Arista Cloud Networks

Arista Networks was founded to deliver software defined cloud networking solutions for large datacenter and high-performance computing environments. The award-winning Arista 7500 Series introduced in April 2010 set new records for maximizing datacenter performance, efficiency and overall network reliability with port count densities that enabled consolidation of what was traditionally three-tier Core/Aggregation/Access designs into two-tier Spine/Leaf designs. The introduction of Arista 7500E Series line cards and fabrics in 2013 provided a 3x increase in capacity/density/performance on the Arista 7500 Series, cementing three-tier designs as designs of the past.

With the introduction of the Arista 7300 Series switches, Arista provides the flexibility to further collapse network layers with the introduction of the single-tier Spline (combined Spine/Leaf) network design.

This whitepaper provides an overview of Arista's Cloud Network designs and how Arista provide unprecedented scale, performance and density without proprietary protocols or lock-ins or forklift upgrades.

Key Points of Arista Designs

All Arista Universal Cloud Network designs revolve around these nine central design goals:

1. *No proprietary protocols or vendor lock-ins*. Arista believes in open standards. Our proven reference designs show that proprietary protocols and vendor lock-ins aren't required to build very large scale-out networks
2. *Fewer Tiers is better than More Tiers.* Designs with fewer tiers (e.g. a 2-tier Spine/Leaf design rather than 3-tier) decrease cost, complexity, cabling and power/heat. Single-tier Spline network designs don't use any ports for interconnecting tiers of switches so provide the lowest cost per usable port. A legacy design that may have required 3 or more tiers to achieve the required port count just a few years ago can be achieved in a 1 or 2-tier design.
3. *No protocol religion.* Arista supports scale-out designs built at layer 2 or layer 3 or hybrid L2/L3 designs with open multi-vendor supported protocols like VXLAN that combine the flexibility of L2 with the scale-out characteristics of L3.
4. *Modern infrastructure should be run active/active.* Multi chassis Link Aggregation (MLAG) at layer 2 and Equal Cost Multi-Pathing (ECMP) at layer 3 enables infrastructure to be built as active/active with no ports blocked so that networks can use all the links available between any two devices.
5. *Designs should be agile and allow for flexibility in port speeds.* The inflection point when the majority of servers/compute nodes connect at 1000Mb to 10G is between 2013-2015. This in turn drives the requirement for network uplinks to migrate from 10G to 40G and to 100G. Arista switches and reference designs enable that flexibility
6. *Scale-out designs enable infrastructure to start small and evolve over time.* A two-way ECMP design can grow from 2-way to 4-way, 8-way, 16-way and as far as a 32-way design. An ECMP design can grow over time without significant up-front capital investment.
7. *Large Buffers can be important.* Modern Operating Systems, Network Interface Cards (NICs) and scale-out storage arrays make use of techniques such as TCP Segmentation Offload (TSO), GSO and LSO. These techniques are fundamental to reducing the CPU cycles required when servers send large amounts of data. A side effect of these techniques is that an application/ OS/ storage that wishes to transmit a chunk of data will offload it to the NIC, which slices the data into segments and puts them on the wire as back-to-back frames at line-rate. If more than one of these is destined to the same output port then microburst congestion occurs.

   One approach to dealing with bursts is to build a network with minimal oversubscription, overprovisioning links such that they can absorb bursts. Another is to reduce the fan-in of traffic. An alternative approach is to deploy switches with deep buffers to absorb the bursts results in packet drops, which in turn results in lower good-put (useful throughput).
8. *Consistent features and OS.* All Arista switches use the same Arista EOS. There is no difference in platform, software trains or OS. It's the same binary image across all switches.
9. *Interoperability.* Arista switches and designs can interoperate with other networking vendors with no proprietary lock-in.

Design Choices - Number of Tiers

An accepted principle of network designs is that a given design should not be based on the short-term requirements but instead the longer-term requirement of how large a network or network pod may grow over time. Network designs should be based on the maximum number of usable ports that are required and the desired oversubscription ratio for traffic between devices attached to those ports over the longer-term.

If the longer-term requirements for number of ports can be fulfilled in a single switch (or pair of switches in a HA design), then there's no reason why a single tier spline design should not be used.

