



Cisco UCS Integrated Infrastructure For Big Data And Analytics with Cloudera and Apache Spark

Bringing the Power of Real-Time Analytics to Big Data

Highlights



Comprehensive Integrated Infrastructure for Big Data and In-Memory Analytics

- The Cisco UCS Integrated Infrastructure for Big Data and Analytics offers high performance, capacity, and scalability for Apache Spark with Cloudera Enterprise.
- It offers proven, high-performance linear scalability and easy scaling of the architecture with single- and multiple-rack deployments.



Easy Deployment

- Cisco UCS Manager simplifies infrastructure deployment with an automated, policy-based mechanism that helps reduce configuration errors and system downtime.



Simplified Management

- Deploy Cisco UCS Director Express for Big Data quickly and easily for big data infrastructure with one-click provisioning, installation, and configuration. Used in combination with Cloudera Manager, a holistic interface that provides end-to-end system management and detailed and precise visibility and control over every part of an enterprise data hub, the solution makes cluster management simple and straightforward.



Flexible Big Data Platform

- Cisco UCS Integrated Infrastructure for Big Data And Analytics on Cloudera Enterprise allows you to deploy Apache Spark in standalone cluster mode or together with Hadoop and leading NoSQL deployments.



Real-Time Data Processing with Spark Streaming

- By adding Spark Streaming and Apache Kafka on Hadoop deployments to Cisco UCS Integrated Infrastructure for Big Data and Analytics, you enable stream analytics which ingest data in small batches and perform transformations on them.



Batch and Real Time Stream Data Processing with Apache Spark

For long, most of big data technologies have been dealing only with batch processing which usually meant hours or days to unlock the value of big data. While companies are realizing the potential of batch processing of big data, companies are now collecting more data than ever and want to extract value from that data in real time. Sensors, Internet of Things (IoT) devices, social networking, and online transactions are all generating data that needs to be captured, monitored, and rapidly processed to make data-based decisions instantly to provide sentiment and exploratory analytics, trigger any alerts, etc.

Apache Spark is a fast, general-purpose engine for large-scale data processing. With Spark, more enterprises are adopting Hadoop and gaining the capability to process a much wider set of workloads, including streaming and machine learning.

Spark handles most of its operations in memory, dramatically accelerating application performance. This capability enables a whole new set of interactive applications that allow analysts and data scientists to perform experiments more quickly and boost their productivity. Spark can also be used on top of existing Hadoop deployments to quickly implement new features. It is highly compatible with the Hadoop ecosystem and can use data in Hadoop Distributed File System (HDFS) and run under Hadoop 2.0 YARN. The Spark core is complemented by a set of powerful, higher-level libraries that can be used transparently in the same application. These libraries currently include SparkSQL, Spark Streaming, MLlib (for machine learning). Additionally, Spark supports a variety of popular development languages, including R, Java, and Scala. And the basic abstraction in Spark is the Resilient Distributed Dataset which is an immutable, fault-tolerant, distributed collection of objects that can be operated on in parallel.

Here are some real-world examples of the ways that Apache Spark is currently being used:

- Simple, fast extract, transform, and load (ETL) processing: Data can be processed into the required format without the need for intermediate write-to-disk operations, and it can be cleaned and aggregated in memory before the final disk write operation.
- Real-time actions: Anomalous behavior is detected in real time, and downstream actions are performed accordingly. For example, credit-card transactions occurring in a different location than expected can generate actions such as fraud alerts, or transmission of device failure data by IoT sensors can trigger recovery processes.
- Data enrichment: By joining live data with a cached static data set, data is enriched with more information, supporting use of a more comprehensive feature set in real time.
- Exploratory analytics: Events related to a specific time window can be grouped together and analyzed and used by data scientists to update machine learning models within Spark.

