



WHITE PAPER

UNIFIED FABRIC: BENEFITS AND ARCHITECTURE OF VIRTUAL I/O

In today's data center, applications are driving increased demand for server processing and I/O, all at a time when budgets are shrinking. To grow to meet demand, the data center manager is forced to deploy servers that come in fixed packages and prevent resources from being shared. This limitation duplicates, isolates, and wastes resources, increasing the network's overall complexity and cost. It also prevents data center architects from taking full advantage of the trend toward commoditization in the server industry. In response, datacenter architects have begun making changes to consolidate and share resources across multiple customers and applications. This enables utility computing, or the ability to dynamically commission and decommission these shared resources.

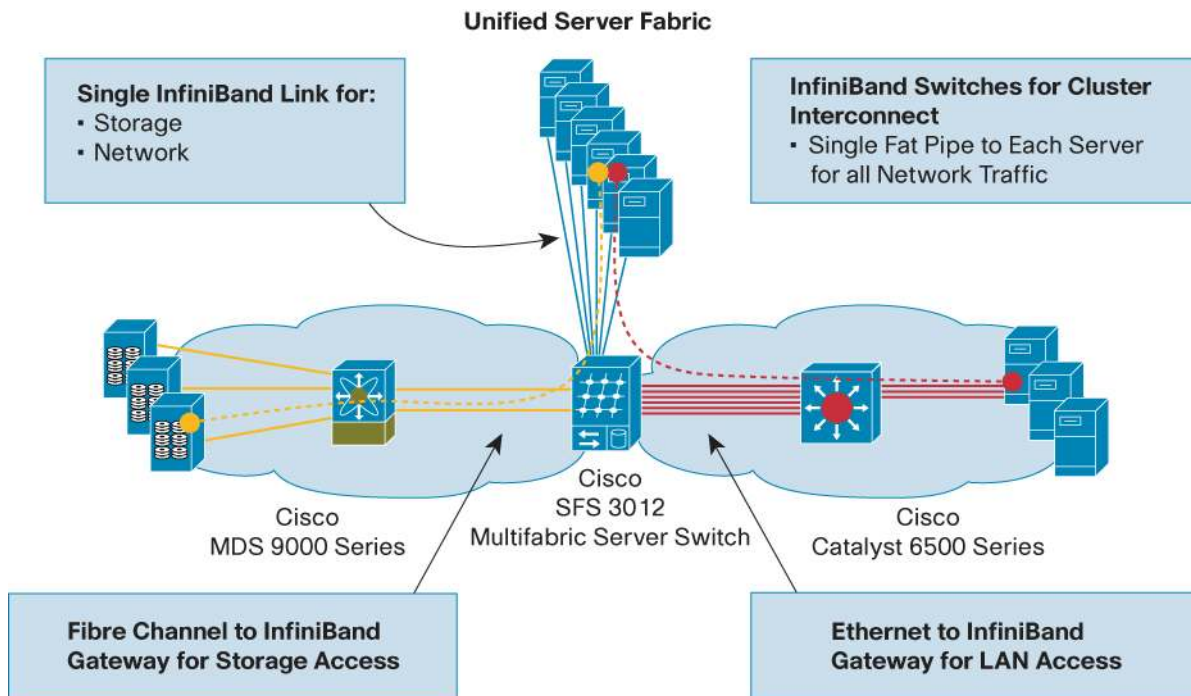
Server I/O architecture is one major architectural component that can be shared in this manner. Typically, a conventional server today is deployed with multiple network adapters that serve three basic I/O requirements: LAN/WAN, storage area network (SAN), and interprocess communications. Servers can be deployed with multiple Ethernet network interface cards (NICs), Fibre Channel host bus adapters (HBAs), and sometimes dedicated clustering interconnects. For mission-critical applications such as databases, servers can require many expansion slots. In a large data center, managing these servers and their cables can be difficult and costly, hampering the ability to change quickly to meet business demands.