Spline Network Designs

Spline designs collapse what have historically been the spine and leaf tiers into a single spline. Single tier spline designs will always offer the lowest capex and opex (as there are no ports used for interconnecting tiers of switches), the lowest latency, are inherently non-oversubscribed with at most two management touch points. Flexible airflow options (front-to-rear or rear-to-front) on a modular spline switch enable its deployment in server/compute racks in the datacenter, with ports on the same side as the servers with airflow that matches the thermal containment of the servers.

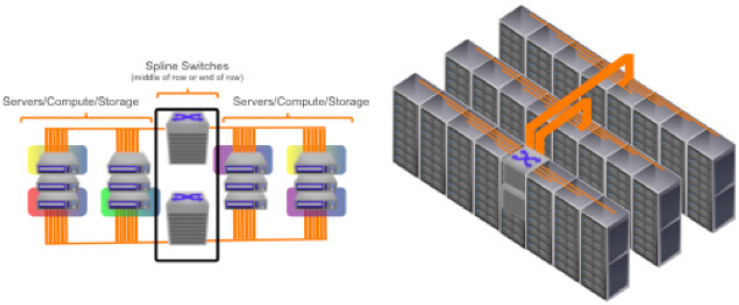

Figure 1: Arista Spline single-tier network designs provide scale up to 2,000 physical servers (49 racks of 1U servers)

Arista 7300 Series (4/8/16 slot modular chassis), Arista 7250X Series (64x40G to 256x10G 2U Fixed switch) and Arista 7050X Series (32x40G to 104x10G+8x40G) switches are ideal for spline network designs providing for 104 to 2048 x 10G ports in a single switch, catering for datacenters as small as 3 racks to as large as 49 racks.

Spine/Leaf Network Designs

For designs that don't fit a single tier spline design then a two-tier spine leaf design is the next logical step. A two-tier design has spine switches at the top tier and leaf switches at the bottom tier with Servers/compute/storage always attached to leaf switches at the top of every rack (or for higher density leaf switches, top of every N racks) and leaf switches uplink to 2 or more spine switches.

Scale out designs start with one pair of spine switches and some quantity of leaf switches. A two-tier leaf/spine network design at 3:1 oversubscription has for 96x10G ports for servers/compute/storage and 8x40G uplinks per leaf switch (Arista 7050SX-128 – 96x10G : 8x40G uplinks = 3:1 oversubscribed).

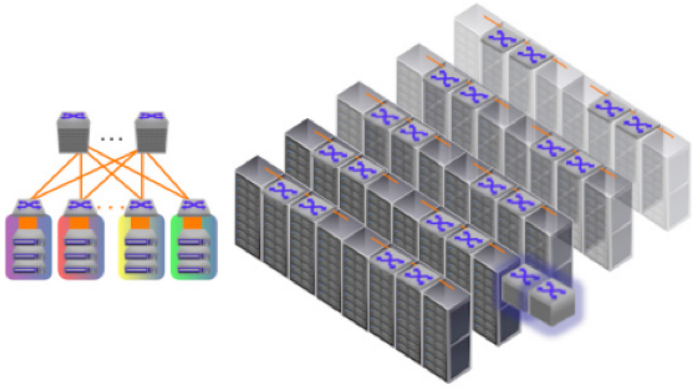

Figure 2: Arista Spine/Leaf two-tier network designs provide scale in excess of 100,000 physical servers

Two-tier Spine/Leaf network designs enable horizontal scale-out with the number of spine switches growing linearly as the number of leaf switches grows over time. The maximum scale achievable is a function of the density of the spine switches, the scale-out that can be achieved (this is a function of cabling and number of physical uplinks from each leaf switch) and desired oversubscription ratio.

Either modular or fixed configuration switches can be used for spine switches in a two-tier spine/leaf design however the spine switch choice locks in the maximum scale that a design can grow to.

Layer 2, Layer 3 or Layer 3 with VXLAN Overlay

Layer 2 or Layer 3

Two-tier Spine/Leaf networks can be built at either layer 2 (VLAN everywhere) or layer 3 (subnets). Each has their advantages and disadvantages.

Layer 2 designs allow the most flexibility allowing VLANs to span everywhere and MAC addresses to migrate anywhere. The downside is that there is a single common fault domain (potentially quite large), and as scale is limited by the MAC address table size of the smallest switch in the network, troubleshooting can be challenging, L3 scale and convergence time will be determined by the size of the Host Route table on the L3 gateway and the largest non-blocking fan-out network is a spine layer two switches wide utilizing Multi-chassis Link Aggregation (MLAG).

