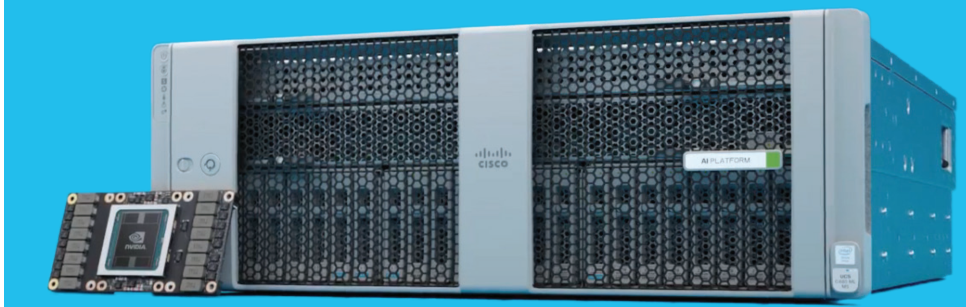
SAS® Visual Data Mining and Machine Learning on Cisco UCS C480 ML M5 Rack Server

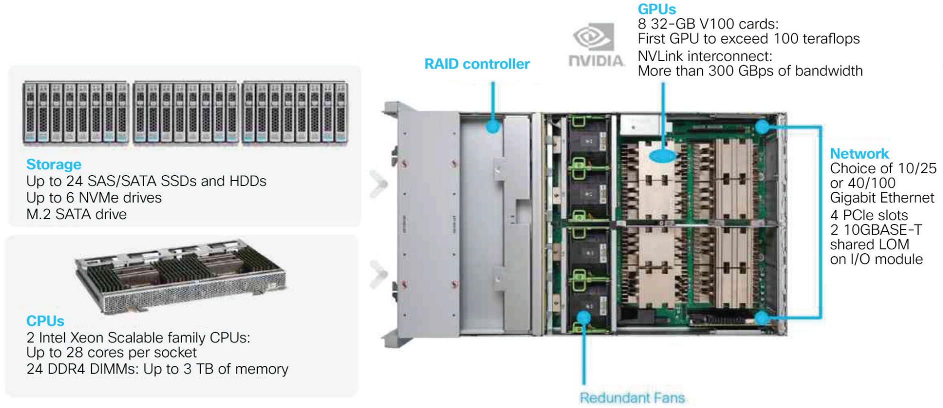
Real-Time Image and Video Analytics Training Platform Solution

Contents

Cisco UCS C480ML



NVIDIA Tesla V100



Storage
Up to 24 SAS/SATA SSDs and HDDs
Up to 6 NVMe drives
M.2 SATA drive

CPUs
2 Intel Xeon Scalable family CPUs:
Up to 28 cores per socket
24 DDR4 DIMMs: Up to 3 TB of memory