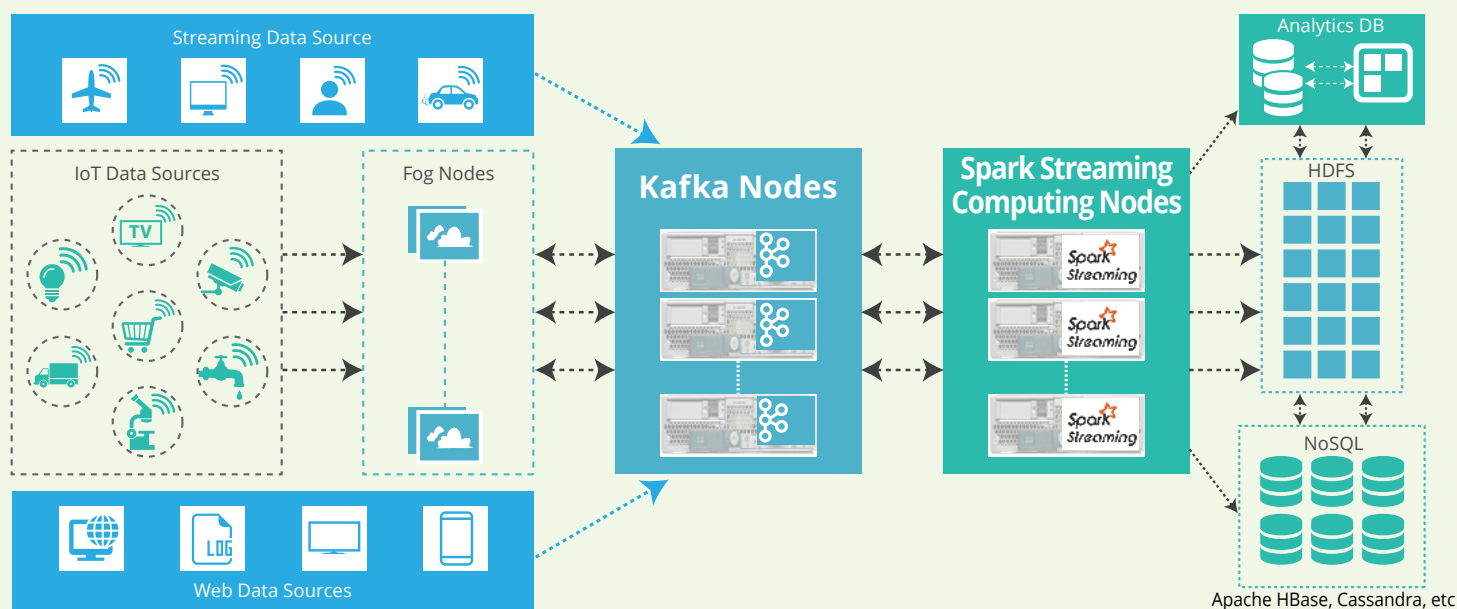
Figure 1 shows the most typical use cases for Spark and Spark Streaming. It shows the flow of data from various data sources to Fog and Kafka nodes, and then to Spark, and then farther downstream to HDFS, NoSQL, SQL databases, Elastic, Solr, and other systems for additional processing.

Cisco UCS Integrated Infrastructure for Big Data and Analytics

Organizations today must ensure that the underlying physical infrastructure can be deployed, scaled, and managed in a way that is agile enough to change as workloads and business requirements change. The Cisco Unified Computing System™ (Cisco UCS®) has redefined the potential of the data center with its revolutionary approach to integrated infrastructure to meet the business needs of IT innovation and acceleration. Cisco UCS Integrated Infrastructure for Big Data and Analytics provides an end-to-end architecture for processing high volumes of real-time or archived data, both structured and unstructured. At the same time, it transparently integrates relevant complex capabilities to deliver an enterprise-class offering with high performance and scalability as applications demand.

Two common reference architectures for Spark on Cisco UCS are available on Cisco UCS C-Series Rack Servers: one adds Spark processing on Hadoop infrastructure, and the other enables stream processing with Kafka or similar technologies. The basic building blocks for these configurations are described in the following sections and shown in Figure 2.

