# NVIDIA GRID APPLICATION SIZING FOR AUTODESK REVIT 2016

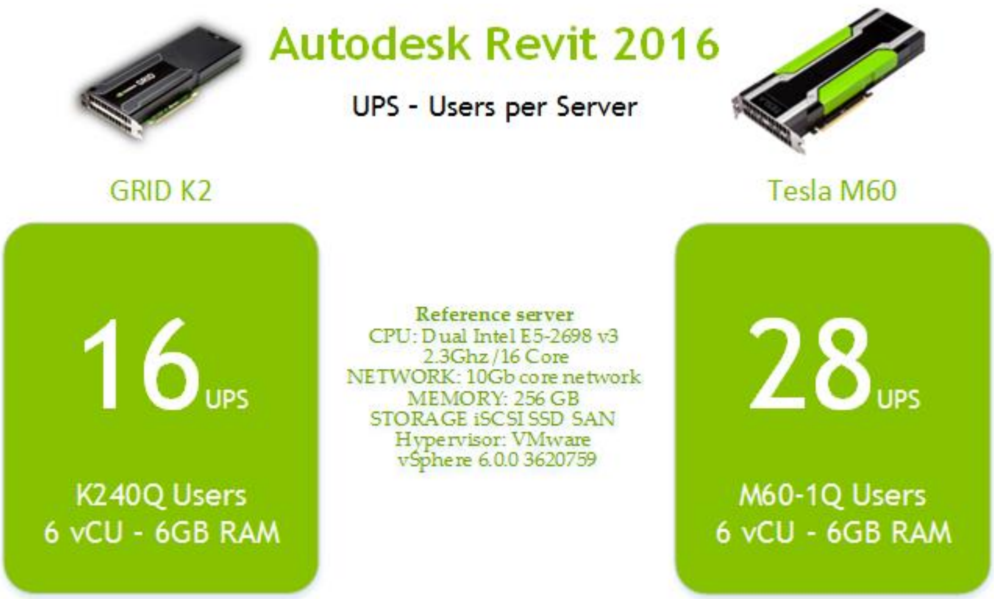
BPG-08489-001 | March 2017

**Best Practices Guide**

# TABLE OF CONTENTS

# USERS PER SERVER (UPS)

The purpose of this guide is to give a detailed analysis of how many users organizations can expect to get per servers based on performance testing with the Autodesk Revit 2016. The NVIDIA Performance Lab worked in cooperation with the Autodesk team to determine the maximum recommended number of users for the reference server configuration. Testing for this guide is based on the industry-standard RFO benchmark to determine the maximum number of Autodesk Revit users per server (UPS) that NVIDIA GRID® can support. To provide customers with a reference point, we have included testing for the latest generation of NVIDIA GRID solution, NVIDIA GRID Virtual Workstation software on NVIDIA Tesla™ M60, as compared to the previous generation, GRID K2. Based on extensive testing, NVIDIA GRID provides the following performance and scalability recommendation.



Reference server
CPU: Dual Intel E5-2698 v3
2.3Ghz /16 Core
NETWORK: 10Gb core network
MEMORY: 256 GB
STORAGE iSCSI SSD SAN
Hypervisor: VMware
vSphere 6.0.0 3620759

Figure 1 Autodesk Revit 2016 Users per Server for GRID K2 and Tesla M60

The maximum number of users per server is based on performance and scalability metrics for Autodesk Revit 2016 configured to perform high workloads concurrently while maintaining reasonable usability.

! **NOTE:** THESE NUMBERS ARE INTENDED TO BE USED AS GENERAL GUIDANCE FOR A MAXIMUM NUMBER OF USERS PER HOST. CHANGES IN SERVER CONFIGURATION AND DIFFERENCES IN USAGE OF THE SOFTWARE WILL CAUSE PERFORMANCE TO VARY.

