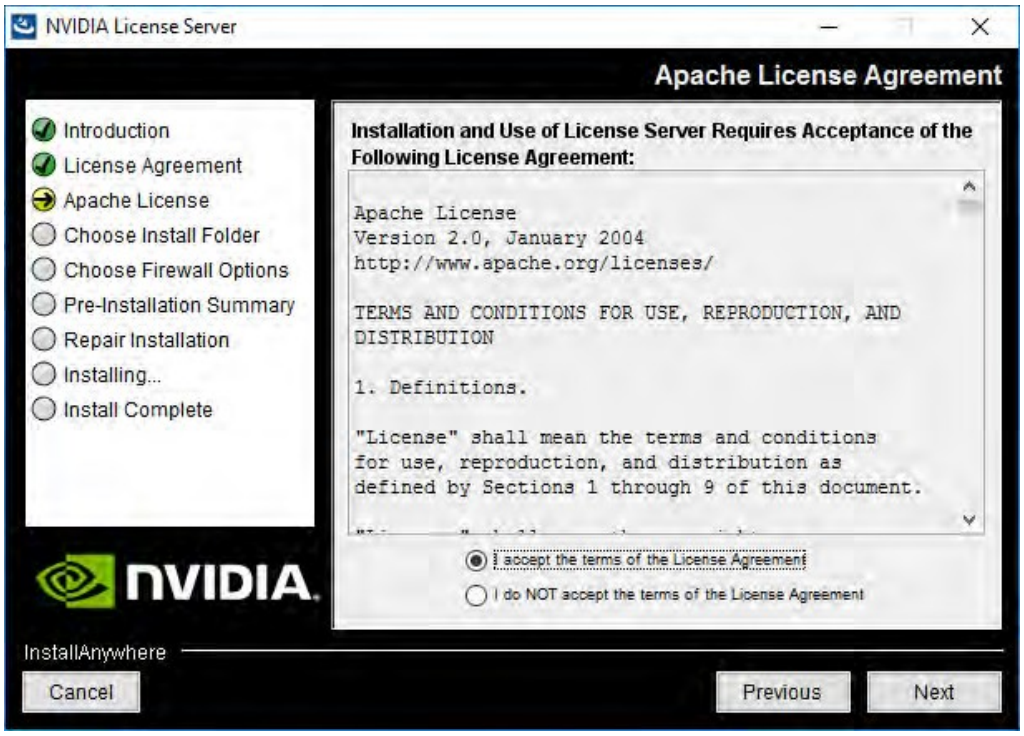

3. Accept the license agreement and click Next (Figure 22).

**Figure 22.** NVIDIA License Agreement page



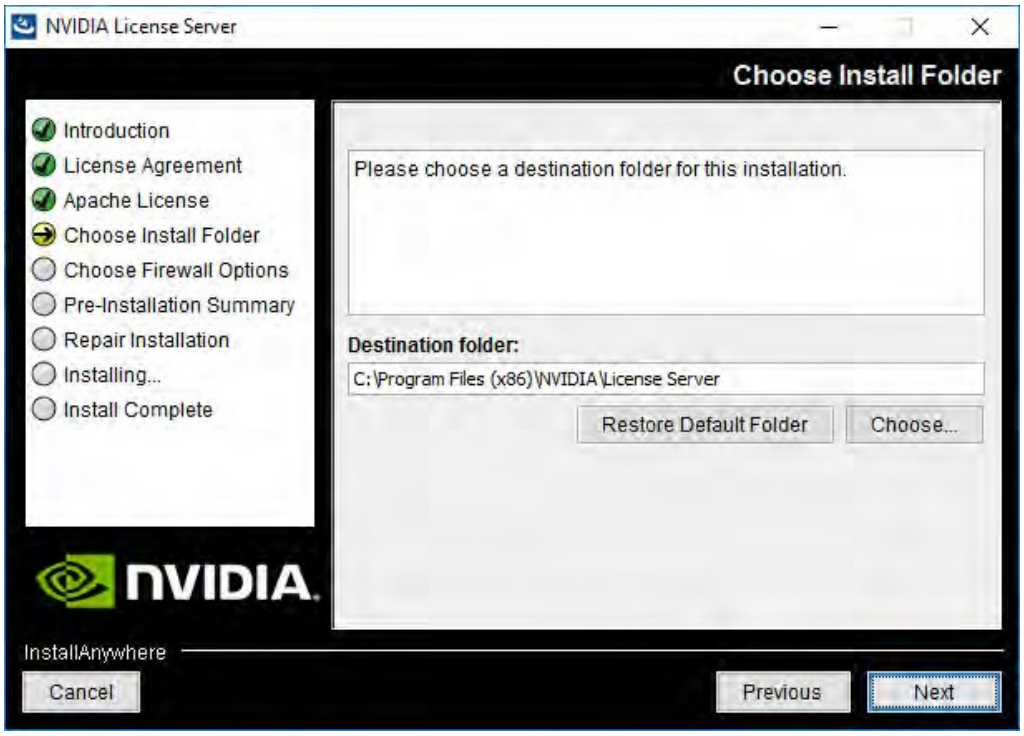

4. Accept the Apache license agreement and click Next (Figure 23).

**Figure 23.** Apache License Agreement page



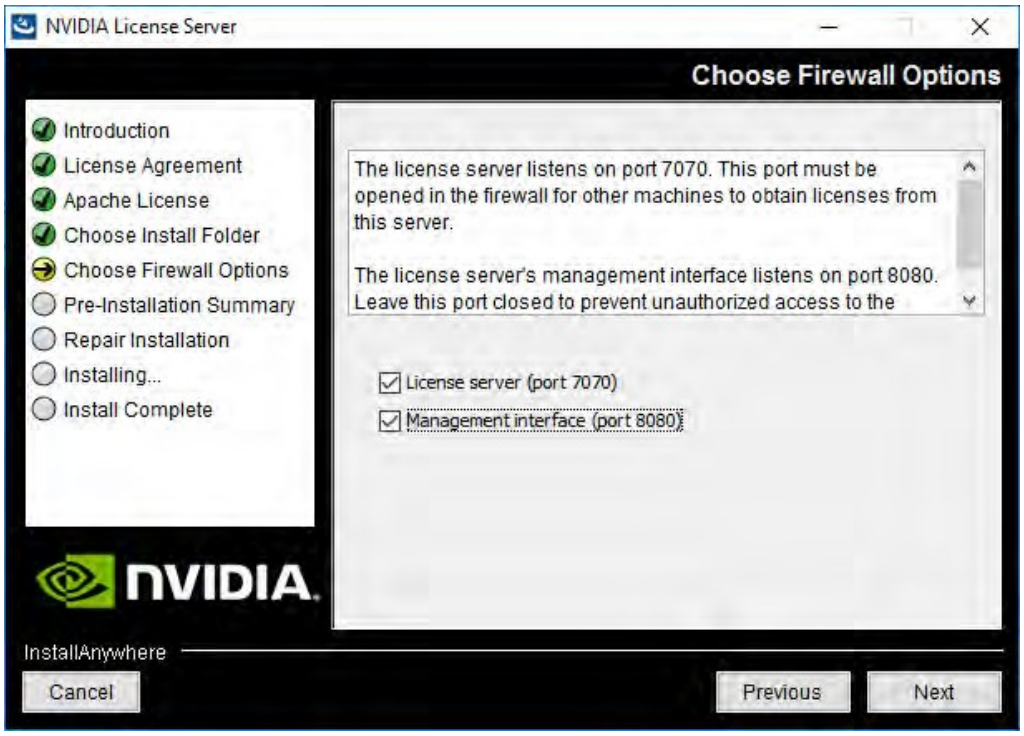

5. Choose the desired installation folder and click Next (Figure 24).

**Figure 24.** Choosing a destination folder




6. The license server listens on port 7070. This port must be opened in the firewall for other machines to obtain licenses from this server. Select the “License server (port 7070)” option.
7. The license server’s management interface listens on port 8080. If you want the administration page accessible from other machines, you will need to open up port 8080. Select the “Management interface (port 8080)” option.
8. Click Next (Figure 25).

**Figure 25.** Setting firewall options




9. On the Pre-installation Summary page, click Install (Figure 26). Installation will automatically progress without user input (Figure 27).

**Figure 26.** Pre-Installation Summary page

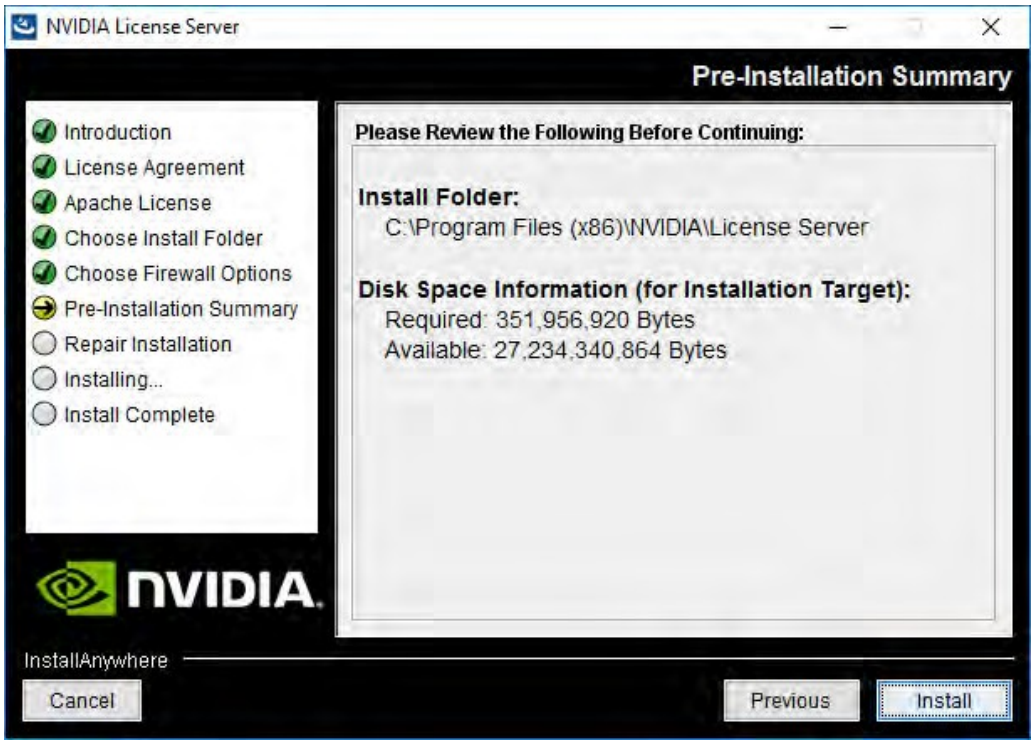
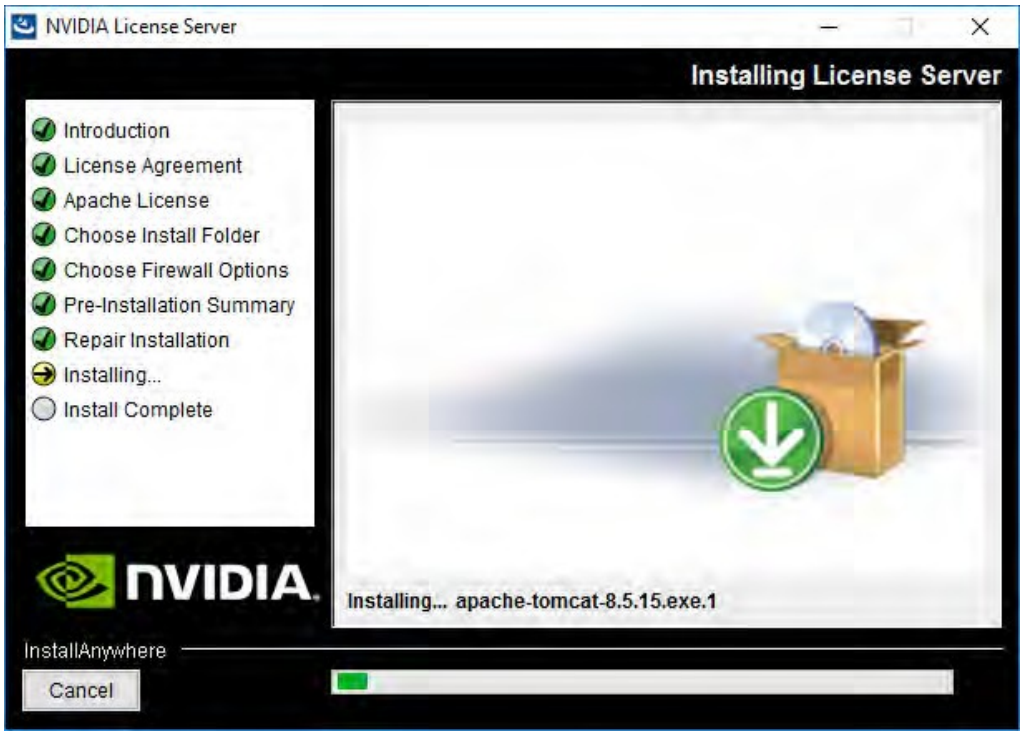
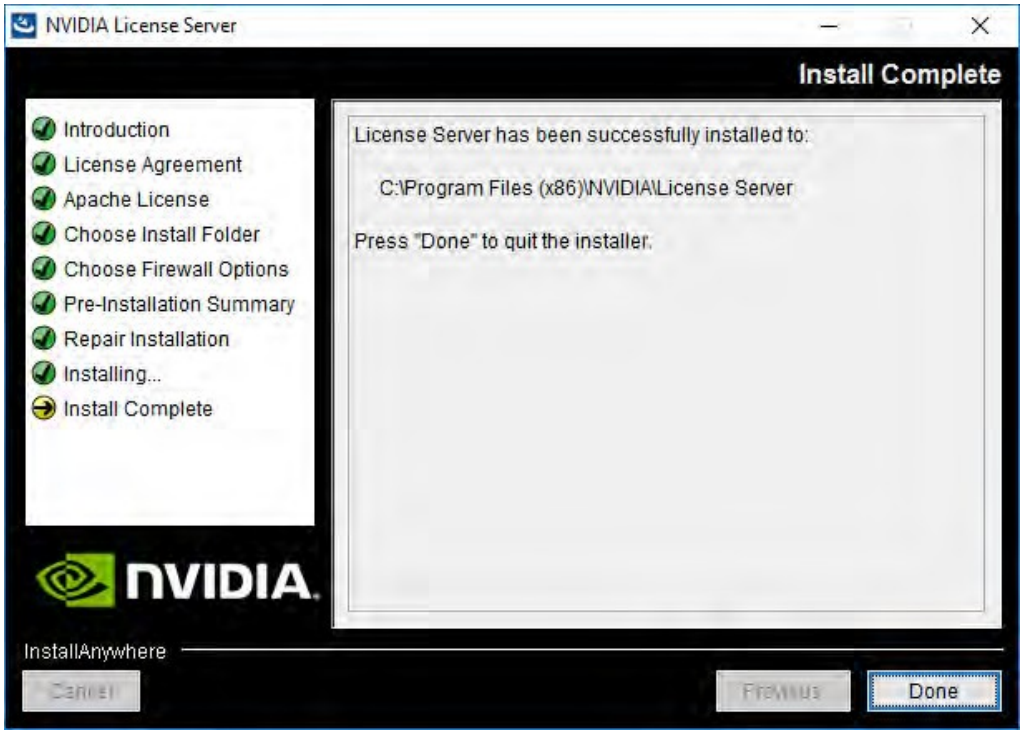



