White Paper

Reconsidering Physical Topologies with 10GBASE-T

10GBASE-T and twisted-pair cabling can dramatically lower the CAPEX of interconnect in the data center. Interconnect includes both the connection between the Top of Rack switch (ToR) and the server and the connection between the ToR and the spine switch. 10GBASE-T lowers costs by replacing expensive SFP+ DAC cabling with affordable UTP cabling and by enabling architects to optimize physical topologies.

May 2013

10GBASE-T Complements Terabit-Class Switches to Permit Lower Interconnect Costs

When leaf ToR switches are physically located on top of the rack, optical 40GbE switch-to-switch links are typically required. Optical 40GbE links have two expensive QSFP optical modules with their interconnecting fiber. These links can represent the highest cost component of the interconnect.

When leaf ToR switches are moved off the top of the rack and placed next to the higher-layer spine switches in a networking aisle, the optical links can be replaced with affordable Quad Small Form-Factor Pluggable (QSFP) Direct Attach Copper (DAC) cabling. 10GBASE-T UTP cabling can then be used to span the longer distance between the server and the leaf switch (formerly the leaf ToR switch).

Current Best Practice

The current paradigm for the construction of data centers is to create a pod of computing capacity to facilitate East-West (E-W) traffic, maximize the efficiency of the pod, and then replicate the pod as capacity is increased. The most common topology includes the use of leaf ToR switches with a full mesh connection to their corresponding spine switches, providing E-W interconnect within the pod. This E-W interconnect in the spine is typically created using 40GBASE-SR optical links. The leaf ToR switches reside on the top of each rack (42RU is a common height), connecting to the 40 servers within the rack with SFP+ DAC cables. The North-South (N-S) traffic into and out of the pod is sourced from the leaf ToR switches, utilizing approximately 10% of the leaf ToR bandwidth in this example. This example also assumes each server has a single network connection.

The technologies available today work well with this logical topology. Current 640 Gbps switching silicon supports the leaf ToR switch with 40 10GbE downlinks, 4x 40GbE uplinks into the spine, and 2x 40GbE uplinks for N-S traffic. Using a single-level spine switch constructed from the same 640 Gbps switching silicon, a pod of 640 servers (16 racks x 40 servers per rack) managing primarily E-W traffic can be constructed with the following configuration:

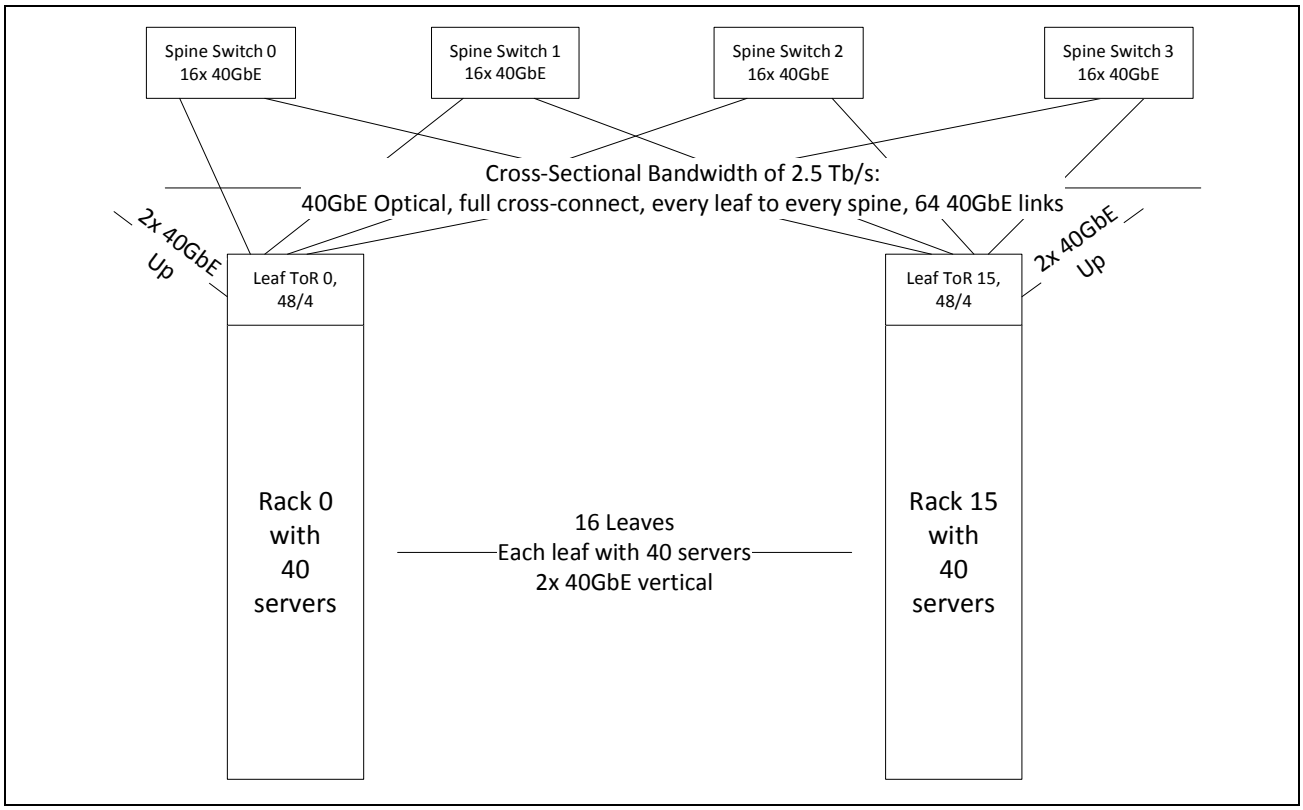
- Four 1RU spine switches, each supporting 16x 40GbE using a 640G switch chip
- Four 40GbE uplinks for each leaf ToR switch (one to each spine switch)
- 40 10GbE downlinks for each leaf ToR switch connected to 40 servers within the rack
- Four spine switches located in any available space (no interconnect is required between the spine switches.)

