Analyzing the Performance Impact of GPU Power Level Using SPEChpc 2021

Planning / Implementation

The power level of a GPU (graphics processing unit) can have a significant impact on its performance. GPU power level refers to the amount of power supplied to the GPU by the system. Generally, increasing the power level can result in higher GPU clock speeds and better performance, but it also increases power consumption and can generate more heat. To analyze the performance impact of GPU power level, we conduct the SPEChpc 2021 benchmark tests using various GPU power levels and measure the resulting performance. To avoid other factors that may impact the performance, we've ensured the CPU and memory are under good status without throttle and consistent ambient temperature when adjusting the GPU power level.

SPEChpc 2021 benchmark

To address the dramatically increase in workload in the High-Performance Computing (HPC) area, more and more modern HPC systems are built with heterogeneous architecture, which means accelerators such as GPU are part of the system to help improve the overall system performance. However, the heterogeneous design needs compiler evolution to overcome the portability challenge across both homogeneous and heterogeneous systems.

In addition, such heterogenous designs also increase complexity and poses challenges to performance evaluation. The High-Performance Group (HPG) under the Standard Performance Evaluation Corporation (SPEC) organization has developed industry-standard HPC benchmark called SPEChpc 2021 to support multiple host and accelerator programming model for modern HPC systems. The SPEChpc 2021 suite supports pure MPI, MPI+OpenMP, MPI+OpenMP target offload, MPI+OpenACC to address the majority type of heterogeneous HPC systems.

The following table lists all the sub-benchmark names, implementation language and each application area.

Table 1. Sub-benchmarks of SPEChpc 2021

Application Name	Language	Area
LBM	C	Computational fluid dynamics
SOMA	C	Physics / Polymeric systems
Tealeaf	C	Physics / High energy physics
Cloverleaf	Fortran	Physics / High energy physics
Minisweep	C	Nuclear engineering - Radiation transport
POT3D	Fortran	Solar hhysics
SPH-EXA	C++14	Astrophysics and Cosmology
HPGMG-FV	C	Cosmology, Astrophysics, Combustion
miniWeather	Fortran	Weather

To fit different cluster sizes, the SPEChpc 2021 provides four suites tiny, small, medium and large that includes different workload sizes as shown in the following table.

Table 2. Four suites of SPEChpc 2021 benchmark

Suite	Description
Tiny	The Tiny workloads use up to 60GB of memory and are intended for use on a single node using between 1 and 256 ranks.
Small	The Small workloads use up to 480GB of memory and are intended for use on one or more nodes using between 64 and 1024 ranks.
Medium	The Medium workloads use up to 4TB of memory and are intended for use on a mid-size cluster using between 256 and 4096 ranks.
Large	The Large workloads use up to 14.5TB of memory and are intended for use on a larger cluster using between 2048 and 32,768 ranks.

For more information about the SPEChpc 2021, visit the SPEChpc 2021 home page: https://www.spec.org/hpc2021

ThinkSystem SR655 V3

The experiment performed on the Lenovo ThinkSystem SR655 V3, which is a 1-socket server that features the AMD EPYC 9004 "Genoa" family of processors. With up to 96 cores per processor and support for the new PCIe 5.0 standard for high performance GPU, the SR665 V3 provides the best system performance a 2U form factor.

Figure 1. Lenovo ThinkSystem SR655 V3

For more information about SR655 V3, see the Lenovo Press product guide: https://lenovopress.lenovo.com/lp1610-thinksystem-sr655-v3-server

The configuration used for the experiment consisted of the following:

- 1x Lenovo ThinkSystem SR655 V3 server
- 1x AMD EPYC 9654P Processor (96 cores, 2.45 GHz)
- 192 GB memory (12x 16GB RDIMMs running at 4800 MHz)
- 1x 480 GB SATA 2.5" SSD
- 1x NVIDIA Tesla H100 80GB
- Red Hat Enterprise Linux Server release 8.6, Kernel