**Figure 27.** Installing the license server



10. When the installation process is complete, click Done (Figure 28).

**Figure 28.** Installation Complete page

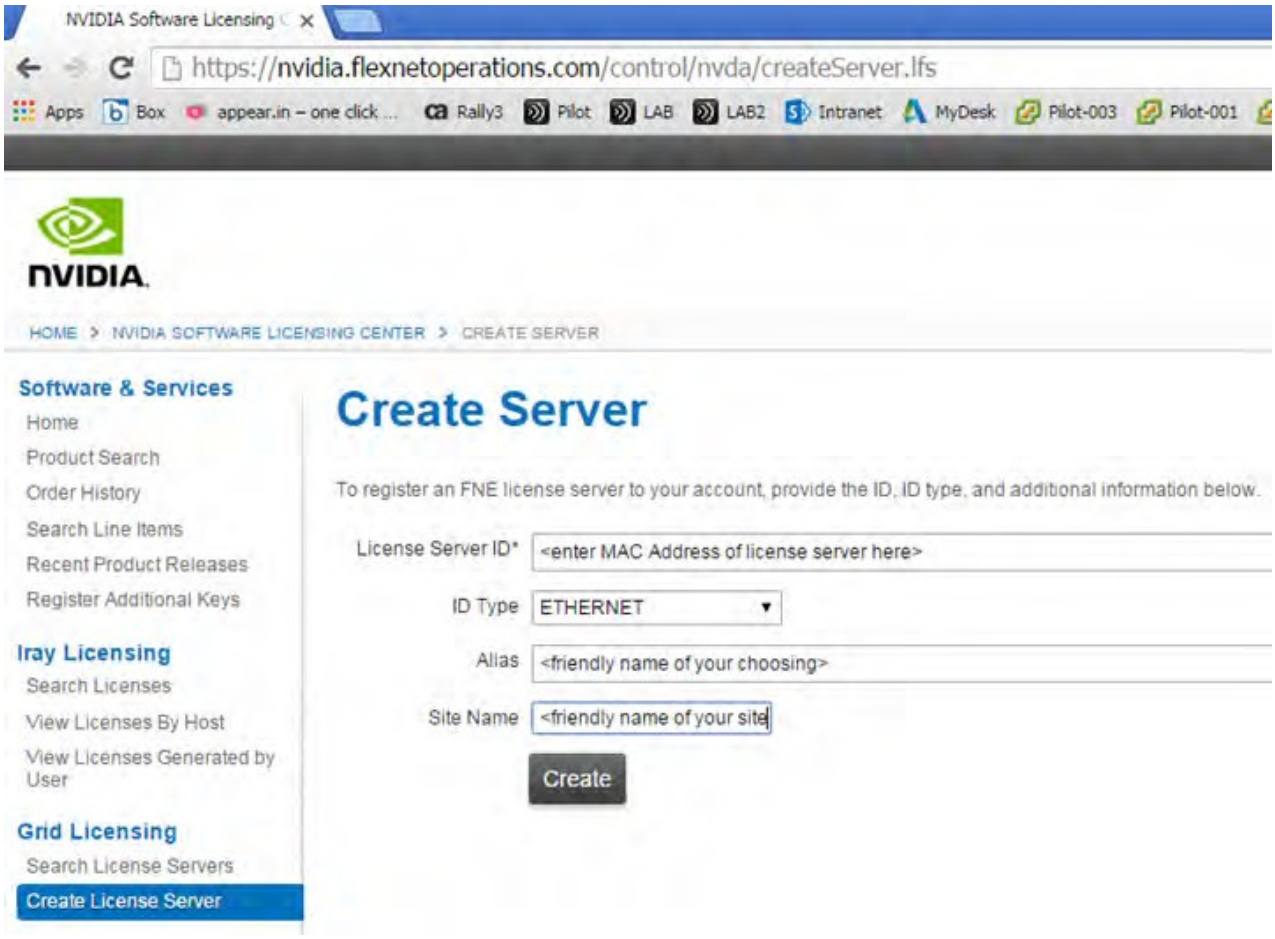



### Configure the NVIDIA GRID 5.0 license server

Now configure the NVIDIA GRID license server.

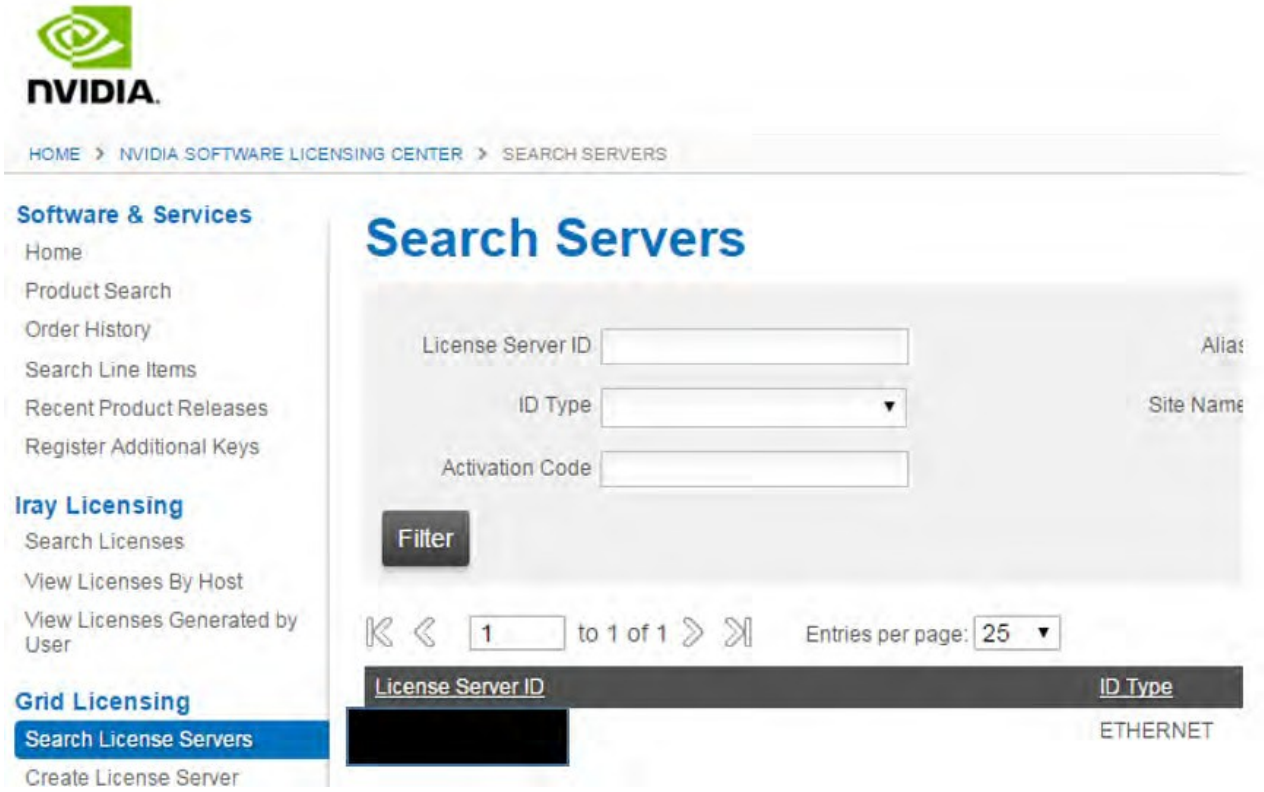
1. Log in to the license server site with the credentials set up during the registration process at nvidia.com/gridevaL. A license file is generated from https://nvidia.flexnetoperations.com.
2. After you are logged in, click Create License Server.
3. Specify the fields as shown in Figure 29. In the License Server ID field, enter the MAC address of your local license server's NIC. Leave ID Type set to Ethernet. For Alias and Site Name, choose user-friendly names. Then click Create.

**Figure 29.** Creating the license server



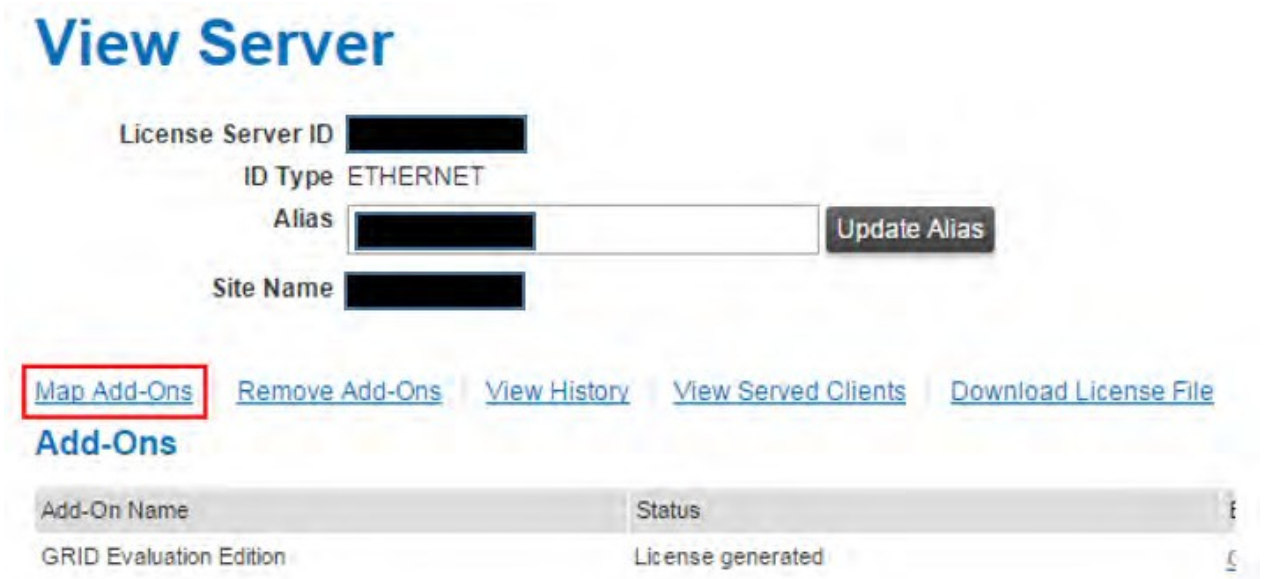

4. Click the Search License Servers node.
5. Click your license server ID (Figure 30).

**Figure 30.** Selecting the license server ID



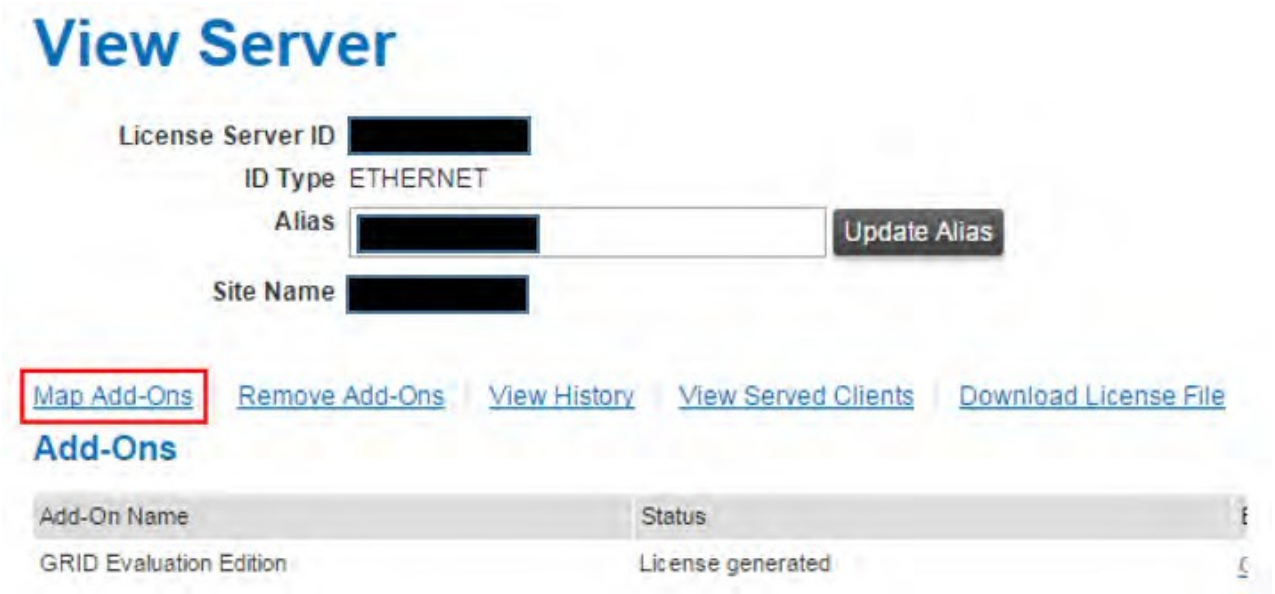

6. Click Map Add-Ons and choose the number of license units out of your total pool to allocate to this license server (Figure 31).

**Figure 31.** Choosing the number of license units from the pool



After the add-ons are mapped, the interface will look like Figure 32, showing 128 units mapped, for example.

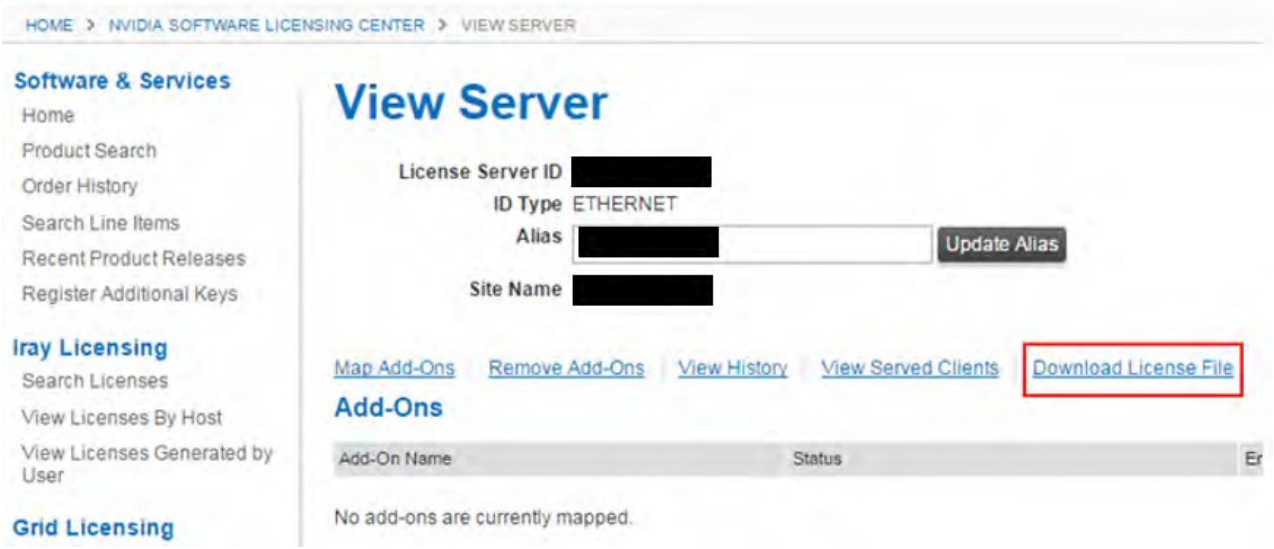
**Figure 32.** View Server page after the add-ons are mapped




7. Click Download License File and save the .bin file to your license server (Figure 33).

**Note:** The .bin file must be uploaded to your local license server within 24 hours of its generation. Otherwise, you will need to generate a new .bin file.

**Figure 33.** Saving the .bin file

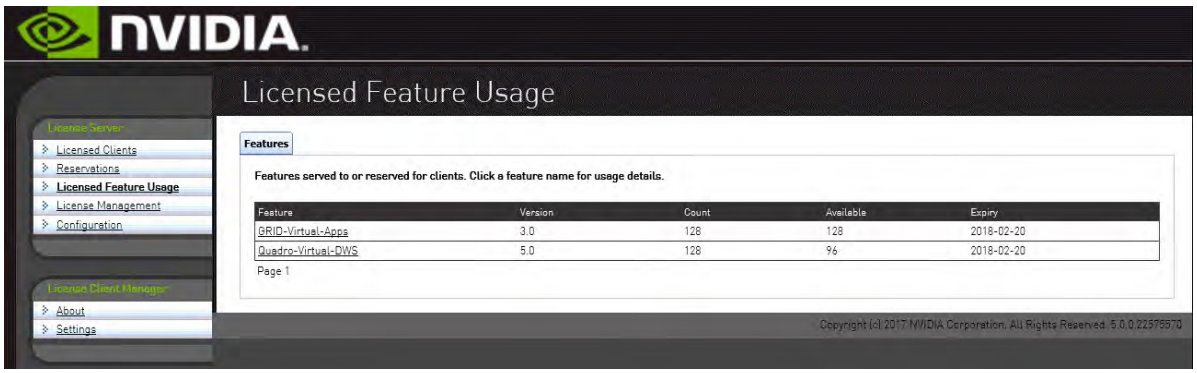



8. On the local license server, browse to http://<FODN>:8080/licserver to display the License Server Configuration page.
9. Click License Management in the left pane.
10. Click Browse to locate your recently download .bin license file. Select the .bin file and click OK.

