



WHITE PAPER



10 Gigabit Ethernet Fabric Delivers High Performance at Sandia National Labs

Results from Sandia CBench benchmark tests demonstrate higher throughput using Dynamic Congestion Avoidance from Fortinet compared to statically routed InfiniBand

Overview

Designing and building HPC clusters can be a difficult task given the necessity for cluster architects to balance ease of construction and maintainability against the compelling need to achieve scalable applications performance. Given such constraints, it is generally accepted within the cluster community that the simplest way to build clusters is to use Ethernet as the underlying network; both for cluster management and for any application-level interconnect solution.

While Ethernet currently remains the only viable solution for cluster management infrastructure, its historical lack of performance has forced cluster builders to seek alternative solutions for the interconnect layer. Following current common practice for cluster design, most cluster architects consider InfiniBand to be the automatic choice for an interconnect solution when designing cluster systems that require high bandwidth and low latency in the MPI stack.

However, InfiniBand can often suffer performance deterioration due to its inability to adaptively manage network traffic, and in this case study we demonstrate that it is fairly easy to generate conditions in an InfiniBand network, using Sandia's CBench test suite, where network congestion manifests itself in a significant way. Moreover, by performing tests on a relatively small system of 128 nodes, it is apparent that the effects of static routing congestion are not limited to large clusters, but affect even moderately sized systems.

Congestion problems are overcome in a 10 Gigabit Ethernet fabric using vScale™ technology from Fortinet that features Dynamic Congestion Avoidance. By monitoring one-way latencies within the fabric, Dynamic Congestion Avoidance continuously assesses the current state of the fabric, detects conditions leading to congestion, and actively reroutes traffic away from active paths in the fabric to less busy paths. In this study, we run the same CBench tests on the same 128node cluster, first using InfiniBand and then through a FortiSwitch-1000 Ethernet Fabric Switch, where we find that congestion is almost entirely eliminated.

These results show that Dynamic Congestion Avoidance satisfies both cluster management and application-level interconnect requirements.

Dynamic Congestion Avoidance

Fortinet has a unique solution for handling congestion in 10 Gigabit Ethernet fabrics. Fortinet removes the limitation of single path routing imposed by the use of the spanning tree algorithm, employing its vScale technology to build multistage switching fabrics, or multipath routed networks.

Using probe packets, Dynamic Congestion Avoidance measures one-way network latencies between any two-edge ports in the fabric, and hence detects traffic buildup or congestion. Dynamic Congestion Avoidance proactively reroutes traffic, directing it in a load balancing fashion, away from congested paths to alternate faster paths in order to relieve congestion in the fabric and maximize overall throughput. Moreover, it does this at low latency and high rate: in a single 144port FortiSwitch-1000 platform, the highest switch latency between any two ports is 1.6 μ S. Fortinet switches can also be combined to form larger single fabrics, at which the highest latency between any two points on that fabric rises to only 6.4 μ S.

CBench Rotate Tests

As one of the world's major users of HPC technology and of large scale clusters in particular, Sandia National Labs continually evaluates the performance of production systems for use in solving the scientific problems it has been charged to investigate. Given wide-ranging application requirements, the task of determining which cluster systems fit particular applications is difficult, especially when technologies change and evolve as rapidly as they do in the cluster space. To aid this effort, Sandia has developed an extensive suite of benchmarks called CBench to help determine systems performance over a wide range of architectural and computational scenarios.

For the tests outlined in this document, we focus on the CBench rotate tests.

The network congestion we characterize in this exercise is a condition that is usually associated with Ethernet networks, but affects even statically routed, full bisection networks such as InfiniBand and Myrinet. Static routing enforces a single route between any two given nodes, which is then followed by communications between those two nodes. In essence, a statically routed network consists of a set of fixed routes between nodes. Given that almost all interconnect switch networks or fabrics are built as multistage networks, we can find many independent routes that share stages with each other. It is this property that is the underlying cause of any congestion we may see.