The interconnect within this pod includes the following:

  – 640 servers with 640 DAC copper cables between the servers and the leaf ToR switches
  – 16 leaf ToR switches, each with 2x 40GbE for N-S traffic and 32 40GBASE-SR optical modules
  – 64 40GbE links between the leaf ToR switches and the spine switches, for a total of 128 40GBASE-SR optical modules

In this configuration, the physical and logical topologies are equivalent. The leaf ToR switch resides on the top of each rack, and the interconnect is supported with DAC connections to each switch. See Figure 1 on page 3.

Figure 1: 2.5 Tbps Cross-Sectional Bandwidth Pod with DAC Connections



Scaling to Terabit Switch Technology

The 640 Gbps switches currently used for leaf ToR switches are being replaced with 1.28 Tbps (1280 Gbps) switches. Meanwhile, the bandwidth requirements of servers are not projected to grow at a high rate. 10GbE is only now becoming mainstream, and is projected to remain the primary server interconnect for one or two generations.

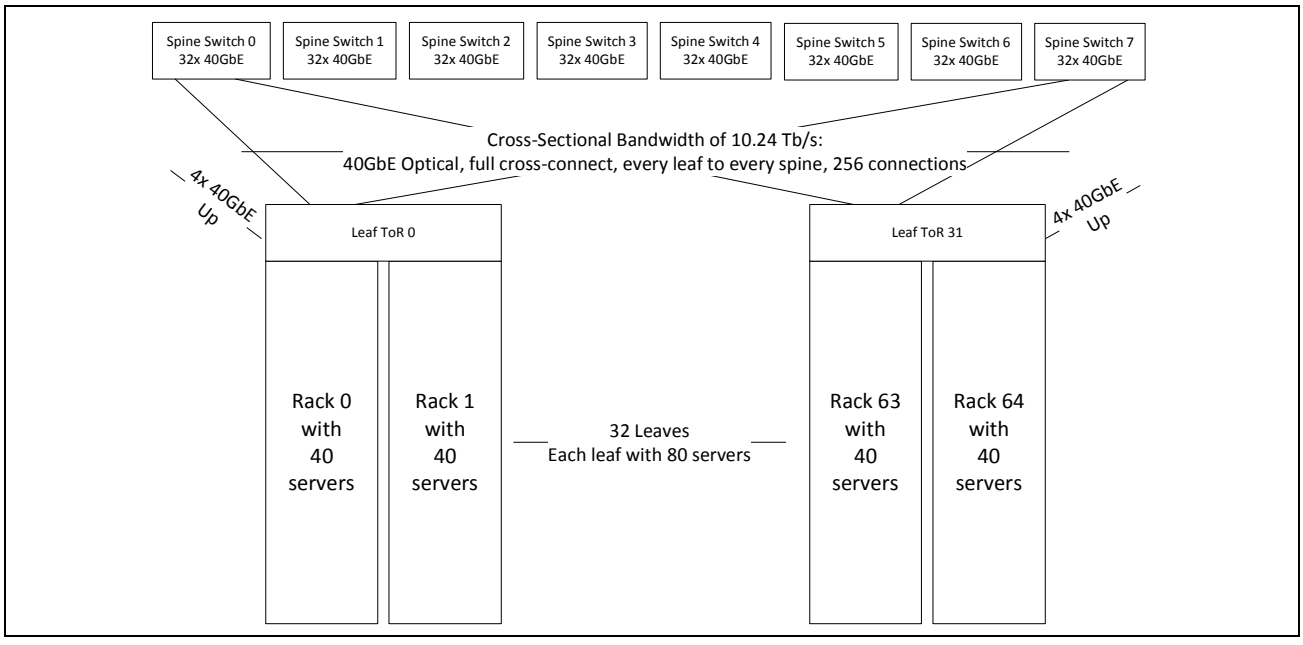
The prototypical switching topology shown in Figure 1 on page 3 can easily be scaled with the new 1.28 Tbps switches in the following configuration:

- Each spine switch becomes a 32x 40GbE switch.
- Each leaf ToR switch has eight 40GbE uplinks, one routing to each spine switch.
- Each leaf ToR switch has 80 10GbE downlinks, connected to 80 servers in two racks (40 servers per rack).
- Each leaf ToR switch directs 4x 40GbE for N-S traffic.

The interconnect in this pod configuration includes the following:

  – 2560 servers (64 racks x 40 servers per rack) with 2560 DAC cables between the servers and the leaf ToR switches
  – 256 40GbE links between the leaf ToR switches and the spine switches, for a total of 512 40GBASE-SR optical modules with a cross-sectional bandwidth of 10.24 Tbps

Figure 2: 10 Tbps Cross-Sectional Bandwidth Pod with DAC



The doubling of the switch bandwidth quadruples the size of the pod, with the number of servers growing from 640 to 2560 servers. As the leaf ToR switch begins to support more servers than fit within a single rack, the increased switch bandwidth challenges the terminology: Is the switch still a ToR switch if it sits on top of multiple racks?

Advances in 10GBASE-T as an Interconnect

Direct Attach Copper (DAC), consisting of a pair of twinax cables terminated in SFP+ modules, proved to be a viable interconnect option when 10GbE was initially rolled out to the network. 10GBASE-T was early in its development cycle and was not a competing technology at the time.

Moore's Law benefited 10GBASE-T just as it continues to benefit switch chips. As a result, switches and adapter cards with 10GBASE-T are now available in the same densities and form factors as SFP+ based platforms.

10GBASE-T has inherent advantages over DAC for conventional ToR topologies. One primary advantage is the lower cost of 10GBASE-T UTP cabling. Table 1 reflects costs for UTP patch cords and DAC cables. The values in Table 1 reflect very aggressive costs for DAC cables which do not reflect channel markup. These prices may not be available to most enterprises.

Table 1 looks at the cost impact of swapping out DAC for 10GBASE-T, with all other considerations being equal. The data is based on the maximum size pod of 2560 servers using 1.28 Tbps switches. Although server to switch cabling is not the primary expense in the overall costs of the pod, switching to 10GBASE-T can lower the cost of this cabling by 75%. Since the cost of DAC cables varies based on how the cables are purchased, actual costs for DAC can be two or three times the values shown when cables are procured from OEMs or through distribution, rather than purchased directly from low-cost manufacturers.

Table 1: Comparison of Costs for UTP and DAC

Number of Cables	*Cost for Cat6A Patch Cord (2m)*	*Cost for DAC Cable (2m)*	*Total Cost for UTP*	*Total Cost for DAC*
2560	$5	$20	$12.8K	$51.2K

A second advantage to 10GBASE-T is its interoperability. Many OEMs configure SFP+ cages to require authorized modules or cables, complicating the use of DAC cables. 10GBASE-T has no such limitation and, therefore, provides data center designers with maximum selection flexibility.