INTRODUCING THE SERVER SWITCH

The Cisco® Server Fabric Switch enables utility computing by dramatically simplifying the data center architecture. It creates a unified, "wire-once" fabric that aggregates I/O and server resources. With the unified fabric, instead of servers having many cables coming out of them, the server switch connects every server with a single high-bandwidth, low-latency network cable (two cables for redundancy). This setup aggregates Ethernet, Fibre Channel, and clustering interconnects into a 10-Gbps InfiniBand cable. The server switch then connects servers to a pool of shared Fibre Channel and Ethernet ports over line-rate gateways and creates virtual I/O subsystems on each host, including virtual HBAs and virtual IP interfaces. Servers can then share a centralized pool of Ethernet and Fibre Channel ports that can be upgraded and serviced without affecting running applications. Similar to how a SAN creates a pool of shared storage that can be managed independently of the servers themselves, virtual I/O creates an independently managed pool of Fibre Channel and Ethernet I/O.

Aggregating the server's I/O resources saves significant capital expense. Consolidating resources over the unified fabric eliminates costs of underutilized Fibre Channel HBAs and NICs as well as associated cabling complexity. Instead of being designed to accommodate bandwidth peaks using a dedicated switch port for each host, a data center can share remote Fibre Channel and Gigabit Ethernet ports, enabling network designs based on average load across multiple servers. This can save up to 50 percent of the cost of the I/O associated with a server. Also, by eliminating multiple adapters and local storage by introducing a single high-bandwidth, low-latency connection, the size of the server is driven only by CPU and memory requirements. This often results in a reduction in the size and cost of the server as well as in its space, power, and cooling needs, resulting in immediate return-on-investment savings of up to 50 percent.

Figure 1. Unified Server Fabric



Virtualizing I/O on the server also makes it possible to aggregate multiple servers by changing server identities rapidly based on time of day. By simply changing the server-to-storage mappings stored in the server switch, physical machines can switch rapidly between different operating systems and applications. Everything unique about a server is stored in the fabric, and the physical server is simply another resource to be assigned. This creates a new level of flexibility, because servers are no longer tied to physical locations. Using virtual I/O in combination with Cisco VFrame Server Fabric Virtualization Software, administrators can create business policies that repurpose servers based on time of day, CPU or application load, or other metrics.

Virtual I/O has two major components: (1) virtual IP interfaces and an InfiniBand-to-Ethernet gateway and (2) virtual HBA and an InfiniBand-to-Fibre Channel gateway. The administrator installs an InfiniBand driver package on the host that includes an IP-to-InfiniBand (IPoIB) driver, SCSI driver (called SCSI RDMA Protocol or SRP), and other RDMA protocols. The server uses the IP and SCSI drivers to communicate through the gateways, bridging IP subnets and allowing hosts to access Fibre Channel-attached storage.