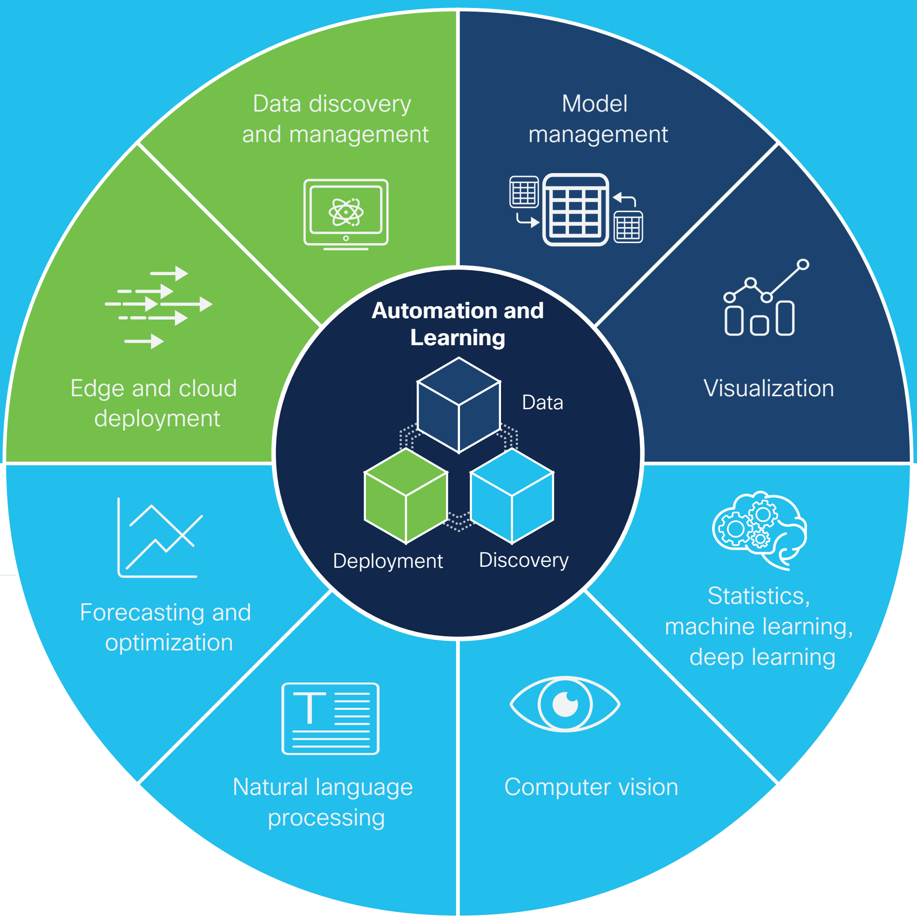
RAID controller

GPUs
8 32-GB V100 cards:
First GPU to exceed 100 teraflops
NVLink interconnect:
More than 300 GBps of bandwidth

Network
Choice of 10/25
or 40/100
Gigabit Ethernet
4 PCIe slots
2 10GBASE-T
shared LOM
on I/O module

Redundant Fans

SAS Viya Platform



Data discovery and management
Model management
Edge and cloud deployment
Visualization
Automation and Learning
Data
Deployment
Discovery
Forecasting and optimization
Statistics, machine learning, deep learning
Natural language processing
Computer vision

Faster decision making through faster training of computer vision models

The SAS and Cisco® combined real-time image and video analytics training platform solution enables quicker decision making through faster training of computer vision models, from testing, development, and training to inference.

Figure 1 shows the Cisco options that support the computer vision modeling from start to finish. Figure 2 provides a detailed view of the way that SAS® Viya® processes image and video data. Streaming images are handled by SAS Event Stream Processing.

The combination of Cisco UCS and SAS Viya accelerates computer vision innovation by providing a ready-to-use solution.

- SAS Viya and Cisco UCS C480ML servers together provide a comprehensive solution.
    - The solution offers a comprehensive platform for video and image analytics, including biomedical image analytics.
    - The GPU enables faster video and image analytics based on deep learning.
    - The solution provides training workloads at scale for very large data sets such as those needed for medical imaging, retail, and manufacturing data.
- Take advantage of the workload capabilities of the Cisco UCS C480ML, with 8 GPUs and 112 CPU cores.
    - Track and monitor workload performance on the GPUs and CPUs.
    - Simulate multiple users by simultaneously running multiple jobs.
    - Monitor GPU temperature, GPU memory, and processor use.
    - Unify and simplify management with customer choice and a validated design.
- Extend SAS Viya computer vision and deep learning expertise to medical imaging.
    - The solution enables object detection and image processing and recognition.
    - Implement supervised, unsupervised, and semisupervised transfer and deep learning.