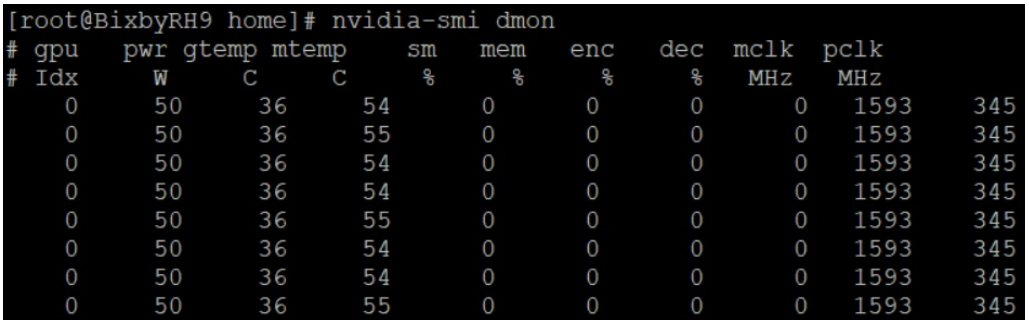
Profiling SPEChpc 2021 sub-benchmarks

Targeting the artificial intelligence (AI), high-performance computing (HPC), and data analytics, the NVIDIA H100 80G PCIe 350W TDP (Thermal Design Power) GPU is composed of multiple GPU Processing Clusters (GPCs), Texture Processing Clusters (TPs), Streaming Multiprocessors (SMs), and memory controllers. The NVIDIA H100 GPU consists of:

The NVIDIA PCIe Gen 5 board form-factor H100 GPU includes the following units:

- 60 MB L2 Cache
- 80 GB HBM3
- 8 GPCs each contains 9 TPCs, total 72 TPCs
- 2 SMs per TPC, total 114 SMs
- Fourth Generation NVLink for cross GPU connection

NVIDIA provides powerful diagnostic tool called `nvidia-smi` for user to monitor GPU status, including SM utilization, SM clock frequency, memory footprint, memory operating clock, power consumption and so on.



```
[root@BixbyRH9 home]# nvidia-smi dmon
# gpu   pwr gtemp mtemp    sm   mem   enc   dec  mclk  pclk
# Idx     W     C     C     %     %     %     %   MHz   MHz
    0    50    36    54     0     0     0     0  1593   345
    0    50    36    55     0     0     0     0  1593   345
    0    50    36    54     0     0     0     0  1593   345
    0    50    36    54     0     0     0     0  1593   345
    0    50    36    55     0     0     0     0  1593   345
    0    50    36    54     0     0     0     0  1593   345
    0    50    36    54     0     0     0     0  1593   345
    0    50    36    55     0     0     0     0  1593   345
```

Figure 2. The nvidia-smi command with dmon argument use to monitor GPU status

Using the nvidia-smi command, we sampled the NVIDIA H100 GPU every second while running the SPEChpc 2021 benchmark to profiling the runtime behavior as documented in the following sections.

GPU utilization and frequency

Most sub-benchmarks fully utilize the GPU compute power, and the Streaming Multiprocessors (SM) utilization rate reaches 100% during the execution. Only x32's utilization goes down to around 50% due to poor parallelization.

Although the utilization rate is high, the SM operating frequency varies among different benchmarks, the 505, 513, 528, 532 and 535 benchmarks can't keep maximum frequency most of the time during the run.

