Redbooks
ibm.com/redbooks

Cognitive Computing

Featuring the IBM Power System AC922

Ivaylo Bozhinov
Boran Lee
Gustavo Santos

Analytics

Power Systems

IBM

Redpaper

IBM Redbooks

Cognitive Computing Featuring the IBM Power System AC922

October 2019

REDP-5555-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (October 2019)

This edition applies to IBM Power System AC922 models GTH and GTX for Cognitive Solutions.

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION 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. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®
IBM®
IBM Cloud™
IBM FlashSystem®
IBM Spectrum®
IBM Spectrum Conductor®
IBM Spectrum Storage™
IBM Watson®
LSF®
Power Architecture®
POWER8®
POWER9™
PowerHA®
PowerVM®
Redbooks®
Redbooks (logo)  ®
Watson™

The following terms are trademarks of other companies:

Intel, Intel Xeon, Intel logo, Intel Inside logo, and Intel Centrino logo are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

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

Preface

This IBM® Redpaper publication describes the advantages of using IBM Power System AC922 for cognitive solutions, and how it can enhance clients' businesses.

In order to optimize the hardware and software, IBM partners with NVIDIA, Mellanox, H2O.ai, SQream, Kinetica, and other prominent companies to design the Power AC922 server, specifically enhanced for the cognitive era. Most of its outstanding hardware features, such as NVIDIA NVLink 2.0 and PCIe 4.0, are described in this publication to illustrate the advantages that clients can realize in comparison with IBM competitors.

We also include a brief description about what cognitive computing is, and how to use IBM Watson® Machine Learning cognitive solutions to bring more value to your business ecosystem. Additionally, we show performance charts that show the advantages of using Power AC922 versus x86 competitors. In the last chapter, we describe the most remarkable use cases in which IBM solves real problems using cognitive solutions.

This IBM Redpaper publication is aimed at IT technical audiences, especially decision-making levels that need a full look at the benefits and improvements that an IBM Cognitive Solution can offer. It also provides valuable information to data science professionals, enabling them to plan their modeling needs. Finally, it offers information to the infrastructure support group in charge of maintaining the solution.

Authors

This paper was produced by a team of specialists from around the world working at IBM Redbooks, Austin Center.

Ivaylo Bozhinov is a technical support professional for the Power System hardware division in Sofia, Bulgaria. He has been with IBM since 2015, and has more than 8 years of experience in IT. He holds a degree in Information Technology from the State University of Library and Information Technology. His main focus is on AI, Big Data, Deep Learning solutions, and IBM Hybrid Cloud.

Boran Lee is a Client Technical Specialist for Cognitive Systems at IBM Systems in Korea. She has 9 years of experience in the field of IT. She holds a degree in Computer Science from Soongsil University. She has 6 years of working experience as a technical sales representative on Power Systems and IBM AIX®, VIOS, IBM PowerVM®, PowerVC, and IBM PowerHA® in multiple accounts in Korea. However, she is also expanding her career into Cognitive Systems, including Power AC922, Linux, and IBM Spectrum® Computing. Her areas of expertise include HPDA, HPC, and modern data platforms.

Gustavo Santos is an IBM Power Systems Consultant and Business Development Manager at IBM. He has been with IBM since 1997. He has 21 years of experience in IBM Power Systems, Cognitive Solutions, and Cloud Architecture. He holds a degree in Systems Engineering from Universidad Abierta Interamericana. During the last 4 years, he worked as a Power Systems Consultant, and during the last year he worked as a Business Development Manager, to create new offers for clients and add value to the IBM Solutions portfolio.

The project that produced this publication was managed by:
Scott Vetter, PMP

Thanks to the following people for their contributions to this project:

Mariano Batista, Carlos Cabañas, Scott Campbell, Glen Corneau, Luis F Armenta Garza, Tom Heller, Yesenia Jimenez, Anto A John, Rajaram B Krishnamurthy, Andrew Laidlay, Nin Lei, Atit Patel, Brandon Pederson, Amanda J Quartly, Marcos Quezada, Steven Roberts, Bill Terry, Maria Ward

Now you can become a published author, too

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time. Join an IBM Redbooks® residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us.

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form:

  ibm.com/redbooks

- Send your comments in an email:

  redbooks@us.ibm.com

- Mail your comments:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction to cognitive computing

This publication describes the IBM definition of cognitive computing. It also covers some brief technical aspects of both software and hardware, focusing mainly on the value add and performance improvements that an IBM Power System AC922 server can offer.

This chapter includes the following topics:

1.1 Definition of cognitive computing

Cognitive computing is a new way to solve human problems by bringing technology closer to the way humans handle information, using it to perform analysis and make inferences to finally deliver results, and then using those results, and their level of accuracy, as feedback to improve the whole process.

Cognitive computing addresses the traditional problems that humans have, but changes the method of addressing them. The old approach was trying to rethink the problem in a series of steps that can then be systematically and repetitively programmed to get an always correct and true result.

Now, cognitive systems approach the human method of solving those problems, using similar mechanisms to learn, obtain the right information, analyze it, and provide an answer. In this case, although the answer is correct, it is more importantly the most trusted one possible.

In this learning process, the context takes a relevant value, and substantially modifies the results for the same problem. This shows a fundamental feature of cognitive systems, which is that they are *contextual*. This feature is what enables these systems to vary the responses depending on the time, place, or population of study, even for the same problem or hypothesis.

At the same time, cognitive systems must be highly *adaptive*, and learn quickly from changes in the environment. This adaptability is another fundamental characteristic of cognitive systems. Furthermore, adaptability is linked to another key aspect of cognitive computing, which is that it is *iterative* or repetitive.

Systems that learn and reformulate answers based on problem, context, and environment should ask the correct questions, and repeat their learning processes when the problems raised are fuzzy or ambiguous. This iteration should take into account the previous results and responses, and use them comparatively to improve the capabilities of the cognitive system and its learning model.

Finally, cognitive computing must be *interactive*, because it inevitably needs to connect and bond with the environment, whether other cognitive and traditional services or systems, and must also interact with humans to obtain or provide answers and results.

This interactivity also changes radically from traditional systems. Cognitive systems approach human behavior, and commonly offer mechanisms of capturing and obtaining information similar to that used by people. For example, a cognitive system could use natural language processing to interpret what we talk about, the context in which we say something, and even the mood or the voice tone.

This is one of the reasons why cognitive systems are linked to methods of connecting with humans that are closer to the natural process of people than to the natural process of systems. A wide range of techniques are used to capture information through cameras, videos, written or spoken dialogue, body language interpretation, emotional capture, and many other methods of interaction typical of humans.

The paradigm shift is evident as the way that the process links with cognitive systems also changes. The process becomes more natural, and it is the cognitive solution that approaches the human method and not the human who has to rethink the problem in order to adapt and code it in a program that solves it by systematic and predefined steps.

Note: IBM participated in creating a standard definition for cognitive computing as a member of the cognitive computing consortium. For more information and to read this definition go to:

https://cognitivecomputingconsortium.com/resources/cognitive-computing-defined/#1467829079735-c0934399-599a

1.2 What is IBM cognitive computing

Based on the previous definition, cognitive computing is a new way to solve human problems from real life by trying to use the same tools that humans use. Cognitive computing uses these tools to understand, reason, learn, and interact, then to give a solution, feedback, and learn again.

Cognitive systems differ from current computing applications in that they move beyond tabulating and calculating based on preconfigured rules and programs. Although they are capable of basic computing, they can also infer and even reason based on broad objectives.

Therefore, a cognitive solution provides help and answers to problems in different ways, and we will not always receive the same answer for the same question. As in other aspects of our lives, answers rely on a broad number of variable conditions. These conditions change our point of view and consequently our answers.

Cognitive computing systems and solutions pursue the same objectives, and they are prepared to learn in the same way, as a human does. However, cognitive computing has the advantages of processing speed, storage capacity, and broadband connections.

IBM cognitive computing systems are empowered and optimized to give the maximum performance when used to process new cognitive models. Power AC922 is a cognitive system with the most current capabilities. Modeling and learning that takes hours in other systems can take only minutes or even seconds in this super high-performing cognitive computing server.

There is also a family of software products based on open source code, and IBM optimized libraries and applications to maximize server capacity utilization. Whether you are running deep learning or machine learning, artificial intelligence or natural language processing models, this is the best server for those workloads.

Cognitive computing is here, but it is not designed to replace humans. Rather, it will empower humans by giving them augmented cognitive capabilities.

1.3 IBM cognitive solutions

Deep learning is a highly complex method of learning that is used by cognitive systems to analyze extremely large amounts of information, and produce much more accurate models. Deep learning is a subset of cognitive learning.

Machine learning is part of the learning methods of cognitive systems, and also is a subset of the artificial intelligence methods currently in use. It uses other libraries and mathematical models that differ from those used in deep learning processes.

Artificial intelligence is what we understand as the complete process of obtaining, ingesting, analyzing, and presenting a response to a set of information, by using a model trained for that purpose. In these cases, the model presents us with certain results with a precision value defined by the model that we have trained. This varies depending on the type, the number, and the context of the samples.

The artificial intelligence then produces results with a certain known level of accuracy, and displays these results along with the deviation and accuracy of the answer.

Finally, cognitive systems combine all this and add more complex models that, in addition to getting answers, weigh the results and offer us the best option within the spectrum of possible solutions to our problem. To perform these tasks, *cognitive computing* uses methods that are much more similar to the human process to interpret information and add variable, contextual conditions to the final results. This context impacts the chosen answer.

Figure 1-1 shows the cognitive computing components.

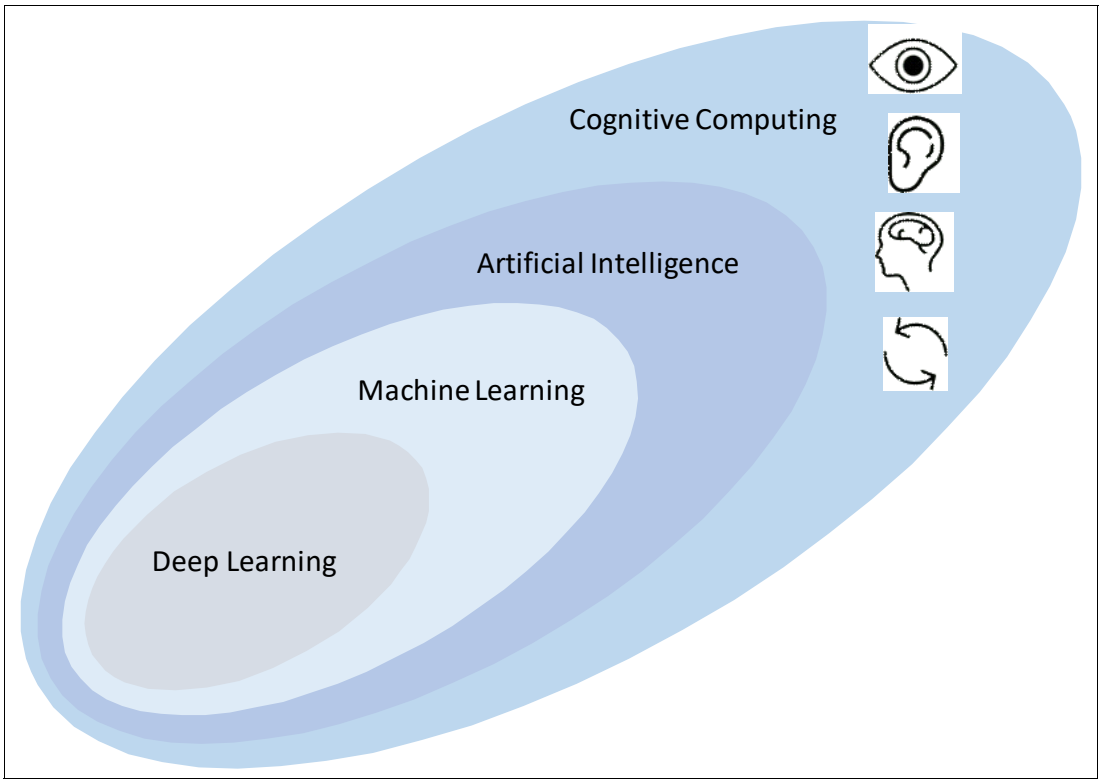

Figure 1-1 Cognitive computing, artificial intelligence, machine learning, and deep learning

1.3.1 Watson Machine Learning

Watson Machine Learning is a complete implementation of a cognitive computing system, ready to build, train, and deploy machine learning and deep learning models. Watson can be deployed either from an automated training process or from manual and completely controlled model creation.

Watson Machine Learning software is based on open source software. However, it is enhanced to be easy to use and mainly to take advantage of the Power AC922, currently the most powerful hardware to run IBM Cognitive Solutions.

Figure 1-2 on page 5 shows a summary of the Watson Cognitive software solutions.

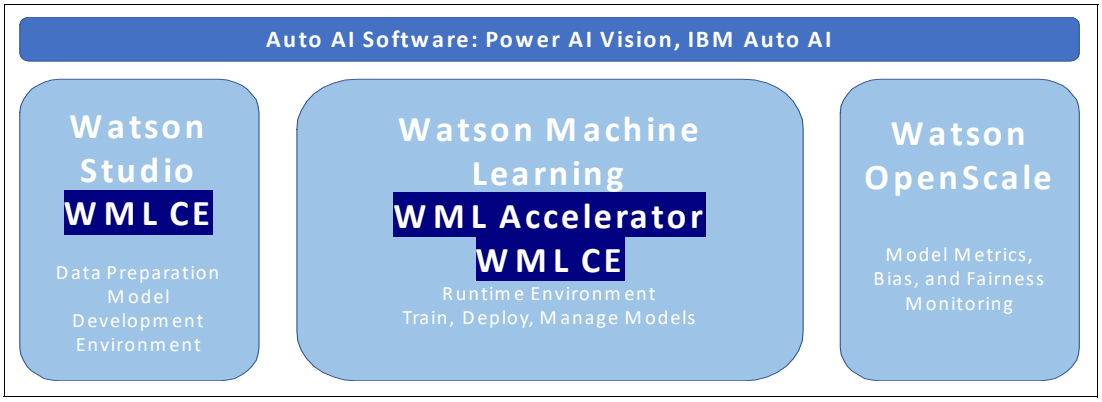

Figure 1-2 Watson Cognitive software solutions

This publication is focused on the Watson Machine Learning software solutions, specifically designed and improved to run on the Power AC922 server. Watson Machine Learning is available in two different bundles:

- Watson Machine Learning Accelerator
- Watson Machine Learning Community Edition

Note: For more information about Watson Machine Learning Accelerator, see an overview:
https://www.ibm.com/support/knowledgecenter/SSFHA8_1.2.1/wmla_overview.html

See the following release for more information about Watson Machine Learning Community Edition:
https://developer.ibm.com/linuxonpower/deep-learning-powerai/releases/

1.3.2 IBM PowerAI Vision

IBM PowerAI Vision (PowerAI Vision) is an image and video analysis tool. It uses the most advanced deep learning tools and techniques to perform image and video analysis. You can use PowerAI Vision to easily label images or videos and use this tags to find valuable information.