Reconsidering the Physical Topology

10GBASE-T has traditionally been used in structured cabling topologies. The cabling is laid in cable trays as part of the structure before the equipment is installed. The cables are connected at each end to patch panels, and patch cords then complete the link. Alternately, the server to leaf switch connection can be a longer point-to-point connection using UTP cabling with no patch panels. This configuration may require additional management, but it offers a lower CAPEX than the patch panel configuration.

Using 10GBASE-T, the physical topology of the ToR can be redesigned without changing the logical topology. For example:

- 10GBASE-T UTP cabling can be used as the interconnect between the servers and the leaf switches. The 64 racks of 40 servers each can be placed in four rows, with all server interconnects being terminated at a networking aisle where the switches reside. In this case, the maximum required cable length is less than 30m. See Figure 3 on page 6.

- The 32 leaf switches and the eight spine switches can be located in close proximity to each other. This enables the 40GbE full mesh cross-connect to be constructed with 40GbE DAC, and it eliminates the 40GBASE-SR optics in the spine interconnect.

Figure 3: 10 Tbps Cross-Sectional Bandwidth Pod with UTP

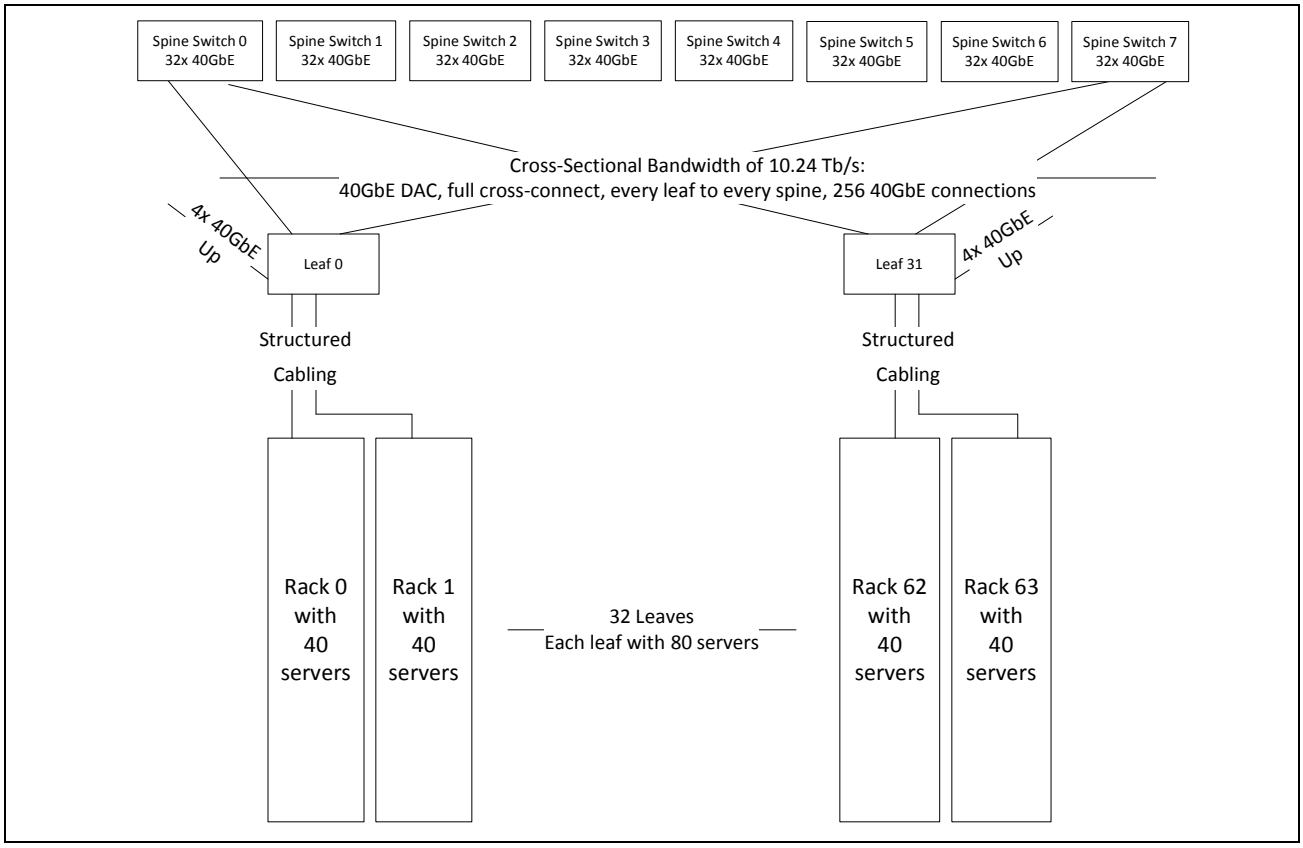
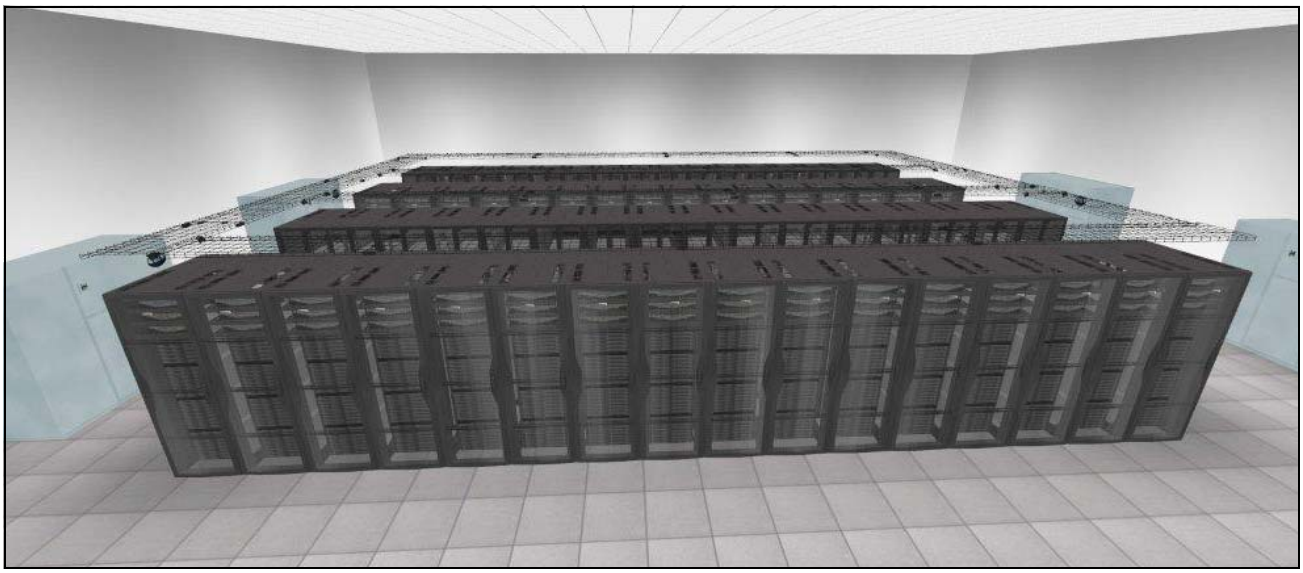