# TECHNOLOGY OVERVIEW

## AUTODESK REVIT 2016 APPLICATION

Autodesk Revit is Building Information Modeling (BIM) software with features for the following elements of building design and construction:

- Architectural design
- Mechanical, engineering, plumbing (MEP) design
- Structural engineering
- Construction

When architecting your NVIDIA GRID environment for Revit, you must consider both the GPU and the CPU.

- Revit requires a GPU as users rotate, zoom, and interact with drawings.
- Revit creates a heavy CPU load as it manages all the elements of a drawing through a database.

  Because it uses a database, Revit needs high performance storage to function properly. The heaviest Revit CPU usage occurs during data-rich operations such as opening and saving files, and updating models.

## NVIDIA GRID PLATFORM

NVIDIA re-defined visual computing by giving designers, engineers, scientists, and graphics artists the power to take on the biggest visualization challenges with immersive, interactive, photorealistic environments.

NVIDIA GRID exploits the power of NVIDIA Tesla GPUs to deliver virtual workstations from the data center or the cloud. Architects, engineers, and designers are now liberated from their desks and can access their graphics-intensive applications and data anywhere.

The NVIDIA Tesla M60 GPU accelerator works with NVIDIA GRID software to provide the industry's highest user performance for virtualized workstations, desktops, and applications. This solution allows enterprises to virtualize any application—including professional graphics applications—and deliver them to any device anywhere.

Since its first release in 2013, NVIDIA GRID has supported GPU cards based on two generations of GPU hardware architecture:

- GRID K1 and K2 GPU cards based on the NVIDIA Kepler™ architecture
- Tesla M6, M10, and M60 GPU cards based on the NVIDIA Maxwell™ architecture

NVIDIA GRID has seen considerable software innovation to continue to drive the best performance and density on the market.

## Software and Hardware Used in the Tests

The tests described in this guide are based on the following combinations of software and hardware:

- VMware Horizon running the first-generation NVIDIA GRID K2 GPU
- VMware Horizon and NVIDIA GRID Virtual Workstation software running on the second-generation Tesla M60 GPU

As shown in Table 1, using the latest generation provides better performance and scalability, and the ability to take advantage of new features and functionality of the software.

### Table 1 Comparison of GRID K2 and Tesla M60

| Feature | NVIDIA GRID K2 | NVIDIA Tesla M60 |
|---|---|---|
| GPU architecture | NVIDIA Kepler | NVIDIA Maxwell |
| GPUs per card | 2 | 2 |
| Maximum users per card | 16 (8 per GPU) | 32 (16 per GPU) |
| NVIDIA CUDA cores | 3,072 NVIDIA Kepler cores (1536 per GPU) | 4096 NVIDIA CUDA Cores (2048 per GPU) |
| GPU memory | 8 GB of GDDR5 memory (4 GB per GPU) | 16 GB of GDDR5 memory (8 GB per GPU) |
| H.264 1080p30 streams | No H.264 support | 36 |
| Max Power Consumption | 225 W | 300 W |
| Thermal Solution | Active/Passive | Active/Passive |
| Form Factor | PCIe 3.0 Dual Slot | PCIe 3.0 Dual Slot |

## Hardware Encoding with NVENC and VMware Blast Extreme

NVIDIA and VMware have been working together for several years to improve the virtualized computing user experience and enable a completely new class of virtual use cases. NVIDIA was the first vendor to enable hardware-accelerated graphics rendering in VMware Horizon View. NVIDIA then enabled the first virtualized graphics acceleration in VMware Horizon View with GRID.

The VMware Blast Extreme protocol, which was released in VMware Horizon 7, enables NVIDIA GRID to offload the H.264 processing from the CPU to the GPU. This offloading frees resources for use by internal applications, increasing user density and application responsiveness. The H.264 codec lowers the demand on network infrastructure, enabling organizations to reach more users over greater network lengths.

VMware offers multiple protocols that are designed for different workloads. The choice of protocol may impact performance, density, image quality, and other factors. Therefore, you must select the best protocol for the needs of your organization. For more information about Horizon with Blast Extreme, refer to Blast Extreme Display Protocol in Horizon 7.