Figure 1. Cisco options for computer vision modeling

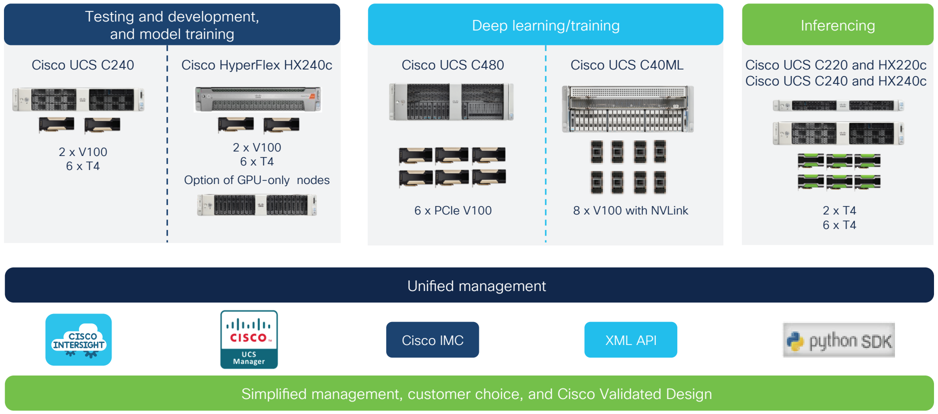
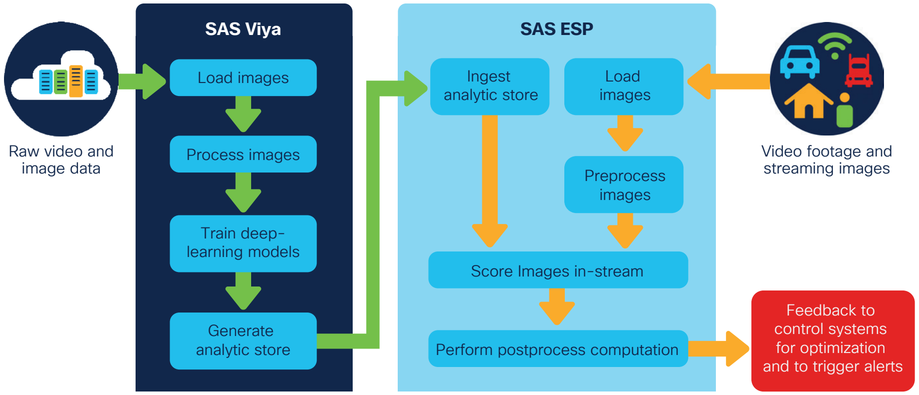


Figure 2. SAS Viya processing of image and video data



Solution performance

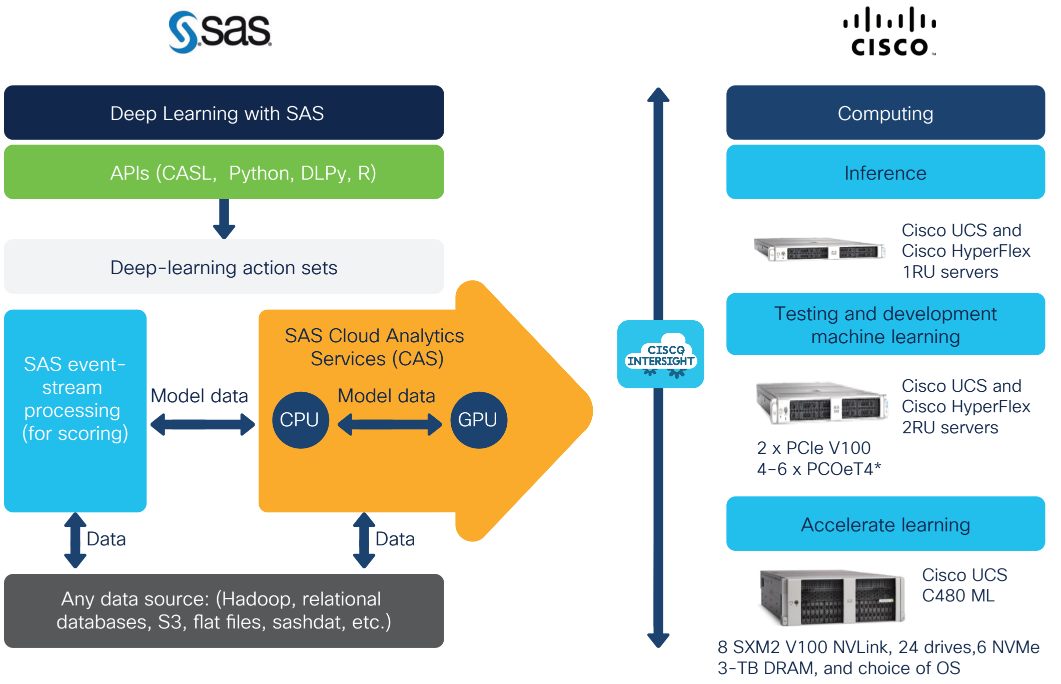
Medical centers and hospitals are now turning to computer vision technology to facilitate the assessment process and augment the work of radiologists. Analysis of images using computer vision requires the training of deep learning models. Deep learning is data intensive and often requires thousands of labeled or pre-identified images for training. In addition, images are often not in a final state useful for training and must be preprocessed to create images optimal for input into the training model. Previous research has shown that the optimal data set size for automated classification of chest radiographs is about 20,000 images, with marginal increases in model accuracy up to about 200,000 images.1