11. Click Upload. The message “Successfully applied license file to license server” should appear on the screen (Figure 34). The features are available (Figure 35).

**Figure 34.** License file successfully applied

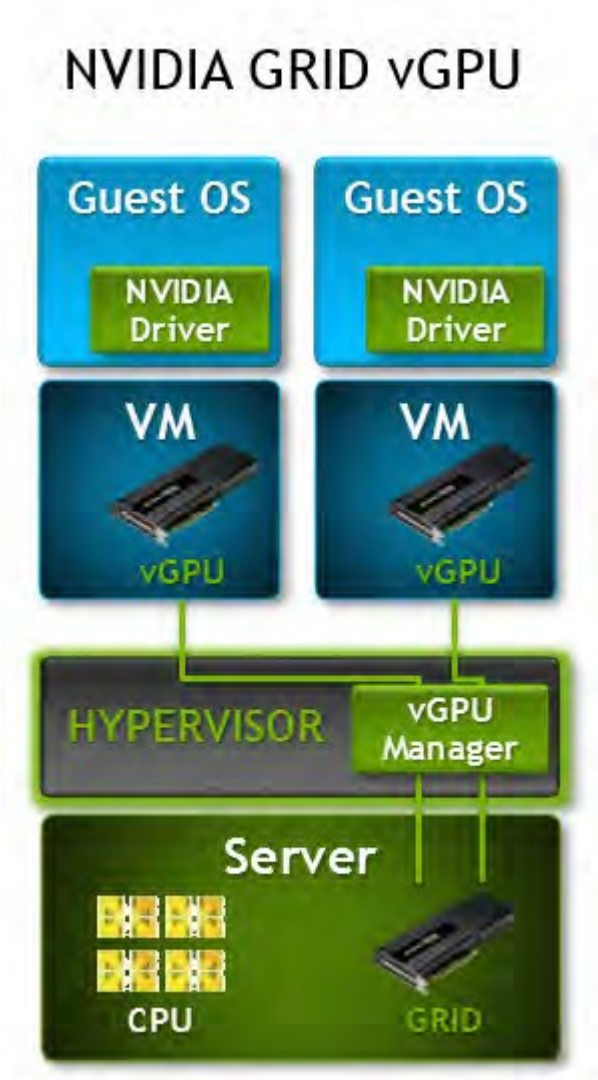

**Figure 35.** NVIDIA license server with features available for use




## Install NVIDIA GRID software on the VMware ESX host and Microsoft Windows virtual machine

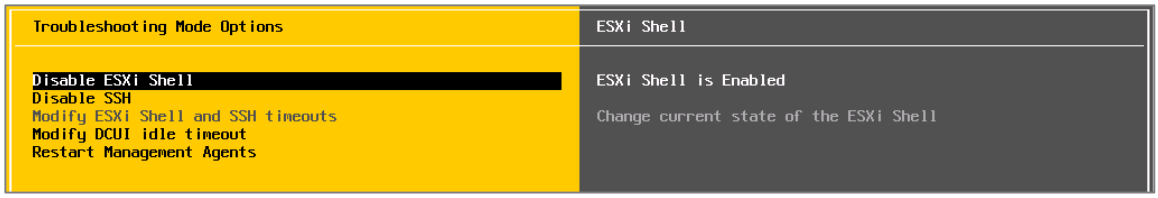
This section summarizes the installation process for configuring an ESXi host and virtual machine for vGPU support. Figure 36 shows the components used for vGPU support.

**Figure 36.** NVIDIA GRID vGPU components




1. Download the NVIDIA GRID GPU driver pack for VMware vSphere ESXi 6.5.
2. Enable the ESXi shell and the Secure Shell (SSH) protocol on the vSphere host from the Troubleshooting Mode Options menu of the vSphere Configuration Console (Figure 37).

**Figure 37.** VMware ESXi configuration console



3. Upload the NVIDIA driver (vSphere Installation Bundle [VIB] file) to the /tmp directory on the ESXi host using a tool such as WinSCP. (Shared storage is preferred if you are installing drivers on multiple servers or using the VMware Update Manager.)
4. Log in as root to the vSphere console through SSH using a tool such as Putty.

**Note:** The ESXi host must be in maintenance mode for you to install the VIB module. To place the host in maintenance mode, use the command esxcli system maintenanceMode set -enable true.

5. Enter the following command to install the NVIDIA vGPU drivers:

   esxcli software vib install --no-sig-check -v /<path>/<filename>.VIB

The command should return output similar to that shown here:

```
# esxcli software vib install --no-sig-check -v /tmp/NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-
1OEM.650.0.0.4598673.vib
Installation Result
   Message: Operation finished successfully.
   Reboot Required: false
   VIBs Installed: NVIDIA_bootbank_NVIDIA-VMware_ESXi_6.5_Host_Driver_384.99-
1OEM.650.0.0.4598673
   VIBs Removed:
   VIBs Skipped:
```

**Note:** Although the display shows "Reboot Required: false," a reboot is necessary for the VIB file to load and for xorg to start.

6. Exit the ESXi host from maintenance mode and reboot the host by using the vSphere Web Client or by entering the following commands:

   **#esxcli system maintenanceMode set -e false**
   **#reboot**

7. After the host reboots successfully, verify that the kernel module has loaded successfully using the following command:

   **esxcli software vib list | grep -i nvidia**

The command should return output similar to that shown here:

```
# esxcli software vib list | grep -i nvidia
NVIDIA-VMware_ESXi_6.5_Host_Driver  384.99-1OEM.650.0.0.4598673          NVIDIA
VMwareAccepted    2017-11-27
```

**Note:** See the VMware knowledge base article for information about removing any existing NVIDIA drivers before installing new drivers: http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434.

8. Confirm GRID GPU detection on the ESXi host. To determine the status of the GPU card's CPU, the card's memory, and the amount of disk space remaining on the card, enter the following command:

   **nvidia-smi**

   The command should return output similar to that shown in Figure 38, 39, or 40, depending on the card used in your environment.

**Figure 38.** VMware ESX SSH console report for GPU P40 card detection on Cisco UCS C240 M5 Rack Server

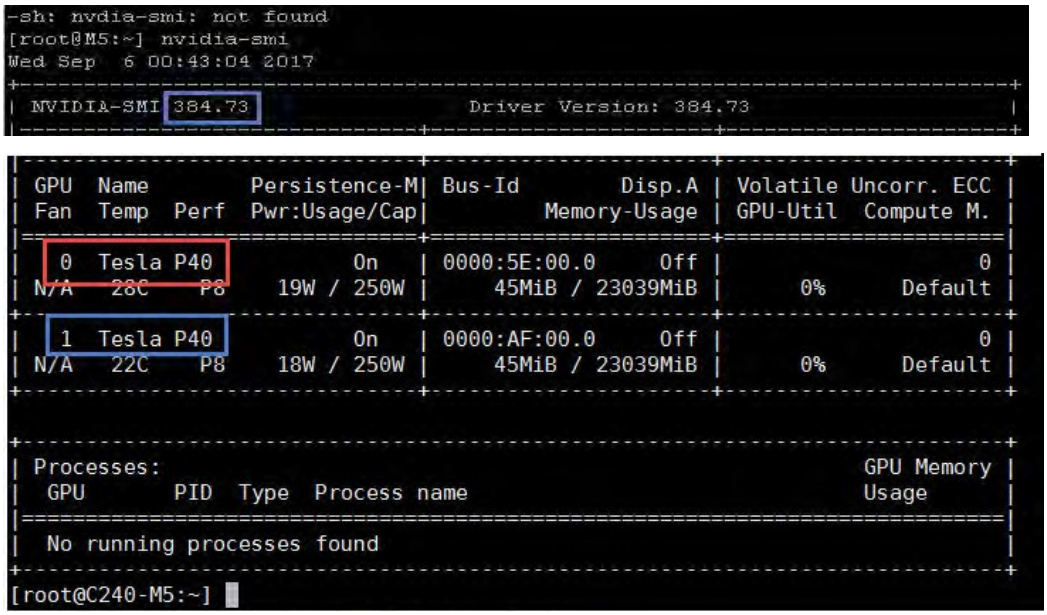
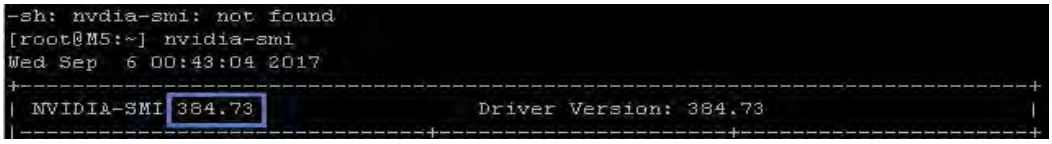



**Figure 39.** VMware ESX SSH console report for GPU M10 card detection on Cisco UCS C240 M5 Rack Server

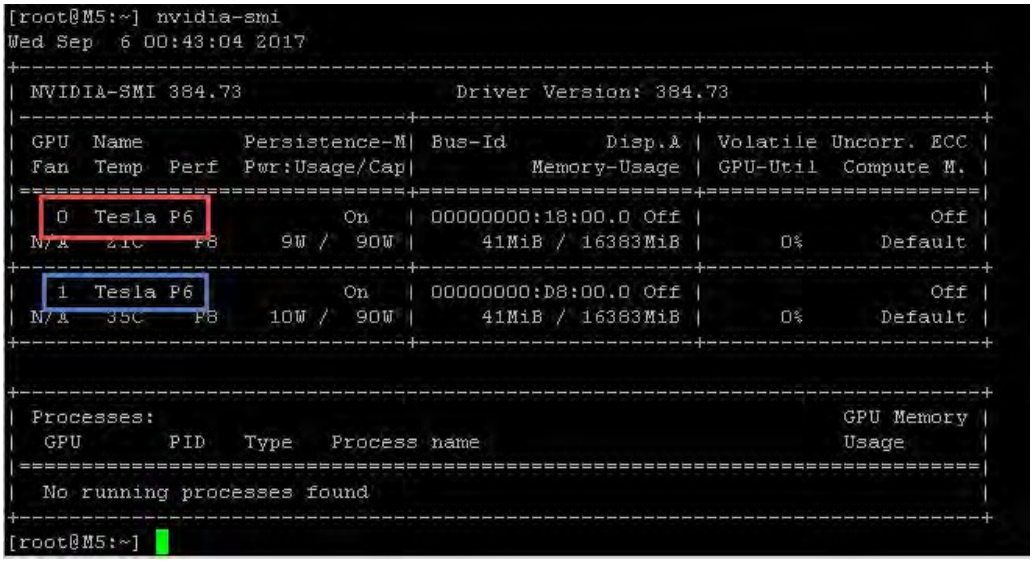


```
+-----------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|=============================+======================+======================|
|   0  Tesla M10           On  | 0000:60:00.0     Off |                  N/A |
| N/A   29C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   1  Tesla M10           On  | 0000:61:00.0     Off |                  N/A |
| N/A   30C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   2  Tesla M10           On  | 0000:62:00.0     Off |                  N/A |
| N/A   26C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   3  Tesla M10           On  | 0000:63:00.0     Off |                  N/A |
| N/A   26C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   4  Tesla M10           On  | 0000:88:00.0     Off |                  N/A |
| N/A   27C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   5  Tesla M10           On  | 0000:89:00.0     Off |                  N/A |
| N/A   28C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   6  Tesla M10           On  | 0000:8A:00.0     Off |                  N/A |
| N/A   25C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+
|   7  Tesla M10           On  | 0000:8B:00.0     Off |                  N/A |
| N/A   24C    P8    10W /  53W |      18MiB /  8191MiB |      0%      Default |
+-----------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

**Figure 40.** VMware ESX SSH console report for GPU P6 card detection on Cisco UCs B200 M5 Blade Server

```
-sh: nvdia-smi: not found
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```




**Note:** The NVIDIA system management interface (SMI) also allows GPU monitoring using the following command: nvidia-smi –l (this command adds a loop, automatically refreshing the display).

### NVIDIA Tesla P6, P40, and M10 profile specifications

The Tesla P6 and P40 cards each have a single physical GPU, and the Tesla M10 card has multiple physical GPUs. Each physical GPU can support several different types of vGPU. Each type of vGPU has a fixed amount of frame buffer space, a fixed number of supported display heads, and a fixed maximum resolution, and each is targeted at a different class of workload. Table 6 lists the vGPU types supported by GRID GPUs.

For more information, see http://www.nvidia.com/object/grid-enterprise-resources.html.

**Table 6.** NVIDIA GRID 5 user profile specifications for NVIDIA Tesla cards

| End-user profile | GRID<br>Virtual App (vApp) profiles | GRID<br>Virtual PC (vPC) profiles | Quadro<br>Virtual Datacenter Workstation (vDWS) profiles |
|---|---|---|---|
| 1 GB | • P6-1A<br>• M10-1A<br>• P40-1A | P6-1B<br>M10-1B<br>P40-1B | • P6-1Q<br>• M10-1Q<br>• P40-1Q |
| 2 GB | • P6-2A<br>• M10-2A<br>• P40-2A | – | • P6-2Q<br>• M10-2Q<br>• P40-2Q |
| 3 GB | P40-3A | – | P40-3Q |
| 4 GB | • P6-4A<br>• M10-4A<br>• P40-4A | – | P6-4Q<br>M10-4Q<br>P40-4Q |
| 6 GB | P40-6A | – | P40-6Q |
| 8 GB | • P6-8A<br>• M10-8A<br>• P40-8A | – | P6-8Q<br>M10-8Q<br>P40-8Q |

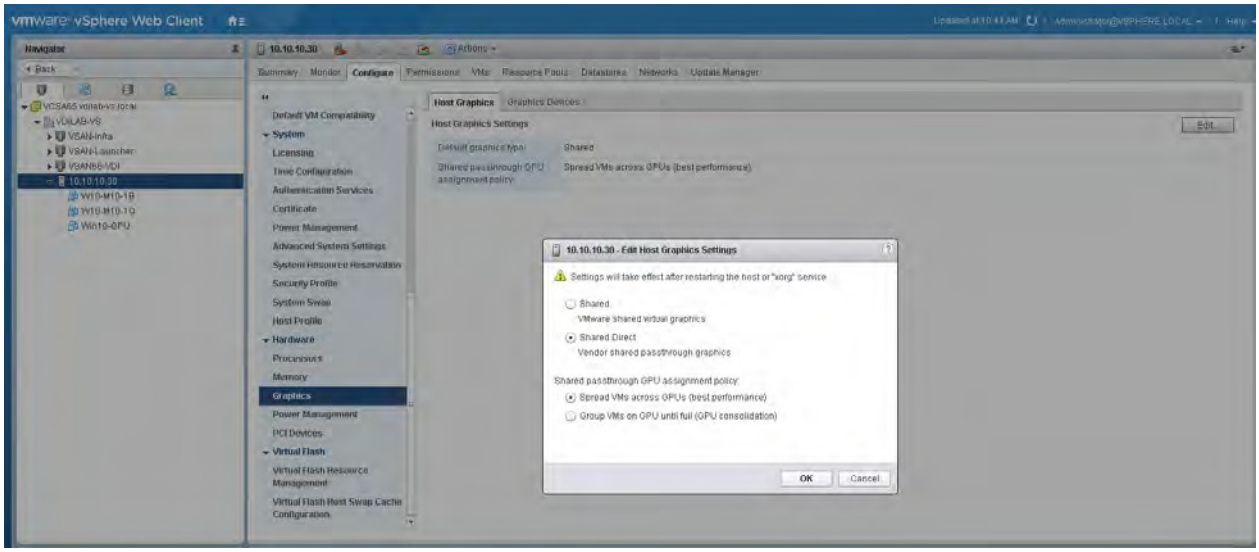
| End-user profile | GRID<br>Virtual App (vApp) profiles | GRID<br>Virtual PC (vPC) profiles | Quadro<br>Virtual Datacenter Workstation (vDWS) profiles |
|---|---|---|---|
| 12 GB | P40-12A | _ | P40-12Q |
| 16 GB | P6-16A | _ | P6-16Q |
| 24 GB | P40-24A | _ | P40-24Q |
| Pass-through | _ | _ | _ |

### Prepare a virtual machine for vGPU support

Use the following procedure to create the virtual machine that will later be used as the VDI base image.