The following are the main features of PowerAI Vision:

- Streamlined model training

  You can use existing models that are already trained as a starting point to reduce the time required to train models and improve trained results.

- Single-click model deployment

  After you create a training model, you can deploy an API with one click. You can also develop applications based on the model that you deployed.

- Data set management and labeling

  You can manage data that is not labeled, and label data.

- Video object detection and labeling assistance

  Videos that you import can be scanned for objects, and the objects can be automatically labeled.

Note: For more information about IBM PowerAI Vision, see the following vision site:
https://www.ibm.com/us-en/marketplace/ibm-powerai-vision

1.4 Third party cognitive solutions

There are other cognitive solutions from-third party providers in the artificial intelligence, deep learning, and machine-learning ecosystem. The solutions listed in the next sections are available to run on the Power AC922, using the IBM POWER9™ processor, the hardware stack, and the GPU improvements on it. See the following list for third-party cognitive solutions to run on the Power AC922 server:

- H2O Driverless AI (3.2.1, “H2O Driverless AI and Power AC922” on page 30)
- Sqream DB (3.2.2, “SQream DB and Power AC922” on page 31)
- Kinetica (3.2.3, “Kinetica and Power AC922” on page 33)

1.5 Power AC922 end-to-end

Modern artificial intelligence, high-performance computing, and analytic workloads are driving an ever-growing set of data-intensive challenges, that can only be met with accelerated infrastructure. To help meet these demands, IBM designed the Power AC922.

Power AC922 is a system that not only provides the hardware, but also includes all the software stack needed for data science professionals. Power AC922 is specifically optimized for IBM Watson Machine Learning Accelerator, which is a set of software solutions, such as IBM Watson Machine Learning Community Edition, IBM Spectrum Conductor®, and IBM Spectrum Conductor Deep Learning Impact. Power AC922 provides an end-to-end cognitive platform for data science professionals, adding IBM premium support for the whole stack, including the open source frameworks.

IBM is committed to bring the best solutions and top innovations to clients. For several years, IBM Power Systems has worked closely with different companies, such as Mellanox and NVIDIA, to develop an industry-leading accelerated hardware, and the result is the Power AC922 server. Valuable collaboration has also occurred with Hortonworks, MongoDB, and EnterpriseDB Postgres to optimize their software running on Power System hardware. In addition, IBM is collaborating with the following cutting-edge artificial intelligence leaders:

- Anaconda
- Galvanize
- MapD
- ScyllaDB
- Nimbix
- Kinetica
- SQream DB
- H2O Driverless AI

The Power AC922 provides a complete set of end-to-end solutions for data professionals.This is a breed of enterprise-class server designed for the artificial intelligence era and created for end-to-end machine learning, deep learning, and artificial intelligence solutions.

The entire hardware and software system is offered by a single company, IBM.

The open ecosystem that IBM has designed is represented by Power AC922, together with NVIDIA and the entire software stack that is supported, such as Watson Machine Learning Accelerator, H2O Driverless AI, Watson Machine Learning Community Edition, SnapML, SQream DB, and Kinetica.

As with so many other solutions for cognitive computing, Power AC922 is helping businesses around the world to gain a competitive advantage by bringing in operational efficiencies, reaching out to new customer segments, and helping them to differentiate themselves. This kind of end-to-end solution greatly reduces complexity, costs, resources, and time.

The starter kit to run IBM Cognitive Solutions brings the power of the following components:

- 2 Power AC922 servers
- 4 NVIDIA V100 GPUs on each server
- Watson Machine Learning Accelerator software

For more information about the recommended hardware and software architecture read “Recommended architecture and hardware configuration” on page 29.

2

IBM Power System AC922 for cognitive computing

This chapter provides an overview of the most important hardware features on Power AC922. This chapter guides you through what you need to know in terms of hardware components and the most outstanding features in the system.

This chapter includes the following topics:

Today's challenges demand innovation. 68% of business leaders believe clients will demand more expandibility from artificial intelligence in the next 3 years, according to an IBM Institute for Business Value survey.

Every minute, an average of 500 hours of videos are uploaded to YouTube, 450,000 tweets are released on Twitter, and 2.5 million posts are made on Facebook. Contemporary data flow requires a full system design, end-to-end, open platform including not only hardware but software stack, all together.

Recent research showed that 38% of IT leaders consider the infrastructure the most strategic component. Power AC922 features breakthrough hardware enhancements, specifically designed for exactly that purpose, analyzing Big Data workloads in the artificial intelligence cognitive era.

Fact: IBM inventors were granted more than 1,600 artificial intelligence patents merely in 2018, making the company one of the most prominent in the artificial intelligence innovations space.

2.1 Key hardware components

Power AC922 is the next generation of the IBM POWER9 processor-based systems, specifically designed for deep learning, artificial intelligence, high-performance data analytics, and high-performance computing.

This contribution characterizes the data movement innovations of the Power AC922 nodes, delivered by IBM to Oak Ridge National Labs and Lawrence Livermore National Labs as part of the 2014 Collaboration of Oak Ridge, Argonne, and Livermore (CORAL) joint procurement activity. With a single high-performance computing system able to perform up to 200 PF of processing with access to 2.5 PB of memory, this architecture motivates a careful look at data movement. Power AC922 system with NVIDIA V100 GPUs have cache line granularity, more than double the bandwidth of PCIe Gen3.

The system is co-designed with OpenPOWER Foundation part of the Linux Foundation as of 20th of August 2019 and is in 2Us of rack space size.

Note: The goal of the OpenPower Foundation is to create an open ecosystem, using the IBM Power Architecture® to share expertise, investment, and server-class intellectual property to serve the evolving needs of the clients.

Recent publications have considered the challenge of movement in and out of the high bandwidth memory in an attempt to maximize GPU utilization and minimize overall application wall time. Power AC922 implements 128 B cache line coherency between the IBM POWER9 processor elements and the NVIDIA V100 processor elements. Power AC922 with six NVIDIA V100 GPUs is shown in Figure 2-1.

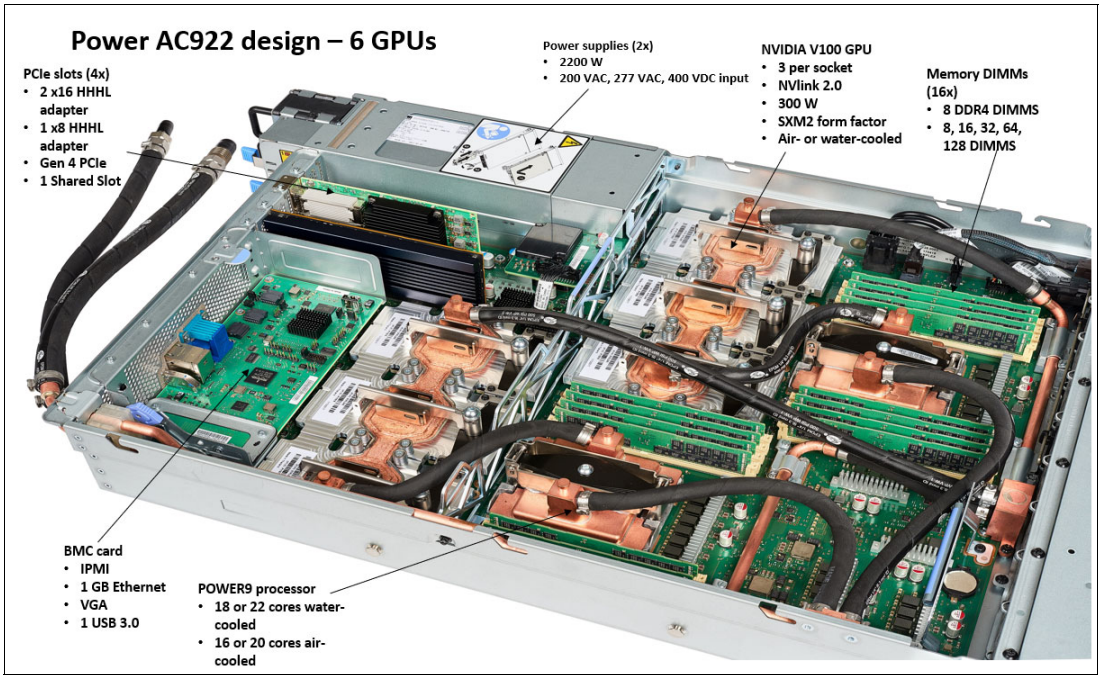

Figure 2-1 Power AC922 with 6 GPUs

The most significant features in Power AC922 system are:

- Up to 20 IBM POWER9 processor
- DDR4 Memory up to 2 TB
- Coherence shared RAM across CPUs and GPUs
- Up to 6 NVIDIA V100 GPUs
- CAPI/FPGA features
- PCIe Gen4 support

2.2 Software supported on the Power AC922

This section provides the list of the most significant cognitive solutions running on Power AC922 and some additional software:

- IBM Watson Machine Learning Accelerator

  Some of the following frameworks and tools are included in the bundle:

  - Tensorflow LMS
  - NVIDIA TensorRT
  - Distributed Deep Learning (DDL)
  - IBM enhanced Caffe with LMS
  - Snap Machine Learning (SnapML) on Apache Spark
  - PyTorch LMS
- Watson Machine Learning Community Edition

  Some of the following frameworks and tools are included in the bundle:

  - Tensorflow LMS
  - NVIDIA TensorRT
  - Distributed Deep Learning (DDL)
  - IBM enhanced Caffe with LMS
  - Snap Machine Learning (SnapML) on Apache Spark
  - PyTorch LMS
- IBM Spectrum Conductor
- IBM Spectrum Conductor Deep Learning Impact
- IBM LSF® Workload Manager
- NVIDIA Telsa CUDA recommended driver level 396.44 or later, or minimum driver level 396.26 from the CUDA 9.2 toolkit
- Red Hat Enterprise Linux and Ubuntu Server are both supported as operating systems
- The following additional software is supported:
  - Kubernetes
  - Docker
  - Singularity
  - Anaconda

2.3 Outstanding features

This section describes the most outstanding hardware features in Power AC922 and why the system is the best choice for cognitive solutions.

The biggest advantage of the Power AC922 is elimination of the PCIe bottleneck between CPU-GPU and GPU-GPU due to its unique architecture.

Power AC922 was the first server that could leverage system memory from the GPU side for 2 TB+ per node. PyTorch tests with Large Model Support running 1000 iterations finished in 49 minutes on the Power AC922 with four NVIDIA V100 GPUs. The same task in a Xeon X86 based 2640 v4 with 4 NVIDIA V100 GPUs needs 3.1 hours. NVIDIA NVLink 2.0 delivers 150 GBps CPU to GPU high-speed communication enabling storing a full hash table in CPU memory and transferring pieces to GPU for fast operations when needed, as shown in Figure 2-2.

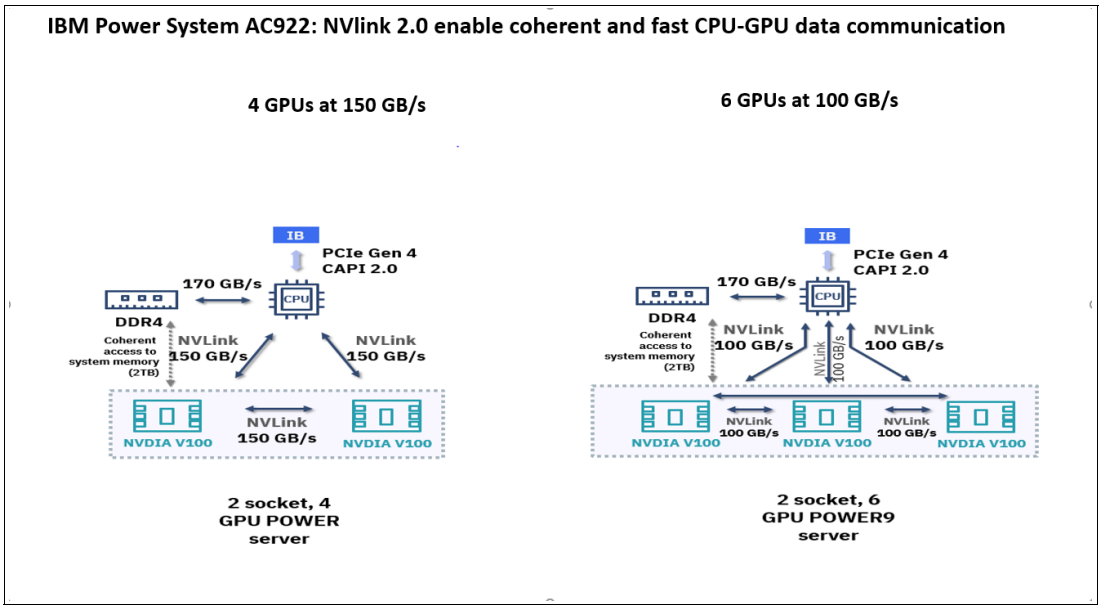

Figure 2-2 Power AC922 NVlink 2.0 features

There are some cases when the model size becomes extremely large as a result of the input data, such that it is not possible to keep the entire model in the GPU. In that case, the client has to keep the model and data in the system memory connected to the CPU and move over pieces at a time to the GPU. This comes at the cost of bottleneck communication between CPU and GPU thorough PCI-Express connection.