The modeling workflow used for this solution uses SAS Visual Data Mining and Machine Learning to prepare the NLST-CT image data set and train the deep learning model. This data set contains CT scan images of patients who underwent a screening trial for lung cancer. The data set contained approximately 30,000 images. A ResNet50 convolutional neural network (CNN) was used to train the data to identify the relevant features. A Cisco Validated Solution was created to track the workload performance of the eight NVIDIA graphics processing units (GPUs) configured on the Cisco UCS® C480 ML M5 Rack Server.

With SAS Visual Data Mining and Machine Learning and the Cisco UCS C480 ML M5 Rack Server, you can take advantage of the synergy between the two to rapidly train, deploy, and score deep learning models.

The SAS Platform architecture for deep learning (Figure 3) uses massively parallel processing and parallel symmetric multiple processors (SMPs) with multiple threading for extremely fast processing. One of the GPU processors is provided with SMP servers. Real-time training and scoring are supported by SAS Event Stream Processing. The architecture is enabled by the computing capabilities provided by the Cisco Unified Computing System™ (Cisco UCS) servers, which support deep learning all the way from training, testing, and development to inference.

Figure 3 SAS Platform architecture for deep learning


sas
Deep Learning with SAS
APIs (CASL, Python, DLPy, R)
Deep-learning action sets
SAS event-stream processing (for scoring)
Model data
SAS Cloud Analytics Services (CAS)
CPU
Model data
GPU
Data
Data
Any data source: (Hadoop, relational databases, S3, flat files, sashdat, etc.)
CISCO INTERSIGHT
cisco
Computing
Inference
Cisco UCS and Cisco HyperFlex 1RU servers
Testing and development machine learning
Cisco UCS and Cisco HyperFlex 2RU servers
2 x PCIe V100
4-6 x PCOeT4*
Accelerate learning
Cisco UCS C480 ML
8 SXM2 V100 NVLink, 24 drives,6 NVMe
3-TB DRAM, and choice of OS

[1] Assessment of Convolutional Neural Networks for Automated Classification of Chest Radiographs, Jared A. Dunnmon, Darvin Yi, Curtis P. Langlotz, Christopher Ré, Daniel L. Rubin, and Matthew P. Lungren,

Figures 4 and 5 show the workflow performance after the Cisco Validated Design is implemented. Figure 4 shows that the number of images processed per second varies greatly with the number of GPUs. Without a GPU, only 0.75 image is processed per second. The addition of one GPU increases that number to 22.29 images per second. With eight GPUs, 112.84 images are processed per second: a fivefold increase in the number of images.