Using the CBench rotate tests, we have a way to exercise underlying network infrastructure at an application level, rather than a system level, and expose any associated performance issues, specifically those issues related to false congestion caused by routing in a given network fabric.

The CBench rotate tests are derived from the com and laten tests, which are part of LLNL's Presta benchmark suite. These tests are very simple: they test bandwidth and latency of MPI-based communications between pairs of nodes connected through an interconnect. It is possible to specify multiple pairs of nodes in these benchmarks in order to see how they might interfere with each other when communicating concurrently, and this would highlight some important limitations

of the underlying infrastructure. For example, when running com on multiple pairs of nodes in a standard Ethernet switch, one might see an aggregate bandwidth measurement that is smaller than the measurement for a single pair multiplied by the number of pairs in the test. The reason for that performance degradation would be either congestion or a switch limitation.

The tests rotate and rotate_latency take this notion further by dynamically changing which nodes form pairs, so that the traffic originating from a node is forced to take different routes through the fabric. The aim of this is to take into account the effects and overhead of dynamic or multipath routing. This is essential in establishing guidelines for estimating the performance of MPI-based applications that strongly leverage collective communications such as all-to-all and reductions.

Performance Testing

In testing of this kind, one of the things we are most interested in discovering is how such benchmarks respond to scaling, even in the confines of the relatively small cluster we used to perform the tests. This is important because the switch fabrics we tested are not constructed from large monolithic switches, but are instead constructed from many smaller switches combined together in a FatTree or Clos architecture. In the case of both the InfiniBand switch we tested, and the Fortinet switch, the smallest switch chip is a 24port unit.

It should be noted that at the port layer, any given switch must divide its network links between edge ports that connect to the outside network or downlinks in the internal network, and uplink ports that connect internally to other switch chips in order to form a larger network or fabric. This ratio is generally 1:1 between downlinks and uplinks at the edge, but can be uneven (3:2 or 2:1) at higher levels of the network fabric.

In practical terms, this means that for any problem size greater than the number of downlinks on an edge switch chip, the problem must populate more than one switch chip. Therefore, the shortest path between some pairs of nodes must cross at least one switch boundary, if not more.

This need to direct traffic between switch chips is at the root of the performance issues we examine in this study.

