Lenovo EveryScale Design Architecture for WEKA Storage

Planning / Implementation

The technological advances of today have created a unique set of challenges for data-intensive applications, such as those used for artificial intelligence (AI), machine learning (ML), deep learning (DL) and high-performance computing (HPC) within, for example, financial analytics, genomics and life sciences. These complex applications require maximum IO performance, but legacy storage solutions were not built to handle the scale of these workloads.

The WEKA® Data Platform is uniquely built to solve the storage challenges of leading-edge applications. WEKA eliminates the complexity and compromises associated with legacy storage (DAS, NAS, SAN) while still providing the enterprise features and benefits of traditional storage solutions, all at a fraction of the cost. WEKA is designed to meet the stringent storage demands of data-intensive workloads and accelerates the process of obtaining insight from mountains of data.

This technical briefing accompanies the Lenovo HPC / AI EveryScale WEKA Storage solution brief and provides details on obtaining highest performance for the system.

Performance disclaimer: To demonstrate some of the technical considerations, some performance data is presented in this technical brief. These are not a commitment of performance for a Lenovo EveryScale WEKA Storage system. Performance will vary based on selection of NVMe drive and high-speed network configuration. Please contact your local Lenovo technical rep to discuss requirements and to assist in performance projections/requirements.

Client mount configuration

WEKA has very few tuning parameters to change the way in which it works, typically the small number of tuning parameters that are available are enabled using mount options when mounting the filesystem on a WEKA client node. The most significant mount options to consider are related to the number of cores which are dedicated for the WEKA client. These cores are used when using DPDK mode and give the highest performance for the client system.

The following graph shows the impact when a 100GbE connected client is configured using 0, 1 or 2 dedicated cores. Similar behavior (though with higher performance) can be seen when using 200GbE or HDR InfiniBand networks.

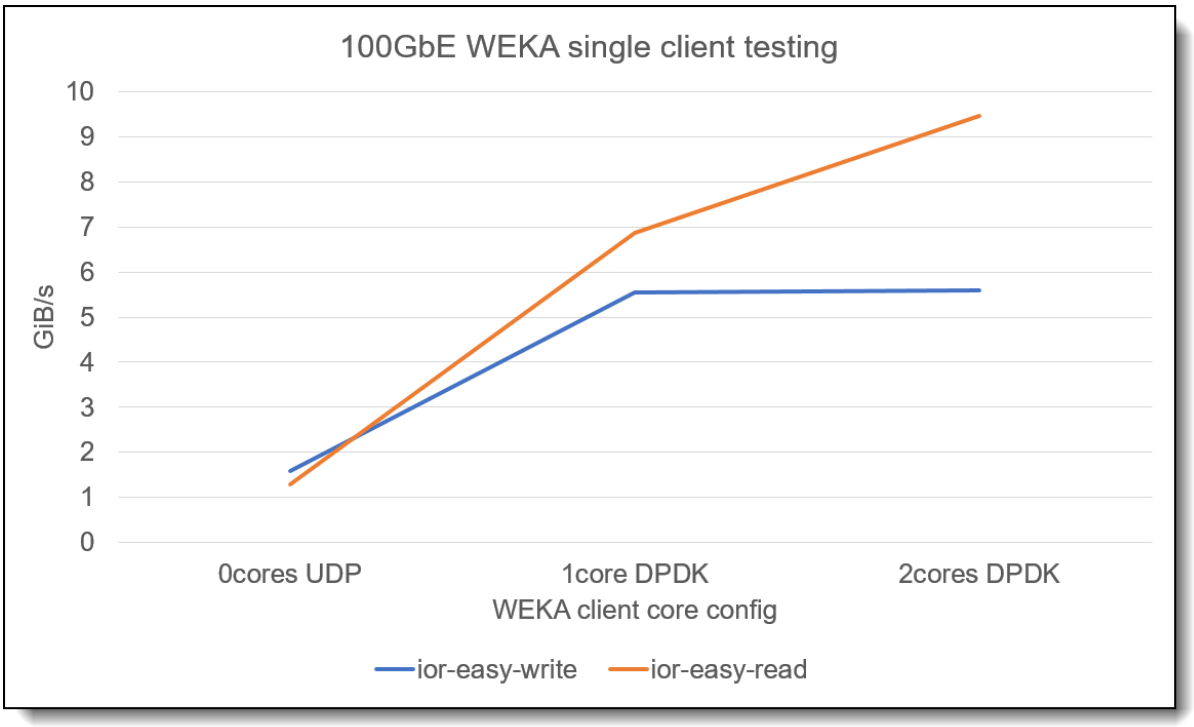

Figure 1. Single node performance impact with dedicated cores

From this graph, the use of dedicated client cores improves the performance of the WEKA client. The selection of 1 or 2 dedicated cores will depend on the bandwidth of the client network and on the IO requirements of the client workload. Typically, 2 client cores give a good choice

Once cores are dedicated to the WEKA client, they are not available for use in computational applications running on the client system. However, given high core count CPUs typically available in modern HPC systems, the requirement to dedicate cores for IO performance is unlikely to be an issue overall.