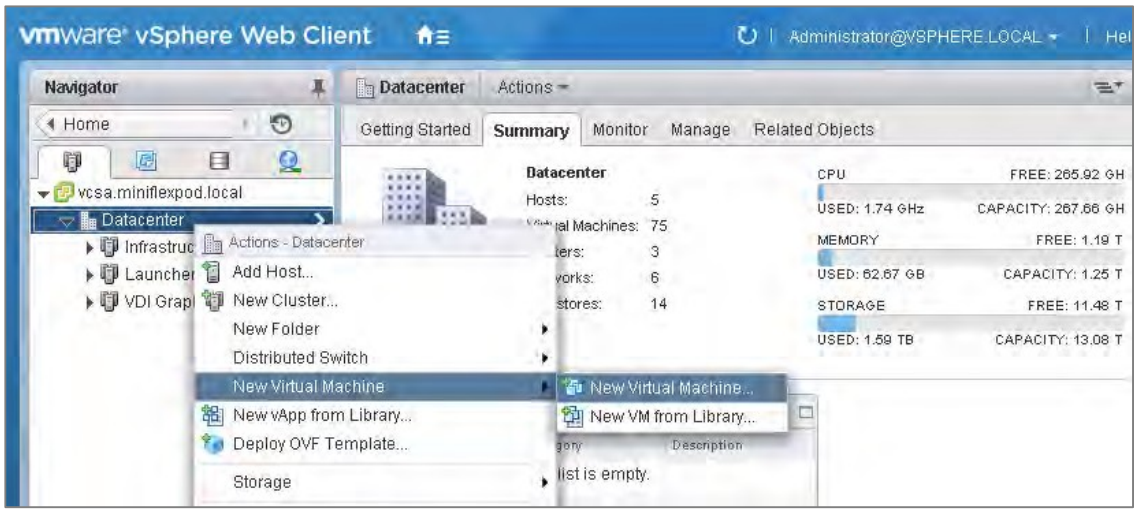
1. Select the ESXi host and click the Configure tab. From the list of options at the left, choose Graphics > Edit Host Graphics Settings. Select Shared Direct "Vendor shared passthrough graphics" (Figure 41). Reboot the system to make the changes effective.

**Figure 41.** Edit Host Graphics Settings window



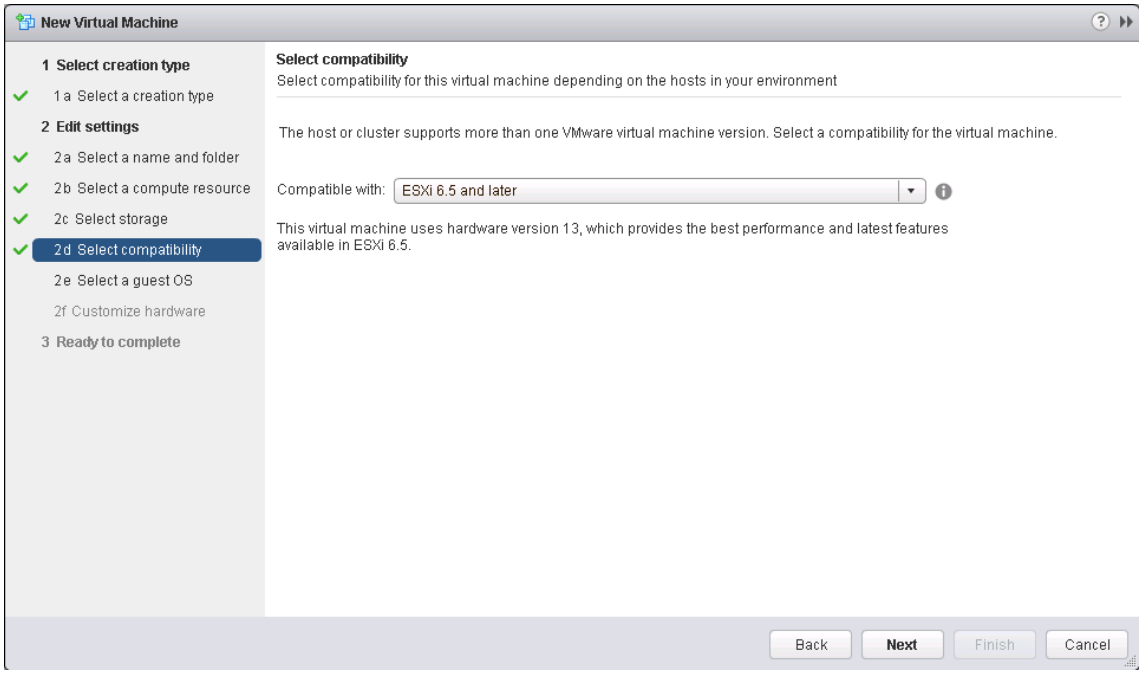

2. Using the vSphere Web Client, create a new virtual machine. To do this, right-click a host or cluster and choose New Virtual Machine. Work through the New Virtual Machine wizard. Unless another configuration is specified, select the configuration settings appropriate for your environment (Figure 42).

**Figure 42.** Creating a new virtual machine in VMware vSphere Web Client




3. Choose “ESXi 6.0 and later” from the “Compatible with” drop-down menu to use the latest features, including the mapping of shared PCI devices, which is required for the vGPU feature (Figure 43). This document uses “ESXi 6.5 and later,” which provides the latest features available in ESXi 6.5 and virtual machine hardware Version 13.

**Figure 43.** Selecting virtual machine hardware Version 11 or later




4. To customize the hardware of the new virtual machine, add a new shared PCI device, select the appropriate GPU profile, and reserve all virtual machine memory (Figures 44 and 45).

**Note:** If you are creating a new virtual machine and using the vSphere Web Client's virtual machine console functions, the mouse will not be usable in the virtual machine until after both the operating system and VMware Tools have been installed. If you cannot use the traditional vSphere Web Client to connect to the virtual machine, do not enable the NVIDIA GRID vGPU at this time.

**Figure 44.** Adding a shared PCI device to the virtual machine to attach the GPU profile

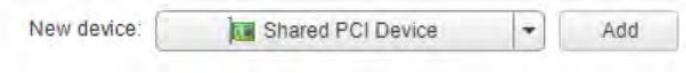
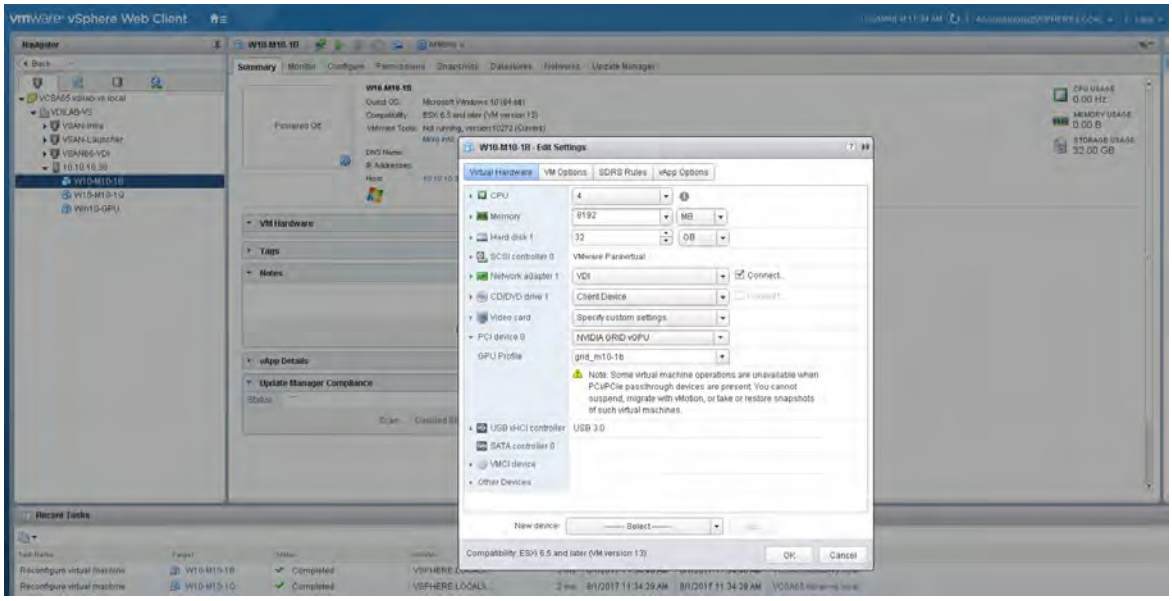



**Figure 45.** Attaching the GPU profile to a shared PCI device

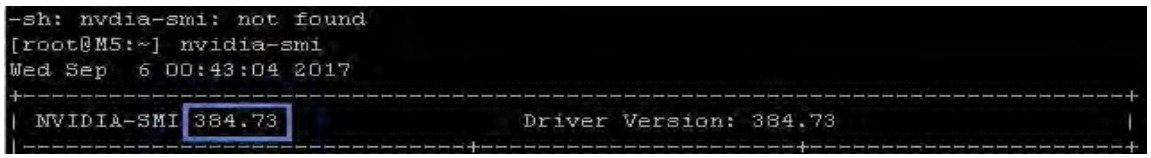


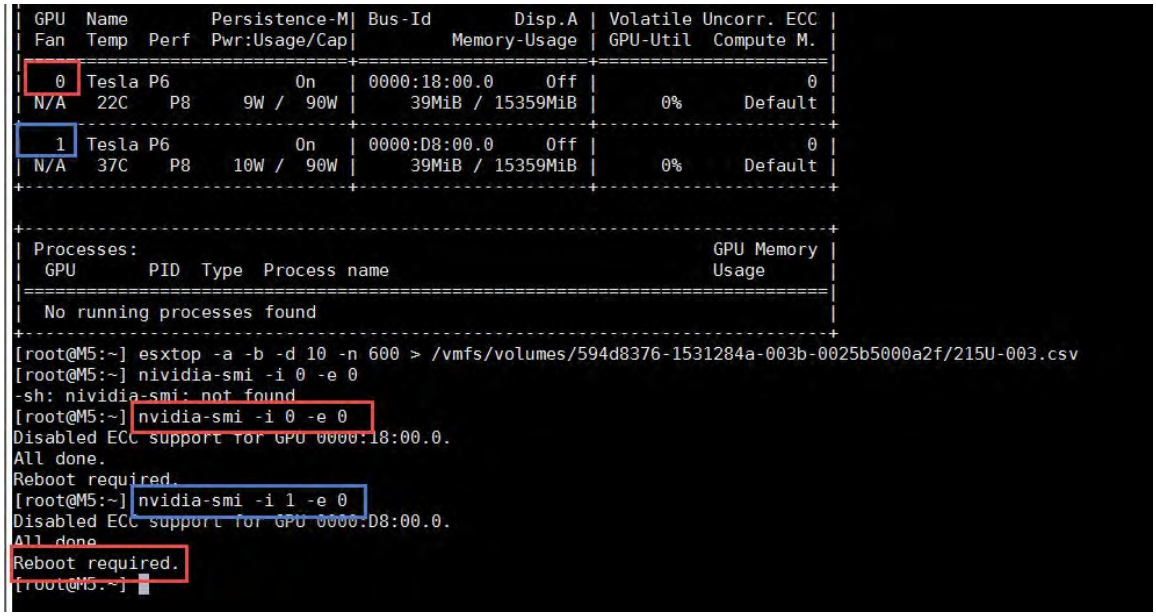

5. A virtual machine with a vGPU assigned will not start if ECC is enabled. If this is the case, as a workaround disable ECC by entering the following commands (Figure 46):

```
# nvidia-smi -i 0 -e 0
# nvidia-smi -i 1 -e 0
```

**Note:** Use -i to target a specific GPU. If two cards are installed in a server, run the command twice as shown in the example here, where 0 and 1 each specify a GPU card.

**Figure 46.** Disabling ECC

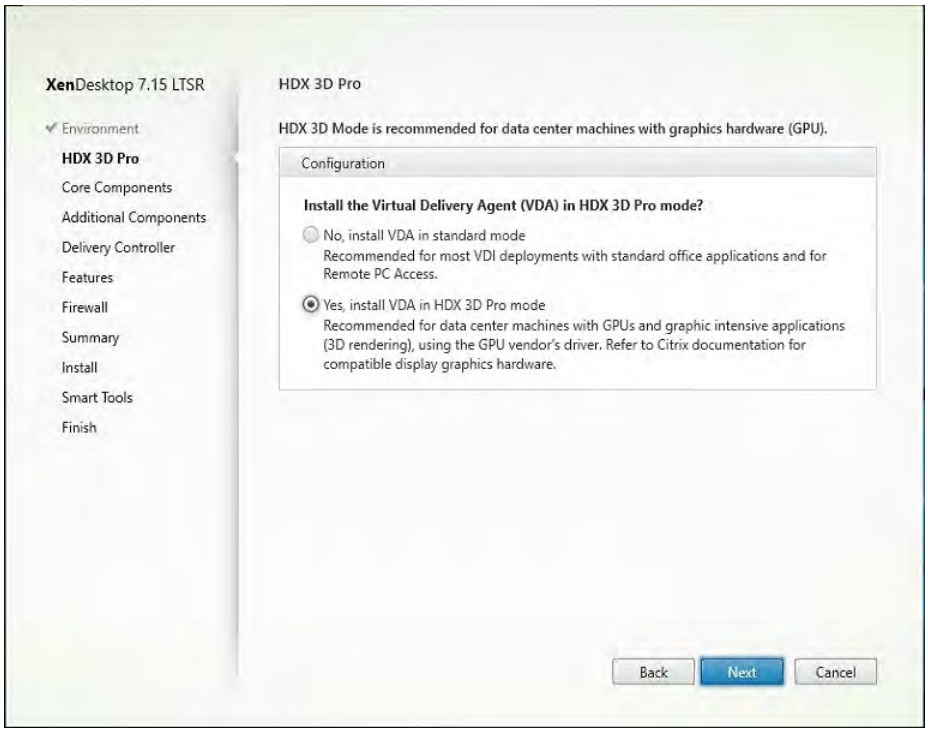





6. Install and configure Microsoft Windows on the virtual machine:
   a. Configure the virtual machine with the appropriate amount of vCPU and RAM according to the GPU profile selected.
   b. Install VMware Tools.
   c. Join the virtual machine to the Microsoft Active Directory domain.
   d. Install or upgrade Citrix HDX 3D Pro Virtual Desktop Agent.

   - When you use the installer's GUI to install a VDA for a Windows desktop, select Yes on the HDX 3D Pro page (Figure 47).

**Figure 47.** Selecting HDX 3D Pro during VDA installation




When you use the command-line interface (CLI) to install the VDA, include the /enable_hdx_3d_pro option with the XenDesktop VdaSetup.exe command.

- To upgrade HDX 3D Pro, uninstall both the separate HDX 3D for Professional Graphics component and the VDA before installing the VDA for HDX 3D Pro. Similarly, to switch from the standard VDA for a Windows desktop to the HDX 3D Pro VDA, uninstall the standard VDA and then install the VDA for HDX 3D Pro.

e. Optimize the Windows OS. VMware OSOT, the optimization tool, includes customizable templates to enable or disable Windows system services and features using LoginVSI recommendations and best practices across multiple systems. Because most Windows system services are enabled by default, the optimization tool can be used to easily disable unnecessary services and features to improve performance.

**Note:** VMware OSOT b1090 with the Windows 10 – LoginVSI template was used for the purposes of this document.

Table 7 shows differences in applied optimizations in the master image used for vGPU-enabled desktops.