# TESTING METHODOLOGY

This section describes the tests performed and the method of testing used to determine sizing and server loads.

## THE PERFORMANCE ENGINEERING LAB

The mandate of the NVIDIA GRID Performance Engineering Team is to measure and validate the performance and scalability delivered by the NVIDIA GRID platform, namely GRID software running on Tesla GPU's, on all enterprise virtualization platforms. It is the goal of the Performance Engineering Team to provide proven testing that gives NVIDIA's customers the ability to deliver a successful deployment.

The NVIDIA Performance Engineering Lab holds a wide variety of different OEM servers, with varying CPU specifications, storage options, client devices, and network configurations. This lab of enterprise virtualization technology provides the Performance Engineering team with the capacity needed to run a wide variety of tests ranging from standard benchmarks to reproducing customer scenarios on a wide range of hardware.

None of this work is possible without the cooperation of ISVs, OEMs, vendors, partners, and their user communities to determine the best methods of benchmarking in ways that are both accurate and reproducible. These methods will ultimately assist mutual customers of NVIDIA and other vendors and OEMs to build and test their own successful deployments. In this way, the Performance Engineering Team works closely with its counterparts in the enterprise virtualization community.

# RECOMMENDED REVIT PHYSICAL SYSTEM CONFIGURATION

Physical system requirements for Autodesk Revit 2016 are listed on the Revit product page. Testing focuses on recommended specifications when feasible. The goal is to test both performance and scalability, maintaining the flexibility and manageability advantages of virtualization without sacrificing the performance end users expect from NVIDIA powered graphics.

It has been well documented that storage performance is key to providing high performance graphics workloads, especially with many users and ever-growing file or model sizes. In the NVIDIA Performance Engineering Lab, a 10G iSCSI-connected all flash SAN from Dell EMC XtremeIO was used. At no time in these tests were IOPS an issue, but note that as you scale to multiple servers hosting many guests, IOPS needs to be monitored.

# TYPICAL REVIT WORKSTATION BUILDS

Autodesk delivers a recommended hardware specification to help choose a physical workstation. These recommendations provide a good starting point from which to start architecting your virtual desktops. Your own tests with your own models will determine if these recommendations meet your specific needs.

Table 2 lists server configuration used for the benchmark tests described in this guide. This configuration is based on testing results from the RFO benchmarks, input from VMware, and feedback from mutual customers.

### Table 2 Server Configuration for Benchmark Testing of Revit 2016

| Configuration | GRID K2 | Tesla M60 |
|---|---|---|
| **VMware Configuration** | | |
| VMware software | VMware Horizon 7 or later<br>VMware vSphere 6 or later | |
| Virtual machine operating system | Microsoft® Windows® 7 SP1 64-bit:<br>Windows 7 Enterprise, Ultimate, or Professional | |

| Configuration | GRID K2 | Tesla M60 |
|---|---|---|
| **Host Server Specification** | | |
| CPU<br>(Haswell, Intel® Xeon E5 v3, or greater recommended) | 2.6 GHz Intel® Xeon E5-2697 v3 | 2.3 GHz Intel® Xeon E5-2698 v3 |
| Memory | 256 GB | 256 GB |
| Networking | 1 Gb VM<br>10 Gb storage | 1 Gb VM<br>10 Gb storage |
| **Virtual Machine Settings** | | |
| Memory | 8-12 GB RAM | 16-32 GB RAM |
| vCPUs | 6 vCPUs | 6 vCPUs |
| Graphics adapter | K240Q | M60-1Q |
| Internet connection | Required for license registration and prerequisite component download | |
| End user access | Each client computer should have the latest VMware® Horizon Client installed. | |

# AUTODESK REVIT BENCHMARK METRICS (RFO)

Autodesk provides a tool called AUBench which, when combined with the scripts provided by the Revit Forums community, creates a benchmark called RFO. RFO interacts with the application and an accompanying model to run several tests, then checks the journal for time stamps, and reports the results. The benchmark is available from the RFOBenchmark thread on the Revit Forum.

