

Cisco UCS Infrastructure for Video Analytics with Awiros Operating System



Solution highlights

- Cisco UCS M5 servers are designed to deliver great performance, expandability, and efficiency for storage, I/O-intensive infrastructure and AI/ML workloads.
- Cisco UCS C240 M5 servers from the fifth generation compute portfolio, are well-suited for AI/ML model training/inferencing with the support of up to two PCI Express (PCIe) GPU adapters for industry leading GPUs of NVIDIA. Cisco C240 M5 servers are best suited from data collection, data preparation, training and analysis at the edge.
- Cisco C480 ML M5 rack server, developed in partnership with NVIDIA, a leader in AI computing, supports eight NVIDIA Tesla V100 Tensor Core GPUs with NVIDIA NVLink interconnect. The V100 is the world's first GPU to break the 100 teraflops barrier of deep learning performance with a whopping 640 Tensor Cores. NVLink provides 10x the bandwidth of PCIe and connects all of the GPUs in a point-to-point network (hybrid cube mesh) that provides optimal performance for these super-fast GPUs.
- Cisco's AI/ML/DL solutions are delivered as part of an integrated system that supports processing data whether it is at the edge or in the datacenter regardless of which server in the portfolio is the best match to solve business problems.
- With Cisco solutions, you can power artificial intelligence (AI) workloads at scale and help extract more intelligence out of data to make better decisions in real time. The system also provides maximum performance that's easy to consume as part of the UCS platform with the industry's only uniform, cloud-powered and automated operations model.
- Awiros is universally adoptable by any video surveillance hardware vendor and easily scalable with respect to adding the number of cameras to an existing location or adding the number of applications on a single camera or multiple applications on multiple cameras.
- Awiros stands apart from the other industry-leading Intelligent Video Analytics by giving you all the right tools including a suite of video decoders, a database manager, a unified user-interface and deployment manager to easily transform your ideas into market ready solutions.
- Key platform elements include Awiros Appstack—an app repository for ready-to-deploy apps, Awiros Grid—for viewing events of interest from all of the active apps, Awiros Coach—an Automated Training Orchestration Module (ATOM) to annotate and label objects to train the underlying models and Awiros Console—a comprehensive dashboard tool for visualization and reporting of data.
- Awiros has been tested to work seamlessly on the Cisco UCS C240 M5 servers and these servers together with Awiros power the data center with an industry-leading operating system, making it AI-ready from day one.

Video Analytics on Cisco UCS Platform

Challenge

With public security becoming a primary concern for governments and enterprises across the world, the installed base of surveillance cameras is growing at tremendous pace. It is expected that there will be one surveillance camera for every humans in the next few years. More than 2000 petabytes of data is recorded globally using the video surveillance cameras on a daily basis, and most of it remains raw! These unprocessed videos need to be analyzed manually and retrospectively, to retrieve any meaningful information. Reviewing an event to find evidence can take hours or even days.

By deploying video analytics software, end users can leverage specific video data and convert it into actionable intelligence for functions such as marketing, health and safety and personnel management. Video analytics can be employed for real-time threat detection and prevention use cases.

Advancements in computer vision and artificial intelligence technologies provide an opportunity to extract more value out of the video content captured and stored through the surveillance system. However, this worldwide massive surveillance adoption also has generated a demand for enterprise-class, easy-to-use, and highly integrated video intelligence solutions that can be deployed as per custom requirement of the governments, law enforcement agencies and enterprises.

The Solution

Awiros is a unified video intelligence platform that converts raw video from surveillance cameras into actionable data in real time. Designed as an operating system, Awiros supports a wide variety of video intelligence applications (apps) through its AppStack, where any app on any camera can be deployed seamlessly.


Cisco Unified Computing System™ (Cisco UCS®) provides the most ideal platform to power Awiros for various target use cases because of its flexibility of components, compact form factor, and NVIDIA GPU support.

Cisco and Awiros engineering have collaborated extensively to install and test the performance of Awiros OS on Cisco UCS servers with NVIDIA Tesla V100 GPU cards.


Along with Cisco’s customers, many partners from across the technology ecosystem have propelled the Cisco UCS platform, collaborating to deliver ever more capable solutions. Cisco and Awiros have collaborated to develop the best-in-class Video Analytics AI/ML solution, a fully baked platform that accelerates machine learning with the Awiros suite of Intelligent Video Analytics product.

Why Awiros?