**Table 7.** Optimization differences

| Optimization | Description | No GPU | vGPU |
|---|---|---|---|
| Software Rendering Internet Explorer | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Hardware Acceleration Office 14 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Hardware Acceleration Office 15 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |

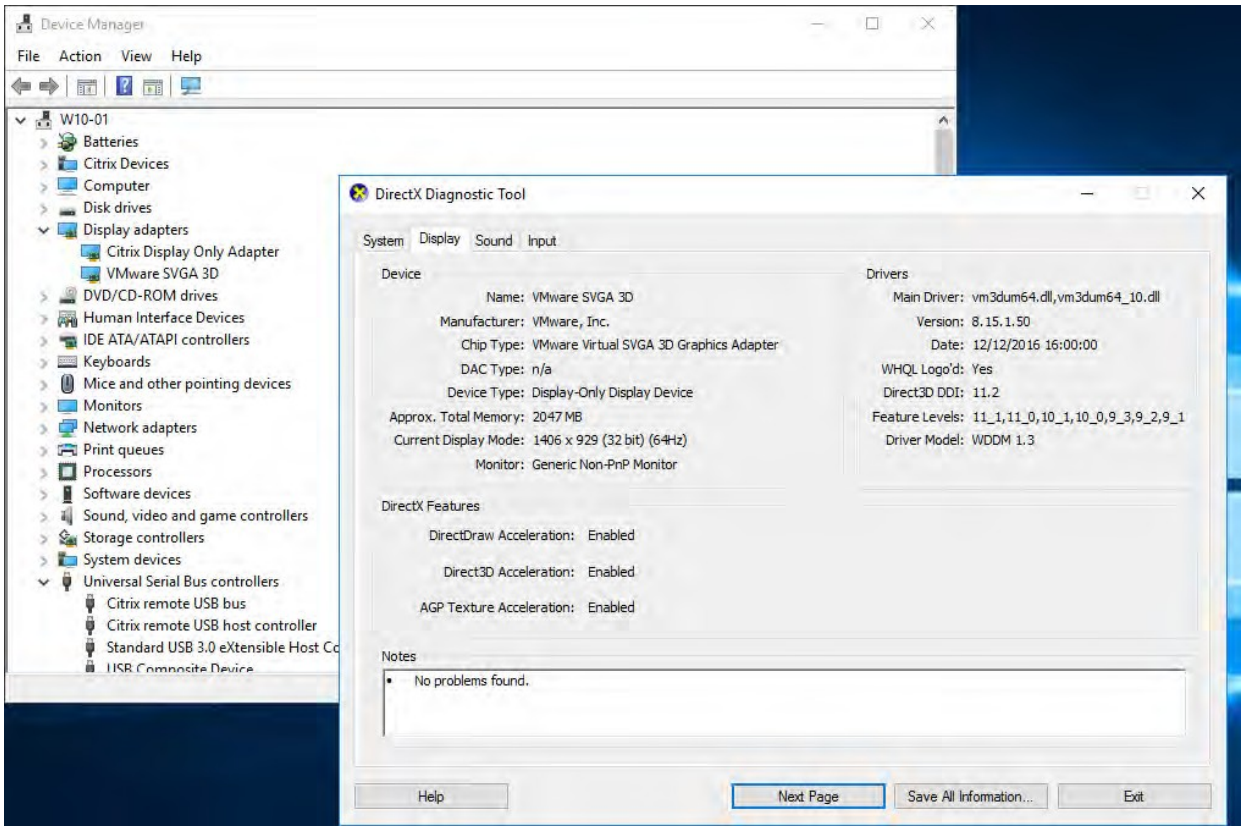
| Optimization | Description | No GPU | vGPU |
|---|---|---|---|
| Disable Animations Office 15 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Hardware Acceleration Office 16 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |
| Disable Animations Office 16 | Uses software rendering instead of hardware rendering. Enable this option if no vGPU is present. | Applied | Not applied |

### Install the NVIDIA vGPU software driver

Use the following procedure to install the NVIDIA GRID vGPU drivers on the desktop virtual machine. To fully enable vGPU operation, the NVIDIA driver must be installed.

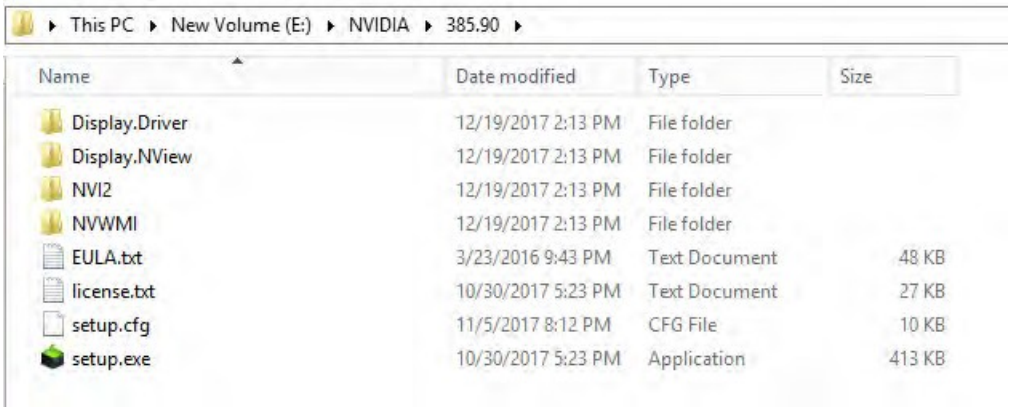
Before the NVIDIA driver is installed on the guest virtual machine, the Device Manager shows the standard VGA graphics adapter (Figure 48).

**Figure 48.** Device Manager before the NVIDIA driver is installed




1. Copy the Microsoft Windows drivers from the NVIDIA GRID vGPU driver pack downloaded earlier to the master virtual machine.
2. Copy the 32- or 64-bit NVIDIA Windows driver from the vGPU driver pack to the desktop virtual machine and run setup.exe (Figure 49).

**Figure 49.** NVIDIA driver pack




**Note:** The vGPU host driver and guest driver versions need to match. Do not attempt to use a newer guest driver with an older vGPU host driver or an older guest driver with a newer vGPU host driver. In addition, the vGPU driver from NVIDIA is a different driver than the GPU pass-through driver.

3. Agree to the NVIDIA software license (Figure 50).

**Figure 50.** Agreeing to the NVIDIA software license

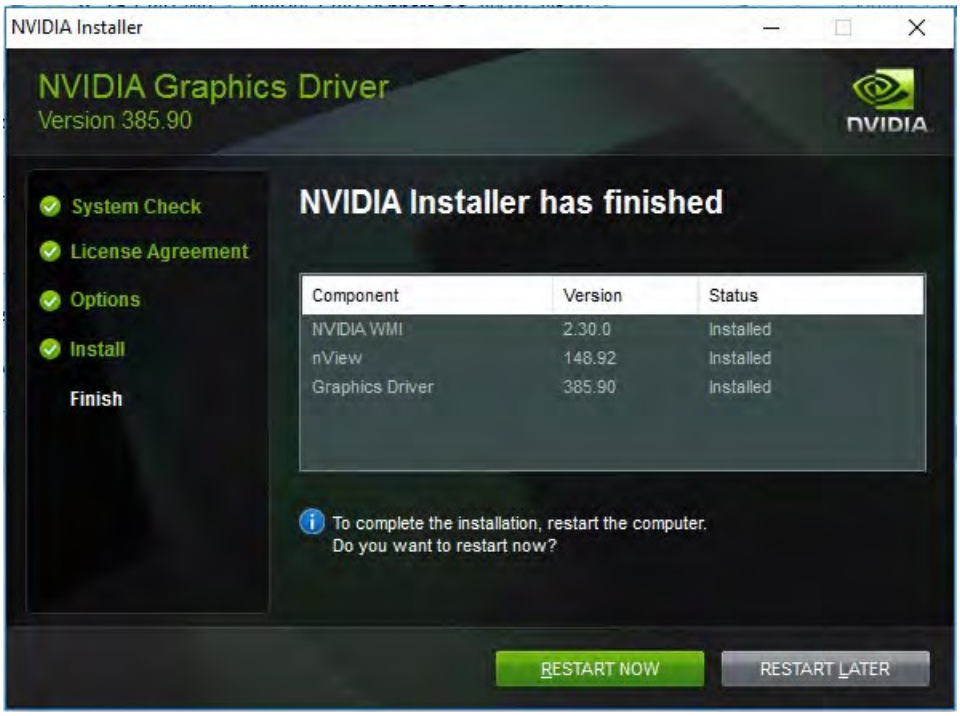

4. Install the graphics drivers using the Express or Custom option (Figures 51 and 52). After the installation has completed successfully, restart the virtual machine (Figure 53).

**Note:** Be sure that remote desktop connections are enabled. After this step, console access may not be available for the virtual machine when you connect from a vSphere Client.

**Figure 51.** Selecting the Express or Custom installation option


**Figure 52.** Components to be installed during NVIDIA graphics driver custom installation process

**Figure 53.** Resarting the virtual machine

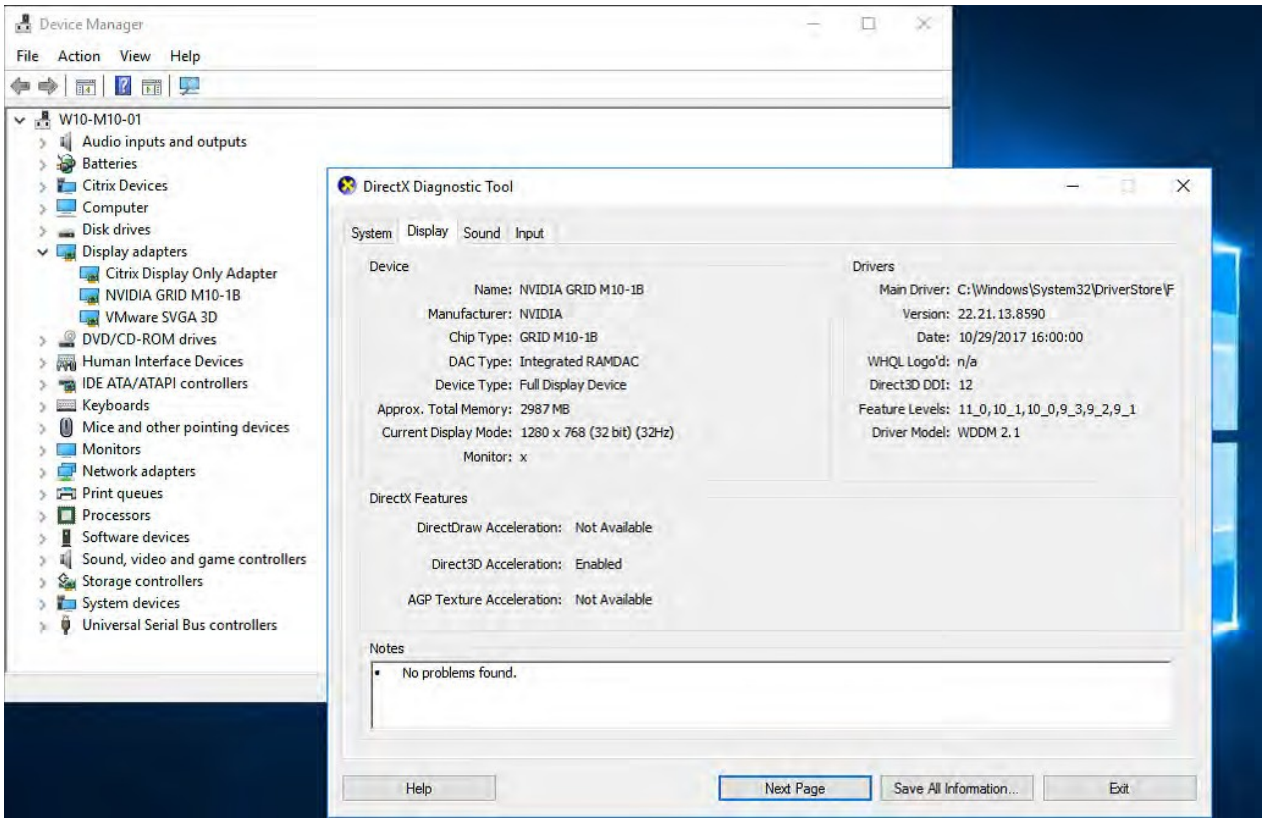



### Verify that applications are ready to support the vGPU

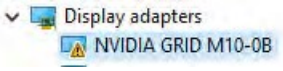
Verify the successful installation of the graphics drivers and the vGPU device.

Open Windows Device Manager and expand the Display Adapter section. The device will reflect your chosen profile (Figure 54).

**Figure 54.** Verifying the driver installation




**Note:** If you see an exclamation point as shown here, a problem has occurred.




The following are the most likely the reasons:

- The GPU driver service is not running.
- The GPU driver is incompatible.

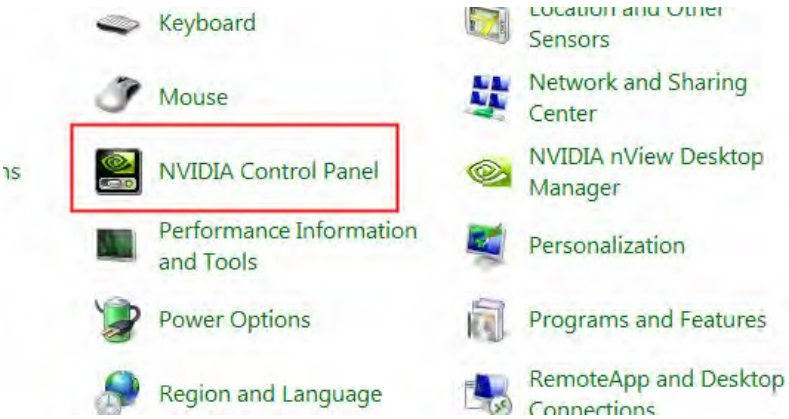
### Configure the virtual machine for an NVIDIA GRID vGPU license

You need to point the master image to the license server so the virtual machines with vGPUs can obtain a license.

**Note:** The license settings persist across reboots. These settings can also be preloaded through registry keys.

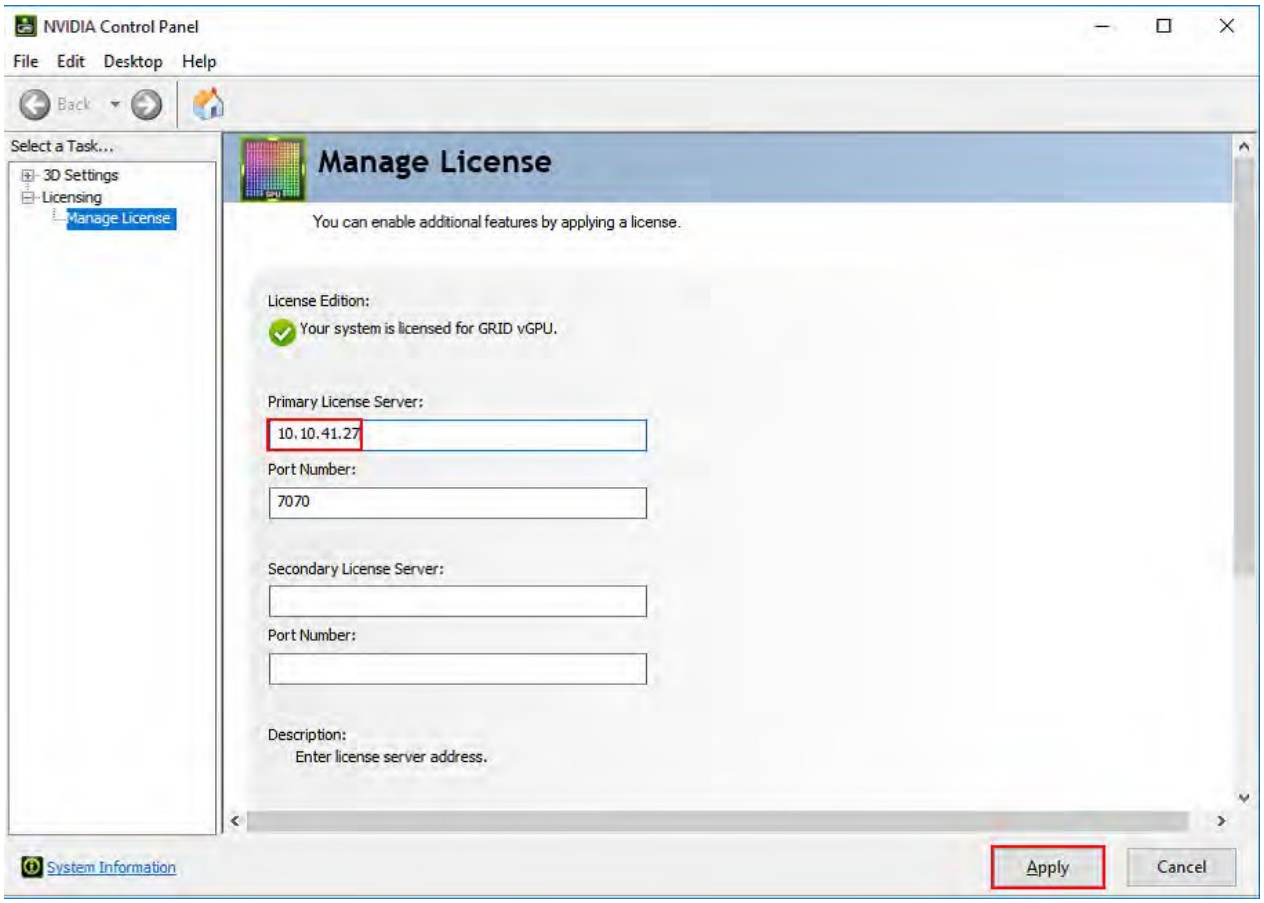
1. In the Microsoft Windows Control Panel, double-click NVIDIA Control Panel (Figure 55).

