Cisco UCS C845A M8 Rack Server

© 2026 Cisco and/or its affiliates. All rights reserved.

Contents

Product overview 3
Features and benefits 7
Product specifications 8
Ordering information 10
Warranty information 10
Product sustainability 10
Product environmental information 11
Cisco and partner services 11
Cisco Capital 11
For more information 11
Document history 12

© 2026 Cisco and/or its affiliates. All rights reserved.

The Cisco UCS® C845A M8 Rack Server/RTX PRO Server is a highly scalable, flexible, and customizable AI system based on the NVIDIA MGX reference design for accelerated computing. With support for two (2) to eight (8) NVIDIA or AMD PCIe GPUs including NVIDIA RTX PRO 4500, RTX PRO 6000D, and 6000 Blackwell Server Edition GPUs and NVIDIA AI Enterprise software, it delivers high performance for a wide range of AI workloads and is optimized for enterprise AI inferencing, agentic AI, and the NVIDIA AI Data Platform in addition to generative AI fine-tuning and RAG.

Product overview

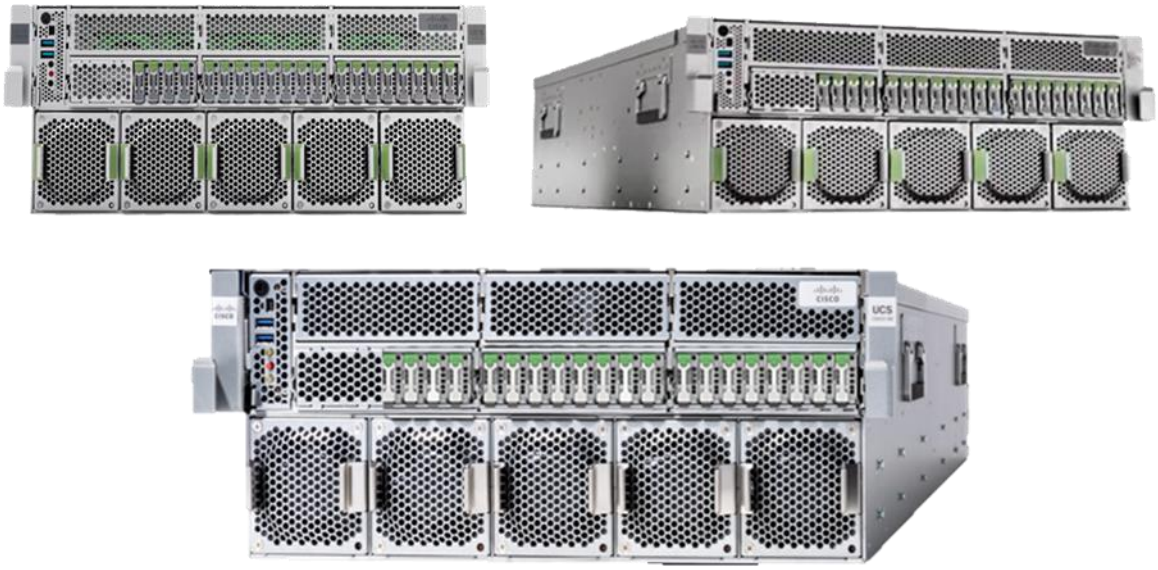


Figure 1.
Cisco UCS C845A M8 Rack Server

The Cisco UCS C845A M8 Rack Server/RTX PRO Server is designed to address the most demanding AI workloads. Now an integral component of Cisco® AI PODs (Cisco Validated Designs for AI) and a foundational compute building block for the Cisco Secure AI Factory with NVIDIA, the Cisco UCS C845A M8 provides a robust foundation for modern AI infrastructure, allowing organizations to easily scale their AI capabilities with confidence—moving from AI pilot to production with secure, simple, and scalable infrastructure that accelerates time to first token and lowers cost per token from core to edge.

With support for 2, 4, 6, or 8 NVIDIA GPUs—including the NVIDIA RTX PRO 4500 (32GB), RTX PRO 6000D (84 GB), and RTX PRO 6000 Blackwell (96 GB), H100 NVL, H200 NVL, L40S, and also supports AMD Instinct MI210 and MI350P GPUs, along with AMD Pensando Pollara 400G —this system offers unparalleled flexibility to meet the diverse needs of enterprises. Combined with up to 8 TB of DDR5-6400 system memory and Gen5 E1.S NVMe storage (up to 15.3 TB per drive), the platform keeps large models, vector indexes, and KV caches close to the GPU, sustaining higher tokens-per-second under real production inferencing workloads. Leveraging the sophistication of the MGX modular reference design, this platform is also future-ready, with next-generation NVIDIA GPUs expected to seamlessly integrate as they become available.