Strict and relaxed ordering of iWarp packets

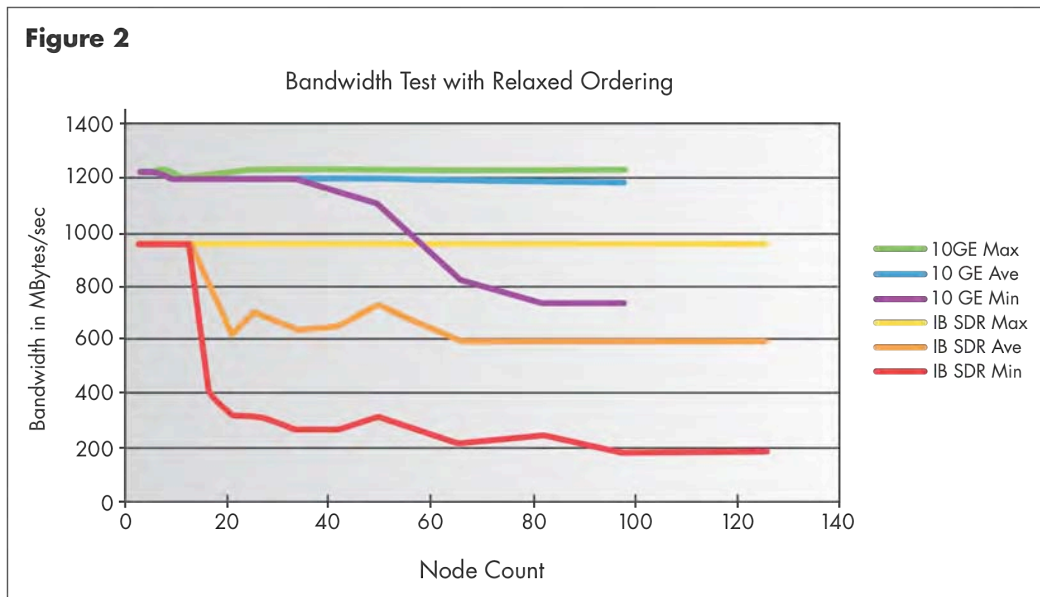
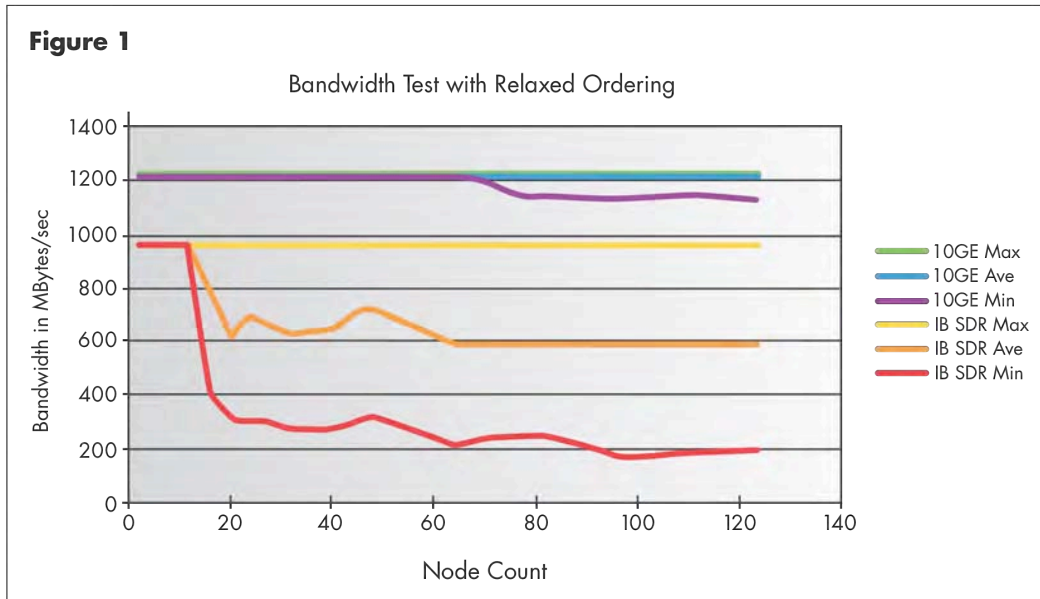
The CBench benchmarks are constructed with MPI and the OFED 1.2 stack. For the OFED stack, the MPI layer is built over a lower level protocol called iWarp, which delivers low latency and high bandwidth using the TCP/IP transport protocol over Ethernet.

One of iWarp's important features is outoforder messaging. An iWarp-enabled device is capable of receiving a packet divided into subpackets at the source, arriving out of sequence, and will combine them at the receiving end to form a complete packet. This feature is called relaxed ordering. Forcing the receiver to read subpackets in sequence is called strict ordering. The benefit of relaxed ordering is that the network card can start sending subpackets for the whole packet without waiting, and therefore take advantage of more aggressive multipath routing.

The fact that a network card is iWarpenabled is transparent to the FortiSwitch-1000 switch which implements managed, multipath routing no matter what the ordering.