Therefore, data science professionals have to either choose between very large training times, or compromise by either reducing the size of the images and loss of accuracy or tiling the image, which makes it hard to train a model based on the input data set. This requirement led to the role of Large Model Support in Power AC922 with NVIDIA V100 GPUs.

Power AC922 enables clients to overcome this limitation for large models. The IBM POWER9 processor has the high-speed, next generation NVIDIA NVLink 2.0 direct interface embedded in the processor chip, which enables direct communication between the POWER9 processors and the NVIDIA V100 GPUs at 150 GBps.

IBM utilized this CPU-GPU NVIDIA NVLink 2.0 connection to build a module named Large Model Support that comes into our Watson Machine Learning Accelerator deep learning enterprise software distribution. For example, TensorFlow Large Model Support (TFLMS) is a Python module that provides an approach to training large models and data that cannot normally be fit in to GPU memory.

TFLMS takes a computational graph defined by users, and automatically adds swap-in and swap-out nodes for transferring tensors from GPUs to the host and vice versa. During training and inferencing this makes the graph execution operate like operating system memory paging. The system memory is effectively treated as a paging cache for the GPU memory, and tensors are swapped back and forth between the GPU memory and CPU memory.

We use the IBM Power AC922 server for our model training and evaluation. The AC922 features POWER9 CPUs and high-speed NVLink 2.0 connections to NVIDIA Tesla V100 GPUs. When this is combined with a high-speed memory bus between the CPU and system memory, it becomes feasible to utilize TFLMS to swap tensors between GPU and system memory. The NVLink 2.0 connections allow 75 GBps communication in each direction between CPU and GPU.

When this is compared to the 32GBps PCI Gen3 communication in traditionally connected GPUs, it is easy to understand how tensor swapping over a low bandwidth connected GPU would lead to an extreme degradation of model training performance. See Figure 2-3 for a comparison of NVlink 2.0 on x86 versus POWER9.

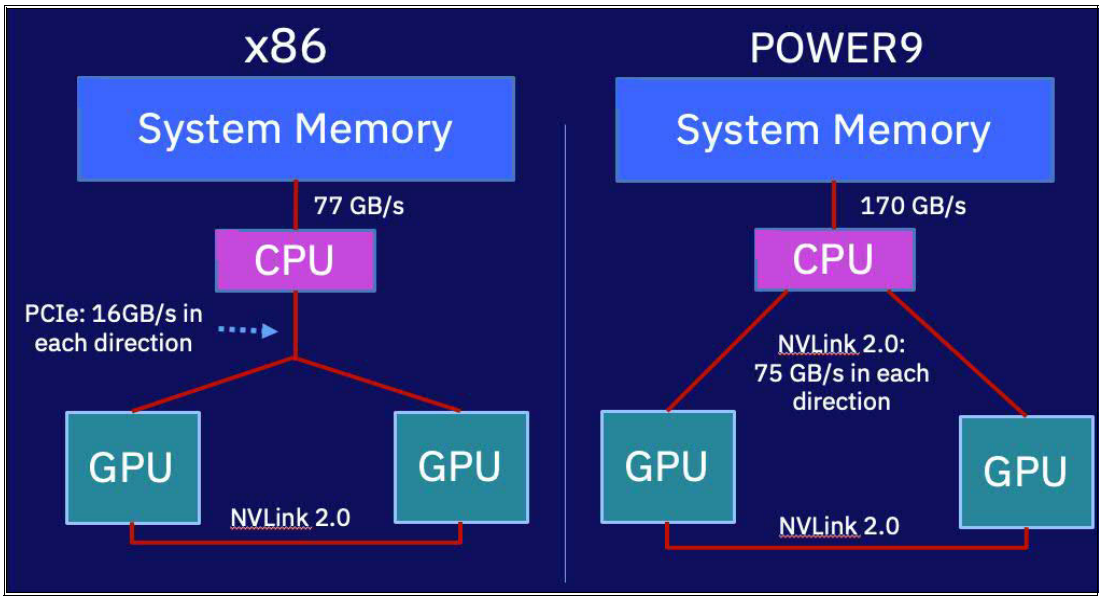

Figure 2-3 NVlink 2.0 on x86 versus NVlink 2.0 POWER9

Power AC922 has the most efficient scaling due to superior I//O between CPU-GPU, and between nodes. Combined with the POWER9 processor and PCIe Gen4, this is the best choice for anyone interested in extreme enterprise artificial intelligence and deep learning solutions.

3

Cognitive solutions

This chapter provides information about the cognitive solutions that can run on the IBM Power System AC922.

There are various solutions made for cognitive computing. In particular, solutions designed to work with specific accelerators to maximize processing performance and effectiveness are essential to achieving cognitive computing. Power AC922 can execute these solutions effectively because of the well-designed system elements. In addition to the cognitive solutions provided by IBM, open source and third-party solutions are included in the scope.

This chapter includes the following topics:

3.1 Cognitive solutions from IBM

This section highlights IBM Cognitive Solutions that can take advantage of the Power AC922 server, and introduces the benefits and the recommended architecture.

IBM offers several solutions that cover the various levels and ranges of capabilities you need for cognitive computing. Each solution clearly achieves its own purpose from a different perspective, and offers unique value for users accessing cognitive computing. Some solutions, in particular, can be combined with the Power AC922 server for superior performance.

It includes the following software:

- IBM Watson Machine Learning Accelerator
- IBM Watson Machine Learning Community Edition
- IBM PowerAI Vision

3.1.1 IBM Watson Machine Learning Accelerator

This section includes an overview of Watson Machine Learning Accelerator and explains the benefits of executing this solutions on the Power AC922 system.

Watson Machine Learning Accelerator is part of the IBM Watson Machine Learning family, and brings data science professionals and people in IT staff access to use the most advanced techniques for deep learning and machine learning.

Watson Machine Learning Accelerator is available to run on both Power Server and x86 server architectures. Executing Watson Machine Learning Accelerator on Power AC922 servers leverages the outstanding features within the server, mainly the second-generation NVIDIA NVLink 2.0, providing up to 150 GBps of bidirectional bandwidth between GPUs, and the CPU.

Watson Machine Learning Accelerator is improved and enhanced to take advantage of the key software and the hardware features of Power AC922. The enhanced functions of Watson Machine Learning Accelerator include Distributed Deep Learning capabilities. These capabilities include training models using more than one node, such as the use of SnapML software libraries to accelerate machine learning process.

SnapML libraries leverage NVIDIA NVLink 2.0 connections between GPU and CPU, and high bandwidth CPU memory to system memory. This enables up to 5.6 × the performance. In addition, memory consistency allows accelerated applications to leverage and use system memory as if it were GPU memory. Therefore, an application can have standard memory available to the GPU, because it is configured inside the GPU. This memory works in addition to the hardware-based memory that came with the GPU itself.

See Figure 3-1 for specific bandwidth speeds.

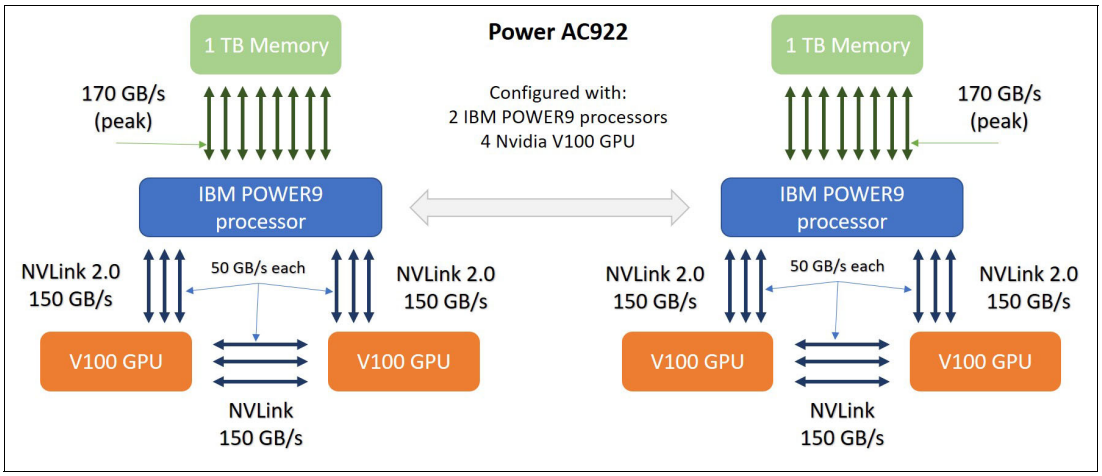

Figure 3-1 Power AC922 memory and NVIDIA NVLink 2.0 bandwidths

Because Watson Machine Learning Accelerator includes Distributed Deep Learning software, it allows growth and expandability of the solution by adding more Power AC922 nodes to the artificial intelligence cluster. This option gives flexibility and keeps your budget adjusted to the project lifecycle. You don't need hundreds of servers to start modeling, but you can add as many as needed without interrupting or reconfiguring the whole environment.

Performance advantages on Power AC922 against x86

Power AC922 is the most advanced hardware currently available to execute cognitive solutions from IBM and third-party providers. Through the following sections we show some of the laboratory testing made with the Power AC922 system, running different tasks and compared to x86 servers.

Power AC922 with PCIe Gen4 and NVIDIA NVLink 2.0

Power AC922 include PCIe Gen4 and NVIDIA NVLink 2.0. This new features resolve the PCI-E bottleneck for the code and with the POWER9 processor and NVIDIA NVLink 2.0 the transfer data rate can be 5.6x faster than the CUDA Host-Device Bandwidth of tested x86 platforms.

There is no code changes required to leverage NVIDIA NVLink 2.0 capability. However the application performance could be further increased with application code optimization. See Figure 3-2 for test results.

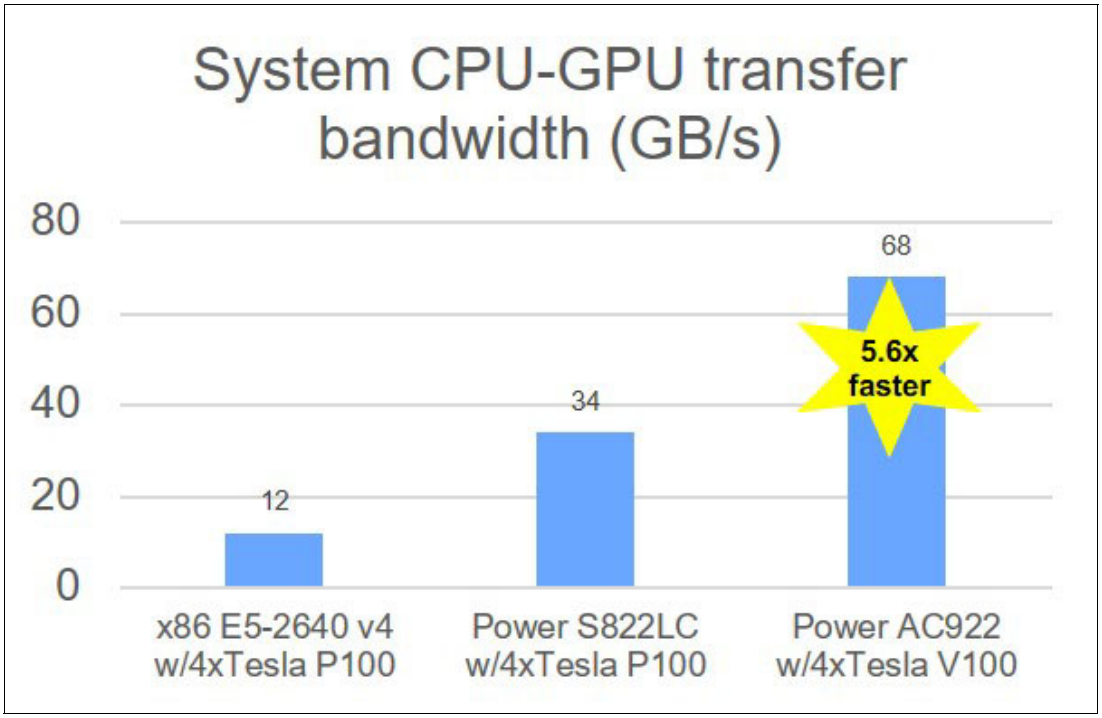

Figure 3-2 System CPU to GPU transfer bandwidth

Note: Results are based on IBM internal measurements running the CUDA H2D/D2H bandwidth test.

IBM Hardware: Power AC922; 3240 cores (2 x 20c chips), POWER9 processor with NVIDIA NVLink 2.0; 2.25 GHz, 1024 GB memory, 4xNVIDIA V100 GPU; RHEL 7.4 for Power LE (POWER9).; Power S822LC for HPC; 20 cores (2 x 10c chips), IBM POWER8® with NVIDIA NVLink 1.0; 2.86 GHz, 1024 GB memory, NVIDIA P100 GPU, RHEL 7.3.

Competitive HW: 2x Xeon E5-2640 v4; 20 cores (2 x 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 512 GB memory, 4xNVIDIA P100 GPU, Ubuntu 16.04.

Cognitive on Power AC922 with NVIDIA V100 GPU

Power AC922 with NVIDIA V100 GPU delivers 3.7 × reduction in artificial intelligence model training versus tested x86 systems. NVIDIA V100 GPU comes with 32 GB of hardware-based memory that works in addition to the Power AC922 server memory.

Maximize research productivity by running training for medical or satellite images with Caffe with Large Model Support on Power AC922 with NVIDIA V100 GPUs. Large Model Support manages data into system memory and caches it to the GPU, delivering larger models even above the hardware memory within the GPU. With this new feature, you can run larger models just by having more system memory. See Figure 3-3 on page 19 for a graphic about Large Model Support.