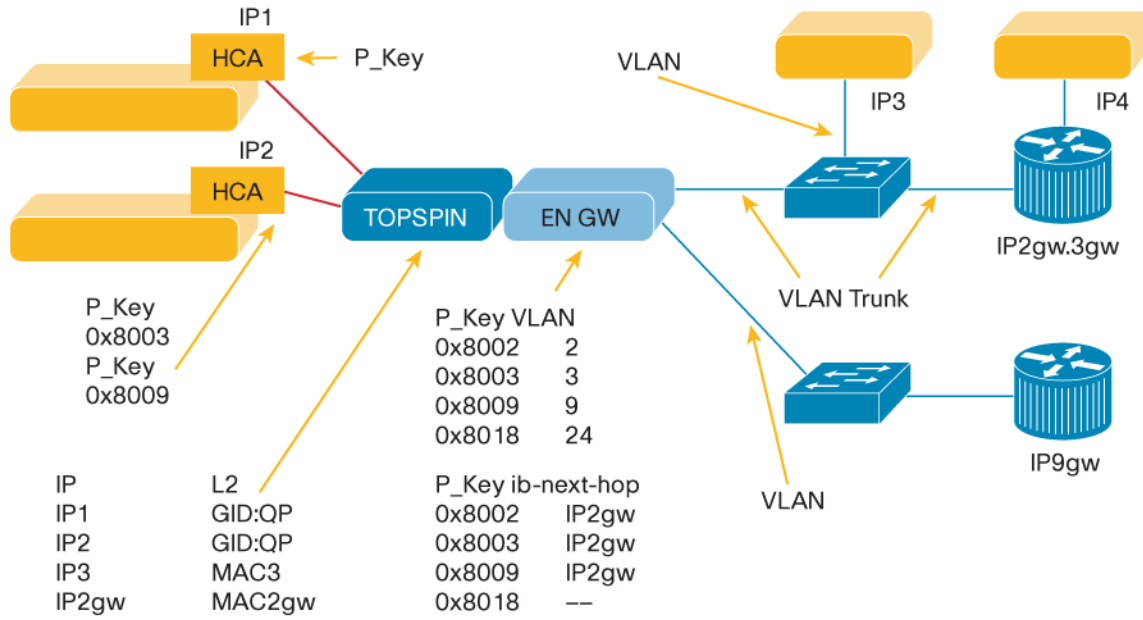
Virtual IP Interfaces

To allow servers to communicate directly with existing IP servers, administrators create virtual IP interfaces on InfiniBand-attached servers by loading an IPoIB driver. Although the server may not be physically connected directly to the LAN/WAN, this interface can transparently communicate to other Ethernet-attached hosts using the InfiniBand-to-Ethernet gateway.

IPoIB is used within the InfiniBand fabric for standard IP-based communications as well as address lookups. IPoIB is also translated across the Ethernet gateway to IP over Ethernet (IPoE) to provide Layer 2 bridging based on IP address (not MAC address). IP addresses are lookup keys in the forwarding tables, and IP addresses are used to make forwarding decisions. Administrators create bridge groups for bridging between InfiniBand and Ethernet subnets, which translates the IPoIB frames to IPoE frames. A VLAN is mapped 1:1 to an InfiniBand P-Key, which provides similar VLAN security. VLAN 802.1Q trunking is also supported. The Ethernet gateway presents its MAC address to the next-hop switch or router. The individual MAC address of the host is not exposed. Thus, tools that use MAC addresses must use the alternative client ID. For example, to support Dynamic

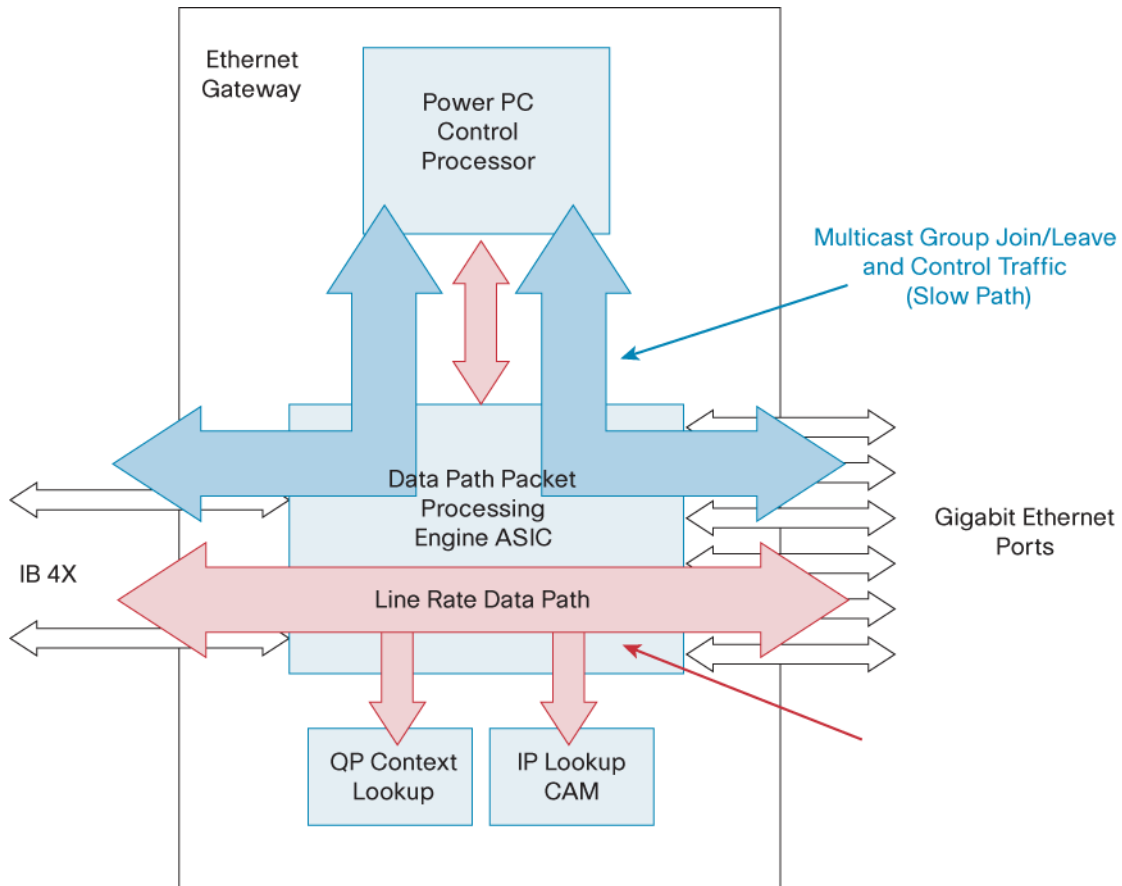
Host Configuration Protocol (DHCP) over the gateway, the DHCP server must use “client ID” as the identifier instead of the MAC address (see Figure 2).