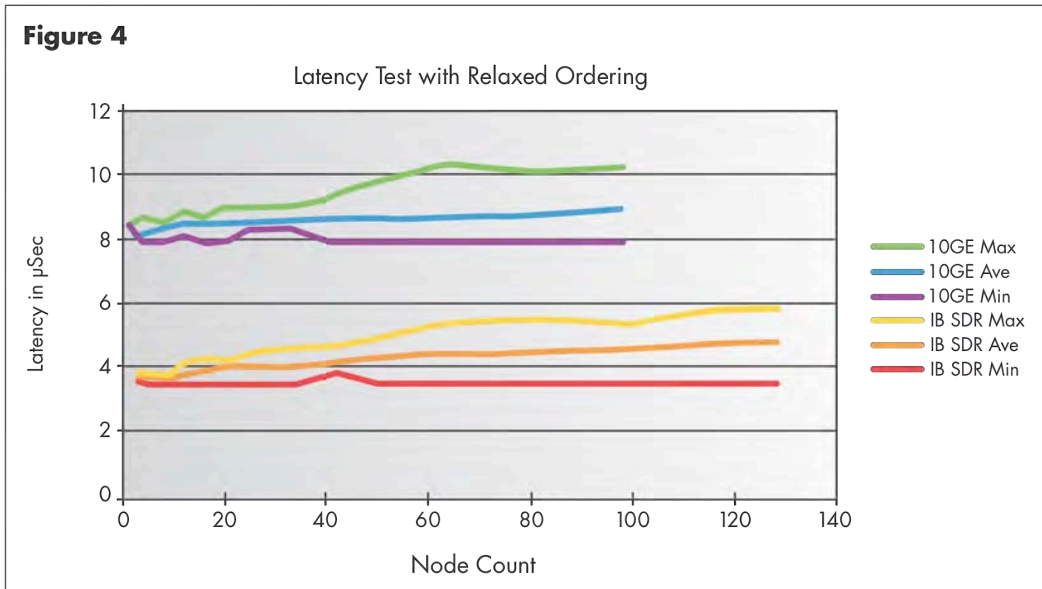
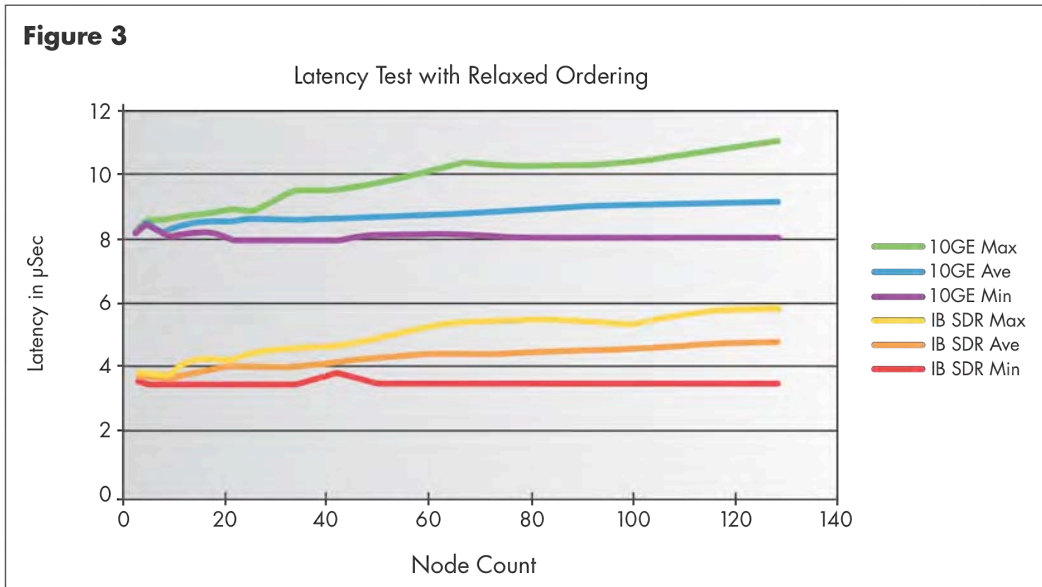
Bandwidth test results

The following figures illustrate results of bandwidth tests, graphed to compare bandwidth in MB/sec with node count. Each figure shows both InfiniBand (IB) network results and 10 Gigabit Ethernet (10GE) network results.