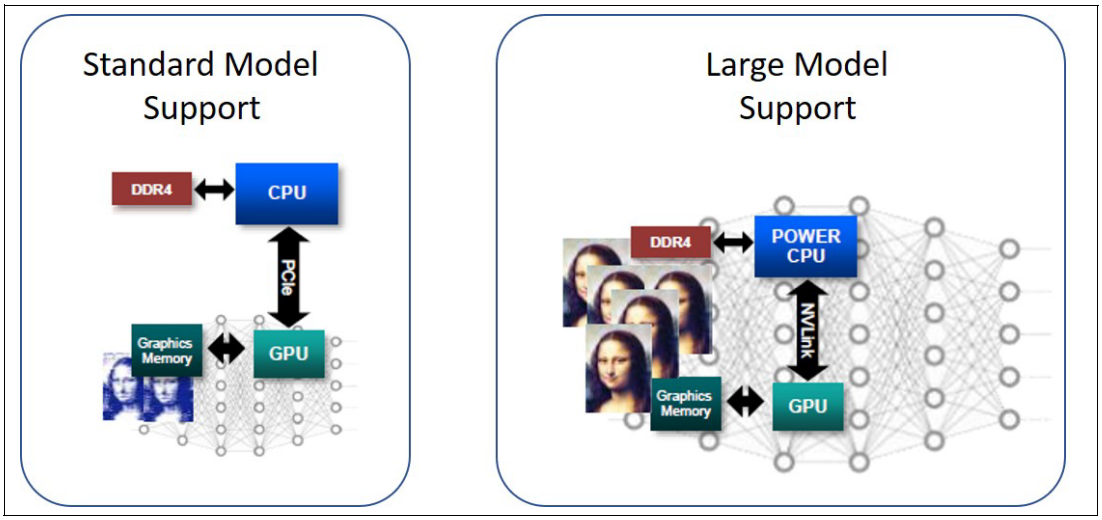

Figure 3-3 Large model support and standard model support

Note: For more information about Large Model Support, read the Developer Portal FAQ: https://developer.ibm.com/linuxonpower/deep-learning-powerai/faq/#tab_Q9

Watson Machine Learning Accelerator performance

Power AC922 with Watson Machine Learning Accelerator together deliver increased functionality:

- More models trained in the same time, improving data science professionals productivity.
- 31% reduction in training time of concurrent experiments versus tested x86 systems based on 4 jobs running concurrently on 3 GPUs, in a 2-node (4 GPUs each) cluster. See Figure 3-4 for test results.

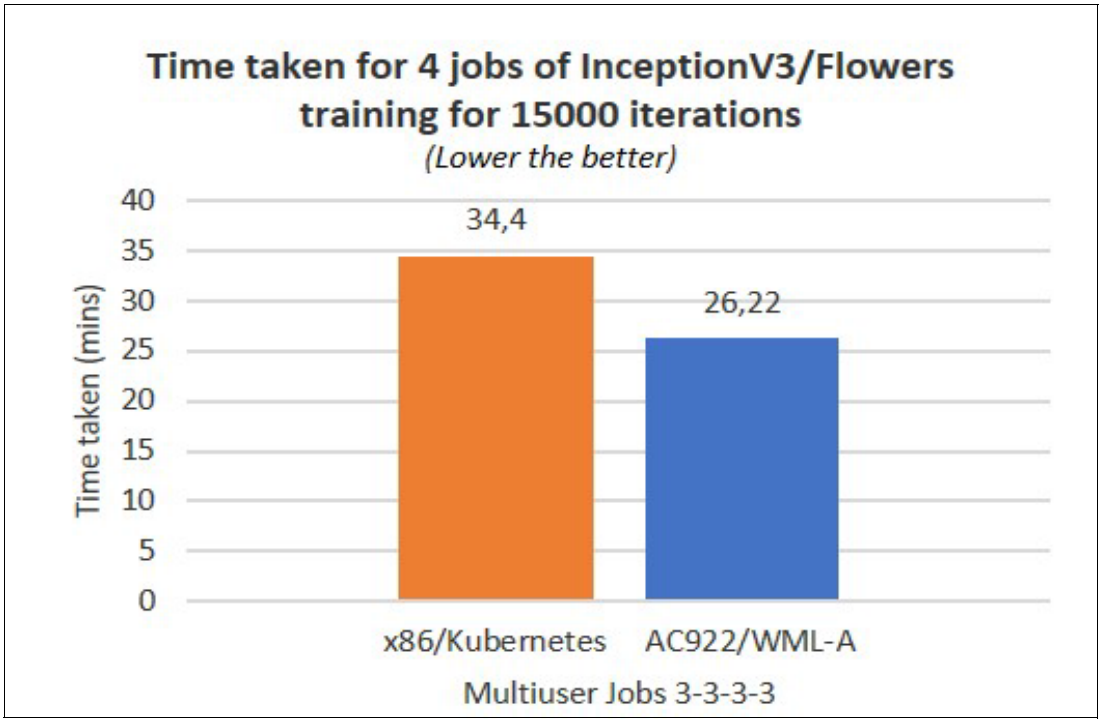

Figure 3-4 Time taken to run four jobs

Watson Machine Learning Accelerator optimizes multi-tenancy across the available resources, delivering the following functionality:

- 33% improvement in resource utilization versus tested x86 systems based on 4 jobs running concurrently on 3 GPUs, in a 2-node (4 GPUs each) cluster. See Figure 3-5 for test results.
- Automated scheduling with Watson Machine Learning Accelerator provides more efficient resource utilization.
- Automated elasticity with fair-share policy across the resources enables improved completion times by shortening or eliminating wait times.

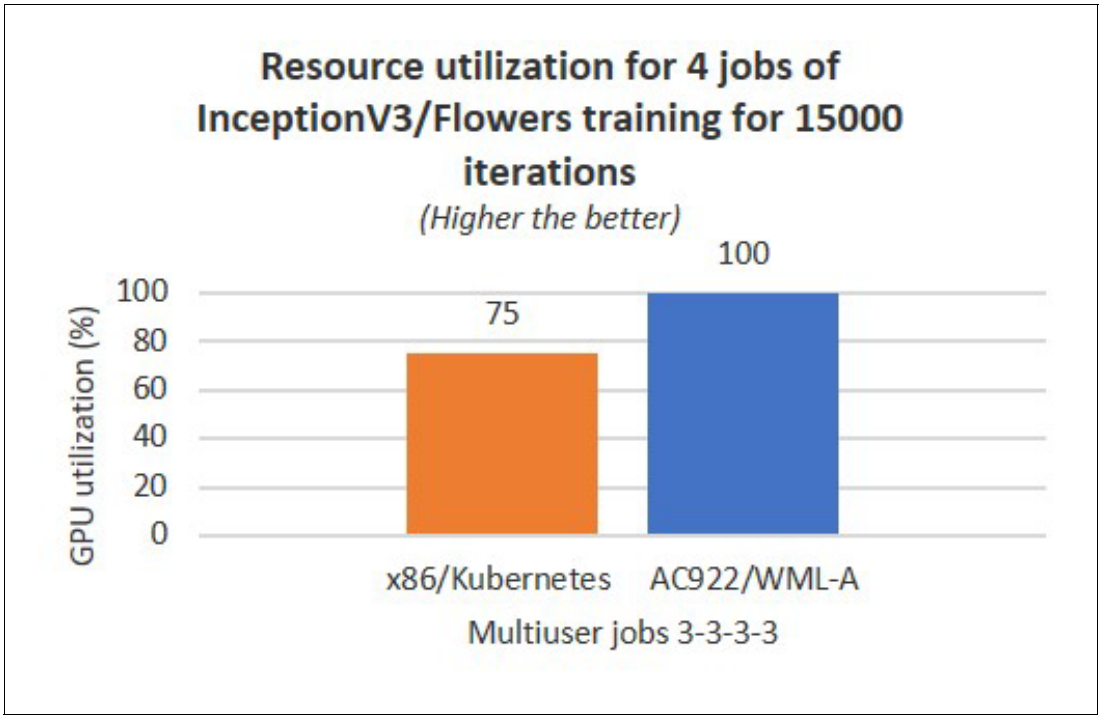

Figure 3-5 Resource utilization to run four jobs

Note: Results are based on IBM internal measurements running 15,000 iteration training of an InceptionV3 model (mini-batch size = 32 per GPU) on Flowers dataset. Conducted under laboratory conditions; individual results can vary based on workload size, use of storage subsystems, and other conditions.

Power AC922: 40 cores (2 x 20c chips), POWER9 processor with NVIDIA NVLink 2.0; 3.8 GHz, 1 TB memory, 4 × NVIDIA V100 GPU; Red Hat Enterprise Linux 7.5 for Power Little Endian (POWER9) with CUDA 9.2/ CUDNN 7.2.1; Watson Machine Learning Accelerator v1.1.1.

Competitive stack: 2x Xeon(R) Gold 6150; 36 cores (2 x 18c chips); 2.70 GHz; 512 GB memory, 4 × NVIDIA V100 GPU, Ubuntu 16.04.4 with CUDA.9.1/ CUDNN 7.1.2; NGC image:nvcr.io/NVIDIA/tensorflow Version: 18.08-py2; Kubernetes v1.11.2.

Power AC922 running training with Caffe and Large Model Support

Power AC922 with NVIDIA V100 GPUs running training with Caffe and Large Model Support can deliver the following results:

- 3.8 × time reduction versus tested x86 systems in runtime of 1000 iterations running on competing systems to train on 2 k × 2 k images. See Figure 3-6 for graphics about Caffe performance runtimes.

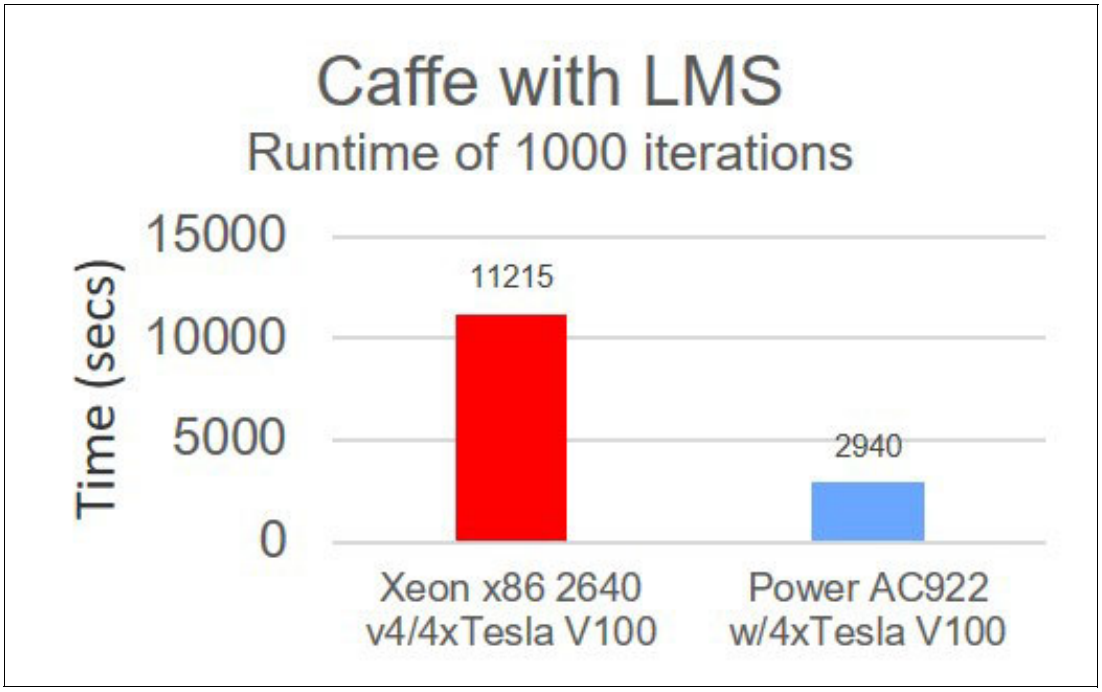

Figure 3-6 Caffe performance chart

Note: Results are based on IBM internal measurements running 1,000 iterations of Enlarged GoogleNet model (mini-batch size=5) on Enlarged Imagenet Dataset (2240 × 2240).

Power AC922::40 cores (2 × 20c chips), POWER9 processor with NVIDIA NVLink 2.0; 2.25 GHz, 1024 GB memory, 4 × NVIDIA V100 GPU; Red Hat Enterprise Linux 7.4 for Power Little Endian (POWER9) with CUDA 9.1/ CUDNN 7.

Competitive stack: 2 × Xeon E5-2640 v4; 20 cores (2 × 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 1024 GB memory, 4 × NVIDIA V100 GPU, Ubuntu 16.04. with CUDA.9.0/ CUDNN 7.

Software: IBM Caffe with Large Model Support Source code
https://github.com/ibmsoe/caffe/tree/master-lms

Power AC922 running TensorFlow 1.10

Time comparison of TensorFlow with Large Model Support on Power AC922 and x86 with one Nvidia V100 GPU can deliver the following performance:

- 2.4 × better in epoch time compared to x86/V100 1GPU with 3DUnet CNN model for image segmentation with a patch size of 192^3 and batch size of 1. See Figure 3-7 on page 22 for test results.