Figure 1: Lambda Architecture with Spark, Spark Streaming, Kafka



Cisco UCS 6200 and 6300 Series Fabric Interconnects

Cisco UCS 6300 and 6200 Series Fabric Interconnects provide high-bandwidth, low-latency connectivity for servers, with integrated, unified management provided for all connected devices by Cisco UCS Manager. The Cisco UCS 6300 Series Fabric Interconnects are a core part of Cisco UCS, providing low-latency, lossless, 10 and 40 Gigabit Ethernet, Fibre Channel over Ethernet (FCoE), and fibre channel functions with management capabilities for systems deployed in redundant pairs. Cisco® fabric interconnects offer the full active-active redundancy, performance, and exceptional scalability needed to support the large number of nodes that are typical in clusters serving big data applications. Cisco UCS Manager enables rapid and consistent server configuration using service profiles and automates ongoing system maintenance activities such as firmware updates across the entire cluster as a single operation. Cisco UCS Manager also offers advanced monitoring with options to raise alarms and send notifications about the health of the entire cluster.

Cisco UCS C-Series Rack Servers

Cisco UCS C240 M4 Rack Servers support a wide range of computing, I/O, and storage-capacity demands in a high-density, compact design. The server uses dual Intel® Xeon® processor

E5-2600 v4 series CPUs and supports up to 768 GB of main memory and a range of hard-disk drive (HDD) and solid-state disk (SSD) drive options. The performance-optimized option supports 24 small-form-factor (SFF) disk drives, and the capacity-optimized option supports 12 large-form-factor (LFF) disk drives. This server can be used with the Cisco UCS Virtual Interface Card (VIC) 1227 or 1387, depending on the fabric interconnect that is being used. The VIC 1227 is designed to optimize high-bandwidth, and low-latency cluster connectivity. The VIC 1387 offers dual-port Enhanced Quad Small Form-Factor Pluggable (QSFP+) 40 Gigabit Ethernet and FCoE in a modular-LAN-on-motherboard (mLOM) form factor.

Cisco UCS Integrated Infrastructure for Big Data and Analytics for Lambda Architecture

In use cases for Spark Streaming with Apache Kafka, Kafka can be deployed on four to eight additional nodes, depending on the data requirements for streaming. The Kafka nodes can be managed using Cloudera Manager, but they need not be part of the Hadoop cluster: that is, they don't need to be running any other Hadoop services (Figure 3).

Figure 2: Cisco UCS Integrated Infrastructure for Big Data and Analytics

2 x Cisco UCS
6296UP Fabric
Interconnects

64 x Cisco UCS C240
M4 Servers

16 x 10 Gigabit Links

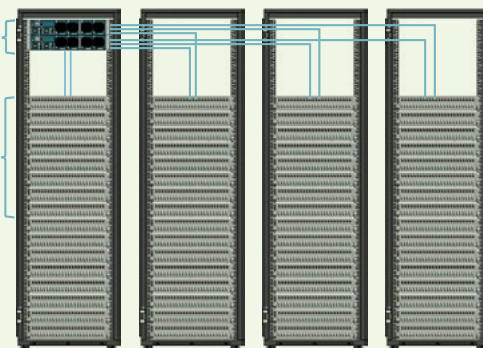
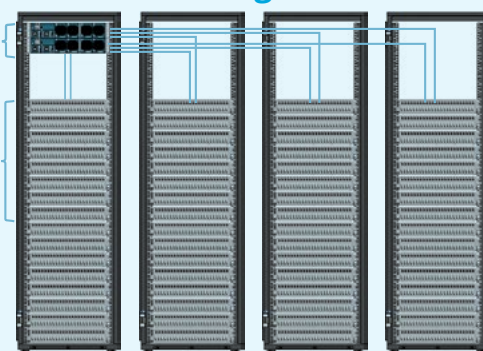


Figure 3: Cisco UCS Integrated Infrastructure for Big Data and Analytics with Apache Kafka

2 Cisco UCS 6296UP
Fabric Interconnects

Kafka Nodes
8 Cisco UCS C240 M4

16 x 10 Gigabit Links



Hadoop Nodes
Cisco UCS C240 M4

Hadoop Clusters
(NameNodes, Resource Manager, Data Nodes, Spark Executors)

To scale the streaming architecture while using Kafka, follow the scaling and sizing guidelines in Table 1. The table provides guidelines for Kafka storage for various servers/drives and replication factors and network bandwidth parameters.

Note: Time taken for filling one server = $\sim((\text{Total storage} / \text{Network bandwidth}) / 3600)$.