These tests are designed to represent user activities and are broken down as follows:

- Model Creation and View and Export of Benchmarks
  - Opening and Loading the Custom Template
  - Creating the Floors Levels and Grids
  - Creating a Group of Walls and Doors
  - Modifying the Group by Adding a Curtain Wall
  - Creating the Exterior Curtain Wall
  - Creating the Sections
  - Changing the Curtain Wall Panel Type
  - Exporting all Views as PNGs
  - Exporting Some Views as DWGs
- Render Benchmark
  - Render

- GPU Benchmark[1] with Hardware Acceleration
  - Refresh Hidden Line View ×12 - with Hardware Acceleration
  - Refresh Consistent Colors View ×12 - with Hardware Acceleration
  - Refresh Realistic View ×12 - with Hardware Acceleration
  - Rotate View ×1 - with Hardware Acceleration
- GPU Benchmark[1] Without Hardware Acceleration
  - Refresh Hidden Line View ×12 - Without Hardware Acceleration
  - Refresh Consistent Colors View ×12 - Without Hardware Acceleration
  - Refresh Realistic View ×12 - Without Hardware Acceleration
  - Rotate View ×1 - Without Hardware Acceleration

## TEST WORKLOADS

To ensure you will be able to reproduce the results described in "Test Results," on page 11, the Revit Forums RFO Benchmark workload was deliberately chosen and simultaneous tests were executed. As a result, all virtual desktops in the test perform the same activities at the same time. A "peak workload" should be unrepresentative of real user interaction but shows the number of users per host when the load on the shared resources is highest and, therefore, represents the most extreme case of user demand.

If you plan to run these tests, consider these aspects of the test workloads:

- **Sample workload.** RFO provides their workload, which is a set of models, for testing with.
- **Scripting.** As RFO is historically designed for single physical workstation testing, there is no built-in automation for multi-desktop scalability testing.
- **Think time.** Pausing between tests simulates human behavior.
- **Staggered start.** Adding a delay to the beginning of each test offsets the impact of running the tests in unison, again, simulating human behavior.
- **Scalability.** In general, the tests are run with 1 virtual desktop, then 8 virtual desktops, and finally 16 virtual desktops. These test runs get a baseline of results and accompanying logs (CPU, GPU, RAM, networking, storage IOPS, and so forth.).

[1] For the hardware acceleration comparison only.

## HOW THE TESTS WERE RUN

1. Virtual machines were created with a standard configuration to determine the threshold of acceptable performance.
2. The RFO benchmark was run with an individual VM on each GRID K2 and Tesla M60 system to determine the peak performance when there was no resource contention.
3. To determine acceptable performance at scale, a 25% increase in the test time required for the total of the tests below was added to the value obtained in the previous step.
4. The threshold of performance in virtual environment testing was done successfully across multiple applications.
5. After the performance threshold for Revit 2016 from the single VM tests was determined, it was used to indicate peak user density for all systems in this guide.

## TEST LIMITATIONS

By design, this type of peak performance testing leads to conservative estimates of scalability. Automated scalability testing is typically more aggressive than a typical user workflow. In most cases, rendering requests are unlikely to be executed by 10 users simultaneously or even to the degree that was replicated in multiple test iterations.

Therefore, the results from these automated scalability tests can be considered a worst-case scenario. They indicate likely minimums for rare conditions to serve as safe guidelines. In most cases, a host should be able to support more than the number of VMs indicated by the test results.

The scalability in typical or even heavily loaded data centers is likely to be higher than the test results suggest. The degree to which higher scalability would be achieved depends on the typical day-to-day activities of users, such as the amount of time spent in meetings, the length of lunch breaks or other breaks, the frequency of multitasking, and so forth.

# TEST RESULTS

The test results show execution times for performing various operations under varying conditions and compare CPU utilization against GPU utilization. The recommendations for the number of users per system suggested by these results is a balance between the need for greatest scalability and the performance expectations of users. Note that your users, your data, and your hardware will impact these results and you may decide that a different level of performance or scalability is required to meet your individual business needs.