Layer 3 designs provide the fastest convergence times and the largest scale with fan-out with Equal Cost Multi Pathing (ECMP) supporting up to 32 or more active/active spine switches. These designs localize the L2/L3 gateway to the first hop switch allowing for the most flexibility in allowing different classes of switches to be utilized to their maximum capability without any dumbing down (lowest-common-denominator) between switches.

Layer 3 designs do restrict VLANs and MAC address mobility to a single switch or pair of switches and so limit the scope of VM mobility to the reach of a single switch or pair of switches, which is typically to within a rack or several racks at most.

Layer 3 Underlay with VXLAN Overlay

VXLAN complements the Layer 3 designs by enabling a layer 2 overlay across layer 3 underlay via the non-proprietary multi-vendor VXLAN standard. It couples the best of layer 3 designs (scale-out, massive network scale, fast convergence and minimized fault domains) with the flexibility of layer 2 (VLAN and MAC address mobility), alleviating the downsides of both layer 2 and layer 3 designs.

VXLAN capabilities can be enabled in software through hypervisor-resident virtual switches as part of a virtual server infrastructure. This approach extends layer 2 over layer 3 but doesn't address how traffic gets to the correct physical server in the most optimal manner. A software-based approach to deploying VXLAN or other overlays in the network also costs CPU cycles on the server, as a result of the offload capabilities on the NIC being disabled.

Hardware VXLAN Gateway capability on a switch enables the most flexibility, greater scale and traffic optimization. The physical network remains at layer 3 for maximum scale-out, best table/capability utilization and fastest convergence times. Servers continue to provide NIC CPU offload capability and the VXLAN Hardware Gateway provides layer 2 and layer 3 forwarding, alongside the layer 2 overlay over layer 3 forwarding.

Arista EOS Foundation Features That Enable These Designs

Arista's scale-out cloud network designs are underpinned on a number of foundation features of Arista's award-winning Extensible Operating System:

Multi Chassis Link Aggregation (MLAG)

MLAG enables devices to be attached to a pair of Arista switches (an MLAG pair) with all links running active/active. MLAG eliminates bottlenecks, provides resiliency and enables layer 2 links to operate active/active without wasting 50% of the bandwidth as is the case with STP blocked links. L3 Anycast Gateway (Virtual ARP / VARP) with MLAG enables the L3 gateway to operate in active/active mode without the overhead of protocols like HSRP or VRRP.

To a neighboring device, MLAG behaves the same as standard link aggregation (LAG) and can run either with Link Aggregation Control Protocol (LACP) (formerly IEEE 802.3ad, more recently IEEE 802.1AX-2008) or in a static 'mode on' configuration.

The MLAG pair of switches synchronize forwarding state between them such that the failure of one node doesn't result in any disruption or outage as there are no protocols to go from standby to active, or new state to learn as the devices are operating in active/active mode.

Zero Touch Provisioning (ZTP)

ZTP enables switches to be physically deployed without any configuration. With ZTP, a switch loads its image and configuration from a centralized location within the network. This simplifies deployment, enabling network engineering resources to be used for more productive tasks by avoiding wasting valuable time on repetitive tasks such as provisioning switches or requiring network engineers to walk around with serial console cables.

An extension to ZTP, Zero Touch Replacement (ZTR) enables switches to be physically replaced, with the replacement switch picking up the same image and configuration as the switch it replaced. Switch identity and configuration aren't tied to switch MAC address but instead are tied to location in the network where the device is attached (based on LLDP information from neighboring devices). While a hardware failure and RMA is not likely to be a common event, ZTR means that in this situation the time-to-restoration is reduced to the time it takes for a new switch to arrive and be physically cabled, and is not dependent on a network engineer being available to provide device configuration, physically in front of the switch with a serial console cable.

VM Tracer

As virtualized datacenters have grown in size, the physical and virtual networks that support them have also grown in size and complexity. Virtual machines connect through virtual switches and then to the physical infrastructure, adding a layer of abstraction and complexity. Server side tools have emerged to help VMware administrators manage virtual machines and networks, however equivalent tools to help the network administrator resolve conflicts between physical and virtual networks have not surfaced.