© 2026 Cisco and/or its affiliates. All rights reserved.

As part of Cisco AI PODs, the C845A M8 is designed to accelerate workloads such as large-scale deep learning, Large Language Model (LLM) training, model optimization, and AI inference. It supports hybrid AI workflows across cloud and edge environments, making it the perfect solution for enterprises working on cutting-edge applications such as AI agents, conversational AI, intelligent search, and virtual desktops. Enhanced by 5th Gen AMD EPYC "Turin" CPUs and NVIDIA BlueField-3 DPUs, this server eliminates performance bottlenecks, ensuring optimal CPU and GPU utilization for data-intensive workloads.

With seamless integration into Cisco Intersight®, customers can manage their entire infrastructure—including compute, storage, and networking—through a unified interface. The Cisco UCS C845A M8, now part of Cisco AI PODs, is the ideal platform for organizations seeking to scale their AI initiatives with confidence and efficiency.

As a building block of the Cisco Secure AI Factory with NVIDIA, the C845A M8 inherits distributed security fused into every layer, protecting models, agents, workloads, and infrastructure with Cisco AI Defense, Hybrid Mesh Firewall, Isovalent®, and Hypershield™. NVIDIA BlueField-3 DPUs with crypto acceleration offload security policy enforcement from CPU and GPU, preserving cycles for AI processing while keeping the entire stack hardened and observable through Splunk® Enterprise Security and Splunk Observability Cloud.

The C845A M8 is engineered to align with the NVIDIA AI Data Platform reference design, connecting enterprise data, vector stores, and retrieval pipelines to GPU-accelerated inferencing, so customers can build agentic AI and RAG applications on their own data with the performance, governance, and security required for production. Combined with NVIDIA NIM microservices, NeMo, and Blueprints, the platform compresses time to first token and shortens the path from AI pilot to enterprise deployment.

What use cases does the Cisco UCS C845A M8 address?

The Cisco UCS C845A M8, a highly scalable and customizable server integrated into Cisco AI PODs, is engineered to drive a multitude of AI workloads. Its flexible GPU configurations enable it to address the most demanding AI challenges, including large deep learning, Large Language Model (LLM) training, model fine-tuning, large model inferencing, and Retrieval-Augmented Generation (RAG).

The platform's versatility is further enhanced by its support for various GPUs, each optimized for specific market needs:

NVIDIA H100 NVL:

- Large-scale High-Performance Computing (HPC) and AI workloads.
- Advanced AI research and high-performance computing.
- Generative AI inference for large language models.
- High-performance AI training.

NVIDIA H200 NVL:

- High-performance LLM inference.
- Generative AI training and fine-tuning.

© 2026 Cisco and/or its affiliates. All rights reserved.

NVIDIA RTX Pro 6000 Blackwell Server Edition:

- **Agentic and physical AI:** powering autonomous systems, smart factories, and robotics for real-time decision-making.
- **Advanced scientific computing and rendering:** accelerating complex simulations, medical imaging, and engineering analysis across hybrid environments.
- **High-fidelity 3D graphics and video:** driving content creation, post-production, and immersive VR/AR experiences.
- **Hybrid AI/ML workflows:** enabling seamless training, inferencing, and AI-assisted graphics across on-premises and cloud.
- **Edge-to-core AI applications:** supporting real-time AI at the edge with centralized management and cloud integration.

NVIDIA RTX PRO 6000D Blackwell Server Edition:

- **Enterprise AI inferencing at scale:** serving high-concurrency LLM, multimodal, and agentic workloads with 84 GB of GPU memory, balancing throughput and cost per token for production deployments.
- **NVIDIA AI Data Platform workloads:** accelerating retrieval-augmented generation, vector search, and enterprise data pipelines on the customer's own data.
- **Cost-optimized GenAI deployments:** a strong price/performance option for fine-tuning and inferencing where 96 GB of HBM is more than the memory that is required.

