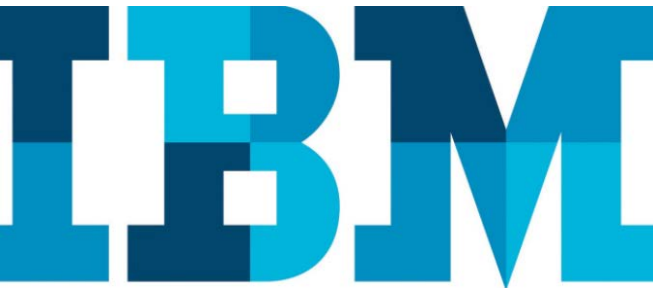
May 2020

Driverless AI on IBM Power Systems

IBM POWER9 processor-based server capacity planning guide

Table of contents



Driverless AI is a solution for data scientists and data analysts to conduct artificial intelligence (AI) projects faster and more efficiently by using automation and state-of-the-art computing power to accomplish AI-related tasks in minutes or hours that normally can take months for humans to complete. IBM® POWER9™ processor-based servers with GPUs accelerate Driverless AI. This white paper provides guidance for server sizes and configuration options based on the expected usage.

Single-server capacity planning

Among the many IBM POWER9 processor-based servers that support GPUs, IBM Power® System AC922 server is considered to be the best matched server for Driverless AI. Table 1 shows three configuration sizes: Minimum, growth, and maximum. In general, the number of simultaneous users determines the hardware configuration required for running Driverless AI on IBM Power servers. The Power AC922 server (based on the POWER9 processor) and memory features are not upgradable in the field; thus, it is essential to choose the correct configuration when making your initial system order.

Hardware	Minimum configuration	Growth configuration	Maximum configuration
POWER9 server • Type-Model	IBM Power AC922 • 8335-GTH • Air-cooled	IBM Power AC922 • 8335-GTH • Air-cooled	IBM Power AC922 • 8335-GTX • Water-cooled
Processors, cores	Two 16-cores	Two 20-cores	Two 22-cores
Memory	256 GB	Up to 2 TB	Up to 2 TB
Storage	1 960 GB SSD + 1 960 GB SSD for redundancy	Add up to two NVMe cards	Add up to two NVMe cards
Network adapter	One 10 GbE	Two or more 10 GbE	Two or more 10 GbE
GPUs: NVIDIA Tesla V100 GPU Accelerators with NVLink	Two	Four	Six (requires water cooling)
Number of simultaneous users	Two	Four	Six