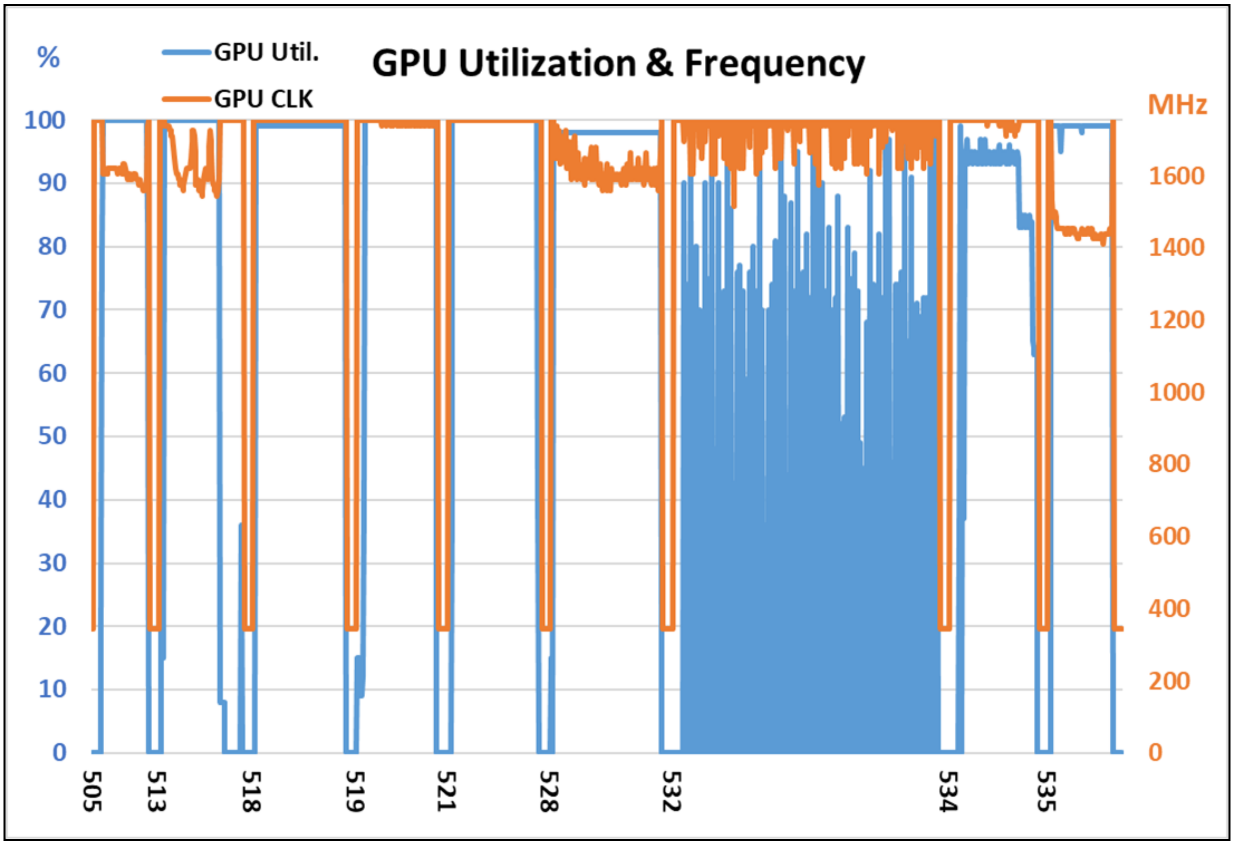


Figure 3. GPU utilization and frequency when running SPEChpc 2021

GPU memory Utilization and frequency

The GPU memory utilization rate indicates the percent of time over the past sample period during which global (device) memory was being read or written. The 518, 519, 528 and 535 are memory bandwidth hungry so the memory utilization is above 90% during the run. The 505, 513, 521 and 534 consume less memory bandwidth, the utilization is around 40% to 70%. Since the less parallelization the memory utilization for the 532 is low (around 10% to 20%).