Latency test results

The following figures illustrate results of latency tests, graphed to compare latency in μs with node count. Each figure shows both InfiniBand (IB) network results and 10 Gigabit Ethernet (10GE) network results.



Analyzing the Results

When we examine the results for `rotate_latency`, we find that the changes in latency as the test scales to more nodes are superficially very similar for 10 GE and InfiniBand. However, even at this small scale, significant differences in scalability can be discerned: by comparing average latency for problem sizes of 2 and 124 nodes, for 10 GE (Figure 3), this measurement increases from 8.0 μ S to 9.2 μ S (15%), whereas for InfiniBand, the change is from 3.5 μ S to 4.75 μ S (36%).

As a percentage, the increase in average latency time is over twice for InfiniBand than it is for 10 GE. While this isn't a significant issue for the cluster size we're using for testing, it's a strong indicator that we can expect to see a substantial decline in performance and scalability for InfiniBand at larger cluster sizes due

to issues caused by routing. Other studies¹ have demonstrated that for large-scale clusters using InfiniBand performance degradation in short packet communication becomes substantial.

When we look at the rotate bandwidth test (Figure 1), things look dramatically different. In this case there is a substantial drop in worst case and average bandwidth for InfiniBand above 12 nodes, to the point that at 64 nodes, the average bandwidth is approximately 55% of the peak bandwidth. The 12node threshold is significant as it represents the point at which the benchmark test exceeds a single switch chip, and makes clear that the congestion we see is indeed caused by static routing. For 10 GE, by contrast, we see that the best case and average bandwidth results remain very close, all the way to 124 nodes, with no significant reduction in performance, and that Fortinet's Dynamic Congestion Avoidance technology is more than capable of counteracting the deficiencies of static routing.

The rotate test also shows the benefit of Dynamic Congestion Avoidance when testing the strict and relaxed ordering policies for iWarp. For `rotate_latency`, changing the policy from strict to relaxed makes no difference, as each sent packet is small, and therefore is transmitted intact. But for rotate, traffic is dominated by large packets, which get broken into smaller sub-packets by the sender. For 10 GE, we see a drop in worst-case bandwidth numbers for the strict ordering case starting at around 40 processes, and a significant drop at 64 nodes. But the average bandwidth stays very close to the best bandwidth result, suggesting that worst-case instances are infrequent enough to have no bearing on overall benchmark performance. If we then adopt the relaxed ordering policy, the drop in performance from best case to worst case almost vanishes. This specific test demonstrates the ability of Dynamic Congestion Avoidance to detect and react quickly to network traffic, and to dynamically load balance traffic across multiple paths to achieve the highest possible throughput.

Implications for Real-World applications

While synthetic benchmarks are useful in highlighting certain features or considering specific issues, it is crucial that the data and knowledge extracted from the process can be applied to end-user applications.

It is clear that the performance degradation described above applies to a large range of end-user scenarios. In fact, it might be easier to point out the cases that are immune. Regular HPC problems that use fixed, low path-count routing such as simple multi-grid and finite difference solvers, or some simple Monte Carlo based applications won't often encounter this problem. Almost all other HPC problems will be affected to a lesser or greater extent.

For parallel applications that use algorithms such as FFTs or matrix solvers, which handle collective or global communications, the susceptibility to performance degradation is clear. Because the MPI topology of those applications is essentially an all-to-all ring, all paths are equal, no single path between two nodes will dominate, and single-path statically routed networks such as InfiniBand will not adaptively route.

A similar problem occurs in a second class of applications, where the MPI topology changes with solver phases within the application, even though each topology instance is easily mapped onto a statically routed network. There are many codes like these, the most important of which are crash simulation codes (such as Dyna3D), and multi-physics codes like those used to model nuclear reactors and aircraft/automobile engines.