**Figure 55.** Choosing NVIDIA Control Panel




2. Select Manage License from the left pane and enter your license server address and port. Click Apply (Figure 56).

**Figure 56.** Managing your license.



## Verify vGPU deployment

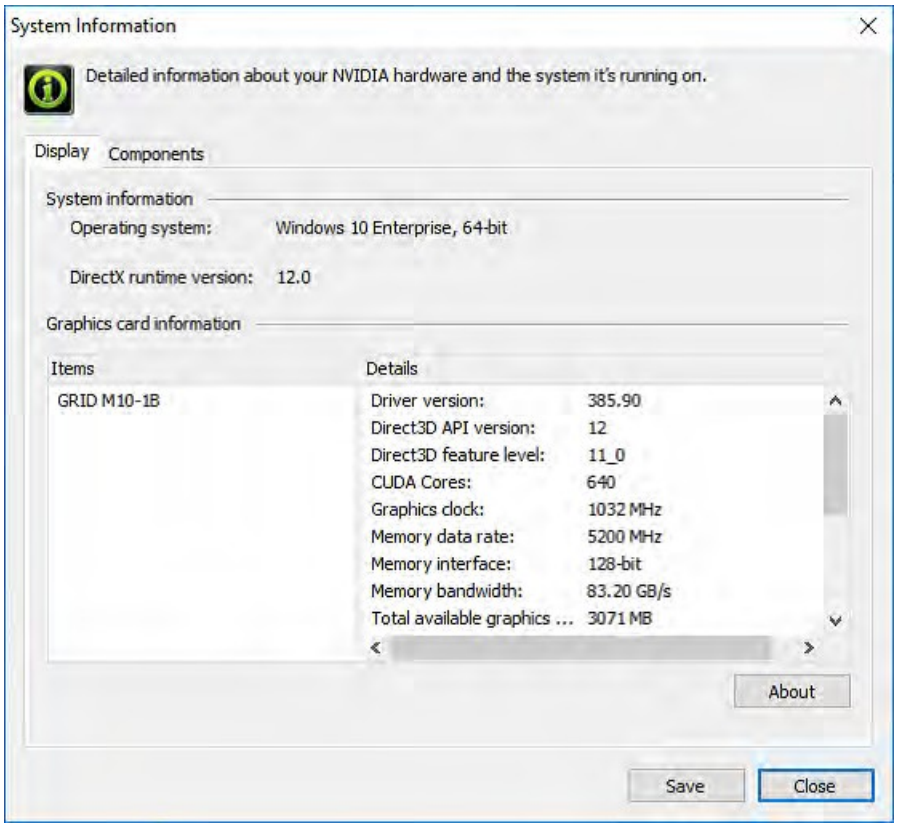
After the desktops are provisioned, use the following steps to verify vGPU deployment in the Citrix XenDesktop environment.

### Verify that the NVIDIA driver is running on the desktop

Follow these steps to verify that the NVIDIA driver is running on the desktop:

1. Right-click the desktop. In the menu, choose NVIDIA Control Panel to open the control panel.
2. In the control panel, select System Information to see the vGPU that the virtual machine is using, the vGPU's capabilities, and the NVIDIA driver version that is loaded (Figure 57).

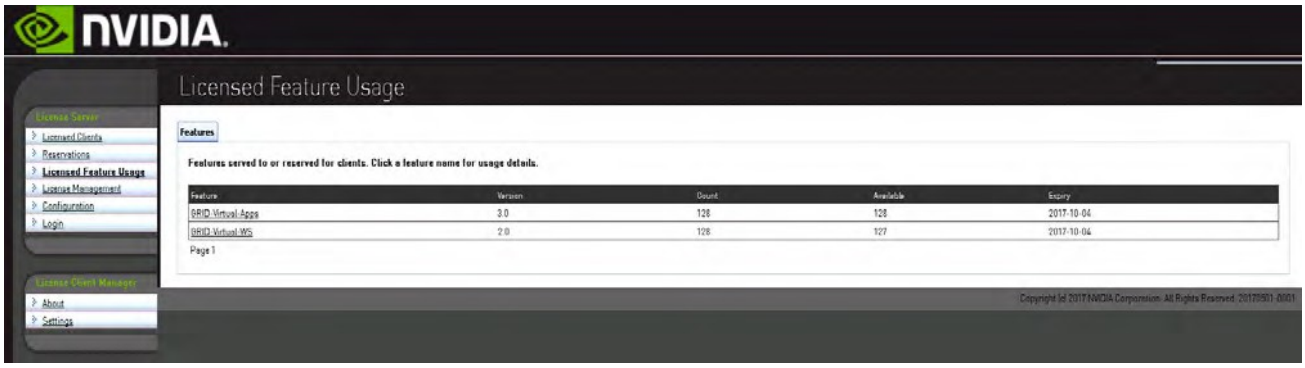
**Figure 57.** NVIDIA Control Panel System Information window




### Verify NVIDIA license acquisition by desktops

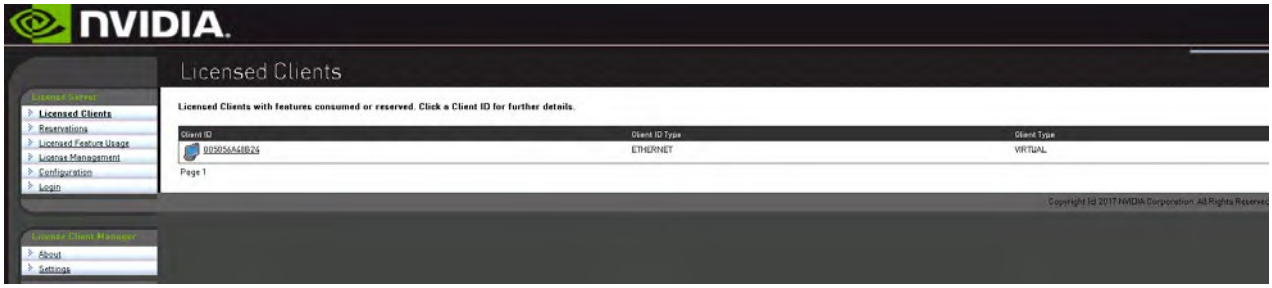
A license is obtained before the user logs on to the virtual machine after the virtual machine is fully booted (Figure 58).

**Figure 58.** NVIDIA License Server: Licensed Feature Usage window




To view the details, select Licensed Clients in the left pane (Figure 59).

**Figure 59.** NVIDIA License Server: Licensed Clients window

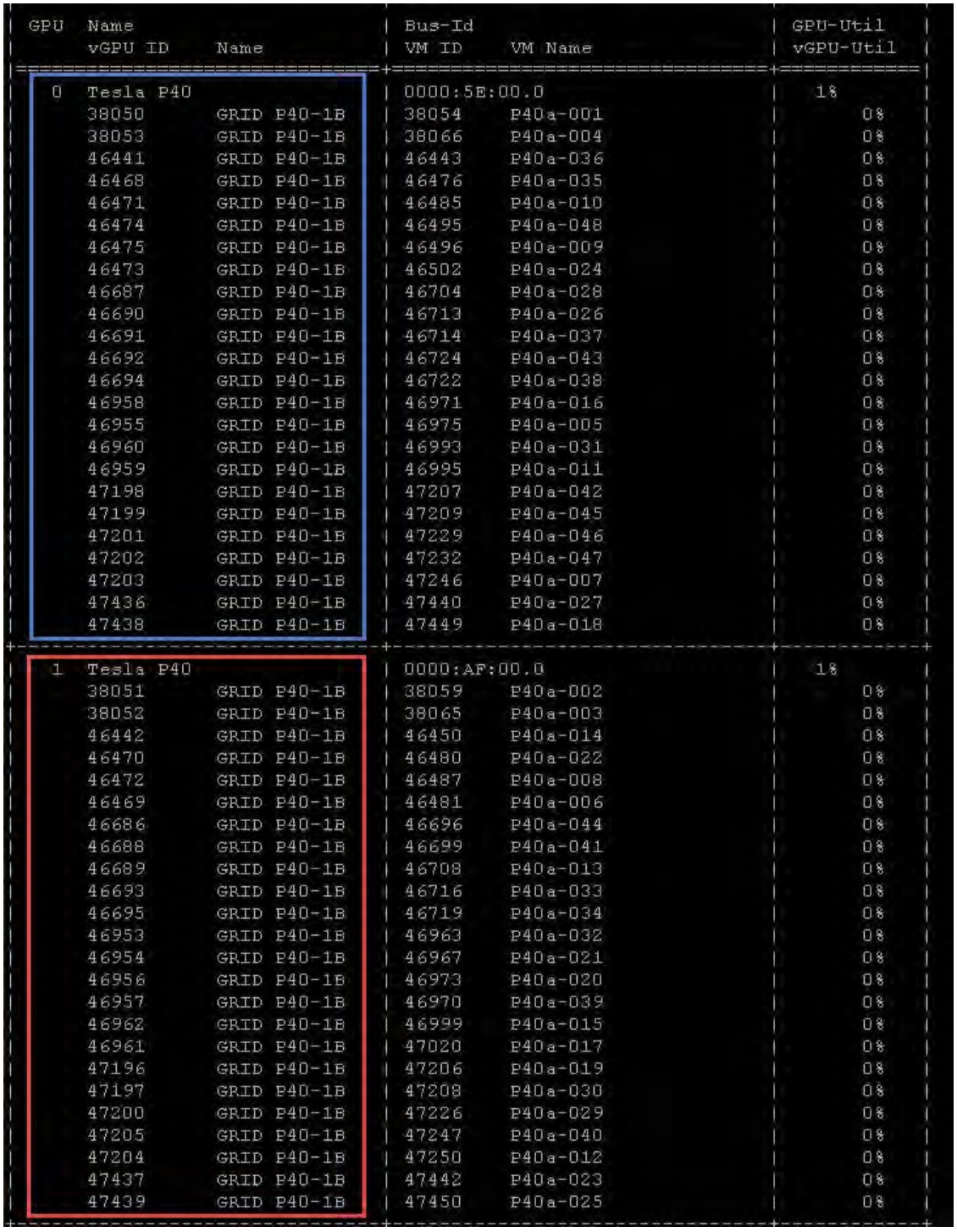



## Verify the NVIDIA configuration on the host

To obtain a hostwide overview of the NVIDIA GPUs, enter the nvidia-smi command without any arguments (Figures 60, 61, and 62).

**Figure 60.** The nvidia-smi command output from the host with two NVIDIA P40 cards and 48 Microsoft Windows 10 desktops with P40-1B vGPU profile

```
[root@M5:~] nvidia-smi
Wed Sep  6 00:43:04 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.73                 Driver Version: 384.73                    |
|-------------------------------+----------------------+----------------------+
```



| GPU | Name<br>vGPU ID | Name | Bus-Id<br>VM ID | VM Name | GPU-Util<br>vGPU-Util |
|---|---|---|---|---|---|
| 0 | Tesla P40 | | 0000:5E:00.0 | | 1% |
| | 38050 | GRID P40-1B | 38054 | P40a-001 | 0% |
| | 38053 | GRID P40-1B | 38066 | P40a-004 | 0% |
| | 46441 | GRID P40-1B | 46443 | P40a-036 | 0% |
| | 46468 | GRID P40-1B | 46476 | P40a-035 | 0% |
| | 46471 | GRID P40-1B | 46485 | P40a-010 | 0% |
| | 46474 | GRID P40-1B | 46495 | P40a-048 | 0% |
| | 46475 | GRID P40-1B | 46496 | P40a-009 | 0% |
| | 46473 | GRID P40-1B | 46502 | P40a-024 | 0% |
| | 46687 | GRID P40-1B | 46704 | P40a-028 | 0% |
| | 46690 | GRID P40-1B | 46713 | P40a-026 | 0% |
| | 46691 | GRID P40-1B | 46714 | P40a-037 | 0% |
| | 46692 | GRID P40-1B | 46724 | P40a-043 | 0% |
| | 46694 | GRID P40-1B | 46722 | P40a-038 | 0% |
| | 46958 | GRID P40-1B | 46971 | P40a-016 | 0% |
| | 46955 | GRID P40-1B | 46975 | P40a-005 | 0% |
| | 46960 | GRID P40-1B | 46993 | P40a-031 | 0% |
| | 46959 | GRID P40-1B | 46995 | P40a-011 | 0% |
| | 47198 | GRID P40-1B | 47207 | P40a-042 | 0% |
| | 47199 | GRID P40-1B | 47209 | P40a-045 | 0% |
| | 47201 | GRID P40-1B | 47229 | P40a-046 | 0% |
| | 47202 | GRID P40-1B | 47232 | P40a-047 | 0% |
| | 47203 | GRID P40-1B | 47246 | P40a-007 | 0% |
| | 47436 | GRID P40-1B | 47440 | P40a-027 | 0% |
| | 47438 | GRID P40-1B | 47449 | P40a-018 | 0% |
| 1 | Tesla P40 | | 0000:AF:00.0 | | 1% |
| | 38051 | GRID P40-1B | 38059 | P40a-002 | 0% |
| | 38052 | GRID P40-1B | 38065 | P40a-003 | 0% |
| | 46442 | GRID P40-1B | 46450 | P40a-014 | 0% |
| | 46470 | GRID P40-1B | 46480 | P40a-022 | 0% |
| | 46472 | GRID P40-1B | 46487 | P40a-008 | 0% |
| | 46469 | GRID P40-1B | 46481 | P40a-006 | 0% |
| | 46686 | GRID P40-1B | 46696 | P40a-044 | 0% |
| | 46688 | GRID P40-1B | 46699 | P40a-041 | 0% |
| | 46689 | GRID P40-1B | 46708 | P40a-013 | 0% |
| | 46693 | GRID P40-1B | 46716 | P40a-033 | 0% |
| | 46695 | GRID P40-1B | 46719 | P40a-034 | 0% |
| | 46953 | GRID P40-1B | 46963 | P40a-032 | 0% |
| | 46954 | GRID P40-1B | 46967 | P40a-021 | 0% |
| | 46956 | GRID P40-1B | 46973 | P40a-020 | 0% |
| | 46957 | GRID P40-1B | 46970 | P40a-039 | 0% |
| | 46962 | GRID P40-1B | 46999 | P40a-015 | 0% |
| | 46961 | GRID P40-1B | 47020 | P40a-017 | 0% |
| | 47196 | GRID P40-1B | 47206 | P40a-019 | 0% |
| | 47197 | GRID P40-1B | 47208 | P40a-030 | 0% |
| | 47200 | GRID P40-1B | 47226 | P40a-029 | 0% |
| | 47205 | GRID P40-1B | 47247 | P40a-040 | 0% |
| | 47204 | GRID P40-1B | 47250 | P40a-012 | 0% |
| | 47437 | GRID P40-1B | 47442 | P40a-023 | 0% |
| | 47439 | GRID P40-1B | 47450 | P40a-025 | 0% |

**Figure 61.** The nvidia-smi command output from the host with two NVIDIA P6 cards and 32 Microsoft Windows 10 desktops with P6-1B vGPU profile