Figure 5 shows a similar trend for model training time. Adding one GPU dramatically reduces training time from 10,700 seconds to 235 seconds. The use of eight GPUs reduces the time to 72 seconds. Overall, the training time was reduced from about 173 minutes to a little more than one minute for a dataset size of 4.6GB. This dramatic reduction in training time can help boost data scientists productivity, allowing them to experiment with models until they find the one best able to solve the problem at hand. The figure shows the effect of data size on the training time. As the data size increases with no GPU, training time increases dramatically. With the 1 and GPU and 8 GPU, training time drastically reduces even as the data size increases.

The main findings from the workload performance evaluation for training deep learning models are the following:

- The use of GPUs makes a dramatic difference in training time, achieving over 90 percent efficiency.
- Training time continues to decrease as you add more GPUs.
- Processing time decreases but at a decreasing rate as you continue to add GPUs.
- Multiple GPU workloads are specific and time sensitive.
- Workload management is critical to optimal use of the hardware. In the testing described here, GPU utilization started at 0 and increased to 98 percent when all eight GPUs were in use. And even though the GPU utilization was at 98 percent, the temperature of each of these GPUs was at an optimal level because of the unique cooling system design of the Cisco UCS 480 ML M5.
- The Cisco UCS C480 is an excellent choice for running multiple machine learning processes at the same time.

Typically, GPUs are associated only with model training efficiency because model scoring is inherently less complex. However, the workload performance evaluation described here shows that the use of a GPU provides a clear advantage for scoring, with about an 80 percent gain in efficiency.

Figure 4 Workflow performance in images per second by number of GPUs

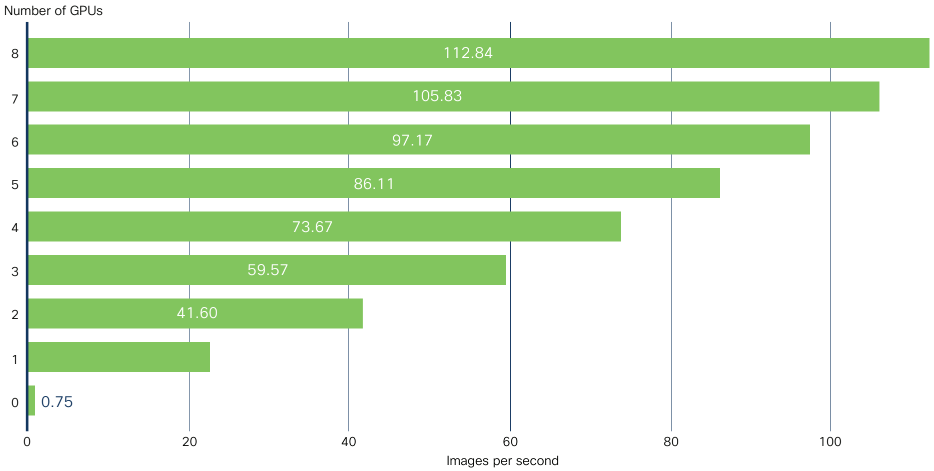
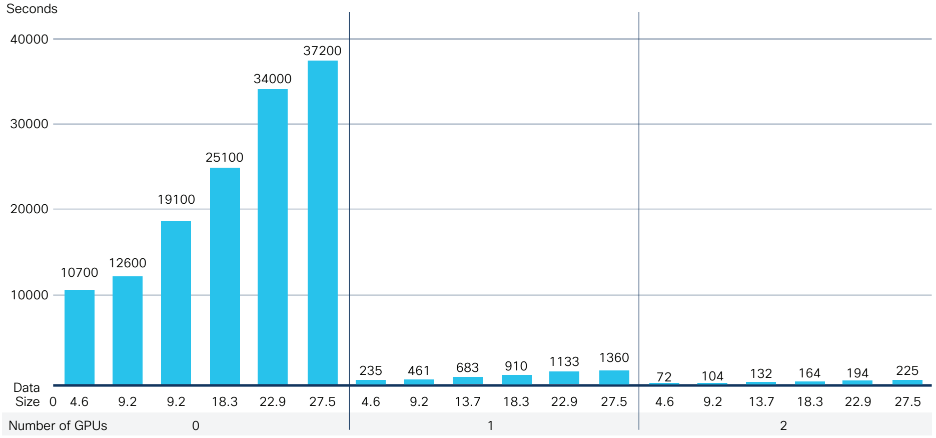