The most important end-user scenario is also the most widely found: the use of general-purpose production clusters. In a production cluster, multiple users run many application instances on subsections of the cluster. Even though individual instances may be well behaved, and, in theory, independent from each other, when they are run in concert on a cluster these instances tend to interfere with each other because they share resources such as the network switch. This situation is exacerbated by the addition of cluster-wide file I/O. The overall effect is that the traffic in the cluster is significantly more irregular compared to when the cluster runs a single instance of an application, and it is this irregularity that causes congestion.

Conclusion

We have shown congestion to be a significant issue for InfiniBand; even at the small scale at which testing has occurred. It is important to note that the conditions that lead to this kind of congestion are very common in HPC, which suggests that this issue cannot be avoided and that its consideration is fundamental to the design of cluster systems. Moreover, as interconnect fabrics get bigger, the number of stages in the fabric also increases, which suggests, in agreement with common anecdotal evidence, that the effects of congestion will become more pronounced at scale, rather than less. This congestion establishes an upper limit to scalability in those fabrics, and that, in turn, limits the size of problems that can be tackled and limits the kind of science that can be studied.

By using Dynamic Congestion Avoidance based on vSCALE technology from Fortinet, we can almost completely eliminate congestion caused by static routing, and by doing so we show that it is fairly easy to achieve better performance and scalability in a 10 GE fabric compared to InfiniBand.

For more information about Fortinet interconnect technologies, please contact your Fortinet sales representative.

System Configuration

The tests were performed on the Talon cluster at Sandia National Labs, which is a 128-node Linux cluster consisting of the following hardware and software:

Compute Node

- Dell 1850, dual socket Xeon 3.60 GHz, 6 GB Memory – Operating system CentOS 4U3
- Linux kernel version: 2.6.942.0.2.ELsmp
- 10 Gigabit Ethernet – Chelsio S310ECX, Copper CX4 interface
- Firmware Version: T 4.0.0 • Driver Version: 1.0.087
- InfiniBand – Mellanox InfiniHost SDR network card
- IBGold 1.83 driver and firmware stack • MPI layer: OFED 1.2pre

Note that the OFED layer used for the testing is a prerelease version of OFED 1.2 containing a number of fixes made by the team at Sandia National Labs. Not all of these fixes were included in the OFED 1.2 release, but are expected to be included in version 1.2.1, or whatever the next bug fix release is called. Also note that OFED 1.2 contains low-level driver software for InfiniBand and Chelsio 10 GE, which is not included in the standard driver releases. This includes the RDMA iWarp layer used in this testing exercise.

10 Gigabit Ethernet Switch

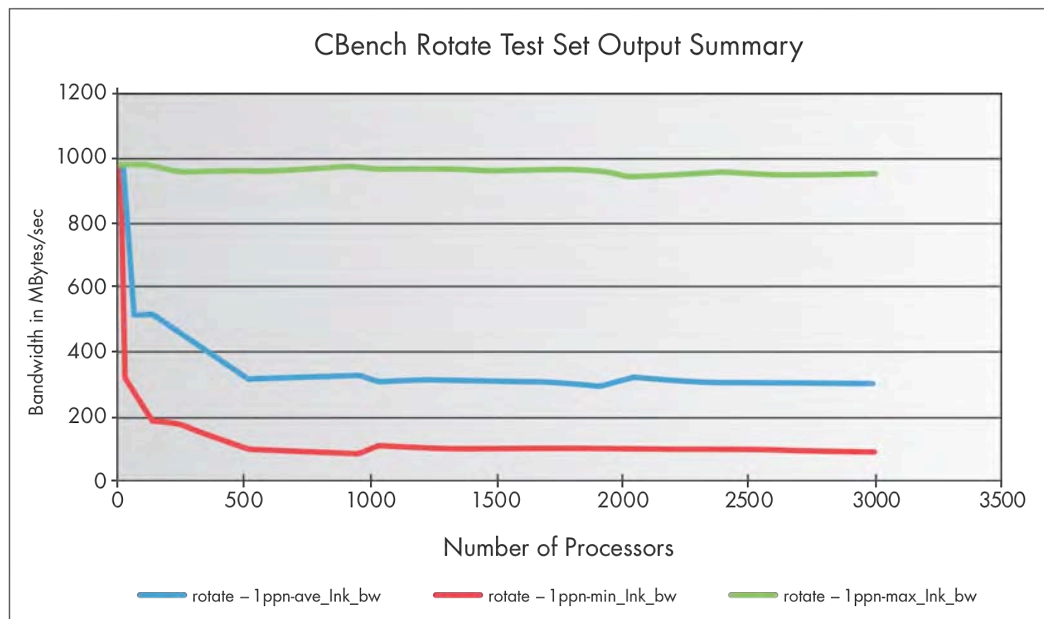
Fortinet FortiSwitch-1000 Ethernet Fabric Switch, 144 ports. Copper CX4 interface