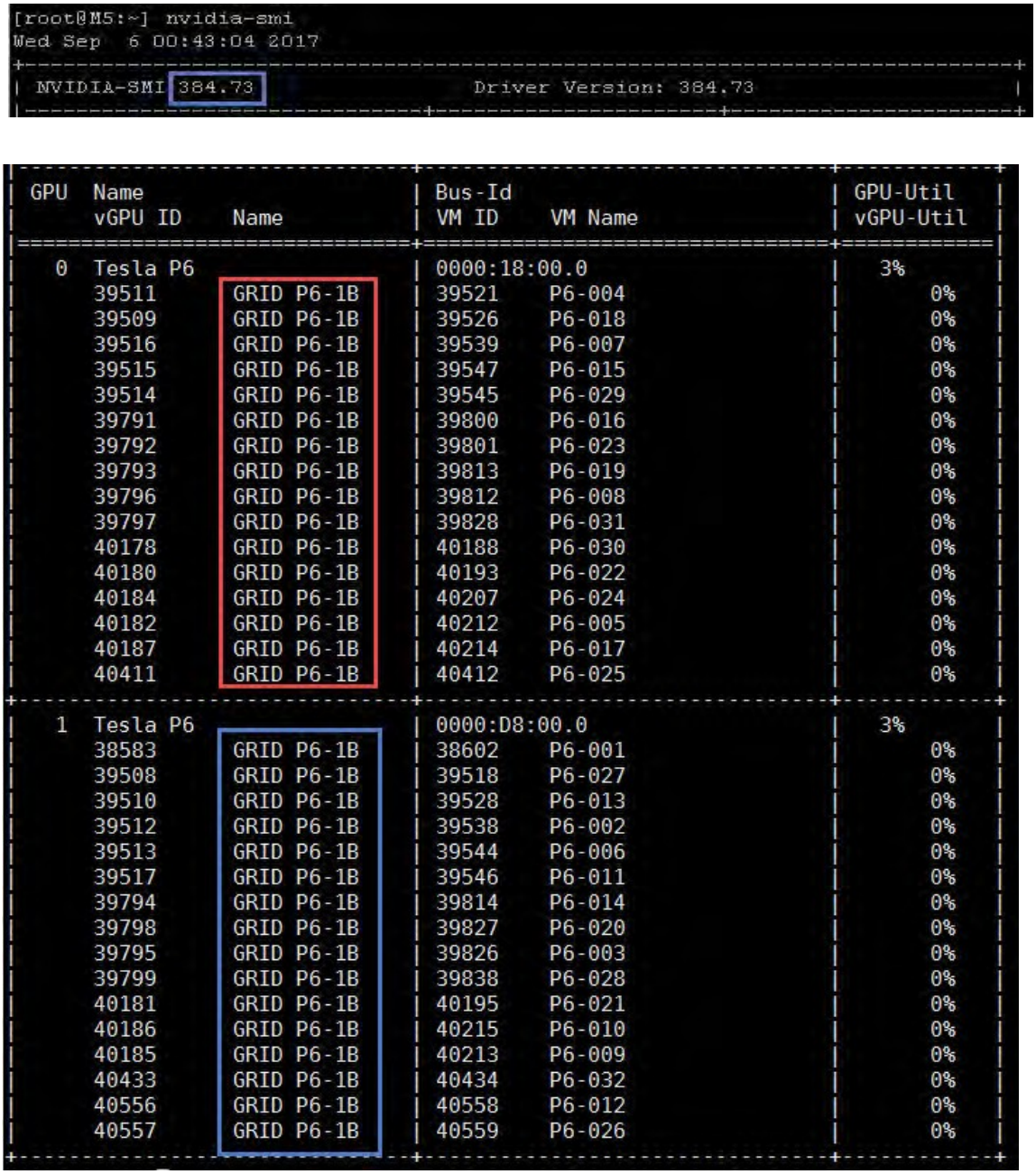
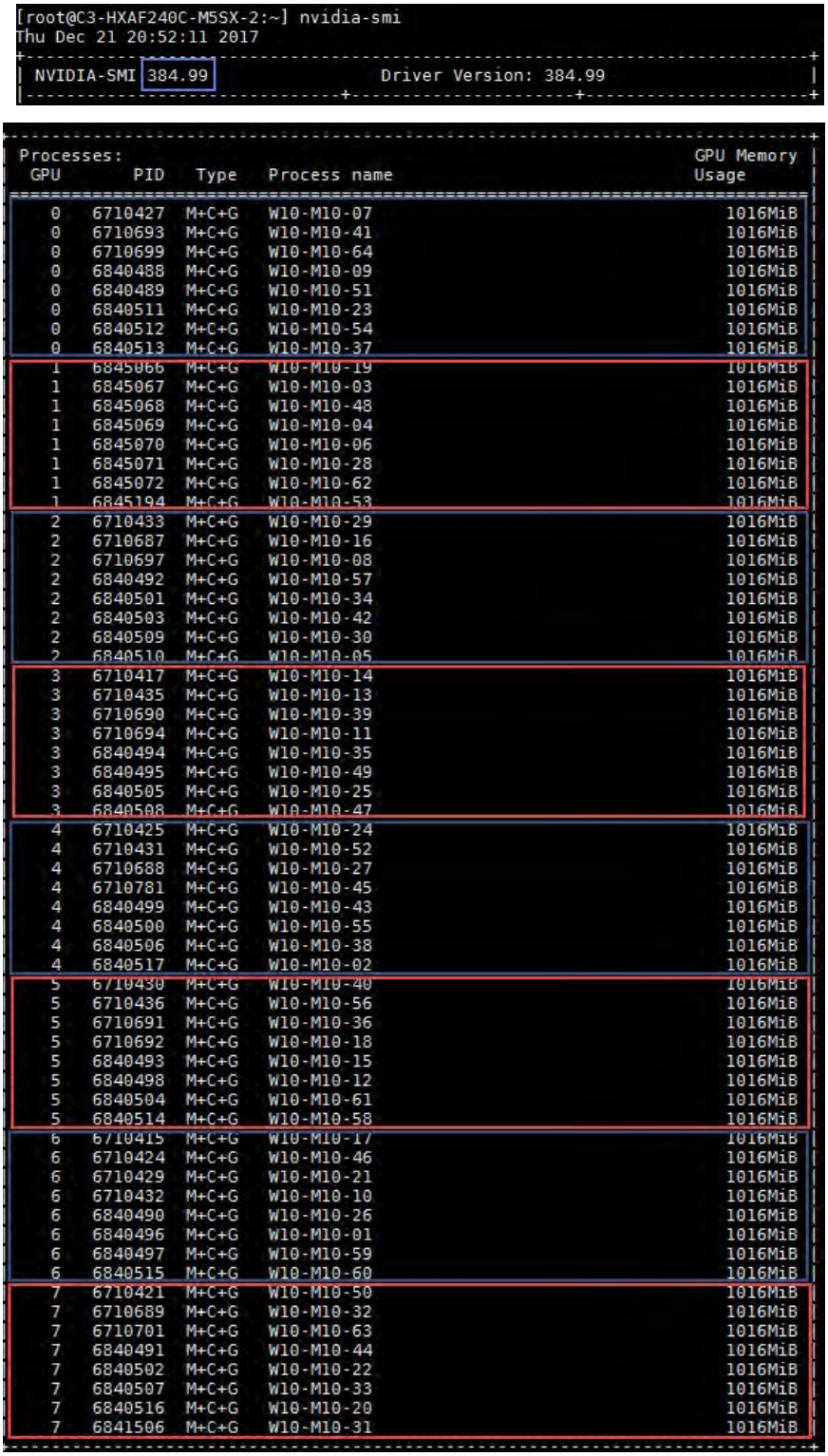


**Figure 62.** The nvidia-smi command output from the host with two NVIDIA M10 cards and 64 Microsoft Windows 10 desktops with M10-1B vGPU profile



```
[root@C3-HXAF240C-M5SX-2:~] nvidia-smi
Thu Dec 21 20:52:11 2017
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 384.99                 Driver Version: 384.99                    |
|-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID   Type   Process name                             Usage      |
|=============================================================================|
|    0   6710427   M+C+G   W10-M10-07                                1016MiB |
|    0   6710693   M+C+G   W10-M10-41                                1016MiB |
|    0   6710699   M+C+G   W10-M10-64                                1016MiB |
|    0   6840488   M+C+G   W10-M10-09                                1016MiB |
|    0   6840489   M+C+G   W10-M10-51                                1016MiB |
|    0   6840511   M+C+G   W10-M10-23                                1016MiB |
|    0   6840512   M+C+G   W10-M10-54                                1016MiB |
|    0   6840513   M+C+G   W10-M10-37                                1016MiB |
|    1   6845066   M+C+G   W10-M10-19                                1016MiB |
|    1   6845067   M+C+G   W10-M10-03                                1016MiB |
|    1   6845068   M+C+G   W10-M10-48                                1016MiB |
|    1   6845069   M+C+G   W10-M10-04                                1016MiB |
|    1   6845070   M+C+G   W10-M10-06                                1016MiB |
|    1   6845071   M+C+G   W10-M10-28                                1016MiB |
|    1   6845072   M+C+G   W10-M10-62                                1016MiB |
|    1   6845194   M+C+G   W10-M10-53                                1016MiB |
|    2   6710433   M+C+G   W10-M10-29                                1016MiB |
|    2   6710687   M+C+G   W10-M10-16                                1016MiB |
|    2   6710697   M+C+G   W10-M10-08                                1016MiB |
|    2   6840492   M+C+G   W10-M10-57                                1016MiB |
|    2   6840501   M+C+G   W10-M10-34                                1016MiB |
|    2   6840503   M+C+G   W10-M10-42                                1016MiB |
|    2   6840509   M+C+G   W10-M10-30                                1016MiB |
|    2   6840510   M+C+G   W10-M10-05                                1016MiB |
|    3   6710417   M+C+G   W10-M10-14                                1016MiB |
|    3   6710435   M+C+G   W10-M10-13                                1016MiB |
|    3   6710690   M+C+G   W10-M10-39                                1016MiB |
|    3   6710694   M+C+G   W10-M10-11                                1016MiB |
|    3   6840494   M+C+G   W10-M10-35                                1016MiB |
|    3   6840495   M+C+G   W10-M10-49                                1016MiB |
|    3   6840505   M+C+G   W10-M10-25                                1016MiB |
|    3   6840508   M+C+G   W10-M10-47                                1016MiB |
|    4   6710425   M+C+G   W10-M10-24                                1016MiB |
|    4   6710431   M+C+G   W10-M10-52                                1016MiB |
|    4   6710688   M+C+G   W10-M10-27                                1016MiB |
|    4   6710781   M+C+G   W10-M10-45                                1016MiB |
|    4   6840499   M+C+G   W10-M10-43                                1016MiB |
|    4   6840500   M+C+G   W10-M10-55                                1016MiB |
|    4   6840506   M+C+G   W10-M10-38                                1016MiB |
|    4   6840517   M+C+G   W10-M10-02                                1016MiB |
|    5   6710430   M+C+G   W10-M10-40                                1016MiB |
|    5   6710436   M+C+G   W10-M10-56                                1016MiB |
|    5   6710691   M+C+G   W10-M10-36                                1016MiB |
|    5   6710692   M+C+G   W10-M10-18                                1016MiB |
|    5   6840493   M+C+G   W10-M10-15                                1016MiB |
|    5   6840498   M+C+G   W10-M10-12                                1016MiB |
|    5   6840504   M+C+G   W10-M10-61                                1016MiB |
|    5   6840514   M+C+G   W10-M10-58                                1016MiB |
|    6   6710415   M+C+G   W10-M10-17                                1016MiB |
|    6   6710424   M+C+G   W10-M10-46                                1016MiB |
|    6   6710429   M+C+G   W10-M10-21                                1016MiB |
|    6   6710432   M+C+G   W10-M10-10                                1016MiB |
|    6   6840490   M+C+G   W10-M10-26                                1016MiB |
|    6   6840496   M+C+G   W10-M10-01                                1016MiB |
|    6   6840497   M+C+G   W10-M10-59                                1016MiB |
|    6   6840515   M+C+G   W10-M10-60                                1016MiB |
|    7   6710421   M+C+G   W10-M10-50                                1016MiB |
|    7   6710689   M+C+G   W10-M10-32                                1016MiB |
|    7   6710701   M+C+G   W10-M10-63                                1016MiB |
|    7   6840491   M+C+G   W10-M10-44                                1016MiB |
|    7   6840502   M+C+G   W10-M10-22                                1016MiB |
|    7   6840507   M+C+G   W10-M10-33                                1016MiB |
|    7   6840516   M+C+G   W10-M10-20                                1016MiB |
|    7   6841506   M+C+G   W10-M10-31                                1016MiB |
+-----------------------------------------------------------------------------+
```

## Additional configurations

This section presents additional configuration options.

### Install and upgrade NVIDIA drivers

The NVIDIA GRID API provides direct access to the frame buffer of the GPU, providing the fastest possible frame rate for a smooth and interactive user experience.

### Use Citrix HDX Monitor

Use the Citrix HDX Monitor tool (which replaces the Health Check tool) to validate the operation and configuration of HDX visualization technology and to diagnose and troubleshoot HDX problems. To download the tool and learn more about it, go to https://taas.citrix.com/hdx/download/.

### Optimize the Citrix HDX 3D Pro user experience

To use HDX 3D Pro with multiple monitors, be sure that the host computer is configured with at least as many monitors as are attached to user devices. The monitors attached to the host computer can be either physical or virtual.

Do not attach a monitor (either physical or virtual) to a host computer while a user is connected to the virtual desktop or the application providing the graphical application. Doing so can cause instability for the duration of a user's session.

Let your users know that changes to the desktop resolution (by them or an application) are not supported while a graphical application session is running. After closing the application session, a user can change the resolution of the Desktop Viewer window in Citrix Receiver Desktop Viewer Preferences.

When multiple users share a connection with limited bandwidth (for example, at a branch office), Citrix recommends that you use the "Overall session bandwidth limit" policy setting to limit the bandwidth available to each user. This setting helps ensure that the available bandwidth does not fluctuate widely as users log on and off. Because HDX 3D Pro automatically adjusts to make use of all the available bandwidth, large variations in the available bandwidth over the course of user sessions can negatively affect performance.

For example, if 20 users share a 60-Mbps connection, the bandwidth available to each user can vary between 3 and 60 Mbps, depending on the number of concurrent users. To optimize the user experience in this scenario, determine the bandwidth required per user at peak periods and limit users to this amount at all times.

For users of a 3D mouse, Citrix recommends that you increase the priority of the generic USB redirection virtual channel to 0. For information about changing the virtual channel priority, see Citrix article CTX128190.

## Use GPU acceleration for Microsoft Windows Server DirectX, Direct3D, and WPF rendering

DirectX, Direct3D, and WPF rendering are available only on servers with a GPU that supports display driver interface (DDI) Version 9ex, 10, or 11.

### Use OpenGL Software Accelerator

OpenGL Software Accelerator is a software rasterizer for OpenGL applications such as ArcGIS, Google Earth, NeHe, Maya, Blender, Voxler, CAD, and CAM. In some cases, OpenGL Software Accelerator can eliminate the need to use graphics cards to deliver a good user experience with OpenGL applications.

**Note:** OpenGL Software Accelerator is provided as is and must be tested with all applications. It may not work with some applications and is intended as a solution to try if the Windows OpenGL rasterizer does not provide adequate performance. If OpenGL Software Accelerator works with your applications, you can use it to avoid the cost of GPU hardware.

OpenGL Software Accelerator is provided in the Support folder on the installation media, and it is supported on all valid VDA platforms.

Try OpenGL Software Accelerator in the following cases:

- If the performance of OpenGL applications running on virtual machines is a concern, try using the OpenGL accelerator. For some applications, the accelerator outperforms the Microsoft OpenGL software rasterizer that is included with Windows because the OpenGL accelerator uses SSE4.1 and AVX. The OpenGL accelerator also supports applications using OpenGL versions up to Version 2.1.
- For applications running on a workstation, first try the default version of OpenGL support provided by the workstation's graphics adapter. If the graphics card is the latest version, in most cases it will deliver the best performance. If the graphics card is an earlier version or does not deliver satisfactory performance, then try OpenGL Software Accelerator.
- 3D OpenGL applications that are not adequately delivered using CPU-based software rasterization may benefit from OpenGL GPU hardware acceleration. This feature can be used on bare-metal devices and virtual machines.