Cloudera Enterprise with Apache Spark

Cloudera is the leading provider of enterprise-ready, big data software and services. Cloudera Enterprise includes the market-

leading Hadoop distribution (Cloudera Distribution Including Apache Hadoop [CDH]), a sophisticated administration and management tool (Cloudera Manager), a native end-to-end data governance solution, comprehensive security tool sets, and technical support (Figure 4). Together, Cisco and Cloudera provide organizations with an enterprise-ready data management platform, as well as management integration with an enterprise application ecosystem. They transparently combine to provide a uniquely capable, industry-leading architectural platform for Hadoop-based applications.

Table 1: Scaling and Sizing Guidelines for storage on Kafka nodes






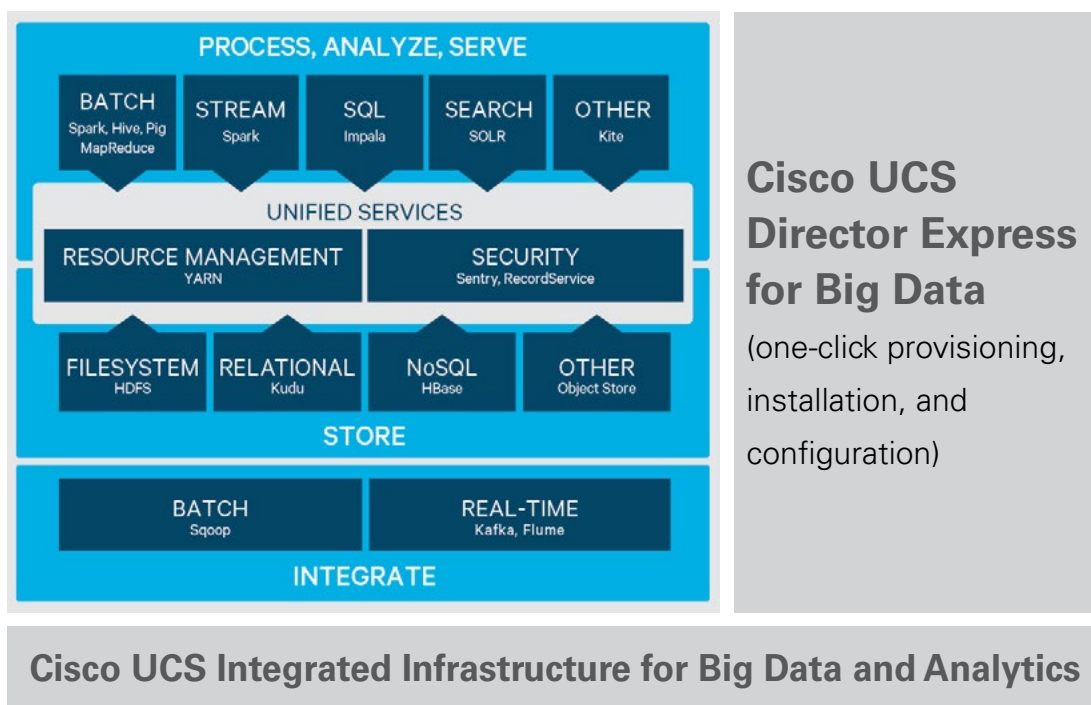
|  Network Bandwidth |  Server Type |  Total Usable Storage |  Time Taken to Fill One Server |  Total Servers (3-Way Replicated Data and Full Network Utilization) |
|--|--|---|---|--|
| 10 Gbps (1.25 GBps) | Cisco UCS C240 M4 (SFF) with 1.8 TB drives | ~40,800 GB | ~9 Hours | ~9 servers for storing 1 day's data |
| 40 Gbps (5 GBps) | Cisco UCS C240 M4 (SFF) with 1.8 TB drives | ~40,800 GB | ~2.3 Hours | ~30 servers for storing 1 day's data |
| 10 Gbps (1.25 GBps) | Cisco UCS C240 M4 (LFF) with 6 TB drives | ~72,000 GB | ~16 Hours | ~6 servers for storing 1 day's data |
| 40 Gbps (5 GBps) | Cisco UCS C240 M4 (LFF) with 6 TB drives | ~72,000 GB | ~4 Hours | ~18 servers for storing 1 day's data |

Figure 4: Cloudera Enterprise



Industry-leading Cloudera products and solutions enable businesses to deploy and manage Apache Hadoop and related projects, manipulate and analyze data, and keep that data secure and protected.