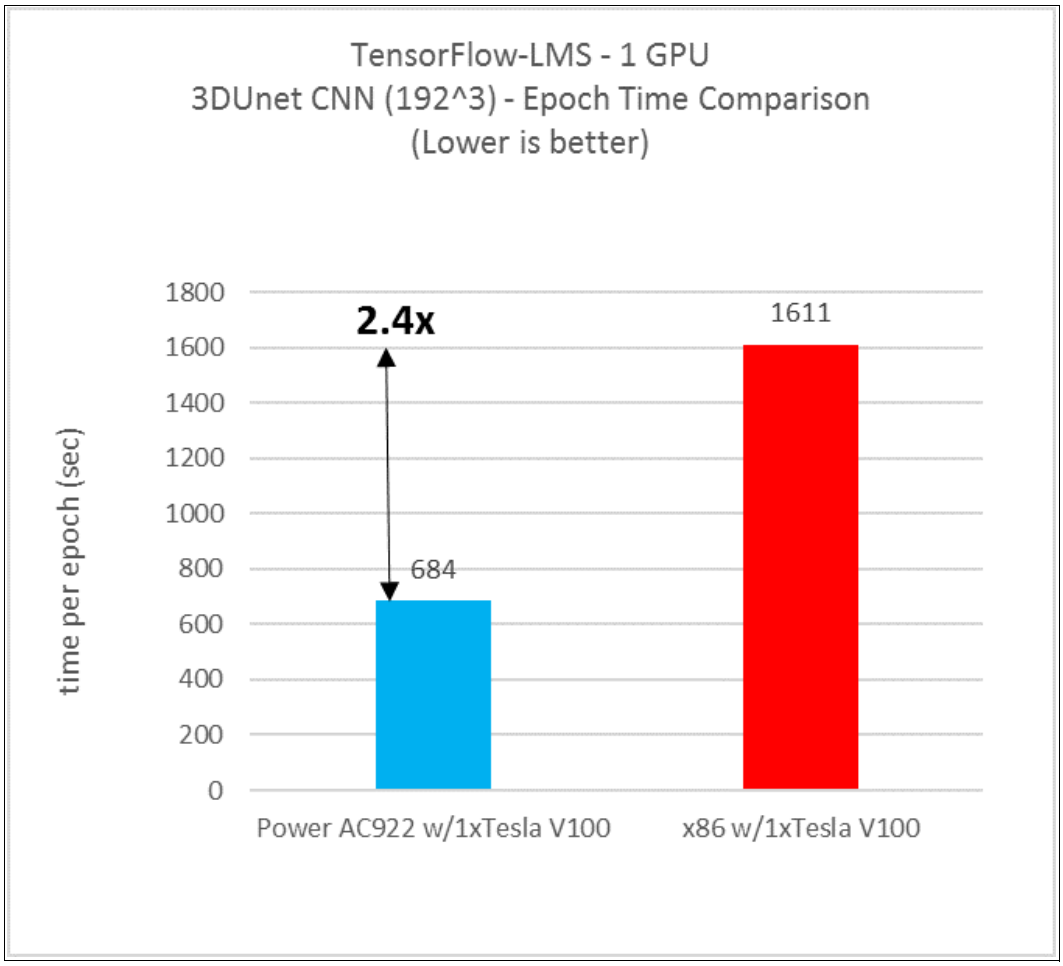

Figure 3-7 TensorFlow 1.10 with Large Model Support (1GPU)

Notes:

Hardware Stack: IBM Power AC922; 40 cores (2 x 20c chips), POWER9 processor with NVLink 2.0; 2.25 GHz, 512 GB memory, 4 × Nvidia V100 GPU (16 GB), RHEL7.5 for P9, CUDA 9.2/396.44, CuDNN7.2.1.

Competitive Stack: x86 server (Intel Xeon); 40 cores (2 × 20c chips); 2.20 GHz; 512 GB memory, 8 × Nvidia V100 GPU (16 GB), Ubuntu 16.04, Cuda9.0/384.145, CuDNN7.2.1.

Framework on P9(AC922)/V100: TensorFlow 1.10 with LMS from PowerAI 1.5.3.

Framework on x86: TensorFlow 1.10 standard distribution with LMS code.

Model on IBM AC922: 3DUnet CNN model was modified to be run on AC922 using TFLMS Keras Callback and use DDL for scaling to 4 GPUs.

Model on x86: 3DUnet CNN model was modified to scale to multiple GPUs using Horovod distributed training framework for TensorFlow and Keras. The model also uses TFLMS Keras Callback to enable LMS Tensor Swapping.

The Keras 3DUnet CNN model was written to process the TCGA and MICCAI BraTS 2017 datasets [12]. BraTS 2017 dataset is preprocessed and converted to `.h5` files. The dataset has 285 images (subjects): 228 (80%) for training and 57 (20%) for validation.

Time comparison of TensorFlow with Large Model Support on Power AC922 and x86 in a multi-GPU scenario with four Nvidia V100 GPUs optimized with Distributed Deep Learning can deliver:

- 3.6 × better in epoch time compared to x86 with four Nvidia V100 GPUs with 3DUnet CNN model for image segmentation with a patch size of 192^3 and batch size of 1. See Figure 3-8 for test results.

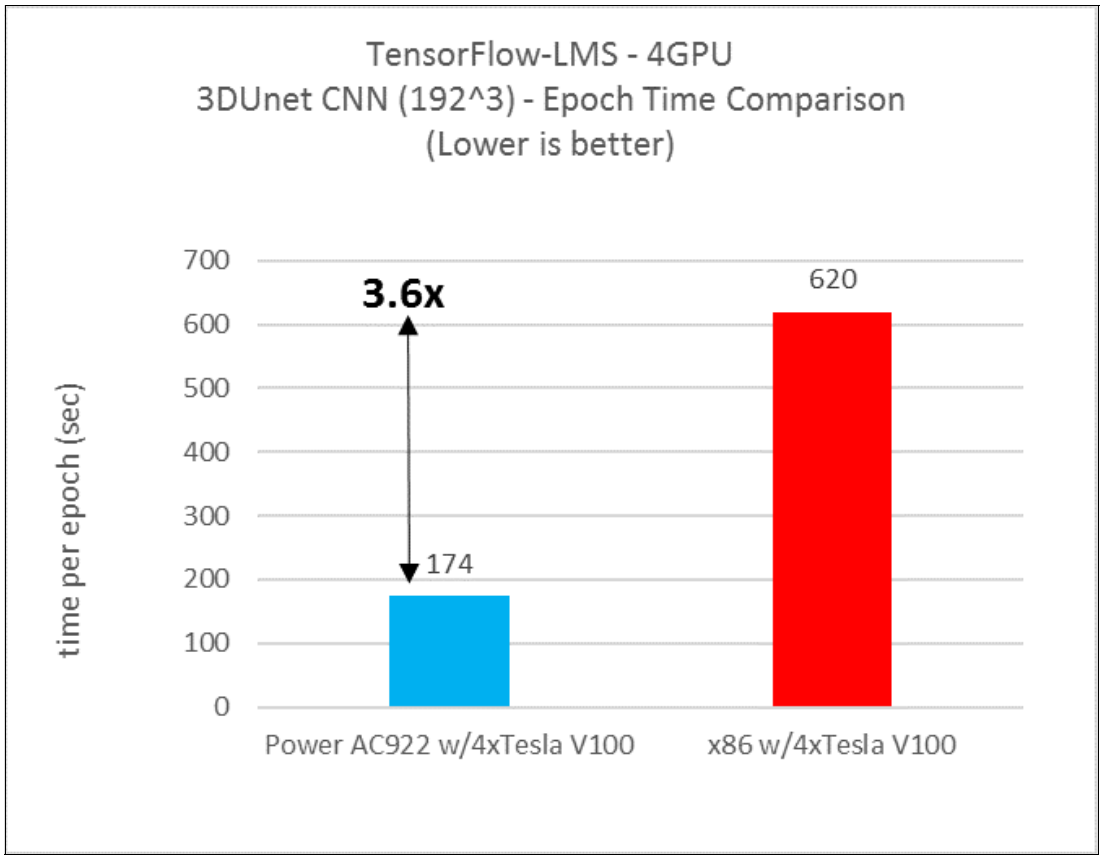

Figure 3-8 TensorFlow 1.10 with Large Model Support (4 GPU)

Notes:

Hardware Stack: IBM Power AC922; 40 cores (2 × 20c chips), POWER9 processor with NVLink 2.0; 2.25 GHz, 512 GB memory, 4 × Nvidia V100 GPU (16 GB), RHEL7.5 for P9, CUDA 9.2/396.44, CuDNN7.2.1.

Competitive Stack: x86 server (Intel Xeon); 40 cores (2 x 20c chips); 2.20 GHz; 512 GB memory, 8x Nvidia V100 GPU (16 GB), Ubuntu 16.04, Cuda9.0/384.145, CuDNN7.2.1.

Framework on P9(AC922)/V100: TensorFlow 1.10 with LMS from PowerAI 1.5.3.

Framework on x86: TensorFlow 1.10 standard distribution with LMS contrib code.

Model on IBM AC922: 3DUnet CNN model was modified to be run on AC922 using TFLMS Keras Callback and use DDL for scaling to 4 GPUs.

Model on x86: 3DUnet CNN model was modified to scale to multiple GPUs using Horovod distributed training framework for TensorFlow and Keras. The model also uses TFLMS Keras Callback to enable LMS Tensor Swapping.

The Keras 3DUnet CNN model was written to process the TCGA and MICCAI BraTS 2017 datasets [12]. BraTS 2017 dataset is preprocessed and converted to `.h5` files. The dataset has 285 images (subjects): 228 (80%) for training and 57 (20%) for validation.

Time comparison of TensorFlow with Large Model Support on Power AC922 with four Nvidia V100 GPUs versus x86 with eight Nvidia V100 GPUs can deliver:

- 2.1 × better in epoch time compared to x86/V100 8GPUs with 3DUnet CNN model for image segmentation with a patch size of 192^3 and batchsize of 1. See Figure 3-9 for test results.

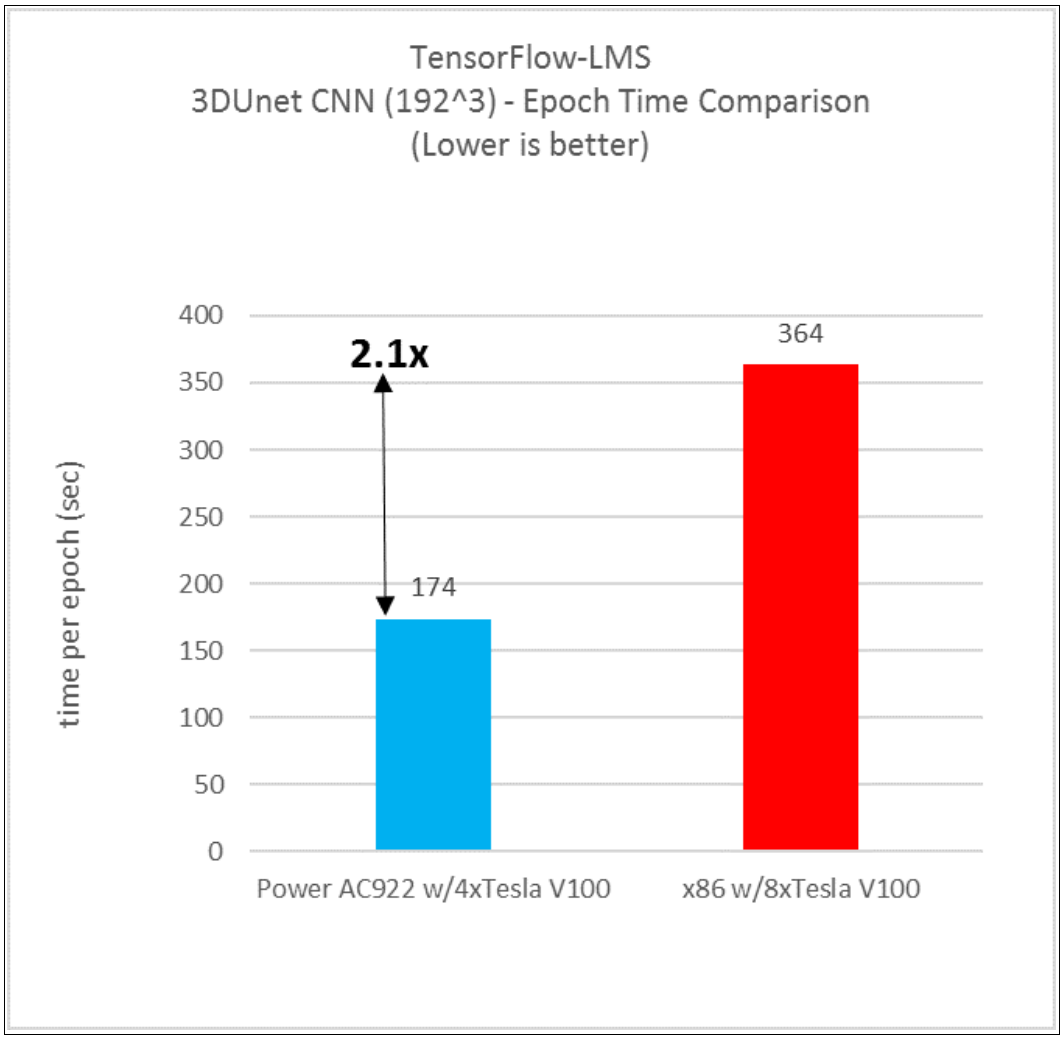

Figure 3-9 TensorFlow 1.10 with Large Model Support (4 GPU P9 vs 8GPU x86)

Notes:

Hardware Stack: IBM Power AC922; 40 cores (2 × 20c chips), POWER9 processor with NVLink 2.0; 2.25 GHz, 512 GB memory, 4x Nvidia V100 GPU (16GB), RHEL7.5 for P9, CUDA 9.2/396.44, CuDNN7.2.1.

Competitive Stack: x86 server (Intel Xeon); 40 cores (2 × 20c chips); 2.20 GHz; 512 GB memory, 8x Nvidia V100 GPU (16GB), Ubuntu 16.04, Cuda9.0/384.145, CuDNN7.2.1.

Framework on P9(AC922)/V100: TensorFlow 1.10 with LMS from PowerAI 1.5.3.

Framework on x86: TensorFlow 1.10 standard distribution with LMS contrib code.

Model on IBM AC922: 3DUnet CNN model was modified to be run on AC922 using TFLMS Keras Callback and use DDL for scaling to 4 GPUs.