Figure 2. Bridging IP from InfiniBand to Ethernet



The InfiniBand-to-Ethernet gateway is based on specialized chipsets that perform this bridging at cut-through line rate on all six Gigabit Ethernet ports (Figure 3). The gateway has two distinct paths for data and control, a “slow path” PowerPC processor, and a “fast path” hardware packet processing engine. Based on this fast path, a single gateway is capable of sustaining 11 million 64-bit packets per second. Similarly, multicast packet processing is handled by the fast path by looking up IP-to-InfiniBand multicast group mappings and changing packet headers to InfiniBand multicast addresses, which are also handled in hardware. Multicast forwarding performs at 8 million packets per second per gateway, incurring negligible latency (about 3 microseconds per 64-byte packet). Multicast joins and leaves are handled in the slow path. In comparison, software-based gateways perform at lower levels of 200,000 packets per second with latency on the order of milliseconds.

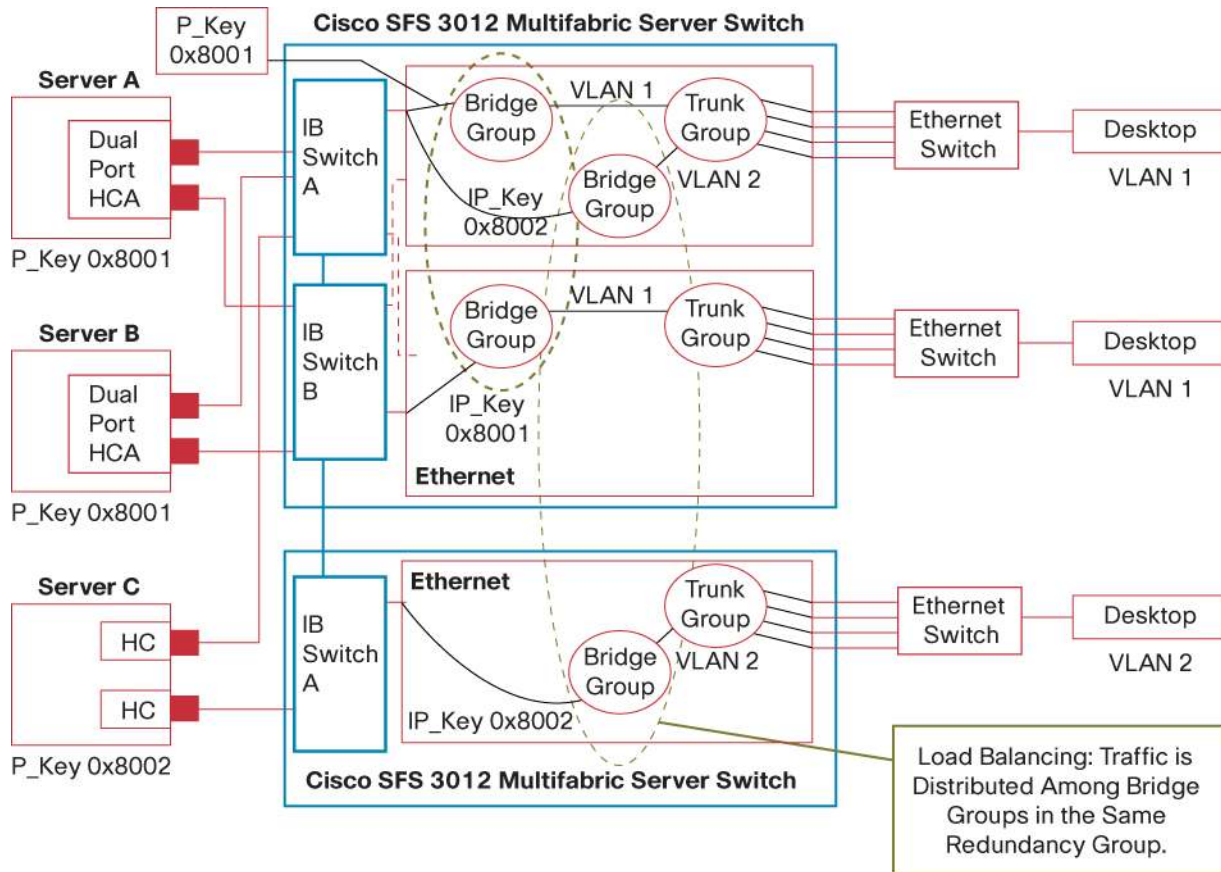
Figure 3. Cisco InfiniBand-to-Ethernet Gateway Architecture