Cloudera provides the following products and tools:

- **Cloudera Enterprise:** Cloudera Enterprise includes the Cloudera distribution of Apache Hadoop and other related open-source projects, including Spark. Cloudera Enterprise also provides security and integration with numerous hardware and software solutions.
- **Apache Spark:** An integrated part of Cloudera Enterprise, Spark is an open standard for flexible in-memory data processing for batch, real-time, and advanced analytics. Cloudera is committed to adopting Spark as the default data processing engine for analytics workloads.
- **Cloudera Manager:** This sophisticated application is used to deploy, manage, monitor, and diagnose problems






with Cloudera deployments. Cloudera Manager provides the Admin Console, a web-based user interface that makes administration of any enterprise data simple and straightforward.

- **Cloudera Navigator:** This end-to-end data management tool for the Cloudera Enterprise platform enables administrators, data managers, and analysts to explore the large amounts of data in Hadoop. The robust auditing, data management, lineage management, and lifecycle management in Cloudera Navigator allow enterprises to adhere to stringent compliance and regulatory requirements.

Reference Architecture

Cisco UCS Integrated Infrastructure for Big Data and Analytics offers several configurations to meet a variety of computing and storage requirements, shown in Table 2.

Table 2: Cisco UCS Integrated Infrastructure for Big Data and Analytics Options

|  Performance Optimized Option 1 (UCS-SL-CPA4-P1) |  Performance Optimized Option 2 (UCS-SL-CPA4-P2) |  Performance Optimized Option 3 (UCS-SL-CPA4-P3) |  Capacity Optimized Option 1 (UCS-SL-CPA4-C1) |  Capacity Optimized Option 2 (UCS-SL-CPA4-C2) |
|--|--|---|--|---|
| Connectivity: <ul style="list-style-type: none"> • 2 Cisco UCS 6296UP 96-Port Fabric Interconnects | Connectivity: <ul style="list-style-type: none"> • 2 Cisco UCS 6296UP 96-Port Fabric Interconnects | Connectivity: <ul style="list-style-type: none"> • 2 Cisco UCS 6332 Fabric Interconnects | Connectivity: <ul style="list-style-type: none"> • 2 Cisco UCS 6296UP 96-Port Fabric Interconnects | Connectivity: <ul style="list-style-type: none"> • 2 Cisco UCS 6296UP 96-Port Fabric Interconnects |
| Scaling <ul style="list-style-type: none"> • Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches | Scaling <ul style="list-style-type: none"> • Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches | Scaling <ul style="list-style-type: none"> • Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches | Scaling <ul style="list-style-type: none"> • Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches | Scaling <ul style="list-style-type: none"> • Scales up to 1000s of servers using Cisco Nexus 9000 or 7000 series switches |
| 16 Cisco UCS C240 M4 Rack Servers (SFF), each with: <ul style="list-style-type: none"> • 2 Intel Xeon processor E5-2680 v4 CPUs (14 cores on each CPU) • 256 GB of memory • Cisco 12-Gbps SAS Modular RAID Controller with 2-GB flash-based write cache (FBWC) • 24 x 1.2-TB 10000-rpm SFF SAS drives (460 TB total) • 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot • Cisco UCS VIC 1227 (with 2x10 Gigabit Ethernet SFP + ports) | 16 Cisco UCS C240 M4 Rack Servers (SFF), each with: <ul style="list-style-type: none"> • 2 Intel Xeon processor E5-2680 v4 CPUs (14 cores on each CPU) • 256 GB of memory • Cisco 12-Gbps SAS Modular RAID Controller with 2-GB flash-based write cache (FBWC) • 24 x 1.8-TB 10000-rpm SFF SAS drives (691 TB total) • 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot • Cisco UCS VIC 1227 (with 2x10 Gigabit Ethernet SFP + ports) | 16 Cisco UCS C240 M4 Rack Servers (SFF), each with: <ul style="list-style-type: none"> • 2 Intel Xeon processor E5-2680 v4 CPUs (14 cores on each CPU) • 256 GB of memory • Cisco 12-Gbps SAS Modular RAID Controller with 2-GB flash-based write cache (FBWC) • 24 x 1.8-TB 10000-rpm SFF SAS drives (691 TB total) • 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot • Cisco UCS VIC 1387 (with 2 x 40 Gigabit Ethernet SFP+ ports) | 16 Cisco UCS C240 M4 Rack Servers (LFF), each with: <ul style="list-style-type: none"> • 2 Intel Xeon processor E5-2620 v4 CPUs (8 cores on each CPU) • 128 GB of memory • Cisco 12-Gbps SAS Modular RAID Controller with 2-GB flash-based write cache (FBWC) • 12 x 6-TB 10000-rpm LFF SAS drives (1152 TB total) • 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot • Cisco UCS VIC 1227 (with 2x10 Gigabit Ethernet SFP + ports) | 16 Cisco UCS C240 M4 Rack Servers (LFF), each with: <ul style="list-style-type: none"> • 2 Intel Xeon processor E5-2620 v4 CPUs (8 cores on each CPU) • 256 GB of memory • Cisco 12-Gbps SAS Modular RAID Controller with 2-GB flash-based write cache (FBWC) • 12 x 8-TB 7200-rpm LFF SAS drives (1536 TB total) • 2 x 240-GB 6-Gbps 2.5-inch enterprise value SATA SSDs for boot • Cisco UCS VIC 1227 (with 2x10 Gigabit Ethernet SFP + ports) |