When using the WEKA client with SLURM as an HPC scheduler, it is necessary to inform SLURM that these cores are no longer available for computation work on the client. The following configuration parameters to set SLURM Core Specialization should be used in the slurm.conf file:

```
TaskPluginParam=SlurmdOffSpec
AllowSpecResourcesUsage=yes
```

The TaskPluginParam=SlrumdOffSpec instructs SLURM that the daemon process is run away from the specialized cores, the use of AllowSpecResourcesUsage permits users of the scheduler to use core-spec configuration in their client job script.

The following line use for the job script when submitting the job to the SLURM cluster:

```
#SBATCH --core-spec=2
```

This line instructs SLURM that two specialized cores are required on the compute node and prevents those cores from being attached to the CGROUP for the compute job or from being used for the slurmd processes. Internally SLURM uses a different core numbering mechanism to the output of lscpu, however by default the numbered highest core on CPU1 will be assigned first, followed by the highest numbered core on CPU0. To ensure consistency across different types of nodes, it is important to ensure that the WEKA client mount uses these core numbers. An example bash script to determine which cores to use:

```
core1=$(lscpu -a --extended | egrep ':0\s+' | tail -1 | awk '{ print $1 }')
core2=$(lscpu -a --extended | egrep ':1\s+' | tail -1 | awk '{ print $1 }')
declare -r CORES=core=${core1},core=${core2}
```

This example script determines the highest core numbers from CPU0 and CPU1 on a two-socket system, these core numbers are then used to mount the WEKA storage system, for example:

```
mount -t wekafs -o net=ib0,${CORES} weka0601/MyWEKAFS /mnt/MyWEKAFS
```

Finally, when using specialised cores with WEKA and SLURM, it is necessary to make a change to the WEKA client configuration file (/etc/wekaio/weka.conf) changing the value of isolate_cpusets to false. Once this is done, the weka-agent should be restarted with systemctl restart weka-agent.

An additional mount option that may be appropriate in some environments us the use of “forcedirect”. By default, the WEKA client software uses some caching and the forcedirect option disables all such caching. This can result in improved data throughput but can result in some impact to small IO or metadata operations. When using WEKA in a new environment, evaluation of performance using both with and without the forcedirect option should be used to identify the best configuration for the environment. For many HPC users, it is recommended to use the forcedirect client mount option.

Networking configuration

WEKA software can utilize multiple network adapters within a server with highest performance seen when RDMA enabled adapters are used. In the Lenovo EveryScale WEKA Storage solution, NVIDIA Networking VPI adapters are recommended. These devices are supported by WEKA and are also recommended for use in the client systems accessing the storage system.

When mounting the storage system, it is important that the RDMA enabled adapter name is passed to the WEKA client mount options. Similarly, the adapters should be attached to the WEKA containers running on the storage servers. WEKA currently supports at most 2 RDMA adapters in the storage servers, for maximum performance when used with the Lenovo ThinkSystem SR630 V2 server, these should be placed in Slot 1 and Slot 3. This ensures that one adapter is attached to each NUMA domain within the server.

Where an object backend is used with WEKA, or where multi-protocol access (e.g., NFS, SMB or Object) is required, a third adapter installed into Slot 2 may be used to provide dedicated network connectivity for those services.

Storage performance

With a six-node cluster using the Lenovo ThinkSystem SR630 V2 and using HDR InfiniBand as the Interconnect, performance can be expected to be as high as 5.1 million read IOPs and up to 228GB/s of streaming read performance.

Topics in this section:

UEFI firmware settings

To achieve maximum performance for WEKA on the Lenovo ThinkSystem SR630 V2 server, it may be necessary to adjust some UEFI settings. Change the system Operating Mode to “Maximum Performance” and then reboot the system. The Operating Mode is a UEFI macro which automatically changes several settings to give maximum performance.

When using the Lenovo OneCLI tool to change UEFI settings, set:

```
OperatingModes.ChooseOperatingMode: Maximum Performance
```

It is essential that the system is rebooted once the Operating Mode has been changed before proceeding to change other firmware settings.