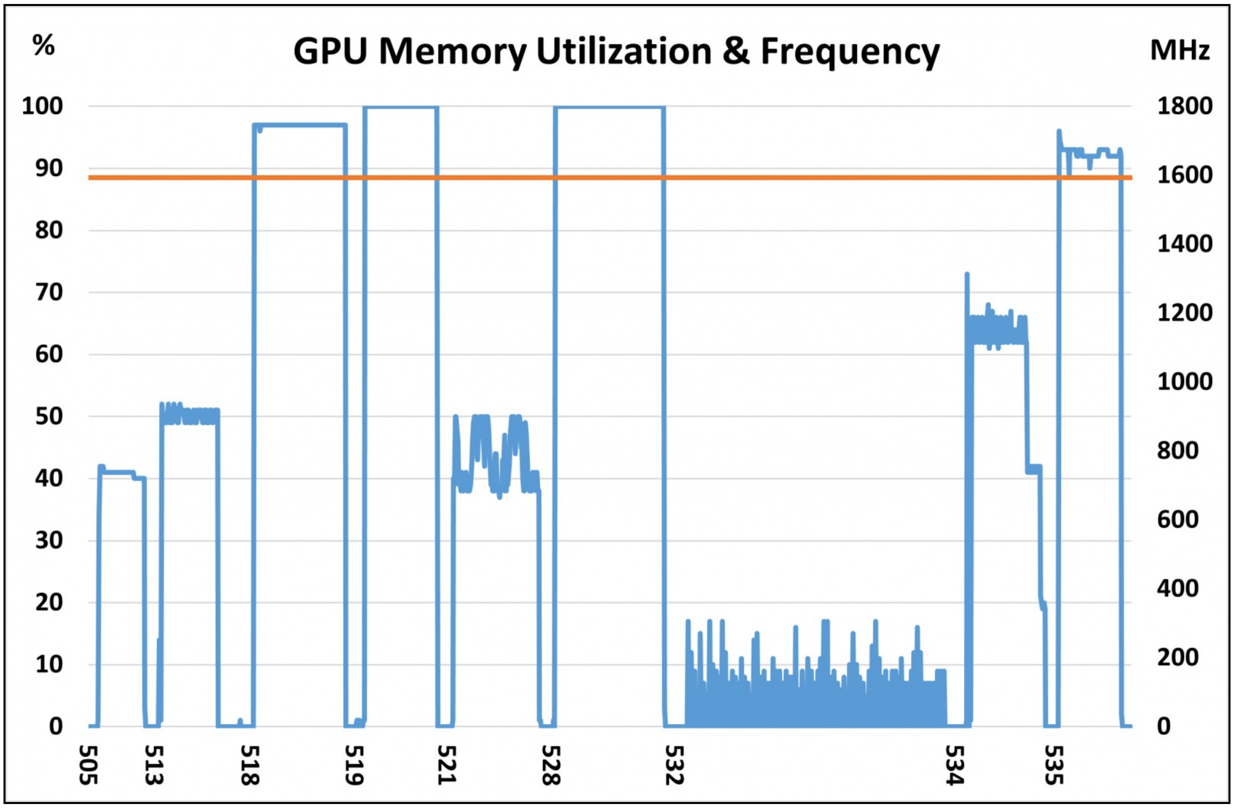


Figure 4. GPU memory Utilization and frequency of SPEChpc 2021

GPU power consumption

The chart below shows the power consumption without power level limitation, where the benchmark is able reach maximum 350W thermal design power (TDP) of the NVIDIA H100 GPU.

The 505, 513, 528 and 535 consume the maximum power during the benchmark execution, the 532's power consumption goes up and down during the run because of low parallel optimization, and the power level for other benchmarks are range from 260W to 330W.