Awiros Operating System can handle any scale or complexity. Awiros seamlessly handles the development of apps for their consumption by users, with unmatched flexibility and robustness.

-
- 


Develop

 - Awiros provides all the necessary resources to transform your AI algorithms into deployable AI products.
 - **Awiros App Development SDK:** Awiros C++/Python SDK allows you to create and customize applications.
 - **Awiros Coach/ Model Training Module:** Awiros Coach allows you to annotate, train and retrain your models through a user-friendly interface.
 - **Awiros API Integration:** Awiros’ Data Exchange Server (DES) provides an easy-to-use plug-in and hooks for making REST API calls to third-party systems.
 - 

Deploy

Dynamically choose any app from the Awiros App Stack for deployment on any video source to seamlessly deploy on the edge, on cloud, on-premises or as a hybrid model.
 - 

Scale

Awiros enables you to conveniently scale across hardware, video sources and apps.
 - 

Manage

Awiros lets you allocate resources across locations, apps and video sources dynamically.

Cisco UCS for AI/ML

Cisco is expanding its Unified Computing System (UCS) to help organizations take full advantage of the growing demand for machine learning and artificial intelligence applications.

Cisco UCS C240 M5 Rack Server is a 2-socket, 2-Rack-Unit (2RU) rack server offering industry-leading performance and expandability. It supports a wide range of storage and I/O-intensive infrastructure workloads, from big data and analytics to collaboration. Built on the new second-generation Intel® Xeon® Scalable processors, UCS C240 M5 servers are ready to take on even more workloads, with up to double the memory capacity of previous systems. These servers are best for high-performance, graphics-intensive, video intelligence kind of applications and tailor made for edge-based analytics.

The Cisco UCS C480 ML M5 Rack Server is a purpose-built server for Deep Learning. It is storage, I/O optimized and designed for the most compute-intensive phase of the AI and ML lifecycle to deliver an industry-leading performance for deep learning training and inferencing applications. Video analytics being one of the most compute-intensive applications it can be easily scaled-up and scaled-out by deploying on C480 ML M5 rack servers. The deep learning framework employed for image analysis at real-time requires a combination of CPU and GPU resources coupled with a large I/O bandwidth. Cisco UCS C480 ML M5 offers flexible options for CPU, memory, networking, and storage while providing outstanding GPU acceleration.

With C480 ML M5 servers, Cisco offers a complete array 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.

The combination of NVIDIA GPUs and Cisco servers deliver unparalleled acceleration and manageability, combined with world-class support to easy installation and operation. By incorporating the NVIDIA NVLink architecture into the UCS form factor, the C480 ML M5 is able to provide highly-optimized GPU-to-GPU communication. Cisco UCS C480 ML M5 with the NVIDIA Tesla V100 Tensor Core GPUs provides maximum performance that’s easy to consume as part of the UCS platform with the industry’s only uniform, cloud-powered, automated operations model.

Cisco high-computing ML servers facilitate the AI models that would 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 and Cisco is arming IT with the scalable solution for deep learning at enterprise scale.

Awiros Intelligent Video Analytics

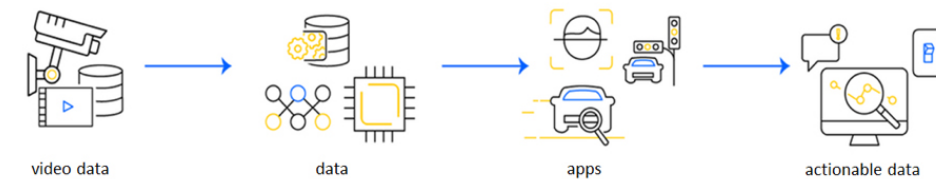
Real-time video data is the most prevalent and wide-spreading content type today. It forms more than 80 percent of the entire Internet traffic. Video intelligence is the science of using computer vision and artificial intelligence technologies to generate insights and notifications from video content from surveillance cameras or stored video files. Different industries may leverage specific video-intelligence “apps” for their own use cases. Figure 1 shows some of the most popular apps.

Figure 1. Various video intelligence apps for different target industries



Awiros is the first of its kind, a full-scale operating system for video intelligence that enables you to create, test, train, and deploy “apps” for image and video processing.