Once the Operating Mode has been changed and the system rebooted, the following UEFI settings should also be changed:

```
Processors.HyperThreading: Disabled
DevicesandIOPorts.SRIOV: Enabled
DiskGPTRecovery.DiskGPTRecovery: Automatic
Power.WorkloadConfiguration: I/O sensitive
```

Kernel boot parameters

For best performance on the Lenovo ThinkSystem SR630 V2 system the following boot-time kernel parameters are recommended to be set:

```
clocksource=tsc tsc=reliable nomodeset intel_iommu=off intel_idle.max_cstate
=0 processor.max_cstate=0 numa_balancing=disable
```

These can be set using the grubby tool, for example:

```
grubby –update-kernel ALL –args="clocksource=tsc tsc=reliable nomodeset inte
l_iommu=off intel_idle.max_cstate=0 processor.max_cstate=0 numa_balancing=di
sable"
```

The following configuration file should also be created to change the behaviour of the i2c driver and prevent issues with WEKA when accessing NVMe drives using SPDK.

/etc/modprobe.d/i2c_i801.conf:

```
# disable IRQ reservation on i801_smbus driver
options i2c_i801 disable_features=0x10
```

Once set, the server should be rebooted to take effect.

WEKA Server Configuration

The Lenovo ThinkSystem SR630 V2 server is a 2-socket platform. Performance engineering in the Lenovo HPC Innovation center has shown that performance can be significantly boosted using multi-container backends with WEKA. In this configuration, two WEKA containers are deployed per server, each with 15 cores allocated. As the distribution of the NVMe drives in the Lenovo ThinkSystem SR630 V2, it is essential that core IDs are mapped to the container and that drives attached to the NUMA node are also mapped to those cores.

Where a single container configuration is used or consideration of core and drive allocation are not made, it is unlikely that the maximum performance of the system can be achieved. Lenovo Professional Services are available to deploy and configure the Lenovo EveryScale WEKA Storage system. Tuning of core allocations for frontend and compute cores requires some careful consideration as the selection of these values impacts performance of metadata and multi-protocol access and the exact configuration will depend on the customer deployment.

Streaming data transfer size

In many HPC filesystems, the concept of block-size is important to consider, and streaming IO should be aligned to ensure data transfers align with the filesystem block size. With WEKA, there is no concept of filesystem block size and the impact of using different IO transfer sizes for streaming data is negligible. The following graph shows an analysis using the IOR benchmark tool and shows the impact of changing the transfer size when running multiple threads over multiple client nodes:

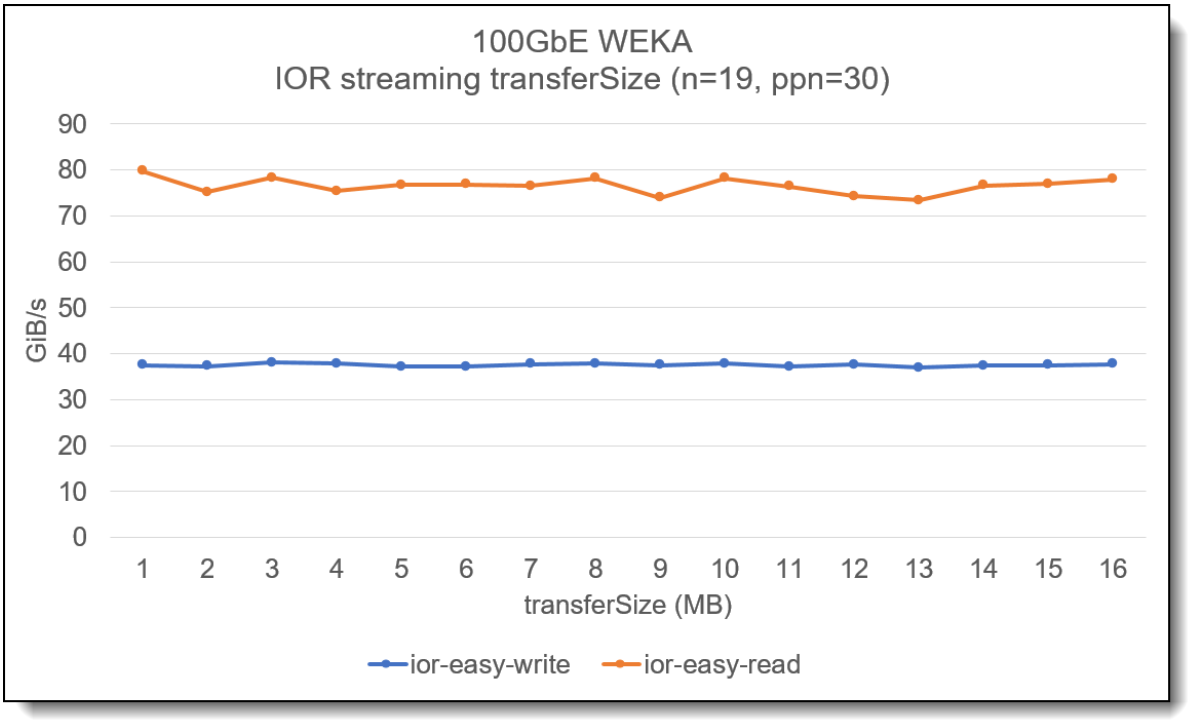

Figure 2. Impact of changing transfer size for streaming IO

Note: This graph is not intended to show maximum performance of the system but is used to demonstrate the impact of changing the transfer/IO size when performing streaming reading or writing to the Lenovo EveryScale WEKA Storage solution.

Single node streaming performance

As with many parallel filesystems, to obtain maximum single client performance, it is necessary to increase the number of threads on the client system before maximum performance is obtained.

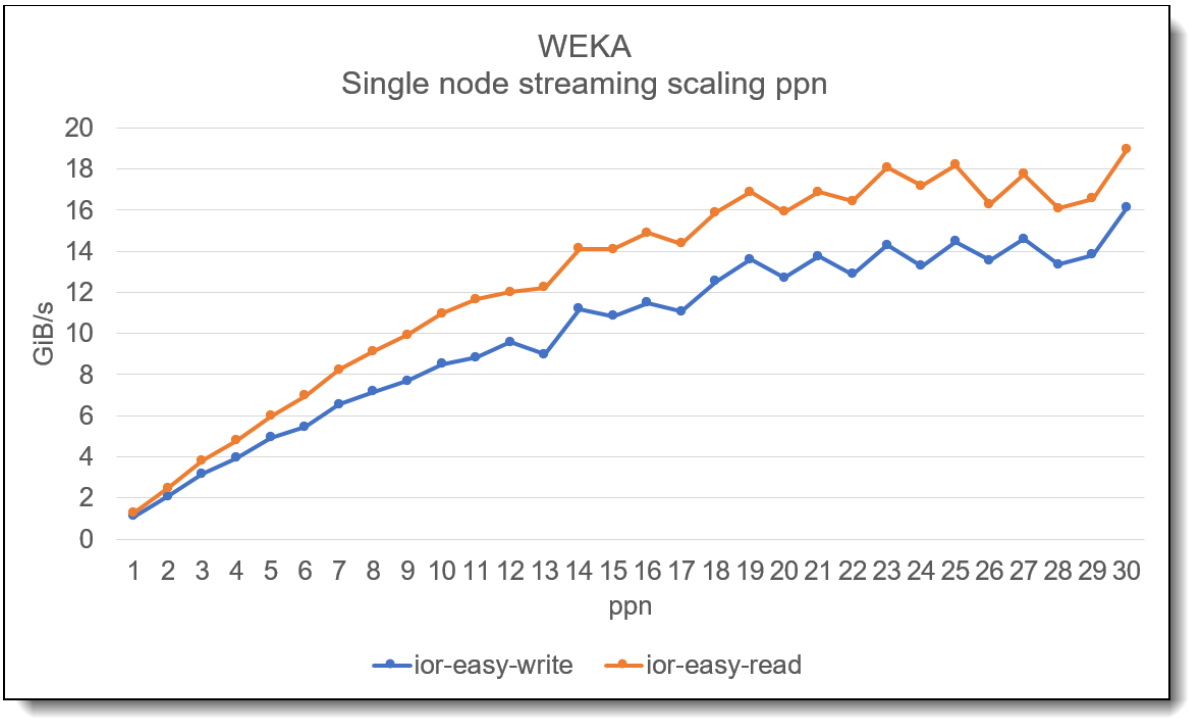

Figure 3. Single node performance scaling with increased thread counts

The graph above shows how increasing the number of threads when using IOR on a single client will scale. In the tests depicted above, a single iteration of IOR is used to show overall scaling. Where multiple iterations are used, it is likely the line would appear less “bumpy”.

In addition to high performance streaming data, WEKA also provides excellent random 4K performance.