Model on x86: 3DUnet CNN model was modified to scale to multiple GPUs using Horovod distributed training framework for TensorFlow and Keras. The model also uses TFLMS Keras Callback to enable LMS Tensor Swapping.

The Keras 3DUnet CNN model was written to process the TCGA and MICCAI BraTS 2017 datasets [12]. BraTS 2017 dataset is preprocessed and converted to .h5 files. The dataset has 285 images (subjects): 228 (80%) for training and 57 (20%) for validation.

Note: For more information about these tests, read the following Developer Portal article: https://developer.ibm.com/linuxonpower/2018/12/19/performance-of-3dunet-multi-gpu-model-for-medical-image-segmentation-using-tensorflow-large-model-support/

Power AC922 running Car-Parrinello molecular dynamics (CPMD)

The Power AC922 reduces the waiting time and improves computational chemistry simulation execution time (CPMD). See Figure 3-10 on page 27 for graphics about CPMD performance. The test results shows that Power AC922 can deliver the following performance factors:

- 2.9 × reduction in execution time compared to tested Xeon x86 systems
- 2.0 × eduction in execution time compared to prior generation Power S822LC for HPC

Power AC922 with NVIDIA NVLink 2.0 unlocks the performance of a GPU-accelerated version of CPMD by enabling lightning fast CPU-GPU data transfers and can deliver:

- 3.3 TB of data movement required between CPU and GPU
- 70 seconds for NVIDIA NVLink 2.0 transfer time versus 300+ seconds for traditional PCIe bus transfer time

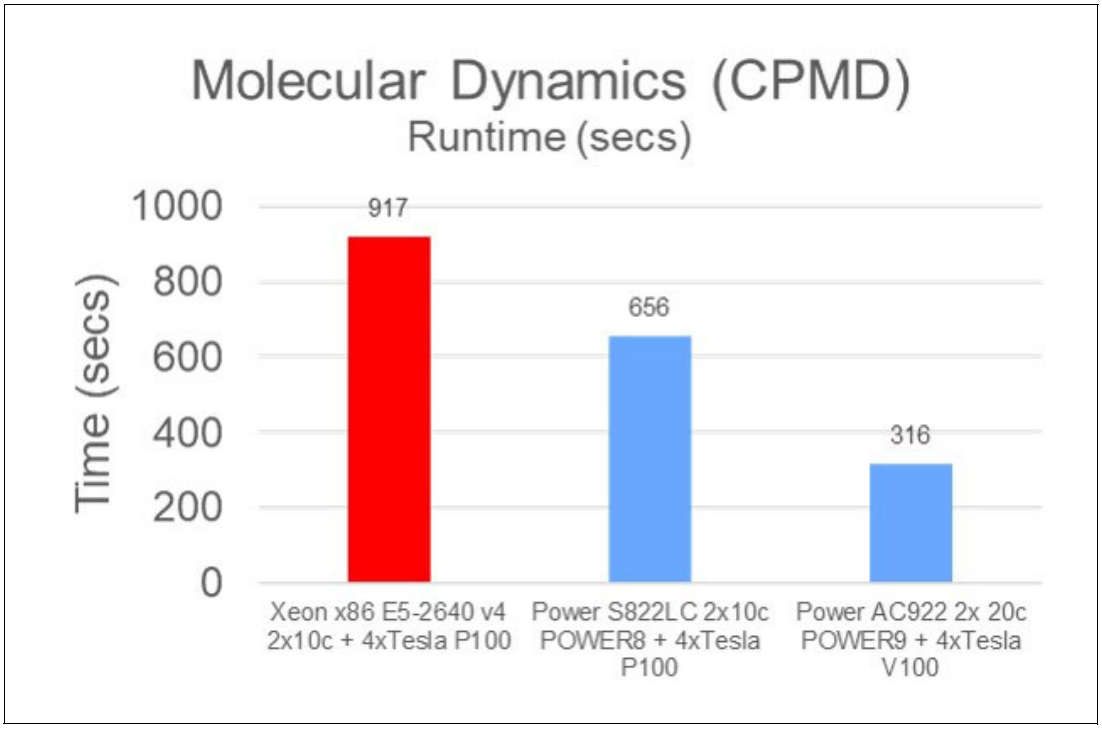

Figure 3-10 Molecular Dynamics (CPMD) performance chart

Note: All results are based on running CPMD, a paralelized plane wave/pseudo potential implementation of Density Functional Theory Application. A Hybrid version of CPMD (for example, MPI + OPENMP + GPU + streams) was implemented with runs are made for 256-Water Box, RANDOM initialization. Results are reported in Execution Time (seconds). Effective measured data rate on PCIe bus of 10 GBps and on NVIDIA NVLink 2.0 of 50 GBps.

Power AC922: 40 cores (2 × 20c chips), POWER9 processor with NVIDIA NVLink 2.0; 2.25 GHz, 1024 GB memory, 4 × NVIDIA V100 GPU; Red Hat Enterprise Linux 7.4 for Power Little Endian (POWER9) with ESSL PRPQ; Spectrum MPI: PRPQ release, XLF: 15.16, CUDA 9.1.

Power S822LC for HPC: 20 cores (2 × 10c chips) / 160 threads, POWER8 with NVIDIA NVLink 1.0; 2.86 GHz, 256 GB memory, 2 × 1 TB SATA 7.2K rpm HDD, 2-port 10 GbEth, 4 × NVIDIA P100 GPU; RHEL 7.4 with ESSL 5.3.2.0; PE2.2; XLF: 15.1, CUDA 8.0.

2x Xeon E5-2640 v4: 20 cores (2 × 10c chips) / 40 threads; Intel Xeon E5-2640 v4; 2.4 GHz; 256 GB memory, 1 × 2 TB SATA 7200 RPM HDD, 2-port 10 GbE; 4 × NVIDIA P100 GPU; Ubuntu 16.04 with OPENBLAS 0.2.18, OpenMPI: 1.10.2, GNU-5.4.0, CUDA-8.0.

Power AC922 running Watson Studio Local

Power Systems Cluster with Power LC922 (CPU optimized) and Power AC922 (GPU accelerated) provides an optimized infrastructure for IBM Watson Studio Local. Accelerating data science professionals productivity and drive faster insights with Watson Studio Local on Power AC922. See Figure 3-11 on page 28 for performance results.

Power AC922 completes running GPU accelerated K-means clustering with 15 GB data in half of the time than tested x86 systems (Skylake 6150 with NVIDIA GPUs), delivering:

- 2 × faster insights for GPU accelerated K-means clustering workload than Intel Xeon SP Gold 6150 based servers.

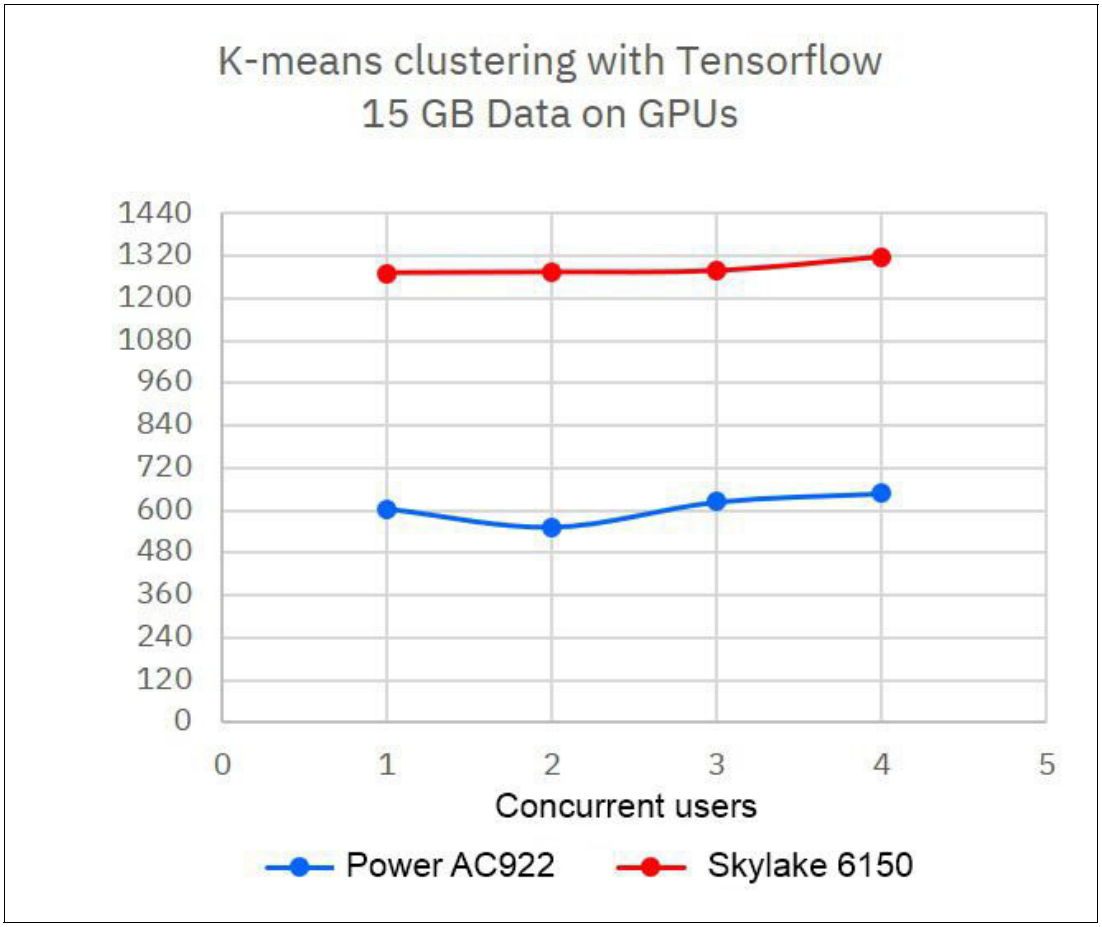

Figure 3-11 K-means clustering with Tensorflow performance chart

Note: Based on IBM internal testing of the core computational step to form 5 clusters using a 5270410 × 301 float64 data set (15 GB) running the K-means algorithm using Apache Python and Tensorflow. Results valid as of 6/13/2018 and conducted under laboratory conditions with speculative execution controls to mitigate user-to-kernel and user-to-user side-channel attacks on both systems. Individual results can vary based on workload size, use of storage subsystems, and other conditions.

Power AC922: (2 × 20-core/3.78 GHz and 4 NVIDIA V100 GPUs with NVIDIA NVLink 2.0) using 2 × 960 GB SSD, 10 GbE 2-port, RHEL 7.5 LE for Power9. Total 1 TB memory, each user assigned 180 GB in DSXL.

Competitive stack: 2-socket Intel Xeon SP (Skylake and 4 NVIDIA Tesla V100 GPUs) Gold 6150 (2 × 18-core/2.7 GHz) using 2 × 960 GB SSD, 10 GbE 2-port and RHEL 7.5. Total 768 GB memory, each user assigned 180 GB in DSXL.

Apache, Apache Python, and associated logos are either registered trademarks or trademarks of the Apache Software Foundation in the United States and/or other countries. No endorsement by The Apache Software Foundation is implied by the use of these marks.

Note: In this test DSXL refers to Data Science Experience Local what was renamed, rebranded and improved to become IBM Watson Studio Local.

Recommended architecture and hardware configuration

Watson Machine Learning Accelerator recommended architecture begins with the PowerAI Starter Kit. This kit comes with all of the required features, including software and hardware to start creating and training machine learning models. See Figure 3-12 for a complete list of features included in the kit.

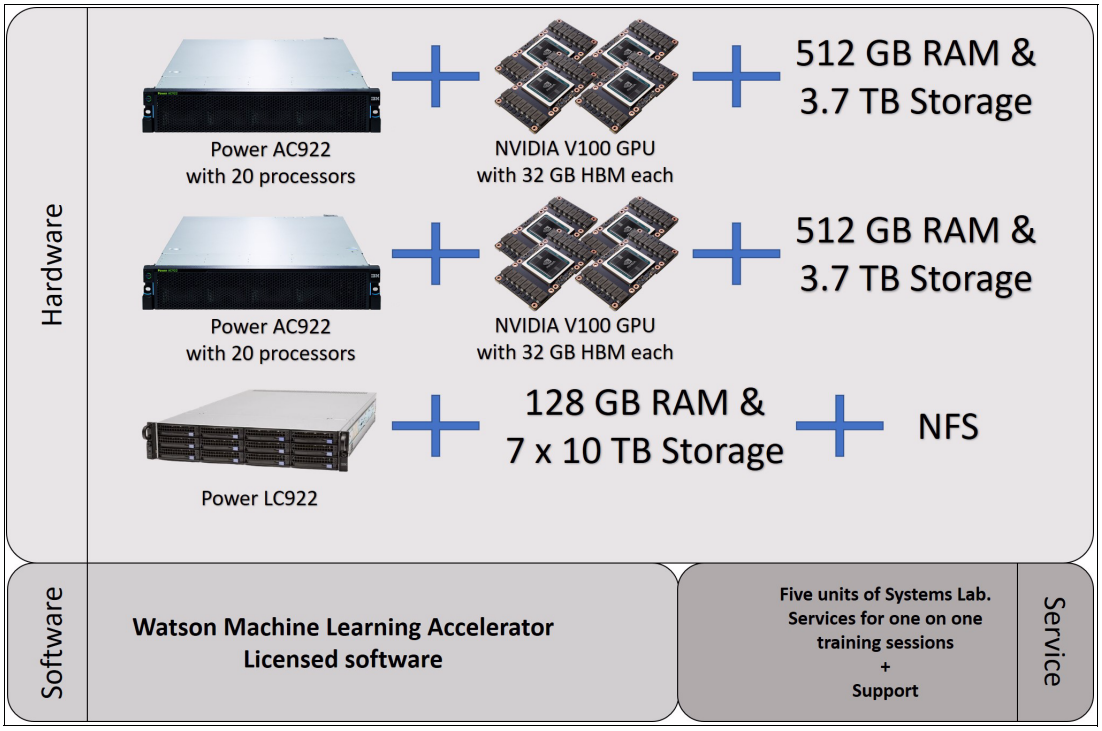

Figure 3-12 Watson Machine Learning Accelerator Starter Kit

This starter kit enables data science professionals to start creating models and training them. When the requirements mature, the infrastructure is also capable of growing by adding more Power AC922 nodes to the cluster.