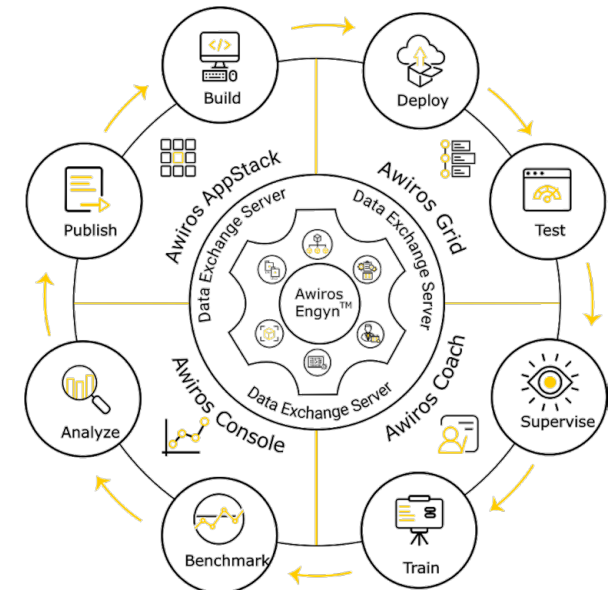
Figure 2. Video Intelligence: converting Video into actionable data



The development and deployment of a video intelligence application requires a platform with a comprehensive toolkit to enable you to manage the entire lifecycle. Awiros is the first such platform that has all the modules and components ranging from development of apps, to training of deep-learning models, to deployment of video intelligence on a large-scale infrastructure.

You can use the four key modules of the Awiros technology stack:

- AppStack:** Awiros AppStack is the repository of apps and has all the necessary tools for a developer to create and publish the apps. The AppStack gives you the flexibility to review and select a suitable app for your own use case.
- Grid:** Grid allows you to manage and deploy apps onto cameras; it provides all the events and notifications. Grid is the active interface of Awiros that provides real-time visibility of the status of each app with relation to cameras in the system.
- Coach:** Artificial Intelligence requires continuous learning and improvement, which Coach enables. Coach includes an annotation and labeling tool and a batched training module that uses the existing and user-uploaded data in the Awiros system to develop new models and weights.
- Console:** Console is the data-analytics dashboard that gives you all the key data points from the entire system of resources: cameras, apps, and events.



Cisco Awiros reference architecture

You can install Awiros on a wide range of hardware devices ranging from small low-powered computing platforms for processing video at the edge to an enterprise rack-mounted server for centralized processing of videos from a large number of video sources in a data center. For ease of installation and management, Awiros software stack is containerized with docker containers that in turn run on a Linux server; for example, Ubuntu server or Red Hat Enterprise Linux.

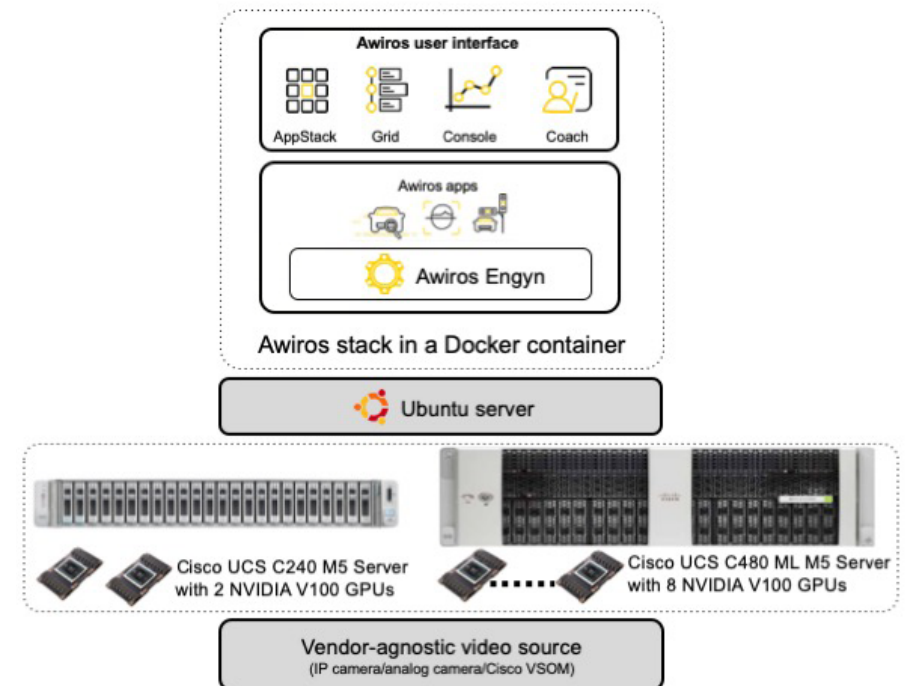
To set up the Awiros platform on either Cisco UCS 240 M5 or Cisco C480ML M5 Server, install Ubuntu server. On the Ubuntu server download and install the Awiros docker image on the Ubuntu server. The docker image contains all the key libraries and prerequisites for Awiros. NVIDIA drivers and toolkit is installed natively on the Ubuntu server to allow Awiros to connect to the GPU cards. When the docker image is ready, install the latest package components of Awiros Stack.