The Cisco UCS reference architectures for Cloudera Enterprise support the massive scalability that Hadoop enterprise deployments demand. The configuration described in this document can be extended to support up to 80 servers with a pair of 96-port Cisco UCS fabric interconnects. Multiple Cisco UCS domains, with up to thousands of servers, can be supported using Cisco Nexus® 9000 or 7000 Series Switches.

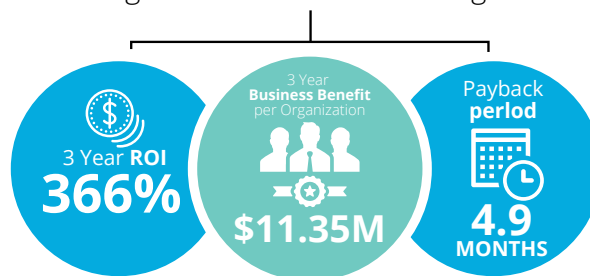
Conclusion

Apache Spark on Cisco UCS Integrated Infrastructure for Big Data and Analytics provides a comprehensive platform that can deliver industry-leading performance. This solution enables organizations to deploy Spark easily with the simplified deployment model of Cisco UCS and expand the solution on demand to support powerful big data analytics processing with Spark.

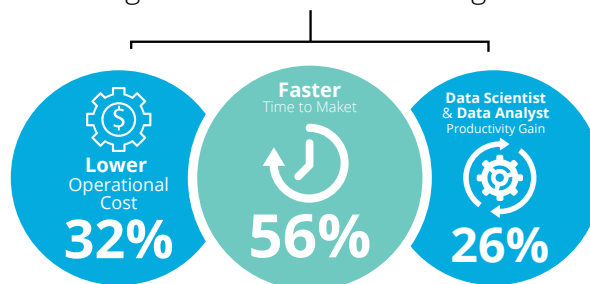
For More Information

- For more information about Cisco UCS big data solutions, see www.cisco.com/go/bigdata.
- For more information about Cisco UCS Integrated Infrastructure for Big Data, see <http://blogs.cisco.com/datacenter/cpav4/>.
- For more information about Spark on Cloudera, see <https://www.cloudera.com/products/apache-hadoop/apache-spark.html>.

Business Value Summary for Cisco UCS Integrated Infrastructure for Big Data



Business Value Summary for Cisco UCS Integrated Infrastructure for Big Data



Business Value Benefits -Average Annual Benefits per Cisco UCS Server

| Business Productivity | IT Staff Productivity | IT Infrastructure Cost Reduction |
|-----------------------|-----------------------|----------------------------------|
| \$29,654 | \$3,861 | \$123 |



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 www.cisco.com/go/offices.



Cisco and the Cisco Logo are trademarks of Cisco Systems, Inc. and/or its affiliates in the U.S. and other countries. A listing of Cisco's trademarks can be found at 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. (1005R)