3.2 Third-party cognitive solutions

In addition to the solutions provided by IBM, there are also other third-party software solutions that enable cognitive computing in a variety of ways. In particular, as computing using accelerators becomes an essential element for analyzing a large amount of data quickly, software that provides a new concept of processing different from conventional multiple processing is emerging.

Designed as the premier GPU acceleration platform, the Power AC922 server offers powerful infrastructure for computing with accelerators. As discussed in Chapter 2, “IBM Power System AC922 for cognitive computing” on page 9, these are the factors that eliminate I/O bottlenecks and share memory between GPUs and CPUs to handle data-intensive workloads.

Third-party solutions include the following software:

- H2O Driverless AI
- SQream DB
- Kinetica

Note: More information about ISVs software is available on the *IBM Power Systems testmonials* website:

https://developer.ibm.com/linuxonpower/isv-testimonials/

3.2.1 H2O Driverless AI and Power AC922

This section describes H2O Driverless AI running on Power AC922 and all of the benefits that clients can gain from this collaboration.

Every day, a new type of technology is being introduced to the world. Not a single company can keep up with the speed of all these emerging technologies, which is why it is very important to partner with the right company that can bring you to the next level.

H2O.ai is a well-known Machine Learning and Data Science platform, which has built an impressive reputation among enterprise clients for innovations in artificial intelligence and commitment to the open source artificial intelligence community. H2O Driverless AI is H2O.ai’s solution for automated machine learning. It simplifies many of the data science machine learning tasks.

The data can be uploaded from IBM cloud, Hadoop, or Power AC922. The data visualization is automatic. H2O is an open-source, machine learning platform, while H2O Driverless AI is the enterprise equivalent with automatic machine learning features.

The great thing about H2O driverless AI is how easily it can be installed. Similar to its name, it can be deployed as a docker image or an RPM package. Since early 2019, H2O open source and H2O Driverless AI are available on IBM Cloud™ Private Platform.

Companies that want to scan huge data pools of textual and numeric information can take advantage of advanced performance boost when pairing H2O Driverless AI with a Power AC922 system. That performance boost is due to the following factors:

- 2.6× more RAM for Big Data Scale
- 9.5× Max I/O bandwidth
- 30× faster NVIDIA/Power System GPU accelerated Machine Learning
- 2× data ingestion speed
- H2O open source libraries integrated into IBM's Watson Studio Analytic solution

Clients are able to access the open source H2O Python modules, H2O Flow, and R Modules from within the Watson Studio Solution. As a result, clients benefit from greater flexibility and wider choice.

Note: H2O.ai was named a leader among the 16 vendors included in Gartner's 2018 Magic Quadrant for Data Science Platforms.

3.2.2 SQream DB and Power AC922

This section introduces you to SQream DB provided by SQream Technology. In particular, it discusses the benefits and recommended architecture of running SQream DB on the IBM Power System AC922. SQream develops and markets SQream DB, a software-defined GPU data warehouse, which has the hardware-accelerated coprocessors as a key component in making more data more accessible.

According to SQream, SQream DB helps today's large companies access more data they collect and discover faster, and more business insights hidden within them. Basically, it adopts columner DB and shows better performance than CPU-based processing methods due to data injection and analysis processing methods using the GPU. It was built to harness the raw brute-force power and high throughput capabilities of the GPU, with MPP-on-chip capabilities and a fully relational SQL database.

SQream DB is not an in-memory database, or a translation layer for Hadoop. It is its own database, designed for larger-than-memory, constantly growing data. To effectively handle large amounts of data, storage is optimized for columnar partitioning, and all nodes require an architecture that shares data. Figure 3-13 shows the process of loading data and processing queries from SQream DB. For more information, see the official website:

http://www.sqream.com/

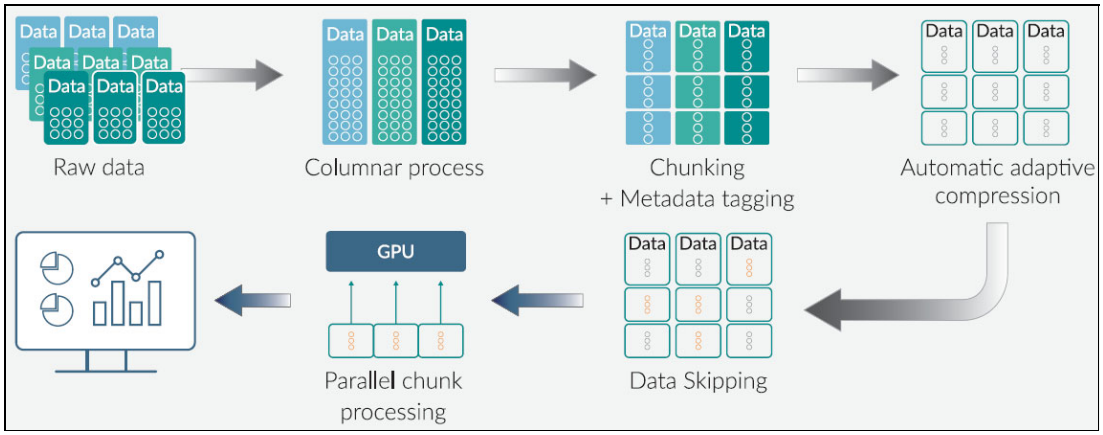

Figure 3-13 How SQream DB Works: Load-and-Go high-throughput architecture

Performance results on Power AC922 against x86

SQream announced at the OpenPOWER Foundation Summit in Europe, that SQream DB includes optimized support for the IBM POWER9 multi-core architecture. At the same time, SQream recalled that running SQream DB on the IBM Power System AC922 server can provide the following performance benefits over x86 systems:

- 3.7× faster query processing time on Power AC922 than x86 system
- 1.7× faster data load time on Power AC922 than x86 system

SQream DB actively uses the GPU for processing queries and loading data. That is, wherever the data that needs to be processed in SQream DB is initially loaded in system memory, it must be replicated in GPU memory. Minimizing the time it takes to move data is a key factor in improving performance. The Power AC922 features a unique architecture with NVIDIA NVLink 2.0 between the CPU and GPU, up to eight large memory channels, and the latest PCI Gen4 that can handle twice the I/O of the previous generation. The major components of the Power AC922 eliminate the bottleneck of performance bottlenecks, minimizing the data movement time mentioned previously and bringing a substantial effect.

As shown in Figure 3-14, running the SQream DB on the Power AC922 server provides approximately 3.7 times faster query processing performance than the x86 system without any changes or tuning in the TPC-H benchmark environment. This is because, as explained previously, the unique elements only provided in the Power AC922, such as the NVIDIA NVLink 2.0 between the CPU and GPU, played a significant role in improving performance.

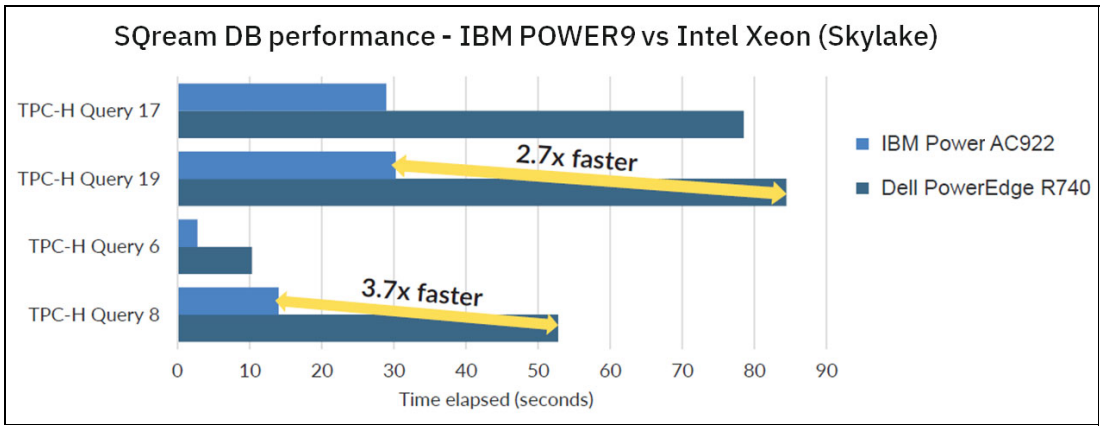

Figure 3-14 TPC-H benchmark results of Power AC922 vs x86 server

Another test result was that when running SQream DB on the Power AC922 server to load 6 billion TPC-H records, due to the superior memory channel compared to the x86 system and the enhanced I/O system, including NVIDIA NVLink 2.0, PCIe Gen4. You can see that the loading completes approximately 1.7 times faster, as shown in Figure 3-15.

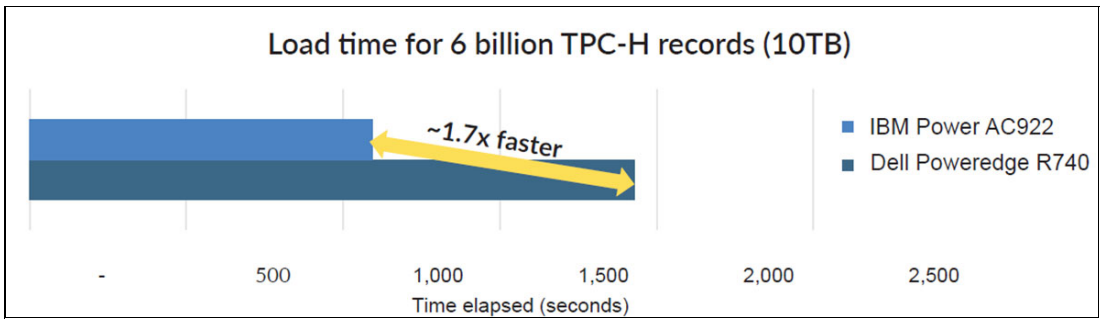

Figure 3-15 Load time comparison of TPC-H records

These results show that SQream DB is a data-intensive workload that takes full advantage of the power of IBM POWER9 processor and NVIDIA GPUs with NVLink 2.0 technology. It enables faster processing, flexibility, and cost-effective analysis of massive datasets when running on the Power AC922.

Note: On the official website of the OpenPower Foundation Summit in Europe, you can find the original content SQream used during the seminar:

https://openpowerfoundation.org/wp-content/uploads/2018/10/David-Leichner.IBM-OpenPOWER-SQream-POWER9.pdf

Recommended architecture and Hardware configuration

The following are four factors recommended as an infrastructure architecture for building SQream DB at scale:

- Compute nodes

  Power AC922 with NVIDIA V100 (32 GB HBM)

- Massive I/O networking

  To process massive I/O networking, it is required to secure large bandwidth of both north/south (server-to-storage) traffic and east/west (server-to-server) traffic. Mellanox Infiniband can be a good option to prepare a large network.

- Shared-data architecture

  IBM Spectrum Scale provides an essential environment for sharing and accessing the same data across all compute nodes. It is an enterprise-grade parallel file system that provides superior resiliency, scalability, and control. IBM Spectrum Scale delivers scalable capacity and performance to handle demanding data analytic, content repositories and technical computing workloads.

- Storage with High-throughput/capacity

  IBM Spectrum Storage™ for AI, IBM Flash Storage

Note: SQream published a brochure introducing SQream DB on Power AC922:

https://info.sqream.com/hubfs/pdf/SQream%20DB%20for%20Power%20Systems.pdf

3.2.3 Kinetica and Power AC922

This section introduces you to Kinetica provided by Kinetica DB, Inc. In particular, it discusses the benefits and recommended architecture of running Kinetica on the Power AC922.

Kinetica is an in-memory-based, GPU-accelerated database that allows you to further accelerate your analysis using the GPU for parallel computing. Kinetica processes SQL queries that analyze billions of rows in microseconds, and simply prepares and injects data without extracting indexes.

It also injects and processes data at the same time, making it easy to perform SQL queries on streaming and geospatial data. Another feature of Kinetica is the ability to instantly translate temporal, geospatial, and streaming data into visuals, and to present it on a visual dashboard. Kinetica's in-memory, distributed image processing, and rendering capabilities seamlessly support this visual foresight. For more information, see the official website:

https://www.kinetica.com/

Performance results on Power AC922 against x86

IBM shows two test results to verify the effectiveness of running Kinetica on IBM Power Systems:

- IBM POWER8 delivers 2.5× better performance than x86 with PCIe Gen3 ×16 when runnning Kinetica.
- IBM POWER9 delivers 1.8× better performance than IBM POWER8 even when running Kinetica.

The first test results were presented by IBM at the 2016 GPU Technology Conference, which compares Power S822LC for HPC servers with NVIDIA NVLink 1.0, and x86 systems with PCIe Gen3 x16. In those tests, Power S822LC for HPC Server with IBM POWER8 processors delivers approximately 2.5× the throughput of x86, as shown in Figure 3-16.

Looking closely at each query time, we can see that on x86 servers the data transfer time, the longest process of the entire query time, has been reduced by 65% in Power 2822LC for HPC. These results show that NVIDIA NVLink has a significant impact on overall performance improvement when processing queries in Kinetica.