NVIDIA RTX PRO 4500 Blackwell Server Edition:

- **Department-scale AI and intelligent automation:** powering agentic workflows, industrial analytics, and real-time decision support in compact servers.
- **Mainstream scientific computing and simulation:** accelerating engineering analysis, digital twins, and data-heavy modeling with strong performance per watt.
- **Professional visualization and content pipelines:** enabling high-quality 3D visualization, design review, and video workflows for distributed teams.
- **Hybrid AI/ML workflows:** supporting efficient fine-tuning, inference, and AI-assisted graphics across on-prem and cloud environments.
- **Edge-to-core AI deployments:** delivering low-latency AI at remote sites with low power budget requirements, centralized policy, monitoring, and lifecycle management.

© 2026 Cisco and/or its affiliates. All rights reserved.

AMD Instinct MI350P

- **Enterprise inference on existing infrastructure:** dual-slot drop-in card for standard air-cooled servers, deploying inference on premises inside current power, cooling, and rack constraints. amd
- **RAG and agentic AI pipelines:** sized for small, medium, and large models in inference and RAG workflows, with 144GB of HBM3E per card keeping larger models and longer context resident on a single GPU. amd
- **Multi-tenant and virtualized AI serving:** up to four 36GB GPU partitions per card for isolated inference tenants, with VMware Cloud Foundation, Nutanix, and Red Hat AI platform support announced at launch. amd
- **High-throughput low-precision serving:** native MXFP6 and MXFP4 support plus sparsity acceleration for cost-per-token reduction on quantized production models. amd
- **Distributed and edge inference:** scale-out across standard chassis rather than dedicated accelerator platforms, described by Akamai as suited to distributed inference where intelligence sits closer to users. amd
- **Secondary technical compute:** FP64 and FP32 matrix throughput available for mixed AI and simulation environments, though this is a capability, not a positioning claim.

AMD Instinct MI210:

- **High-Performance Computing (HPC):** accelerating scientific research, simulations, and complex modeling.
- **Energy-efficient AI/ML workloads:** supporting deep learning training and inferencing with a focus on power efficiency.
- **Large-scale data analytics:** speeding up data processing and analytics for enterprise applications.
- **Hybrid AI inference:** providing balanced compute and efficiency for inferencing tasks in diverse environments.
- **Specialized simulation workloads:** enhancing performance for engineering, manufacturing, and bioinformatics simulations.

NVIDIA L40S:

- Generative AI foundation model fine-tuning.
- Deployment of intelligent chatbots and search tools.
- Language processing and conversational AI.
- Graphics, rendering, and NVIDIA Omniverse applications.
- Virtual desktop infrastructure.

© 2026 Cisco and/or its affiliates. All rights reserved.

Features and benefits

Table 1. Summary of features and benefits of the Cisco UCS C845A M8 Rack Server

Feature	Benefit
Scalable and flexible	• Rapid deployment of AI and accelerated computing technologies • Suited for any data-center environment
Built on NVIDIA MGX design	• Modular and future-ready design that can adapt to evolving technology needs • Next-generation GPUs can be integrated without purchasing a new platform
Supports NVIDIA RTX PRO Blackwell Server Edition GPUs	• **Platform flexibility without compromises:** One UCS server platform supports RTX PRO 4500 and RTX PRO 6000 Blackwell Server Edition at full power, so you can standardize on a single bill of materials while scaling from cost-efficient deployments to maximum GPU performance as needs grow. • **Predictable performance and uptime at scale:** Full-power GPU operation with enterprise-grade cooling, power delivery, and validated configurations delivers consistent throughput for AI and visualization workloads, reducing throttling risk and simplifying operations across fleet deployments.
Dense design	• Drive 8x NVIDIA or AMD GPUs, 2x AMD Turin CPUs, and 5x PCIe slots in a 4RU chassis
Cisco Intersight support	• See, control, and automate your compute, storage, and networking infrastructure throughout its lifecycle
E1.S drives for local storage	• Increased storage density, improved thermal management, and hot-swappable drives • High performance, lower power consumption, and compact form factor
Addition to NVIDIA MGX reference design	• Improved power delivery, fewer PCBs, and improved cable routing
Cisco Secure AI Factory with NVIDIA building block	• Distributed security across the AI stack with Cisco AI Defense, Hybrid Mesh Firewall, Isovalent, and Hypershield • End-to-end observability through Splunk Enterprise Security and Splunk Observability Cloud • Faster path from AI pilot to production: secure, simple, scalable from core to edge
Optimized for enterprise inferencing and the NVIDIA AI Data Platform	• Up to 96 GB of GPU memory per RTX PRO 6000 and up to 8 TB of DDR5-6400 system memory keep larger models, longer contexts, and more concurrent users resident, higher throughput, and lower cost per token. • Aligned with NVIDIA AI Data Platform reference design and NVIDIA NIM, NeMo, and Blueprints for agentic AI and RAG on enterprise data • Gen5 E1.S NVMe (up to 15.3 TB per drive) keeps vector indexes and KV caches close to the GPU for sustained tokens-per-second.