Figure 4: Pod with 2560 Servers, 64 Racks of 40 Servers Each (Source: Panduit)



UTP cabling costs are higher with the longer point-to-point cables, increasing from approximately $5 for the patch cord to $40 for the longer cable. Structured cabling will have a higher premium. However, because the 40GbE optics are eliminated from the mesh cross-connect, the 10GBASE-T topology offers a tremendous overall cost reduction, as shown in Table 2.

Table 2: Total Interconnect Costs

Details	*SFP+ ToR*	*10GBT ToR*	*10GBT Structured*
Leaf Switches	32	32	32
Downlinks per Leaf Switch	80	80	80
Max Number of Downlinks	2560	2560	2560
DAC, Patch, or Structured Link[a]	$20	$5	$30
Cost for Connecting Servers	$51,200	$12,800	$76,800
Uplinks per Leaf	8	8	8
Total Uplinks	256	256	256
Cost per 40GbE SR Optics Module[b]	$500	$500	$0
Cost per 40GbE Cabling[c]	$50	$50	$40
Cost for Spine Interconnect	$268,800	$268,800	$10,240
Total Interconnect Cost	**$320,000**	**$281,600**	**$87,040**

a. $20 for a DAC, $5 for a UTP patch cord, $30 for a UTP connection to server
b. $500 for 40GBASE-SR module
c. $50 for the optical cable between modules, $40 for cost of QSFP DAC

The major cost factor is replacing the 40GbE optics with 40GbE DAC. Table 2 assumes the QSFP 40GBASE-SR modules are $500 each (Crehan Research estimates the weighted average retail ASP is $900; high-volume direct purchases may cost under $200), the corresponding optical cabling (either point-to-point or structured cabling) is $50, and the QSFP DAC cabling is $40. (This cabling can be used when all the switches are collocated.) Topologies consisting of different elements may naturally have different costs.