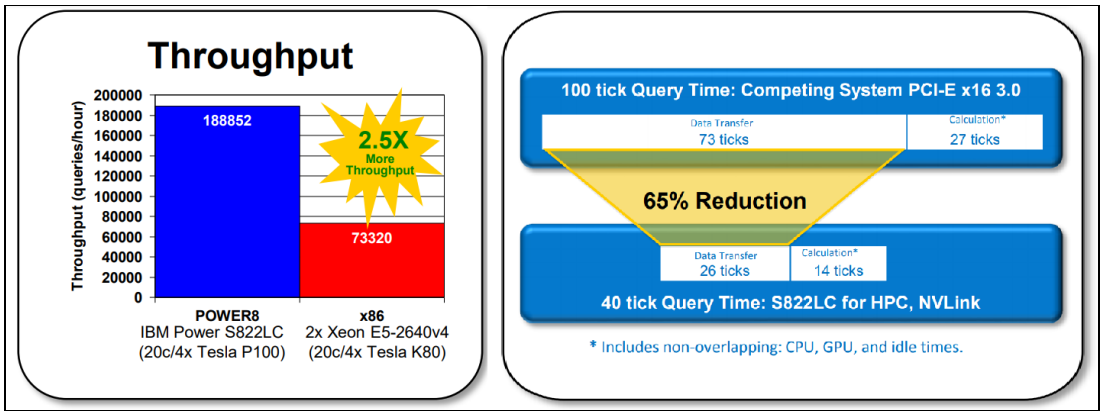

Figure 3-16 Application performance with Kinetica on IBM POWER8 with NVIDIA NVLink 1.0 *versus Kinetica on x86 server with PCIe Gen3 x16*

The second test compares throughput performance with Kinetica on IBM POWER9 with NVIDIA NVLink 2.0 and Kinetica on IBM POWER8 with NVIDIA NVLink 1.0, as shown in Figure 3-17. The test results show that the Power AC922 delivers 1.8× more queries per minute than the Power S822LC for HPC server.

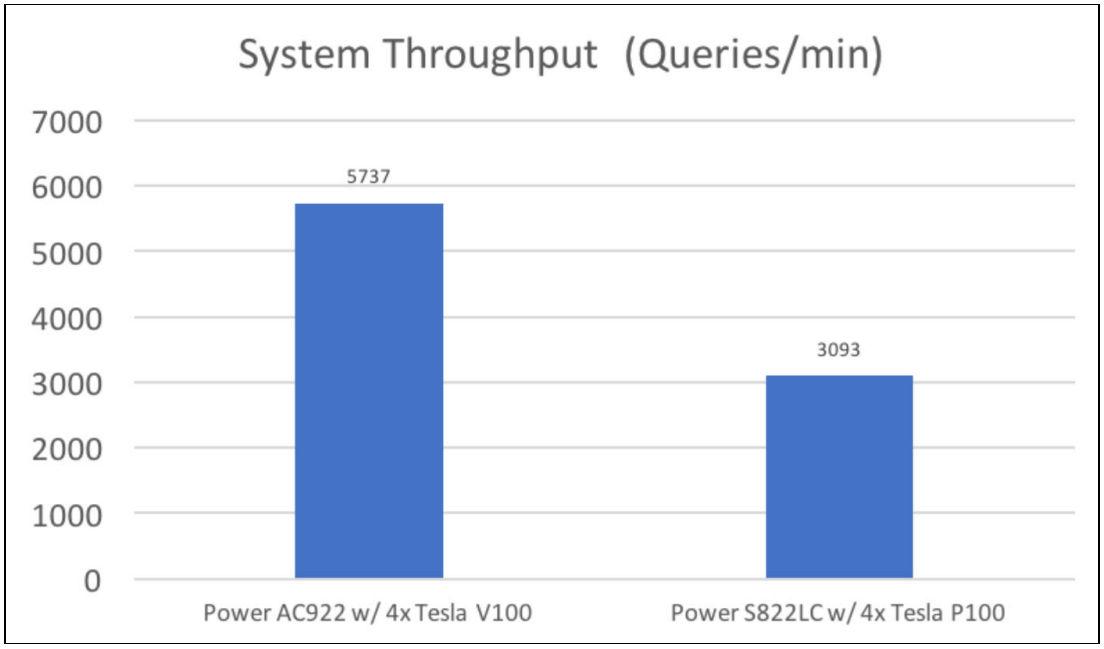

Figure 3-17 Application performance with Kinetica on IBM POWER9 with NVIDIA NVLink 2.0 *versus Kinetica on IBM POWER8 with* NVIDIA NVLink 1.0

In both tests, you can see that Kinetica gains additional performance benefits from the IBM POWER9 architecture, which offers NVIDIA NVLink 2.0 uniquely between CPUs and GPUs. In addition, the fact that the IBM POWER9 processor is equipped with NVIDIA NVLink 2.0, which is approximately 1.85× better per CPU and GPU segment than NVIDIA NVLink 1.0, is an important factor for further performance improvements in Kinetica.

Note: Throughput results are measured by Kinetica for filtering Twitter Tweets, based on running a "`Filter by geographic area`" query. See the IBM Website for more information:

`https://developer.ibm.com/linuxonpower/perfcol/perfcol-bigdata/`

Recommended architecture and Hardware configuration

You can refer to the system overview that Kinetica requires Power AC922 with:

- 2 IBM POWER9 processors and 4 NVIDIA V100 GPUs
- 1 TB memory
- 2 SFF (HDD/SSD), SATA Up to 7.7 TB Storage
- Support 1.6 TB and 3.2 NVMe adapters
- Redundant hot swap power supplies

4

Use cases

This chapter includes the following client use cases:

4.1 Watson Machine Learning Accelerator on Power AC922

The following section describes a Banking solution.

4.1.1 Industry: Banking

The following scenario desribes a solution for the banking industry.

Challenge

The client was looking for a solution that could improve managing risk, financial reporting, monitoring, and preventing fraud and financial crimes. The big challenge here was that the client was dealing with a vast amount of data, because it was crucial to keep the client's historical details. Another important part of the solution was that it needed to be very time sensitive because it could cause big financial, regulatory, and reputational impact.

IBM solution

IBM provided a solution that includes Power AC922 and Watson Machine Learning Accelerator. This was the best solution for the client because Power AC922 together with NVIDIA V100, NVLink 2.0, and PCIe Gen4.0 brought much better performance. IBM won this bid due to a best-in-class, high-performance computing infrastructure with unmatched performance and reliability. The IBM team showed extraordinary attention to details for the client requirements to not only be met, but exceeded. Power AC922 and Watson Machine Learning Accelerator won the competition against x86.

Business values

The new solution addressed all of the client's challenges, providing faster development, increased speed of execution, more straightforward management, as well as automation and organizational efficiencies. The client can see improved performance on their critical open-source, data-centric workloads, and increased availability with reduced costs.

4.2 PowerAI Vision on Power AC922

The following section describes a health care solution.

4.2.1 Industry: Health care

The following scenario desribes a solution for the health care industry.

Challenge

Hundreds of thousands of lives and billions of dollars could be saved by detecting disease sooner and reducing unnecessary overtreatment. For example, it is critical to distinguish between people with stage 2 tumors who need chemotherapy to improve survival and to avoid unnecessary chemotherapy. Each time that the researchers had to look at a tissue sample under a microscope and compare specific tissues was essential for producing results, but it was time-consuming.

Clients wanted to apply artificial intelligence to this visual inspection process and needed technology portability to test not only one specific cancer but other types of cancer. There were also requirements that the service be delivered quickly, accurately, and be scalable.

IBM solution

Working with an IBM Business Partner, the client continued development of image analysis algorithms to enhance its existing platform, based on IBM PowerAI Vision. Running on Power AC922 accelerated servers hosted by an IBM Business Partner, these enhancements use deep learning models to augment their proprietary data and assays, to classify resected tissue samples.

Power AC922 servers pair POWER9 processors and NVIDIA V100 GPU with NVIDIA NVLink to provide massive throughput capability for high-performance computing, deep learning, and artificial intelligence workloads, as shown in Figure 4-1:

- Power AC922 Server
- PowerAI Vision

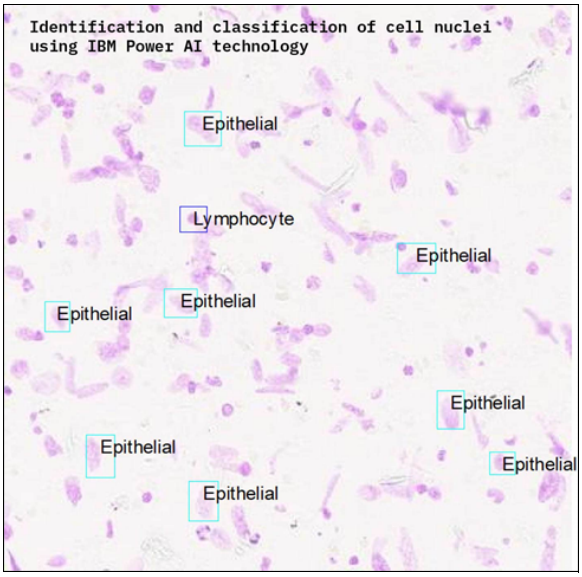

Figure 4-1 Example of identification and classification of cell nuclei using IBM PowerAI Vision

Business values

The client worked with IBM Business Partner to deploy IBM PowerAI Vision running on IBM Power systems, accelerating digital image analysis of tumors to identify risk of some cancer recurrence. According to an announcement from the client, the project has made the following improvements:

- Analyze images more than 10× faster to test more patients at the same time
- Activating the new standard of care improves patient outcomes and increases efficiency
- Shorter time-to-market and accelerated scaling make the client able to deliver products internationally

Note: More detailed Case Study content can be found on the following IBM website:

https://www.ibm.com/case-studies

4.3 H2O Driverless AI on Power AC922

The following section describes a financial services solution.

4.3.1 Industry: Financial services

The following scenario desribes a solution for the financial services industry.

Challenge:

The company was exploring new ways to enhance the client experience and improve competitiveness in the market. As part of the AI transformation journey, the client planned to set up an AI lab with use cases around predictive maintenance, cyber security, and credit scoring enhancements.

IBM solution

The IBM team showed PoC how easy and effective H2O Driverless AI is while running on Power AC922. The IBM team proved that running H2O Driverless AI on Power AC922 provides significant, better performance and scalability compared to the client's current x86 infrastructure.

The team also emphasized the collaboration and innovation between IBM and NVIDIA, leveraging Summit, the world's most powerful and smart AI supercomputer, as a reference. The client agreed that H2O Driverless AI running on Power AC922 was the better choice. The client decided to purchase the solution.

Business values

Feature engineering, model tuning, and visualization are handled automatically by the software. The client improved training times of deep learning frameworks by nearly 4× and data movement between the POWER9 CPUs and NVIDIA V100 GPUs up to 5.6× compared to the PCIe Gen3 buses used in x86 systems.

4.4 SQream DB on Power AC922

The following section describes a mobile carrier solution.

Industry: Mobile Carrier

The following scenario desribes a solution for the cell phone industry.

Challenge

The client has had difficulty running a rapidly growing data store. Although Hadoop storage collected various log data and call quality data, there was no environment for proper analysis. For analysis, it took a very long time to load data into the memory of an existing analytics system, and the amount of data that could be accessed at one time was very limited. This made it difficult for clients to perform their tasks, including proper call quality analysis.

IBM Solutions

IBM worked with SQream to create an Power AC922 server for GPU-based accelerated computing, a storage environment for sharing data, and an InfiniBand EDR network communication environment capable of handling large I/O. The Power AC922 server maximizes the processing performance of SQream DB based on NVIDIA NVLink 2.0, and it can be equipped with a large memory so that SQream DB can secure up to 2 TB of file cache area and use it to improve performance.

In addition, each IBM POWER9-based server is equipped with a PCIe Gen4-based InfiniBand EDR adapter to secure a 100 GBps bandwidth between the Power L922 and the Power AC922 server that provide file services via IBM Spectrum Scale. The IBM FlashSystem® 9100, which is based on NVMe modules, ensures stability and outstanding performance:

- Power AC922 with 4 NVIDIA V100 GPUs (32 GB HBM)
- Power L922 with IBM Spectrum Scale Licenses
- FlashSystem 9100
- Mellanox Infiniband EDR Switch

Business values

With the introduction of the Power AC922 and SQream DB, the client has the environment to handle complex queries that could not be performed previously. Today, the range of data used in traditional analytics is increasing from daily, to weekly, to months. Based on the newly acquired analysis environment, we have established new challenges related to new call quality analysis and monitoring that were not feasible before, and are currently working on those.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM Power System AC922 Technical Overview and Introduction,* REDP-5494
- *IBM PowerAI: Deep Learning Unleashed on IBM Power Systems Servers*, SG24-8409
- *AI and Big Data on IBM Power Systems Servers*, SG24-8435

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM Knowledge Center

  http://www.ibm.com/support/knowledgecenter/
- IBM Knowledge Center: IBM Power Systems Hardware

  https://www.ibm.com/support/knowledgecenter/POWER9/p9hdx/POWER9welcome.htm
- IBM Case Studies

  https://www.ibm.com/case-studies
- IBM cognitive computing standard:

  https://cognitivecomputingconsortium.com/resources/cognitive-computing-defined/#1467829079735-c0934399-599a
- For more information about Watson Machine Learning Accelerator, read:
  https://www.ibm.com/support/knowledgecenter/SSFHA8_1.2.1/wmla_overview.html
- For more information about Watson Machine Learning Community Edition, read:
  https://developer.ibm.com/linuxonpower/deep-learning-powerai/releases/
- For more information about PowerAI Vision visit:
  https://www.ibm.com/us-en/marketplace/ibm-powerai-vision
- BM Caffe with Large Model Support Source code
  https://github.com/ibmsoe/caffe/tree/master-lms
- For more information about performance test tests read**:**
  https://developer.ibm.com/linuxonpower/2018/12/19/performance-of-3dunet-multi-gpu-model-for-medical-image-segmentation-using-tensorflow-large-model-support/

- SQream information:

  https://openpowerfoundation.org/wp-content/uploads/2018/10/David-Leichner.IBM-OpenPOWER-SQream-POWER9.pdf
- SQream published a brochure introducing SQream DB on Power AC922.

  https://info.sqream.com/hubfs/pdf/SQream%20DB%20for%20Power%20Systems.pdf
- For more information on Kenetica, see:

  https://www.kinetica.com/

  https://developer.ibm.com/linuxonpower/perfcol/perfcol-bigdata/

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5555-00

ISBN 0738458112

Printed in U.S.A.