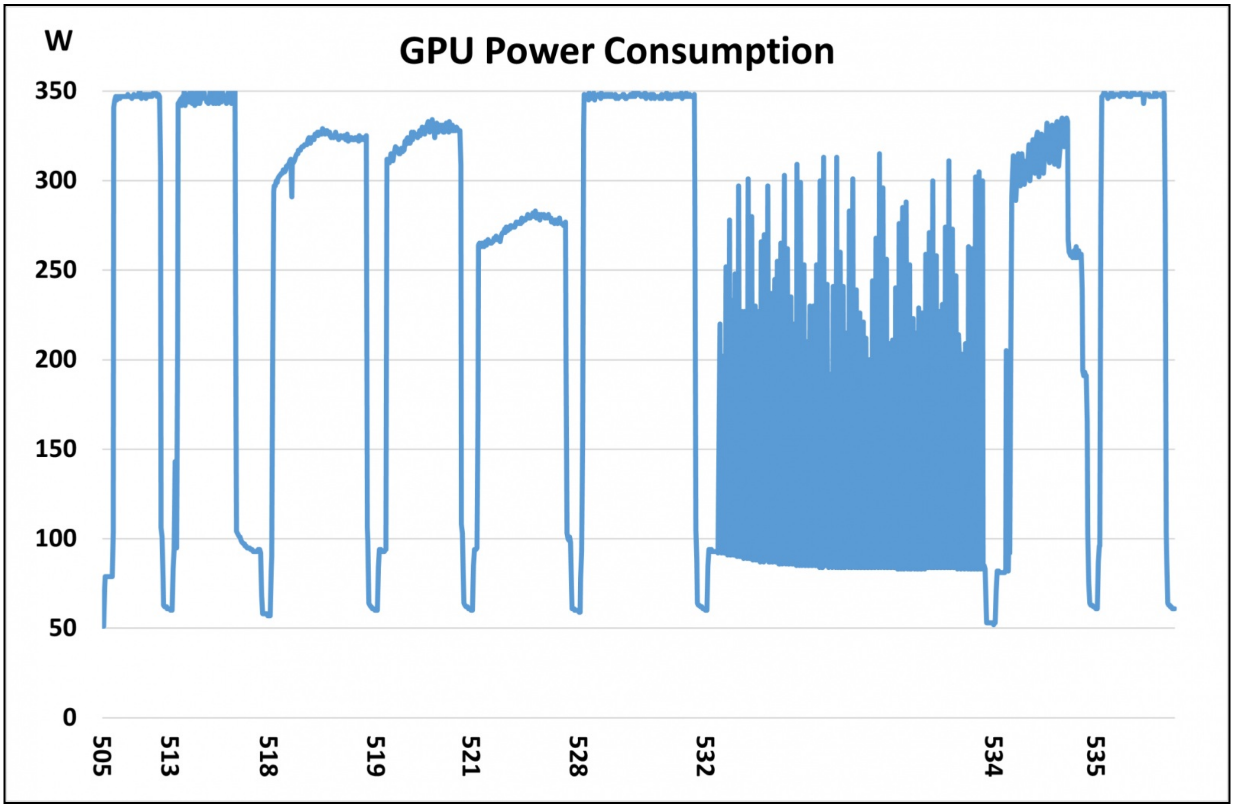


Figure 5. GPU power consumption of SPEChpc 2021

SPEChpc 2021 performance under different GPU power levels

The chart below illustrates related performance in percentage of each sub-benchmark of the SPEChpc 2021 under different power level of the H100 GPU from highest 350W to lowest 200W with 25W (about 7%) for each step.

The 505, 513, 528 and 535 are the most power sensitive sub-benchmark in the SPEChpc 2021 benchmark suite, 18% to 32% performance drop when power level set from 350W to 200W. At the other end of the scale, the 518, 519, 532 and 534 remain at least 87% of performance even if the power level drop 43%.