Solution components

Hardware/Software Components	Version	Description
Cisco UCS C240 M5 server	BIOS - 4.0.1c CIMC - 4.0.1a	Cisco UCS M5 server used for the validation
NVIDIA Tesla V100	32 GB	NVIDIA GPU card
Cisco Modular RAID Controller	12 G SAS	RAID Controller card
Storage	2 x 1.2 TB	HDDs for data
M.2 SATA SSD	240 GB	Modular M.2 or Secure Digital (SD) cards that can be used for boot
Memory	128 GB	DDR4 RAM
CPU	6134	Intel® Xeon® 2 x 8 core CPU
Ubuntu operating system	18.04 LTS	The base operating system for supporting the Awiros OS Instance
NVIDIA CUDA	9.2	NVIDIA GPU computation library for floating point operations
NVIDIA drivers	410.72	NVIDIA driver for the OS
NVIDIA cuDNN	7.5.1	Deep neural network library from NVIDIA

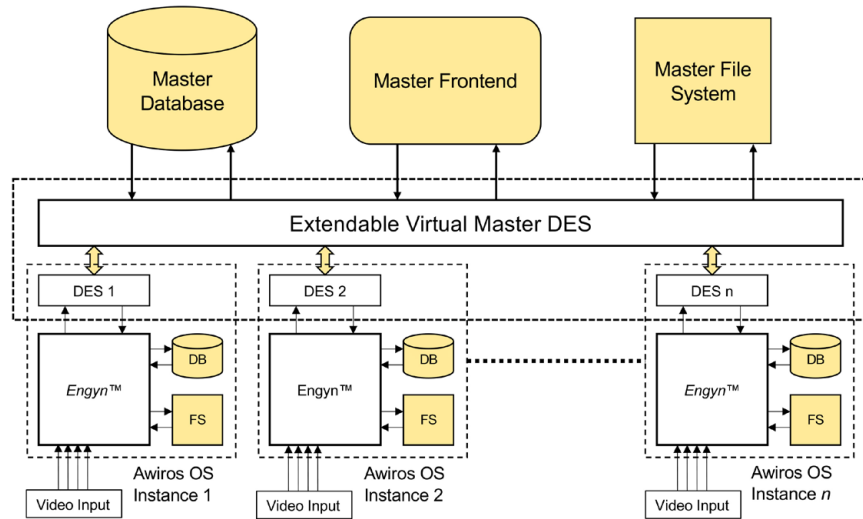
Hardware/Software Components	Version	Description
Docker daemon	18.06.1-ce, build e68fc7a	Docker daemon for supporting Awiros Docker Image
Awiros Docker image	v2.1	Awiros Docker image containing all supporting libraries and drivers
Awiros Engyn	v1.02	The kernel of the Awiros platform powering the computer vision and deep-learning based video processing
Awiros DES software	v1.00	The Data Exchange Server handling all the state and logistics of data between the user, Engyn, database and the file-system
Awiros FE package	v2.3	The complete user interface package

Figure 3. Cisco Awiros reference architecture



When the installation process is complete, create a user account to start management of cameras, apps, and other resources. The Awiros user interface provides complete visibility of all the cameras and video sources in the system, and it can integrate with any Video Surveillance Management Systems to connect to the video feeds.

Figure 4. Awiros architecture



Awiros is a fully replicable OS that can scale across multiple servers, locations and geographies. To handle a large number of cameras in a data center, Awiros is installed on individual servers as independent instances. One of the instances is then configured as the master instance that acts as an interface to all the other instances. You can assign more than one Cisco UCS C240 M5/ UCS C480 ML M5 Server (DES) with Awiros stack to handle workload of a large number of camera streams. Data Exchange Server (DES) is a consolidated management layer for handling data exchange with user-interface, database and the Awiros Engyn. Awiros Engyn is composed of a set of libraries built to support various functions of video intelligence apps. Engyn also provides a uniform abstraction of frameworks such as OpenCV, Dlib, TensorFlow and Keras for the apps.