Figure 5 Model training time in run-time seconds by number of GPUs



Benefits

The SAS and Cisco solution offers these benefits:

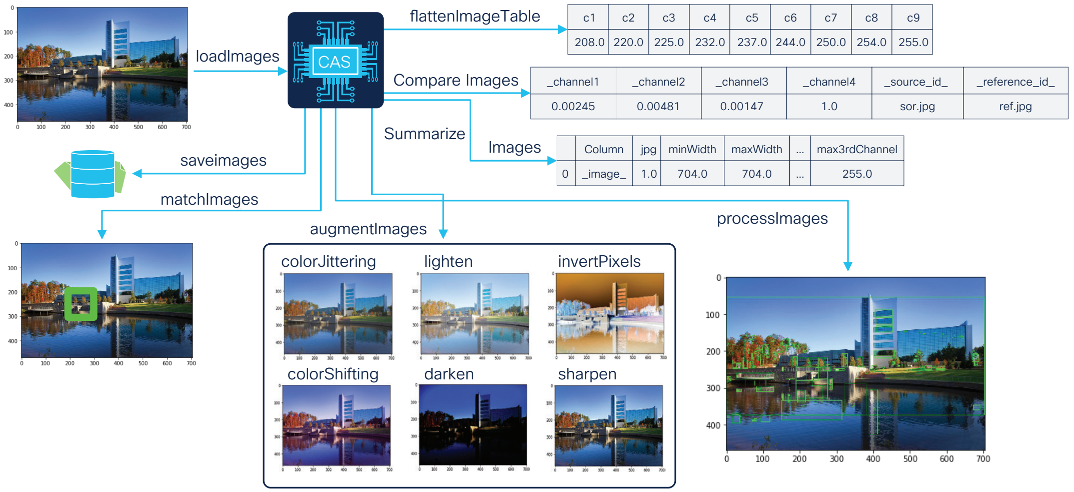
- Quickly deploy your biomedical image analysis infrastructure with a purpose-built server for deep learning.
- Experience super-fast deep learning model training times through the use of eight NVIDIA modules.
- Simplify model training with pretrained SAS deep learning models.
- Support end-to-end analysis of biomedical images with an automated data science and machine learning platform.
- Gain performance and capacity. The Cisco UCS C480 ML M5 offers flexible options for CPU, memory, networking, and storage while providing outstanding GPU acceleration.
- Resolve the uncertainties of the machine learning software ecosystem with validated solutions.
- Simplify operations with the Cisco Intersight™ platform to extend accelerated computing to the locations where it is needed.

Use artificial intelligence to simplify and enhance the radiology workload

Biomedical imaging, such as CT scans, PET scans, and MRIs, are noninvasive tests used by clinicians to diagnose and monitor patient health. These biomedical images show a visual representation of the human anatomy and can facilitate quantitative measurements. However, the number of images generated from these tests is huge, and the resulting workload of reviewing the images and providing an assessment is very time intensive for the radiologist. In addition, often these tests require other clinicians to serve as peer reviewers. The SAS and Cisco solution uses artificial intelligence (AI) to augment the radiologist's work not only by simplifying the workload but also by acting as a second opinion during the assessment.