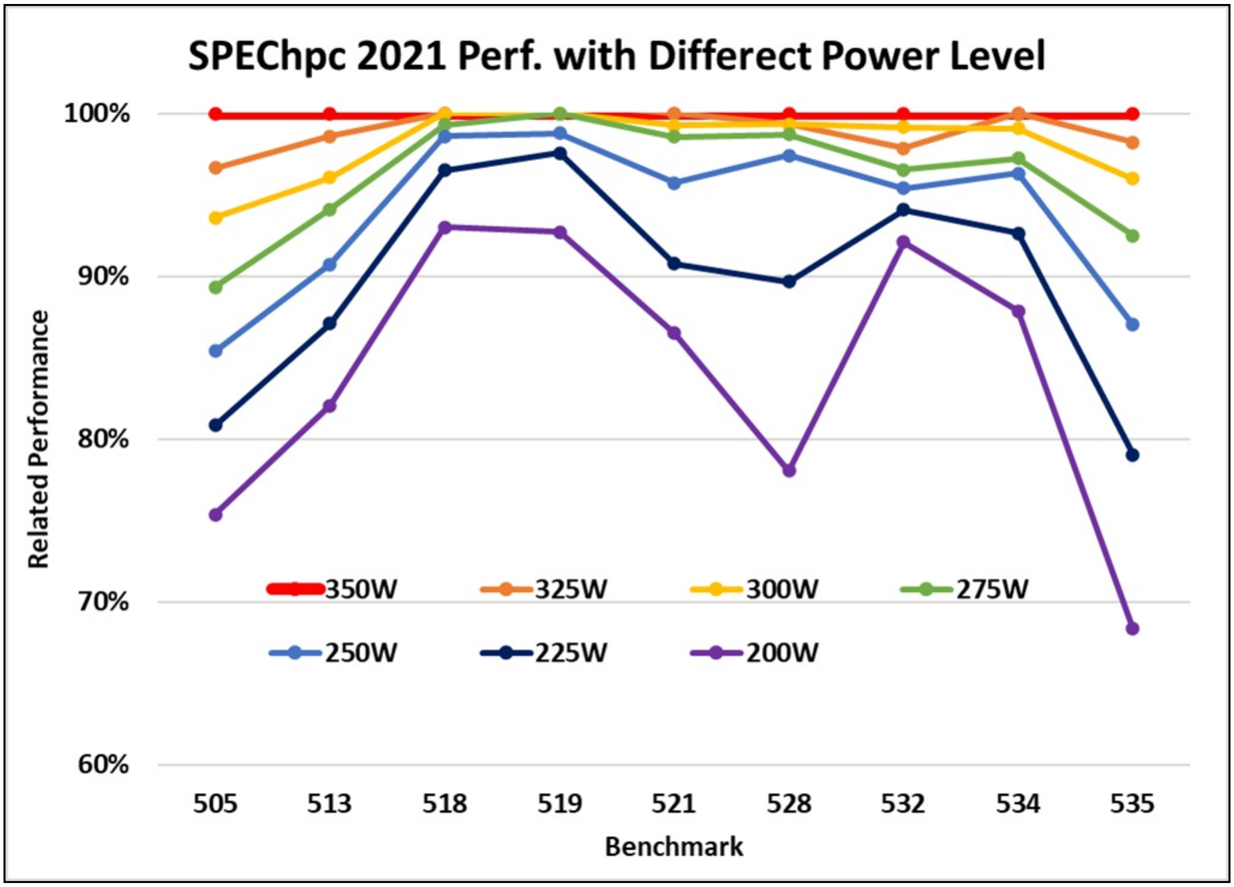


Figure 6. Relative performance of SPEChpc 2021 under different GPU power levels

Conclusion

The chart below combines the power level drop ratio and performance decreasing rate into one graph. Obviously, the slop of power level drop is much higher than performance drop, which means increasing the GPU TDP is not efficient way to improve the workload performance. In other words, a decrease in the TDP helps the GPU reach higher performance per watt if power consumption is more critical than absolute performance for the data center. It's important to find a balance between performance and power consumption to ensure optimal performance.