## Test and evaluation notes

The Login VSI test framework with a custom power user workload was used to simulate users running graphics-intensive workloads and high-definition video content.

Figures 63 and 64 show differences in frames-per-second (FPS) rates during tests between desktops configured with and without a vGPU.

**Figure 63.** BouncingObjects.htm FPS on desktop without vGPU

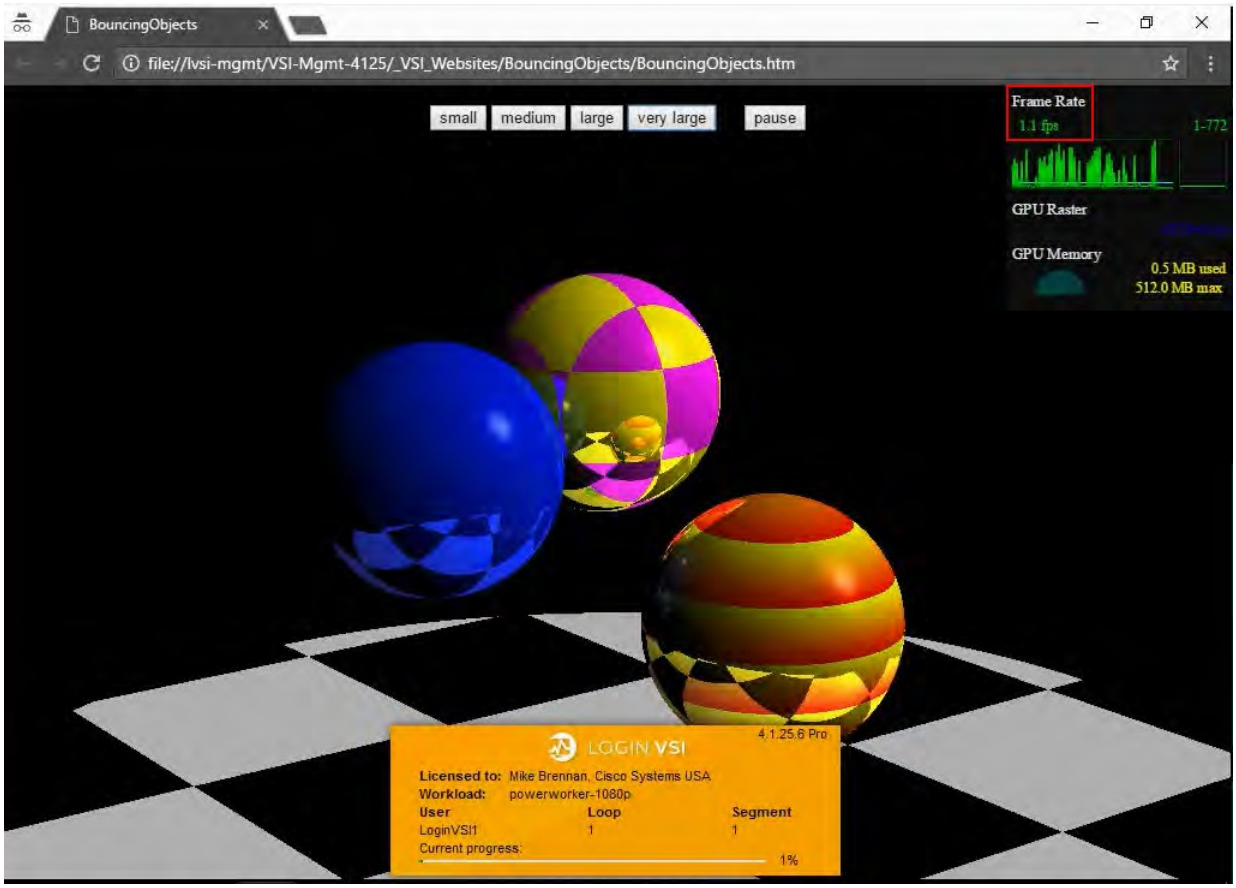
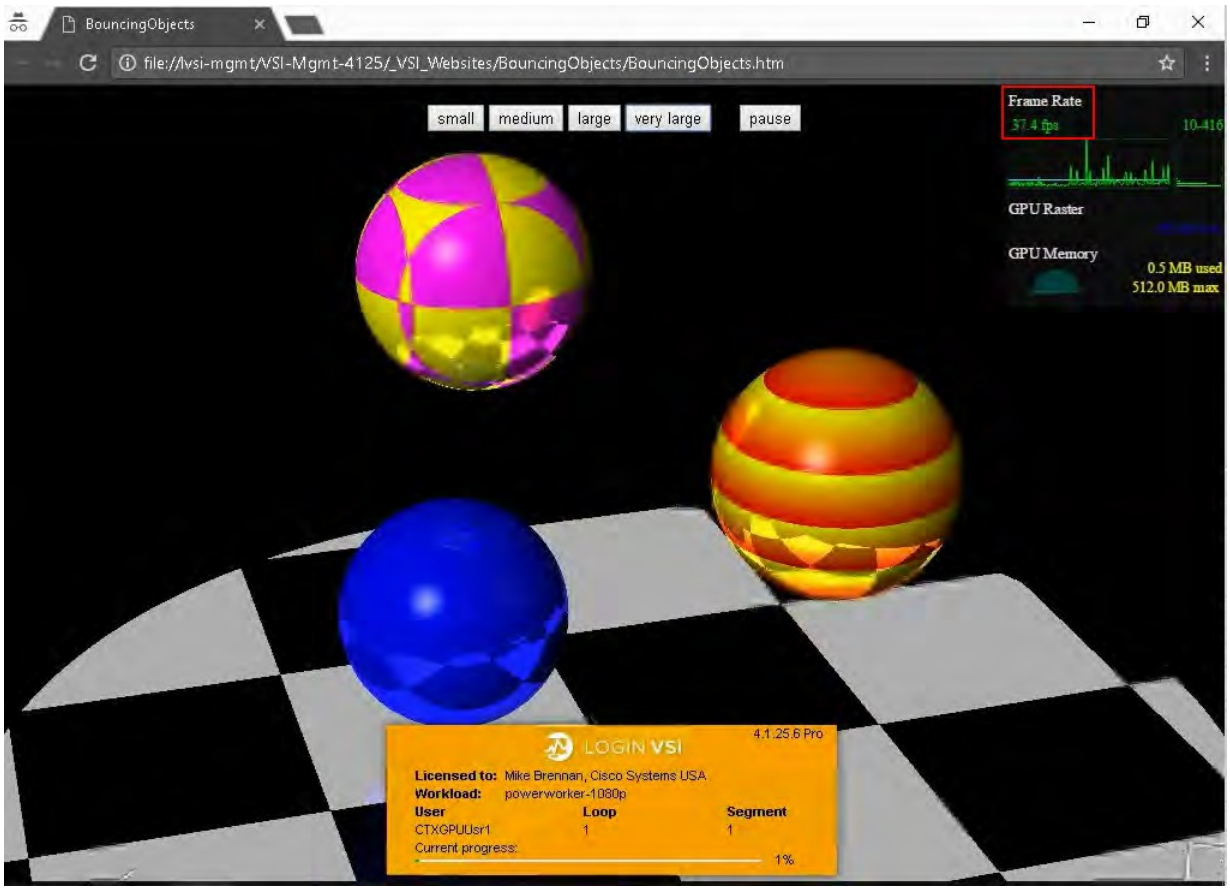


**Figure 64.** BouncingObjects.htm FPS on desktop with vGPU




Figures 65 and 66 provide examples of the CPU utilization seen during tests from the hosts running the same number of Microsoft Windows 10 desktops with and without a vGPU.

**Figure 65.** CPU utilization from the host with two NVIDIA P40 cards running 48 Microsoft Windows 10 desktops with P40-1B vGPU profile and the host without GPU running 48 Microsoft Windows 10 desktops

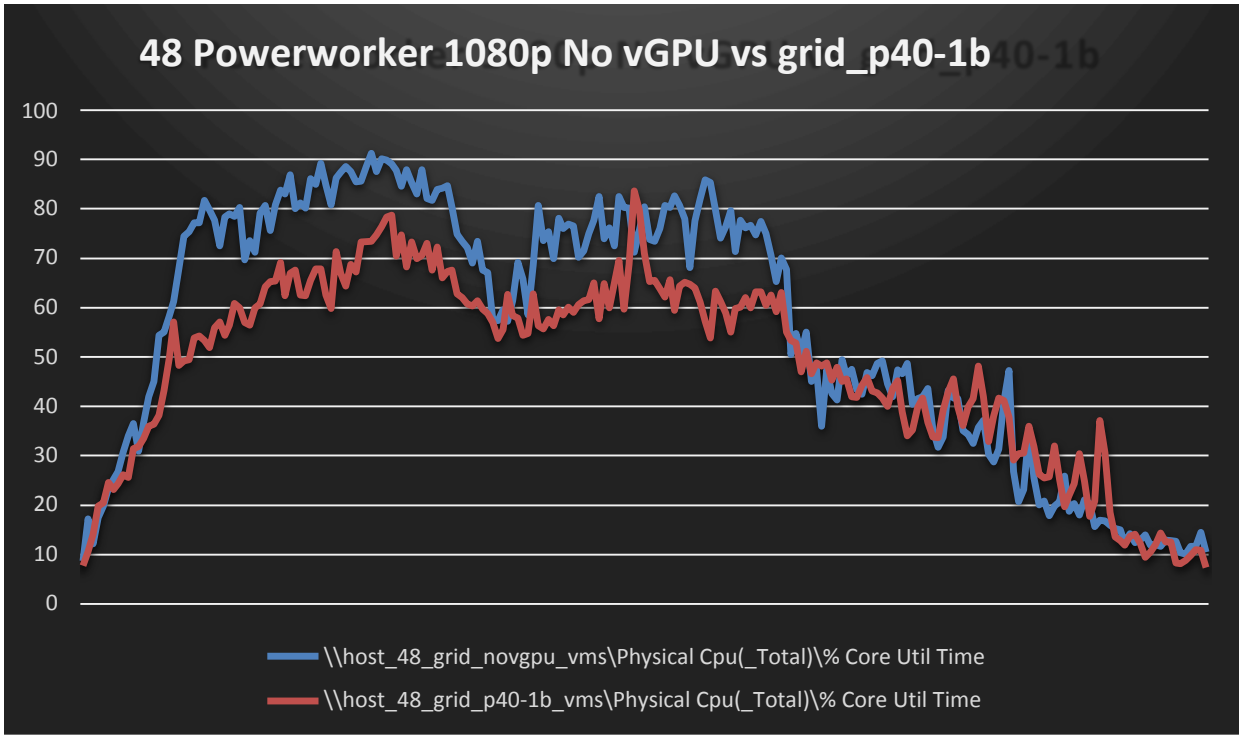
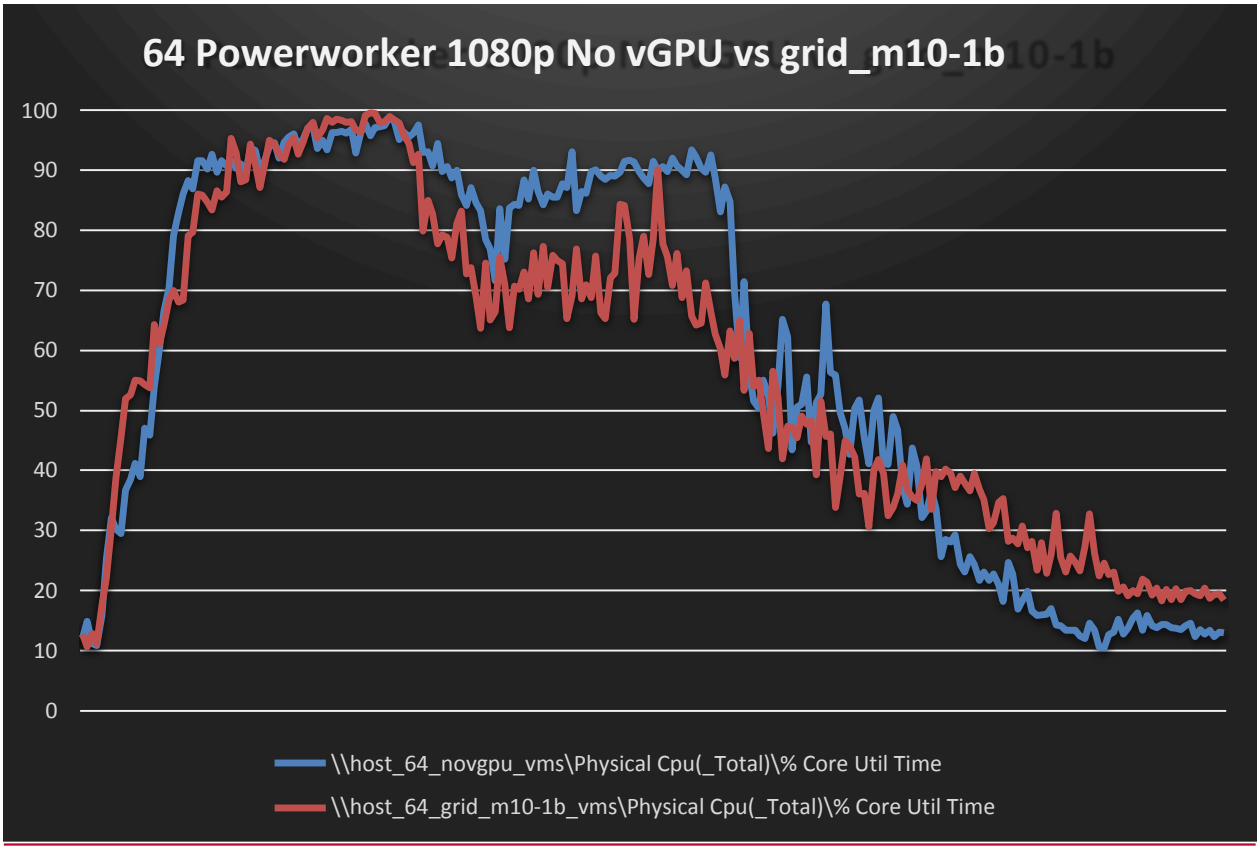


**Figure 66.** CPU utilization from the host with two NVIDIA M10 cards running 64 Microsoft Windows 10 desktops with M10-1B vGPU profile and the host without GPU running 64 Microsoft Windows 10 desktops




## Conclusion

The combination of Cisco UCS Manager, Cisco HyperFlex 2.6, Cisco UCS C240 M5 Rack Servers and B200 M5 Blade Servers, and NVIDIA Tesla cards running VMware vSphere 6.5 and Citrix XenDesktop 7.15 provides a high-performance platform for virtualizing graphics-intensive workloads.

By following the configuration guidance in this document, our customers and partners can be assured that they are ready to host the growing list of graphics applications that are supported by our partners.

## For more information

- Cisco UCS C-Series Rack Servers and B-Series Blade Servers:
  - http://www.cisco.com/en/US/products/ps10265/
- Cisco HyperFlex hyperconverged servers:
  - https://www.cisco.com/c/en/us/products/hyperconverged-infrastructure/hyperflex-hx-series/index.html
- NVIDIA:
  - http://www.nvidia.com/object/grid-technology.html
  - http://docs.nvidia.com/grid/latest/pdf/grid-software-quick-start-guide.pdf
  - http://docs.nvidia.com/grid/latest/pdf/grid-vgpu-release-notes-vmware-vsphere.pdf

- Citrix XenApp and XenDesktop 7.15:
  - https://docs.citrix.com/en-us/xenapp-and-xendesktop/7-15-ltsr.html
  - https://www.citrix.com/products/xenapp-xendesktop/hdx/hdx-3d-pro.html
  - http://blogs.citrix.com/2014/08/13/citrix-hdx-the-big-list-of-graphical-benchmarks-tools-and-demos/
- Microsoft Windows and Citrix optimization guides for virtual desktops:
  - http://support.citrix.com/article/CTX125874
  - https://support.citrix.com/article/CTX216252
  - https://labs.vmware.com/flings/vmware-os-optimization-tool
- VMware vSphere ESXi and vCenter Server 6.5:
  - http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2033434
  - http://pubs.vmware.com/vsphere-6-5/index.jsp
  - https://docs.vmware.com/en/VMware-vSphere/index.html

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at www.cisco.com/go/offices.



Printed in USA C11-740243-00 02/18