© 2026 Cisco and/or its affiliates. All rights reserved.

Product specifications

Table 2. Specifications of the Cisco UCS C845A M8 Rack Server

Specification	Cisco UCS 845A M8 Details
CPU	Dual AMD™ Turin CPUs, up to 400W TDP each. Each CPU supports 3-link XGMII. • Supported CPUs: ◦ AMD 9575F 3.3GHz 400W 64C/256MB Cache DDR5 6000MT/s (CAI-CPU-A9575F) ◦ AMD 9475F 3.65GHz 400W 48C/256MB Cache DDR5 6000MT/s (CAI-CPU-A9475F) ◦ AMD 9375F 3.85GHz 320W 32C/256MB Cache DDR5 6000MT/s (CAI-CPU-A9375F) ◦ AMD 9655 2.6GHz 400W 96C/384MB Cache DDR5 6000MT/s (CAI-CPU-A9655) ◦ AMD 9555 3.2GHz 360W 64C/256MB Cache DDR5 6000MT/s (CAI-CPU-A9555) ◦ AMD 9455 3.15GHz 300W 48C/256MB Cache DDR5 6000MT/s (CAI-CPU-A9455) ◦ AMD 9355 3.55GHz 280W 32C/256MB Cache DDR5 6000MT/s (CAI-CPU-A9355)
System Memory	Up to 32 DDR5 DIMMs: • 32GB DDR5-6400 RDIMM 1Rx4 (16Gb) (CAI-MRx32G1RE5) • 64GB DDR5-6400 RDIMM 2Rx4 (16Gb) (CAI-MRx64G2RE5) • 96GB DDR5-6400 RDIMM 2Rx4 (24Gb) (CAI-MRx96G2RF5) • 128GB DDR5-6400 RDIMM 2Rx4 (32Gb) (CAI-MR128G2RG5) • 256GB DDR5-6400 RDIMM 4Rx4 (32Gb) (CAI-MR256G4RG5)
System Control	Dedicated RJ-45 Ethernet port which provides physical access to the server's baseboard management controller (BMC)
GPU	**Two, four, six, or eight of the following PCIe GPUs:** • NVIDIA RTX Pro 4500 GPU, 165W, 32GB, FHFL, 1-slot (CAI-GPU-RTXP4500) • NVIDIA™ RTX Pro 6000 GPU, 600W, 96GB, FHFL, 2-slot (CAI-GPU-RTXP6000) • NVIDIA RTX Pro 6000D GPU, 600W, 84GB, FHFL, 2-slot (CAI-GPU-RTXP6000D) • NVIDIA L40S: 350W, 48GB, 2-slot FHFL GPU (CAI-GPU-L40S) • NVIDIA OEM H200-NVL GPU 600W, 141GB, 2-slot FHFL (CAI-GPU-H200-NVL). This GPU supports NVIDIA's 2-Way and 4-Way NVLink GPU interconnect. • NVIDIA NVL-2 Way Bridge for H200 GPU (CAI-NVL2-H200) • NVIDIA NVL-4 Way Bridge for H200 GPU (CAI-NVL4-H200) • AMD™ Instinct MI210 GPU, 300W, 64GB, FHFL, 2-slot (CAI-GPU-MI210). This GPU supports AMD's 2-Way and 4-Way GPU interconnect. • AMD Infinity 2-Way Bridge for the AMD MI210 GPU (CAI-INF2-MI210) • AMD Infinity 4-Way Bridge for the AMD MI210 GPU (CAI-INF4-MI210) • AMD™ Instinct MI350P* GPU, 600W, 144GB, FHFL, 2-slot (CAI-GPU-MI350P) **Note** If your system will use this GPU, or another single-width GPU, make sure to obtain a perforated filler panel by ordering (UCSC-PCIE-FH). The number of perforated filler panels you will need equals (16 - x) where x is the number of single-width GPUs.