Table 1. Driverless AI single-server hardware configuration

Minimum single-server configuration

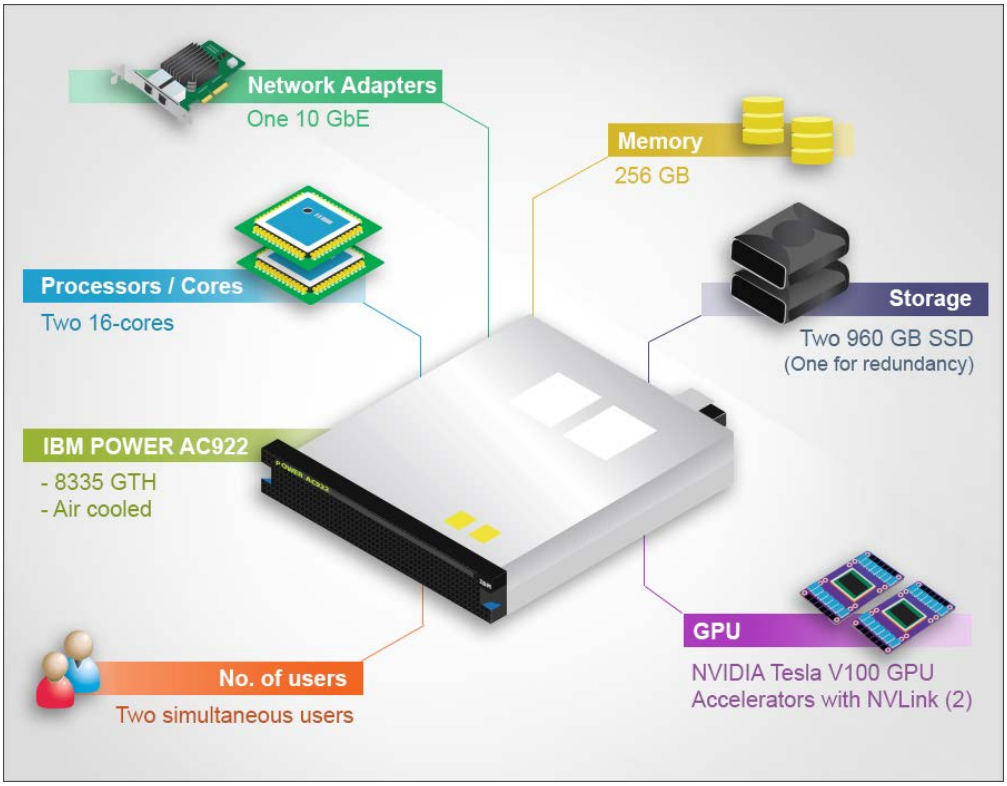


Figure 1. Minimum single-server configuration

The minimum single-server configuration is essentially the smallest Power AC922 configuration that you can order. The minimum configuration can support approximately two simultaneous users. However, the actual number of users will vary based on the type of user (single user or super user), degree of exclusivity to GPUs, the type of data experiment (that is, basic, real-time, or compute-intensive workloads) being run, and time-to-completion needs. This minimum configuration offers one 16-core CPU and one GPU for processing power to each of the two users. One 960 GB TB solid-state drive (SSD) is recommended in addition to a second 960 GB SSD for data redundancy.

Growth single-server configuration

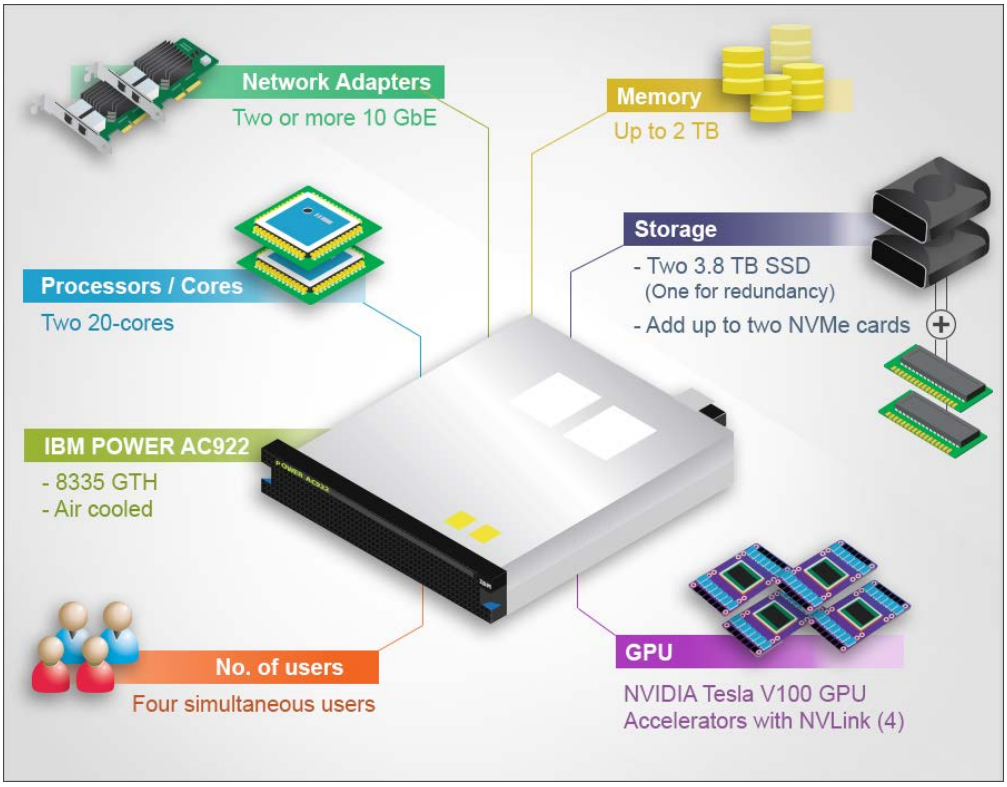


Figure 2. Growth single-server configuration

The growth single-server configuration uses the same Power AC922 server type and air-cooled model as the minimum configuration. The growth configuration provides the ability to choose the configuration that is needed to meet future growth needs within the air-cooled server type.

The growth configuration offers more processor cores and GPUs along with more memory, storage, and network adapters. The Power AC922 server has a total of four PCIe card slots across storage and network. You can add up to two NVMe cards for additional performance based on the data set size and data ingestion requirements. The recommended disk space is 10 times the data set size. Two network adapters are recommended. However, as an alternative, if only one network adapter is needed, a third NVMe card could be added.

The growth configuration can support approximately four simultaneous users. However, the actual number of users will vary based on the type of user, the type of data experiments being run, and the actual configuration features chosen. The growth configuration offers each of the users one GPU for processing power.