## TEST THRESHOLD TIMES

The RFO Benchmark does not currently exercise some of Revit's newest GPU capabilities and was built to push the limits of dedicated hardware as opposed to the shared resources of VDI. Therefore, the decision was made to stop testing when the test times had increased to **125%** of values measured on a single virtual workstation.

Table 3 shows the test threshold times for the NVIDIA GRID K2 and Tesla M60 cards.

### Table 3 Test Threshold Times

| Measurement | K2 | M60 | Comments |
|---|---|---|---|
| 1 VM total test time | 209.47 | 199.80 | |
| Threshold (total test time X 125%) | 261.84 | 239.76 | Maximum threshold for all tests = 261.84 |

## GRID K2 TEST EXECUTION TIMES

Table 4 shows test execution times for the GRID K2 system with 8, 12, and 16 VMs. Note that at the maximum possible number of users (16 by design), the total test time (248.90) is still less than the threshold value of 261.84. Therefore, **16** users can comfortably reside on the same GRID K2 server.

### Table 4 GRID K2 Test Execution Times

| Operation | Value for 8 Users (VMs) | Value for 12 Users (VMs) | Value for 16 Users (VMs) |
|---|---|---|---|
| Opening and loading the custom template | 5.99 | 5.90 | 6.91 |
| Creating the floors levels and grids | 13.75 | 13.13 | 13.77 |
| Creating a group of walls and doors | 28.66 | 30.16 | 31.69 |
| Modifying the group by adding a curtain wall | 55.30 | 58.17 | 62.23 |
| Creating the exterior curtain wall | 15.95 | 17.02 | 17.82 |
| Creating the sections | 9.85 | 10.07 | 10.54 |
| Changing the curtain wall panel type | 5.59 | 5.72 | 6.08 |
| Exporting all views as PNGs | 36.54 | 39.27 | 42.23 |
| Exporting some views as DWGs | 42.53 | 46.10 | 57.63 |
| Total | 214.14 | 225.53 | 248.90 |

## TESLA M60 WITH NVENC TEST EXECUTION TIMES

Table 5 shows the test execution times when the Tesla M60 is accessed through the VMware Blast Extreme protocol. Note that the number of users per host is **increased to 28**.

This improvement is attributable to the improved Tesla M60 GPU and the use by Blast Extreme of the NVIDIA Video Encoder (NVENC) technology. The use of Blast Extreme allows Horizon to offload encoding of the H.264 video stream from the CPU to dedicated encoder engines on the Tesla GPUs, freeing up this much-needed resource for general computing purposes.

The video codec is a very important piece in delivering a remarkable user experience because it impacts many factors, such as like latency, bandwidth, frames per second (FPS), and other factors. Using H.264 as the primary video codec also allows VMware to use millions of H.264 enabled access devices to offload the encode-decode process from the CPU to dedicated H.264 engines on NVIDIA GPUs.

Table 5 Tesla M60 with NVENC Test Execution Times

| Operation | Value for 28 Users (VMs) | Value for 30 Users (VMs) |
|---|---|---|
| Opening and loading the custom template | 5.41 | 8.96 |
| Creating the floors levels and grids | 13.80 | 15.48 |
| Creating a group of walls and doors | 31.78 | 34.69 |
| Modifying the group by adding a curtain wall | 63.02 | 66.31 |
| Creating the exterior curtain wall | 17.72 | 19.32 |
| Creating the sections | 10.52 | 10.87 |
| Changing the curtain wall panel type | 6.25 | 6.30 |
| Exporting all views as PNGs | 41.88 | 45.38 |
| Exporting some views as DWGs | 64.49 | 73.80 |
| Total | 254.87 | 281.11[2] |

# TEST EXECUTION TIMES WITH AND WITHOUT HARDWARE ACCELERATION

The RFO Benchmark also provides results for testing with only CPU, and not using GPU at all. For these particular tests, the GPU is ignored by the software if present, and only CPU is used to graphics processing. This test has been included in the RFO benchmark for illustrative purposes, and we present the results here for the same reasons. It is, in our opinion, highly unlikely that anyone would seriously consider using a graphics intensive application such as Revit without the benefit of a GPU. In the event that someone does, the following tables illustrate the potential pitfall of doing so.