© 2026 Cisco and/or its affiliates. All rights reserved.

Specification	Cisco UCS 845A M8 Details
LAN	One OCP 3.0 SFF PCIe Gen5 x8 NIC (CPU0) with two 10GbE RJ-45 Ethernet
Management	Supported through the DC-SCM card, which features an AST2600 BMC Security is offered through a version 3.0 trusted platform module (TPM). One OCP 3.0 10GBaseT module offers server host management.
Storage	Up to 20 E1.S NVMe PCIe Gen5 SSDs. • 1.9TB E1.S 15mm Kioxia™ XD7P Hg Perf Med End Gen4 1X NVMe (CAI-NVES1T9K1V) • 3.8TB E1.S 15mm Kioxia XD7P Hg Perf Med End Gen4 1X NVMe (CAI-NVES3T8K1V) • 7.6TB E1.S 15mm Kioxia XD7P Hg Perf Med End Gen4 1X NVMe (CAI-NVES7T6K1V) • 1.9TB E1.S 15mm Kioxia XD8 Hg Perf Med End Gen5 1X NVMe (CAI-NVES1T9K2V) • 3.8TB E1.S 15mm Kioxia XD8 Hg Perf Med End Gen5 1X NVMe (CAI-NVES3T8K2V) • 7.6TB E1.S 15mm Kioxia XD8 Hg Perf Med End Gen5 1X NVMe (CAI-NVES7T6K2V) • 1.9TB E1.S 15mm Sandisk SN861 Hg Perf Med End Gen5 1X NVMe (CAI-NVES1T9D1V) • 3.8TB E1.S 15mm Sandisk SN861 Hg Perf Med End Gen5 1X NVMe (CAI-NVES3T8D1V) • 7.6TB E1.S 15mm Sandisk SN861 Hg Perf Med End Gen5 1X NVMe (CAI-NVES7T6D1V) • 15.3TB E1.S 15mm Sandisk SN861 Hg Perf Med End Gen5 1X NVMe (CAI-NVES15T3D1V) Boot Drives: A Cisco Boot-Optimized RAID controller with two M.2 SATA boot drives is also supported in either few the following configurations. • 240GB M.2 SATA Micron G2 SSD (CAI-M2-240G) • 960GB M.2 SATA Micron G2 SSD (CAI-M2-960G)
Expansion Slot	Five FHHL PCIe Gen5 x16 slots One OCP 3.0 slot
Networking	Five PCIe x16 full height, half length (FHHL) slots for single-slot NICS or data processing units (DPUs) **North-South Networking** • NVIDIA MCX755106AS-HEAT 2x200GbE QSFP112 Gen5x16, PCIe VPI NIC (CAI-P-N7D200GFO) • NVIDIA OEM CX713104AS-ADAT 4x25GbE SFP56 Gen4x16, PCIe NIC (CAI-P-N7Q25GFO) • NVIDIA BF-3 B3220 DPU 2x200G QSFP112, Crypto Disabled (CAI-P-N3220) • NVIDIA BF-3 B3220 DPU 2x200G Crypto enabled (CAI-P-NC3220) • Intel X710-DA2 (2x10GbE) (RJ45 OCP 3.0) (CAI-O-ID10GC) • AMD Pensando Pollara* 1x400G (CAI-P-APS400GF) **East-West Networking** • NVIDIA OEM BlueField-3 B3140H SuperNIC 1x400G Crypto Disabled (CAI-P-N3140H) • NVIDIA OEM MCX715105AS-WEAT 1x400GbE QSFP112 PCIe Gen5 NIC, Crypto Disabled (CAI-P-N7S400GFO)

© 2026 Cisco and/or its affiliates. All rights reserved.