If link aggregation of the Ethernet ports is desired to merge the Ethernet ports into one logical link, bridge groups may also be associated with trunk groups. Link groups can be configured to have one to six links or ports. All traffic is distributed between the ports of the trunk group based on six different distribution algorithms, including source/destination IP, source/destination MAC, and Round Robin. If one link fails, the bandwidth of the trunk group is reduced but traffic is unaffected, and when a link recovers, its bandwidth is restored.

Load balancing between discrete gateways and switches is done by assigning multiple bridge groups into redundancy groups. In redundancy groups configured for failover, traffic is not passed on backup bridge groups. In redundancy groups configured for load balancing, all bridge groups are passing traffic. Otherwise, the two modes of operation are similar. Redundancy groups can be created using multiple gateways across multiple chassis. Load distribution is conversation based and is based on source/destination hardware addresses and source/destination IP addresses. Figure 4 describes how load balancing and VLAN-to-partition mappings work across Ethernet gateways.

Figure 4. InfiniBand-to-Ethernet Gateway Load Balancing



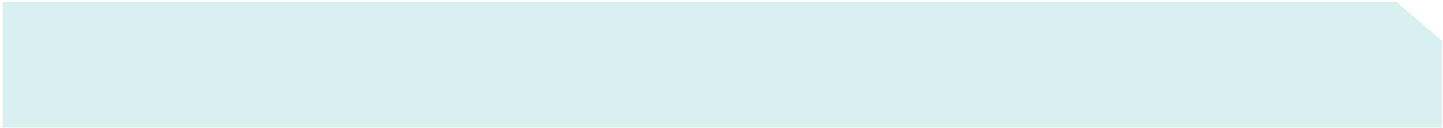
To configure the IPoIB driver on the host, the administrator configures an interface on the ib0 and ib1 ports on the HCA similar to how eth0 and eth1 ports would be configured on an Ethernet NIC. In Linux, the administrator configures it with `ifconfig`, and in Windows, the administrator can configure the interfaces with the Device/driver dialog in the Control Panel. The InfiniBand partition is also configured at this location. Using IPoIB, existing IP-based applications and tools work using existing sockets libraries. Networking diagnostics tools such as ping and traceroute work, and server switches can integrate with existing network management tools such as CiscoWorks, Tivoli, Unicenter, and Openview using Simple Network Management Protocol (SNMP).

Virtual HBA

To transparently connect to Fibre Channel SANs, administrators create virtual HBAs on each server by loading a storage driver on InfiniBand-attached servers. Although the server may not have a Fibre Channel HBA, it appears directly attached to the Fibre Channel network using the InfiniBand-to-Fibre Channel gateway module.

Administrators run SRP for storage communications within the fabric. SRP is also translated to Fibre Channel Protocol (FCP) across the Fibre Channel gateway to allow SRP initiators (hosts) to talk to FCP targets (storage devices).

The combination of the SRP driver and the Fibre Channel gateway allow InfiniBand-attached hosts to appear direct-attached on the InfiniBand fabric as NL-port attached devices using the 2-port Fibre Channel gateway, or devices attached to a virtual arbitrated loop. Using the 4-port Fibre Channel gateway, the gateway ports appear as E-Ports.



Every server can potentially share every Fibre Channel port attached to the InfiniBand fabric. The combination of the SRP driver and Fibre Channel gateway create a talk-through model, creating a unique World-Wide Node Name (WWNN) for each host, and unique World-Wide Port Names for every Fibre Channel port the initiator is discovered through.