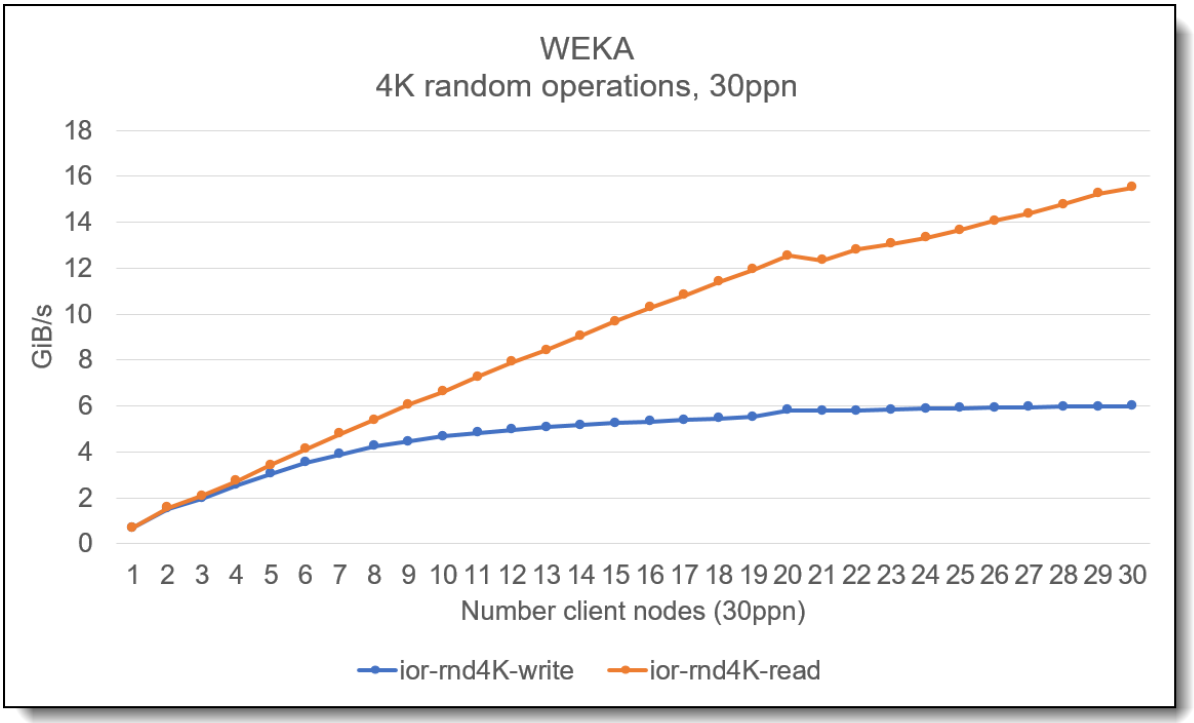

Figure 4. 4K random IO operations

The graph above shows the performance of the storage system when increasing the number of client nodes with each node running 30 threads. The graph shows that 4K random read operations scale well as the number of client nodes increases. Whilst the graph shows aggregate 4K random bandwidth, this is typically an “IOPS” bound test and shows that WEKA is capable of high IOPS performance.

Summary

This technical brief is intended to provide some insight into WEKA scaling and performance capabilities and to provide some advice on approaches to configuring the cluster and client mount options.

WEKA can provide both high performance streaming bandwidth and high-performance random IO workloads. Typically, once a Lenovo EveryScale WEKA Storage system has been deployed, there are very few tuning parameters that need to be considered and with a small amount of analysis of workload, the optimal configuration options can quickly be determined.

For more information, see the solution brief, Lenovo High Performance File System Solution with WEKA Data Platform:
https://lenovopress.lenovo.com/lp1691-lenovo-high-performance-file-system-solution-with-weka-data-platform

Authors

Simon Thompson is a senior manager for HPC Storage & Performance and leads the HPC Performance and Benchmarking team within the WW Lenovo HPC team. Prior to joining Lenovo, Simon spent 20 years architecting and delivery research computing technologies.

Steve Eiland is the Global HPC/AI Storage Product Manager responsible for the integration of all storage products for the HPC/AI Lenovo community. He is focused on addressing vertical storage markets and providing TOP performance, cutting edge, storage solutions within the global Lenovo HPC team. Prior to joining Lenovo, Steve worked in various technology related companies as a Field Application Engineer, Sales Engineer, and Business Development Manager, supporting products in the areas of storage, storage interconnect, IPC (inter processor communication), and compute.

Related product families

Product families related to this document are the following:

- Software-Defined Storage

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, LP1700, was created or updated on February 27, 2023.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1700
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1700.

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:

Interconnect® is a trademark of IBM in the United States, other countries, or both.

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