Computer image processing is complex and requires computational time for preprocessing the images. Figure 6 shows the types of tasks required for image preprocessing. After images are loaded (loadImages), they commonly are augmented (augmentImages) with changes in brightness and contrast, sharpening, cropping, resizing, etc. so that the images are standardized. Images may also be flattened (flattenImageTable) to reduce file size. Other tasks include comparison of images across image tables, summarization of image characteristics, and matching of images against a table of images.

Figure 6 Examples of tasks required for image preprocessing



c1	c2	c3	c4	c5	c6	c7	c8	c9
208.0	220.0	225.0	232.0	237.0	244.0	250.0	254.0	255.0

_channel1	_channel2	_channel3	_channel4	_source_id_	_reference_id_
0.00245	0.00481	0.00147	1.0	sor.jpg	ref.jpg

	Column	jpg	minWidth	maxWidth	...	max3rdChannel
0	_image_	1.0	704.0	704.0	...	255.0

Cisco UCS C480 ML M5 AI platform

The Cisco UCS C480 ML M5 Rack Server is specifically built for deep learning. It is storage and I/O optimized to deliver industry-leading performance for training models. The Cisco UCS C480 ML M5 delivers outstanding storage expandability and performance options for standalone or Cisco UCS managed environments in a 4-rack-unit (4RU) form factor. The Cisco UCS C480 ML M5 offers flexible options for CPU, memory, networking, and storage while providing outstanding GPU acceleration. The Cisco UCS C480 ML M5 is designed for the most computation-intensive phase of the AI and machine learning process: deep learning. This server integrates GPUs and high-speed interconnect technology with large storage capacity and up to 100-Gbps network connectivity.

With Cisco UCS C480 ML M5 servers, Cisco offers a complete range of computing options sized to each element of the AI lifecycle: data collection and analysis near the edge, data preparation and training in the data center core, and real-time inference at the heart of AI. Cisco's cloud-based management makes it easy to extend accelerated computing to the right locations across an increasingly distributed IT landscape.

The combination of NVIDIA GPUs and Cisco servers delivers outstanding acceleration and manageability along with world-class support to ease installation and operation. Cisco UCS C480 ML M5 servers with NVIDIA Tesla V100 Tensor Core GPUs provide excellent performance that is easy to consume as part of the Cisco UCS platform with the industry's only uniform, cloud-powered, automated operations model.

The Cisco UCS C480 ML M5 offers these features and benefits:

- **GPU acceleration**: Eight NVIDIA Tesla V100 SXM2 32-GB modules are interconnected with NVIDIA NVLink technology for fast communication across GPUs to accelerate computing. NVIDIA specifies TensorFlow performance of up to 125 teraflops per module, for a total of up to 1 petaflop of processing capability per server.
- **Internal NVLink topology**: NVLink is a high-speed GPU interconnect. Eight GPUs are connected through an NVLink cube mesh. Each NVLink interconnect is capable of 25 GBps of send and receive processing, for a total bandwidth of about 300 GBps among the eight GPUs.

Cisco high-computing machine learning servers enable AI models that would otherwise consume weeks of computing resources to be trained in a few hours. With this dramatic reduction in training time, a whole new world of problems will now be solvable with AI. Cisco is providing IT with a scalable solution for deep learning at enterprise scale. All Cisco servers–Cisco UCS B-Series Blade Servers, C-Series Rack Servers, and S-Series Storage Servers and Cisco HyperFlex™ servers–can be managed with a single tool: the Cisco Intersight platform. The Cisco Intersight platform provides cloud-based system management augmented by analytics and machine learning. It enables organizations to achieve greater automation, simplicity, and operational efficiency. It provides a holistic and unified approach to managing distributed computing environments regardless of the server form factor, workload, or location.