Table 6 and Figure 2 show the value of using a GPU in the VM with Revit 2016. CPU only response times (“without hardware acceleration”) are increased by a factor of ten or more.

For all cases, CPU-only tests (without hardware acceleration) for a single VM take more time than 16- and 24-VM tests with a GPU (with hardware acceleration).

[2] Exceeds the threshold of 261.84.

Table 6 Test Execution Times with and Without Hardware Acceleration

| VM Configuration | Refresh Hidden Line view × 12 | Refresh Consistent Colors view × 12 | Refresh Realistic view × 12 | Rotate view × 1 |
|---|---|---|---|---|
| 1 VM without hardware acceleration (no GPU) | 79.24 | 80.77 | 83.21 | 47.76 |
| 16 VMs with hardware acceleration (with GPU) | 9.19 | 8.64 | 11.36 | 3.23 |
| 24 VMs with hardware acceleration (with GPU) | 13.32 | 11.49 | 14.86 | 4.32 |

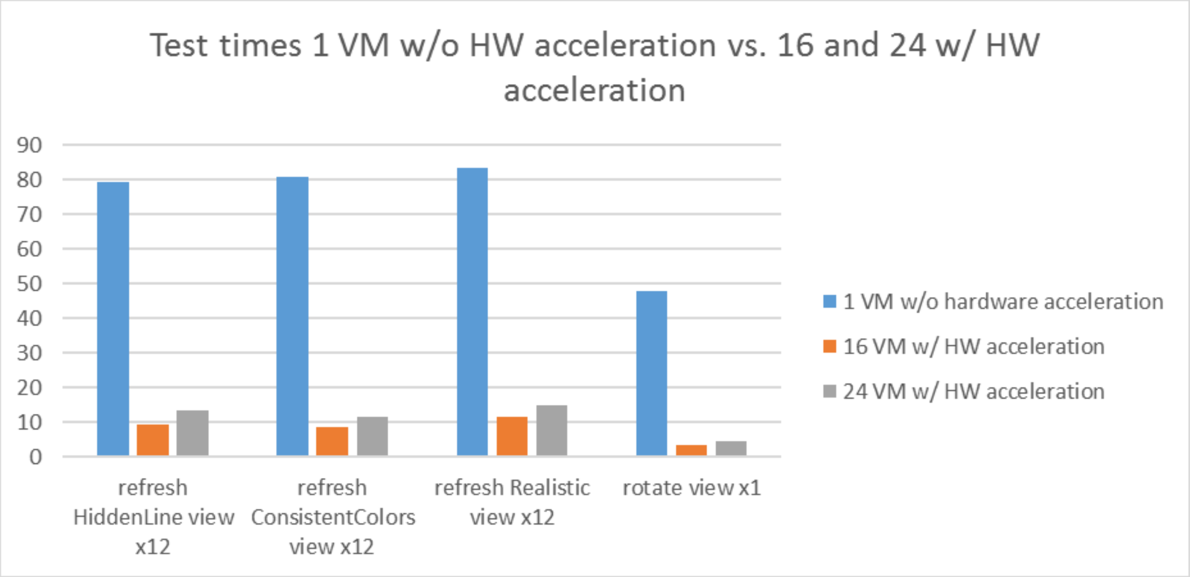


Figure 2 Test Execution Times with and Without Hardware Acceleration

# CPU UTILIZATION AGAINST GPU UTILIZATION

Figure 3 shows for the M60-1Q GPU profile that the CPU becomes a limiting factor long before the GPU resources are exhausted.

Revit requires significant CPU resources so investing in more cores can yield greater performance and scalability. For medium-to-large models, M60-1Q performance will be better for a real-use scenario than a M60-0Q  profile, which doesn’t provide enough frame buffer for satisfactory performance for Revit workloads. However, your results will vary. Therefore, you must test with your own models to ensure the most accurate results.