Maximum single-server configuration

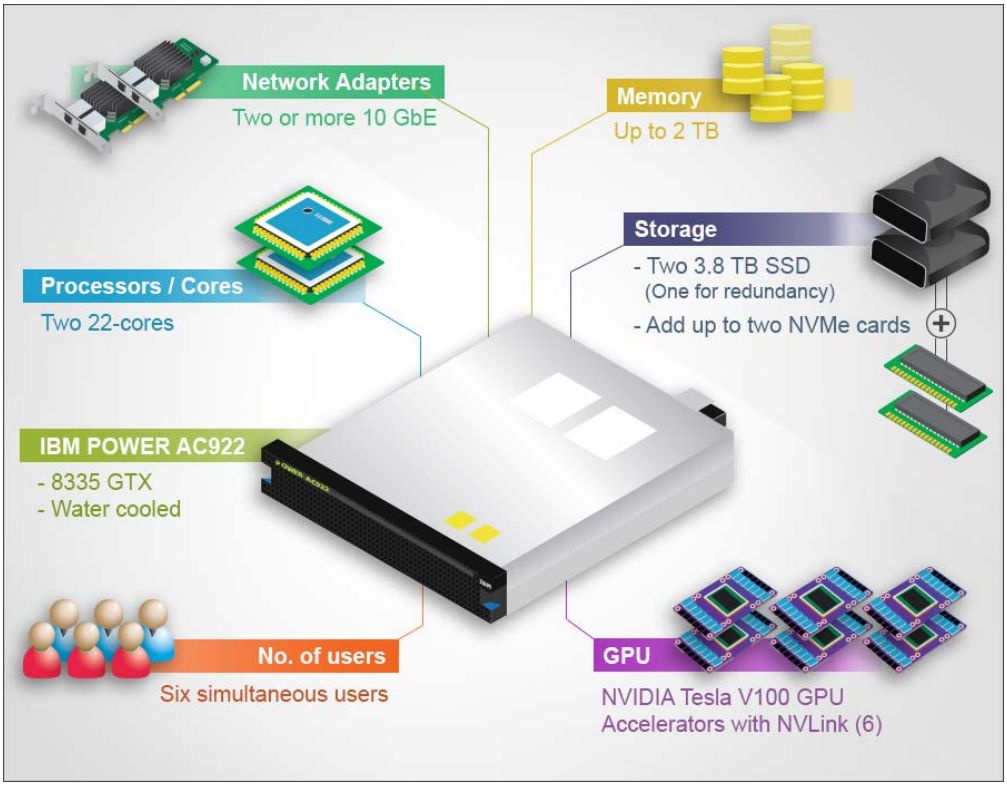


Figure 3. Maximum single-server configuration

The maximum single-server configuration uses the same Power AC922 server type, but requires the water-cooled server model because of the number of GPUs and the heat energy they produce when running. If your data center has water cooling capabilities, you can consider this configuration to maximize your AI capabilities within a single-server configuration.

The maximum configuration offers the most processor cores and GPUs along with more memory, storage, and network adapters. The Power AC922 server has a total of four PCIe card slots across storage and network. You can add up to two NVMe cards for additional performance based on the data set size and data ingestion requirements. The recommended disk space is 10 times the data set size. Two network adapters are recommended. However, as an alternative, if only one network adapter is needed, a third NVMe card could be added.

The maximum configuration can support approximately six simultaneous users. However, the actual number of users will vary based on the type of user, the type of data experiments being run, and the actual configuration features chosen. This maximum configuration offers each of the users one GPU for processing power.

Summary

Driverless AI is supported on Red Hat Enterprise Linux (RHEL) 7.5 for Power LE (POWER9) on IBM Power AC922 servers and other IBM POWER9 processor-based servers that support GPUs. Several configuration size options are available depending on the number of simultaneous users running data experiments. Larger or more powerful configuration features may be required as users become more advanced and data experiments become more compute-intensive. You can select the appropriate configuration size based not only on current requirements but also how usage might grow over time. The Power AC922 server processor and memory features are not upgradable. So, it is key to estimate future usage per server when making purchase decisions.

Get more information

- IBM Power System AC922
- IBM and H2O.ai Strategic Partnership
- H2O.ai Driverless AI
- H2O.ai Driverless AI installation instructions for IBM Power
- Welcoming H2O Driverless AI to the PowerAI ecosystem
- IBM and H2O.ai combine forces to provide machine learning on IBM Power Systems

About the author

Beth L. Hoffman is an Executive IT Specialist and Solution Architect in the IBM Cognitive Systems ISV Ecosystem Technical Development organization. She leads the enablement of AI and analytics solutions on IBM Power Systems for the past 10 years. Beth has more than 15 years of experience consulting and collaborating with key software solution companies. You can reach Beth at bethvh@us.ibm.com or www.linkedin.com/in/bethhoffmanibm

IBM Systems
3039 Cornwallis Road
RTP, NC 27709

Produced in the United States of America

Please recycle