Arista VM Tracer provides this bridge by automatically discovering which physical servers are virtualized (by talking to VMware vCenter APIs), what VLANs they are meant to be in (based on policies in vCenter) and then automatically apply physical switch port configurations in real time with vMotion events. This results in automated port configuration and VLAN database membership and the dynamic adding/removing VLANs from trunk ports.

VM Tracer also provides the network engineer with detailed visibility into the VM and physical server on a physical switch port while enabling flexibility and automation between server and network teams.

VXLAN

VXLAN is a multi-vendor industry-supported network virtualization technology that enables much larger networks to be built at layer 2 without the inherent scale issues that underpin large layer 2 networks. It uses a VLAN-like encapsulation technique to encapsulate layer 2 Ethernet frames within IP packets at layer 3 and as such is categorized as an 'overlay' network. From a virtual machine perspective, VXLAN enables VMs to be deployed on any server in any location, regardless of the IP subnet or VLAN that the physical server resides in.

VXLAN provides solutions to a number of underlying issues with layer 2 network scale, namely:

- Enables large layer 2 networks without increasing the fault domain
- Scales beyond 4K VLANs
- Enables layer 2 connectivity across multiple physical locations or pods
- Potential ability to localize flooding (unknown destination) and broadcast traffic to a single site
- Enables large layer 2 networks to be built without every device having to see every other MAC address

VXLAN is an industry-standard method of supporting layer 2 overlays across layer 3. As multiple vendors support VXLAN there are subsequently a variety of ways VXLAN can be deployed: as a software feature on hypervisor-resident virtual switches, on firewall and load-balancing appliances and on VXLAN hardware gateways built into L3 switches. Arista's approach to VXLAN is to support hardware-accelerated VXLAN gateway functionality across a range of switches: Arista 7150 Series, Arista 7050X Series, Arista 7200 Series, Arista 7300X Series and Arista 7500E Series. These platforms support unicast and multicast VXLAN gateway capabilities that can be orchestrated through non-proprietary and open standard APIs such as eAPI OVSDB or configured statically. This open approach to hardware VXLAN gateway capabilities provides end users choice between cloud orchestration platforms without any proprietary vendor lock-in.

LANZ

Arista Latency Analyzer (LANZ) enables tracking of network congestion in real time before congestion causes performance issues. Today's systems often detect congestion when someone complains, "The network seems slow." The network team gets a trouble ticket, and upon inspection can see packet loss on critical interfaces. The best solution historically available to the network team has been to mirror the problematic port to a packet capture device and hope the congestion problem repeats itself.

Now, with LANZ's proactive congestion detection and alerting capability both human administrators and integrated applications can:

- Pre-empt network conditions that induce latency or packet loss
- Adapt application behavior based on prevailing conditions
- Isolate potential bottlenecks early, enabling pro-active capacity planning
- Maintain forensic data for post-process correlation and back testing

Arista EAPI

Arista EOS API (eAPI) enables applications and scripts to have complete programmatic control over EOS, with a stable and easy to use syntax. eAPI exposes all state and all configuration commands for all features on Arista switches via a programmatic API.

Once eAPI is enabled, the switch accepts commands using Arista's CLI syntax, and responds with machine-readable output and errors serialized in JSON, served over HTTP or HTTPS. The simplicity of this protocol and the availability of JSON clients across all scripting languages means that eAPI is language agnostic and can be easily integrated into any existing infrastructure and workflows and can be utilized from scripts either on-box or off-box.

Arista ensures that a command's structured output will always remain forward compatible for multiple future versions of EOS allowing end users to confidently develop critical applications without compromising their ability to upgrade to newer EOS releases and access new features.

OpenWorkload

OpenWorkload is a network application enabling open workload portability, automation through integration with leading virtualization and orchestration systems, and simplified troubleshooting by offering complete physical and virtual visibility.

- Seamless Scaling - full support for network virtualization, connecting to major SDN controllers
- Integrated Orchestration - interfaces to VMware NSX™, OpenStack, Microsoft, Chef, Puppet, Ansible and more to simplify provisioning
- Workload Visibility to the VM-level, enabling portable policies, persistent monitoring, and rapid troubleshooting of cloud networks

Designed to integrate with VMware, OpenStack and Microsoft OMI, Arista's open architecture allows for integration with any virtualization and orchestration system.