Power

10GBASE-T requires more power than the equivalent SFP+ 10GbE DAC interconnect. The power delta is approximately 1.5W/port in 28 nm technology, based on an estimate of the net power requirement difference between a 10GBASE-T PHY and an XFI to SFI+ PHY. Changing from a ToR-based configuration with 10GbE DAC to a leaf-based configuration with 10GBASE-T increases the power requirement for the server interconnect by 7.7 KW (2560 x 2 x 1.5W), and decreases power for the 40GbE interconnect (as 40GbE DAC replaces 40GBASE-SR QSFP optics) by 0.77 KW (256 x 2 x 1.5W), for a net difference of 7.0 KW. This increased power must be factored into the ROI as a recurring cost.

Note, however, that the net increase in overall power is for a pod of 2560 servers and 40 switches. If each server and switch consumes an estimated 300W, the pod consumes approximately 780 KW. Therefore, the power impact of the PHY selection represents less than 1% of the total required power. The power required for 10GBASE-T decreases with the use of Energy Efficient Ethernet™ (EEE) technology.

Cabling Considerations

10GBASE-T was originally defined to enable 100m channel lengths, similar to the previous three generations of BASE-T. This required the creation of a robust new UTP cable. 10GBASE-T silicon works reliably at 100m when 23AWG Category 6A (Cat6A) cable is deployed. However, 23AWG Cat6A cable has a relatively large diameter which can be an impediment in data center design. This 23AWG Cat6a cable will prove to have a great deal of margin at shorter distances.

Several cabling suppliers provide Cat6A cabling at a reduced diameter, using a finer gauge of wire such as 26AWG or smaller[1]. Cable diameter can also be reduced using a screened cable construction such as F/UTP or F/STP. Data center designers can select the smallest diameter and lowest cost solution that meets the 100m Cat6A electrical specifications. Since the 10GBASE-T PHY supports a 100m link using 23AWG Cat6A cable, smaller gauge cabling (26 or 28AWG) at a reduced length can be used, whereby the excess margin is traded off for smaller diameter. Regardless of the cable gauge, the overall electrical parameters as measured with field test equipment must remain within specification.

Standard Cat6 cable can also be used with 10GBASE-T. However, Cat6 cable is subject to Alien NEXT (ANEXT) crosstalk between links, particularly in the structured cabling topology which uses longer cables. Cat6A cabling is not significantly impacted by ANEXT crosstalk, and is therefore the recommended cable type for structured cabling. Direct connect cables can leverage the smaller diameter of Cat6 cables, where the shorter distance easily meets the ANEXT requirements.

1. The higher the number, the smaller the cable diameter.

Fast Forward to 40GBASE-T

The IEEE has begun developing a specification for 40GBASE-T. The 10GBASE-T topology recommended in this paper will scale efficiently to meet the requirements of server to leaf connections that are four times the current speed, and spine switches that offer four times the current density. Total interconnect costs will be minimized as described when the leaf to spine interconnect migrates to DAC, and structured cabling is used for leaf to server connections.

Summary

10GBASE-T continues to reduce the implementation power required in the data center. Today, 48-port 10GBASE-T 1RU switches are readily available with four to six QSFP uplinks. Many users have compared 10GBASE-T with DAC, and have already begun to deploy 10GBASE-T with ToR switches. While this provides a significant CAPEX cost savings, the savings may be much greater with the full capability of 10GBASE-T: support for links of 30m or more, colocation of all the switches in a networking aisle, and replacement of the 40GbE optical links with 40GbE DAC. These cost savings can be further amplified when the implications of next generation 1.28 Tbps switching silicon are considered.

Connecting everything®

BROADCOM

BROADCOM CORPORATION
5300 California Avenue
Irvine, CA 92617

Phone: 949-926-5000
Fax: 949-926-5203
E-mail: info@broadcom.com
Web: www.broadcom.com

84848-WP100-R May 2013