Specification	Cisco UCS 845A M8 Details
	• NVIDIA BF-3 B3140H SuperNIC 1x400G Crypto enabled (CAI-P-NC3140H) • NVIDIA OEM CX713104AS-ADAT 4x25GbE SFP56 Gen4x16, PCIe NIC (CAI-P-N7Q25GFO) • AMD Pensando Pollara* 1x400G (CAI-P-APS400GF)
System Cooling	Five individual 80 mm fans
Power	Up to four 3.2KW MCRPS hot-swappable PSUs with N+1 redundancy
System BIOS	AMI™ BIOS
System Software	AMD ROCm™ Software Ubuntu / Red Hat Enterprise Linux (operating system)

Ordering information

Part #	Product description
UCS-MGPUM8-MLB	Cisco UCS-845A M8 Rack Server chassis Major Line Bundle (MLB) This MLB consists of the server node (UCSC-C845A-M8) with software and Cisco Intersight options. Use this PID to begin a new configuration.

For more information, check the Ordering Guide: Cisco UCS C845A M8 Rack Server Ordering Guide

Warranty information

Cisco UCS C845A M8 Rack Servers have a three-year Next-Business-Day (NBD) hardware warranty and a 90-day software warranty.

Product sustainability

Information about Cisco's Environmental, Social, and Governance (ESG) initiatives and performance is provided in Cisco's CSR and sustainability reporting.

Table 3. Cisco environmental sustainability information

Sustainability topic		Reference
General	Information on product-material-content laws and regulations	Materials
	Information on electronic waste laws and regulations, including our products, batteries, and packaging	WEEE Compliance
	Information on product takeback and reuse program	Cisco Takeback and Reuse Program
	Sustainability inquiries	Contact: csr_inquiries@cisco.com
Material	Product packaging weight and materials	Contact: environment@cisco.com

© 2026 Cisco and/or its affiliates. All rights reserved.

Product environmental information

Product environmental information for users per Commission Regulation (EU) 2019/424

https://www.cisco.com/web/dofc/25836410.pdf

Cisco and partner services

Cisco and our industry-leading partners deliver services that accelerate your transition to Cisco UCS solutions for AI and High-Performance Computing (HPC). Cisco Unified Computing Services™ can help you create an agile infrastructure, accelerate time to value, reduce costs and risks, and maintain availability during deployment and migration. After deployment, our services can help you improve performance, availability, and resiliency as your business needs evolve, and help you further mitigate risk.

For more information, visit https://www.cisco.com/go/unifiedcomputingservices.

Cisco Capital

Flexible payment solutions to help you achieve your objectives

Cisco Capital® makes it easier to get the right technology to achieve your objectives, enable business transformation and help you stay competitive. We can help you reduce the total cost of ownership, conserve capital, and accelerate growth. In more than 100 countries, our flexible payment solutions can help you acquire hardware, software, services and complementary third-party equipment in easy, predictable payments. Learn more.

Our experts recommend

- BIOS Performance and workload: Tuning guide for Cisco UCS M8 Platforms White Paper
- Cisco AI PODs At-a-Glance
- Cisco UCS Servers with AMD EPYC

For more information

MLperf Inference report UCS C845A M8 with NVIDIA H200 NVL and L40S GPUs
https://www.cisco.com/c/en/us/products/collateral/servers-unified-computing/ucs-c-series-rack-servers/mlperf-ucs-c845a-m8-rack-server-wp.html.

© 2026 Cisco and/or its affiliates. All rights reserved.

Document history

New or Revised Topic	Described In	Date
Update with AMD MI350P GPUs and Pendando Pollara NICs	Cisco UCS C845A M8 Data Sheet	August 2026
New Product overview, specifications and Ordering information	Cisco UCS C845A M8 Data Sheet	July 2025
Added RTX PRO 6000D GPU; DDR5-6400 memory up to 8 TB (256GB DIMMs); Gen5 E1.S NVMe up to 15.3 TB; crypto-enabled BlueField-3 DPU options; Windows Server 2022/2025 support; Cisco Secure AI Factory with NVIDIA and NVIDIA AI Data Platform positioning	Cisco UCS C845A M8 Data Sheet	April 2026

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **https://www.cisco.com/go/offices**.

Cisco and the Cisco logo are trademarks or registered trademarks of Cisco and/or its affiliates in the U.S. and other countries. To view a list of Cisco trademarks, go to this URL: https://www.cisco.com/go/trademarks. Third-party trademarks mentioned are the property of their respective owners. The use of the word partner does not imply a partnership relationship between Cisco and any other company. (1110R)

Printed in USA C78-4905401-06 08/26

© 2026 Cisco and/or its affiliates. All rights reserved.