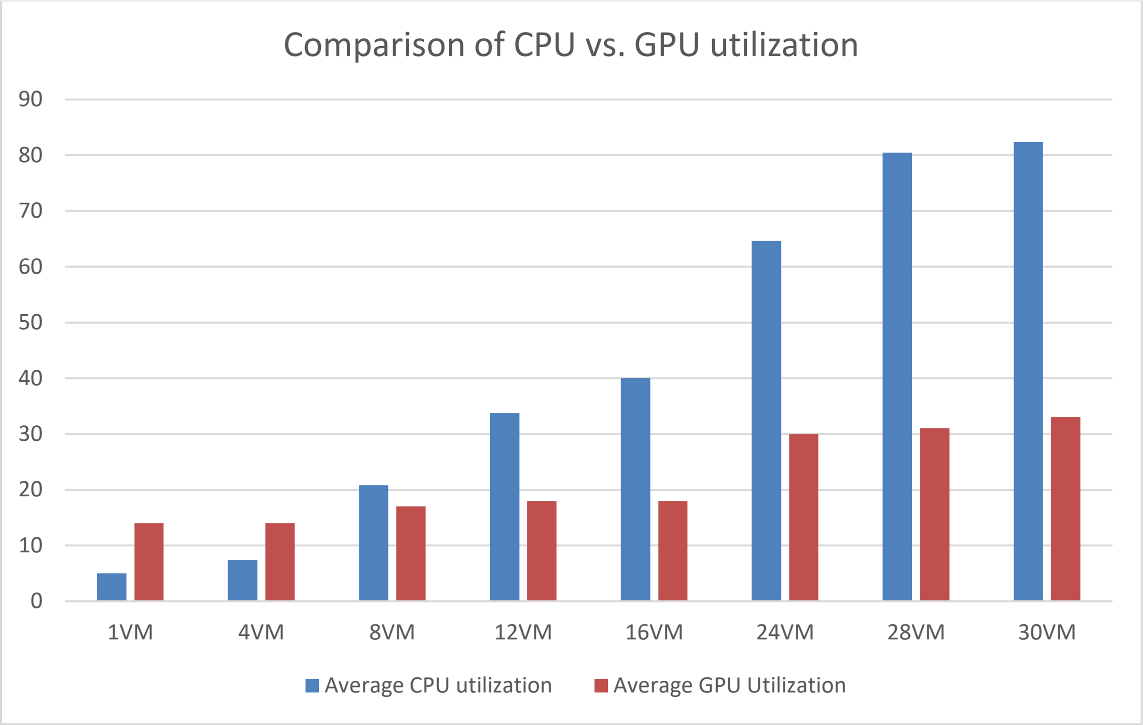


Figure 3 CPU and M60-1Q GPU Utilization

# CONCLUSION

The test results show that with NVIDIA GRID K2 GPUs, one reference server can support **up to 16 users without exceeding the 125%** threshold set by our testing team. With NVIDIA Tesla M60 GPU accelerators with VMware Horizon and Blast Extreme, **28 users per server is achievable.**

As testing has shown, Autodesk Revit is CPU intensive because it uses database software to manage the elements of a drawing. Therefore, the more complex the model, the more CPU intensive the software will become. When calculating scalability, be aware that the CPU will become the bottleneck before the GPU.

In practice, consolidating similar loads onto servers containing GPUs is preferable to running mixed workloads. Plan for these systems to run Revit or other GPU intensive loads, leaving servers that lack GPUs for loads that are not GPU intensive.

## Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, “MATERIALS”) ARE BEING PROVIDED “AS IS.” NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

## HDMI

HDMI, the HDMI logo, and High-Definition Multimedia Interface are trademarks or registered trademarks of HDMI Licensing LLC.

## OpenCL

OpenCL is a trademark of Apple Inc. used under license to the Khronos Group Inc.

## Trademarks

NVIDIA and the NVIDIA logo are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

## Copyright