Smart System Upgrade

Smart System Upgrade (SSU) reduces the burden of network upgrades, minimizing application downtime, and reducing the risks taken during critical change controls. SSU provides a fully customizable suite of features that tightly couples datacenter infrastructure partners, such as Microsoft, F5, and Palo Alto Networks with integration that allows devices to be seamlessly taken out or put into service. This helps customers stay current on the latest software releases without unnecessary downtime or systemic outages.

Network Telemetry

Network Telemetry is a new model for faster troubleshooting from fault detection to fault isolation. Network Telemetry streams data about network state, including both underlay and overlay network statistics, to applications from Splunk, ExtraHop, Corvil and Riverbed. With critical infrastructure information exposed to the application layer, issues can be proactively avoided.

Openflow and Directflow

Arista EOS supports OpenFlow 1.0 controlled by OpenFlow controllers for filtering and redirecting traffic. Arista EOS also supports a controller-less mode relying on Arista's DirectFlow to direct traffic to the SDN applications (for example, TAP aggregators). This lets the production network run standard IP routing protocols, while enabling certain flow handling to be configured programmatically for SDN applications.

Arista EOS: A Platform for Stability and Flexibility

The Arista Extensible Operating System, or EOS, is the most advanced network operating system available. It combines modern-day software and O/S architectures, transparently restartable processes, open platform development, an un-modified Linux kernel, and a stateful publish/subscribe database model.

At the core of EOS is the System Data Base, or SysDB for short. SysDB is machine generated software code based on the object models necessary for state storage for every process in EOS. All inter-process communication in EOS is implemented as writes to SysDB objects. These writes propagate to subscribed agents, triggering events in those agents. As an example, when a user-level ASIC driver detects link failure on a port it writes this to SysDB, then the LED driver receives an update from SysDB and it reads the state of the port and adjusts the LED status accordingly. This centralized database approach to passing state throughout the system and the automated way the SysDB code is generated reduces risk and error, improving software feature velocity and provides flexibility for customers who can use the same APIs to receive notifications from SysDB or customize and extend switch features.

Arista's software engineering methodology also benefits our customers in terms of quality and consistency:

- Complete fault isolation in the user space and through SysDB effectively convert catastrophic events to non-events. The system self-heals from more common scenarios such as memory leaks. Every process is separate, with no IPC or shared memory fate-sharing, endian-independent, and multi-threaded where applicable.
- No manual software testing. All automated tests run 24x7 and with the operating system running in emulators and on hardware Arista scales protocol and unit testing cost effectively.
- Keep a single system binary across all platforms. This improves the testing depth on each platform, improves time-to-market, and keeps feature and bug resolution compatibility across all platforms.

EOS provides a development framework that enables the core concept of Extensibility. An open foundation, and best-in-class software development models deliver feature velocity, improved uptime, easier maintenance, and a choice in tools and options.

Arista EOS Extensibility

Arista EOS provides full Linux shell access for root-level administrators, and makes a broad suite of Linux based tools available to our customers. In the spirit of 'openness' the full SysDB programming model and API set are visible and available via the standard bash shell. SysDB is not a "walled garden" API, where a limited subset of what Arista uses is made available. All programming interfaces that Arista software developers use between address spaces within EOS are available to third party developers and Arista customers.

Some examples of how people customize and make use of Arista EOS extensibility include:

- Want to back up all log files every night to a specific NFS or CIFS share? Just mount the storage straight from the switch and use rsync or rsnapshot to copy configuration files
- Want to store interface statistics or LANZ streaming data on the switch in a round-robin database? Run MRTG right on the switch.
- Like the Internet2 PerfSonar performance management apps? Just run them locally.
- Want to run Nessus to security scan a server when it boots? Create an event-handler triggered on a port coming up.
- Using Chef, Puppet, CFEngine or Sprinkle to automate your server environment? Use any or all of these to automate configuration and monitoring of Arista switches too.
- Want to PXE boot servers straight from the switch? Just run a DHCP and TFTP server right on the switch.