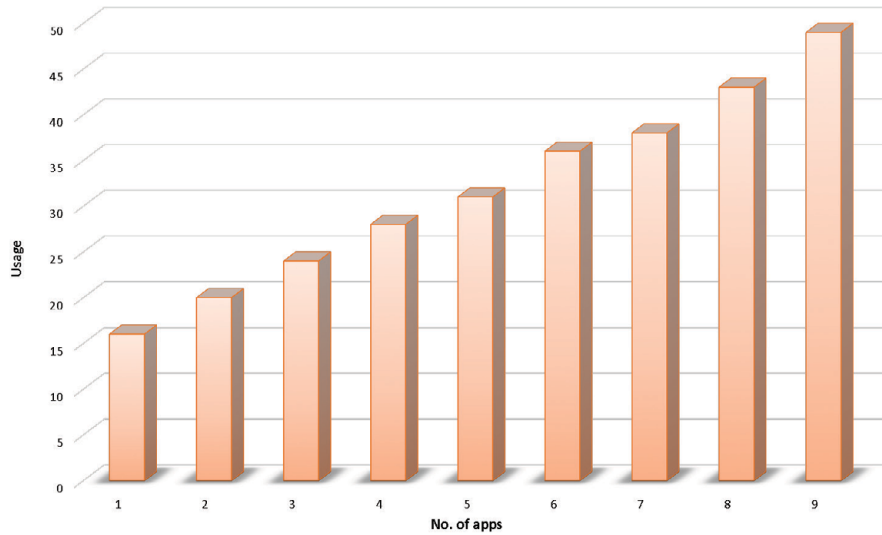
Cisco servers can be suitably placed in both scale-up architectures, where GPU scale is required and scale-out architectures, where compute systems are required to scale horizontally. Enterprise business looking for GPU density can leverage the C480 ML M5 high-computing servers to see best possible model training and inferencing for their AL/ML business use-cases. The total number of channels (cumulative number of apps running on the camera feeds) scales with the number of servers linearly. This allows for dynamic allocation and use of distributed hardware resources through the Awiros platform.

Solution Validation

We have done the solution validation with Cisco UCS 240 M5 servers. In the validation, we observed the GPU and CPU utilization with various Awiros applications. Also, we scaled-up the test environment with the second NVIDIA Tesla V100 GPU card and observed that the Awiros apps utilized both the GPUs in a Round-Robin fashion. We also validated a scale-out test by adding the second Cisco UCS C240 M5 server. We observed the GPU utilization and the way the apps were load balanced between the two servers. Cisco UCS C240 M5 servers with more cores, storage, GPUs and PCIe slots are well suited for video intelligence at a large scale in segments such as smart cities and enterprise security. With Awiros installed on the Cisco UCS 240 M5, running tens of applications simultaneously is easy.

For the solution validation, some of the Awiros apps that we have used include: face recognition, intrusion detection, people tracking, and speed detection. Awiros allows you to run apps on the CPU and, if available, on the GPUs to accelerate the processing of videos using state of the art deep-learning frameworks. For example, executing 10 instances of deep-learning based apps on CPU take up about 50 percent of the CPU resources as shown in **Figure 5**. As a rule of thumb, each instance of the deep-learning app uses about 2 virtual CPU cores, and Cisco UCS 240 M5 has 32 virtual cores. Figure 5 shows a linear increase in the CPU utilization w.r.t to the increasing number of app instances.

Figure 5. The CPU usage for deep-learning apps with the increase in number of app instances on Awiros Stack on UCS 240 M5

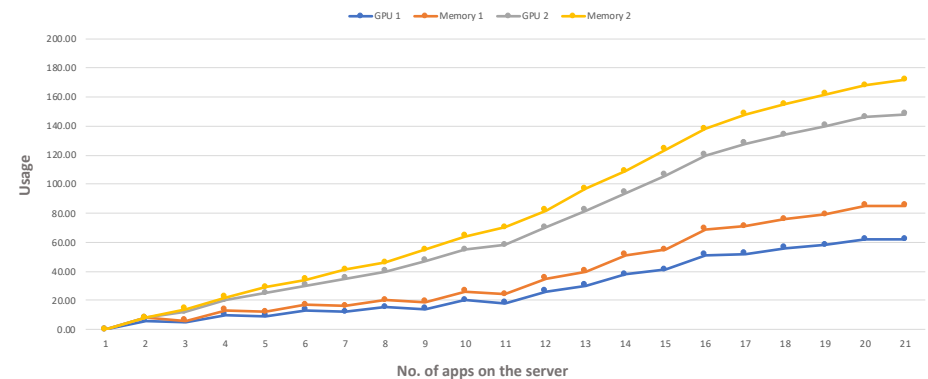


The performance evaluation of the Awiros Operating System and its video intelligence apps was done on the configuration outlined in the previous sections with up to 20 video streams from different sources with the parameters listed in the table:

Resolution	1280 x 720
Encoding	H.264
Frames per second	25 fps
Type of applications	All the application instances used deep-learning based inference engines for detection of objects, Awiros multi-object tracking the module and management engine at every frame: face detection, people tracking and speed detection.