SAS® Visual Data Mining and Machine Learning

SAS Visual Data Mining and Machine Learning provides the following main capabilities:

- Easy-to-use analytics: Best-practices templates enable a quick, consistent start to model building and help ensure consistency among the analytics team. Analytical capabilities include clustering, different types of regression analysis, random forest models, gradient boosting models, support vector machines, natural language processing, and topic detection.
- Computer vision and biomedical imaging: You can acquire and analyze images with model deployment on server, edge, and mobile devices. The solution supports end-to-end flow processing for biomedical image analysis, including annotation of images.
- Deep learning with Python and Open Neural Network Exchange (ONNX) support: Python users can access high-level APIs for deep learning functions within Jupyter notebooks through the SAS deep learning with Python (DLPy) open source package on GitHub. DLPy supports ONNX, allowing easy movement of models between frameworks.
- Highly scalable in-memory processing: Concurrently access data in memory in a secure, multiuser environment. Data and analytical workload operations can be distributed across nodes, in parallel, and multithreaded on each node for very fast speeds.

SAS and Cisco Services boost solution success

Cisco accelerates business agility and infrastructure efficiency with IT expertise, technical services, and machine learning training. We offer Cisco UCS lifecycle services directly and through certified partners to deliver faster data center transformation.

The SAS Artificial Intelligence Center of Excellence (SAS AI CoE) is a group of PhD-level experts in AI, machine learning, natural language processing, computer vision, optimization, and simulation who are focused exclusively on customer implementations. This group is highly attuned to customer needs and combines a business-focused mindset with deep technical expertise to address business challenges and conduct assessments to uncover innovative opportunities that have business value.

Improving developer productivity

If you can provide your data scientists with a prebuilt image processing infrastructure with a machine learning and data science platform sitting on top of it, then you have already improved their productivity. Such a solution frees valuable time that data scientists can use to iterate and experiment with their computer vision models, thus allowing them to create innovative solutions to the problems before them.

To be able to process such large volume of images in real time, a purpose-built, easy-to-manage GPU-based platform is needed. This platform should meet not only the needs of data scientists but also the needs of IT engineers for ease of deployment and manageability.

The Cisco and SAS advantage: Right-sized solutions for every phase of AI and machine learning projects

To be able to process the large images used by radiologists and run multiple iterations applying various transformations, you need a suitable platform that can scale and that is also easy to manage. Cisco offers infrastructure solutions that are sized appropriately for every phase of AI and machine learning projects across the organization: from testing and development, to model development, to training, and to inferencing. The combined SAS and Cisco solution enables data scientists to train and tune machine learning models using a right-sized computing platform, while supporting IT in its deployment of a distributed, next-generation application backed by AI.

Cisco Capital financing to help you achieve your objectives

Cisco Capital® financing can help you acquire the technology you need to achieve your objectives and stay competitive. We can help you reduce capital expenditures (CapEx), accelerate your growth, and optimize your investment dollars and ROI. Cisco Capital financing gives you flexibility in acquiring hardware, software, services, and complementary third-party equipment. And there's just one predictable payment. Cisco Capital financing is available in more than 100 countries. Learn more.

220087-5 02/20

Get started now

See the following resources for additional information:

- For more information about SAS computer vision technology, see https://www.sas.com/en_us/offers/19q2/seeing-is-believing-computer-vision-from-sas.html.
- For more information about SAS Visual Data Mining and Machine Learning, see https://www.sas.com/en_us/software/visual-data-mining-machine learning.html.
- For more information about the SAS AI CoE, see https://www.sas.com/content/dam/SAS/en_us/doc/servicebrief/sas-artificial-intelligence-coe-109593.pdf.
- For an introduction to deep learning for computer vision with a guide to build deep learning models using SAS, see https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/deep learning-with-sas-109610.pdf.
- For more information about the complete portfolio of AI and machine learning solutions on Cisco UCS, see http://www.cisco.com/go/ai-compute.
- For more information about the Cisco UCS 480 ML M5 platform, see https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-741211.html.
- For Cisco UCS C480 ML M5 performance characterization, see https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/whitepaper-c11-741689.pdf.
- SAS Deep Learning with Python (DLPy) open source package on GitHub, here is the link for that: https://github.com/sassoftware/python-dlpy.