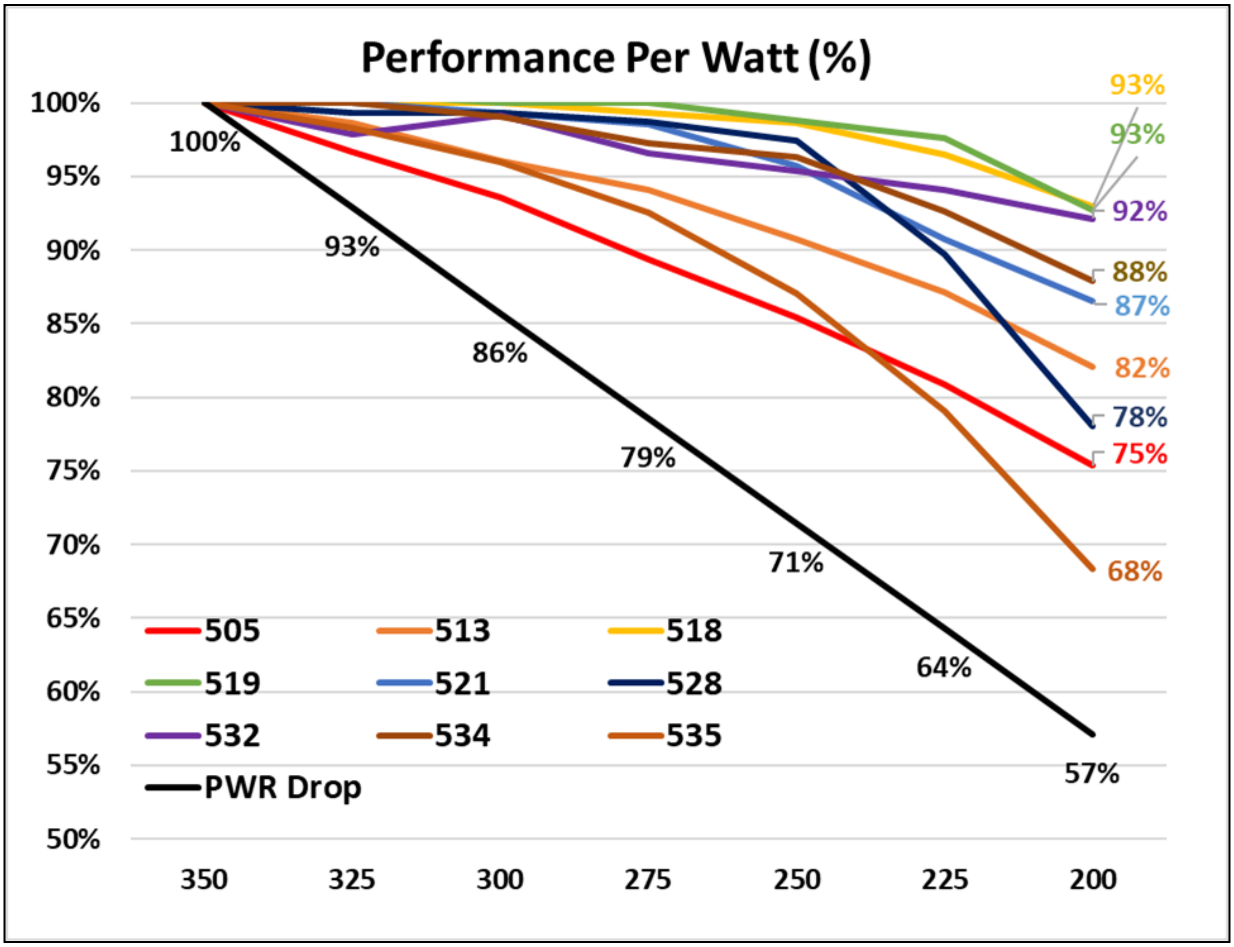


Figure 6. Power level drop ratio versus performance decrease rate

Authors

Jimmy Cheng is a performance engineer in the Lenovo Infrastructure Solutions Group Laboratory in Taipei Taiwan. Jimmy joined Lenovo in December 2016. Prior to this, he worked on IBM POWER system assurance and validation, ATCA system integration, automation development as well as network performance. Jimmy holds a Master's Degree in Electronic and Computer Engineering from National Taiwan University of Science and Technology in Taiwan, and a Bachelor's Degree in Computer Science and Engineering from Yuan-Ze University, Taiwan.

William Wu is a Principal Engineer and HPC system Architect in the Lenovo Infrastructure Solutions Group Laboratory in Taipei Taiwan. He has rich industrial experience including design digital IC, embedded system as well as high-density servers. His recently focus is on HPC system architecture research and development. William holds a Master's Degree from National Chung Hsing University.

Related product families

Product families related to this document are the following:

- GPU adapters
- SPEChpc Benchmark Results

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1706, was created or updated on March 13, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1706
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1706.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ThinkSystem®

The following terms are trademarks of other companies:

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

SPEC® and SPEChpc™ are trademarks of the Standard Performance Evaluation Corporation (SPEC).

Other company, product, or service names may be trademarks or service marks of others.