If you're not comfortable running code on the same Linux instance as what EOS operates on we allow guest OSs to run on the switch via KVM built in. You can allocate resources (CPU, RAM, vNICs) to Guest OSs and we ship switches with additional flash storage via enterprise-grade SSDs.

Other Software Defined Cloud Networking (SDCN) Technologies

In addition to the EOS foundation technologies outlined, Arista Software Defined Cloud Networking (SDCN) incorporates various other technologies that enable scale-out automated network designs. Some of these other technologies include:

- Advanced Event Monitoring (AEM)
- Automated Monitoring/Management
- Arista CloudVision

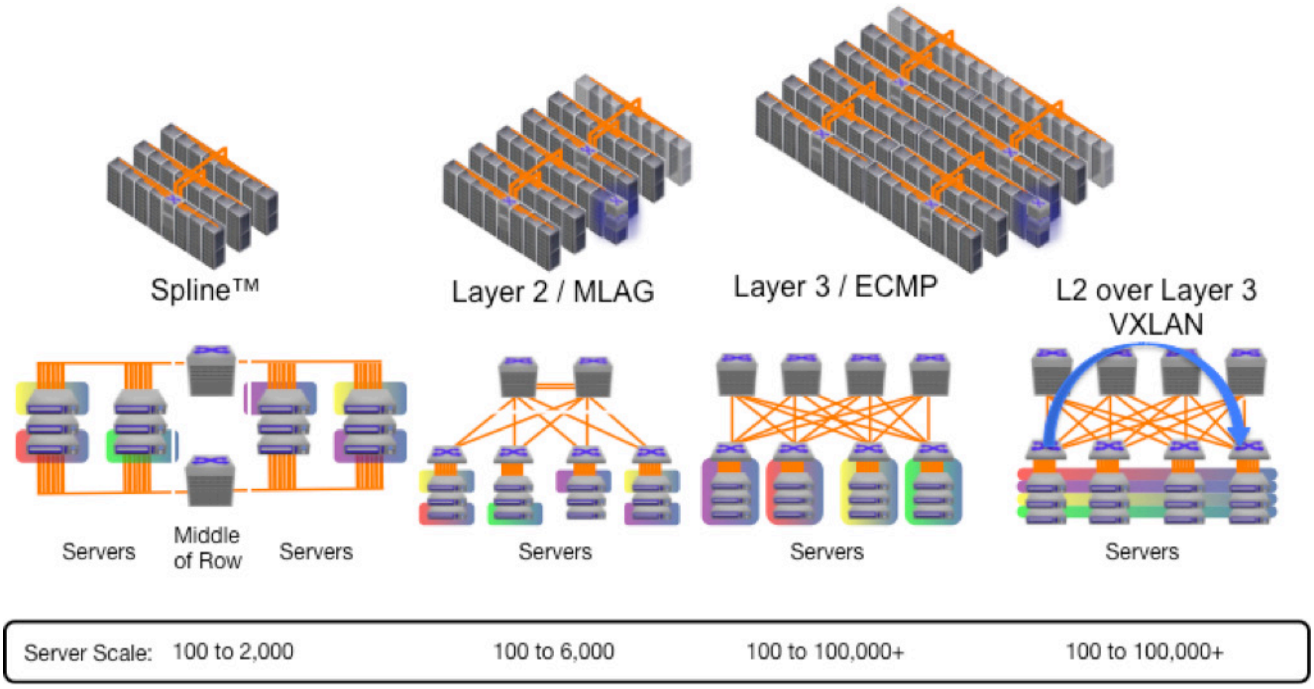

Figure 3: Arista Cloud network designs: Single tier SplineTM and Two Tier Spine/Leaf, 100 to 100,000+ ports

Conclusion

Arista's cloud networks take the principles that have made cloud computing compelling (automation, self-service provisioning, linear scaling of both performance and economics) and combine them with the principles of Software Defined Networking (network virtualization, custom programmability, simplified architectures, and more realistic price points) in a way that is neither proprietary nor a vendor lock-in.

This combination creates a best-in-class software foundation for maximizing the value of the network to both the enterprise and service provider datacenter: a new architecture for the most mission-critical location within the IT infrastructure that simplifies management and provisioning, speeds up service delivery, lowers costs and creates opportunities for competitive differentiation, while putting control and visibility back in the hands of the network and systems administrators.

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A & B, 11th Floor
Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard
#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062