To reduce much of this complexity, administrators can use Cisco VFrame Server Fabric Virtualization Software to manage server access to pools of Fibre Channel ports. This adds the option of creating Fibre Channel port pools. When a virtual server group is assigned to a port pool, all of the mappings are managed transparently to the user.

Fibre Channel sessions are then load-balanced across these ports, allowing a server or group of servers to communicate to multiple storage targets across multiple ports, effectively increasing available aggregate bandwidth and breaking the 1:1 binding that typically exists between a server and its dedicated SAN port. To add more bandwidth, an additional hot-plug expansion module can be added to the multifabric server switch, and Fibre Channel sessions can be dynamically load-balanced to include the new gateway.

SUMMARY

By simplifying the server I/O architecture to a single high-speed pipe, the unified fabric enables the data center administrator to simplify the server itself, reducing costs associated with underutilized ports and management costs associated with those ports. This also simplifies the server itself, allowing administrators to size hardware based on the CPU and memory, not the number of expansion slots. In turn, this further enables the trend toward server commoditization and in combination with Cisco's VFrame technology, the foundation for true utility computing in the data center.

**Corporate Headquarters**

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-4000
800 553-NETS (6387)
Fax: 408 526-4100

European Headquarters

Cisco Systems International BV
Haarlerbergpark
Haarlerbergweg 13-19
1101 CH Amsterdam
The Netherlands
www-europe.cisco.com
Tel: 31 0 20 357 1000
Fax: 31 0 20 357 1100

Americas Headquarters

Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134-1706
USA
www.cisco.com
Tel: 408 526-7660
Fax: 408 527-0883

Asia Pacific Headquarters

Cisco Systems, Inc.
168 Robinson Road
#28-01 Capital Tower
Singapore 068912
www.cisco.com
Tel: +65 6317 7777
Fax: +65 6317 7799

Cisco Systems has more than 200 offices in the following countries and regions. Addresses, phone numbers, and fax numbers are listed on **the Cisco Website at www.cisco.com/go/offices.**

Argentina • Australia • Austria • Belgium • Brazil • Bulgaria • Canada • Chile • China PRC • Colombia • Costa Rica • Croatia • Cyprus
Czech Republic • Denmark • Dubai, UAE • Finland • France • Germany • Greece • Hong Kong SAR • Hungary • India • Indonesia • Ireland • Israel
Italy • Japan • Korea • Luxembourg • Malaysia • Mexico • The Netherlands • New Zealand • Norway • Peru • Philippines • Poland • Portugal
Puerto Rico • Romania • Russia • Saudi Arabia • Scotland • Singapore • Slovakia • Slovenia • South Africa • Spain • Sweden • Switzerland • Taiwan
Thailand • Turkey • Ukraine • United Kingdom • United States • Venezuela • Vietnam • Zimbabwe

Copyright © 2005 Cisco Systems, Inc. All rights reserved. CCSP, CCVP, the Cisco Square Bridge logo, Follow Me Browsing, and StackWise are trademarks of Cisco Systems, Inc.; Changing the Way We Work, Live, Play, and Learn, and iQuick Study are service marks of Cisco Systems, Inc.; and Access Registrar, Aironet, ASIST, BPX, Catalyst, CCDA, CCDP, CCIE, CCIP, CCNA, CCNP, Cisco, the Cisco Certified Internetwork Expert logo, Cisco IOS, Cisco Press, Cisco Systems, Cisco Systems Capital, the Cisco Systems logo, Cisco Unity, Empowering the Internet Generation, Enterprise/Solver, EtherChannel, EtherFast, EtherSwitch, Fast Step, FormShare, GigaDrive, GigaStack, HomeLink, Internet Quotient, IOS, IP/TV, iQ Expertise, the iQ logo, iQ Net Readiness Scorecard, LightStream, Linksys, MeetingPlace, MGX, the Networkers logo, Networking Academy, Network Registrar, Packet, PIX, Post-Routing, Pre-Routing, ProConnect, RateMUX, ScriptShare, SlideCast, SMARTnet, StrataView Plus, TeleRouter, The Fastest Way to Increase Your Internet Quotient, and TransPath are registered trademarks of Cisco Systems, Inc. and/or its affiliates in the United States and certain other countries.

All other trademarks mentioned in this document or Website 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. (0502R) 205410.BA_ETMG_JL_9.05