However, the performance and the number of the deep-learning video intelligence apps that the UCS 240 M5 can support increases significantly with GPUs. Figure 6 shows the usage of processing and memory of each of the NVIDIA V100 GPUs on the server. With up to 20 deep-learning apps, each of the 2 GPUs uses about 50 percent of the processing power. In a real scenario, the server with 2 GPUs can support up to 50 applications simultaneously. The graph in **Figure 6** outlines the pattern of resource usage with respect to the number of apps on the Cisco UCS 240 M5. The memory usage remains relatively constant as the Awiros stack optimizes memory by reusing the deep-learning models loaded in the memory, and CPU usage increases linearly. Awiros' resource manager employs a round-robin scheduling mechanism in the background to assign apps to GPUs, and it maintains a record of the deep-learning Models loaded on each of the GPUs to optimize the memory usage. After a model is loaded on the memory of a GPU, the resource manager ensures that any new app using the same model does not reload the model to avoid redundancy in memory usage. This step helps assure that the usage of memory grows slowly compared to the processor usage.

Figure 6. The GPU processing and memory usage with the increase in number of deep-learning applications on the Awiros Stack running on UCS 240 M5



The Cisco UCS 240 M5 performance results provide great insights into the capability of Awiros to support apps on the runtime while optimizing the resource usage. Furthermore, for large-scale applications, you can use more than one server to dynamically allocate applications to computing resources to achieve the desired scalability. The architecture of multi-server Awiros deployment has been already described in the previous section. The performance of the multi-server deployment follows a pattern similar to that of the single device, because the Awiros resource manager evenly allocates apps to the computing platform.

Conclusion

Cisco UCS 240 M5 Servers deliver capabilities that extend the power and simplicity of unified computing for data-intensive workloads, applications at the edge, big data analytics, video intelligence, machine learning etc. Solution validation testing showed that you can run the video intelligence apps from the Awiros AppStack on a single Cisco UCS 240 M5 Server with a single NVIDIA V100 GPU, scaled-up with the second NVIDIA V100 GPU on the same server and scaled-out with the second Cisco C240 M5 Server (DES) and all these tests showed a linear scale of video intelligence apps. Awiros allows you to use the available hardware resources in the best possible way to empower the camera surveillance system. Cisco UCS AI/ML system with NVIDIA GPU along with Awiros provides a broader spectrum of features and unmatched performance to deliver enterprise-class video intelligence solutions. Cisco's AI/ML push targets one of the top priorities for businesses, which is to learn from the growing amounts of data, the fuel for AI and ML, to expand their competitive edge.

About Authors

Sindhu Sudhir—Technical Marketing Engineer, Cisco Systems Inc.

Sindhu Sudhir is part of Cisco UCS Solutions Engineering team. Her primary focus has been on Container technologies and infrastructure automation on Cisco UCS platform. In her current role, she is also working on AI/ML and video analytics solutions on UCS platform.

Vadiraj Bhatt—Principal Engineer, Cisco Systems Inc.

Vadi leads the UCS Analytics and BigData solutions in India. He has 25+ years of experience in developing and architecting Enterprise solutions. In his current role, he is leading Smart City and video analytics solutions development on UCS platform.

Vikram Gupta—Chief Architect and Founder, Awiros

Vikram Gupta is the founder and the chief architect of Awiros. Vikram is responsible for defining the vision, the product plan and the overall architecture of the core technology powering Awiros. Prior to founding Awiros, Vikram was working on developing an operating system for supporting multiple applications on embedded platforms while doing his PhD from Carnegie Mellon University.

References:

- For more information about the Cisco UCS C240 M5 Server, visit <https://www.cisco.com/c/dam/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/c240m5-sff-specsheet.pdf>
- <https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/datasheet-c78-739279.html>
- For more information about the Awiros Video Intelligence Operating System please visit www.awiros.com