InfiniBand Switch

Topspin TS740, 288 ports

Appendix: Large Scale Effects

The following graph, extracted from the presentation by Leining and Seager², shows the effects of the congestion we've discussed in this paper, at a larger scale. The bandwidth test shows the degradation of average bandwidth from around 600 Mbytes/sec at 128 nodes, to 300 Mbytes/sec at 500 nodes and beyond. This represents a significant obstacle to the kind of applications scalability needed if today's important HPC problems are to become tractable.



FortiGuard® Security Subscription Services deliver dynamic, automated updates for Fortinet products. The Fortinet Global Security Research Team creates these updates to ensure up-to-date protection against sophisticated threats. Subscriptions include antivirus, intrusion prevention, web filtering, antispam, vulnerability and compliance management, application control, and database security services.

FortiCare™ Support Services provide global support for all Fortinet products and services. FortiCare support enables your Fortinet products to perform optimally. Support plans start with 8x5 Enhanced Support with “return and replace” hardware replacement or 24x7 Comprehensive Support with advanced replacement. Options include Premium Support, Premium RMA, and Professional Services. All hardware products include a 1-year limited hardware warranty and 90-day limited software warranty.

FORTINET

GLOBAL HEADQUARTERS

Fortinet Incorporated
 1090 Kifer Road, Sunnyvale, CA 94086 USA
 Tel +1.408.235.7700
 Fax +1.408.235.7737
www.fortinet.com/sales

EMEA SALES OFFICE – FRANCE

Fortinet Incorporated
 120 rue Albert Caquot
 06560, Sophia Antipolis, France
 Tel +33.4.8987.0510
 Fax +33.4.8987.0501

APAC SALES OFFICE – SINGAPORE

Fortinet Incorporated
 61 Robinson Road, #09-04 Robinson Centre
 Singapore 068893
 Tel +65-6513-3730
 Fax +65-6223-6784

Copyright© 2009 Fortinet, Inc. All rights reserved. Fortinet®, FortiGate®, and FortiGuard® are registered trademarks of Fortinet, Inc., and other Fortinet names herein may also be trademarks of Fortinet. All other product or company names may be trademarks of their respective owners. Performance metrics contained herein were attained in internal lab tests under ideal conditions. Network variables, different network environments and other conditions may affect performance results, and Fortinet disclaims all warranties, whether express or implied, except to the extent Fortinet enters a binding contract with a purchaser that expressly warrants that the identified product will perform according to the performance metrics herein. For absolute clarity, any such warranty will be limited to performance in the same ideal conditions as in Fortinet's internal lab tests. Fortinet disclaims in full any guarantees. Fortinet reserves the right to change, modify, transfer, or otherwise revise this publication without notice, and the most current version of the publication shall be applicable. Certain Fortinet products are licensed under U.S. Patent No. 5,623,600.